VISUALIZING TABULAR DATA

AND A DEEP DIVE INTO DATA CHARTS

Petra Isenberg

Inria informatics mathematics

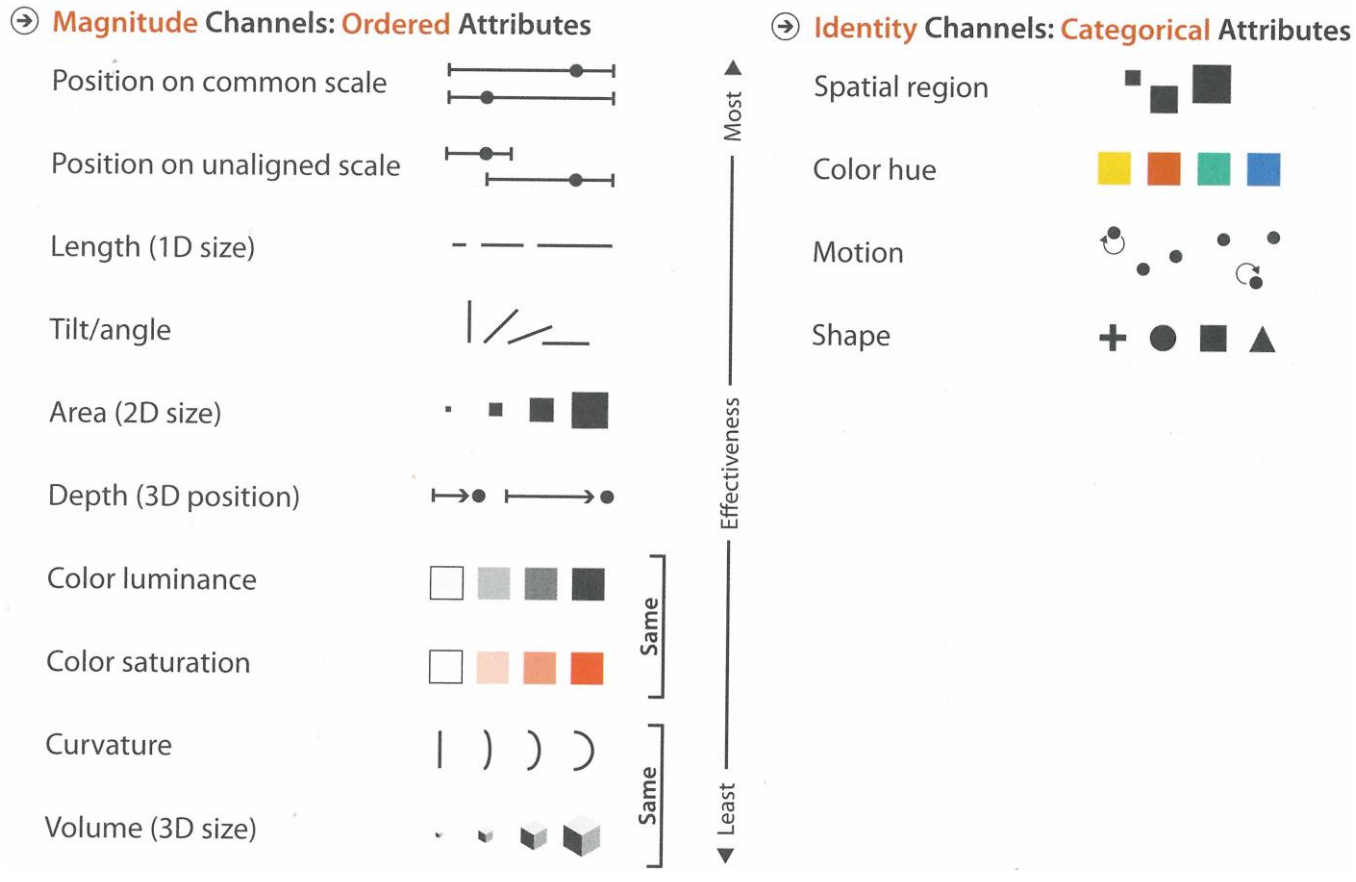
université PARIS-SACLAY

RECAP

you have learned about

- visual channels and marks
- that their perceptual properties matter

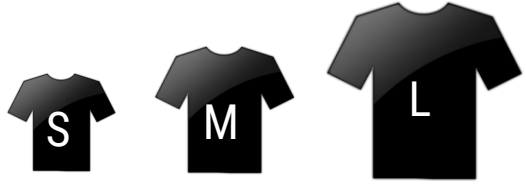
RECAP


Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Most
Effectiveness
Least
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape

RECAP

DATA TYPES

ORDINAL (ranking)



NOMINAL (categorical)

QUANTITATIVE (numerical)

TODAY

How to turn something like this … into a data representation

Vélib' : Disponibilité temps réel

Informations | Tableau | Carte | Analyse | Export | API

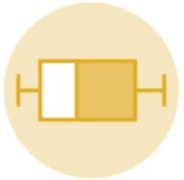
	Code de la station	Nom de la station	Etat des stations	Etat du Totem	Nombres de bornes en station	Nombre de bornes disponibles	Nombre de vélo mécanique	Nombre vélo électrique	Achat possible en station (CB)
1	11037	Faubourg Du Temple - Republique	Close	yes	39	36	2	1	yes
2	11104	Charonne - Robert et Sonia Delauney	Operative	yes	20	17	2	1	no
3	14111	Cassini - Denfert-Rochereau	Close	yes	25	25	0	0	yes
4	12109	Mairie du 12ème	Operative	yes	30	30	0	0	no
5	5110	Lacépède - Monge	Operative	yes	23	16	6	1	yes
6	17038	Grande Armée - Brunel	Operative	yes	62	40	20	2	yes
7	10152	Gare du Nord - Place de Valenciennes	Operative	yes	25	18	6	1	yes
8	13007	Le Brun - Gobelins	Operative	yes	48	47	1	0	yes
9	41301	Bois de Vincennes - Gare	Operative	yes	51	39	8	4	yes
10	31024	Romainville - Vaillant-Couturier	Operative	yes	38	35	2	1	no
11	15028	Grenelle - Dr Finlay	Operative	yes	71	63	7	2	yes
12	16118	Michel-Ange - Parent de Rosan	Operative	yes	26	25	0	1	no
13	20035	Pyrénées - Ménilmontant	Operative	yes	26	24	0	2	no
14	10027	Dunkerque - Alsace	Operative	yes	18	10	8	0	no
15	8048	Marceau - Chaillot	Operative	yes	21	8	9	4	no
16	14013	Liard - Amiral Mouchez	Operative	yes	1	1	1	0	yes
17	5024	Place Monge	Close	yes	21	21	0	0	no
18	7018	Ségur - d'Estrées	Operative	yes	19	8	10	1	no
19	10029	Dunkerque - Rocroy	Operative	yes	23	16	6	1	no
20	8009	Gare Saint-Lazare - Isly	Operative	yes	27	14	10	3	yes
21	8036	Lisbonne - Monceau	Operative	yes	33	7	25	1	yes
22	17040	Pereire - Ternes	Operative	yes	48	44	2	2	yes
23	31708	Noisy le Sec - Jean-Baptiste Clément	Operative	yes	22	21	0	1	no
24	10105	Mazagran - Bonne Nouvelle	Operative	yes	26	15	9	2	yes

HOW TO CHOOSE?


Violin


Density


Histogram


Boxplot


Ridgeline


Line plot


Area


Stacked area


Streamchart


Scatter


Heatmap


Correlogram


Bubble


Connected scatter


Density 2d


Barplot


Spider / Radar

Wordcloud


Parallel


Lollipop


Circular Barplot


Treemap


Doughnut


Pie chart


Dendrogram


Circular packing

https://d3-graph-gallery.com/

FOLLOW A SYSTEMATIC PROCESS

Understand the **data**

what do you have & what does it mean

Understand the **task**

what are the low-level questions and high-level goals

Understand the **stakeholder**

get to know people's background, needs, expectations

Understand the **visualization**

what representation will help fulfill your stakeholders' tasks, questions, interests

QUESTIONS

Often you start from a higher-level question

How can you break it down into something concrete?

The concrete is sometimes called a **task**

The process is sometimes called an **operationalization**

WHO ARE THE BEST MOVIE DIRECTORS?

What is the problem with this question?

WHO ARE THE BEST MOVIE DIRECTORS?

What is the problem with this question?

The answer depends on who is asking

WHO ARE THE BEST MOVIE DIRECTORS?

What is the problem with this question?

The answer depends on who is asking

- A film student?
  She might want to make sure she's talking about someone important
- A hiring manager?
  He might want to hire someone upcoming (but still cheap)
- A journalist?
  Might want to have a list of the "best" directors (we need to find out what "best" means still)

WHO ARE THE BEST MOVIE DIRECTORS?

What is the problem with this question?

The answer depends on who is asking

- A film student?
  She might want to make sure she's talking about someone important
- A hiring manager?
  He might want to hire someone upcoming (but still cheap)
- A journalist?
  Might want to have a list of the "best" directors (we need to find out what "best" means still)

PROXIES

What is “best” ?

We need a partial (imperfect) representation about what the journalist cares about

→ **ideally you should talk to them**

PROXIES

What is “best” ?

Best = has directed many good movies
Good = IMDB rating > 9.8
Many = more than 5

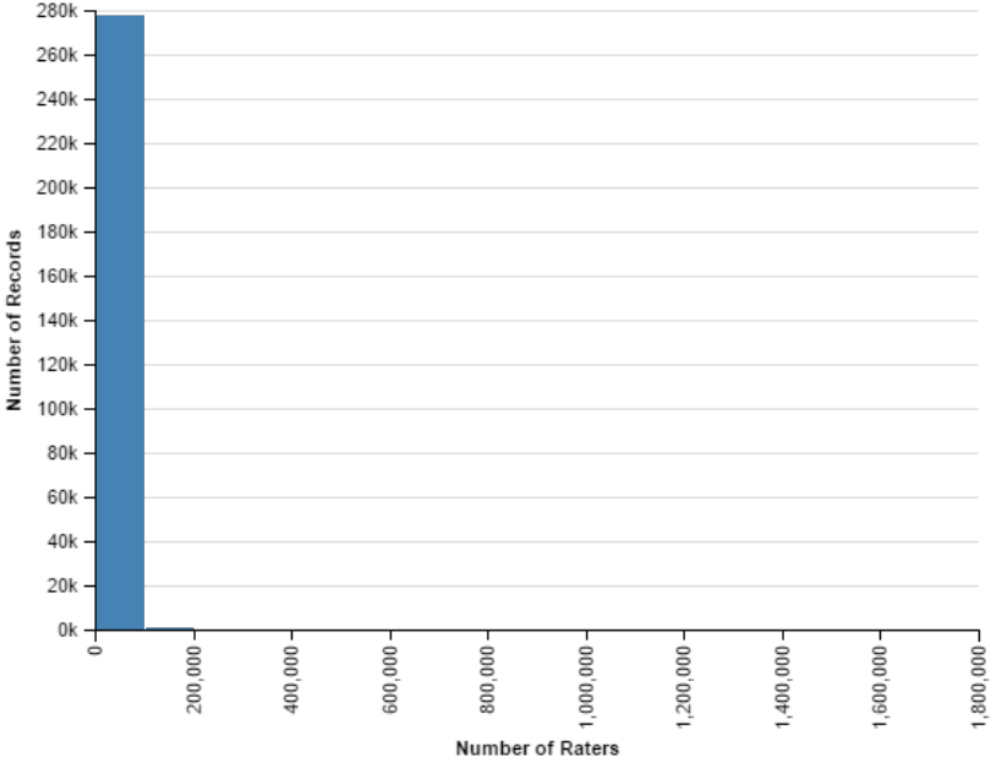
How do you choose these numbers?
Look at your data!

SHOULD WE CARE ABOUT ALL MOVIES?

id	raters	score	title	director
0	12	6.4	#1 (2005)	Breen, James (V)
1	35	6.0	#1 Serial Killer (2013)	Yung, Stanley (I)
2	5	5.8	#137 (2011)	Elliott, Frances
3	11	7.4	#140Characters: A Documentary About Twitter (2…	Beasley, Bryan (I)
4	23	6.7	#30 (2013)	Wilde, Timothy
…	…	…	…	…

What is the rating distribution? Your first opportunity for visualization!

HISTOGRAM



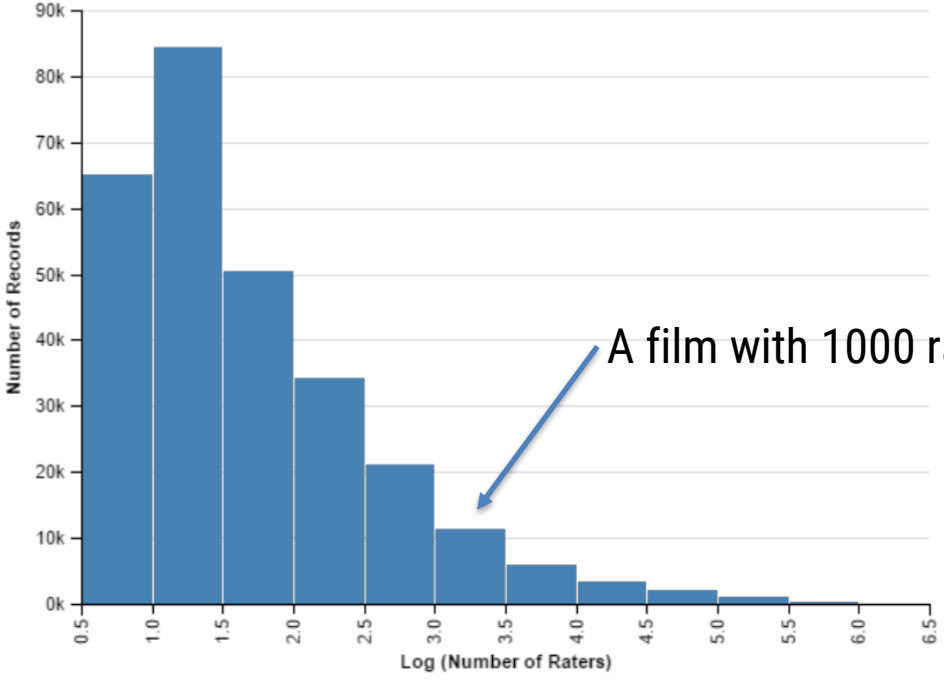
A visualization of one numeric variable

Group the variable into bins

Count the number of occurrences per bin

HISTOGRAM OPERATIONS

Take the logarithm & then bin

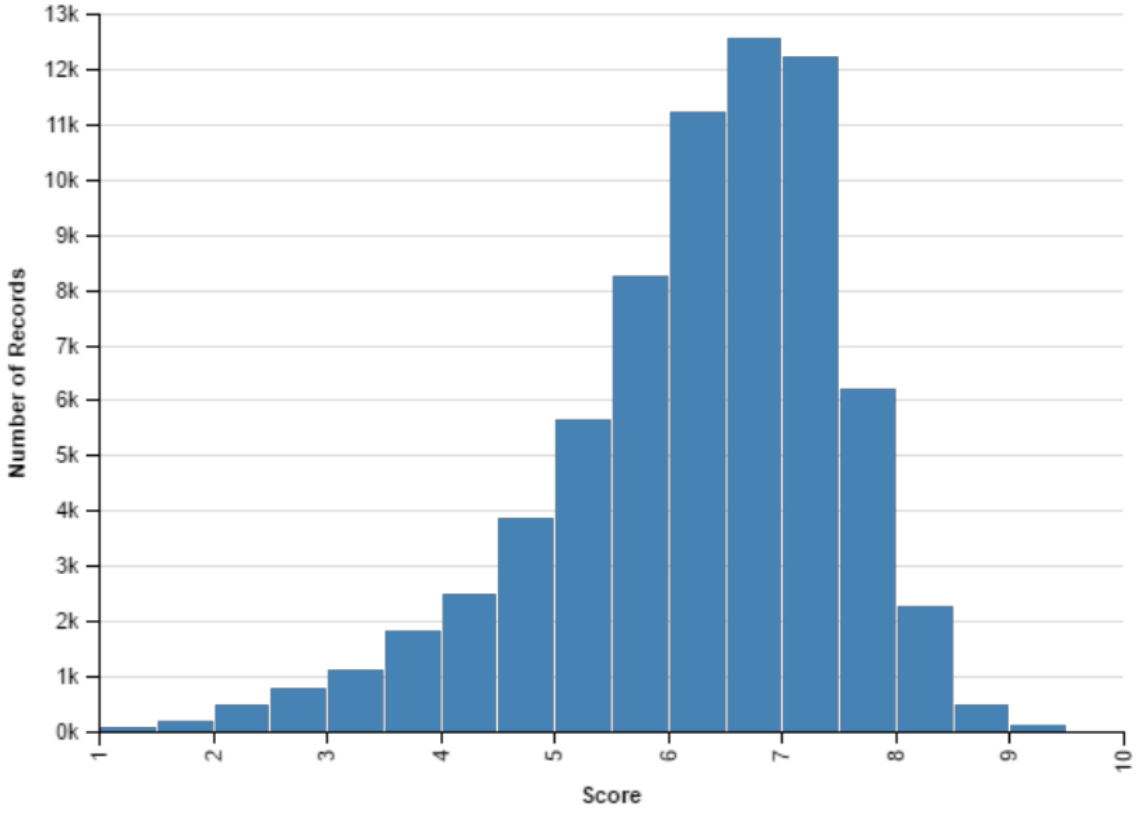


Problem:
Most people don't reason well with logs

Where are movies with 1000 ratings?

Augmenting the plot with means and medians would be helpful

FILTER & PIVOT



We need to narrow down what “good” means

Filter: remove movies which are not known (rated only a few times)

Pivot: change the axes (no scores)

BREAKING DOWN A TASK

Objects: Entities that make up the dataset

movies or directors

Measures: Variable measured for the objects

quality of a director, score for a movie

Grouping: attribute that partitions the data

gender of the director

Actions: what is done with the data

compare, identify, characterize

LETS START VISUALIZING

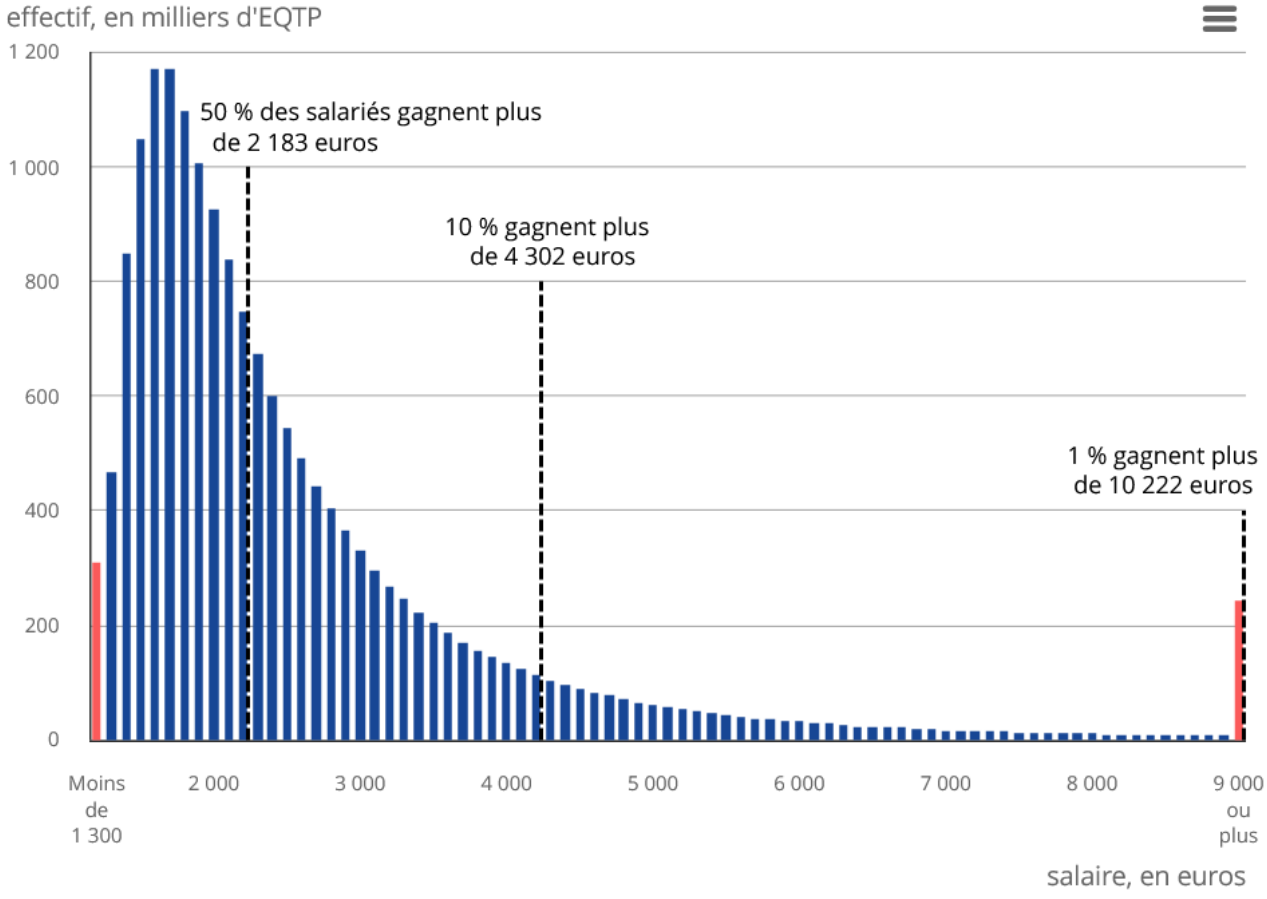
Once you know what you want to do

How is a measure distributed?

Are there values that recur?

Are there outliers?

