Fundamental Concepts

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

About these slides

These slides are companion material for the book

Data Science Project: An Inductive Learning Approach

Prof. Dr. Filipe A. N. Verri

`https://leanpub.com/dsp`

All intellectual content comes from the book and is not AI-generated. Slides were produced with the assistance of Claude Code.

The simple believes everything,
but the prudent gives thought to his steps.

— Proverbs 14:15 (ESV)

Overview

Contents

- Data science definition
- The data science continuum
- Fundamental data theory

Objectives

- Define data science
- Present the main concepts about data theory

Data science definition

Definitions in the literature

- **Zumel & Mount:**[1] cross-disciplinary practice drawing on data engineering, statistics, data mining, ML, and predictive analytics
- **Wickham & Grolemund:**[2] discipline that transforms raw data into understanding, insight, and knowledge
- **Hayashi:**[3] not only unifies statistics and data analysis, but intends to analyze and understand actual phenomena with data

[1] N. Zumel and J. Mount (2019). ***Practical Data Science with R***. 2nd ed. Shelter Island, NY, USA: Manning.

[2] H. Wickham, M. Çetinkaya-Rundel, and G. Grolemund (2023). ***R for Data Science: Import, Tidy, Transform, Visualize, and Model Data***. 2nd ed. O'Reilly Media.

[3] C. Hayashi (1998). **"What is Data Science? Fundamental Concepts and a Heuristic Example".** In: *Data Science, Classification, and Related Methods.* Ed. by C. Hayashi et al. Tokyo, Japan: Springer Japan, pp. 40–51. ISBN: 978-4-431-65950-1.

Our definition

Definition: Data science

Data science is the study of knowledge extraction from measurable phenomena using computational methods.

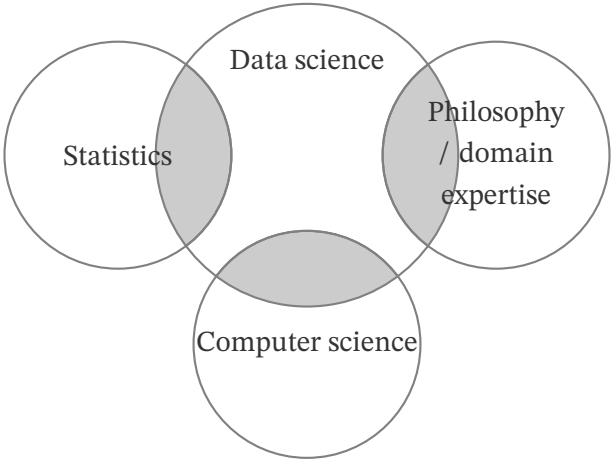
Key terms:

- **Computational methods** — use computers to handle data
- **Knowledge** — information humans can understand and apply
- **Measurable phenomena** — events we can quantify
- **Raw data** — collected directly, not yet transformed

Data science vs conventional sciences

- Conventional: named after object of study (biology $\rightarrow$ life)
- Data science: studies *how* to extract knowledge from data
- Similar to "computer science" — not the study of computers
- Conventional paradigm: model-driven (observe, hypothesize, validate)
- Data science paradigm: data-driven (extract knowledge from data)
- We give data the opportunity to **surprise us**

My view of data science



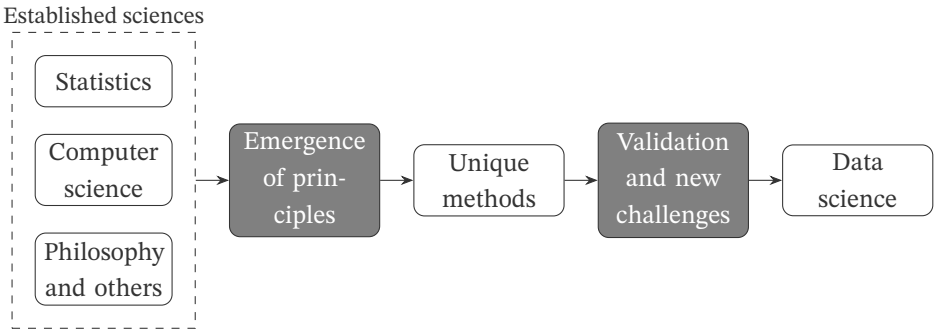
An entirely new science — its basis comes from other sciences, but its object of study is particular enough to raise new questions.

The data science continuum

From borrowed methods to a new science

- Object of study in data science is not new
- Key factor: social demand and importance of data
- “Data is the new oil”
- Methods develop $\rightarrow$ experiments validate $\rightarrow$ credibility grows
- Academic recognition: new courses and programs
- Unique questions emerge, distinct from parent disciplines

The data science continuum



New questions in data science

- How can we guarantee that the data is reliable?
- How can we collect data without biasing conclusions?
- How can we guarantee that data usage is ethical?
- How can we present results to non-experts?

Fundamental data theory

Phenomena and philosophy

- Phenomenon: any observable event or process
- **Ontology** — study of being, existence, categories
- **Epistemology** — study of knowledge and justification
- **Logic** — study of reasoning and inference

Understanding phenomena requires both general philosophical knowledge and domain expertise.

Aristotle's categories

- Aristotle (384–322 BC) — first systematic classification
- Ten categories: substance, quantity, quality, relation, place, time, …
- Practical view: focus on the world we can perceive
- Foundation of logical reasoning and scientific classification
- Still used in computer and data systems

Why ontology matters for data science

- Describing = reducing complexity to simple pieces
- Phenomena $\approx$ substance; data $\approx$ properties, relations, states
- Identifying entities and their properties $\rightarrow$ better data collection
- Understanding logical limitations avoids errors
- Common mistake: assuming columns with the same name have the same meaning

Measurements

- Data science focuses on **measurable phenomena**
- Data collection: systematic gathering of data on a phenomenon
- Systematic = planned, with understood consequences
- **Sampling bias** — influence of the collection method on conclusions
- Data storage: digitally storing collected data

Data types

- Data types must correctly reflect the source phenomenon
- Data types restrict the operations we can perform
- Foundations from computer science:
  - Algorithms and data structures
  - Databases
- Concepts are independent of programming language or RDBMS

Reasoning

- **Deductive reasoning** — from general rules to specific conclusions
- **Inductive reasoning** — from specific observations to general rules
- Data science relies on **inductive reasoning**
- Descartes: algebra to mechanize reasoning
- Leibniz: universal algebraic language for logical methods

Knowledge extraction methods

- **Statistics** — collection, organization, analysis, interpretation
- **Machine learning** — algorithms that learn from data automatically
- **Operations research** — computational methods to optimize decisions
- Domain-specific methods: bioinformatics, geoinformatics, ...
- Each method has its own assumptions and limitations

Data preprocessing

- When data do not match the method's requirements
- **Data cleaning** — detecting and correcting corrupt data
- **Data transformation** — converting formats or types
- **Data enhancement** — integrating data from different sources
- Methods are usually robust to imperfections, but preprocessing helps

Takeaways

- Data science is a new science that studies knowledge extraction from measurable phenomena using computational methods
- The data science continuum: from borrowed methods to a distinct discipline
- Understanding phenomena, measurements, and knowledge extraction is fundamental

Questions?