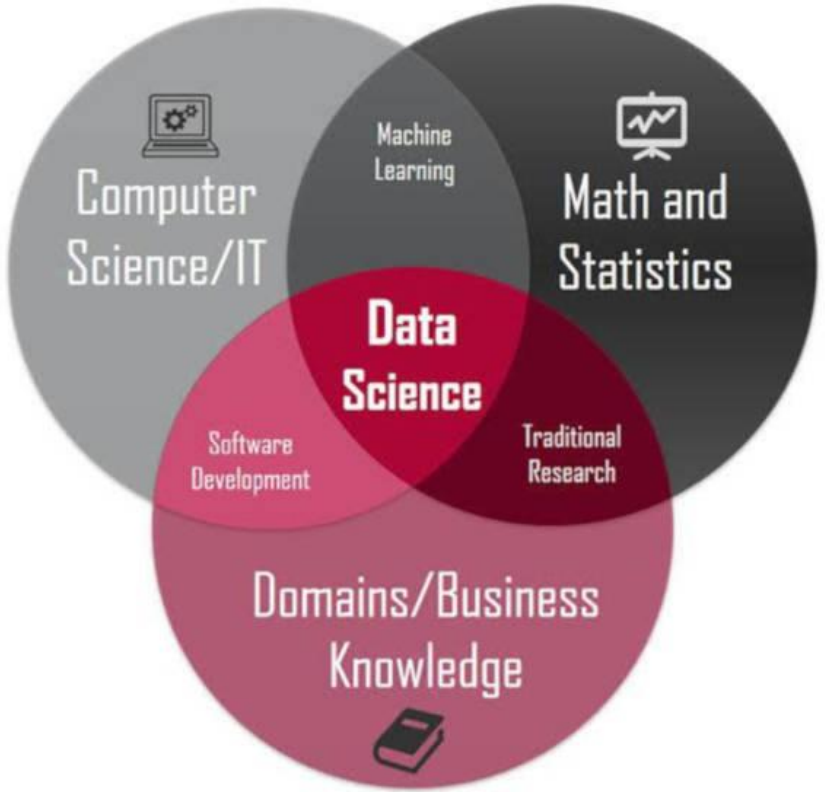
INTRO DATA SCIENCE

LECTURE 1



Course goals

- Je begrijpt wat het domein van Data Science inhoudt.
- Je begrijpt de rol van de Data Scientist, hun proces en hun taken.
- Je kunt de dataset analyseren, ontbrekende waarden identificeren en geschikte voorbewerkingstechnieken toepassen om de data klaar te maken voor modellering.
- Je kunt de zakelijke doelstellingen van een machine learning-component formuleren en de relevante gestructureerde dataset correct laden en verkennen.
- Je kunt een geschikt machine learning-model selecteren, trainen en valideren, en de prestaties evalueren met relevante meetmethoden.

Grading: Portfolio 1/2

- You will build up your **Data Science portfolio** with assignments which you will work on during and after class.

- Assignments will be a combination of
  - Individual and group-assignments
  - Fixed format and free choice
  - Code and documentation

- You are free to add any Data Science-related work you performed to your portfolio, even if it was not part of any assigment. This purely optional but will have a positive contribution to your final grade. Not sure if something is Data Science-related? Ask the teachers.

- Portfolio assignments will be identified by this icon: 

Grading: Portfolio 2/2

- Assessment
  - You explain your portfolio will during the assessment
  - Think about questions as Why are your doing it? What are your findings? Is it what you expected etc.

Portfolio assignment 1

- What is Data Science?
  - Split yourselves in groups of 3-4 students.
  - 10 min: Do online research, individually, to answer the question “What is Data Science?”*
  - 10 min: Create a PowerPoint slide with a summary of your results*

- Now or outside training: Same assignment but the question is “What does a Data Scientist do?”*

