Rail Data Science

Handout - Introduction to Data Science

Prof. Dr. Raphael Pfaff
Lehrgebiet Schienenfahrzeugtechnik

May 23, 2023

Data science as a field has far more applications than quality management, however quality management professionals take huge advantages from the application of data science tools and methods in their daily and non-daily work.

For this reason, enjoy this brief introduction and the more thorough hands-on training provided in the sequel!

What is Data Science?

Data science is our means of taming unstructured information and gathering insight. - Matthew Mayo, KDnuggets

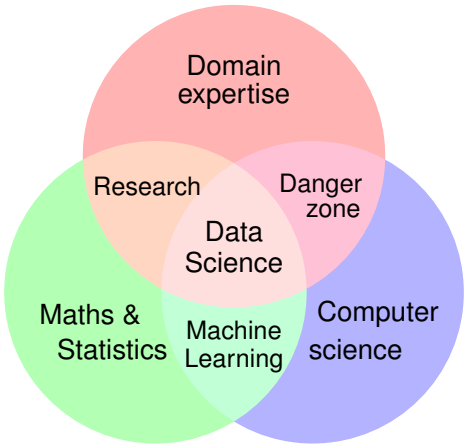
- Interdisciplinary field of
  - Systems,
  - Methods and
  - Processs to extract insight or knowledge from data.
- Term coined in 2001, gained popularity in 2010
- Integrates:
  - Data Engineering
  - Scientific Method
  - Mathematics
  - Statistics
  - Advanced Computational Methods
  - Visualisation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

How does Data Science integrate to Mechanical Engineering?

You may ask yourself: what is the point in learning to code in Python and apply data science tools, I am a mechanical engineer?

The short answer is: you will probably need it.

The longer answer is: your domain expertise, coupled with the willingness to dive into data, makes the difference!



Why Data Sciene?

In a company context, you frequently act without proper information, although it is there - just nobody takes the time to analyse it. As soon as you show up to a board meeting or similar with data, people tend to trust your claims and you may be the one who is heard.

As Deming (the guy with the PDCA circle...) said:

> **In God we trust, all others bring data.**

Data science puts you in a position to:

- 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?

You are about to become engineers bearing the second highest education level in the EU (and world wide, for most countries) - perhaps you should bear in mind that the additional income and employer attractiveness come with some expectation: you are supposed to create value commensurate with your salary expectation and also your education level. Data Science may be helpful in this context.

- 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

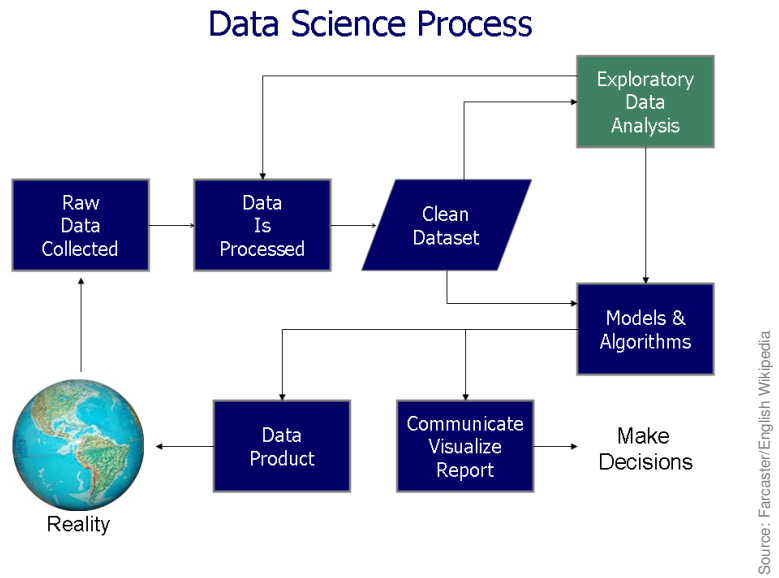
The data science process starts with *real world data*, which is messy, has missing values and all other ugly properties. So most of the time, you need to start with data cleaning and some exploration of the data set, mostly in the form of graphical analysis and looking at aggregated data.

As soon as you know something about your data, you are in a position to develop models and algorithms on your data set.

The result of both may be

1. a data product, e.g. a web app for data analysis or, more frequently in day-to-day work,
2. some form of communication, e.g. slides, a report, etc.,

as displayed below:



How to get started

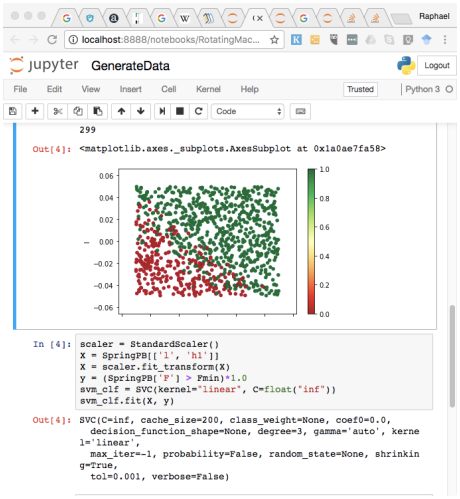
In this module, we want you to get your hands *dirty* on data, linked to quality-related problems. For this, you need some tools available locally. So, perhaps start early - mostly, installing anaconda is not a problem, sometimes however, it is made difficult by some system setups.

The actual steps will be subject of the more in-depth units to follow soon.

- 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
  - ...



For installation, visit: `https://www.anaconda.com/products/individual`https://www.anaconda.com/products/individual

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**