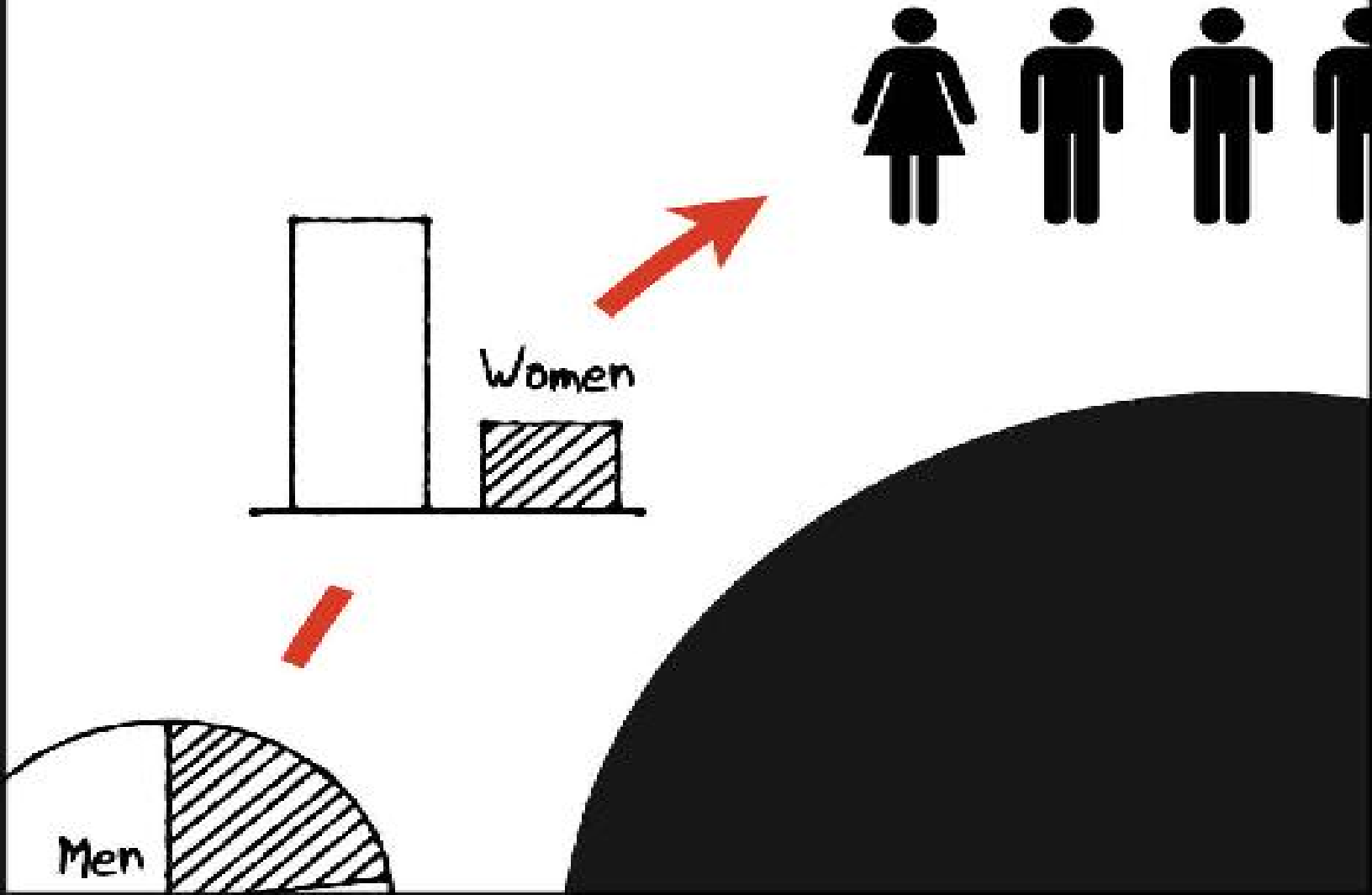
Jose Berengueres, Marybeth Sandell & Ali Fenwick

Introduction to

Data Visualization & Storytelling

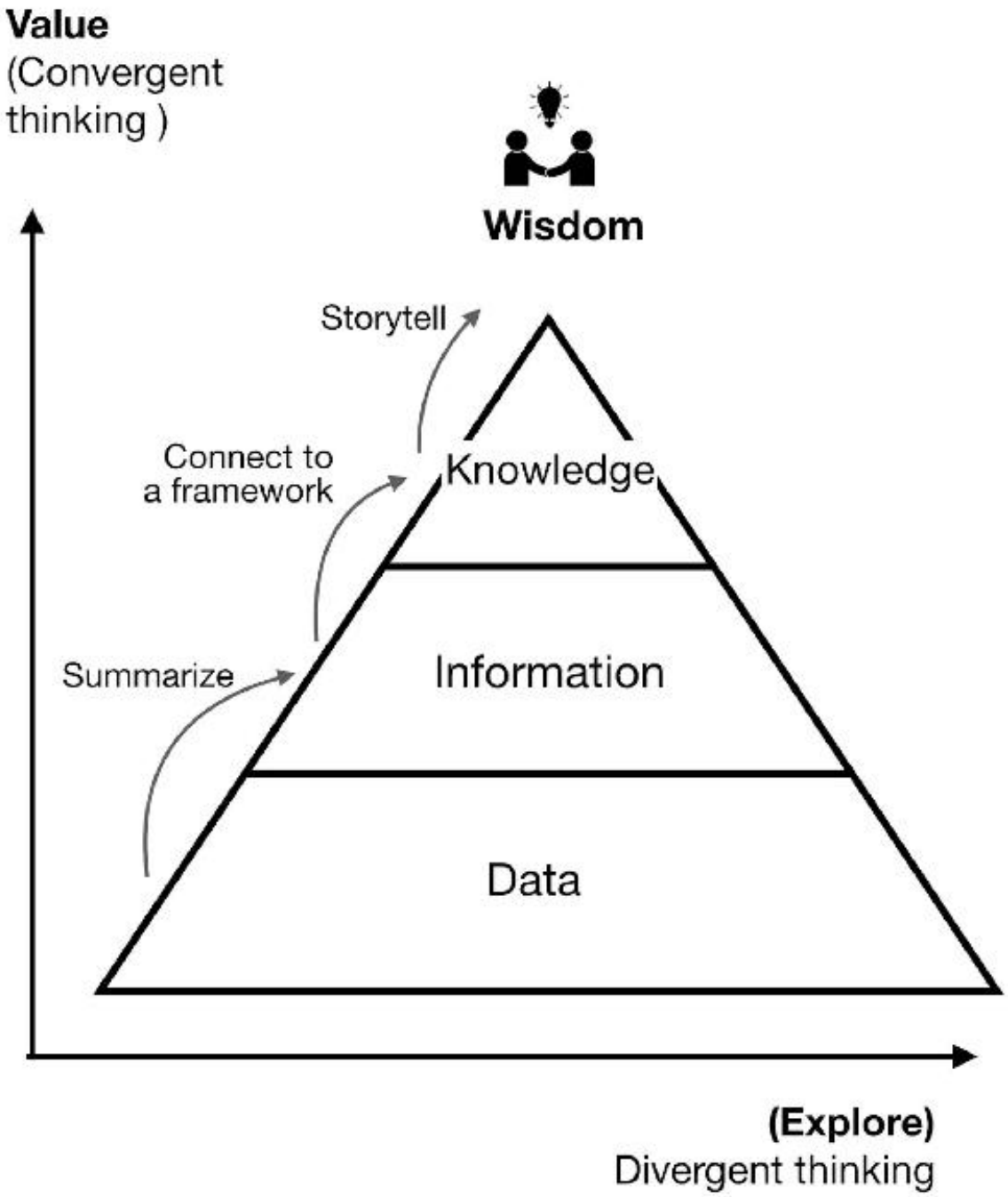


Introduction to Data Visualization & Storytelling
A Guide For The Data Scientist

Jose Berengueres
with
Marybeth Sandell &
Ali Fenwick

2nd Edition
Type setting by SNASK

Data, value creation & thinking modes



TOC

Preface

It was May 2018 when I received an email from Kaggle.com, a data science community where people all over the world compete in data modelling challenges. Every year since 2017, Kaggle has surveyed its users and this year they decided to organize a storytelling competition and offer cash prizes for the best survey visualization. I was curious to see what people had submitted so far and sifted through the entries. I was impressed at how much time and effort had been put into analyzing the data. Some charts required laborious data wrangling, others crafty SQL inner joins, and some Python sorcery. And yet, the charts were not doing justice to the compelling findings and the stories that could be told. No doubt, some were creative, extensive in length, exhaustive in the exploration. However, there were also many unimpressive charts. Why was so much IQ not producing more enticing visualizations?

It's about the awareness

80% of the data scientists graduating today will do so without having received any formal education on storytelling[1], and 60% of them place data visualization at the bottom of their priority list of skills to have. In addition, the fact that most data science is taught at STEM schools, not Art schools, does not help either. (Really, STEM schools should aspire to be STEAM schools, to include Art). In order to improve the quality of data visualization, there must be a change in how data scientists are trained along with a **mind shift** regarding the imperative of good story telling. However, to make effective visualizations, art sensibility is not all that is needed.

Death by default settings

A second cause, is that the gallery of default chart styles of *Microsoft Excel is unlikely to match* what is required to tell the story[2] and is laboriously hard to adjust. For example, a single bar chart in Excel has a whopping +129 configurable options: color fill options, axis options, line

thickness, line types, legend positioning options, scale options, bar type, spacing options. Each option has between 2 and 10 possible values. This is a 10 followed by 38 zeroes. Because of the tricky laws of aesthetics[3], very few of these combinations will produce the Aha! chart that will dazzle the audience, and even if it could, it is far too laborious for most people. To put it into perspective, there are more possible combinations than stars in the Milky Way and if we spent just 3 seconds tweaking each of the 129 combinations it would take 6 minutes and 578 mouse clicks to adjust them all. It does not help that most visualization software in use today (this applies to ggplot too) was developed by CS graduates with little training on the basics of color theory, information design or visual communication. The exception was Steve Jobs, who took calligraphy lessons and credited that for the Mac's beautiful fonts. See Walter Isaacson's bio on Jobs as a flashback on how much the field has improved[4].

Start with why

A third cause, is that to produce a meaningful visualization we need to know why we are visualizing. To get to the why, it's important to explore the story that needs to be told and be able to make it relevant.

To succeed at transforming data into a compelling and relevant story, it helps to **connect** the data to a context, metaphor or mental framework (frameworks from Economics, Art, and Sociology are particularly useful). In order to make these connections, a cross-functional education is necessary. Unfortunately, this is not the norm resulting in loads of very interesting data failing to become useful knowledge. To address this, in the following pages you will find a set of principles by example that I wish I had learned in grad school.

Happy visualizations!
Jose Berengueres
Stockholm, Oct 25th 2019

How to use this book

What will you learn

1. **Identify** the role of a narrative in a chart
2. **Transform** data into information
3. **Synthesize** knowledge
4. **Apply** visual thinking tools to the decision-making process
5. **Select** visual communication techniques to persuade

Updates to this edition

This edition adds two new chapters, one on bias by Marybeth Sandell and one on Psychology by Ali Fenwick. Whether you are using this book as course material for a data visualization course, in journalism, or in a data science minor; this is a solid foundation before getting hands on with Tableau, Excel or Python Pandas. We have also upgraded all the images to 600 DPI.

Exercises

There are several exercises in this book so it is helpful to have paper and black fiber tip pen 0.5-0.7mm such as Pilot EF or a Pilot Sign Pen available.

Slides

This book was born out of workshop full color slides that you can use in the classroom. To get them send an email to jse@ieee.org with subject "DATAVIZ101" and link to your verified purchase **review** on Amazon.

Chapter 1. Stories, Data & Narratives

– Finding the purpose of the chart –

Fig 1 *These three magazines publish some of the most influential charts in the newsstand. None of them is made with Excel*[5].

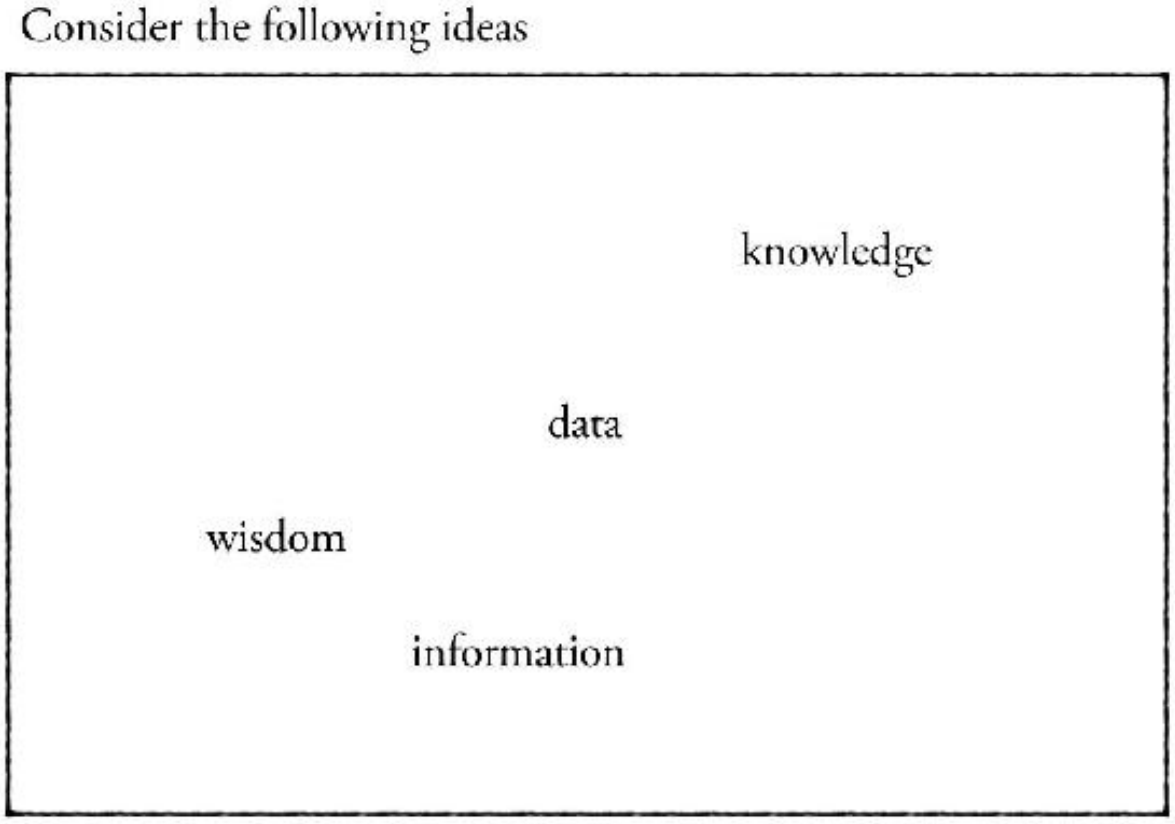
Before you start

Before visualizing, there is data. However, before working with data it is important to understand what knowledge is. What is knowledge to you? In this chapter, you will learn three things; (i) the difference between: data, information and knowledge; (ii) What wisdom is; and (iii) What the process to produce knowledge is.

What is data?

Exercise

What is data to you? We always open data visualization workshops with this exercise.



Now, order them as logically as possible

_____ > _____ > _____ > _____

Fig 2 *An exercise to understand what data is by relating it to knowledge.*

Data has many definitions. However, to understand, there's nothing like doing an exercise. There are various ways to learn something. By listening, by talking about it, by reading about it and, by doing it. Of those four, doing has the highest recall rate. Can you order the four words (data knowledge information wisdom)? Most students will come up with the following ordering:

data > information > knowledge > **wisdom**

Once the order of the words has been agreed, we can discuss the ordering criteria. Why did we order them the way we did? This is a great conversation starter. To ground the conversation, it further helps to list the attributes of the words at the extremes. What are the attributes that **distinguish** data vs. wisdom?

Data vs. wisdom

Exercise

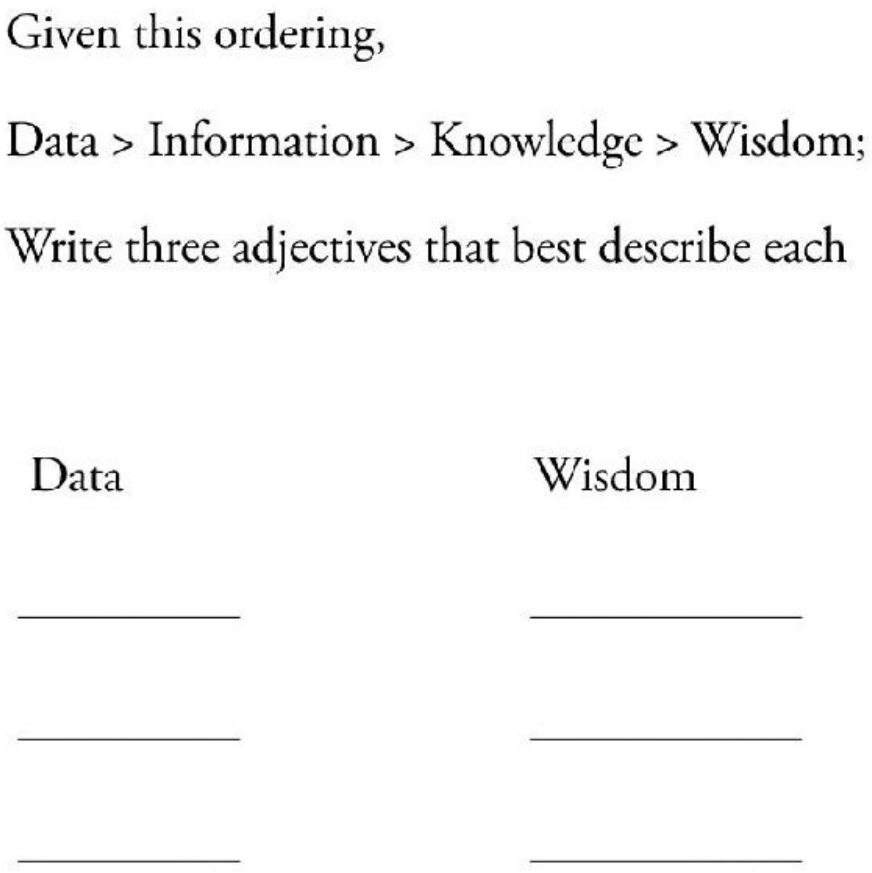


Fig 3 *An exercise used to understand the arrow of value between data and wisdom*[6].

Solution

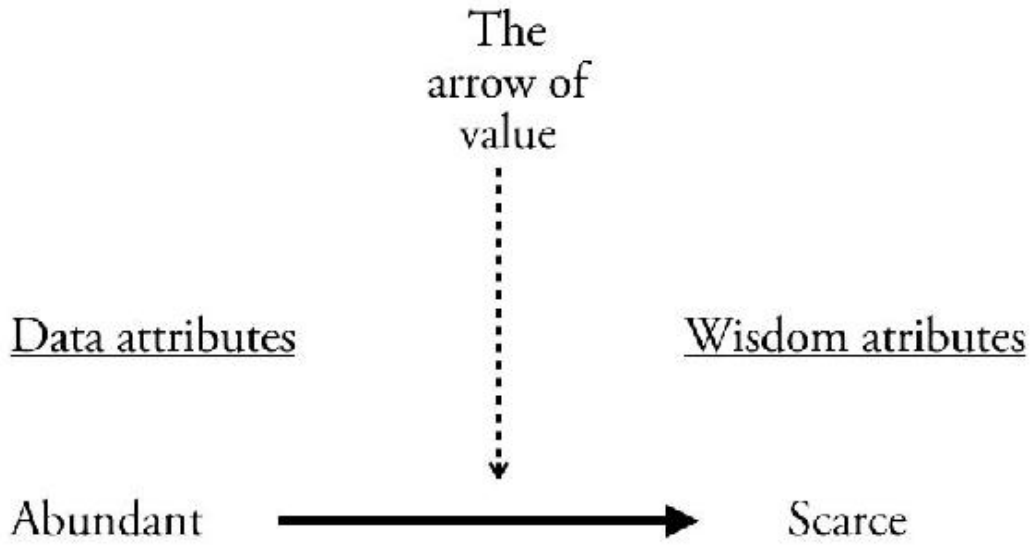


Fig 4 *The arrow of value.*

What is wisdom?

Data is **many** while wisdom is usually characterized by **few**. Data is abundant, wisdom is scarce. Value is closely correlated with scarcity too. This exercise is great to clarify the pervasive confusion between data, information, knowledge and its relationship to value, scarcity and wisdom.

How is wisdom made?

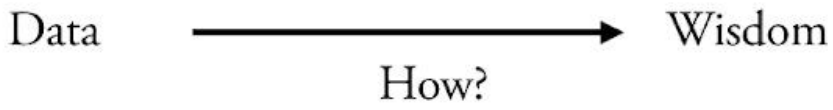


Fig 5 *This arrow represents a process.*

Jackie Chan says in one of his films, “information is not knowledge, and knowledge is not wisdom”. But what is wisdom? Is wisdom just knowledge in context? Is wisdom meta knowledge? — knowledge about knowledge? And more importantly, knowing in which situation to apply a given knowledge? Even if the definition is not universal, what we are more interested in here is how to transform knowledge into wisdom. Why? Because it is a high added value activity and one of the reasons (if not the only reason) why companies employ data scientists. One way to arrive at wisdom is the **Synthesis process** — the dialectic combination of thesis and antithesis into a higher truth.

Match the words

Exercise

Connect each keyword with its corresponding image. Time 1 minute.

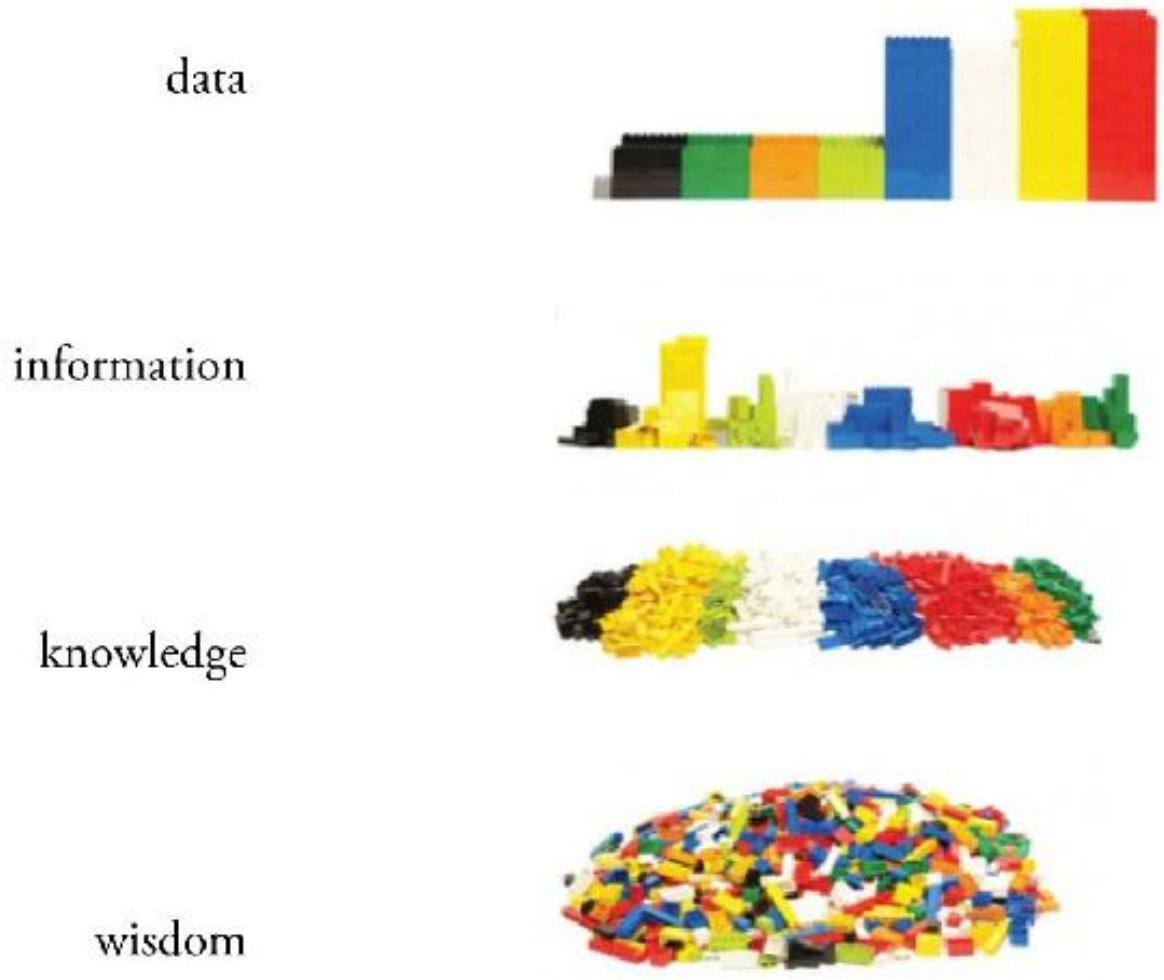


Fig 6 *An exercise used to understand what wisdom is. Lego image source: LinkedIn, anonymous mem.*

Solution

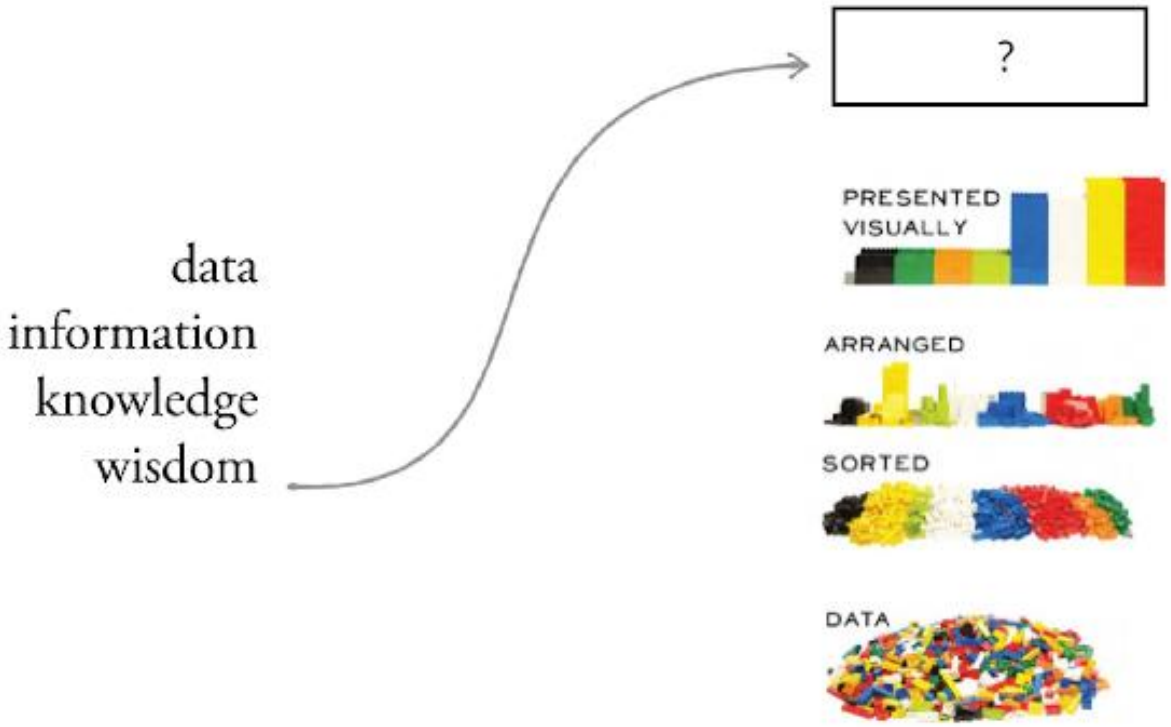


Fig 7 *To solve this exercise, think outside of the box.*

In this exercise, students must match the words to a Lego configuration. Each Lego piece represents some piece of data. The Lego exercise is a great way to clarify that wisdom (and to some extent knowledge) is not just data arranged and visually presented. Let us zoom into this case. Imagine you are the C.E.O. of Lego and this information has been presented to you. Can you map the words to the pictures? Where does wisdom map to? In this exercise wisdom is orphan to drive home the point that wisdom is something more than presenting and arranging data. In other words, wisdom is knowing where to apply knowledge. Example of wisdom: “We do not have any pink Lego pieces in this set. Why is this? Are we blind to some important customer segment? Are sales not optimal because of that?” Note how valuable knowledge emerges when we connect our information (there are no pink bricks) to other existing knowledge (gender studies).

Information vs. meaning

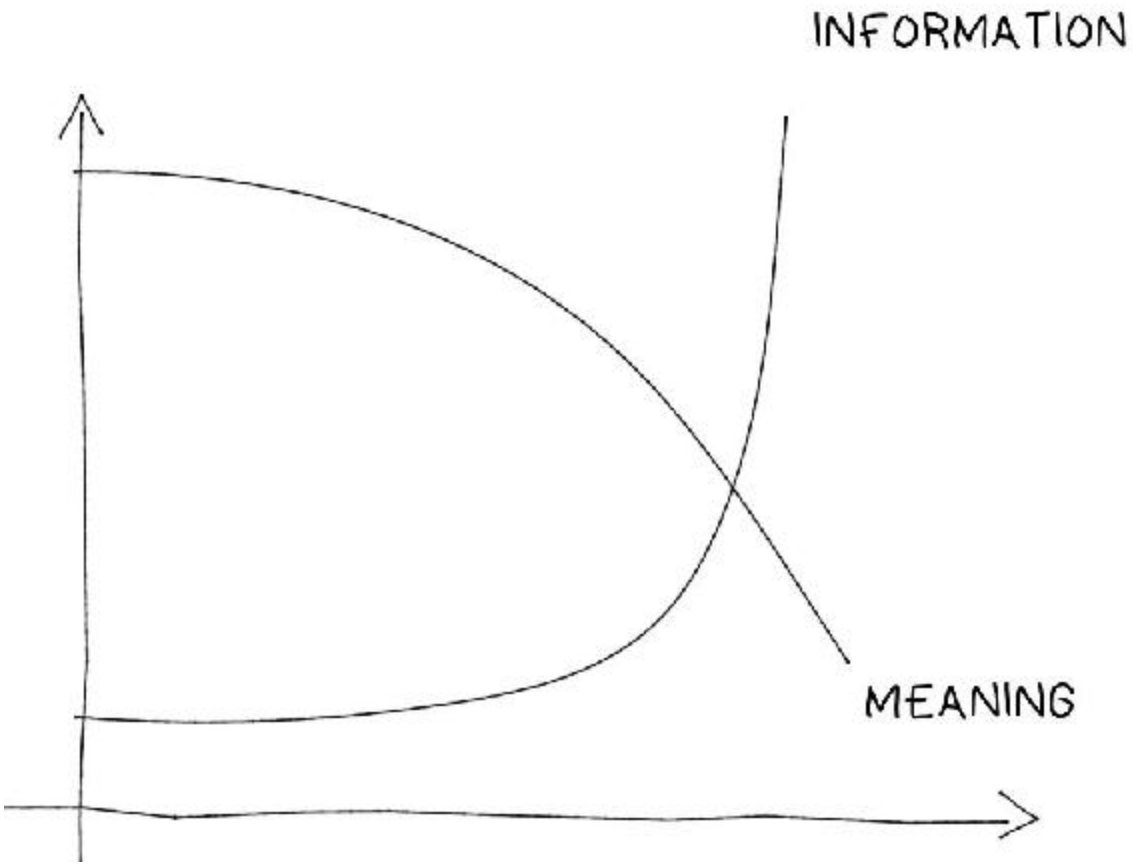


Fig 8 *"There is no peace in more".*

Meaning vs. Information

A version of this chart was originally posted by Hugh McLeod, the author of "Ignore Everybody". When I first saw this chart, I felt mesmerized by it. Then I tried to find flaws in the chart. But there are no flaws. This chart is a masterpiece. I was just flummoxed at the simplicity of it[7]. The y-axis is the Shannon Information Entropy, measured in number of bits. The x-axis, is an ordered sequence (not continuous, and has no units). It is an ordering of cases (visualization examples) that the author ranked in descending order of meaning along the x-axis; for each element two dots on the y-axis are plotted. Then the author drinks his own medicine and

simplifies the chart to reduce all non-essential information to deliver the Aha moment: Less (info) is more (meaning). The info line is called sometimes hockey stick. It is an exponential function. The meaning line fits a downward arch. What this chart means is that it is **not** possible to have meaning with information overload.

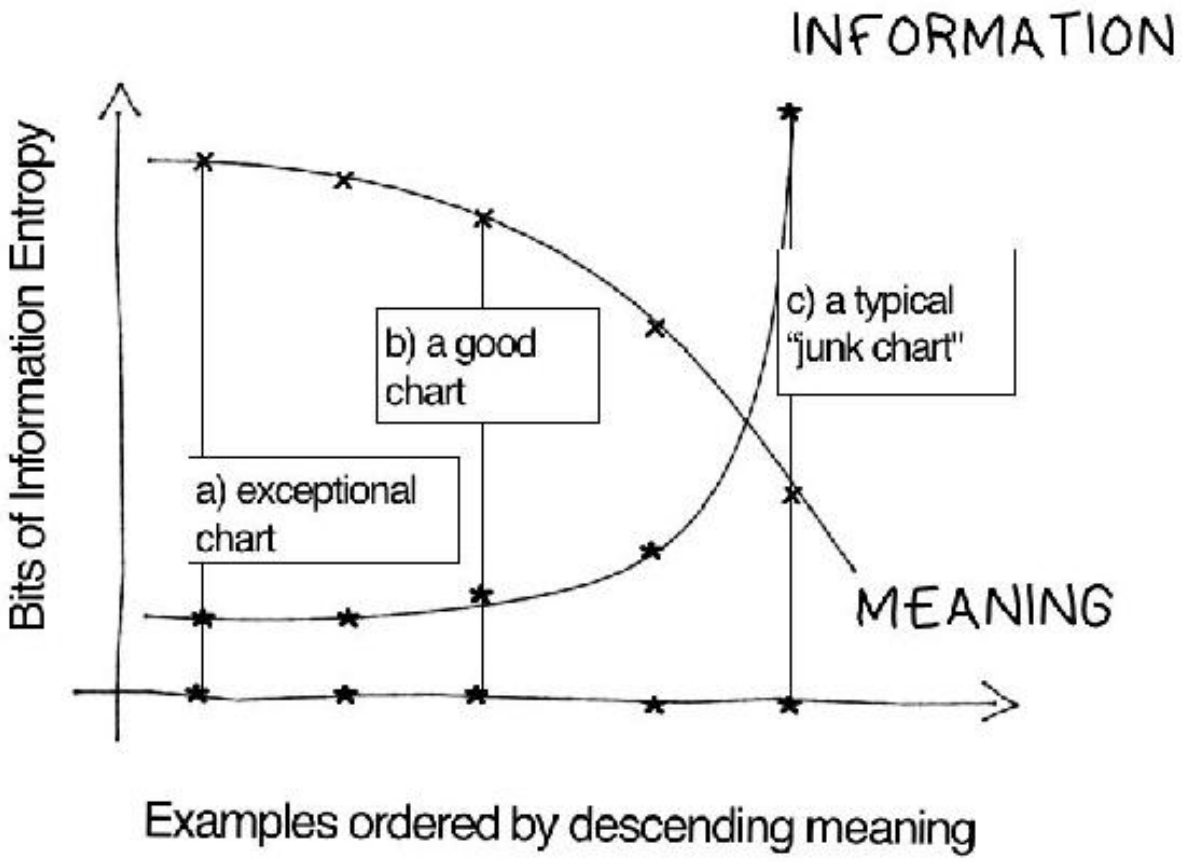


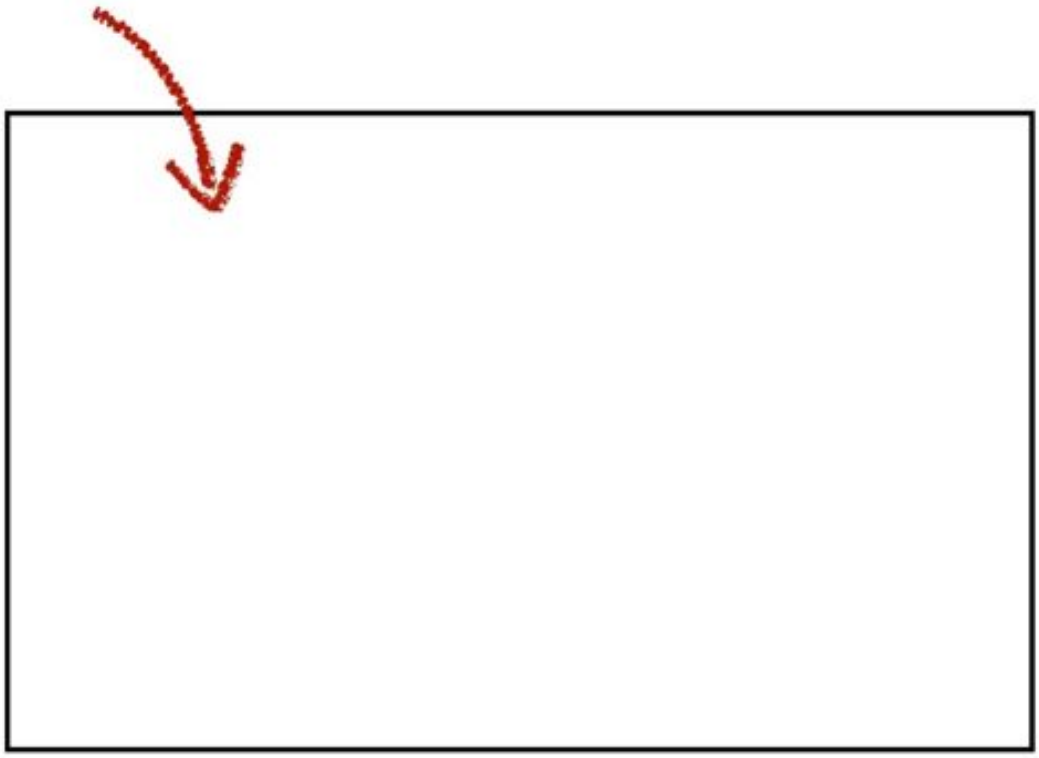
Fig 9 *A scientific deconstruction of the McLeod's chart.*

Exercise

Summarize the previous 20 pages in 20 words or less. Time 3 minutes.

_____

Now, visualize it



Solution

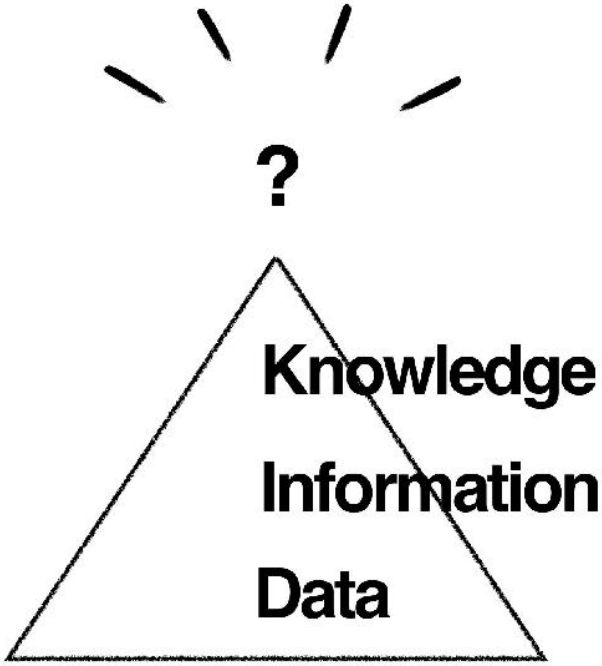


Fig 10 *A visual metaphor based on the DIKW model .*

Wrapping-up

When delivering a workshop, this exercise is a great way to bring attention to the point of summary vs. synthesis. A summary is a mere reduction process whereas synthesis is demonstrating an understanding of the subject by relating it to other subjects, ultimately adding value through connective thinking.

In Fig 10, we summarized the relationship between data and wisdom by way of the pyramid metaphor. Elements at the top of the pyramid are valuable, scarce and hard to carry to the top because they work against the force of gravity. This is a great example of a visual summary of the chapter while also a good example of synthesis.

Now that we've learnt the difference between data and knowledge, and how to transform knowledge into wisdom, let's look at the role of narratives in charts.

Narratives & stories

Exercise

Fill in the blank with a verb. Example: “Data _____ stories”. Time 3 minutes.

