www.datatrain.cc

Contents

Chapter 1

Welcome!

What this is, and what it isn't

This is not a textbook or an encyclopedia. This is not a reference manual. It is not exhaustive or comprehensive. We barely mention statistical tests at all. There is almost no theory. In fact, this curriculum is biased, non-representative, and incomplete – *by design.*

So what is this? This guide is an *accelerator,* an *incubator* designed to guide you along the most direct path from your first line of code to becoming a capable data scientist. Our goal is to help you through the most dangerous period in your data science education: your very first steps. The first three weeks. That is when 99% percent of people give up on learning to code.

But it doesn't need to be this way. We made this book to reach more than just the 1%.

We have based our approach on three core premises:

Premise 1: We learn best by doing. Our goal is to get you *doing* data science. We will keep theory and detail to a minimum. We will give you the absolute basics, then offer you exercises and puzzles that motivate you to learn the rest. Then, once you've been *doing* data science for a bit, you soon begin *thinking* like a data scientist. By that, we mean tackling ambiguous problems with persistence, independence, and creative problem solving.

Premise 2: We learn best with purpose. Once you gain comfort with the basic skills, you will be able to start working on real data, for real projects, with real impact. You will start to *care about what you are coding.* And that is when the learning curve *skyrockets* – because you are motivated, and because you are learning *reactively*, instead of preemptively. Our goal is to get you to the point of take-off as quickly as possible.

Premise 3: A simple toolbox is all you need to build a house. Once you become comfortable with a few basic coding tools, you can build pretty much anything. The toolbox doesn't need to be that big; if you know how to use your tools well, and if you have enough building supplies (i.e., data), the possibilities are limitless.

Who this is for

The target audience for these tutorials is the *rookie*: the student who *wants* to work with data but has *zero* formal training in programming, computer science, or statistics.

What you will learn

- The **Core theory** unit establishes the conceptual foundations and motivations for this work: what data science is, why it matters, and ethical issues surrounding it: the good, the bad, and the ugly. Don't slog through this all at once. Sprinkle it in here and there. The most important thing, at first, is to start writing code.

The next several units comprise a *core* curriculum for tackling data science problems:

- The **Getting started** unit teaches you how to use `R` (in `RStudio`). Here you will add the first and most important tools to your toolbox: working with variables, vectors, dataframes, scripts, and file directories.
- The **Basic `R` workflow** unit teaches you how to bring in your own data and work with it in `R`. You will learn how to produce beautiful plots and how to reorganize, filter, and summarize your datasets. You will also learn how to conduct basic statistics, from exploratory data analyses (e.g., producing and comparing distributions) to significance testing.

For these first two units, we encourage you to take on these modules one at a time, in the exact order they are presented: we put a lot of thought into what we included in these modules (and what we did not).

- The **Review exercises** unit provides various puzzles that allow you to apply the basic `R` skills from the previous unit to fun questions and scenarios. In each of these exercises, questions are arranged in increasing order of difficulty, so that beginners will not feel stuck right out of the gate, nor will experienced coders become bored. This is where you really begin to cut your teeth on real-world data puzzles: figuring out how to use the `R` tools in your toolbag to tackle an ambiguous problem and deliver an excellent data product.
- The **Reproducible research** unit equips you with basic tools needed for truly reproducible data science: documenting your research and code with `Markdown`; weaving together your code and your reporting with `RMarkdown`; allowing users to explore the data themselves with an interactive `Shiny` dashboard or web app; and sharing your code and tracking versions of your code using `Git`.

Who are we?

www.datatrain.global.

Part I

Core theory

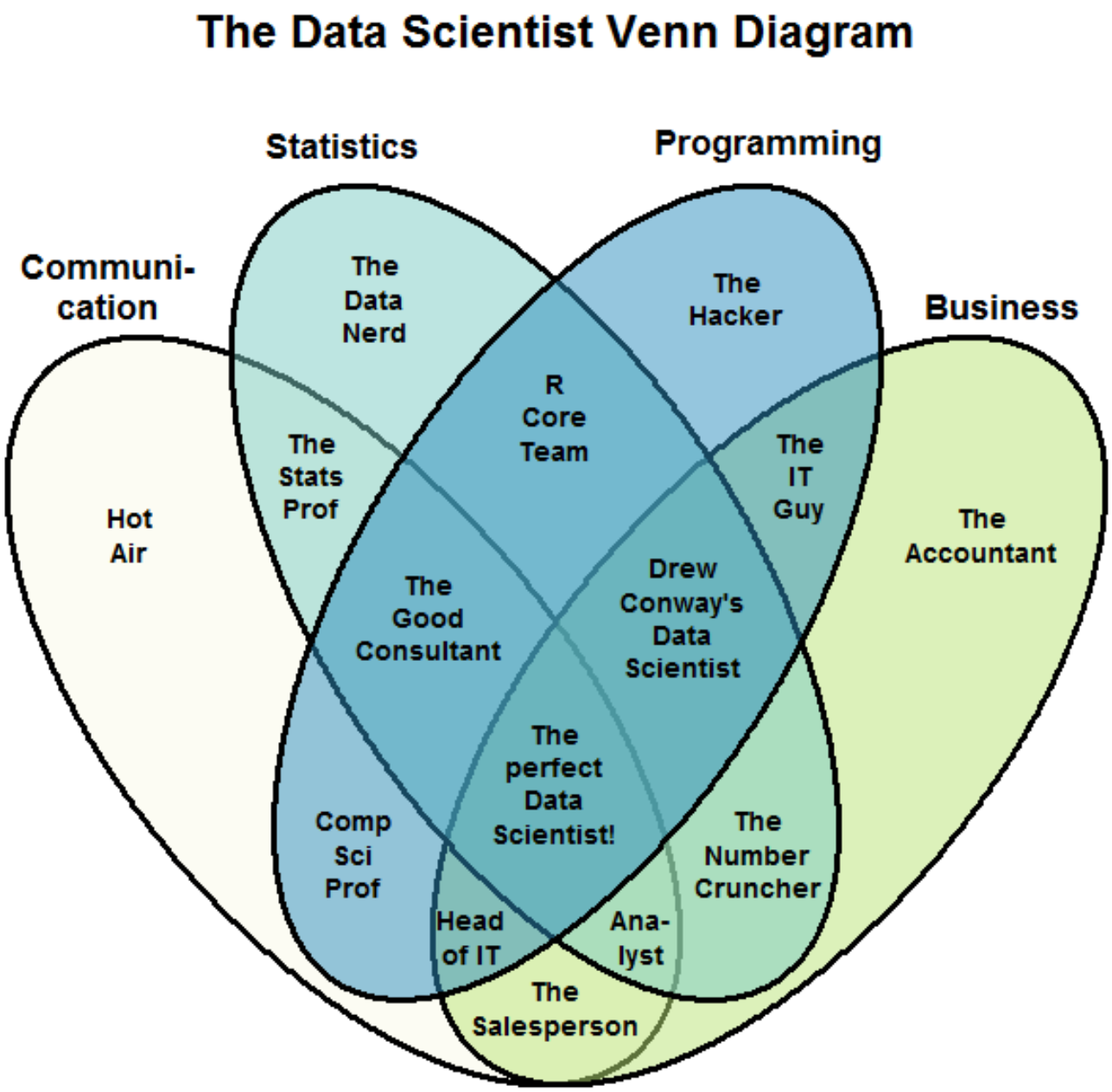
Chapter 2

What is data science?

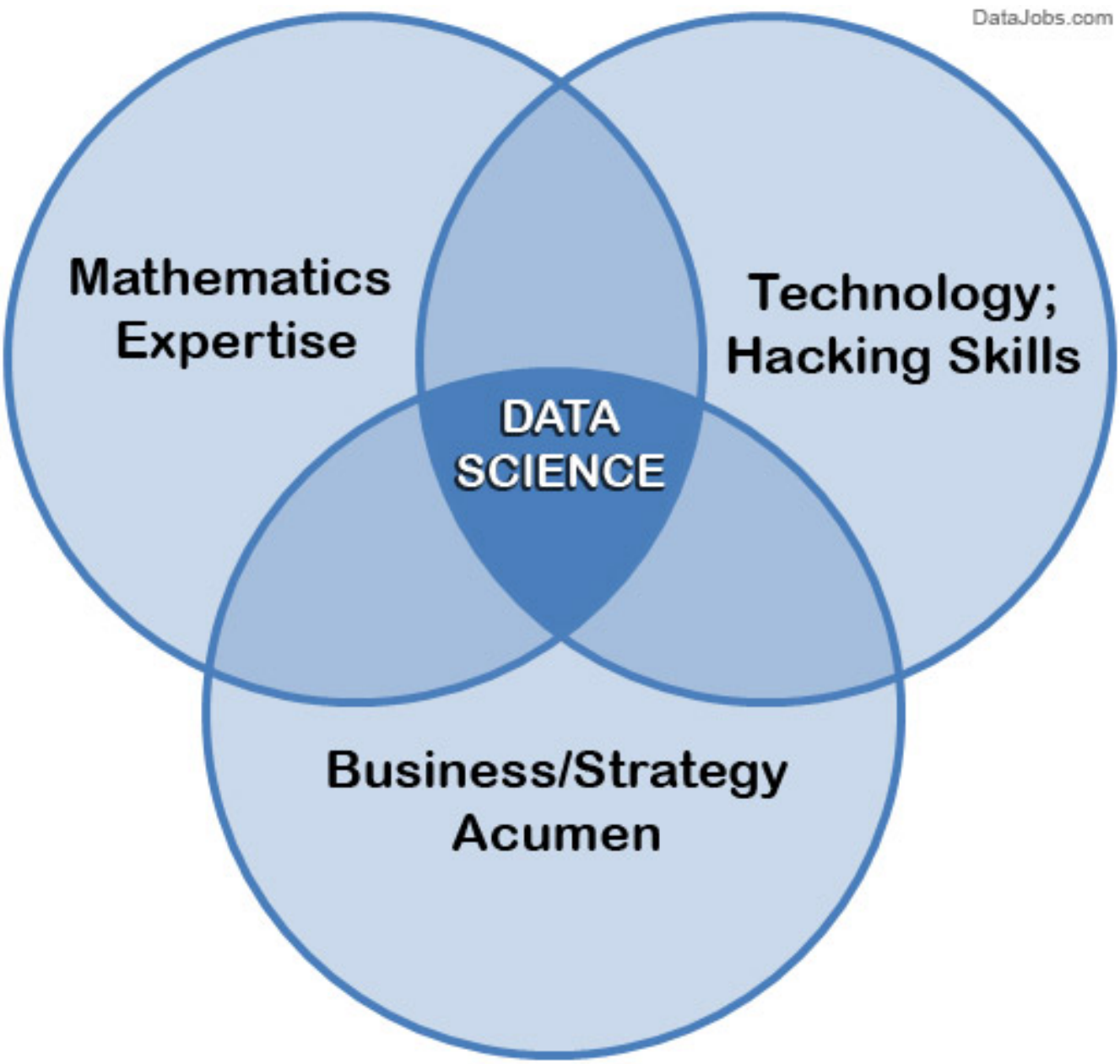
The definition of data science is a moving target. Thirty years ago (1990), 'data science' was an uncommon term that essentially just meant statistics. Twenty years ago (2000), the phrase mainly referred to querying SQL databases. Fifteen years ago (2005), it was "dashboards" and "predictive analytics". Ten years ago (2010), it was 'big data' and 'data mining'. Nowadays folks think of A.I. and machine learning.

In 10 years? Who knows.

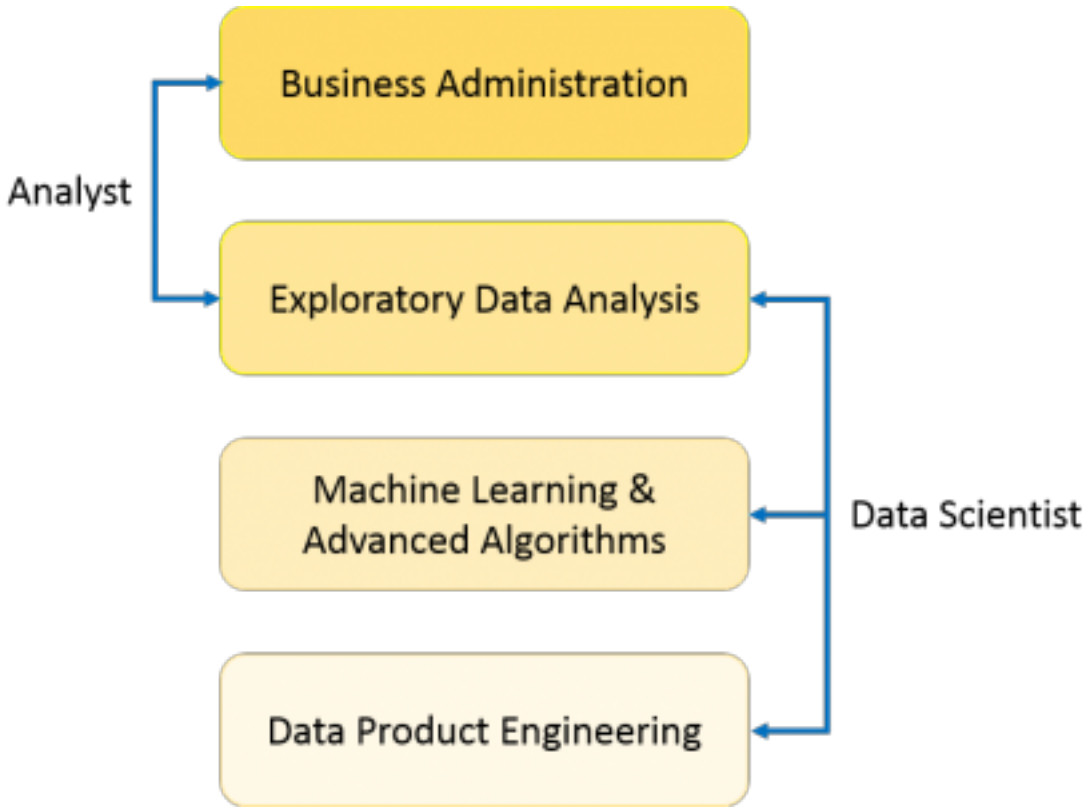
So what is it? There are many definitions out there. Search the internet for the answer and you will find complex diagrams, such as this one, suggesting that a data scientist is someone who has the right blend of programming skills, statistical knowledge, communication ability, and business acumen:



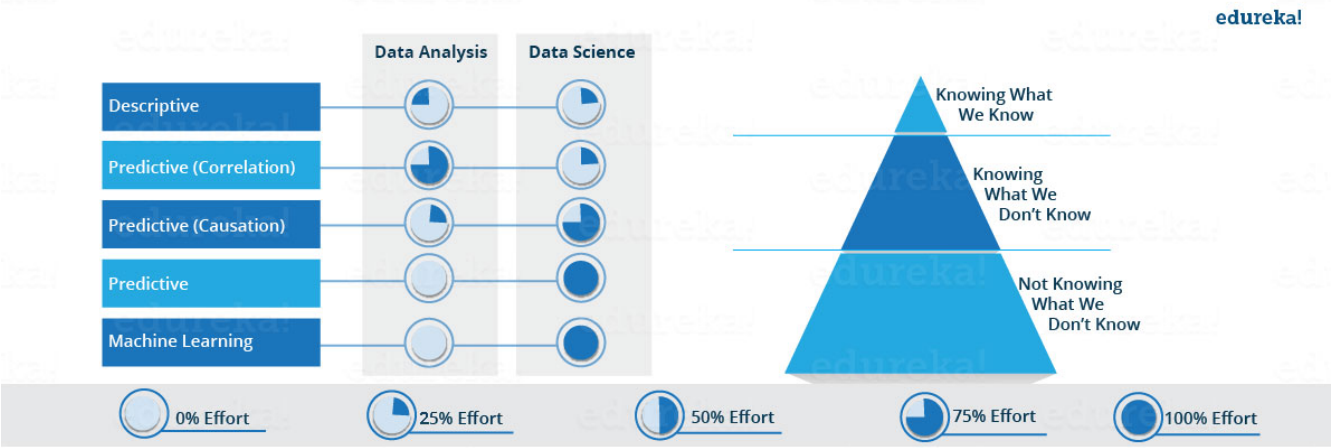
… or here is a more popular, simpler version of the same diagram:



Some argue that data science is simply an extension of statistics. You will also find attempts to distinguish between categories of data science, or to draw lines around what data science is and what it is not. A classic example is the bizarre delineation corporations draw between a data *scientist* and a data *analyst*:



... or ...



Our take? Those definitions are useful, interesting, and to some degree accurate. *But* data science is too new, and too fluid, to be fixed into some static definition. So, to keep our definition accurate, we'll keep it broad:

Data science is simply "doing science with data". And for our purposes, the only difference between our definition and the definition of science itself is not in the word "data" – since nowadays all scientists are, to some degree, "data scientists" – but rather in the word "doing". Data science is about *doing* stuff with data – about *making a difference with data.* And that's what this course is going to be about. *DOING.*

But we'll go one step outward. Data science is not just the combination of academic disciplines like stats and business strategy. Good data science also needs to involve (1) **domain knowledge** (i.e., familiarity with the problem being solved), (2) **a bias to real-world effects** rather than theoretical frameworks, and (3) **a desire to work in the real world**. To do so, data scientists generally need to be effective communicators and have an iterative mentality: they try something, evaluate its effects, try something else, and repeat.

Our definition is very broad, we know. We consider the "analyst" working in business intelligence to be a data scientist; and so too do we think that a data scientist could be an engineer who is processing large amounts of data to extract basic trends. Again, data scientists are those who *do science with data*. That's a lot of people.

In our experience, the best data scientists aren't simply the best programmers or best statisticians; the best ones are the people who consider themselves to be *something else first.* They are the journalists, artists, epidemiologists, psychologists, historians, environmentalists, sports analysts, and political commentators who *also* know how to work with data. In other words, the best data scientists are the ones already out there, on the ground, already embedded in the system they want to improve, positioned perfectly to get the right data, to ask the right questions, and to actually *do* something with insights from the data. Again, data science is about *DOING*.

To summarize, data science is about applying data to problems. It is impact-driven, transdisciplinary, and suited to well-rounded, multi-dimensional professionals.

What is the data life cycle?

There is a misperception about data science work that it is largely or even exclusively interpretative: that is, a data scientist looks at a big set of data and builds a fancy statistical model, then a light bulb goes off in her head, she has some insight, and then acts on that insight.

The reality is data science is much more than that. And most of data science is a combination of *(a)* getting data ready for analysis, *(b)* hypothesis testing, and *(c)* figuring out what to do with the results of *a* and *b*. That is, data science in practice is generally not some artesenal genius staring at a table of numbers until "insight" magically occurs. Rather, it is a lot of work, a lot of structured theories which can be confirmed or falsified, and a lot of *imagination* applied to the task of implementation.

In other words, data goes through a whole *lifecycle* of which analysis is just a small part.

What is the data lifecycle? Here's how we conceptualize it:

0. Observation
1. Problem identification & definition
2. Question formation
3. Hypothesis generation
4. Data collection
5. Data processing

This step is usually the most intensive. Half the battle is wrangling raw data and making it ready for a visualization or a hypothesis test. Note that this step has *nothing to do with statistical tests* – data science is not the same as statistics!

