# The DSPN Vision

## Goal of this document

This document was authored by the Data Science Platform Netherlands (DSPN) to answer the following questions:

- What is data science (and how is it different from AI)?
- Why is data science relevant, in science and society?
- What are the main data science challenges in research and education?
- What will DSPN do in the Dutch data science community to address these?

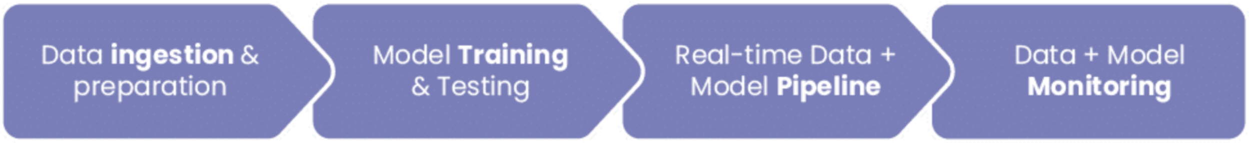
## What is Data Science?

Data Science is the discipline to apply, study, improve, and advance the methods and techniques to extract knowledge and insights from data in the awareness of the standards and requirements of the application context.[1] In 2001, William S. Cleveland was the first to define data science as a new field of study, in his Bell Labs technical report intended as an action plan for the practicing data analyst.[2] More than a decade later, the society has embraced this call for experts who combine a strong mathematical background with a solid understanding of computer science.

### Data Science and AI

Data Science is a fundamental enabler of Artificial Intelligence (AI). In the past decade there has been tremendous growth and attention for AI and, within that, Machine Learning (ML). Data Science is an integrated field that includes the use of ML, and some parts of AI, but is broader: it includes all the computer science necessary to process data, and do so at scale, and in a continuous fashion. Further, ML is not the only technique used for extracting knowledge and insights from data; alternative techniques are data mining, graph analytics, stream processing and analytical query processing. Thus, AI and Data Science have parts in common (notably, ML), and they are neighbouring fields that depend on each other. Data Science, for instance, increasingly depends on AI as many of the techniques and systems it comprises are now being enhanced with ML (e.g., by improving analytical query optimization using ML).

**Figure 1:** Typical data science workflow using ML


Data **ingestion** & preparation
Model **Training** & Testing
Real-time Data + Model **Pipeline**
Data + Model **Monitoring**

[1] https://web.archive.org/web/20210922130130/http://dstf.acm.org/DSReportDraft2Full.pdf

[2] https://web.archive.org/web/20060111162626/http://cm.bell-labs.com/cm/ms/departments/sia/ doc/datascience.pdf.

But *ML also depends on data science*. Figure 1 above shows a data science workflow using ML. The first step for a data scientist is to set up a data processing pipeline to ingest (explore, integrate and combine) multiple data sources and then do extensive preparation (data cleaning and transformation). This step notoriously takes >80% of effort and, these operations often require scalable data science systems (e.g. Spark)  in order to complete in a reasonable time. Only after thus acquiring clean first data to analyze, ML can be performed, to build and test a model. When a model is good enough, that version is taken into a real-time pipeline, which is again a data science infrastructure. Here, new data is periodically and automatically fed through the data ingestion and preparation pipeline set up previously, and the model is evaluated (scored), to make data- and ML-driven decisions. In order to guarantee quality outcomes, there is infrastructure that monitors both the operations in the pipeline itself, as well as the model features and all their data dependencies.

