HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

IT4142E
Introduction to Data Science

Chapter 5: Introduction to Data Visualization

Lecturer:

Muriel VISANI: murielv@soict.hust.edu.vn

Acknowledgements:

Khoat Than
Viet-Trung Tran

Department of Information Systems
School of Information and Communication Technology - HUST

Contents of the course

- Chapter 1: Overview
- Chapter 2: Data scraping
- Chapter 3: Data cleaning, pre-processing and integration
- Chapter 4: Introduction to Exploratory Data Analysis
- Chapter 5: Introduction to Data visualization
- Chapter 6: Introduction to Machine Learning
  - Performance evaluation
- Chapter 7: Introduction to Big Data Analysis
- Chapter 8: Applications to Image and Video Analysis

Goals of this chapter

Goal	Description of the goal
M1	**Understand and be able to design and manage the systems which are based on Data Science (DS)**
M1.3	Be able to design systems based on DS in their future organizations
M2	**Identify and manage the opportunities from DS to boost the existing organizations, or develop new organizations**
M2.1	Understand and promote the use of DS to support their organizations
M2.2	Identify the (possible) impacts of Data Science on their organizations

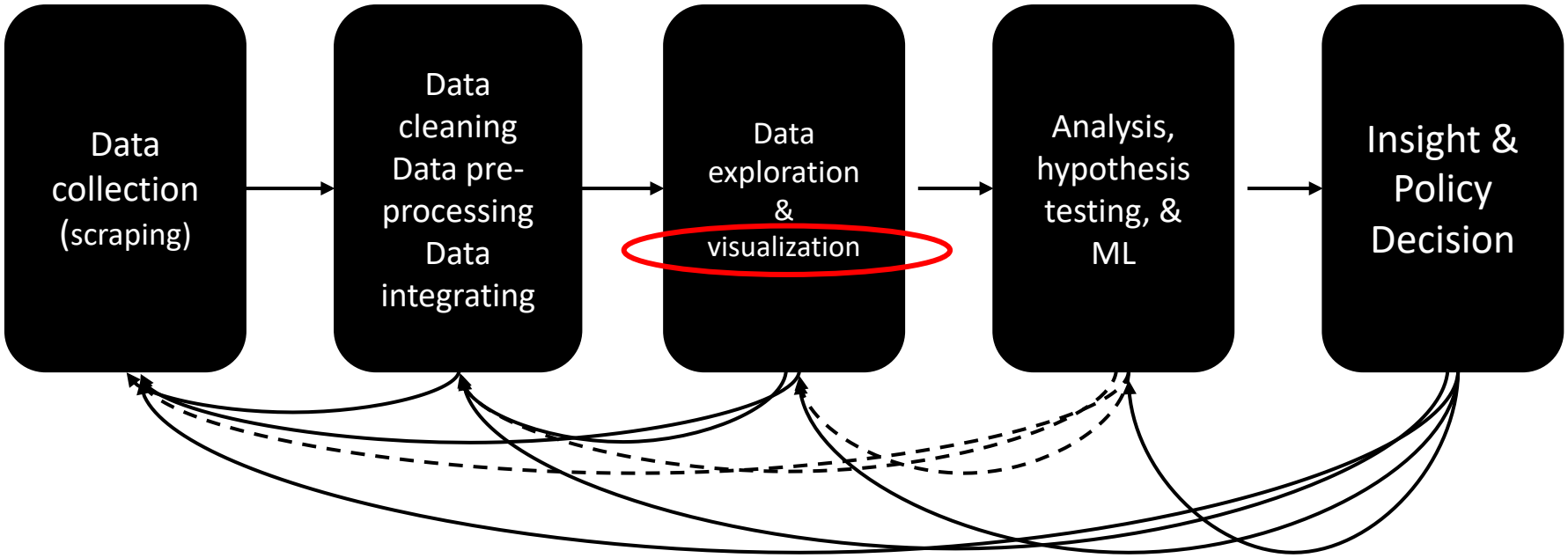
Contents of this chapter

- **Chapter 5: Introduction to Data Visualization**
  - Introduction
    - Definition of data visualization
    - Diversity of possible visualizations
    - Different types of visualization (objective-oriented)
    - Note about the capstone project / your future job
    - Different types of data sets and data visuals
    - Different possible visualization graphics
  - Examples of data visualization charts
    - Numeric variables
    - Categorical variables
    - Mix of numeric variables + categorical variables
    - Special case of time-dependent variables
    - Special case of geographical data
    - Special case of time-space visualization
    - Tree and graph / network visualization
    - Other special cases
  - Technical tools for data visualization
  - Summary
  - Homework

Introduction

Definition of data visualization

Recall: DS methodology



Definition of data visualization

- Data visualization is a technique to communicate insights from data through visual representation
- Main goal: « explore » large datasets into visual graphics
  - For easier understanding of each variable
  - For easier understanding of complex relationships between the variables
- Data visualization is very linked to EDA
  - Some basic (univariate and bi-variate) graphics were seen in Chapter 4
- In this lecture, we are going to talk about:
  - more refined visualization graphics
  - how to choose the right chart for your data, what you want to show, and your public

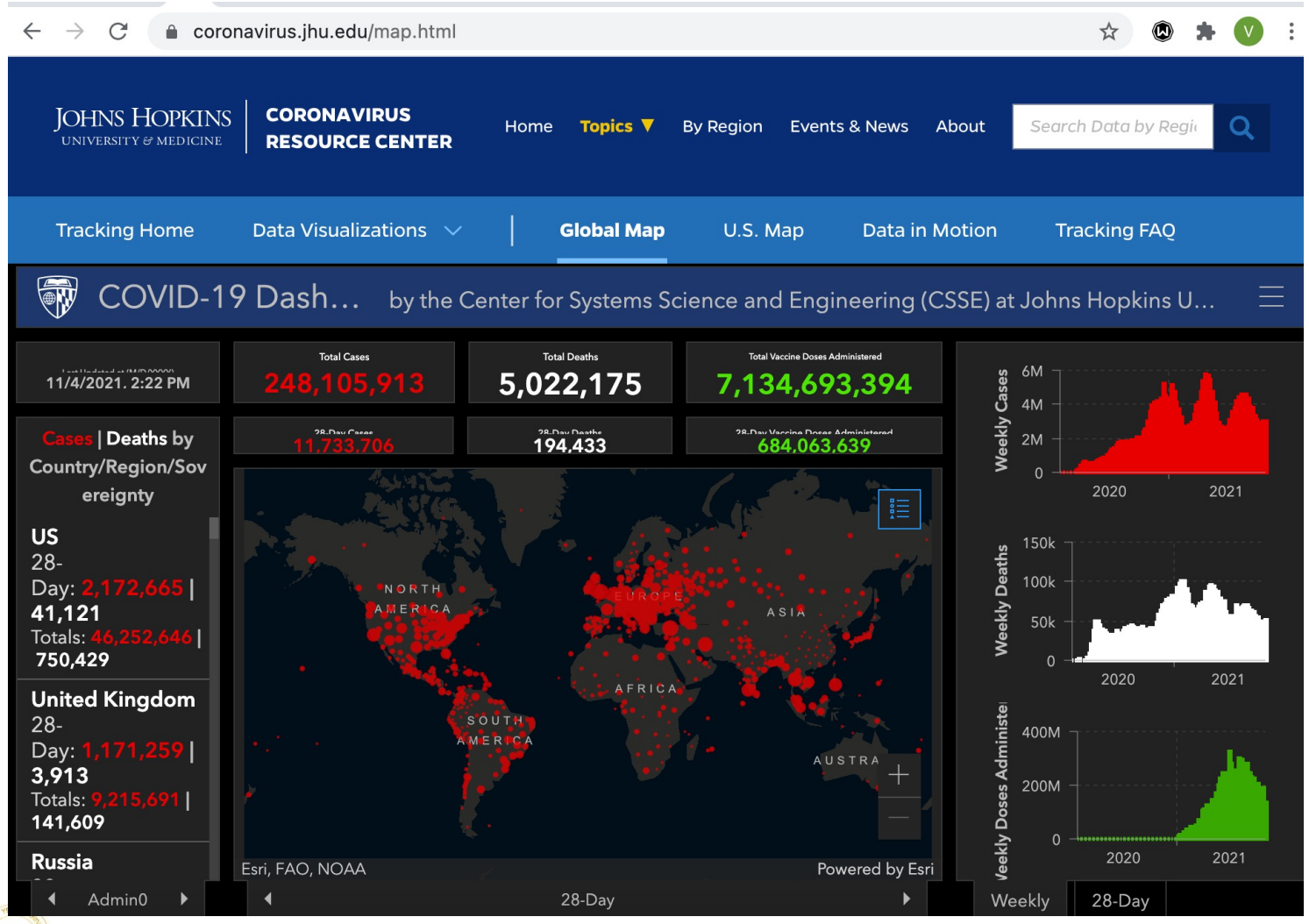
How to choose a good data visualization

- There are many possible visualizations for the same data
- Some data is so complex that it is impossible to have it all summarized with only 1 visualization
  - Some visualizations might even be misleading
    - Interesting article about this subject: http://www.ericson.net/content/2011/10/when-maps-shouldnt-be-maps/
- Choosing THE best visualization might not be easy
  - This choice has to be made according to the author…
  - ... to the type of data (of course!)…
  - … but also according to the target public's background
  - We're going to focus on that later in this lecture

Why visualizing data?

- 3 main reasons
  1. Summarize information
  2. Facilitate reasoning / analysis about information
  3. Present information
     - Highlight some aspects of the data
     - Share and illustrate your opinion
       - Explain it / convince people (storytelling)
     - Collaborate with peers

Why visualizing data? *1- to summarize info*



© Mike Kelley – Photoviz http://shop.gestalten.com/photoviz.html

Why visualizing data? *2- to facilitate reasoning*

John Snow, 1854

Why visualizing data? *3- to present information*

- In a possible interactive way

How's life?

http://www.oecdbetterlifeindex.org/#/11111111111

Introduction

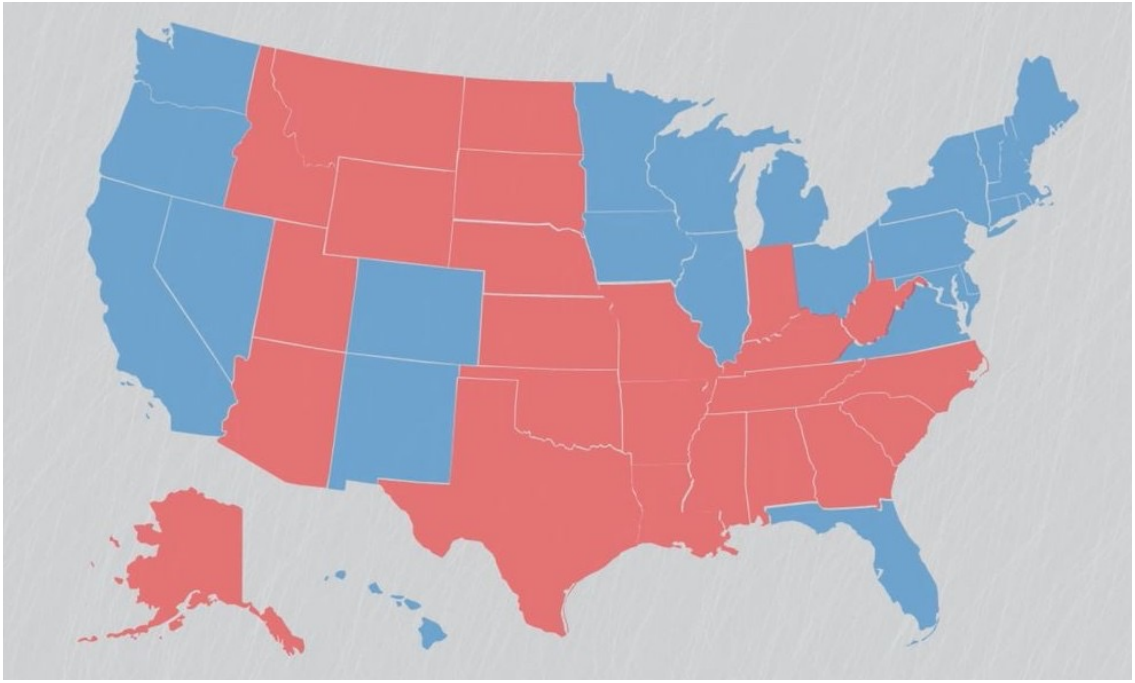
Diversity of possible visualizations

Diversity of possible visualizations

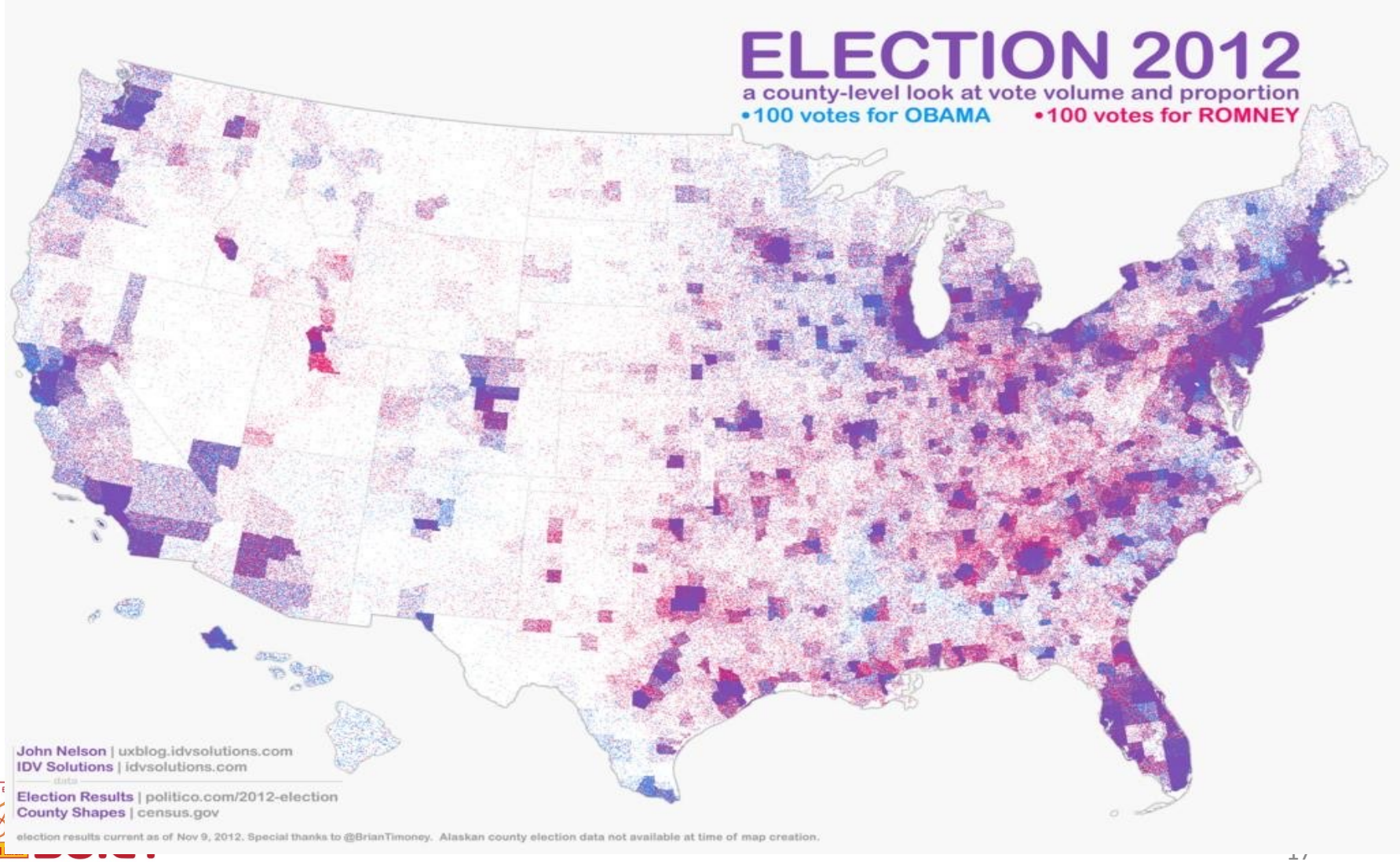
- Example: which visualization do you prefer for the numbers 75 and 37?

Let's criticize: 2012 US election map

- Who is this visualization for?
- What question does it answer?
- Why do you like this visualization?
- Why don't you like it?
- How could you enhance it?

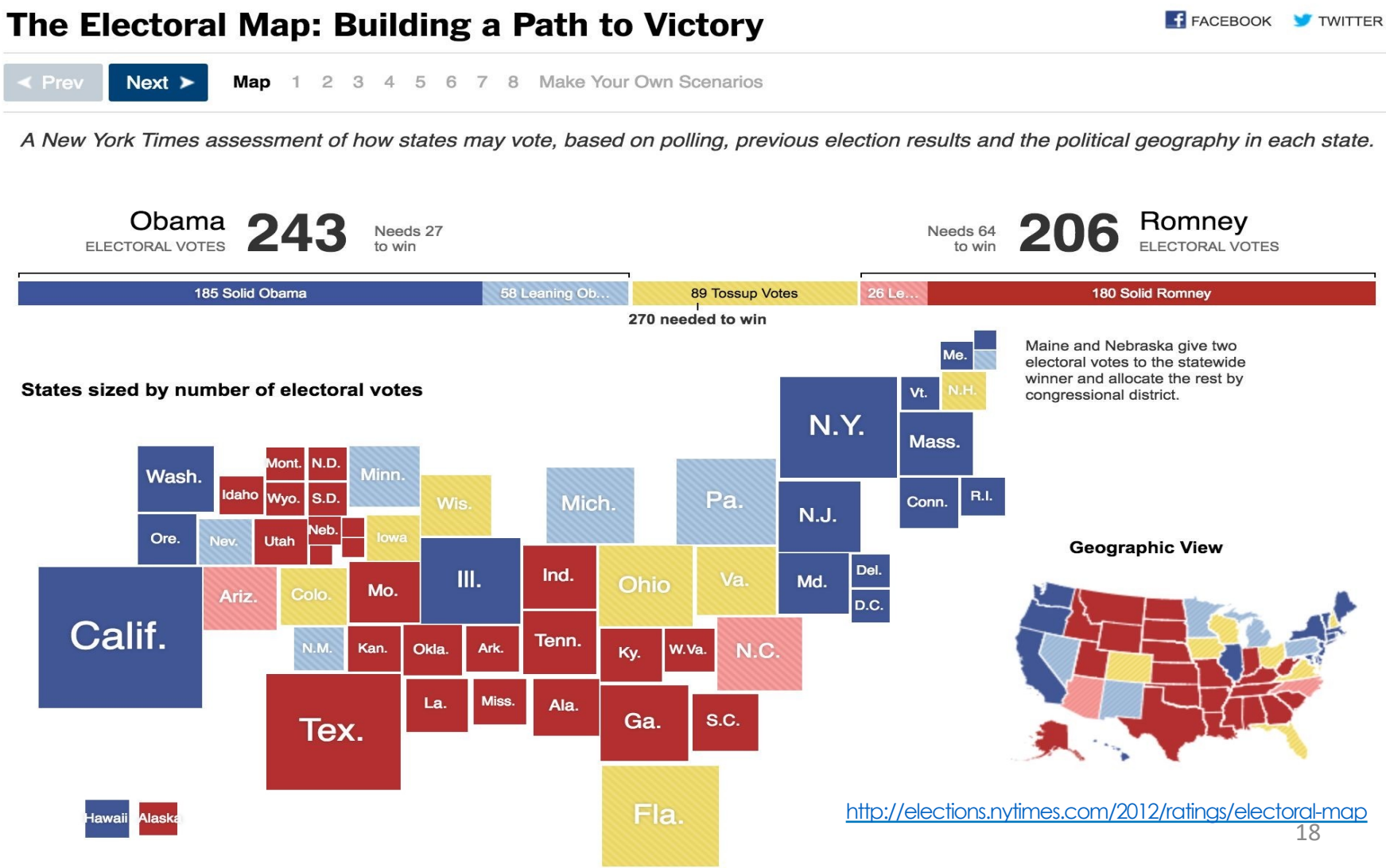


Alternative US election maps



https://www.flickr.com/photos/idvsolutions/8182119174/sizes/k/in/photostream/

Alternative US election maps

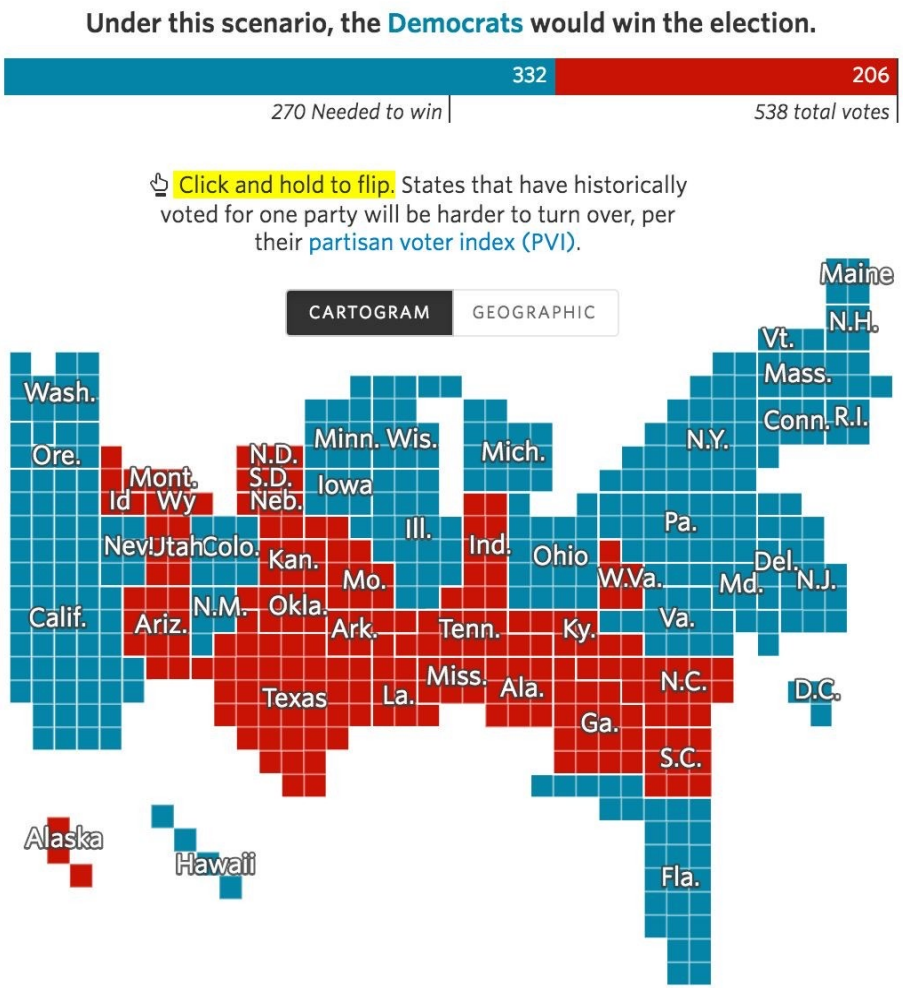


http://elections.nytimes.com/2012/ratings/electoral-map

Alternative US election maps: predictions

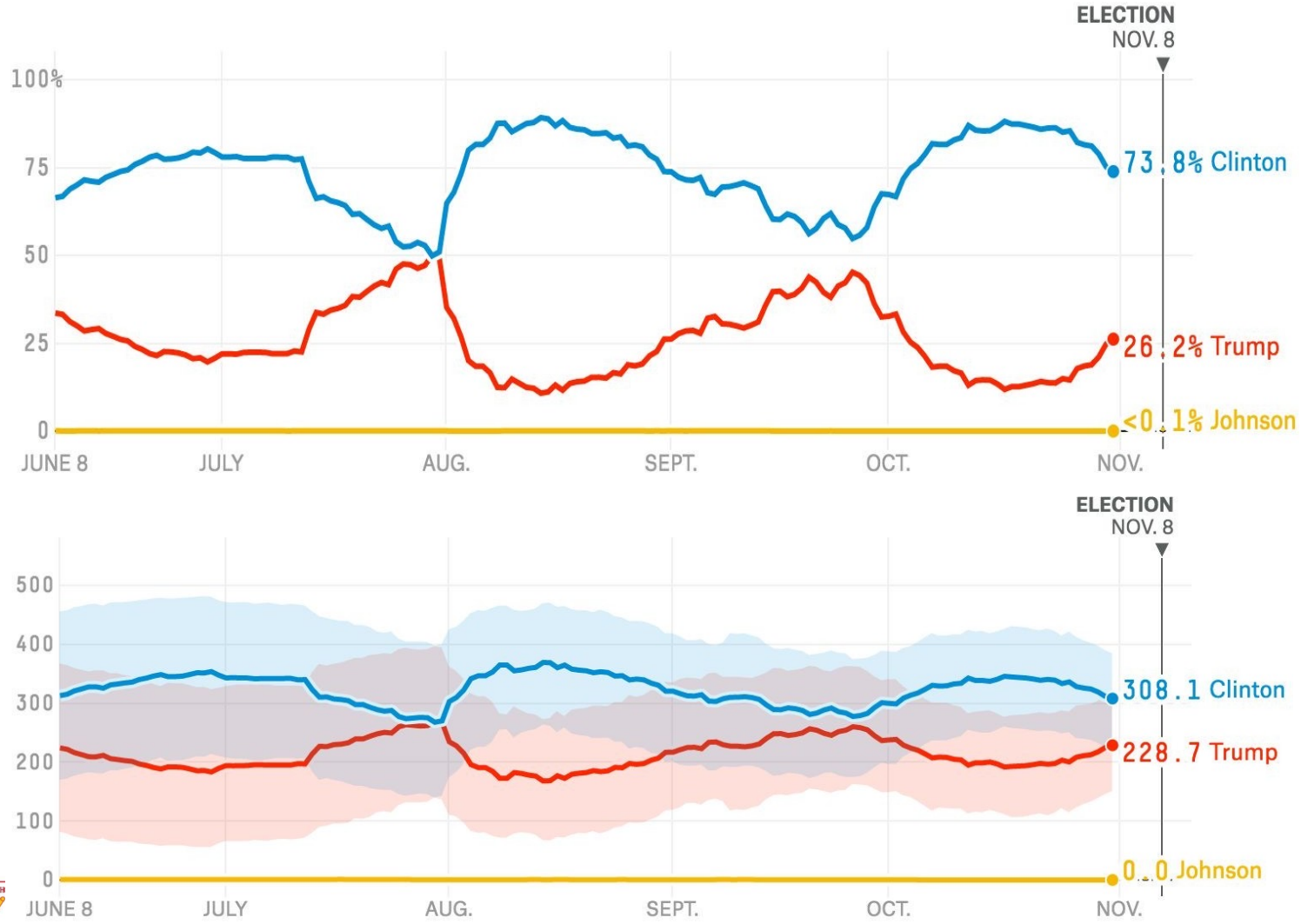
PREDICT THE OUTCOME

A presidential election is a set of 51 contests, in each state plus the District of Columbia, to determine which candidate can build a majority in the Electoral College. Use this map to draw your own path to victory. Click on a state to forecast which political party will carry its electoral votes—it takes 270 votes to win. We've shown how each state voted in the 2012 election. We've also made it easy to flip battleground states and harder to change states that reliably support the same party—click and hold in order to flip those states. You can opt for a traditional map or a cartogram, which shows each state's true weight in the electoral vote. Below are different ways to look at this year's electoral landscape, which may guide your own projections.



http://graphics.wsj.com/elections/2016/2016-electoral-college-map-predictions/

Alternative US election maps: predictions



com/2016-election-

Conclusion

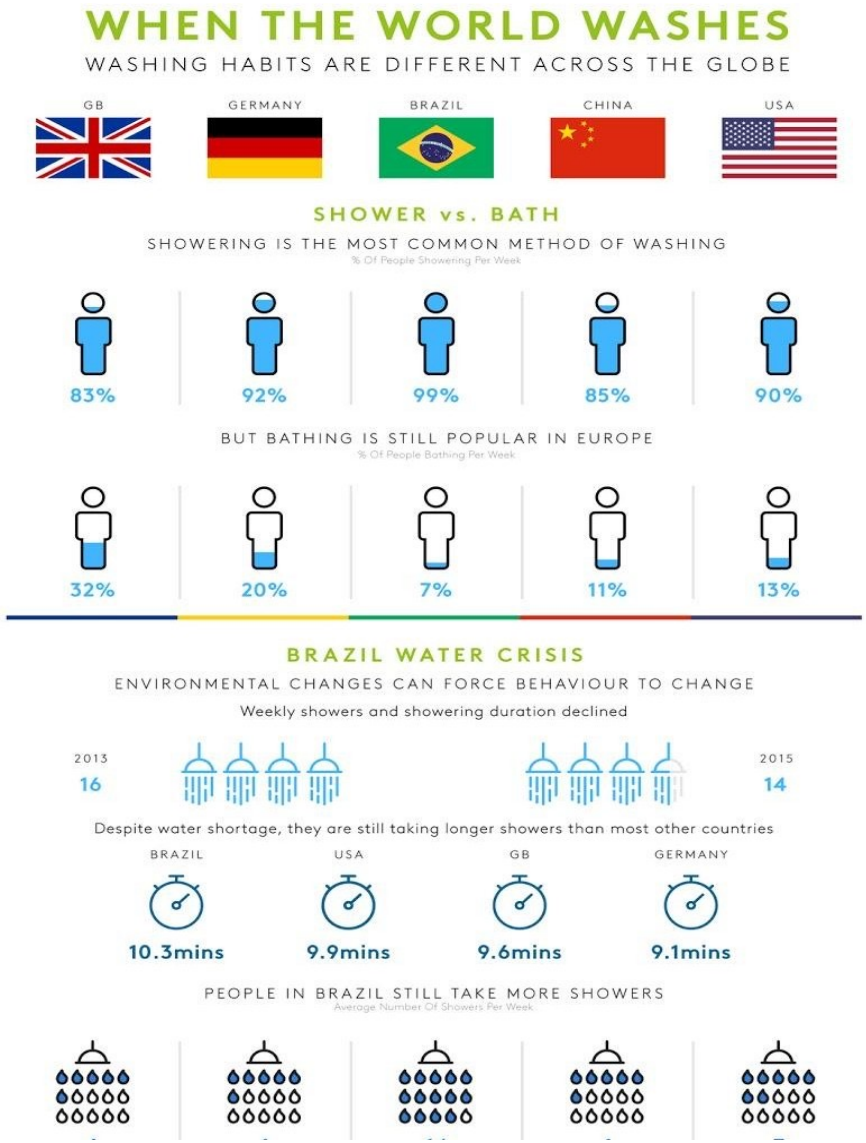
- There are many possible visualizations for the same data
- Some data is so complex that it is impossible to have it all summarized with only 1 visualization
  - Some visualizations might even be misleading
    - Interesting article about this subject: http://www.ericson.net/content/2011/10/when-maps-shouldnt-be-maps/
- Choosing THE best visualization might not be easy
  - This choice has to be made according to the author…
  - ... to the type of data (of course!)…
  - … but also according to the target public's background

Introduction

Different types of visualization (objective-oriented)

Infographics

- Infographics: graphic visual representations of information, data, or knowledge intended to present information quickly and clearly



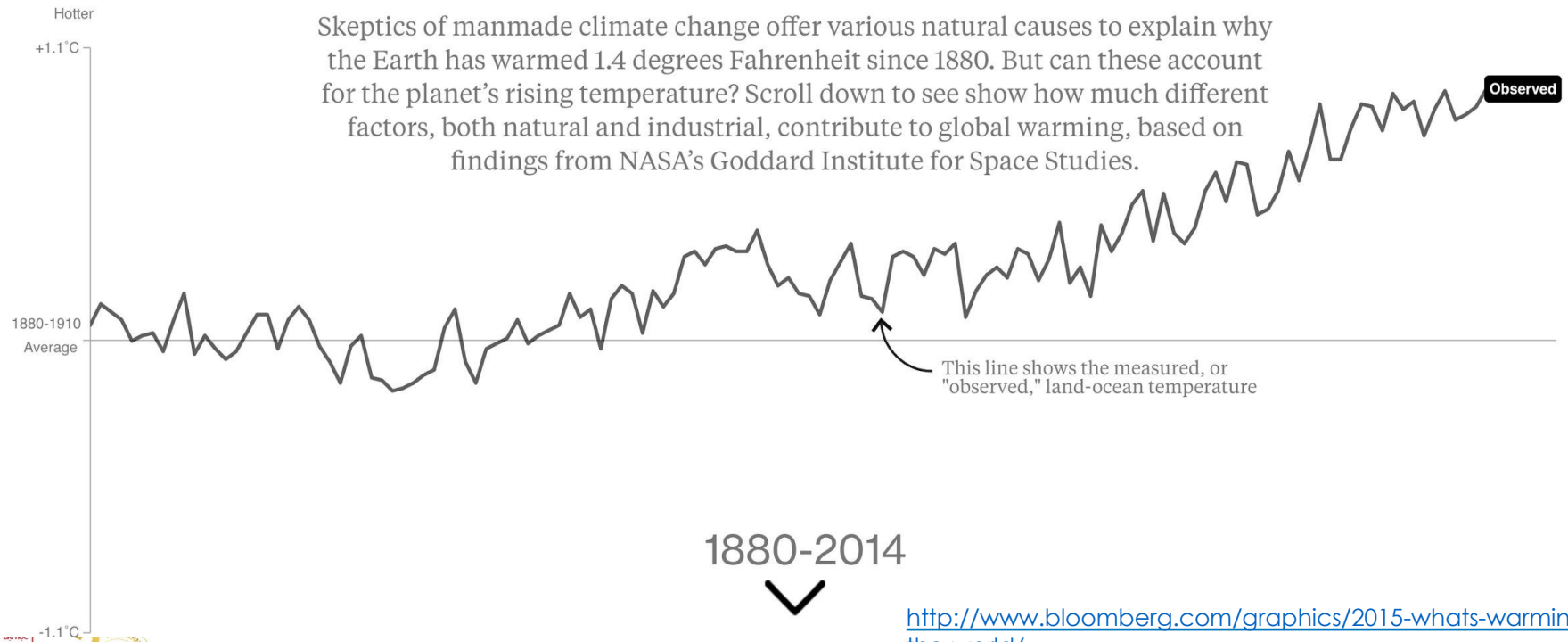
Storytelling

Bloomberg

What's Really Warming the World?

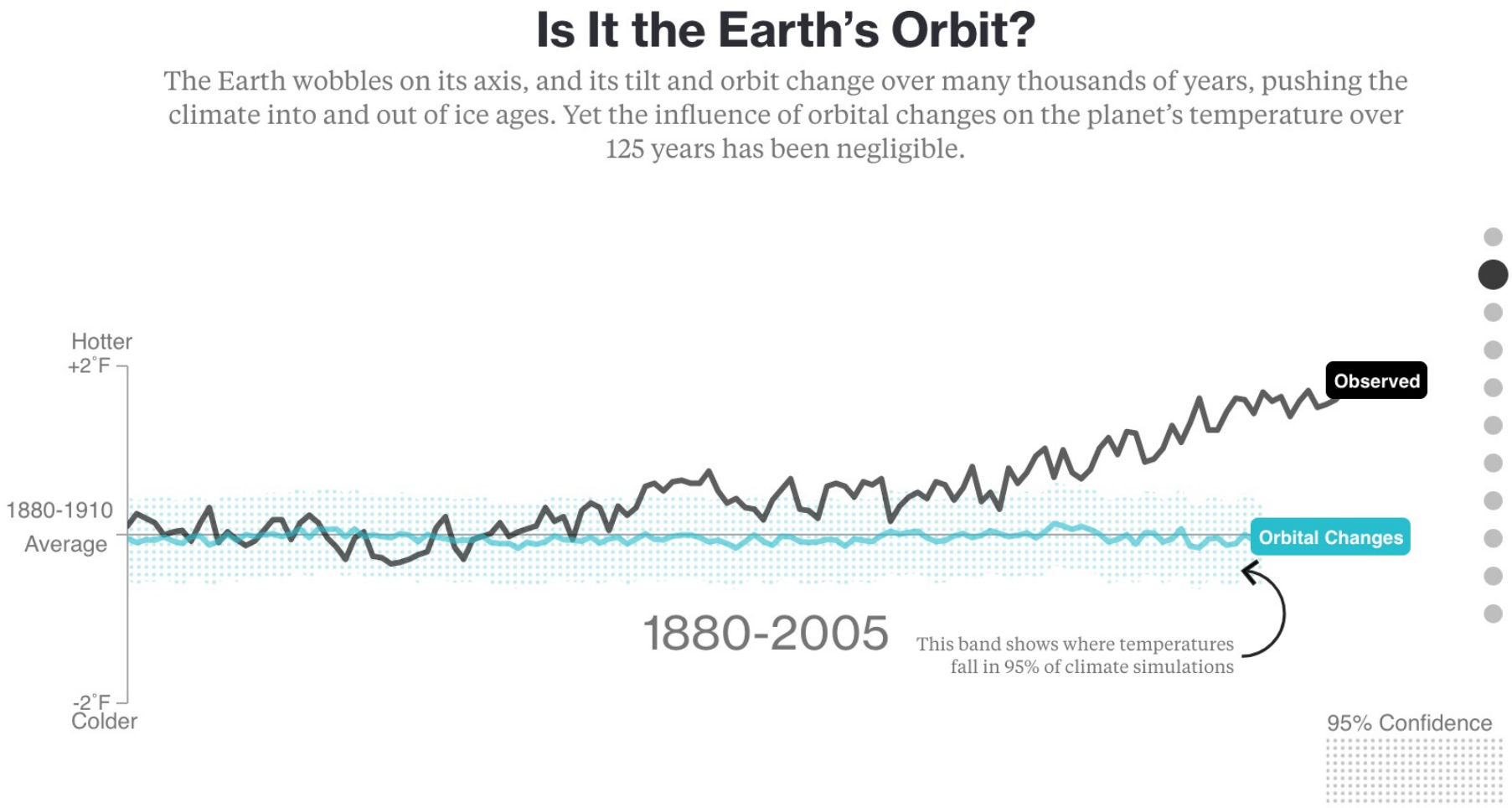
By Eric Roston and Blacki Migliozzi | June 24, 2015

Skeptics of manmade climate change offer various natural causes to explain why the Earth has warmed 1.4 degrees Fahrenheit since 1880. But can these account for the planet's rising temperature? Scroll down to see show how much different factors, both natural and industrial, contribute to global warming, based on findings from NASA's Goddard Institute for Space Studies.



