



PRINCIPLES OF DATA SCIENCE

JOHN P DICKERSON

Lecture #10 – 10/31/2018

CMSC641

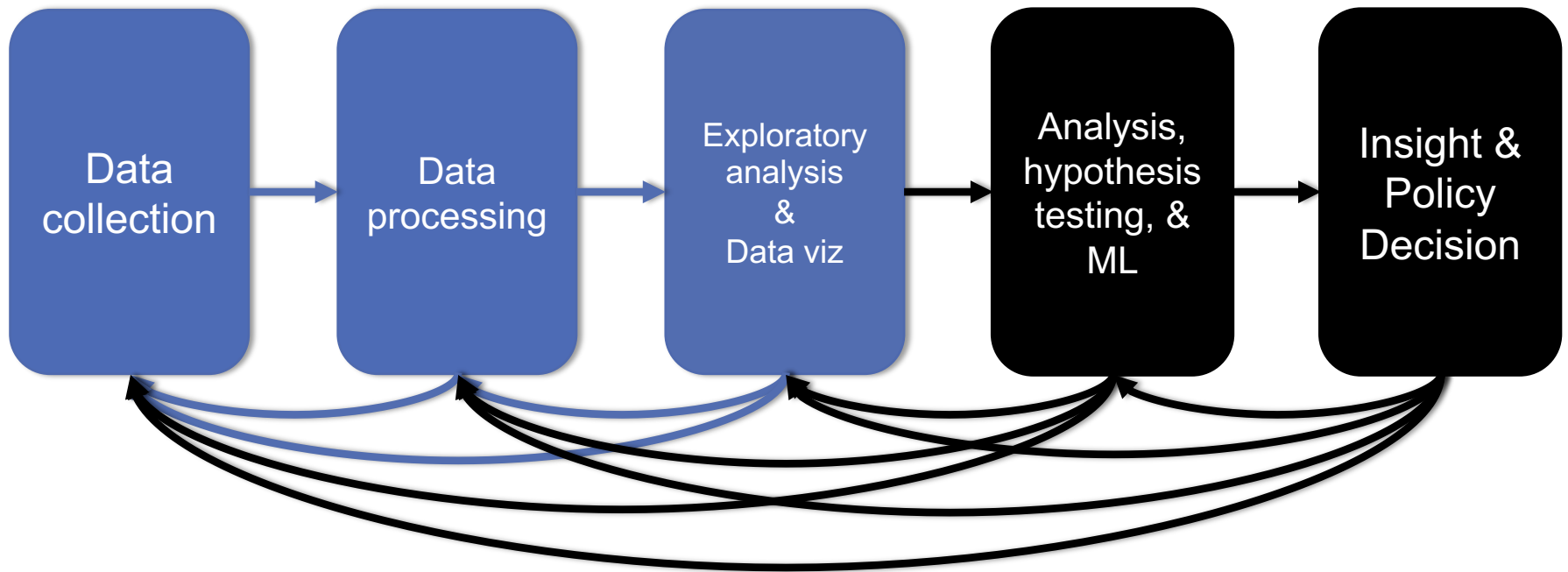
Wednesdays

7:00pm – 9:30pm



COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

TODAY'S LECTURE



TODAY'S LECTURE

Overview

A visualization framework

A visualization workflow

Evaluation research of information visualization

Quick Matplotlib/Seaborn overview

Some extra resources



Huge thanks: AI Laboratory at Arizona State University

WHAT IS VISUALIZATION?

Visualization is any technique for creating images, diagrams, or animations to **communicate a message**.

Visualization through visual imagery has been an effective way to communicate both abstract and concrete ideas since the dawn of humankind.

- Cave paintings
- Hieroglyphs
- Maps
- ...



Message: "Don't touch the spiky hedgehogs" ...

BRIEF HISTORY

First known map: 12th century [Tegarden 1999]

First presentation graphics: William Playfair (1786)

Multidimensional representations appeared in 19th century
[Tufte 1983]

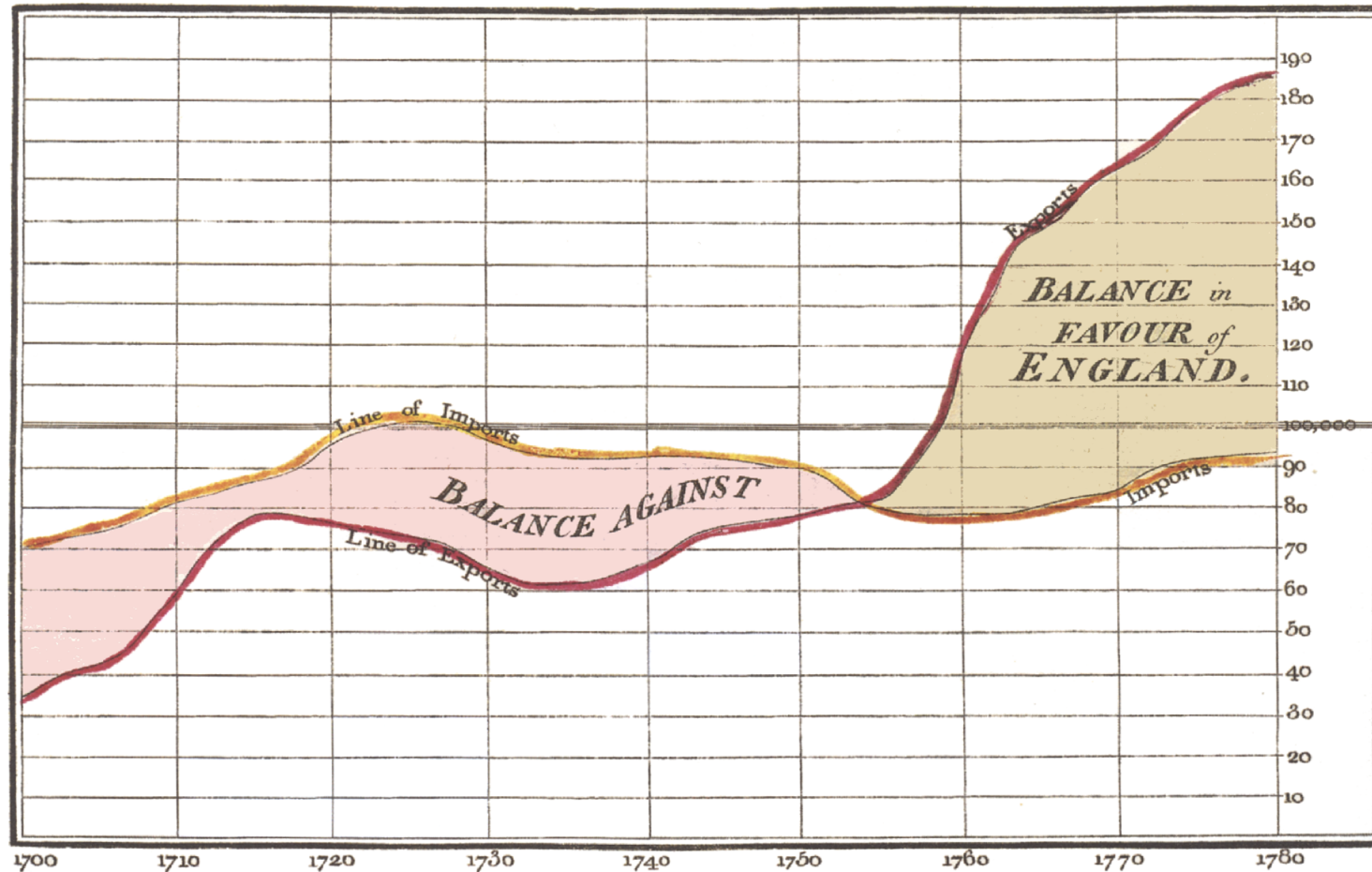
Examples:

- William Playfair (1786, 1821) – founder of graphical methods for statistics & economist
- John Snow (1854) – Cholera Map in London
- Charles Joseph Minard (1861) – Napoleon and The Russian Campaign of 1812

“... the best statistical graphic ever drawn.” -- Tufte

BRIEF HISTORY

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



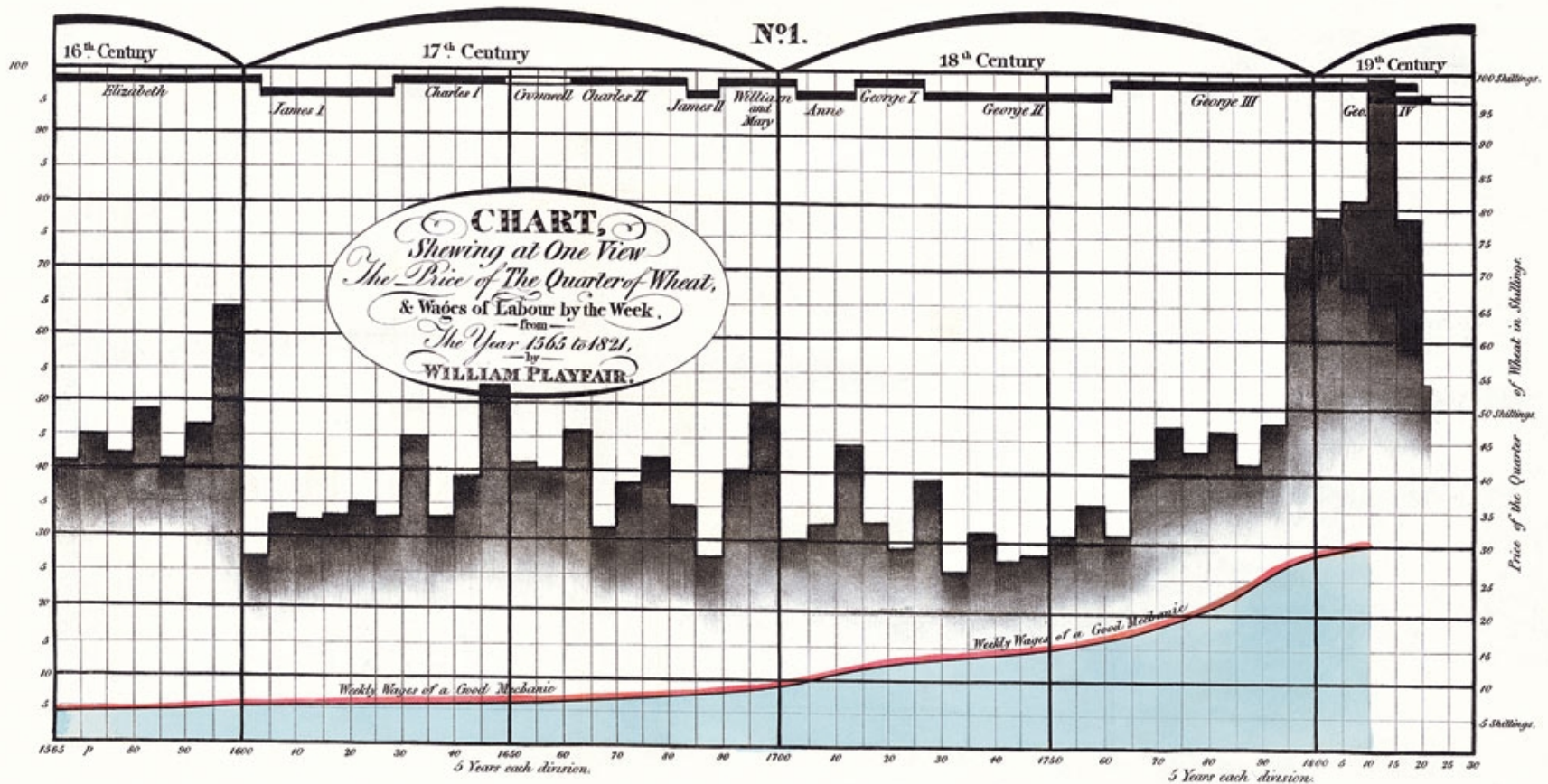
The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 18th May 1786, by W^m Playfair

Nesle sculpt 352 Strand, London.

William Playfair (1786)

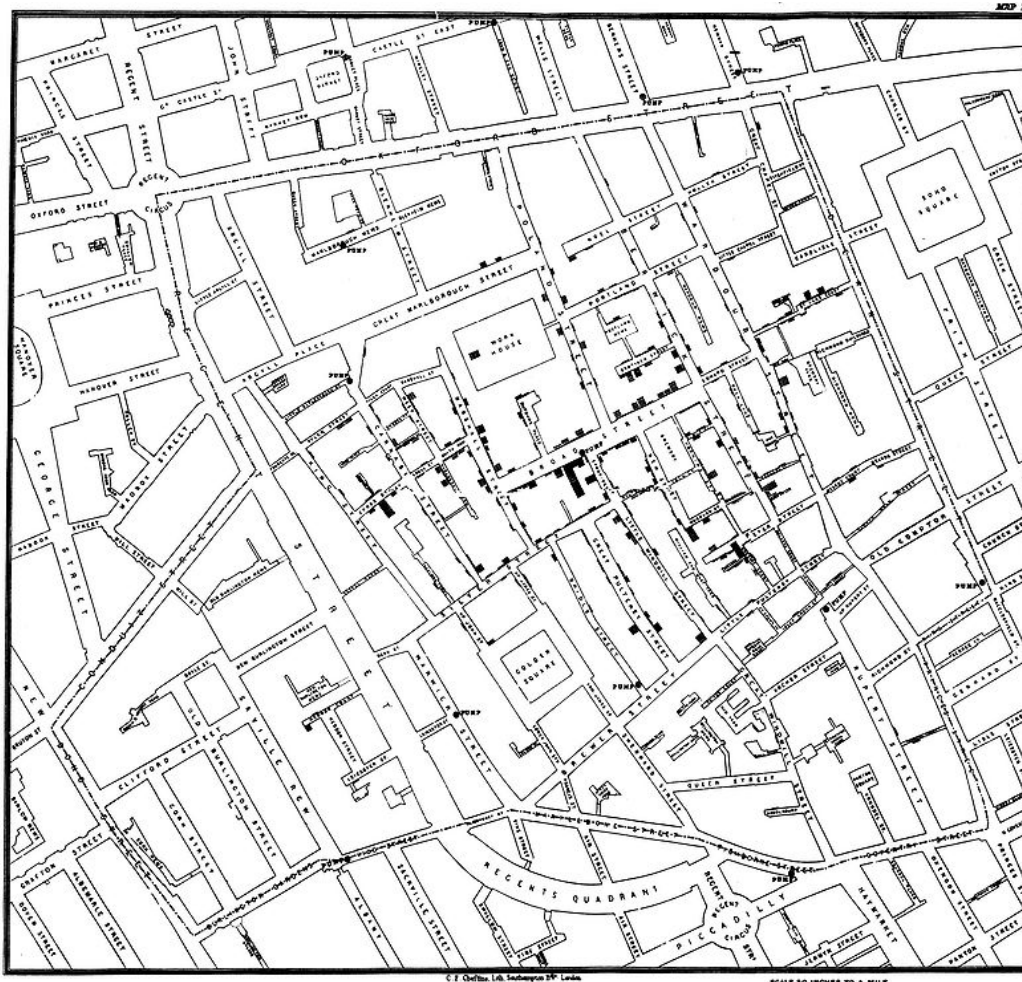
BRIEF HISTORY



- Time-series comparison, aimed to show wheat price declined regarding the increase of wages
- Integration of bar charts and line graph

