Elaborare i dati raccolti negli ambienti di produzione per creare valore e strutturare conoscenza: *metodologie data-driven, sfide e opportunità*

Tania CERQUITELLI
Department of Control and Computer engineering, Politecnico di Torino, Italy

1

Outline

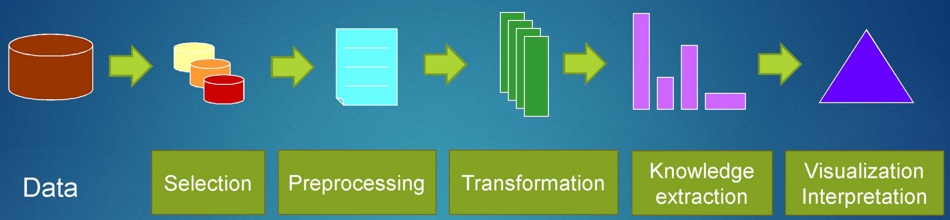
- Knowledge discovery process
- Predictive maintenance in Industry 4.0
- Self-tuning methodologies for predictive maintenance
- Semi-supervised data labelling
- Concept drift management
- Technological aspects
- Open issues

2

Knowledge discovery process

3

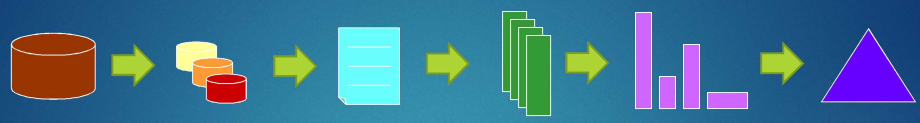
Knowledge extraction process



4

Elaborare i dati raccolti negli ambienti di produzione per creare valore e strutturare conoscenza

Knowledge extraction process



Data → Selection → Preprocessing → Transformation → Knowledge extraction → Visualization Interpretation

Knowledge

↓

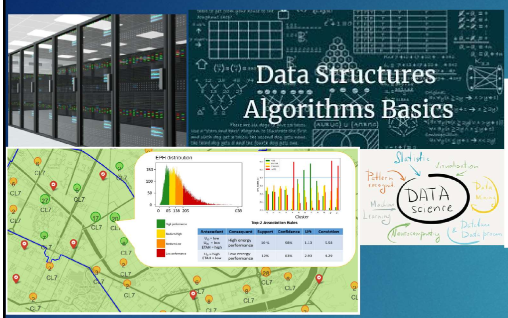
Value

KDD

***K**nowledge **D**iscovery from **D**ata*

5

KDD from Industrial data: two key roles



DATA SCIENTIST

- Design **innovative and efficient algorithms**
- Select the **optimal techniques** to address the challenges of the analysis
- Identify the best **trade-off** between knowledge quality and execution time

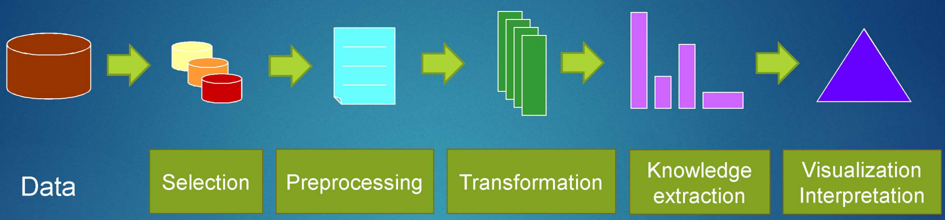
DOMAIN EXPERT

- Support **the data pre-processing** phase
- **Assess** extracted **knowledge**
- Strong involvement in the algorithm definition phase which should **respect/include physical laws** and correctly **model physical events**

6

Elaborare i dati raccolti negli ambienti di produzione per creare valore e strutturare conoscenza

