Data Interpretation, Visualization & Story-telling (Lecture 02)

Dr. Oul Han

Special thanks to:
Claudia Wagner (http://claudiawagner.info/)

What is data visualization?

- Visual display of measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

Source: The Visual Display of Quantitative Information, Edward R. Tufte

Why do we visualize data?

- Being able to visualize data & tell stories with it is key to turning it into information that can be used to drive better decision making

Visualizations should:

- show the data
- induce the viewer to think about the substance
- avoid distorting what the data has to say
- present many numbers in small space
- make large data sets coherent
- encourage eye to compare different pieces of data
- serve a clear purpose: description, exploration, tabulation or decoration
- be closely integrated with statistical & verbal description of data set

Source: Edward R. Tufte The Visual Display of Quantitative Information

Visualizations should:

- show the data
- induce the viewer to think about the substance
- avoid distorting what the data has to say
- present ma
- make large
- encourage eye to compare different pieces of data
- serve a clear purpose: description, exploration, tabulation or decoration
- be closely integrated with statistical & verbal description of data set

Visualizations reveal data!

Source: Edward R. Tufte The Visual Display of Quantitative Information

In this class

Interpretation

Visualization

Story-telling

Data Interpretation

Data can be biased too!

- Random sample of Twitter users
  - Random sample of tweets from public timelines
  - Problem: More active users are more likely to be included
- Friendship paradox
  - Select a random sample of people and ask them to list the people they know. Contact a sample of the listed friends and repeat the survey.
  - Problem: People with more friends are more likely to show up in the friend lists which we generate at the first stage

False conclusion - Correlation vs. Causation

- A study found that the profession with the lowest average age of death was student.
- What can be concluded from that?
  - Being a student makes you die young?
  - Being a student means you are young?
- Amount of ice cream consumed per day is highly correlated with number of drownings per day.
- Why?
  - Both variables are correlated with the daily temperature

Source: “Teaching Statistics: A Bag of Tricks," by Gelman and Nolan (2002)

False conclusions - Conditional Probability

- A study found that only 1.5% of drivers in accidents reported that they were using a cell phone, whereas 10.9% reported that they were distracted by another occupant in the car
- Can we conclude that using a cell phone safer than speaking with another occupant?
- Can we conclude that using a cell phone safer than speaking with another occupant
  - P(cellphone | accident) != P(accident | cellphone)
  - Compare P(accident|cellphone) and P(accident|occupant)
  - We need to know the prevalence of cell phone use
  - It is likely that much more people talk to another occupant in the car while driving than talking on the cell phone

Jessica Utts, What Educated Citizens Should Know about Statistics and Probability, The American Statistician, Vol. 57, No. 2 (May, 2003), pp. 74-79

False conclusions - Ecological fallacy

- Ecological fallacy
  - fallacy in the interpretation of statistical data that occurs when inferences about the nature of individuals are deduced from inferences about the group to which those individuals belong.
- Illiteracy rates in each US state and the proportion of immigrants per state
- Negative Correlation of -0.53
  - The greater the proportion of immigrants in a state, the lower its average illiteracy
- Can we conclude that immigrants are more likely to not be illiterate?
  - No, When individuals are considered, the correlation was +0.12 — immigrants were on average more illiterate than native citizens

Robinson, W.S. (1950). "Ecological Correlations and the Behavior of Individuals".
AmericanSociological Review (American Sociological Review, Vol. 15, No. 3) 15 (3): 351–357.

False conclusions - Simpson's paradox

- Simpson's paradox
  - A trend appears in different groups of data but disappears or reverses when these groups are combined
- Gender bias among Graduate School admissions to UC, Berkley
- Men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

False conclusions - Simpson's paradox

- Examining the individual departments, the pooled and corrected data showed a "small but statistically significant bias in favor of women."

