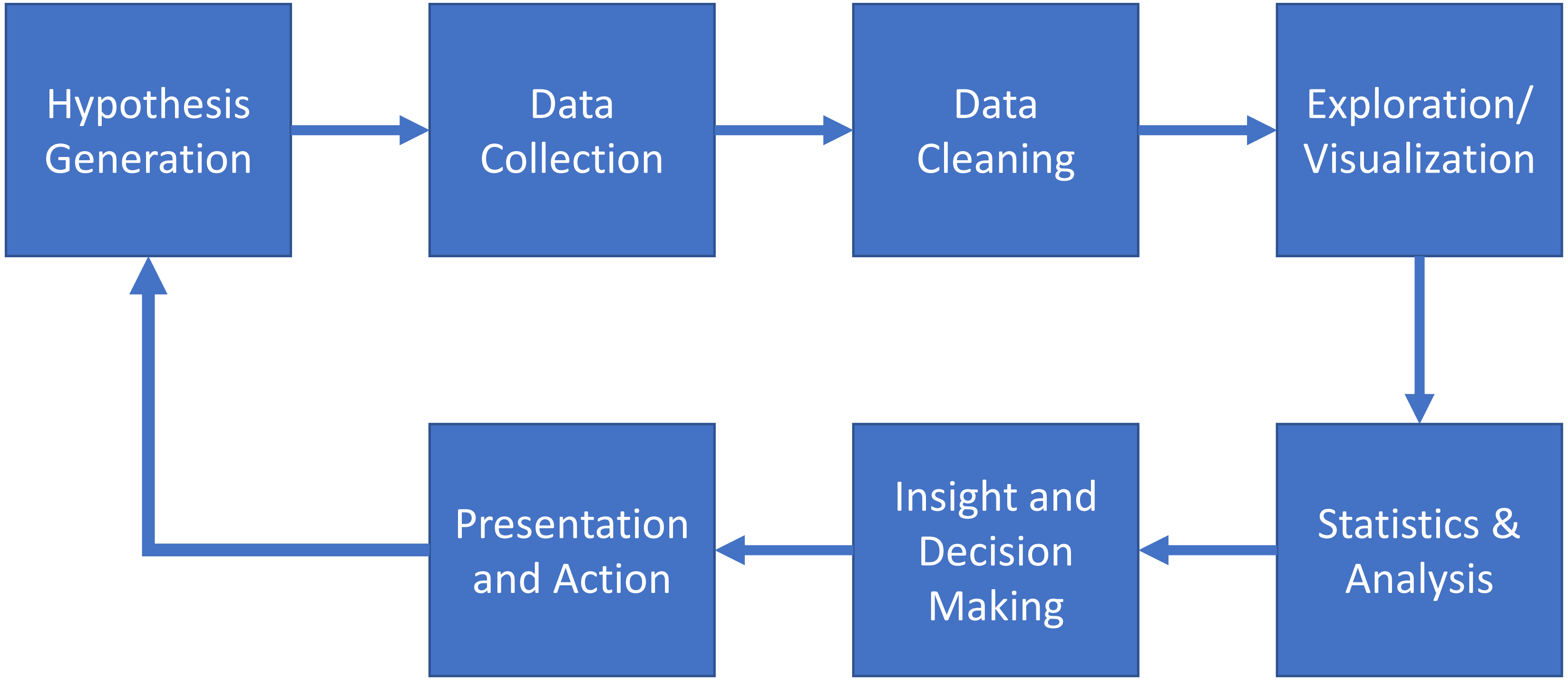
Data Science/Analysis Process



Data Visualization

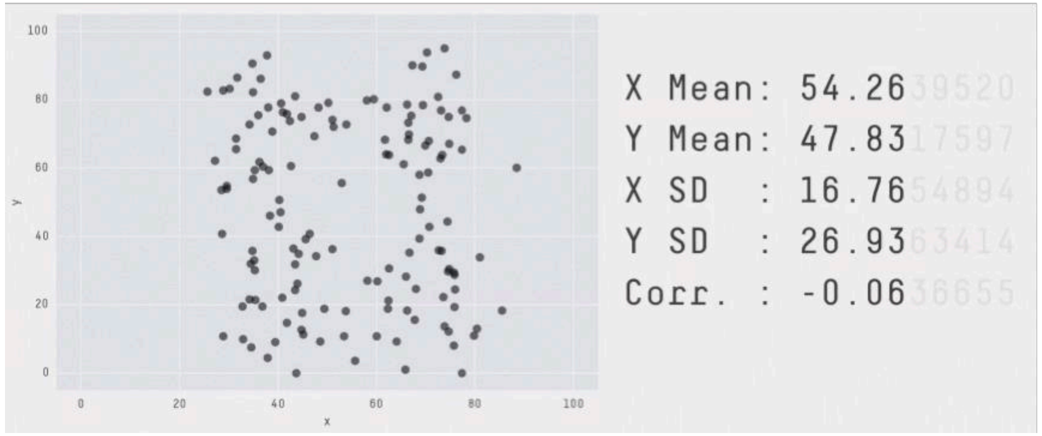
Two types:

- Data Exploration
- Data Presentation

You can’t identify trends in data unless you can see the trends to know what to look for

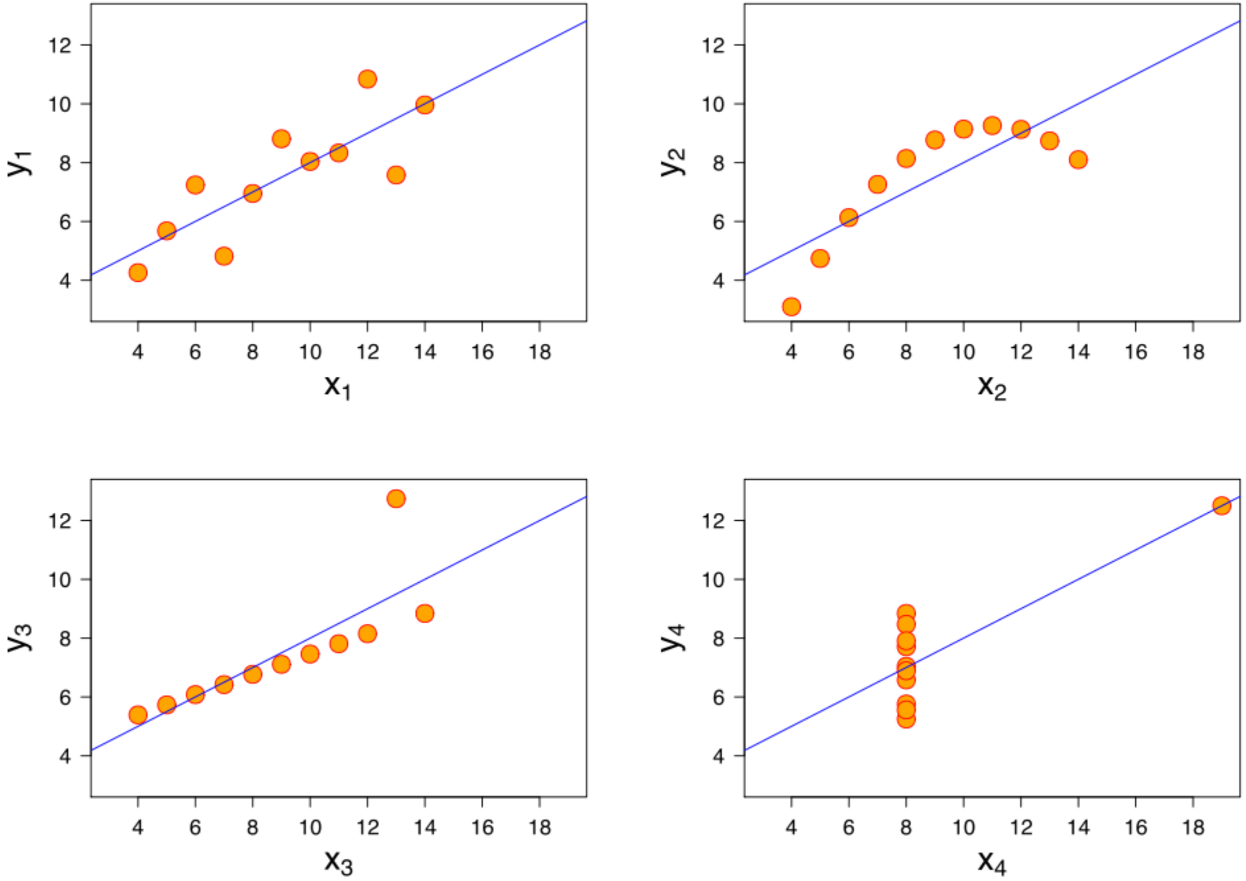
Graphical Exploration

Often presents a better view of your data (although less quantitative) than numerical statistics



```
X Mean: 54.2639520
Y Mean: 47.8317597
X SD  : 16.7654894
Y SD  : 26.9363414
Corr. : -0.0636655
```

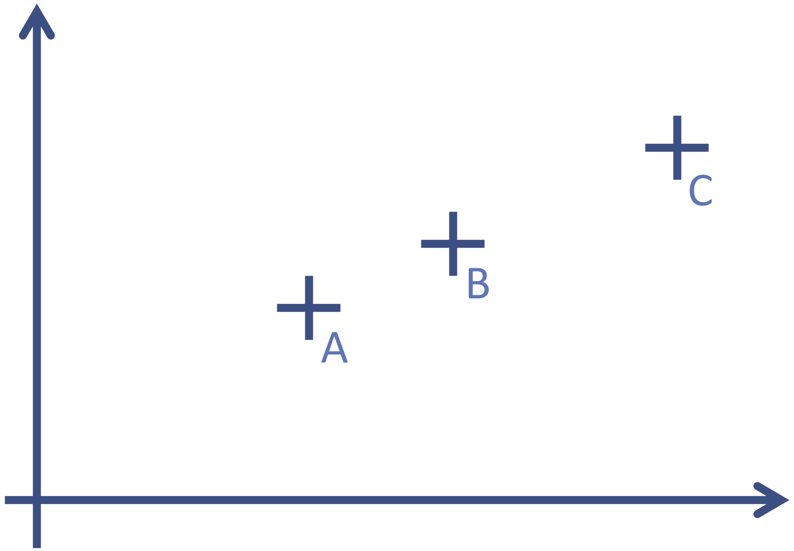
Same Statistics, Very Different Pictures



Visual Encodings

Visual language is a **sign system**

- Images perceived as a set of signs
- Sender **encodes** information in signs
- Receiver **decodes** information from signs



- A, B, C are distinguishable
- B is between A and C
- BC is twice as long as AB

The Brain and Visualizations

How many 3’s?

