

Visualisation and data analysis for journalism studies

Rafal Urbaniak and Nikodem Lewandowski
(University of Gdansk)
<https://rfl-urbaniak.github.io/teaching/>
rfl.urbaniak+teaching@gmail.com

The plan

- Motivations, goals, game rules
- Some history
- The role of perception
- Getting started with R, RStudio and ggplot2
- More on what to show
- Focus
- Epistemic problems
- Technical and mathematical problems
- Statistical learning and probabilistic thinking
- Statistical and analytical blunders
- Basics of Bayesian thinking
- Linear models
- Causality and variable selection

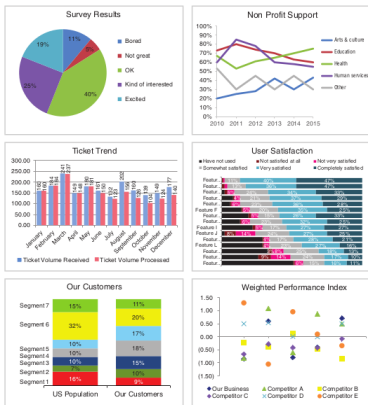
Motivations

- It's too easy to generate tables and visualisation.
- This makes communication harder!

Motivations

- It's too easy to generate tables and visualisation.
- This makes communication harder!

Bad graphs everywhere!



Lack of background

- We learn some math at school.
- We learn some arts at school.

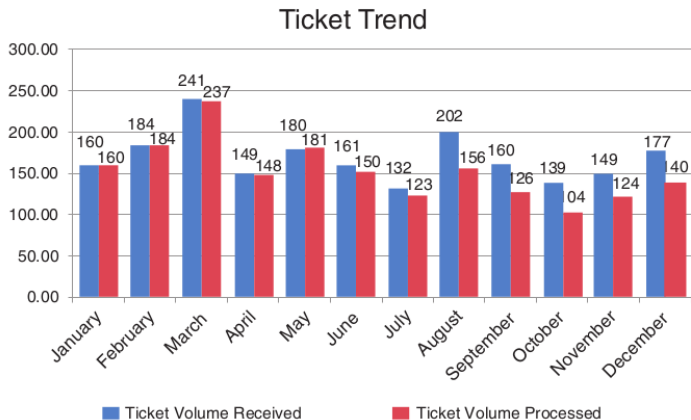
Lack of background

- We learn some math at school.
- We learn some arts at school.

Problem

We never learn to put them together, and think they're opposite.

Some examples



Cole Nussbaum [4]

Some examples

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



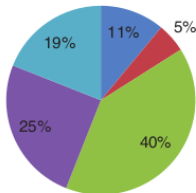
Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

Some examples

Survey Results

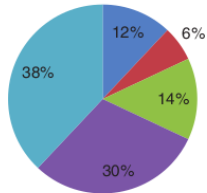
PRE: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



POST: How do you feel about doing science?

■ Bored ■ Not great ■ OK ■ Kind of interested ■ Excited



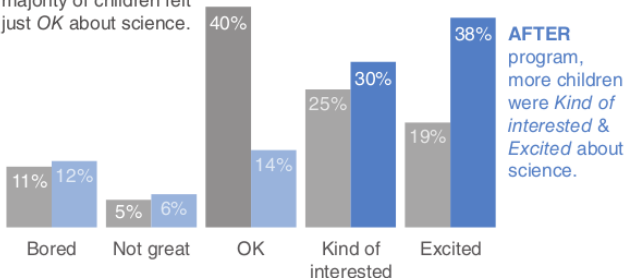
Cole Nussbaum [5]

Some examples

Pilot program was a success

How do you feel about science?

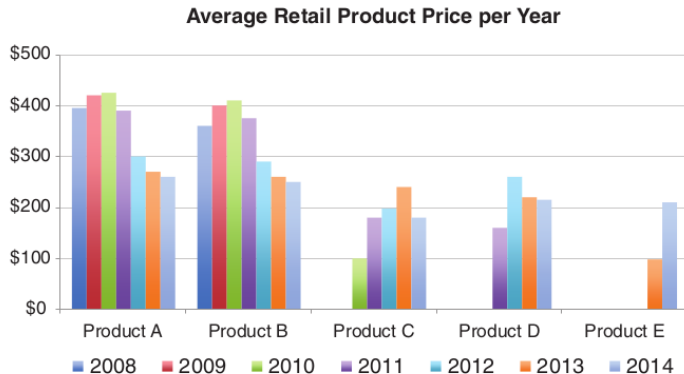
BEFORE program, the majority of children felt just *OK* about science.



Based on survey of 100 students conducted before and after pilot program (100% response rate on both surveys).

Cole Nussbaum [5]

Some examples

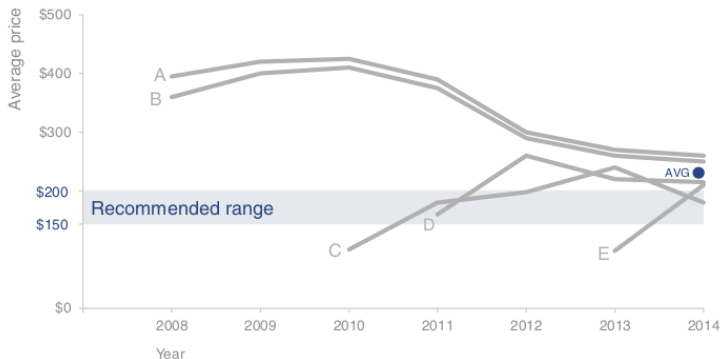


Cole Nussbaum [6]

Some examples

To be competitive, we recommend introducing our product *below* the \$223 average price point in the **\$150–\$200 range**

Retail price over time by product



Cole Nussbaum [6]

Goal

- To understand psychological factors that guide various visualization choices
- To be able to properly analyze data yourself (at a decent level, or at least to understand some of the complexities involved)
- To be able to visualize your data insights so that they clearly convey your message
- To be able to work in R, a statistical programming language

Rules: final grade

Final test: 60 points (optional)

- multiple choice with penalty points

Project: 60 points (optional)

- two-three pages of meaningful text with at least two visualizations, bonus points for animations
- everything prepared in R markdown
- feedback loop: idea -> draft -> feedback -> revisions -> f2 -> r2

Tutorial performance: 60 points (optional)

- If you complete a free-fall exercise without much help, show us, get some points!

Final grade

As if out of 100.

Contact

Updates - only here!

<https://rfl-urbaniak.github.io/teaching/>

Contact - only here!

rfl.urbaniak+teaching@gmail.com



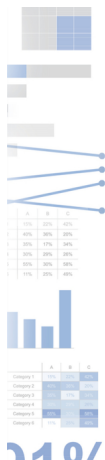
Avoiding Data Pitfalls

How to Steer Clear of Common Blunders
When Working with Data and Presenting
Analysis and Visualizations

Ben Jones

WILEY

Sources



cole nussbaumer knaflic

storytelling with data

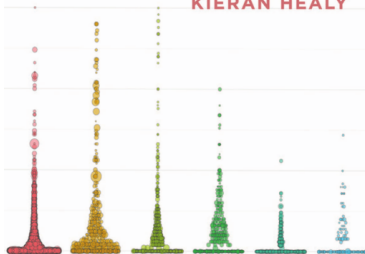
a data
visualization
guide for
business
professionals

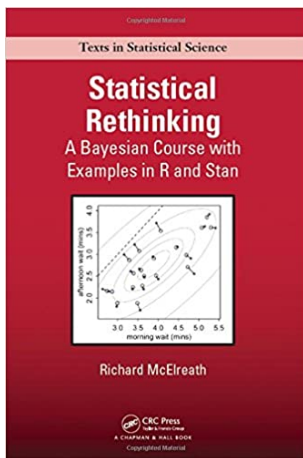
WILEY

DATA VISUALIZATION

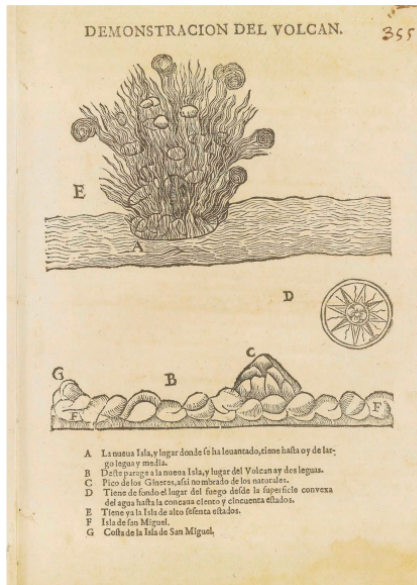
A PRACTICAL INTRODUCTION

KIERAN HEALY





Precursors



The Daily Courant.

Saturday, September 12. 1702.

LONDON, Sept. 12.

THIS Defeat of the Duke of Ormond in the Bay of Cadix being the most considerable Enterprise since the English were driven out of this last Spanish Town; will come under the Publick eye to be here print'd with a Description of that Island and City, which may be of some use for the clearing and clearing of the Island since it came along and that a further report from thence; than will say what else we could have out of the last Foreign News and News Letters, were they not already given what a great matter it is.

THE Head of Cádiz lies between the Mouth of the River Guadalequivi, and the Strait of Gibraltar, on the Coast of Andalucia: It is join'd to the Sea Land by a Bridge of 700 Paces long, call'd La Puente de Leon. The Distance between this

Is Mary and the Bridge is about 12 Miles; from the Bridge to the City of Cadix is likewise about 12 Miles: The Rocks call'd the Diamond and La Puente make the Entrance into the Bay pretty dangerous: The Bay is about 12 Leagues in Circumference and a Bread; but the narrowest part (not Side of which is defended by a Fort call'd the Puerto, and the other by the Fort of St. Juan,) is not above a Mile and an half more: This Part of the Island on which the Town stands is defended towards the Sea by the Fort of St. Philip, the Fort of St. Sebastian, the Walls of the Town, (which surround the Fort) are call'd the Tabic, and by Army Rocks: And there are very strong Fortifications to secure the Passage to the Town by the narrow Neck of Land that runs from the larger Part of the Island to the City: The Fort of St. Catherine which is situate by the Daily Banks between Port St. Mary and the Bay. There are a great many Churches in Cadix, which is well built, very rich, and full of Inhabitants.

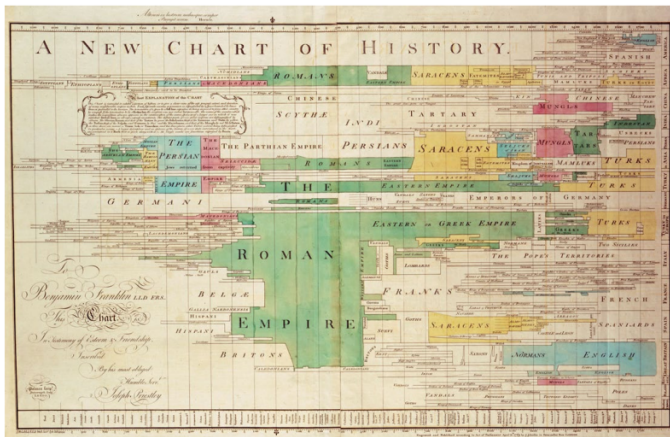


- A. A. The Bay of Cadix.
 B. B. The Diamond.
 C. C. La Puente.
 D. D. Port St. Mary.
 E. E. Fort Leon.
 F. F. La Puente de Leon.
 G. G. Puerto.
 H. H. The Fortifications on the Neck of Land by Port St. Philip.

- I. I. The City.
 K. K. Fort St. Philip.
 L. L. Fort Leon.
 M. M. Cadix.
 N. N. Fort of St. Sebastian.
 O. O. St. Catherine.
 P. P. The Head of the Bay.
 Q. Q. A Head of the Bay.

Precursors

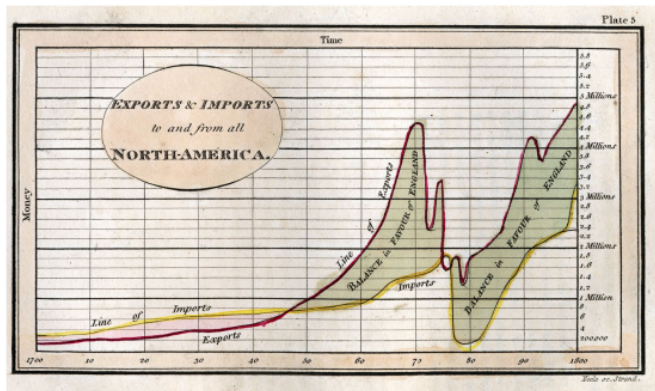
Joseph Priestley (1733-1804)



A new chart of history, 1769

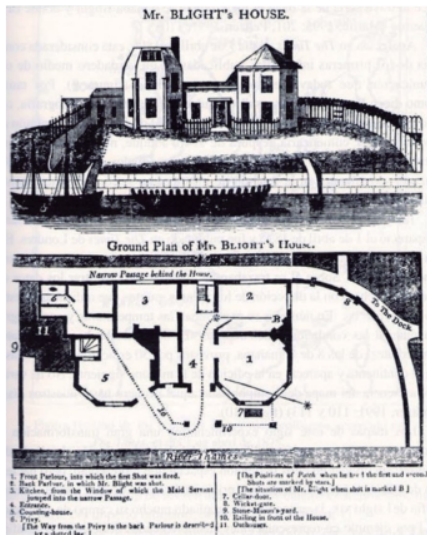
Precursors

William Playfair (1759-1823)



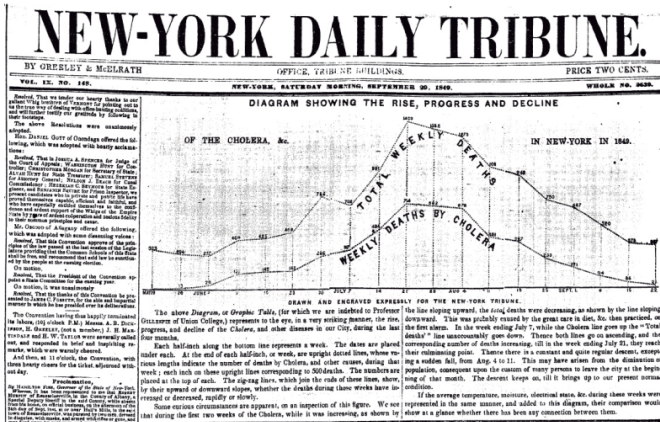
Statistical breviary, 1801

Precursors



A murder case coverage in *The Times*, 1806

William Mitchell Gillespie (1816-1868)

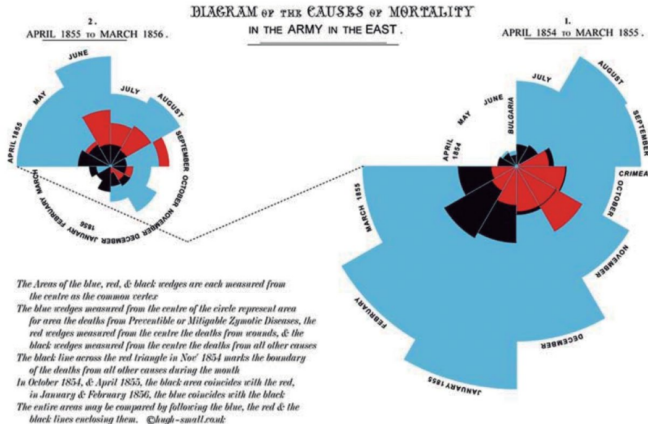


XIXth century explosion

Reasons

- modern nation-states with increased interest in collecting economic and demographic data
- descriptive statistical methods used before in physical sciences began to be used in social sciences (e.g. Adolphe Quelet, Francis Galton)
- dawn of new sciences, such as epidemiology

Florence Nightingale (1820-1910) and the Crimean war

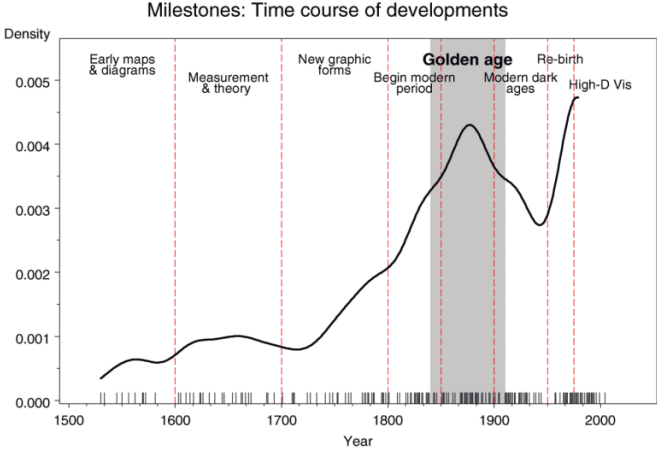


Causes of Mortality, 1856

John Snow (1813-1858) and cholera in London



Modern dark ages in statistics



Number of visualization historical landmarks per year, *Friendly 2008*

The pictorial turn in newspapers

Newspapers became a prime site where visual art and popular forces met and made their peace, and news contributed to the fullness of modernism as it arrived in the twentieth century [...] During the century, the newspapers in the study shifted from the abundant complexity of the Victorian era to the fixed simplicity of modernism. They adopted all the specific forms commentators identified with the modern style: fewer columns, prominent illustrations, horizontal layout, and simplified headline typography. (Barnhurst & Nerone 2001)

Yellow kid journalism (1895-1898)

Say what?

Sensational journalism in the circulation war between Joseph Pulitzer's *New York World* and William Randolph Hearst's *New York Journal* (Pulitzer tried to be more content-based but circulation shrank)



Yellow Kid, *New York World* and *New York Journal*

Yellow kid journalism (1895-1898)

\$50,000 REWARD.—WHO DESTROYED THE MAINE?—\$50,000 REWARD.

EDITION FOR GREATER NEW YORK

NEW YORK JOURNAL AND ADVERTISER


DESTRUCTION OF THE WAR SHIP MAINE WAS THE WORK OF AN ENEMY.

\$50,000!
\$50,000 REWARD!
for the Detection of the Perpetrator of the Maine Outrage!

Assistant Secretary Roosevelt Convinced the Explosion of the War Ship Was Not an Accident.

\$50,000!
\$50,000 REWARD!
for the Detection of the Perpetrator of the Maine Outrage!

The Journal Offers \$50,000 Reward for the Conviction of the Criminals Who Sent 250 American Sailors to Their Death. Naval Officers Unanimous That the Ship Was Destroyed on Purpose.



NAVAL OFFICERS THINK THE MAINE WAS DESTROYED BY A SPANISH MINE.

Hidden Mine or a Sunk Torpedo Believed to Have Been the Weapon Used Against the American Man-of-War—Officers and Men Tell Thrilling Stories of Being Blown into the Air Amid a Mass of Shattered Steel and Exploding Shells—Survivors Brought to Key West Scout the Idea of Accident—Spanish Critics Protest. Two Months' War Cabinet Orders a Searching Inquiry—Journal Sends Divers to Havana to Report Upon the Condition of the Wreck. Was the Vessel Anchored Over a Mine?