6. Model building / hypothesis testing

Note that this step is usually where the scientific method stops. In science, once you analyze your test, you interpret your results and loop back to the beginning of the data cycle. But in *applied data science*, there are a few more steps:

7. Operationalization: This means determining how best to incorporate the data insights into operations.

8. Communication / dissemination

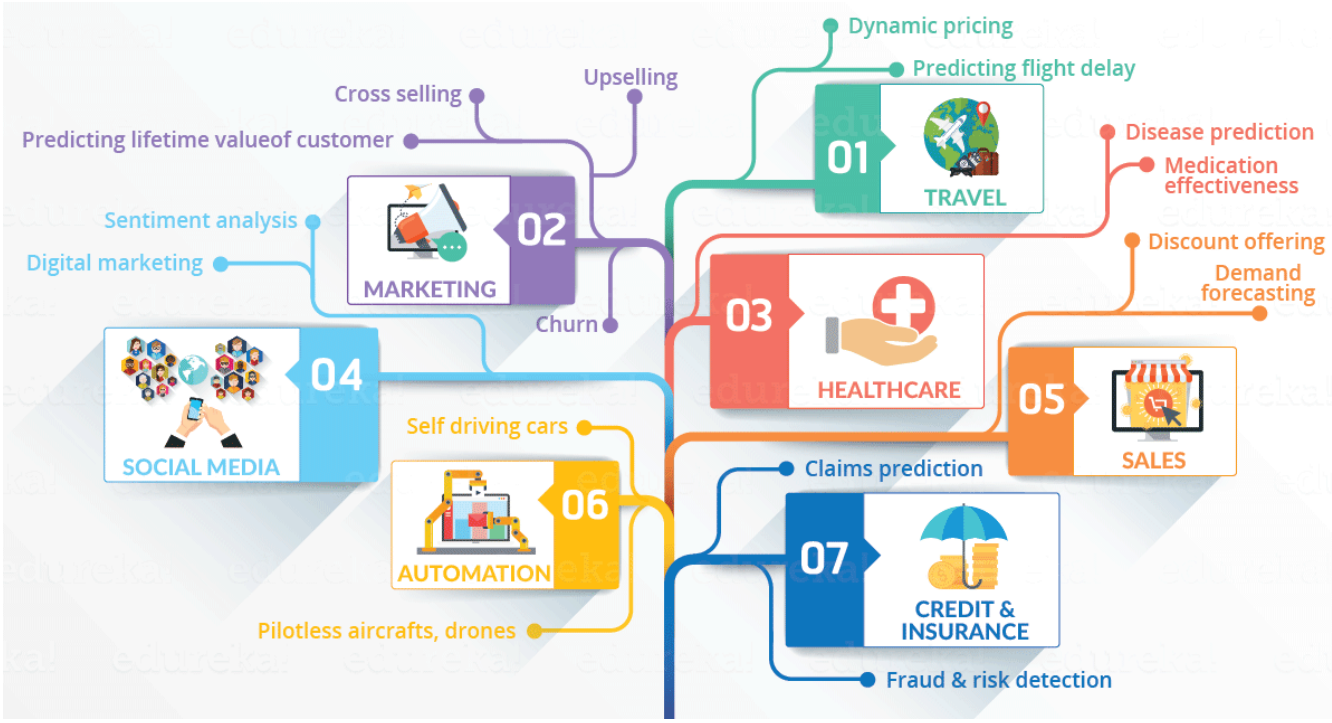
9. Action: This means actually implementing the change.

10. Observation: back to the beginning of the cycle.

Again, the above should look a lot like the scientific method. The main differences are (a) "data processing", which in reality takes up most of any data scientist's time, (b) the bias towards action, and (c) the iterative / looped nature of the lifecycle.

Data science 'in the wild'

Enough theory. What do data scientists actually do? Again, you can search for an answer online and find complex diagrams like this one:



But to capture every problem that data scientists are working on, this diagram would have to be even more convoluted and complex. Data scientists are working on a *ton* of problems.

The most stereotypical data science problems tend to involve advertising, social media, and corporate profiteering:

- Targeted advertising

- Social media feed optimization (getting you to scroll just a little further)
- Facial recognition (automated tagging at Facebook)
- Voice recognition ('Hey, Siri!', 'Alexa!')
- Making video games more fun / addictive

- Dynamic airline pricing

- Search autocomplete

- Autocorrect

- Virtual assistants

These are the kinds of problems that the best-paid data scientists in the world are working to solve. Right now there are thousands of programmers in Mountain View, Cupertino, and elsewhere in the Bay Area (and New York, and London, and Beijing) trying to solve the problem of you not spending enough time on social media.

Maybe you care about these problems, maybe you don't. Maybe they make you indignant or angry. Maybe you find it *problematic* that these things are even considered problems at all. As far as we're concerned, it is deeply unfortunate that our highest-paid data scientists are focusing on problems like these.

But take heart – there are plenty of other data scientists out there working on *actual* problems that are actually *important*:

- Identifying disease through imagery
- Automating identification of credit card fraud
- Filtering spam with malware or viruses.
- Preventive maintenance at nuclear facilities
- Improving chemotherapy dosage
- Increasing voter turnout
- Improve matchmaking systems (liver transplants, love, etc.)
- Measuring deforestation with satellite imagery.
- Efficient and equitable vaccine distribution
- Identifying tax evaders
- Predictive policing
- Storm surge forecasting
- Identifying and removing child pornography from the internet
- Surveilling emergency rooms to predict disease outbreaks
- Detecting fake news
- Increasing accountability and legitimacy of carbon markets
- Quantifying the likelihood of recidivism to prevent over-incarceration

The list goes on. The number of worthwhile problems waiting for data scientists is limitless, there are data scientists working on problems like these right now, and the demand for civic-minded data scientists is immense.

All of this matters for a lot of reasons. The first is that data science is not always a good thing; it can be weaponized by corporations and governments in spite of the public interest, and for that we need to be very careful about how we use it and how we teach it.

But the second reason this matters is that data science can be an *equally powerful force for social good*. We can use data science to make progress on the most urgent and injurious social and environmental problems of our time.

However – and this is the third reason all this matters – data science can only achieve social good *if* we recruit students to its ranks who are values-driven, civic-minded, and committed to using data science for good.

Fourth, and finally, this matters because the Facebook data scientists are using the exact same principles and basic tools as the non-profit data scientists. At their core, the foundational skillsets are the same.

And that's what this book is all about.

Chapter 3

The reproducibility crisis

The crisis

There is a crisis in the sciences: the reproducibility crisis. It is also known as the replication crisis. This refers to the fact that many scientific studies have been impossible to reproduce, calling into question the validity of those studies' findings.

This crisis began in the mid-2000's, when psychologists realized they could not reproduce most of their colleagues' results. They tried to repeat the experiments, following the methods step-by-step, but failed to get the same results. This was enormously unsettling for psychologists, and it cast major doubts upon the validity of psychological theory.

The realization that much of published research is not actually reproducible soon spread to medical research...


PLOS MEDICINE

advanced searc

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

75,226 Save | 5,909 Citation | 2,720,175 View | 7,825 Share

...then it sprung up in marketing...


The Desperate Need for Replications

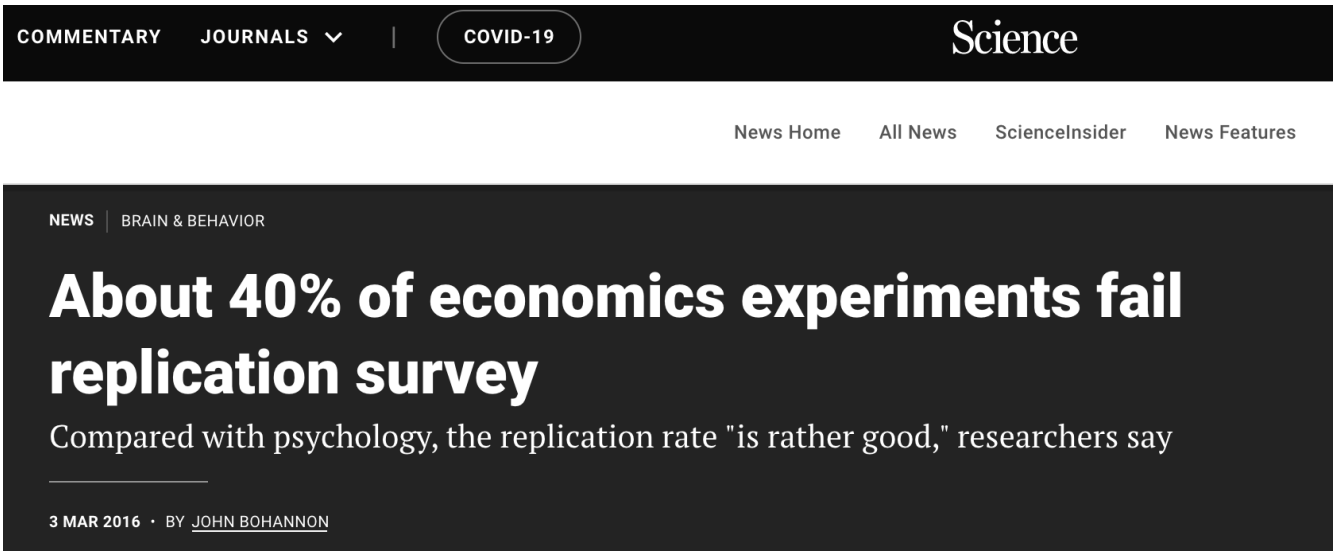
John E. Hunter

Journal of Consumer Research, Volume 28, Issue 1, June 2001, Pages 149–158,

https://doi.org/10.1086/321953

Published: 01 June 2001

...and economics...



...and the sports sciences...

...and the life sciences too:

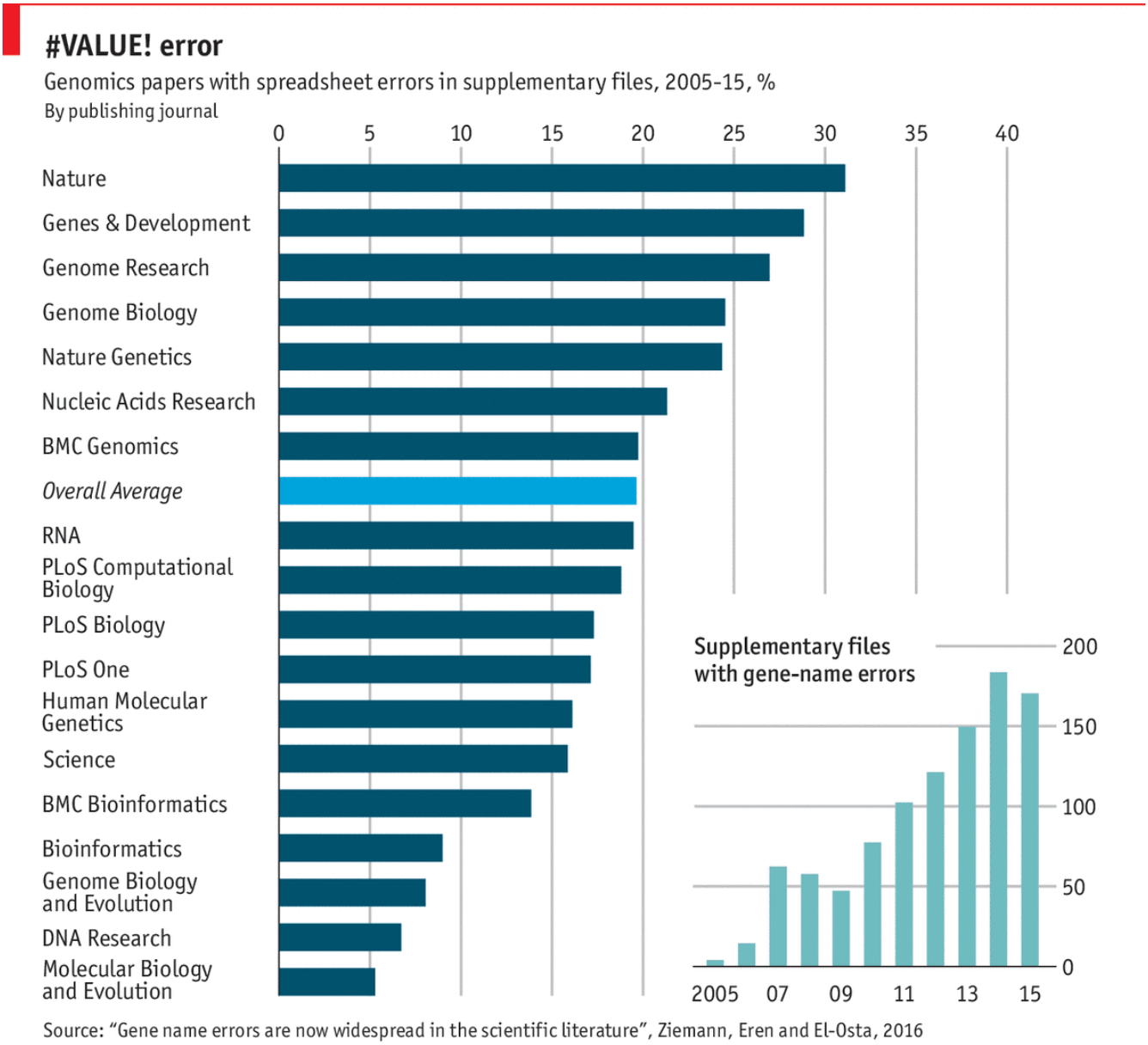
For a complete history of the crisis, check out this article from Wikipedia.

Why is this happening? There are many reasons. Many studies, particularly those in psychology and the social sciences, involve small cohorts of participants. When sample sizes are low, results may not be representative of underlying truths.

On rare occasions it is intentional and fraudulent: scientists face pressure to publish interesting results, so much so that they might fabricate or filter their data to make their results significant.

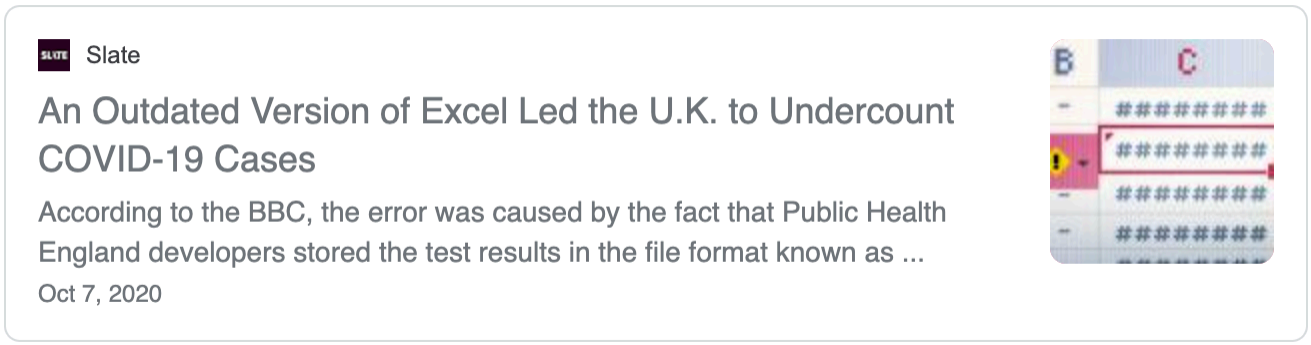
But the most common causes of reproducibility failure are, by far, (1) **poorly documented steps in data processing** – if you don't know how exactly the authors of a paper formatted their raw data to prepare them for analysis, you simply can't reproduce the analysis – and (2) simple, **honest mistakes**, such as typos in spreadsheets.

Consider this summary of the reproducibility crisis from *The Economist*. A scary percentage of genomics studies have simple spreadhseet errors:



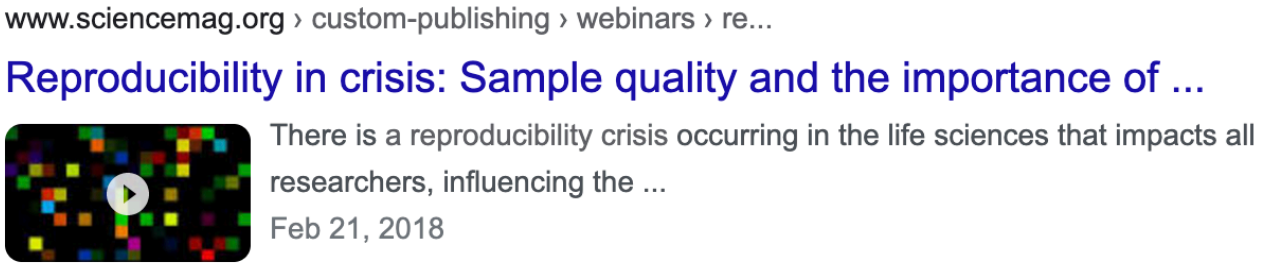
This is a big deal: if a significant part of science is *wrong*, then what do we know? How can we be sure what we know is right? How can we build off of previous research? How can we distinguish valid science from the rest? If science can't be trusted, what value does it have for society? What kind of *damage* is it doing to society?

This crisis is ongoing, and it is impacting our handling of the COVID-19 pandemic. On October 5, 2020, the world learned that 16,000 COVID-19 cases disappeared from the UK's public health database due to a simple glitch in *Microsoft Excel*.

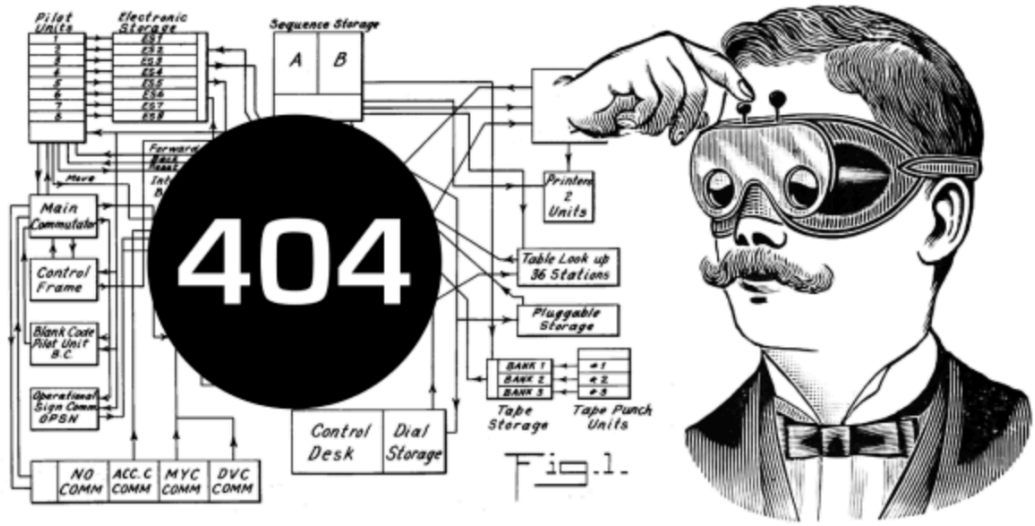


That event demonstrated that the reproducibility crisis is not just an academic concern. It can have serious and potentially deadly consequences for the public.

But there are **silly examples** of the replication crisis, too. Perhaps our favorite is this: in August 2021, when we Google'd "reproducibility crisis", one of the top search results is this video from *Science*, the world's most prestigious scientific journal:



But when we click on this link, here's what we see:



Hmmm ... this doesn't *look* like science.

It seems you're in search of a page that doesn't exist, or may have moved. You can use the Back button in your browser to return to the page that brought you here, or search for your missing page.

If you'd like to visit a page that has plenty of science on it, please visit our homepage.