1281768756138976546984506985604982826762

9809858458224509856458945098450980943585

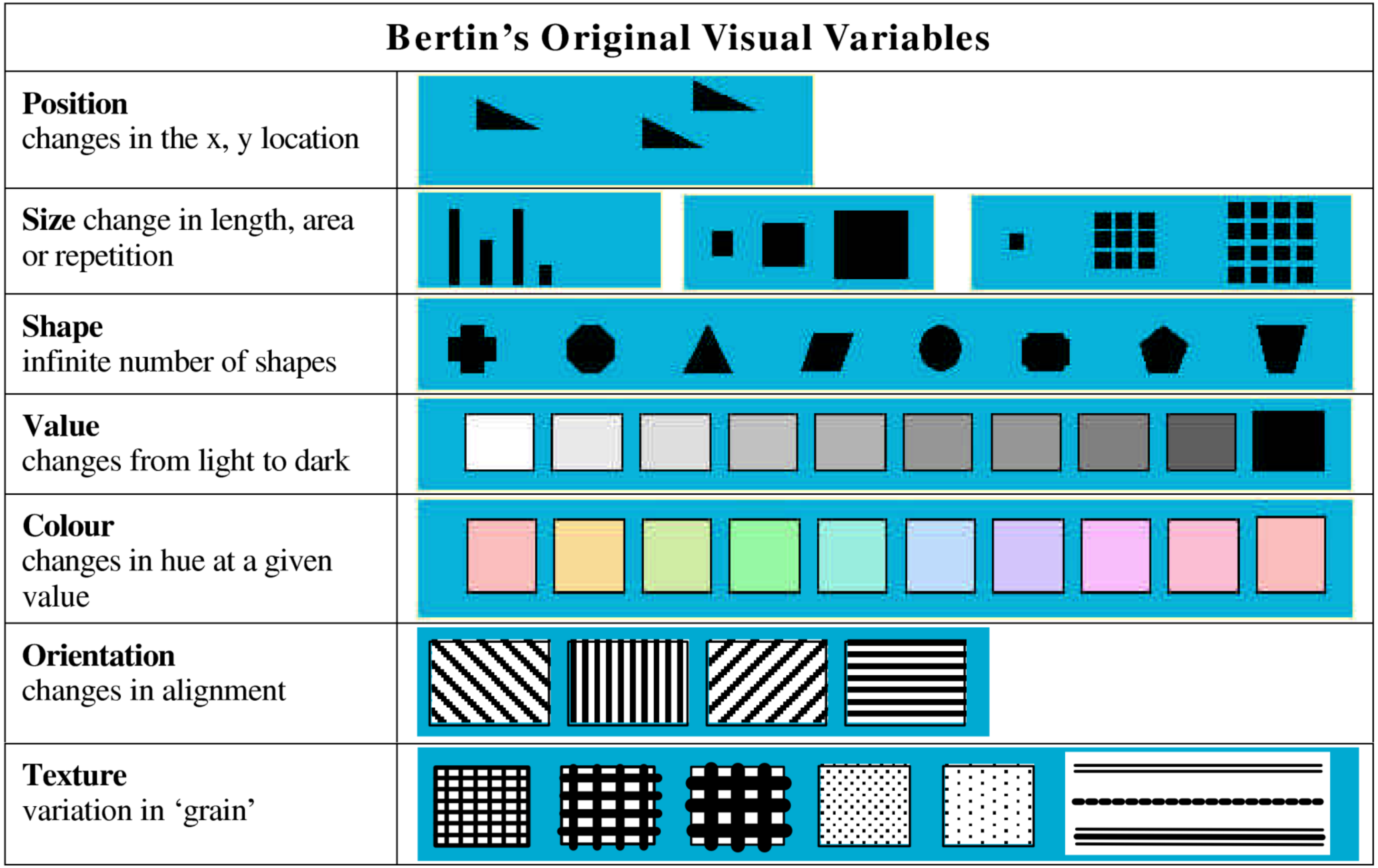
9091030209905959595772564675050678904567

8845789809821677654876364908560912949686

How many 3’s?

12817687561389765469845069856049828267629
80985845822450985645894509845098094358590
91030209905959595772564675050678904567884
57898098216776548763649085609129496 86

Visual Variables



Bertin's Original Visual Variables	
Position changes in the x, y location	
Size change in length, area or repetition	
Shape infinite number of shapes	
Value changes from light to dark	
Colour changes in hue at a given value	
Orientation changes in alignment	
Texture variation in 'grain'	

Types of Data

Categories (labels) =, ≠

- Fruits: apples, oranges, grapes

Ordinal (ordered categories) =, ≠, <, >, ≤, ≥

- Quality of meat: A, AA, AAA

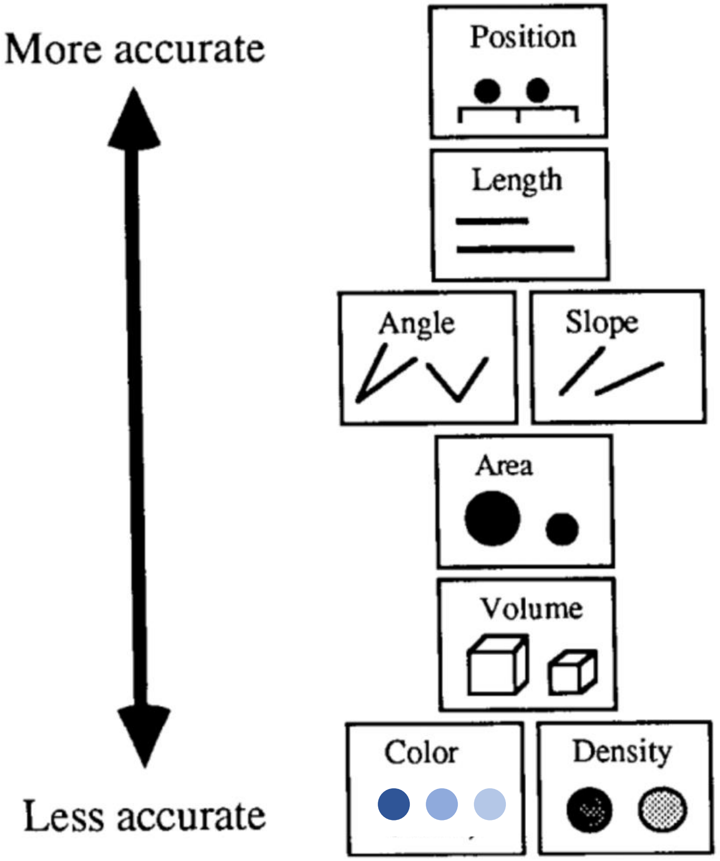
Quantitative (numbers) =, ≠, <, >, ≤, ≥, +, -, *, /

- Dates: January 3rd, 1932; Oct 18, 1981
- Temperature (Celsius)
- Length, Mass
- Temperature (Kelvin)

When to Use Visual Variables

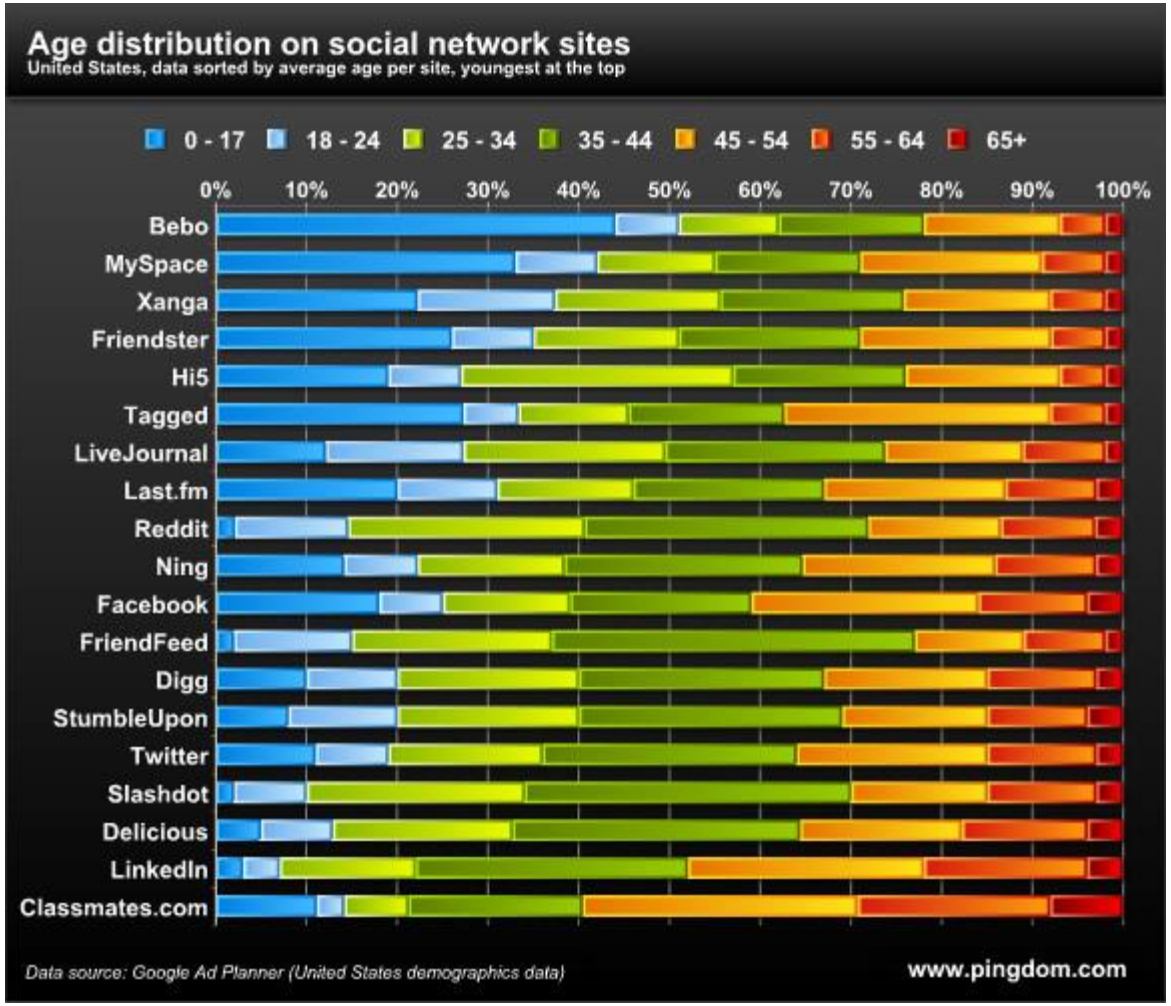
	Categorical	Ordinal	Quantitative
Position	Yes	Yes	Yes
Size	Yes	Yes	Yes
Value	Yes	Yes	Sometimes
Texture	Yes	Sometimes	
Color	Yes	Sometimes	
Orientation	Yes		
Shape	Yes		

How accurately can we detect visual differences?



Correct Use of Visualization

Correct Use of Bar Chart



Andrei Pandre

Incorrect Use of a Bar Graph

Bar Length has No Meaning

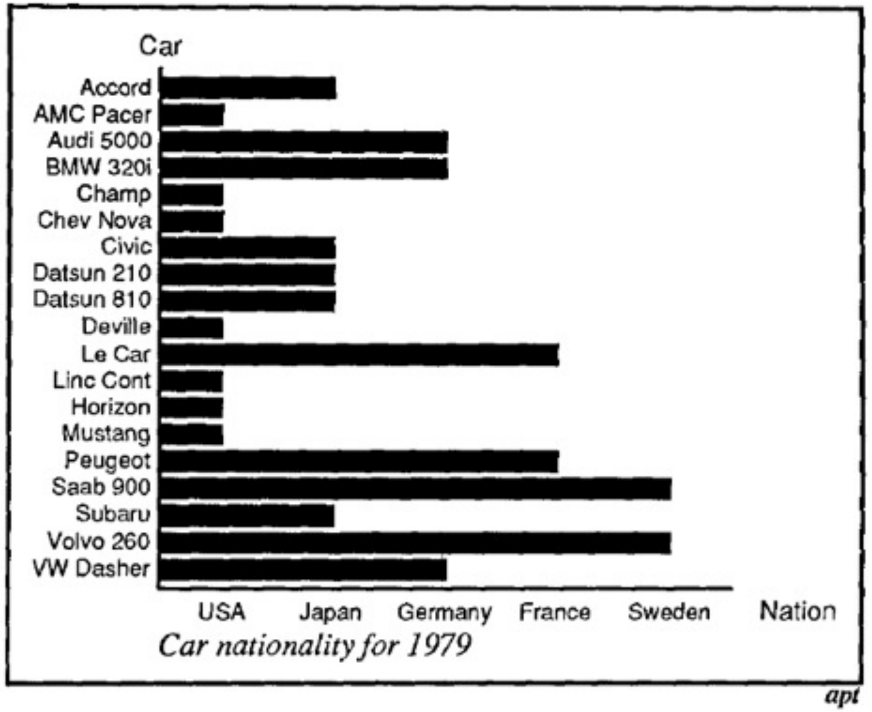
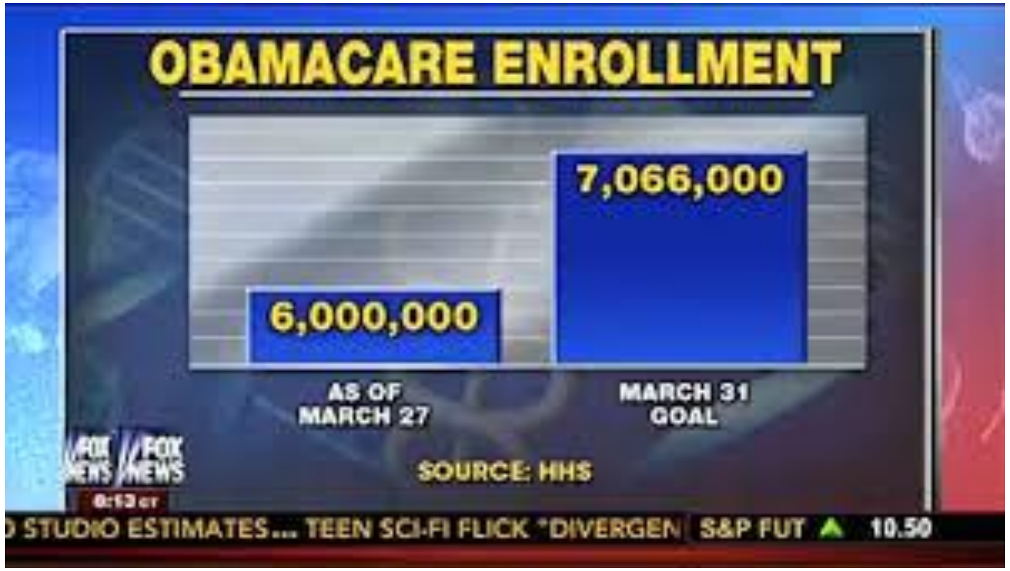


