

Data Science sous R

Introduction to Data Science

Alessandro Leite

October 10, 2024

1 Introduction

2 Data science and the role of data and algorithms in today's world

3 Data mining

4 Data science life cycle

5 Impact of data science

WHO AM I ?!

- ▶ **Lecturer:** Alessandro Leite, PhD
 - ▶ Researcher @ Inria-Saclay, LISN, Paris-Saclay University
- ▶ **How to reach me:**
 - ▶ Email
 - ▶ Microsoft Teams course space

- ▶ Use command-line tools in a Linux or Windows environment

- ▶ Use command-line tools in a Linux or Windows environment
- ▶ Statistics: linear regression, interpretation of a hypothesis test, interpretation of a plot

- ▶ Use command-line tools in a Linux or Windows environment
- ▶ Statistics: linear regression, interpretation of a hypothesis test, interpretation of a plot
- ▶ If you lack some of them, it's your responsibility to make sure you fix it: we'll provide some pointers to help

- ▶ Lectures are compulsory

What you need to do

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position

What you need to do

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position
- ▶ We learn data science by doing

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position
- ▶ We learn data science by doing
- ▶ What you do have to do:

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position
- ▶ We learn data science by doing
- ▶ What you do have to do:
 - ▶ Assignments

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position
- ▶ We learn data science by doing
- ▶ What you do have to do:
 - ▶ Assignments
 - ▶ Practical evaluation

- ▶ Lectures are compulsory
- ▶ You should embrace an active learning position
- ▶ We learn data science by doing
- ▶ What you do have to do:
 - ▶ Assignments
 - ▶ Practical evaluation
 - ▶ Mini-project

An illustration on a dark blue background with diagonal lines of white binary code (0s and 1s). In the foreground, two stylized figures are climbing a bar chart. The chart has four bars of increasing height, colored from left to right: light blue, red, dark red, and dark blue. The first figure, wearing a blue suit, is on the first bar. The second figure, wearing a blue jacket and red pants, is on the second bar, reaching up. A white line graph with three circular markers is overlaid on the first two bars. The background below the chart is a lighter blue with pixelated, cloud-like shapes.

The data deluge

1 Introduction

2 **Data science and the role of data and algorithms in today's world**

- Veridical Data Science
- Evaluating and building trustworthiness through critical thinking
- Case study: 2019-20 Australia bushfires
- Evaluating and building trustworthiness through the PCS framework

3 Data mining

4 Data science life cycle

5 Impact of data science

The practice of data science has a long history

- ▶ The practice of data science—*using data to answer real-world domain questions*—has existed under different names, including “**data analysis**” and “**applied statistics**”.
- ▶ Modern data processing started with the **punch card created** by Herman Hollerith (1860–1929) to help process the overwhelming amount of data from the **1890 US Census**.
- ▶ Punch cards were fundamentally an early version of the binary data storage format still used in today’s computers.
- ▶ Hollerith’s machine led to the development of one of the four companies that later combined to become today’s IBM.
- ▶ Check Donoho’s article “50 Years of Data Science¹” to learn more about the history and formation of data science.
- ▶ The trick is not mastering programming but rather learning to think about data operations.

¹David Donoho. “50 years of data science”. In: *Journal of Computational and Graphical Statistics* 26.4 (2017), pp. 745–766.

