Erasmus Mundus Joint Master Degree

Big Data Management and Analytics

Detailed Course Description

Academic Year 2017-2018

Contents

University: Université Libre de Bruxelles (ULB)
Department: Faculté des Sciences Appliquées
Course ID: ADB (INFO-H-415)
Course name: Advanced Databases
Name and email address of the instructors: Esteban Zimányi (ezimanyi@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh415
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:

Today, databases are moving away from typical management applications, and address new application areas. For this, databases must consider (1) recent developments in computer technology, as the object paradigm and distribution, and (2) management of new data types such as spatial or temporal data. This course introduces the concepts and techniques of some innovative database applications

Learning outcomes:

At the end of the course students are able to

- Understand various different technologies related to database management system
- Understand when to use these technologies according to the requirements of particular applications
- Understand different alternative approaches proposed by extant database management systems for each of these technologies
- Understand the optimization issues related to particular implementation of these technologies in extant database management systems.

Readings and text books:

- R.T. Snodgrass. *Developing Time-Oriented Database Applications in SQL*, Morgan Kaufmann, 2000
- Jim Melton, Alan R. Simon. *SQL: 1999 - Understanding Relational Language Components*, Morgan Kaufmann, 2001
- Jim Melton. *Advanced SQL: 1999 - Understanding Object-Relational and Other Advanced Features*, Morgan Kaufmann, 2002
- Shashi Shekhar, Sanjay Chawla. *Spatial Databases: A Tour*, Prentice Hall, 2003.

Prerequisites:

- Knowledge of the basic principles of database management, in particular SQL

Table of contents:

- Active Databases
  Taxonomy of concepts. Applications of active databases: integrity maintenance, derived data, replication. Design of active databases: termination, confluence, determinism, modularisation.
- Temporal Databases
  Temporal data and applications. Time ontology. Conceptual modeling of temporal aspects. Manipulation of temporal data with standard SQL. New temporal extensions in SQL 2011.
- Object-Oriented and Object-Relational Databases
  Object-oriented model. Object persistance. ODMG standard: Object Definition Language and Object Query Language. .NET Language-Integrated Query: Linq.
  Object-relational model. Built-in constructed types. User-defined types. Typed tables. Type and table hierarchies. SQL standard and Oracle implementation.
- Spatial Databases
  Application domains of Geographical Information Systems (GIS), common GIS data types and analysis. Conceptual data models for spatial databases. Logical data models for spatial databases: rastor model (map algebra), vector model (OGIS/SQL1999). Physical data models for spatial databases: Clustering methods (space filling curves), storage methods (R-tree, Grid files).

Assessment breakdown:

75% written examination, 25% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: Faculté des Sciences Appliquées
Course ID: DBSA (INFO-H-417)
Course name: Database Systems Architecture
Name and email address of the instructors: Stijn Vansummeren (Stijn.vansumeren@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh417
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 12h.
- Projects: 24h.

Goals:

In contrast to a typical introductory course in database systems where one learns to design and query relational databases, the goal of this course is to get a fundamental insight into the implementation aspects of database systems. In particular, we take a look under the hood of relational database management systems, with a focus on query and transaction processing. By having an in-depth understanding of the query-optimisation-and-execution pipeline, one becomes more proficient in administering DBMSs, and hand-optimising SQL queries for fast execution.

Learning outcomes:

Upon successful completion of this course, the student:

- Understands the workflow by which a relational database management systems optimises and executes a query
- Is capable of hand-optimising SQL queries for faster execution
- Understands the I/O model of computation, and is capable of selecting and designing data structures and algorithms that are efficient in this model (both in the context of datababase systems, and in other contexts).
- Understands the manner in which relational database management systems provide support for transaction processing, concurrency control, and fault tolerance

Readings and text books:

- Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom. *Database Systems: The Complete Book*, Prentice Hall, second edition, 2008.
- Raghu Ramakrishnan, Johannes Gehrke. *Database Management Systems*. McGraw-Hill, third edition, 2002.

Prerequisites:

- Introductory course on relational databases, including SQL and relational algebra
- Course on algorithms and data structures
- Knowledge of the Java programming language

Table of contents:

- Query Processing
  With respect to query processing, we study the whole workflow of how a typical relational database management system optimises and executes SQL queries. This entails an in-depth study of:
  - translating the SQL query into a “logical query plan";
  - optimising the logical query plan;
  - how each logical operator can be algorithmically implemented on the physical (disk) level, and how secondary-memory index structures can be used to speed up these algorithms; and
  - the translation of the logical query plan into a physical query plan using cost-based plan estimation.
- Transaction Processing
  - Logging
  - Serializability
  - Concurrency control

Assessment breakdown:

70% written examination, 30% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: Faculté des Sciences Appliquées
Course ID: DW (INFO-H-419)
Course name: Data Warehousing
Name and email address of the instructors: Toon Calders (toon.calders@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh419
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:

Relational and object-oriented databases are mainly suited for operational settings in which there are many small transactions querying and writing to the database. Consistency of the database (in the presence of potentially conflicting transactions) is of utmost importance. Much different is the situation in analytical processing where historical data is analyzed and aggregated in many different ways. Such queries differ significantly from the typical transactional queries in the relational model:

1. Typically analytical queries touch a larger part of the database and last longer than the transactional queries;
2. Analytical queries involve aggregations (min, max, avg, ...) over large subgroups of the data;
3. When analyzing data it is convenient to see it as multi-dimensional.

For these reasons, data to be analyzed is typically collected into a data warehouse with Online Analytical Processing support. Online here refers to the fact that the answers to the queries should not take too long to be computed. Collecting the data is often referred to as Extract-Transform-Load (ELT). The data in the data warehouse needs to be organized in a way to enable the analytical queries to be executed efficiently. For the relational model star and snowflake schemes are popular designs. Next to OLAP on top of a relational database (ROLAP), also native OLAP solutions based on multidimensional structures (MOLAP) exist. In order to further improve query answering efficiency, some query results can already be materialized in the database, and new indexing techniques have been developed.

The first and largest part of the course covers the traditional data warehousing techniques. The main concepts of multidimensional databases are illustrated using the SQL Server tools. The second part of the course consists of advanced topics such as data warehousing appliances, data stream processing, data mining, and spatial-temporal data warehousing. The coverage of these topics connects the data warehousing course with and serves as an introduction towards other related courses in the program. Several associated partners of the program contribute to the course in the form of invited lectures, case studies, and "proof of technology" sessions.

Learning outcomes:

At the end of the course students are able to

- Explain the difference between operational databases and data warehouses and the necessity of maintaining a dedicated data warehousing system
- Understand the principles of multidimensional modeling
- Design a formal conceptual multidimensional model based on an informal description of the available data and analysis needs
- Implement ETL-scripts for loading data from operational sources a the data warehouse
- Deploy data cubes and extract reports from the data warehouse
- Explain the main technological principles underlying data warehousing technology such as indexing and view materialization.

Readings and text books:

- Matteo Golfarelli, Stefano Rizzi. *Data Warehouse Design: Modern Principles and Methodologies.* McGraw-Hill, 2009
- Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen. *Multidimensional Databases and Data Warehousing.* Morgan and Claypool Publishers, 2010
- Ralph Kimball, Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, third edition, Wiley, 2013.
- Esteban Zimányi, Alejandro Vaisman. *Data Warehouse Systems: Design and Implementation.* Springer,

2014

Prerequisites:

- A first course on database systems covering the relational model, SQL, entity-relationship modelling, constraints such as functional dependencies and referential integrity, primary keys, foreign keys.
- Data structures such as binary search trees, linked lists, multidimensional arrays.

Table of contents:

There is a mandatory project to be executed in three steps in groups of 3 students, using the tools learned during the practical sessions, being SQL Server, SSIS, SSAS, and SSRS. Below is the succinct summary of the theoretical part of the course:

- Foundations of multidimensional modelling
- Dimensional Fact Model
- Querying and reporting a multidimensional database with OLAP
- Methodological aspects for data warehouse development
- Populating a data warehouse: The ETL process
- Using the data warehouse: data mining and reporting

Assessment breakdown:

75% written examination, 25% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: Faculté des Sciences Appliquées
Course ID: BPM (INFO-H-420)
Course name: Business Process Management
Name and email address of the instructors: Toon Calders (toon.calders@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh420
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Assignments and project: 12h.

Goals:

This course introduces basic concepts for modeling and implementing business processes using contemporary information technologies. The first part of the considers the *modeling* of business processes, including the control flow, and the data and resource perspectives. *Petri nets* will be used as a theoretical underpinning to formalize the different workflow patterns and unambiguously define the semantics of the different constructions in the workflow modelling languages. The workflow languages *Yet-another-workflow-Language* (YAWL) and the *Business Process Modelling and Notation* (BPMN) will be introduced in detail, as well as the main characteristics of the *Business Process Execution Language* (BPeL) for the composition of web services, and *Event-Driven Process Chains* (EPCs).

The second part of the course then goes into the analysis, simulation, verification, and discovery of workflows. Static techniques to verify properties such as soundness and the option-to-complete at model level will be studied, as well as dynamic properties such as the compliance of an event log with respect to a given model. For the discovery of workflows, an overview of the main process mining techniques will be discussed.

During the course the students have to perform a couple of modelling assignments in YAWL and BPMN. In the final project, students build a prototype system enacting one of the workflow modelled in their modelling assignments.

