Computational Analysis of Big Data

Fall 2017

Copenhagen

3 Credits

Major Disciplines: Computer Science. Mathematics.

Faculty Member: Henrik Pilegaard

Program Director: Iben de Neergaard, Vestergade 10 V23, idn@dis.dk

Program Coordinator: Louise Bjerre Bojsen, Vestergade 10 V23, lbb@dis.dk

Program Assistant: Jenny Han, Vestergade 10 V23, yh@dis.dk

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 investigate the topic of big data 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, short technical seminars, and hands-on problem solving projects where you 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 use-cases.

Prerequisites:

One year of introduction to Computer Science and an introduction to probability theory or statistics at university level. Experience with imperative or functional programming is essential and knowledge of algorithms and data structures is strongly recommended.

Learning Objectives:

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

- Recognize problems that benefit from a Big Data approach
- Select approaches to Big Data problems
- Select tools to facilitate Big Data approaches
- Compose tools into systems that automate Big Data solutions
- Critically evaluate your choices and solutions from both a technical and an ethical perspective

The course will have the following (partially overlapping) parts:

1. Introduction to Big Data - when and why is data 'big' and what can we do with it? 1 session.

2. Basic tools - how do we compute in the cloud and use python? 2 sessions.

3. Data acquisition - where does Big Data come from? 2 sessions.

4. Storing Big Data - how do we store Big Data for various purposes? 4-5 sessions

5. Streaming Big Data - what do we do when data cannot be stored? 1-2 sessions

6. 3.Analysing Big Data - how do we obtain value from Big Data? 5-6 sessions.

7. Lab work on programming project. 6 sessions.

Course Elements:

During the semester, we will touch upon the following technical and non-technical elements:

1. Cloud computing
2. Programming in Python
3. Web scraping
4. Web service APIs
5. SQL and noSQL databases
6. MapReduce processing
7. Streaming algorithms
8. Full text search and analytics engines
9. Graph databases
10.Bayesian analysis
11.Recommendation systems
12. Deep learning
13. Entity extraction
14. Legal considerations in Big Data 15.Ethical considerations in Big Data

Faculty:

Henrik Pilegaard, Ph.D., is the founder of **hifishark.com**, a search engine and social network platform for the High Fidelity interested. He is a former assistant professor at DTU Compute, Technical University of Denmark, where he worked on formal modelling languages and associated formal analysis techniques. He has also spent a significant amount of time at Kapow Software (now Kofax Kapow) working on the most advanced web automation platform currently available.

Required texts:

Most of the learning will be based on up to data online-resources, in the form of tutorials and product documentation. Since we will be using Python for the programming exercises, we will use one or both of the following to cover the basics:

Dmitry Zinoview: *Data Science Essentials in Python: Collect - Organize - Explore - Predict - Value,* The Pragmatic Programmer, 2016

Joel Grus: *Data Science from Scratch: First Principles with Python,* 1st Edition, O'Reily, 2015

Approach to Teaching:

In this course, we will learn mostly by doing. There will be a short lecture at the beginning of each session, but mainly to kick-start and inspire further dialogue and discussion about the topic at hand. Both teaching sessions and exercise classes will allow and encourage free and unrestricted collaboration amongst all students. My aim is for the course to be more of a Makerspace environment than a traditional classroom environment.

Expectations of the Students:

The proposed format for the course makes it important that all student engage actively in the sessions. In particular, I expect you to prepare for each class by studying the recommended material. When preparing, make notes of your observations and questions and bring them to class. This will give us material to generate conversation. Make sure to record the sources of your notes, so that you can reference them in class.

In my view a successful outcome of the course, i.e. that all participants get to understand and appreciate the concepts in Big Data, is a shared responsibility. Therefore, I will reward class behavior that is beneficial to the learning of others and supportive of the teaching format - participating in discussions, asking questions, answering questions, giving and seeking help to/from others during exercise classes.

Field Studies:

Big Data is a 'virtual' concept and, consequently, there is not much point in visits to physical locations. Instead, we set aside all available time for practical exploration of the virtual world. The field studies will therefore be conducted as 'hackathons', i.e. prolonged sessions of working with our programming project.

Assignments and Evaluation:

During the course, you will be expected to complete a programming project that integrates three or more of the outlined course parts. An acceptable project, e.g., where data is collected, analyzed, and the analysis results are used for real time exploration, would require elements of data acquisition, storing for analysis, and storing for real time use. Another acceptable project, where fast data are analyzed and the results are used for real time exploration, would require elements from sata acquisition, dealing with fast data, and storing for real time use. You will be allowed to define your own project, but you can also get assistance from the instructor.

The programming projects will be a group effort. At the end of the project, you will be asked to hand in a concise technical paper on your project. This paper will also be a group effort, and must include a critical evaluation of your own work. Furthermore, each of you will individually to give a presentation followed by a short QA session on your project in class. Finally, you will individually to write a peer review of two projects that you have not participated in, where you are expected to critically evaluate technical, legal, and ethical aspects of the projects.

During the programming projects, you are allow to consult freely with any of the other students and the instructor. Contributions from other students, however, must be noted with citations in your final paper, as required by academic standards. Contributions to your presentations must be acknowledged as well. 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 you efforts will weigh as follows:

Participation: 20% (includes class/exercise/project behavior beneficial to the learning of others)
Programming project/final paper: 40%
Programming project/final presentation: 20%
Written peer review of two other projects: 20%

Course Structure:

W1-1: Introduction to Big Data.
W1-2: Cloud computing platforms
W2-1 Introduction to Python
W2-2: Performing web scraping using python
W3-1: Using web service APIs from Python W3-2:
Lab work on programming project W4–1: SQL
databases
W4-2: NoSQL databases

W5-1: Lab work on programming project
W5-2: MapReduce
W6-1: Streaming Algorithms

W6-2: Lab work on programming project
W7-1: Lab work on programming project
W7-2: Full text analytics engines and Bayesian analysis W8-
1: Graph databases and recommendation services

W8-2: Deep learning and entity recognition
W9-1: Lab work on programming project W9-2:
Lab work on programming project
W10-1: Ethical and Legal considerations in Big Data

W10-2: Case study - competitive pricing - Final paper due
W11-1: Student presentations
W11-2: Student presentations
W12-1: Reserved for buffer/Student’s choice of subject
W12-2: Reserved for buffer/Student’s choice of subject - Peer review due

This syllabus is subject to change.