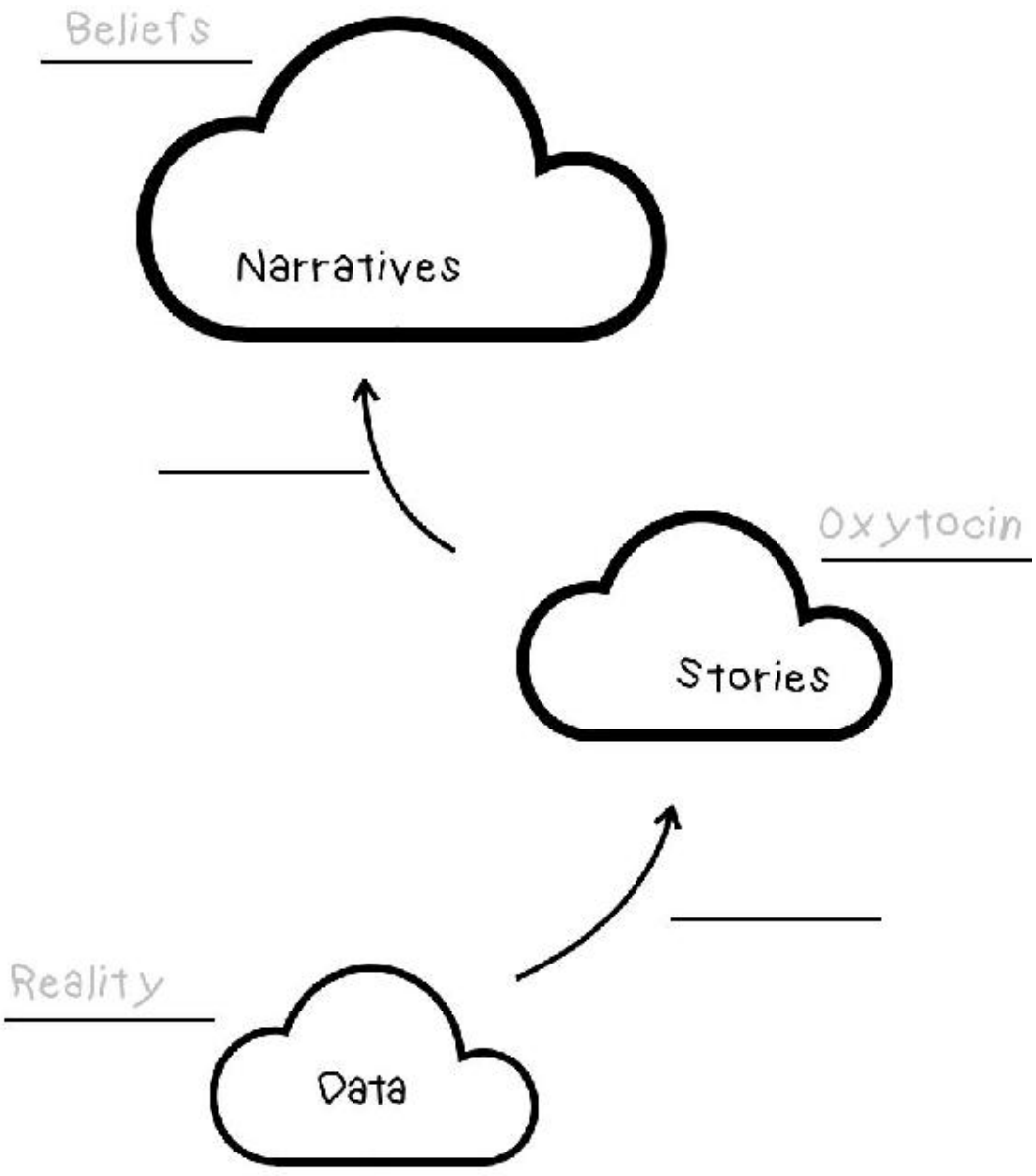


Fig 11 *An exercise used to understand the difference between narratives and stories.*

Solution

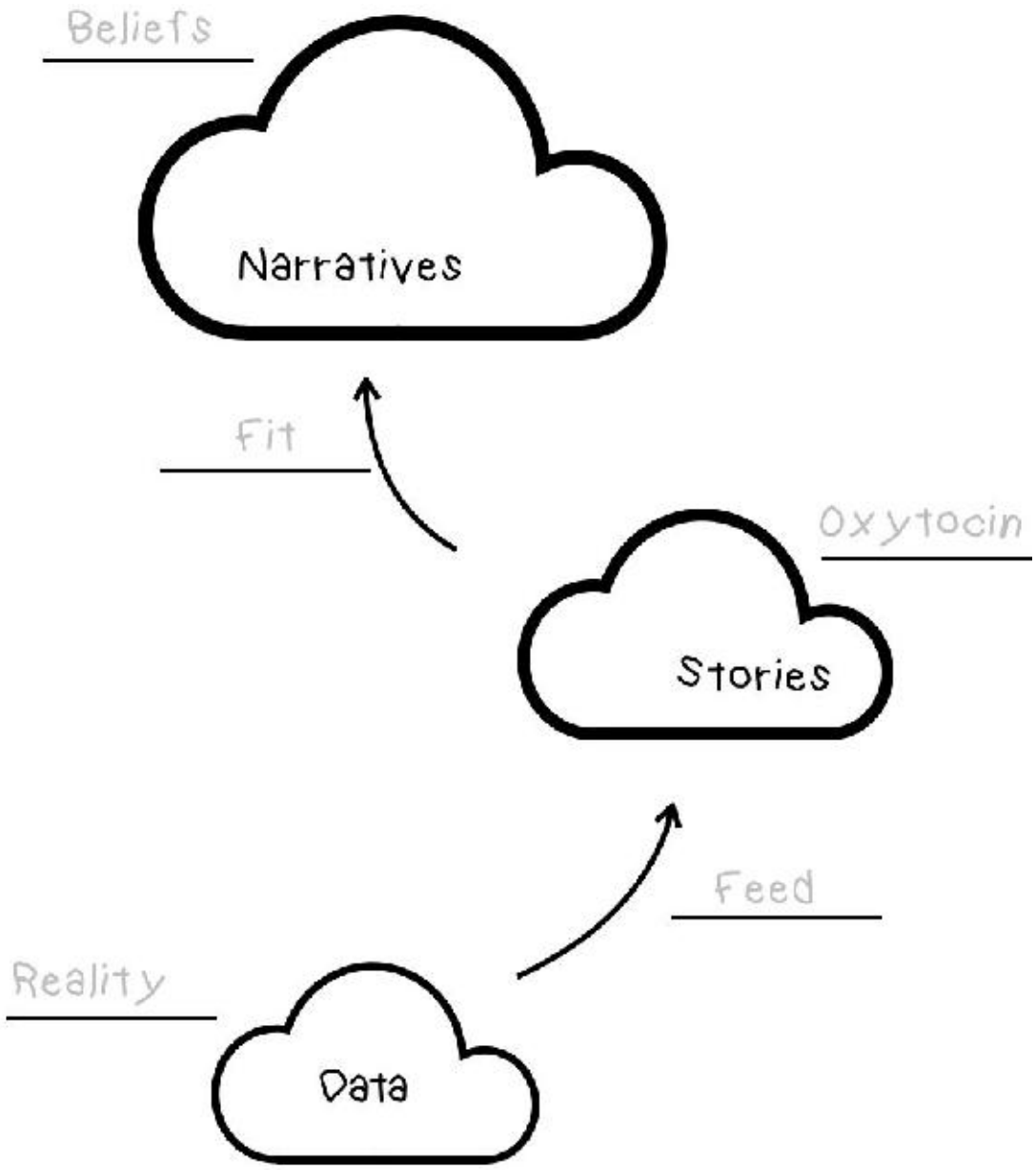


Fig 12 *The cloud template, use when logic is fuzzy*[8].

Stories

To understand what a narrative is, we first need to understand what a story is. A story is an account of events. We humans love stories. Why?

Telling and consuming stories is addictive. For example, listening to an Aesop fable, reading a book, and watching a movie, all those release oxytocin in the brain— the feel-good hormone. That is why people get addicted to Netflix, Venezuelan soap operas and fiction books. Oral storytelling is thought to be the earliest method for sharing narratives. From an anthropological perspective, during most childhoods, narratives are used to shape children on proper behavior and values. This is usually done through tales.

Narratives

A narrative is a set of beliefs, values or *worldview*. Therefore, the chosen narrative interprets the story (and consequently the underlying data / reality). An example of a narrative popular in European culture is that kids, especially young girls, should not trust strangers. A story that promotes that narrative is the story *Little Red Riding Hood*, a tale from the 10th Century. In fables, a narrative, is also made explicit at the end of the tale as in the moral of the story. Another example of a narrative is FUD — Fear, Uncertainty and Doubt. It is also known as a disinformation strategy[9] used to thwart change to the status quo. It is said that IBM was one of the first companies listed on the Dow Jones to use FUD openly. An exemplification was, "*Nobody gets fired for buying an IBM*".

Connection to Aristotle

Because the goal of a story is to persuade; narratives, stories and data are related to the three modes of persuasion of Aristotle — The *Ethos*, the *Pathos* and the *Logos*[10]. The narrative is related to the *Ethos* (to appeal to the ethical values), the story is related to the *Pathos* (to appeal to the emotions), and the data that supports the story is related to the *Logos* (to appeal to logic).

Exercise

Identify the narrative, the story, the data and call to action in this photo. Time 2 minutes.

Fig 13 A Brexit Bus, UK. Original photo by David Beeson.

Solution

Fig 14 *Annotations by a student.*

At first glance, The Brexit bus story is an example of a narrative where an out-group steals from an in-group. But it is more than that. In fact, the bus appeals to so many people because it is connected to four narratives: (i) appeal to in-group tribal instincts to fight "outsiders"; (ii) doing for

common good is ethical; (iii) fighting unfairness is ethical too; and (iv) maximizing the utility of a resource is common sense.

These narratives were brilliantly embodied in the Brexit bus. The statistics and words printed on the bus were later found to be lies, but it did not matter. This story checked all the boxes of the pro-Brexit voters and mobilized them.

Is your narrative ethical?

Narratives can be evaluated from an ethical perspective. There are four ethical frameworks[11] commonly used. Each one maximizes some different ethical criteria. For each framework, the most ethical choice is the one that maximizes a different policy.

The four frameworks

In the so-called **Utilitarian** framework, the most ethical choice is the one that optimizes global happiness. In the **Common** good framework, the most ethical choice is the one that optimizes the Well-being of society. In the **Fairness** framework, the most ethical choice is the one that optimizes equity. In the **Virtues** framework, the most ethical choice is the one that aligns best with a set of predefined values.

Equity vs. equality

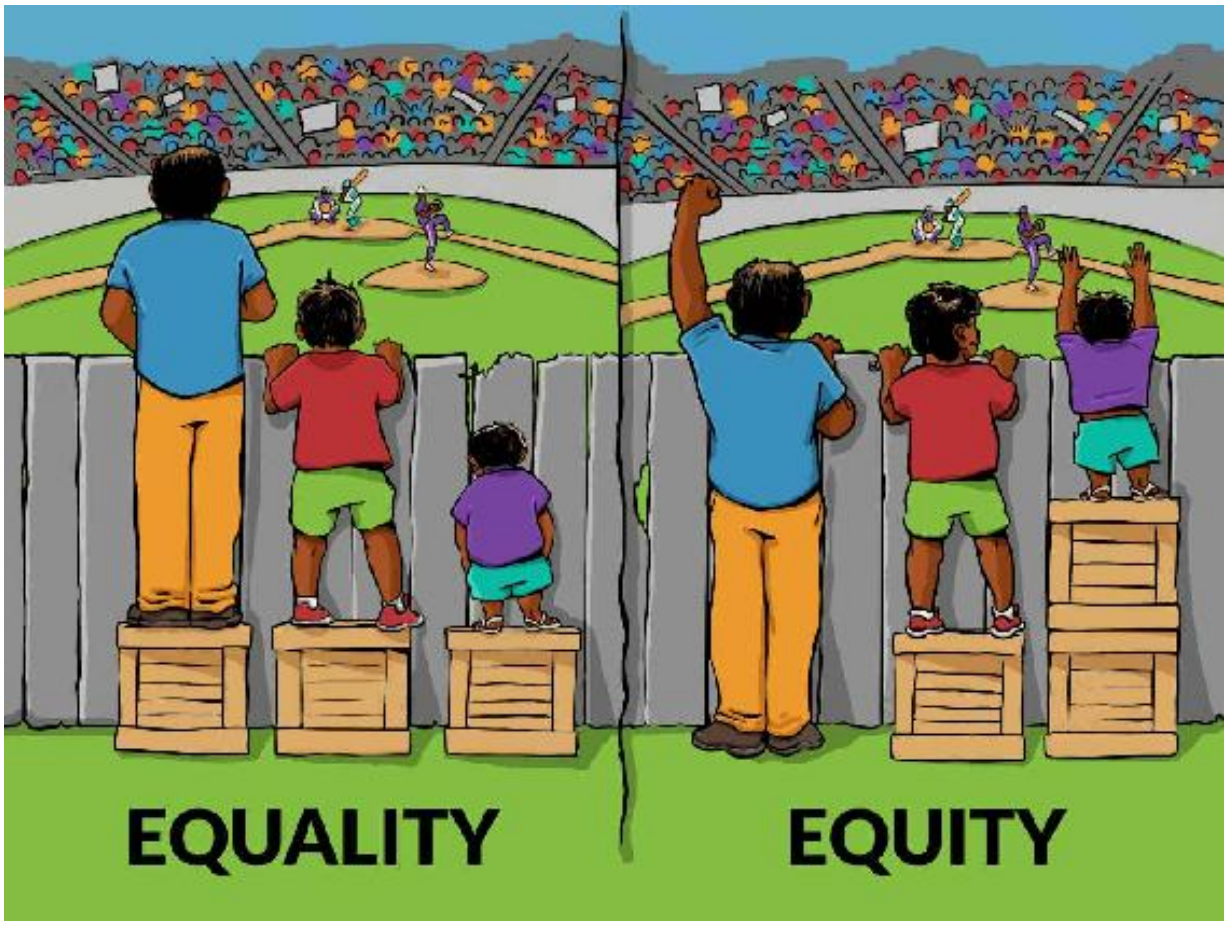


Fig 15 *Equity is not the same as equality. "Interaction Institute for Social Change". Source: Angus Maguire.*

Examples of famous virtue frameworks

- Liberté, égalité, fraternité[12]
- The ten commandments
- The Bushido

What are ethical dilemmas?

The so-called ethical dilemmas arise when conflict between the approaches appears. In addition, dilemmas can also arise when different regions in the brain (amygdala vs. frontal lobe) evaluate a situation with opposed outcomes. (See Trolley dilemma[13]).

Choices & ethics

To be ethical one needs to first consider at least two choices. Many times, we end up with suboptimal choices simply because alternatives were not even considered. Before disseminating a visual, always consider **two** alternatives. Then apply an appropriate framework to rank the alternatives. (If something goes wrong at least you can show you followed a process).

Elements of effective visualization

Exercise

Summarize the relation between story, narrative and data in a visual titled **Elements of effective visualization** that uses a plane as a metaphor. Fill in the gaps. Time 3 minutes.

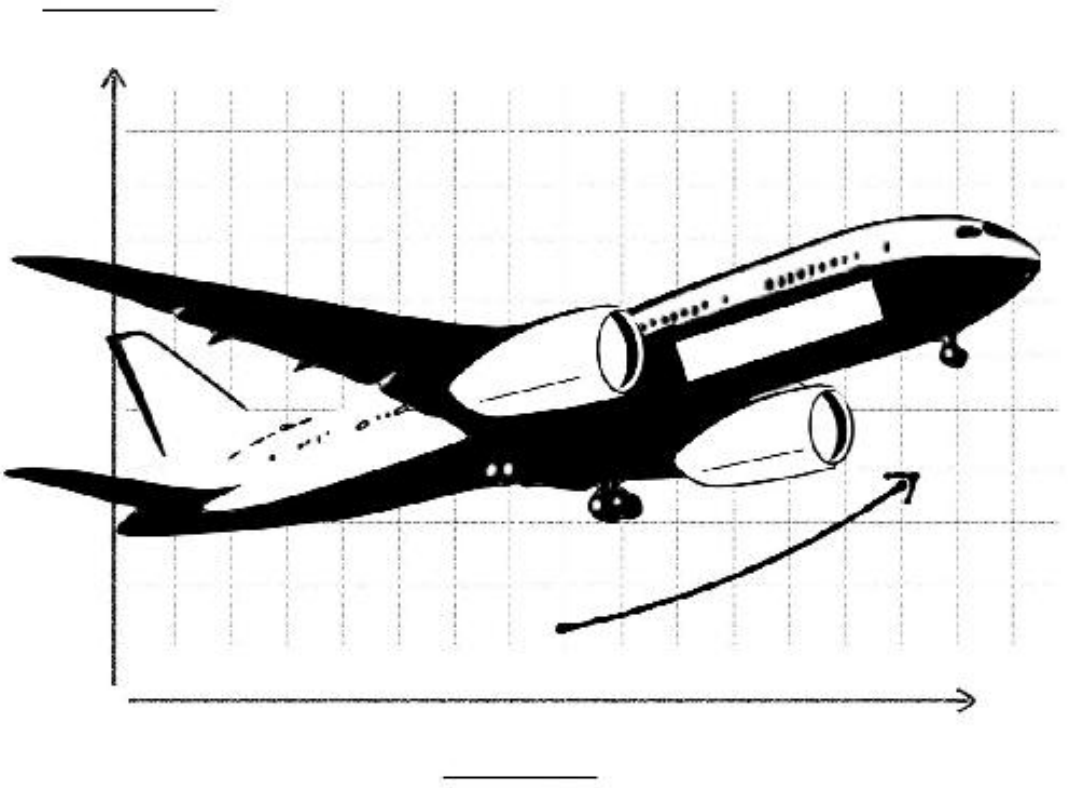


Fig 16 *Visual metaphors + annotations, so powerful.*

(Solution on the next page)

Elements of effective visualization

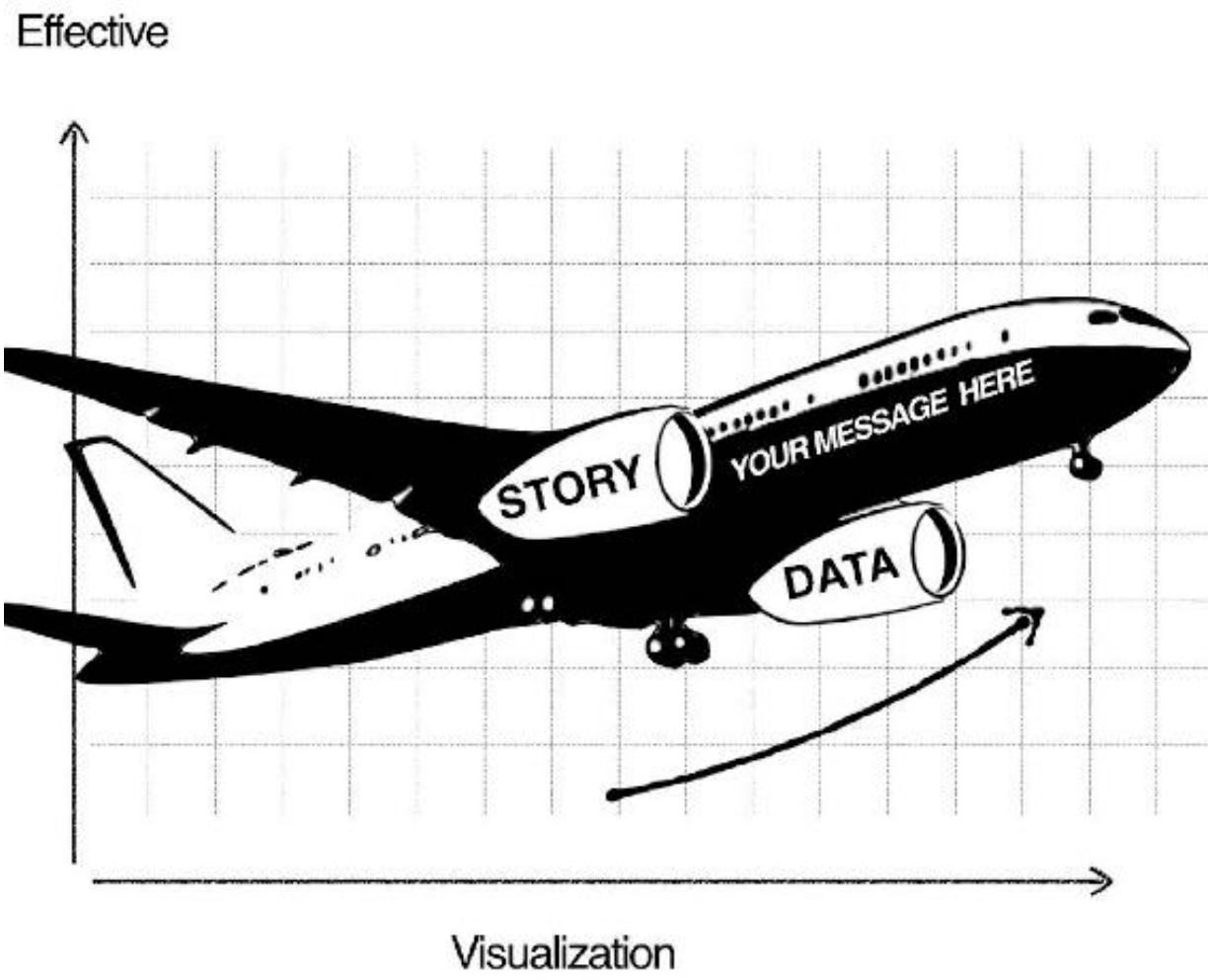


Fig 17 *Don't let words like 'narrative' get in the way of a great story.*

According to this visual, to persuade others you will need:

1. A why or a narrative (message)
2. A story (to help assimilate the message)
3. Data to give credibility to the story
4. A way to visualize the story (metaphors help)

The payload is the narrative (your message), and you are the pilot.

Quiz: Narratives & stories

True or False? Time 5 minutes.

1. A story is an account of events. [True / False]
2. Story consumption releases oxytocin. [T/F]
3. The purpose of a story is to advocate (transmit, perpetuate) a narrative (a belief, a culture, an ideology). [T/F]
4. A narrative does not stick to the human brain, it cannot go viral, it is not easy to remember. [T/F]
5. If your story (chart, visualization) advocates for no narrative, it will feel like it has no purpose. [T/F]

(Solution in the next page)

Solution

1. A story is an account of events [T/F]. True. To gain perspective on this question also see Lisa Feldman work on the Constructed Model of the World as Reality.
2. Story consumption releases oxytocin [T/F]. True. To learn more about the biochemistry of storytelling and its function in evolution.
3. The purpose of a story is to advocate (transmit, perpetuate) a narrative (a belief, a culture, an ideology) [T/F]. True.
4. A narrative does not stick to the human brain, it cannot go viral, it is not easy to remember [T/F]. True. The message of a narrative in its succinct form is usually not viral. Stories in form Ads, movies and books are more suited to go viral.
5. If your story (chart, visualization) advocates for no narrative, it will feel like it has no purpose [T/F]. True.

Quiz: Storytelling climate change

Time 15 minutes. Online search: allowed. In August 2019 Greta Thunberg set sail from Europe to a UN climate conference on the other side of the Atlantic. Identify the data, the story, and the narrative in this story.

Fig 18 *Unlike Al Gore in 2006, Greta Thunberg needed no charts to get her message across* [14].

Solution to Storytelling climate change

Narrative (example of answers)

- The underdog wins[15]
- Role reversal
- Dysfunctional family
- Climate change is an emergency

Story

(Abridged speech[16]). “You might be grown-ups but you are not mature enough to understand this emergency. If you did, you would not jet to the conference like you do. You could Skype, or travel like me to reduce your carbon footprint. Hence, you (not me) is behaving like the immature kid.”

Data

- Look at the big waves behind me, I am serious, this is dangerous
- A carbon neutral sail-ship = it’s possible to reduce carbon footprint
- The situation is bad enough that I had to skip school classes
- Air jet-set travel produces CO2 but there are alternatives look at me

Chapter 2. Visualizing Information

– How to transform data into information –

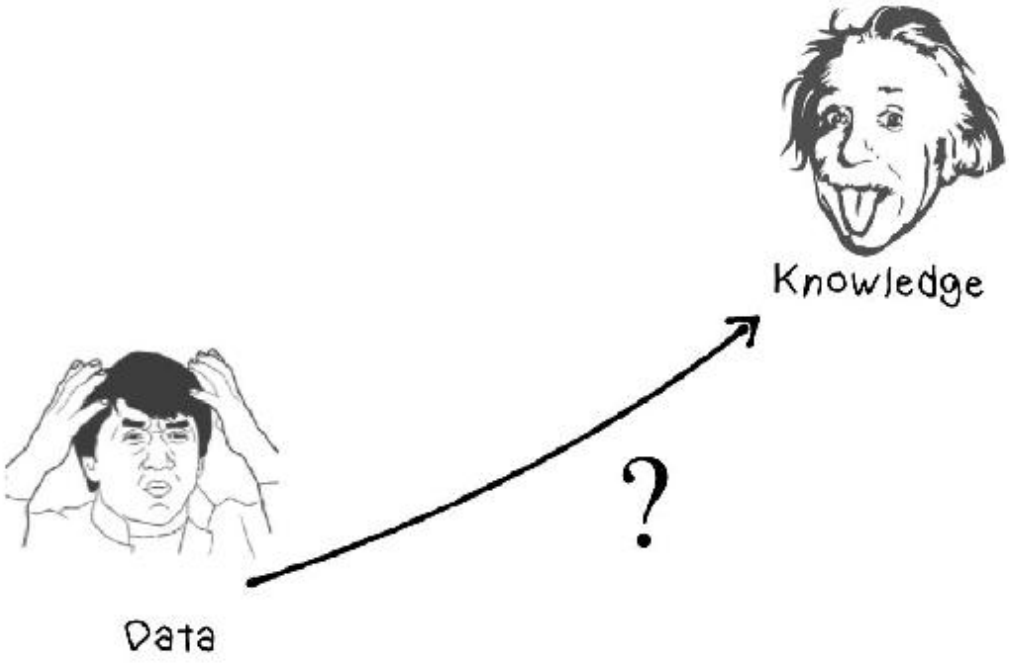


Fig 19 *Creating knowledge from data, the secret to winning the Nobel prize?*

In this chapter, you will learn to transform data into information, (a prerequisite to produce knowledge). In other words, how to make your charts more useful by displaying information **effectively**.

"The foundation of a useful chart is seldom graphic design"

The bulk of published books on data visualization focus on how to build charts and how to make them readable. A few of them spend a considerable number of pages listing all the types of charts available out there and use

the good chart / bad chart template to teach how a little tweak can make or break the readability of a chart. Others give color advice and how important it is not to clutter your chart with colors, (see also the term *junkcharts* coined by Tufte). This is helpful for improving readability and aesthetics but not so helpful in transforming information into knowledge (aka wisdom, prescriptive analytics). And while poor color choices can kill the readability of any chart, focusing on chart aestheticism is the equivalent to teaching about the importance of font types to someone that just wants to become a writer. After all, the book Harry Potter did not become a bestseller because of the font type they used. In the same vein, the typical root cause of a poor chart is failing to transform data into (meaningful) knowledge, not using the wrong aesthetics. Let’s see an example of transforming information into knowledge with data about gender distribution from the Kaggle 2018 survey.

Visualizing gender

Exercise

How would you **visualize** the following gender breakdown of data scientists? Time 3 minutes.

Survey responses	
Female	16.8%
Male	81.4%
NS	1.4%
ND	0.3%

Solution

Solution with Matplotlib

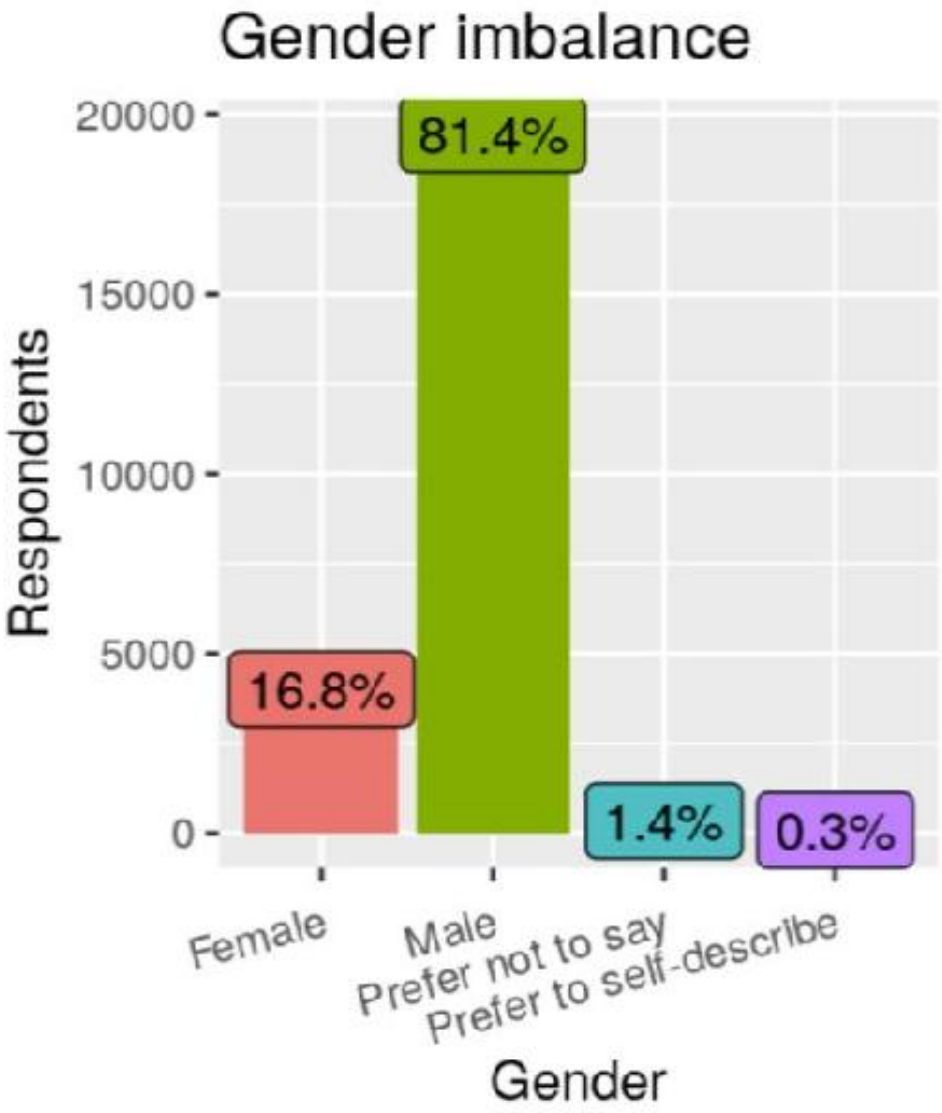


Fig 20 *This default Matplotlib chart uses* ***four*** *different font sizes and* ***six*** *colors (red, green, turquoise, purple, black, grey).*

Exercise

Is the figure above data, information or knowledge? Why? Reason your answer. Time 2 minutes.

Solution

It is just information. It is not knowledge because it is not significantly *more* useful than the original raw data.

Reflection

Unless you are in a preliminary Exploratory Data Analysis (EDA), it is not a good idea to disseminate a chart unless there is a clear why (narrative) for the chart. And even if you produce many charts as a part of an EDA, resist the temptation to show all of them. In this case, we are asked to visualize the gender distribution of the respondents of the 2018 Kaggle survey — one of the largest data science communities worldwide. Gender was one of the 30+ questions on the survey which was answered by about 30k respondents. Fig 20 is the default settings chart produced using the popular python library *Matplotlib*. This chart is perfectly fine. It is informative, but there is no message, there is no why. It lacks a purpose. Why? One reason is that it is not connected to any narrative. Another reason is that it does not increase our knowledge. Is it helping us to become wiser? Is it facilitating the prescriptive analytics function? How would you make this chart more useful?

Exercise

Draw here at least **three** alternative charts to Fig 20? Time 3 minutes. (Solution in the next page).

The Chart-narrative fit

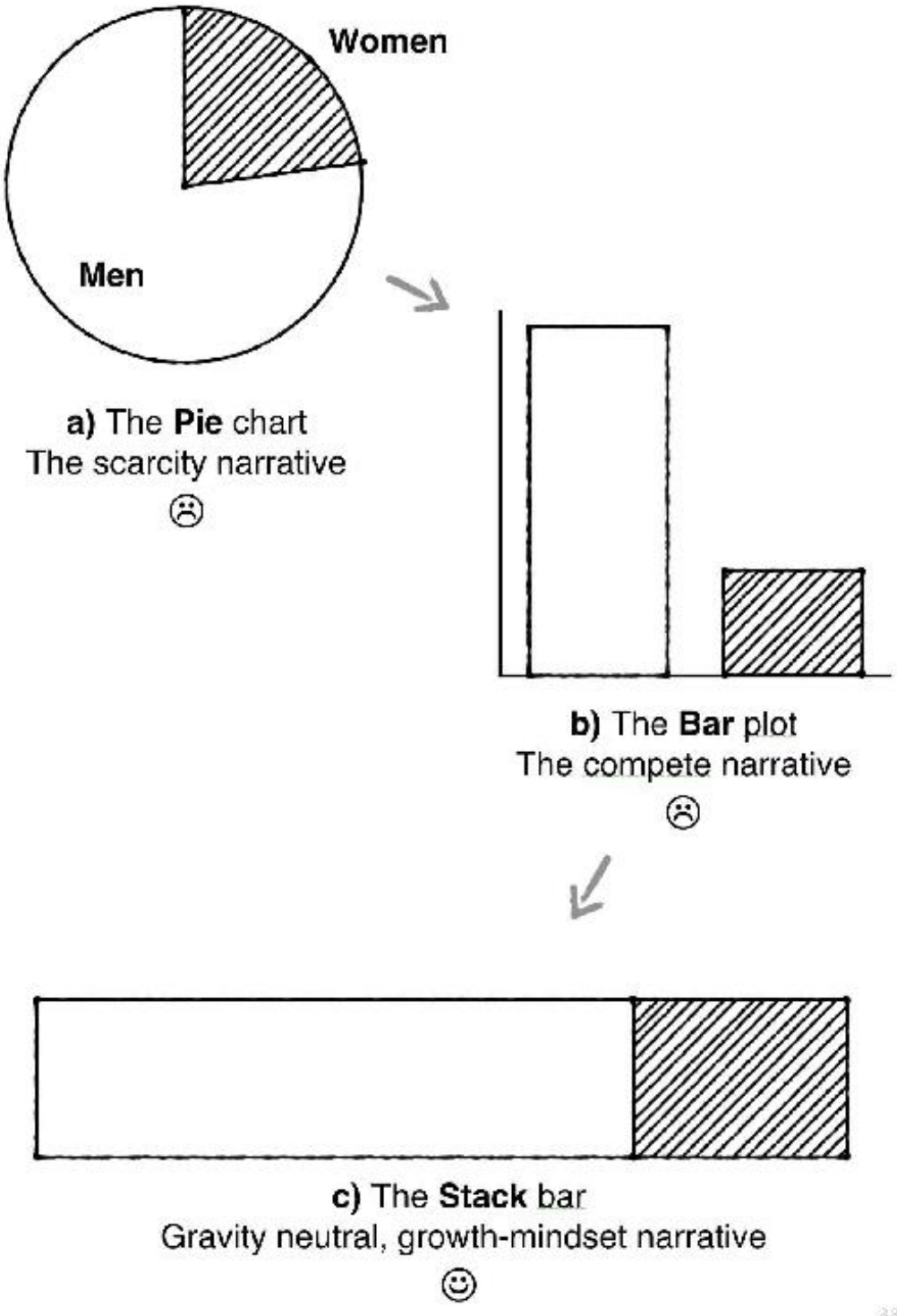


Fig 21 *Three ways to visualize the same data.*

Choosing the right chart

Fig 21 shows the same information as Fig 20 in three different ways, a) pie chart, b) bar chart, c) stack bar chart horizontal. Let's consider the process we followed to produce them.

Simplify

We consider men and women only and ignore the other two outlier groups.

Avoid colors

The men are represented in white and the women in dashed area. We avoided color, because we can. Colors are loaded with symbolisms[17], greyscale less so. Conjugating colors is a very subtle art that is easy to underestimate. Why use more than six colors without proper graphic design training? We have also removed the scales, figures and the axis labels. We traded-off a bit of accuracy for a big gain in clarity.

Consider a narrative

Let's assume we are advocating for a gender equity narrative. We define the equity narrative as a world view in which the world will be a better place if there is less gender ratio imbalance[18].

Check the chart-narrative fit

Once a narrative is set, choosing a **compatible** chart template is the next step to transform information into knowledge. In this case, pies and bars are a poor choice. Why? Because pie charts are connected to the narrative of how much pie each group gets (finite resources), it is a confrontational narrative that undermines the narrative of gender equity. The bar plot chart is also a poor choice because it is connected to the narrative of competing and to other win-lose narratives such as scarcity and the view that the men and women compete. Using them risks undermining any growth mindset[19] or equity (win-win) narrative that you might be advocating for. The stack bar is a better choice. It is horizontal so it is not subject to the gravity metaphor. (See gravity & charts). Now that we have found a fit between the chart and the narrative, let's design the chart in ways that are easier to

assimilate by humans. We call this charts human centered charts. Human centered charts leverage the same principles as human centered design (placing the user at the center of the design process).

Human-centered charts

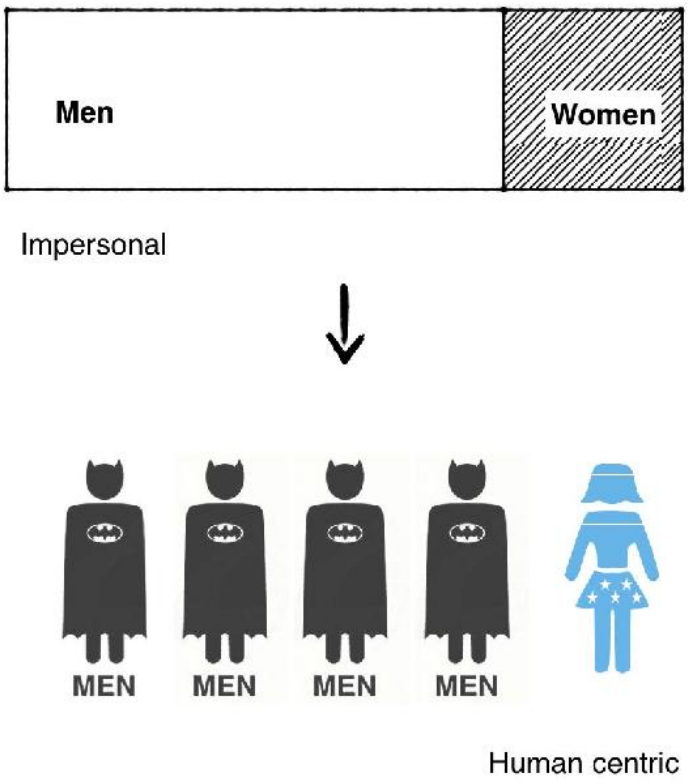


Fig 22 *Four Batmen and a Wonder woman make this chart easier to remember.*

Using superheroes is a one way to visualize the heavy[20] topic of gender equity in STEM dominated environments. We also use the iso-measure. In an iso-measure chart, each unit (in this case a superhero) represents the same quantity (in this case, about 5,000 respondents of the survey). We also reduced the superhero count to the minimum possible (five persons) which means we approximate the original percentages to a 1:4 female to male ratio. This chart can also be read as follows: on average, for every 5 people in a team (room, meeting), one will be female.

Checking the chart-narrative fit

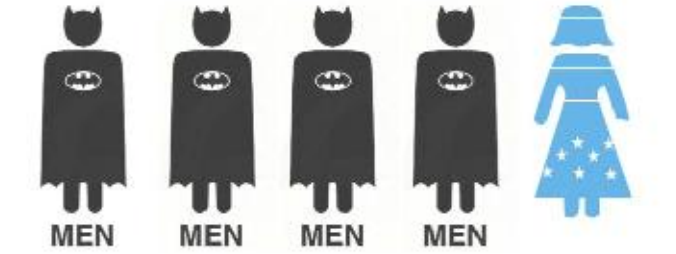
Given the same gender equity narrative used earlier, let's see how the design choices we made fit such narrative.

- A superhero's job is to make the world a better place, the equity narrative is about making the world a better place, this playful theme will help the reader of the chart remember it.
- The iso-measure is humans, like the respondents
- The number of heroes is less than seven, (we are respecting the 7 chunks rule and not overloading the reader with information)
- We use humor[21] to improve the communication effectiveness.

Leveraging humor

Many charts are impersonal because we cannot relate to them. We solved that with the superheroes. See also user personas in Ch. 6) However, if in addition we want the audience to remember the chart, we can use humor or an insider joke as in, data scientists *are* superheroes because they have to "wrangle" with data, see the term *data wrangling.*

Sexism in your chart?



Use of mini skirt might be considered sexist because it perpetuates harmful gender stereotypes

Fig 23 *Always check for blind spots.*

Chart check list

- ☑ Type of chart compatible with the narrative (message)
- ☑ No color overload (3 colors max)
- ☑ One chart, one message
- ☑ Metaphors aligned with the narrative
- ☑ I don’t need to read the caption to understand the chart
- ☑ The caption is used as a synthesis opportunity
- ☑ The caption does not explain the chart again
- ☑ Bias checked by third party

How to bias check? It is important to check for blind spots. Charts are no different. It is prudent to ask for bias check to a diversity of people, ideally with different backgrounds. See Chapter 5 for more on bias.

Gravity & charts

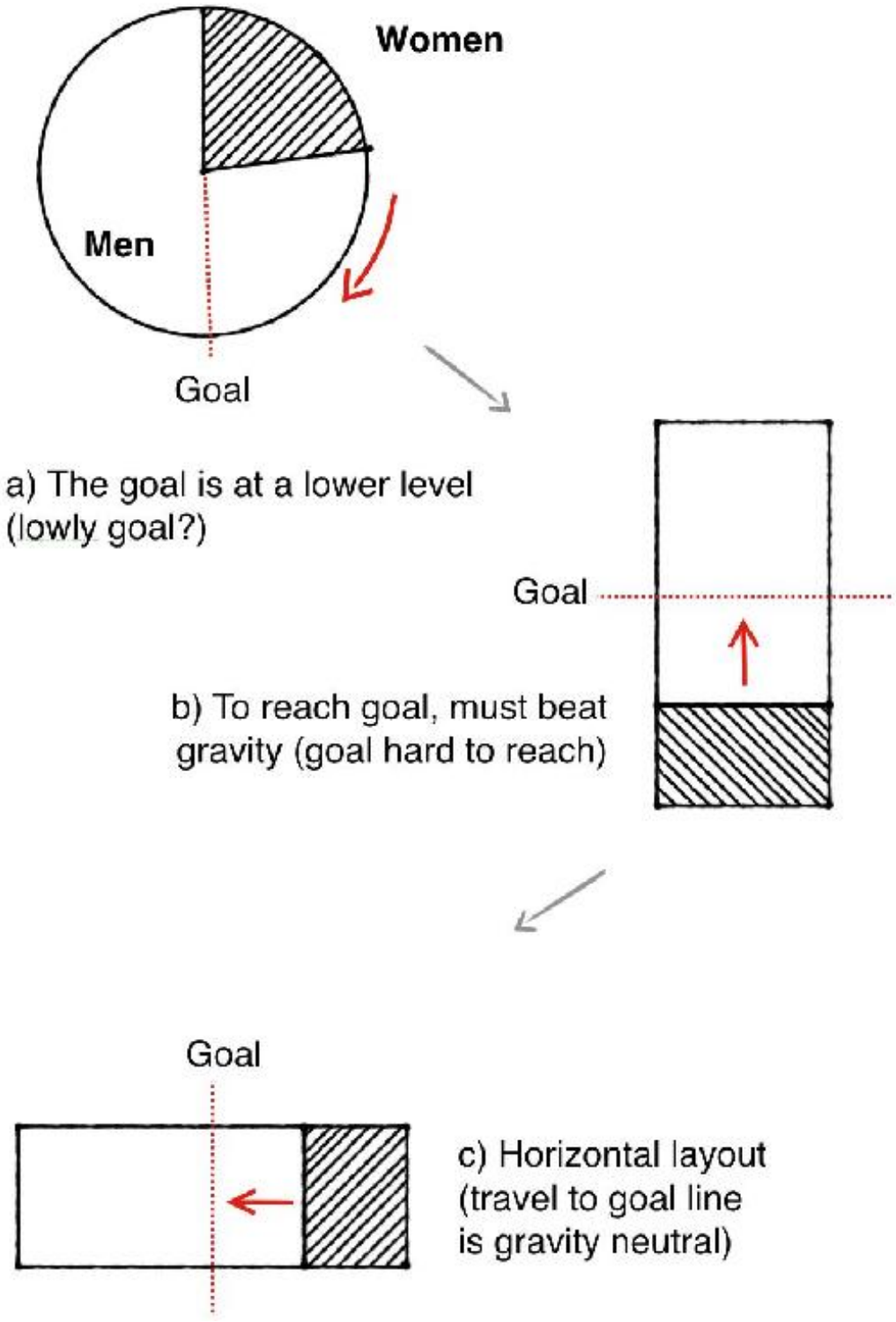


Fig 24 *Gravity shapes everything on Earth, including how we interpret charts.*

How gravity affects goals

Let’s assume your charts have convinced your organization that achieving gender parity at work is a good goal and after a board meeting a goal of reaching a 50% female to male ratio has been set. How to visualize it to **persuade** and rally the rest of the organization? The arrows point to the goal. Note how the alignment of the arrows with the direction of gravity influences how the goal achievability is perceived. In the pie chart, the downward arrow has a negative connotation. For the vertical stack bar the upward arrow against gravity makes the goal appear hard to achieve. The horizontal stack bar is gravity neutral and the freest of connotations that might distract from the narrative.

