COMP9321:
Data services engineering

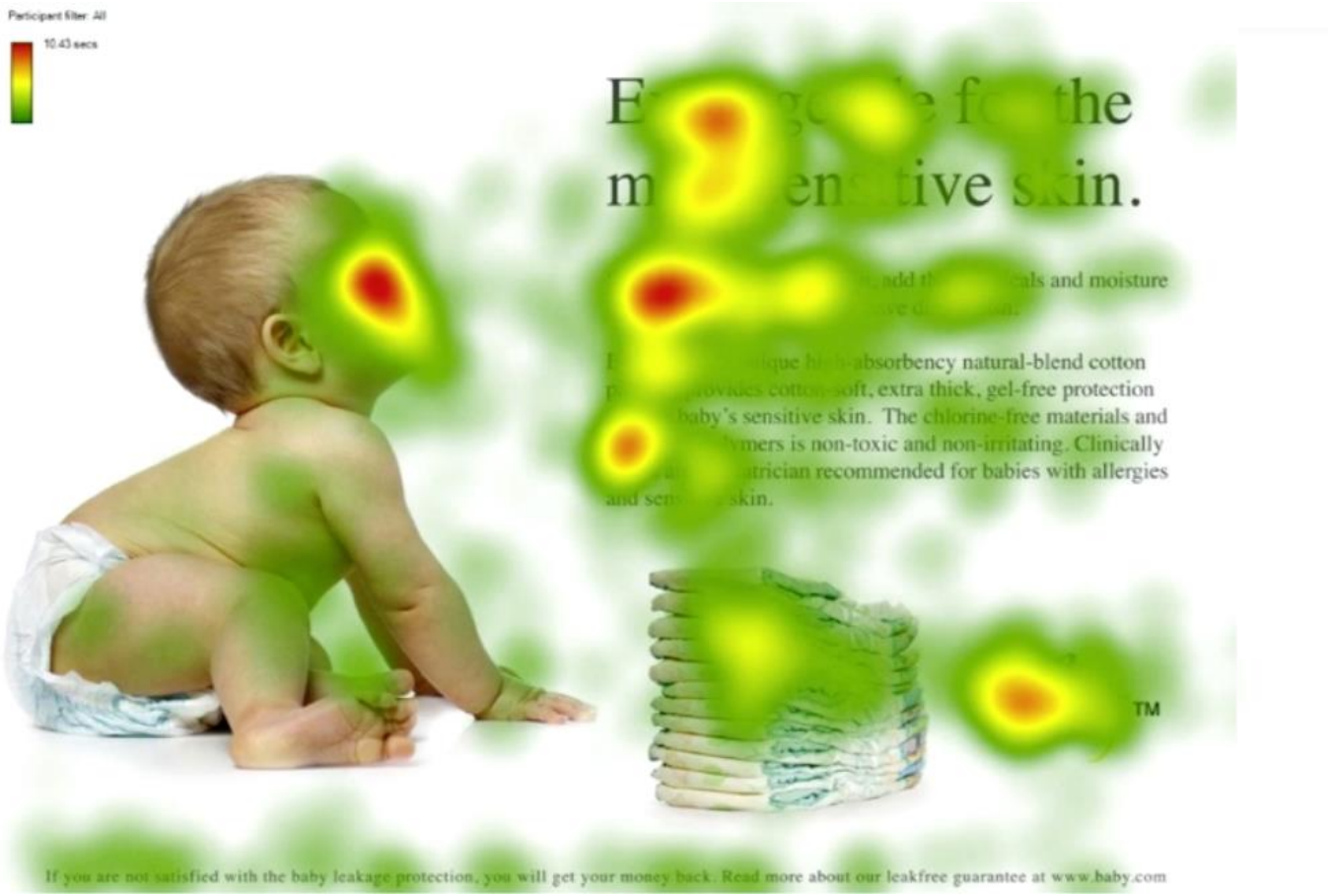
Week 5: Data Visualisation (Principles and Basic Techniques)

Semester 2, 2018,

By Mortada Al-Bana, CSE UNSW

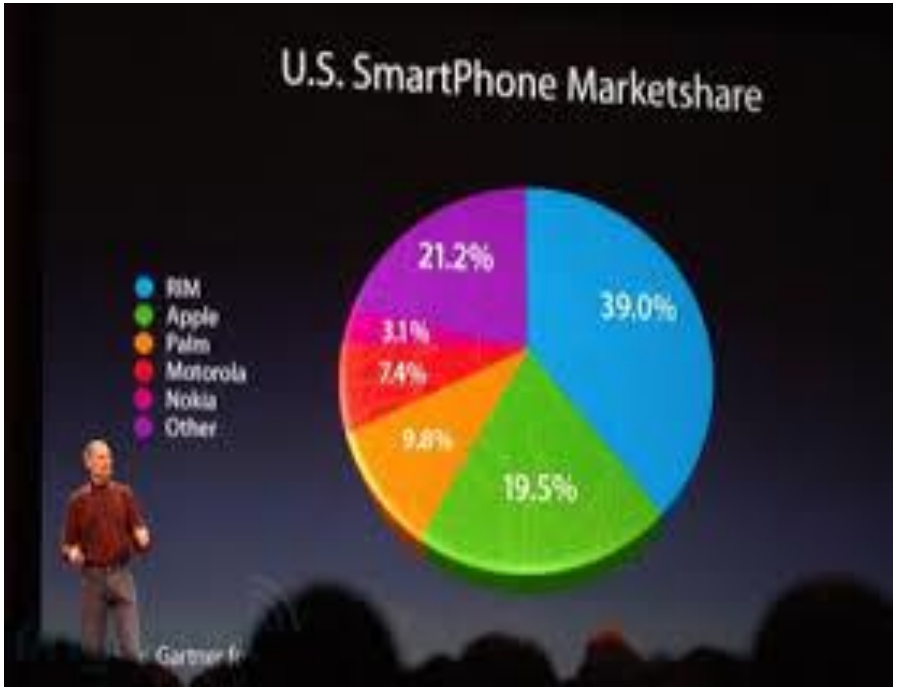
Visualisation isn't just about graphics

Highly competent visualisation "tricks" can affect what you see and what you pay attention to:



When Steve Jobs says …

- Again … playing with human visual perception
- 21.2% versus 19.5% slices of the pie



Apple SmartPhone Market Update, Macworld keynote (2008)

https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S

In this lecture …

Often, visualization is considered highly specialized area and getting it right takes skilled knowledge from multi disciplinary areas (statistics, graphic design, understanding of human perception of visual elements, or sometimes understanding of human psychology)

In most cases, the topic will be a course by itself…

Of course, we cannot get to that level of details here … But if you are doing some data analysis or writing an application that is data-driven, most likely you'd run into some visualization tasks.

What I intend to convey in this lecture is two parts:

First : the basic principles of 'good and competent' visualization

Second: Introduction to Matplotlib library, we actually get to see the basic building blocks of visualization techniques which are useful for further exploration of the area if you are interested …

Presentation as a service

Of course, there is another important point to make.

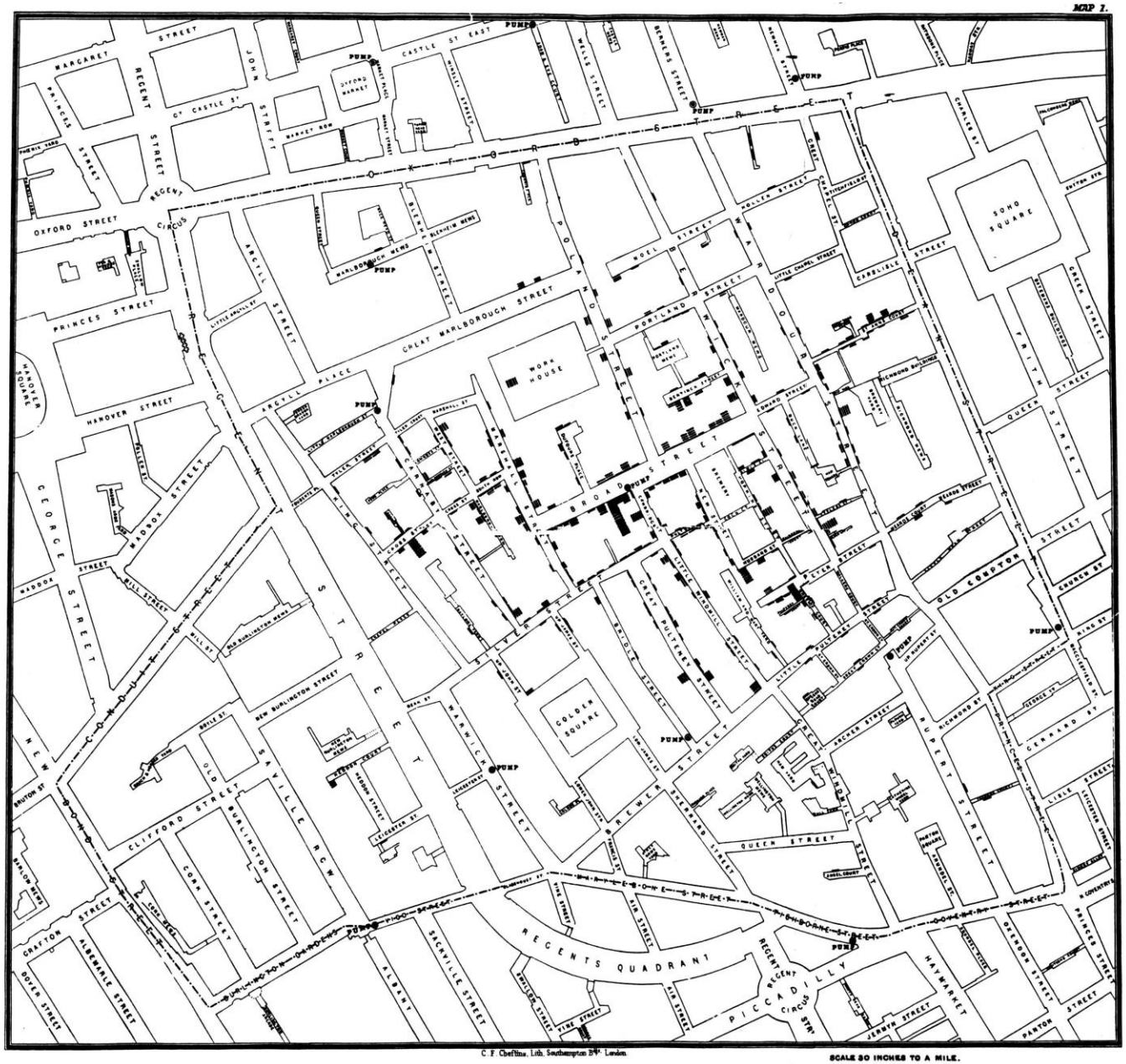
Once you know how to visualise data, what comes naturally after that is to offer that knowledge as a service (API)

It's known as "Presentation" or "Visualisation" as a service.

"Visualisation on the Cloud"

- e.g., https://www.highcharts.com
- e.g., Google Map (good example of presentation as a service)
- On request, its response can contain either "presentation logic + data", or already visualised data in graphics/HTML

So for this course context, the concept of "presentation as a service" is more interesting than the visualisation techniques themselves.



What is Visualisation

Visualization transforms data into images that effectively and accurately represent information about the data.

Insight

https://www1.udel.edu/johnmack/frec682/cholera/ **(John Snow Map of Cholera 1854)**

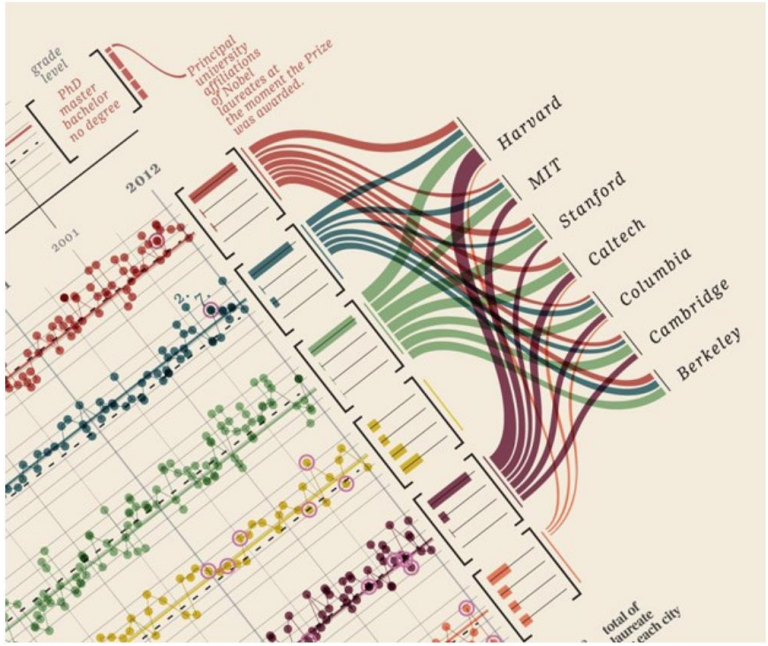
What is visualisation

Three types of goals for visualization – ... to explore

Nothing is known … Visualisation is used for data exploration/discovery …

e.g., A Visual History of Nobel Prizes and Notable Laureates, 1901-2012

https://www.brainpickings.org/2012/11/29/giorgia-lupi-noble-prizes-visualization/



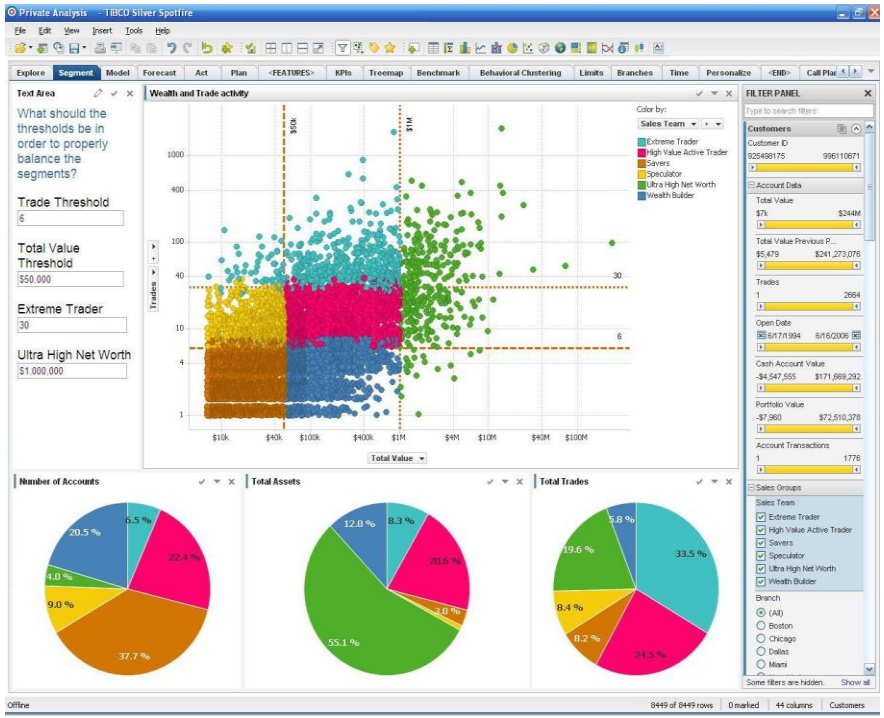
https://www.safaribooksonline.com/library/view/designing-data-visualizations/9781449314774/ch01.html

