Vis Responsibly

Considerations for making accurate data displays

By Erica Gunn

BU Spark CS 506, 10.26.2020

Agenda

Part I: Why data vis?

Part II: What *is* data vis?

Part III: Common mistakes to avoid

Conclusions and takeaways

Part I:
Why data vis?

Origins of visualization — How data visualization, the physical and social sciences and statistical methods have co-evolved.

Visualization as record

The practice of counting underpins almost every aspect of civilization

- Count and make comparisons
- Demonstrate consensus and influence
- Understand resources
- Create strategy
- Support negotiation

Counting has social power and supports cultural and technological advancement.

Counting is Power

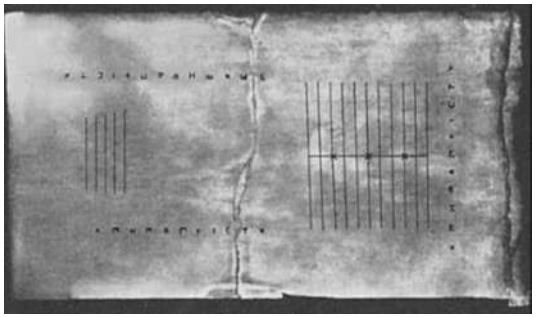
Trade

Tally sticks
41,000 BC Lebombo bone
23,000 BC Ishango bone

Votes

Counting stones – voting
500 BC Wine cup depicting the Trojan War

Taxes



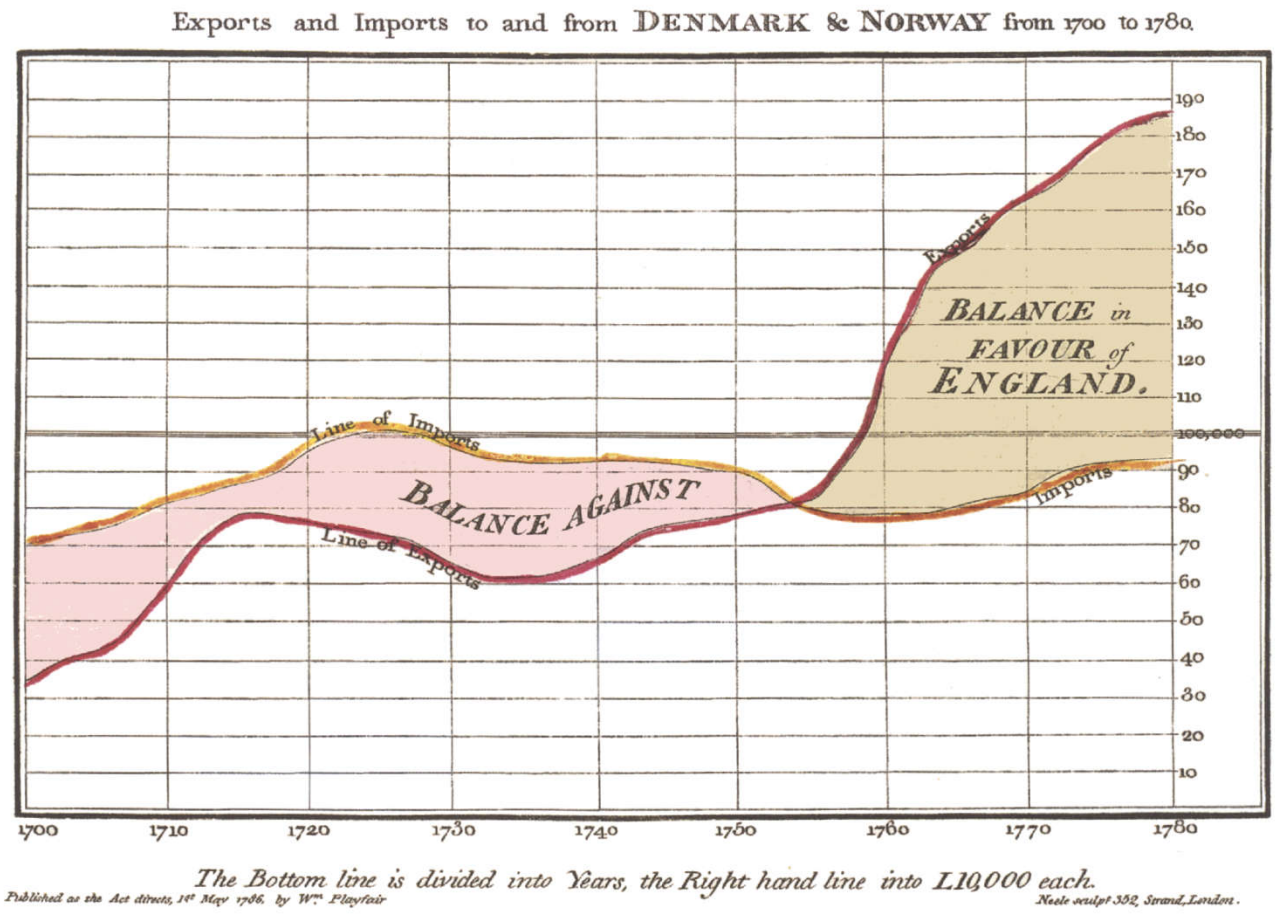
First abacus – Salamis Greece
300 BC Counting device for calculating trade deals that drove prosperity in the ancient world.

4000 BC First census held in Babylon

Money, taxes, budgets, armies, navigation, warfare, trade…all of these things depend on counting, and have driven technological change throughout human history.

Adding a national scale

"Statistics" was coined in 1750 as a word for the tabulation of numbers about the state.



Statistical graphics were first pioneered by William Playfair, who invented the bar, line, area, and pie charts.

This chart is from his Commercial and Political Atlas, published in 1786.

Managing Complexity

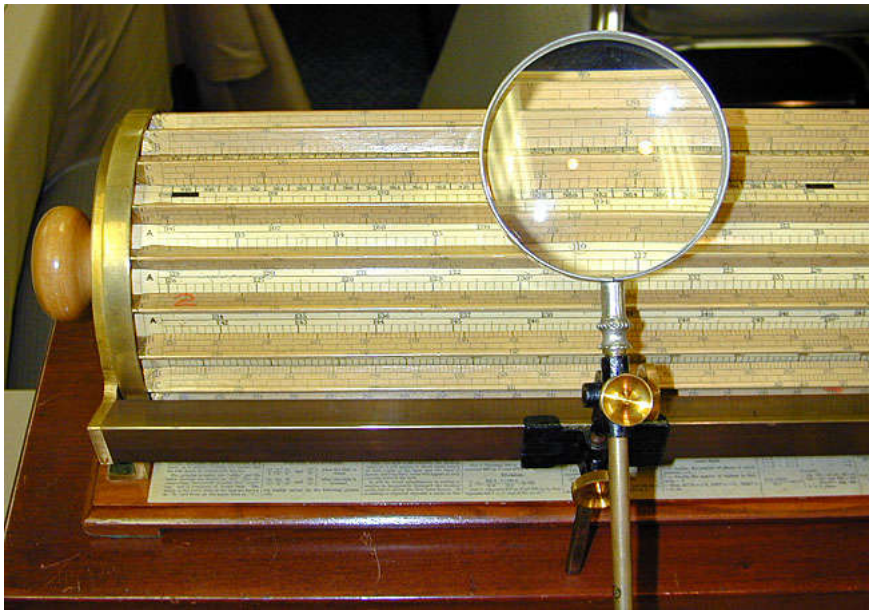
Data visualization is a method for understanding complex systems

- See patterns that weren't obvious before
- Understand complex systems and predict results
- Perform advanced calculations
- Drive innovation

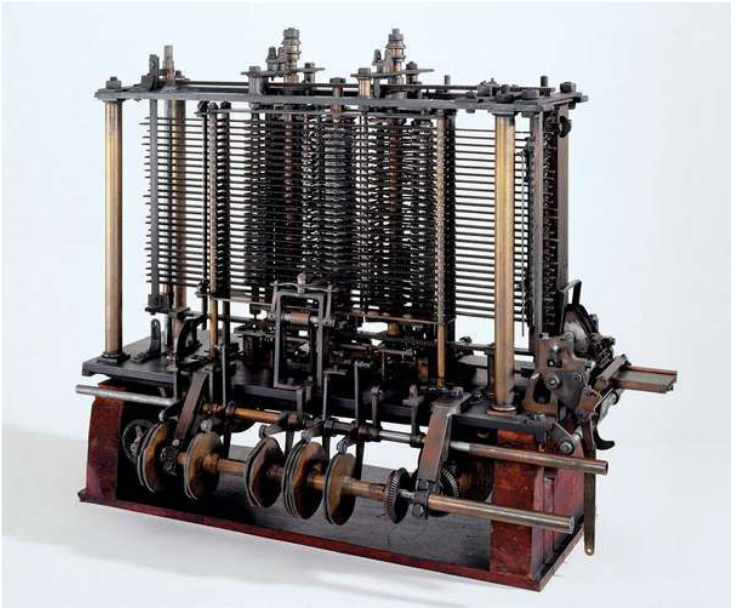
As technology and society became more sophisticated, we needed better tools to understand them.

Technology facilitates counting

Visual and mechanical inventions supported more complex calculation



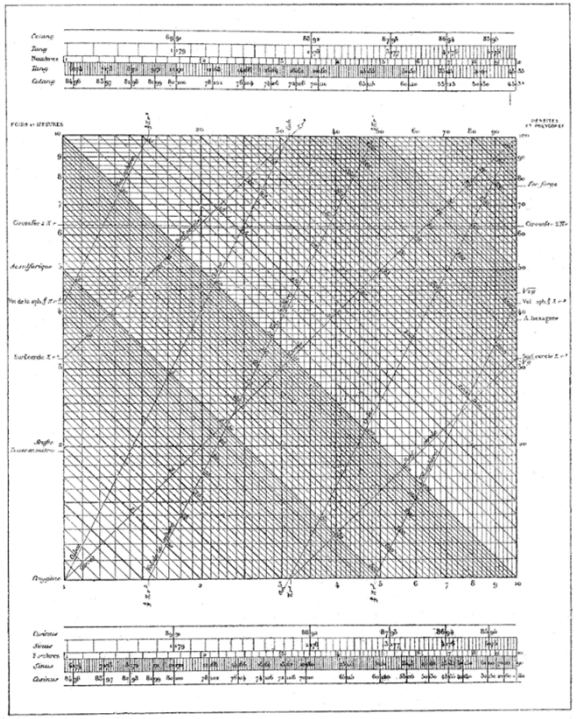
Slide Rule
1622 Advanced calculations
for structural engineering



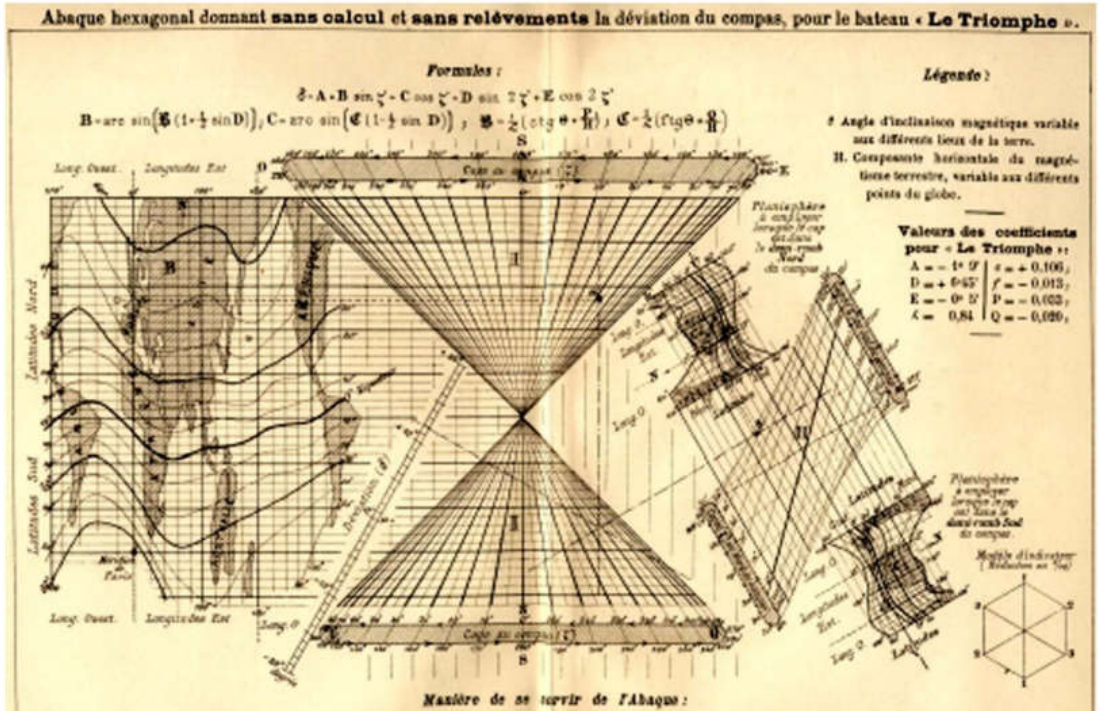
Analytical Engine
1834 Mechanical computer
designed by Charles Babbage
and Ada Lovelace.

Graphics simplify calculation

Visual forms made calculations faster, simpler, and easier to understand



“Universal computer”
1844 Leon LaLanne develops a log-log plot to calculate inverses of trigonometric functions.



“Hexagonal abacus”
1885 Nomograms are a graphical matrix method for calculating multiple variables simultaneously. This one shows compass readings for a ship’s journey.

Data Vis as Argument

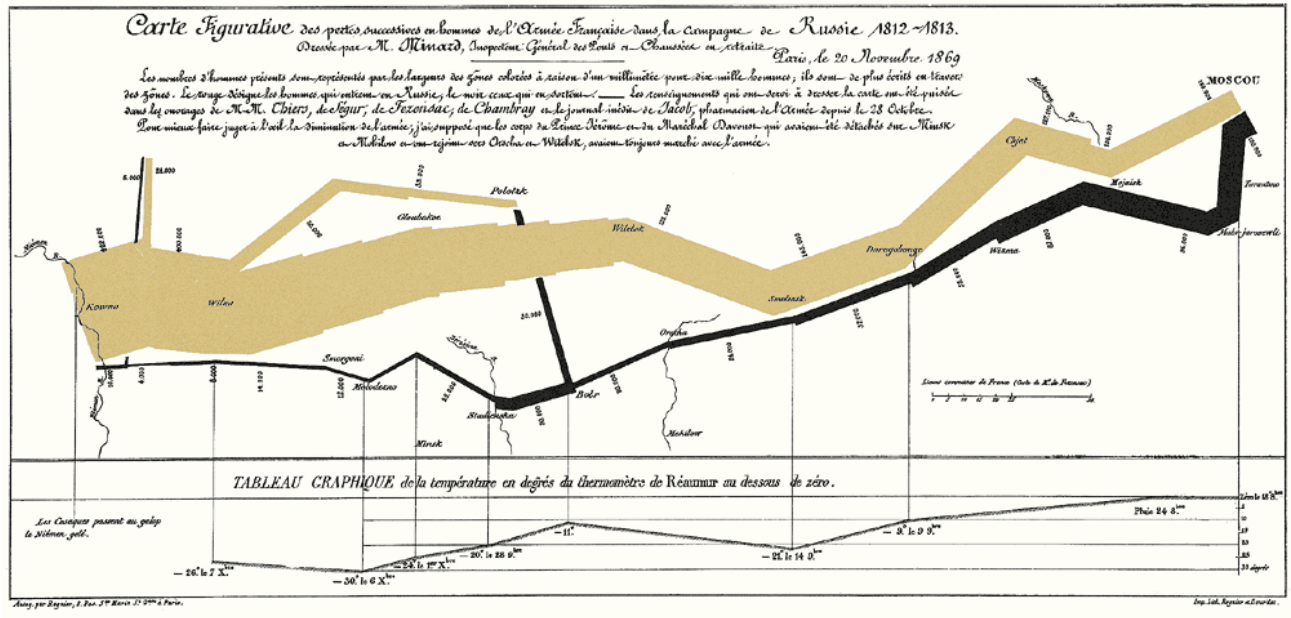
Data visualization is a method of communication

- Data as evidence
- Share a perspective or point of view
- Document large-scale patterns or trends
- Support a theory or idea

Data vis can clarify and support an argument, and played a large role in shaping social conversations.

