OXILATE

Operational eXcellence by Integrating Learned information into AcTionable Expertise

Labeled in ITEA3, a EUREKA cluster, Call 5

ITEA3 Project Number 18023

<D1.3 – Overview of the state of the art

Due date of deliverable: Feb 2021
Actual date of submission: 28-Feb-2021

Start date of project: 1 March 2020	**Duration:** 36 months
Organisation name of lead contractor for this deliverable:	SII Concatel
Author(s):	Elio Saltalamacchia et al.
Status:	Final
Version number:	V1.0
Submission Date:	28-Feb-2021
Doc reference:	OXILATE _D1.3_Overview of the state of the art_V1.0.docx
Work Pack. / Task:	WP1
Description: ***(max 5 lines***	D1.3 will deliver a document highlighting the current state of the art in the technologies that will be investigated and implemented throughout the course of the Oxilate project, namely Machine Learning, Distributed Ledger Technology, Data Management, Intelligent Digital Twins, Generative Engineering, and Mobile / Wearable technology as well as an overview of the applicable regulations and security concerns facing the Oxilate.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Nature:	<Use one of these codes: **R**=Report, **P**=Prototype, **D**=Demonstrator, **O**=Other>		
Dissemination Level:	**PU**	Public	
	PP	Restricted to other programme participants	
	RE	Restricted to a group specified by the consortium	
	CO	Confidential, only for members of the consortium	**X**

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
V0.1	23/12/2020	First draft	Draft	all@oxilate.eu
V0.2	20/01/2021	2nd draft	Draft	all@oxilate.eu
V1.0	28/02/2021	Approved by PMT, to be submitted to ITEA3	Final	all@oxilate.eu

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Table of Contents

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

1. Glossary

AI	Artificial Intelligence
AR	Augmented Reality
BI	Business Intelligence
CAE	Computer-Aided Engineering
COTS	Commercial Off-The-Shelf
CPS	Cyber Physical System
DCS	Distributed Control Systems
DevOps	Development + Operations
DLT	Distributed Ledger Technology
ECM	Enterprise Content Management
ETL	Extract – Transform - Load
IGT	Imaging Guided Therapy
IoT	Internet of Things
ML	Machine Learning
NLP	Natural Language Processing
OPEX	Operation Expenses
PLM	Product Lifecycle Management
PoC	Proof of Concept
SLA	Service Level Agreement
SIEM	Security Information and Event Management
SVM	Software Vulnerability Management
SWO	Service Work Order
TTM	Time To Market
UI	User Interaction
VR	Virtual Reality
XR	eXtended Reality

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

2. Executive Summary

2.1 Work Package Objectives

Work Package 1 concerns the use case definition and demonstration of the Oxilate project. The project's technical work packages (WP 2, WP3, and WP 4) will be defined by the industrial cases investigates in WP1. The large industrial partners will set out use cases, such as Industry 4.0 process automation, insurtech and CAE/PLM software tools, with which the tool providers will formulate the requirements and plan of approach for their research and development, thus, setting mutual expectations for the project.
This work package anticipates the development of several tangible demonstrators which will be used to align collaborations throughout the project. This includes preliminary demonstrators explaining the ideas set forward, intermediate demonstrators where initial project progress is shown and finally a set of demonstrators in which the project results are integrated into products where feasible.
Finally, WP1 will explore new business models and additional service offerings based on the insights gained throughout the Oxilate project.

2.2 Scope of Deliverable

D1.3 will deliver a document highlighting the current state of the art in the technologies that will be investigated and implemented throughout the course of the Oxilate project, namely Machine Learning, Distributed Ledger Technology, Data Management, Intelligent Digital Twins, Generative Engineering, and Mobile / Wearable technology as well as an overview of the applicable regulations and security concerns facing the Oxilate.

Each partner will contribute with an overview of the current state of the art of each technology relevant to their development of the project, including where the technology stands in terms of research and development and, where it is relevant, key developments, breakthroughs or innovations. As each topic is immense and volumes of information could be produced, each partner will focus solely on developments considered relevant to the Oxilate project.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

3. Introduction

3.1 Key Outcomes of OXILATE

OXILATE is the successor of the successful ITEA 14035 REFLEXION project which supported a revolutionary change in the way of working of the high-tech systems industry's R&D by introducing and integrating widespread available data analytic solutions from the open source / data science communities. OXILATE focuses on the complementary integration of expert knowledge to develop widely available support and tools for professionals with the objective of empowering them to transform their respective business activities, making them more proactive and effective, and to create direct business value over the whole product life-cycle they serve.

Oxilate is focused on providing support for well-performing systems fully integrated into the operational workflow through the development and use of multiple "actionable" (independent) tools, such as data-driven tools, methods, processes, models and / or platforms which adopt the principle of integrating expert knowledge with data analysis on operational data. The expected result of Oxilate will be the empowerment of professionals to be more proactive and effective through the use of "actionable" tools, transforming business processes into their respective phase of the product life cycle. The expected results of the project can be classified as technical, application oriented or business oriented.

3.1.1 Technical

The technical outcome of Oxilate will include 'Actionable' tools which help OEMs to improve their business activities. This can include, for example, supporting products in service and thereby accelerating their competitiveness. For SMEs this will involve targeting improvements of their technology platforms and service proposition portfolios will them a competitive lead in their market.

- Coordinated data and knowledge platforms supporting business processes in all phases of their life-cycles with key differentiators: heterogeneous, flexible, adaptable (static and at 'run-time'), data and insight driven, scalable, IT-infrastructure independent, easy useable by 3rd parties by intent-based (cloud-based) APIs
- Digital twin technology concepts for control systems mobilizing heterogeneous expert- and data-driven knowledge – *Valmet Automation, University of Oulu, M-Files, Intopalo Digital, Atostek,* , TurkTraktor and IND Bilisim.
- Data & insight driven proactive diagnostic / root cause analysis tools integrating 3rd line support knowledge and augmented by AI / ML models. *SII Concatel, TyP*
- Decentralized blockchain-based AI agent networks enabling smooth communication to the customers providing fully customized, personalized services accommodating rapid changing customer life-styles – *II-Broker, SII Concatel, TyP*
- Intelligent information management platforms connecting systems with structured and unstructured data utilizing a metadata-driven architecture, including features for automatic extraction and enrichment of metadata, e.g. with AI technologies – *M-Files, University of Oulu*
- Various service technology platform offerings with customer-oriented smart features in the field of expert knowledge capturing, diagnostics, root cause analysis, BI, visual conformance checking, process prediction, smart monitoring, smart reporting, operation & services monitoring and prediction, multi-layered user interfaces for business data visualization – *Atostek, TyP, Turkgen, Semantik*

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

3.1.2 Application

Targeting applied methodology innovations are to be rolled out in the broader European industrial ecosystem.

- Digital user assistance integrating expert knowledge and user interaction data to provide autonomous expert-level support– Siemens Industry Software, SII Concatel, Turkgen, Semantik
- Service (positioning) discovery mechanisms to identify new business opportunities based on the analysis of use data – Valmet Automation
- Systematics to support technical departments of strawberry farms to keep their fleet of agricultural robots up and running – Octinion
- Insights on mobilizing heterogeneous knowledge through digital twins in industry contexts - University of Oulu, Istanbul Okan University (subcontracted by TurkTraktor)

3.1.3 Business

The business-related outcomes of Oxilate will be related directly to business opportunities and have been defined by the consortium partners as:

- Improved insights in the company service portfolios, expanding the service offerings related to the digital transformation –*Valmet Automation*
- Manufacturing costs improvements via better production planning, raw material yield, energy efficiency, timely maintenance – *TurkTraktor*
- Enrichment of current business processes and consolidation of new, enhanced, short-cycle services – *ii-Broker, IND Bilisim*
- Enhancements of the consulting, development, and solution portfolios – Atostek, Intopalo Digital, M-Files, Octinion, TyP, Siemens Industry Software, SII Concatel, Valmet Automation, Turkgen, Semantik

3.1.4 Societal

It is envisaged that one outcome of Oxilate project will be the creation of high-quality jobs to adopt changes in the way we work, especially in terms of developing ‘actionable’ tools which integrate data analytics on operational data and expert knowledge.

3.2 Use case overview

The goal of OXILATE is to support professional customer-oriented business activities, to deal with customer demands typically claiming costly and scarce R&D assets, this will lead to a new set of data- and knowledge-driven methods, techniques and business processes providing key advantages for European high-tech industry in exploring and improving new opportunities for existing products and services, as well as supporting the development of new business propositions.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

3.2.1 Belgium

The Belgian/Flemish consortium consists of Siemens Industry Software NV (SISW; large industry) and Octinion (SME), the former being an engineering software tool provider, the latter being an innovative company in mechatronic product development. Together, SISW and Octinion represent two essential elements in the value chain of mechatronic system design and operations. While being on different ends of this value chain, both partners identify a common need for research and innovation into data-driven tools that, by combining expert R&D knowledge with operational data, are capable of drastically lowering the required expertise level of the (unexperienced) operator (simulation model developer; service technician) to perform his/her daily activities.

Succeeding in defining these tools will assist operators in solving a much wider set of complex problems, hence drastically lowering the need for costly and scarce senior R&D expertise to solve day-to-day problems. These topics fall neatly within the broader scope of topics addressed by the Oxilate project.

3.2.2 Finland

Finnish IT industry seeks constantly new markets for intelligent knowledge creation and management, mobilization, and digitalized service solutions to support a wide variety of high-tech system industries (such as industrial plants, industrial automation, telecommunications, etc.) and related export-ready services in international business ecosystems. Seeking growth in the existing and emerging international markets of high-tech systems and related service business requires ever more competitive and innovative digital services to support the system life-cycle services.

The OXILATE project develops data-driven tools, methods, processes and, altogether, actionable toolboxes which incorporate expert knowledge of high-tech systems with data analytics on operational data, to empower professionals in their respective business activities, such as field service support, predictive diagnostics, and training.

The Finnish consortium will contribute to the OXILATE project by developing means to mobilize knowledge for field services of high-tech systems. The public funding is required to boost the consortium's co-operation, both in the existing Finnish ecosystem and with the other international partners, to research exportable service concepts and solutions.

3.2.3 Spain

The goal of OXILATE is to support professional customer-oriented business activities and to deal with customer demands typically claiming costly and scarce R&D assets. This will lead to a new set of data- and knowledge-driven methods, techniques and business processes providing key advantages for the European high-tech industry in exploring and improving new opportunities for existing products and services as well as supporting the development of new business propositions.

The fundamental challenges highlighted in the Oxilate project are echoed in the InsureTech sector: Increase of complexity due to changing business models; Need for evolvability; Globalization of the supply chain and Emerging data. The Spanish consortium will have a key focus on the InsureTech

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

sector and on the creation of a new AI/ML-based Predictive Product Design and Support system based on efficient services assimilating data from diverse sources: operational data, CRMs and Customer Support Services. Agile development of new wearable and smartphone-based solutions and services will be supported by the system and new AI/ML based services and tools will be employed to retrieve key data throughout the product lifecycle.

3.2.4 Turkey

Turkish consortium also will continue to work on new ICT solution in OXILATE project to serve in a manner to support the goods or services of the lcoal companies by using high-tech software solutions with innovative approach. Turkgen a bid-data analyse company, in cooperation with Semantik, a NLP company, will collaborate to deliver a new solution which is unique in local and international markets.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

4. Machine Learning - NLP

4.1 Intro to SotA ML (NLP)

This section takes a look at Neural Networks as they apply to the OXILATE project, I.e. a look at the state of the art of NN for NLP. This section looks at the progression of NN, starting with RNNs, then a look at LSTMs and investigating the latest in Transformers and attention mechanisms.

4.1.1 NLP State of the Art

Convolutional Neural Networks (CNNs)

In essence, a CNN is a deep learning network which can take an image as an input, assign and importance to aspects of/in that image using weights and biases and then assign a probability as to the classification of that object. They are designed to be able to handle image processing much more efficiently than other NNs. CNNs are adept at classifying images by using convolutions in 3 dimensional space to detect features and then assign probabilities as to what the detected images are/belong to. They typically work by applying filters, known as Kernels, to an images by sliding the filter across the images in the form of a sliding window. The filter is then used on the neighbouring area, allowing the network to learn features from comparing data with neighbouring nodes. CNNs are said to work in 3D space as the usually take as an input the Width, Height and Depth of the image being analysed – the depth in this instance being the colour channels (Red, Green and Blue); so a 32x32 pixel image in RGB would be represented as 32x32x3 (Stanford University, 2015 - 2019).

Research continued but commercial applications were not brought to market

Recurrent Neural Networks (RNN's)

Neural Networks were first described in terms in the field of Neuroscience as early as 1943 by Warren S. McCullough and Walter H. Pitts [1]. The pair were attempting to show that the human brain can be considered a computing device and in doing so, started the cybernetics movement, creating the McCulloch-Pitts Neuron. This was an important step forward for the field of computer science as it essentially established the notions of "Finite Automata" and Logic Design [4], however the idea of Neural Networks fell out of popularity (perhaps because they were computationally less powerful than Turing Machines) and research dried up until it was restarted again in the 1980s.

Recurrent neural networks (RNNs) use time series data or sequential data to resolve ordinal or temporal problems. These types of problems are commonly tackled using NLP techniques and as such, RNNs have played a large part in the development of NLP. Tasks such as language translation, speech recognition, and image captioning all fall under the banner of natural language processing and are often solved using RNNs. In real life, we can find RNNs in applications such as voice activated assistants (Alexa, Siri, Google Assistant etc.) as well as in translation services (Translate and DeepL etc.)
Artificial neural networks first mimicked the features of mammalian

Recurrent neural networks utilize training data to learn, just as feedforward and convolutional neural networks (CNNs) do, however RNNs are distinguished by the fact that they have a "memory" built in.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Whilst traditional deep neural networks (DNN) work on the assumption that inputs and outputs are independent of each other, RNNs use information from previous inputs to influence current inputs and outputs. In short, this means that outputs in a RNN depend on the sequence of previous inputs.

Given this explanation, it is easy to see how RNNs have their application in NLP.

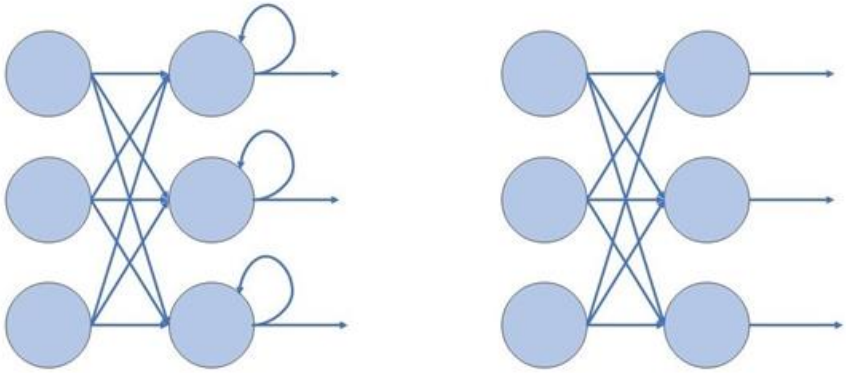


Figure 1: Comparison of Recurrent Neural Networks (on the left) and Feedforward Neural Networks (on the right) [16]

LSTM's

First described in a 1997 paper entitled "Long Short-term Memory" (Long Short-temr Memory, 1997) by Sepp Hochreiter and Jürgen Schmidhuber, Long Short-term Memory networks (called LSTMs) are a type of RNN with the ability to learn long-term dependencies. They were proposed as a way of overcoming the problem of vanishing or exploding gradients: i.e. that back-propagated error signals decay exponentially in the various network layers or they explode exponentially in the same (IDSIA) (Hochreiter, 1991)[1].

Unfortunately, the range of contextual information that standard RNNs can access is in practice quite limited. The problem is that the influence of a given input on the hidden layer, and therefore on the network output, either decays or blows up exponentially as it cycles around the network's recurrent connections. This shortcoming … referred to in the literature as the vanishing gradient problem … Long Short-Term Memory (LSTM) is an RNN architecture specifically designed to address the vanishing gradient problem. [20]

To this end, LSTMs are designed explicitly to avoid the long-term dependency problem as they have the ability to remember information for extended time periods.

LSTMs can be considered as more sophisticated RNNs which employ a gated mechanism and therefore gain the name, Gated Recurrent Networks. A more recent development in this field is the Gated Recurrent Unit (GRU), another type of RNN which is simpler than LSTMs, and are therefore quicker to train, but which offer similar performance.

An in depth description of the difference between GRUs and LSTMs is can be found in the 2014 paper, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modelling (Junyoung Chung, 2014)

[1] Diploma Thesis original in German, however the mathematics is international

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Graph Convolutional Networks (GCN)

Euclidian Geometry deals with the distance between points, for examples shapes in 2D space or 3D space. For many applications Euclidian Geometry works perfectly, for example in image processing where pixels in an image have concrete 2D coordinates. However in the real world there are many examples of non-Euclidian data which also has to be processed for successful machine learning cases. These data include structures such as meshes, point clouds, trees and graphs which can be found in social media, images, document classification, e-commerce etc. Trees in particular grow exponentially and are therefore non-Euclidian. Hierarchies are another data structures not represented in Euclidian space and therefore are not suitable for CNNs or RNNs. This is explained in great detail in two key papers from 2017: Poincaré Embeddings for Learning Hierarchical Representations [23] and Neural Embeddings of Graphs in Hyperbolic Space [24] which promoted the use of hyperbolic space for graph representation learning in ML [25].

Graph Neural Networks first came to prevalence in 1997 when authors Sperduti et al. applied neural networks to directed acyclic graphs by using a “so called "generalized recursive neuron", which is essentially a generalization to structures of a recurrent neuron” (Supervised neural networks for the classification of structures, 1997). Inspired by this work, many early studies helped define what became known as Recurrent Graph Neural Networks (RecGNN) which work by learning a target node’s representation though the iteratively propagating a neighbour’s information until a stable fixed point is reached. This method is however computationally expensive.

Inspired by the success of CNNs in computer vision, researchers applied the idea of convolutions to graph networks and developed the concept of Convolution Graph Neural Netowrks (ConvGNN). As with CNNs, Graph Convolutional Networks also use convolutions through the application of filters or kernels in the form of a set of weights, however unlike CNNs, GCNs are used for non-Euclidian or non-structured data.

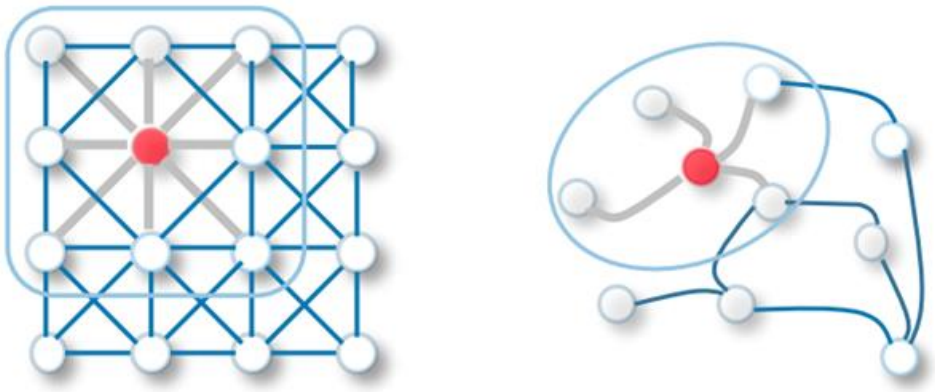


Figure 2:The left image shows a 2D Convolution. Analogous to a graph, each pixel in an image is taken as a node where neighbours are determined by the filter size. The 2D convolution takes the weighted average of pixel values of the red node along with its neighbours. The neighbours of a node are ordered and have a fixed size. The image on the right is a Graph Convolution. To get a hidden representation of the red node, one simple solution of the graph convolutional operation is

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

to take the average value of the node features of the red node along with its neighbours. Different from image data, the neighbours of a node are unordered and variable in size. [[27]

Graph autoencoders (GAEs) are a type of GNN and can be described as deep neural architectures which map nodes into a latent feature space and decode graph information from latent representations. GAEs have generally two key uses: learning network embedding and new graph generation. GAEs can learn the generative distribution of multiple graphs by encoding graphs into hidden representations and then decoding a graph structure given hidden representations, proposing a new graph in a sequential or global manner. Sequential approaches generate a graph by suggesting nodes and edges step by step whereas Global approaches output a graph all at once.

A final type of GNN found in the literature is the Spatial-temporal graph neural network (STGNN). STGNNs aim to learn hidden patterns from spatial-temporal graphs. STGNNs consider spatial dependency and temporal dependency at the same time, most often using graph convolutions to capture spatial dependency and RNNs or CNNs to model temporal dependency (A Comprehensive Survey on Graph Neural Networks, 2021). STGNNs find their applications in a number of applications, examples from the literature include forecasting (Diffusion convolutional recurrent, 2018), driver maneuver anticipation (patio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, 2018), as well as human action recognition (Spatial temporal graph convolutionalnetworks for skeleton-based action recognition , 2018).

For further reading on the timeline of GNN research with links to related papers, Wu et. Al provides the complete and comprehensive guide summarized in Table 1 with citations in Annex 1

Category		Publications
Recurrent Graph Neural Networks (RecGNNs)		Annex 1 – [1] – [4]
Convolutional Graph Neural Networks (ConvGNNs)	Spectral methods	Annex 1 – [5] – [11]
	Spatial methods	Annex 1 – [12] – [31]
Graph Autoencoders (GAEs)	Network Embedding	Annex 1 – [32] – [37]
	Graph Generation	Annex 1 – [38] – [43]
Spatial-temporal Graph Neural Networks (STGNNs)		Annex 1 – [44] – [50]

Table 1: Extensive reading list for GNNs [27]

After an extensive and exhaustive study of the literature, the authors Wu et. Al. (A Comprehensive Survey on Graph Neural Networks, 2021) suggests that further investigation in the field of of GNNs should focus on 4 key areas, Model depth, Scalability trade-off, Heterogeneity and Dynamicity:

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Model depth

The literature shows that as ConvGNNs get deeper, i.e they increase the h number of convolutional layers, their performance drops drastically raising the questions of whether perusing deeper ConvGNNs is still a good strategy for learning graph data.

Scalability trade-off

The scalability of GNNs is gained at the price of corrupting graph completeness. How to trade-off algorithm scalability and graph integrity could be a future research direction.

Heterogeneity

Heterogeneous Graphs often contain different types of nodes and edges, or different forms of node and edge inputs, such as images and text. Most GNNs assume that graphs are homogeneous, however it is still difficult to directly apply current GNNs to heterogeneous graphs, and therefore methods to handle heterogeneous graphs should be an area of future research.

Dynamicity

Nodes or edges may appear or disappear and node/edge inputs may change time by time, meaning that graphs are essentially dynamic in nature. In order to adapt to this dynamicity new graph convolutions are needed beyond the solutions offered by STGNNs, which seldom consider how to perform graph convolutions in the case of dynamic spatial relations.

Attention Mechanisms

One drawback to RNNs is that fact that they process information sequentially. The problem with this “sequential” processing means that information that is not strictly sequential is not well handled. By using mechanisms such as forget gates, reset gates, update gates etc, LTSMs and GRUs provide a method of ensuring that only important information is passed along from one step to the next.

Another type of RNN called Bidirectional RNNS (BiRNNs) provide a mechanism of analysing prior and subsequent information at each step in order to get around the “strictly sequential” (Venkatachalam, 2019) problem, however another problem arises. Namely the fact that the longer the sequence length, the less likely it is that the context will be captured correctly. Put another way, “the more updates are made to the same vector, the higher the chances are the earlier inputs and updates are lost” (Venkatachalam, 2019) , as shown in Figure 3 where an example of a sentence translation is given.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

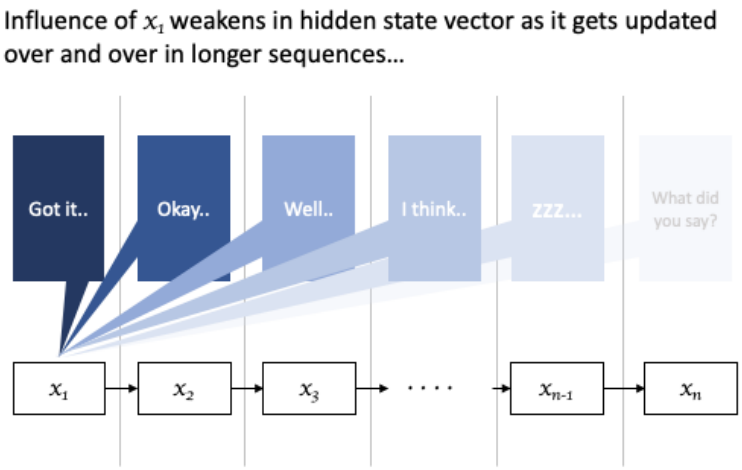


Figure 3: Context weakens with longer sentences *[31]*

This problem was highlighted in a paper entitled On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, the authors state that "the neural machine translation performs relatively well on short sentences without unknown words, but its performance degrades rapidly as the length of the sentence and the number of unknown words increase." (On the Properties of Neural Machine Translation: Encoder–Decoder Approaches, 2014)

Attention Networks were developed in order to solve this problem. First proposed in a paper from 2015, authors Bahdanau et. Al. (NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE, 2015), Attention Networks involve an Encoder-Decoder RNN[2] model that moves away from the fixed-length vectors used by other encoder-decoder models and allows for "a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly" (NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE, 2015). The result was what became known as an Attention Network.

In essence, Attention Networks do not have a single context vector like other RNNs may have and instead have a distinct context vector for each output step. This type of attention mechanism is known as a global attention mechanism and can become computationally very expensive. Local attention networks pay attention only to a given number of the hidden vectors around the hidden vector being considered. First proposed in (Manning, 2015), the authors purported that it offered greater than state of the art performance.

Attention networks are largely based on RNN networks and as such, suffer from the problem of not being able to process information in parallel – thus leading to their complexity with large amounts of data. CNNs can be used to get around this problem however, one problem persists with CNNs for

[2] Encoder-Decoder networks are common in language translation as they used an RNN to understand the context of the sentence being translated (the encoder) and a separate RNN to translate the original sentence into another language (decoder).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

translations and that is that it is difficult to figure out dependencies when translating. For this, we need to look at Transformers.

Transformers

As described in a 2017 paper from Google, Vaswani et. Al. the Transformer is: “a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality” [35]. In this paper the authors had introduced the world to a new type of model and a new type of attention – Self-attention. Although still employing the encoder-decoder mechanism, the authors proposed a radical new 6 Layer (each with 2 sublayers) architecture. It is the first sublayer which enables the model to process all input words at once, allowing it to model the relationships between all the words in the sentence. This ability enables transformers to model “long range” dependencies much faster that CNN and RNN models.

The architecture proposed by Vaswani et. Al. is shown in Figure 4 below.

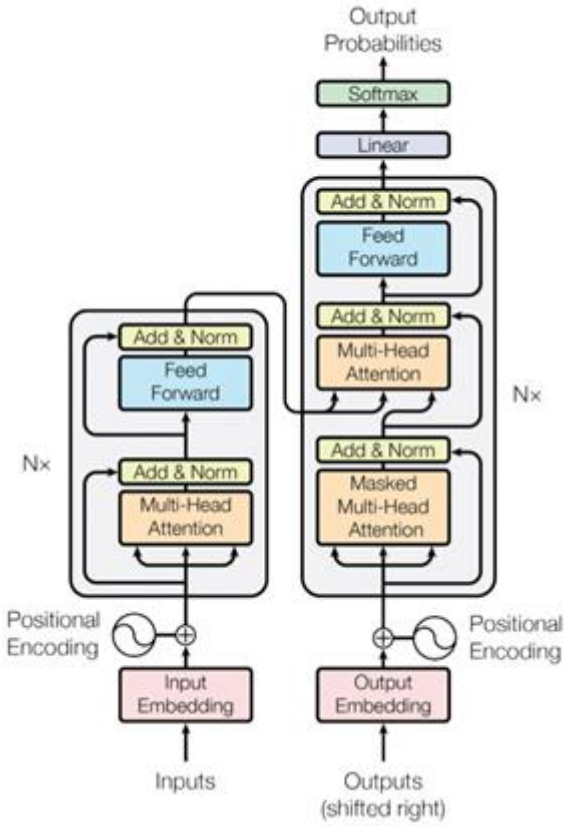


Figure 4: The Transformer model’s novel architecture*[35]*

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

BERT

The work by Vaswani et. Al. led to 2018 (1) paper entitled BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (J. Devlin, 2019) which set the ML community a buzz due to its state-of-the-art results across variety of Natural Language Processing tasks, including Question Answering (SQuAD v1.1) and Natural Language Inference (MNLI) (Horev, 2018) among others.

The **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT) [3] architecture is novel in that it applies the Transformer's bidirectional training abilities to language modelling. That is to say it applies pre-trained language representations using transformers to first pre-train a model and then fine tune it.

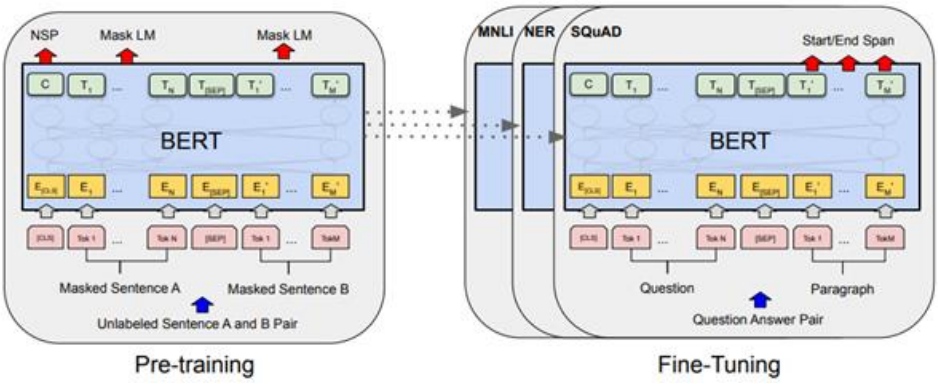


Figure 5: The pre-training and Fine-tuning steps from the BERT *[36]*

(i)Masked Language Model (MLM) by masking & predicting 15% of the input tokens of sentences/paragraphs, and with (ii) Next sentence prediction (NSP) because many NLP tasks depends on understanding relationship between sentences.
In the Pre-training stage, the model is trained using unlabelled data from Wikipedia and BookCorpus and two unsupervised tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). MLM works by randomly masking 15% of the words which appear in the given sentence and then the system attempts to predict them using their context, whereas NSP works by replacing the next sentence with another random sentence from the corpus (50%) in order to train a model that is capable of understanding the relationship between sentences.

In the Fine-tuning stage, the model is initialized with the pre-trained parameters which are then fine-tuned for the desired downstream task. The fine-tuning process will be use case specific and will result in a slightly different BERT on each occasion, however there is no need to retrain the model and the fine-tuning process is fast.

[3] The Bidirectional part of the name BERT comes from the fact that Transformers read all the words in a sequence at once, not sequentially (Left to Right) as with other models, and therefore can ascertain the context of the sentence. Although it is called bidirectional, non-directional may be a more correct description.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

An in-depth discussion of how BERT achieves this can be found in the original paper [36], along with explanations of how it was trained; both of which are considered beyond the scope of this document.

The literature shows that BERT has already had myn practical applications including:

1. BERTSUM (Lapata, 2019) – used as a general framework for both extractive and abstractive text sumarisation models
2. Google Smart Search (Nayak, 2019) – Understanding the context of sentences to better understand search queries
3. SciBERT (SciBERT: A Pretrained Language Model for Scientific Text, 2019) A tool for large-scale knowledge extraction and machine reading of published scientific documents. SCIBERT was developed using 1.4 million papers from semantic scholar and was found to outperform BERT base on several scientific and medical NLP tasks.
4. ClinicalBERT (Huang, 2019) models Clinical Notes and Predicts Hospital Readmission using contextual embeddings of Clinical texts/notes. ClinicalBert outperforms baselines on 30-day hospital readmission prediction saving money, time, and more importantly, lives.

ELECTRA

Whilst BERT uses MLM for training, a new model proposed in 2020 called ELECTRA (ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators, 2020) uses a novel method called replaced token detection (RTD). RTD works by replacing some input token with plausible alternative sampled from a small generator network, then, “instead of training a model that predicts the original identities of the corrupted tokens, a discriminative model is trained that predicts whether each token in the corrupted input was replaced by a generator sample or not”. As the task is defined over all input tokens rather than just the small subset that was masked out, which, the authors report is a much more efficient pre-training task than MLM and results in contextual representations which substantially outperform the ones learned by BERT (given the same model size, data, and compute). The authors also report that out-performance when compared to BERT is significant: “The gains are particularly strong for small models; for example, we train a model on one GPU for 4 days that outperforms GPT (trained using 30x more compute) on the GLUE natural language understanding benchmark. Our approach also works well at scale, where it performs comparably to RoBERTa and XLNet while using less than 1/4 of their compute and outperforms them when using the same amount of compute.”

GPT

Generative Pre-trained Transformers (GPT) are a transformer method which took the NLP world by storm. Prior to GPT, NLP models were trained using supervised learning on a particular task, for example sentiment classification or textual entailment. This required large amounts of labelled data and had to be trained for each new use case, i.e. was not able to generalise for new tasks. GPT proposes a generative learning model which is trained on unlabelled data, then fine-tuned for each task. Using this method, the researchers were able to achieve Zero-shot learning. GPT (what is no known as GPT1) was described in 2018 in a paper entitled Improving Language Understanding by Generative Pre-Training (Alec Radford, 2018). This was greatly improved on by GPT2 (Alec Radford, 2019) which was then in turn updated in in 2020 with GPT3 (Tom B Brown, 2020), perhaps the most powerful NLP tool to date.

