Data Science

Introduction to Data Science

KEA – Data Science Elective PBA – Fall 2019 – Henrik Strøm

September 5, 2019

COPENHAGEN SCHOOL OF DESIGN AND TECHNOLOGY

Table of Contents

About You and Me

Henrik Strøm

- Lecturer at KEA
- Ph.D. Fellow at Aalborg University
- M.Sc. in Engineering
- Master in Information and Communication Technologies
- Specialize in machine learning and data science
- Worked in Denmark, China, Thailand; projects in USA, Singapore
- Worked in major enterprise industry, financial sector, start-up
- Been at KEA since May 1st, 2018

What about you guys – who are you, and what are your expectations?

Quick Poll

How many of you ...

- Know Python?
- Have taken class in artificial intelligence or machine learning?
- How many of you know about dimensionality reduction, cluster analysis, learning algorithms, how to handle bias and variance problems, and statistics?
- Have taken a high level math class?

About the Course

Objectives

The objective of the module is to qualify the student to apply data science methods in a structured manner, extract inferential knowledge from a dataset, and make probabilistic forecasts. In addition, the student must be able to report on the findings and make use of visualization.

About the Course

Knowledge

The objective is to give the student knowledge of:

- various definitions of the field of data science
- designing experiments for data collection
- using a framework for numerical computations
- using a framework for general data analysis
- using a framework for inferential statistical analysis
- using a framework for probabilistic forecasting
- a variety of data analysis algorithms and their applications

About the Course

Skills

The objective is that the student will have acquired the ability to:

- collect data from a variety of sources
- organize data to prepare for analysis
- explore data to gain insights
- apply basic inferential statistics
- make forecasts using probabilistic machine learning tools
- use methods to facilitate reproducibility
- communicate findings in a written report, using visualizations

About the Course

Proficiency
The objective is that the student will have acquired proficiency in participating and contributing in data science projects.

Concepts and understanding is more important than syntax.

Course Workload

- This course is 10 ECTS points = 280 hours workload.
- It is important that you attend class – reading the material is not enough.
- Course routine and repetition.

Mandatory Assignment and Exam

- Two written mandatory assignments: peer-review and hand-in.
- The peer-review is mandatory, you can't take the exam unless the mandatory assignments are accepted.
- Oral exam based on project.

Defining Data Science

Data science is gaining more and more and widespread attention, but no consensus viewpoint on what data science is has emerged. As a new science, its objects of study and scientific issues should not be covered by established sciences. In the present paper, data science is defined as the science of exploring datanature. We believe this is the most logical and accurate definition of data science (...) – (Zhu and Xiong, 2015)

Data Science Research Topics

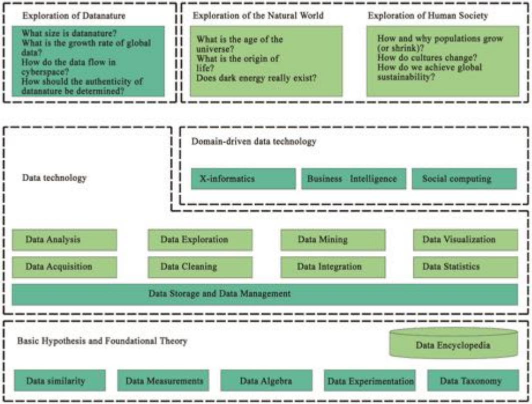
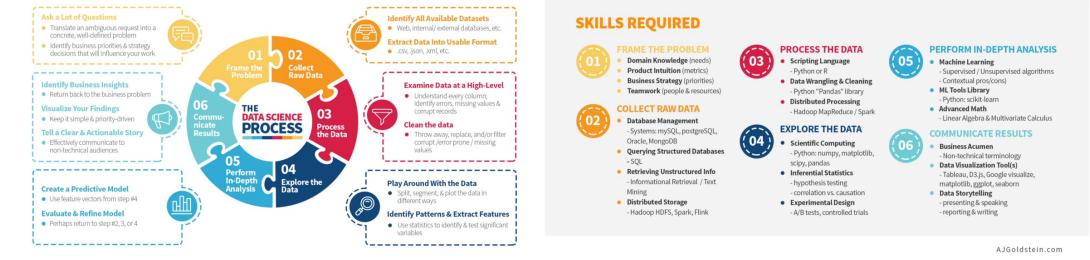

