Curriculum “Data and Knowledge: Science and Technologies”

Objectives/visions - *Public interest in data science and big data is mounting as data-driven decision making becomes visible in everyday life. Society shifted from being predominantly “analog” to “digital” in just a few years. Companies, organizations, and people are “always on”. The Internet of Things (IoT) is contributing significantly to this expansion: Our homes, cars, factories, and cities are getting “smarter” by exploiting the data that is collected from smaller and smaller devices at any time, at any place and about everything. This makes it possible to record and analyze the behavior of people, machines, and organizations.*

While these big data are expected to foster wellbeing, social development and the economy, they are produced at a rate that is far greater than that of the computing power to process them and we miss most of the analytical tools, models and skills to make them usable. This means that we are facing the challenge of providing algorithms, models, methodologies, tools, technologies and competences to acquire, store, process, analyze, search and mine these data. Developments in these areas will have significant impacts onto scientific, business and social applications in many diverse fields such as web search, online social networks, banking, manufacturing (industry 4.0), mobility and transportation, health care and genomics, policy making, education, retail, and so on. These developments are also rapidly changing the way we do business, socialize, conduct research, and govern society.

Unfortunately we are experiencing a severe shortage of scientists and professionals able to master the models, the algorithms and the technologies to turn data into meaningful information, with a strong computer science background. So there is no surprise that small and global companies are seeking for professionals with these skills, ranging from software to data-intensive companies; from telecom to telematics providers; from retail companies to energy providers; from insurance companies to statistics institutes; from pharmaceutical to health-care companies; up to the huge universe of small and medium enterprises as well as start-ups that are developing products deploying Data in every business domain. Given these premises, the curriculum has been designed with the aim of educating the next generation of "data architects" and “software and algorithm engineers” endowed with deep computational, methodological and modeling skills that will allow them to design and implement the future data-intensive algorithms, tools and platforms, as well as master the cutting-edge technologies for big-data analytics, such as Hadoop, Spark, together with the mainstream tools for data and text mining, machine learning, artificial intelligence, complex system modeling and mining spurring from e.g. genomic, web, business, industrial or social applications.

Studies plan

	Course name	**CFU**
	57 CFU OF:	
1	Algorithm engineering (con WTW e ICT)	9
2	Data mining (con WBI per 6 CFU)	9
3	Advanced data bases (con WBI)	9
4	Information retrieval (con WTW)	6
5	Bioinformatics	6
6	Distributed systems: paradigms and models (con WTW)	9
7	Computational mathematics for learning and data analysis (con AI)	9
8-11	**30 CFU (2 da 9 e 2 da 6)OF:**	
	ICT infrastructures (ICT)	6
	ICT risk assessment (ICT)	9
	Mobile and cyber physical systems (ICT)	9
	Machine learning (AI)	9
	Human languages technologies (AI)	9
	Big data analytics (WBI)	6
	Social and ethical issues in computer technology	6
	Peer to peer systems and blockchains (ICT)	6
	Scientific and large data visualization (CNR)	6
	33 CFU OF:	
12	Free choice	9
	Thesis	24

Syllabus

Algorithms engineering [9 CFU]

Study, design and analyze advanced algorithms and data structures for the efficient solution of combinatorial problems involving all basic data types, such as integer sequences, strings, (geometric) points, trees and graphs. The design and analysis will involve several models of computation — such as RAM, 2-level memory, cache-oblivious, streaming — in order to take into account the architectural features of modern PCs and the availability of Big Data upon which algorithms could work on. We will add to such theoretical analysis several engineering considerations spurring from the implementation of the proposed algorithms and from experiments published in the literature

- *Design of algorithms for massive datasets: disk aware or cache oblivious*
- *Design of advanced data structures in hierarchical memories for atomic or string data*
- *Data compression for structured and unstructured data*
- *Algorithms for large graphs*
- *Engineering considerations about the implementation of algorithms and data structures*

Data mining [9 CFU]

This course provides a structured introduction to the key methods of data mining and the design of knowledge discovery processes. Organizations and businesses are overwhelmed by the flood of data continuously collected into their data warehouses as well as sensed by all kinds of digital technologies - the web, social media, mobile devices, the internet of things. Traditional statistical techniques may fail to make sense of the data, due to the inherent complexity and size. Data mining, knowledge discovery and statistical learning techniques emerged as an alternative approach, aimed at revealing patterns, rules and models hidden in the data, and at supporting the analytical user to develop descriptive and predictive models for a number of challenging problems.

