CSE/ISE 332
Introduction to Visualization

Applications and Basic Tasks

Klaus Mueller

Computer Science Department
Stony Brook University

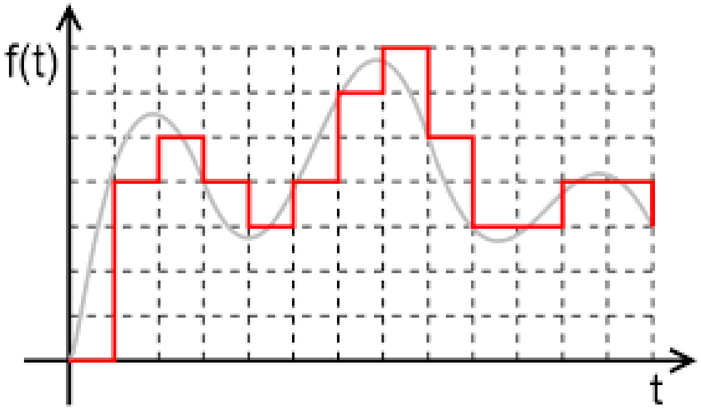
Lecture	Topic	Projects
1	Intro and logistics	
2	Basic visualizations and tasks, data types, examples, ethical considerations	
3	Data preparation (cleaning, imputation, data set integration)	Project #1 out
4	AI-assisted coding for VIS applications (design, debugging, refactoring)	
5	Big data and data reduction (distance/sim metrics, intro to clustering)	
6	High-D data: concept, subspaces, dimension reduction, PCA	
7	Cluster analysis: hierarchical, density, model, embedding, temporal	Project #2(a) out
8	Perception and cognition (human visual system, color, contrast)	
9	Visual design and aesthetics	
10	Visualization of multivariate and high-D data: linear methods, projections	
11	Vis. of multivariate and high-D data: non-linear methods, embeddings	Project #2(b) out
12	Visualization and AI: mutual support and capabilities (VIS4AI, AI4VIS)	
13	Principles of interaction: drive what is visualized, analyzed & how (HCI4VIS)	
14	Midterm recap	
15	Midterm	
16	Visual analytics, human-centered AI, mixed-initiative & collaborative VA	Capstone proposal due
17	Visualization system design & evaluation: nested model, study designs	
18	Visualization of hierarchical data	
19	Visualization of maps and data with geo-reference	
20	Vis. of time-varying, time-series, streaming data, progressive visualization	
21	Critiquing visualization: Tufte's views, integrity, uncertainty, responsibility	Capstone prelim report due
22	Design of effective infographics	
23	Foundations scientific & medical visualization, computer graphics	
24	Intro to volume rendering	
25	Scientific visualization	
26	Capstone project demos (4-5 minutes + 2 minutes Q&A)	All capstone materials due
27	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
28	Capstone project demos (4-5 minutes + 2 minutes Q&A)	
Final	Final exam	

Variable Types

Numerical Variables

Numerical variables

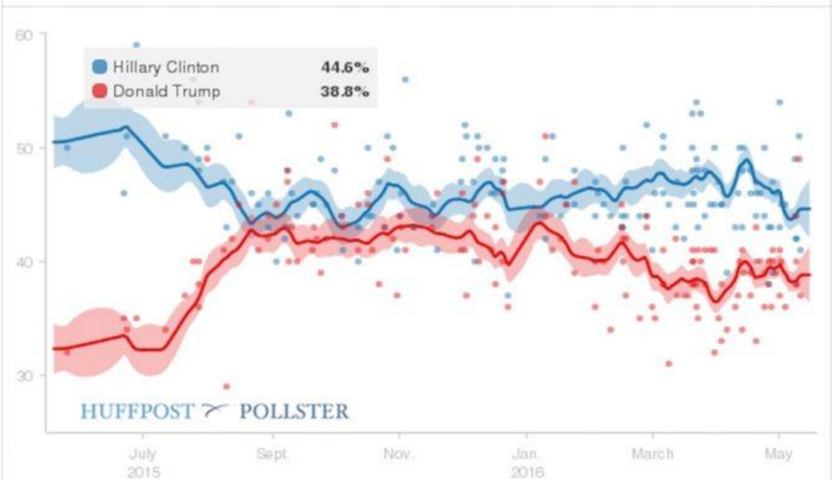
- measure a **quantity** as a number
- like: ‘how many' or 'how much'
- can be continuous (grey curve)
- or discrete (red steps)



Numerical Variables

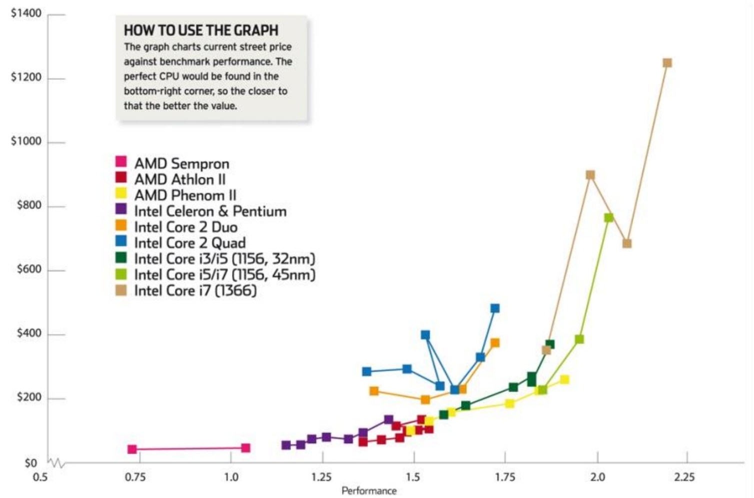
Most often the x-axis is ‘time’

- provides an intuitive & innate ordering of the data values
- the majority of people expect the x-axis to be ‘time’



But ‘time’ is not the only option

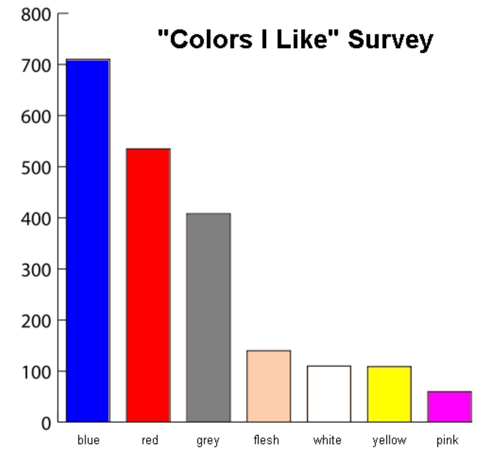
- engineers, statisticians, etc. will be receptive to this idea
- can you think of an example?



Categorical Variables

Categorical variables

- describe a **quality** or characteristic
- like: 'what type' or 'which category'

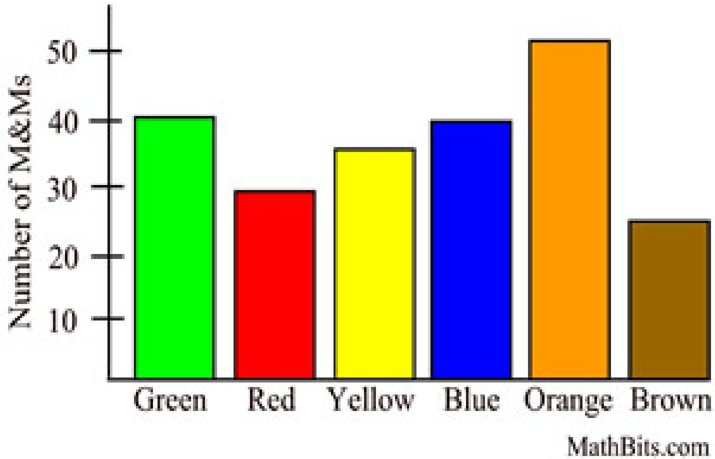


- can be **ordinal** = ordered, ranked (distances need not be equal)
  – clothing size, academic grades, levels of agreement

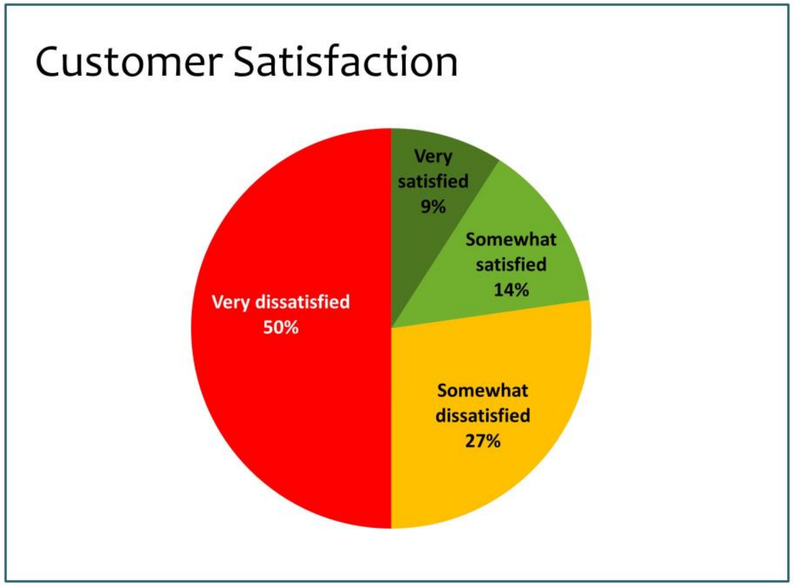
- or **nominal** = not organized into a logical sequence
  – gender, business type, eye color, brand

Categorical Variables

Usually plotted as bar charts or pie charts



Number of Colors in Bag of M&M Candies



??

??

nominal

ordinal

but of course, you can plot either of them in either of these two representations

Numbers are Good

But not everything is expressed in numbers

- images
- video
- text
- web logs
- …

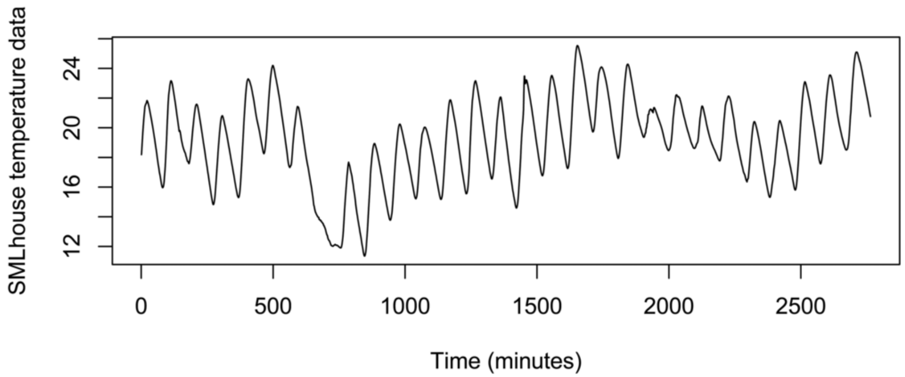
Do feature analysis to turn these abstract things into numbers

- a vector of numbers, to be concrete
- then apply your analysis as usual
- but keep the reference to the original data so you can return to the native domain where the analysis problem originated

Sensor Data

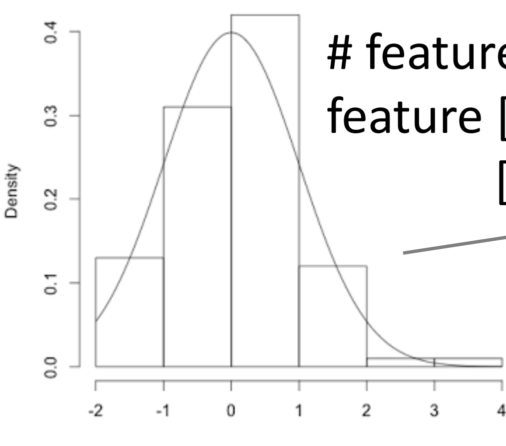
Characteristics

- often large scale
- time series

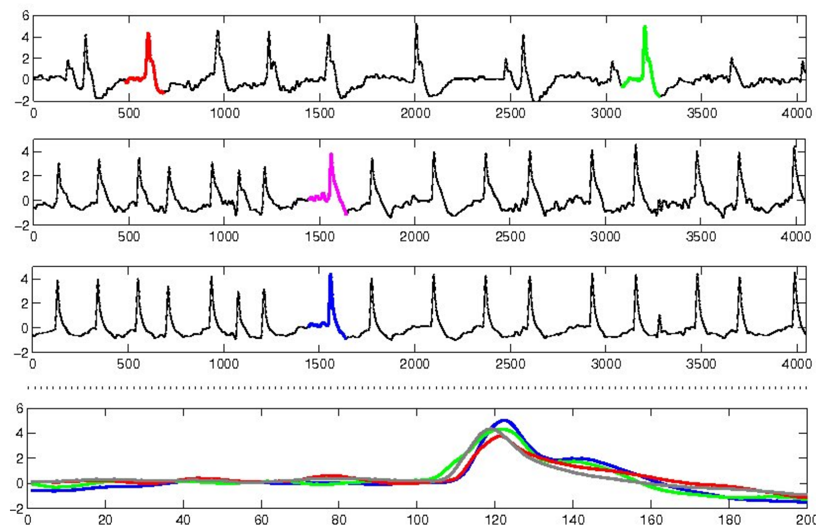


Feature Analysis

- example: Motif discovery
- encode into 5D data vector



features discovered in stream
feature [F1, F 2, F3, F4, F 5]
[12, 3, 41, 12, 5]

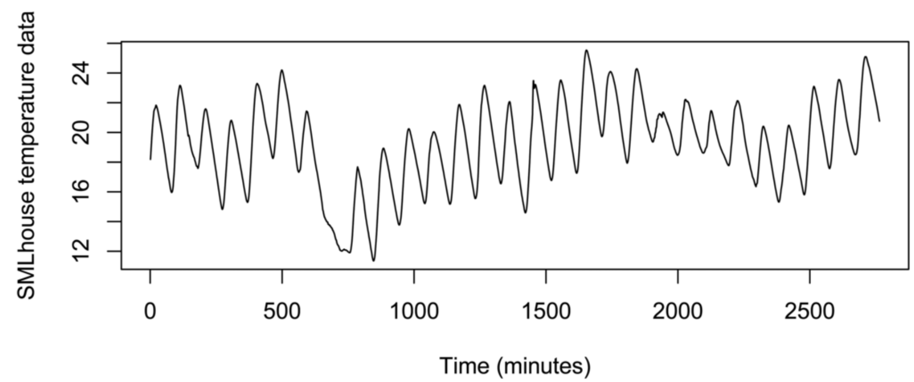


Five different known motifs

Sensor Data

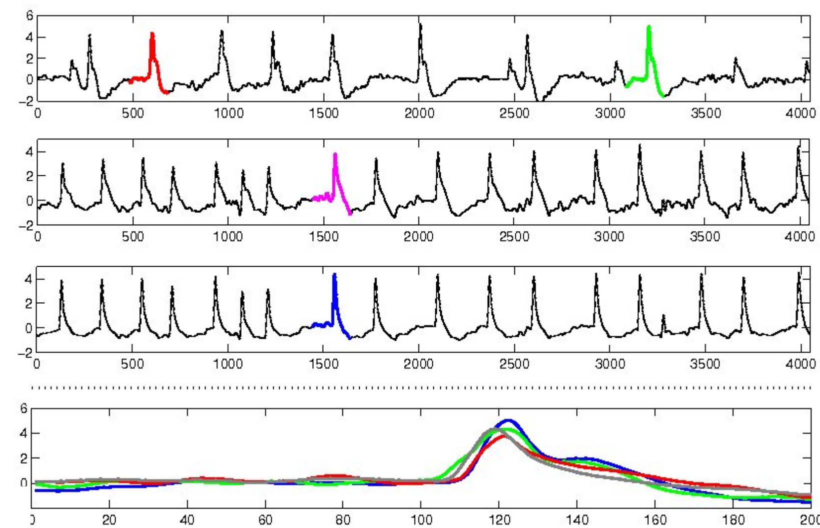
Characteristics

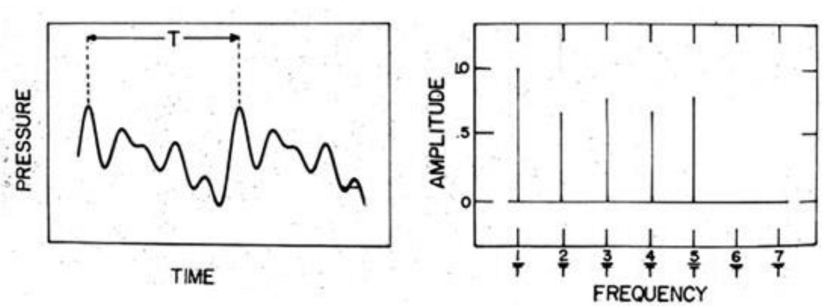
- often large scale
- time series



Feature Analysis

- Fourier transform (FT, FFT)
- Wavelet transform (WT, FWT)





Fourier transform
Store spectrum into a vector

Image Data

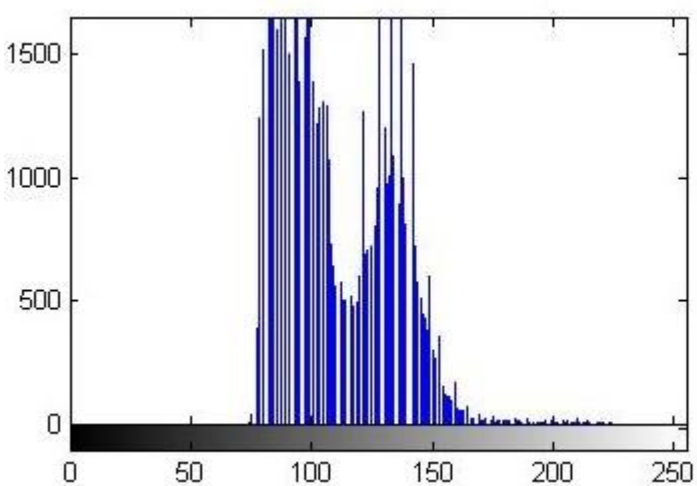
Characteristics

- array of pixels
- representable as a vector of length [width x height]

Feature Analysis

- value histograms
- encode into a 256-D vector

histograms



[0, 0, 0, …., 10, …, 1200, …..]

