Theses

Data Base and Data Mining Group of Politecnico di Torino

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

General information

- Duration: 4-6 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.surname@polito.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, ...

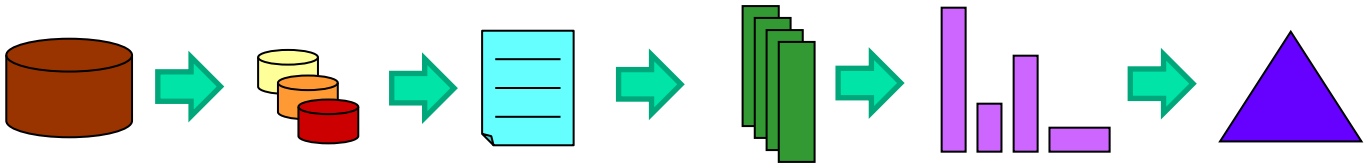
Data science pipeline

- 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
- *Theory-guided data science*, blending data-driven modeling with domain-provided knowledge

Big data mining

- Study of innovative, parallel, and distributed data mining approaches for
  - Pattern mining algorithms
  - Clustering techniques
  - Classification algorithms
- Design and development of novel parallel algorithms based on the Spark framework
- Exploitation of the novel algorithms for 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

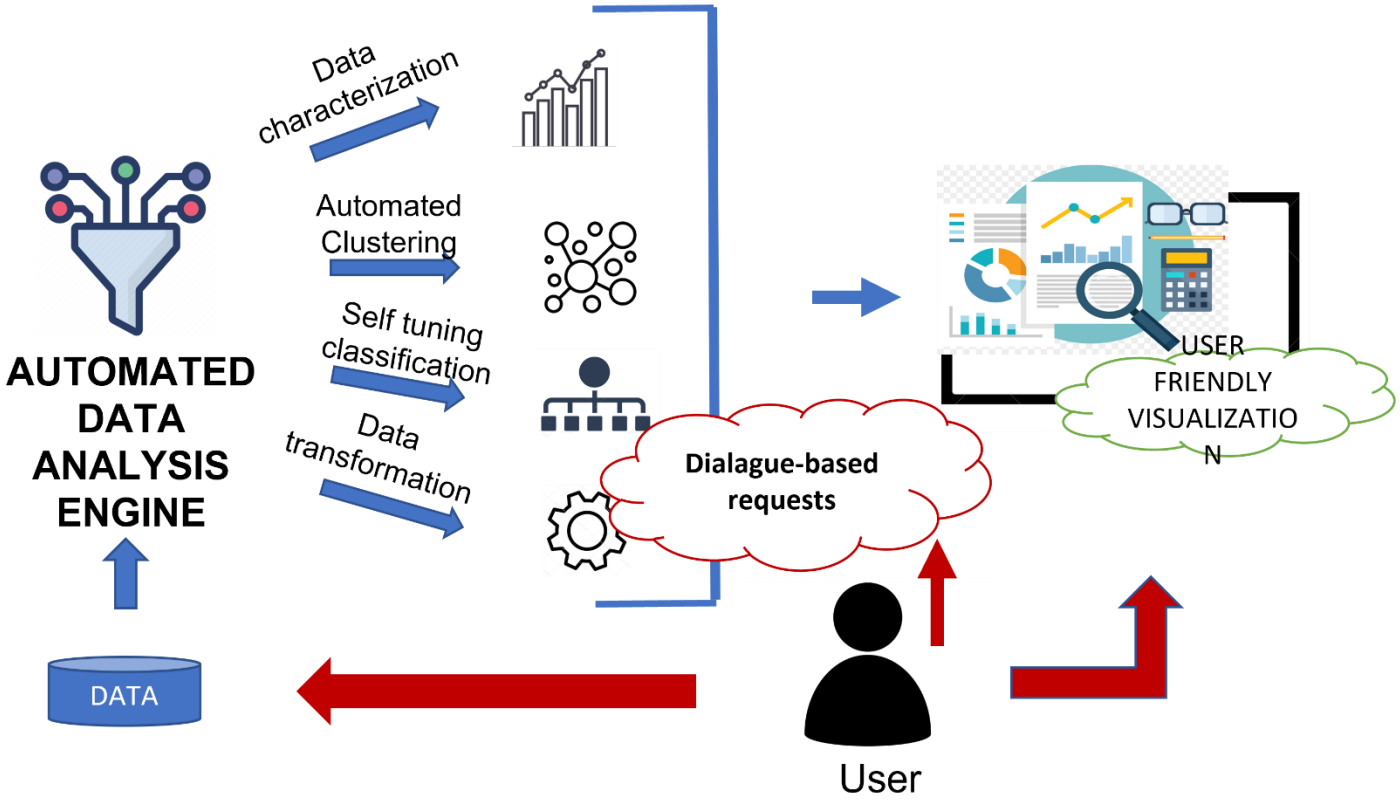
Automation in the data analytics process