Data Vis in Politics

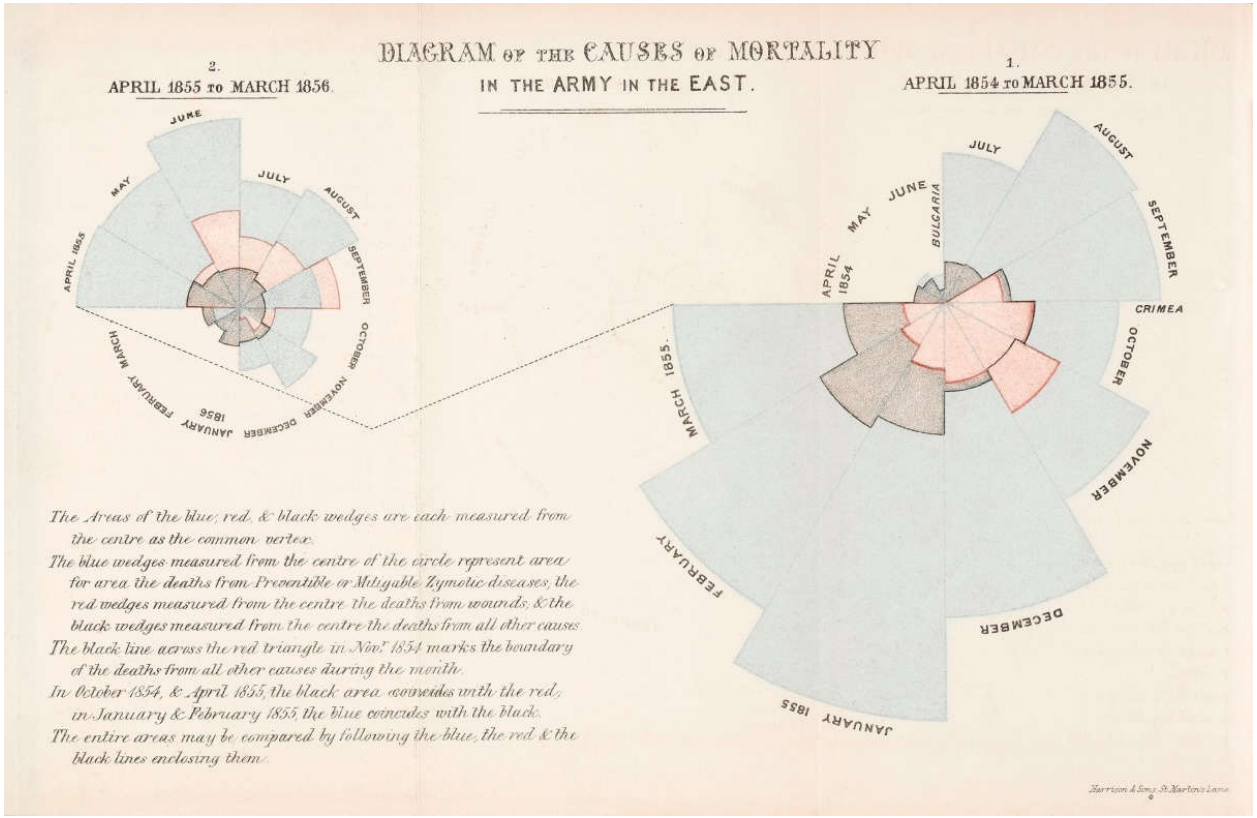
Charts and graphs were used as a rhetorical tool to support political arguments.



Russia campaign
1812 Charles Minard drew this map to represent the military losses during Napoleon's campaign to Russia. Of 442,000 men, only 10,000 returned.

Data Vis for Social Change

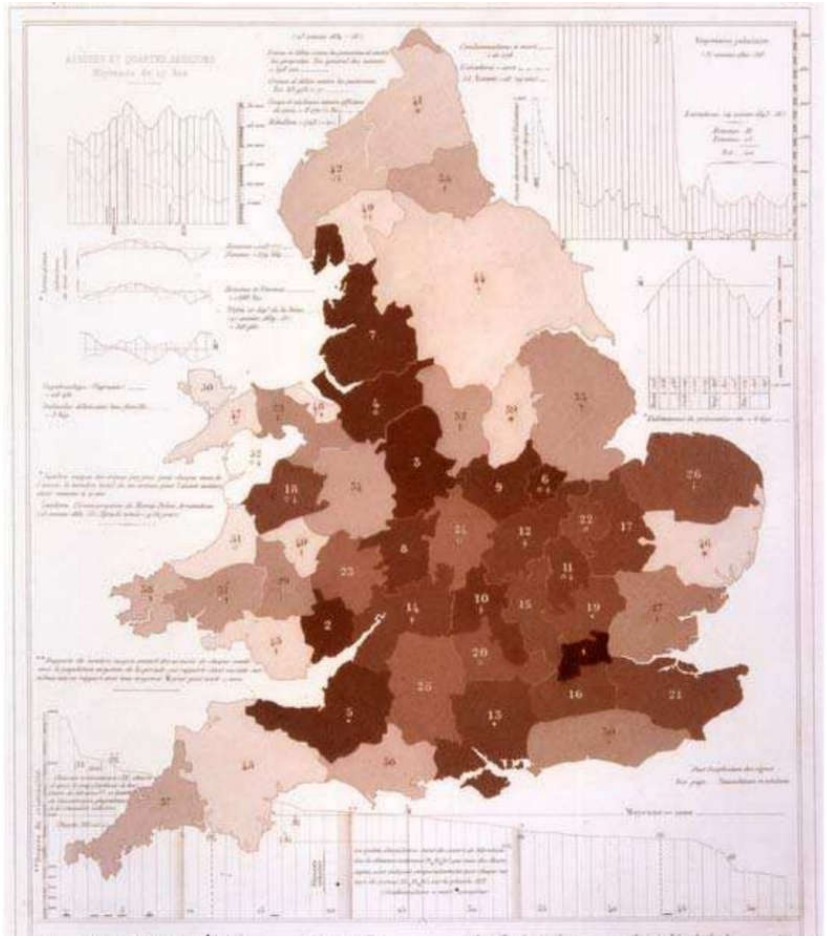
Charts supported new social initiatives as well



Causes of Mortality
1850 Florence Nightingale used visualizations of mortality in soldiers returning from the Crimean war to argue for better medical care. She is widely credited with founding the profession of nursing.

Foundations of Political Science

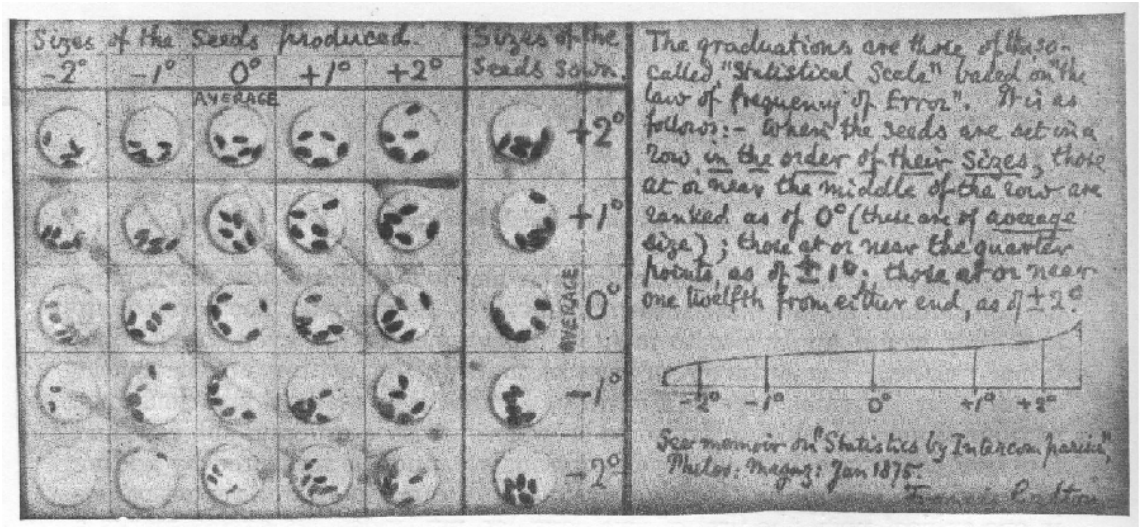
Growing out of statistics, “political arithmetic” became a way of trying to address social problems.



Plot of crime in England
1864 Francis Guerry began mapping out statistical data to understand crime and other social problems, founding the discipline of social science.

Getting Smarter About Data

More sophisticated statistical methods were developed to validate and measure the quality of relationships by mathematical means



Regression
1875 Francis Galton described regression to the mean while interpreting his work on genetic variation in sweet peas.

1888 Galton invented correlation as a way of checking the results of proportional scaling used in visualizations and analysis at the time.

Data visualization took on a secondary role in supporting an argument, rather than being the argument itself.

"Pictures of data became considered—well, just pictures: pretty or evocative perhaps, but incapable of stating a 'fact' to three or more decimals. At least it began to seem this way to many statisticians and practitioners."

Michael Friendly, Golden Age of Statistical Graphics

Understanding Significance

Sure, that vis can help you see something. But is it meaningful?

CRYSTAL GROWTH & DESIGN

ARTICLE

pubs.acs.org/crystal

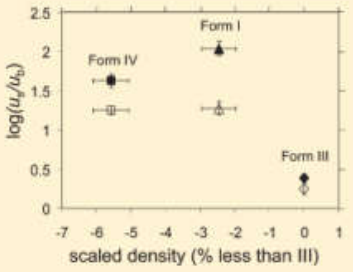
Does Crystal Density Control Fast Surface Crystal Growth in Glasses? A Study with Polymorphs

Published as part of a virtual special issue of selected papers presented at the 2010 Annual Conference of the British Association for Crystal Growth (BACG), Manchester, UK, September 5−7, 2010

Erica M. Gunn, Ilia A. Guzei, and Lian Yu*

School of Pharmacy and Department of Chemistry, University of Wisconsin - Madison, Madison, Wisconsin 53705, United States

ABSTRACT: As organic liquids are cooled to become glasses, crystal growth at the free surface can be substantially faster than in the interior, a phenomenon uncommon for other materials and for which different explanations exist. We have measured the surface and bulk growth rates of three polymorphs in carbamazepine glasses. Crystal density has no controlling effect on the extent to which surface crystal growth is enhanced over bulk crystal growth, in contradiction to models that relate fast surface crystal growth to the release of crystallization-induced tension.



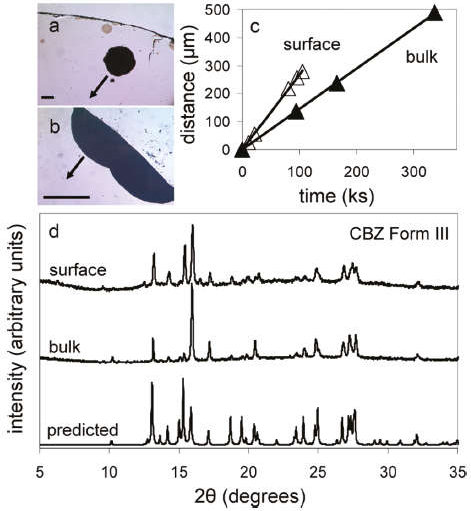


Figure 3. Form III crystals growing (a) at the surface and (b) in the bulk at 313 K. Scale bar = 200 μm. Arrows indicate growth directions. (c) Distance of growth vs time for crystals in (a) and (b). (d) Observed and predicted XRD patterns of Form III crystals.

Without statistical analysis, there is no way to tell if the relationships that you see are real. In science, seeing is never be sufficient for believing.

Data Vis & Mathematical Analysis

Visualizations and statistics go hand in hand. You need a balance of the two techniques for different tasks, but they can never stand alone.

Data Vis

- Communication
- Pattern finding
- Exploration
- Methods development
- Uncovering details hidden by the analysis

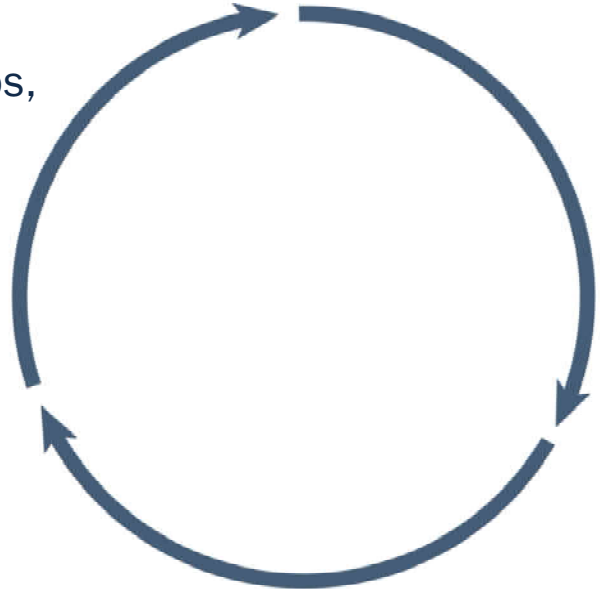
Statistics

- Quality
- Rigor/Significance
- Validation of systematic methods
- Processing efficiency
- Analytical science

Vis & Analysis Cycle

You need different things at different stages of the discovery process

Communicating results
Demonstrating relationships, explaining connections, sharing insight



Developing methods
Exploring patterns, finding anomalies, identifying relationships, building a hypothesis

Applying methods
Consistency, accuracy, reliability. Interpreting results. Analytical rigor and repeatable methods.

Big Data Revolution

Data has become the currency of our lives. It is more and more necessary to translate data for non-technical audiences to understand.

