Week1: Intro

Hakan Mehmetcik

Collective Statistical Illiteracy: Understanding the Challenge

In contemporary society, almost every news article or broadcast includes some form of scientific or numerical data. Headlines routinely mention statistics about public health, economic indicators, environmental changes, and more. However, a significant portion of the audience struggles to interpret these numbers and to understand how they were derived. This difficulty can have profound implications for personal decision-making, organizational strategy, and public policy.

The concept of **collective statistical illiteracy** highlights this widespread inability to grasp the meaning behind statistical facts and figures. When individuals lack the skills to critically evaluate quantitative information, they are more susceptible to misunderstanding important issues, making poorly informed decisions, or being swayed by misleading statistics.

> ***Statistical*** *thinking will one day be as necessary for efficient citizenship as the ability to read and write.*
>
> H.G. Wells (1903, paraphrased by S.S. Wilks, see link)

Why Study Data Science?

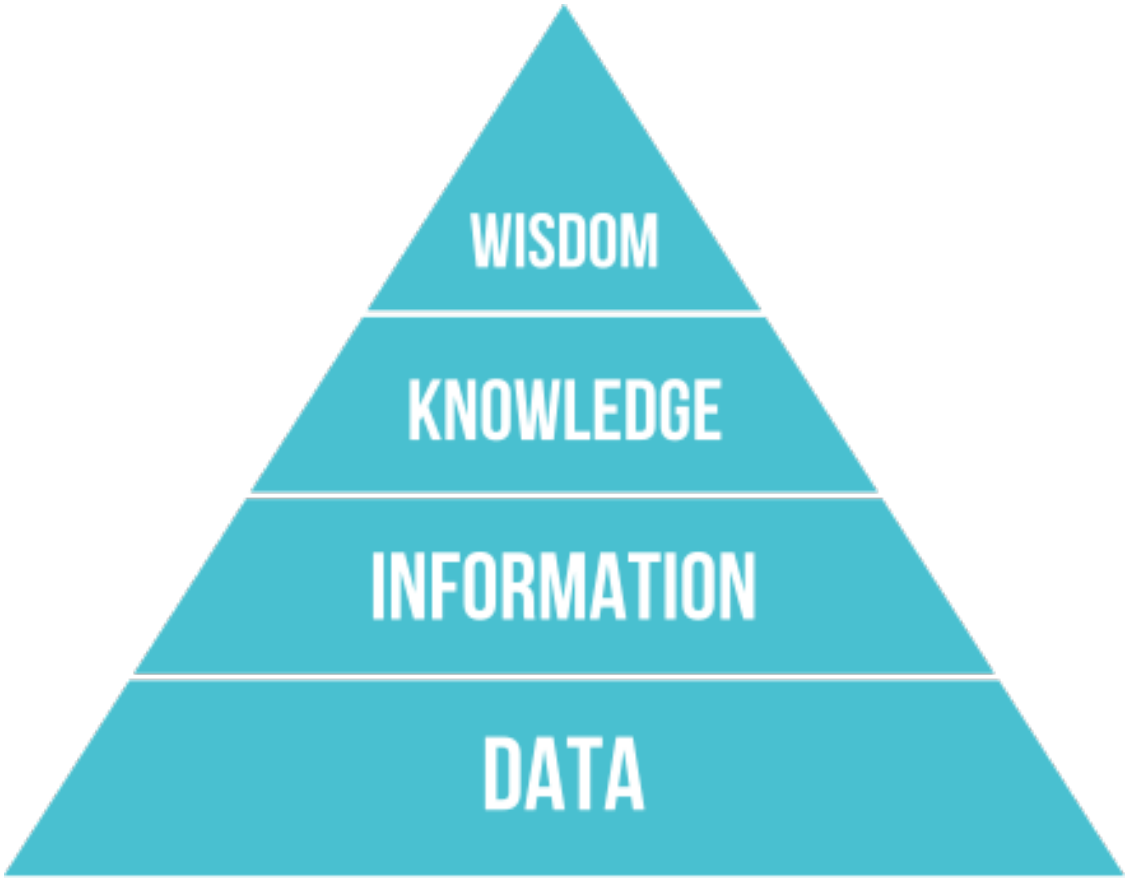
Data science drives innovation and decision-making in virtually every industry—from healthcare and finance to social media and marketing. By combining computational methods with statistical analysis, data science uncovers hidden patterns and insights that help organizations craft evidence-based strategies, predict future trends, and make more informed decisions. As the volume of data grows exponentially, the ability to collect, process, and interpret this information becomes ever more critical.

How Is Data Science Different from Math and Statistics?