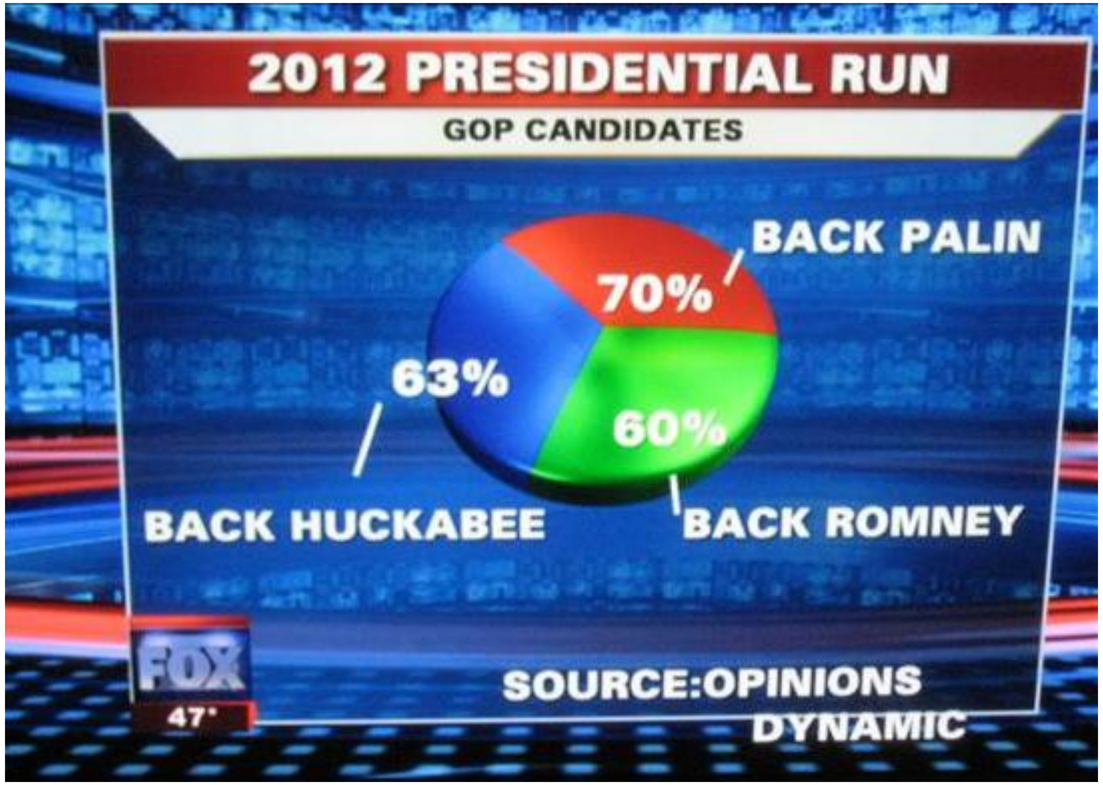
Fig. 11. Incorrect use of a bar chart for the *Nation* relation. The lengths of the bars suggest an ordering on the vertical axis, as if the USA cars were longer or better than the other cars, which is not true for the *Nation* relation.

Incorrect Use of a Bar Graph

Proportion of Bars is Misleading



Incorrect Use of a Pie Chart



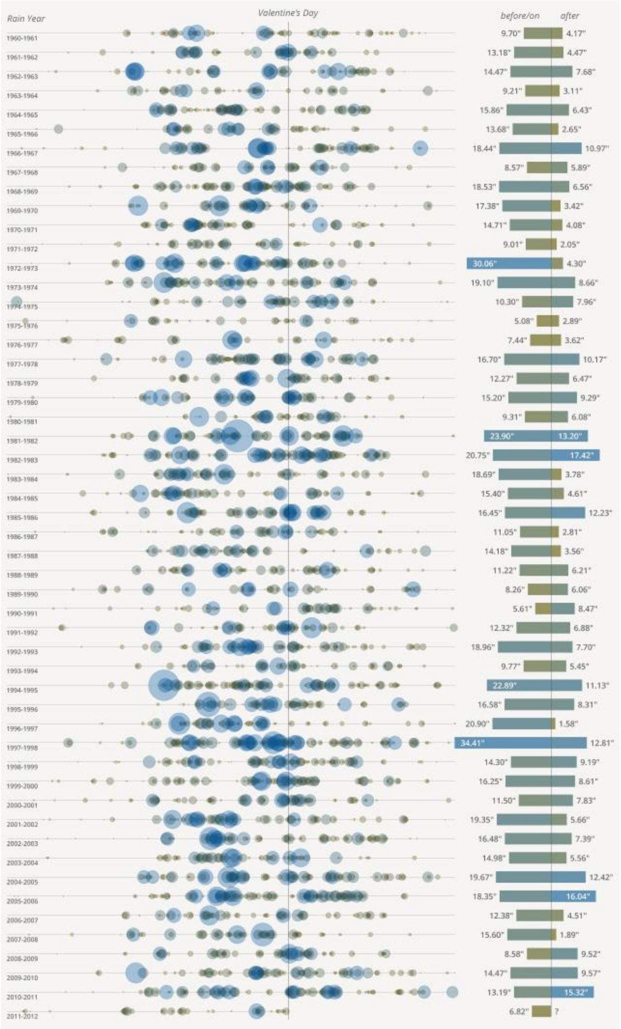
Examples of Pretty Good Visualizations

Find the visual variables…

Find the Visual Variables

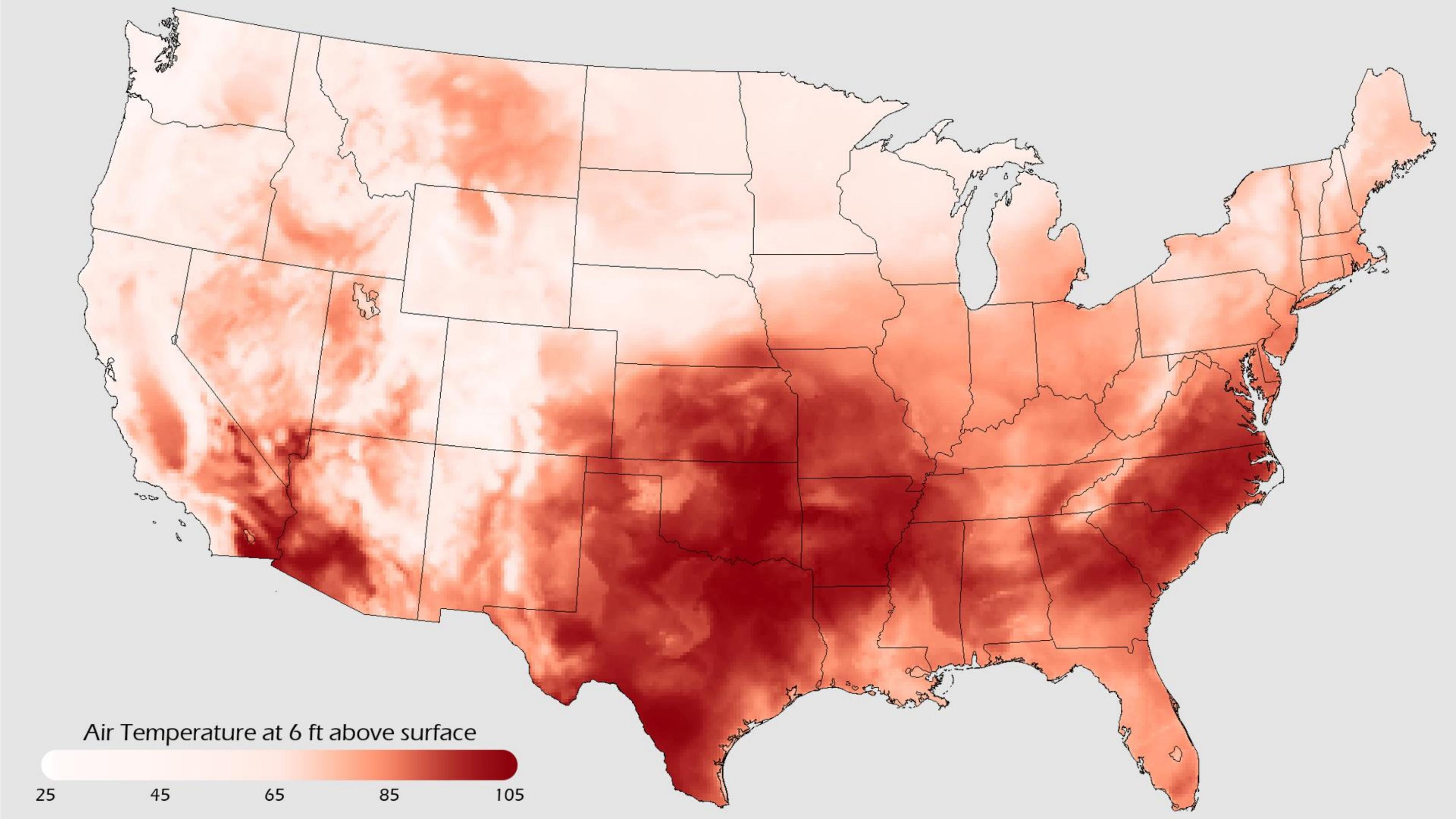
Rain in San Francisco every year from 1960-2011
July through June
Centered on Valentines Day

What visual variables are used?



