# Theses


Daniele Apiletti, Elena Baralis, Luca Cagliero, Tania Cerquitelli,
Silvia Chiusano, Paolo Garza, Danilo Giordano, Alessandro Fiori

# General information

- Duration: 5-6 months full time
  - equivalent overall duration if part time
- *Internal* thesis
  - cooperation on active research topic or research project
  - good programming and analytical skills required
  - supervised by a group member
  - can work at home or in our lab (LAB5)
- *External* thesis (stage)
  - supervised by external tutor

To get more info on specific topics

please contact the reference person of the thematic area of interest by email (*name dot surname at polito dot it*)

# Main topics

- Data mining and machine learning algorithms
  - Design and implementation of novel ML algorithms
- Data science pipeline
  - Design, personalization and implementation of KDD processes in diverse application areas
    - Industry, health, finance, …
- Big data analytics
  - Design of scalable data mining and machine learning algorithms
  - Design of scalable KDD processes
- Database management
  - Data warehouse and NoSQL data modeling

# Data mining and machine learning algorithms

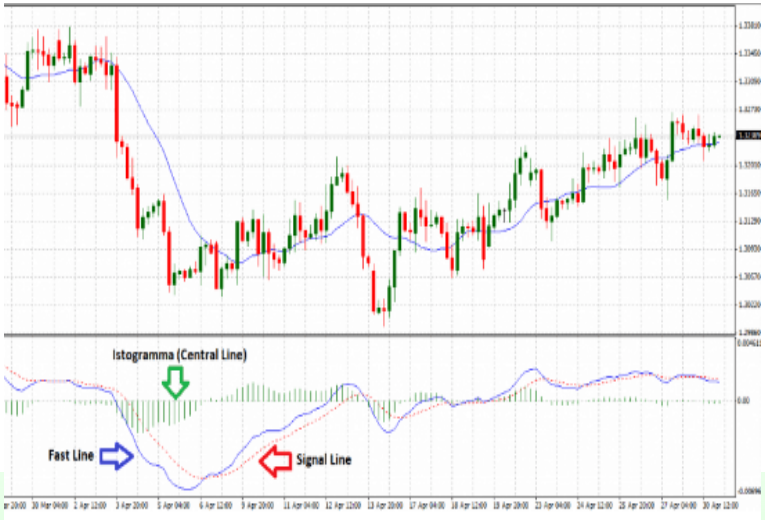
- *Clustering* and *semi-supervised clustering*
- *Scalable* data mining *algorithms* for *big data*
  - E.g., (approximate) clustering
- *Time series* analysis
  - Forecasting models, trend detection
- *Textual data* analysis
  - Summarization, clustering, classification
- *Predictive maintenance* for
  - Industrial processes, robots, automotive components, ...

# Data science pipeline

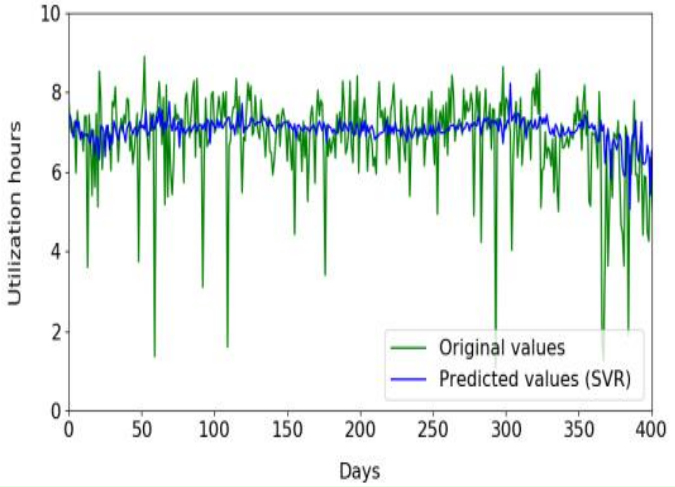
- Design and implementation of *personalized KDD processes*
- Black-box *prediction interpretation*
- *KDD process automation* by means of *self-tuning* and *concept drift detection* techniques
- *Semantic enrichment* by means of entity recognition and latent-based models

# Multivariate time series forecasting

- *Predict the future values* of a time series (e.g., physiological signals, historical stock prices) based on the analysis of *multiple series* (e.g., consider all the stocks of the same sector)
- *Extract* ad hoc features summarizing series *trends* (e.g., moving averages) and the related context (e.g., news articles, social data)