William Playfair (1821)

BRIEF HISTORY



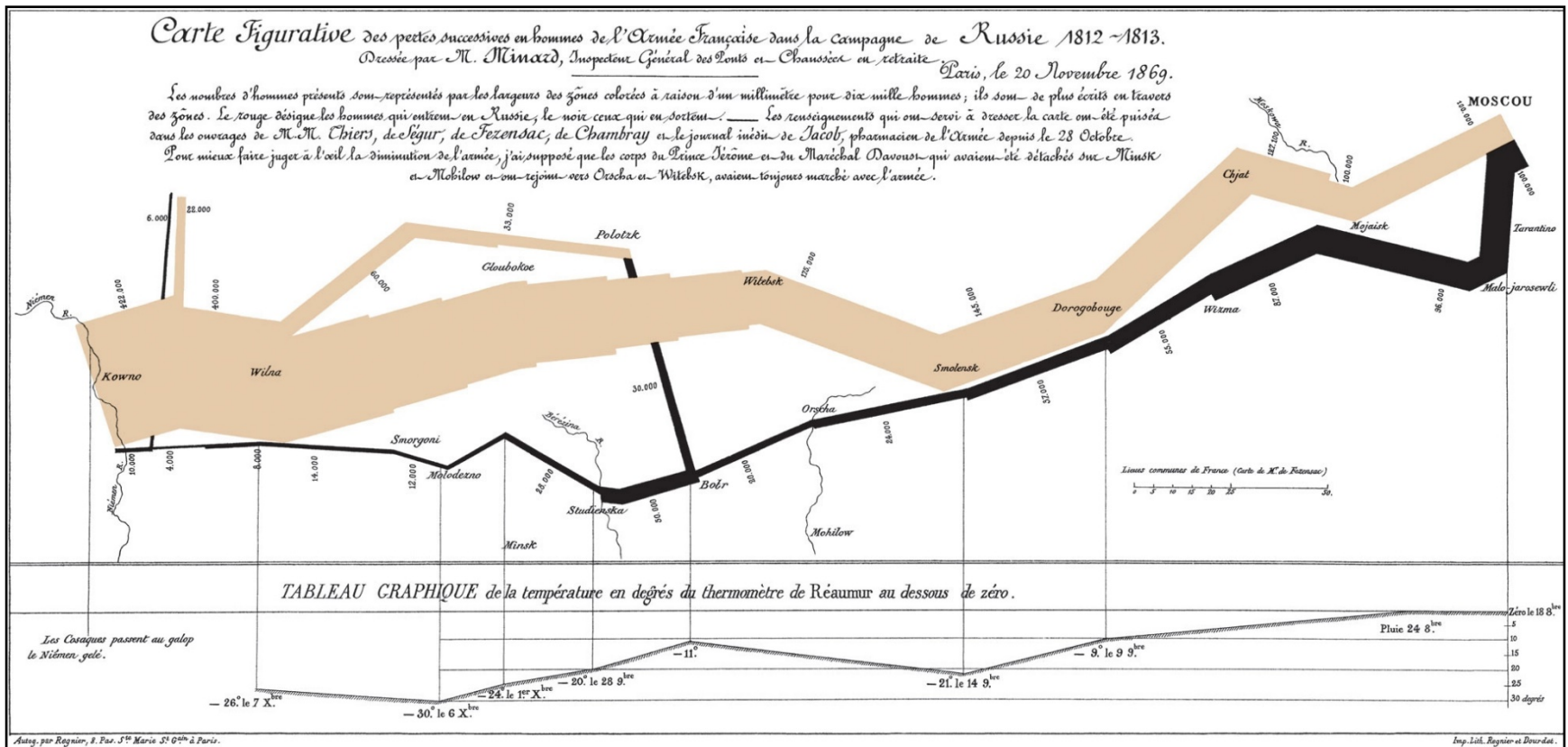
John Snow mapped the cholera cases on the map. The cases were clustered around the pump in Broad street.

- Each death case is a bar
- “Spot map” – Geo-spatial based mapping

John Snow (1854)



BRIEF HISTORY

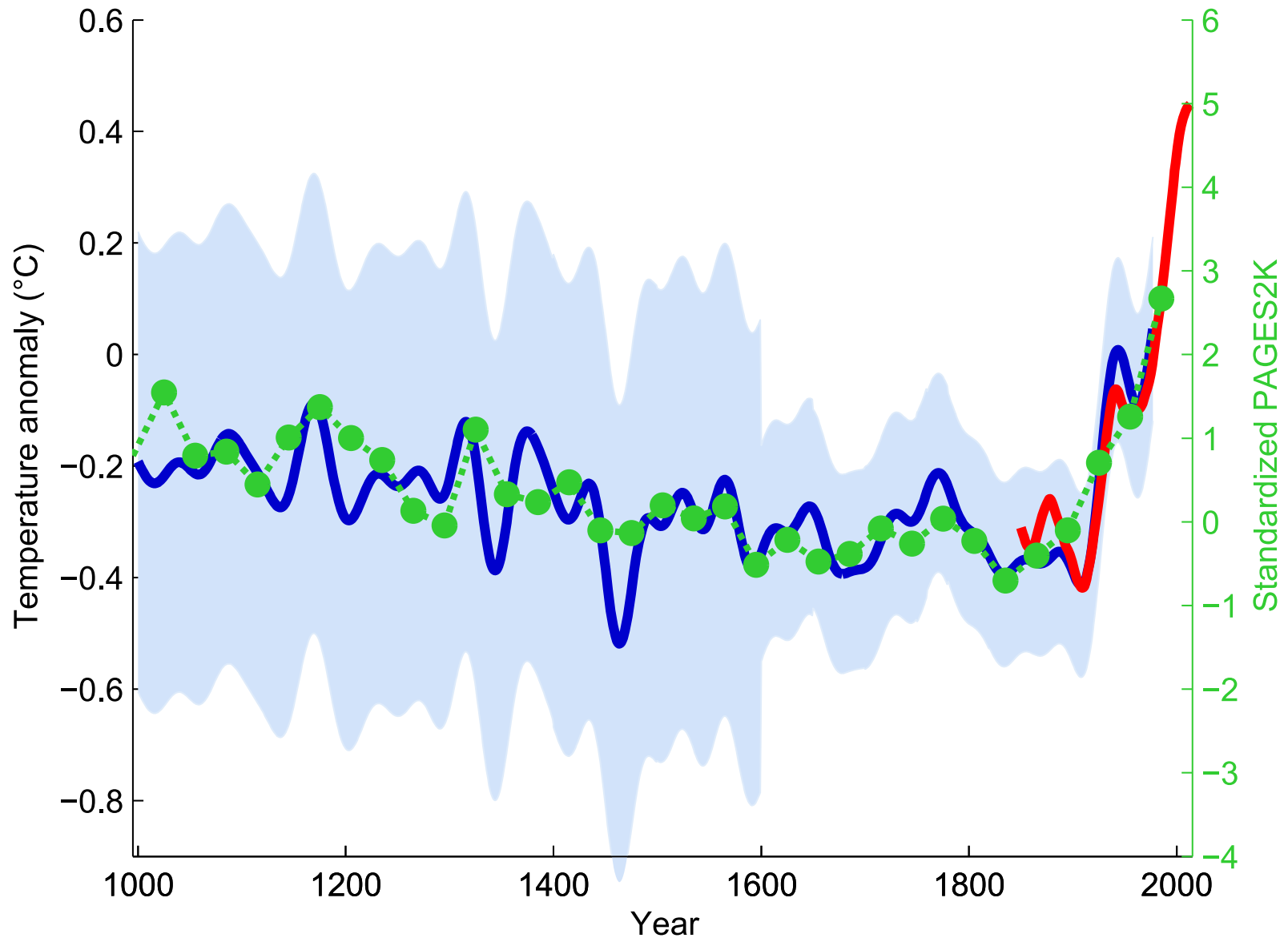


In this visualization, Minard visualized:

- Napoleon's marching and retreat routes
- Army Size
- Temperature during retreat

Charles Joseph Minard (1861)

Hockey Stick Graph: Temperature Change over last 1000 years



BRIEF HISTORY

Modern Visualization

- 1985: NSF Workshop on Scientific Visualization
- 1987: Special issue of Computer Graphics on Visualization in Scientific Computing
- 1990: IEEE 1st visualization conference
- 2000s: Public media started to integrate infographics into TV news, newspaper/ magazine publication
- 2004: Pak Chung Wong and J. Thomas (2004). "Visual Analytics". in: IEEE Computer Graphics and Applications, Volume 24, Issue 5, Sept.-Oct. 2004 Page(s): 20–21.
- Visualizations become popular on social networks
- Techniques such as HTML5, CSS3, D3 enabled interactive visualizations on mobile devices

VISUALIZATION CLASSIFICATION

Visualization can be classified as software visualization, scientific visualization, and information visualization.

- **Software Visualization**
 - Software visualization helps people understand and use computer software effectively [Stasko et al. 1998]
- **Scientific Visualization**
 - Visualizing three-dimensional phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth [Friendly 2008]

VISUALIZATION CLASSIFICATION

- **Information Visualization**

- Information visualization is the two-way and interactive interface between humans and their information resources. Visualization technologies meld the human's capacity with the computational capacity for analytical computing. (P1000 report)
- Information visualization concentrates on the use of computer-supported tools to explore large amount of abstract data. The abstract data include both numerical and non-numerical data, such as text and geographic information.

VALUE OF INFORMATION VISUALIZATION

Exploring information collections becomes increasingly difficult as the volume grows

With minimal effort, the human visual system can process a large amount of information in a parallel manner

The occurrence of advanced graphical software and hardware enables the large-scale visualization and the direct manipulation of interfaces

GOALS FOR INFORMATION VISUALIZATION

Provide insight

- Explain data to solve specific problems
- Support the analytical task, showing the comparison or causality
- Explore large data sets for better understanding

Relieve the cognitive overload

- Conveying abstract information in intuitive ways

HOW TO ORGANIZE VISUALIZATIONS?

By

- User task type?
- User insight need?
- Data to be visualized?
- Data transformation?
- Visualization techniques?
- Visual mapping transformation?
- Interaction techniques?
- ...?

DIFFERENT QUESTION TYPES



Terabytes of data

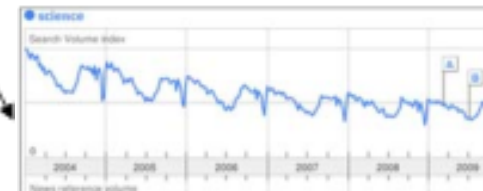
Descriptive &
Predictive
Models



Find your way



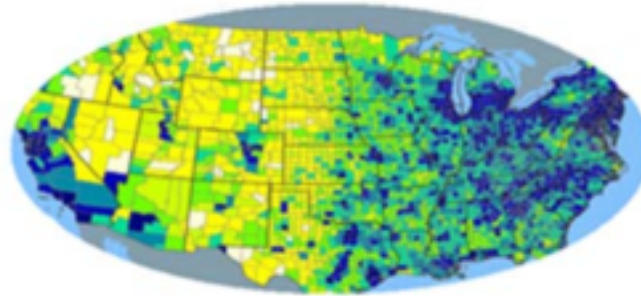
Find collaborators, friends



Identify trends

DIFFERENT LEVELS OF ANALYSIS

Macro/Global
Population Level



Meso/Local
Group Level



Micro
Individual Level



FRAMEWORK BY BÖRNER (2014)

Level of Analysis

Type/Question	Micro/Individual (1-100 records)	Meso/Local (100-10k records)	Macro/Global (10,000+ records)
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of USA, all of science.
Temporal Analysis (When)	Funding portfolio of one individual	Mapping topic bursts in 20 years of PNAS	113 years of physics research
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a state's intellectual landscape	PNAS publications
Topical Analysis (What)	Base knowledge from which one grant draws	Knowledge flows in chemistry research	VxOrd/Topic maps of NIH funding
Network Analysis (With Whom)	NSF Co-PI network of one individual	Co-author network	NIH's core competency

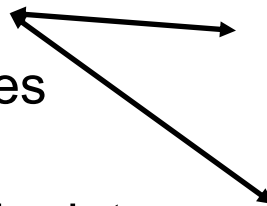
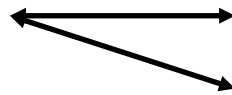
COMPARING TWO WORKFLOWS

