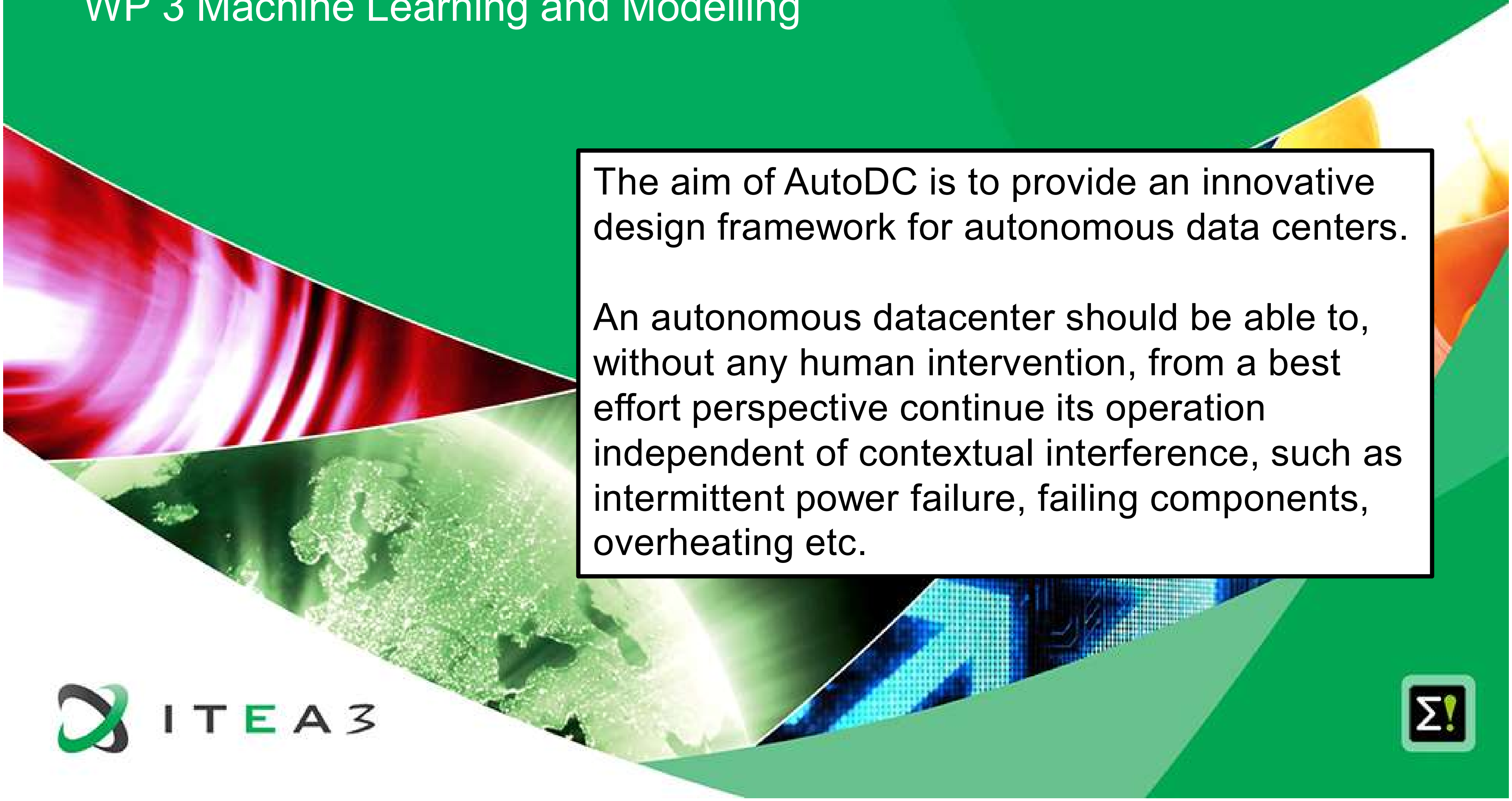
AutoDC - Autonomous “Smart” Datacenters for long term deployment

WP 3 Machine Learning and Modelling



The aim of AutoDC is to provide an innovative design framework for autonomous data centers.

An autonomous datacenter should be able to, without any human intervention, from a best effort perspective continue its operation independent of contextual interference, such as intermittent power failure, failing components, overheating etc.