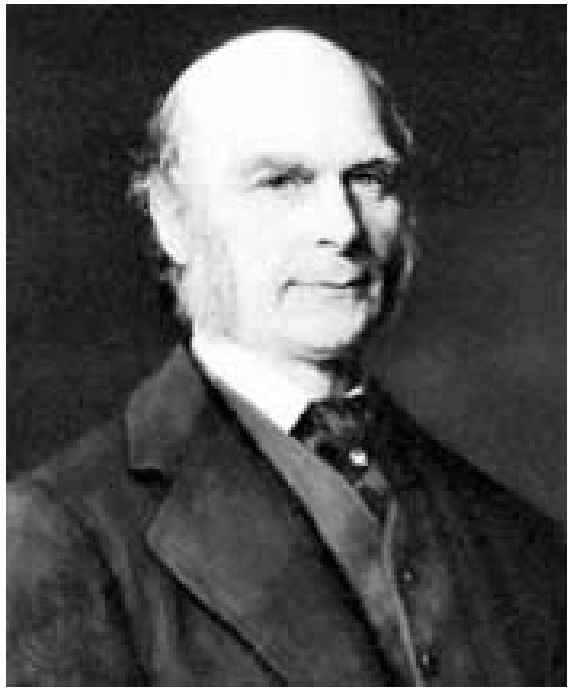
Just because you plot it, doesn't mean it's real.

Just because it's real doesn't mean that you know what it means.

We need new methods of analysis, and new mechanisms to extract and validate data, to understand what our eyes are telling us when we look at a chart or graph.

Data vs. Interpretation

Data is meaningless without analysis, and interpretation is always human. The same mind who gave us statistics also gave us eugenics.



"This is precisely the aim of Eugenics. Its first object is to check the birth-rate of the Unfit, instead of allowing them to come into being, though doomed in large numbers to perish prematurely. The second object is the improvement of the race by furthering the productivity of the Fit by early marriages and healthful rearing of their children. Natural Selection rests upon excessive production and wholesale destruction; Eugenics on bringing no more individuals into the world than can be properly cared for, and those only of the best stock."

— *Francis Galton,*
Memories of My Life, 1908

As data practitioners, it is *always* our primary responsibility to ensure that data interpretation and the analysis is valid.

Part II:

What *is* data visualization?

What is data visualization?

- Makes abstract data visible
- Encodes information
- Converts data channels (variables) into visual form
- Uses marks to represent the data values
- Supports a user task

From Data To Chart

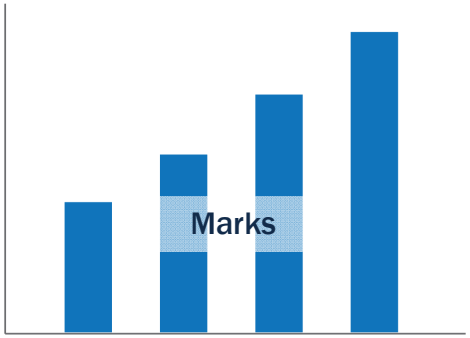
Objects or Observations

Data (measurement)

	A	B
1	Age (y)	Height (in)
2	4	40
3	8	51
4	12	58
5	16	67.5
6		

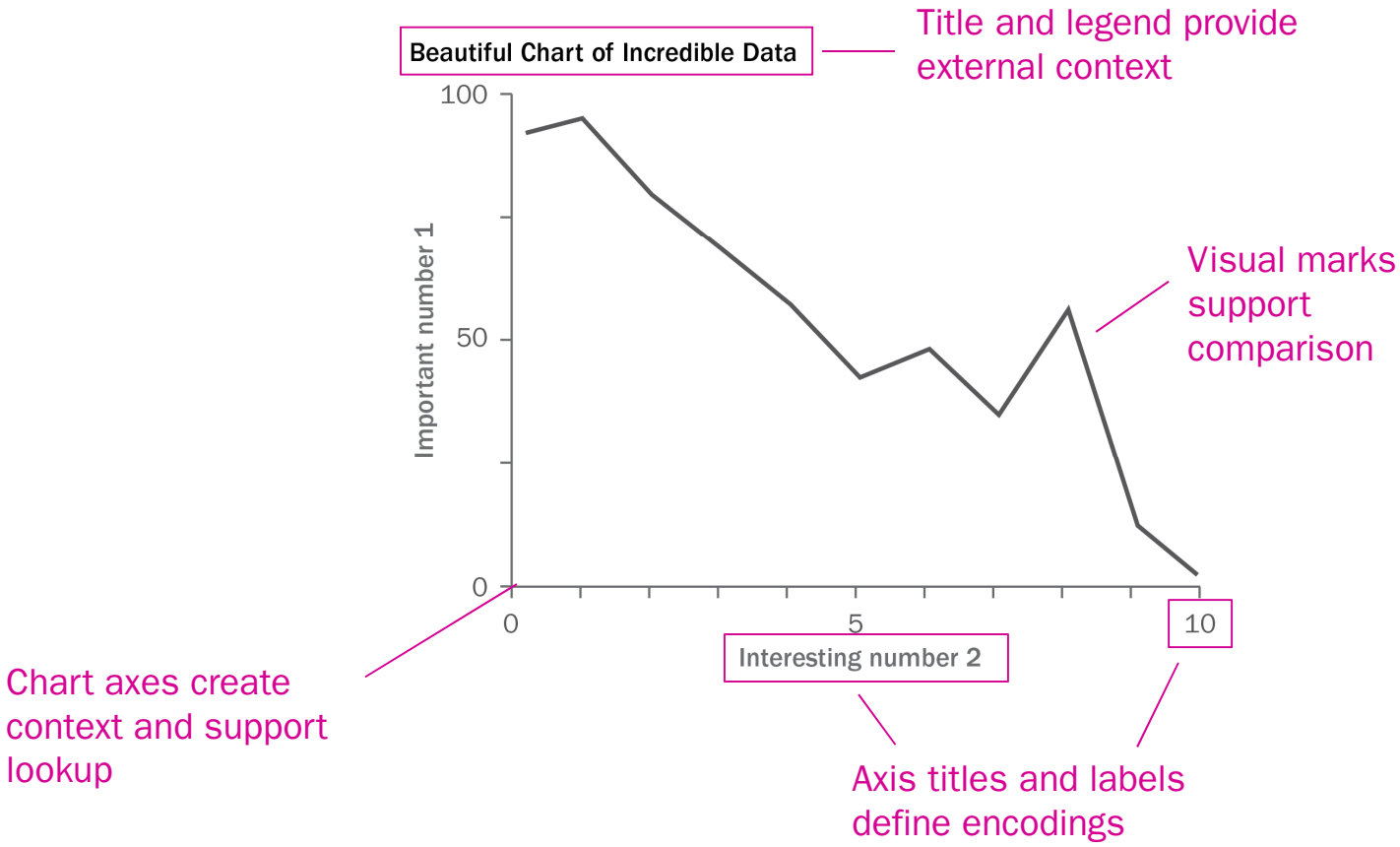
Channel and Encoding

Chart



Anatomy of a Chart

Every part of a chart has a job to do



Part II:
Matching Chart to Task

- Best practices
- Connecting chart to task
- Limitations of visualization

Things to ask yourself:

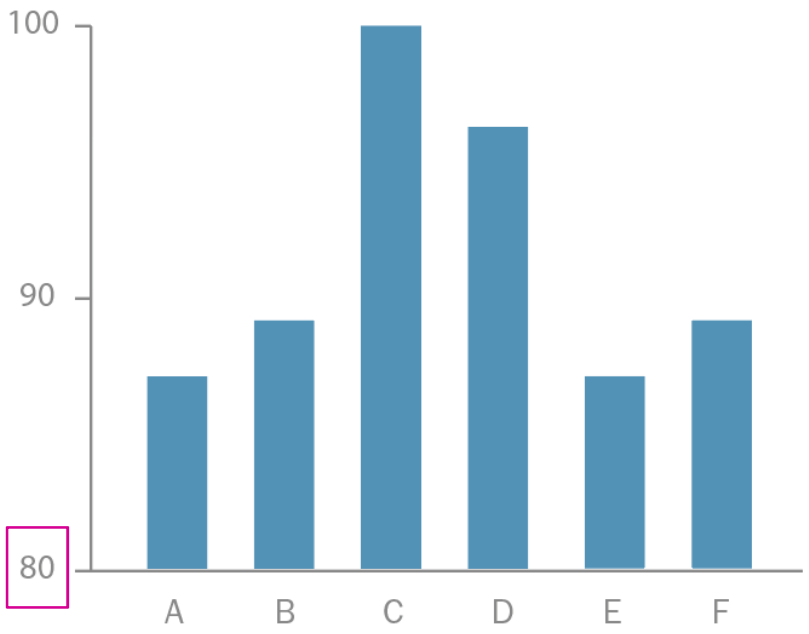
- What's your purpose?
- Who is it for?
- What are you trying to show?
- What do people need to see to understand?
- What makes sense for your data?
- Which chart supports the user task?
- How can you use design principles to clarify your representation?

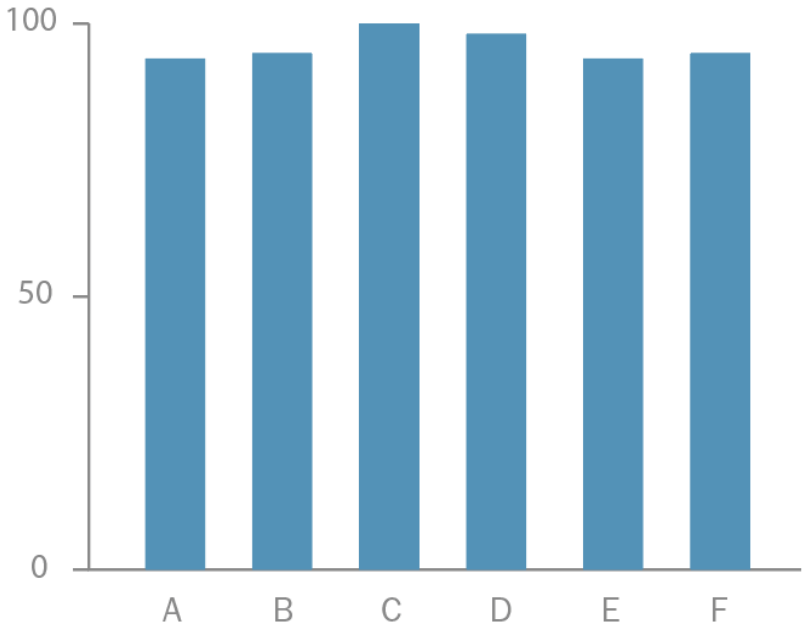
Data visualizations should:

1. Accurately represent information,
   without distortion or undue emphasis.

Nonzero Y Values

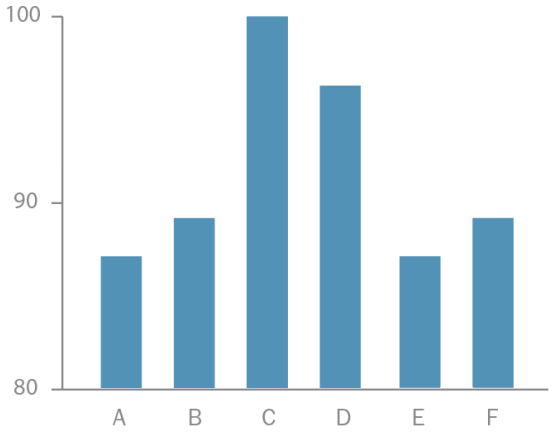
Bar chart axes should always start at zero

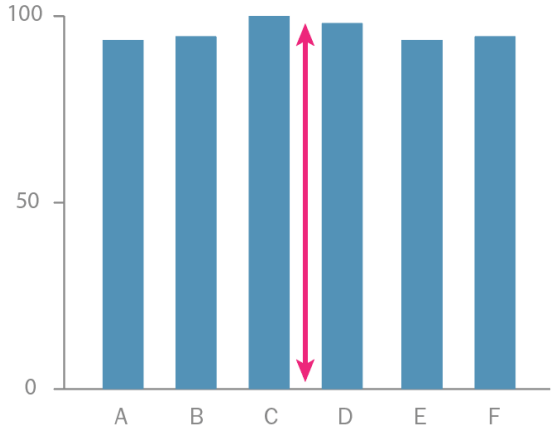


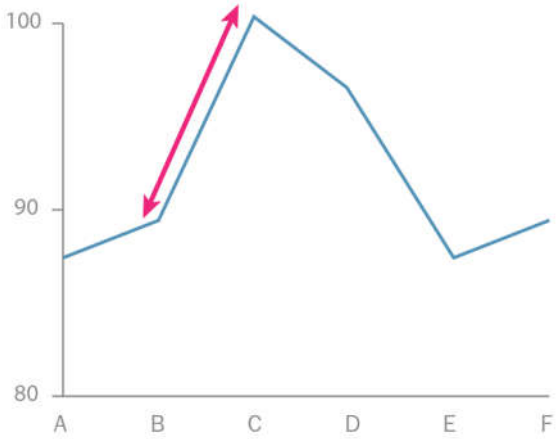


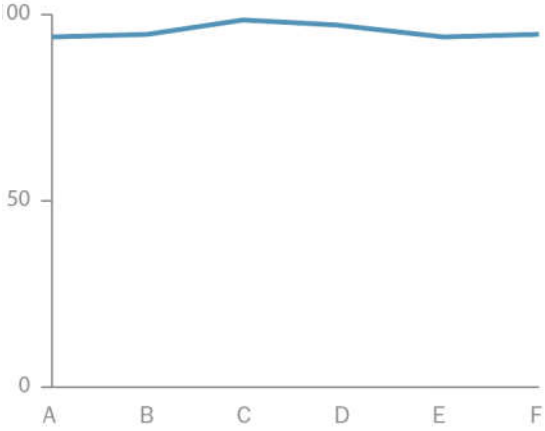
Marks Matter

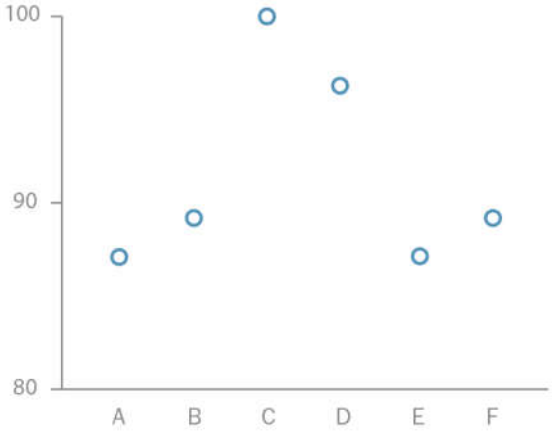
Best practices optimize for what the user needs to read the chart





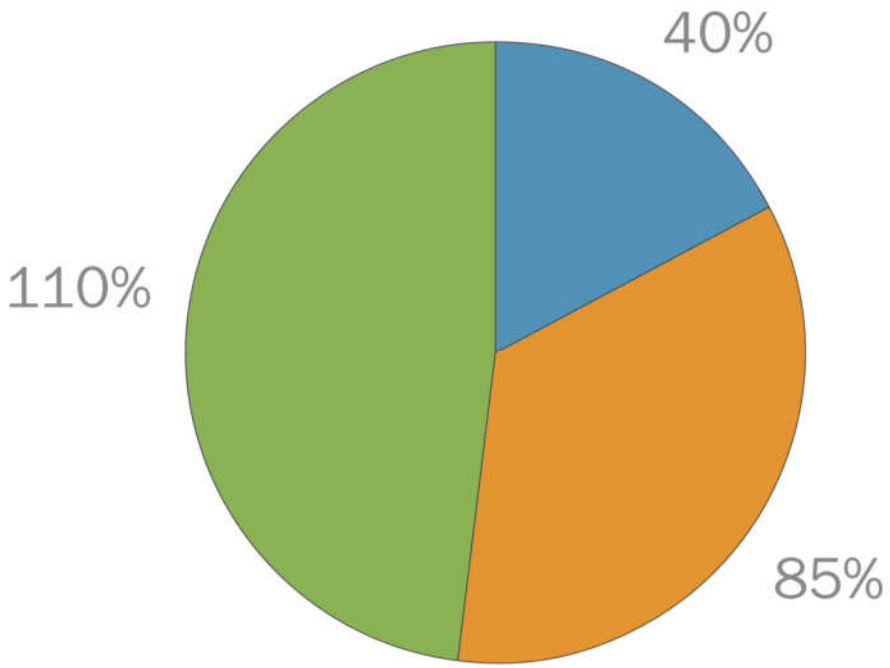






Percentage Scaling

Percentages in a pie chart must always add up to 100

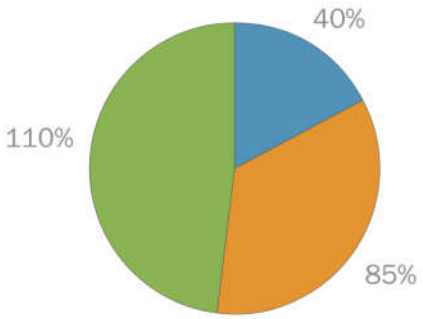


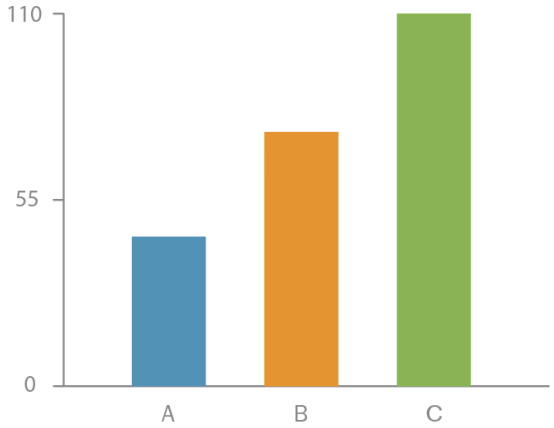
The number that I read should not contradict what my eyes see!