Börner (2014) – Visual Insights

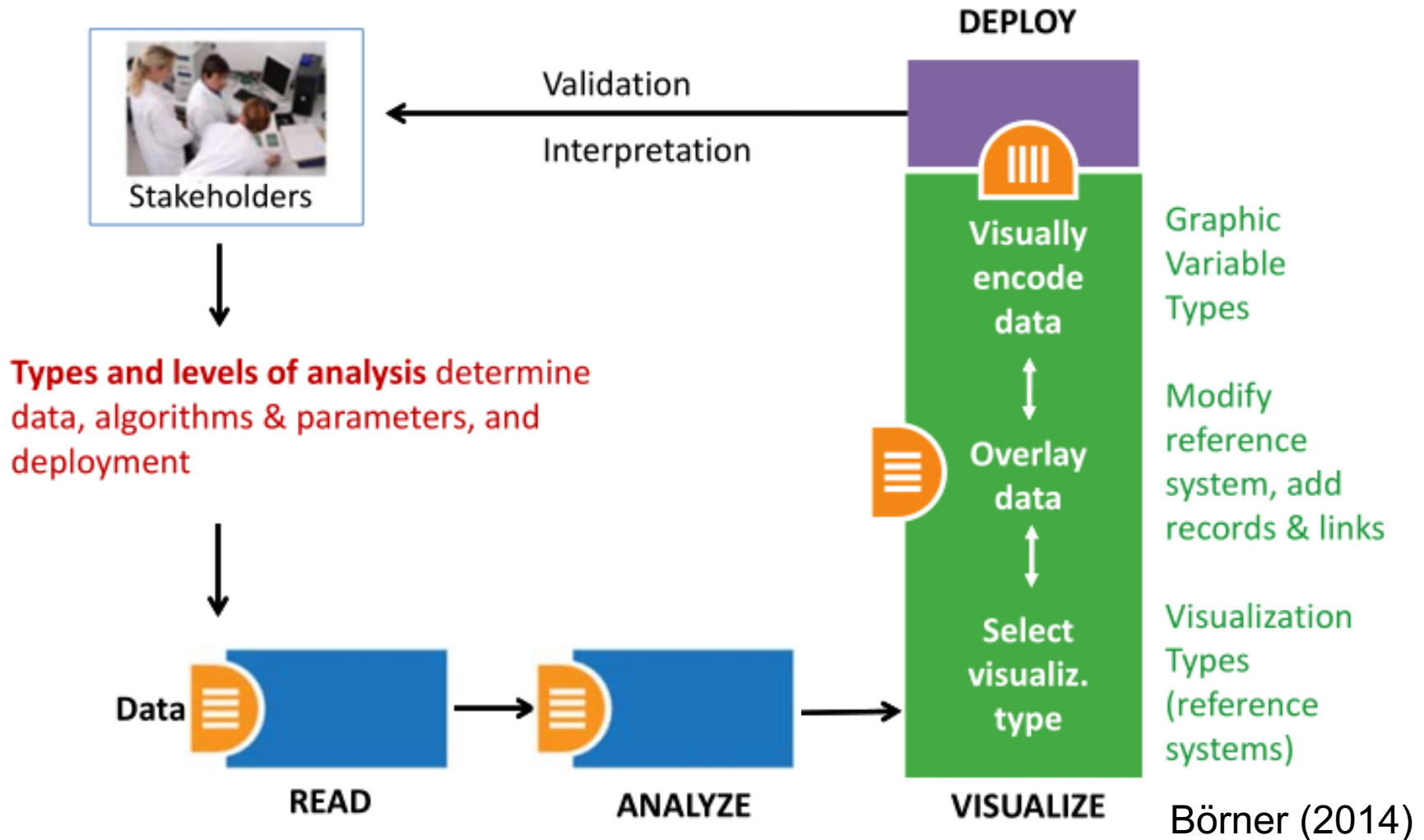
- Read data
- Analyze data
- Visualize
 - Select vis. types
 - Overlay data
 - Visually encode data
- Deploy

Fry (2007) – Visualizing Data

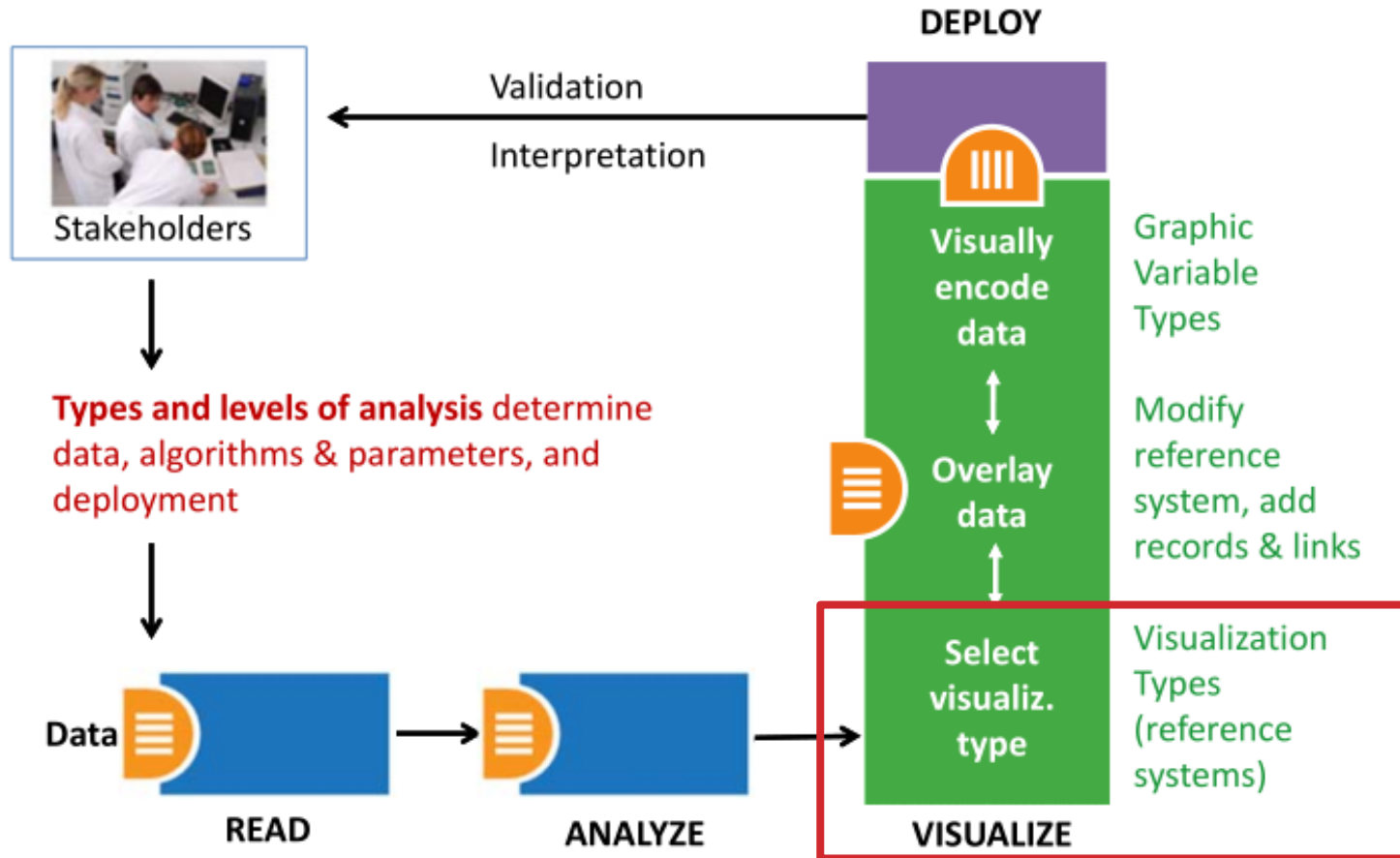
- Acquire
- Parse
- Filter
- Mine
- Represent
- Refine
- Interact



NEEDS-DRIVEN WORKFLOW



NEEDS-DRIVEN WORKFLOW: SELECT VISUALIZATION TYPE



VISUALIZATION TYPES



Shneiderman (1996) proposed seven types of representation methods:

1-D

- To represent information as one-dimensional visual objects in a linear [Eick et al. 1992; Hearst 1995] or a circular [Salton et al. 1995] manner.
- Textual documents, source codes, name lists, ...

2-D

- To represent information as two-dimensional visual objects
- Geographic maps, floor plans, newspaper layouts, ...

3-D

- To represent information as three-dimensional visual objects
- Molecules, human bodies, buildings and etc.

Multi-dimensional

- To represent information as multidimensional objects and projects them into a three-dimensional or a two-dimensional space.
- Relational and statistical database.

This slide is a terrible example of visualizing information ...

VISUALIZATION TYPES



Shneiderman (1996) proposed seven types of representation methods (cont'd):

Temporal

- To represent information based on temporal order
- Medical records, project management, historical presentations, ...

Tree

- To represent hierarchical relationship

Network

- To represent complex relationships that a simple tree structure cannot capture

VISUALIZATION TYPES

Börner (2014) emphasized five major types:

Charts

- No reference system—e.g., Wordle.com, pie charts

Tables

- Categorical axes that can be selected, reordered; cells can be color coded and might contain proportional symbols. Special kind of graph.

Graphs

- Quantitative or qualitative (categorical) axes. Timelines, bar graphs, scatter plots.

Geospatial maps

- Use latitude and longitude reference system. World or city maps.

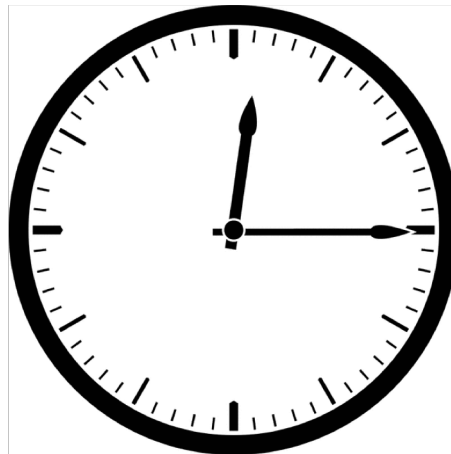
Network graphs

- Node position might depends on node attributes or node similarity.
- Tree graphs: hierarchies, taxonomies, genealogies.
- Networks: social networks, migration flows.

VISUALIZATION TYPE SELECTION: TEMPORAL

Temporal data analysis and visualization answer the **when** question, and can help to :

- Understand the temporal distribution of datasets
- Identify growth rates, latency to peak times, or decay rates
- See patterns in time-series data, such as trends, seasonality, or bursts.



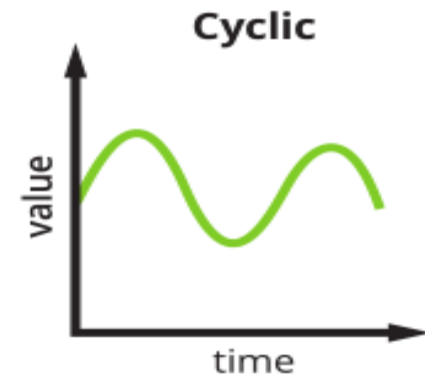
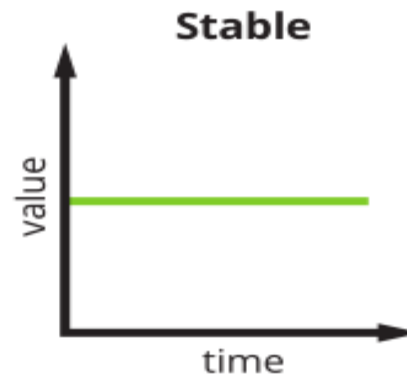
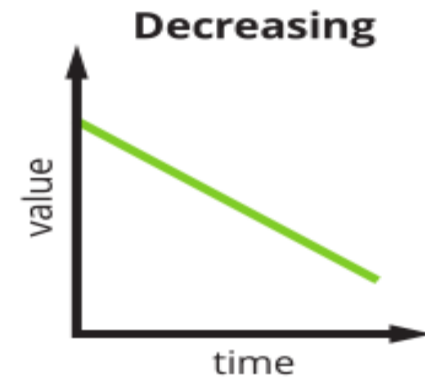
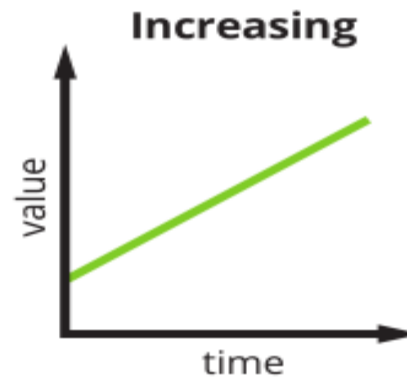
VISUALIZATION TYPE SELECTION: TEMPORAL

Temporal trends

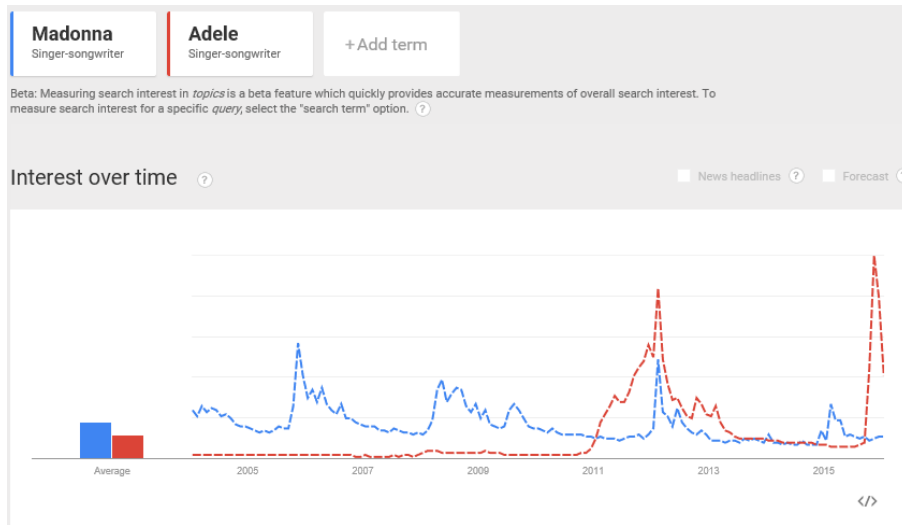
- Increasing
- Decreasing
- Stable
- Cyclic

Visualization Types

- Line graph
- Stacked graph
- Temporal bar graph
- Histograms
- Small multiples

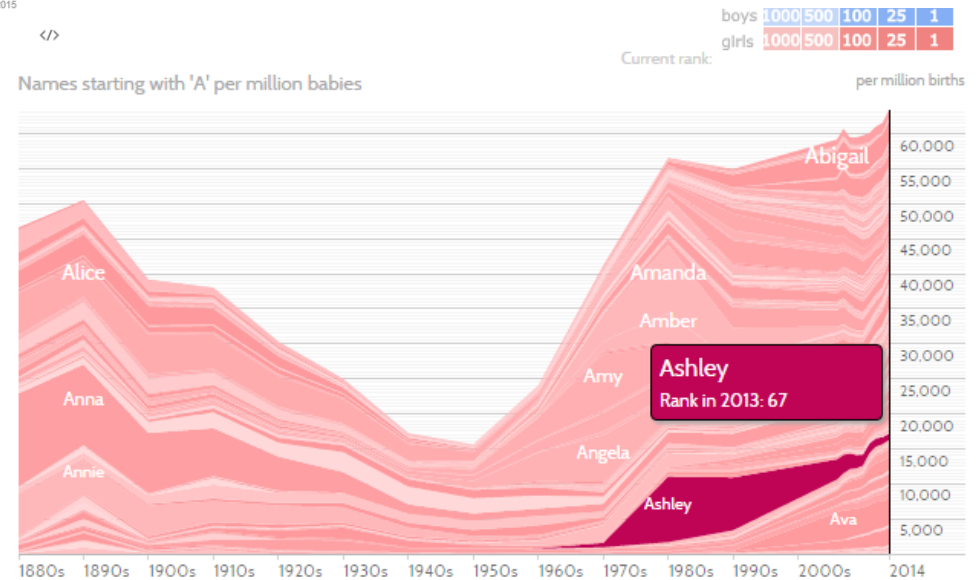


VISUALIZATION TYPE SELECTION: TEMPORAL



Line graph shows the trends over time.

