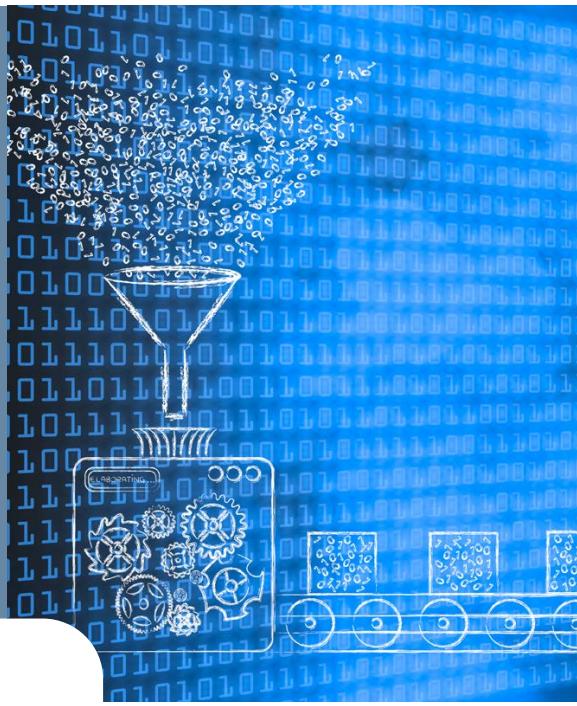
Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

From data to AI

White paper by
Daniel Keim, Kai-Uwe Sattler
Working Group Technological Enablers and Data Science



Executive Summary

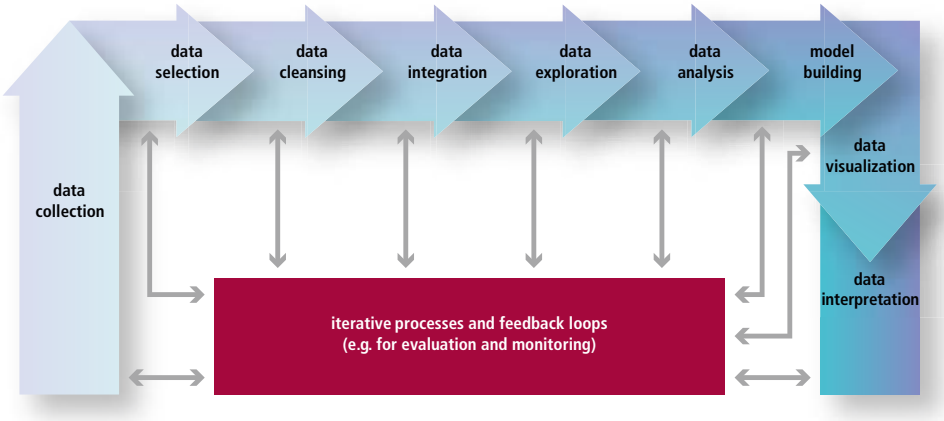
Whether satellite images as data sources for navigation systems or vacation photos on social media platforms – unimaginably large amounts of new data are generated every day. Data has therefore become a central resource in our increasingly digitalized world. Comprehensive data management and the ability to make data accessible for analysis in the first place are thus an important prerequisite for gaining valuable scientific insights and generating beneficial applications for business and society. The interdisciplinary branch of research data science – the management and analysis of data – is therefore already considered one of the most important key disciplines for science and business. The availability of data and data management is a central pre-condition especially when applying self-learning systems.

Data science focuses on how data is processed, prepared and analysed. Science-based methods, processes, algorithms and systems can extract patterns and knowledge from structured and unstructured data. Data science methods are thus considered to be pioneers for scientific findings in a wide range of research fields, such as climate research, astronomy, materials science, chemistry and medicine. They also enable applications that improve our daily lives, such as the use of navigation systems or language assistants. These methods are also considered a requirement for realizing the potential of AI applications, for example in product manufacturing, logistics or customer management.