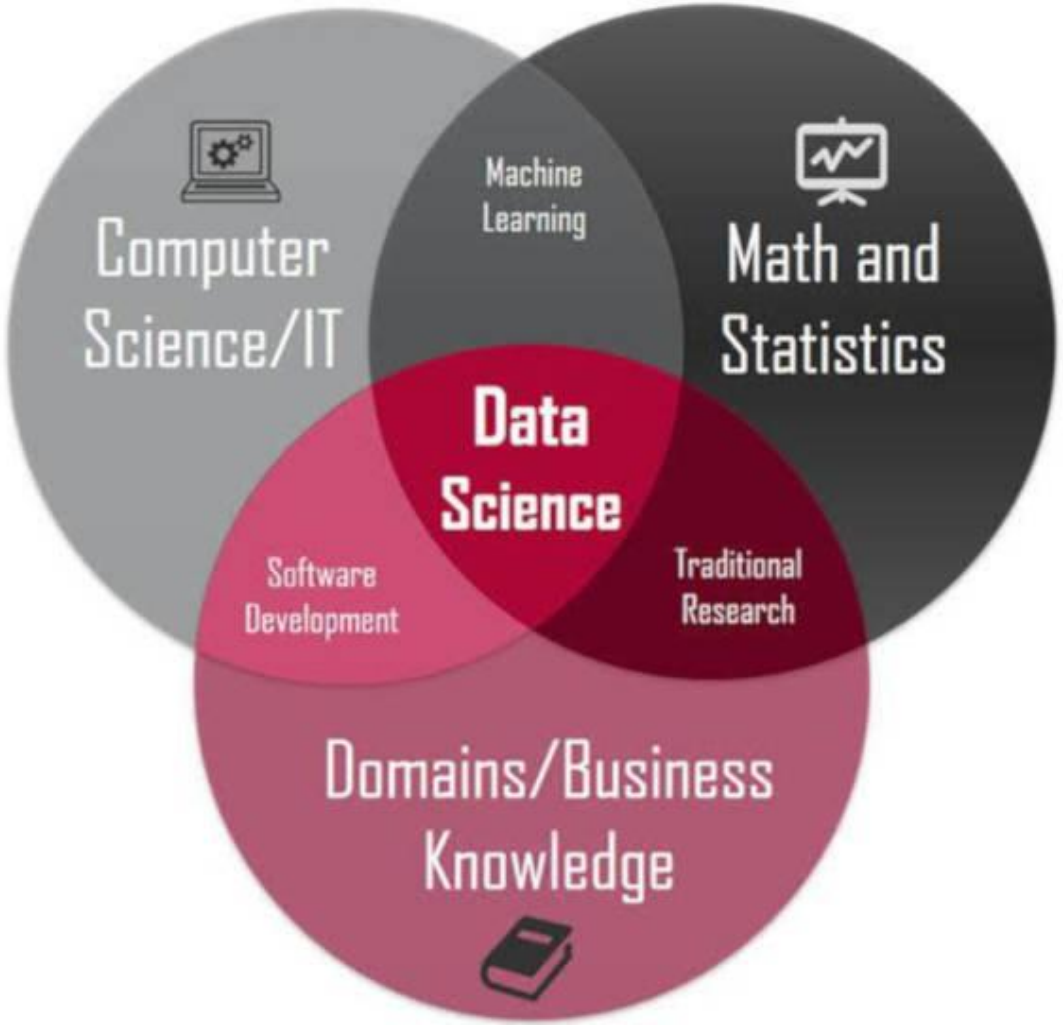
*Do not use Avans school materials as a source for this assignment

What is Data Science?

- "Data science is an **inter-disciplinary field** that uses scientific methods, processes, algorithms and systems to **extract knowledge and insights from** many structural and unstructured **data**… Data science is a 'concept to unify statistics, data analysis and their related methods' in order to 'understand and analyze actual phenomena' with data. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, domain knowledge and information science."
  - Wikipedia