Knowledge extraction process: Innovations

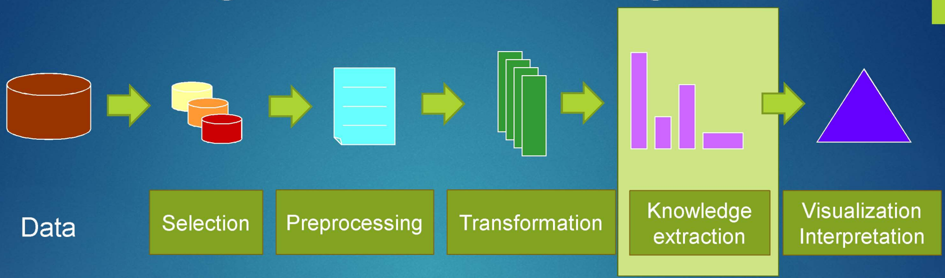


Innovations in the data analytics process

- **Tailor** the **analytic** steps to the different key aspects of **industrial data**
- **Design ad-hoc data transformation strategies to capture different facets of data**
- **Automate** the data analytic workflow to **reduce the manual user interventions**
- Translate the domain-expert knowledge into **automated procedures**
- Design **informative dashboards** to support the translation of the extracted knowledge into effective actions

7

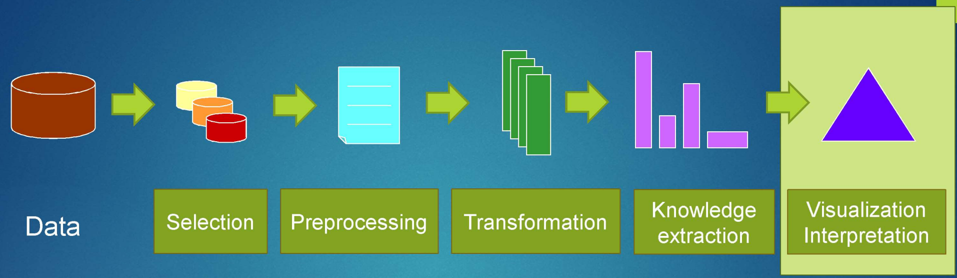
Knowledge extraction algorithms



- **Predictive algorithms**
  - **Predictive maintenance**
- **Clustering algorithms**
  - **Data labeling to support predictive analytics**
- **Anomaly detection techniques**

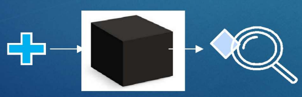
8

Knowledge visualization and interpretation



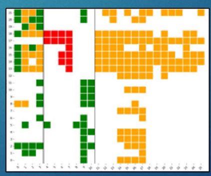
Interpretation of models and results

- Identification of algorithms and methodologies for semantic transparency of Machine Learning models
- Explanation methods for explaining individual predictions of black box models