Chart wars: Pies vs. Bars

Fig 25 *Musk vs. Bezos. Two visions of space exploration. Two ways to visualize altitude.*

When to use pies

There is a fundamental difference between circular charts and bar charts. The brain is sensitive to angular change and (by comparison) numb to linear change[22]. This is particularly true when considering motion, and sensitivity to small changes. If in your narrative, highlighting minute changes in a variable is important for the story, then circular pie charts (speed gauges) are the way to go. If on the contrary, too much attention to change is a distraction, avoid pie charts. Compare for yourself. In the Blue origin cast, the attitude change is barely noticeable. Whereas in the *SpaceX* cast, it pops during all the cast.

- Blue Origin cast: http://bit.ly/2NHycmf
- SpaceX cast: http://bit.ly/2XwXYxY

Quiz: Making useful charts

True or False? Time 10 minutes.

1. A bar chart is a great way to visualize the odds of the Casino Roulette. [T/F]
2. Humans are more sensitive to pie charts than bar chars. However, in animated charts where the ratios change, these changes are more obvious in a bar chart. [T/F]
3. A drawback of the iso-measure chart is that one cannot use metaphors or personas with it. [T/F]
4. A way to turn information into knowledge is to use a palette with the minimum amount of colors possible. [T/F]
5. The purpose of an EDA is to tell a story about the data. [T/F]

(Solution in the next page)

Solution

1. A bar chart is a great way to visualize the odds of the Casino Roulette [T/ **F**]. **False**. A bar chart is linked to the compete narrative. The roulette is a zero-sum game. A Pie chart, an **iso-measure** chart or a photo of the roulette itself communicates odds more clearly.
2. Humans are more sensitive to pie charts than bar charts. However, in animated charts where the ratios change, those changes are more obvious in a bar chart. [T/F]. **False**. They are more sensitive to pies/ needles in **both** situations.
3. A drawback of the iso-measure chart is that one cannot use metaphors or personas with it. [T/F]. **False**. It is the **opposite**. The iso-measure lends itself to metaphors.
4. A way to turn information into knowledge is to use a palette with the minimum amount of colors possible [T/F]. **False**. **Avoiding** information overload is a pre-condition for knowledge. However, the fundamental transform is usefulness. (Usually, by connecting it to other knowledge).
5. The purpose of an EDA is to tell a story about the data [T/F]. **False**. The **main** purpose of an Exploratory Data Analysis is to aggregate and visualize basic statistical information.

Chapter 3. Knowledge

– How to create knowledge with frameworks –

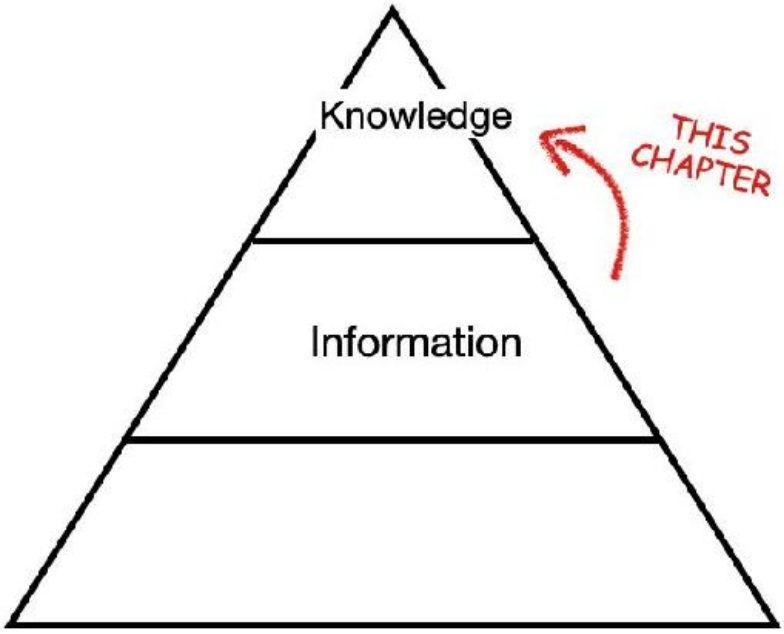


Fig 26 *This chart uses the gravity metaphor.*

In Chapter 1, we saw that charts with purpose have a *message*. In Chapter 2, we saw an example of how to transform gender data into information. In this chapter, we will learn how to synthesize information into knowledge. If we were in an English language class, this skill would be the equivalent of writing the book synthesis. Your data is the book; the chart is the synthesis. An effective way to do this is by using reference

frameworks, summarization techniques and visual metaphors. Let’s see an example that uses age data from a survey.

Mental frameworks

Let's take a look at the chart in Fig 27. How many chunks of information can you count? It has information overload. Let's focus on the color palette for instance, a rainbow. However, a rainbow does not convey any meaning here. On the contrary, by using 12 colors, we have increased the information overload by a whopping **12 memory *chunks*** with a zero gain in meaning, (see information vs. meaning in Ch. 1).

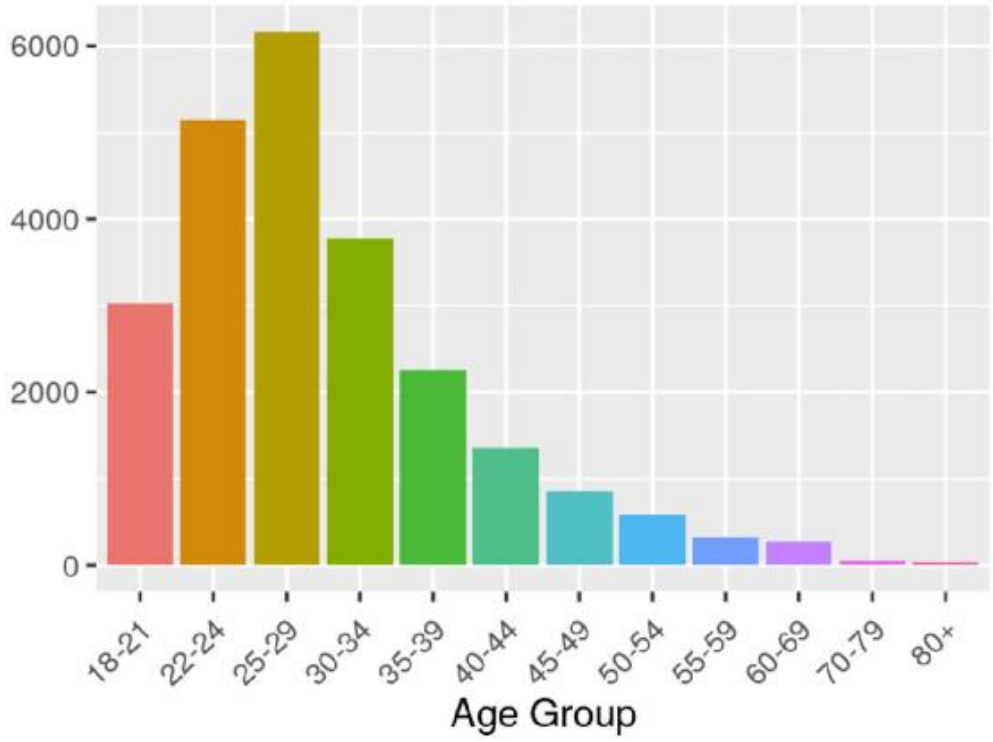


Fig 27 *A ggplot default visualization of the age of the respondents of the Kaggle 2018 survey. Color scale: Rainbow, 12 colors.*

Exercise

How would you make the previous chart more useful? We can start by reducing the information overload. Draw solutions. Time 2 minutes.

Solution

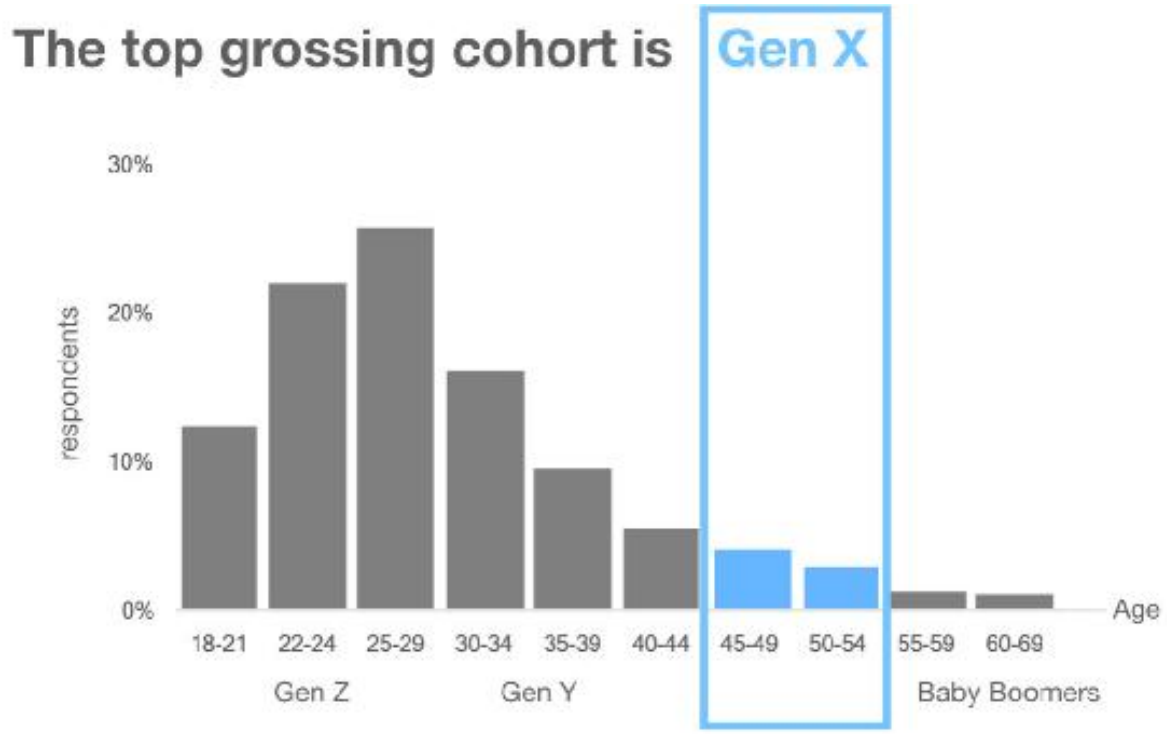


Fig 28 *Same information with a narrative and linked to a reference framework, closer to knowledge.*

What can we learn from senior data scientists?

Like Fig 27, Fig 28 is user distribution by age too. However, here we use a two-color scheme to highlight which age-group won most competitions per user. However, just a few too many age bins can overwhelm any reader. A way to declutter and structure the bins into **usable** information is to reduce their numbers and group them in a familiar, relatable form. One way to do this is to reduce age groups to generation groups. In this case, we used a reference framework many are familiar with: Generations in the workforce. It comprises the gen X, Y, Z and the Boomer[23]. Furthermore, we are interested to see which group is the most productive in terms of competitions and cash prizes per user. Because everyone belongs to a

generation this chart can become very personable. What can we learn from the wisdom that each generation offers? Source: 2018 Kaggle Survey Q2 What is your age?

Generation & work-ethic attribute framework

- The **Baby Boomers**, born 1946 – 1964, “often branded workaholics[24]”
- **Gen X**, born 1967 – 1977, “this generation works to live and carry with them a level of cynicism”
- **Gen Y**, “Millennials” born 1980 – 2000, “considered the most educated and self-aware generation in employment”
- **Gen Z**[25], born after 2000

Narrative

Ageism is the narrative that says things like “old people do not have energy to be entrepreneurial and cannot innovate”[26]. However, a few papers [27] have disproved this claim. The chart advocates for the opposite narrative, “(in data science), older generations are as productive as younger generations”.

Knowledge creation

Note how the key step to create meaning (knowledge) is not only to summarize and declutter, but to find **where** the information is most useful and then by linking it to **that** context (reference framework). In this case, Generations in the workplace and productivity. Another way to create meaning is by way of visual metaphors such as the pyramid. Let’s see an example based on salary information.

Visualizing inclusion

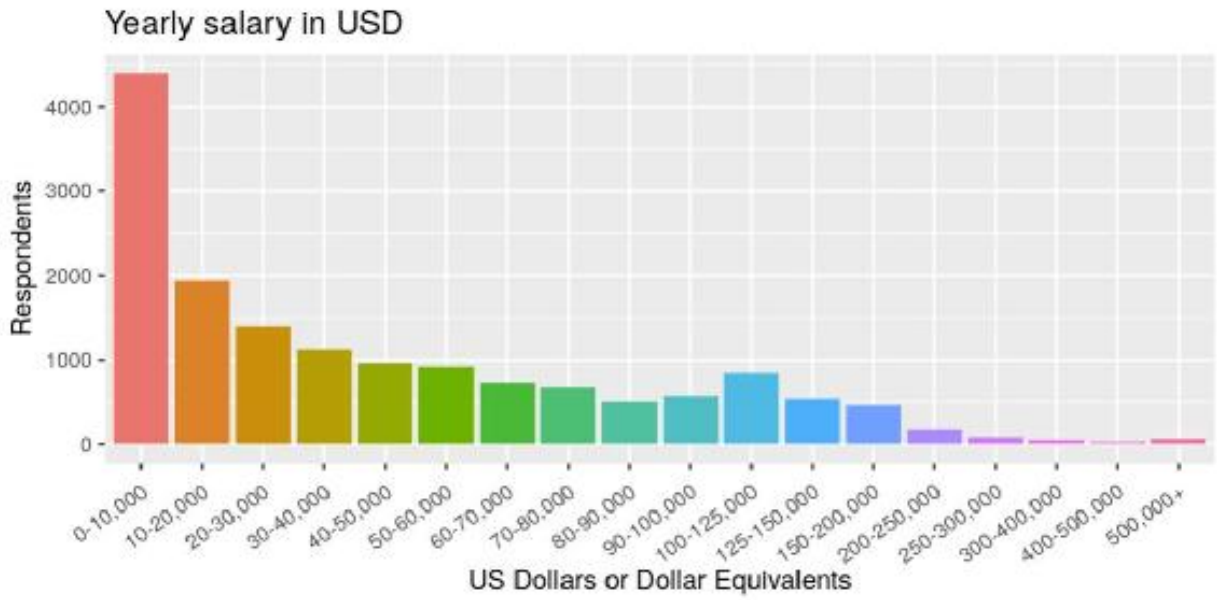


Fig 29 *A default ggplot2 visualization of the salaries of the Kaggle 2018 Survey. Source Q9 What is your current yearly compensation.*

Exercise

In the previous section, we saw an example of meaning creation by way of connecting information to a reference framework. Now let’s do the same and, in addition, let’s apply a visual metaphor. Let’s look at salary data from the same 2018 survey. Is Fig 29 data information or knowledge?

Given an inclusion narrative, how would you create a more useful chart? Time 6 minutes. Hint: If the chart was a building where would highly paid individuals own apartments?

Intermediate step

(Fig 29 rotated counter clockwise 90 degrees)

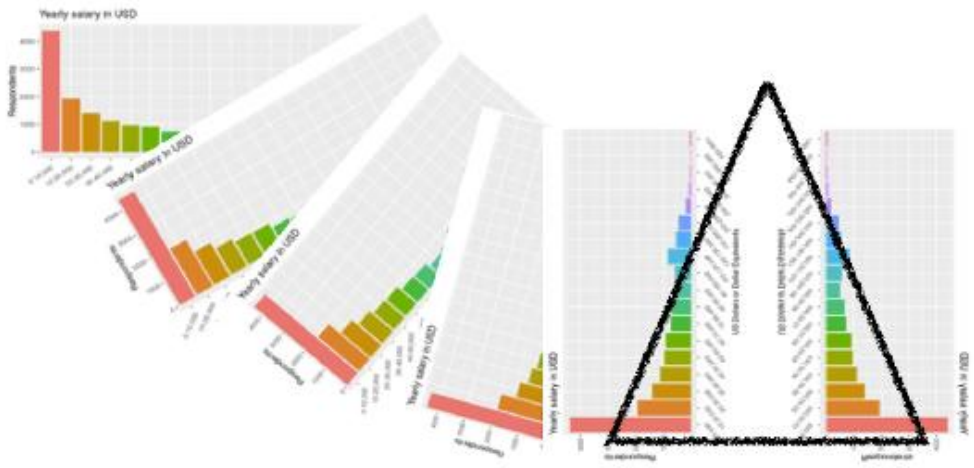


Fig 30 *Does your chart fit any visual metaphor?*

Solution to inclusion

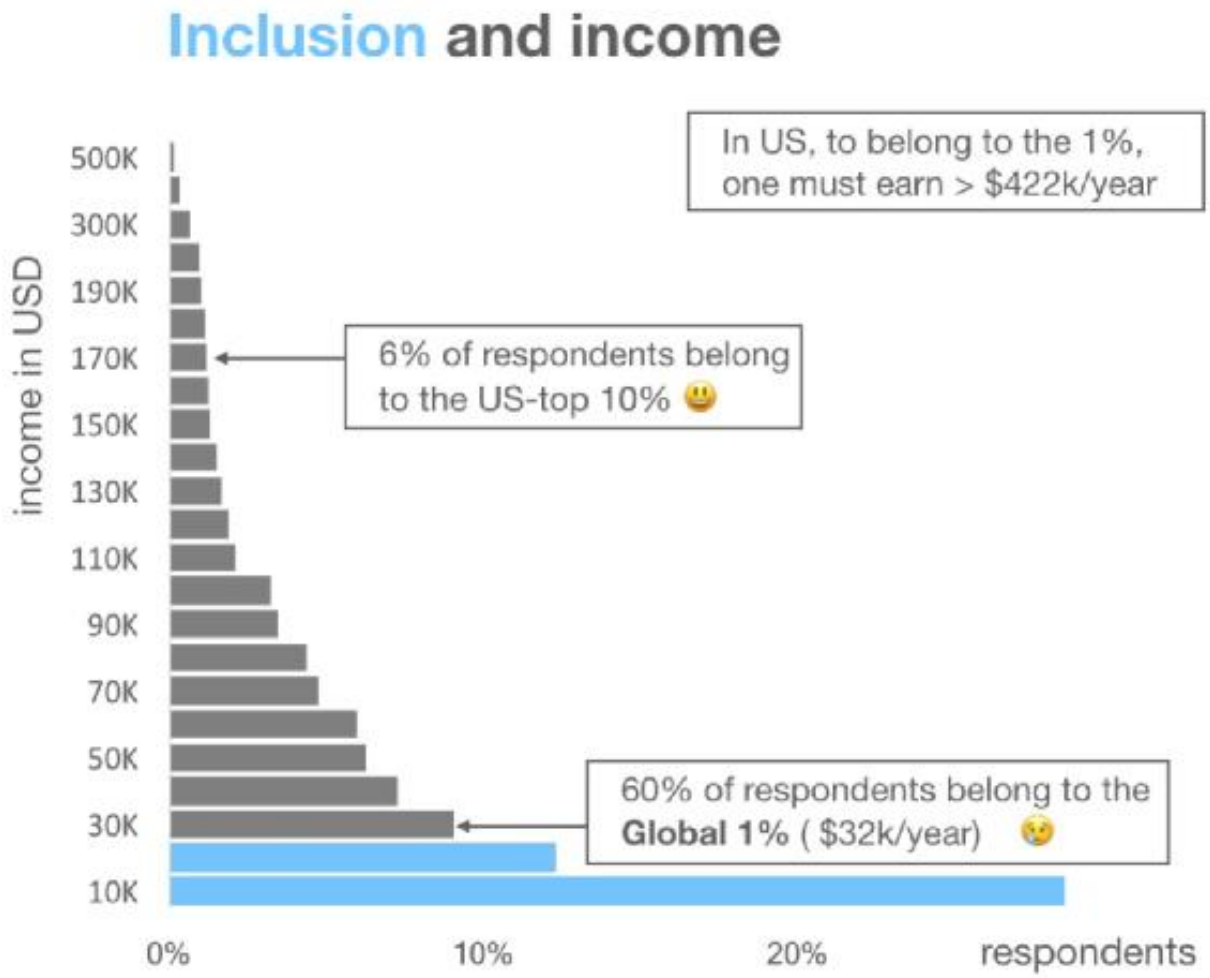


Fig 31 *Using text inside charts is a great way to defuse ambiguity.*

Earlier we mentioned the importance of the chart-narrative fit. Fig 31's narrative is about the digital divide. How inclusive is data science as a community? As we just saw in the previous section, one way to create knowledge is to relate our information to existing reference frameworks. Generations in the workforce was an example of a popular framework to think about age and workforce. Helping the reader **situate** the new information into existing frameworks makes the new information easier to assimilate, use and recall. Here, we apply the reference framework of

income percentiles – a common analysis framework used by economists and fit it with the pyramid a metaphor that represent hierarchies. See also the #onepercent movement. In the US, to belong to the 1% elite, one needs to earn more than $422k per year[28]. About 23 respondents declared that they do. In addition, about 6% declared they belong to the 10% percentile, a very inclusive number because 6% is similar to 10%. The 10% percentile income is about $166k in US[29], so if the sample reflects the distribution found in society, it means it is at least somehow inclusive. We add a smiley emoji to reassure the reader that yes, this is good. However, those numbers are for US household incomes. When we look globally, the 1% percentile threshold is $32k per year. This puts 60% of the respondents in the top 1%. 60% is very different from 1%, so globally this data point does not support inclusiveness because it does not reflect the global distribution. Aha moment. One way to create such moments in the story is to A/Bify the story by switching between two points of view. We just saw how powerful visual metaphors can be. Let’s see three more examples.

Winner takes all

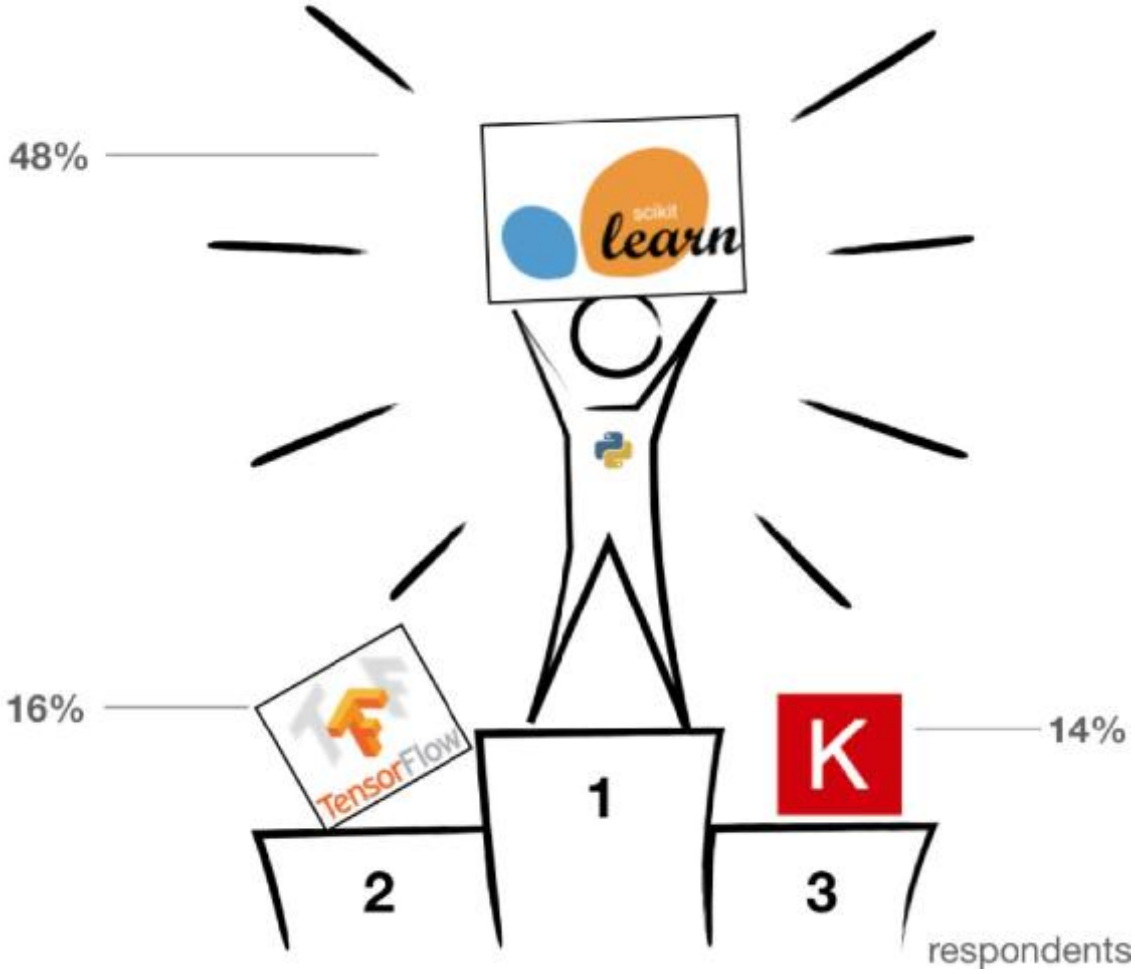


Fig 32 *The top three ML libraries.*

Here, we are visualizing the data science libraries the respondents use. Using survey Q20: Of the choices that you selected in the previous question, which ML library have you used the most? Given a winner takes all narrative (so common in the software world), what visual metaphor can we apply? This chart is an example of less is more. In this case, Sci-Kit (a famous scientific Python library) has a 48% share, Google's TensorFlow

has a 16%, followed by Keras 14%. Let's see how this visual is connected to the narrative.

Chart-narrative fit

In scenarios with strong network externalities at play such as social network, a phone OS or an Olympic race; being on the podium (being first) has a disproportionate effect on the reward. In such cases, the winner-takes-all narrative is in place. Anthropomorphizing the ranking with a podium conveys a memorable narrative and affordance — the glory the winner deserves for the great utility this library provided to the community. This narrative is also connected to other memes famous in the software world such as the developer’s glory. (See S. Balmer in “Developers, developers, developers, developers”).

All or nothing

In the previous section, we visualized data about the most popular ML libraries with a winner takes all narrative, here we do the same with a different narrative.

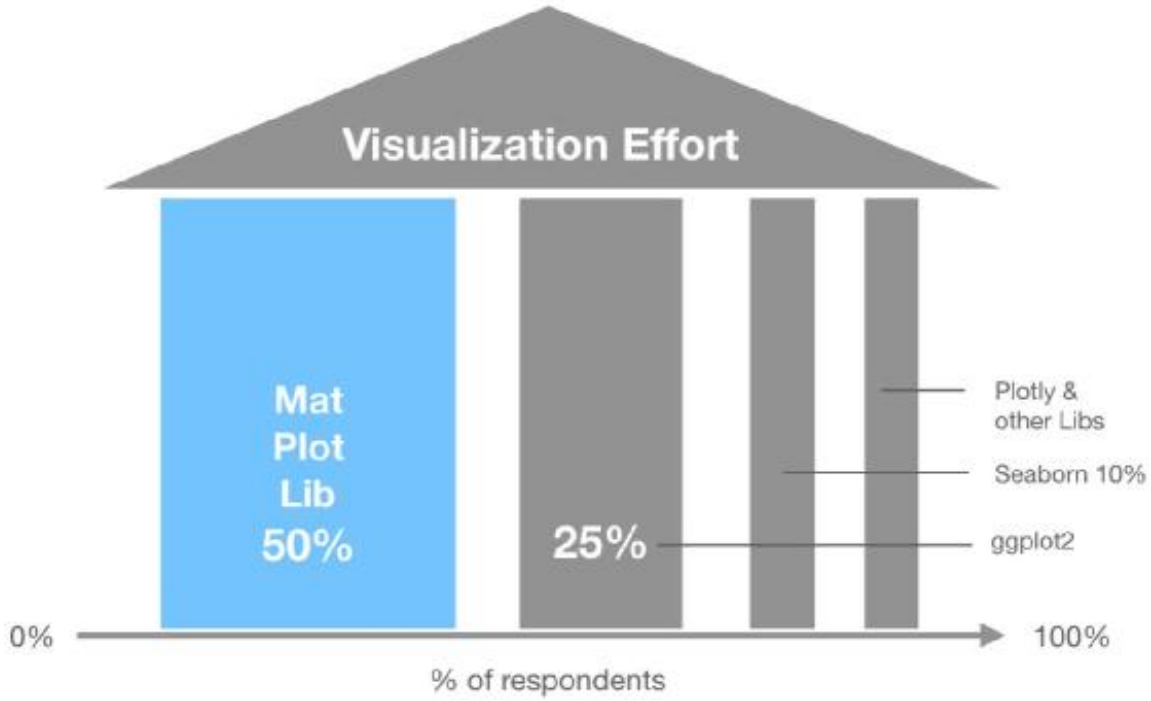


Fig 33 *The house of Shiva. When colored areas occupy large areas, use 50% grey and pastel colors instead of 100% solid bold colors.*

House of Shiva

Fig 33 is a combination of, (i) a chart template called Marimekko, with (ii) a symbolic chart called *House of Shiva*. The House of Shiva is used to emphasize all or nothing relations. The metaphor is that the roof falls if just one column collapses.

Symbolism

The columns support the visualization efforts of the community (roof load, common good). The width of the "columns" expresses how much

work/load each column supports. Grey columns on the right represent other less mainstream libraries such as: D3, Shiny, Bokeh, Leaflet, Lattice. Source: Survey Q22 Which specific data visualization library or tool have you used the most?

Metaphor

The goal is the roof. As with a house, the integrity of it becomes clearly impaired if one column is weak.

Narrative

The narrative here is that non-mainstream visualization libraries are important but with different degree. Note here that if we had used a pie-chart we would have conveyed a win-lose scarcity narrative, not faithfully representing the win-win ethos of the open source movement.

The BRICS framework

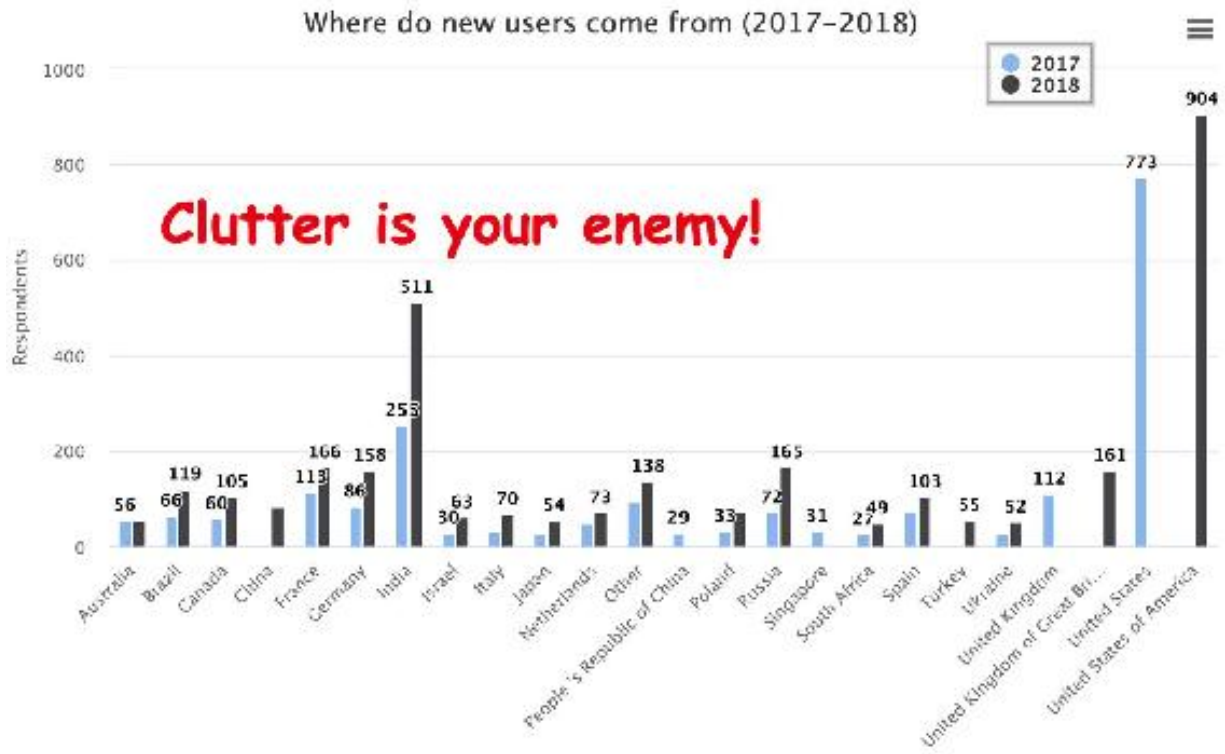


Fig 34 *An example of meaning loss by information overload. Source: http://bit.ly/2K7ZLBk*

A common way to summarize international data is to count grouping by country. However, this often leads to clutter because there are more than 200 countries in the world. Another common way is to use a geographical world map, and to modulate the color of the country with the count. However, some countries are very large in extension while other countries become a pixel on the screen. This is not great for usability.

Exercise

Summarize the previous chart. There are more than 200 countries. Only 24 are shown in this view[30]. How would you go about this? Draw solutions. Time 5 minutes.

Solution to BRICS

If we look at the 2018 survey and compare it to the 2017, there is an increase of 1145 new respondents that identify as “data scientist”. So, where do new data scientist users come from? Fig 34 displays 2018 and 2017 data. However, there are too many countries for a human to make sense of it! Remember that a human brain’s working memory is limited to 5-7 chunks [31]. This means we cannot juggle more than seven countries at the same time and neither should your chart. What would Marie Kondo do?

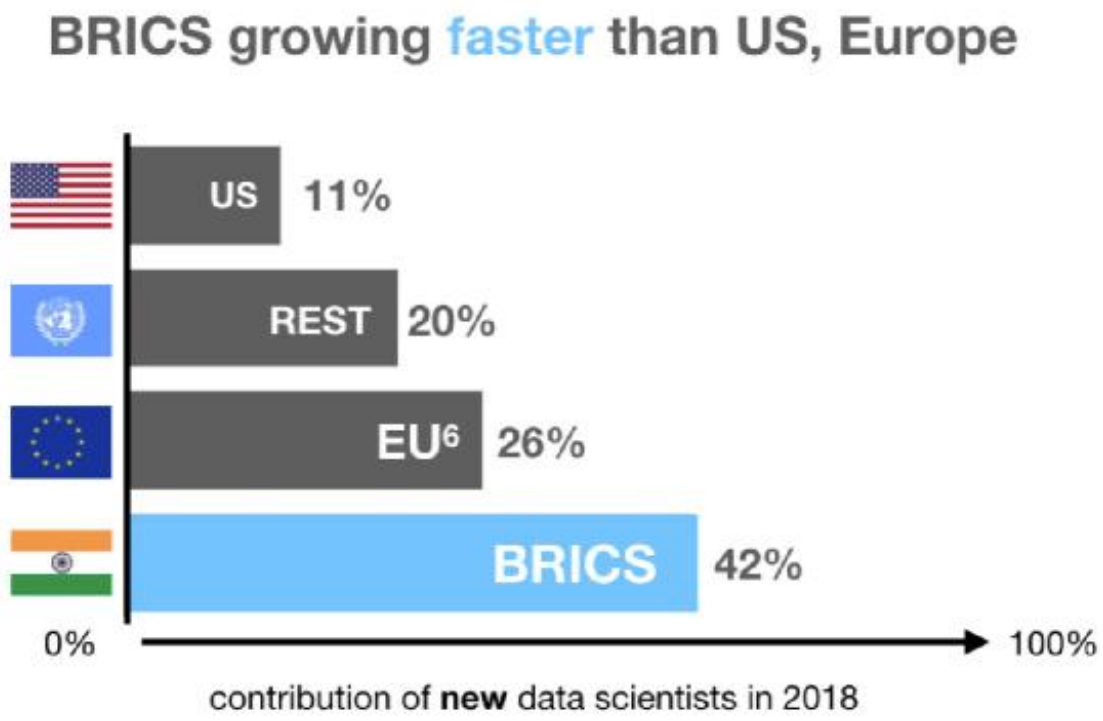


Fig 35 *This chart was made with PowerPoint because it was faster than tweaking the parameters of ggplot. Notice how the golden ratio is used across the chart.*

Use Mega Regions

One way to summarize in a more humane way is to group countries in economic mega regions that have a certain degree of homogeneity (cultural or economic). In this case, we apply the BRICS framework. BRICS stands for Brazil, Russia, India and China. It is a grouping coined by Goldman

Sachs to classify countries with similar economic indicators under one label. (The term “BRIC” was coined in 2001 by then-chairman of Goldman Sachs Asset Management, Jim O’Neill). This framework provides a meaningful way to group countries by bagging them by economic and social development affinity. In this case, we chose four groups: the US, Europe, BRICS and the rest of World. When we do that, we see that not only is BRICS the top contributor to growth with 42% of total growth for 2018, but it was also the fastest growing among the big three. In 2018, in the category “users that define themselves as data scientist”, Europe added 302 users, US 131, the rest of the world: 231, and BRICS: 481. By 2020 the growth of BRICS will outnumber Europe and US combined. Data source: we forked and modified a snippet of the code from Kaggle user ash316 and took the top 20 countries whose respondents identified as “data scientist”. EU-6 means the top 6 European countries.

Aesthetics

Note how different font sizes are used in the chart. How the horizontal layout neutralizes the gravity metaphor. And how flags and labels are used. The individual countries growth added together should stack up to 100%, What narrative would we fall into if the author had used a pie or stack bar?

Narrative

The narrative here is that to be wise, one must look where the ball is going to be not at where it is now. Don’t look at absolute numbers of 2018 only, look at growth. Of course, an alternate narrative is that emerging countries are catching up in strategic fields such as data science.

Quiz: Visual summaries

True or False? Time 5 minutes.

1. Visualizing 200 country names in an 800x1200 display is ridiculous because there are not enough pixels. [True / False]
2. Visualizing 200 countries in a chart violates the 7 chunks limit rule. [T/F]
3. Using PowerPoint or Adobe to build a chart one needs is professional. [T/F]
4. To visualize ‘a winner takes all situation’ we can use a Marimekko chart. [T/F]
5. House of Shiva is best to visualize all or nothing relations. [T/F]

(Solution in the next page)

Solution

1. Visualizing a chart with 200 country names in a 800x1200 display is ridiculous because there are not enough pixels [True / False]. False. It is ridiculous because a chart with 200 text labels is information overload unusual.
2. Visualizing 200 countries in a chart violates the 7 chunks limit rule [T/F]. True.
3. Using PowerPoint or Adobe to build a chart one needs is professional [T/F]. True. That is how Bloomberg Businessweek charts are built.
4. To visualize a ‘winner takes all’ situation we can use a Marimekko chart [T/F]. False. Podium 10x is clearer in this case.
5. House of Shiva is best to visualize ‘all or nothing relations’ [T/F]. True. If one column is missing the roof falls.

Chapter 4. Charts to Think

– How to use charts to make wiser decisions –

So far, we have seen examples of how to transform data into information and information into knowledge. Now let's consider prescriptive analytics — the charts the board room uses to decide what to do in policy and decision making. Unfortunately, policy challenges are complex to deal with. Particularly, when the variables involved have unexpected dependencies that are not properly understood, or even known!

A way to deal with these complex problems is to simplify them by describing them in a more humane way. One popular way, is to transform the problem into a map — a process called mapping. Mapping can be done if we just use two dimensions to describe the problem. Once in 2D, the canvas becomes a design space where we can search for solutions. Because we are in a map setting, we can leverage the extraordinary spatial cognitive abilities that have been gifted to us[32]. Glorious examples of "mapping" are war rooms, the BCG growth-share matrix, Gantt charts, The Business Model Canvas, Kanban boards, Gartner's Magic quadrant and Wardley maps — to date, the most advanced thinking tool to think about strategic innovation.

Let's start with 2D mapping. Imagine you are a high official in charge of innovation in Singapore. You have only been given Fig 34's data (number of data scientists per country). How would you use this data to inform the next innovation policy looking forward? Hint: A first step is to create situational awareness. One way is to rank the countries, so we can see where we stand. The second step is to use a meaningful indicator in the x-axis. (Solution in the next section).

