CSE 332
Introduction to Visualization

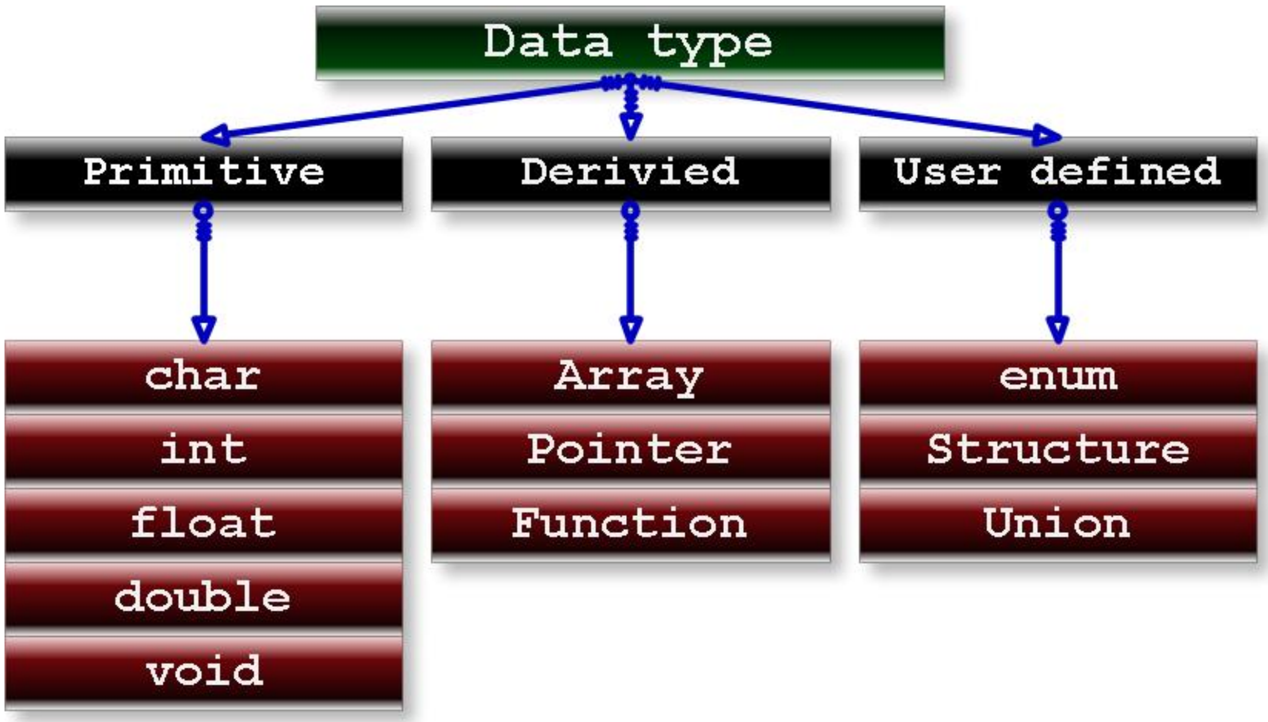
Data Types & basic Applications

Klaus Mueller

Computer Science Department
Stony Brook University

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, data types	
3	Basic tasks	Project 1 out
4	Data preparation and representation	
5	Data reduction, notion of similarity and distance	
6	Dimension reduction	
7	Introduction to D3	Project 2 out
8	Visual perception and cognition	
9	Visual design and aesthetic	
10	Visual analytics tasks	
11	Cluster analysis	
12	High-dimensional data, dimensionality reduction	
13	Visualization of spatial data: volume visualization intro	Project 3 out
14	Introduction to GPU programming	
15	Visualization of spatial data: raycasting, transfer functions	
16	Illumination and isosurface rendering	
17	Midterm	
18	Scientific visualization	
19	Non-photorealistic and illustrative rendering	Project 4 out
20	Midterm discussion	
21	Principles of interaction	
22	Visual analytics and the visual sense making process	
23	Visualization of graphs and hierarchies	
24	Visualization of time-varying and streaming data	Project 5 out
25	Maps	
26	Memorable visualizations, visual embellishments	
27	Evaluation and user studies	
28	Narrative visualization, storytelling, data journalism, XAI	

Data Types Every CS Person Knows



Data Types in Visual Analytics

Numeric

Categorical

Text

Time series

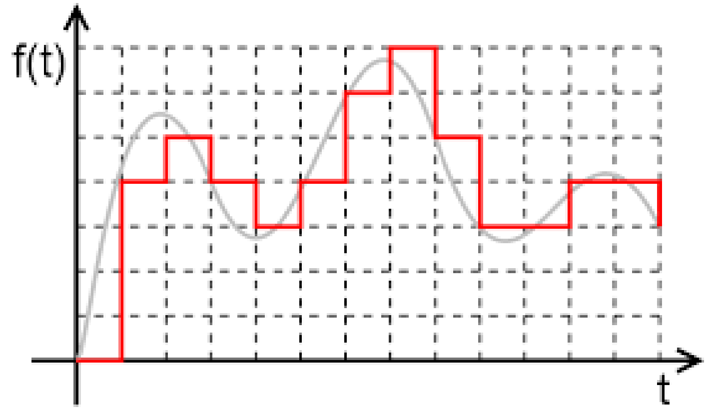
Graphs and networks

Hierarchies

Variables in Statistics

Numeric variables

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



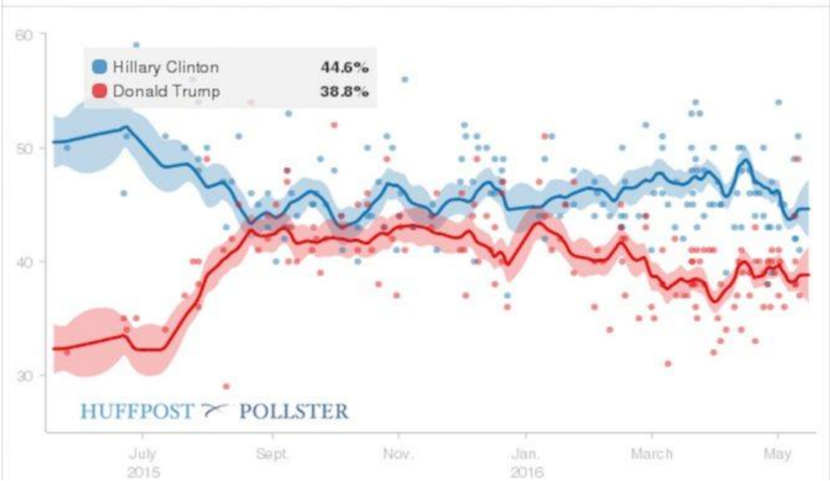
Categorical variables

- 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

Numeric Variables

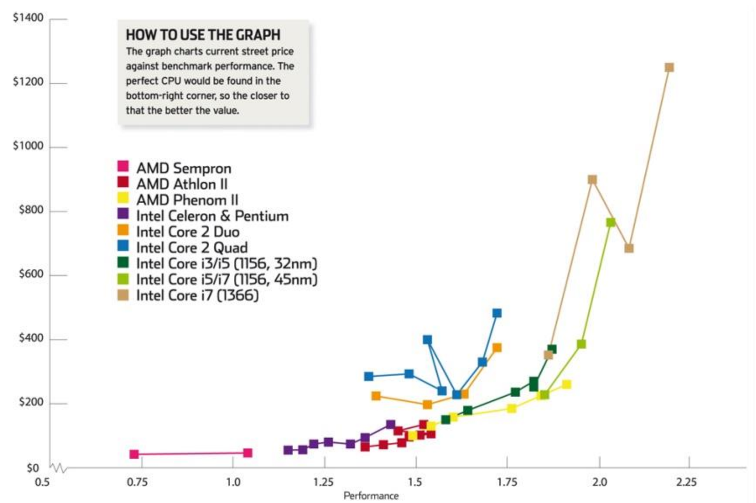
Most often the x-axis is ‘time’

- 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

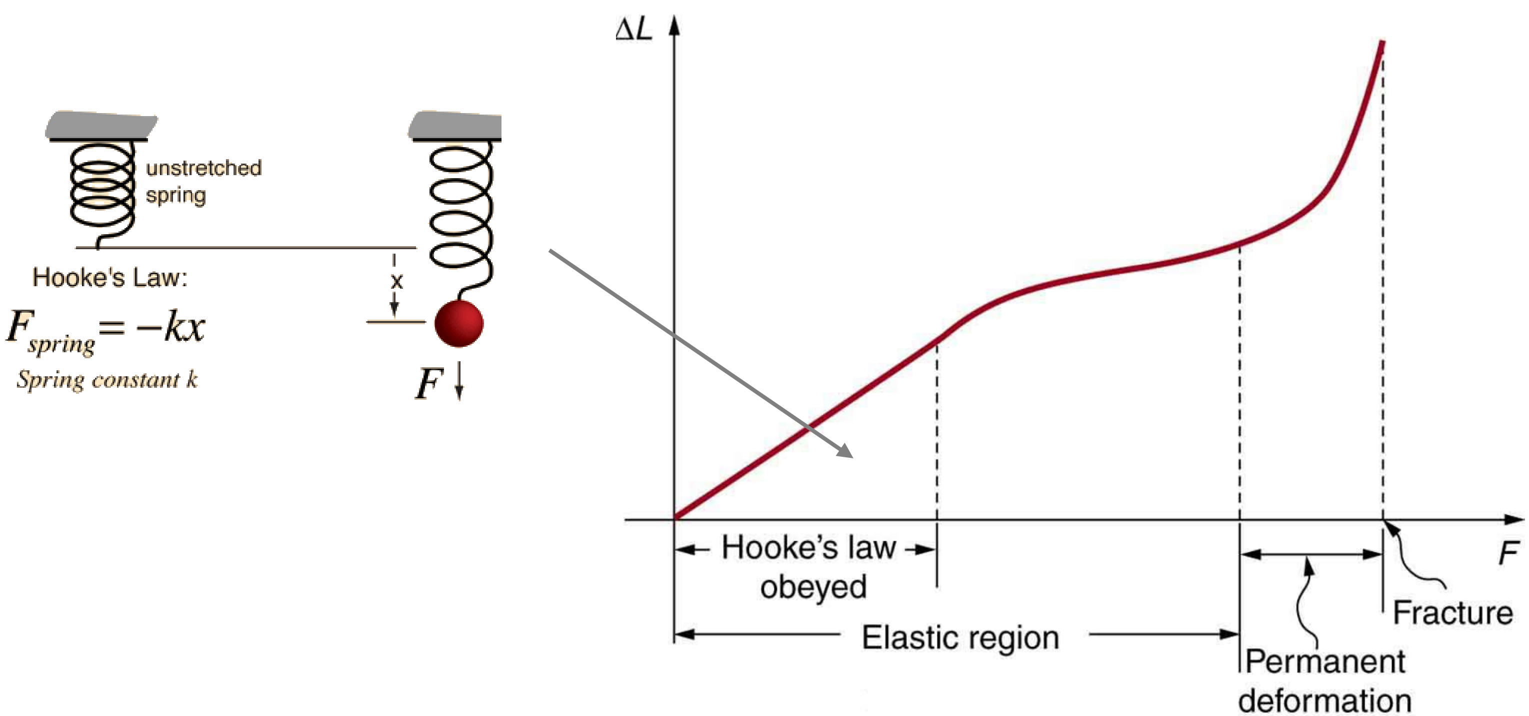
- engineers, statisticians, etc. will be receptive to this idea
- can you think of an example?