HISTOGRAM

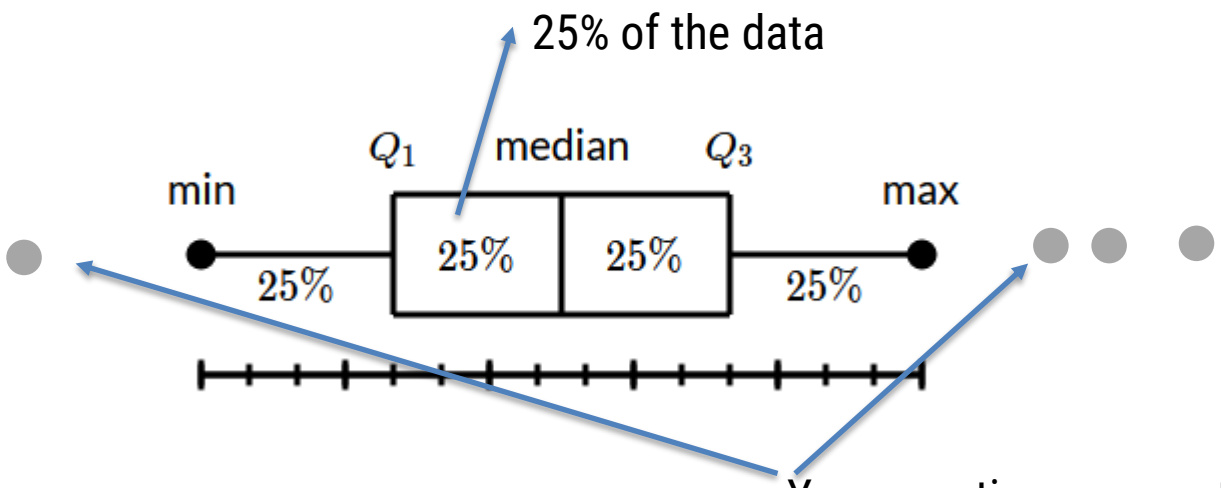


Salaries in the French private sector in 2023
https://www.insee.fr/fr/statistiques/8270416#onglet-2

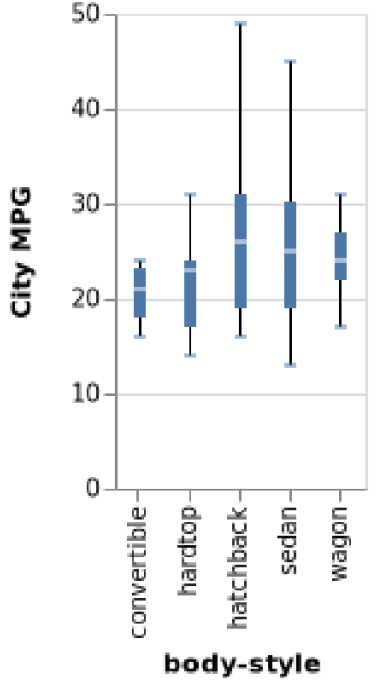
Here:
Continuous quantitative variable

Choosing effective bins is crucial!

BOX PLOTS



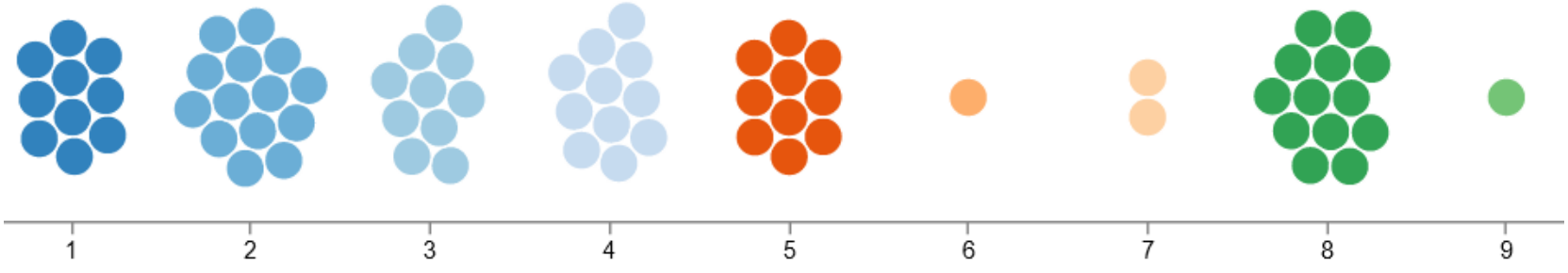
You sometimes see outliers here as dots



Compare multiple distributions

Hides underlying number of entries
Careful when reading boxplots – it's sometimes mean & standard deviation

BEE SWARM PLOTS



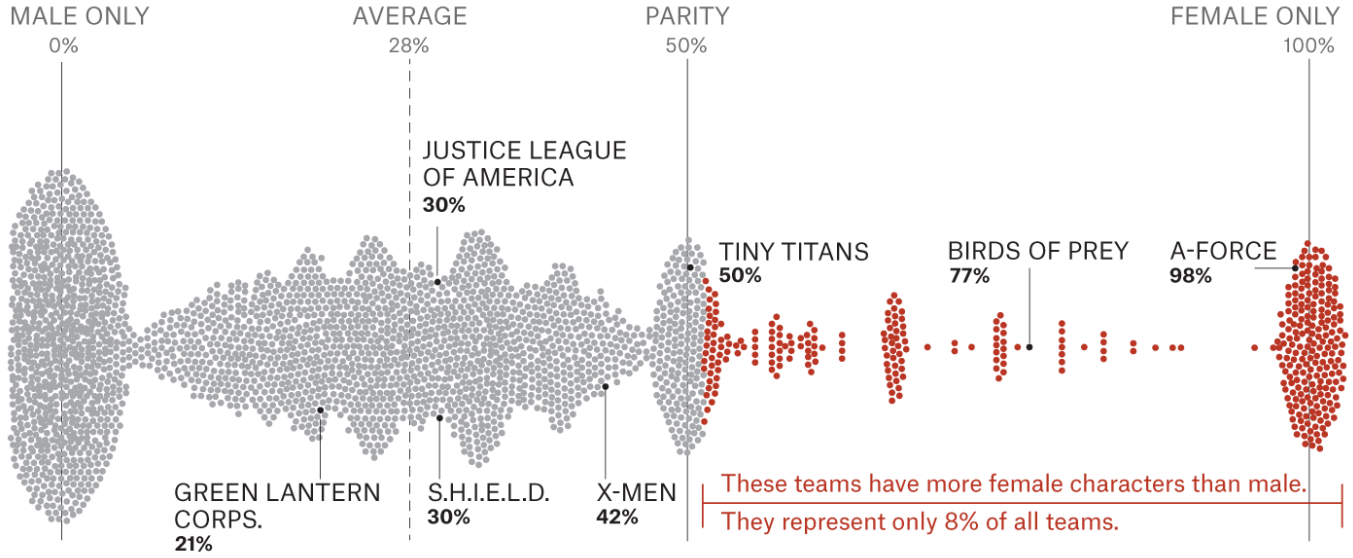
Shows the size of a group of items by clustering the data point per category value

(e.g. one circle per student vote)

ThePudding

Analyzing the Gender Representation of 34,476 Comic Book Characters

Female percentage of every team
Each dot represents one of 2,862 teams in DC and Marvel.



Statistical Computing and Graphics

Violin Plots: A Box Plot-Density Trace Synergism

Jerry L. HINTZE and Ray D. NELSON

Many modifications build on Tukey's original box plot. A proposed further adaptation, the violin plot, pools the best statistical features of alternative graphical representations of batches of data. It adds the information available from local density estimates to the basic summary statistics inherent in box plots. This marriage of summary statistics and density shape into a single plot provides a useful tool for data analysis and exploration.

KEY WORDS: Density estimation; Exploratory data analysis; Graphical techniques.

1. INTRODUCTION

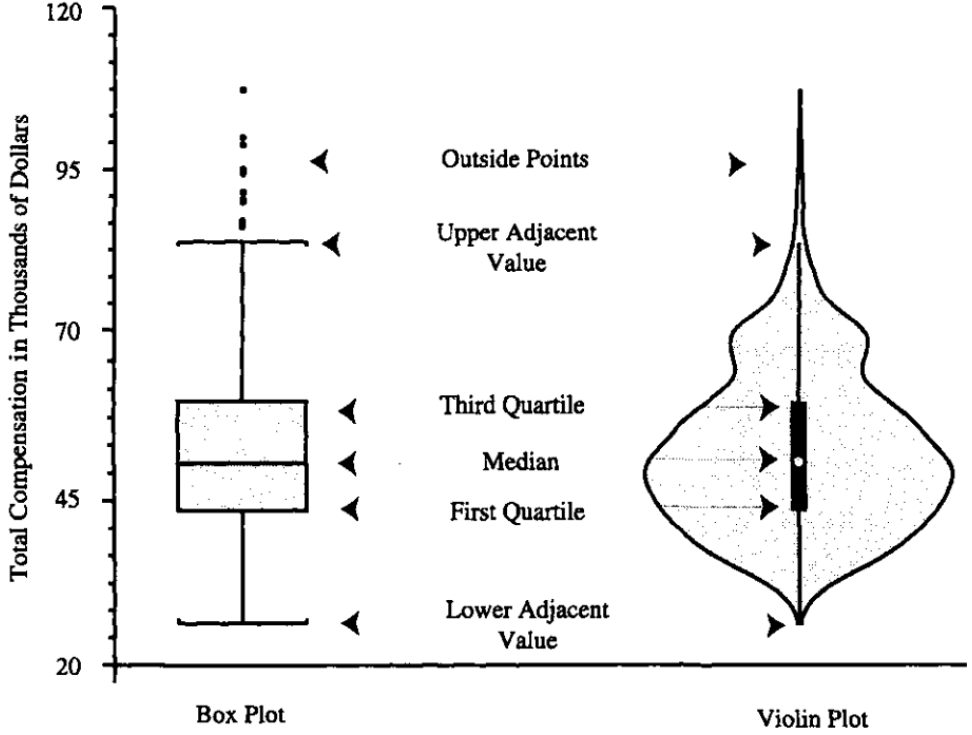
Many different statistics and graphs summarize the characteristics of single batches of data. Descriptive statistics give

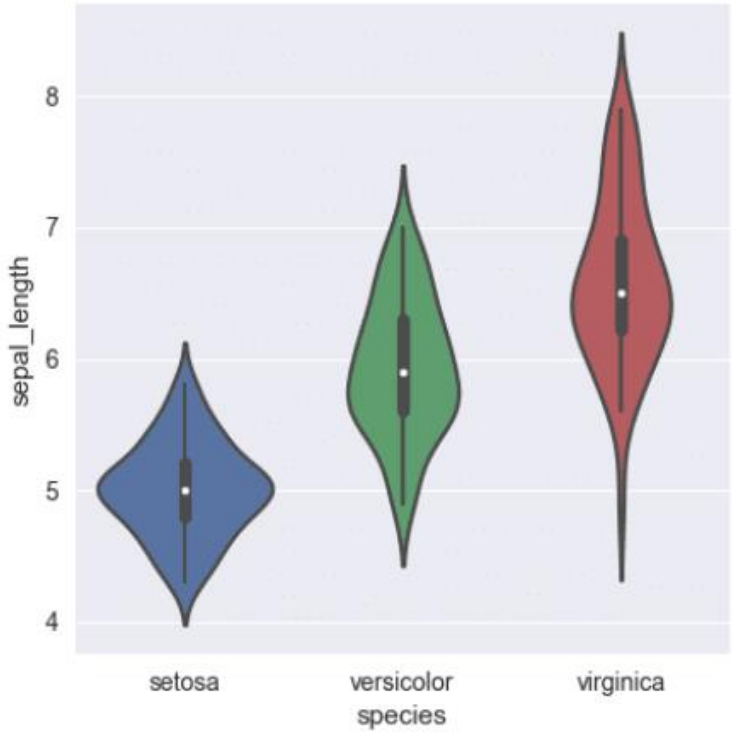
Hoaglin (1981); Chambers, Cleveland, Kleiner, and Tukey (1983); Frigge, Hoaglin, and Iglewicz (1989), and others.

Box plots show four main features about a variable: center, spread, asymmetry, and outliers. As an example, consider the box plot in Figure 1 for the data published by Hamermesh (1994). The ASA Statistical Graphics Section's 1995 Data Analysis Exposition analyzes these data, which report compensation of professors from all academic ranks in the United States. The labels in the diagram identify the principal lines and points which form the main structure of the traditional box plot diagram. As shown, the violin plot includes a box plot with two slight modifications. First, a circle replaces the median line which facilitates quick comparisons when viewing multiple groups. Second, outside points which are traditionally classified as mild and severe outliers, are not identified by individual symbols.

The density trace supplements traditional summary statis-

VIOLIN PLOTS & BOX PLOTS

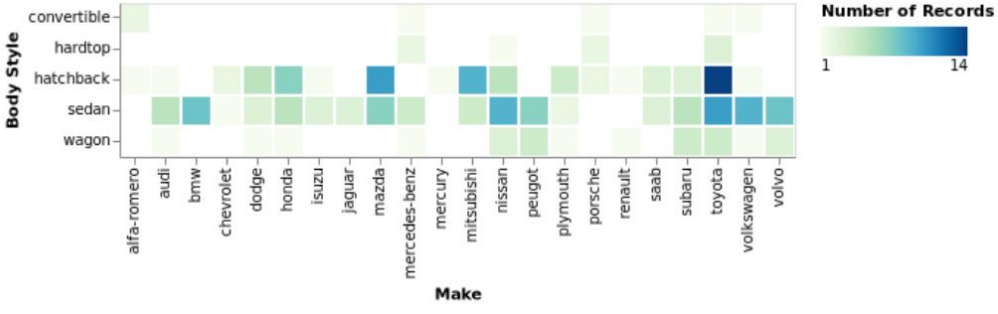




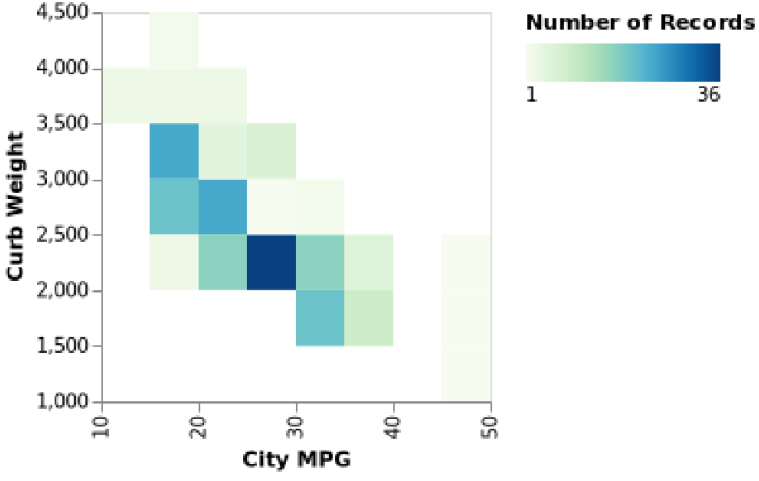
smoothed density of points over a window

https://www.python-graph-gallery.com/violin-plot/

DENSITY PLOT



Categorical density plot

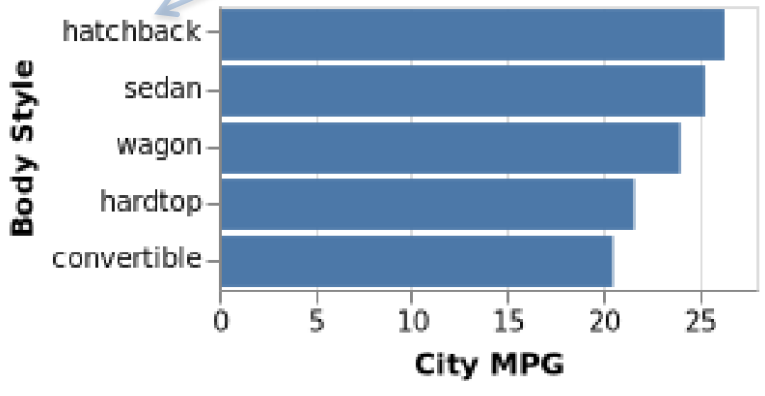


Continuous density plot

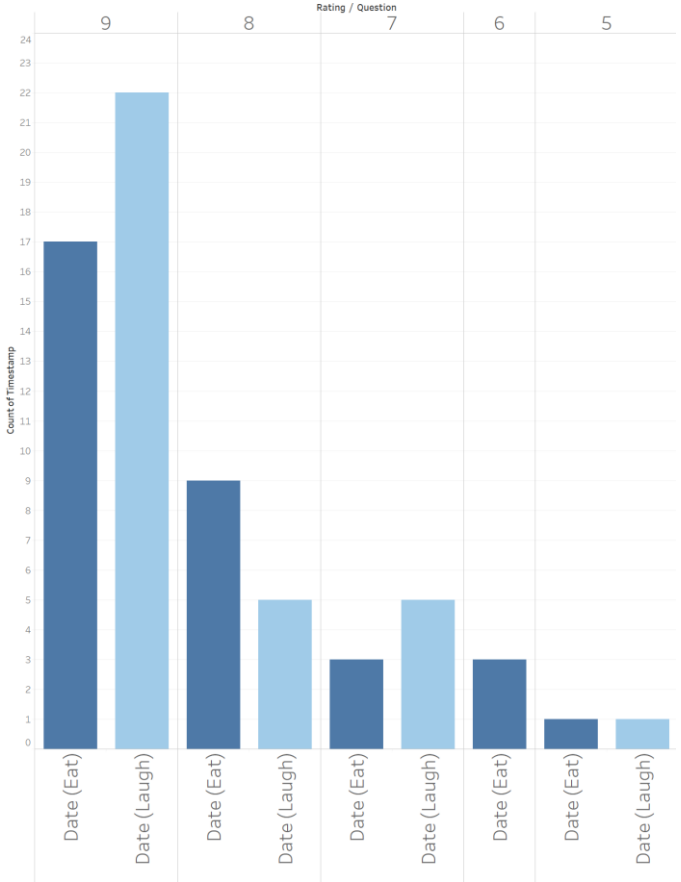
How do groups differ from each other?

BAR CHART

DATA	one quantitative value attribute, one categorical attribute
TASK	lookup and compare values
SCALE	key attribute: dozens to hundreds of levels



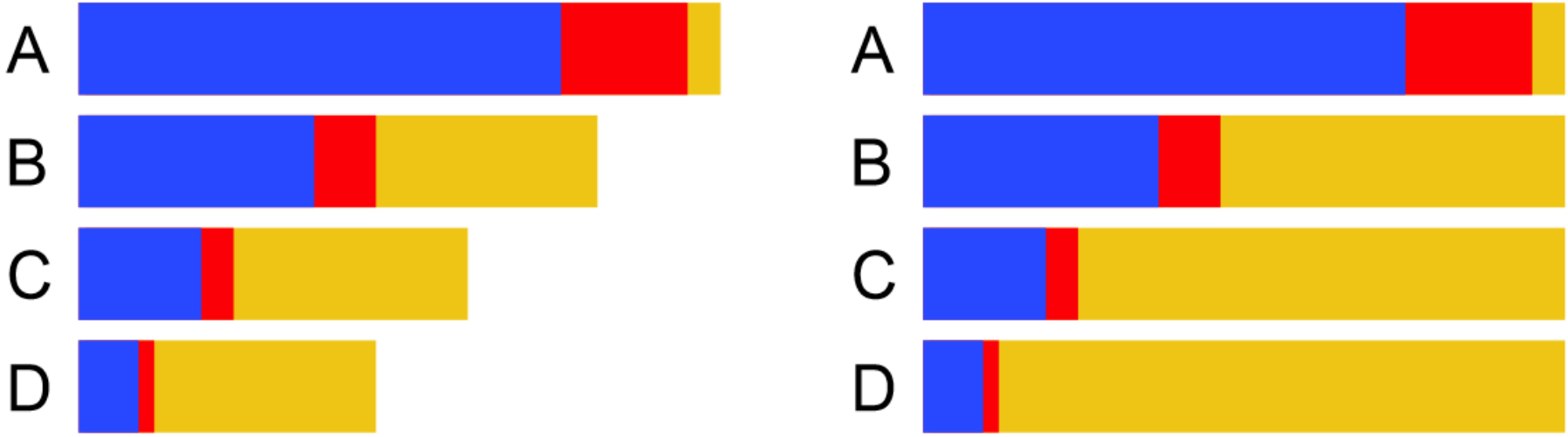
GROUPED BAR CHARTS

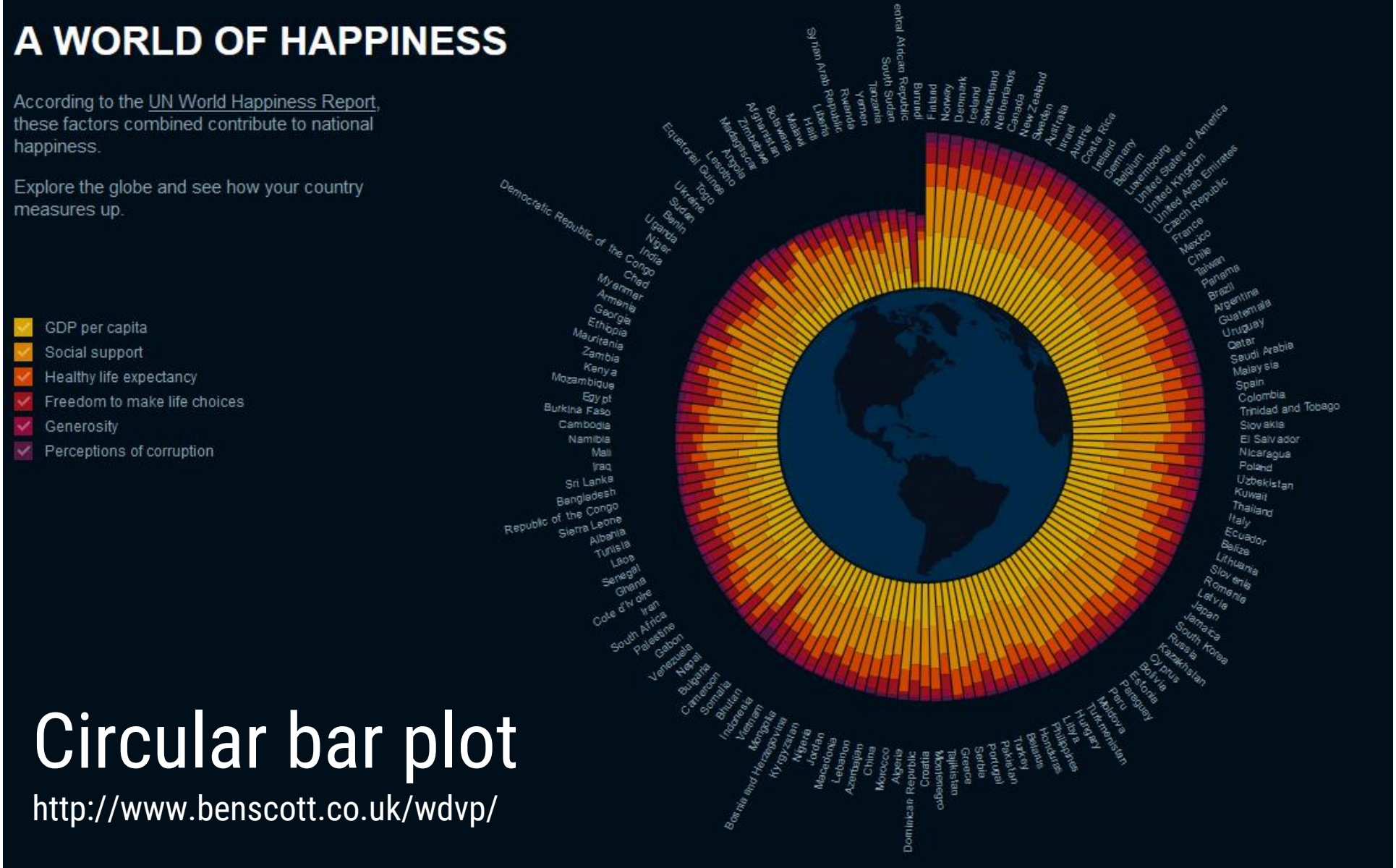


- Sometimes called clustered bar chart
- Bar contains categories of items, grouped by another category

What question can we answer with this chart?