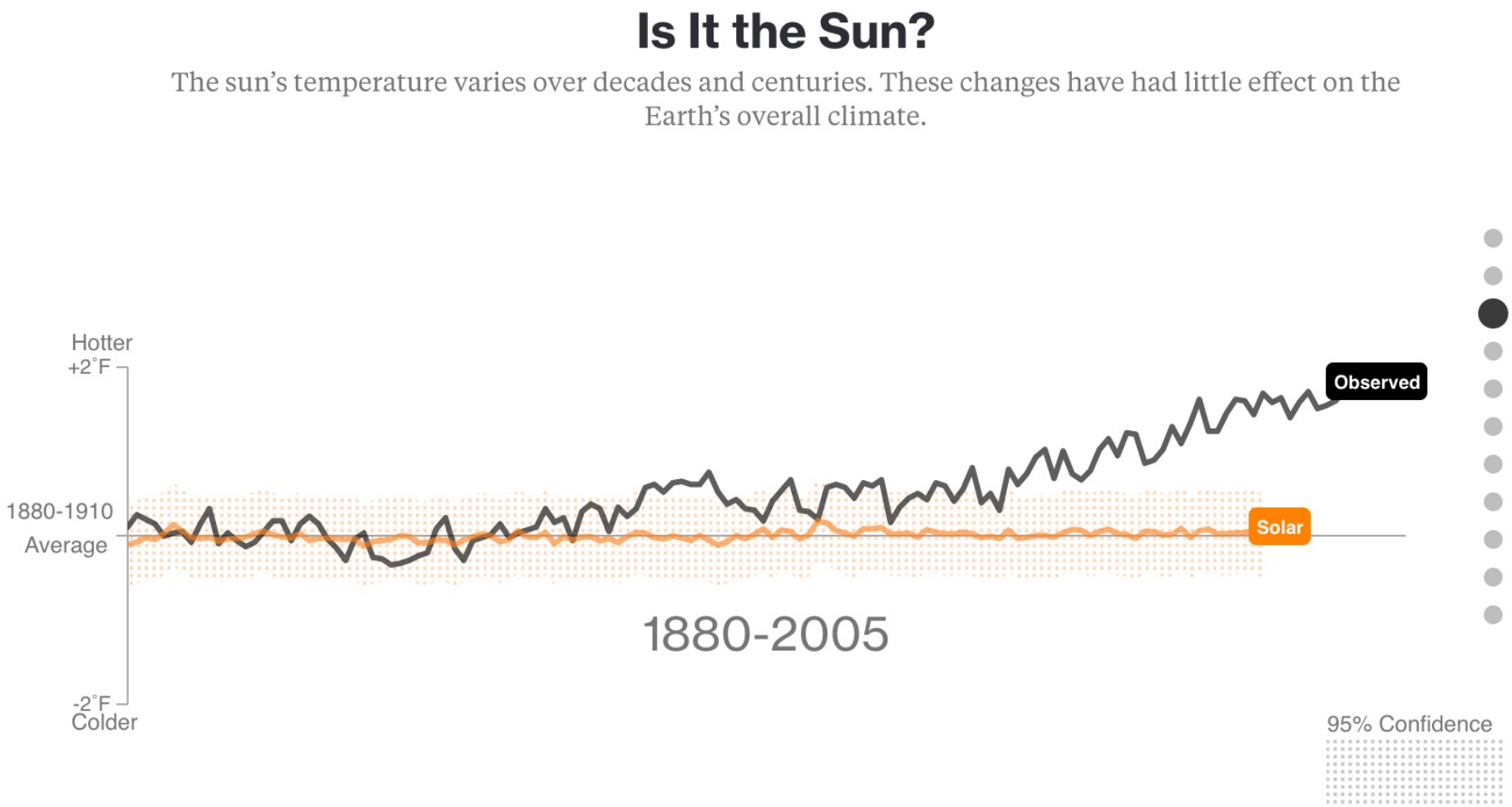
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



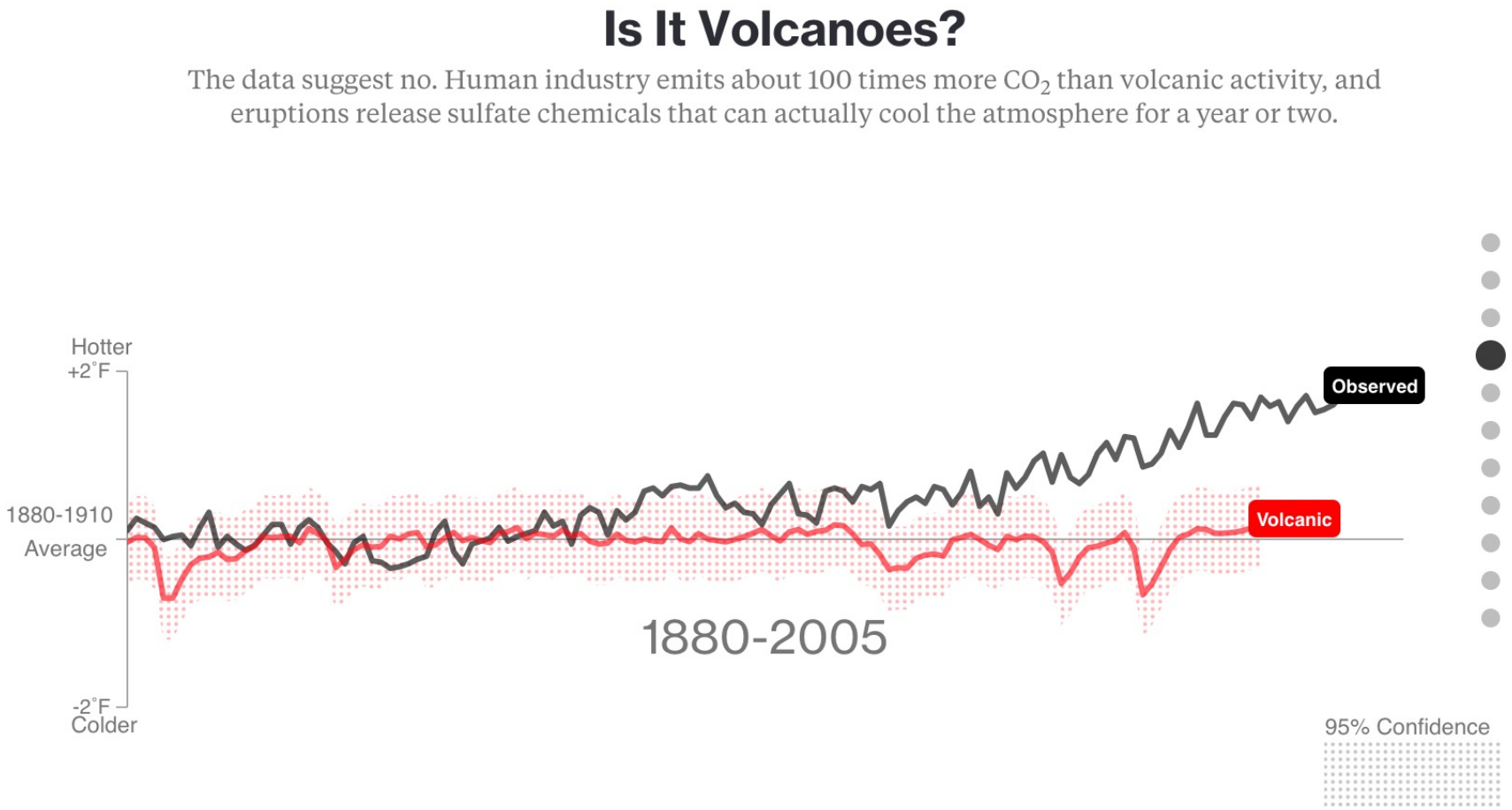
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



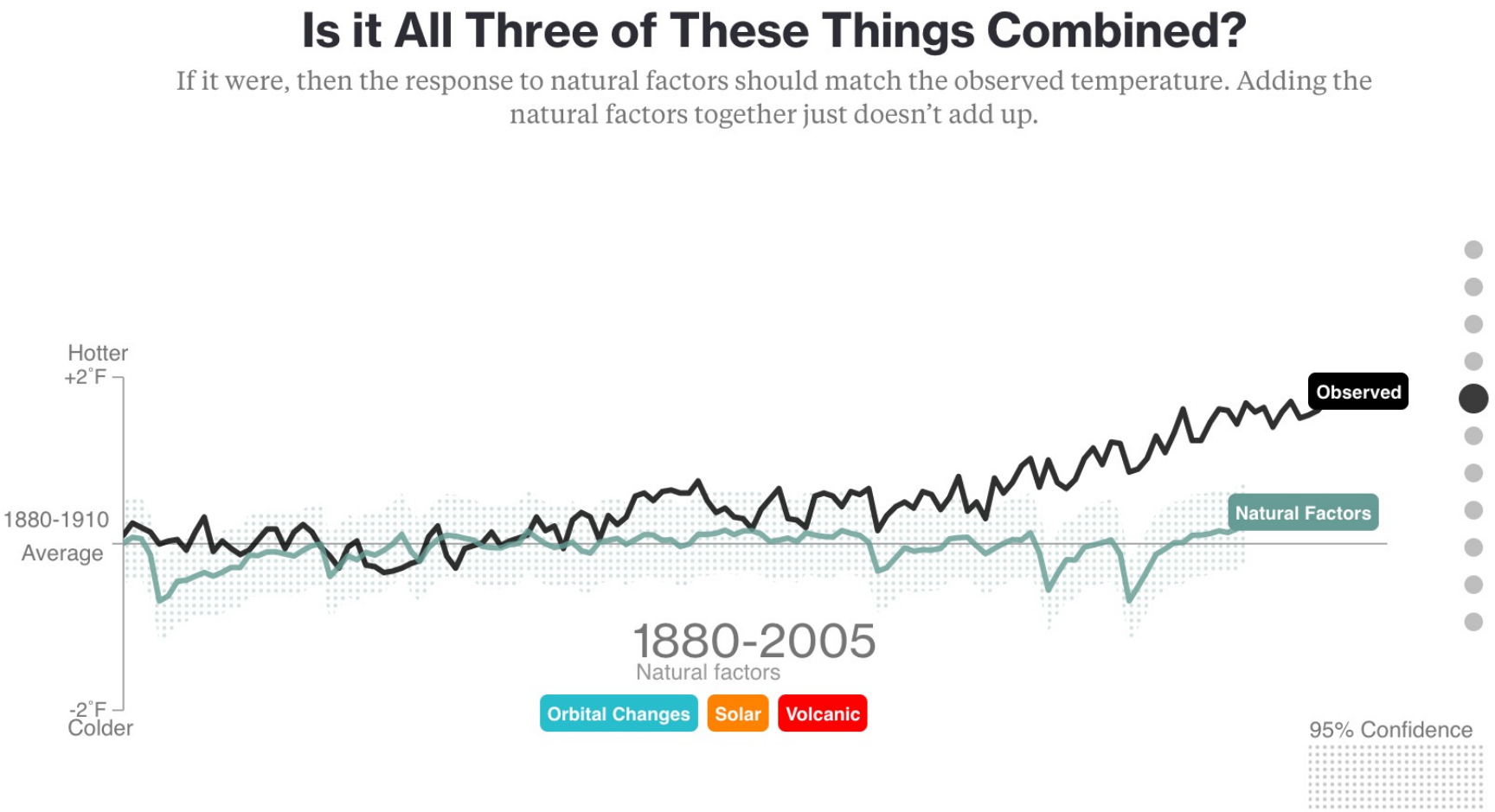
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



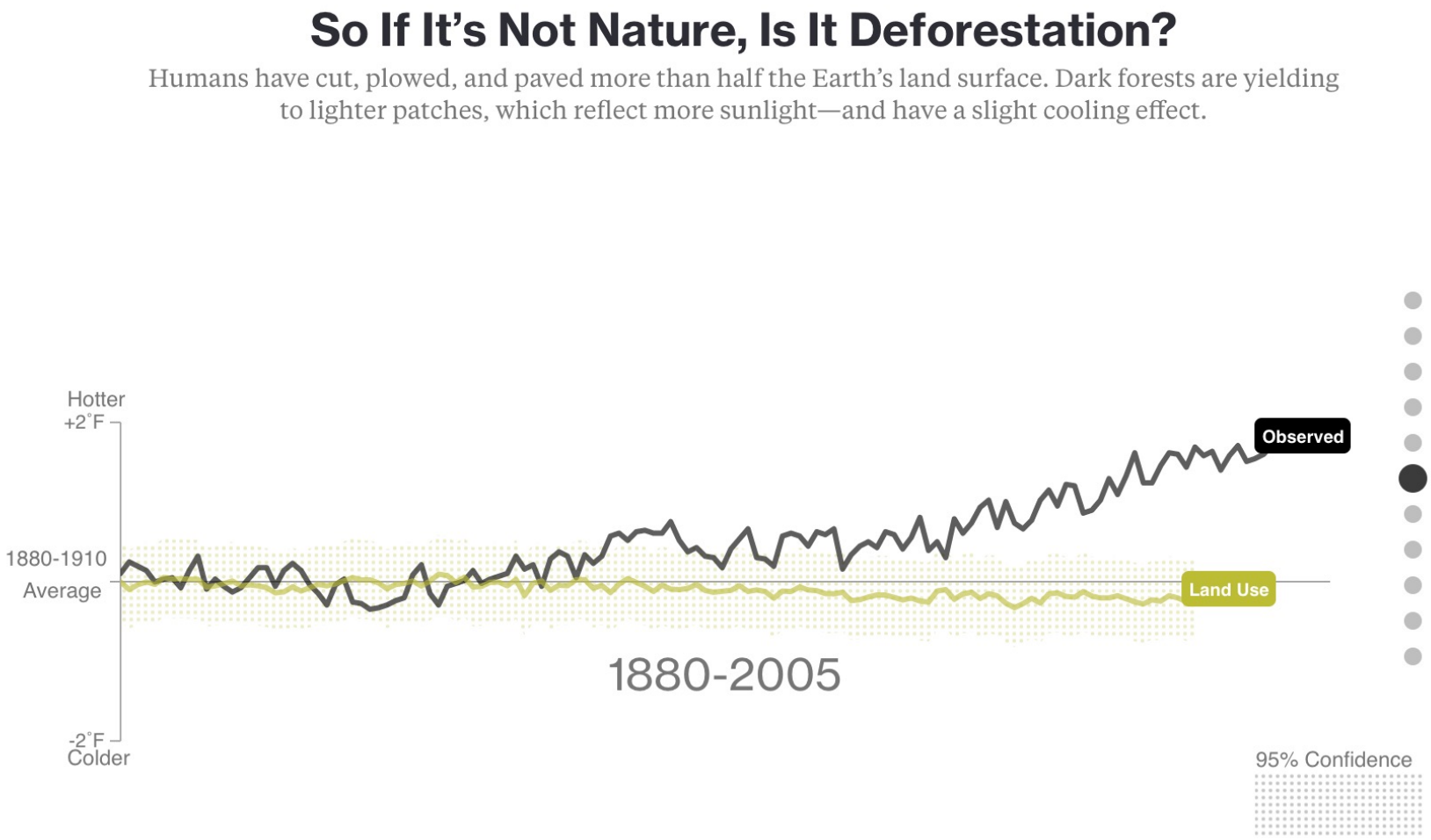
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



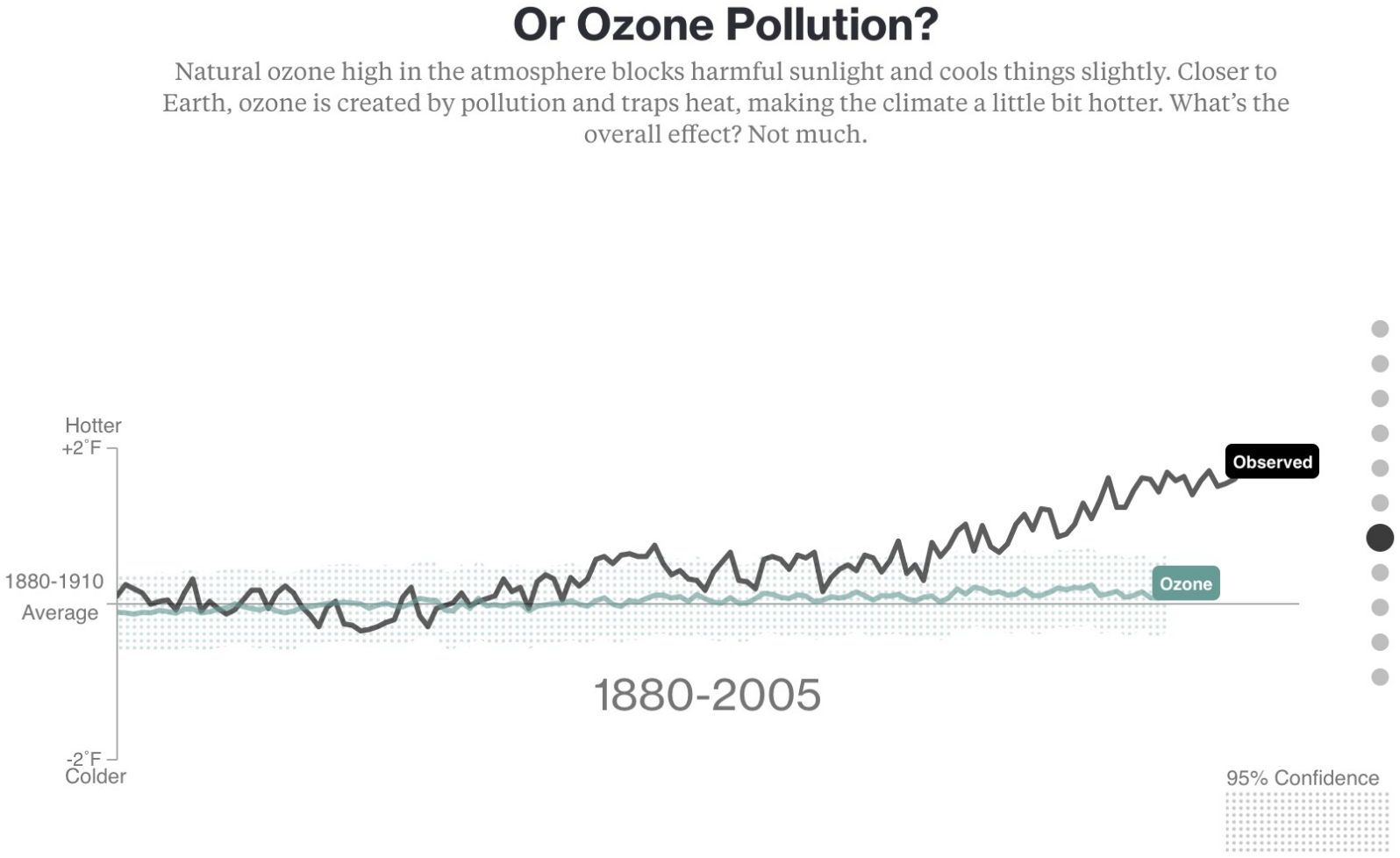
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



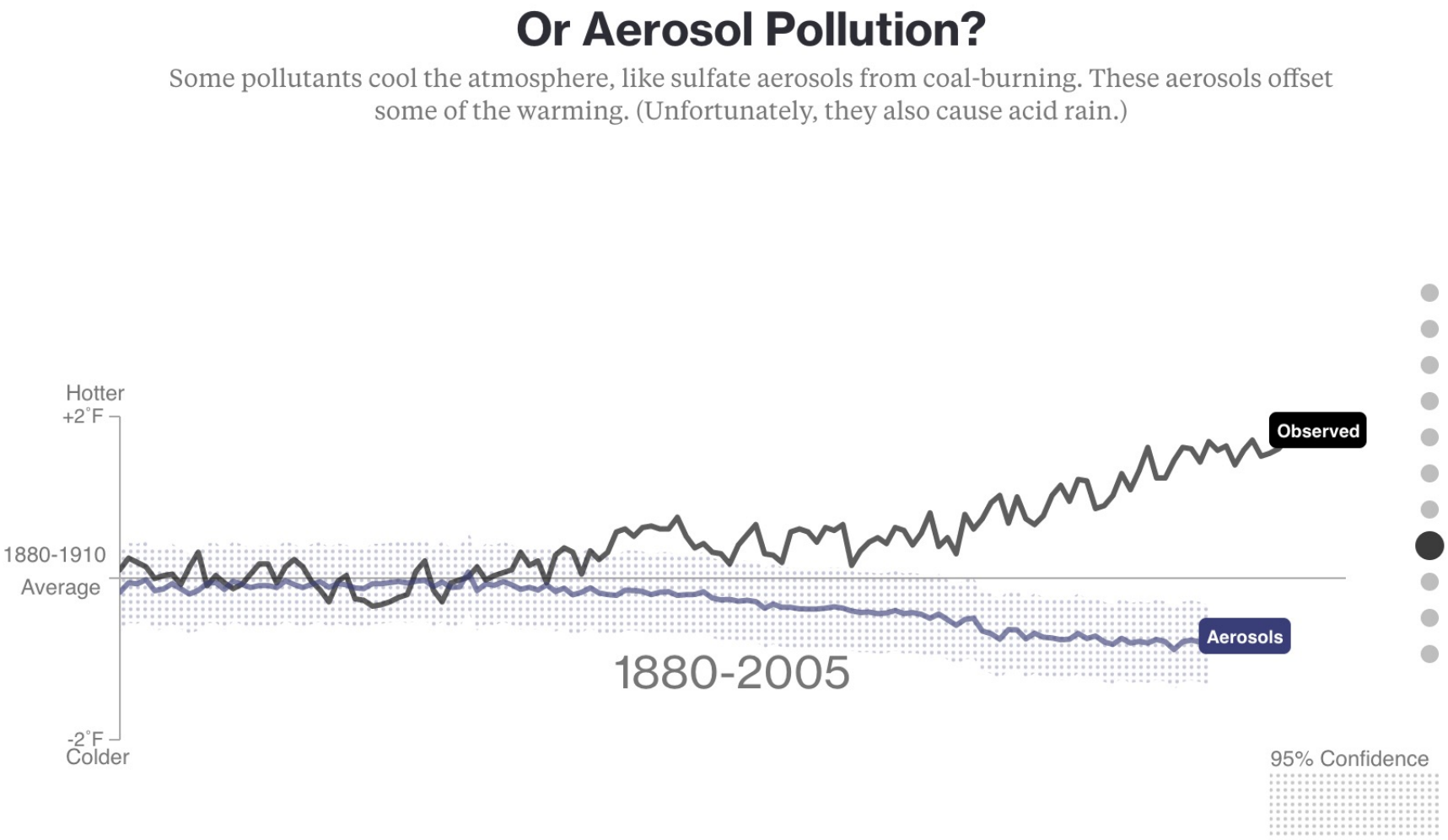
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



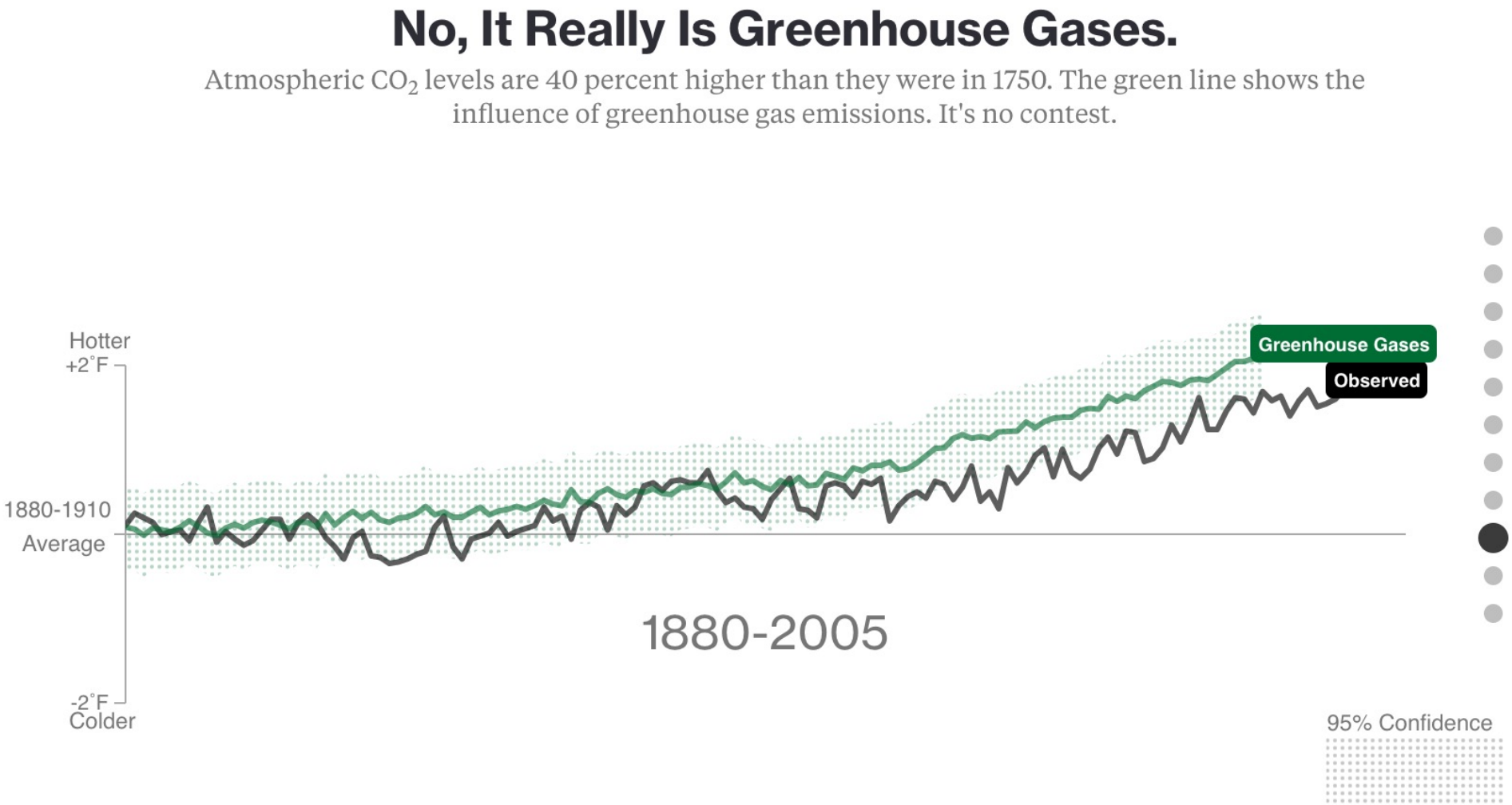
http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling



http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling

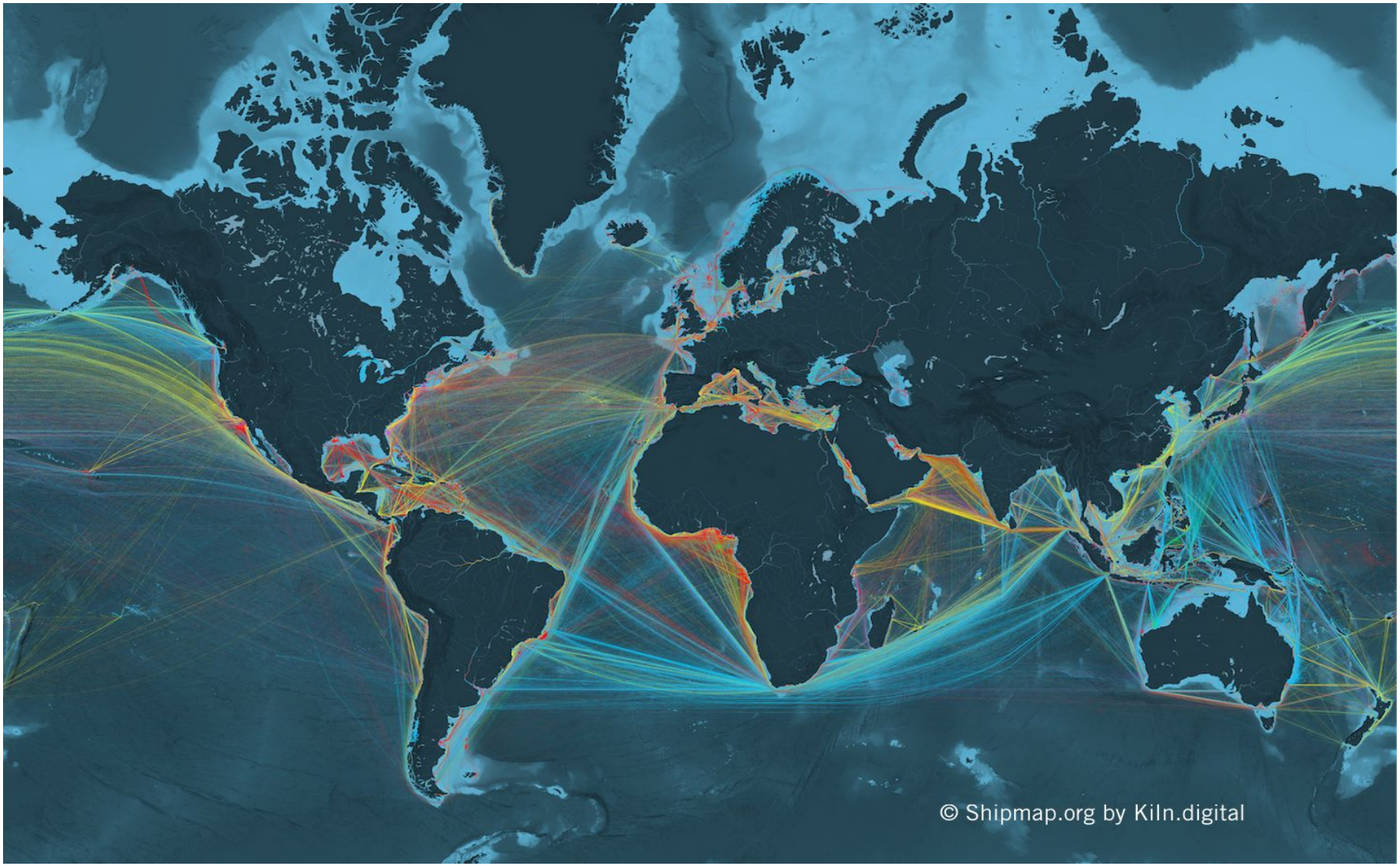


http://www.bloomberg.com/graphics/2015-whats-warming-the-world/

Storytelling – let's think for a minute

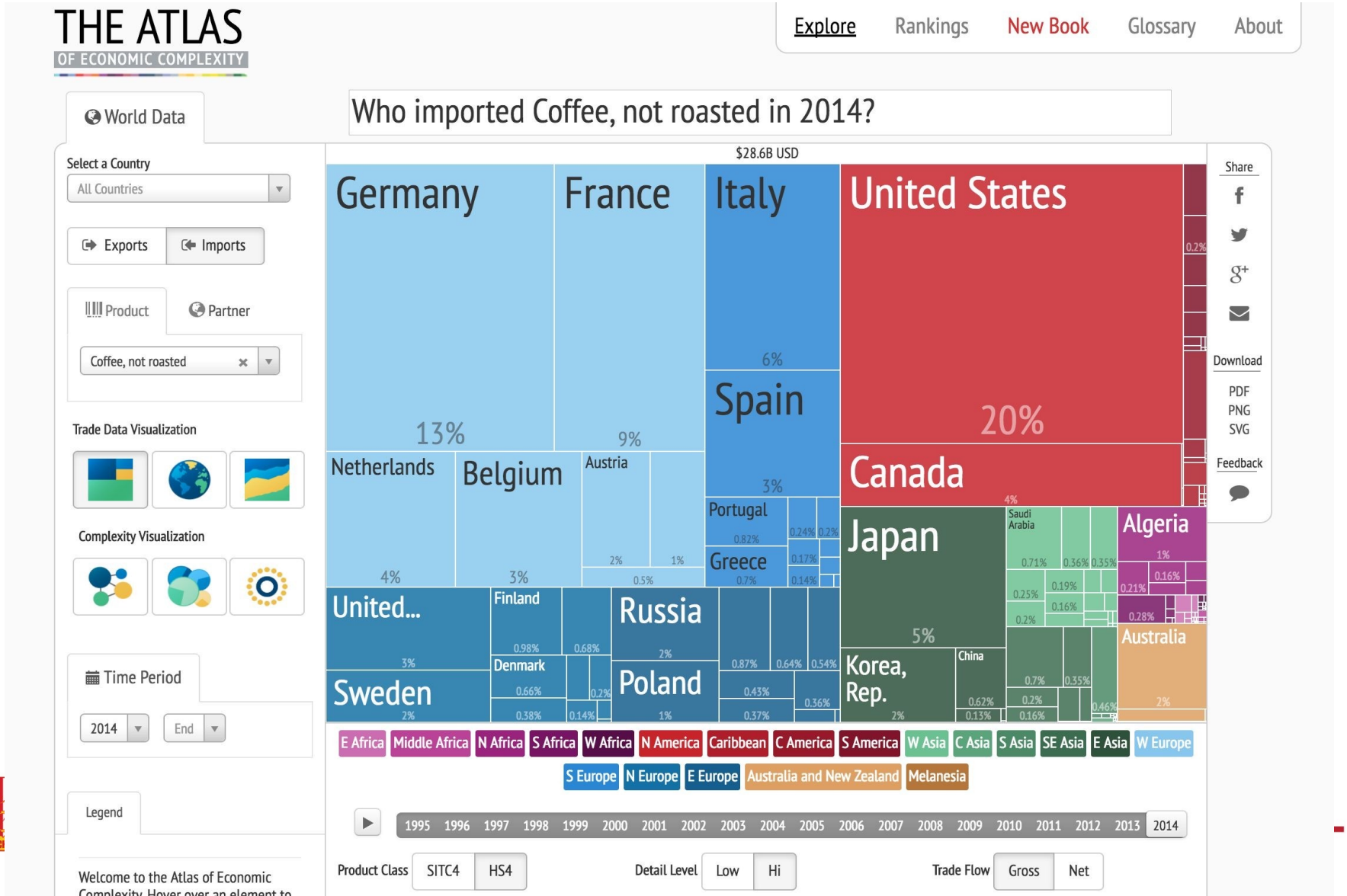
- Do you think that the above storytelling is solid, statistically speaking?
- Do you think it is convincing, for a non-statistician audience?
- Note: many other more solid research clearly shows the link between CO2 emissions and global warming, but they might not be as convincing / understandable for a general audience