BACK TO HOME

The 'reproducibility' movement

Because of this crisis, there has emerged a much needed move to make all science "reproducible". This means making sure that someone else can copy what you did, and get the same results. This is important for identifying scientific fraud, of course, but also for helping us to overcome human bias, mistakes, wishful thinking, etc. Reproducibility is not just a "nice-to-have"; in modern science (and data science), it's a "must".

Good data science must be reproducible. The idea is that work done by scientist A is "reproducible" by scientist B. In other words, if the findings of the research are of any generalizable value, then the results of two scientists working on the same problem should be identical (or very high in agreement). In practice, this means using data and code in a structured, well-documented, accessible, clear way, and ensuring that others can do the same.

Reproducible research also means using tools that others can easily use, and methods that others can easily copy. Programming languages like `R` and `Python` are ideal for this.

Reproducible research matters for lots of reasons:

1. Because making your work reproducible means that *you* will have less problems returning to that work at a later time.

2. Because making your work reproducible means that *others* can collaborate with you, help you, error-check you, and build on your work.

3. Because making your work reproducible means you are fighting the plague of irreproducible results which have characterized the replication crisis.

Making your work reproducible is going to be a bit more work, but it's not optional. And there are tools and best practices in place to make it as painless as possible. Basically, reproducible research involves the following:

- Using code to format and manage your data instead of spreadsheet software such as *Excel* or *GoogleSheets*, since those products will not keep a step-by-step record of each thing that you do. When you code, each command line is both an action you take to process your data *and* a record of what that action is.
- Coding with free, open-source tools, such as `R`.
- For any specific niche task in your analysis, such as processing a batch of images, using other open source tools (e.g., `ImageJ`) that can be used free-of-charge by anyone with an internet connection anywhere.
- Documenting *everything* you do with the data, by commenting your code thoroughly and by creating "Wiki" pages for your projects.
- Making your code open source and freely available online.
- Making your data open source (while protecting privacy and confidentiality of participants).
- Providing tools, such as `Shiny` in `R`, that allow others to explore your data themselves, rather than trusting your own narratives about the data.
- Using tools for generating reports, such as `Rmarkdown`, that remove the 'middle-man' and avoid potential typo's and fabrications.
- Collaborating openly with others.

You will be learning how to do all of these things in this course. We are going to focus on *reproducible research*, *literate programming*, *documentation*, and other components of data science (and research in general) which ensure that (a) our methods and findings can be easily sanity-tested by others, and (b) we set ourselves and our projects' up for future collaborations, hand-offs, and expansion.

Why `R`?

This course is largely about learning to *do*, and will largely use `R`. `R` is not the only tool in the data scientists' toolbox (there are many), but it's a good one, is extremely popular, there is almost nothing you cannot do with it, it can be applied to many fields, and – most importantly – it is a free, open-source tool with an active open-source coding community. The millions of `R` users worldwide emulate the spirit of reproducible research we are trying to advocate for here.

A final thought

A research article about *results* is advertising, not scholarship.

Scholarship is an article with transparent, reproducible methods.

Chapter 4

Data ethics

A few principles

This orientation to the principles of data ethics is not going to be adequate or sufficient. We just need to provide enough context for you (1) to appreciate the limitless complexity and uncertainty of many ethical issues in data science, and (2) to start exploring the complex ethical scenarios below on your own or in dialogue with others.

In most frameworks for data ethics, three foundational principles are used to help us think about whether certain research actions are ethical. Those three principles are:

1. Respect for persons and their autonomy: participation must be based upon informed consent, and privacy must be honored at all times. Immature or incapacitated persons must be protected as they mature or heal.

2. Beneficence: in our work with data, potential risks are minimized while potential benefits are maximized.

3. Justice: Benefits and risks are distributed equally across groups of people. A classic way of asking whether something is 'just' is asking using John Rawls' concept of the **'veil of ignorance'**: pretend you have no information at all about your circumstances or your place in the social order: You don't know your place of birth, year of birth, sex, skin color, language, religion, immigration status, health conditions, or anything else. In other words, you have no information whatsoever that might introduce bias into the way you think about the world. Free of circumstantial bias, what arrangements would you choose to put in place to maximize fairness and fortune for all, and to minimize the chances that you would get screwed by the system?

These principles can guide us as we navigate ethically ambiguous scenarios. When we ask whether something is ethical, we are asking whether all of these principles are upheld. We could also be asking whether the violation of one of these principles might be justified by upholding another in an impactful way.

The question, 'Is something ethical?' is usually not easy to answer, particularly when it comes to the use of data in tackling social problems. It is important to note that reasonable people regularly disagree on these ideas; that is why we have committees and drawn-out processes for obtaining permission to use data in research and commerce.

So why do these principles matter? Because without them, we would not be able to have conversations about the ethics of difficult situations. We need articulated principles that we can point to and debate together. Principles like these allow you to have an account for why you feel the way you do about a certain issue. Without that account, we can't learn from each other's perspectives.

Note also that these principles were designed with **individual human subjects** in mind. It is an open question of active debate how exactly these principles can be applied to **communities of individuals** all at once – what exactly does it mean for a group to consent to something? Does every single individual need to consent? The majority? – or how they translate to our treatment of **non-human communities**: animals, plants, and places.

Let's stop there and explore some concrete scenarios. For a better orientation to ethical precepts underlying issues of data ethics, this chapter by Shannon Vallor is the best open-source resource that we have been able to find. Many of the case studies and scenarios presented below are adapted from that chapter.

Warm up scenarios

Practice applying the above principles to these scenario questions. For each scenario, describe your **opinion vector** (the *direction* of your opinion – yes or no – and the *strength* of your opinion).

Location tracking
Is it ethical for Google to track and store your location information in order to monitor traffic and operating hours of local businesses? Such traffic information is known to help direct emergency service vehicles along the safest and fastest route.

Targeted advertising
Is it ethical for internet search engines to tailor advertisements according to your search history?

Dynamic pricing
Is it ethical for airfare search engines to adjust ticket prices according to your recent search history?

Social media scrolling
Is it ethical for Instagram to count how many milliseconds you spend on each post, then use that info to develop a strategy for getting you to spend more time on its app?

Controversial content
You are a data scientist at Facebook. Based on your analyses of user data, you have discovered that when you show readers sensational or hyperbolic content, such as someone ranting that a new vaccine is an attempt at government-subsidized mind control, the readers stay on Facebook longer and scroll through more contet. Since that translates to profits, is it OK for your team to increase the amount of sensational content in users' feeds?

Case studies

Use these case studies to reflect upon and discuss the ambiguity, complexity, and dangers of data ethics issues.

The Facebook 'Social Contagion' Study

In 2014, data scientists from Facebook published an article in a prestigious academic journal. In this article, they demonstrated that the emotions and moods of users could be manipulated by toggling the amount of positive or negative content in their feeds. They found that these emotional effects would then be passed to other users in the social network; in other words, emotions and moods could be seeded and were 'contagious'. To carry out this research, they manipulated the Facebook feeds of 689,000 users.

PNAS Proceedings of the National Academy of Sciences of the United States of America

Keyword, Author,

Home | Articles | Front Matter | News | Podcasts | Authors

RESEARCH ARTICLE

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

+ See all authors and affiliations

PNAS June 17, 2014 111 (24) 8788-8790; first published June 2, 2014; https://doi.org/10.1073/pnas.1320040111

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

This article has Corrections. Please see:

Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks - July 03, 2014

Correction for Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks - July 03, 2014

Article | Figures & SI | Info & Metrics | PDF

Significance

We show, via a massive (N = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

Facebook did not receive specific consent for this study from its users. Instead, the company argued that the purpose of the study was consistent with the user agreement already in place: to give Facebook knowledge it needs to provide users with a positive experience on the platform.

Discuss:

1. In what ways, specifically, did Facebook violate basic principles of data ethics with this study? Enumerate each violation individually.

2. Can a convincing argument be made that justifies this study? What are the strongest arguments in its favor?

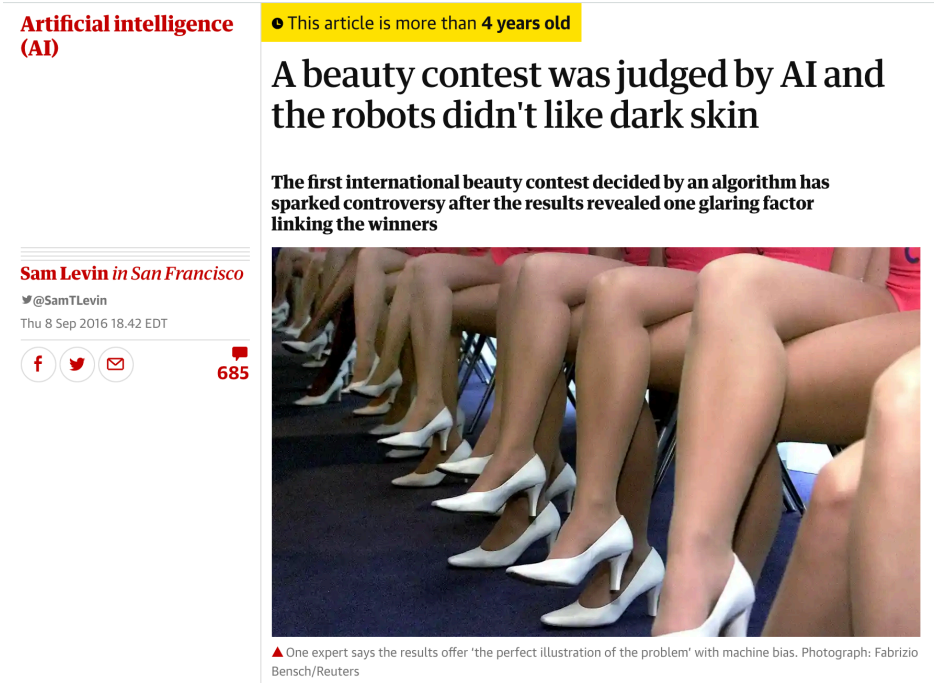
3. What are some things that Facebook could have done differently to handle this situation more ethically?

4. Who exactly should be held morally accountable for any harms caused by this study? The data scientists employed to analyze and publish the data? Mark Zuckerberg? All Facebook employees? Facebook stock holders? Are Facebook users accountable at all?

Machine bias: Beauty contests & recidivism

Machine learning (ML) algorithms are developed by 'training' models on known datasets. The models are then used to predict values in other datasets. For example, if you label cars in a batch of photos then use them to train a ML model on that labeled dataset, you can then use that model to identify cars in thousands of other photos.

Sounds neat, but this means that ML models are only as good as the data they are trained upon, and often those training datasets are created by human labelers who carry unkown or unspoken biases. A classic example is the Beauty.AI beauty contest that occurred in 2016. A ML algorithm was trained on a large set of human-labeled photos of women, then women around the world were invited to submit selfies in a global beauty contest. A key advertising hook for the contest was that the 'robot jury' – the ML algorithm – would be fully impartial and fair. But the results revealed that the ML model was racist: 75% of the 6,000 contestants were white, but 98% of the 44 winners were white. How did this happen? The people who labeled the training set of photos carried implicit bias.



Artificial intelligence (AI)

This article is more than **4 years old**

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

Sam Levin *in San Francisco*

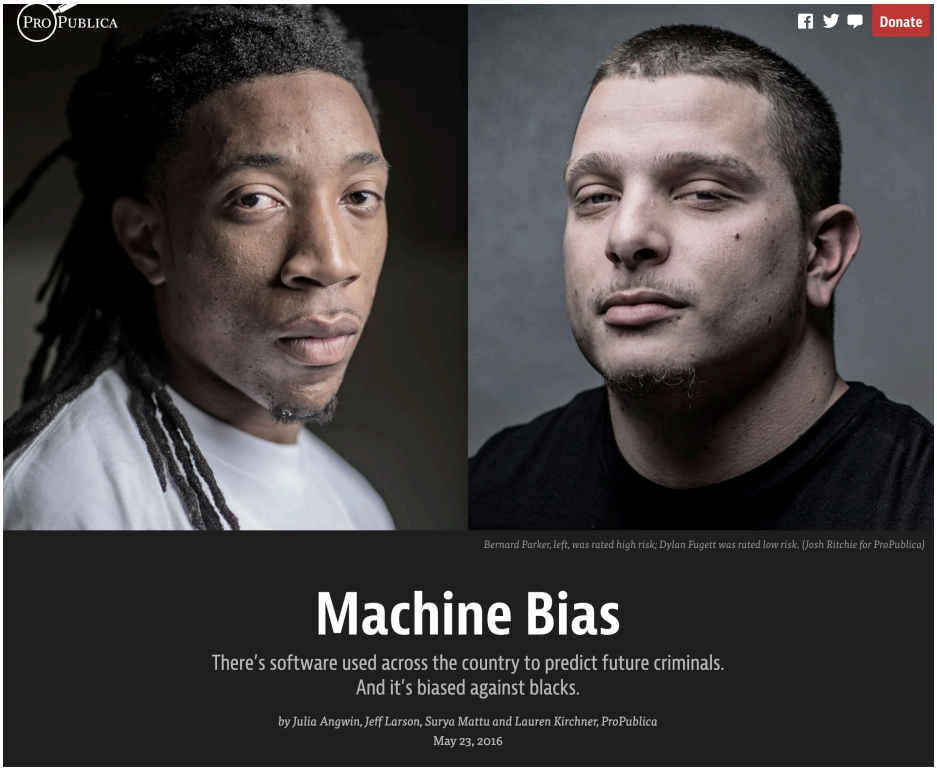
@SamTLevin

Thu 8 Sep 2016 18.42 EDT

685

One expert says the results offer 'the perfect illustration of the problem' with machine bias. Photograph: Fabrizio Bensch/Reuters

Debacles like this can have much more serious consequences. Court systems use ML models to estimate the risks that a convicted criminal will commit more crimes once they are released from prison. But retrospective studies have shown that these models consistently and incorrectly label black prisoners as more dangerous and more likely to return to prison at a later date. Most of the risk assessments being used today have not received adequate validation, even though they are spitting out predictions that can destroy lives, families, and communities.



Discuss:

1. How might bias have entered the training datasets for these ML models, if the people labeling the data did not deliberately intend to exhibit prejudice against African Americans?

2. The ML models used to predict recidivism are imperfect and inherently prejudiced, but it is not clear whether it would be better to leave decisions of sentencing, bail amounts, and prisoner support services to individual humans in the court and penal systems. Would it be better to stop all use of ML evaluations, or should they be kept in place until better models are developed?

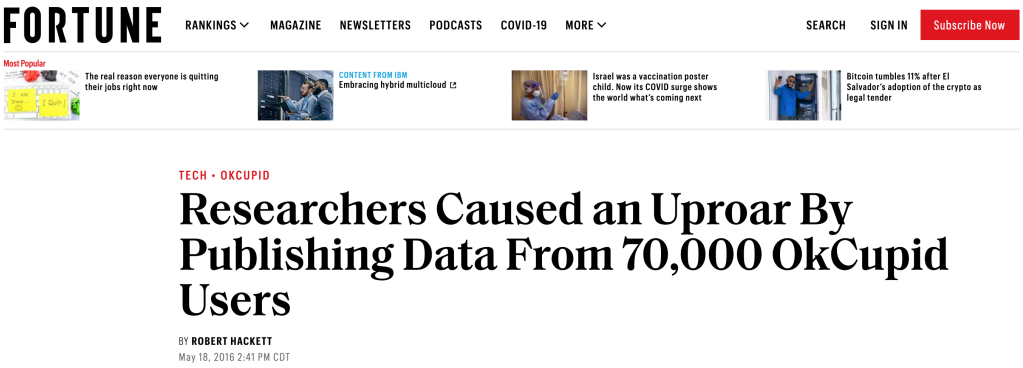
3. Returning to the beauty contest debacle. The attraction of a 'robot jury' compelled people to seek out a single, simplistic definition of beauty, and to place all contestants on the same spectrum of beauty scores. Other than the racial bias baked into their model, what other problems is there with this endeavor? Articulate and enumerate as many as you can. What do those problems tell you about other ethical and humanistic dangers inherent to data science?

Web-scraping OKCupid

In 2016, Danish researchers used new web scraping and text mining software to inventory the user profiles of 60,000 users on the online dating site OkCupid. Their goal was to use this dataset to test for correlations between 'cognitive ability' and sexual orientation, religious affinity, and other personality traits. They chose these user profiles because they were publicly available online to anyone who wanted to sign up for a free account with OKCupid.

When they published their paper, they included a spreadsheet of the 60,000 user profiles in the supplementary material for their article. They had removed the first and last names of the users, but kept everything else, including username, location, sexual orientation, religious orientation, sexual habits, relationship fidelity, drug use history, political views, and more.

The backlash was immediate. When asked why they did not attempt to deidentify or anonymize the data any further, the researchers responded that the data were already public. In the paper itself, the authors wrote: "Some may object to the ethics of gathering and releasing this data … However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely represents it in a more useful form."



Discuss:

1. What do you make of the authors' argument, that the data were already public, posted freely by the users themselves, so how can this be an issue of privacy or consent?
2. If you disagree with the authors' argument, explain how the users might reasonably object to the authors' actions.
3. What kind of risks did the authors impose upon the users of OKCupid? Are any of those risks new, or were they all present when the users decided to make a profile that could have been accessed by any other user?
4. Does it make an ethical difference that the authors accessed publicly available data in a novel way (web-scraping software) and to a much greater extent (harvesting 60,000 profiles at once) than individual users are typically able to do?
5. Does the software developer of the web-scraping tool bear any responsibility for this scandal? Does he have any ethical obligations regarding how his tool is used, and by whom?

More scenarios

Airport screening
A data scientist has come up with a model that prioritizes security screening at airports according to various passenger characteristics. If using 'place of origin' as a predictor in this model improves the model's predictive performance at identifying passengers who are security threats, is it ethical to include that variable in the model and screen certain passengers disproportionately?

Supporting struggling college students
Your college has developed a model that predicts dropouts. It identifies students at high-risk of dropping out and alerts offices that can direct additional support and resources to these students. Your college has found that this model performs better when it includes the student's place of origin, sex, and race as predictors. Is it ethical to implement this model and divert resources accordingly?

Robot cashiers
Self-checkout stations in grocery stores are convenient, but they take jobs away from workers who may not have many other employable skill-sets. When you are checking out, is it more ethical to use the checkout aisles with human cashiers, even if the line is longer and going more slowly?

Electric cars
Cars running on fossil fuels are bad for the environment, but at least they can be serviced by car mechanics who don't necessarily need a college degree.

Electric cars reduce carbon emissions, but they are replacing car mechanics with computer scientists and softwar engineers, all of which require extensive undergradate and post-graduate education. Are electric cars a net social good?

Smartphone app for monitoring cough
A tech start-up has developed an app that can track the prevalence of cough in a network of smartphones. Cough is an important indicator of disease, and cough also helps to spread certain diseases more quickly, such as TB and COVID-19. This app has great potential to help public health officers in the fight against some of the deadliest respiratory diseases. The app works thanks to sophisticated machine learning algorithm for detecting coughs within continuous recordings. That algorithm is currently private and propriety. Do you agree that this cough monitoring app is a good idea, and that public health officials should promote its use?

Automated suicide prevention system
A large internet search engine has developed a model that can predict whether someone is likely to inflict self-harm or attempt suicide based upon their recent search history. This model is 75% accurate. This company would like to set up an automatic emergency alert system, in which local social service providers are notified about at-risk users in their area. They want to automatically enroll users in this service. Is this an ethical feature to add to their product?

Malaria medicine distribution
Your company is trying to distribute a new malaria medicine in a remote region of Africa without primary care clinics, where tens of thousands people die from malaria each year. This medicine is highly effective, but it is also known to cause birth complications. You need to ensure that it is not administered to pregnant women. Your team's plan is to go door to door and distribute the medicine to women who say they aren't pregnant.

