Outline of Lecture 01

✓ Introduction
✓ What is data science?
✓ 10 success stories of data science
  • Data science in Estonia
  • Terminology: data mining, data science, ...
  • What can you learn in this course?
  • Organisational information about this course
Data science in Estonia
Estonian data science community


- [https://www.facebook.com/groups/datasci.ee/](https://www.facebook.com/groups/datasci.ee/)  
  September 2017 – 400+ members  
  September 2020 – 1500+ members
Data science in Estonia: Research
Data science research in Estonia

- **University of Tartu**
  - **Institute of Computer Science**
    - Professors working in fields related to data science
      - Raul Vicente, Prof of Data Science
      - Satish Srirama, Research Prof of Big Data
      - Mark Fišel, Prof of Natural Language Processing
      - Jaak Vilo, Prof of Bioinformatics
      - Marlon Dumas, Prof of Information Systems
    - Associate professors:
      - Meelis Kull, Huber Flores, Dirk Oliver Theis, Kuldar Taveter, Alex Nolte...
    - Senior research fellows:
      - Raivo Kolde, Sven Laur, Leopold Parts, Hedi Peterson, Jaan Aru, Radwa El Shavi, Ahmed Awad, ...
    - Research fellows, lecturers, PhD students, Master students, ...
      - Institute of Mathematics and Statistics
      - ...

- **Tallinn University of Technology**
- ...

Meelis Kull - Autumn 2020 - LTAT.02.002 – Intro to Data Science - Lecture 01
Data science in Estonia: 4 examples from industry
Leap into the future

AI and machine learning solutions built exclusively for your business...
Insights from Mobile Big Data

A technological platform and methodology for processing Mobile Big Data for human mobility monitoring, analyses, and statistical indicators.
Introduction

What is data science?

10 success stories of data science

Data science in Estonia

- **Terminology:** data mining, data science, ...

- What can you learn in this course?

- Organisational information about this course
Terminology: data mining, machine learning, artificial intelligence, data science?
Terminology

• Terms:
  – Data mining (DM)
  – Machine learning (ML)
  – Artificial intelligence (AI)
  – Data science (DS) – the newest term

• Meaning?
  – Highly overlapping terms
  – No agreed boundaries
  – Used with narrow or broad meanings
  – Used with technical or popular meanings
  – Slightly changing in time
Meanings in this course

• Data mining:
  – Discovering patterns from data

• Machine learning:
  – Algorithms that learn from data

• Artificial intelligence:
  – Algorithms of intelligent machines

• Data science:
  – Science about how to use data
Theoretical and Applied DS

• Theoretical data science
  – Science about how to use data (science of data)
    • Academic subject (like computer science or statistics)

• Applied data science is:
  – Applying scientific methods to data
    (solving domain-specific tasks in a data-driven way)
    • Academic or industrial subject

• Data science:
  – Can mean either theoretical data science or applied data science and one should guess from the context
Other related terms

• Pattern recognition
• Big data
• Business intelligence
• Statistics
• Knowledge discovery from databases
Different views on terminology

Sources:
http://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html
Outline of Lecture 01

✓ Introduction
✓ What is data science?
✓ 10 success stories of data science
✓ Data science in Estonia
✓ Terminology: data mining, data science, …

• What can you learn in this course?
• Organisational information about this course
What can you learn in this course?
Cannot teach all tools of data science!
A PERIODIC TABLE OF VISUALIZATION METHODS

Note: Depending on your location and connection speed it can take some time to load a pop-up picture.

version 1.5

© Ralph Langner & Martin Wegner - www.vismethods.org
What can you learn in this course?

- Types of approaches in data science
- Key steps in a data science project
- Understand some important algorithms
- Learn to use Python for data science
- Get some practice in working with data
- Think like a data scientist
- Gain confidence to using data in the future in your work, whatever your work will be
Topics in LTAT.02.002 Intro to Data Science

• Introduction
• First look at the data
• Tableau and visualisation
• Frequent pattern mining
• Relations of attributes, clustering and dimensionality reduction
• Machine learning 1: Introduction
• Machine learning 2: Classification
• Machine learning 3: Regression
• Machine learning 4: Deep learning
• Computational statistics
• CRISP-DM, Graph mining and NLP
• Databases and privacy
Outline of Lecture 01

✓ Introduction
✓ What is data science?
✓ 10 success stories of data science
✓ Data science in Estonia
✓ Terminology: data mining, data science, …
✓ What can you learn in this course?