- **Interdisciplinary Approach**: While mathematics and statistics focus on theoretical models and the science of uncertainty, data science integrates computer programming, domain knowledge, and data engineering alongside statistical techniques.
- **Practical, Data-Driven Focus**: Data scientists often work with real-world data that may be messy or unstructured. They prioritize building reproducible workflows and scalable analyses, which is not the central concern in pure math or traditional statistics.
- **Technology and Tools**: Data science heavily relies on programming languages (e.g., Python, R) and tools (e.g., machine learning frameworks, data visualization platforms) to handle large, complex datasets—capabilities not typically emphasized in standard math or statistics coursework.

Data science—the science of extracting meaningful information from data. As such, data science is a broader field than statistics, encompassing data gathering, preparation, modeling, visualization, computing, and meta-analysis of data science itself.

> **Note**
>
> An assumption underlying these efforts is that data is the foundation of various information types that can eventually be turned into knowledge and wisdom. Figure below shows their arrangement in a hierarchical structure of a DIKW pyramid:



The key distinction between the lower and the upper layers of the pyramid is that data needs to address some hypothesis or answer some question, and has to be interpreted and understood to become valuable. Here to say, another reason for the close interaction between data and theory is that we need theoretical models for understanding and interpreting data. When analyzing data, we are typically interested in the underlying mechanisms (i.e., the causal relationships between variables). Importantly, any pattern of data can be useless and misleading, when the data-generating process is unknown or ignored. Knowledge or at least assumptions regarding the causal process illuminate the data and are required for its sound interpretation. The importance of theoretical assumptions for data analysis cannot be underestimated (see, e.g., the notion of causal models and counterfactual reasoning in Pearl & Mackenzie, 2018). Thus, **pitting data against theory is nonsense**: Using and understanding data is always based on theory.

Our aims is to prepare **well-documented and reproducible** analysis since data literacy is the ability and skill of making sense of data. This includes numeracy, risk-literacy, and the ability of using tools to collect, transform, analyze, interpret, and present data, in a transparent, reproducible, and responsible fashion.

Information is what we want, but data are what we've got. The techniques for transforming data into information is the ***data science!***

Data scientists, therefore, are individuals who strive to transform the plentiful data available today into actionable information, which often appears to be in short supply

Think with data = desire to solve problems using data

A data scientist is "a knowledge worker" who is principally occupied with analyzing complex and massive data resources.

Demand for data skills is strong! Data Scientist is one of the best job in the world since 2016.

The Key components that are part of ***data acumen*** include mathematical, computational, and statistical foundations, data management and curation, data description and visualization, data modeling and assessment, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving.

Therefore, contemporary data science requires **tight integration of statistical, computational, and communication skills.**

Data Wrangling a process of preparing data for visualization and other modern techniuqes of statistical interpretations.

> **Note**
>
> **Recommended Background Readings**
>
> 1. **Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2021). *Modern Data Science with R.***
>   - **Chapter 1: Prologue: Why Data Science?**
>   - This chapter introduces the motivation behind data science as a discipline. It discusses how the field has evolved, its interdisciplinary nature, and why it is relevant today. The reading helps students understand the **philosophical and practical reasons for studying data science**.
> 2. **Donoho, D. (2017). "50 Years of Data Science." *Journal of Computational and Graphical Statistics*, 26(4), 745-766.**
>   - A historical perspective on the evolution of **data science as a field**, emphasizing how it extends beyond traditional statistics.
>   - Discusses the distinction between **data science and statistics**, highlighting the importance of computational tools.

- Advocates for a more **inclusive and interdisciplinary** approach to data science education.

3. **De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., & others. (2017). "Curriculum Guidelines for Undergraduate Programs in Data Science."** ***Annual Review of Statistics and Its Application*****, 4, 15-30.**

    - Provides an overview of the **core competencies needed for data science education**, including programming, statistical inference, and domain expertise.
    - Defines **guidelines for structuring an undergraduate data science curriculum**, which is useful for understanding how your course fits within broader educational frameworks.
    - Emphasizes the importance of **ethics, reproducibility, and communication skills** in data science training.

What is data?

Today, the manner in which we extract meaning from data is different in two ways—both due primarily to advances in computing:

- we are able to compute many more things than we could before, and,
- we have a *lot* more data than we had before.

the traditional two-dimensional representation of data: rows and columns in a data table, and horizontal and vertical in a data graphic. For instance, if someone aimed to collect or compare health-related characteristics of some people, he or she would measure and record these characteristics in some file. In such a file of health records, some values may identify persons (e.g., by name or some ID code), while others describe them by numbers (e.g., their age, height, etc.) or various codes or text labels (e.g., their address, profession, diagnosis, notes on allergies, medication, vaccinations, etc.). The variables and values are typically stored in tabular form: If each of the table's rows describes an individual person as our unit of observation, its columns are called *variables* and the entries in the cells (which can be referenced as combinations of rows and columns) are *values*. The contents of the entire table (i.e., the observations, variables, and values) are typically called "data".