Although extremely powerful and capable of creating high quality textual content, the authors of the paper describing GPT3 admitted that when forming long sentences it starts to lose coherency and can begin to repeat sentences over and over again. Given its ability to generate high quality text, there are

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

ethical concerns with GPT3, namely that it could be used for nefarious ends (phishing, fraud etc). There is also a concern in that the model inherits the biases of the langue it was trained on – meaning that it can inherit traits such as gender, ethnicity, race or religion bias and as such the another suggest caution with its use. At the time of writing this document, it is not open to the public and only available in limited format for research purposes.

Transformer-XL

One problem with sequential data like languages is that of long-range dependence, i.e. that one word used in the current sentence could refer to something mentioned previously in the text, perhaps even hundreds of words previously. Whilst gated RNNs and gradient clipping (Graves, 2013) have been proposed to solve this issue, they have been shown to be less than optimal at dealing with long texts, for example 200 characters or more (Urvashi Khandelwal, 2018). Transformers are used to better capture long-term dependencies however they are currently implemented with a fixed-length context (Rami Al-Rfou, 2018), where text structures are truncated into fixed-length segments, usually of the order of a few hundred characters, and then each segment is processed separately (Le, 2019). This introduces two problems, (1) the algorithm can only deal with dependencies shorter than the fixed length, and (2) the fixed length doesn't respect sentence boundaries and can therefore result context fragmentation.

Transformer-XL was proposed, which uses two techniques to get around the fixed-kength context problem: a segment-level recurrence mechanism and a relative positional encoding scheme.

The **segment-level recurrence mechanism** works by fixing and caching the representations computed for the previous segment in training which can then be reused as an extended context when the model processes the next new segment, resolving the context fragmentation issue and providing necessary context for tokens in the front of a new segment.

The recurrence mechanism is made possible using a novel relative positional encoding scheme, called **Relative Positional Encodings**, which uses fixed embeddings with learnable transformations instead of learnable embeddings, making it much more generalizable to longer sequences at test time. As a result, Transformer-XL has been shown to have a much longer effective context than a vanilla Transformer model at evaluation time.

The authors conclude that Transformer-XL has three key benefits:

- Transformer-XL learns dependency that is about 80% longer than RNNs and 450% longer than vanilla Transformers, which generally have better performance than RNNs, but are not the best for long-range dependency modelling due to fixed-length contexts (please see our paper for details).
- Transformer-XL is up to 1,800+ times faster than a vanilla Transformer during evaluation on language modelling tasks, because no re-computation is needed (see figures above).
- Transformer-XL has better performance in perplexity (more accurate at predicting a sample) on long sequences because of long-term dependency modelling, and also on short sequences by resolving the context fragmentation problem. (Le, 2019)

Transformer-XL is envisioned to bring significant improvements in areas such as improving language model pre-training methods (BERT), generating realistic long articles as well as finding applications in image and speech domains (Zihang Dai, 2019).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

4.1.2 Transfer Learning for NLP

Although previous models covered in this chapter, such as BERT, dealt, implicitly, with the concept of Transfer Learning, this section will go into a little more depth on the concept.
Transfer learning, as the name would suggest, involves taking knowledge gained by solving one problem and "transferring" it to solve another problem. This is in essence how models like BERT function, with their pre-trained models then being used in the fine-tuning steps. One key advantage to transfer learning is that models do not need to be retrained from scratch for new applications, saving resource and time.
Leaning heavily on enhancements in the computer vision space, 2018 saw new NLP models created which pushed the state of the art. Namely, ELMo (Embeddings from Language Models) (Peters, 2018), ULMFiT (Universal Language Model Fine-tuning for Text Classification) (Jeremy Howard, 2018), and GPT (Generative Pretraining) (Alec Radford, 2018)
ELMo is a shallowly bidirectional language model, which is a two-layer LSTM and is pre-trained on the 1 Billion word data set (Al., 2020). It uses a Shallowly Bidirectional Language Model and included Feature Extraction to select the adopted model.

The strength of **ELMo** is that it adds contextualisation to words and as such can create unique, character-based, and deep word representations which are adopted to each task, a significant advancement on previous tasks.

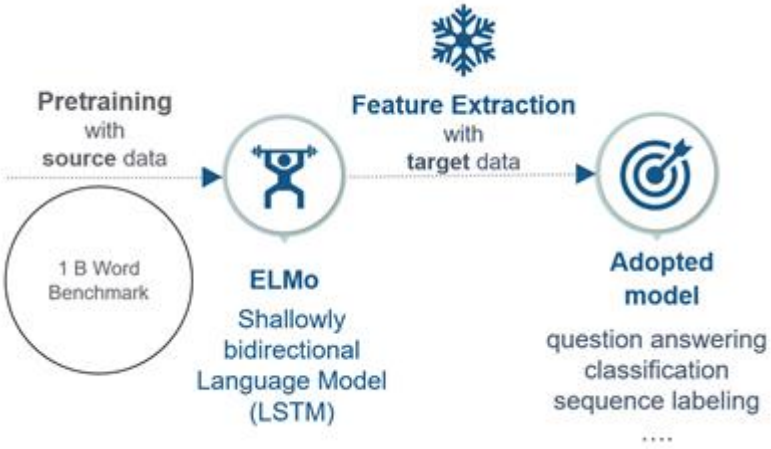


Figure 6: Conceptual Model of ULMFiT showing Pre-Training and Adoption steps with Feature Extraction [50]

The **ULMFiT** model introduced concepts such as fine-tuning, which noticeably lowered the error rate in text classification by 18%-24% (Al., 2020). ULMFiT is based on an AWD-LSTM (ASGD Weight-Dropped LSTM) model and contains three distinctive stages of pre-training and adoption: Language Model pre-training, Language Model fine-tuning (Adoption I) and Classifier fine-tuning (Adoption II) (Al., 2020).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

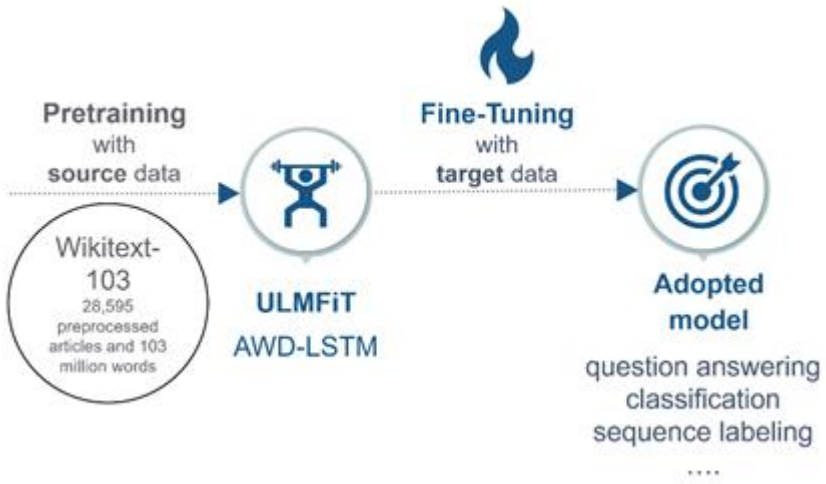


Figure 7: Conceptual Model of ULMFiT showing Pre-Training and Adoption steps with fine-tuning [50]

GPT in similar to ELMo however it uses a multi-layer transformer decoder instead of the LTSM that ELMo uses. ELMo is trained on word embeddings and GPT uses fine-tuning (like ULMFiT).

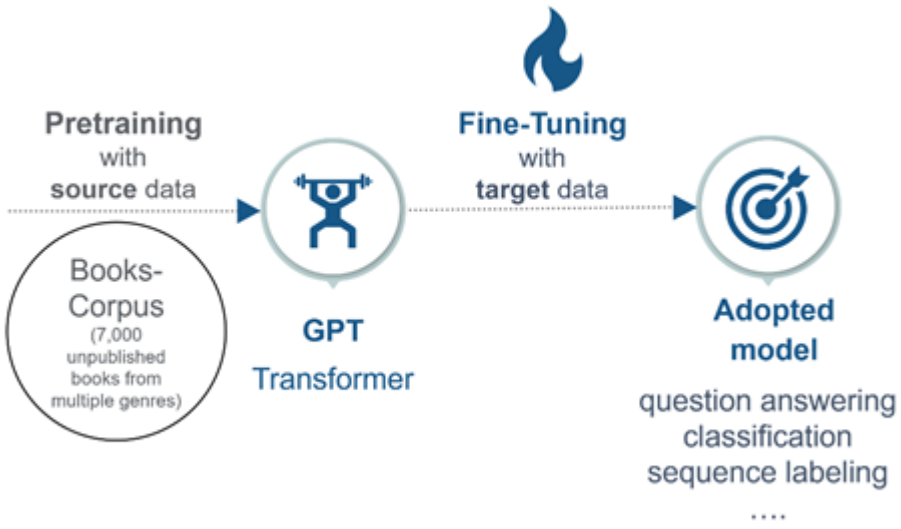


Figure 8: Conceptual Model of ULMFiT showing Transformer Pre-Training and Adoption steps with fine-tuning [50]

4.1.3 Evolutionary learning

Evolutional Algorithms (EA) are randomized optimization methods, a heuristic-based approach to solving problems and are employed in cases where finding a solution is not possible in polynomial time, i.e. problems that are NP Hard or involve massive amounts of computational time to resolve. They can be used by themselves or as a mean to find a starting point for other algorithms.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

The concept originated from the work John Holland from the University of Michigan in the late 1960s when he successfully introduced the concept of sexual reproduction to EA, which has since then evolved somewhat.

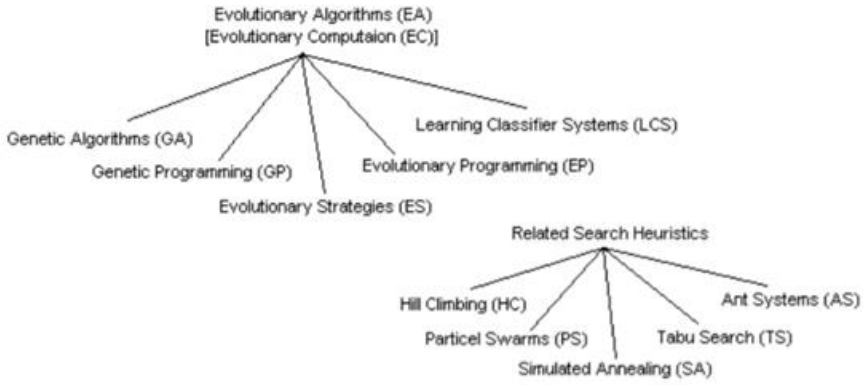


Figure 9:Classification of EA methods **(Streichert, 2005)**

Genetic Algorithms (GA) are one type of EA and they earn the title "genetic" as they mimic the process of natural selection in that they operate in four phases: initialization, selection, genetic operators, and termination.

- **Initialization**
  A genetic algorithm starts with an initialisation – a random pool (called a population) of solutions (called members).
- **Selection**
  One the population has been created, its members are evaluated for fitness against the Fitness Function and a selection of the best fitting members is selected. The Fitness Function is specific to each task and should accurately represent the data in each case.
- **Genetic Operators**
  Just as with real life genetics, two solutions are chosen and combined, mixing their characteristics, in order to create an off-spring solution – known as the next generation. This step is called Crossover. In order to avoid becoming trapped in a local extrema, the algorithm introduces a probabilistic change to the child – known as a mutation.
- **Termination**
  All good things must come to an end, and as such, when the algorithm has reached either its maximum run time or have achieved a required threshold of performance, the final solution is retuned.

EAs find their uses in many scenarios, not least in FinTech where they can be used for portfolio selection, time series prediction as well as to generate trading rules (Streichert, 2005). They have also been successfully applied in the medical field to identify key metabolites for osteoarthritis (Ting Hu, 2018) and also find their applications in statistical natural language processing (How evolutionary algorithms are applied to statistical natural language processing, 2007)

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

4.1.4 NER – Name entity recognition

Named Entities are real world entity that can either be physical, like a product, person, address etc. or can be abstract such as a time or a date. Named entity recognition (NER) is an NLP technique used to identify such entities in a given text and then classify these entities into predefined categories. Sometimes called Entity Identification or Entity Extraction, it is considered sub-task of information extraction and can be used to extract data such as names of people, organizations, times, dates and locations as well as percentages, product numbers, serial numbers or monetary values from text.

The first attempts at NER were rule based, using domain-specific gazetteers and syntactic-lexical patterns (Nltk.org). Unsupervised approaches to NER included techniques such as inverse document frequency or noun phrase chunking. Supervised learning approaches have also been applied to BER problems where features are designed to represent key features in the documentation, then an ML model is trained to recognise these features. These attempts have included “semi-supervised learning techniques, which extracts useful features from transferring information from resource-rich language toward resource-poor language (Semi-supervised learning for named entity recognition using weakly labeled training data,, 2015) as well as Distantly Supervised NER (Distantly Supervised NER with Partial Annotation Learning and Reinforcement Learning, 2018), in order to automatically populate annotated training data without human cost.Another advance for NER was Distributed Word Representation (Distributed representations of words and phrases and their compositionality, 2013) (Word2Vec etc)

In recent years, Deep Learning approaches have been applied to NER to great success. One key problem this solves is that the algorithm can now define features for a data set, thus avoiding any bias created by the programmer when manually creating features and being able to learn much more complex and intricate features from data via non-linear activation functions (A Survey on Deep Learning for Named Entity Recognition, 2020). Attention models have also added a boost to the state of the art performance of NER with great advances in the field being accomplished with BERT models. The following is a guide on how to implement NER with BERT (Sterbak, 2020).
LUKE (Language Understanding with Knowledge-based Embeddings) is a recent extension of the BERT method outlined in the aforementioned guide, and is achieved by tweaking BERT's Masked Language Model (MLM) and adding a new entity-aware self-attention mechanism (Ikuya Yamada, 2020).

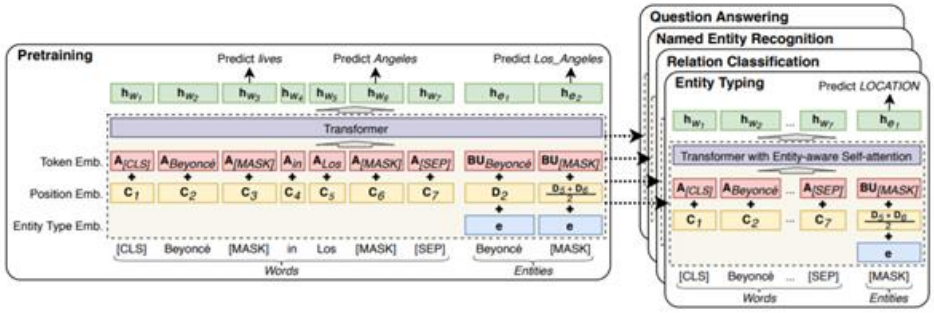


Figure 10: Architecture of LUKE using the input sentence “Beyonce lives in Los Angeles. ´ ” LUKE outputs contextualized representation for each word and entity in the text. The model is trained to

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

predict randomly masked words (e.g., lives and Angeles in the figure) and entities (e.g., Los Angeles in the figure). Downstream tasks are solved using its output representations with linear classifiers.

In order to train a NER model, a data set is required: Several of the most widely used datasets are summarised below in Figure for reference.

Corpus	Year	Text Source	#Tags	URL
MUC-6	1995	Wall Street Journal	7	https://catalog.ldc.upenn.edu/LDC2003T13
MUC-6 Plus	1995	Additional news to MUC-6	7	https://catalog.ldc.upenn.edu/LDC96T10
MUC-7	1997	New York Times news	7	https://catalog.ldc.upenn.edu/LDC2001T02
CoNLL03	2003	Reuters news	4	https://www.clips.uantwerpen.be/conll2003/ner/
ACE	2000 - 2008	Transcripts, news	7	https://www.ldc.upenn.edu/collaborations/past-projects/ace
OntoNotes	2007 - 2012	Magazine, news, web, etc.	18	https://catalog.ldc.upenn.edu/LDC2013T19
W-NUT	2015 - 2018	User-generated text	6/10	http://noisy-text.github.io
BBN	2005	Wall Street Journal	64	https://catalog.ldc.upenn.edu/LDC2005T33
WikiGold	2009	Wikipedia	4	https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
WiNER	2012	Wikipedia	4	http://rali.iro.umontreal.ca/rali/en/winer-wikipedia-for-ner
WikiFiger	2012	Wikipedia	112	https://github.com/xiaoling/figer
HYENA	2012	Wikipedia	505	https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/hyena
N^3	2014	News	3	http://aksw.org/Projects/N3NERNEDNIF.html
Gillick	2016	Magazine, news, web, etc.	89	https://arxiv.org/e-print/1412.1820v2
FG-NER	2018	Various	200	https://fgner.alt.ai/
NNE	2019	Newswire	114	https://github.com/nickyringland/nested_named_entities
GENIA	2004	Biology and clinical text	36	http://www.geniaproject.org/home
GENETAG	2005	MEDLINE	2	https://sourceforge.net/projects/bioc/files/
FSU-PRGE	2010	PubMed and MEDLINE	5	https://julielab.de/Resources/FSU_PRGE.html
NCBI-Disease	2014	PubMed	1	https://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/
BC5CDR	2015	PubMed	3	http://bioc.sourceforge.net/
DFKI	2018	Business news and social media	7	https://dfki-lt-re-group.bitbucket.io/product-corpus/

Figure 11: List of annotated datasets for English NER. "#Tags" refers to the number of entity types. (A Survey on Deep Learning for Named Entity Recognition, 2020)

The success of any NER model is defined against three evaluation metrics, Precision, Reaclll and F-1 Score. These are complex metrics which often makes evaluation of NER performance a non-trivial task. F taken from (Singh, 2021) offers an explanation of these evaluation methods:

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

- **Precision**: Precision is defined as below $Precison = \frac{TP}{TP+FP}$, where TP = True Positive, i.e. entities that are recognized by NER and match the ground truth. FP = False Positive, i.e. entities that are recognized by NER but do not match the ground truth. Precision measures the ability of a NER system to present only correct entities.
- **Recall**: Recall is defined as below $Recall = \frac{TP}{TP+FN}$, where FN = False Negative, i.e. entities annotated in the ground which that are not recognized by NER. Recall measures the ability of a NER system to recognize all entities in a corpus.
- **F-1 Score**: F-1 score is the harmonic mean of precision and recall. i.e. $F-1Score = 2 * \frac{Precision*Recall}{Precision+Recall}$ Since most NER systems involve multiple entity types, it is often required to assess the performance across all entity classes. Two measures are commonly used for this purpose: the macroaveraged F-1 score and the micro-averaged F-1 score. The macro-averaged F-1 score computes the F-1 score independently for each entity type, then takes the average (hence treating all entity types equally). The micro-averaged F-1 score aggregates the contributions of entities from all classes to compute the average (treating all entities equally).

Figure 12: An explanation of the most commonly used evaluation metrics for NER systems

NER has several uses in the businesses including:

1. Ticket Categorisation in Customer Support to automate repetitive customer services requests to make resolution faster, or to extract pertinent information from customer support tickets such as serial numbers and product names etc.
2. Customer Feedback analysis to gain insight into how the market perceives a company's products, including analysis negative and positive feedback for problem resolution or product improvement suggestions.
3. Recommender Systems uses NER to classify items, such as the genre of movies or music in streaming services in order to recommend similar items to customers.
4. NER is also often used in HR where employers have to analyse 1000s of CV for new jobs openings. CVs are notoriously "unstructured" as each one presents its info in a different manner, NER helps to automatically extract important information such as qualifications, name, email, experience etc.

There are various opens source projects which can be easily integrated into any project with the use of an API, such as,

- The Stanford Named Entity Recognizer (SNER) which is JAVA based: https://nlp.stanford.edu/software/CRF-NER.shtml
- SpaCy (Python framework) known for its speed and ease of use: https://spacy.io/usage/linguistic-features#named-entities,
- Apache OpenNLP: https://gitbox.apache.org/repos/asf?p=opennlp.git
- General Architecture for Text Engineering (GATE), developed in JAVA and widely used by the scientific community: https://github.com/GateNLP, and
- The Natural Language Toolkit (NLTK) https://www.nltk.org, libraries for Python.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

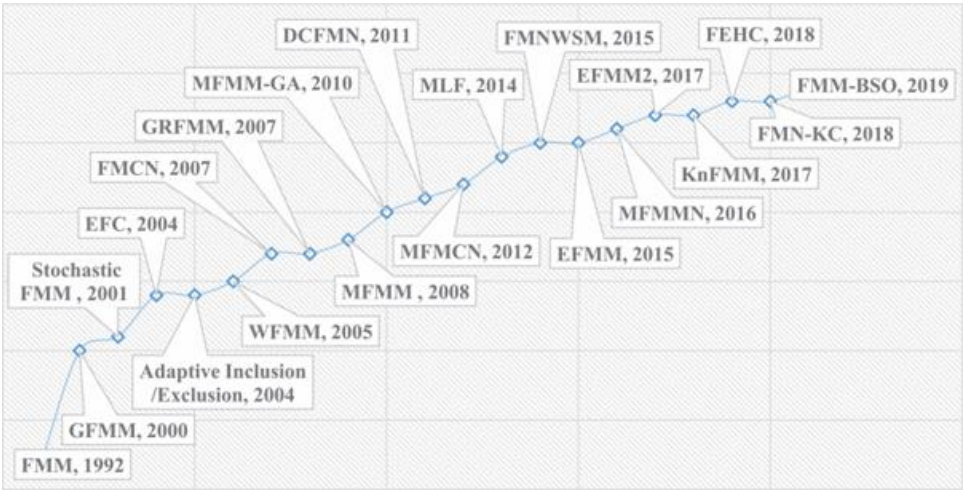
4.1.5 Fuzzy Classifiers

Real life is filled with uncertainty so it is no surprise that it is no different when considering classification tasks. Objects can be correctly classified into one or more categories as it is often the case that there is some uncertainty about which category they fit into. If a particular day has periods of both rain and sunshine, can it be classified as a rainy day or a sunny day? To the same extent, survey questions with multiple answers such as “Strongly Like, Like, Dislike or Strongly Dislike” show degrees of membership to one group or another. Strongly Like and Strongly Dislike could be considered 1 and 0, whereas Like and Dislike could be considered a percentage on the scale between 1 and 0, for example 0.66 and 0.33 respectively.

This idea of a “Fuzzy Set” was first described by Zadeh in 1965 as “a class of objects with a continuum of grades of membership. Such a set is characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one” (Fuzzy sets, 1965) One common example of this is the concept of young.

In terms of ML, a classifier is an algorithm which assigns a class label to an object based on a description of that object. The object description is a vector detailing the relevant features needed for classification. Standard classification is sometimes referred to as Crisp Classification as objects are categorised in specific categories. Fuzzy classification assigns what are called soft-labels, that is to say a label will have a weighting or a percentage of membership of a particular category.

Pattern classification is central to decision making in ML and the Fuzzy Min-Max (FMM) neural network is considered one of the most useful NNs for pattern classification. It combines the operations of an artificial neural network and fuzzy set theory into a common framework (A Critical Review on Selected Fuzzy Min-Max Neural Networks and Their Significance and Challenges in Pattern Classification, 2019). The literature shows that there have been many versions of FMMs developed in the last 30 years and research continues.



Authours Alhroob et. AL. provide a thorough investigation of these differeing methods including their merits and demerits in their paper A Critical Review on Selected Fuzzy Min-Max Neural Networks and Their Significance and Challenges in Pattern Classification. The authors conclude their work by stating that the performance of FMM variants can be enhanced by addressing FMM limitations and propose that future research concentrate on new models with improvements such as

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

1. Developing news learning methods to eliminate the expansion coefficient defined by users, stating that “here is a need to eliminate the user defined coefficient during the learning process to generate accurate hyperboxes decision boundaries and decrease the misclassification rate”
2. A new method for the overlap test to investigate the possibility of overlapped regions among hyperboxes from different classes and avoid their occurrence, and
3. A new contraction process to avoid membership ambiguity of overlapped regions

4.1.6 Link Prediction

Link prediction is used to predict possible links in networks or to find missing links in networks where data is scarce. This often takes the form of prediction of new contacts on social media, new products on ecommerce marketplaces or even prediction connections between species in biology. It has also been applied to criminal networks to find missing links between actors where not all the data is known (Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis, April 22, 2016).

Prediction methodologies can be local, for example based on nearest neighbour analyses, or global, considering the entire graph and the distances between edges. As each graph grows differently, different link prediction algorithms will work differently on each graph.

One good reference for a comparison of the performance or several LP methods is a paper entitled “A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks” in which the authors study several methods as well as evaluation results and links to several benchmark datasets for further testing. To recommendations the authors make as to areas for further study include transfer learning to allow this filed to apply to many more domains as well as including “information such as hierarchical descriptions between entities/relations in KG, textual Internet information, and even the extracted information from other KGs, can also be applied to refine the representation performance of the embedding models (A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks, 2020).

4.1.7 Knowledge Creation

One definition of Knowledge creation is that it is the act of making knowledge, created by individuals available, amplifying it in social contexts, and selectively connecting it to the existing knowledge in the organization (Perspective-tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory, 2009). Put another way, knowledge is described as a continuous process of combining, transferring, and converting different kinds of knowledge. Within businesses it is used to share and generate ideas with the ultimate goal of giving the organisation a competitive edge. These new ideas and concepts are generated when explicit knowledge (easily attainable knowledge) comes into contact with tacit knowledge (knowledge contained in the mid of someone else, not easily shareable). Knowledge Management is the science dedicated to this field and it has led to what are known as Knowledge Economies: an economy where production and implementation of knowledge is paramount to success.

A framework is provided by authors Arling et. Al. in which offers a rubric against which both old and new KM initiatives can be assessed to determine whether they are capable of generating new knowledge (Facilitating new knowledge creation and obtaining KM maturity, 2011).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

4.1.8 NLG (Natural Language Generation)

Natural Language Processing is focused on how to get computers to process natural human language (written or spoken word as opposed to computer code) and turn that input into a specific output. Natural language understanding (NLU) and Natural Language Generation (NLG) can be considered as the parts of this process that respectively understand what is being input and then output a response understandable by humans.

The roots of NLG go back to early attempts to mimic human speech (Building Applied Natural Language Generation Systems, 1995) where three distinct stages were laid out: Document planning (content planning, document outlining), Microplaning (word choice, aggregation to expand the document and referring expressions), and Realisation (conversion of specifications to text).

Early approaches used templates to generate documents however “true” or “real” NLG uses dynamic documentation creation. One of the earliest approaches was “Fill-in-the-gap” approaches where a template was used and the gaps well filled accordingly. Web Scripting were an expansion of this and used web templates to automatically generate custom web page. Use cases include SERPs (Search Engine Results Pages) in a wide level of scenarios, ecommerce, search engines, flight and hotel search etc.

A forward leap was taken from template based NLG to dynamic NLG with Dynamic Sentence Generation at a micro level and Dynamic Document Creation at a macro level.

A recent area of research in NLG is End-to-End learning where all layers of the network are trained simultaneously, as shown in Figure 17 below.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

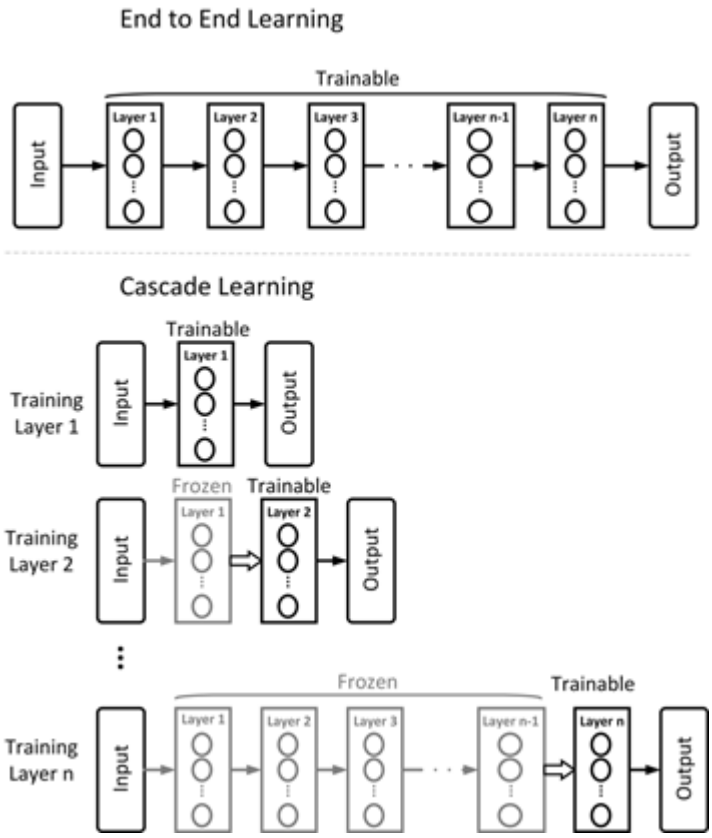


Figure 13: End-to-End (E2E) learning versus Cascade Learning (Transfer Learning Across Human Activities Using a Cascade Neural Network Architecture, 2019)

4.1.9 Chatbots (ES)

Virtual assistants like Amazon's Alexa, Apple's Siri, Google Assistant and Microsoft's Cortana have become commonplace in the technology we use day-to-day, i.e. our tablets, smartphones, smart home devices and our computers and show the potential of Conversational Agents when it comes to interacting with humans. As well as providing assistance at home or on the go, Chatbots have also become commonplace in business scenarios, such as customer service, where clients queries can be easily and swiftly answered by a chatbot or directed to a human operator for further processing.

Chatbots can be "task focused" (i.e. "Please chose one of the following options") directing users to a particular response or outcome or can be much more open-ended and conversational in style.

The earliest examples of chatbots were simple rule-based programmed responses based on templates and predefined answers, however advances in technology have produced chatbots that truly are conversational and can be used in a wide range of applications.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

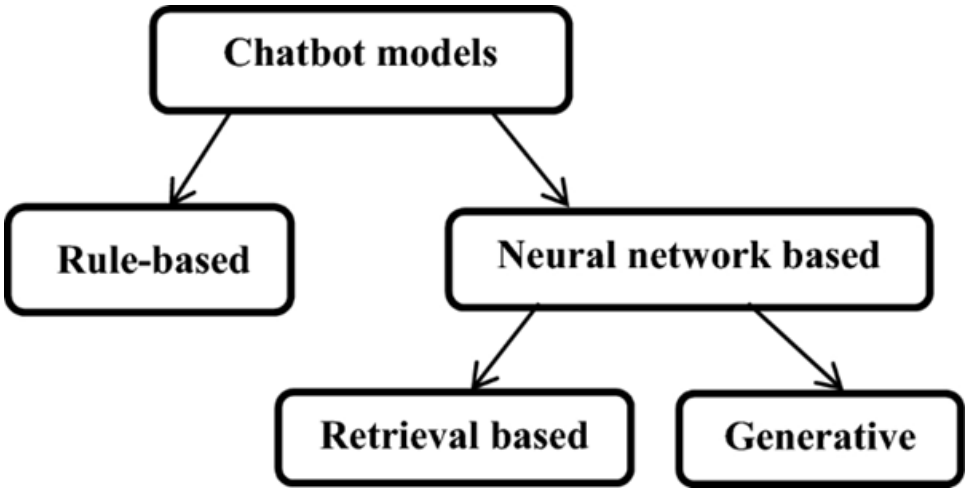


Figure14: Evolution of Chatbots from rule-based to NN Based

Chatbots can generally be of two formats: task-orientated or open-domain. Task-oriented chatbots are very efficient at handling queries from within one particular domain and are adept in situations such as handling reservations or receiving bookings etc. Open-domain chatbots are much more conversational and find their usage in a wider range of applications, such as e-learning, ecommerce, medicine, marketing, customer care and personal assistants, among others. They are much more human-like in this interaction and in the best of cases they are indistinguishable from a human[4].

In some cases, language "tricks" are directly embedded in the programing of a chatbot to offer the bot pre-defined "human-sounding" responses and make it appear more human. These can include some hard-coded pattern matching, pre-defined details about its birth, age, parents, preferences, or even an entire back-story and in some cases, common typing errors and fake keystrokes.

Rule-based Chatbots

In a 1998 paper entitled Introducing MegaHAL (Alder, 1998), a Markov Chain based Model (i.e. – where the probability of the next output is calculated based on the preceding one) was used to create a chatbot called MegaHAL. The authors, Hutchens and Adler, concluded that
"MegaHAL uses a technique which differs significantly from that used by previous entrants to the Loebner contest. It has been submitted in 1998 for the purpose of demonstrating a different method of simulating conversation. Although its replies are occasionally lucid, MegaHAL is most definitely not an Artificial Intelligence; we must be careful not to read too much into what it says."

One important development in the field of conversational agents was that of Artificial Intelligence Markup Language (AIML).

[4] The Turing Test, named after one of the founding fathers of computer science, Alan Turing, attributes intelligence to any system that is indistinguishable from human beings. Each year the Loebner Prize is award to the chatbot that is the deemed to be the most human-like. This prize has been won by Mitsuku (Kuki) for the previous 4 years - 2016-2019. This software can be demoed here: https://chat.kuki.ai/

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

AIML was developed by Dr. Richard S. Wallace along with the Alicebot free software community (AliceBot) between 1995-2000 and is based on a non-XML grammar also called AIML. In a 2003 book Wallace sets out The Elements of AIML Style (Wallace, 2003). Described as an open standard scripting language for chatbots, the AIML foundation (Artificial Intelligence Markup Language) created a simple language for creating chatbots that is easily understandable, interpretable and expandable.