WP3 – Machine Learning and Modelling

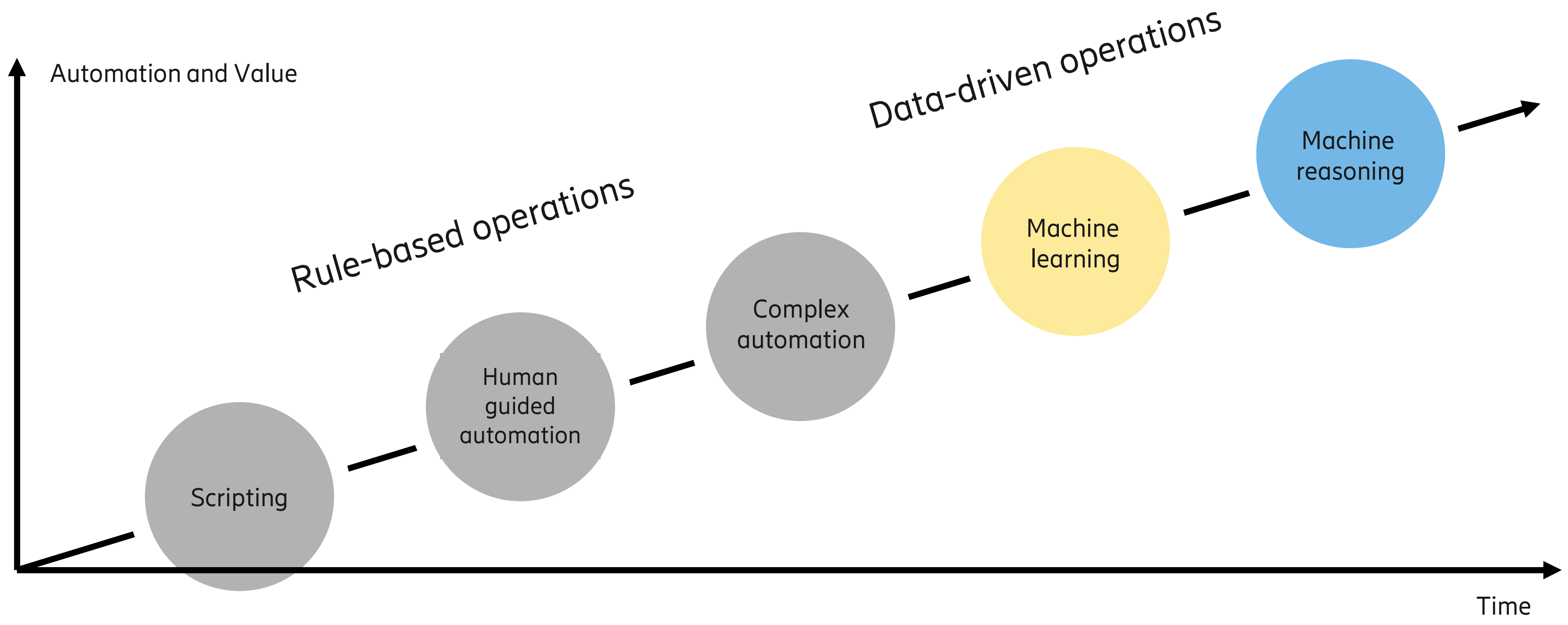
- Challenges
  - Complexity of data centers and operations is increasing
  - Massive growth in data volume, both data-in-flight and data-at-rest
  - Human management is time consuming, expensive, and not fast enough

- Work package innovates on machine learning and model-based technologies to support **automation, management,** and **sustainability** of DCs

- Leveraging data from all components in a DC

Towards autonomous data centers

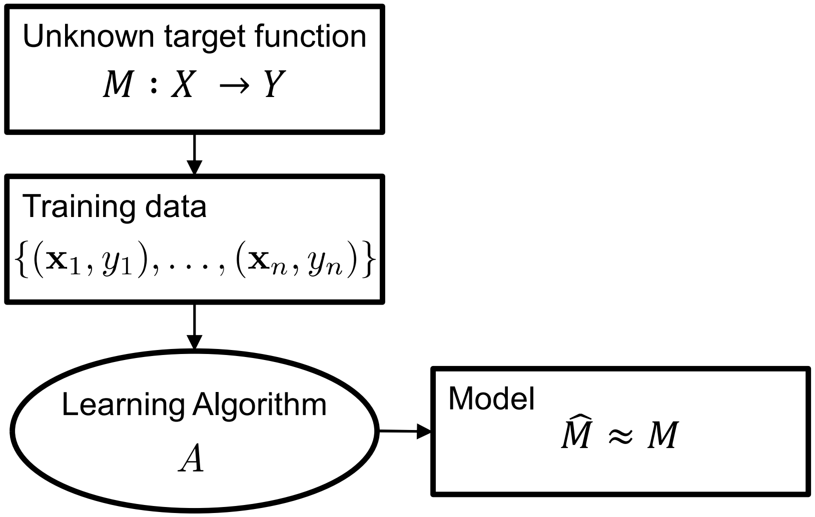
PROJECT IDEA NUMBER: 37



Machine Learning in a nutshell

- What
  - Algorithms that train model from data to make predictions or decisions
  - Rather than following only explicitly programmed instructions
  - Data + Algorithm ➔ Model

- Properties
  - Can learn complex relationships in vast amounts of data
  - Adaptive to changing environments
  - Technology enabler for automation



Unknown target function $M : X \rightarrow Y$

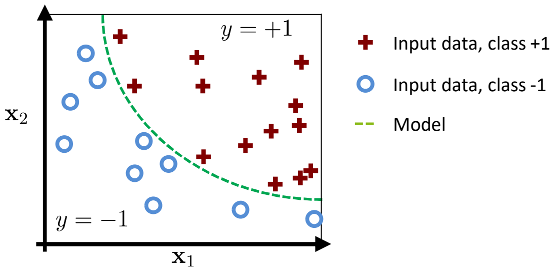
Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

