DATA SCIENCE OVERVIEW

Data Science Tools and Methods

Marcus Birkenkrahe

July 27, 2022

Contents

- You are like Pythagoras (570-495 B.C.)
- He saw nature as a structured system of numbers
- Data science is (still) the "sexiest" job of the century

Figure 1: Modern picture of Pythagoras of Samos (ca. 570-495 BC)

1 WHAT WILL YOU LEARN?

- How and why data science is so popular
- What skills you need to do data science
- Which problems data science can solve
- What data scientists do all the time

2 DATA SCIENCE AS A FIELD OF SCIENTIFIC STUDY

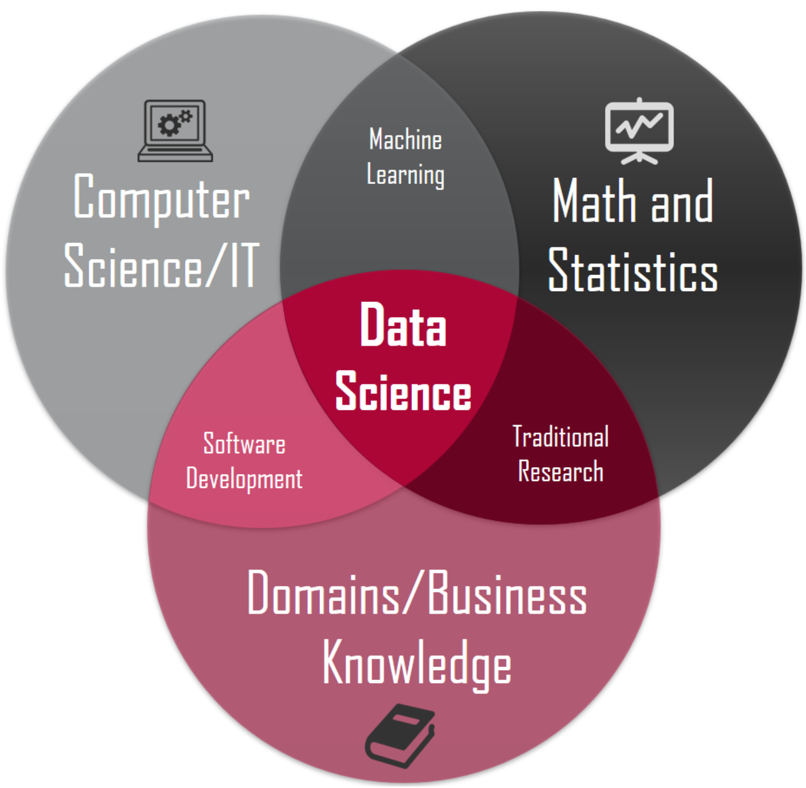


Figure 2: Data science as an interdisciplinary field (Source: Barber, 2018)

- Vast, new, hard to define, highly interdisciplinary field

- 'Data science' is the science of "that what is given" (data)
- My definition: **data + code + stats = story**

  In this lesson, I am giving an oveRview[1] of several aspects of data science. Though young as a declared field (2012), it is a field both ill-defined and (or perhaps because of it) vast and hard to pin down. This outline will be applied rather than scholarly, focusing on applications and practice rather than concepts or theory.

  In its name, "data science" carries both aspects of science and craft: the 'science' part is responsible for the modus operandi, which is informed by statistics and math, systematic and logically rigorous. The 'data' part relates mostly to craft: the ability to extract insights from data using computing tools. Most data scientists are more occupied by and with the craft part than with the science part (cp. Kozyrkov 2018).

  Hence, data science so far is a typical support science. It supports other, more established disciplines in the natural and in the social sciences. Prominent examples are: economics, genomics, and epidemiology.

  The need to use the data "to tell a story" sets data science apart from both traditional data craft and science. It is the reason why visualization techniques and theory ("grammar of graphics", cp. Sarkar 2018) play such an important role.

  I would argue that data science is most successful when supporting fields that themselves are interdisciplinary and therefore need a higher degree of communication across different cultures of science and practice. This is the quasi-definition that I came up with while preparing these notes:

  [RAW] **DATA** + [LITERATE] **CODE** + [APPLIED] **Stats** = [DECISION] **STORY**

  Why? Because data always come in "raw" form and have to be wrestled with. To do this, you need to be able to code (a little anyway). But in order to achieve the main goal, namely add value, process-oriented science has to come in, most importantly through systematic methods and the accompanying processes, to validate insights and help communicate results. Well, so far, so good.

[1] "oveRview" is a joke, not a mistake. Capitalizing the letter "r" in a seemingly random fashion is a common in-joke in the R community. Adding and/or capitalizing the letter "r" is also used to name R software packages, as in: `learnr`, `magrittr`, or `fasteR`.

In the following lecture, I will focus on four aspects of data science: the popularity it currently enjoys (and has enjoyed for the past 10 years), the skills required to "do data science", and the processes or activities involved in doing it. We will look at each of these with some examples.

At the end of each chapter, you'll find a quick challenge ("youR tuRn"[2]) - this is usually just a question related to the text. Sample answers and hints to challenges are gathered at the end of the document.

3 HOW POPULAR IS DATA SCIENCE?

Figure 3: Selfie by Cristina Zaragoza (Unsplash)

- How would you try to find out how popular data science is?

4 WAYS TO EXPLORE POPULARITY

`../img/2_4th_july.gif`

- **Search** (how? where?)

[2]Norman Matloff used this title for small challenges throughout his excellent short course on R, and I have borrowed and "R-ified" it.

- Find relevant **models** (how?)
- Generate **primary** data (how?)
- Use **secondary** data (how?)

Question: Can you think of any issues with these methods?

1. Search - where? How?
   - Google (Scholar) - disadvantage of Google searches?
   - arxiv.org
   - data science blogs (R-Blogger, Towards Data Science, Analytics Vidhya, R Weekly, DataCamp)
2. Find relevant models - what is that?
   - Metaphors are models
   - Mathematical model may not exist
   - Example for such a model
3. Generate primary data
   - Which measures are used?
   - Which methods are used?
4. Look secondary data
   - public?
   - Valid?
   - How do you validate?

Example: social networking analysis - Predicting Tie Strength (2009). Paper: `https://1drv.ms/b/s!AhEvK3qWokrvqz6uRFcluk1LEOW5`

This paper uses a model to distinguish between weak and strong ties (with over 85% accuracy) based on a parametrization (= features to establish splitting the data) and a linear model (= assumption that the predictive variables are linearly correlated). Data science is used to address questions hidden in the data, such as how users relate to one another in social media, how they behave, perhaps even why they do what they do (= statistical inference).