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	373	6%	341	**7%**

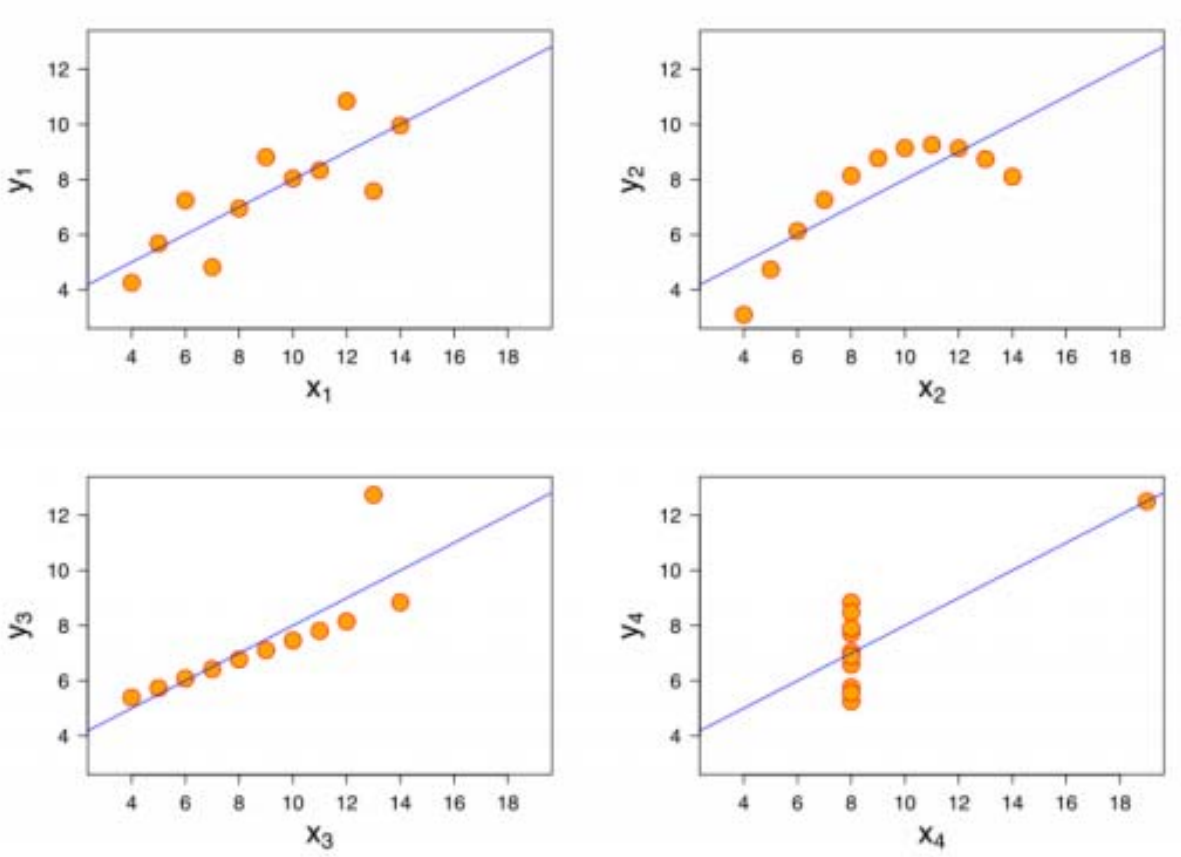
Anscombe's quartet

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

Anscombe's quartet



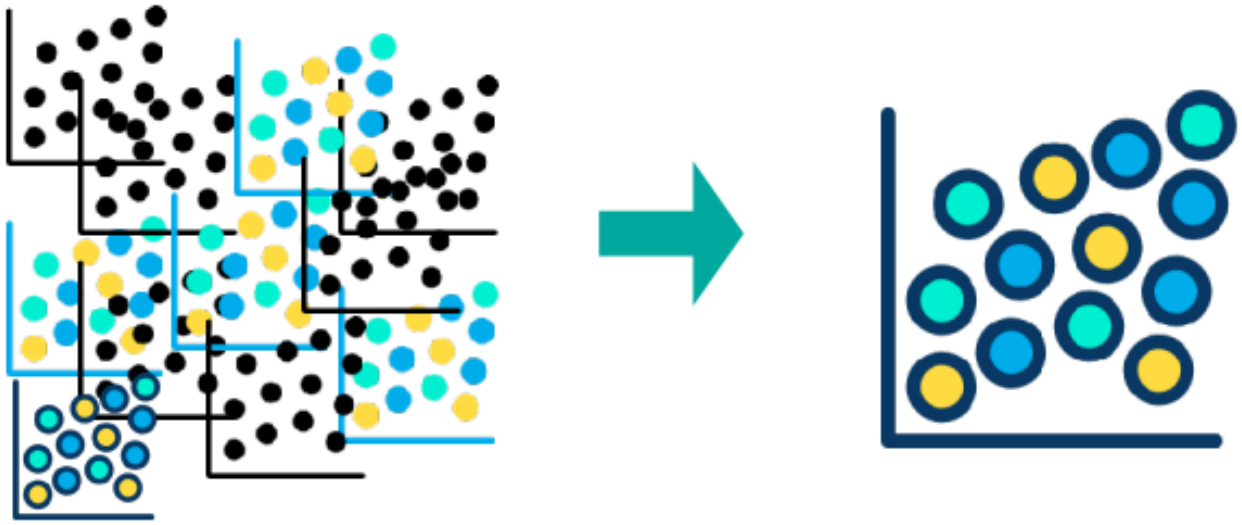
Data Visualization

Data visualization

- Intersection of science and art
  - Science - best practices and guidelines
  - Art - Aesthetics, design and story

Sources: “Storytelling with data” Cole Nussbaumer Knaflic,
“The Visual Display of Quantitative Information” Edward R. Tufte
Image source: https://www.flaticon.com/

Exploratory vs. explanatory analysis



Exploratory analysis

Explanatory analysis

Effective visual communication with data

- Understand the context
- Choose an appropriate visual display
- Eliminate clutter
- Focus attention where you want it
- Think like a designer
- Tell a story

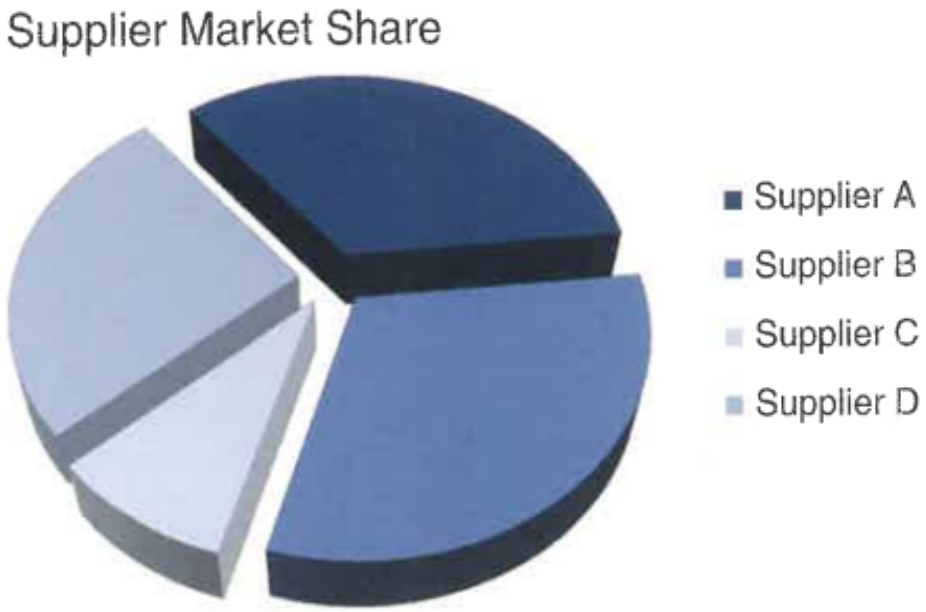
Importance of context

- Before starting, ask yourself
  - Who is your audience?
  - What do you need them to know or do?
- Situational context
  - Audience
  - Communication mechanism
  - Desired tone

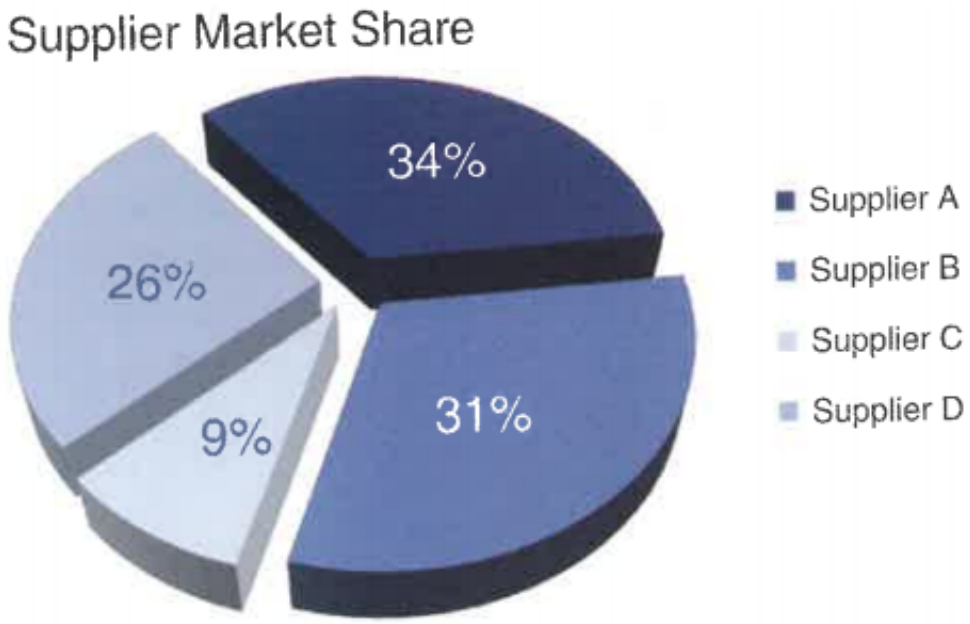
Choosing an effective visual

- What is the best way to show the data you want to communicate?
- Common types of visuals
  - Simple text
  - Table
  - Heatmap
  - Line graph
  - Bar chart
- Avoid
  - Pie/donut charts
  - 3D visualizations