STACKED BARS VS. NORMALIZED STACKED BARS

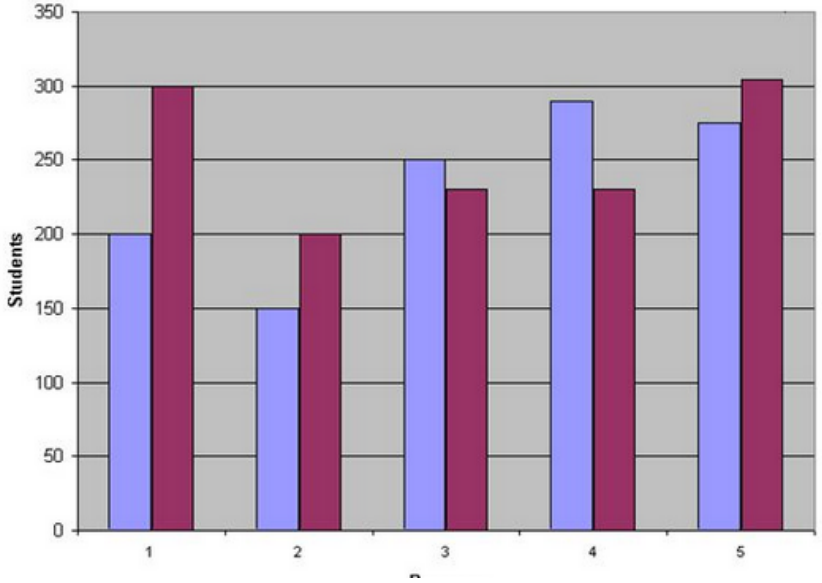


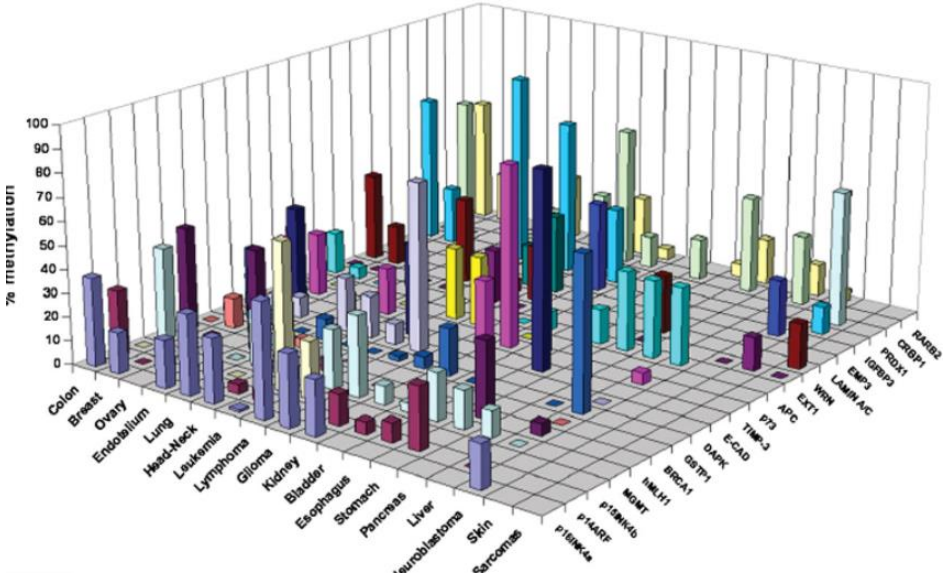


Circular bar plot

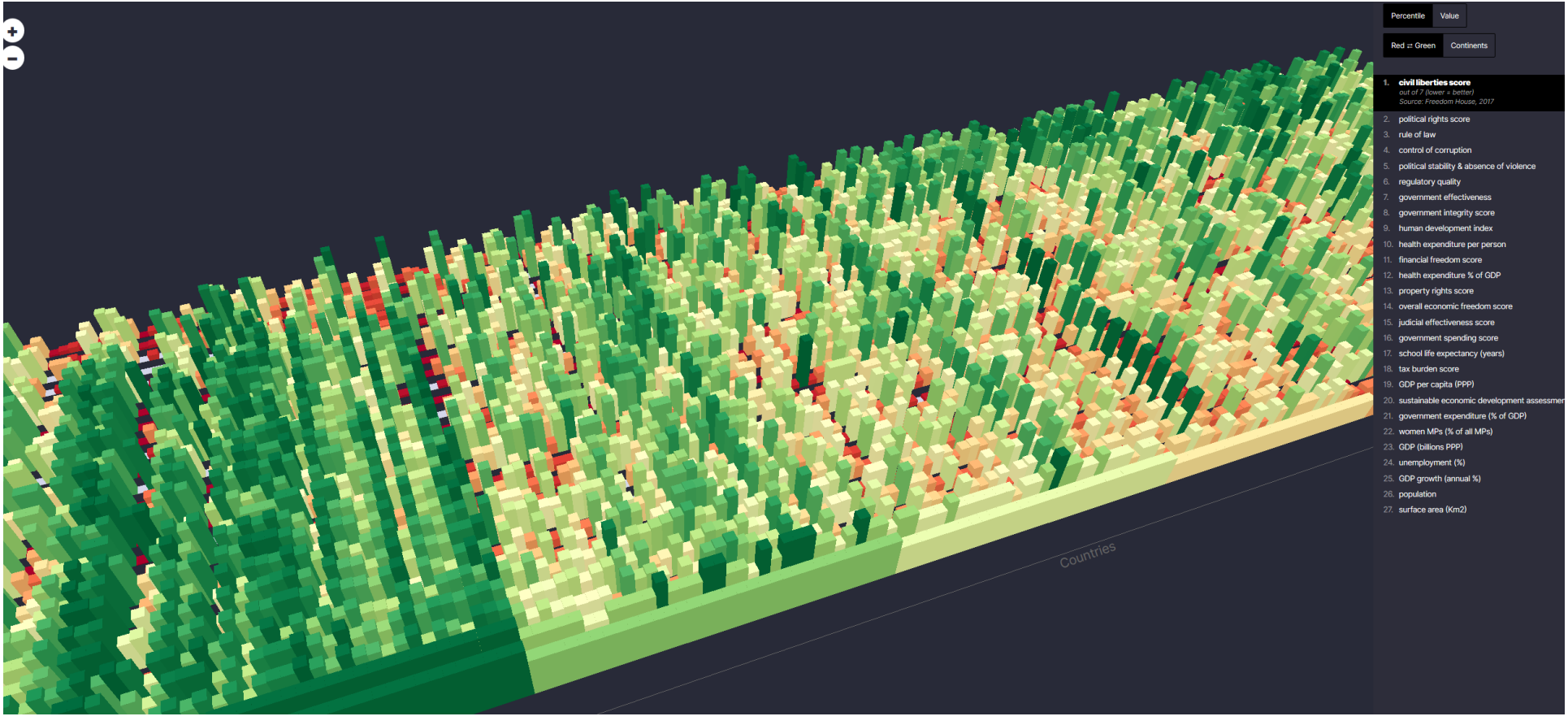
http://www.benscott.co.uk/wdvp/

BAR CHARTS - WHAT TO CONSIDER

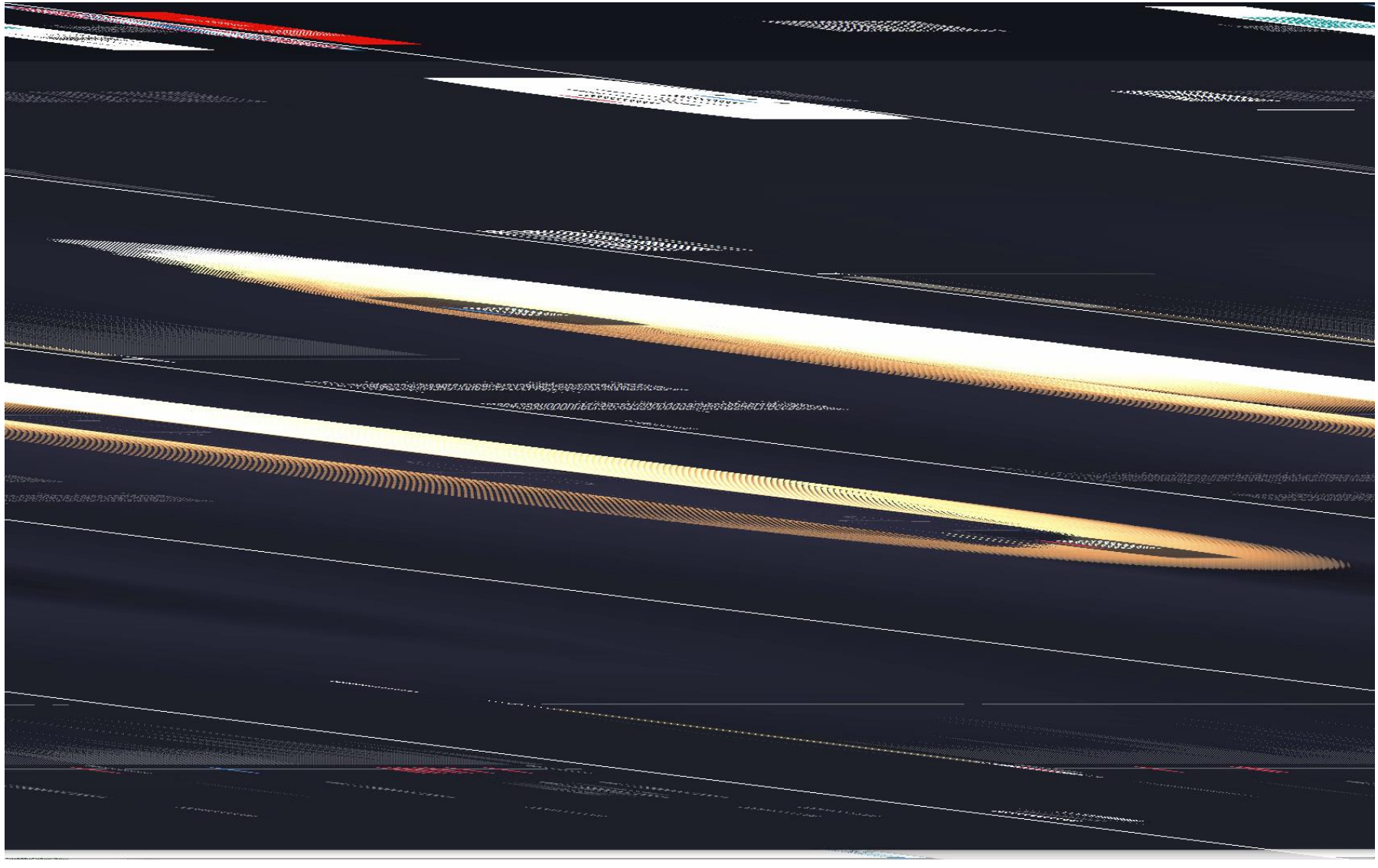




Be careful with 3D bar charts. Only use them when you know what you are doing.

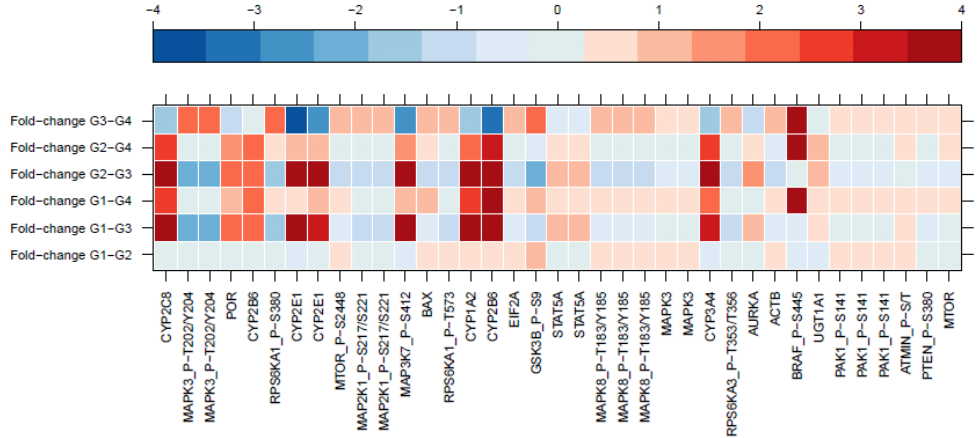


https://wattenberger.com/wdvp



HEATMAP

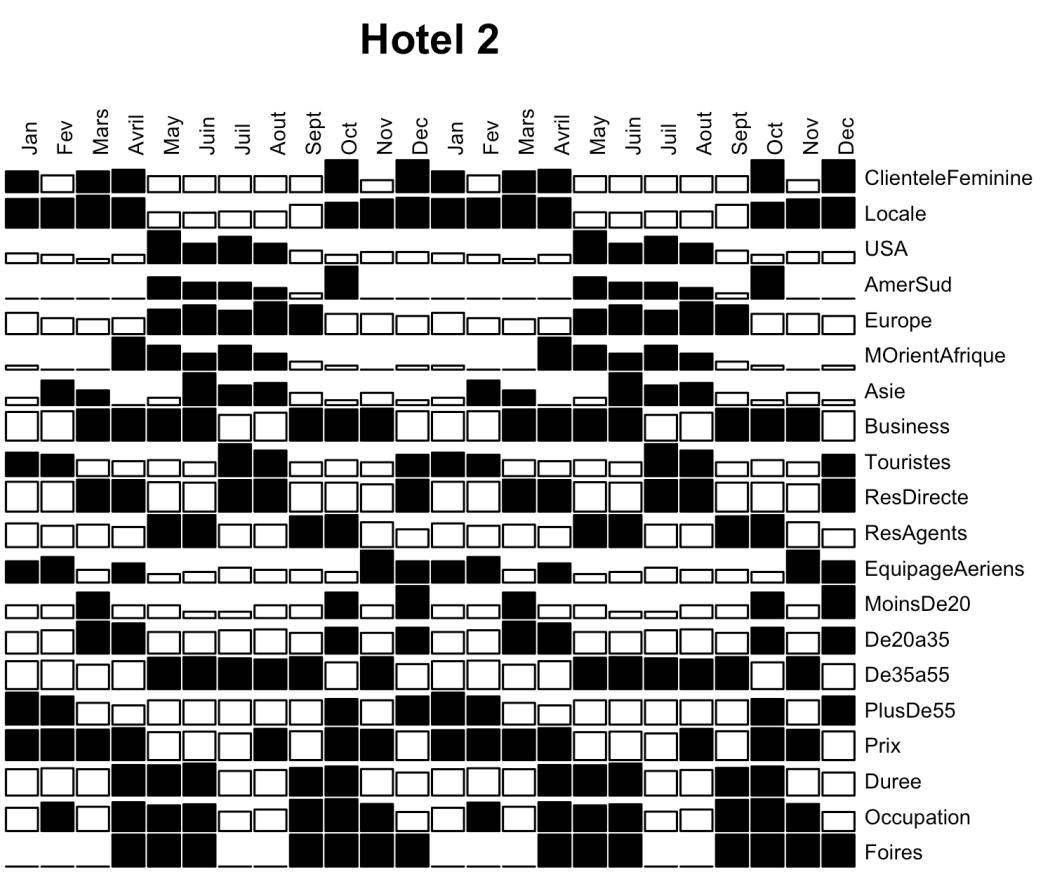
TASK	find clusters, outliers; summarize
SCALE	items: ~1 million (on 1000x1000px), categorical attribute levels: hundreds, quantitative attribute levels: 3-11



Order is very important!

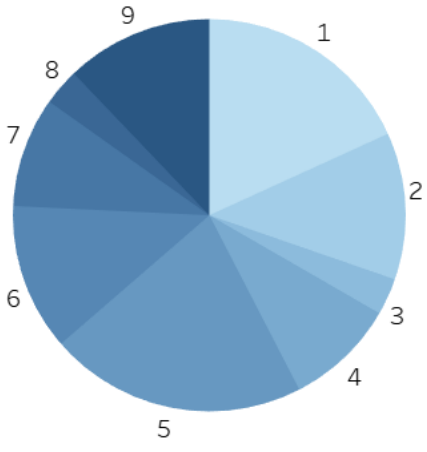
Uses a measure rather than just count (like the density plot)

HEATMAP

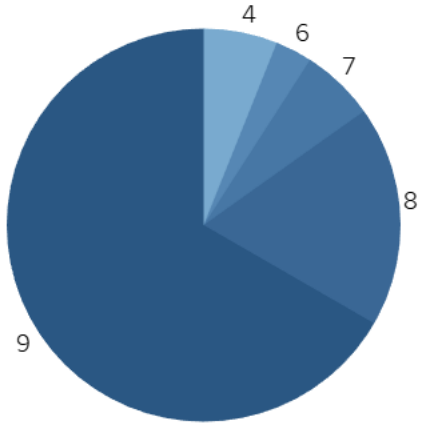

Hotel 2
Jan
Fev
Mars
Avril
May
Juin
Juil
Aout
Sept
Oct
Nov
Dec
Jan
Fev
Mars
Avril
May
Juin
Juil
Aout
Sept
Oct
Nov
Dec
ClienteleFeminine
Locale
USA
AmerSud
Europe
MOrientAfrique
Asie
Business
Touristes
ResDirecte
ResAgents
EquipageAeriens
MoinsDe20
De20a35
De35a55
PlusDe55
Prix
Duree
Occupation
Foires

PIE CHART

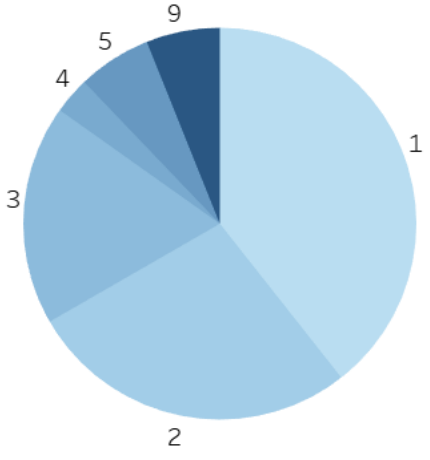
Bar (Argue)



Bar (Laugh)

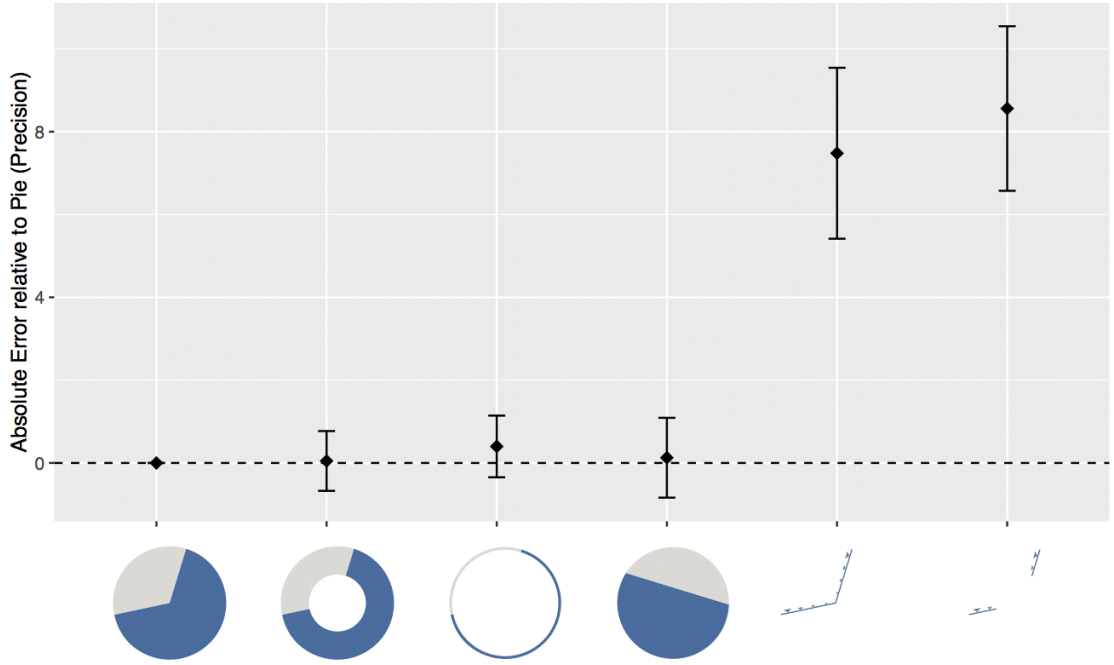


Bar (Run)



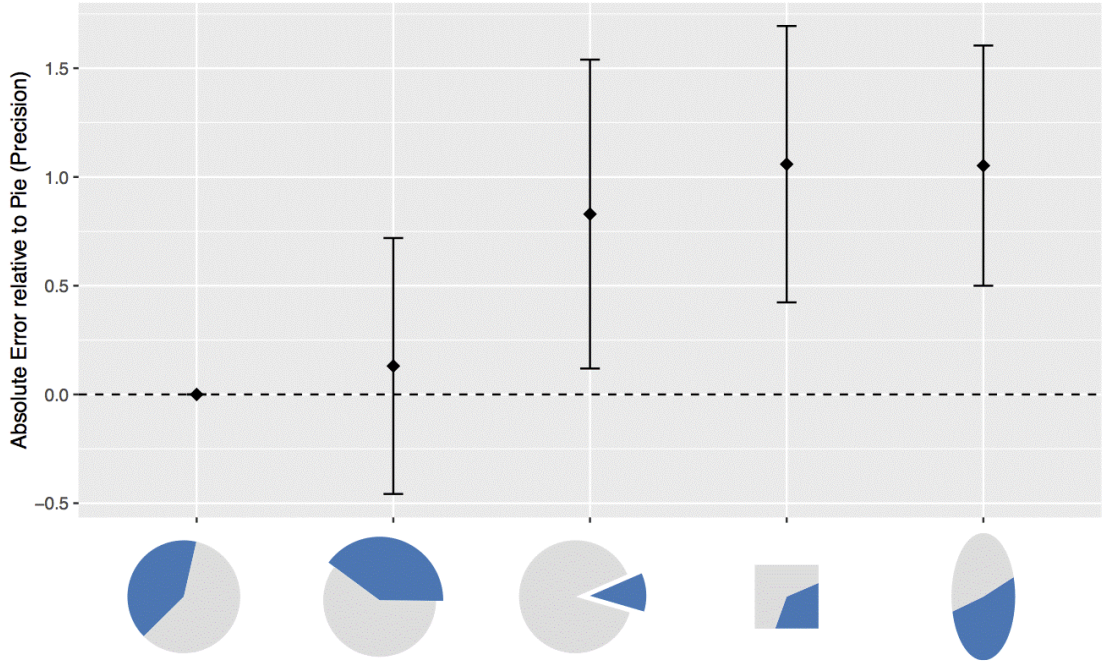
Often not recommended because of angle encoding
BUT: not clear that people actually read angles

HOW DO PEOPLE READ PIE CHARTS?