Figure 4: "Data science" searches, 07/22 (Google trends)

5 WORLDWIDE SEARCHES 2004-2021

- 2012: Data Scientist: The Sexiest Job of the 21st Century
- 2022: Data Scientist: Still the "Sexiest Job of the 21st Century"?

What do you think has changed since 2012?

In the graph from trends.google.com, "numbers represent search interest relative to the highest point on the chart for the given region [worldwide] and time [since logging trends in 2004]." The trend increased is noticeable. It peaked in March 2022 (Source: Google Trends).

In October 2012, almost 10 years ago, Davenport and Patil published "Data Scientist: The Sexiest Job of the 21st Century" and put the term on the map.

What has changed since 2012?

1. (According to Davenport/Patil, 2022)
    - Demand in 2012 restricted to a few cities, startups, tech firms
    - Data scientists in 2012 were science PhDs, exceptional at math, who knew how to code
    - Data scientists now need to develop AI modls
    - By 2019, postings on Indeed had risen by 256%
    - Projected 15% increase from 2019 to 2029
    - Lack of "data-driven cultures" (no use for data insights)
    - Turnover is high (data scientists often don't stay long)
    - Data science is better institutionalized (= widely accepted)

- Diversification and proliferation of roles (many skills needed)
- Changes in technology (like AutoML, MLOps tools)
- Need for an ethical dimension widely acknowledged (politicized)

2. Other changes that might have affected data science:

    - COVID-19 pandemic (2020-2022)
    - Rise of cloud computing, quantum computing, deep learning
    - Political divide deepened (immigration, abortion, gun laws)

6 THE DEFINITION OF SEXY (FOR SCIENTISTS)

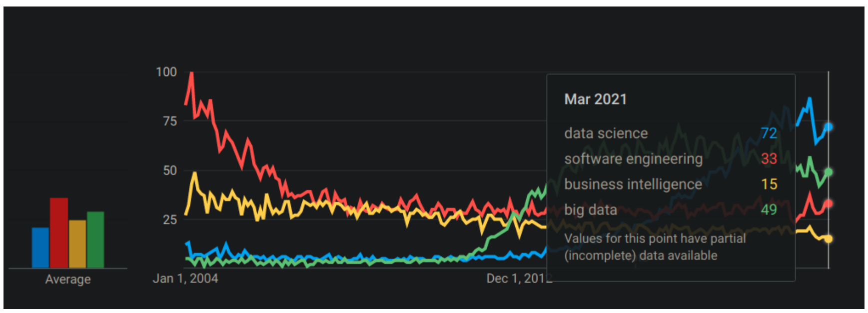
Figure 5: Richard Feynman with drums (ca 1964)

»The best data scientists are product and process innovators and sometimes, developers of new data-discovery tools. That is the definition of sexy.« -Gil Press (Forbes, 09/27/12)

7 POPULARITY CONTEST

What do you think: which of these terms is most searched?[3]

1. Big data?
2. Business intelligence?
3. Software engineering?
4. Data science?

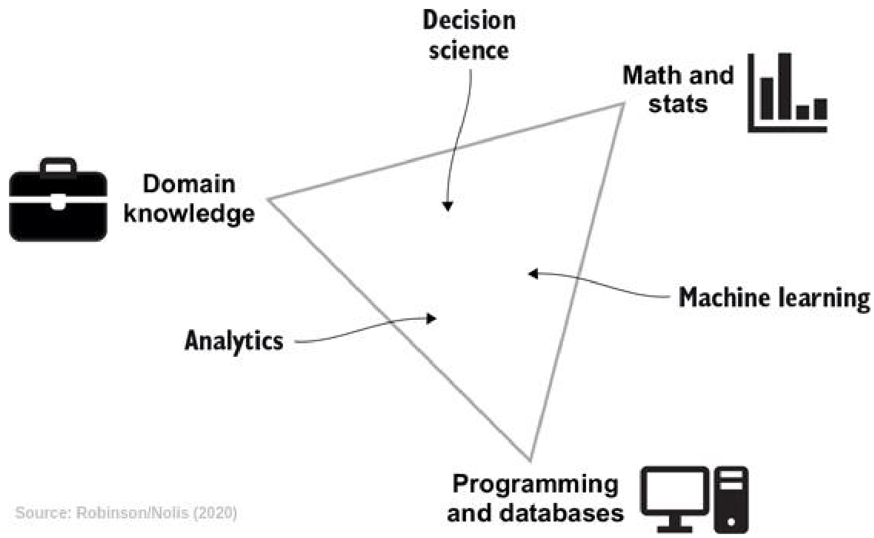


How do you like the visualization?

1. Bar chart (averages) difficult to read (percentages missing)
2. List follows the search order, not the results
3. Grid lines (vertical lines) could improve reading

[3]Notice that this list amounts to a visualization, too, because it suggests an ordering, which is supported by the language: "big data" sounds like it should be at the top. In fact, "data science" takes the top spot, and both BI and Software Engineering were a lot more popular in the past.

8 WHAT ARE DATA SCIENCE SKILLS?



Can you give some examples for any of these skills?

- What do you know for example if you have "domain knowledge"?
- Which professional activities correspond to "math and stats"?
- What kind of "programming" would you have to do?

The three skill areas in figure 8 from Robinson/Nolis (2020) give rise to different tasks and problem settings:

Skill	Sample area	Sample activity	Sample analysis
Domain knowledge	Marketing	Analyze customer data	What do customers like?
	Education	Learner data	How did students learn?
	Finance	Investment data	Which stock performed?
Coding & databases	R, Python, SQL	Analzye/automate/query	Count customers by type
	Cloud computing	Share data and code	Work in virtual teams
	RStudio, Emacs	Improve your workflow	Create a notebook[4]
	Package creation	Write new functions	Distribute package
Maths & stats	Data structure	Data wrangling	Check data tidyness
	Model building	Linear regression	Fit line graph to data
	Distribution	Check significance	Apply t-test[5]

Between two of these areas each are application areas:

1. Domain knowledge and statistics support decision science. See infographic (source: Bobriakov 2019).
2. Data analytics are the result of applying database programming (e.g. with SQL) to domain knowledge problems(this is also sometimes called 'business intelligence' or BI).
3. Programming, maths and statistics give rise to various machine learning (ML) techniques concerned in particular with prediction and pattern recognition.

9 WHAT ABOUT YOUR SKILLS? WHAT ARE THEY?

Practice: fill this Kanban board for your own skills (if any)!