Image Data

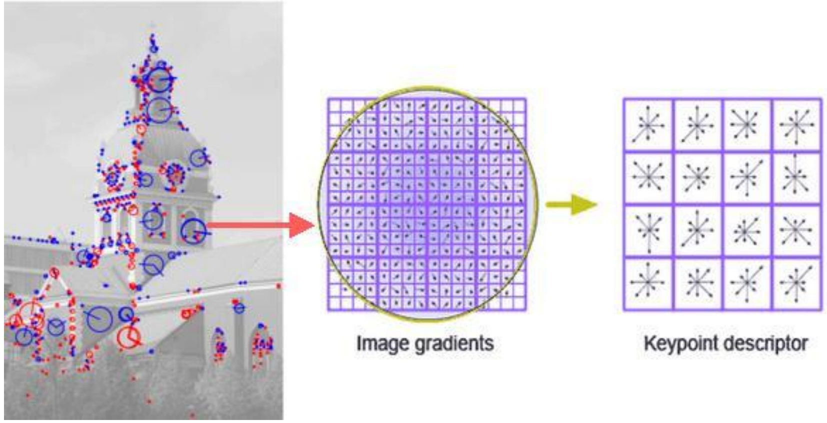
Characteristics

- array of pixels
- representable as a vector of length [width x height]

Feature Analysis

- value histograms
- gradient histograms
- FFT, FWT
- Scale Invariant Feature Transform (SIFT)
- Bag of Features (BoF)
- visual words

histograms



SIFT

Video data

Characteristics

- essentially a time series of images

Feature Analysis

- many of the above techniques apply albeit extension is non-trivial

Text Data

Characteristics

- often raw and unstructured

Feature analysis

- first step is to remove stop words and stem the data
- perform **named-entity recognition** to gain atomic elements
  - identify names, locations, actions, numeric quantities, relations
  - understand the structure of the sentence and complex events
- example:
  - Jim bought 300 shares of Acme Corp. in 2006.
  - $[\text{Jim}]_{\text{Person}}$ bought $[\text{300 shares}]_{\text{Quantity}}$ of $[\text{Acme Corp.}]_{\text{Organiz.}}$ in $[\text{2006}]_{\text{Time}}$
- distinguish between
  - application of grammar rules (old style, need experienced linguists)
  - statistical models (Google etc., need big data to build)

Text to Numeric Data

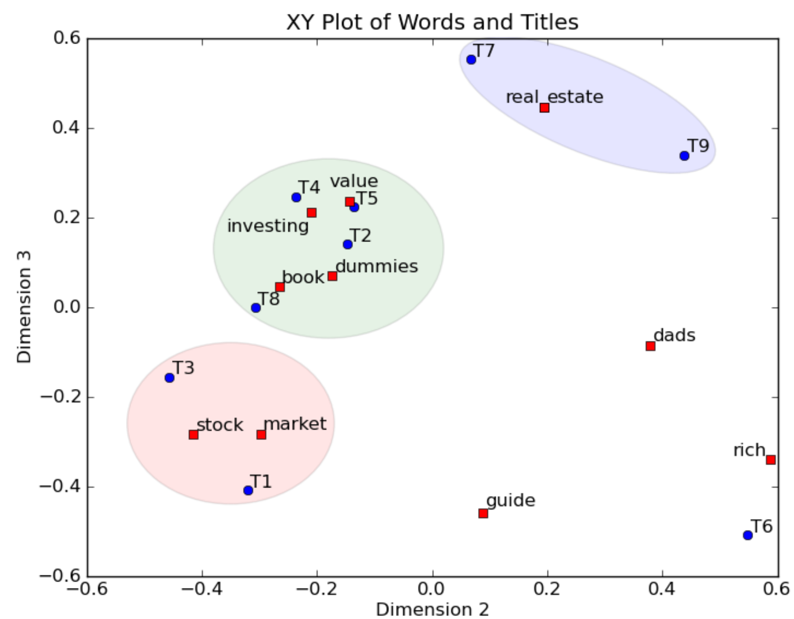
Create a term-document matrix

- turns text into a high-dimensional vector which can be compared
- use Latent Semantic Analysis (LSA) to derive a visualization

Index Words	Titles								
	T1	T2	T3	T4	T5	T6	T7	T8	T9
book			1	1					
dads						1			1
dummies		1						1	
estate							1		1
guide	1					1			
investing	1	1	1	1	1	1	1	1	1
market	1		1						
real							1		1
rich						2			1
stock	1		1					1	
value				1	1				

Term-Document Matrix

LSA $\longrightarrow$

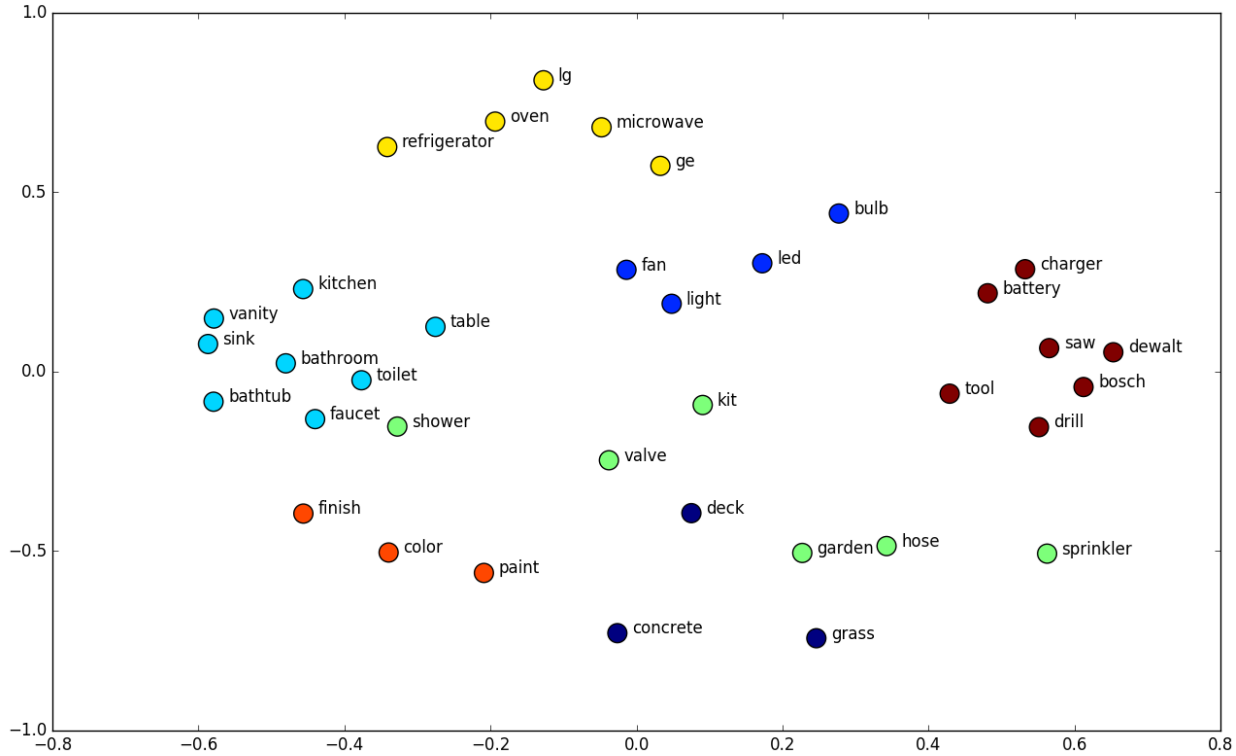


Word/document cluster

Word Embedding

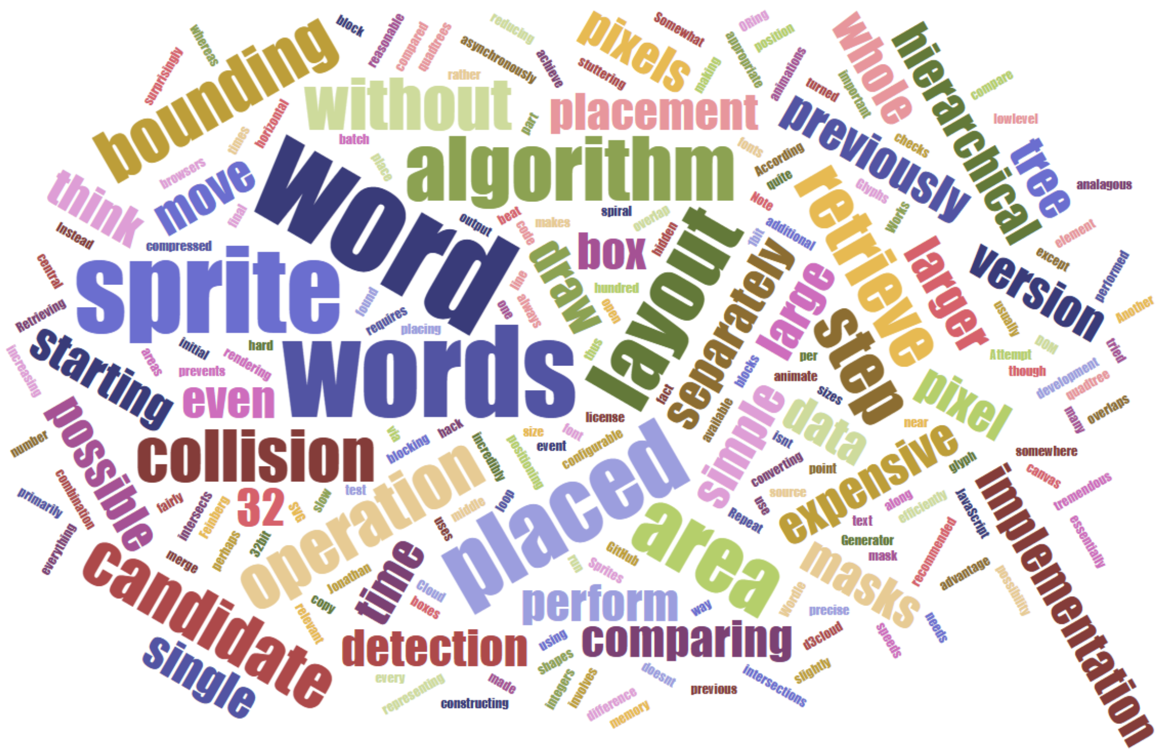
Train a shallow neural network (NN) on a corpus of text

- the NN weight vectors encode word similarity as a high-D vector
- use a 2D embedding technique to display



Word Cloud

Maps the frequency of words in a corpus to size



https://www.jasondavies.com/wordcloud/

Other data

Weblogs

- typically represented as text strings in a pre-specified format
- this makes it easy to convert them into multidimensional representation of categorical and numeric attributes

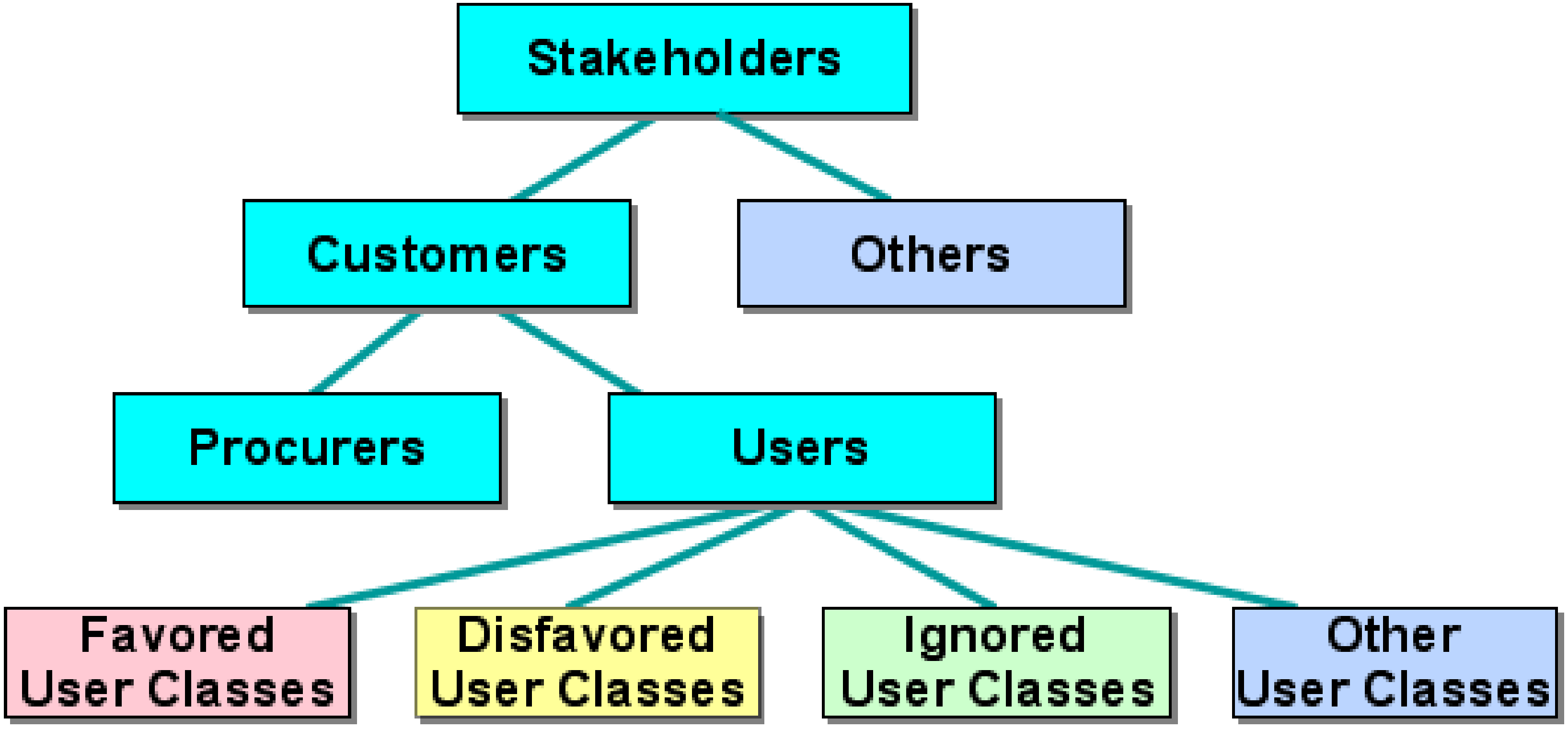
Network traffic

- characteristics of the network packets are used to analyze intrusions or other interesting activity
- a variety of features may be extracted from these packets
  - the number of bytes transferred
  - the network protocol used
  - IP ports used

