CS 109: Data Science

Vis. Goals, Data Types, Statistical Graphs

Marc Streit
mstreit@seas.harvard.edu

Based on material created by Hanspeter Pfister

Marc Streit

mstreit@seas.harvard.edu

PhD in Computer Graphics from Graz University of Technology, Austria

Assistant Prof. at Johannes Kepler University Linz, Austria

Leading visualization group

Visiting Scholar @ HMS CMBI Park Lab in 2012

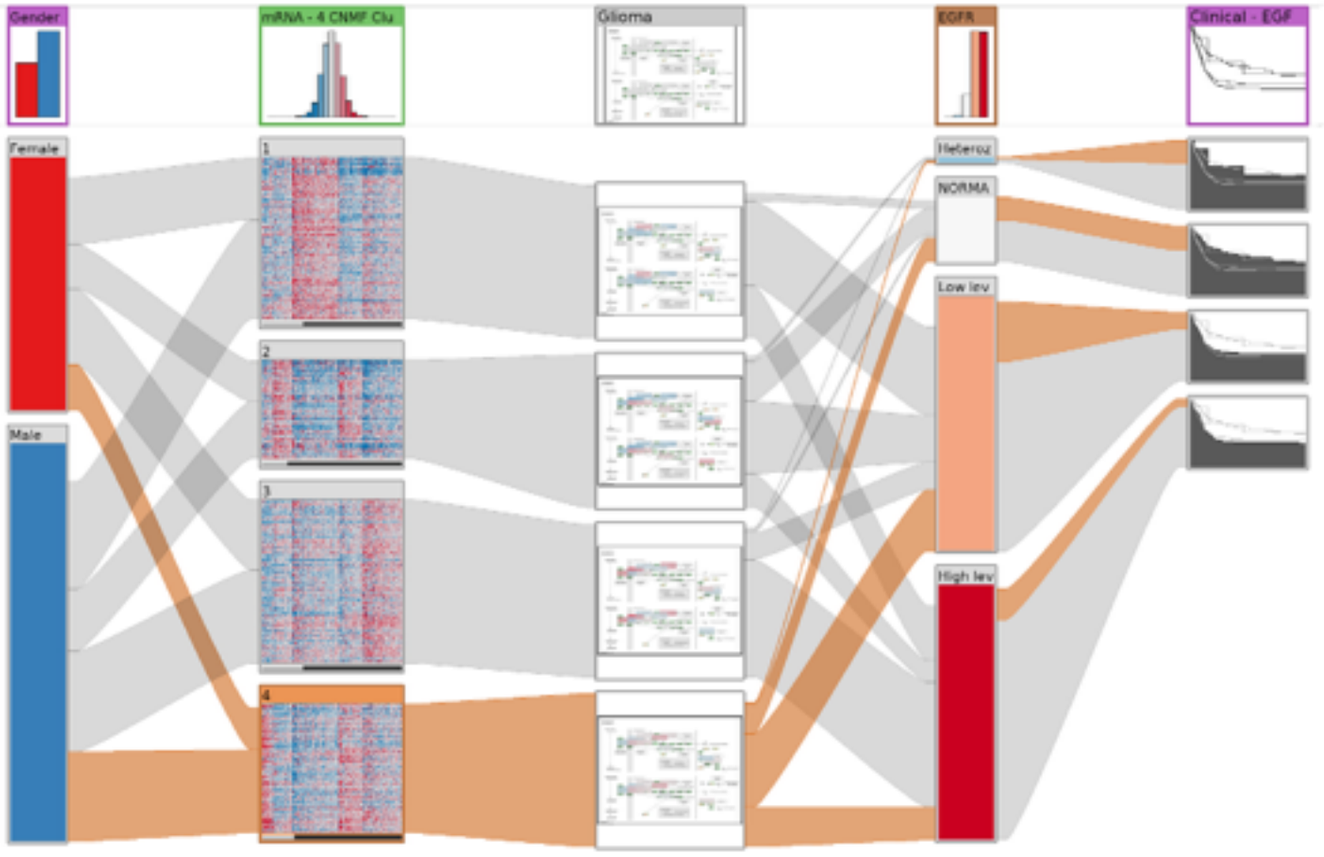
Visiting Professor @ Harvard SEAS Pfister Lab

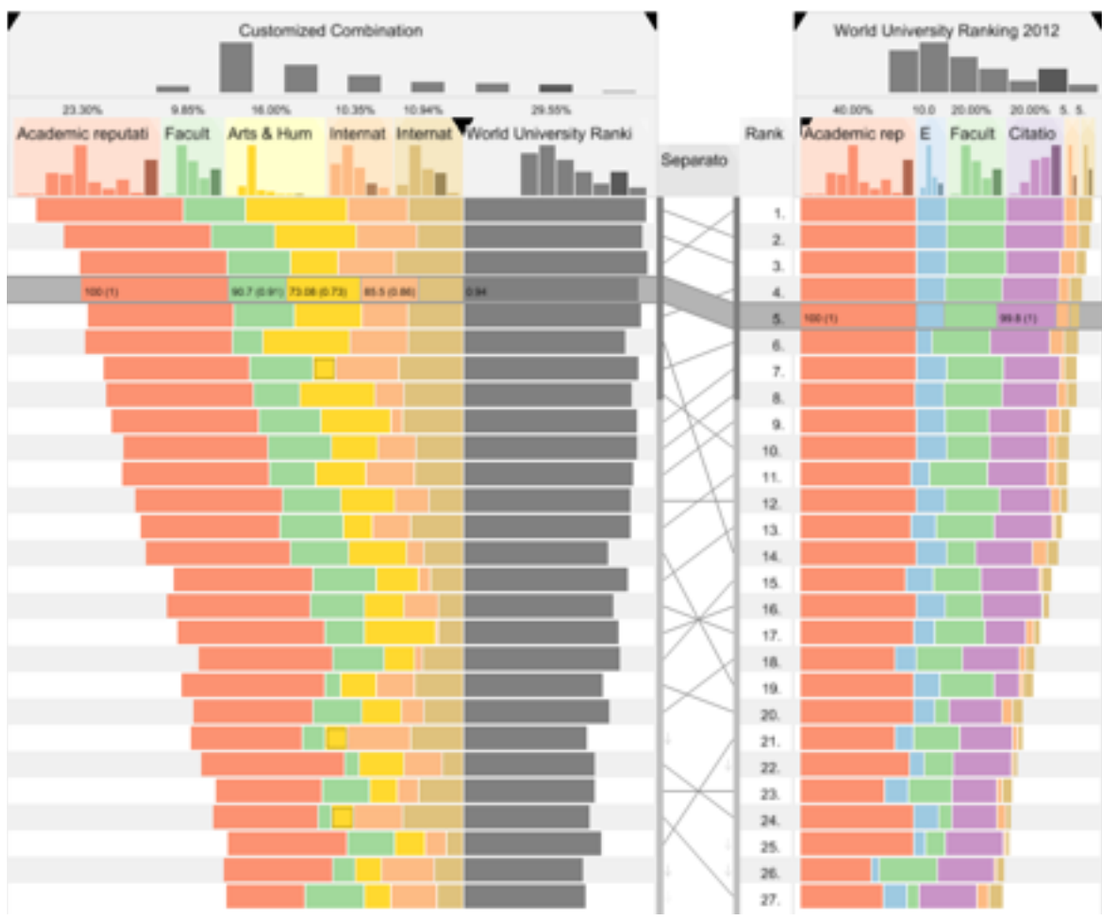
Research Focus

Information Visualization / Visual Analytics

Biological Data Visualization

Caleydo Framework (www.caleydo.org)





This Week

HW0 - due today Tuesday (not graded)

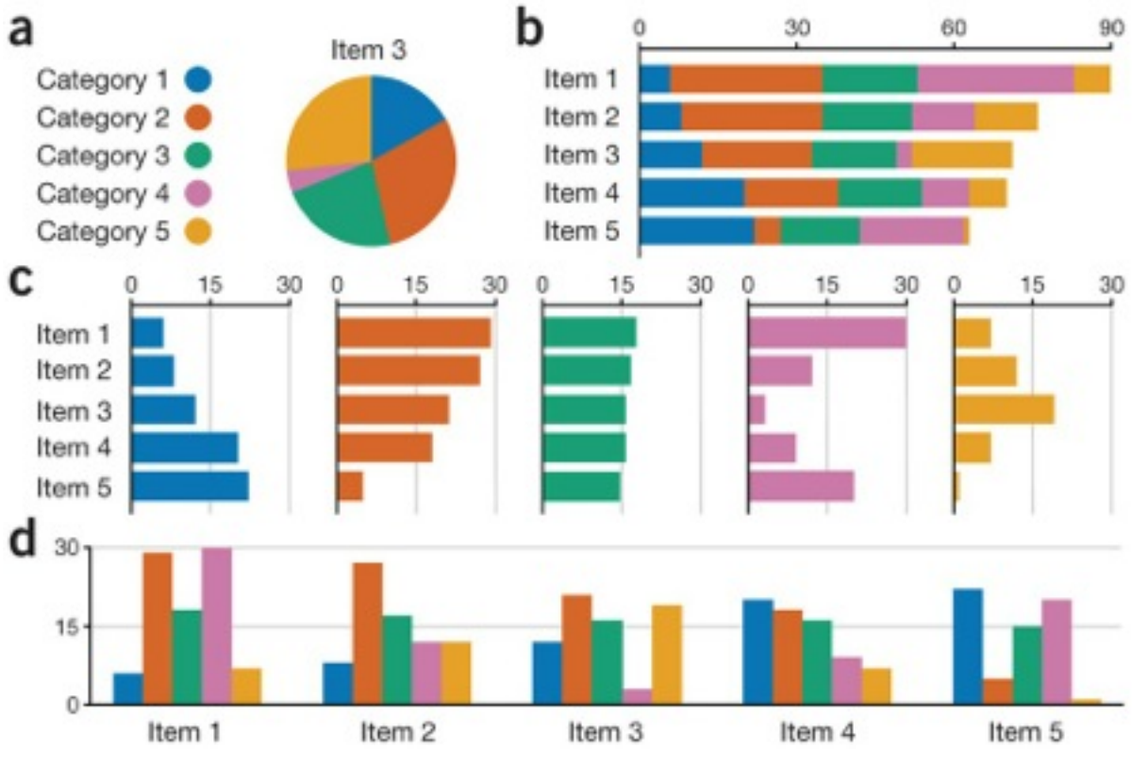
HW1 - due Thursday, Sept. 18 - start now!

Outline

Vis Goals

parch	fare	embarked	class	who
0	7.25	S	Third	man
0	71.2833	C	First	woman
0	7.925	S	Third	woman
0	53.1	S	First	woman
0	8.05	S	Third	man
0	8.4583	Q	Third	man
0	51.8625	S	First	man
1	21.075	S	Third	child
2	11.1333	S	Third	woman
0	30.0708	C	Second	child
1	16.7	S	Third	child
0	26.55	S	First	woman
0	8.05	S	Third	man
5	31.275	S	Third	man
0	7.8542	S	Third	child
0	16.0	S	Second	woman
1	29.125	Q	Third	child
0	13.0	S	Second	man
0	18.0	S	Third	woman
0	7.225	C	Third	woman
0	26.0	S	Second	man
0	13.0	S	Second	man
0	8.0292	Q	Third	child

Data Types
Dimensionality



Graph Types

Visualization Goals and Process

Last Week:

Anscombe's Quartet

Same mean, variance, correlation, and linear regression line

Anscombe's Quartet: Raw Data

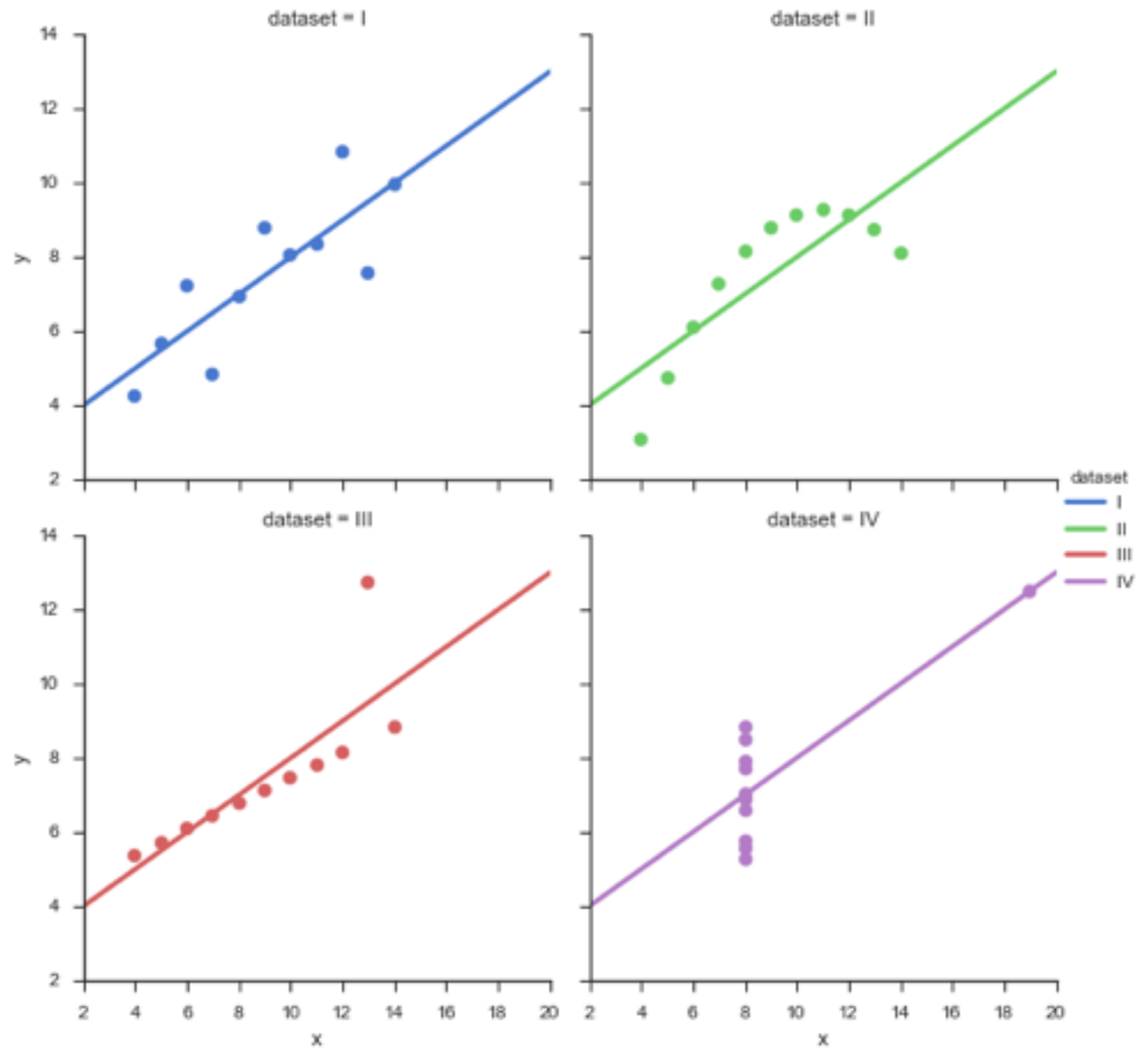
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.		0.816		0.816		0.816		0.816

Anscombe '73

Last Week:

Anscombe's Quartet

Same mean, variance, correlation, and linear regression line



http://web.stanford.edu/~mwaskom/software/seaborn/examples/anscombes_quartet.html

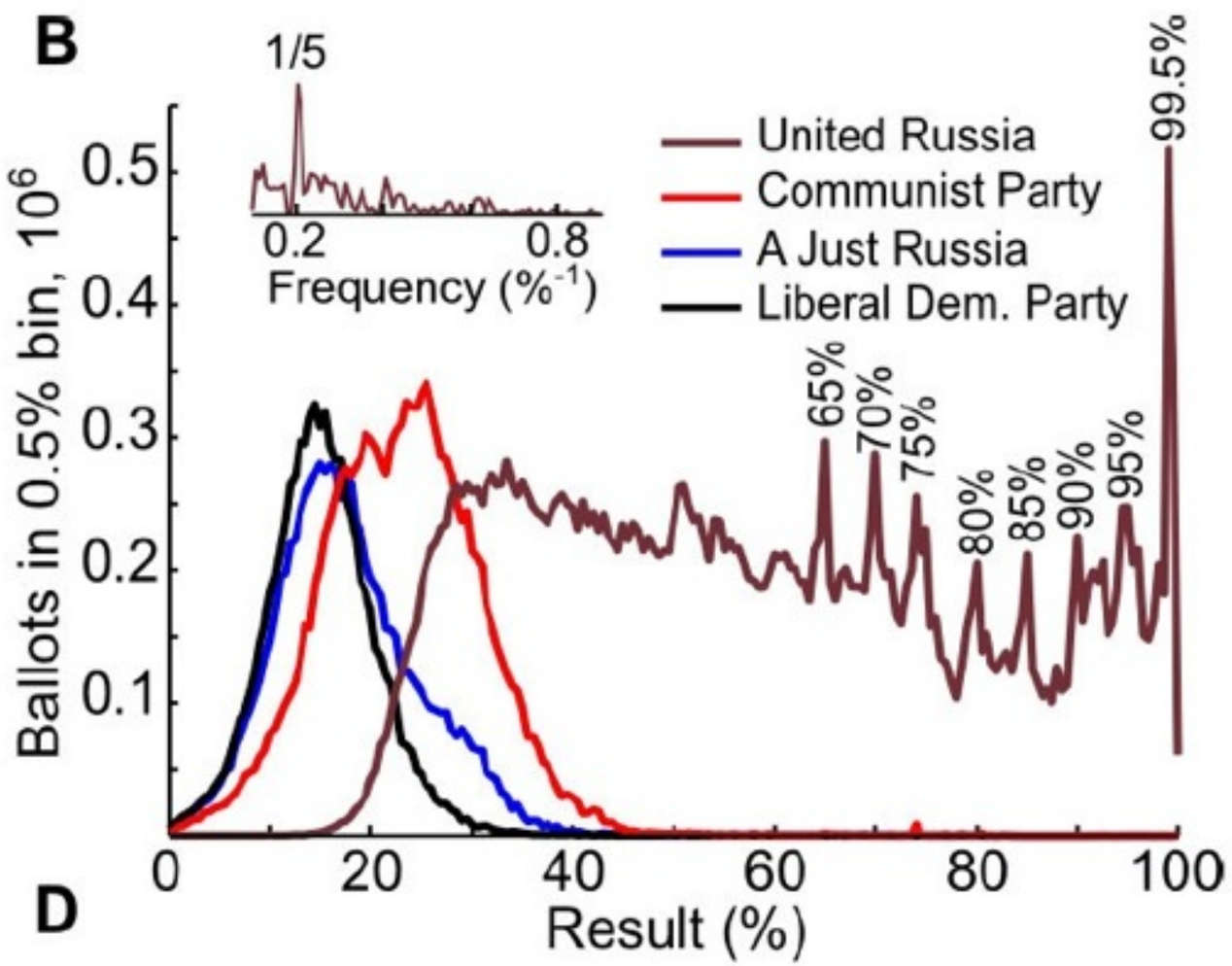
Anscombe '73

Last Week: Normal Distribution

Example: Russian President Elections

"We do not trust Churov
[the head of the committee]!
We trust Gauss"