- *Fundamentals of data mining and of the knowledge discovery process from data.*
- *Design of data analysis processes.*
- *Statistical exploratory analytics for data understanding.*
- *Dimensionality reduction and Principal Component Analysis.*
- *Clustering analysis with centroid-based, hierarchical and density-based methods, predictive analytics and classification models (including decision trees, bayesian, rule-based, kernel-based, SVM, random forest and ensemble methods), pattern mining and association rule discovery.*
- *Validation and interpretation of discovered patterns and models within statistical frameworks.*
- *Design and development of data mining processes using state of the art technology, including KNIME, Python, and R, within a wrap-up project aimed at using and possibly modifying the DM tools and libraries learned in class.*

Advanced data bases [9 CFU]

Database systems occupy a central position in our information-based society, and computer scientist and database application designers should have a good knowledge about both the theoretical and the engineering concepts that underline these systems to ensure the application performance desired. The student who completes the course successfully will demonstrate advanced technical knowledge of the main issues related to the implementation of both classical centralized relational database systems for operational and OLAP processing and of recent advances in non-relational data models (columnar, document, key-value, graph) and scalable distributed architectures. The skills provided will make the student a sophisticated developer of high-performance database applications.

- *Internals of relational database management systems.*
- *Data Warehousing management systems and On-Line Analytical Processing.*
- *Extract-Transform-Load and query/reporting in OLAP systems.*
- *Beyond SQL: NoSQL data management systems for big data.*
- *Distributed data processing and the Map-Reduce paradigm*

Information retrieval [6 CFU]

Study, design and analysis of IR systems which are efficient and effective to store, process, mine and search big data coming from textual as well as any unstructured domain. Some attention will be also posed onto data that can be

modeled as labeled graphs, and into software systems that process those kinds of data. We will adopt mainly an algorithmic approach in studying and analyzing the components of a search engine and the algorithmic techniques, which are now ubiquitous in most modern IR applications.

- *The modules constituting a search engine: algorithms and data structures*
- *Recommendation systems, advertising and other IR applications*
- *Text mining and text annotation*
- *Clustering, classification, compression, sketching and other IR technicalities*

Bioinformatics [6 CFU]

This course has the goal to give the student an overview of algorithmic methods that have been conceived for the analysis of genomic sequences, and to be able to critically observe the practical impact of algorithmic design on real problems with relevant applications. *The exam, besides the obvious goal to evaluate the students understanding of the course contents, is additionally meant as a chance to learn how a scientific paper is like, and how to make an oral presentation on scientific/technical topics, as well as to design it for a specific audience.*

- *A brief introduction to molecular biology*
- *Sequences Alignments*
- *Pattern Matching*
- *Fragment Assembly*
- *New Generation Sequencing*
- *Motifs Extraction*

Distributed systems: paradigms and models [9 CFU] (vedi WTW)

Computational mathematics for learning and data analysis [9 CFU]

The course introduces some of the main techniques for the solution of numerical problems that find widespread use in fields like data analysis, machine learning, and artificial intelligence. These techniques often combine concepts typical of numerical analysis with those proper of numerical optimization, since numerical analysis tools are essential to solve optimization problems, and, vice-versa, problems of numerical analysis can be solved by optimization algorithms. The course has a significant hands-on part whereby students learn how to use some of the most common tools for computational mathematics; during these sessions, specific applications will be briefly illustrated in fields like regression and parameter estimation in statistics, approximation and data fitting, machine learning, artificial intelligence, data mining, information retrieval, and others.

- Multivariate and matrix calculus
- Matrix factorization, decomposition and approximation
- Eigenvalue computation
- Nonlinear optimization: theory and algorithms
- Least-squares problems and data fitting
- MATLAB and other software tools (lab sessions with applications)

Computational Modelling of complex systems [6 CFU] (non attivato)

The objective of this course is to train experts in systems modelling and analysis methodologies. Of course, this will require understanding, to some degree of detail, the mathematical and computational techniques involved. However, this will be done with the aim of shaping good modellers, that know the advantages/disadvantages/risks of the different modelling and analysis methodologies, that are aware of what happens under the hood of a modelling and analysis tool, and that can develop their own tools if needed.

The course will focus on advanced modelling approaches that combine different paradigms and analysis techniques: from ODEs to stochastic models, from simulation to model checking. Case studies from population dynamics, biochemistry, epidemiology, economy and social sciences will be analysed. Moreover, synergistic approaches that combine computational modelling techniques with data-driven methodologies will be outlined.

- *Modelling with ODEs: examples*
- *(Timed and) Hybrid Automata: definition and simulation techniques*
- *Stochastic simulation methods (Gillespie's algorithm and its variants)*
- *Hybrid simulation methods (stochastic/ODEs)*
- *Rule-based modelling*
- *Probabilistic/stochastic model checking: principles, applicability and tools*
- *Statistical model checking*
- *Process mining (basic notions)*

Scientific and large data visualization [6 CFU]

Scientific Visualisation is an area concerned with the visualisation of large and complex data sets, where the data might come from experiments or computations. Visualisation is a way, in many cases the only possible way, to achieve insight and knowledge inside large structured amount of data. The course will discuss discrete models for data representation in low dimensional spaces, scalar and vectorial data in 2D, 3D and for temporal series and algorithms for processing and visualizing massive datasets.

- *Rendering algorithms for massive volume data.*
- *Volume rendering: ray-tracing, splatting, texture based. Isosurface reconstruction.*
- *Transformation of discrete volume data to polygonal representations.*
- *Sampling and simplification techniques*
- *Multiresolution methods for massive 3D visualization*