Lecture 01: Introduction to Data Science

Meelis Kull

meelis.kull@ut.ee

Autumn 2022
Getting to know each other

• Meelis Kull
  – Lectures

• Victor Henrique Cabral Pinheiro
  – Practice groups 1, 2, 3, 7 (Eng)

• Markus Kängsepp
  – Practice group 4 (Est)

• Ingvar Baranin
  – Practice group 5 (Est)

• Friedrich Krull
  – Practice group 6 (Est)

• Anna Aljanaki
  – Practice group 8 (Est)
Meelis Kull

- Associate Professor in Machine Learning (masinõppe kaasprofessor)
- 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
  – 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?

• **Current main project:**
  – Contextual uncertainty and representation learning in machine perception
Lecture format

• There are more than 200 of us but questions are still welcome:
  – Unmute and say that you have a question
  – Or ask it in the Zoom chat
  – Keep in mind that we cannot spend more than 10 minutes on questions/answers without going over the regular 90 minutes

• We will have anonymous poll questions during the lecture:
  • Please try to participate in these, this can help learning
  • https://participant.turningtechnologies.eu/polling/ids2022
Have you used classroom answering systems in any courses yet?

https://participant.turningtechnologies.eu/polling/ids2022

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

https://participant.turningtechnologies.eu/polling/ids2022

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 student
F. PhD student
G. Other
Students from different fields

- Computer Science
- Conversion Master in IT
- Computer Engineering
- Mathematics
- Engineering and Technology
- Actuarial and Financial Engineering
- Bioengineering
- Physics, Chemistry and Material Science
- Robotics and Computer Engineering
- Biology and Biodiversity Conservation
- Botany and Ecology
- Gene Technology
- Geography
- Science and Technology
- Economics and Business Administration
- Innovation and Technology Management
- Quantitative Economics
- Teaching Natural and Exact Sciences at Lower Secondary School
- Special Education and Speech Therapy
- Semiotics and Culture Studies
- Estonian and Finno-Ugric Linguistics
- Romance Studies
How many lectures are you planning to attend live?

https://participant.turningtechnologies.eu/polling/ids2022

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

58% said they plan to attend all lectures.
What would be your preference for lectures?

https://participant.turningtechnologies.eu/polling/ids2022

A. Online  
B. Physical @ Delta  
C. Either is fine  
D. I don’t know
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
• Since 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?

https://participant.turningtechnologies.eu/polling/ids2022

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](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 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 xooS 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

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, econometrics, data warehousing and big data systems. Marketing Distillery is the primary data science department in PaulSiian’s, STS, Social, CRM and brand.

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.

**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, Qlik, Power BI

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

**COMPUTER SKILLS & TOOLS**
- 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

MarketingDistillery.com is a group of practitioners in the area of e-commerce, marketing and e-commerce solutions. Our fields of expertise include marketing, analytics, customer insights, and data science. We help companies improve their business performance through data-driven insights and decision-making.

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

DOMAIN KNOWLEDGE & SOFT SKILLS

🌟 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 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 xooS 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.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, econometrics, data warehousing and big data systems, marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Downloaded from: http://www.ciselab.org/
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/
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 data science master curriculum in the University of Tartu
  – Or study computer science or statistics and make sure you learn the other field also enough
  – Or do PhD in computer science or statistics
Many data scientist jobs available

• In Estonia, for example:
Many data scientist jobs available

• In the world:

new

Data Scientist and Data Science Researcher
GCHQ 4.2 ★
Cheltenham

£37,639 - £51,710 a year

Data Scientist Requirements

• Knowledge of or/and experience working in either Data Science or software engineering and talking to experts in both fields.
• Knowledge of machine learning training, deployment, and maintenance.
• An understanding of exploratory data analysis, statistical reasoning, communicating about data, and data-driven decision making.
• Familiarity with modern software engineering best practices (e.g. agile) and understand how they could be applied to data science projects.
• The technical knowledge to take a prototype ML capability (e.g., in a Jupyter notebook) and build it into a functional system component.
• Proficiency with essential technologies including but not limited to – Python, R, Kubernetes, Docker, JavaScript, Java.
• Familiarity with other useful technologies- ELK, AWS would be beneficial.

https://www.indeed.co.uk/Data-Scientist-jobs
Many data scientist jobs available

• In the world:

Data Scientist and Data Science Researcher
GCHQ 4.2 ★
Cheltenham

£37,639 - £51,710 a year

Data Science Researcher Requirements

We are looking for talented researchers to help us stay at the cutting edge of data science, machine learning and AI. We have positions available to match a range of experience levels, but you will have a strong background in one or more Data Science disciplines. You’ll be using your technical expertise to develop new approaches to complex problems that you won’t find anywhere else. With a diverse set of customers from across our mission no two projects are the same and you’ll be delivering high-quality research either individually or as part of a team. You’ll work closely with other agencies, organisations, and commercial partners; as well as working with organisations that contribute to our corporate social responsibility agenda which may include the environment and inspiring future generations. A positive team player, you’ll help to build a diverse, supportive and effective Data Science Research community, and develop and mentor the next generation of data scientists.

- Experience working in artificial intelligence, machine learning or data science in a research capacity.
- A deep understanding of one or more relevant technologies, e.g. neural networks, statistical methods etc.
- The ability to solve complex problems, both through personal research and leadership of research teams.
- Experience working with academia, implementing and building on published research.
- Proficiency with essential technologies, ideally Python and common data science libraries, e.g. PyTorch.
- The ability to work with customers to turn research into mission applications and identify new opportunities for research.

https://www.indeed.co.uk/Data-Scientist-jobs
Many data scientist jobs available

- In the world: https://www.indeed.co.uk/Data-Scientist-jobs

**Junior Data Scientist**
Mosaic Smart Data
London EC2M

- Full-time

- Developing new data science models (from inception to rollout), by communicating with clients, performing exploratory data analysis and generating informative...