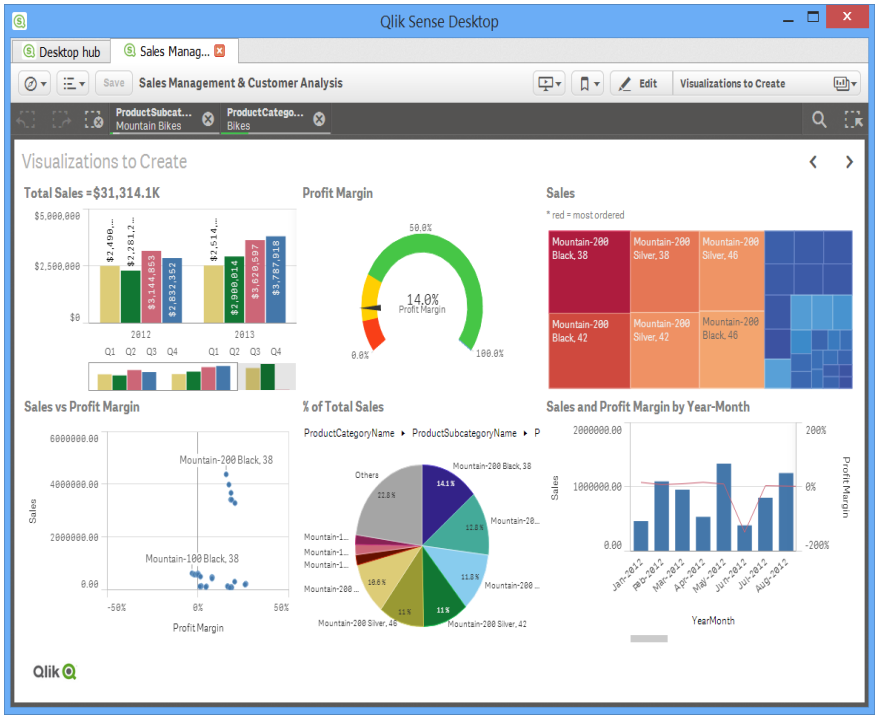
What is visualisation

Three types of goals for visualization – ... to analyse

You have some hypotheses. Visualisation is used for Verification or Falsification

You'd normally use what is classified as "data analysis and visualisation" tools (- a range of data source connectivity, built-in quick visualisation graphs, etc.)





https://www.safaribooksonline.com/library/view/designing-data-visualizations/9781449314774/ch01.html

UNSW SYDNEY

What is visualisation

Three types of goals for visualization – ... to explain

You do know what the data contains … Visualisation is used for “effective” communication of “results” – Making them clear for the audience

https://www.safaribooksonline.com/library/view/designing-data-visualizations/9781449314774/ch01.html

Data Visualisation

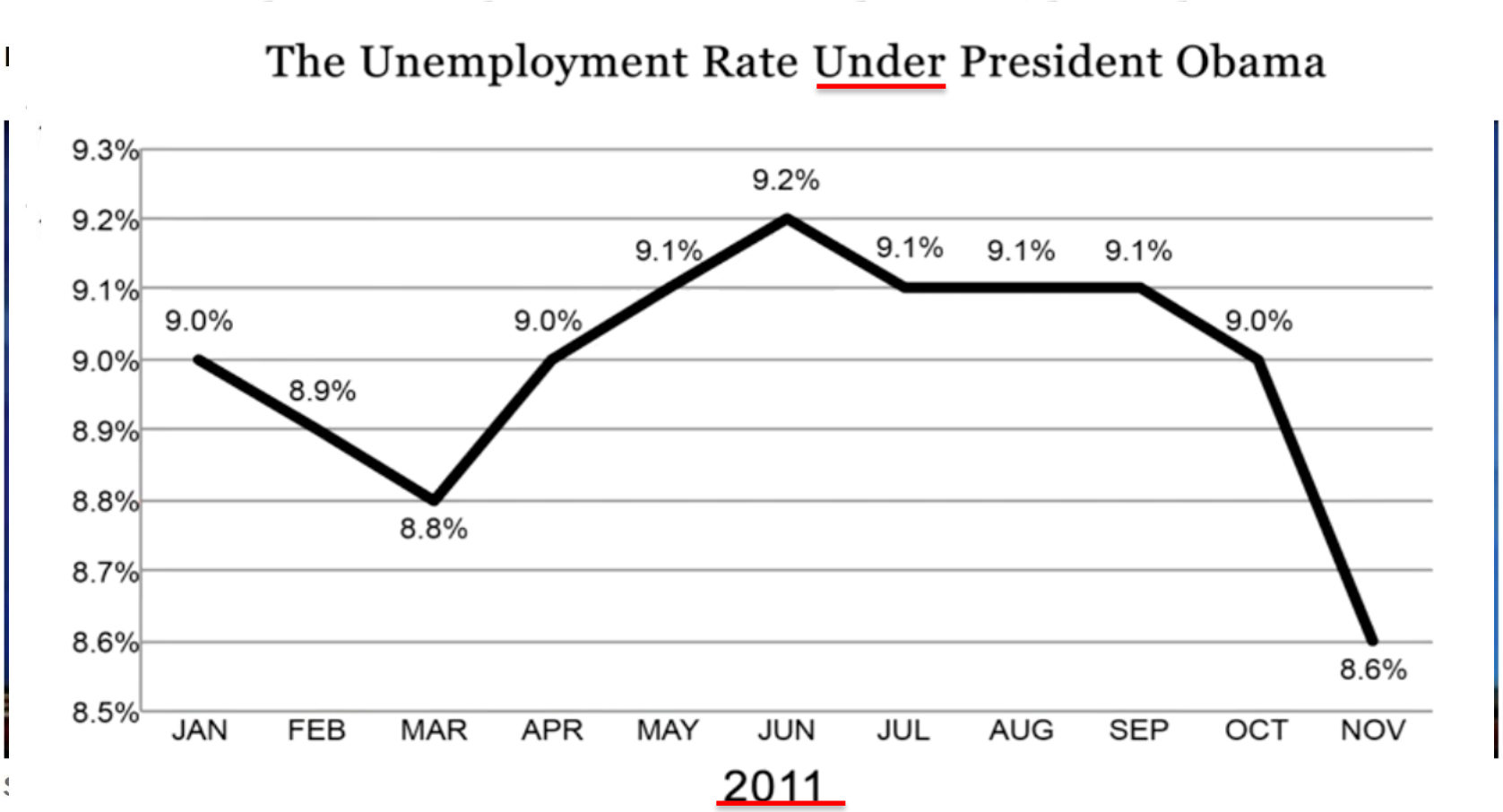
Referring to any visual representation of data that is:

- algorithmically drawn (may have custom touches but is largely rendered with the help of computerized methods);
- easy to regenerate with different data (the same form may be repurposed to represent different datasets with similar dimensions or characteristics);
- often aesthetically simple (data is not decorated); and
- relatively data-rich (large volumes of data are welcome and viable, in contrast to infographics).

https://www.safaribooksonline.com/library/view/designing-data-visualizations/9781449314774/ch01.html

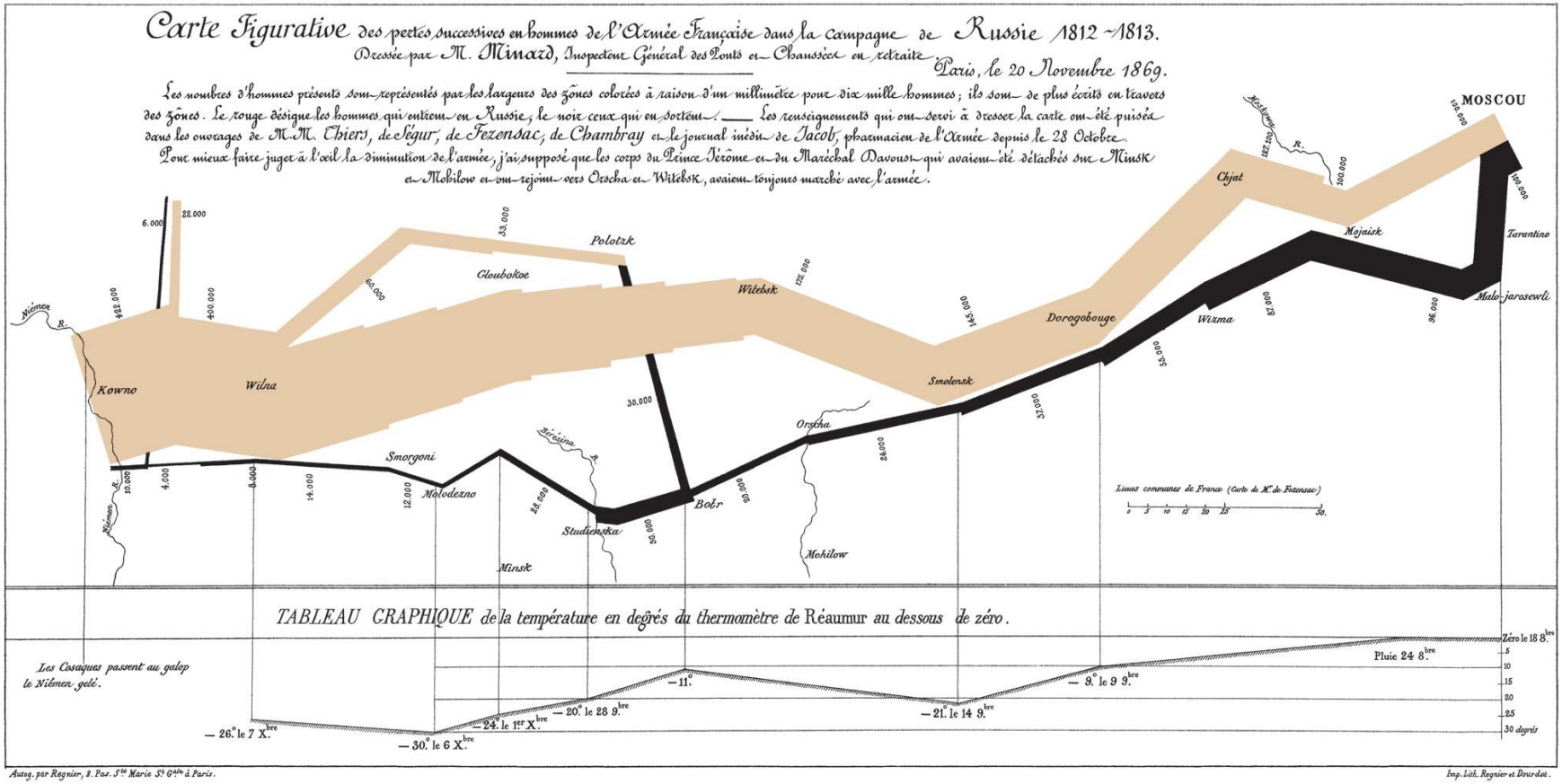
So what makes a good visualisation

Accuracy, Story, Knowledge: Aim to create a visualisation that are accurate, tell a good story, and provide real knowledge to the audience.



So what makes a good visualisation

Accuracy, **Story**, Knowledge: Aim to create a visualisation that are accurate, tell a good story, and provide real knowledge to the audience.



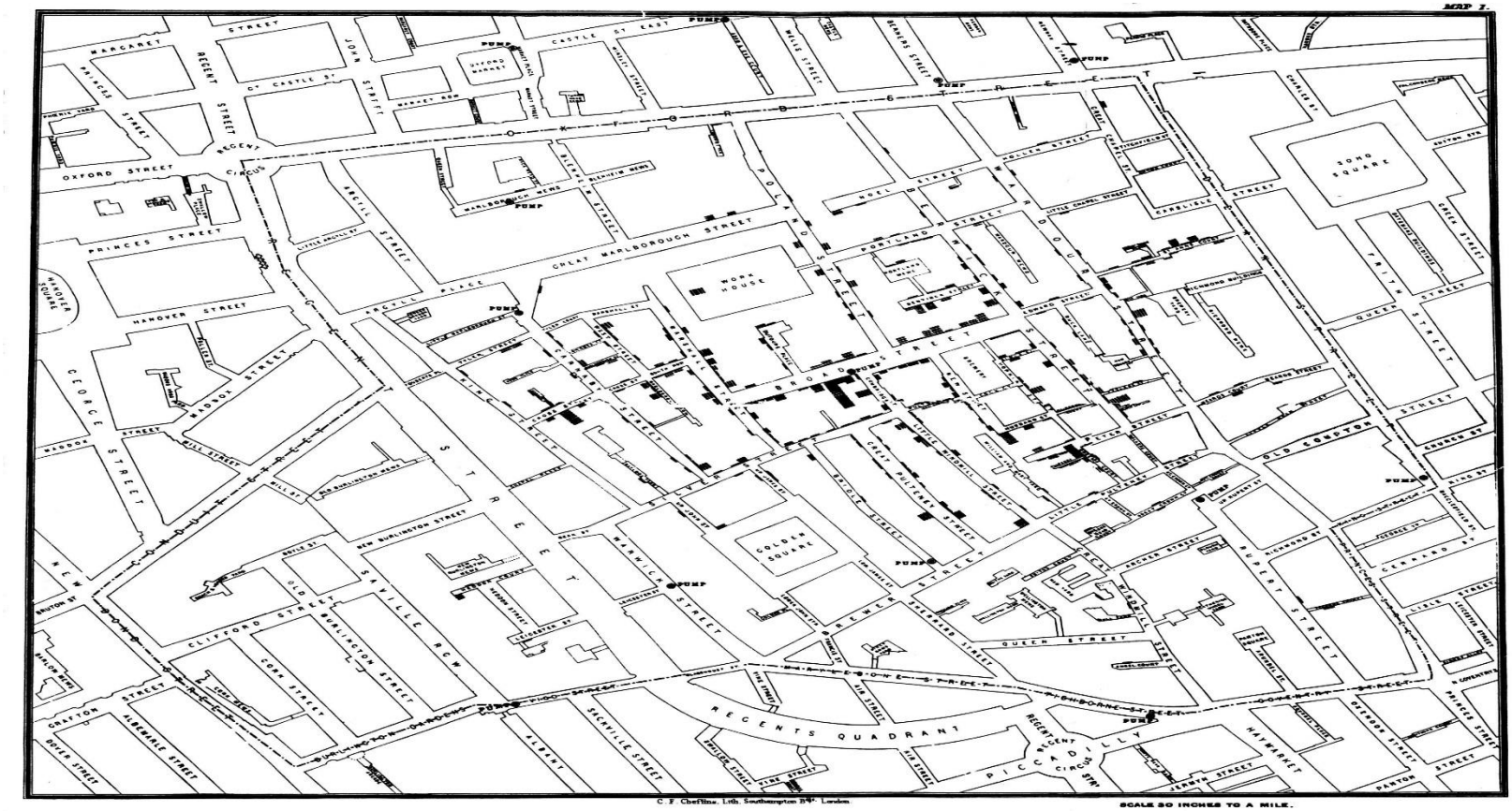
The Minard Map - "The best statistical graphic ever drawn"

https://en.wikipedia.org/wiki/Charles_Joseph_Minard **(Charles Minard Map)**

Learning Data Visualization by Bill Shanders, Lynda.com

So what makes a good visualisation

Accuracy, Story, **Knowledge**: Aim to create a visualisation that are accurate, tell a good story, and provide real knowledge to the audience.