Stacked graph illustrates individual and total trends.



VISUALIZATION TYPE SELECTION: TEMPORAL

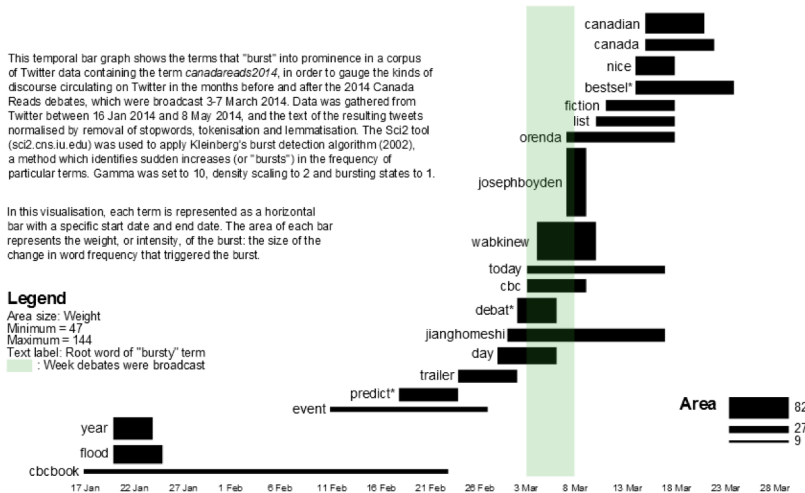
Prominent terms in tweets containing the term *canadareads2014*

This temporal bar graph shows the terms that "burst" into prominence in a corpus of Twitter data containing the term *canadareads2014*, in order to gauge the kinds of discourse circulating on Twitter in the months before and after the 2014 Canada Reads debates, which were broadcast 3-7 March 2014. Data was gathered from Twitter between 16 Jan 2014 and 8 May 2014, and the text of the resulting tweets normalised by removal of stopwords, tokenisation and lemmatisation. The Sci2 tool (sci2.cns.iu.edu) was used to apply Kleinberg's burst detection algorithm (2002), a method which identifies sudden increases (or "bursts") in the frequency of particular terms. Gamma was set to .10, density scaling to 2 and bursting states to 1.

In this visualisation, each term is represented as a horizontal bar with a specific start date and end date. The area of each bar represents the weight, or intensity, of the burst; the size of the change in word frequency that triggered the burst.

Legend

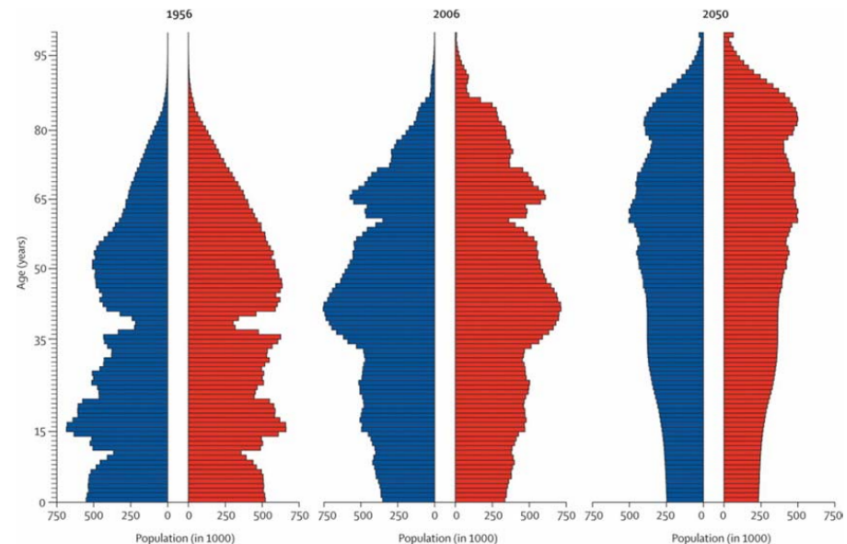
Area size: Weight
Minimum = 47
Maximum = 144
Text label: Root word of "bursty" term
Green bar: Week debates were broadcast



Anouk Lang, University of Strathclyde • @a_e_lang • <http://aelang.net> • Produced with funding from the AHRC's Cultural Value Project

Temporal bar graph shows begin, end, and properties of events.

Histograms represent the number of observations of a certain value have been made.

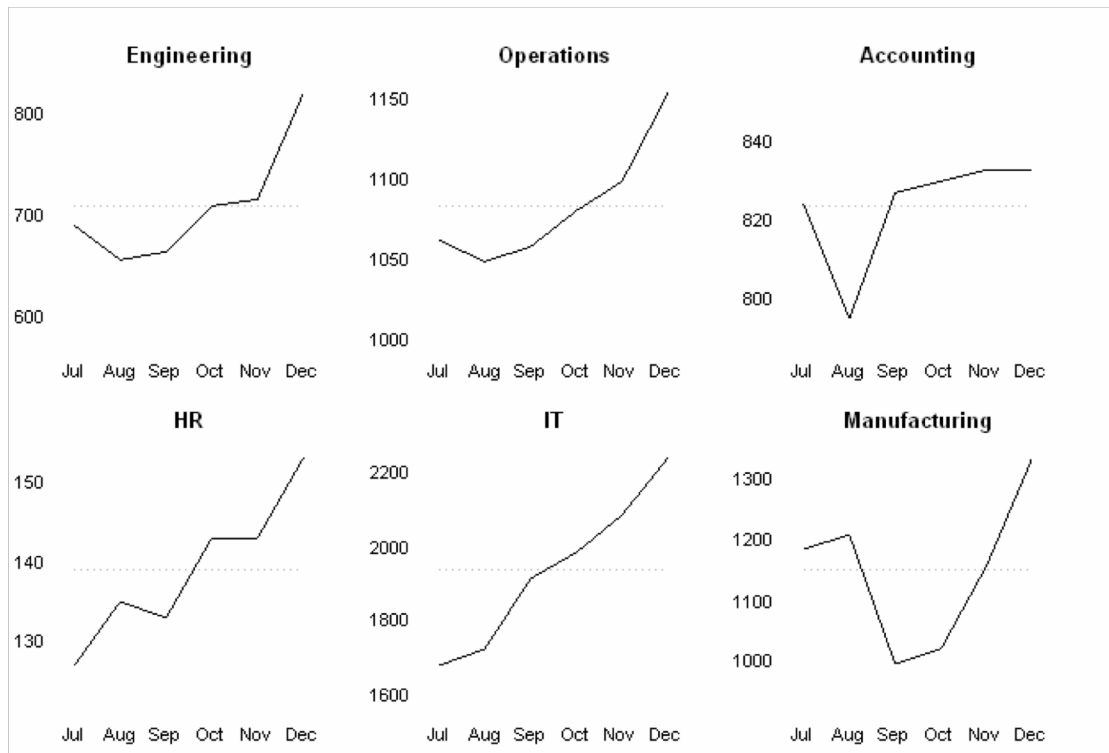


Christensen, Kaare, Gabriele Doblhammer, Roland Rau, and James W. Vaupel. "Ageing Populations: The Challenges Ahead." *The Lancet* 374 (9696): 1196-1208.

VISUALIZATION TYPE

SELECTION: TEMPORAL

A small multiple (sometimes called trellis chart, lattice chart, grid chart, or panel chart) is a series of similar graphs or charts using the same scale + axes, allowing them to be easily compared. [Tufte 1983]



VISUALIZATION TYPE

SELECTION: GEOSPATIAL

Geospatial data analysis and visualization originated in geography and cartography, but are increasingly common in statistics, information visualization, and many other areas of science.

The analyses aim to answer **where** questions that use location information to identify their position or movement over geographic space.



VISUALIZATION TYPE SELECTION: GEOSPATIAL

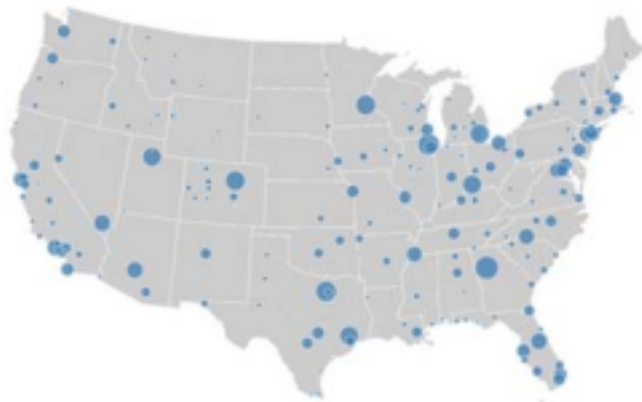
General Map Types

- General reference maps
- Topographic maps
- Thematic maps: Emphasize the spatial distribution of one or more geographic variables
 - Physio-geographical:
 - Maps showing the natural features of the earth's surface.
 - Socio-economic:
 - Maps showing political boundaries, population density, or voting behavior.
 - Technical
 - Maps showing navigation routes.

VISUALIZATION TYPE

SELECTION: GEOSPATIAL

Proportional symbol map represents data variables by symbols that are sized, colored, etc. according to their amount. Data is (or can be) aggregated at points within areas.

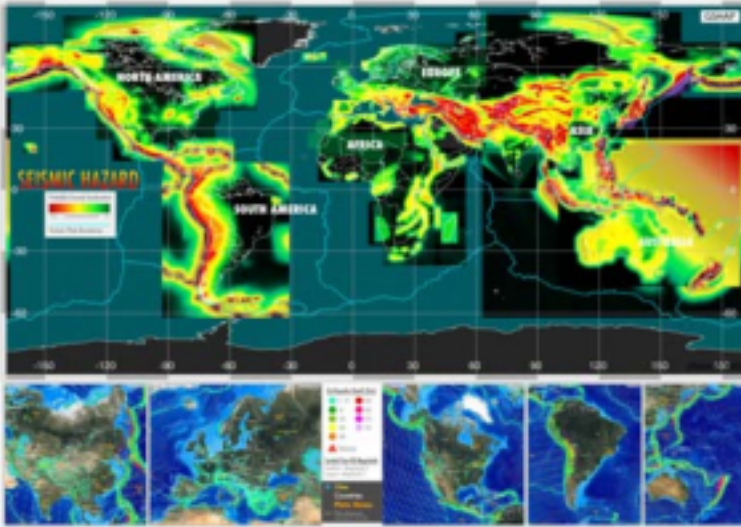


Choropleth map represents data variables such as densities, ratios, or rates by proportionally colored or patterned areas.

- Mapped according to a prearranged key, e.g., districts or states



VISUALIZATION TYPE SELECTION: GEOSPATIAL



Heat (isopleth) maps represent continuous data variable values by colors. Heat maps might show color-based contour lines that connect points of equal value or value-by-area maps.

Cartograms are not drawn to scale. Instead they distort geographical areas in proportion to data values. Familiarity with region is necessary. Mostly used for world, continental, and country maps.

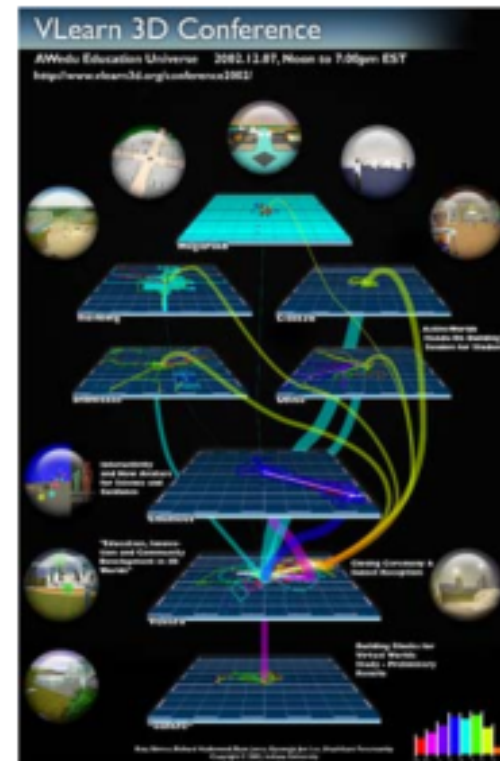


VISUALIZATION TYPE SELECTION: GEOSPATIAL



Flow maps show paths that (in)tangible objects take to get from one geospatial place to another. Variables such as capacity or maximum speed are encoded proportionally by line width/color.

Space-time cubes display entities, locations, and events over time.