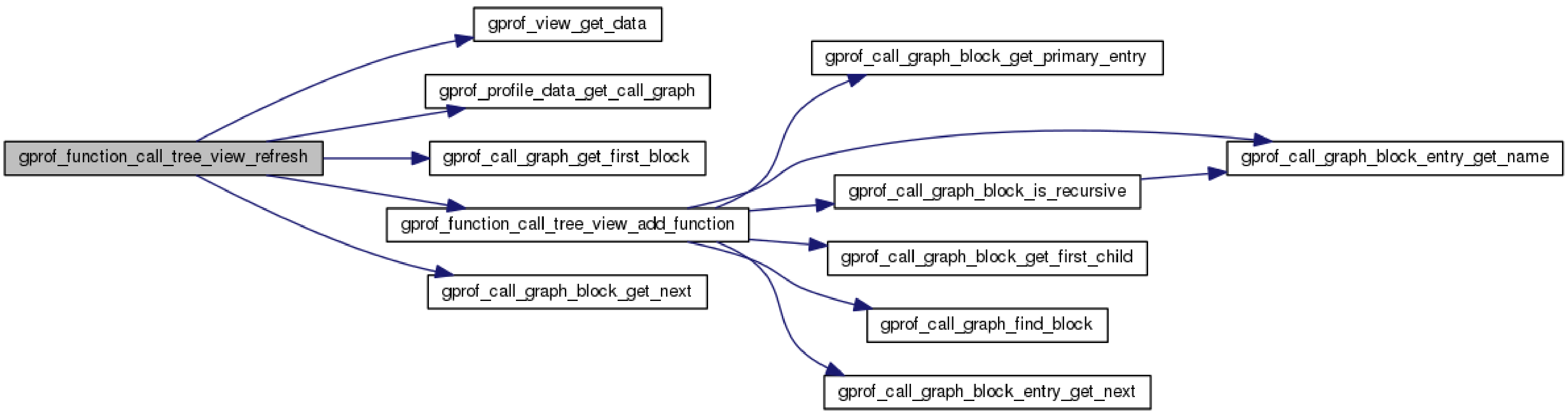
Let's Look at Some Essential Graphical Representations

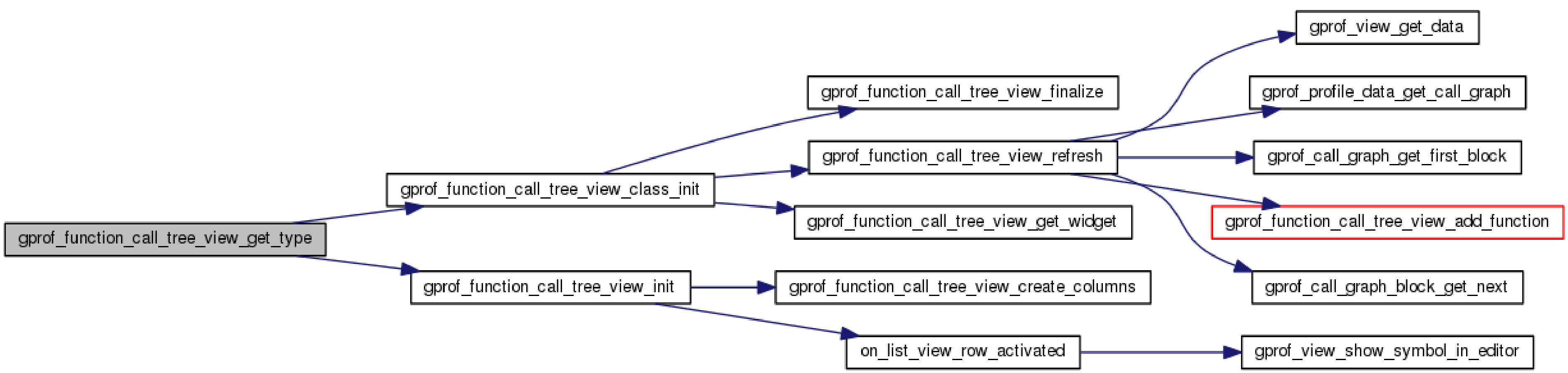
And Do Some Advertising for D3

Stakeholder Hierarchy

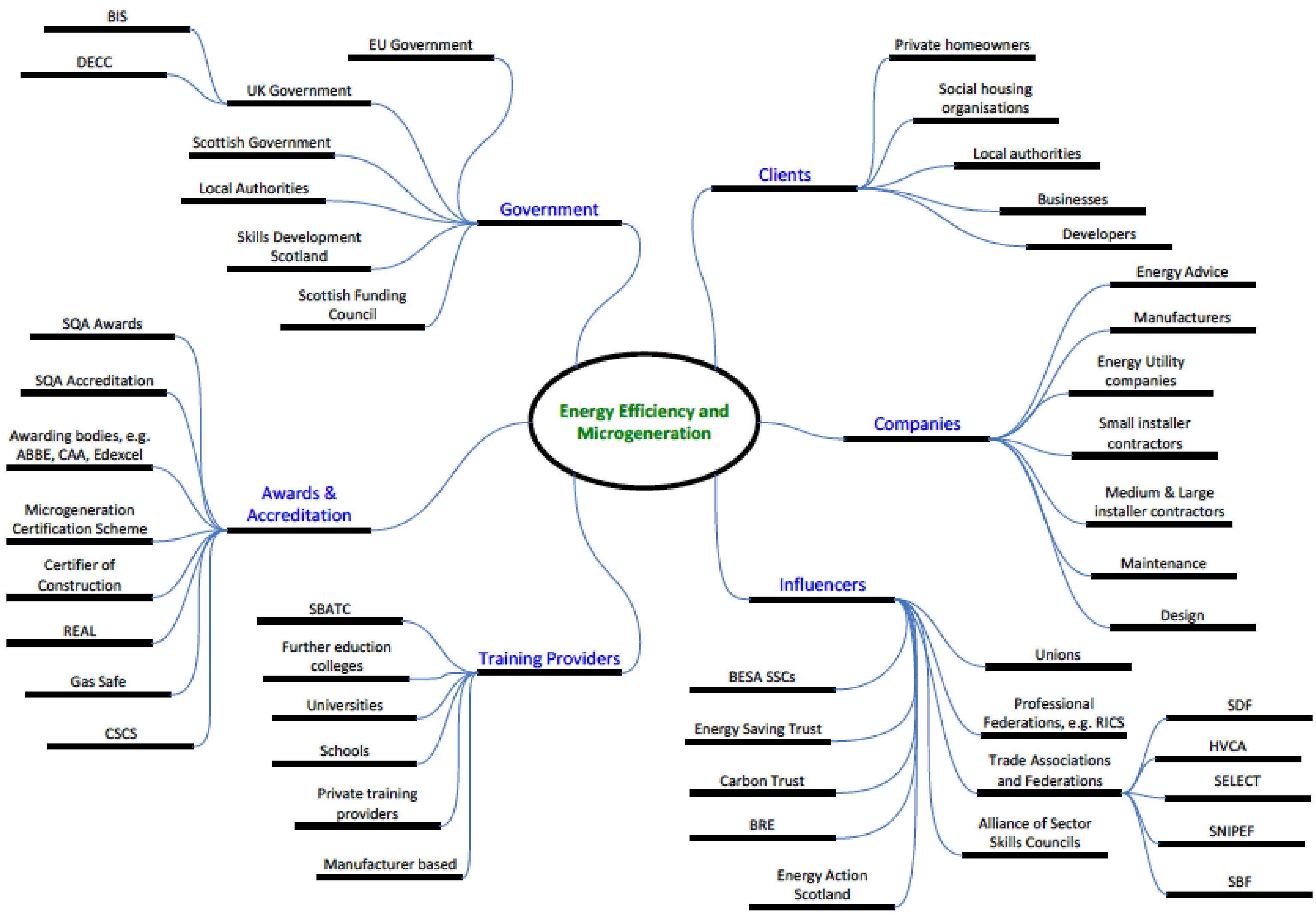


FUNCTION CALL TREE





More Complex Stakeholder Hierarchy



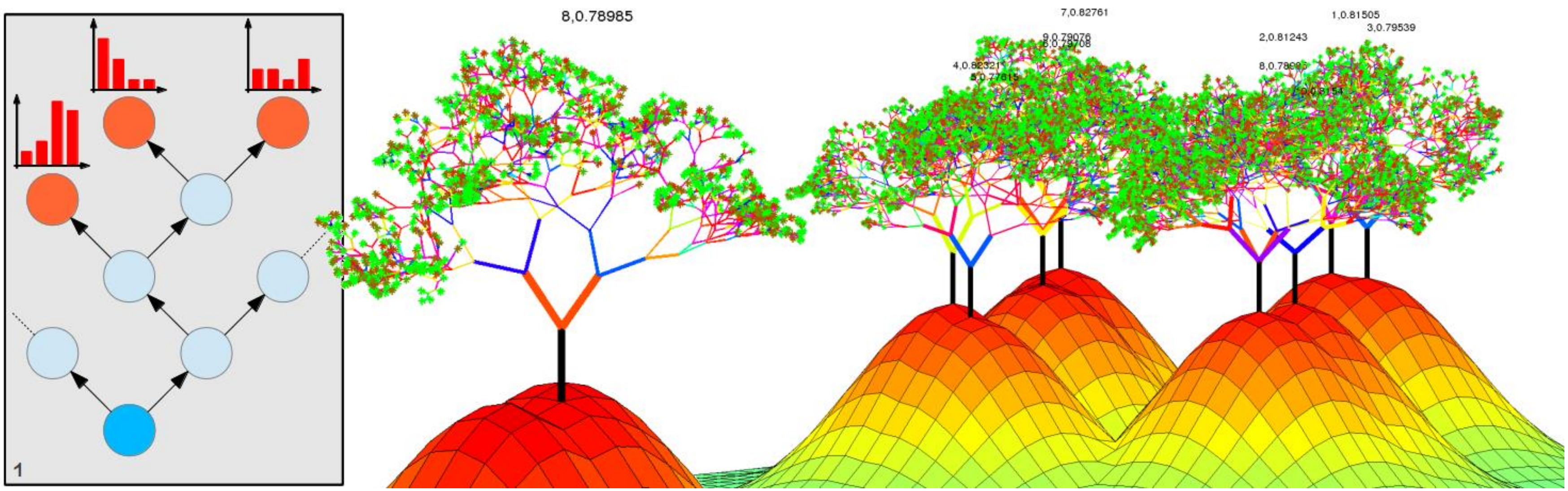
Hierarchies

Questions you might have

- how large is each group of stakeholders (or function)?
  - tree with quantities
- what fraction is each group with respect to the entire group?
  - partition of unity
- how is information disseminated among the stakeholders (or functions)?
  - information flow
- how close (or distant) are the individual stakeholders (functions) in terms of some metric?
  - force directed layout

Invoke Nature

More scalable tree, and natural with some randomness



from: [A small guide to Random Forest]

Collapsible Tree

A standard tree, but one that is scalable to large hierarchies

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

Zoomable Partition Layout

A tree that is scalable and has partial partition of unity

https://mbostock.github.io/d3/talk/20111018/partition.html

Sunburst

More space efficient since it's radial, has partial partition of unity

https://observablehq.com/@d3/zoomable-sunburst

Bubble Charts

No hierarchy information, just quantities

https://observablehq.com/@d3/bubble-chart

Circle Packing

Quantities and containment, but not partition of unity

http://mbostock.github.io/d3/talk/20111116/pack-hierarchy.html

Treemap

Quantities, containment, and full partition of unity

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

Chord Diagram

Relationships among group fractions, not necessarily a tree

https://observablehq.com/@d3/chord-diagram

Hierarchical Edge Bundling

Relationships of individual group members, also in terms of quantitative measures such as information flow

http://mbostock.github.io/d3/talk/20111116/bundle.html

Collapsible Force Layout

Relationships within organization members expressed as distance and proximity

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

Voronoi Tessellation

Shows the closest point on the plane for a given set of points… and a new point via interaction

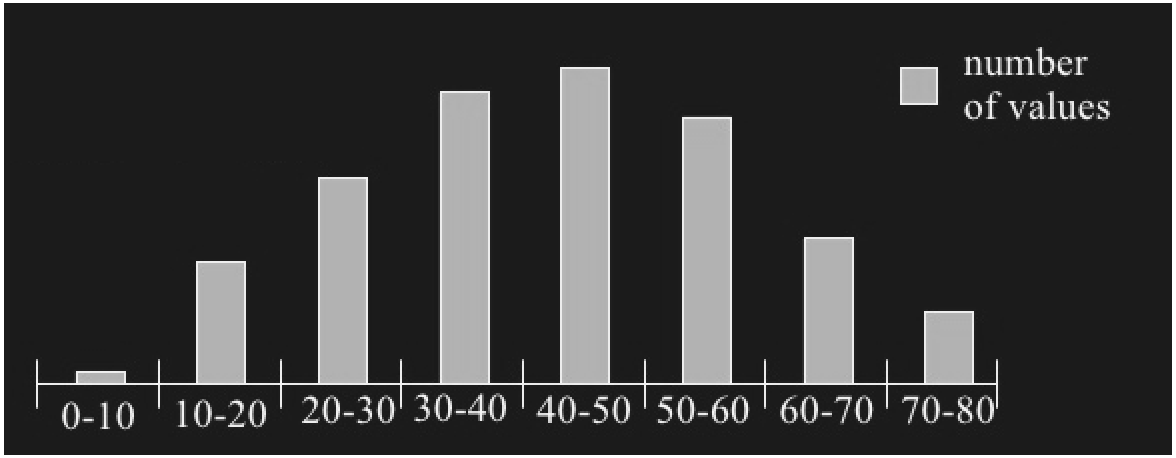
https://observablehq.com/collection/@d3/d3-delaunay

Data Type Conversions and Transformation

Numeric to Categorical Data: Discretization (1)

Solution 1:

- divide the numeric attribute values into φ **equi-width** ranges
- each range/bucket has the same width
- example: customer age

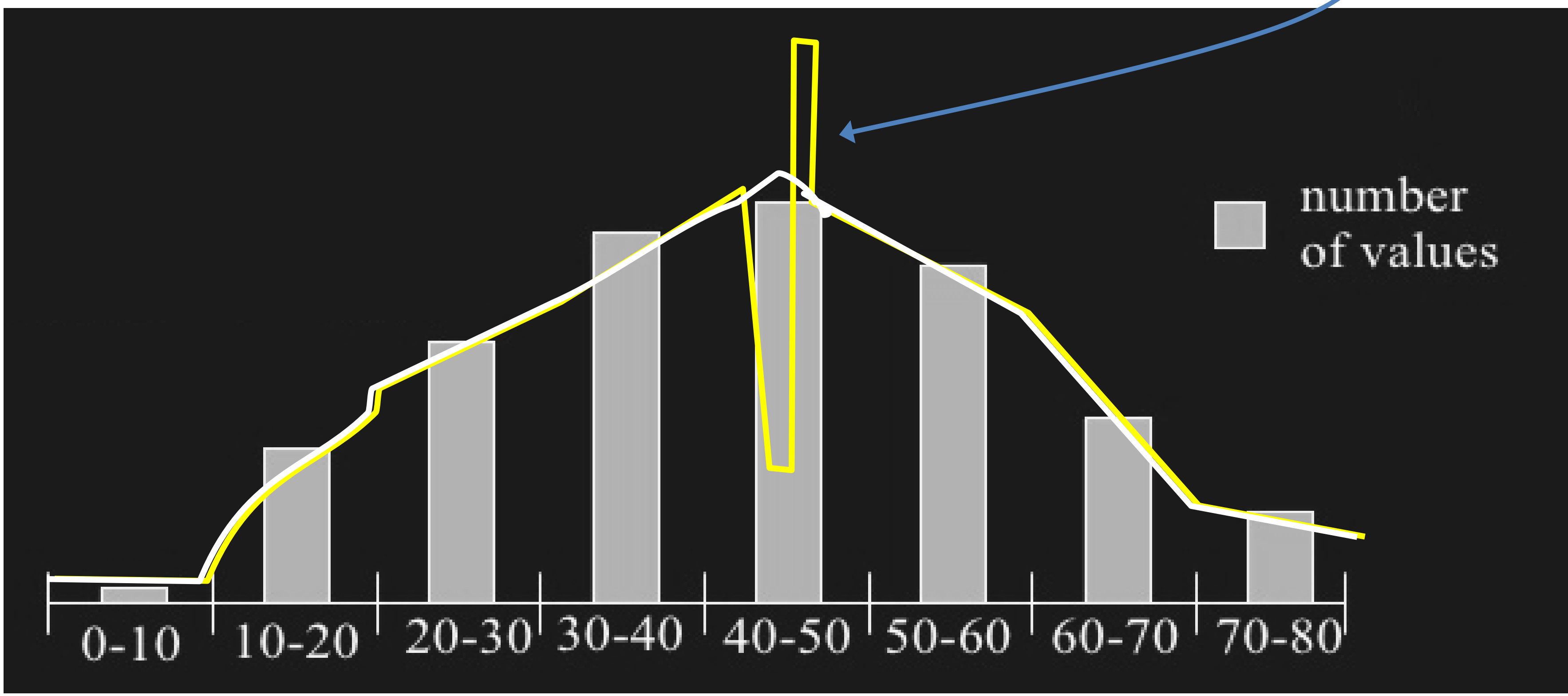


- what is lost here?