Image from http://nl.livejournal.com/1082778.html



Visualization Goals

Presentation

Known facts about data

Task: Communicate results

Exploration

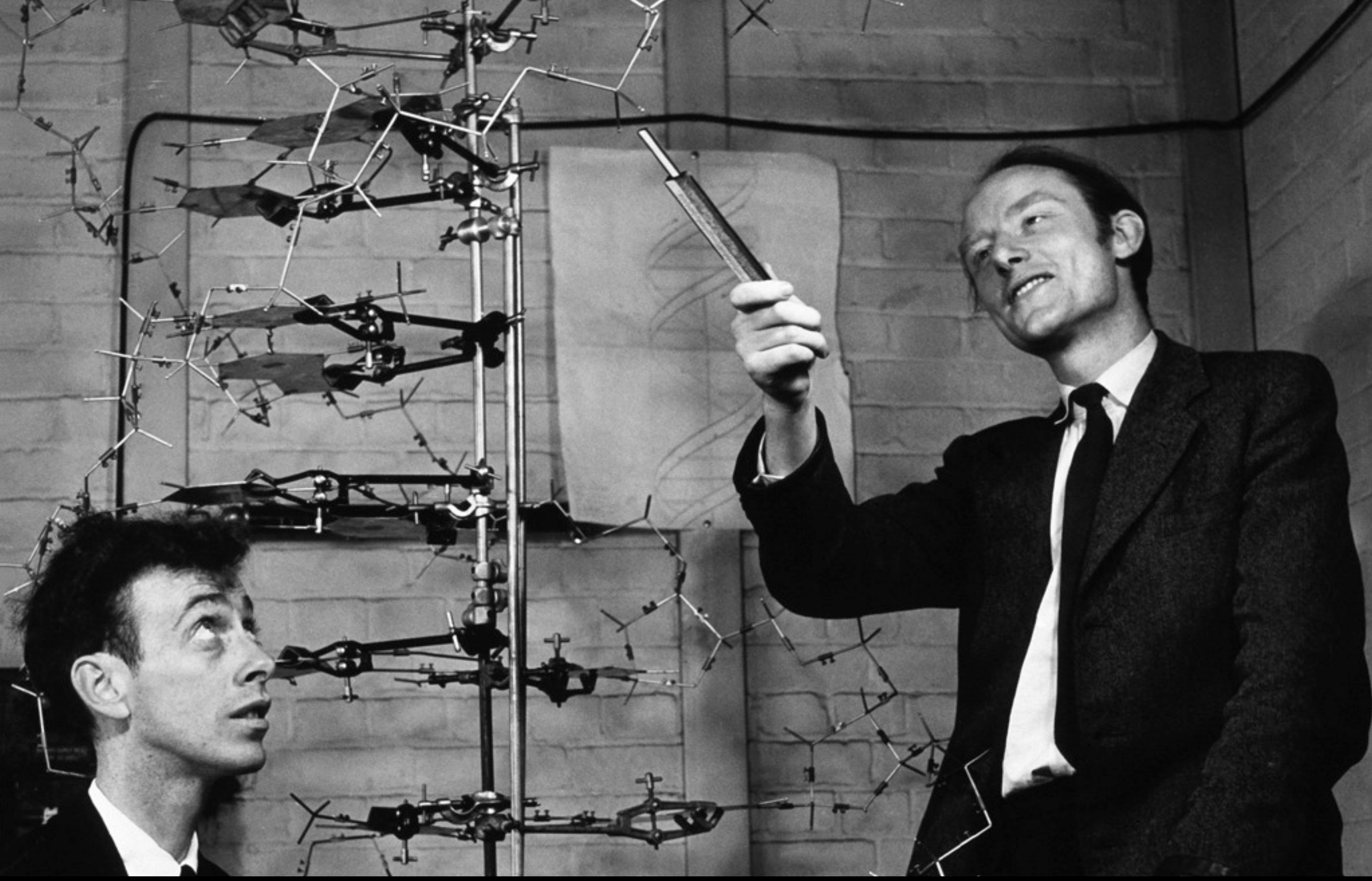
Data without hypothesis

Task: Generate hypothesis

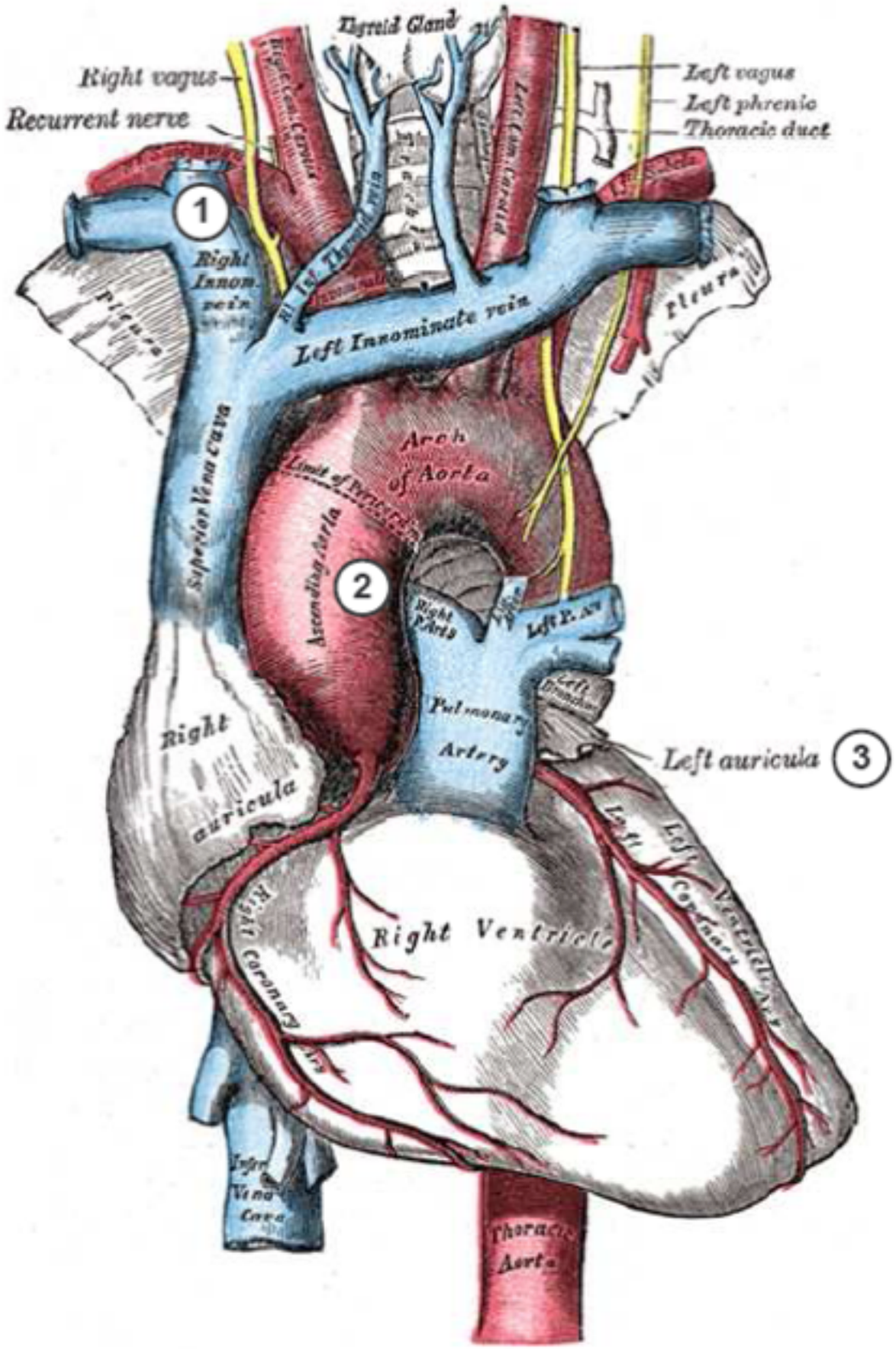
Confirmation

Hypothesis is given

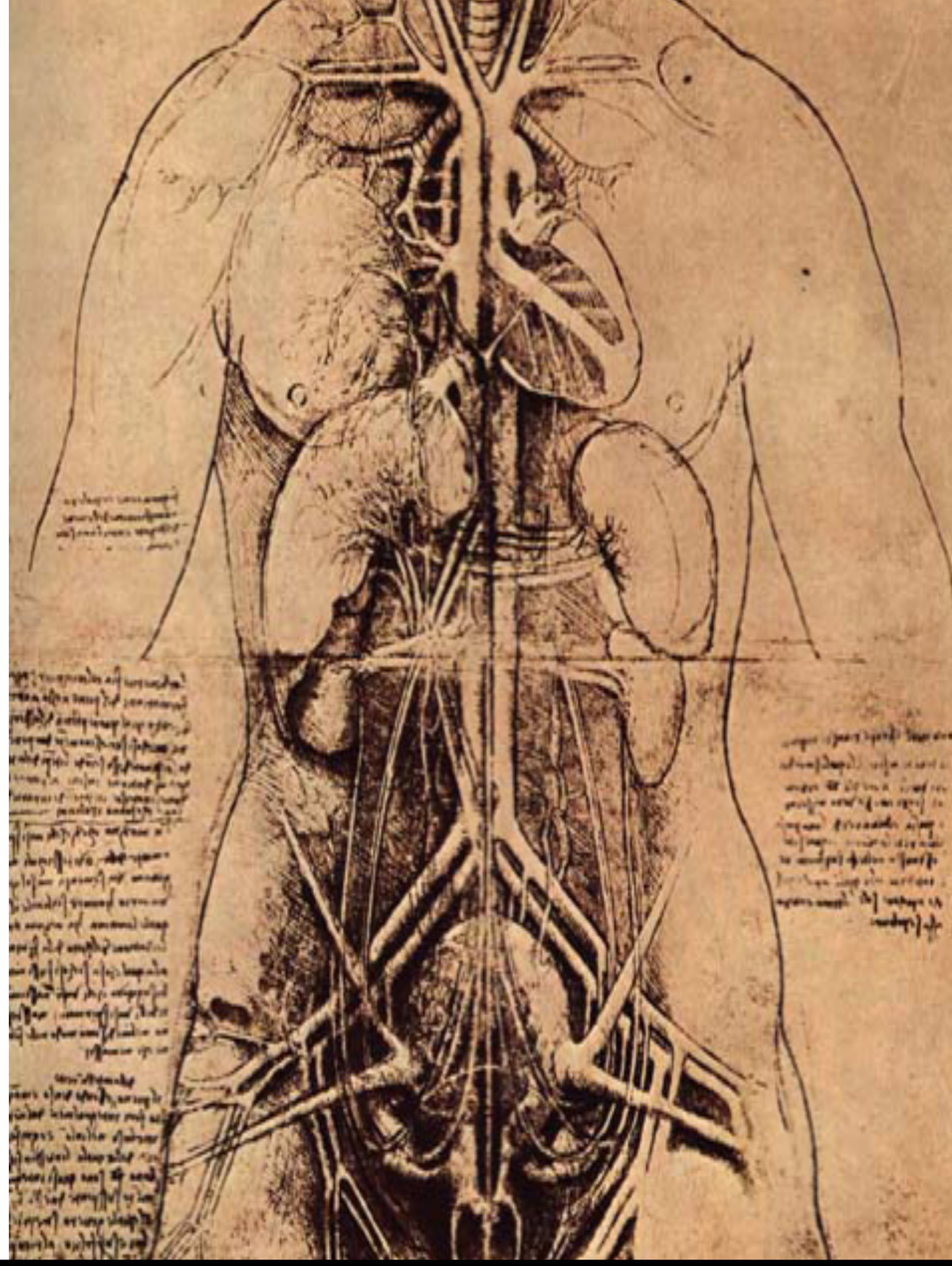
Task: Verify / falsify hypothesis



Presentation

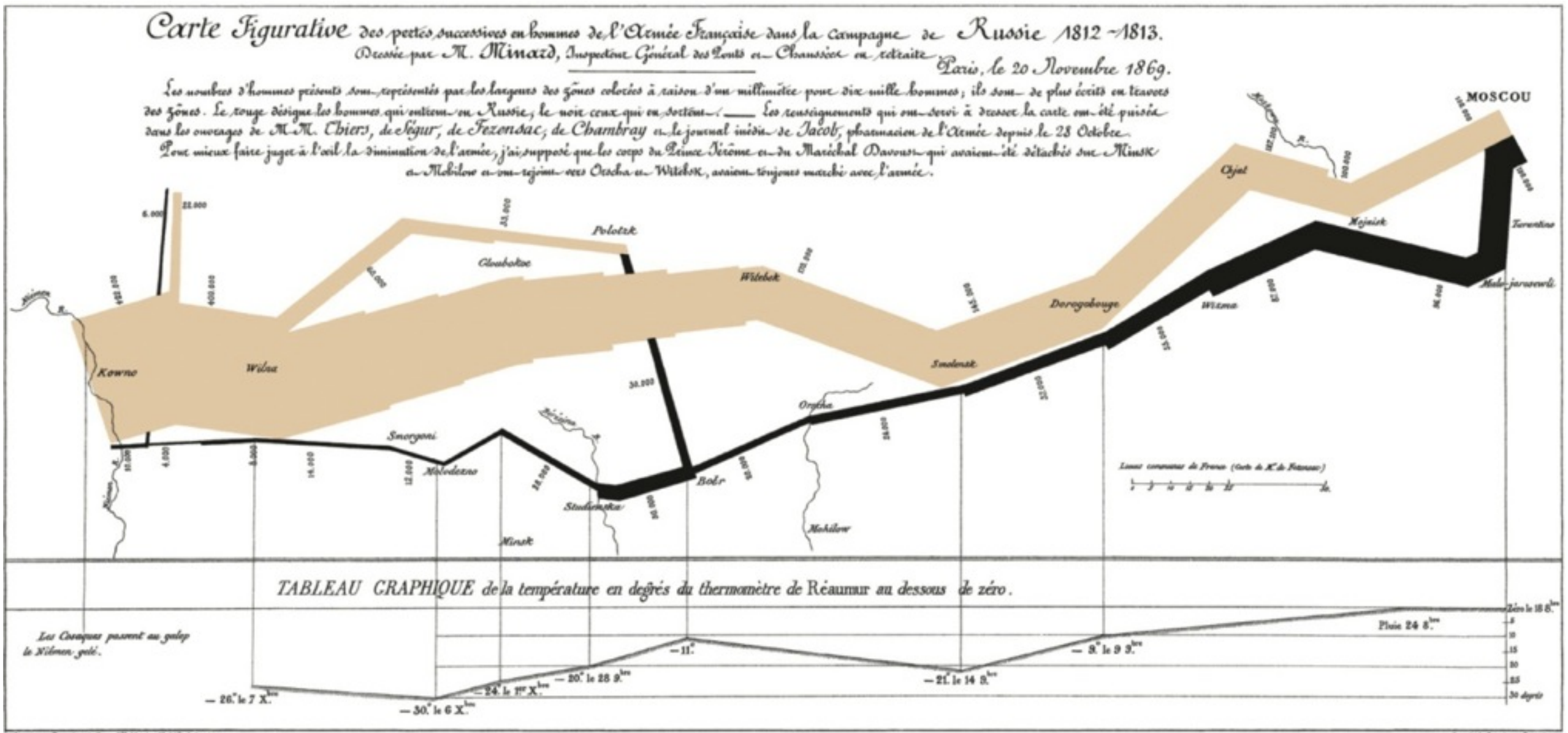


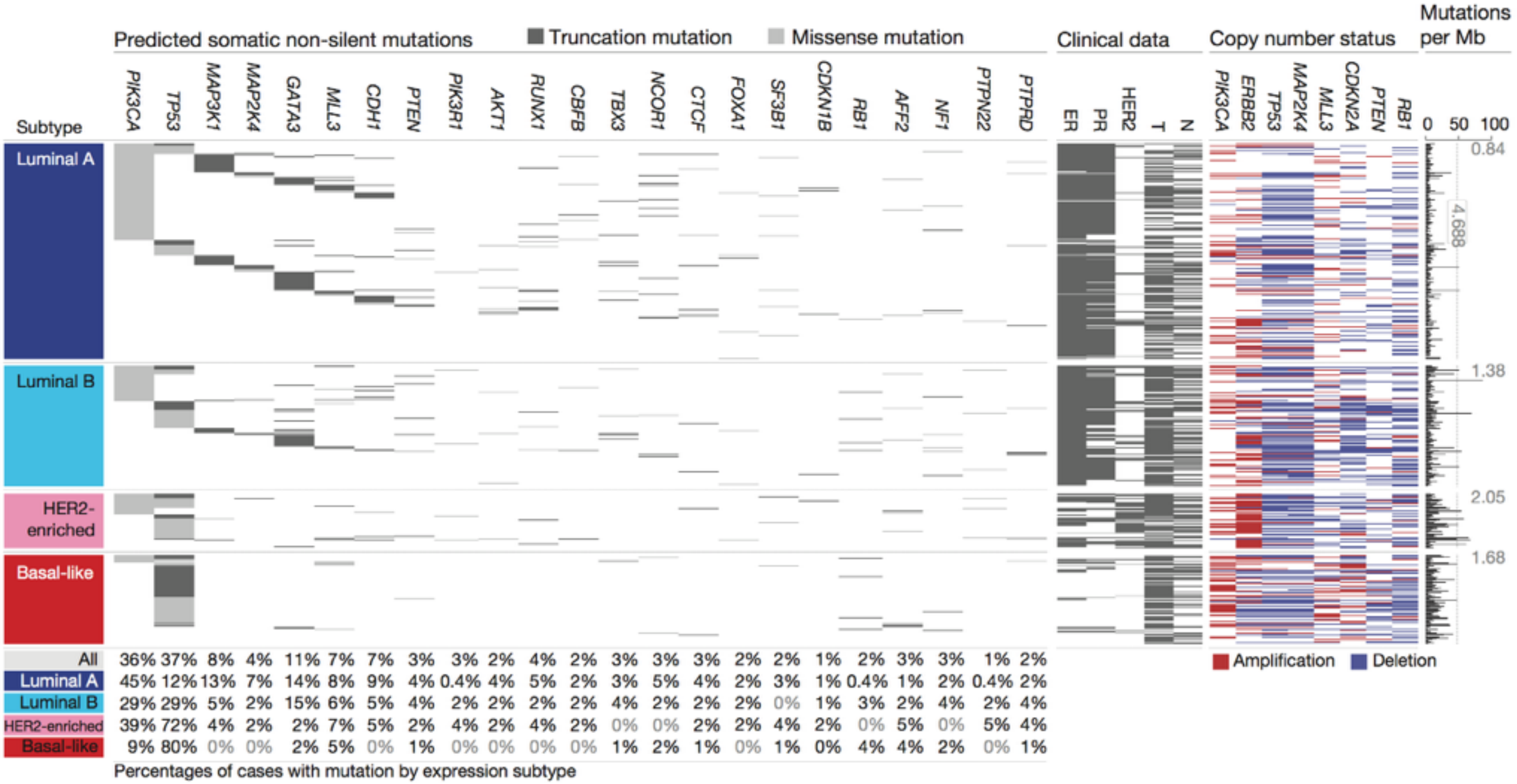
Henry Gray, 1918
„Anatomy of the Human Body“



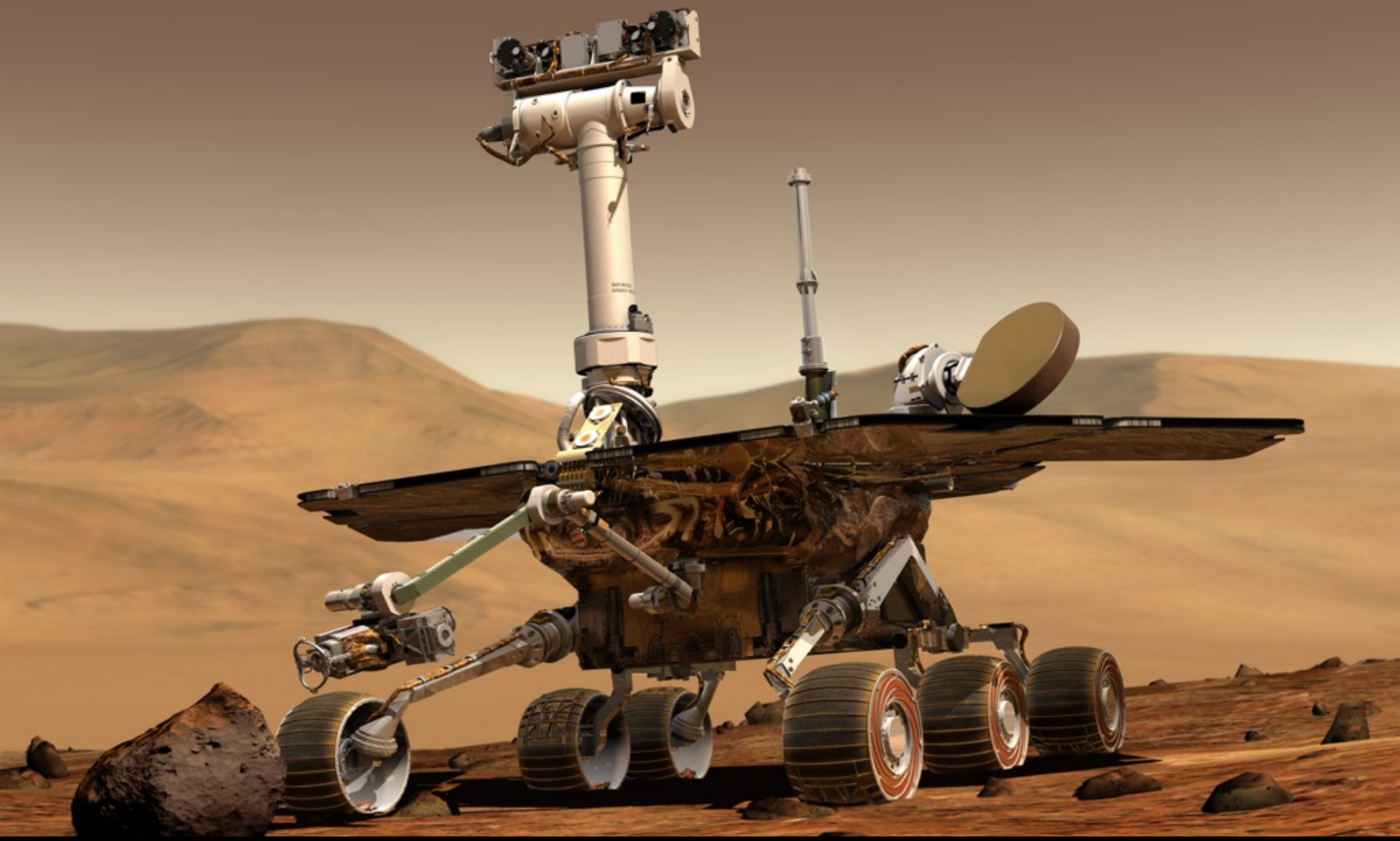
Drawing of a female body
Leonardo da Vinci, ~ 1510

Minard's Map
Napoleon's March on Moscow





Heterogeneous Heatmap
TCGA Paper, Nature 2012



Interactive

Exploration / Confirmation

Exploratory Data Analysis (EDA)

"The greatest value of a picture is when it forces us to notice what we never expected to see."

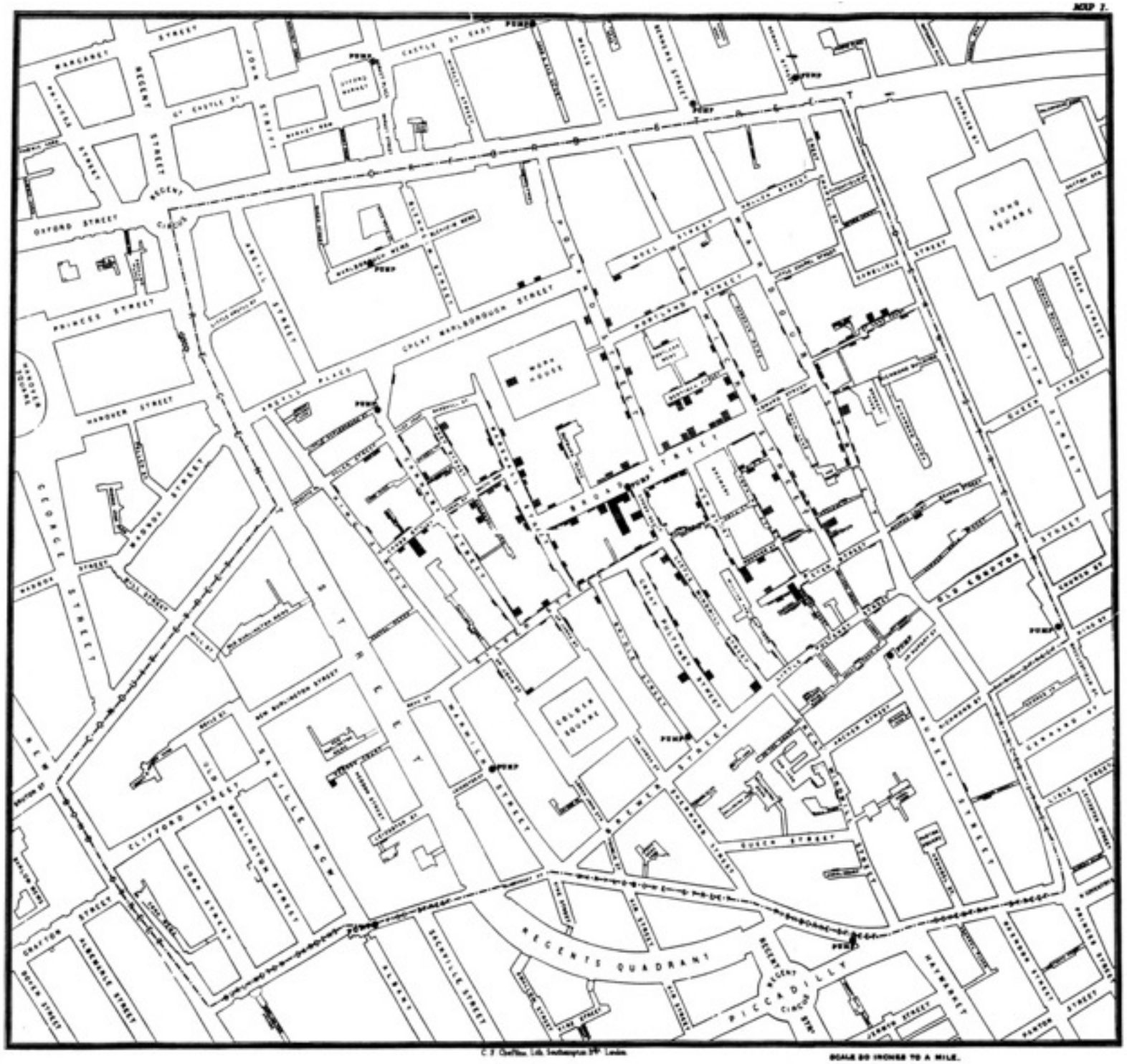
-John Tukey (1915 - 2000)

EDA Definition

An approach of analyzing data to summarize their main characteristics without using a statistical model or having formulated a prior hypothesis.