Use rankings to create situational awareness

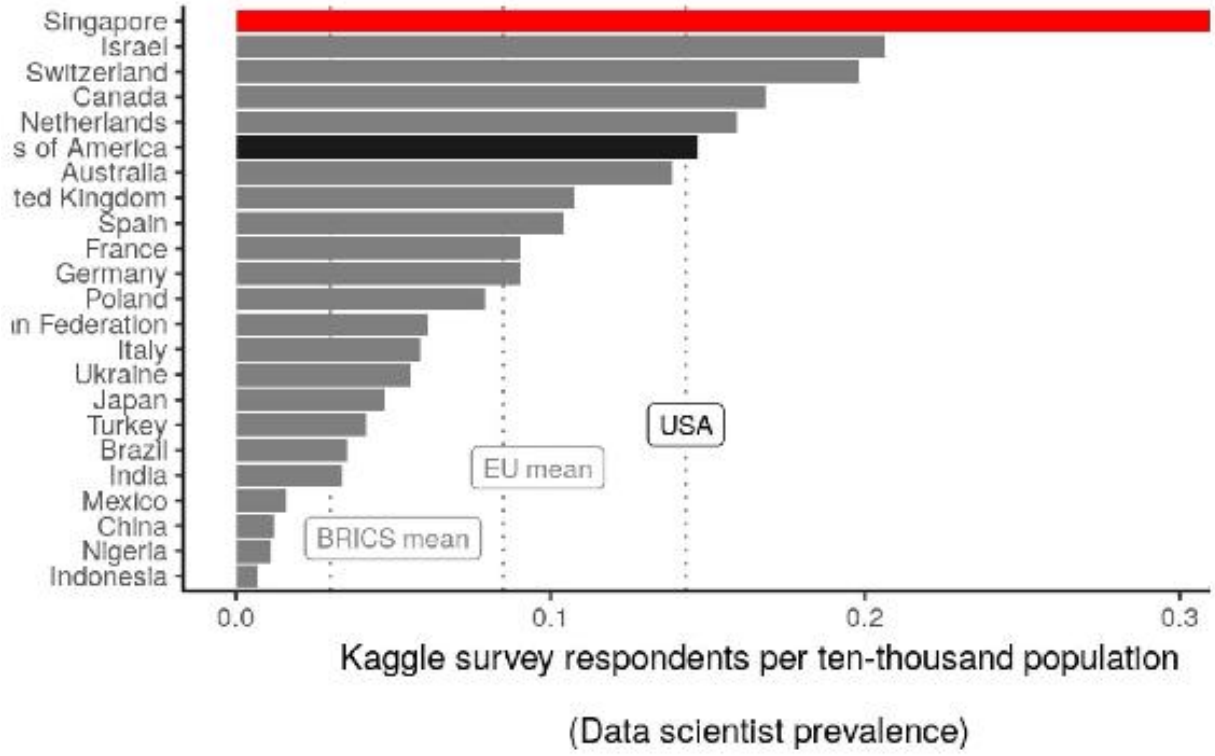


Fig 36 *Two is the maximum amount of colors you should use in a chart. Grey does not count.*

It's per capita or die

As the economist and comedian Harald Eia politely implies in his Oslo TEDx Talk, when comparing countries we are hardly ever interested in absolute numbers. In fact, any non-per-capita measure is rather useless. Yet, how often we forget this! Charts not normalized per country population are setting themselves up for stereotypes and unhealthy narratives. From the 2018 Kaggle data science survey, it is possible to count the number of data scientists in each country, and most participants rendered such charts. But how useful is that? We already know that China and the US are large countries, so in absolute numbers they will also have a larger quantity of data scientists. That is expected. However, if it is expected it means we already knew it. The more expected, the less information (100% expected

means zero information). What is useful, is to look at per capita measures. How do countries compare on data scientist density? In this chart, we highlight Singapore in red and the US (home to the largest survey community, like a center of mass) in black so the reader has a reference point. This chart has a lot going on:

- USA mean: 0.14 per 10,000
- EU6 mean*: 0.09 per 10,000
- BRICS mean*: 0.03 (5x less than US)

Mind the gap

Mind the gap is a common strategy to think about differences between categories in the data, in this case, countries. Thinking about why the gap exists can help explain the reality that the chart is representing. For example, a linguist might think about the gap in terms of English proficiency and its correlation to the prevalence of data scientists. Is the language barrier an explanatory factor for the gap? What are the policy implications? Note: The BRICS, and EU6 mean is mean of country means, not weighted by respondents. Source: World Bank Population Data 2016, Survey Q11 Current country of residence.

Aesthetics

This color scheme is called the red on grey, it is my favorite scheme for charts. Unlike, other schemes such as purple on grey, it is gender neutral. However, for it to work, the red surface must be kept to a minimum, otherwise it comes across as strident.

Narrative

Michael Porter’s Competitive Advantage of Nations.

Exercise

Let’s take Fig 36 chart a step further. Earlier we saw that frameworks can help the reader make sense of new information. Using the *global innovation*

index, how would you relate it to the prevalence of data scientists? Use 2D mapping. Time 5 minutes. (Solution in the next section).

The design space

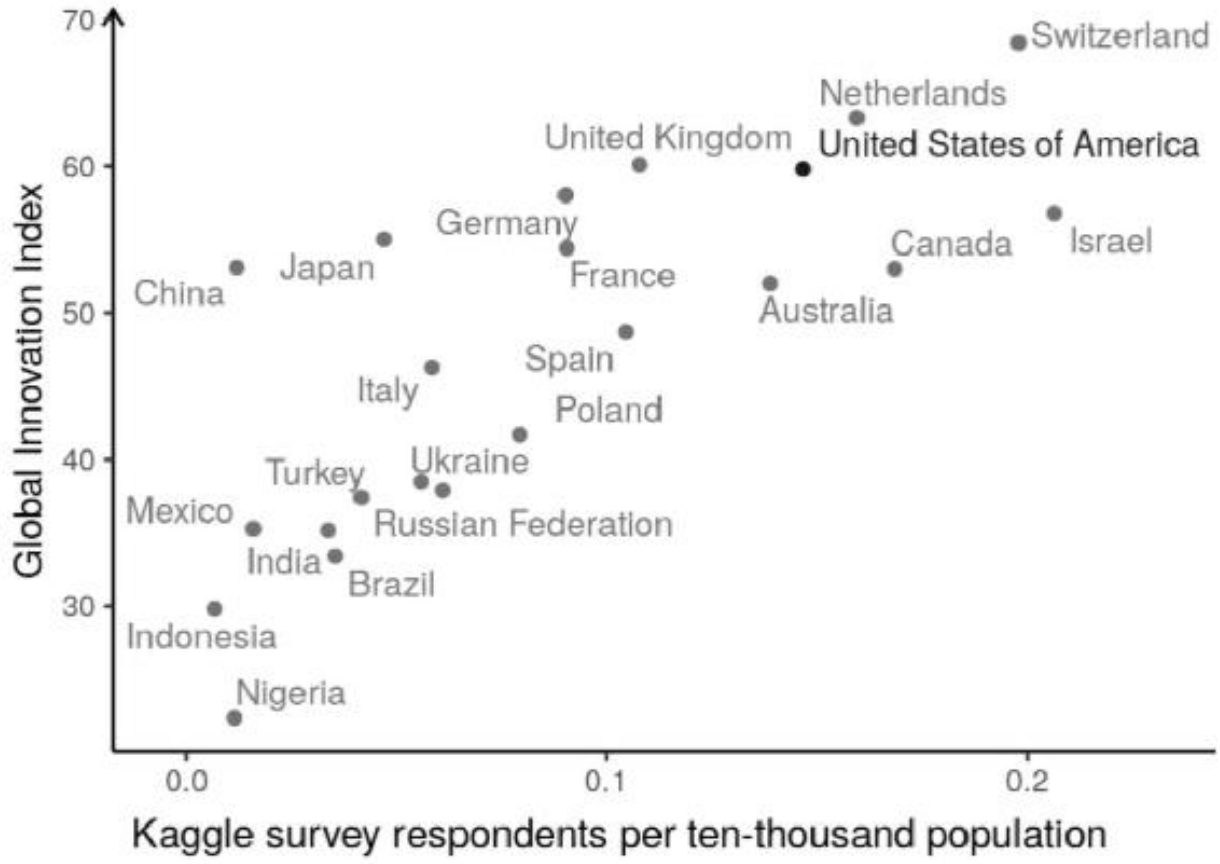


Fig 37 *A scatter plot between two correlated variables always yields a similar cloud.*

As Mr. Wardley would say — When you need to understand the territory it helps to have a map. Here we use 2D mapping by scattering the countries along two dimensions[33]. The technique of projecting into two dimensions has been successfully used in famous charts such as Wardley Maps, the BCG growth share matrix, The Urgent-Important matrix and Gartner's magic quadrant. This map can be used to cluster countries by policy to help elucidate success factors that influence the position in the map (See also Gapminder).

Narrative

Porter’s Competitive Advantage of Nations.

About the Innovation Index

Every year, INSEAD MBA, Cornel University and the WIPO publish the Global Innovation Index. In 2018, the most innovative country was Switzerland. A Spearman rank correlation between GII and user prevalence yields 79%.

Exercise

Let’s take this chart a step further. One of the most valuable skills is prediction. Given Fig 37, can you predict where Japan will be 10 years from now? Use a linear regression. Time 5 minutes. (Solution in the next section).

Forecasting with mean-reversion

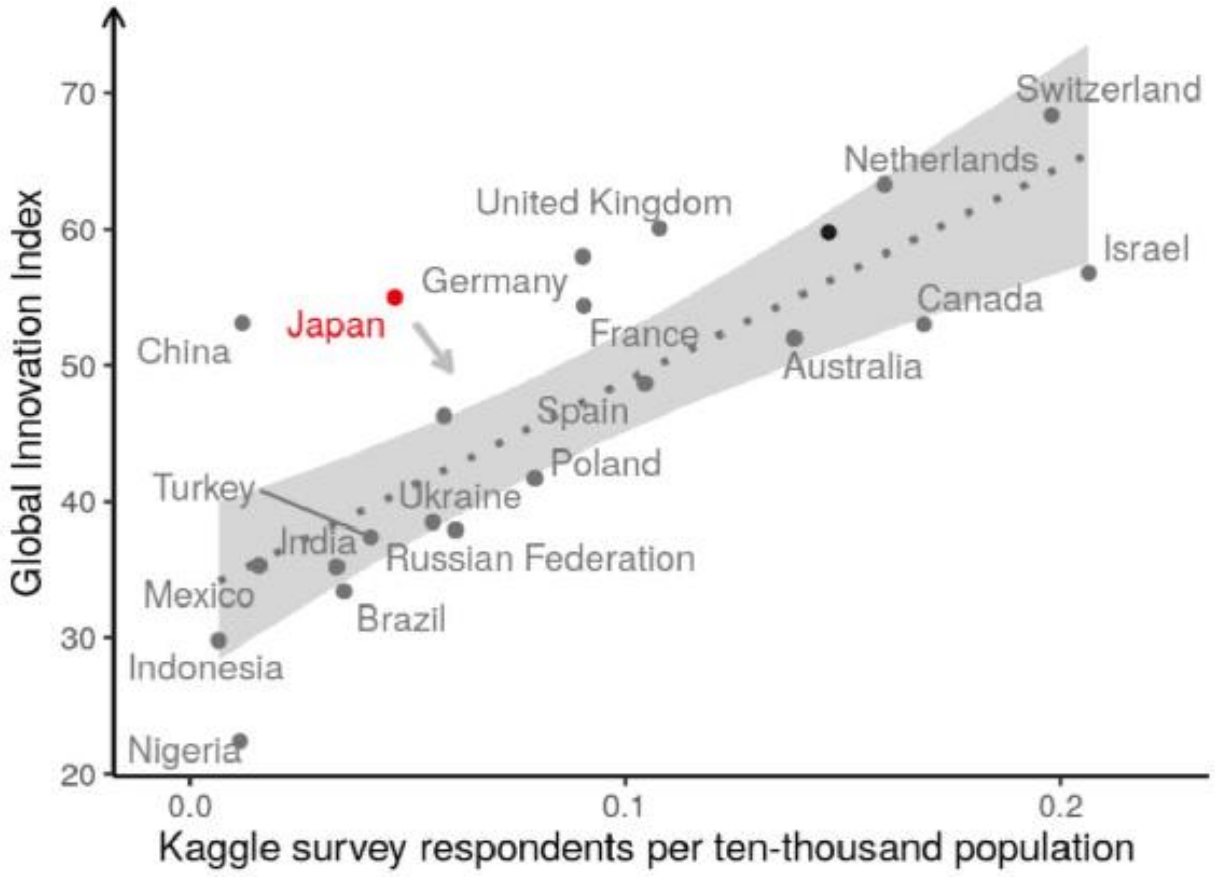


Fig 38 *Mean reversion, always right in the long run?*

Here, we just added a regression line and removed the outlier Singapore. The 95% standard error margin is shown in grey. Some countries are below and some above. Highlighted in red is Japan, as an outlier with high in Innovation Index (y-axis) but low in x-axis relative to peers. Let's assume that the principle of mean reversion applies here as a baseline predictor and a hidden hand continually pushes countries towards the mean (dotted line). The principle of mean reversion is based on the idea that there are no

permanent competitive advantages to either companies (See introduction chapter in Blue Ocean Strategy) or nations. It has shown its worth, particularly in finance. For example, in betting on the composition of the DOW JONES, very few companies have what it takes to last long in the Dow Jones. Of the original members of the index formed in 1896, only GE remains.

Reflection

What can we forecast about the 2019 GII rank position of Japan? Applying the principle of mean reversion, it is unlikely that Japan will increase its rank because it is already high. Even if Japan catches up in data scientist prevalence, it is likely that it will still go down (towards the mean). Indexes are just weights. Assuming the Data Science weight in the innovation economy will only increase in the coming decades and that the GII index calculation method will be updated accordingly, what countries are more likely to improve their “nominal" ranking in 2019? When the GII index weights are rebalanced, is it likely that countries Canada, Australia will jump a few places? Source: Global Innovation Index 2018, World Bank Population Data 2016, Q11 - Current country of residence.

A note on the origin of linear regression

The name linear regression as in the line that minimizes the sum of the square of the errors, was popularized in a paper where the principle of “regression to the mean” was verified in how offspring height is related to the parent’s height. Spoiler alert! Only 60% of the offspring height is explained by the parents’ heights. The rest is explained by the mean of the race. Which means that the Mean reversion principle applies in height with a 40% influence approximately. However, the mathematical method is completely unrelated to any concept of regression. The paper got famous and the regression word stuck to the method. A great trick question is to ask students to explain why linear regression is called linear regression. I am always amazed at the inventiveness of some students[34].

The design space in business

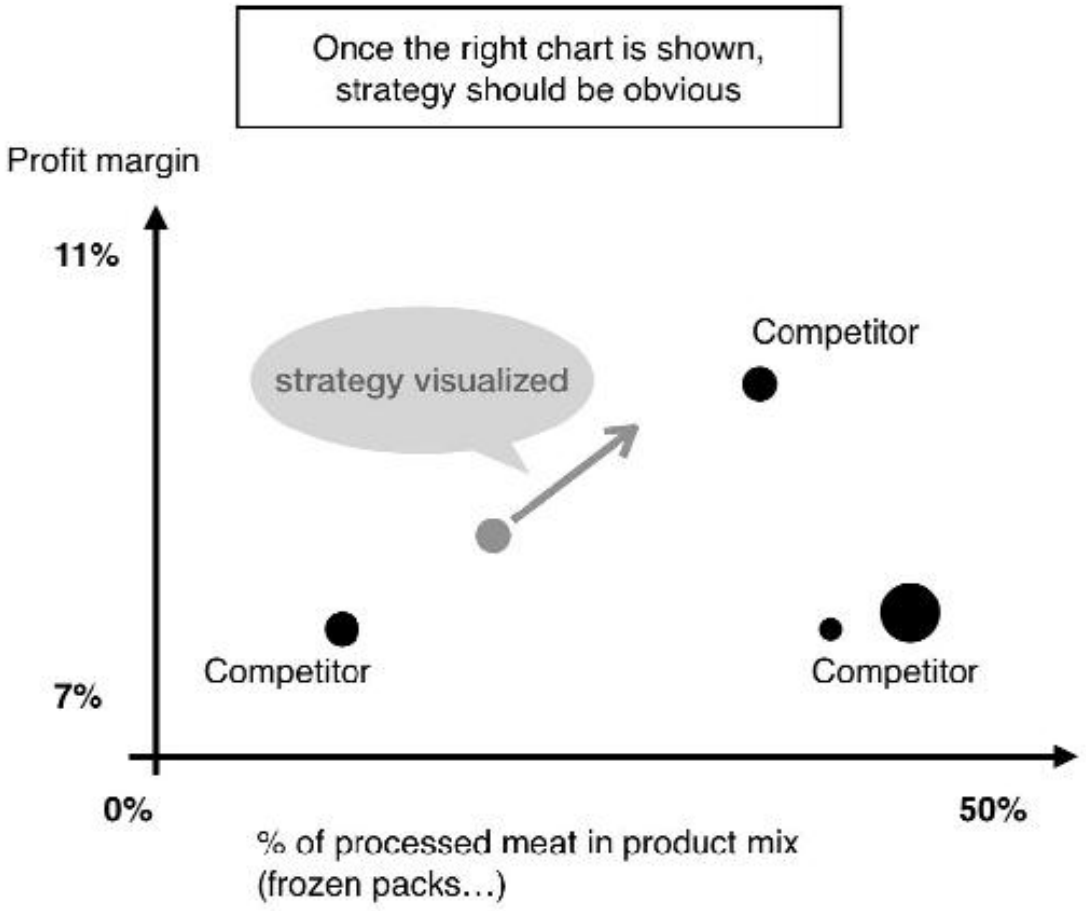


Fig 39 *Can you visualize a problem as a map?*

Reframe it

One of the most important roles of a data scientist, is to realize when the customer cannot articulate his own needs (see *Jobs-to-be-done* theory). This skill is what distinguishes the A+ data scientist from the rest. The chart here is adapted from the book *The Accidental Investment Banker*. The author, a banker, came up with it during a business engagement. He used it to map out the M&A strategy for a client. Once he made this chart, everybody in the room could visualize where value was. In his book, he credits this chart as an important moment in his career.

The Gap Matrix

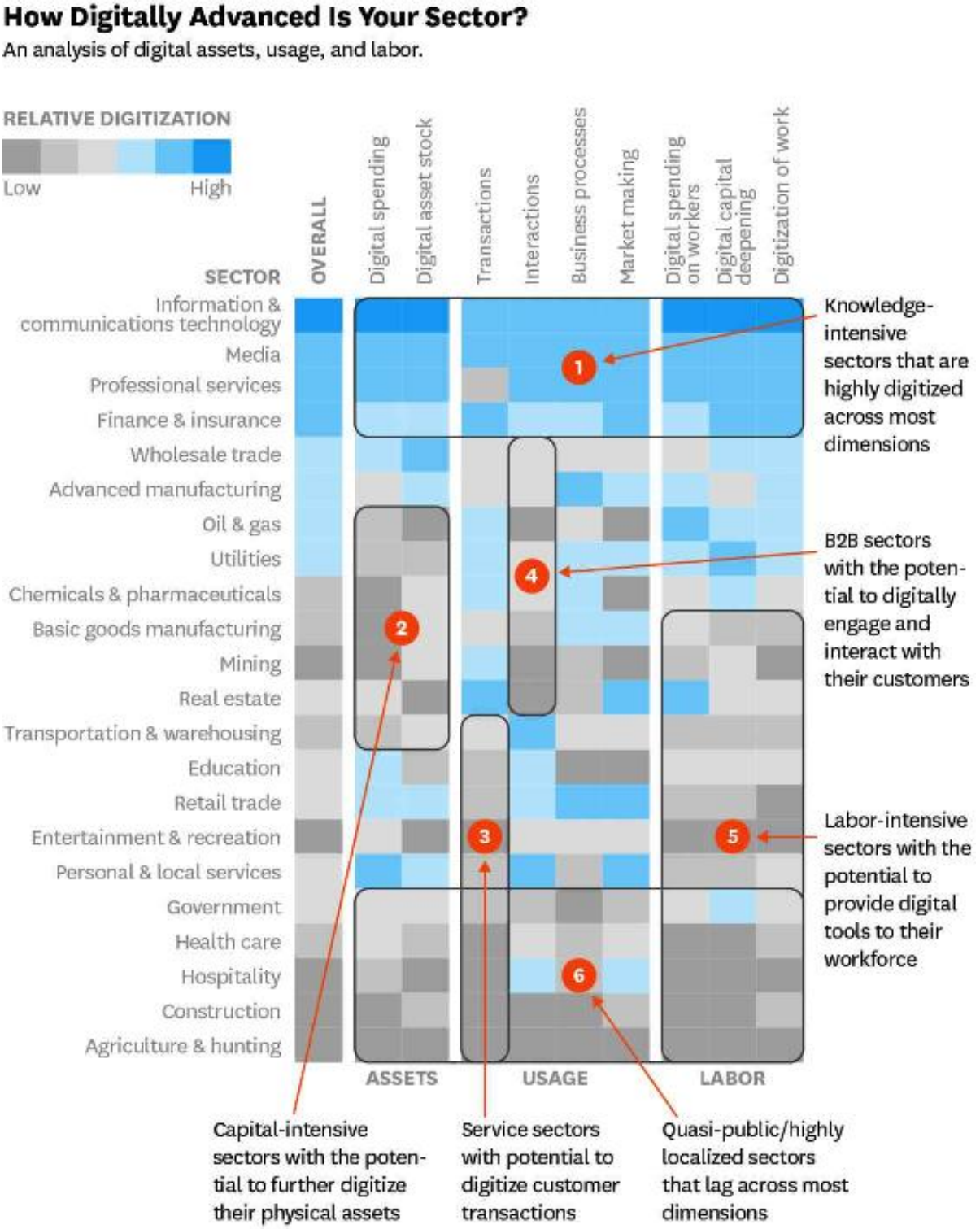


Fig 40*a Business Innovation is sometimes as easy as finding a white space. Source: McKinsey Global Institute*[35]

Finding gaps in the market

Another use of 2D design space is to find overlooked or underserved segments. We can apply this method to matrixes too. Examples are, a company-employee skills matrix, technology roadmaps, and the innovation matrix. In Fig 40, McKinsey took a sectorial view. However, other viewpoints can yield discoveries too. For example: instead of viewing by sector we could view by software vendor (Salesforce, JIRA, Autodesk…).

The Periodic table

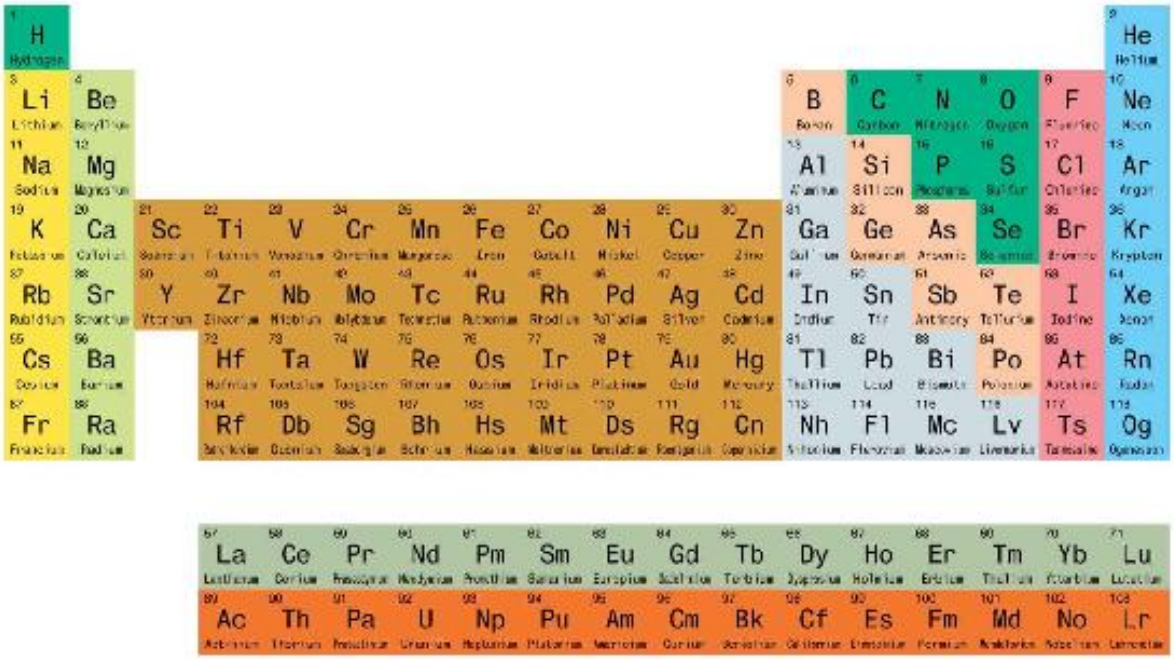


Fig 40*b The most useful visualization in the history of Science? Source: Bloomberg BusinessWeek.*

A famous application of the Gap Matrix is Mendeleev's 1869 Periodic Table of Elements. In 1869, 150 years ago, Dmitri Mendeleev published a periodic table of the chemical elements in based on properties that appeared with some regularity as he laid out the elements from lightest to heaviest. When Mendeleev proposed his periodic table, he noted gaps. At the time, only 63 of the 118 known elements identified today were known. Then he predicted the properties of five undiscovered elements -- a genius coup for any young scientist[36]. Mendeleev visual predictions spurred a discovery race. Shockingly, he never received the Nobel prize – a testament to scientific rivalries, but also of how visual thinking has been historically despised by an academia dominated by word oriented thinking.

The Innovation Matrix

The Innovation Matrix is a knowledge discovery tool – Leo Tschirsky

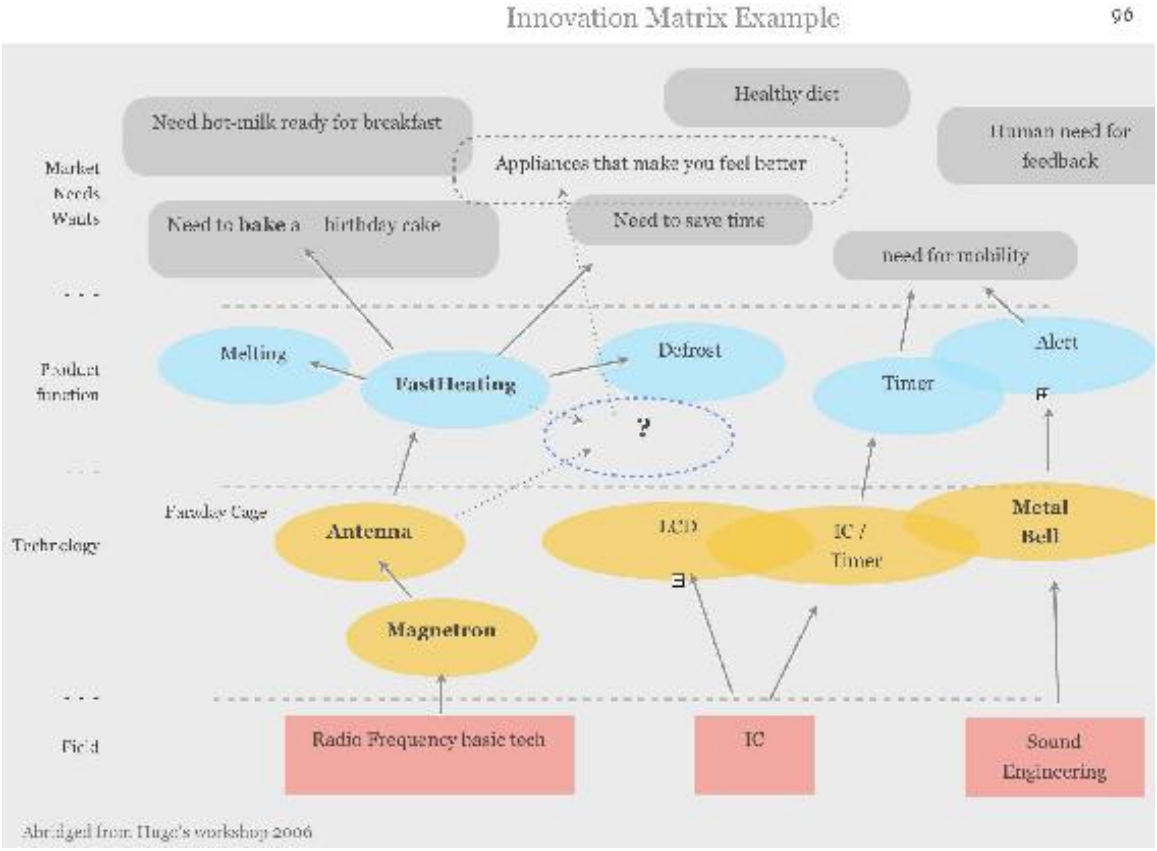


Fig 41 *The Innovation Matrix (IM)*[37].

Leo Tschirsky, Professor Emeritus of Business Management at the Swiss Federal Institute of Technology (ETH), facilitated this microwave oven workshop in Tokyo Institute of Technology around 2006. The **IM** helps to formalize and organize functional relations between:

- Market needs
- Product functions
- Technology
- Science fields

How to use

1. **Draw a matrix format by rows**

- Write the market needs (Why people buy ovens)
- Product functions (heat, boil...)
- Technologies behind those functions (Magnetron, LCD...)
- Basic science field supporting those technologies. (RF, IC)

2. Link concepts with arrows

- Clarify
- Seek deep truths
- Use 5-Why root cause analysis, jobs-to-be-done theory

Once your product is clear and mapped out...

3. Innovate

Now that you have a clear picture of relationships between value, customer needs, costs and technology. You are in a better position to innovate using a variety of techniques such as:

- Brainstorming
- Planning an ideo style shopping cart workshop
- Using Edward deBono creativity tools
- Finding gaps
- Serving new needs with exiting functions

Exercise

Groups of four. Time 20 minutes. Think about this microwave and its components…

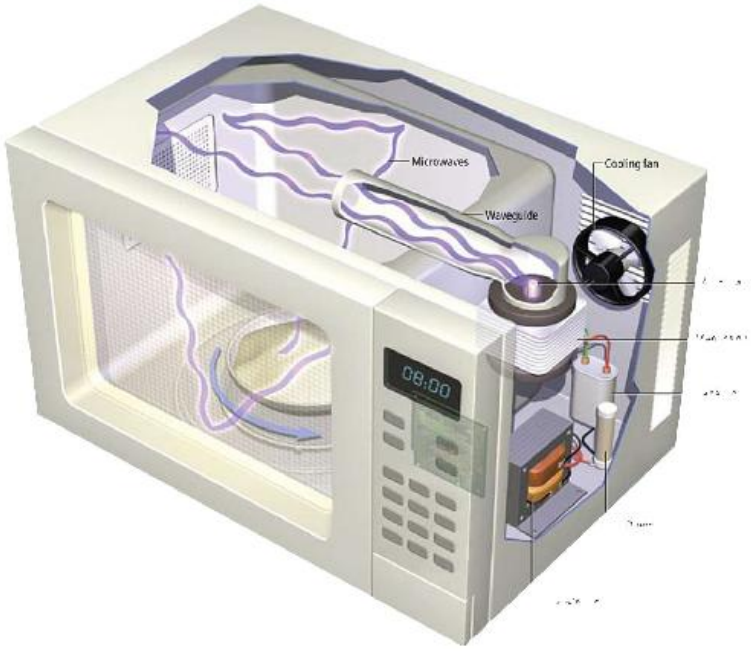


Fig 42 *The cyclotron is the element that generates the microwaves.*

You have been hired by a microwave oven brand. Recently, due to Chinese competition, the oven margins are paper thin so the survival of the company might very well depend on you coming up with a new oven design for which customers want to pay more. **Required**. Use the IM to innovate the microwave.

Solution

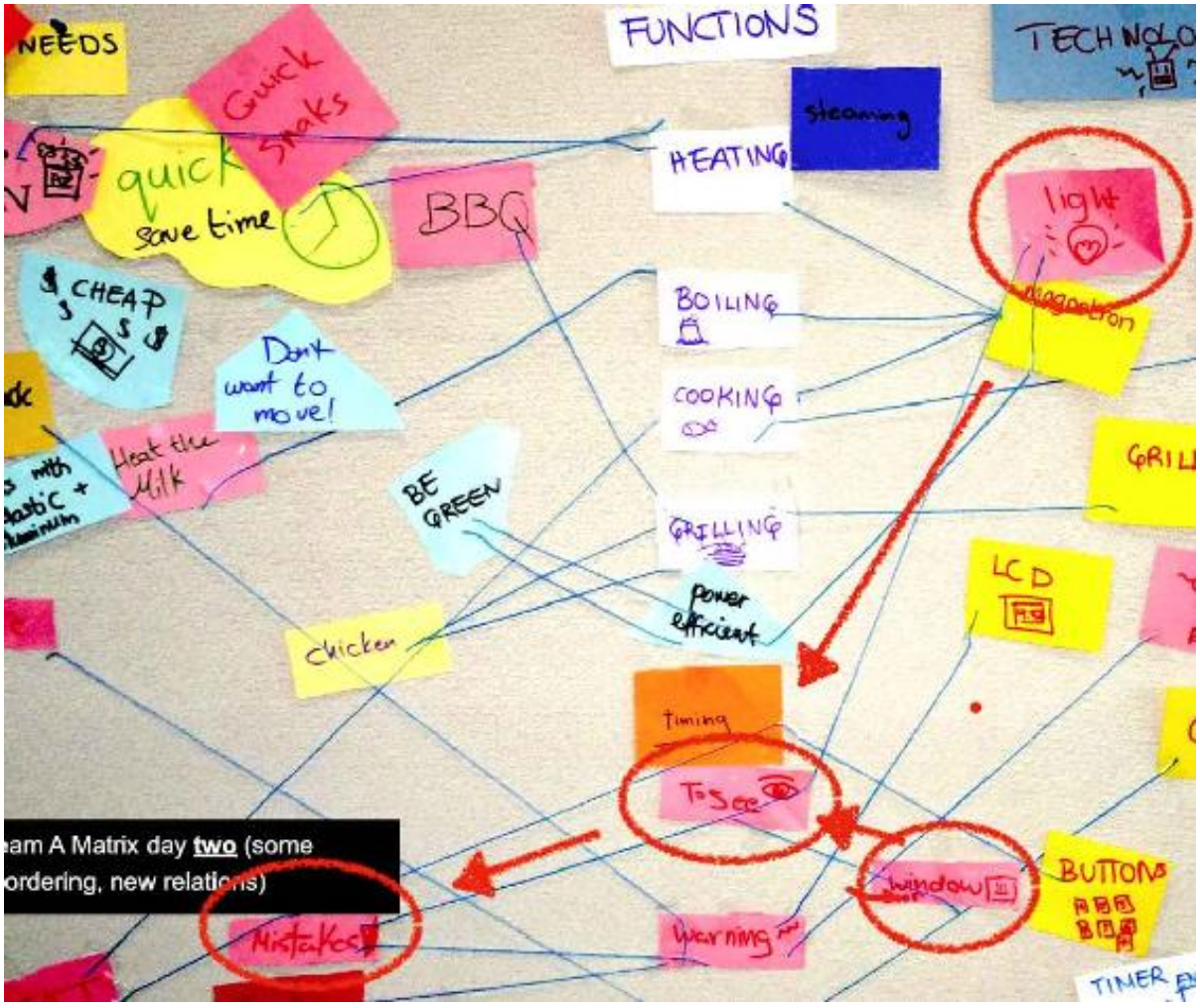


Fig 43-44 *Functional Innovation example, can we replace light and window by a cheaper component with same function?*

Light and oven window seem two unrelated technologies in an oven. However, they serve a common need: The need to check for mistakes! ***Cost of (Window +Light) > Cost of (X)?*** Consider the cost of a window + light. Their sole purpose is now clear. Can X do their function better, cheaper? Let's find X!

Wardley maps

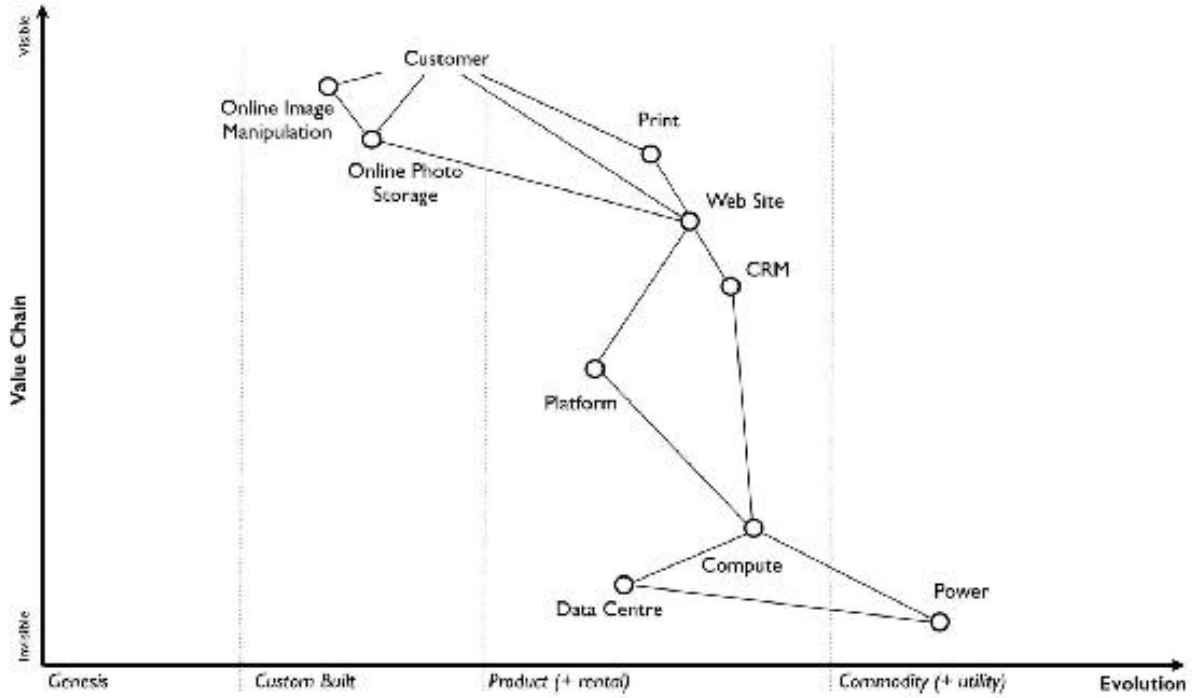


Fig 45 *A Wardley map, Simon Wardley* *CC BY-SA 4.0* .

A Wardley map can be thought of as a 2D mapping of commoditization vs value chain, or as an evolution of Leo Tschirsky's Innovation Matrix. In any case, when there are dilemmas about whether to outsource or not, this is the go to tool in 2019. So far we have seen a few visualization tools used in business and policy making. However, visualization is also used in coaching. Let's see an example called the wheel of life.

Wheel of Life

"A chart to think about your life goals"

Amy is a working, single mom who just resigned from a Fortune 500 job. A successful career woman in her own right, Amy just left her ascending executive role because she wanted to travel less and spend more quality time with her 14-year-old teenage daughter. I am sitting with Amy. We are meeting to talk about life, career and charts, yes charts! Amy is about to show me a chart I have never seen before. She calls it the ***Wheel of Life***. This chart, she tells me, was instrumental in helping her realize that she needed to change her life.

How to use

Fig 46 *Start with a dot in the middle of a paper.*

First, you draw spokes. Eight spokes of a wheel. Each spoke represents a different category of your life and will help you measure your satisfaction in each area of your life. The first one is Money (How satisfied are you with the money you have saved/make?) Second, Career (How satisfied are you with your path, progress and current career?). Third, Wellness (both spiritual/mental and physical). Then, Friends & Family, Love, Fun, Physical Environment (Do you like the country, city/ house/ neighborhood you are in?), and finally, spiritual and personal growth. We put a grade on each

category marking a dot on the spoke on a scale of 1 to 10, 1 being at the center and 10 being away from the center, and then we connect the dots.

Connecting the dots

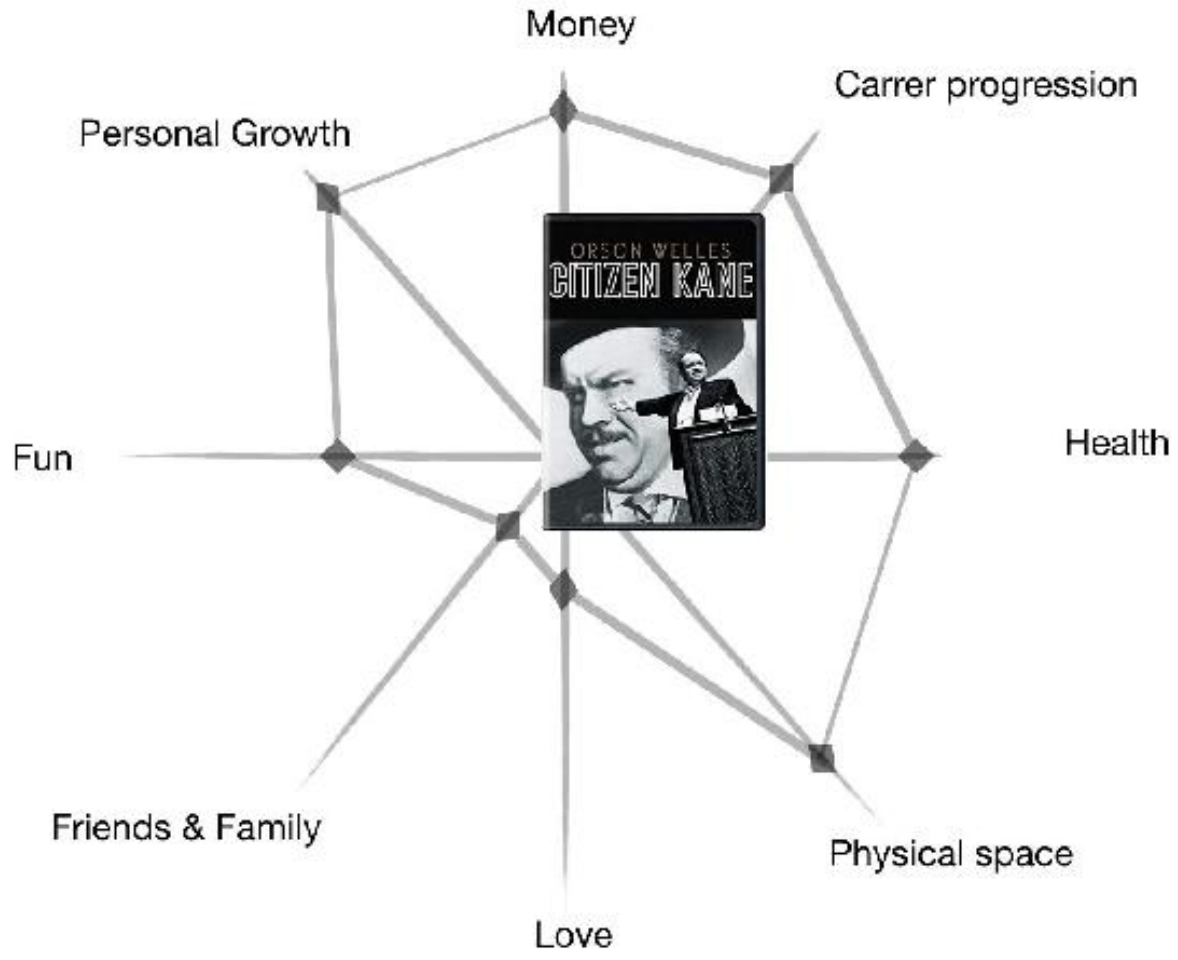


Fig 47 *Citizen Kane-type wheel of Life.*

As I connect the dots of my chart, a chill goes up my spine. My wheel of life is not round! It looks like an asteroid and has a big dip where the spoke of love is. Is that bad? “Oh, well” I think, mortified, I guess I need to reprioritize my life.

An x-ray of your life

Then Amy leans over and says, “This is the snapshot of your life. I learnt this exercise at a drop-in class at Stanford. It turned out to be one of the most useful things I learnt there because it helped me visualize my **blind spots**. Many of my classmates did this chart too, and they were just as shocked as you are now...” Then she adds, “The typical comment of my classmate was: Here I am, focusing on my career and personal development, as I have done all my life, while equally or more important parts of my life are being neglected.” The wheel of life together with other visualization tools such as *the good time journal* are part of the **design your life** trend where design principles of iteration, visualization and design thinking in general are applied to coaching. The results are spectacular. To learn more, the book by Atari co-founder *Designing your life* is a great start.