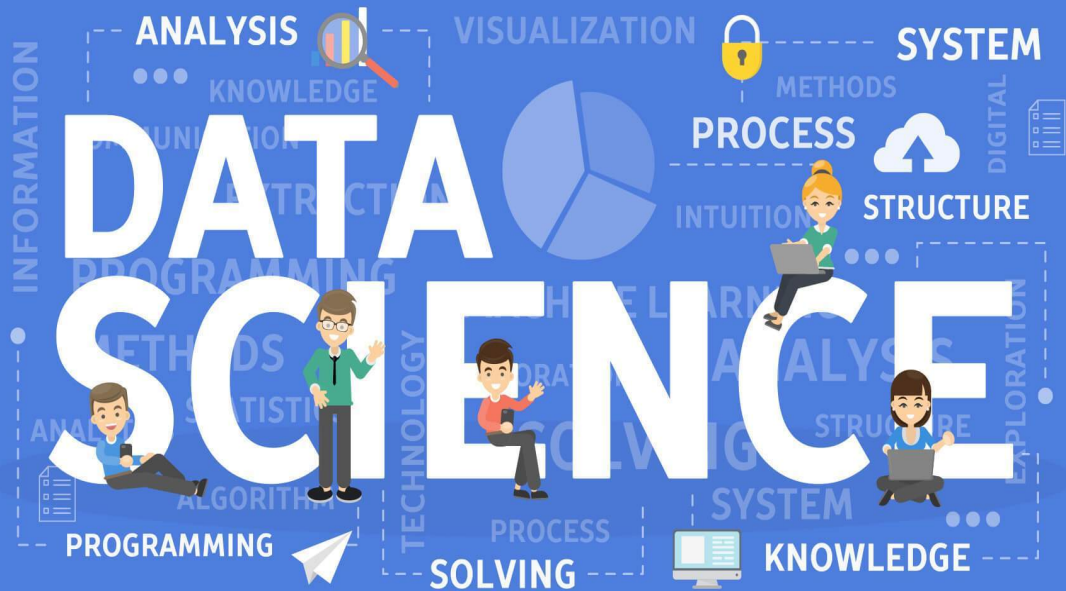
- ▶ Manufacturing, retail, healthcare, agriculture, and other sectors are using data-driven insights to respond to changes in markets, improve decision-making, and operate more efficiently
- ▶ Many groups cannot take advantage because they **lack** the **skilled staff** and resources needed
- ▶ There are several challenges in filling the workforce gap between academia and industry
- ▶ Often, the **data collected by organizations are simply not useful**, or the **data analytics skills** needed to use the data are **unavailable**

*"It's hard to overstate the need for people with data science acumen and some data-specific skills. ... The best talents will go to the best tech-oriented companies, so it's hard for small and mid-size companies to recruit and retain the best and brightest." - Mark Daniel Ward, a professor at Purdue and director of The Data Mine*²

² beta.nsf.gov/science-matters/developing-21st-century-data-science-workforce



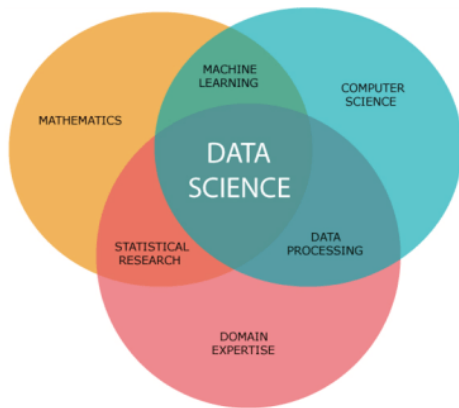
What is Data Science?





**Is Data Science an old
wine in a new bottle?**

What is data science?



- ▶ **Data science** lies at the intersection of:
 - ▶ **Computer science**: the development and use of computational techniques to solve problems and answer questions
 - ▶ **Statistics/mathematics**: the development and use of mathematical and statistical techniques to solve problems and answer questions
 - ▶ **Domain expertise**: knowledge of background information that underlies the problem or question being asked.

What is data science?

Data science unifies statistics, data analysis, computer science, and their related methods to understand and analyze actual phenomena with data^a

^aCao, “Data science: a comprehensive overview”; Donoho, “50 years of data science”.

- ▶ A **data scientist** is a professional who writes code and combines it with statistical knowledge to create insights from data and then communicate their findings to a range of audiences to enhance understanding and inform domain decision-making.

What is data science?

Data science unifies statistics, data analysis, computer science, and their related methods to understand and analyze actual phenomena with data^a

^aCao, “Data science: a comprehensive overview”; Donoho, “50 years of data science”.

- ▶ A **data scientist** is a professional who writes code and combines it with statistical knowledge to create insights from data and then communicate their findings to a range of audiences to enhance understanding and inform domain decision-making.
- ▶ Consider a **project** that aims to **predict** which hospital **patients** are at **high risk** of **readmission** using information from **electronic medical records**.