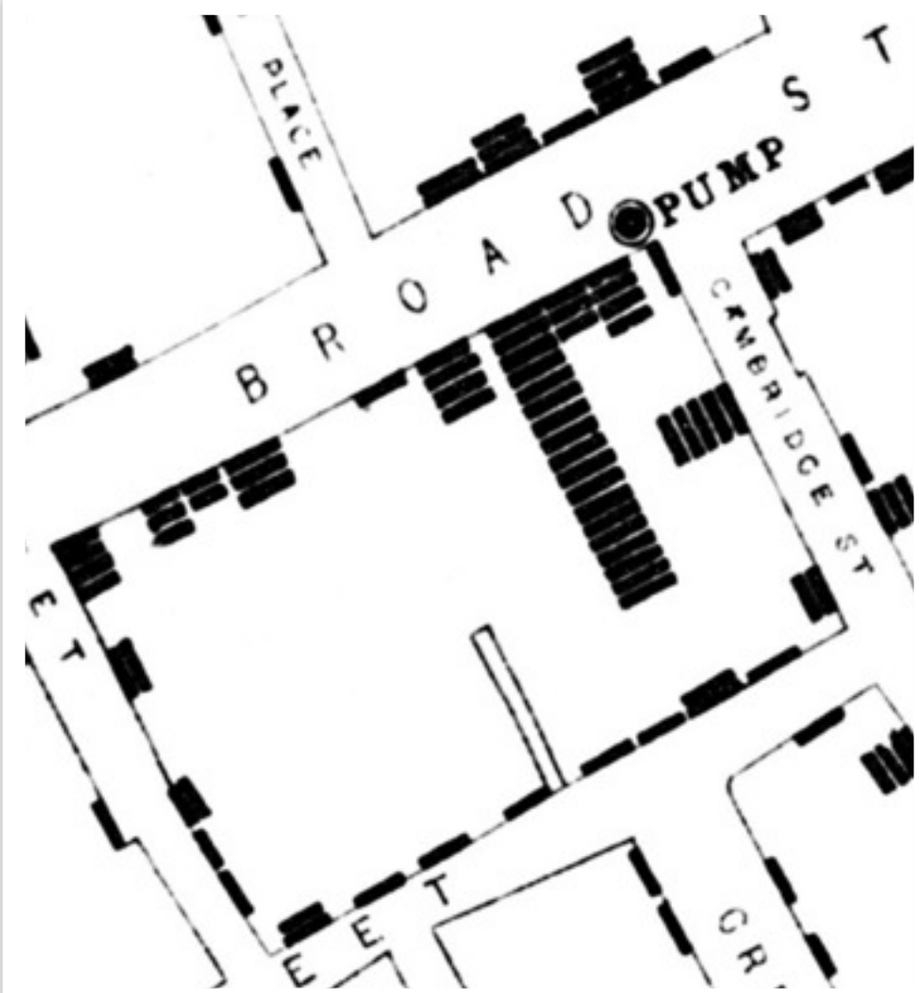
Exploratory data analysis was promoted by John Tukey to encourage statisticians to **visually examine their datasets**, to **formulate hypotheses** that could be tested on new datasets. [wikipedia]

Detect the expected – discover the unexpected



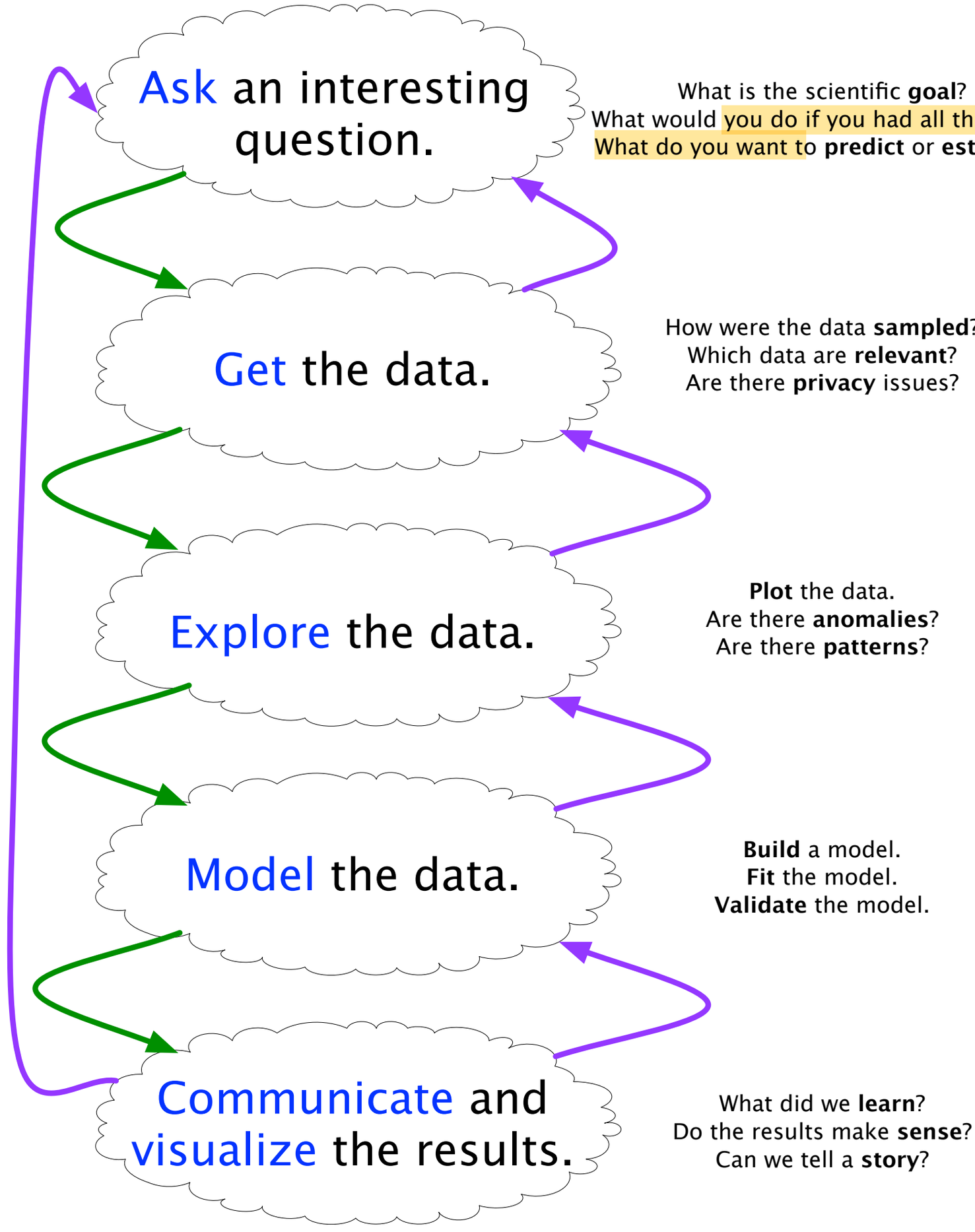
Wikimedia Commons

John Snow
(1813 – 1858)



The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka” but “That’s funny...”

- Isaac Asimov (1920–1992)


Ask an interesting question.
What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?
Get the data.
How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?
Explore the data.
Plot the data.
Are there **anomalies**?
Are there **patterns**?
Model the data.
Build a model.
Fit the model.
Validate the model.
Communicate and visualize the results.
What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Example: Antibiotics

Will Burtin, 1951

Effectiveness of Antibiotics

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

Data & Questions

What are the data types?

What are possible questions?

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

Data

Antibiotic name [string]

Gram staining? [pos/neg]

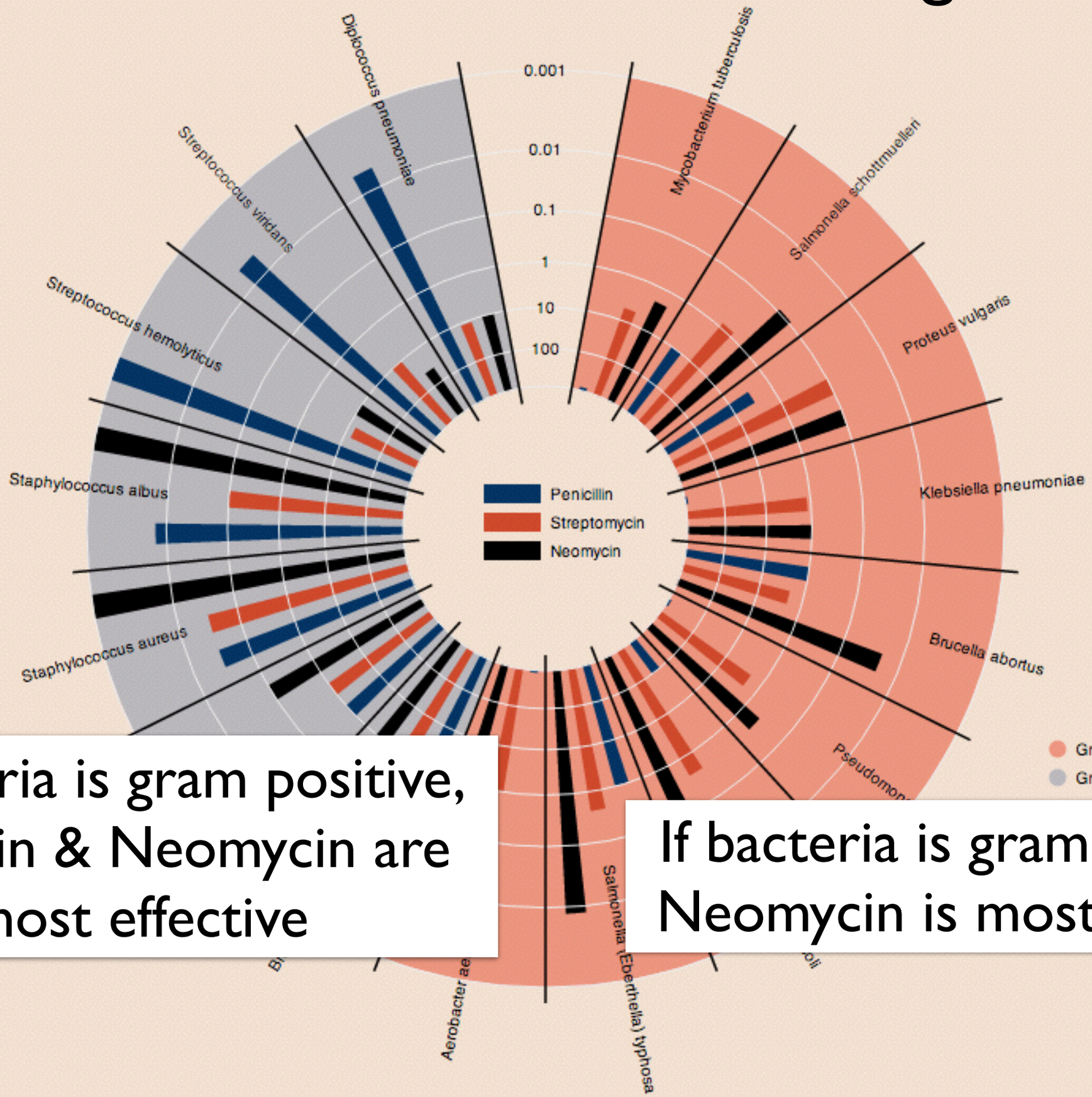
Minimum inhibitory concentration (mg/ml) [float]
(lower == more effective)

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+

What Questions?

Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

How effective are the drugs?

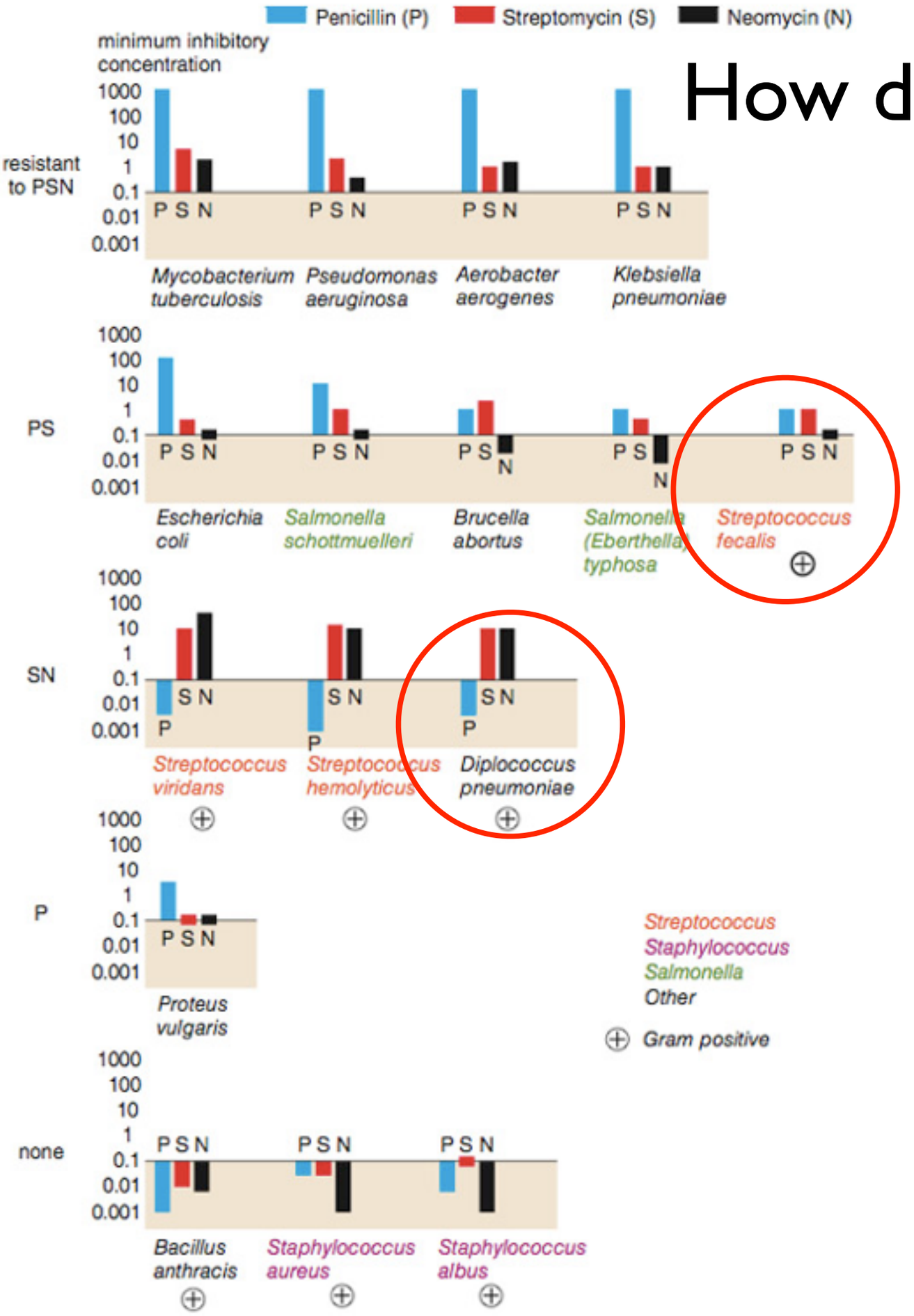


If bacteria is gram positive, Penicillin & Neomycin are most effective

If bacteria is gram negative, Neomycin is most effective

M. Bostock, Protovis after W. Burtin, 1951

How do the bacteria compare?

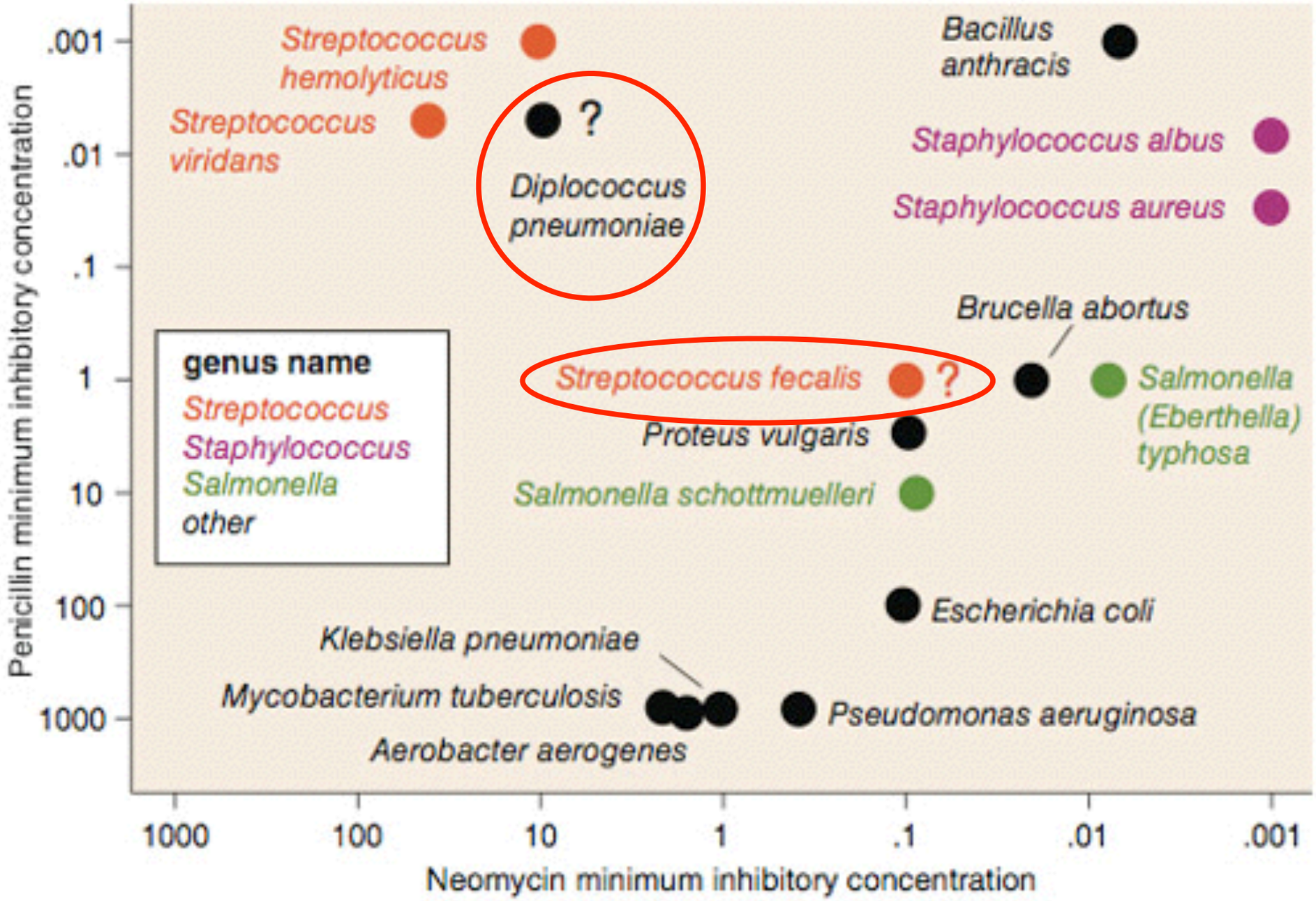


Not a streptococcus!
(realized ~30 years later)

Really a streptococcus!
(realized ~20 years later)

Wainer & Lysen, "That's funny..."
American Scientist, 2009
Adapted from Brian Schmotzer

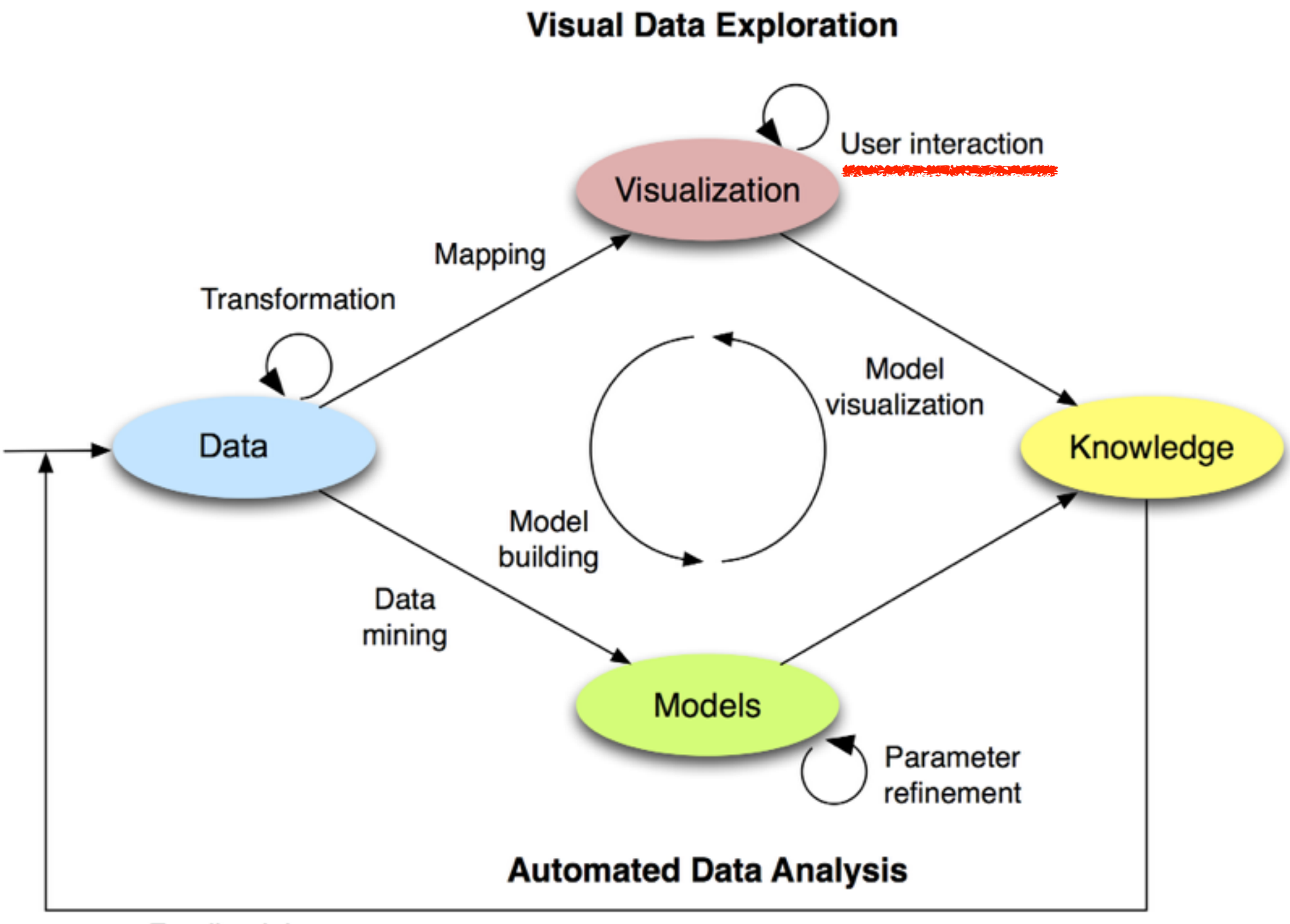
How do the bacteria compare?