Learning Algorithm A

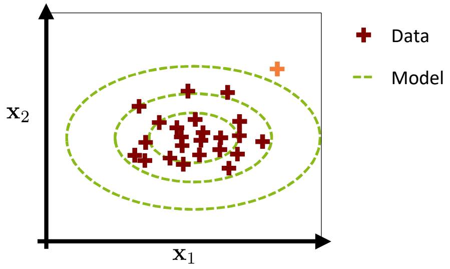
Model $\widehat{M} \approx M$

Machine Learning in a nutshell

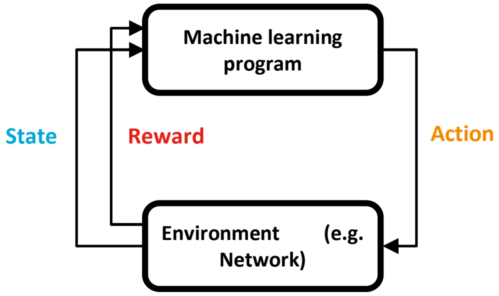
- Supervised learning
  - Given input, output pairs, learn to predict for new data
  - Tasks: Classification, Regression, Ranking
  - Examples: KPI prediction, energy consumption, ...



- Unsupervised learning
  - Given only inputs, find structure
  - Tasks: Clustering, Anomaly Detection
  - Examples: Detect abnormal DC events, grouping of alarms and errors, ...



- Reinforcement learning
  - Receive feedback in the form of rewards
  - Task: Actions in environment to maximize cumulative (future) reward
  - Examples: Play GO, control of DC infrastructure, antenna tilting in 5G



WP3 – Work areas

- ML models enabling autonomous DCs
  - Energy efficiency
  - Anomaly detection
  - Performance optimization
  - Resource optimization
  - Root-cause analysis/troubleshooting
  - Strategic business support
  - Prolonged lifetime

- Nonstationary learning and scalability
  - Techniques for coping with ML in dynamic environments
  - Scalability and overhead

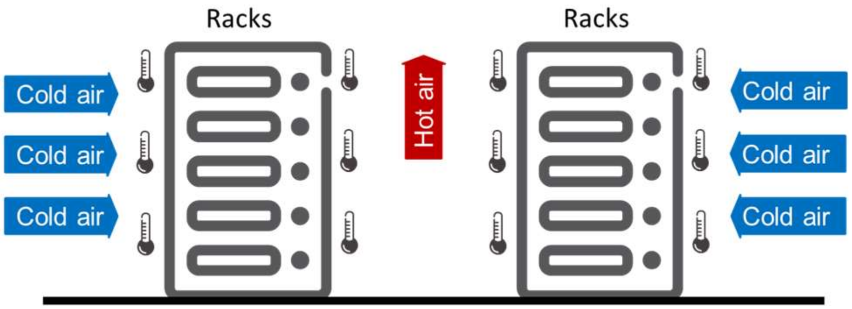
- Integration of ML and model-based techniques

Simulation models for EDGE DC cooling management

- Towards energy efficient autonomous DCs

- Predicting DC temperatures using ML
  - Data-driven model using LSTM neural network
  - Implemented using Keras/TensorFlow
  - The model was trained with 90 hours of data from the edge DC at RISE I.C.E.
  - The model was tested by using 1 hour of data to initialize the LSTM network, and then predicting 9 hours into the future, while changing the setpoints hourly.

- Model is an enabler for improved control strategies and fault detection



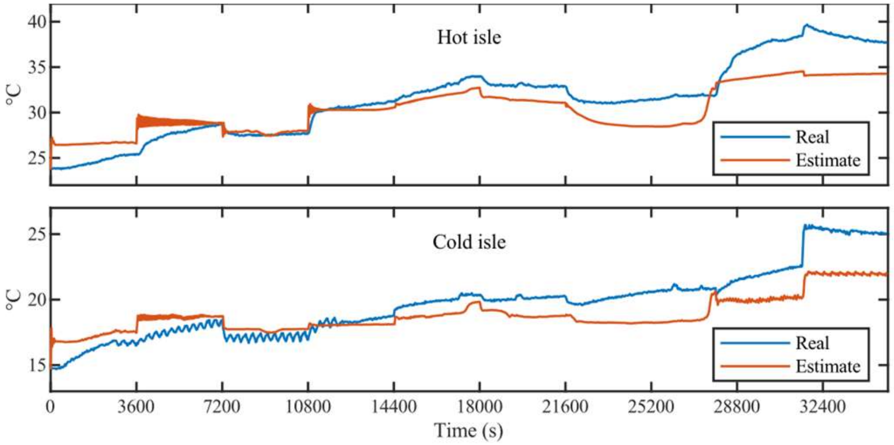


Figure: Hot and cold isle temperature predictions from the simulation model test.

