Lecture 01: Introduction to Data Science

Meelis Kull

meelis.kull@ut.ee

Autumn 2019
Getting to know each other

• Meelis Kull
  – Lecture, Thursday 14-16 @ Liivi 2-111
  – Practice group 1, Mon 16-18 @ Liivi 2 – 206
  – Practice group 3, Wed 12-14 @ Liivi 2 – 404

• Laura Ruusmann
  – Practice group 2, Mon 16-18 @ Liivi 2 – 403
  – Practice group 6, Tue 16-18 @ Liivi 2 – 224

• Markus Kängsepp
  – Practice group 4, Wed 12-14 @ Ülikooli 17 – 220
  – Practice group 5, Tue 16-18 @ Liivi 2 – 207
Meelis Kull

• Associate Professor (dentsent) in Machine Learning
• 1998-2011 studied computer science at Tartu:
  – BSc:
    • Supervisor: prof. Mati Tombak
    • Thesis field: theoretical computer science
  – MSc, PhD:
    • Supervisor: prof. Jaak Vilo
    • Thesis field: data mining and bioinformatics
• 2013-2017 Postdoctoral researcher at Bristol, UK
  – Project 1: context change and model reuse in machine learning - http://reframe-d2k.org
  – Project 2: analysis of data from smart homes - http://www.irc-sphere.ac.uk
• Main scientific interests:
  – Machine learning (including deep learning), artificial intelligence, data science
  – Mostly theoretical, but also applications
  – Please contact me for potential supervision if you aim at an excellent thesis in any of these fields

• Examples of studied scientific questions:
  – Why are predictions from machine learning methods (or AI) often over-confident?
  – What can we do about this?
  – How to reuse knowledge learned in earlier tasks?
Have you used classroom answering systems in any courses yet?

A. Yes, the same clickers!
B. Yes, but a different system
C. No
D. Don’t know
Please tell me about yourself

A. 1\textsuperscript{st} year bachelor student  
B. 2\textsuperscript{nd} year bachelor student  
C. 3\textsuperscript{rd} year bachelor student  
D. 1\textsuperscript{st} year master student  
E. 2\textsuperscript{nd} year master master student  
F. PhD student  
G. Other
How many lectures are you planning to attend?

A. All lectures
B. Most lectures
C. Half of lectures
D. Less than half of lectures
E. Probably none after this
### 179 registered students (from SIS)

<table>
<thead>
<tr>
<th>Program</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer Science 2nd year bachelor</td>
<td></td>
</tr>
<tr>
<td>Bioengineering 1st year master</td>
<td></td>
</tr>
<tr>
<td>Foreign/Visiting student</td>
<td></td>
</tr>
<tr>
<td>Computer Science 3rd year bachelor</td>
<td></td>
</tr>
<tr>
<td>Computer Science 1st year bachelor</td>
<td></td>
</tr>
<tr>
<td>Innovation and Technology Management master</td>
<td></td>
</tr>
<tr>
<td>Science and Technology bachelor</td>
<td></td>
</tr>
<tr>
<td>Conversion Master in IT</td>
<td></td>
</tr>
<tr>
<td>Physics, Chemistry and Material Science bachelor</td>
<td></td>
</tr>
<tr>
<td>Mathematics bachelor</td>
<td></td>
</tr>
<tr>
<td>Geoinformatics for Urbanised Society master</td>
<td></td>
</tr>
<tr>
<td>Geography doctoral</td>
<td></td>
</tr>
<tr>
<td>Computer Science 1st year master</td>
<td></td>
</tr>
<tr>
<td>Chemistry 2nd year master</td>
<td></td>
</tr>
<tr>
<td>Software Engineering master</td>
<td></td>
</tr>
<tr>
<td>Robotics and Computer Engineering master</td>
<td></td>
</tr>
<tr>
<td>Quantitative Economics master</td>
<td></td>
</tr>
<tr>
<td>Molecular Biosciences master</td>
<td></td>
</tr>
<tr>
<td>Mathematical Statistics master</td>
<td></td>
</tr>
<tr>
<td>Mathematical Statistics bachelor</td>
<td></td>
</tr>
<tr>
<td>Materials Science master</td>
<td></td>
</tr>
<tr>
<td>Economics doctoral</td>
<td></td>
</tr>
<tr>
<td>Computer Science doctoral</td>
<td></td>
</tr>
<tr>
<td>Computer Science 2nd year master</td>
<td></td>
</tr>
<tr>
<td>Actuarial and Financial Engineering master</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Counts are indicated by the length of the line.*
LTAT.02.002 Introduction to Data Science

- 2018 renamed from MTAT.03.183 Data Mining
- 2004-2017 spring taught by prof. Jaak Vilo
  - I studied this course during its first year 2004
- 2017 autumn taught by Meelis Kull
- Goals:
  - [2004] To learn the basics of the Data Mining and Knowledge Discovery processes and main algorithms used
  - [2018] To get acquainted with the basic concepts and principles of data science and learn to carry out a simple practical data science project
Data mining vs Data science?