But in this region, cultural attitudes to pregnancy, and the notion of sharing your pregnancy status with a stranger, are very sensitive. Daughters and wives may not feel safe to answer such questions truthfully.

Your team has to choose between (1) taking women's responses at their word, (2) avoiding the pregnancy issue by only distributing it to men, (3) not distributing the medicine at all, (4) some as-yet-unknown solution. What do you do?

User accountability
Let's say that you have disagreed with one or more of the claims about social media in the 'Warm-Up Scenarios' section above. Is it ethical for you to continue to use Google, Instagram, or Facebook?

Chapter 5

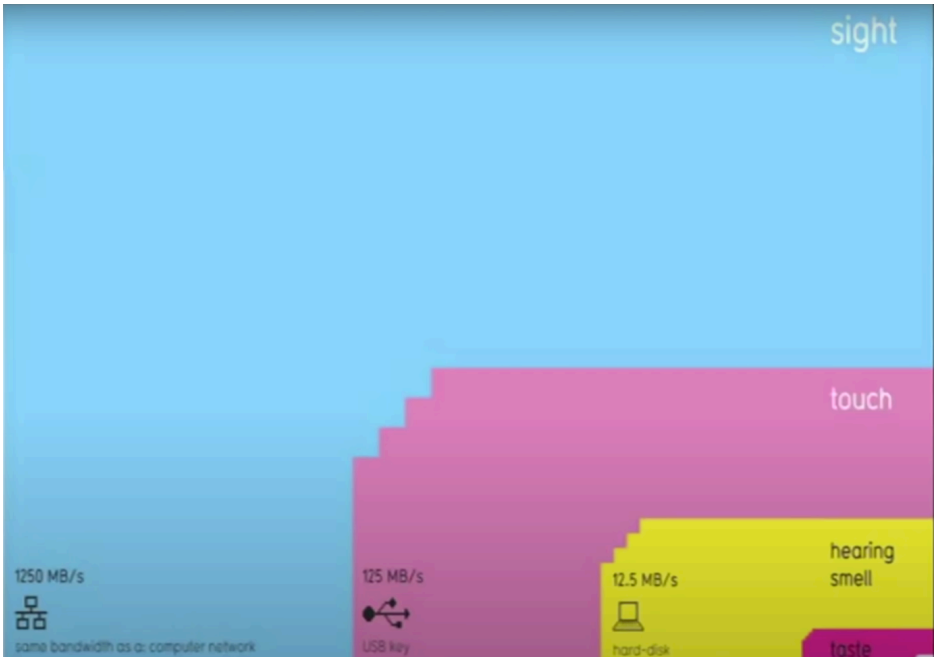
Visualizing data

This course will focus heavily upon visualizing data in plots, maps, and dashboards. If there is anything you take from this course, it will be this: you will be able to take data and make some pretty pictures. *And that's not trivial.*

Why? Because **humans are wired to process information through *pictures*.** We can translate images into meaning with amazing speed.

The value of *data viz*

This screenshot, from David McCandless's TED talk about the beauty of data visualization, depicts how quickly each of our senses can process information.



This plot's punchline: When we process data with our eyes – with *pictures* – we can take in a lot of information all at once.

Let's try this out with an example. Below this paragraph is another paragraph describing a painting. Try this: scroll down quickly, look at this paragraph for just *one second*, then keep scrolling until it is out of view.

Go!

One-second paragraph:

Fog rises from the evergreen forest of a distance mountain range. A whitewater creek cascades down a streambed with large, rounded boulders, arriving at a broad flatwater pool where ducks are milling. There is one group of four and another group of two. On the shore near the ducks, a wooden dinghy is tied up to a small dock with eight pilings. The dock lead to a path through more round peddles and tall grass, past a chair and a fire ring, and continues uphill to a small cabin. The evening sun and sparse fairweather clouds are reflected in the cabin's large, multi-paned windows under the small front porch. A cobblestone chimney on the side of the cabin has a whisp of smoke rising from it. The steep roof implies that this cabin is designed to withstand heavy snow. Tall evergreens tower over the diminutive cabin; the cabin seems to be placed up against a forested hillside. There are only a few deciduous trees in view, and their leaf colors – combined with the lack of snow in the distant mountains – imply that the time of year is early fall.

End of paragraph.

OK. Now try to answer these simple questions:

- What was this a painting of, in general? Can you describe the scene?
- What details do you recall?

OK. Try this next: the actual painting is at the very end of this chapter. Scroll down to it quickly, look at this painting for just *one second*, then scroll back up to this spot in the module.

Ready? Go!

< Return to this line! >

Now try to answer those same questions above. What was this a painting of? Did you catch any more details? Was there anything in the water? Was there smoke coming out of the chimney? What time of day was it?

Which type of visual information was easier for you to process quickly? Text, or a picture?

Think about the profound *differences* in these two forms of visual communication:

When we read text, we are working outward, from individual details to the big picture: we process each individual word, understand their individual meanings, understand their meanings in the context of each individual sentence, then use all of the information to step back and imagine the scene based on the details.

In contrast, when we look at a picture, we are working inward, from the big picture down to the details. We understand the scene first, then we start exploring the finer points. And, since each finer point is interpreted from within the context of the bigger picture, we can make sense of the details much more efficiently.

Pictures communicate data. **This is why data *science* and data *visualization* nearly always go hand in hand.**

Data scientists use visualizations both to communicate their insights externally, e.g., to the public in a *Twitter* post, but also internally: when they are working with the data themselves. A data scientist's workflow is peppered with data visualization, because – again – visualizing your data is the most effective way of making sense of it: Download the data, then visualize it. Do something to the data, then visualize what you've done. Repeat, then visualize, then repeat again.

The point here is that great data visualizations are not simply pretty. Much more importantly, they are *effective* too. They are the best means you have of conveying insights from your data to someone else.

A final thought: Keep in mind that plots can be effective and misleading at the same time. There is a politics to plots and maps; they can have agendas, and they can manipulate viewers into interpreting the data in certain ways. So, it is incomplete for us to say simply that a good plot is an effective plot. Here's a better definition: *a good plot is one that is both effective and fair.*

So, when you are viewing other people's plots and making plots of your own, keep these five rules in mind:

1. A bad plot is an ineffective one, even if it is beautiful.

2. A good plot is an effective plot.

3. A great plot is one that is both effective *and* beautiful.

4. If you ever have to make a trade-off between effectiveness and beauty, sacrifice beauty.

5. Any plot that misleads or manipulates the viewer is bad, no matter how effective or beautiful it is.

Before you begin evaluating the plots in the gallery below, enjoy this excellent talk by the Egyptian data scientist, **David McCandless**, about the **beauty of data visualization** (link here).

Plot gallery

What follows is a gallery of plots: some good, some bad, and some ugly too. Let's use these to explore what works and what doesn't. For each plot, ask yourself three questions:

- What makes this plot good (as in, effective and not misleading)?

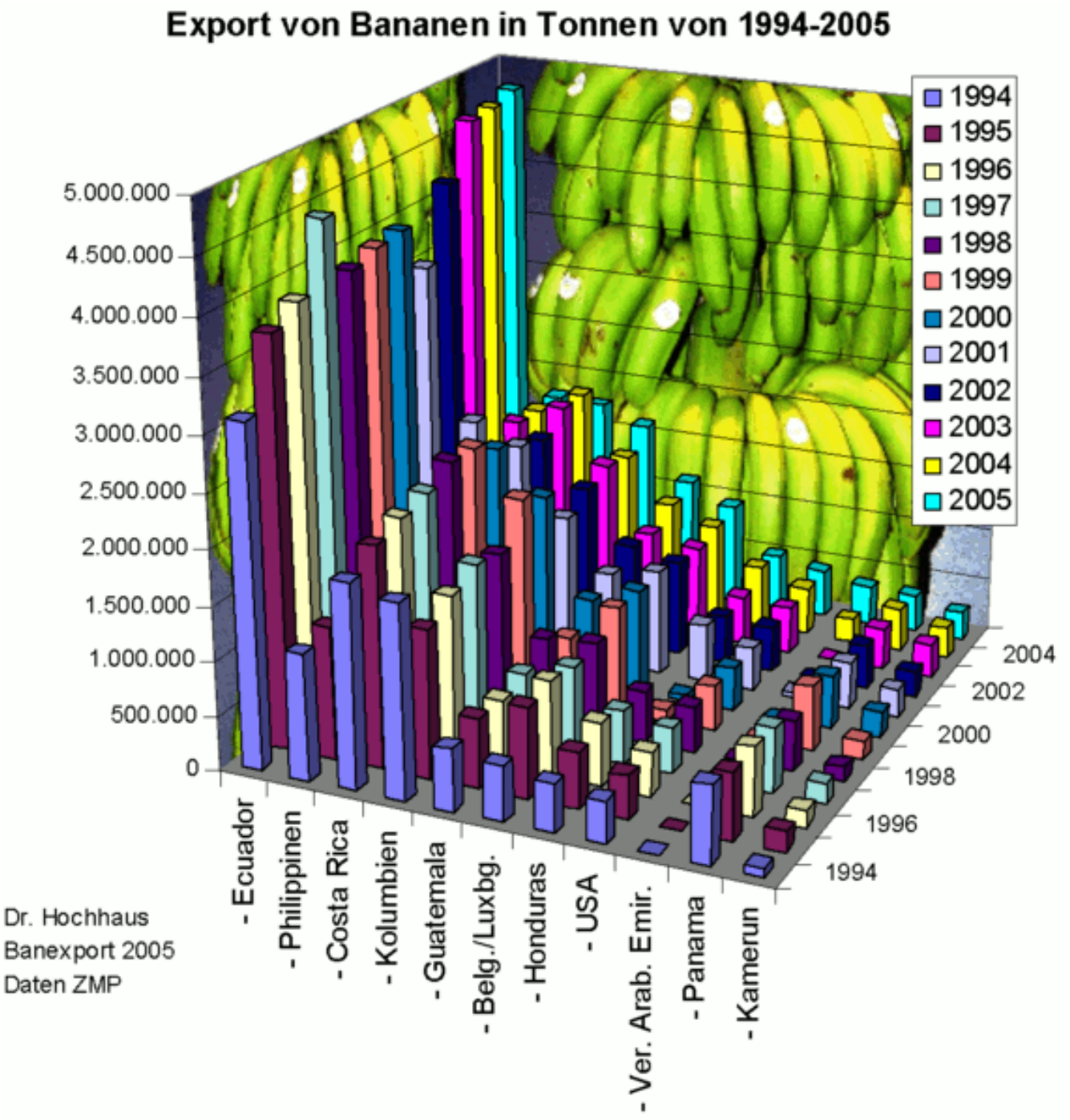
- What makes this plot bad (ineffective and/or misleading)? How could the plot be improved?

- What might make this plot prettier?

The point of this is not to make fun of others for their plots. The point is to learn from their choices. Because plot technique matters. Data science is about communication, action, and impact. You will spend so much time working on an analysis, and you are gonna go through all the work of making a plot. What a shame if the end product undermines all of that hard work!

Instructor tip:

Let the students offer feedback on each plot before you fill in the gaps with your own opinions and the ideas offered below each plot.

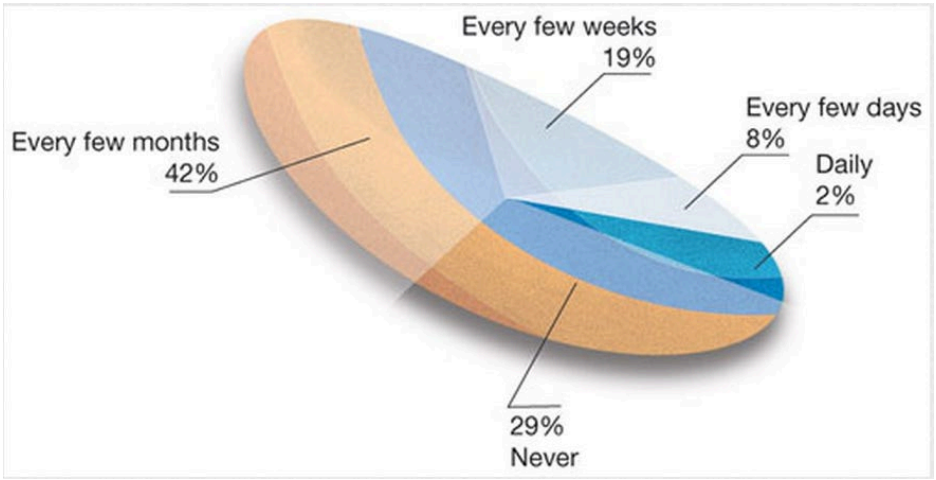


- **Good:** Frankly, there is not much good about this plot. Yes, it has a lot of information, but this crosses the line into information overload. It is so convoluted and difficult to interpret that we quickly lose interest in spending time exploring its details.

- **Bad:** The 3D perspective (1) makes it almost impossible to compare bar heights, (2) causes alot of the bars in the back to be hidden, and (3) adds needless complexity.

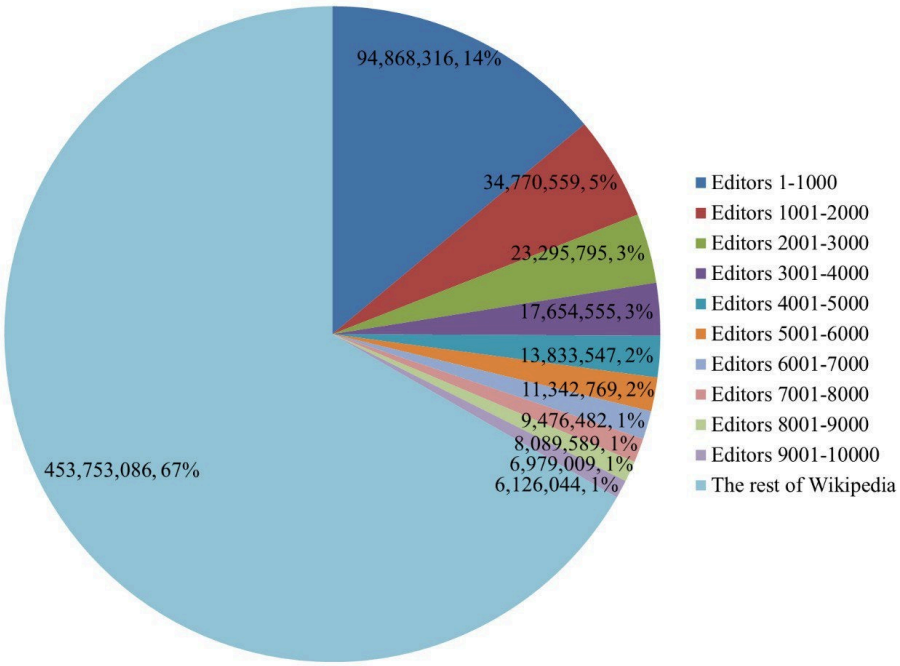
- **Bad:** The 3D perspective makes it almost impossible to compare bar heights.
- **Bad:** The colors representing each year do not follow a logical sequential flow; years are sequential, and colors can be too (think the ROYGBIV rainbow sequence).

- **Ugly:** The bananas! Sure, this plot has to do with banana exports, but those banana pictures don't represent anything at all about the data and they make everything else convoluted. Plus, it's cheesy.



We don't think this plot is good or pretty.

- **Bad:** This is an unfamiliar plot format; is it a pie chart? A blood platelet? A pickle?
- **Bad / Ugly:** Why is it lopsided or rotated? That has nothing to do with the data.
- **Bad:** What is this plot even about? There are no context clues whatsoever. Titles and labels, in moderation, can be really helpful.
- **Bad / Ugly:** What are the colors representing? They seem to have no relation at all to the pie slices. Very confusing.
- **Bad:** The slices do not seem to represent the percentages accurately. The 8% slice does not look four times larger than the 2% slice.



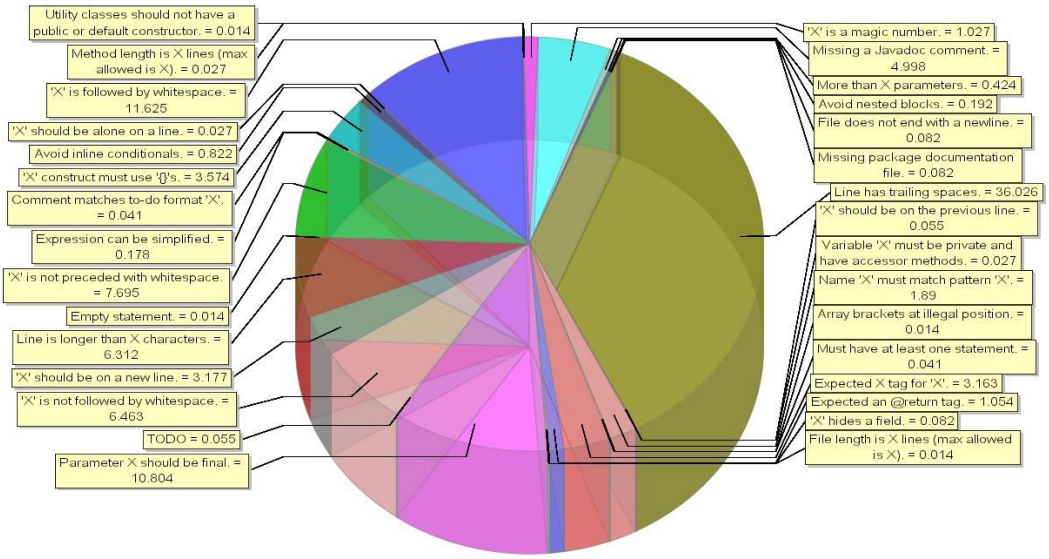
Here is another pie chart that isn't very effective.

- **Bad:** Lots of significant digits in these number. Instead of forcing viewers to read numbers like 453,753,086, why not display 454 M?

- **Bad:** The percentages next to the other numbers make it even harder to read.
- **Bad:** Superimposing text on top of the pie sclices makes it impossible to use the slices for their intended purpose: visually comparing the size of subgroups in the data.
- **Bad:** There are so many color-coded slices that it takes far too long to understand the details.
- **Bad:** One reason it takes so long is that the text is redundant: don't put a lot of text on the pie *and* put a lot of text in your legend. Figure out a way to point to each pie slice with a line, then have all the info for that slice in the same spot.