- In which domain do you have knowledge?
- Which (non-trivial) decisions have you made?
- What do you know in maths and stats
- Which programming/database languages/systems do you know?
- Which process analytics tools have you used?
- What are you skills in machine learning?

  Compare: "My IT skill stack"[6]

1. Problem solving skills:
    - Understand the problem: the conditions, the unknowns, the data. Of these, I am particularly good with data.

[4] A data science notebook is a "literate programming" artifact. This concept goes back to 1984 (Knuth 1984). Today, there are plenty of commercial notebook implementations for many different programming languages (see Myers 2020 "primer").

[5] "A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features." (Source)

[6] Written in August 2020 for students of an MA international business program at the Berlin Professional School.

- Design a plan of attack (e.g. by modeling - abstracting from the details to identify one or more routes or options)
- Carry out the plan of attack: this is execution. Probably my least favorite part (often, when I see the solution path, I get bored). But I can do it, and it's satisfying to finish something.
- Look back, review and discuss your solution. I am especially good at this type of postmortem analysis - it's probably what I use most when it comes to teaching stuff.

2. Computational thinking skills
    - 10 programming languages - recommended: SQL and R
3. Data literacy skills
    - Wikipedia definition is not bad: "Ability to understand, create, and communicate data as information." (I.e. structured data)
    - Use of visualization and storytelling techniques
    - Business process modeling
4. Communication skills
    - team / leadership experience
5. Tool skills
    - I love tools
    - In my courses usually use about 20 different IT tools

10 WHAT ARE TECHNICAL DATA SCIENCE SKILLS?

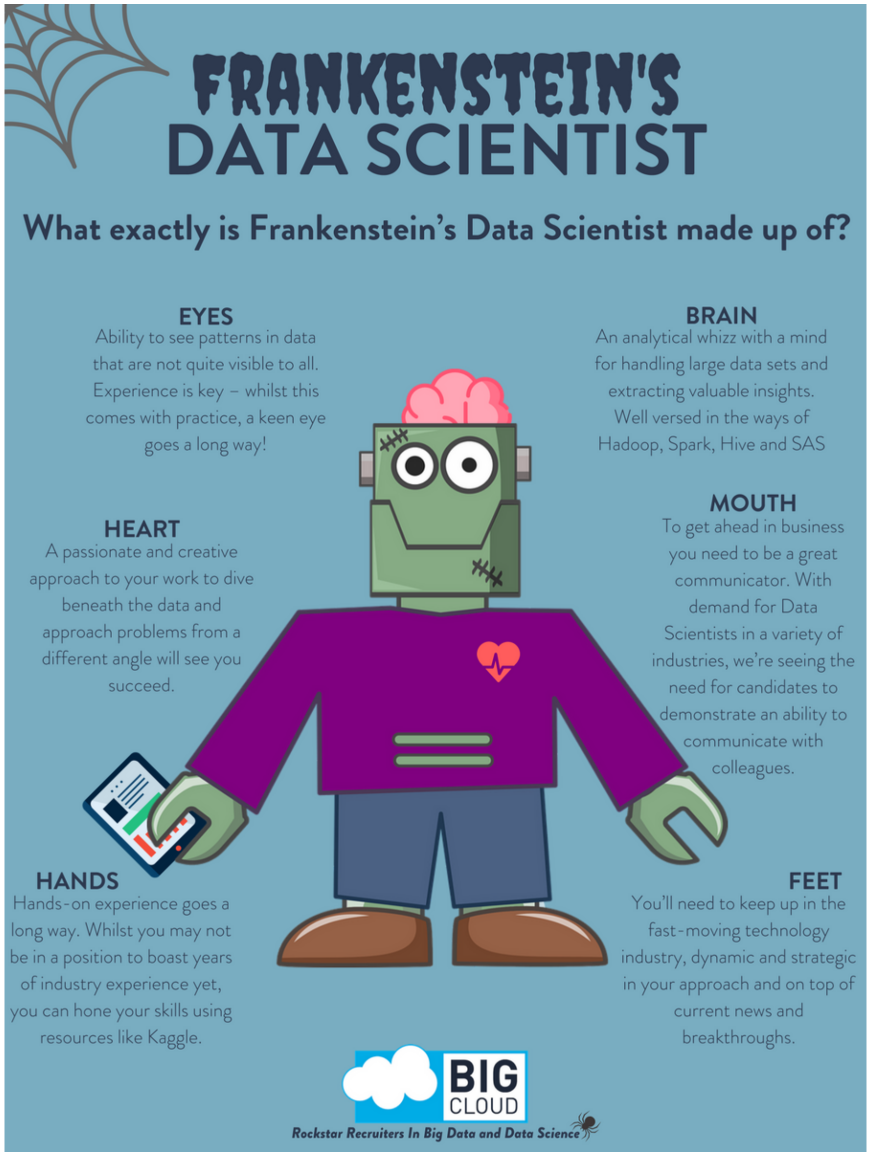
R	Apache Spark	Apache Pig
Python	NoSQL databases	Tableau
Apache Hadoop	Cloud computing	iPython notebooks
MapReduce	D3	GitHub

Have you heard of any of these?

Tip: when you come across products you don't know, make it a habit to look them up - knowing the names and what they stand for will help you anchor yourself in anything you read, and the most important products, which are most talked about, are often talked about for a reason - e.g. because they represent an innovation and/or an advantage. By knowing the products, you can also learn something about the innovation. This dependency on products also shows that both computer and data science are crafts.

TOOL	PURPOSE	TOOL	PURPOSE
D3.js	Visualization	Apache Hadoop	distributed computing
Apache Spark	Analytics engine	MapReduce	Google scalability
Apache Pig	Analytics platform	NoSQL	Unstructured big data
Tableau	Visualization	iPython nb	Literate Programming
GitHub	Version control		

11 WHAT IS "FRANKENSTEIN'S DATA SCIENTIST" MADE OF?

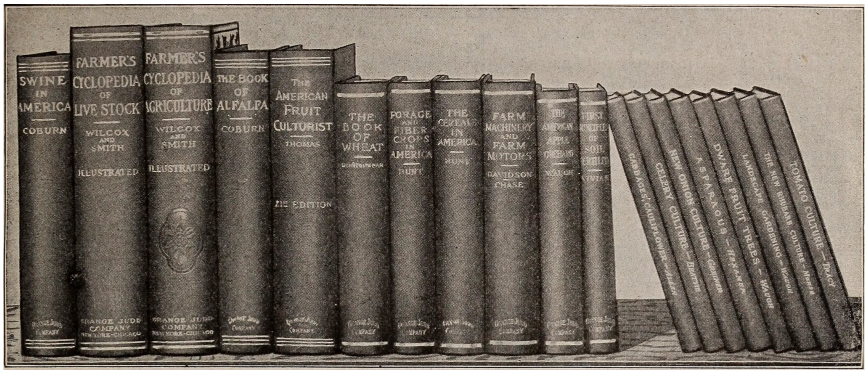


Source: datasciencecentral.com

"Frankenstein's monster" (based on the novel by "Frankenstein, or The Modern Prometheus", by Mary Shelley, 1818) is used in figure 11 as a metaphor for a working data scientist. it is a rich metaphor with many connotations.

