A Brief History of Data Science

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.

“Begin at the beginning,” the King said gravely, “and go on till you come to the end: then stop.”

— Lewis Carroll, Alice in Wonderland

Overview

Contents

- The term “data science”
- Timeline and historical markers

Objectives

- Understand the history of the term
- Recognize major milestones
- Identify important figures

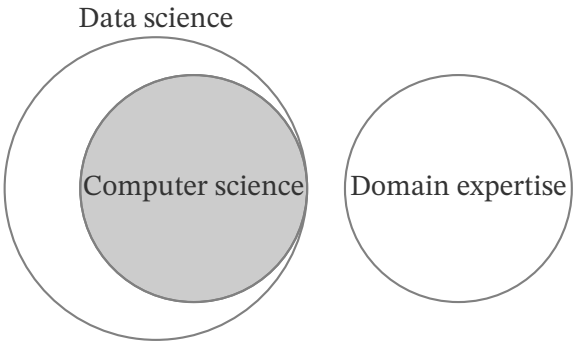
The term “data science”

Peter Naur (1928–2016)

- Danish computer scientist and mathematician
- Coined “data science” in the 1960s[1]
- Suggested replacing “computer science” with **datalogy**
- Emphasis on data as a fundamental component
- Data science deals with data; meaning is delegated to other fields

[1] P. Naur (1974). ***Concise Survey of Computer Methods.*** Lund, Sweden: Studentlitteratur. ISBN: 91-44-07881-1. URL: http://www.naur.com/Conc.Surv.html.

Naur's view of data science



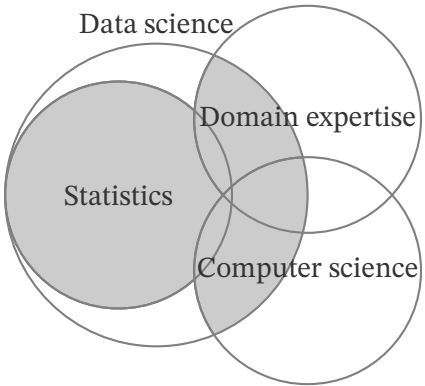
Data science studies techniques to deal with data,
but delegates the meaning of data to other fields.

William Cleveland (born 1943)

- American statistician
- Used “data science” in 2001 to name a new discipline[2]
- Plan to enlarge the major areas of statistics
- Data science = “modern” statistics + computing + domain expertise
- Credited with defining data science as used today

[2] W. S. Cleveland (2001). **“Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”.** In: *ISI Review*. Vol. 69, pp. 21–26.

Cleveland's view of data science



Statistics enlarged by computer science and domain expertise.

Buzzword or a new science?

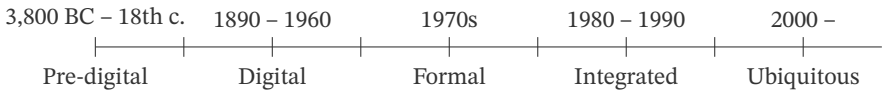
- No consensus on the definition of data science
- Most usages: rough reference to data-driven techniques or marketing
- Naur and Cleveland: enlarged scope of an existing science
- Counter-argument: social and economic demand supports a distinct science
- Many “data scientist” roles and degree programs emerging

Timeline and historical markers

Two timelines

- **Data handling** — sources, collection, organization, storage, transformation
- **Data analysis** — knowledge extraction and learning from data

Timeline of the ages of data handling



Pre-digital age

- Lebombo bone (∼40,000 years old) — earliest tally stick[3]
- First known census: 3,800 BC, Babylonian Empire[4]
- Sumerian archaic writing (∼3,500 BC) — recording transactions[5]
- Storage evolution: clay tablets, papyrus, codex, paper, printing press
- **Florence Nightingale** (1820–1910)
  - Used statistics to influence public opinion
  - Developed the polar area diagram
  - Standardized healthcare data collection

[3] P. B. Beaumont and R. G. Bednarik (2013). In: *Rock Art Research* 30.1, pp. 33–54. DOI: 10.3316/informit.488018706238392.

