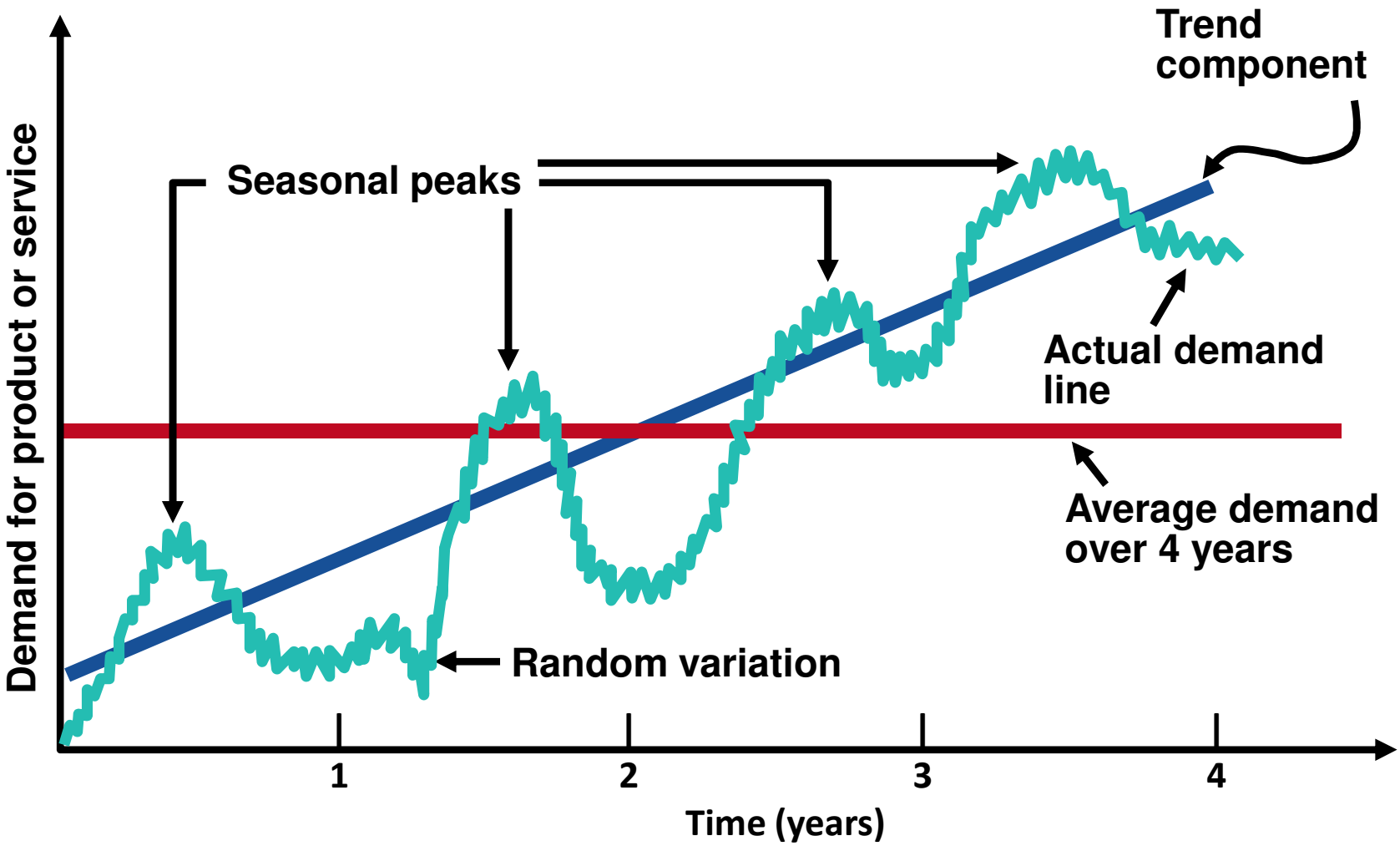
Data Visualization and Visual Analysis

Recap from last week

Historical patterns



Recap from last week: Seasonality

- Step 1. Calculate the average demand for each year
- Step 2. Calculate seasonal indexes
- Step 3. Average the indexes
- Step 4. Forecast demand for the next year
- Step 5. Multiple next year's average seasonal demand by each average seasonal index

Quarter	Year 1	Seasonal Index	Year 2	Seasonal Index	Avg. Index	Year3
Fall	24000	**1.2**	26000	**1.24**	**1.22**	**26840**
Winter	23000	**1.15**	22000	**1.05**	…	**…**
Spring	19000	**0.95**	19000	…		
Summer	14000	**0.7**	17000	…		
Average	**20000**		**21000**			**22000**

Recap from last week

- Collaborative filtering: recommend items by only **users' past behavior**
- Advantage: we don't need to have additional information about the users or content of the items
  - Users' **rating** or **purchase history** is the only information that is needed to work
- Input: Rating matrix
  - Users rate (rank) items (purchased, watched)
  - Explicit ratings: entered by a user directly
  - Implicit ratings: inferred from other user behavior
    - e.g. the amount of time users spent on a webpage
    - e.g. the number of times users listen to a song

Movie	Alice (1)	Bob (2)	Carol (3)	Dave (4)
Love at last	5	5	0	0
Romance forever	5	?	?	0
Cute puppies of love	?	4	0	?
Nonstop car chases	0	0	5	4
Swords vs. karate	0	0	5	?

3803ICT course structure

W1. Introduction to Data Analytics

Data Preparation and Preprocessing

W2. Data Preparation and Preprocessing

Data Analysis and Interpretation

W3. Exploratory Data Analytics

W4. Statistical Data Analytics

W5. Predictive Data Analytics

Visualization

W6. Data Visualization

Analysis of special types of data

W7. Time Series

W8. Textual Data

W9. Graph Data

Analysis with big data infrastructure

W10. Distributed Data Analysis

W11. Cloud-based Data Analysis

W12. Revision

Learning Objectives

At the end of this lecture, you should be able to:

- Identify **building blocks** and **principles** of data visualization (OPTIONAL)
- Devise techniques in
  - **spatial data** visualisation
  - **text data** visualisation
- Apply the principles of **dashboard** design for visual analytics
- Handle Big Data

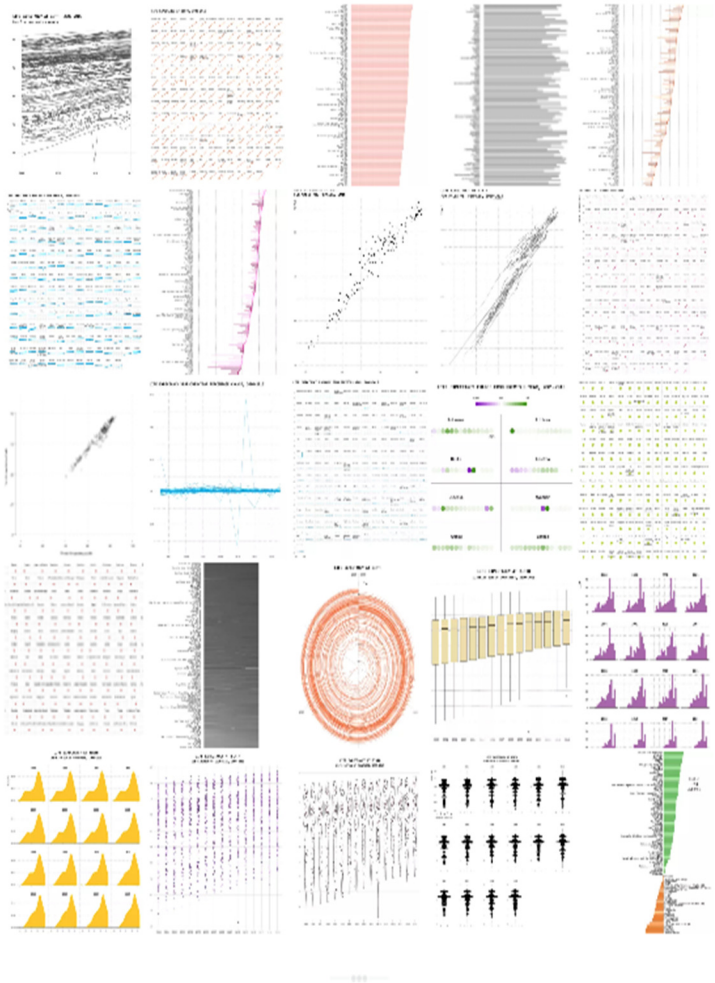
Data Visualization

- "*... finding the artificial memory that best supports our natural means of perception.*" [Bertin 1967]

- "*Transformation of the symbolic into the geometric*" [McCormick et al. 1987]

- "*The use of computer-generated, interactive, visual representations of data to amplify cognition.*" [Card, Mackinlay, & Shneiderman 1999]

One Dataset, Visualized 25 Ways

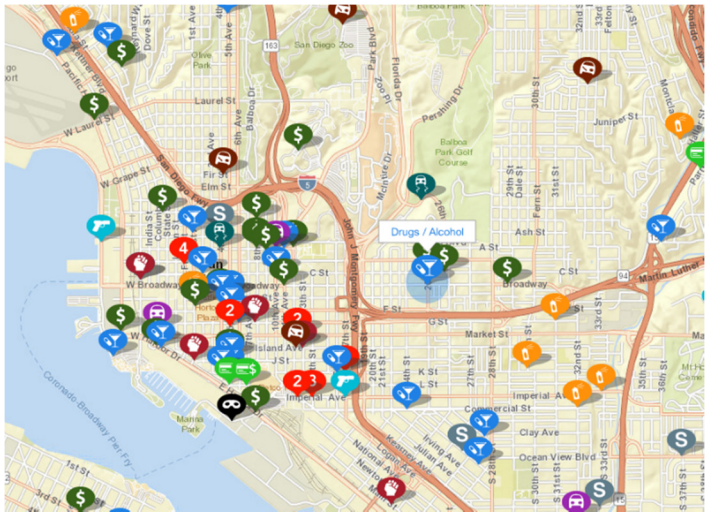


"You must help the data focus and get to the point. Otherwise, it just ends up rambling!"

http://flowingdata.com/2017/01/24/one-dataset-visualized-25-ways/

From Data Visualization to Visual Analytics

- Map-based analytics, such as CrimeMapping



https://www.crimemapping.com/map

- Interactive Education
  - The famous Gapminder Video, Hans Rosling: 200 Countries, 200 Years, 4 Minutes. https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo

- Future of Journalism: e.g. NY Times
  - NY Times Interactive Visualizations (recession/recovery 2014). http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html
  - And 2014 “the year in interactive storytelling”. http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html?_r=0

Outline

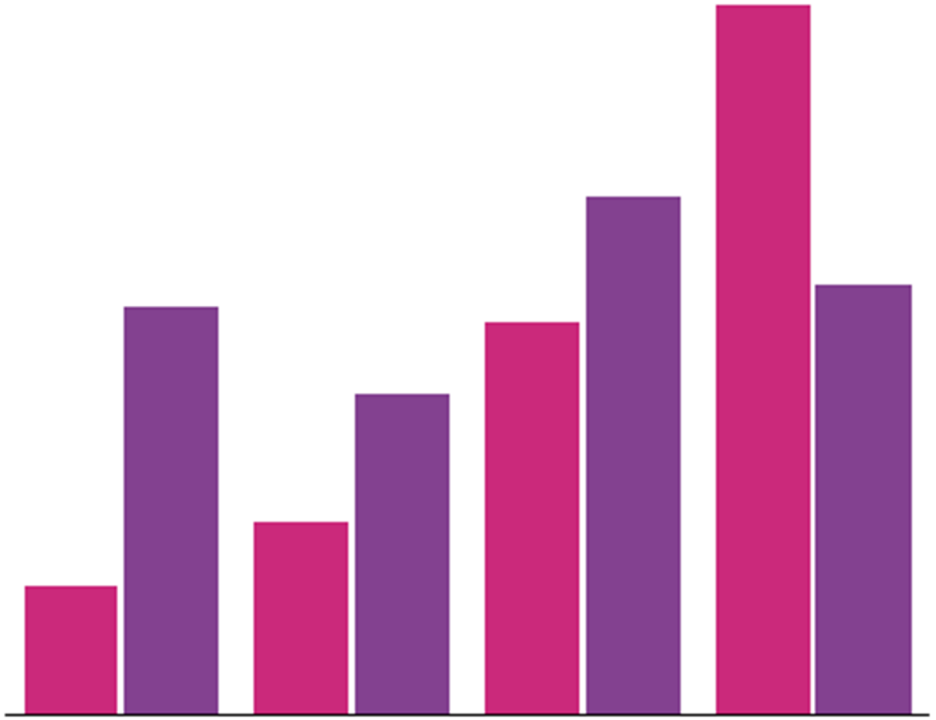
I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
   - Spatial Data Visualization
   - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

I. Building blocks of a data visualization

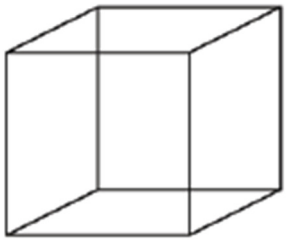
- Every data viz can be broken down into combinations of building blocks:
  - Graphical elements
  - A way of controlling their appearance

Graphical elements

- **Marks**: basic geometric elements to represent items or links
- **Channels**: visual variable, change the appearance of marks based on attributes



Marks for items

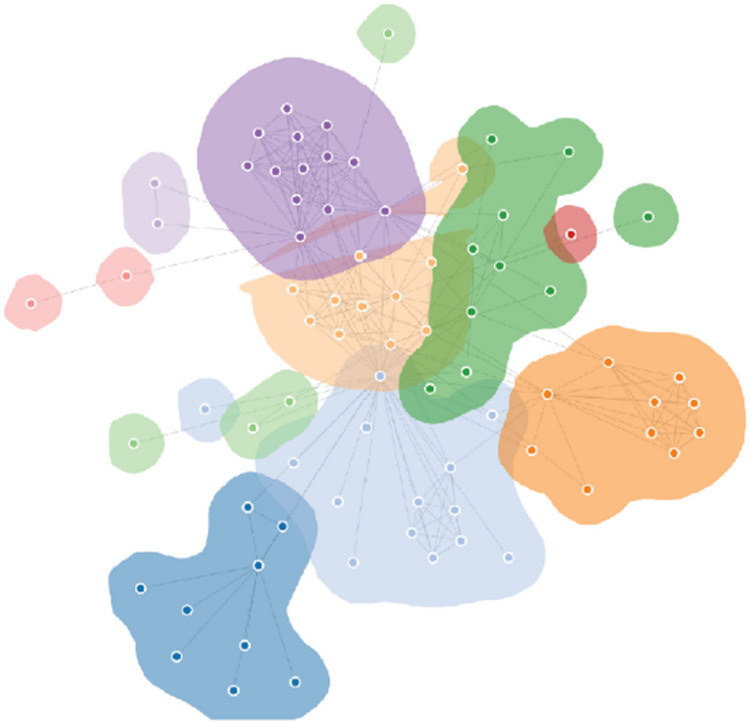
- Points (OD):  sense of place
- Lines (1D):  sense of length and direction
- Areas (2D):  sense of space and scale
- Objects (3D):  sense of volume

Marks for links

- Containment (enclosure):

- Connection:

Example: communities



http://p3.johnathanmercer.com/

Channel types

- Identity channels (What?):
  - E.g. **Categorical** (= nominal)

- Magnitude channels (How much?):
  - e.g. Ordered
    - **Ordinal**

    - **Quantitative** (= numeric)

Refresher

- Categorical (= nominal)
  - hair colour
- Ordinal
  - Completely disagree, disagree, neutral, agree, completely agree
- Quantitative (= numeric)
  - Discrete
    - number of students, months in year
  - Continuous
    - weight, time period

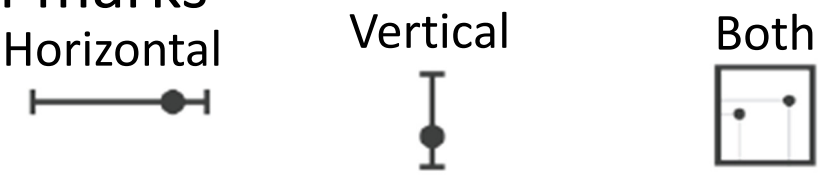
Channel properties

Property	Description
Selectivity	Can we spot the difference between marks?
Associativity	Can we group similar marks together?
Quantitativity	Can we measure the difference between two marks?
Ordering	Can we order marks?
Drawability count	How many unique marks can we make?

Channels (visual attributes)

Control appearance of marks

- Position:

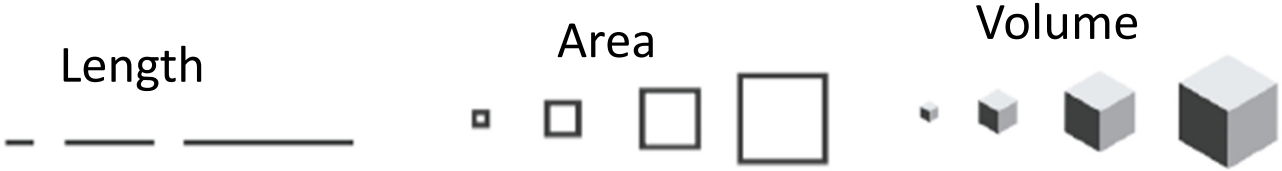


- Color:

- Shape:

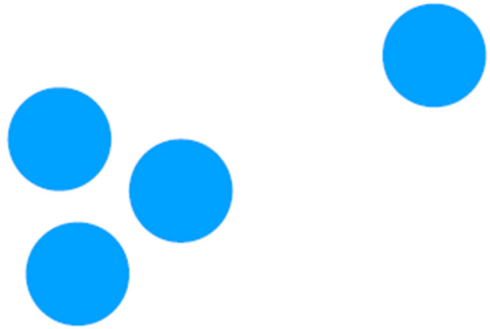
- Tilt/Angle:

- Size:



Position

- Pros:
  - Strongest channel
  - Works for all data types
- Cons:
  - Cluttering
  - Not always available (spatial dataset)



Selectivity	yes
Associativity	yes
Quantitativity	yes
Ordering	yes
Drawability count	huge

Color

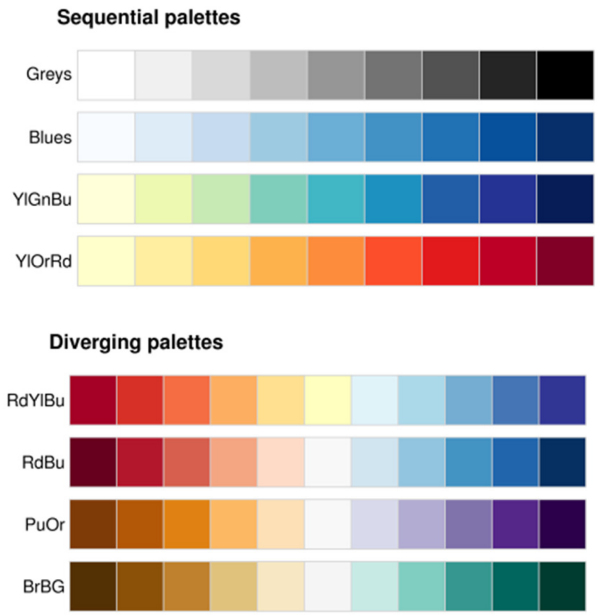
- Pros:
  - Good for categorical data (identify channel)
- Cons:
  - Limited number of classes (7-10)
  - Doesn't work for ordered data
  - Color maps are hard to design

Selectivity	yes
Associativity	yes
Quantitativity	no
Ordering	no
Drawability count	low

Luminance, saturation

- Pros:
  - OK in addition of position, length and size

- Cons:
  - Many shades are not easily distinguishable

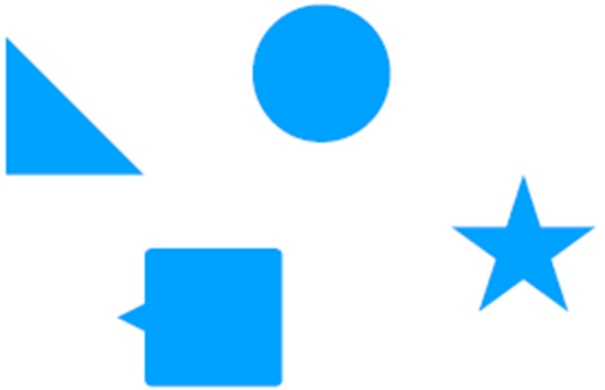


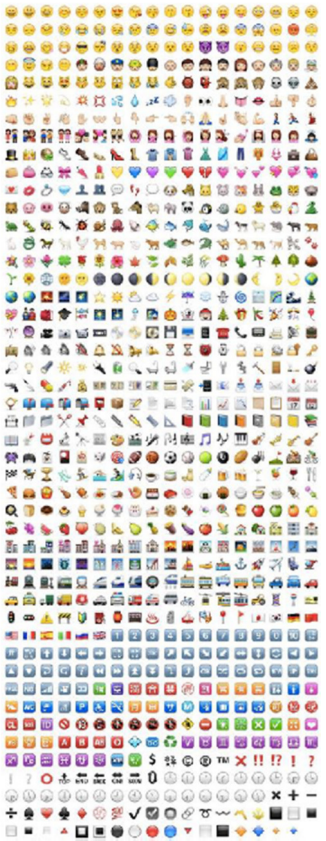
https://betterfigures.org/2015/06/23/picking-a-colour-scale-for-scientific-graphics/

Selectivity	yes
Associativity	yes
Quantitativity	OK-ish
Ordering	yes
Drawability count	low

Shape

- Pros:
  - Excellent to recognize many classes
- Cons:
  - No grouping
  - No ordering

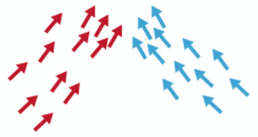




Selectivity	yes
Associativity	low
Quantitativity	no
Ordering	no
Drawability count	high

Tilt/Angle

- Pros:
  - Good for some special cases
  - Can be used to augment other channels
- Cons:
  - Limited use-cases
  - Medium properties:
    - not clear for all people
    - might be a mess for some data



wind direction

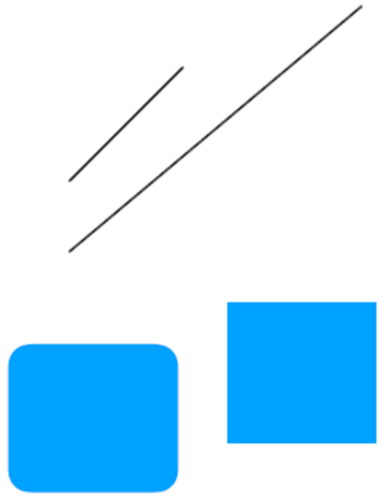
complex mark with angle

pair-wise similarity

Selectivity	medium
Associativity	medium
Quantitativity	medium
Ordering	medium
Drawability count	medium

Length, area, volume

- Pros:
  - Easy to spot the longest line
  - OK to find the largest area
- Issues
  - Hard to find the biggest volume