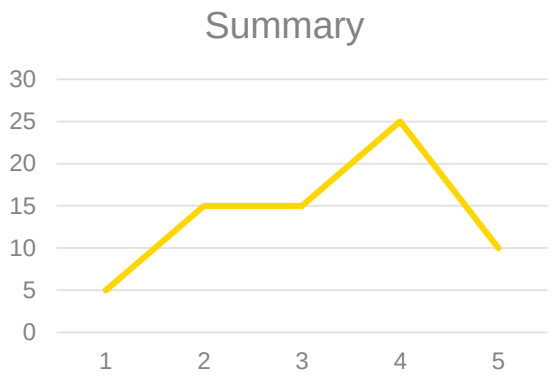


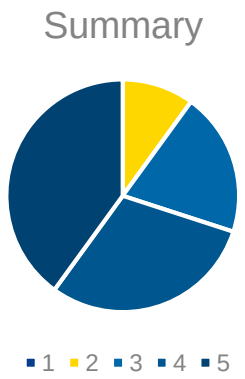
Widely attributed to creating the field of Epidemiology

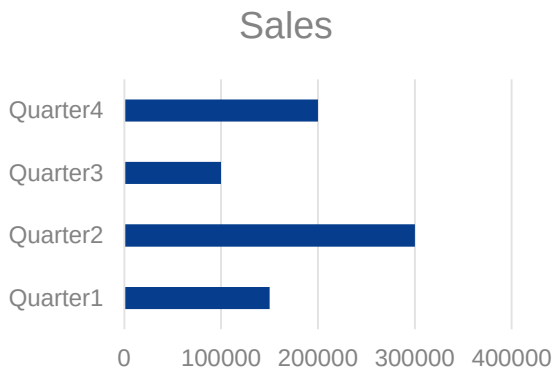
Some of the basics … “Charts vs. Graphs”

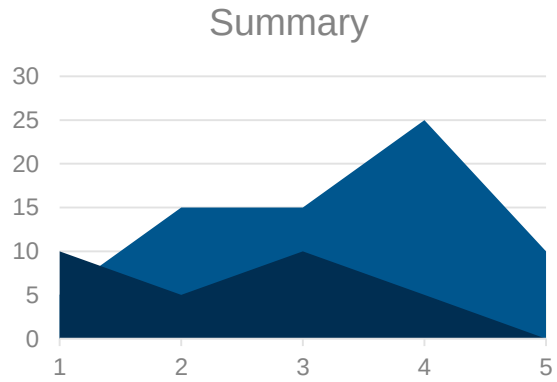
Often used interchangeably, but they are different in terms of the “visualisation techniques involved”.

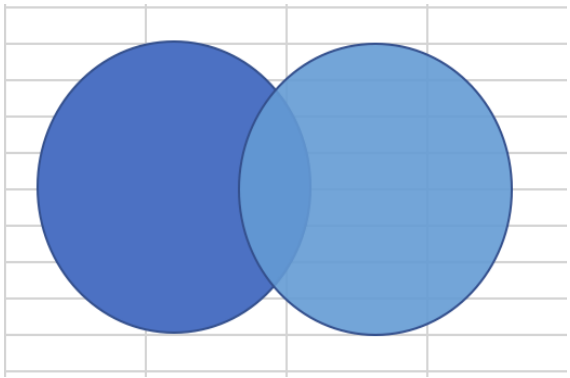
	Lights	Desk	Table	Monitor
Exhibit 1	✘	✓	✘	✘
Exhibit 2	✓	✓	✓	✘
Exhibit 3	✓	✘	✘	✓
Exhibit 4	✓	✓	✓	✓
Exhibit 5	✓	✓	✘	✘
Exhibit 6	✘	✘	✘	✓
Exhibit 7	✓	✓	✘	✓
Exhibit 8	✓	✘	✓	✘
Exhibit 9	✓	✘	✓	✘
Exhibit 10	✓	✓	✘	✘
Exhibit 11	✓	✓	✘	✓
Exhibit 12	✓	✘	✓	✓
Exhibit 13	✘	✘	✘	✘







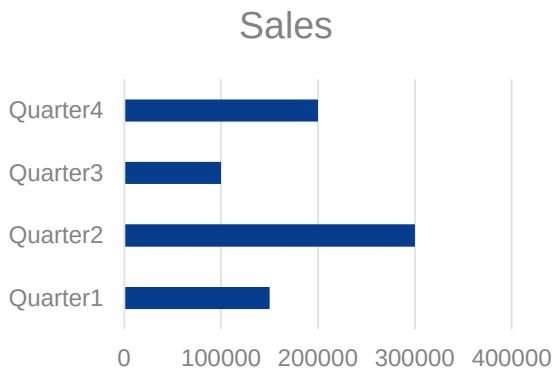




Data Visualization Best Practices: Killer Infographics by Amy Balliett,

Some of the basics … “Charts vs. Graphs”

	Lights	Desk	Table	Monitor
Exhibit 1	✘	✓	✘	✘
Exhibit 2	✓	✓	✓	✘
Exhibit 3	✓	✘	✘	✓
Exhibit 4	✓	✓	✓	✓
Exhibit 5	✓	✓	✘	✘
Exhibit 6	✘	✘	✘	✓
Exhibit 7	✓	✓	✘	✓
Exhibit 8	✓	✘	✓	✘
Exhibit 9	✓	✘	✓	✘
Exhibit 10	✓	✓	✘	✘
Exhibit 11	✓	✓	✘	✓
Exhibit 12	✓	✘	✓	✓
Exhibit 13	✘	✘	✘	✘



Both rely on an established, repeated pattern to show data

e.g., Bar: repeating equal width rectangles along a scale of information

e.g., Comparison Chart: repeating Tick/Cross along a scale of information

Graphs: rely on X or Y or both axes to make sense. At least one of these axes is numeric. Graph draws correlation between these axes by plotting points along the grid

Charts: not restricted by X/Y axes, not necessarily numerical

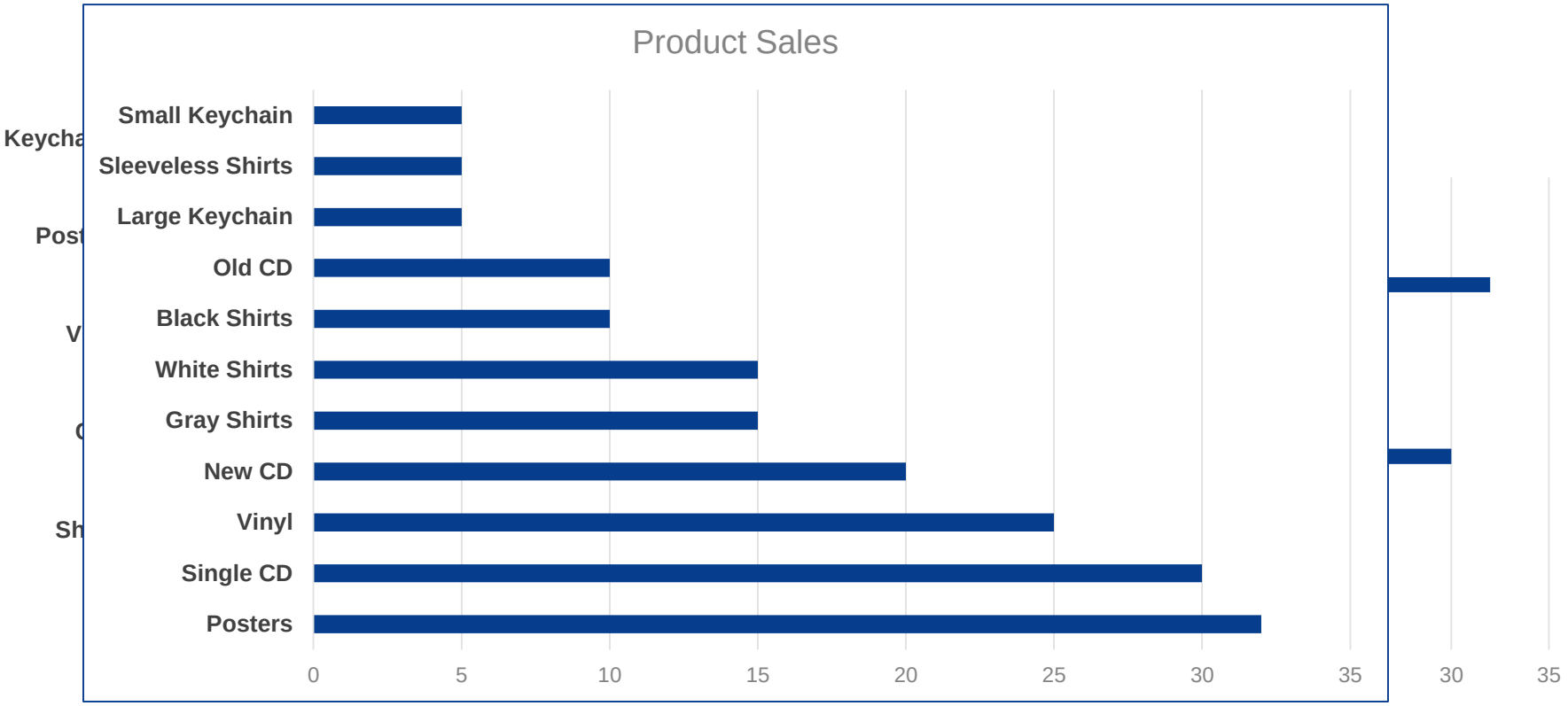
Which ones are charts? Which ones are graphs?

Data Visualization Best Practices: Killer Infographics by Amy Balliett,

Some of the basics … Organising data

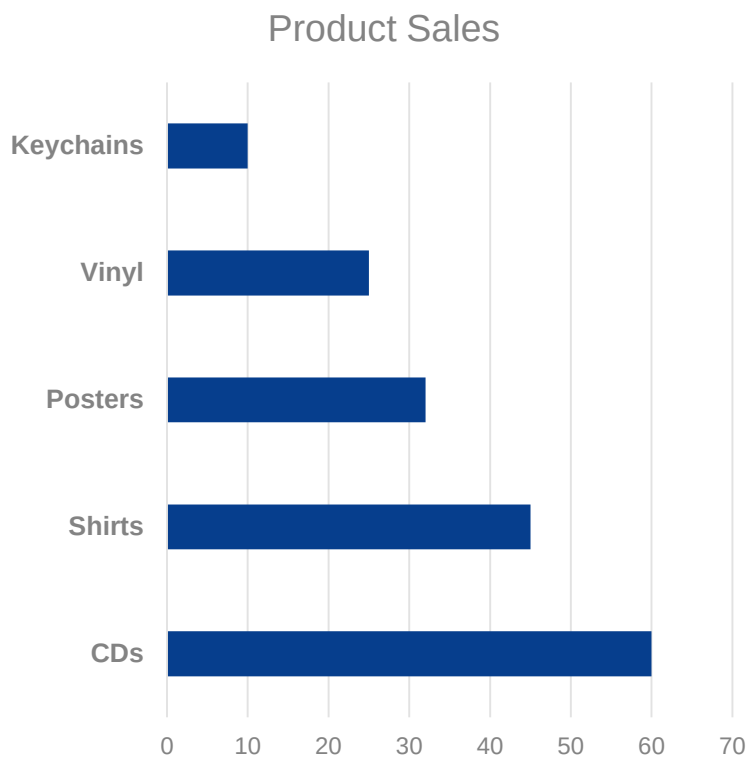
Imagine a merchandise sales figure by an artist:

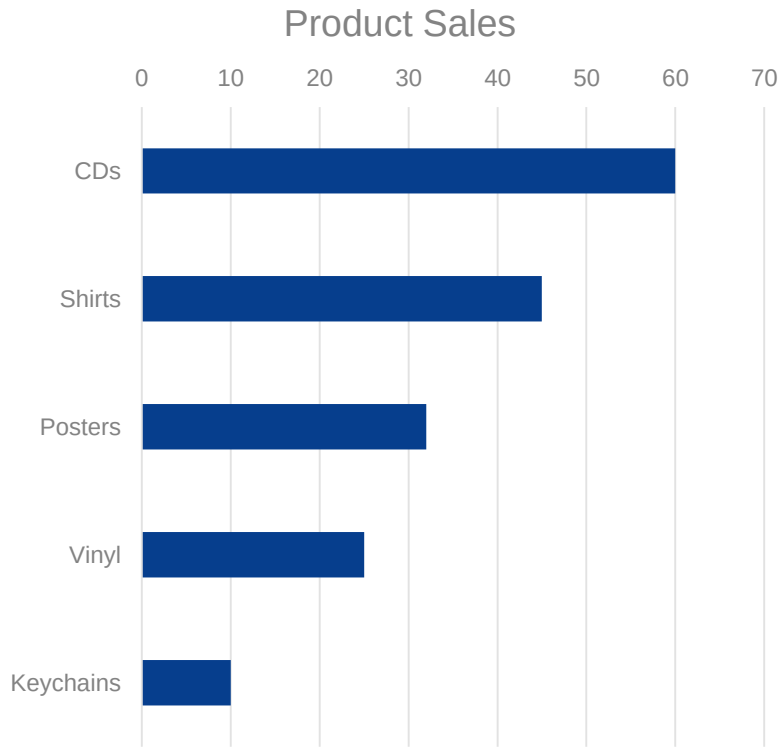
T-shirts: 45, CD: 60, Vinyl: 25, Posters: 32, Keychains: 10