VISUALIZING UNCERTAINTY



What you show
Credit: Alberto Cairo

VISUALIZING UNCERTAINTY



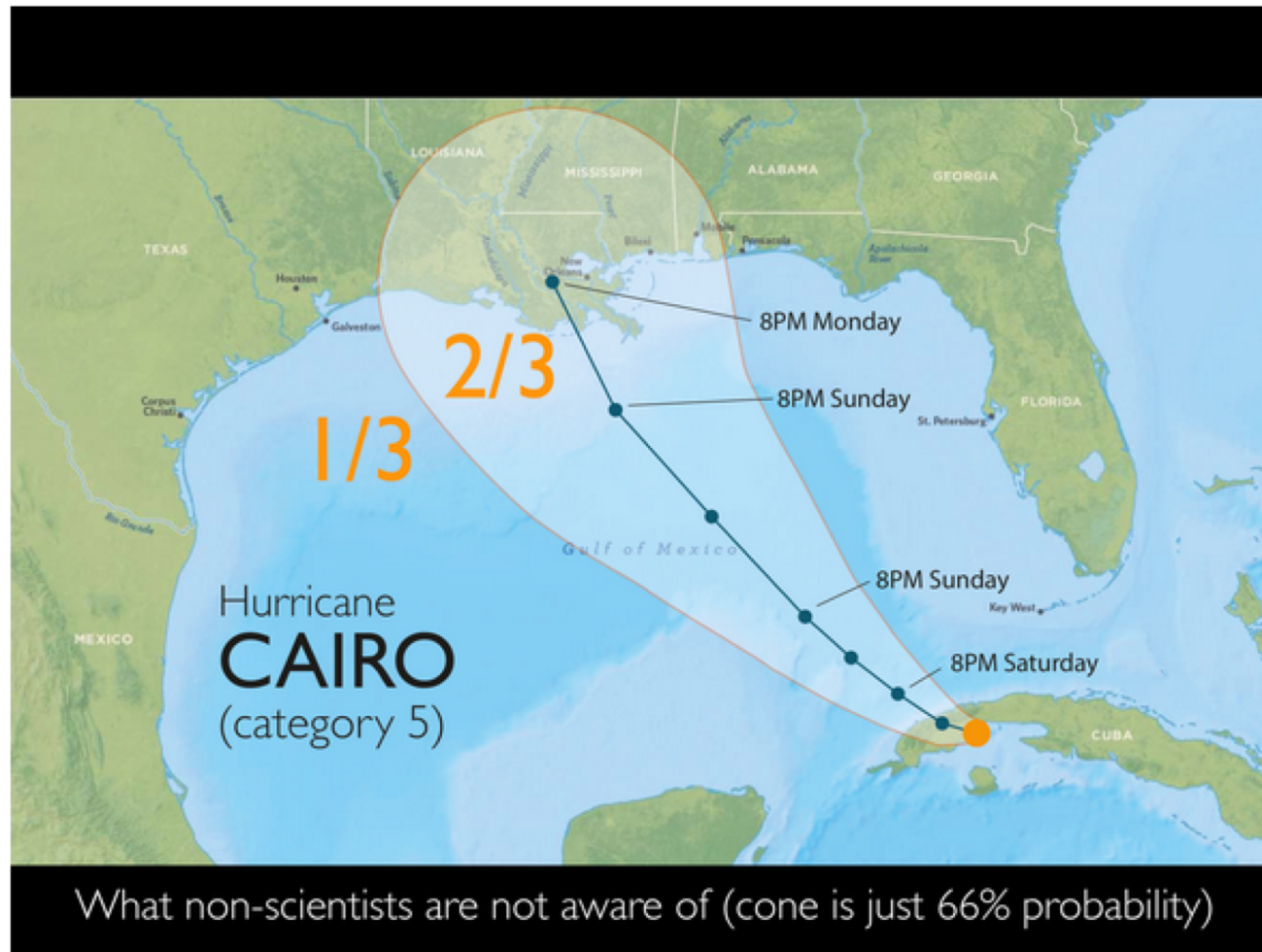
What the cone is based on
Credit: Alberto Cairo

VISUALIZING UNCERTAINTY



What I think some people see
Credit: Alberto Cairo

VISUALIZING UNCERTAINTY



What non-scientists are not aware of (cone is just 66% probability)

Credit: Alberto Cairo

VISUALIZING UNCERTAINTY



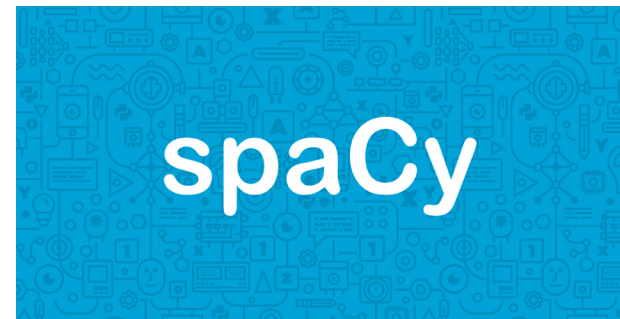
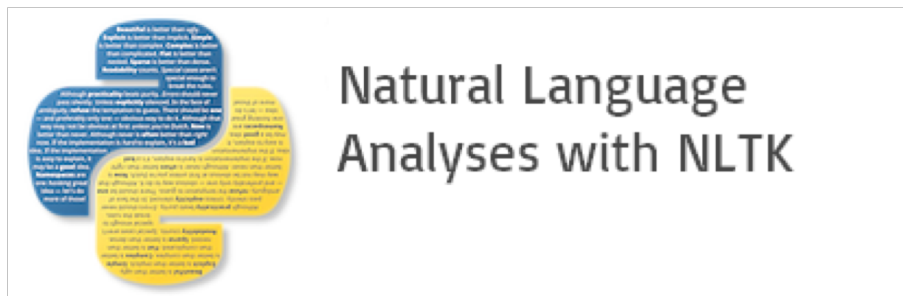
What we could be showing instead
Credit: Alberto Cairo

VISUALIZATION TYPE SELECTION: TOPICAL

To answer the **what** question, we will be using texts to identify major topics, their interrelations, and their evolution over time at different levels of analysis – micro to macro.

To generate visualizations from text, text processing or natural language processing is needed to generate qualitative or quantitative features of the text.

- NLP lectures talk about some of this



VISUALIZATION TYPE SELECTION: TOPICAL

Representations of topical data:

- Charts: Word cloud, text overlay
- Tables: GRIDL, Periodic table
- Graphs: circular visualization, crossmaps
- Geospatial maps: SOM maps
- Network graphs: Tree visualizations, word co-occurrence networks, concept maps, science map overlays

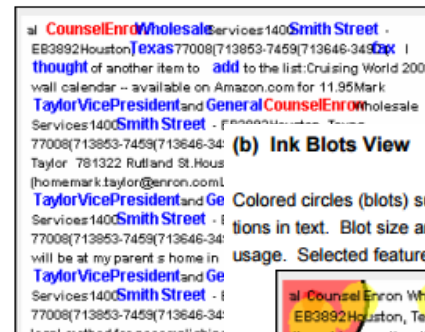
VISUALIZATION TYPE SELECTION: TOPICAL



Text Overlay [Abbassi and Chen 2008]

(a) Text Annotation View

Feature occurrences are highlighted in blue. The selected bag-of-words feature is highlighted in red (CounselEnron).



4: (b) Ink Blots View

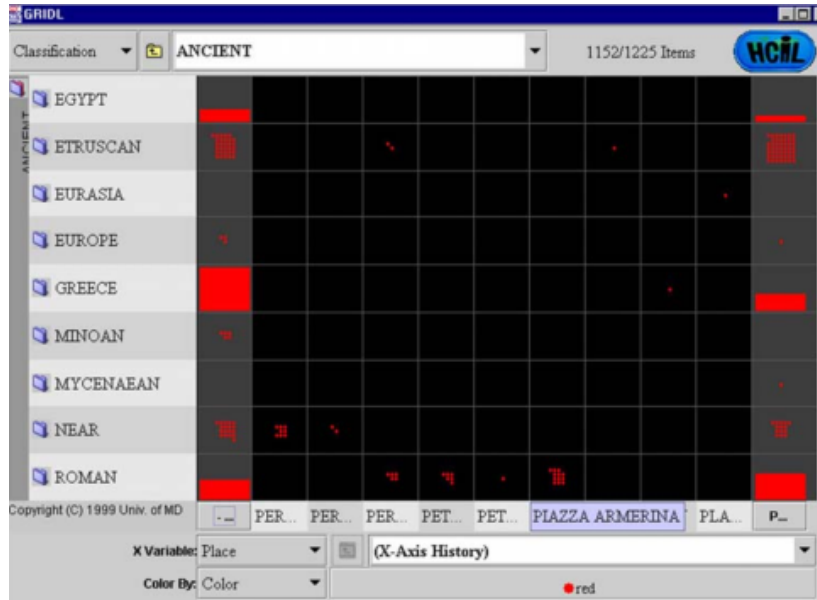
Colored circles (blots) superimposed onto feature occurrence locations in text. Blot size and color indicates feature importance and usage. Selected feature's blots are highlighted with black circles.



Word clouds visualize the frequency of the words.

- IMDB movie titles word cloud created with Wordle.

VISUALIZATION TYPE SELECTION: TOPICAL

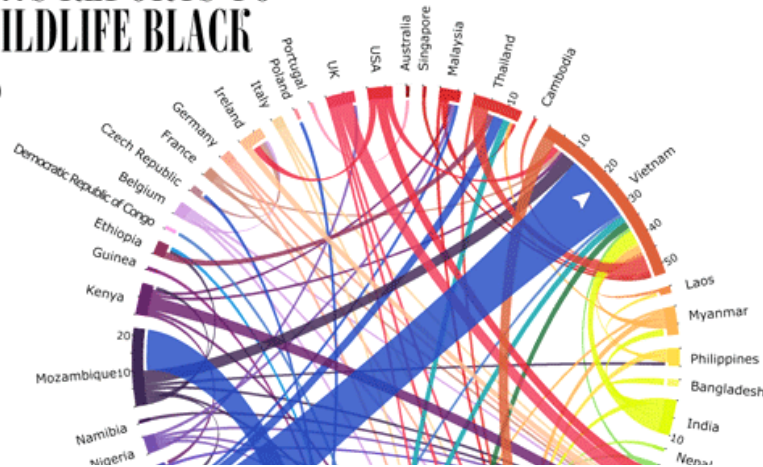


GRIDL uses categorical and hierarchical axes to support categorical zooming.

A Periodic Table is used to organize elements sharing similar properties.

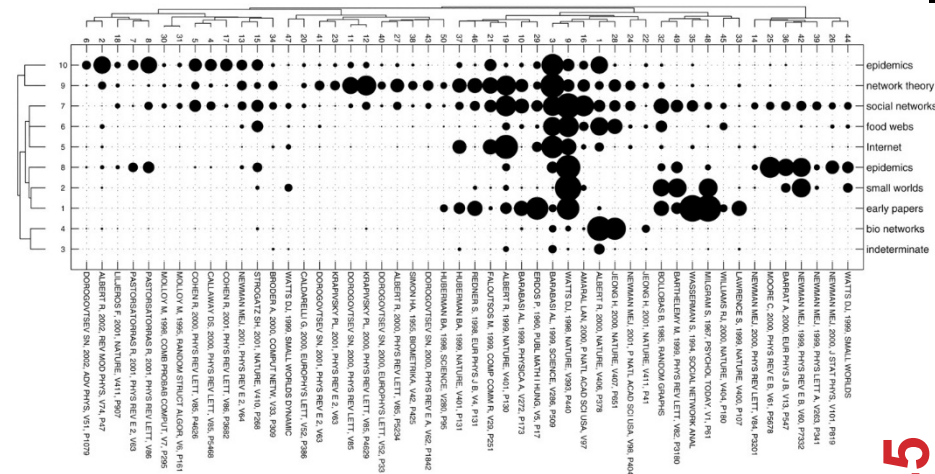
Group →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
↓ Period																		
1	1 H																	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Uut	114 Fl	115 Uup	116 Lv	117 Uus	118 Uuo
			57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb		
			89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No		

USING NEWS REPORTS TO TRACK WILDLIFE BLACK MARKETS

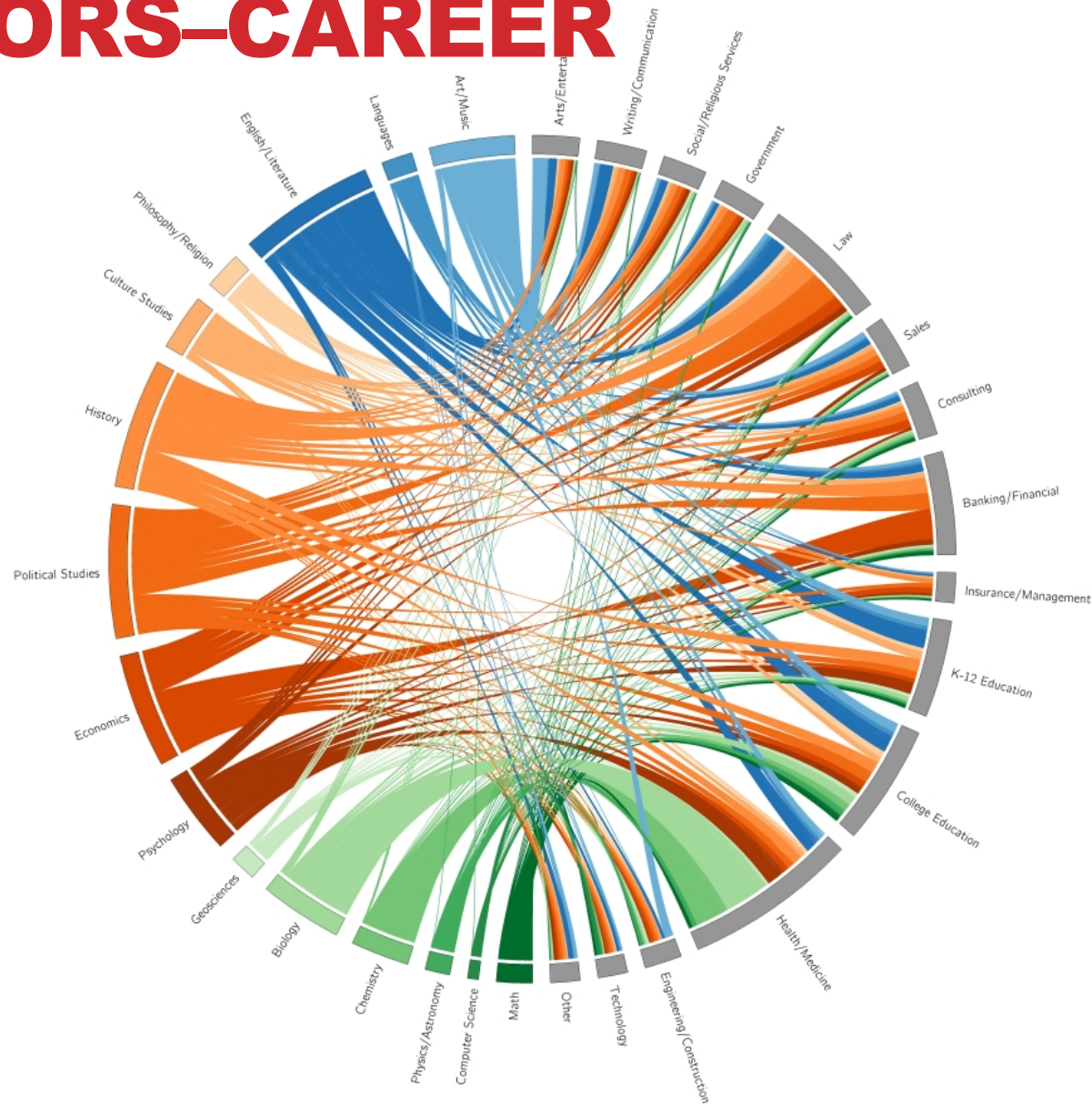


Crossmaps visualize multiple and overlapping relations among entity types in collections of journal articles.