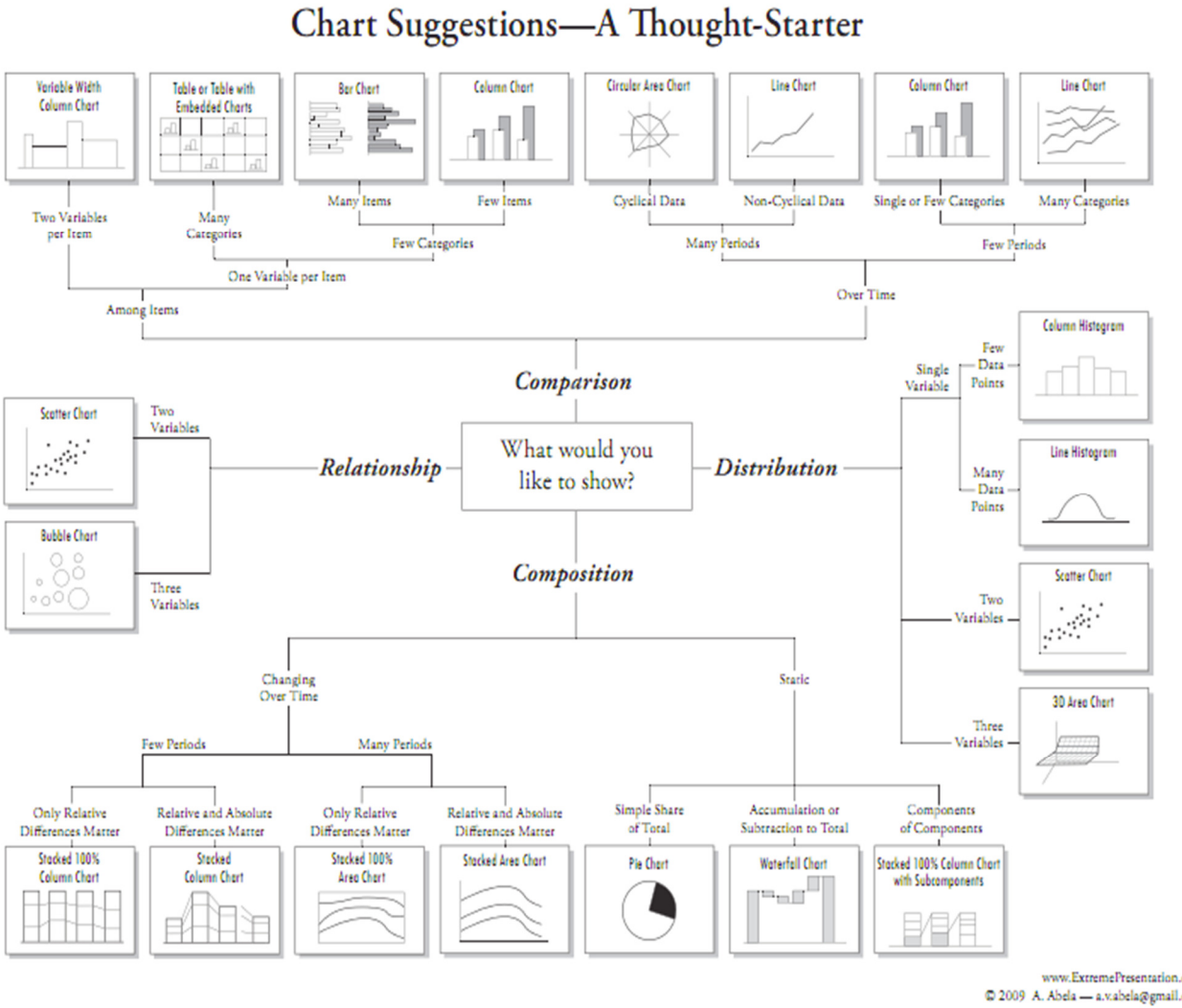


1D length

Selectivity	yes
Associativity	yes
Quantitativity	yes
Ordering	yes
Drawability count	high

Charts and Graphs (revisit)

- https://www.tableau.com/solutions/gallery



Outline

I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
   - Spatial Data Visualization
   - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

II. Design principles to make it great

Great visualizations:

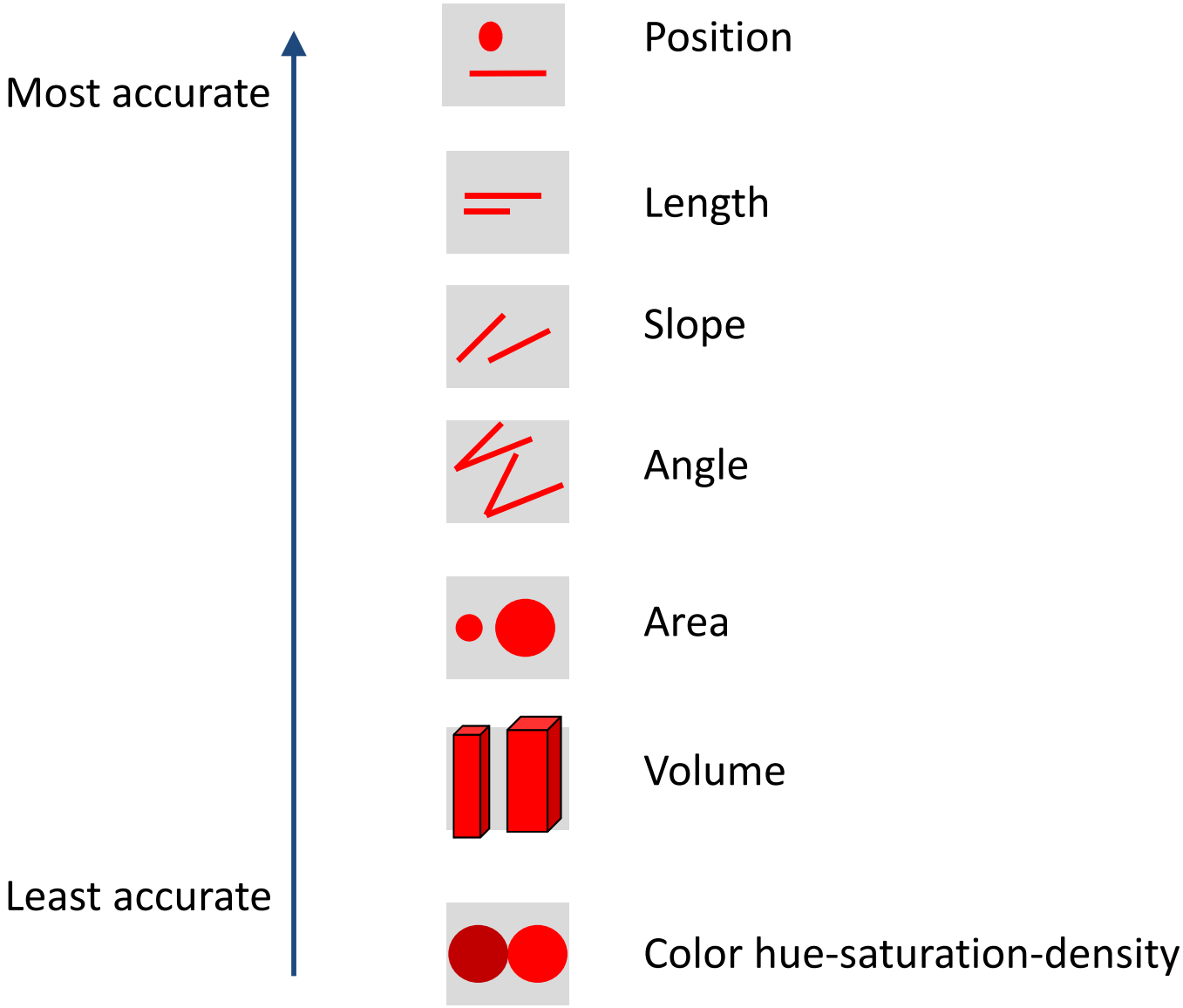
- Can turn data insights into action
- Goal of data visualization is impact, not numbers

1. Design principles:
   a) Metrics of success
   b) Guidelines for visual elements: Mark, Color, Size, Text
2. Create effective views
3. Great visualization checklist

1.a) Design principles: metrics of success

- **Expressiveness**: the visual encoding should only express the information contained in the dataset attributes
- **Effectiveness**: the importance of the attribute should match the noticeability of the channel.
  - Accuracy: How well can a user read the information in the channel?
  - Discriminability: How easily can we perceive differences between attribute levels?
  - Separability: Can we use one channel independently of another? Do they interfere?
- Influential factors:
  - Alignment
  - Distractors
  - Distance
  - Common scale

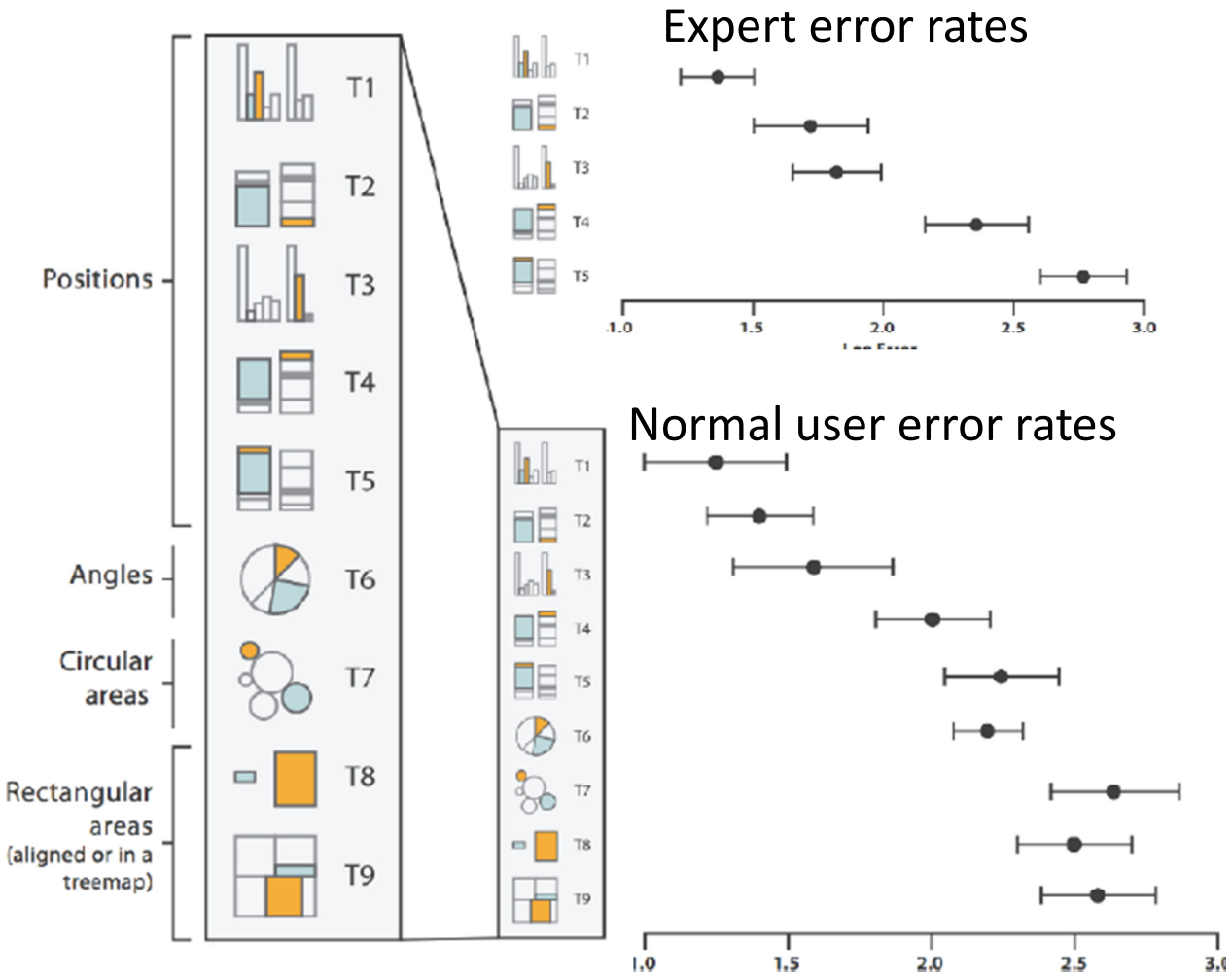
Accuracy



Cleveland, McGill (1984)

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

Discriminability



Expert error rates

Human perceives better with simple charts [5]

Normal user error rates

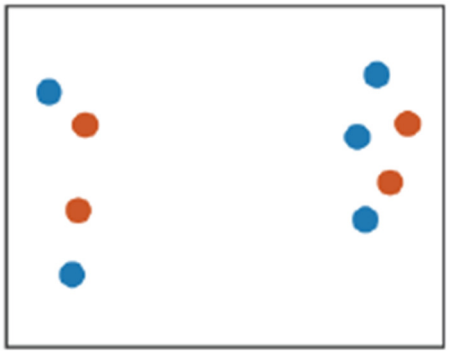
Human perceives better with simple charts [5]

Perception test

Separability of attributes

- According to Gestalt Psychology principles (1912):

Position + Color



Fully separable

Size + Color

Some interference

Width + Height

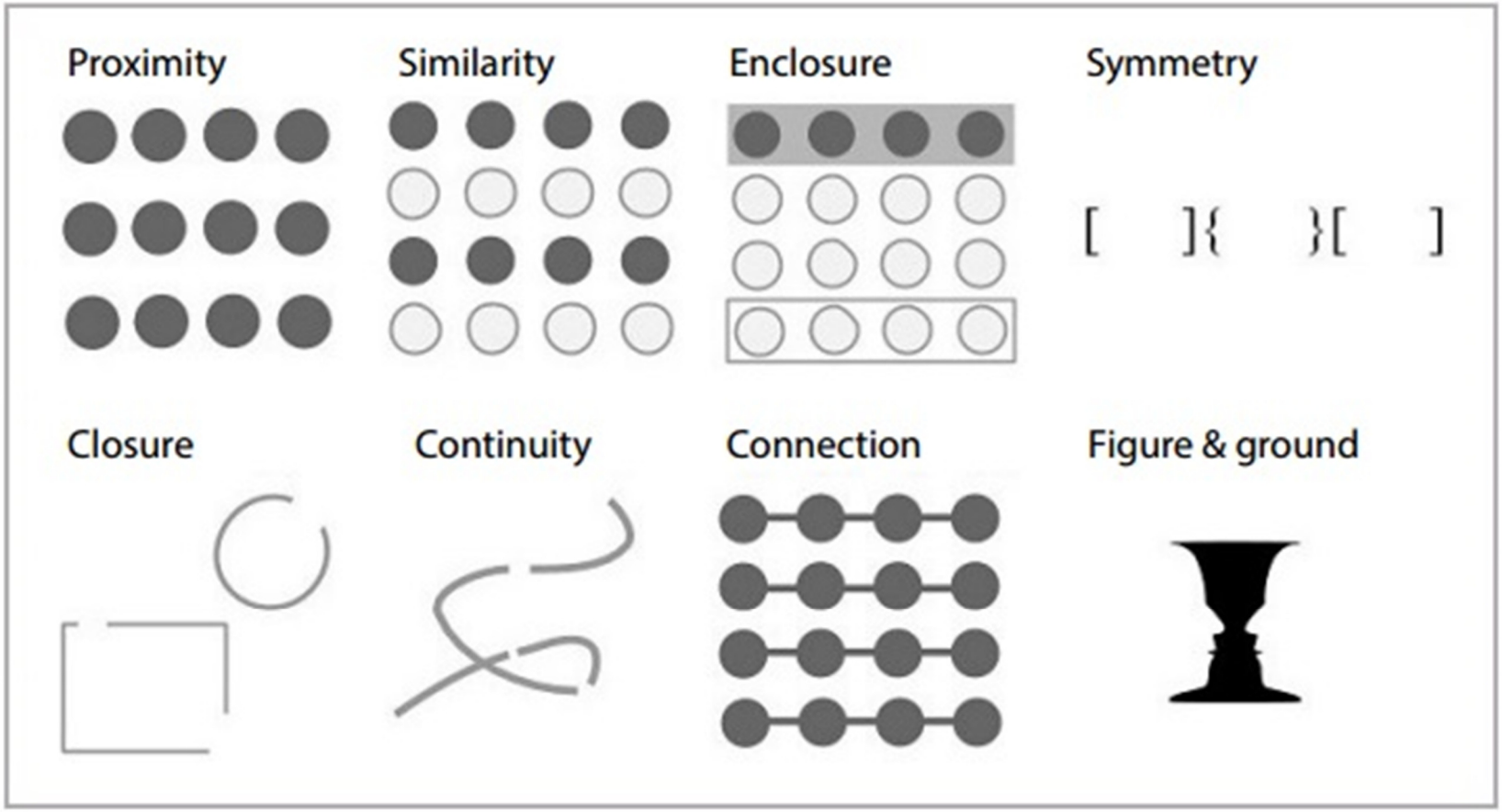
Some/significant inference

More colors

Major interference

1.b) Design principles: Guidelines

Mark (from Gestalt principles):

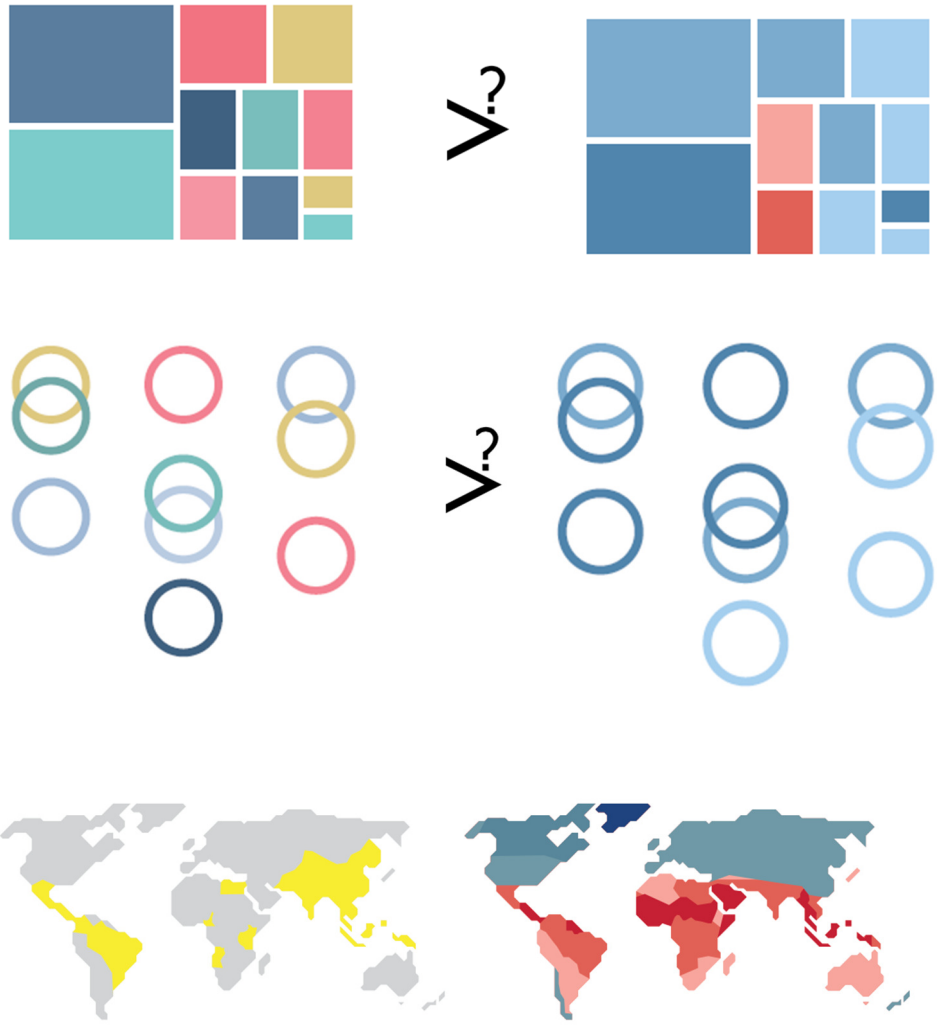


Source http://blog.fusioncharts.com/2014/03/how-to-use-the-gestalt-principles-for-visual-storytelling-podv/

Color

- Highlight specific insights or identify outliers
  - Differentiable:
    - Don't use similar colors, or too many colors
    - Don't re-use colors for different dimensions or measures on the same dashboard
  - Measurable:
    - Does the color scale match my data?
    - Does the color move from light to dark, or is it stepped to best represent what you're measuring?
  - Relatable:
    - Semantically-resonant colors help people process information faster
      - E.g. use yellow to depict bananas, red to represent heat

Depends on specific data!