*Stock price forecasting*




*Prediction of vehicle utilization hours*

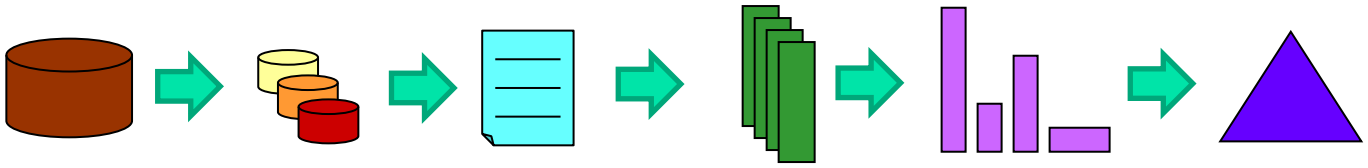
# Big data mining

- Study of innovative, parallel, and distributed data mining approaches for
  - Pattern mining algorithms
  - Clustering techniques
  - Classification algorithms
  - Summarization algorithms

  to efficiently gain interesting insights from huge data volume
- Design and development of novel cloud-based data mining services based on
  - HADOOP and Spark frameworks
  - MapReduce paradigm


- Exploitation of the cloud-based services for novel big data analytics applications (e.g., network traffic data, fraud detection, social networks)
- Analysis modules based on HADOOP and Spark Ecosystems