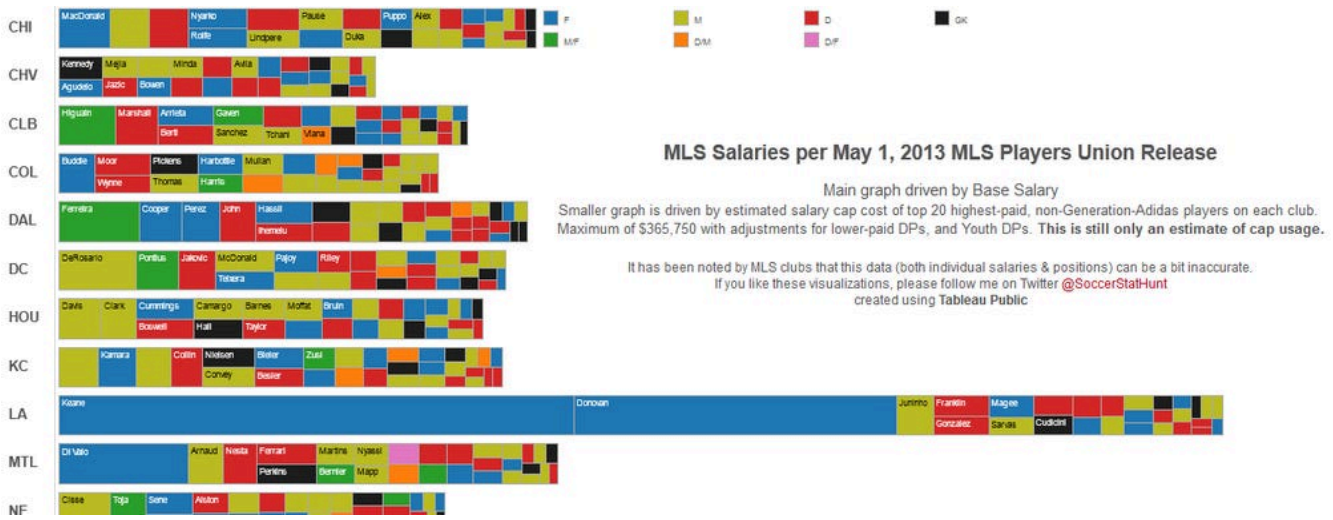
- **Bad / Ugly:** The dark text on top of dark colors is hard to read.

Pie charts are *super* common in media, but they are actually an infamously bad form of data visualization. That's because the human eye is much worse at comparing *areas*, such as the size of a pie slice, than they are at comparing *heights*. To make matters worse, we are worse at comparing areas for non-rectangular shapes, like a pie slice, than for squares or rectangles. So: avoid pie charts.



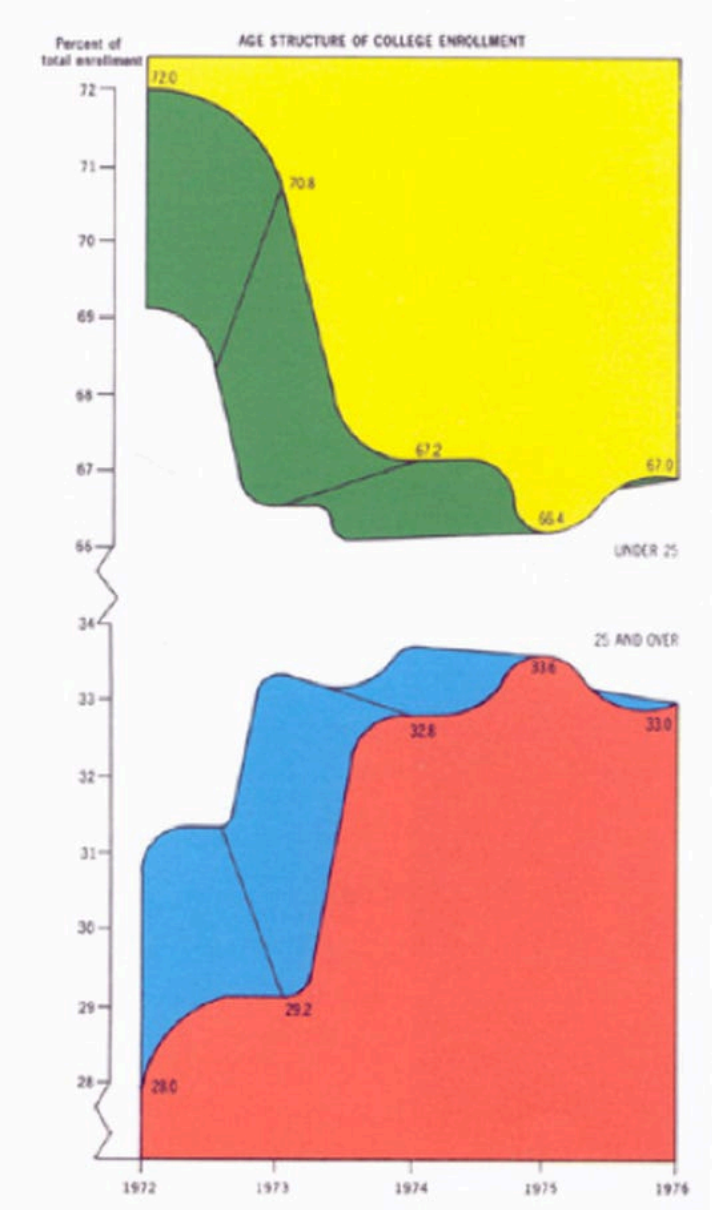
- **Bad:** Pie chart.
- **Bad:** There is no reason for this to be 3D. The third dimension has nothing to do with the data.
- **Bad:** There is no reason for this to be semi-transparent. It just makes everything even more convoluted.
- **Bad:** The text is way too small to read.
- **Bad:** On a related note, there is way too much text.

- **Bad / Ugly:** The yellow text boxes and dark lines around them add uninformative junk to this plot. If all lines unrelated to data or labels were removed, this chart would be more intelligible.



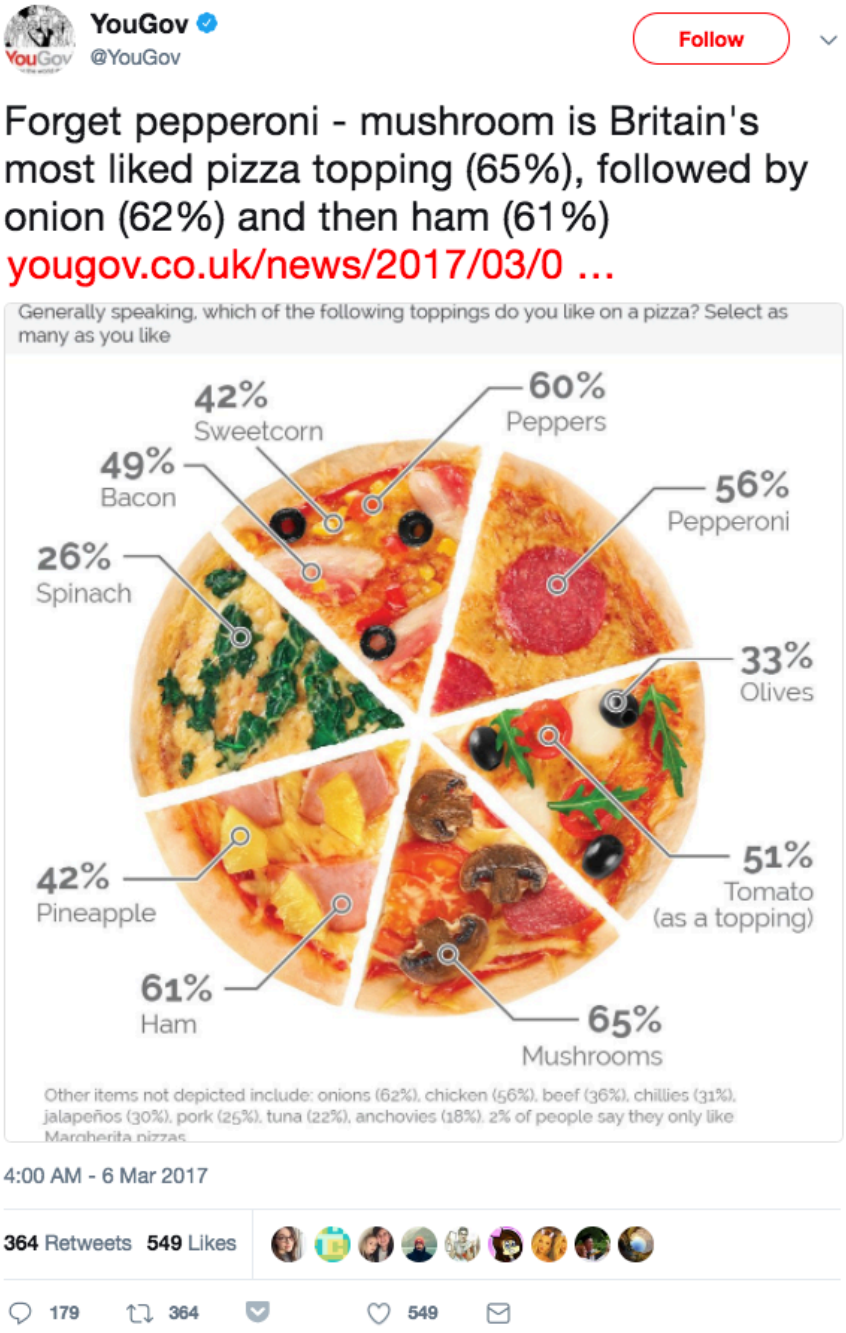
- **Bad:** Information overload.
- **Bad:** It is not clear whether the repeat use of colors in each row conveys any meaning, or if it is just a random recycling of colors.
- **Bad:** The text is too small to read.

- **Bad:** Abbreviations are not explained.



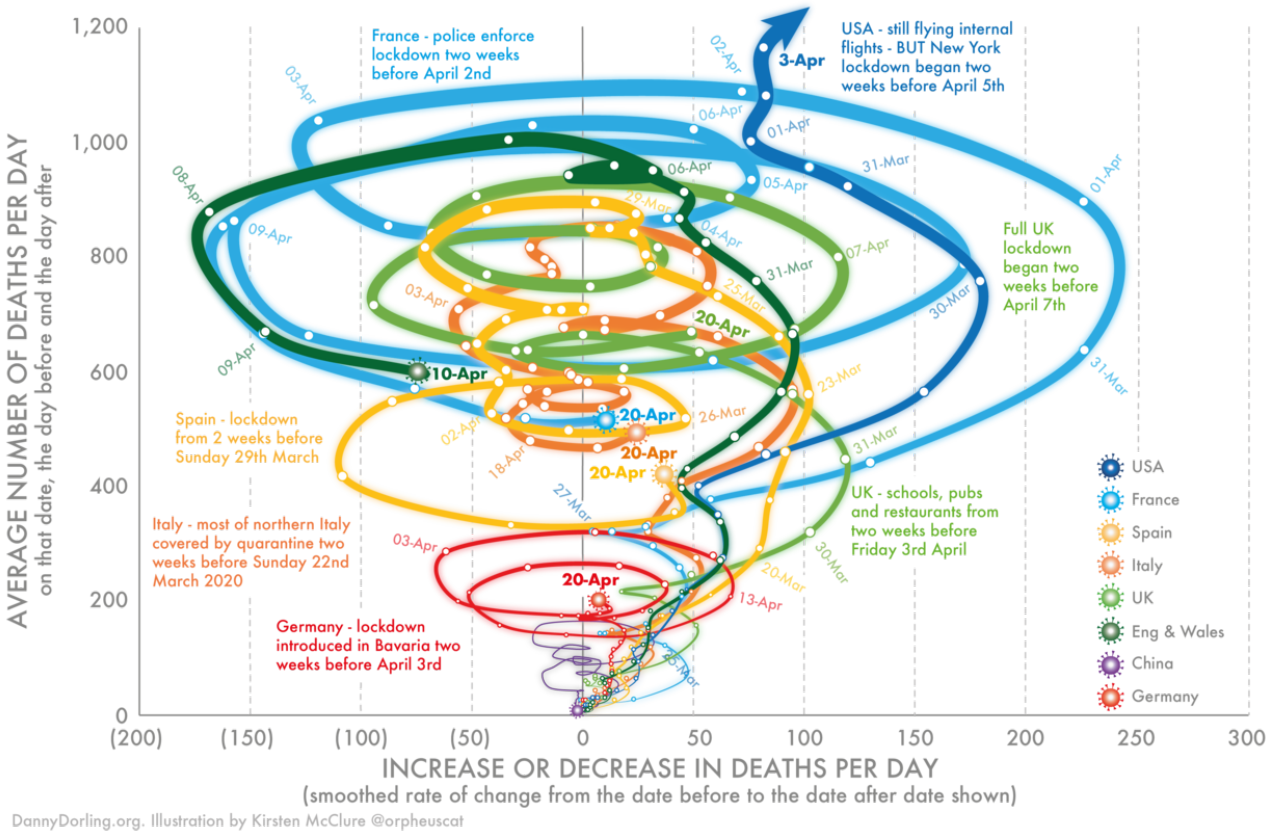
- **Bad:** The text is very small.
- **Bad / Ugly:** The third dimension, with a weird perspective effect added, has nothing to do with the data and makes this plot difficult to understand. Should you pay attention to the edge in the distance or the line in the foreground? Are they the same?

- **Bad:** The y axis is crazy! (1) The scale break is confusing. (2) The attempt to plot these two subgroups as separate trends belies the fact that one is just the remainder of the other: the two curves sum to 100%. (3) By attempting to place the two trends on the same proportional scale, this plot gives the visual impression that the changes over time are really extreme.
- **Bad:** The trend lines are unnecessarily curved. The plot's authors probably only have data for each semester, but smoothed lines give the impression that they have more data.

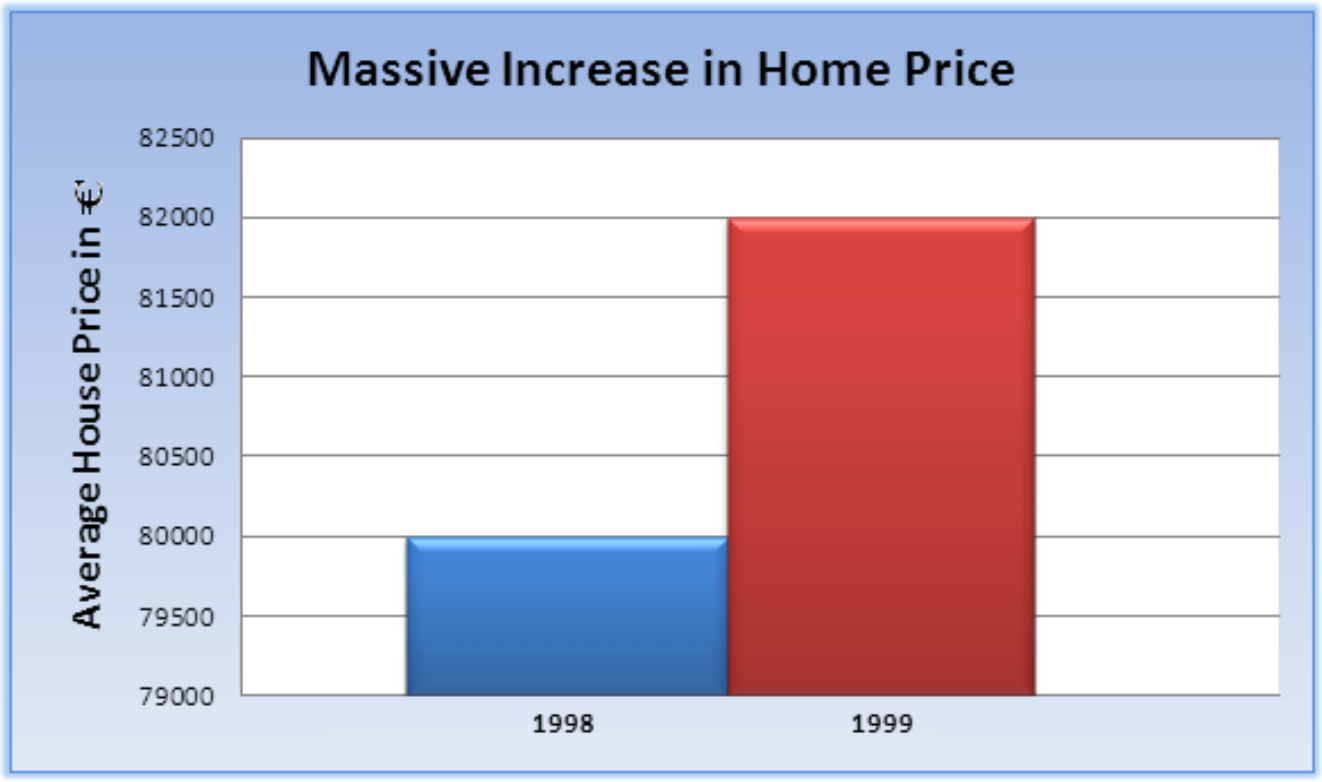


- **Bad:** These percentages sum to more than 100%. That needs to be explained, or avoided.

- **Bad:** The use of a pizza pie, though creative, implies that the size of the pie slices will have something to do with the data. They don't.
- **Bad:** Check out the fine print on the bottom. Several toppings were left off this chart, and some of them were quite popular; for example, onions had 62% popularity and chicken was 56% popular. Why are they not on this chart but spinach (26%) is? The arbitrary exclusion of categories should immediately make viewers suspicious: how are these decisions being made?



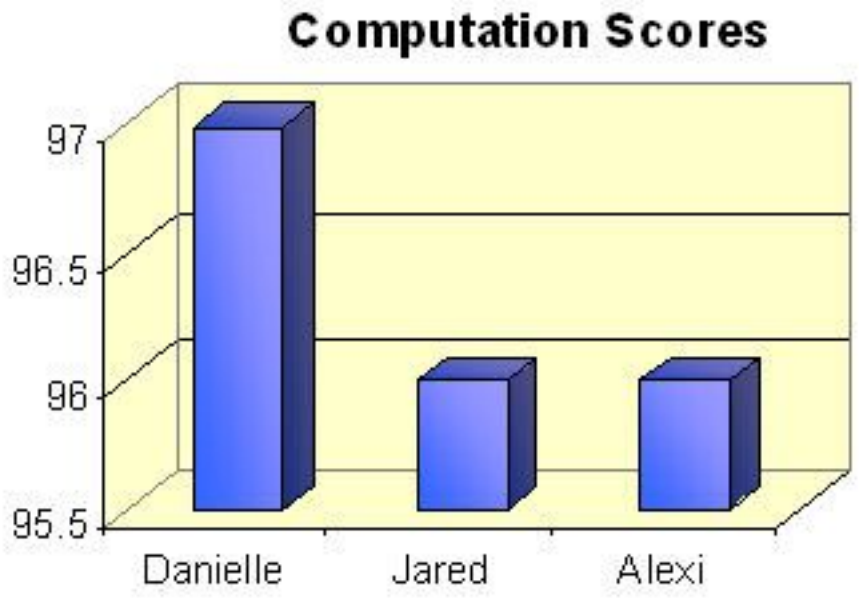
This is another case of information overload, and another case of a creative visualization that doesn't really help us make sense of what's going on. If we wanted to understand the take-a-way message or punchline from this plot, we're not sure we'd be able to. We could stare at this for minutes and still not understand what we are looking at, and it is so complicated that we would rather ignore this plot than go through that effort.



- **Bad:** The y axis range makes the difference in home prices sound a lot larger than it actually is. Based on the height of these two bars, you would expect the 1999 price to be 3x the the 1998 price when, in actuality, it is larger by 2.5%.