Cartography



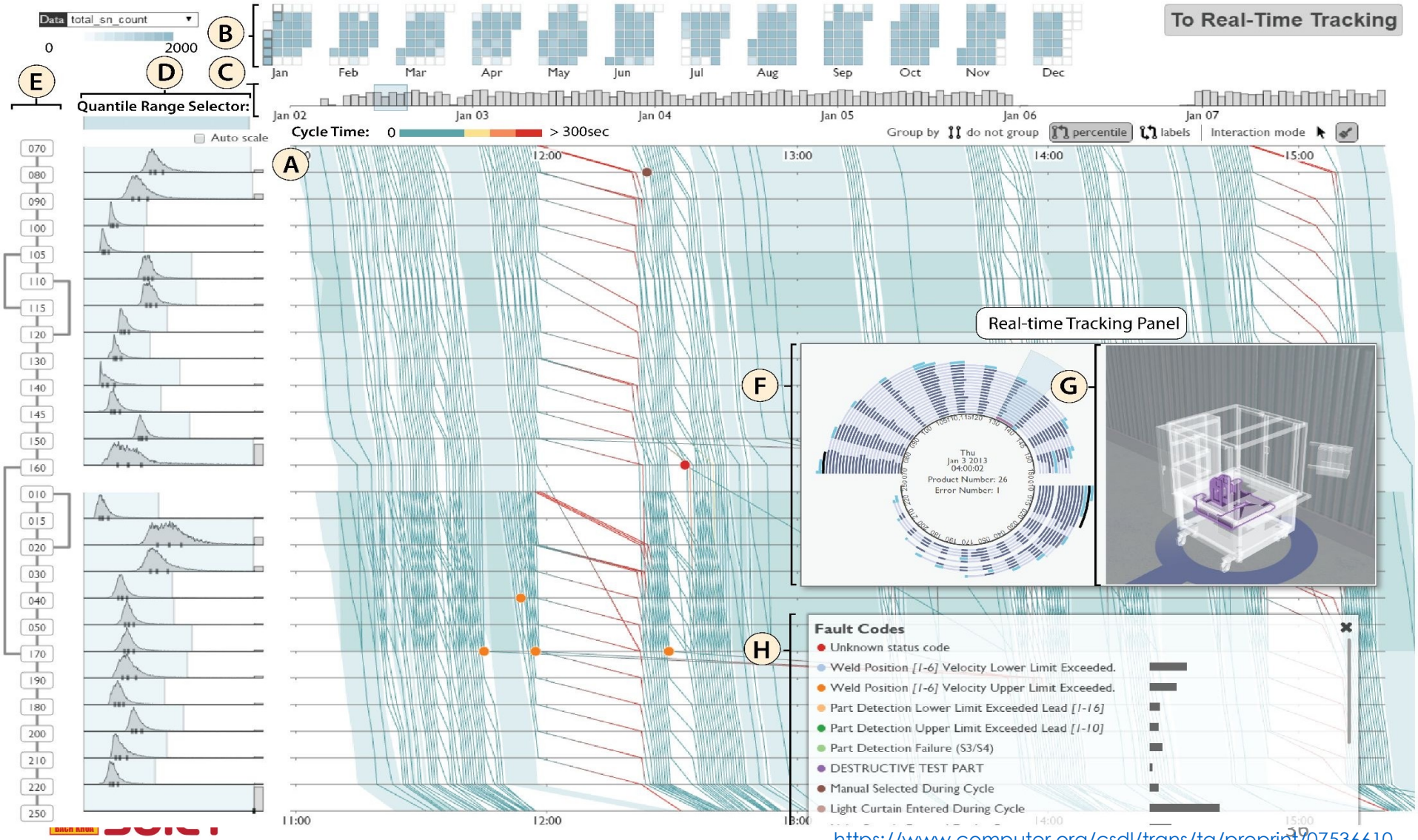
Information visualization

- As synthetic as infographics: but with more info



Visual analytics

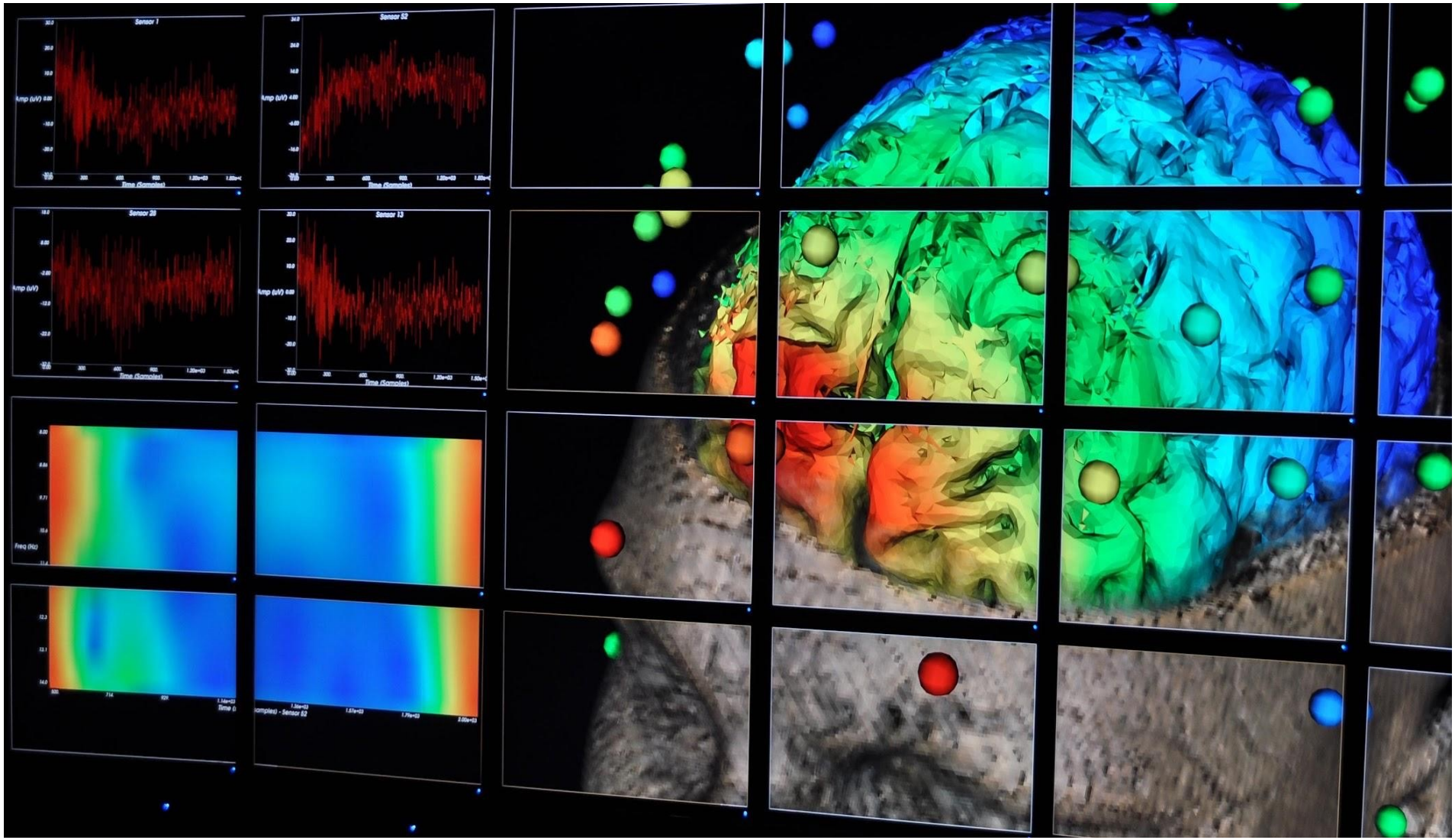
- Audience: mostly for statisticians / computer scientists



https://www.computer.org/csdl/trans/tg/preprint/07536610-abs.html

Scientific visualization

- Audience: mostly for specific experts (medical doctors, traders)



VisTrails
https://www.nsf.gov/discoveries/disc_images.jsp?cntn_id=114322&org=NSF

Introduction

Note about the capstone project / your future job

Visualization is important for data scientists!

- As a future data scientist, you need to be able to
  - Understand the data for yourself
  - Transfer your knowledge of the data to others
    - Visual representations can greatly help you and your audience!
    - 1 graphic = a lot of words / numbers…
- So, this chapter of data visualization is **very important**
  - Often, informaticians do not care too much about visualization
  - But, as a data scientist, you **must** care!
    - Communication is a big part of the data scientist's job!

Note about the project

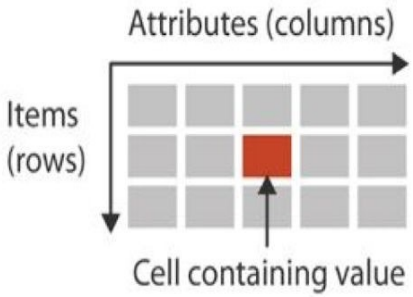
- Especially for the students who chose the topic on EDA, data visualization is a **big part** of the capstone project's report and presentation
- The difficulty is that, choosing the appropriate chart(s) is not always easy…
  - Depends on the type(s) of data
  - Depends on what you want to show
- In this lecture, I will not give an exhaustive list of all possible charts, nor the Python functions (not enough time)
  - You'll have to learn how to program them on your own
- But, I will give you:
  - The methodology to choose the type of chart to use
  - Some resources which list many many types of charts and the Python libraries, with examples of code
  - Many ideas of original charts to use for your future communications

Introduction

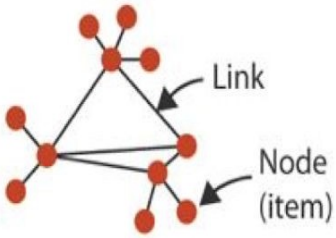
Different types of data sets and data visuals

Different types of data sets

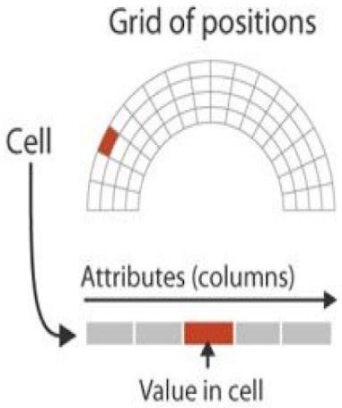
➔ Tables
(numeric/categorical variables)



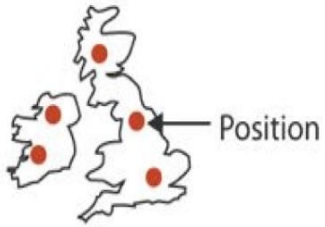
➔ Networks
(graphs)



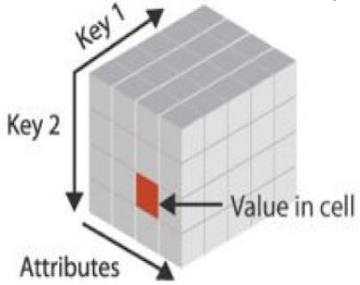
➔ Fields (Continuous)



➔ Geometry (Spatial)

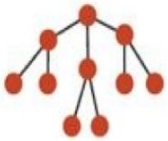


➔ *Multidimensional Table*
(data cubes in OLAP)



(numeric/categorical variables)

➔ *Trees*

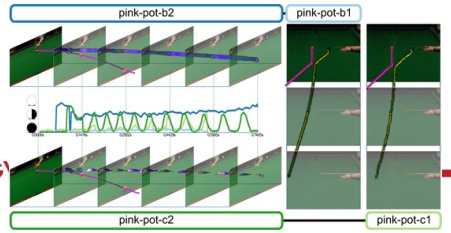


➔ Text

Blablablabla
Blobloblo…

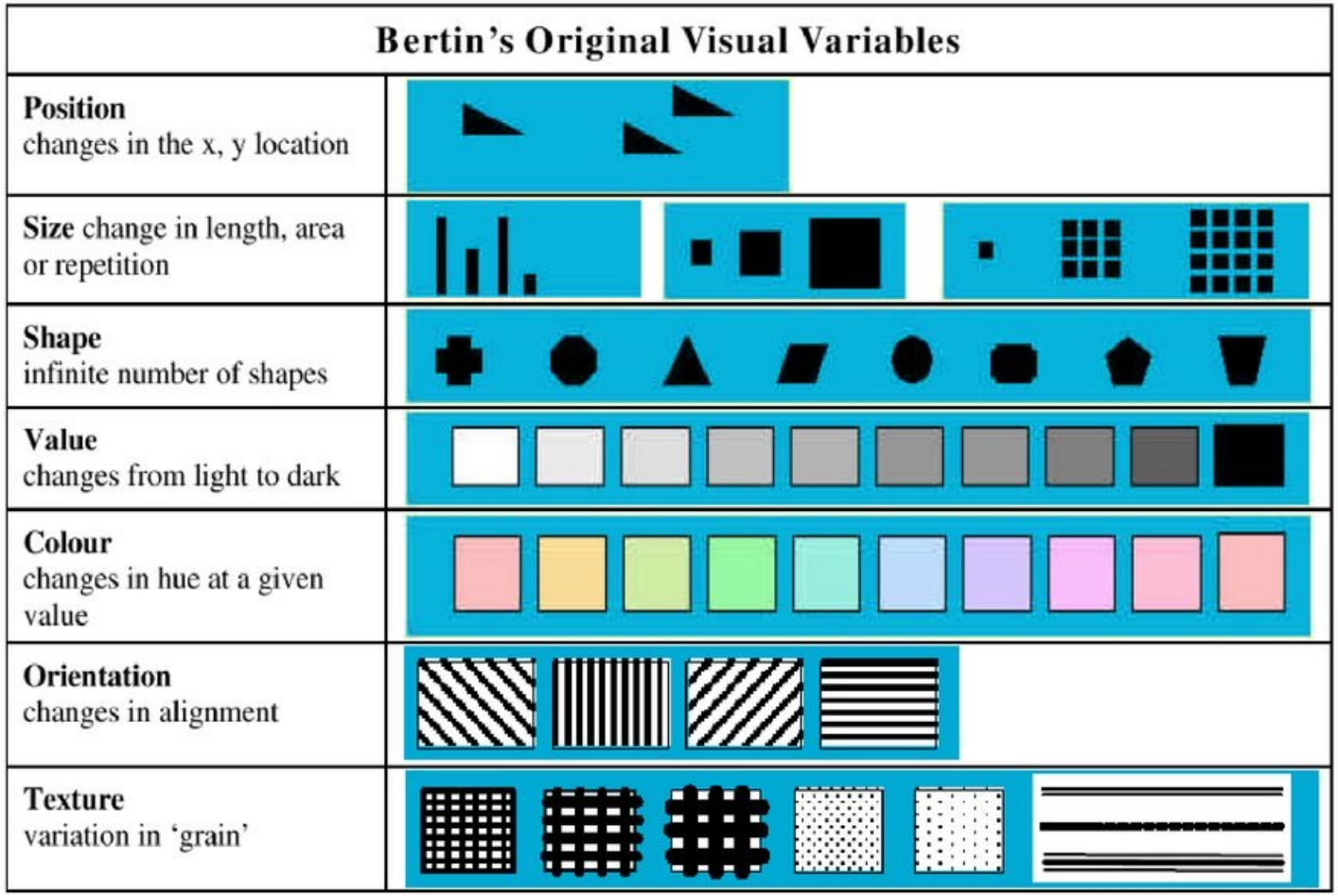
➔ Images

➔ Sequences of images / videos



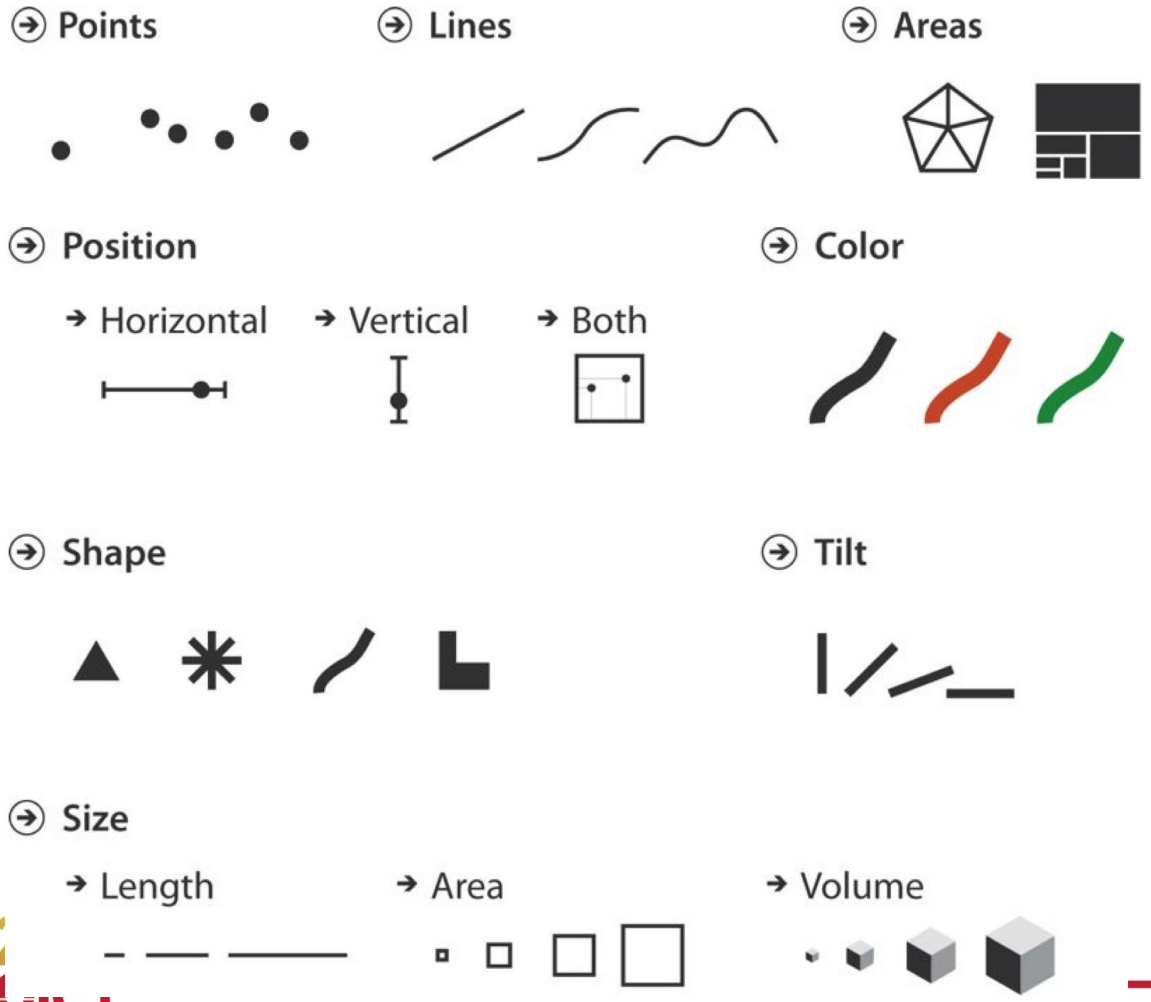
Different types of visuals to be used

- In cartography
  - Jacques Bertin's variables



Different types of visuals to be used

- More generally…

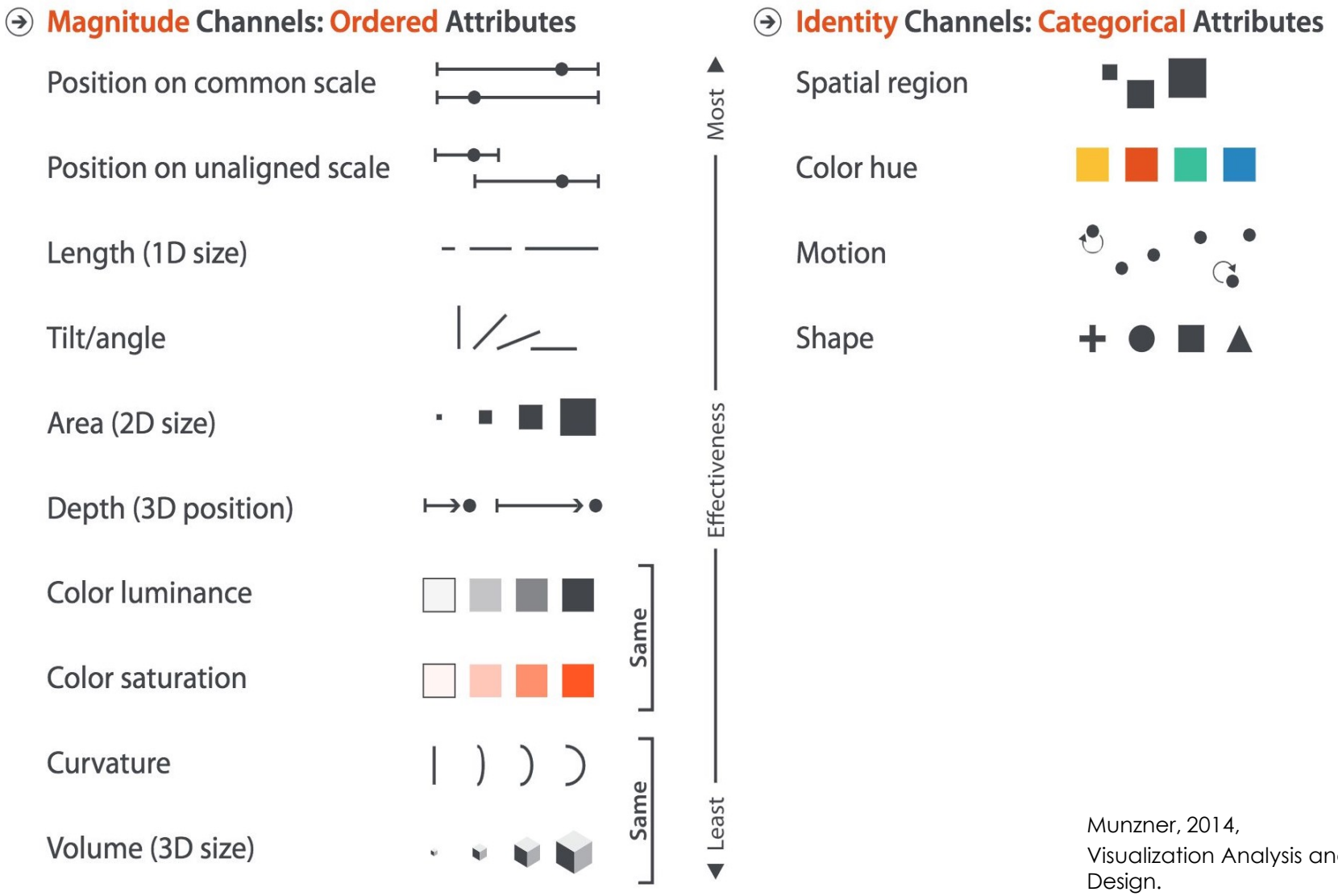


Munzner, 2014, Visualization Analysis and Design.

Different types of visuals to be used

For numeric and categorical, ordinal variables

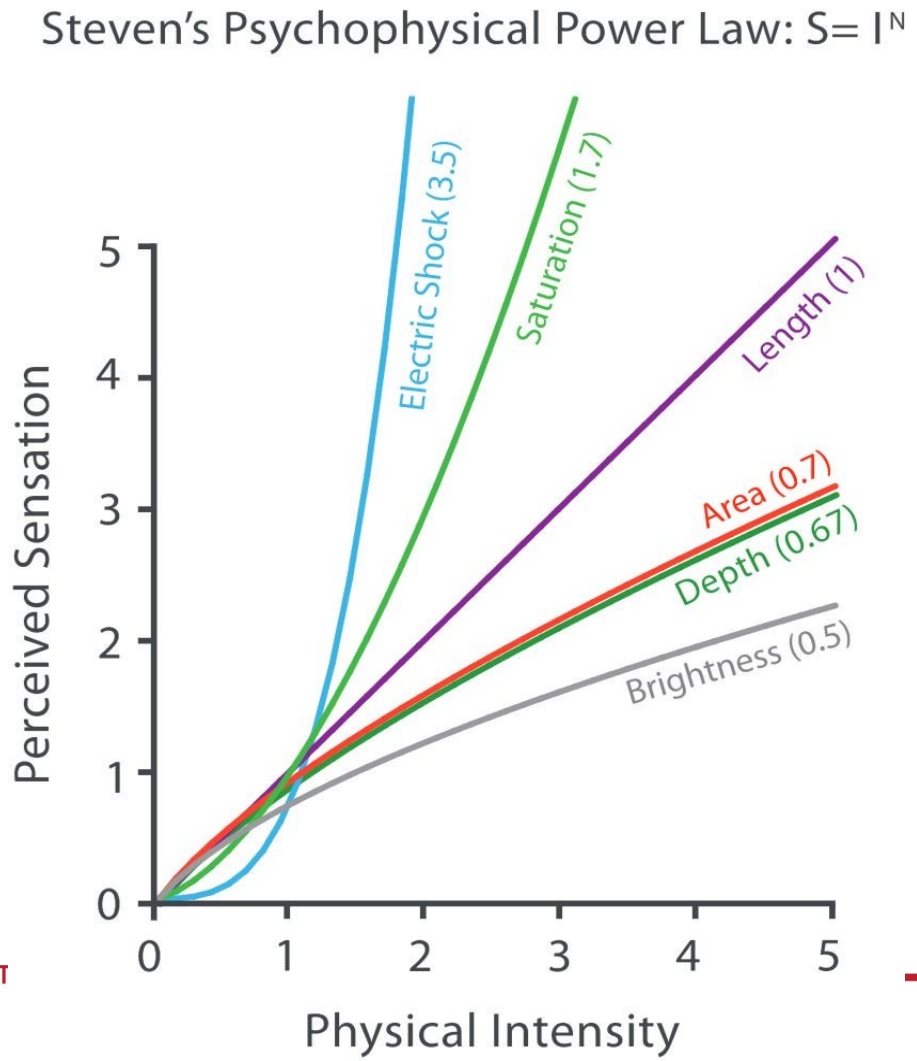
For categorical, nominal variables



Munzner, 2014, Visualization Analysis and Design.

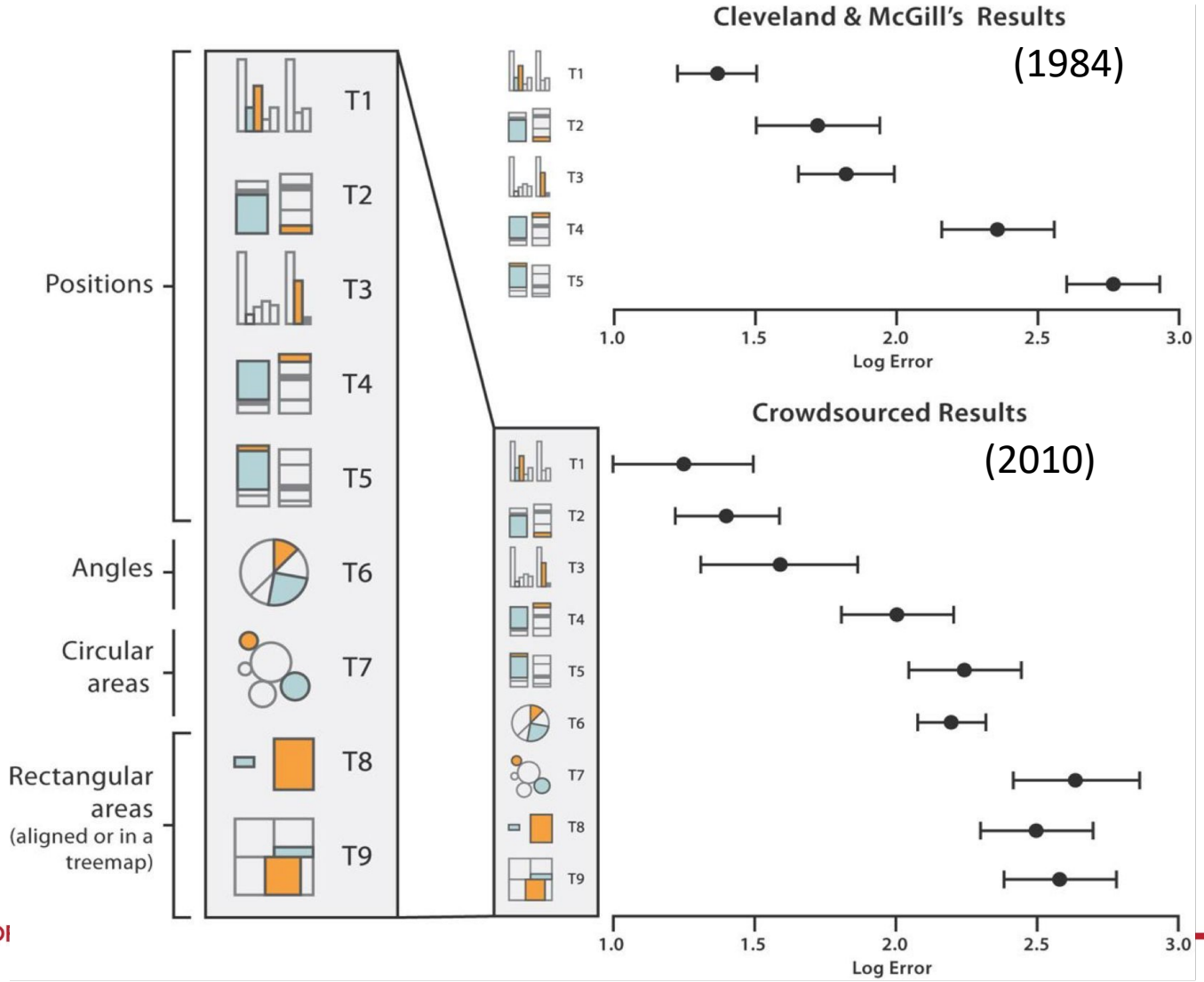
Psycho-visual effectiveness of the visuals

- About human perception…



Psycho-visual effectiveness of the visuals

- About human perception…



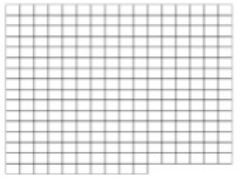
Cleveland and McGill, 1984
Heer and Bostock, 2010

Introduction

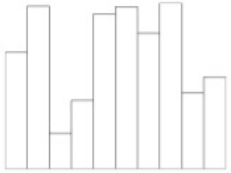
Different possible visualization graphics

Different visualization graphics

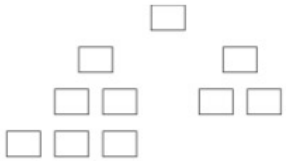
- Simple graphics



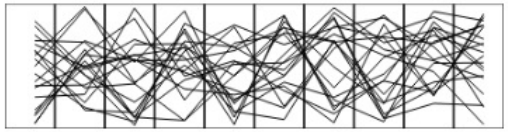
grid / isotype



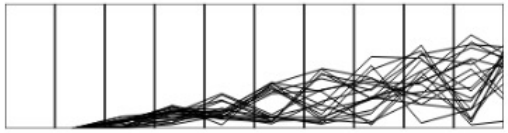
bar chart



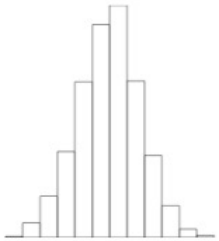
tree layout



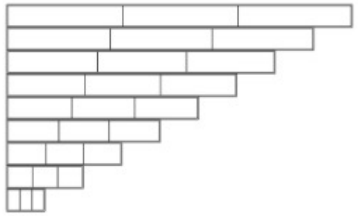
parallel coordinates



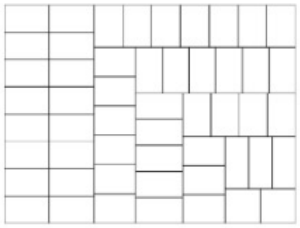
line chart



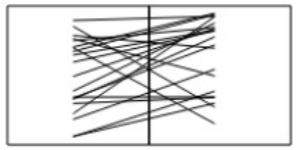
histogram



stacked chart



treemap



slope graph

Different visualization graphics

- For a very complete list of possible graphics, with comprehensive definitions
  - https://datavizcatalogue.com

Choosing the right chart for your data

- There are so many possible charts!
  - Graphs
  - Plots
  - Maps
  - Diagrams
  - ...

Choosing the right chart for your data

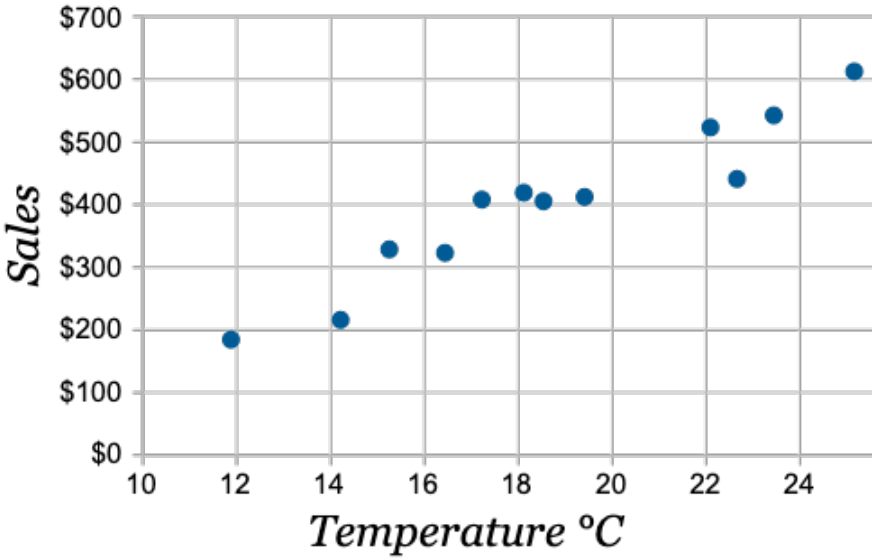
- In order to choose the right chart for your data (1/2):
  - **First**, you need to wonder what is the **type** of your variables
    - Numeric variable(s) only
    - Categorical variable(s) only
    - A mix of numeric and categorical variable(s)
    - Special case of geographical data
    - Special case of graphs / networks
    - Special case of time series
    - Other types of data (mix of time/geographical data, words, images, …)
  - **Second**, you need to wonder **how many** variables you want to visualize
    - 1 variable: univariate
    - 2 variables: bivariate
    - 3 variables: trivariate
    - More than 3 variables: multivariate

Choosing the right chart for your data

- In order to choose the right chart for your data (2/2):
  - **Third**, you need to wonder **why** you need that chart:
    - To summarize the data / understand the data
    - To show the relationship between different variables?
    - To compare several data samples?
  - **Fourth**, you need to know who will be your audience
    - General public?
    - Your classmates?
    - A boss / decision-maker of your enterprise?

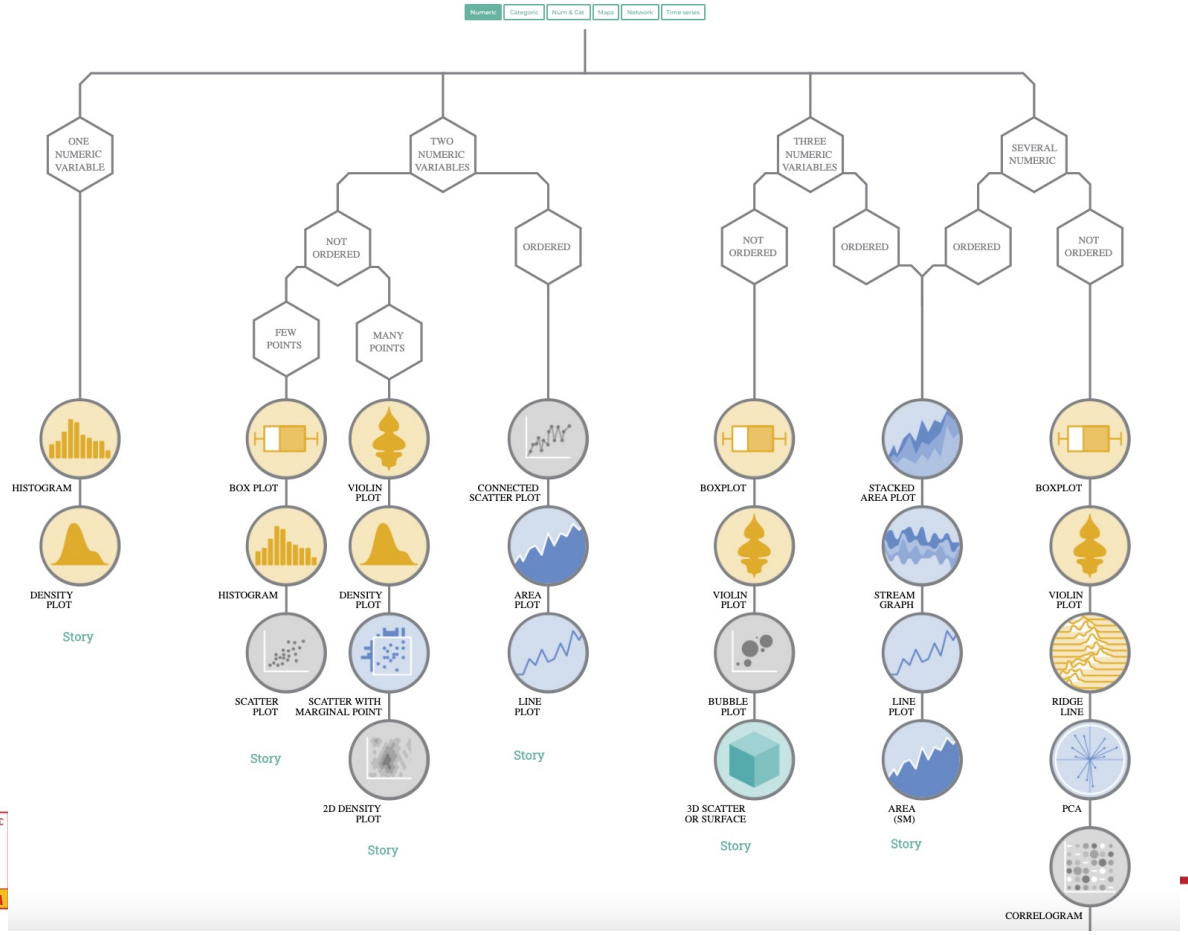
Example: scatter plot

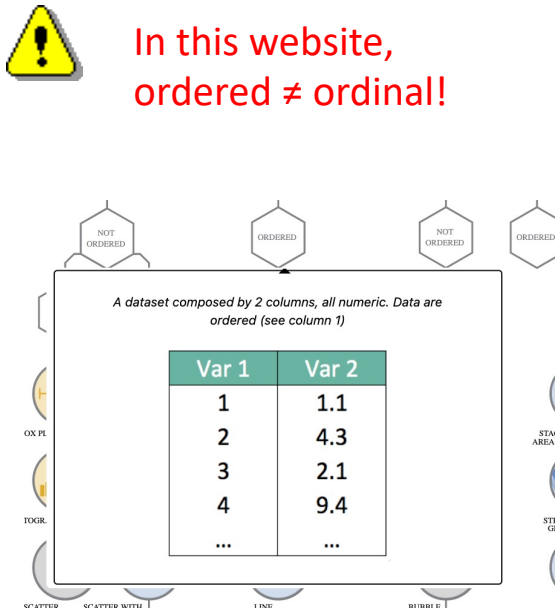
- Circle the good answer in the three lists
- What is the **type** of your variables
  - Numeric variable(s) only
  - Categorical variable(s) only
  - A mix of numeric / categorical variable(s)
  - Special case of geographical data
  - Special case of graphs / networks
  - Special case of time series
  - Other types of data (words, images, …)
- **How many** variables to visualize
  - 1 variable: univariate
  - 2 variables: bivariate
  - 3 variables: trivariate
  - multivariate
- **Why** you need that chart?
  - To summarize the data / understand the data distribution
  - To show the relationship between different variables?
  - To compare several data samples?