Wainer & Lysen, "That's funny..."
American Scientist, 2009

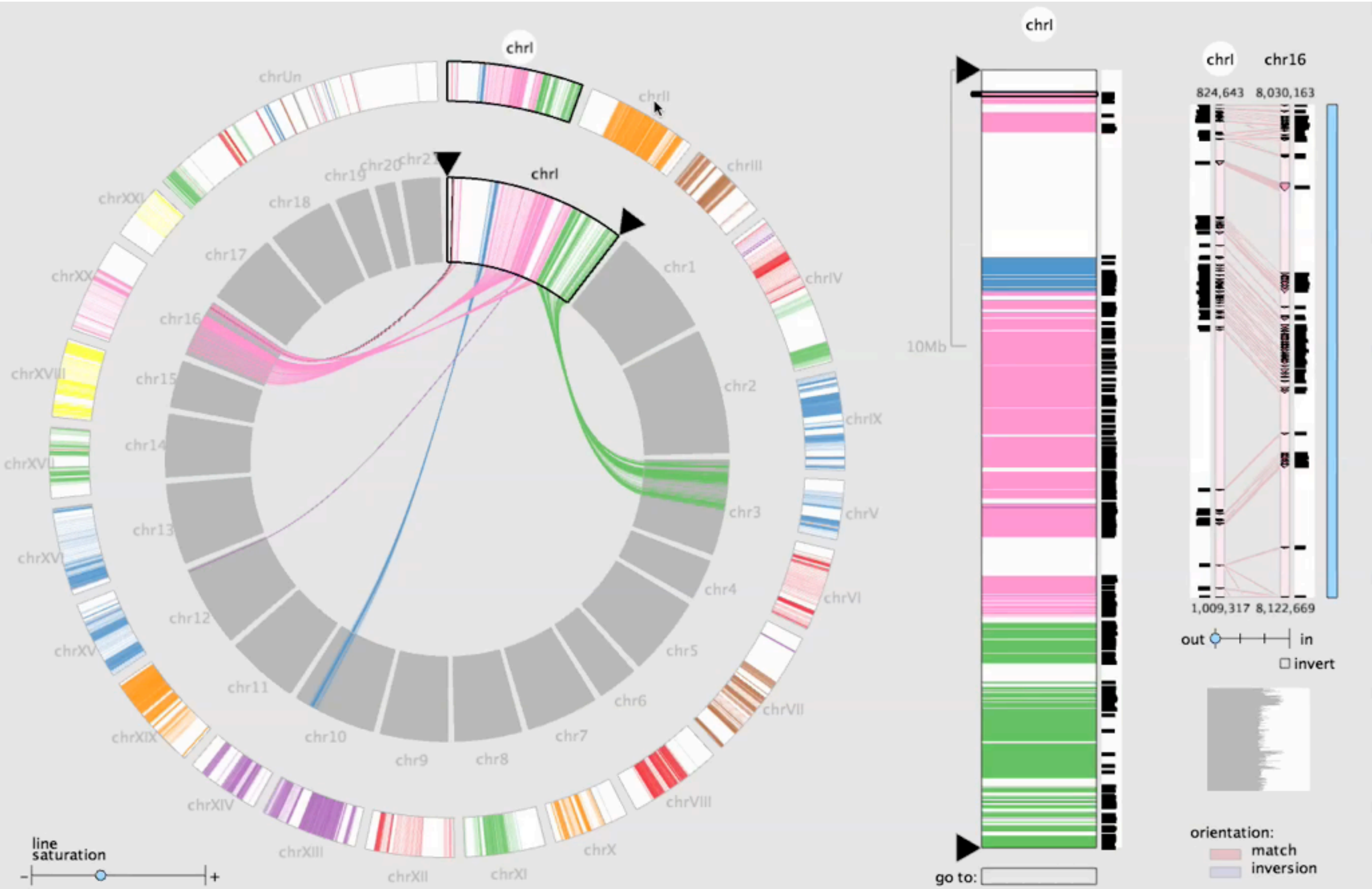
So far everything was static…

Visual Analytics Process



Keim, et al. 2010

Example: MizBee [Meyer et al. 2009]



Data Types

Ben Shneiderman, 1996

1D (sequences)

Temporal

2D (maps)

3D (shaped)

nD (relational)

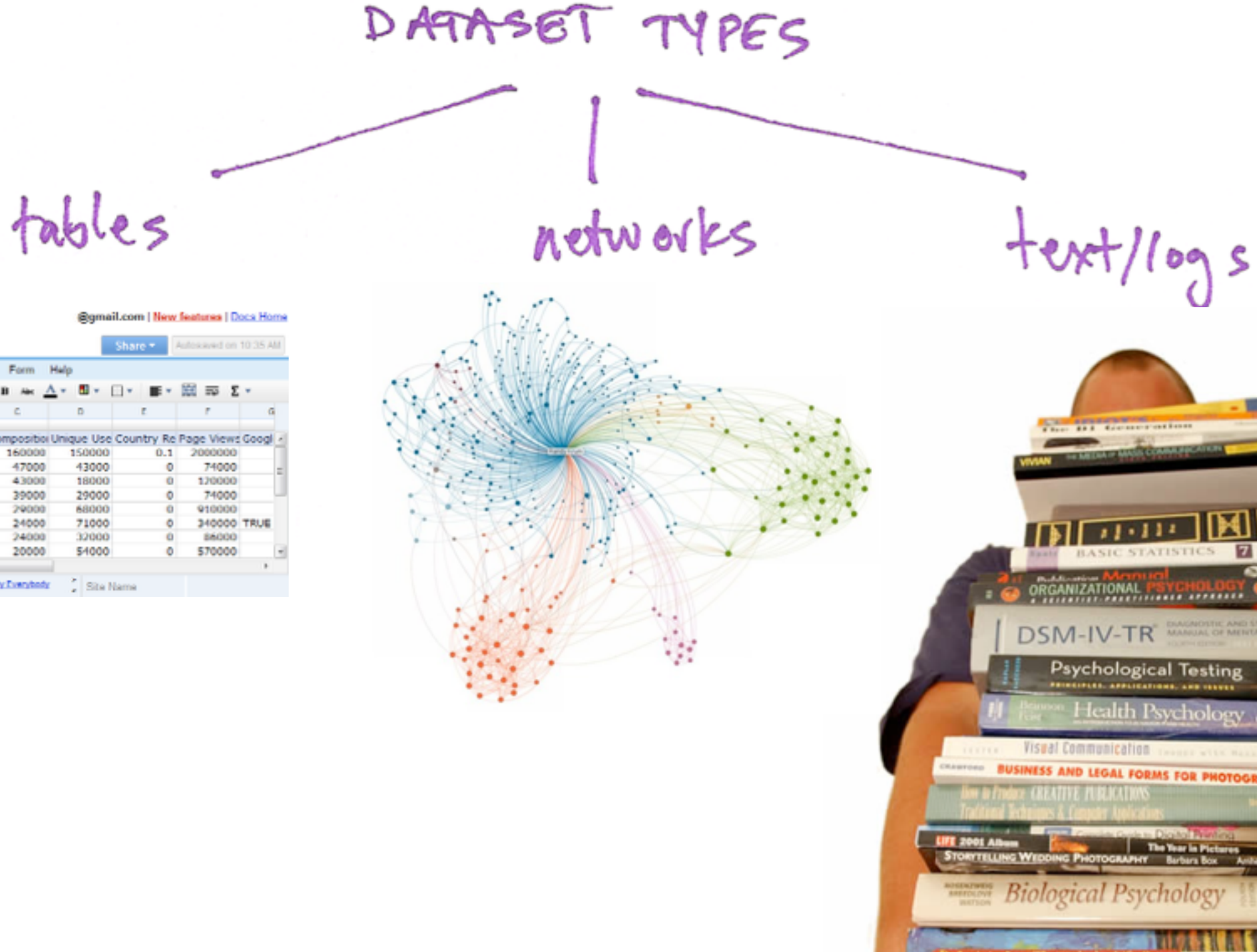
Trees (hierarchical)

Networks (graphs)

Others?

The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization [Shneiderman, 96]

Tamara Munzner, 2013



Semantics vs. Types

Data Semantics: The real-world meaning

e.g., company name, day of the month, person height, etc.

Data Type: Interpretation in terms of scales of measurements

e.g., quantity or category, sensible mathematical operations, data structure, etc.

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

FOR SEVEN YEARS A COMMITTEE of the British Association for the Advancement of Science debated the problem of measurement. Appointed in 1932 to represent Section A (Mathematical and Physical Sciences) and Section J (Psychology), the committee was instructed to consider and report upon the possibility of "quantitative estimates of sensory events"—meaning simply: Is it possible to measure human sensation? Deliberation led only to disagreement, mainly about what is meant by the term measurement. An interim report in 1938 found one member complaining that his colleagues

by the formal (mathematical) properties of the scales. Furthermore—and this is of great concern to several of the sciences—the statistical manipulations that can legitimately be applied to empirical data depend upon the type of scale against which the data are ordered.

A Classification of Scales of Measurement

Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads

Stevens' 4 scales of measurements

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

On the theory of scales and measurements [S. Stevens, 46]

Data Types

Nominal (Categorical) (N)

Are = or ≠ to other values

Apples, oranges, bananas, ...

Ordinal (O)

Obey a < relationship

Small, medium, large

Quantitative (Q)

Can do arithmetic on them

10 inches, 23 inches, etc.

On the theory of scales and measurements [S. Stevens, 46]

Data Types

Q - Interval (location of zero arbitrary)

Dates: Jan 19; Location: (Lat, Long)

Like a geometric point. Cannot compare directly.

Only differences (i.e., intervals) can be compared

Q - Ratio (zero fixed)

Measurements: Length, Mass, Temp, ...

Origin is meaningful, can measure ratios & proportions

Like a geometric vector, origin is meaningful

On the theory of scales and measurements [S. Stevens, 46]

Data Types

N - Nominal (labels)

Operations: =, ≠

O - Ordinal (ordered)

Operations: =, ≠, >, <

Q - Interval (location of zero arbitrary)

Operations: =, ≠, >, <, +, − (distance)

Q - Ratio (zero fixed)

Operations: =, ≠, >, <, +, −, ×, ÷ (proportions)

On the theory of scales and measurements [S. Stevens, 46]

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Example: Titanic Dataset

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third				Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First				Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Semantics

Example: Titanic Dataset

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First				Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third				Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Item

Example: Titanic Dataset

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0	1	1	16.7	S	Third	child	False	G	Southampton	yes	False
1	1	female	58.0	0	0	26.55	S	First	woman	False	C	Southampton	yes	True
0	3	male	20.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male	39.0	1	5	31.275	S	Third	man	True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Attribute
aka
Feature

Example: Titanic Dataset

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0					False		Cherbourg	yes	False
1	3	female	4.0	1	1					False	G	Southampton	yes	False
1	1	female	58.0	0	0				n	False	C	Southampton	yes	True
0	3	male	20.0	0	0					True		Southampton	no	True
0	3	male	39.0	1	5					True		Southampton	no	False
0	3	female	14.0	0	0	7.8542	S	Third	child	False		Southampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

1 = Quantitative

2 = Nominal

3 = Ordinal

survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	1	female	35.0	1	0	53.1	S	First	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Third	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Third	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	First	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Third	child	False		Southampton	no	False
1	3	female	27.0	0	2	11.1333	S	Third	woman	False		Southampton	yes	False
1	2	female	14.0	1	0	30.0708	C	Second	child	False		Cherbourg	yes	False
1	3	female	4.0									outhampton	yes	False
1	1	female	58.									outhampton	yes	True
0	3	male	20.									outhampton	no	True
0	3	male	39.									outhampton	no	False
0	3	female	14.									outhampton	no	True
1	2	female	55.0	0	0	16.0	S	Second	woman	False		Southampton	yes	True
0	3	male	2.0	4	1	29.125	Q	Third	child	False		Queenstown	no	False
1	2	male		0	0	13.0	S	Second	man	True		Southampton	yes	True
0	3	female	31.0	1	0	18.0	S	Third	woman	False		Southampton	no	False
1	3	female		0	0	7.225	C	Third	woman	False		Cherbourg	yes	True
0	2	male	35.0	0	0	26.0	S	Second	man	True		Southampton	no	True
1	2	male	34.0	0	0	13.0	S	Second	man	True	D	Southampton	yes	True
1	3	female	15.0	0	0	8.0292	Q	Third	child	False		Queenstown	yes	True

Nominal / Ordinal = Dimensions

Describe the data, independent variables

Quantitative = Measures

Numbers to be analyzed, dependent variables

Data vs. Conceptual Model

Data Model: Low-level description of the data

Set with operations, e.g., floats with +, -, /, *

Conceptual Model: Mental construction

Includes semantics, supports reasoning

Data	Conceptual
1D floats	temperature
3D vector of floats	space

Data vs. Conceptual Model

From data model...

32.5, 54.0, -17.3, … (floats)

using conceptual model...

Temperature

to data type

Continuous to 4 significant figures (Q)

Hot, warm, cold (O)

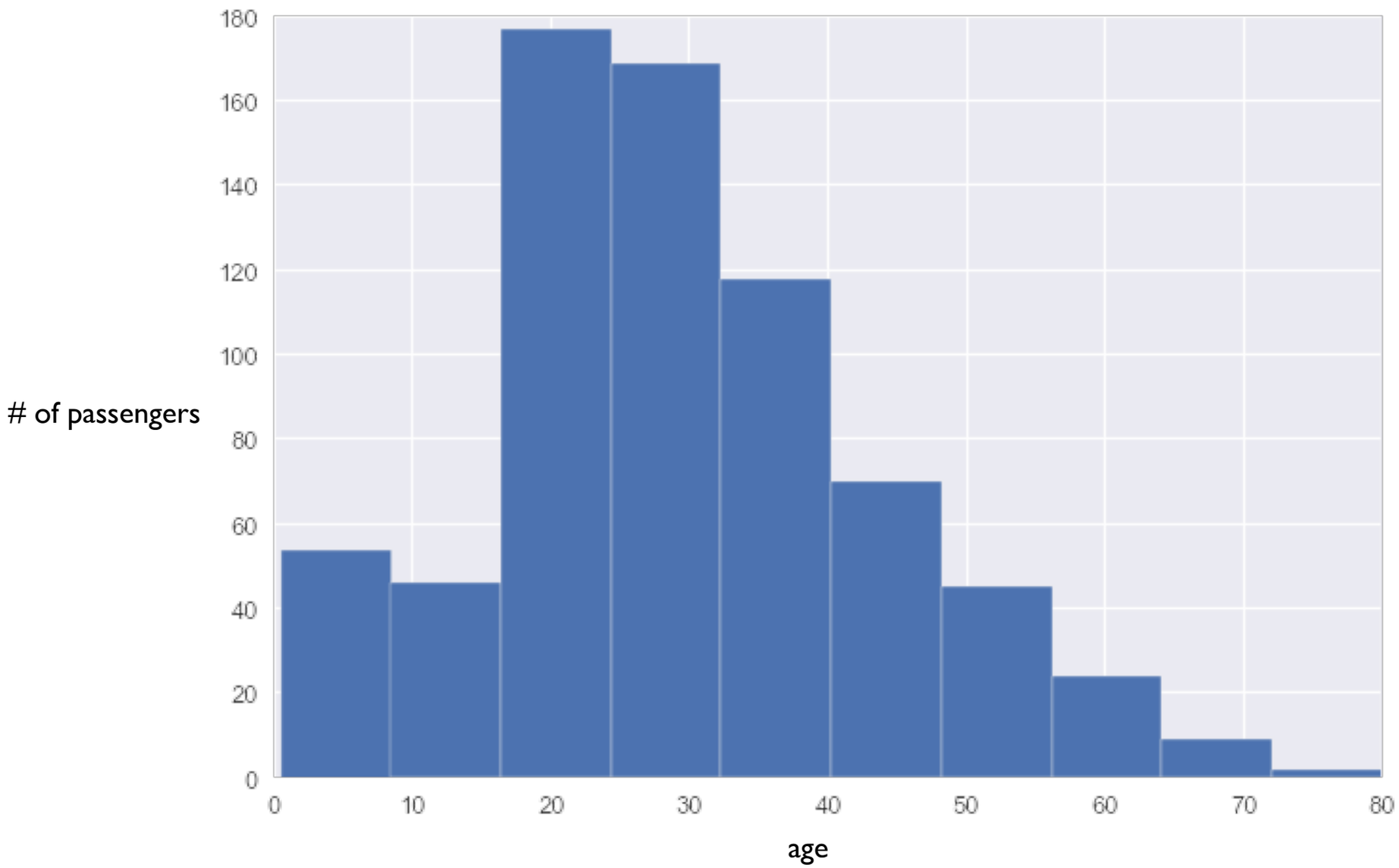
Burned vs. Not burned (N)

Based on slide from Munzner

Data Dimensions

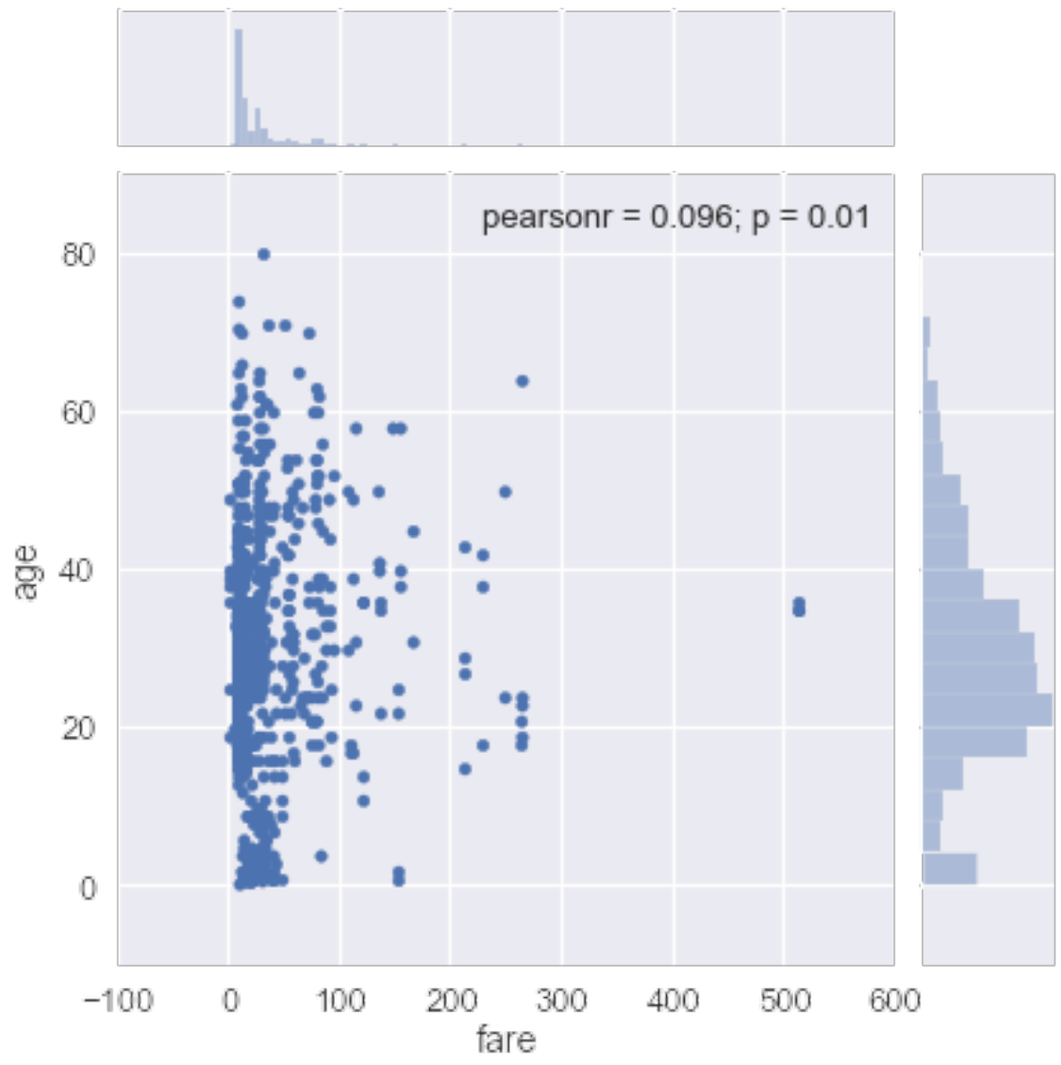
Univariate Data

age
22.0
38.0
26.0
35.0
35.0
54.0
2.0
27.0
14.0
4.0
58.0
20.0
39.0
14.0
55.0
2.0
31.0
35.0
34.0
15.0