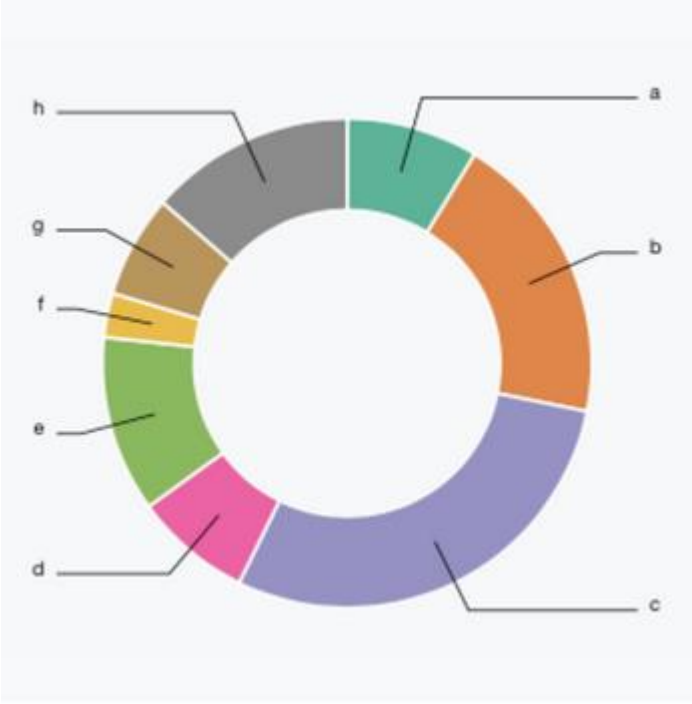
https://eagereyes.org/blog/2016/an-illustrated-tour-of-the-pie-chart-study-results

HOW DO PEOPLE READ PIE CHARTS?



https://eagereyes.org/blog/2016/an-illustrated-tour-of-the-pie-chart-study-results

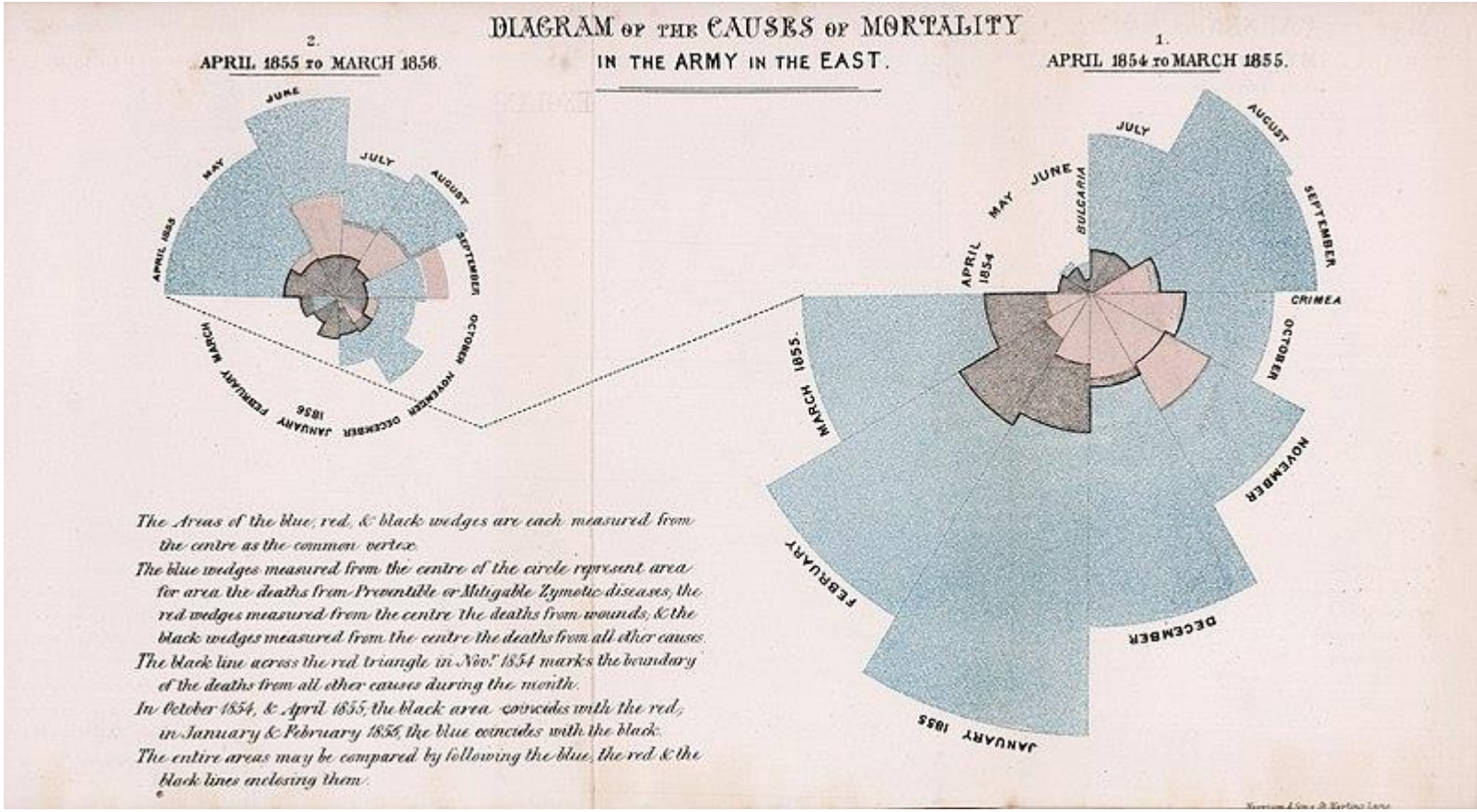
VARIATION: DONUT CHART



Sometimes useful to display outside of something else

https://www.d3-graph-gallery.com/donut.html

POLAR AREA CHARTS



Florence Nightingale

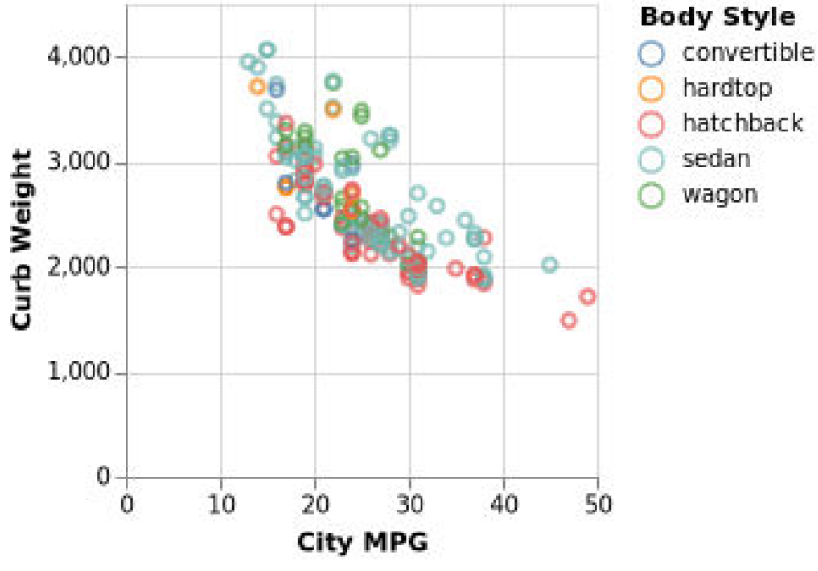
Also called
Nightingale rose

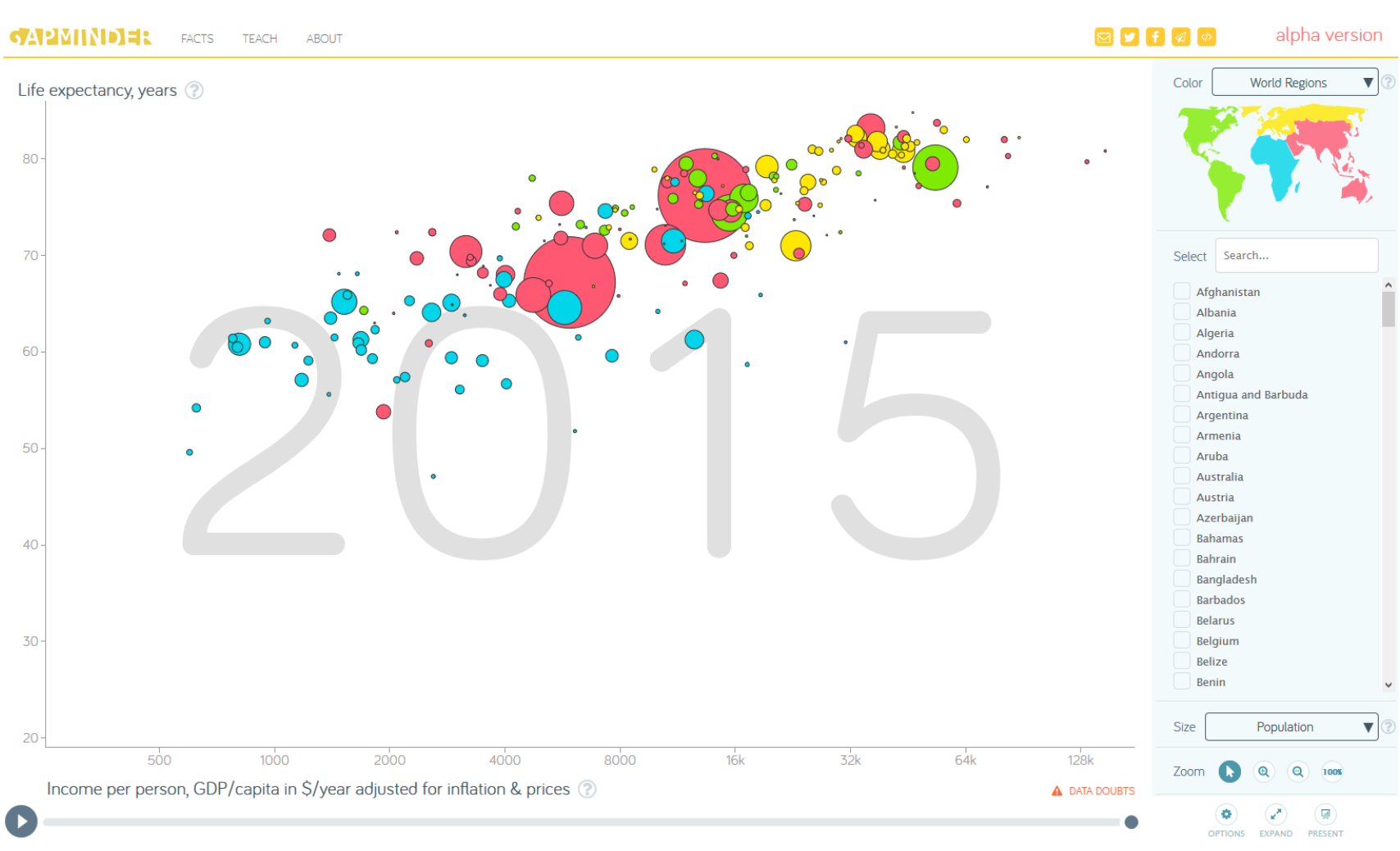
https://upload.wikimedia.org/wikipedia/commons/1/17/Nightingale-mortality.jpg

Do individual items fall into groups?
Is there a relationship between attributes of items?

SCATTERPLOTS

- two quantitative values
- horizontal and vertical spatial dimensions
- mark type = point



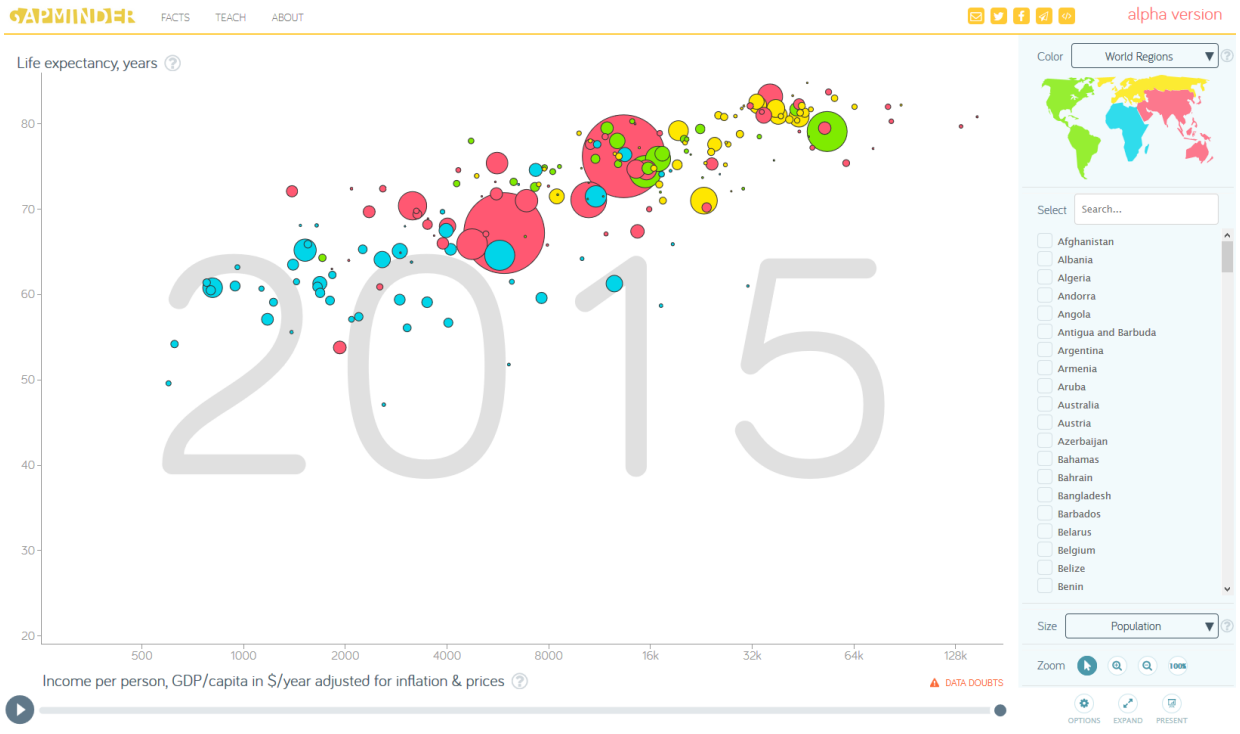


when marks are sized, the chart is often called a **bubble chart** or bubble plot

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen#t-6682

TASKS

- find trends
- find outliers
- show distribution
- show correlation
- locate clusters

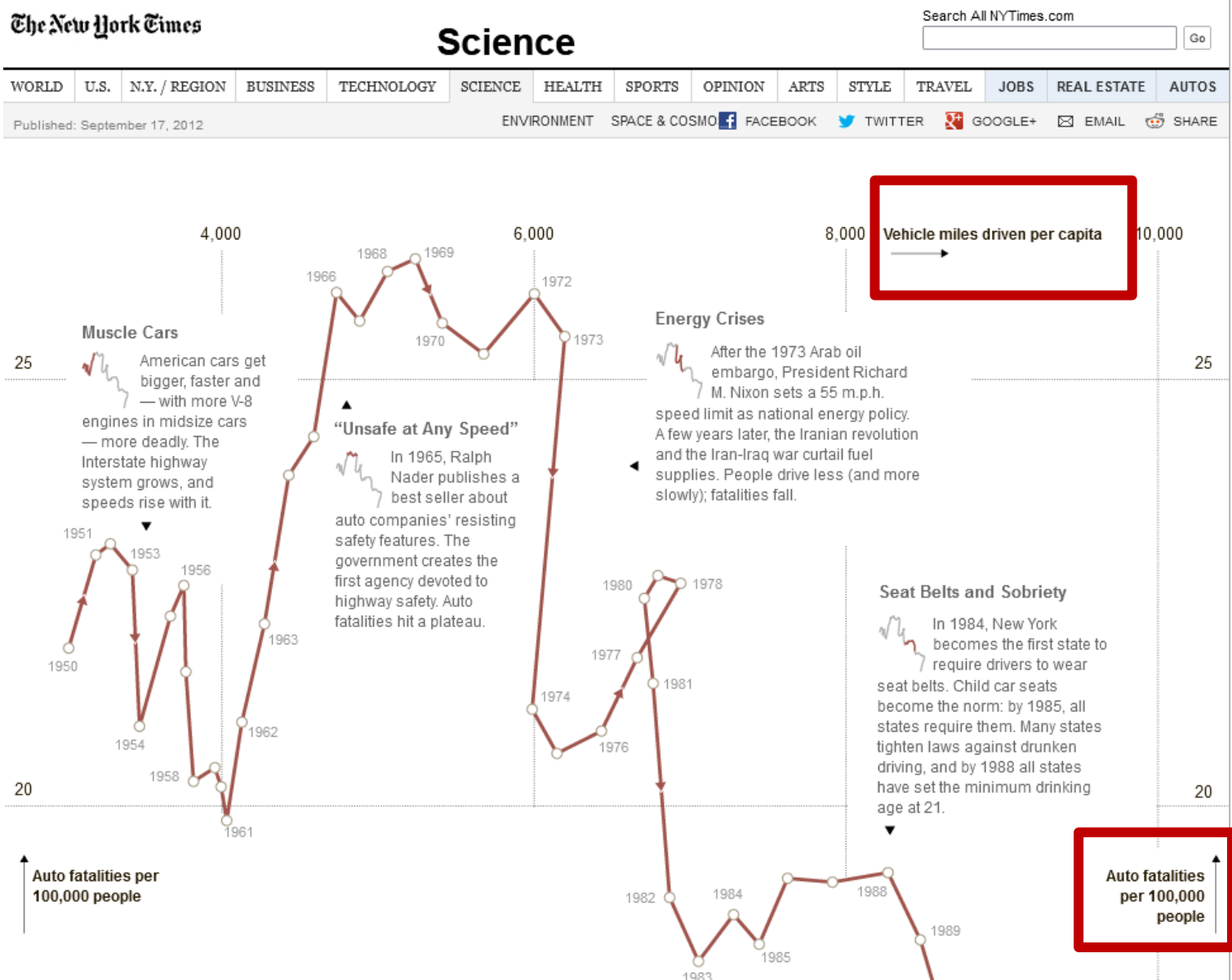


CONNECTED SCATTERPLOT

Good for storytelling

Confusing for analysis

http://steveharoz.com/research/connected_scatterplot/



SCATTERPLOT MATRIX

This idea scales relatively well

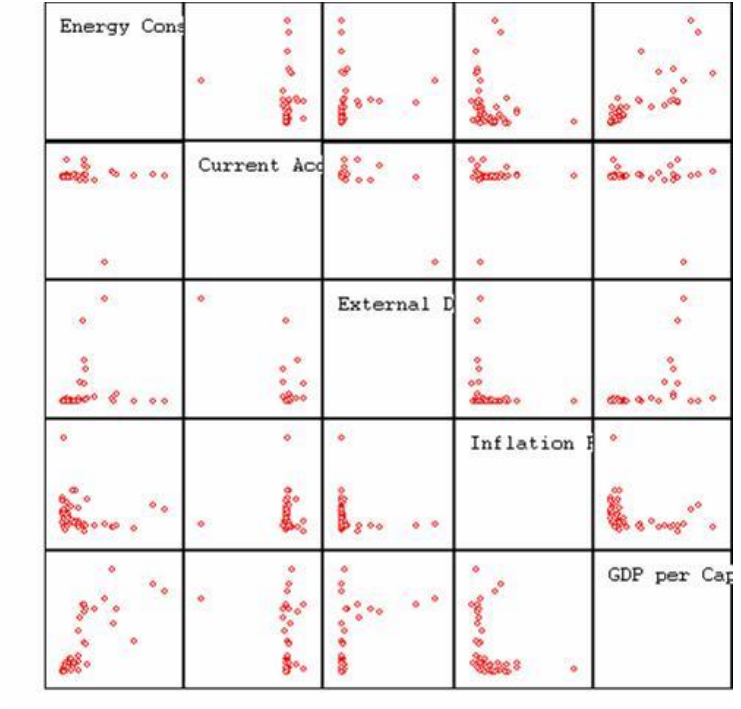
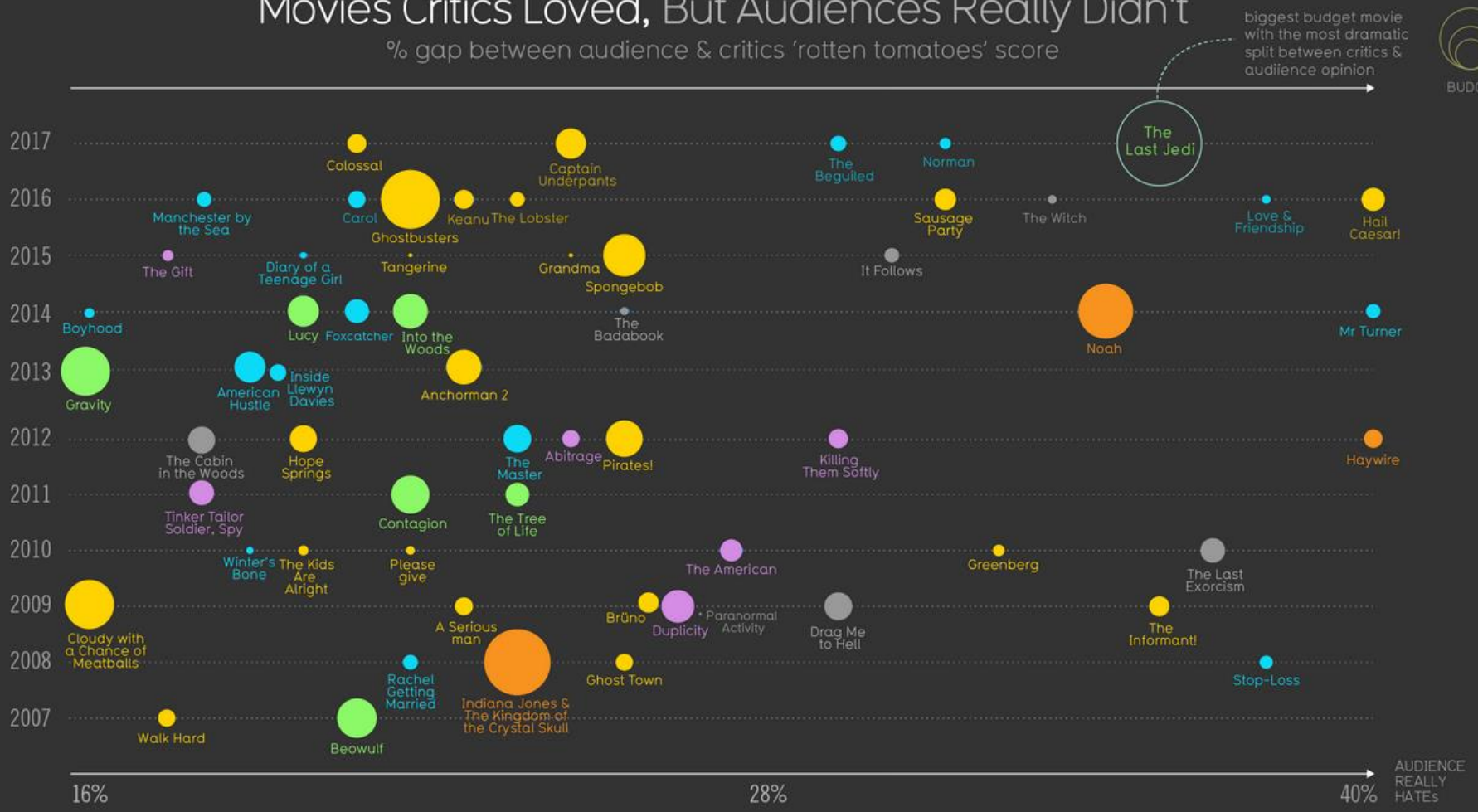