Numeric Variables

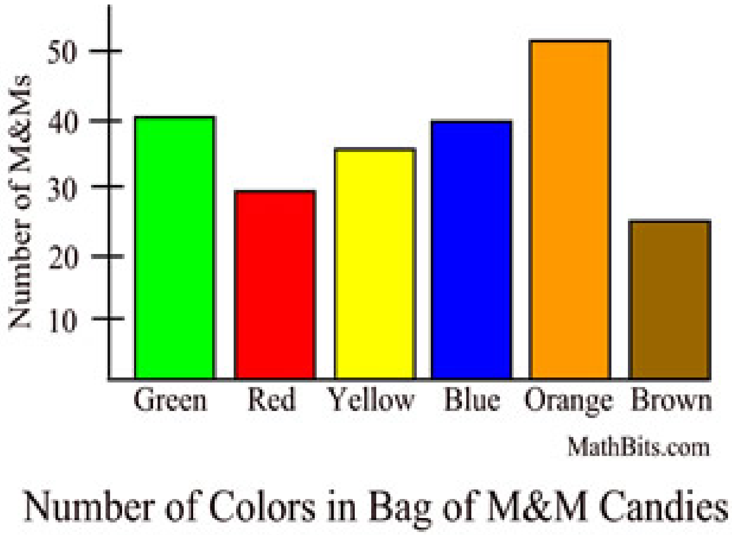
Another plot where 'time' is not the x-axis

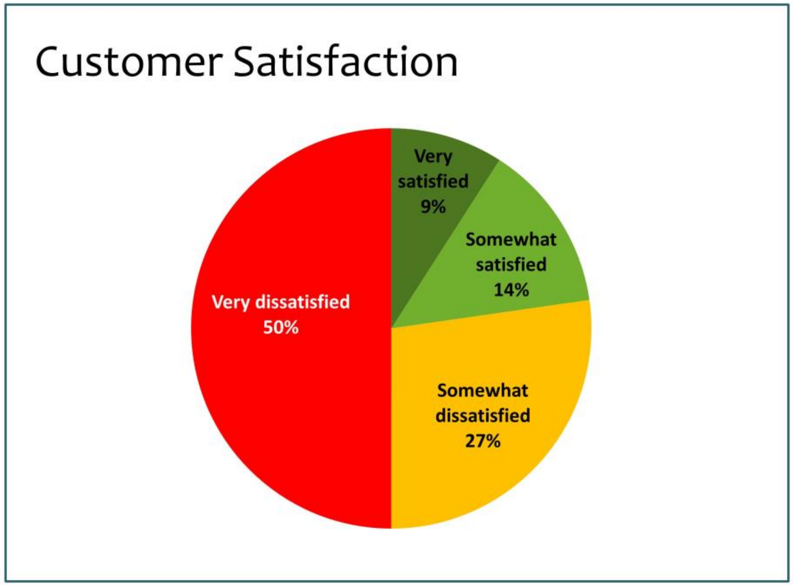
- from the engineering / physics domain
- in some sense, it tells a story



Categorical Variables

Usually plotted as bar charts or pie charts





?? ??

nominal

ordinal

NUMBERS ARE GOOD

But not everything is expressed in numbers

- images
- video
- text
- web logs
- …

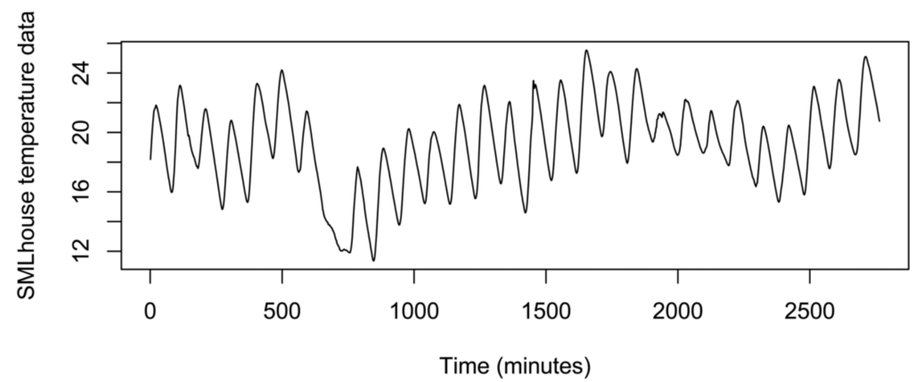
Need to do feature analysis to turn these abstract things into numbers

- 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

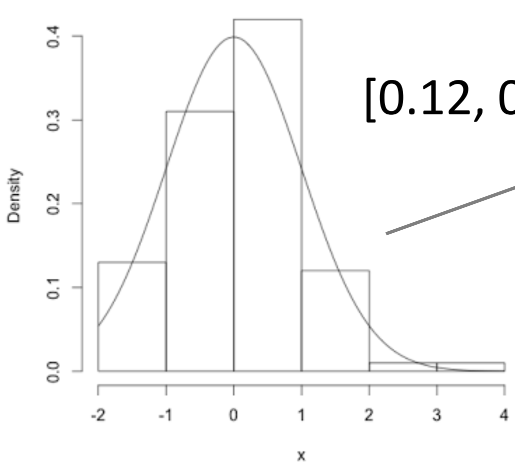
Characteristics

- often large scale
- time series

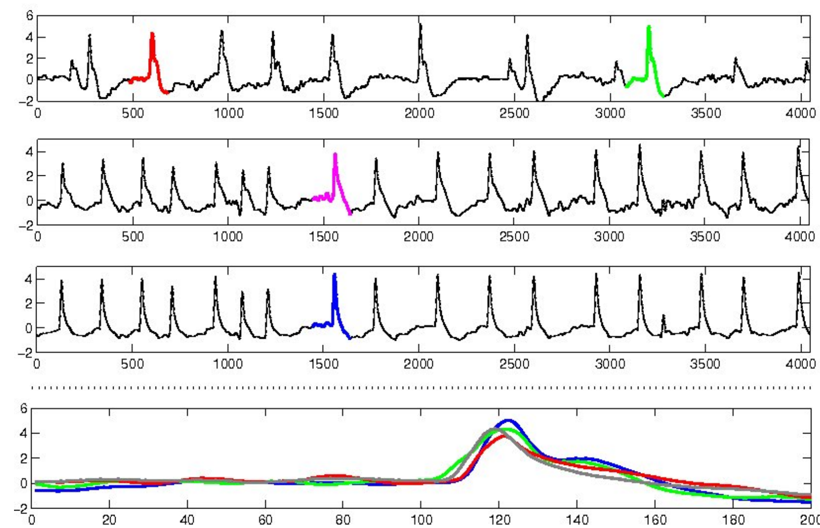


Feature Analysis

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



[0.12, 0.3, 0.41, 0.12, 0.02]



Motif discovery

Image Data

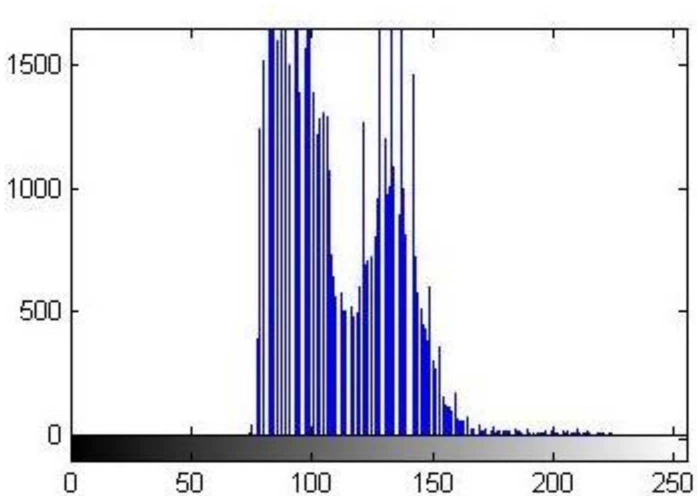
histograms

Characteristics

- array of pixels

Feature Analysis

- example: value histograms
- encode into a 256-D vector



[0, 0, 0, …., 10, …, 1200, …..]

Video data

Characteristics

- essentially a time series of images

Feature Analysis

- many of the image techniques apply but 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.
  - [Jim]$_{\text{Person}}$ bought [300 shares] $_{\text{Quantity}}$ of [Acme Corp.]$_{\text{Organiz.}}$ in [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