Choosing the right chart for your data

- There is a website that is **extremely** useful, and well done:
  - https://www.data-to-viz.com/
- Example, for numeric variables:

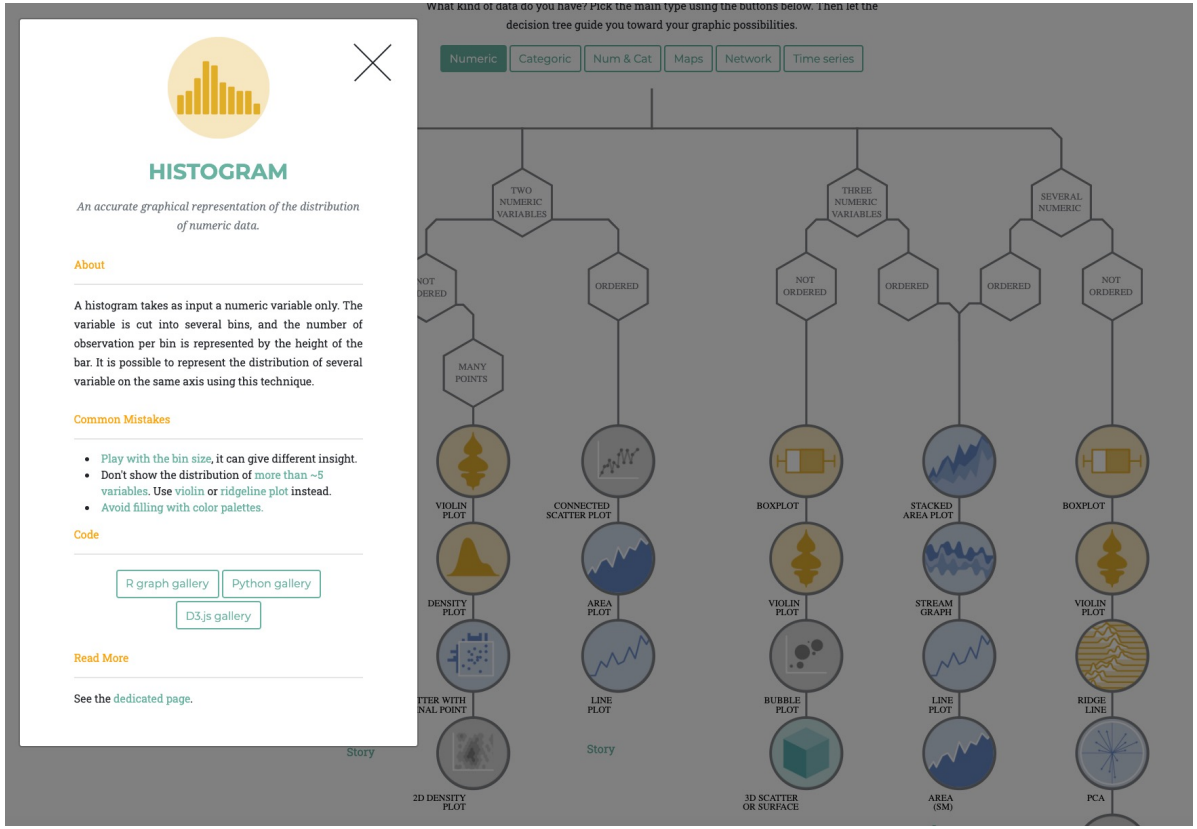




In this website,
ordered ≠ ordinal!

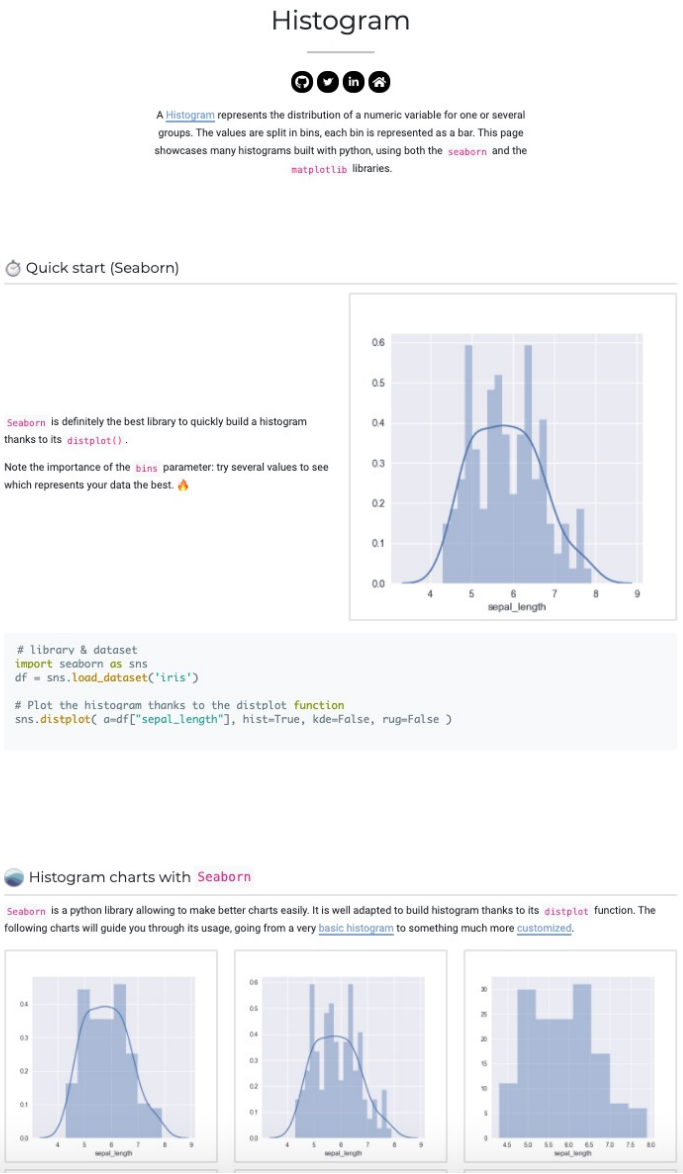
Choosing the right chart for your data

- There is a website that is **extremely** useful, and well done:
  - https://www.data-to-viz.com/
- Example, for histograms:



Choosing the right chart for your data

- There is a website that is **extre**
  - https://www.data-to-viz.com/
- Example, for histograms
  - When I click on « Python gallery »



In this lecture, I will

- Explain some of the graphics that are difficult to understand on your own
- Explain some original graphics that cannot easily be found from websites such as data-to-viz
- I will organize the charts depending on the type(s) of variables

  
  N.B. One chart can be used for multiple types of variables / multiple objectives -> classification is not exclusive
- I'll **not** give all possible charts, nor their code(s)
- You'll have to study them on your own, and come up with your questions during the next lecture

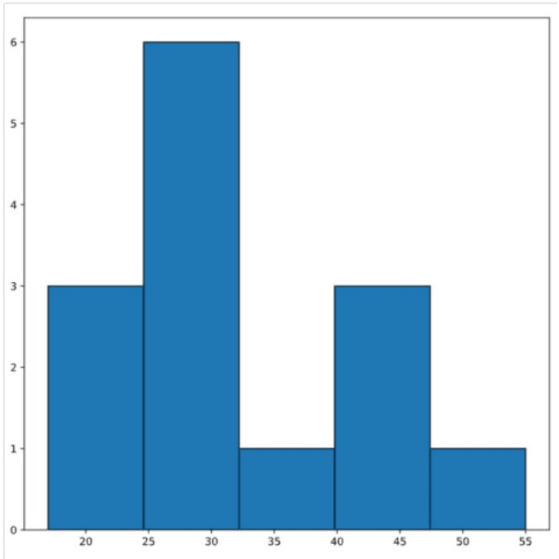
Examples of data visualization charts

Numeric variables

Univariate graphics for numeric variable summarization

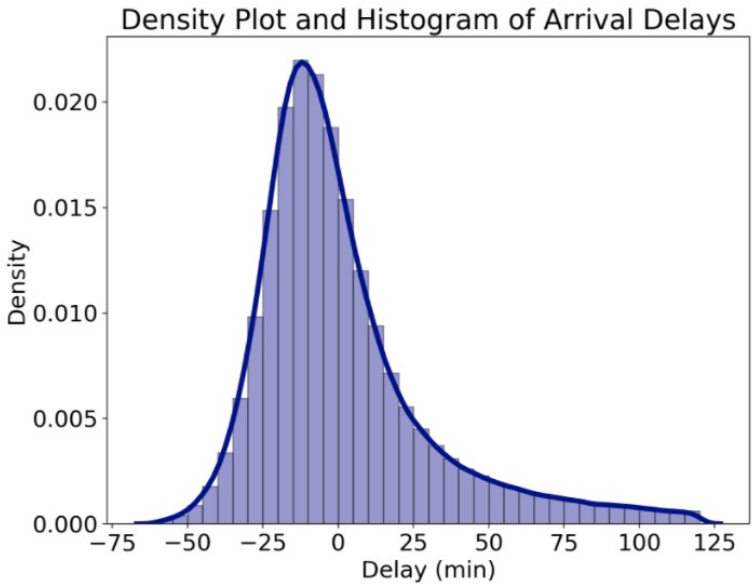
- For showing/understanding the data distribution of 1 numeric variable

Histogram



Density plot

It uses a kernel density estimate to show the probability density function of the variable: https://www.data-to-viz.com/graph/density.html#:~:text=A%20density%20plot%20is%20a,used%20in%20the%20same%20concept.
It is a smoothed version of the histogram and is used in the same concept

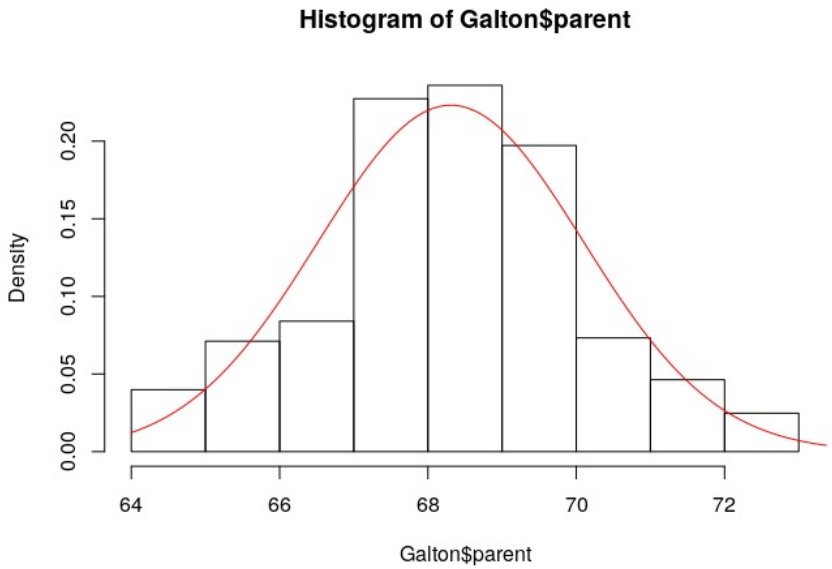


Possible univariate graphics for numeric variable summarization

- For showing/understanding the data distribution of 1 numeric variable

Density curve with histogram

For comparing the data distribution to a theoretical model (*e.g.* Gaussian with same mean and variance)



Do **not confuse** between density plot and density curve with histogram!!!
- bear different information
- used for different purposes

Univariate graphics for numeric variable summarization

- For showing/understanding the data distribution of 1 numeric variable
- Other possible univariate graphics include:

Boxplot

Violin plot=
Boxplot+ data distribution (density plot) -> better than boxplot *e.g.* when more than one peak

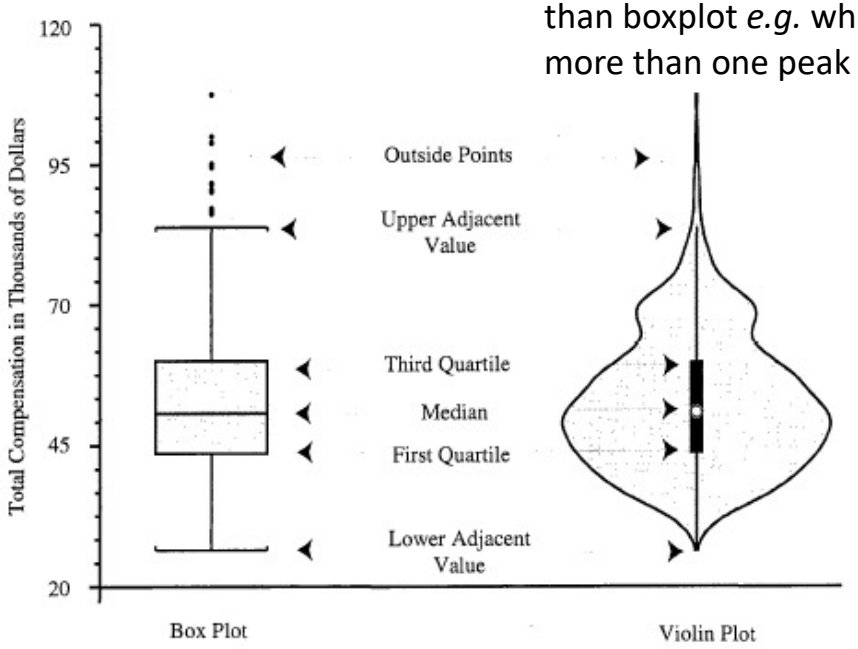
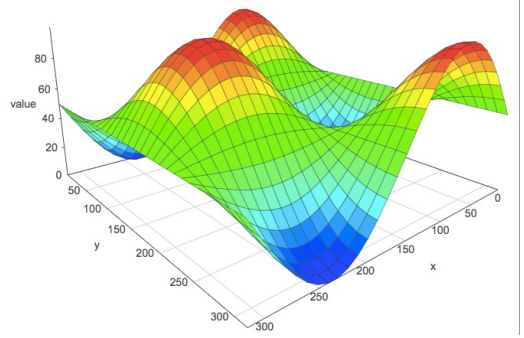


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

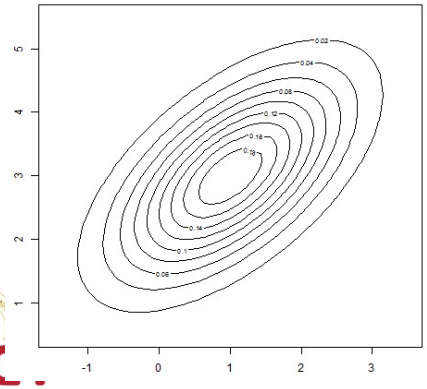
Possible multi-variate graphics for numeric variable summarization

- For showing/understanding the data distribution of multiple numeric variables

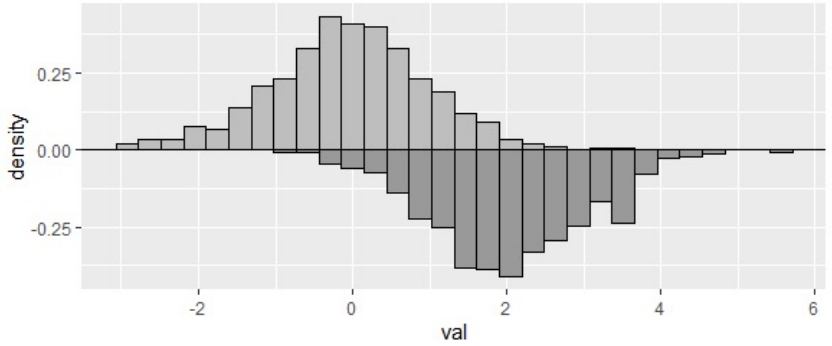
3D scatterplot (3 variables)



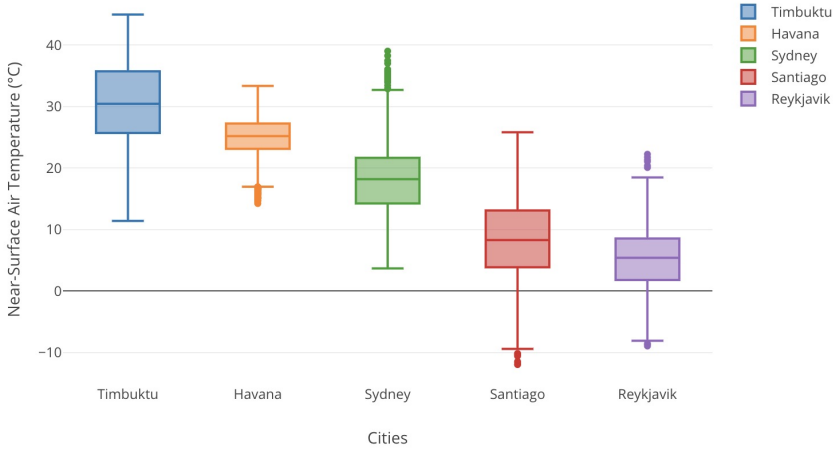
Contour plot (3 variables)



Mirror histograms (2 variables)

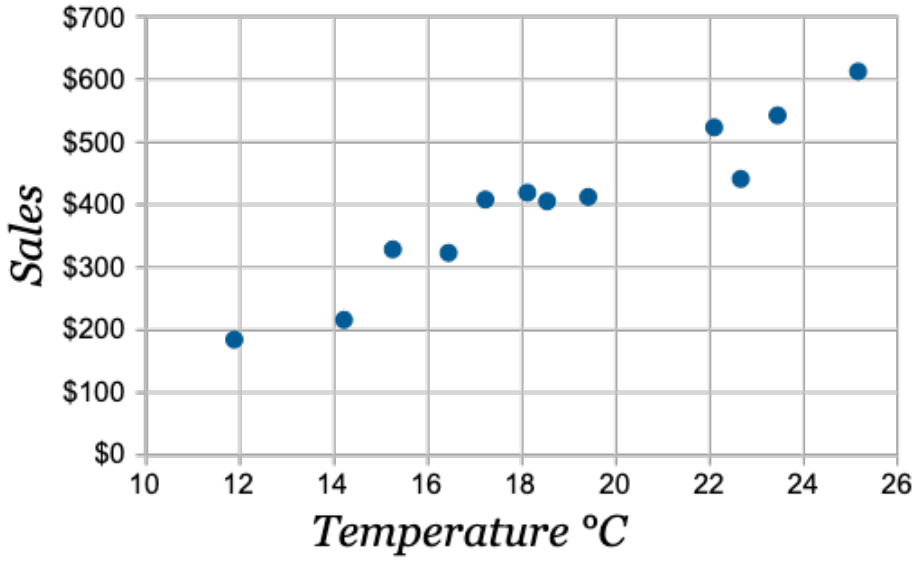


Box plots (multiple variables)

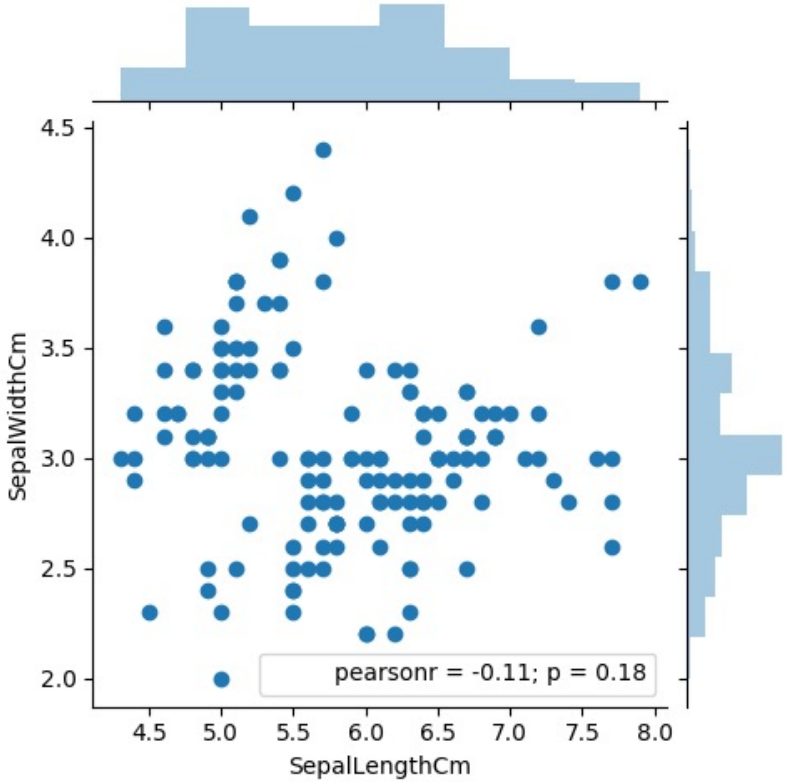


Possible multi-variate graphics for studying the link between multiple numeric variables

- Scatterplot



- Link+summary:
  - Marginal histogram

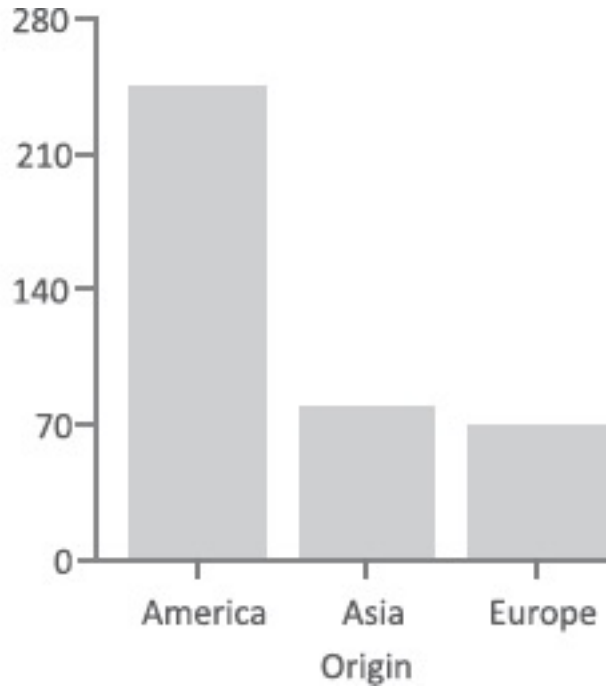


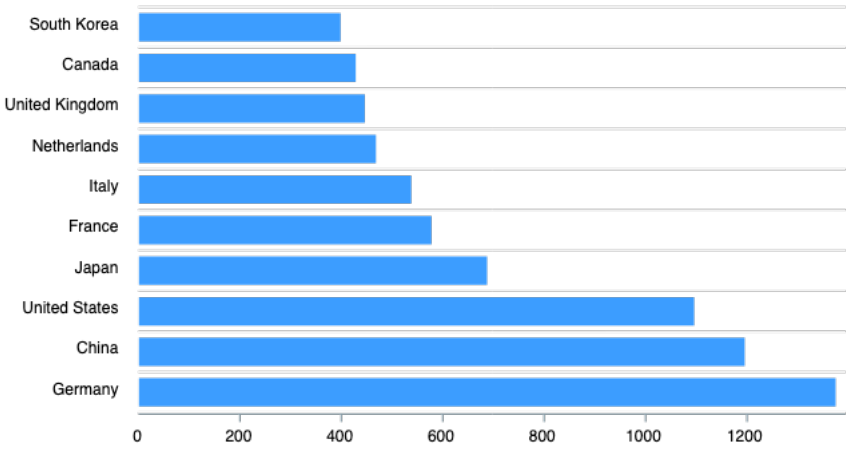
Examples of data visualization charts

Categorical variables

Univariate graphics for categorical variable summarization

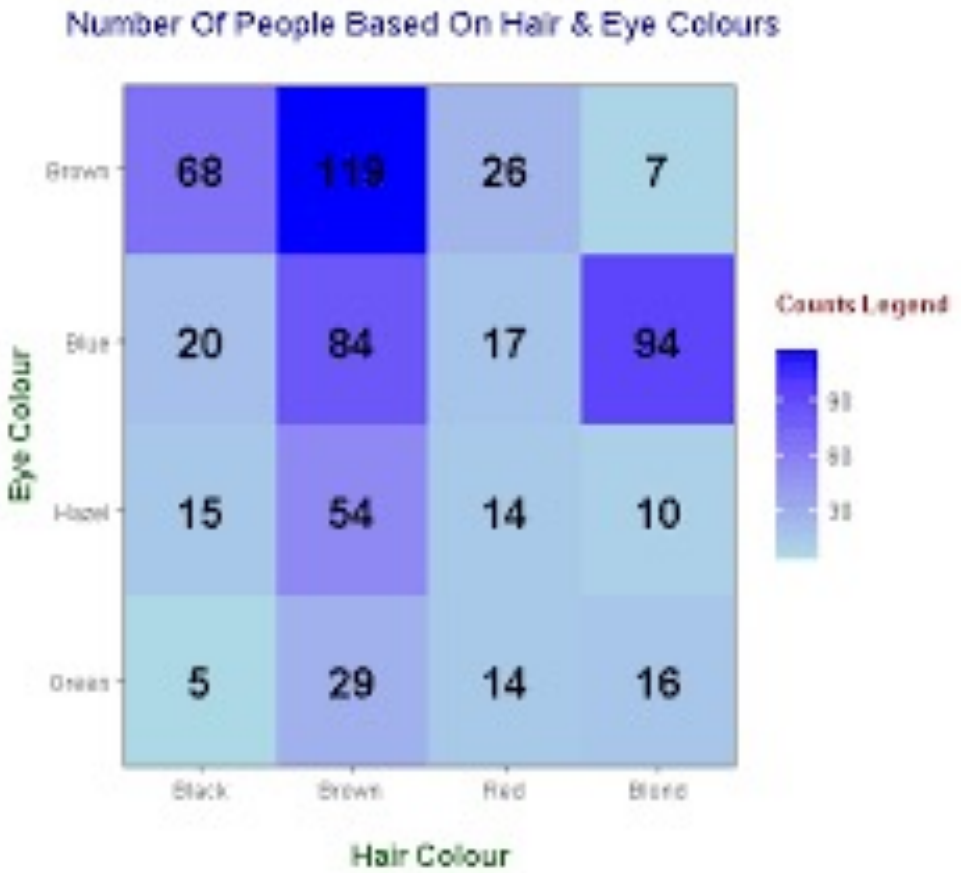
- For showing/understanding the data distribution of 1 categorical variable
  - Bar charts, also called barplots (vertical / horizontal)





Multivariate graphics for categorical variables

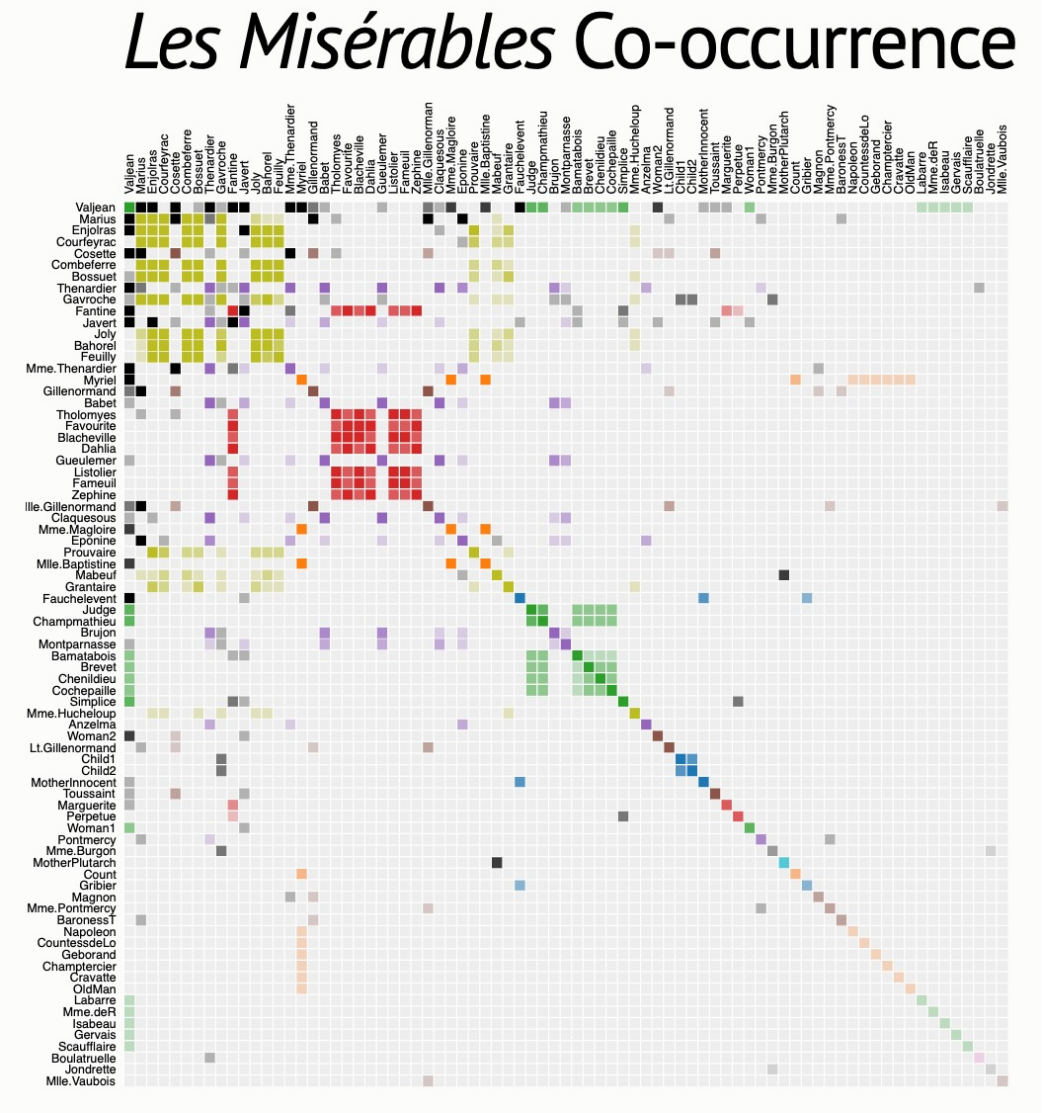
- Graphical representation of a contingency matrix
  - Can be with, or without, the numbers in the cells
  - Usually, to study the link between multiple variables
  - Special case of a **heatmap**



Multivariate graphics for categorical variables

- Another graphical representation of a contingency matrix
  - Co-occurrence matrix
    - Usually, to study the link between multiple categorical variables
    - Can also be used to study the link between multiple numeric, discrete variables (with finite possible values)
      - For describing an image texture, for instance

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



Possible data visualization charts

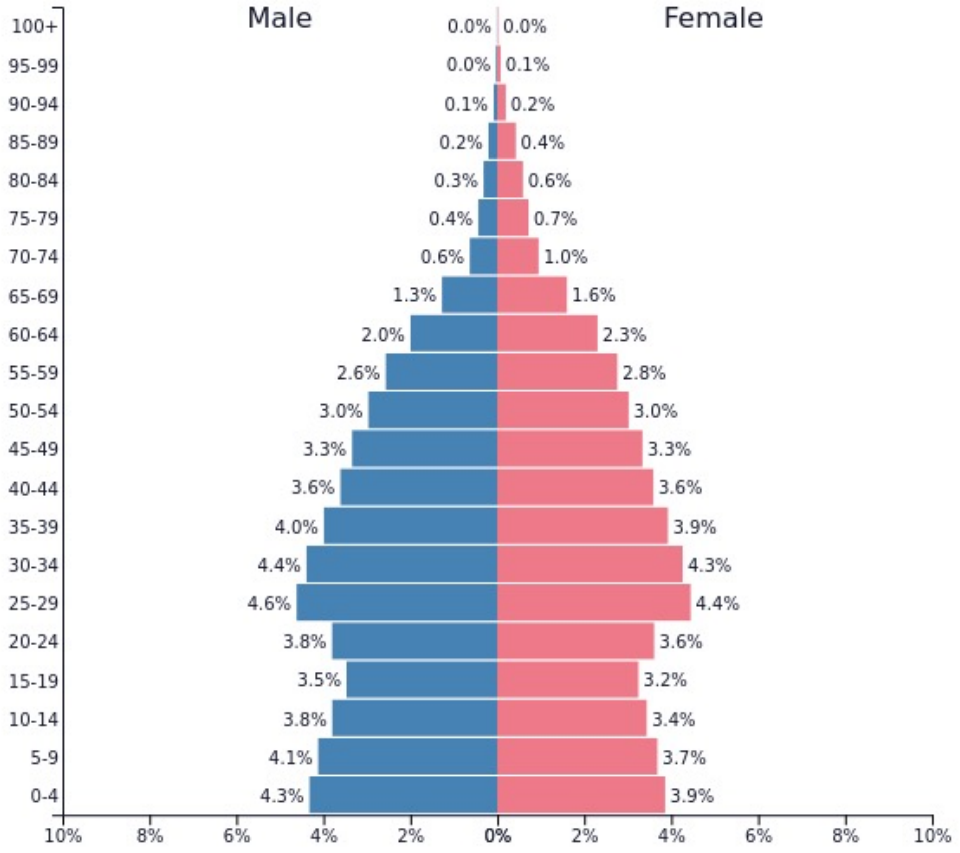
Mix of numeric variables + categorical variables

Multivariate graphics for a mix of numeric / categorical variables

- In the case of a mix between numeric + categorical variables, graphics are especially useful for EDA
  - Because statistics are limited / difficult to understand

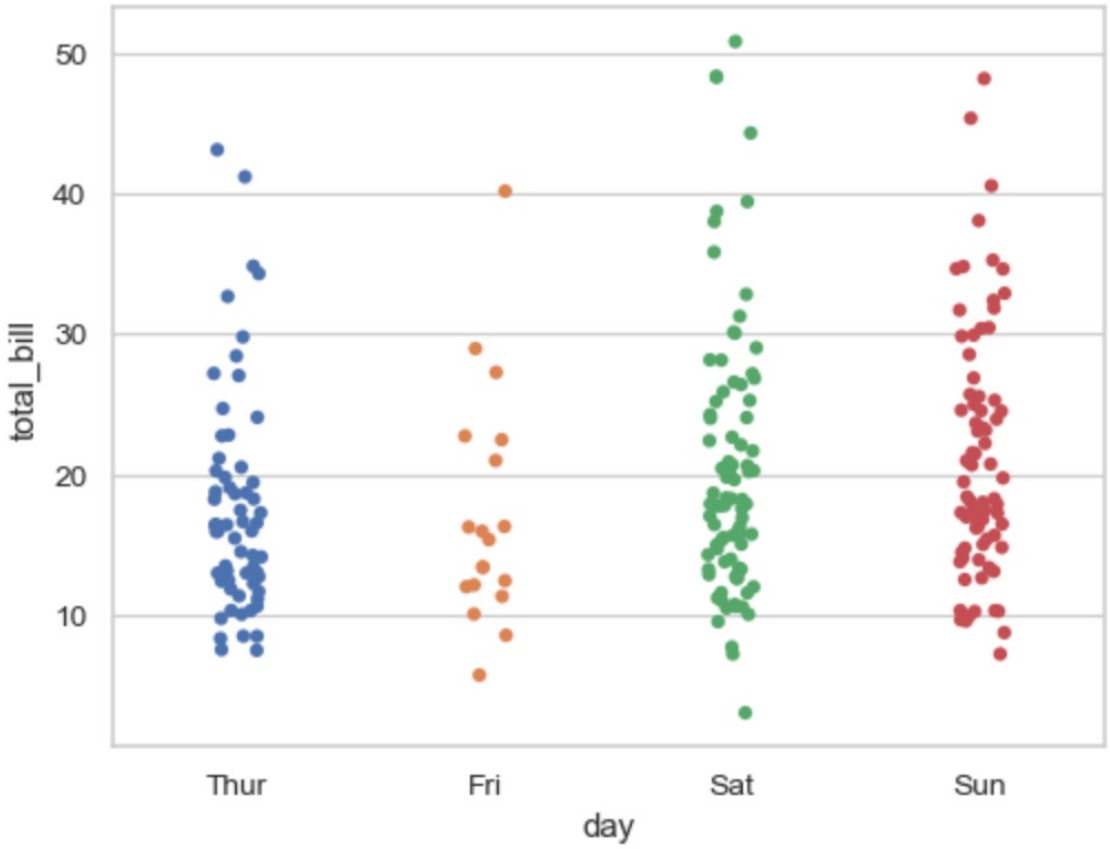
Multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between 1 numeric + 1 categorical variable
  - **Population pyramid**



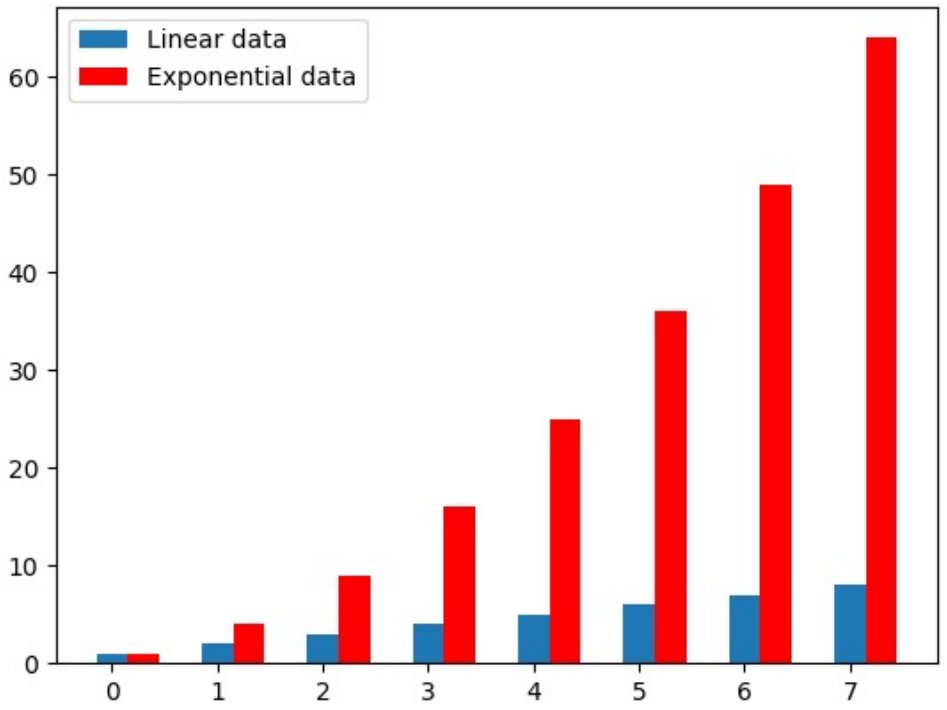
Possible multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between 1 numeric + 1 categorical variable
  - **Strip plot**