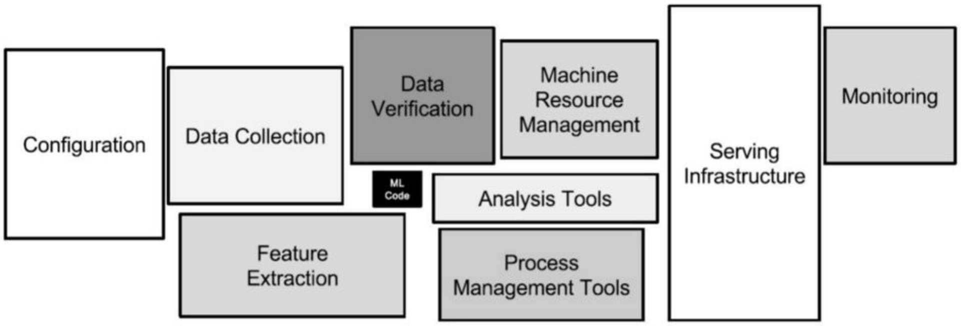


**Figure 2:** Only a small fraction of real-world ML-driven data science systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding data systems infrastructure is vast and complex[3].

As illustrated by Figure 2 in a publication[3] by Google, ML thus depends on data science for collecting data as input for training ML models, verifying its quality, and for extracting features to shape multiple data sources e.g. into a matrix representation. Once deployed, the quality of the input data and features needs permanent data and model quality monitoring and all of this processing needs to happen at scale, requiring resource management and (often, distributed) serving infrastructures. Therefore, the success of ML in moving from a very powerful technology into a tool that can be robustly deployed to solve problems, depends on data science infrastructure. Figure 1 illustrates that the common experience that >80% of ML project effort goes into data preparation (i.e. data science) also translates into the fact that an even greater majority of software infrastructure in an industrial ML deployment consists of data science infrastructure. Further, while Big Tech is able to construct the boxes in Figure 2, using access to top human talent, advances in data science technology should aim to make robust ML deployment more broadly attainable.

## Societal Impact of Data Science

As it has become possible in the past decades to collect data cheaply, all organizations have started to do so. By retaining more historical data, but also by collecting more data, e.g. instrumenting existing work processes with data collection. As such, the importance of data science extends to all sectors of society, and therefore (basic) skills and understanding of data science also need to extend to all professions. In

[3] Sculley, David, et al. "Hidden technical debt in machine learning systems." *Advances in neural information processing systems* 28 (2015): 2503-2511.

that sense, in most places where statistics is considered to be part of an educational curriculum, we expect over time this to be augmented into data science (of which statistical methods are yet another component).

Apart from including data science in curricula of all disciplines, we also anticipate these disciplines to start changing their practices and processes, to make data collection, refinement, processing and analysis an integral part of better achieving their goals. A few examples are: the medical sciences (e.g. to use data to predict and treat conditions better and more timely), the humanities (e.g. using natural language processing and knowledge graphs to better understand literary data sources), engineering (e.g. deployment of sensors to instrument factory processes and analyze data in real-time to optimize these processes but also perform predictive equipment maintenance).

In any discipline, it is possible to come up with such examples. In fact, the ability to leverage these data opportunities is what will make or break company competitiveness, but arguably also the environment, and the responsible use of data science will strongly influence the quality of democracy and of society as a whole. Because of the necessity of dealing with data in different contexts, of different types, in different locations, data science also needs to be tailored and developed in myriad ways.

Applying data science in society is generating significant growth in the IT sector of The Netherlands, with many professionals and companies incorporating data science into their products and services. Also, a new sector specializing in creating technology to do data science is emerging. Examples are the entrance of large data companies like Uber, Booking, Google AI and Databricks[4] in The Netherlands, but also many start-ups. Growing this data technology ecosystem from the academic perspective, via research, education and community-building, is a goal of DSPN.