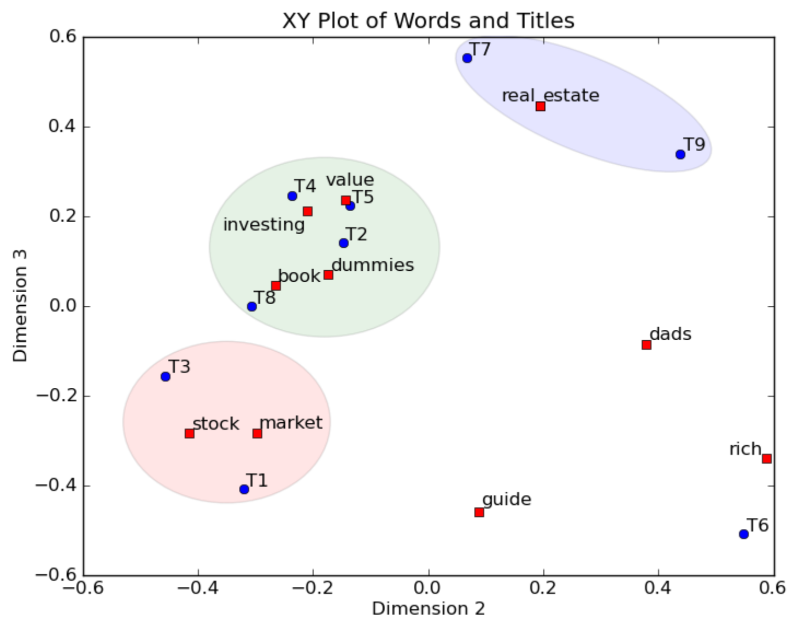
Create a term-document matrix

- 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

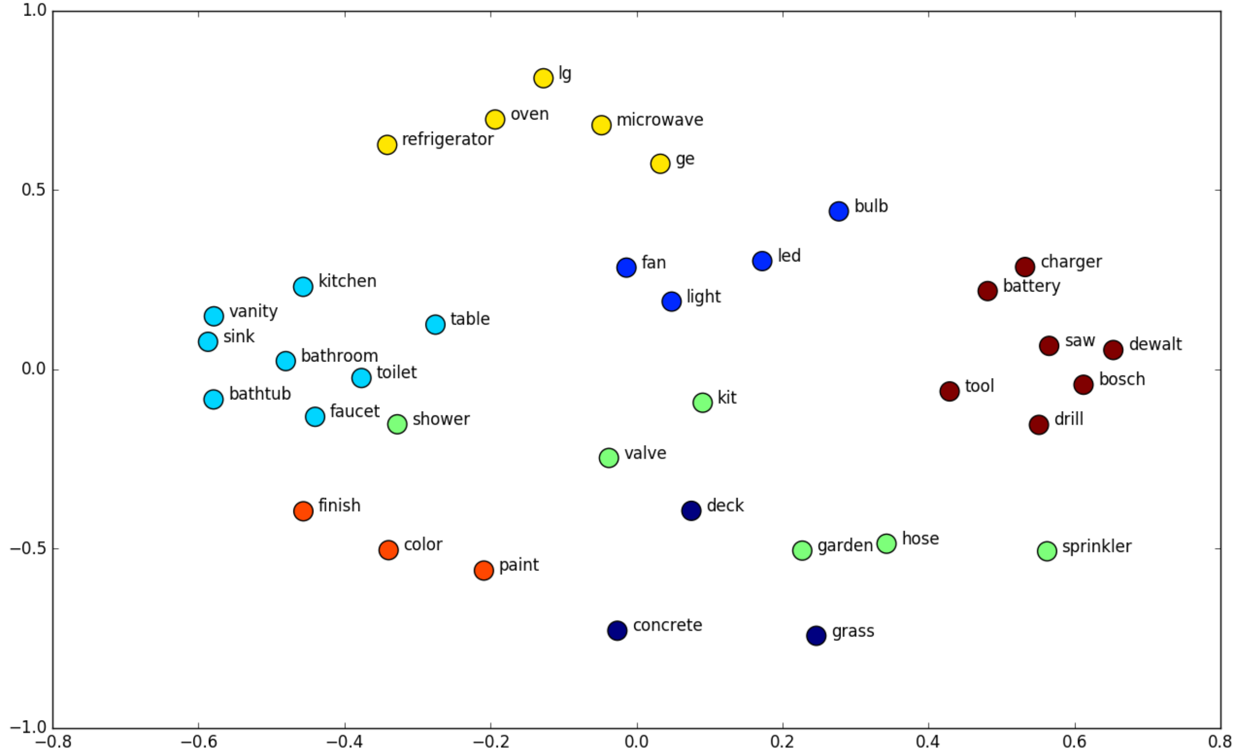


Word/document cluster

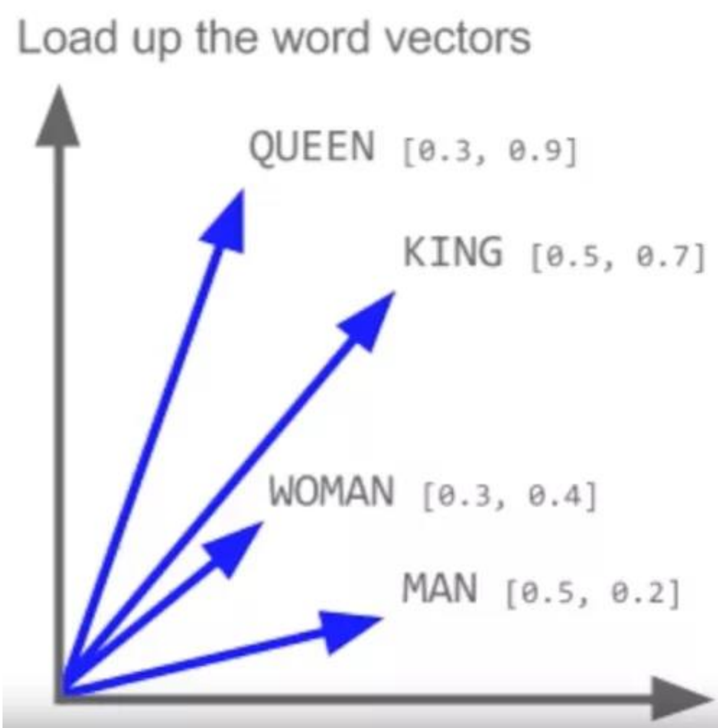
Word Embedding

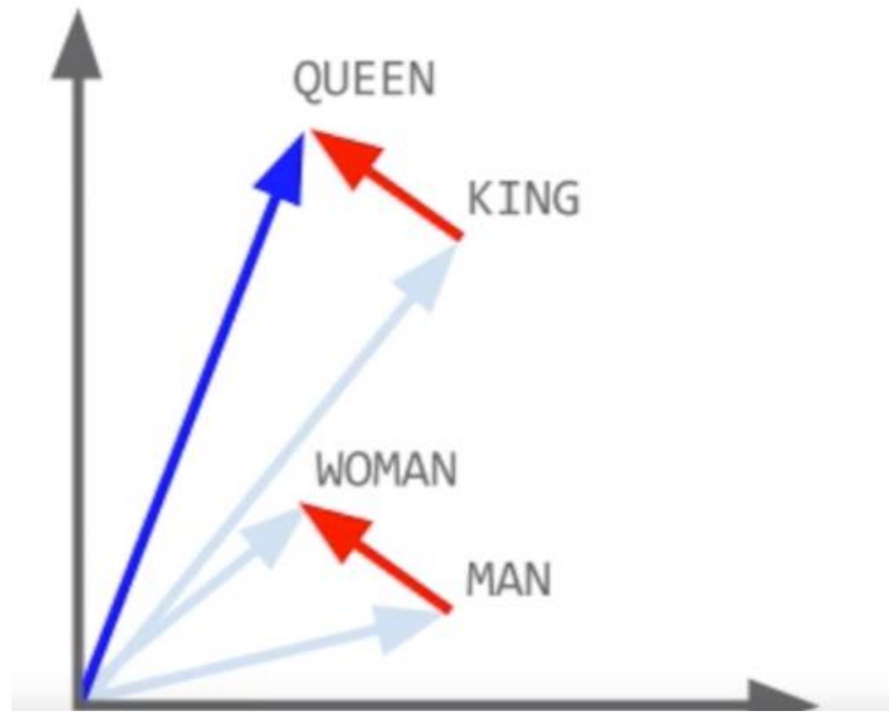
Train a shallow neural network (NN) on a corpus of text