- "Eyes": experience with detecting data patterns. to do this actually with your eyes is unlikely - you need some tools for that, but you also need experience to know which tools will work. example: `head(dataset)` only prints the first 6 rows of a dataset giving you an idea of the type of data in the dataset.
- "Heart": passion for and creativity with data. "passion" is perhaps more relevant for the data's origin and for what you can do with well interpreted data - namely change the world! example: hans rosling's gapminder animations (and his passionate storytelling, demonstrated e.g. in Hans Rosling's TED videos.
- "Hands": domain knowledge gained by working in an industry for years, supported by activity in communities like kaggle (owned by google since 2017), which hosts datasets, notebooks and ml competitions.
- "Brain": analytical mindset and knowledge of analysis tools (none of the tools mentioned here, hadoop, spark, hive - a data warehouse - or sas - another statistical analysis workbench - are necessary - they are merely nice to know). how do you know that you have this kind of brain? e.g. if you enjoy getting quantitative (number-based) answers and if you like visualizations of complex or complicated data (like the gapminder data). also, if you like programming or maths, you've likely got such a brain.
- "Mouth": communication with colleagues - but not only. in fact, especially being able to communicate with people who are not your colleagues (so they are perhaps very different from you) is key. this is another way of saying that you need to be able to "tell a story" after data analysis (e.g. Prevos 2020).
- "Feet": data science is a very fast-moving technology field, especially its "machine learning" offshoot (which is not part of this course) - cp. Kozyrkov 2019. you need to keep on top of the available information. at the same time, there is too much to take in and digest - this means that it is very important to have a sound understanding of the foundations of data science.

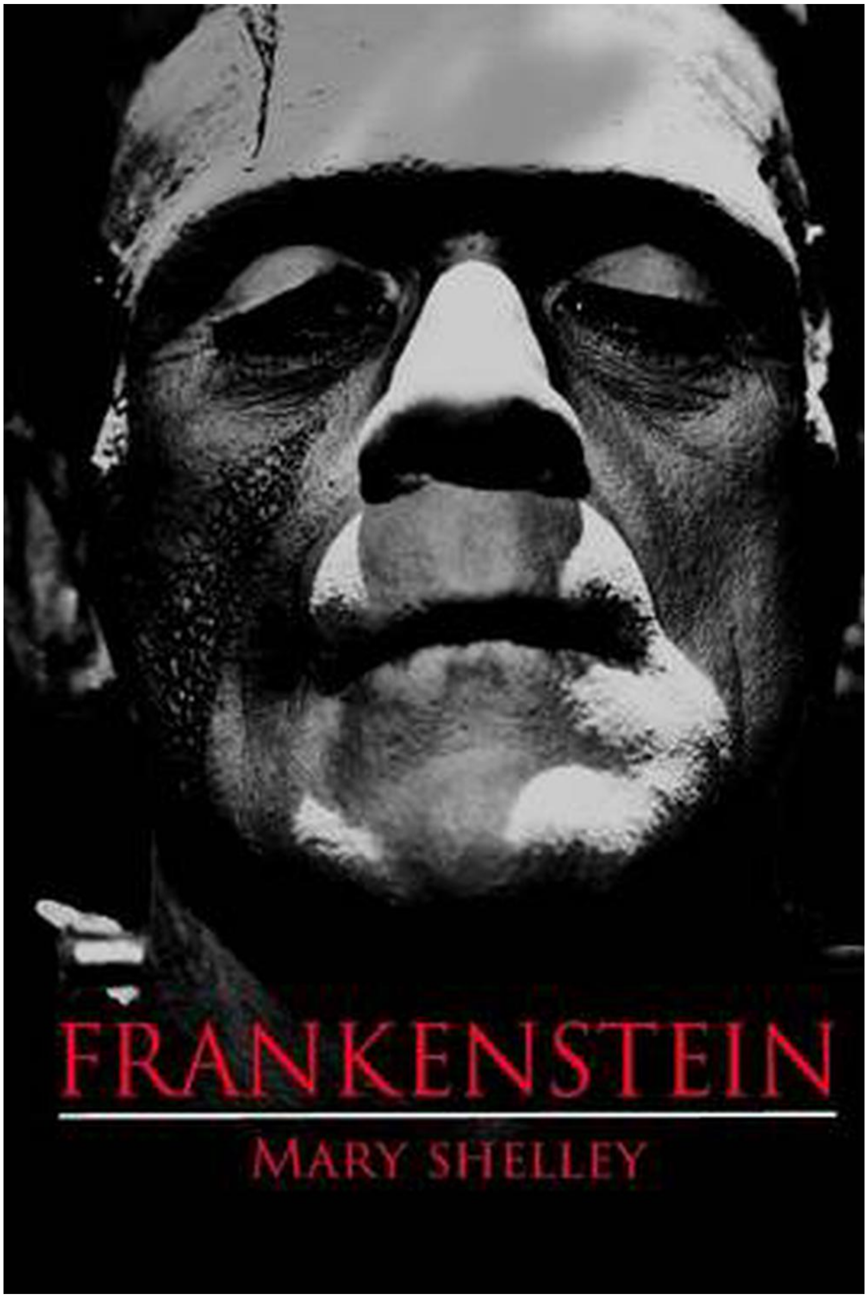
12 DO YOU HAVE A BRAIN FOR NUMBERS?



- What if you don't have a "brain for numbers"?
- What if graphs scare you because of the underlying math?
- What if you like novels but hate manuals?
- What if you actually hate computers and machines?