Image Source: Wikipedia

Movies Critics Loved, But Audiences Really Didn't

% gap between audience & critics 'rotten tomatoes' score



Comedy Sci-fi / Fantasy Action / Adventure Drama Thriller Horror

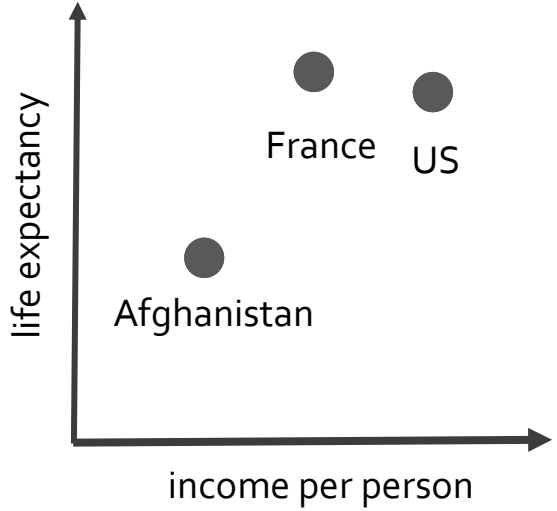
source: RottenTomatoes.com

data: bit.ly/IIB_MovieLove

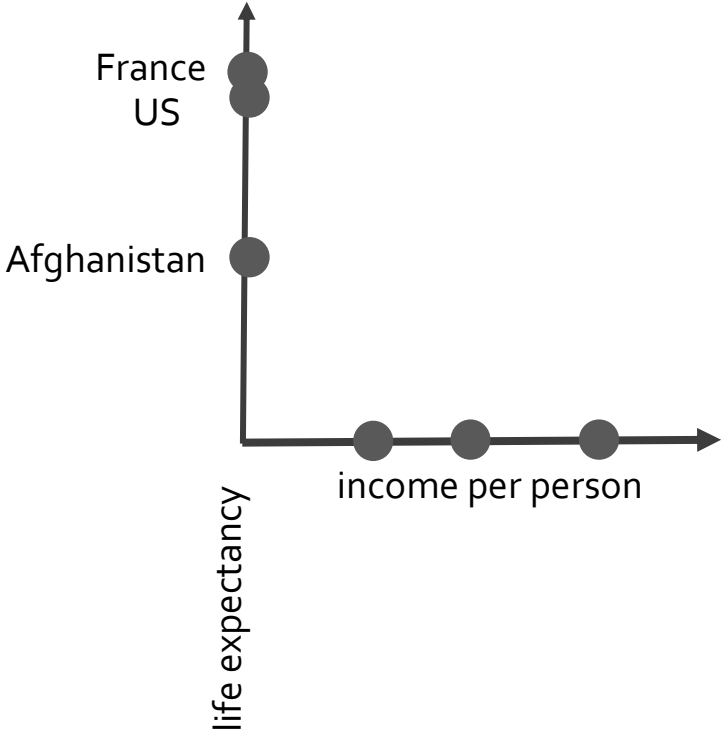
informationisbeautiful.net

PARALLEL COORDINATES

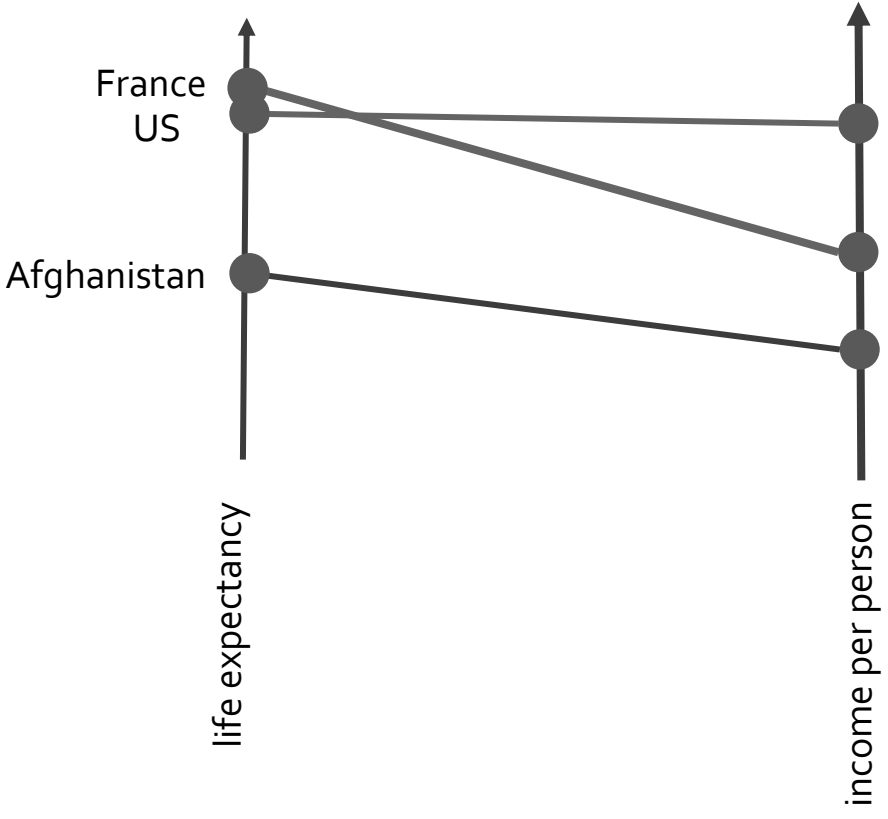
Back to our original example



PARALLEL COORDINATES



PARALLEL COORDINATES



show correlations between neighboring axes

MULTIDIMENSIONAL DETECTIVE

Alfred Inselberg*, Multidimensional Graphs Ltd†
&
Computer Science Department
Tel Aviv University, Israel
aiisreal@math.tau.ac.il

Abstract

The display of multivariate datasets in parallel coordinates, transforms the search for *relations* among the variables into a 2-D pattern recognition problem. This is the basis for the application to *Visual Data Mining*. The Knowledge Discovery process together with some general guidelines are illustrated on a dataset from the production of a VLSI chip. The special strength of parallel coordinates is in modeling **relations**. As an example, a simplified Economic Model is constructed with data from various economic sectors of a real country. The visual model shows the interelationship and dependencies between the sectors, circumstances where there is competition for the same resource, and feasible economic policies. Interactively, the model can be used to do trade-off analyses, discover sensitivities, do approximate optimization, monitor (as in a Process) and Decision Support.

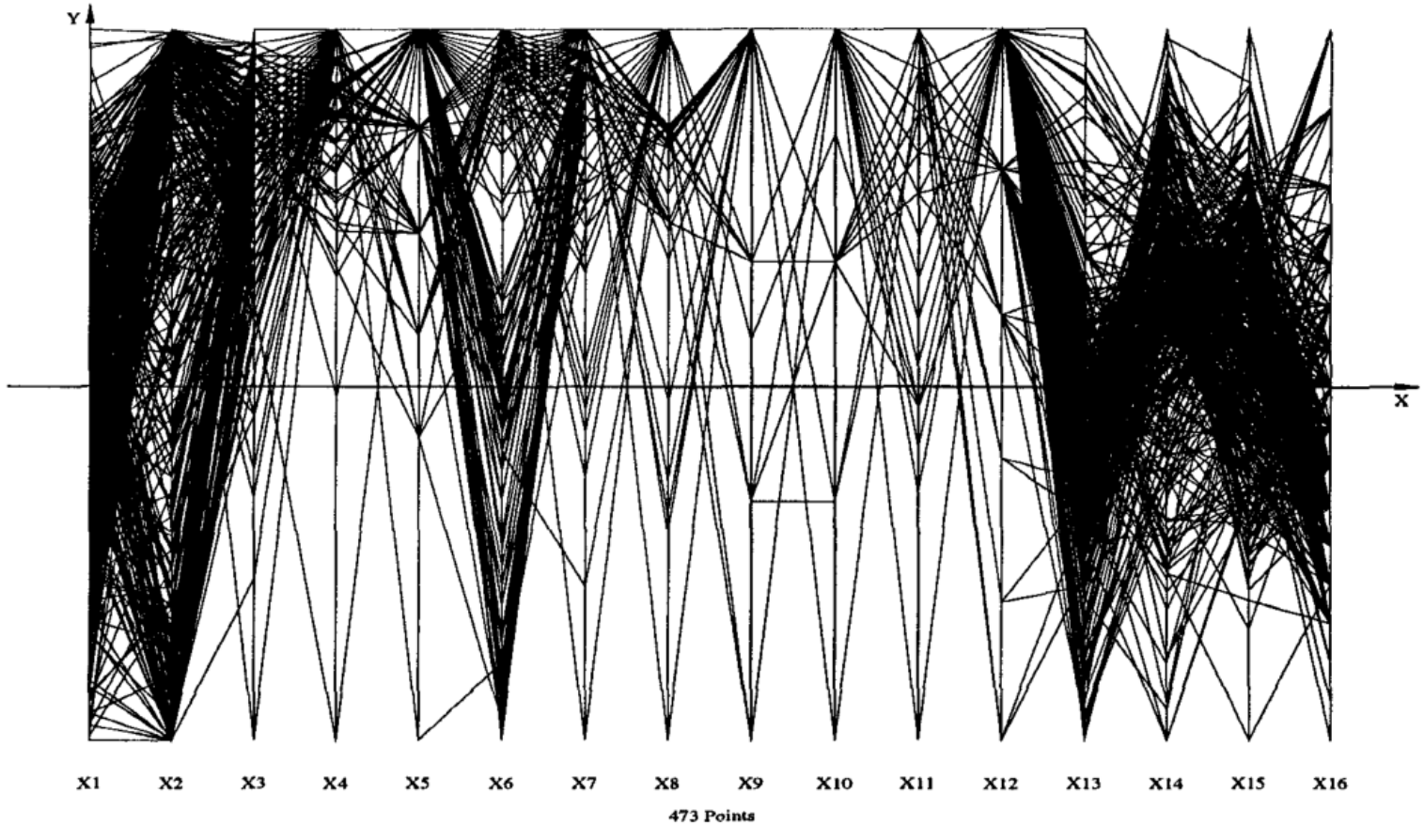
Introduction

In Geometry parallelism, which does not require a notion of angle, rather than orthogonality is the more fundamental concept. This, together with the fact that orthogonality "uses-up" the plane very fast, was the inspiration in 1959 for "Parallel" Coordinates. The systematic development began in 1977 [4]. The goals of the program were and still are (see [6] and [5] for short reviews) the visualization of multivariate/multidimensional problems without loss of information and having the properties:

1. Low representational complexity. Since the number of axes, N equals the number of dimensions (variables) the complexity is $O(N)$,
2. Works for any N,
3. Every variable is treated uniformly (unlike "Chernoff Faces" and various types of "glyphs"),
4. The displayed object can be recognized under projective transformations (i.e. rotation, translation, scaling, perspective),
5. The display easily/intuitively conveys information on the properties of the N-dimensional object it represents,
6. The methodology is based on rigorous mathematical and algorithmic results.

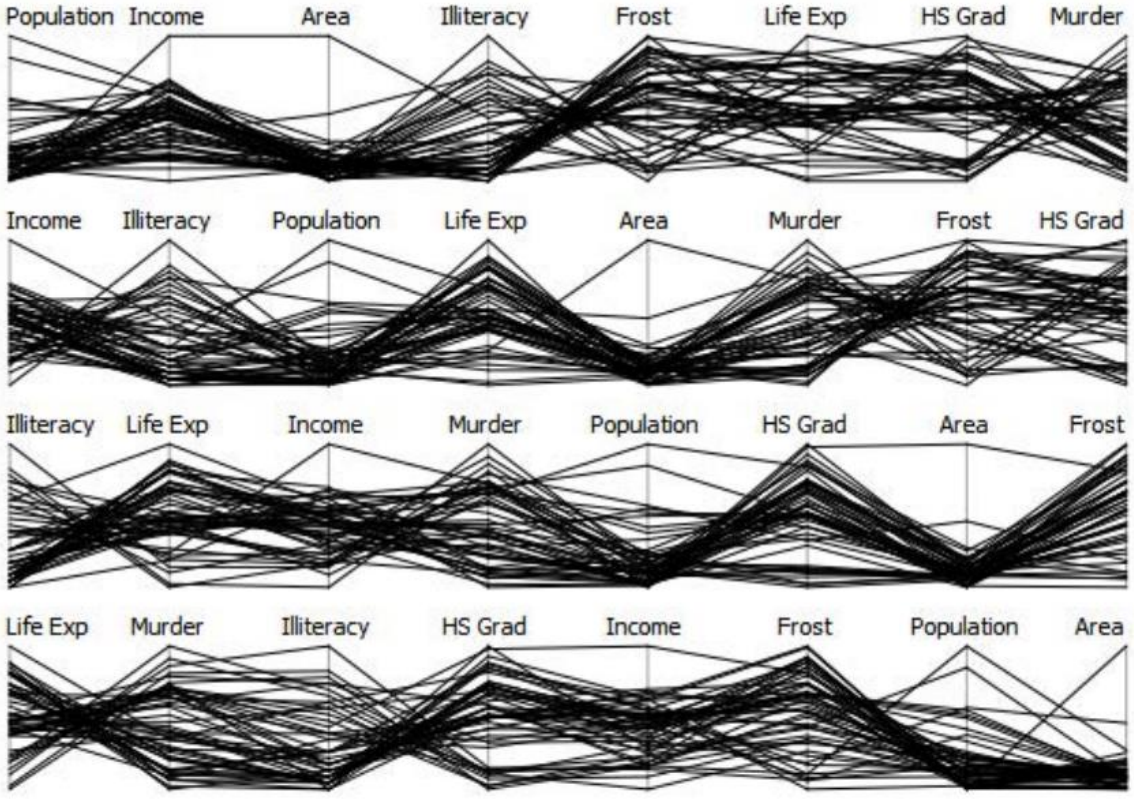
Parallel coordinates (abbr.||-coords) transform multivariate relations into 2-D patterns, a property that is well suited for Visual Data Mining.

*Senior Fellow San Diego SuperComputing Center
†36A Yehuda Halevy Street, Raanana 43556, Israel



Original Example from Inselberg 1997

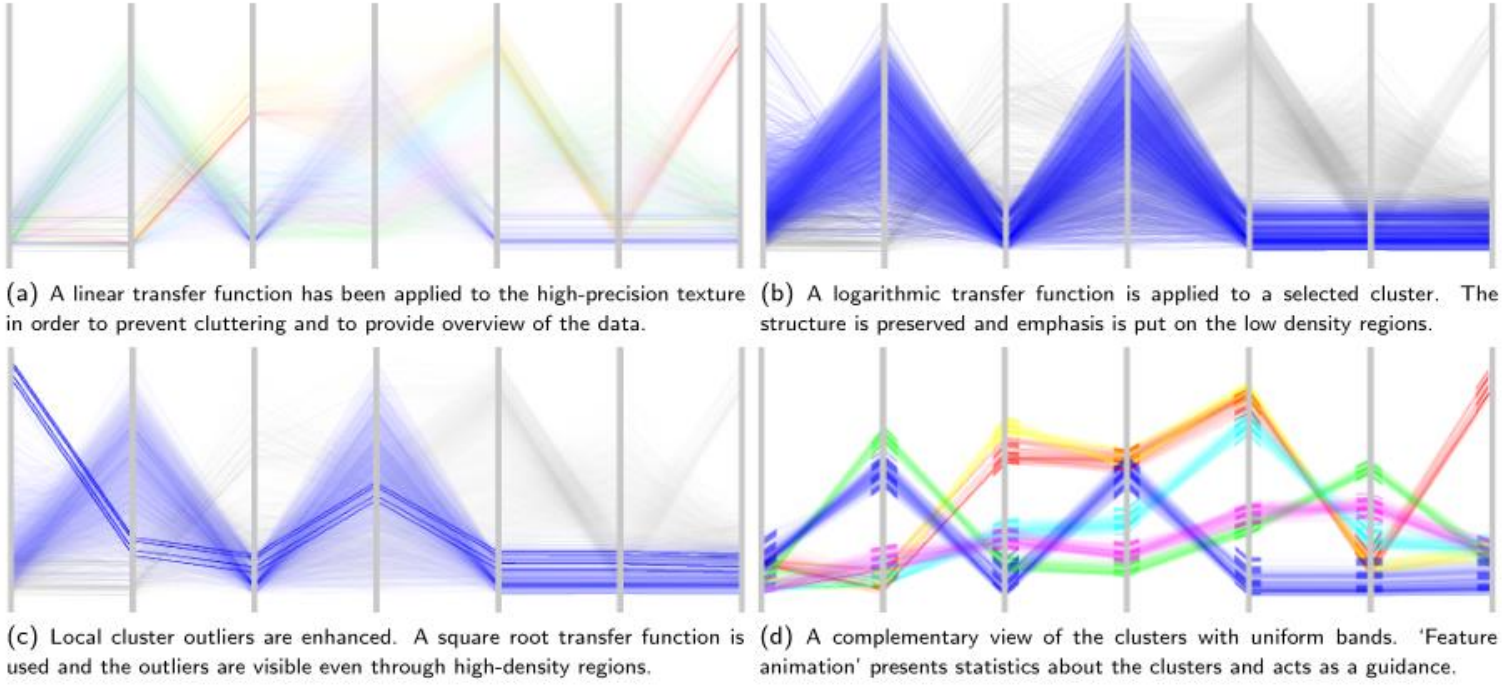
THE ORDER OF AXES MATTERS



Eurographics 2013, STAR Report
J. Heinrich, D. Weiskopf

REDUCE CLUTTER - HIGHLIGHT CLUSTERS

Lots of work on this. For example:



(a) A linear transfer function has been applied to the high-precision texture in order to prevent cluttering and to provide overview of the data.

(b) A logarithmic transfer function is applied to a selected cluster. The structure is preserved and emphasis is put on the low density regions.