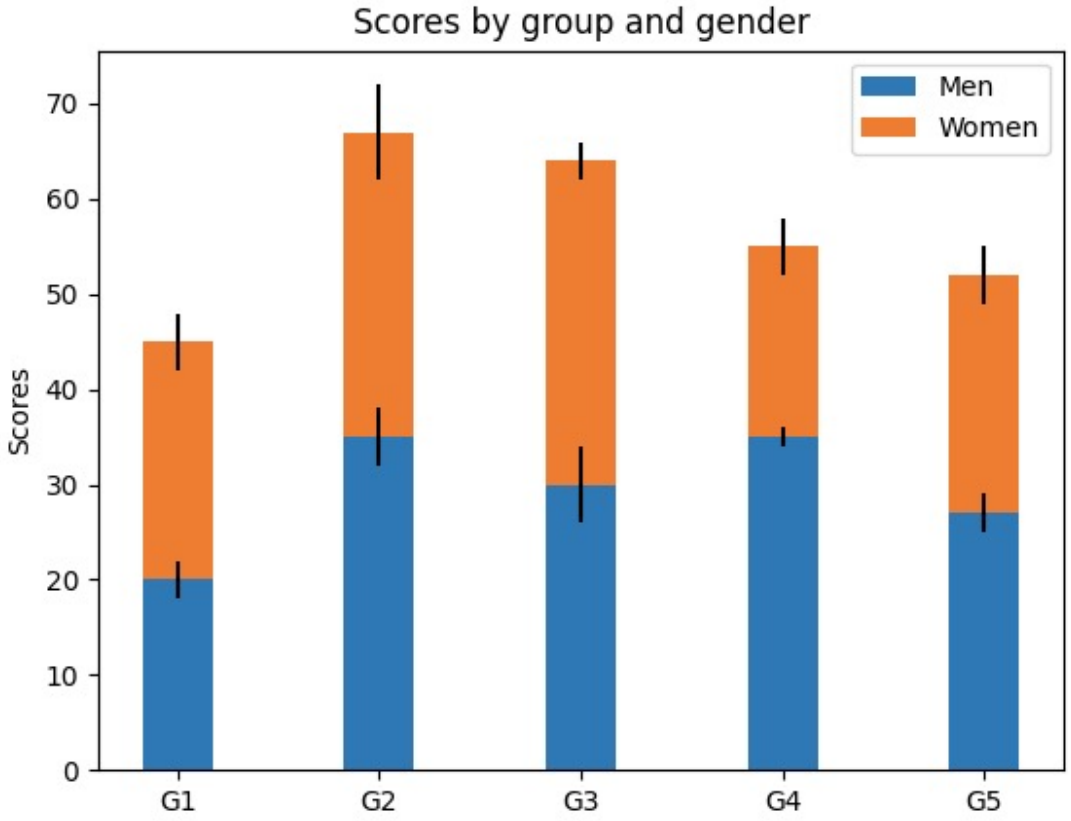
Possible multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between 1 numeric + multiple categorical variables
  - **Vertical bar chart**



Multivariate graphics for a mix of numeric / categorical variables

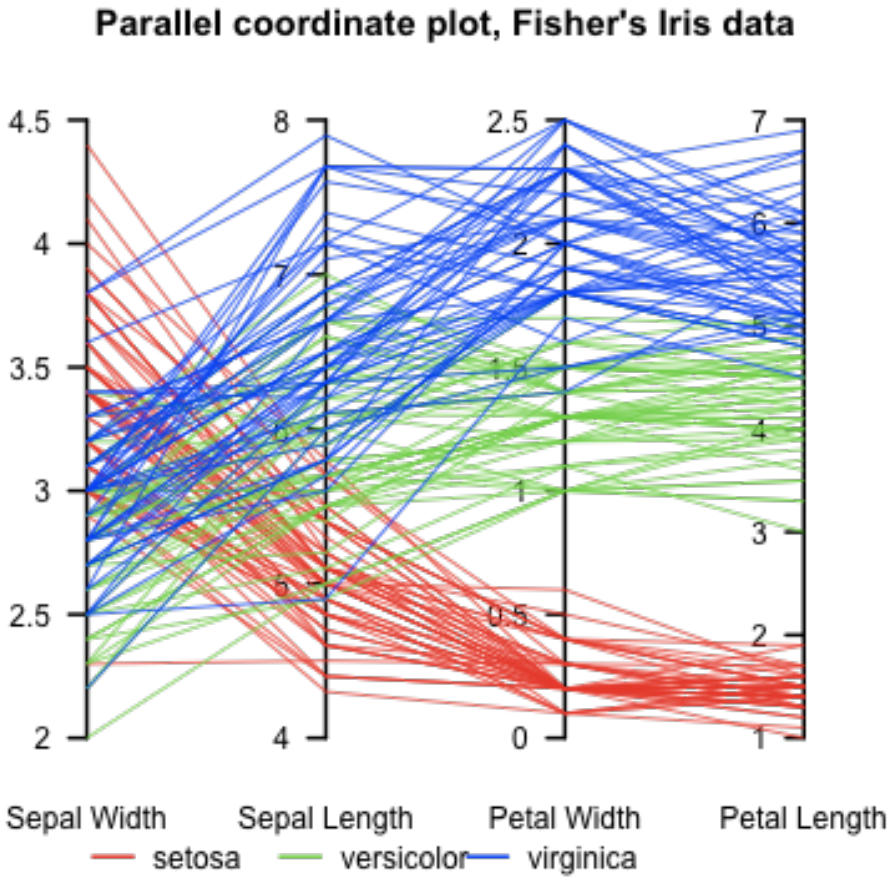
- For showing / understanding the data distributions / links between 1 numeric + multiple categorical variable
  - **Stacked bar charts**



Multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between multiple numeric + 1 categorical variable

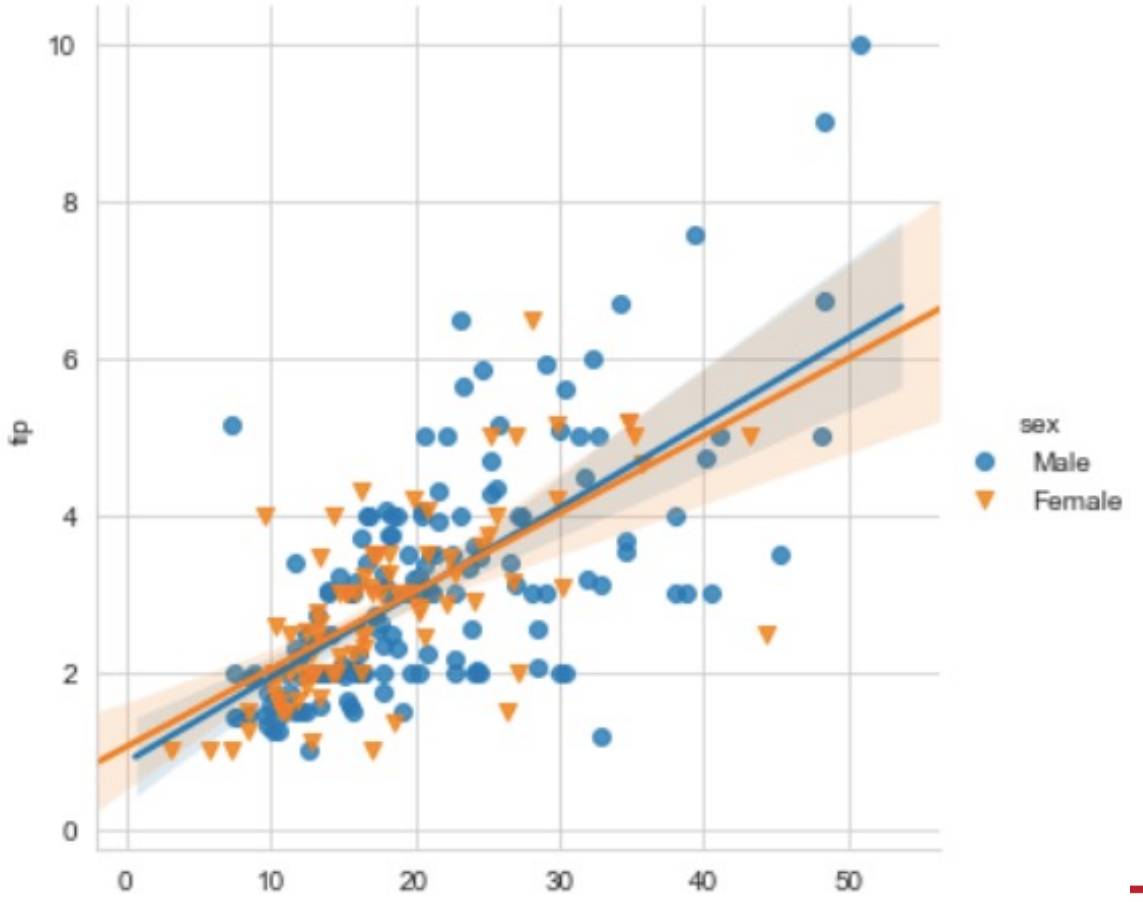
  - **Parallel coordinates plot**
    - Each vertical, parallel axis corresponds to a numeric variable
    - A point in *n*-dimensional space (here, 1 flower) is represented as a **polyline**
      - the position of the vertex on the i-th axis corresponds to the value of the i-th attribute for this record
    - It might be interesting to try different axis arrangements



Multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between multiple numeric + 1 categorical variable

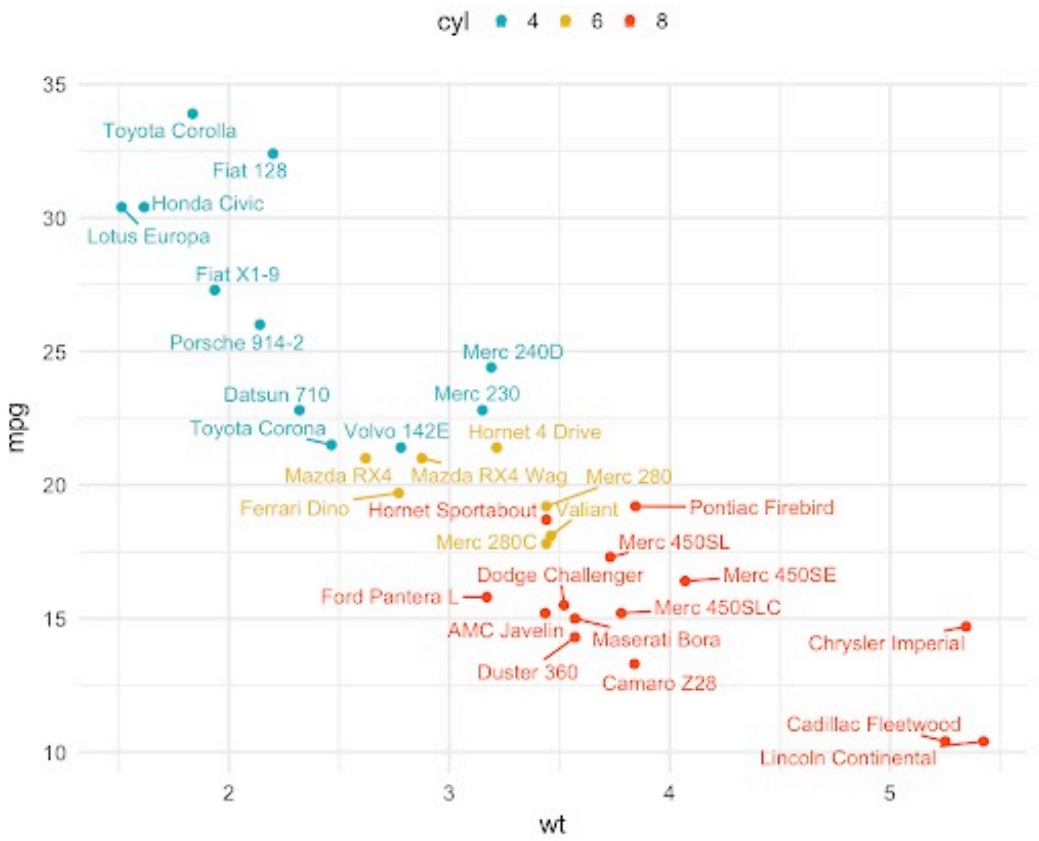
  - **Regression plot**



Multivariate graphics for a mix of numeric / categorical variables

- For showing / understanding the data distributions / links between multiple numeric + 1 categorical variable

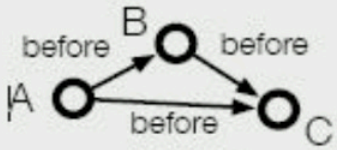
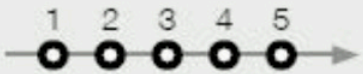
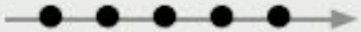
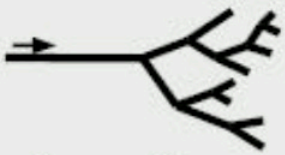
  - **Scatter plot with label**
    - Useful for clustering



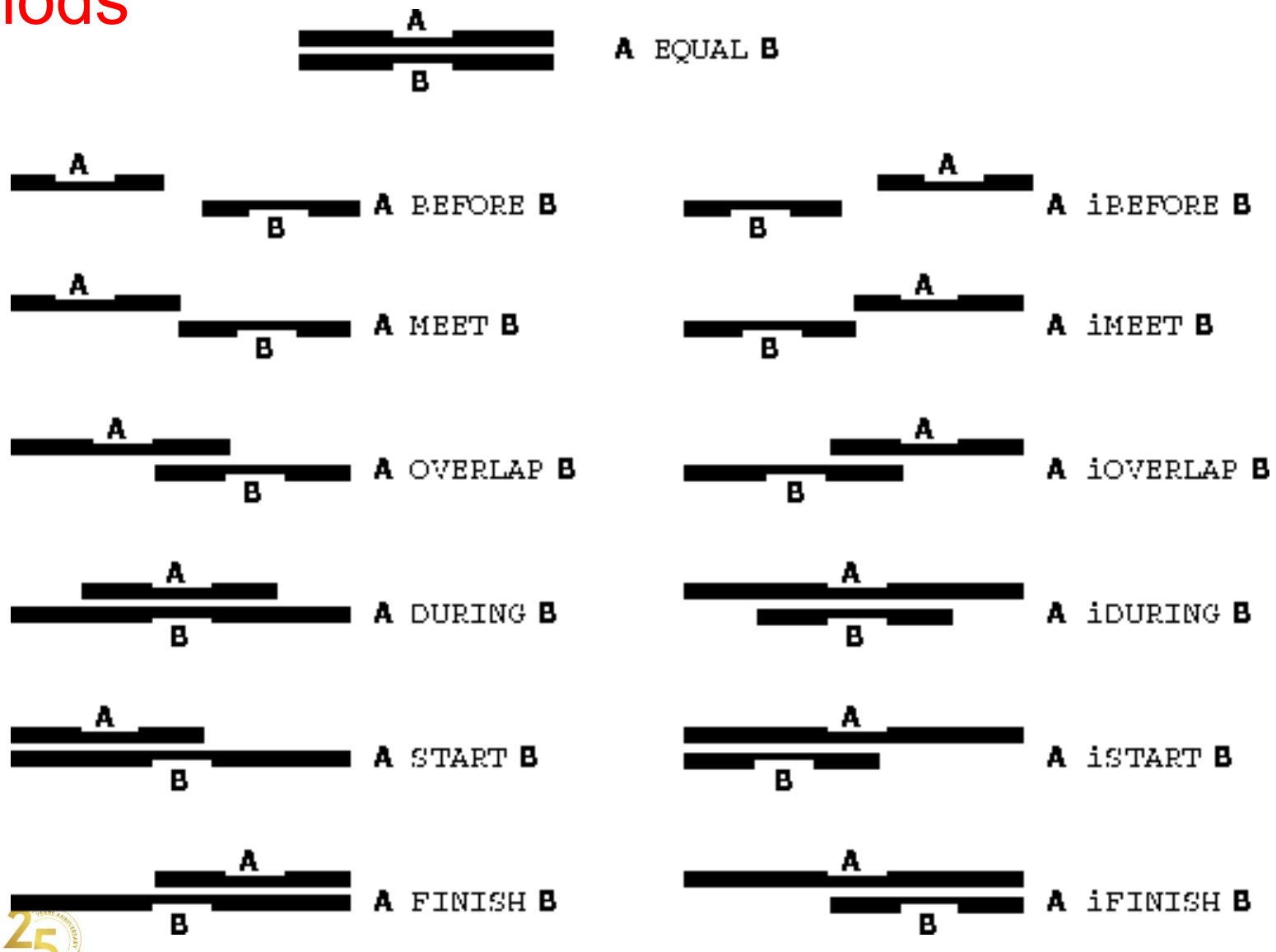
Examples of data visualization charts

Special case of time-dependent variables

Examples of visuals for time visualization

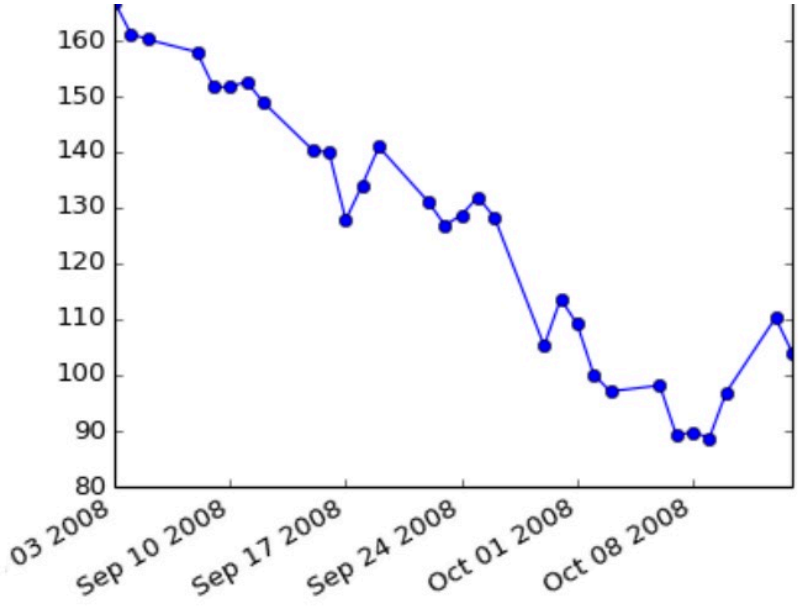
scale	 ordinal	 discrete	 continuous
scope	 point-based	 interval-based	
arrangement	 linear	 cyclic	
viewpoint	 ordered	 branching	 multiple perspectives

Multivariate graphics for summarization of time periods

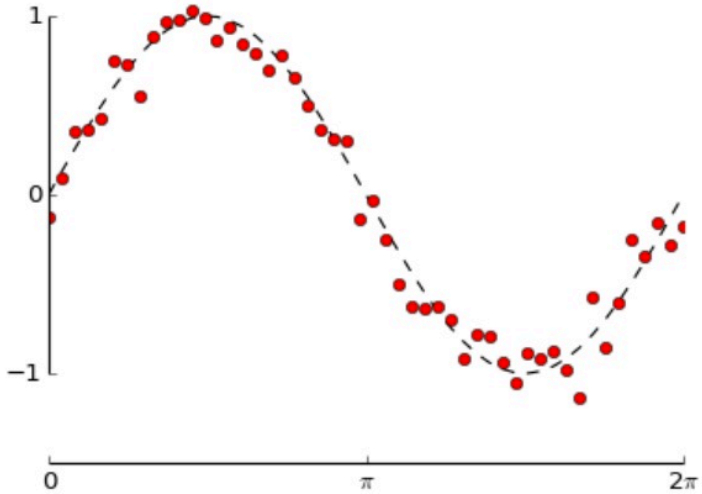


Multivariate graphics for summarization of time series (1 numeric variable + time)

Run sequence plot



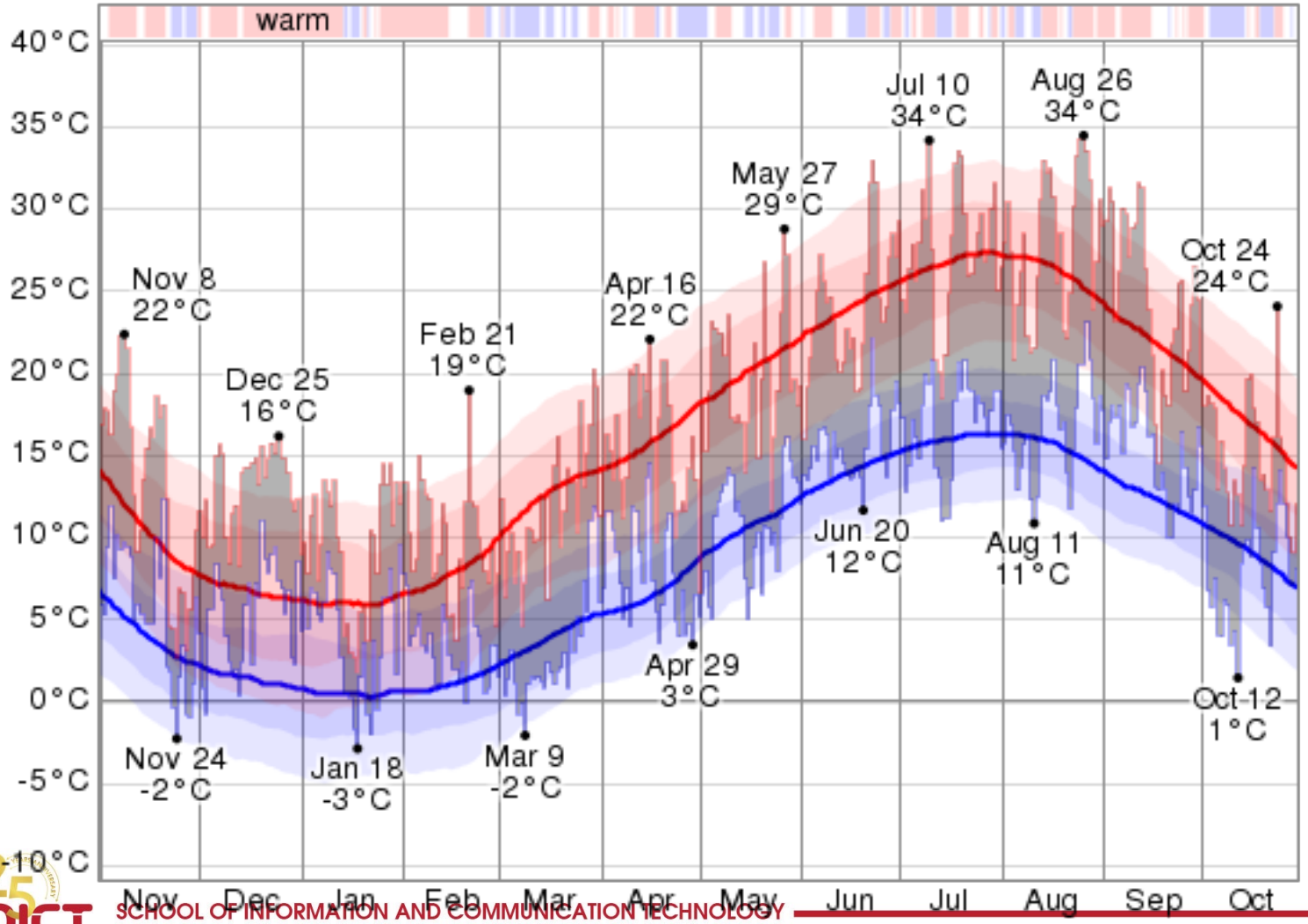
Run sequence plot + B-spline smoothing



matplotlib gallery

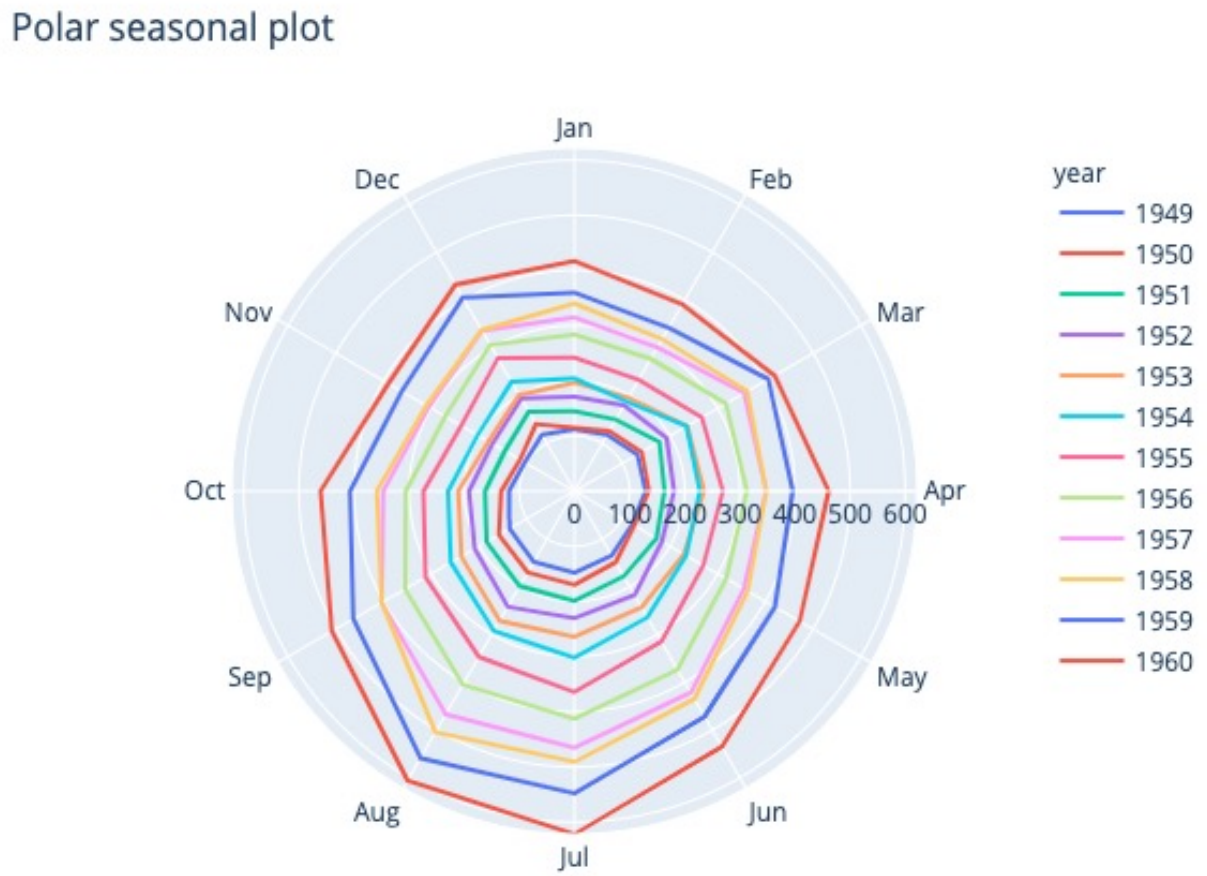
Multivariate graphics for summarization of time series (1 numeric variable + time)

Run sequence plot + average + confidence interval



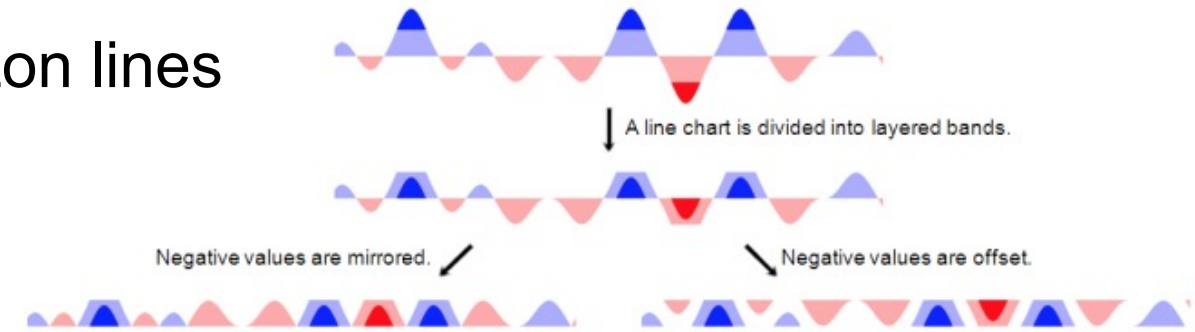
Multivariate graphics for summarization of time series (1 numeric variable + time)

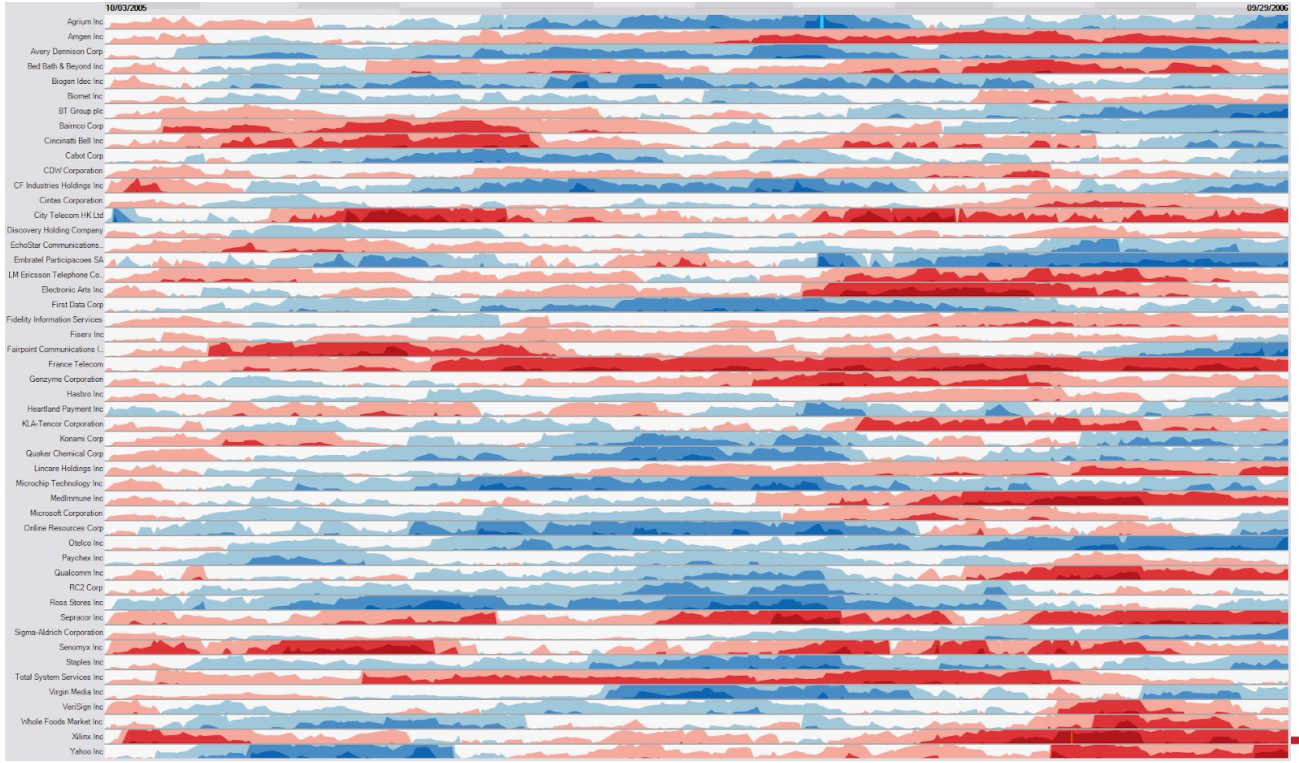
- Polar seasonal plot (number of flight passengers, per year)
  - Especially useful when the values of the numeric variable increase with time



Multivariate graphics for summarization of time dependent variables (multiple numeric variables + time)

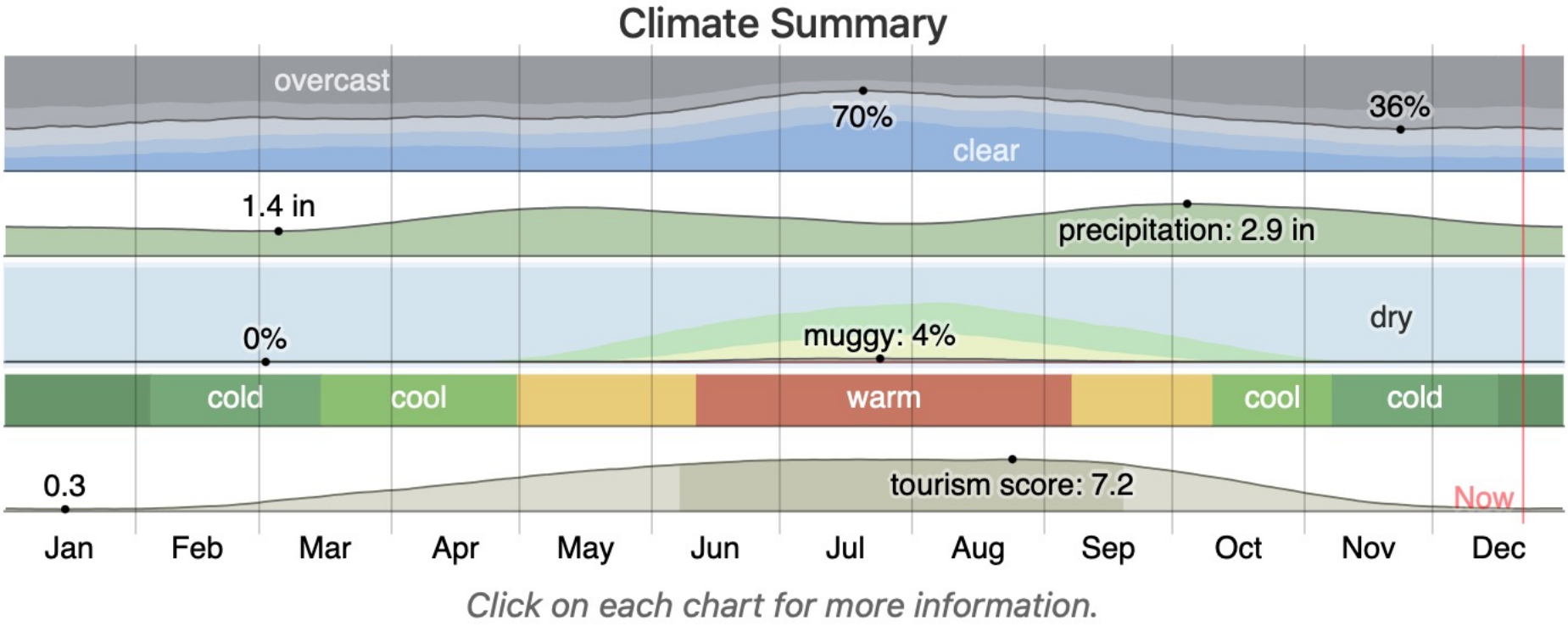
- Horizon lines





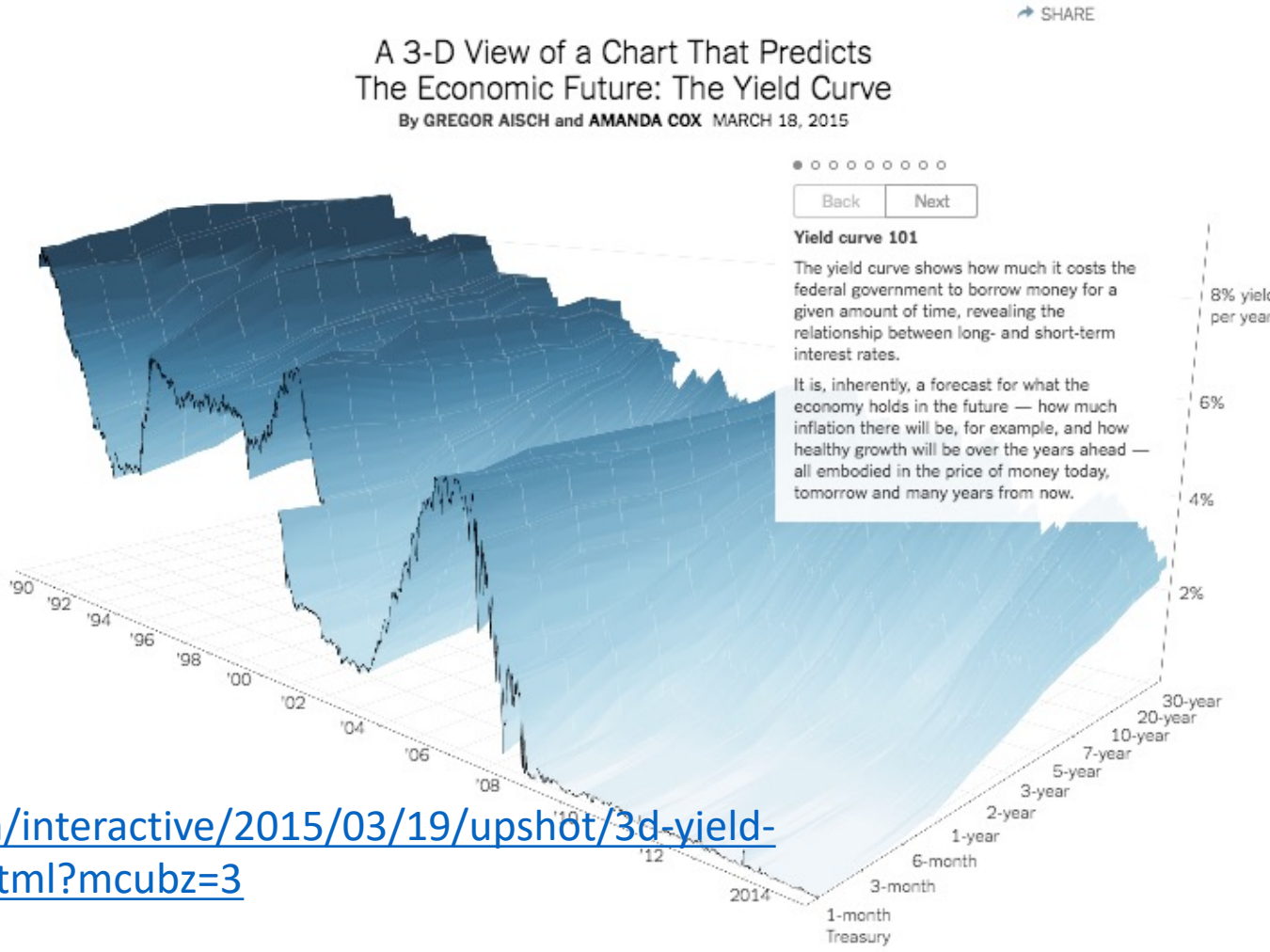
Multivariate graphics for summarization of time dependent variables (multiple numeric variables + time)