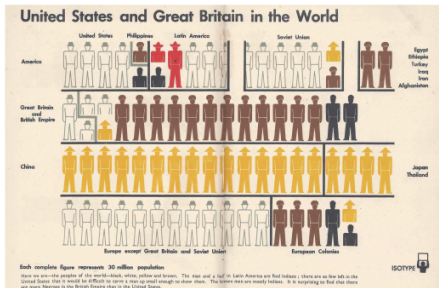
BY CAPTAIN J. J. PALMER, U. S. N.

Washington, agents can see that Captain Rogers had started some such story as a hidden mine. The English light ends was used all day yesterday.

The sinking of Maine in the bay of Havana (notice the Spanish mine), *New York Journal*, Feb. 17, 1898

Viennese Museum for Society and the Economy (1924)

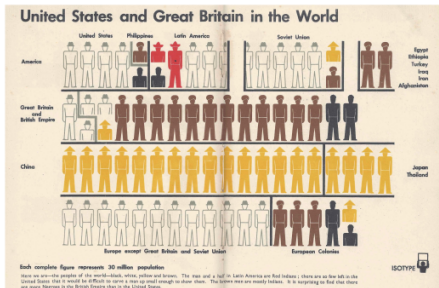
Facts for the uneducated



ISOTYPE, universal visual language by Neurath, Arntz and Reidemeister

Viennese Museum for Society and the Economy (1924)

Facts for the uneducated



ISOTYPE, universal visual language by Neurath, Arntz and Reidemeister

The "Bible"

Pictographs and Graphs: How to Make and Use Them, Modley & Lowenstein, 1952



A page from Fortune, 1929

Birth of USA Today (1982)

WEATHER ACROSS THE USA

HOW TO USE THIS PAGE

The map and table show the high temperature and the amount of precipitation expected for each area during the next 24 hours. The legend shows the symbols for each weather condition.

LEGEND

High Temperature
Precipitation

Clear
Partly Cloudy
Cloudy
Thunderstorm
Rain
Snow
Sleet
Fog
Haze
Ice

FOUR-DAY HIGHLIGHTS

For more details on the weather conditions in your area, see the forecast for your area on page 10.

Thunderstorm makers

As the sun beats down on the Southern Plains, the air is becoming increasingly unstable. This is the perfect recipe for the formation of thunderstorms. The air is becoming increasingly unstable because of the combination of warm, moist air from the Gulf of Mexico and dry, hot air from the desert. This combination creates a perfect storm for the formation of thunderstorms.

Tornadoes possible in Southern Plains

As the sun beats down on the Southern Plains, the air is becoming increasingly unstable. This is the perfect recipe for the formation of tornadoes. The air is becoming increasingly unstable because of the combination of warm, moist air from the Gulf of Mexico and dry, hot air from the desert. This combination creates a perfect storm for the formation of tornadoes.

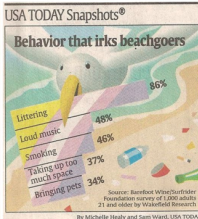
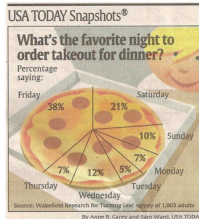
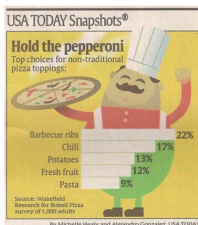
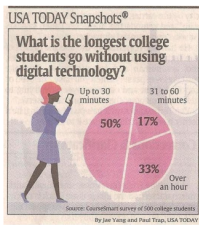
WORLD WEATHER TODAY

Area	High	Low	Wind	Clouds	Precip.
ALASKA	40-50	20-30	10-20	Partly cloudy	None
ARIZONA	80-90	50-60	10-20	Clear	None
ARKANSAS	70-80	40-50	10-20	Partly cloudy	None
CALIFORNIA	60-70	40-50	10-20	Partly cloudy	None
COLORADO	60-70	40-50	10-20	Partly cloudy	None
CONNECTICUT	60-70	40-50	10-20	Partly cloudy	None
DELAWARE	60-70	40-50	10-20	Partly cloudy	None
FLORIDA	80-90	60-70	10-20	Partly cloudy	None
GEORGIA	70-80	50-60	10-20	Partly cloudy	None
ILLINOIS	60-70	40-50	10-20	Partly cloudy	None
INDIANA	60-70	40-50	10-20	Partly cloudy	None
IOWA	60-70	40-50	10-20	Partly cloudy	None
KANSAS	70-80	50-60	10-20	Partly cloudy	None
KENTUCKY	60-70	40-50	10-20	Partly cloudy	None
Louisiana	70-80	50-60	10-20	Partly cloudy	None
Maine	60-70	40-50	10-20	Partly cloudy	None
Maryland	60-70	40-50	10-20	Partly cloudy	None
MASSACHUSETTS	60-70	40-50	10-20	Partly cloudy	None
Michigan	60-70	40-50	10-20	Partly cloudy	None
Minnesota	60-70	40-50	10-20	Partly cloudy	None
MISSISSIPPI	70-80	50-60	10-20	Partly cloudy	None
MISSOURI	60-70	40-50	10-20	Partly cloudy	None
Montana	60-70	40-50	10-20	Partly cloudy	None
Nebraska	60-70	40-50	10-20	Partly cloudy	None
NEVADA	80-90	50-60	10-20	Clear	None
New Hampshire	60-70	40-50	10-20	Partly cloudy	None
New Jersey	60-70	40-50	10-20	Partly cloudy	None
New Mexico	80-90	50-60	10-20	Clear	None
New York	60-70	40-50	10-20	Partly cloudy	None
North Carolina	70-80	50-60	10-20	Partly cloudy	None
North Dakota	60-70	40-50	10-20	Partly cloudy	None
OHIO	60-70	40-50	10-20	Partly cloudy	None
Oklahoma	70-80	50-60	10-20	Partly cloudy	None
Oregon	60-70	40-50	10-20	Partly cloudy	None
Pennsylvania	60-70	40-50	10-20	Partly cloudy	None
Rhode Island	60-70	40-50	10-20	Partly cloudy	None
South Carolina	70-80	50-60	10-20	Partly cloudy	None
South Dakota	60-70	40-50	10-20	Partly cloudy	None
Tennessee	70-80	50-60	10-20	Partly cloudy	None
Texas	80-90	60-70	10-20	Partly cloudy	None
Utah	60-70	40-50	10-20	Partly cloudy	None
Vermont	60-70	40-50	10-20	Partly cloudy	None
Virginia	60-70	40-50	10-20	Partly cloudy	None
Washington	60-70	40-50	10-20	Partly cloudy	None
West Virginia	60-70	40-50	10-20	Partly cloudy	None
Wisconsin	60-70	40-50	10-20	Partly cloudy	None
Wyoming	60-70	40-50	10-20	Partly cloudy	None

A revolutionary weather map

Birth of USA Today (1982)

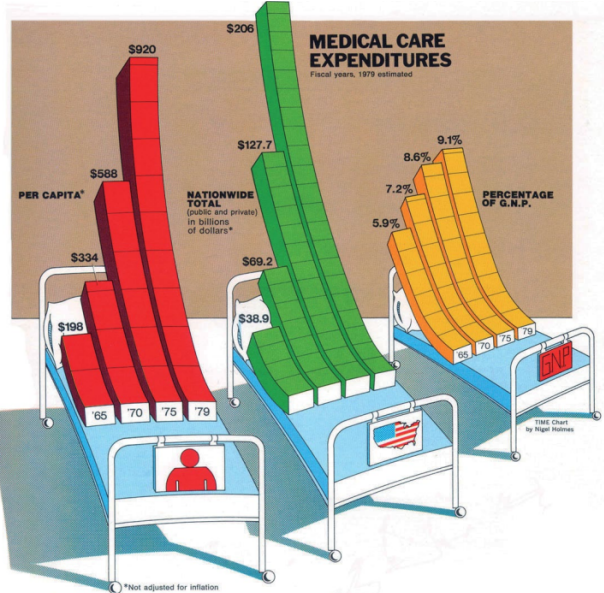
- its success expanded the use of graphics in print publications
- tilted the stylistic balance towards the pictorial and lighthearted
- art training, no quantitative expertise
- in 1984 60% of 156 newspapers reported an increased use of news graphics, and an additional 22% said that they had just incorporated them into their pages



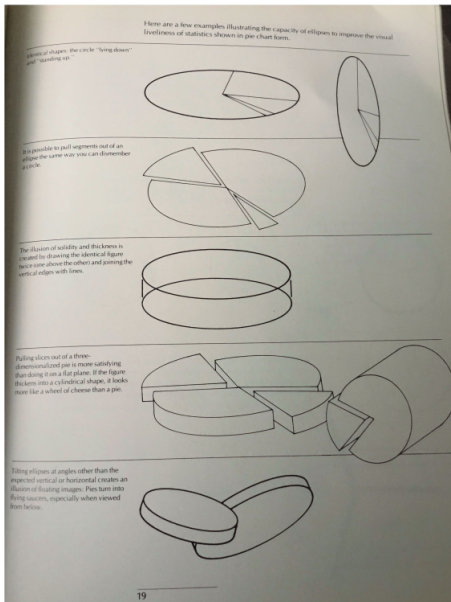
What's the problem?

*Nearly all those who produce graphics for mass publication are trained exclusively in the fine arts and have had little experience with the analysis of data [...] Illustrators too often see their work as a exclusively artistic enterprise—the words "creative", "concept", and "style" combine regularly in all possible permutations, a Big Think jargon for the small task of constructing a time-series a few data points long. Those who get ahead are those who beautify data, never mind statistical integrity.
[Edward Tufte 1983]*

Nigel Holmes

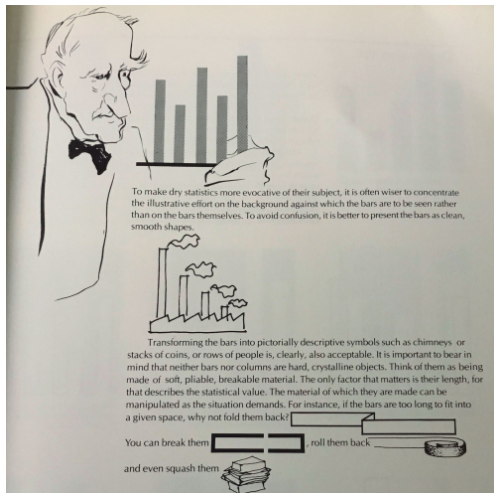


As long as the artist understands that the primary function is to convey statistics and respect that duty, then you can have fun (or be serious) with the image: that is, the form in which those statistics appear. Boredom is as much a threat in visual design as it is elsewhere in art and communication. The mind and eye demand stimulation and surprise.



To make dry statistics more evocative of their subject, it is often wiser to concentrate the illustrative effort on the background against which the bars are to be seen rather than on the bars themselves, [...] transforming the bars into pictorially descriptive symbols such as chimneys or stacks or coins, or rows of people is, clearly, also acceptable [...] The material of which they are made can be manipulated as the situation demands. For instance, if the bars are too long to fit into a given space, why not fold them back? You can break them, roll them back and even squash them.

(Jan. V. White, 1984)

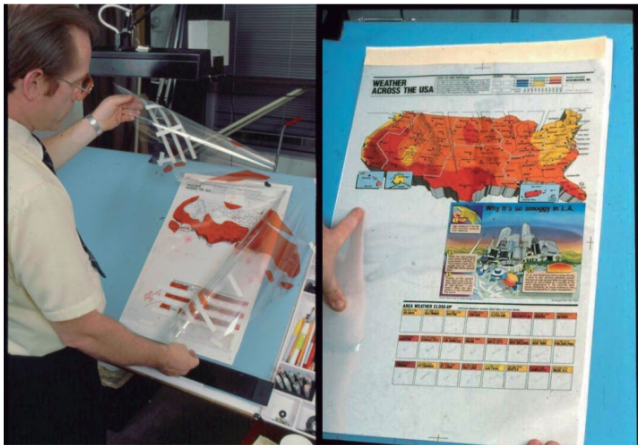


Computer-age graphics



George Rorick, hand-made visualization, 11 a.m. to 6 p.m.

Computer-age graphics



George Rorick, hand-made visualisation, 11 a.m. to 6 p.m.

Computer-age graphics

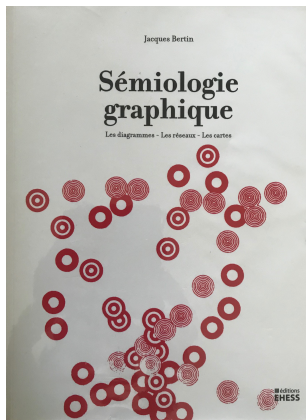
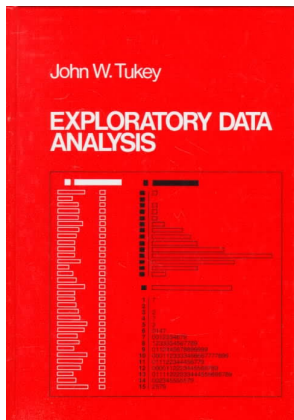
- Apple, 1984
- PostScript & Adobe Illustrator, 1987 (raster vs. vector files)
- Adobe Photoshop, 1989

We went from some very nice illustrated graphics to some very poor computer-generated graphics, but that was the limitations of the technology, and it took about at least five years, maybe more, before we started to see the computer graphics start to rise up in quality.

John Grimwade (check out his website!)

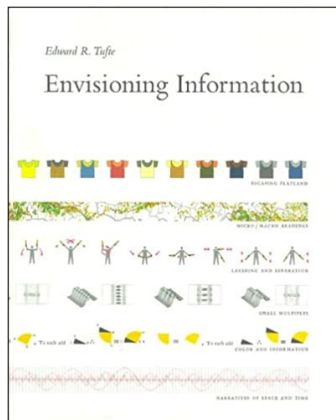
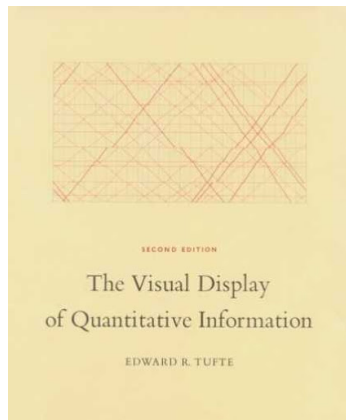
Backlash against cartoons

Tukey 1977, Bertin 1967



Backlash against cartoons

Tufte 1983, 1990



Backlash against cartoons

Sometimes decoration can help editorialize about the substance of the graphic. But it is wrong to distort the data measures—the ink locating values of numbers—in order to make an editorial comment or fit a decorative scheme.
(Tufte 1983: 59)

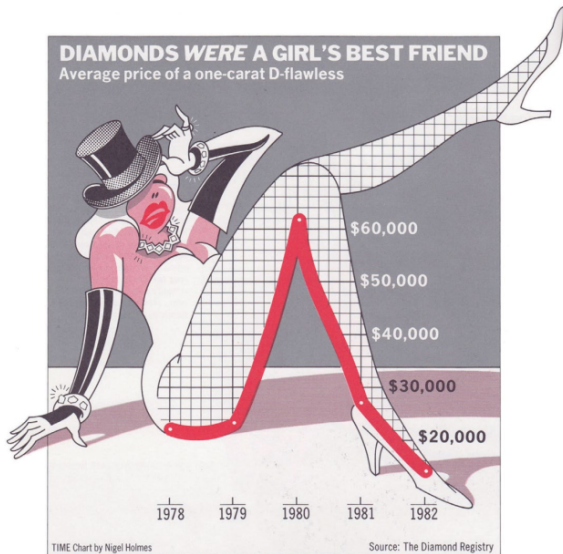
Backlash against cartoons

If you belong to the school of people who believe that charts should only present statistics in the most straightforward, plain way, with no other visual help to the reader, for example, than the bar of the bar chart, the line of the fever graph, the circle of the pie chart, or the rules of the table, then move on to another part of the book [...] Boredom is as much a threat in visual design as it is elsewhere in art and communication. The mind and eye demand stimulation and surprise [...] Even a smile will encourage a reader to look into the statistics he or she might not have thought of reading in a less embellished chart. (Holmes 1984: 72)

Backlash against cartoons

Too many data presentations [...] seek to attract and divert attention by means of display apparatus and ornament. Chartjunk has come to corrupt all sorts of information exhibits and computer interfaces (Tufte 1990: 33)

Backlash against cartoons



Holmes' chart in the *Times* magazine

Backlash against cartoons

Consider this unsavory exhibit at right —chockablock with cliché and stereotype, coarse humor, and a content-empty third dimension. Is it the product of a visual sensitivity in which a thigh-graph with a fishnet-stocking grid counts as Creative Concept. [...] Lurking behind chartjunk is contempt for both information and for the audience. Chartjunk promoters imagine that numbers and details are boring, dull, and tedious, requiring ornament to enliven. Cosmetic decoration, which frequently distorts the data, will never salvage an underlying lack of content. If the numbers are boring, then you've got the wrong numbers. Credibility vanishes in clouds of chartjunk; who would trust a chart that looks like a video game? (Tufte 1990: 34).

Backlash against cartoons

*Graphical competence demands three quite different skills: the substantive, statistical, and artistic. Yet now [in the early 80s] most graphical work, particularly at news publications, is under the direction of but a single expertise —the artistic. Allowing artist-illustrators to control the design and content of statistical graphics is almost like allowing typographers to control the content, style, and editing of prose.
(Tufte 1983: 87).*

Recent developments

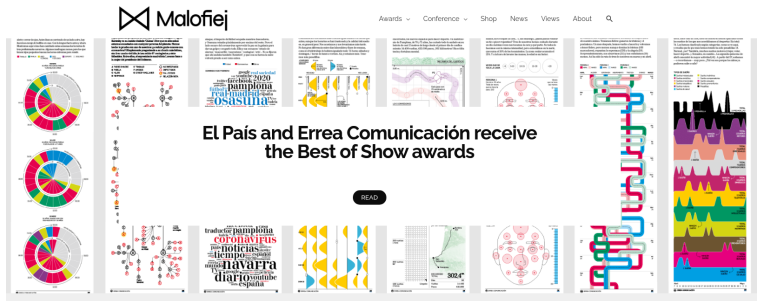
Recent developments

Geek takeover

- more information density and more data
- visualization desks more independent from arts departments
- the 90s and early 2000s: illustration-driven explanations, sometimes supplemented by small and straight-forward statistical graphs and data maps
- today, the balance has shifted to presentations that rely mainly on the visual display of data, both quantitative and qualitative
- often, no longer detached “graphics departments”. Data journalists, nerd journalism!