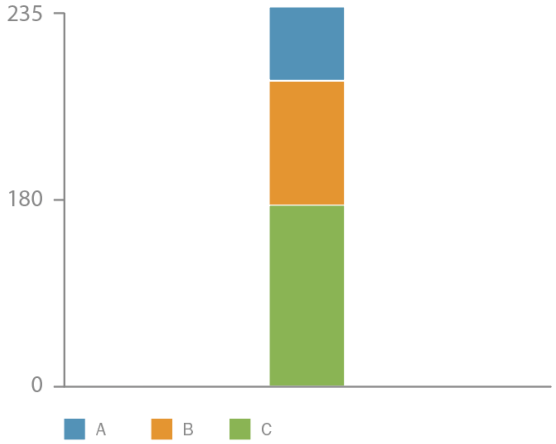
1 pie = 100% Values = 235% ???

Percentage Scaling Improved

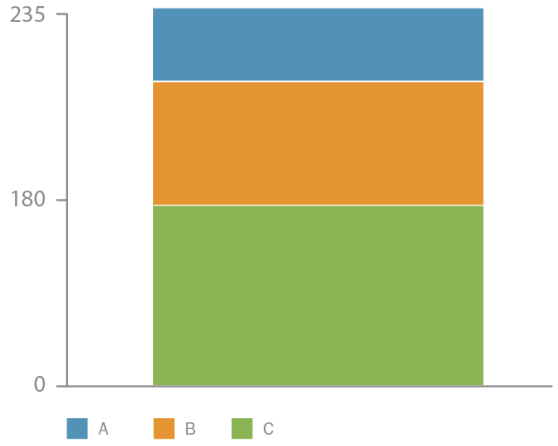
Bars don't add up to a whole, so they can represent more than 100%







Be careful with style choices; even small changes can distort a user's perception of the chart data.



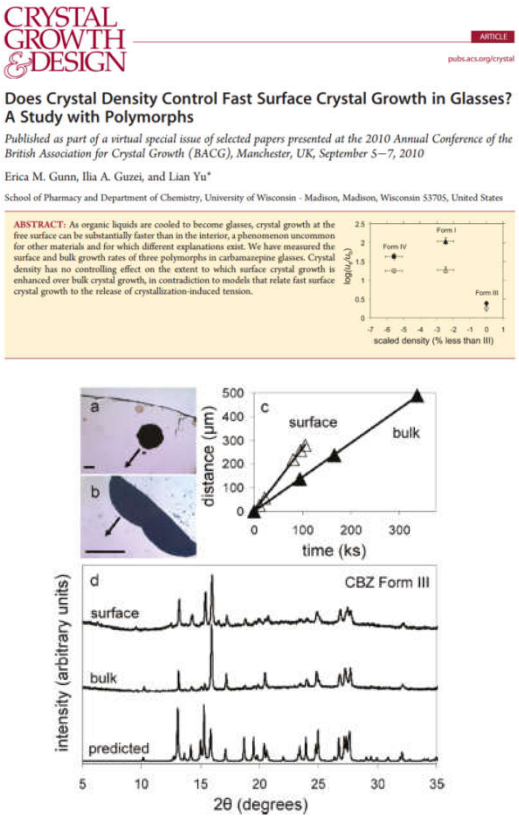
2. Have a purpose.

Audience: Who am I talking to?

Context: What kind of information do they need?

Know Your Audience

A chart's purpose dictates appropriate design choices



CRYSTAL GROWTH & DESIGN

ARTICLE

pubs.acs.org/crystal

Does Crystal Density Control Fast Surface Crystal Growth in Glasses? A Study with Polymorphs

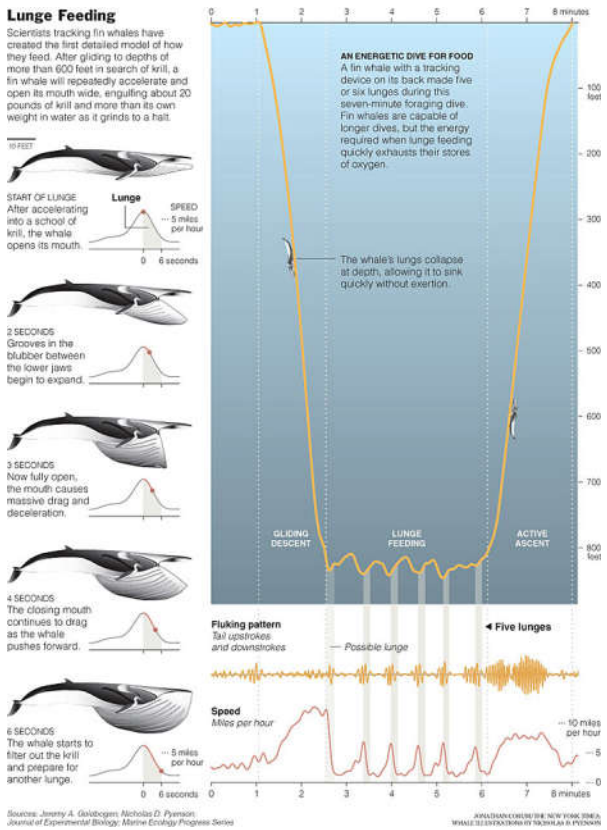
Published as part of a virtual special issue of selected papers presented at the 2010 Annual Conference of the British Association for Crystal Growth (BACG), Manchester, UK, September 5−7, 2010

Erica M. Gunn, Ilia A. Guzei, and Lian Yu*

School of Pharmacy and Department of Chemistry, University of Wisconsin - Madison, Madison, Wisconsin 53705, United States

ABSTRACT: As organic liquids are cooled to become glasses, crystal growth at the free surface can be substantially faster than in the interior, a phenomenon uncommon for other materials and for which different explanations exist. We have measured the surface and bulk growth rates of three polymorphs in carbamazepine glasses. Crystal density has no controlling effect on the extent to which surface crystal growth is enhanced over bulk crystal growth, in contradiction to models that relate fast surface crystal growth to the release of crystallization-induced tension.

Figure 3. Form III crystals growing (a) at the surface and (b) in the bulk at 313 K. Scale bar = 200 μm. Arrows indicate growth directions. (c) Distance of growth vs time for crystals in (a) and (b). (d) Observed and predicted XRD patterns of Form III crystals.



- Tone
- Sequence
- Level of complexity
- Quantitative accuracy
- Compactness
- Level of abstraction
- Annotations
- Details/emphasis
- User controls

https://www.nytimes.com/2007/12/11/science/11gulp.html

Know Your Message

What kind of a point are you trying to make?

Jill Pelto (www.jillpelto.com)

3. Have a perspective.

Who measured the data?

What perspective does it reflect?

Who is represented, how, and why?

Data is *Always* Socially Situated

Georgia Lupi: http://giorgialupi.com/bruises-the-data-we-dont-see/
Johanna Drucker: Graphesis. Visual Forms of Knowledge Production

4. Help people understand something about the data that they might not otherwise have seen.

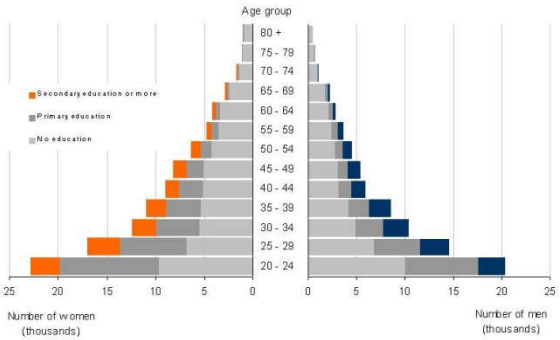
5. Support a user task.

What does your user need to do, or see?

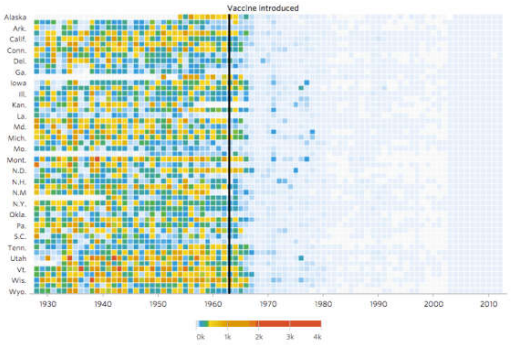
Different Charts for Different Tasks

Charts can support different tasks

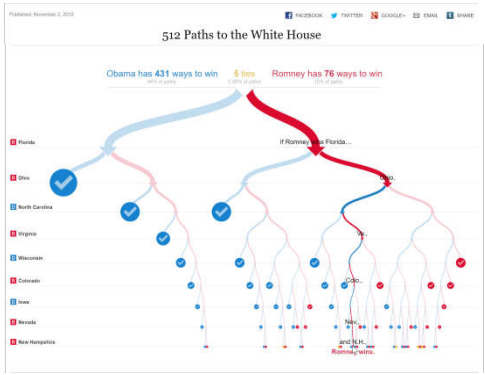
Compare objects side by side



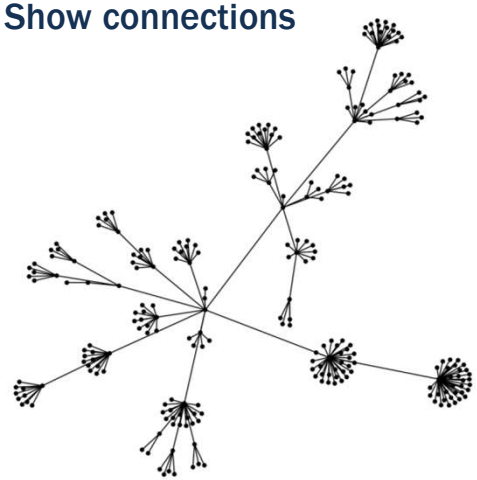
Group things together



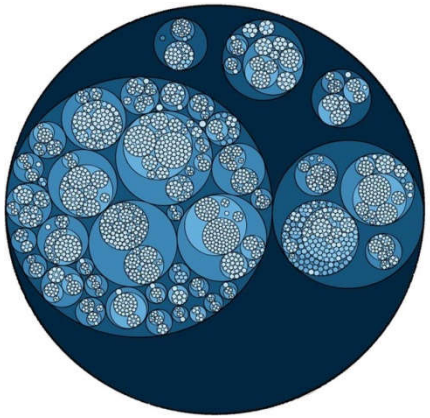
Narrate a sequence of events



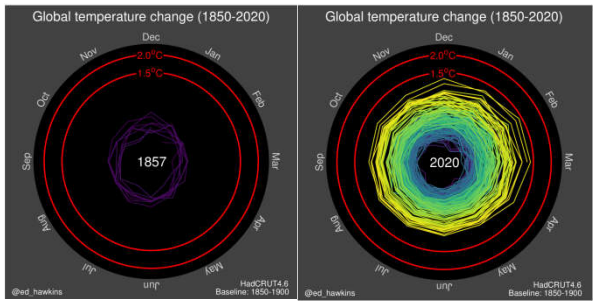
Show connections



Show membership



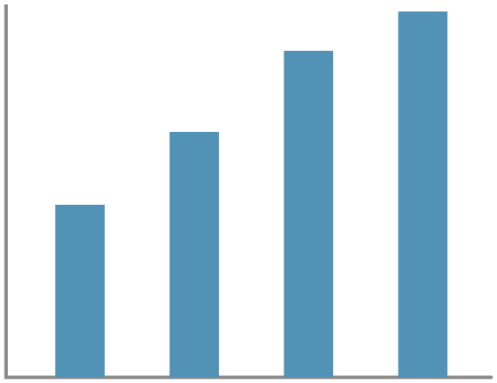
Explain how things change



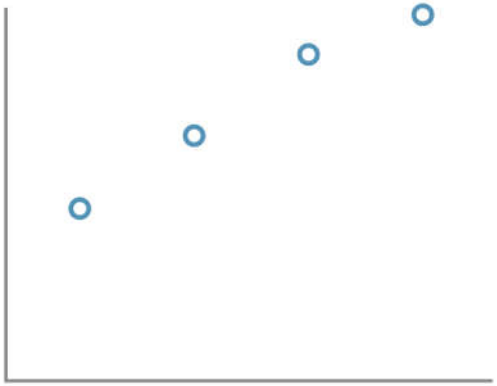
Marks Matter

How you draw the data affects what you see

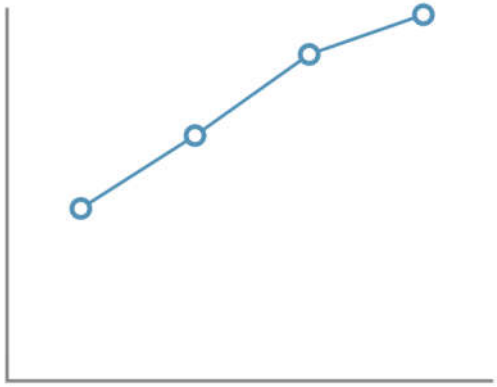
Compare bar height



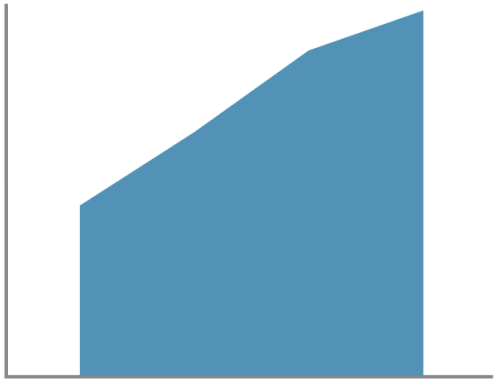
Read dot values



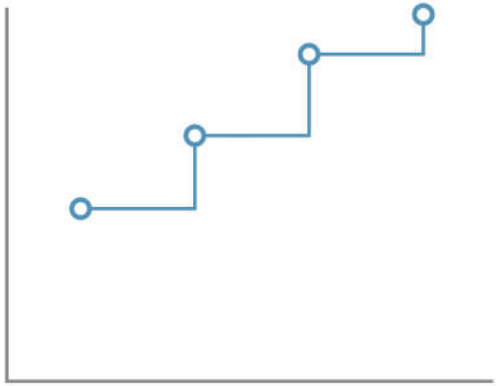
See change btwn points



Focus on area under curve



See size of change



Mistakes to Avoid

Unsuitable encodings

Calculation distortions

Abstraction and representation errors

Aggregation and denominator errors

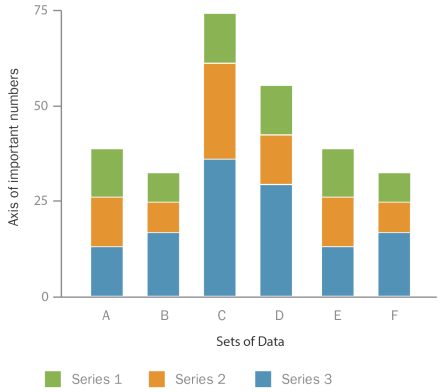
Unsuitable encodings

Encodings have different strengths and weaknesses. The method for representing the data should reflect the purpose of the chart.