Informative dashboard

- Visualization methodologies



9

Predictive maintenance in Industry 4.0

10

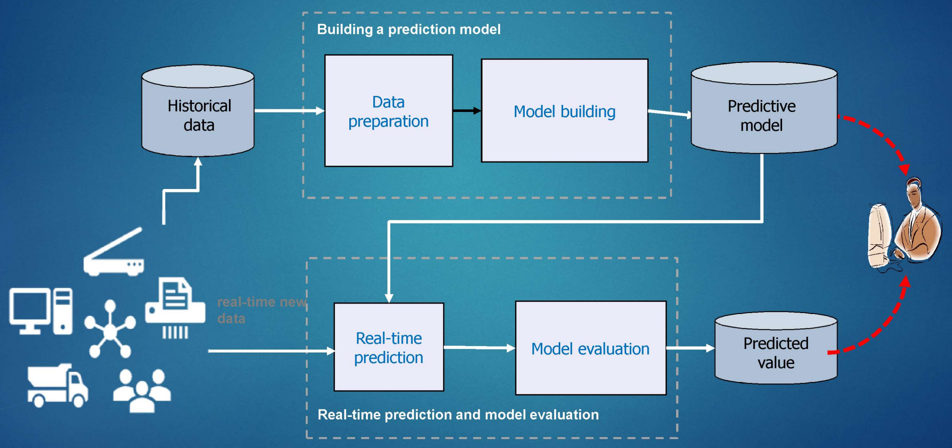
Predictive maintenance in Industry 4.0

- **Intelligent techniques** to identify **symptoms of imminent machine failure** before their actual occurrence
  - It combines **physical models** of complex devices (machines, robots, conveyors, etc.) together with **data-driven algorithms** to effectively support smart predictive diagnostics (prognostics)
- It anticipates failures and estimates the **Remaining Useful Lifetime** (RUL)
  - It exploits innovative **analytics methods** to forecast the **future evolution of machine degradation**
- **On-line data** collected in the factory characterizing the current dynamics of the process/machine from the

- Some of the **most common needs of manufacturing enterprises**
  - compatibility with both the on-premises and the in-the-cloud environments
  - exploitation of reliable and largely supported Big Data platforms
  - easy deployment through containerized software modules
  - virtually unlimited horizontal scalability
  - fault-tolerant self-reconfiguration