Recent developments

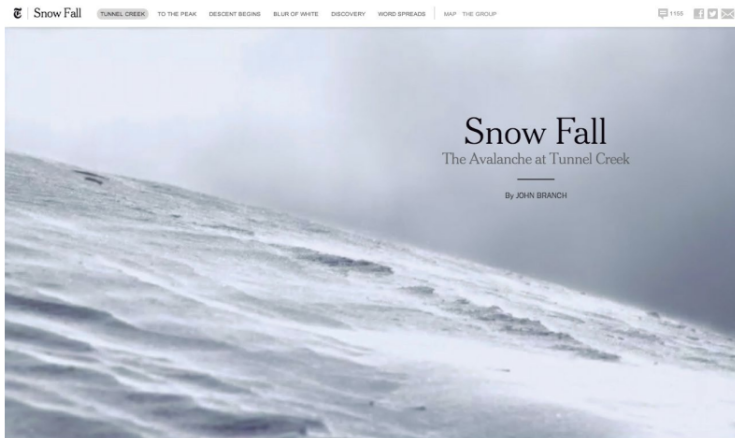
Check out Malofiej awards (1992)



Malofiej awards website

Recent developments

Example (“new era”, 3 mln. in no time)



Snowfall at NY Times

Recent developments




Example (most popular piece in Times, 2013)

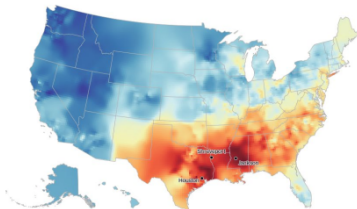
How Y'all, Youse and You Guys Talk

What does the way you speak say about where you're from?
Answer all the questions below to see your personal dialect map.

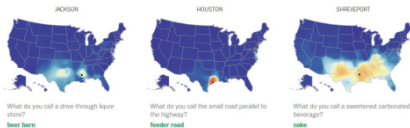
Your Map

See the pattern of your dialect in the map below. Three of the most similar cities are shown.

Least similar Most similar Show least similar SHARE YOUR MAP:   



These maps show your most distinctive answer for each of these cities.



How Y'all quiz, NYT

For the tutorial

Complete the introductory instructions about github, bring a flash drive!

Lecture 2

The role of perception

Exploratory data visualisation

Look at the data!

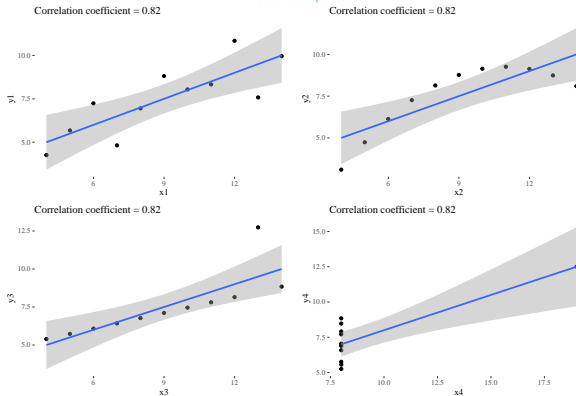
- understand and learn the structure
- obtain insights to pursue

Exploratory data visualisation

Look at the data!

- understand and learn the structure
- obtain insights to pursue

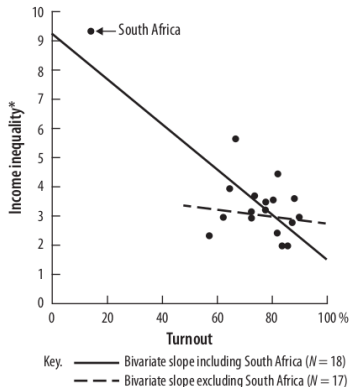
Anscombe's quartet



Exploratory data visualisation

Income and voter turnout

Jackman (1980) on Hewitt (1977). The original paper had argued for a significant association between voter turnout and income inequality based on a quantitative analysis of eighteen countries.



Jackman's illustration of outlier impact

Chartjunk?

Data-to-ink ratio

- Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design.
- [It] consists of complex ideas communicated with clarity, precision, and efficiency.
- [It] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
- [It] is nearly always multivariate. And graphical excellence requires telling the truth about the data.

(Tufte 1983, 51)

Chartjunk?



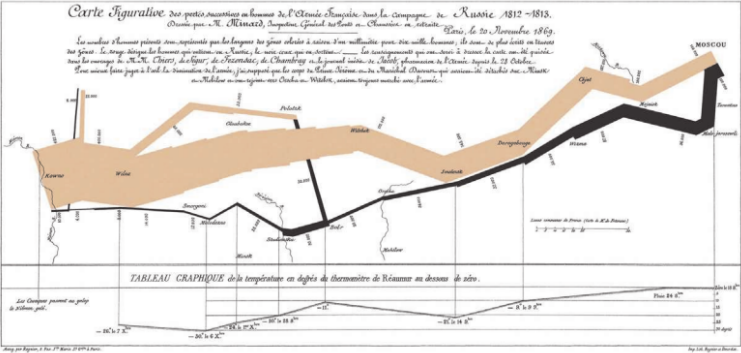
Example of chartjunk

Chartjunk?



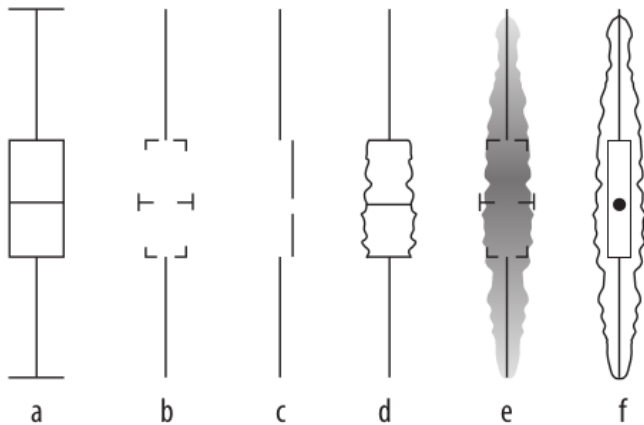
Holmes's *Monstrous costs* are more readily recalled (Bateman et al. 2010)

In contrast



Minard's visualisation of Napoleon's retreat

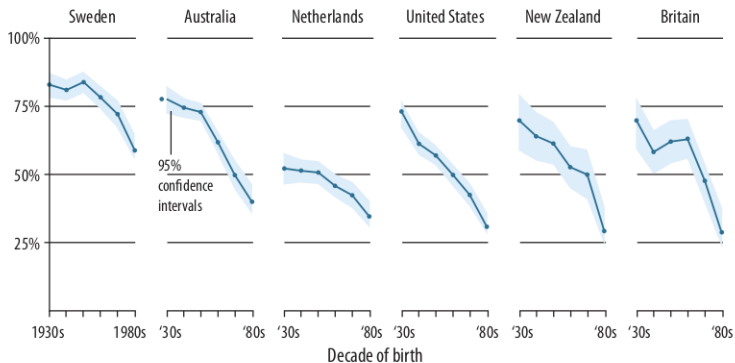
Golden middle?



E. W. Anderson et al. (2011) found that Tufté's (C) proved to be the most cognitively difficult for viewers to interpret.

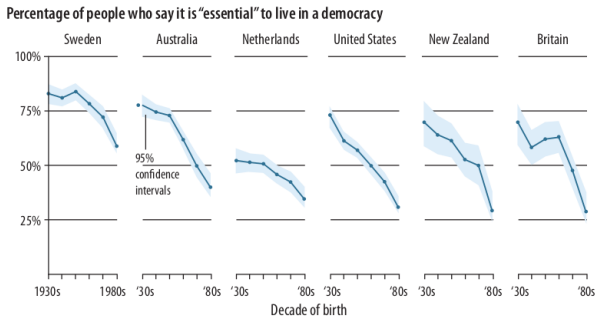
Bad data

Percentage of people who say it is "essential" to live in a democracy



"How Stable Are Democracies?" Warning Signs Are Flashing Red, The Times, 2016

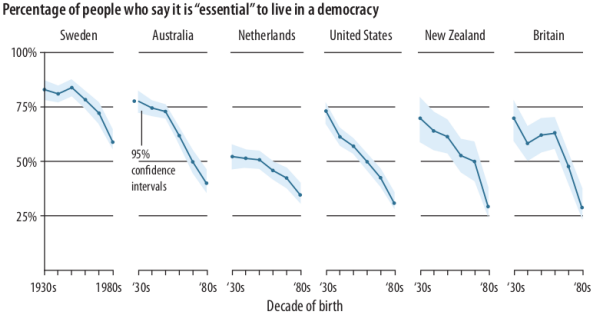
Bad data



"How Stable Are Democracies?" Warning Signs Are Flashing Red, The Times, 2016

- cross-sectional rather than longitudinal (line graph suggests otherwise)!
- Seems like people were asked "is it essential to live in democracy"?

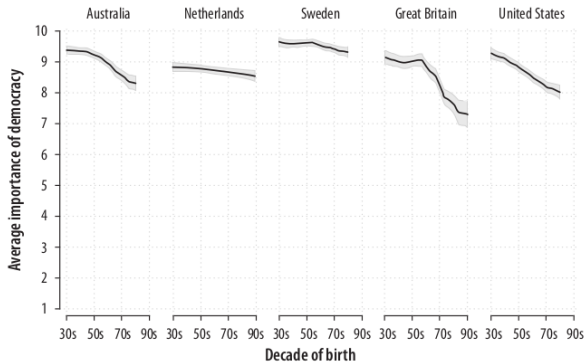
Bad data



“How Stable Are Democracies?” Warning Signs Are Flashing Red, The Times, 2016

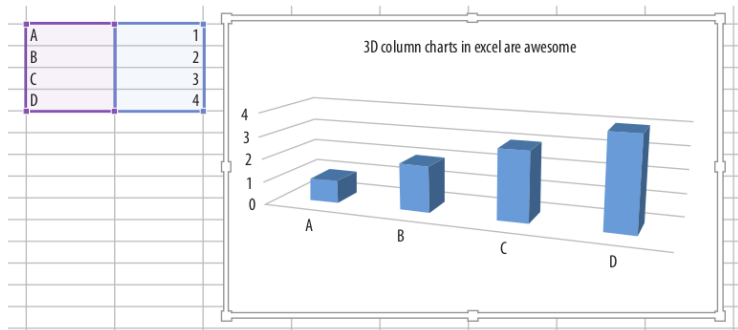
- cross-sectional rather than longitudinal (line graph suggests otherwise)!
- Seems like people were asked “is it essential to live in democracy”?
- In fact, 10-point scale, lines for those who gave 10s.

Bad data



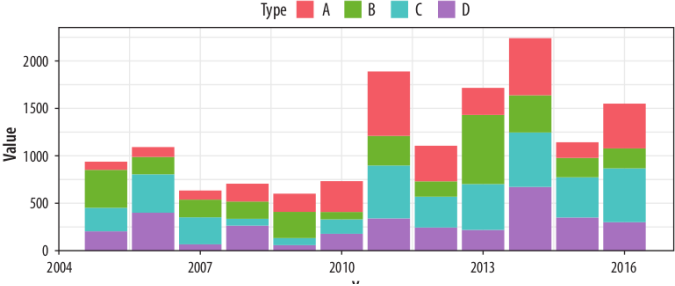
Erik Voeten: same data, mean responses

Bad perception



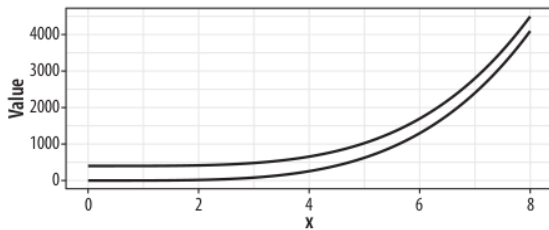
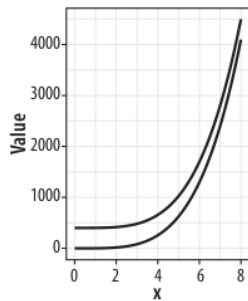
A default bar graph in Excel

Bad perception



Junk free, still hard to interpret

Bad perception

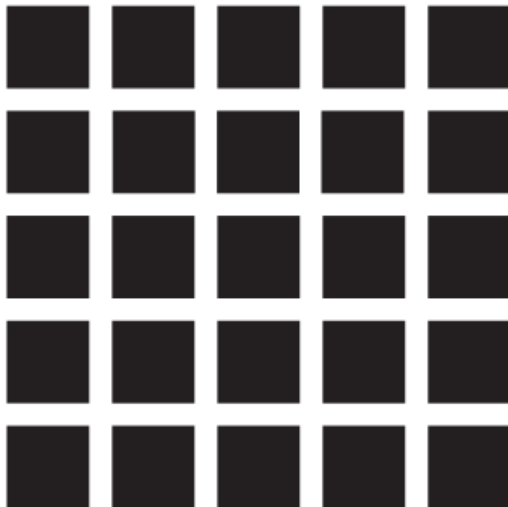


William S. Cleveland's example of the impact of the aspect ratio (no real convergence)

Perception and data visualisation

Edges

Make some thinks easier to see. Even if they're not there.



Hermann's grid effect (1870): blobs at intersections

Perception and data visualisation

Edges

Make some things easier to see. Even if they're not there.



Mach bands: where do you see more contrast?

Perception and data visualisation

Edges

Make some things easier to see. Even if they're not there.

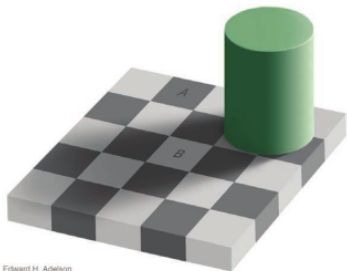


Mach bands: where do you see more contrast?

- same shade of grey is perceived differently depending on background
- distinguishing shades of brightness is not uniform either (we better distinguish dark shades)

Perception and data visualisation

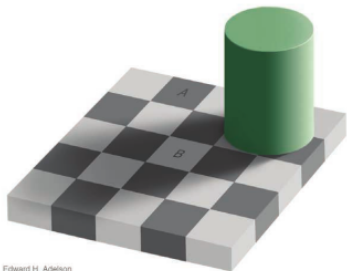
Attraction to edges



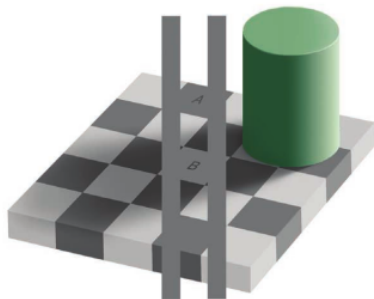
Edward H. Adelson

Perception and data visualisation

Attraction to edges



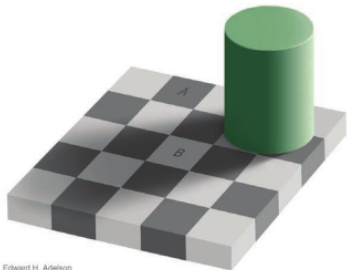
Edward H. Adelson



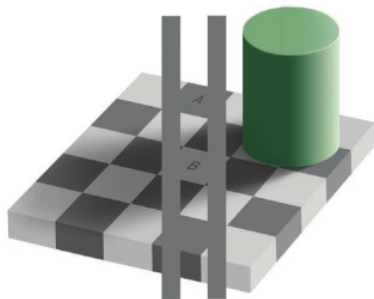
Adelson's checkershadows

Perception and data visualisation

Attraction to edges



Edward H. Adelson



Adelson's checkershadows

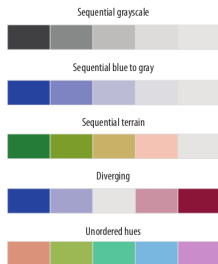
Not like magic trick!

After I explain, you still cannot stop seeing these.

Using colors

Three components

- luminance (conventionally: brightness)
- hue (conventionally: color)
- chrominance/chroma (conventionally: intensity)

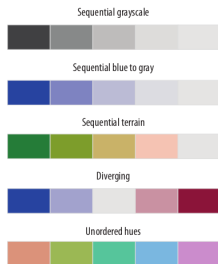


Luminance, Luminance + chroma, all, diverging with a neutral point, unordered

Using colors

Three components

- luminance (conventionally: brightness)
- hue (conventionally: color)
- chrominance/chroma (conventionally: intensity)



Luminance, Luminance + chroma, all, diverging with a neutral point, unordered

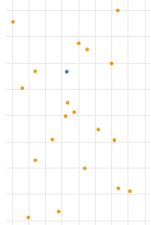
Question

How to meaningfully map data to colors, avoiding blinding the color-blind, and without introducing confusion?

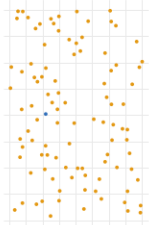
Preattentive search

Preattentive search

Color only, $N = 20$



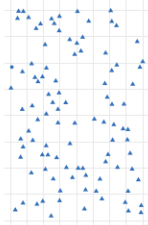
Color only, $N = 100$



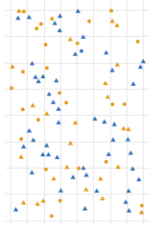
Shape only, $N = 20$



Shape only, $N = 100$

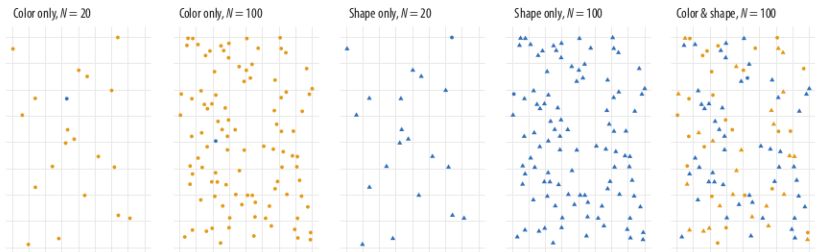


Color & shape, $N = 100$



Find the blue circles

Preattentive search

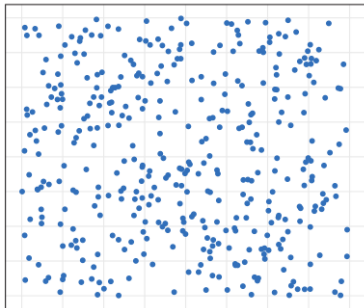


Find the blue circles

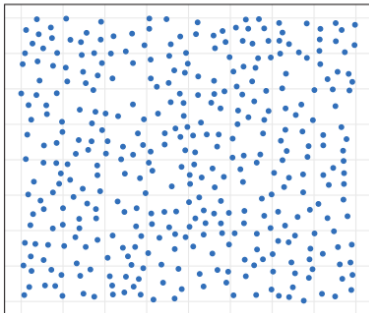
- shape and color are two distinct channels
- pop-out on the color channel is stronger
- dual channels slow people down

Looking for structure

Poisson

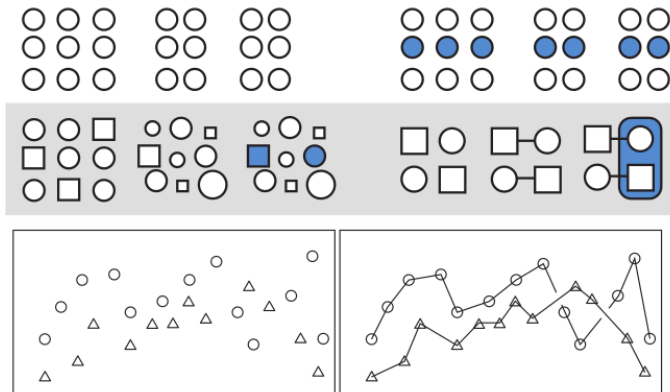


Matérn



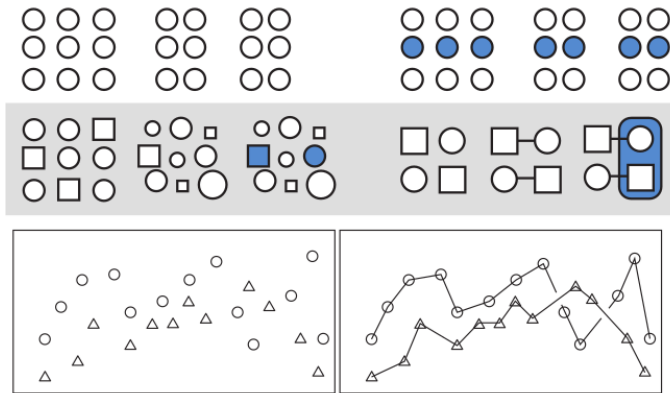
Which is more random?