Data science unifies statistics, data analysis, computer science, and their related methods to understand and analyze actual phenomena with data^a

^aCao, “Data science: a comprehensive overview”; Donoho, “50 years of data science”.

- ▶ A **data scientist** is a professional who writes code and combines it with statistical knowledge to create insights from data and then communicate their findings to a range of audiences to enhance understanding and inform domain decision-making.
- ▶ Consider a **project** that aims to **predict** which hospital **patients** are at **high risk** of **readmission** using information from **electronic medical records**.
- ▶ The project's overall **goal** is to extract **readmission information** from **electronic medical records** and **communicate** this information to the doctors.

Data science unifies statistics, data analysis, computer science, and their related methods to understand and analyze actual phenomena with data^a

^aCao, “Data science: a comprehensive overview”; Donoho, “50 years of data science”.

- ▶ A **data scientist** is a professional who writes code and combines it with statistical knowledge to create insights from data and then communicate their findings to a range of audiences to enhance understanding and inform domain decision-making.
- ▶ Consider a **project** that aims to **predict** which hospital **patients** are at **high risk** of **readmission** using information from **electronic medical records**.
- ▶ The project's overall **goal** is to extract **readmission information** from **electronic medical records** and **communicate** this information to the doctors.
- ▶ Then, medical **doctors decide whom to discharge**.

What is data science?

Data science unifies statistics, data analysis, computer science, and their related methods to understand and analyze actual phenomena with data^a

^aCao, “Data science: a comprehensive overview”; Donoho, “50 years of data science”.

- ▶ A **data scientist** is a professional who writes code and combines it with statistical knowledge to create insights from data and then communicate their findings to a range of audiences to enhance understanding and inform domain decision-making.
- ▶ Consider a **project** that aims to **predict** which hospital **patients** are at **high risk** of **readmission** using information from **electronic medical records**.
- ▶ The project's overall **goal** is to extract **readmission information** from **electronic medical records** and **communicate** this information to the doctors.
- ▶ Then, medical **doctors decide whom to discharge**.
- ▶ Consequently, the project's **goal** is **not** to **extract readmission** information from electronic medical records.

What is data science used for?

1 Descriptive analysis:

- ▶ gain **insights** into **what happened** or what is happening
- ▶ characterized by **data visualizations** including bar charts, line graphs, and tables

2 Diagnostic analysis

- ▶ detailed data study to **understand why** something happened
- ▶ characterized by techniques such as **drill-down**, **data discovery**, **data mining**, and **correlations**

3 Predictive analysis

- ▶ uses historical data to **forecast** data patterns
- ▶ characterized by techniques such as **machine learning**, **pattern matching**, and **predictive modeling**

4 Prescriptive analysis

- ▶ **predicts** and **suggests** an optimum response for the **predicted outcome**
- ▶ characterized by the usage of **graph analysis**, **simulation**, **complex event processing**, **neural networks**, and **recommendation engines**.

Data science is more than the intersection of different domains

- ▶ It **integrates statistical** and **computational thinking** into **real-world domain problems** in science, technology, and beyond.
- ▶ Data science projects are grounded in real-world problems.
- ▶ Data science involves programming, but it is **more** than just **programming**.
- ▶ It is critical that data scientists work **side-by-side** with domain experts to ensure that their data-driven **results** provide useful, ethical, and trustworthy solutions to real-world **domain problems**.
- ▶ Most data science projects aim to use the insights gained from data analyses to help us make decisions in the real world.
 - ▶ Demographers conduct and analyze surveys on members of a human population of interest to help guide policy decisions
 - ▶ Ecologists observe and analyze data on animal and plant distributions and behaviors to help make conservation decisions
 - ▶ Financial analysts evaluate financial market trends to help make financial decisions

Real-world data science projects are challenging

- ▶ Real-world data science projects usually start with a vague domain question
- ▶ You must answer with a messy dataset riddled with ambiguities and errors (and that was most certainly not collected with your particular project in mind)
- ▶ It involves embarking on a winding journey of analyses whose underlying assumptions don't quite fit the data you have
- ▶ In the end, what you usually have to show for your months (or years) of hard work is a set of depressingly inconclusive results.