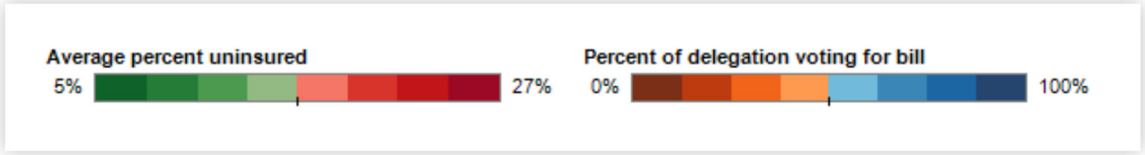


Color (cont'd)

- Choose colors based on the information you want to convey: Sequential, Diverging, Categorical

- Use online resources to discover and record your color schemes: Color Brewer, Kuler, Colour Lovers
  - Where possible, use your organization's palette

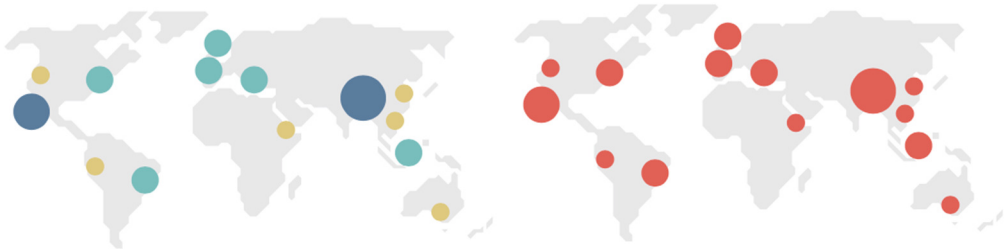
- Try to use no more than two color palettes. Use non-overlapping scales

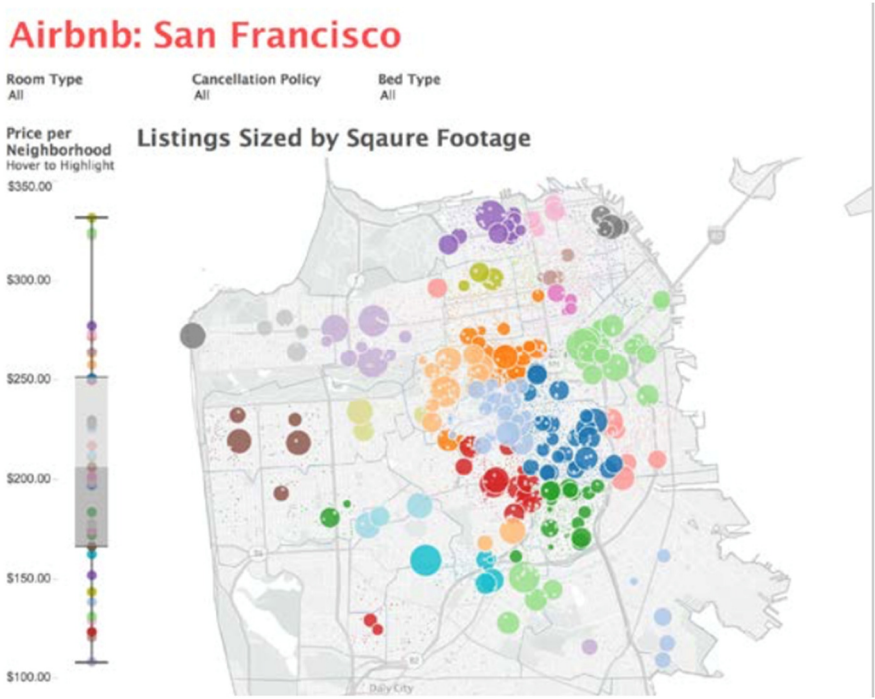


- Use light color for background

Size

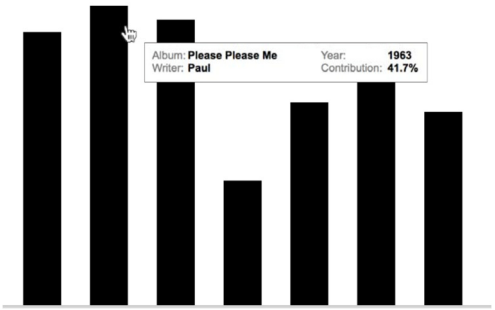
- Use size to draw emphasis to your key message
- E.g. bold shapes and colors might work well with bar charts and area charts

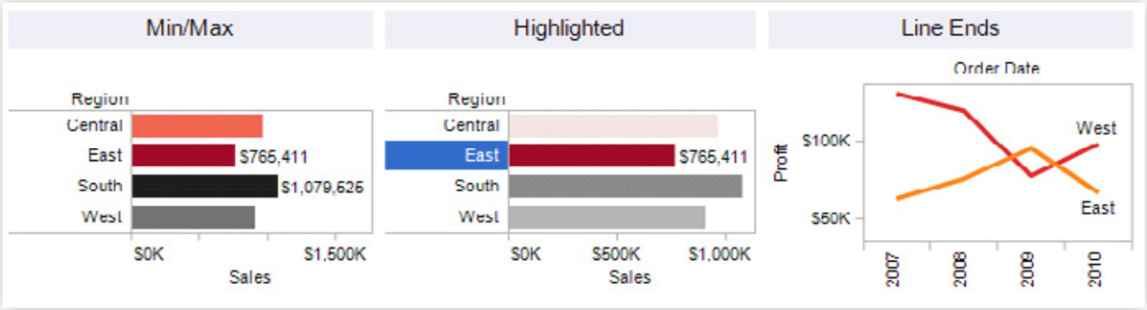




Text

- Readability is essential,
  - Make the most important information stand out
  - Use readable fonts
  - Use tooltips
- Titles: Keep them short, but powerful
- Labels: Find the sweet spot
- Axes: make intelligent axes



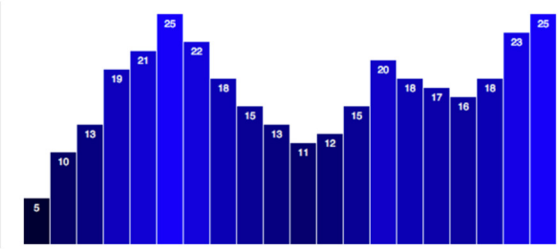


2. Create effective views

- Emphasize the most important data
- Orient your views for legibility
- Organize your views
- Avoid overloading your views
- Limit the number of colors and shapes in a single view

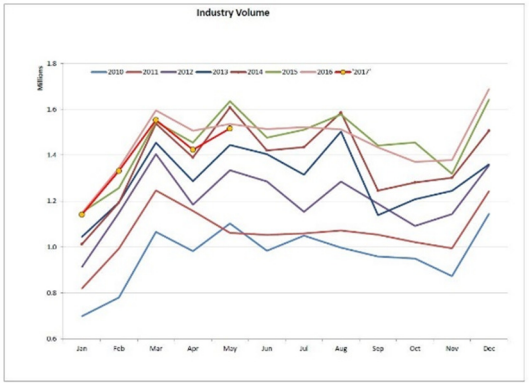
Bad practices

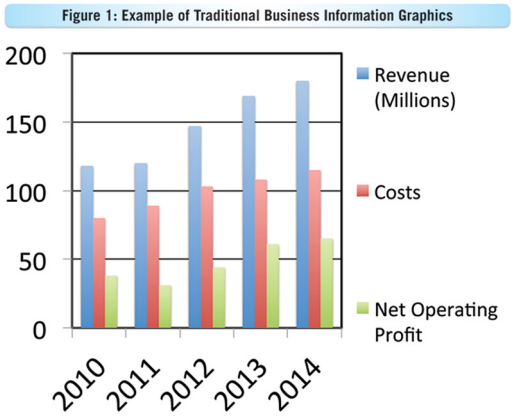
- Redundant encoding



Length + color

- Cluttered space + no clear relationships





- Too many colors

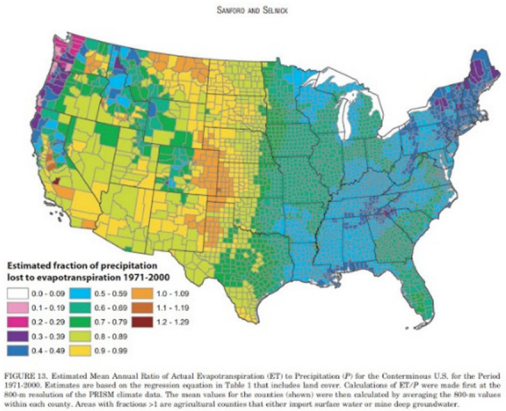
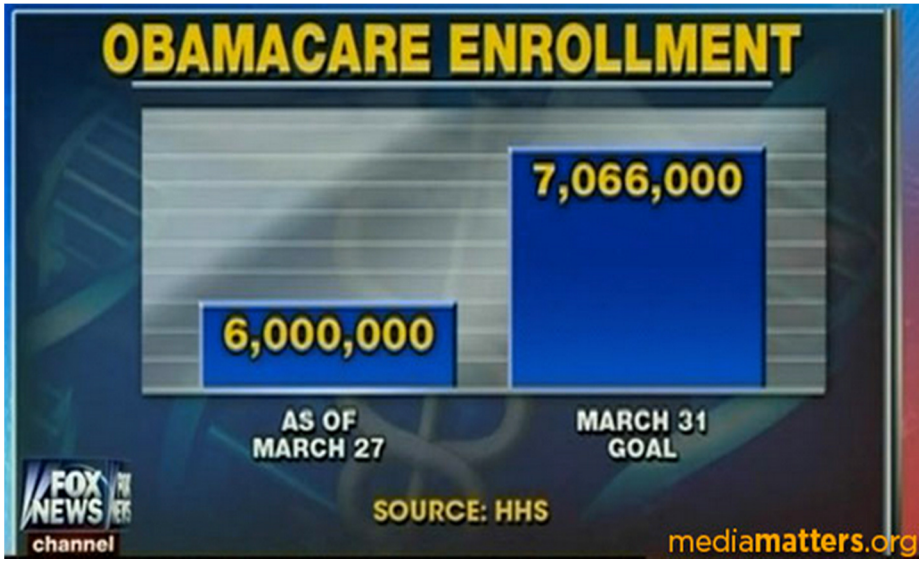


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

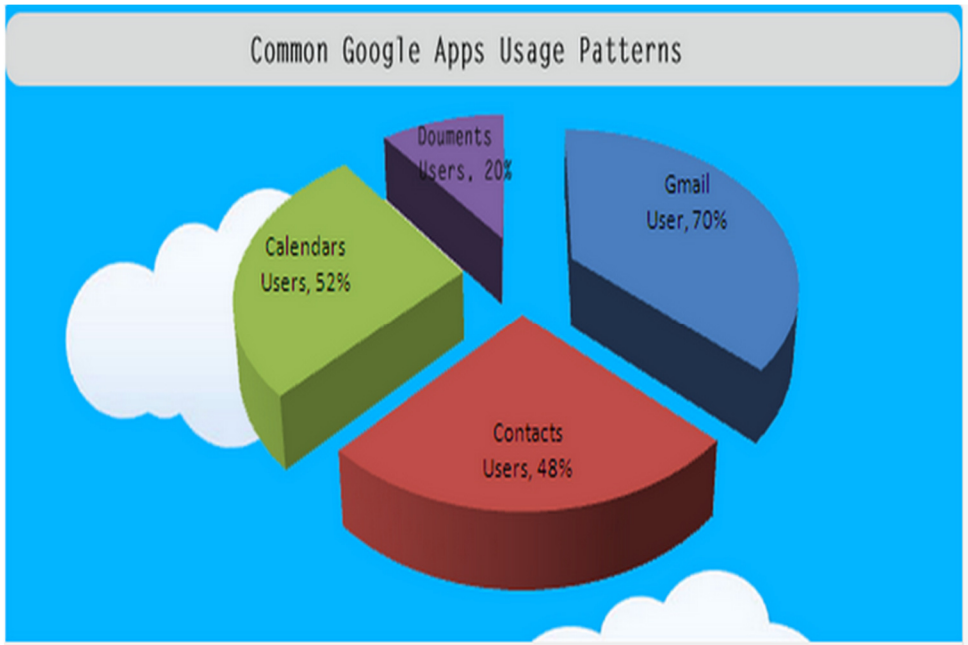
http://cliffmass.blogspot.ch/2013/06/evaporation-versus-precipitation-which.html

Bad practices (courtesy of viz.wtf)

- Misleading Scale



- Pie in the Sky?



Best practices

Line/Bar

Length
Position

1 quantitative attr.
+ 1 categorical attr.

Point

Position

2 quantitative attr.

Point

Position
Color

2 quantitative attr.
+ 1 categorical attr.

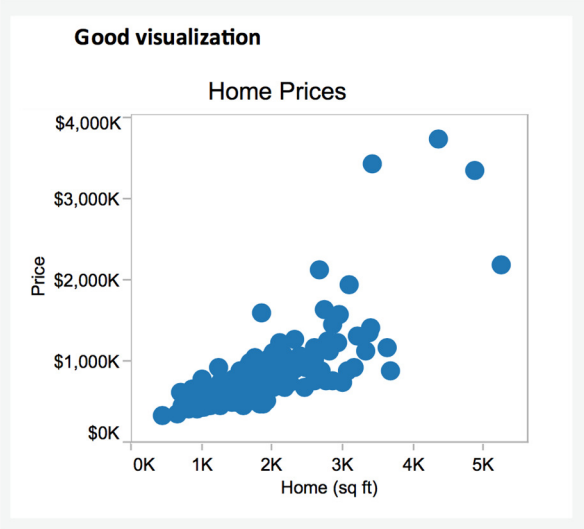
Point

Position
Color
Size

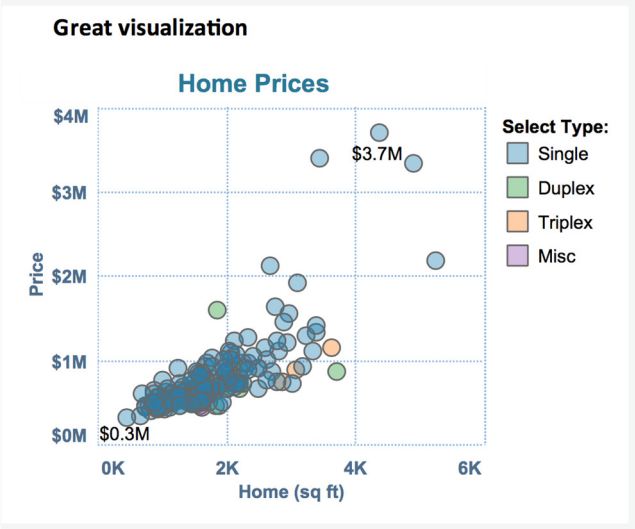
3 quantitative attr.
+ 1 categorical attr.

