Data Science Basics

RBS Greenhouse

RTU Riga Business School Greenhouse program

Greenhouse.rbs.lv

About Me

Valdis Saulespurens

30+ Years programming

Lector RTU RBS

Reasearcher National Library of Latvia

valdis.saulespurens@rtu.lv

Data Lake

Data Science intro

Data science is like being a detective, but instead of solving crimes, you're uncovering hidden truths and secret patterns in mountains of data!

You'll use all kinds of cool tools and techniques from statistics, computer science, artificial intelligence and your own expertise to explore data from all sorts of sources. And the best part is you get to use all of that information to make predictions and decisions that can change the world!

Data Science definition

Data science - interdisciplinary field that involves using scientific methods, processes, systems, and infrastructures to extract knowledge and insights from structured and unstructured data.

Applications of Data Science

Forecasting the likelihood of severe storms using data from weather sensors.

Applications of Data Science

– Using data from GPS and social media to recommend the best places to eat or hang out with friends in a city.

Applications of Data Science

- Using data from fitness trackers and wearables to monitor personal health and make personalized workout and nutrition recommendations.

Applications of Data Science

– Predicting product preferences using data from social media

Applications of Data Science

– Using data from online
  learning platforms to
  identify which students
  are struggling and
  providing targeted
  support to help them
  succeed.

Applications of Data Science

– Using data from traffic cameras and sensors to optimize traffic flow and reduce congestion on roads and highways.

Applications of Data Science

– Using data from social media to understand public opinion and sentiment on different issues and topics.

Data Science in high school

THE ABOVE LIST IS JUST SMALL SAMPLING, THERE ARE MANY MORE!

WHAT COULD BE SOME POTENTIAL USE OF DATA SCIENCE IN HIGH SCHOOL?

Sub-disciplines of Data Science

Machine learning:

a subfield of data science that focuses on developing algorithms and models that enable computers to learn from data and make predictions or decisions without being explicitly programmed to do so.

ML has multiple subfields such as:

Supervised learning, unsupervised learning and..

Sub-disciplines of Data Science

Deep learning:

a subset of machine learning that involves the use of neural networks to model complex patterns in data

type of mathematical model inspired by the way the human brain works.

Neural networks are composed of layers of interconnected nodes, or artificial neurons, that process and transmit information. Each layer of neurons can learn to extract a different level of abstraction from the input data. The more layers a neural network has, the "deeper" it is, thus the term "deep learning".

Sub-disciplines of Data Science

Natural language processing (NLP):