- Interactive
  - https://weatherspark.com/y/50604/Average-Weather-in-Lyon-France-Year-Round



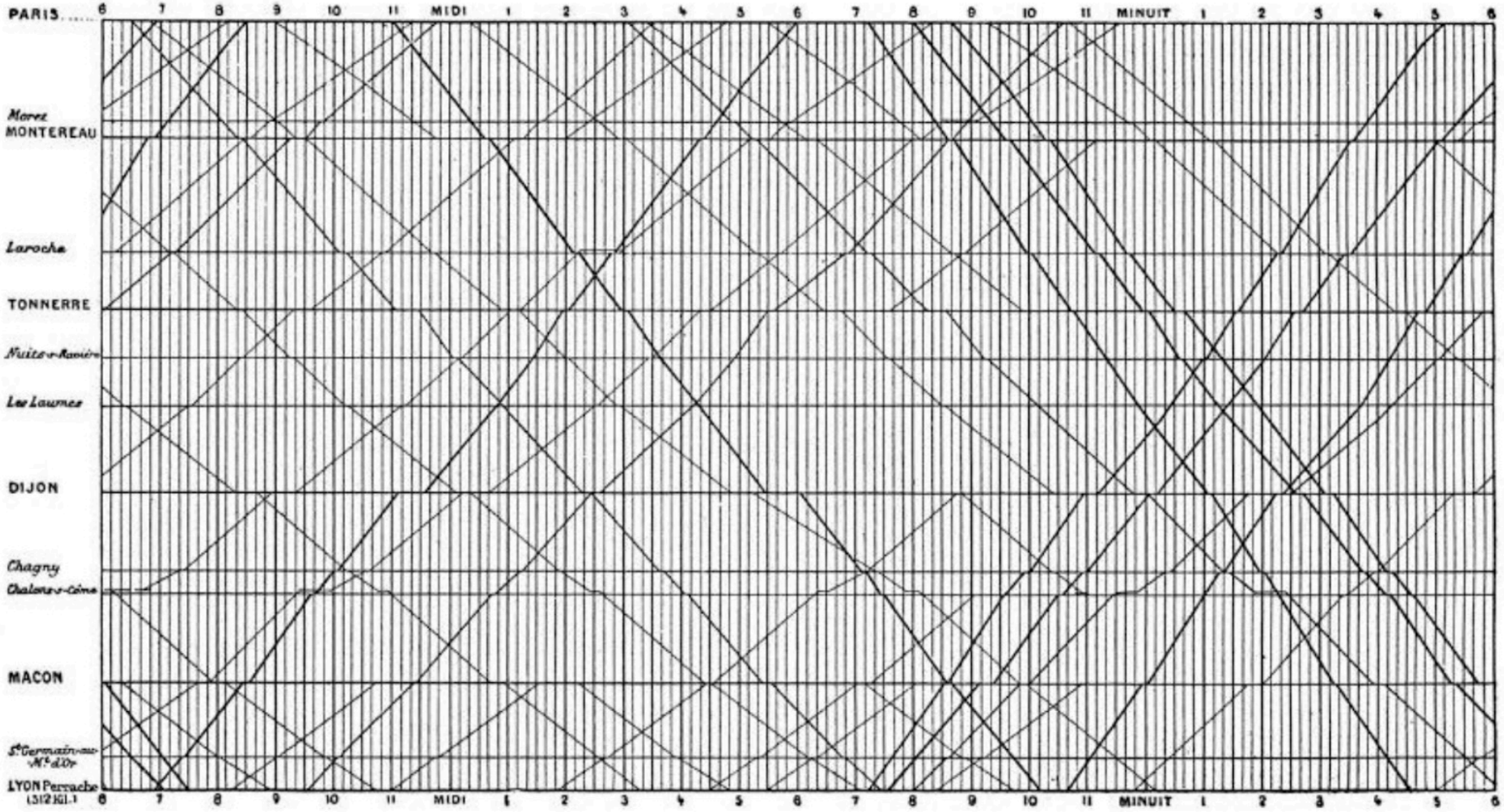
Multivariate graphics for studying the link between multiple time dependent variables (multiple numeric variables+time)

- 3D views



https://www.nytimes.com/interactive/2015/03/19/upshot/3d-yield-curve-economic-growth.html?mcubz=3

Multivariate graphics for summarization of time dependent variables (1 numeric variable + 1 categorical variable + time)

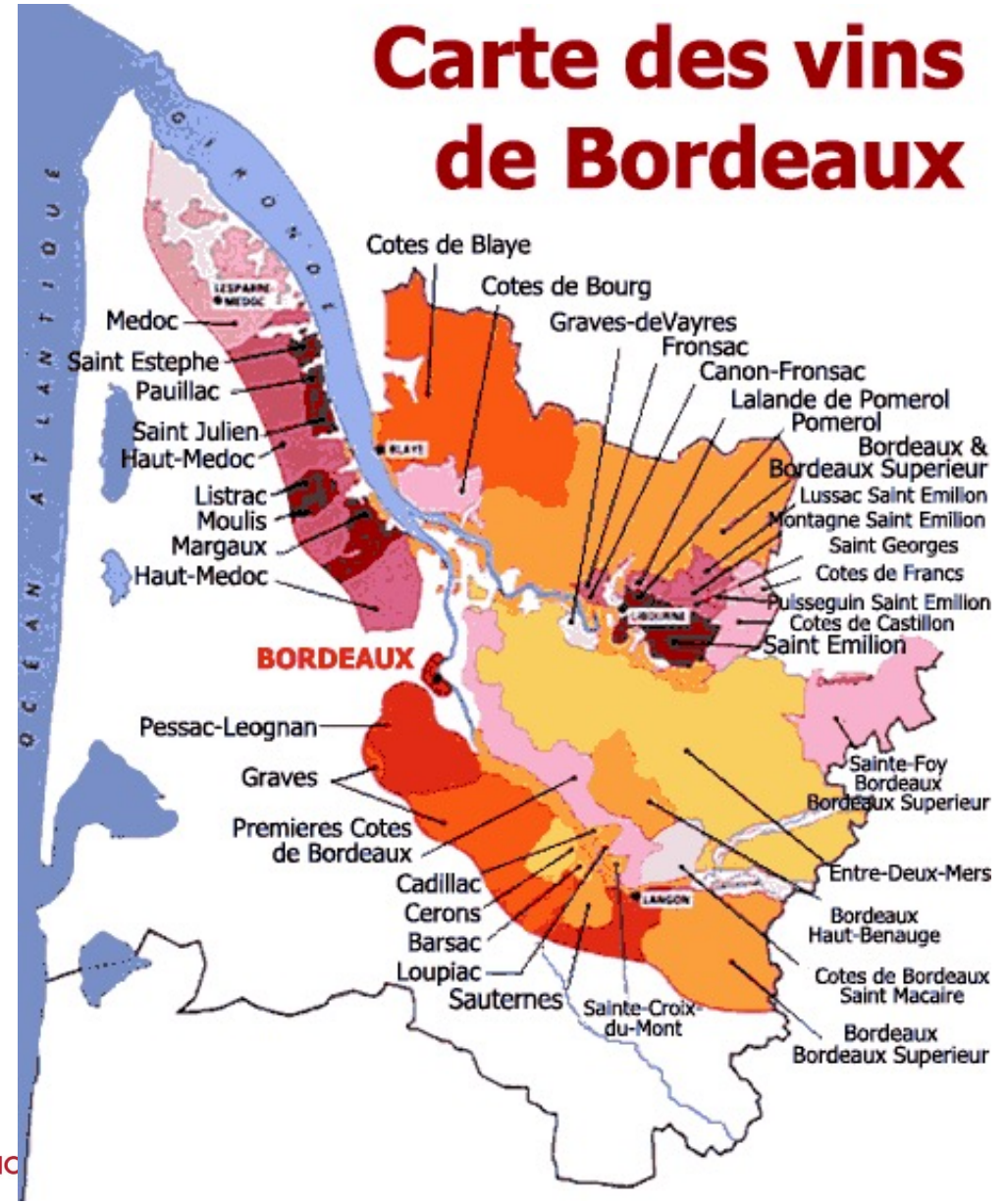


Multivariate graphics for summarization of time dependent variables (1 numeric variable + 1 categorical variable + time)

Possible data visualization charts

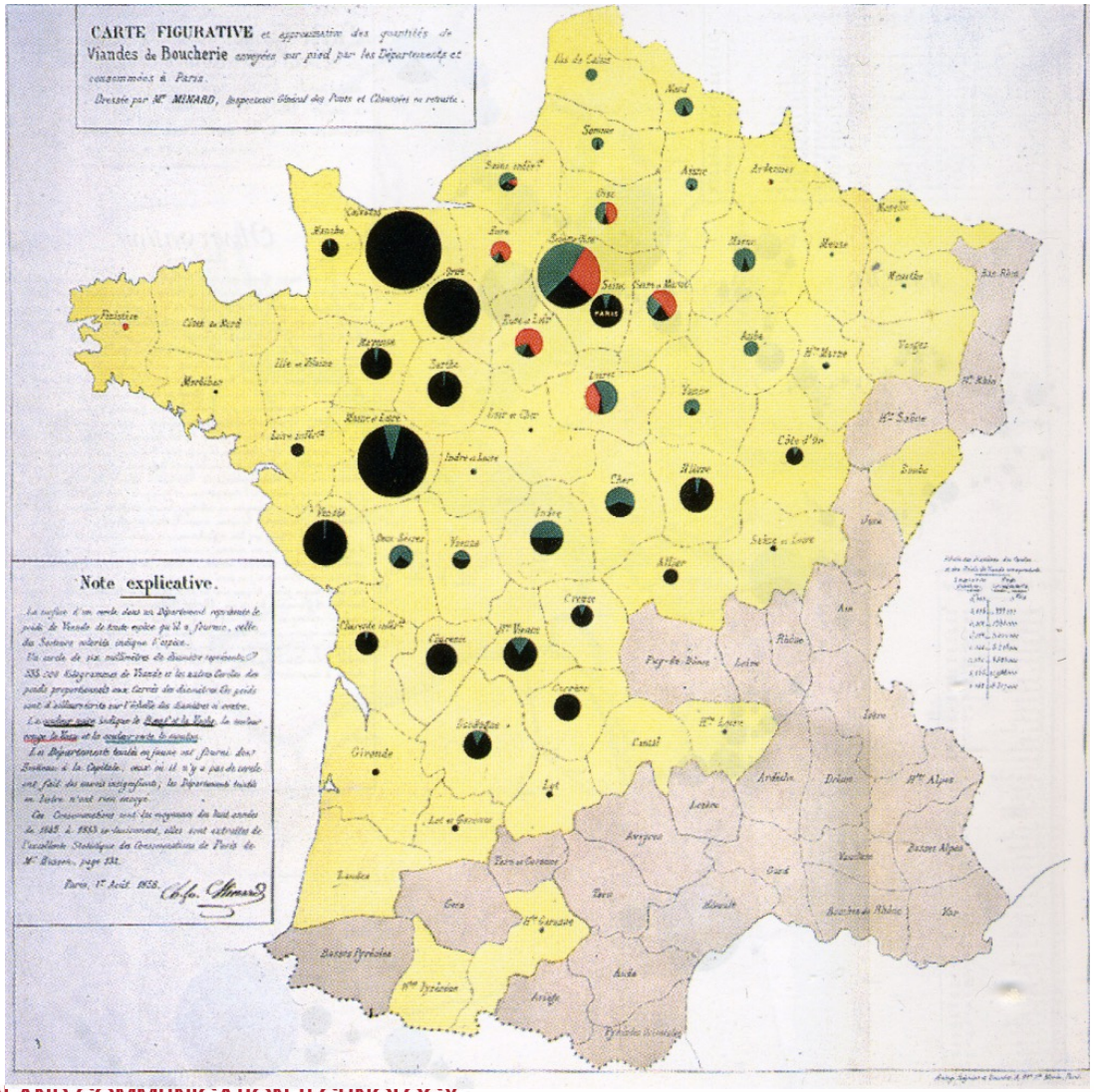
Special case of geographical data

Multivariate graphics for geographical + categorical variables



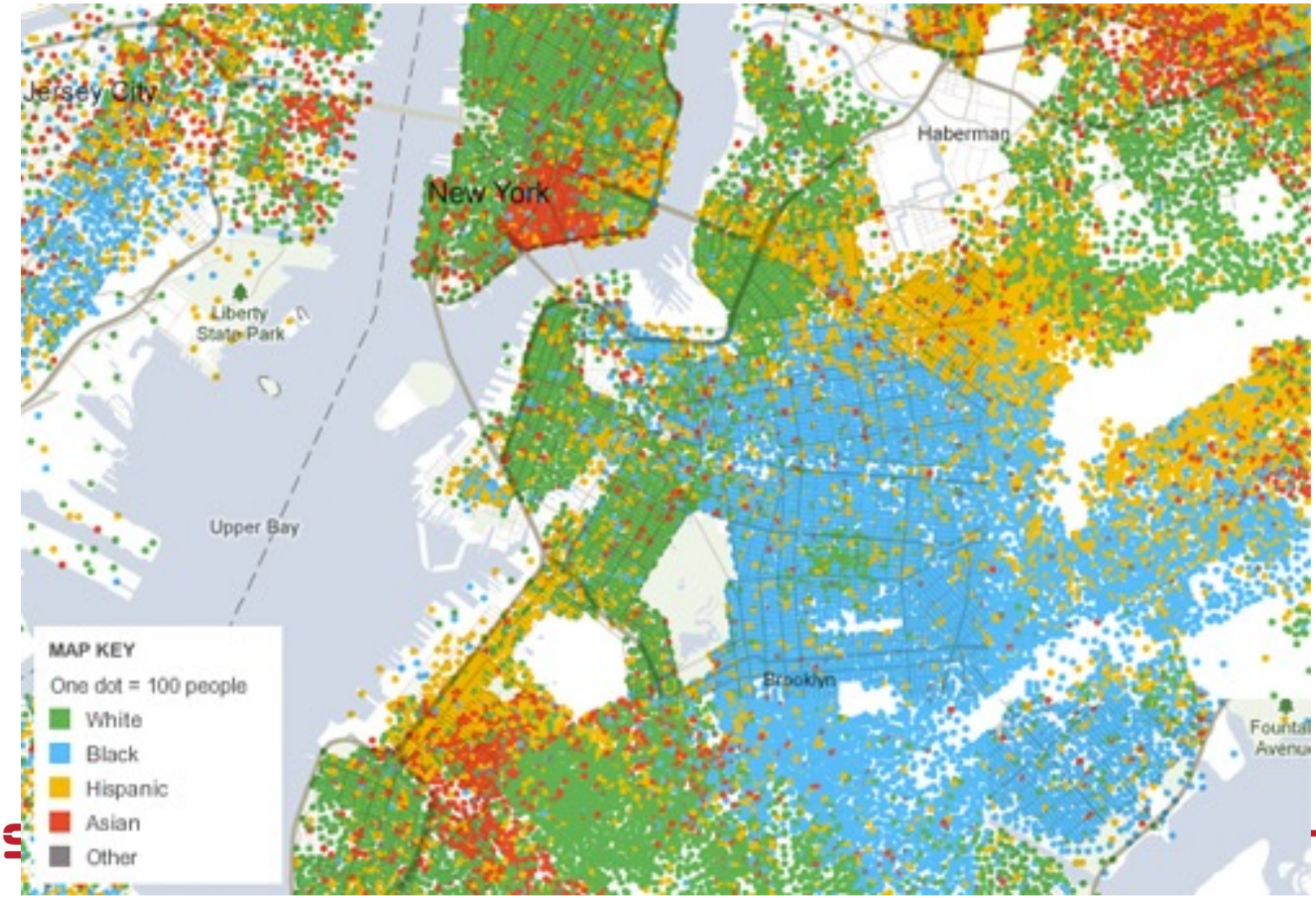
Multivariate graphics for geographical + categorical variables

- Early days



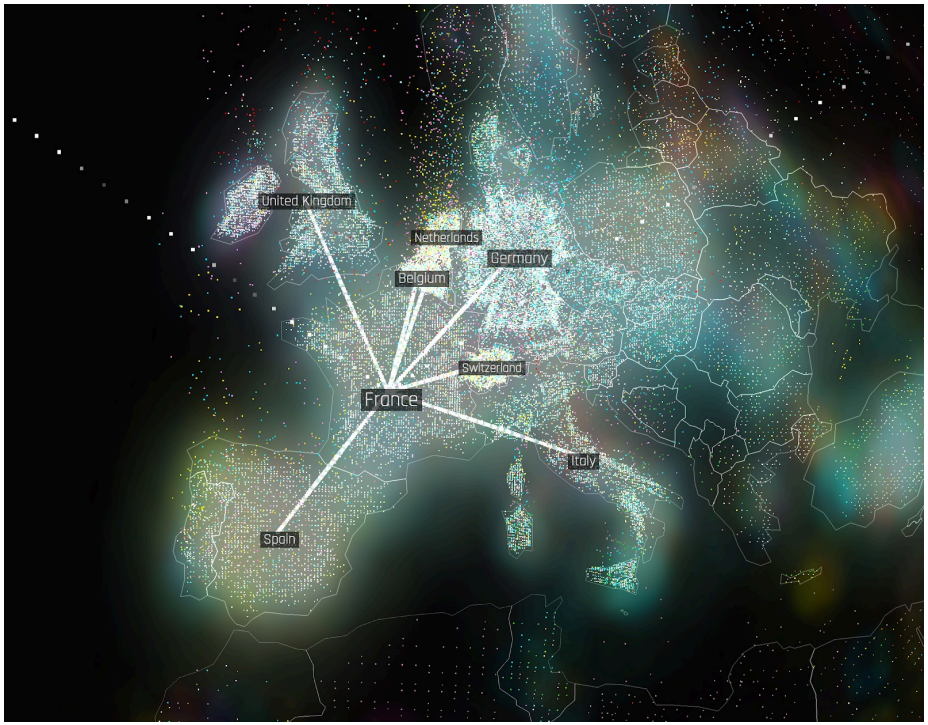
Multivariate graphics for geographical + categorical variables

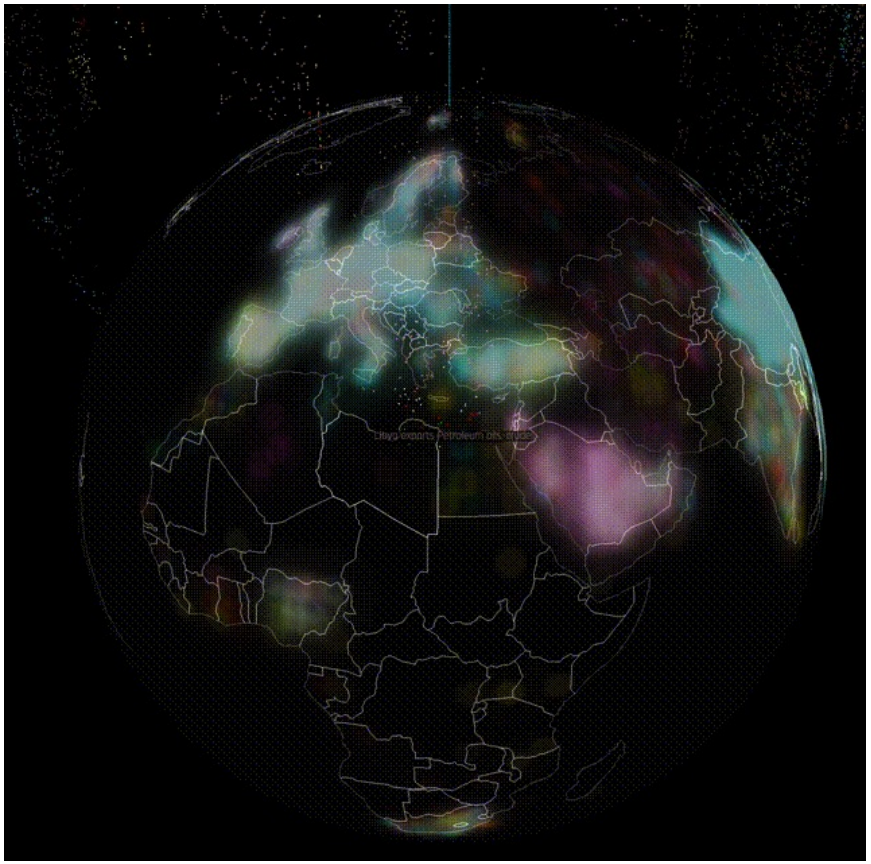
- Dot-based maps: early days / paper



Multivariate graphics for geographical + numeric variables

- Dot-based maps: nowadays / computer

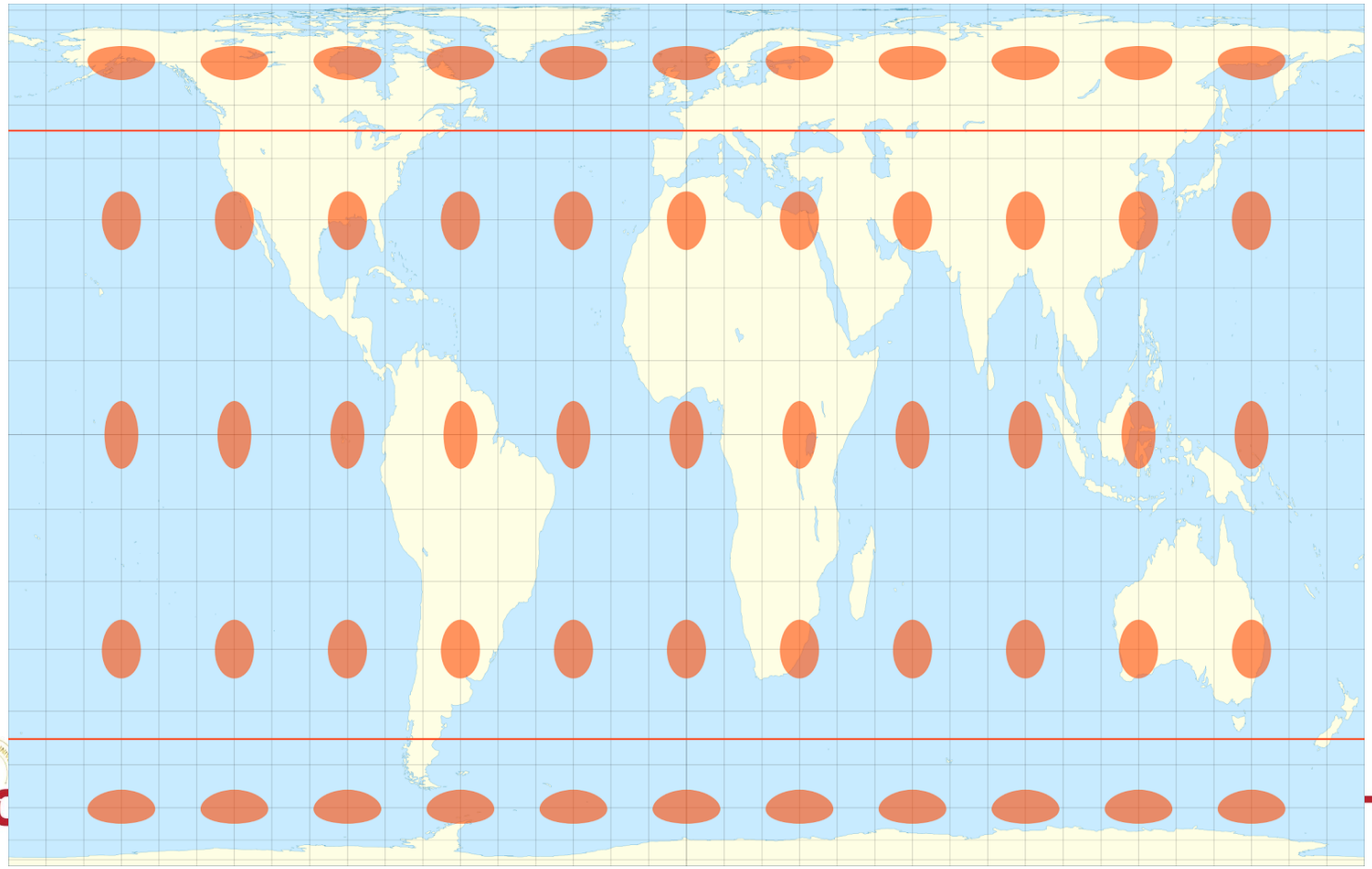




http://globe.cid.harvard.edu/

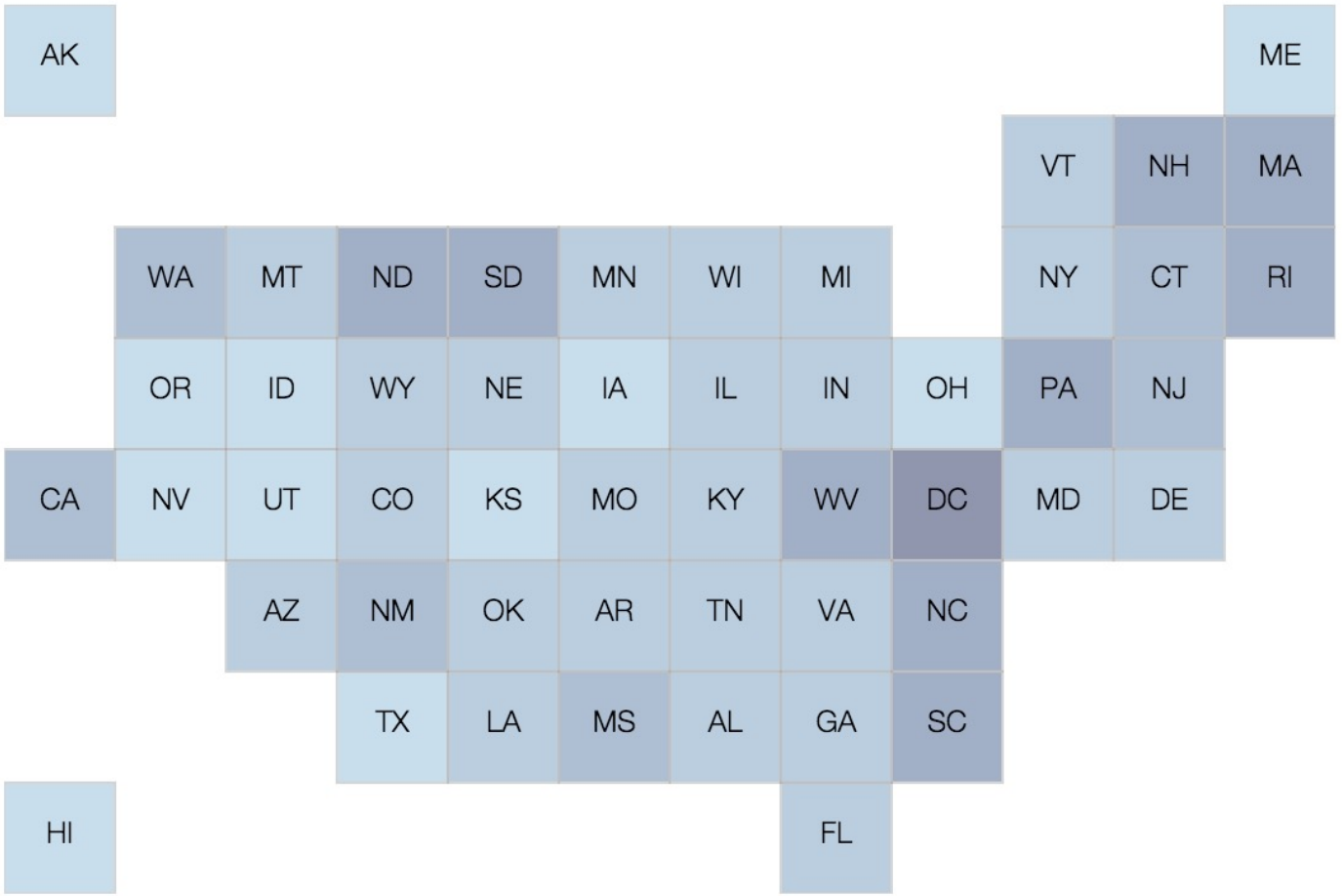
Comments on space visualization

- Issues with projections
  - Peter's projection



Comments on space visualization

- Grid-based projections

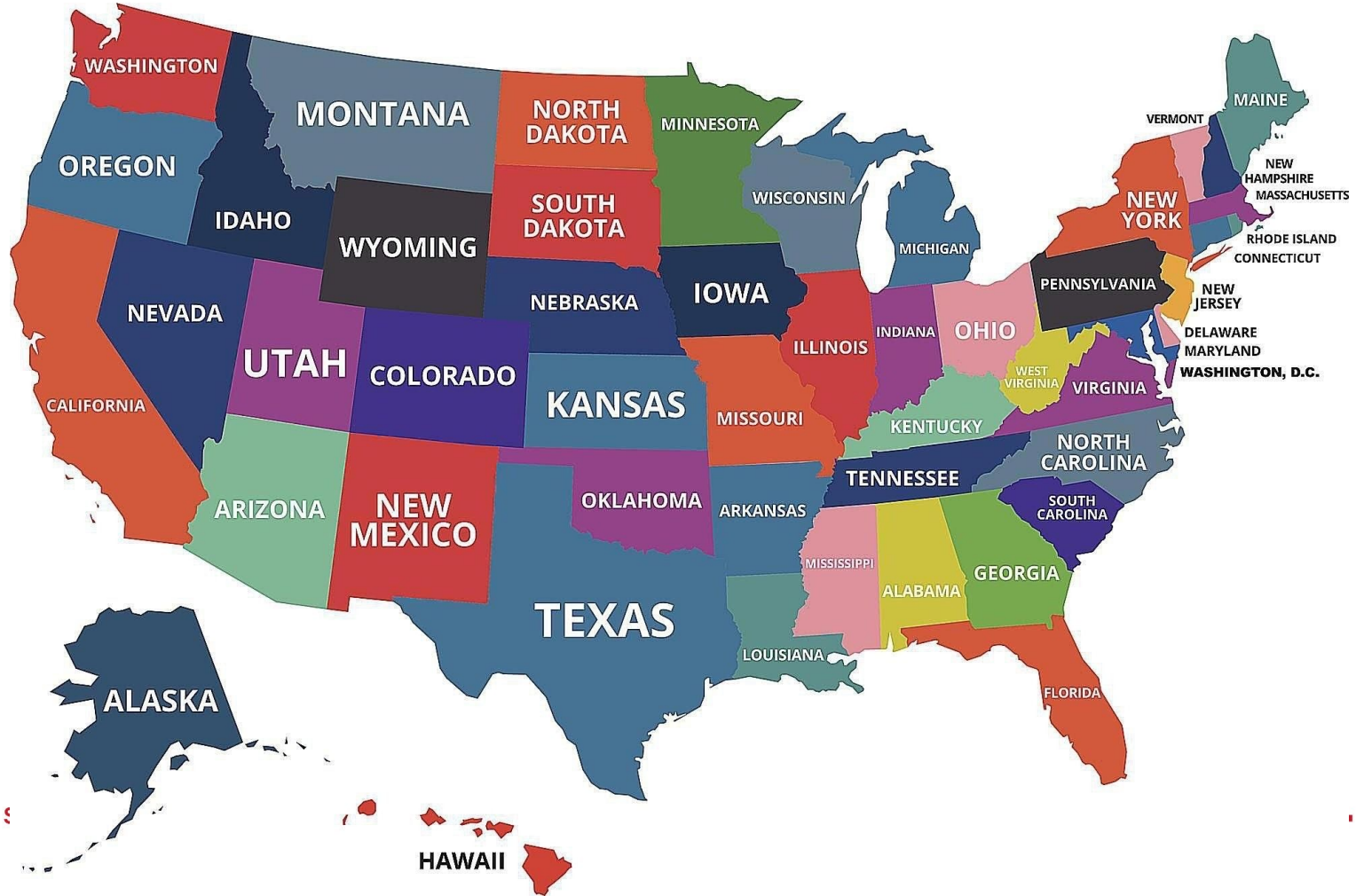


Comments on space visualization

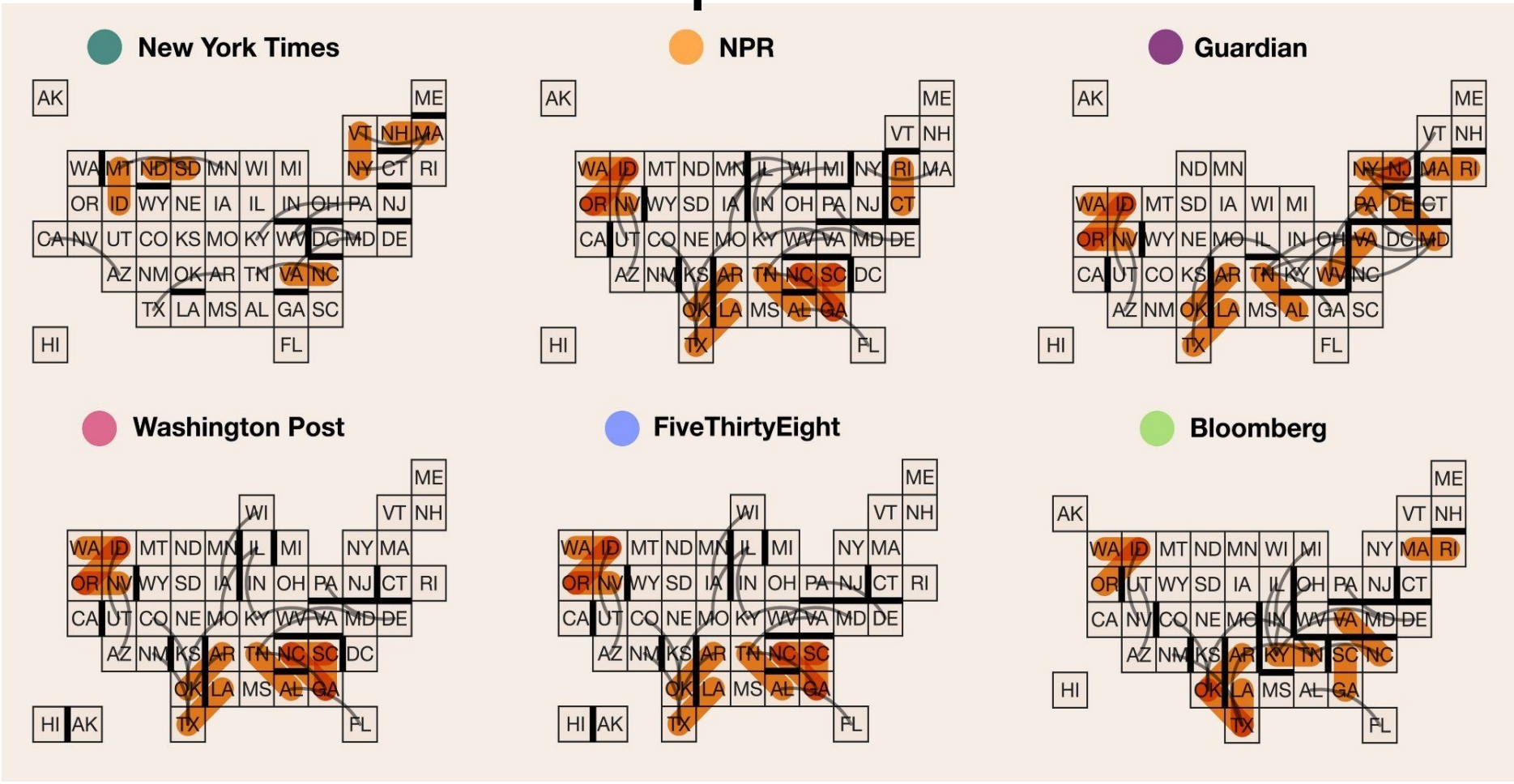
- Grid-based projections

Comments on space visualization

- But often, defining the grid is not so easy!!!
  - Can you imagine having to cut Hoan Kiem district into a grid???