The basis of many data science applications are process chains, which cover the steps of data acquisition, selection, cleansing, integration, exploration, analysis and modelling up to data visualization and interpretation. These processes are either specified explicitly (procedural or declarative) and then executed automatically or rather implicitly in an interactive and explorative way. Frequently,

these processes are not static procedures, such as extract, transform, and load (ETL) processes in data warehousing, but interactive processes that require human intervention and decisions („human-in-the-loop"). In some cases, however, these also consist of iterative processes with feedback loops in order to change data, methods or parameters if necessary, to obtain new training data or to update models with new data. Such interventions need continuous monitoring and evaluation of each individual steps' results, e.g. the quality of the input data or the quality of the model.

data science processes



During such processes a variety of methods from different areas are used. Examples are:

- signal processing methods, e.g. for filtering data
- data integration methods
- statistical methods for characterizing data, deriving key figures or features, time series analysis etc.
- data visualization methods for visual data exploration
- data mining and machine learning methods for data cleansing (e.g. replacement of missing values, duplicate detection)
- methods for the actual modelling

The selection and acquisition of data – for example as training data – plays an important role in these processes just as ensuring data quality. In addition to data profiling and data cleaning, integrating and enriching data is important in order to access and merge the often heterogenous data. In this context, semantic technologies such as knowledge graphs are also used, the facts of which are used, among other tasks, to validate, evaluate and annotate data. Data collections and knowledge bases such as Wikidata, Freebase, DBpedia or YAGO provide extensive facts, e.g. for data preparation.

The field of data engineering covers specifically methods and processes for the acquisition, management, storage, preparation, enrichment and provision of data. It focuses on questions of high-performance and reliable infrastructures for data management. These infrastructures are fundamental for the

efficient support of data science processes and methods for the management and preparation of data and models. Data engineering is therefore used in the following as a generic term for data management, data integration and data preparation. In addition, visual exploration and visual analysis of data and models is essential (Visual Analytics) to assess the data quality and models, to interact effectively with data and models or to derive new training data.

In order to understand data science applications, assess their quality and thus confirm their trustworthiness, we need a deep understanding of each individual component of a data science process and an equally deep understanding of the interaction of these components. It is therefore important to understand the process chain as a whole.

Among the central methods and techniques that are fundamental for data science – and thus also for Artificial Intelligence – are:

- **data management:** This includes storage technologies, data structures, techniques for the physical design of databases such as caching, replication, indexing, data compression, partitioning for distributed and parallel data storage as well as precomputation and materialization of frequently used data. This also applies to the management of training data and models.

- **data preparation:** Techniques for data profiling and data cleaning using statistical procedures and methods of machine learning, techniques of data integration via the combination and linking of data up to heterogeneous systems.

- **performance and scaling:** Data management systems are optimized for high-performance and scalable processing of large amounts of data. On the one hand this is achieved by distribution and parallelization of the processing. In addition to distributed or parallel database systems, this also includes platforms for distributed processing in cluster environments such as Apache Spark or the successful German development Flink as well as data infrastructures for cloud computing. There is also a close connection to distributed machine learning. On the other hand, this includes the efficient use of modern hardware (e.g. in the form of main memory databases, the use of GPU- or FPGA-based accelerators, as well as multi-core and cluster systems). It also includes the online processing of data streams, such as those found in the Internet of Things, Industry 4.0-scenarios or social media platforms. Increasingly, special hardware architectures such as tensor computing units and neuromorphic systems play an important role, too.

- **optimization and execution of complex processes:** The development and execution of complex data processing pipelines is supported by declarative query processing including analytics methods (e.g. online analytical processing, OLAP) and data flow models (as in Spark or Flink). This also applies to the support of continuous machine learning. While using AI applications, data is generated which can be used as training data for continuous learning systems.