a subfield of data science that deals with the interaction between computers and human languages, and includes tasks such as language translation, text summarization, and sentiment analysis.
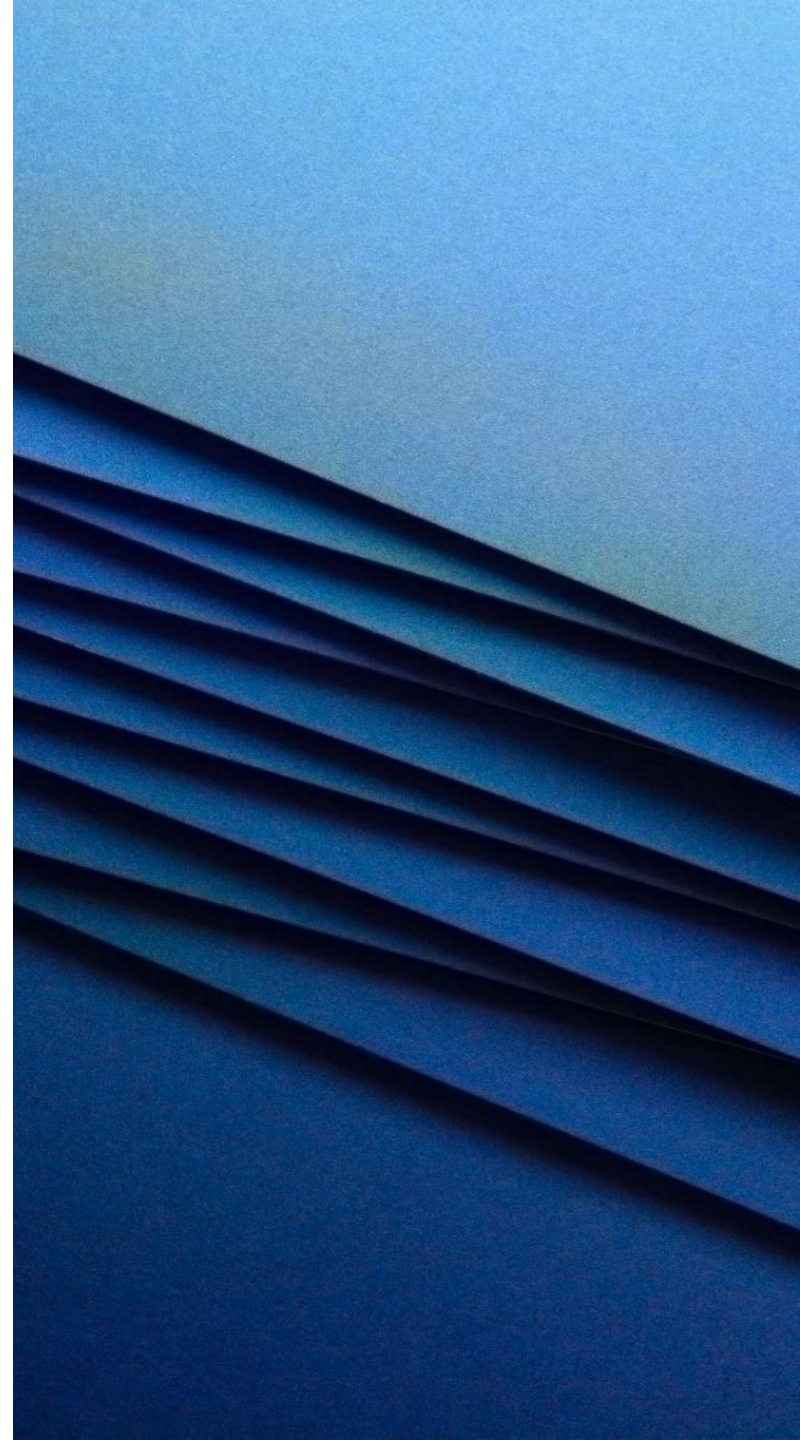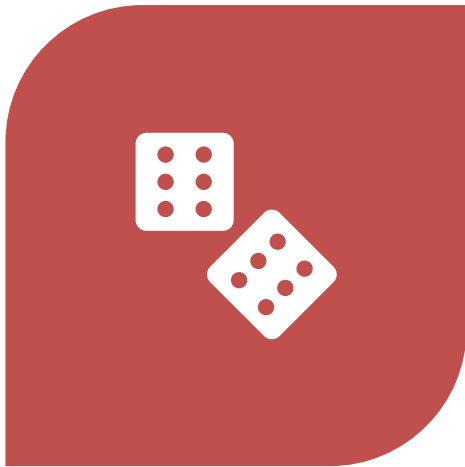
Sub-disciplines of Data Science

Computer vision:

a subfield of data science that deals with the use of computer algorithms to understand and interpret visual data, such as images and videos.

Sub-disciplines of Data Science

PREDICTIVE MODELING:

A SUBFIELD OF DATA SCIENCE THAT USES STATISTICAL AND MACHINE LEARNING TECHNIQUES TO BUILD MODELS THAT CAN PREDICT FUTURE EVENTS OR OUTCOMES BASED ON HISTORICAL DATA.

Sub-disciplines of Data Science

Recommender Systems:

a subfield of data science that builds systems that can predict what a user may be interested in based on their past interactions and behaviors.

Sub-disciplines of Data Science

Anomaly detection:

a subfield of data science that deals with identifying unusual or abnormal patterns in data, which can be useful for detecting fraud, errors, or other outliers.

Sub-disciplines of Data Science

Data visualization:

a subfield of data science that deals with the design and creation of visual representations of data, such as charts, graphs, and maps, to communicate information and insights effectively.

The process - the "pipeline"

1. Collecting data
2. Cleaning and preparing the data
3. Analyzing the data - Exploratory Data Analysis - Visualization
4. Using the insights gained to make decisions or predictions - building a model
5. Deployment - "shipping" - making an app/program/
6. Monitoring for results / changes / improvements -> Back to Squares (1,2,3,4,5)

Difference between Data Scientist and Data Engineer

Data scientists and data engineers are both important roles in the field of data science, but they have different responsibilities and focus on different aspects of the data pipeline.

Data Scientist

Data scientists are responsible for analyzing and interpreting data, and using that information to make predictions or decisions. They use statistical and machine learning techniques to build predictive models, and they often work closely with stakeholders to understand their needs and communicate the results of their analysis.

Data Engineer

Data engineers, on the other hand, are responsible for the technical infrastructure and tools that support data science work. They design, build and maintain the systems that collect, store, and process data, and they ensure that data is accurate, reliable, and easily accessible to data scientists and other stakeholders. They work with technologies like Hadoop, Spark, and SQL databases to extract, transform and load the data.
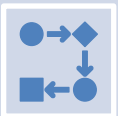
Both roles often overlap.