Can you still have a "brain for data science" like Dr. Frankenstein's monster here? (Hint)

13 WHAT ARE METAPHORS GOOD FOR?

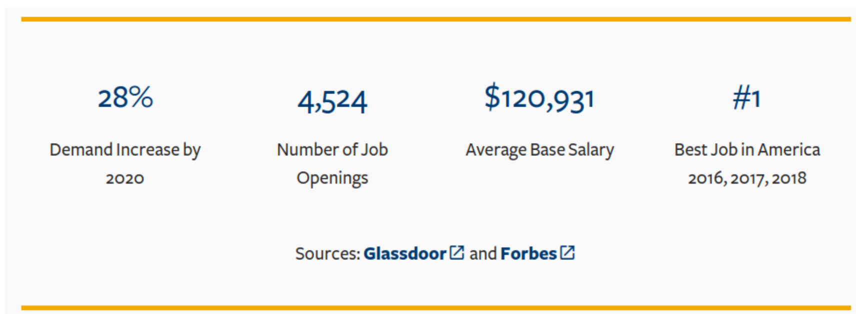


- What are the connotations of "Frankenstein's Data Scientist"?

- Do you find this metaphor apt or not?
- Which metaphor would you have chosen?

. . . youR tuRn: What are the connotations of using "Frankenstein's monster" as a metaphor for "data scientist"? Metaphors are especially important when definitions are not easily forthcoming, are confused or not standardized (all of which is the case for data science). Metaphors are a type of model. (Hints)

14 WHAT'S THE (US) JOB MARKET FOR DATA SCIENTISTS LIKE?



Challenge: search a job portal for "data scientist".

The value of statistics like shown in figure 14 depends on the exact definitions of the job, on the ability of business to recruit exactly for what they want etc. I have personally not spoken to any recruiter about this - I only read career-related blogs and looked at statistics like these (published by Berkeley School of Information 2020, a site that is interested in attracting data science students, therefore highly biased). However, as a rule, you can never go wrong with growing your skill stack, especially with regard to STEM skills, and within these especially with regard to your ability to analyse data quantitatively - which is what data science boils down to. For more details on "data science careers", see Robinson/Nolis (2020).

Mathematics, especially statistics, programming and databases are the skill-based disciplines that you need to master. Having said that: "mastering" could easily take not one, but several life times, and you need to begin somewhere. If you do this in earnest, you'll soon find that you start learning faster and faster the more connections with what you already know you can make.] Here is a (free) book called, incidentally, "Foundations of Data Science" (Blum et al 2015, 466 p.). It includes some geometry, graph theory, linear algebra, markov chains, and a variety of algorithms for "massive data problems" like streaming, sketching and sampling.

15 JOB PROFILES (ACCORDING TO DATACAMP)

Data Engineer	Data Analyst	Data Scientist	Machine Learning Scientist
Store and maintain data	Visualize and describe data	Gain insights from data	Predict with data
SQL + Java/Scala/Python	SQL + BI Tools + Spreadsheets	Python/R	Python/R

- Who would you rather be?
- Why?
- Which job is most in demand?

Introductory DataCamp courses on data science "for everyone" (that is, without being tied to one of the three dominant languages - Python, R, or SQL), contain a job profile section to help users find their professional data science niche.

The figure 15 shows four such profiles from a 2020 course. What is notably missing here is the maths and/or CS or software engineering knowledge required or desirable to fill these roles. But there are also people who say that you best come to a firm as a general-purpose computer scientist and then learn any of these on the job depending on the needs and the available experience.

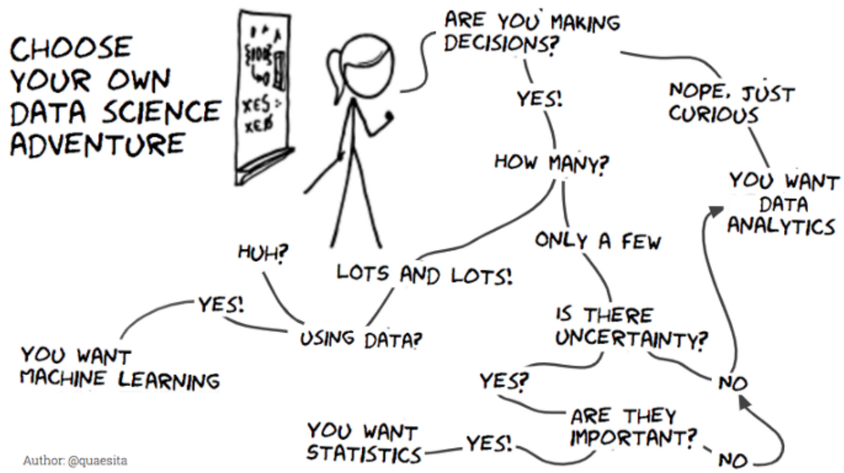


Figure 6: Cartoon by Cassie Kozyrkov (@quaesita)

16 WHAT ARE TYPICAL DATA SCIENCE PROBLEMS?

- **Data analytics**: explorative or explanatory
- **Machine learning**: many decisions involving big data
- **Statistics**: few important decisions with high uncertainty

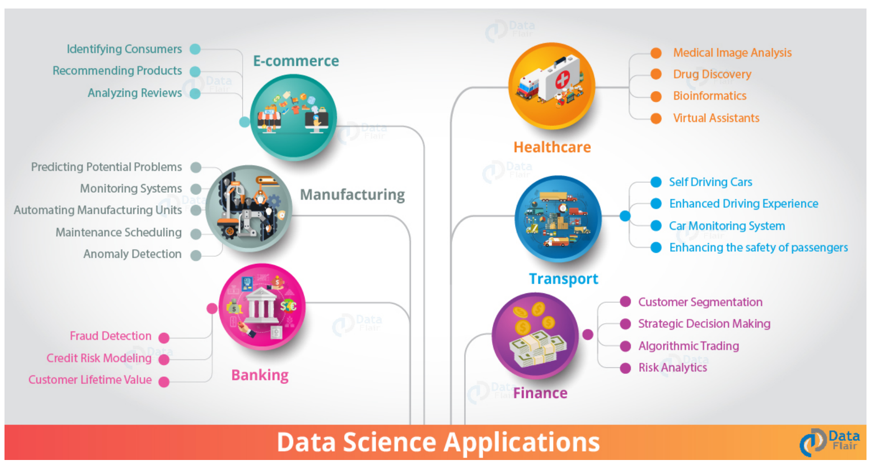
The cartoon in figure 6 is by Google's head of "decision intelligence", Cassie Kozyrkov (2018). She has a specific, business- and decision-oriented idea of the purpose of data science, which I share: data science is there to help you make decisions. The option tree shown distinguishes three sub-fields of data science: data analytics, statistics and machine learning. It asks if you're "making decisions" at the start (many, few, hardly any), it quickly focuses on the type of data (few vs big) and the 'uncertainty' and 'importance' of the decisions. This is still a data-centric, not a decision-centric taxonomy. A focus on the latter would allow for many more options (e.g. strategic vs. tactical, organizational vs. managerial, routine vs. exceptional decisions etc.) Hence, for decision science, this kind of breakdown is not very useful.

