A data science overview

Marcus Birkenkrahe

Winter 2020

Contents

1 [raw] DATA + [code] + SCIENCE [methods] = Value

- ☐ How popular is data science?
- ☐ What skills do you need?

- ☐ What problems are solved with it?
- ☐ What is the data science process?

In this talk, I am giving an overview of several aspects of data science. Though young, 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.

I want to 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.

2 How popular is data science?

In this 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 February 2020 (Source).

See also: Challenge (1)

3 What skills do you need to do it?

The three skill areas give rise to different tasks and problem settings:

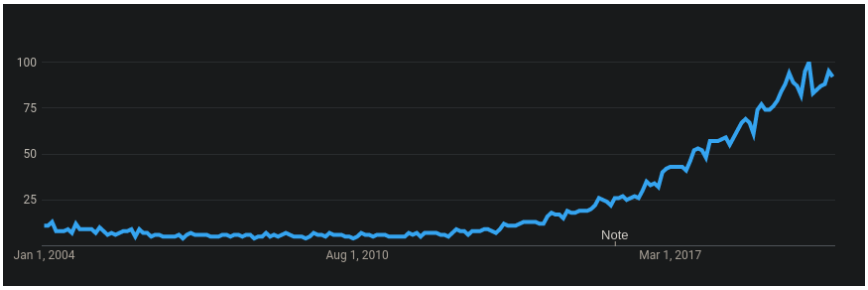


Figure 1: Google Trends, August 2020

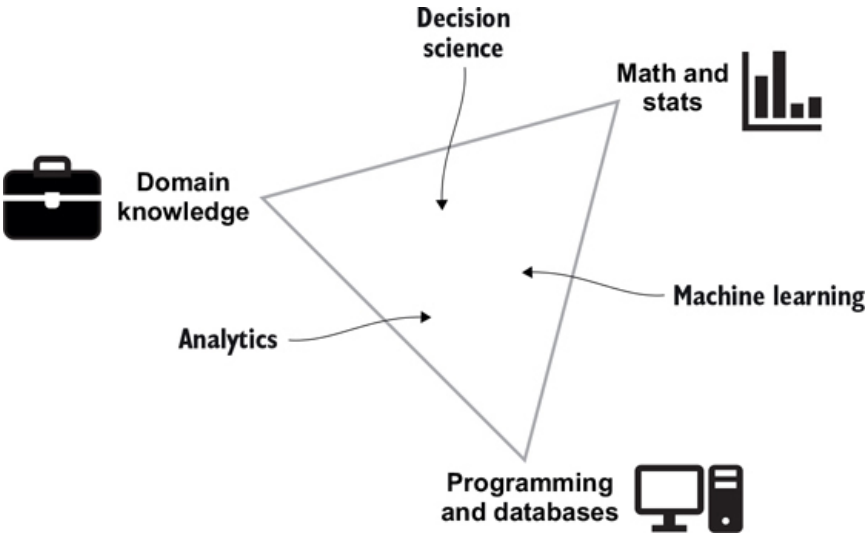


Figure 2: Robinson/Nolis, 2020

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[1]
	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[2]

Between two of these areas each are application areas: (1) domain knowledge and statistics support decision science. See infographic (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.

See also: Challenge (2)

4 Which technical skills are required? [9/9]

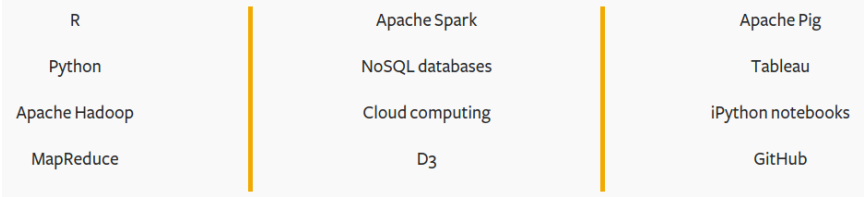


Figure 3: UC Berkeley School of Information, 2020

A little background on the names mentioned:

[1] 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").

[2] "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)

- D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS.

- Apache Hadoop "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 "MapReduce is 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 "is 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 are 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 "is 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 notebooks (now "Jupyter Notebook"): "interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media." Source: jupyter.org. Part of Anaconda. See also: Google Colaboratory.
- GitHub (owned by Microsoft): "is 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.