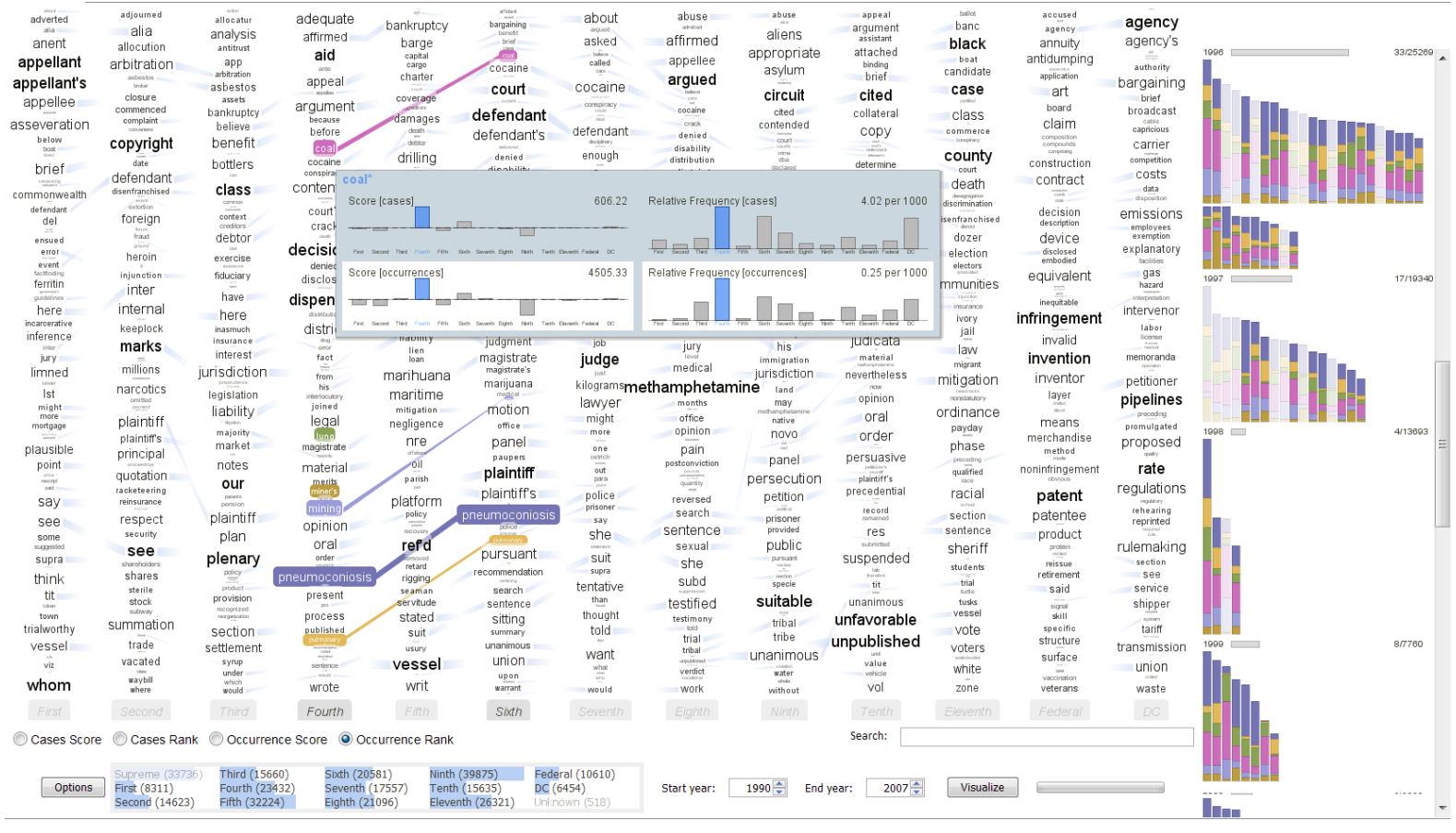
(c) Local cluster outliers are enhanced. A square root transfer function is used and the outliers are visible even through high-density regions.

(d) A complementary view of the clusters with uniform bands. 'Feature animation' presents statistics about the clusters and acts as a guidance.

Revealing Structure within Clustered Parallel Coordinates Displays, InfoVis 2005

COMBINE WITH OTHER VISUALIZATION TECHNIQUES

Parallel Tag Clouds to Explore Faceted Text Corpora (Collins et al., VAST 2009)



THERE IS MUCH MORE ON THIS…

Start here if you want more information

State of the Art of Parallel Coordinates

J. Heinrich and D. Weiskopf

Visualization Research Center, University of Stuttgart

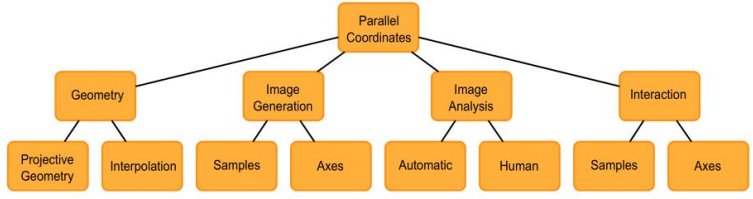
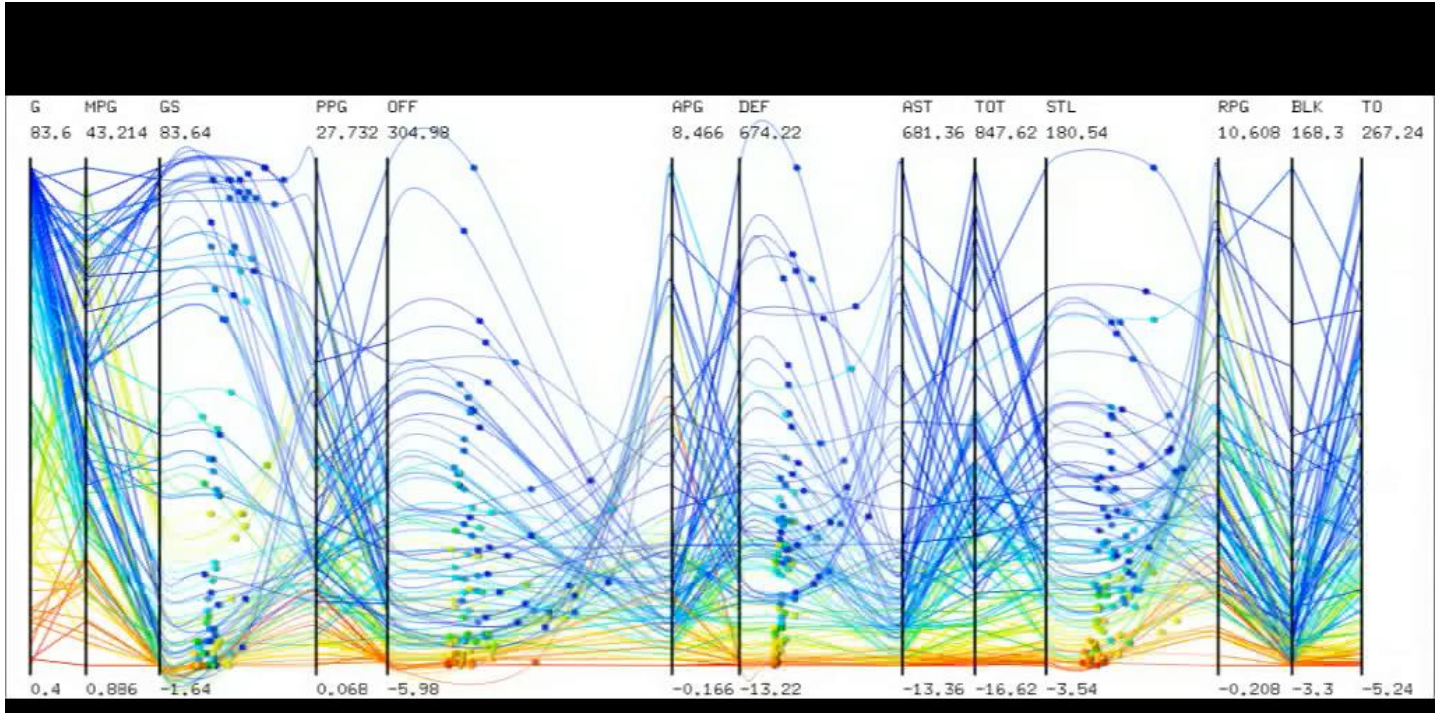


Figure 1: Taxonomy of topics for parallel coordinates in the scientific literature. The first-level nodes each represent a section in this paper, where the scope and definition of each topic will be explained.

Abstract

This work presents a survey of the current state of the art of visualization techniques for parallel coordinates. It covers geometric models for constructing parallel coordinates and reviews methods for creating and understanding visual representations of parallel coordinates. The classification of these methods is based on a taxonomy that was established from the literature and is aimed at guiding researchers to find existing techniques and identifying white spots that require further research. The techniques covered in this survey are further related to an established taxonomy of knowledge-discovery tasks to support users of parallel coordinates in choosing a technique for their problem at hand. Finally, we discuss the challenges in constructing and understanding parallel-coordinates plots and provide some examples from different application domains.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation



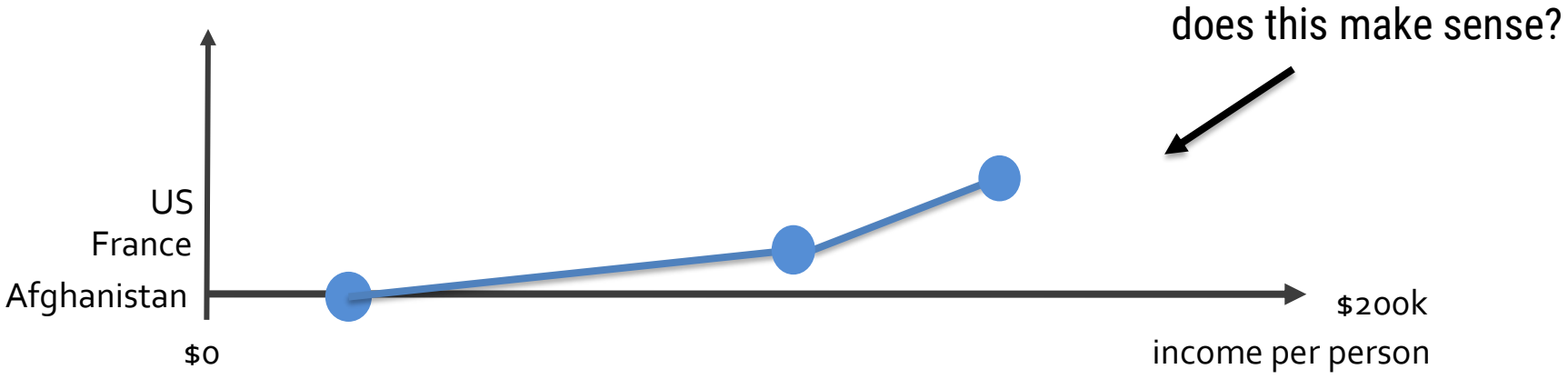
Scattering Points in Parallel Coordinates (SPPC)

How does an attribute vary continuously?

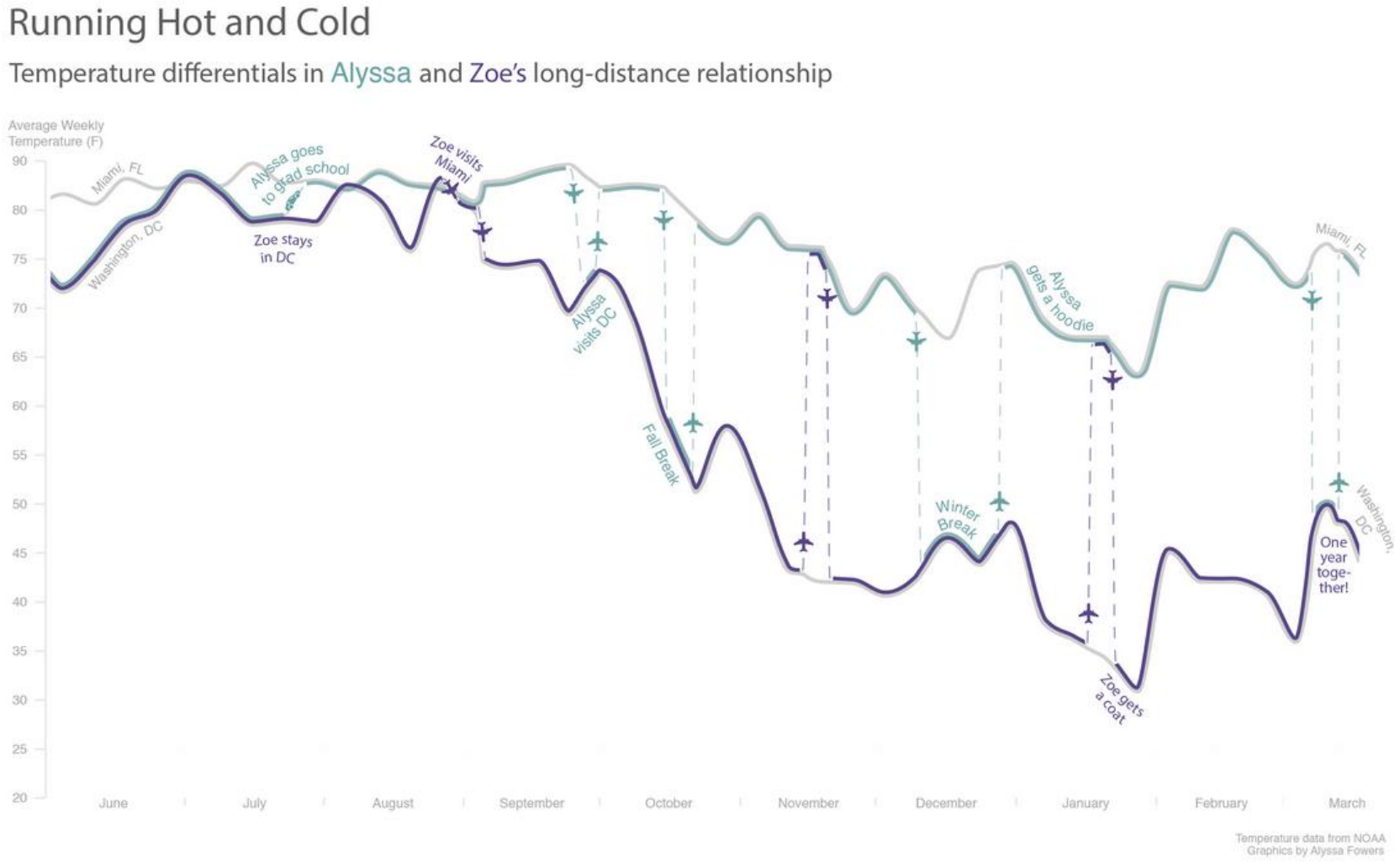
SHOW TRENDS - LINE CHARTS

Connect values with lines

Emphasize and order

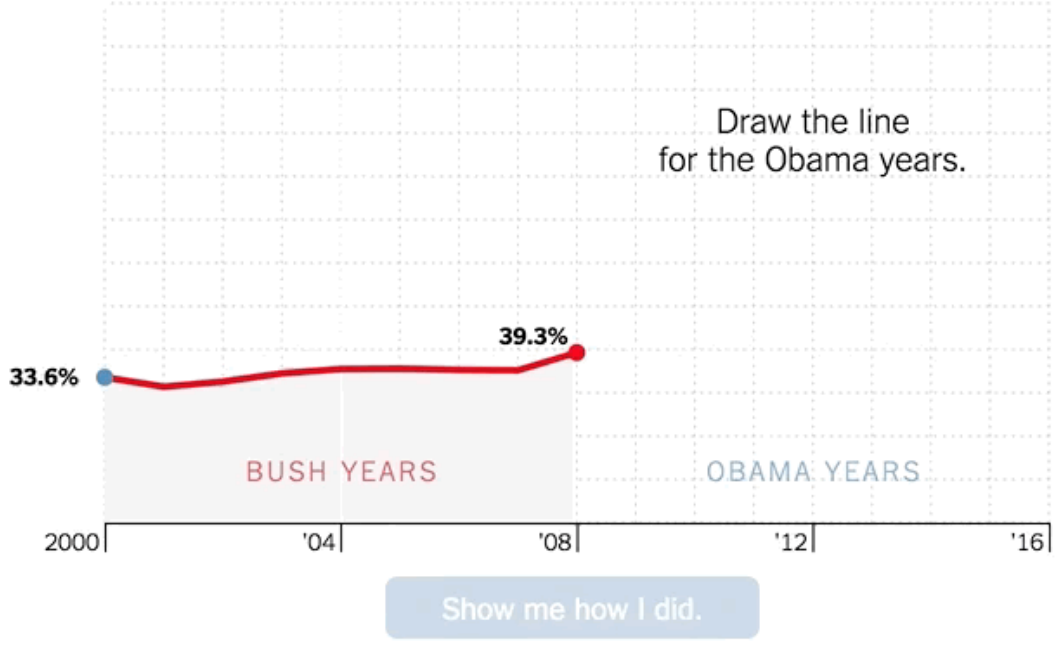


Often used with time

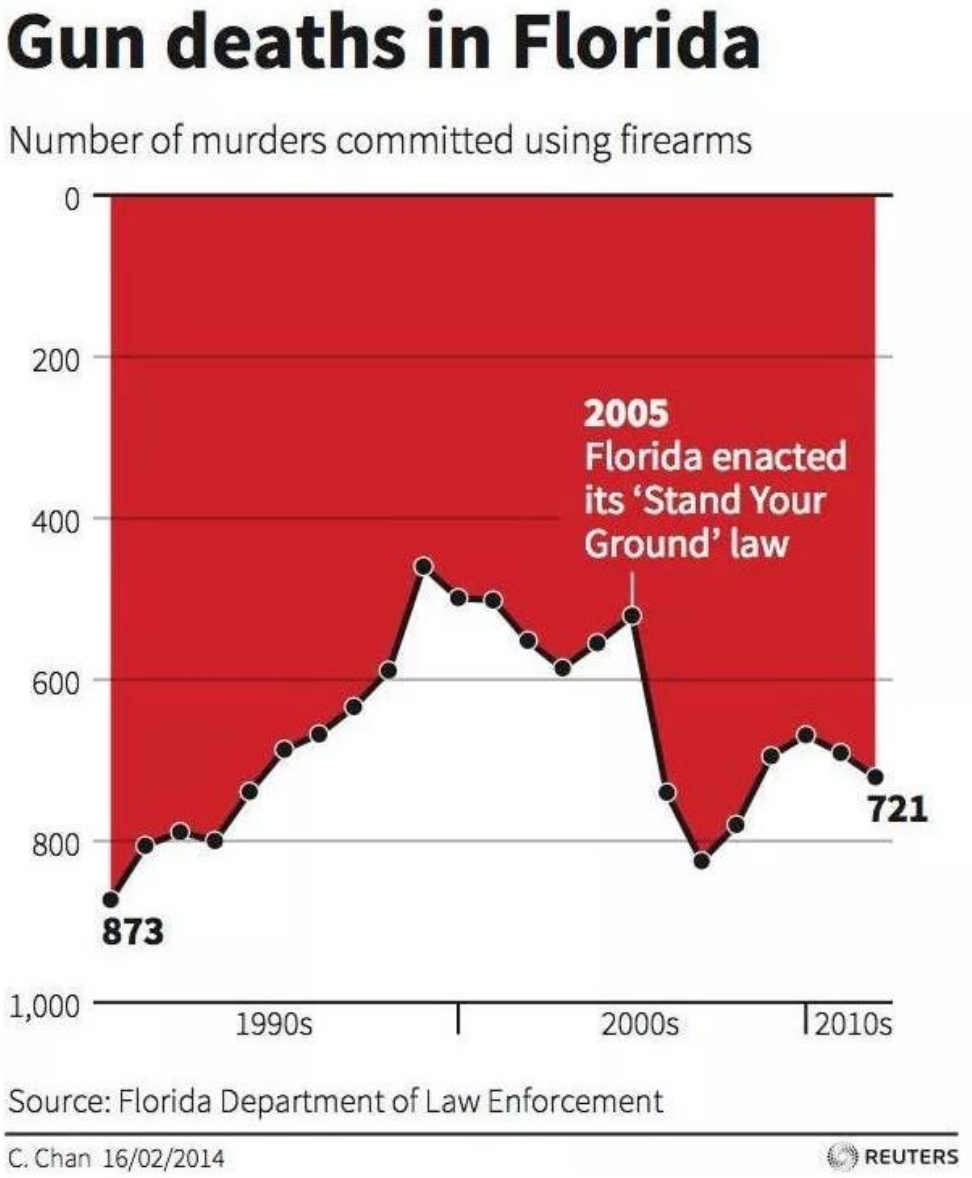


https://www.alyssafowers.com/portfolio/#/running-hot-and-cold/

Under Mr. Obama, the **national debt** as a percentage of the gross domestic product ...



Numbers



Take a moment to look at this chart

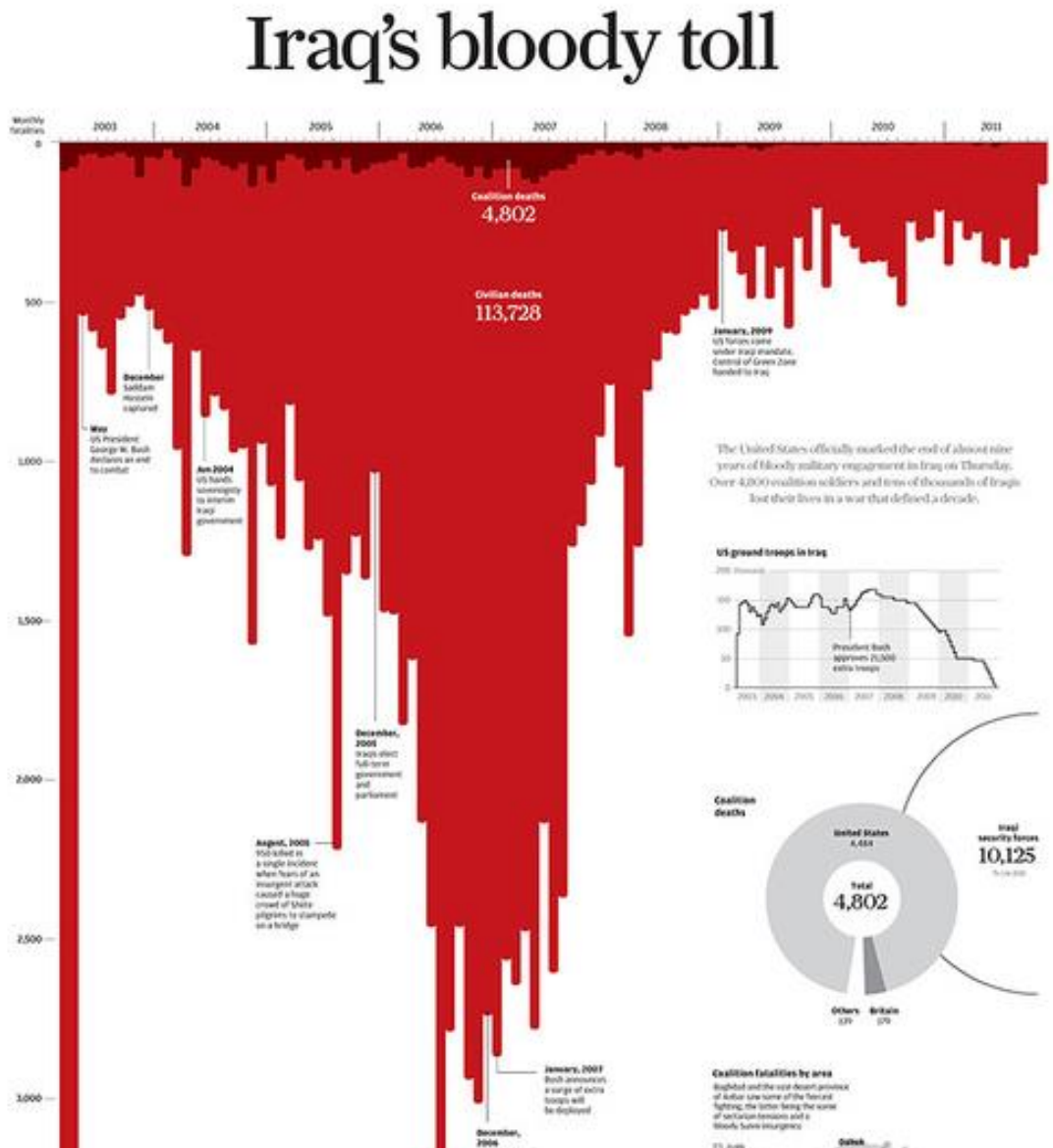
What does it tell us?