Veridical Data Science

👍 “Veridical data science is the practice of conducting data analysis while making human judgment calls and using domain knowledge to extract and communicate useful and trustworthy information from data to solve a real-world domain problem^a.”

^aBin Yu and Rebecca L Barter. *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press, 2024.

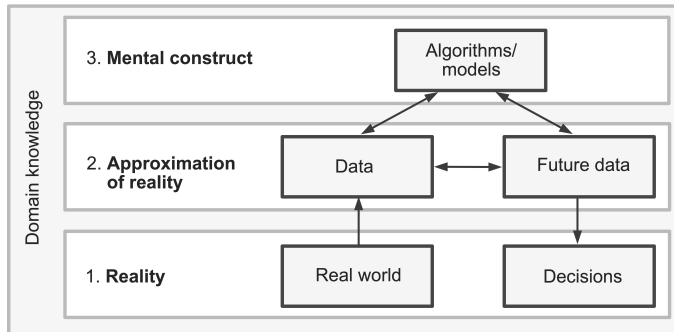



Figure 1: The three realms: reality, an approximation of reality (where our data lives), and the mental construct realm (where our analyses and algorithms live). Source: [Yu and Barter, 2024]

- ▶ **Critical thinking** skills and a **domain knowledge** are fundamental for pointing out potential **weaknesses** in the **data analysis** and **detecting biases** that might sneak into the data collection and analysis pipeline.
- ▶ A data scientist well versed in critical thinking and domain knowledge, possessing a reasonable understanding of the algorithms being used, and embracing the **predictability, computability**, and **stability (PCS)** principles is less likely to fall into the trap of unintentionally lying with data³.

³Yu and Barter, *Veridical data science: The practice of responsible data analysis and decision making*. 

Question to catch inconsistencies and errors into data analyses

- ▶ Questions about the domain problem:
 - ▶ Is the question being framed the question you really care about?
 - ▶ Is the question ethical?
- ▶ Questions about the data:
 - ▶ Where did the data come from?
 - ▶ Is the source of the data trustworthy?
 - ▶ Is the data relevant to the question being asked and any future setting to which the results might be applied?
 - ▶ What do the values within the data correspond to in the real world?
- ▶ Questions about the analysis, algorithms, and results:
 - ▶ What do the results mean in the context of reality?
 - ▶ Do the analysis results make sense?
 - ▶ Do the analysis and algorithms involve making any assumptions about the data or domain?
 - ▶ Are the assumptions reasonable?
 - ▶ Do results answer the primary project question?
 - ▶ Are the results being communicated accurately?

 Whenever possible, do the “shoe leather” work⁴.

⁴Freedman, “Statistical models and shoe leather”.

Confirmation bias

- ▶ It occurs when we fail to critically scrutinize results because they match what we expect to find.
- ▶ As human beings, we are a lot more likely to double-check and scrutinize our results when they don't match our expectations.

Desirability bias

- ▶ It occurs when we try to find the results that we think other people expect us to find.

⚠ Although it is impossible to diminish the risks of confirmation and desirability bias, we can develop a **healthy skepticism** of every result we obtain, regardless of whether it matches what we or other people expected or wanted to find.

- ▶ 2019–20 Australian bushfire season ravaged over 42 million acres of the east coast for many months during late 2019 and 2020⁵.
- ▶ Many scientists, including Professor Chris Dickman at the University of Sydney, estimated that **“over 1 billion animals perished in the fires.”**
- ▶ Let us try to understand where this 1 billion number might have come from.

⁵ tinyurl.com/sav4vzj5

Asking questions about the data

- ▶ How do you imagine that the data that underlies the statement that “1 billion animals perished in the fires” were collected?
 - ▶ Was there a team of scientists manually counting every single animal that died?
 - ▶ Does the term “animal” include insects, such as ants and flies?
- ▶ A text from the University of Sydney states that Professor Dickman based his estimations on a detailed study by [Johnson et al., 2007] also coauthored by Professor Dickman.