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



Word Embedding Algebra





gender = WOMAN – MAN
QUEEN = KING + gender

QUEEN = KING – MAN + WOMAN

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

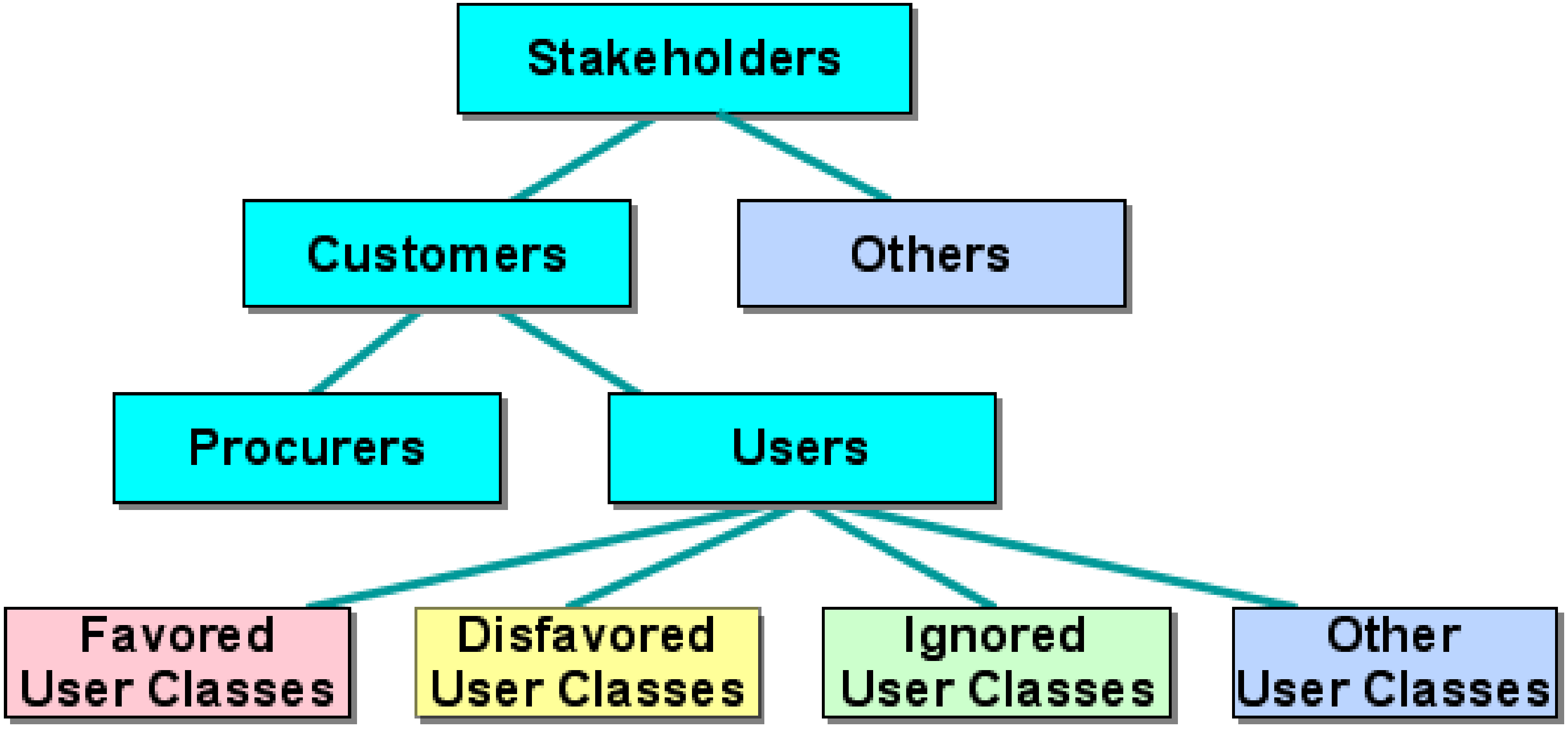
Network traffic

- 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

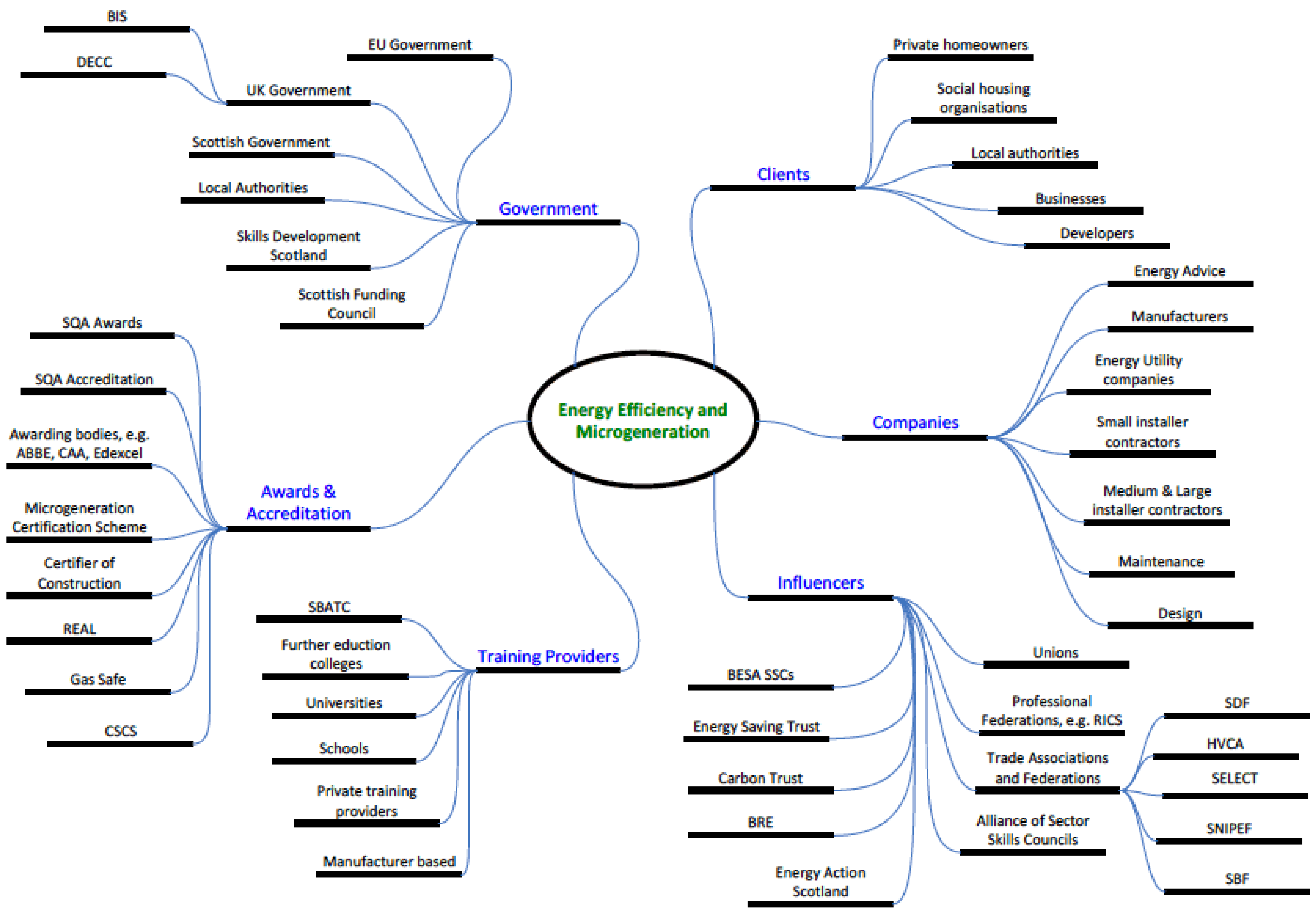
Let's Look at Some Essential Graphical Representations