Affiliated industrial partners of the Erasmus Mundus project will be involved in the course in the form of invited lectures, case studies, and "proof of technology" sessions. These lectures complement the academic coverage of the topic with a more business-oriented perspective and form a nice addition to provide a more complete picture of the Business Processing Modeling landscape.

Learning outcomes:

At the end of the course students are able to

- Explain the business process management cycle;
- Recognize and report the value and benefit as well as the limitations of the automation and management of a concrete business process for an organization;
- Identify business processes amenable for automation in a concrete business context;
- Design a formal model of the business process based on an informal description;
- Evaluate the correctness of a specification using formal analysis and simulation;
- Implement the formal description of a model into a business process management tool, including resource management;
- Analyze an existing business process management solution, and, based on this analysis, propose optimizations to improve its performance.

Readings and text books:

- Mathias Weske. *Business Process Management: Concepts, Languages, Architectures*. Springer. 2007
- Arthur H. M. ter Hofstede, Wil M. P. van der Aalst, Michael Adams, Nick Russell (editors). *Modern Business Process Automation: YAWL and its Support Environment*. Springer, 2009.
- Wil van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer, 2012.

Prerequisites:

- Basic programming skills: variables, control structures such as loops and if-then-else, procedures, object-oriented notions such as classes and objects, ...
- Set theory (Notions such as set, set operations, sequence, multiset, function) and logics (mathematical notation and argumentation; basic proofs)

- Basic graph theory (notions such as graphs, reachability, transitivity, ...)
- Experience with modelling languages such as UML and ER diagrams is recommended.

Table of contents:
There is a mandatory project, split into several tasks during the whole period of the course offering, to be realized individually for the initial 3 assignments, and in group for the final assignment. The theoretical part of the course is dedicated to topics that allow the students to successfully carry out the project. Below is a high-level overview of the theoretical part of the course:

- Short overview of enterprise systems architecture and the place of business process management systems in it. The BPM life cycle.
- Modelling business processes: modelling the control flow, data and resource perspective.
- Enacting the business process models.
- Static and dynamic verification of process models; conformance checking.
- Discovering process models and other properties of processes through process mining.

Assessment breakdown:
50% oral examination, 50% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: Faculté des Sciences Appliquées
Course ID: DM (INFO-H-423)
Course name: Data Mining
Name and email address of the instructors: Hatem Haddad (hatem.haddad@ulb.ac.be)
Web page of the course: To be created
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:

Data Mining aims at finding useful regularities in large structured and unstructured data sets. The goal of this course is to get a fundamental understanding of popular data mining techniques strengths and limitations, as well as their associated computational complexity issues. It will also identify industry branches which most benefit from Data Mining such as health care and e-commerce. The course will focus on business solutions and results by presenting case studies using real (public domain) data. The students will use recent Data Mining software.

Learning outcomes:

At the end of the course students are able to

- Establish the main characteristics and limitations of algorithms for addressing data mining tasks.
- Select, based on the description of a data mining problem, the most appropriate combination of algorithms to solve it.
- Develop and execute a data mining workflow on a real-life dataset to solve a data-driven analysis problem.
- Use recent data mining software for solving practical problems.
- Identify promising business applications of data mining.

Readings and text books:

- David J. Hand, Heikki Mannila, Padhraic Smyth. *Principles of Data Mining*. MIT Press, 2001.
- Delmater Rhonda, Hancock Monte. *Data Mining Explained*. Digital Press, 2001.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining*. Pearson Education (Addison Wesley), 2006.

Prerequisites:

- Programming experience
- Data structures
- Algorithms, complexity

Table of contents:

- Data mining
  - Data mining and knowledge discovery
  - Data mining functionalities
  - Data mining primitives, languages, and system architectures
- Data preprocessing
  - Data cleaning
  - Data transformation
  - Data reduction
  - Discretization and generating concept hierarchies
- Data mining algorithms
  - Motivation and terminology
  - Different algorithm types
- Classification and Clustering
  - Classification: SVM classifier
  - Clustering: K-means, Latent Semantic Analysis

- Mining Associations and Correlations
  - Item sets
  - Association rules
  - Generating item sets and rules efficiently
  - Correlation analysis
- Advanced techniques, data mining software, and applications
  - Text mining: extracting attributes (keywords), structural approaches (parsing, soft parsing).
  - Bayesian approach to classifying text
  - Web mining: classifying web pages, extracting knowledge from the web
  - Recommendation systems
  - Data mining software and applications

Assessment breakdown:
75% written examination, 25% project evaluation

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: BDM
Course name: Big Data Management
Name and email address of the instructors: Alberto Abelló (aabello@essi.upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 30 h.
- Laboratories: 15 h.
- Self-Study: 105 h.

Goals:
The main goal of this course is to analyze the technological and engineering needs of Big Data Management. The enabling technology for such a challenge is cloud services, which provide the elasticity needed to properly scale the infrastructure as the needs of the company grow. Thus, building on top of the knowledge acquired in the course DBSA in the first semester, students will go deeper in advanced data management techniques (i.e., NOSQL solutions) that also scale with the infrastructure. Being Big Data Management the evolution of Data Warehousing, such knowledge is assumed in this course, which will specifically focus on the management of data Volume and Velocity.

On the one hand, to deal with high volumes of data, we will see how a distributed file system can scale to as many machines as necessary. Then, we will study different physical structures we can use to store our data in it. Such structures can be in the form of a file format at the operating system level, or at a higher level of abstraction. In the latter case, they take the form of either sets of key-value pairs, collections of semi-structured documents or column-wise stored tables. We will see that, independently of the kind of storage we choose, current highly parallelizable processing systems use a functional programming model (typically based on Map and Reduce functions), whose processing framework can rely on temporal files (like Hadoop MapReduce) or in-memory structures (like Spark).

On the other hand, to deal with high velocity of data we need some low latency system which processes either streams or micro-batches. However, nowadays, data production is already beyond processing technologies capacity. More data is being generated than we can store or even process on the fly. Thus, we will recognize the need of (a) some techniques to select subsets of data (i.e., filter out or sample), (b) summarize them maximizing the valuable information retained, and (c) simplify our algorithms to reduce their computational complexity (i.e., doing one single pass over the data) and provide an approximate answer.

Finally, the complexity of a Big Data project (combining all the necessary tools in a collaborative ecosystem), which typically involves several people with different backgrounds, requires the definition of a high level architecture that abstracts technological difficulties and focuses on functionalities provided and interactions between modules. Therefore, we will also analyse different software architectures for Big Data.

This course participates in a joint project conducted during the second semester together with VBP, SDM and CC. In VBP, the students will come up with a business idea related to Big Data Management and Analytics, which will be evaluated from a business perspective. In the three other courses the students have to implement a prototype meeting the business idea created. In CC they will be introduced to the main concepts behind large-scale distributed computing based on a service-based model and will have to choose the right infrastructure for their prototype. In BDM and SDM they will be introduce to specific data management techniques to deal with Volume and Velocity (BDM) and Variety (SDM). As final outcome, a working prototype must be delivered.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Knowledge
  - Understand the differences and benefits of in-memory data management.
  - Understand the execution flow of a distributed query.
  - Identify the difficulties of scalability and parallelization.
- Skills
  - Design a distributed database using NoSQL tools.
  - Produce a functional program to process Big Data in a cloud environment.
  - Manage and process a stream of data.
  - Design the architecture of a Big Data management system.

Readings and text books:

- M. Tamer Özsu, Patrick Valduriez. *Principles of Distributed Database Systems*, Springer, 2011.
- Ling Liu, M. Tamer Özsu. *Encyclopedia of Database Systems*, Springer, 2009.
- Pramod J. Sadalage, Martin Fowler. *NoSQL Distilled*, Addison-Wesley, 2013.
- Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira. *Hadoop Application Architectures*, O'Rilley, 2015.
- Hasso Plattner, Alexander Zeier. *In-Memory Data Management*, Springer, 2011.
- Matei Zaharia. *An Architecture for Fast and General Data Processing on Large Clusters*, ACM Press, 2016.
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of Massive Datasets*, http://www.mmds.org/
- Charu C. Aggarwal (Ed.). *Data Streams*, Springer, 2007.

Prerequisites:

- Advanced Databases (ADB)
- Database Systems Architecture (DBSA)
- Data Warehousing (DW)

Table of contents:

Assessment breakdown:
30% written examination, 35% problem classes and laboratories, 30% project, 5% peer marking

University: Universitat Politècnica de Catalunya
Department: Department of Service and Information System Engineering
Course ID: SDM
Course name: Semantic Data Management
Name and email address of the instructors: Oscar Romero (oromero@essi.upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 30 h.
- Laboratories: 15 h.
- Self-study: 105 h.

Goals:

Big Data is traditionally defined with the three V's: Volume, Velocity and Variety. Big Data has been traditionally associated with Volume (e.g., the Hadoop ecosystem) and recently Velocity has earned its momentum (especially, with the arrival of Stream processors such as Spark Streaming). However, even if Variety has been part of the Big Data definition, how to tackle Variety in real-world projects is yet not clear and there are no standarized solutions (such as Hadoop for Volume or Streaming for Velocity) for this challenge.