“The figures quoted by Professor Dickman are based on a 2007 report for the World Wide Fund for Nature (WWF) on the impacts of land clearing on Australian wildlife in New South Wales⁶.”

“The figure includes mammals (excluding bats), birds, and reptiles and does not include frogs, insects, or other invertebrates.”

- ▶ The data that underlie this study is relevant, but there is no doubt that it is limited. It is clear that Professor Dickman did his best with an existing dataset that he was already familiar with since no other one was available.
- ▶ At the end of the day, the best data is just the data you have.

⁶University of Sydney. *More Than One Billion Animals Impacted In Australian Bushfires*. tinyurl.com/2usszt6x. 2020.

- ▶ Considering that the data were based only on a small localized region, how did Professor Dickman and his collaborators use the data to approximate the number of animals that died across the entire affected region?
“The authors of that report obtained estimates of mammal, bird, and reptile population density in NSW and then multiplied the density estimates by the areas of vegetation approved to be cleared [Sydney, 2020].”
- ▶ This strategy ignores the differences in animal densities across the entire affected area.
- ▶ And it also ignores how these densities may have changed over time.
- ▶ Given that no additional data were available, what else could have been done?
- 👍 ▶ When criticizing someone else’s study, keep in mind that behind every data-driven conclusion, there is a well-intentioned human being with feelings.
- ▶ In the case of Professor Dickman’s study, it was the **first high-profile study** that even attempted to quantify the impact of the fires on animal populations.

PCS framework

- ▶ The predictability, computability, and stability (PCS) framework provides guidelines for empirically evaluating the trustworthiness of evidence obtained from data at every step of the data science life cycle.
- ▶ Implementing the PCS principles involves using computational explorations to ensure that the data analysis results are predictable and stable^a

^a[Bin Yu and Rebecca L Barter](#). *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press, 2024.

- ▶ The three PCS framework comprises three principles:
 - ▶ **Predictability**
 - ▶ **Computability**
 - ▶ **Stability**
- ▶ Predictability and stability assessments provide evidence for the trustworthiness of our data-driven results.
- ▶ They don't enable us to prove that our data-driven results are trustworthy.
- ▶ Absolute proof requires completely exhaustive predictability and stability assessments that consider all possible alternative data and analytic scenarios, and it is an impossible task.
- ▶ They are more closed to “Popperian falsification” principle as defined by the philosopher Karl Popper in the 1930s⁷.

⁷Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

Predictability

- ▶ Data-driven results are predictable if they can be shown to reemerge in new, relevant scenarios.
- ▶ Strong evidence of predictability can be obtained by demonstrating that the results hold in the context of the actual future or external data to which you will apply your results.
- ▶ In the absence of available future/external data, a validation set surrogate can be used to evaluate the results.
- ▶ In the absence of available future/external data, the recommended technique for generating surrogate external/future datasets is to split your data into a training set and a validation set.

The roles of the training, validation, and test sets

- ▶ When we cannot access future/external data for predictability evaluation, we can **split** your data into training or test or training, validation, and test sets.
- ▶ The **training set** is used for **exploring underlying patterns** and relationships in the data and **training algorithms**
- ▶ The **validation set** is used to evaluate the results and algorithms by serving as a **surrogate** for future data.
- ▶ The **test set** serves as a final surrogate for future data
- ▶ The data should be split in such a way (e.g., using a random, group-based, or time-based split) that the **relationship** between the training data and the testing and validation data **reflects** the relationship between the **current** and **future data** to which the results and algorithms will be applied.
- ▶ It is strongly recommended to **split** the data into training, validation, and test datasets **before** start exploring and analyzing the data to prevent **data leakage**⁸.

⁸Shachar Kaufman et al. "Leakage in data mining: Formulation, detection, and avoidance". In: *ACM Transactions on Knowledge Discovery from Data* 6.4 (2012), pp. 1–21.