Does the encoding work?

Additional material:

https://policyviz.com/2020/06/16/the-inverted-vertical-axes/

https://doi.org/10.1109/TVCG.2021.3088343

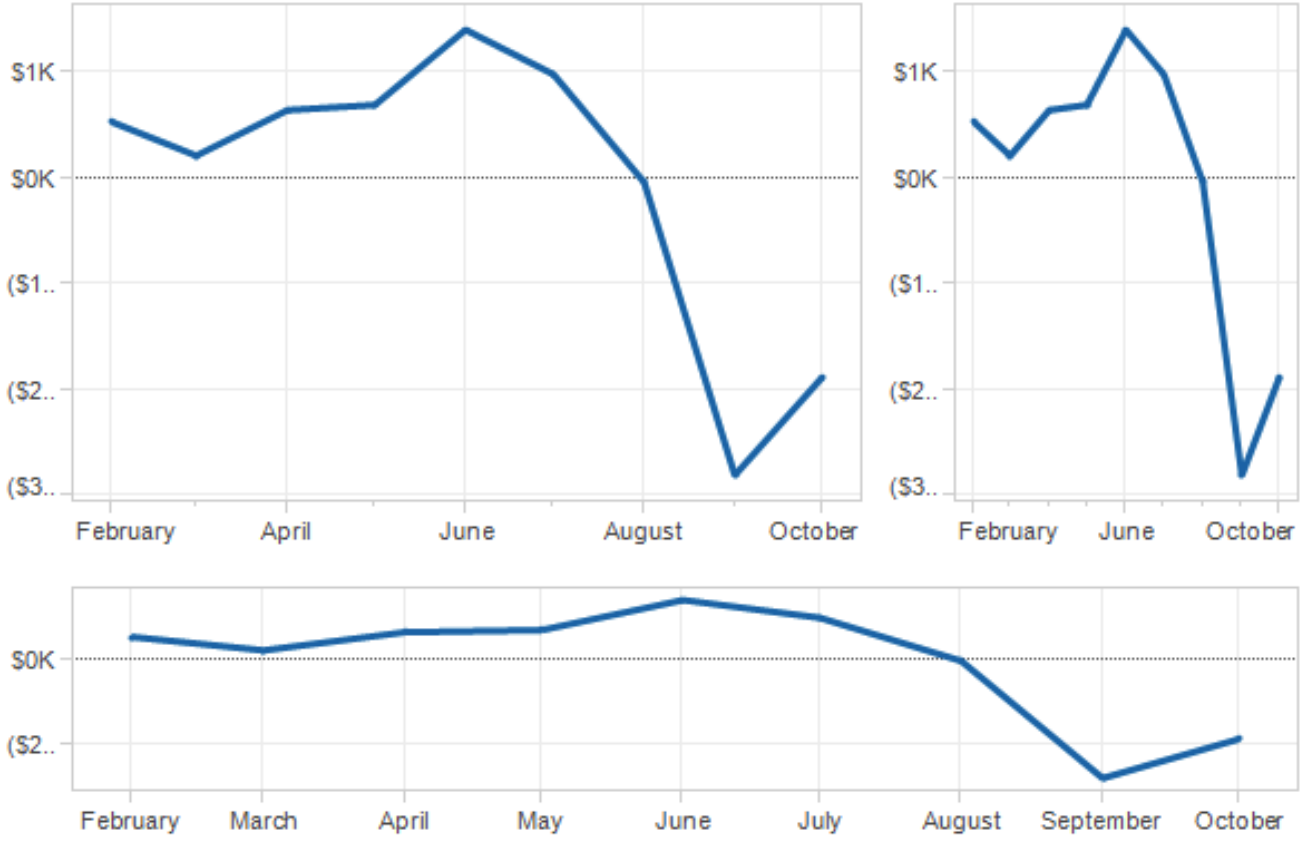


This graphic was created to mark the end of the United States' military engagement in Iraq in 2011. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in the war.
One deliberate design choice with this graphic was the visual metaphor of blood. This striking visual would hopefully draw the reader into the graphic.

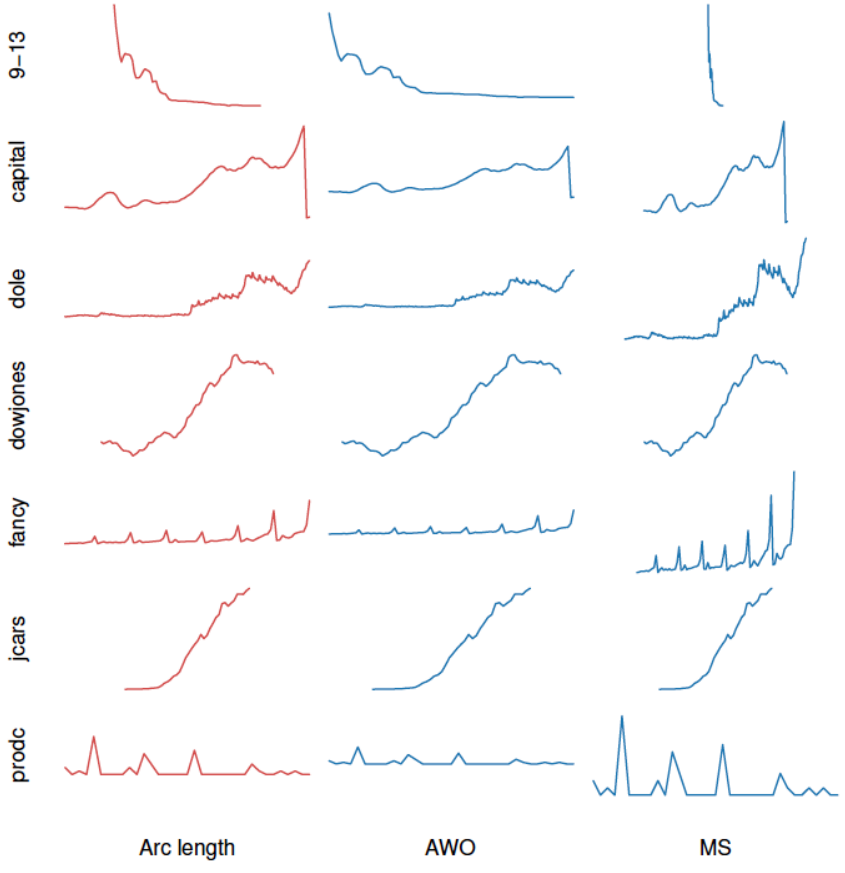
By Simon Scarr
South China Morning Post

http://www.simonscarr.com/iraqs-bloody-toll

ASPECT RATIO SELECTION



MANY METHODS

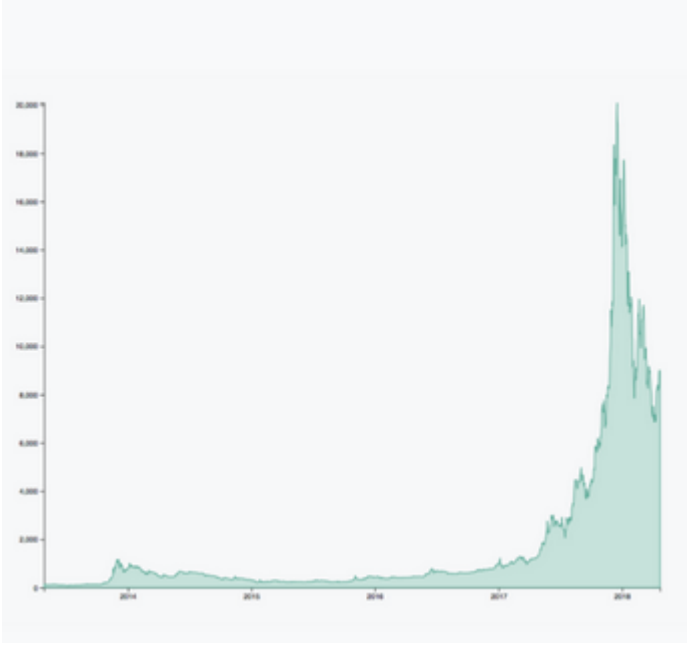


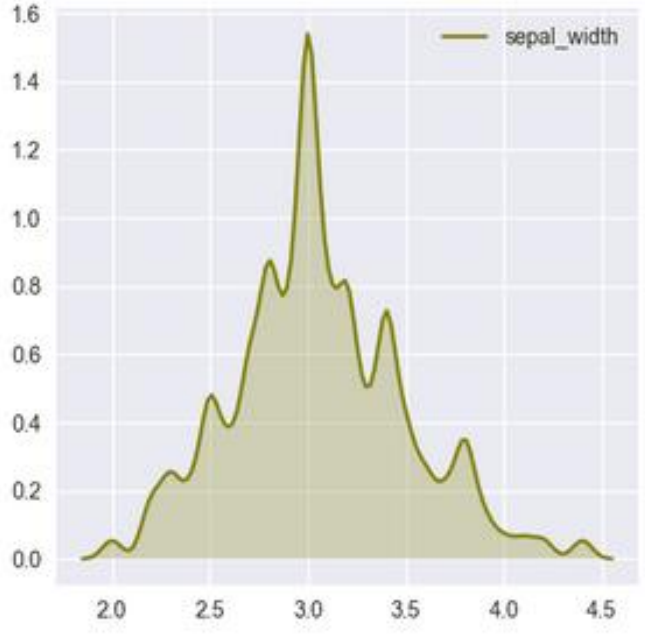
Practical advice:

CHOOSE AN ASPECT RATIO THAT EMPHASIZES THE IMPORTANT DETAILS FOR YOUR TASK

[TALBOT ET AL, 2011]

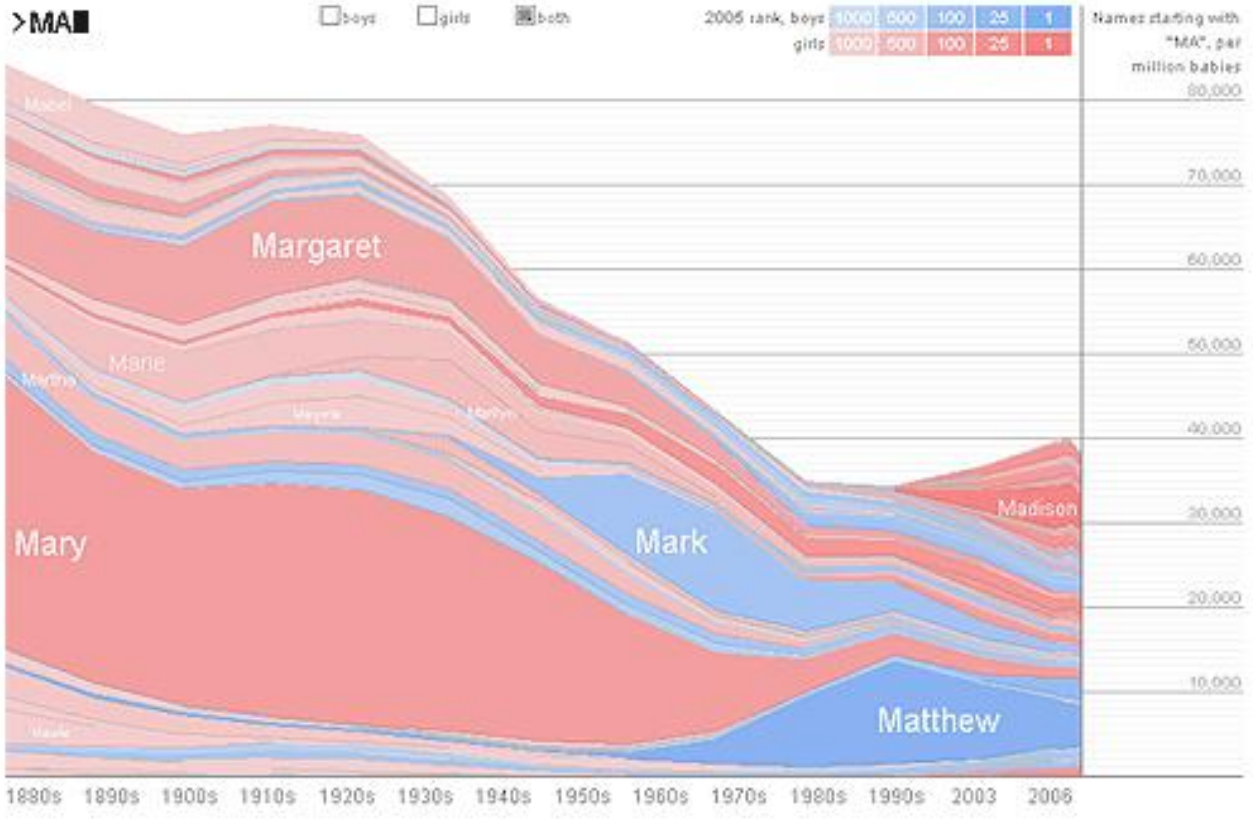
VARIATIONS: AREA CHART





Cut or don’t cut the y-axis?

VARIATIONS: STACKED AREA CHART

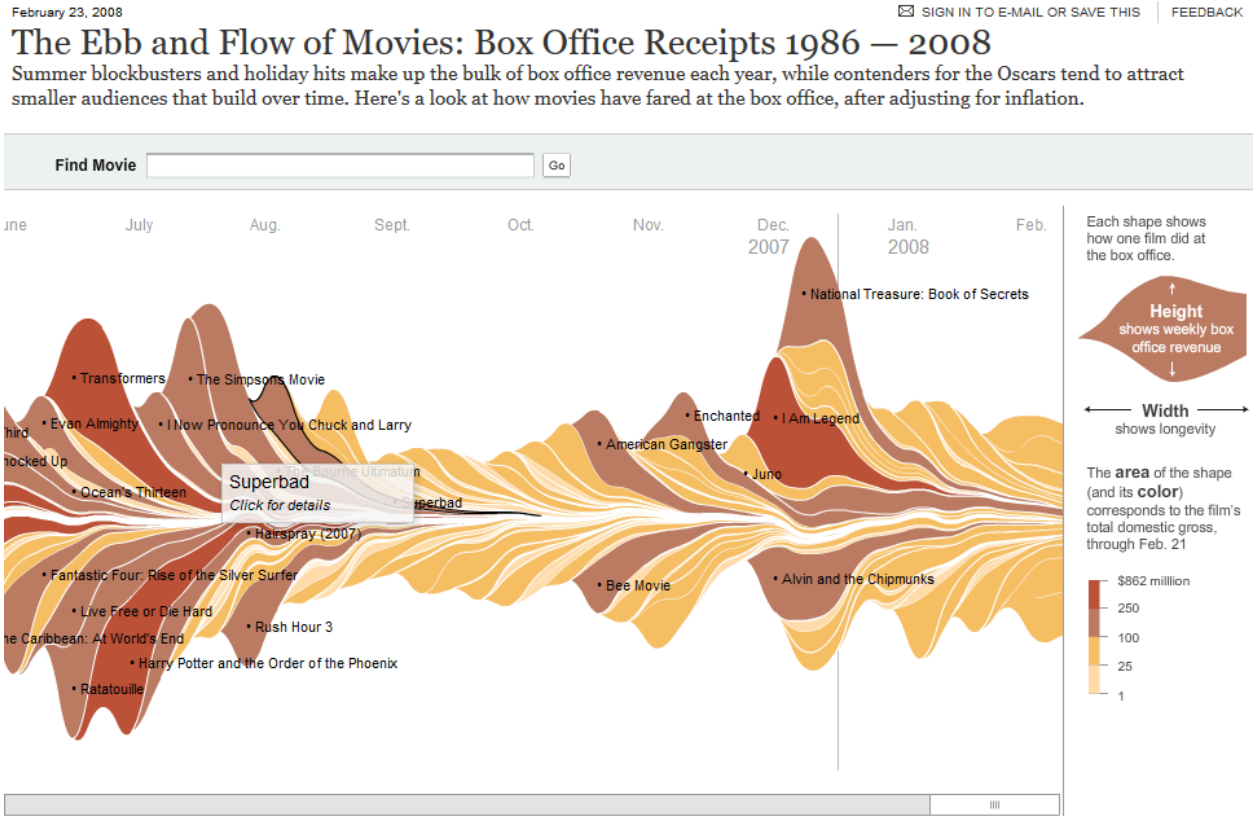


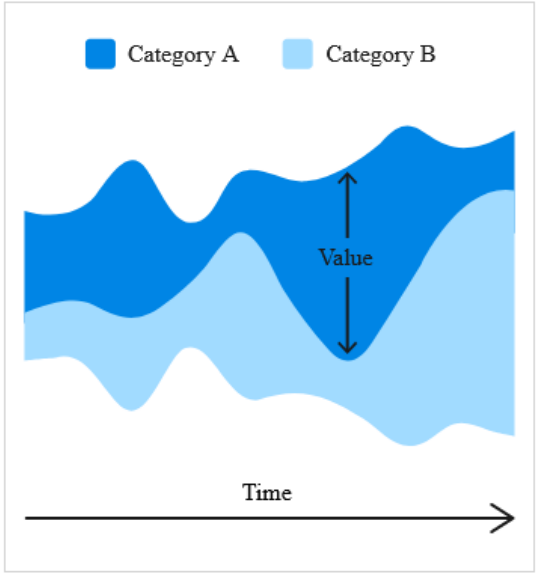
http://www.bewitched.com/namevoyager.html

Drawbacks:
Evolution of individual categories hard to see

Advantage:
Study of the whole
Relative proportions per category

STREAMGRAPH

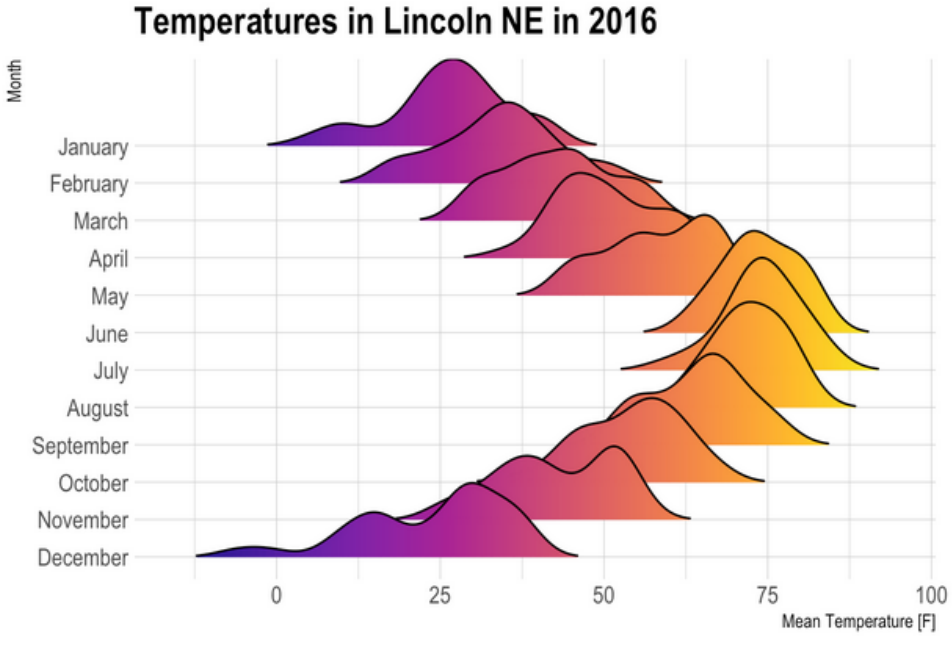




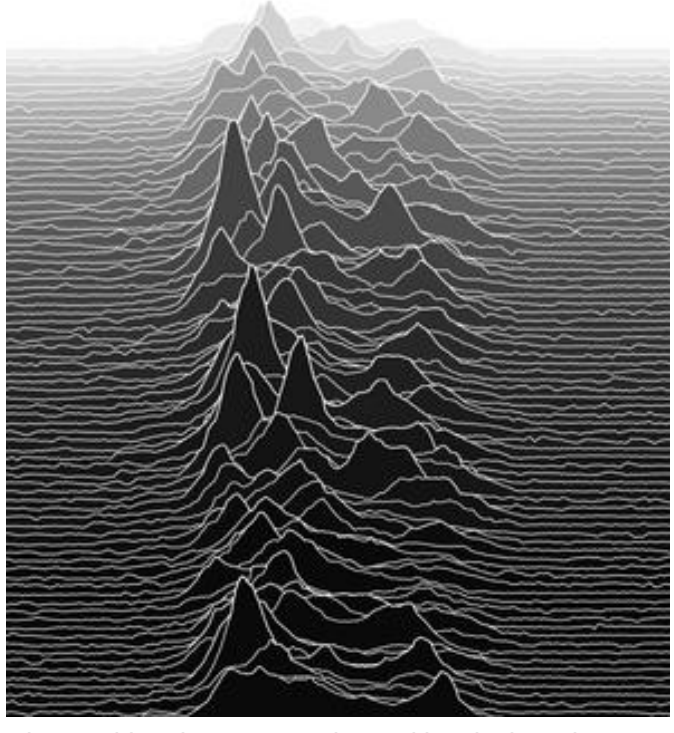
https://datavizcatalogue.com/methods/stream_graph.html

RIDGELINE PLOTS

Small Multiple Area Charts



https://www.data-to-viz.com/graph/ridgeline.html



https://blog.datawrapper.de/weekly-ridgeline-plot/

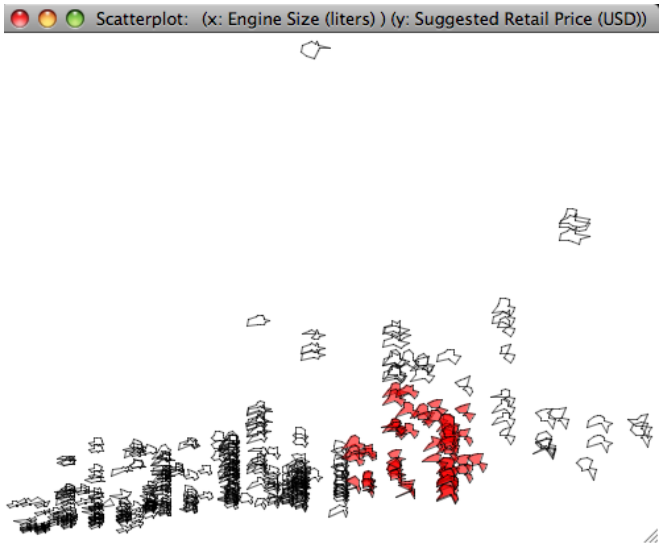
How are objects related to each other?

