MTAT.03.183 - Data Mining - Lecture 01

Introduction to the Data Mining Course

Meelis Kull

meelis.kull@ut.ee

Autumn 2017
Getting to know each other

• Meelis Kull
  – Lecture, Monday 10-12 @ Liivi 2-111
  – Practice group 1, Monday 12-14 @ Liivi 2-402

• Mari-Liis Allikivi
  – Practice group 2, Monday 14-16 @ Liivi 2-122

• Dmytro Fishman
  – Practice group 3, Monday 12-14 @ Liivi 2-405
  – Practice group 4, Monday 14-16 @ Liivi 2-405
Meelis Kull

• Associate Professor (dotsent) in Data Mining
• 1998-2011 studied informatics at Tartu:
  – BSc:
    • Supervisor: prof. Mati Tombak
    • Thesis topic: theoretical informatics
  – MSc, PhD:
    • Supervisor: prof. Jaak Vilo
    • Thesis topics: 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, artificial intelligence, data mining, data science, applications in smart homes, health, smart electric grids, etc.
- **Please contact me if looking for thesis topics or supervision**

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 combine data from many sources to improve predictions?
Is it a good idea to use clickers in data mining lectures?

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

A. 1st year master student
B. 2nd year master student
C. PhD student
D. Bachelor student
E. Other
I will probably attend …

A. All lectures
B. Most lectures
C. Half of lectures
D. Less than half of lectures
E. Probably none after this

![Bar chart showing the distribution of responses.]

- All lectures: 47
- Most lectures: 26
- Half of lectures: 5
- Less than half of lectures: 2
- Probably none after this: 4

Meelis Kull - Autumn 2017 - MTAT.03.183 - Data Mining - Lecture 01
131 registered students (from SIS)

<table>
<thead>
<tr>
<th>Curriculum</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software Engineering</td>
<td>57</td>
</tr>
<tr>
<td>Computer Science</td>
<td>52</td>
</tr>
<tr>
<td>N/A</td>
<td>9</td>
</tr>
<tr>
<td>Conversion Master in IT</td>
<td>2</td>
</tr>
<tr>
<td>Semiotics and Culture Studies</td>
<td>1</td>
</tr>
<tr>
<td>Robotics and Computer Engineering</td>
<td>1</td>
</tr>
<tr>
<td>Political Science</td>
<td>1</td>
</tr>
<tr>
<td>Molecular and Cell Biology</td>
<td>1</td>
</tr>
<tr>
<td>Media and Communication</td>
<td>1</td>
</tr>
<tr>
<td>Mathematics and statistics</td>
<td>1</td>
</tr>
<tr>
<td>Materials Science</td>
<td>1</td>
</tr>
<tr>
<td>Information and Knowledge Management</td>
<td>1</td>
</tr>
<tr>
<td>Excellence in Analytical Chemistry</td>
<td>1</td>
</tr>
<tr>
<td>Economics</td>
<td>1</td>
</tr>
<tr>
<td>Actuarial and Financial Engineering</td>
<td>1</td>
</tr>
</tbody>
</table>

count
MTAT.03.183 Data Mining

• Taught since 2004 by prof. Jaak Vilo
  – [Link](http://www.quaretec.com/u/vilo/edu/2004-05/Andmekaevandus/)
  – I studied in this course during its first year
  – Now taking teaching over from Jaak

• Goals:
  – [2004] To learn the basics of the Data Mining and Knowledge Discovery processes and main algorithms used
  – [2017] This course aims at introducing students to the concepts and principles of data mining and knowledge discovery theory and practice
Plan to rename the course into **Data Science**

**Difference of data mining and 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
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
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 modelling
- 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 ‘as’ 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. Tableau, Tableau

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

### Skillsets

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

**Marketing Distillery**

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

- 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 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 modelling
- 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 xoo’S 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.: Tableau, D3, Qlikview

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/

Meelis Kull - Autumn 2017 - MTAT.03.183 - Data Mining - Lecture 01
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.
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/
This course could be your first step towards becoming a data scientist

Provides a short introduction to many of the above topics

Many more steps to be taken
Many data scientist jobs available

• In Estonia, for example:

What we offer you:
• Salary of €3000+
• Highly qualified colleagues
• Small development team
• Flexitime and mobility
• Participation in the decision-making process
• Cozy office at the city center

What you get to do:
• Perform data wrangling and analysis tasks
• Create statistical and machine learning models in multiple technologies to support client goals
• 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:

Junior Data Scientist -

- Central London / West End (WC1), C
- From £40,000 to £50,000 per annum
- Data Idols
- Permanent
- Expires in 2 days

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

what

Data Scientist

job title, keywords or company

Find Jobs

Jobs 1 to 10 of 3,231
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:

- Many different **forms of data**
- Many ways of **organising** the data
- Many ways of **querying** the data
- Many **tasks** to solve with data
- Many **algorithms** to solve these tasks
- Some ways of **visualising** the results
- Some **software tools** that allow to do all this
This course will not teach you

- How to build AI
  - However, data mining is 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
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

[Bar chart showing the distribution of responses]
Who is the perfect data scientist?

The Data Scientist Venn Diagram

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
10 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
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/forecast/multimodel
Weather Forecasting Now

<table>
<thead>
<tr>
<th></th>
<th>Sat</th>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIRLAM11 53m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>ICON13 71m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>ARPEGE11 52m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>ARPEGE40 62m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>ICON7 48m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>GFS22 52m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>NEMS12 51m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>NEMS2-12 51m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>NMM12 64m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>NEMS30 63m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>GEM25 72m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
<tr>
<td>NEMS2-30 63m asl</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
<td>☁️</td>
</tr>
</tbody>
</table>

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:
  - $2 \times 10^{22}$ bytes

Sources:
- 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**

Source: https://en.wikipedia.org/wiki/Watson_(computer)
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

• 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](https://www.youtube.com/watch?v=TEiaIfMuydQ&t=4s)
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
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

• http://datasci.ee/ [in Estonian]

Data Science Estonia | Eesti andmeteaduse kommuun

• https://www.facebook.com/groups/datasci.ee/ [Closed Group] 397 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
      – 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, …
    • Research fellows, lecturers, PhD students, Master students, …
  – Institute of Mathematics and Statistics
  – …

• Tallinn University of Technology
  • …
Data science in Estonia: 3 examples from industry
STACC is the leading data science center in Estonia that develops machine learning models and data analytics solutions.
THE FUTURE OF DATA-DRIVEN ENERGY

Planet OS provides big data infrastructure to help renewable energy companies transform the way data is used in their organizations. From operators in control rooms to executives in boardrooms, our specialized applications to integrate, exchange, and visualize data for all stakeholders make renewable energy more competitive.
THE FUTURE OF DATA-DRIVEN ENERGY

AN API FOR THE PLANET

Discover and access data from the fastest-growing data catalog of weather, climate and environmental data.

Select from over 2,100 variables...
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

- **By above definitions and course contents the course should be called Data Science**
Science of / with Data

• Science of Data:
  – Science about how to use data
  – Science of Data = Data Science

• Science with Data:
  – Doing science with the help of data
  – Science with Data ≠ Data Science
  – At least according to my terminology

• Data scientist in industry
  – Seldom researches data science
  – Sometimes researches science with data
  – Usually just applies data science methods
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
I am developing an algorithm to recognise birds from photos

A. This is data science research
B. This is not data science research
C. Not sure whether it is data science research
I am applying an algorithm to find anomalies in data

A. This is data science research
B. This is not data science research
C. Not sure whether it is data science research
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

CROSS INFRASTRUCTURE/ANALYTICS

FRAMEWORK
QUERY/DATA FLOW
DATA ACCESS
COORDINATION
STREAMING
LOGICAL/LOG ANALYSIS VISUALIZATION COLLABORATION SECURITY

OPEN SOURCE

DATA SOURCES & APIS

IOT
FINANCIAL/ECONOMIC DATA
ARCH

DATA RESOURCES

FIRSTMARK
CREATIVITY VENTURE CAPITAL

UNIVERSITY OF TARTU
Institute of Computer Science

Meelis Kull - Autumn 2017 - MTAT.03.183 - Data Mining - Lecture 01

V2 - Last updated 5/3/2017
© Matt Turck (@matturck), Jim Hao (@jinhao), & FirstMark (@firstmarkcap) matturck.com/bigdata2017
A PERIODIC TABLE OF VISUALIZATION METHODS

Data Visualization
Visual representations of quantitative data in schematic form (either with or without cues)

Strategy Visualization
The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.