11

Predictive maintenance approach



12

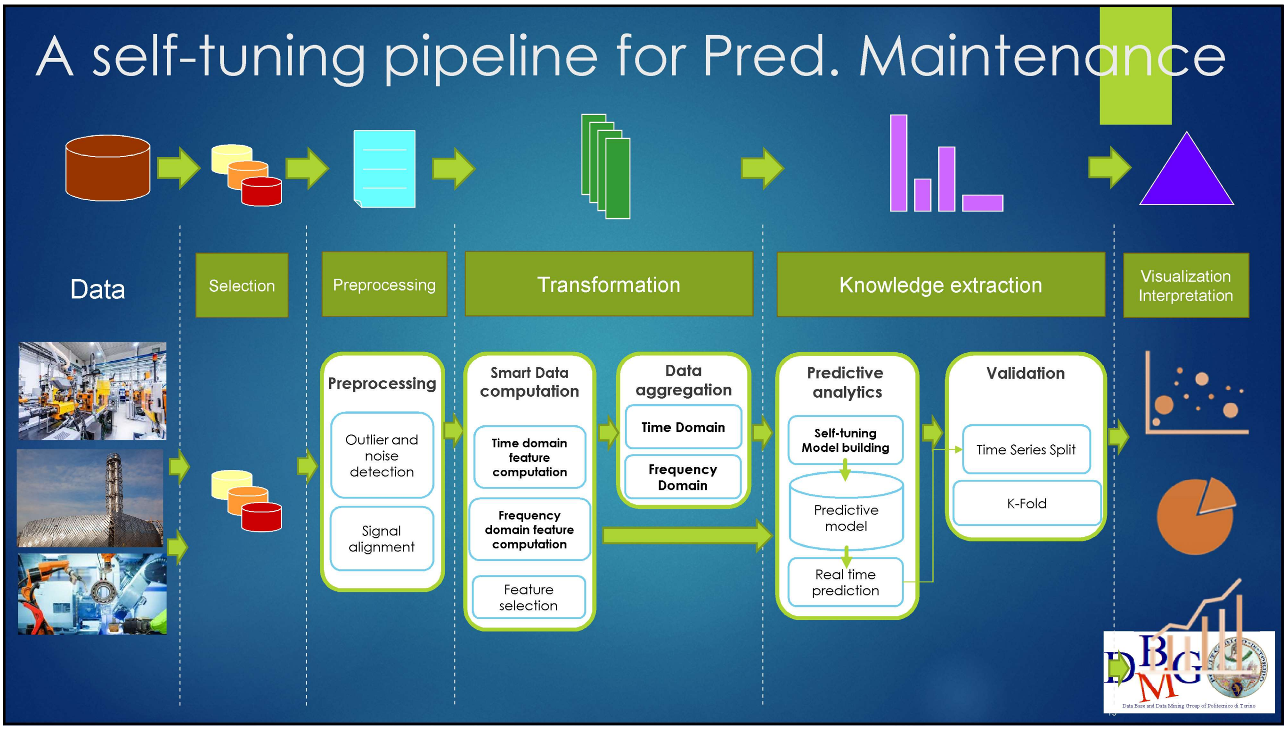
Main issues

- **Tailoring the KDD process to predictive maintenance requires a lot of expertise**
- **Identifying the best data preprocessing approach is very challenging**
  - Summarize long time-dependent series through ad-doc statistics capturing the main key features for the prediction
  - Use-case dependent
- **A variety of state-of-the-art algorithms is available**
  - Data driven methodologies possibly enriched with physical models
- **Each algorithm is characterized by many different input-parameters**

13

Self-tuning methodologies for predictive maintenance

14



15

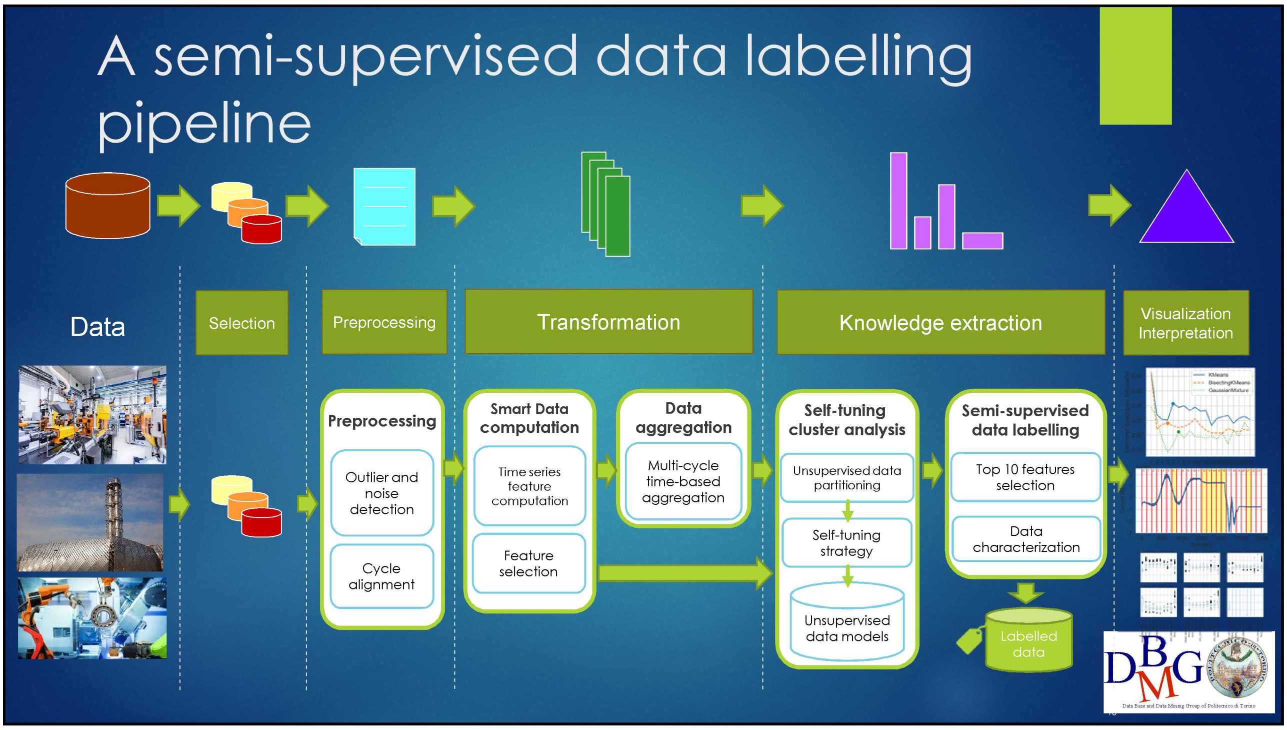
Key components

- **Ad-hoc data preparation and data aggregation techniques tailored to predictive maintenance issues**
- **Self-tuning strategy to offload the data scientist from**
  - setting the specific algorithm parameters
  - selecting the best algorithm
- **Transparent mining (i.e., native interpretable models) allows the final user to know why prediction outcomes have been taken**
  - Needed to target specific problems in the production processes and trigger precise corrective actions.
  - Black box models: very accurate models, but do not permit the final user to have a deep understanding of causes of the predictions