Stephen Von Worley

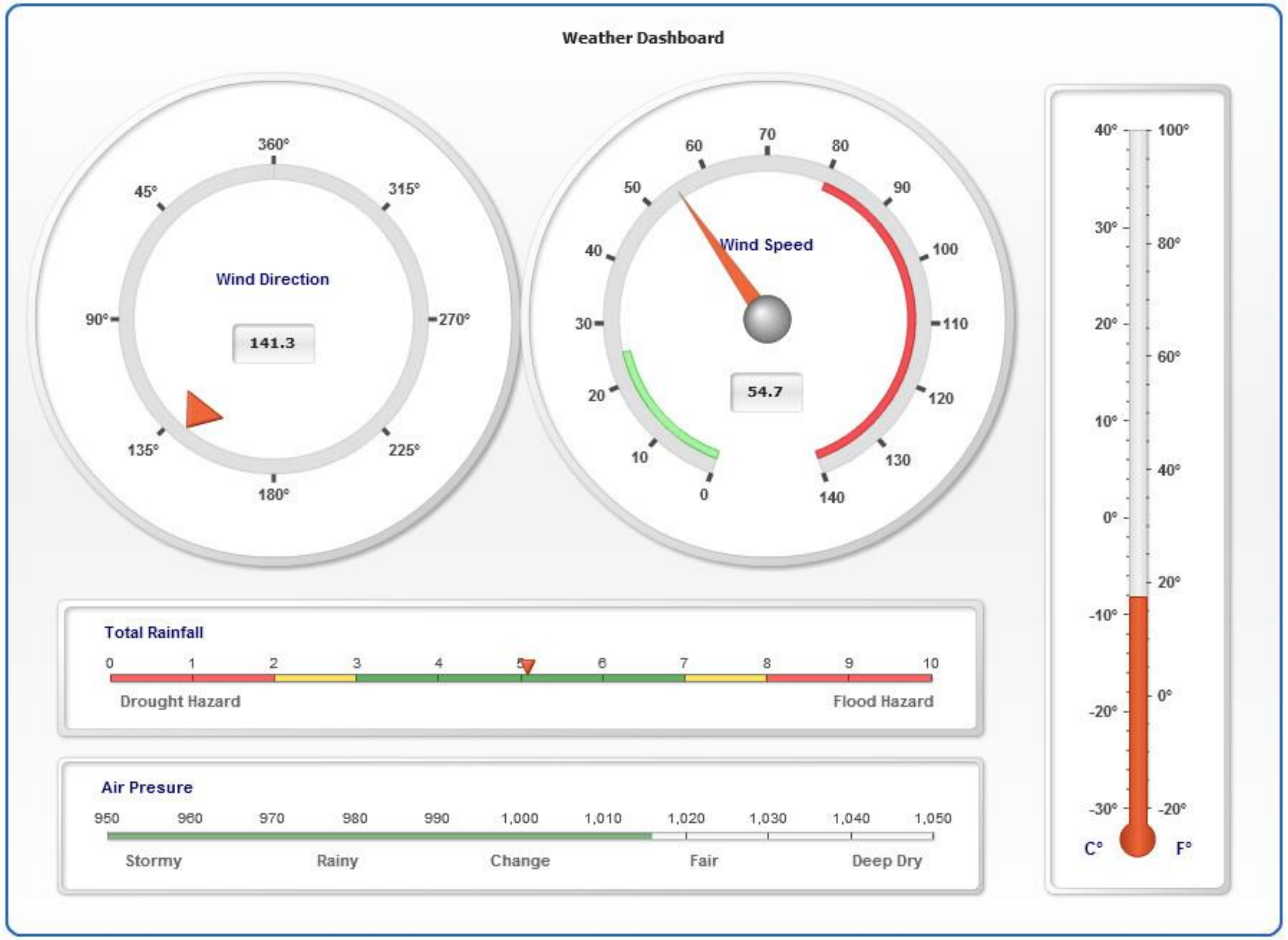
Find the Visual Variables



NOAA, July 12, 2014

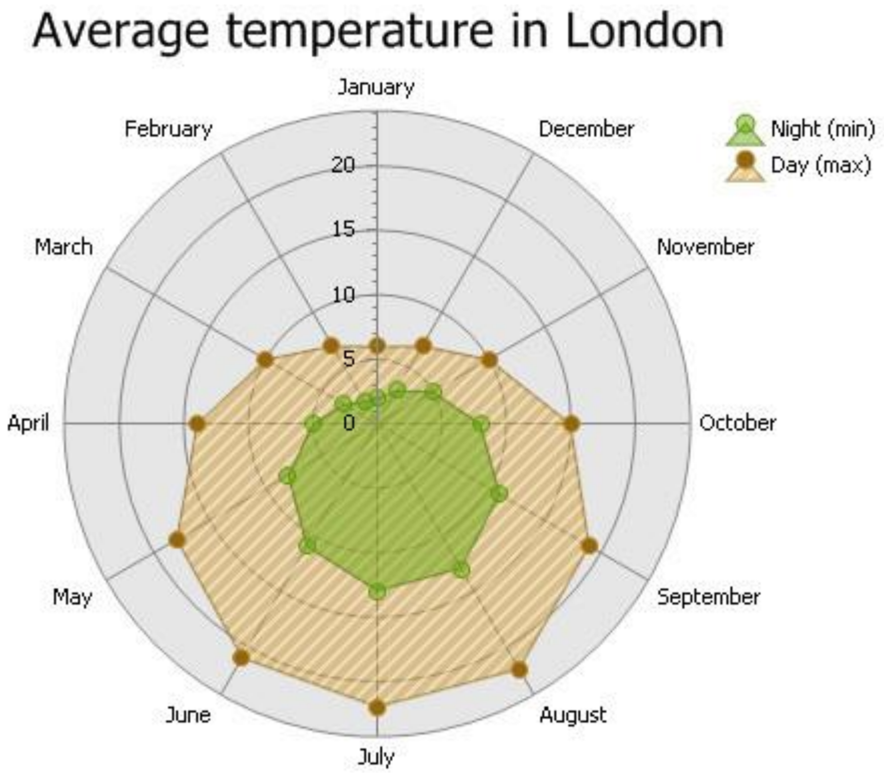
Find the Visual Variables

Weather Dashboard Analogy to a Car Dashboard



Andrei Pandre

Find the Visual Variables



Circular Area Chart – Where Values are Centered

Andrei Pandre

Choosing Visualizations

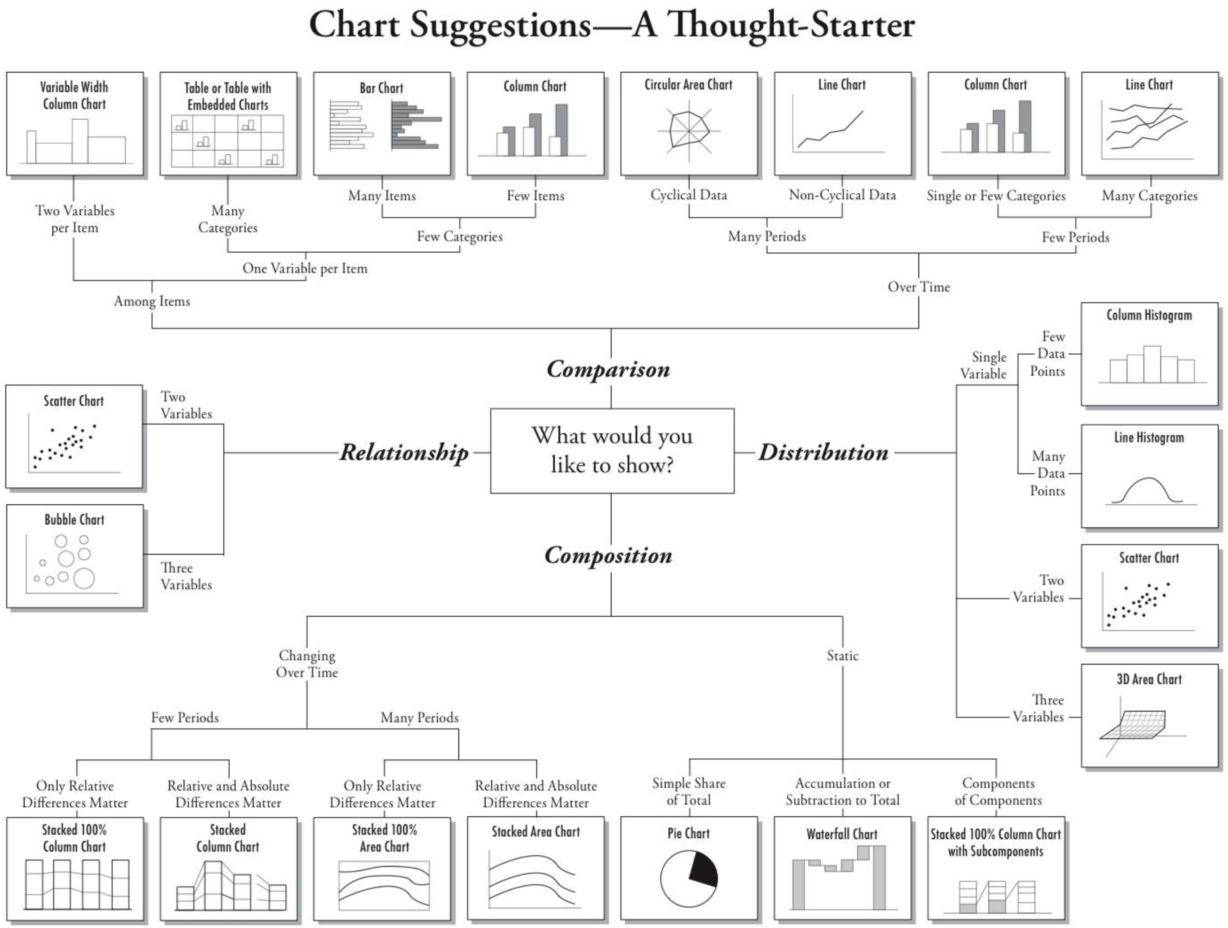


Chart Suggestions—A Thought-Starter

© 2006 A. Abela — a.v.abela@gmail.com

Visualizing Data

Types of visualizations

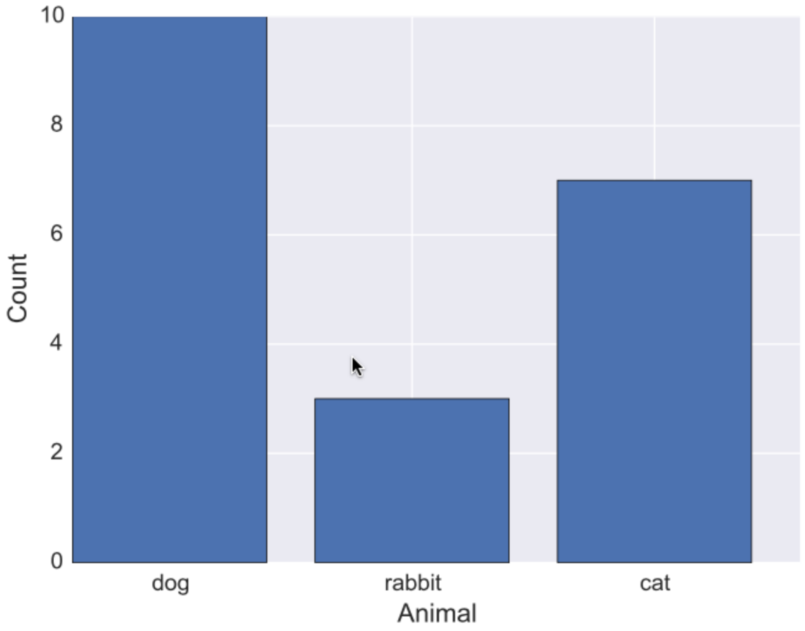
- Histograms
- Scatterplots
- Bar Charts
- Stacked Bar Charts
- Pie Charts
- Time Series
- Decision Trees, Flow Charts, etc

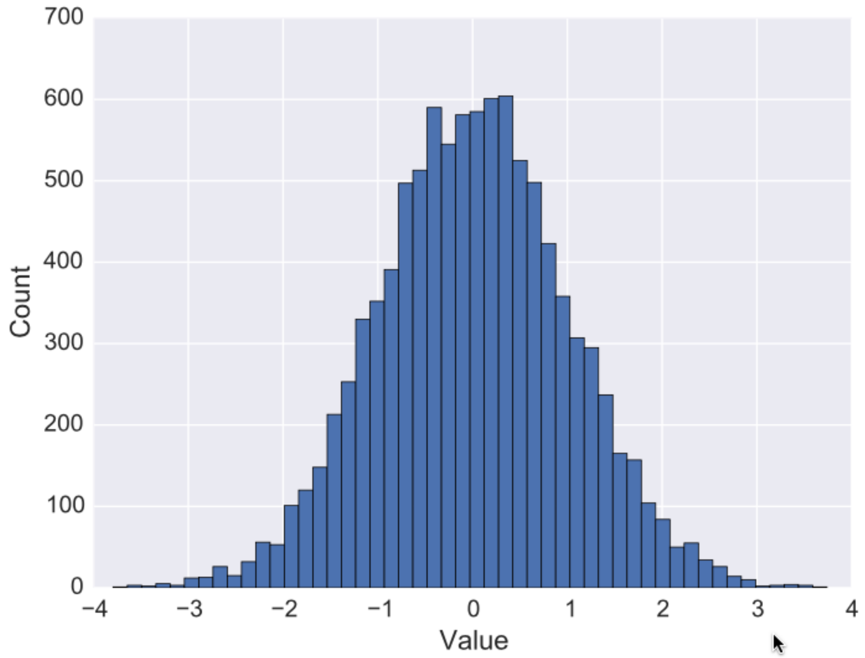
Visualizing 1 Dimensional Data

- “I want to know how many of each product type are in my data”
- “I want to know the proportion of people who have cats in my data”