  See also: Challenge (3)

5 What exactly is Frankenstein's data scientist made up of? [6/6]

"Frankenstein's monster" (based on the novel by "Frankenstein, or The Modern Prometheus", by Mary Shelley, 1818) is used here 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 in this video (gapminder 2014).

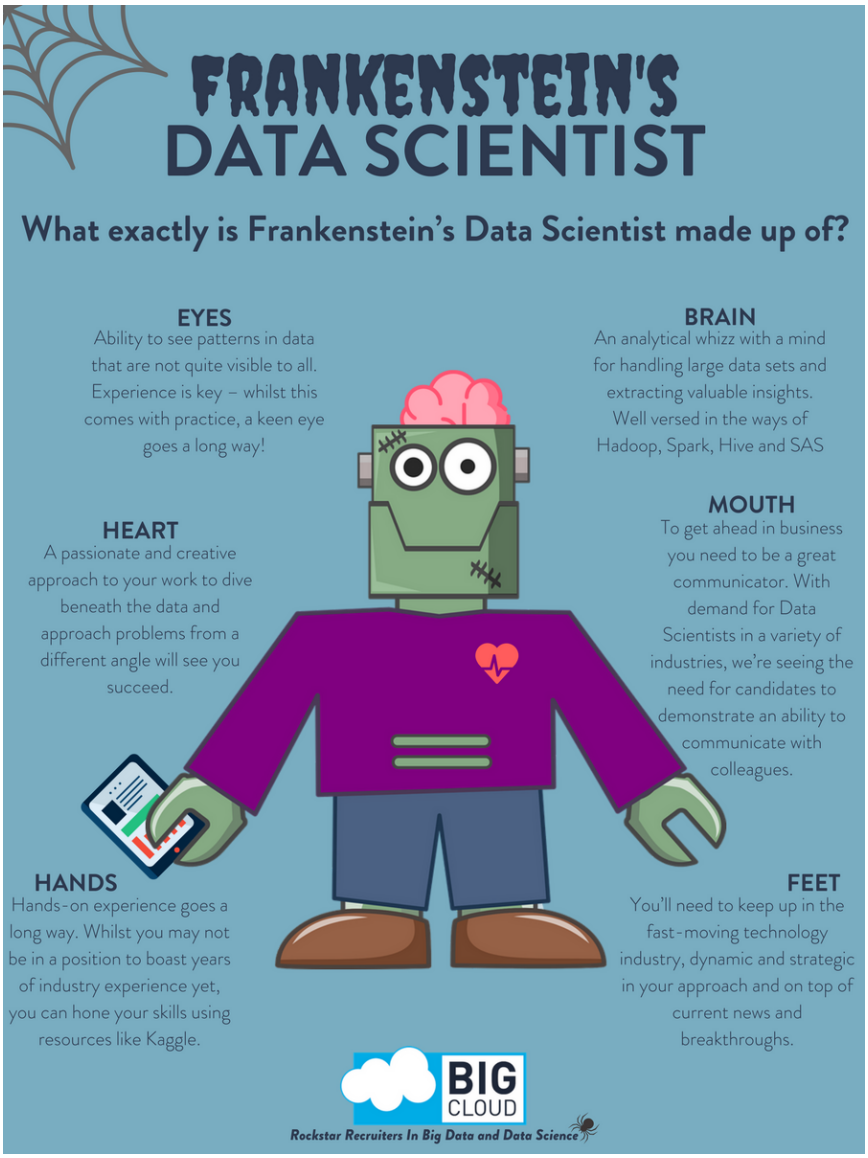


Figure 4: source: bigcloud (bigcloud.io)

- "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.

See also: Challenge (4) and Challenge (5)

6 What's the (US) job market like?

Statistics like these are highly volatile, of course. They depend 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).