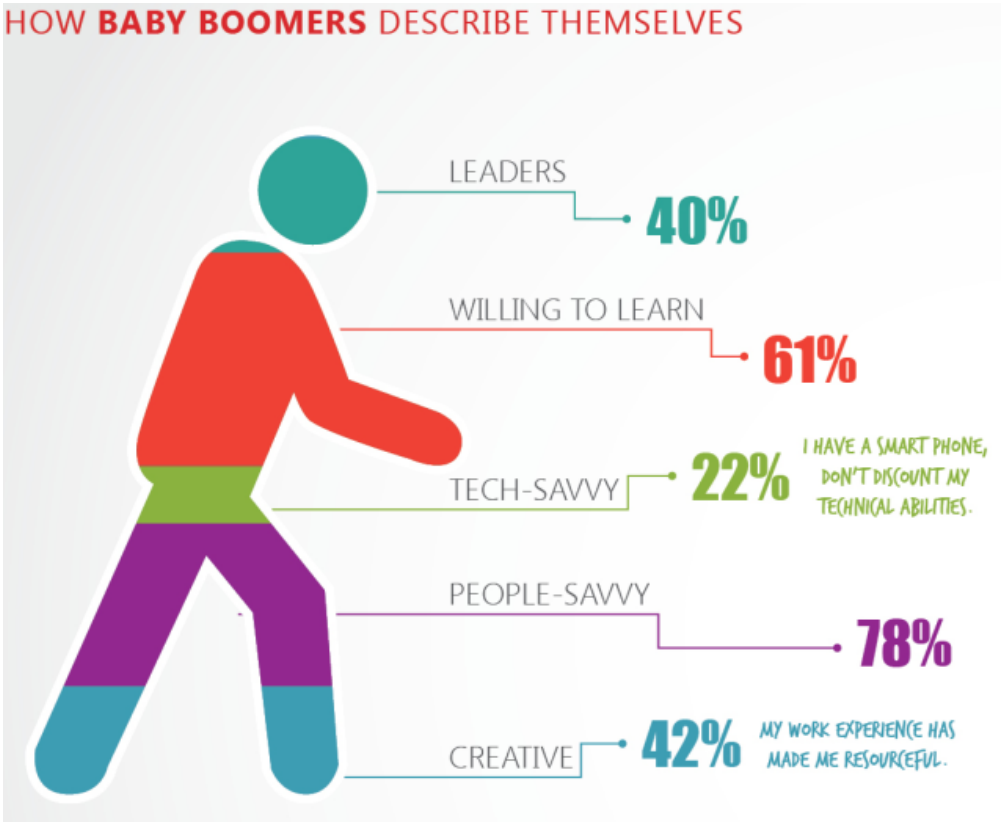
- **Bad:** The use of qualifiers such as "Massive" is generally discouraged. It's a little too controlling. Let the viewers make up their own minds about which differences are substantial and which are trivial.

These two notes are classic indicators of misleading or agenda-driven data visualization.



This is another example of manipulating perception with a sneaky y-axis range.

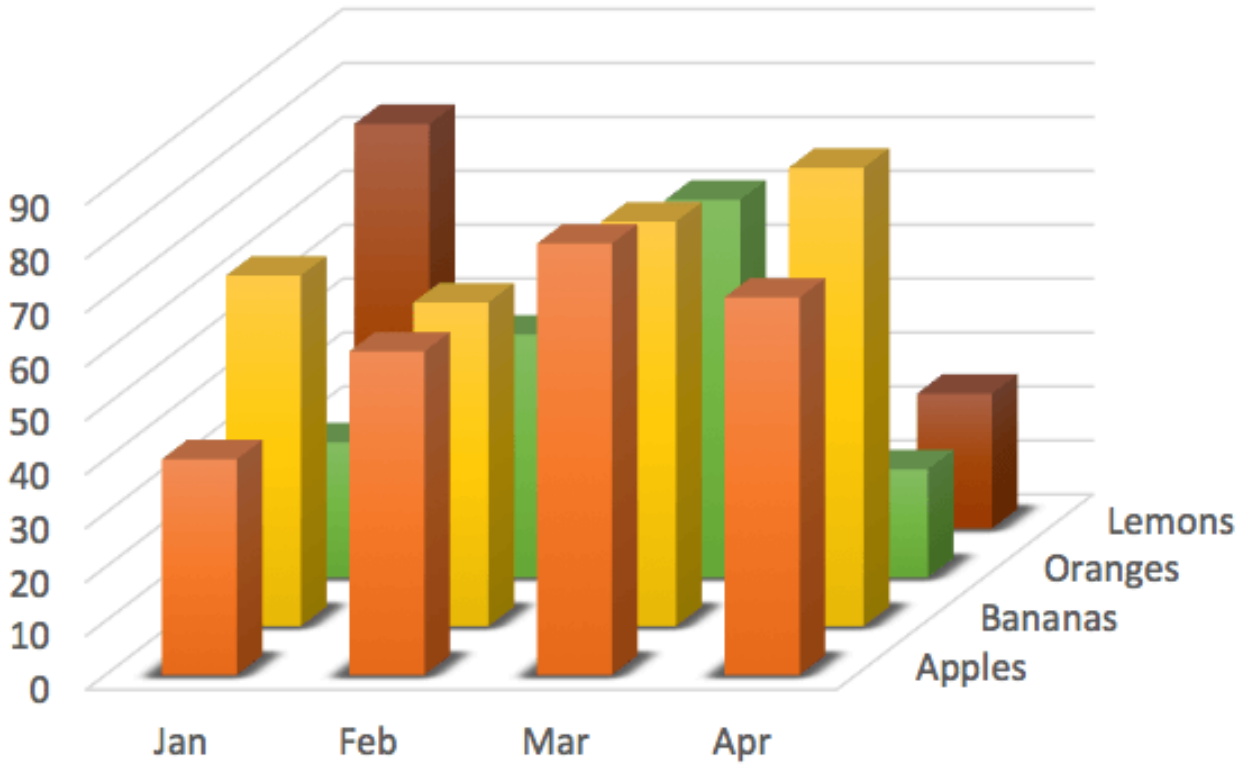
Also, like the examples above, this plot doesn't have to be 3D. The height has a meaning, but what does the depth mean? It is junk, in the sense that it does not add meaning.



- **Bad:** These percentages sum to more than 100%. That needs to be explained, or avoided.

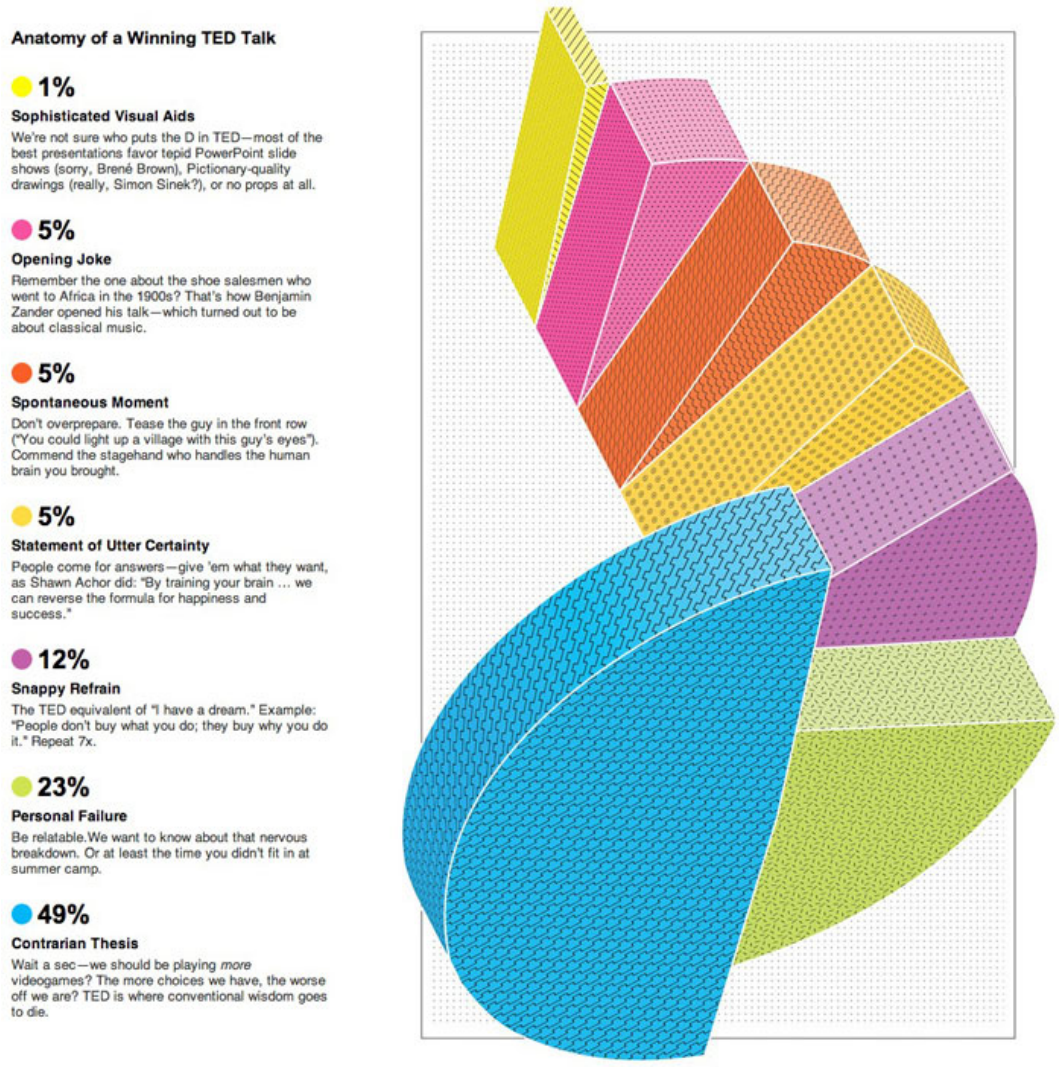
- **Bad:** The rationale for ordering of the categories is not clear. They are not ranked by percentage. Are they supposed to correspond to the body part being pointed to in the figure? Should crotch equal 'tech-savvy'?

- **Bad:** The geometric shape itself (the person figure) does not help at all in visualizing the percentages. Can you compare the area of the feet and the area of the head? The fact that they provide the actual numbers in large bold font suggests that they knew the figure would not be useful as a picture.



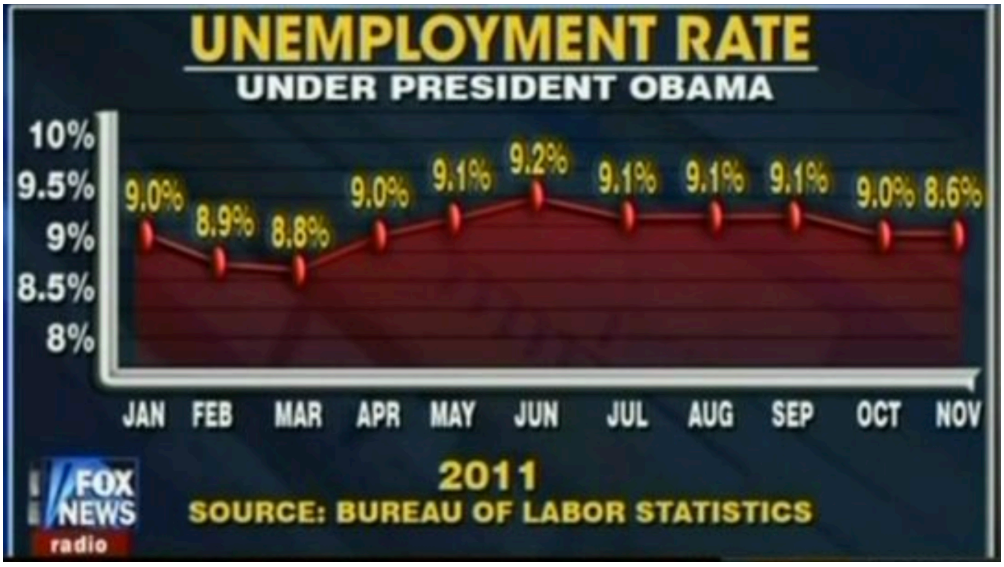
- **Bad:** Some data are completely obscured. What is happening to lemons in February and March?

- **Bad:** Like the infamous banana plot above, the 3D perspective here makes it almost impossible to compare bar heights and adds needless complexity.

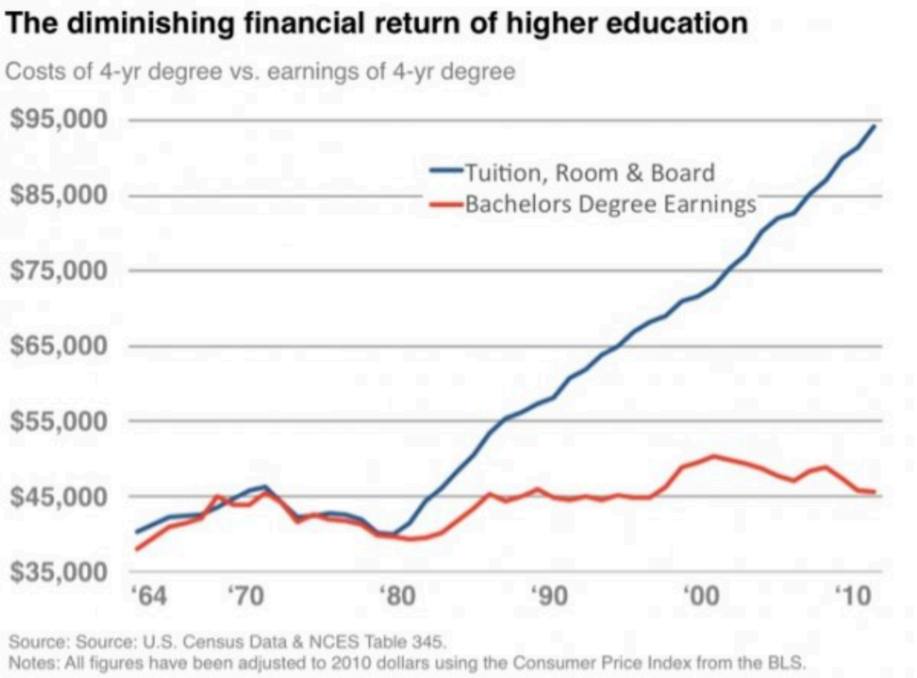


- **Bad / Ugly:** The contortion of this pie chart into a spiraling 3D object is confusing and gratuitous.
- **Bad / Ugly:** Some colors are very similar to each other.

- **Bad:** Pie chart!

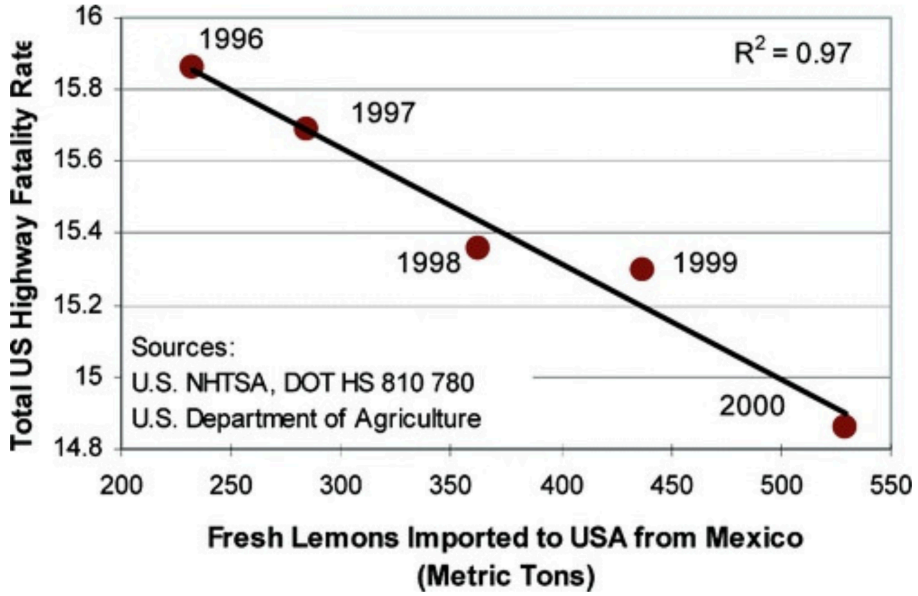


- **Bad:** The small y axis range exagerrates changes in the unemployment rate.
- **Bad:** The final data point on this plot is wrong! 8.6% should not be at the same height as 9.0%.
- **Bad:** The title of this plot implies that the data will show unemployment trends throughout the Obama administration, which began in 2009, but this data is for 2011 only.

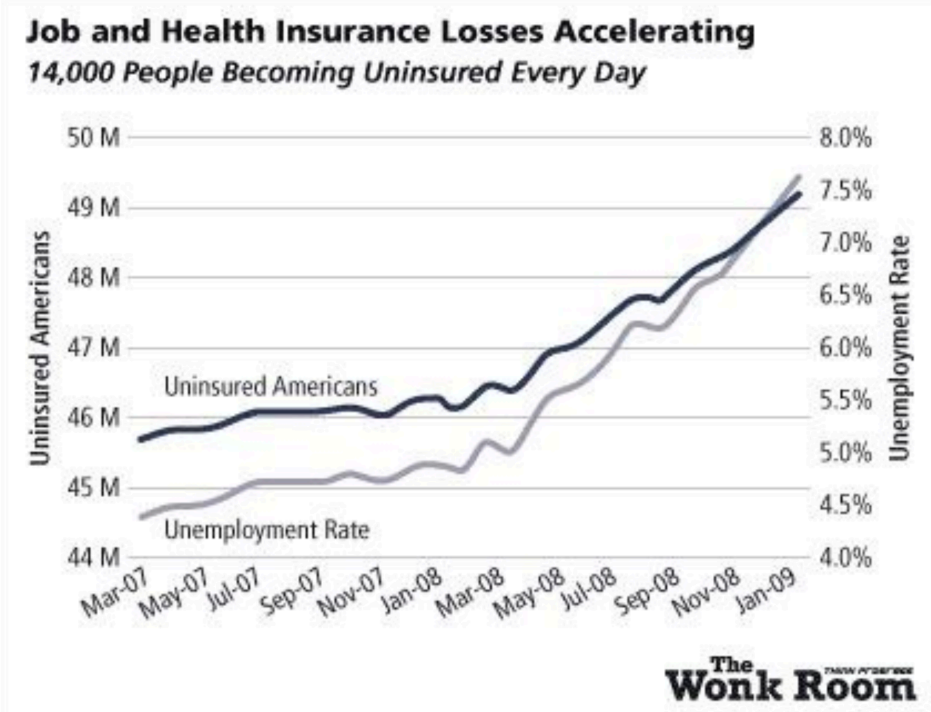


- **Good:** All in all, this is a very well made plot. Unnecessary lines and text are kept to a minimum; the title and subtitle are clear. The text size is appropriate. The source of the data is provided.
- **Bad:** This plot is misleading, because it is plotting data with different units on the same y axis. The blue line is the average *four-year* cost of a college degree. The red line is average *annual* salary for someone with a Bachelor's degree.

How should these data be plotted differently in order to correctly explore whether a four-year degree is still worthwhile in terms of its benefits to a 30-year career?