- **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

Conversational data science

Objective: making data analysis accessible to all and **democratizing the data science**, through:

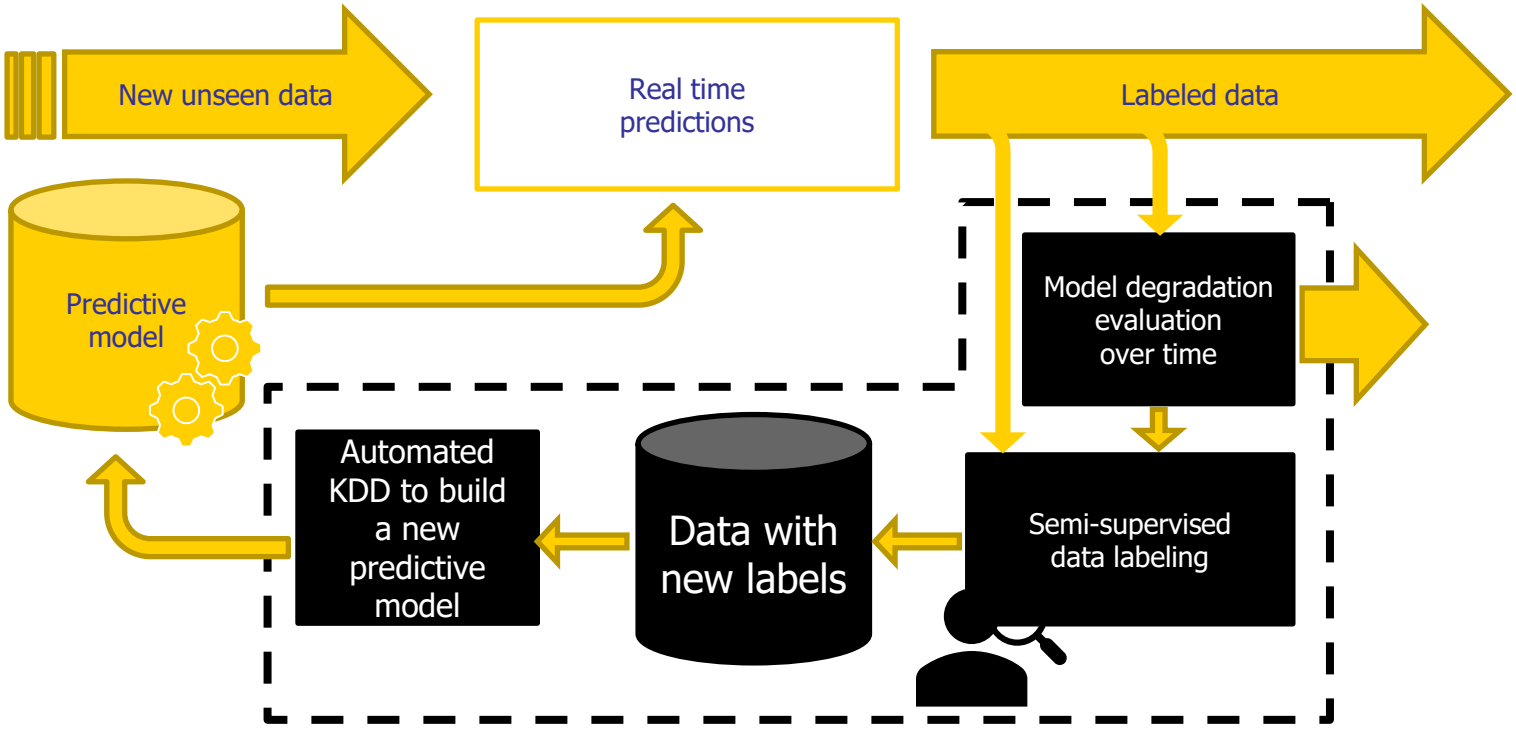
- self-tuning algorithms
- friendly interaction between user and a data analysis engine through a **conversational approach**
- complexity of the analysis hidden to the user



Opens issue:

- define ad hoc self tuning strategies for each considered algorithm
- offer a friendly and dialogue-based interaction to the user, properly handling its requests
- choose the most suitable visualization to show the results to the user