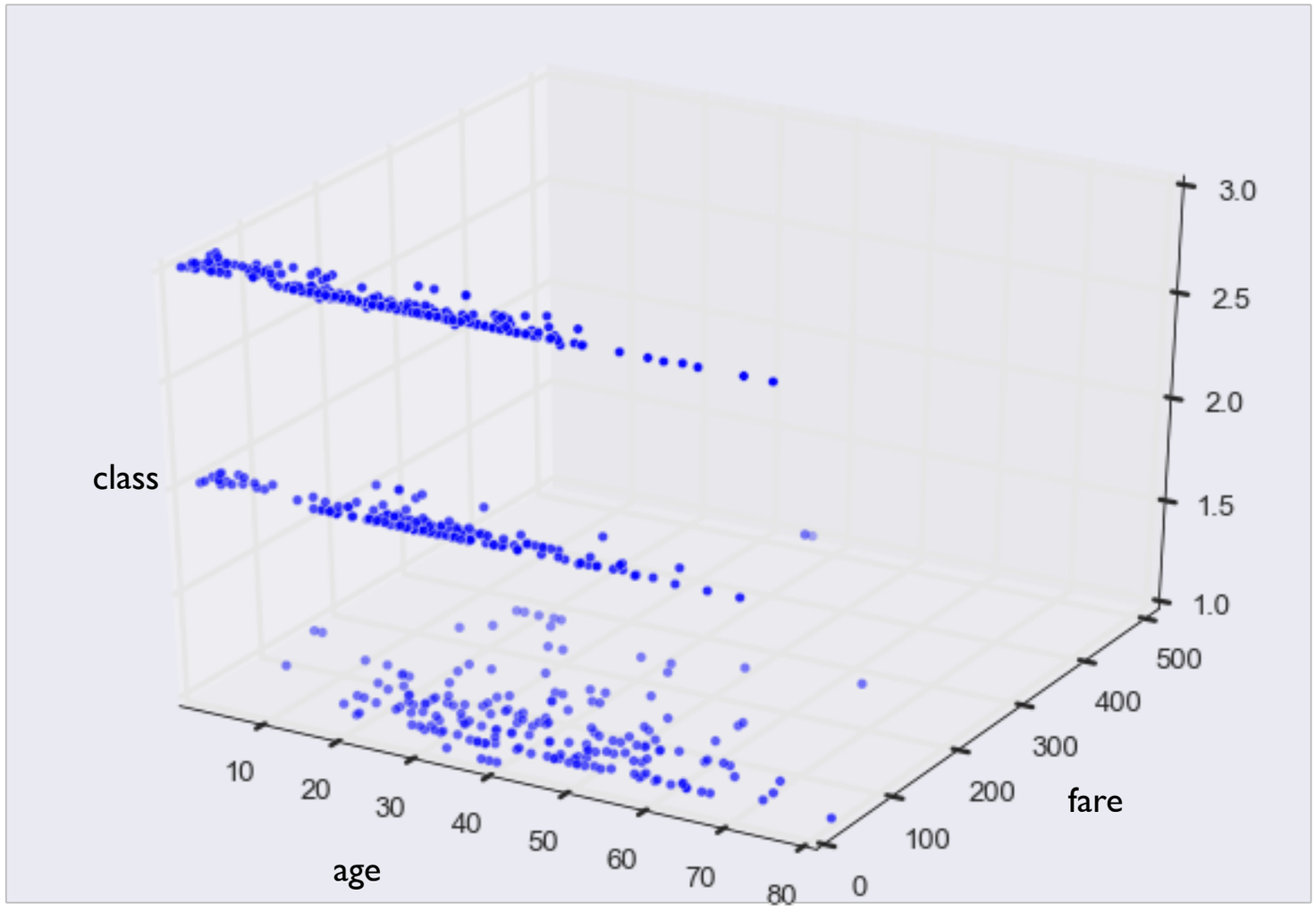
age	fare
22.0	7.25
38.0	71.2833
26.0	7.925
35.0	53.1
35.0	8.05
	8.4583
54.0	51.8625
2.0	21.075
27.0	11.1333
14.0	30.0708
4.0	16.7
58.0	26.55
20.0	8.05
39.0	31.275
14.0	7.8542
55.0	16.0
2.0	29.125
	13.0
31.0	18.0
	7.225
35.0	26.0
34.0	13.0
15.0	8.0292

Bivariate Data



age	fare	class
22.0	7.25	Third
38.0	71.2833	First
26.0	7.925	Third
35.0	53.1	First
35.0	8.05	Third
	8.4583	Third
54.0	51.8625	First
2.0	21.075	Third
27.0	11.1333	Third
14.0	30.0708	Second
4.0	16.7	Third
58.0	26.55	First
20.0	8.05	Third
39.0	31.275	Third
14.0	7.8542	Third
55.0	16.0	Second
2.0	29.125	Third
	13.0	Second
31.0	18.0	Third
	7.225	Third
35.0	26.0	Second
34.0	13.0	Second
15.0	8.0292	Third

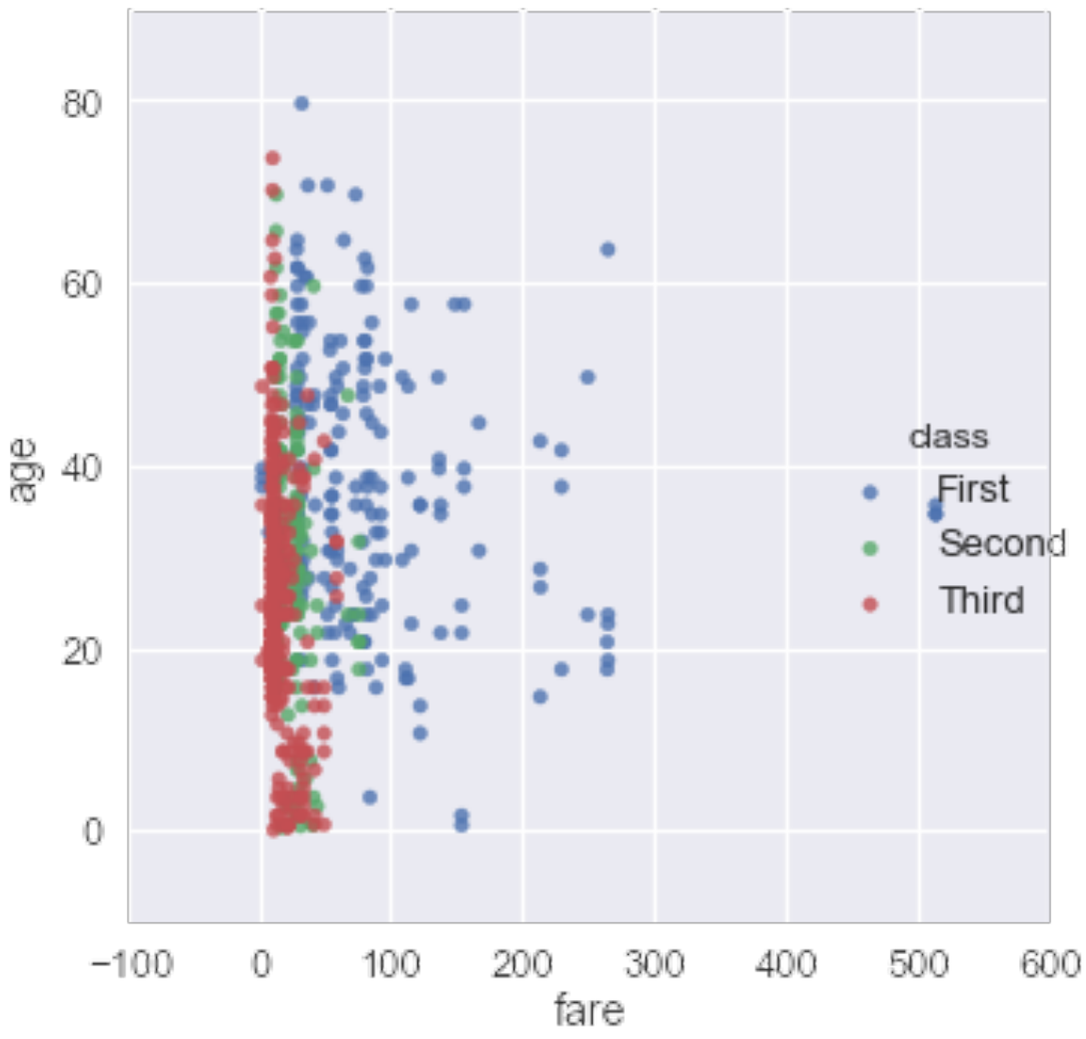
Trivariate Data



Do NOT use 3D scatterplots!

age	fare	class
22.0	7.25	Third
38.0	71.2833	First
26.0	7.925	Third
35.0	53.1	First
35.0	8.05	Third
	8.4583	Third
54.0	51.8625	First
2.0	21.075	Third
27.0	11.1333	Third
14.0	30.0708	Second
4.0	16.7	Third
58.0	26.55	First
20.0	8.05	Third
39.0	31.275	Third
14.0	7.8542	Third
55.0	16.0	Second
2.0	29.125	Third
	13.0	Second
31.0	18.0	Third
	7.225	Third
35.0	26.0	Second
34.0	13.0	Second
15.0	8.0292	Third

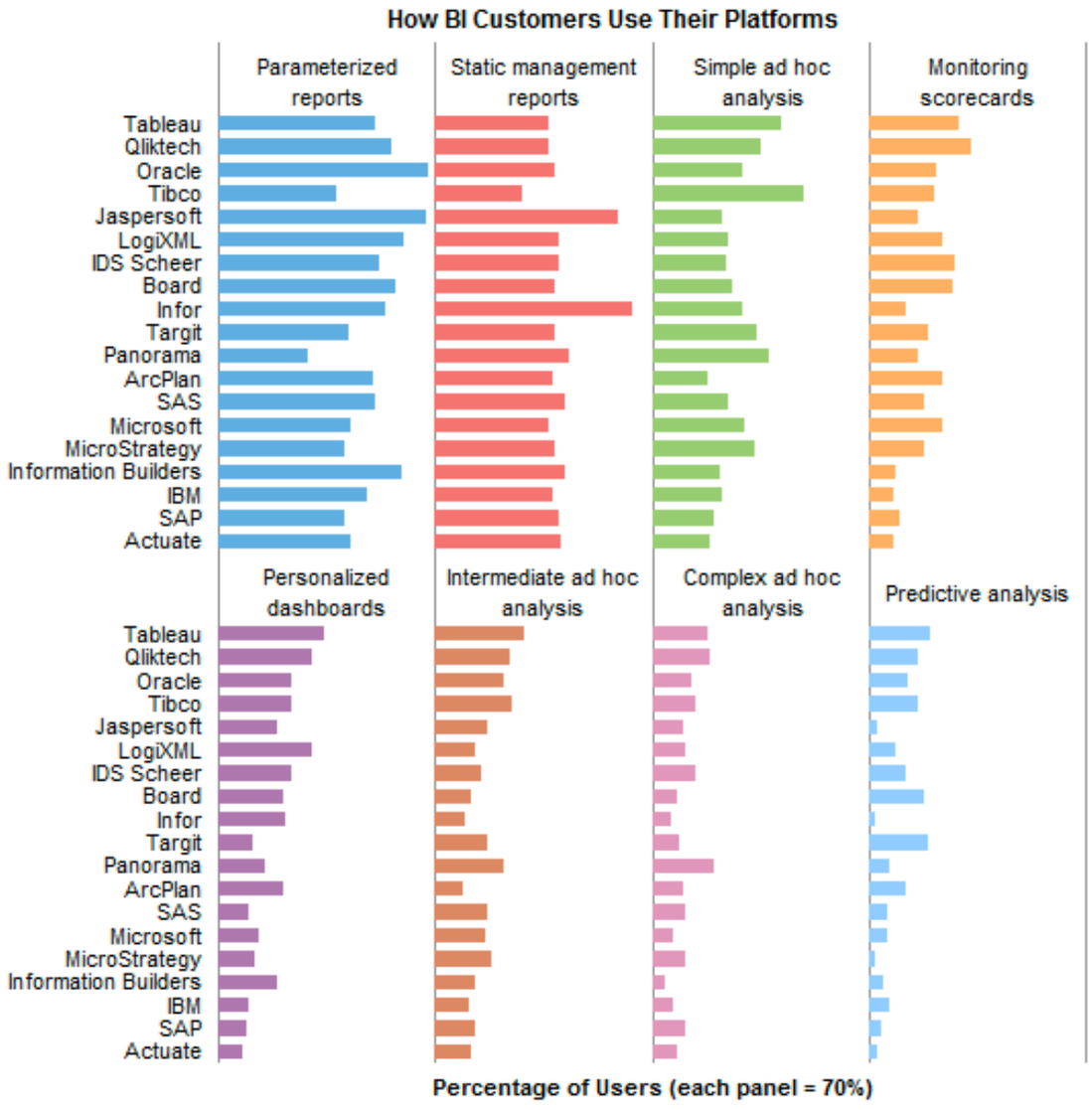
Trivariate Data

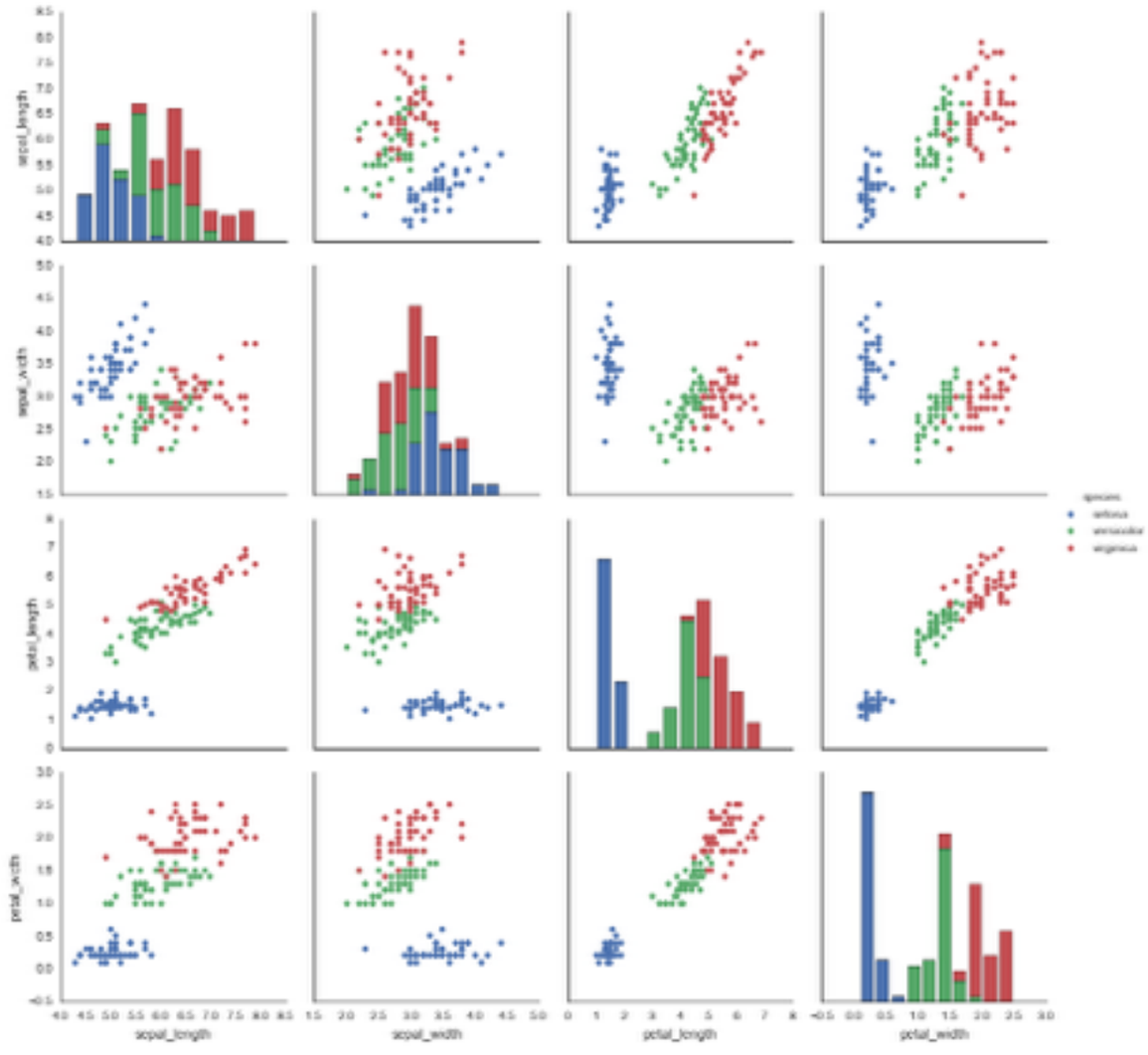


Map the third dimension to some other visual attribute

Multivariate Data

Give each attribute its own display (small multiples)





More next Tuesday!

Data Reduction

Filtering: Eliminate some items or attributes

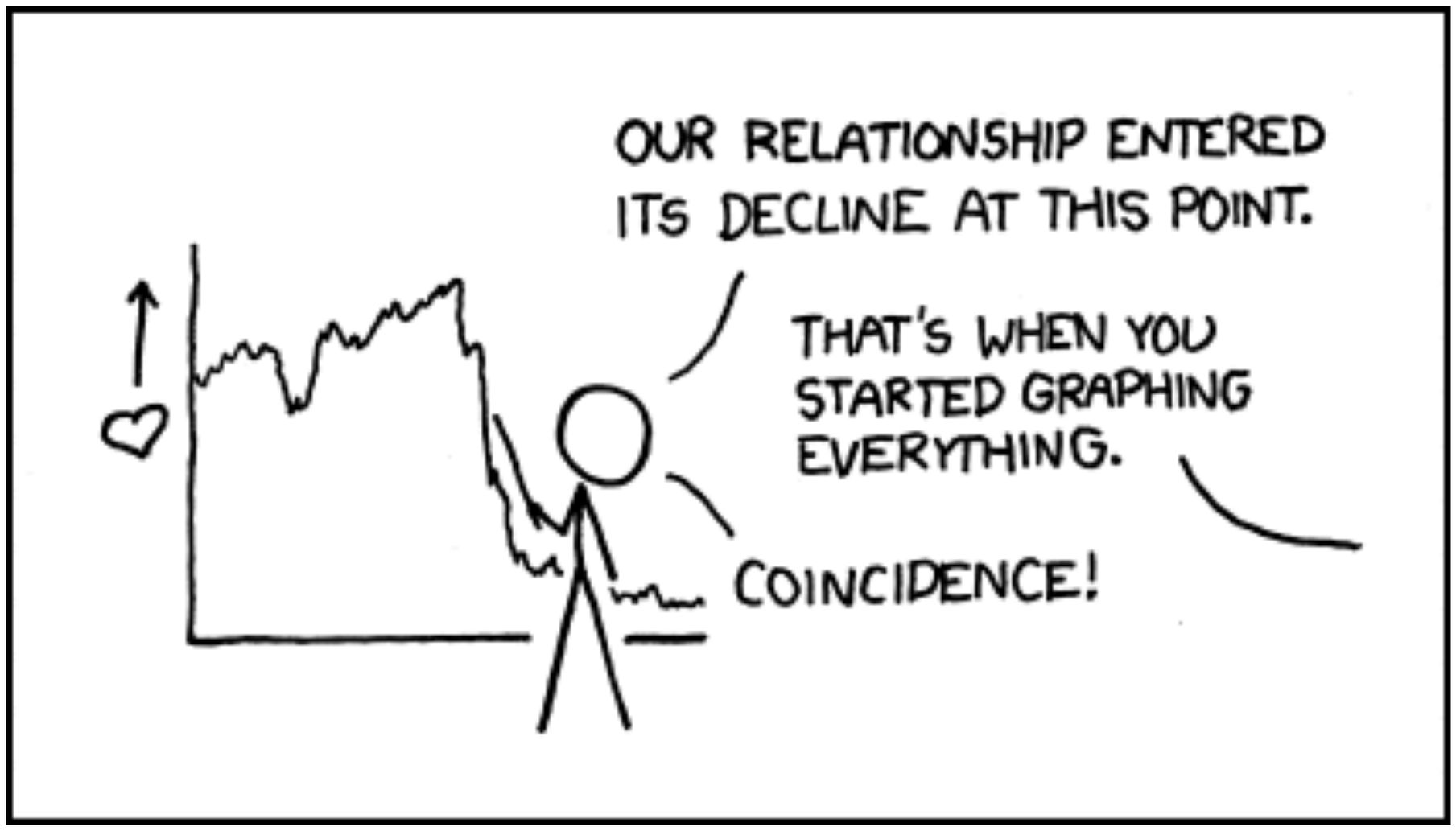
e.g., select range of interest, zoom in, remove outliers, etc.