OntBot was a chatbot developed in 2011 which was based on an ontology model in order to jump a number of existing hurdles: "the need to learn and use chatbot specific language such as AIML, high botmaster[5] interference, and the use of non-matured technology" (OntBot: Ontology based chatbot, 2011).

The 2010 Loebner Prize winner, Suzette, was based on a scripting language called Chatscript, developed by Bruce Wilcox in 2010 (Bruce Wilcox, 2011). Chatscript defined "concepts" of similar words to give context and Word-net Ontologies can be combined with chatscript to give better responses.

Another chatbot approach is with SQL and relational databases. It was introduced in a 2009 paper (An architectural design of Virtual Dietitian (ViDi) for diabetic patients, 2009) and its application in the ViDi (Virtual Diabetes Physician) chatbot was detailed in a 2010 paper (Extension and Prerequisite: An Algorithm to Enable Relations Between Responses in Chatbot Technology, 2010). The author's research was focused on "enabling chatbot to become a search engine that can process the next search with the relation to the previous search output. In chatbot context, this functionality will enhance the capability of chatbot's input processing"

Neural Network based Chatbots

One of the benefits of NN based chatbots is that it is no longer necessary to write rules for the chatbot. The NN can be used to create outputs automatically (generative), to retrieve the outputs from a dataset of pre-existing responses (retrieval) or can be a mix of both (hybrid).

The fact that data can persist through a recurrent neural network means that chatbots can analyse previous inputs and offer outputs more suited to the previous question. Whilst heavily used in tasks such as translation, RNNs suffer from gradient degradations (The Vanishing Gradient problem mentioned previously) and as such are not the first choice for Chatbots. LTSMs or GRUs are a better choice as their memory functions allows for previous input to propagate through the network to influence future answers. This could include remembering previously input information (such as gender, age or any applicable title) to use at a later data to make the conversation much more personal for the human involved.

The encoder-decoder model is popular with NN chatbots and works with an encoder NN which reads the source sentence and transforms it into a fixed-length vector. This vector is then used as the initial hidden state of the decoder which will generate the target sentence. The seq2seq model became the go-to standard for NN based chatbots however they are hard to train and require that all sequences have the same length in order to be encoded in vector format. This is problematic in that as sentences get long, the context can easily be lost – not good for chatbots. For this reason, attention models are

[5] In chatbots using supervised learning, the human doing the supervision is known as the botmaster

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

used. One solution, linked several encoders together to form a Deep seq2seq model, before passing this output to their deciders, improving on performance (A Neural Conversational Model, 2015). (Seq2seq Dependency Parsing, 2018) describes a model for end-to-end seq2seq parsing.

Contextualisation – Using Knowledge Graphs

Contextualisation in KGs is the task of augmenting the current information with additional or useful information. That is, the problem of extracting meaningful and relevant sub-graphs from the current KG. The fact that KGs can be huge makes this a difficult task, i.e. finding and delivering the most valuable information among a vast number of candidates.

One method for this is Neural Fact Contextualisation Method (NFCM) introduced in 2018 by Voskarides et. Al. in their paper Weakly-supervised Contextualization of Knowledge Graph Facts (Weakly-supervised Contextualization of Knowledge Graph Facts, 2018). In the paper they tackled three current issues: entity relationship explanation, distant supervision, and fact ranking. NFCM works by firstly generating a candidate set for a query fact through 1 or 2-hop neighbour analysis then these candidates are ranked using supervised machine learning. NFCM is used to combine human input features with features automatically identified using a deep NN. Distant supervision is used to boost the training data gathered and was shown to significantly outperform several heuristic baselines for this task.

The authors recommend a deeper investigation of how to combine human input with automatically learned features for ranking.

Another method for contextualization was highlighted in a 2019 paper entitled Path-Based Contextualization of Knowledge Graphs for Textual Entailment, in which the authors presented a path selection mechanism to generate sub-graphs that reduce noise as well as retrieving meaningful information from large knowledge graphs. This was achieved by customizing the edge-weights in the graph with costs produced by various heuristic functions. Their work was based on finding paths in a cost-customized external knowledge graph, and building the most relevant sub-graph that connects two sentences P and H, in order to automatically predict the entailment relationship between them (Fadnis, et al., 2019). The concluded by proposing that fully interpreting these paths and ranking them qualitatively based on their value to the NLI task is their next research topic.

4.1.10 Textual & Verbal Command Interfaces

In this section, we describe a structural view of the concepts (see Figure) involved in Natural Language Processing (NLP) solutions, whether it being text extraction or sentiment analysis.
NLP is quickly becoming an essential skill for modern-day organizations to gain a competitive edge. It has become the essential tool for many new business functions, from chatbots and question-answering systems to sentiment analysis, compliance monitoring, and BI (Business Intelligence) and analytics of unstructured and semi-structured content.
Consider all the unstructured content that can bring significant insights – queries, email communications, social media, videos, customer reviews, customer support requests, etc. Natural Language Processing (NLP) tools and techniques help process, analyse, and understand unstructured “big data” in order to operate effectively and proactively.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

FİGURE . - NATURAL LANGUAGE PROCESSİNG PROCESSES WORKFLOW [1]

In many use cases, the content with the most important information is written down in a natural language (such as Turkish, English, German, Spanish, Chinese, etc.) and not conveniently tagged. To extract information from this content you will need to rely on some levels of text mining, text extraction, or possibly full-up natural language processing (NLP) techniques.

Typical full-text extraction for Internet content includes:

- Extracting entities – such as companies, people, dollar amounts, key initiatives, etc.,
- Categorizing content – positive or negative (e.g. sentiment analysis), by function, intention or purpose, or by industry or other categories for analytics and trends,
- Clustering content – to identify main topics of discourse and/or to discover new topics,
- Fact extraction – to fill databases with structured information for analysis, visualization, trending, or alerts,
- Relationship extraction – to fill out graph databases to explore real-world relationships.
- This approach will be used as a basis in Oxilate digital assistance solution.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

5. Data Management

5.1 Intro to SotA of DA techniques

As we connect more and more devices to the internet, or to each other, our capacity to monitor and collect data is increasing exponentially – and therefore so too is the amount of data we capture. This data, being amassed by the myriad connected systems, devices, platforms, users and so on, poses new problems for researchers and interested parties, namely what to do with all of this data?

As its name implies, Big Data is distinct from regular data, not least by its size, i.e. the sheer amount of data captured. In the literature this is often referred to as the "Vs" of Big Data, and there are varying numbers of Vs depending on the author. For example, "big data has three main attributes namely: volume, velocity, and variety" (Big Data LifeCycle: Threats and Security Model, 2015) as described by Alshboul et. Al. In this case, Volume refers to the sheer size of the data being collected, velocity refers to the speed of generating and processing this data and variety refers to the different sources of heterogeneous data that make up Big Data(Social Media, IoT Devices, Wearables, Log Files, Databases and so on and so forth). This is highlighted in Figure 16 below.

Although several studies refer to the 3 Vs, Big Data has also been described as being characterized by at least 5 Vs in other studies:

Big data is different from traditional data. The main differences come from characteristics such as volume, velocity, variety, veracity, value and overall complexity of data sets in a data ecosystem. Understanding these V words provide useful insights into the nature of Big Data (Yildiz, 2019).

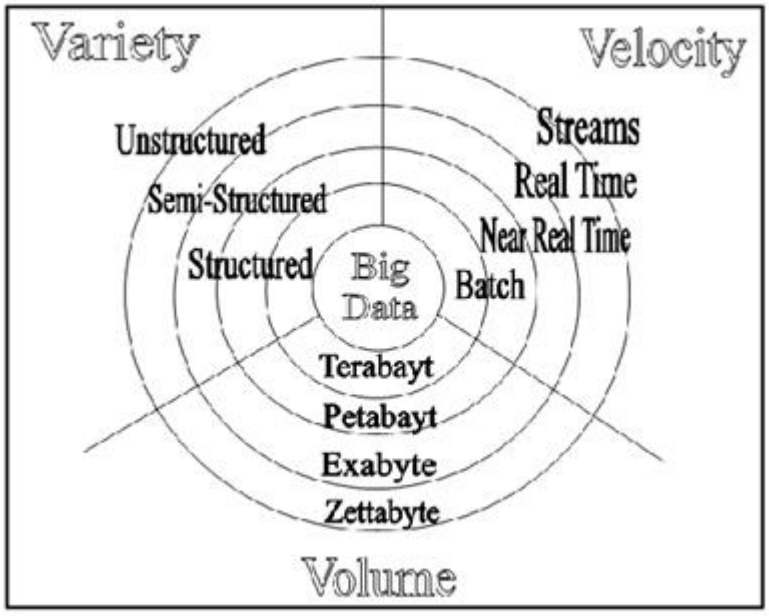


Figure 15: The 3 Vs of Big Data with examples(Big data: A review, 2013)

One of the key issues with data is the issue of Data Governance, i.e., how to safely and securely deal with all of this data in accordance with law and regulations. To this end it is important to approach the

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

problem of Big Data Governance in an organised and systematic manner. This is one of the focuses of the field of Data Science: an attempt to utilize all this available data in manageable, regulated and profitable ways, and by defining a Big Data Lifecycle, data managers can identify data at all the points in its “lifecycle” and assess the correct governance requirements at each stage.

5.2 Big Data Lifecycle

As laws and regulations concerning the handling of data get stricter and impose greater penalties for misuse of said data, it is increasing more important to define a framework for data to analyse what data is being captured, what is being down with this data and where does it end up. With this framework in pace, governance becomes a much clearer task. The Big Data Lifecycle is an attempt to define distinct phases in the “life” of data in order to uncover what the distinct governance requirements are at each stage. However, as with Big Data itself, the Big Data Lifecycle is an evolving terminology and is described in a variety of manners throughout the literature, each method varying slightly from other methods.

In the rest of this section, we will look into some of these methods and explain what they involve.

Perhaps the best know methodology for data mining is the CRoss Industry Standard Process for Data Mining (CRISP-DM). Conceived in 1996, CRISP-DM went on to be implemented by the European Union under the ESPRIT funding initiative and was later championed by five companies: SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA. It was launched in 1999 in order to standardize data mining processes across industries and has gone on to become the most common methodology (SALTZ, 2020) used for data mining, analytics, and data science projects. It offers us a detailed description of the Big Data Lifecycle, as shown in Figur below

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

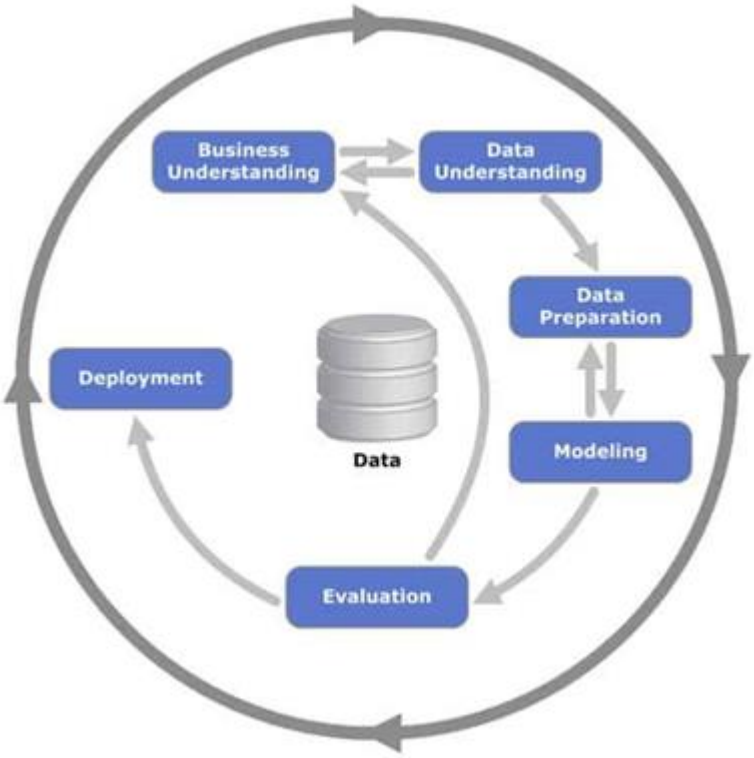


Figure 16: CRISP-DM Big Data Lifecycle (Datascience-pm.com)

Being the most commonly adopted methodology, the rest of this chapter will look into the Big Data Lifecycle through the lens of CRISP-DM.

However, before that it is worth briefly mentioning alternatives to CRISP-DM.

SEMMA

Although not as popular as CRISP-DM, SEMMA from SAS (SAS.com) is still a widely adopted methodology in the field of data science. It gets its name from the data mining definition from the SAP institute; data mining is a process which involves Sampling, Exploring, Modifying, Modelling, and Assessing (SEMMA), highlighted in Figure below.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

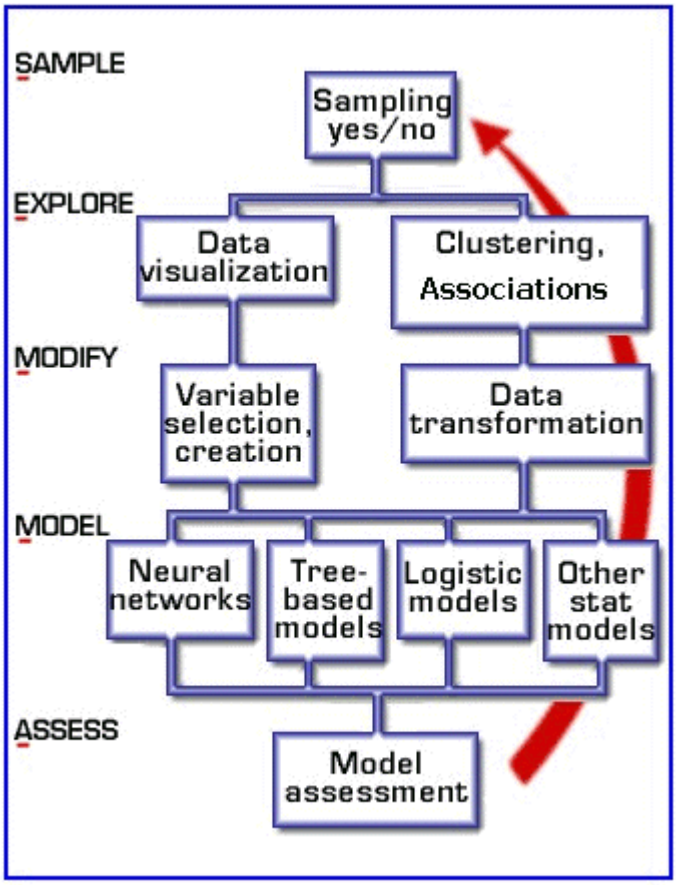


Figure 17: Overview of the SEMMA methodology (SAS.com)

Some of the steps my need to be repeated and it is not necessary to include each and every step in the process.

SAS offer a product called Enterprise Miner (SAS.com) which provides a GUI to guide data engineers through the process of big data analysis which they define as the following (SAS.com):

1. "Sample the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process.
2. Explore the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
3. Modify the data by creating, selecting, and transforming the variables to focus the model selection process.
4. Model the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.
5. Assess the data by evaluating the usefulness and reliability of the findings from the data mining process."

TDSP

Launched in 2016 and often described as a mixture between CRISP-DM and SCRUM, Microsoft's Team Data Science Process (TDSP) takes the methodological approach of CRISP-DM but also

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

considers the team executing the project as well. It is an iterative process and has five distinct stages (Data Science Project Management):

1. „Business Understanding: define objectives and identify data sources
2. Data Acquisition and Understanding: ingest data and determine if it can answer the presenting question (effectively combines Data Understanding and Data Cleaning from CRISP-DM)
3. Modeling: feature engineering and model training (combines Modeling and Evaluation)
4. Deployment: deploy into a production environment
5. Customer Acceptance: customer validation if the system meets business needs (a phase not explicitly covered by CRISP-DM)“

For each of the five stages, Microsoft provides information on what specific objectives are needed to reach each goal, an outline of the specific tasks and guidance on how to complete them as well as providing the deliverables and the support needed to produce them (Microsoft).

The TDSP life cycle is summarized by the following diagram:

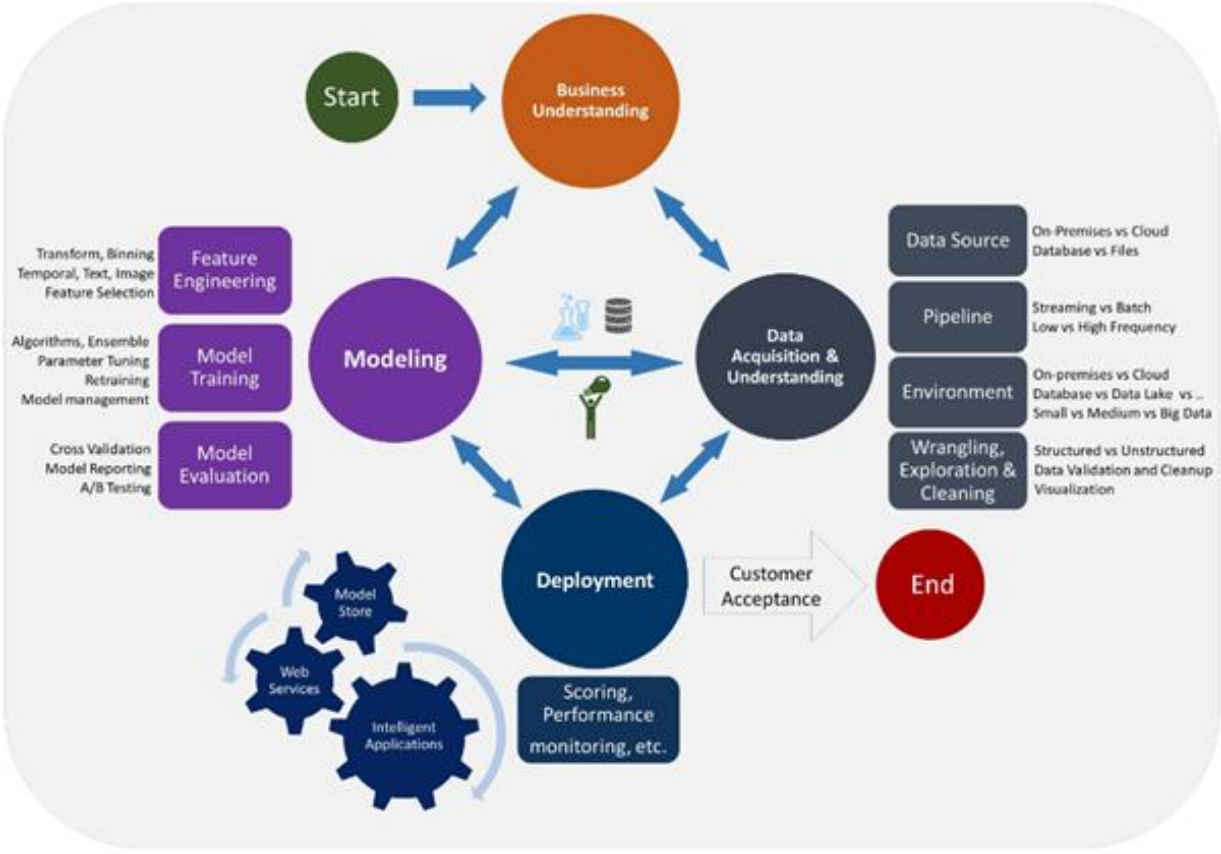


Figure 18: Microsoft Team Data Science Lifecycle (Microsoft)

The rest of this section will be dedicated to the 6 stages laid out in CRISP-DM, the most commonly implemented methodology for data science projects.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

1. Business Understanding

This first phase is a fundamental project management step and deals with identifying objectives and requirements of the project from a business perspective. If can be sub divided into 4 tasks.

- Determine business objectives of customer
- Assess the situation in terms of risk & contingency, requirements, available resource, cost-benefit analysis etc.
- Determine data mining goals to determin what success conditions are, and
- Produce a project plan defining technologies and tool chains etc.

2. Data Understanding

The next stage involves understanding the available data with the goal of identifying and collecting the data sets required for the successful completion of the project. Again, this stage has for distinct tasks:

- Initial Data Collection
- Data description (Format, identities etc)
- Data Exploration
- Data Quality Verification

3. Data Preparation

Data preparation, also referred to as “Data Munging” is often considered one of the most important stages of the project and deals with the preparation of the data sets for modelling. It is broken down into five distinct tasks:

- Data selection – selection of the appropriate data sets and documentation of rational for selection (or for not selecting other sets).
- Data Cleansing – often the most resource intensive part of the project
- Data construction - to derive new attributes from existing data
- Data integration – combining multiple data sources to create new data sets
- Reformatting data where necessary

4. Modelling

This is considered to be the most interesting part of the data science project but is often the shortest in practice. The modelling techniques used in this stage are various and which technique used will depend on the characteristics of the project itself. CRISP-DM recommend iteration of models until the best fit is found. This phase has 4 tasks:

- Modelling Technique / Algorithm Selection
- Test Design – splitting data for training purposes etc.
- Model Build
- Assessment – to determine the success of the chosen model

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

5. Evaluation

Although task 4 of the previous phase deals with assessing the model, this is done from a technical standpoint. This phase deals with evaluation from a business perspective and involves three tasks:

- Evaluate the results against the business criteria to determine success or failure
- Review process to identify errors or missed steps
- Determine next steps, i.e. to continue iterating or to move on

6. Deployment

The final stage is the deployment of the model so that it can be put into production and the business can reap the rewards of this hard work. The specifics of each deployment will vary in complexity depending on the specifics of each project, however there are 4 stages to this phase:

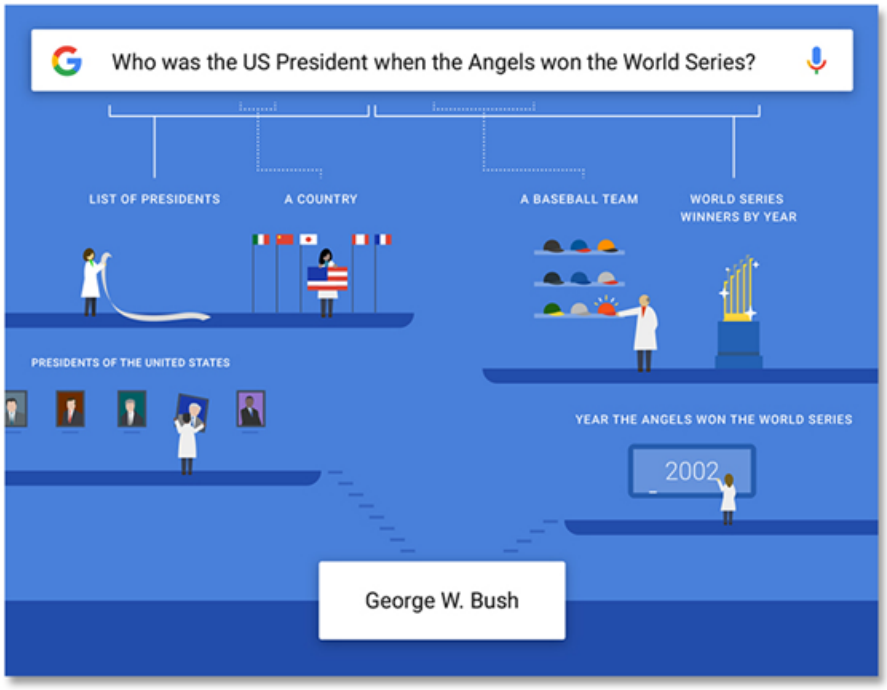
- Deployment Planning
- Monitoring and Maintenance Planning
- Final Reporting
- Project Review

5.3 Knowledge Graphs

Euclidian Geometry deals with the distance between points, for examples shapes in 2D space or 3D space. For many applications Euclidian Geometry works perfectly, for example in image processing where pixels in an image have concrete 2D coordinates. However in the real world there are many examples of non-Euclidian data which also has to be processed for successful machine learning cases. These data include structures such as meshes, point clouds, trees and graphs, which can be found in social media, images, document classification, e-commerce etc. Trees in particular grow exponentially and are therefore non-Euclidian. Hierarchies are another data structures not represented in Euclidian space and therefore are not suitable for CNNs or RNNs. This is explained in great detail in two key papers from 2017: Poincaré Embeddings for Learning Hierarchical Representations (Kiela, 2017) and Neural Embeddings of Graphs in Hyperbolic Space (Neural Embeddings of Graphs in Hyperbolic Space, 2017) which promoted the use of hyperbolic space for graph representation learning in ML (Fred Sala, 2019). Although exact definitions for KG vary, an n inclusive definition can be found in (AIDAN HOGAN, 2021)where the authors define a knowledge graph as “a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities”

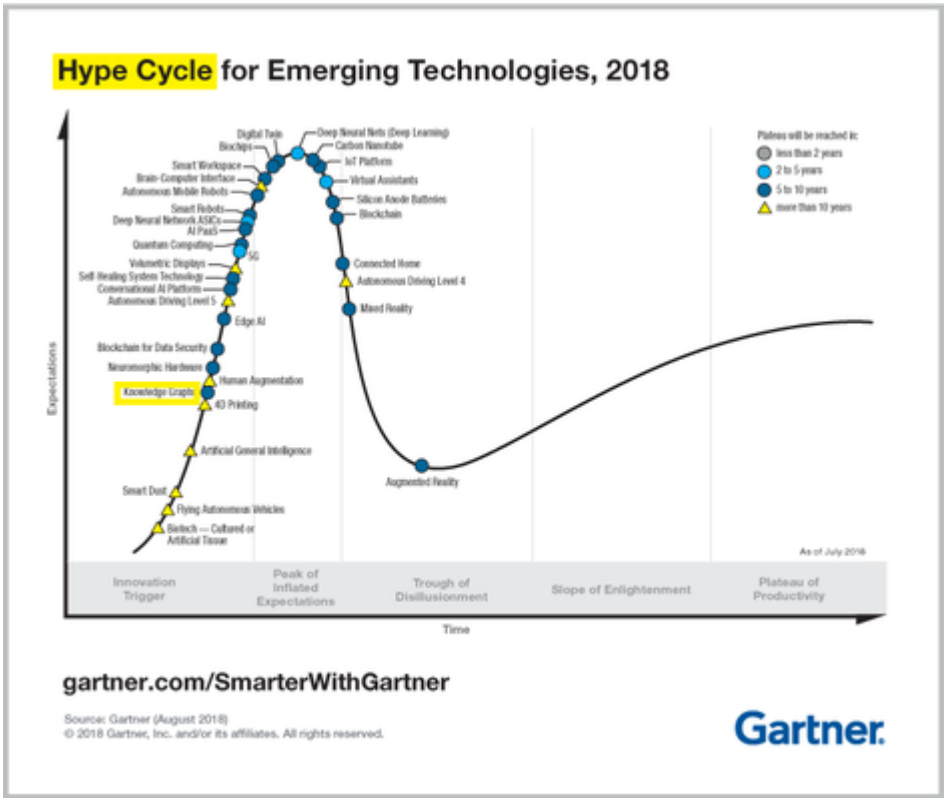
Although not new, the modern discussions of knowledge graphs began with an introduction of the concept by Google (Singhal, 2012) with a discussions of how Google was using knowledge graphs in search: helping to narrow search results to relevant terms, understanding queries in order to offer summarised information relevant to the search and presenting deeper information about the search query.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.



Source: Google

Companies like Airbnb, Facebook and Uber followed suit and in 2018 Gartner confirmed that Knowledge Graphs were ascending the Hype Cycle for emerging technologies, see Figure 21.



This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Figure 19: Although they have been around for 20 years, Knowledge graphs are climbing the hype cycle according to Gartner Research

Knowledge graphs (also called semantic graphs) are a specific subclass of graphs which come with metadata, schema, and global identifier capabilities. Graph-based learning systems are commonly employed in ecommerce recommender systems as they can easily exploit interactions between users and products. In the field of drug discovers, molecules are modeled as graphs, and their bioactivity needs to be identified for drug discovery, a task suitable for application by KGs. Another popular area of application is in the research field where papers are linked to each other via citations. Graphs offer a method of categorizing these papers into different groups. (A Comprehensive Survey on Graph Neural Networks, 2021).

The rest of this section deals with the technologies behind Knowledge Graph implementation and their use in databases, as well as Graph Maintenance.

5.3.1 Knowledge Models

A Knowledge Model is a set containing differing types of concepts, procedures, principles, skills and facts and has a structure that represents the relationships among each item. It typically contains three types of knowledge, Domain knowledge, Inference knowledge and Task knowledge and will also contain a full domain schema diagram (e.g. UML class diagram, Ontology, ER data model), an inference-structure diagram, a list of knowledge roles as well as textual and graphical specifications of the tasks and task methods (Chen-Burger, 2011). Knowledge models are interpretable by computers and can contain knowledge about processes or even information about a specific product.

FOAF

FOAF (friend-of-a-friend) (Dan Brickley, 2000) was started in 2000 and is considered the first Semantic Web project. It is a descriptive language that uses Resource Description Framework (RDF) and the Web Ontology Language (OWL) in order to allow computers to find knowledge through definitions of relationships. For example, it could be used to find all faculty members that teach Computer Science and that are in the same social club as a given person, it could also be used to find all employees who have worked on projects with a specific employee.

In a 2007 essay (Lee, 2007), Tim Berners Lee described FOAF as a “start of the revolution”. In this essay he described the semantic web as a Giant Global Graph, in the RDF creates a graph, rather than a web and it is the knowledge in this graph that is the most interesting aspect of the semantic web.

> In the long-term vision, thinking in terms of the graph rather than the web is critical to us making best use of the mobile web, the zoo of wildy differing devices which will give us access to the system. Then, when I book a flight it is the flight that interests me. Not the flight page on the travel site, or the flight page on the airline site, but the URI (issued by the airlines) of the flight itself. That's what I will bookmark. And whichever device I use to look up the bookmark, phone or office wall, it will access a situation-appropriate view of an integration of everything I know about that flight from different sources. The task of booking and taking the flight will involve many

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

interactions. And all throughout them, that task and the flight will be primary things in my awareness, the websites involved will be secondary things, and the network and the devices tertiary. - Tim Berners Lee

schema.org

Schema.org is an open-source community founded by Google, Microsoft, Yahoo and Yandex with the goal of promoting schemas for structured data on the Internet, on web pages, in email messages, and beyond (Schema.org). It can be used with RDFa, Microdata and JSON-LD encoding, among others and covers entities, relationships between entities and actions. Currently over 10 million sites use Schema.org vocabularies to markup web pages and email messages including numerous applications from Google, Microsoft, Pinterest and Yandex among others.

Vocabularies can be extended by the community in order to provide a decentralized extension of the vocabulary, called Hosted Vocabularies. However, in 2019 the decision was made to simplify Schemas and provide assurances that a vocabulary was in fact part of Schema.org and as such all of the previous hosted extensions have been folded into schema.org.

More information, including extensive documentation and a list of schemas can be found at https://schema.org/

FIBO

FIBO, which stands for the Financial Industry Business Ontology (FIBO) is a conceptual model for business created to describe how all financial instruments, business entities and processes work in the financial industry. It was created by the Enterprise Data Management Council in order to solve the problem of common terms with different meanings and vague definitions that don't capture critical nuances for products and services within the financial services industry and ultimately make it difficult to compare data and services in a meaningful manner. FIBO is described as a "formal model of the legal structures, rights and obligations contained in the contracts and agreements that form the foundation of the financial industry" (EDMCouncil). FIBO is standardized by the Object Management Group (OMG) and is developed as an ontology in the Web Ontology Language (OWL). FIBO is based on Description Logic in order to ensure that each FIBO concept is framed in a way that is unambiguous as well as readable both by humans and machines.

Since January 2020 the FIBO community has used GitHub as a focal point of development and discussion. More info can be found here: https://github.com/edmcouncil/fibo

5.3.2 Multi-model graph databases

There are a number of database types in existence and which database is implemented in any particular situation largely depends on the requirements of the situation. Traditionally most databases in common usage have been relational databases and rely on SQL as a langue for executing transactions, however other models such as NoSQL and Graph databases have become popular in the last few years.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Regardless of the type of database used, all databases implement transactions in order to populate, access and asses the data contained within. These transactions can involve reading or writing data in to tables in the database or can be more complex involving unions and joins of tables to create new relationships between data. In some cases, data security and consistency is the prime concern and therefor these transactions require ACID guarantees (atomicity, consistency, integrity and durability); for example, financial transactions or supply chain updates. Other transactions do not need such strict guarantees and can accept less data accuracy in order to gain speed and scalability (think of social media posts). The latter type of dataset will use a BASE (Basic Availability, Soft-state, Eventual consistency) consistency model where it is ok to use out of date data and give approximate answer are time of request.

Multi-Modal Databases

Different databases are appropriate in differing use cases and as such, different databases are often implemented in order to answer these use cases. The problems with this methodology is that they become complex quickly and suffer a problem known as Polyglot Persistence. In essence, polyglot persistence occurs when using different storage mechanisms to handle storage in one system and can lead to a failure of consistency across these diverse storage systems. One example of a polyglot stack could include Redis or MemSQL as a caching layer, Couchbase or MongoDB for collecting logs, MySQL for metadata, and SOLR for indexing and search (Prasad, 2020). IN this situation it is impossible to provide consistence across all systems and therefore ACID guarantees connot be provided.

Multi-modal databases provide an answer to this problem by combining different types of database model into one integrated database engine, offering a back end to access multiple data models such as Grahp, Relational, Document, Object Oriented etc. They are flexible and as such can reduce complexity and costs. They can also provide ACID guarantees and components can be scaled independently of one another.