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 architectures
  - Commonly used to address NLP tasks
  - Examples: Word2Vec, BERT, BART, PEGASUS, GPT-3
- Current issues
  - Multimodal/multilingual text summarization
    - Extract a concise descriptions of multimodal/multilingual data sources
  - Generative sentiment analysis
    - Apply generative models to address sentiment analysis
  - Automatic text paraphrasing
    - Reformulate text in natural language by conveying the same meaning with different words

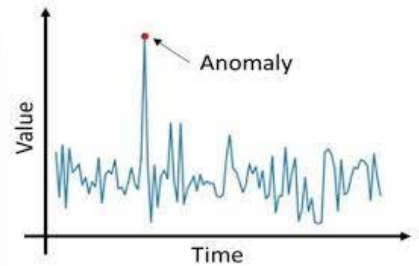
Transformers

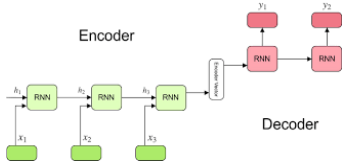
Time series forecasting

- Design multivariate supervised models for predicting time series directions/values
  - Application of regression/classification models in different domains
    - Seq2Seq, Reinforcement Learning, and Transformer models

- Finance
  - Stock and cryptocurrency trading
  - Both Long- and short-term strategies

- Vehicular data
  - Predictive maintenance
  - Anomaly detection





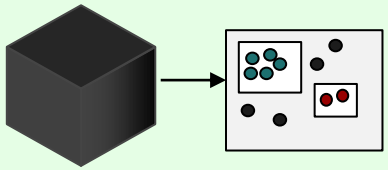
  - Ongoing collaboration with Tierra spa

TIERRA

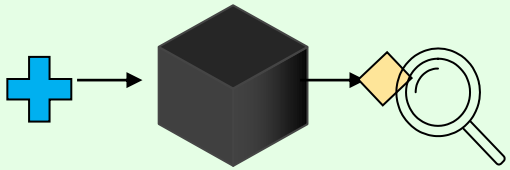
REFERENCE PERSON: PROF. LUCA CAGLIERO

Explainable AI

- Understanding the behavior of machine learning models
  - From the individual and subgroup perspective



Subgroup



Local predictions

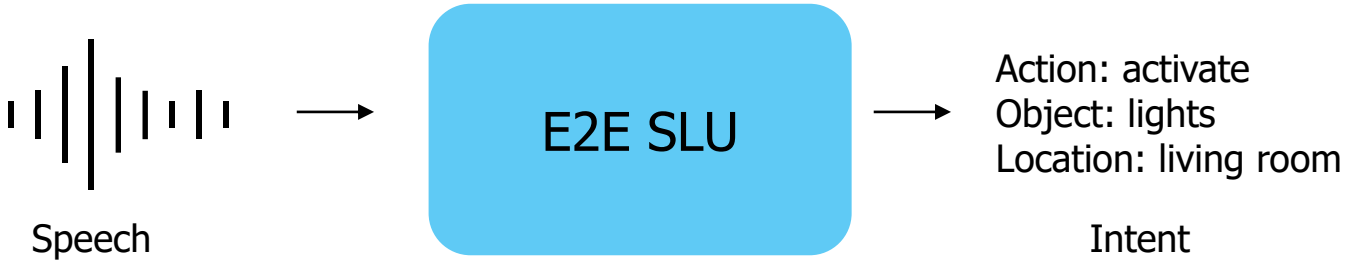
- Exploring the relationship with the desiderata of XAI research

FAIRNESS

ERROR ANALYSIS
& DEBUGGING

Explaining End-to-End SLU models

- Characterizing the behavior of End-to-End (E2E) Spoken Language Understanding (SLU) systems