Some of the basics … Organising data

When possible always order the data … but perception could be also tricky …



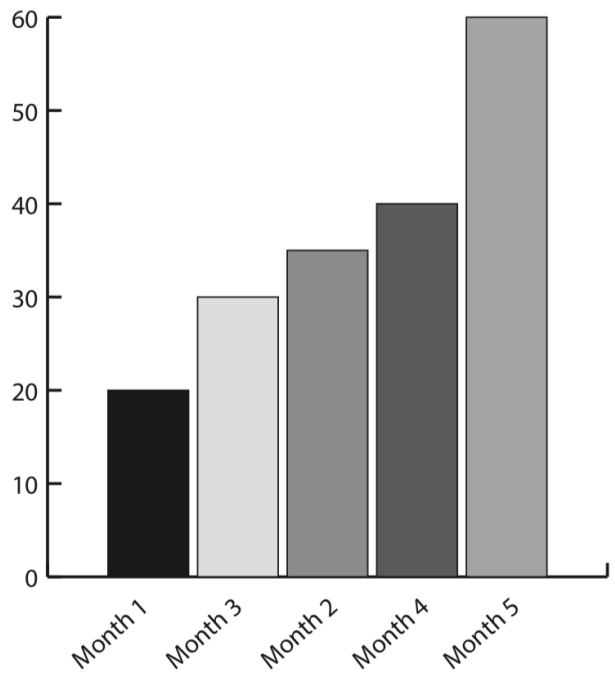


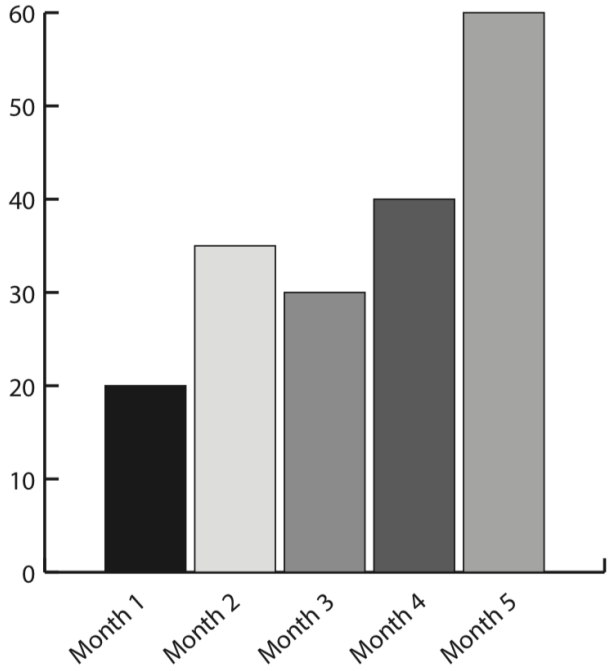
Subjective interpretation: most people read left to right …

Data Visualization Best Practices: Killer Infographics by Amy Balliett,

Some of the basics … Organising data

When possible always order the data … but perception could be also tricky …



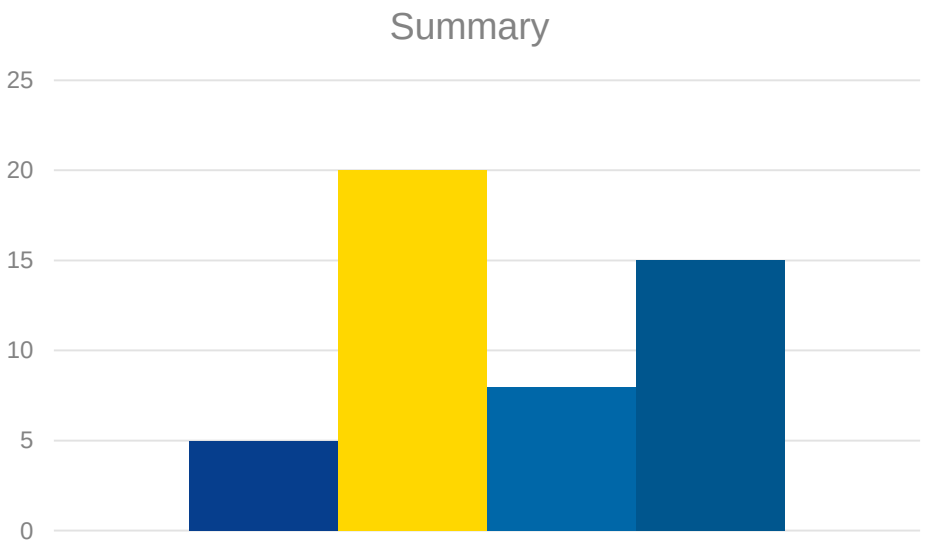


When a line follows a ”timeline”, stick to the timeline …

Data Visualization Best Practices: Killer Infographics by Amy Balliett,

Some of the basics … Colours important

Putting “Form (Prettiness)” before “Function”



Lots of colours could be confusing, distracting from the information

Just because you have millions of colors to choose from

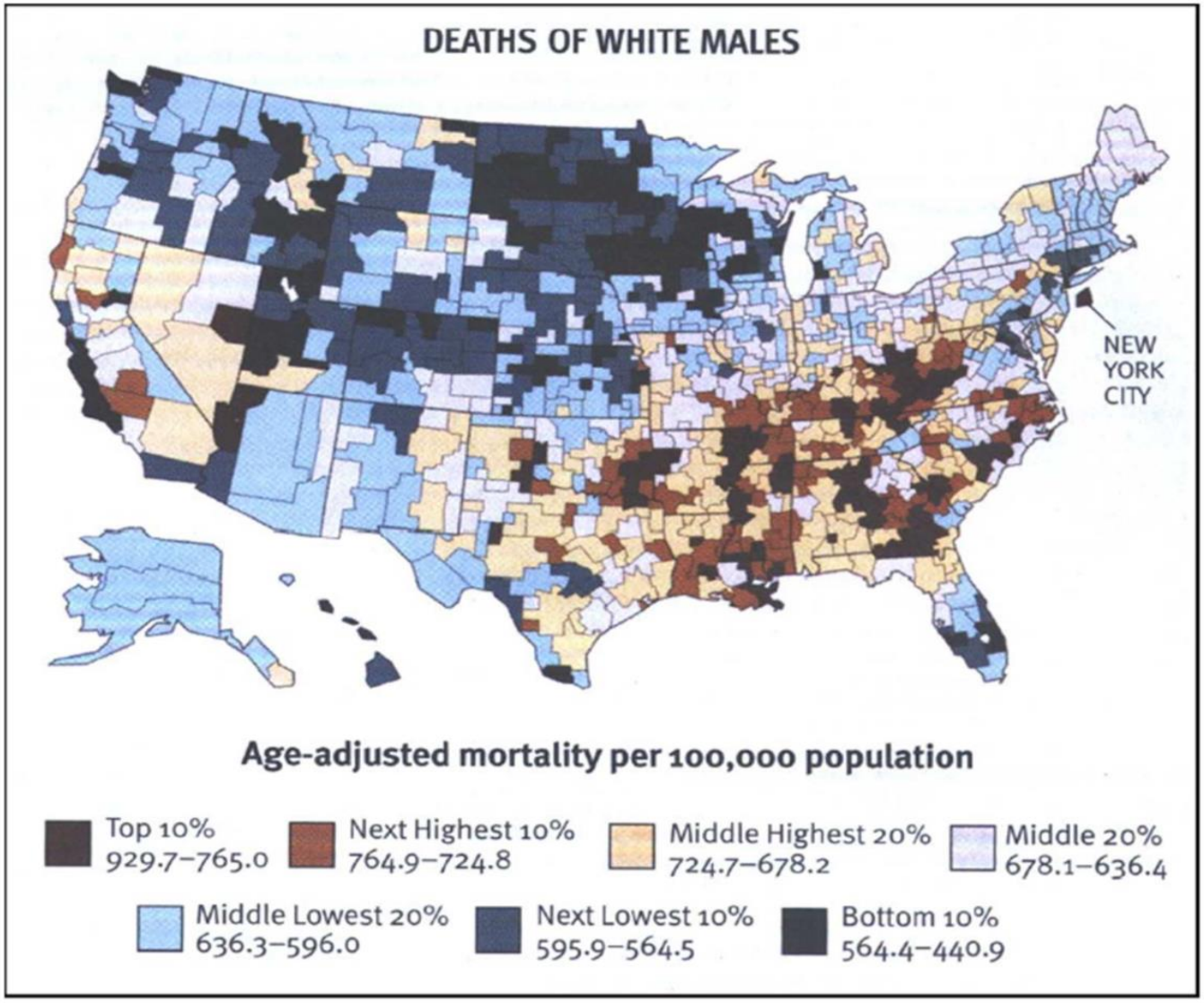
doesn't mean you must use them all •••

Data Visualization Best Practices: Killer Infographics by Amy Balliett,

Some of the basics … Colours important

Not just about limiting colours …

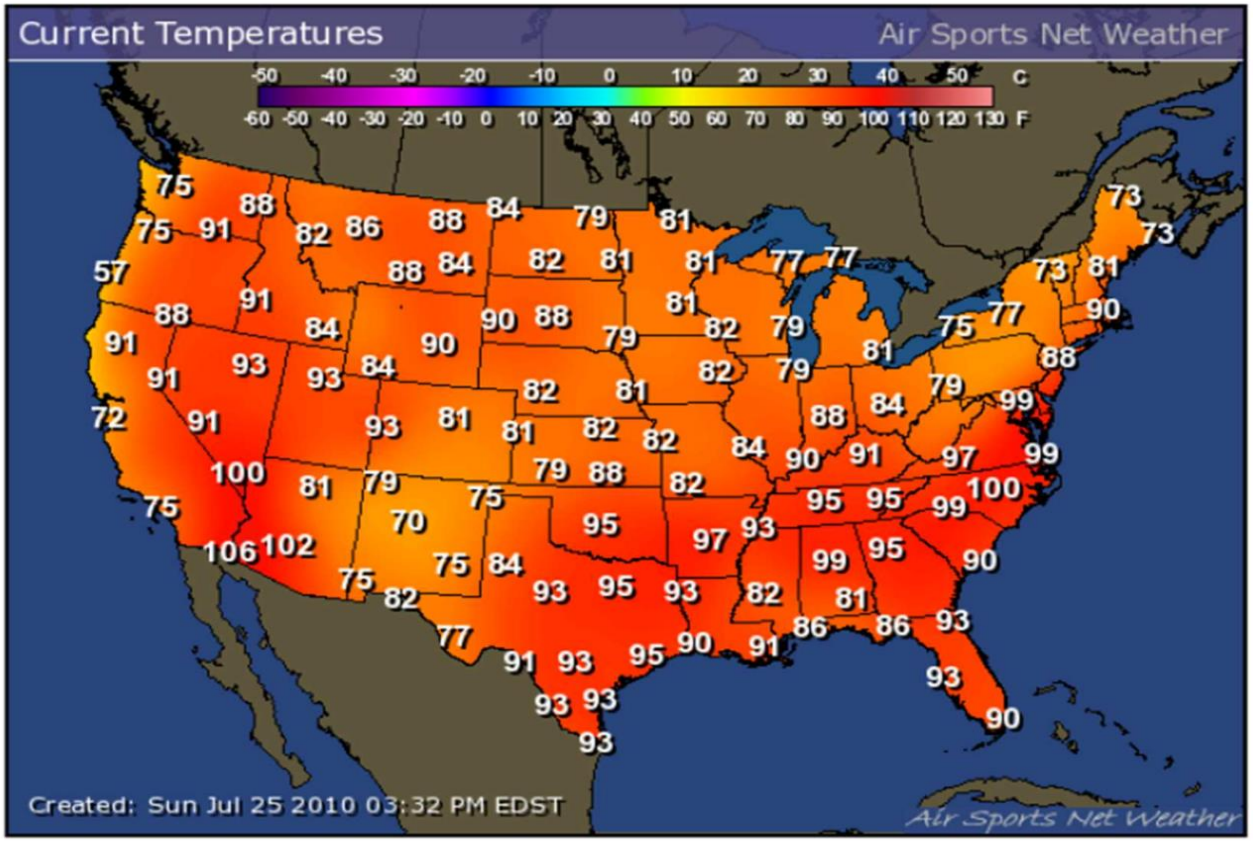
What's wrong with this colour map?



http://www2.cs.uh.edu/~chengu/Teaching/Fall2017/Visualization_fall2017.html

Some of the basics … Colours important

Not a bad choice of color scale,
but the Dynamic Range needs some work



http://www2.cs.uh.edu/~chengu/Teaching/Fall2017/Visualization_fall2017.html

Some of the basics … Colours important

Common advice on using colours:

most people have strong association with pre-established colour meanings. Don't go against them.

Red	Green	Blue
Stop	On	Cool
Off	Plants	Safe
Dangerous	Carbon	Deep
Hot	Moving	Nitrogen
High stress	Money	
Oxygen		
Shallow		
Money loss		

http://www2.cs.uh.edu/~chengu/Teaching/Fall2017/Visualization_fall2017.html

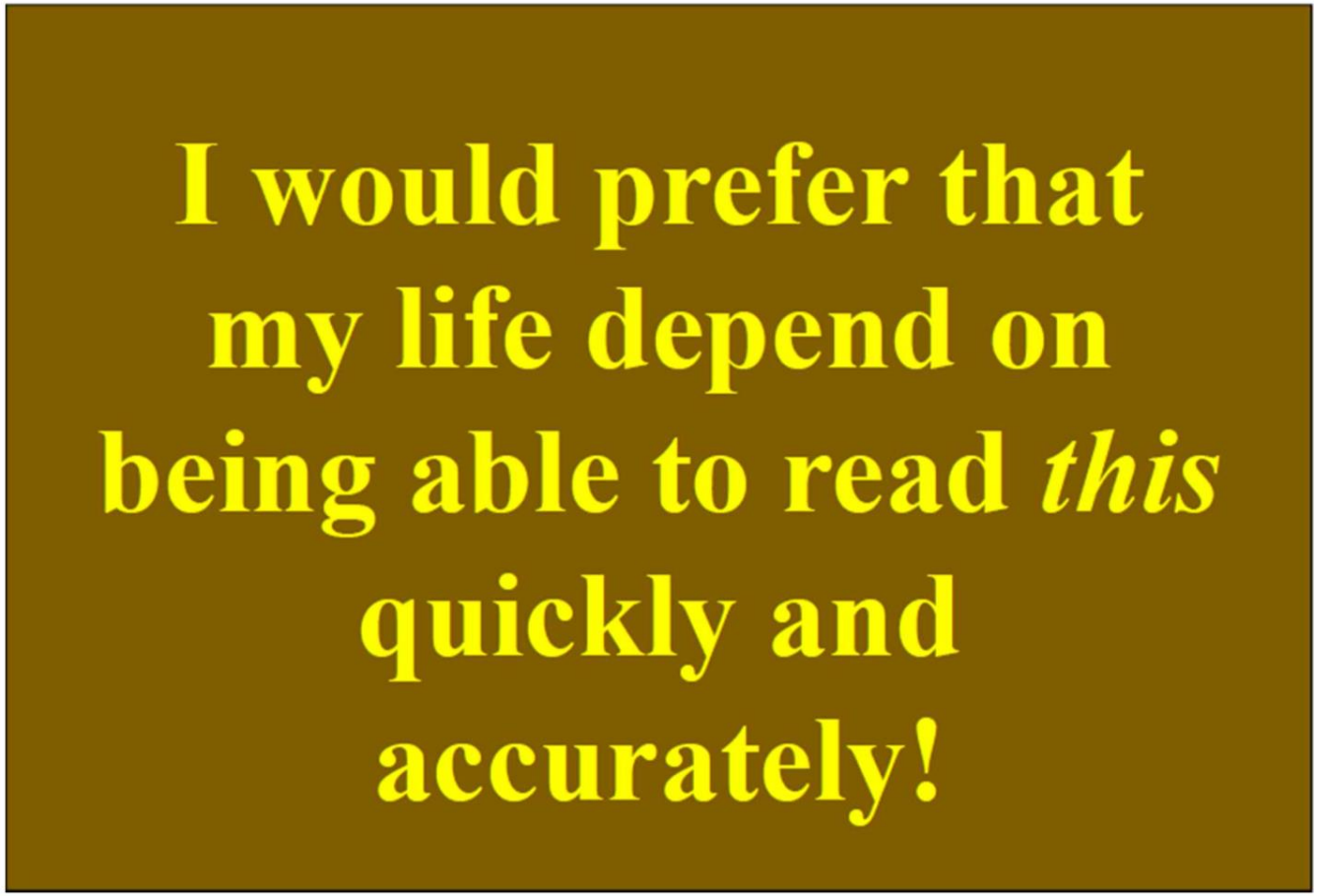
Some of the basics … Colours important

Just to make things interesting …. Colour alone is not the whole picture.