- ArangoDB (https://www.arangodb.com),
- MarkLogic (https://www.marklogic.com),
- OrientDB (https://www.orientdb.org) and
- Azure Cosmos DB (https://azure.microsoft.com/en-us/services/cosmos-db),

are some of the top-rated multi-modal database providers on the market at present.

Graph Databases

Until recently, most of the world's databases were relational and ran on SQL. In 2005 Big Data as a term came into existence (Foote, 2017) and the requirements for data storage and retrieval changed. Hadoop became popular because of its ability to handle structured and unstructured data. Although invented in 1998, it wasn't until after the invention of Big Data that the popularity of NoSQL started to explode given, its ability to handle unstructured web data quickly. Around this time, it was adopted by organizations such as Facebook, Twitter, LinkedIn, and Google in order to analyse their vast quantities of unstructured data.

Graph databases have risen to popularity recently with the AI boom. In traditional relational databases, information is stored in blocks, called tables. It is easy to recall information but the relationships

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

between data is not so clear. It can be accessed through database operators on tables, however at scale this can become a complex and resource intensive task. Graphs, on the other hand have this information readily available and therefore perform much better as data grows large, i.e., Big Data. Whereas relational databases tend to slow down when finding relations between data points at scale, graph databases offer constant performance. What's more, nodes and edges can easily be added to graph databases allowing them to scale and evolve with a project in an agile manner. As graph databases find their use in various cases, one report from Gartner predicts that "The application of graph processing and graph DBMSs will grow at 100 percent annually through 2022 to continuously accelerate data preparation and enable more complex and adaptive data science." (Gartner, 2019). Graph databases are being implemented in use cases as varied as recommender systems for streaming media services, medical research, knowledge management and fraud detection.

- Neo4j - http://neo4j.com/
- ArangoDB (https://www.arangodb.com)
- Dgraph - https://dgraph.io/support
- OrientDB - https://orientdb.com/

Are examples of the top-rated graph databases available at present.

Graph databases as Multimodal databases

Graph databases have been most successfully used in situations where analysis is key – i.e., analytical workloads. This is due to the fact that they are not best suited to IUD (insert, update, delete) transactions, even though they do support them, and are better suited to storing and fetching data. To this end manufactures or relational and NoSQL databases offer incorporated graph data structure as part of a multi-modal database, graph venders do not currently offer backend functions for transactional or streaming workloads.

5.3.3 Graph Maintenance (Graph Merging)

As mentioned in the previous section, graphs are best suited for storage and retrieval operations and not for IUD, however over time graphs need to be maintained or merged. This is not a trivial tasks as different graphs may vary in their terminology when referring to the same entities. Given the complexity of this task, automated techniques such as entity alignment have come to the fore.

Deepmatcher (Jorge Decorte, 2020) is one such technique developed with speed in mind. It is a Python based entity and text matching package which utilises built-in neural networks and utilities to allow users to train and apply state-of-the-art deep learning models for entity matching.

In a 2020 paper (Qi Zhu, 2020), the authors propose a novel collective aggregation function which relieves the incompleteness of knowledge graphs and scales up efficiently with the mini-batch training paradigm and an effective neighbourhood sampling strategy which improved upon the best-performing of ten baseline systems by 10% (Wei, 2020). They created a NN with a self-attention mechanism to learn which attributes of an entity are most important for distinguishing it from entities that look similar. This solution was built on Amazon's DGL – Deep Graph Library (https://www.dgl.ai). The authors in

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

this case reduced computations complexity by adding weighted sampling meaning that although they were comparing two-hop neighbourhoods, the neighbourhood was restricted and therefore the computational load was reduced without affecting performance.

Promising results have recently been published in the field of entity alignment using embedding-based entity alignment methods. These methods represent knowledge graphs as low-dimensional embeddings and perform entity alignments by measuring the similarity between entity embeddings (Embedding-Based Entity Alignment Using Relation Structural Similarity, 2020).

In another paper from 2020, the authors surveyed 23 recent embedding-based entity alignment approaches and categorized them based on their characteristics and the techniques used. They also proposed a new knowledge graph sampling algorithm. The authors propose future research be directed in 4 key areas:

7. Unsupervised entity alignment – to get around the requirement of seed alignment. Active and abductive learning are also proposed to this end.
8. Long-tail entity alignment
9. Large-scale entity alignment to reduce training time on large datasets, and
10. Entity alignment in non-Euclidean spaces (alignment-oriented non-Euclidean KG embedding models)

5.3.4 Visualization techniques

Visualization is a graphical representation of the products or digital environments that can be portrayed either on a physical device or in system.
Enterprises from different industries are taking advantage of Digital Twin (DT)'s ability to realize Industry 4.0 requirements. They acquire data models to provide input for visualization of real-time environments, software analytics, computational intelligence and 3D Representations.

Visualization techniques are used to produce insights for analysis of data models and algorithmic simulations. Visualization is applied in mainly two categories:

1. Equipment/Mechanical Status
2. Process Automation Status

For equipment/mechanical status, virtual modeling concept idealizes CAD (Computer Assisted Design) and CAE (Computer Assisted Engineering) simulations.

For process automation status, data visualization platforms can be used to create or offer capabilities. Data Visualization Techniques and Platforms are summarized in (Toasa1et all, 2018).

Techniques are given as follows in Table II (Toasa1et all, 2018):

1. Autocharting
2. Correlation Matrix
3. Network Diagram
4. Sankey Diagrams
5. Visualization for Mobile Devices

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

6. Word Cloud

Platforms are also given as follows in Table III (Toasa1et all, 2018):

1. Google Analytics
2. SAAS Visual Analytics
3. Sisense
4. Tableau
5. Zoho Reports

These platforms help users to make good graphics in a short time, although most of them are very expensive and difficult to manipulate by non-technical users. They also provide 3D support that leverages the understanding and interpretation of the data, knowledge and information. 3D Visualizers that can import data from 2D CAD models are also spreading to strengthen the analysis of data
In a recent review, (Lim, K.Y.H. et Al. 2020) mentions about AWS tools (Yuqian Lu and Xu 2019) and Elastic Stack (ELK) (Damjanovic-Behrendt and Behrendt 2019) for data analysis and visualization . ELK comprises of ElasticStack, Logstash and Kibanafor searching, analysis and visualizing data in real-time.

5.3.5 VR/AR

We are surrounded by data and the vast quantities of data that are being produced by the second poses the very real problem of analysis. How do we make sense of this data in ways that are intuitive yet provide a deep understanding of what the data has to say? One method that is particularly appealing and intuitive to humans is data visualisation – essentially the representation of data in the form of graphs, charts maps and so on. Data visualisation is nothing new, in fact one of the most common tools for data visualisation in use today is perhaps one of the oldest; Microsoft Excel. Excel offers data visualisation in the form of graphs and charts, such a pivot tables, bar charts and pie charts. Whilst useful for small data sets, Excel gets quickly overwhelmed when it comes to big data.

To process big data more powerful tools are needed and at present there are various options on the market, both proprietary and open source. Examples include:

- Juypter – Open Source (http://jupyter.org)
- Tableau – Paid Software (https://www.tableau.com/)
- Google Chart – Freeware (https://developers.google.com/chart)
- D3.js – JS Library (https://d3js.org/)

Visualization in general helps with understanding data by allowing vast amounts of data to be summarized in visual form with which humans can interact and adjust on the fly (with filters etc.). This trend has not gone unnoticed by Microsoft who in 2018 launched a tool called Charts3D for PC and surface Hub. The product comes from the Microsoft Garage (Microsoft) Innovation program and allows for visualization of big data in the form of a 3D, interactive graph or chart which allows users to zoom, rotate, or pan different graphs and charts using mouse or touch to quickly find the most interesting aspect of their results (Viswav, 2018).

The recent boom in Augmented Reality (AR) and Virtual Reality (VR) technologies has highlighted the aptness of this technology for the task of data visualization. One of the key adopters of this technology is NASA, describing its use in a 2018 paper (Science Data Visualization in AR/VR for Planetary and

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Earth Science, December 2018) as “VR/AR will enable NASA scientists and engineers to more intuitively visualize and manipulate data in many different domains”. The AR/VR team at the NASA Goddard Space Flight Center (GSFC) are developing a tool called PointCloudVR to aid with visualization of space data including “visualizing flow data, such as convection flows or ocean flows”.

Another use for VR/AR visualization techniques is that of visualizing network operations. 3Data, a startup based in Austin, Texas is doing just this, blending AI with AR/VR for advanced visualization and can be used to detect anomalies in data-flows on a network to stop cyber-attacks such as DDoS attacks. More info can be found on their website at https://www.virtualitics.com

A startup founded in 2018 with a grant from AR Company MagicLeap is using data visualization techniques to process signals coming from WiFi and phones and make them visible to users wearing MagicLeap headsets in order to visualize network coverage and eliminate blind spots. They describe the process as using “patent-pending sampling algorithms, data spatialization, and a touch of machine learning” (BadVR , 2019). Since 2018 the company now offers three products, an immersive data platform to visualise

- 5G networks: Immersion can visualize full-scale deployments, coverage models, and even live usage activity.
- Public Safety: visualise and understand the impact of municipal decisions and emergency response, and coming soon,
- Cyber-Security: understand impact immediately, across every channel

More information can be found on their website: https://badvr.com/

Finally, in the finance sector, Boston-based startup Immersion Analytics has created a tool for the Magic Leap that allows users to combine up to 7 dimensions (Database columns) in one single view. They started by targeting the finance sector with a focus on risk management, alpha generation as well as portfolio management and quantitative research. They have since expanded to provide Immersive Analytics for Business Intelligence, Financial Markets, Defence, Cybersecurity and AI Governance. Their software is available in 3D versions on pc and other 2D platforms as well as smartphone and VR/AR devices and integrates with Qlik Sense for business analysts, and MATLAB and Python for data scientists.

More info can be found on their website: https://www.immersionanalytics.com/

5.4 Metadata-driven architectures

5.4.1 Automatic extraction and enrichment of metadata

In the earlier days of automatic data extraction, technologies such as Named Entity Recognition (NER) were used to find occurrences of entity types. This worked well for finding "mentions" of entities within content, but it comes up short when it is critical to understand the context of those entities. For example, a lease document may mention two organizations; NER can find those, but it cannot indicate which is the lessee and which is the lessor. Another example, the same lease document may mention several dates; NER can find those dates, but it does not indicate which is the commencement date, or which is the termination date.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

In the recent years, the state of the art has moved from basic "data extraction" to "metadata extraction", that is, data within context. Various techniques have been employed combining natural language processing, machine learning, and deep learning, to either extract entities within a given context, or to extract entities and then imply their context.

Automatic metadata extraction from textual content is typically approached differently for structured content versus unstructured content. Structured content includes short, semi-tabular content, for example: forms, invoices, and purchase orders. Unstructured content includes typically longer content that contains a significant proportion of natural language, for example: legal documents, reports, and specifications.

Technologies used to extract metadata from structured content are sometimes referred to as data capture or document capture and are often coupled with scanning and OCR technologies. The techniques used for metadata extraction often involve a combination of several heuristics including: prefix and suffix rules, pattern matching (e.g. regular expressions), partial matching, and list matching. Machine learning is also used often, with a feature set that is focused less on the textual content and more on the document itself including: zones, graphical elements, and tabular structures.

Technologies used to extract metadata from unstructured content rely much more heavily on natural language features. Using machine learning, the natural language context of each specific metadata element can be modelled. This model then can be used to find and measure the fitness of candidates in unseen content.

Because of these two very different techniques for structured and unstructured content, often a classifier is used to decide which extraction techniques is warranted. As well, a single piece of content can contain both structured portions and unstructured portions. In this case, the best extraction technique can be predetermined for each specific metadata element and/or the content is subdivided, and each portion is processed independently.

5.4.2 Intelligent information management

The idea of intelligent information management is simple: store, manage and track files and documents with an intuitive software platform. But there are some essential concepts which are important to understanding the world of intelligent information management solutions.

The Evolution of Information Management

From the advent of the first computer ENIAC in 1944 until now, one constant remains – computers need to store data. The original ENIAC machine had 18,000 tubes, but no system to store memory. Later though, paper punch cards were used for external storage and information could be captured and saved.

Fast forward fifty years and the need to store ever-growing volumes of business information gave rise to a class of software solutions designed to do just that. This class of software has gone by many names, each having a slightly different meaning. These terms include:

- Document Management System (DMS)
- Enterprise Content Management (ECM)
- Content Services Platform (CSP)
- Intelligent Information Management (IIM)

DMS, ECM, CSP, IIM… While there are plenty of acronyms, it is worth understanding the evolution represented by each term.

Document Management System

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Before the proliferation of modern computers, companies often employed people to the position of File Clerk – someone that would file documents in cabinets and retrieve them for employees upon request. A DMS can be viewed as a fancy electronic equivalent of a human file clerk.

A DMS refers to software that is used to store, manage, and track electronic documents and electronic images of paper-based information captured using a document scanner. The Association of Intelligent Information Management (AIIM) says of document management systems: "Document management is one of the precursor technologies to content management, and not all that long ago was available solely on a standalone basis like its imaging, workflow, and archiving brethren. It provides some of the most basic functionality to content management, imposing controls and management capabilities onto otherwise 'dumb' documents."

Enterprise Content Management

In contrast to a DMS – which usually means just the software technology – enterprise content management refers to the entire playbook of managing enterprise information. AIIM defines ECM as follows: "Neither a single technology nor a methodology nor a process, it is a dynamic combination of strategies, methods, and tools used to capture, manage, store, preserve, and deliver information supporting key organizational processes through its entire lifecycle."

In essence, ECM refers to an overarching strategy that, firstly, encompasses all different classes of information, not just documents – things like email, graphics, pictures, web content, video, multimedia records, and other resources. Secondly, ECM usually has the added components of document lifecycle (applying timelines to content) and workflows (incorporating files into processes).

Content Services Platform

In 2018, Gartner further expanded the scope of traditional ECM into the term Content Services Platform. According to Gartner, ECM and CSP represent two distinct ideas. This is from their Reinventing ECM research:

"Content services are a set of services and microservices, embodied either as an integrated product suite or as separate applications that share common APIs and repositories, to exploit diverse content types and to serve multiple constituencies and numerous use cases across an organization."

Gartner further delineates three categories of content services: platforms, applications, and components – all interoperable pieces to a holistic content management strategy. In other words, Enterprise Content Management always had the mission of achieving a wide array of content-related operational goals using a centralized platform; Content Services Platforms, on the other hand, embody a new approach – one focused less on storing documents centrally and more on the strategy an enterprise uses to deal with their growing content, data and document needs.

Intelligent Information Management

The latest evolution – called intelligent information management – is where the most modern solutions reside. According to AIIM, IIM effectively combines the following key capabilities:

- **CONTENT SERVICES:** A flexible and modular approach that utilizes content and information wherever and whenever it is needed, independent of where it is stored.
- **PROCESS SERVICES:** Tools that can be delivered with the simplicity of an app, but within a framework that allows the business to remain in control.
- **ANALYTICS SERVICES:** Automated tools to prepare all information – both structured and unstructured – for machine learning.

IIM is not a single repository for information or even a set of standards. IIM systems are flexible in that they connect to existing repositories to present information, no matter where it lives – email, CRM, network folders, etc. Furthermore, the addition of artificial intelligence and machine learning to help classify and manage otherwise unstructured information is a hallmark of IIM.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Key concepts on which Intelligent Information Management is built

Metadata

Metadata is a set of fields that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage information.

The easiest example of metadata is a streaming service like Netflix. Each film or show is tagged with descriptors like title, genre, cast, year, director. The presence of metadata makes it simple for users to search and find the content they want.

The same applies to metadata on files and documents. In an IIM solution, every file has a metadata card – where the information can be classified and described. An invoice, for example, might have a metadata card that includes vendor, invoice date, paid date, cost, description.

Metadata is arguably the most important concept in intelligent information management. It underpins much of the functionality that drives efficient information strategy.

- **CLASSIFY:** Organize documents and files based on what they are.
- **SEARCH:** Users can search with just a keyword and find information, no matter where it is stored.
- **PERMISSIONS & SECURITY:** Define sensitive information and create custom permissions that secure the content.
- **WORKFLOWS:** Define processes like review-approve-sign or learning requirements with metadata.

Folderless

The repository-neutral approach means no more storing files in computer folders and subfolders. It is a novel concept for people that have used the traditional folder structure their entire lives.

Folder-based storage has always relied on each individual employee's organization skills and memory, but the digital information age has pushed even the most proficient employees to their limits. Besides the need to deal with huge quantities of documents and information, they must now anticipate how each document will be searched for by others in the company.

For example, consider the case of saving a new sales document. Should the proposal for Project ABC at customer XYZ's site in Australia be filed under "Proposals" or "Company XYZ" or "international business opportunities"? To make sure that the proposal will show up in various folders, the document author might duplicate the document and store it in each relevant folder.

IDC (International Data Corporation) added up the time wasted on information searches, estimating the cost to be $19,732 per information worker per year. The world is on the verge of a paradigm shift – where file location becomes less important and the content takes center stage.

Artificial Intelligence

Automation can help save time from monotonous manual document handling. AI can help tag information with correct metadata and enrich the metadata for future purposes. As 73% say their organization has no clear guidelines on how documents should be labelled, AI can help keep consistency.

Various Intelligence Services can help automate tedious office work, for example:

- **CLASSIFIER**: Automatically suggests a Class for an object, such as "Template", "Agreement", or "Offer".
- **INFORMATION EXTRACTOR**: Automatically scans through the content of a document to suggest relevant metadata.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

- **IMAGE ANALYSIS**: Automatically examines the image to suggest captions, categories, tags, and facial analysis results as metadata suggestions.

At its best, AI is fully self-learning and no configuration required, right out of the box. It is all about making users' work easier – allowing them to do their job without worrying about the AI itself or annotating documents or configuring it manually.

What if the information management solution could just learn from what users are doing? What if it could just listen and learn? The best of breed AI modules do just that; they automatically dive into a repository full of documents and learn.

Essentially, modern intelligent information management solutions offer ways to automate the time-consuming and often boring document-driven processes with AI. Artificial intelligence systems can easily churn through lots of information to recognize patterns and categories in the data. That ability is put to work to enable new ways to search, find, use, and manage information, and add automated workflows to document management processes.

Structured vs. Unstructured

Not all data is created equal. The data contained in Word documents and PowerPoint presentations is vastly different than point-of-sale data or a phone number directory. Data is classified as structured data vs. unstructured data, and each classification has bearing on how it is collected, processed, and analyzed.

Structured data – or *quantitative* data – is the type of data that fits nicely into a relational database. It is highly organized and easily analyzed. Most IT staff are used to working with structured data.
Structured data typically means things that would sit nicely in a spreadsheet. Examples include:

- Dates
- Phone numbers
- ZIP codes
- Customer names
- Product inventories
- Point-of-sale (POS) transaction information

Its inherent structure and orderliness make it simple to query and analyze. Common applications that rely on structured data in relational databases include CRM, ERP and POS systems.

Unstructured data – or *qualitative* data – is just the opposite. It does not fit nicely into a spreadsheet or database. It can be textual or non-textual. It can be human- or machine-generated. Examples of unstructured data include:

- **Media:** Audio and video files, images
- **Text files:** Word documents, PowerPoint presentations, email, chat logs
- **Email:** There is some internal metadata structure, so it is sometimes called semi-structured, but the message field is unstructured and difficult to analyze with traditional tools.
- **Social Media:** Data from social networking sites like Facebook, Twitter, and LinkedIn
- **Mobile data:** Text messages, locations
- **Communications:** Chat, call recordings

Here is the two-fold, compounding problem: unstructured data is important. The volume of unstructured data is growing – and that growth is accelerating. Right now, IDC suggest that anywhere from 80–90% of data is unstructured.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

With a modern IIM solution, unstructured data becomes accessible, searchable, and available. By applying structure in the form of metadata, companies render that information relevant. Metadata is the key to the castle. It describes what the data is, how it relates to other data, key data points within documents and where in a particular business process that data fits.

When unstructured data is accessible, searchable, available, and relevant, it is converted into information that an enterprise can use to make better decisions. Organizations can essentially exploit the power of unstructured data with a modern IIM solution.

Workflows

Workflows help to free up working time by automating business processes. Automated workflows streamline common business processes – like contract approvals, controlled content, and invoicing – so users can stay productive and ensure compliance.

A modern IIM solution automatically sends a notification when there is a task that needs to be handled, and automatically monitors each step of the business process. All workflow steps are tracked in the version history of each data object, providing audit-proof business while gaining full visibility into all important business processes.

5.4.3 Digital platforms for knowledge utilization

Flexible Deployment: Cloud, On-Premises or Hybrid

What should a preferred IT environment for knowledge utilization look like – cloud-based, on-premises or hybrid? Some organizations need the flexibility of the cloud. Others may be highly regulated and required to maintain sensitive information on-premises for compliance reasons.

With a modern IIM solution, the organizations can choose a combination of secure cloud-based repositories and mix and match them with on-premises servers while enjoying the same user interface and experience across the business. Secure and easy way to connect cloud repositories and systems with on-premises documents and data is also required. There is a clear demand for true hybrid cloud content services platform.

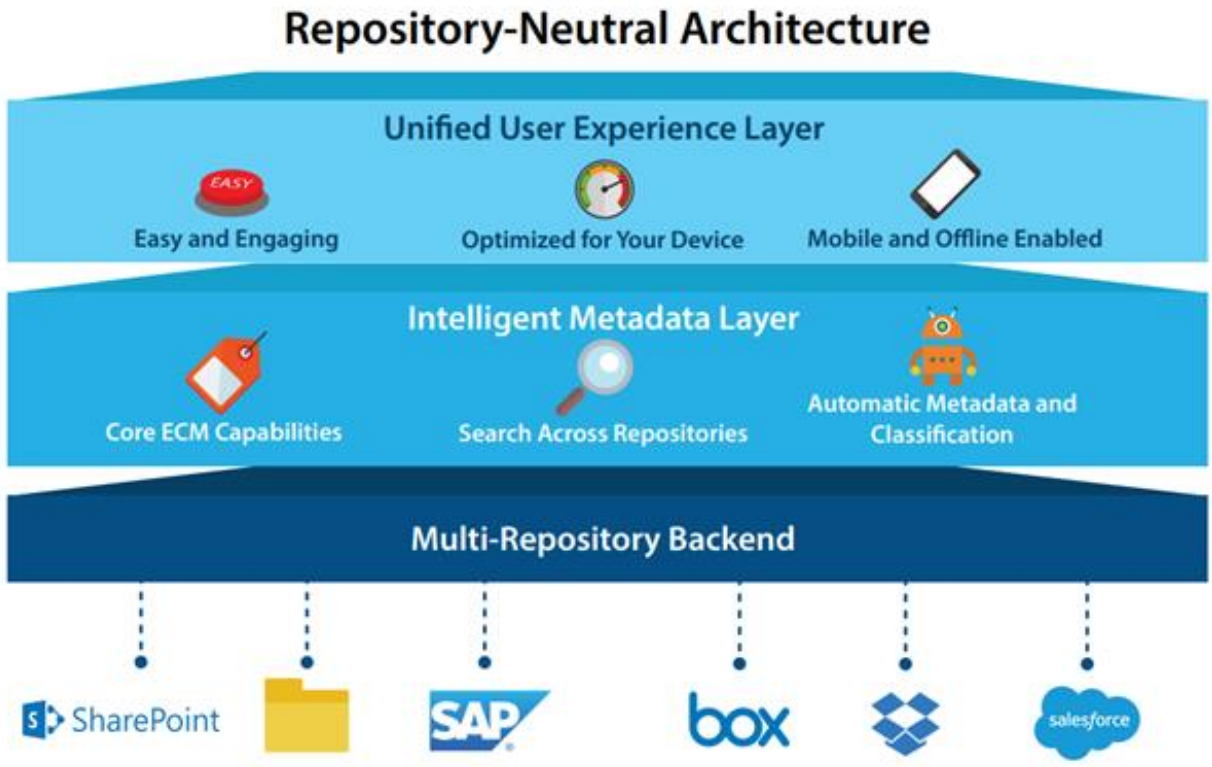
Repository-Neutral

The “old,” or traditional, approach to managing information has been primarily based on where content or information is stored. In fact, for essentially traditional information management systems, to get value from the system, the content has to reside in that system.

As an example, M-Files intelligent information management solution features a repository-neutral intelligent metadata layer that unifies information across the enterprise based on context, not on the system or folder in which it is stored.

This new methodology allows content and data to remain in place, undisturbed, so that users of existing system can continue to work uninterrupted, while also allowing information to be enriched for evolving needs and new use cases.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.



Front-End Neutral

The concept of front-end neutral is where information management capabilities are integrated seamlessly into the user interfaces of other applications – like Office 365, Salesforce, Google Workspace, SharePoint, Teams, NetSuite, SAP, QuickBooks, Workday or Esri ArcGIS, for example. In addition to the context established by metadata and relationships to important business objects like customers, projects and cases, a new layer of context is added to the picture, the context of the user interface of the line-of-business application.

Essentially, it means that businesses recover productivity lost to context switching, which we all have suffered from – navigating from application to application in search of a document. Users of a modern IIM solution can access the entire information ecosystem from within key applications like Salesforce, Microsoft Office, Teams, and others.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

6. DLT

This section of the paper looks at the state of the art of Blockchain. It starts with a definition of the terms DLT and Blockchain (clearing up a common misconception) and following into a description of the essential elements of Distributed Ledger Technology. It then takes a look at what is known as DLT 3.0 – the inclusion of AI on the chain and finishes with a study of Business Process Planning with DLT.

6.1 Intro to SotA DLT (ES)

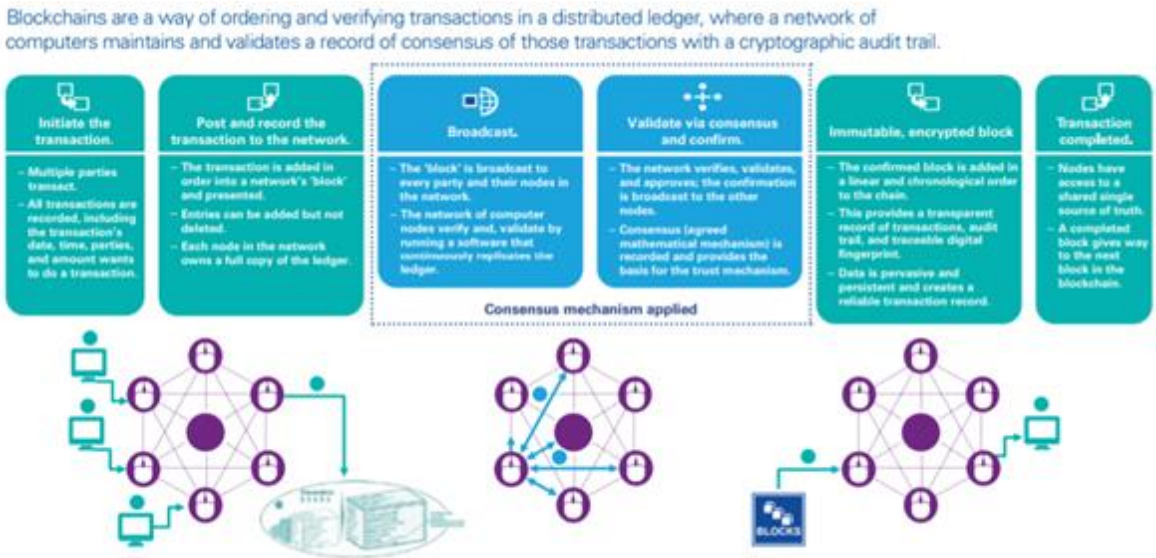
This section provides an overview of Distributed Ledger Technology where relevant to the OXILATE project. It starts with a definition of what DLT is (compared with Blockchain) and then continues with an explanation of what DLT is and the various forms of DLT on the market. A general look into use cases where DLT is most advantageous is provided as well as a brief overview of current DLT providers. This chapter finished off with a look at the latest iteration of DLT, commonly called Blockchain 3.0 and highlights where this technology is going.

6.1.1 DLT Vs Blockchain

Although often used interchangeably, it is a common misconception that Blockchain and Distributed Ledger Technology (DLT) are the same thing, they are not.

1. A distributed ledger is a distributed, decentralised database that is shared among all participants. In essence, a distributed log of records with no one central authority, therefore it is more transparent and offers a higher level of security, being that it is harder to hack or falsify.
2. Blockchain is a type of DLT but it uses a structure of blocks, each block dependant on the previous block, forming a “chain” of data.

A comprehensive description of Blockchain (Sigrid Seibold, 2016) can be found in Figure below:



This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Figure 20: Description of Blockchain in the DLT framework by KPMG

Blockchain is immutable; data can never be erased from the chain. DLT can allow deletions from the database. For this reason, DLT does not have to follow a sequence – block-after-block construction. It also does not have to be token based which Blockchain generally is.

6.1.2 What is DLT

Distributed Ledger

A centralised ledger is perhaps the most common form of ledger and can be found in most institutions and companies operating today. That is to say, it is a ledger saved on a central database and administered by a central authority. Transactions, changes and updates, carried out on the central database are verified by the central authority and users implicitly have to trust this authority. Although centralised databases are efficient and scalable, they can be prone to attack, human error or force majeure. If a centralised database is attacked / hacked, then the database will cease to function or will display fraudulent or corrupted data for all who access it.

On the contrary, a distributed leger is shared among all participants in the network (across multiple sites, institutions, or geographies) and is therefore much harder to attack or corrupt data. It is however less efficient and scalable than a centralised data but as all transactions are verified with all nodes in the network, it is known as a trust-less system. Transactions need to have public "witnesses" to verify that a transaction occurred and was sent by the node claiming to have issued the transaction. The participant at each node of the network can access the recordings shared across that network and can own an identical copy of it and any changes or additions made to the ledger are reflected and copied to all participants in a matter of seconds or minutes. (MAJASKI, 2020). IBM gives the following definition of a Distributed Ledger:

> *A distributed ledger is a type of database that is shared, replicated, and synchronized among the members of a decentralized network. The distributed ledger records the transactions, such as the exchange of assets or data, among the participants in the network (Sloane Brakeville, 2019)*

Consensus Mechanisms

As Distributed Ledgers are not centralised, they cannot rely on a central authority (such as a financial institution) to verify transactions or act as a mediating party. For this reason, a consensus protocol is used between the network's member nodes in order to reach agreement on changes to the ledger.

At the heart of a consensus mechanism is the consensus algorithm and in many ways, this is what distinguishes on DLT from another. The consensus mechanism can be through of as containing two key parts, the Consensus Protocol and the Consensus Algorithm.

- The consensus protocol is a set of rules that need to be followed to arrive at consensus and in essence define who the system operates.
- The consensus algorithm is a set of instructions that produce a given output and need to be followed in the correct order to arrive at consensus.

There are various different consensus mechanisms in the literature - and more are being developed – but they all offer ways to decentralise the transactions, authenticate the identity of network participants,

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

validate the integrity of the transaction, ensure the privacy of the system (only the intended recipient can read the message) and ensures the permanence of the network under one or mode node failures (Sigrid Seibold, 2016).

There are numerous consensus mechanisms in use, including Proof-of-Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Leased Proof of Stake (LPoS) among others.

1. **Proof-of-work** requires all participants in the network to apply "brute-force" computational power to solve a complex mathematical problem and the first participant to solve the problem creates the next block, usually getting a reward – in the form of a token – for their efforts. Bitcoin.org, Etherium.org and Litecoin.org operate in this way.
2. With the **Proof-of-Stake** mechanism, "validators" are chosen at random and then validate the next block. Each validators much "stake" some of their coins in order to be chosen to validate the next block, therefore proving their benevolent intent towards the network. Decreed.org and peercoin.net are systems that rely on Proof-of-Stake. It is worth noting that Ethereumhas plans to move to a PoS model, to improve scalability, with Ethereum2.0 (Kim, 2020)
3. As it requires a lot of coins to enter the lottery to be chosen to write the next block, participants without the required funding can receive their opportunity through delegation. This type of consensus mechanism is known as **Distributed Proof of Stake** and is used by Eos.io and Steemit.com
4. **Leased Proof of Stake** is another mechanism for getting around the fact that it requires a lot of coins to write the next block by allowing participants to lease coins to other participants so that they can write the next block. This improves the consensus mechanism by ensure that it is not always the same participants who are writing the blocks. Wavesplatform.com uses this mechanism.

Hashing

Hashing is an important part of the consensus mechanism as it is used to encode data and ensure its security and validity. In its simplest form, hashing involves generating a code from an input string or set of values based on a mathematical formula.

In Blockchain, for example, the Hash is a fixed length and is generated based on the content of the block, making it unique. It is embedded in the next block written and is subsequently used to generate the following block, and so on. The actual hash function used in Bitcoin is the SHA256 algorithm (Gilbert H., 2003). I this way, each hash is based on all the previous values in the chain and therefore changing one value would mean that all the other values in the entire Blockchain would have to the changed.

Scalability

One of the major criticisms levelled at DLT and Blockchain is that it is not scalable. To be more precise, it cannot be scalable and remain decentralised and secure. On example is DLT for finance where fast transactions per second (TPS) is vital to the scalability of operations. Visa purportedly can handle over 65,000 TPS (Visa) [In reality, however it is estimated to be more like 1700 TPS - based on a calculation derived from the official claim of over 150 million transactions per day (L., 2019)]. Cryptocurrencies

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

like Ethereum can handle 30 TPS and Bitcoin a glacial 7 TPS. It has to be said that Etherium's limit is self-imposed due to gas limits (see section 6.1.8.1 for more details) and some estimates state that Etherium's limit could be pushed to 3000TPS after the Istanbul hard fork (Saint-Leger, 2019) (Mainly due to two Ethereum Improvement Proposals (EIPs) that make zk-SNARKs (iden3.io) cheaper—EIP-1108 (Antonio Salazar Cardozo, 2018) and EIP-2028 (Alexey Akhunov(@AlexeyAkhunov), 2019)). Although EOS uses a different consensus mechanism and therefore was created for faster TPS than other cryptocurrencies, after a year of operation it had grown to 4 terabytes (whilst the entire Ethereumnetwork is on 200 GB) also imposing limits on its scalability.