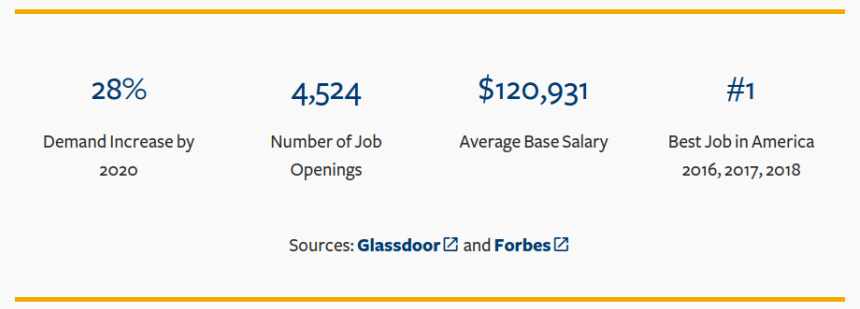


Figure 5: US Jobmarket for Data Science (Source: UC Berkeley, June 2020)

See also: Challenge (6)

7 What problems are solved with it?

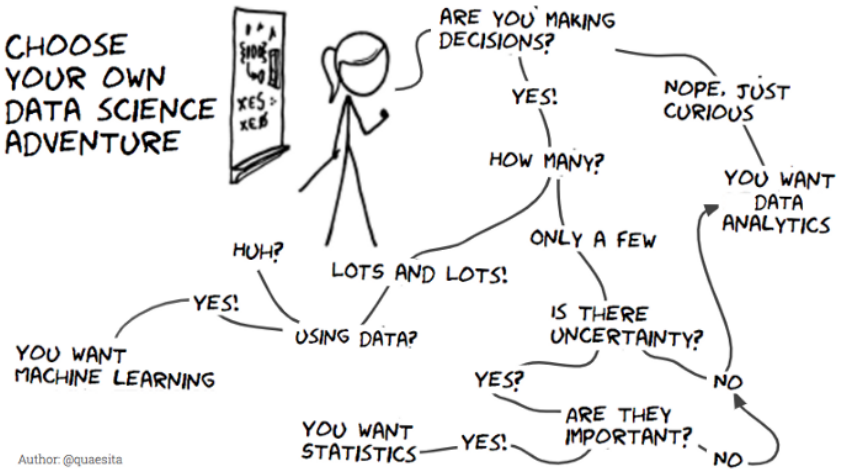


Figure 6: Source: Cassie Kozyrkov (@quaesita)

The cartoon 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.

8 What is the data science process?

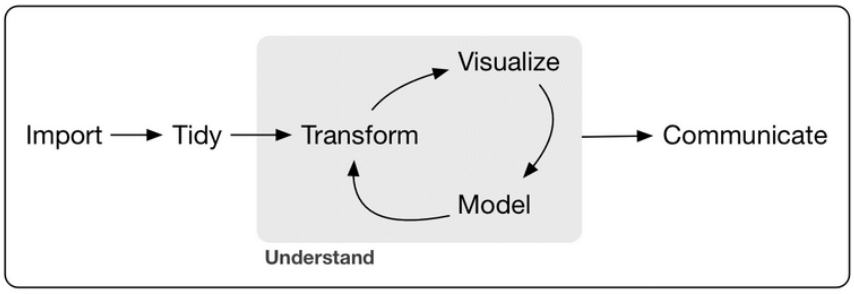


Figure 7: Source: Wickham/Grolemund, 2017

The figure 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:

(1) Sometimes, even data selection is helpful - e.g. to see the variables of a dataset, the `head(dataset)` function, which gives us the first 6 observations for all variables, may suffice.

```
data(mtcars) # load dataset 'mtcars'
head(mtcars) # first six lines of dataset
```

```
:                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
: Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
: Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
: Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
: Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
: Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
: Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

(2) The `str(dataset)` function, which gives an overview of the dataset size and all variables including the data types. Here are examples for all three of these utility functions for the built-in dataset `mtcars` ('Motor Trend Car Road Tests'):