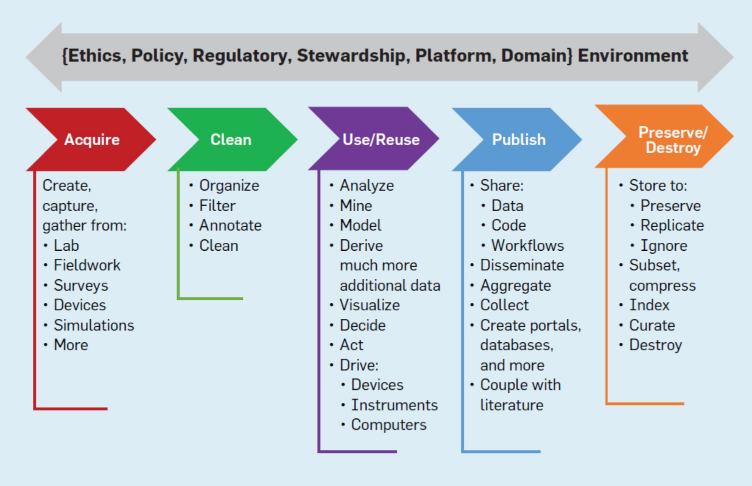


**Figure 3:** The Data Life Cycle[5]

Adoption of data science also produces new societal and organizational challenges. The biggest pitfall of data science is drawing wrong conclusions based on wrong data  (“garbage in, garbage out”). As such, formalizing, judging and improving data quality is an important research direction. Related to this, we

[4] https://web.archive.org/web/20200104141726/https://databricks.com/blog/2019/10/30/why-we-are-investing-100-million-euros-in-our-european-development-center.html

[5] Francine Berman et. al. 2018. Realizing the potential of data science. Commun. ACM 61, 4 (April 2018), 67–72. DOI:https://doi.org/10.1145/3188721

know that biased input data leads to biased outcomes, even if the data is “correct”. Finding ways to make data science Fair, Accurate, Confidential and Transparent (FACT) is going to be pivotal to combat bias, and therefore, have the application of data techniques accepted by society. Making data Findable, Accessible, Interoperable, and Reusable (FAIR) is important to make sure the benefits of data science will be reaped as broadly as possible, also by individuals, small organizations and by people in economically deprived regions.

## Scientific Research Challenges in Data Science

Figure 3 shows the Data Life Cycle, showing again that data science is a broad field. Given the broad need to apply data science, we observe a wide variety of research initiatives in The Netherlands from many disciplines *outside* computer science and mathematics, that carry the labels of Big Data, Data Science and ML. These aim to apply data science in specific domains. DSPN stresses the need to also perform top-notch *computer science* research into the *foundations* of data science. These foundations can be theoretical (new formalisms), algorithmic (new methods) and in systems (new tools). The Netherlands has a strong computer science community performing state-of-the-art data science research. Let us now summarize those areas of data science where there is critical mass in The Netherlands, specifically:

- Data Systems (e.g. CWI, TUe, UU, TU Delft, UvA, UT)
- Data Quality and Data Integration (e.g. CWI, TUe, UU, TU Delft, UvA, UT)
- Knowledge Representation & Provenance (e.g. VU, MU, UvA, UU, CWI)
- Data Mining and Exploration (e.g. UU, TUe, LU, TU Delft)
- Information Retrieval (e.g. UvA, RU, TU Delft, LU)
- Responsible Data Science: FACT & FAIR (e.g. VU, UvA, TUe, TU Delft, UT)
- Data Science in Interaction with Application Context (e.g. RUG, WUR, …)