This trade-of has been described the "Blockchain Trilemma" in that it can only be two of Decentralised, Secure or Scalable (blockescence DLT, 2019). Unfortunately, there cannot be a one-size-fits all solution in DLT and the following paper can be referenced for its illumination of 24 prevalent trade-offs in DLT designs (NICLAS KANNENGIEßER, 2020)

This Trilemma trade-off is not necessarily a shortcoming of DLT but a characteristic, which, forces users to make a decision when implementing a particular DLT. Given that only two of the essential elements of DLT, Decentralisation, Security or Scalability, can be implemented at any given time the question becomes - which one can be traded-off.

To opt for Security and Decentralisation, i.e., all transactions are accurately and securely validated, then Scalability must be sacrificed. This means the network will not be able to operate at high speed (high number of TPS) and therefore will not be very scalable. Bitcoin.org and Etherium.org are examples of this type of Blockchain.
If speed (i.e., Scalability) is a key concern then one way to achieve this is to alter the consensus mechanism (for example DPoS) to ensure that transactions are processed faster however Decentralisation has to be sacrificed as only a proportion of the network participants get to validate transactions and write the next block. EOS.io and Ethereum2.0 are examples of this type of network.

If Decentralisation is a key issue, then Scalability and Decentralisation can be chosen at the expense of Security. This is how the IOTA tangle works – although it has been criticised for therefore having a single point of failure (Babayan, 2019).

As a key barrier to mainstream adoption of Blockchain and DLT, scalability is a problem being tackled on many fronts. Sharding (The ZILLIQA Team, 2017) is one such attempt introduced by Zilliqa (Zilliqa.com), which uses transaction and computational sharding, what it calls Network Sharding (Zilliqa, 2017) to scale to thousands of transactions per second. Essentially, it breaks nodes in to groups which can then process transactions in parallel, achieving much faster speeds. Zilliqa also uses more efficient programming languages such as Scilla (SCILLA) are being implemented to facilitate better smart contract designs. Vyper (Vyper) is another more efficient contract-oriented, pythonic programming language that targets the Ethereum Virtual Machine.

Directed Acyclic Graphs are another technology such as IOTA, are on the cusp of scaling to practical levels for the future Internet of Things. New consensus models such as Proof of Stake, Delegated Proof of Stake, and Proof of Authority are being researched and implemented as new methods of blockchain consensus.
In Bitcoin, off-chain (layer 2) solutions like the Lightning Network are already live and showing promise as future scalable solutions to transaction capacity on the network.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Side Chains

An emerging field of research into Blockchain efficiency is the Side Chain. The side chain is in effect a secondary Blockchain which is connected to the main Blockchain through a “peg” but can have its own consensus protocols, including protocols which are completely different from the main chain's protocol (Amritraj Singh, 2020) (Adam Back, 2014). These chains can be assigned to handle sub-tasks adjacent to the main chain and can be used to improve efficiency; for example a chain to handle high speed transactions and a separate chain to handle complex computations (Lee, 2018). One of the important uses of side chains is Interoperability.

Interoperability

Interoperability is an emerging topic in the world of DLT and has arisen to avoid the siloing of different Blockchains. In most modern computer systems interoperability is an essential element – that is to say that each system can interact with other systems and is not cut-off. Unfortunately most Blockchains operate completely separately from other chains and are therefore siloed, forcing users to choose one technology over the other.

This ability to communicate between chains allows much more functionality and allows actions on one chain to influence actions on another, therefore users are not forced to choose one technology over another. Blockchains can in effect become side chains of other Blockchains. For example, a user can deposit money into a Bitcoin wallet which would then send proof to Ethereum with an order to purchase CryptoKitties[6].

Bloxroute (Bloxroute.com) is an example of a start-up implementing one method of interoperability at the network level.

Directed Acyclic Graphs

Hardware improvements like disk input/output, RAM and bandwidth, can improve on Blockchain issues such as processing speed, however they can actually increase the issue of node syncing – in more concrete terms the problem of orphan nodes. An orphan node occurs when a block or chain is created outside the main chain due to unavoidable network delays. That is, the faster the network the more likely it is that once a block has been created and propagated through the network, newer blocks will have been added to the main chain which do not reference it. When more “honest blocks” get left outside the main chain, the less secure the chain is (Yonatan Sompolinsky, 2015) and (The Bitcoin Backbone Protocol: Analysis and Applications, 2015) discuss this in more detail.

Directed Acyclic Graphs are not new and defined as graphs which have not cycles and directed edges – that is no vertex has a path back to itself. Unlike the Bitcoin Blockchain in which each new block points to the last block of the longest chain, new blocks in a DAG point to all of the last block of the graph – known as the graph tips. As a DAG has no cycles, each node has a history in that node Z can be proven to have been created by node Y and so on. To this end, an ordering protocol is already

[6] CryptoKittes was the first mainstream adoption of the Ethereumnetwork. Created by Canadian studio Axiom Zen it is a game which allows players to create, breed and sell virtual cats – with some cats even selling for $300,000. At one point CryptoKitties accounted for 25% of Ethereumtraffic (Cryptokitties.co).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

established, with the exception of blocks created in parallel. Protocols have been created to deal with these (see SPECTRE (Yonatan Sompolinsky, 2018) and (Yonatan Sompolinsky, 2020)) and thus greatly improve speed and scalability.

Aleph Zero (Alephzero.org) is an example of a DAG based implementation (DAGsim: Simulation of DAG-based distributed ledger protocols, 2018) and boasts a purported 100,000 TPS.

Dispute Resolution / Forking

Given the decentralised nature of DLT, decisions are mostly made by the community and not by a centralized organisation. This means that when disputes arise, for a number of reasons, they can lead to either a split in the community or a split in the Blockchain itself. One of the most famous examples was that of the Bitcoin and Bitcoin Cash split in 2017 (Coindesk.com) where the community disagreed on the size of a block. One part of the community wanted to raise the size of a block from 1MB to 8MB whilst a significant part of the community wanted it to remain at 1MB. This resulted in a Forking of the Bitcoin chain and the creation of a separate chain, the aforementioned Bitcoin Cash. Another famous Forking occurred after an attack on the Ethereum chain which syphoned $50M (Waters, 2016) out of the Decentralized Autonomous Organization (DAO). After this Ethereum created a smart contract which would Hard Fork the chain and re-allocate the funds tied to the DAO so that the original owners could withdraw their funds (FRANKENFIELD, 2021). Not all of the members of the Ethereum community agreed with this change and as a result, Ethereum was hard forked into Ether (ETH) with the original chain remaining as Ethereum Classic (ETC). Forks can be classified into two key types:

1. Hard Fork – where the original chain splits and the new chain is not backwards compatible. That is to say, new blocks cannot validate older blocks. This is usually due to a change in technology (like the block size in Bitcoin Cash) which means the new chain is not compatible with the older chain.
2. Soft Fork – this is a split in the chain which creates two chains, however they are backwards compatible.

Costs

One of the problems holding back wide scale adoption of DLT is that of costs; in terms of energy, resource and transactions.

In terms of resource, this can be thought of as the implementation and maintenance costs or running a distributed ledger. This includes design, development and deployment costs as well as the cost of migrating data onto the Blockchain. After this there will be ongoing maintenance costs and costs for third-party tools (Hosting, Storage, Notification Systems, bug-tracking, uptime monitoring etc. etc).

Another cost is the transaction cost. In order to complete a transaction on the Blockchain the user has to pay a transaction fee. In terms of Bitcoin this fee is for the processing of a payment. In Ethereumthis fee could be used to deploy a smart contract. Although the fee may be fixed its value will fluctuate depending on numerous factors, transaction processing time, network congestion etc.

Yet another consideration of DLT is the environmental impact it can have due to the fact that it is very energy intensive. The Energy Institute at the University of Michigan reported that Bitcoin miners consumed 17 million kWh in one day in 2016 (Hughes-Cromwick) and digiconomist.com reports that

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Bitcoin alone will use a predicted 77.8 TWh of energy in 2021. It also reports that the annualised total footprints equivalencies are the following:

1. The BC Carbon Footprint is 36.95 Mt CO2, Comparable to the carbon footprint of New Zealand,
2. The electrical energy consumed by BC will be Electrical Energy 77.78 TWh, Comparable to the power consumption of Chile, and
3. The electronic waste generated will be 10.98 Kt, Comparable to the e-waste generation of Luxembourg.

Whilst Blockchain is an intriguing technology, many sources have reported that it is too slow and too expensive for inclusion in their business model. A trial project using Blockchain to transfer and settle securities launched in 2016 by Bundesbank, Germany's central bank, and Deutsche Boerse concluded in 2019 (Look, 2016) by claiming that it was slower and more costly than alternative methods: "the prototype "in principle fulfilled all basic regulatory features for financial transactions." Yet while advocates of distributed ledger technology say it has the potential to be cheaper and faster than current settlement mechanisms, Jens Weidmann said the Bundesbank project did not bear those out" (Kaminska, 2019). In another example, Daniel Springer, CEO of DocuSign claimed that Blockchain will still to expensive for their purposes. The company started a trial with Ethereumand found out that transactions cost about $1, whereas the traditional cost of digitally signing documents with DocuSign is around $0.07 (Detrixhe, 2020).

As well as the technical considerations of whether Blockchain is the right technology for any project, and if so, which type of Blockchain would be suitable, consideration must also be given to the cost factors behind implementing DLT in terms of both expense and resource as well as environmental responsibility.

6.1.3 Smart Contracts

A smart contract is in essence, a program that is set to run when a given set of circumstances occur. Nick Szabo, the inventor of a virtual currency called "Bit Gold", defined smart contracts as "computerized transaction protocols that execute terms of a contract" (FRANKENFIELD, 2019). These protocols, or lines of code, are stored on the Blockchain and will execute once predetermined conditions are met (Nigel Gopie, 2018). The goals of smart contracts are to speed up processes and eliminate certain costs. Smart contracts can facilitate, enforce or verify a given transaction automatically without the extra cost and delays of involving third parties. This could mean verifying the identity of a purchaser or their credit worthiness or even arranging insurance for their purchase without involving middlemen and a lengthy settlement process.

Smart contracts are digital and therefore require no paper validation process – making them faster than traditional methods. They can be trusted as they have the inherent security of the Blockchain and they can be cheaper to execute as they involve fewer, or no third parties.
The Accord Project is an example of an open-source project to provide tools for organisations to create legal smart contracts: The Accord Project is an open ecosystem enabling anyone to build smart agreements and documents on a technology neutral platform (AccordProject.org).

Smart Contracts have many use cases in automatically and securely executing contracts in areas such as Insurance, Supply Chains and Logistics, Healthcare, Real Estate and Law with many more use cases being created.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

6.1.4 DAO

DAO stands for Decentralised Autonomous Organisation (Decentralized Autonomous Organizations: Concept, Model, and Applications, Oct. 2019) and is an organisation with no central management and therefore now standard top-down managerial hierarchy. The DAO is governed by a self-enforcing code, the governance rules, act to incentivize and direct a network of participants without centralized intermediaries. The bitcoin network can be considered a DAO as no central organisation controls it and decisions are made by a consensus of over 51% of the network.

DAOs are designed to be transparent by being open source and are incorruptible, at least in theory.

6.1.5 DApps

A decentralized application (dapp) is defined as an application that is built on a decentralized network which combines a smart contract and a user interface, i.e. a frontend, in contrast to a standard App which has backend code running on centralised servers.

The concept of DApps may still be in its infancy and a fast and therefore definitions might vary, DApps do however share certain characteristics: They are decentralised, open source and are incentivised for users. Although an argument can be made that Bitcoin was the first DApp, Ethereum really pioneered the concept by allowing developers from all over the world to code and launch DApps on the Ethereum Platform. One key feature of DApps is that one deployed to Ethereum, the Dapp code can't be taken down. Ethereum defines a DApp as follows:

1. Decentralized that makes them independent, and no one controls them.
2. Deterministic i.e., they perform the same function irrespective of the environment they are executed.
3. Turing complete, which means given the required resources, the Dapp can perform any action.
4. Isolated, which means they are executed in a virtual environment known as Ethereum Virtual Machine so that if the smart contract happens to have a bug, it won't hamper the normal functioning of the blockchain network. (Martin Huschenbett, 2021)

For examples of available DApps, an extended list can be found at https://dappradar.com/

6.2 Types of DLT

Public

Public Blockchains are defined as Open Source and un-permissioned, meaning anyone can participate – they do not have to be chosen, elected or pay a stake to participate. They are based on the PoW consensus mechanism and are widely used. They are distributed, have a high level of security and users can be totally anonymous when using one.

The source code of a public Blockchain can be downloaded by anyone and their machine can then become a Public Node, meaning they can then verify transactions and participate in the consensus

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

mechanism to determine which new blocks are added to the chain. The Blockchain is open to anyone who wants to send transactions and once validated, those transactions are encoded in the public chain. Transactions are also open and transparent and can be read in the public book explorer.

Bitcoin.com, Dash.org, Litecoin.org and Etherium.org are examples of Public Blockchains.

Private

Private networks, as their name suggests, are not open for all to participate but are in fact closed and only open to specific participants. Private Blockchains are not open and transparent and are therefore welcomed by private companies or organisation where data security and privacy issues are paramount. Private Blockchains are governed by the company or organisation who set them up and can therefore be regulated at that organisation sees fit.

Another appealing characteristic of private Blockchains is that transactions feed can remain fixed. As only a limited number of actors can request transitions the demand remains under control and the price per transaction can also be controlled – and not spiral up or down as with public Blockchains.

Because private BCs are often used within companies, they can sacrifice decentralization to gain more speed, meaning they are therefore more scalable. Private Blockchains find their uses in monitoring and executing transactions, vote counting and supply chain management.

Corda.net and Hyperledger.org are two examples of private Blockchains currently in development.

Federated

A third type of Blockchain is the Federated (or Consortium) Blockchain. As with Private Blockchains, Federated Blockchains are not open to the general public and only allow certain participants to take part in the network. They are operated under the leadership of a group or consortium and operate under group rules, for example, in order to achieve consensus, a predetermined number of consortium members must be in agreement.

BankChain is an early example of a Federated Blockchain: formed in 2017 it now has 37 members and 9 live projects. A full list of members can be found at bankchaintech.com.

Another example from the InsureTech sector is B3i Reinsurance, built on the B3i Fluidity platform. They have developed a product called Cat XoL which supports the electronic placement and administration of Property Catastrophe Excess of Loss (Cat XoL) treaties. It was designed to structuring the submission, to handle negotiations, binding, and endorsing as well as technical accounting. More information can be found at http://B3i.tech

6.2.1 Advantages / Disadvantages

There are distinct advantages and disadvantages to using DLT or Blockchain. As mentioned previously, there is no one-size-fits-all Blockchain solution and each individual situation will have to be

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

analysed in order to make the right decision as to whether DLT is the right technology to implement and if so, which type of DLT is correct for those circumstances.

The main advantage to DLT is that fact that it is decentralised, meaning that the middle man, or centralised control can be taken out of the loop. Other advantages to DLT include data immutability, transparency and security.

Disadvantages include complex integration, a required cultural shift in terms of usage, transactions may be slower or more costly than other technologies and it can consume a lot of energy to run a Blockchain. Forking can be a disadvantage when it comes to applications such as finance and insurance as high reliability is required and Forks can diminish reliability and confidence in networks.

A 2018 study on The Advantages and Disadvantages of Blockchain Technology (The Advantages and Disadvantages of the Blockchain Technology, 2018) concluded by admitting that the challenges facing DLT are large but that the advantages currently outweigh the disadvantages.

6.2.2 Use Case Scenarios

This section deals not with specific use cases but rather the generic situations where DLT can sensibly be applied. The following are clear use case scenarios where it makes perfect sense to implement one or other form of DLT:

1. The first instance is where there is an intermediary/middleman or entity that cannot be trusted, therefore Blockchain is used as a shared, secure database.
2. Another case for implementing DLT is when a Blockchain infrastructure has already been implemented and different layers can be built to optimize the current infrastructure
3. Making digital signatures secure and trustable is yet another use case scenario for Blockchain, for example, when documents are moving from one user to another or when they need a trustworthy source of truth (creating a digital twin)
4. A final use case scenario for DLT is that of information curation and governance through tokens. If there are sets of information and different agents looking at this information then it is possible to implement a token system that allows the agents to participate in the voting and curation of information. This is also called DAO, mentioned previously in this chapter.

6.3 DLT Providers

6.3.1 Ethereum

Ethereum was launched in 2015, and is an open-source, blockchain-based, decentralized software platform and programming language. It is used for its own cryptocurrency, called ether and can allows for SmartContracts and Distributed Applications (ĐApps) whilst eliminating downtime, fraud, control, or third party interference.

Ether has two main purposes: it is used as a cryptocurrency and is also used to run applications or to monetize work. According to its creators, Ethereum can be used to trade just about anything has been

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

offered under the Ethereum-Blockchain-as-a-Service (EBaaS) model since 2015 through a collaboration with ConsenSys and Microsoft Azure (Gray, 2019).

Ethereum is one of the largest decentralised cryptocurrencies in the world, second only to Bitcoin, meaning that no central organisation has control over the network. It has built in interoperability meaning that any apps created on Ethereum can connect with numerous other protocols. A number of financial product have been rapidly created and deployed on Ethereum owing to the fact that they can connect to existing infrastructure rather than staring a new protocol from scratch.

Ethereum is not very fast and this lack of speed calls into question its scalability. There are also questions over its immutability after transactions were able to be rolled back following a cyber-attack in 2016. Although this was for the good of the network, it has raised concerns that this could happen again, however unlikely that would be.

In Ethereum, Ether is used to pay miners for computing a transaction. Unlike bitcoin, where there will only ever be a fixed amount of bitcoin issued, Ethereum Gas, Ether or often ETH will only be valuable as long as there are transactions to process.

Given its problematic history with scalability, Ethereum is currently researching and implementing the first stages of scalability solutions. Examples of which are Casper[7] (Ethereum, 2019), Plasma[8] (Joseph Poon, 2017), and their state sharding plans (Njui, 2018) which are all set to be implemented over the course of the next few years.

6.3.2 EOS IO

EOS was built for speed and real world tests show that it can compete 4000 TPS (AlphaZero.org, 2021), a number that is expected to increase with future development. This scalability is attracting developers to EOS from Ethereum to create Dapps and Smart Contracts, which can be created in WebAssembly languages such as like C++, Java and Python. Although Ethereum is implementing changes to achieve faster speeds, it is not expected to achieve these speeds until the end of 2021 at the earliest (Klemens, 2021).

Unlike Ethereum, EOS has no transaction fees and user instead pay for the necessary computing resource – e.g. Computation Bandwidth and State storage. This makes it more user friendly than Ethereum Gas, whose functionality is often not well understood by Ethereum users, one example being the user who paid $2.6M in fees to a mining group in order to move a sum of $130 (Phillips, 2020)

EOS has a total of 21 block producers who use the Delegated Proof of Stake (DPoS) mechanism to add now blocks to the chain. EOS users can delegate, Stake, some of their tokens to a producer and receive a portion of the block reward in return. Current Staking Rewards can be found at https://www.stakingrewards.com/earn/eos

[7] the implementation that will eventually convert Etherium into a Proof of Stake (PoS) blockchain known as Ethereum 2.0

[8] A framework that allows the creation of 'child' blockchains which use the main Ethereum chain as a trust and arbitration layer

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

EOS token can also be used to vote on governance issues like the EOS constitution and can use this vote to freeze hacker's accounts if it is proven they are misusing the network. Although this governance model is quite unique, Binance Research claimed that often power shits to the largest holders of EOS. The article also examines the question as to whether EOS is too centralised.

More information can be found at https://research.binance.com/en/analysis/eos-governance

6.3.3 IOTA Tangle

IOTA uses a DAG consensus mechanism called the Tangle. Each transation in the Tangle is connected to two previous connections thought the hashes in its branch and truck transactions fields, making the Tangle immutable. Nodes in the Tangle are referenced directly by parent node and indirectly by grandparent nodes, forming a chain of reference – the history of the transaction. A transaction can be valid only if it references two other transaction's whose history does not conflict with it. Like and DAG, the Tangle can grow in many directions and transactions can be attached to any part of this graph, meaning that some transactions may lead to inconsistencies. For this reason, all transactions start in a pending state until consensus can be reached as to its validity.

The Coordinator is an application that's run by the IOTA Foundation and whose purpose is protect the Tangle from attacks such as parasite chains. Currently, nodes consider a transaction confirmed if it is directly or indirectly referenced by a transaction that's created and sent by the Coordinator[9] (milestone) (Iota.org).

IOTA is a peer-to-peer distributed network and there are no fees for transactions. For each validated transaction which is added to the tangle, two previous transactions are validated. This makes ITOA extremely scalable in that new transactions lead to faster validations.

Use cases include bIOTAsphere (https://biotasphere.com/) which can be used to buy dynamic car insurance for connected cars.

6.3.4 StreamR

Streamr is a very interesting project in the DLT space and takes the format of an open source, crowdfunded, decentralized platform for real-time data (Streamr.network). It allows data producers to own and trade their data in a P2P real-time manner. Streamr's key features include publishing data feeds to subscribers, allowing subscribers to subscribe to data feeds in order to process them as they occur as well as storing data for later use.

Central to Streamr is the Streamr Network, which is used to transport streams of messages from data publishers to subscribers. These messages flow in real time and are hosted on a distributed network

[9] The Coordinator is an application that's run by the IOTA Foundation and whose purpose is protect the Tangle from attacks such as parasite chains

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

of computers around the world. The Network is a scalable, low-latency transport layer and has been created with Dapps and smart devices in mind - including IoT sensors, smart wearables or connected cars – in order to steam data to the Marketplace. DATAs token are earned by participating nodes in exchange for the bandwidth and validation they provide.

Data owners can trade their data streams in the Marketplace - a web-based application designed for trading real-time data streams - using a publisher/subscriber model where consumers pay for access to this data. Ethereum smart contracts are used to handle terms of use, price schedules as well as time-based access control.

The Data Union is an interesting addition to Streamr currently under construction. In essence, it allows for data providers to group together and aggregate their data streams, making them much more interesting for buyers looking to perform analysis and gain insights. One of the first projects comes from Turkey and allows users to monetize their browser data, as shown in Figure. More info can be found here: https://swashapp.io/.

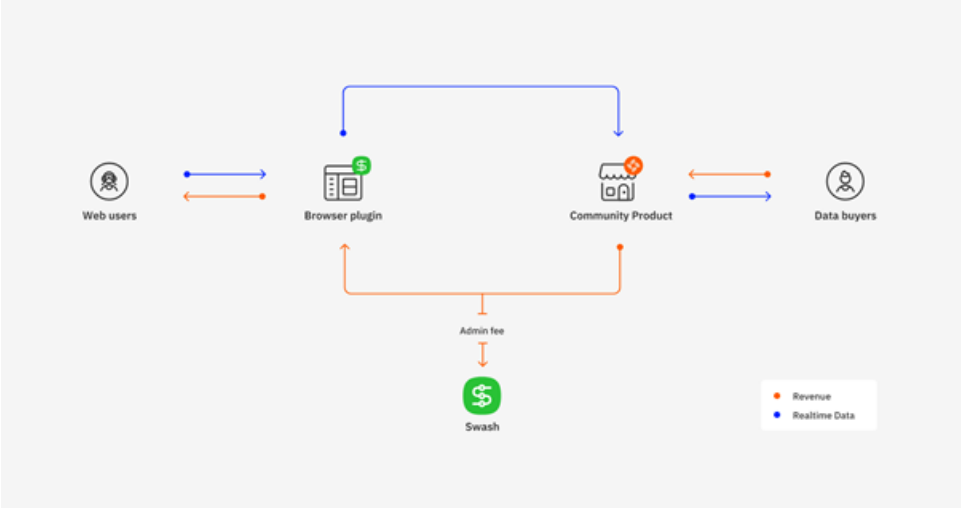


Figure 21: Simple diagram showing how data and Revenue flow with the Swash app.

.

The Streamr website has more case studies of this concept inactions including Tracy – a system used to digitalise Philippine fishermen. It works by adding catch and trade data into a Streamr Marketplace Data Union, where this data will be made available for sale. Banking partner UnionBank will use this data to assess individual fishermen for off-season loans, which are made using their own Peso stablecoin. UnionBank will also provide Know Your Client (KYC) support for the app, and a digital wallet to receive loans and Marketplace revenue (Streamr.network, 2020). The flow of data and payments is shown in Figure below.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

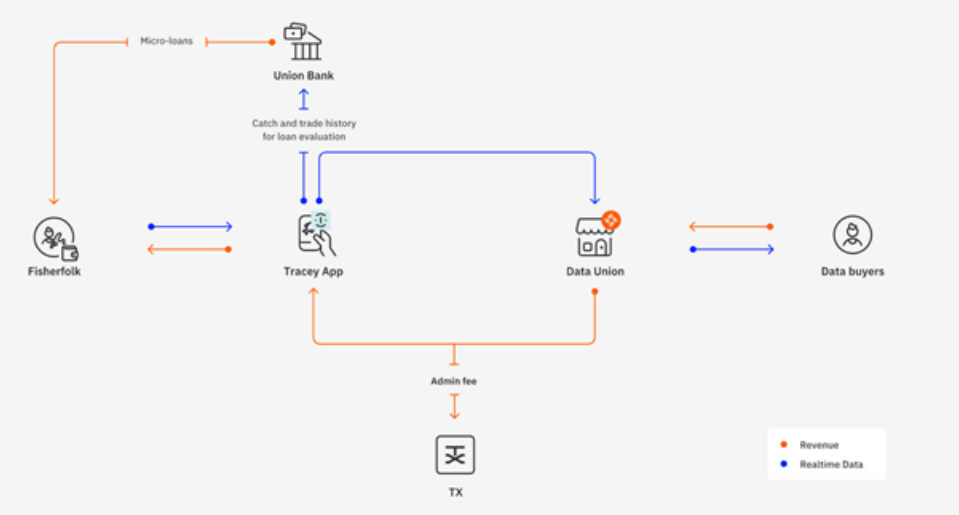


Figure 22: Data and revenue flows for the Tracey project, Source Streamr.network

6.4 Blockchain 3.0: ML + DLT

It is commonly accepted that Bitcoin can be considered Blockchain 1.0 and Ethereum Blockchain 2.0., as has been covered previously in this chapter, and clearly both systems have their pros and cons leading to the fact that neither have made it to wide scale adoption yet. Whilst there is no concrete definition of what Blockchain 3.0 is, it is generally agreed that it is defined by an attempt to overcome the setbacks of Blockchain 1.0 and 2.0 and to give DLT the final push it needs to become mainstream. In particular, Blockchain 3.0 aims to resolve problems such as **Scalability, Interoperability, Governance** and **Privacy**.

As these topics have been covered in their respective sections previously in this chapter, what follows will be an investigation of how DLT technology can be advanced through the use of Artificial Intelligence.

Firstly, it is worth noting that at the time of writing this document, the blending of AI and DLT is nascent and whilst there are a few high-profile interesting using cases emerging, from a business point of view none of these have reached maturity yet. To this end they serve to offer a view of what will be happing in the coming years once these advances reach a higher tech readiness level.

Although the field of AI has made many vast leaps in the last few years, much of it remains inaccessible to those without access to powerful computing resource or extensive training data, which generally takes the format of propriety and centralised databases. Microsoft is attempting to “Democratise AI” by launching a project named Decentralized & Collaborative AI on Blockchain (Decentralized & Collaborative AI on Blockchain, 2019). The code for the open-source Ethereum implementation can be found on GitHub using the following link (Microsoft, 2020).
The goal of the project is to allow users to run ML models with their existing technology, for example on web browsers or on apps on their mobile phones. Microsoft states that this will be ideal for testing AI in scenarios people encounter daily, such as using personal assistants, playing games, or using recommender systems. The fact that the code is run by Smart Contracts means that the users can have a high level of certainty that the code is genuine and that it has not be altered to malicious intent.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

At present it costs a few dollars to deploy an AI solution on the Blockchain and it is free to execute. There is a transaction for updating the AI, significantly less that the cost of deploying, but Microsoft has plans to make this free of charge – through reimbursement or 3rd party fees (Harris, 2019). Figure 19 shows this model in action.

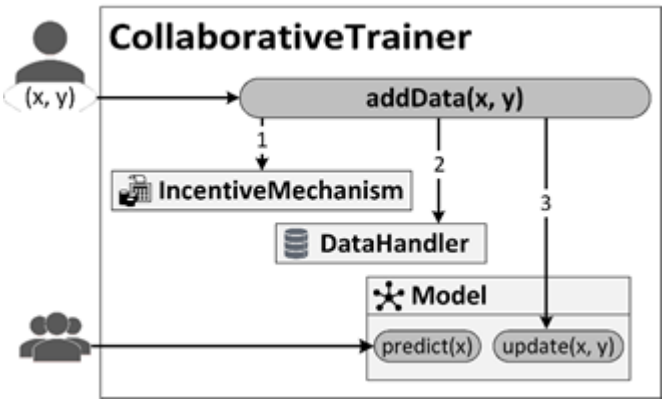


Figure 23: Adding data to a model in the Decentralized & Collaborative AI on Blockchain framework consists of three steps: (1) The incentive mechanism, designed to encourage the contribution of "good" data, validates the transaction, for instance, requiring a "stake" or monetary deposit. (2) The data handler stores data and metadata onto the blockchain. (3) The machine learning model is updated. (Harris, 2019)

Another area of development where AI and DLT converge is in the financial sector, and in particular native digital assets traded on chain by AI. This concept was first highlighted by Chainhaus CEO Jamiel Sheikh in a presentation he gave at MIT entitled Digital Investment Assets: New frontiers for blockchain & AI (Jamiel Sheikh). In the presentation he argues that there are 4 key eras to blockchain development, namely:

- Era 1: Proof of Concepts
- Era 2: Smart Contracts + Tokenization
- Era 3: Native Digital Assets + ML
- Era 4: AI as economic agents

According to Sheikh, we are currently in Era 2, Smart Contracts + Tokenization, but are on the cusp on entering into Era 3, Native Digital Assets, for which ML will be an essential ingredient. At present, securities, physical assets, cash flow rights, or utilities etc. are all tokenised and the token is a digital representation of this asset. This can then be traded on the chain cheaper and faster than through standard means. AI is use here but only in terms of analysis for pattern matching or prediction.
Sheikh argues that the next stage of development, Era 3, will be to create fully digital assets – that is to say the token no longer represents an asset but is in itself an asset. This also means that the business model moves from using the chain to reduce costs to using these digital native assets to generate revenue. These assets will be highly complex and will be created with a mix of human input and AI, or perhaps even just AI given their complexity.

The move to the fourth and final era would involve fully autonomous AI agents not only creating assets but trading them on the chain. This is a move away from Smart Contracts to Smart Agents and would involve DLT becoming and almost invisible layer in the accepted tech stack in much the same way protocols like HTTP and UDP have. This would mean that DLT would be inherent in most technology

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

from web and mobile to wearables to IoT, and that smart agents could trade using strategies that they create, modify and if necessary, eliminate, with minimal intervention from humans.

6.4.1 Transparency

One of the problems with AI is that when networks get very deep, there can be a lack of transparency as to what steps the AI is taking (Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy, 2020). In a lot of situations this is not important but in highly regulated verticals such as medicine, Finance, and Insurance, transparency is vital to widespread AI adoption. This has led to a separate category of AI research called Explainable AI, or XAI (Wojciech Samek, 2017). AI can also take massive amounts of data, analyse it and produce results – for example, Walmart's AI systems use an entire month's worth of transaction data for each of its stores to create recommendations as to where products should be positioned and what products should be stocked (Marr, 2017). The lack of transparency in these processes and the size of the data being used makes auditing such decisions very difficult. DLT can be used in this situation to record data points and decisions so that a virtual "paper-trail" is created. This can then be used to audit the decisions being made, for educational purposes or for purposes such as creating more trust in AI systems among the general public.

Numerai (Numerai, 2017) is an interesting project in this area where a hedge fund has opened up its models to the public in order to be crowdsourced. Data Scientists around the world can used the Blockchain to contribute to algorithms that Numeri will eventually employ in managing their hedge fund. IN return, successful contributors will get rewards in Ether or their own NMR.

6.4.2 Interesting Use Cases

AI as a service seems to be a key area for growth in combination with DLT and there are various companies working in this area. One example is **SingualrityNET** (SingularityNET, 2019), an AI marketplace founded in Amsterdam in 2017 with the goal of providing buyers and sellers access to AI software and hardware service APIs that can be integrated into smart contract templates. SingularityNET see that their network will be filled with AI agents providing "core AI algorithmic services" as well as customer facing "high-level AI services". The later could be examples of AIaaS and would be used to offer AI as Language Processing Services, Image/Video/Audio Processing Services or even to offer access to curated databases for training new AIs.

DeepBrain Chain (DeepBrainChain.org) is a not-for-profit based in Singapore have developed what they call "decentralized, low-cost and privacy-protecting AI computing platform with a full range of related products and services". They have created DeepBrain Chain to be a "decentralized neural network" in which nodes from across their network supply AI companies with computing power and in return receive tokens called DBC. The use DPoS and POI consensus mechanisms and aim to bring value to using data. More specific details can be found in their white paper (DeepBrain Chain).

