EPSILON – European Platform for Data Science: Incubation, Learning, Operations and Network

Training Material for Teaching and Self-Learning

Introduction to Data Science

Module 1/6

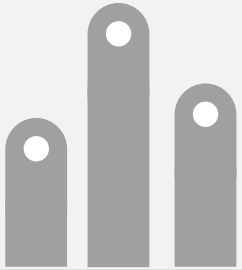
Hochschule Harz
Harz University of Applied Sciences
EPSILON
Funded by the European Union

Agenda

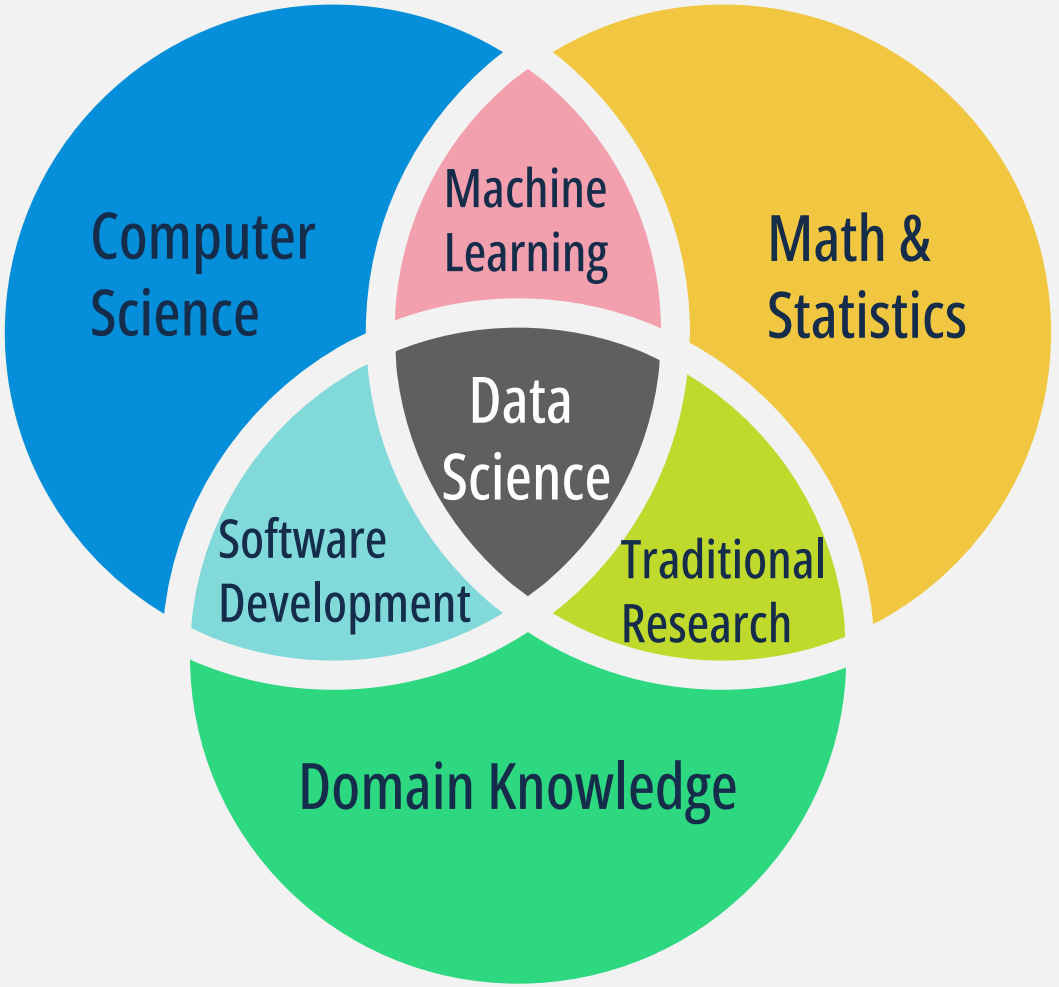
- Elements of Data Science
- Big Data
- Data Processing
- Ethical Issues

Elements of Data Science

An intersection of multiple disciplines



Elements of Data Science



Elements of Data Science

Computer Science

The study of computers and their hardware, software and processing capabilities

Domain Knowledge

Seen in the context of Social Good

The knowledge about the projects **themes**, targets, **goals** and **background knowledge**, can also include knowledge about parties involved and/or effected.