• Data science is a more modern term
• Data science is wider, covering more topics
  – Although, there is no common agreement in terms
• More about the difference:
  – Later during this lecture
  – Later during this course
What is your experience with programming in Python?

A. Fluent in Python
B. Not fluent, but completed a course in Python
C. Not fluent, but learned Python myself
D. Haven’t used Python, but I’m ok in programming
E. I haven’t learned programming
Prerequisites for this course

• I will assume that you know how to program in Python
  – You have successfully finished one programming course that taught Python - ok
  – You learned Python yourself - ok
  – You know some other programming language very well and you will learn Python individually during the few following weeks - ok
  – You have never programmed before – not ok

• I will assume that you understand basic math
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 is Data Science?

• Discussed in a panel at the Opinion Festival 2019 (in Estonian)

Mis on andmeteadus?
Ürituse lava: Tehnoloogiaala

Arutelu juht:
Mart Mägi
Osalejad:
Ene-Margit Tiit, Taivo Pungas, Meelis Kull, Kristjan Vassil, Krista Fischer

Järelikuulamine (1 hour 34 min):
https://soundcloud.com/arvamusfestival/mis-on-andmeteadus
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who 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
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Py
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- 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
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Downloaded from: http://www.ciselab.org/
MODERN DATA SCIENTIST

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

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

Domain Knowledge & Soft Skills
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- 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
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Downloaded from: http://www.ciselab.org/
Work with big data

**Programming & Database**

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package 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 xaaS like AWS
Analyse data

**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
Become domain expert

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative
Communicate the results

COMMUNICATION & VISUALIZATION

🌟 Able to engage with senior management
🌟 Story telling skills
🌟 Translate data-driven insights into decisions and actions
🌟 Visual art design
🌟 R packages like ggplot or lattice
🌟 Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection 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
- Scripting language e.g. Python
- Statistical computing package 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 xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- 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
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Downloaded from: http://www.ciselab.org/
Modern Data Scientist

A data scientist is a job title for an employee or business intelligence (BI) consultant who excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge.

Downloaded from: http://hpc-asia.com/modern-data-scientist/
What is Data Science?

• “Data Science is statistics on a Mac”
• “A data scientist is a statistician who lives in San Francisco”
• “A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician”

• From: https://datascopeanalytics.com/blog/what-is-a-data-scientist/
Downloaded from: http://nirvacana.com/thoughts/becoming-a-data-scientist/
LTAT.02.002 Intro to Data Science

• This course could be your first step towards becoming a data scientist
• Provides a short introduction to many of the above topics

• Want to become a data scientist?
  – Go for the computer science master curriculum in Tartu and take relevant courses along the way
  – Or study statistics and make sure you learn enough computer science along the way
  – Or do PhD in computer science or statistics
Many data scientist jobs available

- In Estonia, for example:

What we offer you:
- Salary of €3000+, based on qualifications
- Highly qualified colleagues
- Small development team
- Flexitime and more
- Participation in training
- Cozy office at the Digicentre
- Frequent all-inclusive lunch

What you get to do:
- Perform data wrangling and analysis
- Create statistical and machine learning models
- Partner with clients to deliver data science solutions, including associated documentation and presentations
- Define activities, scope, and timelines on data science projects
- Help to develop the skills of the entire team as data-savvy leaders who ask great questions and make informed decisions

What helps you succeed:
- Advanced analytical skills in Python and/or R
- Excellent data wrangling skills
- Experience with machine learning
- Expertise in SQL and NoSQL databases
- Excellent oral and written communication skills
- Experience in building/delivering data science products
- Ability to tell business story using data
- Bonus points for experience in recommender systems
Many data scientist jobs available

• In the world:

As a Junior Data Scientist, you will have:
- PhD degree in a Computer Science or Statistics based subject (or similar)
- Knowledge of Python/R
- Excellent communications skills to enable you to talk to both the technical members of staff and the business

Junior Data Scientist - Up to £50,000 DOE - London

Key Skills

- Commercial experience using R or SQL
- Relevant experience working within data analytics
- Someone who can learn new programs and put them to practice quickly
- Someone who is passionate about Big Data & Data Science

https://www.indeed.co.uk/Data-Scientist-jobs
Data scientist: MSc or PhD?