http://www2.cs.uh.edu/~chengu/Teaching/Fall2017/Visualization_fall2017.html

Some of the basics … Colours important

What about correct contrast choice?



http://www2.cs.uh.edu/~chengu/Teaching/Fall2017/Visualization_fall2017.html

Some of the basics … Colours important

Colour blindness is more common than we think … (close to 10% of generic population)

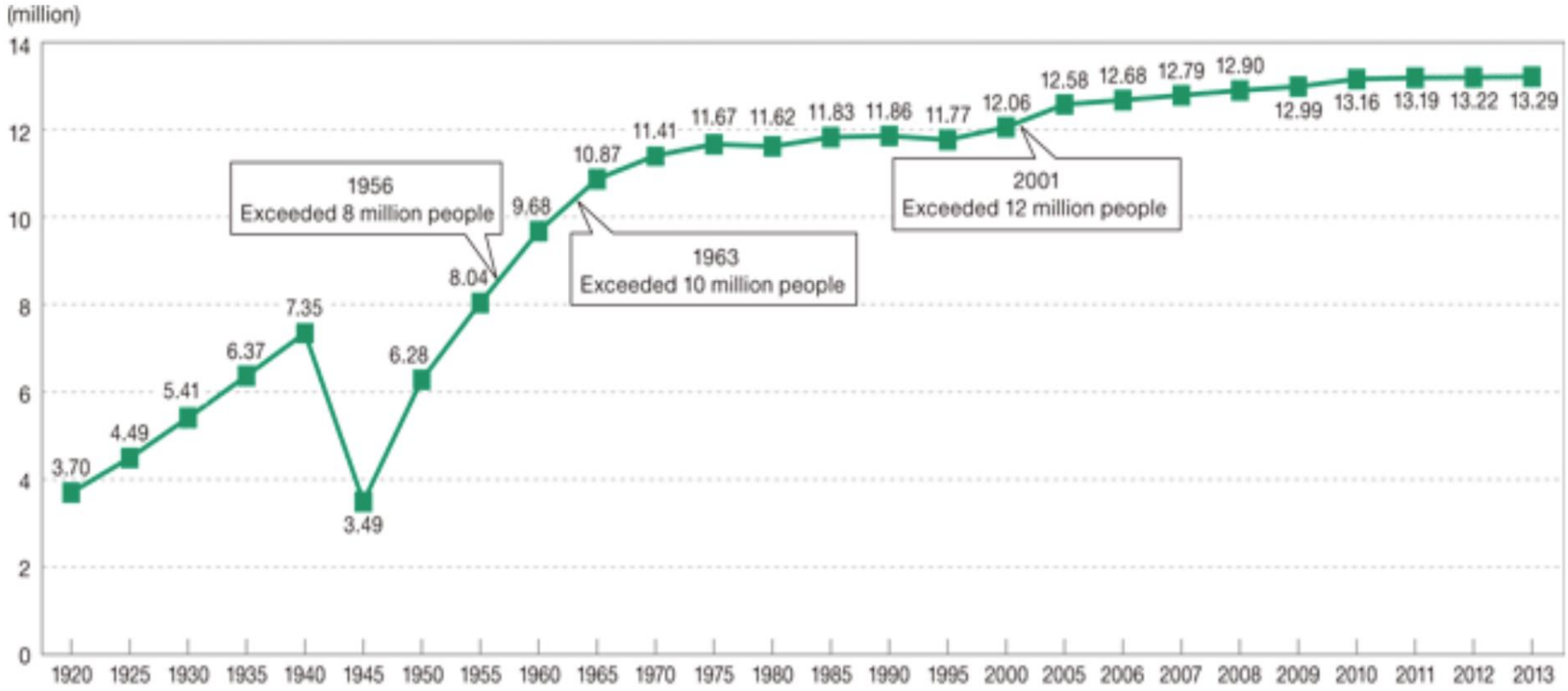
There are resources you could utilise:

e.g., http://colorbrewer2.org/

Tips on increasing accessibility of your visualisation:

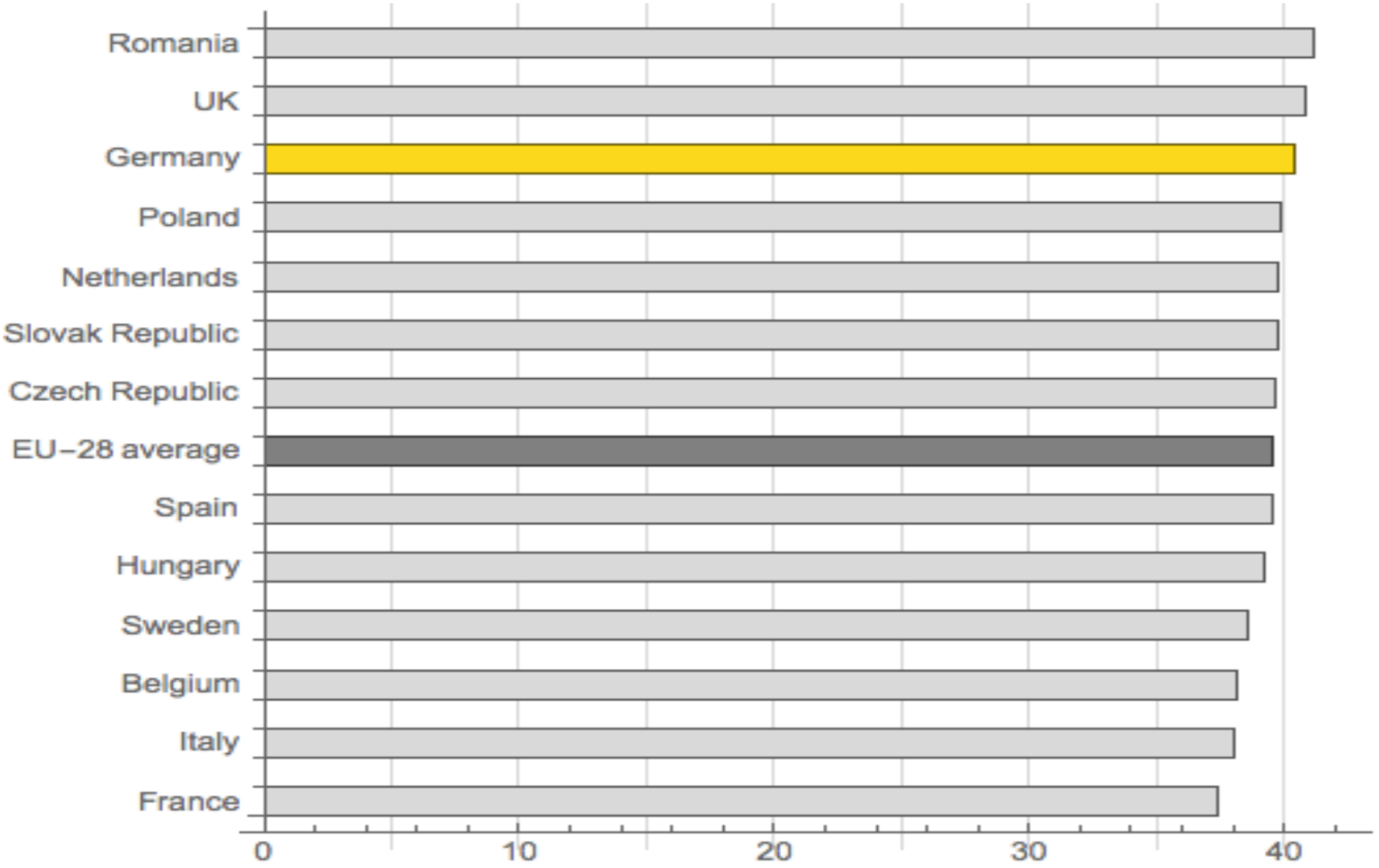
http://blog.usabilla.com/how-to-design-for-color-blindness/

Some of the basics … Keep Scales consistent



https://www.livestories.com/blog/five-ways-to-fail-data-visualization

Some of the basics … Keep scales consistent

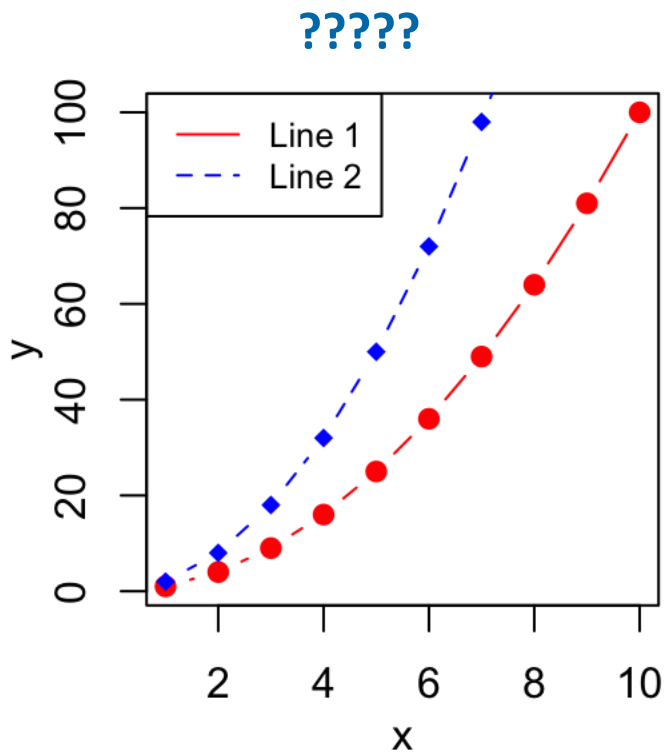


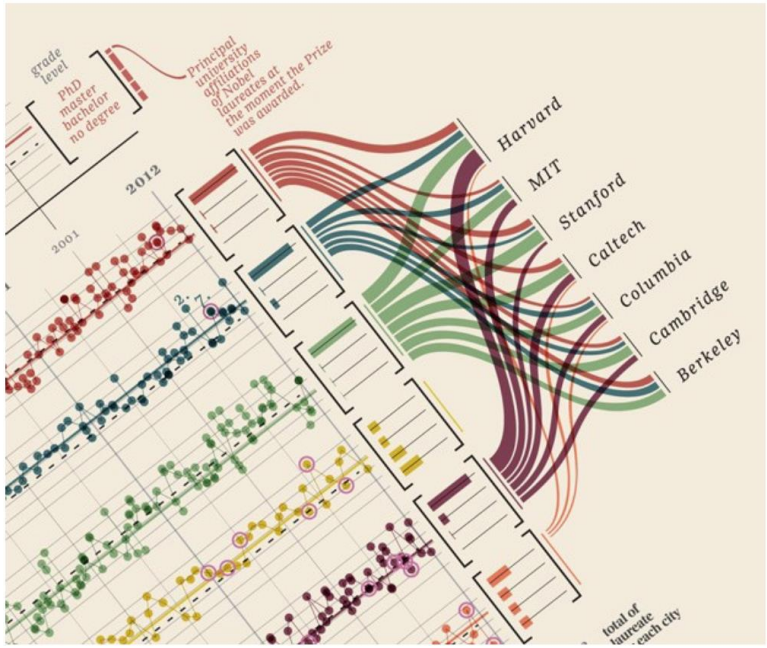
http://callingbullshit.org/tools/tools_misleading_axes.html

Some of the basics … Legend and Sources

It is absolutely necessary to include the sources of your data and correct legends to interpret your visualisation.

Without the legends, your visualisation is basically a “pretty picture” with no meaning





Preparing the data for visualisation

Data almost never comes in the exact form that you need it in

One of the common parts of the data visualisation tasks is to clean and convert the data

Some of the common data adjustments:

- Calculating indexes and ratios
- Calculating percentile
- Aggregating
- Regrouping
- Converting from Excel/CSV to JSON/XML/SQL

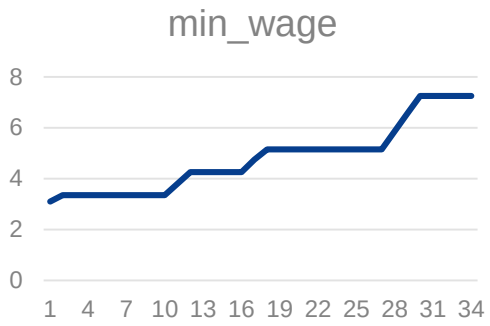
You may need simple tools, or database SQL scripting, programming scripting solutions for these.

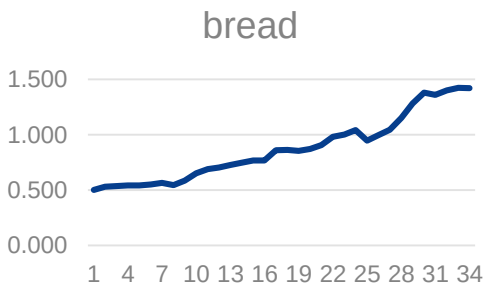
- Excel (available!)
- Tableau (commercial)
- Direct data manipulation with SQL or programming language

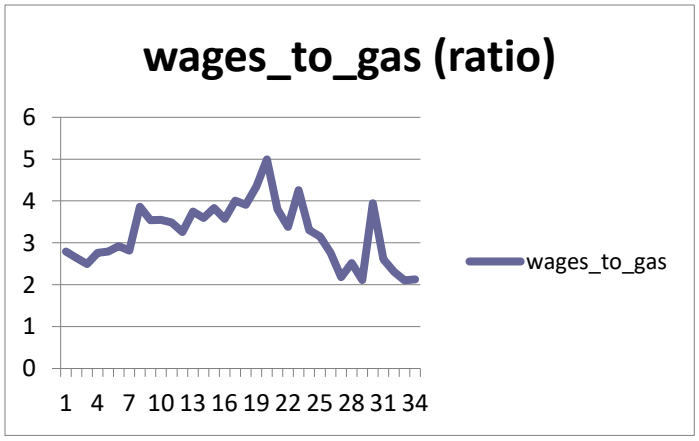
Preparing the data: Indexes and Ratios

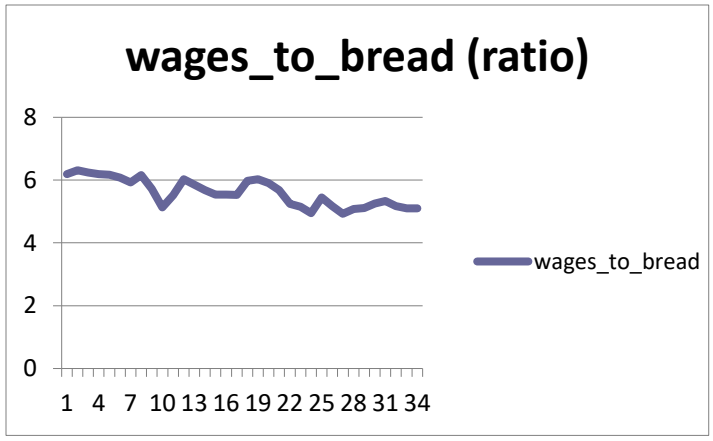
Are we comparing apples with apples (or are they apples and oranges?)