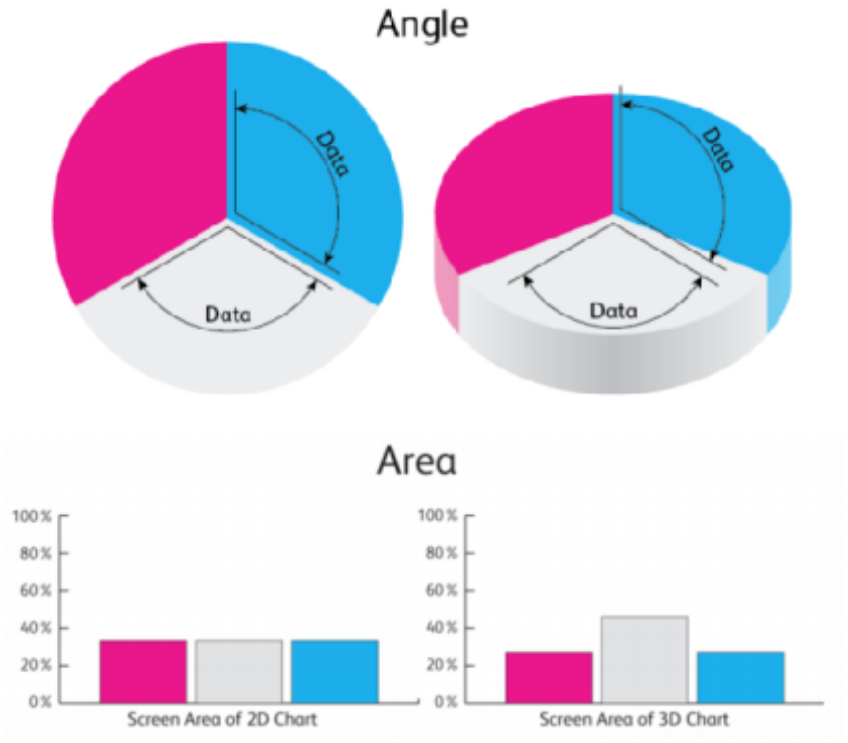
Pie chart + 3D


Supplier Market Share
Supplier A
Supplier B
Supplier C
Supplier D

Pie chart + 3D = Disaster

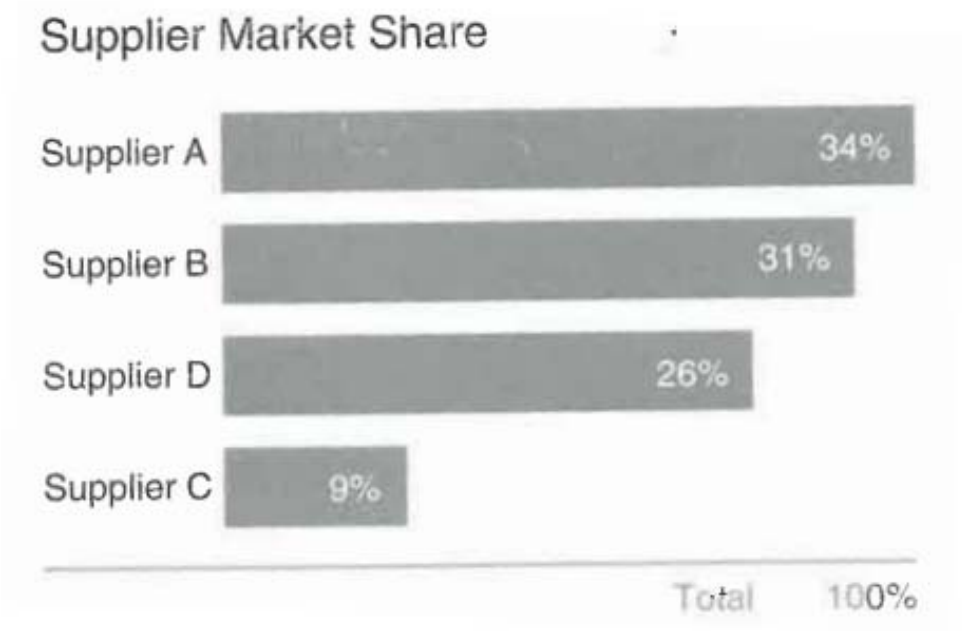


Don’t create 3D out of 2D data!



https://en.rockcontent.com/blog/2ds-company-3ds-a-crowd/

Better option - Bar charts



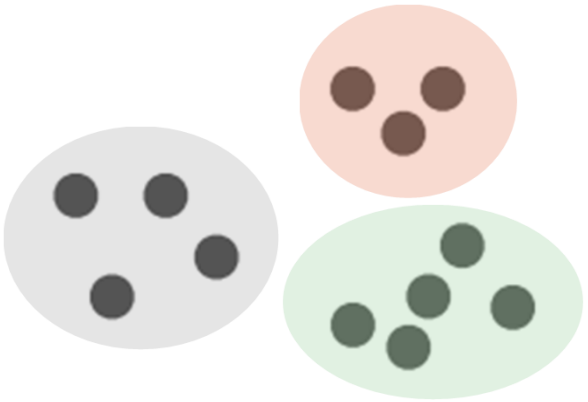
- The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
- Shrinking dollar fallacy

Avoiding clutter

- Everything in your visualization takes up cognitive load of your audience
- Gestalt principles of visual perception

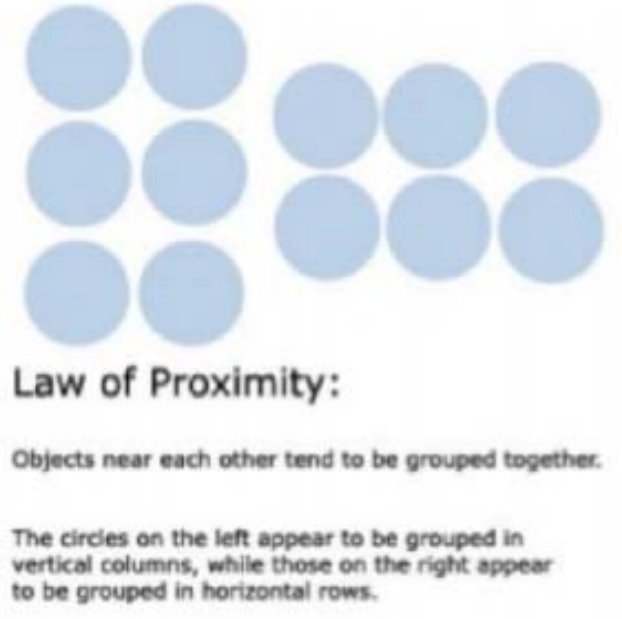
Gestalt principles of visual perception

- Proximity



Gestalt principles of visual perception

- Proximity



Gestalt principles of visual perception

- Similarity

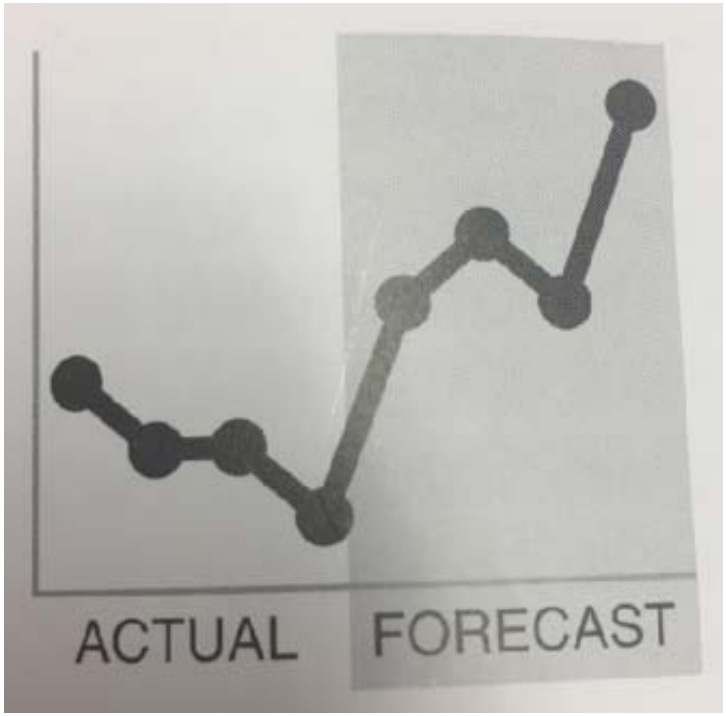
Law of Similarity:

Items that are similar tend to be grouped together.

In the image above, most people see vertical columns of circles and squares.