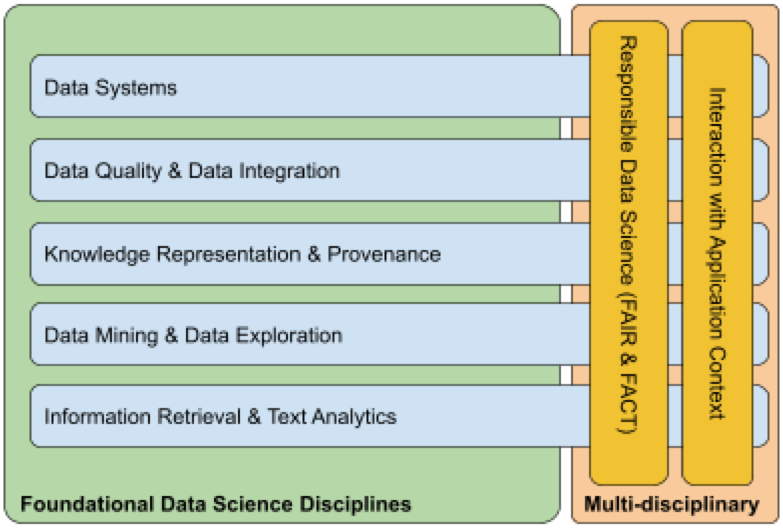

Data Systems
Data Quality & Data Integration
Knowledge Representation & Provenance
Data Mining & Data Exploration
Information Retrieval & Text Analytics
Foundational Data Science Disciplines
Responsible Data Science (FAIR & FACT)
Interaction with Application Context
Multi-disciplinary

**Figure 4:** The Data Science-related research expertise in the Netherlands.

There are two further fields that are part of data science and are related to this vision. Those are:

- Artificial Intelligence (AI), which we did not include because this field is covered well by SIGAI, the sister-SIG of DSPN in IPN, with whom DSPN collaborates in its activities.
- Statistics, which we did not include as it falls under mathematics and is represented by IPN sister PWN. It is an aim of DSPN to collaborate with statisticians and interface with PWN where fitting.

We now summarize the most important research challenges per area of critical mass, covered by DSPN.

**Data systems.** Data systems are a core technology in IT as a whole, but also in data science, where Big Data volumes need to be efficiently captured, stored (typically involving columnar compressed storage) and queried, using efficient execution techniques, as well as automatic query optimization and the required data and algorithms to guide this, in the face of complex data layouts and correlations. We should research the end-to-end data-to-insights pipeline, including understanding systems that go all the way from raw data to an end-user's desired outcome, such as a visualization, or a prediction by a ML model. We should research declarative programming paradigms to specify and optimize all stages of machine learning pipelines (data discovery, data preparation, and model building) to *better support ML with data systems*.

It remains important that data systems are efficient. Hardware will become very heterogeneous (CPUs, GPUs, FPGAs, ASICs), posing a software challenge to make programming them simple and portable. In the cloud, disaggregation of resources for elasticity challenges existing data system assumptions. Both disaggregation and new hardware require system architecture research. Data can be on-premise, in the cloud and on the edge; but in all cases human time to manage it is scarce and expensive. Therefore, we need to research *self-managing data systems*.

A data systems research area is *confidential data processing* without trusting a cloud provider. As we continue to aggregate data, the need to uphold *data privacy* with analytical usage of such data for decision support has emerged as a key challenge. There is a need for novel private data management infrastructures that better uphold privacy of citizens, and creating these will pose systemic and algorithmic challenges. Cryptographic techniques as well as *differential privacy* have emerged as a foundation of privacy-aware data systems. Data sharing across organizations will require new *multi-party computation* techniques to be integrated in data systems.

**Data Quality and Data Integration.** It is important to reduce the >80% effort spent on *data integration and wrangling,* aimed to improve the quality of data and allowing different data sources to be correctly combined. To have confidence in any data-driven decision, there should be a trust on the data on which that decisions were based. For this trust to materialize, the data has to be of high quality, i.e., accurate, complete, consistent and up-to-date. Unfortunately, most real-world datasets have data quality issues with significant consistencies. There is a need to develop tool techniques and methodologies that are able to clean the data from quality issues, but also, in the case this is not possible or is too expensive, to quantify the impact that these issues have to the results of the data analytics. Even after the data has been cleaned, the data has to be maintained. Recent advances in data collection and storage systems have allowed the creation of huge datasets that are often hard to manage. Parts of them have often to be forgotten. Such an action has to be done in ways it does not jeopardize future needs or minimizes the impact to future analytics. The notion of quality extends naturally to the algorithms that are applied on the data. The trust on the final result of any analytic process is equally dependent on the data on which the process is applied as on the process itself.