Information Visualization
The use of interactive visual representations of data to simplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with it.

Metaphor Visualization
Visual metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed.

Concept Visualization
Methods to elaborate (nearly) qualitative concepts, ideas, plans, and analyses.

Compound Visualization
The complementary use of different graphic representation formats in one single scheme or frame.

Process Visualization

Structure Visualization

Overview

Detail

Detail AND Overview

Divergent thinking

Su supply demand curve
Pc performance charting
St strategy map
Oc organization chart
Ho house of quality
Ft fixture tree
Mq magic quadrant
Ld life-cycle diagram
Po Porter’s five forces
S i-cycle
Sm sticker map
Is Ishikawa diagram
Tc technology roadmap

Ed equipment
Pf portfolio
Sg supplier
Mz manufacturer’s
Z Ad zed
De decision
Bm best matrix
Stc strategy
Vc value chain
Hy hypcycle
Sr sterilizer
Ta taut
Sd sardine

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

© Ralph Lengler & Martin J. Eppler; www.visual-literacy.org

version 1.5
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 R programming language
- Learn to use some useful R packages
- 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 MTAT.03.183 Data Mining

• Getting the data
• Preprocessing and exploration
• Pattern mining
• Big data
• Curse of dimensionality
• Clustering
• Classification and regression
• Probabilistic modelling
• Deep learning
• Standards in data mining
• Applications
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:
  – Registration is part of the first homework (HW1)

• Homework submission and feedback:
  – https://gradescope.com/courses/10132
  – You will get invitations to register by today evening
Meeting times

• Lecture:
  – Monday 10-12 @ Liivi 2-111

• Practice sessions (start next week Sept 18):
  – G1 (Meelis Kull): Mon 12-14 @ Liivi 2-402
  – G2 (Mari-Liis Allikivi): Mon 14-16 @ Liivi 2-122
  – G3 (Dmytro Fishman): Mon 12-14 @ Liivi 2-405
  – G4 (Dmytro Fishman): Mon 14-16 @ Liivi 2-405

• Consultation hour (starts this week Sept 13)
  – Wednesday 16-17 @ Liivi 2-611 (voluntary)
This is how the course works

• Monday 10:15-12:00:
  – new material presented in the lecture

• Monday evening:
  – 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

• Wednesday 16:00-17:00 (note the start 16:00)
  – Voluntary consultation session, a chance to ask more

• Sunday 23:59
  – Deadline for submitting homework into Gradescope

• Monday afternoon:
  – Homework solutions discussed in the practice session

• Monday-Sunday:
  – Homework graded by the instructor of your practice group
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 pt
  – Group project and poster presentation (20 points)
  – Written exam (36 points)
• Additional points can be earned from bonus tasks in homeworks
• Attending at least 8 out of 11 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)
Group project

• Teams of 2-3 students
• Every team must present the project as a poster in the poster session on Jan 10, 2018
• Projects start officially towards the end of the course, after regular homeworks end
• You can start forming teams and thinking about potential topics already now!
• More information coming during the course
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 11 practice sessions (22h)
- 25 hours per person on project (25h)
- 15 hours on preparing for the exam (15h)
- Total: 156 hours
Live demonstration of homework upload into Gradescope
Is it a good idea to use clickers in data mining lectures?

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

• It is also possible to use your phone or laptop instead
  – Go to https://responseware.turningtechnologies.eu/responseware/
  – Download the TurningPoint app or use browser
  – Log into session ‘datamining’
  – You can leave name, e-mail etc empty
  – Now you can answer polls, see results
  – Additionally can send me anonymous messages during the lecture (questions / comments)
    • Hopefully I notice them soon on my laptop screen
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