- **ensuring repeatability and traceability:** Especially for data-driven learning processes, it is important to make the modelling process repeatable and traceable, for example to be able to trace incorrect predictions, changes in training data or in processing. For this purpose, data management techniques such as data versioning, time-travel requests or auditing offer solutions.

There are various approaches to use data more efficiently and effectively to benefit society in the future and to promote the understanding of data science processes and data management technologies in our society – a society in which collecting, processing and analysing data is a basis for prosperity, simplification of everyday life and scientific progress. This includes among other things:

- **data literacy:** As already demanded by the German Informatics Society (GI), schools and universities must emphasize data literacy more broadly. This goes far beyond teaching computer science in schools or university programmes related to computer science. It includes providing the opportunities to acquire the skills needed to access, collect and assess the quality of data and to become familiar with tools for data processing and analysis and for the visualization and critical interpretation of results.

- **data science/data engineering training:** More emphasis should be placed on data engineering topics in data science programmes and in computer science courses. This goes beyond the content of classical database lectures, and concerns, for example, data-literacy skills, aspects of data integration and data quality, data exploration and data visualization, alternative data models and paradigms of data processing. The German Informatics Society's working group „Data Science/Data Literacy", in collaboration with Plattform Lernende Systeme, has developed recommendations for university programs and advanced training in the field of data science (Gesellschaft für Informatik, 2019).

- **infrastructure and data:** Successful models should not be adopted without reflection. Even if they are as impressive as the examples and the application of AI labs at Internet companies. In addition to understanding the learning methods and their limitations and knowledge of the training data used, this also requires the possibility to conduct comparable elaborate learning procedures. This demands suitable infrastructures with sufficient storage and computing capacity and data collections (e.g. for training data).

  At the same time, „small-data"-methods must also be given consideration. Areas of application that are particularly important in Germany, such as medical technology or mechanical engineering, are characterized by significantly smaller data volumes. This is particularly true in the area of SMEs, where data are often poorly integrated and small in scale. Infrastructures, research and training programs for the AI sector should take this situation into account.

  Overall, data management technologies play a central role for projects such as European data spaces, which the European Commission, for example, wants to push forward in its data strategy (European Commission, 2020), and for cloud-based data infrastructures such as GAIA-X or data ecosystems in general. Since computing infrastructures are themselves the subject of research, such infrastructures must be strengthened and expanded.

This is necessary because it enables researchers and developers to design underlying data management processes and conduct experiments.

- **research:** In future research programs, such as the AI strategy of the Federal Government or in research projects of companies and research and development institutions, the importance of data engineering as an overall process in connection with machine learning procedures should be taken into account even more. In this way, strengths can be further developed both in the acquisition, selection, exploration and visualization of data and in the use of AI and data management methods in data science processes. Considering the essential role of data management technologies for the data spaces and data ecosystems of the future, an even stronger promotion of relevant fields of research is purposeful.

- **application:** Data engineering plays an important role in putting self-learning systems into practice. The overall process of data engineering and self-learning systems should therefore also be given more consideration in the development and implementation of AI applications in companies by paying special attention to the collection, pre-processing and secure storage of both training data and process data. High-quality data are an essential basis for the successful application of self-learning systems. This is especially true in technical and industrial fields of application such as automation and predictive analytics, but it holds true in the medical field, too. At the same time, such data is highly sensitive. Infrastructures and data ecosystems must therefore take application- and industry-specific solutions into account in an appropriate way.

Imprint
Editor: Lernende Systeme – Germany's Platform for Artificial Intelligence | Managing Office | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Follow us on Twitter: @LernendeSysteme | Status: October 2020 | Image credit: faithie/Adobe Stock/Title

This executive summary is based on the white paper *From data to AI – Intelligent data management as a basis for data science and the use of self-learning systems*, Munich, 2020. The authors are members of the working group Technological Enablers and Data Science of Plattform Lernende Systeme. The original version of this publication is available at: https://www.plattform-lernende-systeme.de/publikationen.html

SPONSORED BY THE

Federal Ministry of Education and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING