

Course Syllabus

[Jump to Today](#)

Computational Analysis of Big Data



Semester & Location:

Fall 2018- DIS Copenhagen

Type & Credits:

Elective Course - 3 credits

Major Disciplines:

Computer Science. Mathematics

Faculty Members:

Ulf Aslak, ulfaslak@gmail.com (<mailto:ulfaslak@gmail.com>)

Program Director:

Time & Place:

Location: Vestergade 10, A22

Thursdays 1:15 - 4:10pm

Course Description

Walmart started using big data even before the term became recognized. Today, industries, governments, social media platforms, finance, and organizations alike use data and analytics to optimize sales, minimize cost, and maximize reach. The ability to do so comes from the power of knowledge-based prediction, with the main goal of turning massive amount of data into actionable information.

In this course, we will learn about Big Data and Data Science from various perspectives and gain hands-on experience with a broad selection of tools and approaches in the context of relevant use-cases.

Classes will be a mix of thematic discussions, hands-on problem solving, and project work in groups. At the end of the course, you will be able to select and use appropriate combinations of tools and approaches to tackle typical problems due to Big Data.

Prerequisites

One year of introduction to Computer Science and an introduction to probability theory, linear algebra or statistics at university level. Practical programming experience is strongly recommended (e.g. in Python/Javascript/Java/C++/Matlab) and prior knowledge of algorithms and data structures is useful.

Learning Objectives

Upon successfully completing the course, the student will be able to:

- Understand how Big Data fits into the context of Data Science
- Select computational tools for performing analysis on Big Data
- Acquire large datasets from online sources and apply Data Science tools
- Extract knowledge and build prediction models using machine learning
- Critically evaluate how analytical tools influence results both from both a technical and ethical perspective

Course overview

The course is rooted in 12 sessions:

1. Coding with data in Python
2. A Data Scientist's most fundamental tools

3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Ethical and legal considerations in Big Data
10. Lab work on project report
11. Lab work on project report
12. Project presentations

Course Elements

The following topics are covered in this course:

- Python programming
- Web scraping
- Natural language processing
- MapReduce
- Machine learning
- Networks
- Legal considerations in Big Data
- Ethical considerations in Big Data

Teacher

Ulf Aslak, PhD researcher at the Copenhagen Centre for Social Data Science, University of Copenhagen, visiting researcher at the Technical University of Denmark (DTU), MScEng in Digital Media Engineering from DTU, former research assistant at the Uri Alon lab at the Weizmann Institute of Science, former Data Scientist at Trustpilot, and seasoned teaching assistant in academia. His current research focuses on pattern detection and behavior prediction in multimodal social network data.

Required texts

Most of the learning will be based on the book *Data Science from Scratch: First Principles with Python, 1st Edition* written by Joel Grus. We will also use the freely available book *Network Science* by Albert-László Barabási. Some learning will also be based on papers, blog posts and videos available online.

Approach to Teaching

The course is designed around the principle of [constructive alignment](http://www.johnbiggs.com.au/academic/constructive-alignment/) (<http://www.johnbiggs.com.au/academic/constructive-alignment/>). The two major components in the course—the

assignments and the final project—implement this principle by stating clear outcome goals of every activity and the course as a whole.

Assignments: Leading up to each session, students are given a "preparation goal" and a suggested list of materials they can use to reach it. Sessions start with a short lecture (less than 1 hour) that introduces the topic of the day, and then students work through a set of technical exercises. The students are required to hand in two assignments throughout the course (40% of their final grade, 20% each), which are composed of selected problems from the exercises they have solved in class. This gives the student a clear outcome goal for each session: "show up prepared and complete the exercises". It gives incentive to prepare and work focussed.

Final project: From the beginning of the course the students are aware that an outcome of the course is a project that, if done well, can add value to their professional portfolio. The project is a small study on some popular topic of their own choosing that they can investigate with data they have scraped or downloaded from the Internet. They submit the project in two parts: First, each team must compose a *proposal video* which demonstrates that they have made a plan for their project and are able to hypothesize about the outcomes. Second, after they have completed their project they must communicate the results in the popular format of a blog post. The proposal video is a fun exercise that serves as a platform for sharing ideas between groups (we view them all in class) but it also forces them to start with a very comprehensive idea of the outcome in mind.

Another small but important component of the teaching approach is peer evaluation. Each student is tasked with reviewing 2 assignments after handing in their own (with or without a group). The reviewing process is anonymous. Using peer evaluations, each hand in gets a lot of varied feedback, and lets students reflect on their own work by reviewing how others solved the same problems. High quality feedback is incentivized by having each reviewee rate their received feedback such as to produce a feedback quality score for every reviewer which, by a small fraction, influences their final grade.