Gestalt inferences



Proximity, similarity, connection, continuity, closure, figure and ground, common fate

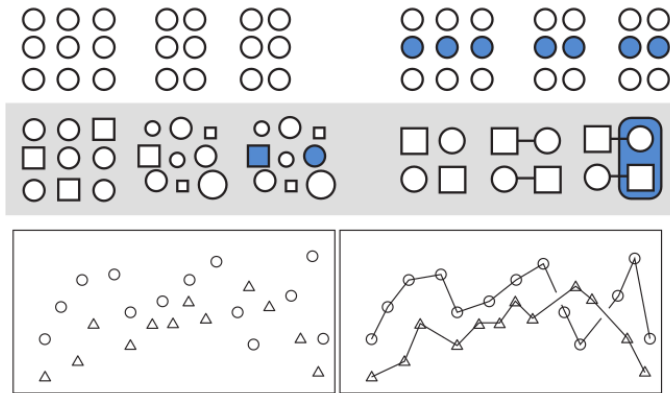
Gestalt inferences



Proximity, similarity, connection, continuity, closure, figure and ground, common fate

- upper left: proximity > shape

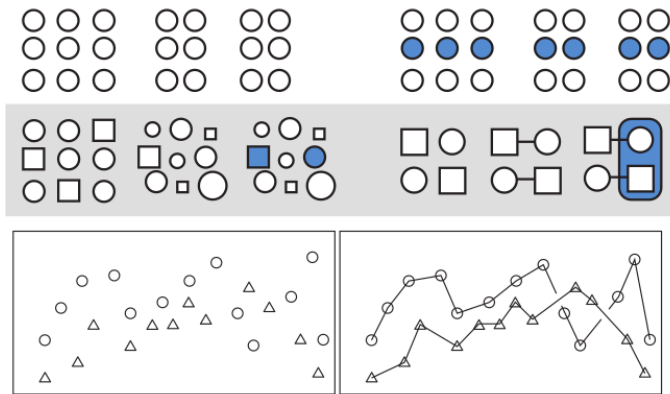
Gestalt inferences



Proximity, similarity, connection, continuity, closure, figure and ground, common fate

- upper left: proximity > shape
- upper right: color > shape, proximity

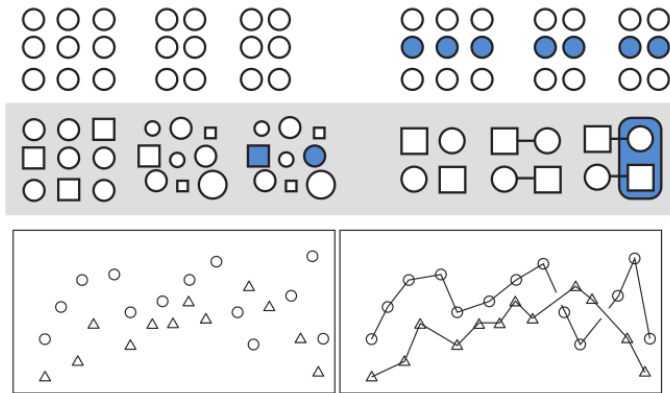
Gestalt inferences



Proximity, similarity, connection, continuity, closure, figure and ground, common fate

- upper left: proximity > shape
- upper right: color > shape, proximity
- middle: left (no clarity), right: connection > shape

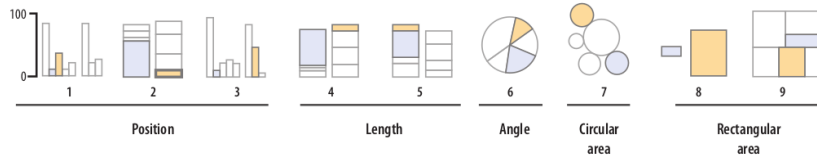
Gestalt inferences



Proximity, similarity, connection, continuity, closure, figure and ground, common fate

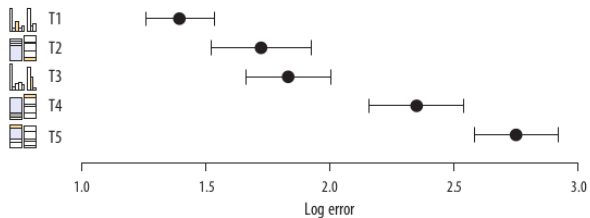
- upper left: proximity > shape
- upper right: color > shape, proximity
- middle: left (no clarity), right: connection > shape
- connection/fate, left-to-right (note continuity)

Impact on graph decoding

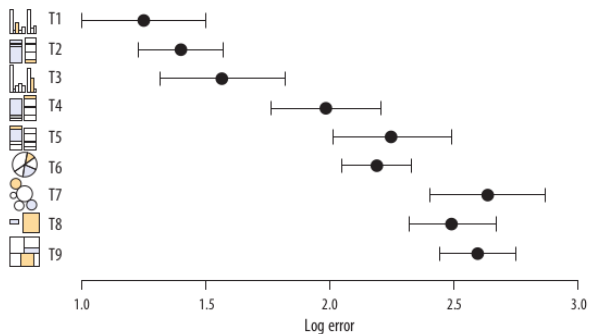


Cleveland & McGill, 1984, 1987, Heer & Bostock 2010

Impact on graph decoding



Crowdsourced results



Impact on graph decoding

- we do best with relative position aligned on a common scale

Impact on graph decoding

- we do best with relative position aligned on a common scale
- when elements are not aligned but still share a scale, comparison is a little harder

Impact on graph decoding

- we do best with relative position aligned on a common scale
- when elements are not aligned but still share a scale, comparison is a little harder
- it is more difficult again to compare the lengths of lines without a common baseline

Impact on graph decoding

- we do best with relative position aligned on a common scale
- when elements are not aligned but still share a scale, comparison is a little harder
- it is more difficult again to compare the lengths of lines without a common baseline
- we misjudge angles and areas

Impact on graph decoding

- we do best with relative position aligned on a common scale
- when elements are not aligned but still share a scale, comparison is a little harder
- it is more difficult again to compare the lengths of lines without a common baseline
- we misjudge angles and areas
- we're even worse with the change of slope

Re-thinking channels

Re-thinking channels

- the channels has to be able to capture the values properly (e.g. avoid gradient scale with categorical data?)

Re-thinking channels

- the channels has to be able to capture the values properly (e.g. avoid gradient scale with categorical data?)
- try to choose the most effective channels (e.g. avoid encoding numbers as areas)

Re-thinking channels

- the channels has to be able to capture the values properly (e.g. avoid gradient scale with categorical data?)
- try to choose the most effective channels (e.g. avoid encoding numbers as areas)
- given a channel, error rate depends on minor choices (e.g. wrong sequence of colors)

Clutter and gestalt

Signal-to-noise ratio

- you're fighting for the viewer's attention!
- eliminate redundant cognitive load!
- Remembering gestalt principles may help here

Proximity



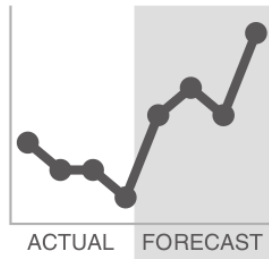
Separate by empty space to group, no need to draw anything more

Similarity



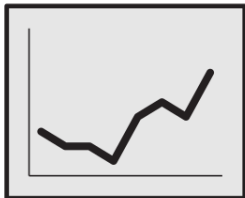
Use similarity to capture additional grouping

Enclosure



Enclosure is even stronger, use sparingly

Closure



Often borders and backgrounds are unnecessary

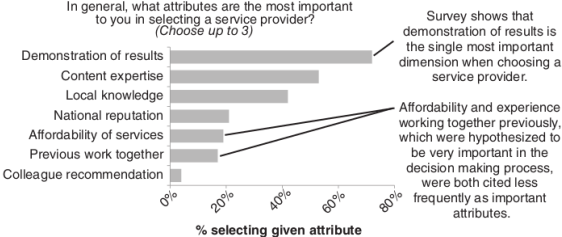
Continuity



Avoid lines which can be obtained by continuity

Lack of visual order

Demonstrating effectiveness is most important consideration when selecting a provider



Data source: xyz; includes N number of survey respondents. Note that respondents were able to choose up to 3 options.

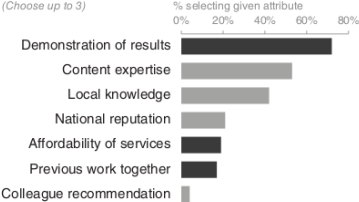
No channels used to introduce order

Lack of visual order

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)



Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

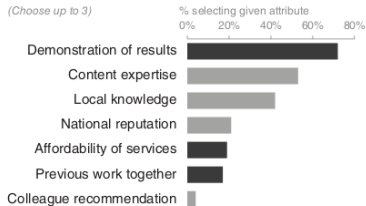
Ordered by various channels

Lack of visual order

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)



Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

Ordered by various channels

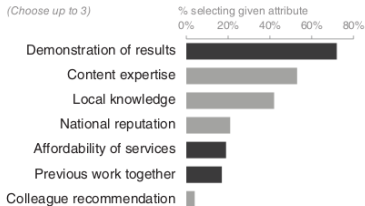
- notice left-to-right, top-to-bottom

Lack of visual order

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)



Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

Ordered by various channels

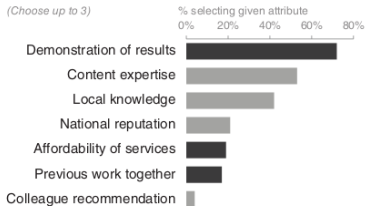
- notice left-to-right, top-to-bottom
- notice how dropping diagonal elements improves clarity

Lack of visual order

Demonstrating effectiveness is most important consideration when selecting a provider

In general, **what attributes are the most important** to you in selecting a service provider?

(Choose up to 3)



Survey shows that **demonstration of results** is the single most important dimension when choosing a service provider.

Affordability and **experience working together previously**, which were hypothesized to be very important in the decision making process, were both cited less frequently as important attributes.

Data source: xyz; includes N number of survey respondents.
Note that respondents were able to choose up to 3 options.

Ordered by various channels

- notice left-to-right, top-to-bottom
- notice how dropping diagonal elements improves clarity
- Same applies to text: the reading of rotated text 45 degrees is 52% slower (text rotated 90 degrees in either direction is 205% slower).

White space

Never add data just for the sake of adding data

Only add data with a thoughtful and specific purpose in mind!

Contrast

It's easy to spot a hawk in a sky full of pigeons, but as the variety of birds increases, that hawk becomes harder and harder to pick out.

(Colin Ware, *Information Visualization: Perception for Design*, 2004)

Contrast

It's easy to spot a hawk in a sky full of pigeons, but as the variety of birds increases, that hawk becomes harder and harder to pick out.

(Colin Ware, *Information Visualization: Perception for Design*, 2004)



What's the lesson here?

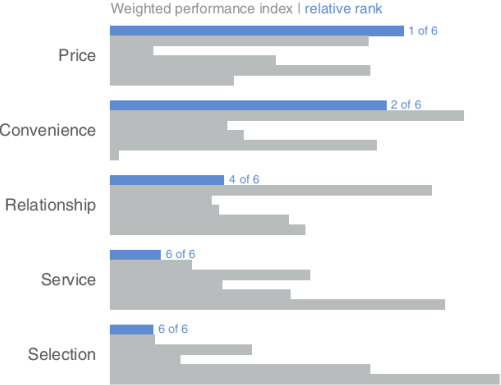
Contrast

It's easy to spot a hawk in a sky full of pigeons, but as the variety of birds increases, that hawk becomes harder and harder to pick out.

(Colin Ware, *Information Visualization: Perception for Design*, 2004)

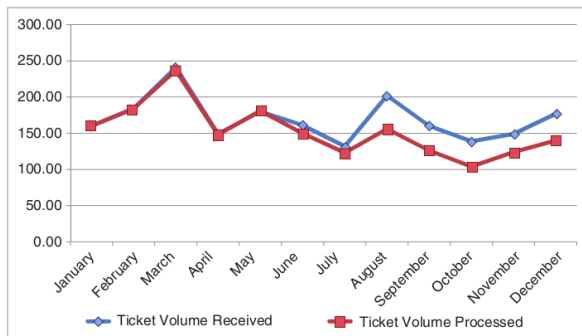
Performance overview

- **Our business**
- Competitor A
- Competitor B
- Competitor C
- Competitor D
- Competitor E



Better?

Decluttering: a case study



Initial visualization

Decluttering: a case study

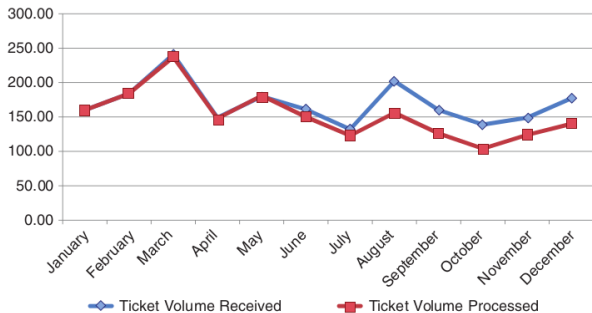
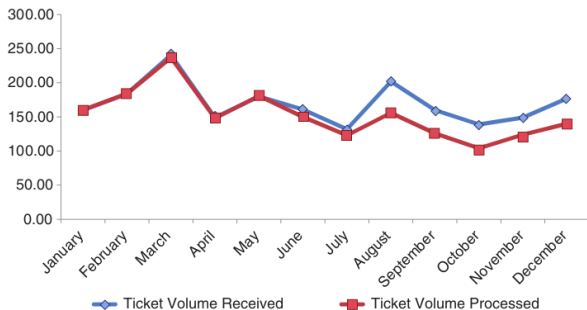


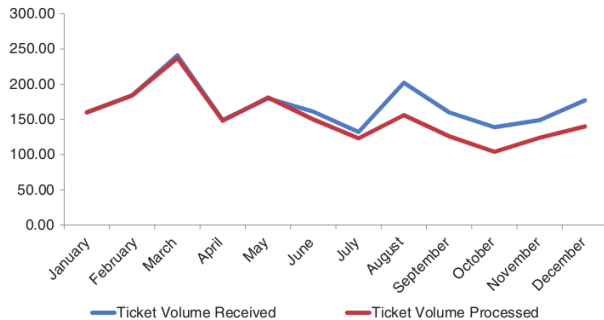
Chart borders were redundant

Decluttering: a case study



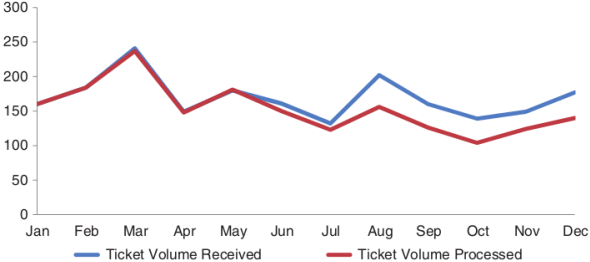
Grid lines only if specific values are essential

Decluttering: a case study



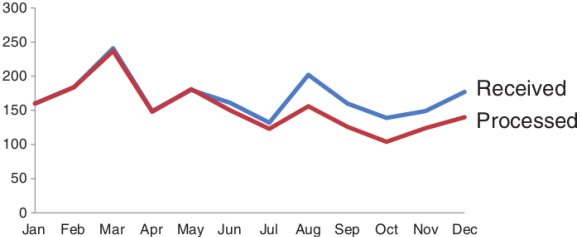
Data markers add no content

Decluttering: a case study



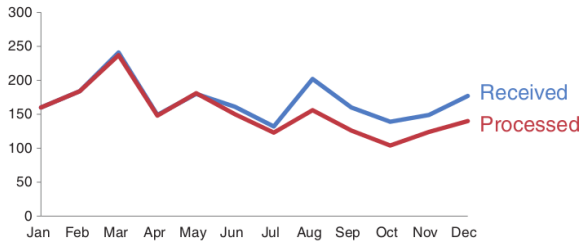
Clean up axis labels

Decluttering: a case study



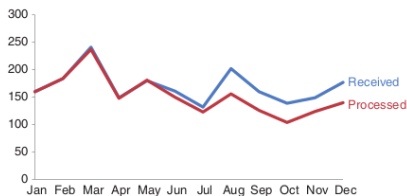
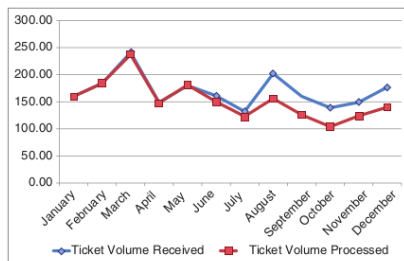
Label data directly

Decluttering: a case study



Leverage consistent colors

Decluttering: a case study



Before & after

Getting started with R, RStudio and ggplot2

More on what to show

Importance of context

Exploratory visualisation

- Not much care to the fine details
- Multiple visualizations for yourself before you find the pearl

Importance of context

Exploratory visualisation

- Not much care to the fine details
- Multiple visualizations for yourself before you find the pearl

Explanatory visualisation

- Don't show them everything!
- Focus on key messages and polish their presentation

Who, what, how

Who are you addressing?

- Find common ground, identify how much you can assume
- Communicating to too many disparate audiences you will fail
- Do they think you know what you're doing, or do you have to convince them?

Who, what, how

What do you want them to learn?

- First, three-minute story: before producing a graph, come up with a short elevator pitch for what you want to convey

Who, what, how

What do you want them to learn?

- First, three-minute story: before producing a graph, come up with a short elevator pitch for what you want to convey
- Next: a big picture statement: articulate your unique point of view, convey what's at stake, make it a complete sentence

Who, what, how

What do you want them to learn?

- First, three-minute story: before producing a graph, come up with a short elevator pitch for what you want to convey
- Next: a big picture statement: articulate your unique point of view, convey what's at stake, make it a complete sentence
- **Only then**, prepare the visualization, keeping these in mind

Who, what, how

How will you communicate?

- Live presentation?
- Written text?
- just the visualization?

Who, what, how

How will you communicate?

- Live presentation?
- Written text?
- just the visualization?
- The less control you have, the more details you need!

Who, what, how

How will you communicate?

- Live presentation?
- Written text?
- just the visualization?
- The less control you have, the more details you need!

If talking

Know your stuff and practice, practice, practice! Never read!

Choosing the visual

Embarrassment of riches

Out of hundreds of methods, only 10-20 are really good.
The rest is fluff.