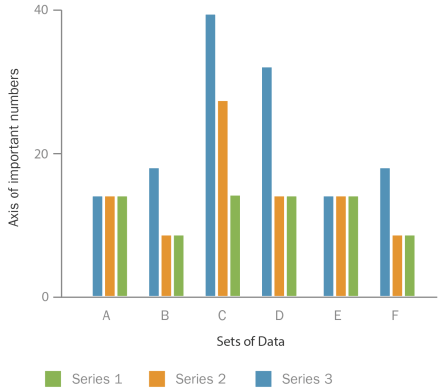
Charts Facilitate Comparison

Choosing the right encoding helps the user to understand your data.

Stacked bar chart

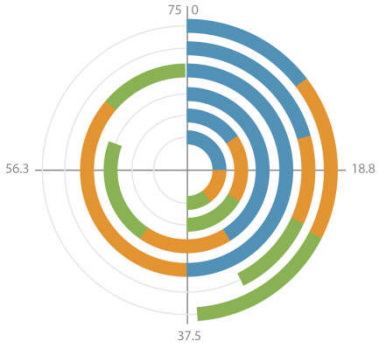


Grouped bar chart

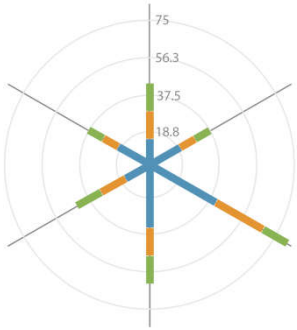


Small multiples bar charts

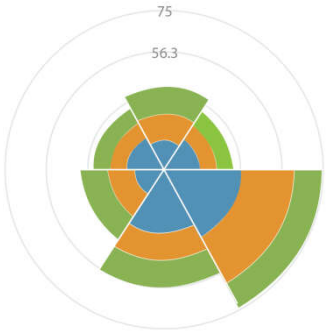
Racetrack chart



Radial bar chart



Rose diagram (length)

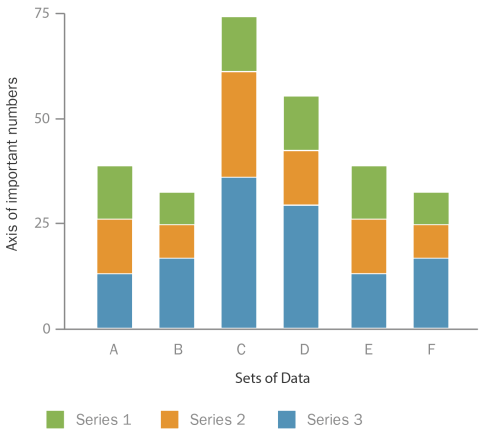


Rose diagram (area)

Assess Fitness for Task

Use task information to evaluate the different chart options

Stacked bar chart



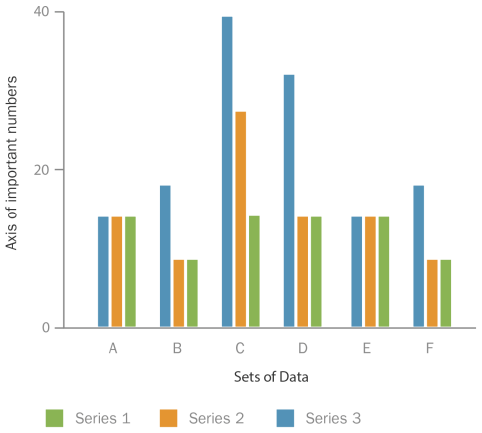
Measure Values

Compare Heights

- of bars
- of stacks
- of series

Compare proportions

Grouped bar chart



Measure Values

Compare Heights

- of bars
- of stacks
- of series

Compare proportions

Small multiples bar charts

Measure Values

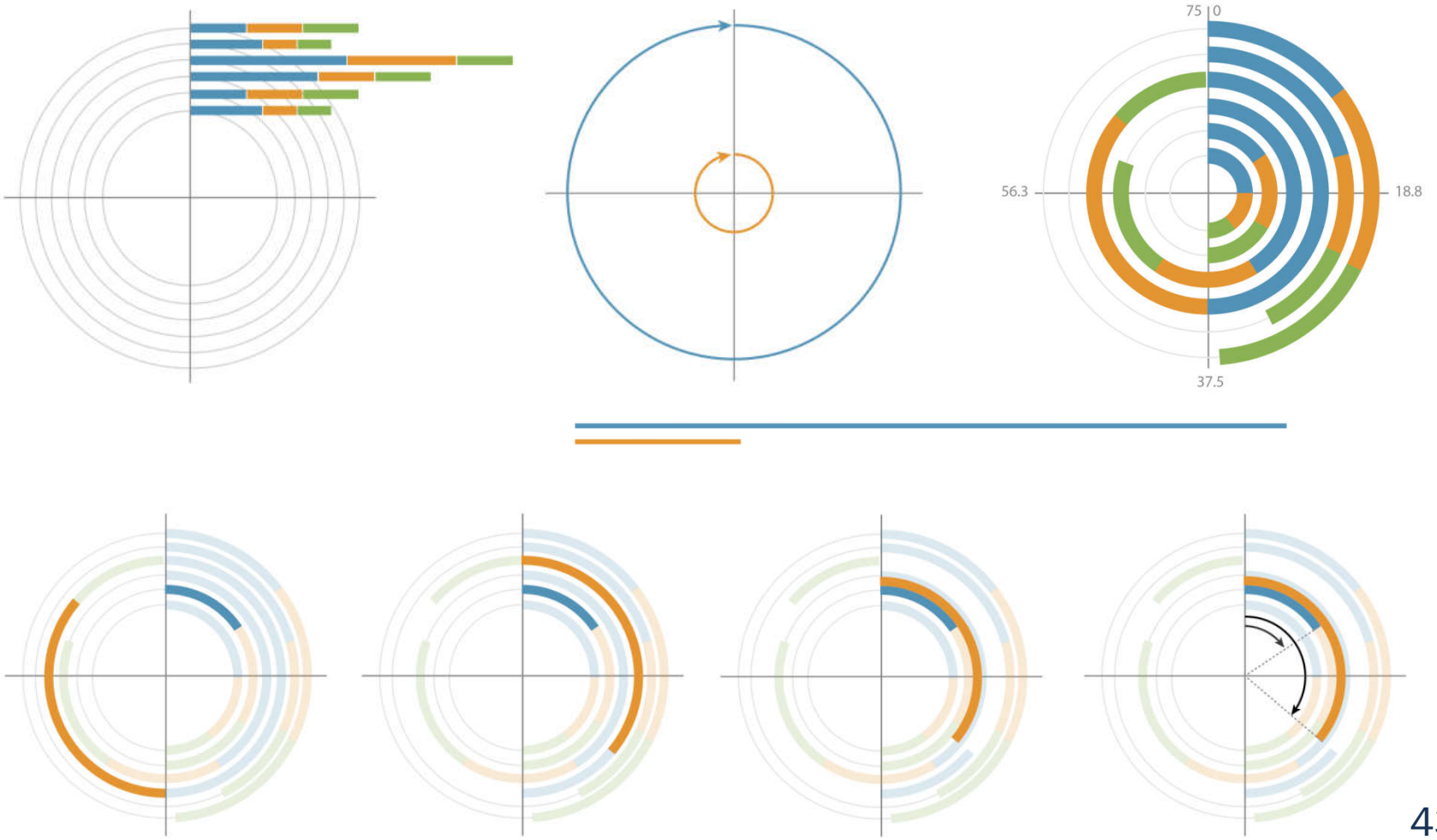
Compare Heights

- of bars
- of stacks
- of series

Compare proportions

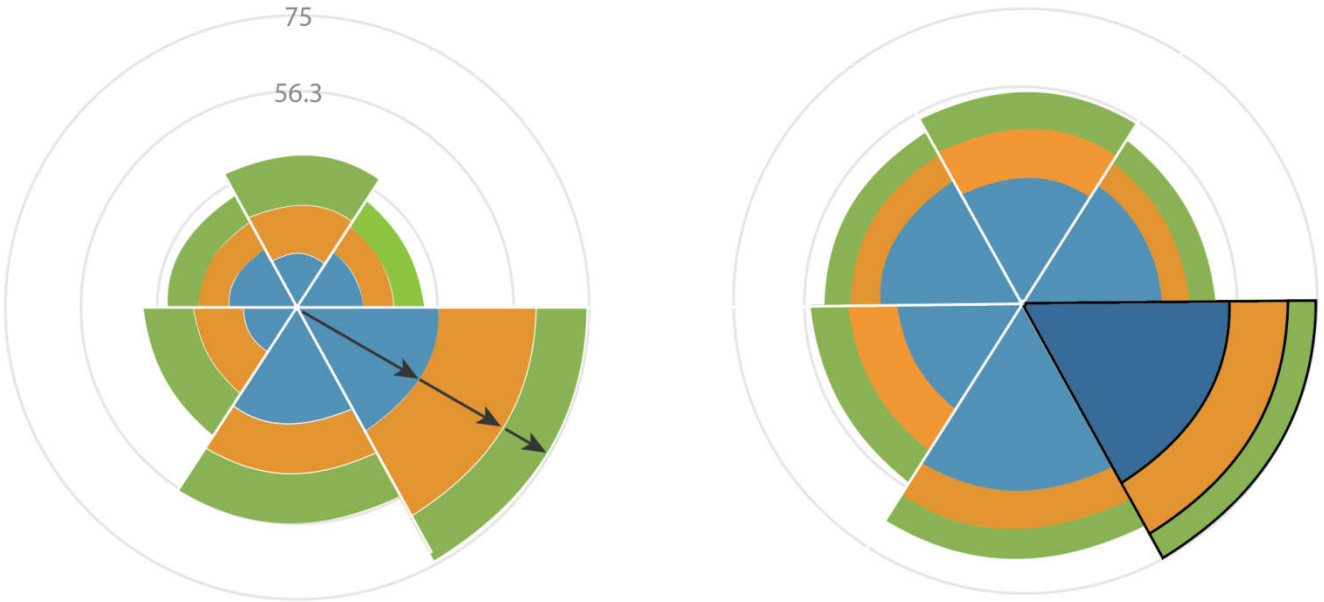
Spotlight: Racetrack Chart

The racetrack chart is a bar chart stretched around a circle



Spotlight: Rose Diagram

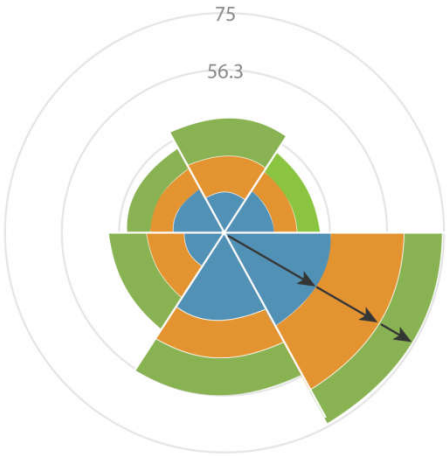
The rose diagram can be drawn as a stretched radial bar chart (length encoding), or as a stacked pie chart (area encoding)

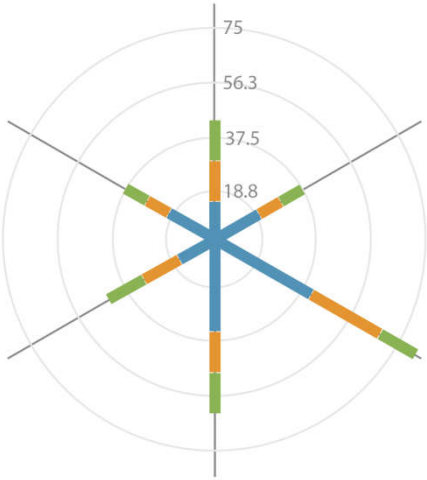


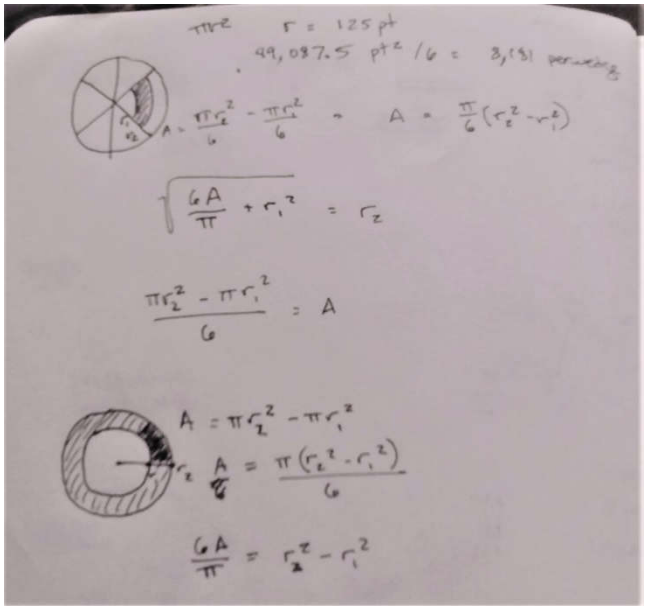
It's hard for a user to tell how to interpret the graph: they have to guess which method you used.