Histograms

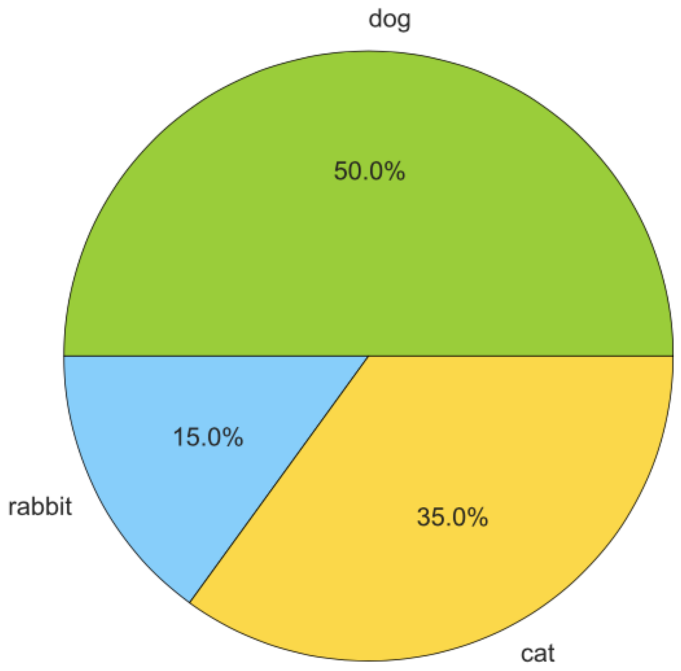
Counts (y axis) per category or value range (x axis)





Pie Chart

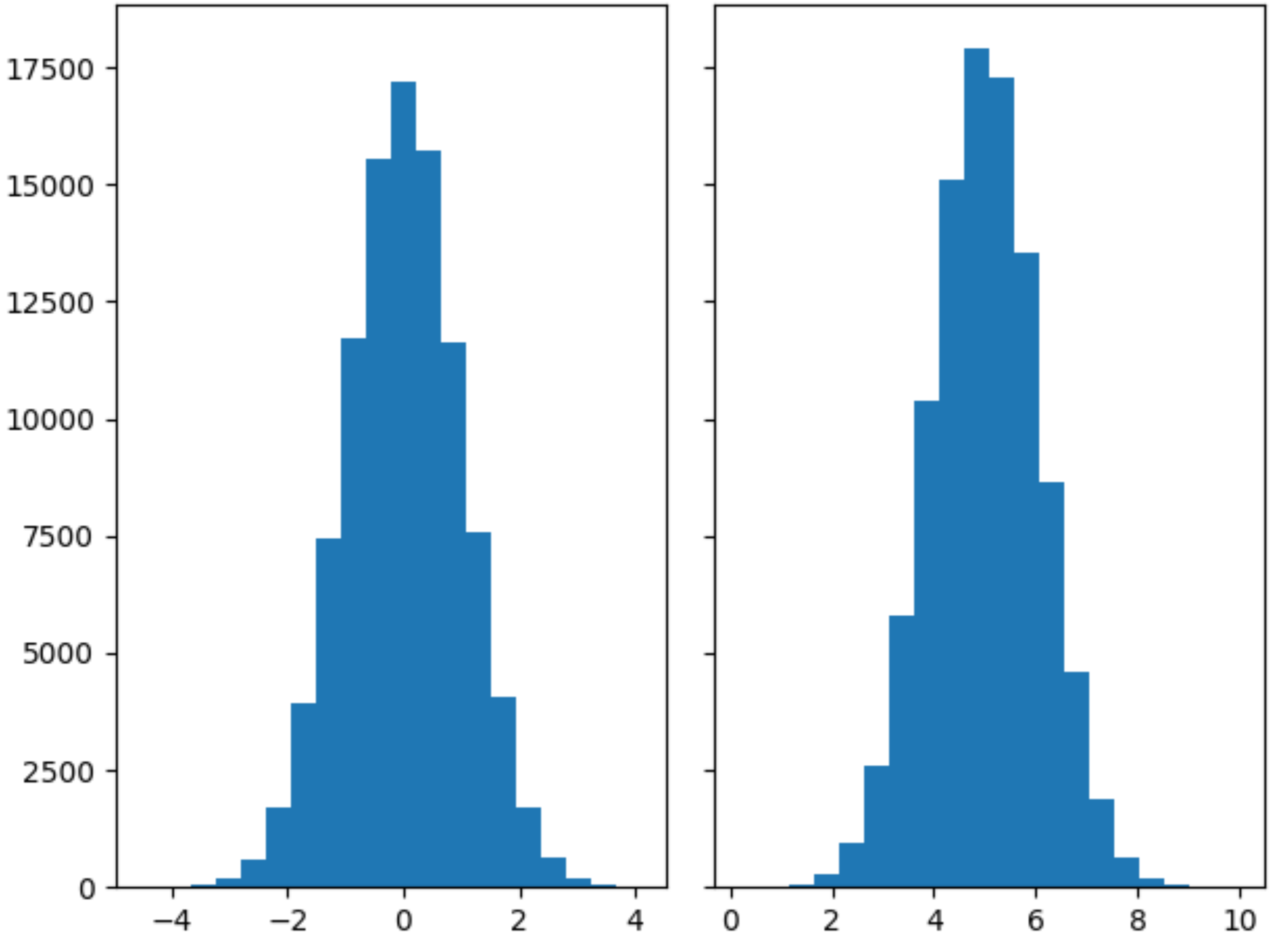
Proportion of the whole count



Histogram Matplotlib

```
# From Matplotlib website
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import colors

N_points = 100000
n_bins = 20

# Generate a normal distribution, center at x=0 and y=5
x = np.random.randn(N_points) #random data
y = .4 * x + np.random.randn(N_points) + 5 #shifted random

# Make 1 row and 2 columns (where the y axes are the same)
fig, ax = plt.subplots(1, 2, sharey=True, tight_layout=True)
# We can set the number of bins with the 'bins' argument
ax[0].hist(x, bins=n_bins)
ax[1].hist(y, bins=n_bins)

plt.show()
```

Matplotlib

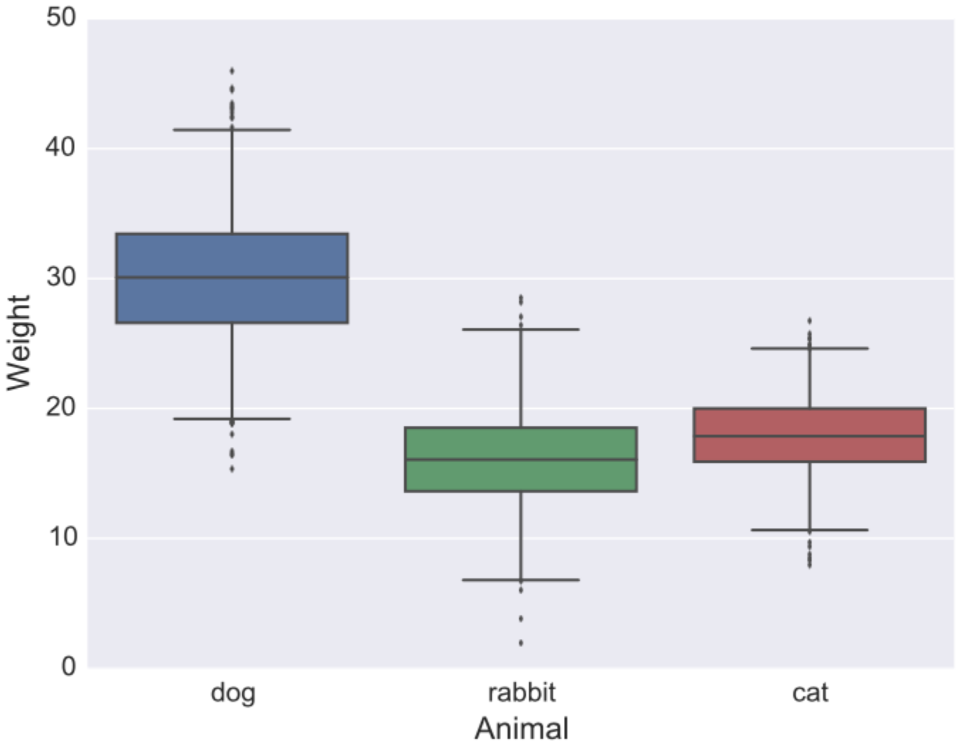


2 Dimensional Data

- ”I want to know the **cost** of each **product category** that we have”
- “I want to know the **weight** of the animals that people own, by **category**”
- ”I want to know how the **size** of the product affects **the cost of shipping**”