16


Semi-supervised data labelling
DBMG

17


A semi-supervised data labelling pipeline
Data
Selection
Preprocessing
Transformation
Knowledge extraction
Visualization Interpretation
Preprocessing
Outlier and noise detection
Cycle alignment
Smart Data computation
Time series feature computation
Feature selection
Data aggregation
Multi-cycle time-based aggregation
Self-tuning cluster analysis
Unsupervised data partitioning
Self-tuning strategy
Unsupervised data models
Semi-supervised data labelling
Top 10 features selection
Data characterization
Labelled data
DBMG

18

Key components

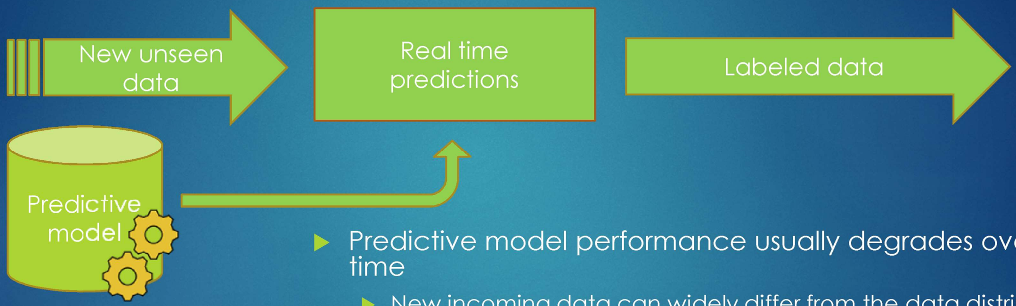
- **Self-tuning strategy to offload the data scientist from**
  - setting the specific algorithm parameters
  - selecting the best algorithm
- **To help the domain expert to easily define the data labelling to production cycles/machines in the plants**
  - The top-k smart data features to focus only on the most relevant features
  - Boxplot distribution for the top-k features
  - Few representative samples for each cluster are manually inspected
  - Most relevant sub-cycles are highlighted