# Automated data science pipeline

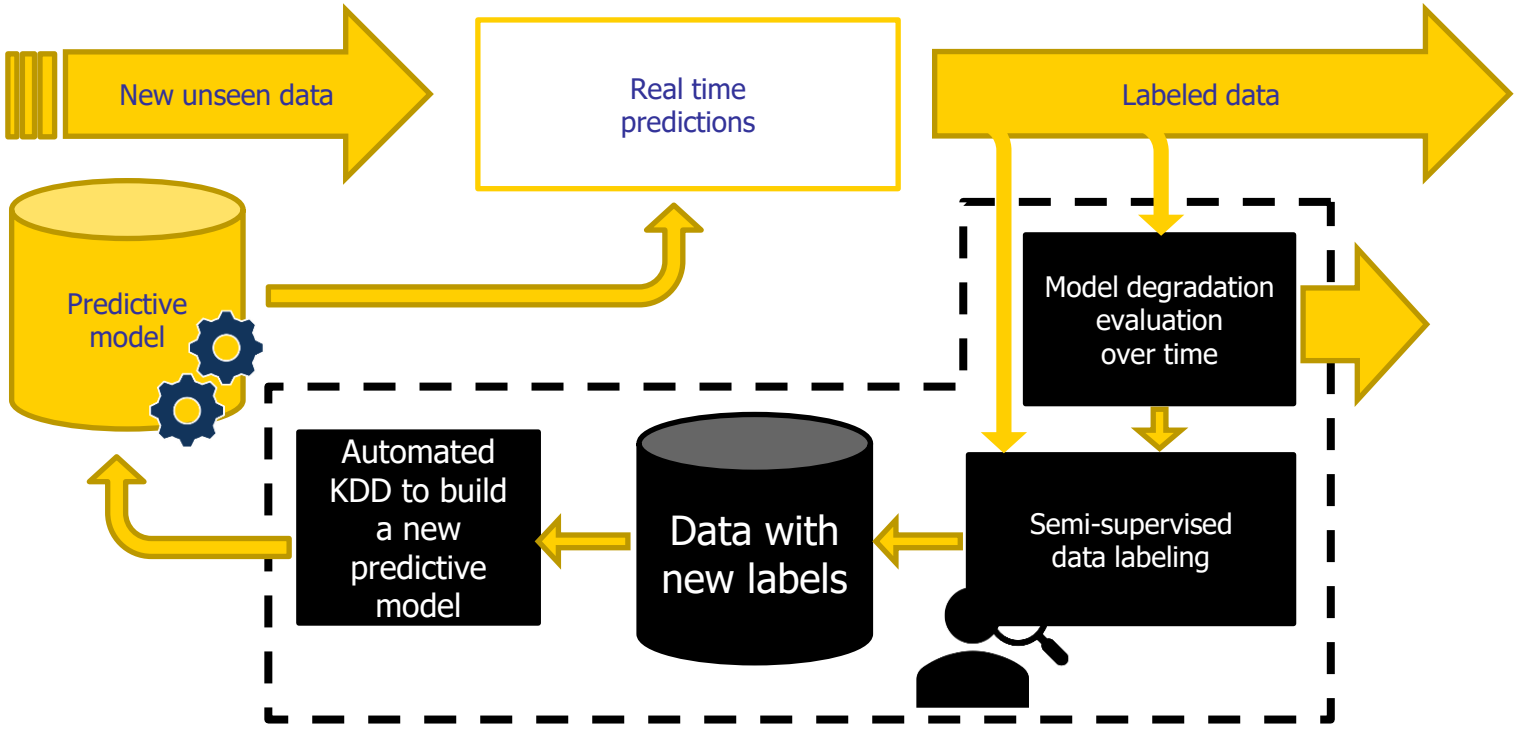


Data | Selection | Preprocessing | Transformation | Knowledge extraction | Visualization Explanation

## Automation in the data analytics process

- **Tailor** the **analytic** steps to the different key aspects of the **data under analysis**
- **Automate** the analytic workflow to **reduce manual user interventions**
- Translate the domain-expert knowledge into **automated procedures**
- Automatically configure input parameters by means of **self-tuning strategies**
- Design **informative dashboards** and **explanation techniques** to support the translation of the extracted knowledge into effective actions

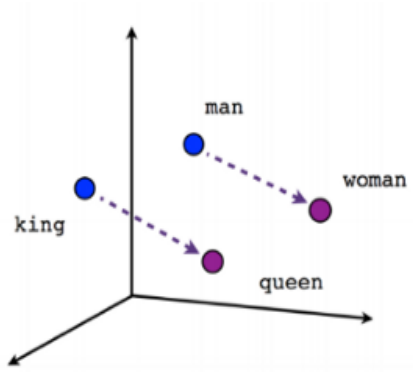
# Automated concept drift management




- 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

# Deep Natural Language Processing

- Vector representations of text data
  - Trained using Deep Learning models
  - Commonly used to address NLP tasks
  - Examples: Word2Vec, FastText, BERT, GPT-3
- Open issues
  - Text generation
  - Semantic specialization of distributional word vectors
- Methods
  - GAN
  - Seq2Seq models

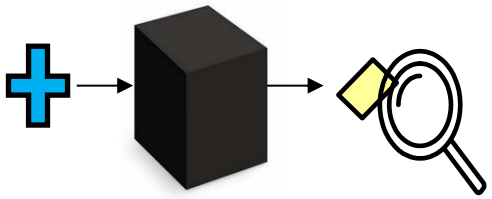


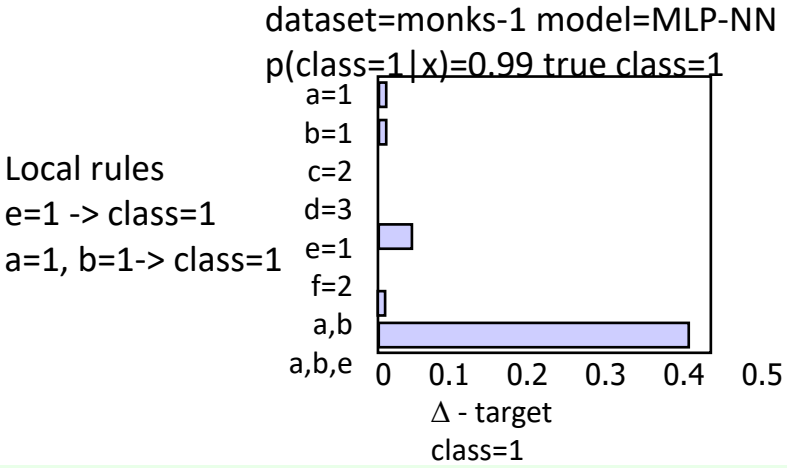
# Text summarization

- Problem
  - identification of salient knowledge from news articles, scientific publications, learning materials, social data
  - generation of sound and easy-to-read summaries of large document collections
- Current issues
  - Fine-tuning of Deep NLP-based solutions to specific domains
    - E.g., BERT, BART
  - Cross-lingual summarization
    - Summaries of collections of documents written in different languages
- Methods
  - Neural summarization
  - Itemset-based summarization