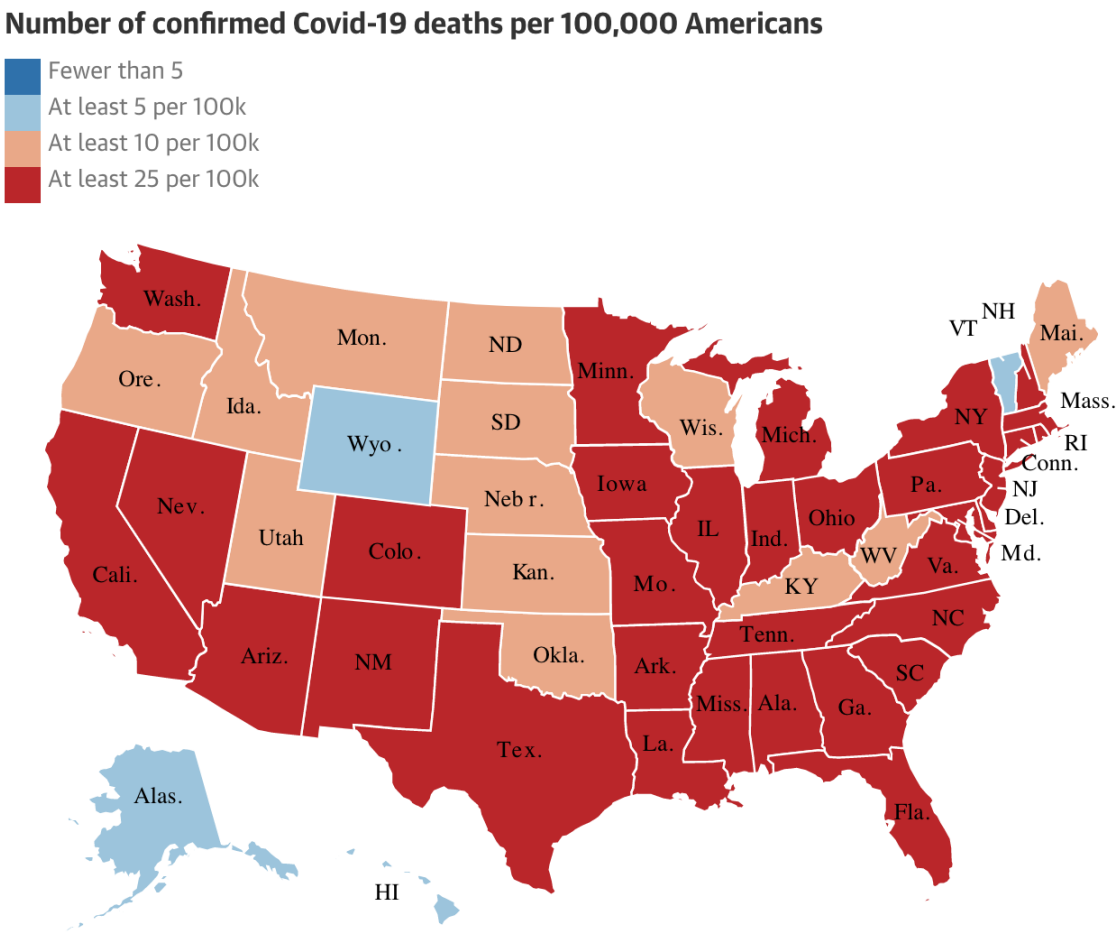


This is a classic example of a *spurious correlation*: two trends that are correlated but have absolutely nothing to do with each other.



This is another example of a beautiful plot, but it is also an example of how a plot's message can be coaxed by

manipulating y axis scales. If the curve for Uninsured Americans were plotted on the same percentage scale as the Unemployment Rate, it would not seem to be accelerating so rapidly.

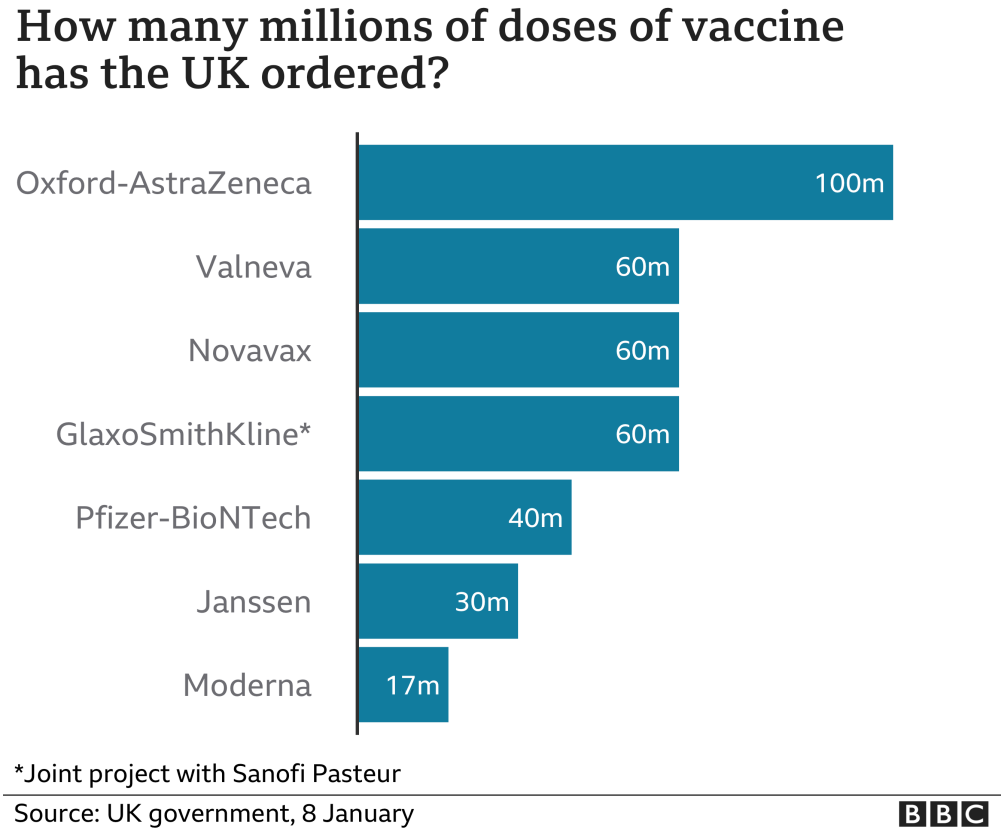


- **Good:** This is a clean and simple plot, more or less, without too much text.
- **Good:** The color scale relates to only 4 categories: that is simple enough to make sense of quickly.

- **Good:** The color palette follows an intuitively sequential trend: Blue = low/no bad; Red = high/severe.
- **Bad:** The abbreviation system for states is inconsistent. Some use initials, some use abbreviations; some use a period, some don't.

- **Bad:** The lowest color category is not used in this plot, and could be removed to increase simplicity.

- **Bad:** Showing a map of the U.S. states with this particular color scale can be confusing: if you had to guess what this chart is about, you would probably assume it is an electoral map.

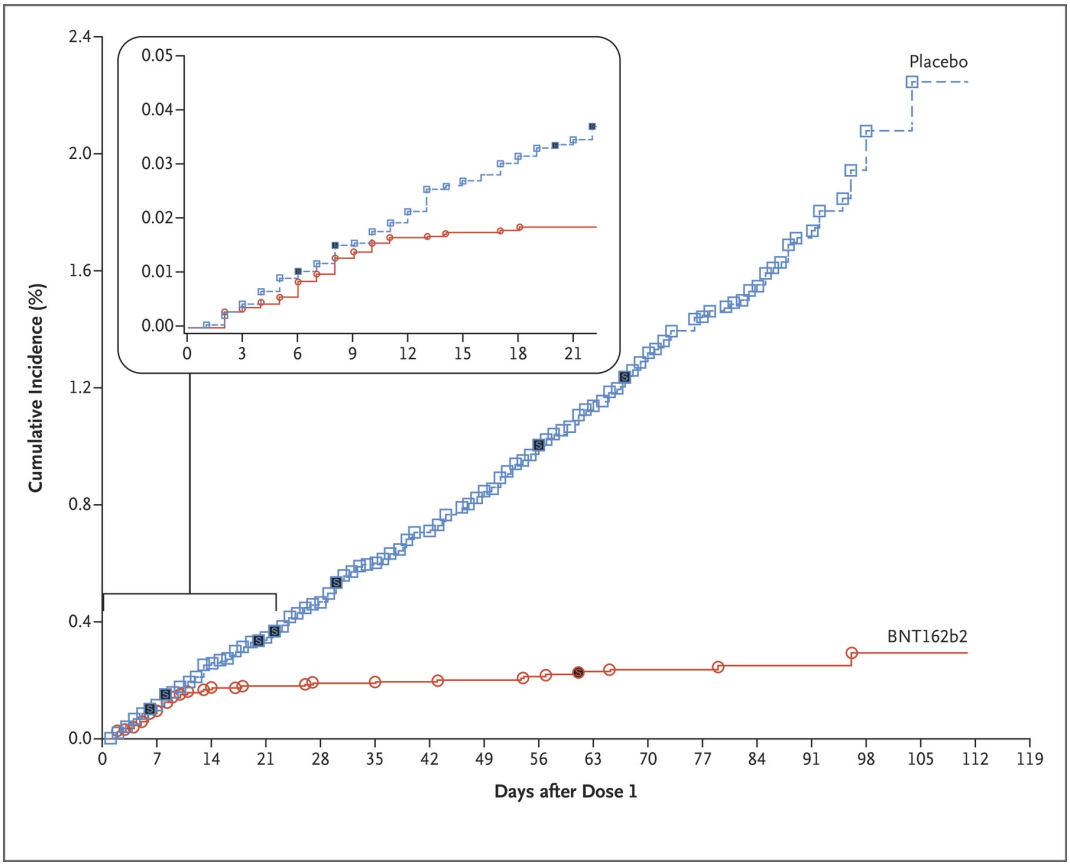


- **Good:** This is a nice plot. It is simple and junk-free. No unnecessary text or axes.
- **Good:** Large font, strong color contrast.

- **Good:** The axis starts at zero, and the bars widths are proportional to the data, e.g., the 30m bar is half the width of the 60m bar.

- **Good:** Since this plot is so simple, the actual numbers for each bar can be included without cluttering the plot.

Effectiveness of COVID-19 vaccine BNT162b2:

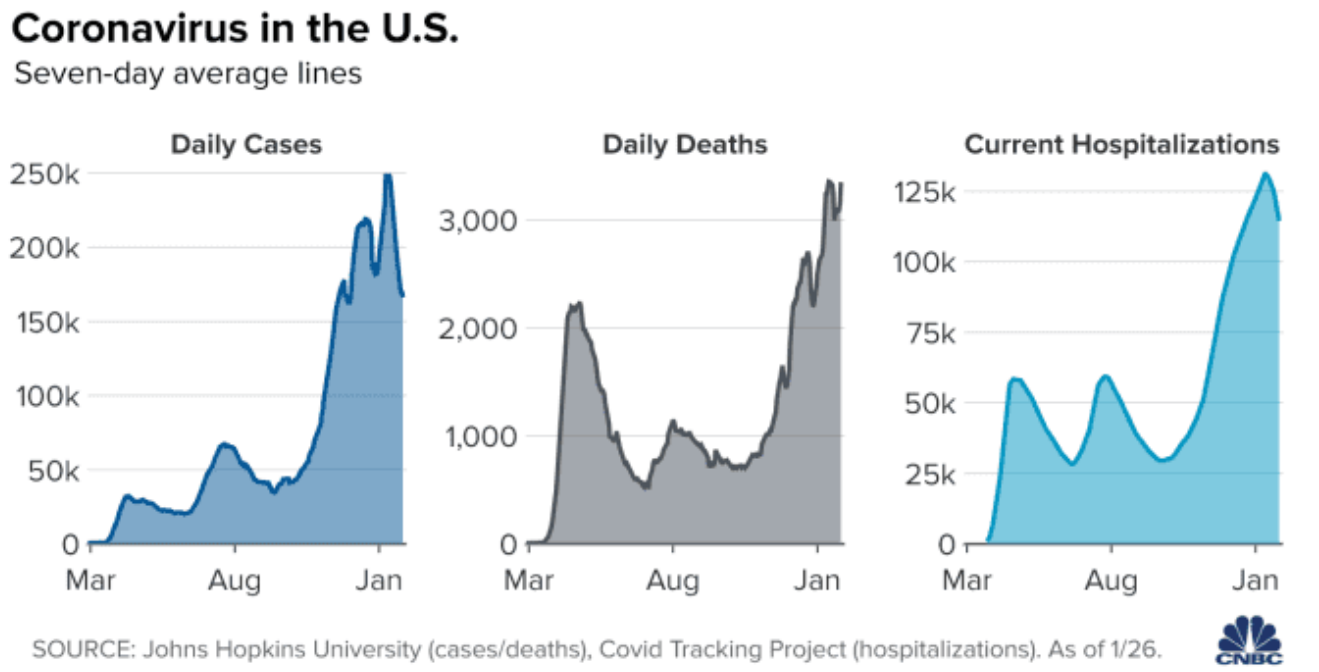


- **Good:** This is a simple plot that tells a good story: a week or so after receiving a dose, vaccinated participants contracted COVID-19 at a much lower rate than those who received the placebo.

- **Bad:** The labels are not clear to viewers who are not clinical virologists. To every extent possible, data visualizations should be inclusive and inviting. Don't make someone feel stupid by forcing them to look at their plot.

- **Bad:** The overlay that zooms in on the first three weeks makes this plot a bit cluttered. We might suggest plotting those first three weeks in a separate plot, adjacent to this one but not embedded within it.

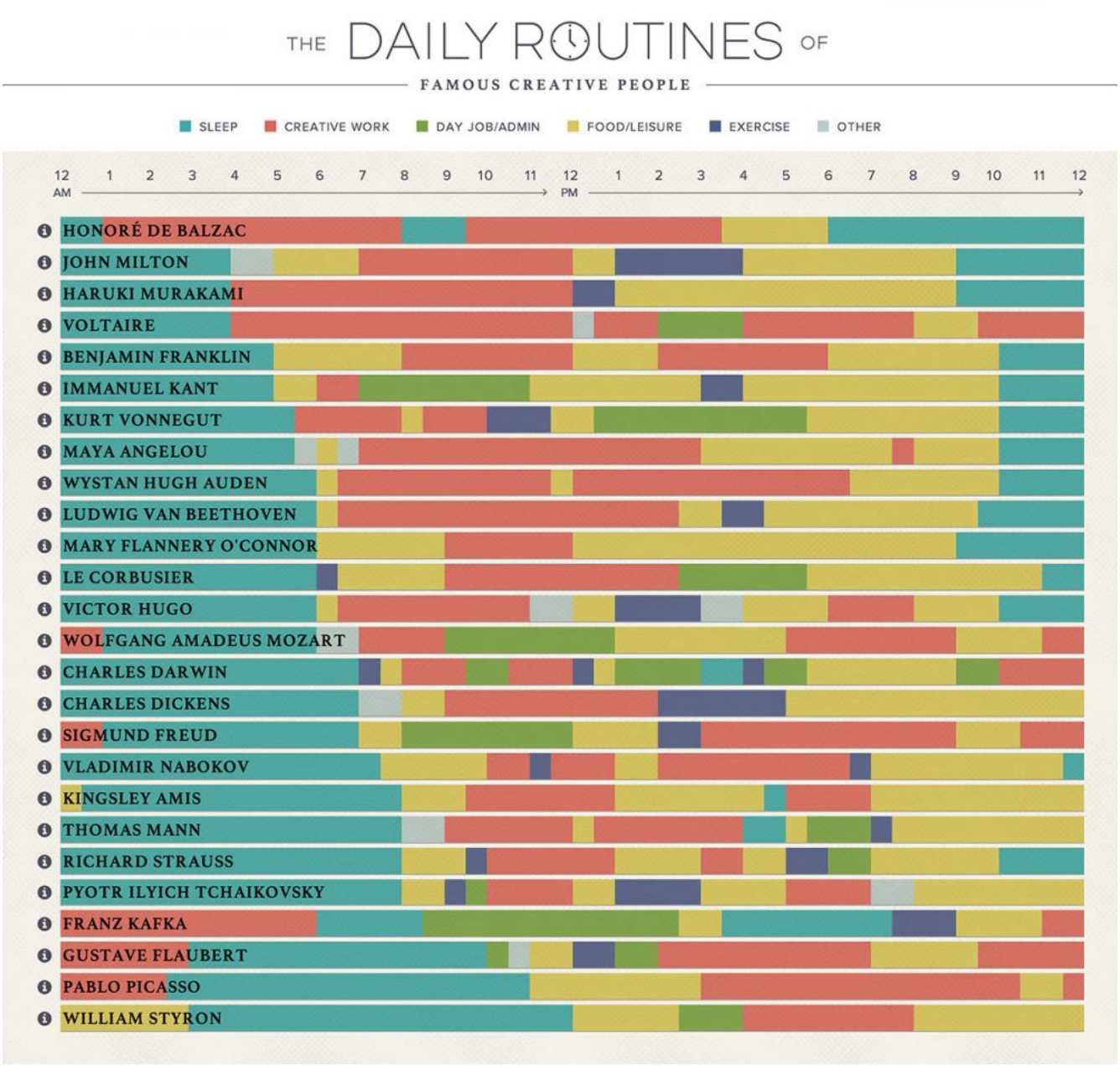
- **Ugly:** This is an effective plot, but it is not beautiful. What would you do to make this plot more beautiful without compromising its message?



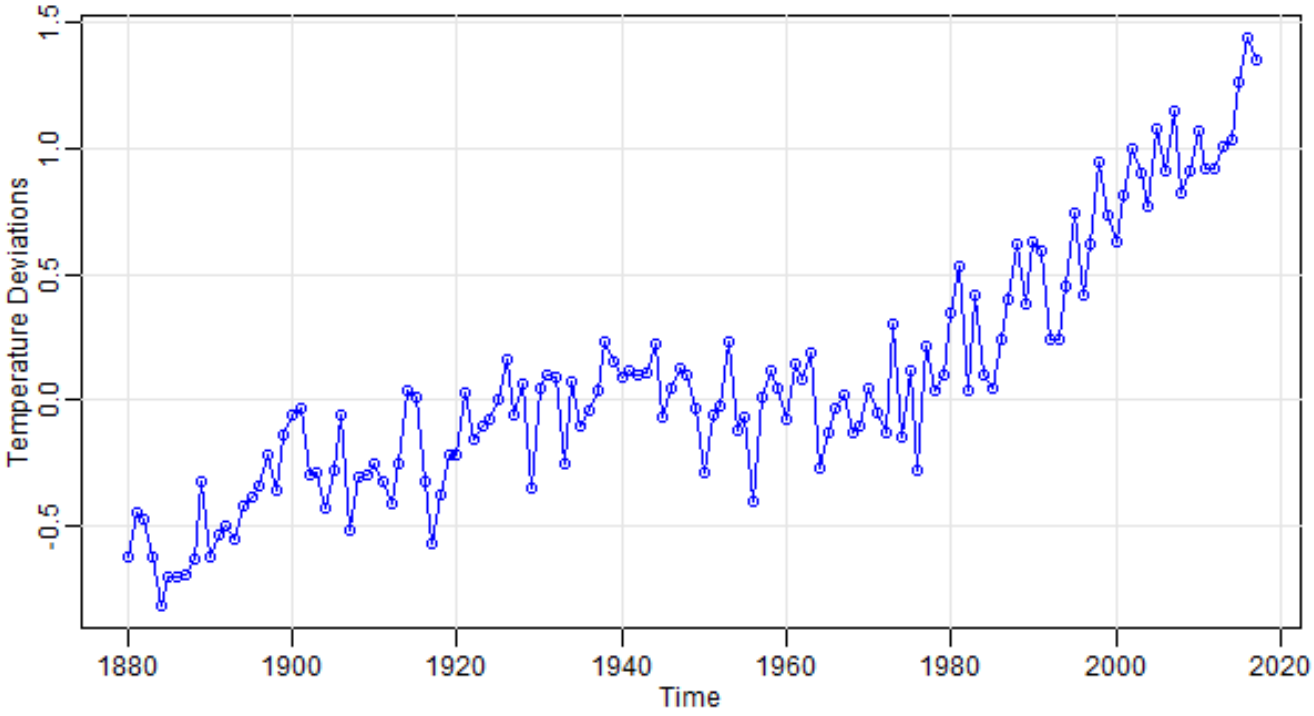
- **Good:** These are clear and simple plots with an obvious take away: daily COVID-19 case counts correspond to deaths and hospitalizations.

- **Good:** No extraneous labels or lines. These plots are chart-junk free.