19

Concept drift management

20

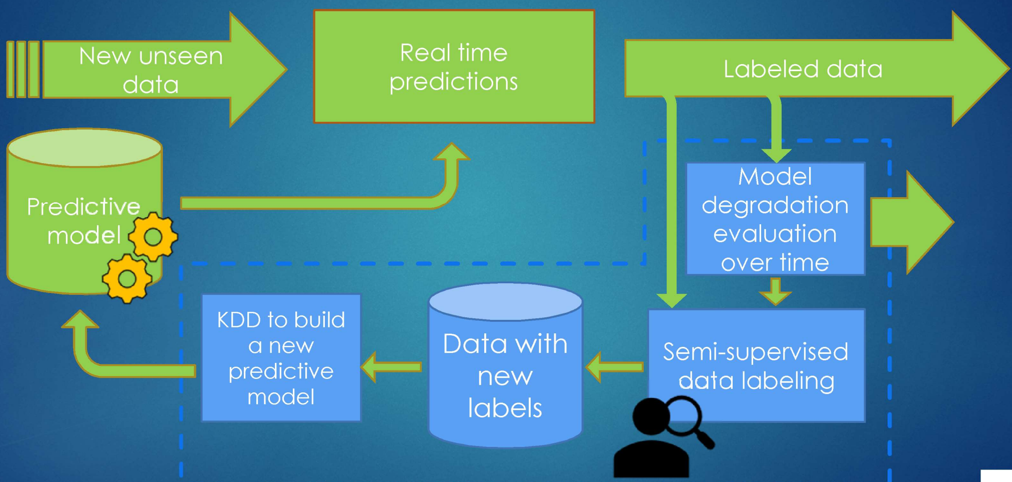
Concept drift management



- Predictive model performance usually degrades over time
  - New incoming data can widely differ from the data distribution on which the model was trained
  - Not all possible classes (labels) are known at training time
  - Real time predictions performed on new unseen data may be misleading or totally wrong

21

Automated concept drift management



22

Key components

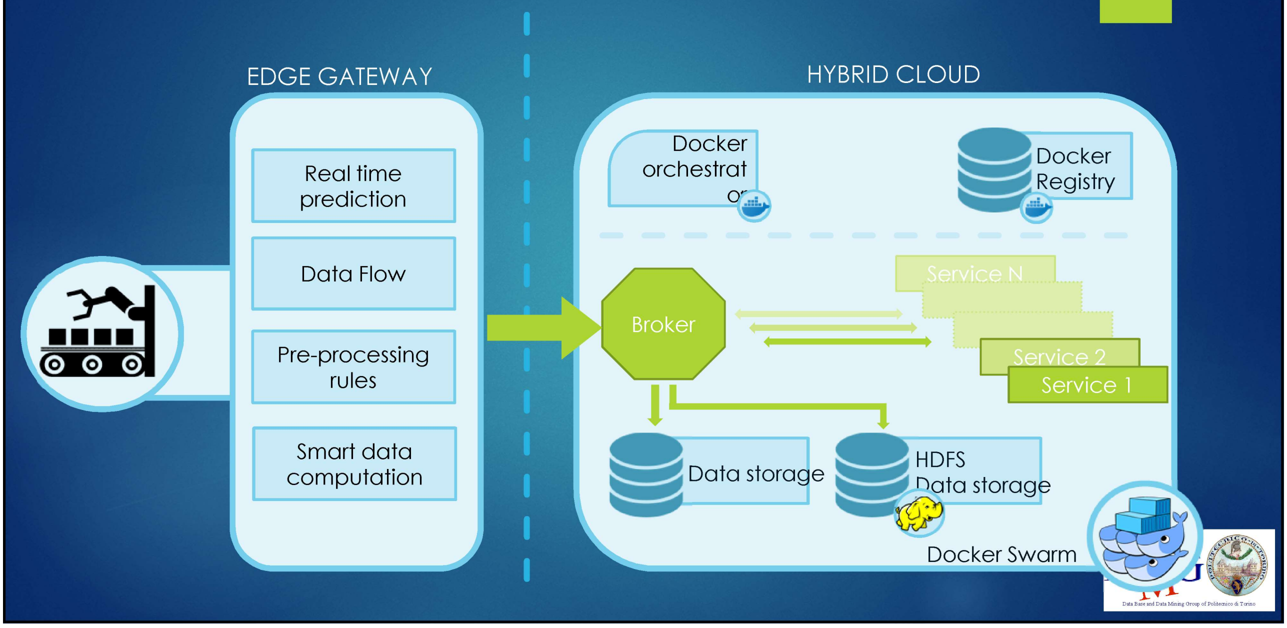
- **Automatic triggering of the predictive model retraining only when necessary**
  - Towards real-time evaluation
- **Unsupervised approach, without the ground-truth labels for the newly classified samples**
  - Different (scalable) quality metrics for drift detection

23

Technological aspects

24

A cloud-to-edge architecture



25

- Fast and general engine for **large-scale data processing**
  - Batch processing
  - Streaming processing
  - 10 to 100 times faster than Hadoop MapReduce
- **Advanced Analytics**
  - Map-Reduce, SparkSQL, MLlib **(machine learning)**
  - Spark **Streaming,** Spark Graphix (graph support)
- **Unified Engine,** Spark can run on top of
  - Hadoop, Cassandra, Amazon S3
  - Mesos, Standalone
- **In-memory data sharing**
  - Good for iterative, interactive and event stream processing tasks.
- Active, expanding user **community**