[4] C. G. Grajalez et al. (2013). **"Great moments in statistics".** In: *Significance* 10.6, pp. 21–28. DOI: 10.1111/j.1740-9713.2013.00706.x.

[5] G. Ifrah (1998). ***The Universal History of Numbers, from Prehistory to the Invention of the Computer.*** First published in French, 1994. London: Harvill. ISBN: 1 86046 324 x.

Digital age

- Punched cards — earliest famous usage by Bouchon (1725)
- 1890 US Census — first machine-readable punched cards
- **Herman Hollerith** (1860–1929) — tabulating machine, later IBM
- ENIAC (1945) — first electronic general-purpose computer
- UNIVAC I — used for the 1950 US Census
- Digital computers: exponential growth in capture and storage

Formal age

- **Edgar F. Codd** (1923–2003) — relational model (1970)
- Data organized in tables (relations)[6]; rows = records, columns = attributes
- Minimizes redundancy, improves integrity (normalization)
- Led to SQL (1974) — standard language for relational databases
- New challenge: aggregating data from different sources

[6] E. F. Codd (1970). **"A Relational Model of Data for Large Shared Data Banks"**. In: *Commun. ACM* 13.6, pp. 377–387. ISSN: 0001-0782. DOI: 10.1145/362384.362685.

Integrated age

- ETL (Extract, Transform, Load) process
- Data warehousing concept (late 1980s, IBM)
- **Ralph Kimball** vs **Bill Inmon** — bottom-up vs top-down
- Both agree: data warehouses are the foundation for BI and analytics
- Walmart case study (early 1990s) — single source of truth

Ubiquitous age — Key figures

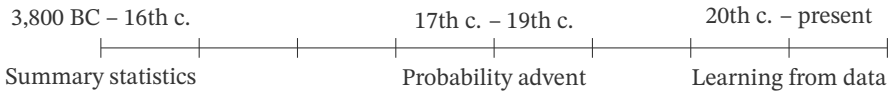
- **Cerf & Kahn** — TCP/IP protocols
- **Tim Berners-Lee** — World Wide Web
- **Jobs, Wozniak, Gates** — personal computers
- **Page & Brin** — Google
- **Zuckerberg** — Facebook and social media

Ubiquitous age — Technologies

- NoSQL databases — unstructured, semi-structured, structured data
- Five V's of big data: Volume, Velocity, Variety, Veracity, Value
- **Cutting & Cafarella** — Apache Hadoop (2006), HDFS, MapReduce[7]
- Distributed computing frameworks (Hadoop, Spark)
- IoT — proliferation of data sources

[7] J. Dean and S. Ghemawat (Jan. 2008). **"MapReduce: simplified data processing on large clusters".** In: *Commun. ACM* 51.1, pp. 107–113. ISSN: 0001-0782. DOI: 10.1145/1327452.1327492.

Timeline of the ages of data analysis



Summary statistics

- Earliest form of statistical analysis
- Term “statistics” refers to analysis of data *about the state*
- Central tendencies (e.g., arithmetic mean)
- Variability (e.g., range)

Probability advent

Foundations of modern probability theory (17th century):

- Blaise Pascal (1623–1662)
- Pierre de Fermat (1601–1665)
- Christiaan Huygens (1629–1695)
- Jacob Bernoulli (1655–1705)

Led to the field of **statistical inference**.

Bayes' rule

- **Thomas Bayes** (1701–1761)
- Calculates conditional probabilities from evidence
- Foundation of **learning from evidence**[8]
- Naïve Bayes classifiers likely used since the 18th century

[8] T. Bayes (1763). **"An Essay towards Solving a Problem in the Doctrine of Chances".** In: *Philosophical Transactions of the Royal Society of London* 53. Communicated by Richard Price, pp. 370–418.