Choosing the visual

91%

Simple text



Scatterplot

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

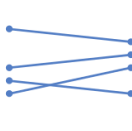
Table



Line

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

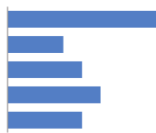


Slopegraph

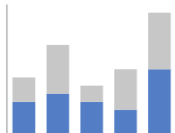
Choosing the visual



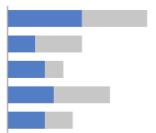
Vertical bar



Horizontal bar



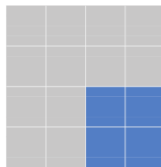
Stacked vertical bar



Stacked horizontal bar



Waterfall



Square area

Simple text

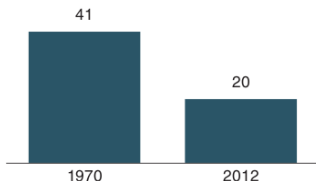
Key strategy

- Focus on the number(s)
- Perhaps add a few supporting words
- Messing with more you will lose the oomph

Simple text

Children with a "Traditional" Stay-at-Home Mother

% of children with a married stay-at-home mother with a working husband



Note: Based on children younger than 18. Their mothers are categorized based on employment status in 1970 and 2012.

Source: Pew Research Center analysis of March Current Population Surveys Integrated Public Use Microdata Series (IPUMS-CPS), 1971 and 2013

Adapted from PEW RESEARCH CENTER

Stay-at-home moms (original)

- Lots of space lost on graphing two data points
- Lot of detailed commentary that can be said, moved to a footnote or the figure description
- What do you think about “The number of children having a traditional stay-at-home mom decreased more than 50% between 1970 and 2012”?

20%

of children had a
traditional stay-at-home mom
in 2012, compared to 41% in 1970

Stay-at-home moms, remade

Tables

Good for

- communicating to a mixed audience whose members might be interested in different rows
- multiple different units of measurement

Tables

Good for

- communicating to a mixed audience whose members might be interested in different rows
- multiple different units of measurement

Bad for

- Live presentation
- A more narrative take

Tables

Key rule

Let the data get the attention

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	\$X.X	Y%	Z,ZZZ
Group 2	\$X.X	Y%	Z,ZZZ
Group 3	\$X.X	Y%	Z,ZZZ
Group 4	\$X.X	Y%	Z,ZZZ
Group 5	\$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	\$X.X	Y%	Z,ZZZ
Group 2	\$X.X	Y%	Z,ZZZ
Group 3	\$X.X	Y%	Z,ZZZ
Group 4	\$X.X	Y%	Z,ZZZ
Group 5	\$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	\$X.X	Y%	Z,ZZZ
Group 2	\$X.X	Y%	Z,ZZZ
Group 3	\$X.X	Y%	Z,ZZZ
Group 4	\$X.X	Y%	Z,ZZZ
Group 5	\$X.X	Y%	Z,ZZZ

Avoid heavy borders

Heatmap

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

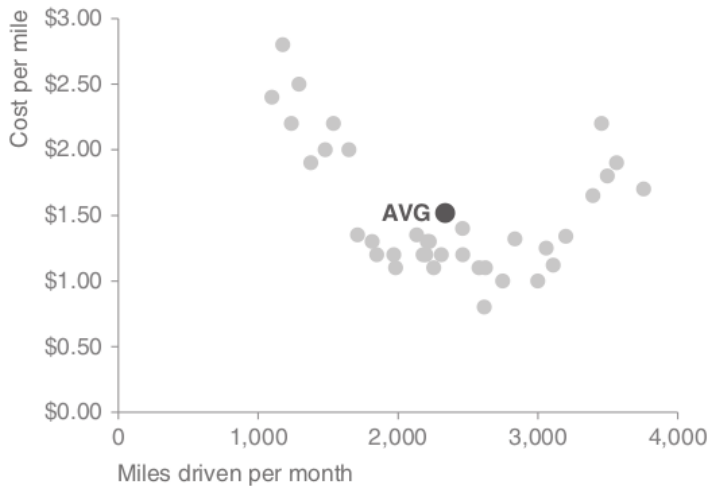
LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Leverage color saturation to convey relative magnitude

Scatterplot

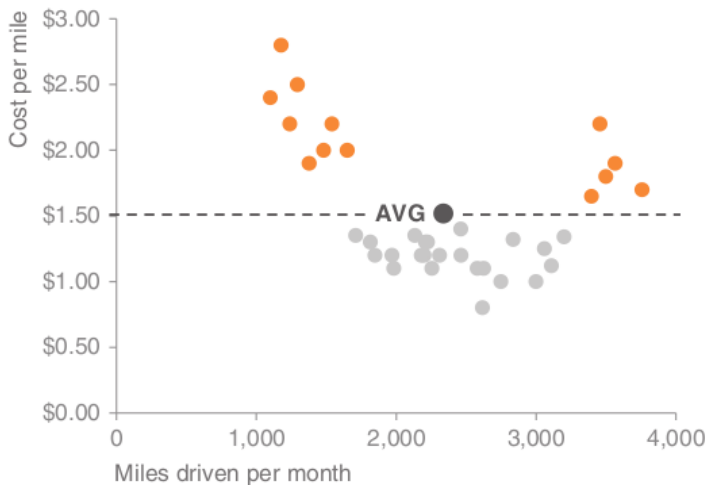
Cost per mile by miles driven



Original scatterplot

Scatterplot

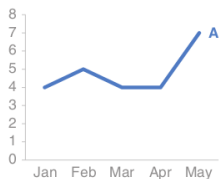
Cost per mile by miles driven



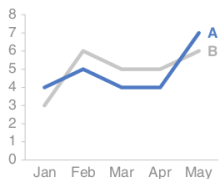
Use colors and a line to make a point

Line graph

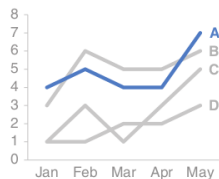
Single series



Two series



Multiple series

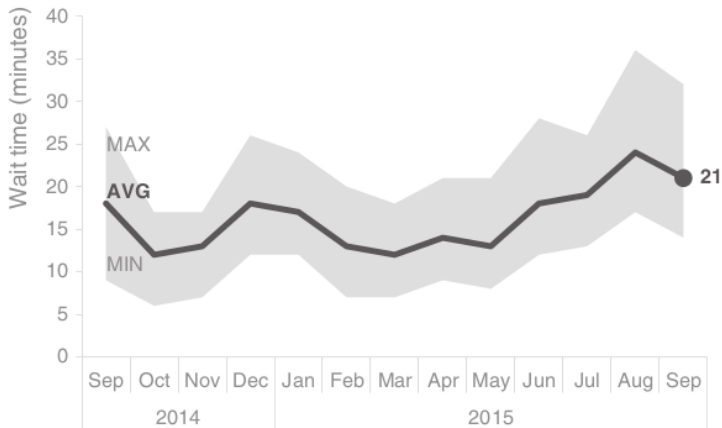


Single or multiple series with color for emphasis, note consistent intervals

Line graph

Passport control wait time

Past 13 months



If showing a summary with a range, be clear about what you're showing

Slopegraph

Employee feedback over time



Use for two time periods or paired sets of for comparison

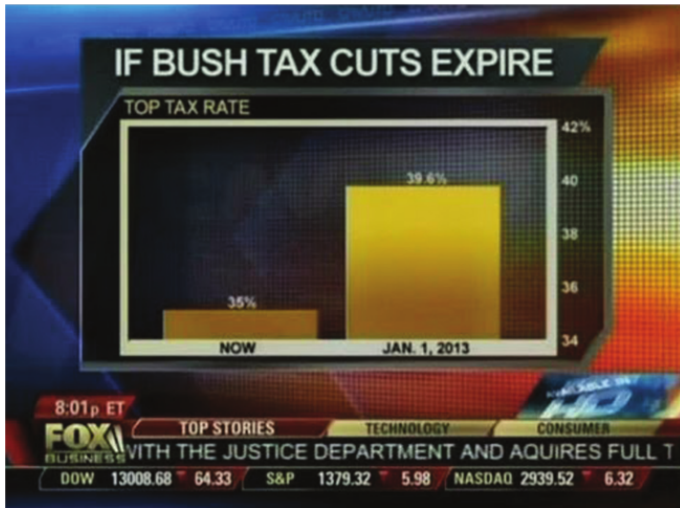
Slopegraph

Employee feedback over time



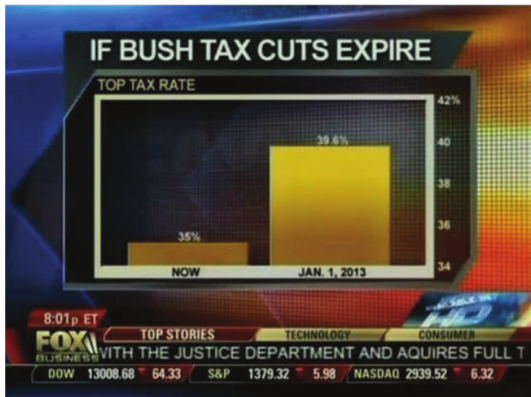
Use color for emphasis

Barplots for categorical data



Lack of zero baseline leads to false visual comparison (Fox News)

Barplots for categorical data



The visual increase is 460%, the actual increase is 13%, lie ratio of 35.38

$$35 - 34 = 1$$

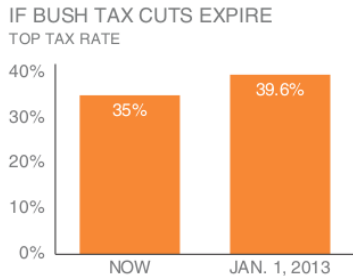
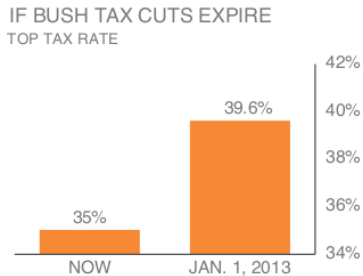
$$39.6 - 34 = 5.6$$

$$5.6 - 1 = 4.6$$

$$4.6/1 = 4.6$$

$$(39.6 - 35)/35 = .13$$

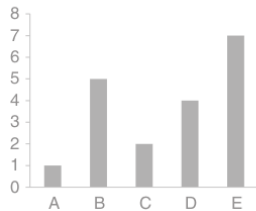
Barplots for categorical data



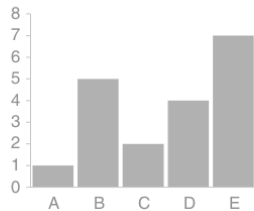
Note y axis moved to the left, labels pulled inside

Barplots for categorical data

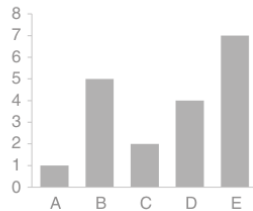
Too thin



Too thick



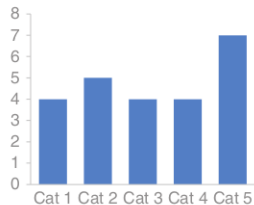
Just right



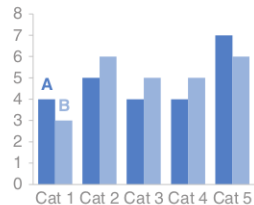
Balancing the width

Vertical bar chart

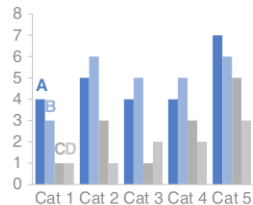
Single series



Two series



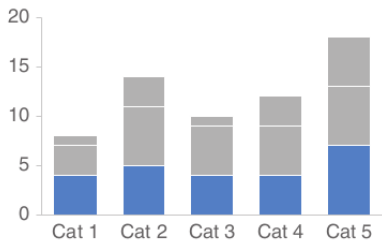
Multiple series



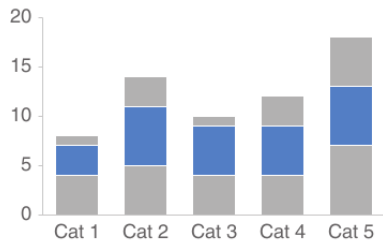
Adding series becomes messy; if you really do this, use color for emphasis

Stacked bar chart

Comparing **these** is easy



Comparing **these** is hard

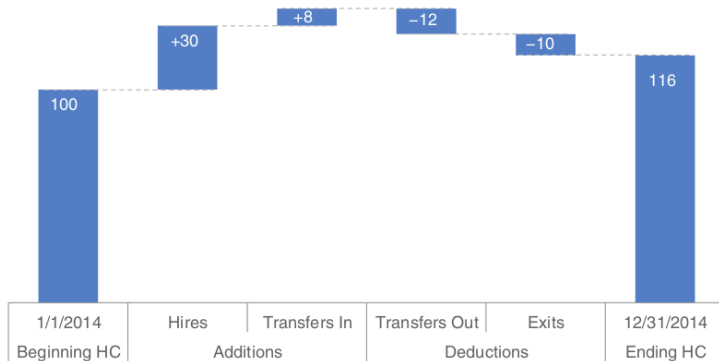


Only if you really care about the total

Waterfall chart

2014 Headcount math

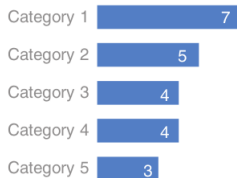
Though more employees transferred out of the team than transferred in, aggressive hiring means overall headcount (HC) increased 16% over the course of the year.



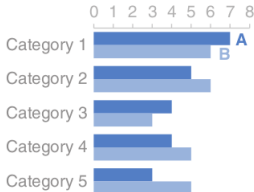
If you want to focus on intermediate changes

Horizontal barplot

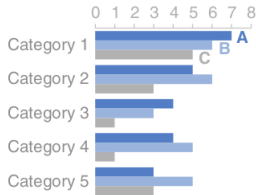
Single series



Two series



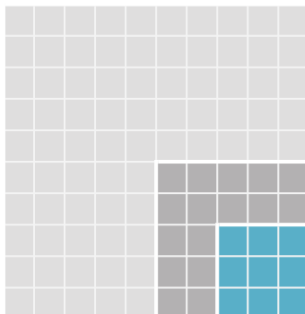
Multiple series



Easy to read if category names are longer

Area graph

Interview breakdown



Out of every **100**
phone screens...

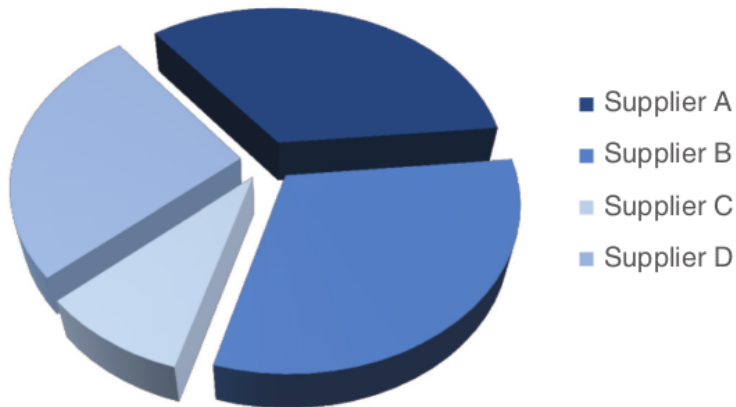
we bring **25**
candidates onsite
for interviews...

and
extend 9 offers.

Avoid, unless you visualize vastly different numbers

Pie charts are evil

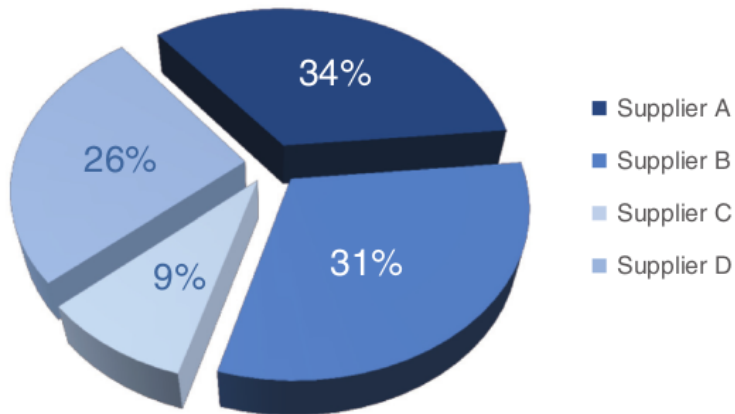
Supplier Market Share



Which supplier is the largest? What's your percentage estimate?

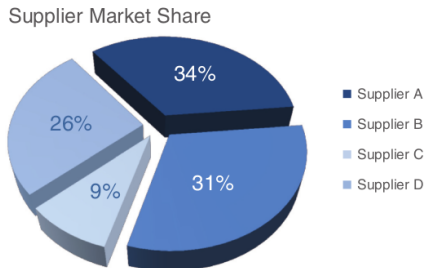
Pie charts are evil

Supplier Market Share



Now with labels

Pie charts are evil



Now with labels

What's wrong?

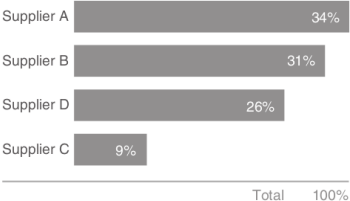
- Don't use 3D!
- Even without 3D, we're bad with angles!
- If you need the labels to avoid confusion, the visualization failed

Pie charts are evil

Supplier Market Share



Supplier Market Share

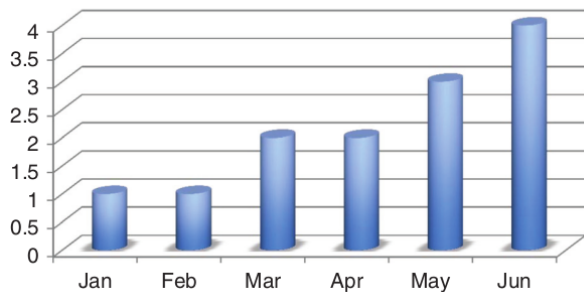


Total 100%

What to do instead

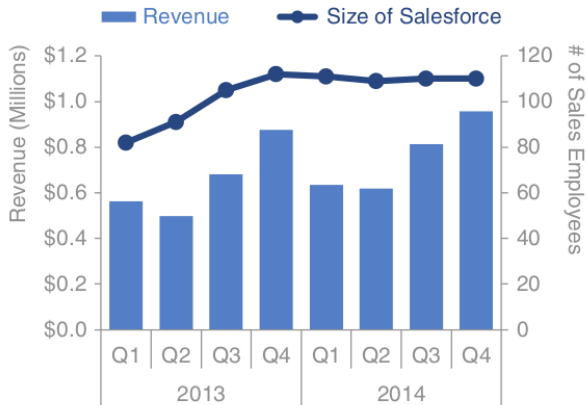
Don't use 3D

Number of issues



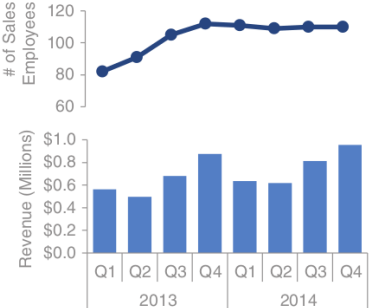
What are the actual values?

Don't use secondary y-axis



This is hard to read without confusion

Don't use secondary y-axis



Label directly or pull apart using the same x-axis; note you imply a connection!