30+ days ago · More...
Many data scientist jobs available

In the world:

https://www.indeed.co.uk/Data-Scientist-jobs

Junior Data Scientist
Mosaic Smart Data
London EC2M

What We Are Looking For
Required

- Masters or PhD in a STEM subject or industry experience to compensate
- Good programming skills in Python
- Experience with coding best practices, test-driven development, code review and Git
- Some experience with data science, finance or both
- Some understanding of machine learning techniques and algorithms (e.g. k-NN, Naive Bayes, Decision trees, neural networks)
- Some experience with common data science toolkits (e.g. Pandas, NumPy, SciKit Learn, TensorFlow)
- Active team member willing to help and mentor colleagues across disciplines
- Data-oriented personality with the ability to solve complex problems in a pragmatic way
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

https://participant.turningtechnologies.eu/polling/ids2022
Data scientist must be particularly good in ... 

https://participant.turningtechnologies.eu/polling/ids2022

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

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
- **2017 AlphaGo Zero:**
  - Learning from scratch, without data from human play
- **2017 AlphaZero:**
  - Same model architecture for chess, go, shogi
- Example of generating synthetic data and then learning from it

Source: [https://en.wikipedia.org/wiki/AlphaGo](https://en.wikipedia.org/wiki/AlphaGo)
Story 11: Foundation Models
Example using a foundation model: Flamingo
Sources: [https://en.wikipedia.org/wiki/Foundation_models](https://en.wikipedia.org/wiki/Foundation_models)
Bommasani et al 2021: On the Opportunities and Risks of Foundation Models
Foundation Models

- The revolution fuelled by transformers (attention-based neural network models)
- 2018: GPT by OpenAI
- Later: BERT, GPT3, CLIP, DALL-E, Flamingo, Florence
- Typically predict what should come next in a sequence (text, images, …)
- This is an example of AI that you can build upon

Sources: https://en.wikipedia.org/wiki/Foundation_models
Bommasani et al 2021: On the Opportunities and Risks of Foundation Models
https://arxiv.org/abs/2108.07258
Some messages from 11 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
• Foundation models – **build-upon** AI
I learned the most from story ...

https://participant.turningtechnologies.eu/polling/ids2022

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

26% 15% 14% 14%

7% 4% 6% 8% 4%
More data science stories?

• 10 most popular TED talks about data science:

• Best TED talks for 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

https://participant.turningtechnologies.eu/polling/ids2022
I would now like a break of ...

https://participant.turningtechnologies.eu/polling/ids2022

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

101 votes

- 65% for 0 minutes
- 15% for 5 minutes
- 3% for 10 minutes
- 3% for 15 minutes
- 14% for 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


Data Science Estonia | Eesti andmeteaduse kommuun

- [https://www.facebook.com/groups/datasci.ee/](https://www.facebook.com/groups/datasci.ee/) · Private group
  - September 2017 – 400+ members
  - September 2022 – 1800+ 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
      – Ahmed Awad, Prof of Big Data
      – Mark Fišel, Prof of Natural Language Processing
      – Jaak Vilo, Prof of Bioinformatics
      – Hedi Peterson, Prof of Bioinformatics
      – Marlon Dumas, Prof of Information Systems
    • Associate professors:
      – Jaan Aru, Radwa El Shavi, Huber Flores, Raivo Kolde, Meelis Kull, Sven Laur, Leopold Parts, Dirk Oliver Theis, Kuldar Taveter, …
    • Research fellows, lecturers, PhD students, Master students, …
      – Institute of Mathematics and Statistics
      – …
  • Tallinn University of Technology
  • …
Data science in Estonia: 4 examples from industry
Building **AI-driven organizations**

Solving business problems with advanced AI and machine learning solutions for public and private sectors.

Get in touch

- 20 Projects in 20+ countries
- 75+ AI projects delivered
- 85% 85%+ project success rate
Statistical insights from mobile data

Positium provides innovative data solutions and analytics from mobile positioning data in mobility, tourism and population.
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-multiprofessional.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

<table>
<thead>
<tr>
<th>Data Visualization</th>
<th>Strategy Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual representations of quantitative data in schematic form (either with or without axes)</td>
<td>The systematic use of complementary visual representations in the analysis, development, documentation, communication, and implementation of strategies to organizations.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information Visualization</th>
<th>Metaphor Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>The use of interactive visual representations of data to simplify cognition. This means that the data is transformed into images, which are mapped to screen space. The image can be changed by users as they proceed working with it.</td>
<td>Visual metaphors position information graphically to organise and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept Visualization</th>
<th>Compound Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.</td>
<td>The complementary use of different graphic representation formats in one single scheme or frame.</td>
</tr>
</tbody>
</table>

---

**Note:** Depending on your location and connection speed, it can take some time to load a pop-up picture. [Source](www.visual-literacy.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://campuswire.com/c/GAF3154B4](https://campuswire.com/c/GAF3154B4)
  – Electronic communication is only through Campuswire
    (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 = midday (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
  – except for medical reasons (contact the practice group’s teacher)
• 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 19, at noon = midday (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

*In case of pandemic restrictions we will either have an online exam or cancel the exam and multiply the points from the homeworks and project by 1.67
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
• 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/ids2021

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

- Absolutely: 63%
- Probably good: 27%
- Not sure yet: 6%
- Probably bad: 0%
- Bad idea: 3%
- Prefer using some other...: 1%
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