Time-based split

- ▶ For time series data sets, we cannot randomly split the observations into train, validation, and test sets
- ▶ The measures are not exchangeable.
- ▶ In this case, you can use the earliest 60 percent of the data as the training set and split the later 40 percent of the data into validation and test sets.
- ▶ For instance, imagine we have 10 years of stock market data.
- ▶ We can use the stock prices from the first 6 years as a training set and then use stock prices from the final 4 years for the validation and test datasets for evaluation.
- ▶ After finishing our evaluations, we can re-train the algorithm using all of the available data to ensure that the model captures the most up-to-date trends.

Group-based split

- ▶ For data having natural grouping information (e.g., a dataset with information from patients from several different hospitals), and if we want to train to a new group of data points (e.g., to patients in a different hospital), then we have
 - ▶ to choose a random subset of 60 percent of the groups (e.g., hospitals, rather than patients) to form the training dataset
 - ▶ to split the remaining groups (e.g., hospitals) between the validation and test datasets.
- ▶ This ensures that every data point from the same group will be contained entirely within the training dataset, the validation dataset, or the test dataset rather than split across them and best reflects the process of applying the algorithm to data points in a new group (e.g., patients in a new hospital).

Random split

- ▶ Use a purely random split to form the validation and test datasets when the data points are more or less exchangeable
- ▶ For example, when people are randomly included in a survey, the new data points to which the results will be applied are similar to the current data points.

Stability

- ▶ Data-driven results are stable if they tend not to change across reasonable alternative perturbations throughout the data science life cycle.
- ▶ Stability analysis aims to explore many relevant **sources** of the **uncertainties** that are associated with the results.
- ▶ **Uncertainty** refers to how the results might have looked different under various scenarios.
- ▶ The common sources of uncertainty include the data collection process, the data cleaning and pre-processing tasks, and the choice of used algorithm.
- ▶ They are commonly named **aleatoric** and **epistemic** uncertainty.
- ▶ The major goal is thus to assess the stability of the results across reasonable perturbations to the data collection process, our own data cleaning and pre-processing judgment calls, and our algorithmic choices.

- 1 Introduction
- 2 Data science and the role of data and algorithms in today's world
- 3 Data mining**
- 4 Data science life cycle
- 5 Impact of data science

The knowledge discovery in databases (KDD) strategy

► CRoss-Industry Standard Process for Data Mining (CRISP-DM)

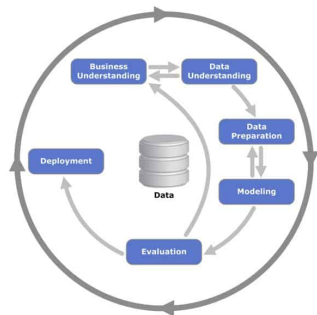
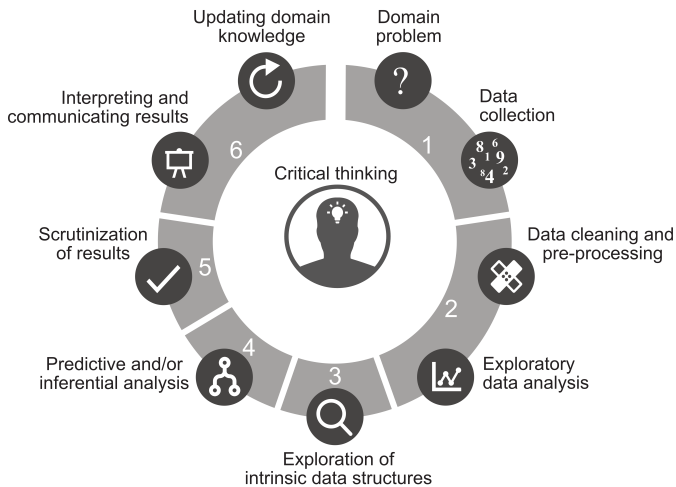


Figure 2: CRISP-DM Life Cycle

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Select Data <i>Rationale for Inclusion/Exclusion</i>	Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Clean Data <i>Data Cleaning Report</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goals <i>Data Mining Goals Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Construct Data <i>Derived Attributes Generated Records</i>	Build Model <i>Parameter Settings Models Model Descriptions</i>	Determine Next Steps <i>List of Possible Actions Decision</i>	Produce Final Report <i>Final Report Final Presentation</i>
Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Integrate Data <i>Merged Data</i>	Assess Model <i>Model Assessment Revised Parameter Settings</i>		Review Project Experience <i>Documentation</i>
		Format Data <i>Reformatted Data Dataset Dataset Description</i>			