Leverage focus

Preattentive attributes

756395068473

658663037576

860372658602

846589107830

Count threes here

Preattentive attributes

756**3**9506847**3**

65866**3**0**3**7576

860**3**72658602

8465891078**3**0

Count threes now

Preattentive attributes



Orientation



Shape



Line length



Line width



Size



Curvature



Added marks



Enclosure



Hue



Intensity



Spatial position



Motion

Various preattentive attributes

Preattentive attributes in text

No preattentive attributes

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Color

What are we doing well? Great Products. **These products are clearly the best in their class.** Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Bold

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Italics

What are we doing well? Great Products. These products are clearly the best in their class. *Replacement parts are shipped when needed.* You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.
You have a great company – keep up the good work!

Notice the difference in grade of attention

Preattentive attributes in text

Size

What are we doing well? Great Products. These products are the best in their class. Replacement parts are shipped when needed. You sent gaskets

without me having to

ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours. You have a great company – keep up the good work!

Outline (enclosure)

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.

You have a great company – keep up the good work!

Separate spatially

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask.

Problems are resolved promptly.

Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours. You have a great company – keep up the good work!

Underline (added marks)

What are we doing well? Great Products. These products are clearly the best in their class. Replacement parts are shipped when needed. You sent me gaskets without me having to ask. Problems are resolved promptly. Bev in the billing office was quick to resolve a billing issue I had. General customer service exceeds expectations. The account manager even called to check in after normal business hours.

You have a great company – keep up the good work!

Notice the difference in grade of attention

Preattentive attributes in text

What are we doing well?

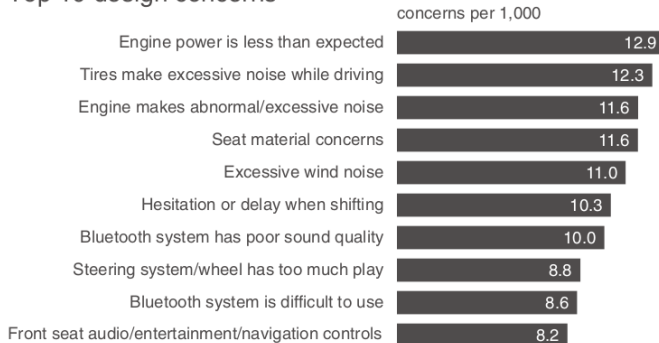
Themes & example comments

- **Great products:** "These products are clearly the best in class."
- **Replacement parts are shipped when needed:**
"You sent me gaskets without me having to ask, and I really needed them, too!"
- **Problems are resolved promptly:** "Bev in the billing office was quick to resolve a billing issue I had."
- **General customer service exceeds expectations:**
"The account manager even called after normal business hours.
You have a great company - keep up the good work!"

Create visual hierarchy

Preattentive attributes in graphs

Top 10 design concerns



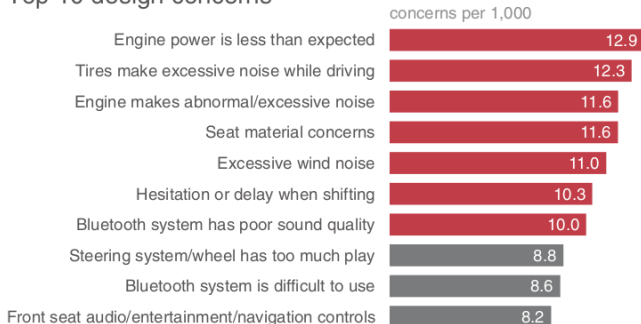
Original graph

Preattentive attributes in graphs

7 of the top 10 design concerns have 10 or more concerns per 1,000.

Discussion: is this an acceptable default rate?

Top 10 design concerns

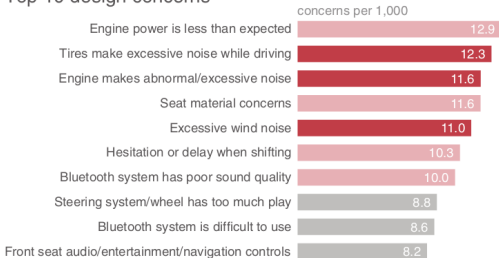


Show information with color

Preattentive attributes in graphs

Of the top design concerns, three are noise-related.

Top 10 design concerns



Comments indicate that **noisy tire issues** are most apparent **in the rain**.

Complaints about **engine noise** commonly cited **after the car had not been driven for a while**.

Excessive **wind noise** is noted primarily in **freeway driving at high speeds**.

Even more focus with a hierarchy

Preattentive attributes in graphs

Country Level Sales Rank Top 5 Drugs

Rainbow distribution in color indicates sales rank in given country from #1 (red) to #10 or higher (dark purple)

Country	A	B	C	D	E
AUS	1	2	3	6	7
BRA	1	3	4	5	6
CAN	2	3	6	12	8
CHI	1	2	8	4	7
FRA	3	2	4	8	10
GER	3	1	6	5	4
IND	4	1	8	10	5
ITA	2	4	10	9	8
MEX	1	5	4	6	3
RUS	4	3	7	9	12
SPA	2	3	4	5	11
TUR	7	2	3	4	8
UK	1	2	3	6	7
US	1	2	4	3	5

Top 5 drugs: country-level sales rank

RANK

1	2	3	4	5+
---	---	---	---	----

COUNTRY | DRUG

	A	B	C	D	E
Australia	1	2	3	6	7
Brazil	1	3	4	5	6
Canada	2	3	6	12	8
China	1	2	8	4	7
France	3	2	4	8	10
Germany	3	1	6	5	4
India	4	1	8	10	5
Italy	2	4	10	9	8
Mexico	1	5	4	6	3
Russia	4	3	7	9	12
Spain	2	3	4	5	11
Turkey	7	2	3	4	8
United Kingdom	1	2	3	6	7
United States	1	2	4	3	5

Use colors sparingly, after exploratory analysis

Preattentive attributes in graphs

A simple test

- Create your visual
- Close your eyes or look away
- Look back at it: where are your eyes drawn first?

Preattentive attributes in graphs

Things to pay attention to

- use colors consistently: change in colors suggests change in meaning!
- 8% of men and .5% of women are colorblind (no shades of red/ no shades of green)
- use vischeck.com to simulate what a colorblind person would see

Epistemic problems in data analysis

Key epistemic problems

Epistemology

The branch of philosophy that deals with the nature, origin, and scope of our knowledge.

Key epistemic problems

Epistemology

The branch of philosophy that deals with the nature, origin, and scope of our knowledge.

The usual epistemic flaws

- Assuming that the data we are using is a perfect reflection of reality
- Forming conclusions about the future based on historical data only
- Seeking to use data to verify a previously held belief rather than to test it to see whether it's actually false

Why care?

Car driving

We don't need to know how the car works to drive it!

Why care?

Car driving

We don't need to know how the car works to drive it!

Data analysis

This is more like cooking, you need to know what goes in and how it's combined!

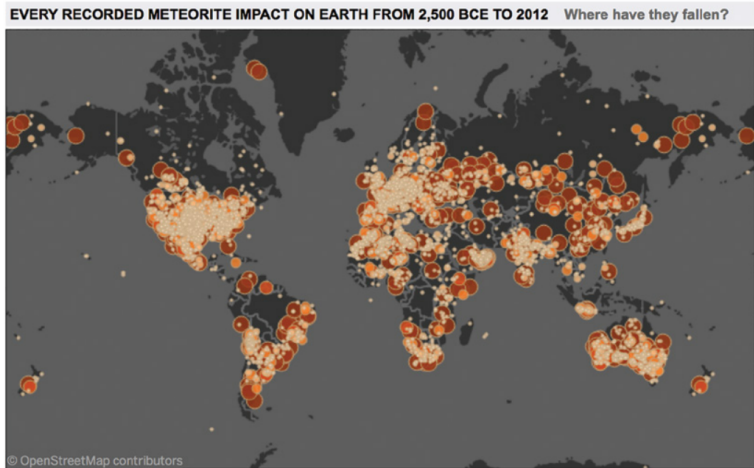
Data-reality gap

Examples

- It's not crime, it's reported crime.
- It's not the outer diameter of a mechanical part, it's the measured outer diameter.
- It's not how the public feels about a topic, it's how people who responded to the survey are willing to say they feel.

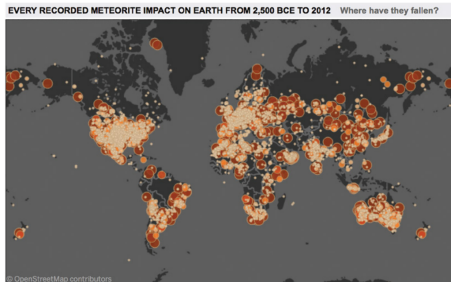
Meteorites

The Meteorological Society provides data for 34,513 meteorites that struck the surface of the earth between 2500 BCE and 2012.



Meteors landing (map by Ramon Martinez)

Meteorites

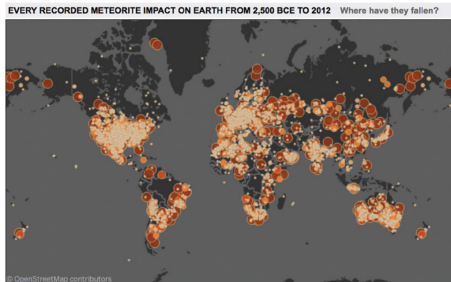


Meteors landing (map by Ramon Martinez)

Question

Why this doesn't tell us where meteorites are more likely to strike the Earth?

Meteorites



Meteorites landing (map by Ramon Martinez)

Question

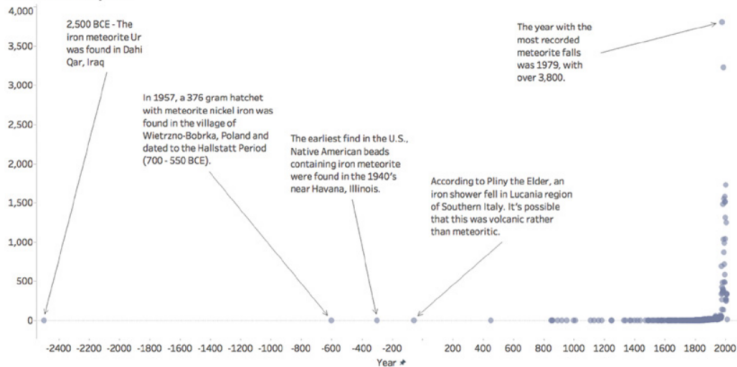
Why this doesn't tell us where meteorites are more likely to strike the Earth?

Answer

It tells us where meteorites are more likely to have fallen (in the past), and were observed by someone who reported it to someone who recorded it faithfully.

Meteors

Meteorite Falls by Year



Reported meteors landing in time

Earthquakes

The United States Geological Survey provides an Earthquake Archive Search.



Earthquakes

The United States Geological Survey provides an Earthquake Archive Search.

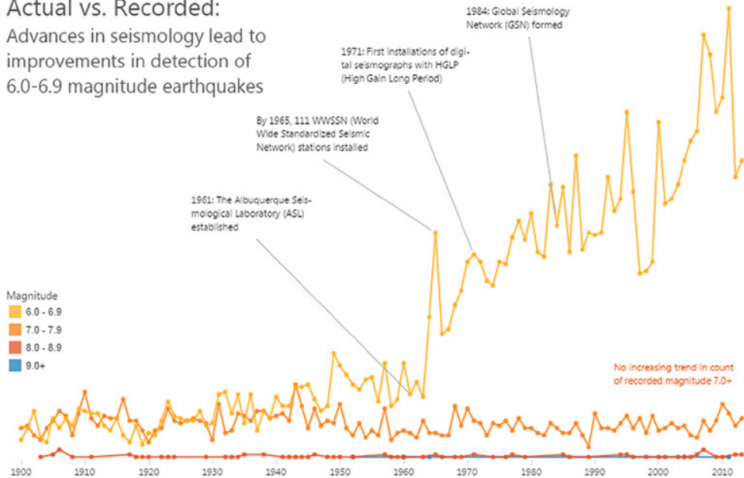


Question

Why isn't this a cause for alarm?

Earthquakes

Actual vs. Recorded:
Advances in seismology lead to
improvements in detection of
6.0-6.9 magnitude earthquakes



Sources | Data: <http://earthquake.usgs.gov/earthquakes/search/>, Dates: <http://pubs.usgs.gov/fs/2011/3065/pdf/FS11-3065.pdf>

Bicycles

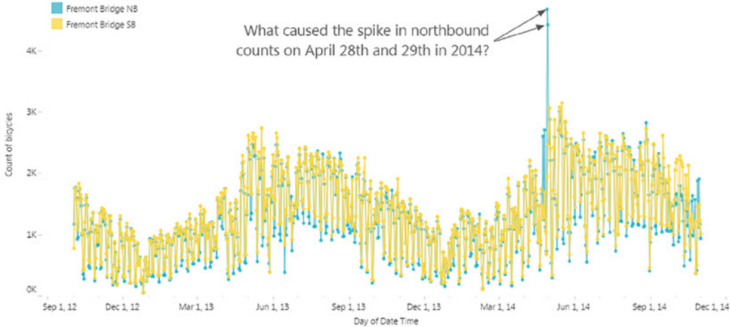
The City of Seattle Department of Transportation has installed two inductive loops on the pedestrian/bicycle pathways of the bridge.



Fremont Bridge, Seattle (the most opened drawbridge in the United States, 35/day)

Bicycles

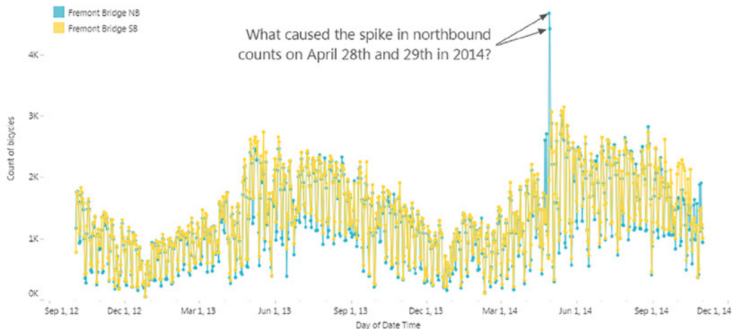
Fremont Bridge Bike Counter Time Series, Oct 2012 - Oct 2014



Data source: http://www.seattle.gov/transportation/bikecounter_fremont.htm

Bicycles

Fremont Bridge Bike Counter Time Series, Oct 2012 - Oct 2014

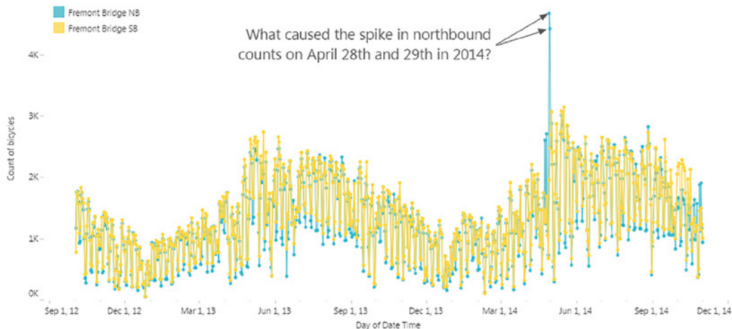


Data source: http://www.seattle.gov/transportation/bikecounter_fremont.htm

Think!

Bicycles

Fremont Bridge Bike Counter Time Series, Oct 2012 - Oct 2014



Data source: http://www.seattle.gov/transportation/bikecounter_fremont.htm

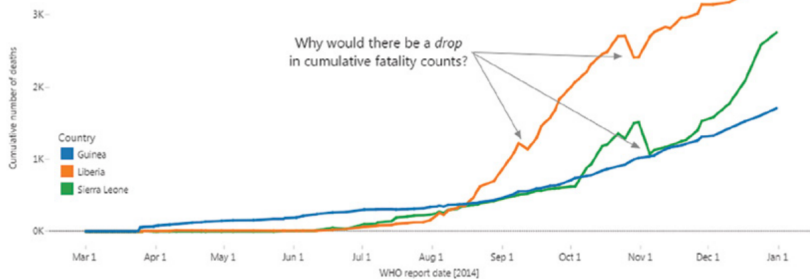
Think!

Equipment error

Now the dataset is fixed by averaging.

Ebola

Ebola deaths in West Africa, 2014

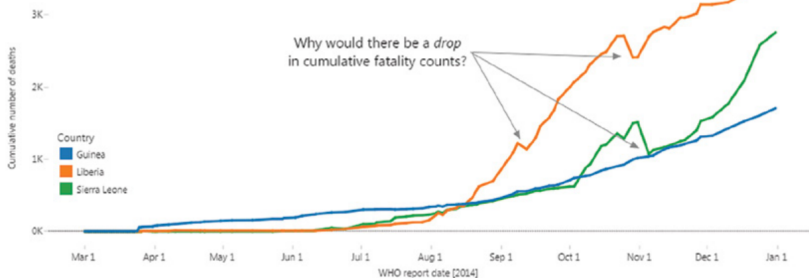


Data Source: <http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/cumulative-cases-graphs.html>

WHO fatalities count

Ebola

Ebola deaths in West Africa, 2014



Data Source: <http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/cumulative-cases-graphs.html>

WHO fatalities count

Important distinction

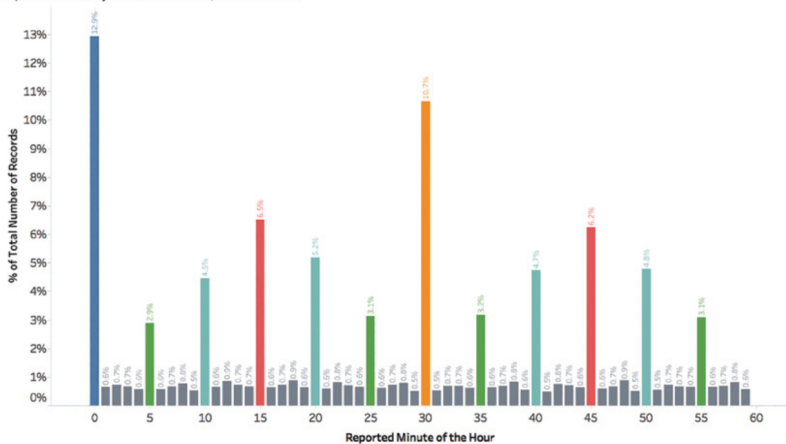
Suspected/probable/confirmed.