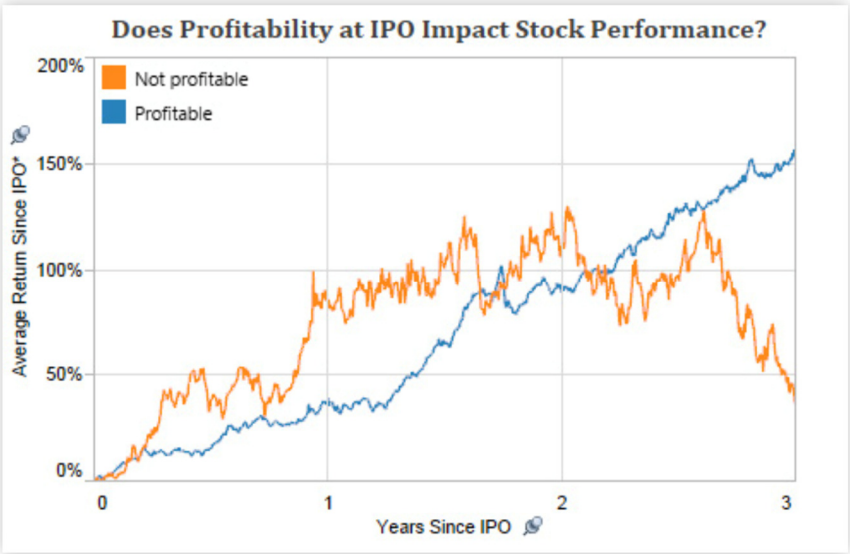
Best Practices

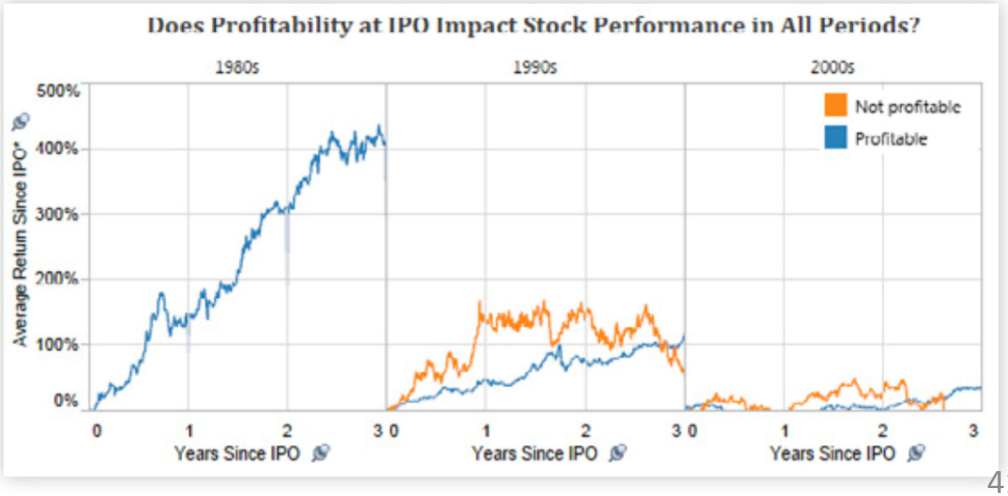
Good



Great

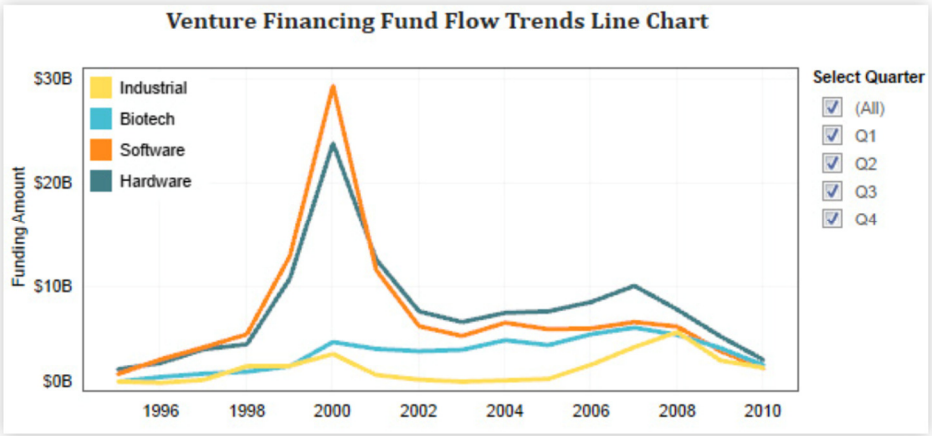




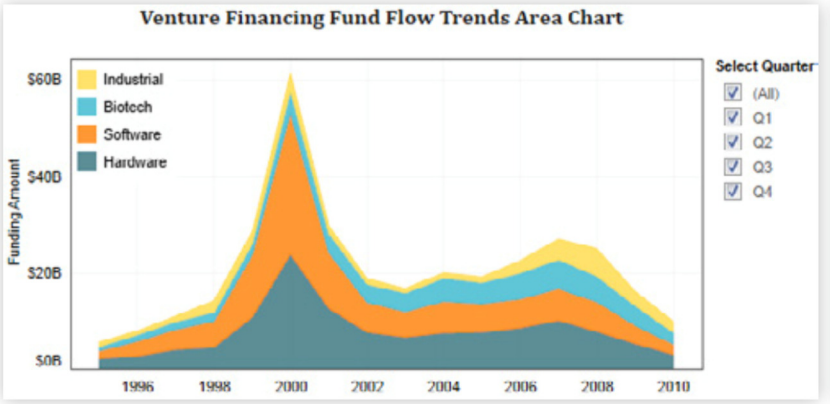


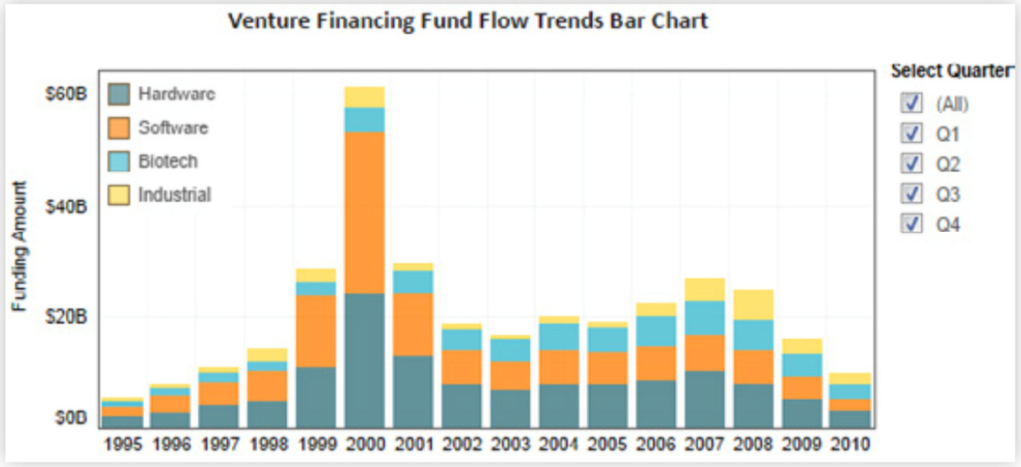
Best Practices

Good



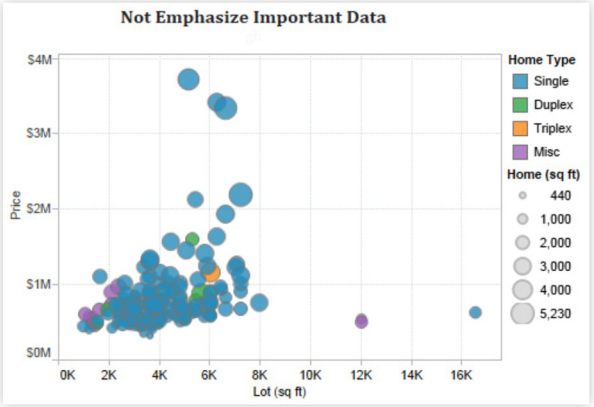
Great

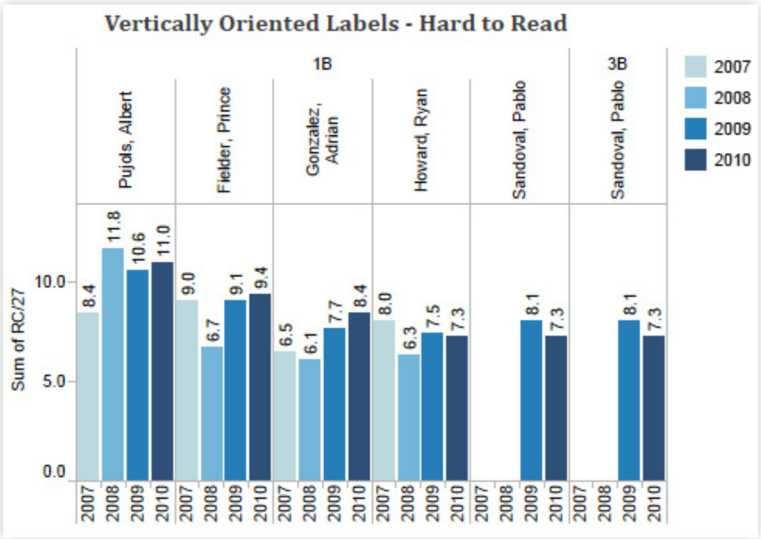




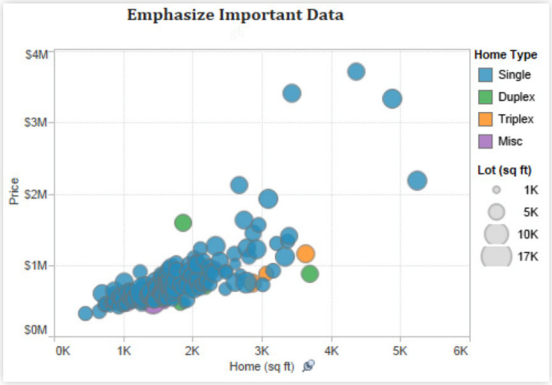
Best Practices

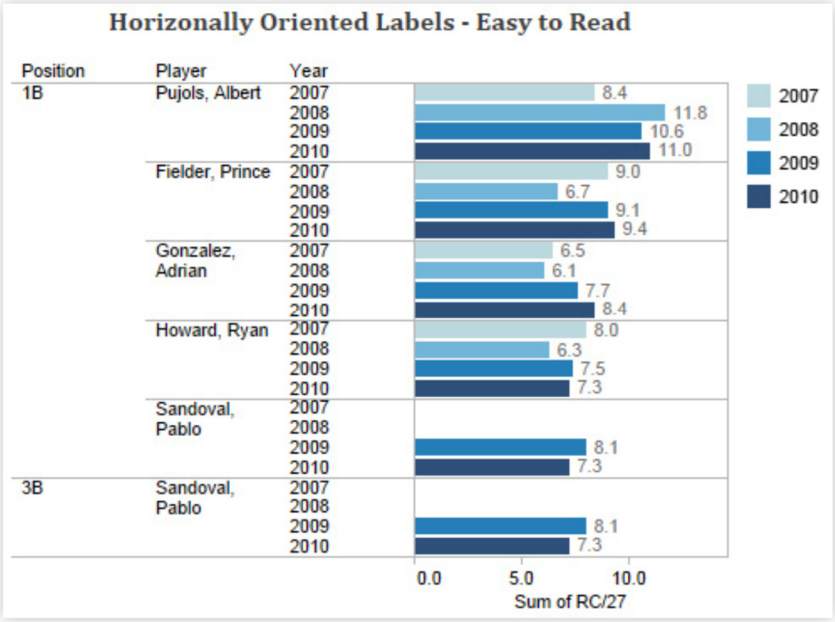
Good





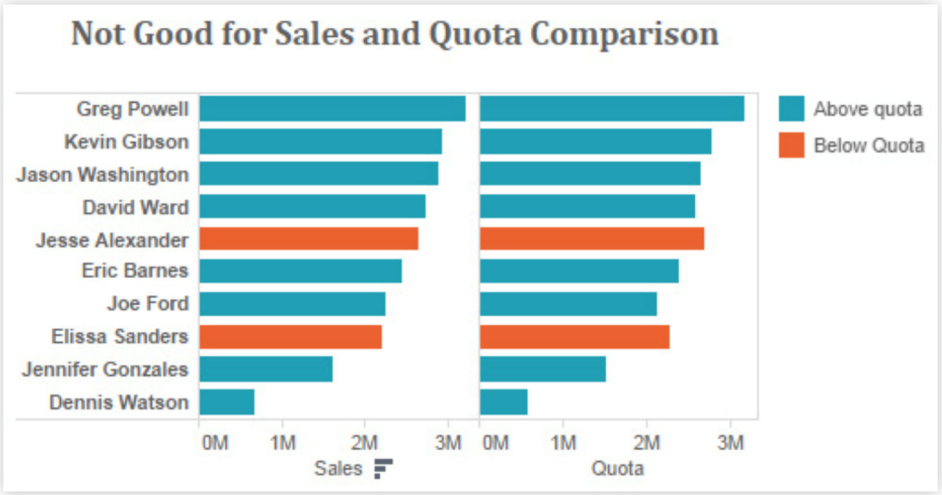
Great

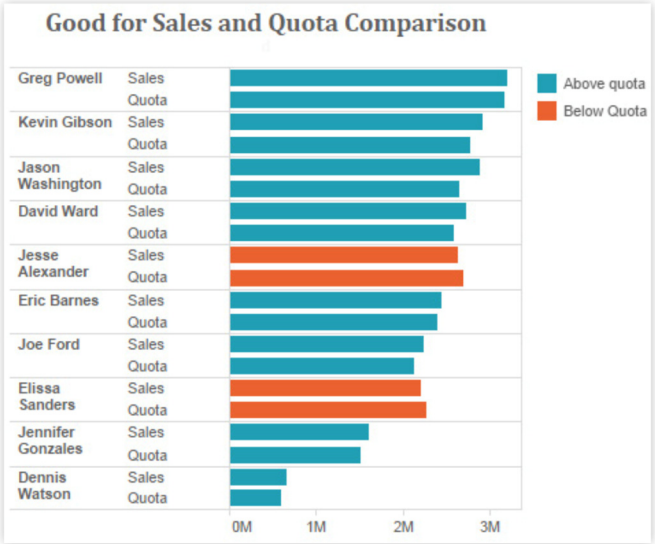




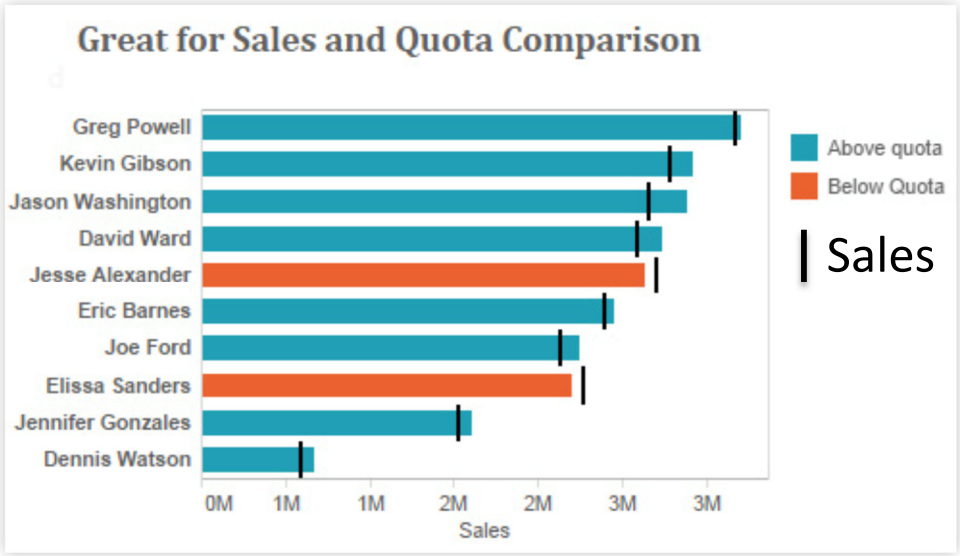
Best Practices

Good



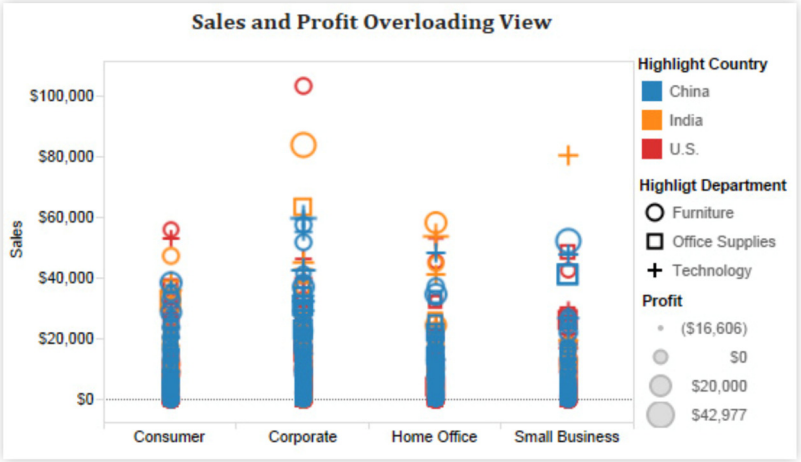


Great

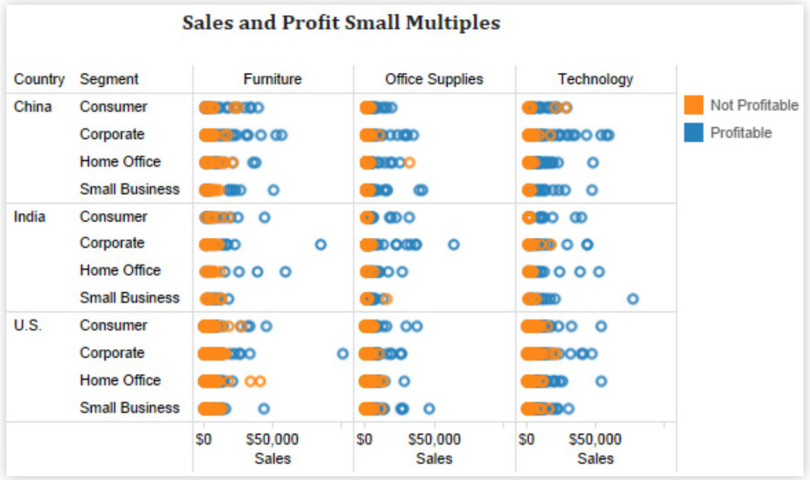


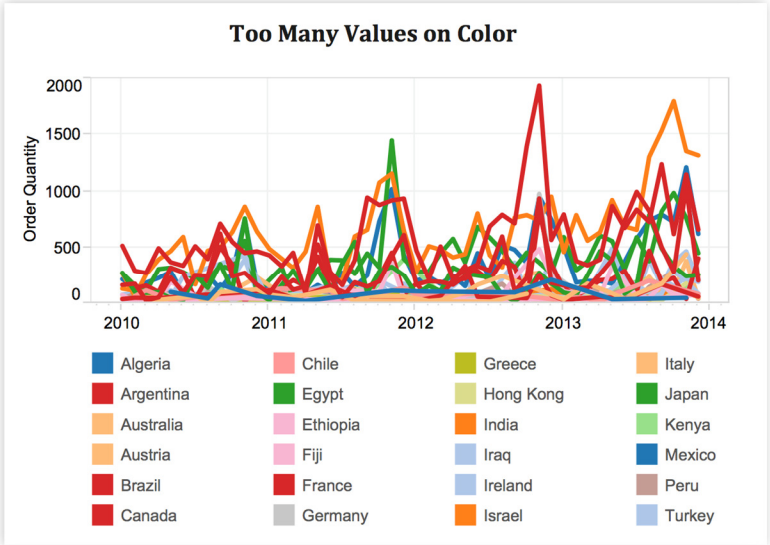
Best Practices

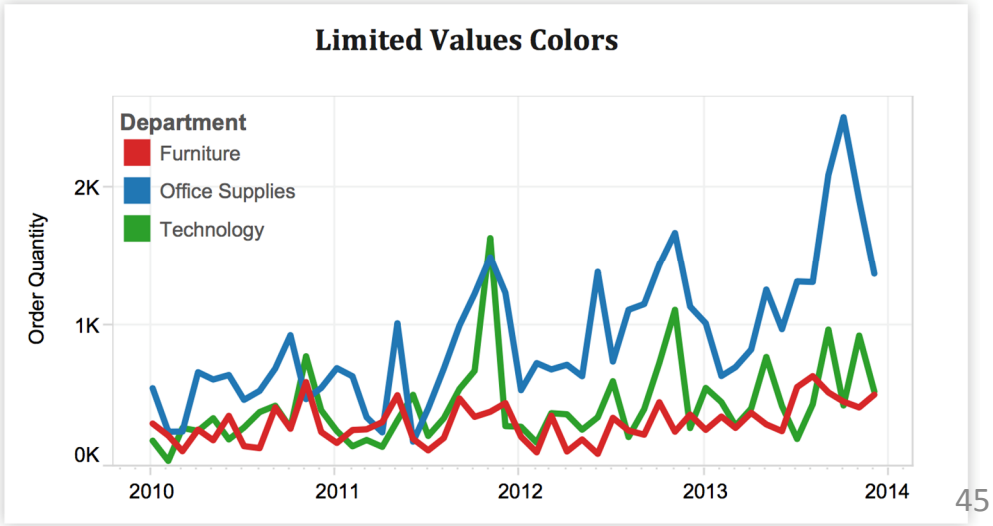
Good



Great







3. Great visualization checklist

- **What questions are you trying to answer?**
  - Does this visualization answer all of your questions?
  - Is the purpose of the visualization clearly explained in its title or surrounding text?
  - Can you understand the visualization in 30 seconds or less, without additional information?
  - Does your visualization include a title? Is that title simple, informative, and eye-catching?
  - Does your visualization include subtitles to guide your viewers?

3. Great visualization checklist (cont'd)

- **Do you have the right chart type for your analysis?**
  - What types of analyses are you performing?
  - Have you selected the most suitable chart type(s) for your types of analyses?
  - Have you considered alternative chart types that could work better than the ones you have chosen?

3. Great visualization checklist (cont'd)

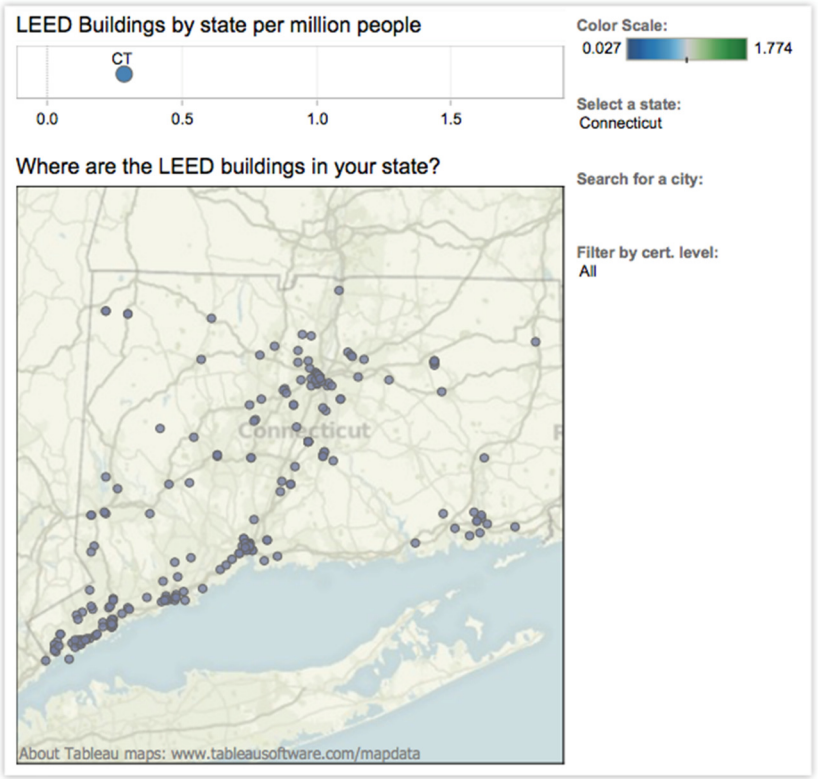
- **Are your views effective?**
  - Are your most important data shown on the X- and Y-axes and your less important data encoded in color or shape attributes?
  - Are your views oriented intuitively—do they cater to the way your viewers read and perceive data?
  - Have you limited the number of measures or dimensions in a single view so that your users can see your data?
  - Have you limited your usage of colors and shapes so that your users can distinguish them and see patterns?

Outline

I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
   - Spatial Data Visualization
   - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

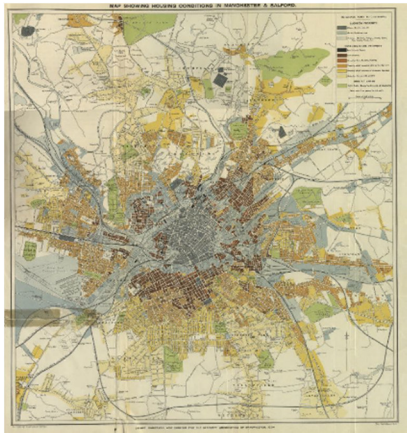
Spatial Data Visualization

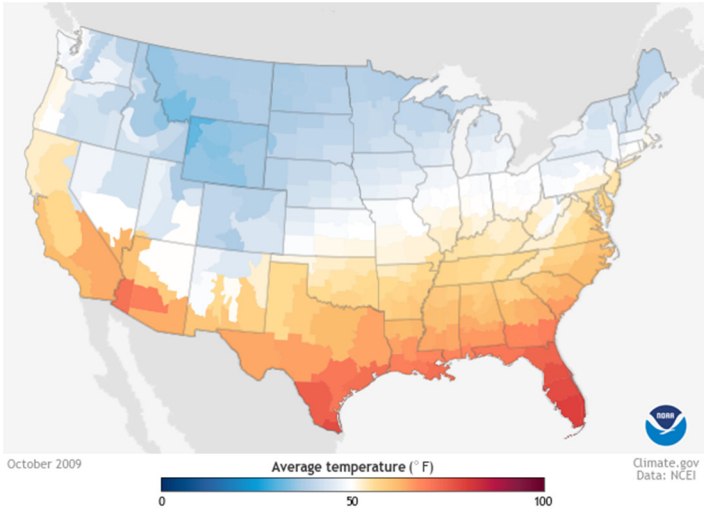
- Regional analysis: visualize data on **geographical maps** to answer location specific questions, or aid geographical exploration
  - Examples: Insurance claims by state, product export destinations by country, car accidents by zip code, custom sales territories.
- Tips:
  - Use maps as a filter for other types of charts, graphs, and tables.
  - Layer bubble charts on top of maps
  - Showing geocoded data: postal codes, country names, etc.



What is a map

- A map is a special visualization of spatial data that requires:
  - Geographic coordinates
  - Location reference system
  - Attributes linked to a particular location
- What tasks:
  - Find Location/Feature (country, city, street)
  - Give directions (find route)
  - Compare attributes associated to locations (temperature, traffic, polls, vote, etc.)





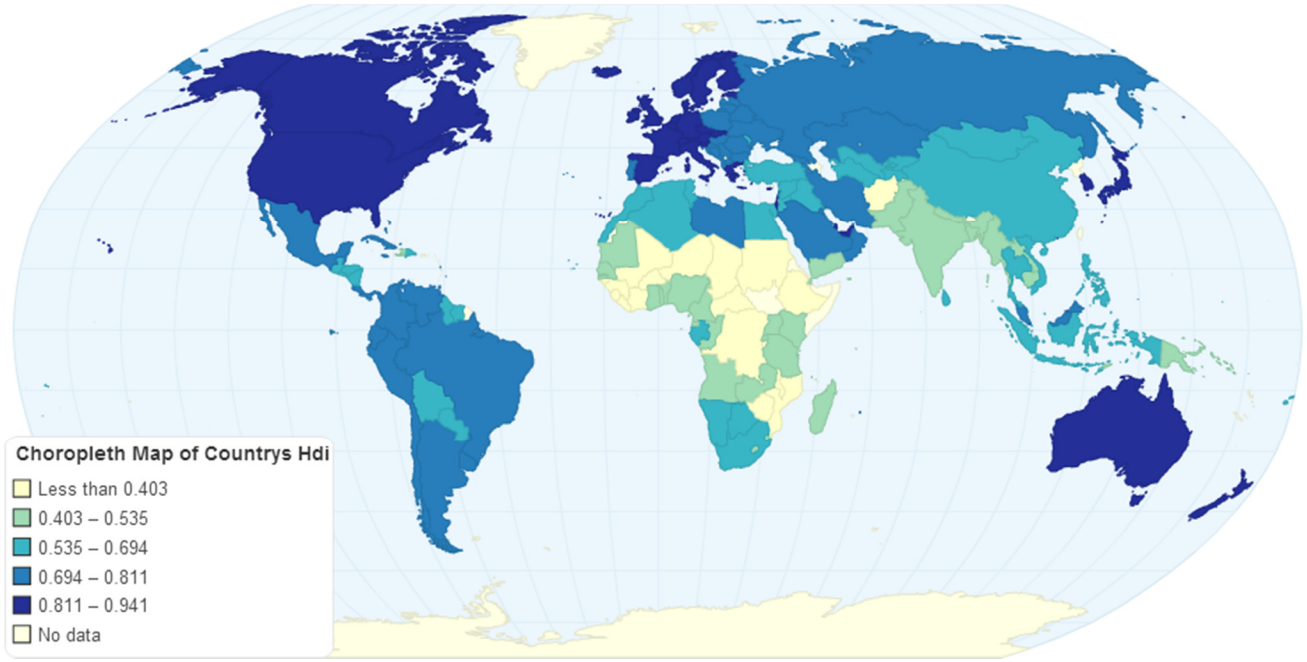
https://www.climate.gov/maps-data/data-snapshots/averagetemp-monthly-cmb-2009-10-00?theme=Temperature

Map Visualization Techniques

1. Choropleth maps
2. Heat maps
3. Contour maps
4. Cartogram
5. Dot distribution maps
6. Flow maps

1. Choropleth maps

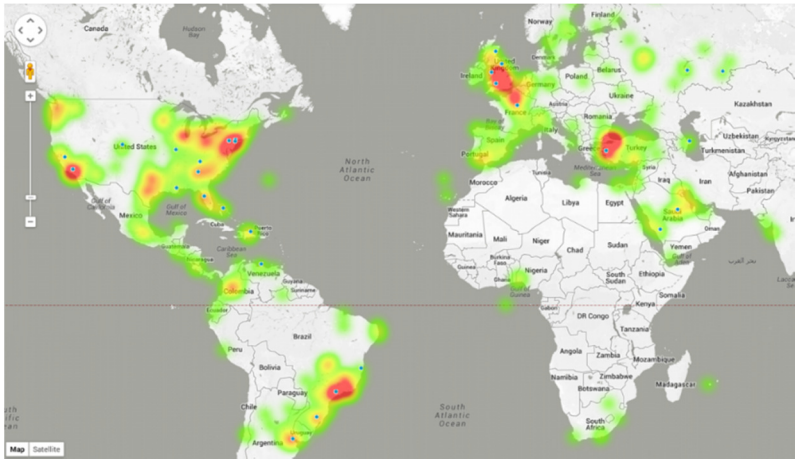
- Areas are shaded or patterned in proportion to the measurement of attributes
- Each spatial region is filled with an uniform color (or pattern)
- Best used when you have defined areas
- Best used when you can group the data into 3-6 categories.



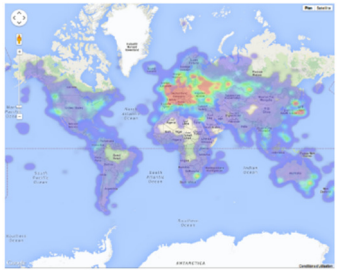
http://chartsbin.com/view/36449

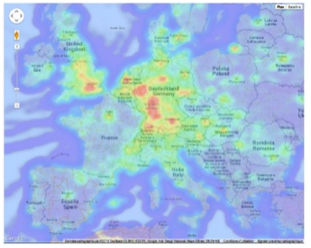
2. Heat maps (aka density visualization)

- Represent continuous data variable values by colors
- The coloration a heat map does not correspond to geographic boundaries
- Typical types:
  - Zoom-dependent: fixed radius, color mixed into a big circle when zooming-out (e.g. heatmap layer on Google Maps)
  - Zoom-independent: maintain the density shape when zooming-out



Geotagged tweets on a Google Map
https://cdn.blog.safe.com/wp-content/uploads/2014/04/2014-04-16_09-58-28-1024x590.png

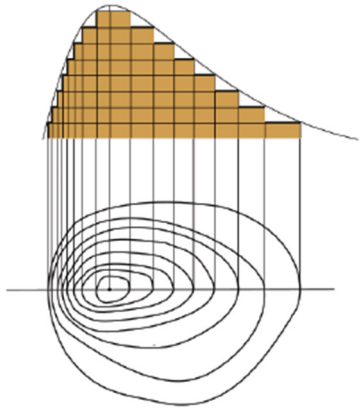




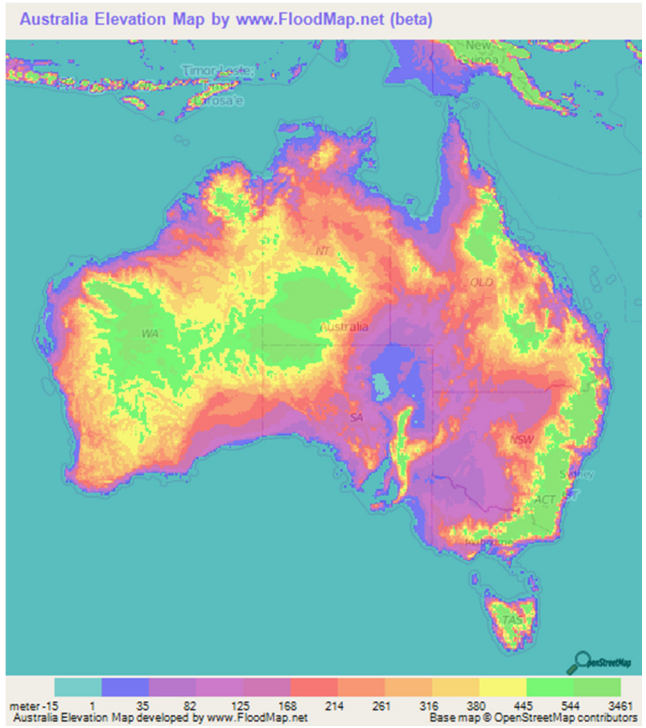
Perrot, Alexandre, et al. "Large interactive visualization of density functions on big data infrastructure." *Large Data Analysis and Visualization (LDAV), 2015 IEEE 5th Symposium on*. IEEE, 2015.