Figure 3: CRISP-DM **tasks** in **bold**, and **outcomes** in *italic* [Ncr et al., 1999]

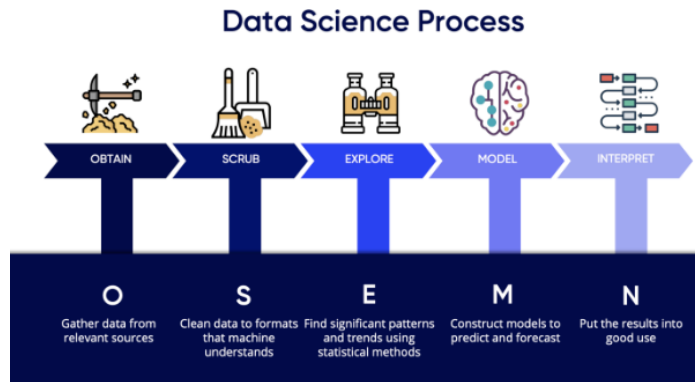
- 1 Introduction
- 2 Data science and the role of data and algorithms in today's world
- 3 Data mining
- 4 Data science life cycle**
- 5 Impact of data science



- 1** Identify a domain problem, formulate a relevant question, and collect data.
- 2** Cleans, pre-processes, explores the data
- 3** (Optional) Explores the intrinsic structure of the data
- 4** Trains a machine learning algorithm or conducts inferential analysis
- 5** Evaluates the results
- 6** Communicates the findings and updates the domain knowledge with what was learned.

Figure 4: Stages of a data science project. Source: [Yu and Barter, 2024]

- ▶ The data science life cycle (DSLC) describes the non-linear problem-specific path every data science project takes.

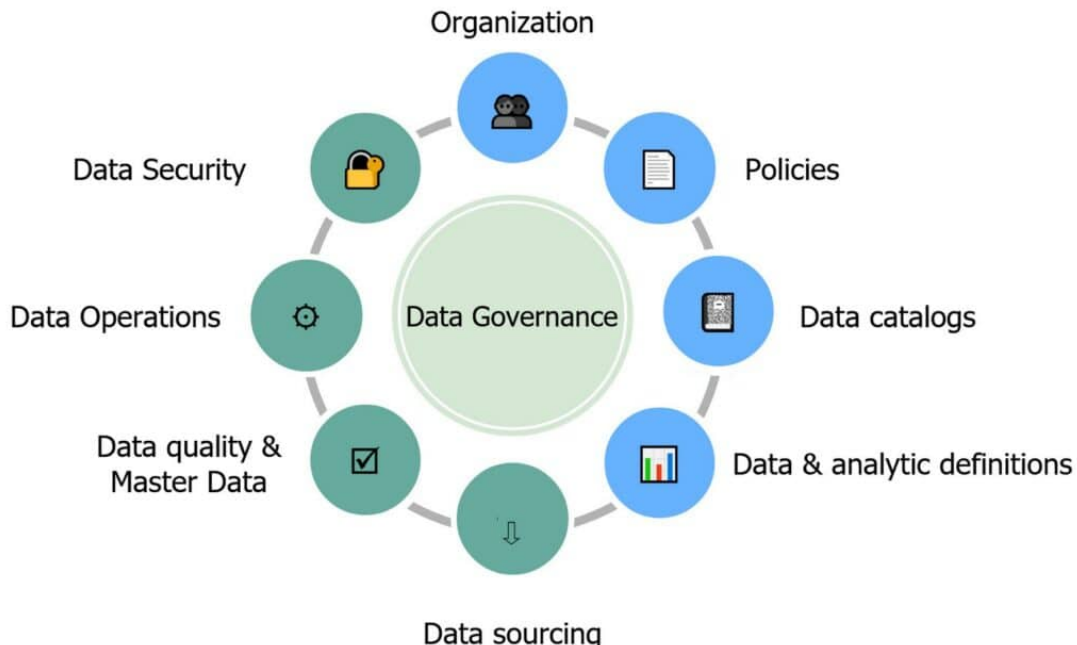


- ▶ This multi-stage process can be attributed to Box⁹ and Cox and Snell¹⁰.