The dominance of "big data" has also been doubted, especially when it comes to making (business) decisions. "Small [not big] data" (Saklani, 2017) and "thick [qualitative, descriptive] data" may be just as good depending on what you want to know. The article by Chiu (2020) is a bit of a history hack (in the scholarly sense) but it raises some good points.

Brandon Rohrer, [then] a data scientist at Microsoft, has addressed this question in a 3-part series of short articles (Rohrer, 2015a, 2015b, 2015c). His examples are a more specific, especially because he also says which family of algorithms match which type of data-related question. It is too early for us to discuss his taxonomy but at the end of the course, you should have a better idea about what you can do with data science tools.

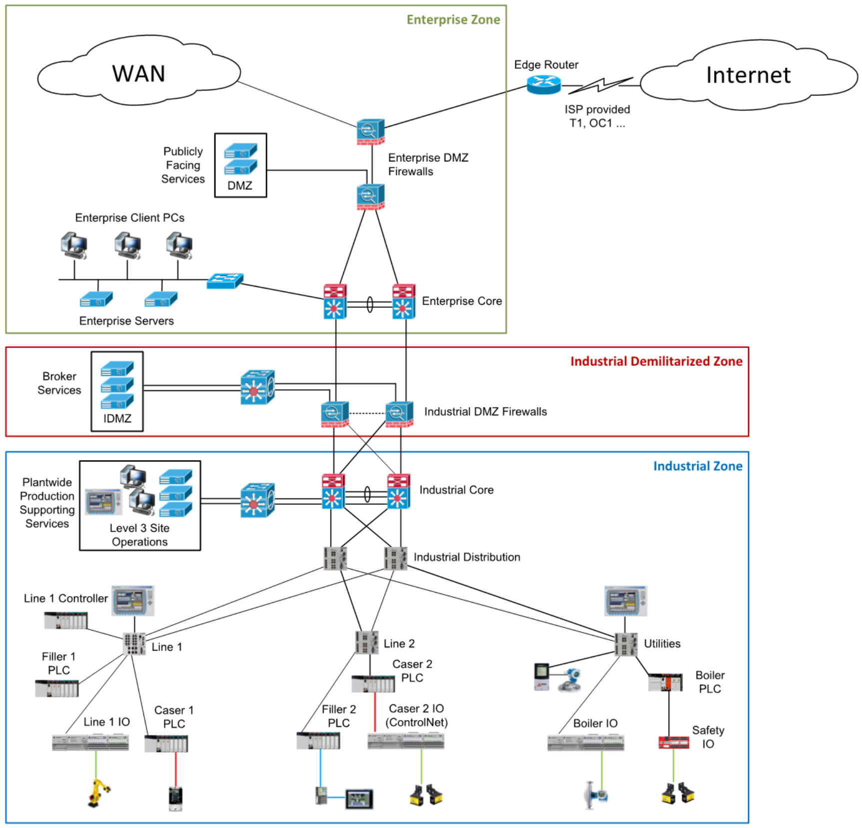
. . . youR tuRn: Think about any decision you make - what are the steps you go through? Do they amount to a "data science adventure" as shown in the figure 6 - why (or why not)? (Hint)

17 DATA SCIENCE APPLICATIONS



Source: data-flair.training

18 EXAMPLE 1: CYBERSECURITY



Source: Industrial Cybersecurity (2017)

- Problem: how to secure critical digital infrastructure
- Solution: Industrial Control System
- Data science: EDA (user data), simulation (sample data)

19 EXAMPLE 2: TIME SERIES ANALYSIS & TEXT MINING

```
Jul 16 09:10:11 linux systemd-timesyncd[1219]: Synchronized to time server [2001:67c:1560:8003::c8]:123 (ntp.ubuntu.com).
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: TCP connection done (i'm the existing device)
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: Starting server ssl (I'm the client TCP socket)
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: TCP connection done (i'm the existing device)
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: Starting server ssl (I'm the client TCP socket)
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: Socket succesfully stablished an SSL connection
Jul 16 09:11:41 linux kdeconnectd.desktop[6764]: kdeconnect.core: It is a known device "Xperia L2"
```

- Data: Linux /var/log/syslog event log
- Problem: Textual time series data
- Solution: Text or process mining of the event log data

All system components continuously write data protocols in the form of simple event logs, which you can view easily on Linux systems e.g. on Ubuntu. Check available system logs with `ls -la /var/log/`. Figure 19 shows a sample section from my computer's system log in `/var/log/syslog`.

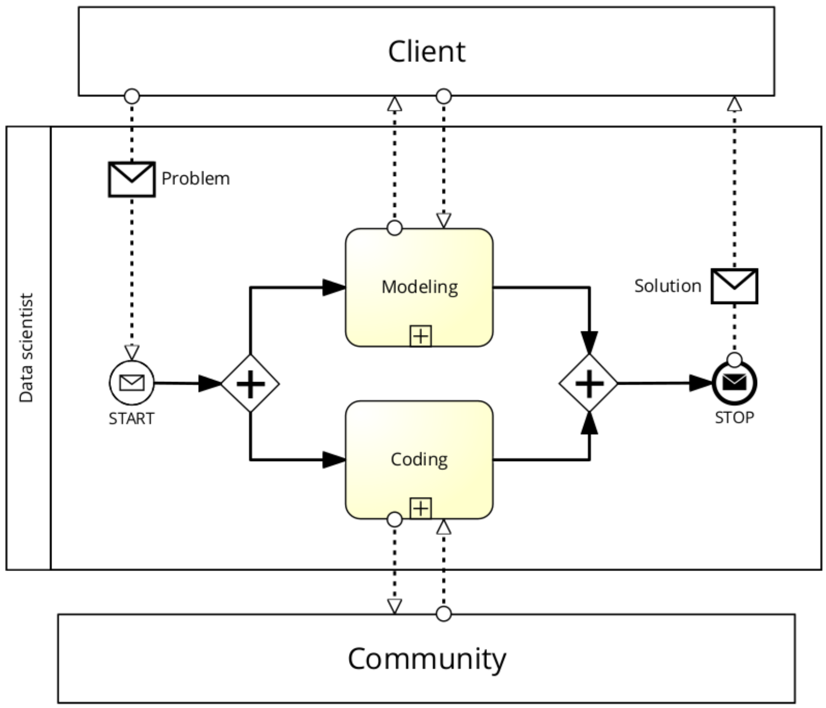
This excerpt shows how and when the computer clock was set remotely, and the starting of various servers and one socket where my mobile phone ("Xperia L2") was connected.