William Chen, Data Science Manager at Quora

Updated Apr 11 · Upvoted by Jalem Raj Rohit, Sr. Data Scientist at Episource and Lili Jiang, Data Scientist at Quora

*Originally Answered: Does one need to do a MS/PHD to be a data scientist?*

**No, you do not need either, but you need the right background.**

Candidates coming out of certain MS / PhD programs may have advantages in data science because at least one of these are true: They:
- do research involving programming and large datasets
- have gathered statistical and data intuition through their work
- show resilience when asking / answering hard questions
- can explain the motivations and reasoning behind their work
- are able to think critically about hard problems
- can learn and adapt quickly

If you want to be a data scientist in a company or a team where the product is primarily based on data science, then the bar tends to be a lot higher, and a Masters/PhD is typically required.
This course teaches you to:

- List and explain DS subfields, concepts, tasks
- Perform a **descriptive analysis** of a given dataset
- Choose and implement appropriate **visualisations**
- Apply common **clustering** methods, interpret results
- Choose and apply basic **machine learning** algorithms to build predictive models for classification, regression
- Use basic **evaluation** methods to measure predictive performance of models obtained by machine learning
- Interpret the results from a **statistical test**
- Plan and implement a simple **practical DS project**
This course **will not** teach you

- How to build AI
  - However, working with data is also important for AI
  - [LTAT.01.003 Artificial Intelligence](#)

- How to train deep neural networks
  - However, we will learn about what they are
  - [LTAT.02.001 Neural Networks](#)

- How to make machines learn by themselves
  - Well, actually we do have some lectures about that
  - [MTAT.03.227 Machine Learning](#)

- How to do a proper statistical analysis
  - However, some basic statistics will be used
  - [MTMS.02.059 Probability and Mathematical Statistics](#)
Which term is currently more popular?

A. Data science
B. Data mining
C. Equally popular

Response Counter

0 0 0

Data science  Data mining  Equally popular
Data scientist must be particularly good in …

A. Programming & Databases
B. Math & Statistics
C. Domain Knowledge & Soft Skills
D. Communication & Visualisation
E. All of the above

Response Counter
Who is the perfect data scientist?

The Data Scientist Venn Diagram

- Statistics
  - The Data Nerd
  - The Stats Prof
  - Comp Sci Prof
  - Head of IT
  - The Analyst
  - The Salesperson
  - The perfect Data Scientist!

- Programming
  - R Core Team
  - The Hacker
  - The IT Guy
  - Drew Conway's Data Scientist
  - The Number Cruncher

- Business
  - The Accountant

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
Ten success stories of data science
Story 1: Cholera
Year 1854, London

- On the night of 31 Aug 1854, after a hot day:
  - Many people became violently ill of cholera in Soho, London

- Following few days:
  - 89 people died
  - 75% of the population left this area
  - Dr John Snow started an investigation

- 7 days later, on 7 Sept 1854:
  - Dr John Snow solved the mystery and convinced the authorities of the best action to take
Dr John Snow’s map of cholera
Dr John Snow’s map of cholera
Dr John Snow’s map of cholera
A modern update to the map
Sept 7, 1854 (7 days after outbreak)

- Based on data, Dr John Snow identifies the Broad Street Pump as the source of cholera
- He shows his **data visualisation** and convinces the authorities to remove the handle of the pump to stop usage
- Infections dropped at once
- Read more at: [http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html](http://www.ph.ucla.edu/epi/snow/snowcricketarticle.html)
Story 2: Weather forecasting
Year 1950, Weather Forecasting

• First successful numerical prediction of 24h in advance, taking almost 24h to calculate

• People involved:
  – US meteorologists
  – John von Neumann (mathematician)
  – Klara Dan von Neumann (programmer)

• Computer:
  – ENIAC

• More information:
Weather Forecasting Now

Source: https://www.meteoblue.com/en/weather/14-days/tartu_estonia_588335
Story 3: Databases
Databases