3. Contour maps

- Also known as Isarithmic or Isopleth
- Depict smooth continuous phenomena
- **Contour line:** curve connecting points where the function has the same particular value.
- Needs a lot of points to be precise
- Can be combined with heatmap

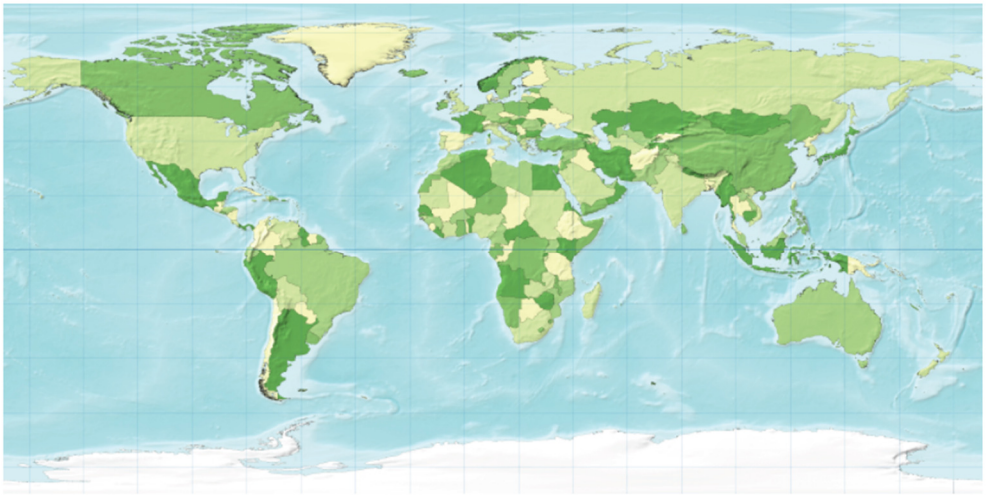


LINES OF MAGNETIC DECLINATION
Edmund Halley, 1701



4. Cartogram

- Distort the shape of geographic regions to encode another variable in the spatial area



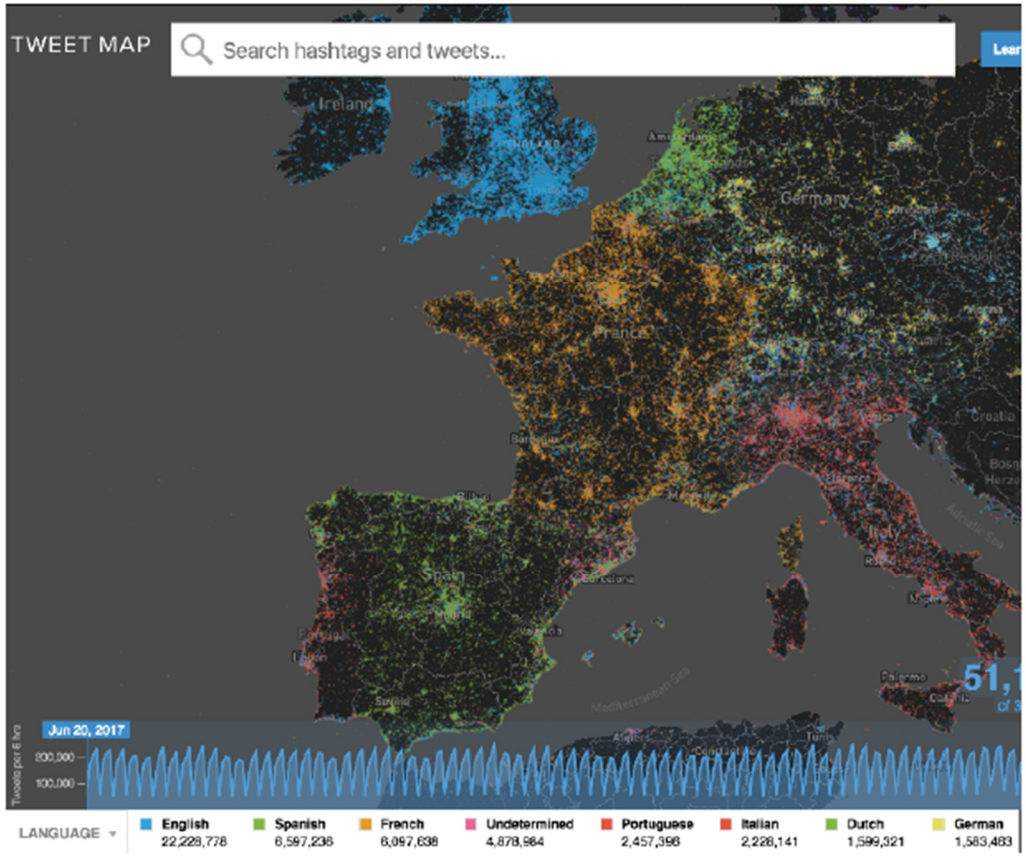
Population on normal world map

Population cartogram

http://www-personal.umich.edu/~mejn/cartograms/

5. Dot distribution maps

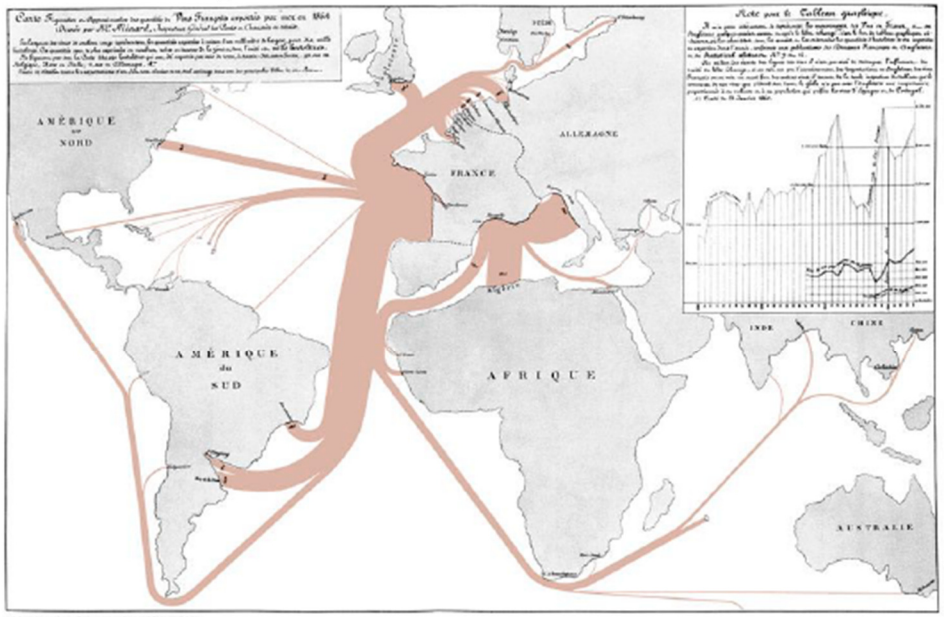
- Reveal spatial distribution using the point mark
- Size is important
  - Too small: hardly perceived
  - Too big: wrong impression of densities



https://www.mapd.com/demos/tweetmap/

6. Flow maps

- Multivariate representation of movement in space
- Properties:
  - Spatial position
  - Line width
  - Color



Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard*, 1845-1869, a portfolio of his work held by the Bibliothèque de l'Ecole Nationale des Ponts et Chaussées, Paris.

Minard's map of French wine exports for 1864

World heroin movements (Source: CIA)

Outline

I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
   - Spatial Data Visualization
   - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

Textual Data Visualization

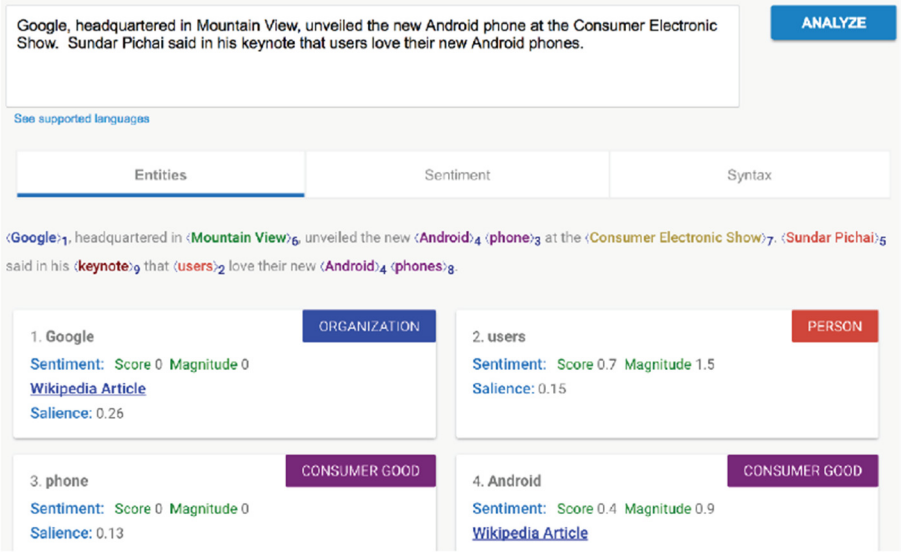
- Why visualize text?
  - **Faster understanding:** get quick insights on what I am reading
  - **Comparison:** compare document collections, or inspect evolution of collection over time
  - **Clustering (grouping):** classification or topic extractions
  - **Correlation:** compare patterns in the current text to other datasets

Text datasets?

- Forum articles
- Social comments (Facebook, Twitter, Reddit)
- Chats
- Social profiles

Text pre-processing

1. **Tokenization**: decompose text into meaningful units
   - Segment text into terms: “a sentence” → [“a”, “sentence”]
   - Remove **stop words**: a, after, before, such, then, the, etc.
   - Remove numbers and symbols: #love, @potus, yo!!!!?
   - **Named Entity Recognition**: Google, U.S.A
     - Label named entities in a text
     - Further analysis: check co-occurrences in short term window (5 words)



https://cloud.google.com/natural-language/

Text pre-processing

1. Tokenization: decompose text into meaningful units
2. **Token normalization**
   - Stemming: find the root word (e.g. use is the root word of usage, user, etc.)
   - Lemmatization: transform a word to simple form (e.g. use is the simple form of used, uses, using, etc.)
   - Unordered word model: bag of words
     - N-gram (e.g. capture New South Wales with N=3)

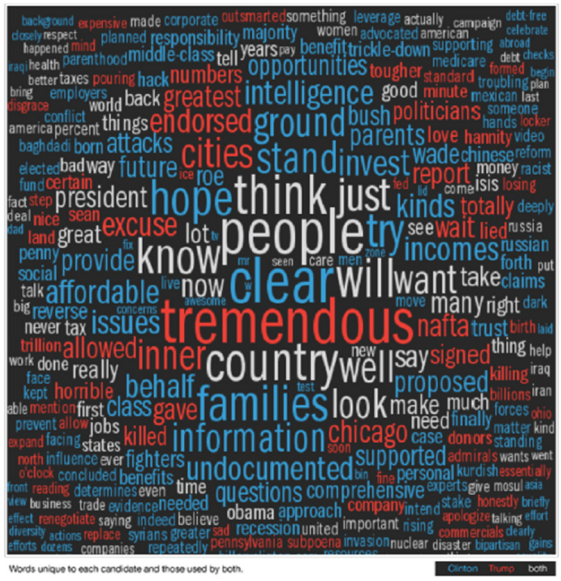
Common text visualizations

- Word cloud
- Term frequency
- Advanced viz

Word cloud

- **Pros:**
  - Can help with an initial query
    - e.g. identifying the main words
- **Cons:**
  - Bad visual encoding (size vs. position)!
  - No structure
  - Inaccurate size encoding (long words look bigger)
  - Layout is unstable, difficult to compare two tag clouds

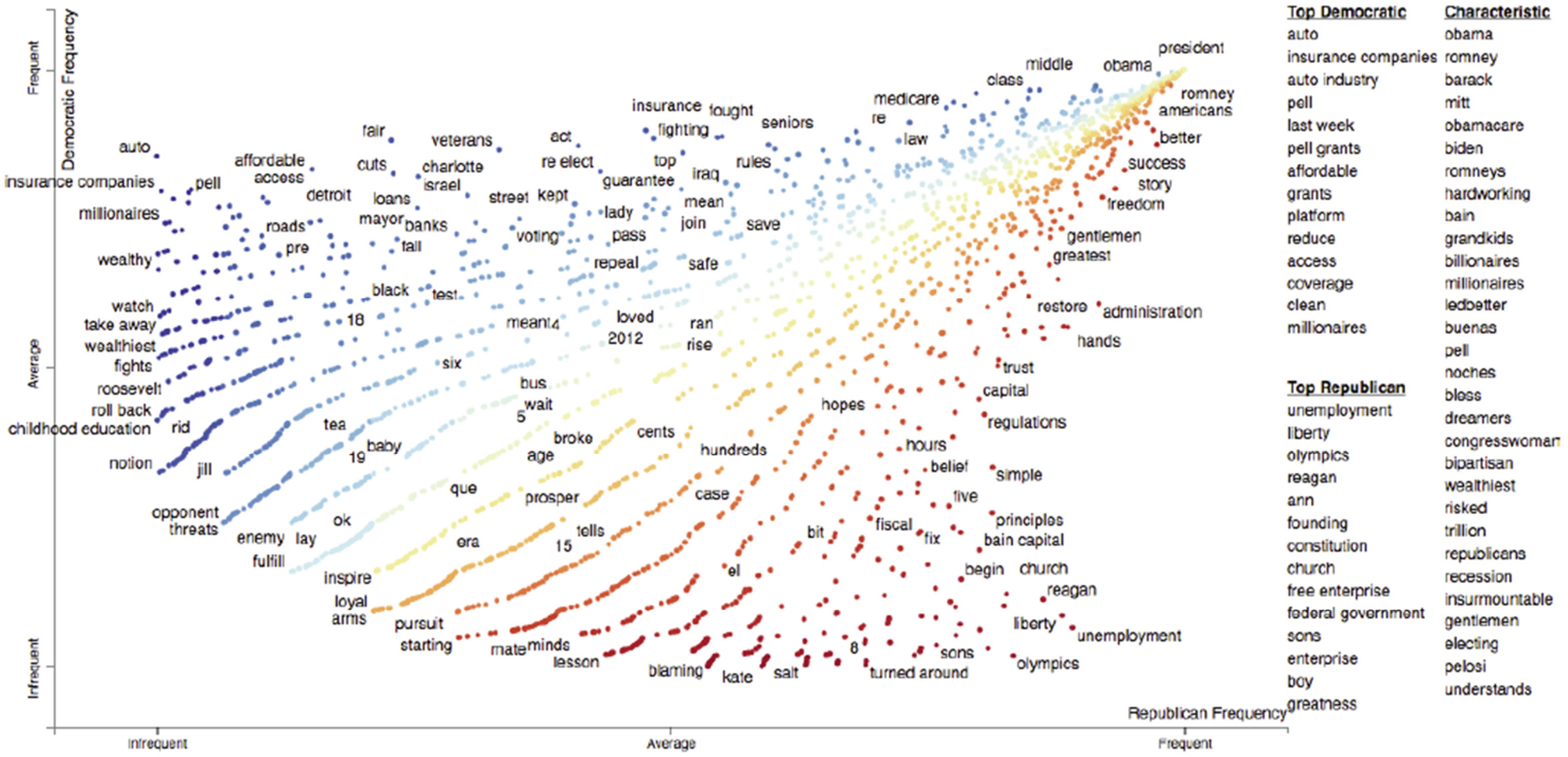
http://blog.jazzfactory.in/2010/11/create-free-tag-clouds-with-wordle.html



http://mkweb.bcgsc.ca/debates2016/?debate=clinton-trump-04

Term frequency

- Count the average frequency of each word across documents



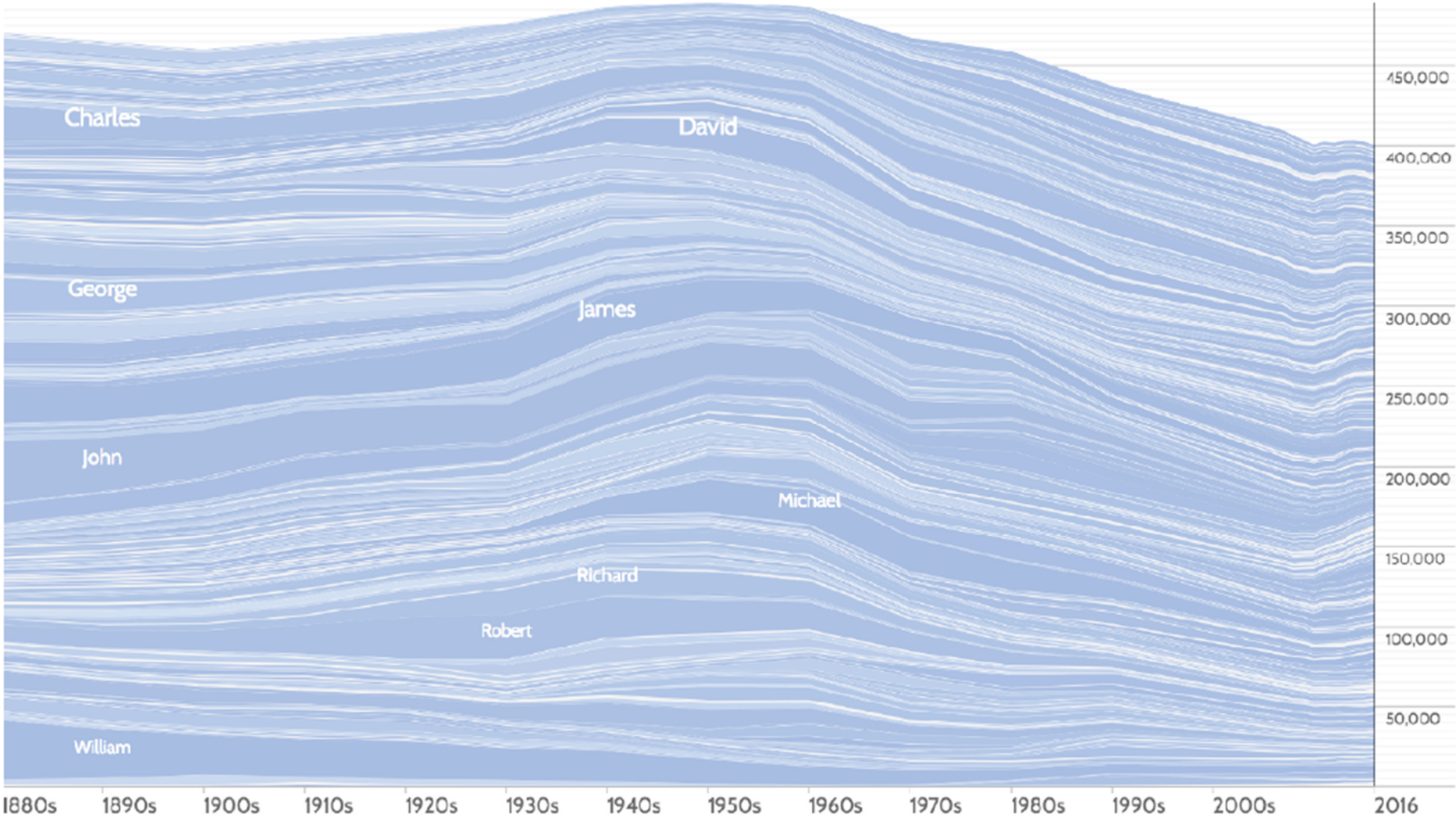
Top Democratic
auto
insurance companies
auto industry
pell
last week
pell grants
affordable
grants
platform
reduce
access
coverage
clean
millionaires

Top Republican
unemployment
liberty
olympics
reagan
ann
founding
constitution
church
free enterprise
federal government
sons
enterprise
boy
greatness

Characteristic
obama
romney
barack
mitt
obamacare
biden
romneys
hardworking
bain
grandkids
billionaires
millionaires
ledbetter
buenas
pell
noches
bless
dreamers
congresswoman
bipartisan
wealthiest
risked
trillion
republicans
recession
insurmountable
gentlemen
electing
pelosi
understands

https://jasonkessler.github.io/Conventions-Visualization.html

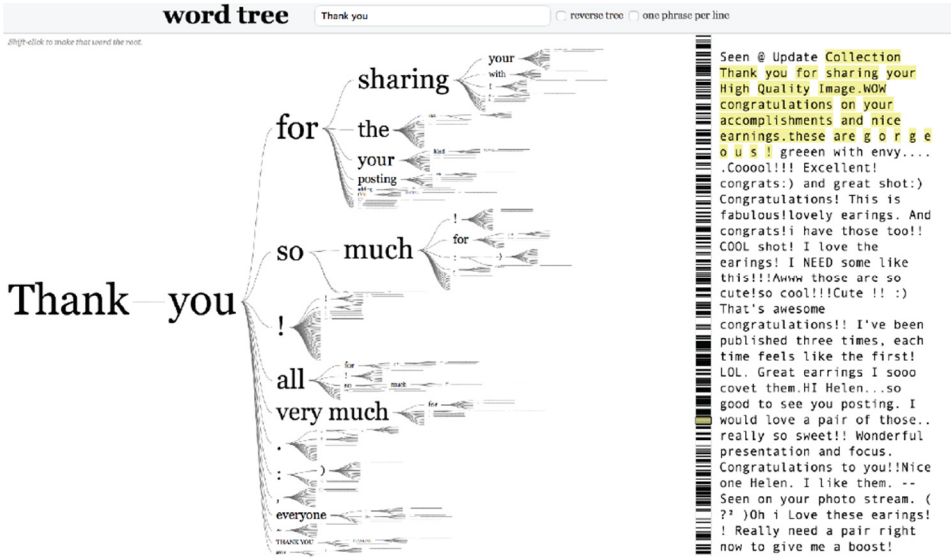
Advanced text visual analytics: Combined with other charts



Baby Name Wizard

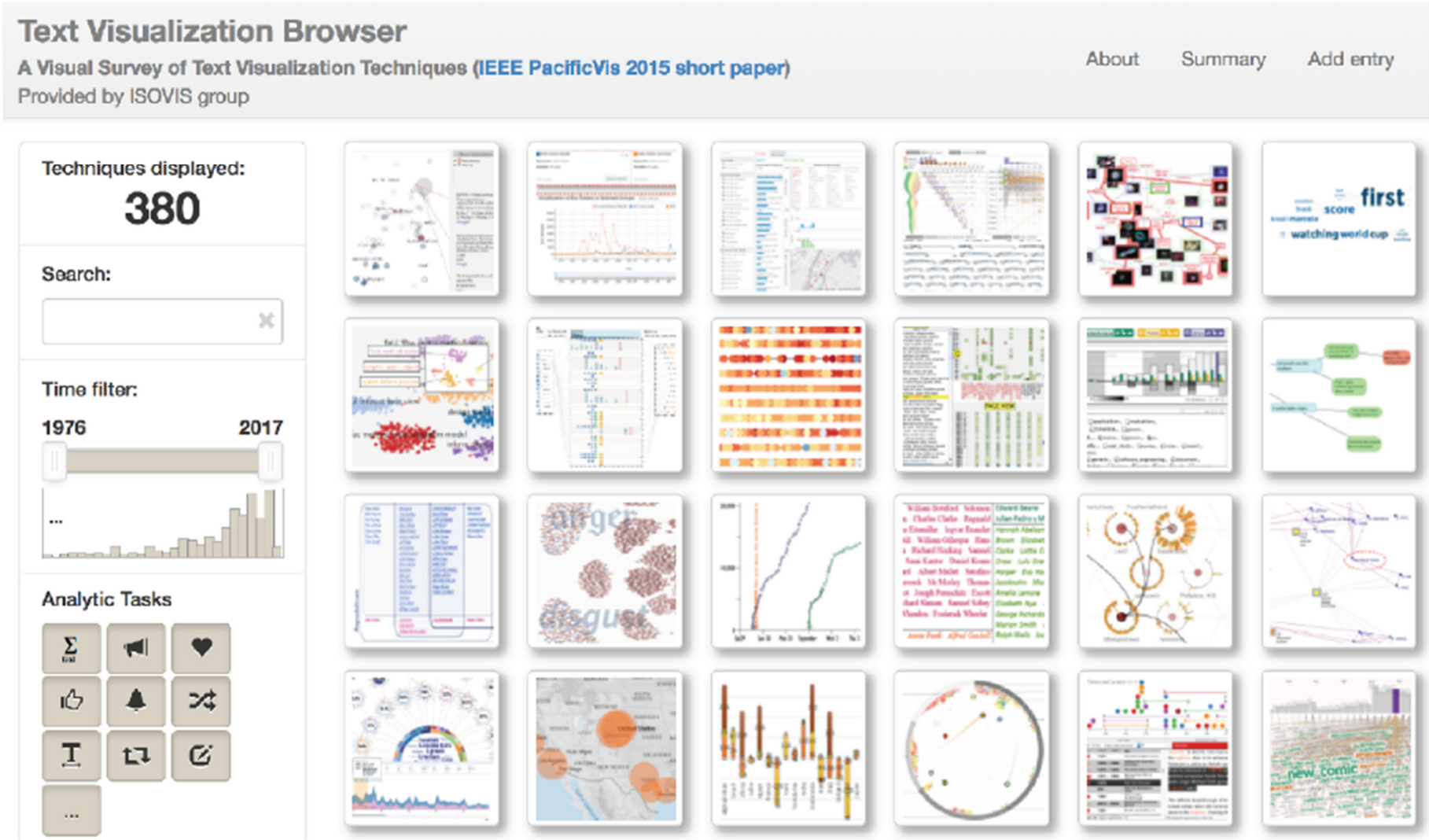
Advanced text visual analytics: WordTree