Gauss' method of least squares

- **Carl Friedrich Gauss** (1777–1855)
- Developed circa 1794 for calculating the orbit of Ceres[9]
- Beginning of **regression analysis**
- Shift: solving overdetermined systems using data

[9] C. F. Gauss (1809). ***Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium.*** Latin. Hamburg, Germany: Friedrich Perthes und I. H. Besser.

Playfair's data visualization

- **William Playfair** (1759–1823)
- Invented line, area, bar chart, pie chart, circle graph[10]
- Graphical representation of economic data
- Changed how we communicate data insights

[10] W. Playfair (1786). ***The Commercial and Political Atlas.*** Representing, by Means of Stained Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England during the Whole of the Eighteenth Century. London, England: Debrett; Robinson; and Sewell.

Learning from data

- 20th century onward
- Development of **learning machines**
- Shift from fitting theoretical models to extracting knowledge from data
- Algorithms that learn with minimal human intervention
- Fueled by advances in computation and data storage

Fisher's discriminant analysis (1930s)

- **Sir Ronald A. Fisher** (1890–1962)
- Linear functions to separate classes of objects[11]
- Foundation of classification and dimensionality reduction
- Highlighted the importance of **feature selection**

[11] R. A. Fisher (1936). **"The Use of Multiple Measurements in Taxonomic Problems".** In: *Annals of Eugenics* 7.2, pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

Shannon's information theory (1940s)

- **Claude Shannon** (1916–2001)
- Quantification, storage, and communication of information[12]
- Key concepts: entropy, mutual information, information gain
- Data as a sequence of symbols that can be compressed
- Foundation of several machine learning algorithms

[12] C. E. Shannon (1948). **"A mathematical theory of communication".** In: *The Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

K-Nearest Neighbors (1951)

- **Evelyn Fix** (1904–1965) and **Joseph Hodges Jr.** (1922–2000)
- Non-parametric method for classification and regression[13]
- Shift from parametric to **non-parametric methods**
- Intuitive: similar objects likely belong to the same class

[13] E. Fix and J. L. Hodges (1989). ***Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties.*** Tech. rep. 3, pp. 238–247. URL: http://www.jstor.org/stable/1403797.

Rosenblatt's perceptron (1960s)

- **Frank Rosenblatt** (1928–1971), psychologist
- First model of a **learning machine**[14]
- Could learn simple tasks (AND, OR)
- Foundation of artificial neural networks
- Minsky & Papert (1969): limited to linearly separable problems
- Contributed to the first AI winter

[14] F. Rosenblatt (1958). **"The perceptron: a probabilistic model for information storage and organization in the brain."**. In: *Psychological review* 65.6, p. 386.

Hunt's inducing trees (1966)

- **Earl Hunt** — inducing decision trees from data
- Based on information entropy
- Precursor of Quinlan's ID3 algorithm[15]
- Intuitive, interpretable, symbolic rules

[15] E. B. Hunt, J. Marin, and P. J. Stone (1966). ***Experiments in Induction***. New York, NY, USA: Academic Press; J. R. Quinlan (1986). **"Induction of Decision Trees"**. In: *Machine Learning* 1, pp. 81–106. URL: `https://api.semanticscholar.org/CorpusID:13252401`.

Tukey's exploratory data analysis (1977)

- **John W. Tukey** (1915–2000)
- Countered dominance of confirmatory statistical methods
- Advocated informal study of data before formal modeling
- Introduced box plot, stem-and-leaf display
- Philosophy: data analysis should include open-ended exploration[16]

[16] J. W. Tukey (1977). ***Exploratory data analysis.*** Pearson, p. 712. ISBN: 978-0201076165.

Empirical risk minimization (1960s–1980s)

- **Vladimir Vapnik** (born 1936) and **Alexey Chervonenkis** (1938–2014)
- VC entropy and VC dimension (1968)[17]
- ERM principle — foundation of statistical learning theory
- Theoretical framework for understanding learning from data

[17] V. Vapnik and A. Chervonenkis (1968). **"On the uniform convergence of relative frequencies of events to their probabilities".** In: *Doklady Akademii Nauk USSR*. vol. 181. 4, pp. 781–787.