Math & Statistics

Algorithms and mathematical formulars required to analyse data.

Belford & Tucker 2023/Egger 2020, p. 32f

Elements of Data Science

Machine Learning

A subfield of Artificial Intelligence that utilizes algorithms trained on data sets to create models for analysis.

Software Development

Drafting, coding, providing and/or supporting of computer programms.

Traditional Research

Studying a subject in detail in order to discover new information or reach a new understanding.

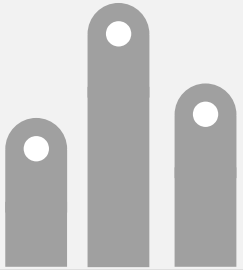
Data Science

Scientific methods, algorithms and systems to extract knowledge or insights from big data.

„The term *Data Science* was created in the early 1960s to describe a new profession that would support the understanding and interpretation of large amounts of data. […] In the past 30 years, Data Science has quietly grown to include businesses and organizations worldwide. […] During its evolution, Data Science use of **big data** was not simply a „scaling up“ of the data, but included shifting to new systems for processing data and the ways data gets studied and analysed.“

Big Data

What makes Big Data “Big”? – A set of parameters



Big Data

- Volume
- Velocity
- Variety

> Traditional Big Data parameters

- Veracity
- Variability/Volatility

> Common, additional parameters

Egger 2020, p. 62f

Traditional parameters

Big Data

- **Volume**
- Velocity
- Variety

- Veracity
- Variability/Volatility

The quantitative amount of data that can be collected through a variety of sources.

Including, but not limited to: transactions, industrial equipment, mobile phones and sensory data.

Egger 2020, p. 62

Traditional parameters

Big Data

- Volume
- **Velocity**
- Variety

- Veracity
- Variability/Volatility

The speed, with which data is produced. Some forms of data collection might require near real-time processing in order to be properly captured.

Egger 2020, p. 62f

Traditional parameters

Big Data

- Volume
- Velocity
- **Variety**

- Veracity
- Variability/Volatility

The amount of different data formats such as text, images, videos and sensory data.

Egger 2020, p. 62

Additional parameters

Big Data

- Volume
- Velocity
- Variety

- **Veracity**
- Variability/Volatility

Data needs to be quality checked. The quality of an analysis is based on the data used. If incorrect data is fed into the system, the end result might be skewed if not outright wrong.

Egger 2020, p. 63

Additional parameters

Big Data

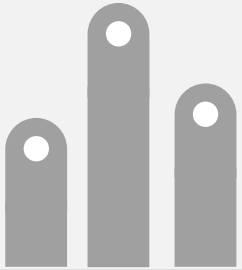
- Volume
- Velocity
- Variety

- Veracity
- **Variability/Volatility**

This refers to potential fluctuation of an incoming data stream that, for example could be influenced by weather or supply and demand. Best case would be a stable flow of data.

Data Processing

...because the quality of the results depends on the quality of the initial data.



Data Processing

Problem Framing

- What is the problem?
- What are related questions regarding the problem?
- What is the goal?

Data Processing

Data Gathering

- What kind/type of data can I collect?
- How do I collect the data?
- What could be a suitable data structure?

Egger 2020, p. 64f

Data Cleaning and Preparation

- How are data errors, missing data or extreme outliers handled?
- Sorting data into useful or unnecessary data with regard to the problem.
- Ensure a clear structure of the data base.

Data Processing

Data Analysis

- Data exploration and analysis.
- What method is suitable for the data? – e.g. visual data techniques, statistical models, machine learning algorithms…
- Extraction of meaningful data.