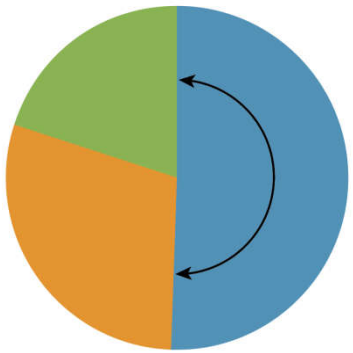
Spotlight: Rose Diagram

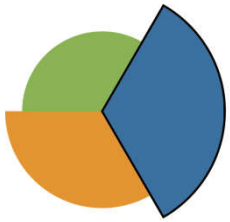
At first glance, the rose diagram looks like a radial bar, but it's actually much more closely related to the pie chart





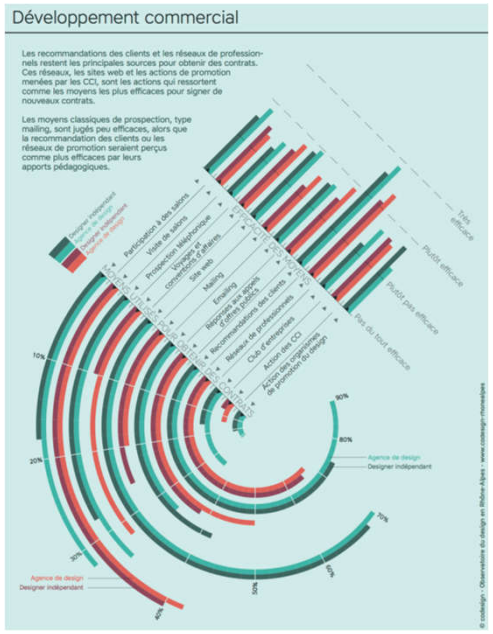




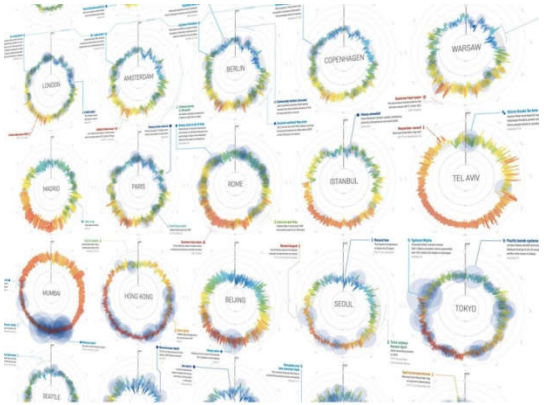


Know your purpose

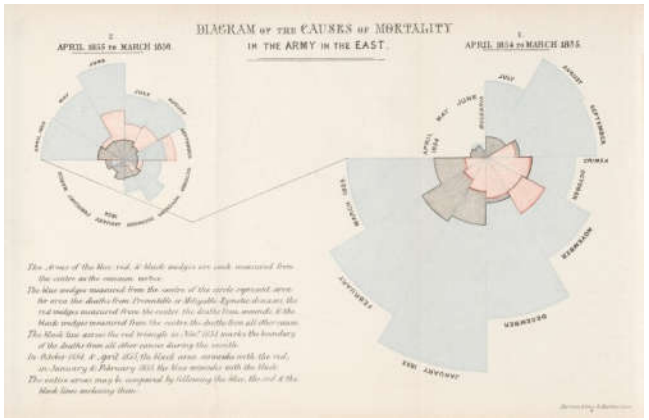
The “right” encoding ultimately depends on your purpose for the chart



http://www.corp-lab.com/codesign/codesign.pdf



http://www.weather-radials.com/



https://edspace.american.edu/visualwar/nightingale/

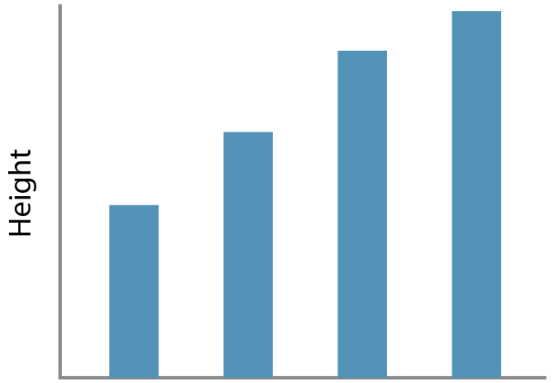
Understand the strengths and weaknesses of your charts, and make sure that they show an appropriate picture of the data.

Context distortions

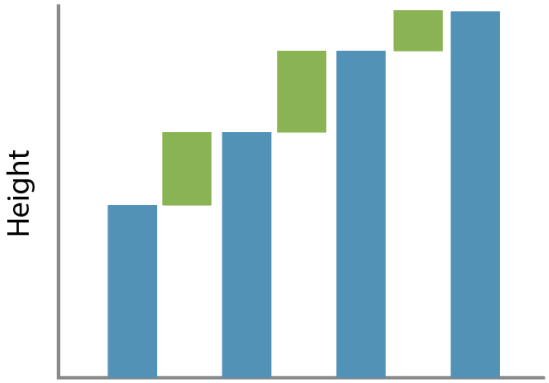
Sometimes, the way we calculate the data distorts what users can see.

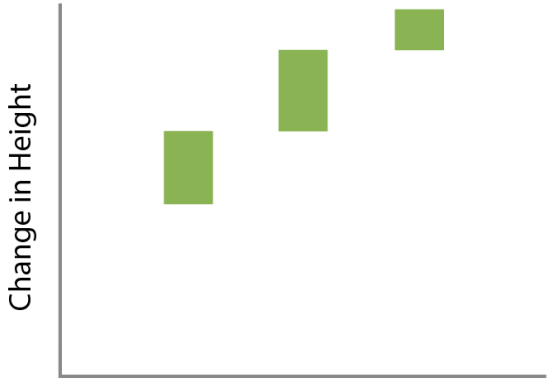
Change the Channel

Redefining a channel can emphasize different aspects of the data



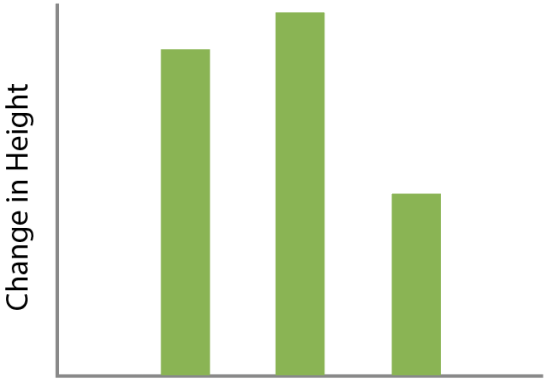
How tall is Bob?





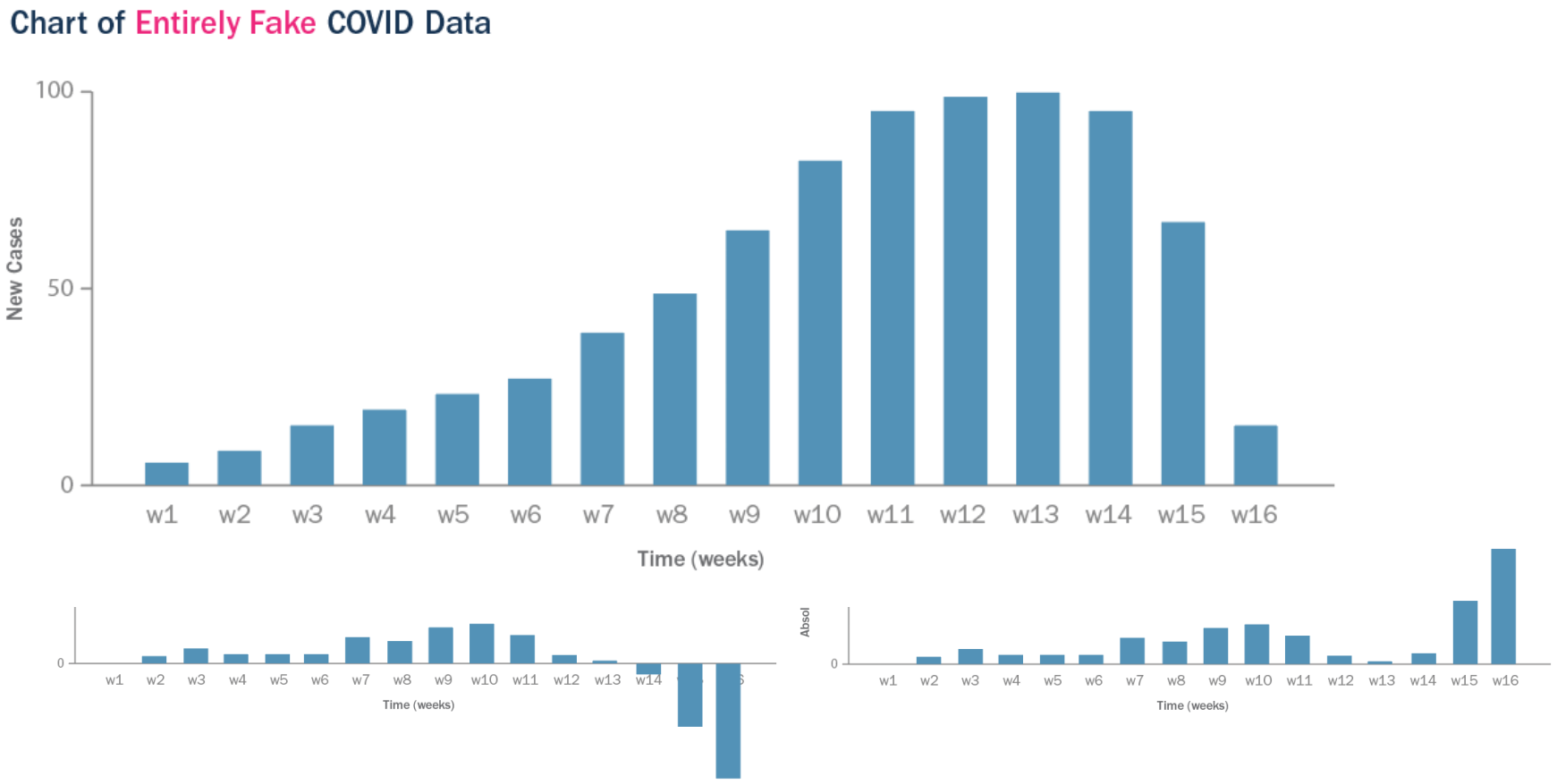
How much has Bob grown this year?

Be careful with derived statistics; they can add unintended emphasis



Case Study: (Fake) COVID data

Derivative data can hide context and change conclusions

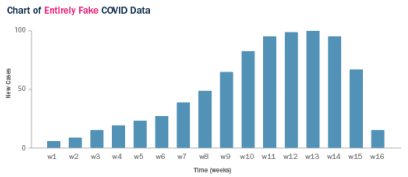


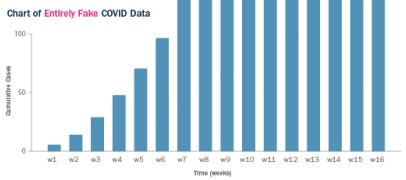
Beware the Delta

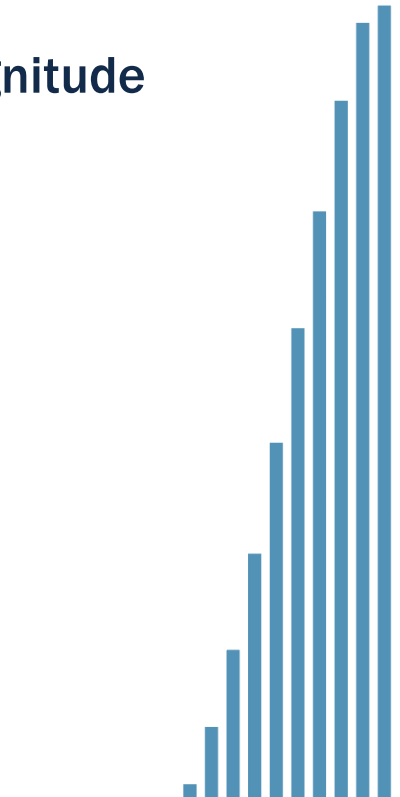
Looking only at change statistics can mask the magnitude of the underlying data

New cases *is itself* a "delta" statistic

The total number of cases keeps going up, even when the change in cases is going down.







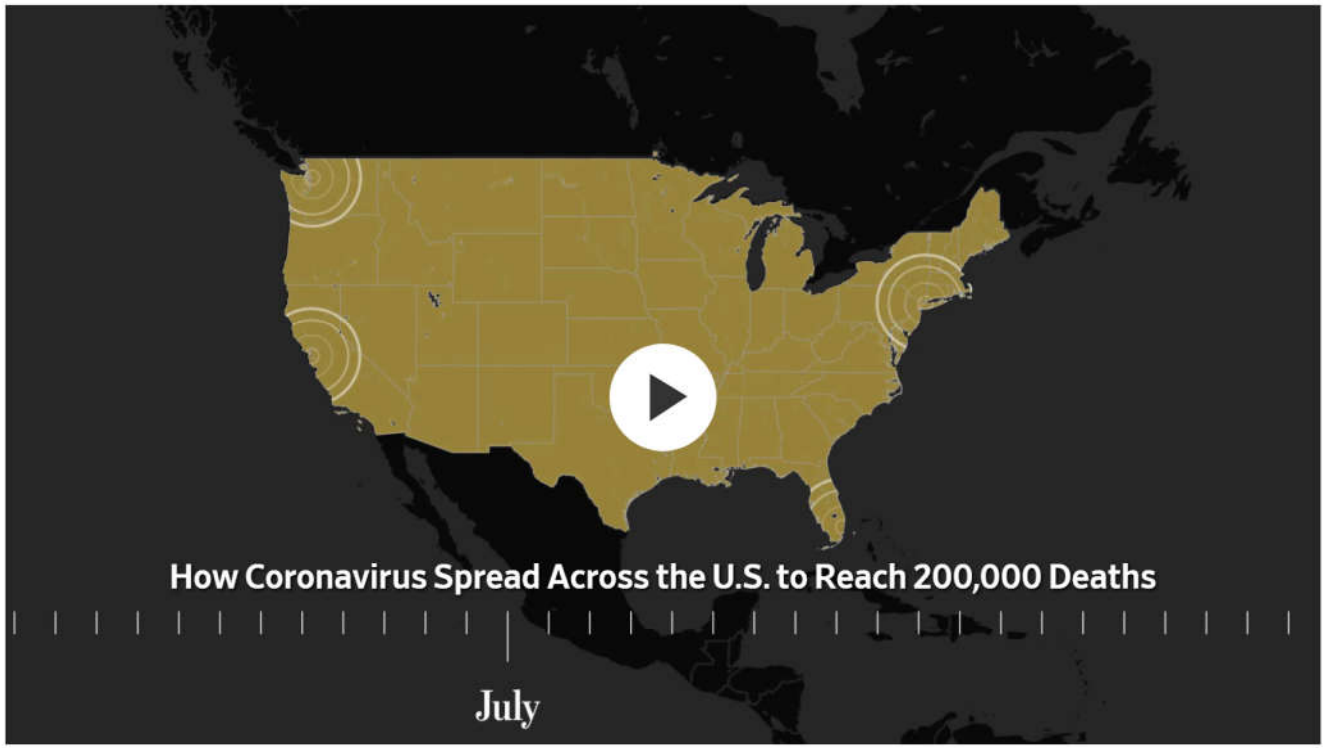
Check your calculations for unintended consequences, and provide additional context where necessary.