Expectations of the Students

Students are expected to reach the preparation goal leading up to each session. Students who have little or no experience coding in Python should either follow a Python tutorial before the course starts, or prepare to invest some hours getting up to speed with the language once we start. Students should have a working laptop computer. It is advised that each machine has a least 4 GB of RAM and a reasonable processor (if it's bought after 2012 you should be fine). The Unix operating system is preferred (OSX and Linux), but not a necessity.

Field Studies

During the course there is allocated time for two one-day field studies. In the first field study, we will do a hackathon competition, where students compete to create the best performing prediction model on a selected dataset from [Kaggle](https://www.kaggle.com/) [\(https://www.kaggle.com/\)](https://www.kaggle.com/). In the second hackathon we will visit a local company that works with Big Data.

Assignments and Evaluation

During the course you will hand in two assignments containing selected exercises solved in class. Furthermore, you will complete a larger project that uses tools which have been taught in the class. An acceptable project will cover e.g. data scraping and analysis. You will be allowed to define your own project, but you can also get assistance from the teacher.

Both project and assignments are group efforts. The teacher will rate all the assignments, but you will also participate using the peer evaluation system Peergrade.io, where each handin is double-blind peer-reviewed by 3-4 students which, together with the teacher's evaluation composes indicators towards the final grade. This creates more and fairer feedback for each group as well as evaluation that is less sensitive to mistakes. Students' overall feedback quality is taken into account during grade evaluation.

During the programming projects, you are allowed to consult freely with any of the other students and the instructor. Contributions from other students, however, must be acknowledged with citations in your final report, as required by academic standards. Contributions to your presentations must similarly be acknowledged. Needless to say, the right to consult does not include the right to copy — programs, papers, and presentations must be your own original work.

When assigning the final grades, your efforts will weigh as follows:


- Participation: 15% (includes class/exercise/project behavior that is beneficial to the learning of others)
- Mandatory assignments: 40% (two hand-ins, each accounting for 20%)
- Final project: 35% (10% proposal video, 25% project report and presentation)
- Overall peer feedback quality: 10%






















Academic Regulations

Please make sure to read the [Academic Regulations \(https://disabroad.org/copenhagen/student-resource/academic-regulations/\)](https://disabroad.org/copenhagen/student-resource/academic-regulations/) on the DIS website. There you will find regulations on:

- [Course Enrollment and Grading \(https://disabroad.org/copenhagen/student-resource/academic-regulations/course-enrollment-grading/\)](https://disabroad.org/copenhagen/student-resource/academic-regulations/course-enrollment-grading/)
- [Attendance \(https://disabroad.org/copenhagen/student-resource/academic-regulations/attendance-policies/\)](https://disabroad.org/copenhagen/student-resource/academic-regulations/attendance-policies/)
- [Coursework, Exams, and Final Grade Reports \(https://disabroad.org/copenhagen/student-resource/academic-regulations/coursework-exams-final-grade-reports/\)](https://disabroad.org/copenhagen/student-resource/academic-regulations/coursework-exams-final-grade-reports/)

Course Summary:

Date	Details	
Thu Aug 23, 2018	 W1.1-2: Coding with data in Python	1:15pm to 4:10pm

Date	Details	
Thu Aug 30, 2018	 W2.1-2: A Data Scientist's most fundamental tools	1:15pm to 4:10pm
Thu Sep 6, 2018	 W3.1-2: Getting data—scraping and APIs	1:15pm to 4:10pm
Thu Sep 20, 2018	 W4.1-2: Machine learning 1	1:15pm to 4:10pm
Thu Sep 27, 2018	 W5.1-2: Machine Learning 2	1:15pm to 4:10pm
Wed Oct 3, 2018	 Field Study 1: Kaggle hackathon	8:30am to 12:30pm
	 Assignment 1	due by 11:59pm
Thu Oct 4, 2018	 W6.1-2: Networks	1:15pm to 4:10pm
Wed Oct 24, 2018	 Night seminar	6pm to 8pm
Thu Oct 25, 2018	 W7.1-2: Natural language processing	1:15pm to 4:10pm
Thu Nov 8, 2018	 W8.1-2: Crunching data with MapReduce	1:15pm to 4:10pm
Wed Nov 14, 2018	 Assignment 2	due by 11:59pm
Thu Nov 15, 2018	 W10.1-2: Screening of proposal videos (Project A due)	1:15pm to 4:10pm
Wed Nov 28, 2018	 Field Study 2	1pm to 5pm
Thu Nov 29, 2018	 W11.1-2: Project work	1:15pm to 4:10pm
Thu Dec 6, 2018	 W12.1-2: Final project presentations (Project B due)	1:15pm to 4:10pm
	 Participation grade	
	 Peer feedback	
	 Project: Blog post & repost	
	 Project: Presentation	
	 Project: Proposal video	
	 Roll Call Attendance	