And Do Some Advertising for D3

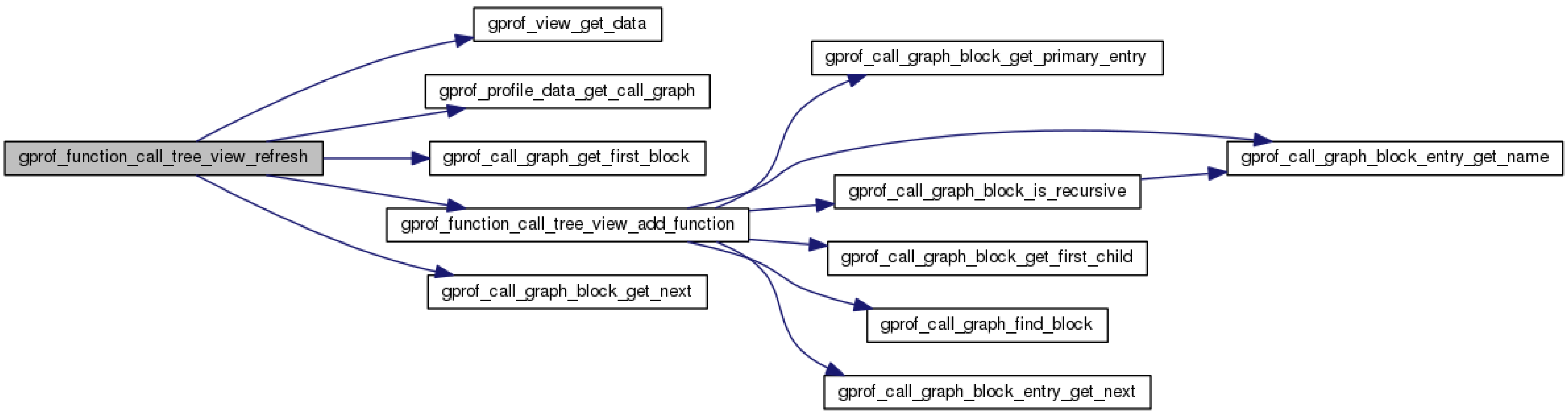
Stakeholder Hierarchy

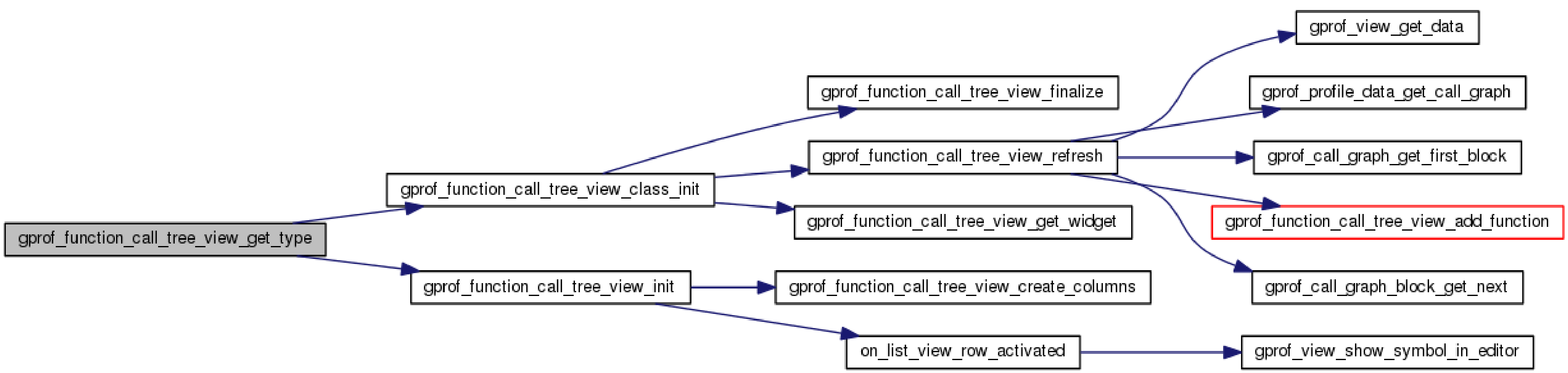


More Complex Stakeholder Hierarchy



Function Call Tree





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

http://animateddata.co.uk/lab/d3-tree/

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

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

Sunburst

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

https://www.jasondavies.com/coffee-wheel/

http://bl.ocks.org/kerryrodden/7090426

Bubble Charts

No hierarchy information, just quantities

http://bl.ocks.org/mbostock/4063269

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

http://bl.ocks.org/mbostock/4062006

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

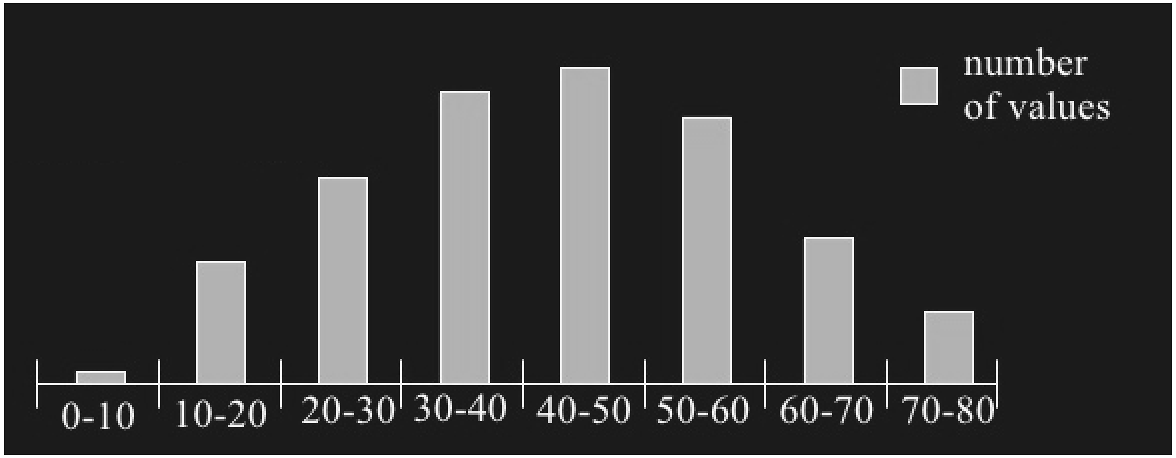
http://bl.ocks.org/mbostock/4060366

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

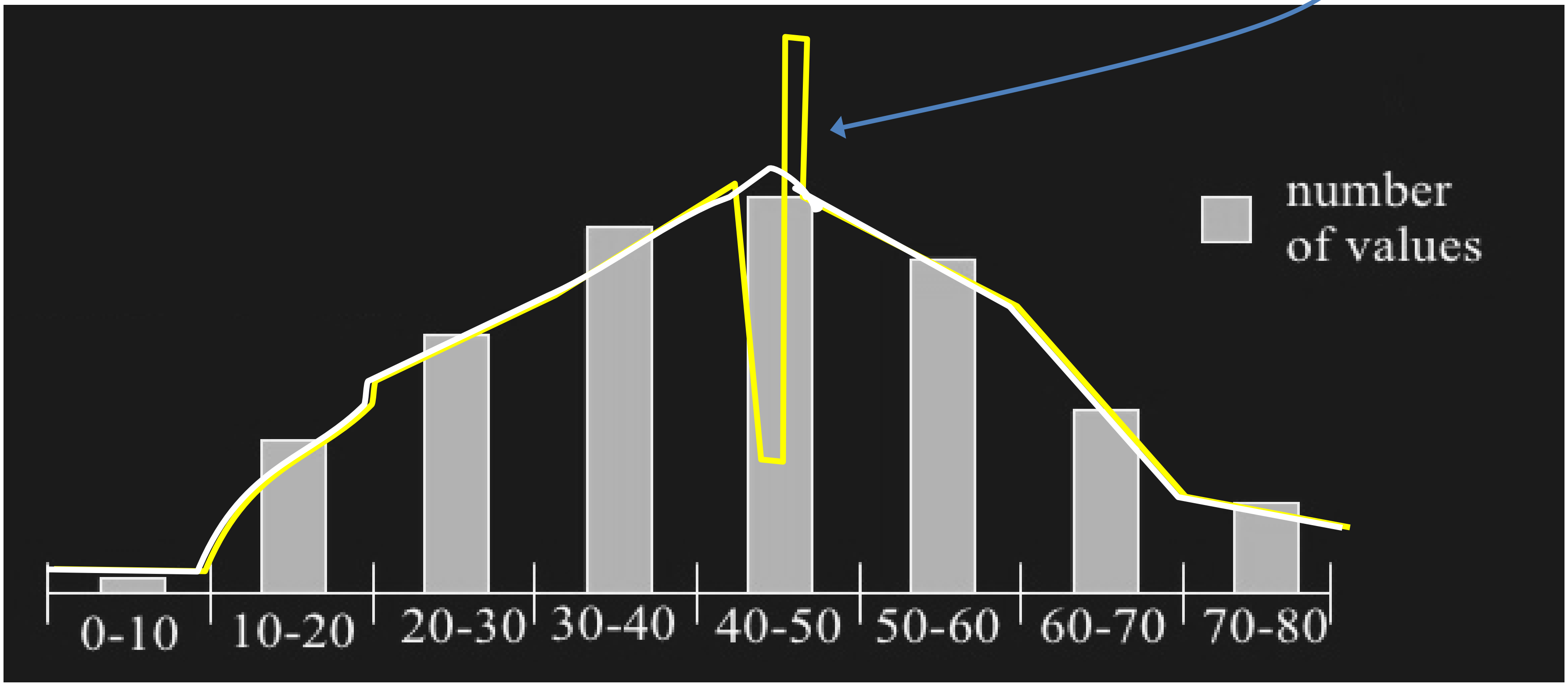


- what is lost here?

Problem With Equi-Width Histogram

Age ranges of customers could be unevenly distributed within a bin

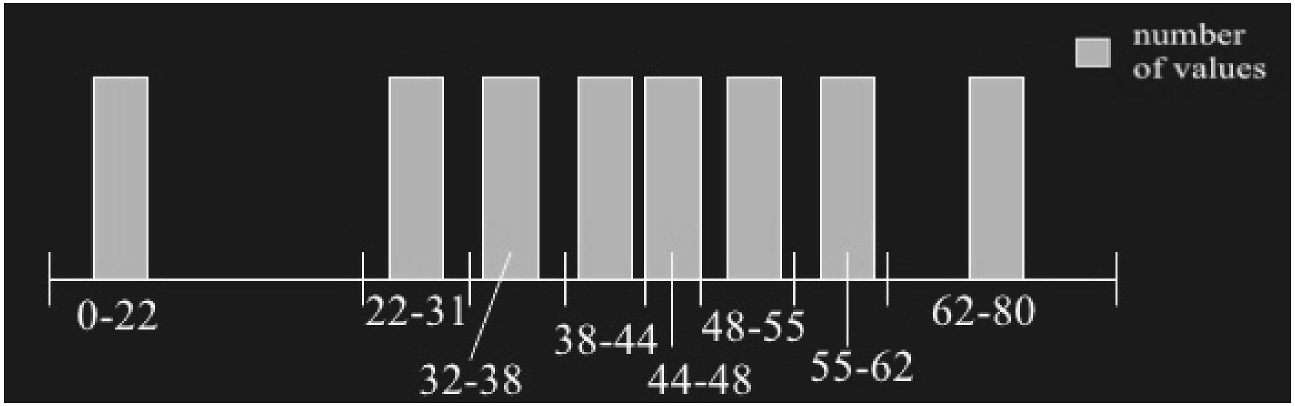
- this could be an interesting anomaly



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:

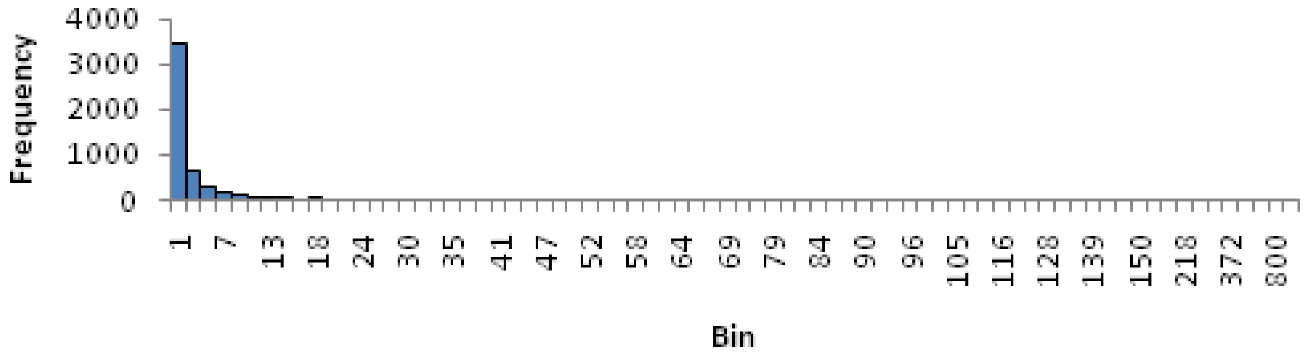


- what is the disadvantage here?
- extra storage needed: must store the start/end value for each bin

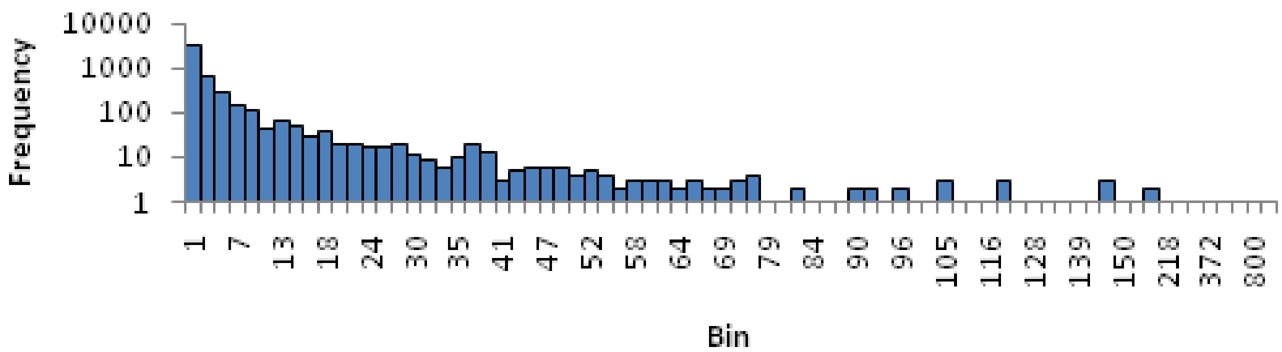
Numeric to Categorical Data: Discretization (3)

Solution 3:

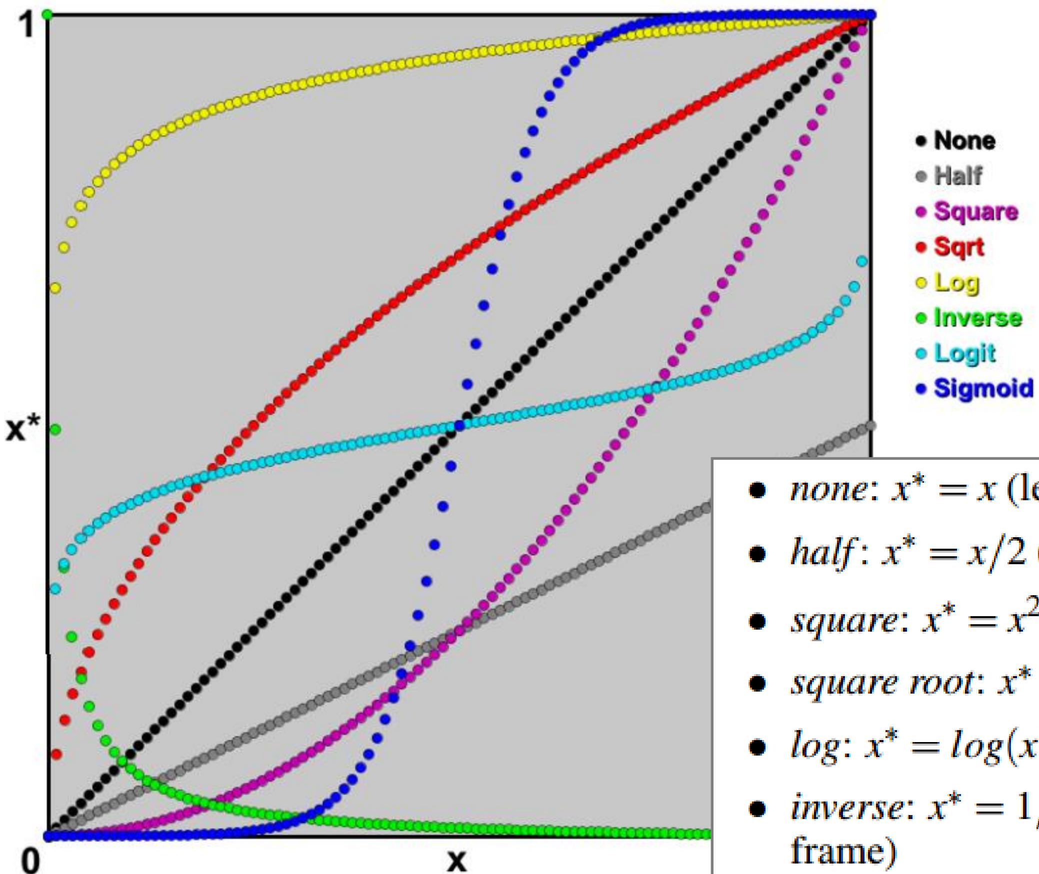
- what if all the bars have seemingly height
- or are dominated by one large peak