THE WALL STREET JOURNAL.

U.S.

New U.S. Covid-19 Cases Top 80,000 to Reach a Single-Day Record

Cases spreading in remote areas as well as cities that have already battled virus; hospitalizations also rising



As the number of U.S. coronavirus deaths surpassed 200,000, public-health experts point to a series of missteps and miscalculations in the country's response. Here is a look back at how the U.S. became the center of the global pandemic. Photo Illustration: Carter McCall/WSJ

By Jennifer Calfas and Sarah Krouse
Updated Oct. 24, 2020 8:09 am ET

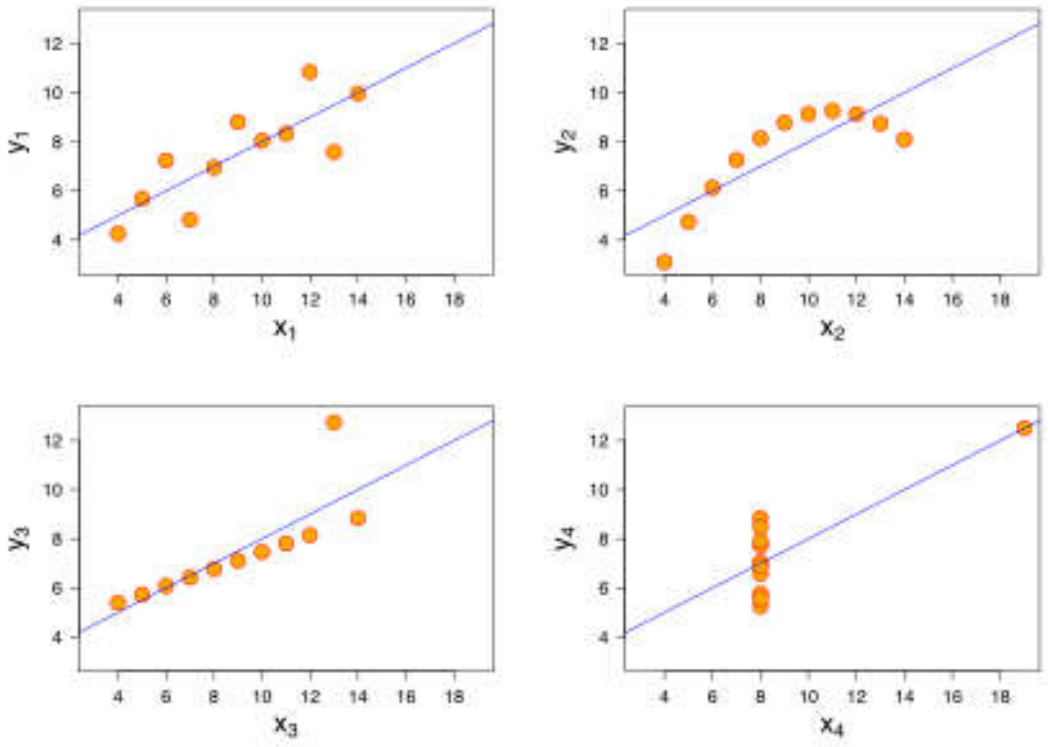
https://www.wsj.com/articles/coronavirus-latest-updates-10-23-2020-11603442326

Representation errors

Sometimes, the way we report the data hides or distorts important information, leading to the wrong conclusion.

Hiding the Distribution

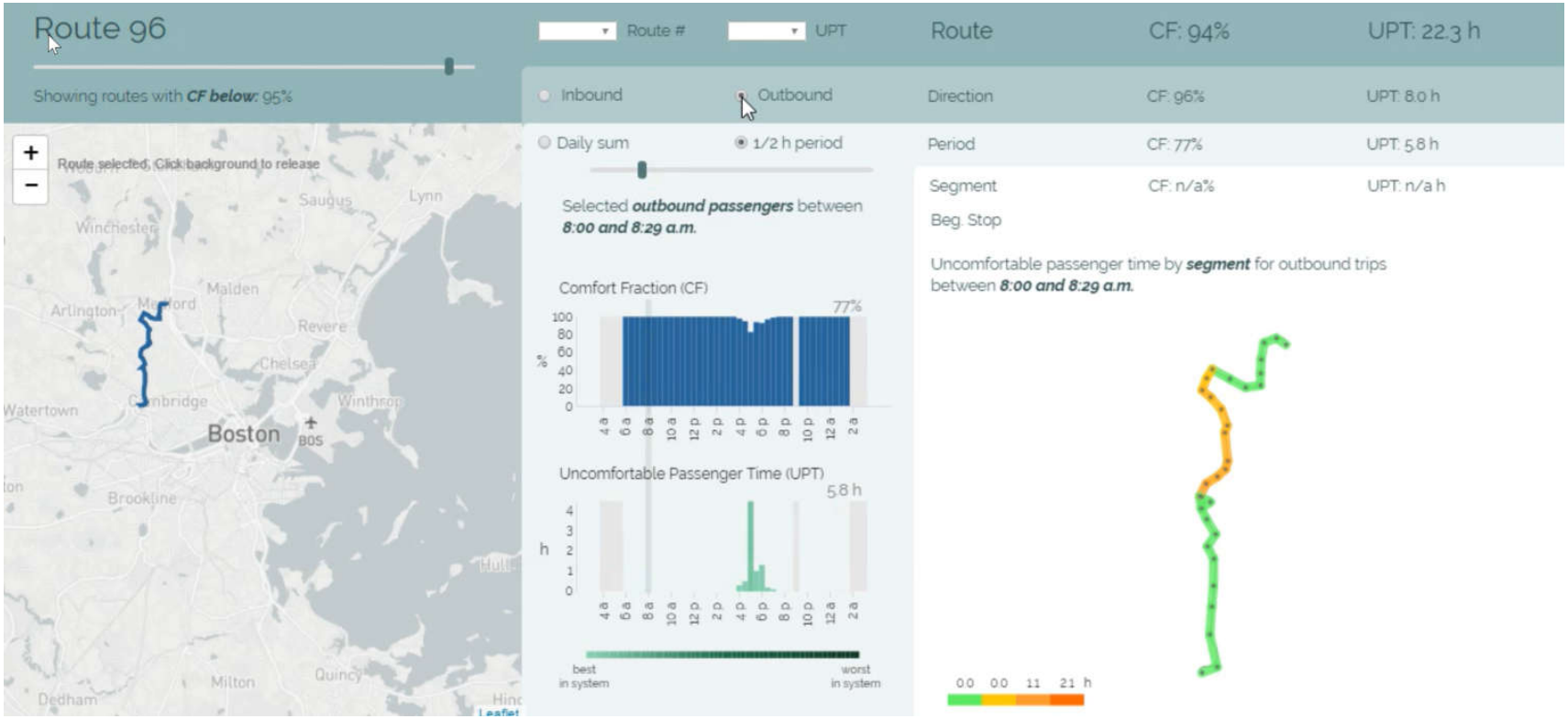
Anscombe's Quartet: The descriptive statistics for these charts are exactly the same, but the data points are quite different.



Look at the underlying data points, and run quality checks on your analysis.

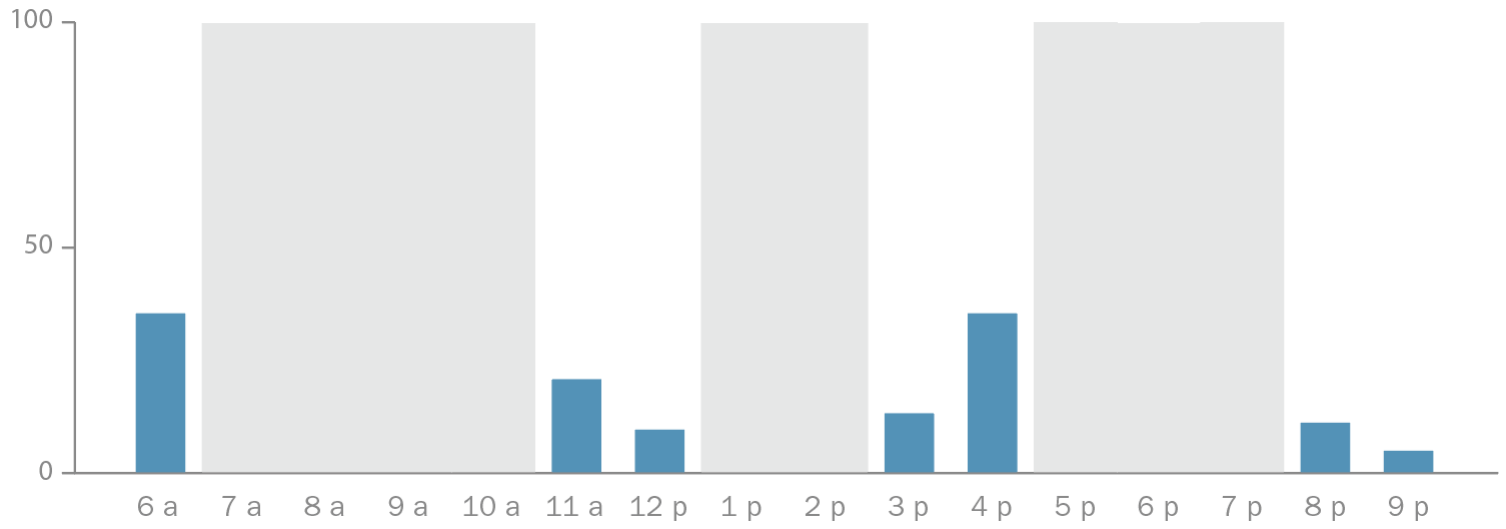
Null Data Errors

In some visualizations, failing to distinguish between zero and null can lead to errors.



Find the Missing Bars

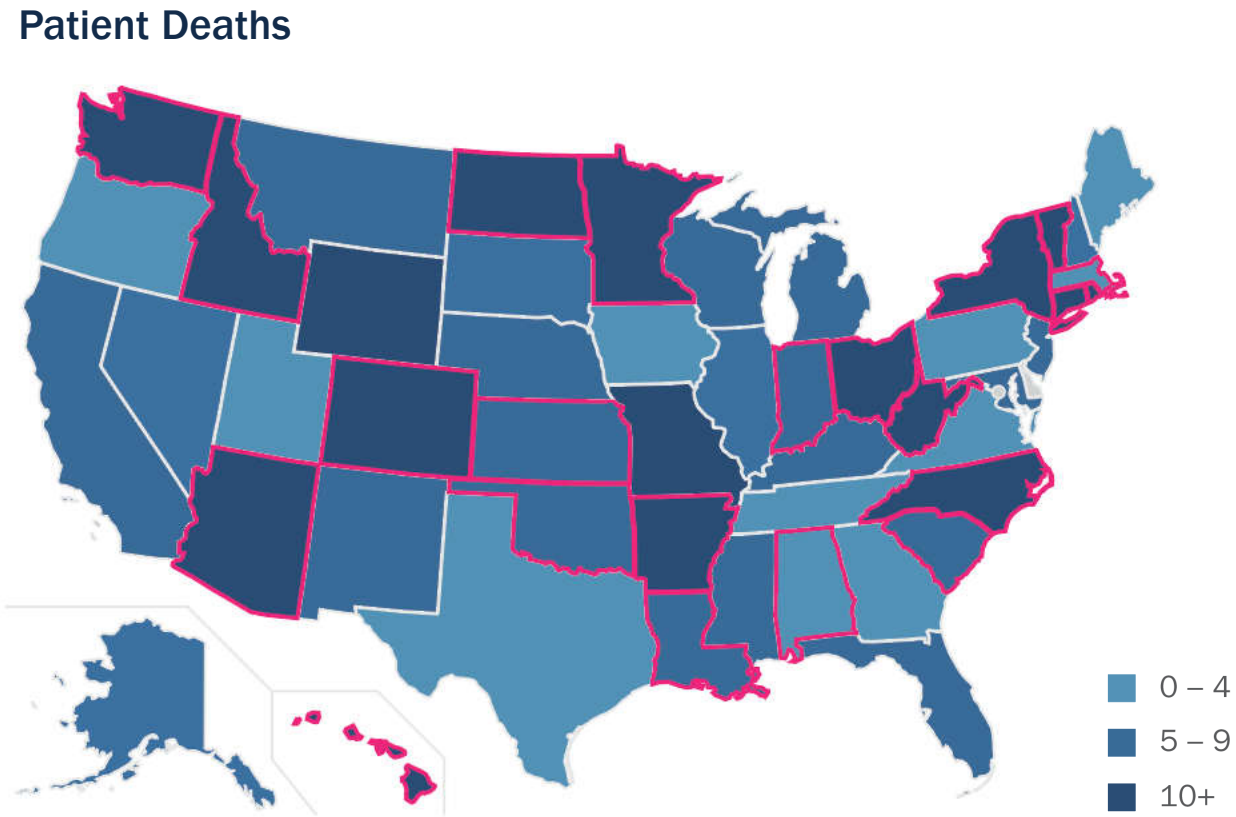
Look for the data that's missing from your representation



There is a big difference between zero and "I didn't measure."

Nulls in Representational Graphics

Maps are particularly prone to problems with null data



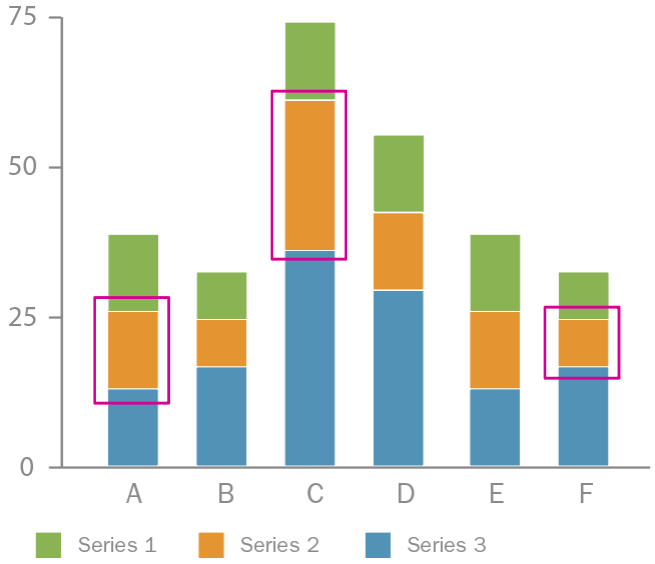
Think about what your encodings imply

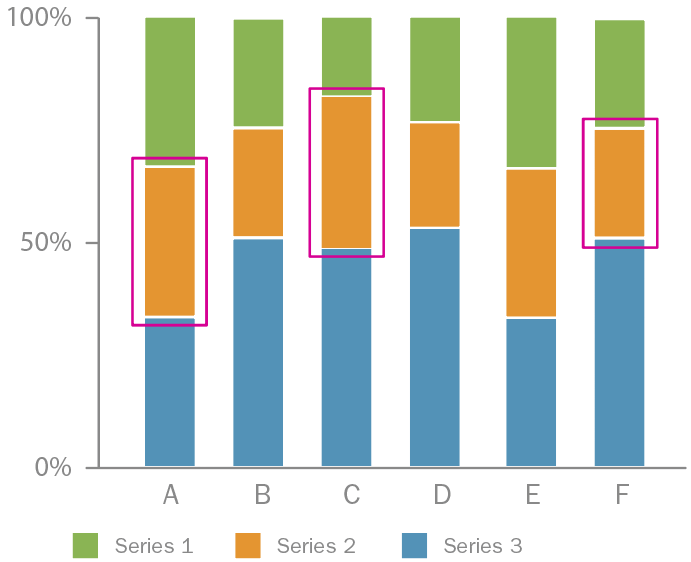
Denominator errors

Data scaling can distort data values and interfere with a user's understanding.