• 1970: relational model of data proposed
• 1974: standard query language (SQL) was developed
• 2000s: NoSQL and NewSQL systems

Sources:
https://en.wikipedia.org/wiki/SQL
https://en.wikipedia.org/wiki/Database
More data than A4 papers ever fit

- $10^{13}$ trees on Earth
- $10^5$ A4 sheets from 1 tree
- $10^2$ rows of text per A4
- $10^2$ bytes per row of A4 (assuming 1 byte per character)
- Papers from all trees can fit:
  - $10^{22}$ bytes
- Currently data in the digital world:
  - $3 \times 10^{22}$ bytes (30 zettabytes)

Sources:
- [https://www.quora.com/How-many-trees-are-there-on-Earth](https://www.quora.com/How-many-trees-are-there-on-Earth)
Story 4: Spreadsheets
Spreadsheet Software

• 1979: VisiCalc
  – First software that combined all essential features of modern spreadsheet applications

• 1985: Microsoft Excel
  – Opened up data processing for all computer users

Source: https://en.wikipedia.org/wiki/Spreadsheet
Story 5: Game of Chess
Chess - IBM Deep Blue

• 1996: IBM Deep Blue wins Garry Kasparov
  – First chess engine to win a chess game against a reigning world champion

• Used data:
  – Opening books with 700k grand master games
  – Endgame database

• Demonstrated the **usefulness of data for AI**

Story 6: Internet search
Year 1997, Internet Search

- The best search engine looked like this:
Then came: Google PageRank

• Google PageRank algorithm:
  – rates pages based on the number and importance of links that point to them

• Uses data from the collective intelligence of the Internet
Story 7: Recommendations
Amazon Company

- 1994: Amazon Company founded
- 2001: Amazon turned its first profit
- Amazon uses data to guess what people want and provide recommendations
  - They even possess a patent that permits them to ship merchandise even before it’s ordered
- Recommendation engines are examples of using data for businesses

Story 8: Autonomous driving
Self-driving cars

• 2005: Stanley wins DARPA Grand Challenge
  – Team from Stanford University
  – Drives autonomously 12km off-road
  – Team lead Sebastian Thrun later started to lead Google Self-Driving car development

• 2016: Tesla accident
  – First casualty during self-driving mode

• Self-driving cars are an example of using data for robotics and automation

Source: https://en.wikipedia.org/wiki/Autonomous_car
Story 9: Question answering
Question answering

• 2011: IBM Watson wins Jeopardy!
  – Competing against former winners
  – Not connected to Internet
  – 4TB disk storage, including Wikipedia

• This is an example of using **natural language as data**

Story 10: Game of Go
AlphaGo of Google DeepMind

• 2016: AlphaGo wins 4-1 over world champion Lee Sedol

• AlphaGo used:
  – Deep neural nets
  – Monte Carlo search
  – Data from human and computer play

• 2017 AlphaGo Zero:
  – Learning from scratch, without data from human play

• Example of generating synthetic data and then learning from it

Source: https://en.wikipedia.org/wiki/AlphaGo
Some messages from 10 stories

- Cholera – data **visualisation matters**
- Weather forecasting – data **affects everyone**
- Databases – **management** of growing data
- Spreadsheets – data **processing for all**
- Game of Chess – data for **AI**
- Internet search – data of **collective intelligence**
- Recommendations – data for **businesses**
- Autonomous driving – data for **robotics** and **AI**
- Question answering – **natural language** as data
- Game of Go – generating **synthetic** data for **AI**
I learned the most from story ...

A. Cholera
B. Weather forecasting
C. Databases
D. Spreadsheets
E. Game of Chess
F. Internet search
G. Recommendations
H. Autonomous driving
I. Question answering
J. Game of Go
More data science stories?

• 10 most popular TED talks about data science:

• [in Estonian] Taivo Pungas “Miks andmeteadus” (Why data science) TEDxYouth@Tallinn
  - https://www.youtube.com/watch?v=TEiaIfMuydQ&t=4s
AI reaches human level in understanding and reasoning about the world in:

A. Less than 5 years
B. 5-10 years
C. 10-20 years
D. 20-30 years
E. 30-50 years
F. 50-100 years
G. More than 100 years
H. Never
I would now like a break of ...

A. 0 minutes
B. 5 minutes
C. 10 minutes
D. 15 minutes
E. Fine with any of the above