Exploration of Datanature
What size is datanature?
What is the growth rate of global data?
How do the data flow in cyberspace?
How should the authenticity of datanature be determined?
Exploration of the Natural World
What is the age of the universe?
What is the origin of life?
Does dark energy really exist?
Exploration of Human Society
How and why populations grow (or shrink)?
How do cultures change?
How do we achieve global sustainability?
Data technology
Domain-driven data technology
X-informatics
Business Intelligence
Social computing
Data Analysis
Data Exploration
Data Mining
Data Visualization
Data Acquisition
Data Cleaning
Data Integration
Data Statistics
Data Storage and Data Management
Basic Hypothesis and Foundational Theory
Data Encyclopedia
Data similarity
Data Measurements
Data Algebra
Data Experimentation
Data Taxonomy

Figure: Data Science research topics. (Zhu and Xiong, 2015)

Elements of Data Science



Ask a Lot of Questions
- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

Identify Business Insights
- Return back to the business problem

Visualize Your Findings
- Keep it simple & priority-driven

Tell a Clear & Actionable Story
- Effectively communicate to non-technical audiences

Create a Predictive Model
- Use feature vectors from step #4

Evaluate & Refine Model
- Perhaps return to step #2, 3, or 4

THE DATA SCIENCE PROCESS

01 Frame the Problem

02 Collect Raw Data

03 Process the Data

04 Explore the Data

05 Perform In-Depth Analysis

06 Commu-nicate Results

Identify All Available Datasets
- Web, internal/ external databases, etc.

Extract Data Into Usable Format
- .csv, .json, .xml, etc.

Examine Data at a High-Level
- Understand every column; identify errors, missing values & corrupt records

Clean the data
- Throw away, replace, and/or filter corrupt /error prone / missing values

Play Around With the Data
- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features
- Use statistics to identify & test significant variables

SKILLS REQUIRED

01 FRAME THE PROBLEM
- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 COLLECT RAW DATA
- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 PROCESS THE DATA
- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 EXPLORE THE DATA
- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

05 PERFORM IN-DEPTH ANALYSIS
- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

06 COMMUNICATE RESULTS
- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

AJGoldstein.com

Figure: Elements of Data Science (Goldstein, 2017)

What's Most Important

- Transparency
- Reproducibility
- Correctness
- Predictive Power
- *Insights!*

Data Science is an application of scientific methods and principles to data processing.

Modern Science

What is the difference between
the ”science” of Artistole
and the science of Galileo Galilei?

Computer Science – Levels of Knowledge

- Axiomatic knowledge
- Scientific Consensus
- Working application of best scientific standards (Development/Engineering)
- Software Construction
- Fooling Around

(Nunamaker Jr, Chen, and Purdin, 1990)

Ethics

No, this is not going to be a social science class, however …

– ethics is important – both in how you conduct your research, and in how you report and communicate it. Data Science is not tabloid journalism or politics.

With great power comes great responsibility: be honest and follow the evidence!

Software Needed for the Course

- Git – use terminal or GUI
- Anaconda (Python 3 edition)
- Course materials: `https://bitbucket.org/henrikstroem/kea-public-datascience-2019-2`

Python Primer

DEMO

References I

- Goldstein, A. J. (2017). *Deconstructing Data Science: Breaking The Complex Craft Into It's Simplest Parts*. URL: https://medium.com/the-mission/deconstructing-data-science-breaking-the-complex-craft-into-its-simplest-parts-15b15420df21.
- Nunamaker Jr, Jay F, Minder Chen, and Titus D M Purdin (1990). "Systems development in information systems research". In: *Journal of management information systems* 7.3, pp. 89–106. URL: http://gkmc.utah.edu/7910F/papers/JMISsystemsdevelopmentinISresearch.pdf.
- Zhu, Yangyong and Yun Xiong (2015). "Defining data science". In: *arXiv preprint arXiv:1501.05039*. URL: https://arxiv.org/abs/1501.05039.