Innoplexus, a German start-up are bringing Blockchain and AI to the field of Drug discovery to help with data security and discovery. They claim that in the research field, many insights are discovered but for a number of reasons are never published and are therefore not reachable by other actors who

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

could make good use of them. These reasons are varied but could include factors such as the experiment failed to prove the experimenter's hypothesis and as such the research was abandoned before being published, the fact that the organisation changed research direction before, during or after the results were discovered or even the fact that the experiment was carried out in an academic setting and the results were simply never published outside of academia. Innoplexus claim that Blockchain can allow them to securely publish this data and because it is on the chain, it is simple to attribute ownership and licensing rights to the original authors. AI can be used in real time to evaluate the artefacts and establish duplicates of similar documents. The platform can also be used to make available other data considered enterprise specific or restricted data and make it discoverable on a company's platform whilst ensuring it remains private. Figure 20 shows an overview of the architecture used by Innoplexus to integrate AI and Blockchain. More info can be found in their white paper (Innoplexus).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

7. Intelligent Digital Twins

7.1 Intro to DT State of the Art

Digital Twin, the term designated to a ubiquitous strategy, owing to the vast array of its applications in various fields, essentially means that it is a virtual representation employed to understand and foretell the behaviour of its physical equivalent. In simpler terms, Digital Twin is a digital replica of a physical product in all its aspects. The latter half of the sentence implies that the digital twin observes the entire life cycle of its physical counterpart through sensors and stores it in the form of data that is thereafter used to simulate the behaviour and even optimize some actions in the behavior if required. (Khasnis, et. all 2020)

The concept of Digital Twin (DT) has emerged together with Industry 4.0. Essentially, “the digital twin is a virtual and computerized counterpart of a physical system that can be used to simulate it for various purposes, exploiting a real-time synchronization of the sensed data coming from the physical system”. (Negri et al. 2017). Another definition of DT by NASA highlights the integration of the best available multi-physics and multi-scale simulation models and sensor data (Shafto et al. 2010). From the literature to date, as stated in (Lim, K.Y.H. et all 2020), DT-related enabling techniques have experienced exponential growth over time and its core idea has been transformed into distinctive concepts. Digital twin technologies indicate an exponential increase within the other technology areas. Sensors become cheaper to procure and communication technology advances on devices. Data computation methods including ML&Analytics enable computational processing effort. Data management and connectivity platforms are being developed and intended to be used in the DT technologies. AR/VR concepts provide extension with the virtualization and simulation services.
Gartner predicts that by 2021 half of all large industrial companies will use digital twins. This is in large part because digital twins are now driving new business models and evolving industrial automation. Despite this, many businesses have been slow to embrace digital twin technology because it is complicated to implement. Once implemented, however, digital twin technology can add value to traditional analytical approaches by improving situational awareness. (https://www.futurithmic.com/2020/04/14/how-digital-twins-driving-future-of-engineering/)

DT can be classified in terms of real-time integration and the intended utilization in product/process lifecycle; Real-time feedback between the model and the physical system establishes a DT, whereas one-way data exchange from real system to model is categorized as a digital shadow and the lack of real-time integration is understood as digital model (Kritzinger et al. 2018). The classification based on the utilization comprises design twins, product or production twins and performance twins. “The digital product twin includes all design artefacts of a product, the digital production twin covers models for the manufacturing process and production system itself and a digital performance twin derives insight from utilization data and analyzes actual performance” (Boschert et al. 2018). In OXILATE, the focus is on performance and product twins with real-time capabilities.

It is foreseen that DT is a key enabling technology in centralized analysis and control of the manufacturing processes (Uhlemann et al. 2017), integrated data analysis, prediction and data visualization (He et al. 2018), among the others. DTs can support in process automation, complex control automation, transferring of product components and information, intelligent prediction and understanding the value chain. In addition, DTs can help in mobilizing the information. The state of the art of these topics are described in the following subsections.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

7.2 DT for Process Automation

Process automation refers to the use of digital technology to perform a process or processes in order to accomplish a workflow or function. Process is composed of process owners, process assets and process tasks performed in a timeline upon completion. It is a digital transformation of an existing process already defined and applied in any industry area.
Within the promising increase in the digital twin technologies development, DT provides capabilities to support the Product Lifecycle Management in industrial IoT world. The most promising capabilities are 3D Representation, Visualization, Data Model, Simulation, and Document Management regarding the Product Lifecycle Management and connected Analytics. They foster the evolution of four classes of digital twins hierarchy. Parts twinning, asset twinning, system twinning and process twinning. (Revathi et. all, 2020)

Process and asset twining is the classes used to enhance process automation development, process and asset twinning. Process twinning is the high level twinning that represents a whole production process and provides insight into the collaboration of all other system. In other words, process twin can be expressed as a set of asset or system twins but they mainly focus on process rather than the equipment. They become more functional and effective when all the system, assets and parts fulfill their requirements which make the level of analyzing the entire process via digital twins much easier. (Revathi et. all, 2020)

Assets are single entities that is managed by processes. During process automations, assets related with the tasks are widely automated and their lifecycle from creation to end is managed.

Digital Twin for Process Automation use cases and applications are quite wide. Automotive Industry, Education, Healthcare, any kinds of Product Industries, Smart Cities, etc. can be listed.

7.3 Transfer and control of product components and information

Turk Traktor is planning to implement new sensors and use existing ones in Automotive Paint Shop area in one of its manufacturing plants to provide near real-time data collection from different manufacturing robots and ovens. There are multiple parameters from these machines to be collected through implemented SCADA systems and directly from sensors to a central server and database. This central database architecture will be constructed with a joint work of Turkgen, Semantik and IND Bilisim. After data collection and preparation for analysis, there will be software led mechanisms to understand data as planned with machine learning algorithms. The challenge will be defining and manipulating the data collected from many parameters. Semantik will employ transformer models for natural language processing from/to messaging commands of users with the machines. IND Bilisim will use data to create and run Digital Twin of Paint Shop process for Turk Traktor.

Transfer and control of components and information can be taken in consideration of Communication and Computation stack layers within the Technology Stack for DT creation. (Lim, K.Y.H. et Al. 2020)

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Data acquisition and transmission are crucial in DT for real-time information flow and connectivity. This section emphasizes on key network architectures, data exchange protocols, as well as middleware platforms used in studies to facilitate information exchange and streaming processing. Network architecture involves integration of protocols and layered network interface through function blocks.
Table 2 highlights the prominent architectures discussed such as multi-tier architecture and others. The OSI model, consisting of 7 layers (physical, data-link, network, transport, session, presentation, application), then establishes the concept of layered network architecture with the use of abstraction layers. These communication protocols are crucial rule sets for machine-to-machine connectivity and information transfer between communicating entities. (Lim, K.Y.H. et Al. 2020)

Author	Architecture	Description
(Hao Zhang et al. 2017) (Leng et al. 2018) (Q. Liu et al. 2018)	J2EE SSH programming architecture	Platform providing functionality for developing multi-tiered and distributed Web based applications
(Arafsha et al. 2019)	Master Slave	Communication model where a device has unidirectional control
(J. H. Lee et al. 2018)	RESTful	Software architectural style for creating web services
(Park et al. 2019)	Service Oriented Architecture	Software design style where services are provided via application components, through communication protocol
(P. Zheng, Lin, et al. Server-Client Computing model in which server 2018)	Server Client	Computing model in which server 2018) architecture manages resources consumed

Table 2: Network architectures

Table 3 highlights data exchange protocols in manufacturing environments used by data acquisition systems for high level DT communication. (Lim, K.Y.H. et Al. 2020)

OSI Layers	**Authors**	**Rule**	**Description**
Application Presentation Session	(Ardanza et al. 2019) (Neill 2016) (Bao et al. 2018) (C. Liu et al. 2019) (Luo et al. 2018) (Y. Zheng et al. 2018)	OPC UA	Machine-to-machine communication protocol for industrial automation
Application Presentation Session	(Park et al. 2019)	SOAP	Messaging protocol specification for exchanging structured information in the implementation of web services in networks
	(Nikolakis et al. 2019)	NTP	Networking protocol for clock synchronization between systems

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Transport	(C. Liu et al. 2018) TCP/IP Communication protocol suite to (Ardanza et al. 2019)	TCP/IP	Communication Protocols
Data Link	(Moreno et al. 2017)	Ethernet/IP	Industrial network protocol widely used in industries including factory, hybrid and process

Table 3: Data exchange protocols

Table 4 summarizes key middleware platforms to enable seamless connectivity without altering infrastructures, allowing easier DT adoption into the current manufacturing ecosystem.

Author	**Tool**	**Description**
(P. Zheng, Lin, et al.)	Amazon EC2	Cloud-based environment for cloud deployable 2018)
(Lovas et al. 2018) (Damjanovic-Behrendt and Behrendt 2019)	Docker	Saas and Paas products that use operation system level virtualization to develop and deliver software in containers
(Damjanovic-Behrendt Kubernetes Open source container orchestration system for and Behrendt 2019)	Kubernetes	Open source container orchestration system
(J. H. Lee et al. 2018)	Spark	General Purpose distributed data processing

Table 4: Middleware platforms

After enabling the transfer protocols, computational models are employed to provide control for batch-oriented and real-time data processing. Extracting practical knowledge from heterogeneous data is challenging and thus, determining the right methodologies and tools for querying and aggregating sensor data is crucial to DT construction.

Machine learning and data processing tools provide a wide range of solutions ranging from analytics to automation and these provide DT with decision-aiding capabilities via enabling tools such as computer vision.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Table 5 summarizes the computational processing tools used in this review. In Table 9, machine learning and analytics methodologies, including statistical kits for optimization are presented. (Lim, K.Y.H. et all 2020)

Author	Tool	Description
(Yuqian Lu and Xu 2019)	AWS Elastic MapReduce	AWS tools for big data processing and MapReduce analysis across Hadoop
((Damjanovic-Behrendt and Behrendt 2019)	TensorFlow	Open-source software library for dataflow and differentiable programming for machine learning applications
(J. H. Lee et al. 2018)	Reduce	Extract feature vectors from time-series data

Table 5: Middleware platforms

Within this scope of work, manufacturing process in the Paint Shop will be observed in detail with the support of sensors and collection of more explicit data from manual operations. In order to realize the output, Turkgen will develop a machine learning based software product, IND will develop the Digital Twin with dashboards and simple simulation with change of parameters to analyse potential changes in the manufacturing for taking actions, Semantik will develop a digital assistant to command and control the process.

7.4 DT for Intelligent Prediction

Predictive ability of the soft sensors, data-driven models and digital twins is an utmost important feature in performance-oriented applications. For example, the intelligent diagnostics with prediction abilities can support industrial operators to prevent equipment breakages, minimize failure rates and to avoid unnecessary shutdowns. The predictions can be used to schedule the equipment maintenance or replacement based on need, thus having positive impact on capital and operational costs. Intelligent prediction of process variables can offer significantly improved possibilities for predictive process monitoring and control. Ultimately, these solutions can be linked to the optimization of the production processes, hence reducing the production losses and out of spec products and minimizing the energy and material usage.
In a recent review, (Carvalho et al. 2019) presented how the machine learning has been employed in predictive maintenance applications, where the degradation process of the system is monitored and predicted before the fault or failure occurs. They pointed out studies where intelligent methods have found to be more useful than statistical approaches. According to the survey, the most popular ML methods have been Random Forest, Neural Networks, Support Vector Machines, and k-means clustering. They also pointed out the public data sets available for develop, test, and compare ML methods. Another important review focused on machinery health prognosis has been performed by Lei et al. (2018). Their systematic review comprises the whole chain from data acquisition to prediction. They also mention number of commercial platforms for machinery prognosis.

One of the fundamental black-box approaches to monitor and predict the production processes is relied on multivariate statistics and dimensionality reduction (Nomikos & MacGregor 1994). Data-driven

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

methods for the predictive process performance monitoring have revcently been reviewed e.g. in (Witt et al. 2019) with a focus in distributed batch production. In continuous production processes, the predictive modelling has also been an important tool for a long time (Chen & McAvoy, 1998).
For example, in chemical industries, the efficiencies often are concentrated to reducing energy costs per product. Higher energy prices strongly contribute to increasing operating and manufacturing costs. Additionally, inefficiencies in energy usage also contribute to higher greenhouse gas emission. It has been concluded that the improvements in energy efficiency require pragmatic and holistic approaches (Drumm et al. 2012). The dynamic losses (difference between the current energy consumption and the historical or theoretical energy consumption) can be estimated from the process data and visualized to the plant operators (Drumm et al. 2012). Other similar approaches can be found e.g. from (Prodanuks et al. 2017, Nikula et al. 2016). The commercial solutions for the energy efficiency monitoring, prediction and optimization involve examples such as the Web-based app and Microsoft .NET based tool by ABB (ABB website). For the manufacturing processes, examples of emerging IoT frameworks for energy consumption and prediction are presented in (Qin et al. 2017, Tan et al. 2017).

7.5 DT for Understanding the value chain components

7.5.1 Value chain and Digital Twin:

For any organization to create and deliver valuable products, operations need to be performed where each operation contains specific values to value the product. Therefore, these operations need to be connected in a series of chain to make the product/service to the processes in which operational values are added to create the product/service.

To go beyond the state of the art value chain thinking (e.g. Porter [1]), Oxilate project focuses on supporting systems which are fully integrated in the customer's (often ill-defined) operating environment and workflow and reacting to the (sometimes unknown) customer needs in an agile manner, closing the gap towards a 'DevOps' way of working. The project will employ 'actionable' tools integrating 1) data analytics on operational data and 2) expert knowledge, made 3) resilient to change, so that the value chain operations can be optimally supported by professionals for their respective business activities (green) and preventing overload of experts (red).

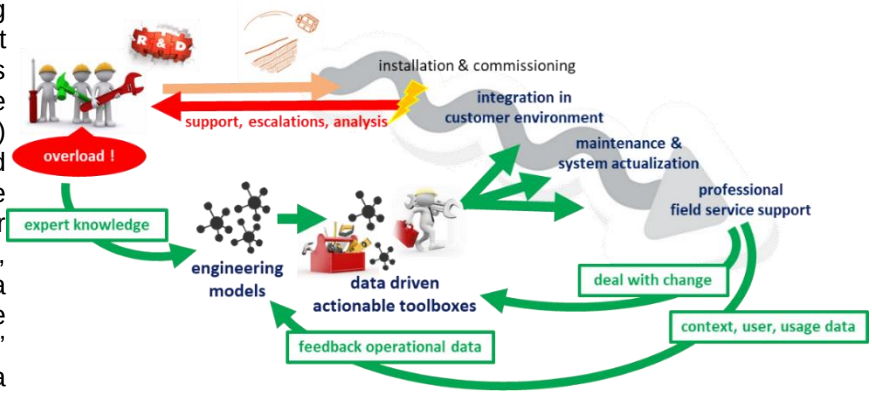


Figure 23: How Oxilate improves value chain operation with Digital Twin

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

7.5.2 Value chain in Industry 4.0, related challenges, and success factors

The decreasing cost of sensors and increasing computational capabilities have enabled three main capabilities: the industrial Internet of Things (IIoT), artificial intelligence (AI), and machine learning (ML). These capabilities make it possible and affordable to create an exact digital replica of something in the physical world. As a direct result of the influence of technology, supply chains are complex and dynamic, and it has never been so important to adopt an agile approach. This is where digital twins come in.

From a supply chain perspective, a digital twin is essentially a virtual supply chain replica that consists of hundreds of assets, warehouses, logistics and inventory positions, and it is gaining more attention in the industry due to improvements in technical and computational capabilities with operations technology (BCG, 2020) [2]. Industry 4.0 is gaining momentum in supply chain amid the COVID-19 environment. In manufacturing industry and the introduction of industry 4.0, according to Imran (2020) [3], business assets, process manufacturers are not isolated; they exist within the context of supply chains and their surrounding business environment. When a process manufacturer and supply chain respond in unison to market signals and disturbances, value chain optimization is possible.

Challenges and success factors of Digital Twin in Industry

According to BCG (2020) [2], to create sustainable change, companies should address three enablers from the start:

- Changing processes so that they can effectively use the new insights.
- Building up the required capabilities and changing the way of working.
- Leveraging a data and digital platform in order to release data from core IT systems and quickly build analytics-based minimum viable products.

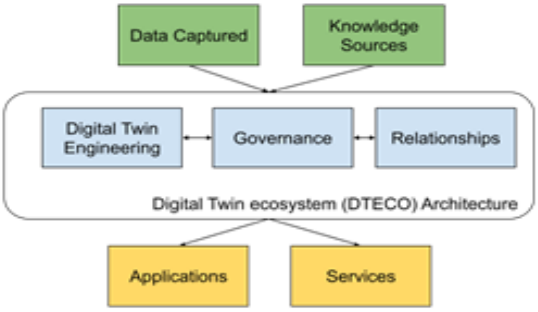
Imran [3] addresses the Critical Success Factors of Digital Twin in supply chain including: IoT Sensors and Data, Simulation Capability, Interoperability, Supply chain practices, and Advanced analytics and visualization

7.5.3 Going beyond value chain: digital twin ecosystem

The data collected from the *IoT Sensors* or real time data from supply chain along with different knowledge sources acquired need to be analyzed to create applications and services for digital twin. This creates an opportunity for platform and application providers to create solutions using the data given by the customers. Thus, in the ecosystem viewpoint, considering the continuous growth of companies' capabilities in analytics, it is important that all stakeholders within the organization and across the industry ecosystem operations are technically and technologically capable of deploying digital twin technology. All stakeholders must be willing to integrate and cooperate to align with the mission. Stakeholders must share the large volumes of data to predict the events and manage disruptions. The DT ecosystem architecture (see Fig 2a) can contain three aspects, the digital twin ecosystem architecture would contain three aspects such as digital twin engineering, governance, and relationship. The digital twin engineering will focus on actual development of digital twin solutions from the capture data. This will include aspects such as software engineering, virtual and augmented reality engineering and so forth. Furthermore, in governance, the business and the management perspective of the ecosystem need to be considered in which it is important to see that the solution data created from the data has a business value along with the management of the overall process in order to maintain smooth working of the ecosystem. Finally, in relationship, it's important to understand what

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

the roles of the different actors are playing in the ecosystems are. For example, one actor could be the customer which is providing the supply chain data. Similarly, platform provider will provide a standard base on which digital twin application could be developed and the page itself is utilized by the application provider. Therefore, it is important to understand the relationships that exist between the different actors working together in the ecosystem to create value to the customer and the user.



	Internal platform	Supply-chain platform	Ecosystem platform
Unit of analysis	Company	Supply chain	Industry ecosystem
Degree of openness	Closed interfaces	Selectively open and closed interfaces	Open interfaces
The level of access	Company capabilities	Supply-chain capabilities	Unlimited number of external sources
Control/governance mechanisms	Managerial hierarchy	Contracts between supply chain companies	Ecosystem governance

Figure 24. (a) Digital twin ecosystem architecture, (b) classifications of the technological platform

Furthermore, Platform's aspect needs to be considered whether internally within the firm or externally with the stakeholders are essential for digital twin technology. Most of them are cloud-based platforms and provide powerful computing infrastructure with analytics capabilities to create a digital twin. These platforms provide capabilities to process effective simulation and scenario evaluation based on real-time data to avoid disruption. Furthermore, building on Gawer's work [4], Xu, Päivärinta & Kuvaja [5] categorizes the platforms into three organizational categories: the internal platform, the supply chain platform and the industry ecosystem platform. The classifications of the technological platform are shown in Fig 2b. The extent of the ecosystem needs to be analyzed in form of three aspect such as internal platform, supply chain platform and ecosystem platform. The three-level hierarchy will enable regarding the business, assess and governance mechanism of the ecosystem as well as type of the digital solution provided to different stakeholders. Moreover, three units of analysis will be focused on the Oxilate project, namely: 1) **DT ecosystem's degree of openness**: The degree of openness is a defining characteristic of an ecosystem; 2) **DT's level of access to capabilities and resources**: The platform's openness does not only affect actors' access to information, naturally, it also influences the platform's access to capabilities and resources as an increasing number of participators means a larger number of source pool and stronger capabilities, such as the open innovation capability in the literature [6]; 3) **DT ecosystem's control/governance mechanisms**: Schulte et al. 2012 [7] suggest virtual factories that combine service-oriented computing and service-oriented workflows with the IoT. In a service-oriented architecture, service providers, service consumers, and virtual factory brokers as essential roles of DT in industry 4.0. The latter role is responsible for managing and controlling the virtual factory and uses services to model manufacturing processes and assemble products based on the results of factories of various business partners. The plug-and-play character of DT and Industry helps companies execute cross-organizational manufacturing processes as if they were running in a single company.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

8. Generative Engineering

8.1 Intro to SotA Generative Engineering (BE)

Generative engineering is a collective name for innovative computational computer-aided engineering (CAE) methods that help supporting engineers in the early steps of (mechatronic) system design. These methods rely on a more abstract description of architecture models, on generic design constraints that translate expert knowledge, and on powerful solvers for creating design variants. The usage of computer-aided generative design tools provides engineers with the possibilities to create much bigger pool of possible solutions, thus the chances for developing the most optimal systems rise. In contrast with traditional engineering process, which require multiple manual calculations, can only deal with simple geometries, and are not prone to elementary errors, software tools are able to cope with complex tasks requiring much lower level of efforts. Currently, we can distinguish between the all-in generative engineering software, that consist of both generative design toolkits, together with design variants solvers, and a CAE software that only proposes one of the mentioned capabilities.

The usage of computer-aided generative engineering approaches for early-phase mechatronic system design requires significant rethinking of traditional engineering processes. These tools are very performant and potentially lead to much more advanced solutions than those obtained by traditional, manual design approaches. Nevertheless, mastering these software tools requires a lot of expertise from a user. This, in turn, implies significant needs for user support and training, before the user becomes efficient and successful in adopting the new generative-engineering paradigm.

Therefore, where the traditional focus in the CAE business is on functionality and performance, the upcoming generation of advanced computer-aided generative tools will require significant advancements in usability and user experience: classical documentation, tutorials, and training modules will not be sufficient to support junior and senior engineers in their daily tasks.

Thus, empowering engineers with intelligent digital assistants, inbuild in computer aided generative software, will take individuals productivity and utilization of this software tools to a next level. This digital assistance will complement traditional user support (e.g., documentation and training modules), which are found to be insufficient in assisting unexperienced users in mastering the powerful but complex new prototype tool.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

9. Mobile / Wearables

It can be argued that wearable technology has been with us since the first eyeglasses in the in the 1200s or the earliest timepieces in the 1500s, however modern wearable technology (more often called Wearables) involves microprocessors and sensors capable of picking up data and transmitting data through the internet or other communication protocol. Modern wearables can be worn (such as smart glasses of smart watches,) or can be embedded into fabric to monitor medical conditions.

Figure 25: Wearables on the body. Source (University of Illinois Chicago, 2019)This chapter will focus on wearables from the top providers as this covers most of the market (Statista, 2021), being Apple, Android, and Samsung.

9.1 Intro to SotA of Frameworks and services (ES)

Android Wear

Development for wearables running on the android platform are developed on Wear OS by Google. It is based the Android SDK and therefore development is similar to other apps in the Android environment however they differ in design and functionality. The official development language for Android is Java however applications can be created with C and C++ using the Android Native Development Kit (NDK). Just like the Android platform, Wear OS contains features such as apps, notifications, and Actions but adds watch faces as well. Wear OS by Google works with phones running Android 6.0+ (excluding Go edition) or iOS 10.0+, however supported features may vary between platforms and countries.

More info can be found at https://developer.android.com/wear

watchOS 7

WatchOS 7, available from the 16 of September 2020, and is the latest operation system with the Apple Wearable watches, at the time of writing this document, and will work with the Apple Watch

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

series 3 to 6. The 7th iteration of their wearable interface includes updates such as watch face sharing, sleep tracking capabilities, handwashing tracking, Siri translation, new Complications[10] and Shortcuts.

watchOS 7 has been problematic and users have reported a series of issues, including forced reboots for Apple Watch 3, increased battery drain on Apple Watch (and iPhone), lost workout data, among other. This has promoted Apple to release a series of updates, the last being version 7.3 in January 2021.

Programming for the Apple Watch is done with Object C as well as SwiftUI for the GUI and Complications.

Samsung Gear

Samsung make one of the most popular smart watches, the Samsung Gear range which was originally based on Google's OS but is now built with Samsung's own OS called Tizen (Tizen.org). Tzen Studio 4.1 was released in December 2020. As with watchOS and Android Wear, Tizen needs to be paired with a compatible smart phone but will work with Apple and Android phones.

9.1.1Data Collection

The Apple watch contains an accelerometer, a gyroscope, a pedometer, an optical heart rate sensor, GPS and force-touch haptic feedback. It has a multi-directional microphone for processing different audio streams at once: The smart assistant Siri uses the microphone to pick up audio cues and can respond through built-in watch speakers. It can send and receive data through 802.11n Wi-Fi, Bluetooth, GPS, and cellular connectivity and can connect with third party apps and other smart devices.

Unlike WatchOS which is only used by Apple devices, Android Wear OS is used by a multitude of devices from many different providers. For this reason, each device will have different sensors and therefore different data collection capabilities. In general Android provides for the collection of data from Motion Sensors, Orientation Sensors and Environmental Condition Sensors.

Supported sensors can include: Axes Sensors to determine position or movement in 3D space as well as "Base Sensors" (Accelerometer, Gyroscope, Barometer and Magnetometer), temperature sensors, lights sensors, heart rate monitors, proximity sensors, pressures sensors and relative humidity sensors. An exhaustive list can be found here: https://source.android.com/devices/sensors/sensor-types

In terms of connectivity, Samsung offers 3G/LTE, Bluetooth 4.2, Wi-Fi b/g/n, NFC and A-GPS/Glonass. It contains an Accelerometer, a Gyroscope, a Barometer, a Heart Rate Monitor and an Ambient Light sensor.

[10] Complications are what apple calls customisations of the Apple Watch face, and offer the user more info about the app, can rapidly launch an app to a specific screen when tapped or perform background tasks to ensure the complication and app are up to date.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Figure 26: Types of data collected by wearables. Source (University of Illinois Chicago, 2019)

9.1.2 Data analysis

An IBM Watson study reported that the average person is likely to generate more than 1 million gigabytes of health-related data (IBM, 2016) in their lifetime. It is the vast quantity of generated data that poses problems to industries such as sports, healthcare and InsureTech: how to leverage a robust big data and analytics model to utilize this data and add real value to the user.

One of the problems with this data is that often first party app providers, such as hospitals, healthcare providers or insurers are not equipped with the right data science teams to analyse and make use of such data and must/should therefore externalize this data as opposed to try to analyse it in-house. This also poses the issue of data security and costs or damages if data breaches occur – with some estimates putting the costs of beaches in the U.S. health care field at $6.2 billion each year (University of Illinois Chicago, 2019). This will require a change in mindset and will require the redesign of research methods and alternatives to clinical trials which would result in the speeding up medical innovations (University of Illinois Chicago, 2019). Another problem is that this data is often not easily integratable into existing systems of workflows. This means that organizations need to build the technological infrastructure necessary to handle this data, not necessarily something that is immediately possible.

Data collected can be processed on-board and shared with connected devices or platforms or streamed directly over WiFi, Bluetooth or mobile connections. Data from users' daily lives can be collected and analysed in order to give an overall picture of their health and wellbeing and can then be used for medical examinations, testing or clinical trials. In the field of insurance, health and life insurance policies could be dynamically adjusted in accordance with the data collected and new risk and pricing models could be created.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

9.2 SotA mobile/wearable Hardware

9.2.1 Devices

Modern wearables such as smart watches are highly functions as well as being stylish, come with a range of apps to run on device or to connect to other smart devices and are more than capable of analysing data using their on-board processers.

The latest Samsung Galaxy smart watch has an Exynos 9110 Dual core processor running at 1.15GHz and 4GB of Internal Memory and wither 1.5GB or 768MB of RAM depending on the model. The Apple Watch contains a S5 SiP with 64-bit dual-core processor and up to 32GB of memory.

Smart Devices

Power

Power is a key concern for smart wearables and as with smart phones, as screens get larger and brighter, and sensors such a GPS, biosensors and heartrate constantly monitoring conditions, the strain on battery life is ever increasing. There is a trade-off between functionally, processing power and battery life. One solution is low voltage circuitry (Ambiq.com, 2021) as well as higher performance but more efficient processors. In parallel, efficiencies in communications protocols such as low energy Bluetooth and ANT+ technology have helped greatly reduce battery drain. Other advances include changing the way sensors perform continuous sampling. Change Point-based Activity Monitoring (CPAM) (Easing Power Consumption of Wearable Activity Monitoring with Change Point Detection., 2020), introduced by Culman et. Al. is an “energy-efficient strategy for recognizing and monitoring a range of simple and complex activities in real time. CPAM employs unsupervised change point detection to detect likely activity transition times”. The authors show that “by adapting the sampling rate at each change point, CPAM reduces energy consumption by 74.64% while retaining the activity recognition performance of continuous sampling”

Range

One consideration of wearable is the range of communication needed. For devices such as Smart Watches which are not mobile enabled, they need to be connected to a smartphone in order to get the required internet connectivity.

Projects such as Helium aim to provide a network of routers connected to private or commercial wifi networks in order to create a “sub-network” for connected devices to communicate on - devices for which it is not feasible to provide mobile connectivity. For example, for a smart dog collar to broadcast the dog’s GPS coordinates to its owner, it must be connected to a network. It would not make sense for the owner to pay for a mobile phone contract just for the dog, but if enough users were connected to the Helium network, then the dog would be able to continuously broadcast its position whilst roaming outside. Users are incentivised to add the Helium router to their home WiFi through the user of Helium’s blockchain token system. More info on the Helium project https://www.helium.com/technology

Connectivity

Modern Wearables have at their disposition a range of connectivity options. The following is a list of the most common:

- Bluetooth – operates on the 2.4GHz band, has a range of 50-150m and can transfer 1Mbps of data

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

- ZigBee – low power low bandwidth protocol operating on the 2.4GHz band, has a range of 10-100m and can transfer 250Kbps
- WiFi operates on the 2.4GHz and 5GHz bands with a range of 50m and can transfer up to a max of 600Mbps
- Ant, Ant+ is a Bluetooth alternative offering lower power consumption than Bluetooth Low Energy (BLE). It operates on the 2.4GHz band and can transfer up to 19.2kbps of data and is often used in communications with sport sensors and smart devices.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

10. Security and Regulations

10.1 InsureTech Regulations

General Data Protection Regulation (GDPR)

Considered the toughest privacy and security law in the world, GDPR [51] was passed by the European Parliament in 2016 and came into effect on the 25th May 2018. Although it was drafted and passed by the European Union (EU), its obligations extend to any company or organization anywhere in the world when they target or collect data related to people in the EU. Violators of the GDPR rules can face tough fines, up to a maximum of €20M or 4% of annual revenues (whichever is greater), and can be order to pay compensation for damages as well.

The GDPR identifies two entities that both have obligations, namely Data Controllers and Data Processors.
Data Controllers determine what data is required, for what purposes and under what conditions it will be collected and processed.
Data Processors are any company processing data on behalf of the data controller.
It is important to note the on entity could be both Data Controller and Data Processor.

GDPR stats that data must be processed according to 7 principles

1. Data processing must be lawful, fair, and transparent to the data subject.
2. Data can only be processes for the legitimate purposes specified explicitly to the data subject when it was collected.
3. Companies should only collect and process as much data as is absolutely necessary for the intended purposes specified when collecting the data.
4. Personal data must be accurate and kept up to date.
5. Personally identifying data can only be stored as long as is necessary to perfom the intended purpose for which it was collected.
6. Integrity and confidentiality must be ensured throughout the entire process (for examples, secure servers, data encryption etc.).
7. Finally, the data controller must be able to demonstrate their company's GDPR compliance.

Although GDPE defines a wealth of legal terms, one criticism often levied is that its terms are somewhat vague leading to confusion as to how to implement its regulations and requirements.

As of 2021, EU businesses have had almost 3 years to become GDPR compliant so there is little confusion around how to implement its requirements. The partners in the OXILATE consortium will also have to adhere to GDPR standards whenever collecting and processing personal data.

Insurance Distribution Directive (IDD)

The goal of Directive (EU) 2016/97 [52], I.e. The Insurance Distribution Directive (IDD), is to harmonise insurance market regulation across the single European market whilst improving consumer protection standards. It is known as a minimum harmonising directive. Meaning that individual member states can build on this regulation to introduce additional provisions, provided that they are consistent with the

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

directive. The IDD regulates how insurance products are designed and distributed in the EU by making sure that distributors take responsibility for consumer outcomes. The IDD defines the information required by customers before signing a contract in an attempt to ensure that products sold to consumers meet their needs and expectations.

It also imposes conduct of business and transparency rules on distributors, clarifies procedures and rules for cross-border business and lays down rules for the supervision and sanctioning of insurance distributors that don't comply with the IDD. [53]

The rules of the IDD apply to the sale of all insurance products. More prescriptive rules apply to distributors selling insurance products that have an investment element, e.g. unit-linked life insurance contracts.

More information and editable templates can be found at the website of the European Insurance and Occupational Pensions Authority (EIOPA) [53]

Packaged Retail and Insurance-based Investment Products (PRIIPs) Regulation

Any investment product sold to retails which is subject to an investment risk is covered by the Packaged Retail and Insurance-based Investment Products (PRIIPs) [54]. The PRIIPs Regulation has applied since 1 January 2018.
PRIIPs include:

- structured financial products, such as options, which are packaged in insurance policies, securities or banking products;
- financial products whose value is derived from reference values such as shares or exchange rates (derivatives);
- closed-ended and open-ended investment funds;
- investment-type insurance products, such as with-profit and unit-linked life insurance and hybrid products; and
- instruments issued by special purpose vehicles. [55]

PRIIPs Regulation was created to help investors gain a better understanding of and ability to compare key features, risk, rewards and costs of different PRIIPs, thus encouraging efficient EU. They do this through access to short and consumer-friendly Key Information Documents (KIDs).