In this course the student will be introduced to advanced database technologies, modeling techniques and methods for tackling Variety for decision making. We will also explore the difficulties that arise when combining Variety with Volume and / or Velocity. The focus of this course is on the need to enrich the available data (typically owned by the organization) with external repositories (special attention will be paid to Open Data), in order to gain further insights into the organization business domain. There is a vast amount of examples of external data to be considered as relevant in the decision making processes of any company. For example, data coming from social networks such as Facebook or Twitter; data released by governmental bodies (such as town councils or governments); data coming from sensor networks (such as those in the city services within the Smart Cities paradigm); etc.

This is a new hot topic without a clear and well-established (mature enough) methodology. The student will learn about semantic-aware data management as the most promising solution for this problem. As such, it will be introduced to graph modeling, storage and processing. Special emphasis will be paid to semantic graph modeling (i.e., ontology languages such as RDF and OWL) and its specific storage and processing solutions.

This course participates in a joint project conducted during the second semester together with VBP, BDM and CC. In VBP, the students will come up with a business idea related to Big Data Management and Analytics, which will be evaluated from a business perspective. In the three other courses the students have to implement a prototype meeting the business idea created. In CC they will be introduced to the main concepts behind large-scale distributed computing based on a service-based model and will have to choose the right infrastructure for their prototype. In BDM and SDM they will be introduce to specific data management techniques to deal with Volume and Velocity (BDM) and Variety (SDM). As final outcome, a working prototype must be delivered.

Learning outcomes:

The student will learn models, languages, methods and techniques to cope with Variety in the presence of Volume and Velocity. Specifically:

- Graph modeling, storage and processing,
- Semantic Models and Ontologies (RDF, RDFS and OWL),
- Logics-based foundations of ontology languages (Description Logics) and
- Specific storage and processing techniques for semantic graphs.

The students will learn how to apply the above mentioned foundations to automate the data management lifecycle in front of Variety. We will focus on semantic data governance protocols and the role of semantic metadata artifacts to assist the end-user.

Readings and text books:

- Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart, textitWeb Data Management, Cambridge University Press, 2011.
- Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, Peter F. Patel-Schneider, *The Description Logic Handbook*, Cambridge University Press, 2010.
- Sven Groppe, *Data Management and Query Processing in Semantic Databases*, Springer, 2011.

Prerequisites:

- Advanced Databases (ADB)
- Data Warehousing (DW)
- Database Systems Architecture (DBSA)

Table of contents:

- Introduction:
  - Variety and Variability. Definition. External (non-controlled) data sources. Semi-structured data models, non-structured data. Schema and data evolution.
  - The data management life-cycle. Challenges when incorporating external (to the organisation) semi-structured and non-structured data.
- Foundations:
  - Graph management: the graph data model, graph databases, graph processing.
  - Semantic-aware management: Ontology languages (RDF, RDFS, OWL). Logical foundations and Description Logics. Storage (triplestores). Processing (SPARQL).
- Applications:
  - Open Data. Linked Open Data.
  - The Load-First Model-Later paradigm. The Data Lake. Semantic annotations and the Semantic-Aware Data Lake paradigm. Semantic data governance.
  - Semantic-aware metadata artifacts to automate data integration, linkage and / or cross of data between heterogeneous data sources.

Assessment breakdown:
30% written examination, 35% exercises and laboratories, 30% project, 5% peer marking

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Computers Architecture
Course ID: CC
Course name: Cloud Computing
Name and email address of the instructors: Angel Toribio (angel.toribio@upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 36 h.
- Lab sessions: 18 h.
- Autonomous work: 94 h.

Goals:

Cloud computing is a service model for large-scale distributed computing based on a converged infrastructure and a set of common services over which applications can be deployed and run over the network. This course about Cloud computing is divided into two main parts, one centered in the key concepts behind the Cloud paradigm and another more practical that deals with the related technologies.

In the first part of this course, the students will learn the principles and the state of the art of large-scale distributed computing in a service-based model. They will look at how scale affects system properties, models, architecture, and requirements. In terms of principles, the course looks at how scale affects systems properties, issues (such as virtualization, availability, locality, performance and adaptation), system models (game-theoretic, economic, evolutionary, control, complexity), architectural models (multi-tier, cluster, cloud), environment and application requirements (such as fault tolerance, content distribution). The first part of this course also reviews the state of the art in resource management of Cloud environments (composed of different types of platforms and organization) to support current applications and their requirements.

In the second part of the course the students will gain a practical view of the current state of the art of the Cloud technology in order to implement a prototype that meets the business idea created for the VBP course. In this part the students will start creating a basic toolbox to getting started in the Cloud. That will prepare them to practice with APIs, the doors in the Cloud. All these things together will allow the students to mining the deluge of data coming from the Cloud or use new advanced analytics services provided nowadays by the Cloud. Finally we will see under the hood of these advanced analytics services in the Cloud, either in terms of software and hardware, to understand how their high performance requirements can be provided.

The goal is to help students become part of this profound transformation that is causing the Advanced Analytics over the Big Data that Cloud is already providing as a service, and encourage their desire to want to delve further into this exciting world of technology and become actively involved.

This course participates in a joint project conducted during the second semester together with VBP, BDM and SDM. In VBP, the students will come up with a business idea related to Big Data Management and Analytics, which will be evaluated from a business perspective. In the three other courses the students have to implement a prototype meeting the business idea created. In CC they will be introduced to the main concepts behind large-scale distributed computing based on a service-based model and will have to choose the right infrastructure for their prototype. In BDM and SDM they will be introduced to specific data management techniques to deal with Volume and Velocity (BDM) and Variety (SDM). As final outcome, a working prototype must be delivered.

Learning outcomes:

Upon successful completion of this course, the student is able to:

- Capability to understand models, problems and algorithms related to distributed systems, and to design and evaluate algorithms and systems that process the distribution problems and provide distributed services.
- Capability to understand models, problems and mathematical tools to analyze, design and evaluate computer networks and distributed systems.
- Capability to analyze, evaluate, design and manage hardware/software of current cloud systems and services.
- Understand how can be used and how can be provided by Cloud systems, the new wave of Advanced Analytics services that nowadays companies and start-ups are requiring.

Readings and text books:

- Research and industrial papers.
- Software documentation.
- API documentation.
- Book chapters.

- Hands-on documentation.

Prerequisites:

- Knowledge about operating systems and networks (including Linux Shell).
- Knowledge about performance measurement methods.
- Programming experience in mainstream programming languages.

Table of contents:

Part I: Cloud Computing fundamentals

- Fundamental concepts: The effect of scale in system properties.
- Issues in large-scale systems: virtualization, service orientation and composition, availability, locality, performance and adaptation.
- Models for large-scale systems: system models for analysis, architectural models and service/deployment models.
- Scaling techniques: basic techniques, scalable computing techniques for architectural models.
- Middleware and Applications: computing, storage, web, content distribution, Internet-scale systems or services.
- Environment and applications requirements.

Part II: Practical view of Cloud Computing

- Big Data Analytics in the Cloud
- APIs: The Doors in the Cloud
- Current required layers in a Big Data Software Stack
- New Software requirements for Advanced Analytics
- New Hardware requirements for Advanced Analytics

Part III: Guest Lectures

- Invited lectures from the industry and academia will illustrate and describe specific aspects. The content will vary depending on visits and availability of invited speakers.

Part IV: Experimental part

- Development of a prototype application using Cloud service offerings such as AWS, Google AppEngine, Open Stack, OpenNebula.
- Development of a prototype application using advanced analytic services either provided in terms of APIs or Software as a Service.

Assessment breakdown:

- 25% Homeworks, continuous assessment and final exam.
- 35% Weekly lab sessions.
- 15% Course project.
- 25% Presentations and attendance.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Management
Course ID: VBP
Course name: Viability of Business Projects
Name and email address of the instructors: Marc Eguiguren (marc.eguiguren@upc.edu)
Web page of the course: To be announced. **Semester**: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 36 h.
- Projects in the classroom: 18 h.
- Projects: 56 h.
- Self-Study: 40 h.

Goals:

University graduates can find themselves in the situation of having to analyse or take on the project of starting their own business. This is especially true in the case of computer scientists in any field related to Big Data Management (BDM) or more generally, in the world of services. There are moments in one's professional career at which one must be able to assess or judge the suitability of business ventures undertaken or promoted by third parties, or understand the possibilities of success of dealing with a Big Data based service. It is for this reason that this subject focuses on providing students with an understanding of the main techniques used in analysing the viability of new business ventures: business start-up or the implementation of new projects in the world of services based on BDM. This project-oriented, eminently practical subject is aimed at each student's being able to draft as realistic a business plan as possible.

This course participates in a joint project conducted during the second semester together with VBP, BDM and CC. In VBP, the students will come up with a business idea related to Big Data Management and Analytics, which will be evaluated from a business perspective. In the three other courses the students have to implement a prototype meeting the business idea created. In CC they will be introduced to the main concepts behind large-scale distributed computing based on a service-based model and will have to choose the right infrastructure for their prototype. In BDM and SDM they will be introduce to specific data management techniques to deal with Volume and Velocity (BDM) and Variety (SDM). As final outcome, a working prototype must be delivered.