- switch to log scaling of the y-value



Other Transformations



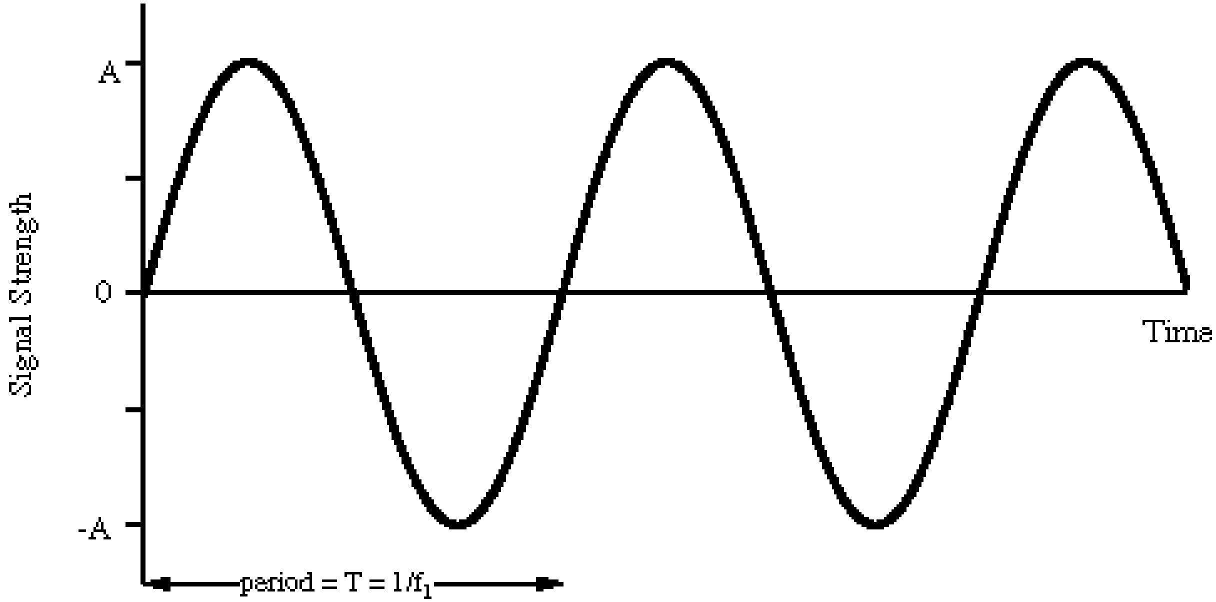
- *none*: $x^* = x$ (leaves points unchanged)
- *half*: $x^* = x/2$ (squeezes all points together)
- *square*: $x^* = x^2$ (pulls points toward left of frame)
- *square root*: $x^* = \sqrt{x}$ (mildly pulls points toward right of frame)
- *log*: $x^* = log(x)$ (strongly pulls points toward right of frame)
- *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

Continuous to Discrete

Why discrete?

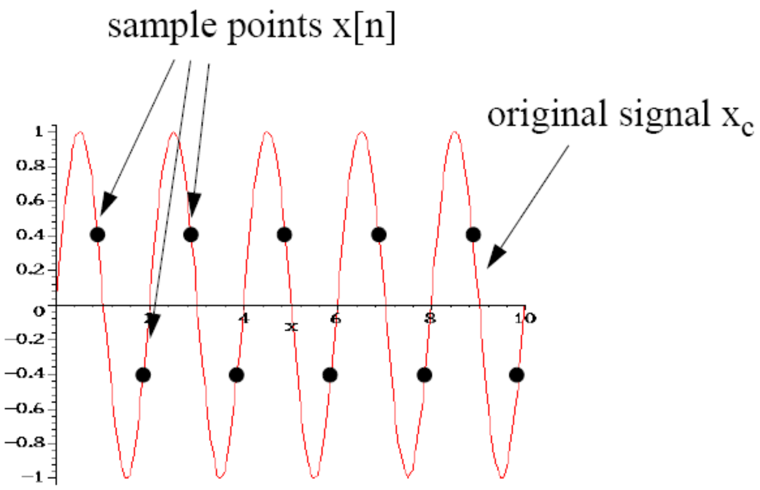
- because we can't store continuous data
- we can only store samples of the continuous data
- how many samples do we need?
- also keep this in mind for data reduction



Continuous to Discrete

Why discrete?

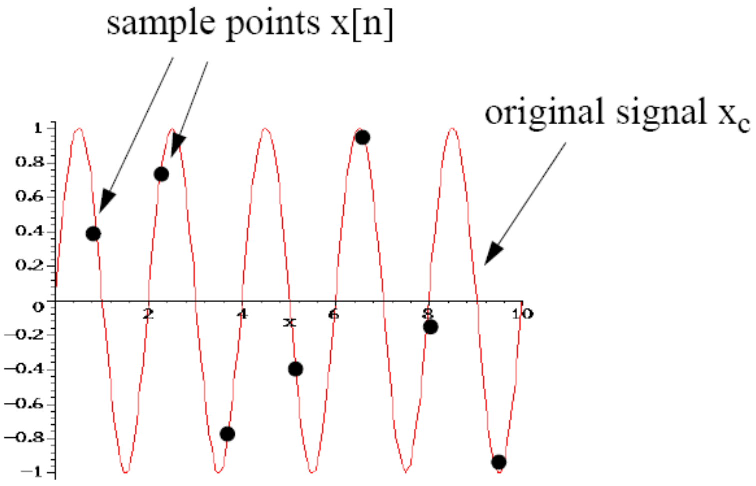
- because we can't store continuous data
- we can only store samples of the continuous data
- how many samples do we need?



Continuous to Discrete

Why discrete?

- because we can't store continuous data
- we can only store samples of the continuous data
- how many samples do we need?



Continuous to Discrete

Why discrete?

- because we can't store continuous data
- we can only store samples of the continuous data
- how many samples do we need?

We need a certain number of samples to represent a continuous phenomenon

- twice as many samples as the highest frequency in the signal
- called the *Nyquist frequency*
- else we get *aliasing*

Practical Implications

Ever tried to reduce the size of an image and you got this?

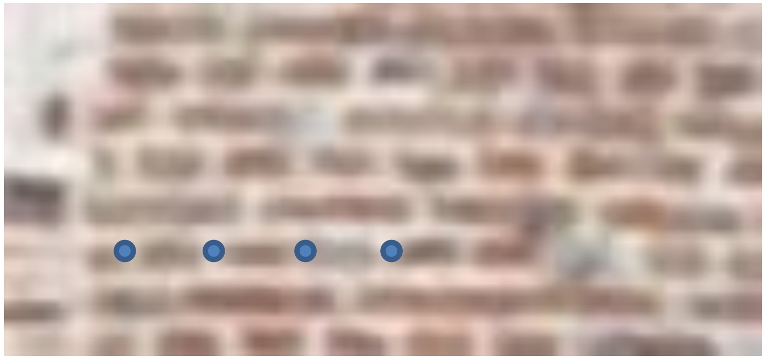
This is aliasing

Practical Implications

But what you really wanted is this:

This is *anti-aliasing*

Why Is This Happening?



The smaller image resolution cannot represent the image detail captured at the higher resolution

- skipping this small detail leads to these undesired artifacts

What Is Anti-Aliasing

Procedure

- either sample at a higher rate
- or smooth the signal before sampling it
- the latter is called *filtering*

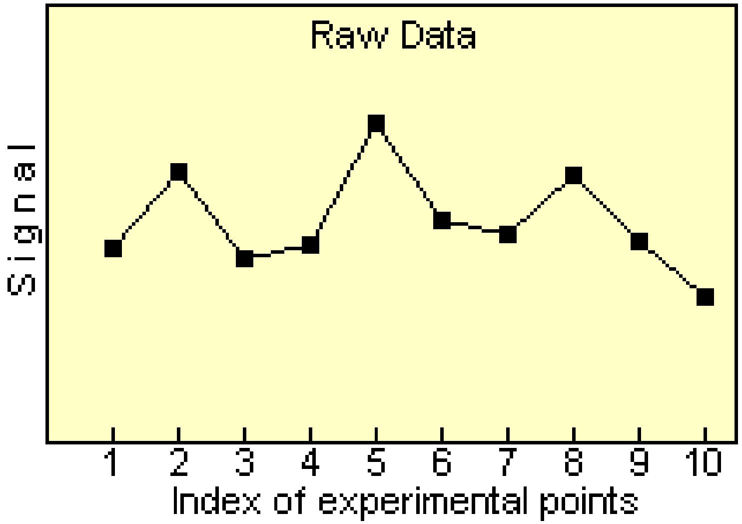
Anti-Aliasing Via Smoothing

Anti-Aliasing Via Smoothing

What is Smoothing?

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



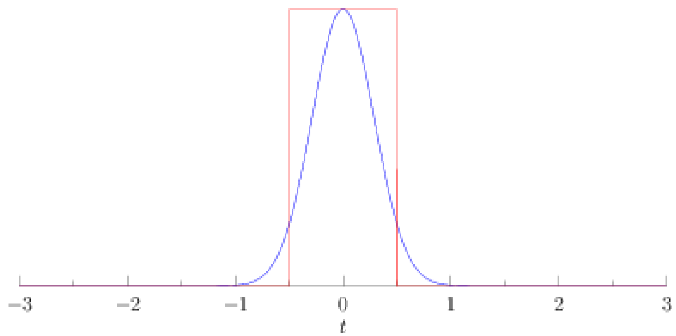
Filters

What is the filter we just used called?