Problem With Equi-Width Histogram

Age ranges of customers could be unevenly distributed within a bin

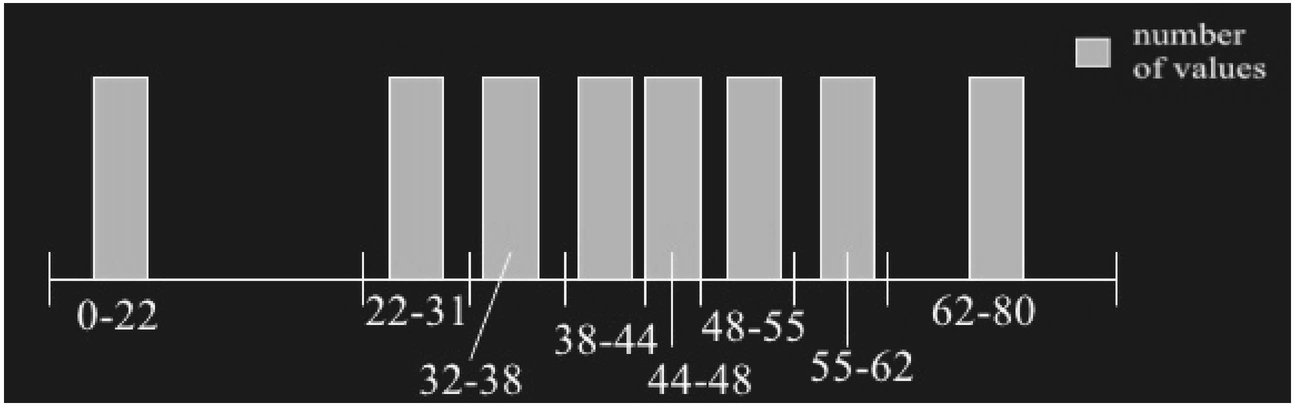
- this could be an interesting anomaly



Numeric to Categorical Data: Discretization (2)

Solution 2:

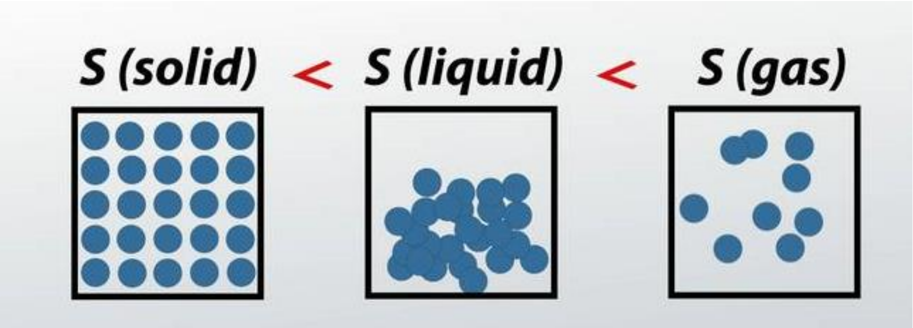
- divide the numeric attribute values into φ **equi-depth** ranges
- same number of samples in each bin
- (again) example: customer age:



- what is the disadvantage here?
- extra storage needed: must store the start/end value for each bin

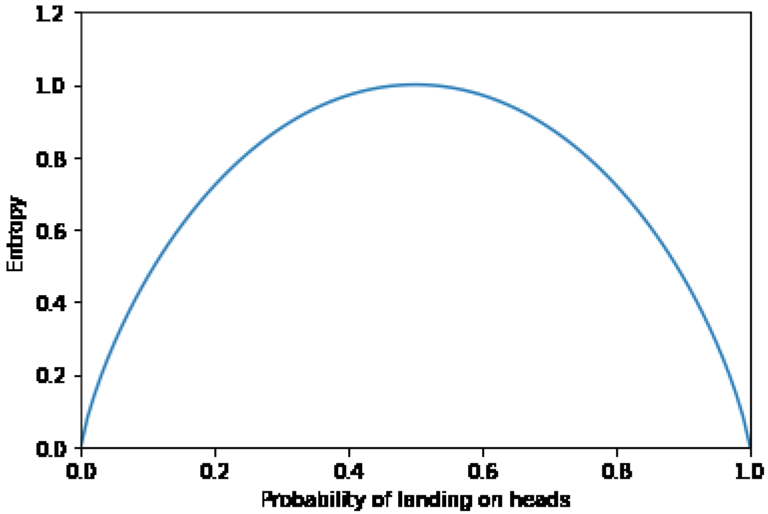
One More Binning Method

Entropy-based binning



Entropy is the amount of surprise to make a certain observation

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i)$$



THE DATA

O-Ring Failure	Temperature
Y	53
Y	56
Y	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
N	70
Y	70
Y	70
Y	70
N	72
N	73
N	75
Y	75
N	76
N	76
N	78
N	79
N	80
N	81

from
https://www.saedsayad.com/supervised_binning.htm

ENTROPY BASED BINNING (EBB)

Aim:

- find the best split so that the bins are as pure as possible,
- -> the majority of the values in a bin have the same class label
- formally, it is characterized by finding the split with the maximal information gain.

Step 1: Calculate "Entropy" for the target.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

O-Ring Failure	
Y	N
7	17

E (Failure) = E(7, 17) = E(0.29, .71) = -0.29 x $\log_2$(0.29) - 0.71 x $\log_2$(0.71) = **0.871**

Entropy Based Binning (EBB)

Step 2: Calculate "Entropy" for the target given a bin.

$$E(S,A) = \sum_{v \in A} \frac{|S_v|}{|S|} E(S_v)$$

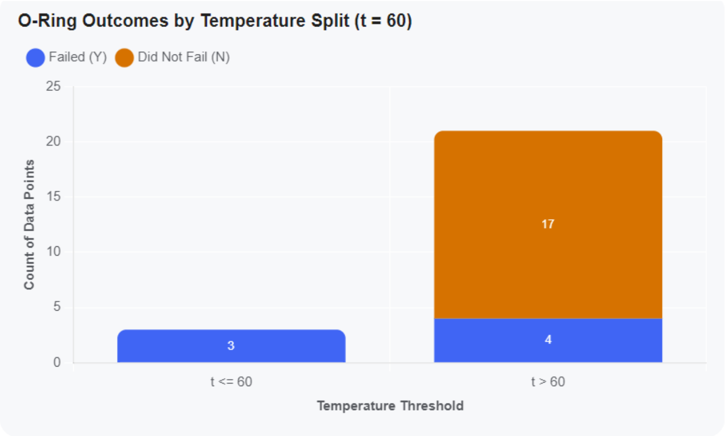
		O-Ring Failure	
		Y	N
Temperature	<= 60	3	0
	> 60	4	17

E (Failure,Temperature) = P(<=60) x E(3,0) + P(>60) x E(4,17) = 3/24 x 0 + 21/24 x 0.7= **0.615**

Step 3: Calculate "Information Gain" given a bin.

$$\textit{Information Gain} = E(S) - E(S,A)$$

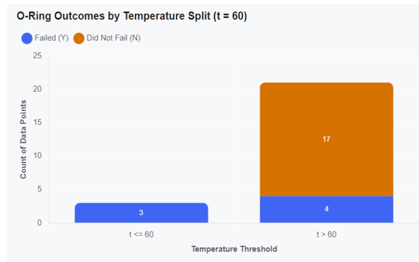
Information Gain (Failure, Temperature) = **0.256**



Entropy Based Binning (EBB)

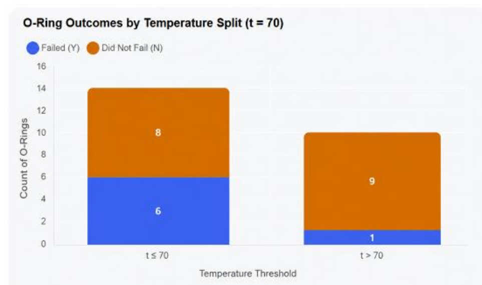
Iterate over further splits for bins with highest entropies

[<=60, >60] turns out to be the best split



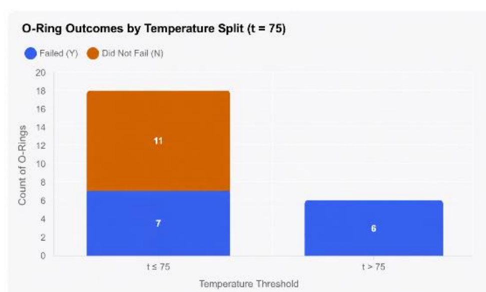
Gain = 0.256

		O-Ring Failure Y	N
Temperature	<=60	3	0
	>60	4	17



Gain = 0.101

		O-Ring Failure Y	N
Temperature	<=70	6	8
	>70	1	9



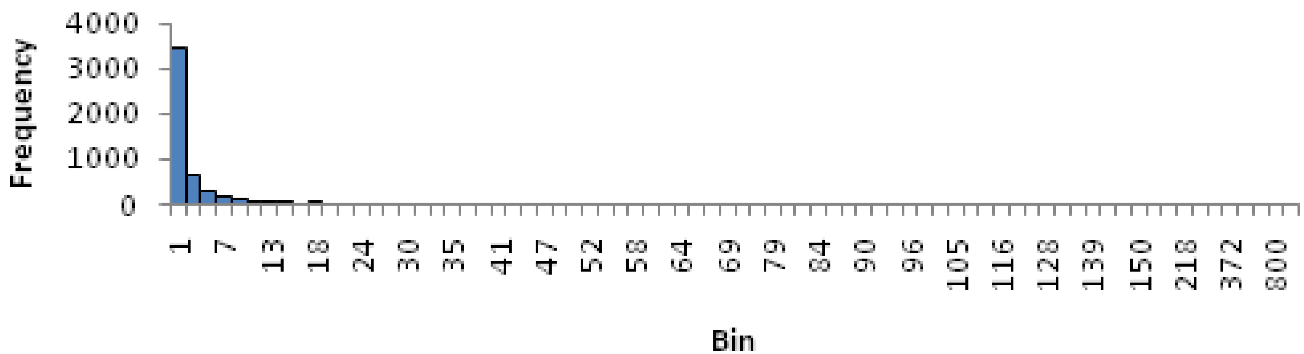
Gain = 0.148

		O-Ring Failure Y	N
Temperature	<=75	7	11
	>75	0	6

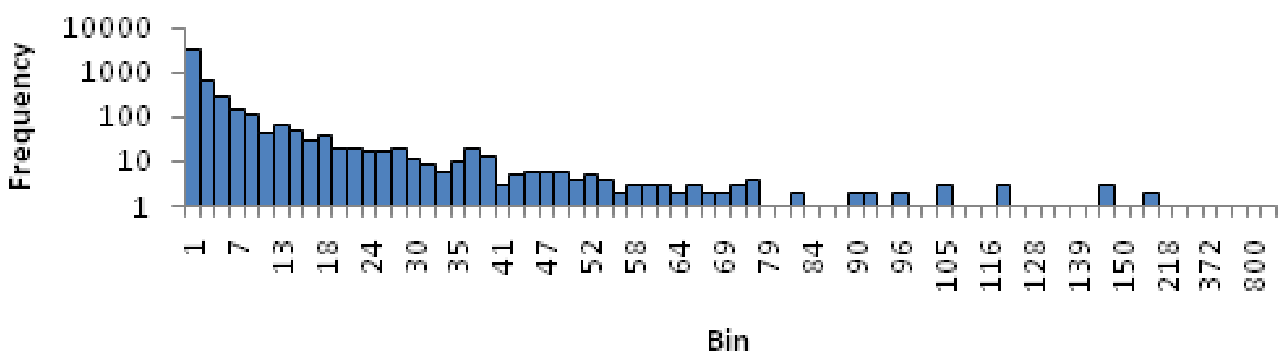
Numeric to Categorical Data: Discretization (3)

Solution 3:

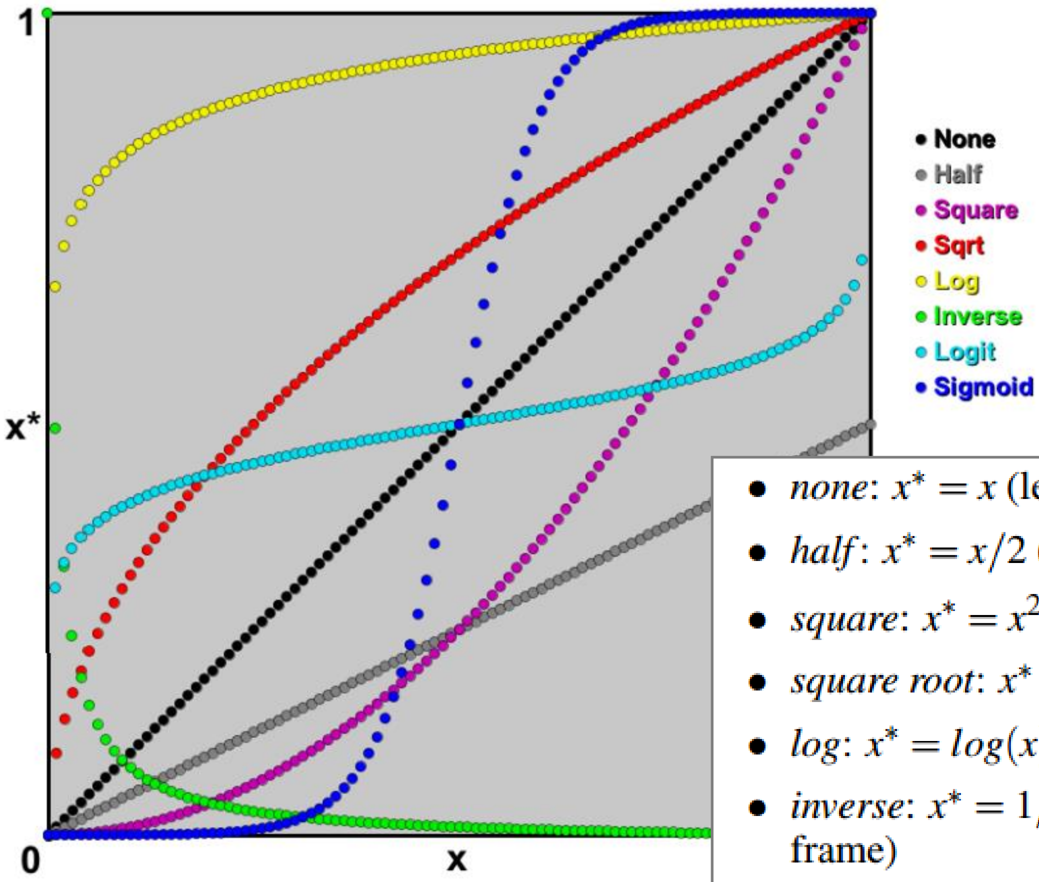
- what if all the bars have seemingly the same height
- or are dominated by one large peak



- switch to log scaling of the y-value



Other Transformations



- *none*: $x^* = x$ (leaves points unchanged)
- *half*: $x^* = x/2$ (squeezes all points together)
- *square*: $x^* = x^2$ (pulls points toward left of frame)
- *square root*: $x^* = \sqrt{x}$ (mildly pulls points toward right of frame)
- *log*: $x^* = log(x)$ (strongly pulls points toward right of frame)
- *inverse*: $x^* = 1/x$ (reverses scale and squeezes points into left of frame)
- *logit*: $x^* = (log(x/(1-x)) + 10)/20$ (squeezes points toward middle of frame)
- *sigmoid*: $x^* = 1/(1 + exp(-20x + 10))$ (expands points away from middle of frame)