Learning outcomes:

Upon successful completion of this course, the student is able to:

- Being able to analyze the external situation to determine business innovative ideas in the field of BDM
- Around an innovative BDM project, being able to build a reasonable and ethically solid business plan
- Building a solid and convincing speech about a business idea and a business plan
- Training the students to build a P&L forecast and a forecasted treasury plan for a starting company
- Understanding and being able to apply the different instruments to finance the company, both debt instruments or private equity and venture capital sources
- Understand and appreciate the role of the entrepreneur in modern society

Readings and text books:

- Rhonda Abrams, Eugène Kleiner. *The Successful Business Plan.* The Planning Shop, 2003.
- Rob Cosgrove. *Online Backup Guide for Service Providers*, Cosgrove, Rob, 2010.
- Peter Drucker. *Innovation and Entrepreneurs.* Butterworth-Heinemann, Classic Drucker Collection edition, 2007.
- Robert D. Hisrich, Michael P. Peter, Dean A. Shepherd. *Entrepreneurship.* Mc Graw Hill, 6^{th} Ed., 2005.
- Mike McKeever. *How to Write a Business Plan.* Nolo, 2010.
- Lawrence W. Tuller. *Finance for Non-Financial Managers and Small Business Owners.* Adams Business, 2008.

Other Literature:

- M. Eguiguren, E. Barroso. *Empresa 3.0: políticas y valores corporativos en una cultura empresarial sostenible.* Pirámide , 2011.
- M. Eguiguren, E. Barroso. *Por qué fracasan las organizaciones: de los errores también se aprende.* Pirámide, 2013.

Prerequisites:

Having some previous knowledge or experience in business administration is an additional asset.

Table of contents: This course focuses on developing a BDM service oriented business plan. So that, the students are expected to reuse and consolidate any previous knowledge on databases, software engineering and BDM obtained in previous courses to develop a comprehensible, sustainable and profitable business.

The course is structured in 14 well-defined stages:

- Introduction to the course and key aspects of a business idea
- The entrepreneur's role in society, characteristics and profile
- Innovation and benchmarking Axis 1) Identification of long-term market megatrends
- Innovation and benchmarking axis 2) Big Data evolution as a source of ideas. Technology applied to industry.
- Axis of innovation and benchmarking 3) ethical business models as a source of innovation and ideas
- From the idea to the company. Contents of the business plan. Market research.
- Competitive advantages. SWOT Analysis
- Marketing plan: strategic marketing for a BDM service company, distribution and product
- Marketing plan: price and promotion strategies
- The human team in a small innovative company
- Different kind of societies. Fiscal basics for entrepreneurs
- Need of resources. Building the balance sheet at the beginning of the company
- Building a forecasted P&L for the first two years. Cash-Flow
- Revising the initial balance sheet and building the forecasted balance sheet for year one
- Treasury plan, Identifying long and short term financial needs
- Conventional long and short term financial instruments
- Private equity: founders, fools, friends & family, venture capital. Their limitations. Cautions to be taken and how they work.
- Presenting the plan to possible simulated or real investors

The business plan is expected to be partially developed in internal activities (under supervision of the teacher), and in external activities, always as teamwork (with no supervision).

Assessment breakdown:
The assessment is based on student presentations and the defence of the business plan before a jury comprising course faculty members and - optionally - another member of the teaching staff or guest professionals such as business angels, investors and successful IT enterpreneurs.

Throughout the course there will be four evaluative milestones:

- presentation of the innovative business model,
- presentation of the marketing plan,
- presentation of the business plan as a whole, that will include an evaluation about ethics and sustainability of the project together with SEAIT,
- analysis of the financial plan and the proposal to investors.

The presentation simulates a professional setting. Accordingly, the following aspects will also be assessed: dress, formal, well-structured communication, etc.

In order to be able to publicly defend the business plan, students must have attended at least 70% of the classes and teams must have delivered on time the activities that have been planned during the course. The plan is the result of teamwork, which will be reflected in the grade given to the group as a whole. Each member of the group will be responsible for part of the project and will be graded individually on his or her contribution.

This approach is designed to foster teamwork, in which members share responsibility for attaining a common objective.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: BDS
Course name: Big Data Seminar
Name and email address of the instructors: Oscar Romero (oromero@essi.upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 2

Course breakdown and hours:

- Lectures: 21 h.
- Autonomous work: 29 h.

Goals:
The students will be introduced to recent trends in Big Data. Seminars will be lectured by guest speakers, who will present business cases, research topics, internships and master's thesis subjects. Also, the three specialisations will be presented and discussed with the students within the seminars umbrella. Students will also perform a state-of-the art research in one topic, which will be presented and jointly evaluated by all partners in the mandatory eBISS summer school.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Read and understand scientific papers
- Develop critical thinking when assessing scientific papers
- Write and explain a state-of-the-art in a rigorous manner
- Elaborate on recent trends in Big Data

Readings and text books:

- Gordana Dodig-Crnkovic, *Theory of Science*, online resource: http://www.idt.mdh.se/kurser/ct3340/archives/ht04/theory_of_science_compendiu.pdf
- Jennifer Widom, *Tips for Writing Technical Papers*, online resource: https://cs.stanford.edu/people/widom/paper-writing.html
- Robert Siegel, *Reading Scientific Papers*, online resource: https://web.stanford.edu/~siegelr/readingsci.htm

Prerequisites:

- No prerequisites

Table of contents:
The seminars content will vary from course to course as they will focus on current hot topics in Big Data.

Assessment breakdown:
50% Seminar evaluations, 50% Poster presentation and state-of-the-art document
Note: Attending the eBISS summer school is a mandatory part of this course

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: DEBD
Course name: Debates on Ethics of Big Data
Name and email address of the instructors: Alberto Abelló (aabello@essi.upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 2

Course breakdown and hours:

- Lectures: 24 h.
- Autonomous work: 26 h.

Goals:
In this course we debate the impact on society of new advances in Big Data. We focus on ethics and the impact of such approaches on society. This course fosters the social competences of students, by building on their acquired oral communication skills to debate about concrete problems involving ethical issues in Big Data. The aim is to start developing their critical attitude and reflection. A written summary of their position is meant to train their writing skills.

During the course sessions the debates discussing innovative and visionary ideas on Big Data will take place. You must read the available material before the debate. Then, during the debate you will be assign to a group: either to defend an idea, or go against it. You may also be asked to moderate the debate. Then, the debate takes place and afterwards, each group needs to write down a report with their conclusions.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Ability to study and analyze problems in a critical mood
- Ability to critically read texts
- Develop critical reasoning with special focus on ethics and social impact
- Develop soft skills to defend - criticize a predetermined position in public
- Improve the writing skills

Readings and text books:

- Rudi Volti, *Society and technological change*, Worth, 2009.
- Richard T. De George, *The Ethics of information technology and business*, Blackwell, 2003.
- David Elliott, *Energy, society, and environment: technology for a sustainable future*, Routledge, 2003.
- Kord Davis, *Ethics of big data*, O'Reilly, 2013.

Prerequisites:

- No prerequisites

Table of contents:
Debates on ethics and social impact of Big Data (some examples):

- Ethics codes
- Right to privacy
- Content piracy
- Social responsibility of IT companies
- Artificial Intelligence and their limits

Assessment breakdown:
80% Debate evaluations, 20% Ethical analysis of a data-based business idea

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: SDS
Course name: Scalable Data Science
[TUB id & course name: AIM-3, Scalable Data Science: Systems & Methods (SDSSM)]
Name and email address of the instructors: Volker Markl (sekr@dima.tu-berlin.de)
Web page of the course: http://www.dima.tu-berlin.de
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Plenary sessions: 60 hrs
- Exercises / practice: 60 hrs
- Preparation & consolidation (incl. literature studies): 60 hrs

Goals:

This course introduces various parallel processing paradigms. The module will focus on mainstream distributed processing platforms and paradigms and learn how to employ these to solve challenging big data problems using popular data mining methods. Students will learn how to implement and employ varying data mining algorithms, such as Naive Bayes, K-Means Clustering, and Page Rank on varying open-source systems (e.g., Apache Hadoop, Apache Flink).

Learning outcomes:

The last decade was marked by the digitalization of virtually all aspects of modern society. Today, businesses, government institutions, as well as science and engineering organizations, among others face an avalanche of digital data on a daily basis. In order to derive insight from all of this data, society needs individuals with a strong foundation in scalable data science. In this course students will learn about popular scalable data analysis systems and scalable data analytics methods and gain practical experience in conducting scalable data science.

Readings and text books:

- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of massive datasets*, http://www.mmds.org/
- Ian H. Witten, Eibe Frank, Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2011.
- Tom White, *Hadoop: The Definitive Guide*, fourth edition, O'Reilly Media, 2015.
- Supplementary reading material may be assigned to complement course lectures.

Prerequisites:

- A database course adressed in a computer science Bachelor's curriculum, like "Information Systems and Data Analysis" or an equivalent, as well as good Java programming skills.
- Basic knowledge in linear algebra, numerical analysis, probability, and statistics.
- It is preferable if students have already completed a machine-learning course.

Table of contents:

Assessment breakdown:

- Homework (30%)
- In-class presentations (20%)
- Written test (50%)

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: MHI
Course name: Management of Heterogeneous Information
[TUB id & course name: AIM-1, Heterogeneous and Distributed Information Systems (HDIS)]
Name and email address of the instructors: Ralf-Detlef Kutsche (sekr@dima.tu-berlin.de)
Web page of the course: http://www.dima.tu-berlin.de
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Final report & presentation: 30 hrs
- Lab exercises including project tasks: 60 hrs
- Plenary sessions of this integrated course: 60 hrs
- Preparation / Consolidation (incl. literature work and seminar presentation): 30 hrs

Goals:
This course addresses master students with a focus on database systems, information management and business intelligence, in order to achieve deep knowledge in modern distributed heterogeneous information infrastructures. Heterogeneity in data models, distributed data organization/software architectures as well as interoperability and middleware platforms for HDIS (FIS, P2P, CS, ...), and, finally their persistency concepts and services will be studied in deep detail. Additional support by semantic concepts (ontologies, ...), model/metamodel and metadata management guides the way towards model-based development of HDIS and their applications in industry and public services.