- it's called a *box filter*

There are other filters

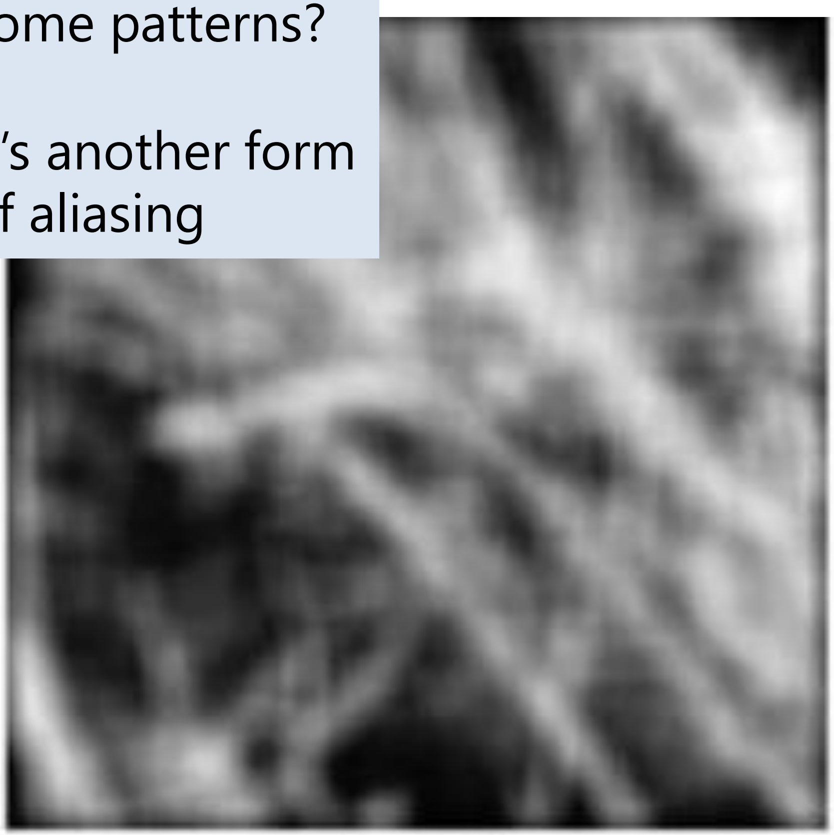
- for example, Gaussian filter
- yields a smoother result
- box filtering is simplest

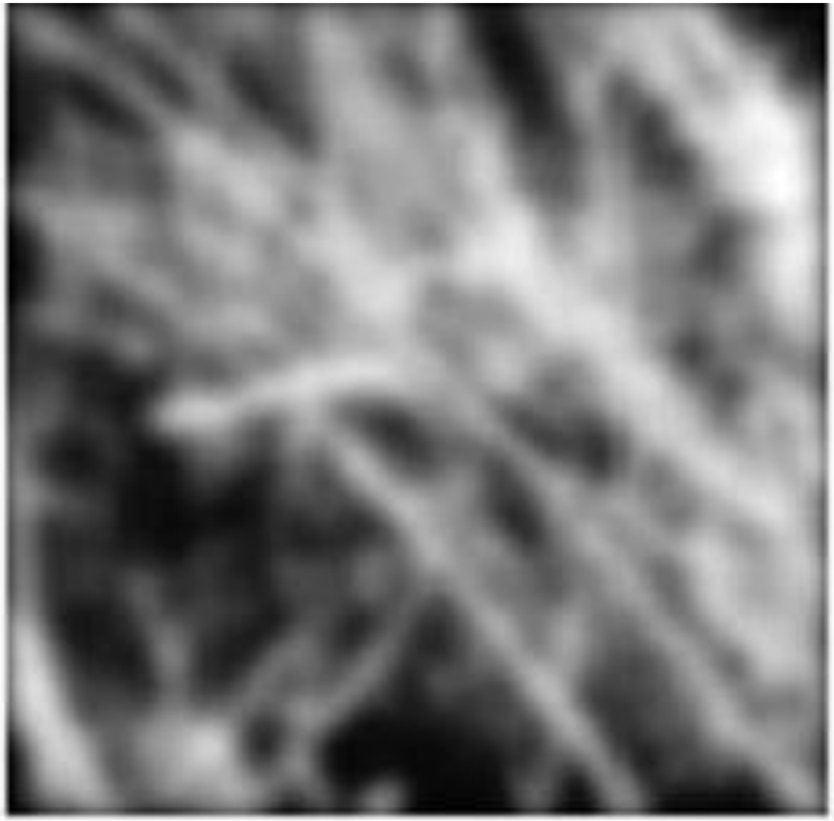


Box Filter vs. Gaussian Filter

Can you see some patterns?

It's another form of aliasing





2D box

2D Gaussian

The Solution

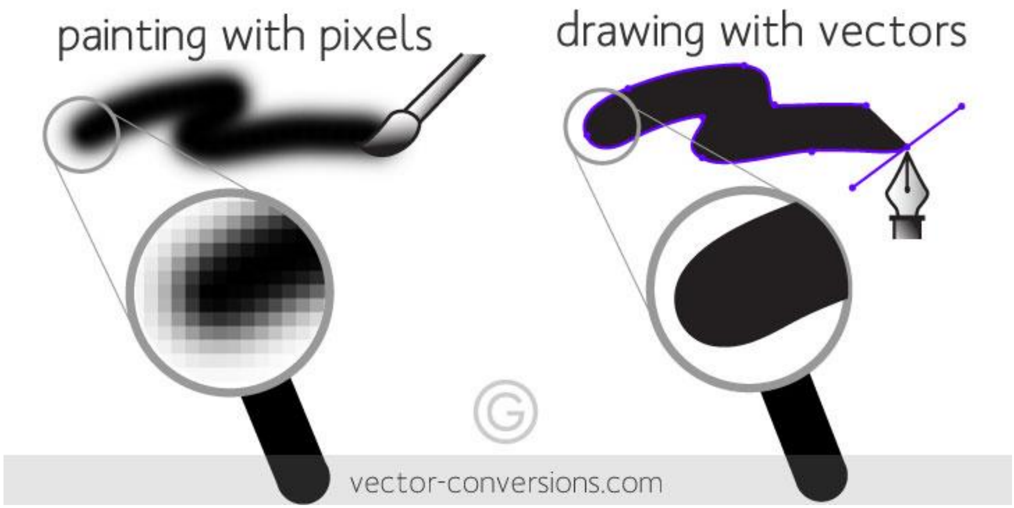
What's the underlying problem?

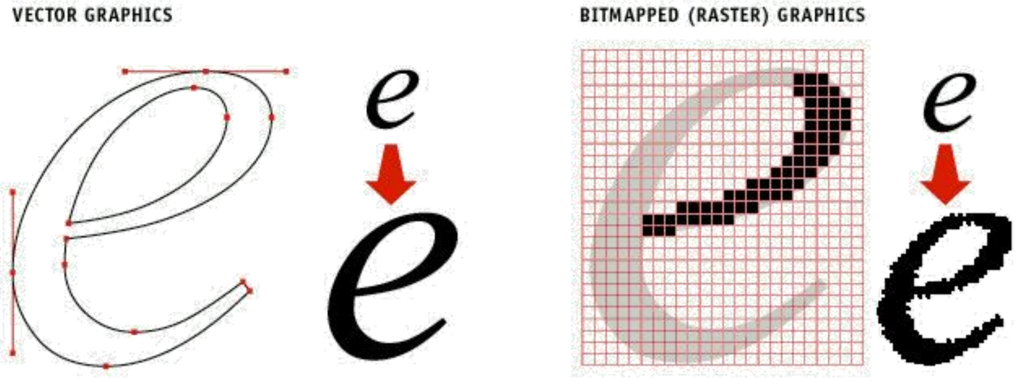
- detail can't be refined upon zoom
- can just be replicated or blurred

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

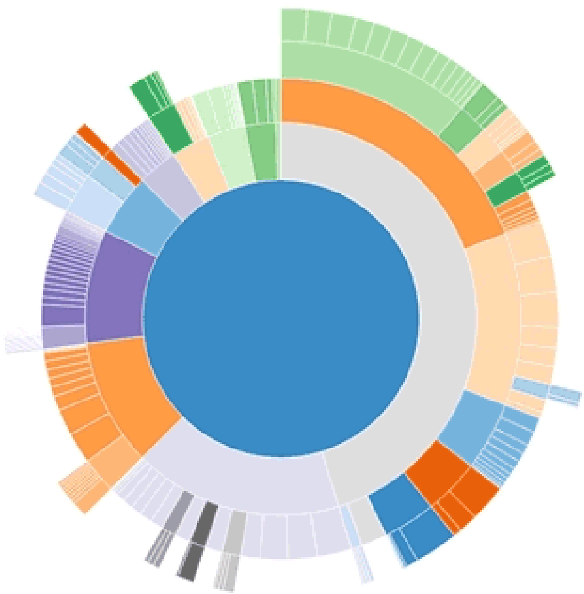
Vector graphics tends to have an “cartoonish” look