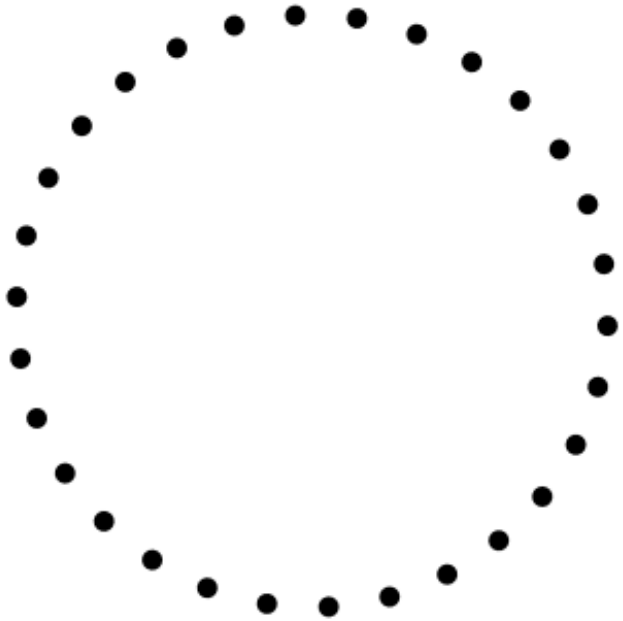
Gestalt principles of visual perception

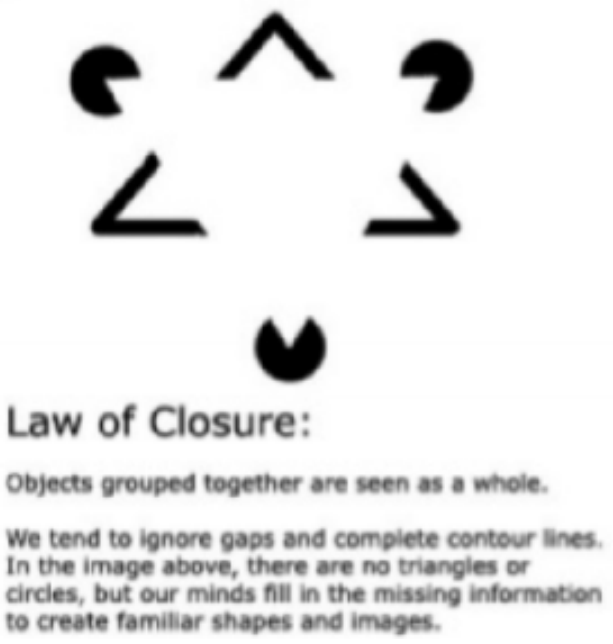
- Enclosure



Gestalt principles of visual perception

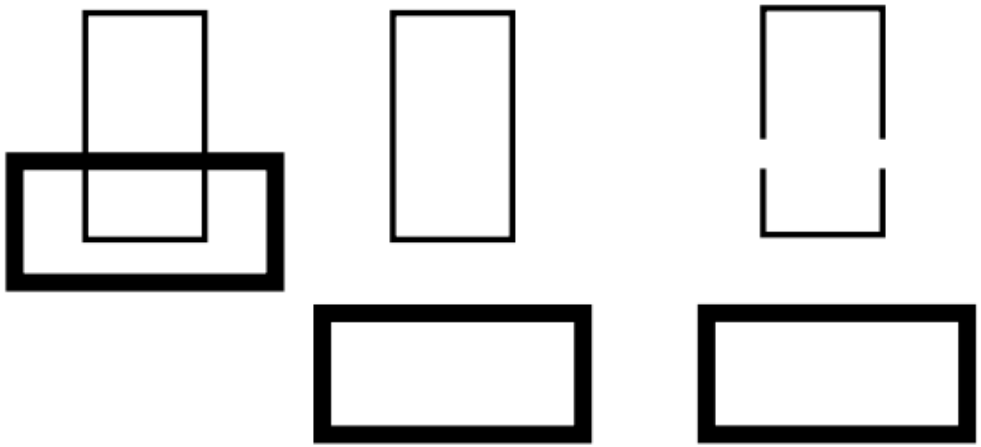
- Closure





Gestalt principles of visual perception

- Continuity



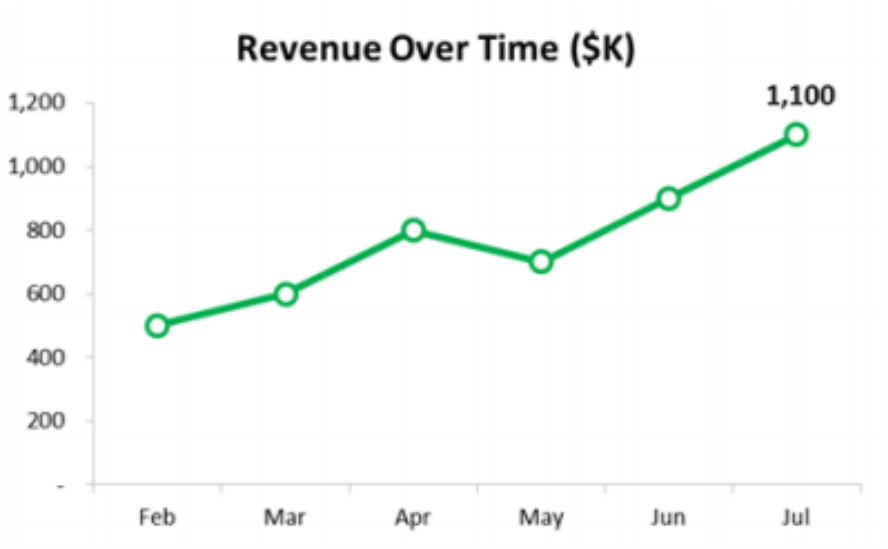
Law of Pragnanz:

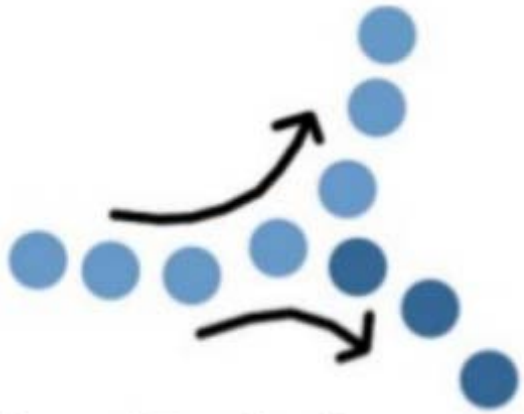
Reality is organized or reduced to the simplest form possible.

For example, we see the image above as a series of circles rather than as many much more complicated shapes.

Gestalt principles of visual perception

- Connection





Law of Continuity:

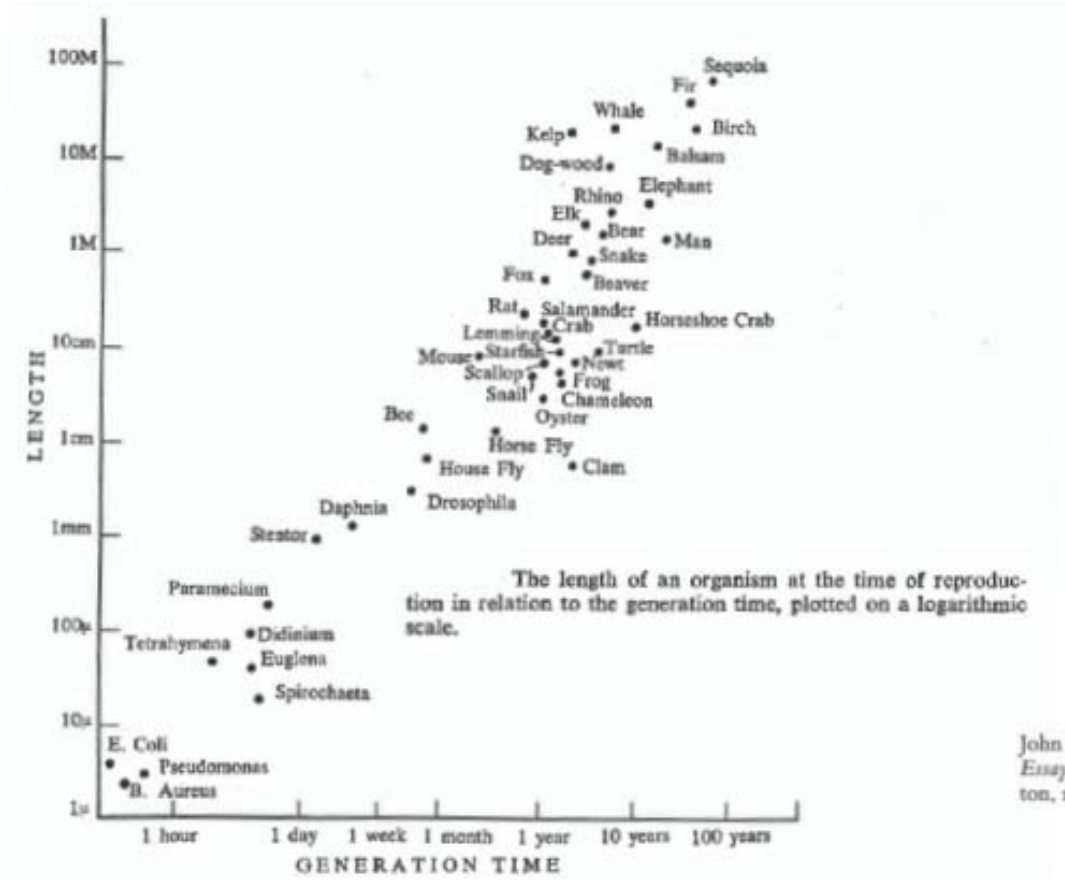
Lines are seen as following the smoothest path.