Learning outcomes:
Upon successful completion of this course, the participants of this course will achieve deep conceptual, methodical, technical and practical knowledge in requirements analysis, design, architecture and development of heterogeneous and distributed information systems. This includes firstly classical knowledge about federated databases and mediator-based information systems (tight or loose coupling wrt. the dimensions of distribution, heterogeneity and autonomy). Secondly, the students can handle different paradigms of heterogeneous information infrastructures and their management (e.g., FedIS, CS, P2P) and interoperability architectures ("middleware"). Finally, modern model-based concepts for the development, integration and evolution of arbitrary information infrastructures, and –under this conceptual frame– model, metamodel, and metadata management as well as semantic concepts will be brought into practical experience by a larger seminar project work.

The course is principally designed to impart: technical skills 50%, method skills 30%, system skills 10%, social skills 10%

Readings and text books:
For each topic during this course appropriate text books, journal and conference papers, PhD dissertations, technical reports, and industrial standards will be used.

Prerequisites:

- Basic modules in a Bachelor Curriculum in Computer Science, Computer Engineering or Business Information Technology
- Particularly knowledge in Database Systems/ Information Modeling and Software Engineering
- Programming skills

Table of contents:

- Foundations / Terminology of HDIS (FDBS, FIS, MBIS)
- Dimensions of HDIS: distribution, heterogeneity, autonomy
- Heterogeneous data models in HDIS: structured, semistructured, unstructured
- Distributed data organisation and software architectures of HDIS (FIS, P2P, CS, ...)
- Interoperability and middleware platforms for HDIS
- Persistency services
- Semantic concepts in HDIS
- Metadata standards and management in HDIS
- Model-based development of HDIS
- Applications from industry and public services

Assessment breakdown:

- Homework presentation (10%)
- Project work: software, documentation, presentation (25%)
- Seminar talk (25%)
- Written seminar report (40%)

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: MDS
Course name: Management of Data Streams
[TUB id & course name: AIM-2, Management of Data Streams]
Name and email address of the instructors: Volker Markl (sekr@dima.tu-berlin.de)
Web page of the course: http://www.dima.tu-berlin.de
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 60 hrs
- Exercises / Practice: 30 hrs
- Elaboration: 30 hrs
- Preparation & consolidation (incl. literature studies and oral presentation) : 60 hrs

Goals:
In this course students will acquire the theory and practical experience for analysing data streams, including windowing operations and data stream mining. They will experiment the combined analysis of data in motion and data at rest.

Learning outcomes:
Through the technological advances in the last few years more and more applications are being created that constantly generate data which is only relevant for a certain time frame. Because of this, this type of application has to be able to handle various streams of data. You will gain conceptual, methodological and practical skills in the area of processing data streams, by using examples from various application areas.

Readings and text books:

- Charu C. Aggarwal (Ed.). *Data Streams: Models and Algorithms*, Springer, 2007.
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of massive datasets*, http://www.mmds.org/

Prerequisites:

- Knowledge about modern modelling languages and about database management

Table of contents:
In recent years, advances in hardware technology have facilitated new ways of collecting data continuously. In many applications such as for instance network monitoring, the volume of such data is so large that it may be impossible to store the data on disk. Furthemore, even when the data can be stored, the volume of the incoming data may be so large that it may be impossible to process any particular record more than once. Therefore, many database operations and data analysis algorithms such as for instance filtering, indexing, classification and clustering become significantly more challenging in this context.
The course has the following main topics:

- Basic conceptual understanding and terminology of data flow management, introduction to data streams, the difference to classical data management, examples (telephone networks, automotive electronics, avionics, medical, transport management, building monitoring, etc.)
- Basic conceptual understanding and terminology of data flow management, introduction to data streams, the difference to classical data management, examples (telephone networks, automotive electronics, avionics, medical, transport management, building monitoring, etc.)
- Data sources, requirements elicitation, requirements structuring, requirements of data stream management systems (DSMS)
- Reference architecture of a DSMS, architecture modeling
- Modeling of the functionality, logical architecture. Description on technical architecture, interface definition, behavior modeling
- Windowing, The Sliding-Window Computation Model and Results
- Data processing in sensor networks, resource utilization, transmission and transfer costs
- Modeling examples (automotive electronics, avionics), prototype systems (Aurora, STREAM, TelegraphCQ), frameworks (Flink, Spark, Storm, Samza, SAMOA).

Assessment breakdown:
Oral exam (100%)

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: BDAP
Course name: Big Data Analytics Project
[TUB id & course name: BDAPRO, Big Data Analytics Project]
Name and email address of the instructors: Volker Markl (sekr@dima.tu-berlin.de)
Web page of the course: http://www.dima.tu-berlin.de
Semester: 3
Number of ECTS: 9

Course breakdown and hours:

- Documentations, presentations: 40 hrs
- Implementation, tests, experiments: 130 hrs
- Participation in meetings: 60 hrs
- Preparation phase and design: 40 hrs

Goals:
In this course you will learn to systematically analyze a current issue in the information management area and to develop and implement a problem-oriented solution as part of a team. You will learn to cooperate as team member and to contribute to project organization, quality assurance and documentation. The quality of your solution has to be proven through analysis, systematic experiments and test cases. Examples of projects carried out in recent semesters are a tool used to analyse Web 2.0 Forum data, an online multiplayer game for mobile phones, implementation and analysis of new join methods for a cloud computing platform or the development of data mining operations on the massively parallel system Hadoop as part of the Apache open source project Mahout.

Learning outcomes:
After the course, students will be able to understand methods for large-scale data analytics and to solve large-scale data analytics problems. They will be capable of designing and implementing large-scale data analytics solutions in a collaborative team.

Readings and text books:

- H. Garcia-Molina, J. D. Ullman, J. Widom: *Database Systems: The Complete Book*, Pearson 2009.
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of massive datasets*, http://www.mmds.org/

Prerequisites:

- Knowledge from a complete Bachelor program like computer science or computer engineering
- Knowledge in linear algebra and statistics
- Solid programming skills in at least one of the following programming languages: Java, C++, Scala, Python
- Basic knowledge in functional programming
- Basic knowledge in distributed source control management systems (Git, Mercurial) and software processes like Scrum

Table of contents:
Both the sciences and industry are currently undergoing a profound transformation: large-scale, diverse data sets - derived from sensors, the web, or via crowd sourcing - present a huge opportunity for data-driven decision making. This data poses new challenges in a variety of dimensions: in its unprecedented volume, in the speed at which it is generated (its velocity) and in the variety of data sources that need to be integrated. A whole new breed of systems and paradigms is currently developed to be able to cope with that these challenges.

The field of Big Data Analytics deals with the technological means of gaining insights from huge amounts of data. Students will conduct projects that deal with applying data mining algorithms to large datasets. For that, students will learn to use so called Parallel Processing Platforms (e.g. Flink, Spark, Haddop, HBase), systems that execute parallel computations with terabytes of data on clusters of up to several thousand machines.

At the start of the project, a student will receive a topic as well as some information material. The team, with the assistance of the lecturer, will decide on a project environment with the suitable tools for team work, project communication, development and testing. Next, the problem will have to be analyzed, modelled and decomposed into individual components, from which tasks are derived that are subsequently assigned to smaller teams or individuals. At weekly project meetings, the project team presents progress and milestones that have been reached. In consultation with the lecturer, it is decided which further steps to take. The project is concluded with a final report, a project poster as well as a final presentation which includes a

demonstration of the prototype.

Assessment breakdown:
The overall grade for the module consists of the results of the course work. The following are included in the final grade:

- Prototype with test cases and documentation (30%)
- Experiment design and execution (20%)
- Intermediate presentation (10%)
- Experiment analysis (20%)
- Final presentation (20%)

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: BDAS
Course name: Big Data Analytics Seminar
[TUB id & course name: BDASEM, Big Data Analytics Seminar]
Name and email address of the instructors: Volker Markl (sekr@dima.tu-berlin.de)
Web page of the course: http://www.dima.tu-berlin.de
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Plenary Sessions: 30 hrs
- Preparation / Individual work: 60 hrs

Goals:
Participants of this seminar will acquire knowledge about recent research results and trends in the analysis of web-scale data. Through the work in this seminar, students will learn the comprehensive preparation and presentation of a research topic in this field. In order to achieve this, students will get to read and categorise a scientific paper, conduct background literature research and present as well as discuss their findings.

Learning outcomes:
After the course, students will be able to critically read and evaluate scientific publications, and to conduct background research. They will be capable of preparing for and giving oral presentations on research topics for an expert audience, of analyzing the state of the art of a research topic, and of summarizing it in a scientific paper. They should also understand techniques used in the scientific community like peer reviews, conference presentations, and defenses of the findings after their presentation, as well as they should understand methods for large-scale data analytics.

Readings and text books:
At the beginning of the semester students will receive a set of primary literature, which consists of a basic item for every participant.