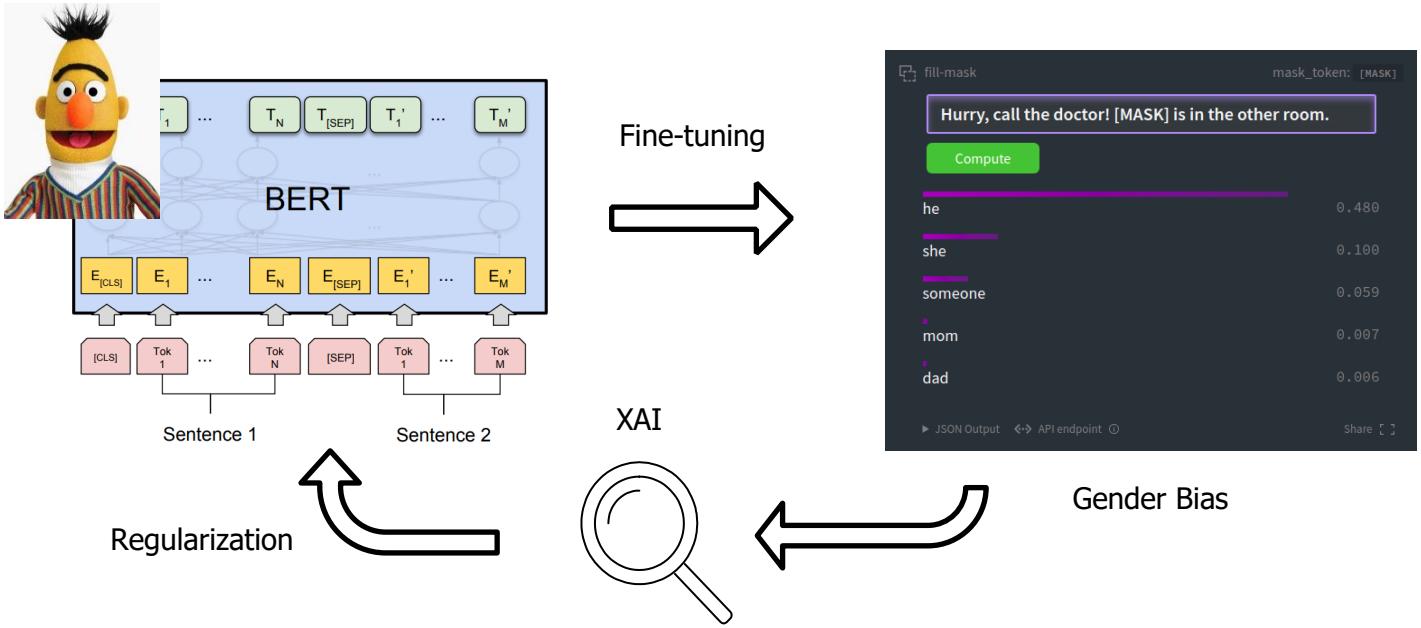


- Understanding system behavior from the subgroup perspective
- Explaining model errors and anomalous behaviors

amazon alexa

Explaining Transformers

- Analyzing and explaining transformer models
- Focus on Natural Language Models

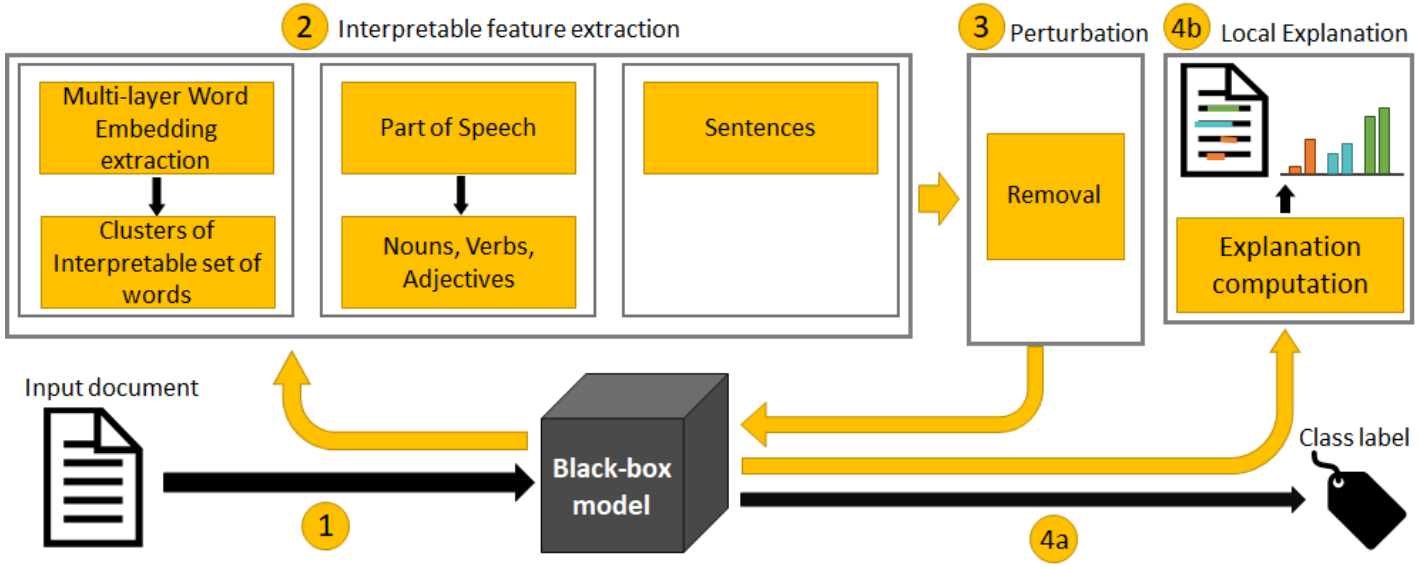


- Characterize shortcomings (e.g., gender bias) and propose regularization techniques