Exercise

Make your Wheel of Life. How would you make this chart more compelling than Fig 47? Time 7 minutes. (Solution in the next page)

Solution

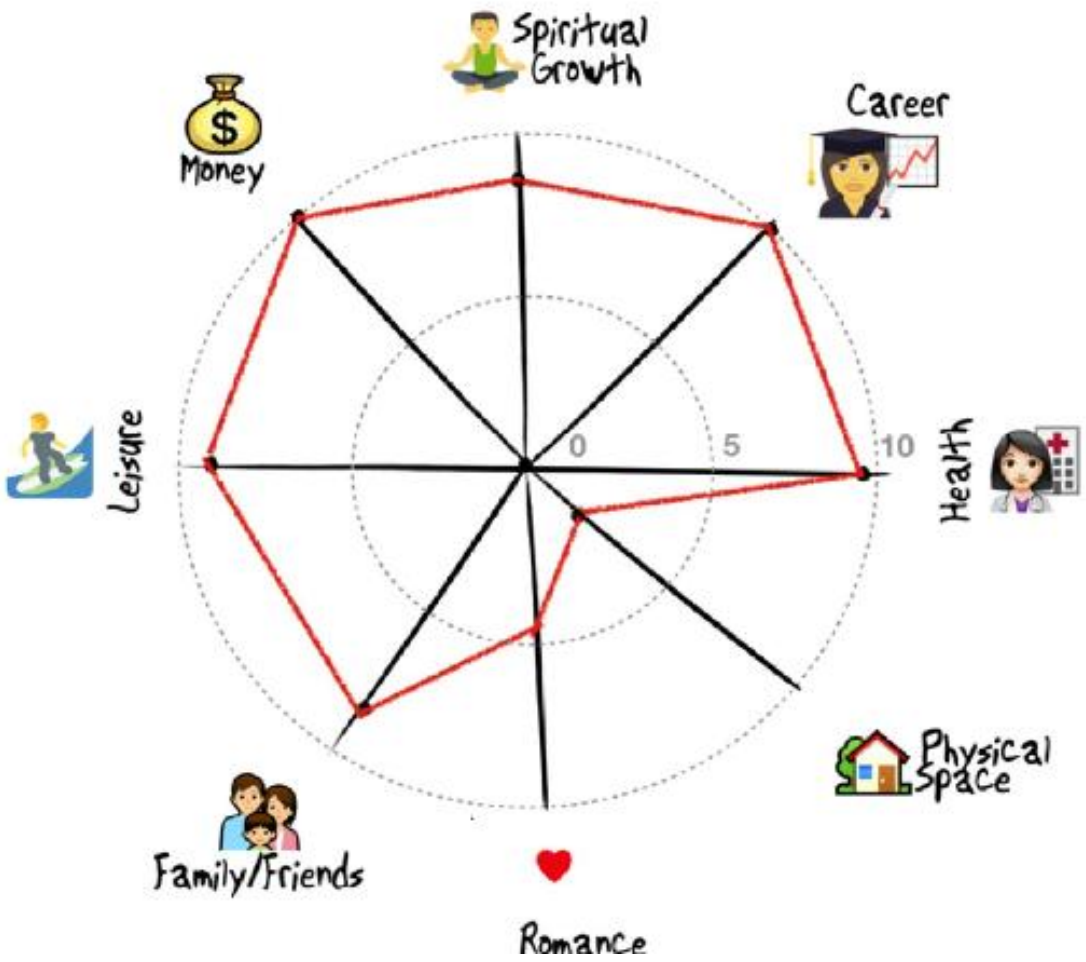


Fig 48 *Icons, use to reduce the time to understand a chart.*

Icons and emoji are an underused resource in chart making. On the other hand, emoji use is correlated with employee engagement

Interactive exploration

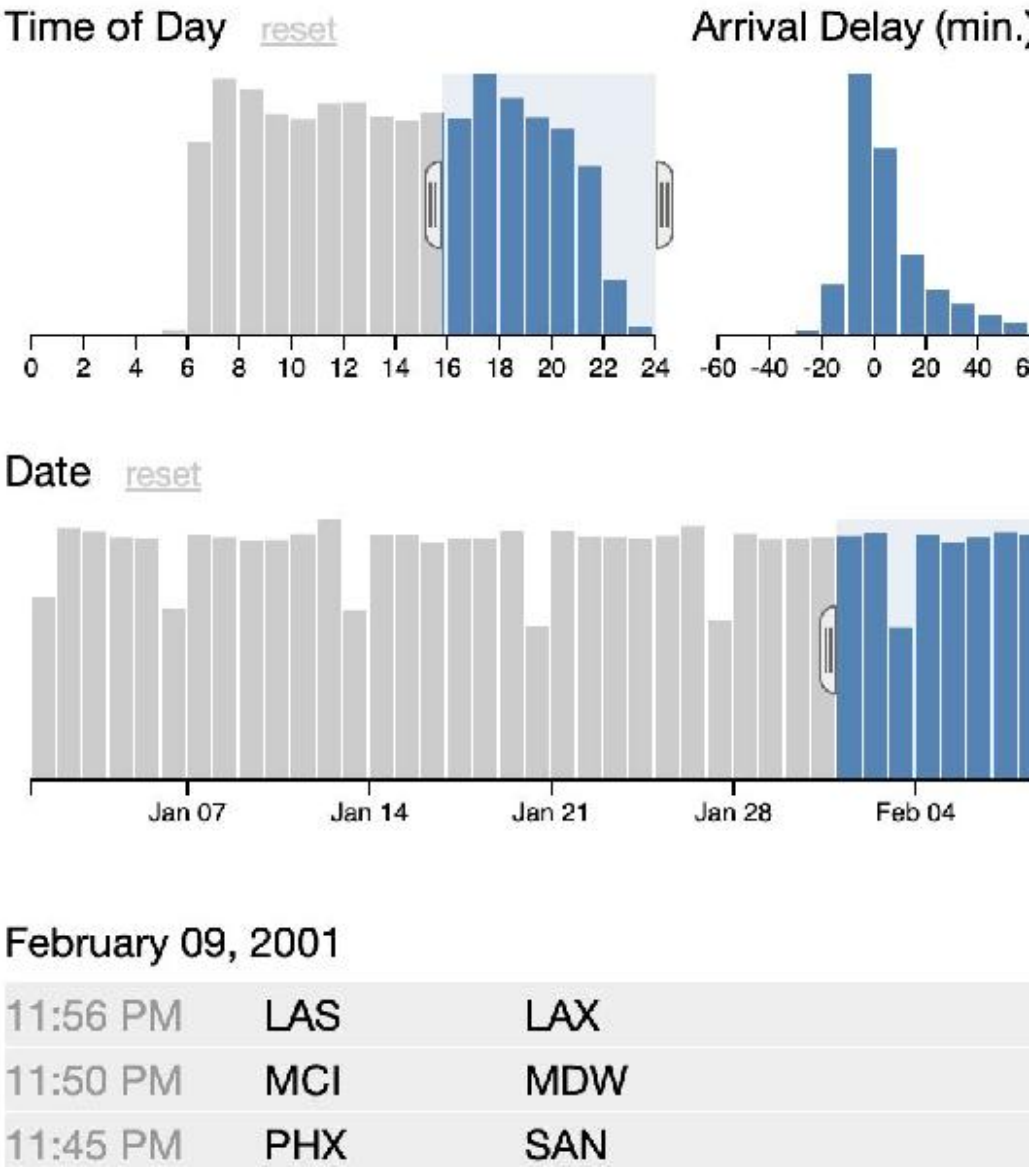


Fig 49 *A screenshot of a real time, visual SQL inner join operation between three tables; Source: Square 2001.*

Square's Crossfilter

Crossfilter is a JavaScript library for exploring large multivariate datasets in the browser. Extremely fast (<30ms), it allows "Doherty

threshold” interaction with coordinated views, even with datasets containing a million or more records; Square built it in 2001 to power analytics for Square Register.

Exercise

Groups of two. Time 12 minutes. Using Square Crossfilter demo, find three insights about air travel. Example: to avoid delays fly in the morning. Visualize your findings.

Quiz: Supporting decisions

True or False? Time 10 minutes.

1. Working with per capita measures is the only rational way to compare countries. [True / False]
2. The principle of mean reversion states that, in the long-run, a ‘hidden hand’ pushes outliers towards the mean. [T/F]
3. The gap matrix is used primarily to discover niche or gaps in the market. [T/F]
4. Wardley maps have rendered Innovation Matrix obsolete. [T/F]
5. The Crossfilter library is an appropriate choice to visualize linear relationships between two variables such as delays and departure time. [T/F]

Solution

1. Working with per capita measures is the only rational way to compare countries. [True / False]. True.
2. The principle of mean reversion states that, in the long-run, a ‘hidden hand’ pushes outliers towards the mean [T/F]. True. There are no long lasting competitive advantages [98].
3. The gap matrix is used primarily to discover niche or gaps in the market [T/F]. True. However, it can also be used in other areas.
4. Wardley maps have rendered Innovation Matrix obsolete [T/F]. False. Innovation Matrix is still better than Wardley to do product innovation and to map a product to its market.
5. The Crossfilter library is an appropriate choice to visualize linear relationships between two variables such as delays and departure time [T/F]. False. A scatter plot is clearer. Crossfilter is great for EDA.

Chapter 5. Making your chart pop

– Tips to make it viral –

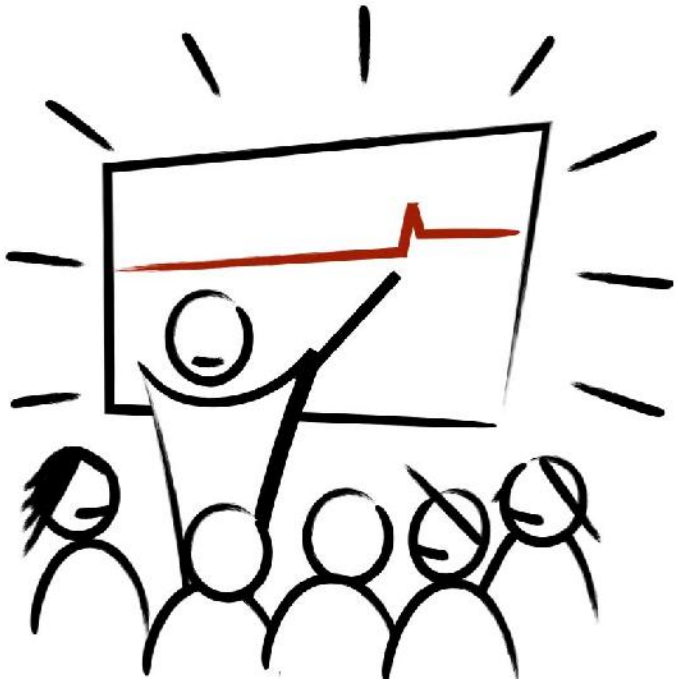
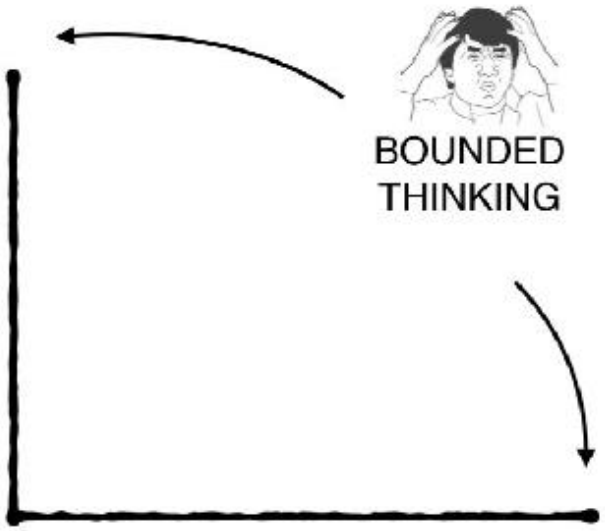


Fig 50 *Jazz hands, use to make a visual compelling.*

In this chapter, we will see hacking strategies to make your chart memorable. Now that you know the principles of meaningful chart making (by reducing information overload, linking to frameworks, fitting narratives and leveraging visual metaphors), let's look at simple tips on how to jazz it up, or in design thinking jargon, make it "pop". Pop means it pops into the eyes. My favorite resources on this are: Dan Roam's books, Information is Beautiful by McCandless, and the Instagram account "chartr". Be careful! Some charts are so beautiful on their own that they are like a piece of art in their own right. In any case, remember, the chart should never be the main character of your story, what matters is the narrative (message) and how well it is communicated to your audience. Unfortunately, aesthetically pleasing visuals and a visual that gets the job done do not always coincide. Incidentally, this is one of the reasons that in an Advertising agency you

will see the art director and account manager always at odds, the art director wants to win ad awards to go to Cannes Lions Creativity Festival, the manager wants effective advertisement. Let’s see of the main techniques use to make a chart “pop”.

Use arrows to unbound your thinking



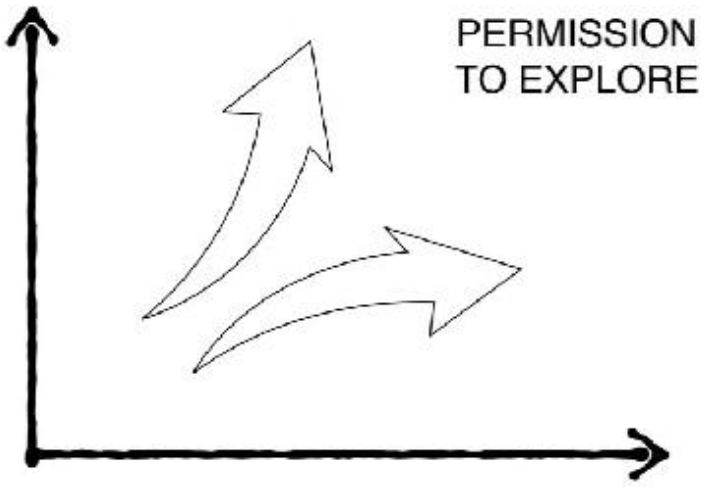


Fig 51 *The arrows give you permission to think beyond.*

Decluttering charts

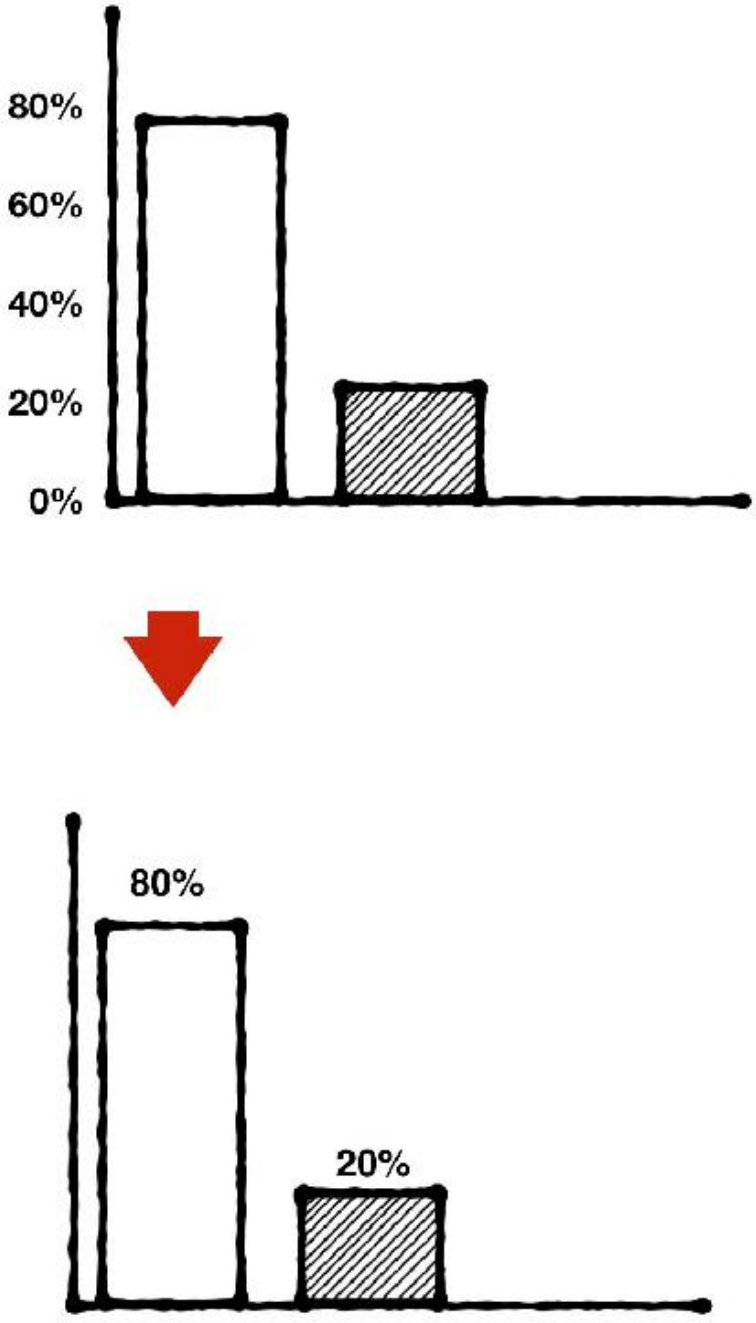


Fig 52 *Marie Kondo applies the principle of throwing away things in one's life that do not spark joy.*

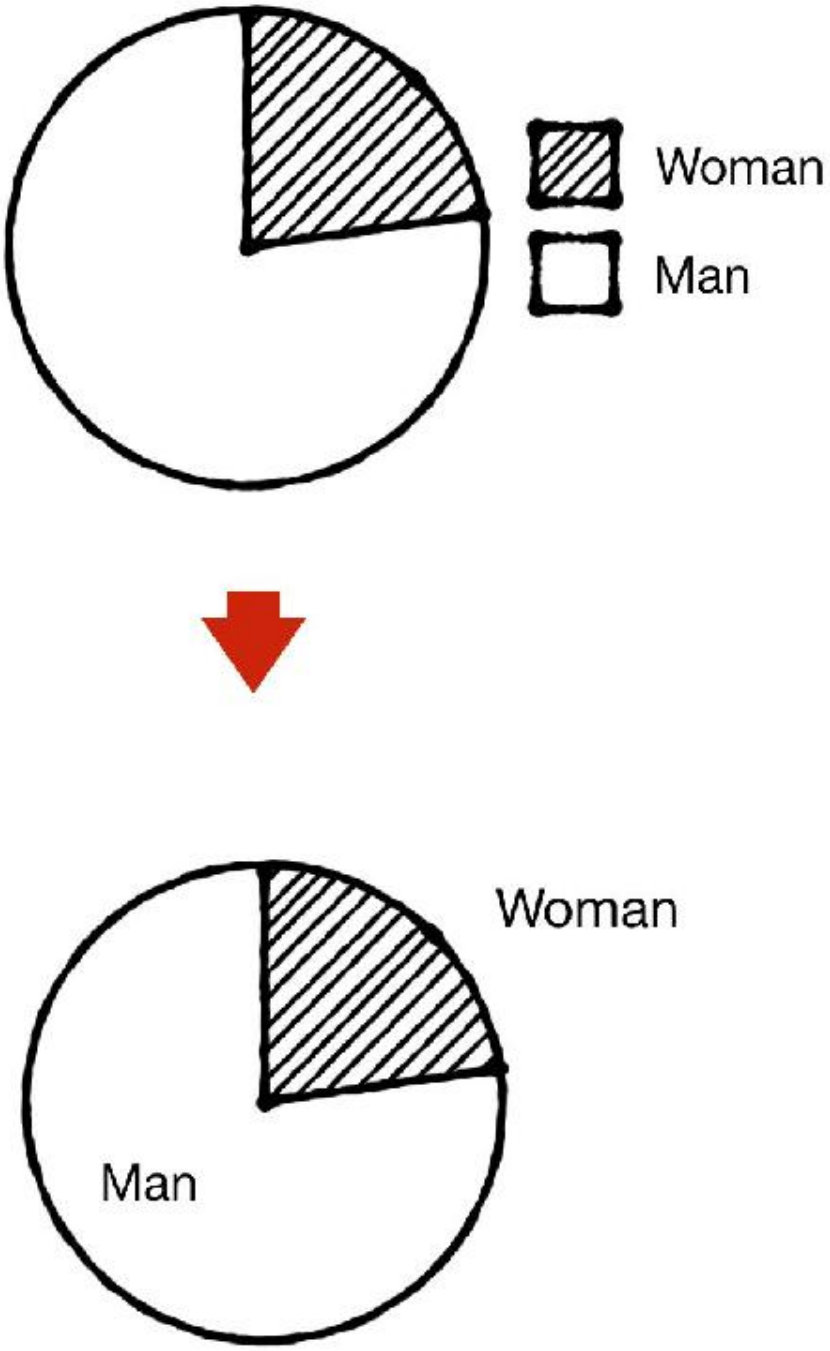


Fig 53 *Animation on decluttering a pie chart bit.ly/2OgCLUO*

Use personas to win over the audience

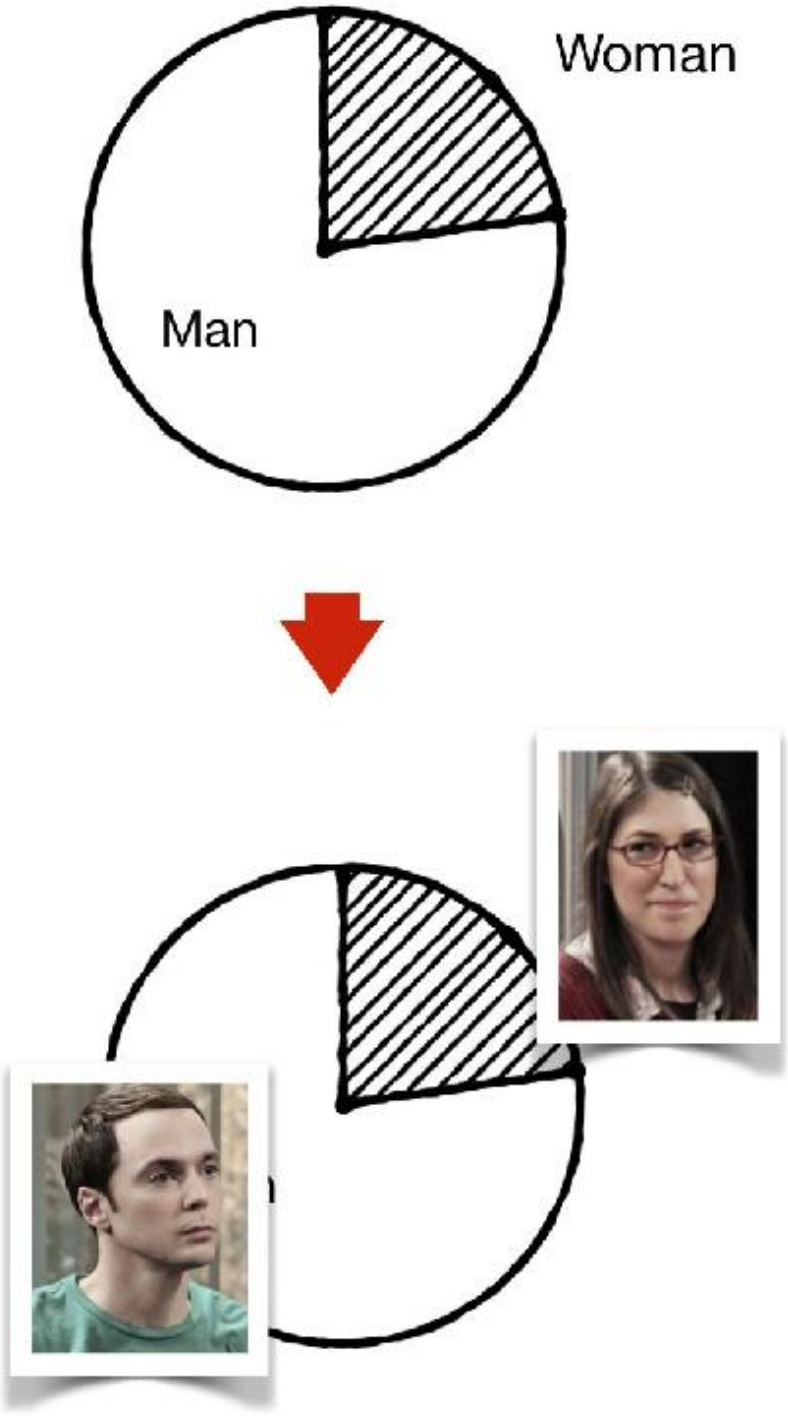


Fig 54 *Humans are adept at recognizing faces.*

Did you know that we can recognize a face faster than many other objects in the world? Use it! In 2007 Honda used this principle when they

designed a motorbike that, from the back, looked like a human face (anthropomorphic).

Visualizing big differences

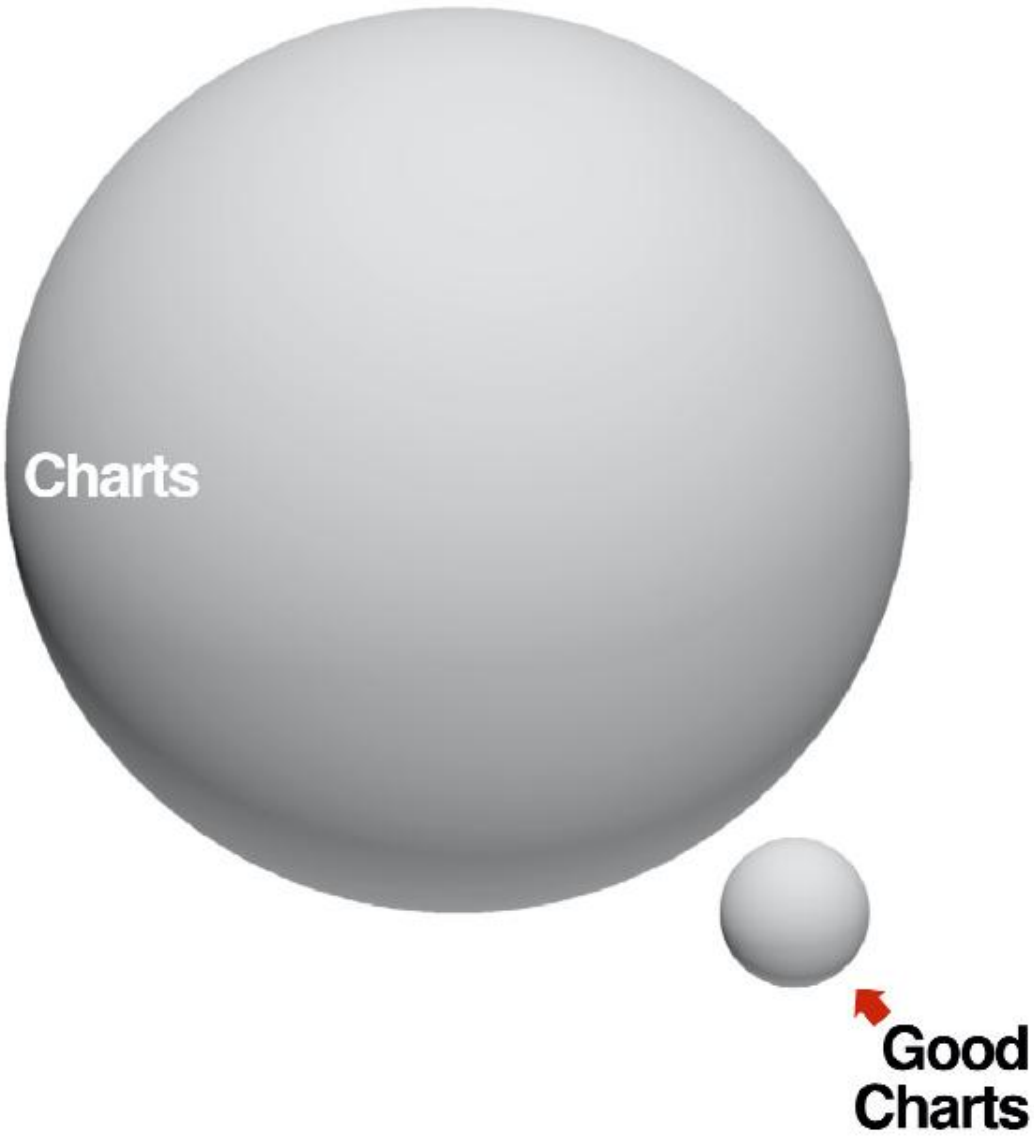


Fig 55 *This chart has an astonishing range of four orders of magnitude.*

How to calculate the dynamic range in a planet chart. In a planet chart, the dynamic range is n^3, where n is the times small planet radius fits in big one. In Fig 56, n=12. Therefore, $n^3 = 12^3 = 1720$. Therefore, the share of the small planet is approximately 1 in (1720+1), 0.05%.

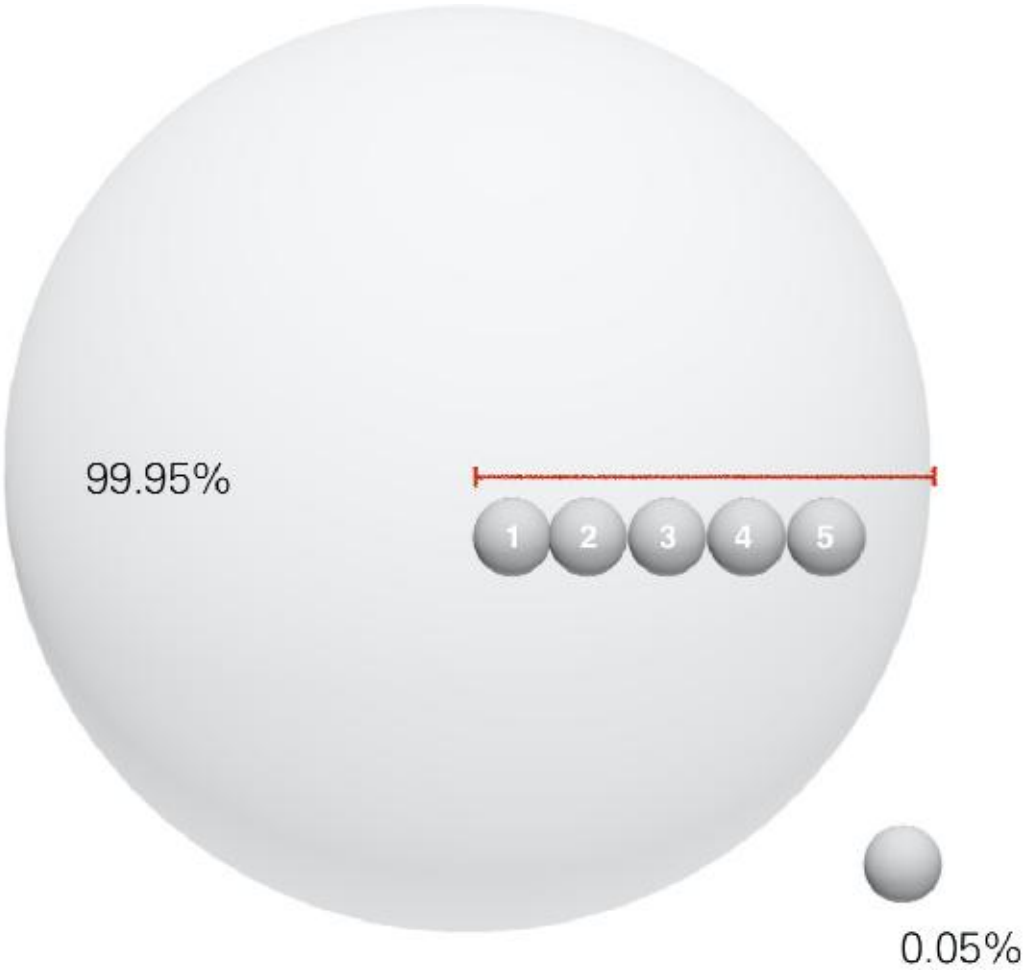


Fig 56 *The small planet fits 1720 times in the volume of the big one.*

Other successful attempts to visualize enormous differences are the famous post “What does a PhD mean to you”, and the 1977 film Powers of Ten; both used the 2D zoom technique to illustrate gargantuan size contrasts. However, of all chart metaphors, the planet metaphor is excellent to visualize vast differences. See the next page to find out why.

Putting the world's money into perspective

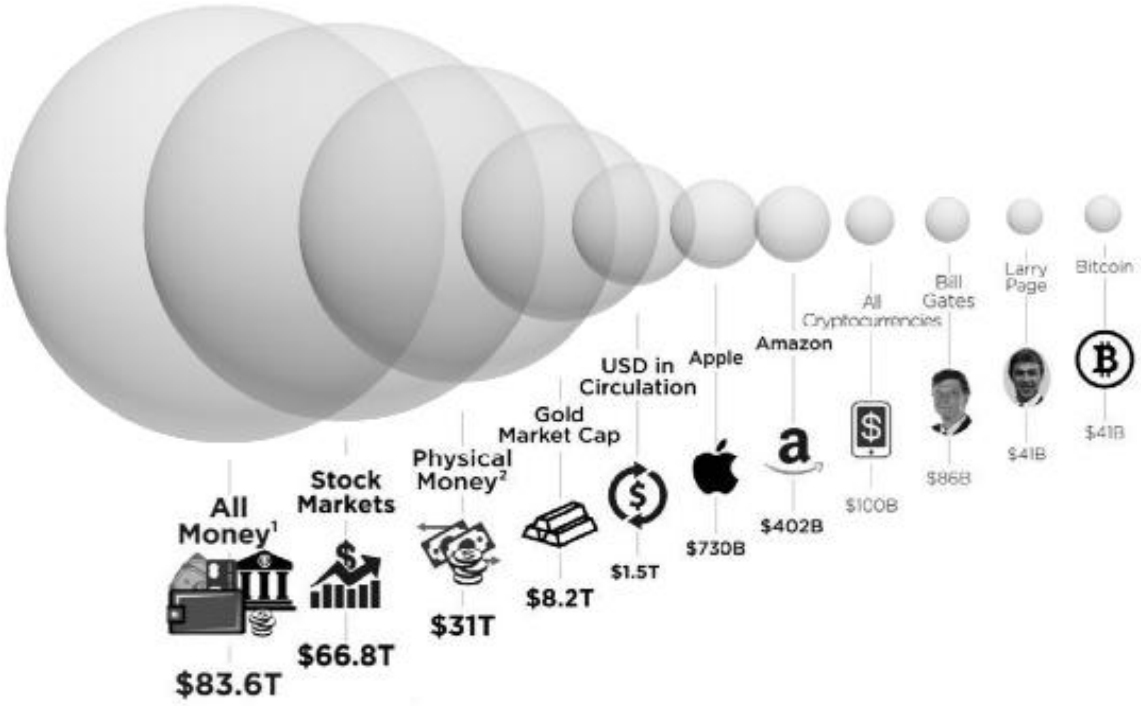


Fig 57 *Humans are evolutionary adapted to estimate mass from height.*

The dynamic rage of this chart is an astonishing **four orders** of magnitude! It can visualize 41bn and something 20,000 times larger. This chart was combined with the Fear Of Missing Out (FOMO) narrative before the Bitcoin crash of January 2018. Original source: HowMuch.net.

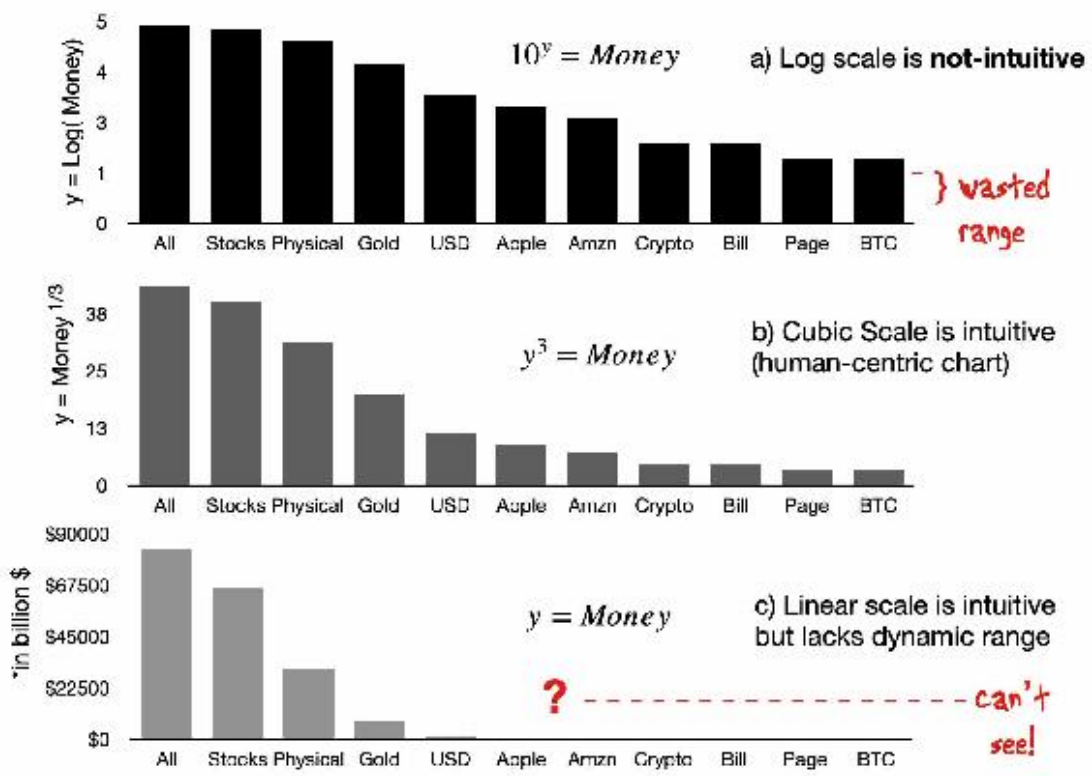


Fig 58 *Comparison between a) Log of value, b) Cubic root of value (Radius of equivalent weight sphere), and c) Linear scale.*

In Fig 58, note how (a) has the most dynamic range but is not-intuitive; (b) is the radius of the sphere equivalent. It would be intuitive if expressed as volume, like in Fig 57; (c) is a linear scale; is intuitive, but lacks dynamic range.

Why 3D spheres work so well

We humans have evolved to estimate weight of an animal from sight. Of course, this was a very useful skill for our ancestors in the Savannah. Notice how much easier it is to understand relative sizes when we use volume, versus any other option. For primates, estimating the weight of a fellow primate visually was a crucial survival skill useful to determine how dangerous the threat of an opponent was before contact. Given most living forms have a similar weight density, a way to do this was by estimating volume. At the same time, we humans are struggle to understand bar charts

when the bars differ more than 2 orders of magnitude. Luckily, if shown 2D projections of 3D objects, most humans can estimate the weight well. This comes in handy to compare magnitudes as different as 3 or 4 orders of magnitude on a flat surface such as this book. Using the cubic relation, a 1 to 10 change in height becomes a 1 to 1,000 change in weight — a great dynamic range.

Log charts

Note that the log plot solves the issue of dynamic range but we humans are not born with *logarithmic* intuition built-in (Fig 58). In other words, a kid will understand the balls, but it takes hours for undergrads to become familiar with semi-log plots.

How many Solar Panels are needed to power the USA?

Fig 59 *In 2017, Elon Musk used such a chart template to advocate for Solar Energy. — It was a flop.*

Fig 59 narrative is scarcity. In 2017, Elon Musk used a chart like this one. He was advocating for Solar Energy. He said, “We just need one pixel of the map covered in panels to power the whole USA, remember just one pixel.” It was a flop. Why? Because it connected to a win-lose narrative. It is also hard to trust what we cannot see (one pixel is not a great visualization). Unfortunately, 2D charts do not have enough dynamic range to visualize differences larger than 2 orders of magnitude. He was trying to visualize 4 orders.

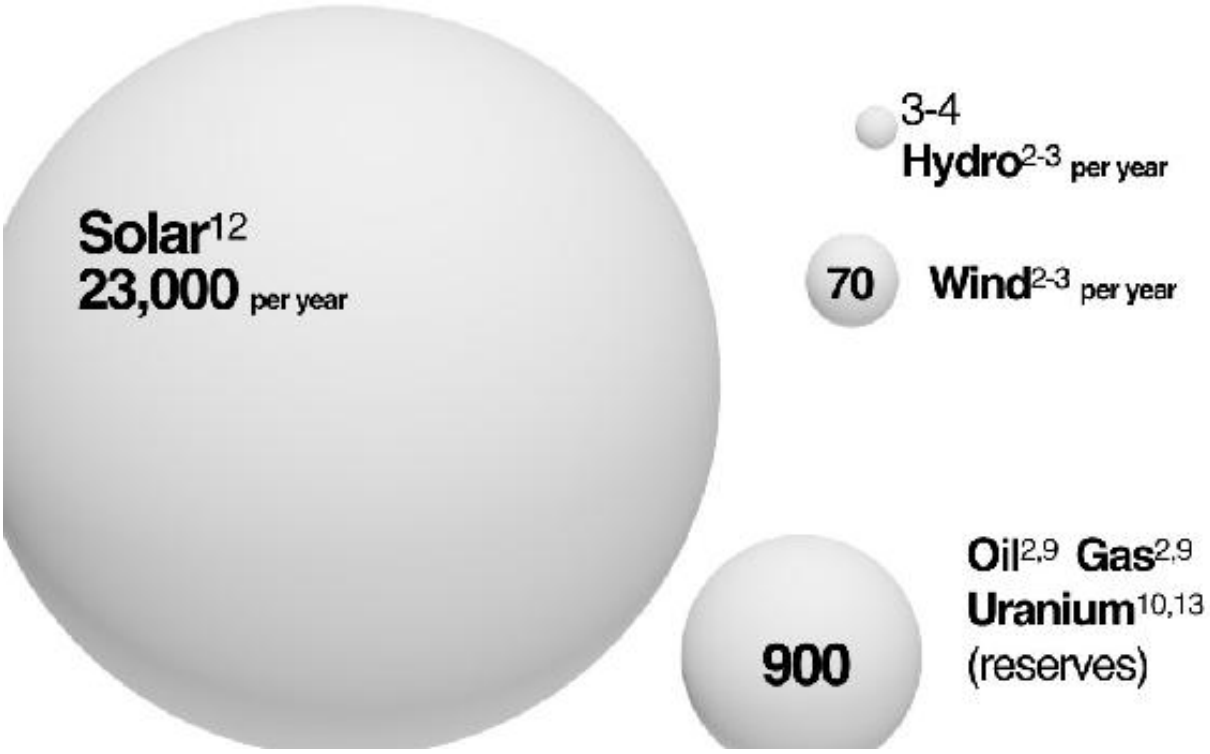


Fig 60 *A chart that uses the growth mindset narrative. Adapted from Q-Cells.*

Fig 60 narrative is the ***growth mindset.*** It visualizes more than 5 orders effortlessly. Circa the year 2005, German solar maker Q-cells used a similar chart in their PR. This chart is more effective and trusted than Fig 59 because it connects us to the growth mindset narrative by visualizing the astonishing abundance of **renewable** energy available.