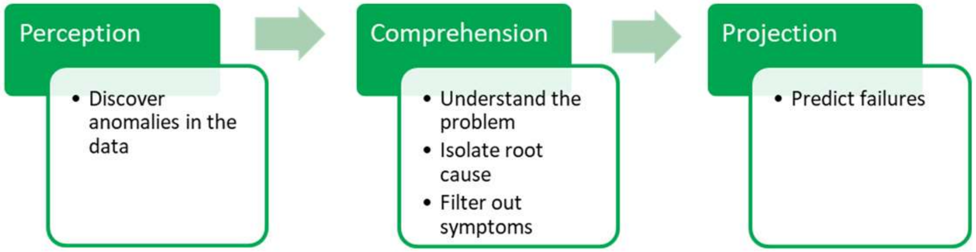
MAE = 1.84 on hot isle
MAE = 1.44 on the cold isle

Anomaly detection and root-cause analysis

PROJECT IDEA NUMBER: 37

- Anomaly detection and RCA is an enabler for autonomous, self-driving DC operations

- The process



Perception

- Discover anomalies in the data

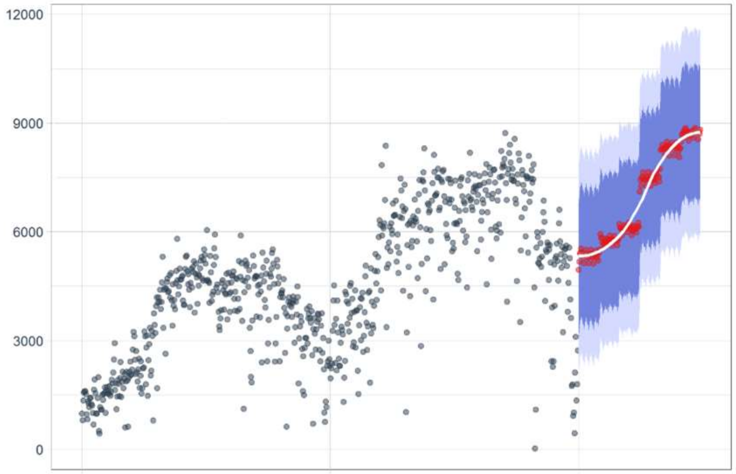
Comprehension

- Understand the problem
- Isolate root cause
- Filter out symptoms

Projection

- Predict failures

- Ongoing work
  - Evaluation of anomaly detection models based on ML for streaming event-based DC data
  - Investigation of methods for determining causal relationships among nodes in multi-dimensional graphs



Example:

- Model data behaviour
- Use model to forecast next value
- Detect if new observed values represent anomalies

Transfer learning for coping with dynamic DC environments

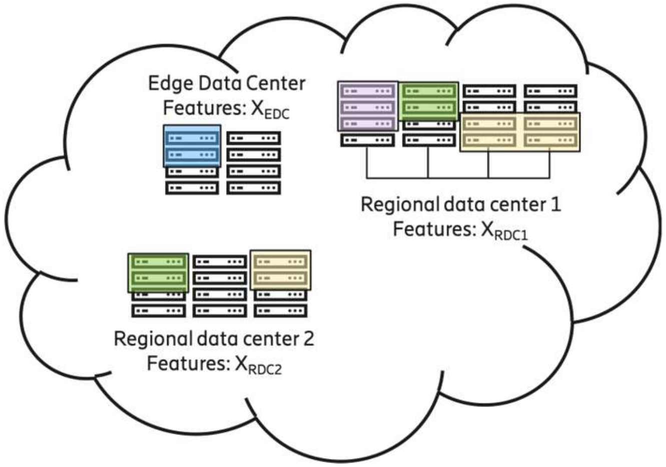
PROJECT IDEA NUMBER: 37

- Dynamic DC environments can reduce ML performance

- Challenges
  - Service containers are distributed, short-lived, and dynamically migrated and scaled
  - DC infrastructure may change over its lifetime
  - Sizeable changes in the execution environment reduces ML performance
  - Not enough time and samples to learn new model for services in transient execution environments

- Approach
  - Reuse model knowledge learned in similar systems
  - Transfer learning to improve ML model performance

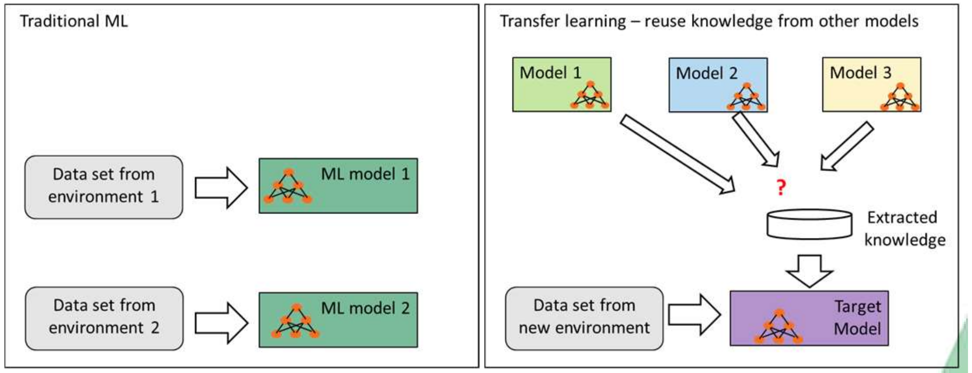
- Ongoing work
  - Models for service performance in DCs
  - How to select a good source model



Edge Data Center
Features: X_{EDC}

Regional data center 1
Features: X_{RDC1}

Regional data center 2
Features: X_{RDC2}



Traditional ML

Data set from environment 1 → ML model 1

Data set from environment 2 → ML model 2

Transfer learning – reuse knowledge from other models

Model 1

Model 2

Model 3

?