Prerequisites:

Table of contents:
Both the sciences and industry are currently undergoing a profound transformation: large-scale, diverse data sets - derived from sensors, the web, or via crowd sourcing - present a huge opportunity for data-driven decision making. This data poses new challenges in a variety of dimensions: in its unprecedented volume, in the speed at which it is generated (its velocity) and in the variety of data sources that need to be integrated. A whole new breed of systems and paradigms is currently developed to be able to cope with that these challenges. The field of Big Data Analytics deals with the technological means of gaining insights from huge amounts of data. In this seminar, students will review the current state of the art in this field.
Students will learn about presentation techniques and guidelines on how to read scientific papers. This is be extended by learning how to write texts specially in the context of the English language. Students should use secondary sources to research the topic assigned to them in the seminar, which should go beyond the supplied primary literature. Next to conventional sources like the internet students are required to use research journals and articles published at information management conferences such as WWW, VLDB, or SIGMOD.

Assessment breakdown:
The exam will be done as a 'portfolio examination', including two deliverable assessments:

- Presentation (50%)
- Written seminar report (50%)

University: Technische Universität Berlin (TUB)
Department: School of Electrical Engineering and Computer Science
Course ID: LSDADM
Course name: Large-Scale Data Analysis and Data Mining
Name and email address of the instructors: Volker Markl (sekr@dima.cs.tu-berlin.de)
Web page of the course: http://www.dima.cs.tu-berlin.de
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 30h.
- Exercises: 30h.

Goals:
The focus is of this module is to get familiar with different parallel processing platforms and paradigms and to understand their feasibility for different kinds of data mining problems. For that students will learn how to adapt popular data mining and standard machine learning algorithms such as: Naive Bayes, K-Means clustering, PageRank, Alternate Least Squares, or other methods of text mining, graph analysis, or recommender systems to scalable processing paradigms. Students will subsequently gain practical experience in how to analyse Big Data on parallel processing platforms such as the Apache Big Data Stack (Hadoop, Giraph, etc.), the Berkeley Big Data Analytics Stack (e.g., Spark) and Apache Flink.

Learning outcomes:
TO BE DONE

Readings and text books:

- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of massive datasets*, http://www.mmds.org/ http://www.mmds.org/
- Ian H. Witten, Eibe Frank, Mark A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*, third edition. Morgan Kaufmann, 2011.
- Tom White. *Hadoop: The Definitive Guide*, third edition. O'Reilly Media, 2012.

For each topic during this course additional research papers and reports will be used.

Prerequisites:

- Database Systems Architecture (DBSA)

Table of contents:
TO BE DONE

Assessment breakdown:
The grade of the module will be composed from the results of the presentation (50%) and the written seminar report (50%) for this presentation.

University: Technische Universiteit Eindhoven (TU/e)
Department: Department of Mathematics and Computer Science
Course ID: BIS
Course name: Business Information Systems
Name and email address of the instructors: Dirk Fahland (d.fahland@tue.nl)
Web page of the course: To be created
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 32 h.
- Instructions: 32 h.

Goals:
In this course students will learn about the modelling, analysis, and enactment of business processes and the information systems to support these processes, understanding the relationship between systems and processes.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Model complex (business) processes, systems and web-service choreographies in terms of classical and Colored Petri nets.
- Translate informal requirements into explicit models.
- Analyze processes (and the corresponding systems) using state-space analysis and invariants.
- Analyze and improve processes (and the corresponding systems) through simulation.
- Suggest redesigns of existing processes (and the corresponding systems).

Readings and text books:

- Wil van der Aalst, Christian Stahl. *Modeling Business Processes: A Petri Net-Oriented Approach*, MIT Press, 2011.

Prerequisites:

- Datamodelling and databases (recommended)
- Programming (recommended)
- Logic and set theory (recommended)
- Automata and process theory (recommended)

Table of contents:
Process-aware information systems (e.g., workflow management systems, ERP systems, CRM systems, PDM systems) are generic information systems that are configured on the basis of process models. In some systems the process models are explicit and can be adapted (e.g., the control-flow in a workflow system) while in other systems they are implicit (e.g., the reference models in the context of SAP). In some systems they are hard-coded and in other systems truly configurable. However, it is clear that in any enterprise, business processes and information systems are strongly intertwined. Therefore, it is important that students understand the relationship between systems and processes and are able to model complex systems involving processes, humans, and organizations.

Assessment breakdown:
70% Written final exam, 20% Assignment(s), 10% Interim examination
Participation in the intermediate exam and the CPN assignment are required for entrance to the written exam. The points for the intermediate exam and the assignment are retained for the retake exam.

University: Technische Universiteit Eindhoven (TU/e)
Department: Department of Mathematics and Computer Science
Course ID: IPM
Course name: Introduction to Process Mining
Name and email address of the instructors: Wil van der Aalst (w.m.p.v.d.aalst@tue.nl)
Web page of the course: To be created
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 12 h. (recorded on video)
- Instructions: 14 h.
- Videos and self-tests of the MOOC https://www.coursera.org/course/procmin are used

A form of blended learning is used: Students learn content online by watching video lectures and homework is discussed in class

Goals:
In this course students will acquire the theoretical foundations of process mining and will be exposed to real-life data sets helping them understand the challenges related to process discovery, conformance checking, and model extension.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Have a good understanding of process mining.
- Understand the role of data science in today's society.
- Relate process mining techniques to other analysis techniques such as simulation, business intelligence, data mining, machine learning, and verification.
- Apply basic process discovery techniques such as the alpha algorithm to learn a process model from an event log (both manually and using tools).
- Apply basic conformance checking techniques (such as token-based replay) to compare event logs and process models (both manually and using tools).
- Extend a process model with information extracted from the event log (e.g., show bottlenecks).
- Have a good understanding of the data needed to start a process mining project.
- Characterize the questions that can be answered based on such event data.
- Explain how process mining can also be used for operational support (prediction and recommendation).
- Use tools such as ProM and Disco.
- Execute process mining projects in a structured manner using the L* life-cycle model.

Readings and text books:

- Wil van der Aalst, *Process Mining: Data Science in Action*, Springer-Verlag, Berlin, 2016.
- Video lectures based on https://www.coursera.org/course/procmin.
- ProM, Disco, and RapidMiner are used for hands-on experiments.
- Provided slides, event logs, exercises, and additional papers.

Prerequisites:

- Basic computer science skills (e.g. bachelor in computer science, mathematics or industrial engineering).

Table of contents:
Data science is the profession of the future, because organizations that are unable to use (big) data in a smart way will not survive. It is not sufficient to focus on data storage and data analysis. The data scientist also needs to relate data to process analysis. Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining. Process mining seeks the confrontation between event data (i.e., observed behavior) and process models (hand-made or discovered automatically). This technology has become available only recently, but it can be applied to any type of operational processes (organizations and systems). Example applications include: analyzing treatment processes in hospitals, improving customer service processes in a multinational, understanding the browsing behavior of customers using a booking site, analyzing failures of a baggage handling system, and improving the user interface of an X-ray machine. All of these applications have in common that dynamic behavior needs to be related to process models. Hence, we refer to this as "data science in action".

The course covers the three main types of process mining. The first type of process mining is discovery. A

discovery technique takes an event log and produces a process model without using any a-priori information. An example is the a-algorithm that takes an event log and produces a Petri net explaining the behavior recorded in the log. The second type of process mining is conformance. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. The third type of process mining is enhancement. Here, the idea is to extend or improve an existing process model using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a-priori model. An example is the extension of a process model with performance information, e.g., showing bottlenecks. Process mining techniques can be used in an offline, but also online setting. The latter is known as operational support. An example is the detection of non-conformance at the moment the deviation actually takes place. Another example is time prediction for running cases, i.e., given a partially executed case the remaining processing time is estimated based on historic information of similar cases.

Process mining provides not only a bridge between data mining and business process management; it also helps to address the classical divide between "business" and "IT". Evidence-based business process management based on process mining helps to create a common ground for business process improvement and information systems development. The course uses many examples using real-life event logs to illustrate the concepts and algorithms. After taking this course, one is able to run process mining projects and have a good understanding of the data science field.

Assessment breakdown:
Written final exam, Assignments

- Final test Introduction to Process Mining
- Assignment 1 Introduction to Process Mining
- Assignment 2 Introduction to Process Mining

University: Technische Universiteit Eindhoven (TU/e)
Department: Department of Mathematics and Computer Science
Course ID: VIS
Course name: Visualization
Name and email address of the instructors: Michel Westenberg (m.a.westenberg@tue.nl)
Web page of the course: http://www.win.tue.nl/~mwestenb/edu/2IMV20_Visualization/
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 14 h.
- Practical work: 14 h.
- Video recordings are available as additional teaching material and are not supposed to replace the lectures.

Goals:
In this course students will learn the theory and practice of data visualization, including topics such as data representation, grid types, data sampling, data interpolation, data reconstruction, datasets, and the visualization pipeline.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Have a good knowledge of the theory, principles, and methods which are frequently used in practice in the construction and use of data visualization applications.
- Design, implement, and customize a data visualization application of average complexity in order to get insight in a real-world dataset from one of the application domains addressed during the lecture.
- Understand (and apply) the various design and implementation trade-offs which are often encountered in the construction of visualization applications.

Readings and text books:

- Lecture slides, selected papers, and assignments.