```
data(mtcars) # load dataset 'mtcars'
str(mtcars)  # variables in dataset
```

```
'data.frame': 32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

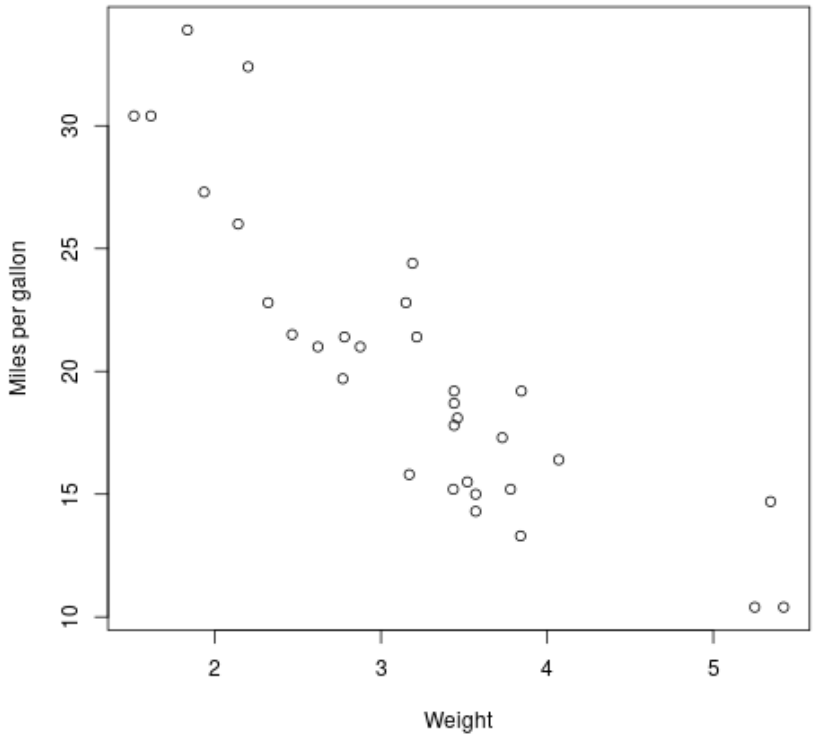
(3) The `summary(dataset)` function, which computes various statistical standard measures for the variables and helps you understand data content.

```
data(mtcars) # load dataset 'mtcars'
summary(mtcars)  # summaries of various model fitting functions
```

```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

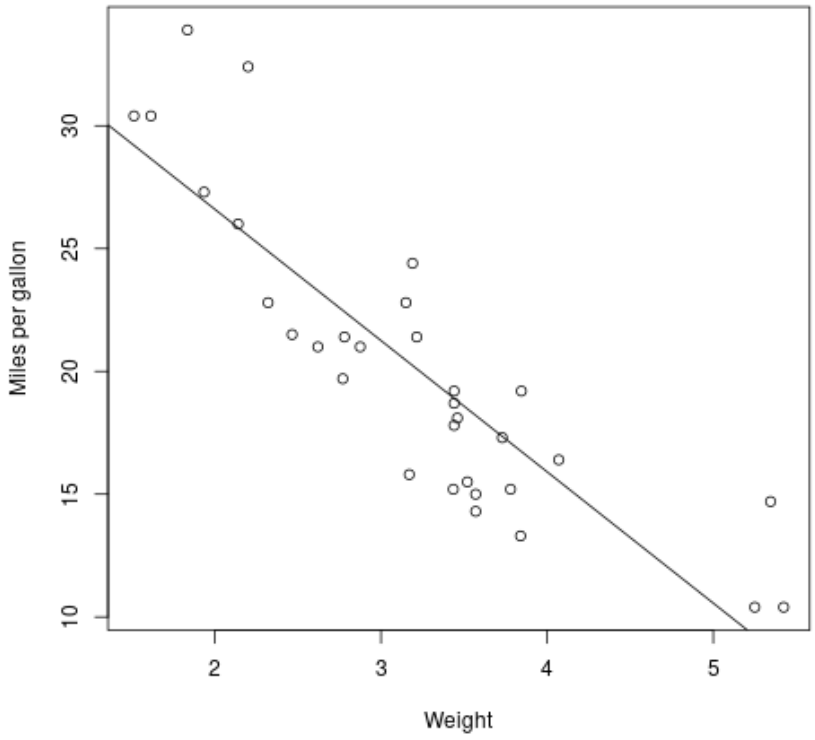
As an example of graphical visualization, let's plot two variables of the `mtcars` dataset against one another: e.g. miles-per-gallon (`mpg`) as a function of the car's weight (`wt`):

```
data(mtcars)
x <- mtcars$wt   # x-axis: weight of each car
y <- mtcars$mpg  # y-axis: miles-per-gallon of each car
plot(x, y, xlab = "Weight", ylab = "Miles per gallon")
```



The last step in the cycle, "model", is technically the most difficult one. It requires at least a basic understanding of algorithms and statistics to be done correctly, and domain knowledge, to be done meaningfully. In our example, we assume that the miles per gallon fall linearly with the weight of the car - the heavier the car, the less economical it will run. The model used here is linear regression, `lm` in R (see Porras 2018 for a tutorial):