Egger 2020, p. 65f

Data Processing

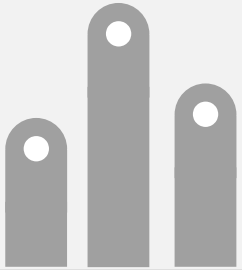
Data Exploitation

- What is the potential impact of the output?
- How to utilize the gathered output?
- What new strategies can be derived from the output?

Egger 2020, p. 66f

Ethical Issues

Because Data Science models are ubiquitous and can impact everyone, they must be examined critically to avoid biased or potentially harmful results.



Definition of Ethics

What is **right** and what is **wrong**?

"The term "*ethics*" comes from the Greek work *Ethos*, which means "*habit*" or "*custom.*" Ethics instructs us on what is right and wrong. [...] Most people associate ethics with morality. A natural sense of what is "*good*". We as humans live in a society, and society has rules and regulations. [...] Ethics deals with feelings, laws and social norms which determine right from wrong." (Majumder 2023)

Importance of Ethics in Data Science

- Data Science in the form of algorithms is omnipresent in the digital world.
- The three largest ethical issues are:

Decisions made based on data	Privacy and Confidentiality of Data	Ownership of Data

The 3 Ethical Questions

Decisions made

Who is **directly** impacted and how?

Who is **indirectly** impacted and how?

→ Talk to, and/or consider everyone who is potentially affected.

Privacy and Confidentiality

How is the privacy of an individuals' data as well as its confidentiality warranted?

What are possible leaks and issues for such?

Ownership

Who owns the rights to the data?

What data is freely available and what data requires either a licence or permission?

Who is accountable?

Sources I

- Egger, R. (2020) *Applied Data Science in Tourism*, Springer, ISBN 978-3-030-88389-8
- Coursera, 1 (last opened 23.11.2023), *What is Machine Learning?*, https://www.coursera.org/articles/what-is-machine-learning
- IBM, 1 (last opened 23.11.2023), *Was ist Softwareentwicklung?*, https://www.ibm.com/de-de/topics/software-development
- Cambridge Dictionary (last opened 23.11.2023), *Research* https://dictionary.cambridge.org/dictionary/english/research
- Foote, K. (16.10.2021) *A Brief History of Data Science*, https://www.dataversity.net/brief-history-data-science/
- Belford, G. & Tucker, A. (last updated 25.11.2023) *Computer Science*, https://www.britannica.com/science/computer-science
- Barak, M. (2020) *The Practice and Science of Social Good: Emerging Paths to Positive Social Impact*, an article in „Research on Social Work Practice Vol 30(2)“ p.139-150 DOI: 10.1177/1049731517745600

Sources II

- Kenton, W. (last updated 26.08.2022) *Social Good: Definition, Benefits, Examples*, https://www.investopedia.com/terms/s/social_good.asp
- Majumder, P. (last updated 09.08.2023) *Ethics in Data Science and Proper Privacy and Usage of Data*, https://www.analyticsvidhya.com/blog/2022/02/ethics-in-data-science-and-proper-privacy-and-usage-of-data/
- Bezuidenhout, L. & Ratti. E. (2021) *What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues*, AI & Societies 36, p.939-953, https://doi.org/10.1007/s00146-020-01112-w
- Scribblr (last opened 30.11.2023) *Types of Bias in Research | Defintion & Examples*, https://www.scribbr.com/category/research-bias/
- Aula, V. & Bowles, J. (09.05.2023) *Stepping back from Data and AI for Good – current trends and ways forward*, Big Data & Society, p.2f, https://doi.org/10.1177/20539517231173901

Open Educational Resources

ATTRIBUTION 4.0 INTERNATIONAL - Deed

- **You are free to:**
- Share - copy and redistribute the material in any medium or format.
- Adapt - remix, transform, and build upon the material for any purpose, even commercially.
- **Under the following terms:**
- Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. If you wish to use this work in a way not covered by the license, please contact:

Harz University of Applied Science
Friedrichstraße 57 – 59
38855 Wernigerode
E-mail: info@hs-harz.de