Extracted knowledge

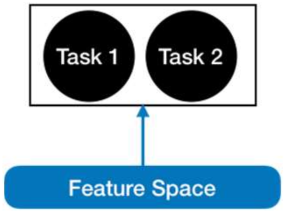
Data set from new environment → Target Model

Coping with large feature spaces and massive data

PROJECT IDEA NUMBER: 37

Scalability challenge

- Impractical to collect and maintain all data features in a DC
- Large data storage
- Cost associated to data transfer, network capacity and energy

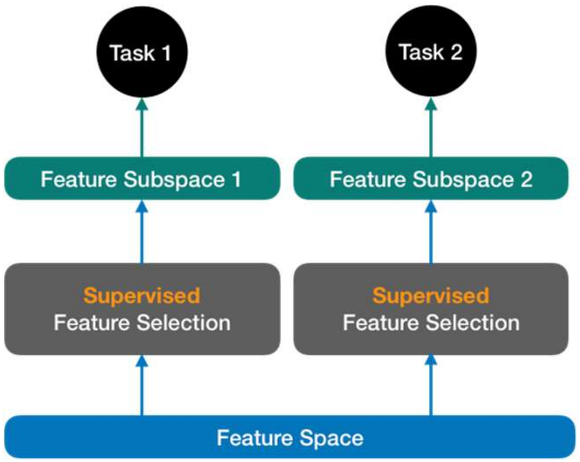

Task 1
Task 2
Feature Space

Feature Selection for improved ML:

- Reduces data dimensionality
- Maintains interpretability (e.g., root-cause analysis)

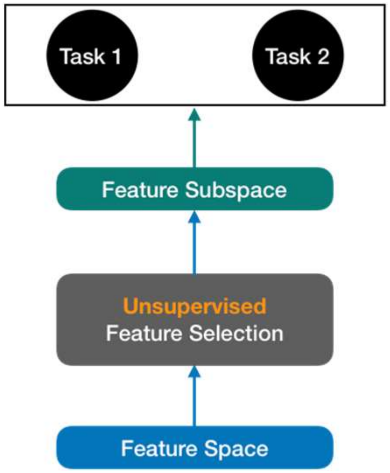
Supervised **Feature Selection**:

- Use-case (task) specific


Task 1
Task 2
Feature Subspace 1
Feature Subspace 2
Supervised Feature Selection
Supervised Feature Selection
Feature Space

Unsupervised **Feature Selection**:

- Use-case (task) specific


Task 1
Task 2
Feature Subspace
Unsupervised Feature Selection
Feature Space

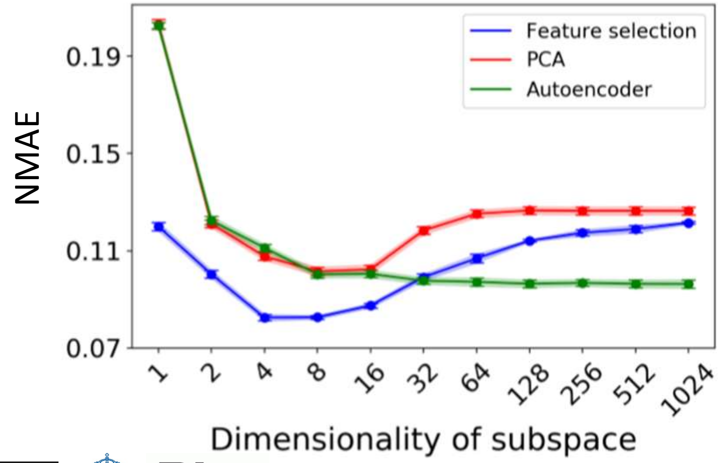
Coping with large feature spaces and massive data