Aggregation: Represent a group of elements by a new derived element

e.g., take average, min, max, count, sum

Attribute aggregation a.k.a. dimensionality reduction

Statistical Graph Types

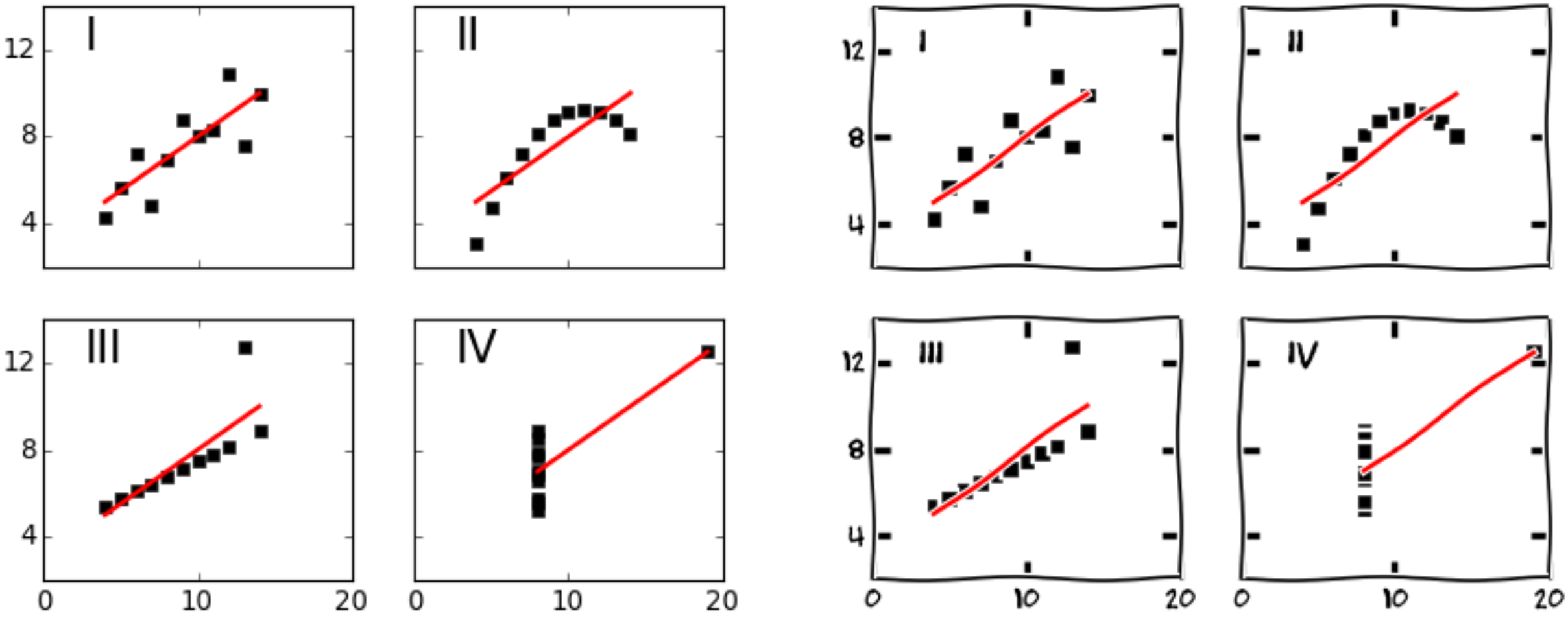


http://xkcd.com/523/

Side Note

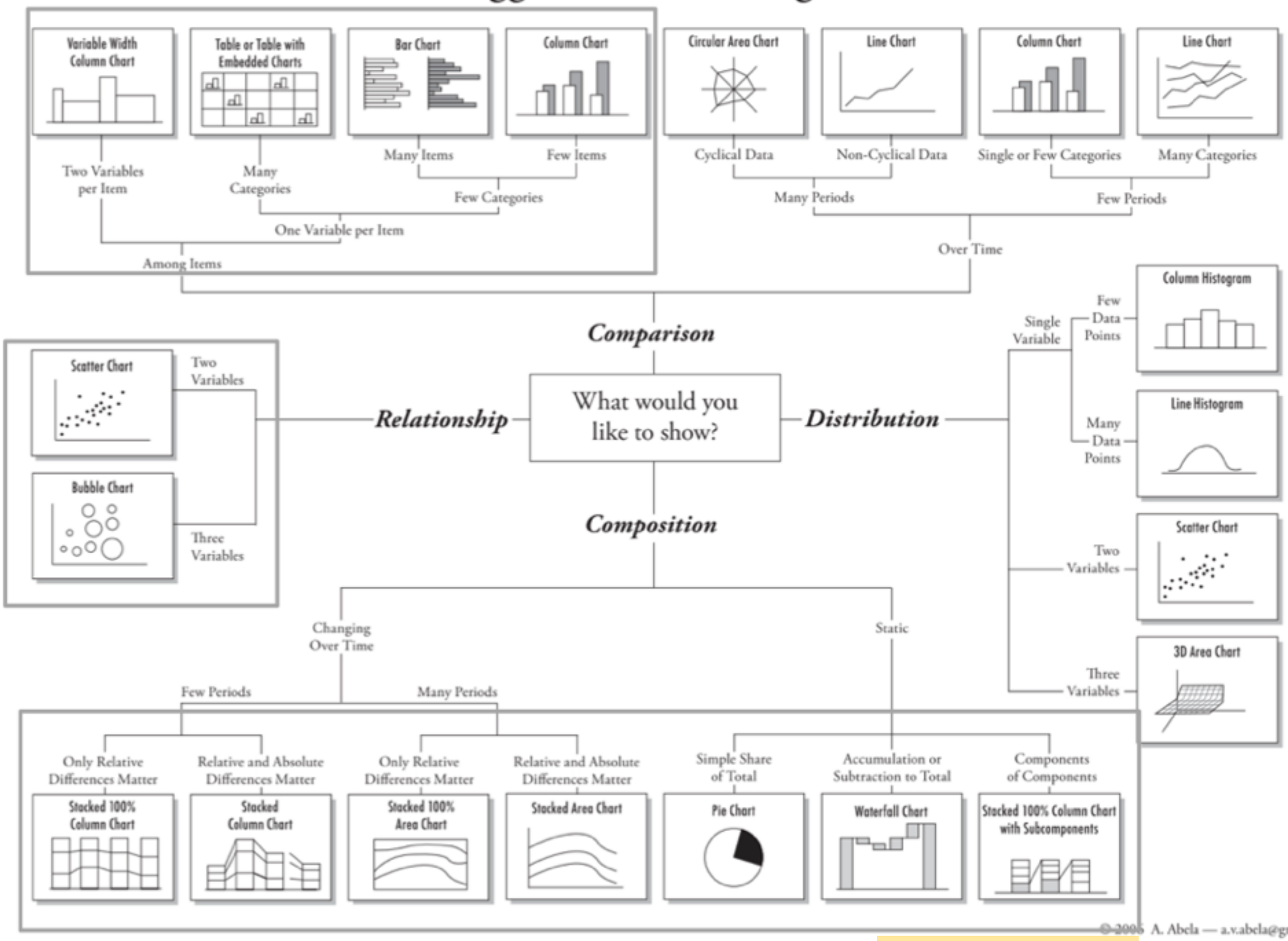
MATPLOTLIB

XKCD-ify your plot



http://matplotlib.org/xkcd

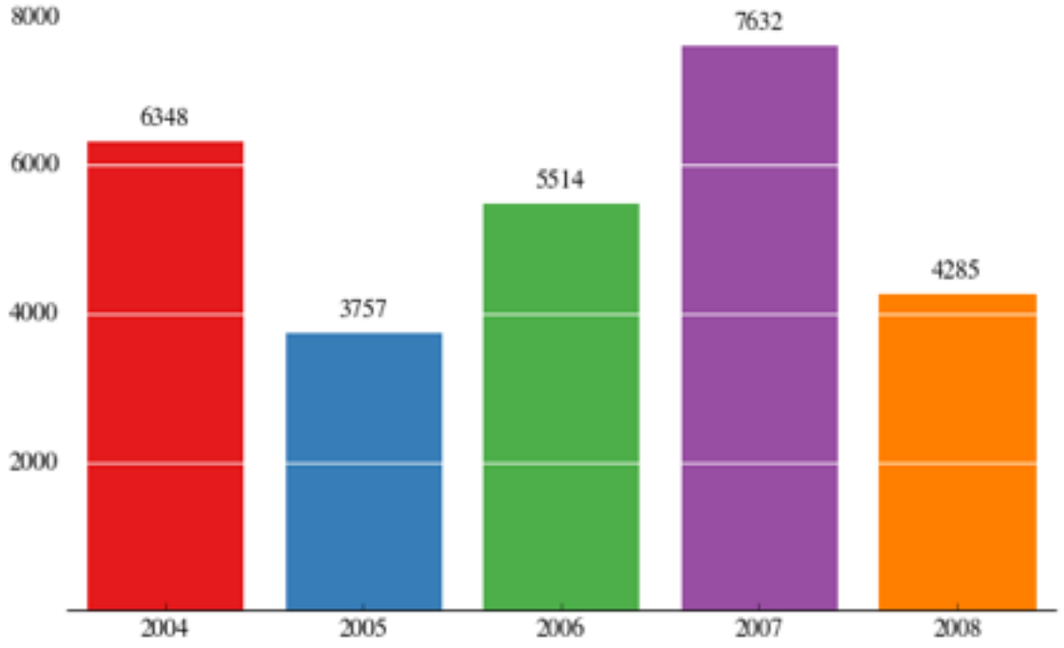
Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

Comparisons

Bar Chart



http://nbviewer.ipython.org/gist/olgabot/5357268

Direction

Graduation rates up in most cities

Graduation rate for principal school district of the largest cities

	2005 RATE	1995-2005 CHANGE
Philadelphia	62.1%	23.2%
Tucson	71.6	22.7
Kansas City, Mo.	53.3	19.7
El Paso	60.6	13.9
Portland, Ore.	68.6	13.1
New York	50.5	12.8
Dallas	50.8	12.7
Columbus	44.7	12.6
Mesa	76.6	12.0
Austin	58.9	11.5
Atlanta	43.5	10.8
Fort Worth	56.5	10.4
Miami	55.9	10.4
Houston	52.9	9.8
Chicago	51.0	9.2
Oakland	50.5	9.2
Virginia Beach	68.5	8.8
Baltimore	41.5	7.7
Denver	58.6	6.9
Detroit	37.5	6.9
San Antonio	47.3	6.4
Phoenix	58.0	5.6
Oklahoma City	47.0	5.3
Indianapolis	30.5	5.3
Milwaukee	41.0	5.2
Sacramento	62.1	4.9
Washington, D.C.	57.6	4.8
Colorado Springs	68.8	4.6
Honolulu	67.4	3.6
Nashville	45.2	3.1
Jacksonville	50.8	0.7

The average high school graduation rate of major cities was 54.7 percent in 2005. Of the 62 percent that improved since 1995, Philadelphia had the highest increase. The rate in Las Vegas decreased the most.

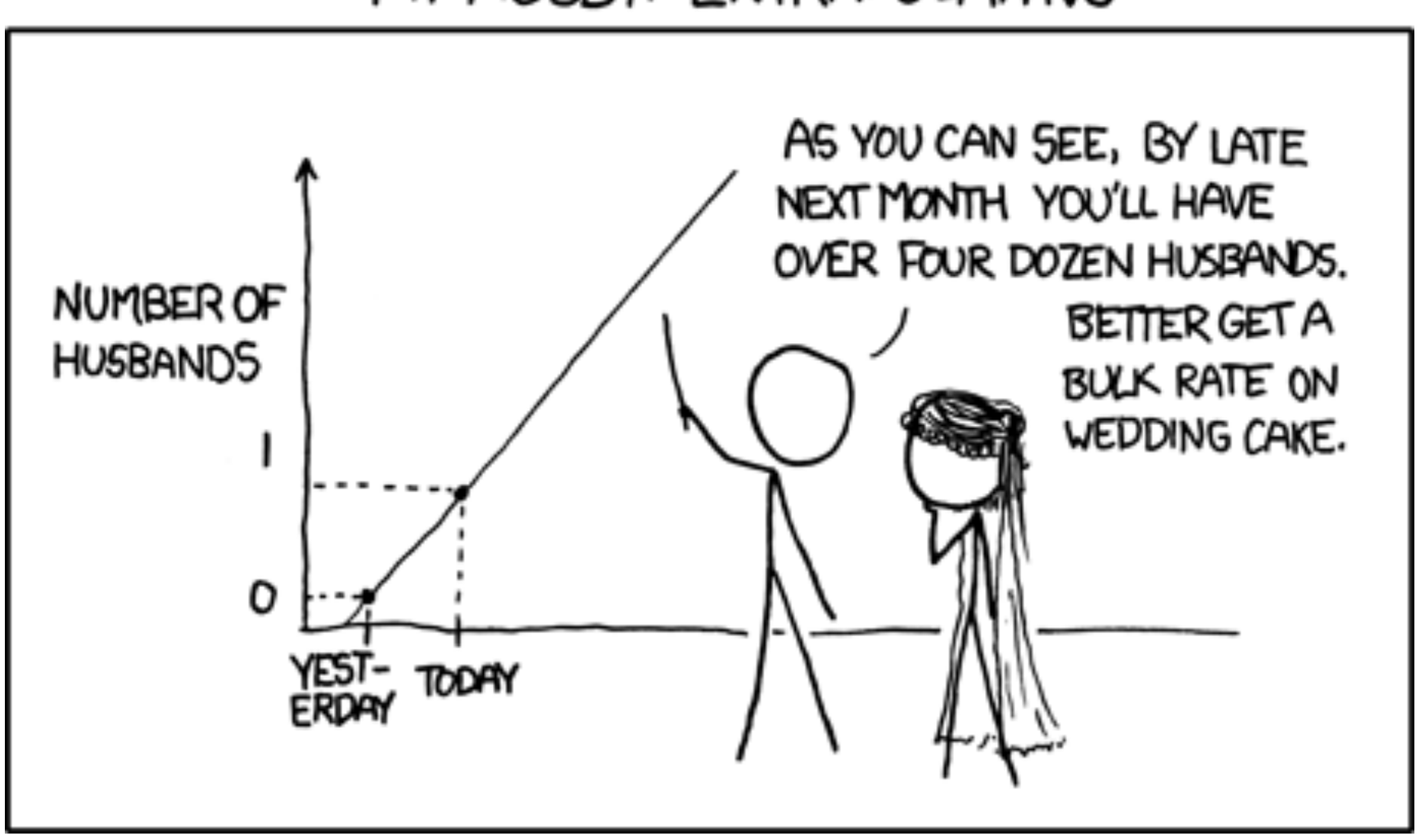
1995-2005 CHANGE	2005 RATE	
-0.3	63.4%	Louisville
-0.7	68.9	Seattle
-1.2	51.2	Memphis
-1.5	51.9	Fresno
-1.7	58.6	Boston
-1.7	45.3	Minneapolis
-1.8	73.3	San Jose
-2.0	48.5	Tulsa
-2.3	60.5	Charlotte
-2.4	63.7	San Diego
-3.6	44.4	Los Angeles
-3.7	64.0	Long Beach
-4.9	34.4	Cleveland
-6.5	57.1	San Francisco
-6.6	49.0	Albuquerque
-11.6	60.3	Arlington, TX
-14.8	49.6	Omaha
-17.6	54.5	Wichita
-23.1	44.5	Las Vegas