The PRIIPs Regulatory Technical Standards (RTSs) [55] defines how information in the KID should be calculated and presented.

Regulations in the Insurance sector are many and their content voluminous. They are ultimately aimed at protecting the consumer however they can often have the effect of confusing them through changing regulations, as shown in the timeline for the PRIIPs framework in Figure 16 or create much more work for brokers and insurance providers as shown in Figure 17

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

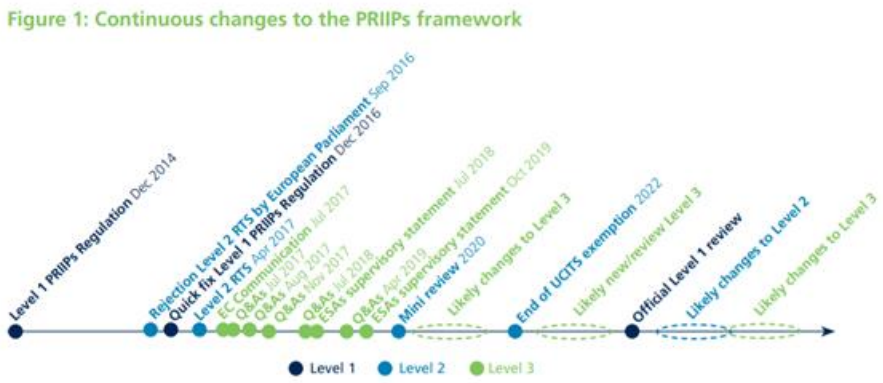


Figure 27: PRIIPs Regulation was intended to make it easier for consumers to compare products and make better-informed decisions, however, the PRIIPs Key Information Document (KID) is difficult to understand and — at times — even misleading. To address the flaws in the KID, the PRIIPs Regulation and its delegated regulations were followed by Commission guidelines, several successive batches of Q&As from the European supervisory authorities and two supervisory statements. Now the delegated regulations are subject to a mini-review ahead of a formal review that could result in further changes to both the Regulation and delegated regulations, most likely necessitating new Level 3 measures. These successive changes to the PRIIPs Regulation and the KID not only result in higher compliance costs for the industry, but also further confuse consumers and reduce their trust in the information they receive [57]

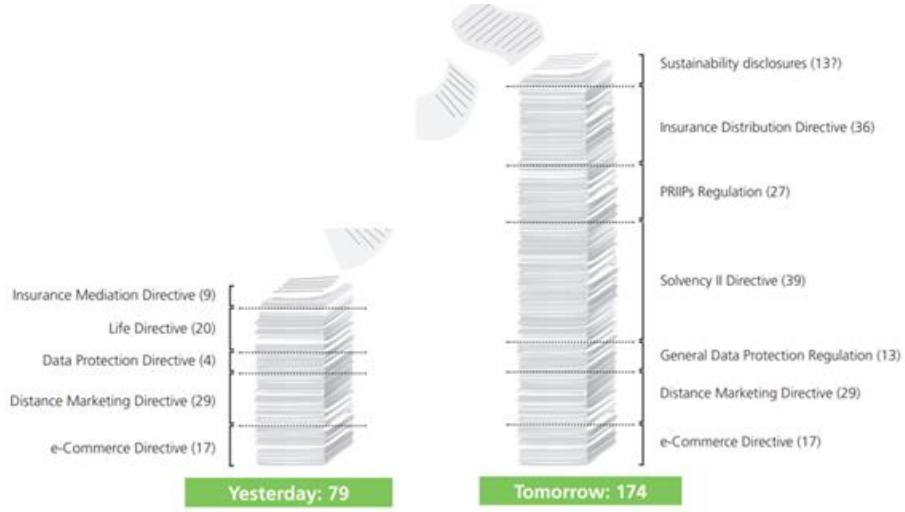


Figure 28:Comparison of EU disclosure requirements for consumers buying a sustainable insurance-based investment product (online sale by a broker, including duplications) current and future [57]

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

10.2 Secure-by-Design in the Industrial Internet of Things (IIoT) context

The advent of fourth industrial revolution is based on the developments of the advanced connectivity and software-based systems. Traditional Industrial Automation and Control Systems (ICAS) and Supervisory control and data acquisition systems were isolated systems from the digital networks at large. Riding on the waves of the Industry 4.0, massive adoption of the Internet of Things (IoT) or known as Industrial Internet of things (IIoT) and the Cyber Physical Systems (CPS) is changing the operational typical paradigm characteristics of the industrial ecosystems [1]. The primary requirement for a typical Industrial control system is monitoring and controlling the physical processes in the industrial context with its heart as Operational Technology (OT). The Operational Technology systems are expected to be highly available and safety critical systems. The introduction of the industry 4.0 is calling for the amalgamation of Operational Technology systems with Information Technology (IT) which was not part of the Industrial ecosystem as a co-habitant technology rather than just for supportive purposes such as enterprise systems [2].

The June 2010 marked the discovery of one of the first highly potent malware with weaponized payload that was targeting the Industrial control systems, STUXNET [3]. Since it was highly publicized, it helped in creating a general awareness about the hazardous nature of software-based malware among the industrial community [4 - 6]. However, there are several reported incidents of software-based attacks and cybersecurity attacks that is affecting the industrial control systems and this number is increasing substantially year after year [7]. This clearly assert the deficiencies in the defensive actions taken by the industrial community that are already aware of the security challenges. In recent years the security by design philosophy has taken more prominence to address the security challenges among many others [8]. In addition to this longer operational life cycle, presence of active legacy systems and mission critical nature of such industrial systems often mandate regulatory requirements such as secure development life cycle [9].

Standards like IEC 62443, NIST 800-82, NERC CIP, IEEE 1686, IEC 63096 are some of the international standards that are defining the standard security requirements for industrial systems. These standards are derived and distilled by the effort of a global community involved in the active development and consumption of the industrial systems related technology and they are also proof of the global acknowledgment of the need for the security in such systems. Recently specifically for industry 4.0 the architectural reference model called RAMI 4.0 is getting significant traction among the adapters of industry 4.0 based industrial systems [9]. Cross-industry standards such as IEC 62443 amends the need for secure development lifecycle which includes secure design, secure implementation, verification and validation, entropy management and product disposal [10]. However, these standards often usually do not provide a step-by-step guideline rather than set of constraint often causing significant difficulty in adoption at various stages. In addition, the highly complex nature and difficulty in training the developers cause formidable challenges that are difficult to overcome [11, 12].

In light of many studies done over the last few years it is generally considered humans as the weakest link in software development or even as end users when it comes to security challenges regardless of the domain. Reasons for these challenges' ranges absence of the security awareness due to lack of training to basic careless mistakes [13 - 15]. With the backing of necessary supportive tools for development it is often seen that the security of the developed systems is improved overcoming the

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

human factor challenges to a certain extend [16]. However, there are challenges in adopting secure development tools for the companies. The companies and the users will take time to adopt to the new changes [17].

An SDLC framework will help an organization to develop software from its origin to its end of life. There are many different types of secure development approaches available and generally from a high-level point of view they all have the common steps as follows.

- Requirement identification
- Architecture Design
- Coding
- Testing
- Production and maintenance of the application.

Currently there are several development secure lifecycle models used in the software industry such as

- McGraw Model
- Microsoft SDL
- CLASP (Comprehensive Lightweight Application Security process)
- TSP Teams Software Process for Secure software Development
- Rational Unified Process Secure (RUPSec)
- Building Security in Maturity Model (BSIMM)
- Software Assurance Maturity Model (OPEN SAMM)
- Appropriate and Effective Security Guidance for the Information Security (AEGIS)
- Secure Software Development Model (SSDM)
- Writing Secure Code
- Waterfall-based software security engineering process model
- Secure Software Development Model (SecSDM)
- Software Development Process Model (S2D-ProM)
- Correctness by Construction (CbyC)
- Security Quality Requirements Engineering (SQUARE) methodology.

One of the primary challenges in using these models is the lack of flexibility in adopting to the organizational requirements [18]. Additionally, these models are targeted for the IT based systems than the OT systems and they have different security priorities [19,20]. At present there is visible gap in OT secure development domain [21].

According to our experience, the members of the R&D organization typically face the following challenges, grouped by the work role:

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Management:

- It may be hard to ensure support from top management for the cyber security spending
- Hiring cyber security competence is hard
- Motivating and engaging the organization to accept the need for change
- It may be hard to ensure the bandwidth for cyber security in practice, given all other business priorities

Cyber security leaders:

- Visibility to the current state of how the process is working may be poor
- Feedback loops from making a change in the process to hearing feedback and seeing the impacts can be too long, as typically feedback may need to be collected with separate surveys or interviews
- It can be tedious to scale the process up to be adopted by more R&D teams
- When the process is maintained and improved, it is hard to roll out the improvements especially when some teams have already used the previous versions of the templates and blueprints.

Project managers, developers, architects, and test engineers:

- It may be too hard or tedious in practice to follow the process. Even knowning where to find the latest templates or what exactly needs to be done in a certain phase
- The process may turn out to be too complex or too abstract to understand
- There may not be enough supporting tools, so cyber security may have to be managed using ad hoc tools
- Sometimes the user experience of the tool portfolio may be poor

Our aim is to exceed the state of the art with the Oxilate research and address the above-memntioned challenges by the following aspects:

- Address the practical challenges related to leadership and adoption of cyber security management better than the state of art. We aim to base this on developing a new service model that supports the organization. For example, the service model might consist of a repeatable adoption method, a training method, a digital model of the process with tooling support and cyber security expert support.
- Digitalize the secure development lifecycle models so that applying the practices would be self-explanatory to the end users. The digital model should also provide contextual advice to the end users so that instead of having to refer to the full process descriptions, the user would be shown only those parts that are relevant to the user's current context.
- Generate visibility, metrics and evidence of practicing the models automatically, just from the end user's usage of the digitalized model.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

11. Bibliography

Easing Power Consumption of Wearable Activity Monitoring with Change Point Detection. **Culman, & Aminikhanghahi, Samaneh & Cook, J.. 2020.** 2020, Sensors.

IDSIA. Sepp Hochreiter's Fundamental Deep Learning Problem (1991). [Online] IDSIA. http://people.idsia.ch/~juergen/fundamentaldeeplearningproblem.html.

A Comprehensive Survey on Graph Neural Networks. **Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang and P. S. Yu. 2021.** 1, 2021, IEEE Transactions on Neural Networks and Learning Systems, Vol. 23, pp. 4-24.

A Critical Review on Selected Fuzzy Min-Max Neural Networks and Their Significance and Challenges in Pattern Classification. **Alhroob, Essam, et al. 2019.** s.l. : IEEE, Apr 18, 2019, IEEE Access.

A learning scheme for asemetric threshold network. **LeCun, Yann. 1985.** 1985, Cognitiva.

A Logical Calculus of the ideas immanent in nervous activity. **Warren S. McCullough, Walter H. Pitts. 1943.** 1943, Bulletin of Mathematical Biophysics, Vol. 5, pp. 115-133.

A Neural Conversational Model. **Oriol Vinyals, Quoc Le. 2015.** 2015. ICML Deep Learning Workshop 2015.

A Novel Connectionist System for Unconstrained Handwriting Recognition. **A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber,. 2009.** 5, 2009, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, pp. 855-868.

A Survey on Deep Learning for Named Entity Recognition. **Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020.** s.l. : IEEE, 2020, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2020.

A Survey on Knowledge Graph Embedding: Approaches, Applications and Benchmarks. **Yuanfei Dai, Shiping Wang, Naixue Xiong, Wenzhong Guo. 2020.** 5, May 2020, Electronics, Vol. 9.

AccordProject.org. Open source software tools for smart legal contracts. *AccordProject.org.* [Online] AccordProject.org. https://accordproject.org/.

Adam Back, Matt Corallo, Luke Dashjr,Mark Friedenbach, Gregory Maxwell,Andrew Miller, Andrew Poelstra, Jorge Timón, and Pieter Wuille. 2014. *Enabling Blockchain Innovations with Pegged Sidechains.* s.l. : Blockstream, 2014.

AIDAN HOGAN, et. Al. 2021. *Knowledge Graphs.* 2021.

Al., Carolin Baker et. 2020. Chapter 7 Transfer Learning for NLP I. *Modern Approaches in Natural Language Processing.* s.l. : Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License., 2020.

Alder, Jason L. Hutchens and Michael D. 1998. Introducing MegaHal. *on Human Computer Conversation, ACL, pp 271-274.* 1998.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training.* s.l. : Open AI, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners.* s.l. : OpenAI, 2019.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Alephzero.org. Alephzero.org. *Alephzero.org.* [Online] Alephzero.org. https://alephzero.org/.

Alexey Akhunov(@AlexeyAkhunov), Eli Ben Sasson (eli@starkware.co), Tom Brand (tom@starkware.co), Avihu Levy (avihu@starkware.co). 2019. eip-2028.md. *Github.* [Online] Github, 05 03, 2019. https://github.com/ethereum/EIPs/blob/77aa54f578b13e15c45d22dc1d5f9d93e231366c/EIPS/eip-2028.md.

AliceBot. AliceBot. *AliceBot.* [Online] AliceBot. https://alicebot.org/.

AlphaZero.org. 2021. What Is The Fastest Blockchain And Why? Analysis of 43 Blockchains. *AlphaZero.org.* [Online] AlphaZero.org, Jan 24, 2021. https://alephzero.org/blog/what-is-the-fastest-blockchain-and-why-analysis-of-43-blockchains/.

Ambiq.com. 2021. Ambiq.com. *Ambiq.com.* [Online] Ambiq.com, 2021. https://ambiq.com/technology/.

Amritraj Singh, Kelly Click, Reza M. Parizi, Qi Zhang, Ali Dehghantanha, Kim-Kwang, Raymond Choo. 2020. Sidechain technologies in blockchain networks: An examination and state-of-the-art review. *Journal of Network and Computer Applications* - . 2020, Vol. 149.

An architectural design of Virtual Dietitian (ViDi) for diabetic patients. **Abbas Saliimi Lokman, Jasni Mohamad Zain. 2009.** s.l. : IEEE, 2009. International Conference on Computer Science and Information Technology, 2009.

Antonio Salazar Cardozo, Zachary Williamson. 2018. https://eips.ethereum.org/EIPS/eip-1108. *EIP-1108: Reduce alt_bn128 precompile gas costs.* [Online] etherium.org, 05 21, 2018. https://eips.ethereum.org/EIPS/eip-1108.

Artificial Intelligence Markup Language. Artificial Intelligence Markup Language. *Artificial Intelligence Markup Language.* [Online] Artificial Intelligence Markup Language. http://www.aiml.foundation/.

Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. **Karl de Fine Licht, Jenny de Fine Licht. 2020.** s.l. : Springer, 2020, AI & SOCIETY , Vol. 35.

Attention Is All You Need. **Al., Ashish Vaswani et. 2017.** Long Beach, CA : s.n., 2017. 31st Conference on Neural Information Processing Systems (NIPS 2017).

Babayan, Davit. 2019. Long-Time Investor: IOTA is Centralized, Single Point of Failure Exists. *NewsBTC.com.* [Online] 2019. https://www.newsbtc.com/news/blockchain/long-time-investor-iota-is-centralized-single-point-of-failure-exists/.

BadVR . 2019. BadVR Awarded Magic Leap Independent Creator Program Grant; Will Data Drive the Future of AR? *PRNewswire.com.* [Online] PR Newswire, Feb 28, 2019. https://www.prnewswire.com/news-releases/badvr-awarded-magic-leap-independent-creator-program-grant-will-data-drive-the-future-of-ar-300803751.html.

Big Data LifeCycle: Threats and Security Model. **Yazan Alshboul, Yong Wang, Raj Kumar Nepali. 2015.** Puerto Rico : s.n., 2015. Twenty-first Americas Conference on Information Systems.

Big data: A review. **S. Sagiroglu, Duygu Sinanc. 2013.** s.l. : IEEE, 2013.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

blockescence DLT. 2019. The Road To Adoption (Part 1): The Scalability Trilemma. *blockescence DLT solutions.* [Online] Medium.com, May 24, 2019. https://medium.com/blockescence-dlt-solutions/the-road-to-adoption-part-1-the-scalability-trilemma-edfdd4dc6a9d.

Bloxroute.com. Bloxroute.com. *Bloxroute.com.* [Online] Bloxroute.com. https://bloxroute.com/.

Bruce Wilcox, Sue Wilcox. 2011. *Suzette, the Most Human Computer* . San Rafael, CA : Telltale Games, 2011.

Building Applied Natural Language Generation Systems. **Ehud Reiter, Robert Dale. 1995.** 1, s.l. : Cambridge University Press, 1995, Natural Language Engineering, Vol. 1.

Chandra, Akshay L. 2018. McCulloch-Pitts Neuron — Mankind's First Mathematical Model Of A Biological Neuron. [Online] Towards Datascience, July 24, 2018. https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1.

Chen-Burger, Yun-Heh. 2011. Knowledge Modelling and Management. *Knowledge Modelling and Management.* [Online] The University of Edinburgh, 02 28, 2011. http://www.inf.ed.ac.uk/teaching/courses/kmm/PDF/5-knowledge-model-20110228.pdf.

Coindesk.com. Bitcoin Cash BCH. *Coindesk.com.* [Online] Coindesk.com. https://www.coindesk.com/crypto/bitcoin-cash.

Cryptokitties.co. Key Information. *Cryptokitties.co/technical-details.* [Online] https://www.cryptokitties.co. https://www.cryptokitties.co/technical-details.

DAGsim: Simulation of DAG-based distributed ledger protocols. **Manuel Zander, Tom Waite, Dominik Harz. 2018.** Toulouse, France : Symposium on Cryptocurrency Analysis (SOCCA), 2018.

Dan Brickley, Libby Miller. 2000. FOAF. *FOAF.* [Online] FOAF, 2000. http://xmlns.com/foaf/spec/.

Data Science Project Management. Team Data Science Process (TDSP). *Data Science Project Management.* [Online] Data Science Project Management. https://www.datascience-pm.com/tdsp/.

Datascience-pm.com. CRISP-DM. *Data Science Project Management.* [Online] Datascience-pm.com. https://www.datascience-pm.com/crisp-dm-2/.

Decentralized & Collaborative AI on Blockchain. **Waggoner, Justin D. Harris Bo. 2019.** s.l. : IEEE, 2019. 2019 IEEE International Conference on Blockchain (Blockchain).

Decentralized Autonomous Organizations: Concept, Model, and Applications. **S. Wang, W. Ding, J. Li, Y. Yuan, L. Ouyang and F. Wang,. Oct. 2019.** 5, s.l. : IEEE , Oct. 2019, Transactions on Computational Social Systems,, Vol. 6.

DeepBrain Chain. *DeepBrain Chain, Artificial Intelligence Computing Platform Driven By BlockChain.*

DeepBrainChain.org. DeepBrainChain.org. *DeepBrainChain.org.* [Online] DeepBrainChain.org. https://www.deepbrainchain.org/.

Detrixhe, John. 2020. The simple reason DocuSign doesn't use blockchain. *https://qz.com/.* [Online] Quartz, Dec 07, 2020. https://qz.com/1942479/docusign-ceo-says-blockchain-is-too-expensive-for-wide-adoption/.

Diffusion convolutional recurrent. **Y. Li, R. Yu, C. Shahabi, and Y. Liu. 2018.** 2018. ICLR.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Distantly Supervised NER with Partial Annotation Learning and Reinforcement Learning. Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, Min Zhang. 2018. Santa Fe, New Mexico, USA : s.n., 2018. Proceedings of the 27th International Conference on Computational Linguistics. pp. 2159–2169.

Distributed representations of words and phrases and their compositionality. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado , Jeffrey Dean. 2013. 2013. NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems. Vol. 2, pp. 3111–3119.

EDMCouncil. About FIBO - The open semantic standard for the financial industry. *EDMCouncil.org.* [Online] EDMCouncil. https://edmcouncil.org/page/aboutfiboreview.

ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. 2020. s.l. : Google Research, 2020. ICLR 2020 Conference.

Embedding-Based Entity Alignment Using Relation Structural Similarity. Peng, Yanhui, et al. 2020. s.l. : IEEE, 2020. IEEE International Conference on Big Knowledge (ICBK).

Ethereum. 2019. Etherium Casper. *Github.* [Online] Etherium, Mar 28, 2019. https://github.com/ethereum/casper.

Extension and Prerequisite: An Algorithm to Enable Relations Between Responses in Chatbot Technology. Abbas Saliimi Lokman, Jasni Mohamad Zain. 2010. 10, 2010, Journal of Computer Science, Vol. 6.

Facilitating new knowledge creation and obtaining KM maturity. Priscilla A. Arling, Mark W.S. Chun. 2011. 15, 2011, Journal of Knowledge Management, Vol. 2.

Fadnis, Kshitij, et al. 2019. *Path-Based Contextualization of Knowledge Graphs for Textual Entailment.* s.l. : arXiv, 2019.

Foote, Keith D. 2016. A Brief History of Artificial Intelligence. [Online] April 5, 2016. https://www.dataversity.net/brief-history-artificial-intelligence/#.

—. 2017. A Brief History of Big Data. *Dataveristy.* [Online] Dataversity.net, Dec 14, 2017. https://www.dataversity.net/brief-history-big-data/.

FRANKENFIELD, JAKE. 2021. Hard Fork (Blockchain). *Investopedia.com.* [Online] Investopedia.com, Jan 28, 2021. Hard Fork (Blockchain).

—. 2019. Smart Contracts. *investopedia.com.* [Online] investopedia.com, Oct 08, 2019. https://www.investopedia.com/terms/s/smart-contracts.asp.

Fred Sala, Ines Chami, Adva Wolf, Albert Gu, Beliz Gunel and Chris Ré. 2019. Into the Wild: Machine Learning In Non-Euclidean Spaces. [Online] Stanford University, Oct 10, 2019. https://dawn.cs.stanford.edu/2019/10/10/noneuclidean/.

Fuzzy sets. panelL.A.Zadeh, Author links open overlay. 1965. 3, 1965, Information and Control, Vol. 8, pp. 338-353.

Gartner. 2019. Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019. *Gartner Newsroom.* [Online] Gartner Research, Feb 18, 2019. https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Gilbert H., Handschuh H. 2003. Security Analysis of SHA-256 and Sisters. [book auth.] Zuccherato R.J. (eds) Matsui M. *Selected Areas in Cryptography.* Berlin : Springer, 2003.

Graves, Alex. 2013. *Generating Sequences With Recurrent Neural Networks.* 2013.

Gray, Marley. 2019. Ethereum Blockchain as a Service now on Azure. *Microsoft Azure.* [Online] Microsoft, Nov 09, 2019. https://azure.microsoft.com/es-es/blog/ethereum-blockchain-as-a-service-now-on-azure/.

Harris, Justin D. 2019. Leveraging blockchain to make machine learning models more accessible. *Microsoft Research Blog.* [Online] Microsoft , Jul 12, 2019. https://www.microsoft.com/en-us/research/blog/leveraging-blockchain-to-make-machine-learning-models-more-accessible/.

Hochreiter, Josef. 1991. DIPLOMARBEIT IM FACH INFORMATIK. [Online] June 15, 1991. http://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf.

Horev, Rani. 2018. BERT Explained: State of the art language model for NLP. [Online] Towards Data Science, Nov 10, 2018. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

How evolutionary algorithms are applied to statistical natural language processing. **Araujo, Lourdes. 2007.** 2007, Artificial Intelligence Review volume, Vol. 20, pp. 275–303.

Huang, Kexin & Altosaar, Jaan & Ranganath, Rajesh. 2019. *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. .* 2019.

Hughes-Cromwick, Ellen. Cryptocurrency Energy Consumption. *Cryptocurrency Energy Consumption.* [Online] Energy INstitute University of Michigan. https://energy.umich.edu/news-events/energy-economics-weekly-briefings/story/cryptocurrency-energy-consumption/.

IBM Cloud Education. 2020. Recurrent Neural Networks. [Online] IBM, 09 14, 2020. https://www.ibm.com/cloud/learn/recurrent-neural-networks.

IBM. 2016. *The future of health is cognitive, Harnessing data and insight to deliver better health, value and.* s.l. : IBM Healthcare and Life Sciences, 2016.

iden3.io. Key concepts. *docs.iden3.io.* [Online] iden3.io. https://docs.iden3.io/#/basics/key-concepts?id=zk-snarks.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, Yuji Matsumoto. 2020. *LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention.* s.l. : arXiv.org, 2020.

Innoplexus. *The convergence of Blockchain and AI in Drug Discovery.*

Iota.org. The Tangle. *docs.iota.org.* [Online] Iota.org. https://docs.iota.org/docs/getting-started/0.1/network/the-tangle.

J. Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT 2019.* 2019.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Jamiel Sheikh. Digital Investment Assets. *Dropbox.com.* [Online] Chainhaus / MIT. https://www.dropbox.com/s/jerlpyk9rx4o68w/Digital%20Investment%20Assets%20Frontiers.pptx?dl=0.

Jeremy Howard, Sebastian Ruder. 2018. *Universal Language Model Fine-tuning for Text Classification.* 2018.

Jorge Decorte, Sidharth Mudgal. 2020. Deepmatcher. *GitHub.* [Online] 2020. https://github.com/anhaidgroup/deepmatcher.

Joseph Poon, Vitalik Buterin. 2017. *Plasma: Scalable Autonomous Smart Contracts.* s.l. : Plasma.io, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio. 2014. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.* s.l. : Universite de Montreal, 2014.

Kaminska, Izabella. 2019. Blockchain officially confirmed as slower and more expensive. *Financial Times.* [Online] FT.com, May 29, 2019. https://www.ft.com/content/fe5b17e1-4040-3249-b473-f3c998c67de9.

Kiela, Maximilian Nickel and Douwe. 2017. *Poincaré Embeddings for Learning Hierarchical Representations.* s.l. : Facebook AI Research, 2017.

Kim, Christine. 2020. Everything You Need to Know About Ethereum 2.0. *coindesk.com.* [Online] coindesk.com, Jul 24, 2020. https://www.coindesk.com/everything-you-need-to-know-about-ethereum-2-0.

Klemens, Sam. 2021. What is EOS? EOS Coin. *Exodus.com.* [Online] Exodus.com, Jan 05, 2021. https://www.exodus.com/blog/what-is-eos-coin/.

L., Kenny. 2019. The Blockchain Scalability Problem & the Race for Visa-Like Transaction Speed. *Towards Data Science.* [Online] Medium.com, Jan 30, 2019. https://towardsdatascience.com/the-blockchain-scalability-problem-the-race-for-visa-like-transaction-speed-5cce48f9d44#:~:text=Visa%20does%20around%201%2C700%20transactions,is%20bottlenecked%20currently%20by%20scalability..

Lapata, Yang Liu and Mirella. 2019. *Text Summarization with Pretrained Encoders.* s.l. : The University of Edinburgh, 2019.

Le, Zhilin Yang and Quoc. 2019. Transformer-XL: Unleashing the Potential of Attention Models. [Online] Google AI, Jan 29, 2019. https://ai.googleblog.com/2019/01/transformer-xl-unleashing-potential-of.html.

Lee, Sherman. 2018. Explaining Side Chains, The Next Breakthrough In Blockchain. *Forbes.com.* [Online] Forbes.com, Feb 07, 2018. https://www.forbes.com/sites/shermanlee/2018/02/07/explaining-side-chains-the-next-breakthrough-in-blockchain/.

Lee, Tim Berners. 2007. Giant Global Graph. *Semantic Web Technologies .* [Online] MIT, 03 27, 2007. https://web.archive.org/web/20160713021037/http://dig.csail.mit.edu/breadcrumbs/node/215.

Lefkowitz, Melanie. 2019. Professor's perceptron paved the way for AI – 60 years too soon. [Online] Cornell University, Sept 25, 2019. https://news.cornell.edu/stories/2019/09/professors-perceptron-paved-way-ai-60-years-too-soon.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Link Prediction in Criminal Networks: A Tool for Criminal Intelligence Analysis. Giulia Berlusconi, Francesco Calderoni ,Nicola Parolini,Marco Verani,Carlo Piccardi. April 22, 2016. s.l. : PLOS ONE, April 22, 2016.

Long Short-temr Memory. Schmidhuber, Sepp Hochreiter and Jürgen. 1997. 8, 1997, Neural Computation, Vol. 9, pp. 1735-80.

Look, Carolynn. 2016. Blockchain Settlement Was Slow, Costly in Trial, Weidmann Says. *Bloomberg.com.* [Online] Bloomberg.com, May 29, 2016. Blockchain Settlement Was Slow, Costly in Trial, Weidmann Says.

MAJASKI, CHRISTINA. 2020. Distributed Ledgers. [Online] Investopedia, May 12, 2020. https://www.investopedia.com/terms/d/distributed-ledgers.asp.

Manning, Minh-Thang Luong Hieu Pham Christopher D. 2015. *Effective Approaches to Attention-based Neural Machine Translation.* Stanford, CA : Stanford University, 2015.

Marr, Bernard. 2017. How Walmart Is Using Machine Learning AI, IoT And Big Data To Boost Retail Performance. *Forbes.* [Online] Forbes, Aug 29, 2017. https://www.forbes.com/sites/bernardmarr/2017/08/29/how-walmart-is-using-machine-learning-ai-iot-and-big-data-to-boost-retail-performance/?sh=33b01e6f6cb1.

Martin Huschenbett, Patrick Gallagher, Sam Richards, Ryan Cordell. 2021. INTRODUCTION TO DAPPS. *Ethereum.org.* [Online] Ethereum.org, Jan 12, 2021. https://ethereum.org/en/developers/docs/dapps/.

Marvin Minsky, Seymour Papert. 1969. *A Review of "Perceptrons: An Introduction to Computational Geometry".* Ithaca, New York 14850 : Department of Theoretical and Applied Mechanics, Cornell University,, 1969.

Microsoft. 2020. Sharing Updatable Models (SUM) on Blockchain. *GitHub.* [Online] Microsoft, Nov 23, 2020. https://github.com/microsoft/0xDeCA10B.

—. The Garage is a program that drives a culture of innovation. *Microsoft.com.* [Online] Microsoft. https://www.microsoft.com/en-us/garage.

—. The Team Data Science Process lifecycle. *Docs.microsoft.com.* [Online] Microsoft.com. https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle.

Nayak, Pandu. 2019. Understanding searches better than ever before. [Online] Google, Oct 25, 2019. https://www.blog.google/products/search/search-language-understanding-bert/.

Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, Kunihiko. 1980. 1980, Biological Cybernetics, Vol. 36, pp. 193–202.

Neural Embeddings of Graphs in Hyperbolic Space. Benjamin Paul Chamberlain, James Clough, Marc Peter Deisenroth. 2017. 2017. 13th international workshop on mining and learning from graphs held in conjunction with KDD.

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. Dzmitry Bahdanau, KyungHyun Cho & Yoshua Bengio. 2015. 2015. ICLR 2015.

NICLAS KANNENGIEßER, SEBASTIAN LINS, TOBIAS DEHLING, and ALI SUNYAEV. 2020. Trade-offs between Distributed Ledger Technology Characteristics. *ACM Comput. Surv.* 2020, Vol. 53, 2.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Nigel Gopie, PhD. 2018. What are smart contracts on blockchain? *IBM Blockchain Blog.* [Online] IBM, Jul 02, 2018. https://www.ibm.com/blogs/blockchain/2018/07/what-are-smart-contracts-on-blockchain/.

Njui, John P. 2018. HOW SHARDING ON THE ETHEREUM NETWORK MIGHT BE AROUND THE CORNER. *Ethereumworldnews.com.* [Online] Ethereumworldnews.com, Sep 15, 2018. https://ethereumworldnews.com/how-sharding-on-the-ethereum-network-might-be-around-the-corner/.

Nltk.org. Natural Language Tool Kit. *Nltk.org.* [Online] Nltk.org. https://www.nltk.org/book/ch10.html.

Numerai. 2017. Numerai. *Numerai.* [Online] Numerai, 2017. https://numer.ai/.

On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. **Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, Yoshua Bengio. 2014.** Doha, Qatar : s.n., 2014. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation.

OntBot: Ontology based chatbot. **Hadeel Al-Zubaide, A. A. Issa. 2011.** Amman, Jordan : IEEE, 2011. International Symposium on Innovations in Information and Communications Technology. pp. 7-12.

patio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. **B. Yu, H. Yin, and Z. Zhu,. 2018.** 2018. IJCAI, .

Perspective-tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. **I. Nonaka, G. von Krogh. 2009.** 3, 2009, Organization Science, Vol. 20, pp. 635-652.

Peters, Matthew & Neumann, Mark & Iyyer, Mohit & Gardner, Matt & Clark, Christopher & Lee, Kenton & Zettlemoyer, Luke. 2018. *Deep contextualized word representations. .* 2018.

Phi, Michael. 2018. Illustrated Guide to LSTM's and GRU's: A step by step explanation. [Online] Medium, Sept 04, 2018. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

Phillips, Daniel. 2020. Someone just made a $2.6 million mistake on Ethereum. *Decrypt.co.* [Online] Decrypt.co, Jun 10, 2020. https://decrypt.co/31830/someone-just-made-a-2-6-million-mistake-on-ethereum.

Polynomial Theory of Complex Systems. **IVAKHNENKO, A. G. 1971.** 4, 1971, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, Vols. SMC-1, pp. 364-378.