The language we're about to use in this course (and in the follow up course on machine learning), R, is well suited for rapid interactive exploration of datasets such as this one. The two immediately relevant problem areas are "text mining" (notice that all system files are human-readable to aid debugging), and "time series analysis" (event logs are time series).

Text mining is considered a part of "Natural Language Processing", and Time Series Analysis is also really important in finance, e.g. when analysing portfolio performance.

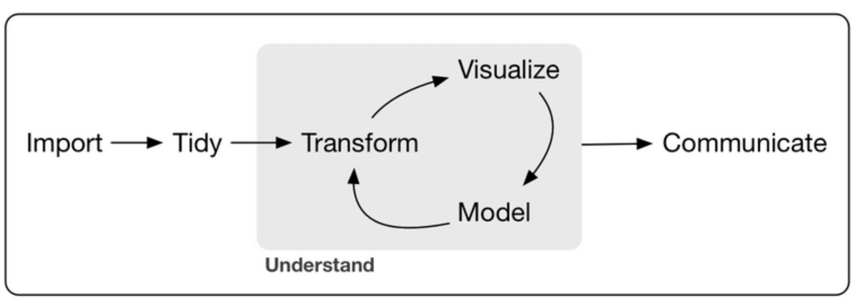
A separate technique (not immediately part of an R programming course) is "process mining".

20 WHAT IS THE DATA SCIENCE PROCESS?



Source: Birkenkrahe (2021)

21 EXPLORATORY DATA ANALYSIS (EDA) PROCESS MODEL

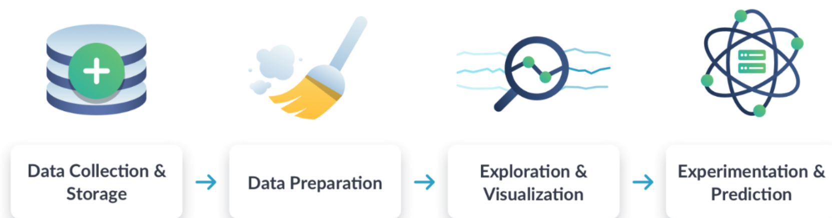


Source: Wickham/Grolemund (2017) - here is my interactive BPMN version.

Figure 20 shows a process that begins with raw data. Such data are usually not formatted as "tidy" data, i.e. "each row represents one observation and columns represent the different variables available for each of these observations" (Irizarry 2020). This is also the tabular format, which is usual for storing data in relational databases for analysis with SQL.

Once we have tidy data, an (often repeated) sub-process begins: "transform" refers to any operation on the dataset that helps us understand the data better. Depending on the size of the data tables, we will use different methods of visualization to make underlying structure visible. But visualization does not always have to be graphical. Let's look at three examples in the next section.

22 DATA SCIENCE WORKFLOW



Source: Data science for everyone (DataCamp)

23 A MODEL FOR LEARNING DATA SCIENCE

- Algorithmic vs heuristic
- Coding vs modeling
- Dashboards vs. Prediction

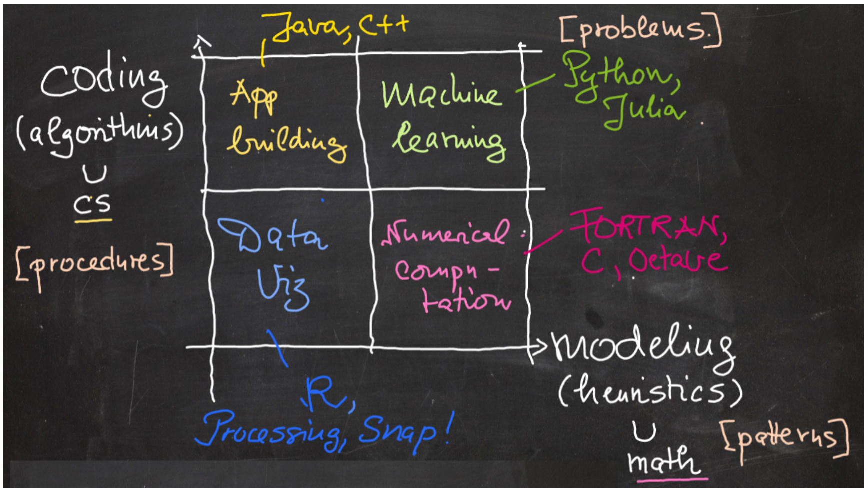
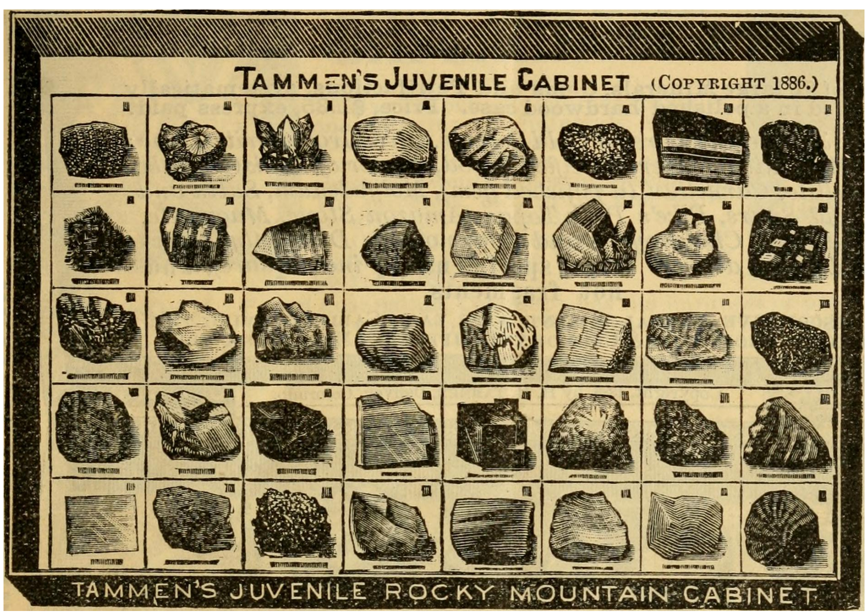


Figure 7: Talk@Lyon College (Birkenkrahe, 2021)

24 CONCEPT SUMMARY



- Data science is used for **decision support**, **process analytics** and **machine learning**.
- Data science makes use of **domain knowledge** - experience in a par-

ticular field of business.

- The *job market(for data science is good.
- The data science **process** includes modeling, visualizing, and communicating data analysis results.