Prerequisites:

- Computer graphics (basic)
- Linear algebra
- Programming (some experience in Java or a similar language)

Table of contents:
The course covers the theory and practice of data visualization. This addresses several technical topics, such as: data representation; different types of grids; data sampling, interpolation, and reconstruction; the concept of a dataset; the visualization pipeline. In terms of visualization application, several examples are treated, following the different types of visualization data: scalar visualization, vector visualization, tensor visualization. Besides these, several additional topics are treated, such as volume data visualization and a brief introduction to information visualization. The techiques treated in the course are illustrated by means of several practical, hands-on, examples.

Assessment breakdown:
50% First assignment, 50% Second assignment
Two assignments covering the core topics of the course.

University: Technische Universiteit Eindhoven (TU/e)
Department: Department of Mathematics and Computer Science
Course ID: SBD
Course name: Statistics for Big Data
Name and email address of the instructors: Edwin van den Heuvel (e.r.v.d.heuvel@tue.nl)
Web page of the course: To be created
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 32 h.
- Practical work: 32 h.

Goals:
In this course students will learn various statistical methods for analysing BD, focusing on analysing temporal observational data, i.e., data that is collected over time without involving well-developed experimental designs.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Have a good knowledge of theoretical and practical aspects of mixed models for different type of outcomes (continuous, categorical, binary).
- Understand how model features will affect the correlations in outcomes and how bias in estimates are affected by confounders.
- Identify the difference between two classes of models: marginal and subject specific models.
- Apply the different methods of estimation (generalized estimating equations and maximum likelihood).
- Apply different approaches to handle missing data (like multiple imputation methods).
- Apply the methods on real data sets, with real problems using statistical software.

Readings and text books:

- Handouts
- Tutorial and overview papers from statistical literature

Prerequisites:

- Knowledge on probability theory and on modeling data with both linear and generalized linear models (preferably 2WS40 and 2WS70).

Table of contents:
The course discusses statistical models for analyzing relative large data sets with complex structured correlated data. Both individual participant data analysis and aggregate data analysis will be discussed. These methods are useful when data sets from different sources are being pooled for analysis. One important topic is how different but similar variables can be harmonized and/or standardized for such a pooled analysis. The family of models that will be typically discussed are mixed models. These models are not just relevant for big data alone, but also useful for small data sets. Missing data problems in relation to mixed models will be discussed in detail, since missing data can really influence conclusions if it is not addressed appropriately. We will also discuss issues that are related to selection bias (e.g. propensity scores and confounding) and causal inference.

Assessment breakdown:
Written final exam, Assignments
Homework assignments. Two assignments on the analysis of data sets with complex correlated data. Another assignment is about a simulation study to investigate the performance of the discussed theory on statistical modeling. Both content and reporting will be evaluated.

University: Technische Universiteit Eindhoven (TU/e)
Department: Department of Mathematics and Computer Science
Course ID: BPAS
Course name: Business Process Analytics Seminar
Name and email address of the instructors: H.A. Reijers (h.a.reijers@vu.nl)
Web page of the course: http://www.win.tue.nl/is/doku.php?id=education:2ii96
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 14 h.

Goals:
This seminar introduces students to research in business process analytics by following lectures by staff members and guest lecturers from industry, analysing master theses, study research papers, and execute a small research project.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Have a good knowledge about the AIS research topics.
- Analyze selected Master theses performed within the AIS group to learn about the research topics the group works on and to get a better idea about âĂIJpatternsâĂİ and âĂIJanti-patternsâĂİ in the execution of a Master project and writing a thesis. The results of this work will be given in a review of a Master thesis and a presentation about it.
- Use existing tools such as ProM, Disco, CPN Tools to work on their research project

Readings and text books:

- Various items (offered via the website or studyweb)

Prerequisites:

- Process modelling (recommended)
- Process mining (recommended)
- Business information systems (recommended)
- Data mining and knowledge systems (recommended)
- Business process management systems (recommended)
- Visualization (recommended)

Table of contents:
Organizations are constantly trying to improve the way their businesses perform. To this end, managers have to take decisions on changes to the operational processes. However, for decision making, it is of the utmost importance to have a thorough understanding of the operational processes as they take place in the organization.

Typically, such operational processes are supported by information systems that record events as they take place in real life in so-called event logs. Process Mining techniques allow for extracting information from event logs. For example, the audit trails of a workflow management system or the transaction logs of an enterprise resource planning system can be used to discover models describing processes, organizations, and products. Moreover, it is possible to use process mining to monitor deviations (e.g., comparing the observed events with predefined models or business rules in the context of SOX).

Process mining is closely related to BAM (Business Activity Monitoring), BOM (Business Operations Management), BPI (Business Process Intelligence), and data/workflow mining. Unlike classical data mining techniques the focus is on processes and questions that transcend the simple performance-related queries supported by tools such as Business Objects, Cognos BI, and Hyperion.

In this seminar, students are introduced to the research being conducted in the Architecture of Information Systems group of this university. Specifically, this seminar focuses on the state-of-the-art research on Process Mining. To emphasize that Process Mining is not only a research area, but has gained increasing interest from Industry, practical application will be considered.

Assessment breakdown:
Assignment(s)

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: DKQ
Course name: Data and Knowledge Quality
Name and email address of the instructors: Verónika Peralta (Veronika.peralta@univ-tours.fr), Béatrice Markhoff (beatrice.markhoff@univ-tours.fr)
Web page of the course:
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 24 h.
- Exercises: 26 h.
- Projects: 10 h.

Goals:
In this course, students will learn the concepts and techniques for assessing and assuring the quality of data and knowledge, by deeply studying key quality dimensions, methodologies, algorithms and specialized tools.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Understand quality terminology and methodologies
- Understand, identify and use major quality dimensions and metrics
- Understand major techniques for assessing and assuring quality
- Propose a quality plan for a given application
- Apply studied techniques for assessing and assuring quality using current tools

Readings and text books:

- Batini, C., Scannapieco, M.: Data and Information Quality: Dimensions, Principles and Techniques. Springer 2016.
- Bleiholder, J., Naumann, F.: "Data Fusion". ACM Computing Survey 2009.
- Lukoianova, Rubin: Veracity Roadmap: Is Big Data Objective, Truthful and Credible?. Advances In Classification Research Online, 24(1), 2014
- Ilyas, I. F., Chu, X.: Trends in Cleaning Relational Data: Consistency and Deduplication. In Foundations and Trends in Databases, Volume 5, Issue 4, 2015
- Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, Peter F. Patel-Schneider, The Description Logic Handbook, Cambridge University Press, 2010.

Prerequisites:

- A first course on database systems covering the relational model, Datalog, SQL, constraints (functional dependencies, primary keys, foreign keys)
- A first course on logics
- Data Warehouses (DW)
- Database Systems Architecture (DBSA)
- Semantic Data Management (SDM)

Table of contents:

- Quality concepts. Notion of data quality. Causes and consequences of bad data quality. Data quality in multi-source context. Data quality in big data context. Data and knowledge quality in semantic web context.
- Quality Dimensions. Quality multi-dimensionality. Key data quality dimensions (accuracy, completeness, freshness, consistency, uniqueness). Emergent quality dimensions (truthfulness, objectivity, credibility).
- Handling data quality. Assessment. Diagnosis. Correction. Prevention
- Integration and Fusion.
- Data integration process. Duplicate detection. Data fusion.
- Methodology Terminology, Goal-question-metric approach. Data profiling.
- Knowledge representation for the semantic web. Description Logics and Datalog+/-. Ontology-based data integration.
- Query result validity Querying the semantic web. Inconsistency-tolerant query answering. Dealing with validity constraint on query results.

- Provenance

Assessment breakdown:
75% written examination, 25% project evaluation

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: UCA
Course name: User-Centric Approaches
Name and email address of the instructors: Patrick Marcel (Patrick.Marcel@univ-tours.fr), Nicolas Labroche (Nicolas.Labroche@univ-tours.fr), Verónika Peralta (Veronika.Peralta@univ-tours.fr), Gilles Venturini (Gilles.Venturini@univ-tours.fr),
Web page of the course:
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 20 h.
- Exercises: 10 h.
- Projects: 20 h.

Goals:
In this course, students will learn the models and approaches needed for user-centric data analysis. They will study preference models and processing, query personalization, recommender systems, and gain practical experience with platforms like Apache Mahout and Weka.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Understand the major theoretical and practical approaches for modeling user interests and preferences
- Understand and apply the major approaches to extract, process and rank according to user preferences and interests
- Identify the major approaches for visual and interactive data exploration
- Develop and tune user-centric applications for exploring, searching and analyzing data sets

Readings and text books:

- Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman. *Mining of massive datasets*, http://www.mmds.org/
- Amy N. Langville, Carl D. Meyer. *Who's #1?: The Science of Rating and Ranking.* Princeton University Press, 2012.

Prerequisites:

- Data Warehouses (DW)
- Data Mining (DM)
- Semantic Data Management (SDM)
- Viability of Business Projects (VBP)

Table of contents:

- User representation
  User models, order basics, preference models: quantitative and qualitative approaches, preference combination, conditional preferences, CP-nets
- Preference extraction
  Implicit feedback, user engagement, trace analysis, preference mining, intent learning, interest learning
- Ranking based on user interests
  Query personalization, query expansion, recommender systems: content-based, collaborative, SVD
- Interactive data exploration and visual mining
  Visual perception, multidimensional data (Chernoff faces, Stick Icons, Scatter plots, Parallel coordinates, Star Plot, Kohonen maps, Radial visualizations, proximity graphs, etc.), hierarchical data and graphs (trees, Force-directed algorithms), temporal data
- Evaluation of user-centric approaches
  Evaluation strategy, metrics for user-centric approaches, benchmarks
- Use cases
  Web usage mining, preference in databases, recommendation for interactive data exploration, intelligent personal assistants

Assessment breakdown:
75% written examination, 25% project evaluation

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: NLP
Course name: Natural Language Processing
Name and email address of the instructors: Agata Savary (Agata.Savary@univ-tours.fr)
Web page of the course:
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 20 h.
- Exercises: 10 h.
- Projects: 20 h.