SOURCE: EPE Research Center

AP

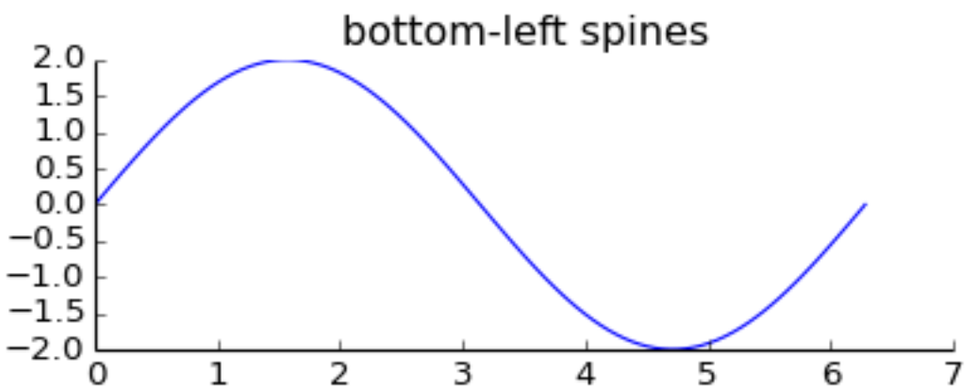
Nicolas Rapp

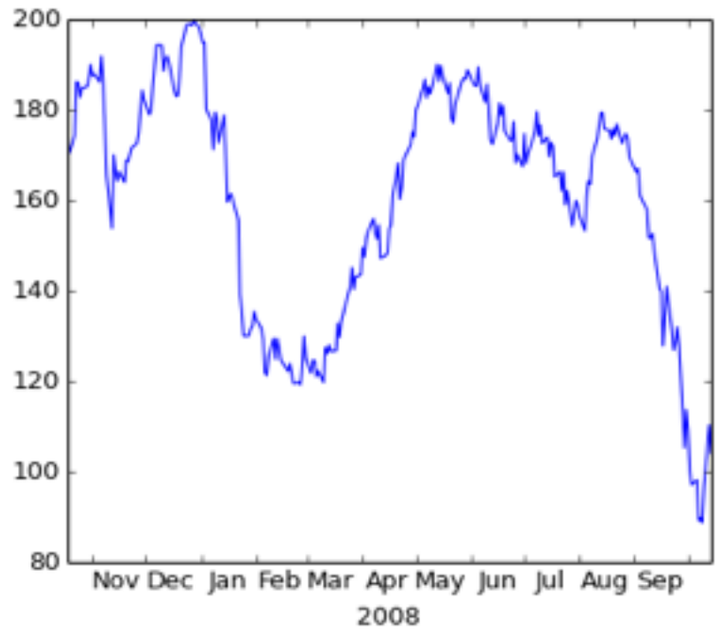
Trends Over Time

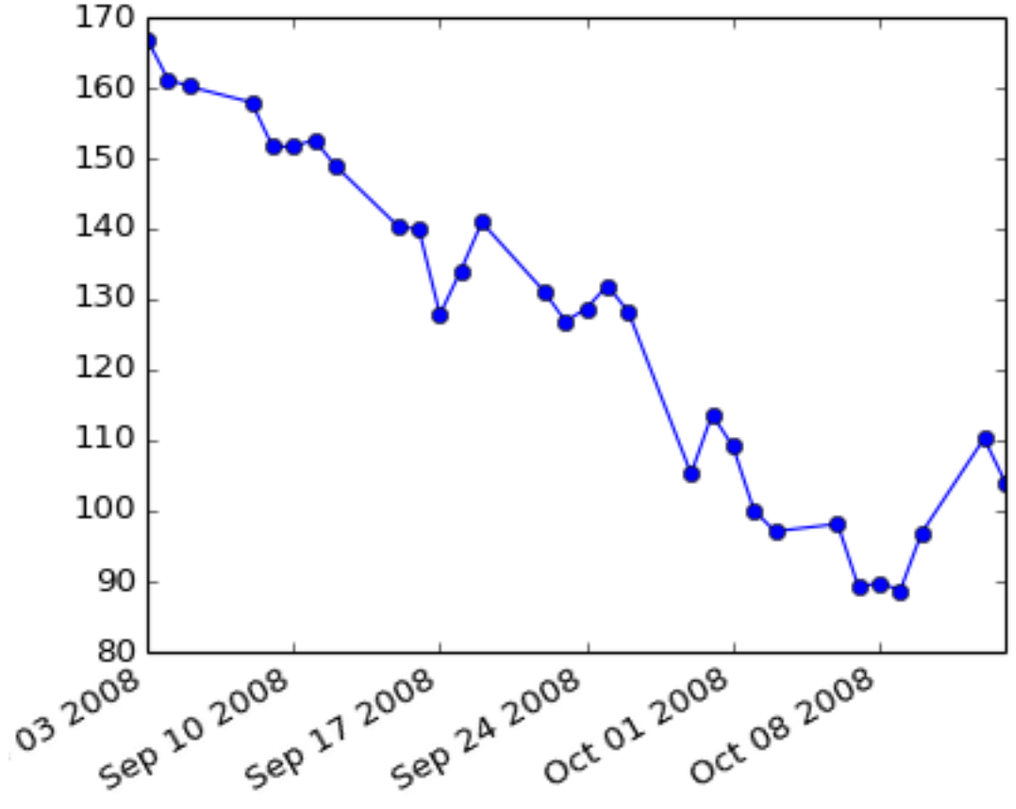


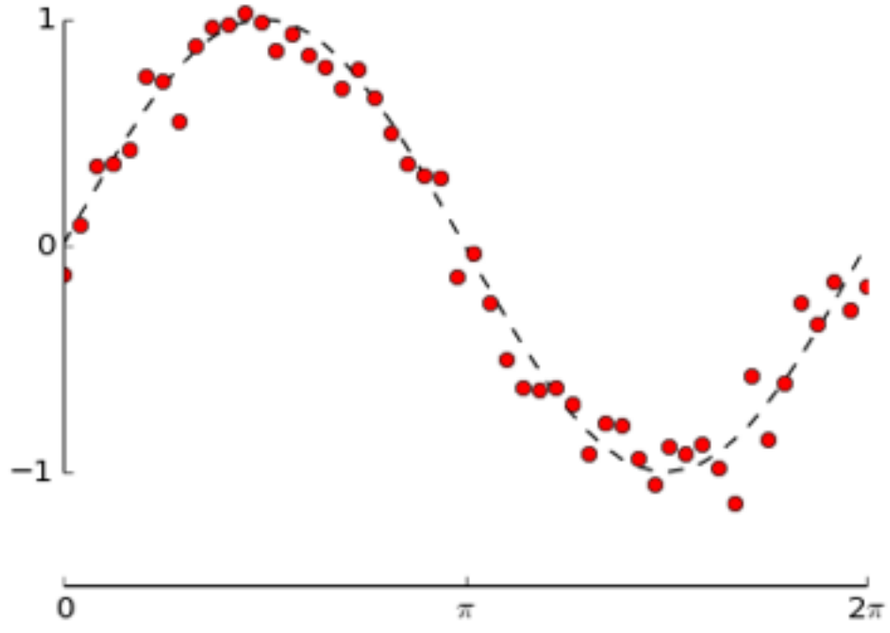
http://xkcd.com/605/

Line Charts



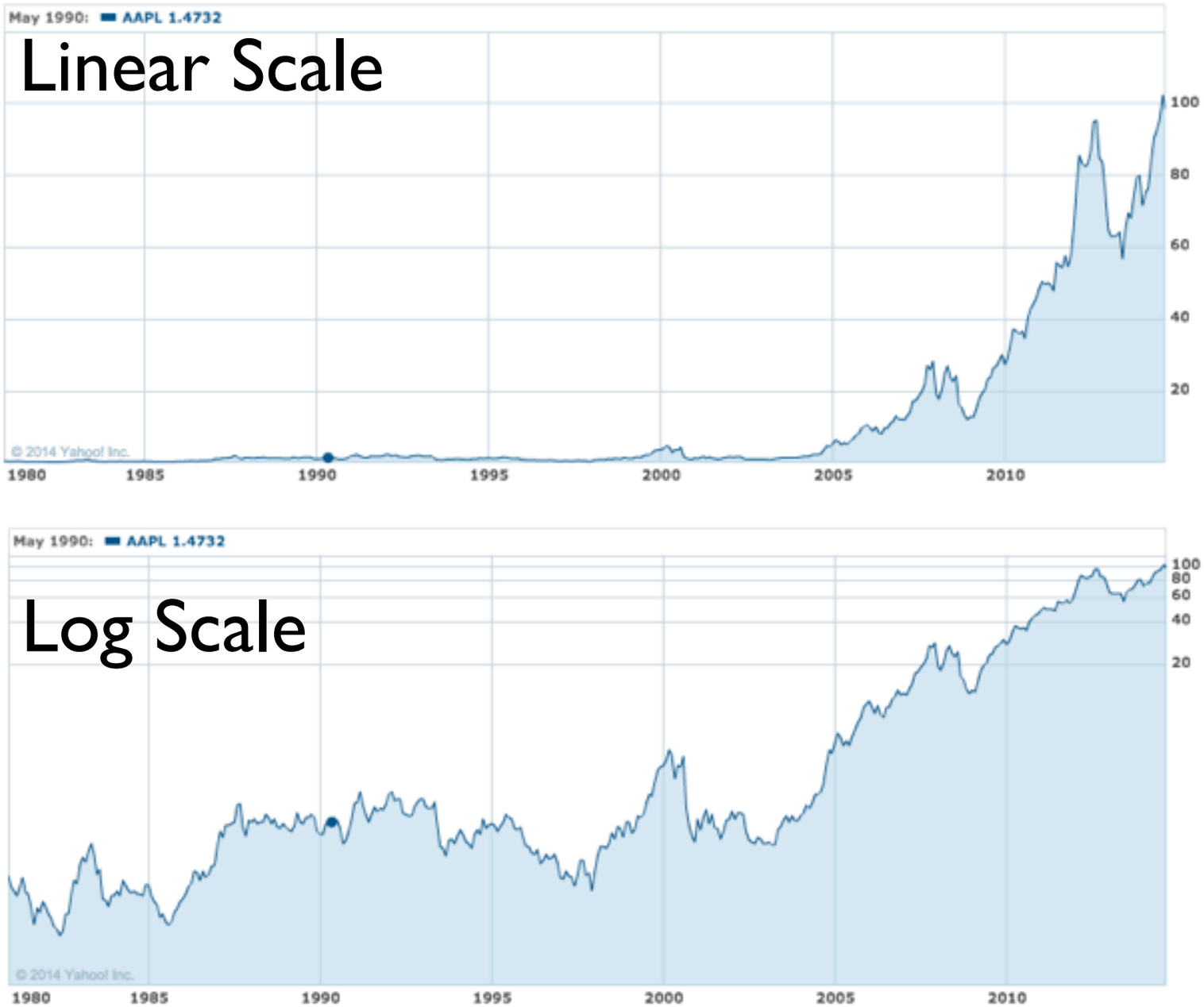


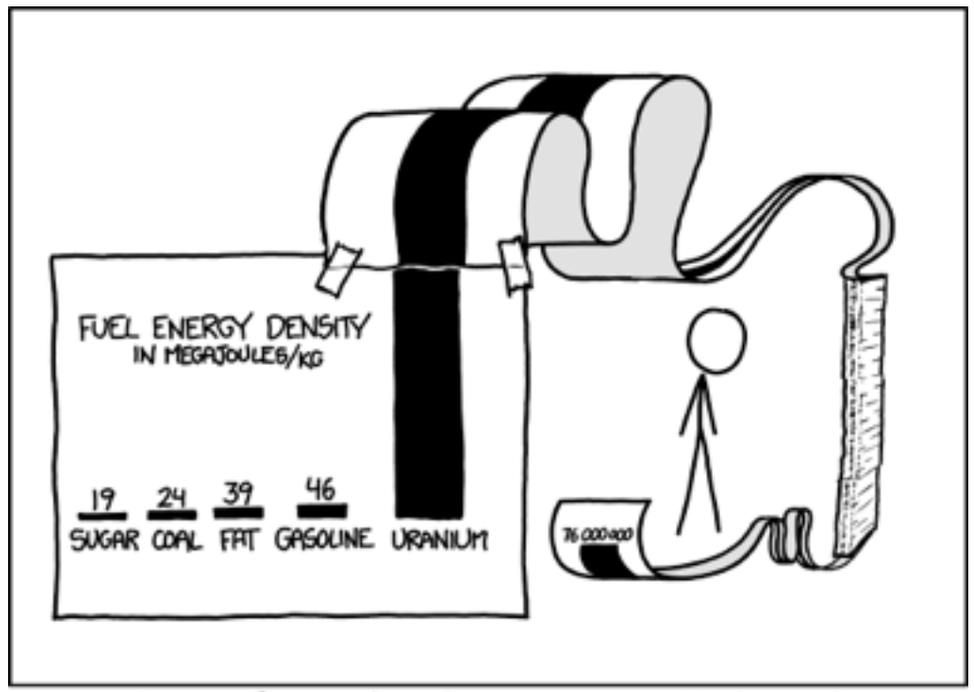




matplotlib gallery

Linear vs. Logarithmic Scale





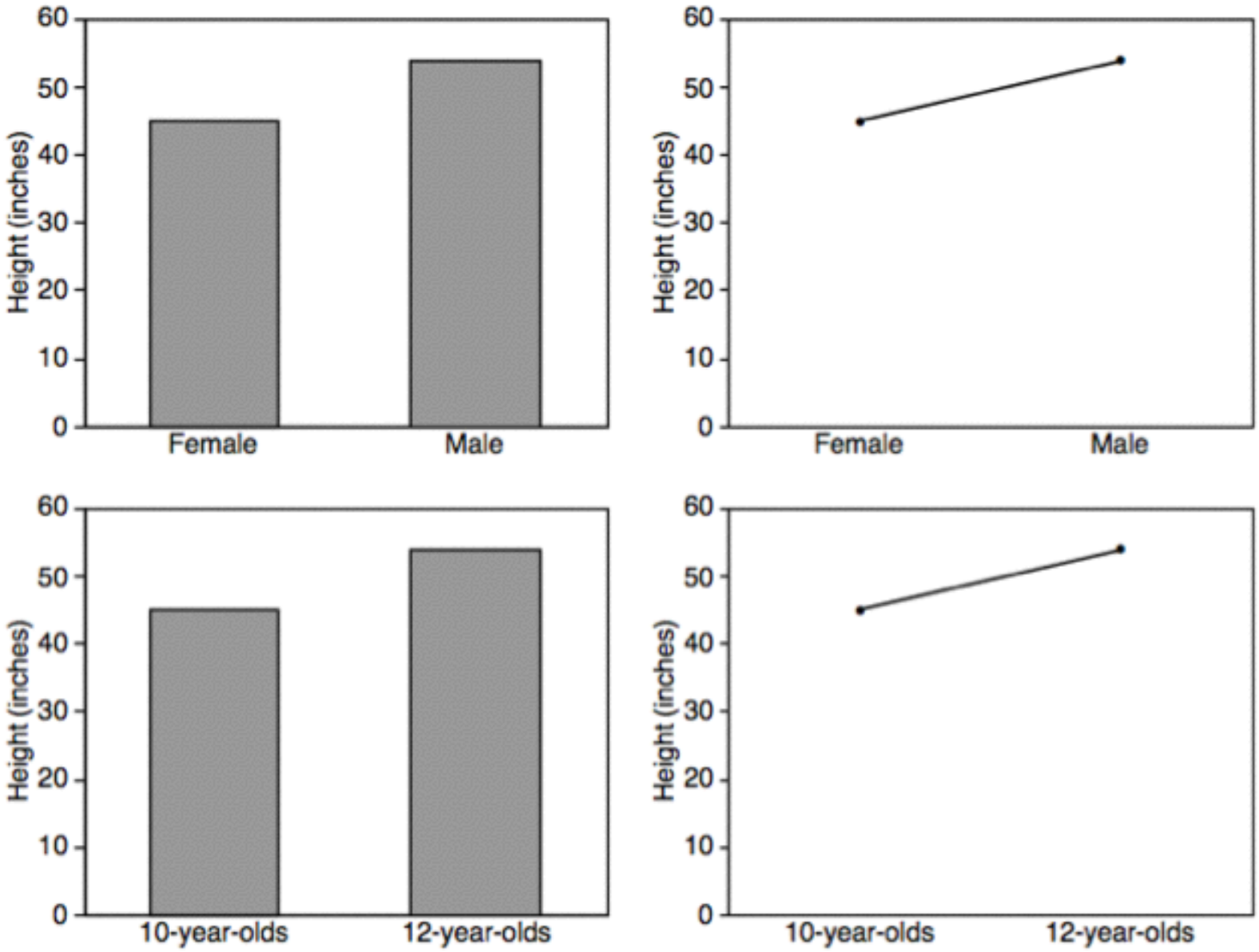
SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

http://xkcd.com/1162/

Apple Stock Price http://finance.yahoo.com/echarts?s=AAPL

Bars vs. Lines

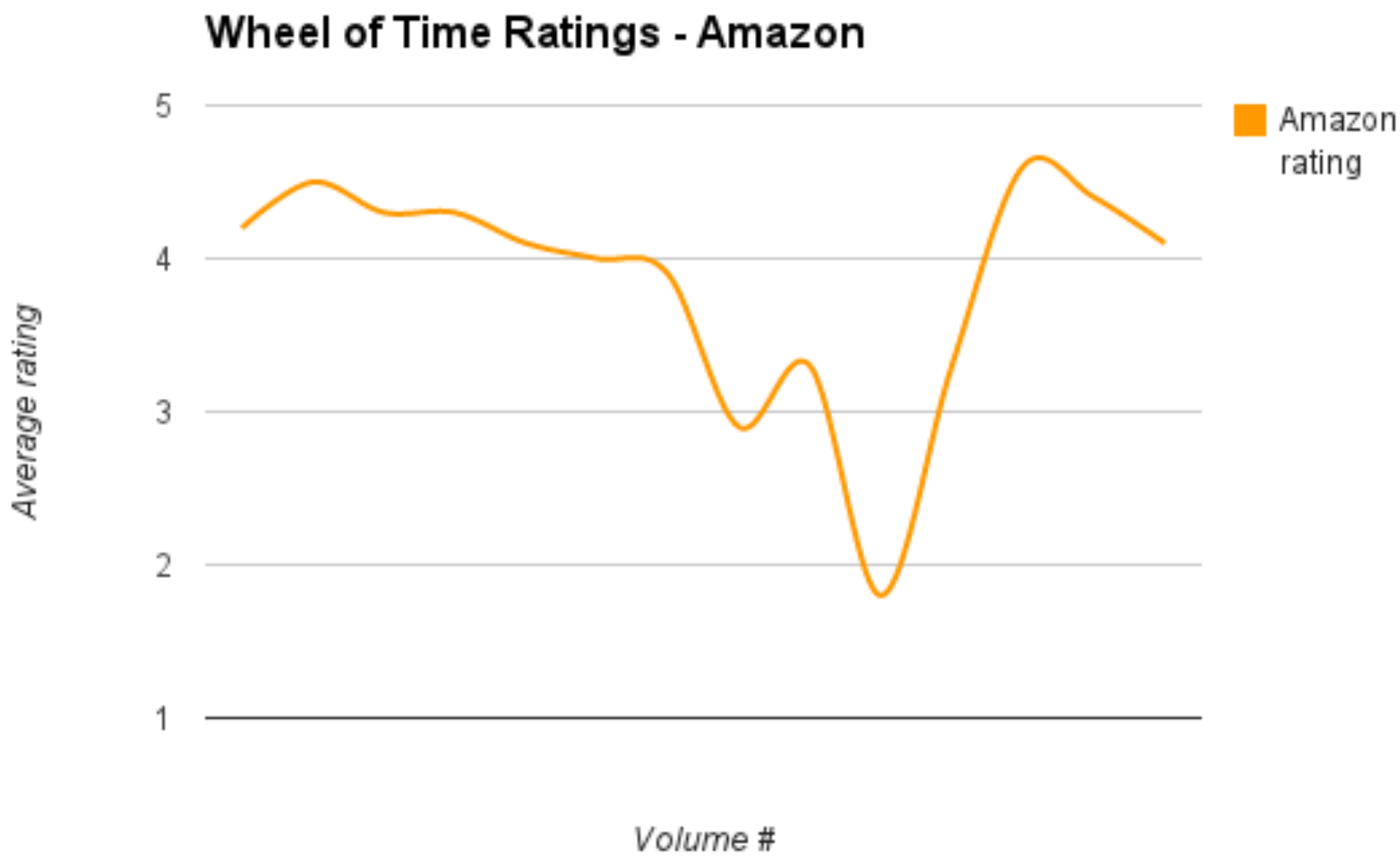
Lines imply connections - do not use for categorical data



Zacks 1999

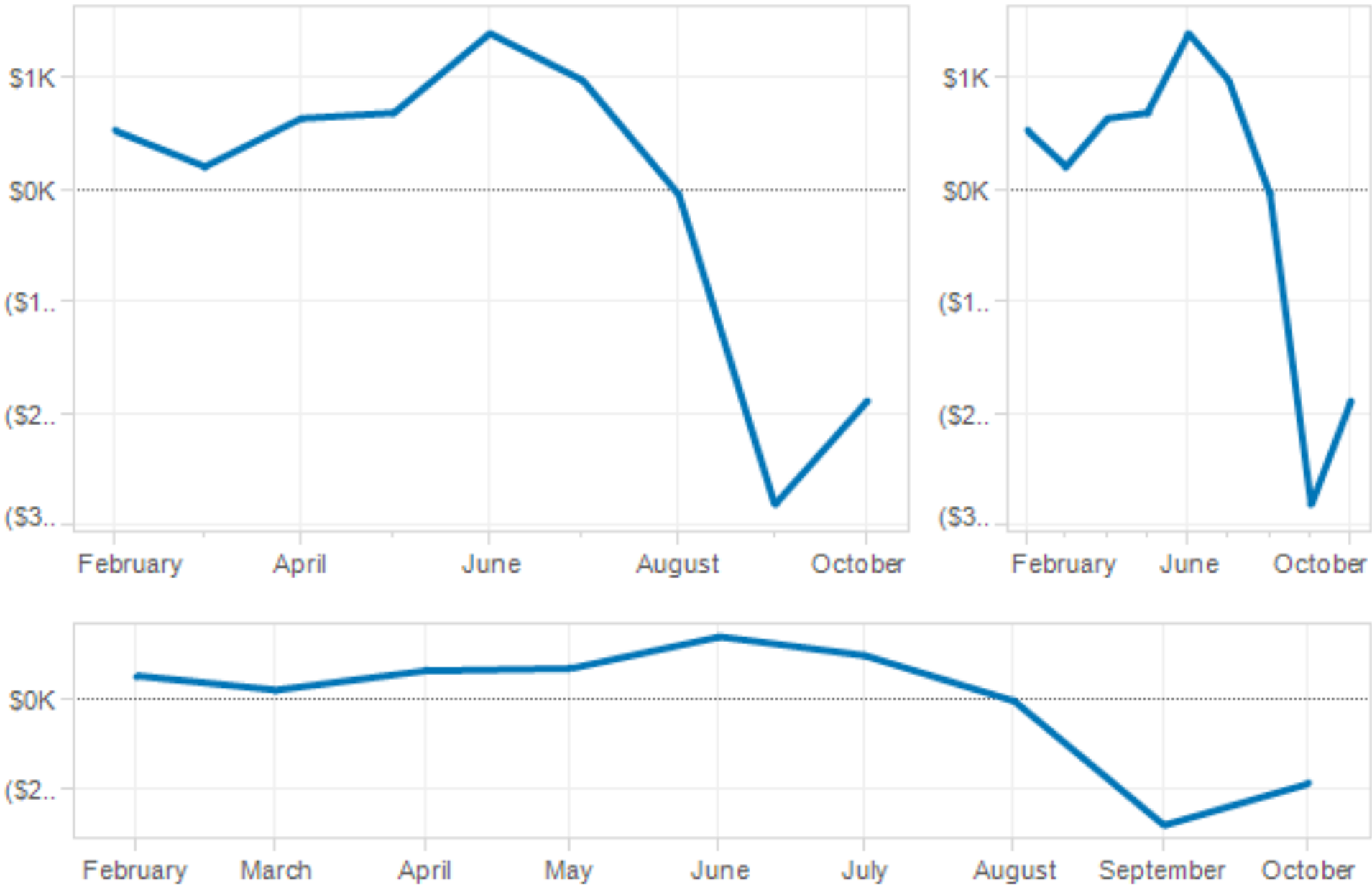
Don’t

Use bar charts to compare book ratings



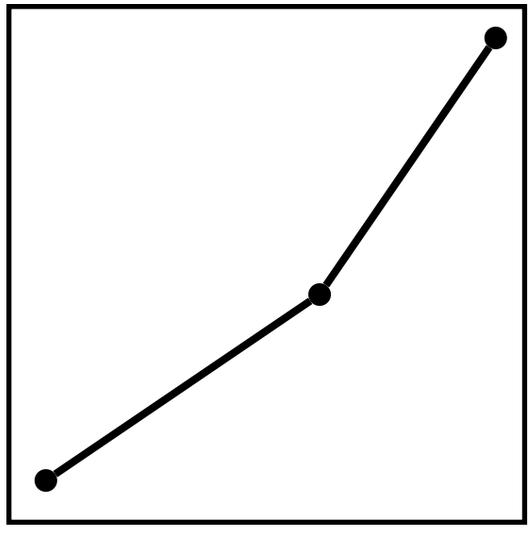
“Visualizing The Wheel of Time: Reader Sentiment for an Epic Fantasy Series”, J. Siddle, Sept 2013

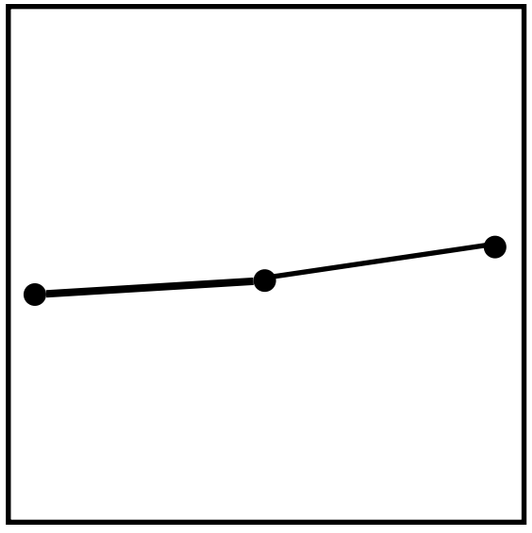
Aspect Ratios

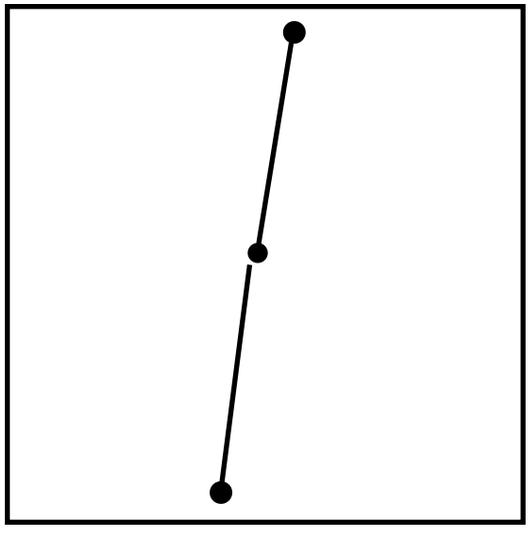


Banking to 45°

Two line segments are maximally discriminable when their average absolute angle is 45° W. Cleveland