- **Pros:**
  - Useful for n-gram visualisation n > 1
  - Show recurrent concordance (common local context) of repeated structures
- **Cons:**
  - Need to choose a starting word
  - Ignore other type of structures (syntactic, lexical)



https://www.jasondavies.com/wordtree/?source=flickr-comments.txt&prefix=Thank%20you%20so%20much

More examples



http://textvis.lnu.se/

Outline

I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
  - Spatial Data Visualization
  - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

IV. From Data Visualization to Visual Analytics

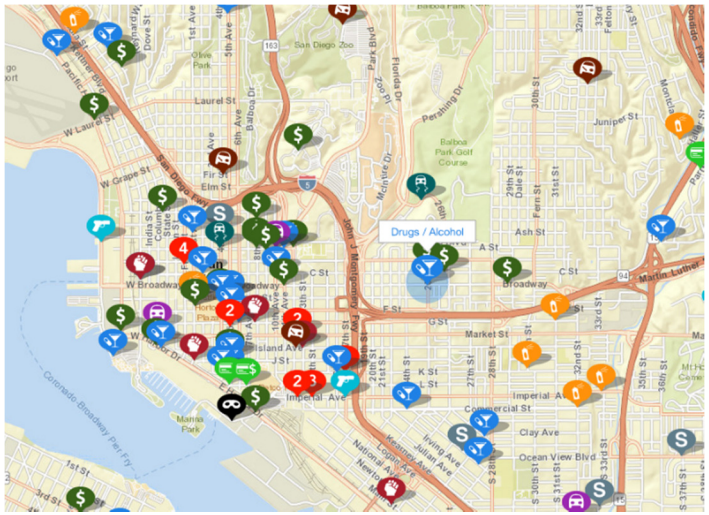
- Interactive viz = Old-fashioned viz + Interaction Scheme
  - Enable **visual analytics** via interactive and reproducible results
  - Easy and fast to develop and customize

- Old-fashioned viz
  - Great for data exploration, developed throughout the last few centuries
  - Rapid data exploration
  - Focus on most important details

- Interactive viz
  - More and more common nowadays. New frameworks are the key enabler.
  - Support multiple analyses
  - Focus on more dimensions

Visual Analytics: Applications

- Map-based analytics, such as CrimeMapping



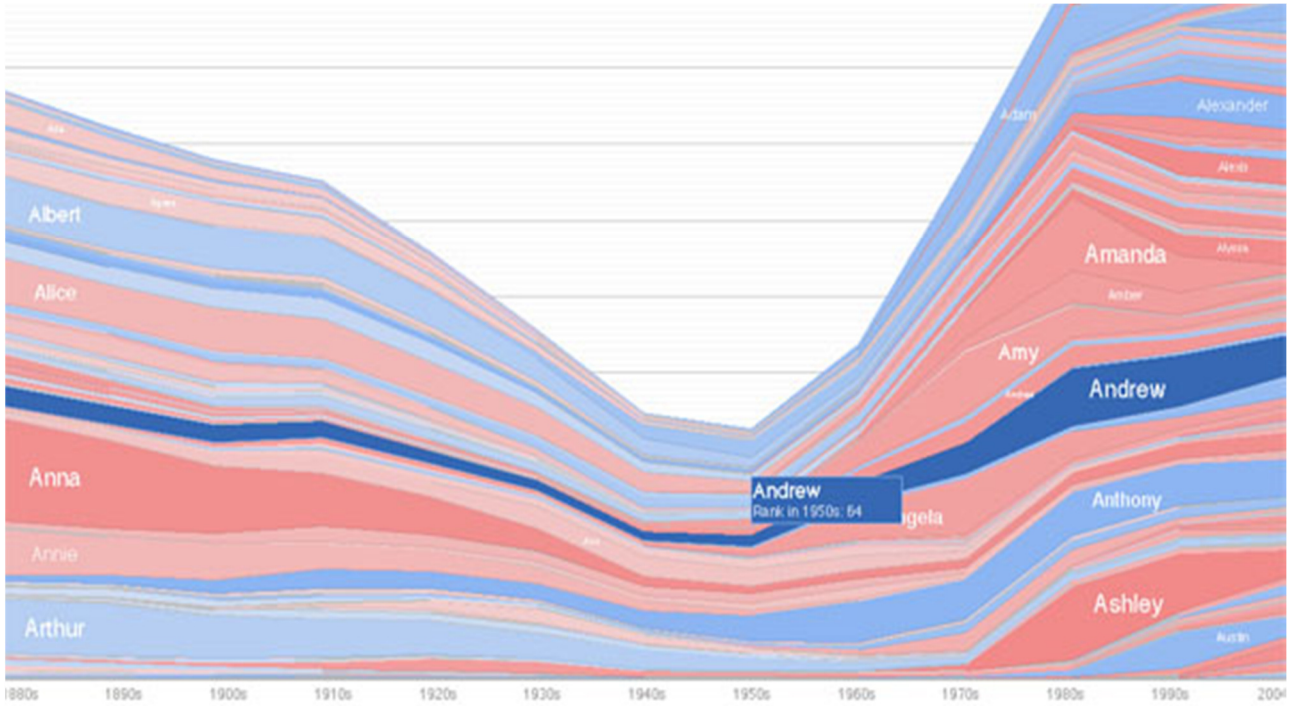
https://www.crimemapping.com/map

- Interactive Education
  - The famous Gapminder Video, Hans Rosling: 200 Countries, 200 Years, 4 Minutes. https://www.youtube.com/watch?feature=player_embedded&v=jbkSRLYSojo

- Future of Journalism: e.g. NY Times
  - NY Times Interactive Visualizations (recession/recovery 2014). http://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html
  - And 2014 "the year in interactive storytelling". http://www.nytimes.com/interactive/2014/12/29/us/year-in-interactive-storytelling.html?_r=0

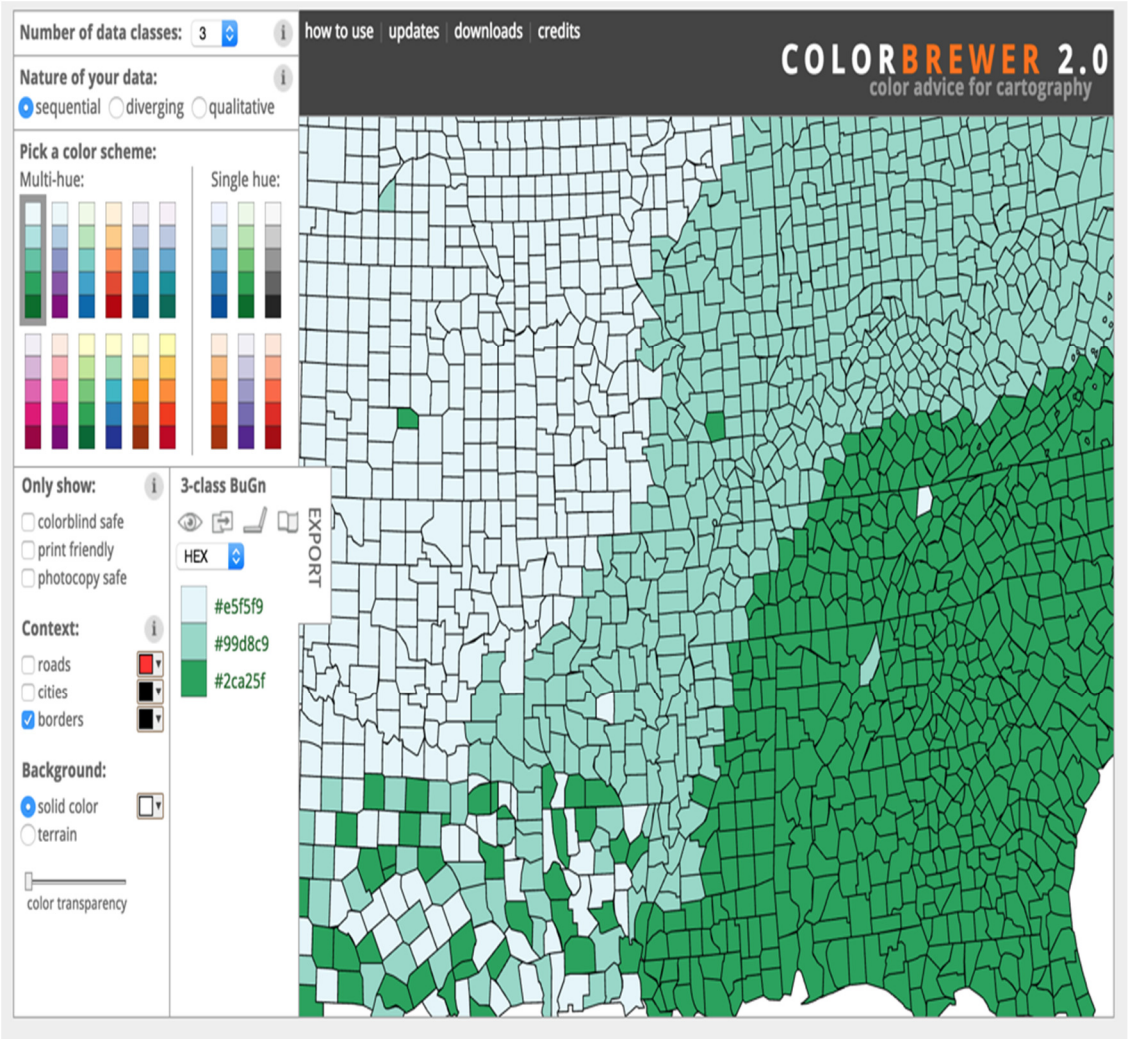
Why Interactive?

- With interactive charts you can keep things very simple by **hiding** and **dynamically revealing** important structure.
- On an interactive chart, you reveal the information most useful for **navigating** the chart.

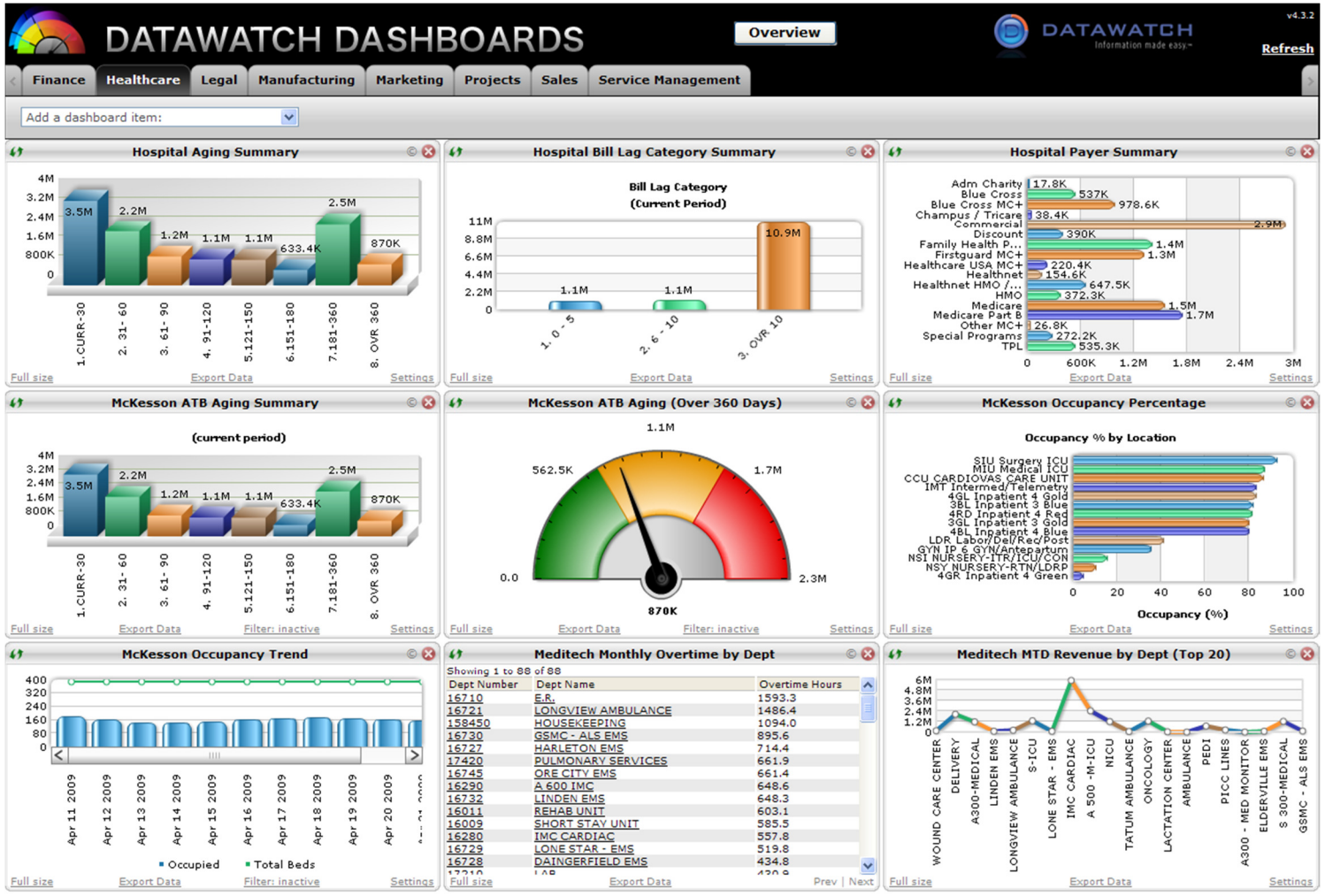


Why interactive?

- A translation between visual information and raw data



Visual Analytics: Dashboards



Source https://www.vocalabs.com/blog/my-dashboard-pet-peeve

Dashboard

- Dashboards are an effective tool for distilling data into actionable insights
- Benefits:
  - Align your organization's efforts, speed up decision-making
  - Track performance outcomes
- Best practices for creating successful dashboards:
  1. Connect to all of your data
  2. Choose metrics that matter
  3. Use better visualizations
  4. Share for collaboration

Practice 1: Connect to all of your data

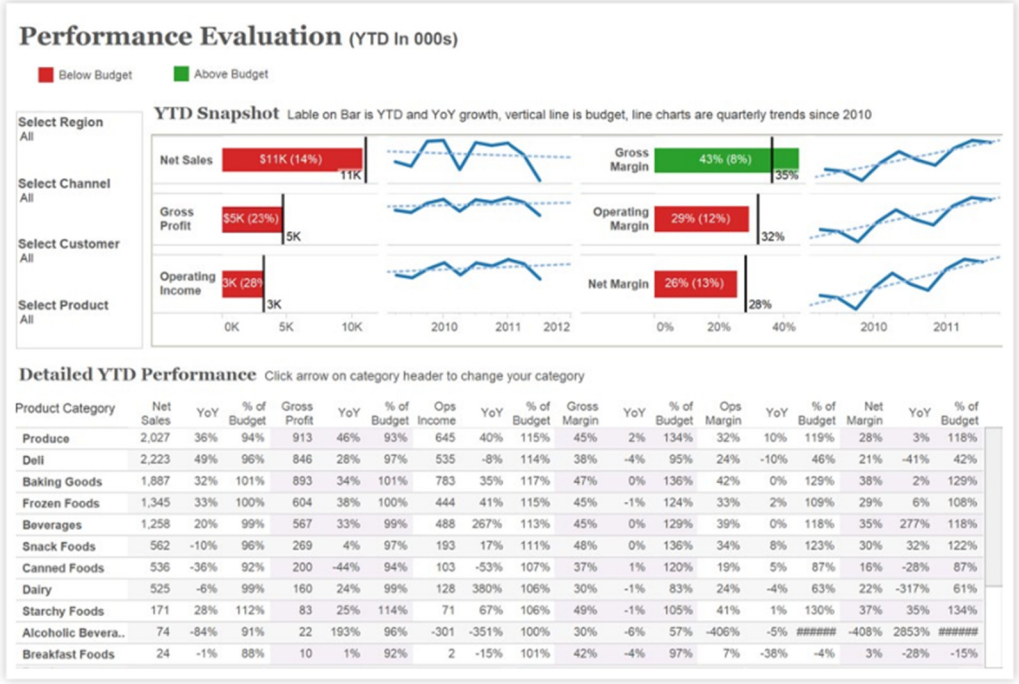
- Requirements:
  - Data is distributed across different places, in or out of an organization
- Solutions:
  - Connect your dashboard to all of your relevant data
  - Publish all dashboards to the cloud
  - Provide secure, convenient access for all employees across different locations and different devices.

Practice 2: Choose metrics that matter

- Requirements:
  - Metrics must be relevant, but should be highly selective
  - Allow users to discover not only what but also why
- Solutions: verify with the following questions
  - How does each metric contribute to your objectives?
  - Can you design a meaningful metric measuring those contributions?
  - Is this metric truly necessary to contribute to the objectives?
  - Can you build a systematic and ongoing means of measurement?
  - Consider incorporating third-party market-share metrics.
  - The litmus test: every metric must connect to your objectives

Practice 2: Choose metrics that matter

- Case study: sales against budget over two years
  - Details metrics match up directly with the company objectives:
    - Answer the growth rates
    - Identify outliers or abnormal trends
    - Check whether performance goals have been met



Performance Evaluation (YTD In 000s)

Below Budget Above Budget

Select Region: All
Select Channel: All
Select Customer: All
Select Product: All

YTD Snapshot Lable on Bar is YTD and YoY growth, vertical line is budget, line charts are quarterly trends since 2010

Net Sales $11K (14%) 11K
Gross Profit $5K (23%) 5K
Operating Income 3K (28% 3K
Gross Margin 43% (8%) 35%
Operating Margin 29% (12%) 32%
Net Margin 26% (13%) 28%

0K 5K 10K 2010 2011 2012
0% 20% 40% 2010 2011

Detailed YTD Performance Click arrow on category header to change your category

Product Category	Net Sales	YoY	% of Budget	Gross Profit	YoY	% of Budget	Ops Income	YoY	% of Budget	Gross Margin	YoY	% of Budget	Ops Margin	YoY	% of Budget	Net Margin	YoY	% of Budget
Produce	2,027	36%	94%	913	46%	93%	645	40%	115%	45%	2%	134%	32%	10%	119%	28%	3%	118%
Deli	2,223	49%	96%	846	28%	97%	535	-8%	114%	38%	-4%	95%	24%	-10%	46%	21%	-41%	42%
Baking Goods	1,887	32%	101%	893	34%	101%	783	35%	117%	47%	0%	136%	42%	0%	129%	38%	2%	129%
Frozen Foods	1,345	33%	100%	604	38%	100%	444	41%	115%	45%	-1%	124%	33%	2%	109%	29%	6%	108%
Beverages	1,258	20%	99%	567	33%	99%	488	267%	113%	45%	0%	129%	39%	0%	118%	35%	277%	118%
Snack Foods	562	-10%	96%	269	4%	97%	193	17%	111%	48%	0%	136%	34%	8%	123%	30%	32%	122%
Canned Foods	536	-36%	92%	200	-44%	94%	103	-53%	107%	37%	1%	120%	19%	5%	87%	16%	-28%	87%
Dairy	525	-6%	99%	160	24%	99%	128	380%	106%	30%	-1%	83%	24%	-4%	63%	22%	-317%	61%
Starchy Foods	171	28%	112%	83	25%	114%	71	67%	106%	49%	-1%	105%	41%	1%	130%	37%	35%	134%
Alcoholic Bevera..	74	-84%	91%	22	193%	96%	-301	-351%	100%	30%	-6%	57%	-406%	-5%	######	-408%	2853%	######
Breakfast Foods	24	-1%	88%	10	1%	92%	2	-15%	101%	42%	-4%	97%	7%	-38%	-4%	3%	-28%	-15%

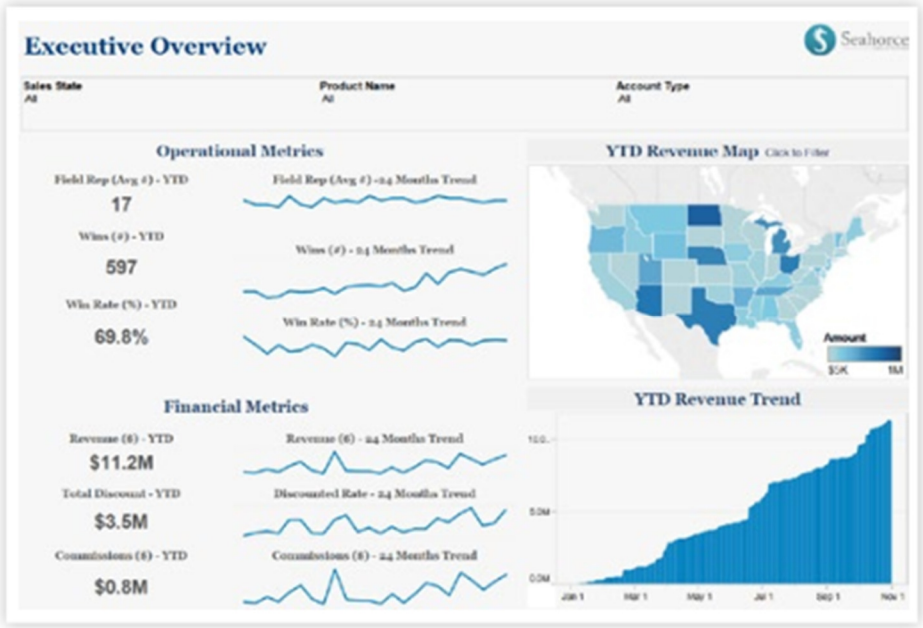
Practice 3: Utilize Better Visualizations

- Requirements:
  - A picture is worth a thousand words
  - Choose the right chart/graph: basic is good, but advanced may be better
- Solutions: fulfill the core selling points
  - Variety:
    - Not everything has to be a pie chart or a bar chart
    - Utilize multiple types of visualizations, colors, and filters
  - Current & Interactive:
    - Real-time data
    - Interactive: filtering views, adjust parameters, drilling down
  - Temporal trends:
    - Forward Looking: support forecasting and planning
    - Support different windows of historical data

Practice 3: Utilize Better Visualizations

- Case study: dashboard for sales executives
  - Scan-friendly overview of key executive metrics: revenue, profit, sales discount, commissions
  - Revenue map + trend chart

  → Monitor individual performance, regional performance, and upcoming opportunities

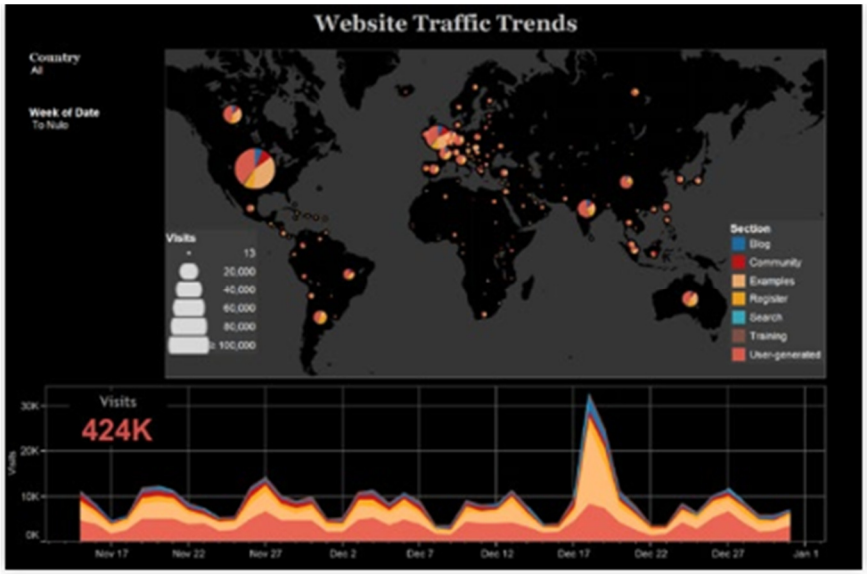


Practice 4: Share for Collaboration

- Requirements:
  - Sharing information is what dashboards are all about
    - Get everyone looking at the same page
- Solutions:
  - Simple browser-based distribution
    - Click and go
    - Pull recent data
  - Embed dashboards in existing tools
  - Test your distribution plan and get feedback

Practice 4: Share for Collaboration

- Case Study: Website Traffic
  - Connect directly with Google Analytics
  - Easily follow a link, click and interact to see trends for a website



Metrics of Success

- Turn insights into action and inspire true innovation
  - Efficiency:
    - Visualize for faster understanding
    - Allow sharing and mobile options for better collaboration
  - Effectiveness:
    - Multiple data sources, multiple metrics, multiple patterns
    - Support both summary views and in-depth views
  - Business readers and knowledge workers ask and answer questions in real-time
    - What number can you hit?
    - What deals can you close?
    - How can you improve your supply chain?
    - How many lives can be saved?
    - What changes can you make this very moment for a better outcome?