Goals:

Computers can do many things faster and more efficiently than humans can, and efficient processing of Big Data is a notable exemplification of this fact. However, there is one particular area where humans have the upper hand, and that's language. Despite recent advances in artificial intelligence, a deep understanding of human language (also called *natural language*) remains a rather distant perspective for computers. This is because computers, in order to be efficient and accurate, need to rely on highly structured, non-ambiguous and explicit data, while the very nature of the human language is to be unstructured, highly ambiguous and often implicit. This fact is the major challenge in a fully efficient exploitation of huge quantities of textual data available in knowledge bases.

Still, the domains of *natural language processing* (NLP) and *computational linguistics* (CL) have made substantial progress in coping with relatively well delimited subproblems of language understanding. Thus, they can contribute to accounting for the challenging aspects of Big Data, such as Variety and Variability, *inter alia*.

The goals of this module are to:

- gain the overall understanding of the unstructured, ambiguous and implicit **nature of the** natural language **data**, and challenges that they pose for computing,
- discover the foundations of **linguistic modeling** usable in NLP applications nowadays,
- understand **basic NLP techniques** underlying NLP processing chains,
- discover state-of-the-art **results** in NLP applications (what is doable and what is too hard on a large scale?),
- understand the **usefulness** of NLP tools **for Big Data**,
- study selected **advanced NLP techniques** in more detail.

Learning outcomes:

Upon successful completion of this course, the student is able to:

- Understand the nature of natural language data and their modeling,
- Manipulate some existing NLP tools, with an understanding of the underlying processes involved,
- Develop sample NLP applications to analyze textual data.

Readings, text books and video lectures:

- Jurafsky, Martin (2009) Speech and Language Processing, Pearson
- Jurafsky, Manning Natural Language Processing, Stanford lecture
- Building Watson - A Brief Overview of the DeepQA Project
- IBM Watson: The Science Behind an Answer
- Speech recognition and machine translation at Microsoft
- POWERSET: Natural Language and the Semantic Web
- NLP at Google
- GATE: Hands-on Natural Language Processing for Information Access Applications

Prerequisites:

- Data Mining (DM)
- Formal languages and compilers (recommended).

Table of contents:

- NLP applications in Big Data: opinion mining, information retrieval, information extraction, ontology extraction and enrichment,

- Nature of language data, foundations of linguistic modeling,
- Key techniques for textual data (tokenizsation, POS tagging),
- State-of-the-art NLP techniques heavily used in applications: named entity recognition, entity linking, shallow parsing,
- Limits of SOA techniques and introduction to more advanced techniques: deep parsing.

This course has some strong links with the parallel course on Advanced Data Mining (semester 3): mining text data and the underlying techniques such as HMM, CRF, LSA, deep learning, neural networks, etc.

Assessment breakdown:
75% written examination, 25% project evaluation

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: ADM
Course name: Advanced Data Mining
Name and email address of the instructors: Arnaud Giacometti (Arnaud.Giacometti@univ-tours.fr)
Web page of the course:
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 12 h.
- Tutorials: 9 h.
- Laboratories: 12 h.
- Projects: 20 h.

Goals:

This course develops from the Data Mining (DM) course at ULB. Students will learn advanced machine learning and data mining techniques like semi-supervised approaches or deep learning, as well as state-of-the-art techniques related to the variety of data. A particular focus will be made on link analysis, social networks analysis, preference mining, usage analysis, topics and trends detection, opinion, and sentiment mining.

Learning outcomes:

Upon successful completion of this course, the student is able to:

- Prepare raw input data, and process it appropriately to provide suitable input for a wide range of data mining algorithms.
- Understand the theoretical background of data mining algorithms.
- Critically evaluate and select appropriate data mining algorithms.
- Apply data mining algorithms, interpret and report the output appropriately.
- More generally, students should be able to actively manage and participate in data mining projects executed by consultants or specialists in data mining

Readings and text books:

- J. Han, M. Kamber. *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2006.
- P.-N. Tan, M. Steinbach, V. Kumar. *Introduction to Data Mining*, Addison-Wesley, 2005.
- D. J. Hand, H. Mannila, P. Smyth. *Principles of Data Mining*, The MIT Press, 2001.

Prerequisites:

- Database Mining (DM)
- Big Data Management (BDM)
- Semantic Data Management (SDM)
- Cloud Computing (CC)

Table of contents:

- Mining spatio-temporal data (time series, data streams, trajectories, ...)
- Mining image databases (neural networks and deep learning)
- Mining the web (analysis of social networks, preference mining, relation and ontology extraction)
- Large-scale data mining (map reduce for data mining, neighbor search in high dimensional data, dimensionality reduction, ...)
- Mining text data (document classification and clustering, topics detection, opinion ming and sentiment analysis, ...)
- Interactive data mining
- Incremental learning
- Semi-supervised methods
- Multimedia and multimodal data processing and analysis

Assessment breakdown:

75% written examination, 25% project evaluation

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: CUAS
Course name: Content and Usage Analytics Seminar
Name and email address of the instructors: Nicolas Labroche (Nicolas.Labroche@univ-tours.fr)
Web page of the course:
Semester: 3
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 20 h.
- Projects: 30 h.

Goals:
In this seminar, students will get in touch with research in the area of content analytics. The research area will be studied in the following ways: (1) Students follow lectures by guest lecturers and experts, (2) Students will study research papers related to some proposed topics, (3) Students will be trained to writing scientific publications, preparing them for their forthcoming Master's thesis.

Learning outcomes:
With this module, students will:

- acquire a good understanding of the state of the art and the next evolutions and challenges in the domain,
- learn to synthesize, organize, and present scientific and technical information related to the domain of content and usage analysis, to make it clear to colleagues, and discuss it objectively

Readings and text books:

- Research articles and white papers in the domain of content and usage analytics.

Prerequisites:

- Database Systems Architecture (DBSA)
- Data Warehouses (DW)
- Data Mining (DM)
- Big Data Management (BDM)

Table of contents:

- Seminars linked to research fields covered in this specialization

Assessment breakdown:
75% written examination, 25% project evaluation

University: Université François Rabelais Tours
Department: Department of Computer Science
Course ID: EDT
Course name: Ethics and Digital Technologies
Name and email address of the instructors: Jean-Yves Antoine (Jean-Yves.Antoine@univ-tours.fr)
Web page of the course:
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 20 h.
- Exercises: 10 h.

Goals:

In this course, students will learn about the ethical implications of the development of digital technologies. They will be trained to detect and investigate the ethical issues raised in their professional activity, and providing methodological clues to answer these issues. The course covers the following aspects: Philosophical foundation of ethics and current ethical approaches, Ethics and technological risk, Ethics and regulation, Ethical responsibility of data scientists, and an overview of ethical issues in big data through case studies (e.g., privacy, neutrality, respect of intellectual property, individual and social impact of digital technologies).

Learning outcomes:

Upon successful completion of this course, the student is able to:

- Evaluate and discuss the impact of a scientific or an engineering work
- Identify the code of ethics (or the ethically related regulation) employed by professional bodies

Readings and text books:

- A. Briggle, C. Mitcham. *Ethics and Science: An Introduction.* Cambridge University Press, 2012.
- P. Oliver. *The student's guide to research ethics*, second edition, McGraw-Hill/Open University Press Maidenhead, 2010.
- B. Latour. *Pandora's Hope. Essays on the reality of science studies.* Harvard University Press, 1999.
- B. Latour, S. Woolgar. *Laboratory Life: the Social Construction of Scientific Facts.* Sage Publications, Beverly Hills, 1979.
- H. Jonas. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*, The University of Chicago Press, 1984.
- C. Nelson, S.R. Peterson. *If You're an Engineer, You're Probably a Utilitarian.* Proc. of the American Society of Civil Engineers: Issues in Engineering, 8(1):13-18, 1982.

Prerequisites:

- None

Table of contents:

- Philosophical foundation of ethics and current ethical approaches : ethics of virtues, deontologic and consequentialist ethics
- Ethics and technological risk : individual, societal and environmental potential impacts.
- Ethics and regulation : soft law vs. hard law, ethical grounding of law (universal principles of human rights)
- Shared responsability : ethics and compliance comitees, ethics guidelines, socially acceptable decision on ethical issues
- Ethics, computer science and data science : case studies
  - history of computer science : an epistemologic study of the evolution of computer uses
  - privacy : personal data protection and individual profiling
  - neutrality : access to information and decision models (fairness accountability and transparency with machine learning techniques)
  - intellectual property and Internet
  - indivudal impacts of digital technologies : cognitive effects, computer and social networks addictions
  - social impacts of digital technologies
- Technical advances and ethics : which relations between human beings, human societies and technique (transhumanism vs. technical scepticism)

Assessment breakdown:
100% Written essay on a specific ethical question concerning digital technologies