In the image above, the top branch is seen as continuing the first segment of the line. This allows us to see things as flowing smoothly without breaking lines up into multiple parts.

Avoiding clutter

- Everything in your visualization takes up cognitive load of your audience
- Gestalt principles of visual perception
- Maximize data-ink ratio (within reason)
- $\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$

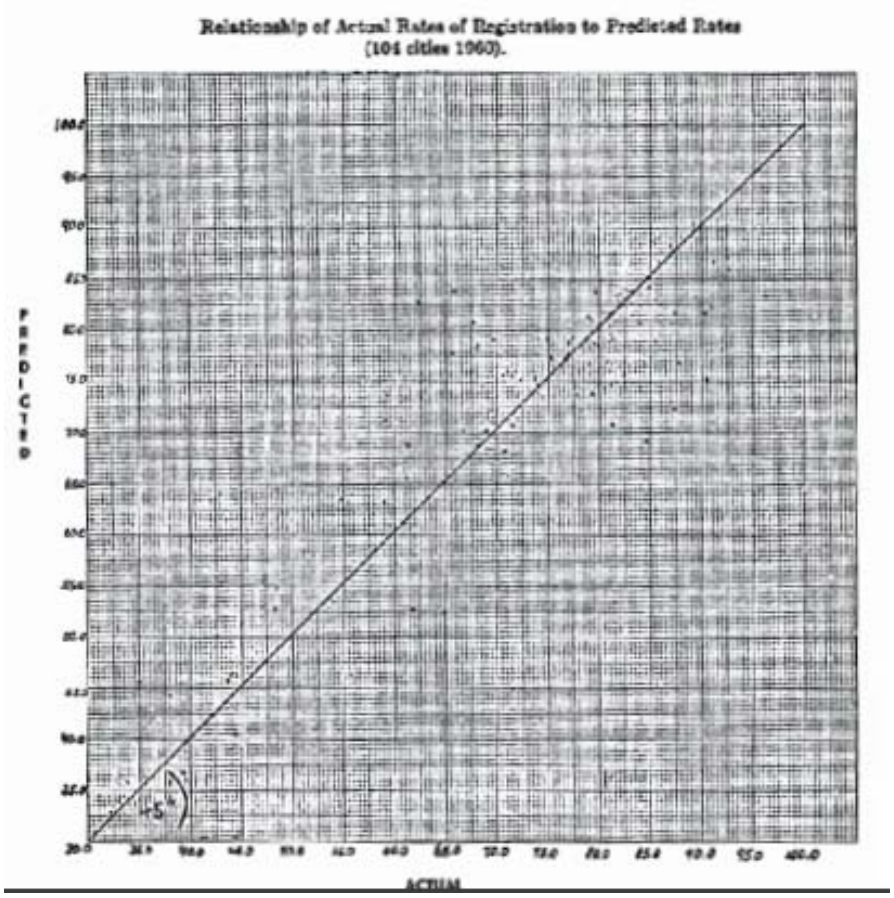
Maximize data-ink ratio



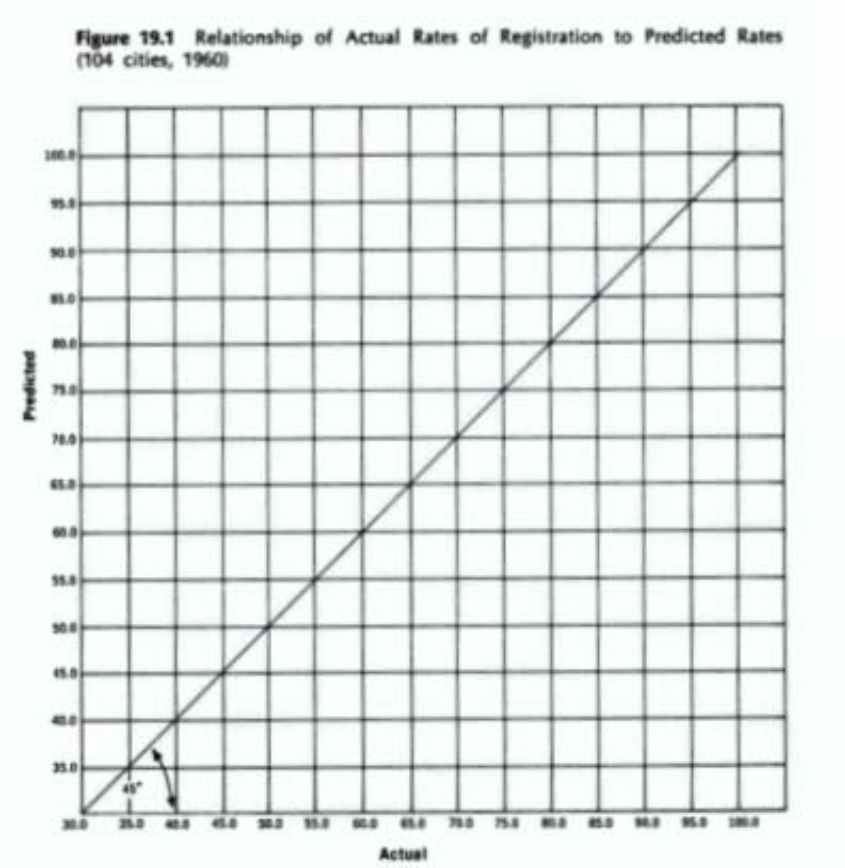
The length of an organism at the time of reproduction in relation to the generation time, plotted on a logarithmic scale.

John Tyler Bonner, *Size and Cycle: An Essay on the Structure of Biology* (Princeton, 1965), 17.