XAI in NLP

Goal: Designing innovative XAI solutions that offer a level of transparency greater than existing methods tailored to NLP

- We currently implemented an XAI solution, named T-EBAnO, producing local and global explanations for deep neural networks in text classification tasks



Open Issues

- A quantitative XAI evaluation methodology
- Comparison with other XAI techniques
- Extending the approach to tasks other than classification (e.g., Q&A)
- Exploit the XAI explanations for re-training the models in case of wrong predictions or correct predictions but for the wrong reason, with the human-in-the-loop feedbacks

REFERENCE PERSON: PROF. TANIA CERQUITELLI

Cross-lingual sentiment embeddings

- Extraction of embeddings based on the sentiment of words
  - e.g. for sentiment analysis tasks
- Not trivial for languages other than English
- Focus on the adoption of *propagation techniques* from English to other languages

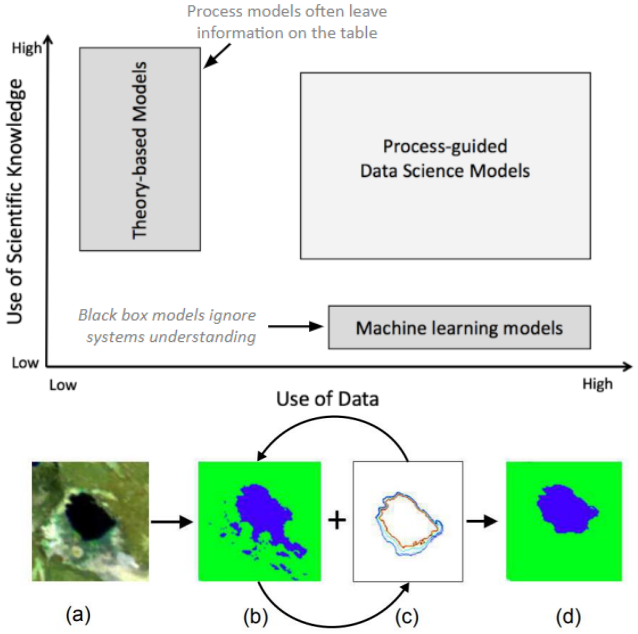
x	*x*'s nearest neighbors	Cosine similarity
excellent	eccellente	0.575
	ottimo	0.513
	apprezzabile	0.369
	buon	0.367
	adatto	0.322

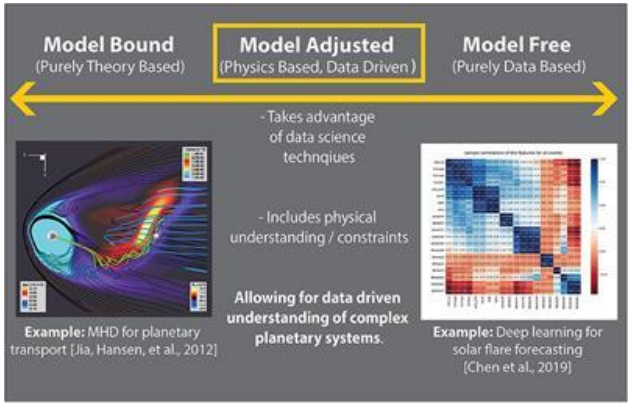
Spatio-temporal data mining

- 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
      - By means of graph-based patterns and ST-GNN (Spatio-Temporal Graph Neural Networks)

Theory-guided Data Science

- Data science models, although successful in several domains, have limited applicability in scientific problems involving complex phenomena
- Theory-guided data science blends data-driven models with domain-provided knwoledge
  - leveraging existing scientific knowledge for improving the effectiveness of data science models
- Examples
  - intelligent supply chains, business process optimizations, medical image recognition
  - exploiting Graph Neural Networks as a mean to model complex systems and inter/intra relationships among them and their components





https://arxiv.org/pdf/1612.08544

Urban Data Science: Mining mobility data

- Profile user activities and interests from mobility data
  - Mobility data provide an implicit feedback on the most common users' urban activities and interests

- Improve mobility-based services
  - Forecast vehicle availability in urban areas

- Analyse user perception of mobility-based services
  - Sentiment analysis on user comments on mobility-based services

Different real open data on mobility based services (e.g., free floating car sharing, e-scoter, bikes) and social networks.

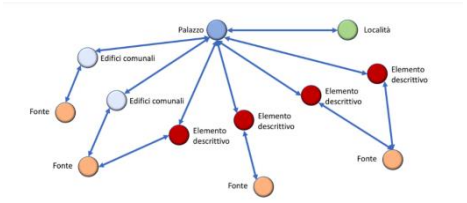
National research project

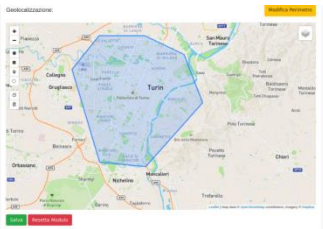
sm2art

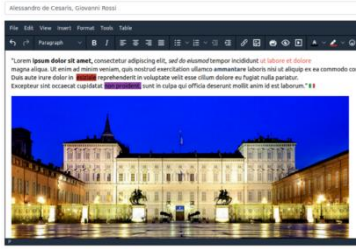
Digital Cultural Heritage

Development of a web-based digital atlas for historical buildings

- Store heterogeneous information (building structure, textual documentation, pictures, video, …) on geo-located historical buildings
- Support geo-referenced queries to retrieve information on buildings located in a geographic area
- Profile user interaction with the system and guide end-user in identifying the information of interest







- Adopted technologies: data repository: MongoDB; backend: Django; frontend: JQuery, Bootrstrap, Leaflet, TinyMCE.

Apps for e-Education

DesignER  (designer.polito.it)

- Web-Based Application for Computer-Aided Design of Relational Databases
- Guide users in the conceptual and logical design of a relational database through a tutoring-based approach
- Profile user design patterns and user mistakes in design process

REFERENCE PERSON: PROF. SILVIA CHIUSANO

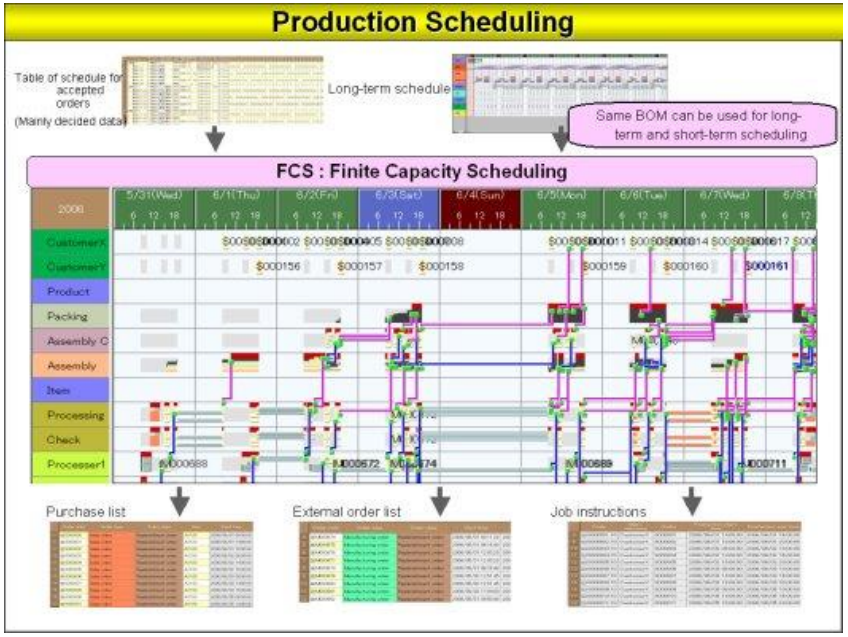
Development of a platform for unplugged coding

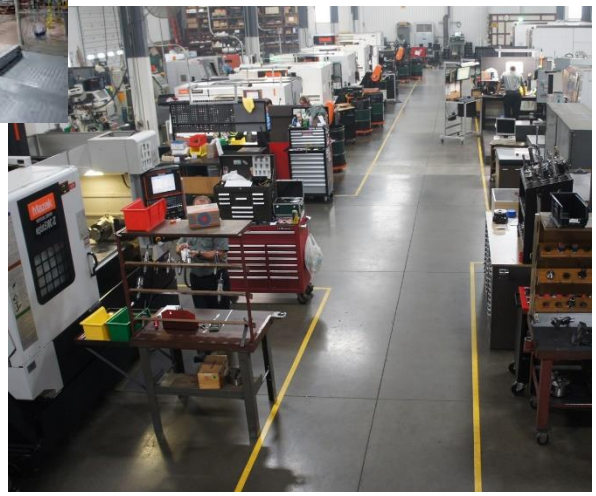
- Coding courses are currently offered at pre-school and primary schools
- this thesis aims at developing a platform to help teachers to realize unplugged coding exercises and to be able to share them with the school community in a simple and intuitive way

REFERENCE PERSON: PROF. ALESSANDRO FIORI

Manufacturing optimization

- Exploration and development of deep-learning techniques to optimize KPIs of future scheduling from historical data of the manufacturing process
  - E.g., reduce the overall duration, maximize throughput, shape energy consumption, etc.

