Theses

Elena Baralis, Silvia Chiusano, Tania Cerquitelli, Paolo Garza, Luca Caglieri, Daniele Apiletti, Alessandro Fiori
General information

- **Duration**: 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 polite dot it)
Main Topics

- Big data and cloud-based data mining services and algorithms
- Database and data mining applications
  - Industrial data analysis
  - Financial data mining
  - E-learning and social network data analysis
  - Energy data mining
  - Medical data mining
  - Data analysis for Smart Cities
  - ....
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**

**REFERENCE PERSON: PROF. PAOLO GARZA**
Design and implementation of an **automatic system** (Mining Advisor) to select for a dataset an **optimal mining algorithm** for a given analysis task based on

- innovative data characterization statistics
- definition/design of mining algorithms (i.e., access methods and mining primitives), possibly disk-based
- algorithm selection strategies exploiting a trade-off between accuracy and exploration time

Different instances of a Mining advisor can be **tailored to different data mining techniques** (e.g., clustering algorithms, pattern discovery)

**REFERENCE PERSON:** PROF. TANIA CERQUITELLI
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

- **Open issues**
  - Cross-lingual summarization
    - Generation of summaries of collections of documents written in different languages
  - Supervised summarization
    - Generation of textual summaries based on *predictive models* learnt from training documents
  - Incremental summarization
    - Efficient *update* of the previously generated summaries as soon as the input documents change

**REFERENCE PERSON:** DR. LUCA CAGLIERO
Interpretability in machine learning

- **Study of techniques for model and prediction explanation**
  - global understanding of how a model works
  - local understanding of why a prediction is made

- **Use of semantic information to enhance classification accuracy**
  - Outlier and error detection in black box models by exploiting knowledge rules inferred from data

**REFERENCE PERSON: PROF. ELENA BARALIS**
Integration and analysis of tourism data

- Online Travel Agencies (OTAs) and review aggregator websites (e.g., TripAdvisor, Airbnb) are profoundly changing the structure of the hospitality industry.
- Integrating and analyzing data acquired from OTAs is crucial for creating and capturing economic value for hotels.

Main goals
- Data crawling from multiple sources
- Data enrichment with semantic information
- Data integration
- Knowledge extraction from hotel reviews using Natural Language Processing techniques

REFERENCE PERSON: DR. LUCA CAGLIERO
Learning analytics

- Measurement, collection, analysis, and reporting of data about **learners** and their contexts, for purposes of understanding and optimizing **learning** and the environments in which it occurs.
  - **Measure learners’ engagement**
    - Questionnaires, serious games, Web-based platforms, MOOCs
  - **Characterize learners’ profiles**
    - Identify recurrent learners’ activities and habits
    - Detect unexpected behaviours

**REFERENCE PERSON: DR. LUCA CAGLIERO**
Vehicular traffic data analysis

- Characterization of industrial vehicles’ usage based on the analysis of BUS CAN bus data, routes, and driver profiles
- Fleet management
  - Optimize routes
- Predictive maintenance
  - Planning of periodic maintenance actions

REFERENCE PERSON: PROF. ELENA BARALIS
Petrographic Data Analysis

- Application of **Data mining, Machine learning,** and **Deep learning** techniques to Petrographic data
- Pore typing image analysis
  - by means of **image clustering** and/or **classification**
- Enhanced **mineralogical analysis** by means of regression techniques based on **transfer learning**
  - infer the results of a mineralogical analysis by exploiting the results of another type of rock analysis

REFERENCE PERSON: PROF. ELENA BARALIS
Focus on vehicle and engine remote diagnosis and prognosis:

“Modern cars are equipped with dozens of sensors, both on the engine and the body, that collect huge amounts of data about the car and its environment. These data is stored on cars and later send to the manufacturer's servers for analyses. In the near future this flow of data will be transmitted constantly and in real time through 4G data channels.

This thesis aims at defining methods to aggregate and analyze data from large car fleets, for car diagnostic and prognostic goals, in batch mode initially, and real time mode later.”

REFERENCE PERSON: PROF. ELENA BARALIS
GM Global Propulsion Systems

Development of a tool for knowledge management in automotive

Mission
Design of high performance database and application to support advanced analytics and search on heterogeneous data produced by GM (pictures, reports, raw data) during the last 10 years of engine development.
Unique opportunity to design an industrial relevant application from scratch.

Additional information
expected duration: 6 months
on-site thesis at GM Global Propulsion System in Torino
possible monthly reimbursement of expenses.