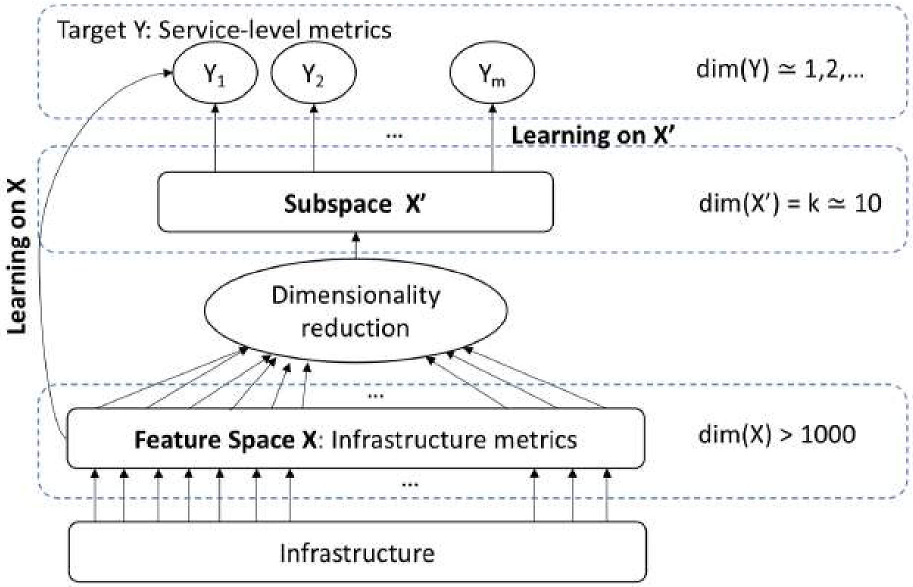
PROJECT IDEA NUMBER: 37

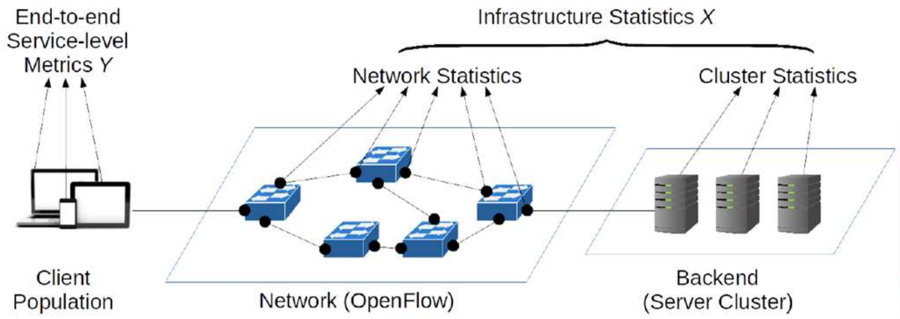
- Models for predicting service-level metrics Y from infrastructure metrics X

- Approach for dimensionality reduction
  - Feature selection
  - Principle component analysis (PCA)
  - Autoencoders

- Results







Exploitation Related Achievements

- Utility (patent)
  - Methods and systems for dynamic service performance prediction using transfer learning, PCT/SE2019/050672
  - Improving performance modeling in dynamic clouds, US 62/770,330

- Academic exam (human capital)
  - Service Metric Prediction in Clouds using Transfer Learning, MSc thesis report, 2019

- Publications (dissemination)
  - Efficient Learning on High-dimensional Operational Data, International Conference on Network and Service Management (CNSM), 2019
  - Digital Twin for Tuning of Server Fan Controllers, IEEE International Conference on Industrial Informatics (INDNI), 2019
  - Performance Prediction in Dynamic Clouds using Transfer Learning, IFIP/IEEE Integrated Network Management (IM), 2019 **(best paper award)**
  - Predicting Distributions of Service Metrics, IFIP/IEEE Integrated Network Management (IM), 2019

Exploitation Related Achievements, cont.

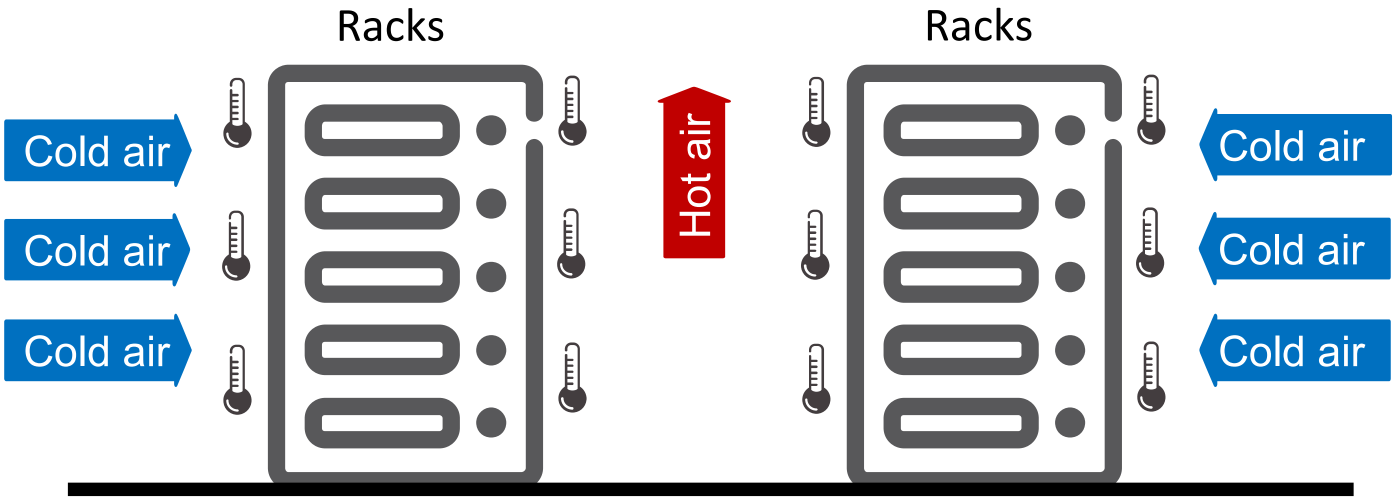
- Collaboration (exploitation)
  - Established collaboration RISE North, Luleå University of Technology, and Ericsson

- Internal (dissemination)
  - Seminar on CDE using MDNs, 2019
  - Seminar on transfer learning, 2019

- Conference (dissemination)
  - Introduction to data-driven engineering of networked systems, Keynote at BlackSeaCom, 2019

Next steps

- Continue ongoing activities
- Use case and data-driven research
  - Define new use cases for automation of DCs
  - Understand data availability


Racks
Racks
Cold air
Cold air
Cold air
Hot air
Cold air
Cold air
Cold air

Presentation structure

PROJECT IDEA NUMBER: 37

- WP3 objectives, what we want to achieve
  - 1
- WP3 structure, i.e. tasks and what they mean, how they contribute to autonomous DCs
  - 2
- Selected description of ERAs
  - 4
- Plans
  - 1
- ERAs
  - 2

Ongoing and planned work in T3.3: Use-case Self-driving Systems

- Goal
  - ***System dynamically configures*** to meet management objectives,
    e.g.: minimize energy usage while 95 percentile of reads must be below 20ms
  - ***System dynamically adjusts*** configuration parameters to changes in environment,
    e.g., dynamic load pattern, query profile changes.