**Knowledge Representation & Provenance.** Knowledge Representation has its roots in expert systems and traditionally concerns curated explicit knowledge, such as embodied by Tim Berners Lee's early vision of the Semantic Web. This vision has been adopted at scale in the form of Knowledge Graphs (KGs). As KGs grow larger, they can no longer be manually constructed, and research is needed into automated extraction of specific knowledge and its context, from large, diverse, and noisy data sources. KGs are therefore also no longer consistent, but fuzzy, and research thus needs to make reasoning probabilistic. Semantics and reasoning are important foundational techniques for answering of complex queries, also posed in natural language. The field uses a diversity in methods to create knowledge, among which reasoning, language modeling, text analytics, and machine learning; as well as a diversity of methods to integrate knowledge into query answering and decision making. On top of that, the ambition to further enhance Machine Learning exposes the question whether integration of explicit knowledge, can lead to better ML performance, combining the fields of explicit knowledge and learned (black box) models.

Explicit knowledge that describes data is a key idea to enable automatic data integration. This not only concerns the semantics of data, but also other meta-information, e.g. describing how data was obtained and its validity and quality. We thus need to study *data provenance* and its application in data science pipelines, which involves tracking, integrating, and analyzing metadata. Provenance is also required for reproducibility. We should research tools and systems that support *data sharing*, including labeling, annotating, exchanging, securing, discovering, and capturing provenance of data. This is needed to support auditing at scale so that checks for legitimate usage can be implemented at possible fine levels of granularity.

**Data Mining & Exploration.** Data Mining studies targeted extraction of knowledge from data. Current research topics include mining of knowledge from heterogeneous data sources, also other than tables, such as multi-modal data (including picture, sound, video, sensors) or graphs. Graph similarity, graph clustering, graph pattern matching, and influence maximization are active topics where more efficient algorithms are advancing the state of the art; where the outcome of the analysis may also be graph-shaped such as in graph neural networks. Foundational challenges in data mining are on *causality* and *data quality*. While it is very hard to disentangle correlation and causality in static after-the-fact data, there are new techniques that make progress on causality. This is important because data science conclusions drawn from correlation rather than causation are a risk to its integrity. The detection and correction of data quality problems (missing/incomplete, noisy, inaccurate data), also called *data cleaning*, and the assessment of data quality and how it affects models and the quality of predictions, and automating this is of crucial importance to enable data science and safeguards it from its biggest pitfall ('garbage in, garbage out').

*Process mining* studies the analysis of large-scale log data, for reconstructing, understanding, and optimizing the processes that created this data. Active topics are the detection of missing information (hidden events), the integration of a diversity of perspectives, and challenges in dealing with missing, noisy and biased data. Finally, *Data Exploration* techniques such as *Visual Analytics* seek to aid humans in understanding data, by (automatic) summarization, discovering outliers and counterfactuals and presenting these to the user interactively, combining foundational research in human-computer interaction, (immersive) data visualization and reality augmentation; with (large-scale) data analytics.

**Information Retrieval & Text Analytics**. Information Retrieval (IR) is the science of searching for information in non-tabular data, typically text documents, but possibly also other data such as video, images or sound; typically ranking data items and presenting a top-N of query answers. Information created by, connected to, or consumed by an individual now resides across a great number of separate information silos: personal devices; the web; file systems; messaging systems and social media; and systems from external parties including doctors, banks, employers and government. Rather than searching information in an equivalent number of independent search systems, future *personal IR* systems should integrate across all of these. They should also support complex, evolving, or long-term information seeking goals such as acquiring broad knowledge either for its own sake or to make an informed decision.

*Conversational Information Seeking (CIS)* concerns research where humans interact with the IR system in a dialog, where the input may be non-textual (spoken, gesture). It may take into account long term user state, user needs beyond topical relevance (e.g. the form of presentation), and permitting initiative to be taken by either the user or the system at different points of time. As described in the sequel, detecting and correcting for bias in data and ranking increasingly is an urgent research topic in IR.