POP QUIZ

What is the main goal of data science?

a) To extract insights and knowledge from data b) To collect and organize data

c) To make predictions and decisions

d) To build predictive models

POP QUIZ

What is one example of a task that falls under the area of Natural Language Processing (NLP)?

a) Image recognition

b) Language translation

c) Anomaly detection

d) Recommender systems

POP QUIZ

What is one example of a task that falls under the area of Computer Vision?

a) Sentiment analysis

b) Object recognition

c) Time series forecasting

d) Recommender systems

ANSWERS

Answers: 1-c 2-d 3-a 4-b 5-b

PS I would debate about 1-c

Tools of the trade

Programming Languages

There are several programming languages that are commonly used in data science.

It's worth noting that a data scientist may not necessarily need to know all of these languages but having a strong knowledge of one or two of these is sufficient, as they are often used in combination with other technologies and tools to perform data science tasks.

Python

Python is a widely-used, high-level programming language that is popular among data scientists because of its simplicity, readability, and wide range of powerful libraries and frameworks for data analysis, modeling, and visualization.

Some of the most commonly used libraries for data science in Python include NumPy, pandas, scikit-learn, and TensorFlow.

R:

R is a programming language and environment for statistical computing and graphics. It is widely used among statisticians, data scientists and academics, and it has a large community that develops and maintains a wide range of packages for data analysis, visualization and modeling.

SQL:

SQL (Structured Query Language) is a domain-specific language used for managing relational databases. It is widely used to access, retrieve and manipulate data stored in databases, which is often a key component of data science projects.

Java and Scala:

Java and Scala are general-purpose programming languages that are widely used in industry. They are popular for big data processing using technologies like Apache Hadoop and Apache Spark.

Julia:

Julia is a high-performance, high-level programming language for technical computing, with syntax that is familiar to users of other technical computing environments. It is designed to be fast and easy to use, and it is gaining popularity among data scientists and researchers.

Text Editors

- Visual Studio Code: Visual Studio Code is a popular text editor that is available for Windows, Mac, and Linux. It is developed by Microsoft and it offers advanced features such as debugging, code completion, and built-in Git support.
- PyCharm: PyCharm is a powerful, cross-platform text editor that is specifically designed for Python development. It offers advanced features such as code completion, debugging, and integration with version control systems.
- RStudio: RStudio is a powerful, cross-platform text editor that is specifically designed for R development. It offers advanced features such as code completion, debugging, and integration with version control systems.
- Many others, try out different ones

Other Software

Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that contain live code, equations, visualizations, and narrative text. It is often used in data science to:

Clean and explore data: Jupyter Notebooks allow data scientists to load, manipulate and visualize data in an interactive way, making it easy to understand the data and identify any issues or outliers.

Develop and test models: Jupyter Notebooks allow data scientists to develop, test and iterate on models in an interactive environment. They can quickly test different algorithms and parameters, and visualize the results.

Communicate and share results: Jupyter Notebooks are a great tool for data scientists to communicate their findings and results to others. They can include text, images, and code in one document, which makes it easy to understand the results and the steps that were taken to produce them.

Collaboration: Jupyter Notebooks are also useful for collaboration and reproducibility, as they allow multiple data scientists to work on the same notebook and keep track of the changes made.

Jupyter Notebooks can be used with different languages like Python, R, Julia, Scala, and many more.

Git and Github

Git is a popular version control system that is widely used in data science for several reasons:

Collaboration: Git allows multiple data scientists to work on the same project and keep track of the changes made by each member.

Reproducibility: By using Git, data scientists can keep track of the changes made to the code, data, and documentation. This allows them to reproduce the results of an analysis or experiment at a later date, and also allows others to reproduce the results.

Backup and archiving: Git allows to store the code, data and documentation in a remote repository, like **GitHub**, GitLab or Bitbucket.

Experimentation: Git allows data scientists to easily test out different ideas and approaches without affecting the main codebase.

Versioning: Git also allows data scientists to save different versions of their code, data and documentation, which is useful when experimenting with different approaches.

Tools of the Future and Present

- Github Copilot – based on GPT
- Stable Diffussion
- OpenAI GPT models such as ChatGPT

- All of these tools helped in making this presentation

Want to get started but can't wait until September?

There are many free or inexpensive online courses that can give you a headstart.

[Mini Data Science course by Google](#)

Books

- [Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 3rd Edition](#)

Questions?

Code Repository for this seminar:
https://github.com/ValRCS/rbs_greenhouse

THANK YOU!