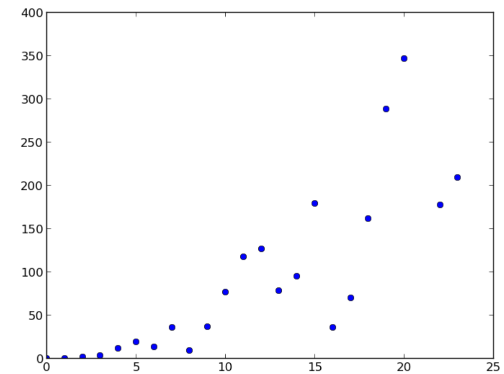
raster graphics

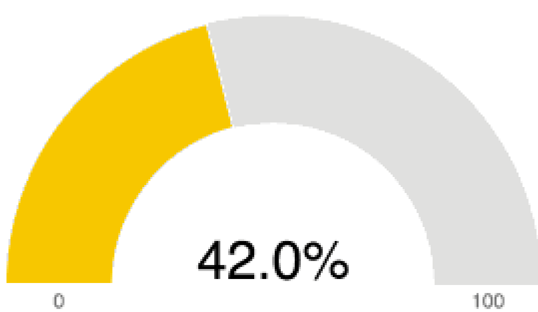
vector graphics

Photographs and Images in SVG

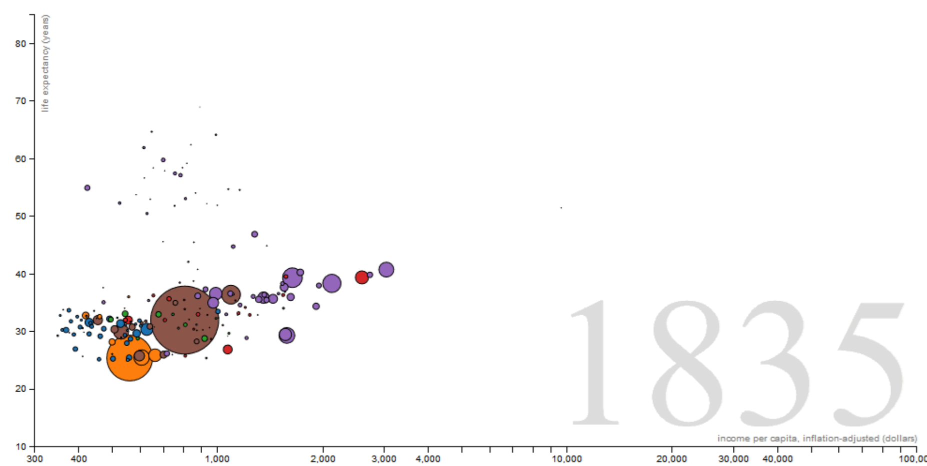
D3 Uses SVG







The Wealth & Health of Nations

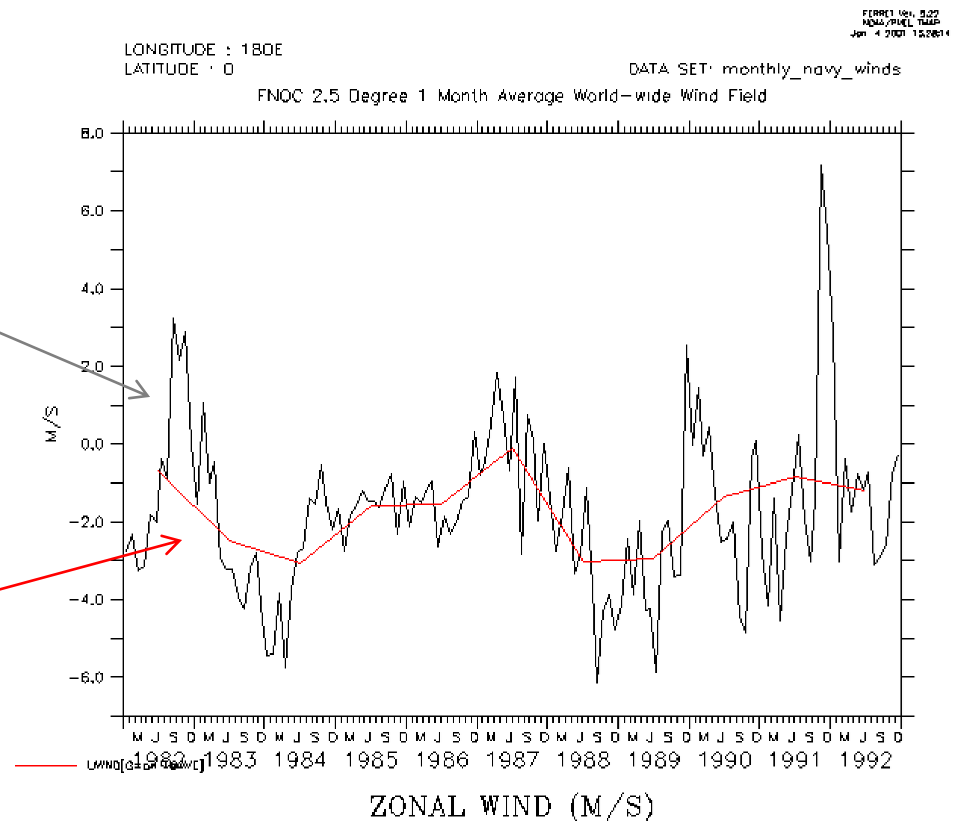


De-Noising

Filtering also eliminates noise in the data

noise

data



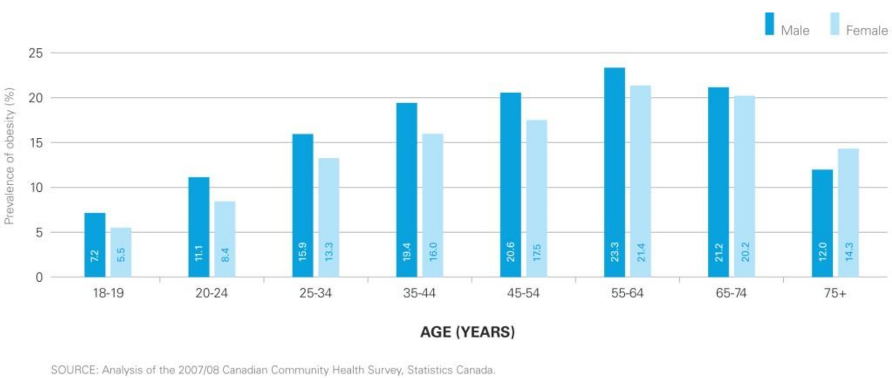
Back to Bar Charts

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

- the bins do the smoothing

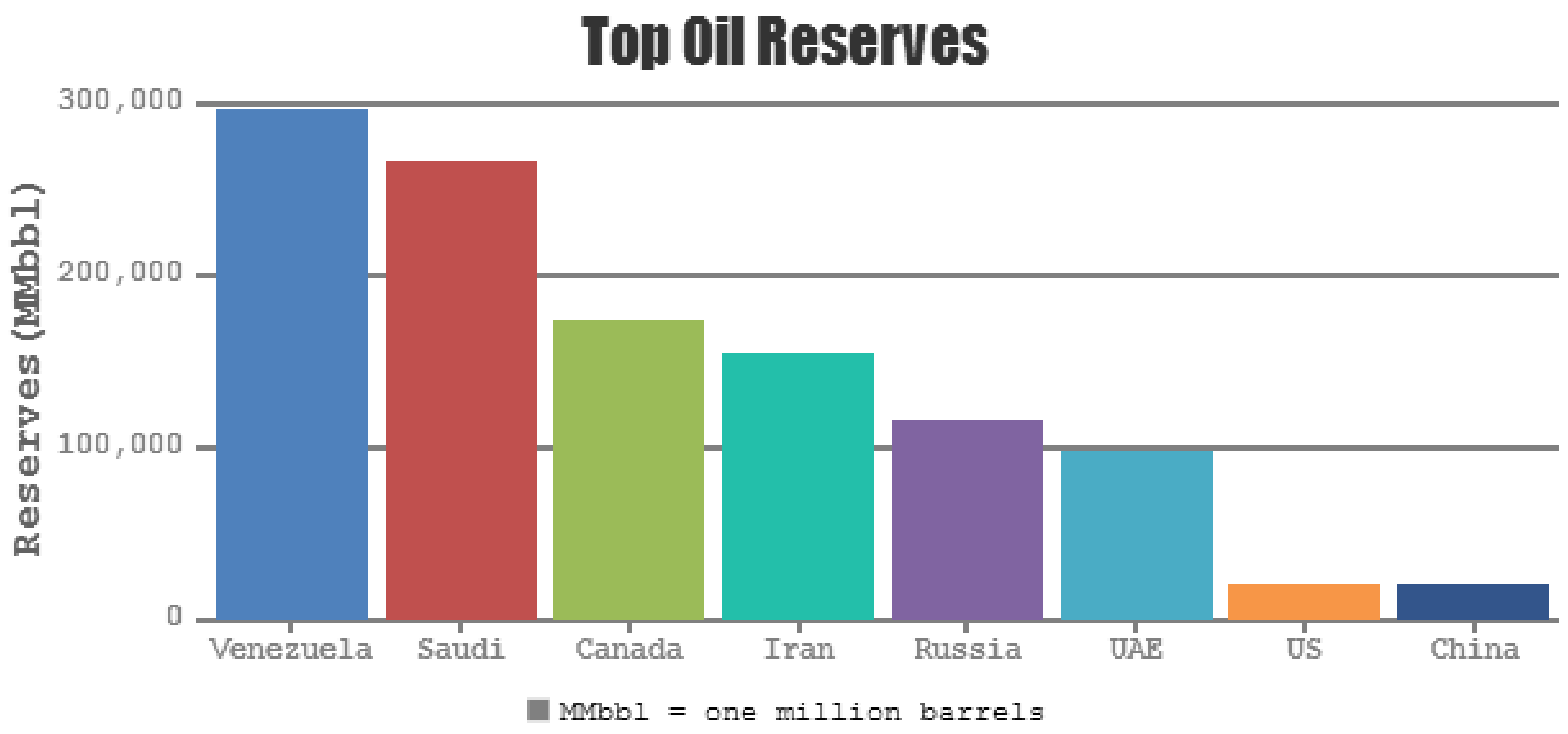
Example:

- obesity over age (group)



Bar Charts

Of course, bar charts can also hold categorical data



Bar Charts in D3

http://bl.ocks.org/mbostock/3885304

Working with bar charts will be your job for Lab 2

- the next two slides offer some help with calculations

Bar Chart 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

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

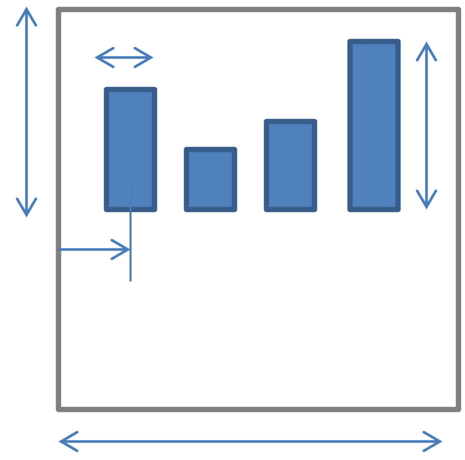
Given a data value *val* increment (++) the bin value

- but first initialize bin val array to 0

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

Bar Chart 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

Project #1

Find some interesting data on the web

- something that challenges and interests you
- there are many data sources on the web
- use google and some imagination

Criteria for selection

- more than 500 data points (observations)
- more than 10 attributes
- the more the better (you can always reduce it)

Deliverables

- 2-page report that describes the data and justifies your choice
- a URL to the data source

Due date

- Tuesday, September 18, 11:59pm

PROJECT #1: DATASET EXAMPLE

Multivariate - Quantitative data and Categorical data

Data Items

	A	B	C	D	E	F	G	H	I
1	**Name**	**Country**	**Miles Per Gallon**	**Accceleration**	**Horsepower**	**weight**	**cylinders**	**year**	**price**
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	Ford Fairmont A	USA	20,2	15,8	85	2965	6	78	2375
13	Ford Fairmont M	USA	25,1	15,4	88	2720	4	78	2275
14	Plymouth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100
25	Toyota Corona	Japan	27,5	14,2	95	2560	4	78	2975

Categorical

Quantitative

Categorical (Ordinal)
Quantitative

Data types

Quantitative (Numerical)

Categorical (Ordinal)

Project #1: Notes on Dataset

Other data types are OK

- text, images, video, logs, etc.
- just convert them to numbers via appropriate mechanism as discussed in class
- must produce a spreadsheet of rows (data items) and attributes (columns)

Categorical data

- color, brand, country, etc.
- convert into numbers by assigning a numerical ID

Questions?

The course has been set up with Piazza

- http://piazza.com/stonybrook/fall2018/cse332/home
- please let me know if you cannot access it

Make use of this handy discussion forum

- ask questions of general interest
- give advice to peers (those who ask questions)
- give general feedback (observe etiquette)
- but obviously, don't provide actual solutions and aid in cheating