Storytelling age-bias with humor

Fig 61 is based on an arrest dataset provided by Minneapolis Police and other sources. It was published by Kaggle in 2018. It shows the correlation between arrests with no charges and the age difference between policeman and subject.

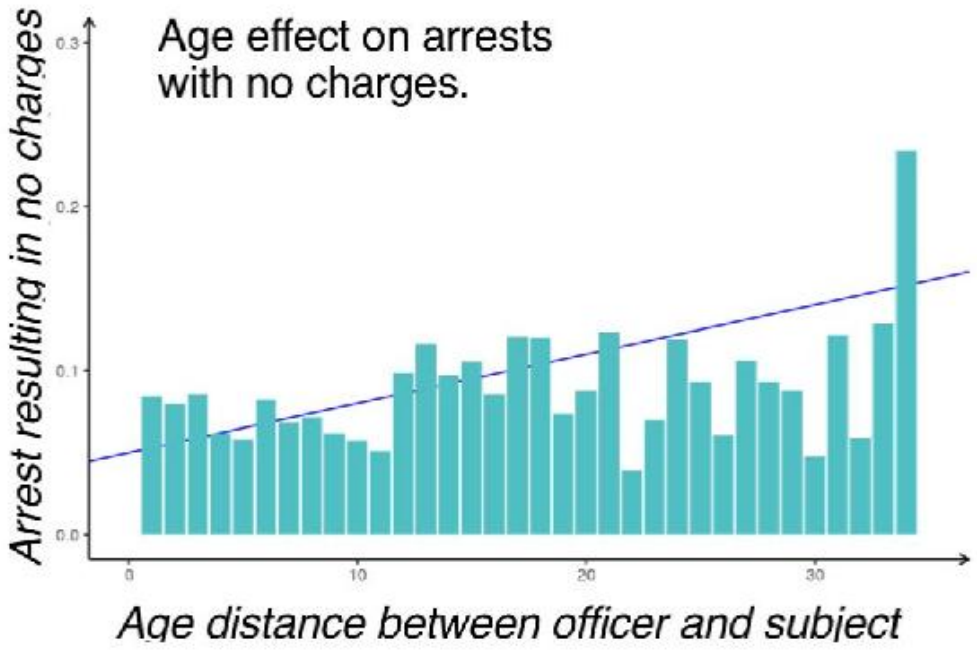


Fig 61 *Of all the biases, age is one of the most pervasive and less talked about.*

Full kernel: http://bit.ly/2JXWP9d. Regression shown in blue. If one considers not being charged as a positive outcome, then one would prefer interacting with an older cop, but if one considers "a no charge arrest" as a negative outcome, then one would prefer interacting with a cop closer to one's age. The lower the age gap, the lower the chance of not being charged.

Exercise

Make this chart more compelling. Time 3 minutes.

Solution

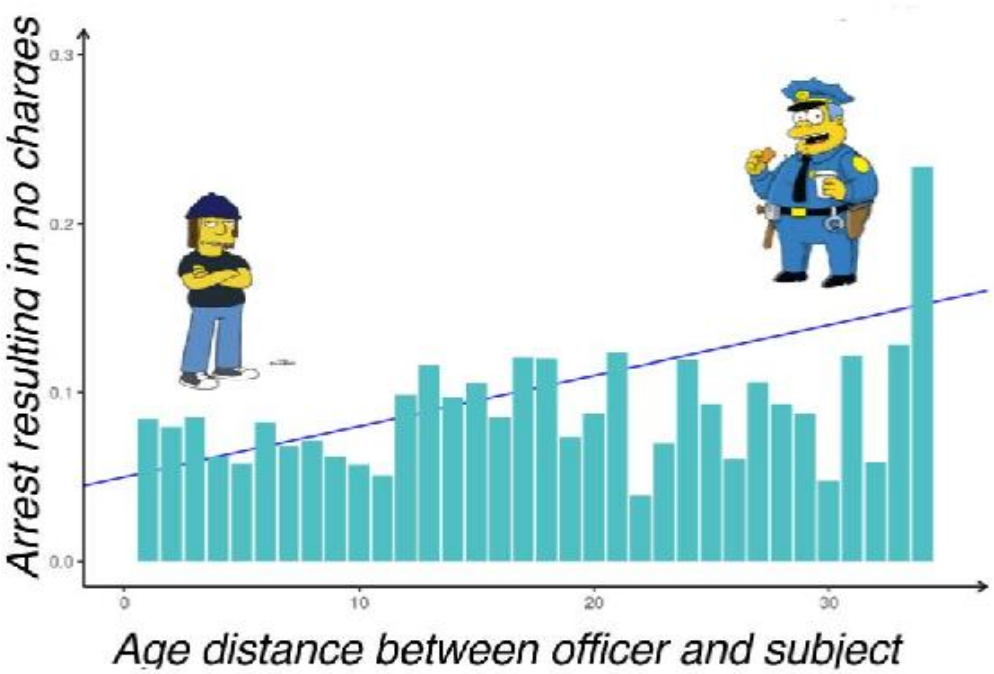


Fig 62 *Use personas to make a chart memorable.*

Use the Golden ratio everywhere

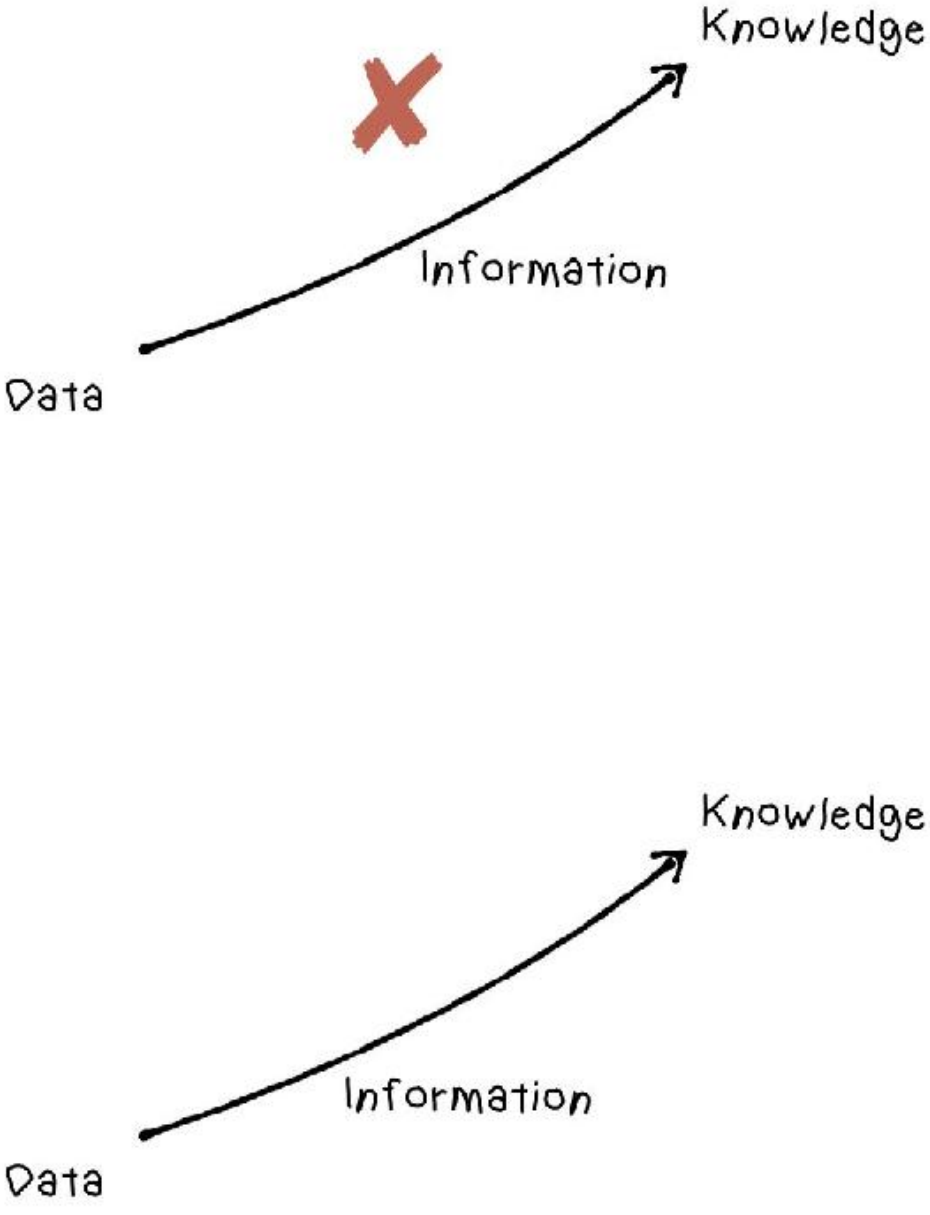


Fig 63 Which one feels more harmonious?

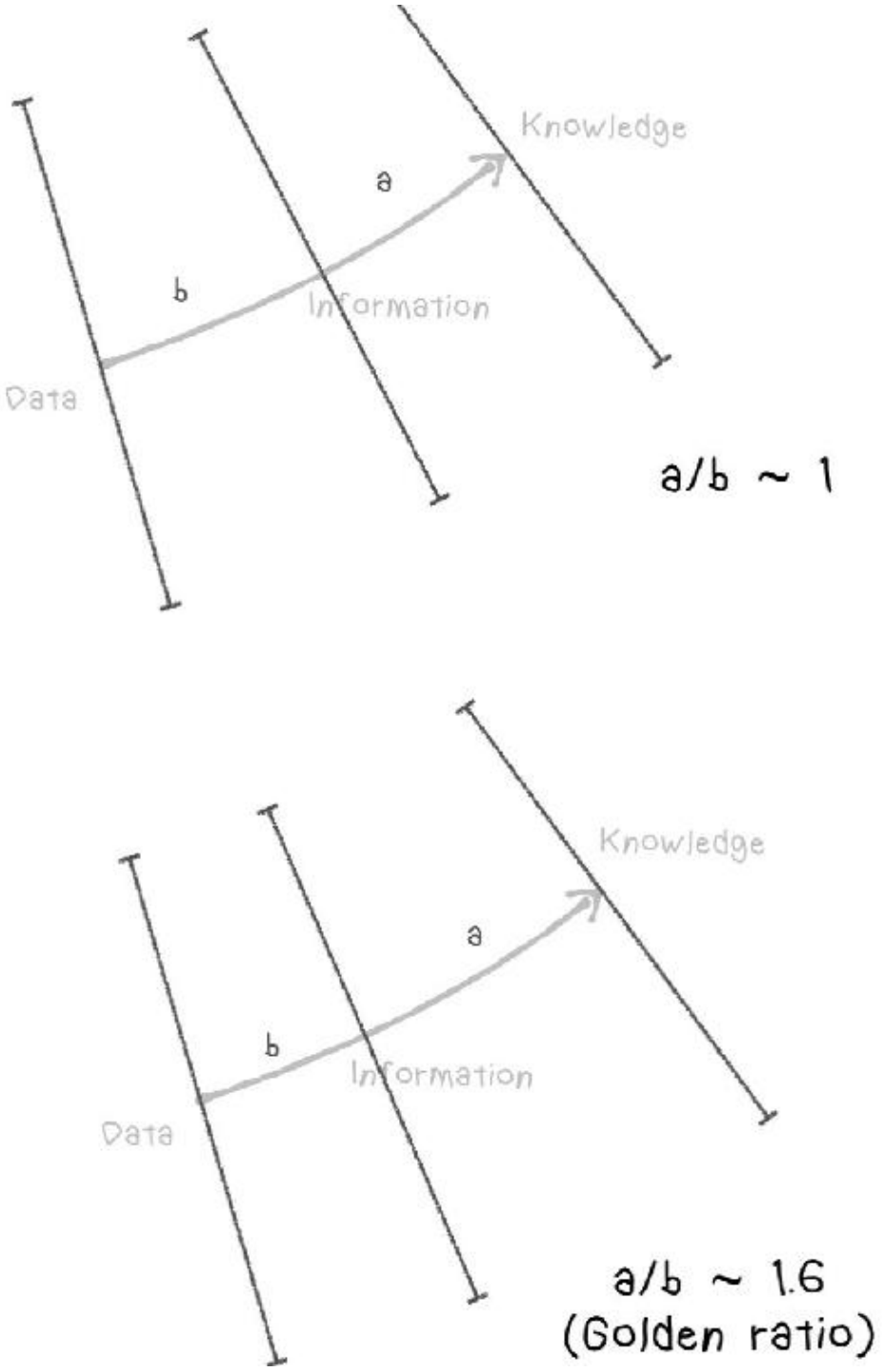


Fig 64 *Never miss a chance to use the golden ratio.*

Did you know that if you show two different cards to a human, she will choose the one whose sides ratio is the closest to the 1.6.

Twists

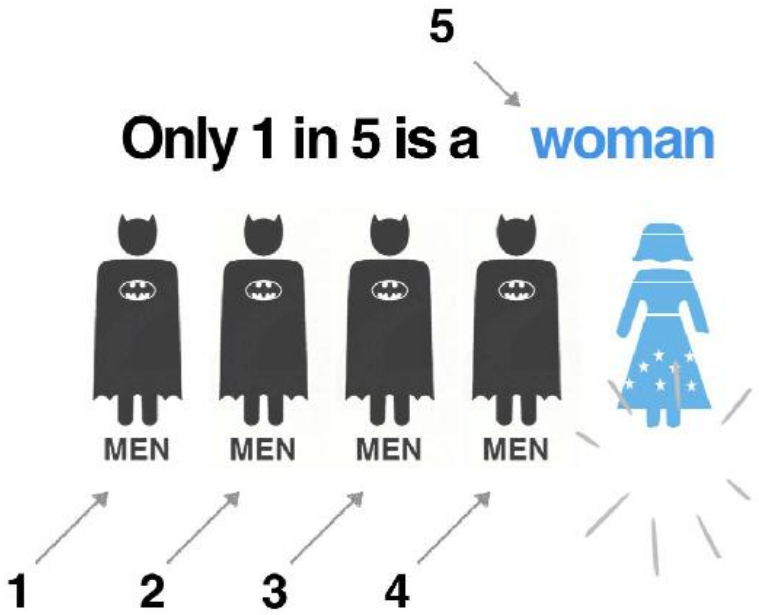


Fig 65 *Use twists to make your chart stick.*

In *improvisation*[38] theatre, we always say that a good story should have a twist. An effective twist it what makes a TikTok video viral, or in business cases, the so called Aha! moment. We can use negative space to create twists in charts. In this case, the reader expects to find the label #5 in a certain place but it is not there! This sends the brain into search mode[39]. When a few milliseconds later the brain finds the missing label, some endorphins are released. A bonus of this twist is that it fail-proofs (Poka-Yoke[40]) the chart — changing the label location makes sure that the reader will not miss the title.

Now that we know the main hacks to make a chart pop, let’s see a few examples.

Quiz: Ice baby

Evaluate the following chart. Then prose a better way to visualize it. **Required**: 300 words essay + two mockup sketches of proposed improved visualization. Time 25 minutes.

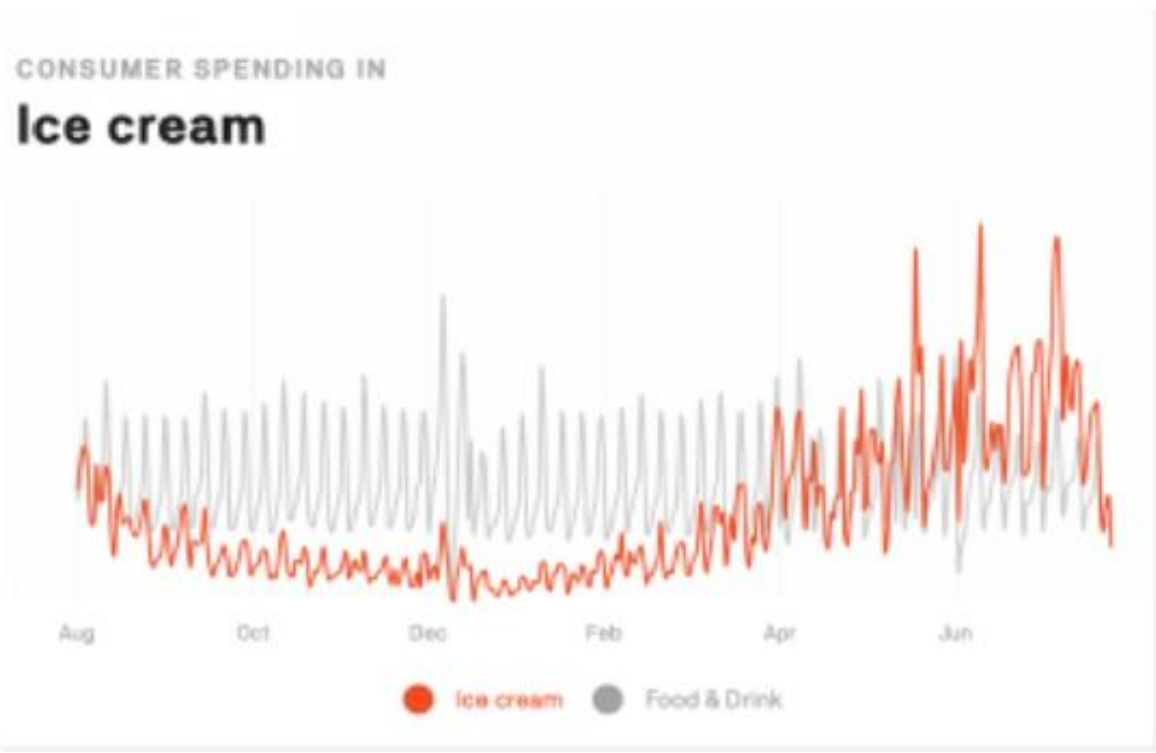


Fig 66 *Sales of ice cream in Sweden, Iceland and Finland. Source: Meniga/Wrapp.*

(Solution in the next page)

Solution

Main observations

- This chart contains useful information such as: weekend sales are higher than Wednesday sales, and Summer sales are higher than Winter. However, the chart time line is not optimal to showcase this.
- To give this chart more utility, connect it to other knowledge. For example, comparing to sales in Spain, (where the Winter: Summer ratio of sales is lower, 4x vs 10x in Sweden).
- This chart lacks narrative. An example of a narrative could be how culture influences consumer habits. Example: “It’s funny actually what happens in Sweden during summer months, we sort of lose our minds. We eat, drink and swim like we never saw the sun - that’s what happens when you live close to the North Pole and July comes.”. Thérèse Lundquist -- Head of Marketing & PR at Wrapp.

Color choices

Orange: cute at first glance, annoying after 20 seconds.

Time line

The data is weekly and yearly periodical, therefore a yearly or weekly periodical time scales would reduce clutter. For example, a weekly radar chart.

Alternative charts

- Radar weekly
- Bar plot monthly, weekly with iso-measures of ice cream
- The grey line is constant therefore carries no meaning (remove?)

Other suggestions

- Normalize per capita & per app user
- Annotate Christmas weekend and other peak days
- Explode into a scatter between temperature (x-axis), sales (y-axis), color of dot is day of week
- Calculate % of sales due to temperature and % sales due to seasonality from a linear regression analysis using as factors summer and weekends.
- Use Crossfilter to let users explore and discover hidden relationships for the variables weekday, month, temperature, sales of ice-cream and individual variance in consumption

Quiz: Global warming

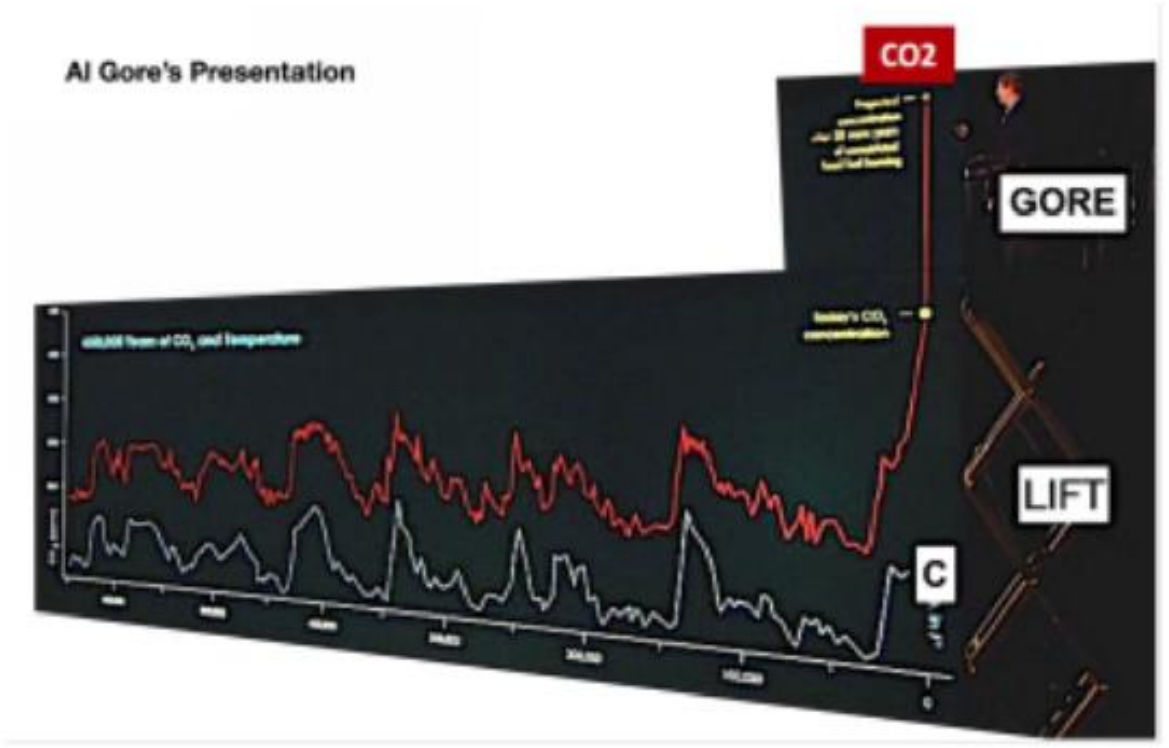


Fig 67 *Al Gore's CO2 Emissions Chart was not viral.*

In 2006 Al Gore's presented this CO2 Emissions Chart. X-axis span is +100 years; In Y-axis, red line is concentration of CO2; the blue line is temperature. Note the high correlation between both. To the right is Al Gore on top of a scissor lift. It was a flop. How would you help Al Gore visualize global warming? Time 25 minutes.

Solution

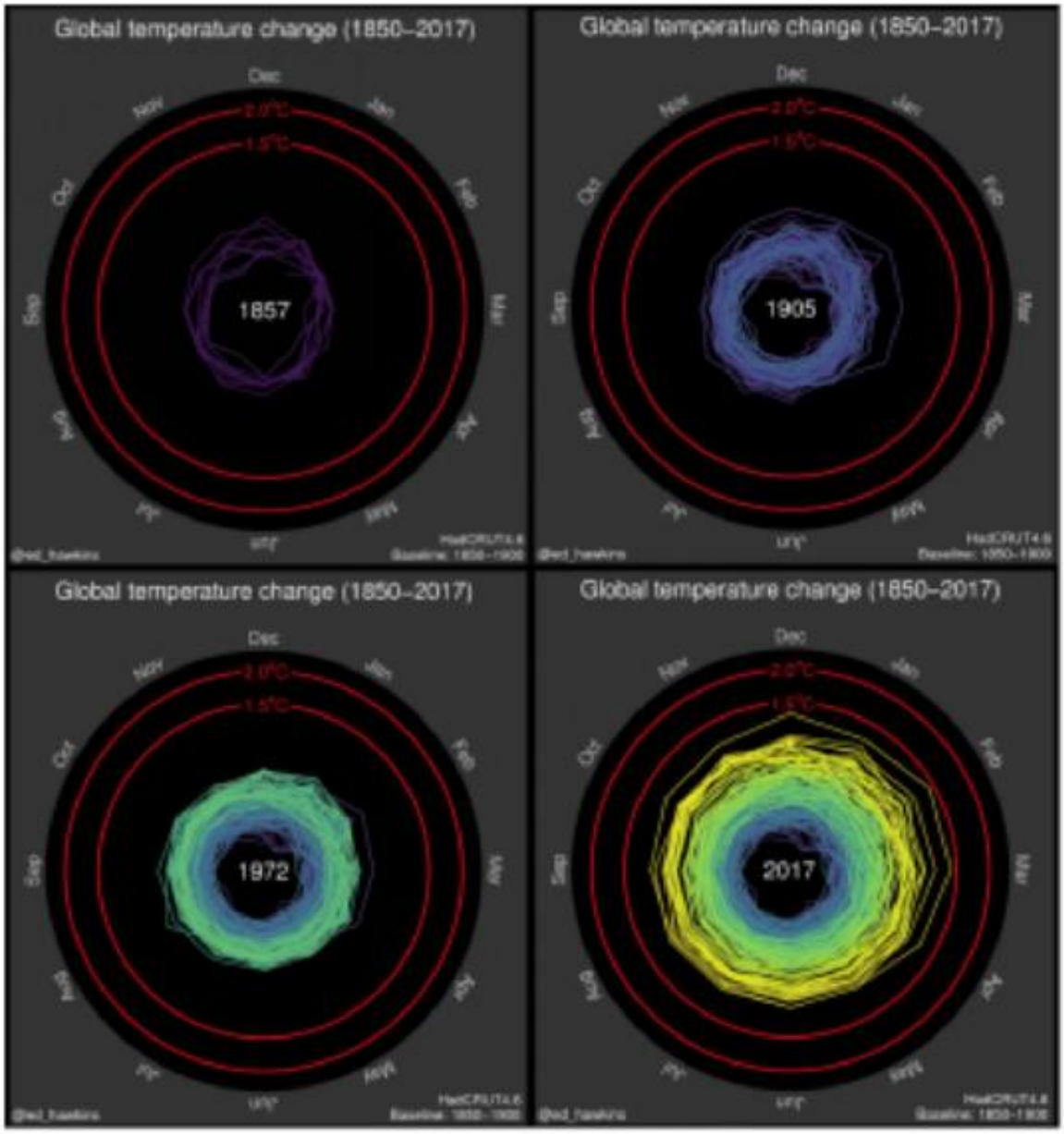


Fig 68 *Ed Hawkins made this* ***spiral chart*** *in May 2016. It went viral in minutes.*

From Pies vs. Bars we know that humans are **more** sensitive to circular than linear change. If we want the chart to align with the narrative that “climate change is an emergency”, then let’s leverage that!.

Quiz: Magic quadrants

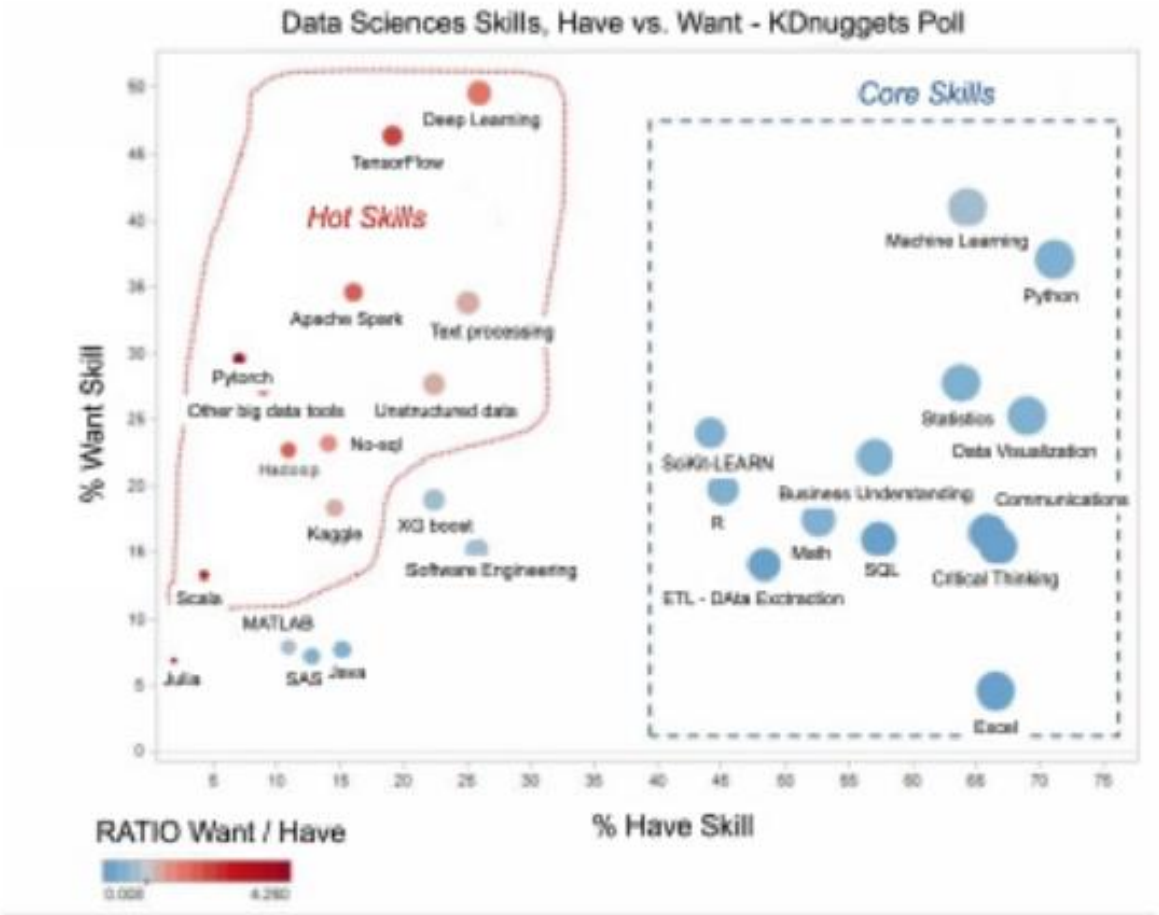


Fig 69 *Most wanted Data Science skills in 2019. Source [KDnuggets](). Image at original resolution (might appear blurry in some devices).*

In 2019, Gregory Piatetsky, editor of kdnuggets, published this chart. It used three visualization principles:

1. **Color** labelling to identify the cold/hot skills,
2. **Dot** size to indicate a magnitude, and
3. **Mapping** to a 2D space (but did not utilize its design potential)

The chart was based on a poll. The poll had just two questions:

1. Which skills/knowledge areas do you currently **have**? and
2. Which skills do you **want** to add or improve?

KDnuggets received 1,500 answers, and the aggregates by skill look as follows.

Table 1. Aggregates of the survey.

Skill	**Have it**	**Want it**	**ratio**
Python	71.2%	37.1%	0.52
Data Visualization	69.0%	**25.3%**	0.37
Critical Thinking	66.7%	15.5%	0.23
Excel	66.5%	4.6%	0.07
…			

How might we visualize this data into a more meaningful way? What design space is most appropriate given the data? Time 4 hours. Hint: See Magic Quadrants.

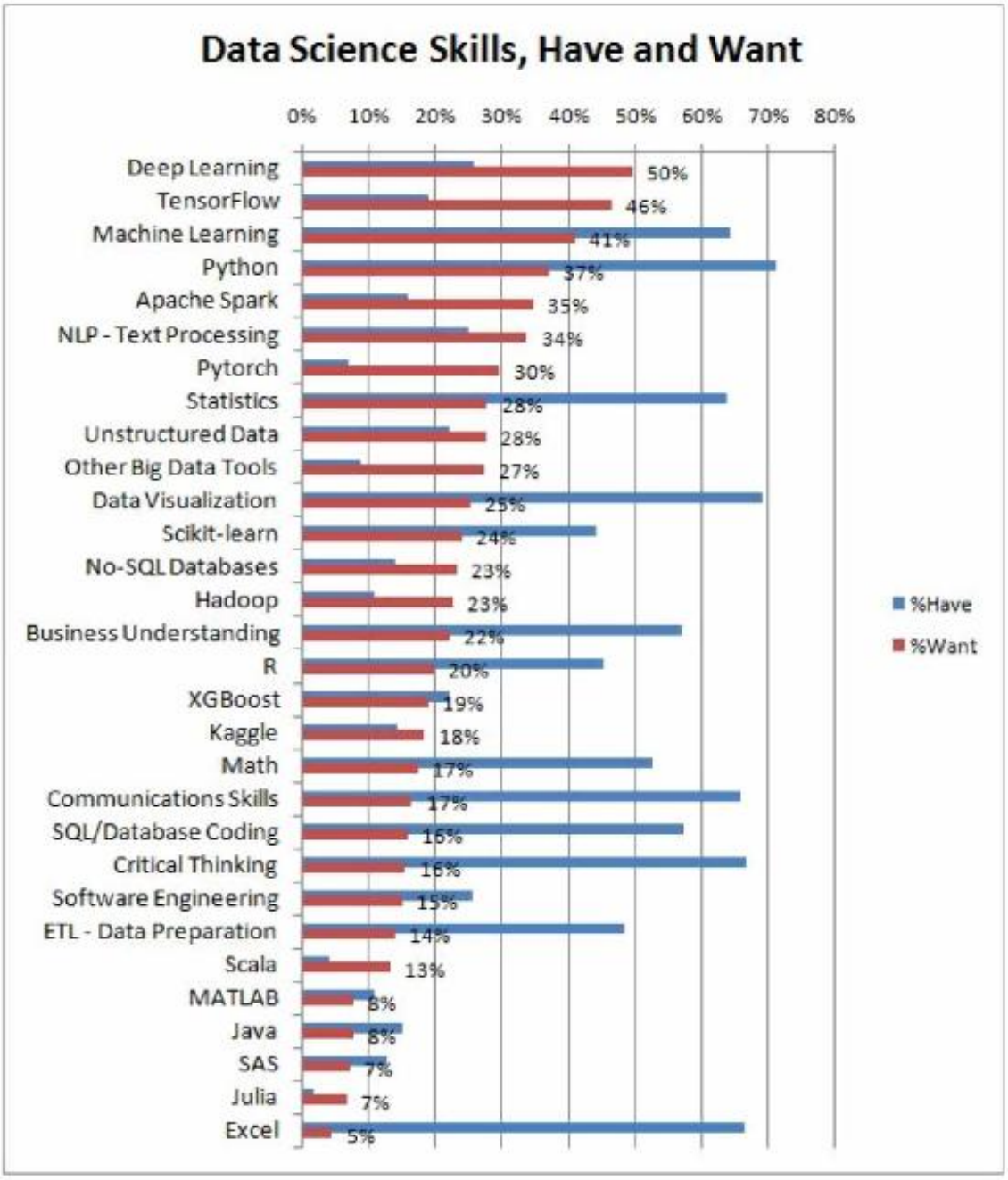


Fig 69b *Fig from Most wanted Data Science skills in 2019. Source [KDnuggets](). Image at original resolution (might appear blurry in some devices).*

Solution

Let’s apply what we have learned so far. (Find a why, transform data into information, synthesise knowledge by linking to frameworks, make it useful for decision making). Before finding a why, let’s first explore the data.

The first instinct is to do a scatter plot to identify interesting clusters. The x-axis can be the percentage of respondents that have a given skill, and the y-axis, percentage of respondents that would like to have that skill (want). However, there are too many data points for a human to make sense of it. It is a textbook case of death by information overload and in Fig 70 we used the Jackie Chan meme to convey it.

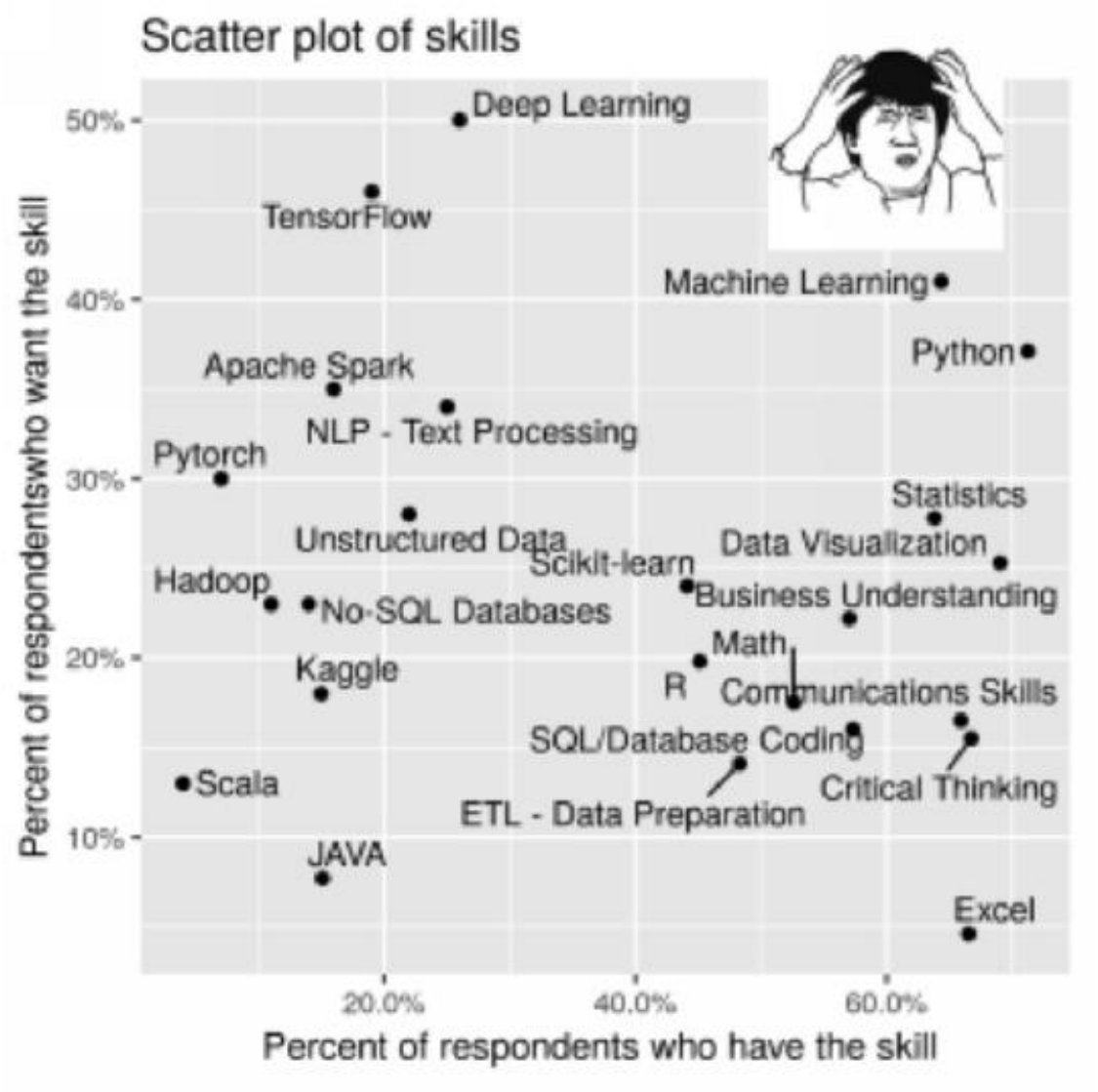


Fig 70 *A victim of information overload?*

How to transform this information into knowledge? Let's take a cue from **Gartner magic quadrant**, a framework used in IT business intelligence. It reduces complexity to human levels by using a quadrant hierarchy (Fig 71).

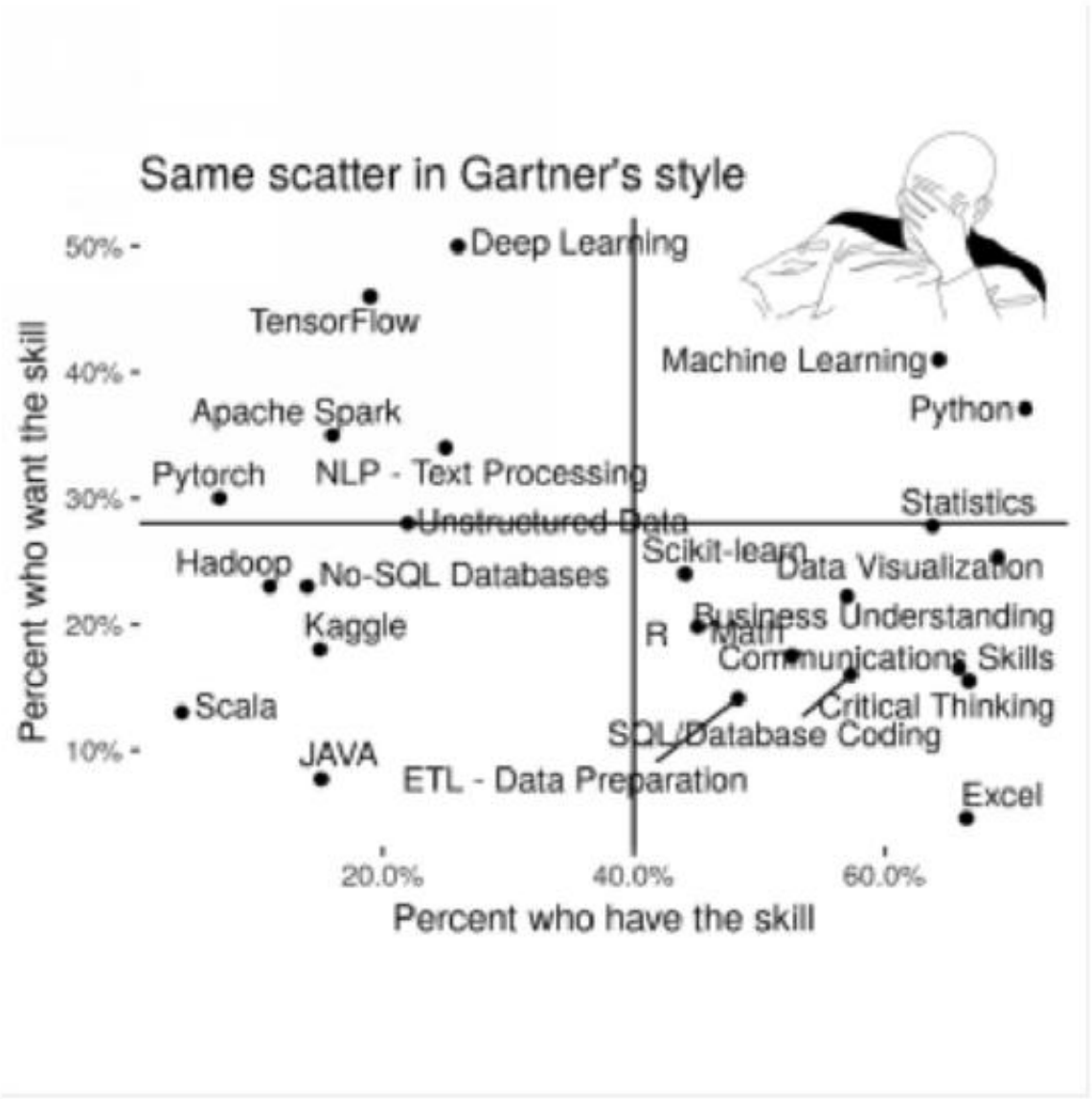


Fig 71 *Gartner uses the quadrants framework to cluster, reduce complexity and make meaning.*

However, Fig 71 is far from ready. The y-axis is aligned with the gravity metaphor (highly wanted, high y). However, the x-axis is not aligned with another unspoken rule, (this one by Guy Kawasaki): “you want (desired goals) to be high and to the right”. In this case, the most desired skill (Deep Learning) is on the wrong side – we need to flip the x-axis, (Fig 72).