Comments on space visualization



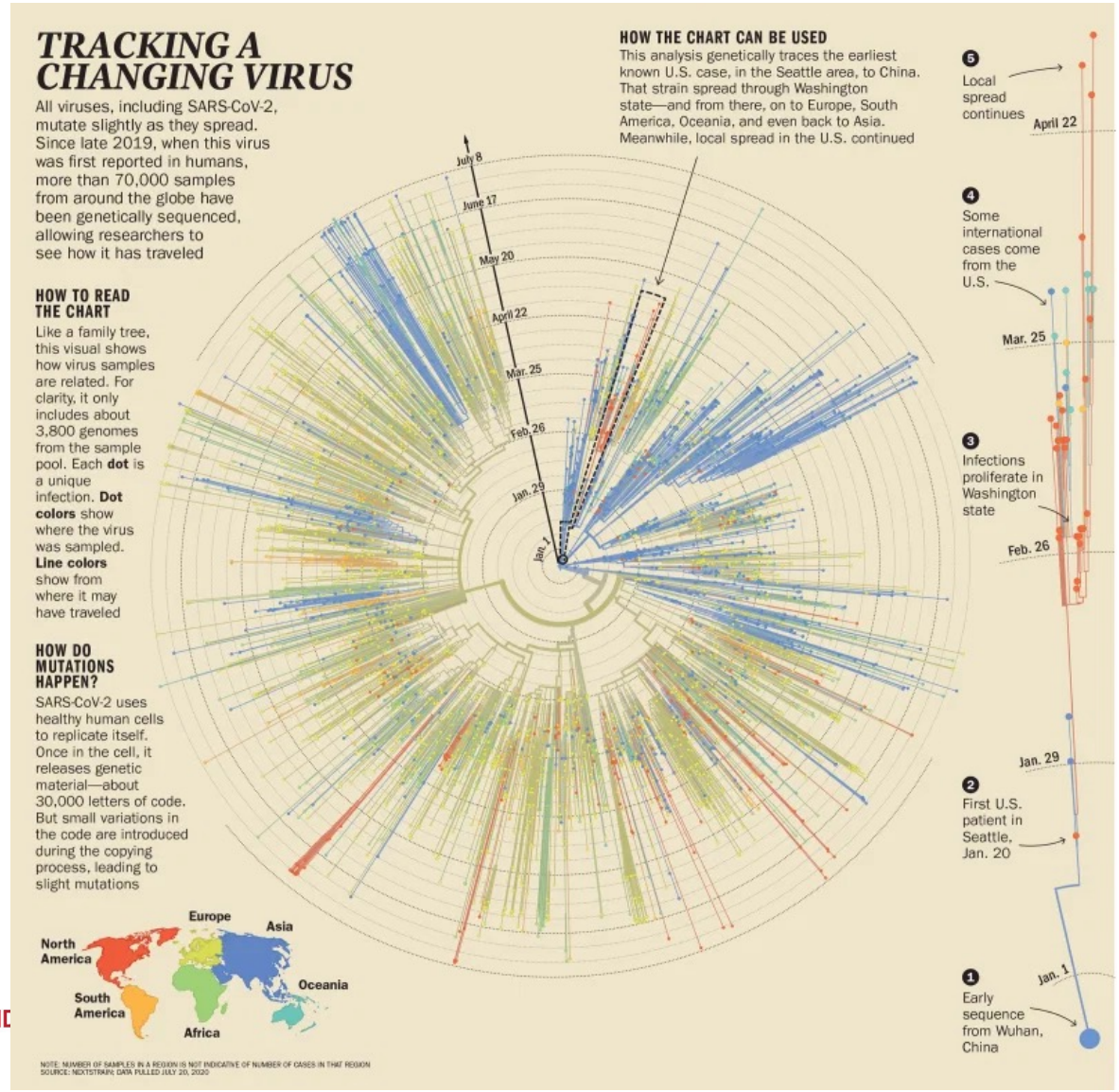
Different US map layouts from six publishers.
Black border = invalid neighbors, Thick orange line = misdirection, Curve line = missing neighbors.
https://medium.com/@kristw/whose-grid-map-is-better-quality-metrics-for-grid-map-layouts-e3d6075d9e80

Possible data visualization charts

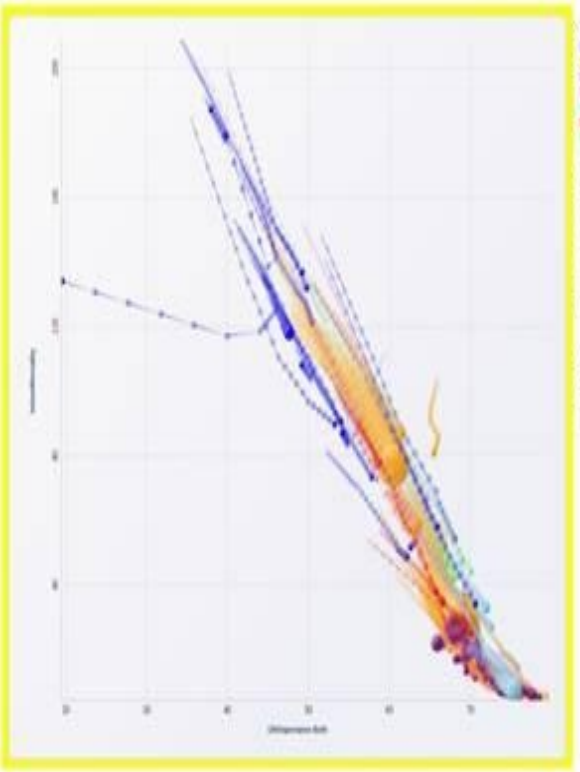
Special case of time / space visualization

Time / space visualizations

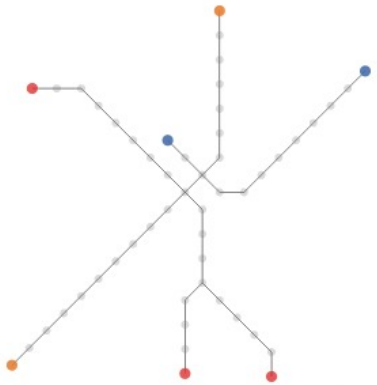
- COVID-19
  - Source:
    - Time magazine



Time / space visualization using animations (1/2)



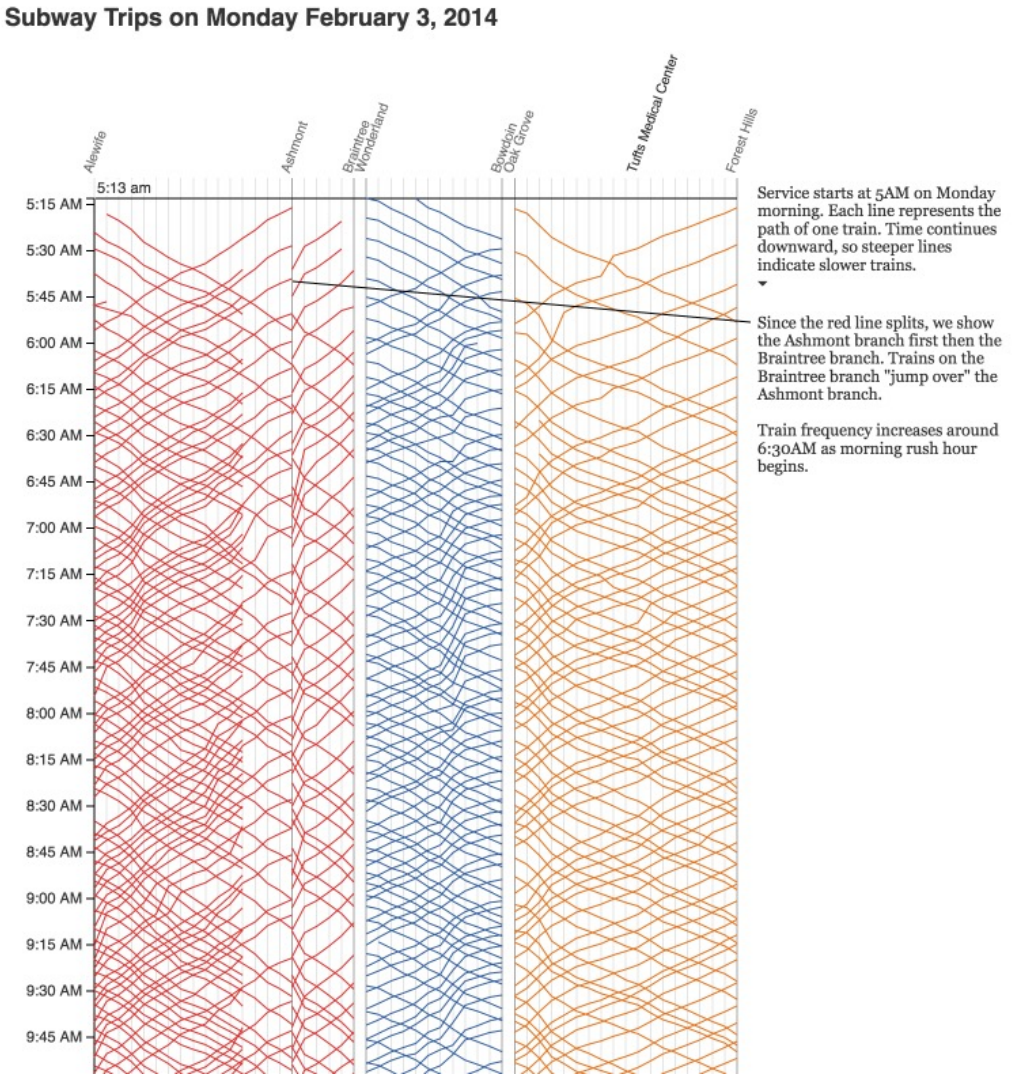
Time / space visualization using animations (2/2)



Locations of each train on the red, blue, and orange lines at 5:13 am. Hover over the diagram to the right to display trains at a different time.

Trains are on the right side of the track relative to the direction they are moving.

See the morning rush-hour, midday lull, afternoon rush-hour, and the evening lull.

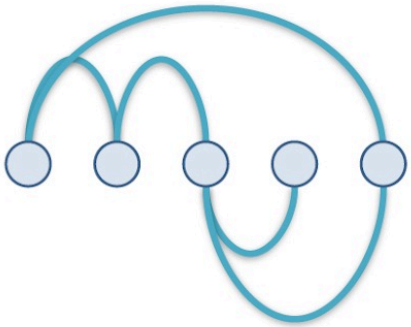


Examples of data visualization charts

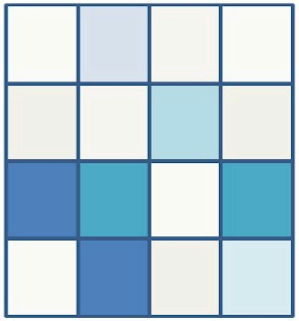
Tree and graphs / networks visualization

Tree and graph visualization

- Different types of tree and graphs visualization



Explicit
(Node-Link)

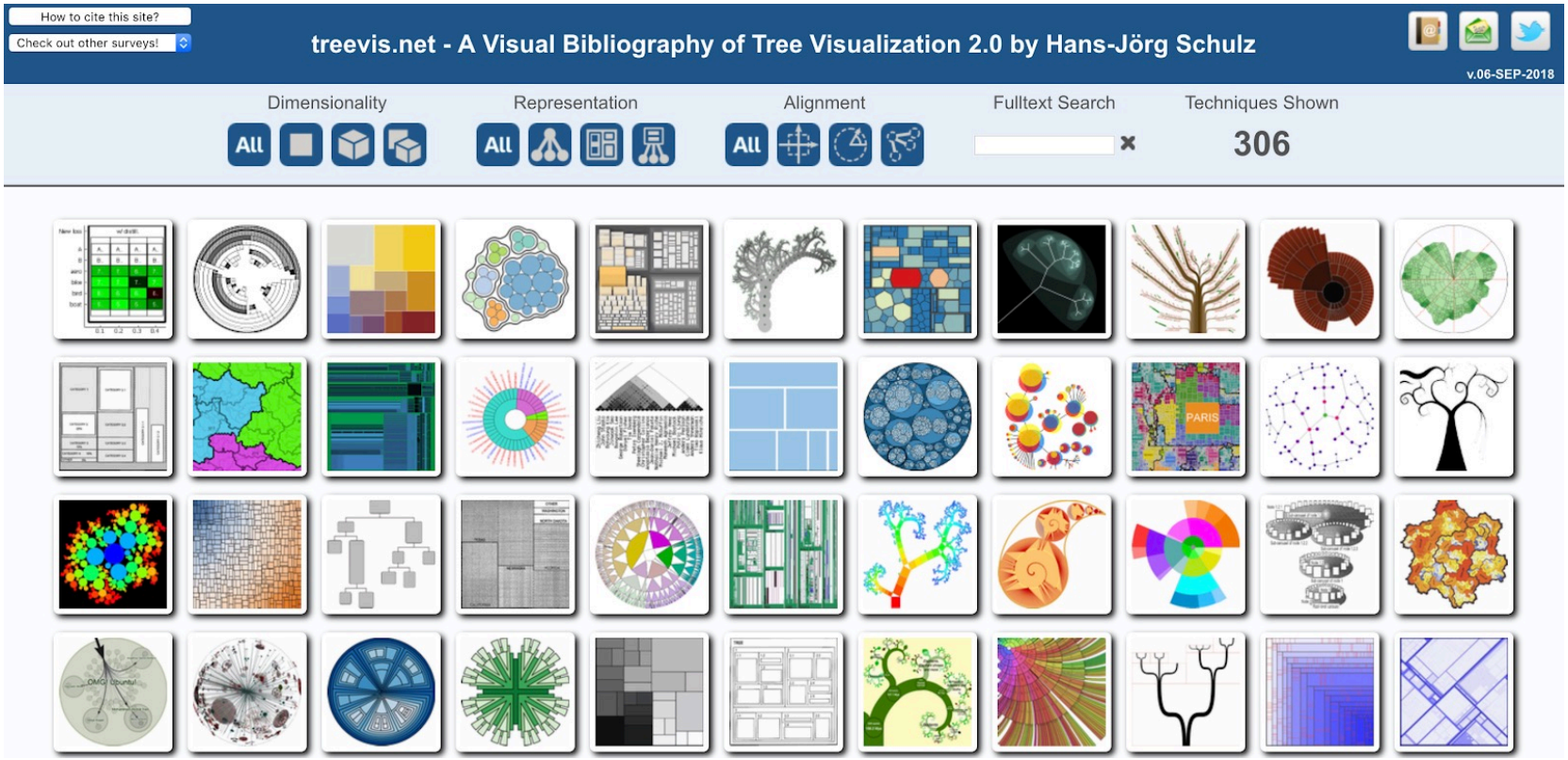


Matrix

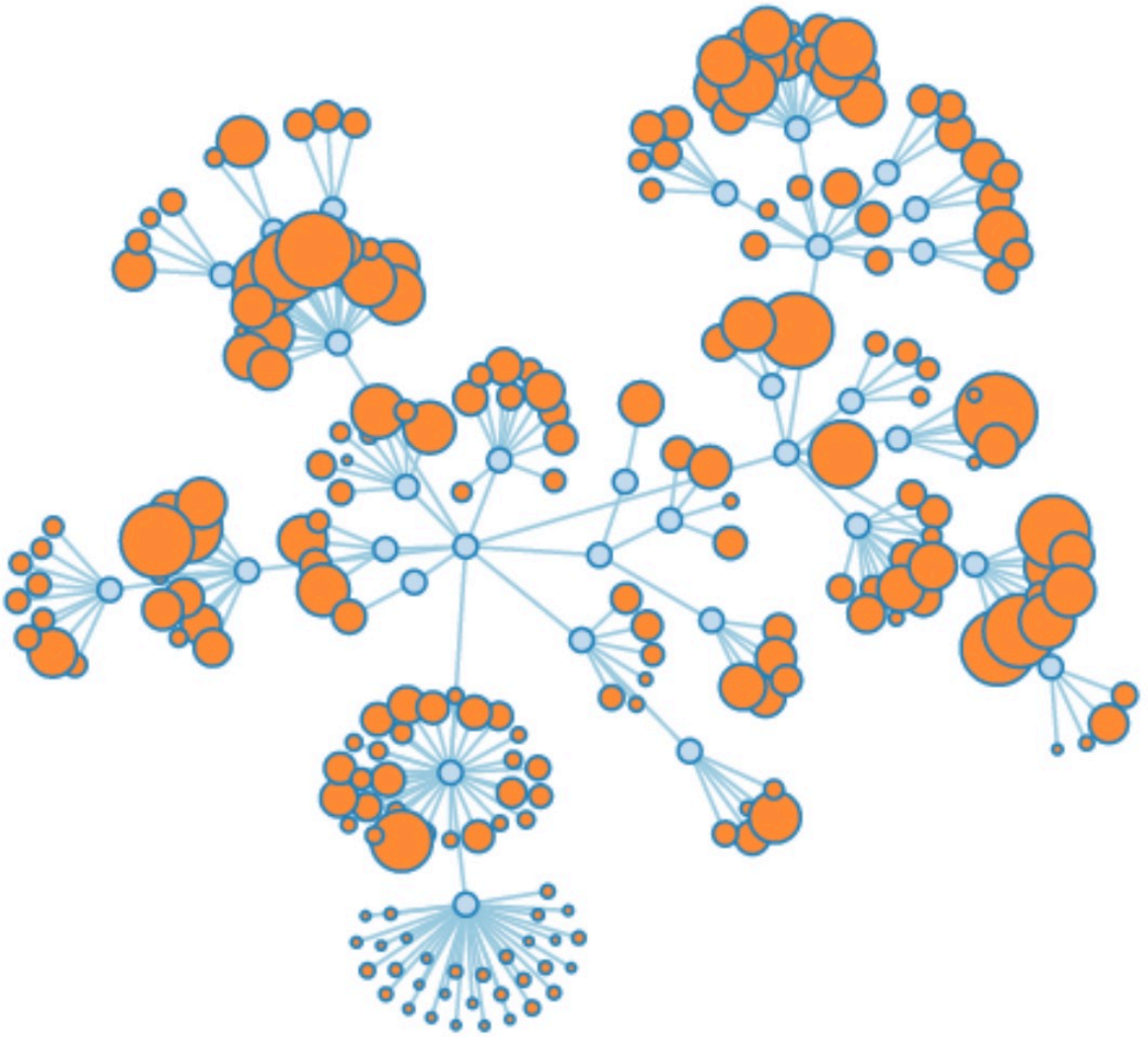
Implicit

Tree visualization

- TreeVis.net
  - Many possibilities: choose according to the audience!!!

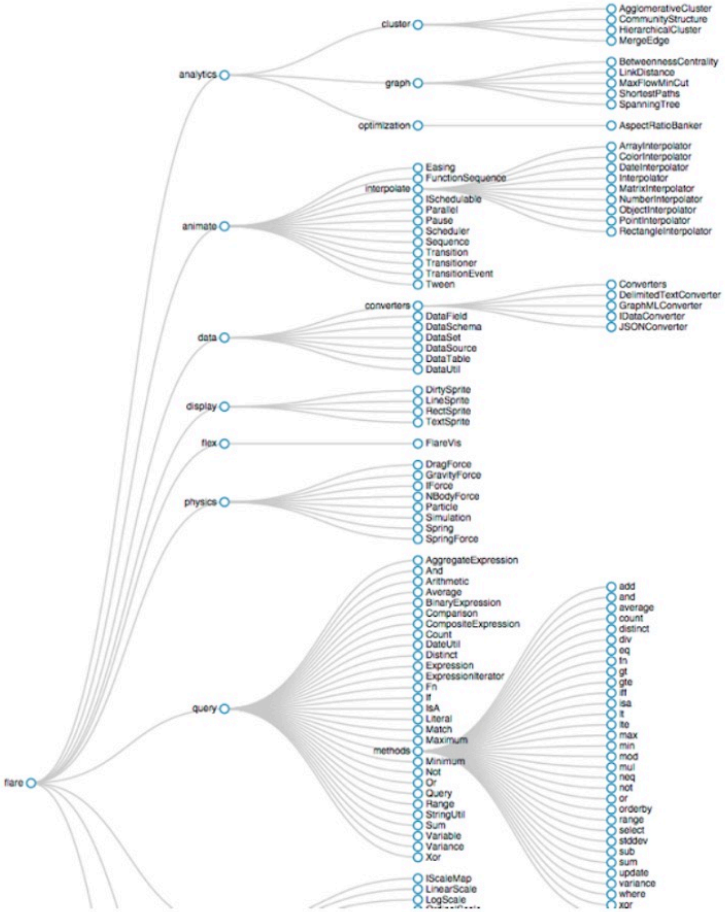


Explicit tree visualization

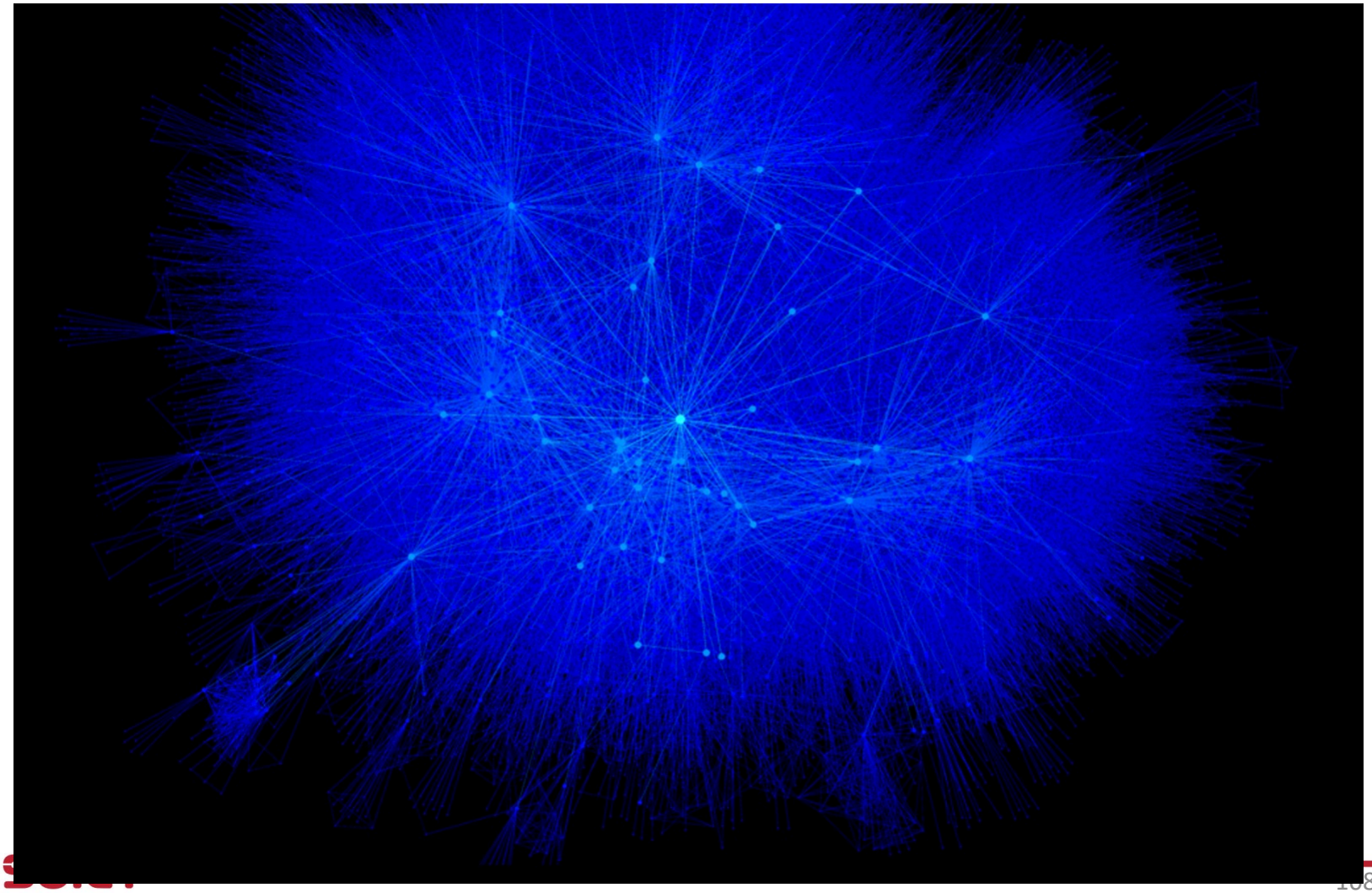


Explicit tree visualization

- Example with text: JSON thesaurus

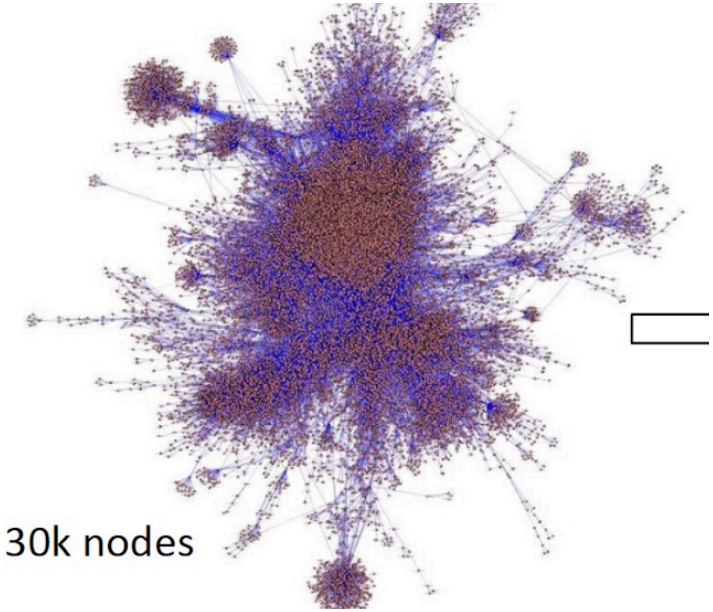


Explicit graph visualization



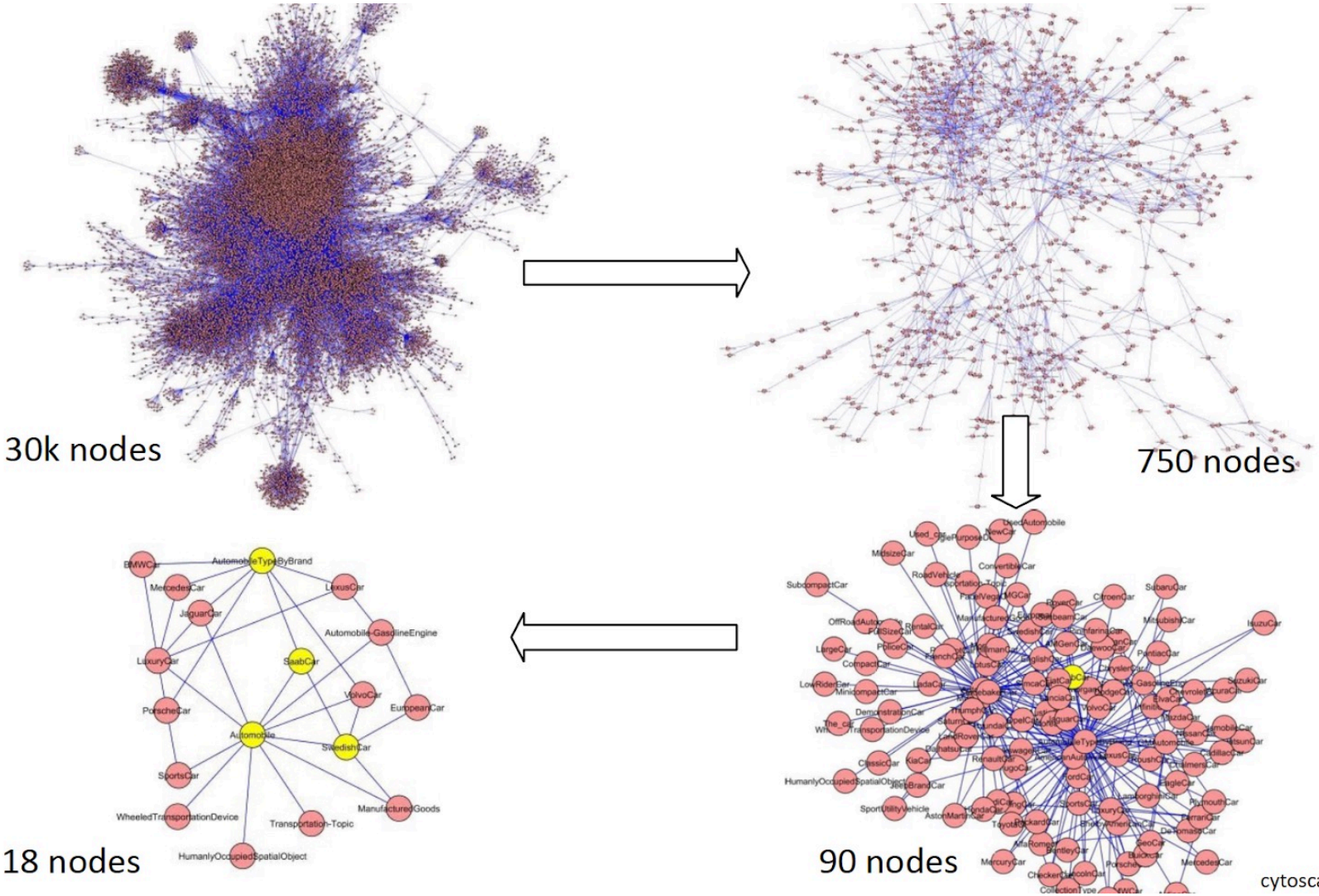
Difficulties with explicit graph visualizations

- Possibilities
  - Reducing the dimensions of the graph by creating “categories”
  - Using interactive visualizations
  - Using implicit visualizations



Difficulties with explicit graph visualizations

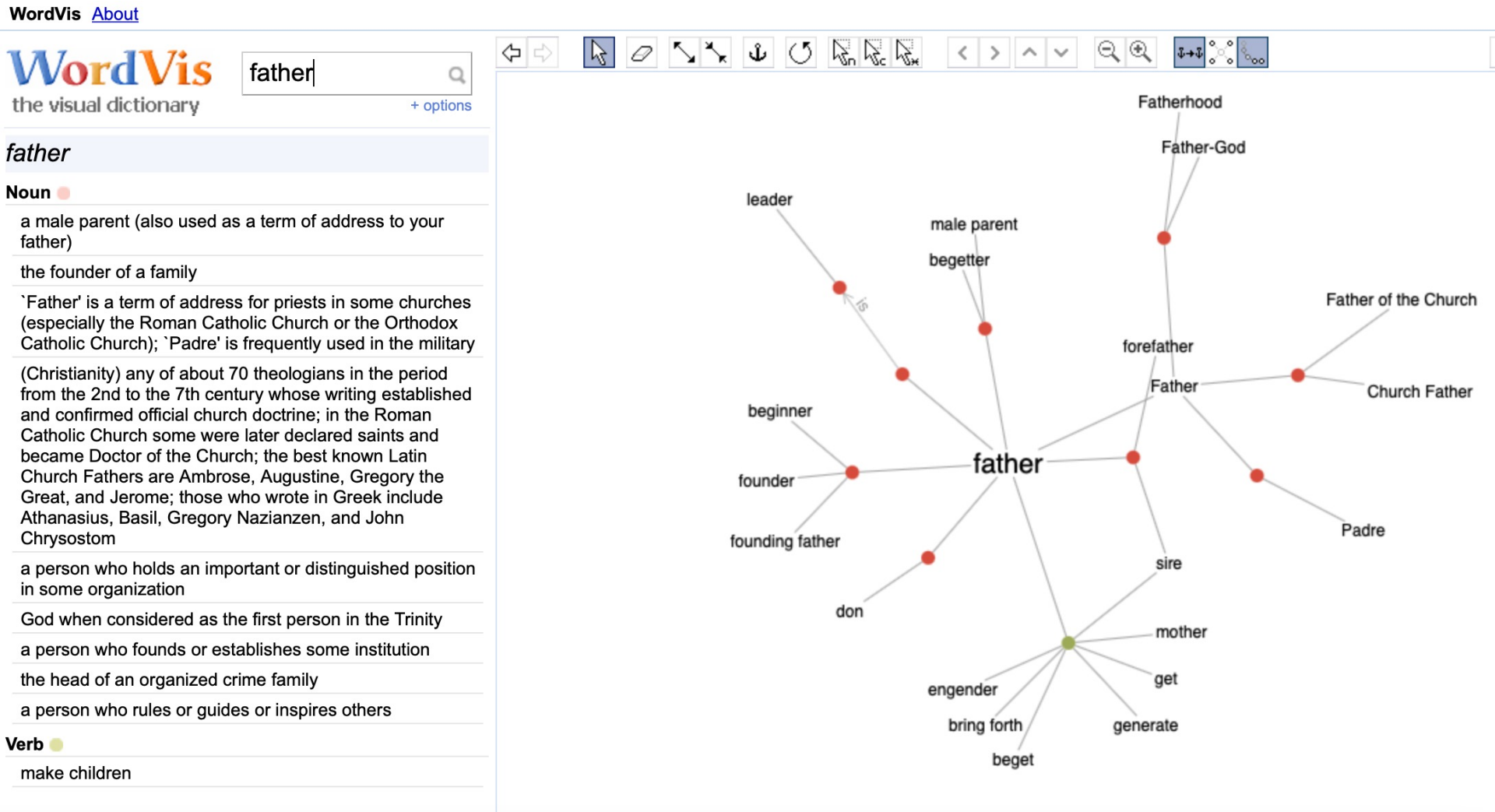
Reducing the dimensions of the graph by creating “categories”



