1. COURSE LOGISTICS
2. WHAT IS DATA SCIENCE?
3. DATA
WHO WE ARE

• Lecturer: Teemu Roos, Associate professor, PhD

• TAs: Chang Rajani, Ioanna Bouri & Ville-Veikko Saari

• How to reach us:
  1. Piazza: piazza.com/helsinki.fi/fall2017/data11001
  2. Email teemu.roos@cs.helsinki.fi, chra@cs.helsinki.fi
  3. Bump into us
  4. Knock on the door
WITH SPECIAL THANKS TO

WRAY BUNTINE
DIRECTOR OF DATA
SCIENCE MSC PROGRAMME
MONASH UNIVERSITY
AUSTRALIA

(FOR LETTING ME TAKE A
LOOK AT HIS
"INTRODUCTION TO DATA
SCIENCE" MATERIALS)
LOGISTICS

• Lectures Mondays 10am-12pm & Tuesdays 4pm-6pm, B123

• Exercise groups – **starting next week**
  1. Tuesdays 12pm-2pm B120 (57 registered)
  2. Thursdays 4pm-6pm B222 (56 registered)
  3. Wednesdays 12pm-2pm C222 (47 registered)
  4. <new group tba> (25 registered)

• 185 registered! Oops.

• If you have problems registering, please contact Reijo Siven <reijo.siven@helsinki.fi>
• Date & time: (this data can be extracted from the department website)

• In the exam, you can have a “cheat sheet”: a single double-sided A4, handwritten (not copied) notes

• Point: you’ll have to summarize the course contents to yourself – often making the notes is more useful than having them in the exam
WHAT YOU NEED TO DO

• Lectures are **not** compulsory – but meant to be useful

• **YOU DON’T LEARN TO DO JUST BY LISTENING**

• Grade = exercises + exam + miniproject

• Alternative way: project + separate exam

• Note about the alternative way:
  – project has to be submitted **a week prior** to the separate exam
  – register to the separate exam online
WHAT YOU SHOULD KNOW (ALREADY)

• Pretty good programming skills
  – no time to learn how to program on this course, sorry
  – language is your choice but we recommend python
  – you can probably pick it (python) up as we go

• Using command-line tools in a Linux environment

• Some statistics:
  – linear regression, interpretation of a hypothesis test, …

• If you’re missing some of these, it’s your responsibility to make sure you fix it: we’ll provide some pointers to help.
OVERVIEW

**THEME 1**
data science, storage, data formats, “wrangling”

**THEME 2**
exploration, visualization

**THEME 3**
statistical methods, machine learning

**THEME 4**
big data frameworks, deep learning

**THEME 5**
data governance, privacy, ethics

**THEME 6**
operationalization
a.k.a. “How to create added value as a data scientist”
WHAT IS DATA SCIENCE?
DEFINITIONS

• “It’s what a data-scientist does.” – circular

• “Machine learning/data mining/statistics.” – too narrow

• “Collecting, manipulating, and analysing data in order to extracting value from it.”

• Wikipedia: “Data Science is the extraction of knowledge from data, which is a continuation of the field of data mining and predictive analytics.”

• NIST Big Data Working Group: “Data Science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.”
THE HYPE

• There’s huge and growing demand especially in business

• Predicting the future is hard, but most likely you’ve made a great choice

• Because of the hype, everyone wants to “own” data science! – many of them are just selling their stuff with a new label

• Here, we mostly ignore the hype and talk about Data Science
WHAT IS A DATA SCIENTIST?

• A Data Scientist can:
  – *understand* the background domain
  – *design* solutions that produce added value to the organization
  – *implement* the solutions efficiently
  – *communicate* the findings clearly (important!)

• Data Scientist is a *practitioner* with sufficient expertise in software engineering, statistics/machine learning, and the application domain.

• [Hans Rosling video](#)
NEXT UP:
DATA
BIG DATA

• TED Talk: “Big data is better data” by Kenneth Cukier

• A crucial part of the rise of Data Science is the steep increase in the amount and availability of data

• Big Data refers not only to the quantity but also to the quality of the data:
  – VOLUME: lots of it
  – VELOCITY: fast (streaming)
  – VARIETY: all kinds, not nice and “clean”
  – VERACITY: can it be trusted?
**KINDS OF DATA**

- **STRUCTURED DATA**
  - lists
  - $n \times p$ tables, arrays
  - hierarchies
    (e.g., organization chart)
  - networks
    (e.g., travel routes, hypertext = links)

- Generic data-interchange formats:
  XML, JSON

- **UNSTRUCTURED DATA**
  - text
  - images
  - video
  - sound

- Often can be made structured by, e.g., parsing language, segmenting images, etc.
**STRUCTURED DATA FORMATS**

- CSV, comma separated values
  
  ```
  sepal_length,sepal_width,petal_length,petal_width,species
  5.1,3.5,1.4,0.2,setosa
  4.9,3,1.4,0.2,setosa
  4.7,3.2,1.3,0.2,setosa
  4.6,3.1,1.5,0.2,setosa
  ```

- Hierarchies, e.g., Newick tree format
  
  ```
  (A,B,(C,D)E)F;
  ```

- Networks, e.g., GraphViz (DOT)
  
  ```
  digraph graphname {
    a -> b -> c;
    b -> d;
  }
  ```
JSON

- Similar to XML but simpler
import json
string = '{"first_name": "Alice", "last_name": "Wu"}'
parsed_object = json.loads(string)

print(parsed_object['first_name'])
Alice

d = {
    'name': 'Alice Wu',
    'titles': ['Dr', 'Prof'],
}

print(json.dumps(d))
{"titles": ["Dr", "Prof"], "name": "Alice Wu"}
XML

- same example:

```xml
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <streetAddress>21 2nd Street</streetAddress>
    <city>New York</city>
    <state>NY</state>
    <postalCode>10021</postalCode>
  </address>
  <phoneNumber>
    <type>home</type>
    <number>212 555-1234</number>
  </phoneNumber>
  <phoneNumber>
    <type>fax</type>
    <number>646 555-4567</number>
  </phoneNumber>
  <gender>
    <type>male</type>
  </gender>
</person>
```
PARSING

• Given a known grammar, unstructured text data can be parsed

• “It ain’t over till the fat lady sings”

  ((it, (ain't, over)), (till, ((the, (fat, lady)), sings)))

• Similarly, images can be segmented into parts