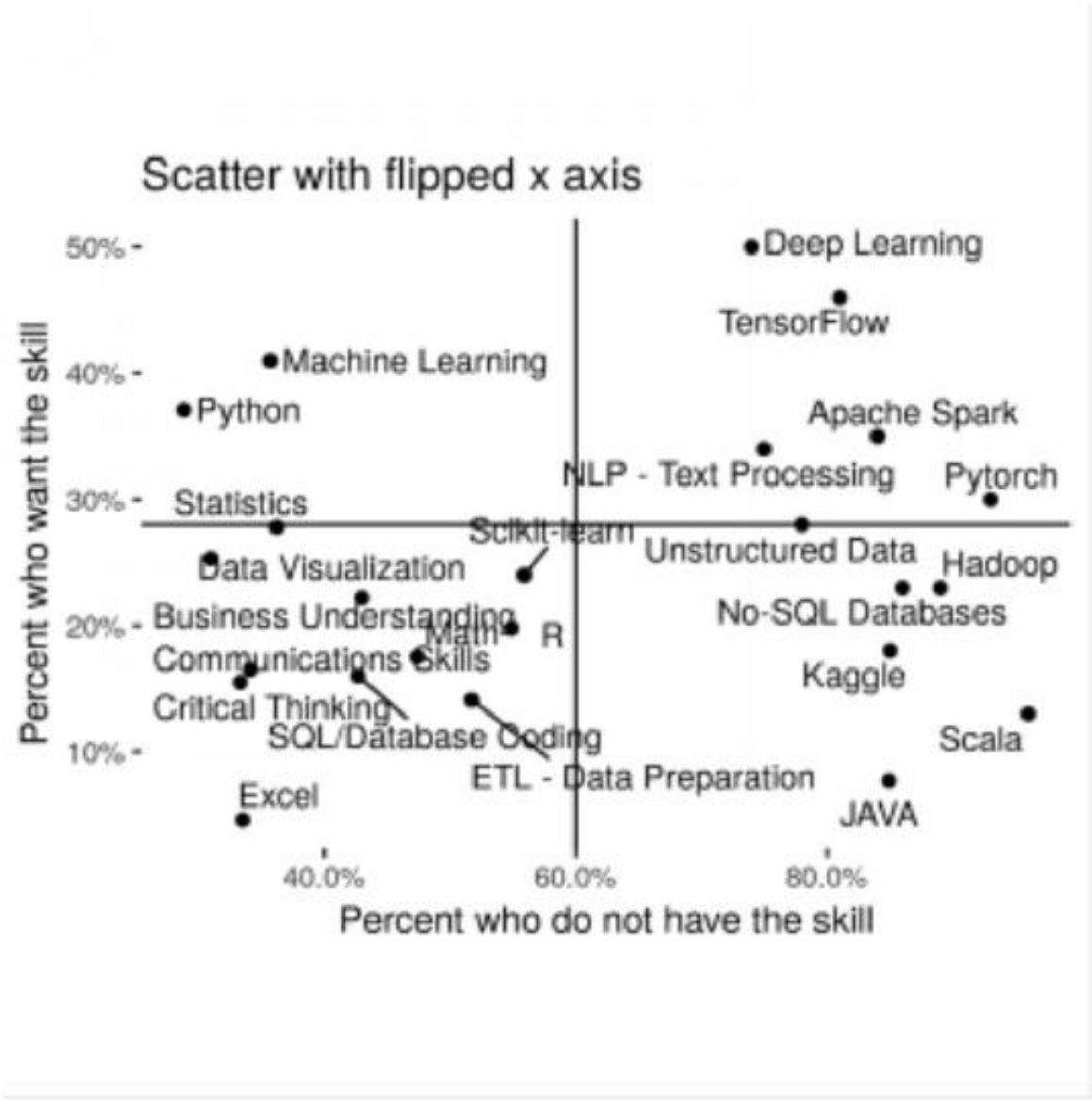


Fig 72 *Goals should be "high and to the right" – Guy Kawasaki.*

If you make a chart and no one remembers it. Did it still happen? In Fig 72, we grouped the skills in four categories but what good are they if no one remembers them? One way to help your audience to remember is personas (memes, in Gen-Z speak). Let's apply user personas. In Fig 73 each quadrant means:

- **Unwanted skills** (Have but, don't want = Excel)

- **No-thank-you skills** (don’t want and don’t have = JAVA)
- **Hot skills** (want but don’t have = Tensor Flow)
- **Loved skills** (want and have = Python)

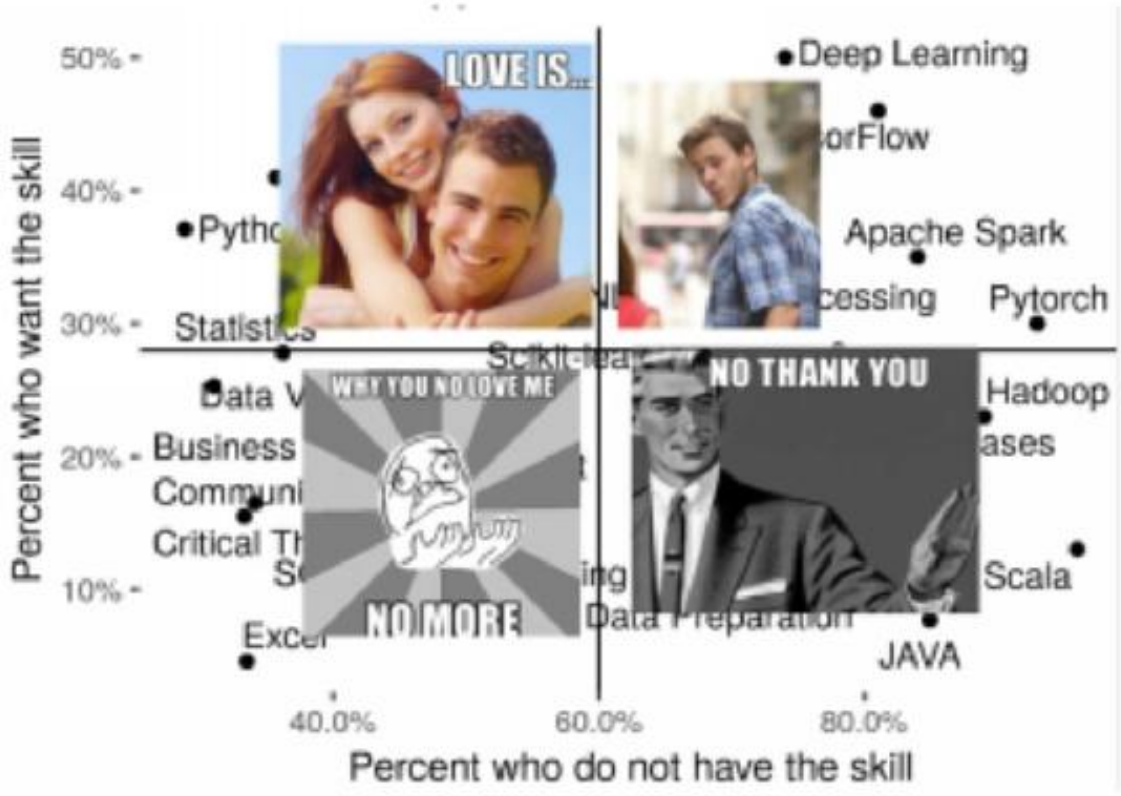


Fig 73 *Pop culture, use to make your chart stick.*

Red pen

Don’t be afraid to red pen your canvas! Captions are an opportunity to clarify meaning and add punch to your story (not everyone is visual). Note how in Fig 74 we broke the symmetry by tilting the “loved” label, that is Feng-Shui for charts. We also added a twist in the Java quadrant by not having a label for it. This ensures that the reader will go to this quadrant after visiting all the others. The label for this quadrant is inside the meme (No thank you).

Layering

Note how we have layered information in hierarchies (meme, quadrant labels, quadrant representative). We have respected the seven-chunk limit in each layer to avoid overload. Meaning was achieved by linking to an

existing framework and organizing the data into quadrants and creating a labelling them.

Narrative

Finally, charts should have a purpose. It is reasonable to go through the process of knowledge creation without knowing why. Once knowledge has been found and visualized, the why will be easy to find. My personal why for this chart is: “I like to see more Python and less Java in my classroom”.

Fig 74 *Can Magic Quadrants make a scatter plot memorable?*

Thinking about strategy

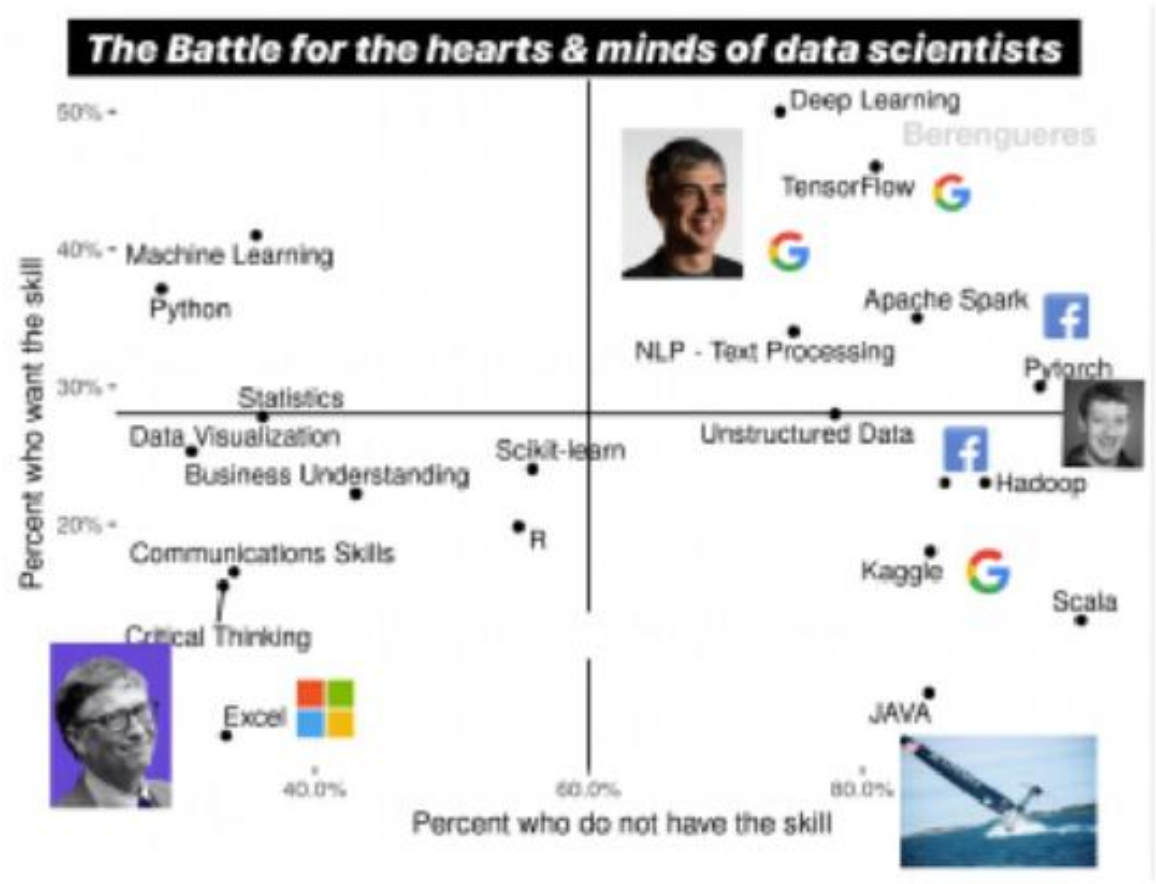


Fig 75 *Salesforce acquired Tableau. Why?*

Now that we created knowledge, can we use this chart as a thinking tool? One way is to imagine contexts where this chart might be useful. Where could this chart be used to create situational awareness? The figurative “cloud wars” between Microsoft and Google are fought via proxies such PowerBI, kaggle and other cloud software lock-in levers. A similar play book developed in the 90’s in the database market. Fig 75 visualizes who sponsors which language to see where allegiances stand.

Quiz: Visualizing gaps

Fig 76 Data science skills - *How to conjugate absolute magnitude and relative differences. Source: KDnuggets. Image at original resolution (might appear blurry in some devices)*

In the same post, KDnuggets posted this chart. The purpose of this chart was to display the **gap** between *have* and *wanted skills* while not losing perspective of the absolute magnitudes involved in each skill. Why does this chart fail? How would you declutter this chart and make it more meaningful without loss of information? (no summarization). Time 1 hour. Hint: See waterfall stack bars with negative components by McKinsey.

Solution

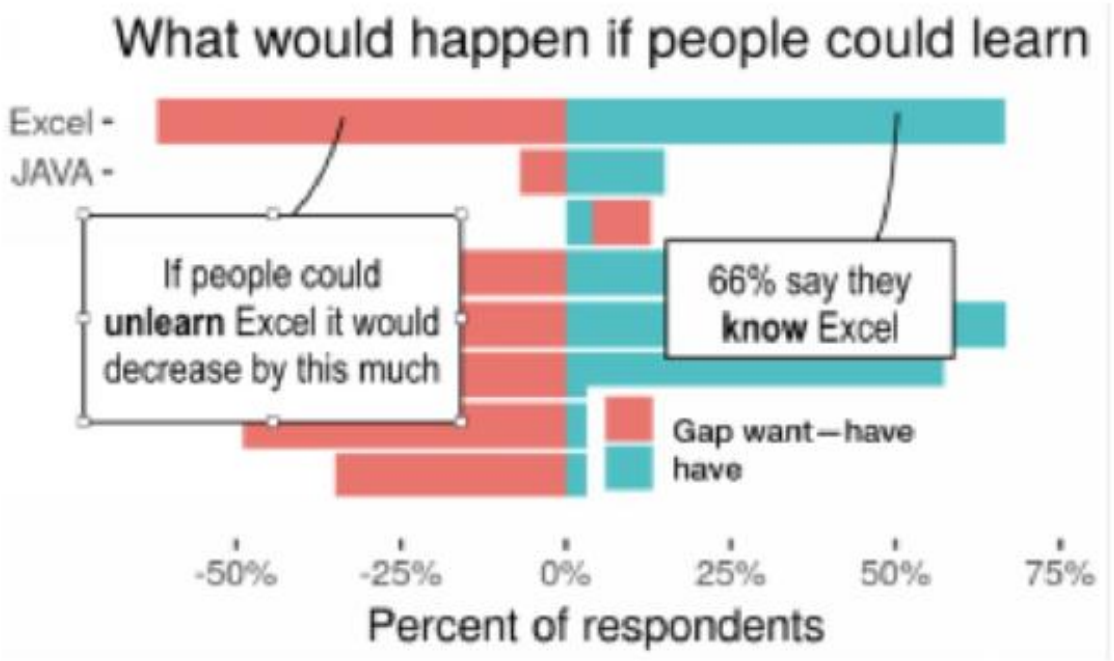


Fig 77 *If people could unlearn things, would they? (Only 3 items are shown due to resolution limits of the device)*

Explanation of the decluttering process

The author wants to display for each skill three attributes:

- **Have** skills
- **Wanted** skills
- Wanted minus have (the gap)

Of these three only two are independent variables. Applying the principle of eliminating superfluous information we display only two. Have (as the underlying) and the gap (want minus the underlying).

Zoom in

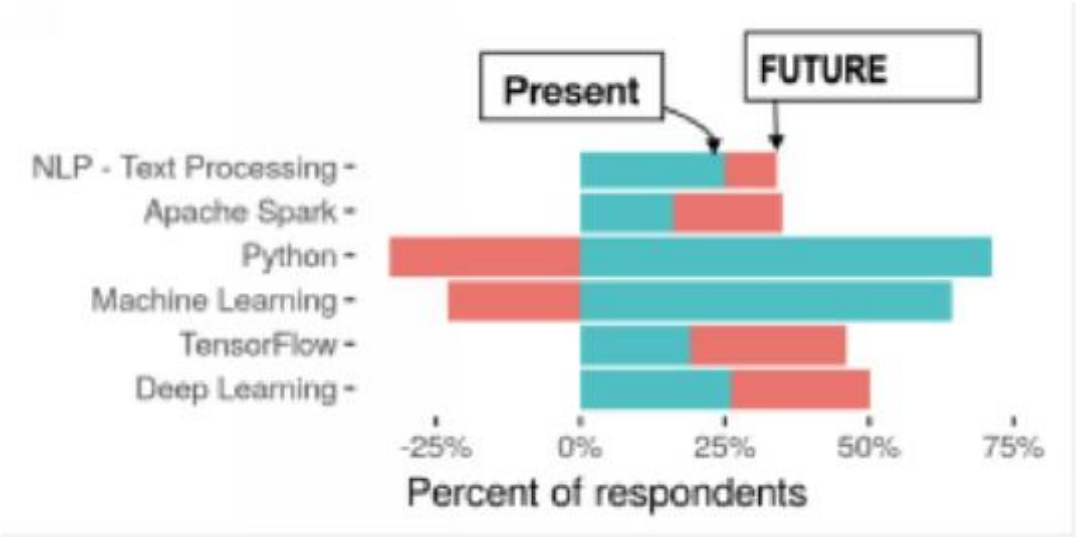


Fig 78 *What if you could learn anything you wanted? (Only 6 items are shown due to resolution limits of the device)*

This chart has now become a **predictor** of what would happen if people learnt what they say they want. *Have* is the current (prevalence) level. *Have + gap* is the future level. In the case of negative gaps, the gap bars are plotted on the other side of the y-axis, (a glitch of the stacking function of ggplot2 or a feature - one cannot unlearn). Note how, from a compete narrative of Fig 71 (competing bars), we have switched to a **growth mindset** narrative with the "what if you could learn anything you wanted". How might we use this chart to prioritize what skills to teach.

Chapter 6 Psychology of Visualization

with Ali Fenwick

Psychology plays a central role in designing effective visuals, especially when it comes to persuasion and influence. This chapter will take a further look at how behavioral principles can be used to influence visual design and data interpretation. Specifically, we will focus on three psychological concepts that apply to visualization: Context, Framing, and Gestalt theory. Familiarizing yourself with these principles and techniques and learning how to apply them effectively will greatly improve how you design persuasive visualizations.

Behavioral science

The field of behavioral science is a broad set of specializations focusing on theory and frameworks ranging from anthropology to psychology. Insights from cognitive psychology and behavioral economic theory can be useful in understanding how visualization techniques can affect human judgment, especially when deciding how to visualize data and how to tell a story. In real-life settings, important decisions are not just based on facts, but also on subjective partialities. As we saw in the previous chapter, the *way* information is presented, rather than *what* is presented, can be equally if not more influential on how people interpret and act upon the data. As Tversky & Kahneman explained in their seminal paper, written over 40 years ago, judgment and decision-making are heavily affected by unconscious factors such as psychological, emotional, social, and contextual factors [41]. Being aware of how these factors affect cognition is a first step toward developing effective and ethical visualizations. In the next section, we will start by exploring the power of context and how it plays a crucial part role in data visualization.

Context

According to Thomas & Millar, context is one of the strongest factors of influence on judgement and decision-making[42]. This is because our brain uses all the senses (e.g. sight, sound, smell) to collect information about the environment to create a coherent story of what is going on. In presentations, for example, visual design, speaker voice, dress, body language, and the physical environment all together communicate a *'single narrative'* to help the brain give meaning to the experience. The brain also tries to make sense of all these inputs by relating them to our existing narratives (e.g. our values, beliefs, and mental frameworks). In other words, individual information is not (and cannot be) processed in isolation or independently from its environment.

When speaking to an audience for the first time, the brain uses all of these inputs to create an opinion about someone or something. The science shows that this happens within the first 11 seconds of your encounter and influences how people will further evaluate you or your presentation. That is why, for example, when I give a presentation to an audience I have never met before, I make sure to rehearse the first minute of my presentation meticulously so that I can recite it word by word during the opening of presentation, I suit up to look the part, make sure to face my audience confidently, and if possible have someone else introduce me. Techniques I know will significantly improve that first impression and consequently my success in delivering an impactful presentation. Sir David Ogilvy, the advertising pioneer, was a master at using context to influence his audience and built his career upon it! In his 1963 biography *'Confessions of an Ad man*' he recounts several techniques he used. Next, let's look at an example of how context, in this case relativity, affects visual perception.

Principle of relativity

Take a look at the Fig 88. Which blue circle is bigger?

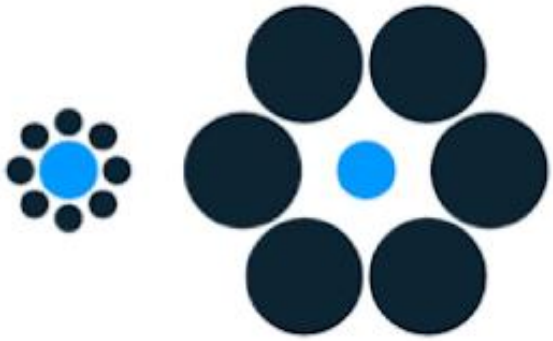


Fig 79 *Context influence on perceived size.*

You can already expect what the correct answer is: they are both the same size. But perception-wise, they are not. The blue circle looks bigger when it is accompanied by smaller circles around it and looks smaller when we put bigger circles around it. Cover the outer balls with your fingers and you see that both blue circles are exactly the same size.

This visual illusion mirrors how our mind interprets the world. It is therefore unwise to ignore the broader impact of contextual factors on your presentations. This effect not only holds true for charts, but also for ephemeral things such as emotions, attitudes, and points of view. This means that our brain uses both the external context (e.g. visuals, sounds, smells) as well as the **internal** context (e.g. emotions, past experiences, desires) to interpret the narrative as it happens.

The mental processing of contextual information, be it visual or not, happens predominantly at the unconscious level. Less than 20% of our

awareness is conscious awareness, which means that factors such as context can easily influence judgment and decision-making without people being aware of it, having both intentional and unintentional effects. Let's see with a food example how context can influence eating behavior unconsciously.

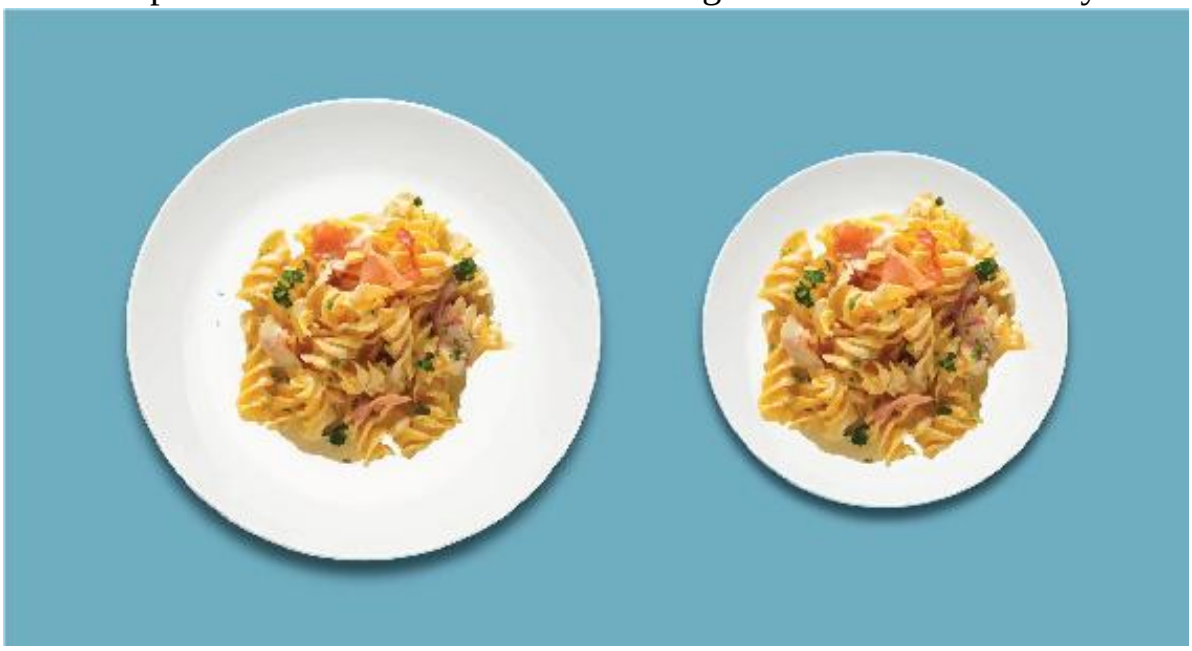


Fig 80 *Professor Wansink studied the effect that the plate's size has on your intake of calories. Source: Pelle Guldborg Hansen, inudgeyou.com*

The same amount of food is presented on two different size dishes; one being bigger that the other. Visually, the smaller dish makes the food look bigger and studies have shown that people who eat from the smaller dish *feel* full quicker and therefore eat less. People eating from the bigger dish tend to eat more than the people eating from the smaller plate. Now think how this applies to charts.

Cognitive Overload

According to Sweller, cognitive overload is a phenomenon during which the brain is unable to process information effectively due to the sheer amount of information presented[43]. Cognitive overload causes mental fatigue and reliance on mental shortcuts (which are subjected to heuristics and biases). It is therefore important that design incorporates visual elements which help overcome cognitive resource limitations and prevent faulty decision-making.

As we saw in the Magic Quadrants in Ch. 5 the main mechanisms to avoid overload are:

1. **Limit** the amount of data presented on a table or slide
2. **Structure** your data into bite-size pieces of relevant information, and
3. **Highlight** key words and phrases in sentences which help the reader pick up the most important parts of the text

Framing

Triggering mental shortcuts in visual design can also be beneficial to improve visual effectiveness. One way mental shortcuts can be triggered in visualization is through framing. Framing is a cognitive bias which affects judgment and decision-making using positive (gain) or negative (loss) messaging (see again Kahneman's work[44]). People tend to be more motivated to take action when messages are framed as either positive (a gain) or negative (a loss) depending on the situation. For example, studies show that people are more likely to seek **risks** when a message is framed as negative or a loss. Here is an example.

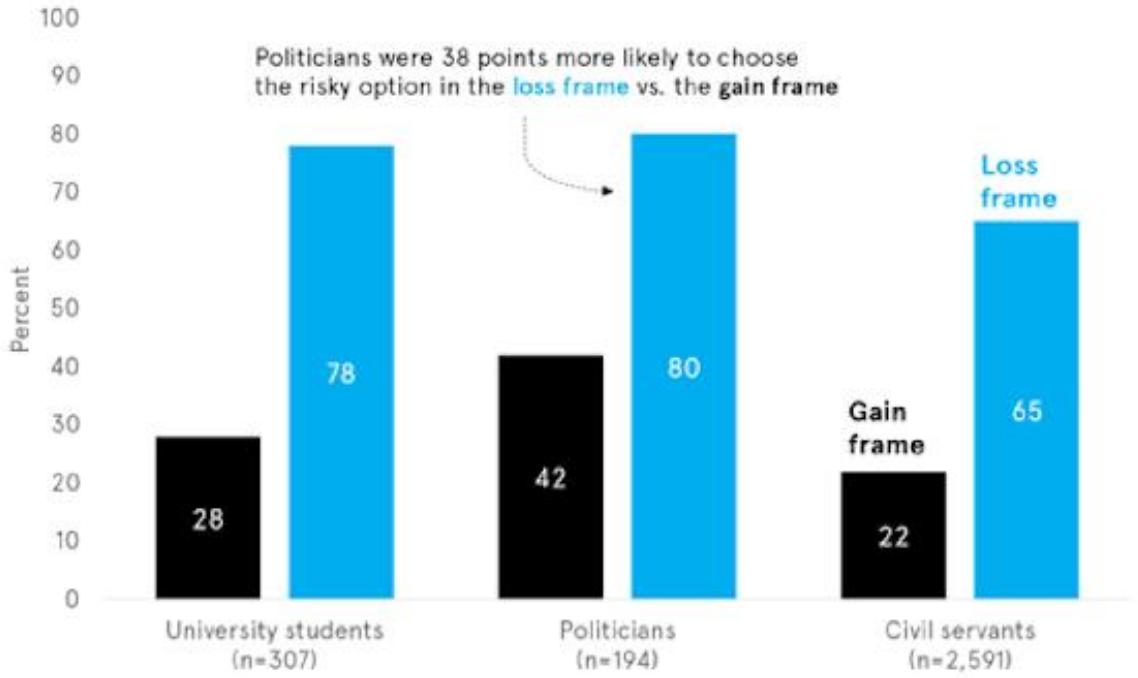


Fig 81 *Avoiding deaths (loss thinking frame) leads to more persuasion than phrasing a policy proposal as "saving lives" (a gain thinking frame). Source: bi.team*

According to the above diagram, government officials were **three** times more likely to take riskier options to combat a disease outbreak when policy design was framed as preventing lives lost versus lives saved. In medical situations, studies show that it wiser to use positive framing, instead of

negative framing, to help patients make better health-decisions for themselves, especially in high-uncertainty situations. Interestingly, recent findings in framing effects have found that certain factors might moderate the impact of framing. For example, a study by Thomas and Millar in 2011 showed that framing effects change with age, affecting decision-making and memory of younger and older people differently. Framing effects have also been found to disappear when information is presented in a second language or a language which does not come natural to the reader. This can be explained by the fact that more deliberate thought is needed to process information in a foreign language which reduces the brain’s tendency to automatic (biased) responses.

Gestalt

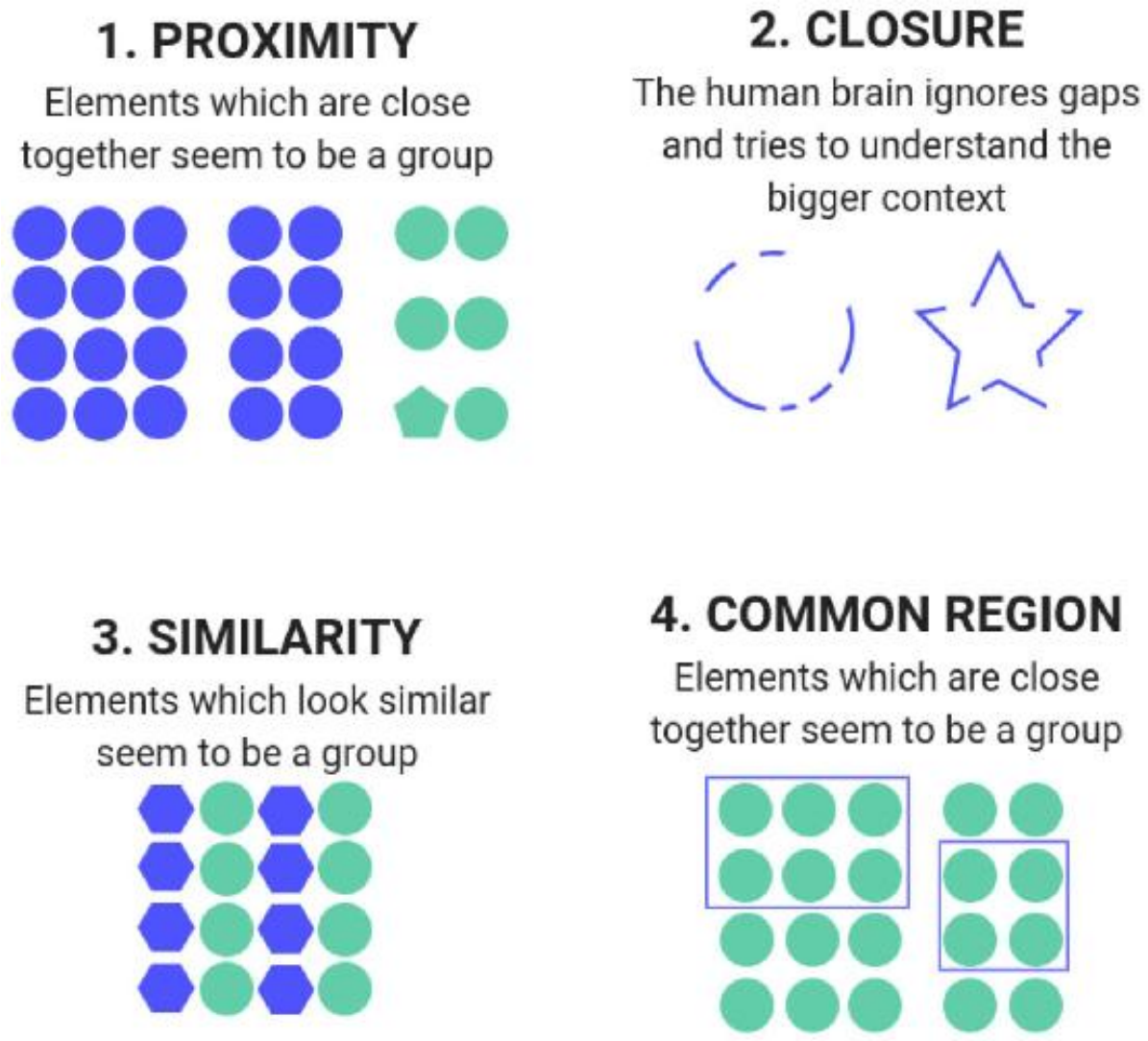


Fig 82 *Four of the eight Gestalt principles. Source: Annemarie Bufe*

In the early 20th century, a group of German and Austrian psychologists provided an understanding of how the human brain interprets and organizes visual elements. Their investigation lead to the creation of the Gestalt Principles for better visual design to reflect the brain's natural way of processing information. Gestalt is a German word which means pattern or a unified whole. Let's take a look at how some of these principles can be applied to data visualization.

The Law of Proximity

The law of proximity reflects the need to group similar items together or in close proximity to each other. Grouping similar items together makes it easier for the reader to understand what is being communicated. The closer parts are placed together, the higher the perceived relatedness. The mean reversion chart, for example, in Chapter 4 (Fig 38), is a great example of the law of proximity. The closer the label “Japan” is to its corresponding dot, the less the confusion and cognitive load. Tip: Watch out not to put unrelated objects too close to each other, as the eye will aim to create relatedness based on the proximity of items presented.

Enclosure law

Enclosing elements together creates relatedness. Apply enclosure to highlight certain elements in a visual. Enclosure could reflect the usage of (dotted) circles on a scatter plot to indicate belonginess amongst specific pieces of information (e.g. clusters, outliers). In the Fig 83-84 below we have added an example of enclosure. The circular enclosures of Fig 84 help reduce cognitive load of Fig 83 and increase its meaning.

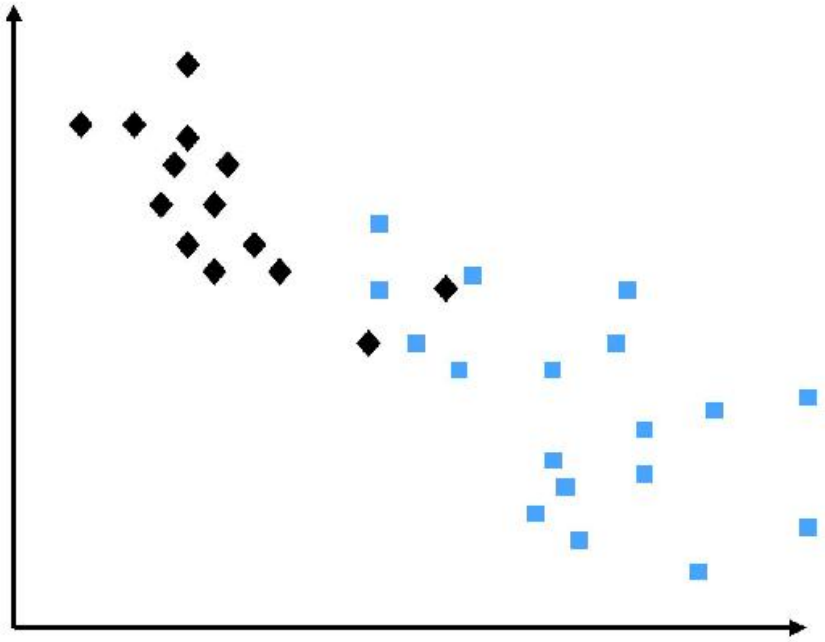


Fig 83 *A scatter plot with two classes of data points.*

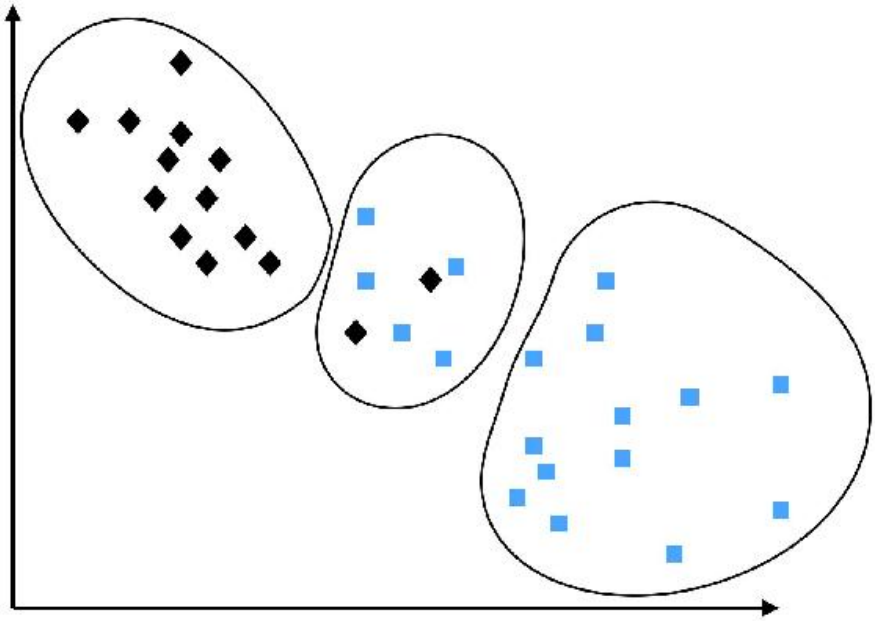


Fig 84 *Enclosures, use to make meaning.*

The Peak–End rule

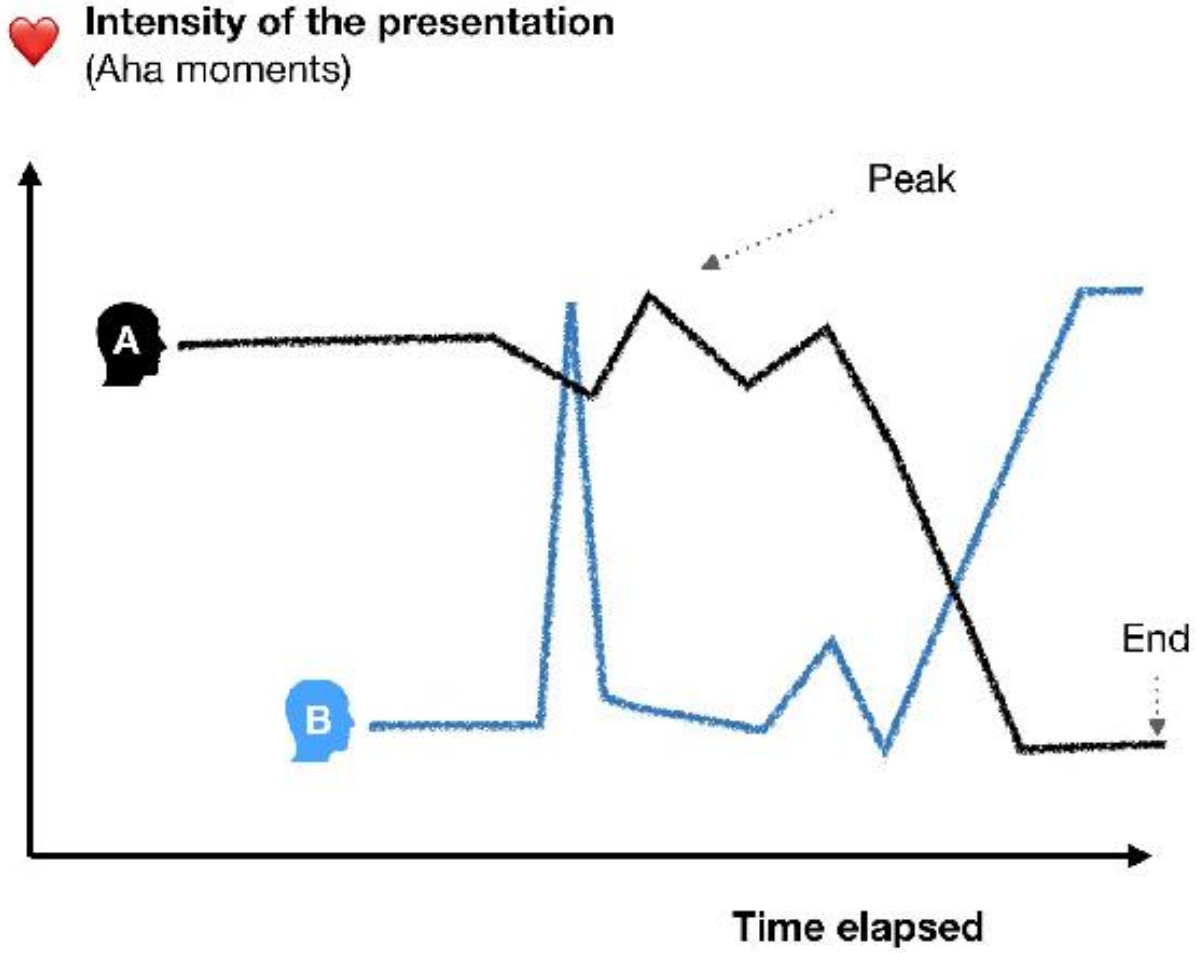


Fig 85 *Peak-end rule, use to maximize impact.*

Finally, let's see a fascinating psychological principle called *the peak–end* rule, a psychological heuristic in which people judge (or remember) an experience largely based on two specific moments: a peak moment somewhere in the middle of the experience and the experience at the end, rather than evaluating the experience in its totality. This surprising effect occurs regardless of whether the experience is pleasant or unpleasant. In other words, when storytelling your visual, the audience will remember the experience more favorably if it ends on a high note (save the loudest bang for last!). Fig 85 shows the intensity of two presentations by Mr. A and Mr.

