Introduction

Evangelos Pournaras, Izabela Moise, Dirk Helbing
Outline

1. Data Science
2. Course Description
Part 1 - Data Science
What is Data Science?

A collection of orchestrated methods from different scientific fields, e.g. statistics, computer science, etc., that provide understanding of domain data and result in data-based products and services.
Is Data Science about Big Data? I

“Remember, the other team is counting on Big Data insights based on previous games. So, kick the ball with your other foot.”
Is Data Science about Big Data? II

Big vs. Small Data

It’s more about using the **right data** and asking the **right questions**!
What about Techno-socio-economic Systems?
ICT & Techno-socio-economic Systems?

- Embedded ICT systems in most societal domains. How?
- Internet of Things, pervasive/ubiquitous computing, advanced networking systems, inter-operability. Result?
- A new explosion of data sources! Opportunities?
- Understanding, improving, managing & sustaining our complex society! Threats?
- Privacy, discrimination, misinterpretations, over-fitting, etc.
"The new hidden cameras will allow us to see if anyone is violating our privacy policy by reading someone else's email."
Threats II

I think we should give it another shot.

We should break up, and I can prove it.

Our relationship

Huh.

Maybe you're right.

I knew data would convince you.

No, I just think I can do better than someone who doesn't label her axes.
Who is a Data Scientist?

- A statistician?
- A computer programmer?
- Both and More!

**Tip**

*Domain knowledge can be more valuable than machine learning, data mining, etc.*
Real-world Profile I

MODERN DATA SCIENTIST

Data Scientist, the most recent job of the 21st century, requires a mixture of multidisciplinary skills ranging from an understanding of mathematics, statistics, computer science, communication, and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Programming language e.g. Python
- Statistical computing packages e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with tools like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visuals and design
- R packages like ggplot or lattice
- Knowledge of any visualization tools e.g. R, D3.js, Tableau

DOMINO Analytics is a group of geeks who can help you with your data science needs. Our team of experts includes marketing, strategy and communication. We help companies build and maintain data science and analytics programs. We have experience in data mining, big data analytics, natural language processing, data warehousing, and big data systems. Visit our website at www.dominorocks.com

DGESS
Evangelos Pournaras, Izabela Moise, Dirk Helbing
Real-world Profile II
More about Data Scientists

More about Data Scientists

"Data scientists’ most basic, universal skill is the ability to write code. This may be less true in five years’ time, when many more people will have the title "data scientist" on their business cards. More enduring will be the need for data scientists to communicate in language that all their stakeholders understand-and to demonstrate the special skills involved in storytelling with data, whether verbally, visually, or - ideally both."
More about Data Scientists

"But we would say the dominant trait among data scientists is an intense curiosity—a desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested."
More about Data Scientists

"A quantitative analyst can be great at analyzing data but not at subduing a mass of unstructured data and getting it into a form in which it can be analyzed."

"A data management expert" might be great at generating and organizing data in structured form but not at turning unstructured data into structured data-and also not at actually analyzing the data."

"And while people without strong social skills might thrive in traditional data professions, data scientists must have such skills to be effective."
Part 2 - Course Description
Course Objectives

Qualify you with knowledge & skills to tackle real-world problems using data.

1. Acquiring domain knowledge and understanding.
2. Better understanding and interpretation of data.
3. Awareness about the applicability of different data science methods.
4. Development of technical skills, e.g. programming, use of different tools, etc.
5. Presenting scientific results both written and orally.
Course Prerequisites

Some programming skills are required, e.g. skills for the material of this course:

1. Java/C++/Python
2. UNIX

Didn’t you have an opportunity to practice this earlier?

No problem, this is a golden opportunity!

Tip

*Programming skills will make you more flexible and efficient data scientist!*
Assessment

- Seminar thesis
- 100% of the grade, no exams!
- Detailed illustration in a next lecture.

Tip

Start early! Give the opportunity for your project and your skills to develop during the course.
Lectures

- Every Monday 17:15-19:00 at LFW B 1
- Participation is not obligatory but **highly recommended**!
- 60 minutes lectures followed by **40 minutes interactive discussions**.
- Opportunity to discuss your project.
- Lectures at:
  
  http://www.coss.ethz.ch/education/datascience.html
Subjects I

1. Computational Social Science Applications - 3 weeks
   - Smart Grids, geolocation, traffic systems, social sensing/mining
   - Tools & platforms: Nervousnet, Twitter, GDELT

2. Data Science Fundamentals - 2 weeks
   - databases, data types, data collection, data pre-processing, plotting, visualization, etc.
   - Tools: Java, AWK, MySQL, Gnuplot, Gephi, etc.

3. Data Mining and Machine Learning - 2 weeks
   - classification, clustering, prediction, neural networks, etc.
   - Tools: Weka

4. Big Data Analytics - 2 weeks
Subjects II

- MapReduce, parallel computing, data streaming, social media, etc.
- Tools: Hadoop, Spark, Mahout, Spark Streaming, Storm, etc.

5. The Nervousnet Hackathon

- Social Sensing & Analytics

6. Other - 2 weeks

- Project presentations
Lectures Outline

Lecture 01 (22.02.16)
Introduction

Lecture 02 (29.02.16)
Applications

Lecture 03 (07.03.16)
Applications

Lecture 04 (14.03.16)
Applications

Lecture 05 (21.03.16)
Data Science Fundamentals

Lecture 06 (04.04.16)
Data Science Fundamentals

Lecture 07 (11.04.16)
Data Mining and Machine Learning

Lecture 08 (25.04.16)
Data Mining and Machine Learning

Lecture 10 (02.05.16)
Big Data Analytics

Lecture 11 (09.05.16)
Big Data Analytics

Lecture 12 (23.05.16)
Oral Presentations

Lecture 12 (30.05.16)
Oral Presentations

Special Lecture (22&23.04.16)
The Nervousnet Hackathon
How to contact us

Communication:

- Discussion session in the course
- E-mail with subject [DATA-SCIENCE-COURSE-2016]<otherinfo> to:
  - Evangelos Pournaras: epournaras@ethz.ch, and/or
  - Iza Moise: imoise@ethz.ch

Supervision - strictly for issues not addressed in the course:

- Mondays 15:00-17:00
  Clausiusstrasse 50 (CLU C 4), 8092, Zurich
Proposed Literature

B. Ellis.

*Real-Time Analytics: Techniques to Analyze and Visualize Streaming Data.*

J. Han.

*Data Mining: Concepts and Techniques.*

T. White.

*Hadoop: The Definitive Guide.*

I. H. Witten, E. Frank, and M. A. Hall.

What is next?

- Seminar thesis
- Examples and applications