Learning Objectives

At the end of this lecture, you should be able to:

- Identify **building blocks** and **principles** of data visualization
- Devise techniques in
  - **spatial data** analytics
  - **text data** analytics
- Apply the principles of **dashboard** design for visual analytics
- Handling Big Data

Outline

I. Building blocks of data visualization (OPTIONAL)
II. Data visualization principles (OPTIONAL)
III. Data Visualization for Special Types of Data
   - Spatial Data Visualization
   - Textual Data Visualization
IV. From Data Visualization to Visual Analytics
V. Big Data Visualization

4-V of Big Data Visualization

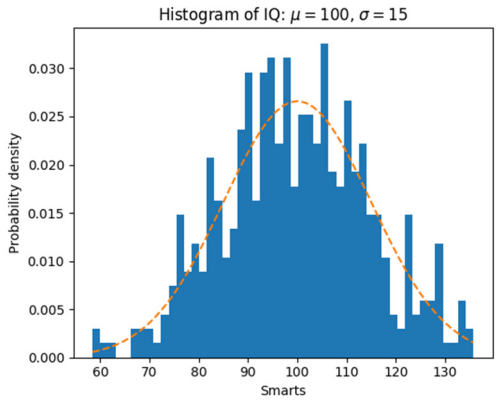
- **Volume**: sampling for data points
- **Variety**: multiple sources, diversity, aggregation
- **Velocity**: sampling for data stream
- **Veracity**: out-of-scope. E.g. credibility analysis techniques.

Big Data Visualization: Volume

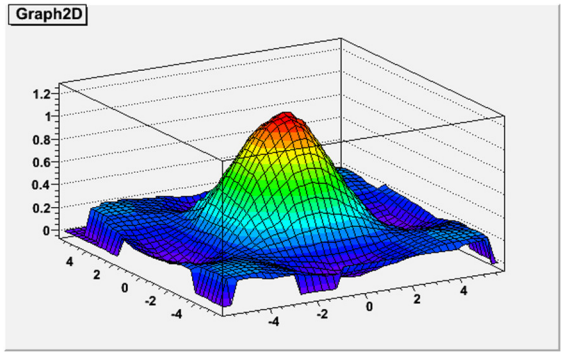
1. **Aggregation**: one visual point represents multiple data points
2. **Sampling**: show only some of the dataset

1. Aggregation

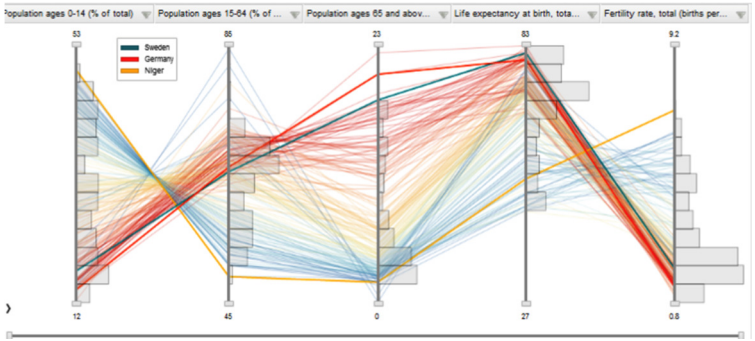
- Bar Chart → Histogram



- Points on a Map, Scatterplot, Heatmap → 2D Histogram



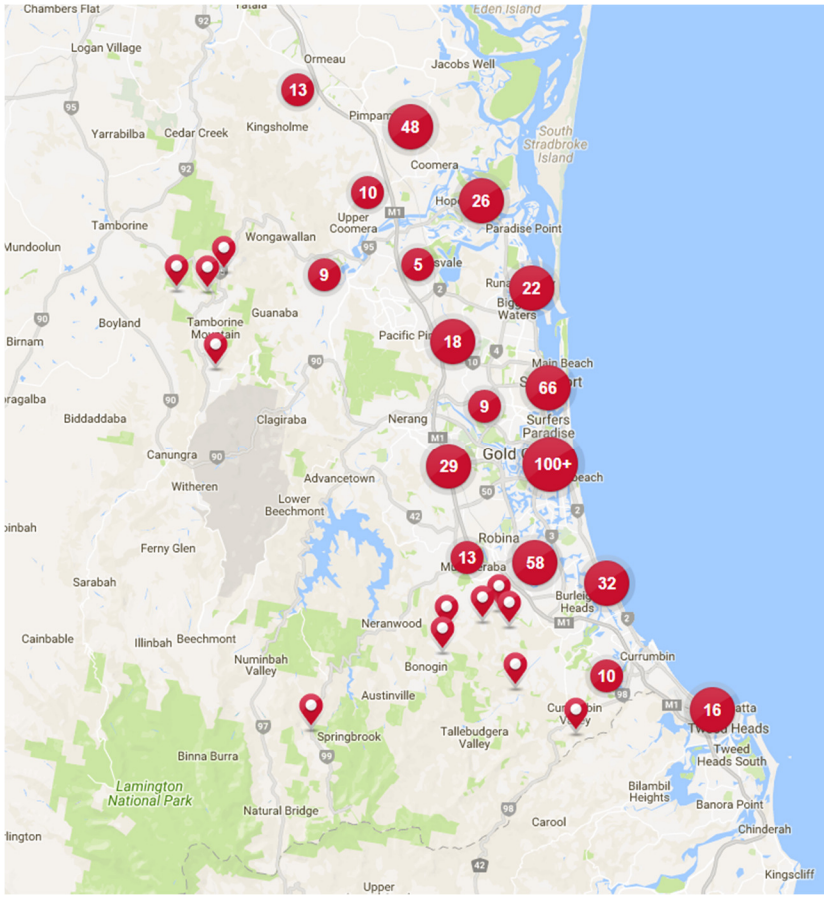
- Parallel Coordinates → Area Para Coordinates



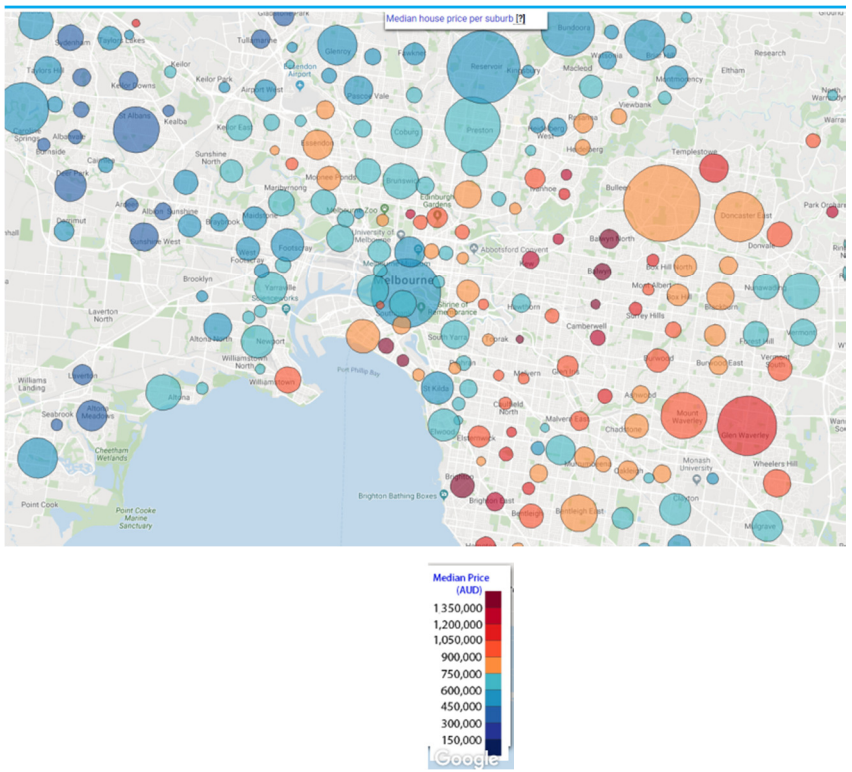
http://mitweb.itn.liu.se/GAV/dashboard/#story=data/Ageing%20Population%20in%20the%20World.xml&layout=(radarchart,pcp)

Aggregation

- Aggregation on Map (use size of the circles)



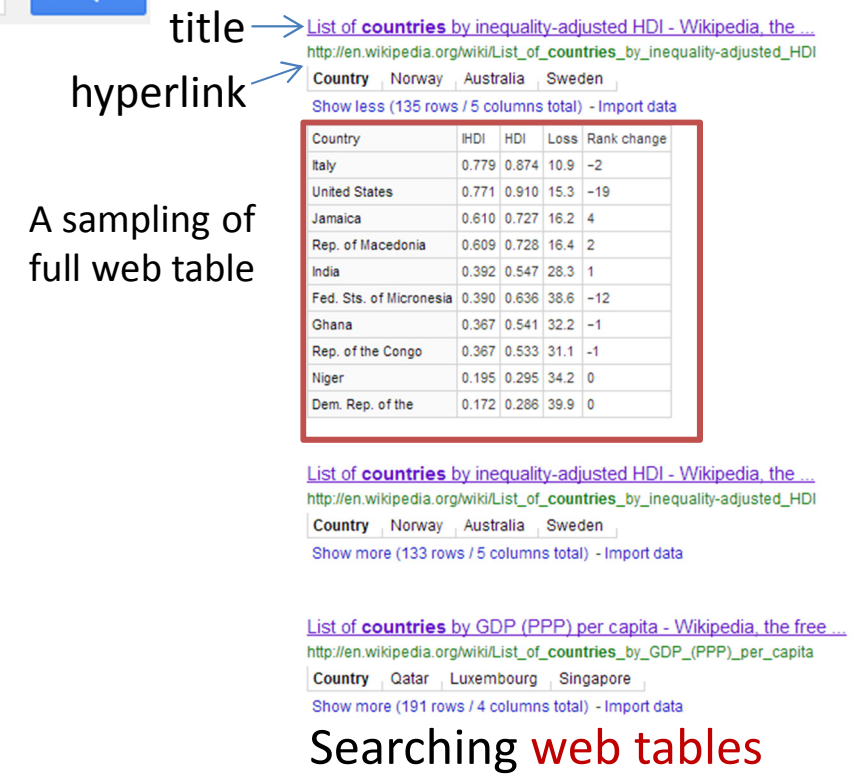
realestate.com.au



2. Sampling

- Show only a subset of data
- Sampling requirements:
  - Small enough for user cognitive load
  - Reflect the properties of original data
- **Application**: Web Table searching

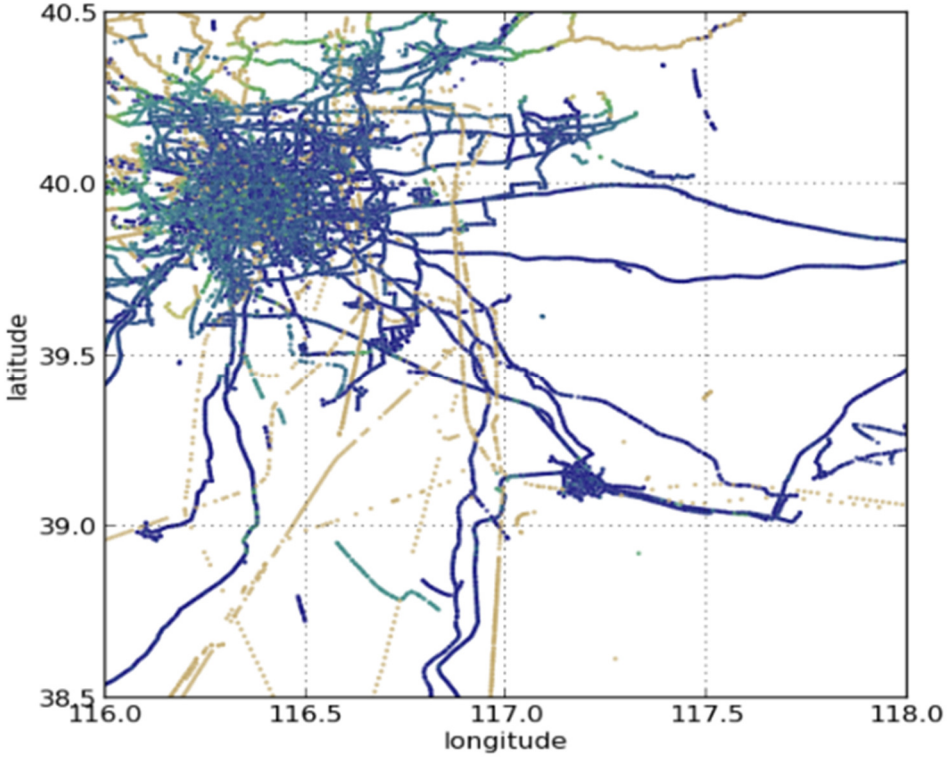
Searching web pages



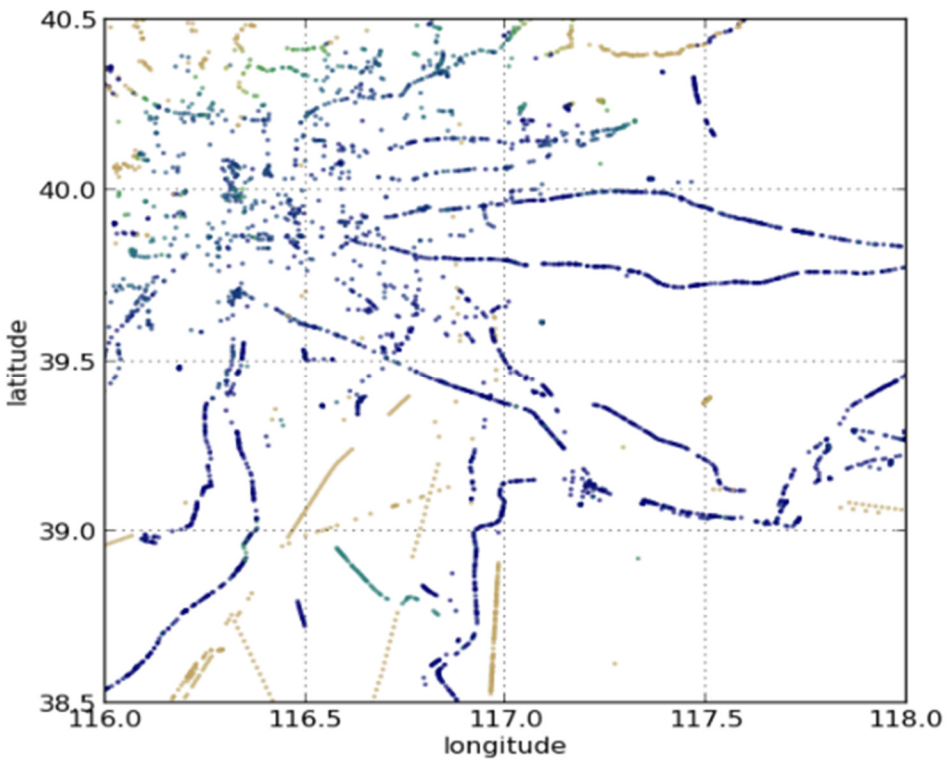
Country	IHDI	HDI	Loss	Rank change
Italy	0.779	0.874	10.9	−2
United States	0.771	0.910	15.3	−19
Jamaica	0.610	0.727	16.2	4
Rep. of Macedonia	0.609	0.728	16.4	2
India	0.392	0.547	28.3	1
Fed. Sts. of Micronesia	0.390	0.636	38.6	−12
Ghana	0.367	0.541	32.2	−1
Rep. of the Congo	0.367	0.533	31.1	-1
Niger	0.195	0.295	34.2	0
Dem. Rep. of the	0.172	0.286	39.9	0

Searching web tables

Sampling: another application



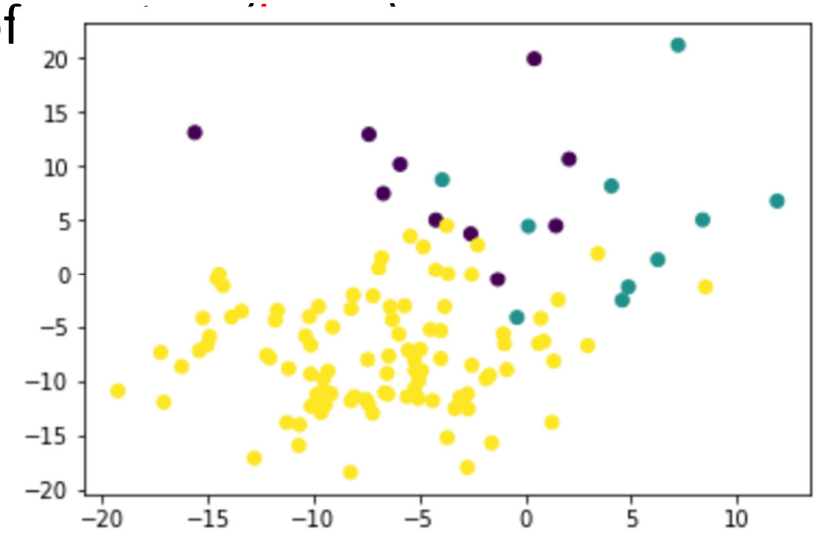
original data



a good sample

Sampling Techniques

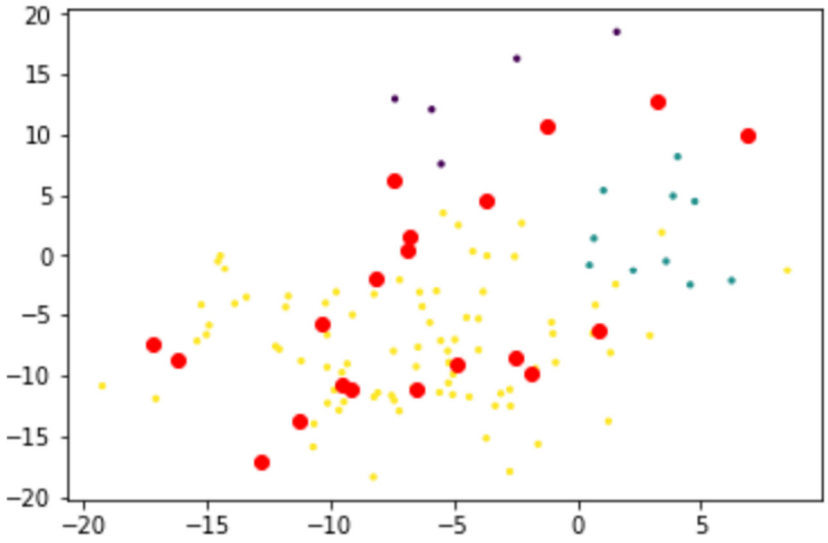
- **Pre-processing**: transform data items to d-dimensional points
  - Define features or reuse attributes
  - Consider each feature/attribute a one dimension
  - Each item is represented by a vector of its feature/attribute values
- **Goal of sampling:**
  - Select k out of



Example: a labeled dataset from scikit-learn

Simple Random Sampling (SRS)

- The probability of selecting every data item is uniform
- Algorithm?:
  - **Input**: a set of original data points D, |D|=n
  - **Output**: a set of samples $S \subseteq D$, with $|S| = k$

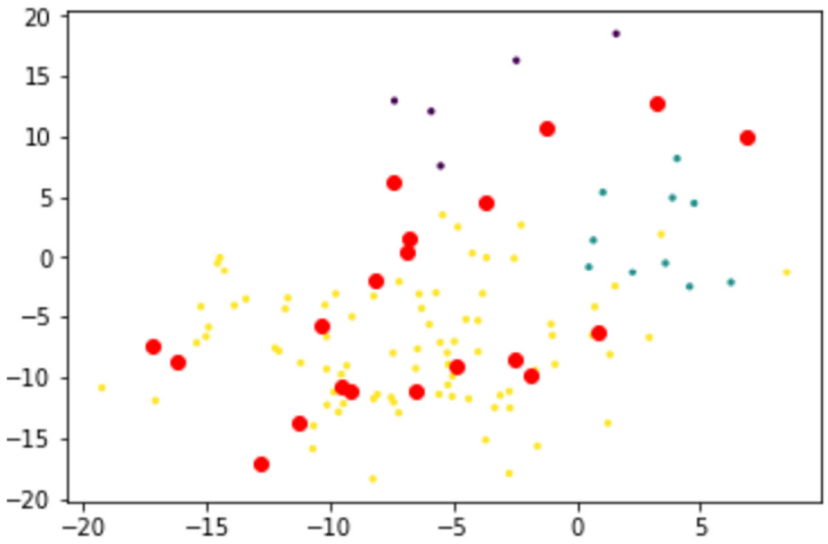


Simple Random Sampling (SRS)