25 REFERENCES

1. Barber M (Jan 14, 2018). Data science concepts you need to know! Part 1. URL: towardsdatascience.com
2. Blum A/Hopcroft J/Kannan R (4 Jan 2018). Foundations of Data Science - Cornell U. Online: cornell.edu.
3. Bobriakov I (16 Apr 2020). Data Science vs. Decision Science [Infographic]. Online: medium.com/@bobriakov.
4. Bolles R and Brooks K (2021). What color is your parachute? Online: `https://www.parachutebook.com/`
5. Chiu J (17 Aug 2020). Why Data Doesn't Have to Be That Big. Online: datacamp.com.
6. Davenport TH/Patil DJ (2012). Data Scientist: The Sexiest Job of the 21st Century. Online: hbr.org.
7. Davenport TH/Patil DJ (July 15, 2022). Is Data Scientist Still the Sexiest Job of the 21st Century? Online: hbr.org.
8. Devlin K (1 Jan 2017). Number Sense: the most important mathematical concept in 21st Century K-12 education. Online: huffpost.com.
9. Gapminder Foundation (15 Dec 2014). DON'T PANIC - Hans Rosling showing the facts about population. Online: youtube.com
10. Grolemund G/Wickham H (2017). R for Data Science. O'Reilly.
11. Irizarry R (2020). Introduction to Data Science. CRC Press.
12. Kozyrkov C (10 Aug 2018). What on earth is data science? Online: hackernoon.com.
13. Kozyrkov C (22 May 2019). Automated Inspiration. Online: Forbes.com]].

14. Knuth D (1992). Literate Programming. Stanford, Center for the Study of Language and Information Lecture Notes 27.

15. Myers A (28 Apr 2020). Data Science Notebooks - A Primer. Online: medium.com/memory-leak.

16. Porras E M (18 Jul 2018). Linear Regression in R. Online: datacamp.com.

17. Prevos P (14 Aug 2020). Storytelling with Data: Visualising the Receding Sea Ice Sheets. Online: lucidmanager.org]].

18. Robinson E/Nolis, J (2020). Build a Career in Data Science. Manning.

19. Rohrer B (2015a). What Can Data Science Do For Me? Online: microsoft.com.

20. Rohrer B (2015b). What Types of Questions Can Data Science Answer? Online: microsoft.com.

21. Rohrer B (2015c). Which Algorithm Family Can Answer My Question? Online: microsoft.com.

22. Saklani P (19 Jul 2017). Sometimes "Small Data" Is Enough to Create Smart Products. Online: hbr.org.

23. Sarkar DJ (12 Sept 2018). A Comprehensive Guide to the Grammar of Graphics for Effective Visualization of Multi-dimensional Data. Online: towardsdatascience.com

24. Scherpereel CM (2006). Decision orders: A decision taxonomy. In: Management Decision 44(1):123-136.

25. Wing JM (2 Jul 2019). The data life cycle. Harvard Data Science Review. Online: hdsr.mitpress.mit.edu.

26 "Your tuRn" (HINTS AND SOLUTIONS)

26.1 Popularity

Check out the seminal article by Davenport/Patil 2012. (At least) one answer is in there.

26.2 Skills

Recently, an MBA student asked me these same questions and here is my answer: "My IT Skill Stack". See also Bolles and Brooks (2021)

26.3 Software

- D3.js, a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS.
- Apache Hadoop, a "software library framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures." (Source: Apache.org)
- MapReduce, "a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop. The term "MapReduce" refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job." Source: IBM. See also: tutorialspoint.
- Apache Spark, "a lightning-fast unified analytics engine for big data and machine learning. It was originally developed at UC Berkeley in 2009." Source: databricks.
- NoSQL "databases, purpose-built for specific data models and have flexible schemas for building modern applications. NoSQL databases are widely recognized for their ease of development, functionality, and performance at scale." Source: AWS.
- Apache Pig, "a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient

property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets. At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin." Source: apache.org. Tutorialspoint.

- Tableau (owned by Salesforce), commercial interactive data visualization software (SQL-based dashboards). Tableau public.

- iPython notebook (now "Jupyter Notebook"), a "interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media." Source: jupyter.org. Part of the Anaconda distribution. See also: Google Colaboratory for a (free) cloud-based version.

- GitHub (owned by Microsoft), "a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code" (Source: kinsta.com) centered on the open-source version control software Git. There are many platforms like GitHub (e.g. GitLab, BitBucket, SourceForge).

  Of these applications, only Git (not GitHub) is really absolutely necessary for a professional data scientist working in teams. Though a working knowledge of the principles behind all of them will be very useful (especially if they come up in interviews). Hence, no reason to be scared.

26.4 Your brain

Other terms for what we're talking about here are: "number sense" (in maths education), or "computational thinking" (in computer science) or, more recently, "data literacy". All of these are relatively new concepts, so feel free to speculate and make up your own mind! Cp. Devlin 2017

26.5 Frankenstein

How do you feel about anything if doing it would turn you into a monster? What kind of monster is Frankenstein (if you didn't read the book or saw the film, I'll tell you: ugly but soulful, loveable and capable of love, too)? What is special about him as a monster in mechanical terms?

26.6 Job market

Mathematics, especially statistics, programming and databases are the skill-based disciplines that you need to master. Having said that: "mastering" could easily take not one, but several life times, and you need to begin somewhere. If you do this in earnest, you'll soon find that you start learning faster and faster the more connections with what you already know you can make.] Here is a (free) book called, incidentally, "Foundations of Data Science" (Blum et al 2015, 466 p.). It includes some geometry, graph theory, linear algebra, markov chains, and a variety of algorithms for "massive data problems" like streaming, sketching and sampling.

26.7 Decisions

The figure (like the underlying article) targets business decisions more than everyday decisions. For business decisions, taxonomies exist, which are generally a lot more complicated than shown here, see e.g. Scherpereel 2006.

26.8 Process

On the surface, Wing's "Data Life Cycle" (2019) has a few more steps (and it is also not a "cycle") - it does not use the artificial (technical) term "tidy" but instead terms that can more easily be understood by practitioners outside of data science. Modeling is not addressed by Wing but instead she puts "management" at the center of the process, right between data-centric and (business) process-centric categories. Another related process model you may have heard of is the "design thinking" process, which plays an important role in innovation and when solving so-called "wicked problems".

26.9 Summary

"The ability to write code" is still the "most basic, universal skill" for a data scientist - which is why learning R is the focus of this introductory course. There are many data science programs at universities now - often offered as minors or as Masters programs for people trained already in maths, computer science, or fields with obvious and current data science applications (like biology). The understanding of a data scientist as a hybrid professional is still very rudimentary.