* Indexes and Ratios allow you to convert the data in a way that makes it easy to look at data side-by-side that is not necessarily easy to do in the original form









min_wage/gas

min_wage/bread

Preparing the data: Calculating Percentile

Calculating percentile makes it easier to compare numbers to each other as part of a whole
-- where you stand compared to the rest of the herd, ***relative standing***

	A	B	C	G
1	Country Name	2016	Rank	Percentile
2	United States	$ 18,624,475,000,000	1	99%
3	China	$ 11,199,145,157,649	2	99%
4	Japan	$ 4,940,158,776,617	3	98%
5	Germany	$ 3,477,796,274,497	4	98%
6	United Kingdom	$ 2,647,898,654,635	5	97%
7	France	$ 2,465,453,975,282	6	97%
8	India	$ 2,263,792,499,341	7	96%
9	Italy	$ 1,858,913,163,928	8	96%
10	Brazil	$ 1,796,186,586,414	9	95%
11	Canada	$ 1,529,760,492,201	10	95%
12	Korea, Rep.	$ 1,411,245,589,977	11	94%
13	Russian Federation	$ 1,283,162,985,989	12	94%
14	Spain	$ 1,237,255,019,654	13	93%
15	Australia	$ 1,204,616,439,828	14	93%
16	Mexico	$ 1,046,922,702,461	15	92%
17	Indonesia	$ 932,259,177,765	16	92%
18	Turkey	$ 863,711,710,427	17	91%
19	Netherlands	$ 777,227,541,581	18	91%
20	Switzerland	$ 668,851,296,244	19	90%

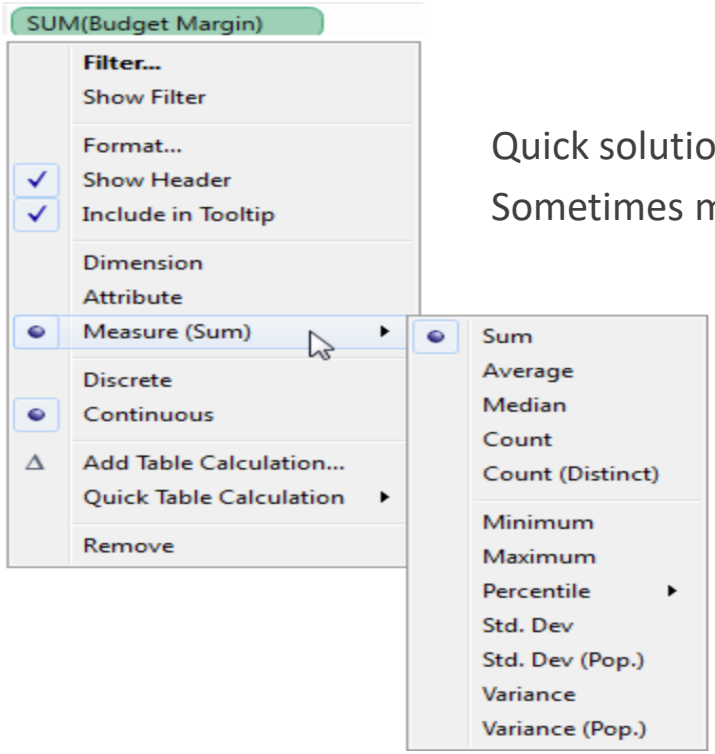
1-(ranking/total county)

1-(1/191)
1-(2/191)
1-(3/191)
…

Preparing the data: Aggregating, Converting Data

Data aggregation is the process where raw data is gathered and expressed in a summary form for statistical analysis

For example, raw data can be aggregated over a given time period to provide statistics such as ***average, minimum, maximum, sum, and count.***



Quick solution by a tool like Tableau/Excel …

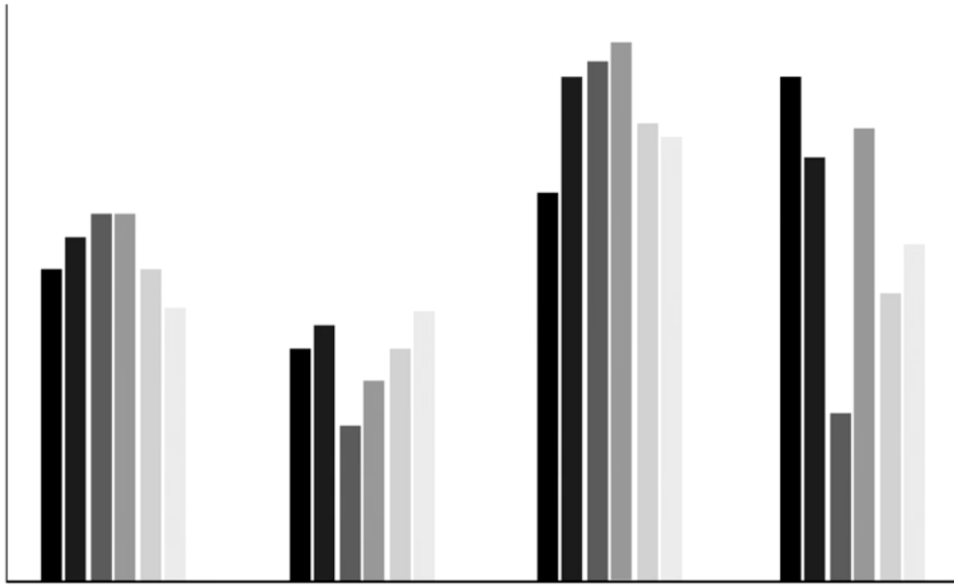
Sometimes manual SQL scripts necessary

The right paradigm

One of the most difficult things to do is figuring out which charts/graphs to use in which situation. I'd say you need to have the basic competency here – and then be aware of good alternatives.

The good old BAR graphs:

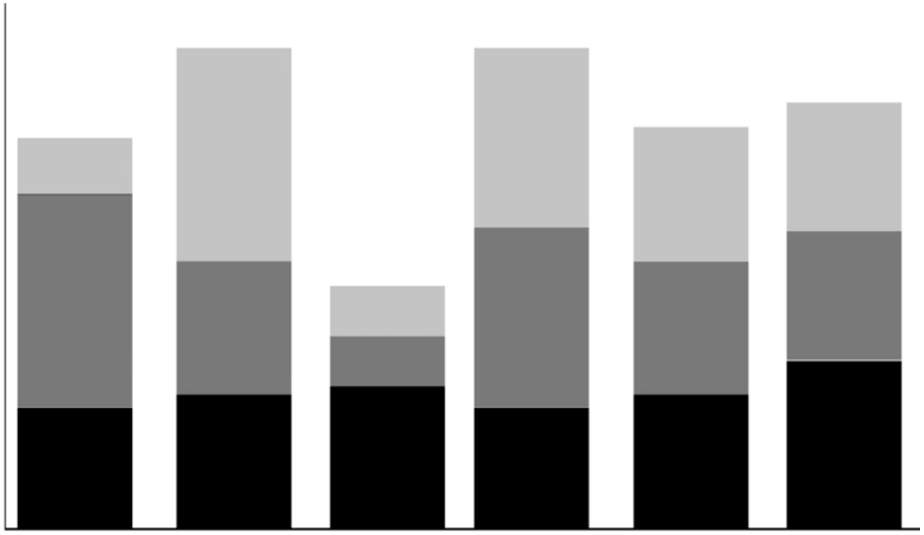
- Highly effective in terms of 'parsing the information'
- Experts say human brain is wired to differentiate these rectangular shapes
- In fact, you should start by asking "why isn't a bar graph enough here?"
- But when there are more variables, many "grouped" bars don't look good



The right paradigm

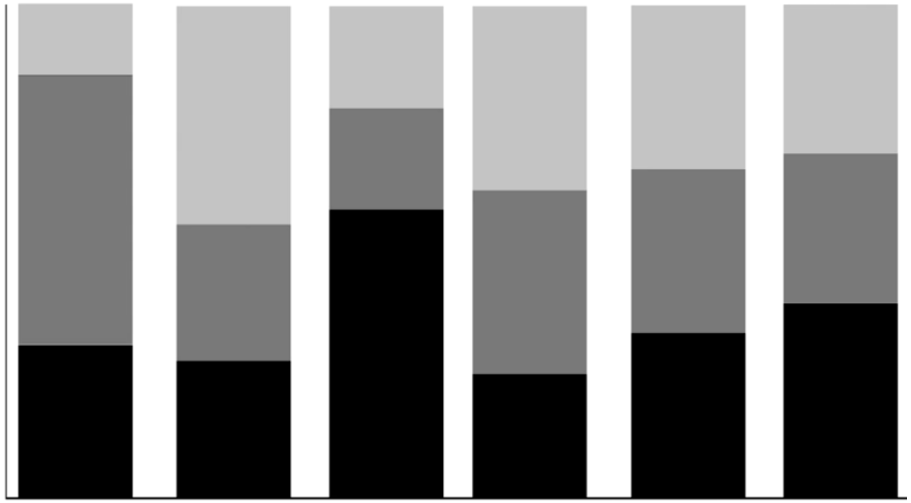
STACKED Bar graphs:

- Compare data within groups
- Whole bar represents the total value of that group, and each segment represents the value within the group

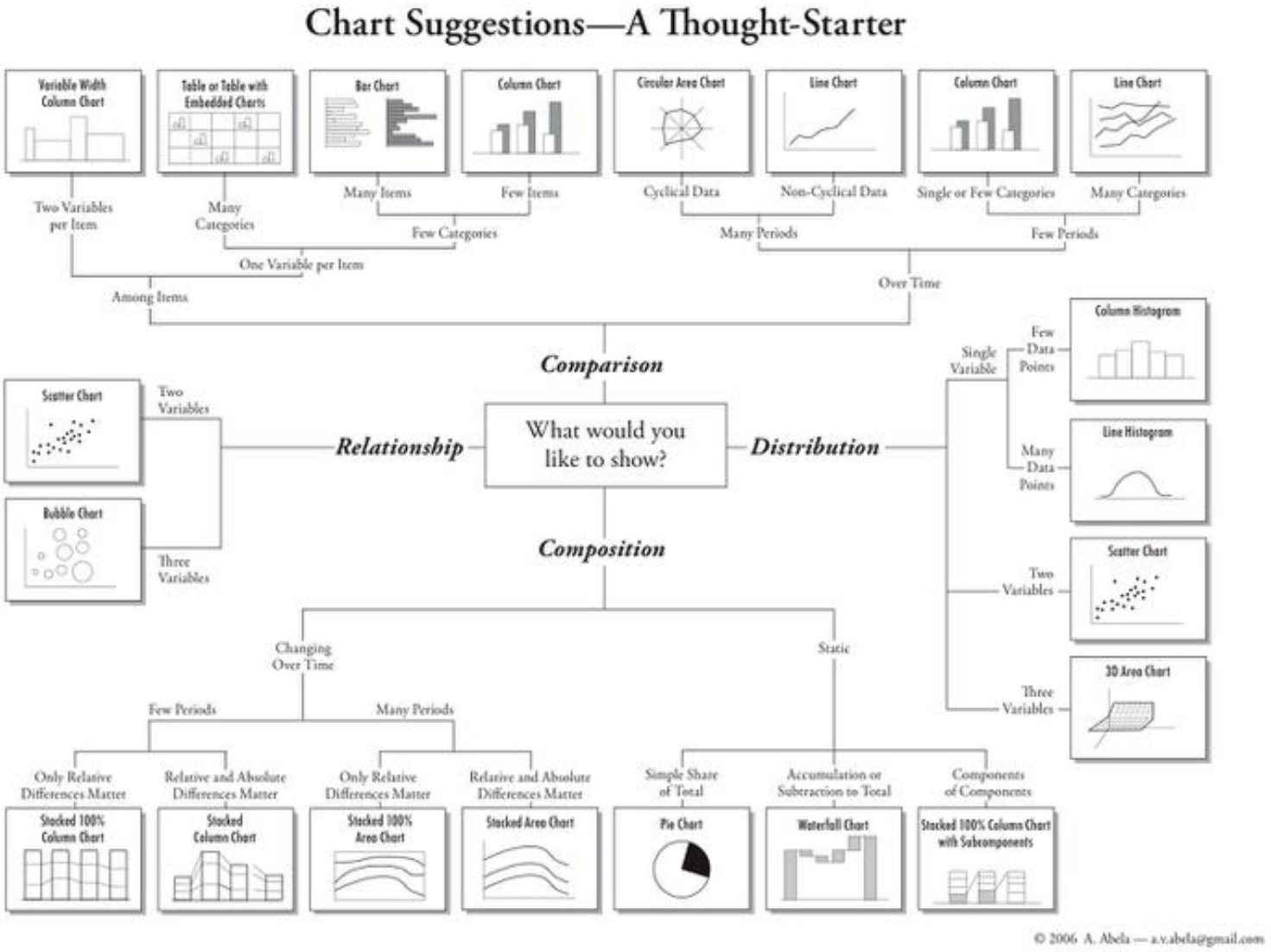


STACKED Percentage Bar graphs:

- If you want to compare relative contribution of each category to the whole ….
- Whole bar = 100%
- Showing relative strength of each category within the whole