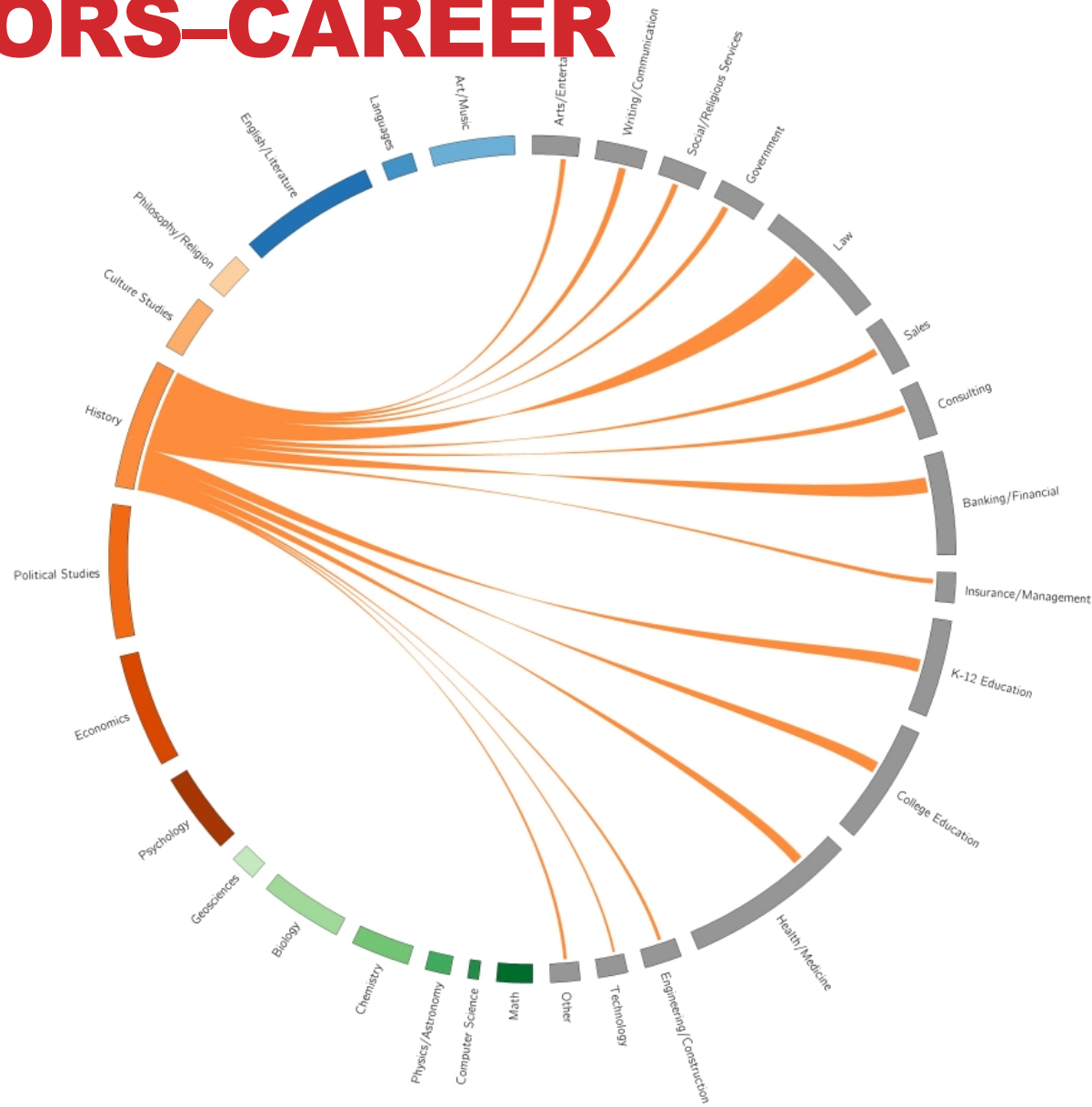
Circular visualization shows the relationships between entities.



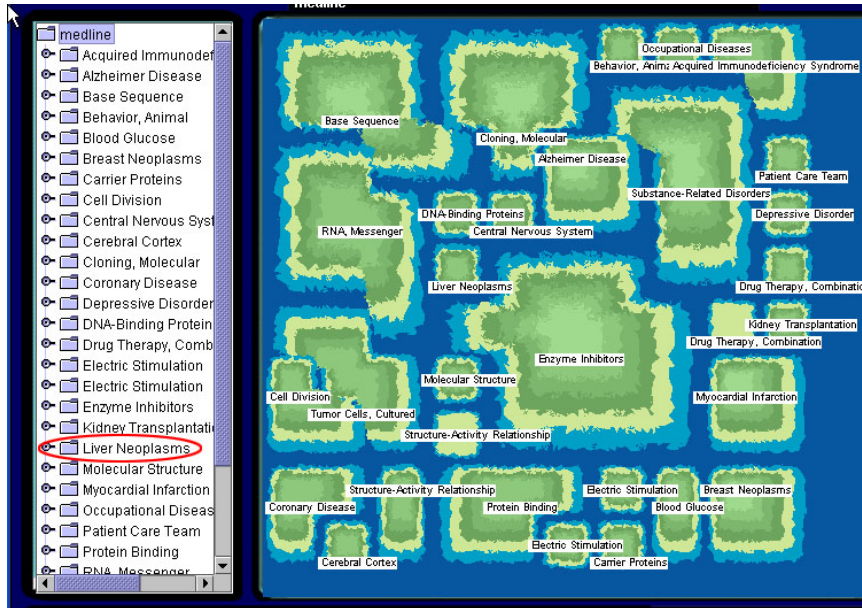
WILLIAMS COLLEGE: MAJORS–CAREER



WILLIAMS COLLEGE: MAJORS-CAREER

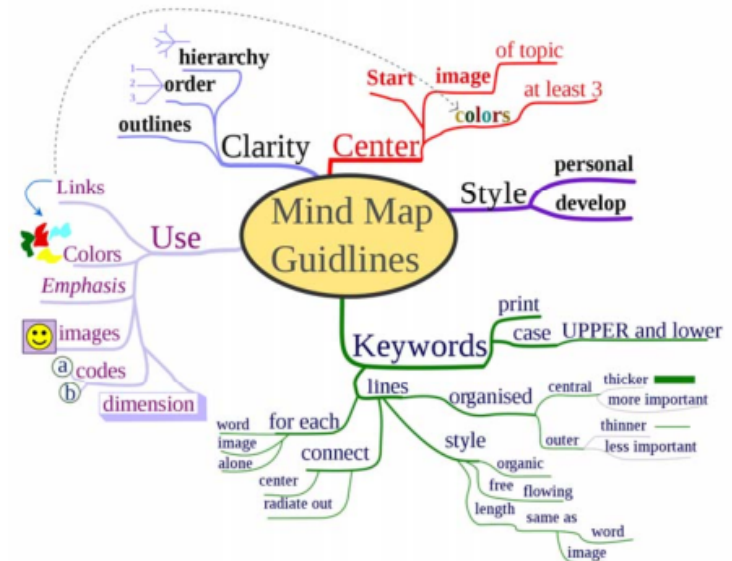


VISUALIZATION TYPE SELECTION: TOPICAL



Self-organizing maps (SOMs) use geospatial metaphors to create abstract 2-D space and map elements onto the space.

Concept maps are network graphs that show the relationships among concepts.

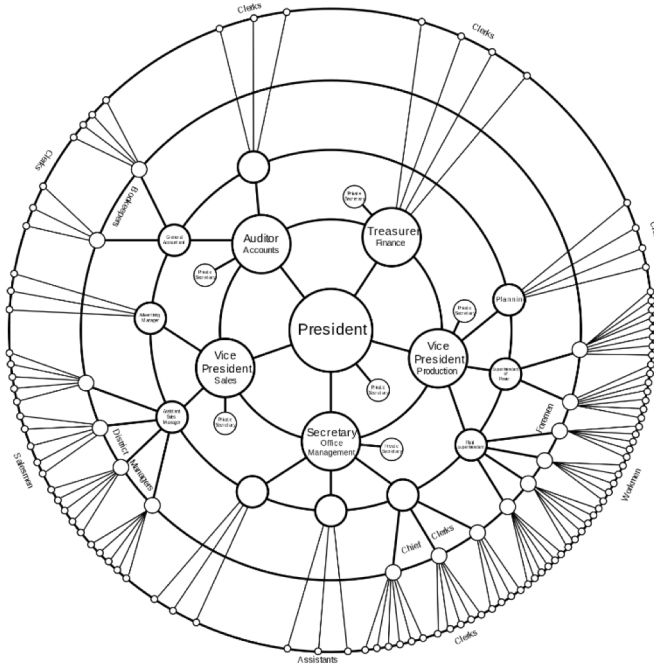


VISUALIZATION TYPE SELECTION: NETWORK

The **with whom** question can be answered by tree and network visualization.

- **Tree visualization** utilizes structural, hierarchical data.
 - Radial Tree Map
 - Treemap
- **Network visualization** can deal with more complex relationships between nodes.
 - Random layout
 - Circular layout
 - Force directed layout
 - Bipartite layout
 - Subway map layout
 - Network overlays on geospatial maps

VISUALIZATION TYPE SELECTION: NETWORK

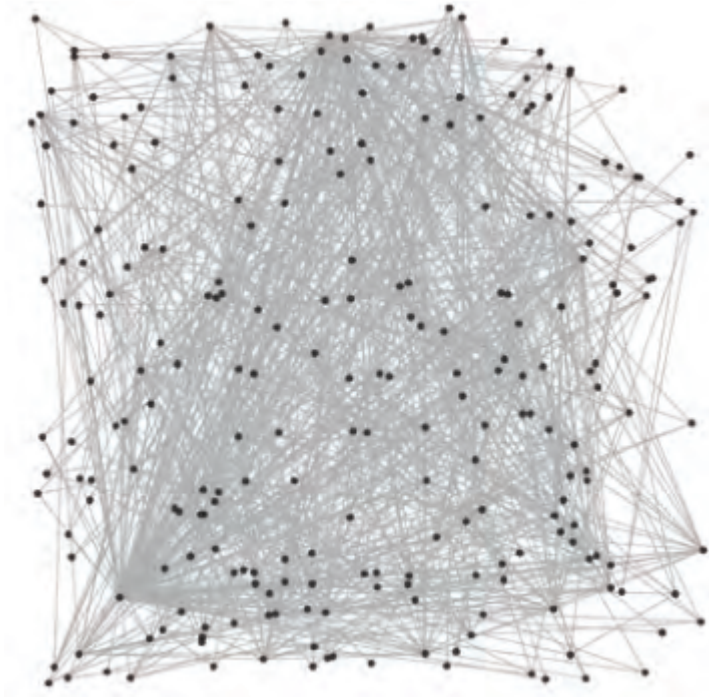


Radial tree maps are a method of displaying a tree structure in a way that expands outwards, radially.

Treemaps are a space-constrained visualization of hierarchical structure



VISUALIZATION TYPE SELECTION: NETWORK

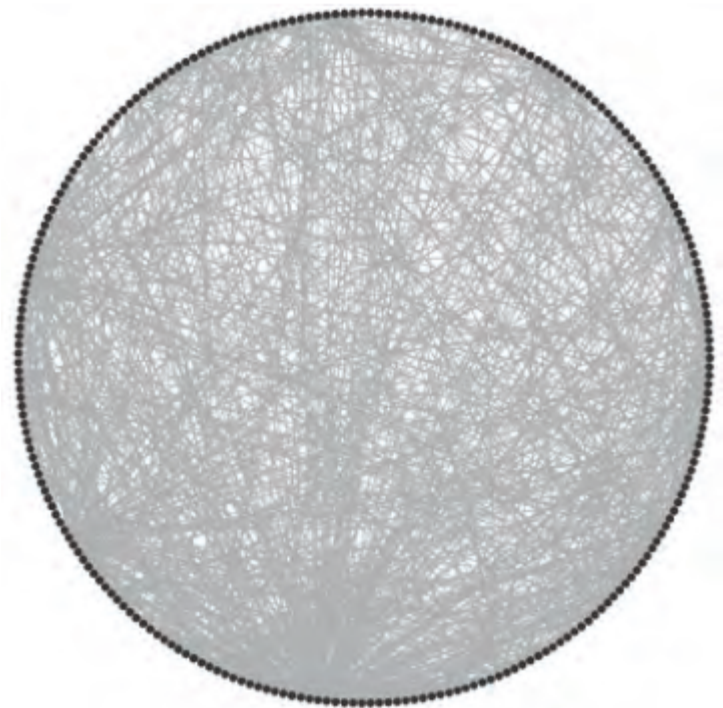


Random layout

- Fast
- Impression of size and density

Circular layout

- Fast
- Can provide sequential information



VISUALIZATION TYPE SELECTION: NETWORK

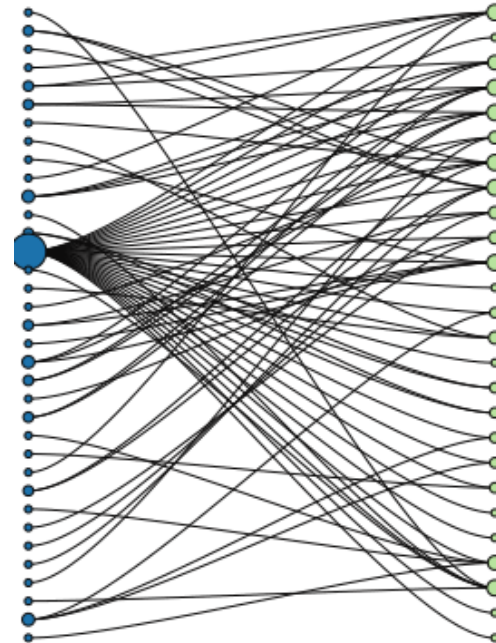


Force-directed layout

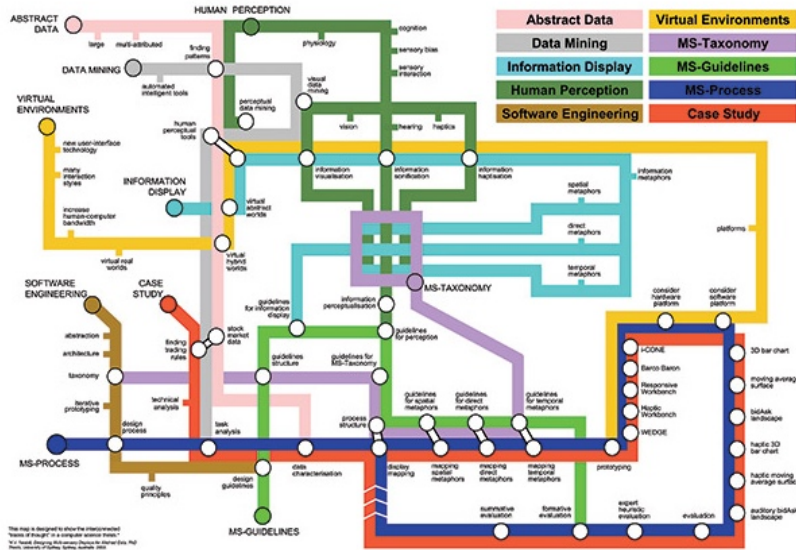
- Utilizes similarity or distance relationships among nodes

Bipartite layout

- Renders networks with two node types as two lists.