• Organisational information about this course
Organisational information about this course
Online information about the course

- **Course homepage:**
  - This is the primary source of all information

- **Course forum:**
  - [https://piazza.com/ut.ee/fall2020/ltat02002](https://piazza.com/ut.ee/fall2020/ltat02002)
  - Electronic communication is only through Piazza (no communication via e-mails)
  - You will get invitations to register by today evening

- **Homework submission and feedback:**
  - Through course homepage, more information given in the practice sessions
This is how the course works

- **Thursday (day 0):**
  - New material presented in the lecture

- **Monday (day 4):**
  - Homework about this material becomes available at the course homepage

- **At all times:**
  - Use course forum to ask and provide help in understanding the lecture and homework tasks

- **Monday/Tuesday/Wednesday (day 4 / day 5 / day 6):**
  - Practice sessions discussing the homework to make sure everyone knows what needs to be done

- **Monday at noon (12:00 of day 11):**
  - Deadline for submitting homework into course homepage

- **Monday/Tuesday/Wednesday (day 11 / day 12 / day 13):**
  - Homework solutions discussed in the practice session

- **The following 2 weeks:**
  - Homework will be graded
Course grading

- The grade is calculated from points (max **100 points**)
- The points can be earned as follows:
  - Homeworks (**40 points**): 10 homeworks, each 4 points
  - Group project and poster presentation (**20 points**)
  - Written exam (**40 points**)
    - In case of pandemic restrictions we replace by online exam or multiply the points from the homeworks and project by 1.67
- Additional points can be earned from bonus tasks in homeworks
- Attending at least 9 out of 12 practice sessions is compulsory:
  - after missing 3 practice sessions each additional missed practice session results in losing 5 points
  - both physical and online presence are acceptable
- In order to pass the course the student must get
  - at least 50% from homeworks (threshold 20 points) and
  - at least 50% from project (threshold 10 points)
  - at least 50% from the exam (threshold 20 points)
Homeworks

• Homeworks are individual
  – You are not allowed to share your homework solutions with anyone else

• You need to submit a Jupyter notebook to the course homepage
  – More detailed instructions will be provided later

• Please be ready to explain your solutions during the practice session to everyone
  – Explaining is an important part of data science!

• First homework deadline:
  – Monday, September 21, at noon (12:00)
Group project

• Teams of 2-3 students
• Every team must present the project as a poster in the poster session in December
• Projects start officially towards the end of the course, with the last homework
• You can start forming teams and thinking about potential topics already now!
• More information coming during the course
Exam

• Exam is in written form

• Two times to choose from and a resit time
  – One in December (for a limited number)
  – Another in January
  – Resit in January
  – …exact dates given soon

• Exam will cover lecture and homework material
  – You are not required to write programs in the exam
Amount of work

- 6 ECTS = 6*26 = 156 hours of intensive work
  - This is an expected average over all students, assuming basic computer science background (programming, mathematics, probabilities & stats)
  - Less background and skills means more hours
- 2 hours per each of 12 lectures (24h)
- 6 hours per each of 10 homeworks (60h)
- 2 hours per each of 12 practice sessions (24h)
- 30 hours per person on project (30h)
- 18 hours on preparing for the exam (18h)
- Total: 156 hours
Working language

• **English by default**
  – There are many who cannot understand Estonian
  – Everyone benefits from practicing presenting in English
  – Not enough resources to run English and Estonian courses separately

• For me and for the University of Tartu it is important that the Estonian students would know both Estonian and English scientific vocabulary

• Therefore, we are running some practice sessions in Estonian
Is it a good idea to use polls in data science lectures?

https://participant.turningtechnologies.eu/polling/ids2020

A. Absolutely!
B. Probably good
C. Not sure yet
D. Probably bad
E. Bad idea
F. Prefer using some other app for feedback
For first practice session

• Please bring your own laptop
• Make sure you have Python3 and Jupyter Notebook installed in your laptop
  – See instructions on the course homepage
• For future practice sessions (can do later):
  – Install Pandas, Scikit-learn
Outline of Lecture 01

✓ Introduction
✓ What is data science?
✓ 10 success stories of data science
✓ Data science in Estonia
✓ Terminology: data mining, data science, …
✓ What can you learn in this course?
✓ Organisational information about this course
Quotes about Data

• “Drowning in Data yet Starving for Knowledge”
  – Source unknown

• “Computers have promised us a fountain of wisdom but delivered a flood of data”
  – William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus

• “Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?”
  – T. S. Eliot

• From: http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-1.html