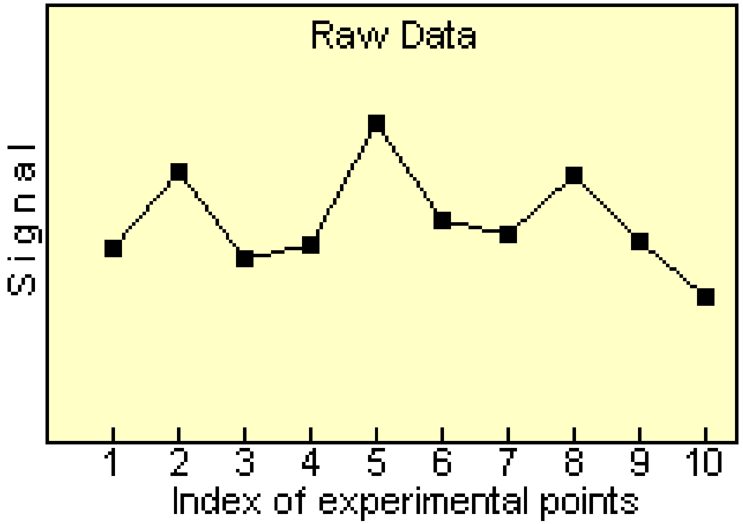
Dang and Wilkinson, "Transforming Scagnostics to Reveal Hidden Features", TVCG 2014

Infinite Zooms

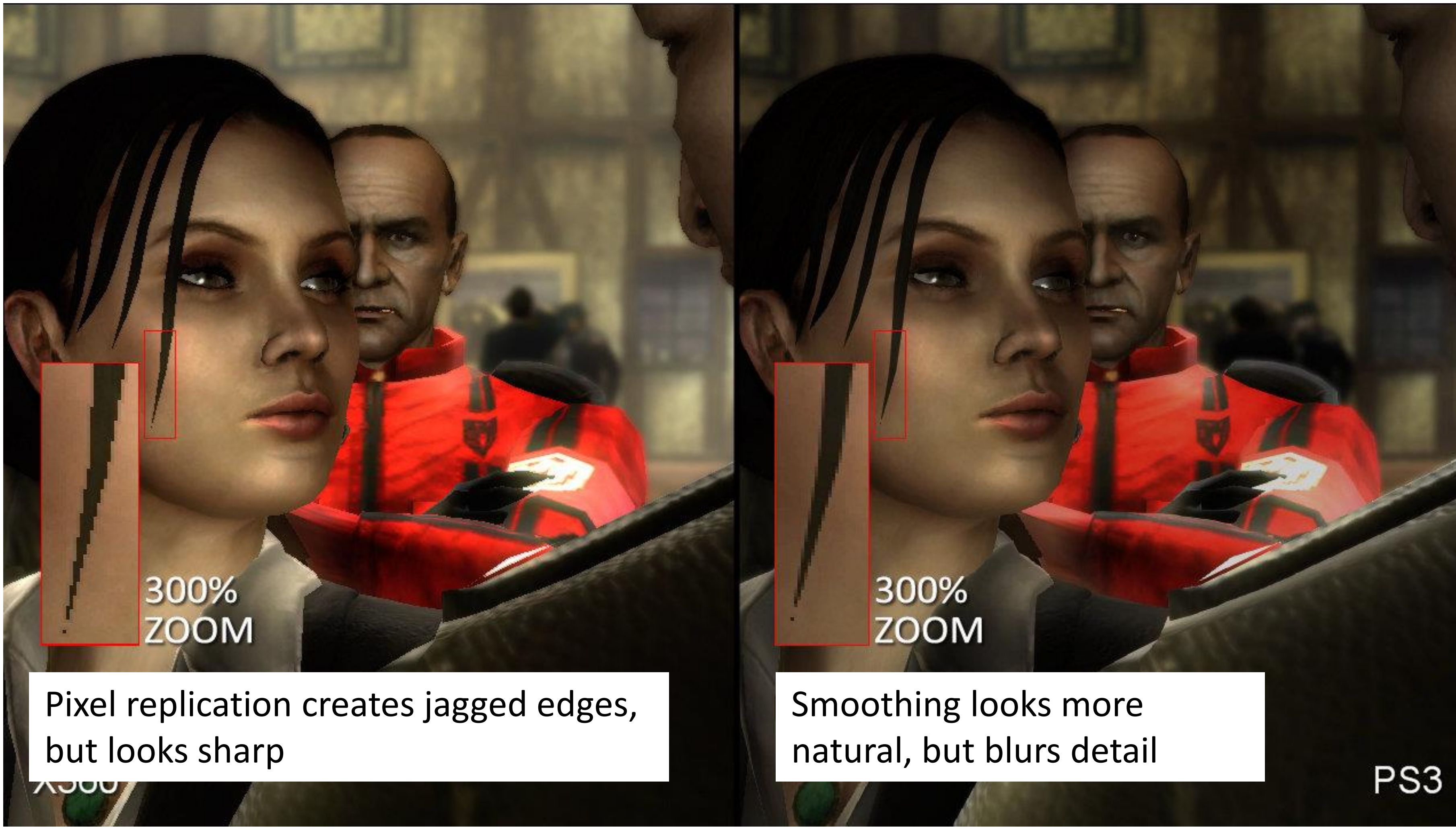
Prelude- How To Smooth a Discrete Signal?

Slide a window across the signal

- stop at each discrete sample point
- average the original data points that fall into the window
- store this average value at the sample point
- move the window to the next sample point
- repeat



Zooming into Images

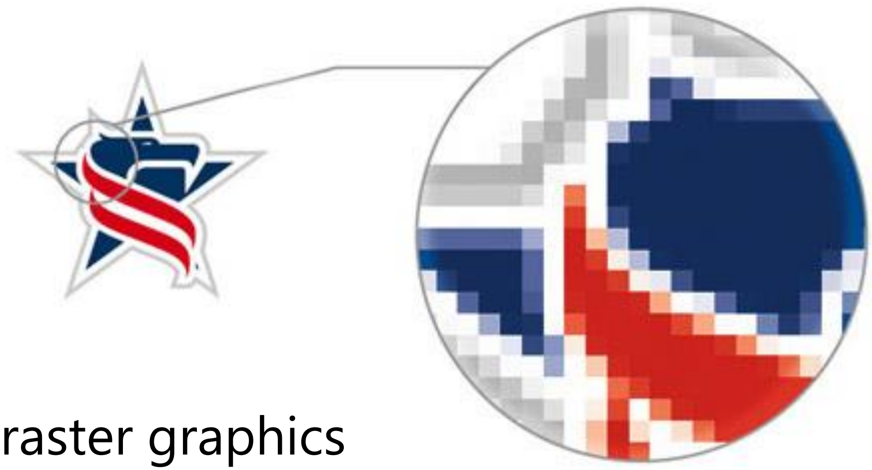


Anti-Aliasing Via Smoothing

The Solution

What's the underlying problem?

- detail can't be refined upon zoom
- can just be replicated or blurred



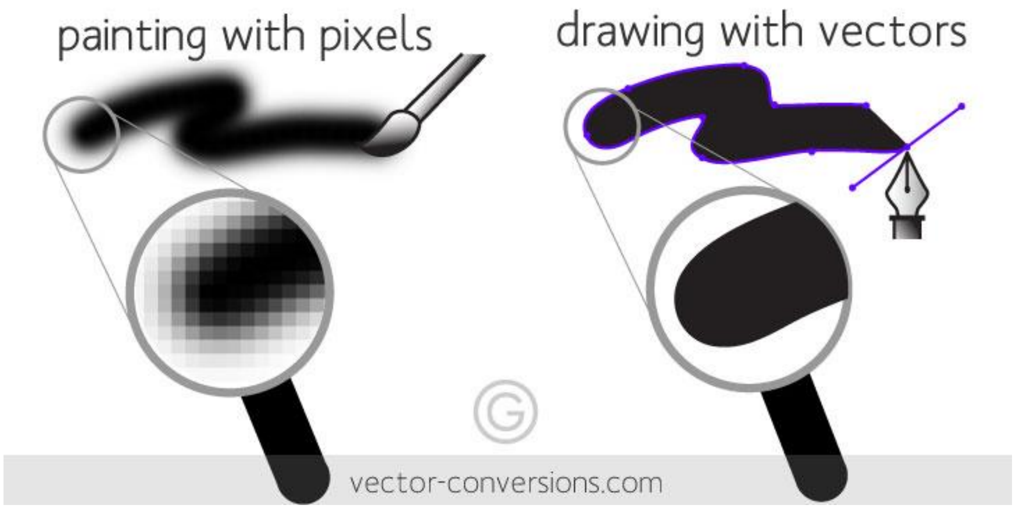
raster graphics

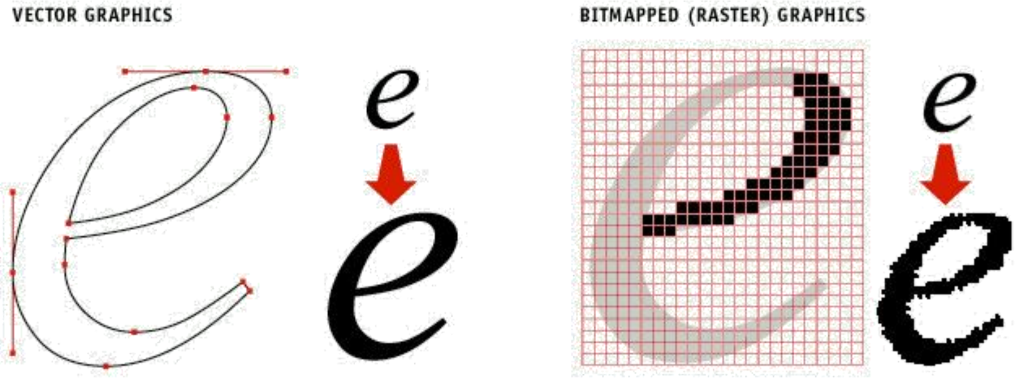
vector graphics

The solution…

- represent detail as a function that can be mathematically refined
- replace raster graphics by vector graphics

Scalable Vector Graphics (SVG)





Photographs and Images in SVG

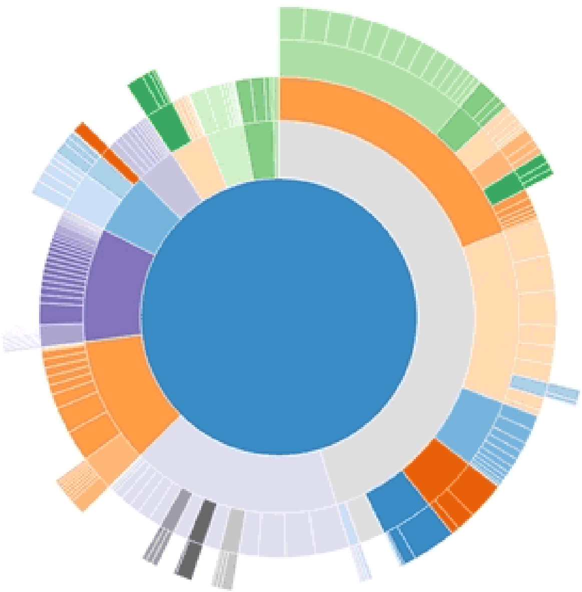
Vector graphics tends to have an "cartoonish" look

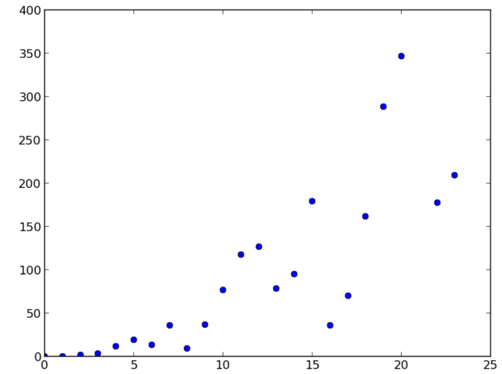
raster graphics

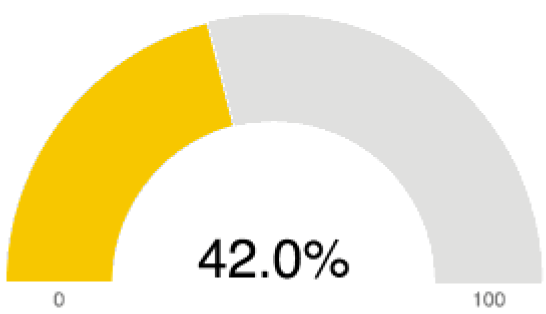
vector graphics

Photographs and Images in SVG

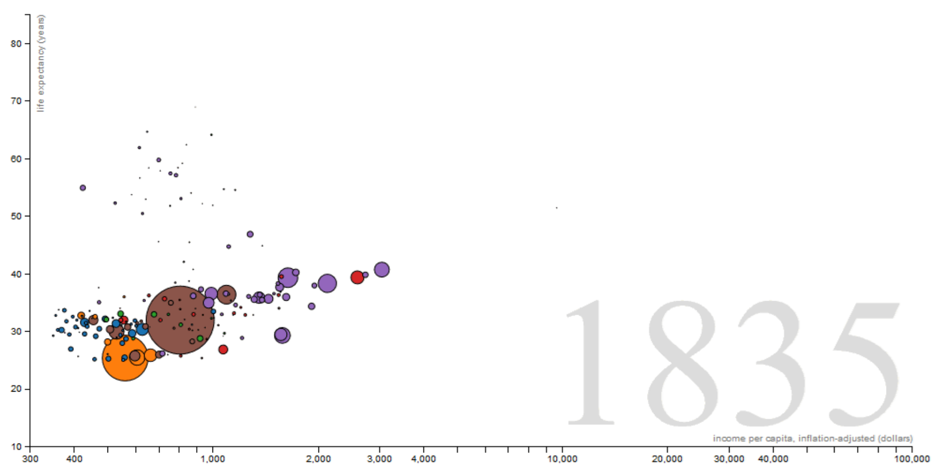
D3 Uses SVG







The Wealth & Health of Nations

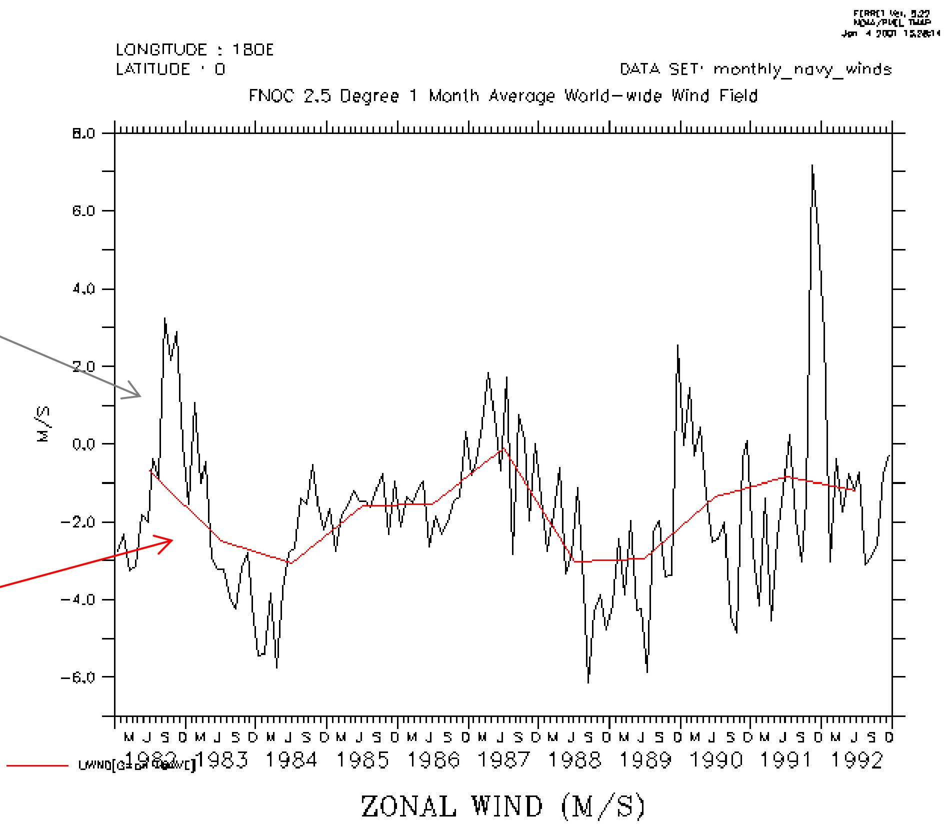


Smoothing for De-Noising

Filtering/smoothing also eliminates noise in the data

noise

data



Let's Talk About Bar Charts

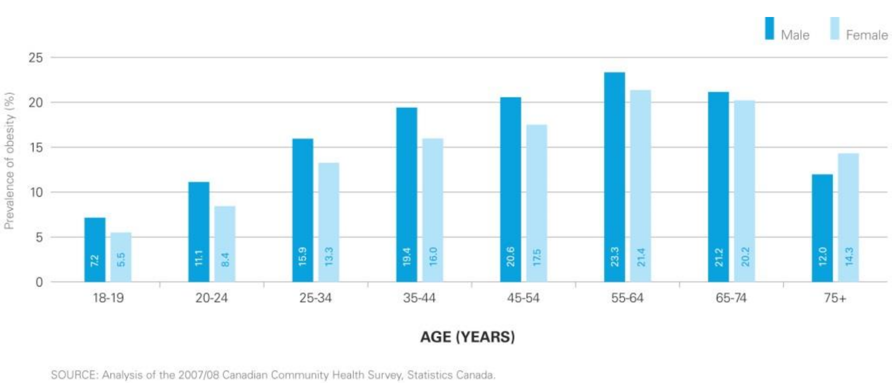
Bar Charts Can Be Data Smoothers

In some ways, bar charts reduce noise and uncertainties in the data

- the bins do the smoothing

Example:

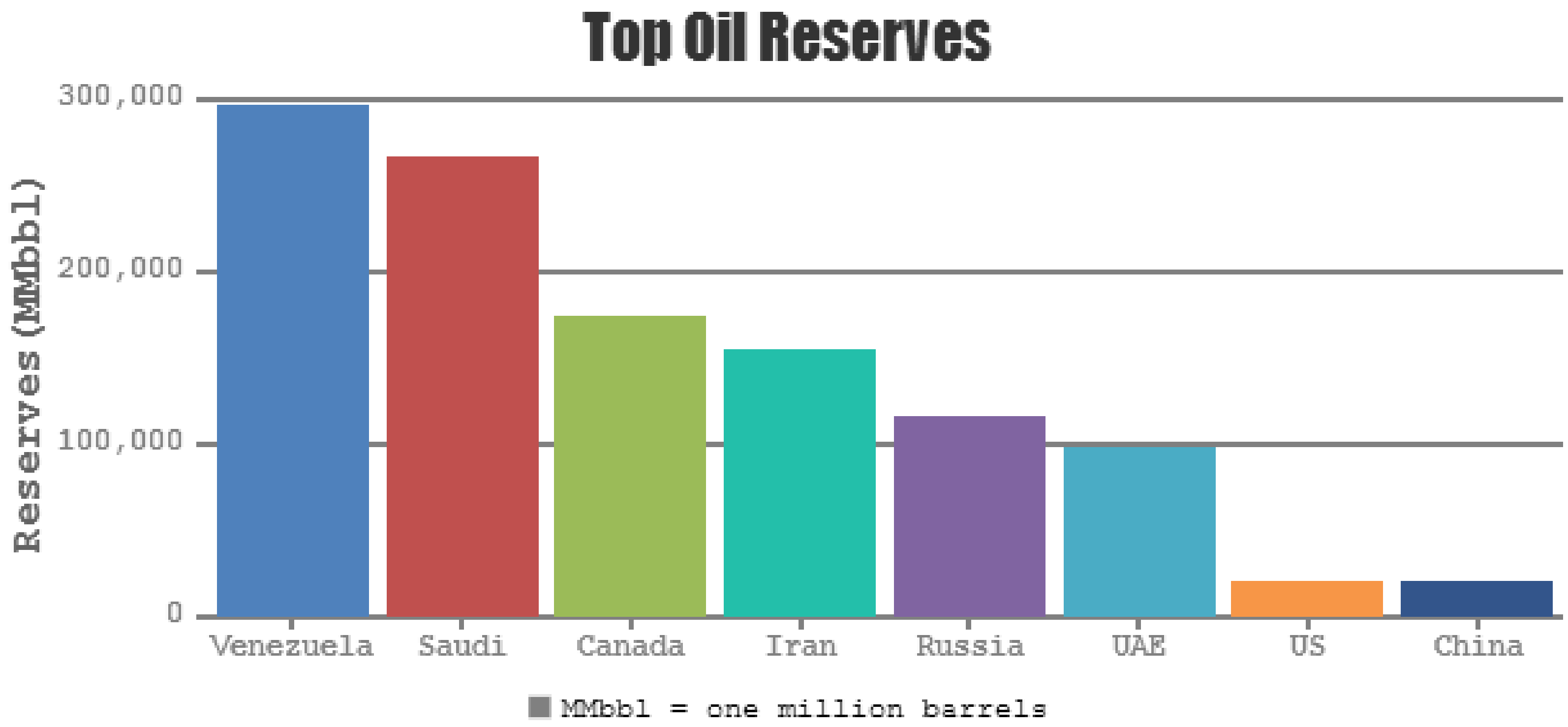
- obesity over age (group)



Categorical Bar Charts

Bar charts that hold categorical data

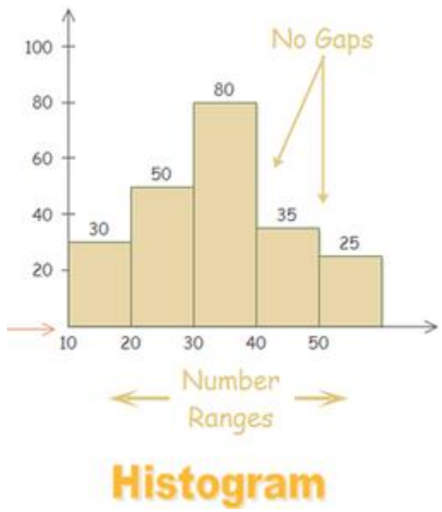
- smoothing by semantic grouping
- for example, Europe vs. {France, Spain, Italy, Germany, …}



Bar Charts vs. Histograms

Histograms

- bars show the frequency of numerical data
- quantitative data
- elements are grouped together, so that they are considered as ranges
- bars cannot be reordered
- width of bars need not be the same



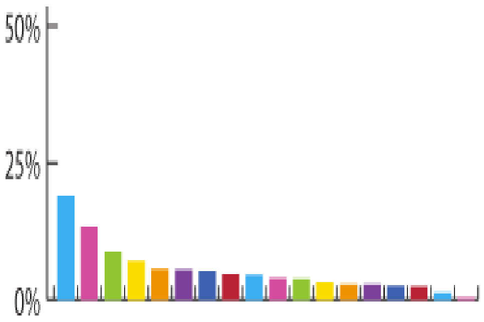
Bar charts

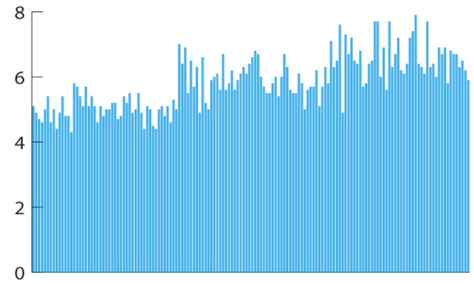
- uses bars to compare different categories of data
- comparison of discrete variables
- elements are taken as individual entities
- bars can be reordered
- width of bars need to be the same

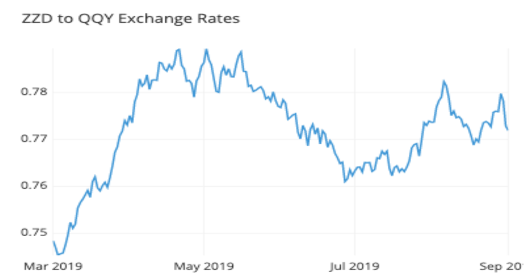
How many Bars in a Bar Chart

How many bars are too many (in a chart)

- if individual categories are the focus? 12 is a good rule
- if the overall trend is the important factor? 50 or even more
- eventually you can switch to a line chart







- sort bars by height and use ‘other’ to aggregate the bar chart tails into a single bar
- find a grouping that can semantically aggregate bars, for example aggregate countries into continents

[more information]()

Bar Charts in D3

https://observablehq.com/@d3/bar-chart/2

Working with bar charts and histograms is the topic of Lab 1

- the next two slides offer some help with calculations

Histogram Calculations – Binning

Determine bin size

- min(data) is optional, can also use 0 or some reasonable value
- max(data) is optional, can also use some reasonable value

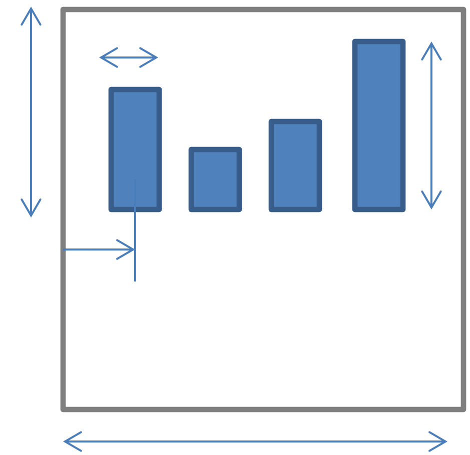
$$bin\ size = \frac{\max(data) - \min(data)}{number\ of\ bins}$$

Given a data value *val* increment (++) the bin value

- but first initialize bin val array to 0

$$bin\ val\ array\left[\left\lfloor \frac{val - \min(data)}{bin\ size} \right\rfloor\right]++$$

Histogram Calculations – Plotting



Determine bin size on the screen

$$bin\ size\ on\ screen = \frac{chart\ width}{number\ of\ bins}$$

Center of a bar for bin with index *bin index*

$$bar\ center\ on\ screen = (bin\ index\ \cdot bin\ size\ on\ screen) + 0.5$$

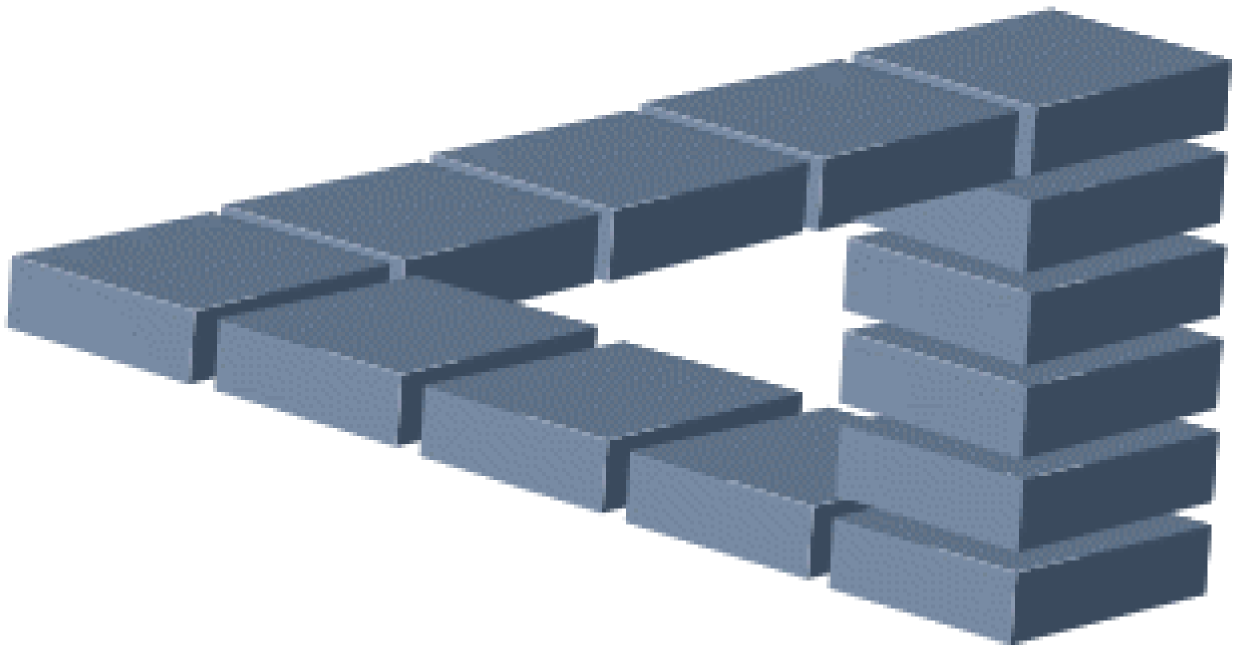
Height of the bar for a bin with index *bin index*

$$bar\ height(bin\ index) = bin\ val\ array(bin\ index) \cdot \frac{chart\ height}{\max(bin\ val\ array)}$$

Do not forget that the origin of a web page is the top left corner

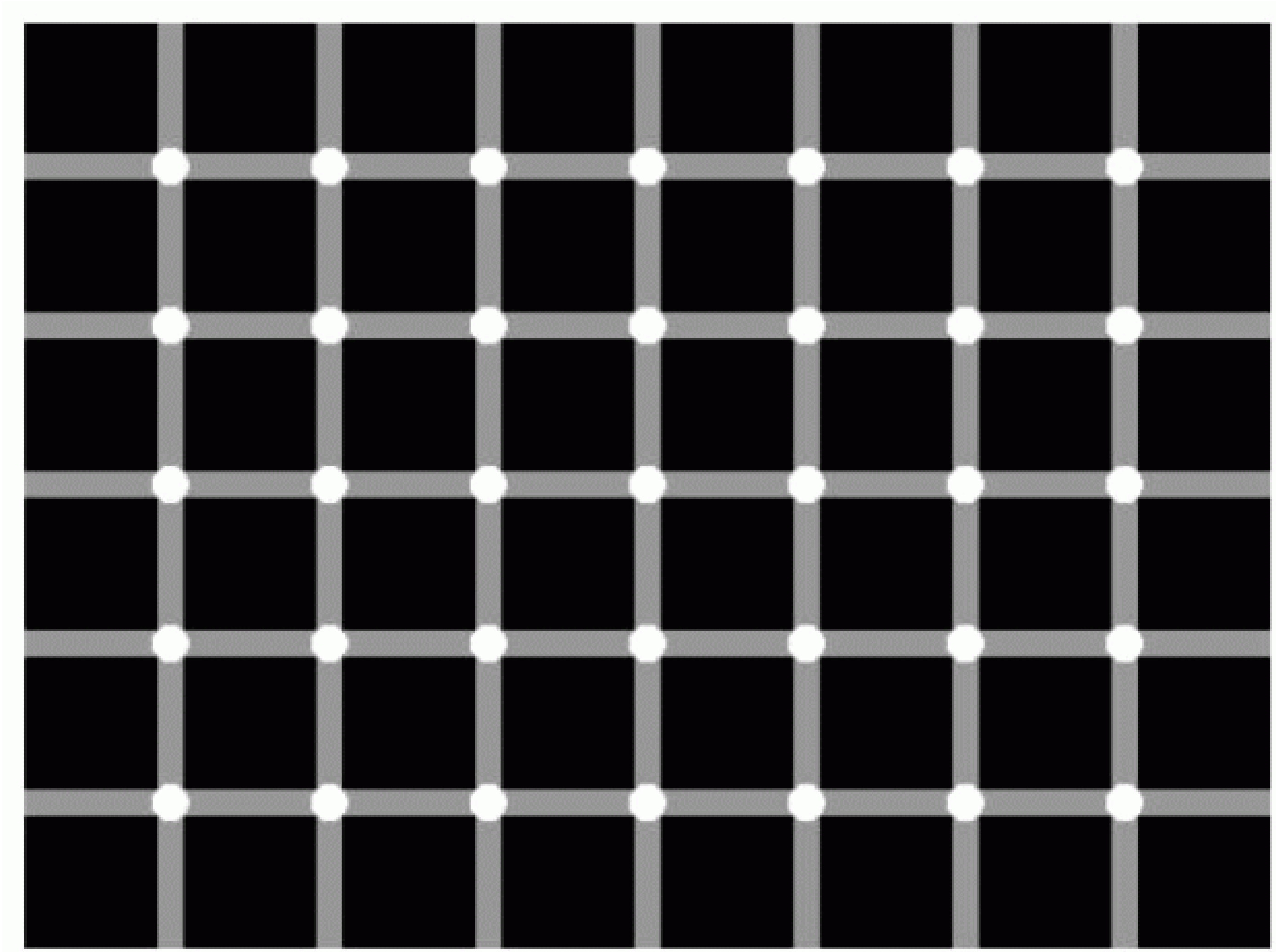
Illusions & Ethical Considerations (Visualizations Can Deceive)