(Networks & Hierarchies get their own lecture)

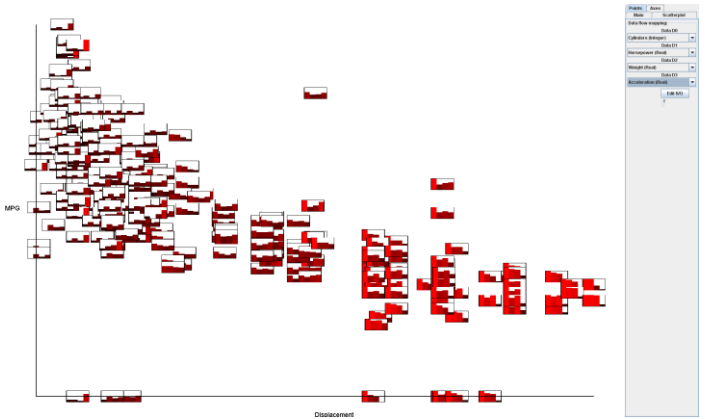
GLYPHS

marks can be replaced with glyphs

glyphs are themselves composed of multiple marks



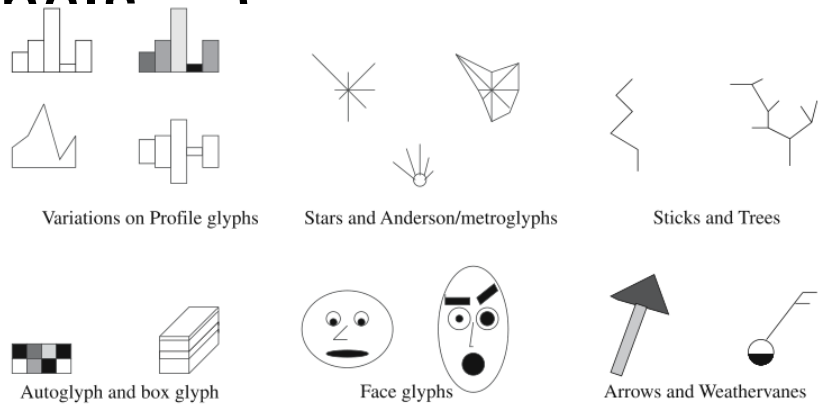
http://rosuda.org/software/Gauguin/gauguin.html



https://engineering.purdue.edu/~elm/projects/gpuvis.html

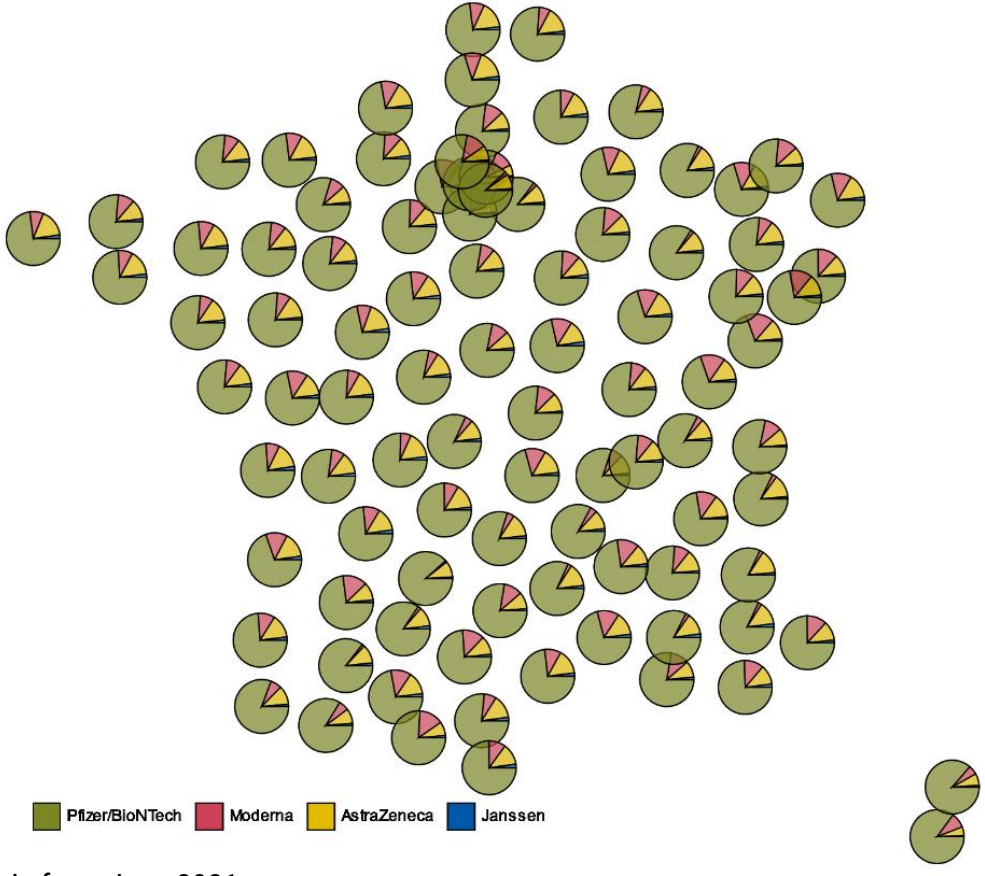
GLYPHS

- Small composite visual representations of multi-dimensional data points
- Characterized generally by lack of reference structures (grid lines, axes labels)



From Ward, 2002
A taxonomy of glyph placement strategies for multidimensional data visualization

MEANINGFUL LAYOUT



this data is from June 2021

EXAMPLE: CHERNOFF FACES

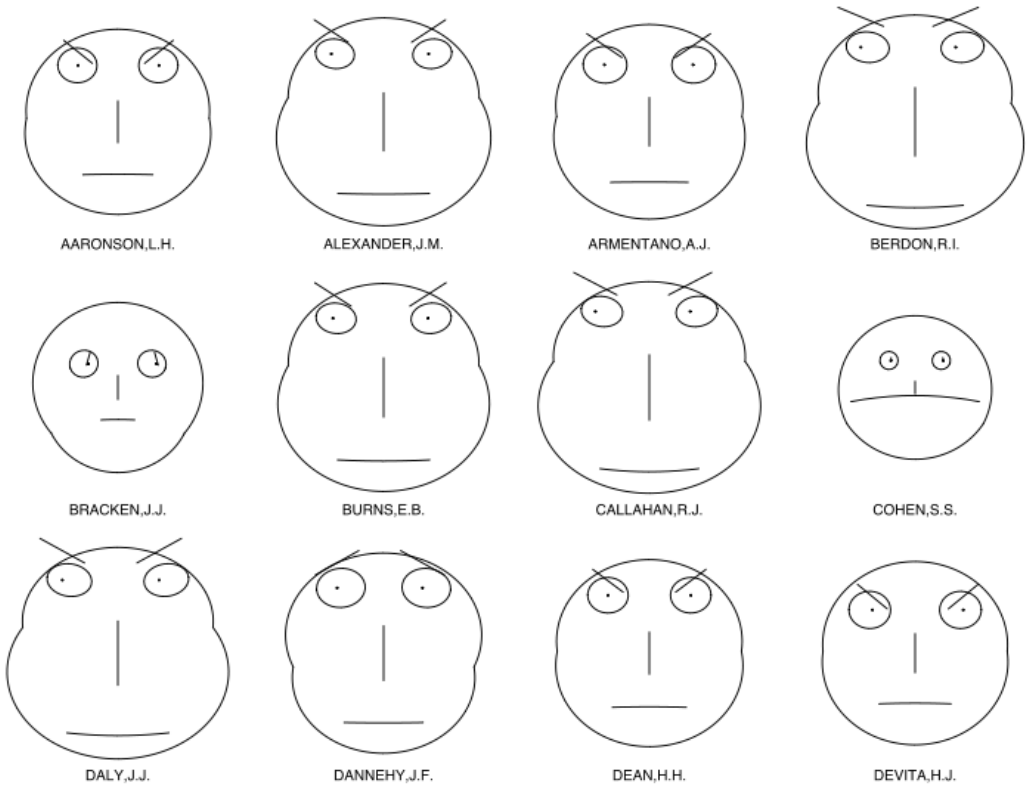
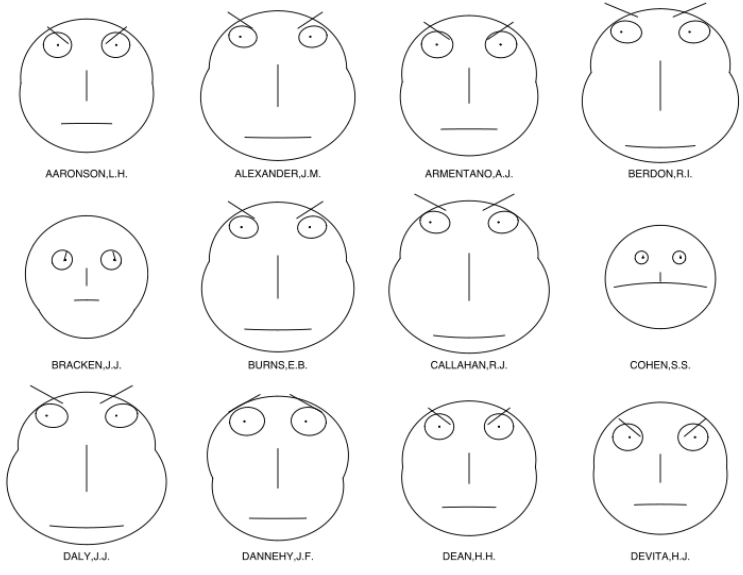


Image source: Wikipedia

Herman Chernoff, The Use of Faces to Represent Points in K-Dimensional Space Graphically, 1973.

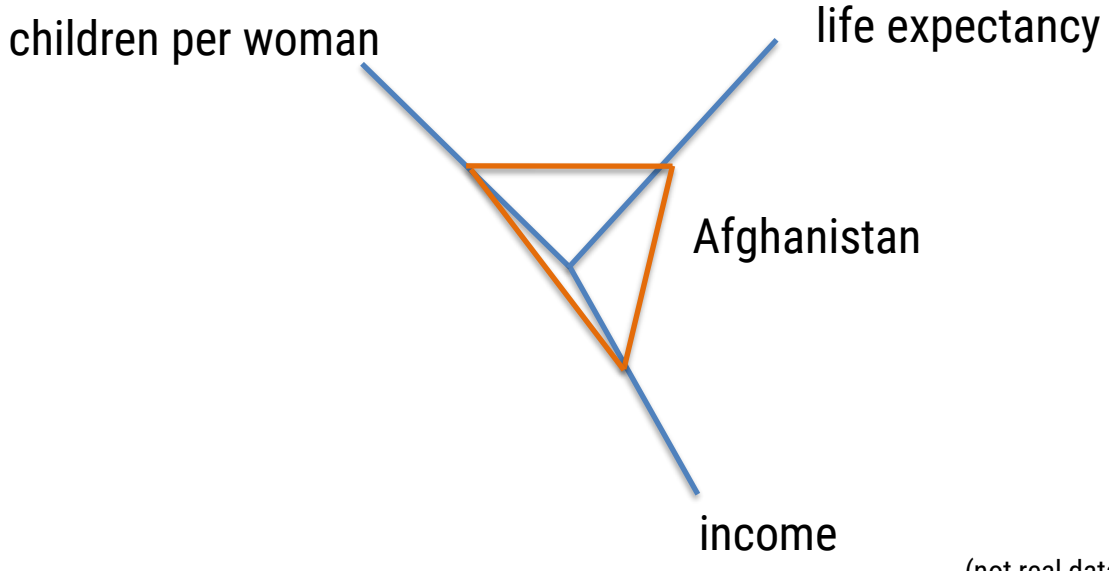
CHERNOFF FACES

- features of a human face encode data values (e.g. slant of eye brows, size of eyes, …)
- reasoning: humans are good at differentiating faces and reading face features
- problem: chernoff faces have generally been found not to be very effective



EXAMPLE: STAR GLYPHS

- Lay out dimension in radial fashion
- Draw each point as a ring



(not real data, here)

Glyphs (small or large) can be good replacements for maps!

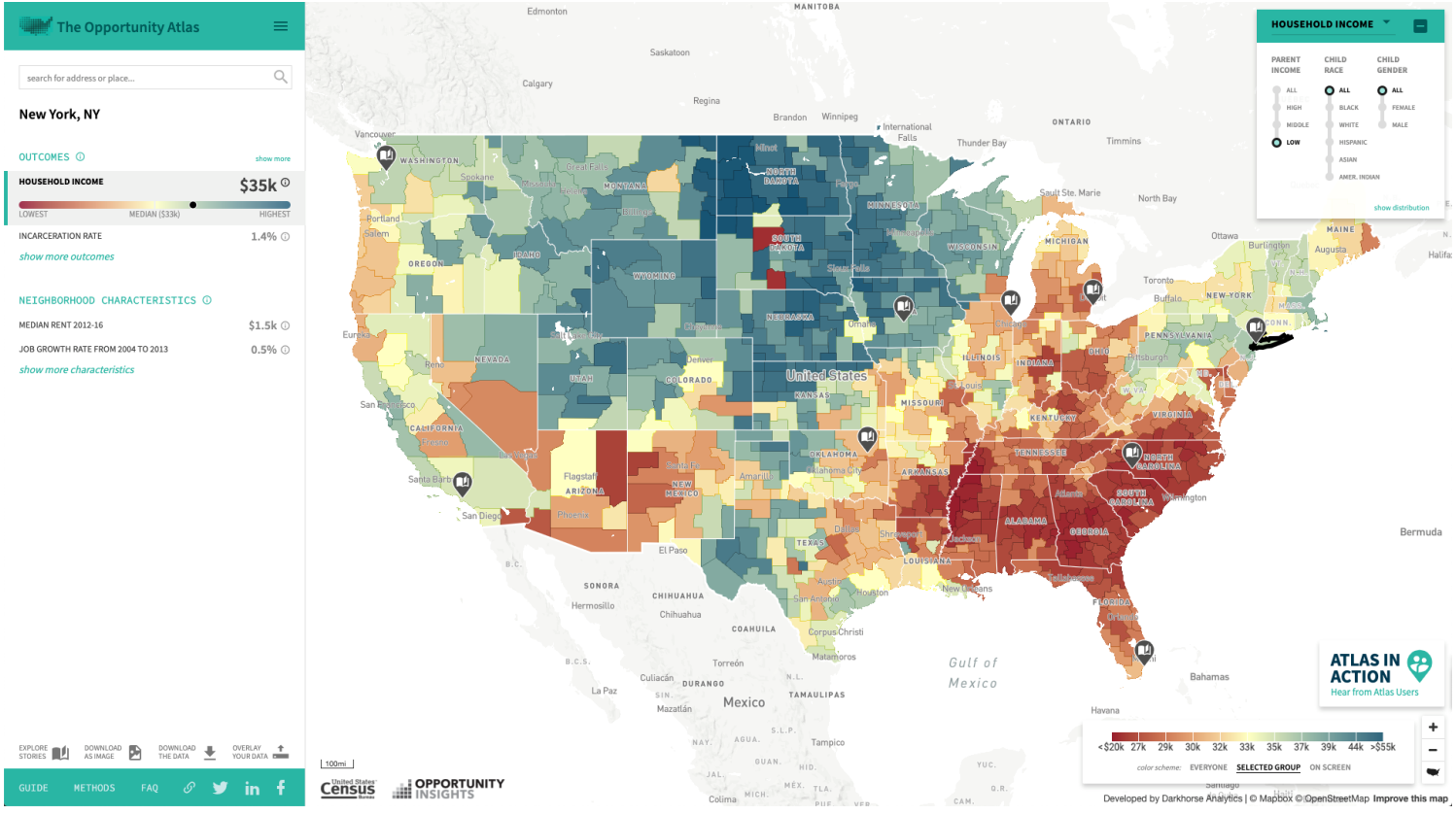
SMALL MULTIPLES

https://pudding.cool/2017/03/incarceration/index.html

Where are objects located?

(We don’t have time for a separate maps lecture ☹)

CLOROPLETH MAP



Uses heatmap idea

What is in this text?

(Sometimes we have time for a text lecture…we will see)

WORD CLOUD


VISUALIZATION
DATA
VISUAL
QUESTION
PROCESS
BOOK
MAKE
HELP
EXPLORATION
REPRESENTATION
OFTEN
TOP
PEOPLE
SENSE
TASKS
SET
ADDRESSED
RIGHT
USING
MANY
USE
EFFECTIVE
TYPES
THINK
TOOLS
STEPS
ALREADY
BOOKS
INSIGHT
WORKED
DEFINED
ATTRIBUTES
TALK
ANALYSIS
CHOOSE
SCIENCE
INSIGHTS
WELL
BROAD
PROBLEMS
WAY
TEAM
PROVIDE
BUSINESS
REQUIRES
LIKE
ANALYSTS
EXAMPLE
BESPOKE
TASK
COMPLEX
ASKING
TOPICS
CHAPTERS
DASHBOARD
ITERATIVE
DISCUSSES
FIGURING
TWO
RATHER
CODE
SPACE
MULTIPLE
INTELLIGENCE
VISUALIZED
ALLOW
DESCRIBES
PRESENTING
UNDERSTANDING
OVERVIEW
THOUGH
ANSWER
GETTING
SOPHISTICATED
CREATING
ILLUSTRATES
DIRECTORS
QUICK
TECHNIQUES
HUMAN
SPECIFIC
OPERATIONALIZATION
VIEWS
TYPE
MOVIE
SYSTEMS
FUNDAMENTALLY
FOUND
IMPORTANT
INSTEAD
BEST
LAY
GIVEN
GAIN
TEACH
SHELF
WHAT
GET

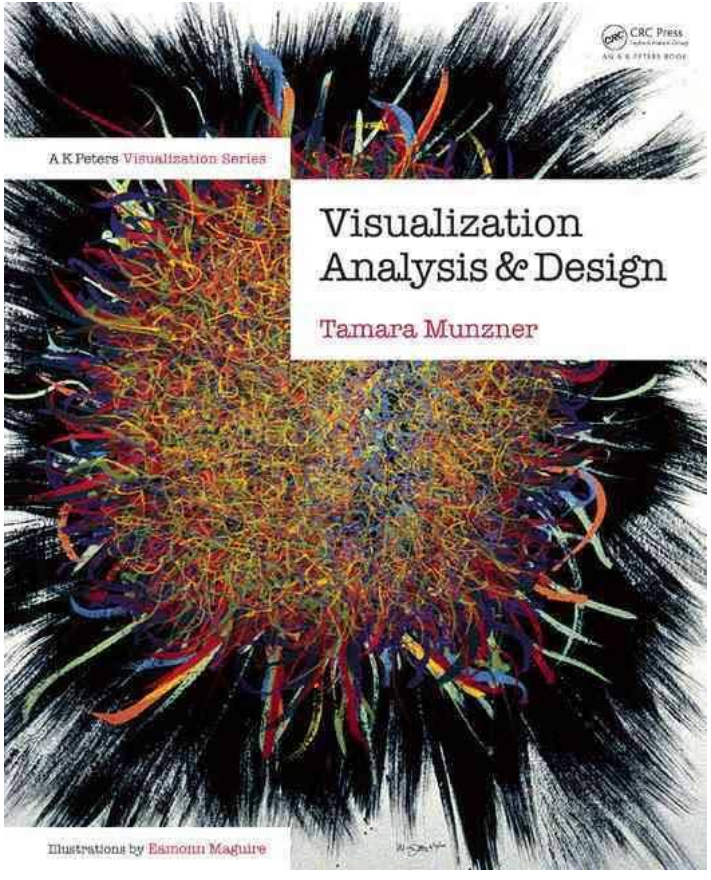
CONVENTIONS

- Check lecture on visual variable effectiveness
- Lines should connect things that go together
- When things are stacked people read the sum
- When things cluster together people compare them
- 3D charts are difficult to get right

CONCLUSIONS

- Find good ways to ask questions about the data
- Find mappings that will support your questions
- Creating effective visualizations is a process
- Enjoy the moment when meaning emerges

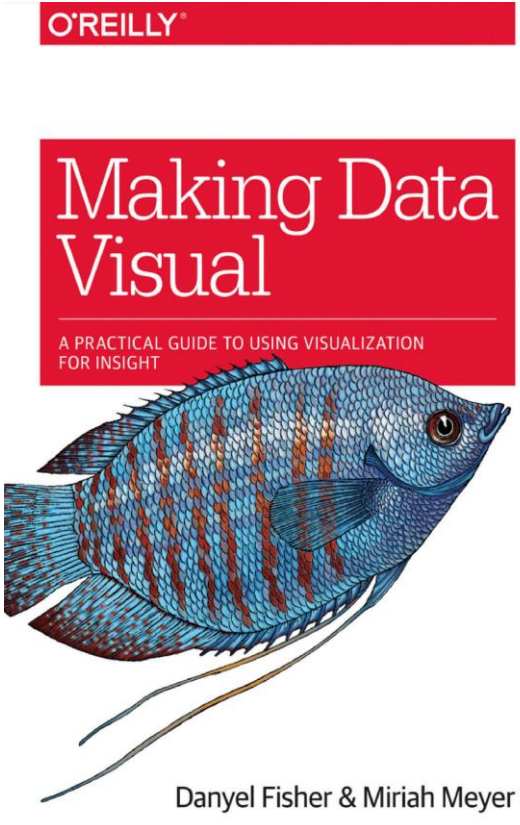
READINGS



https://www.data-to-viz.com

pretty nice overview but don't believe everything

READINGS


O'REILLY
Making Data Visual
A PRACTICAL GUIDE TO USING VISUALIZATION FOR INSIGHT
Danyel Fisher & Miriah Meyer