**Responsible Data Science: FAIR & FACT**. FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability. This makes data more valuable as it is easier to find and easier to combine and integrate thanks to a formal shared knowledge representation. It creates economic opportunities and a more level playing field, as people with the best ideas may access to FAIR data; and is also key to making the results of (data) science transparent and reproducible. DSPN strives to help (research) data become FAIR, by researching, creating and supporting novel infrastructure for this.

Further foundational research in all data science disciplines is focusing on FACT, i.e., questions related to Fairness, Accuracy, Confidentiality, and Transparency. Data science approaches learn from training data while maximizing an objective. However, this does not imply that the outcome is *fair*. The training data may be biased and defining and detecting this remains an open question, as well as correcting for bias. New data science approaches should also provide insight on the *accuracy* of the output. If individuals do not trust the underlying data pipelines, they may worry about *confidentiality* which could prevent them from sharing their data. Data science practices and tools should prevent data confidentiality breaches. Finally, data science should not be viewed as a black box that magically transforms data into insight and value. Many design choices need to be made in a data science pipeline; the journey from raw data to meaningful conclusions involves multiple steps and actors, thus accountability and comprehensibility are key for *transparency*.

**Data Science in Interaction with Application Context:** Data Science has relevance to a wide range of application domains, and its application always happens in a context. The extraction of knowledge from data is done by people applying the techniques. The techniques themselves can never do this. This is a fundamental limitation and the need to interface with other disciplines is crucial in data science. Data science thus demands domain expert knowledge in order to understand the meaning of data and the pipelines constructed for processing them. In addition, specific application domains sometimes require specific and specialized data science techniques, such as algorithms, storage formats and data processing sub-systems. For example, application areas introduce challenges in the form of specialized data formats such as specialized string processing(genomics, DNA) and chemistry (SMILES codes) and biochemistry (FASTA codes). The discipline of data science covers these problems, where the interaction between the clean, idealized world described by the fundamental mathematical models, and the complex, large and diverse universe of real-world data to be handled by efficient, effective, energy-aware actual algorithms provides many exciting challenges.

## Education Challenges in Data Science

The Dutch computer science education system provides in-depth courses in data science in the bachelor, master and postgraduate levels, catering to a wide variety of communities. Here, we outline key educational challenges.

**Data Science Foundations.** Data science education at computer science faculties aims to educate data science professionals with a strong technical background, who are able to design, analyze and optimize complex intricate scalable data science infrastructures. This includes a deep understanding of data science technologies, including the ability to generate new foundational technologies. DSPN sees a role to harmonize education in both bachelor and master level data science curricula by establishing common guidelines and best practices.

**Data Science For All.** At the moment there is a tremendous and unfulfillable demand for data scientists. This phenomenon is a result of the lack of data literacy skills at non-CS science curricula. This high demand for data science skills leads non-CS scientists to study Data Science in CS departments either as a minor or as a master. We believe that in order to satisfy the demand for data scientists across all sciences, data science foundations and skills need to be integrated as part of all science curricula. In most curricula where there is now a statistics course, there is likely a demand for extending it to include data science. The breath of this challenge is daunting, as it is inconceivable that computer science departments who are now already overstretched could provide for this educational demand. Therefore, DSPN proposes to focus on the design of a core curriculum, in collaboration with other departments. International experience has shown that the most successful data science programs in the world are the result of collaboration between multiple organizational units within or across institutions[6].

## The Role of the DSPN Community

As a SIG of IPN, DSPN unites Dutch researchers in data science, with a focus on stimulating foundational research to create systems, techniques and methods to make data science practitioners more effective. Recognizing the computer science core of data science, we do so with computational and algorithmic focus. Data science is and should be well-connected to its application in society and industry and DSPN will foster research and education of people that lead to the creation of data, systems, models, and artifacts that are usable and provide societal and industrial impact.

[6] Francine Berman et. al. 2018. Realizing the potential of data science. Commun. ACM 61, 4 (April 2018), 67–72. DOI:https://doi.org/10.1145/3188721