What do you think about the choice to use three different scales for the y axis? That tends to lead to confusion, but do you think that, in this case, it was justified?



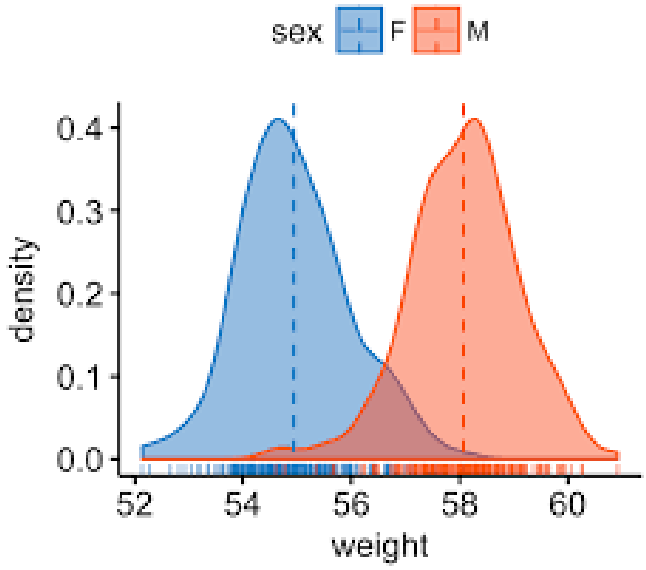
Now this is a data visualization! It is a tad complicated, but it is elegant and fun to explore.



This plot is straightforward and simple.

- What it is showing is fairly self-explanatory, though some viewers might benefit from a more informative y-axis label.
- To help explain the y axis, it may be helpful to include a helper line at 0.0 degrees.

- Are the points really necessary? Since it is fairly clear that this is an annual dataset, those points are just repeating the information contained in the line, aren't they?

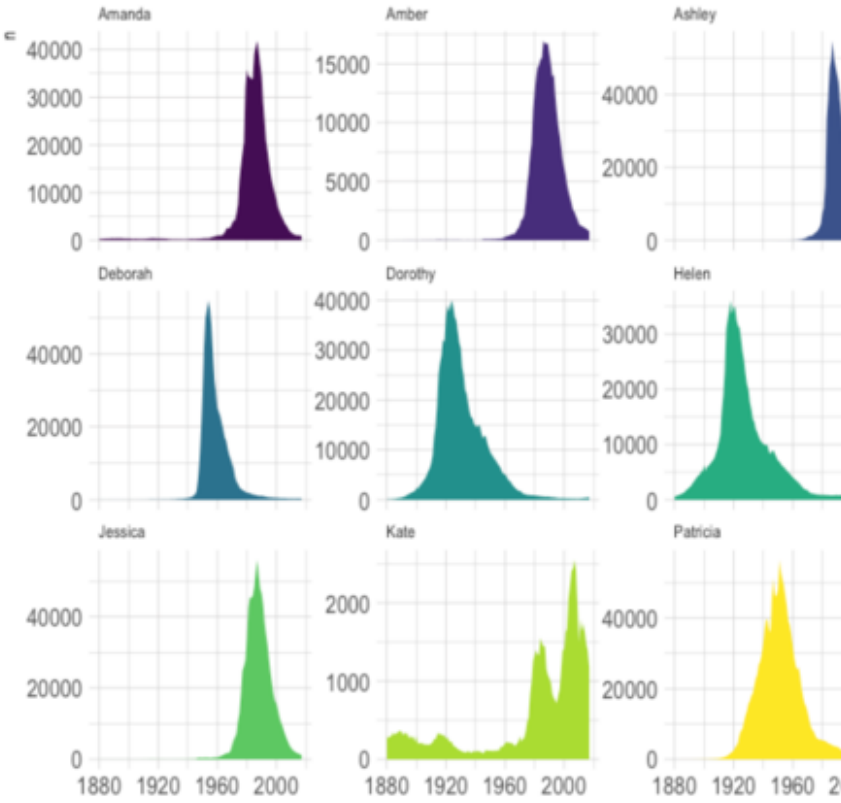


Another straightforward and simple plot, whose meaning is pretty obvious even without a title.

- Some more effort could have gone into the x axis label. Are the units kilograms, or pounds?
- This plot conveys a lot of information really intuitively. You can intuit that the dotted lines are probably mean weights; that the distributions show the range of data.

- The semi-transparent colors make it possible to see how the two distributions overlap. Very helpful!
- It is interesting that the male/female color associations are opposite the convention. Do you think that was intentional?

- Note the little tick marks along the x axis. That is a suble way of indicating sample size; it shows how much data are used to populate each distribution.



Some baby names throughout the last 140 years

- In this plot, chart junk is kept to a minimum, which is nice, but the names are a little too small to read. It is strange that the names are a smaller font size than the axes labels.
- What do you think about the use of color? The colors are pretty, but it is a purely aesthetic choice; the colors don't correspond to anything about the data at all.
- It was an interesting decision to only include x axis labels on the bottom row of plots. In a way this is nice because it (1) reinforces the fact that each plot is using the same x axis range, and (2) removes redundant content from the plot. However, it makes it a bit more laborious to explore the plots in the top row.
- It was also an interesting decision to make the y axis ranges different for each plot. There are trade-offs to this decision. What would be lost if all of these facets were forced to use the same y axis range?

Chart junk

A few times already, we have referred to the concept of chart junk. This refers to the idea that the best plots are the ones that minimize the ink-to-data ratio. In other words, there should be no extraneous or unnecessary ink on your plot.

The chart junk principle applies to both graphical and tabular representations of data. Which of these tables is easier to read?

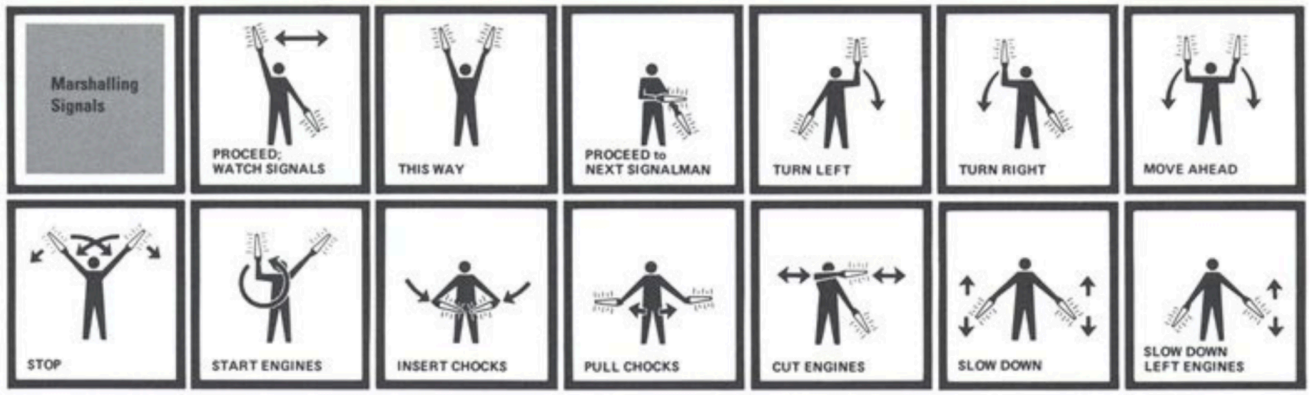
Before

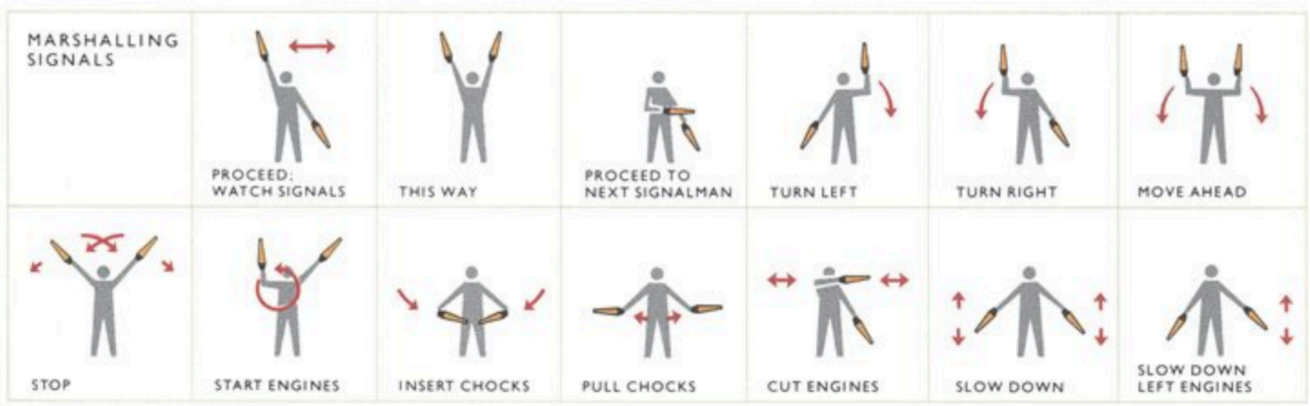
Team	1999	2000	2001	2002	2003	2004	2005	2006
Arizona Diamondbacks	$61,104,250	$72,346,274	$72,505,820	$77,893,950	$80,657,500	$60,521,550	$58,348,000	$56,943,474
Atlanta Braves	$68,134,250	$70,448,200	$74,073,950	$75,379,325	$96,872,425	$79,024,800	$85,148,575	$79,708,500
Baltimore Orioles	$73,057,675	$70,213,300	$62,436,500	$47,260,525	$59,872,650	$45,725,575	$66,580,850	$64,808,674
Boston Red Sox	$55,142,125	$64,950,275	$85,592,900	$90,305,050	$89,473,775	$104,340,450	$108,314,225	$111,204,200
Chicago White Sox	$22,740,725	$26,839,225	$57,743,525	$52,826,700	$49,048,075	$62,704,325	$69,655,550	$98,918,900
Chicago Cubs	$51,889,225	$50,449,450	$61,553,675	$67,581,500	$72,092,250	$78,535,925	$77,866,900	$84,679,625
Cincinnati Reds	$28,588,575	$43,395,550	$43,488,350	$37,542,000	$50,874,800	$39,453,450	$49,719,225	$53,115,000
Cleveland Indians	$40,769,300	$72,962,375	$76,645,825	$65,757,875	$39,382,925	$28,807,750	$38,060,700	$56,795,875
Colorado Rockies	$53,716,200	$54,563,575	$65,803,700	$62,639,850	$55,818,050	$57,670,000	$41,197,425	$34,277,500
Detroit Tigers	$30,460,600	$53,949,225	$44,492,125	$49,150,000	$47,722,125	$41,387,100	$61,600,525	$76,205,625
Florida Marlins	$17,477,775	$17,303,450	$29,586,800	$37,482,075	$43,180,975	$38,998,125	$55,903,675	$14,421,625
Houston Astros	$49,643,225	$47,489,925	$55,909,975	$58,748,825	$67,778,700	$74,666,300	$73,825,975	$88,991,825
Kansas City Royals	$22,794,225	$20,922,325	$30,726,725	$40,718,800	$38,959,125	$39,674,175	$34,149,075	$40,976,700
Los Angeles Dodgers	$70,773,175	$81,596,550	$93,949,100	$91,202,850	$101,824,625	$86,244,550	$87,524,575	$91,830,000
Anaheim/Los Angeles Angels	$39,265,275	$42,886,800	$37,554,975	$55,108,625	$73,177,475	$93,596,925	$81,911,925	$103,625,325
Milwaukee Brewers	$38,329,400	$28,519,800	$39,897,525	$43,351,575	$35,023,275	$27,518,500	$40,234,825	$56,790,000
Minnesota Twins	$18,302,400	$16,884,125	$22,548,800	$38,677,875	$53,466,350	$51,524,050	$52,421,300	$61,355,825
Montreal/Washington Nationals	$14,977,325	$30,006,750	$28,978,750	$34,527,225	$49,960,950	$36,997,925	$40,484,575	$52,722,925
New York Mets	$57,924,475	$79,509,775	$83,191,450	$90,993,850	$100,748,800	$96,758,950	$97,009,400	$97,020,275
New York Yankees	$75,923,625	$79,774,350	$88,541,850	$108,509,125	$133,670,625	$157,616,825	$198,017,725	$177,376,950
Oakland Athletics	$22,340,700	$29,603,075	$31,306,250	$36,740,000	$48,423,875	$55,393,675	$53,720,450	$62,322,050
Philadelphia Phillies	$26,118,625	$40,782,750	$40,061,700	$51,745,625	$61,017,250	$86,314,050	$91,671,075	$81,734,575
Pittsburgh Pirates	$18,498,050	$27,815,700	$42,498,650	$38,405,850	$48,698,300	$29,840,675	$34,047,325	$41,846,200
San Diego Padres	$42,703,875	$45,684,175	$35,493,625	$39,741,200	$37,858,325	$54,639,500	$56,155,175	$62,254,625
Seattle Mariners	$48,941,825	$56,649,050	$87,146,875	$80,282,675	$80,729,400	$72,807,000	$87,096,350	$84,927,400
San Francisco Giants	$44,943,550	$51,670,975	$58,641,350	$72,499,850	$79,184,775	$66,144,500	$86,045,600	$90,862,075
St. Louis Cardinals	$42,313,275	$56,901,725	$66,612,800	$71,248,325	$67,057,175	$75,633,625	$89,730,625	$85,039,625
Tampa Bay Rays	$29,269,400	$50,617,050	$60,881,125	$30,696,425	$19,630,000	$27,321,000	$26,690,675	$31,623,175
Texas Rangers	$71,956,675	$68,073,000	$71,374,525	$90,777,700	$87,106,400	$47,263,775	$46,089,375	$52,796,875
Toronto Blue Jays	$42,797,425	$44,459,925	$67,477,225	$66,262,350	$47,480,550	$48,693,275	$43,621,625	$66,087,575
Average	$43,338,913	$49,875,624	$56,243,678	$59,605,910	$63,877,748	$62,107,276	$66,361,319	$72,051,937

After

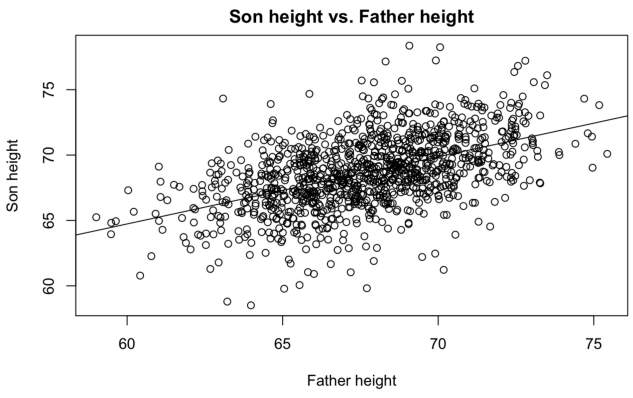
	Average Salary ($ millions)							
Team	1999	2000	2001	2002	2003	2004	2005	2006
Arizona Diamondbacks	61.2	72.3	72.5	77.9	80.7	60.5	58.3	56.9
Atlanta Braves	68.1	70.4	74.1	75.4	96.9	79.0	85.1	79.7
Baltimore Orioles	73.1	70.2	62.4	47.3	59.9	45.7	66.6	64.8
Boston Red Sox	55.1	65.0	85.6	90.3	89.5	104.3	108.3	111.2
Chicago White Sox	22.7	26.8	57.7	52.8	49.0	62.7	69.7	98.9
Chicago Cubs	51.9	50.4	61.6	67.6	72.1	78.5	77.9	84.7
Cincinnati Reds	28.6	43.4	43.5	37.5	50.9	39.5	49.7	53.1
Cleveland Indians	60.8	73.0	76.6	65.8	39.4	28.8	38.1	56.8
Colorado Rockies	53.7	54.6	65.8	62.6	55.8	57.7	41.2	34.3
Detroit Tigers	30.5	53.9	44.5	49.2	47.3	41.4	61.6	76.2
Florida Marlins	17.5	17.3	29.6	37.5	43.2	39.0	55.9	14.4
Houston Astros	49.6	47.5	55.9	58.7	67.8	74.7	73.8	89.0
Kansas City Royals	22.8	20.9	30.7	40.7	39.0	39.7	34.1	40.8
Los Angeles Dodgers	70.8	81.6	93.9	91.2	101.8	86.2	87.5	91.8
Anaheim/Los Angeles Angels	39.3	42.9	37.6	55.1	73.2	93.6	81.9	103.6
Milwaukee Brewers	38.3	28.5	39.9	43.4	35.0	27.5	40.2	56.8
Minnesota Twins	18.5	15.9	22.5	38.7	53.5	51.5	52.4	61.4
Montreal/Washington Nationals	15.0	30.0	29.0	34.5	50.0	36.0	40.5	52.7
New York Mets	57.8	79.5	83.2	91.0	100.7	96.8	97.0	97.0
New York Yankees	75.9	79.8	88.5	108.6	133.7	157.6	198.0	177.4
Oakland Athletics	22.3	29.6	31.3	36.7	48.4	55.4	53.7	62.3
Philadelphia Phillies	26.1	40.8	40.1	51.7	61.0	86.3	91.7	81.7
Pittsburgh Pirates	18.5	27.8	42.5	36.5	48.7	29.8	34.0	41.8
San Diego Padres	42.7	45.7	35.5	35.7	37.9	54.6	56.2	62.3
Seattle Mariners	48.0	56.6	67.5	80.3	80.7	72.8	87.1	84.9
San Francisco Giants	44.9	51.7	58.6	72.5	79.2	66.1	86.0	90.9
St. Louis Cardinals	42.3	56.9	66.6	71.2	67.1	75.6	89.7	85.0
Tampa Bay Rays	29.3	50.6	60.9	30.7	19.6	27.3	26.7	31.6
Texas Rangers	72.0	68.1	71.4	90.8	87.1	47.3	46.1	52.8
Toronto Blue Jays	42.8	44.5	67.7	66.3	47.5	48.1	43.6	66.6
Average Salary	43.3	49.9	56.2	59.6	63.9	62.1	66.4	72.1

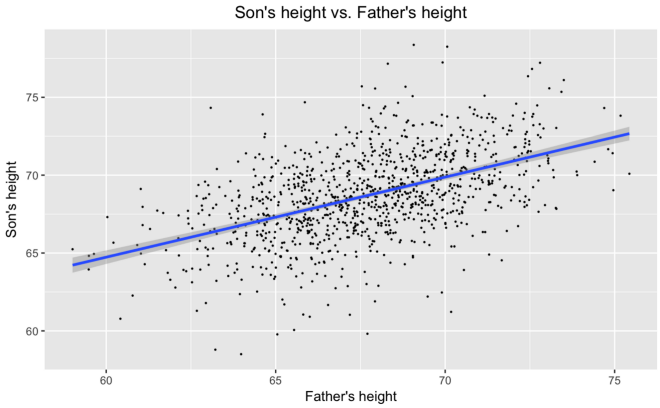
Which of these diagrams is easier to read?





And what about these basic scatterplots? Which is more effective and elegant?





Final thoughts

To get a sense of what can be done with data visualization – and just how enthusiastic data scientists can get about data viz – enjoy this video by the Swedish epidemiologist **Hans Rosling,** the pioneer of **interactive data visualization** (link here).

As you go down the rabbit hole of data visualization, it will be important to become familiar with the work of **Edward Tufte**, the grandfather of thinking about data viz as an art form. This video showcases Tufte and other data scientists who have been inspired by his work (link here).

Finally, this video offers a nice and concise summary of Edward Tufte's principles of data visualization (link here).

(PART) Getting started