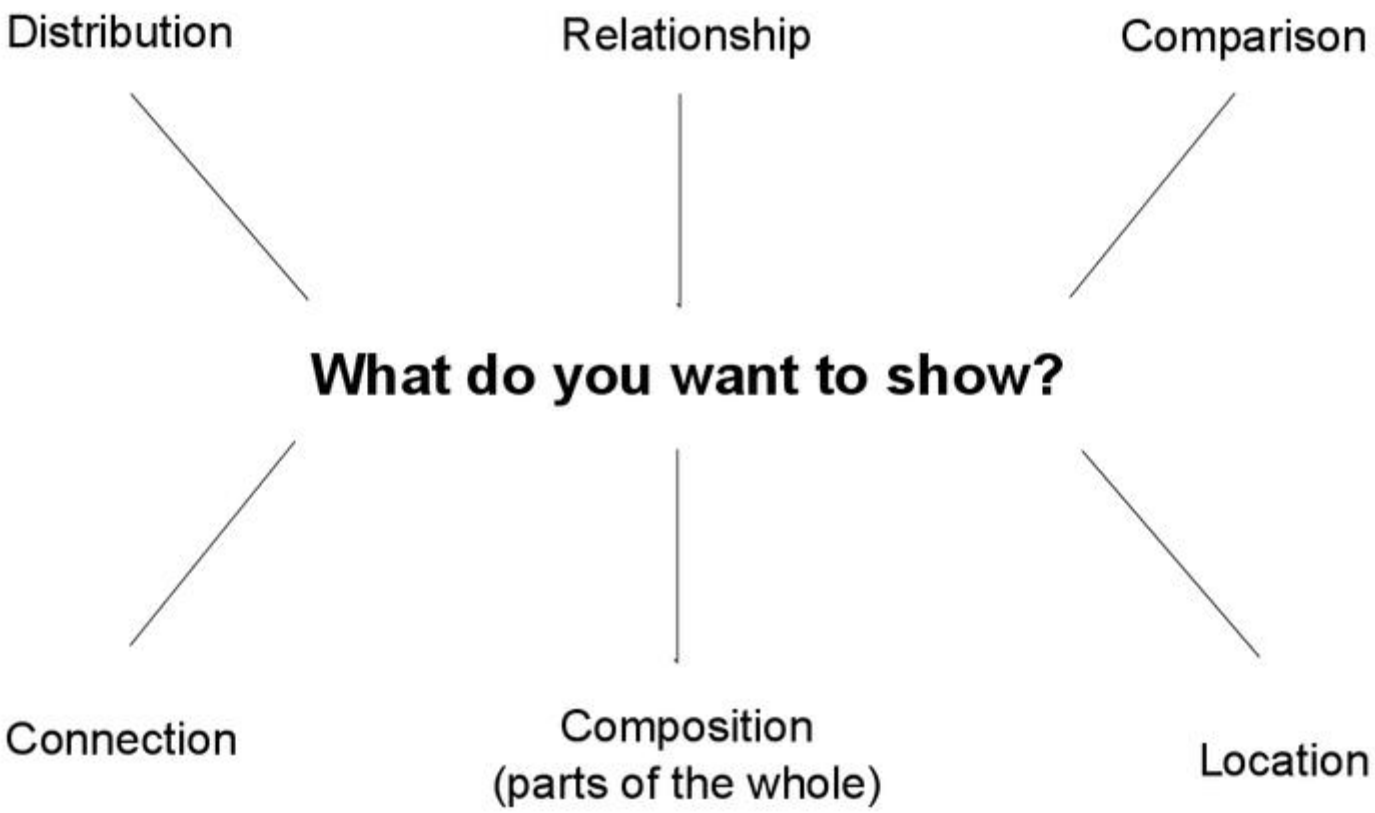


The right paradigm



http://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

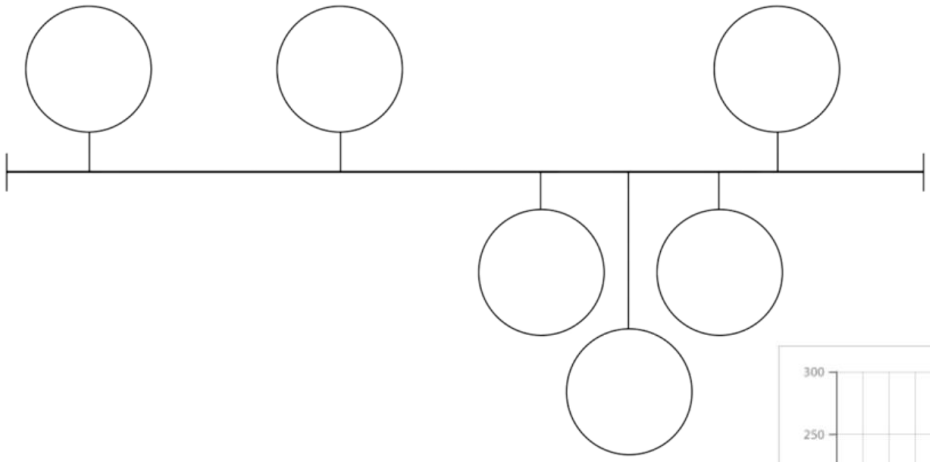
The right paradigm

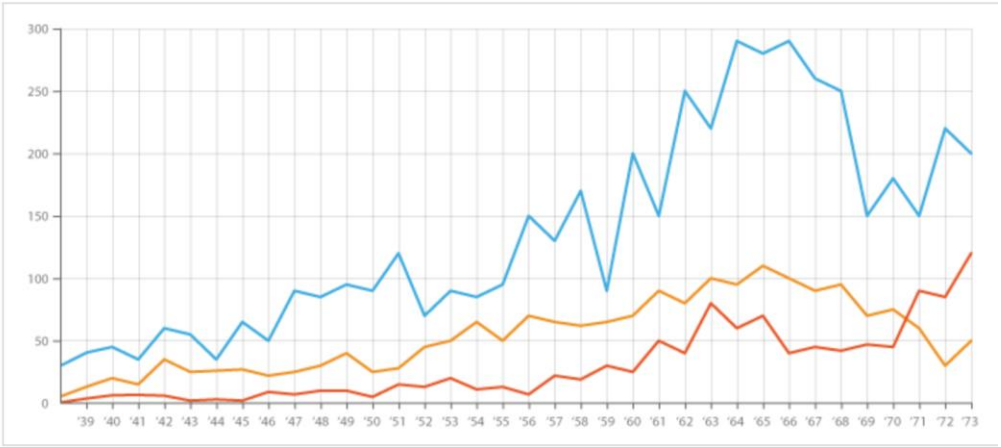


The right paradigm

Line graphs:

- To show values “over time” or a continuous interval
- Stories over “Timeline”

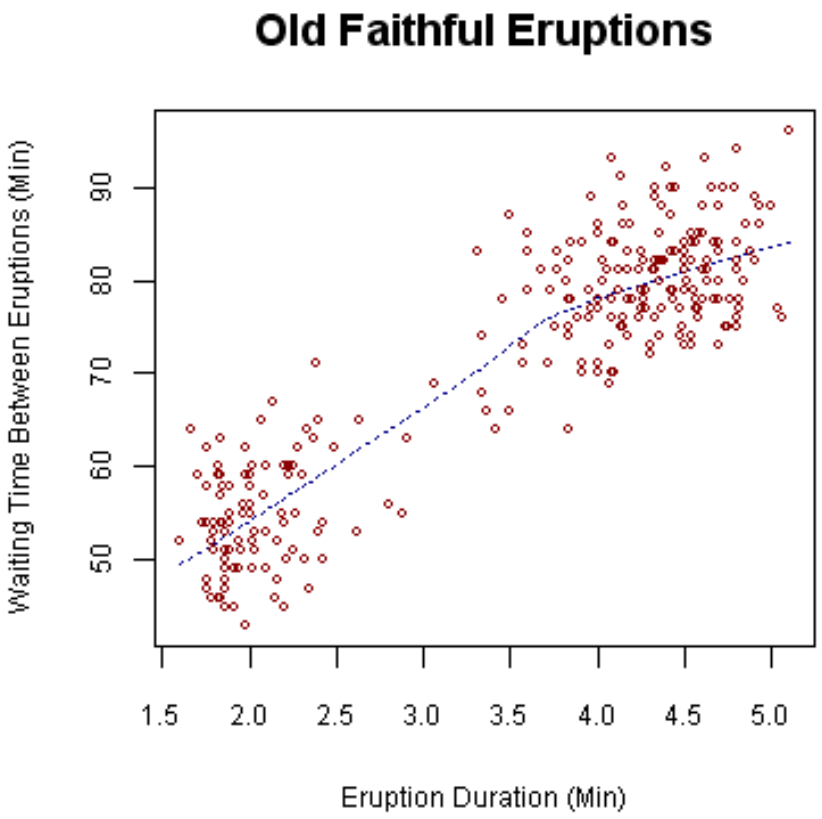




The right paradigm

Scatter Plot graphs:

- To show two variables and their correlations (i.e., X axis vs. Y axis)



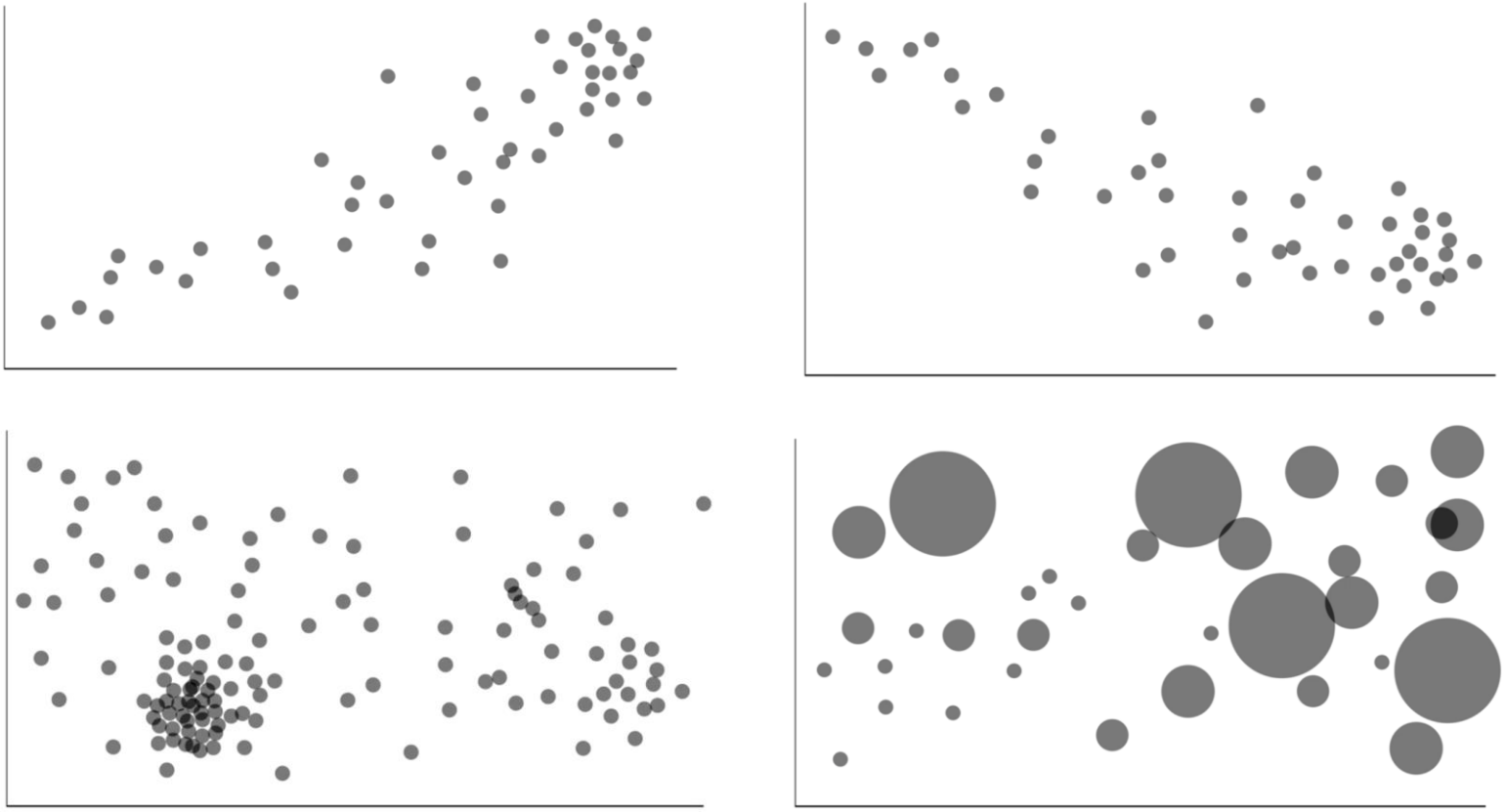
https://en.wikipedia.org/wiki/Scatter_plot

Learning Data Visualization by Bill Shanders, Lynda.com

The right paradigm

Scatter Plot graphs:

- To show two variables and their correlations (i.e., X axis vs. Y axis)



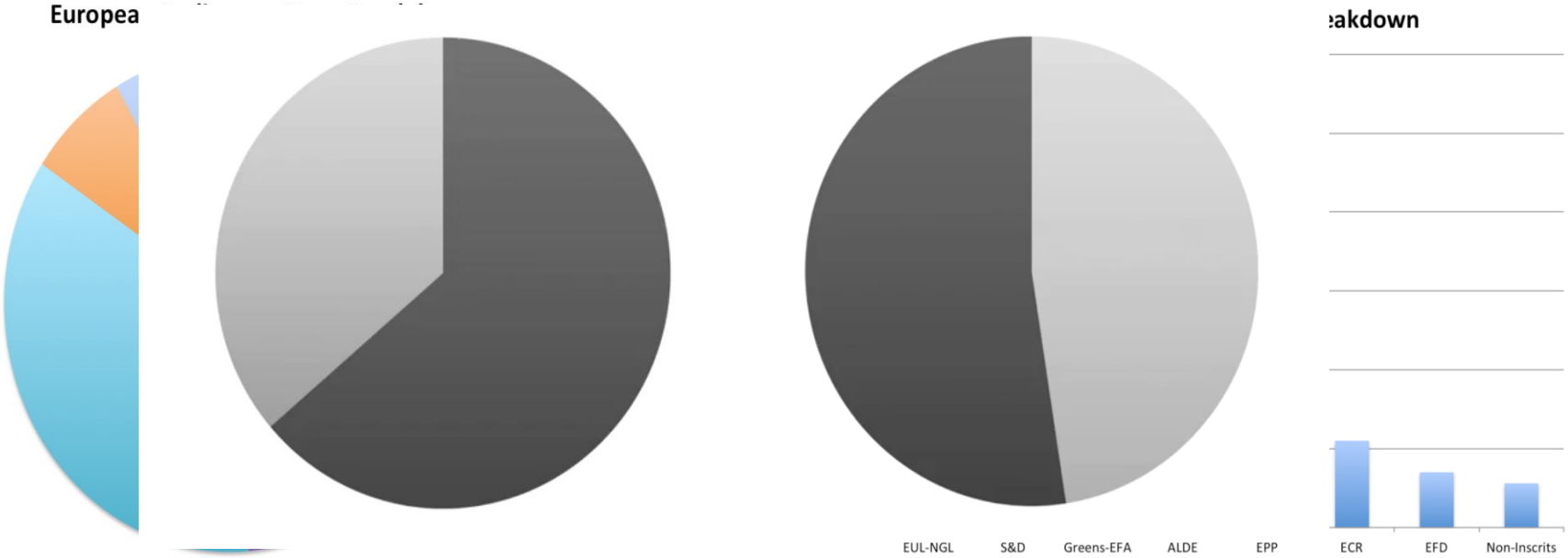
No correlations, but discernible patterns