- Containerization platform
- Microservices architecture deployment
- Faster application development and delivery
- It improves modularity
  - applications easier to understand
- Container isolation makes application portable to any infrastructure
- Open source technology and a modular design
  - easy to integrate into your existing environment

26

Open issues

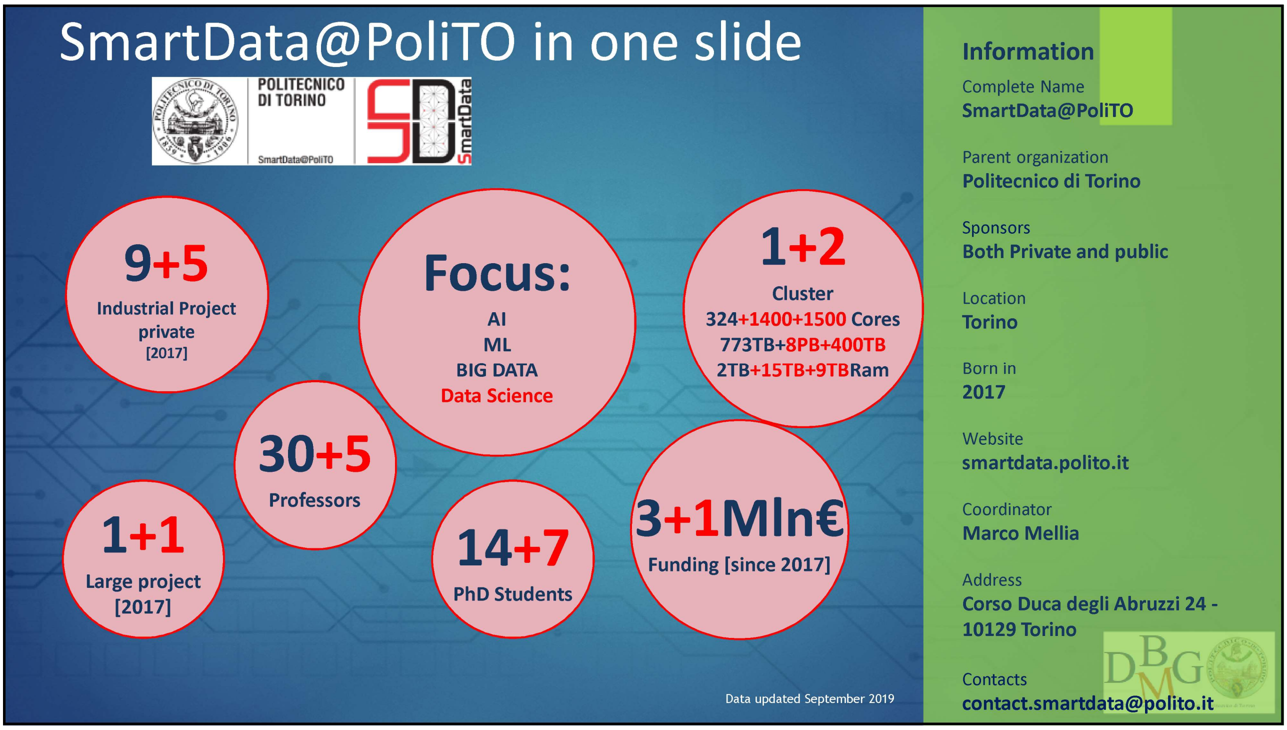
- **The automated predictive analytics pipeline is still a dream**
  - Lack of a **general-purpose** data-driven methodology
  - **Different** data-analytics **solutions** might be required to correctly address **a specific use-case**
  - Specific aspects in the Industry 4.0 context requires specific data-driven solutions
  - Tailored methodologies might be **context-dependent**

