The importance of data and data science

Data - the oil of the 21st century

Data plays a very central and crucial role for many AI solutions and applications.

With the right data, the machine learns to do right and meaningful things.

With the wrong data, the machine will do the wrong things.

Training data for machine learning algorithms

Training data must be correct and consistent, relevant and meaningful. It should neither intentionally nor unintentionally be manipulated, and it should not contain any bias.

The accuracy, precision and usefulness of the results of machine learning algorithms directly depends on the relevance and quality of the data with which they are trained.

Data must be available in critical minimum quantity only then can machine and deep learning algorithms unfold their full effectiveness and power.

Structured vs. unstructured data

Structured data

- This is data that is stored in a structured form in IT systems and underlying databases of the company.
- Data can be displayed in structured forms, such as rows and columns of a table. For example an excel sheet with numbers.

Unstructured data

- This data is not deliberately planned and systematized.
- It may be data that cannot be even displayed in structured forms. These are, for example, data from emails, videos, pictures, voice recordings and the like.

Internal vs. external data

1 Internal data

- Data stored in IT systems and underlying databases **within the company**, meaning **within its own firewalls**.
- For example, data of customers, products, processes and transactions.

2 External data

- Companies have also access to vast amounts of additional **external data outside of their own firewalls**.
- For example, social media related data.

Correlation vs. causal relation

Simple correlation

A pure statistical correlation, that means association, between two variables without a real explainable causal relation

Real causal relation

A relationship between two variables (cause vs. effect) where the cause is partly responsible for the effect, and the effect is partly dependent on the cause

Biased data and false conclusions

Bias is disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief.

In science and engineering, a bias is a systematic error. **Statistical bias** results from an unfair sampling of a population, or from an estimation process that does not give accurate results on average.

If you **intentionally or unintentionally** feed and train an algorithm with biased data, AI will recognize wrong statistical patterns and then may come up with **false conclusions**.

Data Science

The art of analysing data, structuring it, cleaning and processing it, but also evaluating and interpreting it