Size of the dots – third variable

The right paradigm

Pie Charts:

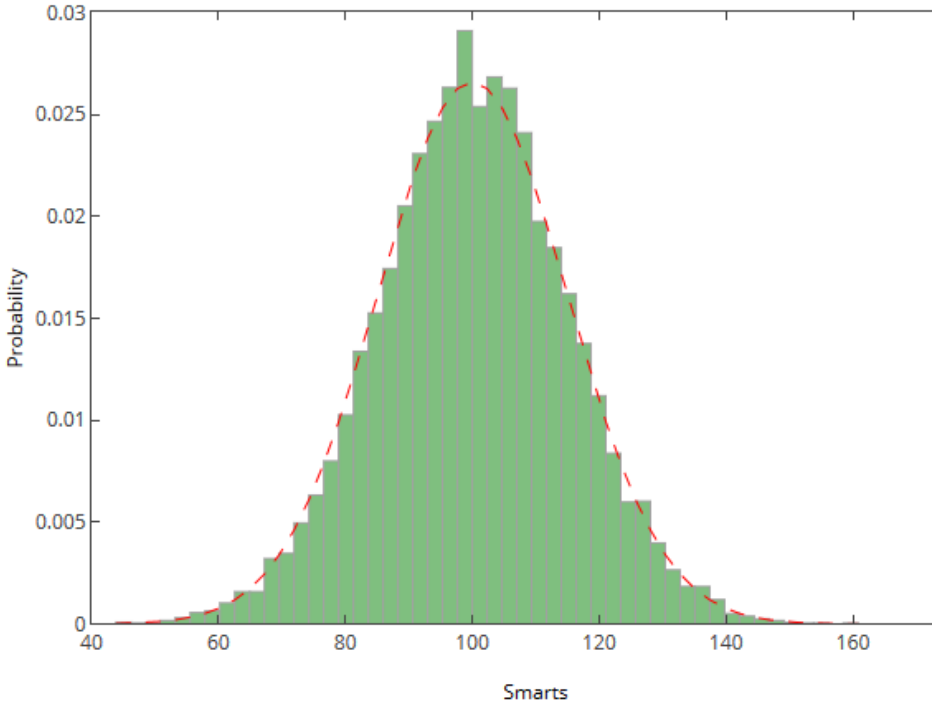
- Although commonly used, not considered an effective form. Human brain is not wired to parse round shape areas and arcs
- Normally other graphs can do the same job (e.g., BAR graphs)
- Maybe OK when showing two variables (<, >, similar, etc.)



The right paradigm

Histograms

- Histograms are useful for viewing (or really discovering)the distribution of data points
- The use of bins (discretization) really helps us see the “bigger picture” where as if we use all of the data points without discrete bins, there would probably be a lot of noise in the visualization, making it hard to see what is really going on



The right paradigm: hierarchical data

To show the “connections” between and the hierarchy of objects.

A tree diagram:

http://mbostock.github.io/d3/talk/20111018/tree.html

- Default for showing hierarchy (e.g., org chart)
- Any situation where you a parent, which has children (and grand children)

A node link diagram:

http://mbostock.github.io/d3/talk/20111116/force-collapsible.html

- Showing a lot of links between objects

Tree map:

http://mbostock.github.io/d3/talk/20111018/treemap.html

- Size of each category

Chord Diagram:

https://bost.ocks.org/mike/uberdata/

Complex data -> very difficult to parse the information (e.g., between dots far apart) Interactivity could help parse (https://bost.ocks.org/mike/fisheye/)

The right paradigm: showing data on maps

On an existing map API like Google API …

Place markers (https://www.latlong.net)

- specific location (e.g., building), centre of a region

Layers (data associated with the regions on a map)

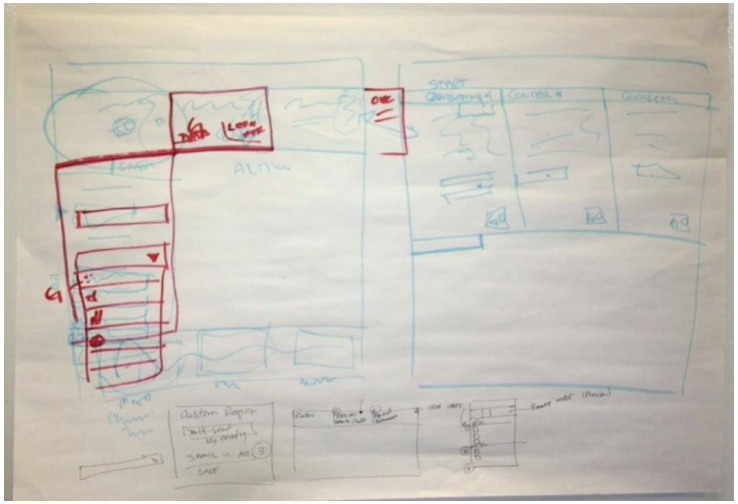
- Point clustering (http://bl.ocks.org/andrewxhill/raw/8360694/)
  - display aggregated number/data points per region
- Choropleth map (http://leafletjs.com/examples/choropleth/)
  - display divided geographical areas or regions that are coloured, shaded or patterned in relation to a data variable.
- Heat map (https://onemilliontweetmap.com/)
- Flow map
  - show the movement of information or objects from one location to another and their amount (thickness of lines, colours)
  - https://datavizcatalogue.com/methods/flow_map.html
  - https://www.iom.int/world-migration

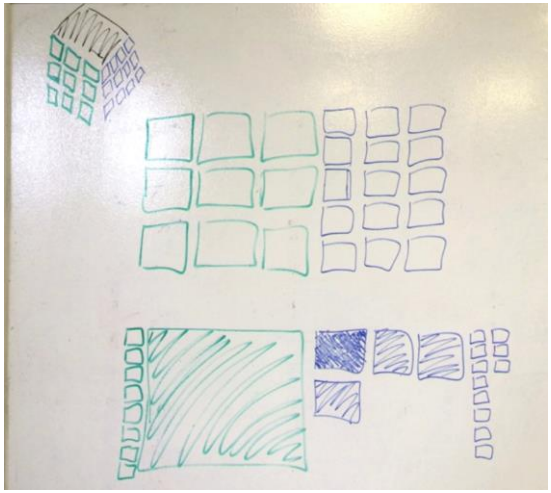
To put it together …

So … there are many many options for visualising data (including a lot of fancy and interactive ones from the latest tools and libraries)

But let's try to have some basic competency on this:

- Accuracy is important, having a clear story to tell is important
- You need to be ready to do some basic data prep and pre analysis before visualisation
- Knowing the right paradigm (form) to use for the story
- Aware of your own limitation as 'non-expert' (visualisation is not easy)





Actually, a lot of experts recommend "sketching the idea out" with pen and paper.

Data Visualization using Matplotlib

- There are many excellent plotting libraries in Python and I recommend exploring more than one in order to create presentable graphics.
- In this course we are going to introduce the Matplotlib library. It is the foundation for many other plotting libraries and plotting support in higher-level libraries such as Pandas.
- Don’t get confused with Matplotlib’s many ways of plotting the same thing. Pandas is our access point

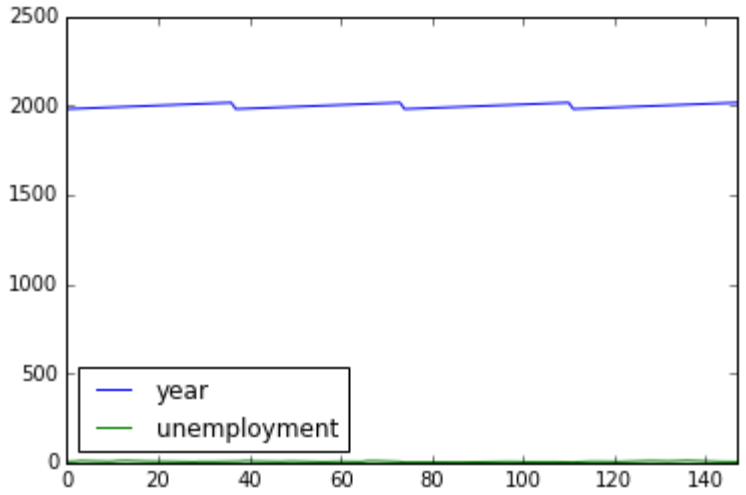
Matplotlib and Dataframes

- Under the hood, pandas plots graphs with the matplotlib library. This is usually pretty convenient since it allows you to just .plot your graphs.
- When you use .plot on a dataframe, you sometimes pass things to it and sometimes you don’t.
- .plot plots the index against every column
- .plot(x='col1') plots against a single specific column
- .plot(x='col1', y='col2') plots one specific column against another specific column

Matplotlib and Dataframes

	country	year	unemployment
35	Australia	2015	6.063658
36	Australia	2016	5.723454
37	USA	1980	7.141667
38	USA	1981	7.600000
39	USA	1982	9.708333

If you use: df.plot()



http://jonathansoma.com/lede/algorithms-2017/classes/fuzziness-matplotlib/understand-df-plot-in-pandas/

Matplotlib and Dataframes

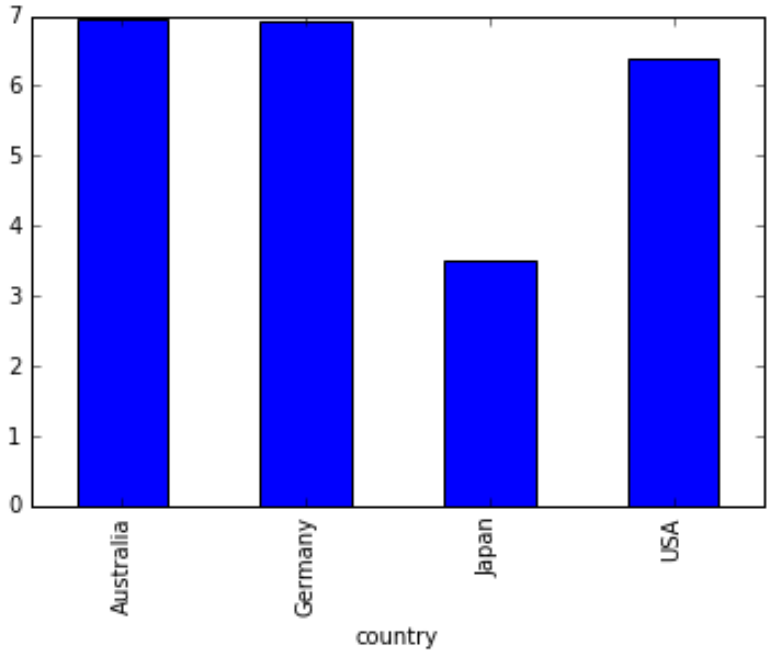
- The major use cases for .plot() is when you have a meaningful index, which usually happens in two situations:
    - You've just done a .value_counts() or a .groupby()
    - You've used .set_index, probably with dates

Matplotlib and Dataframes

- So if you do:

df.groupby("country")['unemployment'].mean().plot(kind ='bar’)

You’ll get:



Matplotlib and Dataframes

- What about (.plot(x='col1', y='col2’))
- Let’s try it for the same data before:

df.plot(x='year', y='unemployment')

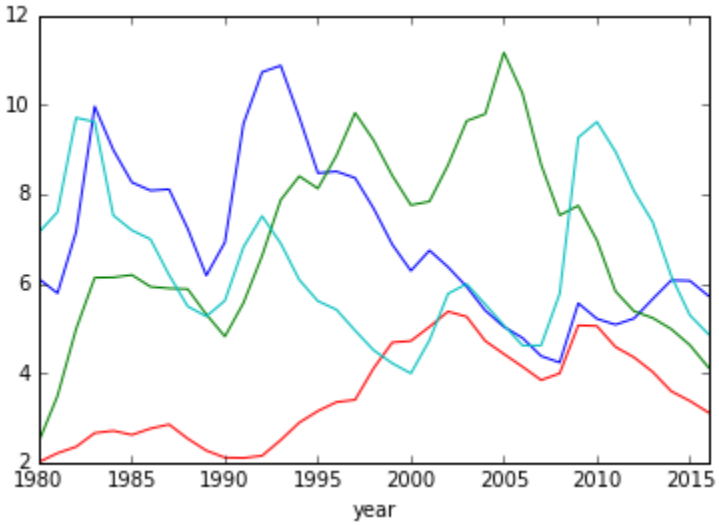
Talk about connected

Matplotlib and Dataframes

- Groupby to do it right.
- Great a single graph

```
fig, ax = plt.subplots()

df.groupby('country').plot(x='year', y='unemployment', ax=ax, legend=False)
```



Matplotlib without dataframes

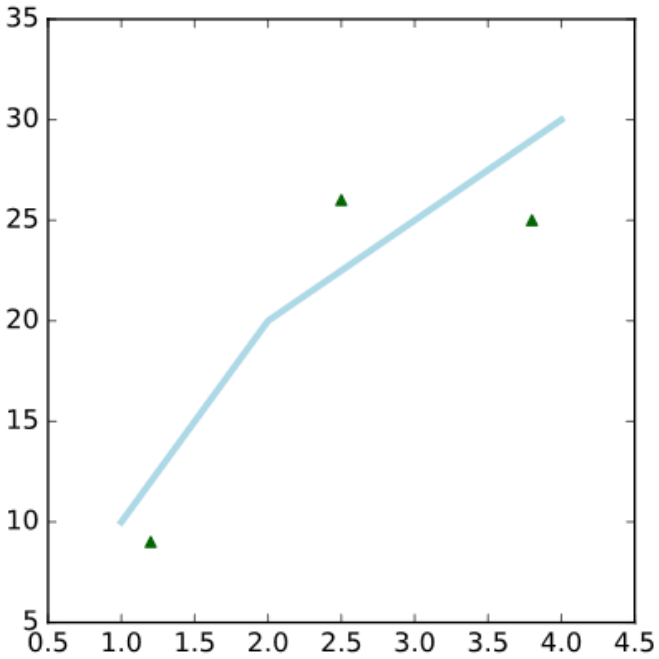
```
import matplotlib.pyplot as plt

plt.plot([1, 2, 3, 4], [10, 20, 25, 30], color='lightblue', linewidth=3)

plt.scatter([0.3, 3.8, 1.2, 2.5], [11, 25, 9, 26], color='darkgreen',
marker='^')

plt.xlim(0.5, 4.5)

plt.show()
```



Matplotlib Conclusion

- Many details and scattered around documentation
- Stackoverflow is King
- DO YOUR LABS

Useful Read

- Book: the Functional Art by Alberto Cairo (Chapter 1,2, and 3)
- http://jonathansoma.com/lede/algorithms-2017/classes/fuzziness-matplotlib/how-pandas-uses-matplotlib-plus-figures-axes-and-subplots/
- https://pythonspot.com/visualize-data-with-pandas/