B. According to this rule, the presentation of Mr. B will be rated nearly as favorably as Mr. A's presentation even though Mr. A presentation was longer and had more *Aha moments*.

Steve Jobs used this rule to maximize the impact of his own presentations by always ending in a high-note. His "*One more thing…*" is now part of the pop culture. Throughout 1998 to 2010, Jobs used it in no less than 25 occasions. Greg Wyatt of *@appleexplained* made an hour-long compilation of such moments. The peak-end rule not only can be applied to presentations but to charts as well. An example is Fig 65 in "twists". There, (at the expense of readability), the author maximized the chart's punch by making sure that the *Aha moment* is not reached until the audience went through all the other elements of the chart first.

Chapter 7. Detecting Bias

with Marybeth Sandell

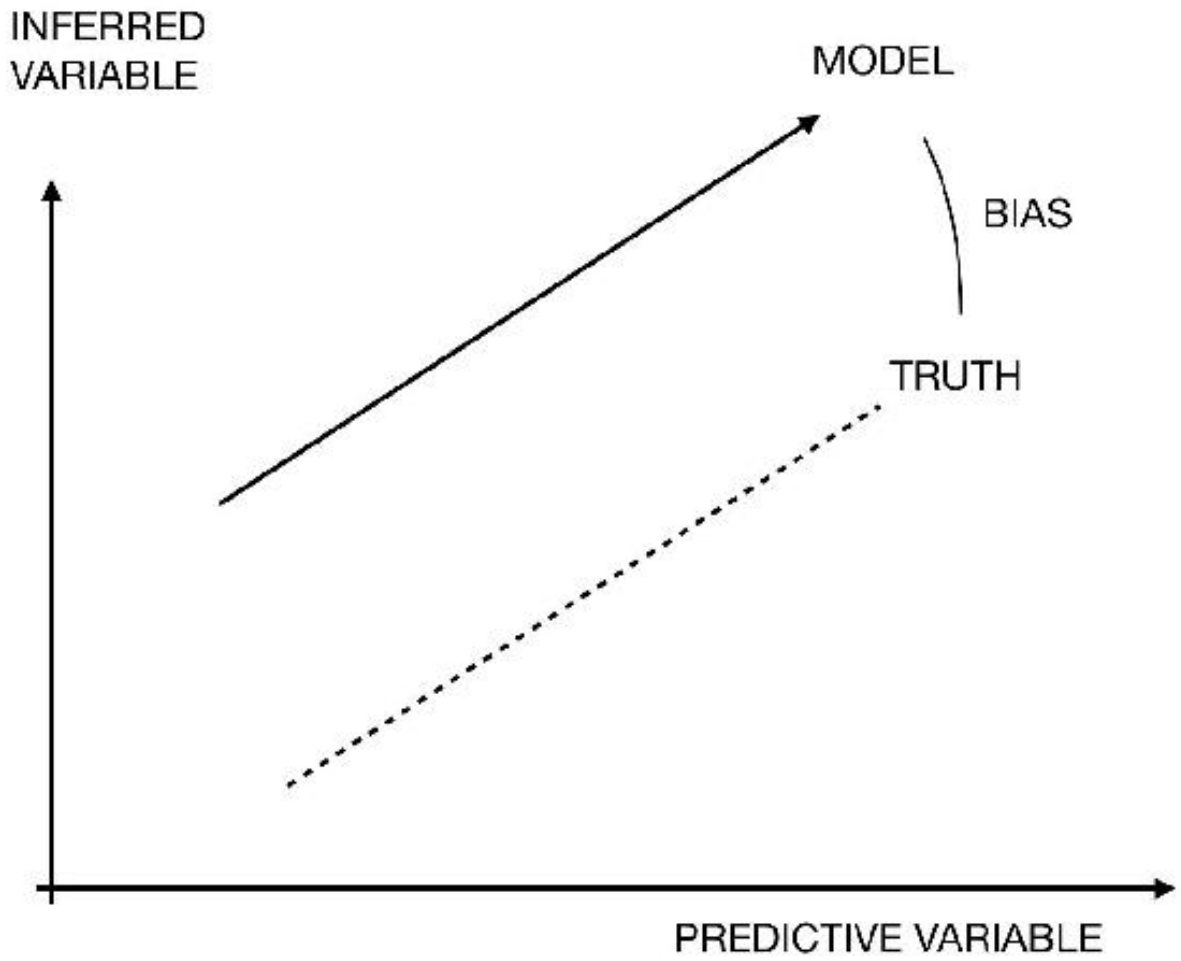


Fig 86 *Bias is unethical when it is unfair, usually against minorities.*

In this chapter, you will learn what bias is and how it can affect a data-driven visual. Bias not only can be sorted by their point of entry (data, story, narrative) but also by the area they exploit in the cognition system (optical illusions, cultural biases). It is easy to assume that bias is intentional. However, bias can emerge for many reasons.

First, bias can be embedded in the data itself, intentionally in the way it is gathered but also accidentally by not realizing what is missing.

Second, bias can appear as the story is crafted. Again, this can be intentional by cherry-picking from existing data, or accidental from cases where not enough time is spent exploring all data available (usually due to time pressure).

Third, it can be embedded in the narrative itself. Often this is intentional, as in propaganda. But it can also be unintentional as in cultural bias.

Types of bias

In broad terms, bias is any systematic error. In other words, a systematic difference between a model and the "truth" it supposedly represents. In social sciences bias is judged to be unethical when it is **unfair** (usually towards a minority). See also ethical frameworks in Ch. 1.

Bias can affect the producer of a visual (as in selection bias in data), but also the consumer of the visual (as in *Groupthink*, and the *hot hand fallacy, and so on*). Psychologists and behavioral economists have identified more than 200 types of cognitive biases. Those can be classified in three groups:

- Belief bias
- Social bias
- Memory bias

In addition, a part of the mentioned cognitive biases, when dealing with data visualization, visual perception biases also apply. Let's see some examples.

Bias in narrative

The broadest forms of unconscious bias are due to unawareness and are so rooted in society they usually are cultural (moral) norms too. Note that not all cultural norms are biased but that most norms evolve slower than society does and thus are usually lagging behind reality. Examples of conscious narrative bias are: Propaganda and disinformation. Typical techniques used are **FUD** (Fear, Doubt and Uncertainty) as seen with the tobacco industry and **FLICC** (Fake lies, Logical fallacies, Impossible

expectation, Cherry-picking and Conspiracy theories) as seen with the climate change denial[45]. Let’s see an example.

Bias in narrative: A balanced meal?

In the second half of the 20th century, a **balanced** diet was assumed to be optimal for health. In school, many kids (myself too) were shown charts with relatively balanced food groups. Yet, other cultures and a few independent research papers show that perhaps that balanced food narrative isn't the healthiest one. For example, Okinawan diet contains less than 5% of animal protein and no milk derivatives. Their diet would be considered dangerously unbalanced by any Western standards. However, Okinawans report one of the healthiest and longest lifespans in the World.

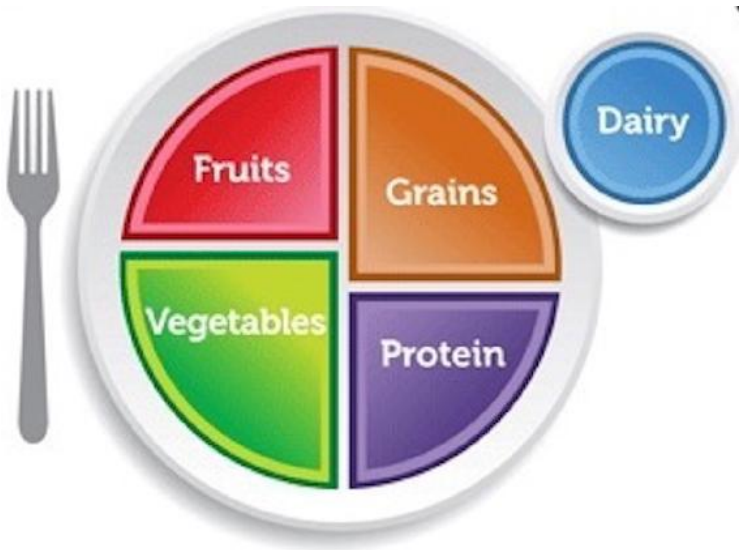


Fig 87 *Logical fallacy? How could this "balanced" diet not be healthy if it is balanced? Source: US Department of Agriculture choosemyplate.org*

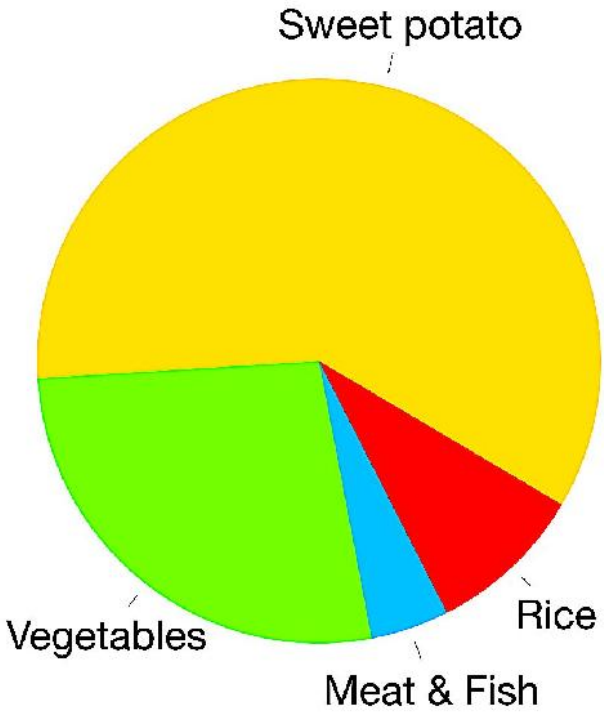


Fig 88 *Okinawa diet in 1950 is correlated to the world's longest and healthiest life spans. Does it disprove the popular belief that an unbalanced*

diet cannot be healthy?

Bias in the story: Hard working Germans?

When telling a visual story, how the data is used and presented can induce conscious and unconscious biases (well known in advertising and neuropsychology). This can be done through the selection and the presentation of data, colors used and so on. Let’s see an example.

Exercise: Hard-working Germans

Fig 89 is a graph created by the German Economic Development Agency. They used EU data to show that Germans work longer (harder) than the average. But how much harder? Can you identify where is the bias? Time 4 minutes.

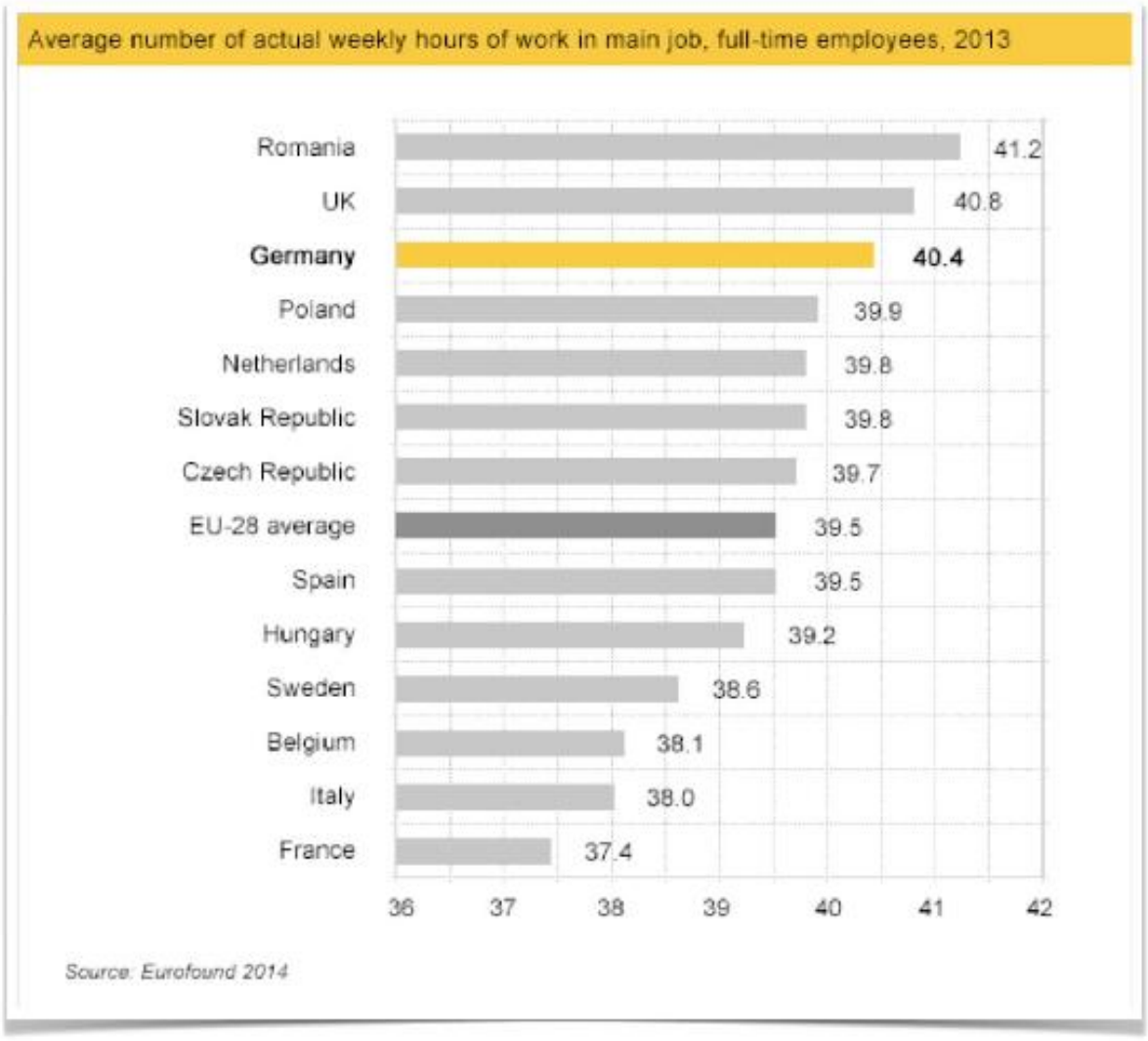


Fig 89 *A partial view?*

Solution

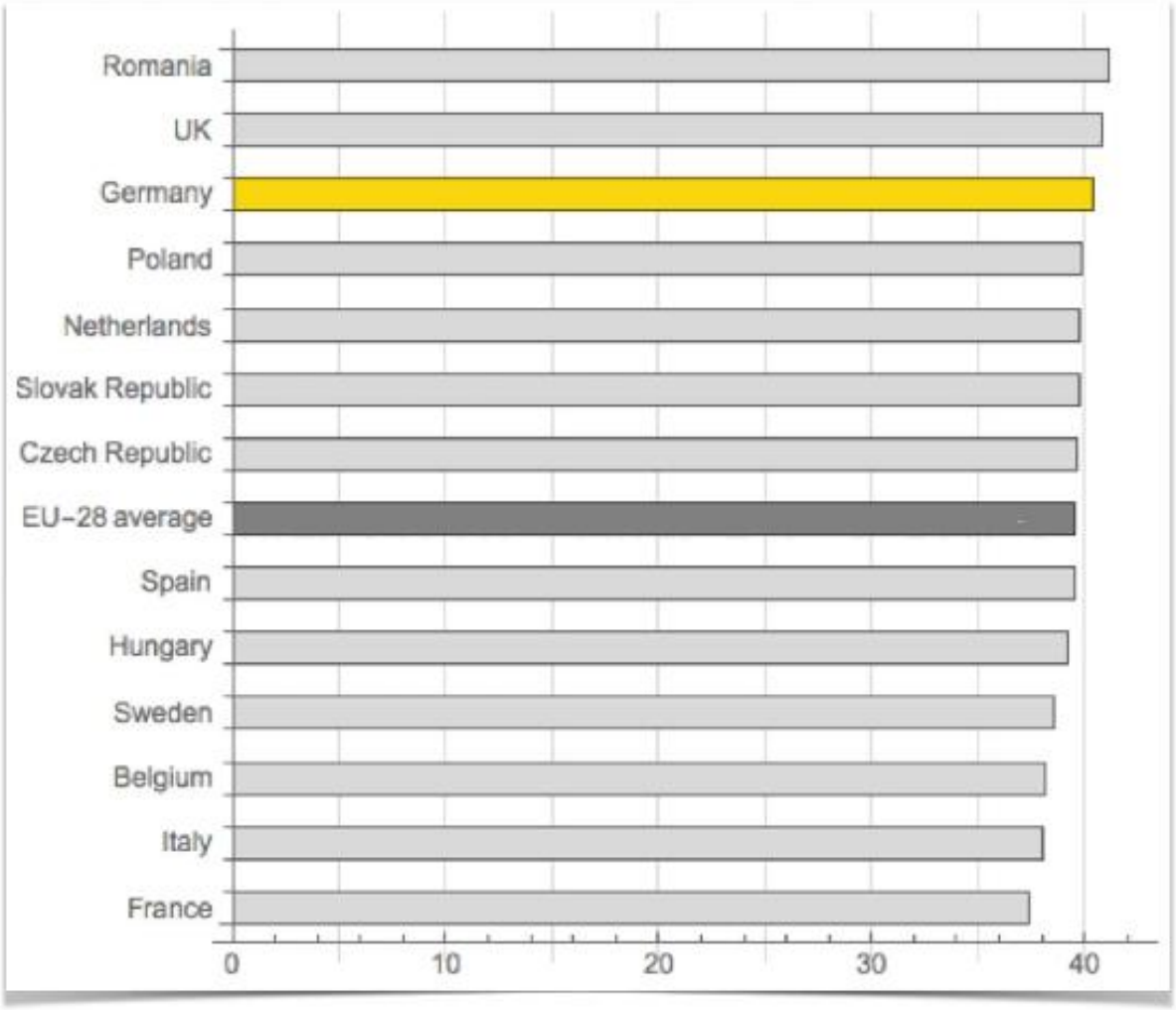


Fig 90 *A fuller view.*

In the chart, Germany optically doubles France. In reality, the difference is less than 10% because the axis starts at 36. A website redrew the graph to start at zero and suddenly, German doesn't stand out much at all. This chart might also suffer from data biases that for a trained statistician are obvious. For example, (due to historical reasons), Germany has more industrial jobs than any other EU country. Germany has more double the amount of

industrial jobs per capita than Spain for instance. If industrial jobs have different regulation or a different mix of full time / part time workforce than other sectors (service jobs) then the chart is not comparing apples to apples. The chart just reflects a different job mix in each country. This type of bias is called data **skewness** and it is one of the common errors of amateur statisticians: to confuse correlation with skewness. Let's learn more about data bias in the next example.

Data bias

In journalism, bias in data comes under three broad umbrellas: selection flaws, skewness and omission. Some examples include: how questions are asked (as in leading questions); bias in hypothesis or assumptions; mathematical errors and survey design flaws. Let's see an example of a famous selection bias[46] discovered by the student Thomas Herndon[47].

Bias in data: Austerity controversy

Growth in a Time of Debt (2010) was a non-peer reviewed paper published by Carmen Reinhart and Ken Rogoff - a former chief economist of the International Monetary Fund. It said that a country's economy slows significantly when debt rises above 90% of GDP. This is not true. However, the paper was used to justify austerity for the sake of the "common good", (see ethical frameworks in Ch. 1). A student trying to replicate the results discovered that from the 20+ countries considered to make the statistic, key countries had been arbitrarily omitted in the calculations. Later, Prof. Michael Ash noted another unbecoming behavior in the weighting average method used. For example: "New Zealand's (a country of population less than 5 million people) that had one single year, (1951), at -8% growth was held up with the same weight as Britain's 20 years at 2.5% growth (population 60 million)". On a dollar basis, that is a bias ratio of 1 to 240. By now we have seen three examples of bias in the story, data and narrative. Now let's see a few more biases often at play in charts.

Other biases

Confirmation bias – we tend to believe statements and data that confirm our beliefs and ignore evidence we dislike, (see previous section).

Picture superiority effect – presentations with charts have more credibility.

Self-generation bias – we tend to remember what we say better than what others say. So, if you ask a question during your visual presentation, and let the audience come up with the conclusion they will remember it more.

Illusion-of-truth bias – we are more likely to identify statements we hear in repetition as true. Just because it's been retweeted by multiple sources doesn't mean it is correct. For example, the previous austerity paper was cited more than 2000 times before the bias was uncovered.

Optical deception – Another way in which visuals can be misleading are optical deceptions. In 2015, in the paper "*How Deceptive are Deceptive Visualizations?*", Dr. Pandey studied how such manipulations are used in charts [48]. As we have seen in the German working hours' example, most of the manipulations in charts consist in playing with axes zoom levels and praying that no one will notice it. Note it is not an excuse to say that we didn't *hide* the scale of the axis, the way Panday measures deception is by measuring what people *understood* after watching a visual, not what its written in the visual.

The Amazon forest fires case

Bias can be found at multiple levels. For example, look back at the fires of the amazon forest in Brazil in the summer of 2019. This story dominated the major news outlets for a few weeks as some readers expressed outrage the impact the fires would have on the earth’s climate. Apply what we learnt so far to find potential biases in the following chart. Time 5 minutes.

Exercise

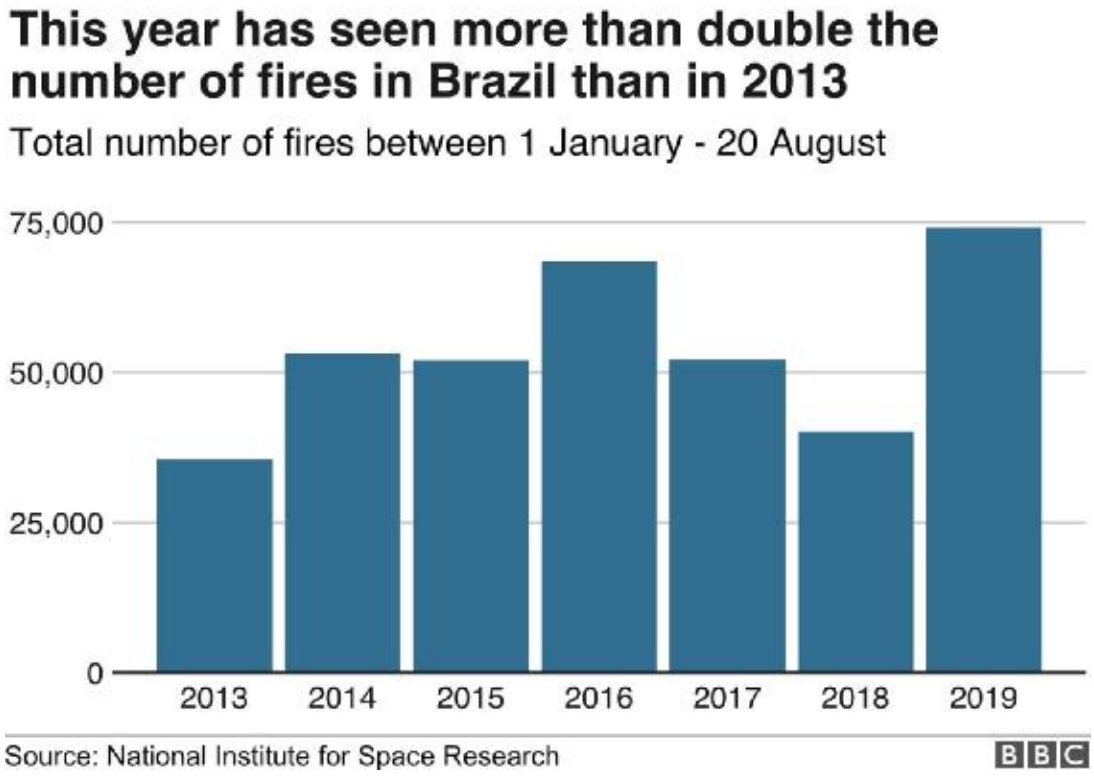


Fig 91 *Where is the bias?*

Solution to fires

Data was central to this story. Just how bad were the fires? There was one main data source — the National Institute for Space Research (INPE[49]). Initially, the news coverage presented data that encouraged the reader to be outraged and suspect a crisis was at hand. The chart displays data on the **number** of fires from 2013 to 2019. The graph leads the reader to think that this was the highest level of fires ever. In fact, CNN on Aug. 22 wrote a story that the forest was “burning at a record rate”. The first sentence of that CNN story said the rate of burning was a record “since INPE began recording tracking fires in 2013”. Between the use of words like ‘double’ and ‘record’ and the use of a visual with the 2019 bar looming over all the other years shown, the narrative was set. To be sure, agencies

like CNN added to their written story lead: “...and scientists warn it could strike a devastating blow to the fight against climate change.”

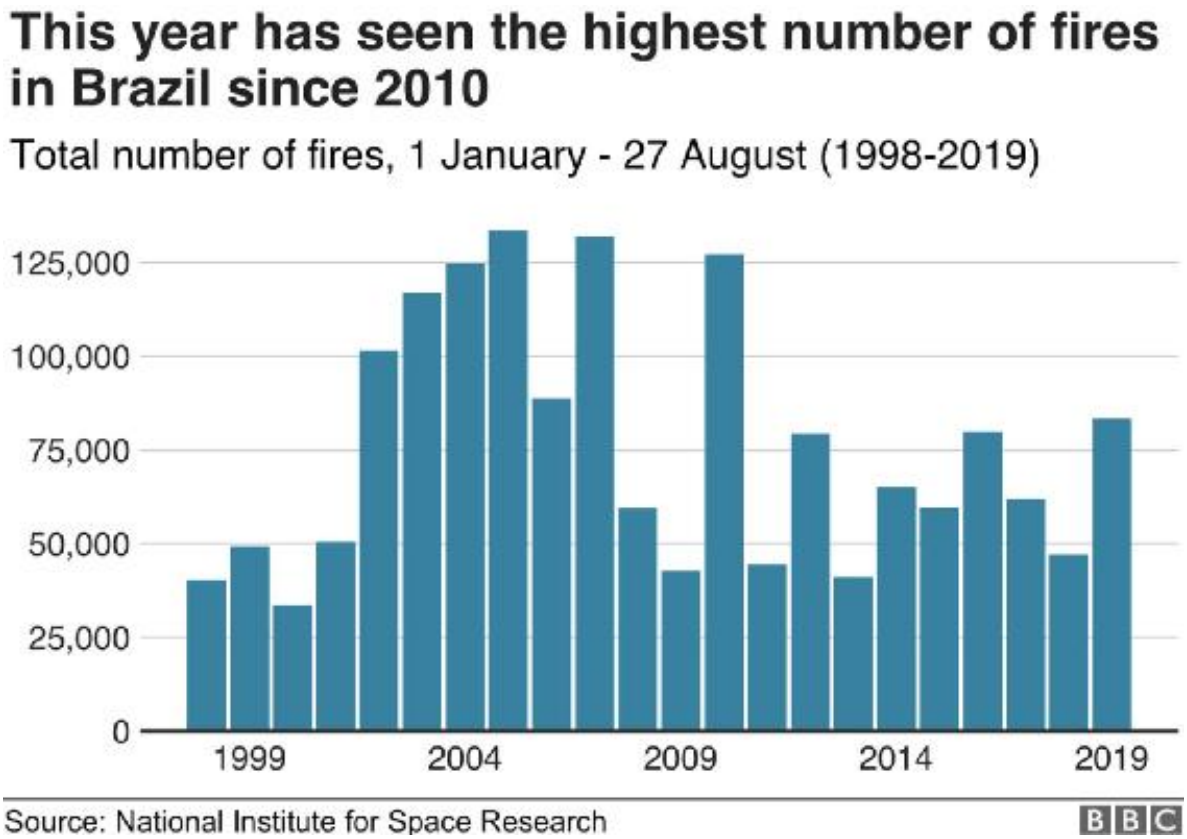


Fig 92 *Good journalism?*

However, **a mark of due diligence is to always ask if there is more data**. Often it is there but not available online, or not free online. It may be in a drawer or behind a paywall. It may exist but the time-starved reporter didn't have time to look, or even possibly intentionally ignored it.

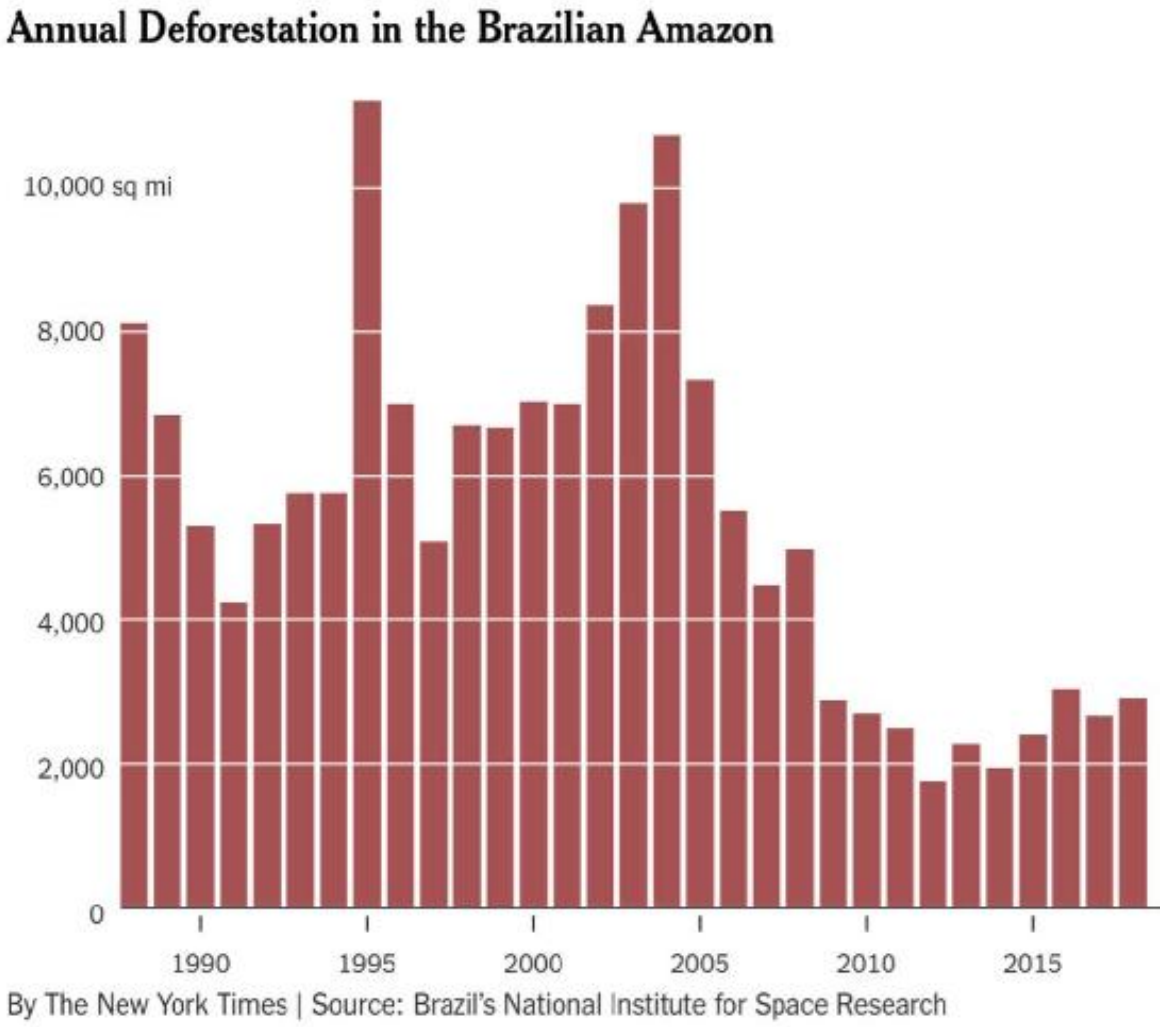


Fig 93 *Better journalism.*

In this case, INSE did indeed have more data regarding number of fires (it was just not online!). The BBC updated its chart to include more years. Here is how it looked when digging further into the non-online archives.

Still, is this the best data we can get? Is there more? Consider what is being measured: **number** of fires. Does this mean that if we light five small fires today and one big one tomorrow, the total number of fires is declining? Then, The New York Times published the chart below. Here, we have the same source, INSE, but the measurement is square miles burned. The

numbers used initially weren’t wrong, but rather they were not complete or fully reflective of the situation, a case of unconscious selection bias.

Due diligence checklist

1. **Pay attention to words**

Any good set of data will offer transparency into the methodology of how the data was gathered. This means paying particular attention to what and how questions asked in surveys or statements made. A red flag is any use of **adverbs** and **adjectives**. They are usually loaded with bias.

2. **Follow the money. Who paid for the research?**

Big tobacco showed us that the organization that pays for the research can control its results. For example, the egg industry lobbyists are paying for research at accredited universities to promote research that says eggs won't boost bad cholesterol in humans. In the 1960s, the sugar industry paid researchers to produce data that made consumers believe fat was a bigger health hazard than sugar. The list goes on. Next time you see research about health or the environment, try to discover the identity of the ultimate financial backer. Follow the money.

3. **Pay attention to the statistical methods used**

As we saw, sometimes the data is being selected to intentionally support a position. After performing some statistical analysis, a good rule of thumb is to always ask a more proficient data scientist to find flaws. It works wonders and one can learn a lot.

4. **Consider the availability of data**

Just because the data isn't publicly online it doesn't mean it is non-existent. Post millennial journalists who were never taught how to do research before the internet existed are particularly vulnerable to this bias.

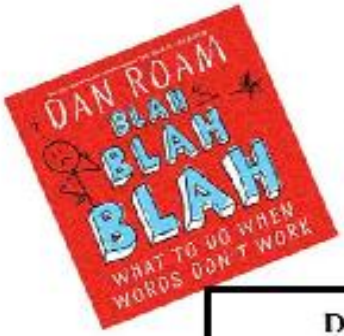
Quiz: Fire Tweets

Fig 94 *Where is the problem in this tweet?*

Answer: The photo is fake. Use Search google for photo to find out that this is Aerial view of the Taim Ecological Station on fire, in Rio Grande do Sul state, southern Brazil, on March 27, 2013. It is not from the Amazon forest fires of 2019. NYTIMES wrote on Aug. 23 about how misleading visuals went viral during the Amazon forest fires.

What to read next?

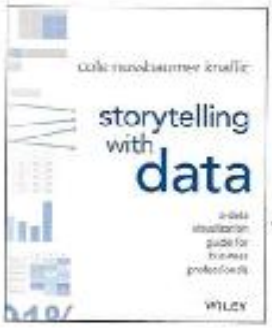
ONE OF THE BEST [60]



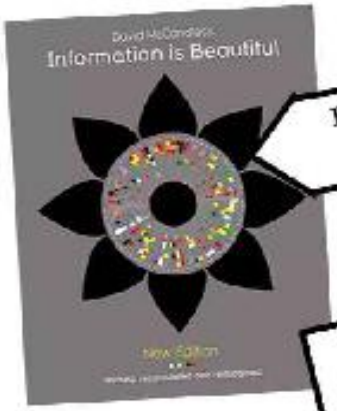
TEMPLATES [65]

DATA-BRIVEN JOURNALISM

THE HERO'S JOURNEY [66]



TO MARIE KONDO CHARTS [63]



INFOGRAPHICS UBER ALLES [67]

FOR BUSINESS [64]

Fig 95 *Direction pole metaphor, more human-centric than a list.*

Acknowledgement

This book received help from many friends. Special thanks to Cheryl Shuttle, Ricardo Bragança, Liz Ortiguera, Agatha Braja, Kumiko Sasaki, Louise Taylor, Natalia Rodriguez, Eric Nylen, Ferran Pujol, Eva Mitchell, Cliff Hazel, James Piecoway, Amal, the Kaggler user *Head or Tails* and Paul. Warm thanks to Anthony Goldblom, Carolina Ampuero, Francois Cadiou, Nicolas Paris, Alex Rios, Steven Anthony, DiDi library, the 2019 digital government conference organizing committee, SNASK, Susanna Vento, Kai Bruns, 710stories, the team at Ivory & Gold, and the Kaggle team. The ebook style is inspired on the 3rd edition of “Don’t make me think’.

References & Footnotes

---

[1] *Own estimates [go back]*

[2] Rose, T. (2016). The end of average: How to succeed in a world that values sameness. Penguin UK. [go back]

[3] Rand, P., 1985. Paul Rand: A designer's art. New Haven: Yale University Press. *[go back]*

[4] Isaacson, Walter. Steve Jobs. 2011. *[go back]*

[5] "It takes too long to style a chart like you want by using R or Excel. At the end, it is faster to use Adobe Illustrator" – Marybeth Sandell, former bureau chief London, Bloomberg. [go back]

[6] Berengueres, J., 2019, June. Visualization & Storytelling Workshop. In 20th Annual International Conference on Digital Government Research (pp. 532-533). ACM. *[go back]*

[7] *See also "how to detect genius ideas" in Ogilvy & Advertising. [go back]*

[8] See also Tanaka, K., 1997. An introduction to fuzzy logic for practical applications. [go back]

[9] *See also the film Merchants of Doubt and Cambridge Analytica scandal. [go back]*

[10] *Gallo, C. 2019. The Art of Persuasion Hasn't Changed in 2,000 Years.* [go back]

[11] *Reynolds, George. Ethics in information technology. Nelson Education, 2011. [go back]*

[12] *Jefferys, S., 2003. Liberté, Egalité and Fraternité at work: changing French employment relations and management. Springer. [go back]*

[13] *Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108. [go back]*

[14] In 2006, Al Gore hired Nancy Duarte to make his famous CO2 chart presentation. A scissor lift was used on stage to show that the CO2 is "off the charts". Al Gore was later criticized by emitting a lot of CO2 due to his jet travelling. [go back]

[15] *See also David and Goliath in the book of Samuel. [go back]*

[16] *Thunberg, G., 2018. Speech at UN Climate Change COP 24 Conference, Poland 2018. Published online on YouTube by Connect4Climate. [go back]*

[17] *Hyman, J., 2006. The objective eye: color, form, and reality in the theory of art. University of Chicago Press. [go back]*

[18] *Schippers, M. C., Scheepers, A. W., & Peterson, J. B. (2015). A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap. Palgrave Communications, 1, 15014. [go back]*

[19] *Dweck, C., 2015. Carol Dweck revisits the growth mindset. Education Week, 35(5), pp.20-24. [go back]*

[20] *Corbett, Christianne, and Catherine Hill. Solving the Equation: The Variables for Women's Success in Engineering and Computing. American Association of University Women. 1111 Sixteenth*

Street NW, Washington, DC 20036, 2015 [go back]
[21] *Norman, D., 2013. The design of everyday things: Revised and expanded edition. Basic books.* [go back]
[22] *Mourant, R.R. and Rockwell, T.H., 1972. Strategies of visual search by novice and experienced drivers. Human factors, 14(4), pp.325-335.* [go back]
[23] *See also "Generations in the workforce"* [go back]
[24] *Sinton, E (2011). 'Baby boomers are very privileged human beings' https://www.telegraph.co.uk/finance/personalfinance/pensions/88409 63/Baby-boomers-are-very-privileged-human-beings.html retrieved October 23, 2013 from www.telegraph.co.uk* [go back]
[25] *] Lukianoff, G. and Haidt, J., 2018. The coddling of the American mind: How good intentions and bad ideas are setting up a generation for failure. Penguin.* [go back]
[26] Palmore, Erdman. Ageism: Negative and positive. Springer Publishing Company, 1999. [go back]
[27] *Backes-Gellner, U. and Veen, S., 2013. Positive effects of ageing and age diversity in innovative companies–large-scale empirical evidence on company productivity. Human Resource Management Journal, 23(3), pp.279-295* [go back]
[28] *https://en.wikipedia.org/wiki/Affluence_in_the_United_States* [go back]
[29] *https://www.epi.org/blog/top-1-0-percent-reaches-highest-wages-ever-up-157-percent-since-1979/* [go back]
[30] *https://www.kaggle.com/harriken/kaggle-journey-2017-2018* [go back]
[31] *Cowan, N., 2010. The magical mystery four: How is working memory capacity limited, and why?. Current directions in psychological science, 19(1), pp.51-57.* [go back]
[32] *Wynn, T., 2002. Archaeology and cognitive evolution. Behavioral and brain sciences, 25(3), pp.389-402* [go back]
[33]. *[go back]*
[34] *Galton, F., 1886. Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, pp.246-263.* [go back]
[35] *Gandhi, P., Khanna, S. and Ramaswamy, S., 2016. Which industries are the most digital (and why). Harvard Business Review, 1.* [go back]
[36] *Stewart, P.J., 2019. Mendeleev's predictions: success and failure. Foundations of Chemistry, 21(1), pp.3-9.* [go back]
[37] Source: Berengueres, J., 2015. The Brown Book of Design Thinking: A workshop based approach. UAE University College. [go back]
[38] *Johnstone, Keith. Impro for storytellers. Routledge, 2014.* [go back]
[39] Norman, D., 2013. The design of everyday things: Revised and expanded edition. Basic books.
[40] *Berengueres, J., 2007. The Toyota production system re-contextualized. Lulu. com.* [go back]
[41] *Tversky, A. and Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. science, 185(4157), pp.1124-1131.* [go back]
[42] *Thomas, A.K. and Millar, P.R., 2011. Reducing the framing effect in older and younger adults by encouraging analytic processing. Journals of Gerontology Series B: Psychological Sciences and*

Social Sciences, 67(2), pp.139-149 *[go back]*
[43] *Sweller, J., 1988. Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), pp.257-285.* *[go back]*
[44] *Kahneman, D. and Tversky, A., 1981. On the study of statistical intuitions (No. TR-6). STANFORD UNIV CA, DEPT OF PSYCHOLOGY.* *[go back]*
[45] *Cook, J., Supran, G., Lewandowsky, S., Oreskes, N. and Maibach, E., 2019. America Misled: How the Fossil Fuel Industry Deliberately Misled Americans about Climate Change.* *[go back]*
[46] *Because the authors of the paper refused to retract their beliefs after the data showed otherwise, this is also evidence of confirmation bias.* *[go back]*
[47] *Herndon, T., Ash, M. and Pollin, R., 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Cambridge journal of economics, 38(2), pp.257-279.* *[go back]*
[48] *Pandey, A.V., Rall, K., Satterthwaite, M.L., Nov, O. and Bertini, E., 2015, April. How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 1469-1478). ACM.* *[go back]*
[49] *The National Institute for Space Research (Portuguese: Instituto Nacional de Pesquisas Espaciais, INPE). It is a research unit of the Brazilian Ministry of Science, Technology and Innovation.* *[go back]*