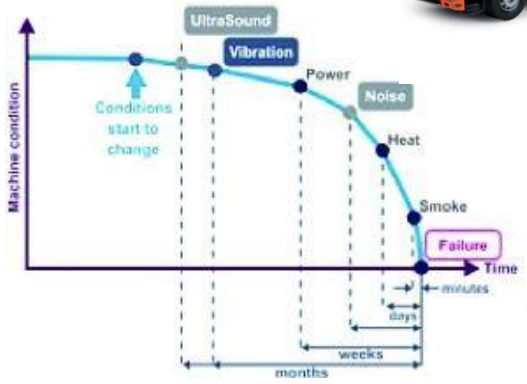


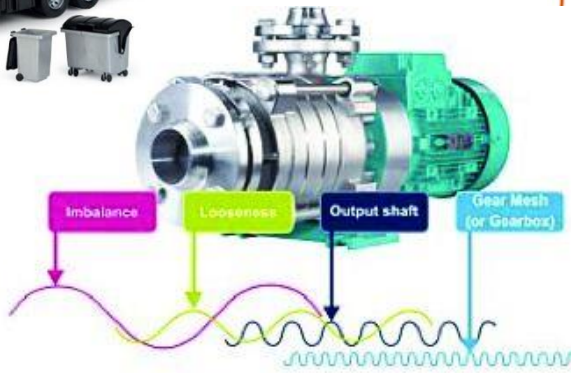


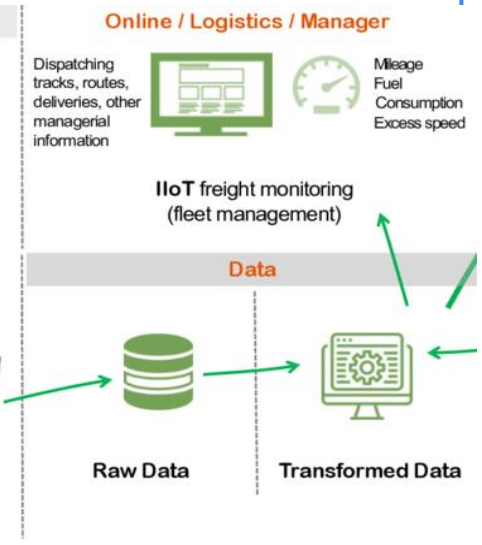
Vehicle sensor-data analysis

- Identifying anomalies and patterns in sensor data collected from vehicles
  - Operational time series to profile vehicle behavior
  - Contextual information, e.g., maintenance interventions
- Example

  optimizing garbage-collection operations and procedures by analyzing sensor data collected from trucks and their maintenance interventions