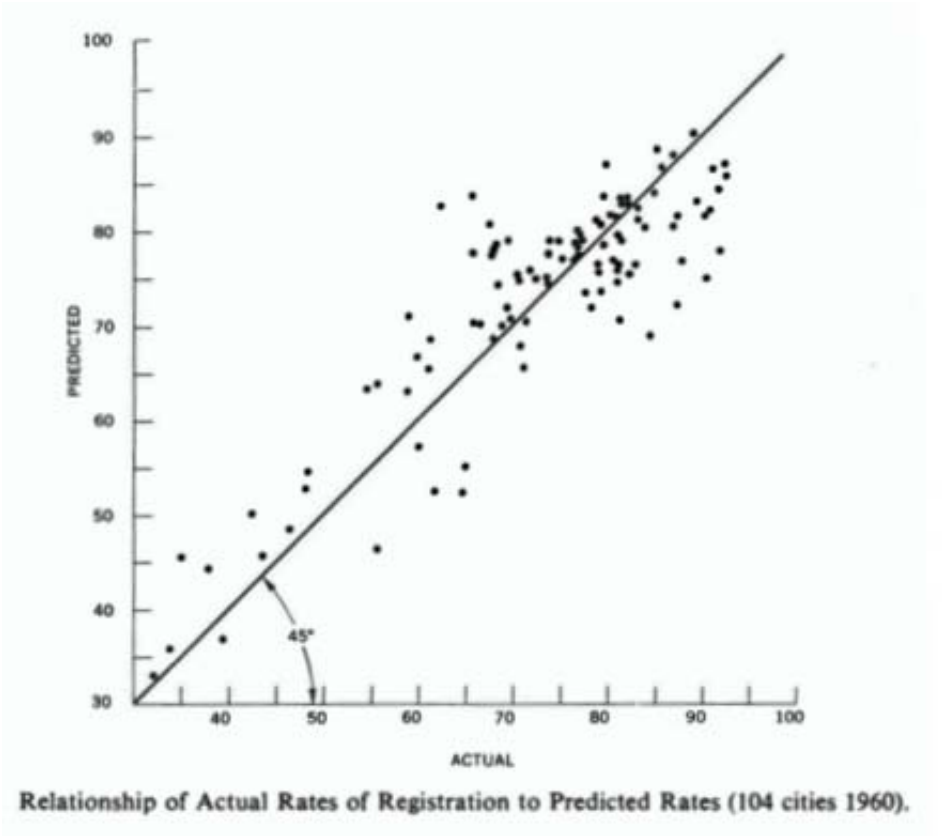
Maximize data-ink ratio



Maximize data-ink ratio



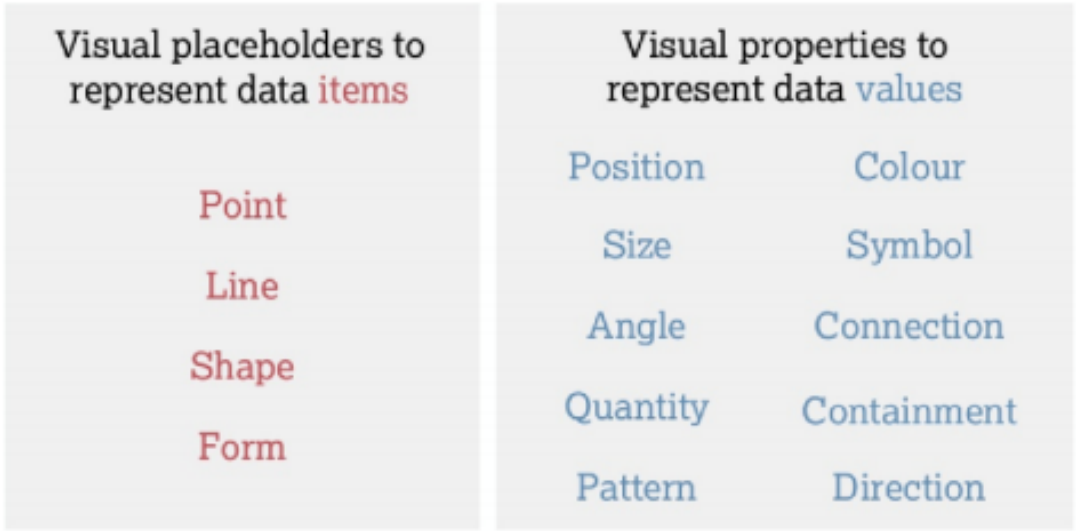
Maximize data-ink ratio



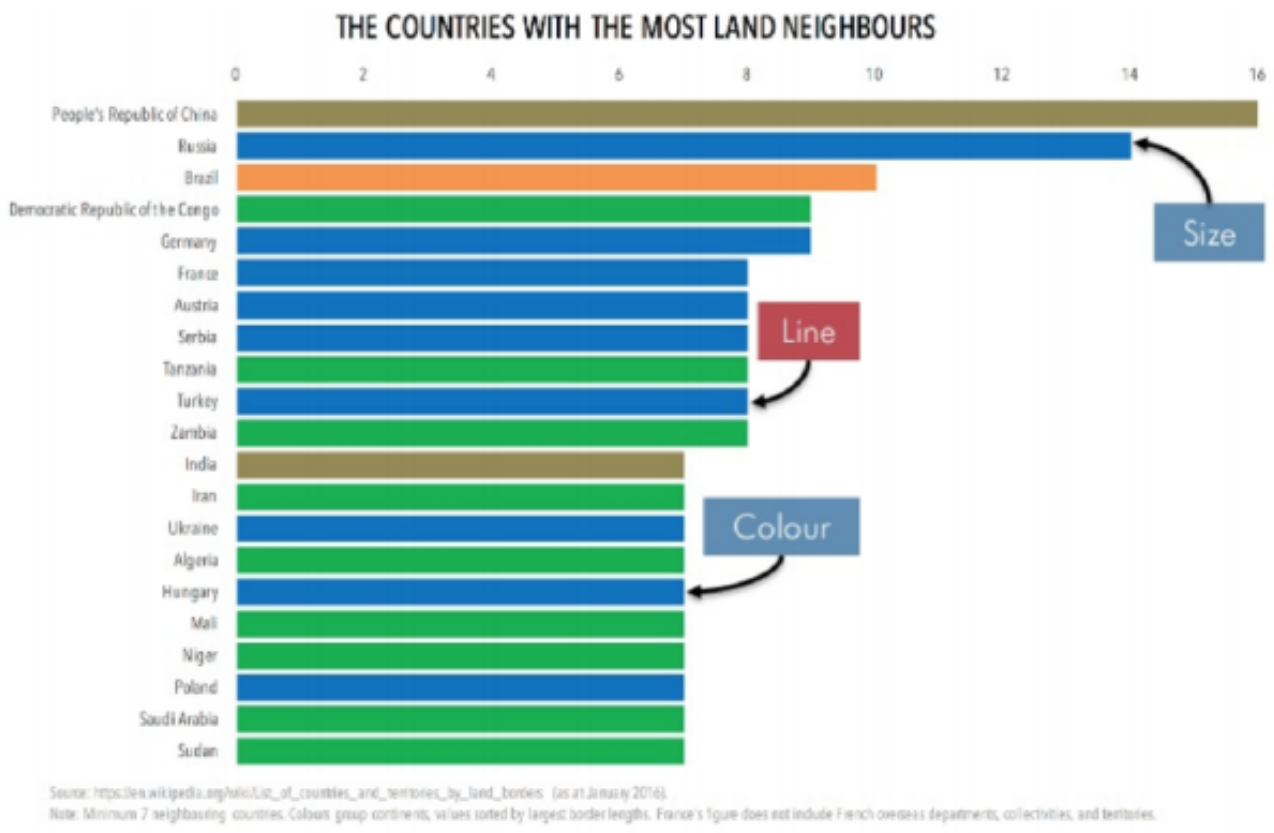
Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