Don’t Do This #3: Percentage scaling

Be careful with percentages in proportional representations in general

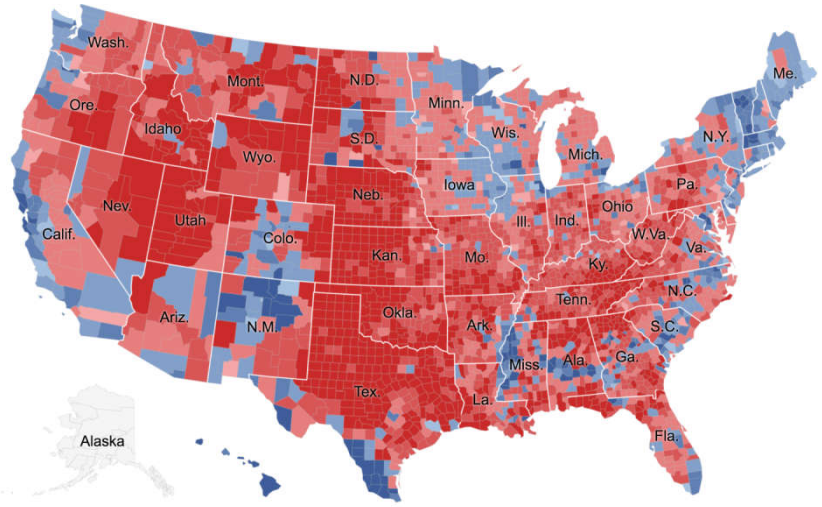




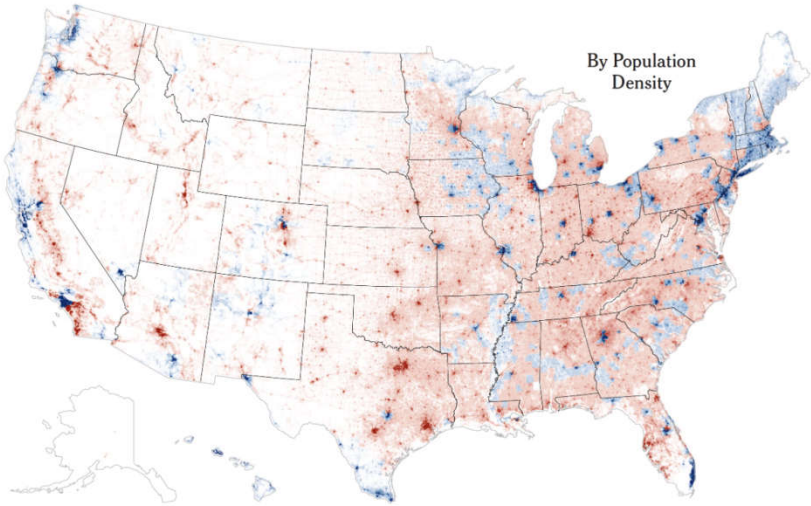
Mixing Area and Value

Choropleth maps mix data value with geographic area

Bin by area



Population density



https://www.nytimes.com/interactive/2016/11/01/upshot/many-ways-to-map-election-results.html

Case Study in Binning

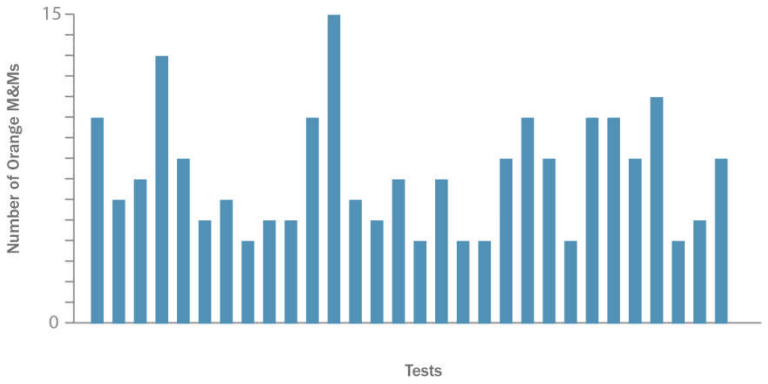
When you're working with binned data, the size of the bin makes a huge difference in what you can see!

What's the average number of orange M&M's in a bag?

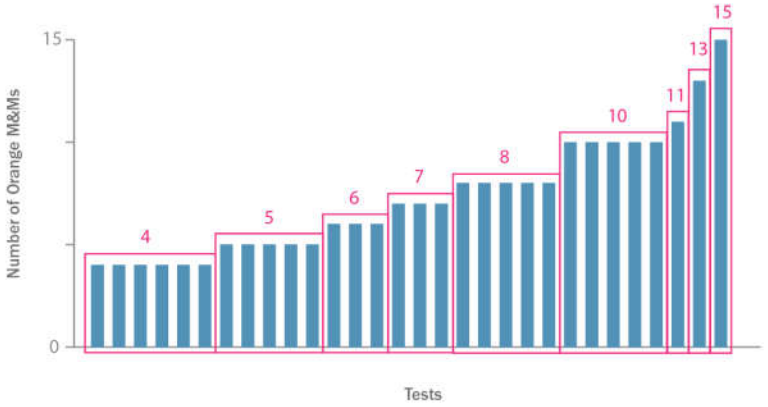
Building a Histogram

A histogram is a way of turning raw observations into frequency counts by binning similar measurements.

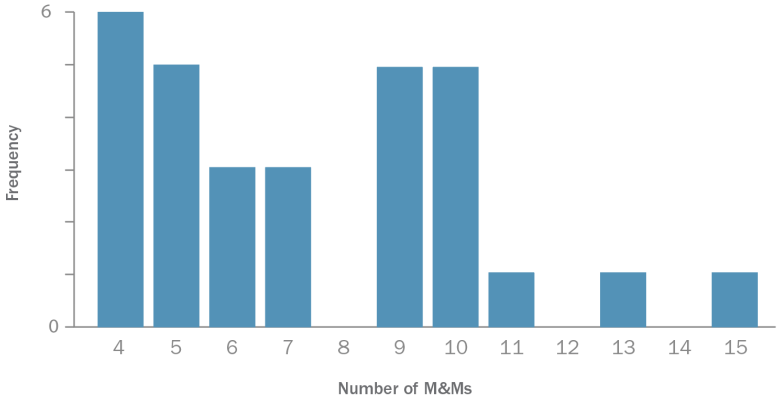
Raw Data (one bar per test)



Sort by value



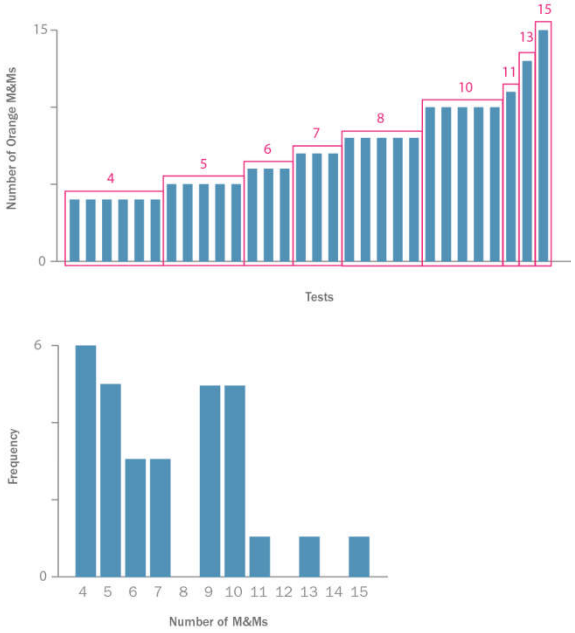
Bin by value (one bar per value)



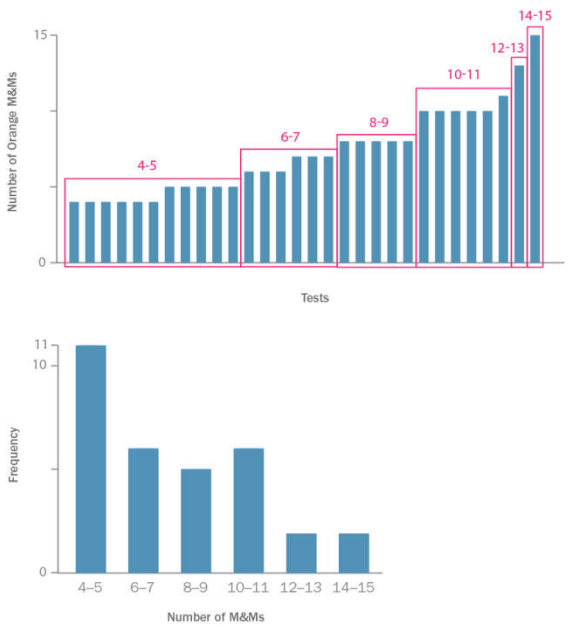
Choosing a Bin Size

The size of the bin makes a huge difference in what you can see!

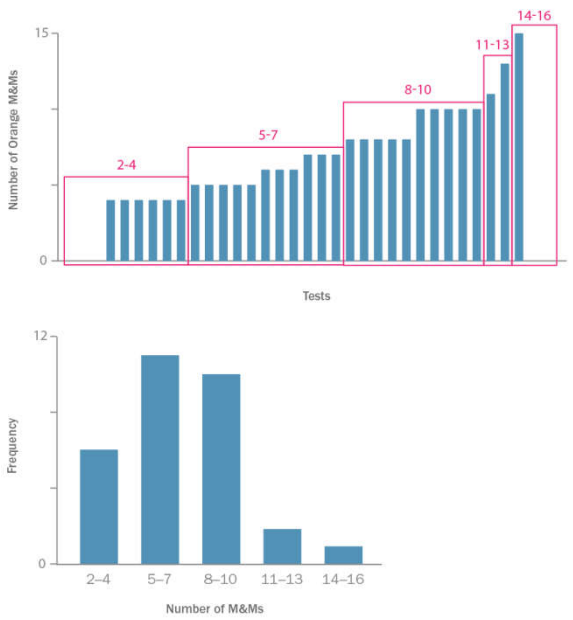
Bin width = 1



Bin width = 2



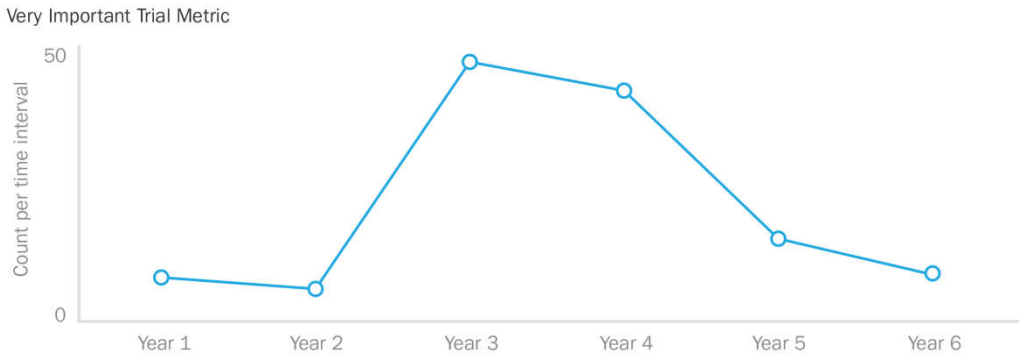
Bin width = 3

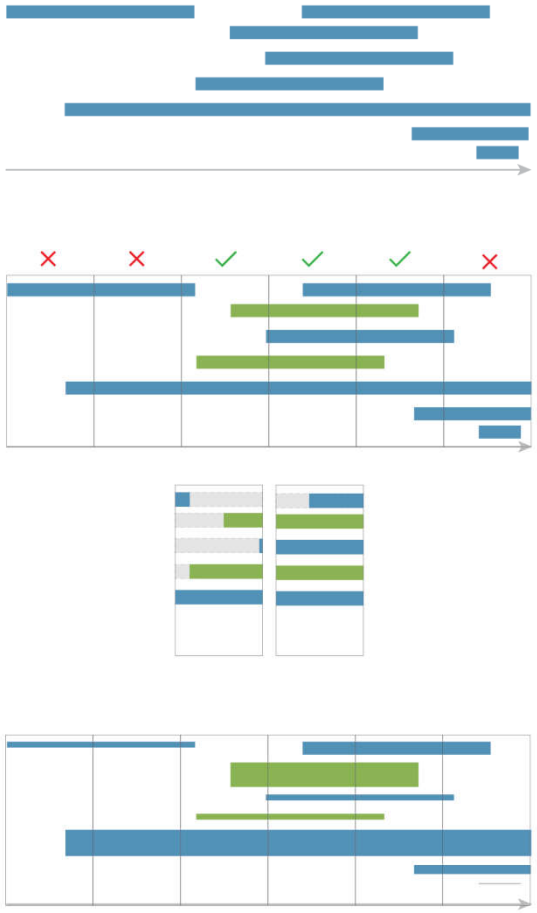


Check that your counting methods make sense

A Case Study in Aggregation

This chart represents counts per time interval, so that a user can see "trends over time." Let's take a closer look at what's going on.





Proposed Solution

Support a summary chart with details in a table to provide context and guide interpretation

Chart shows summary values and statistical distributions

Table provides additional details that allow a user to judge the quality and relevance of data in the chart

Part III:
Leveraging Design Principles

Visualization as a design problem

Tips for displaying your data

Leveraging Design

Basic design principles can make a complicated chart clearer

- Labels
- Animation/introduction
- Keys and legends
- Highlight important data
  - Color
  - Text/object style
  - Visual salience
- Interactions
- Dynamic annotations
- Tooltips to show values

The Bottom Line

You are responsible for ensuring that your visualizations are clear, accurate, and appropriate for the data.

Your professional reputation, the credibility of your profession, and sometimes people's lives depend on it.

It is easy to hide things in the chart or in the model; it is your job to make sure that we never do that.

Summing up: Things to think about

- Does this chart:
  - Show what it needs to show?
  - Support the user task?
- Is it appropriate for the data?
- Are elements of the data or analysis hidden or distorted?
- Could this chart lead to misunderstandings, misinterpretations, or errors?
- How can you use design to clarify and enhance the chart to suit your intended purpose?

Want to learn more?

www.datavisualizationsociety.com

- Slack Group
- Medium Publication
- Tutorials, Fireside chats, Office hours with experts, and more!

Monthly(ish) Boston/virtual meetups
#location-newengland channel

Keep in touch!

www.ericagunn.com

@Erica.Gunn

@EricaGunn
#UXPABos2020

@EricaGunn
#DataVisualizationSociety