- The probability of selecting every data item is uniform
- An algorithm for generate SRS:
  - **Input**: a set of original data points D, |D|=n
  - **Output**: a set of samples $S \subseteq D$, with $|S| = k$
  1. Generate a random permutation
     - for i from 0 to n−2 do
       - j ← random integer such that i ≤ j < n
       - exchange D[i] and D[j]
  2. Return the first k-elements: $S = D[0:k]$
- **A Python implementation**: S = numpy.random.permutation(D)[:k]

Simple Random Sampling (SRS)

- **Properties:**
  - Most of the sampling points will fall into big clusters
  - In extreme cases, small clusters will have no sampling points

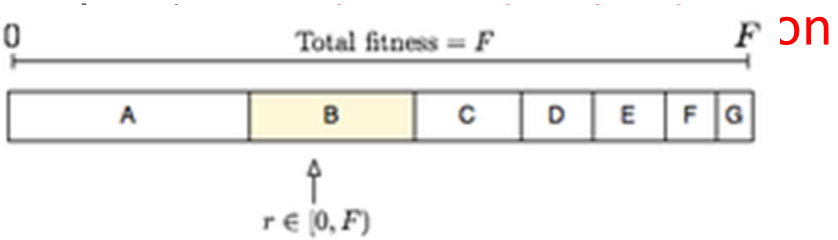


Weighted Random Sampling

- Data items are weighted and the probability of selecting each item is determined by its relative weight.
- Algorithm?
  - **Input**: a set D of n weighted items, each $d_i \in D$ is associated with $w_i \geq 0$
  - **Output**: a set of samples $S \subseteq D$, with $|S| = k$

Weighted Random Sampling

- Data items are weighted and the probability of selecting each item is determined by its relative weight.
- Algorithm:
  - **Input**: a set D of n weighted items
  - **Output**: a set of samples $S \subseteq D$, with $|S| = k$
  1. For $\mathrm{r} = 1$ to k do
     - Update $p_i = \frac{w_i}{\sum_{d_j \in D \setminus S} w_j}$ be the probability of item $d_i \in D \setminus S$ to be selected in round r
     - Randomly select an item $d_i \in$ on [*]
     - Insert it into S
  2. Return S



[*] https://en.wikipedia.org/wiki/Fitness_proportionate_selection

Weighted Random Sampling

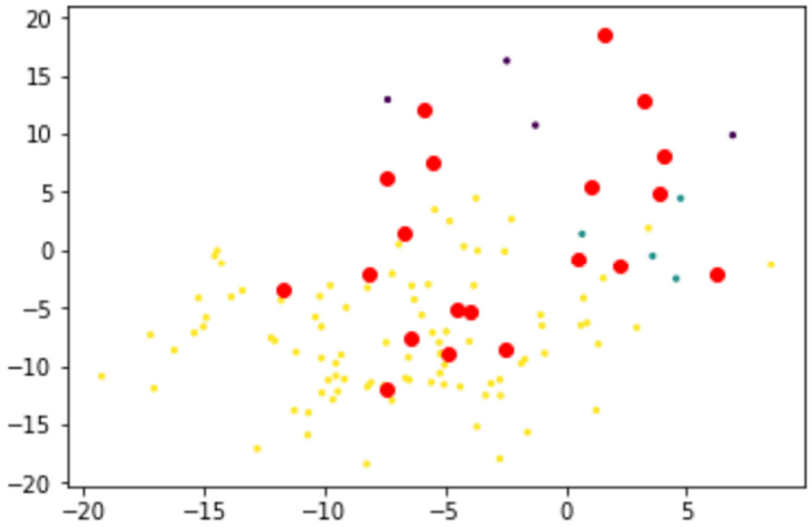
- **Pros**: improve Simple Random Sampling by putting the weights for small clusters
- **Cons**: you have to assign weights for each data point yourself

Stratified Sampling

- **Definition**:
  - A stratified random sample is essentially a series of SRSs performed on subgroups of a given population.
  - The SRS taken within each group in a stratified random sample need not be of the same size
- **General procedure:**
  - Divide data points to group (if not available before-hand)
  - Sample for each group by random sampling
- **Pseudo-code**:
  - sample = SRS(partition1, k_1) ∪ SRS(partition2, k_2) ∪ …

Stratified Sampling

- **Properties:**
  - Uniform distribution across different groups (if we set stratified sample of the same size)



Big Data Visualization: Variety

- Data comes from different sources and needs to be filtered properly
- Application: Facebook News Feed
  1. **Novelty**: show new information
  2. **Relevance**: show feeds from close friends
  3. **Diversity**:
     - Showing most recent feeds is redundant for strangers
     - Show friend is redu

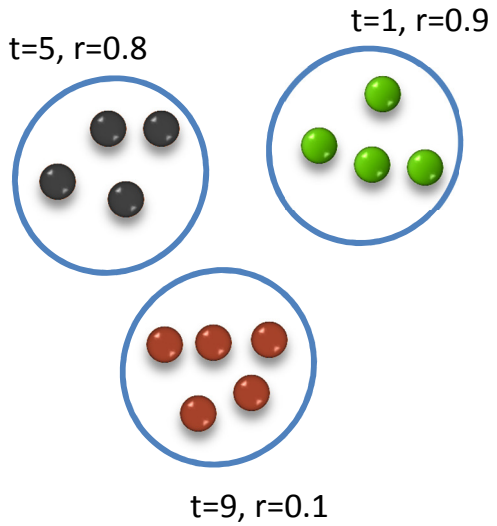
https://blog.bufferapp.com/facebook-news-feed-algorithm

Variety: Techniques

- Formulated as an **optimization problem**:
  - Each post is represented as a data point
  - Each data point has a relevance degree to user interest
  - Each data point is associated by a time index.
  - A similarity function for a pair of posts
- **Objective function**: relevance + novelty – similarity

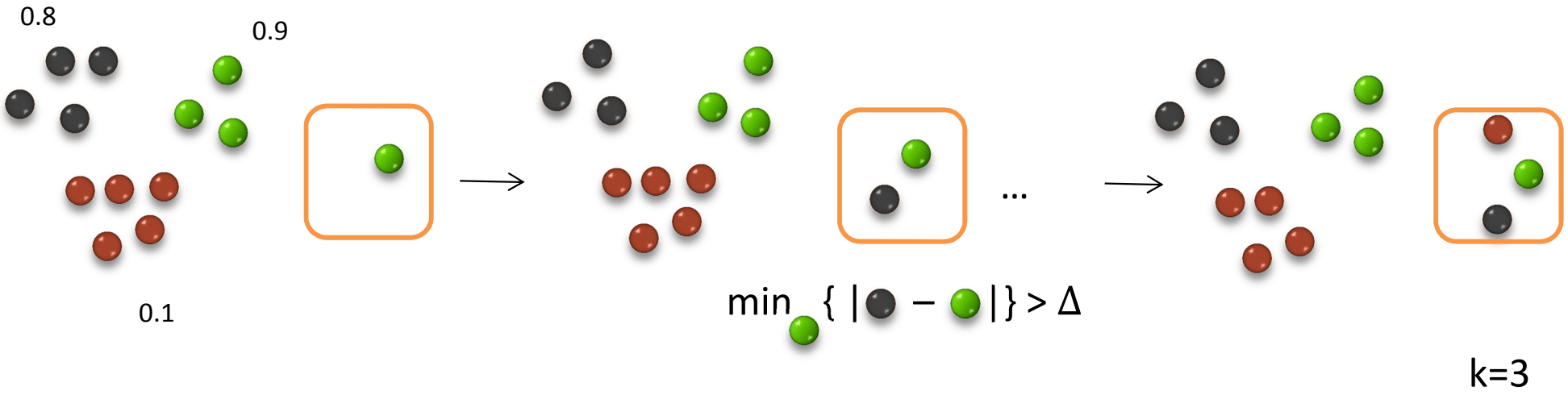
Multi-objective problem

- **Goal:**
  - Select k data items to create an output set
  - Maximize criteria simultaneously:
    - **Novelty**: the more recent the better
    - **Relevance**: relevance score of each data item for a given query
    - **Diversity**: dissimilarity between data items



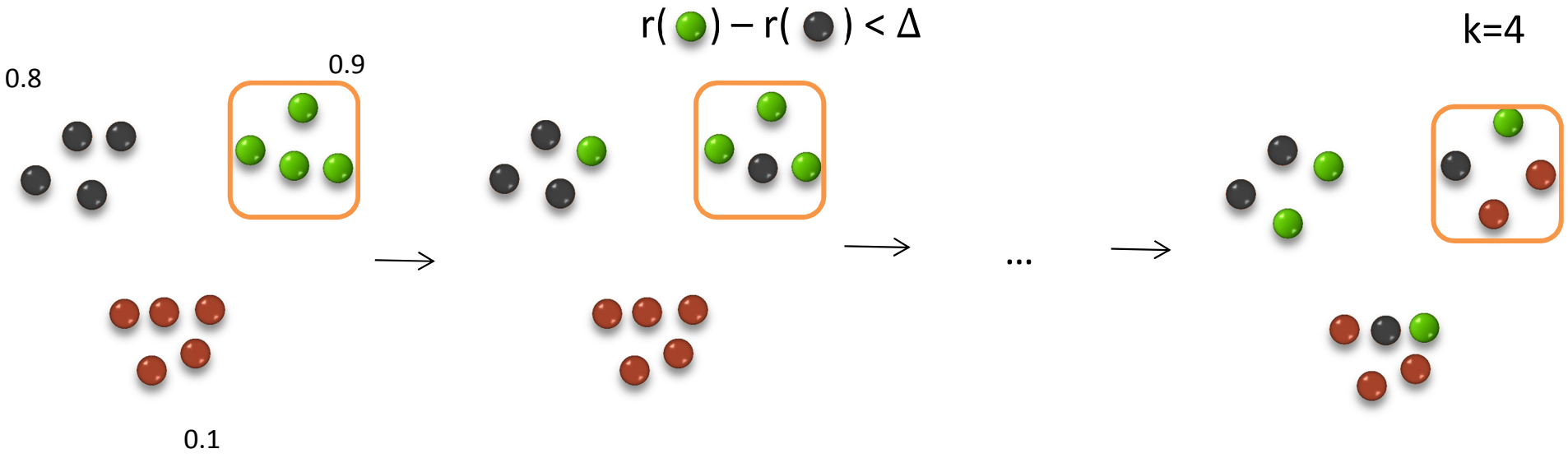
A diversification algorithm

- Motley [15]: constructs the output by incrementally adding items in the decreasing order of relevance and maximizing the minimum dissimilarity.
  1. Traverse items in the decreasing order of relevance
  2. Add an item to output if the minimum dissimilarity with other selected items is larger than a threshold Δ



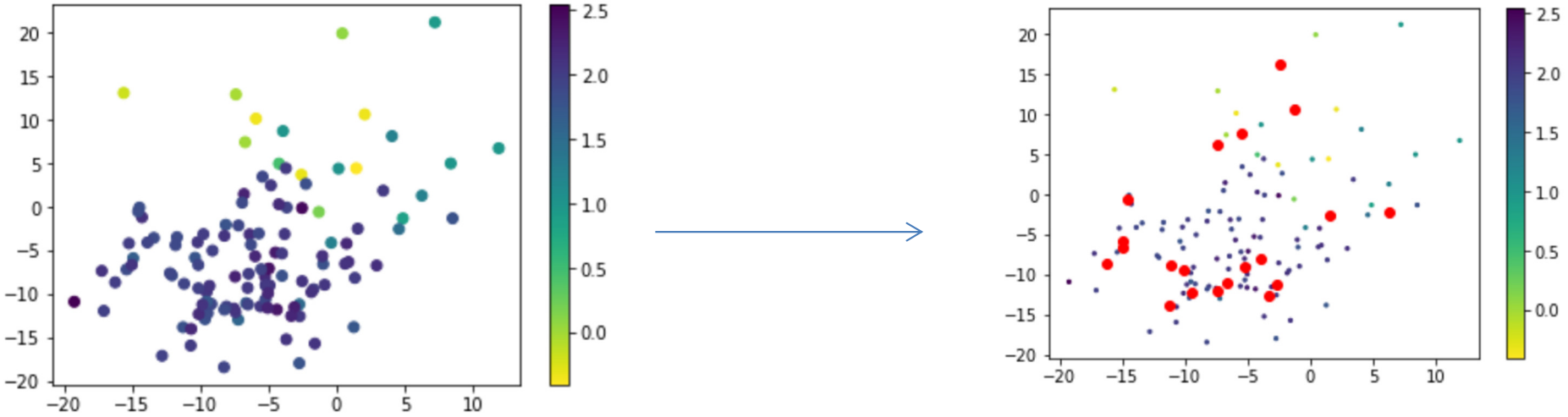
Another diversification algorithm

- Swap [14]: firstly focuses on relevance and then gradually improves diversity by swapping out the least dissimilar items.
  1. Take top-k highest relevant items
  2. Traverse remaining items in decreasing order of relevance
  3. Swap with the median in the output if the relevance drop is less than a threshold Δ



Diversification Algorithm: Example

- Similar to sampling if the relevance information is missing or similar to the clustering structure of the data



Facebook News Feed Algorithm

- Design principles:
  - **Friend and family come first:** The main objective of the News Feed is to connect people with their friends and family. So posts from friends and family are prioritized. After those posts, Facebook found that people want their feed to inform and entertain them.
  - **A platform for all ideas:** Facebook welcomes all ideas while making sure that everyone feels and is safe. They aim to deliver stories that each individual wants to see the most, based on their actions and feedback.
  - **Authentic communications:** Facebook prioritizes genuine stories over misleading, sensational, and spammy ones.
  - **You control your experience:** Individuals know themselves best. So Facebook creates features (such as unfollow and see first) to let people customize their Facebook experience.

https://blog.bufferapp.com/facebook-news-feed-algorithm

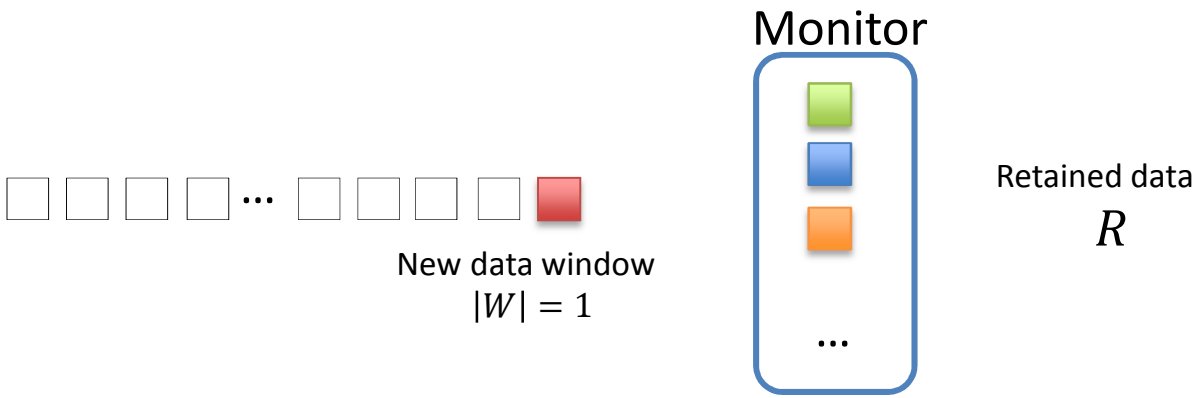
Big Data Visualization: Velocity (*: Advanced)

- **Data** is big but contains many low-valued items: redundant, overlapping, sparse
- Data comes in stream and with high speed
  - E.g. social media, Internet of Things (IoT)
- Data monitoring systems have limited storage

→ Need to effectively decide which old data to replace

Velocity (*)

- Data retaining protocol
  - The monitor system can only visualize a limited number of data items
  - With a new data arrival:
    - Which one of the existing items will be replaced?



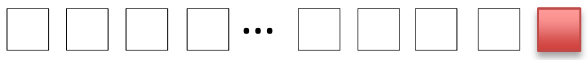
Current Data $D = R \cup W$

Velocity (*)

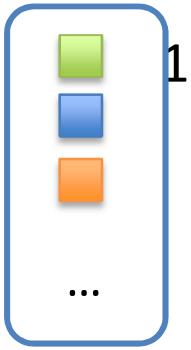
- Data retaining protocol
- Challenges:
  - **Information regret**:
    - Replaced items cannot be recovered
    - New items might become less useful than the old ones

    → Need to quantify the utility of a set of data items
  - **Data has correlations**: processing 1 item at a time loses collective information

    → Need to process data as batch, 1



New data window
$|W| = 1$

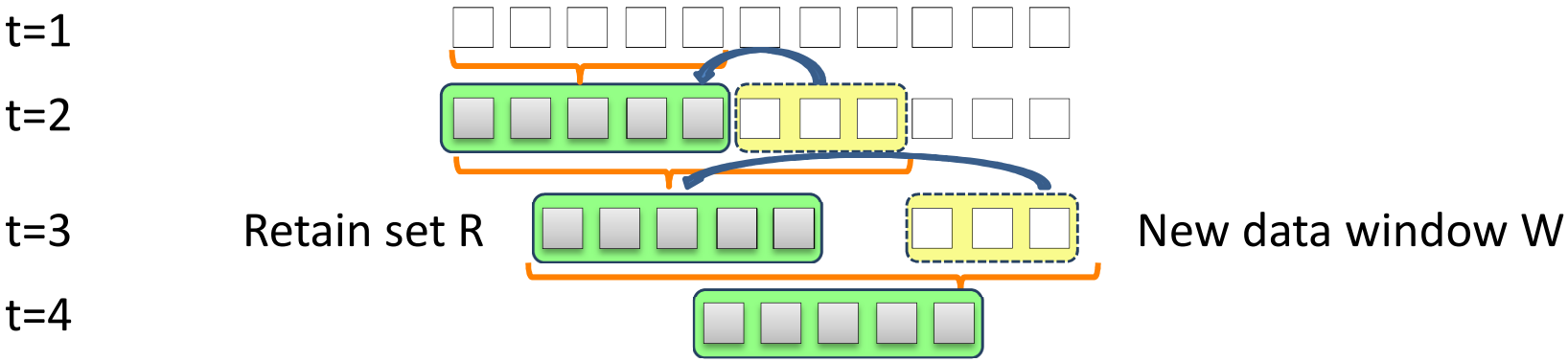


Retained data
R

Current Data $D = R \cup W$

Data Retaining Problem (*)

- The retaining protocol can be formulated as an optimization problem
  - **Optimization problem**: find a k-subset of $D_t = R_t \cup W_t$ with maximal utility
    - $R_{t+1} = argmax_{S \subseteq D_t, |S|=k} f(S)$



Data Retaining Problem (*)

- The retaining protocol can be formulated as an optimization problem
  - **Optimization problem**: find a k-subset of $D = R \cup W$ with maximal utility
    - $argmax_{S \subset D, |S|=k} f(S)$
  - **Utility function**: on a set S of data items
    - $f(S) = 2\sum_{d \in S} q(d)r(d) - \sum_{d,d' \in S} M(d,d')r(d)r(d')$

relevance≠timeliness

diversity

  - where $r(d)$ is the relevance, $M(d,d')$is the similarity between two data items and $q(d)$ is the timeliness of a data item.

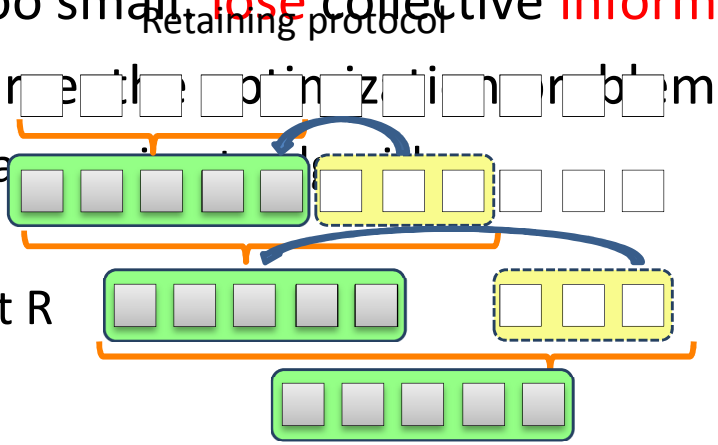
→ first term is about relevance and timeliness of data items

→the second term is about the diversity of data items

Data Retaining Problem (*)

- The retaining protocol can be formulated as an optimization problem
  - **Optimization problem**: find a k-subset of $D = R \cup W$ with maximal utility
    - $argmax_{S \subset D, |S|=k} f(S)$
  - **Utility function**: on a set S of data items
    - $f(S) = 2\sum_{d \in S} q(d)r(d) - \sum_{d,d' \in S} M(d,d')r(d)r(d')$
  - **Trade-offs:**
    - If |W| is too small: lose collective information
    - If |W| is large, the optimization problem becomes NP-hard

      → Need a



Data Retaining: Approximate Algorithm (*)

- **Algorithm**: find a subset $S \subset D, |S| = k$ with maximal utility $f(S)$
  - Compute a ranking score $\mathrm{s}(d) = 2q(d)r(d)$ for each data item
  - Initialize the retain set $S = \emptyset$
  - For k iterations do:
    1. Pick an non-selected item x with maximal ranking, i.e.
       $$\mathrm{x} = argmax_{d \in D \setminus S} s(d)$$
    2. Select x into S, i.e. $S = S \cup \{x\}$
    3. Update the ranking score $\mathrm{s}(d) = s(d) - 2r(x)M(d,x)r(d)$ for remaining data items
  - Return S
- **Complexity**: $O(k^2)$
- **Approximate ratio**: $\left(1 - \frac{1}{e}\right) \approx 0.63$, i.e. the output is guaranteed to have utility $\geq 63\%$ of the optimal solution

Data Retaining: Performance Improvement (*)

- In data stream setting, the update time of approximation algo. is slow:
  - With W new data items, the update time is $O(k \times (k + |W|))$
- Improvement: greedy algorithm [14]
  - Replace an old item in R with a new item in W such that the utility gain is maximal
  - The update time is only $O(k)$
  - However, the utility is sacrificed and the approximation ratio is no longer bounded

References

[1] Benzi Kirell Mael. Data Visualization, Autumn 2017. http://edu.epfl.ch/coursebook/en/data-visualization-COM-480

[2] West Robert. Applied Data Analysis, Autumn 2017. http://edu.epfl.ch/coursebook/en/applied-data-analysis-CS-401

[3] T. Munzner, Visualization Analysis and Design, 2014.

[4] Jacques Bertin, Semilogy of Graphics, 1967.

[5] Heer and Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. 2010

[6] https://www.youtube.com/watch?v=xAoljeRJ3lU&feature=youtu.be

[7] Good Enough to Great: A Quick Guide for Better Data Visualizations

[8] The Power of “Where”

[9] Visual Analysis Best Practices

[10] Visual Analysis for Everyone

[11] 6 Best Practices for Creating Effective Dashboards

[12] The Power of R and Visual Analytics

[13] Lei Yu, Jieping Ye, Huan Liu. Dimensionality Reduction for Data Mining: Techniques, Applications and Trends. 2007.

[14] Yu, Cong, Laks Lakshmanan, and Sihem Amer-Yahia. "It takes variety to make a world: diversification in recommender systems." *Proceedings of the 12th international conference on extending database technology: Advances in database technology*. ACM, 2009.

[15] Jain, Anoop, Parag Sarda, and Jayant R. Haritsa. "Providing diversity in k-nearest neighbor query results." *PAKDD*. Vol. 4. 2004.

[16] Glyph-based visualization http://vis.cs.ucdavis.edu/vis2014papers/VIS_Conference/tutorials/Glyph-based_Visualization/Glyph-Tutorial-vis2014.pdf