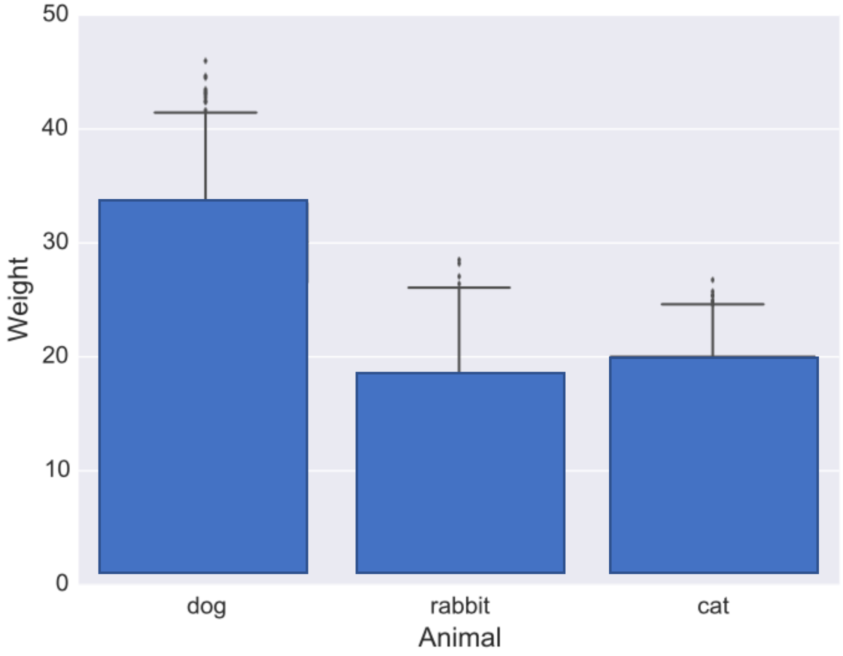
Box and Whiskers Plot

One dimension is a category and one is numeric, shows ranges of values



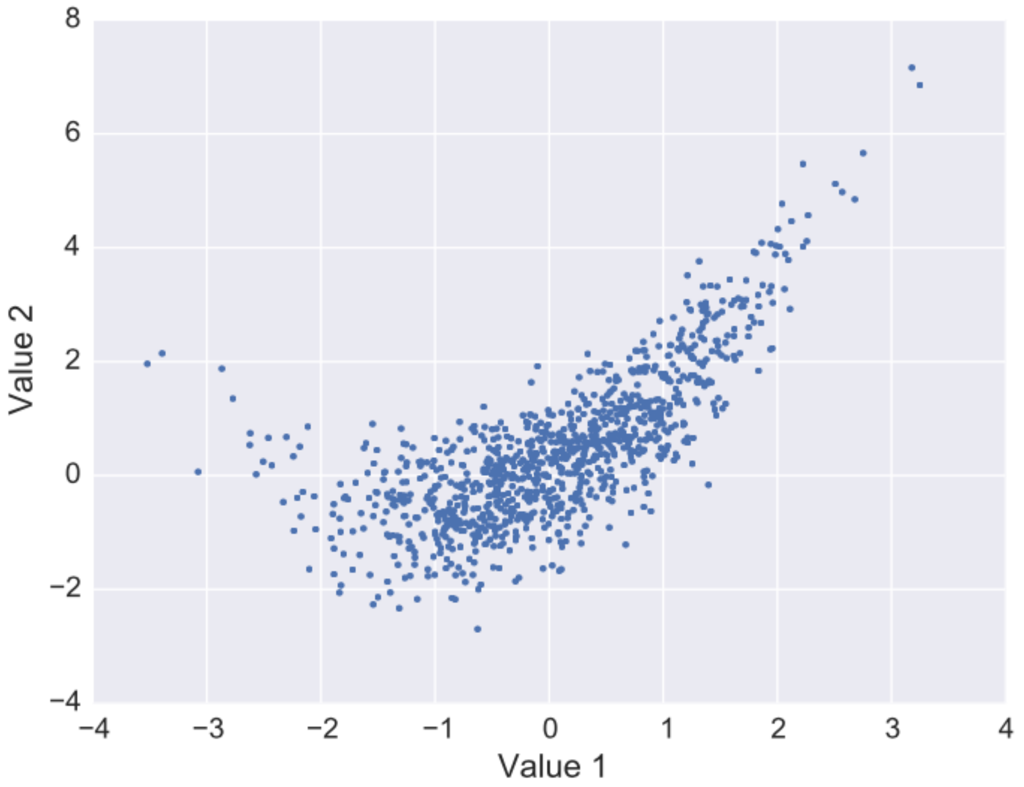
Bar Chart

One dimension is a category and one is numeric, shows AVERAGE of values



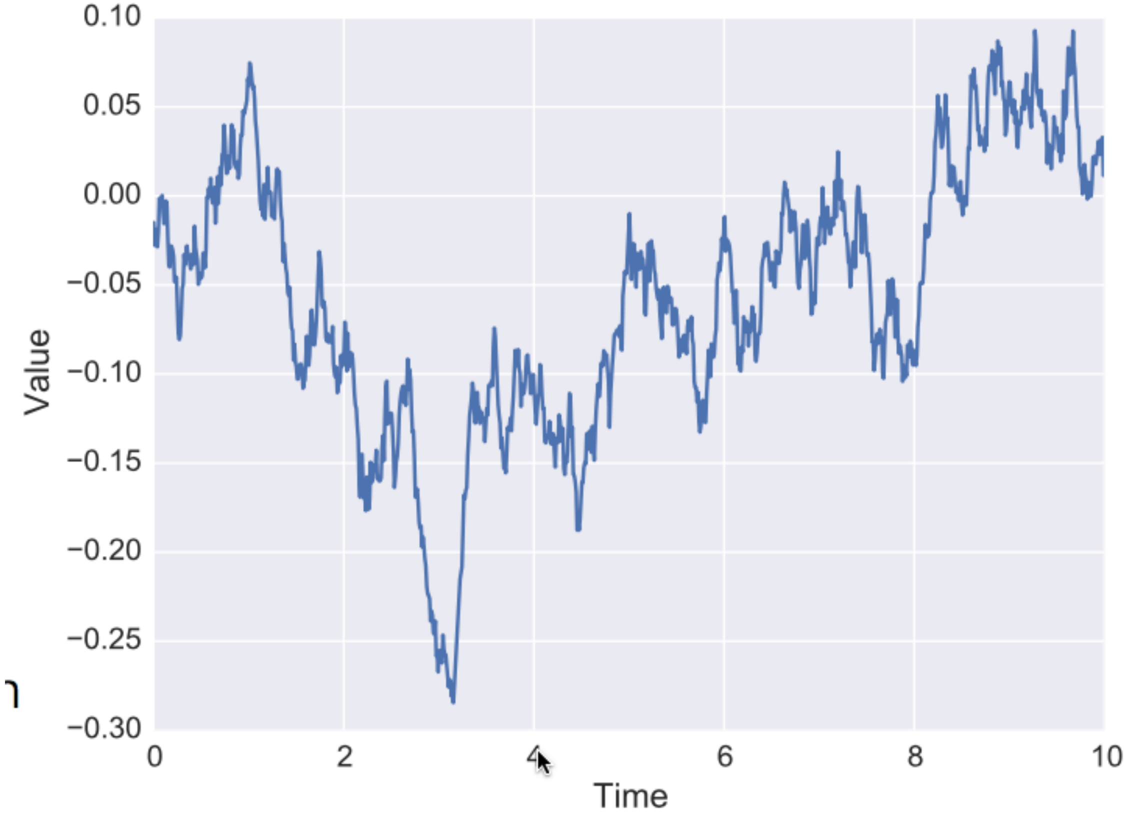
Scatterplot

Two numeric dimensions, shows correlations (or lack thereof)



Line Plot

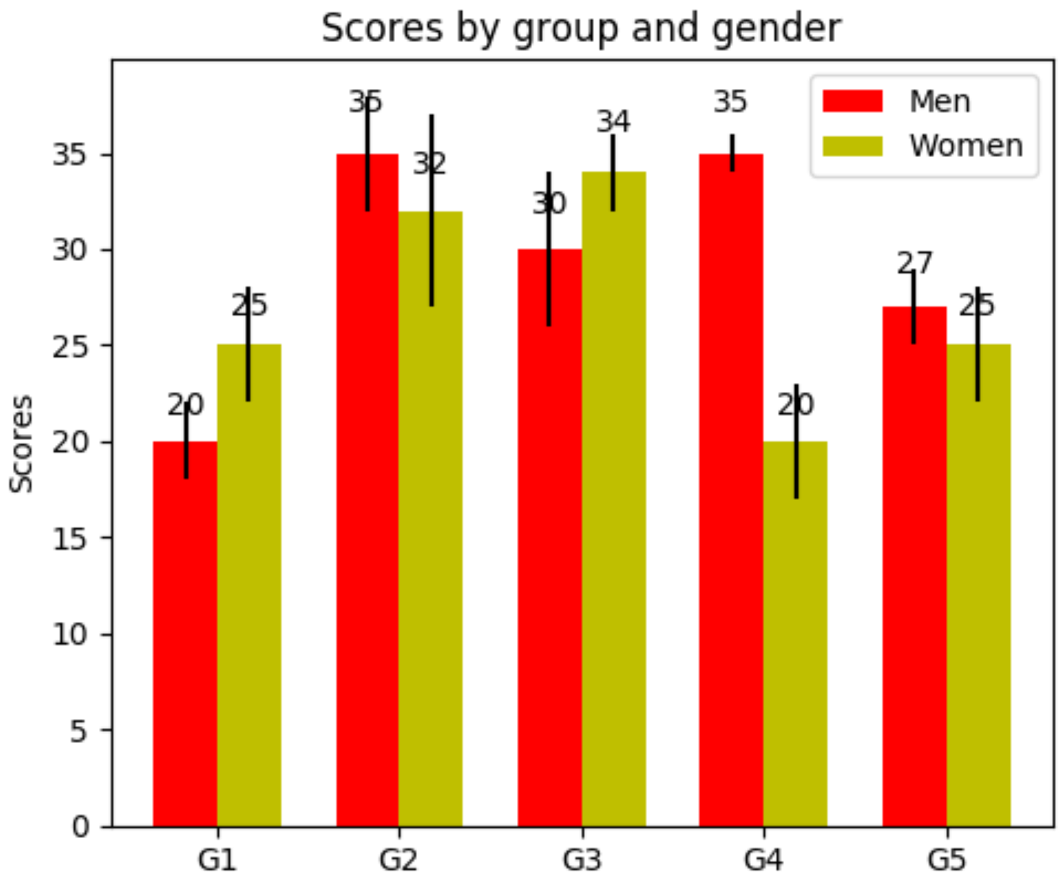
TIME and a numeric dimension



Bar Chart Matplotlib

```python
# From Matplotlib website
import numpy as np
import matplotlib.pyplot as plt

N = 5
men_means = (20, 35, 30, 35, 27) #each number is a mean for a separate bar
men_std = (2, 3, 4, 1, 2)
women_means = (25, 32, 34, 20, 25)
women_std = (3, 5, 2, 3, 3)

ind = np.arange(N) # the x locations for the 5 categories
width = 0.35 # the width of the bars
fig, ax = plt.subplots()
rects1 = ax.bar(ind, men_means, width, color='r', yerr=men_std)
rects2 = ax.bar(ind+width, women_means, width, color='y', yerr=women_std)
# add some text for labels, title and axes ticks
ax.set_ylabel('Scores')
ax.set_title('Scores by group and gender')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels(('G1', 'G2', 'G3', 'G4', 'G5'))
ax.legend((rects1[0], rects2[0]), ('Men', 'Women'))
plt.show()
```

Matplotlib

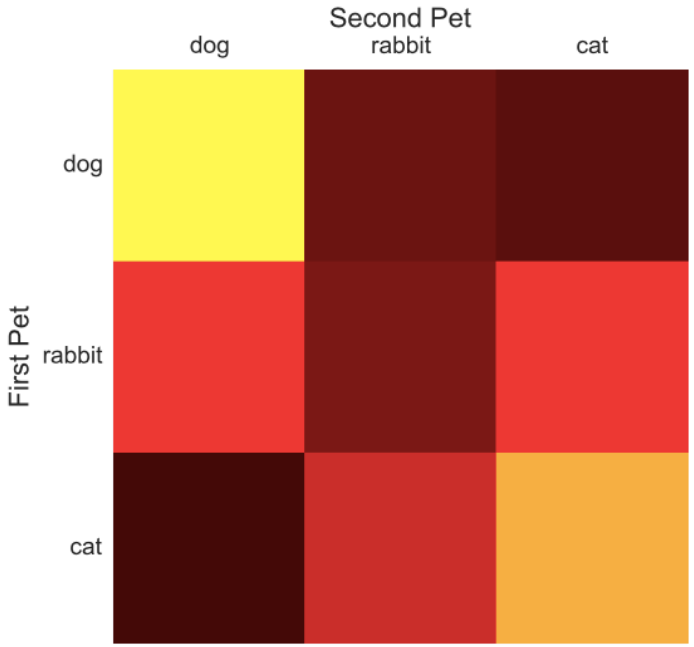

Scores by group and gender
Scores
0
5
10
15
20
25
30
35
G1
G2
G3
G4
G5
20
25
35
32
30
34
35
20
27
25
Men
Women

3 Dimensional Data

- ”I want to know the **cost** and the **development time** by **product category**”
- “I want to know the **weight** of the animals that people own and **cost**, by **category**”
- ”I want to know how the **size** of the product and the **manufacture location** affects **the cost of shipping**”