Non-traditional data types (e.g., geospatial, text, network, "big") and interactive data graphic are the new normal! Thus, we can argue that the gap between generating scientific insights and understanding them is widening.

The increasing complexity and heterogeneity of modern data means that each data analysis project needs to be custom-built. Simply put, the modern data analyst needs to be able to read and write computer instructions, the "code" from which data analysis projects are built.

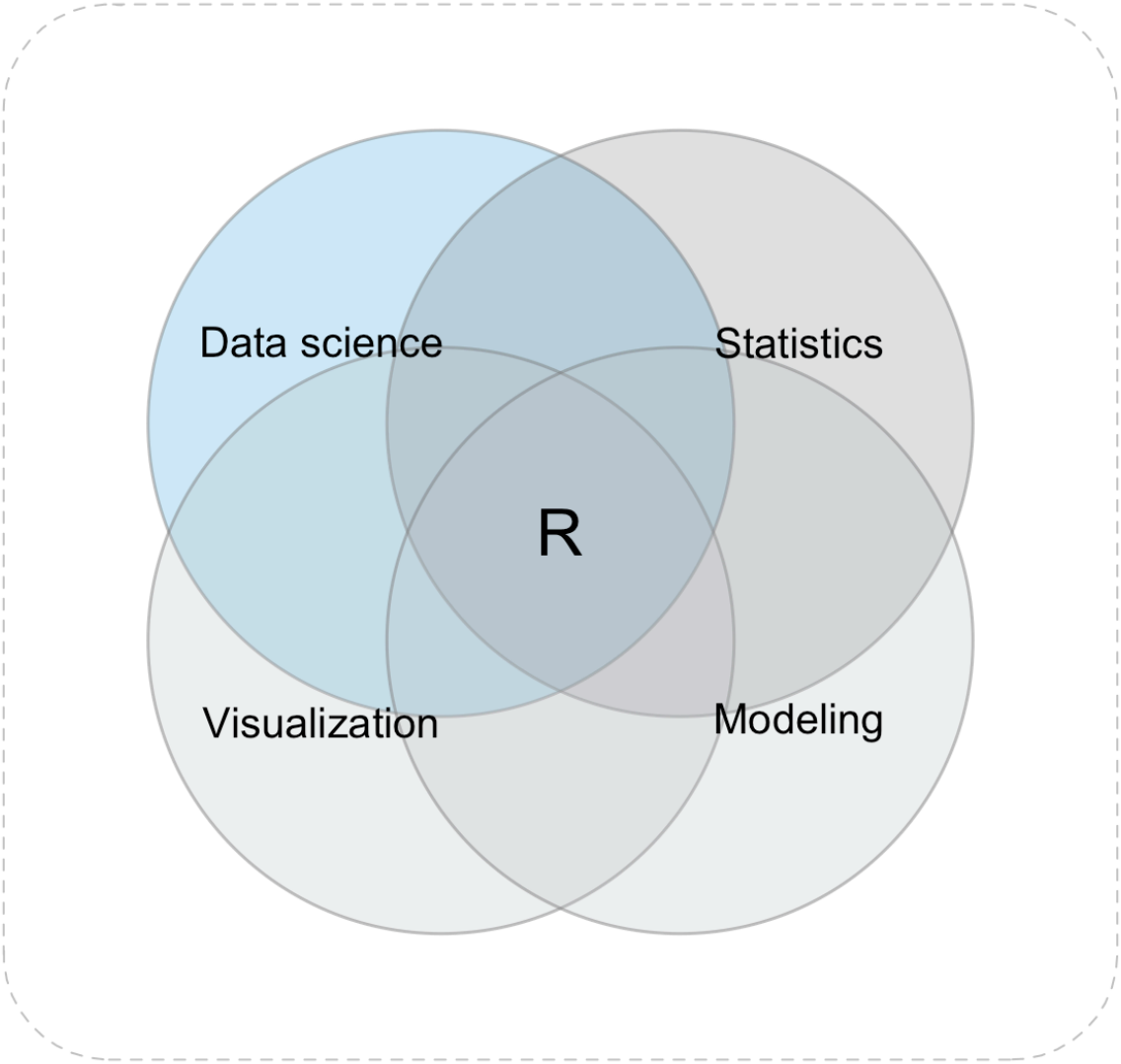
Domain knowledge is always useful in data science since data science is best applied in the context of expert knowledge about the domain from which the data originate. For data scientists of all application domains, creativity, domain knowledge, and technical ability are absolutely essential.