Focus audience's attention

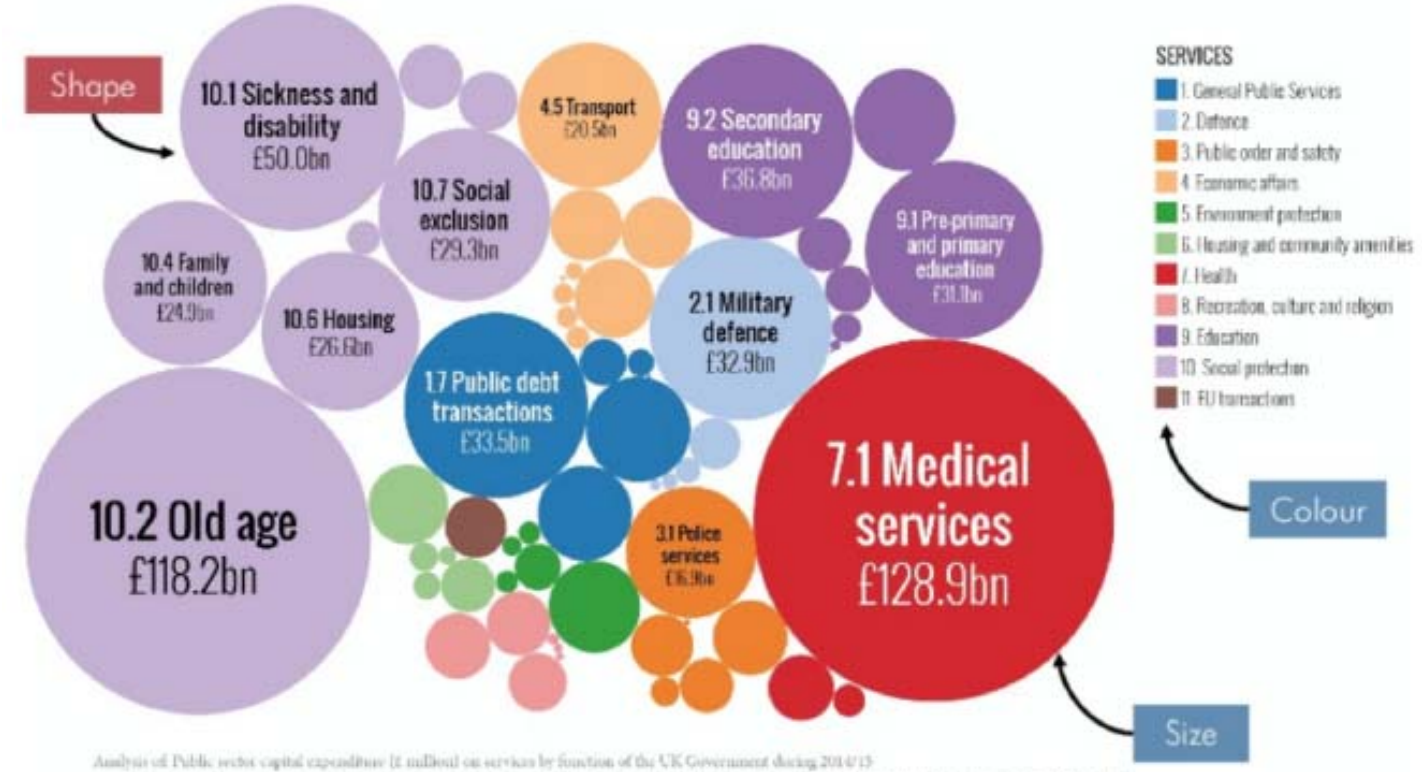
- Human sight & memory
- Pre-attentive attributes
  - Size
  - Color
  - Position on page



Focus audience's attention



Focus audience's attention

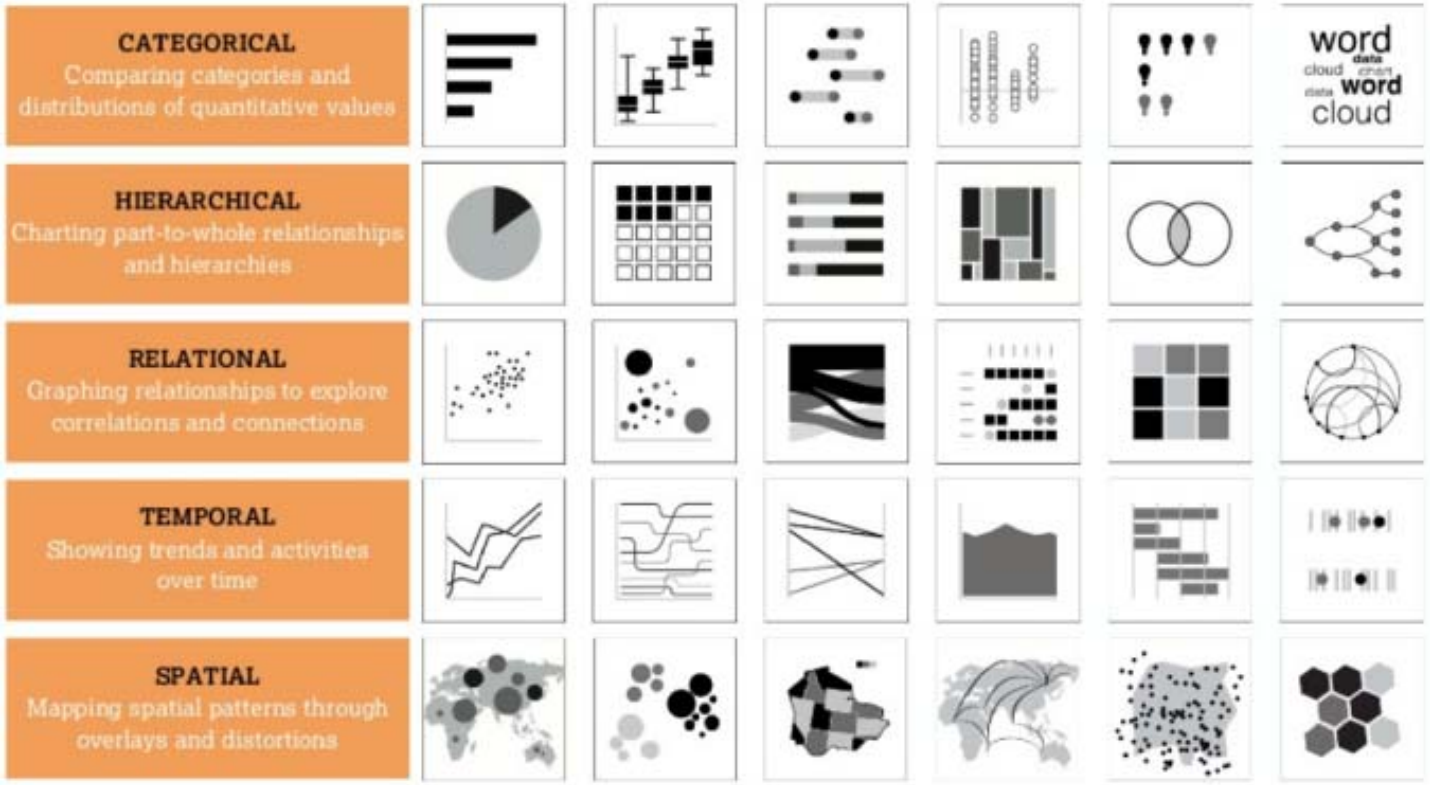


Analysis of Public sector capital expenditure (£ million) on services by function of the UK Government during 2014/15
Data from: HM Treasury, 'Public Expenditure Statistical Analyses 2015' https://www.gov.uk/government/statistics/public-expenditure-statistical-analyses-2015

Think like a designer

- Form follows function
- First decide what we want the audience to do with the data (function) and then create a visualization (form) to help them do it with ease.

How to show what you want to say?



Source: Andy Kirk, Visualizing Data Ltd.

Storytelling with data

Why stories?

"Maybe stories are just data with a soul."