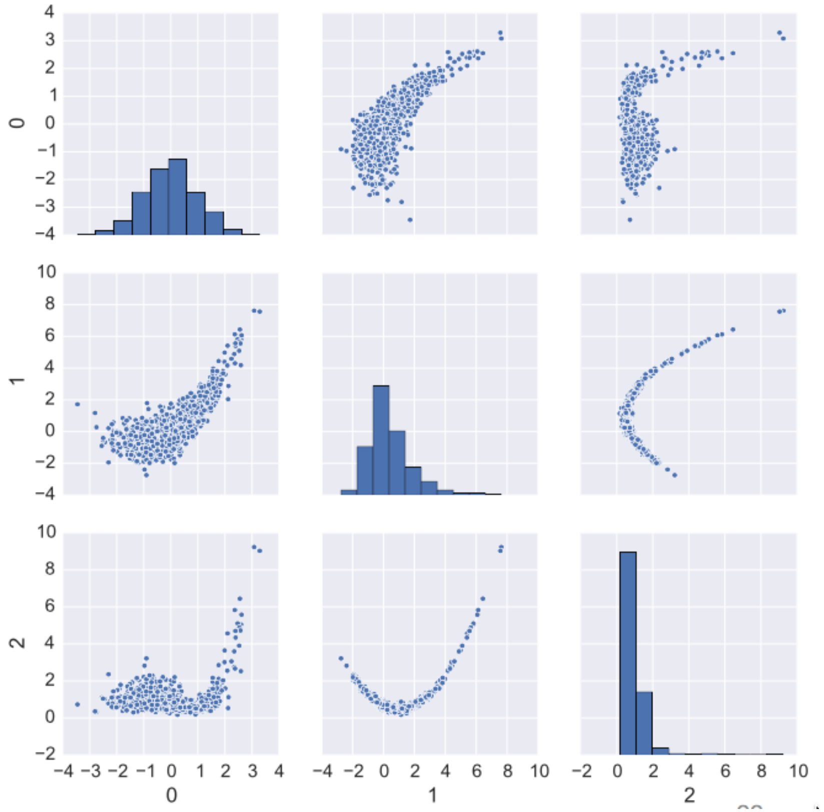
Heatmap

Two categorical variables, color shows numeric value or count



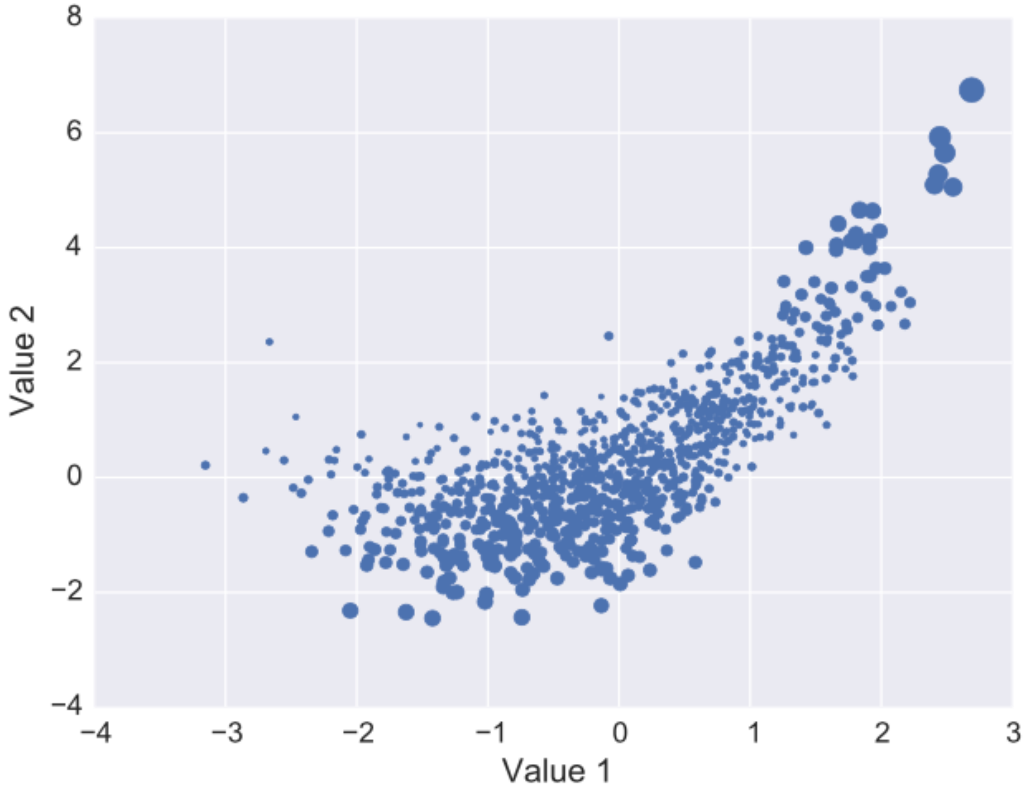
Scatterplot matrix

Histograms on the diagonal scatterplots (or other appropriate plots for each variable)



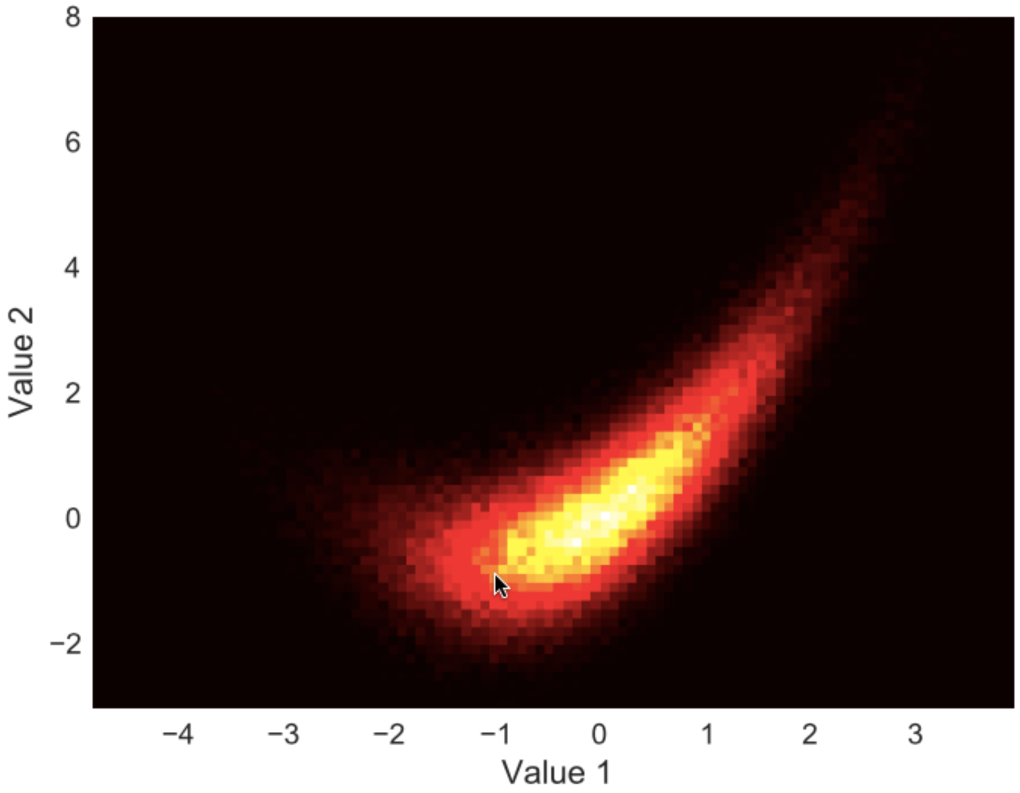
Bubbleplot

Three numeric variables



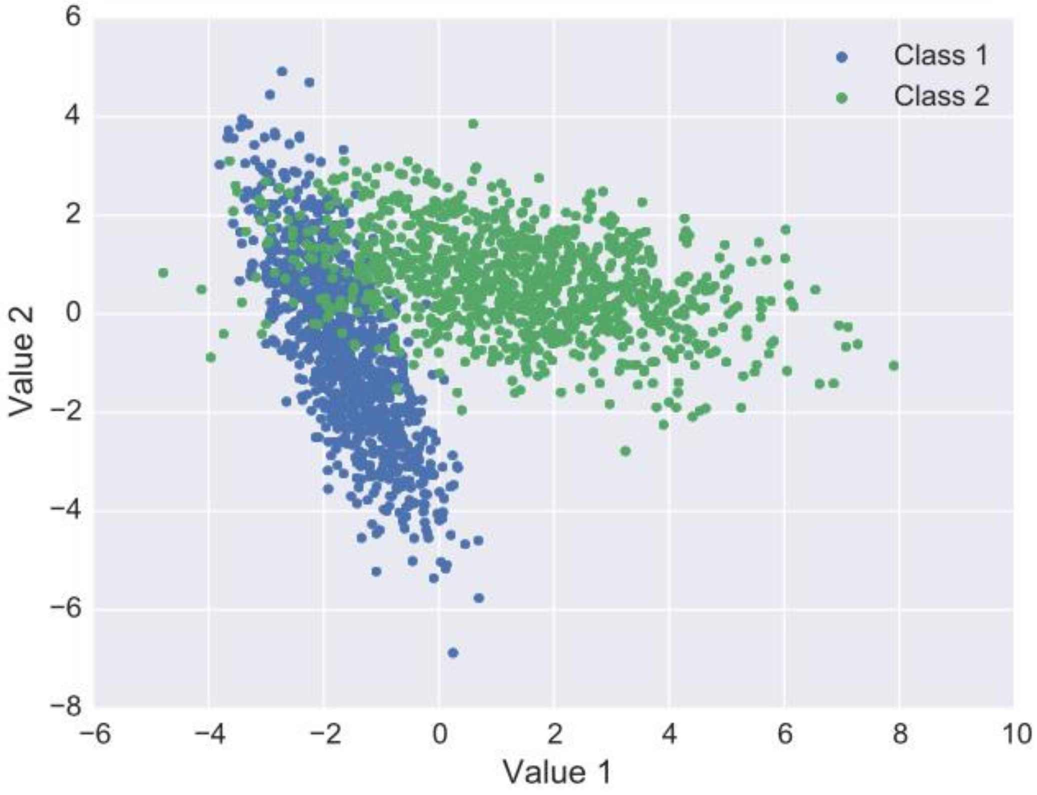
Scatterplot Heatmap

- Three numeric variables



Color Scatterplot

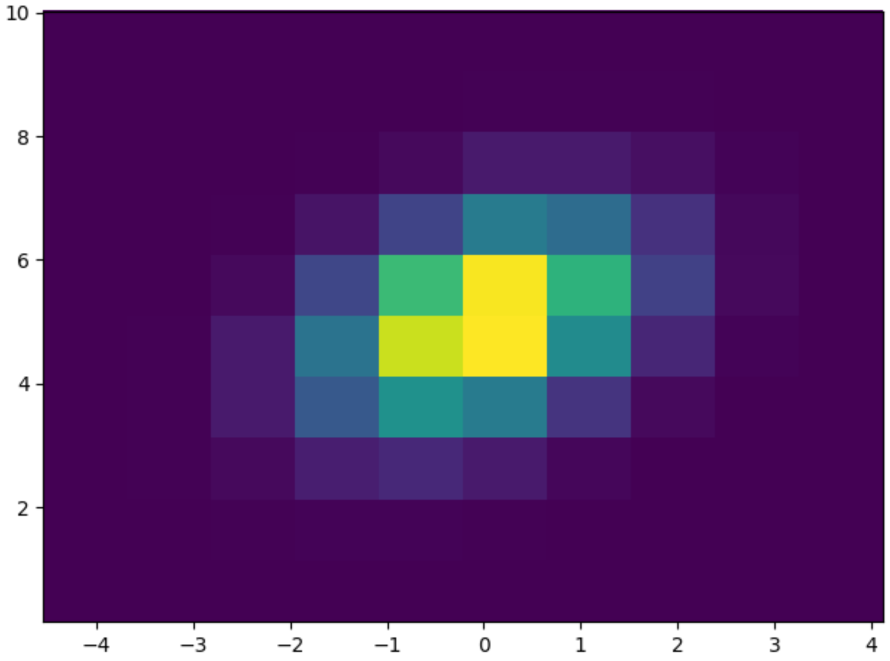
- Two numeric variables and one categorical

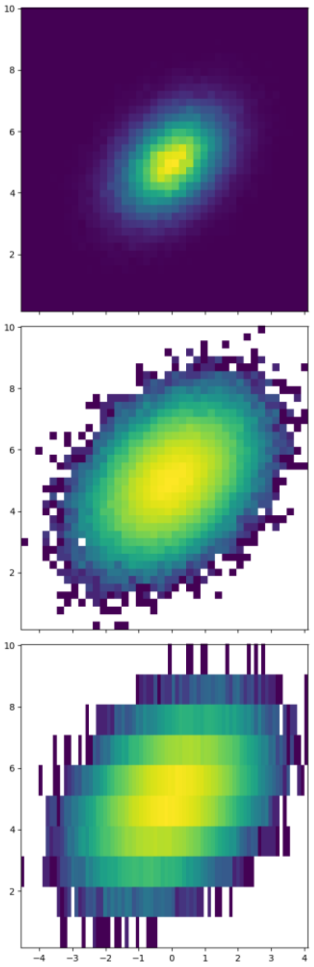


Matplotlib

```python
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import colors

N_points = 100000
n_bins = 20

# Generate a normal distribution, center at x=0 and y=5
x = np.random.randn(N_points)
y = .4 * x + np.random.randn(100000) + 5
fig, ax = plt.subplots(tight_layout=True)
hist = ax.hist2d(x, y)
plt.show()
```