# Explainable AI

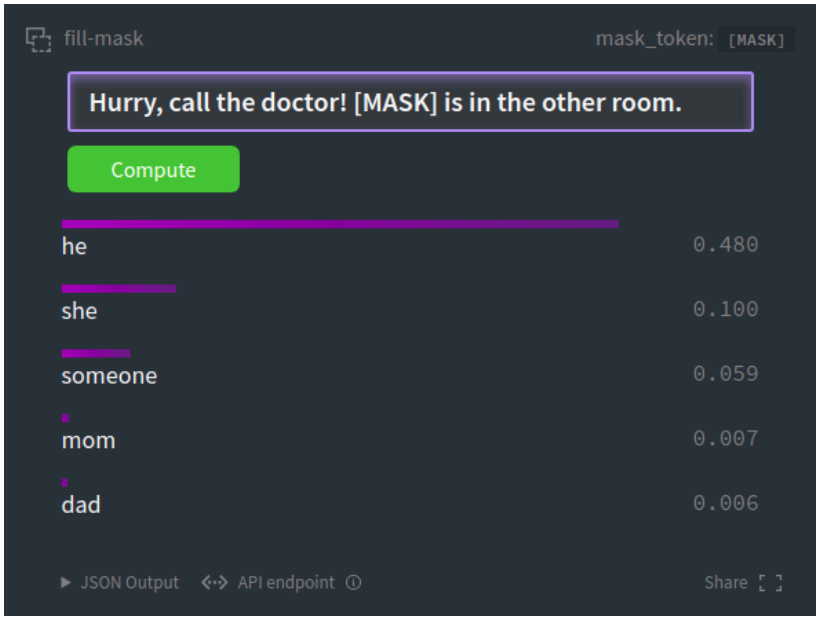
- Model-agnostic explanation methods to explain classifier predictions on single instances
  - Analysis of the joint effect of feature subsets on the prediction
  - Local properties of the original model to be explained are exploited by learning a local rule-based model in its neighborhood
- Dual form of provided explanations
  - Qualitative insight -> local rules
  - Quantitative insight -> prediction difference





# Interpretability of NLP models

- Inspect modern NLP architectures to understand how they encode linguistic features
  - Regularization purposes (e.g. reduction of gender bias)




Mask-fill predictions from Google's BERT

# Vulnerability of interpretability

- Analyze SoTA interpretability techniques to assess vulnerabilities and countermeasures
  - Exploitation of both Offensive and Defensive techniques


From:
*Explanations can be manipulated and geometry is to blame, Dombrowski et al., NIPS 2019*

# Data analytics for healthcare

- Design and development of a smart software component to allow a patient-centered delivery of medical services
  - Collect and integrate heterogeneous data including data on the structures providing medical services, information related to patients, doctors, and staff
  - Study and develop architectural, technological and algorithmic solutions to efficiently manage the above collected data to suggest the optimal medical structure to each patient

Piedmont research project **CANP**

- Physiological data analysis
  - analyze physiological data collected to detect/predict discomfort conditions in automotive environnement
    - characterize discomfort conditions
    - reduce the effect of discomfort conditions

DBMG

REFERENCE PERSON: PROF. SILVIA CHIUSANO

# AI & Financial Data

- Design and implementation of data mining-oriented strategies for financial data analysis
  - Forecasting
    - generate accurate predictions for various markets (stock markets, crypto-currencies) to support long- and short-term quantitative trading strategies
  - Trend detection
    - Apply Seq2Seq models to produce multi-scale descriptions of financial series

# Data analysis for Smart Cities

- Mining urban data to increase the well-being of citizens by improving the efficiency and accessibility of services
  - Analysis of data on citizen mobility in urban area
    - e.g., car pooling and bike sharing systems data to forecast critical situations and characterize the cyclic mobility patterns
  - Analysis of air pollution data on urban area to detect possible critical conditions
  - Analysis of data for citizen security and urban safety
- Different types of data area analyzed as sensor data, open data, social network data, etc.

# Spatio-temporal data mining

- Problem
  - We are overloaded by heterogeneous spatio-temporal data
  - Satellite images and measurements (e.g., Copernicus data)
  - Ground-based sensor measurements
  - Etc.
- Thesis goals
  - Design and implement data mining algorithms for
    - Describing spatio-temporal phenomena
      - By means of sequential and/or graph-based patterns
    - Predicting spatio-temporal events
      - E.g., Spatio-temporal classification algorithms

# Vehicular traffic data analysis

- Characterization of industrial vehicles' usage based on the analysis of BUS CAN bus data, routes, and driver profiles
- Predictive maintenance
  - Predict faults based on the analysis of diagnostic messages
- Profile vehicle duties
  - Cluster vehicle usage data
- Optimize signal transmission
  - Optimize schedules of CAN bus messages

TIERRA
TELEMATICS DESIGN