Resurgence of neural networks (1986)

- **Backpropagation** developed independently by several researchers[18]
- Training of multi-layer neural networks
- Solves nonlinearly separable problems
- Fueled by availability of data and computational power
- Preference for simple algorithms and intuitive models

[18] Y. Le Cun (1986). **"Learning Process in an Asymmetric Threshold Network".** In: *Disordered Systems and Biological Organization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 233–240. ISBN: 978-3-642-82657-3; D. E. Rumelhart, G. E. Hinton, and R. J. Williams (1986). **"Learning representations by back-propagating errors".** In: *Nature* 323.6088, pp. 533–536. DOI: 10.1038/323533a0.

Ensembles (1990s)

- Combine multiple learning machines to improve performance
- **Boosting** — sequential, correcting previous errors[19]
- **Bagging** — parallel, trained with small data variations[20]
- Random forests (Ho, 1995) — bagging[21]
- XGBoost (Friedman, 2001) — gradient boosting[22]

[19] R. E. Schapire (1990). **"The strength of weak learnability".** In: *Machine Learning* 5.2, pp. 197–227. DOI: 10.1007/BF00116037.

[20] L. Breiman (1996). **"Bagging predictors".** In: *Machine Learning* 24.2, pp. 123–140. DOI: 10.1007/BF00058655.

[21] T. K. Ho (1995). **"Random decision forests".** In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1, 278–282 vol.1. DOI: 10.1109/ICDAR.1995.598994.

[22] J. H. Friedman (2001). **"Greedy function approximation: A gradient boosting machine.".** In: *The Annals of Statistics* 29.5, pp. 1189–1232. DOI: 10.1214/aos/1013203451.

Support vector machines (1995)

- **Cortes & Vapnik**[23] — based on VC theory and ERM principle
- Finds optimal separating hyperplane with maximum margins
- Uses Cover's theorem[24] (high-dimensional mapping)
- Practical and efficient learning machines

[23] C. Cortes and V. N. Vapnik (1995). **"Support-vector networks".** In: *Machine Learning* 20.3, pp. 273–297. DOI: 10.1007/BF00994018.

[24] T. M. Cover (1965). **"Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition".** In: *IEEE Transactions on Electronic Computers* EC-14.3, pp. 326–334. DOI: 10.1109/PGEC.1965.264137.

Deep learning (late 2000s)

- Neural networks with multiple layers
- State-of-the-art in computer vision and NLP
- **Bengio, Hinton, LeCun** — 2018 Turing Award[25]

[25] I. Goodfellow, Y. Bengio, and A. Courville (2016). ***Deep Learning***. http://www.deeplearningbook.org. MIT Press.

LUSI learning theory (2010s)

- **Vapnik**[26] — Learning Using Statistical Invariants
- Extension of statistical learning theory
- Based on properties preserved under transformations
- “Complete statistical theory of learning”

[26] V. N. Vapnik and R. Izmailov (2015). **“Learning with Intelligent Teacher: Similarity Control and Knowledge Transfer”.** In: *Statistical Learning and Data Sciences*. Ed. by A. Gammerman, V. Vovk, and H. Papadopoulos. Cham: Springer International Publishing, pp. 3–32. ISBN: 978-3-319-17091-6.

Large Language Models (2017–)

- **Vaswani et al.**[27] (2017) — transformer architecture
- Attention mechanisms replace recurrence and convolution
- Enabled development of LLMs
- OpenAI's ChatGPT (2022) — wide public availability
- Capable of classification and reasoning without task-specific training

[27] A. Vaswani et al. (2017). **"Attention is all you need".** In: *Advances in neural information processing systems* 30.

Takeaways

- We have evolved both in theory and application of data-driven sciences
- No consensus on the definition of data science
- Sufficient evidence to support data science as a distinct science

Questions?