Visualization Can be Deceptive



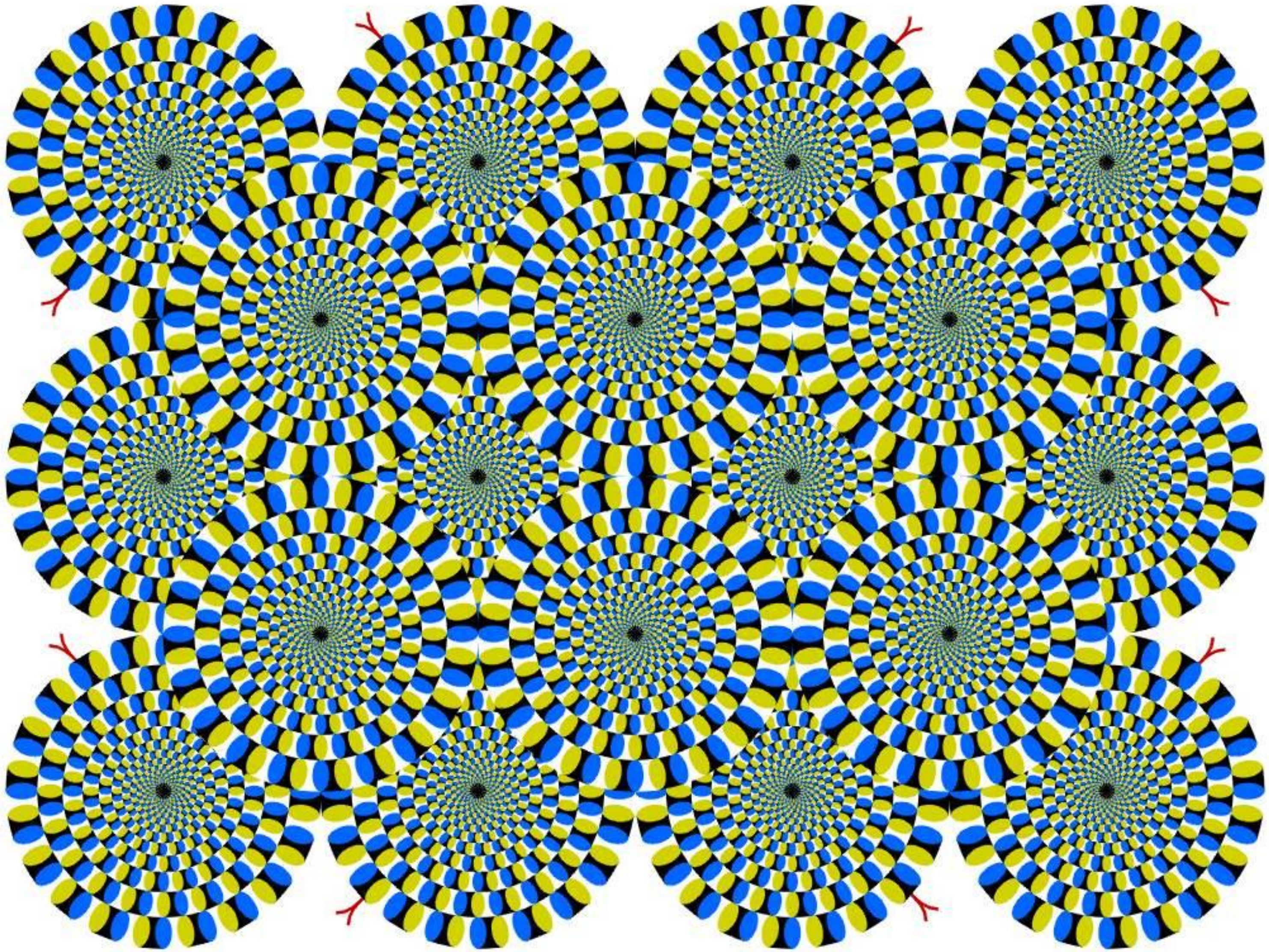
Visualization Can be Deceptive

Visualization Can be Deceptive

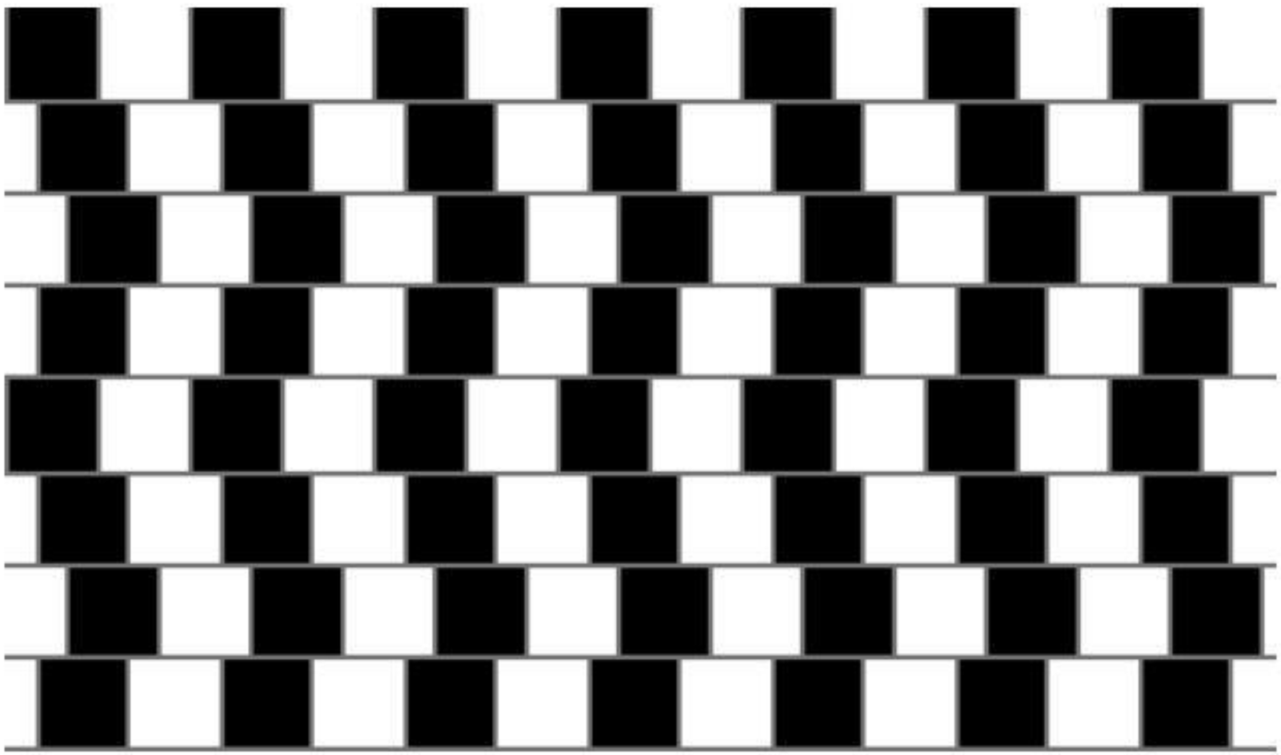


Count the number of black dots

Visualization Can be Deceptive

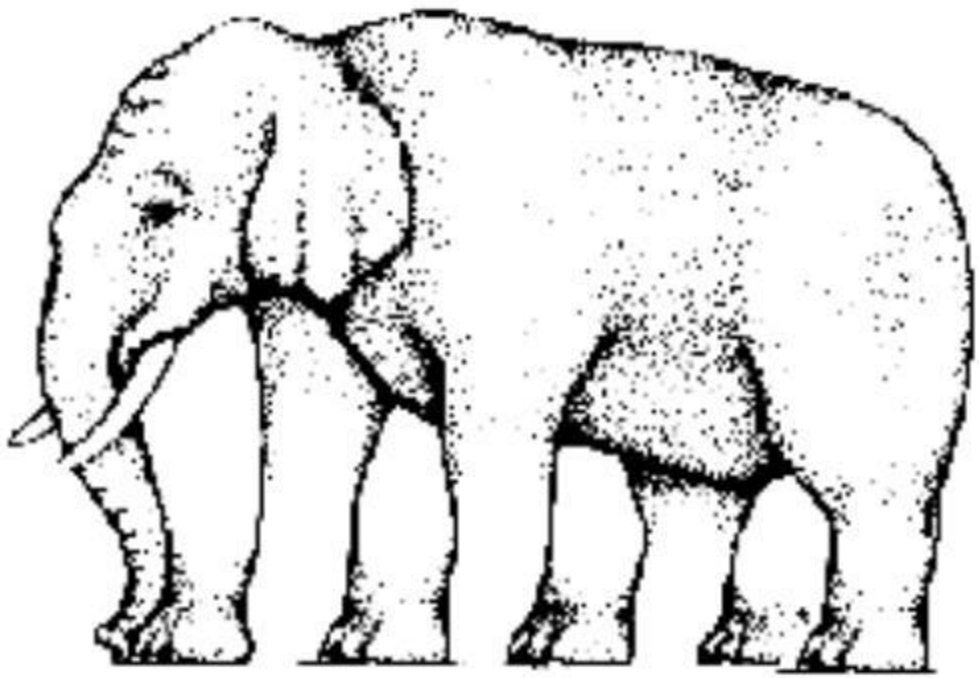


Visualization Can be Deceptive



Are the horizontal lines parallel or do they slope?

Visualization Can be Deceptive

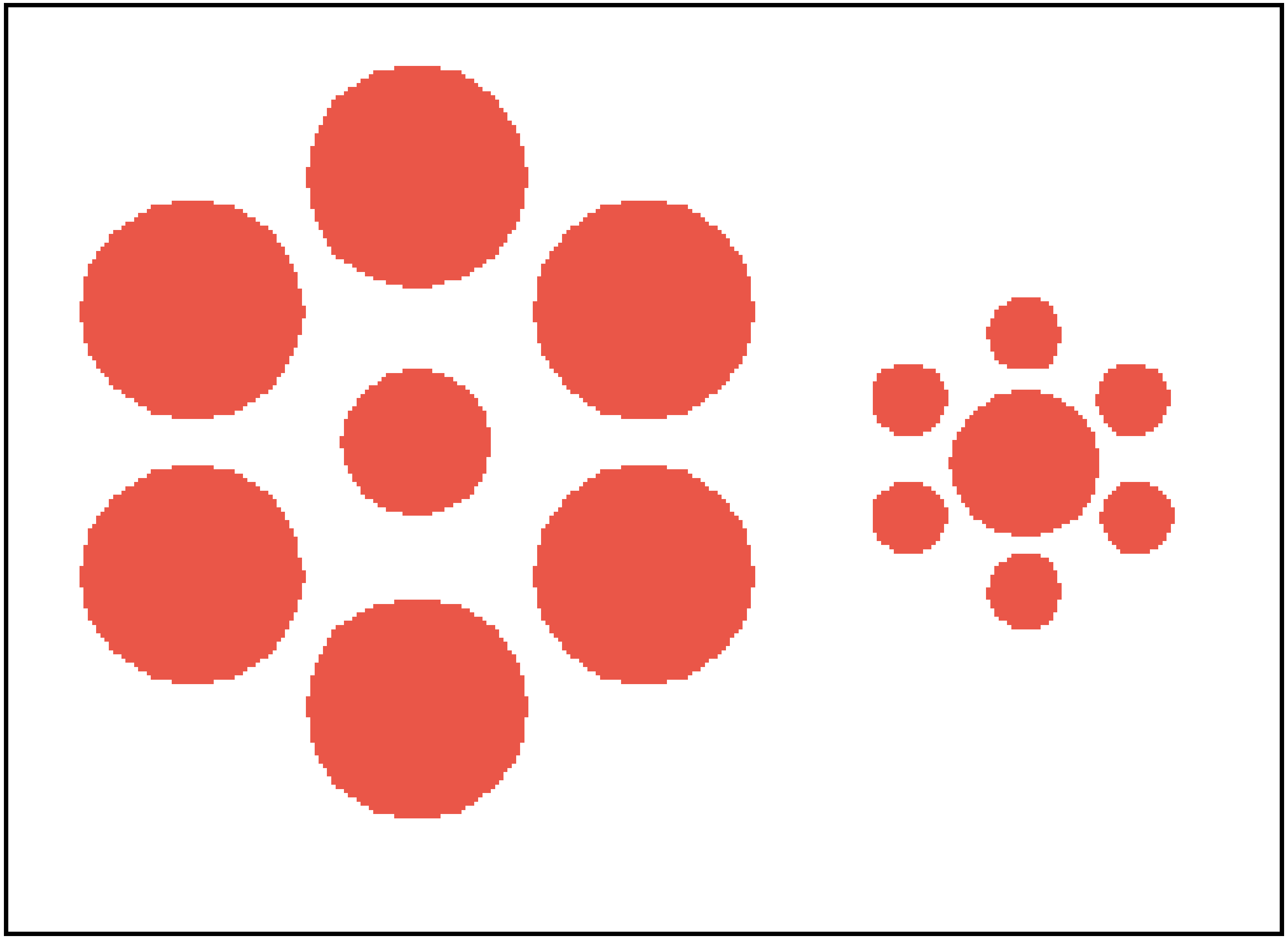


How many legs does this elephant have?

Visualization Can be Deceptive

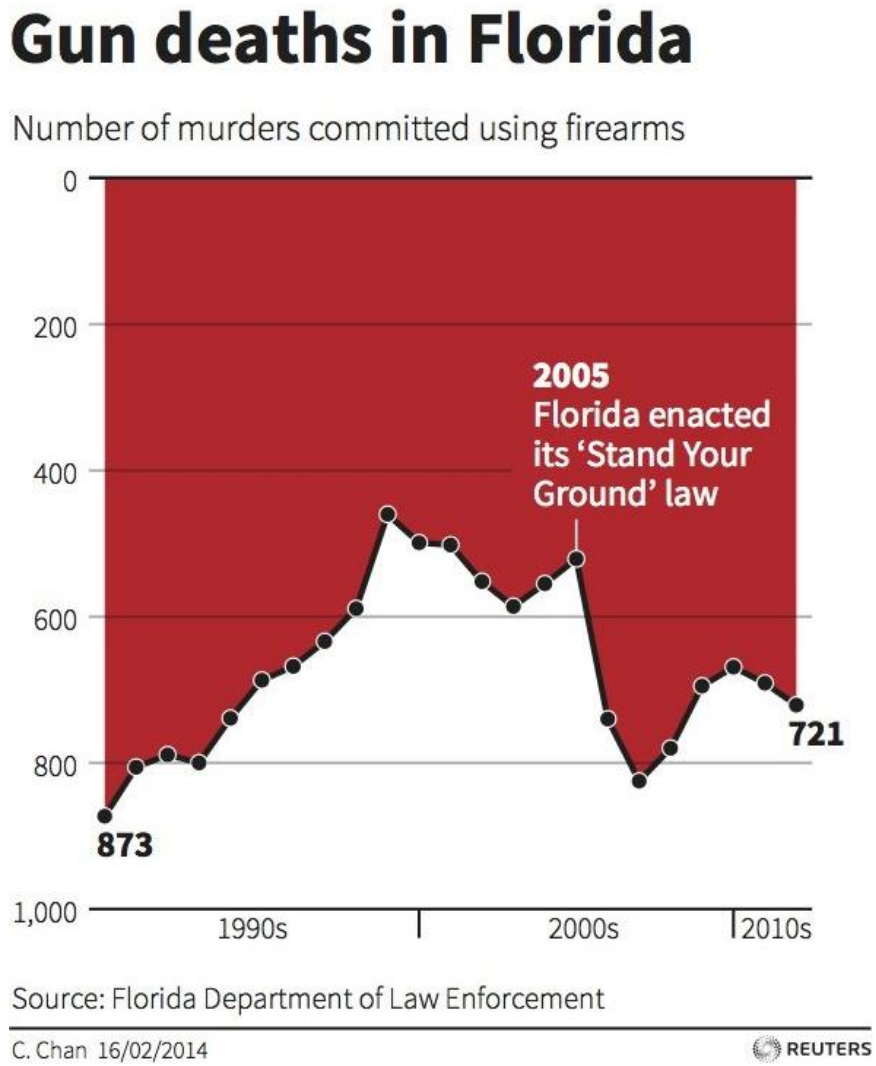
Julian Beever

Visualization Can be Deceptive



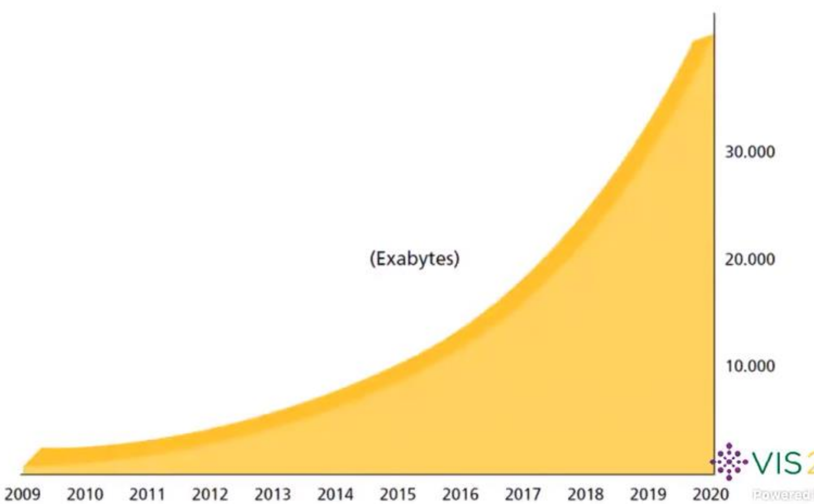
Which circle in the middle is bigger?