REFERENCE PERSON: PROF. ELENA BARALIS
Dynamic Integrated ShopFloor MANagement

- Design and development of software components to collect, integrate and manage a huge amount of data collected during industrial processes and activities and processed in the cloud
  - Identify architectural, technological and algorithmic solutions to efficiently manage huge amounts of data
  - Collect large amounts of heterogeneous data generated through IoT devices, in a continuous, reliable and trusted way, thanks to virtualized software components (e.g., IoT data-connectors)
  - Integrate and store data on distributed systems with ad-hoc indexing strategies to guarantee efficient data management, querying and analytics activities

Piedmont research project, with 28 International partners

REFERENCE PERSON: PROF. TANIA CERQUITELLI
Mining medical data

- **Physiological data analysis**
  - analyze **physiological data** collected during **incremental tests** (e.g., cardiopulmonary exercise testing) commonly used in clinical domain and in sport science
    - **improve** the effectiveness of the reliability/training sessions
    - **predict** the final values of crucial parameters
    - **reduce** test duration and the physical effort for patients/athletes

- **Clinical data analysis**
  - analyze data collected by the healthcare network of an Italian Health Care Center
    - **extract** medical treatments (in terms of performed examinations, prescribed drugs) frequently done by patients
    - **identify** deviation from expected medical treatments according to medical guidelines

REFERENCE PERSON: PROF. SILVIA CHIUSANO
Green data mining

- Joint analysis of
  - Energy consumption logs of residential and public building heating systems and indoor climate conditions
  - Data on the user thermal comfort perception of indoor climate conditions and user feedbacks

- Goals
  - Suggest ready-to-implement energy efficient actions based on innovative and user-friendly indicators
  - Discovery of interesting correlations in the large and heterogeneous amount of available data

REFERENCE PERSON: PROF. TANIA CERQUITELLI
Green data mining

- Joint analysis of energy and water consumption data to efficiently support an intelligent building management system
  - localization of network losses and leaks
  - detection of abnormal consumption
  - characterization of user consumption
  - forecast of energy and water consumption

- Beyond smart-meters and smart thermostats. Cloud-based data-driven energy consumption modeling and forecasting of a grid of customer households
  - Aims at closing the loop inside the on-premises boiler controllers, to reach new levels of efficiency and customer satisfaction

REFERENCE PERSON: PROF. TANIA CERQUITELLI
Predictive maintenance and quality

■ Input
  ■ professional coffee machines telemetry data

■ Output
  ■ prediction of faults in coffee machines
  ■ prediction of coffee quality

■ Methodology
  ■ Exploration analysis
  ■ Feasibility study
  ■ Data mining classification model

REFERENCE PERSONS: PROF. ELENA BARALIS, DR. DANIELE APILETTI
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

REFERENCE PERSON: PROF. SILVIA CHIUSANO
Financial Data Mining

- Design and implementation of data mining-oriented strategies for online trading
  - Problem
    - generation of accurate predictions for various markets (stock markets, crypto-currencies)
  - Scenarios
    - Intraday trading
      - Generation of short-term
    - Long-term portfolio planning
      - Generation of profitable and diversified portfolios

REFERENCE PERSON: DR. LUCA CAGLIERO
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.

REFERENCE PERSON: PROF. SILVIA CHIUSANO
Laboratory Assistant Suite
- modular architecture to manage different kinds of raw experimental data, tracking several laboratory activities, integrate different resources and aid in performing a variety of analyses to extract knowledge related to tumors
- design of model-driven automated GUI generation
- development of infrastructural components (e.g., task scheduler, query designer, dashboard)

Genome analysis and visualization
- analytical algorithms to identify genetic variants from NGS data
- implementation and optimization of analytical algorithms for identification and classification of genetic variants in paired comparative tests
- development of data analysis pipelines in parallel and distributed environment
- design of graphical representations of genomic data (e.g., cBioPortal)

Microarray data analysis
- study of class discovery algorithms (e.g., clustering, bi-clustering)
- identify robust gene markers by means of the integration of several classification methods
- analysis of gene expression values over the time on data derived from xenopatients

REFERENCE PERSON: DR. ALESSANDRO FIORI
Virtual Laboratory @ IRCCS

- Realization of virtual representation of life-science working environments based on 3D interactive models
  - development of a prototypic application for user-friendly management of complex and hierarchical storage systems, by means of 3D realistic representation of the physical containers and their interactions
- Computer vision for sensor-based real-time tracking of laboratory activities
  - development of a prototypic platform for automated monitoring of interactions between users, instruments and experimental materials
  - virtual representation of objects and activities is also foreseen to provide intuitive and user-friendly GUIs

Virtual reality to improve graphical user interfaces usability in laboratory information management systems (LIMS)

Computer vision to improve the efficiency of laboratory data-tracking procedures

Biological laboratories need Next Generation LIMS

REFERENCE PERSON: DR. ALESSANDRO FIORI