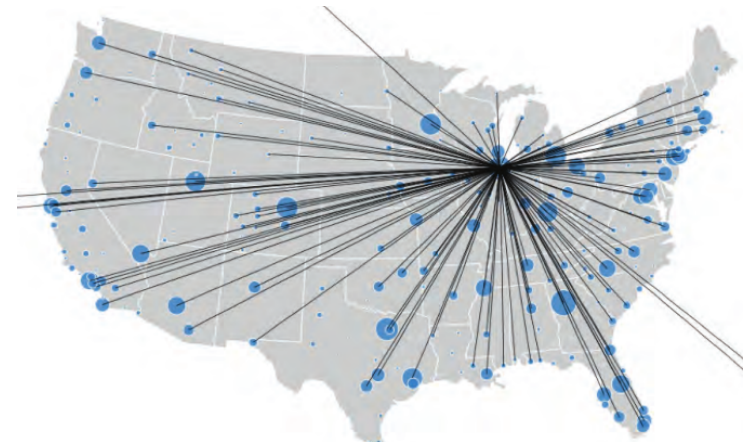


VISUALIZATION TYPE SELECTION: NETWORK



Network overlays on geospatial maps

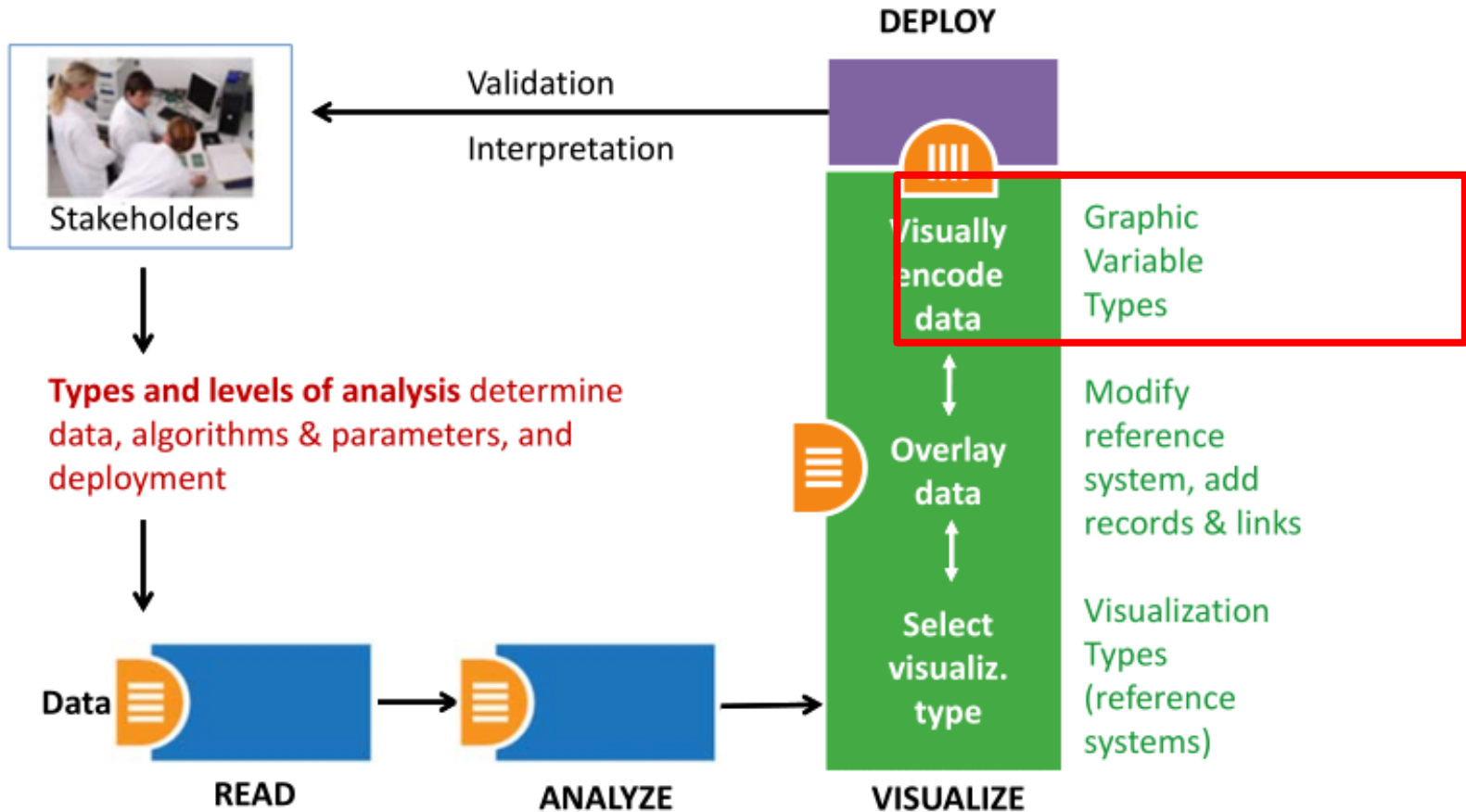
- Use a geospatial reference system to place nodes
- E.g., airport traffic



Subway map layout

- Aim for evenly distributed nodes, uniform edge lengths, and orthogonal drawings.
- E.g. Ph.D. Thesis map (2004)

NEEDS-DRIVEN MODEL: VISUALLY ENCODE DATA



DATA SCALE TYPES

Categorical (nominal): A categorical scale, also called nominal or category scale, is qualitative. Categories are assumed to be non-overlapping.

Ordinal: An ordinal scale, also called sequence or ordered, is qualitative. It rank-orders values representing categories based on some intrinsic ranking but not at measureable intervals.

Interval: An interval scale, also called value scale, is a quantitative numerical scale of measurement where the distance between any two adjacent values (or intervals) is equal but the zero point is arbitrary.

Ratio: A ratio scale, also called proportional scale, is a quantitative numerical scale. It represents values organized as an ordered sequence, with meaningful uniform spacing, and has a true zero point.

More
Qualitative



More
Quantitative

DATA SCALE TYPES: EXAMPLES

Categorical (nominal): ?????????

- Words or numbers constituting the names and descriptions of people, places, things, or events.

Ordinal: ?????????

- Days of the week, degree of satisfaction and preference rating scores (e.g., Likert scale), or rankings such as low, medium, high.

Interval: ?????????



- Temperature in degrees or time in hours. Spatial variables such as latitude and longitude are interval.

Ratio: ?????????


- Physical measures such as weight, height, (reaction) time, or intensity of light; number of published papers, co-authors, citations.

GRAPHIC VARIABLE TYPES

Quantitative

- Position
 - x, y; possibly z
- Form
 - Size
- Color
 - Value (Lightness, Brightness)

 - Saturation (Intensity)

- Texture
 - Pattern
 - Rotation
 - Coarseness
 - Size
 - Density gradient

Qualitative

- Form
 - Shape
 - Orientation (Rotation)
- Color
 - Hue (tint)

- Optics
 - Crispness
 - Transparency
 - Shading

DYNAMIC VISUALIZATION

Previous sections demonstrates the workflow to create a static visualization.

In order to present more information and make visualization **dynamic**, we can show multiple static images side by side or as an animation.

Another way to make dynamic visualization is to introduce interaction into the design process.

- The final mini-project has some of this.
- Your tutorial can have lots of this! E.g., with D3.js.



SHNEIDERMAN'S MANTRA



User-Interface Interaction

- Immediate interaction not only allows direct manipulation of the visual objects displayed but also allows users to select what to be displayed (Card et al., 1999)
- Shneiderman (1996) summarizes six types of interface functionality
 - Overview
 - Zoom
 - Filtering
 - Details on demand
 - Relate
 - History

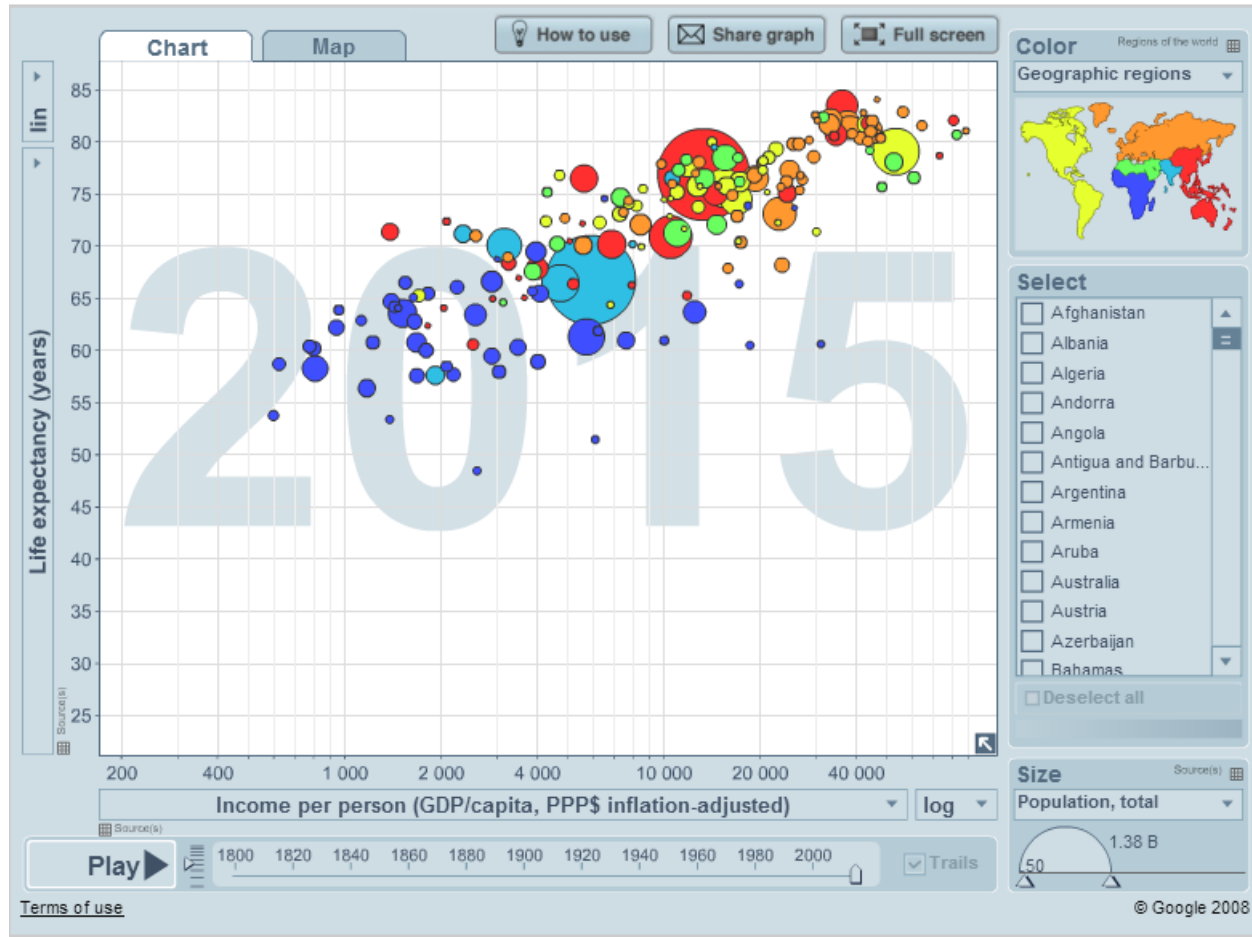
Overview first,
zoom and filter,
then details-on-demand.

TWO INTERACTION APPROACHES

User-interface interaction

- **Overview + detail**
 - First overview provides overall patterns to users; then details about the part of interest to the use can be displayed. [Card et al. 1999]
 - Spatial zooming & semantic zooming are usually used
- **Focus + context**
 - Details (focus) and overview (context) dynamically on the same view. Users could change the region of focus dynamically.
 - Information Landscape [Andrews 1995]
 - Cone Tree [Robertson et al. 1991]
 - Fish-eye [Furnas 1986]

INTERACTIVE VISUALIZATION EXAMPLE



<http://www.gapminder.org/world/>

DESIGN PRINCIPLES

Following design principles can help increasing the quality of visualizations in multiple ways.

- Effective information presentation
- Avoiding distraction and confusion, reducing cognitive load
- Adding aesthetic value

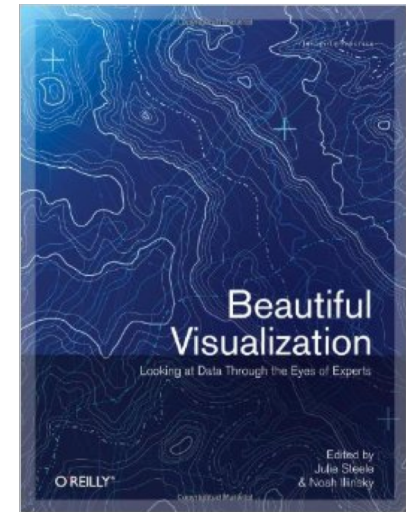
Some selected principles:

- Beauty of Visualization
- Data-Ink Ratio
- Data Density
- Color Heuristics
- Fitts's Law

BEAUTY OF VISUALIZATION

In *Beautiful Visualization* [2010], four criteria were raised to define the “beauty” of a visualization:

- **Informative**
 - Successfully convey information
- **Efficient**
 - Simple, focused, clear and straightforward
- **Aesthetic**
 - Use axes, layout, shape, colors, lines, and typography appropriately
- **Novel**
 - Be creative and attract readers with new design



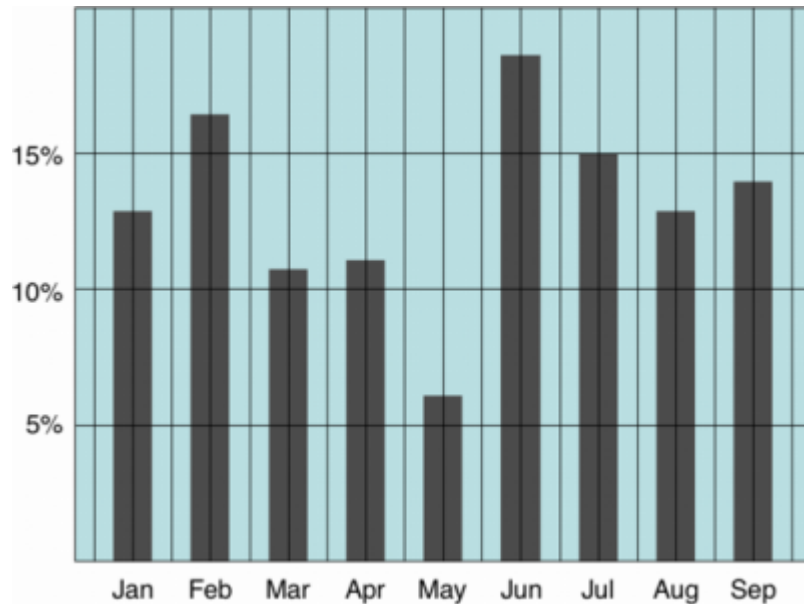
DATA-INK RATIO

Tufte [1983] claims that good graphical representations **maximize data-ink** and erase as much non-data-ink as possible.