A wider perspective so far

- measurement systems change
- definitions change
- missing data
- misclassified data

The fudging

Reported strikes by minute of the hour, non-null values

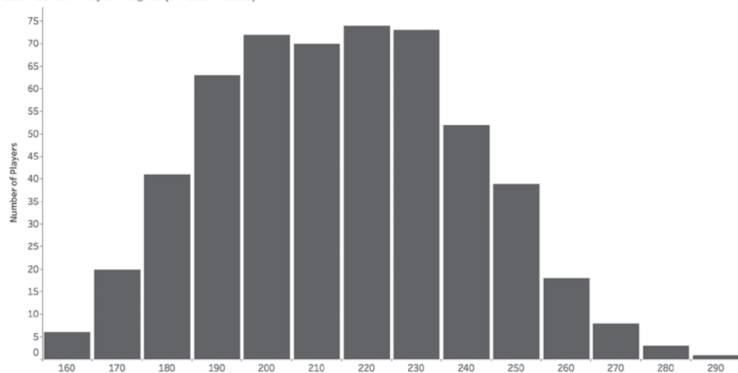


The number of minutes past the hour that pilots provide when they report to the FAA that their aircraft struck wildlife, n= 85k

Note the geometric regularity

The fudging

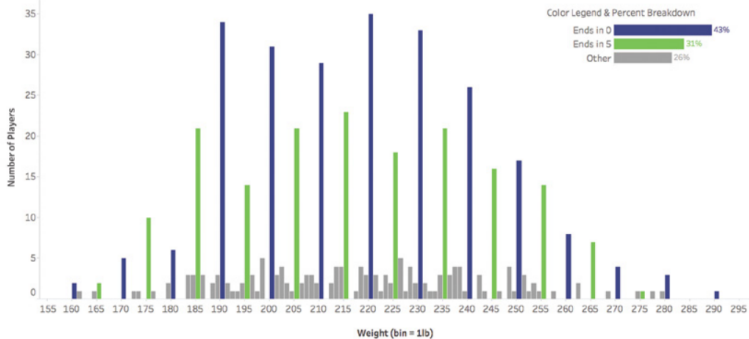
2017-18 NBA Player Weights (bin size = 10lbs)



This looks kinda normal, right?

The fudging

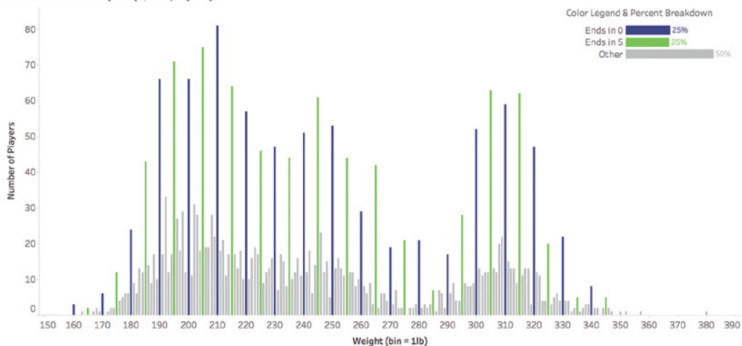
2017-18 NBA Player Weight (in lbs)



How about now?

The fudging

2018 NFL Active Players (2,875 players)

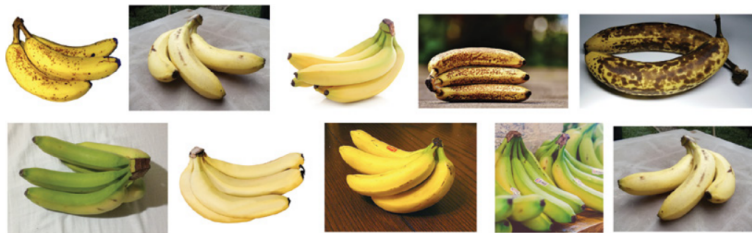


Another example, football players

Inconsistent ratings

The task (Ben Jones)

Rate a series of 10 banana photos on a ripeness scale:
unripe, almost ripe, ripe, very ripe, or overripe

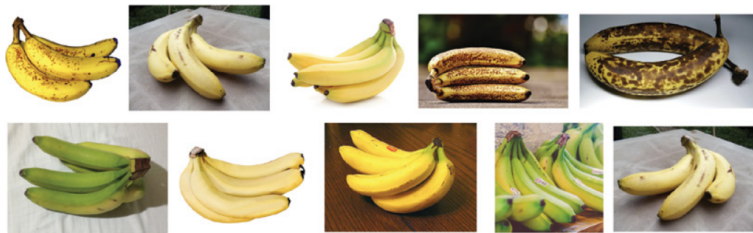


Images tested on 231 respondents; do you see anything tricky here?

Inconsistent ratings

The task (Ben Jones)

Rate a series of 10 banana photos on a ripeness scale:
unripe, almost ripe, ripe, very ripe, or overripe



Images tested on 231 respondents; do you see anything tricky here?

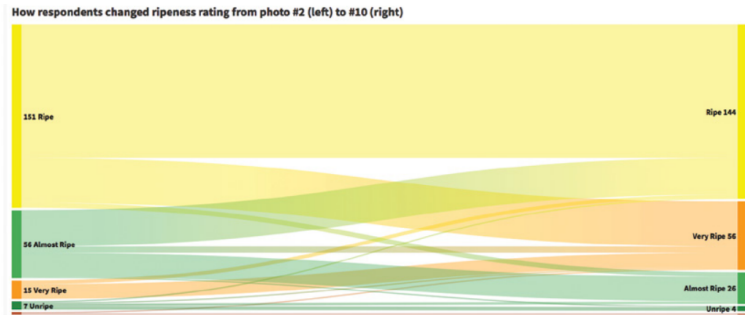
Look at bananas 2 and 10!

Inconsistent ratings



85 respondents had inconsistent ratings for the repeated banana.

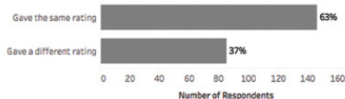
Inconsistent ratings



Sankey diagram of opinion change

Inconsistent ratings

The 10th photo was a mirror image of the 2nd photo. 37% of respondents give the mirror image a different ripeness level than they gave the original one. See how they changed their rating in the table below.



Here's the 2nd photo shown in the set, and how respondents rated it, broken down by how they rated the 10th photo:



Here's the 10th photo shown in the set, and how respondents rated it based on how they rated the 2nd photo:



	Unripe	Almost Ripe	Ripe	Very Ripe	Overripe	Total
Unripe	3	0	0	0	0	7
Almost Ripe	0	70	80	5	0	56
Ripe	0	4	110	37	0	151
Very Ripe	0	0	3	12	0	15
Overripe	0	0	0	0	2	2
	4	26	144	56	1	231

Inconsistent ratings



The ninth banana

General points here

- Our ratings and opinions have a degree of noise in them, even over short time horizons, and that we're possibly influenced to some degree by the context

General points here

- Our ratings and opinions have a degree of noise in them, even over short time horizons, and that we're possibly influenced to some degree by the context
- Every measurement system has some degree of error due to challenges with repeatability and reproducibility.

What to do?

Keep in mind!

Every data point that exists was collected, stored, accessed, and analyzed via imperfect processes by fallible human beings dealing with equipment that has built-in measurement error.

What to do?

Keep in mind!

Every data point that exists was collected, stored, accessed, and analyzed via imperfect processes by fallible human beings dealing with equipment that has built-in measurement error.

Do your homework!

The more we know about these processes—the equipment used, the protocol followed, the people involved, the steps they took, their motivations—the better equipped we will be to assess the data-reality gap.

What to do

Key steps

- Clearly understand the operational definitions of all metrics.

What to do

Key steps

- Clearly understand the operational definitions of all metrics.
- Draw the data collection steps as a process flow diagram.

What to do

Key steps

- Clearly understand the operational definitions of all metrics.
- Draw the data collection steps as a process flow diagram.
- Understand the limitations and inaccuracies of each step in the process.

What to do

Key steps

- Clearly understand the operational definitions of all metrics.
- Draw the data collection steps as a process flow diagram.
- Understand the limitations and inaccuracies of each step in the process.
- Identify any changes in method or equipment over time.

What to do

Key steps

- Clearly understand the operational definitions of all metrics.
- Draw the data collection steps as a process flow diagram.
- Understand the limitations and inaccuracies of each step in the process.
- Identify any changes in method or equipment over time.
- Seek to understand the motives of the people collecting and reporting. Could there be any biases or incentives involved?

What to do

Key steps

- Clearly understand the operational definitions of all metrics.
- Draw the data collection steps as a process flow diagram.
- Understand the limitations and inaccuracies of each step in the process.
- Identify any changes in method or equipment over time.
- Seek to understand the motives of the people collecting and reporting. Could there be any biases or incentives involved?
- Visualize the data and investigate any shifts, outliers, and trends for possible discrepancies.

Confirmation bias

How about...

... we use data to verify our hypotheses?

Confirmation bias

How about...

... we use data to verify our hypotheses?

No!

Focus in finding out what isn't true about our previously held conceptions about the world we live in, and to suggest additional questions for which we don't have any answers yet!

Confirmation bias

The induction step

We often assume that singular statements that we encounter in data verify universal truths, beyond the time, place, and conditions in which data were collected.

- It's not just how many times bikes crossed the Fremont bridge in April 2014, it's how many bikes cross the bridge in general.
- It's not just the preference of certain particular customers, it's the preference of all other potential customers as well.
- It's not just that the pilot manufacturing line had high yields during qualification, it's that the process will also have high yields at full volume production as well.
- It's not just that a particular mutual fund outperformed all others last year, it's that it'll be the best investment going forward.

Unfalsifiability

The problem

Either we form a hypothesis that isn't falsifiable, or we do our best to protect our hypothesis from any possible attempt to show it to be false.

Unfalsifiability

The problem

Either we form a hypothesis that isn't falsifiable, or we do our best to protect our hypothesis from any possible attempt to show it to be false.

Ask yourself

Do we actively seek to prove our own hypotheses to be false, to debunk our own myths, or do we mostly try to prove ourselves right and others wrong?

Leaps in reasoning

The faulty process

1. Basic question \Rightarrow
2. Data analysis \Rightarrow
3. Singular statement \Rightarrow (unaware of the inductive leap)
4. Belief in a universal statement

Leaps in reasoning

The faulty process

1. Basic question \Rightarrow
2. Data analysis \Rightarrow
3. Singular statement \Rightarrow (unaware of the inductive leap)
4. Belief in a universal statement

Example

1. A bicycle counter on the Fremont bridge! Let's learn about ridership in my city.
2. Okay, I found some data from the Seattle Department of Transportation, and it looks like...
3. 49,718 crossed in the eastbound direction, and 44,859 crossed headed west in April 2014.
4. Hmm, so more bicycles cross the bridge headed east than west, then. I wonder why that is? Maybe some riders cross to get to work in the morning but ride the bus home.

Leaps in reasoning

A better process

1. Basic question \Rightarrow
2. Data analysis \Rightarrow
3. Singular statement \Rightarrow
4. Falsifiable universal statement hypothesis \Rightarrow
5. An honest attempt to disprove it

Leaps in reasoning

A better process

1. Basic question \Rightarrow
2. Data analysis \Rightarrow
3. Singular statement \Rightarrow
4. Falsifiable universal statement hypothesis \Rightarrow
5. An honest attempt to disprove it

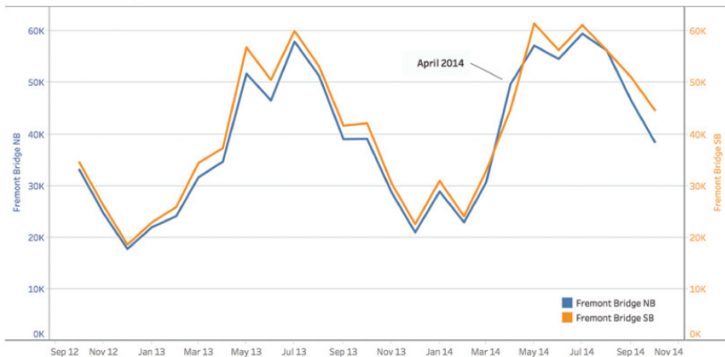
Example

⋮

4. Hmm, so the counters registered higher counts in the eastbound direction as compared to westbound that month. I wonder whether all months have seen higher counts going east as opposed to west?
5. Let me see whether that's not the case.

Leaps in reasoning

Fremont Bridge Bike Counter Measurements



The hypothesis was false, and the differences are minor

Technical and mathematical problems

Data wrangling

What is it?

- Pre-processing raw data to obtain something susceptible to visualisation and analysis.
- Not sexy, but important.
- 50-80% of the work.

Data wrangling

What is it?

- Pre-processing raw data to obtain something susceptible to visualisation and analysis.
- Not sexy, but important.
- 50-80% of the work.

Every data is dirty

- misspelled text values
- date format issues
- mismatching units
- missing values
- null values
- incompatible geospatial coordinate formats
- ...

Data wrangling

The Baltimore City Department of Transportation provides a downloadable record of over 61300 car tow events dating from January 2017 back to October 2012.

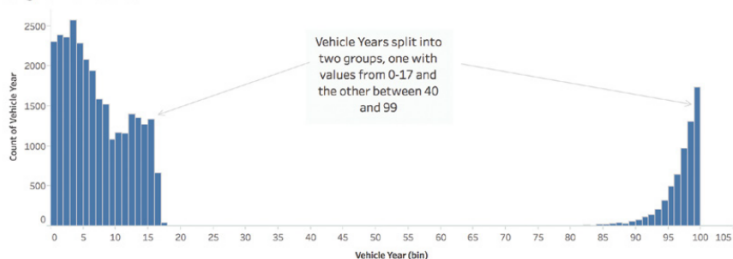
	A	B	C	D	E	F	G	H	I	J	K
1	propertynumber	towedDateTime	vehicleType	vehicleYear	vehicleMake	vehicleModel	vehicleColor	tagNumber	towCompany	towCharge	towedFromLocation
2	P206813	10/23/10 10:50	Car	99	Mercedes	C230	Burg	7EVM54	Jim Elliotts Towing	\$140.00	200 Longwood Rd
3	P206814	10/23/10 11:50	Car	95	Lexus	L5400	Gray	EXV9405	Bermans Towing	\$140.00	700 W Fayette St
4	P206815	10/23/10 11:35	Car	4	Chevrolet	Cavalier	Blue	9ERW87	Frankford Towing	\$130.00	500 Grundy St
5	P206816	10/23/10 12:04	Scoter	8	Velocity		Black		Bermans Towing	\$140.00	2100 North Ave
6	F01135	10/24/10 12:38	Van		LEXUS			9CAA97	City	\$130.00	LI/B W HLIGHTS ST.
7	P206905	10/25/10 11:12	SUV	6	Toyota	RAV4	Blue	410M804	Cherryhill Towing Service	\$140.00	200 Fredhilton Pass
8	P206914	10/25/10 14:49	Car	97	Hyundai	Tiburon	Red	8EEZ91	City	\$140.00	1 N Paca St
9	P207054	10/25/10 14:53	Car	99	Honda	Accord	Burgundy	A219155	Fallsway	\$140.00	600 N Caroline St
10	P209809	12/20/10 8:41	SUV	0	Jeep	Cherokee	White	27415M5	Fallsway	\$130.00	200 Monroe St
11	P209807	12/20/10 16:45	Car	93	Honda	Accord	Brown	4EL575	Fallsway	\$130.00	1400 E Monument St
12	P209808	12/21/10 7:37	Car	95	Bmw	318i	White	4EDT18	Fallsway	\$130.00	100 S Greene St
13	P209775	12/22/10 12:35	Car	98	Pontiac	Grand Prix	Red	3F5H05	City	\$130.00	3719 Greenmount Ave
14	P209776	12/22/10 12:41	Car	0	Nissan	Maxima	Black	9C0CD55	Bermans Towing	\$140.00	1400 Russell St
15	P209777	12/22/10 12:45	Van	97	Mercury	Villager	Green		Bermans Towing	\$140.00	500 N Carey St
16	P209778	12/22/10 13:10	Car	93	Mitsubishi	Diamante	Silver		Aarons Automotive Services	\$130.00	900 E 23rd St
17	P209779	12/22/10 13:26	Pick-up Truc	3	Ford	F350	Black	83S213	Aarons Automotive Services	\$130.00	2100 N Wolfe St
18	P209780	12/22/10 13:30	Van	99	Chevrolet	Astro	White		City	\$130.00	2000 Ellsworth St
19	P209781	12/22/10 13:37	Car	0	Dodge	Stratus	Silver	9FJC68	Frankford Towing	\$130.00	1500 E Belvedere Ave
20	P209782	12/22/10 14:15	Pick-up Truc	91	Ford	F150	Red/Silver	48X235	City	\$130.00	200 S Ellwood Ave
21	P209783	12/22/10 14:26	Car	98	Honda	Accord	Black	9AC4502	Aarons Automotive Services	\$130.00	2800 Harford Rd
22	P209785	12/22/10 14:36	Car	98	Buick	Lesabre	Tan	7AA3187	City	\$140.00	1600 Gaywnn Falls Parkway
23	P209786	12/22/10 14:38	Car	99	Ford	Taurus	Black	7AD3025	Frankford Towing	\$130.00	500 N Luzerne
24	P209788	12/22/10 14:40	Trailer	?	Ez Loader	Hydra-Sports	Silver	AA67474	City	\$130.00	4020 Belle Ave
25	P209784	12/22/10 14:40	Boat	79	Sportcraft	Caprice	White	1703PN	City	\$130.00	4020 Belle Ave
26	P209787	12/22/10 16:57	SUV	5	Lexus	RX330	Silver	33742CB	Frankford Towing	\$130.00	3000 Myfield

Head of the tow data

Data wrangling

Average year of manufacture: 23. What?

Original Vehicle Year

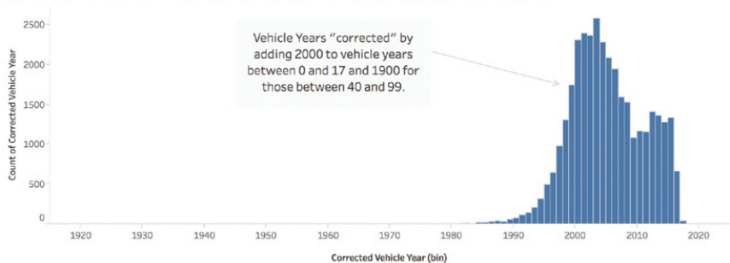


Add2000 to years between 0 and 17 and 1900 to years greater than that

Data wrangling

Long tail

Corrected Vehicle Year: Add 2000 to Years 0-17 and add 1900 to all other years

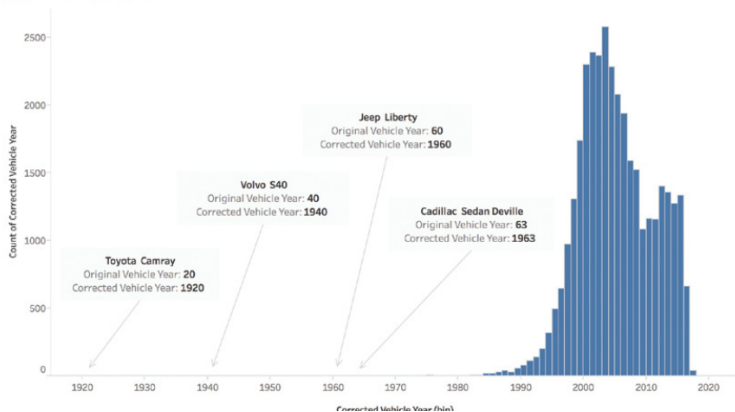


description

Data wrangling

Long tail

Outlier Vehicle Years



Check the outliers

Data wrangling

Misspelled makes



Chevrolet, Chevy, Cheverolet, Chevolet, Peterbilt, Peterbutt, Mitshubishi, Mitsubishit, ...