⁹Box, "Science and statistics".

¹⁰Cox and Snell, *Applied Statistics-Principles and Examples*.

Data governance cares for the data and its subjects



Data Science Roles & How They Interact

Enable data access & utilization & enable value capture

Builds and supports the infrastructure or 'data pipe' and all associated SW engineering infrastructure tasks.

Data Engineer

Core skills:

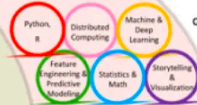


Collect, manage, analyze & visualize data

Build, deploy data infrastructure & architecture

Data Scientist

Core skills:



Ascribes value to raw data through original interpretation & modeling

Use sophisticated methods to interrogate the data

Optimize & enable data for business & functional value capture & value creation

Analysis and interpretation of complex digital data to extract or discover knowledge and assist decision-making.

Hypothesis Development
Monetization
Governance

Journey Maps

Value Streams & Service Maps

Deployment & Integration

Core skills:



ROI, NPV, Domain Expertise

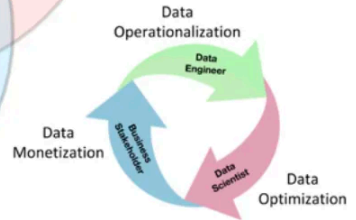
Value Chains

Financial Analysis

Business Stakeholder

Help the business make better decisions through data

Blend of business, analytic and math skills to explore and solve challenges, bridging the data and business communities.



- 1 Introduction
- 2 Data science and the role of data and algorithms in today's world
- 3 Data mining
- 4 Data science life cycle
- 5 Impact of data science**

- ▶ Fields like Physics, Bioinformatics, and Earth Science used big data anyway
 - ▶ they had their own independent data science revolution
- ▶ In other areas, the profile of data-driven science spurred governments to develop cross-disciplinary programs
 - ▶ Alan Turing Institute for data science in the UK
 - ▶ DataIA Institute (dataia.eu) in France
- ▶ Has provided new data sources and tools for collecting data
 - ▶ crowdsourcing
 - ▶ social media
- ▶ Enables citizen to be an actor in science
 - ▶ DataONE (dataone.org)

- ▶ Data, Predictions, and Decisions in Support of People and Society (bit.ly/3ptb9di)
- ▶ AID DATA (aiddata.org): making finance data more accessible
- ▶ Mechanism Design for Social Good (md4sg.com)



- Data science projects fail usually when:
- 1 mismatch between business and technical views
 - 2 Unclear goals
 - 3 Poor data quality
 - 4 Incorrect data processing
 - 5 Usage of not appropriate techniques

- ▶ George E.P. Box. “Science and statistics”. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799
- ▶ David Roxbee Cox and E. J. Snell. *Applied Statistics-Principles and Examples*. CRC Press, 1971
- ▶ C Johnson et al. *Impacts of landclearing; the impacts of approved clearing of native vegetation on Australian wildlife in New South Wales*. Tech. rep. World Wide Fund for Nature, 2007
- ▶ Tony Hey, Stewart Tansley, and Kristin Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009
- ▶ Thomas H Davenport and DJ Patil. “Data scientist”. In: *Harvard business review* 90.5 (2012), pp. 70–76
- ▶ Nate Silver. *The signal and the noise: why so many predictions fail—but some don’t*. Penguin, 2012
- ▶ Longbing Cao. “Data science: a comprehensive overview”. In: *ACM Computing Surveys* 50.3 (2017), pp. 1–42
- ▶ Wil MP Van der Aalst. “Data scientist: the engineer of the future”. In: *Enterprise interoperability VI*. 2014, pp. 13–26
- ▶ Bin Yu and Rebecca L Barter. *Veridical data science: The practice of responsible data analysis and decision making*. MIT Press, 2024