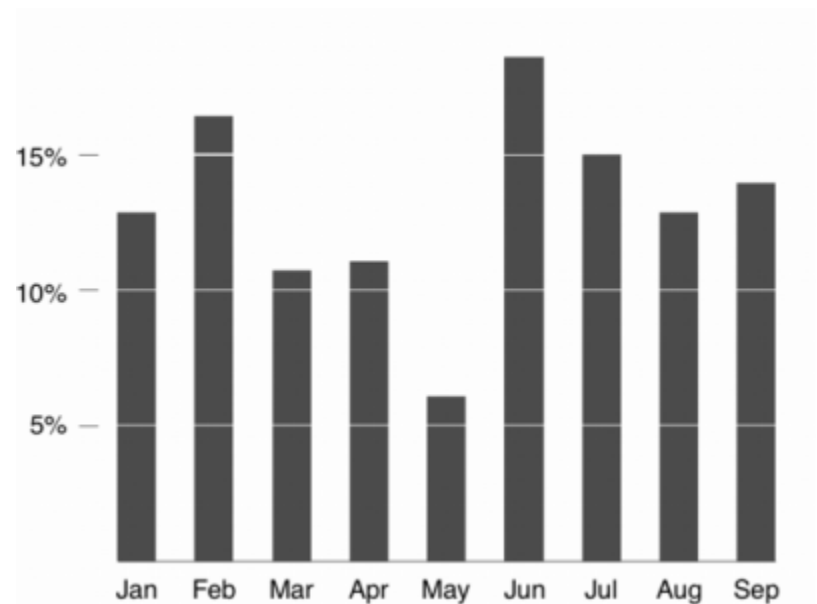
- Non-data-ink: Scales, borders, ...

$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{Data-ink}}{\text{Total ink used to print the graphic}} \\ &= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information} \\ &= 1.0 - \text{proportion of a graphic that can be erased}\end{aligned}$$

DATA-INK RATIO



Low Data-Ink Ratio



High Data-Ink Ratio

DATA DENSITY

The **data density** of a graph is the proportion of the total size of the graph that is dedicated to displaying data [Tufte 1983]

- Similar to Data-Ink Ratio, Tufte prefers visualizations with high Data Density.
- He claims that most graphs can be shrunk way down without losing legibility or information.

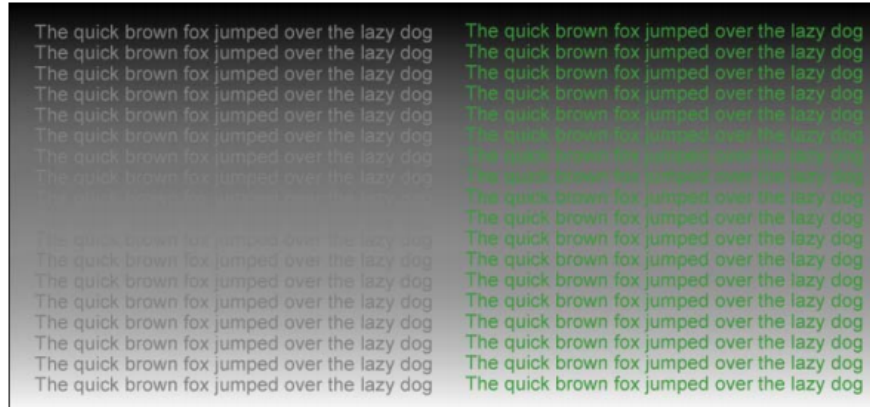
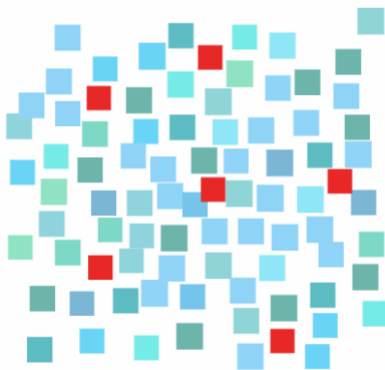
COLOR HEURISTICS

Contrast and **analogy** are the principles that define color design.

- Contrasting colors are different, analogous colors are similar.
- Contrast draws attention, analogy groups.

Color-coded information should be legible.

- Legibility: difference between foreground and background



https://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf

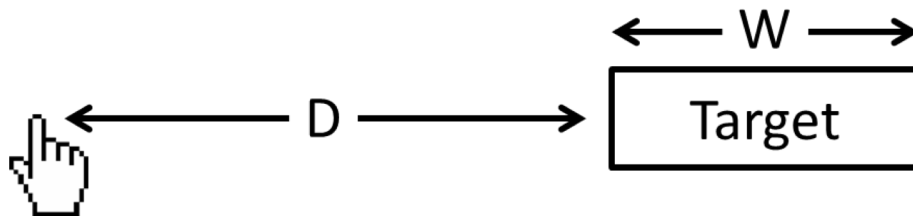
http://cdn2.hubspot.net/hub/111084/file-708877165-pdf/docs/ebooks/eBook-UX-Color-Theory_Applying-Color-Knowledge-to-Data-Visualization.pdf

FITTS' LAW

Fitts described a model of human movement primarily used in HCI and ergonomics. It may apply to interactive visualization design.

$$MT = a + b \cdot ID = a + b \cdot \log_2 \left(\frac{2D}{W} \right)$$

- MT: Movement time
- ID: Index of difficulty
- D: Distance
- W: Width



Fitts' Law indicates that, to shorten the movement time and lower the index of difficulty, an HCI designer should:

- **Shorten the distance to the target**
 - Group functions/buttons together to reduce redundant movements
- **Enlarge the target**

VISUALIZATION LIBRARIES



D3.JS



JavaScript library for manipulating documents and visualizations

- Made interactive via JS, CSS, SVG

Not a new visualization language – built entirely on functionality available in browsers

- So, can require up-to-date browsers ...

```
# Resize pre-defined circle element
d3.selectAll("circle").transition()
  .duration(750)
  .delay(function(d, i) { return i * 10; })
  .attr("r", function(d) {
    return Math.sqrt(d * scale);
  });
```

MATPLOTLIB/PYPLOTL

Standard plotting library for Python

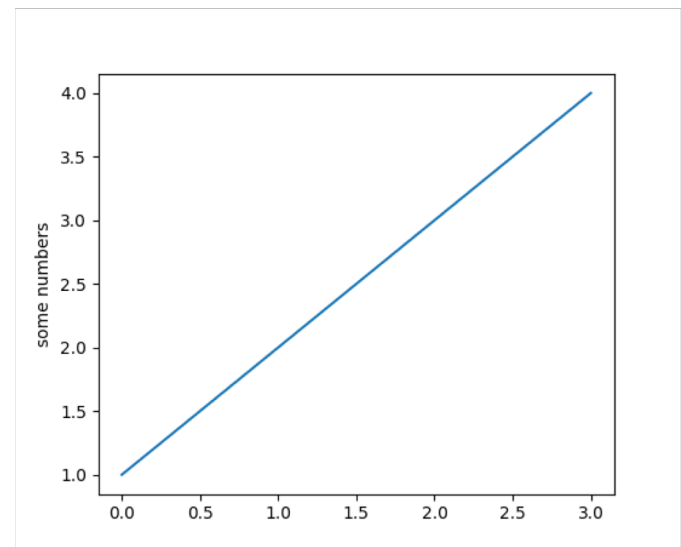
- Meant to mimic Matlab
- Built on Numpy



Quite powerful, but also fairly complex and cludgy

Rendering code serves as backend for newer viz packages

```
# Make a line plot
import matplotlib.pyplot as plt
plt.plot([1,2,3,4])
plt.ylabel('some numbers')
plt.show()
```



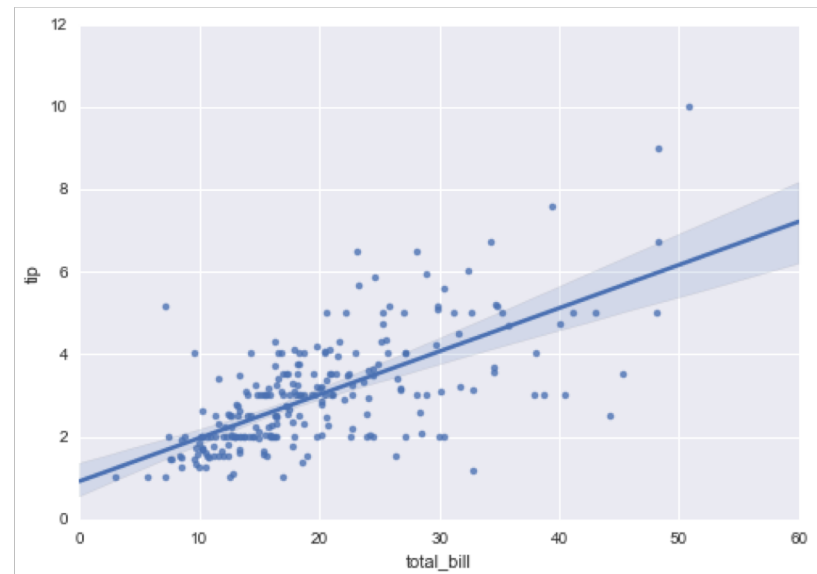
SEABORN

Complements Matplotlib; it's built on top!

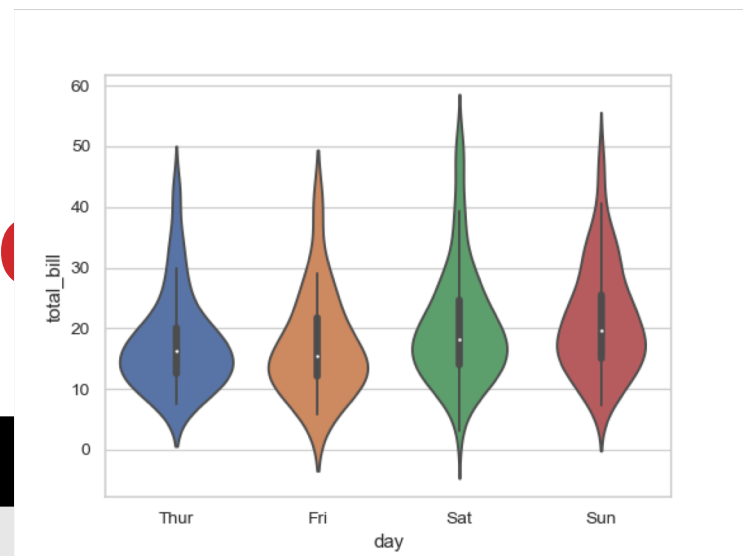
- **Much** better defaults for aesthetics
- Closer ties to numpy, pandas, scipy
- More advanced functionality built in, like clustering, matrix visualization, grids of plots, etc

```
import seaborn as sns
# ... load some data ...

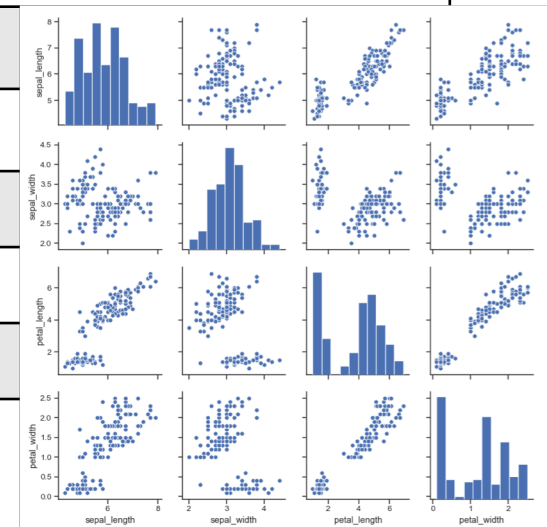
sns.regplot(
    x="total_bill",
    y="tip",
    data=tips);
```



SEABORN: TYPES



Module	Description of module
distplot	Standard histogram
barplot	Estimate of central tendency for a numeric variable
violinplot	Similar to boxplot, also shows the probability density of the data
jointplot	Scatterplot
regplot	Regression plot
pairplot	Pairplot
boxplot	Boxplot
swarmplot	Categorical scatterplot
factorplot	General categorical plot



GGPLOT



Data visualization package for R

- Based on Leiland's "grammar for graphics" paradigm that separates visualization into different semantic components, like layers, scales, etc
- Easily the most used visualization package for R
- Big influence on other plotting libraries and industry (e.g., Leiland went on to a VP role at Tableau)

A well-maintained adaptation for Python is available:

- <https://github.com/yhat/ggpy>
- Works well with Numpy/Pandas stack

UP NEXT:
**STATISTICS &
HYPOTHESIS TESTING**

STATISTICAL INFERENCE

Statistical inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

- Process of going from the world to the data, and then back to the world
- Often the goal is to develop a statistical model of the world from observed data

Conclusion is typically:

- an estimate;
- or a confidence interval;
- or rejection of a hypothesis
- or clustering or classification of data points into groups

BASIC PROBABILITY I

Probability is concerned with the outcome of a trial (also called experiment or observation)

Sample Space: Set of all possible outcomes of a trial Probability of Sample Space = 1

Event is the specification of the outcome of a trial

- For example: Trial = Tossing a coin; Sample Space = {Heads, Tails}; Event = Heads

If two events E and F are independent, then: Probability of E does not change if F has already happened = $P(E)$, i.e., $P(E | F) = P(E)$

Also: $P(E \text{ AND } F) = P(E) * P(F)$

If two events E and F are mutually exclusive, then: $P(E \text{ UNION } F) = P(E) + P(F)$

BASIC PROBABILITY II

Bayes Theorem $P(A | B) = P(B | A) * P(A) / P(B)$

Simple equation, but fundamental to Bayesian inference

Conditional Independence: A and B are conditionally independent given C if: $\Pr(A \text{ AND } B | C) = \Pr(A | C) * \Pr(B | C)$

Powerful in reducing the computational efforts in storing and manipulating large joint probability distributions

Entropy A measure of the uncertainty in a probability distribution

[Wikipedia Article](#)