Some related discussions in this contexts include:

1. The distinction between *raw data* (e.g., measurements, inputs) and *processed data* (results or interpretations): Data are often portrayed as a potential resource: Something that can be collected or generated, harvested, and — by processing — refined or distilled to turn it into something more valuable. The related term *data mining* also suggests that data is some passive raw material, whereas the active processing and interpretation of data (by algorithms, rules, instructions) can generate interest and revenue.
2. The distinction between *data* and *information*: In contrast to data, the term *information* has a clear formal definition (see Wikipedia: Information theory). And while data appears to be neutral, information is usually viewed as something positive and valuable. Hence, when data is being used and perceived as useful, we can generate *insight*, *knowledge*, or perhaps even *wisdom* from it?
3. The field of signal detection theory (SDT, see Wikipedia: Detection theory) distinguishes between *signal* and *noise*. Does *data* include both signal and noise, or should we only count the signal parts as data?

What is Data Science (DS)?

DS is not a single and homogeneous discipline. Instead, it overlaps and is intricately interwoven with several other academic fields and requires corresponding skills. For instance, becoming an expert in DS requires considerable knowledge in statistics, but the discipline of data science is not as mathematical as statistics. Similarly, data science involves computers, but is usually not as much concerned with abstract formalisms as computer science. We could describe DS as applied statistics or applied computer science, but this would suggest that it is a sub-discipline of these other fields — and any serious application ususally assumes knowledge of the discipline's foundations.



As R is defined as a "software environment for statistical computing and graphics" (R Core Team, 2023), it supports and provides tools for all topics mentioned and is thus pictured at the center. However, this central position does not imply that R is the only or a necessary tool for DS (e.g., a popular alternative is the programming language Python).

> **Note**
>
> A striking example of bad software choices is the Public Health England (PHE)'s recent decision to import CSV-files on COVID-19 test results into Microsoft Excel's XLS file format. Due to this file format's artificial limit to a maximum of 65.536 rows of data, nearly 16.000 Covid cases went unreported in the U.K. (which amounts to almost 24% of the cases recorded in the time span from Sep 25 to Oct. 2, 2020).

The skills for successfully dealing with data are not confined to one discipline or talent. As data scientists must discover, mine, select, organize, transform, analyze, understand, communicate and present information, they tend to be generalists, rather than specialists. Beyond a set of skills from a diverse range of areas, getting DS done requires the familiarity with and mastery of suitable tools.

The *same* data can be represented in many *different types* and *shapes*.

Types of data

As variables and values depend on what we want to measure (i.e., our goals), it is impossible to provide a comprehensive list of data types. Nevertheless, computer scientists typically categorize data into different types. This is made possible by distinguishing between different ways in which data is represented. The three most basic types of data are:

1. *Truth values* (aka. *logicals*): either `TRUE` or `FALSE`
2. *Numbers*: e.g., 2,12,√22,12,2
3. *Text* (aka. *characters* or *strings*): e.g., "Donald Duck" or "War and Peace"

Most computer languages have ways to represent these three elementary types. Exactly how truth values, numbers, or text relate to the actual phenomena being described involves many issues of representation and measurement.

In addition to these three basic data types, there are common data types that have precise and widely-shared definitions, like

- *dates* and *times*: e.g., 2025-02-10, `11:55`
- *locations*: e.g., the summit of Mount Everest, or N40 44.9064', W073 59.0735'

but even more potential data types that we could define, if we had or wanted to, for instance

- *temporal terms*: e.g., `Monday`, `noon`, or `tomorrow`
- *visualizations*: e.g., a bar chart, or Venn diagram

As we can express more complex measures in terms of simpler ones (e.g., as numbers or by some descriptive text), we can get quite far by combining our three basic data types (e.g., dates, times and locations can be described as combinations of characters and text).

Shapes of data

Beyond distinguishing data types, we can also ask about the *shape* of data or representations. While it gets challenging to answer questions like "How is the number 2 represented in the brain?" or "What is the shape of yellow?", we can simplify our lives by asking: "In which shape do computers represent data?".

1. *scalars*: e.g., 1 or `TRUE`
2. 1-dimensional vectors and lists: e.g., `1, 2, 3` or `'x', 'y', 'z'`
3. 2-dimensional matrices: e.g., a table of health records for different people

Just as for data types, we can easily extend these basic shapes into more complex data formats, like

- n-dimensional arrays, e.g., the number of Titanic passengers by `age`, `sex`, and `survival`
- non-rectangular data, e.g., a list of sentences, what someone did last summer

> **Note**
>
> The contents are mostly based on **Introduction to Data Science** https://bookdown.org/hneth/i2ds/intro.html