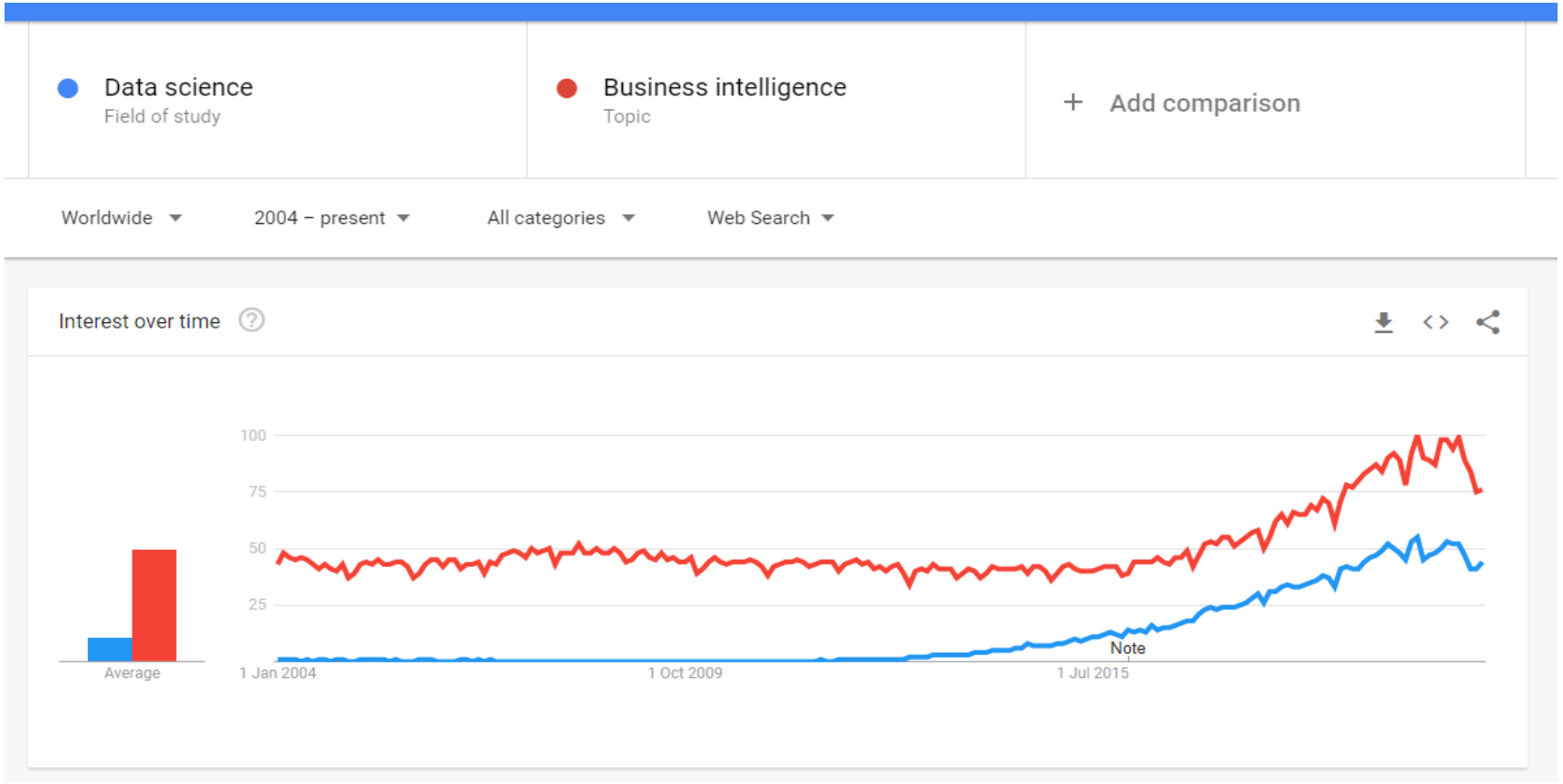
- "Effective data scientists are able to identify relevant questions, collect data from a multitude of different data sources, organize the information, translate results into solutions, and communicate their findings in a way that positively affects business decisions."
  - Berkeley

What is Data science?



Source: https://www.kdnuggets.com/2020/08/top-10-lists-data-science.html

Data Science vs BI



Source: https://trends.google.com/trends/explore?date=all&q=%2Fm%2F0jt3_q3,%2Fm%2F016jq3

Data Science vs BI

Data Science	Business Intelligence
• The data scientist is skilled in technology and content. • The data scientist digs for the right data himself to provide himself with the right information. • Uses the OLTP and OLAP environments as possible sources.	• The ICT owns the technology but not the content. • The ICT person ensures that everyone in an organisation is provided with the right information at the right time. • Setting up and maintaining data warehouse (OLAP) and reports.

