Lecture 01: Introduction to Data Science

Meelis Kull

meelis.kull@ut.ee

Autumn 2018
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


Data Science Estonia | Eesti andmeteaduse kommuun

- [https://www.facebook.com/groups/datasci.ee/](https://www.facebook.com/groups/datasci.ee/)
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
      – Sherif Sakr, Research Prof of Big Data
      – Jaak Vilo, Prof of Bioinformatics
      – Marlon Dumas, Prof of Information Systems
      – Eero Vainikko, Prof of Distributed Systems
    • Associate professors:
      – Mark Fišel, Meelis Kull, Fabrizio Maggi, Satish Srirama, …
    • Senior research fellows:
      – Sven Laur, Leopold Parts, Hedi Peterson, Ahmed Awad, Radwa El Shavi…
    • Research fellows, lecturers, PhD students, Master students, …
  – Institute of Mathematics and Statistics
  – …
• Tallinn University of Technology
• …
Data science in Estonia: 4 examples from industry
STACC is the leading **machine learning and data science** competence center in Estonia that develops artificial intelligence solutions.
SIFR is a company focused on empowering your data using artificial intelligence and big data. We pride ourselves by finding real world business value from AI solutions and we help companies set up data science projects, build AI teams and develop an AI strategy.
Insights from Mobile Big Data

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

BIG DATA LANDSCAPE 2017

INFRASTRUCTURE
- HADOOP ON-PREMISE
  - Cloudera
  - MapR
  - Hadoop
- HADOOP IN THE CLOUD
  - Amazon
  - Microsoft Azure
  - Google Cloud Platform
- STREAMING / IN-MEMORY
  - Spark
  - Hbase
- DATABASES
  - NoSQL
  - SQL
- DATA TRANSFORMATION
  - Talend
  - Pentaho
  - IBM
- DATA INTEGRATION
  - Talend
  - Pentaho
  - IBM
- DATA GOVERNANCE
  - Cloudera
  - IBM
- BI PLATFORMS
  - Tableau
  - Qlik
- VISUALIZATION / ANALYTICS
  - Tableau
  - Qlik

ANALYTICS
- DATA ANALYST PLATFORMS
  - Pentaho
  - Tableau
- ALGORITHMS
  - Spark
  - Hbase
- STATISTICAL COMPUTING
  - SAS
  - R
- BUSINESS INTELLIGENCE
  - Tableau
  - Qlik

APPLICATIONS – ENTERPRISE
- SALES
  - Oracle
  - SAP
- MARKETING – B2B
  - Marketo
  - Salesforce
- CUSTOMER SERVICE
  - Salesforce

APPLICATIONS – INDUSTRY
- ADVERTISING
  - Adobe
  - DoubleClick
- DATA SOURCES & APIs
  - Foursquare
  - Twitter
- MACHINES LEARNING
  - TensorFlow
  - Scikit-learn
- horizontally
  - Google
  - IBM

CROSS-INFRASTRUCTURE/ANALYTICS
- OPEN SOURCE
  - Python
  - R
- FRAMEWORK
  - Hadoop
  - Spark
- QUERY / DATA FLOW
  - SQL
  - NoSQL
- DATA ACCESS
  - MongoDB
  - CouchDB
- COORDINATION
  - Apache

DATA SOURCES & APIs
- HEALTH
  - Apple
  - Garmin
- IOT
  - GE
  - IBM
- FINANCIAL & ECONOMIC DATA
  - Bloomberg
  - Thomson Reuters
- AIR / SPACE / SEA
  - NASA
  - MIT
- PEOPLE & ENTITIES
  - LinkedIn
  - LABC
- LOCATION INTELLIGENCE
  - Google Maps
  - Yahoo!
- OTHER
  - Amazon
  - Google

DATA RESOURCES
- RESEARCH
  - MIT
  - Stanford
- INCUBATORS & ACCELERATORS
  - Y Combinator
  - 500 Startups
- UNIVERSITY OF TAMU
  - Institute of Computer Science

V2 – Last updated 5/3/2017
© Matt Turck (@matturck), Jim Hao (@jimhcao), & FirstMark (@firstmarkcap) matturck.com/bigdata2017

Meelis Kull - Autumn 2018 - LTAT.02.002 – Intro to Data Science - Lecture 01
Cannot teach all visualisation methods!
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

- Lecture 01 - Sept 5 - Introduction
- Lecture 02 - Sept 12 - First look at the data
- Lecture 03 - Sept 19 - Data exploration
- Lecture 04 - Sept 26 - Frequent pattern mining
- Lecture 05 - Oct 3 - Clustering and dimensionality reduction
- Lecture 06 - Oct 10 - Machine learning 1: Introduction
- Lecture 07 - Oct 17 - Machine learning 2: Classification
- Lecture 08 - Oct 24 - Machine learning 3: ROC analysis and regression
- Lecture 09 - Oct 31 - Machine learning 4: Deep learning
- Lecture 10 - Nov 7 - Basic statistics
- Lecture 11 - Nov 14 - Databases and big data
- Lecture 12 - Nov 21 - Applications
- Lecture 13 - Nov 28 - Software tools
- Lecture 14 - Dec 5 - The world of data science
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:
  – 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
Meeting times

• Lecture:
  – Wed 16-18 @ Liivi 2-111

• Practice sessions (start Sept 17 and 18):
  – Group 1, Tue 16-18 @ Ülikooli 17 – 220
  – Group 2, Tue 8-10 @ Liivi 2 – 405
  – Group 3, Wed 8-10 @ Liivi 2 – 405
  – Group 4, Wed 10-12 @ Liivi 2 – 206
This is how the course works

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

- **Thursday (day 1):**
  - 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

- **Tuesday/Wednesday (day 6 / day 7):**
  - Practice sessions discussing the homework to make sure everyone knows what needs to be done

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

- **Tuesday/Wednesday (day 13 / day 14):**
  - 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 (44 points): 11 homeworks, each 4 points
  - Group project and poster presentation (20 points)
  - Written exam (36 points)
- 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
- In order to pass the course the student must get
  - at least 50% from homeworks (threshold 22 points) and
  - at least 50% from the exam (threshold 18 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:
  – September 24, at noon (12:00)
Group project

• Teams of 2-3 students
• Every team must present the project as a poster in the poster session on December 19
• 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
  – Will be held 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 14 lectures (28h)
- 6 hours per each of 11 homeworks (66h)
- 2 hours per each of 12 practice sessions (24h)
- 24 hours per person on project (24h)
- 14 hours on preparing for the exam (14h)
- Total: 156 hours
Is it a good idea to use clickers in data science lectures?

A. Absolutely!
B. Probably good
C. Not sure yet
D. Probably bad
E. Bad idea

Bar chart showing:
- Absolutely: 40
- Probably good: 9
- Not sure yet: 9
- Probably bad: 1
- Bad idea: 1
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