```
data(mtcars)
x <- mtcars$wt   # x-axis: weight of each car
y <- mtcars$mpg  # y-axis: miles-per-gallon of each car
plot(x, y, xlab = "Weight", ylab = "Miles per gallon")
lm_model <- lm(y ~ x, data = mtcars) # create a linear regression lm
abline(lm_model) # add the line for lm to the plot
```



Once we've found a model that seems to fit the data, we understand the data much better than before, when we only had the raw data. We are now ready to communicate our insights. Often, this leads to the need to begin all over again, e.g. gather more or different data, and try different models. See here for a BPMN model in which I have detailed some of the steps of the model by Grolemund and Wickham (2017).

See also: Challenge (8)

9 SUMMARY

- ☐ You can define data science in terms of relevant areas.
- ☐ You can name some data science skills and traits.
- ☐ You can say what data scientists do with data.

- □ R code examples: We have already seen quite a bit of R code in the examples - the why, what and how of R will be the subject of the next lecture. You don't need to remember these now - we'll practice them soon and often enough - but if you do, it might be helpful:

```
data()
head()
str()
summary()
plot()
lm()
abline()
```

After hearing a lot of information, I find it sometimes necessary to anchor myself again. You can do this and test your basic understanding of data science by addressingT Challenge (9).

10 REFERENCES

1. Blum A/Hopcroft J/Kannan R (4 Jan 2018). Foundations of Data Science - Cornell U. Online: cornell.edu.
2. Bobriakov I (16 Apr 2020). Data Science vs. Decision Science [Infographic]. Online: medium.com/@bobriakov.
3. Chiu J (17 Aug 2020). Why Data Doesn't Have to Be That Big. Online: datacamp.com.
4. Davenport TH/Patil DJ (2012). Data Scientist: The Sexiest Job of the 21st Century. Online: hbr.org.
5. Devlin K (1 Jan 2017). Number Sense: the most important mathematical concept in 21st Century K-12 education. Online: huffpost.com.
6. Gapminder Foundation (15 Dec 2014). DON'T PANIC - Hans Rosling showing the facts about population. Online: youtube.com
7. Grolemund G/Wickham H (2017). R for Data Science. O'Reilly.
8. Irizarry R (2020). Introduction to Data Science. CRC Press.
9. Kozyrkov C (10 Aug 2018). What on earth is data science? Online: hackernoon.com.

10. Kozyrkov C (22 May 2019). Automated Inspiration. Online: Forbes.com.

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

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

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

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

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

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

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

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

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

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

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

11 Challenges

1. ☐ How do you explain this curve? What happened in 2012? [Answer: check out Davenport/Patil 2012.]

2. ☐ Where do you have your skills? How do you know that you have that skill? In which area would you like to improve your skills? Recently, an MBA student asked me these same questions and here is my answer: "My IT Skill Stack".

3. ☐ Which software packages are these? What do they do? Can you use them freely? [Answer: see script. 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 computer and data science are crafts.]

4. ☐ What if you don't have a brain for numbers, can you still do data science? What if numbers don't turn you on but instead put you to sleep? What if graphs scare you because you suspect that difficult mathematics is necessary to understand the graph and what's behind it? What if you like novels but don't care for manuals - what if you don't even like computers? Can you still have a "brain for data science"? [Note: other terms for this 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]

5. ☐ What are the connotations of using "Frankenstein's monster" as a metaphor for "data scientist"?

6. ☐ What is the "foundation of data science" in terms of skills? How can you learn them? Check out an online job portal like [Answer: see skills slide - 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. 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.] There is also a (free) book called "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.]

7. ☐ Think about any decision you make - what are the steps you go through? Do they amount to a "data science adventure" and why (or why not)?

8. ☐ Compare the process in the figure with the "data life cycle" by Wing 2019. There is no graphical model so you may have to draw one yourself (e.g. as a BPMN process model). What's the purpose of such a life cycle model? Do you know of any other models like this?

9. ☐ Read the seminal article by Davenport/Patil (2012), which put data science on the map for business. What has changed since then (if anything)? Test their claims (e.g. "there are no university programs offering degrees in data science" - is this still true?). How would you measure the performance of a data scientist? Has the understanding of data science and data scientist ("a hybrid of data hacker, analyst, communicator, and trusted adviser") changed since then?