Data wrangling

Misspelled makes

Cluster & Edit column "vehicleMake"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method **key collision** Keying Function **ngram-fingerprint** Ngram Size **1** **113 clusters found**

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
21	5067	<ul style="list-style-type: none">Chevrolet (3636 rows)CHEVROLET (1366 rows)Chevrolet (30 rows)Chevrolet (7 rows)Chevrolet (5 rows)CHERYVOLET (3 rows)Chevrolet (3 rows)CHVROLET (2 rows)Chevrolet (2 rows)Chevrolet (2 rows)CCHEVROLET (1 rows)CHEVROLET (1 rows)CHEVRLT (1 rows)CHEVROELT (1 rows)CHEVROLET (1 rows)Chevrolet (1 rows)Chevrolet (1 rows)Chevrolet (1 rows)Chevrolet (1 rows)Chvrolet (1 rows)Chvrolet (1 rows)	<input type="checkbox"/>	Chevrolet
19	533	<ul style="list-style-type: none">Mitsubishi (369 rows)MITSUBISHI (132 rows)Mitsubishi (11 rows)Mitsubushi (4 rows)Mitubishi (2 rows)	<input type="checkbox"/>	Mitsubishi

Choices in Cluster
2 — 21

Rows in Cluster
0 — 7700

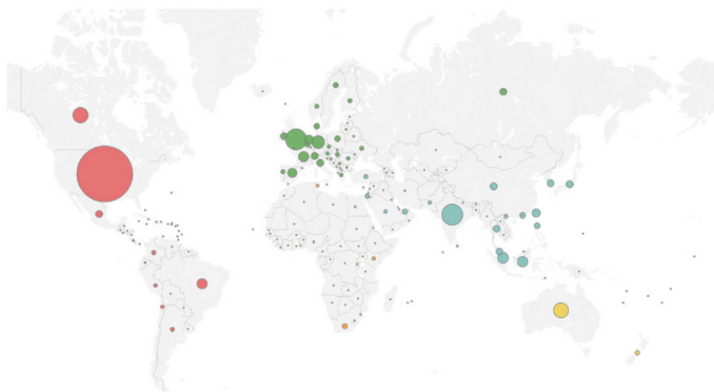
Average Length of Choices
1 — 15

Length Variance of Choices
0 — 1.7

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Open refine: from 899 to 507 makes

Data wrangling



Google Analytics map of website views; say we want to compare to population...

Data wrangling

Two population lists

- World Bank web, 2016 country population
- Wikipedia

Data Set

Number of Sets

3

Section Details

Set 1	Set 2	Set 3
Google Analytics	WorldBank	Wikipedia
180	228	234

So, how many countries are there?

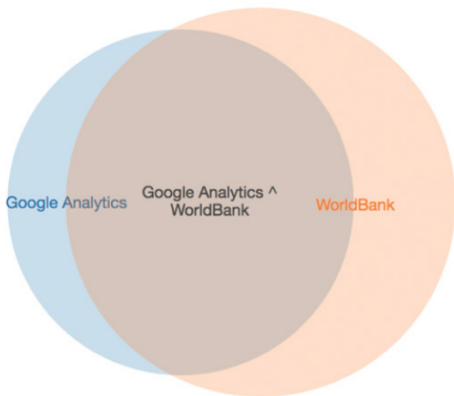
Data wrangling

- WB list contains 82 grouped values, can you do inner join?

Data wrangling

- WB list contains 82 grouped values, can you do inner join?

WorldBank List	Google Analytics List	Pageviews
Null	Antigua & Barbuda	27
	Bahamas	11,881
	Bosnia & Herzegovina	14,400
	Brunei	2,618
	Cape Verde	3,978
	Congo - Brazzaville	817
	Congo - Kinshasa	1,305
	Côte d'Ivoire	2,067
	Czechia	88,218
	Egypt	54,916
	Eritrea	457
	Gambia	330
	Guernsey	694
	Hong Kong	238,493
	Iran	53,667
	Jersey	589
	Kyrgyzstan	212
	Laos	1,627
	Macao	3,959
	Macedonia (FYROM)	4,386
	Martinique	2,043
	Myanmar (Burma)	21,493
	Palestine	1,506
	Polunon	6,170
	Russia	315,740
	Slovakia	34,755
	South Korea	313,568
	St. Kitts & Nevis	477
	Syria	771
	Taiwan	460,819
	Trinidad & Tobago	12,554
	U.S. Virgin Islands	175
	Venezuela	27,805
	Yemen	6,867



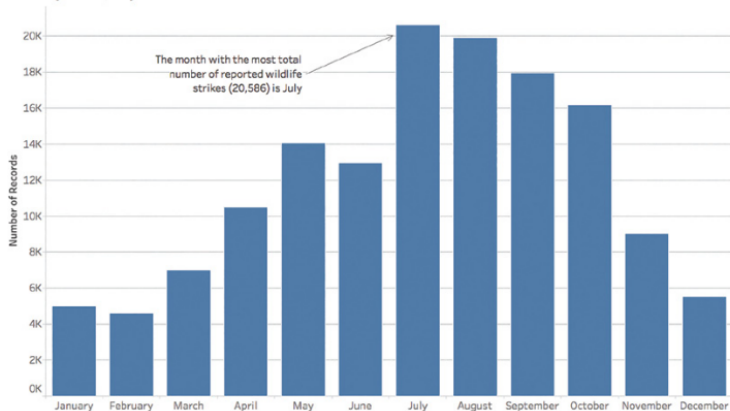
Differences in country lists

Mathematical problems

- Summing quantities to various levels of aggregation, such as buckets of time – the amount of some quantity per week, or month, or year
- Dividing quantities in our data with other quantities in our data to produce rates or ratios
- Working with proportions or percentages
- Converting from one unit of measure to another

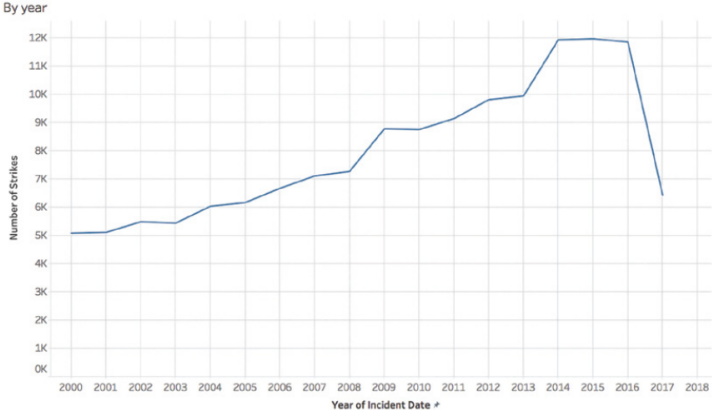
Mathematical problems

Strikes by month, all years



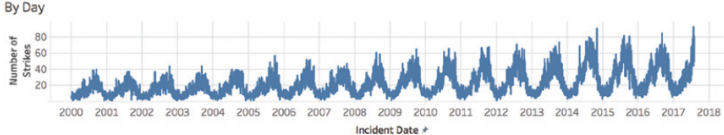
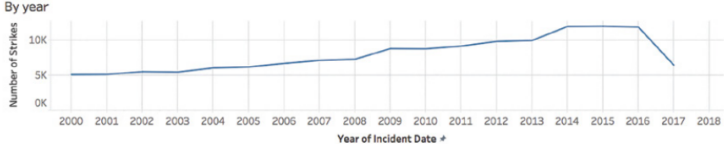
Recorded wildlife strikes by month (raw)

Mathematical problems



Timeline of recorded wildlife strikes

Mathematical problems



Granularity shift reveals the source of the problem

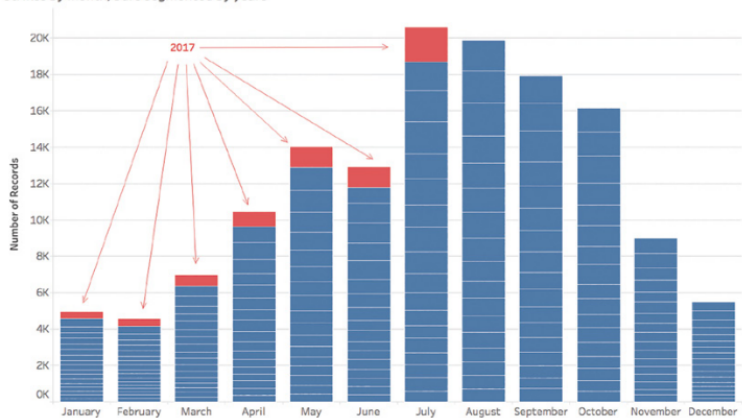
Mathematical problems



Cooks' strait (vs. Abel Tasman, 1642)

Mathematical problems

Strikes by month, bars segmented by years



Strikes again, now with attention

Mathematical problems

Infectious diseases contracted by California residents from 2001 through 2015, Center for Infectious Diseases, California Department of Public Health.

_id	Disease	County	Year	Sex	Count	Population	Rate	Cl.lower	Cl.upper	Unstable
1	Amebiasis	California	2001	Female	176	17339700	1.015	0.871	1.177	
2	Amebiasis	California	2001	Male	365	17173042	2.125	1.913	2.355	
3	Amebiasis	California	2001	Total	541	34512742	1.568	1.438	1.705	
4	Amebiasis	California	2002	Female	145	17554666	0.826	0.697	0.972	
5	Amebiasis	California	2002	Male	279	17383624	1.605	1.422	1.805	
6	Amebiasis	California	2002	Total	424	34938290	1.214	1.101	1.335	
7	Amebiasis	California	2003	Female	127	17782868	0.714	0.595	0.85	
8	Amebiasis	California	2003	Male	261	17606060	1.482	1.308	1.674	
9	Amebiasis	California	2003	Total	388	35388928	1.096	0.99	1.211	
10	Amebiasis	California	2004	Female	101	17968347	0.562	0.458	0.683	

Head of the diseases dataset

Mathematical problems

Question

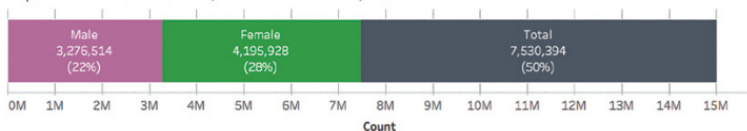
Are there more for male or female?

Mathematical problems

Question

Are there more for male or female?

Reported Infectious Diseases, California Residents, 2001-2015



Something's off! Look at the head again!

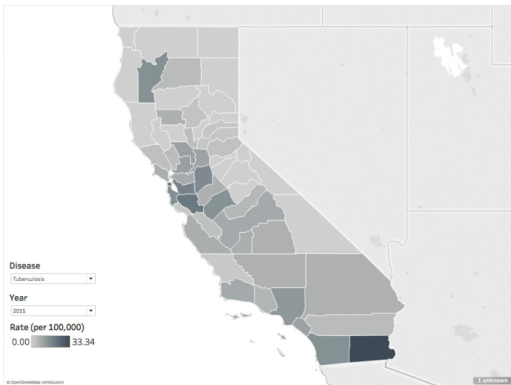
Mathematical problems

How are they distributed in the counties?

Mathematical problems

How are they distributed in the counties?

Reported Infectious Diseases, California Residents, by County

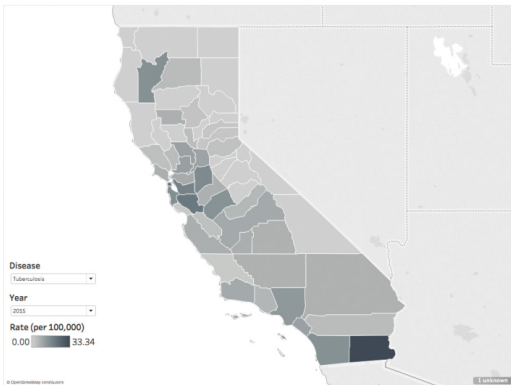


What's "1 unknown"?

Mathematical problems

How are they distributed in the counties?

Reported Infectious Diseases, California Residents, by County

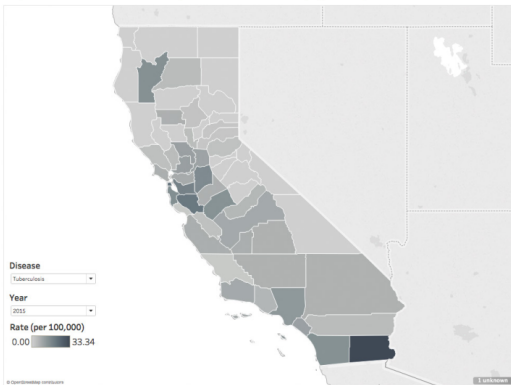


What's "1 unknown"? California!

Mathematical problems

How are they distributed in the counties?

Reported Infectious Diseases, California Residents, by County



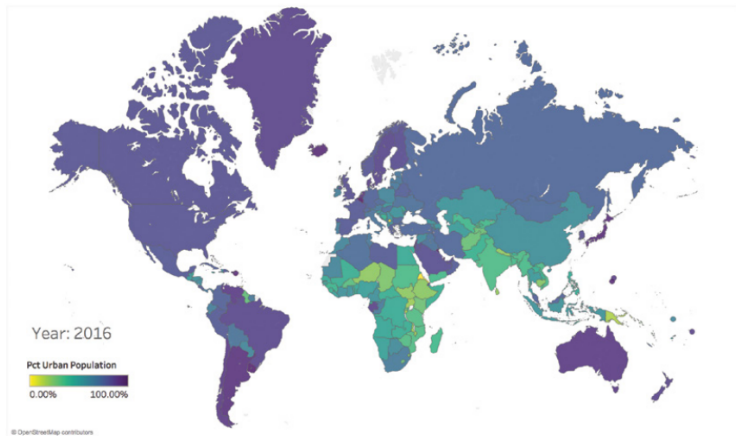
What's "1 unknown"? California!

Wait, so we were. . .

counting twice for each gender, and then twice again for each county!

Mathematical problems

The World Bank data set with estimates of the percent of each country's population that lives in an urban environment. From 33.6% in 1960 to 54.3% in 2016.



Notice Eritrea (Asmara, 650K), St. ! Martin and Kosovo

Mathematical problems

Percent Urban Population, 2016

Region	Country Name	Pct Urban Population
North America	Bermuda	100.00%
	Canada	82.01%
	United States	81.79%

Let's think about North America

Question

How to calculate the percent for the entire region from these three country-level figures?

Mathematical problems

Percent Urban Population, 2016

Region	Country Name	Pct Urban Population
North America	Bermuda	100.00%
	Canada	82.01%
	United States	81.79%
Average		87.93%

Let's average!

Mathematical problems

Percent Urban Population, 2016

Region	Country Name	Pct Urban Population
North America	Bermuda	100.00%
	Canada	82.01%
	United States	81.79%
Average		87.93%

Let's average! Or wait...

Mathematical problems

Percent Urban Population, 2016

Region	Country Name	Pct Urban Population
North America	Bermuda	100.00%
	Canada	82.01%
	United States	81.79%
Average		87.93%

Let's average! Or wait...

$$\text{mean} \left(\frac{\text{urban US}}{\text{total US}}, \frac{\text{urban Canada}}{\text{total Canada}}, \frac{\text{urban Bermuda}}{\text{total Bermuda}} \right) \neq \frac{\text{urban US} + \text{urban Canada} + \text{urban Bermuda}}{\text{total US} + \text{total Canada} + \text{total Bermuda}}$$

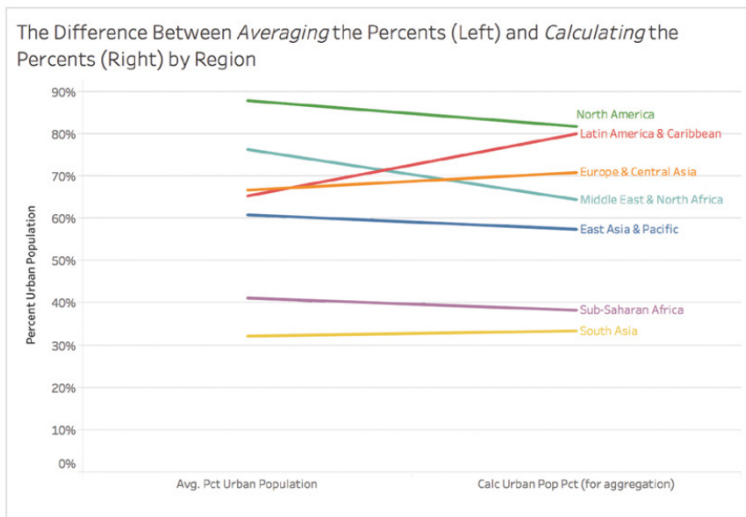
Mathematical problems

Percent Urban Population, 2016

Region	Country Name	Calc Urban Pop Pct (for aggregation)	Total population	Calculated Urban Pop
North America	Bermuda	100.00%	65,376	65,376
	Canada	82.01%	36,264,604	29,739,151
	United States	81.79%	323,127,513	264,279,530
Grand Total		81.81%	359,457,493	294,084,057

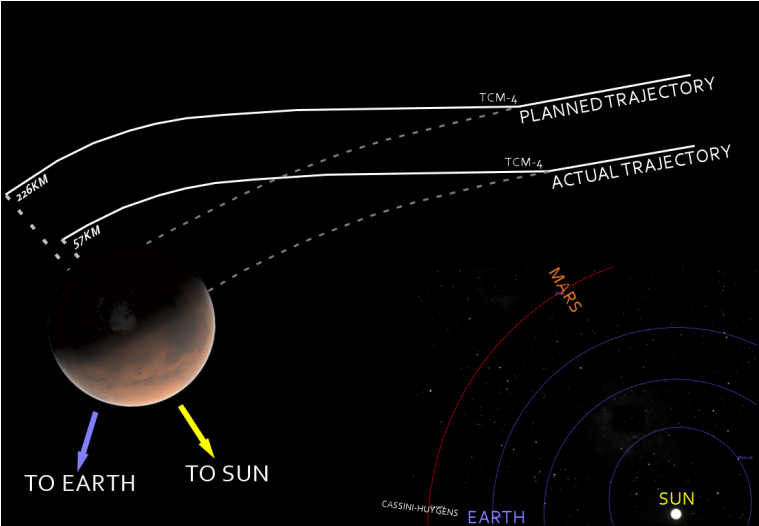
You need the totals before you calculate!

Mathematical problems



A general picture

Mathematical problems



One pound-force second (Lockheed) = 4.45 Newton (NASA); \$327.4 million

Mathematical problems

- cost or revenue with different currencies
- inventory with different units of measure: units, boxes, palettes etc.
- temperatures: Celsius, Fahrenheit, Kelvin
- doing math with any quantity with suffixes such as K or M
- latitude and longitude in degrees minutes seconds (DMS) versus decimal degrees (dd)
- working with 2-D spatial location using cartesian versus polar coordinates
- working with angles in degrees versus radians
- shipping dates when working with calendar days versus business days

Mathematical problems

- cost or revenue with different currencies
- inventory with different units of measure: units, boxes, palettes etc.
- temperatures: Celsius, Fahrenheit, Kelvin
- doing math with any quantity with suffixes such as K or M
- latitude and longitude in degrees minutes seconds (DMS) versus decimal degrees (dd)
- working with 2-D spatial location using cartesian versus polar coordinates
- working with angles in degrees versus radians
- shipping dates when working with calendar days versus business days

Solution

Prepare or read carefully the metadata.