Rail Data Science

Introduction to Data Science

Prof. Dr. Raphael Pfaff

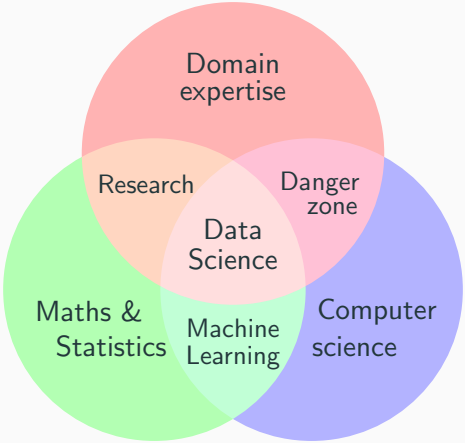
23. Mai 2023

Fachhochschule Aachen

What is Data Science?

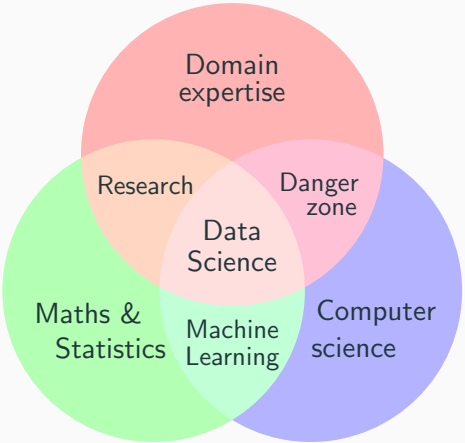
- Interdisciplinary field of
  - Systems,
  - Methods and
  - Processs to extract insight or knowledge from data.
- Term coined in 2001, gained popularity in 2010
- Integrates:
  - Data Engineering
  - Scientific Method
  - Mathematics
  - Statistics
  - Advanced Computational Methods
  - Visualisation
  - Hacker Mindset
  - Domain Expertise

How does Data Science integrate to Mechanical Engineering?



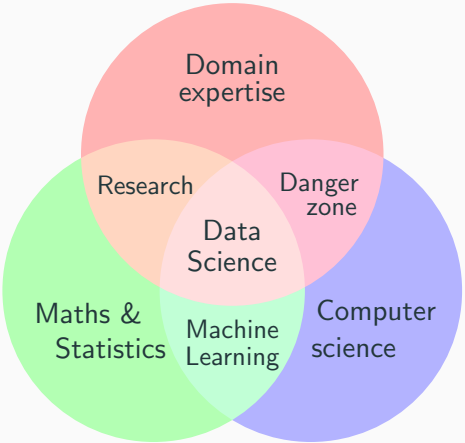
Why Data Sciene?

- Turn data to information
  - Inform decisions
  - Increase insight
- Companies:
  - Collect large amounts of data
  - Do rarely integrate them
  - Frequently decide based on the “gut”

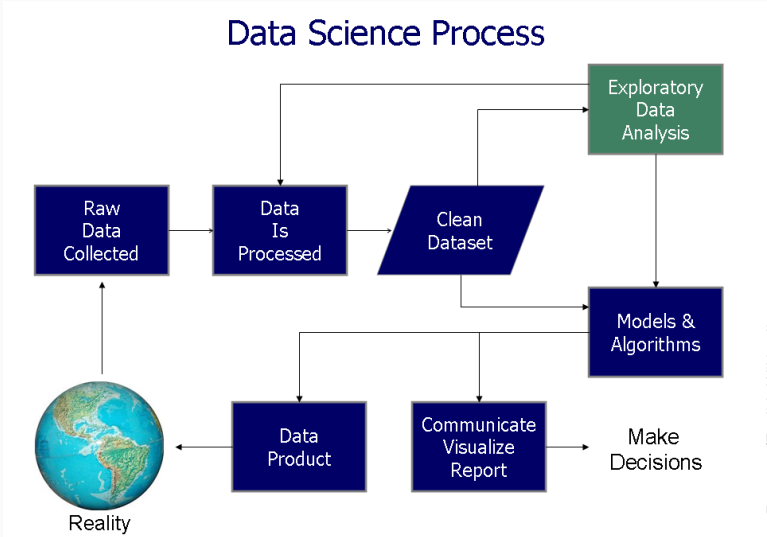


How do you increase Value with Data Science?

- Improve decision making
  - Empower management
  - Supply data driven evidence
- Identify trends and bring to action
- Challenge your colleagues
- Find opportunities for improvement
- Test decisions
- Understand customers



The data science process



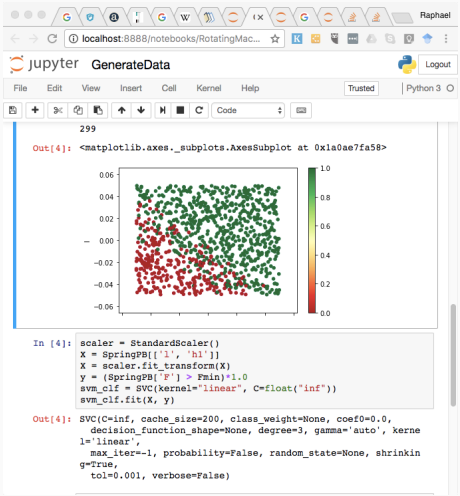
Source: Farcaster/English Wikipedia

How to get started

- Set up your system.
  - Install Anaconda to obtain Python/Jupyter
- Acquire data: start with popular open data sets. Get your company to make data accessible.
- Ingest and transform: figure out the formats and sizes of your data. Find appropriate ways to import or access them.
- Explore the data. Do you already find patterns from just plotting them?
- Try your "toolbox" of methods (or a subset of it that sounds promising).
- Visualise the results. Make your findings convincing to others: colleagues, managers, customers etc.

Tools for Data Scientists

- Programming languages:
  - R
  - Python
  - SAS
  - ...
- Visualisation app:
  - Tableau
- Development Environment (IDE):
  - Jupyter
  - Spyder
- Also potentially:
  - Matlab
  - Scilab



Selected Techniques applied in Data Science

- **Visualisation**
- Regression: Linear, Logistic
- **Density Estimation**
- Confidence Intervals
- Test of Hypotheses
- Pattern Recognition
- Time Series
- **Unsupervised Learning (Clustering)**
- Supervised Learning
- Decision Trees
- **Monte-Carlo-Simulation**
- Bayesian Statistics
- Principal Component Analysis
- **Support Vector Machines**