Complex graph (ontology) explicit, interactive visualization

Using interactive visualizations

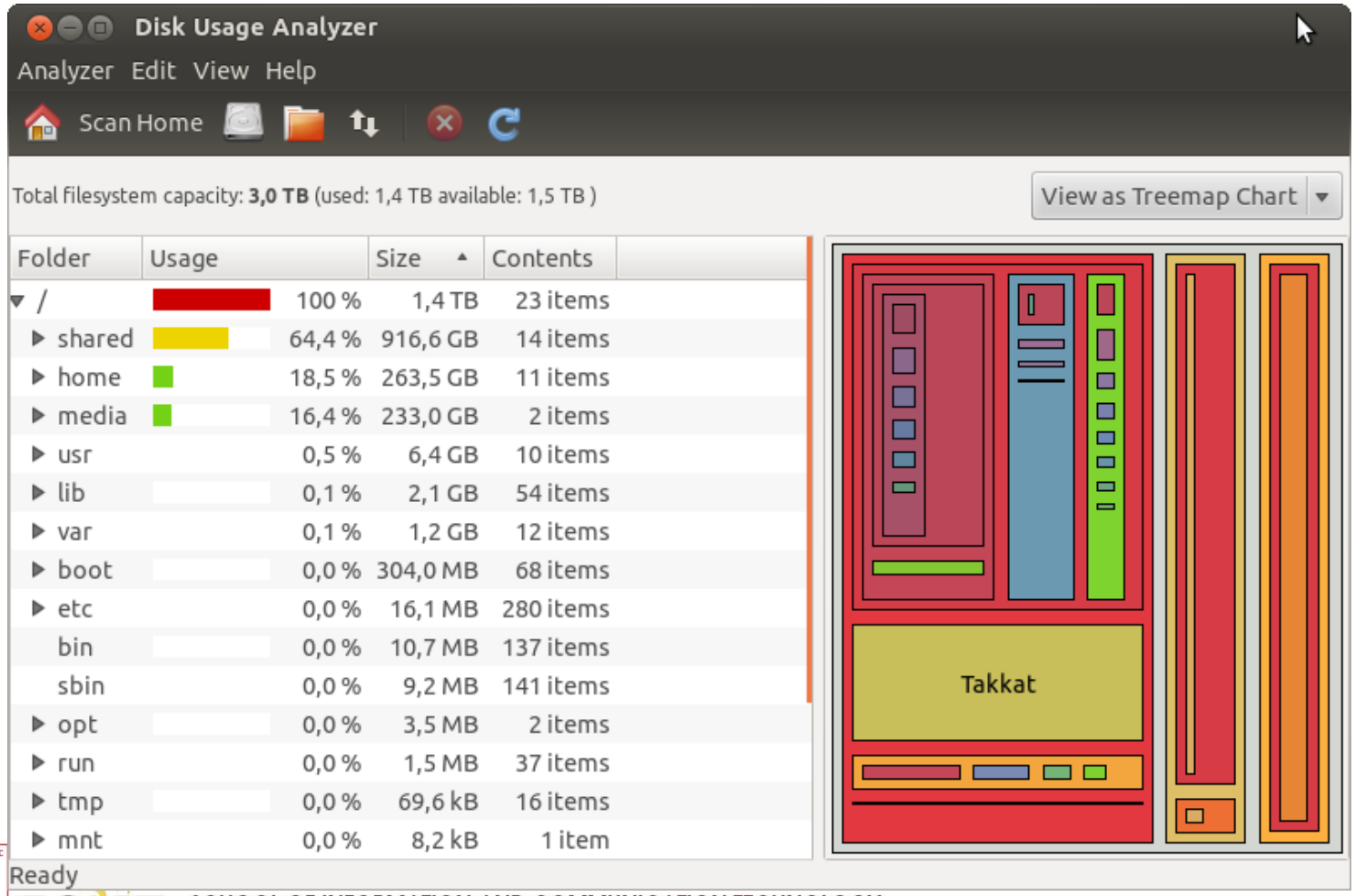
- http://wordvis.com



Implicit tree visualization

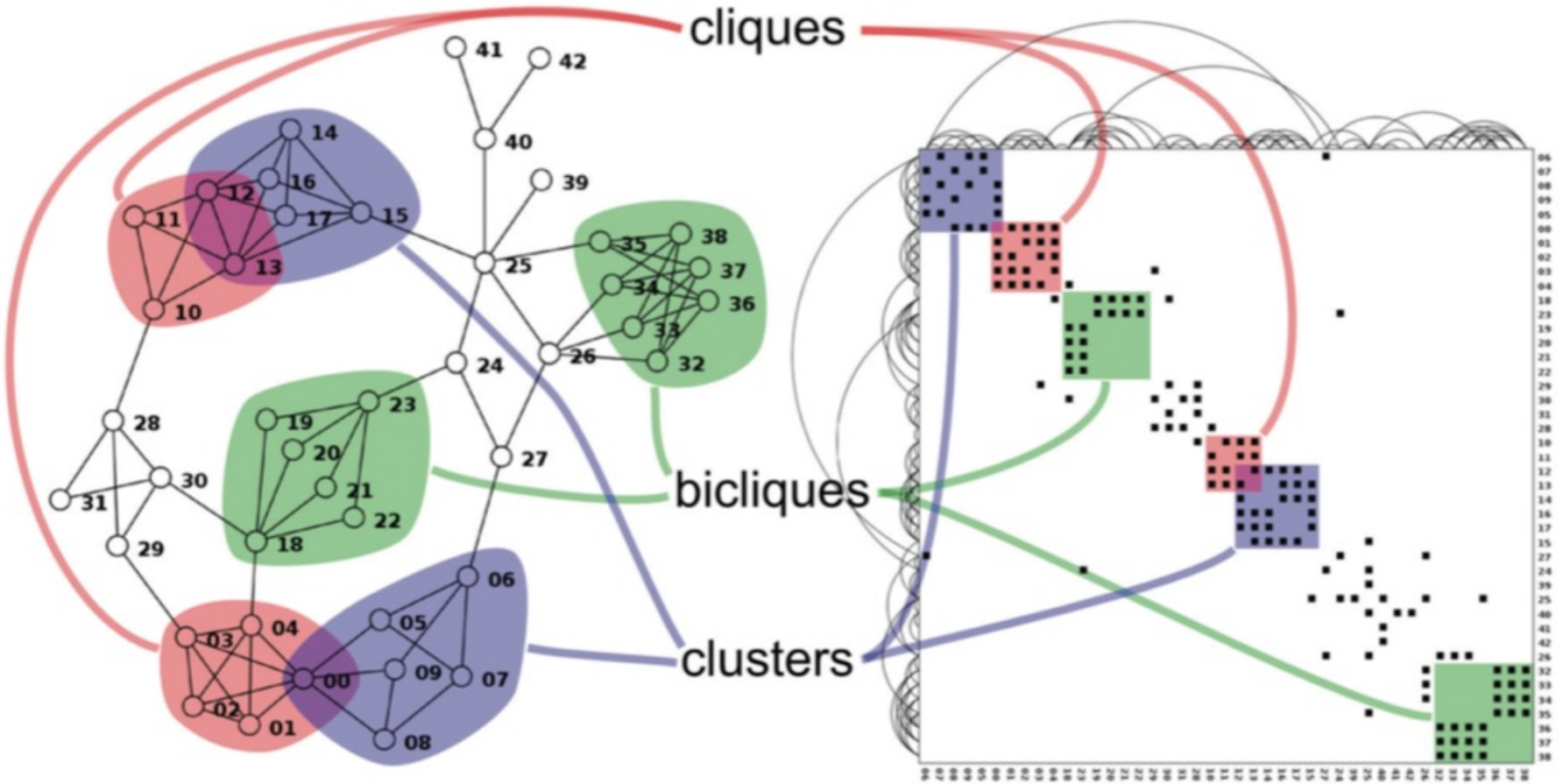
Using implicit visualizations

- Example: treemap



Implicit graph matrix visualization

Using implicit visualizations



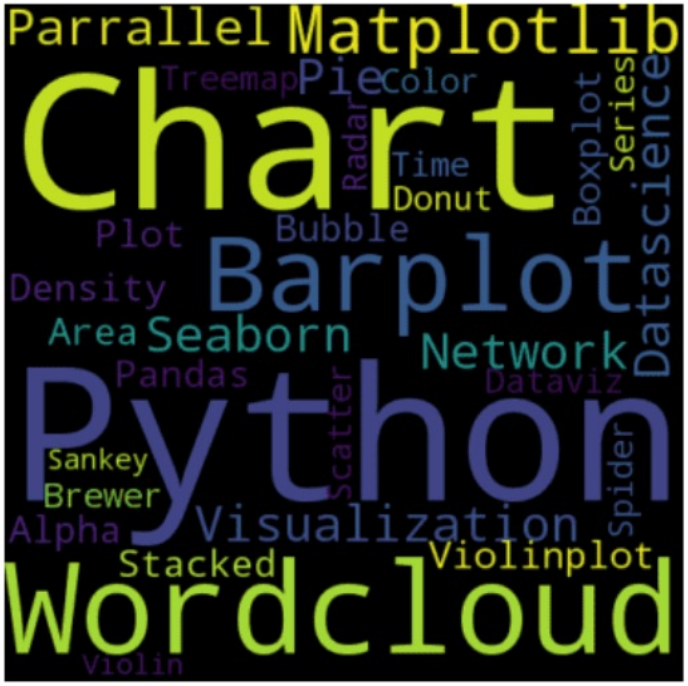
Examples of data visualization charts

Other special cases

Special case of text

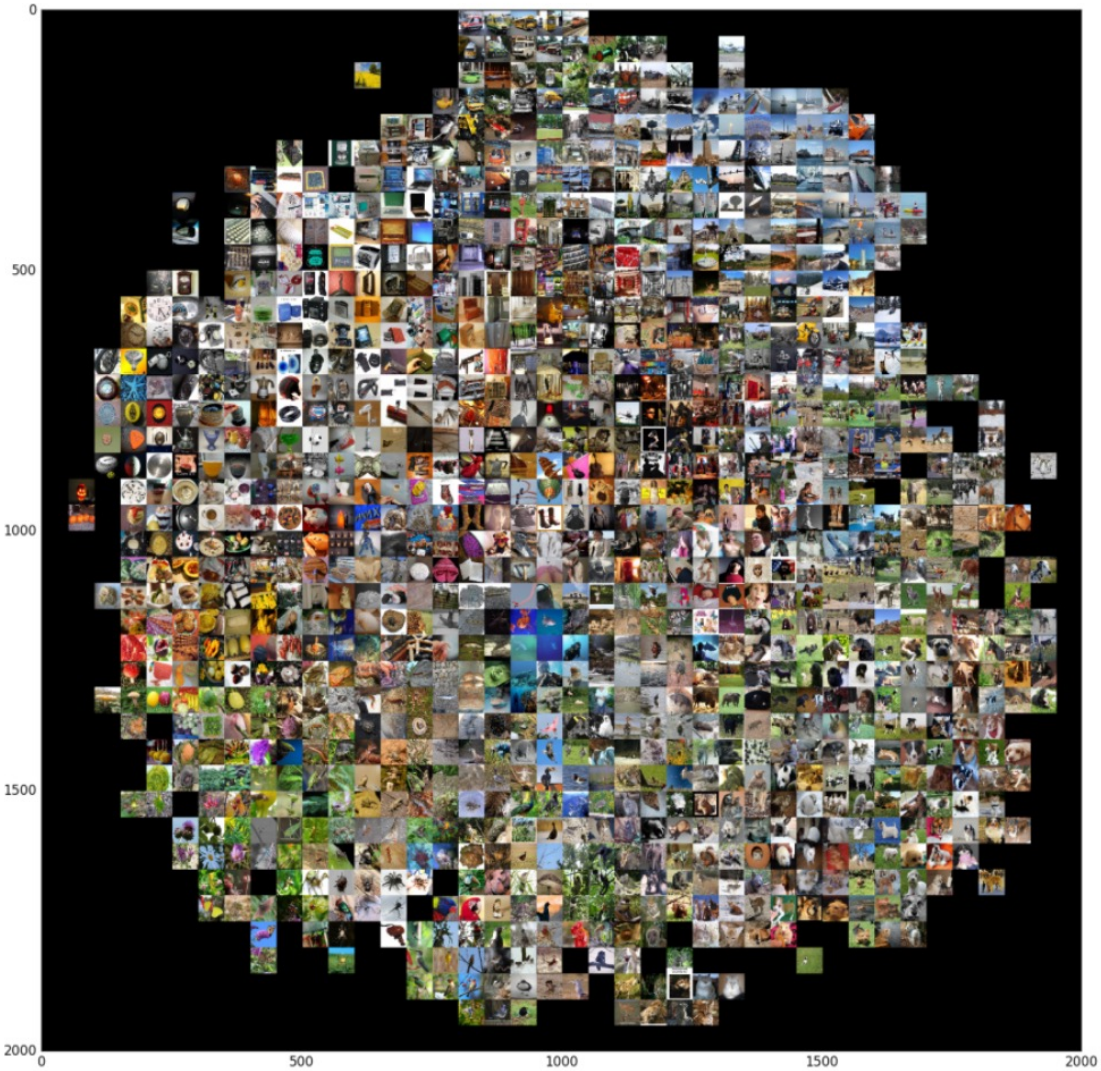
Words can be seen as categorical variables

Example of representation according to the word frequency: wordcloud

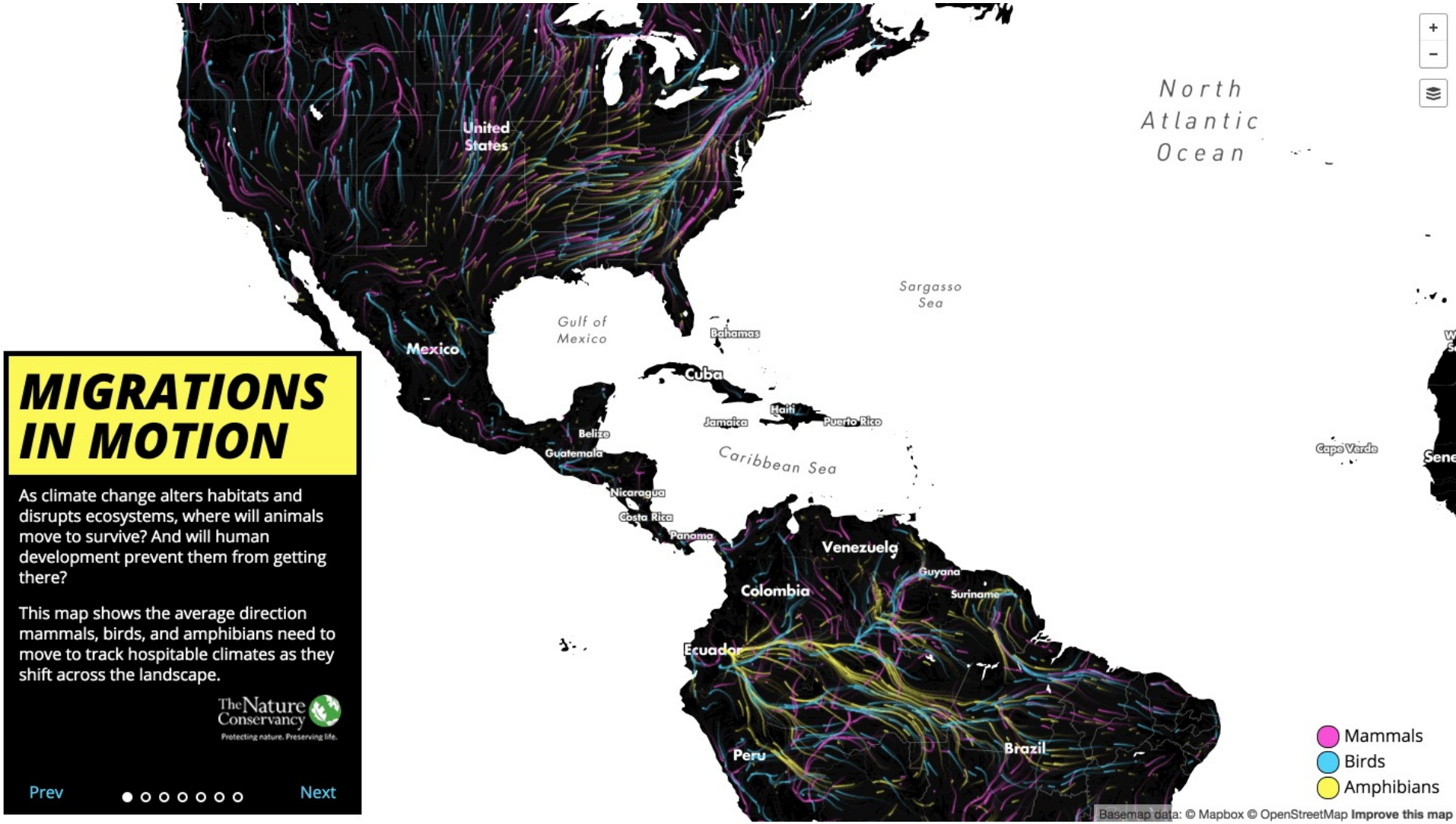


Special case of images

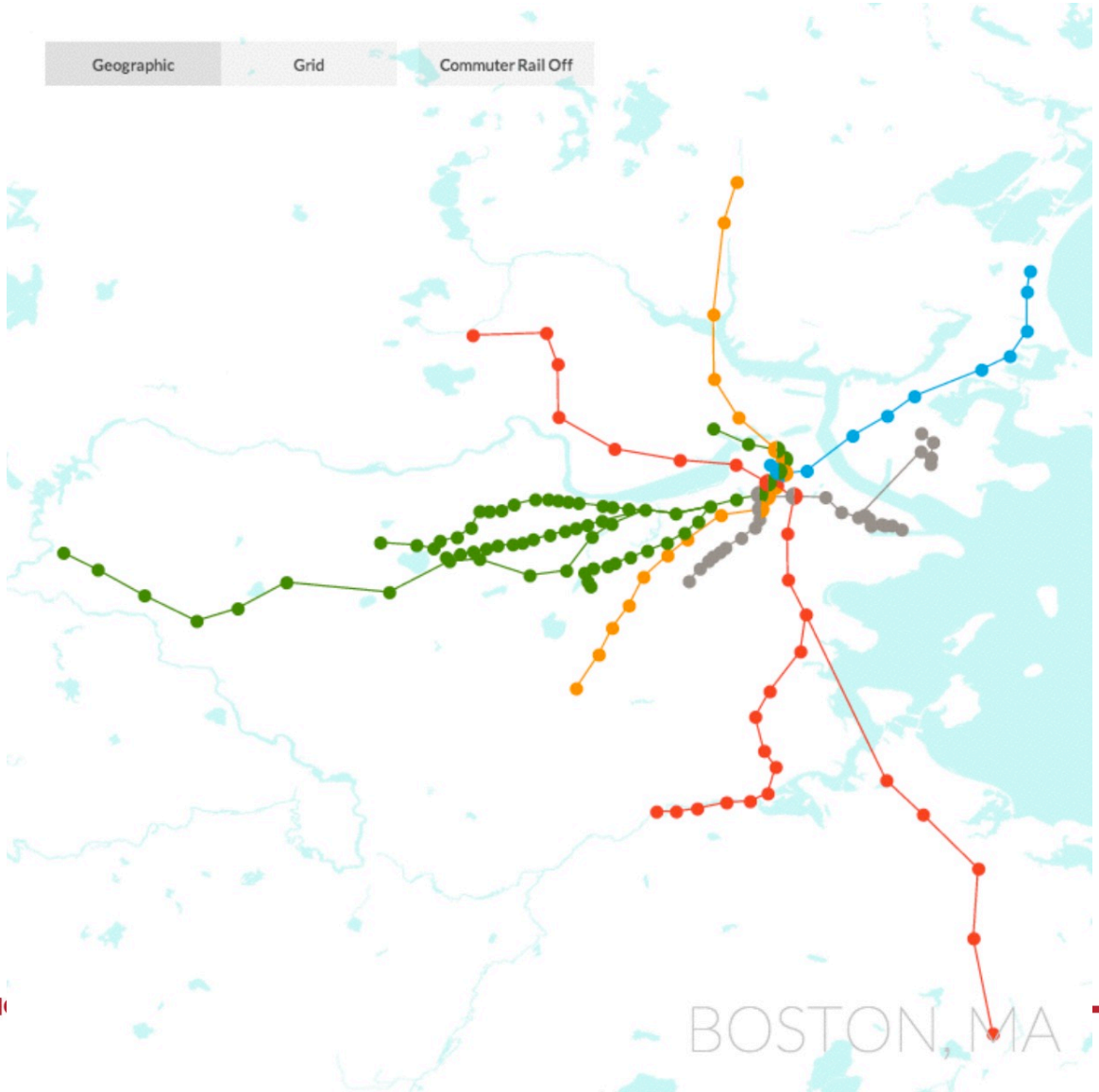
- Images can be seen as a 2D array of numeric variables (pixel values)
  - Usually, millions of variables!
  - **dimensionality reduction, then representation**
    - One can use PCA or tSNE for dim. reduction
    - Example: tSNE, 5000 images



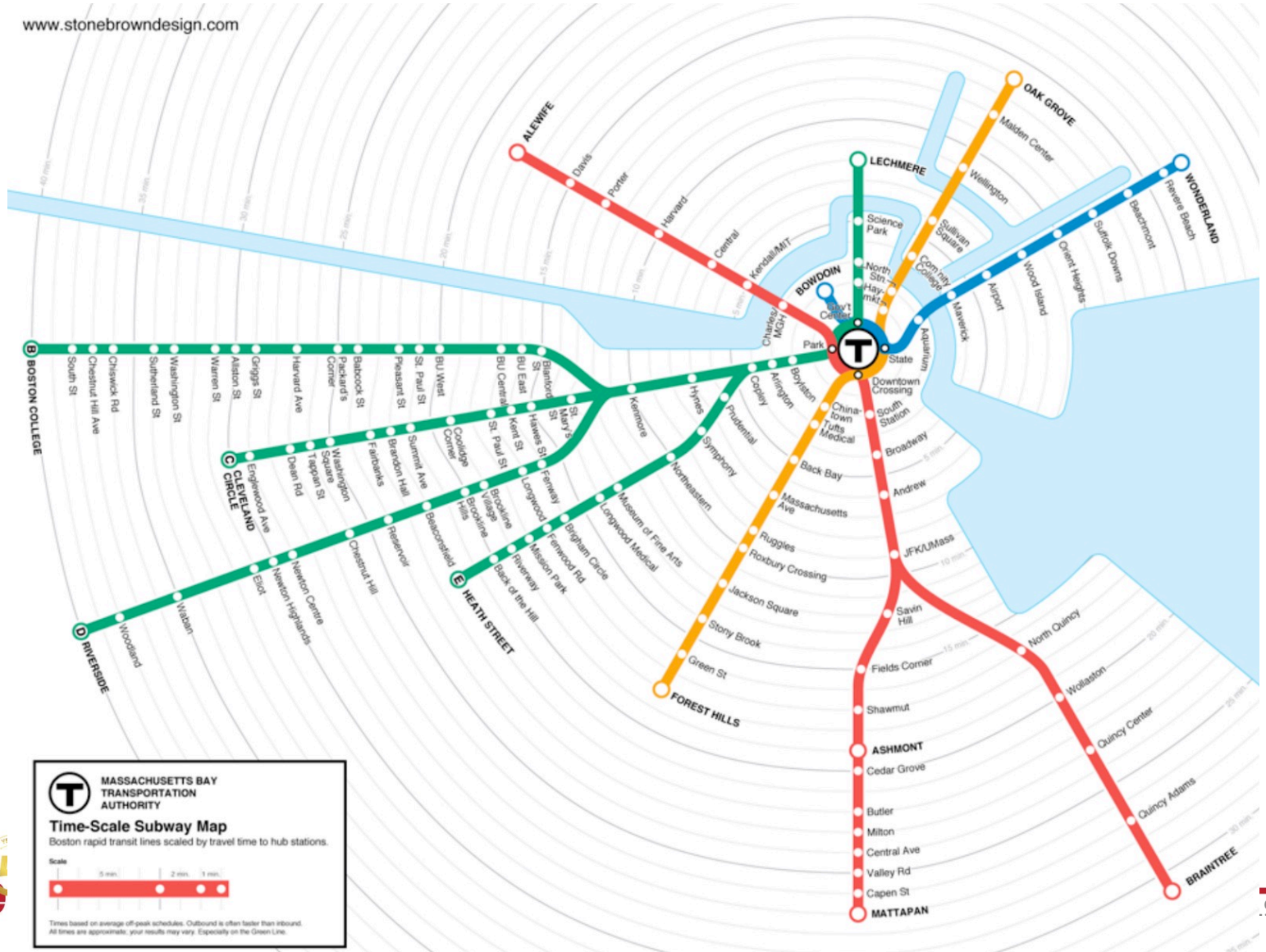
Space + trajectory / flow visualizations



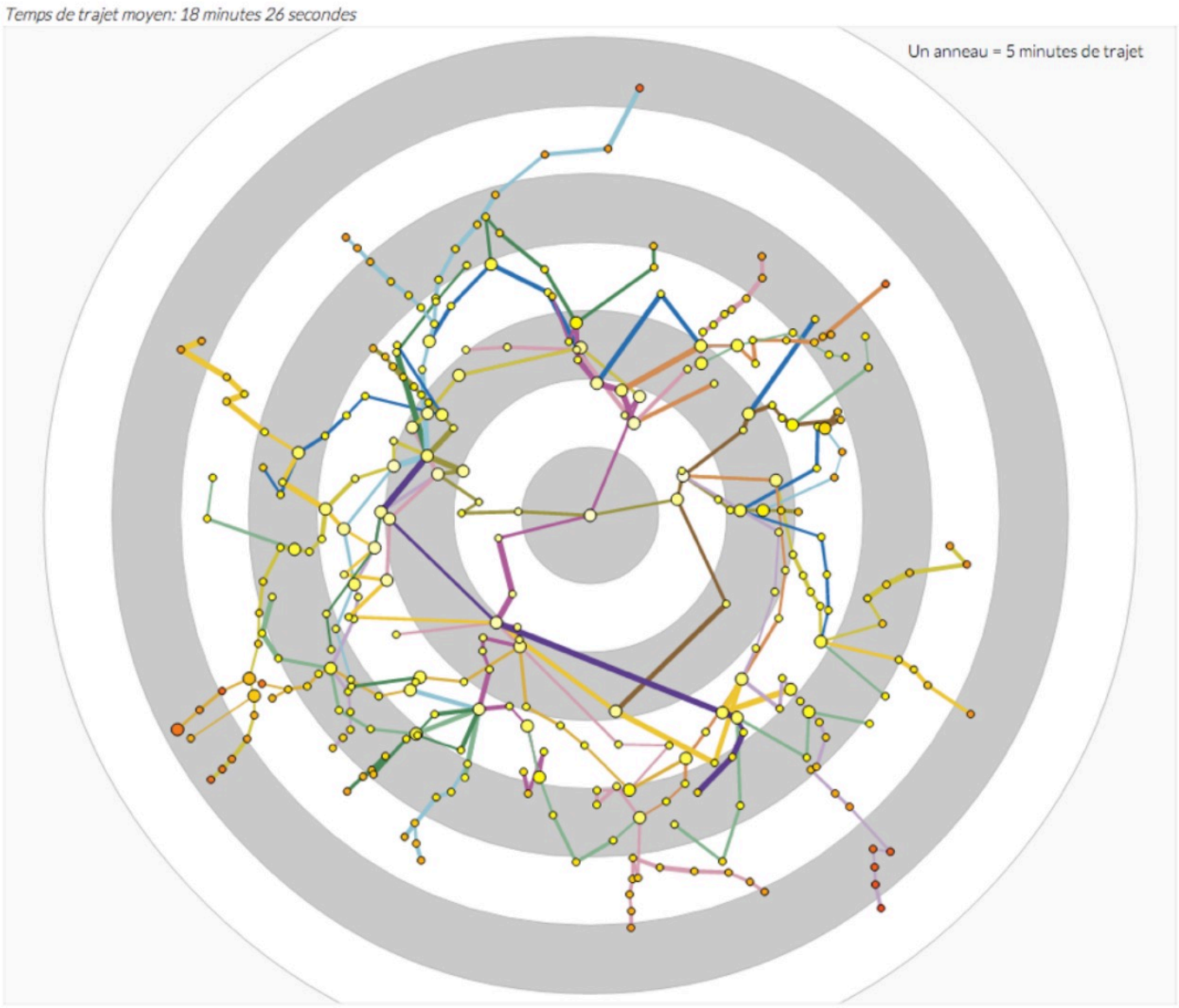
Space + network visualizations



Space + network + time visualizations



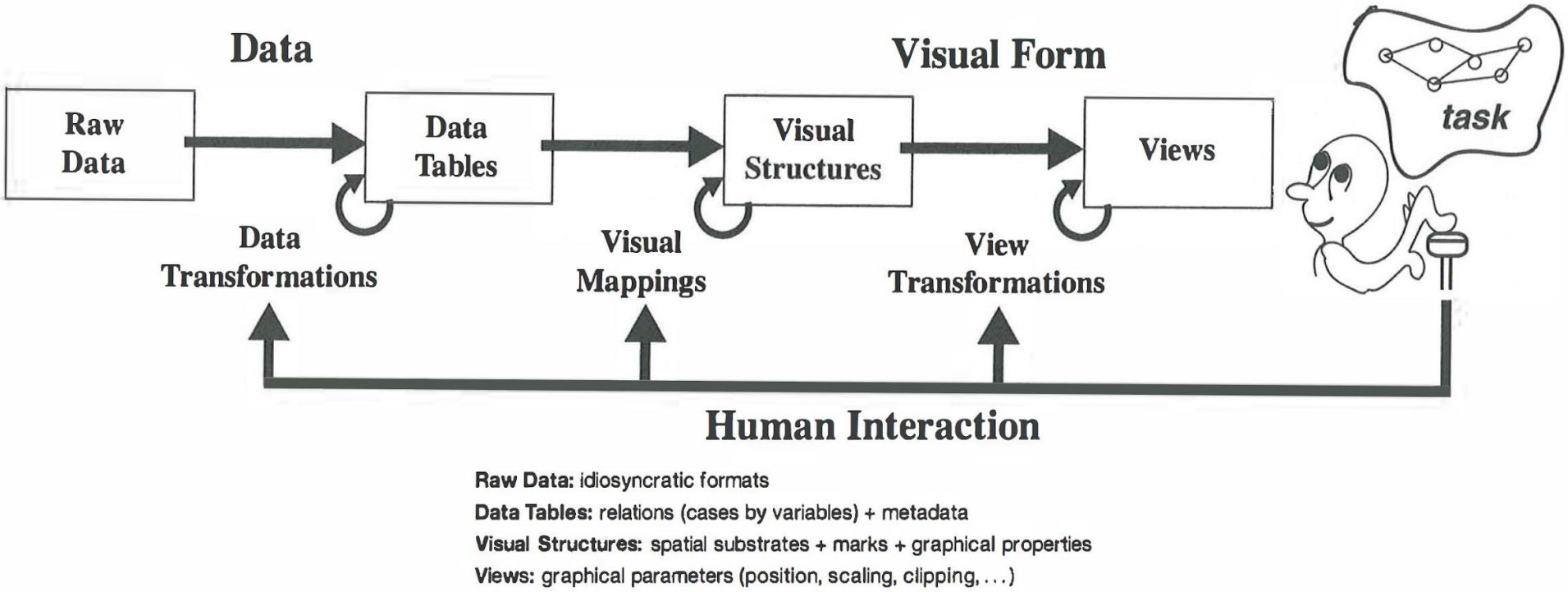
Space + network + time visualizations



Technical tools for data visualization

What do data visualization tools do?

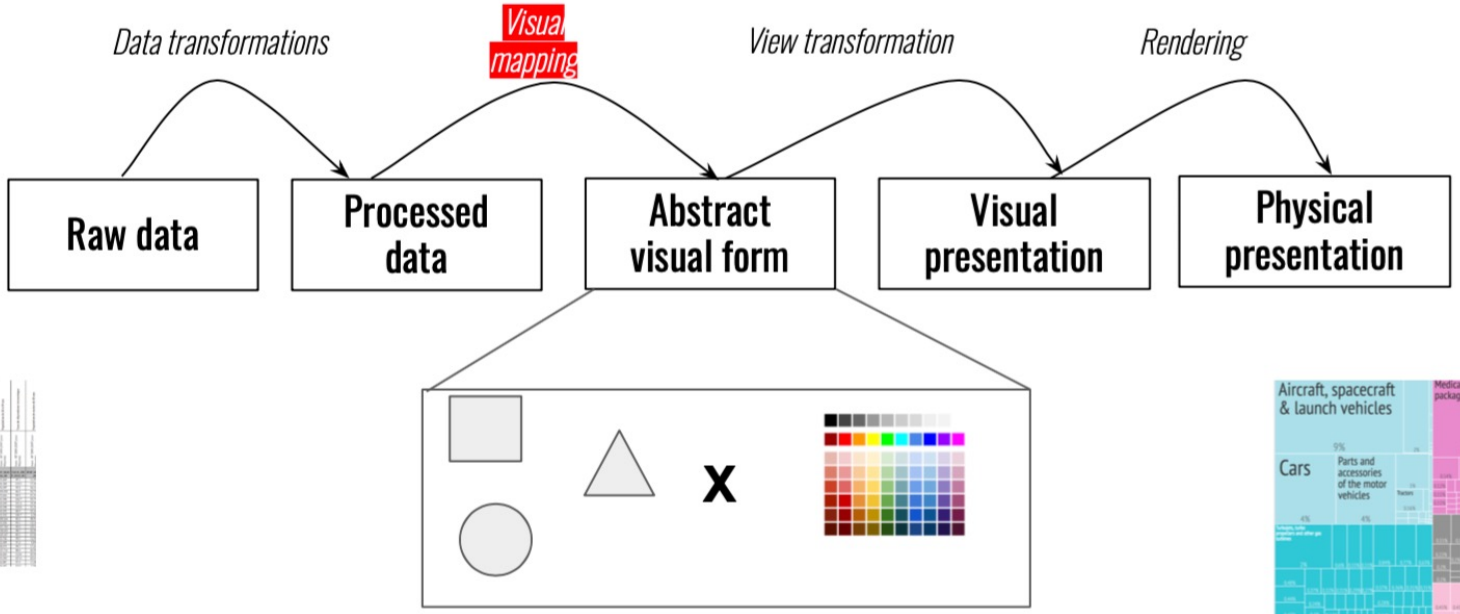
- Usually, all these steps are performed end-to-end by tools / libraries

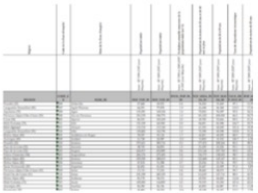


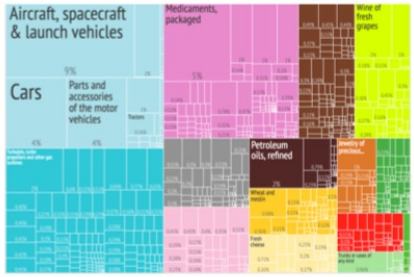
[Card, Mackinlay, Shneiderman, Readings in Information Visualization: Using Vision to Think, 1999]

What do data visualization tools do?

- Usually, all these steps are performed end-to-end by tools / libraries
  - You might still have to pre-process the data before feeding it to the data visualization tool

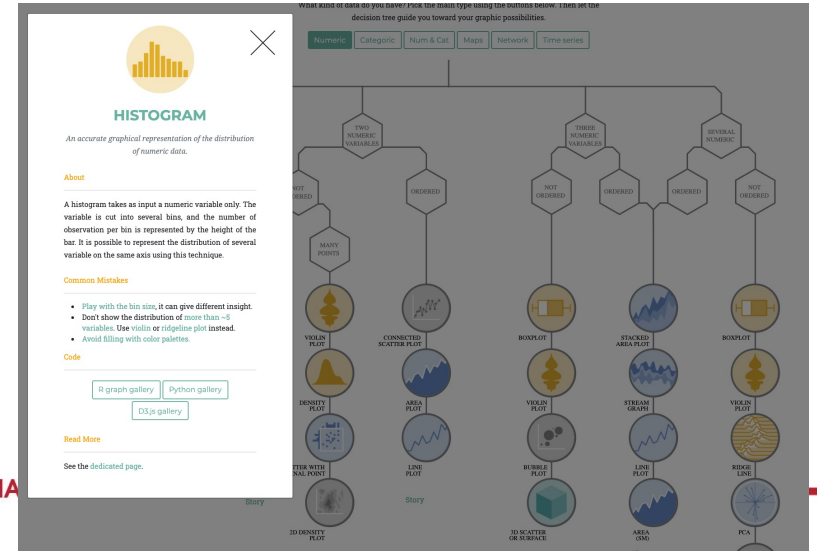






Tools for data visualization

- There are A LOT of tools for data visualization
  - For example, TreeVis for tree visualizations, Roassal for graphs
- There are Python libraries for almost all kinds of visualization (except maybe for the most refined / interactive)
  - Matplotlib for the most basic graphics
  - For the more refined graphics, find the corresponding library on
    - https://www.data-to-viz.com/ -> Python gallery



Tools for data visualization

- Most useful Python libraries for data visualization:
  - Matplotlib
  - Seaborn
  - Plotnine(ggplot)
  - Bokeh
  - pygal
  - Plotly
  - geoplotlib
  - Gleam
  - missingno
  - Leather
  - Altair
  - Folium
- More complete list on: https://mode.com/blog/python-data-visualization-libraries/

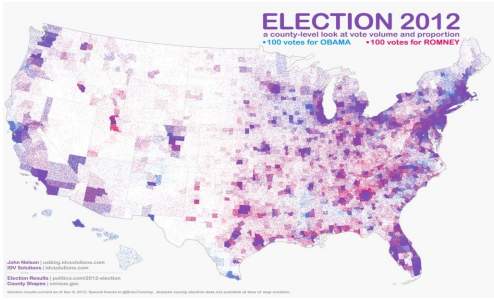
Summary

Summary

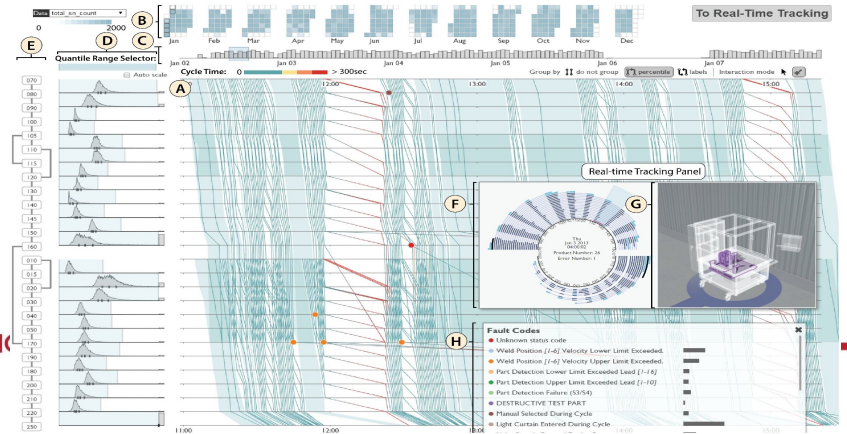
- Data visualization is a **big part** of your future work as a data scientist
  - And, a big part of your capstone project ☺
- The difficulty is that, choosing the appropriate chart(s) is not always easy…
  - Depends on the type(s) of data
  - Depends on what you want to show
  - Depends on your audience

Summary

- After this lecture, you should not make beginner's mistakes any more
  - Example of beginner's mistakes:
    - Use a scatter plot for categorical, nominal variables encoded by numbers
    - Use a vertical bar chart for numeric variables
    - Use visuals that are not easily perceptible by humans



    - Use visuals that are not easily understandable by your audience



In this lecture, you've seen (1/2)

- Different types of visualization (depending on the objective, audience and data)
  - Infographics
  - Story-telling
  - Cartography
  - Information Visualization
  - Visual analytics
  - Scientific visualization

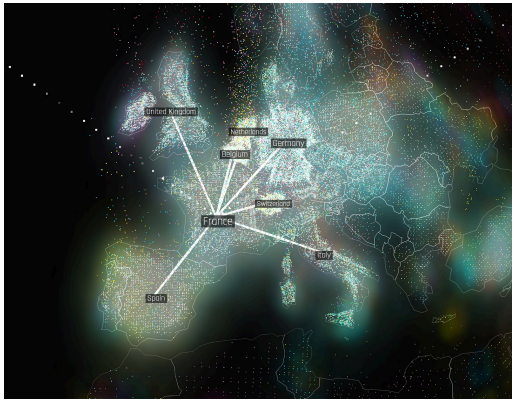
In this lecture, you've seen (2/2)

- Examples of charts / ideas of original charts to use for your future communications
  - Numeric variables
  - Categorical variables
  - Mix of numeric variables + categorical variables
  - Special case of time-dependent variables
  - Special case of geographical data
  - Special case of time-space visualization
  - Tree and graph / network visualization
  - Other special cases
- Technical tools for data visualization: Python libraries

Homework

For next lecture, you'll have to

- Study the charts I gave you in this lecture
  - Visit the source website from which I took the visualization
    - So as to better understand the information in the chart



http://globe.cid.harvard.edu/

- Study the charts presented in https://www.data-to-viz.com/
- Study the charts presented in https://datavizcatalogue.com
- Write down your questions, and ask me during next lecture
  - Only questions about the theory, for programming details you should be autonomous

Questions

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!