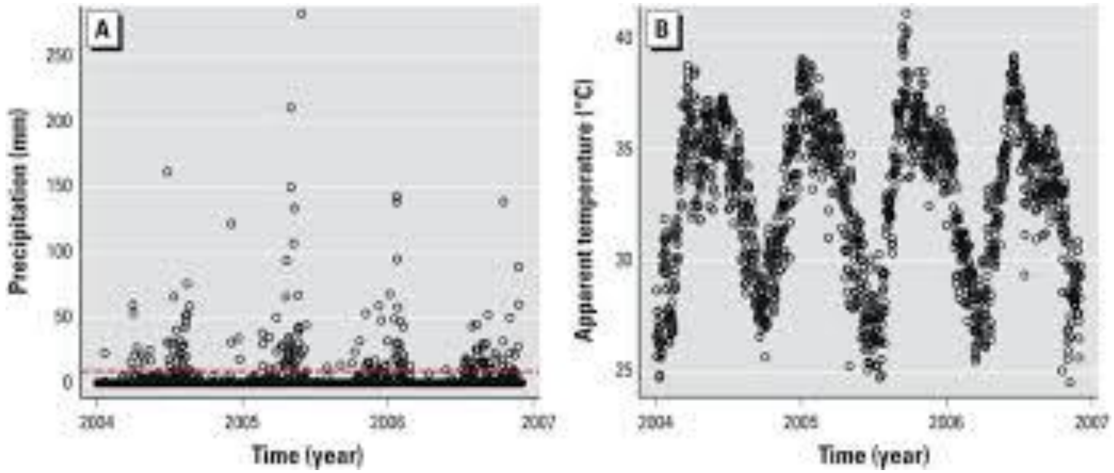
Matplotlib 2D/3D Histogram



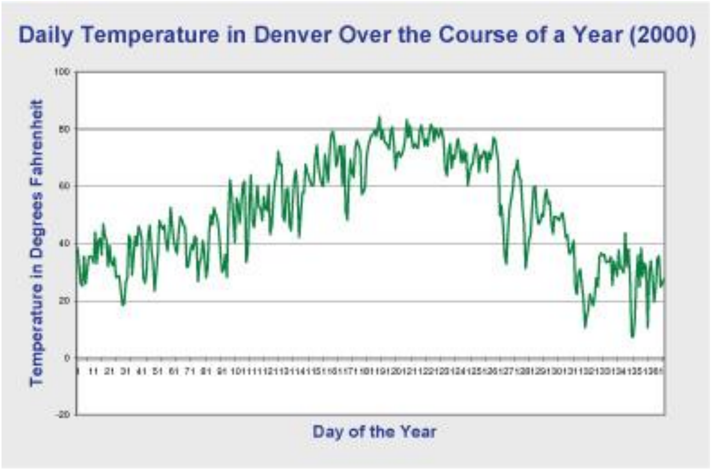


Time Series

Time is x axis, numeric variable on y axis

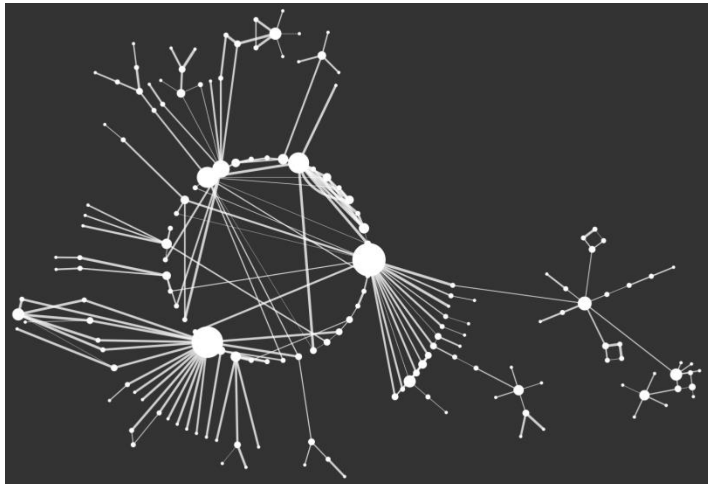


Rain and Temperature in Chennai, India



Temperature in Denver, CO

Visualizing Graphs and Trees



Graph Basics

- Nodes = entities
- Edges = relations

Graph Types

- Graphs generally model relations between data
- Trees represent hierarchies

qiita.com, bigml.com

Graph Visualization Applications

- Tournaments
- Organization Charts
- Genealogy
- Diagramming (e.g., Visio)
- Biological Interactions (Genes, Proteins)
- Computer Networks
- Social Networks
- Simulation and Modeling
- Integrated Circuit Design

Graph Examples and D3 Library

- https://bl.ocks.org/mbostock/4062045
- https://www.jasondavies.com/collatz-graph/

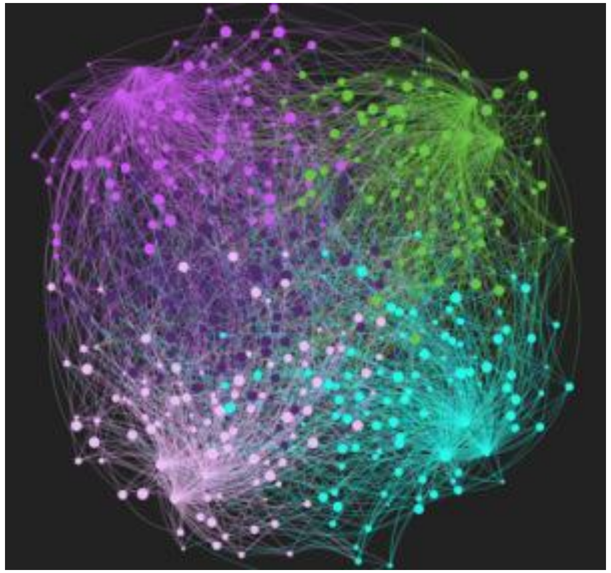
- https://github.com/d3/d3/wiki/Gallery

Graph Spatial Layout

Layout to see all nodes and edges

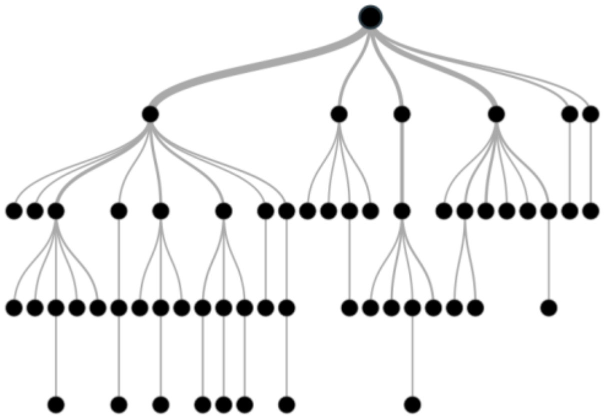
Ideally, also see structure in graph

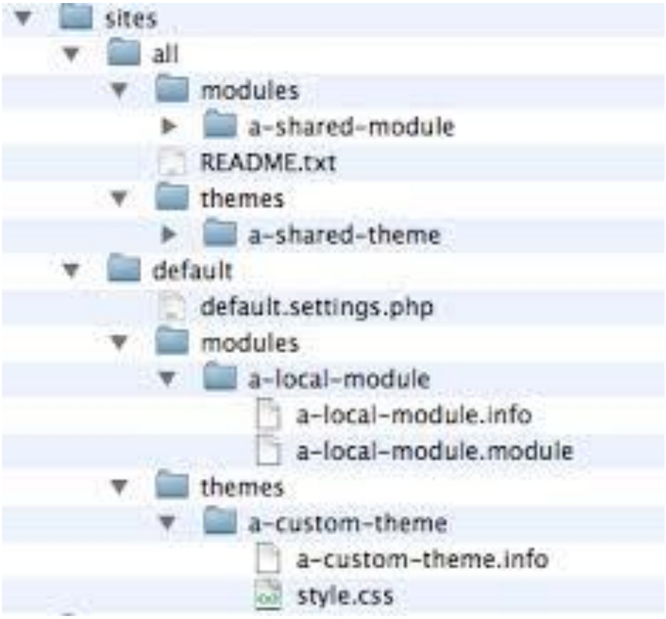
- Connectivity
- Network Distance
- Clustering
- Ordering



Tree Visualization

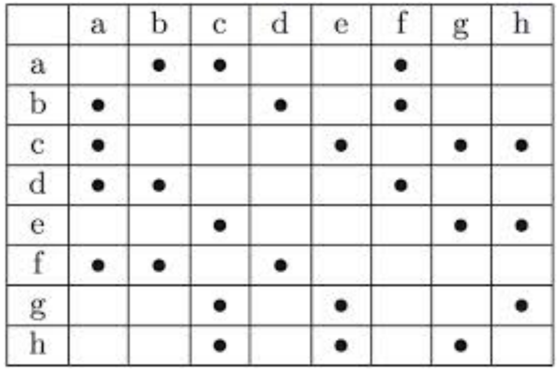
- Indentation
  - Linear list, indentation encodes depth
- Node-link diagrams
  - Nodes connected by lines/curves
- Enclosure diagrams
  - Represent hierarchy by enclosure
- Layering
  - Layering and alignment



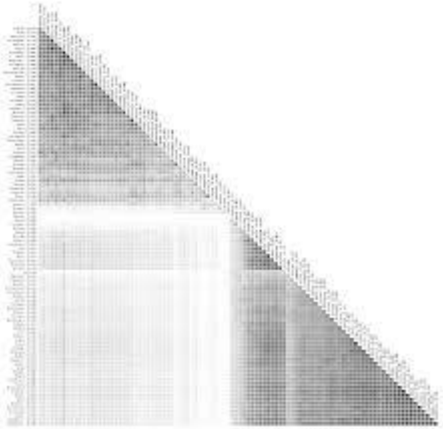


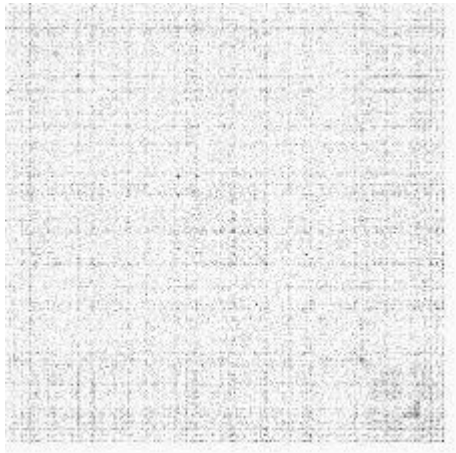
```
▼ sites
    ▼ all
        ▼ modules
            ▶ a-shared-module
          README.txt
        ▼ themes
            ▶ a-shared-theme
    ▼ default
          default.settings.php
        ▼ modules
            ▼ a-local-module
                  a-local-module.info
                  a-local-module.module
        ▼ themes
            ▼ a-custom-theme
                  a-custom-theme.info
                  style.css
```

Adjacency Matrix Visualization



	a	b	c	d	e	f	g	h
a		●	●			●		
b	●			●		●		
c	●				●		●	●
d	●	●				●		
e			●				●	●
f	●	●		●				
g			●		●			●
h			●		●		●	

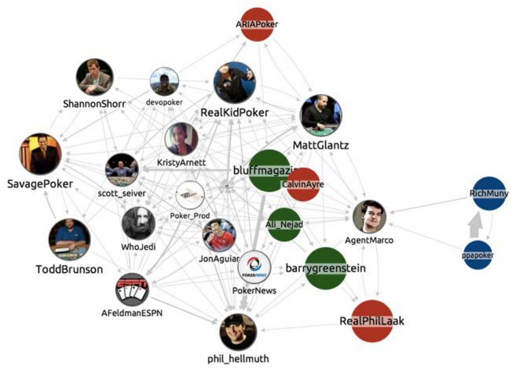




Visualizing Text

- Words are sparse and high-dimensional.
- Word Clouds
- Word Sequences (trees)
- Revision History
- Conversations (graphs)

readable
unaccountable
mind-map
comments
words
support
well
subject
corresponding
consultant
neurological
disappointing
web
manage
another
sources
accordingly
reading
introduction
best
Classic
people
contrary
Semantic
Atlantic
part
implementation
W3
Books
use
attention
meaning
making
annotate
job
plays
showing
computer
post
pains
scientific
technical
used
Firefox
explains
background
can
make
irrelevant
blog
visual
brain
draw
unconcentrated
Zotero
word-play
nice
discussion
sequential
will
Tell
article
just
deal
examples
standardized
sure
go
At
much
works
Approach
idea
Web
associations
artificial
intelligence
soon
literacy
rather
semantic
us
span
Zeix
i.e
way
stressing
really
piles
detail
think
something
clustering
gives
book
show
Part
seeking
printer
imprecise
areas
want
practical
previously
long
sustained
Search
look
results
Difficulties
difficulties
new
allusions
usability
impression
Cluuz
behavior
feel
printed
e.g
rules
mere
engine
Google
contrast
articles
data
books
information
objects
However
different
circuits
Powerset
ranking
completely
woven
many
published
author
format
far
structured
users
matter
enthusiastic
appreciate
one
incidentally
extensive



Takeaways

- The brain sees color, shape, size at different granularities and speeds
  - This affects our ability to distinguish between different parts of a graph
- Use the proper visualization with the good visual features to help a reader understand your graphs