- Brené Brown

Telling stories

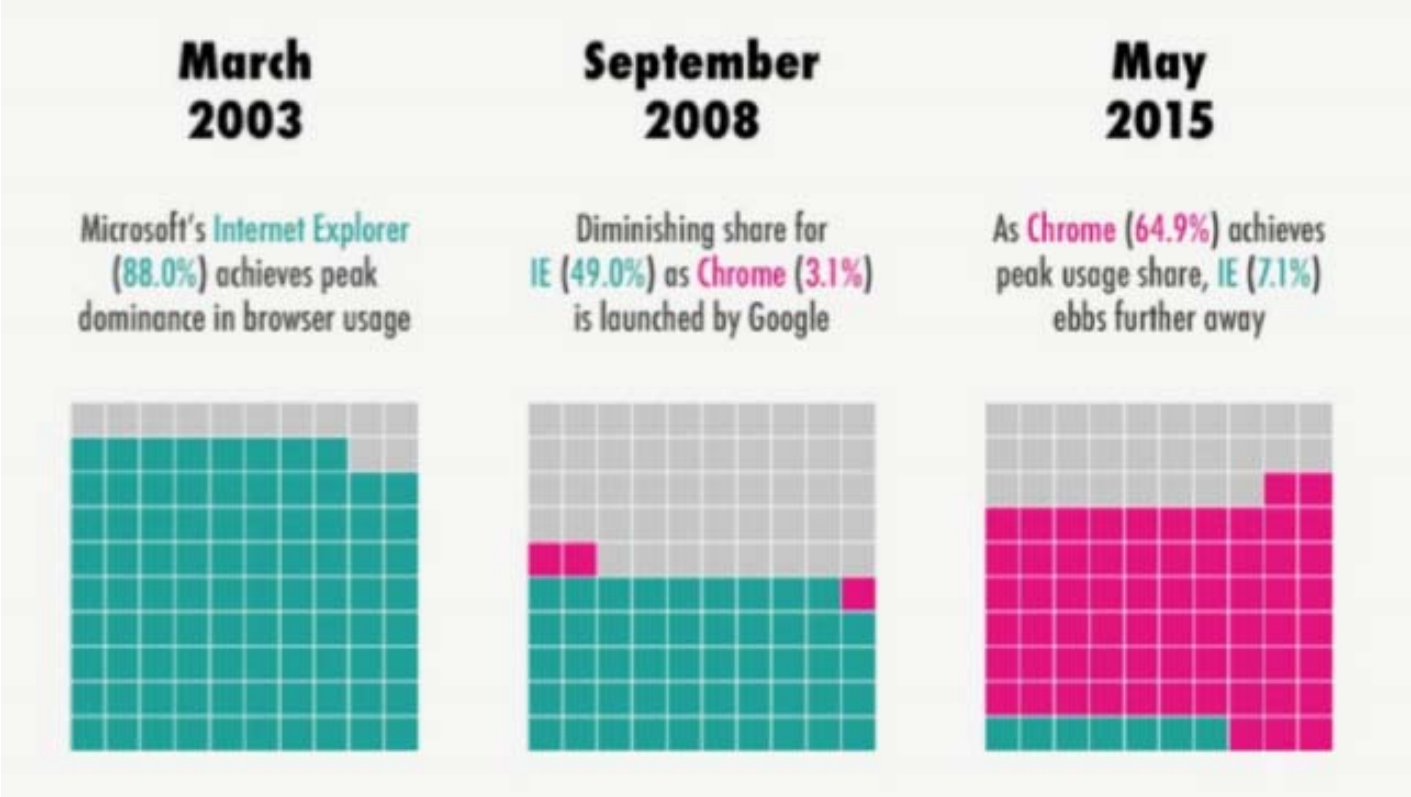
- Stories resonate and stick with us in ways that data alone can not
- Stories have a beginning, middle and end
- Power of repetition, narrative flow etc.
- Good stories drive change

http://www.forbes.com/sites/brentdykes/2016/03/31/data-storytelling-the-essential-data-science-skill-everyone-needs

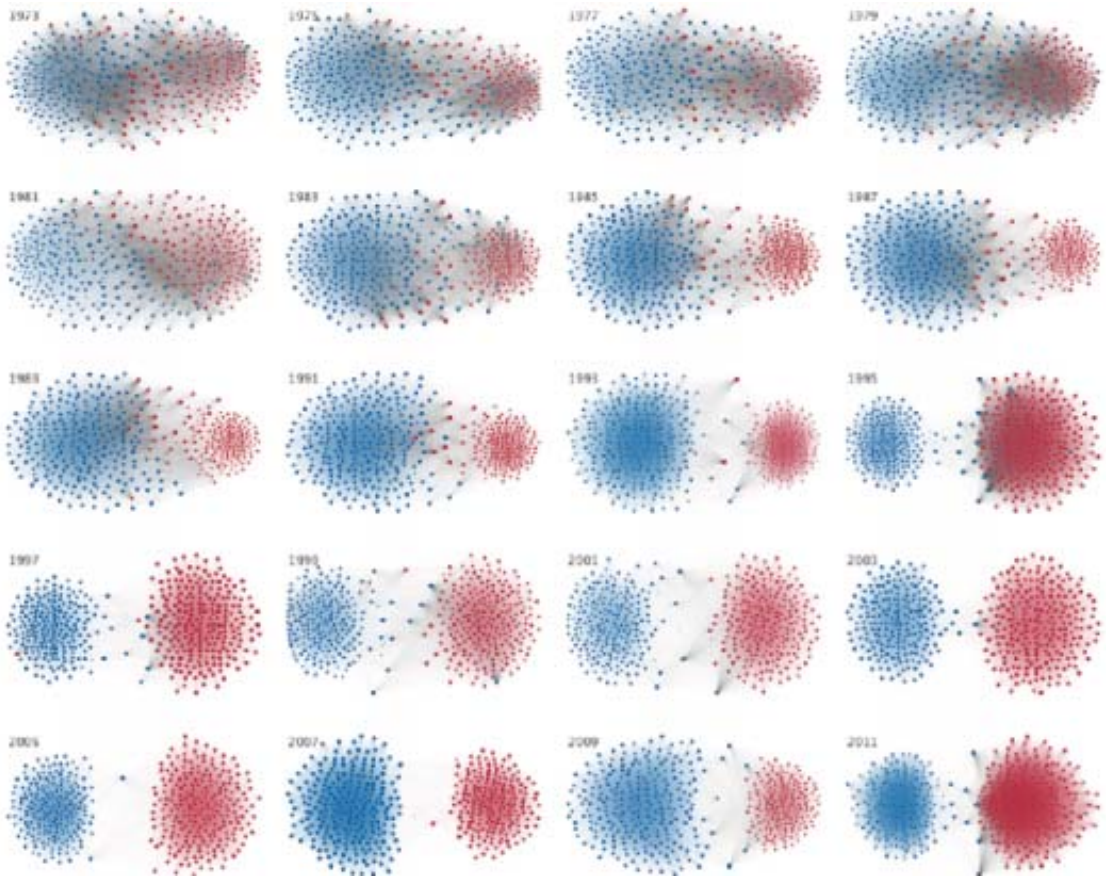
Telling stories

- First decide what's your goal, your main point
- Organize facts into compelling narrative
  - Narrative is "an account of a series of events, facts, etc, given in order while establishing connections between them
- Start with the problem/question
- Attempt to resolve/answer the problem/question
- End with resolution
- Include visualization to support narrative
- Engage the audience & make them think (give them 2+2 not 4)

Stories can emerge from temporal presentations

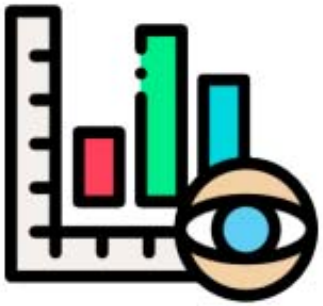


Stories can emerge from temporal presentations



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0123507

In this class… we learnt



Visualization

Interpretation

Story-telling

Image Source: https://www.flaticon.com/

End of Lecture 2

Questions:

Tutorials 2. TA Office hours (book slot)