This course

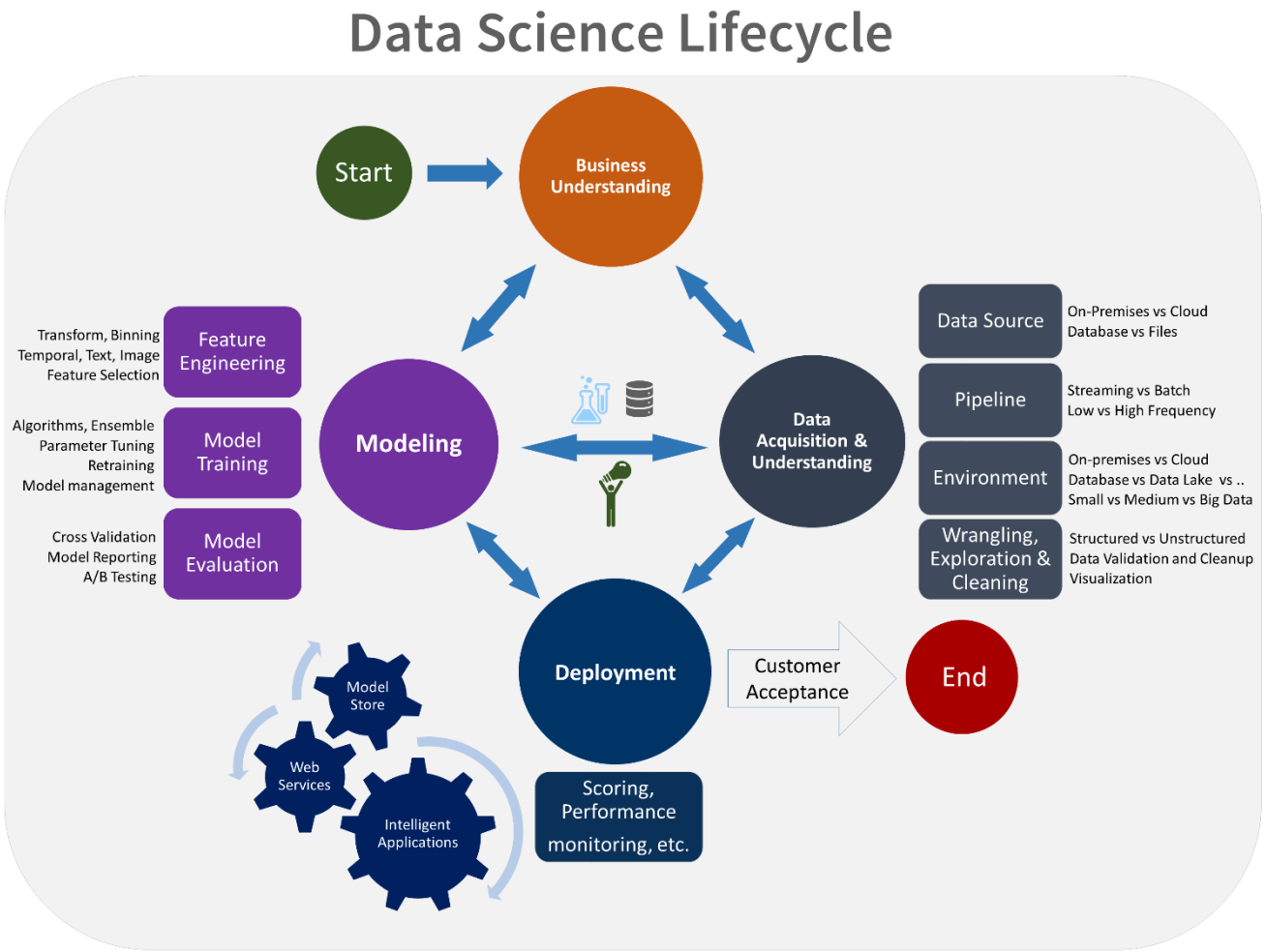
- Walk a mile in the shoes of a data scientist.
  - To improve your abilities to cooperate with data scientists
  - The first steps in your journey as data scientist

What does a Data Scientist do?

- Turns data into actionable insights
  - Explore customer data to extract useful information
  - Identify customers segments based on customer data
  - …
- Create data products:
  - Build a product recommender based on customer & product data
  - Predict customer churn* based on customer data
  - Predict product sales based on sales data
  - …

*churn: the loss of clients or customers.

How does a Data Scientist do this?



Source: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Portfolio assignment 2

- What are the most popular data science tools?
  - You will be split up into breakrooms with 3-4 students each.
  - 10 min: Do online research to answer the question “What are the most popular data science tools?”*
  - 10 min: Create a PowerPoint slide with a summary of your results

- Optional: Same assignment but the question is “What are the characteristics that make these tools popular for data science tasks?”*

*Do not use Avans school materials as a source for this assignment

Data science tools

- Programming languages
  - Python with Jupyter Notebooks
  - Julia
  - R with RStudio
  - …
- Software
  - Anaconda, PyCharm, Cursor, Visual Composer
  - SAS
  - SPSS
  - RapidMiner
  - …
- Important differences
  - Programming Languages vs Software
  - Open-Source/Free vs Proprietary

Data Science tools

- What do we need?
  - Iterative nature of data analysis -> Interactive environment-> REPL/Notebook
  - Custom logic for data manipulation -> Programming language
  - Don’t reinvent the wheel -> Toolbox -> Libraries/packages -> Cheatsheets
  - Visualising the data -> Plotting tools -> GUI

Installing the tools

- https://www.anaconda.com/products/individual#Downloads
- Demo in class