Visualization Can be Deceptive

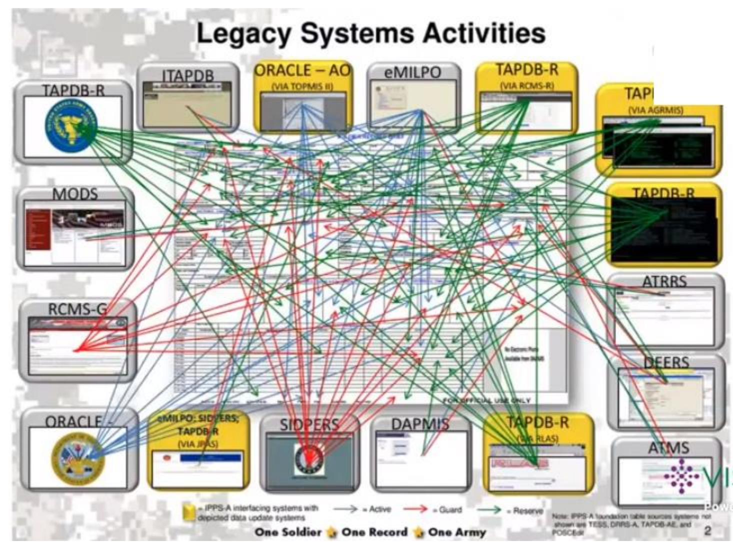


Visualization Can Be BS*

From Michael Correll's alt.vis 2021 talk ([link]())
- Don't relate to the real world
- Don't really help people understand their data
- Don't even have the decency to *lie* to you



"Stock footage chart"

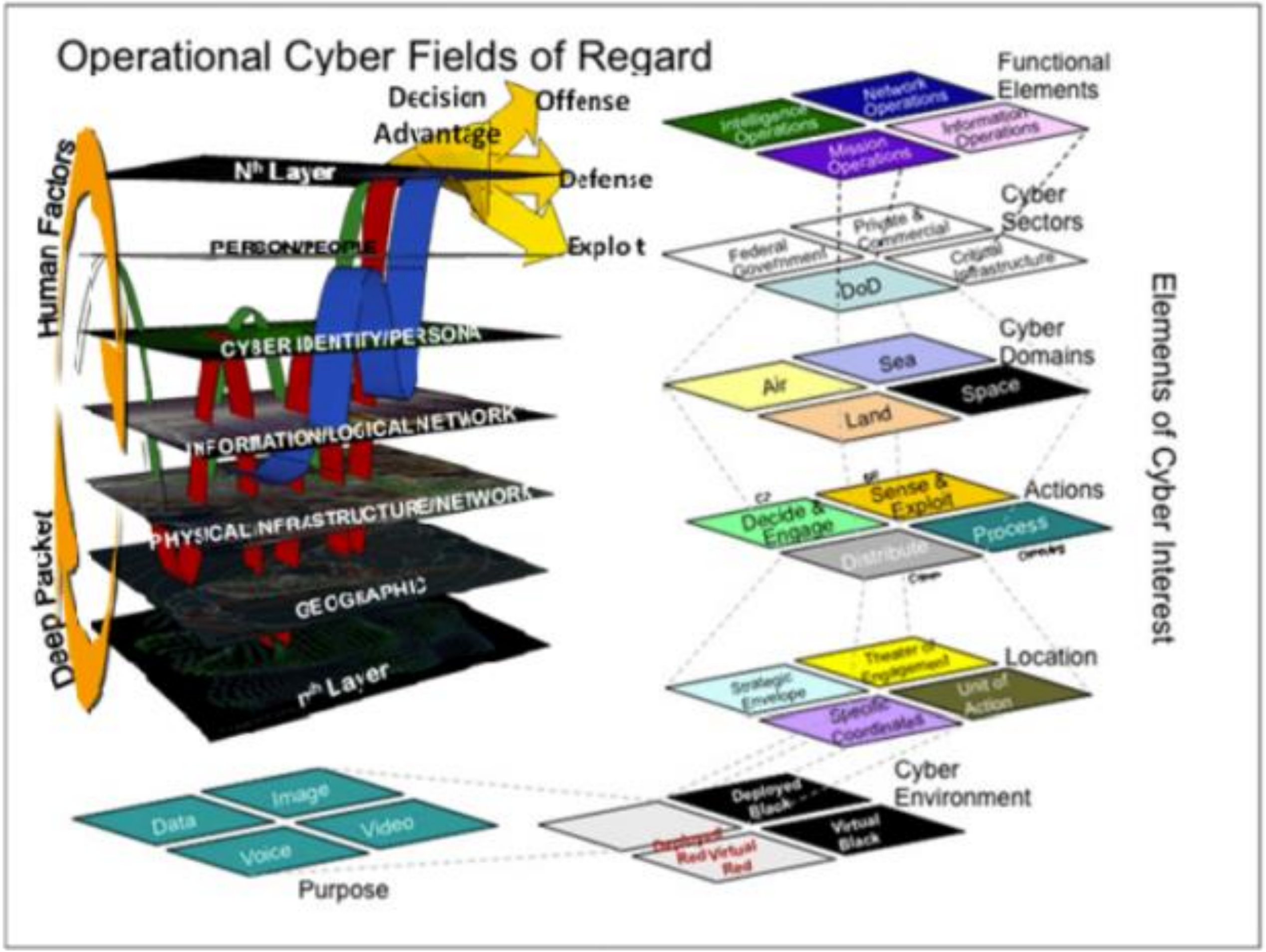


"Novocaine chart"

This stuff is way too complex for you to understand. Aren't you glad there's somebody smart like me taking care of it?

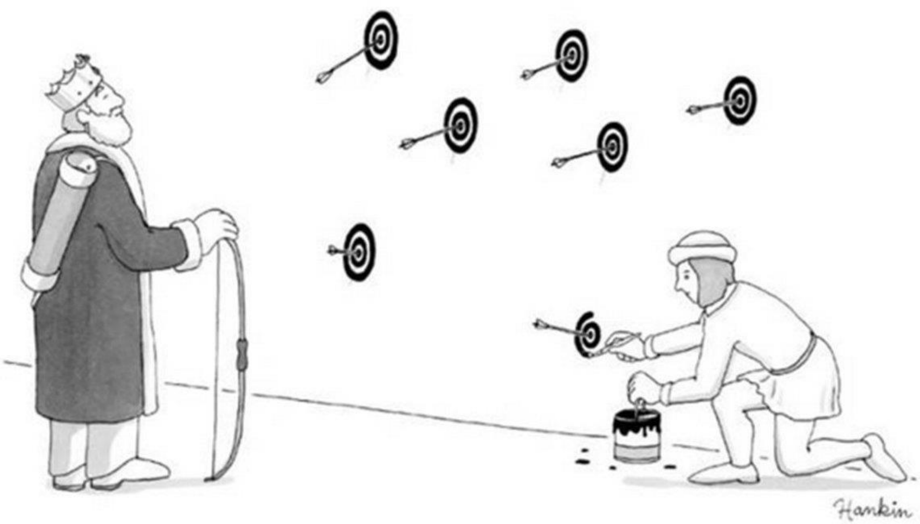
* Bullshit

Visualization Can Be BS

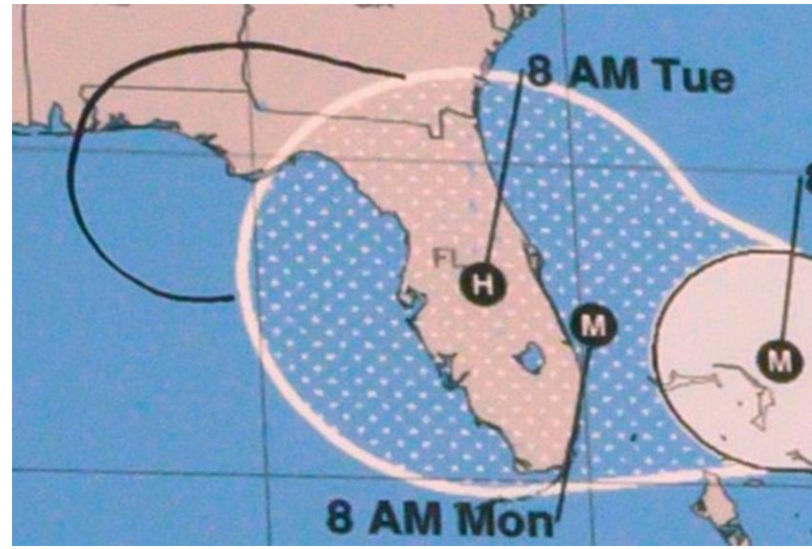


Hwang's @DefenseCharts Twitter account, "dedicated to the presentational aesthetics of the defense-industrial complex"

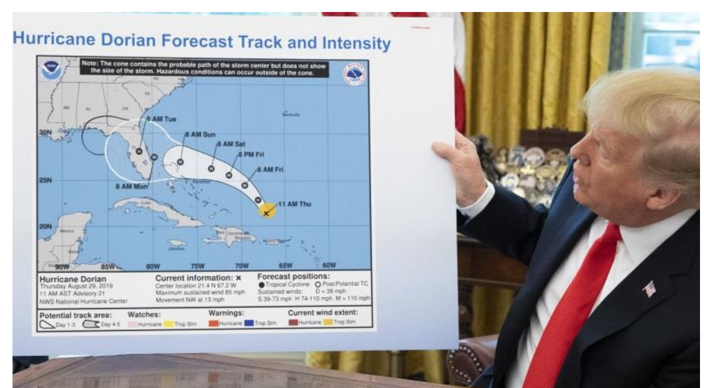
Visualization Can Be BS



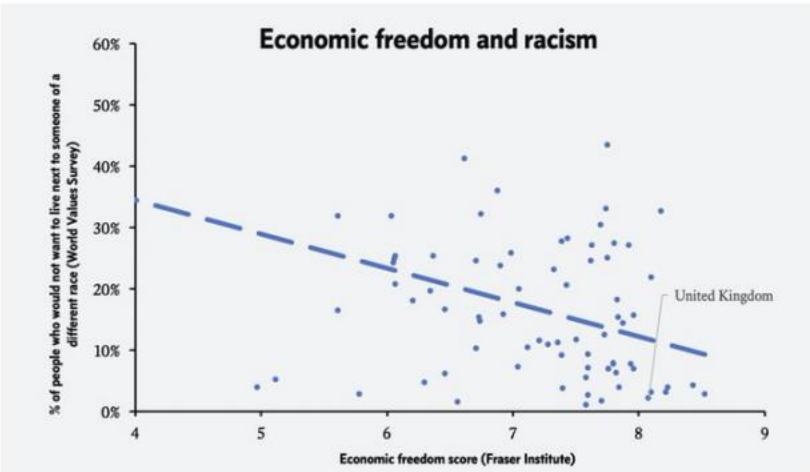
“Texas sharpshooter chart”



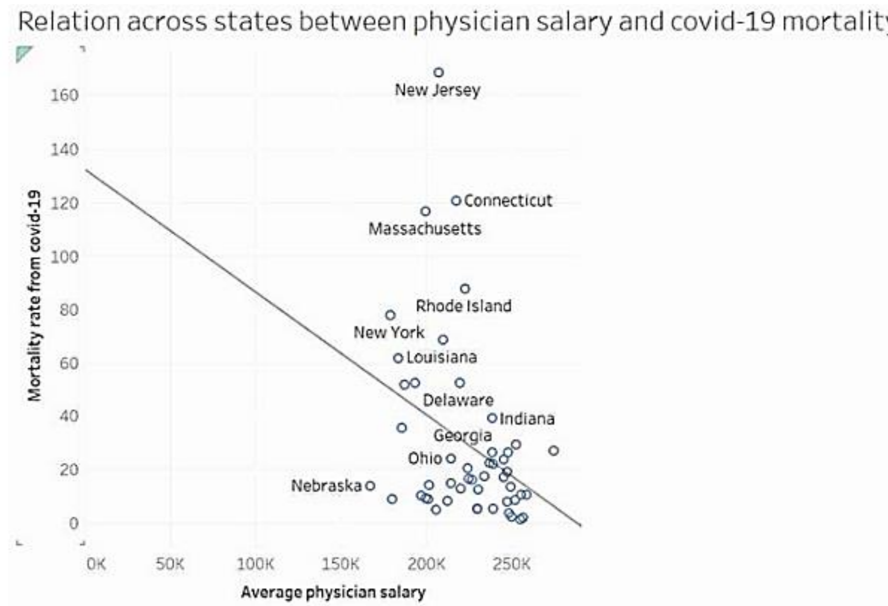
Example: Sharpiegate



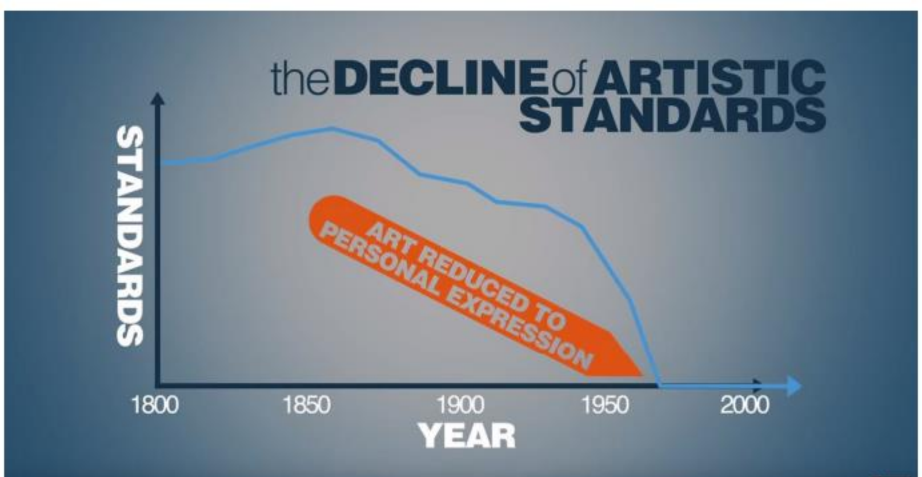
Visualization Can Be BS



To show: "Countries with more economic freedom have less racist attitudes"



To show: "States where physicians are highly paid have lower COVID-19 mortality per capita"



Artificial noise added to make the chart look like there is a complex metric being measured precisely over time (when it is really not)

Detect and Fix Misleading Visualizations

There is an app for that:

- MisVisFix: An Interactive Dashboard for Detecting, Explaining & Correcting Misleading Visualization (Das, Mueller, IEEE VIS 2025)

youtube