Prasad, Pandit. 2020. Reducing IT cost with an effective database strategy. *Panditprasad.medium.com.* [Online] Medium.com, Mar 23, 2020. https://panditprasad.medium.com/reducing-it-cost-with-an-effective-database-strategy-5fe03a169ad0.

Qi Zhu, Hao Wei, Bunyamin Sisman, Da Zheng, Christos Faloutsos, Xin Luna Dong, Jiawei Han. 2020. Collective knowledge graph multi-type entity alignment. *Amazon Science.* [Online] Amazon.com, 2020. https://www.amazon.science/publications/collective-knowledge-graph-multi-type-entity-alignment.

Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo and Llion Jones. 2018. Character-Level Language Modeling with Deeper Self-Attention. [Online] Google AI, Dec 10, 2018. https://arxiv.org/pdf/1808.04444.pdf.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Receptive fields, binocular interaction, and functional architecture. **Hubel, D. H. and Wiesel, T. 1962.** 1962, Journal of Physiology, Vol. 160, pp. 106-154.

Review of State-of-the-Art Design Techniques for Chatbots. **Ritu Agarwal, Mani Wadhwa. 2020.** 1, Jul 29, 2020, SN Computer Science, Vol. 1.

—. **Wadhwa, Ritu Agarwal & Mani. 2020.** Jul 20, 2020, SN Computer Science, Vol. 1.

RI Revisited: Four Years in the Trenches . **Judith Bachant, John McDermott. 1984.** 3, 1984, AI Magazine, Vol. 5 , pp. 21-32.

Rosnay, J. de. 2000. History of Cybernetics and Systems Science. [Online] Principia Cybernetica Web, 2000. http://pespmc1.vub.ac.be/CYBSHIST.html.

Saint-Leger, Sacha. 2019. Istanbul, zkRollup, and Ethereum throughput limits: an analysis. *iden3.io.* [Online] iden3.io, Dec 12, 2019. https://blog.iden3.io/istanbul-zkrollup-ethereum-throughput-limits-analysis.

SALTZ, JEFF. 2020. CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects. *Data Science Project Management.* [Online] Data Science Project Management, Nov 30, 2020. https://www.datascience-pm.com/crisp-dm-still-most-popular/.

SAS.com. Introduction to SEMMA. *https://documentation.sas.com/.* [Online] SAS.com. https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en.

—. SAS® Enterprise Miner™. *SAS.com.* [Online] SAS.com. https://support.sas.com/en/software/enterprise-miner-support.html.

Schema.org. Schema.org. *Schema.org.* [Online] Schema.org. https://schema.org/.

Schmidhuber, Jurgen. 2014. *Deep Learning in Neural Networks: An Overview.* Manno-Lugano, Switzerland : The Swiss AI Lab IDSIA, 2014.

SciBERT: A Pretrained Language Model for Scientific Text. **Iz Beltagy, Kyle Lo, Arman Cohan. 2019.** 2019, EMNLP/IJCNLP 2.

Science Data Visualization in AR/VR for Planetary and Earth Science. **Grubb, T., et al. December 2018.** s.l. : American Geophysical Union, December 2018.

SCILLA. Safe-By-Design Smart Contract Language. *SCILLA.* [Online] SCILLA. https://scilla-lang.org/.

Semi-supervised learning for named entity recognition using weakly labeled training data,. **A. Zafarian, A. Rokni, S. Khadivi and S. Ghiasifard. 2015.** Mashhad, Iran : s.n., 2015. The International Symposium on Artificial Intelligence and Signal Processing (AISP). pp. 129-135.

Seq2seq Dependency Parsing. **Zuchao Li, Jiaxun Cai, Shexia He, Hai Zhao. 2018.** Santa Fe, New Mexico, USA : s.n., 2018. Proceedings of the 27th International Conference on Computational Linguistics. pp. 3203–3214.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Sharma, Ruchika. 2020. How Big Data and Analytics Reshape the Wearable Device Market. *HealthTechAdvisor.com.* [Online] HealthTechAdvisor.com, Oct 27, 2020. https://healthtechadvisor.com/how-big-data-and-analytics-reshape-the-wearable-device-market/.

Sidechain technologies in blockchain networks: An examination and state-of-the-art review. **Amritraj Singh, Kelly Click, Reza M.Parizi, Qi Zhang, Ali Dehghantanha, Kim-Kwang, Raymond Choo. 2020.** s.l. : Elsevier, 2020, Vol. 149.

Sigrid Seibold, George Samman. 2016. *Consensus, Immutable agreement for the internet of value.* s.l. : KPMG, 2016.

Singh, Rohit Kumar. 2021. A 2021 Guide to Named Entity Recognition. *Nanonets.com.* [Online] Nanonets.com, Jan 2021. https://nanonets.com/blog/named-entity-recognition-2020-guide/.

Singhal, Amit. 2012. Introducing the Knowledge Graph: things, not strings. *Google Blog.* [Online] Google, May 16, 2012. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.

SingularityNET. 2019. *SingularityNET: A Decentralized, Open Market and Network for AIs.* Amsterdam : SingularityNET, 2019.

Sloane Brakeville, Bhargav Perepa. 2019. Blockchain basics: Introduction to distributed ledgers. *IBM Developer.* [Online] IBM, June 01, 2019. https://developer.ibm.com/technologies/blockchain/tutorials/cl-blockchain-basics-intro-bluemix-trs/.

Spatial temporal graph convolutionalnetworks for skeleton-based action recognition . **S. Yan, Y. Xiong, and D. Lin,. 2018.** 2018. AAAI,.

Stanford University. 2015 - 2019. CS231n Convolutional Neural Networks for Visual Recognition. [Online] Stanford University, 2015 - 2019. https://cs231n.github.io/convolutional-networks/.

Statista. 2021. Market share of smartwatch unit shipments worldwide from the 2Q'14 to 1Q '20*, by vendor. *Statista.com.* [Online] Statista, Jan 22, 2021. https://www.statista.com/statistics/524830/global-smartwatch-vendors-market-share/.

Sterbak, Tobias. 2020. Named entity recognition with Bert. *depends-on-the-definition.* [Online] https://www.depends-on-the-definition.com/, 04 28, 2020. https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/.

Streamr.network. 2020. Streamer Case Study — Tracey: TOWARDS SUSTAINABLE & TRACEABLE FISHERIES. *Streamr.network.* [Online] Streamr.network, 2020. https://streamr.network/case-studies/tracey/.

—. What is Streamr? *Streamr.network.* [Online] Streamr.network. https://streamr.network/docs/introduction.

Streichert, Felix. 2005. *Introduction to Evolutionary Algorithms.* s.l. : University of Tuebingen, 2005.

Supervised neural networks for the classification of structures. **Starita, A. Sperduti and A. 1997.** 3, 1997, IEEE Transactions on Neural Networks, Vol. 8, pp. 714-735.

The Advantages and Disadvantages of the Blockchain Technology. **Julija Strebko, Andrejs Romanovs. 2018.** 2018. 2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE).

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

The Bitcoin Backbone Protocol: Analysis and Applications. **Juan A. Garay, Aggelos Kiayias, Nikos Leonardos. 2015.** s.l. : Eurocrypt 15, 2015.

The first computational theory of mind and brain: a close look at McCulloch and Pitts's "logical calculus of ideas immanent in nervous activity". **Piccinini, Gualtiero. 2004.** 2004, Synthese, Vol. 141, pp. 175–215.

The ZILLIQA Team. 2017. *The ZILLIQA Technical Whitepaper.* s.l. : Zilliqa.com, 2017.

Ting Hu, Karoliina Oksanen,Weidong Zhang,Ed Randell,Andrew Furey,Guang Sun,Guangju Zhai. 2018. *An evolutionary learning and network approach to identifying key metabolites for osteoarthritis.* s.l. : PLOS Computational Biology, 2018.

Tizen.org. Tizen.org. *Tizen.org.* [Online] Tizen.org. https://www.tizen.org/.

TRAINING A 3-NODE NEURAL NETWORK IS NP-COMPLETE. **Avrim Blum, Ronald L. Rivest. 1992.** 1, 1992, Neural Networks, Vol. 5, pp. 117-127.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Dan. 2020. *Language models are few-shot learners.* s.l. : OpenAI, 2020.

Transfer Learning Across Human Activities Using a Cascade Neural Network Architecture. **Xin Du, Katayoun Farrahi. 2019.** London : s.n., 2019. International Symposium on Wearable Computers (ISWC 2019).

University of Illinois Chicago. 2019. Big Data and Wearable Health Monitors: Harnessing the Benefits and Overcoming Challenges. *Healthinformatics.uic.edu.* [Online] University of Illinois Chicago, 2019. https://healthinformatics.uic.edu/blog/big-data-and-wearable-health-monitors-harnessing-the-benefits-and-overcoming-challenges/.

Urvashi Khandelwal, He He, Peng Qi, Dan Jurafsky. 2018. *Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context.* 2018.

Venkatachalam, Mahendran. 2019. An introduction to Attention, The why and the what. [Online] Towards Data Science, Jun 29, 2019. https://towardsdatascience.com/an-introduction-to-attention-transformers-and-bert-part-1-da0e838c7cda.

Visa. *VISA Fact Sheet - What you need to know about one of the world's largest payments companies.* s.l. : Visa.

Viswav, Pradeep. 2018. Microsoft Garage releases a new data visualization tool for PC and Surface Hub. *MSPoweruser.com.* [Online] MSPoweruser.com, Jul 31, 2018. https://mspoweruser.com/microsoft-garage-releases-a-new-data-visualization-tool-for-pc-and-surface-hub/.

Vyper. Vyper. *Vyper.* [Online] Vyper. https://vyper.readthedocs.io/en/stable/.

Wallace, Dr. Richard S. 2003. *The Elements of AIML Style.* s.l. : ALICE A. I. Foundation, Inc, 2003.

Waters, Richard. 2016. 'Ether' brought to earth by theft of $50m in cryptocurrency. *Financial Times.* [Online] Financial Times, Jun 18, 2016. https://www.ft.com/content/591518a0-34df-11e6-ad39-3fee5ffe5b5b.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Weakly-supervised Contextualization of Knowledge Graph Facts. Nikos Voskarides, Edgar Meij, Ridho Reinanda, Abhinav Khaitan, Miles Osborne, Giorgio Stefanoni, Prabhanjan Kambadur, and Maarten de Rijke. 2018. Ann Arbor, MI, USA : Association for Computing Machinery, 2018. ISBN 978-1-4503-5657-2/18/07.

Wei, Hao. 2020. Combining knowledge graphs, quickly and accurately. *Amazon Science.* [Online] Amazon.com, May 19, 2020. https://www.amazon.science/blog/combining-knowledge-graphs-quickly-and-accurately.

Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller. 2017. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.* 2017.

Yildiz, Dr Mehmet. 2019. *Architecting Big Data Solutions Integrated with IoT & Cloud: Create strategic business insights with agility.* s.l. : Independently published, 2019. 1089121164.

Yonatan Sompolinsky, Aviv Zohar. 2015. *Secure High-Rate Transaction Processing in Bitcoin.* s.l. : FC'15, 2015.

Yonatan Sompolinsky, Shai Wyborski, Aviv Zohar. 2020. PHANTOM and GHOSTDAG A Scalable Generalization of Nakamoto Consensus. [Online] Feb 02, 2020. https://eprint.iacr.org/2018/104.pdf.

Yonatan Sompolinsky, Yoad Lewenberg, and Aviv Zohar. 2018. SPECTRE: Serialization of Proof-of-work Events: Confirming Transactions via Recursive Elections. [Online] School of Engineering and Computer Science, The Hebrew University of Jerusalem, Israel, 2018. https://eprint.iacr.org/2016/1159.pdf.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov. 2019. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.* s.l. : Google AI, 2019.

Zilliqa. 2017. The Zilliqa Design Story Piece by Piece: Part 1 (Network Sharding). *Zilliqa Blog.* [Online] Medium.com, Oct 16, 2017. https://blog.zilliqa.com/https-blog-zilliqa-com-the-zilliqa-design-story-piece-by-piece-part1-d9cb32ea1e65.

Zilliqa.com. Zilliqa.com. *Zilliqa.com.* [Online] Zilliqa.com. https://www.zilliqa.com/.

References

1. Porter, Michael E. "Competitive advantage: creating and sustaining superior performance. 1985." *New York: FreePress* 43 (1985): 214.

2. BCG. “Conquering Complexity in Supply Chains with Digital Twins”. Online (2020): https://image-src.bcg.com/Images/BCG-Conquering-Complexity-in-Supply-Chains-with-Digital-Twins-Jan-2020_tcm50-237911.pdf Accessed on 01.02.2021

3. Ali, Imran: “Managing Disruptions: The Digital Twin For Supply Chain”. Online (2020). https://www.thefuturefactory.com/blog/59 Accessed on 05.02.2021

4. Gawer, A.: Bridging differing perspectives on technological platforms: Toward an integrative framework. Res. Policy. 43, 1239–1249 (2014).

5. Xu, Yueqiang, Päivärinta, Tero and Kuvaja, Pasi. Digital Twins as Software and Service Development Ecosystems in Industry 4.0: Towards a Research Agenda. *Commun. Comput. Inf. Sci.* 1210 CCIS, (2020), 51–64. DOI: https://doi.org/10.1007/978-981-15-7530-3_5

6. Chesbrough, Henry and Vanhaverbeke, Wim. Open Innovation: A New Paradigm for Understanding Industrial Innovation. *Open Innov.* 4, (2005), 1–27. DOI: https://doi.org/citeulike-article-id:5207447

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

7. Schulte, Stefan, Schuller, Dieter, Steinmetz, Ralf and Abels, Sven. Plug-and-play virtual factories. *IEEE Internet Comput.* 16, 5 (2012), 78–82. DOI: https://doi.org/10.1109/MIC.2012.114

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

References for section 10.3

1. Boyes, H., Hallaq, B., Cunningham, J., & Watson, T. (2018). The industrial internet of things (IIoT): An analysis framework. *Computers in industry*, *101*, 1-12.
2. Hahn, A. (2016). Operational technology and information technology in industrial control systems. In *Cyber-security of SCADA and other industrial control systems* (pp. 51-68). Springer, Cham.
3. Stevens, C. (2020). Assembling cybersecurity: The politics and materiality of technical malware reports and the case of Stuxnet. *Contemporary Security Policy*, *41*(1), 129-152.
4. Lu, G., Feng, D., & Huang, B. (2020). Hidden Markov Model-Based Attack Detection for Networked Control Systems Subject to Random Packet Dropouts. *IEEE Transactions on Industrial Electronics*, *68*(1), 642-653.
5. Klebanov, L. R., & Polubinskaya, S. V. (2020). COMPUTER TECHNOLOGIES FOR COMMITTING SABOTAGE AND TERRORISM. *RUDN Journal of Law*, *24*(3), 717-734.
6. Piggin, R. (2014). Industrial systems: cyber-security's new battlefront [Information Technology Operational Technology]. *Engineering & Technology*, *9*(8), 70-74.
7. Pliatsios, D., Sarigiannidis, P., Lagkas, T., & Sarigiannidis, A. G. (2020). A survey on SCADA systems: secure protocols, incidents, threats and tactics. *IEEE Communications Surveys & Tutorials*, *22*(3), 1942-1976.
8. Yu, X., & Guo, H. (2019, August). A survey on IIoT security. In *2019 IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS)* (pp. 1-5). IEEE.
9. Bajramovic, E., Gupta, D., Guo, Y., Waedt, K., & Bajramovic, A. (2019). Security Challenges and Best Practices for IIoT. In *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik–Informatik für Gesellschaft (Workshop-Beiträge)*. Gesellschaft für Informatik eV.
10. International Electrotechnical Commission. IEC 62443-4-1: Security for Industrial Automation and Control Systems – Part 4-1: Secure Product Development Lifecycle Requirements, Tech. rep.. Geneva, Switzerland: IEC; 2018. https://webstore.iec. ch/publication/33615
11. Mosteiro-Sanchez, A., Barcelo, M., Astorga, J., & Urbieta, A. (2020). Securing IIoT using Defence-in-Depth: Towards an End-to-End secure Industry 4.0. *Journal of Manufacturing Systems*, *57*, 367-378.
12. Leander, B., Čaušević, A., & Hansson, H. (2019, August). Applicability of the IEC 62443 standard in Industry 4.0/IIoT. In *Proceedings of the 14th International Conference on Availability, Reliability and Security* (pp. 1-8).
13. Zarour, M., Alenezi, M., & Alsarayrah, K. (2020). Software Security Specifications and Design: How Software Engineers and Practitioners Are Mixing Things up. In *Proceedings of the Evaluation and Assessment in Software Engineering* (pp. 451-456).
14. Mc Mahon, C. (2020). In defence of the human factor. *Frontiers in Psychology*, *11*, 1390.
15. Zhao, J., Xiang, Y., & Liu, F. (2020). Design and Implementation of Marine Information Security Early-Warning System Oriented to Security Elements Association Analysis. *Journal of Coastal Research*, *108*(SI), 266-269.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

16. Chen, L. (2017). Continuous delivery: overcoming adoption challenges. *Journal of Systems and Software*, *128*, 72-86.
17. Howard, M., & Lipner, S. (2006). *The security development lifecycle* (Vol. 8). Redmond: Microsoft Press.
18. de Vicente Mohino, J., Bermejo Higuera, J., Bermejo Higuera, J. R., & Sicilia Montalvo, J. A. (2019). The application of a new secure software development life cycle (S-SDLC) with agile methodologies. *Electronics*, *8*(11), 1218.
19. Murray, G., Johnstone, M. N., & Valli, C. (2017). The convergence of IT and OT in critical infrastructure.
20. Paes, R., Mazur, D. C., Venne, B. K., & Ostrzenski, J. (2019). A guide to securing industrial control networks: Integrating IT and OT systems. *IEEE Industry Applications Magazine*, *26*(2), 47-53.
21. Ginter, A. (2019). *Secure operations technology*. Lulu. com.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

12. Annex 1

Extensive reading list on Graph Neural Networks provided by Wu et. Al. (A Comprehensive Survey on Graph Neural Networks, 2021)

12.1 Recurrent Graph Neural Networks (RecGNNs)

[1] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 61–80, 2009.

[2] C. Gallicchio and A. Micheli, "Graph echo state networks," in IJCNN. IEEE, 2010, pp. 1–8.

[3] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," in Proc. of ICLR, 2015.

[4] H. Dai, Z. Kozareva, B. Dai, A. Smola, and L. Song, "Learning steadystates of iterative algorithms over graphs," in Proc. of ICML, 2018, pp. 1114–1122.

12.2 Convolutional Graph Neural Networks (ConvGNNs)

[5] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," in Proc. of ICLR, 2014.

[6] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," arXiv preprint arXiv:1506.05163, 2015.

[7] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in Proc. of NIPS, 2016, pp. 3844–3852.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in Proc. of ICLR, 2017.

[9] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, "Cayleynets: Graph convolutional neural networks with complex rational spectral filters," IEEE Transactions on Signal Processing, vol. 67, no. 1, pp.

97–109, 2017.

[10] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in Proc. of AAAI, 2018, pp. 3546–3553.

[11] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graphbased semi-supervised classification," in WWW, 2018, pp. 499–508

[12] A. Micheli, "Neural network for graphs: A contextual constructiveapproach," IEEE Transactions on Neural Networks, vol. 20, no. 3, pp. 498–511, 2009.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

[13] J. Atwood and D. Towsley, “Diffusion-convolutional neural networks,” in Proc. of NIPS, 2016, pp. 1993–2001.

[14] M. Niepert, M. Ahmed, and K. Kutzkov, “Learning convolutional neural networks for graphs,” in Proc. of ICML, 2016, pp. 2014–2023.

[15] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in Proc. of ICML, 2017, pp. 1263–1272.

[16] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” in Proc. of ICLR, 2017.

[17] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, “Geometric deep learning on graphs and manifolds using mixture model cnns,” in Proc. of CVPR, 2017, pp. 5115–5124.

[18] H. Gao, Z. Wang, and S. Ji, “Large-scale learnable graph convolutional networks,” in Proc. of KDD. ACM, 2018, pp. 1416–1424.

[19] D. V. Tran, A. Sperduti et al., “On filter size in graph convolutional networks,” in SSCI. IEEE, 2018, pp. 1534–1541.

[20] D. Bacciu, F. Errica, and A. Micheli, “Contextual graph markov model: A deep and generative approach to graph processing,” in Proc. of ICML, 2018.

[21] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung, “Gaan: Gated attention networks for learning on large and spatiotemporal graphs,” in Proc. of UAI, 2018.

[22] J. Chen, T. Ma, and C. Xiao, “Fastgcn: fast learning with graph convolutional networks via importance sampling,” in Proc. of ICLR, 2018.

[23] J. Chen, J. Zhu, and L. Song, “Stochastic training of graph convolutional networks with variance reduction,” in Proc. of ICML, 2018, pp. 941–949.

[24] W. Huang, T. Zhang, Y. Rong, and J. Huang, “Adaptive sampling towards fast graph representation learning,” in Proc. of NeurIPS, 2018, pp. 4563–4572.

[25] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, “An end-to-end deep learning architecture for graph classification,” in Proc. of AAAI, 2018.

[26] Q. Li, Z. Han, and X.-M. Wu, “Deeper insights into graph convolutional networks for semi-supervised learning,” in Proc. of AAAI, 2018.

[27] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, “Hierarchical graph representation learning with differentiable pooling,” in Proc. of NeurIPS, 2018, pp. 4801–4811.

[28] Z. Liu, C. Chen, L. Li, J. Zhou, X. Li, and L. Song, “Geniepath: Graph neural networks with adaptive receptive paths,” in Proc. of AAAI, 2019.

[29] P. Velickovi ˇ c, W. Fedus, W. L. Hamilton, P. Li ´ o, Y. Bengio, and R. D. ` Hjelm, “Deep graph infomax,” in Proc. of ICLR, 2019.

[30] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks,” in Proc. of ICLR, 2019.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

[31] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, “Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks,” in Proc. of KDD. ACM, 2019

12.3 Graph Autoencoders (GAEs)

[32] S. Cao, W. Lu, and Q. Xu, “Deep neural networks for learning graph representations,” in Proc. of AAAI, 2016, pp. 1145–1152.

[33] D. Wang, P. Cui, and W. Zhu, “Structural deep network embedding,” in Proc. of KDD. ACM, 2016, pp. 1225–1234.

[34] T. N. Kipf and M. Welling, “Variational graph auto-encoders,” NIPS Workshop on Bayesian Deep Learning, 2016.

[35] S. Pan, R. Hu, G. Long, J. Jiang, L. Yao, and C. Zhang, “Adversarially regularized graph autoencoder for graph embedding.” in Proc. of IJCAI, 2018, pp. 2609–2615.

[36] K. Tu, P. Cui, X. Wang, P. S. Yu, and W. Zhu, “Deep recursive network embedding with regular equivalence,” in Proc. of KDD. ACM, 2018, pp. 2357–2366.

[37] W. Yu, C. Zheng, W. Cheng, C. C. Aggarwal, D. Song, B. Zong, H. Chen, and W. Wang, “Learning deep network representations with adversarially regularized autoencoders,” in Proc. of AAAI. ACM, 2018, pp. 2663–2671.

[38] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia, “Learning deep generative models of graphs,” in Proc. of ICML, 2018.

[39] J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec, “Graphrnn: A deep generative model for graphs,” Proc. of ICML, 2018.

[40] M. Simonovsky and N. Komodakis, “Graphvae: Towards generation of small graphs using variational autoencoders,” in ICANN. Springer, 2018, pp. 412–422.

[41] T. Ma, J. Chen, and C. Xiao, “Constrained generation of semantically valid graphs via regularizing variational autoencoders,” in Proc. Of NeurIPS, 2018, pp. 7110–7121.

[42] N. De Cao and T. Kipf, “MolGAN: An implicit generative model for small molecular graphs,” ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018.

[43] A. Bojchevski, O. Shchur, D. Zugner, and S. G ¨ unnemann, “Netgan: ¨ Generating graphs via random walks,” in Proc. of ICML, 2018.

12.4 Spatial-temporal Graph Neural Networks (STGNNs)

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

[44] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in International Conference on Neural Information Processing. Springer, 2018, pp. 362–373.

[45] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in Proc. of ICLR, 2018.

[46] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in Proc. of CVPR, 2016, pp. 5308–5317.

[47] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in Proc. of IJCAI, 2018, pp. 3634–3640.

[48] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proc. of AAAI, 2018.

[49] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling," in Proc. of IJCAI, 2019.

[50] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatialtemporal graph convolutional networks for traffic flow forecasting," in Proc. of AAAI, 2019.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

13. Annex 2

Summary of recent research in NN based Chatbots, provided by Agarwal et. Al. (Review of State-of-the-Art Design Techniques for Chatbots, 2020) showing recent advances and offering proposed areas for further research as enhancements. This collection of work ranges from 2015-2018.

Technique employed	References	Possible areas of enhancements	Paper Reference
This paper proposes an encoder-decoder framework for conversational modeling. The attention mechanism is applied to the model, and beam search is used for decoding	Neural responding machine for short-text conversation	Only for short text conversation	Shang L, Lu Z, Li H. Neural responding machine for short-text conversation. 2015; 1577–1586.
This paper focuses on generating context-sensitive responses by encoding past information using embedding based model. This work utilizes Recurrent Neural Network Language Model (RLM) architecture and tons of features are added on top of it	A neural network approach to context-sensitive generation of conversational responses	Bag of words model is used which does not take into account the order within context and message	Galley M, Auli M, Brockett C, Ji Y, Mitchell M, Gao J. A neural network approach to context-sensitive generation of conversational responses. arXiv preprint arXiv:1506.06714. 2015.
HRED architecture is used in which RNN encodes the input utterances. These encoded vectors are used by context RNN as context vector. Finally, GRU is used to encode the structure of input utterance seen so far. The decoder, takes as input, the output of context RNN and with the help of beam search, produces output	Building end-to-end dialogue systems using generative hierarchical neural network models	Generic Responses of 'I don't know' are frequent	Serban IV, Sordoni A, Bengio Y, Courville A, Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. 2015,
Maximum Mutual Information(MMI) has been used in place of likelihood of output, as objective function	A diversity-promoting objective function for neural conversation models	Factors such as grounding, persona and intent have not been covered	Gao J. A diversity-promoting objective function for neural conversation models. 2015.
Apart from encoder, decoder LSTM structure, Intention structure is also included	Attention with intention for a neural network conversation model	Intention-specific	Zweig V. Attention with Intention for a Neural Network Conversation Model. 2015; 1–7.
Persona/artificial agent has been introduced by capturing the speaking style and background	A persona-based neural conversation model	Not able to capture mood, emotions at a particular point of time	Li J, Galley M, Brockett C, Spithourakis GP, Gao J, Dolan B. A persona-based neural conversation model. 2016.
Overcomes the problem of chatbots being passive, i.e. computer takes the initiative and introduces new content. When a stalemate is detected using keywords, named entity recognition is applied on previous conversations. Retrieval and ranking based system	StalemateBreaker : a proactive content-introducing approach to automatic human–computer conversation	Retrieval based system	Li X, Mou L, Yan R, Zhang M. StalemateBreaker : a proactive content-introducing approach to automatic human-computer conversation. arXiv preprint arXiv:1604.04358. 1:2845–2851.
VHRED model is introduced which extends the HRED model by augmenting latent variable at the decoder. The training step is done by maximizing variational lower bound on log likelihood	A hierarchical latent variable encoder-decoder model for generating dialogues	Longer utterances generated every time, even when short replies are expected as well as suitable	Serban IV, Sordoni A, Lowe R, Charlin L, Pineau J. A hierarchical latent variable encoder-decoder model for generating dialogues. 2016.
Reinforcement learning has been applied in conjugation with seq2seq model	Deep reinforcement learning for dialogue generation	Rewards are heuristic and hence	Li J, Monroe W, Ritter A, Galley M, Gao J, Jurafsky D. Deep reinforcement learning

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

		does not lead to an ideal conversation	for dialogue generation. 2016.
Incorporated attention in HRED model. IDF term is used in objective function	An attentional neural conversation model with improved specificity	Larger training data required for better results	Peng B, Zweig G. An attentional neural conversation model with improved specificity. arXiv preprint arXiv:1606.01292. 2016.
This paper suggests that more the number of previous conversation turns, the better the response generated	Neural discourse modelling of conversations	Increasing the value of N has a tradeoff on time and resources	Pierre JM, Butler M, Portnoff J, Aguilar L. Neural discourse modelling of conversations. 2016; 5(6):1–8.
Input to decoder goes through two neural networks, one is encoder network,other is CNN(for learning topic distribution)	Neural contextual conversation learning with labeled question-answering Pairs	Perplexity is low for shorter sentences	Xiong K, Cui A, Zhang Z, Li M. Neural contextual conversation learning with labeled question-answering pairs. 2014, 2016.
First, a keyword is chosen using pointwise mutual information,then the reply is generated by going forward and backward using two RNN	Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation	Shorter replies than seq2seq	Mou L, Song Y, Yan R, Li G, Zhang L, Jin Z. Sequence to backward and forward sequences: a content-introducing approach to generative short-text conversation. 2016.
A response is retrieved and fed to generative part of the model. Then the resultant reply is compared with that of retrieved one by reranking them	Two are better than one : an ensemble of retrieval- and generation-based dialog systems	No mention of results on goal-oriented system	Song Y, Yan R, Li X, Zhao D, Zhang M. Two are better than one : an ensemble of retrieval- and generation-based dialog systems. arXiv preprint arXiv:1610.07149. 2016 ; 1:1–11.
Generator-Discriminator model is introduced first time for NLP in this paper. Generator is seq2seq model and discriminator is used to distinguish dialogs whether they are human or machine generated	Adversarial learning for neural dialogue generation	This model does not perform very well if there is less discrepancy between generated and reference sequences	Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D. Adversarial learning for neural dialogue generation. 2017.
Responses are based on contextual history as well as facts from knowledge base (Amazon, Wikipedia) on top of seq2seq model. Beam search is used along and reranking is done on MMI	A knowledge-grounded neural conversation model	Since it deals with facts as a knowledge base, some facts can be irrelevant or contradictory	Ghazvininejad M et al. A knowledge-grounded neural conversation model. 2017.
RNN is used. Utterance is processed in three ways: using bag of words, embedding and entity extraction, are passed to RNN	Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning	Model need to be deployed in live dialog system	Williams JD, Asadi K, Zweig G. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. 2017.
Conversation is represented using dialog context, response utterance and latent variable. Knowledge-guided Conditional Variational AutoEncoder (CVAE) deployed	Learning discourse-level diversity for neural dialog models using conditional variational autoencoders	Various improvements suggested by author like using deep neural network learning powers, and considering linguistic phenomena	Zhao T, Zhao R, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. 2017.
Common-sense knowledge base is integrated with retrieval-based models	Augmenting end-to-end dialog systems with common-sense knowledge	only for retrieval based scenario	Young T, Cambria E, Chaturvedi I, Huang M, Zhou H, Biswas S. Augmenting end-to-end dialog systems with common-sense knowledge. 2017.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.

Utterance-level LSTM is used. Two strategies employed on top of it. One is policy network and another is Reinforcement Learning	End-to-end optimization of task-oriented dialogue model with deep reinforcement learning	Only updating policy network results in lesser performance improvement as compared to Reinforcement learning	Liu B, Tur G, Hakkani-Tur D, Shah P, Heck L. End-to-end optimization of task-oriented dialogue model with deep reinforcement learning. 2017; 1–6.
Seq2seq model is augmented with memory network that help encode personas(information about themselves)	Personalizing dialogue agents: i have a dog, do you have pets too?	Trained on persona-chat, can be done in a way where model learns and gains persona from chat history itself and remember that	Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing dialogue agents: i have a dog, do you have pets too?. 2018.
The degeneration problem of VAEs has been solved using utterance drop regularization	A hierarchical latent structure for variational conversation modeling	Overfitting in case of Cornell Movie Dialog Dataset	Kim G. A hierarchical latent structure for variational conversation modeling. 2018.
This is a VAE based approach with discrete latent variables. Two models suggested, one DI-VAE (Recognition and Generator network), other is DI-VST (Discrete variational skip-thought)	Unsupervised discrete sentence representation learning for interpretable neural dialog generation	Better context based latent actions learning is possible	Zhao T, Lee K, Eskenazi M. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. 2018.
GAN is being trained within the latent variable space. DialogWAE with Gaussian mixture network performs better than previous models for dialog generation	DialogWAE : multimodal response generation with conditional wasserstein auto-encoder [56]	Intra-distinct scores not better because of long responses	Gu X, Cho K, Ha J, Kim S. DialogWAE : multimodal response generation with conditional wasserstein auto-encoder. 2018; 1–10.
Combination of HRED and GAN, along with teacher forcing	Multi-turn dialogue response generation in an adversarial learning framework	No human evaluation	Olabiyi OO, Salimov A, Mueller ET. Multi-turn dialogue response generation in an adversarial learning framework. 2018.
Diversity in responses has been increased using adversarial training. CNN encoder and LSTM decoder is used	Generating informative and diverse conversational responses via adversarial information maximization	Distributional discrepancy between ground-truth responses and responses generated has not yet been covered	Zhang Y, Gan Z, Brockett C. Generating informative and diverse conversational responses via adversarial information maximization. Nips. 2018.
The encoder part of the transformer has been used in the model. Masking of words procedure has been used, where it can mask existing words or replacing them with random words	BERT: pre-training of deep bidirectional transformers for language understanding	Linguistic phenomena still to be captured	Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2018.

This document and the information contained are the property of the OXILATE Consortium and shall not be copied in any form or disclosed to any party outside the Consortium without the written permission of the Project Coordination Committee, as regulated by the OXILATE Consortium Agreement and the AENEAS Articles of Association and Internal Regulations.