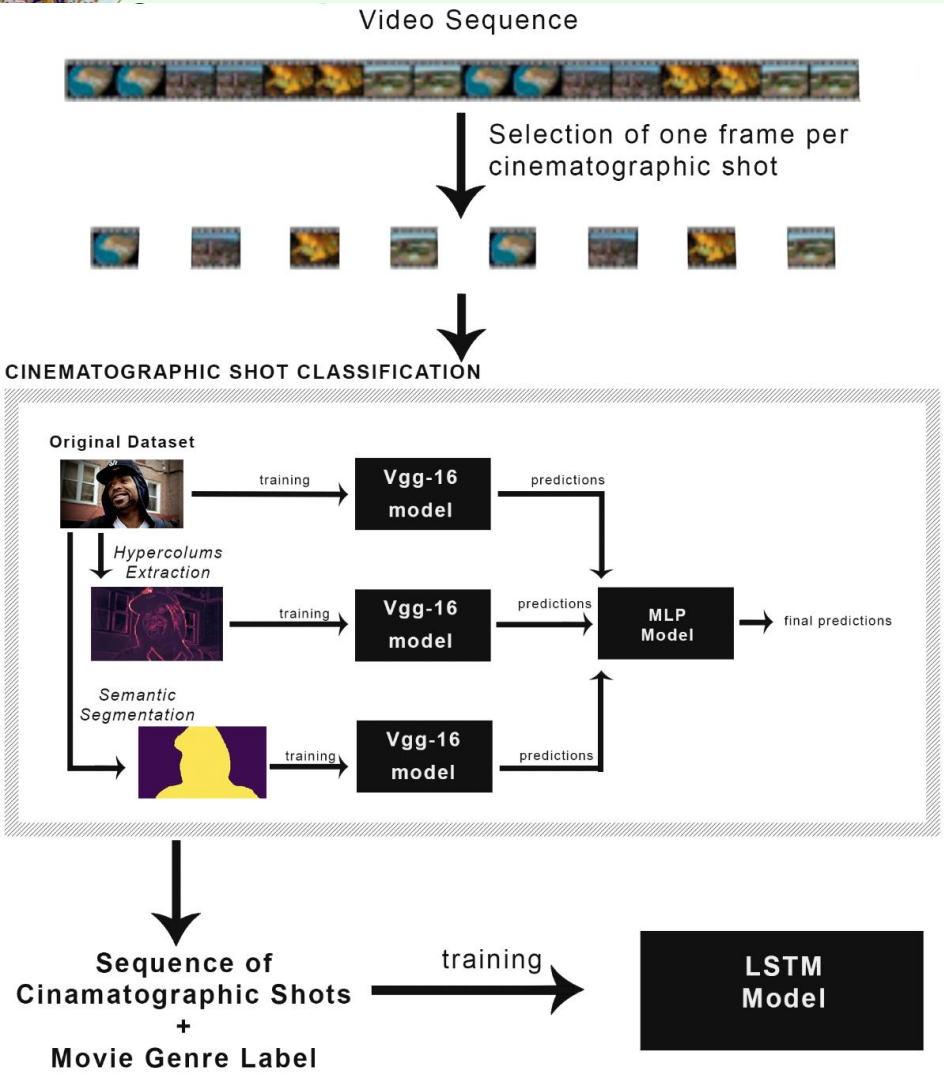






REFERENCE PERSON: TANIA CERQUITELLI

Creativity-injection into AI-powered multimedia storyboards



Research objectives:

- Cinematographic **shot classification**
- Movie Genre Classification based on the editing techniques
- Modeling **user creativity profiling**
- Inferring **creativity-based** production storyboards
- Creativity-based production storyboards with **human-in-the-loop**

Open issues

- Few labeled datasets available
- Unbalanced class distribution
- User creativity is hard to be modeled

REFERENCE PERSON: PROF. TANIA CERQUITELLI

Machine Learning on HPC

Objective: Modeling and predicting the job/activity time on HPC by exploiting machine learning algorithms

CNH INDUSTRIAL

set of variables describing the submitted job

Job Description

set of variables describing hpc resources at the time of submission

Context Variables

set of variables that model all jobs launched with the same software

Model Parameters

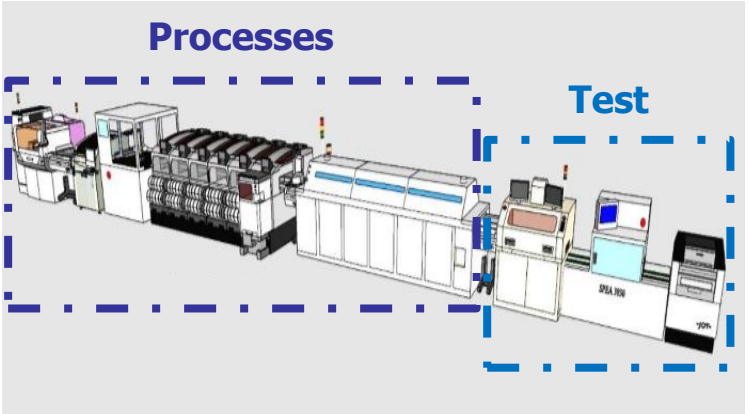
Model

Time Prediction

Data Science Pipeline for Defect Prediction

Thanks to the Industry 4.0 the manufacturing **processes** expose many data about production

- **conditions**: environment conditions, wear and tear of the machines, etc.
- **parameters:** oven temperature, production speed, components characteristics, etc.



Conditions and parameters affect the outcome of the production line and can lead to **defects** in the final product

- defects that are detected only at the end of the production by specific **test procedures**

The thesis aims to

- identify the most important conditions/parameters describing and **healthy**/**defect** production
- create a Machine Learning pipeline to predict the outcome (**healthy**/**defect**) based on the production conditions/parameters

The thesis leverages **real** production line **data**

BITRON

DBMG

REFERENCE PERSON: PROF. DANILO GIORDANO