- Approach
  - Learning ***regression models*** for service-level KPIs prediction and forecasting (prior work)
  - ***Reinforcement learning*** to control resource allocation, horizonatal and vertical scaling
  - Achieving scalability though dimensionality reduction techniques

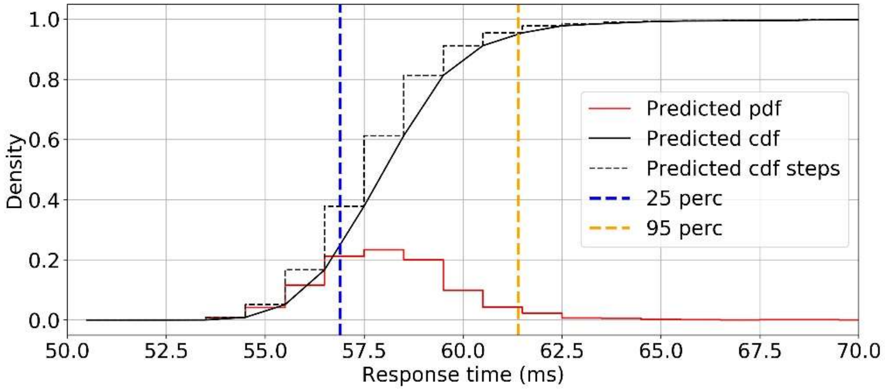
- Research challenges
  - Efficient, ***low-overhead learning and control***: can prediction and reinforcement learning be integrated?
  - How can ***models be reused*** to speed up system configuration and control for new environments (Transfer learning)?

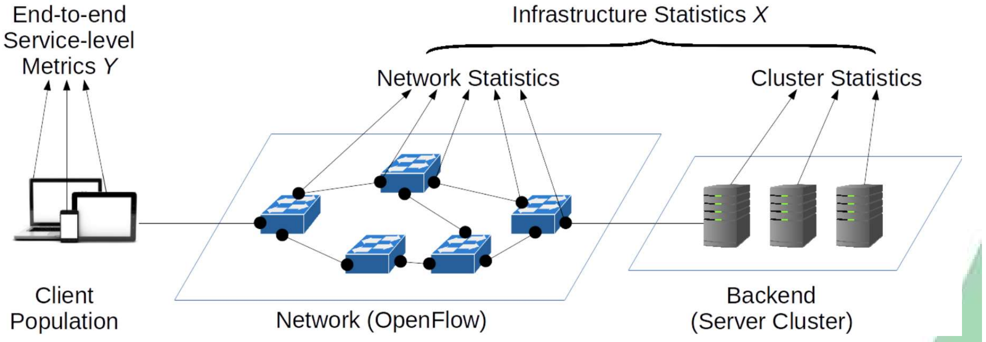
- Concepts are developed and evaluated on new experimental platform

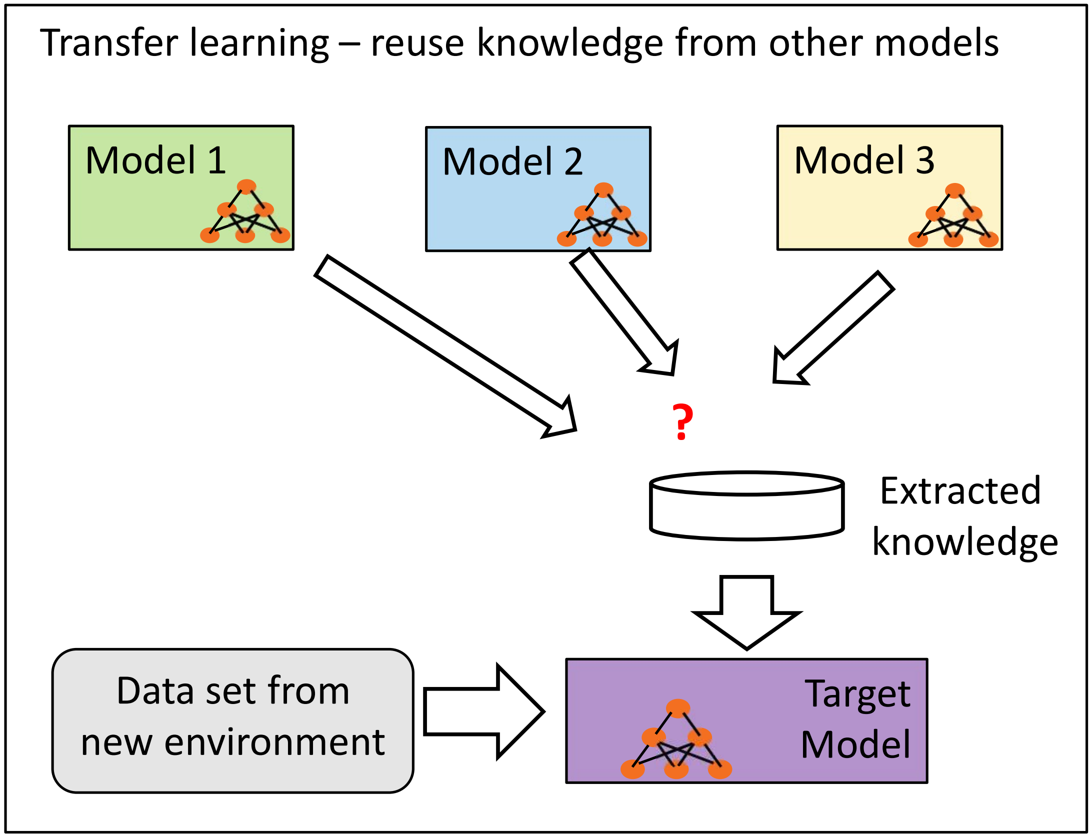
Could be reported, but I think we can focus on other things:

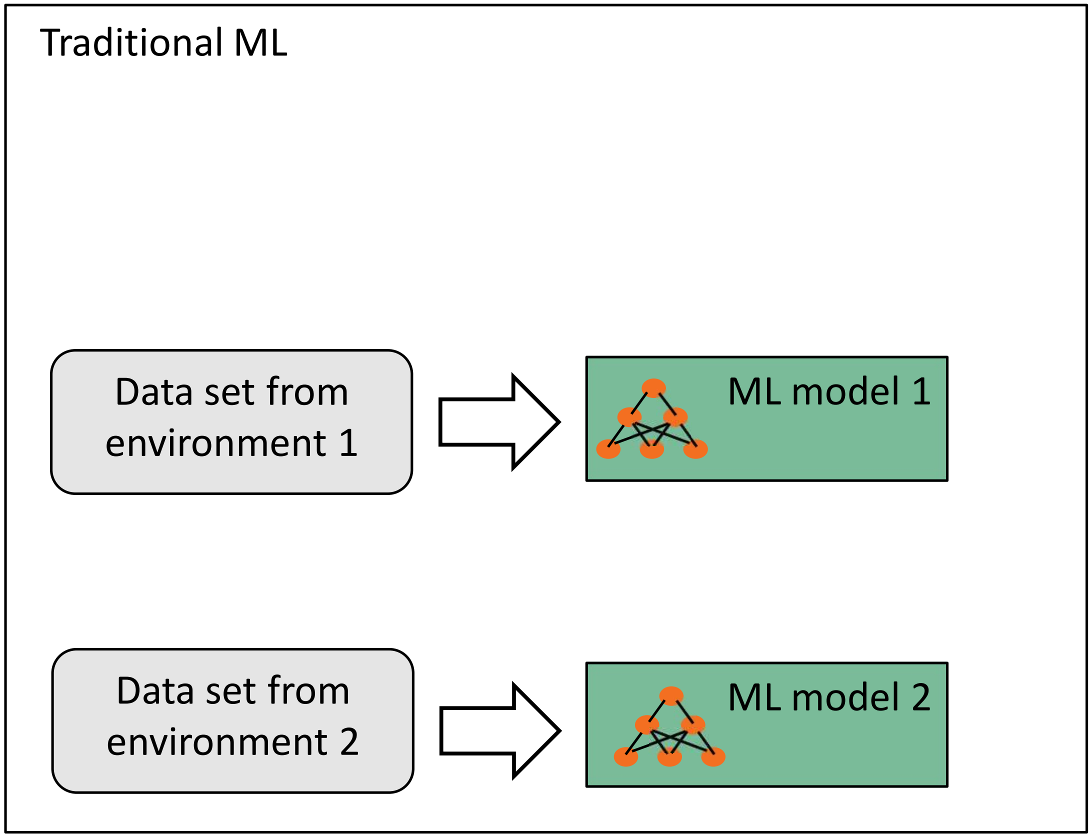
Work in T3.3: Conditional density estimation

- Predicting service KPIs is key for predicting e.g. SLA conformance, failure probabilities, ...
- Machine learning is powerful tool for prediction
- Today most works use point predictions
- Predicting distributions gives a more complete description
- Developed two solutions
  - Mixture density networks with kernels
  - Discretized conditional density using histogram estimators






Transfer learning – reuse knowledge from other models


Traditional ML
Data set from environment 1
ML model 1
Data set from environment 2
ML model 2

AutoDC WP3 tasks

PROJECT IDEA NUMBER: 37

T3.1 – Novel DCM and PDSS solutions

- ML models for strategic business support

T3.2 – Predictive and prescriptive modelling

- ML models for autonomous DC operations
- Energy efficiency, anomaly detection, performance optimization, bottleneck mitigation, ...

T3.3 – Nonstationary learning and scalability

- ML technology for practical DC deployment
  - Techniques for coping with ML in dynamic environments
  - Scalability and overhead

T3.4 – Integration of ML and model-based techniques

- Control theoretical techniques to overcome challenges related to e.g. lack of data