27

Publications

- Tania Cerquitelli, Enrico Macii, Alberto Macii **Elaborare i dati raccolti negli ambienti di produzione, creare valore e strutturare conoscenza: metodi, sfide e opportunità**. Il futuro della fabbrica. La via italiana per il rinascimento della manifattura. Editor: ESTE ISBN: 978-88-98053-32-2
- Tania Cerquitelli, Stefano Proto, Francesco Ventura, Daniele Apiletti, Elena Baralis: **Towards a real-time unsupervised estimation of predictive model degradation.** BIRTE 2019: 5:1-5:6
- Daniele Apiletti, Claudia Barberis, Tania Cerquitelli, Alberto Macii, Enrico Macii, Massimo Poncino, Francesco Ventura: **iSTEP, an Integrated Self-Tuning Engine for Predictive Maintenance in Industry 4.0.** In IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, 2018, Melbourne, Australia, December 11-13, 2018. IEEE 2018, ISBN 978-1-7281-1141-4. pp 924-931 – Best paper Award
- Stefano Proto, Francesco Ventura, Daniele Apiletti, Tania Cerquitelli, Elena Baralis, Enrico Macii, Alberto Macii: **PREMISES, a Scalable Data-Driven Service to Predict Alarms in Slowly-Degrading Multi-Cycle Industrial Processes**. 2019 IEEE International Congress on Big Data, BigData Congress 2019, Milan, Italy, July 8-13, 2019. IEEE 2019, ISBN 978-1-7281-2772-9: pp. 139-143
- Francesco Ventura, Stefano Proto, Daniele Apiletti, Tania Cerquitelli, Simone Panicucci, Elena Baralis, Enrico Macii, Alberto Macii: **A New Unsupervised Predictive-Model Self-Assessment Approach that SCALEs**. 2019 IEEE International Congress on Big Data, BigData Congress 2019, Milan, Italy, July 8-13, 2019. IEEE 2019, ISBN 978-1-7281-2772-9: pp. 144-148
- Tania Cerquitelli, David Bowden, Angelo Marguglio, Lucrezia Morabito, Chiara Napione, Simone Panicucci, Nikolaos Nikolakis, Sotiris Makris, Guido Coppo, Salvatore Andolina, Alberto Macii, Enrico Macii, Niamh O'Mahony, Paul Becker, Sven Jung: **A Fog Computing Approach for Predictive Maintenance**. CAiSE Workshops 2019: 139-147

28


SmartData@PoliTO in one slide
POLITECNICO DI TORINO
SmartData@PoliTO
SmartData
9+5
Industrial Project
private
[2017]
Focus:
AI
ML
BIG DATA
Data Science
1+2
Cluster
324+1400+1500 Cores
773TB+8PB+400TB
2TB+15TB+9TBRam
30+5
Professors
1+1
Large project
[2017]
14+7
PhD Students
3+1Mln€
Funding [since 2017]
Data updated September 2019
Information
Complete Name
SmartData@PoliTO
Parent organization
Politecnico di Torino
Sponsors
Both Private and public
Location
Torino
Born in
2017
Website
smartdata.polito.it
Coordinator
Marco Mellia
Address
Corso Duca degli Abruzzi 24 - 10129 Torino
Contacts
contact.smartdata@polito.it
DBMG

29


POLITECNICO DI TORINO
1859 1906
DANKSCHEEN
BIYAN
SHUKRIA
TINGKI
YAQHANYELAY
TASHAKKUR ATU
GRACIAS
SUKSAMA
EKHMET
THANK
YOU
ARIGATO
SHUKURIA
GOZAIMASHITA
EFCHARISTO
KOMAPSUMNIDA
MARKE
GRAZIE
MEHRBANI
PALDIES
JUSPAXAR
BOLZIN
MERCI
... questions?
Tania CERQUITELLI
tania.cerquitelli@polito.it

30