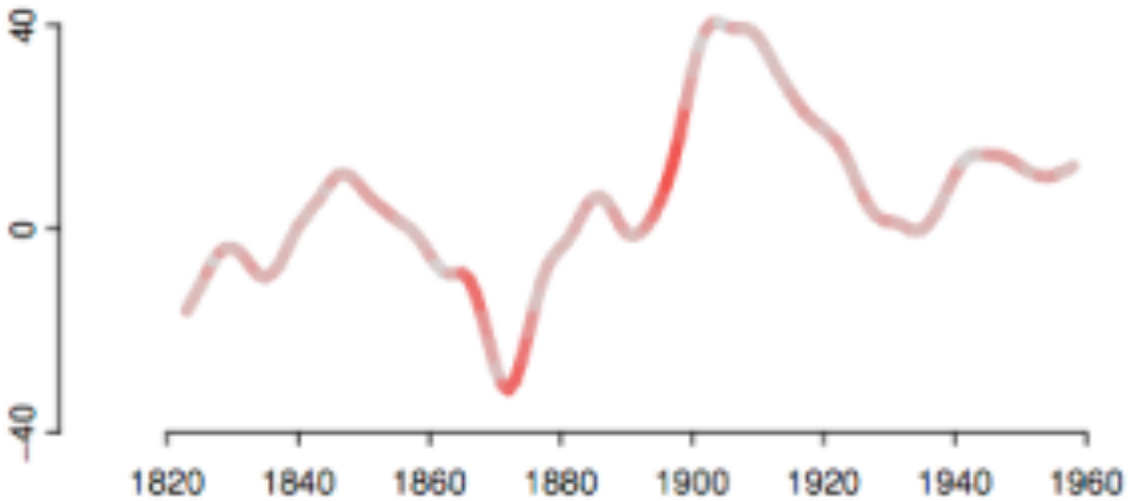




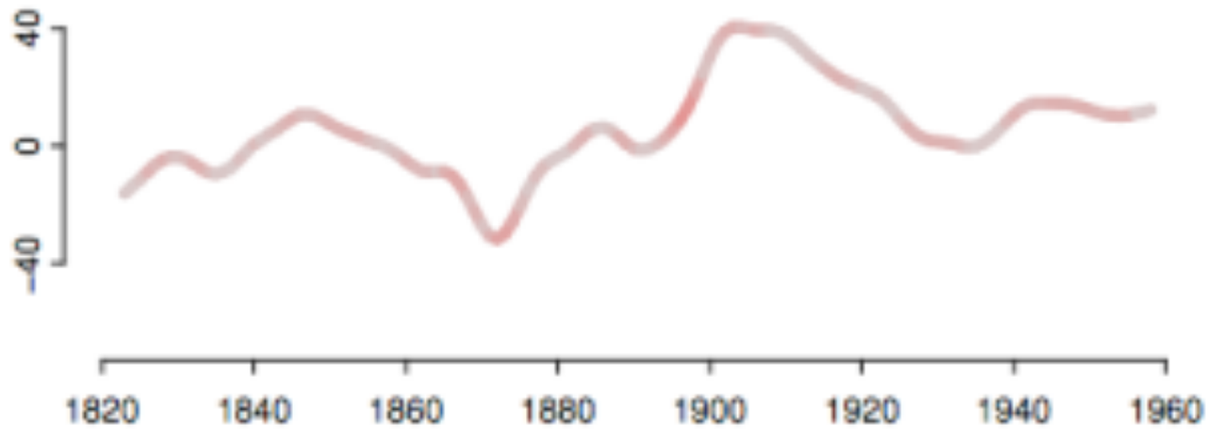


Slide based on J. Heer / M. Agrawala

Banking to 45°



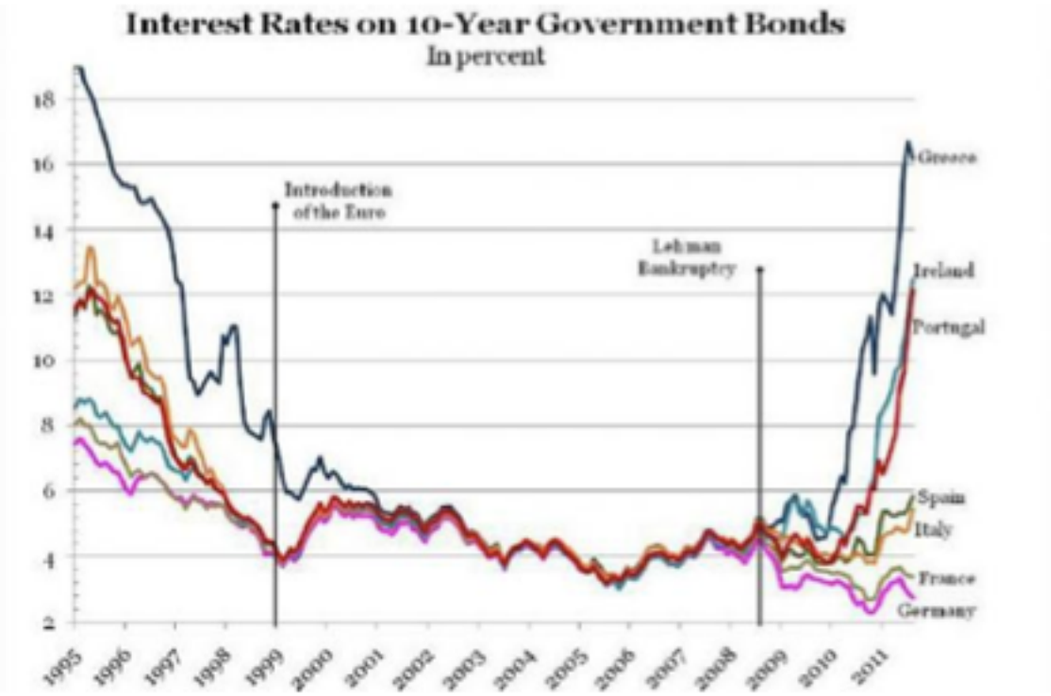
Error Prone

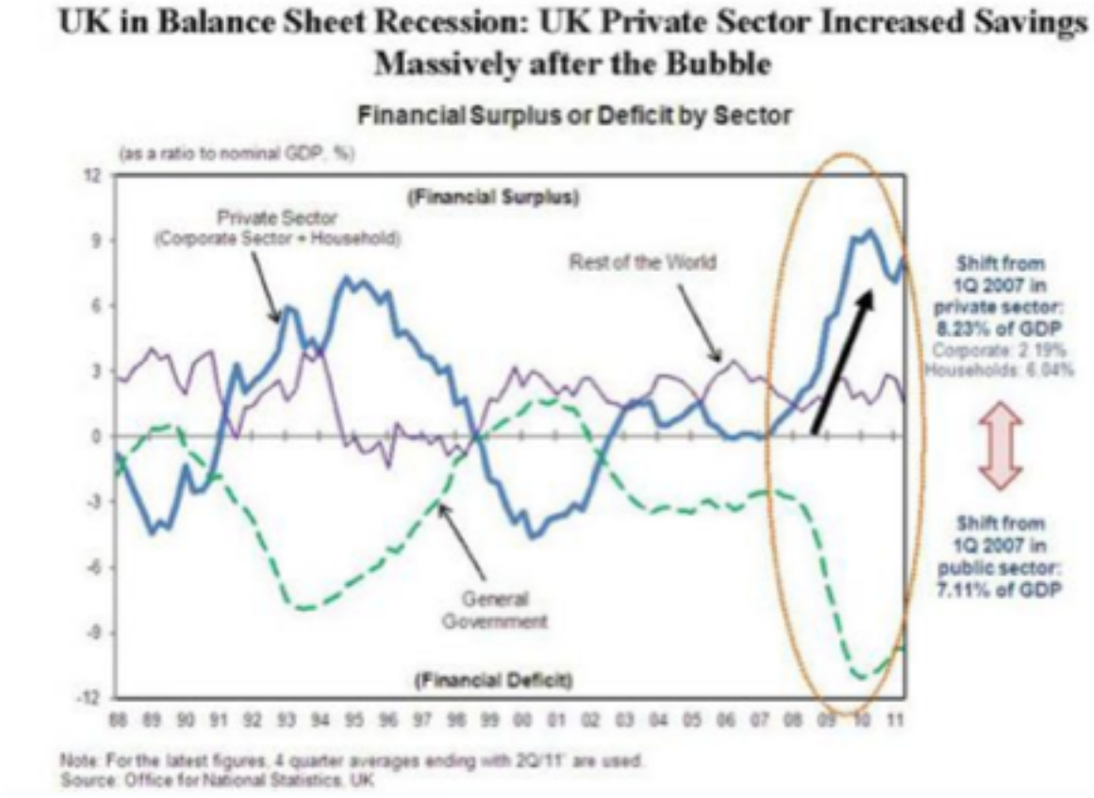


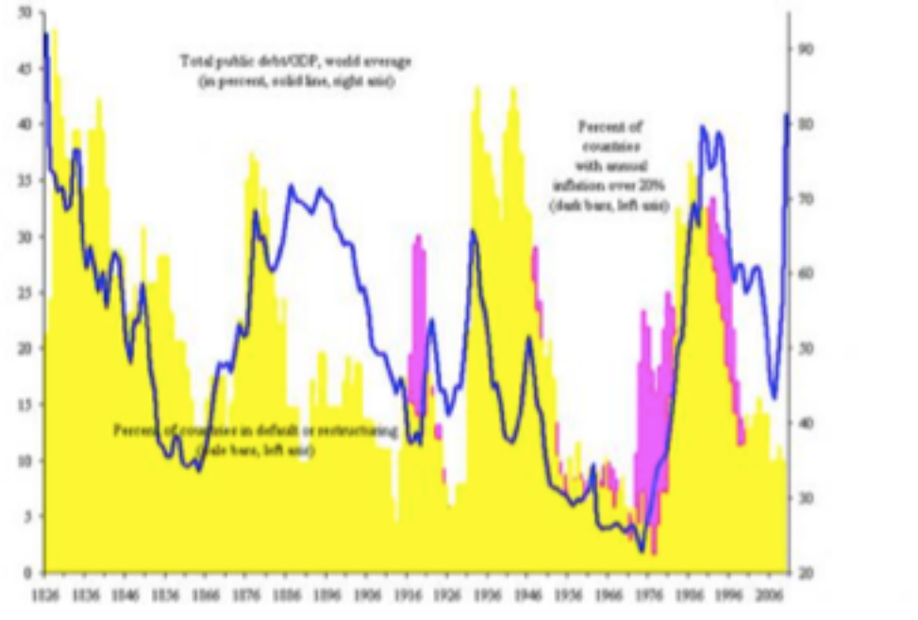
Optimal Aspect Ratio

An Empirical Model of Slope Ratio Comparisons,
Justin Talbot, John Gerth, Pat Hanrahan

Don’t

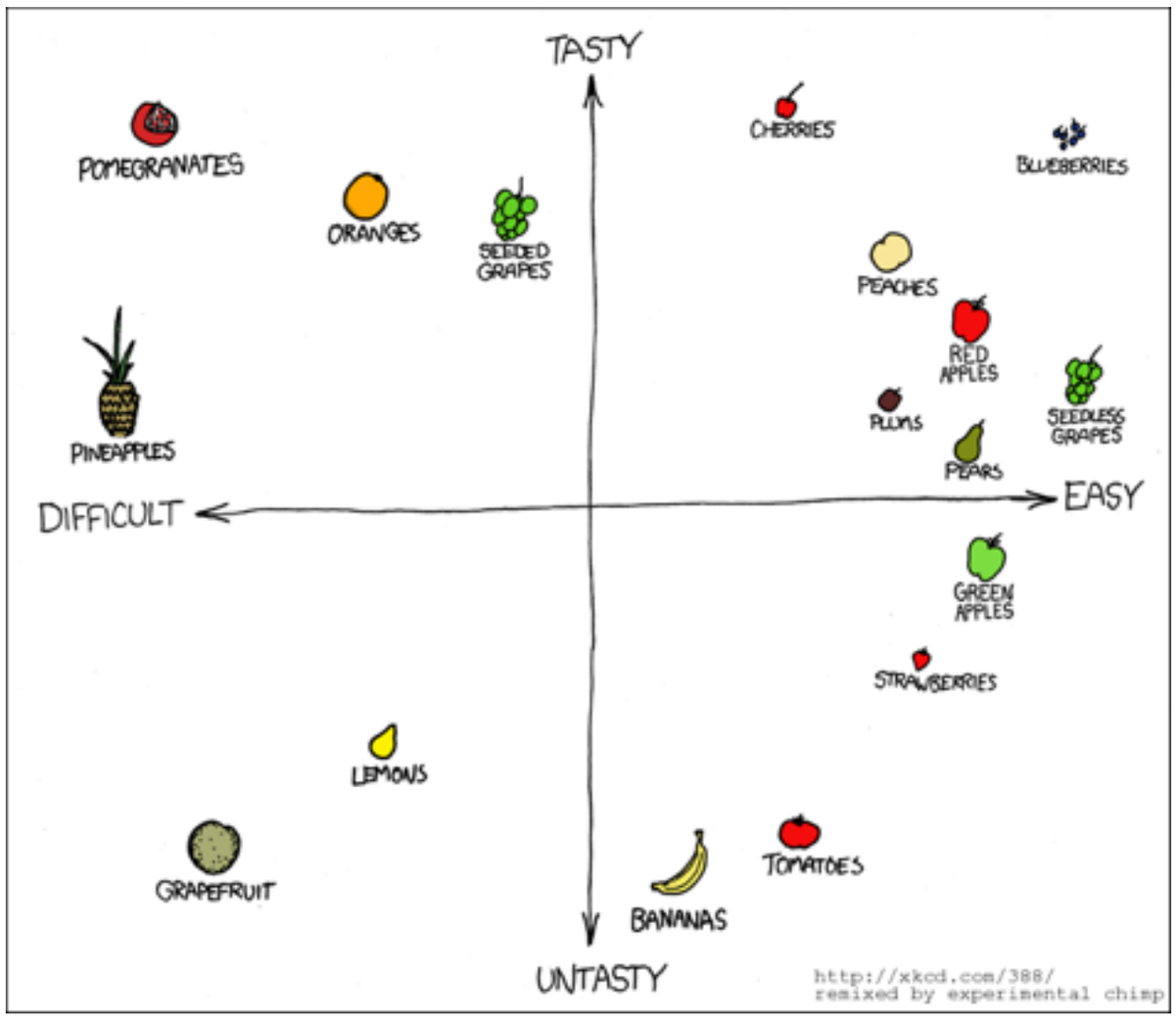

Interest Rates on 10-Year Government Bonds
In percent
Introduction of the Euro
Lehman Bankruptcy
Greece
Ireland
Portugal
Spain
Italy
France
Germany
18
16
14
12
10
8
6
4
2
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

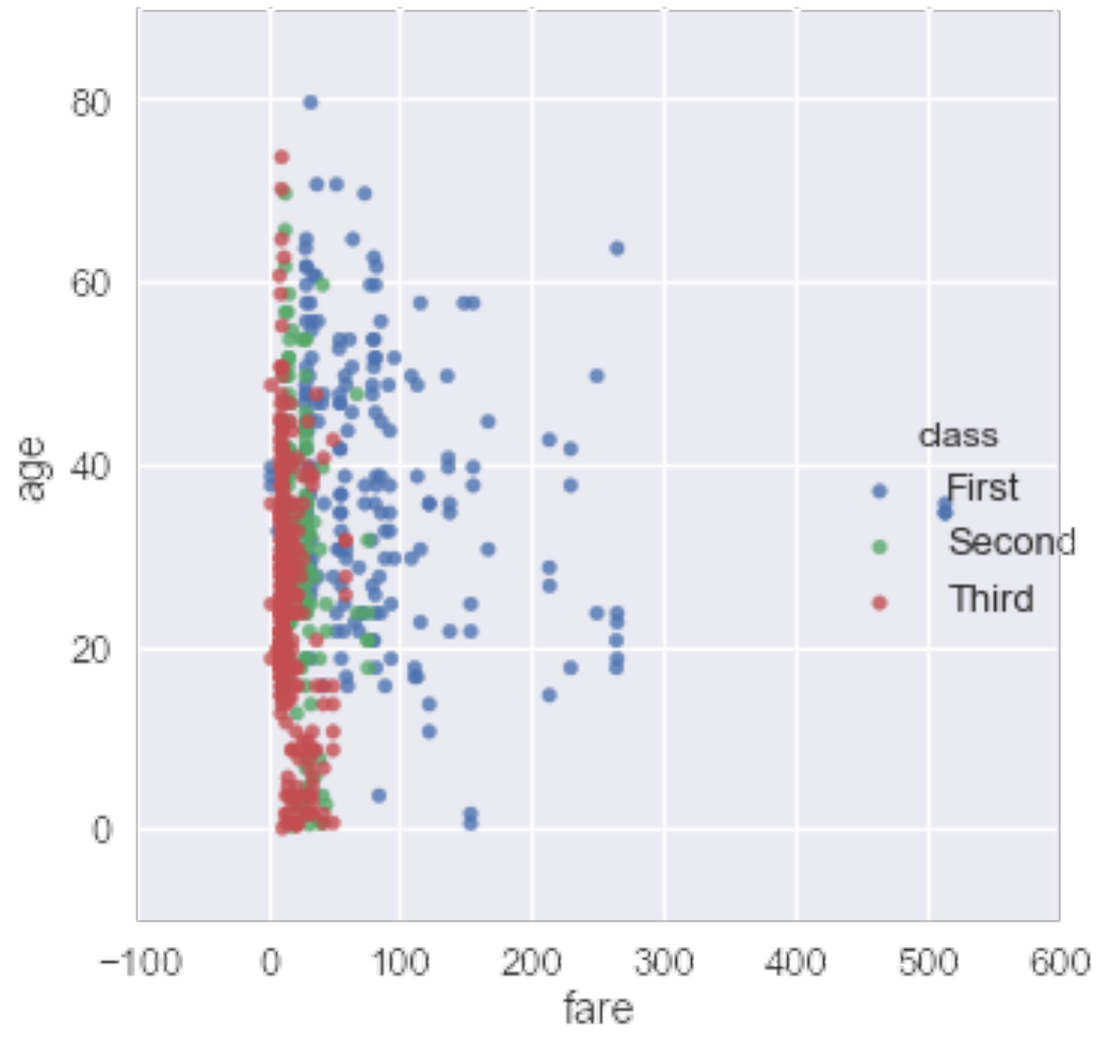

UK in Balance Sheet Recession: UK Private Sector Increased Savings Massively after the Bubble
Financial Surplus or Deficit by Sector
(as a ratio to nominal GDP, %)
(Financial Surplus)
Private Sector
(Corporate Sector + Household)
Rest of the World
General Government
(Financial Deficit)
Shift from 1Q 2007 in private sector: 8.23% of GDP
Corporate: 2.19%
Households: 6.04%
Shift from 1Q 2007 in public sector: 7.11% of GDP
12
9
6
3
0
-3
-6
-9
-12
88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11
Note: For the latest figures, 4 quarter averages ending with 2Q/11' are used.
Source: Office for National Statistics, UK


Total public debt/GDP, world average
(in percent, solid line, right axis)
Percent of countries with annual inflation over 20%
(dark bars, left axis)
Percent of countries in default or restructuring
(dark bars, left axis)
50
45
40
35
30
25
20
15
10
5
0
90
80
70
60
50
40
30
20
1826 1836 1846 1856 1866 1876 1886 1896 1906 1916 1926 1936 1946 1956 1966 1976 1986 1996 2006

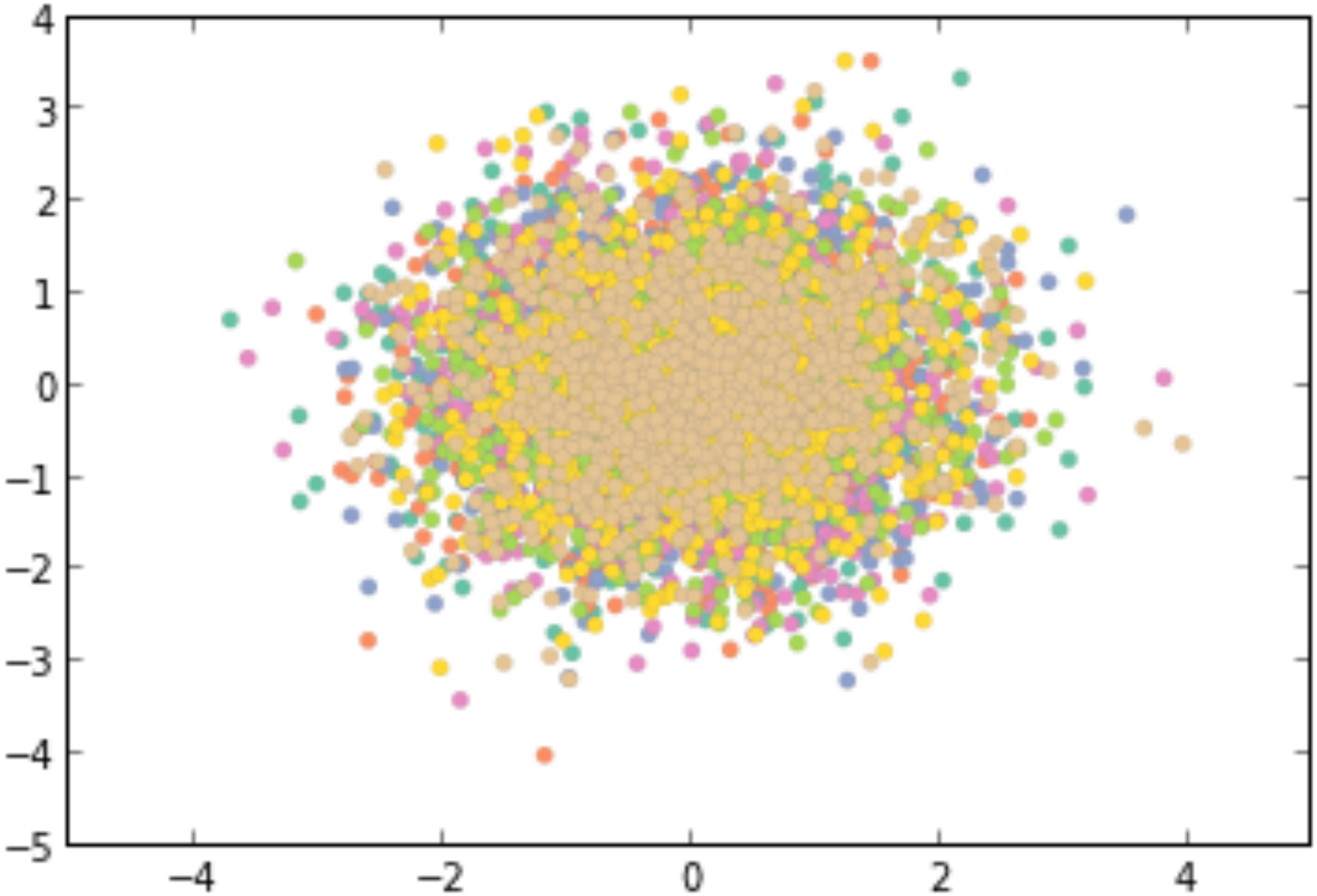
Correlations

Scatterplots





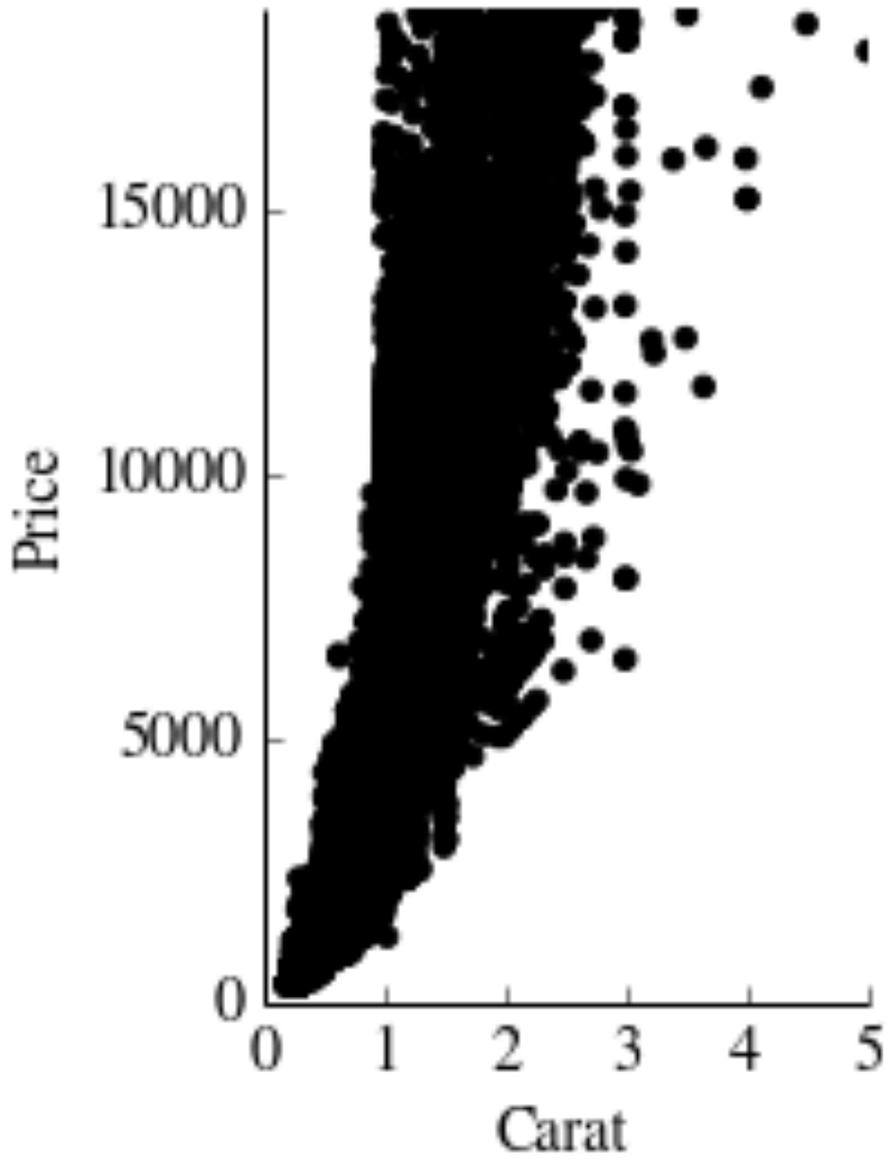
Scatterplots

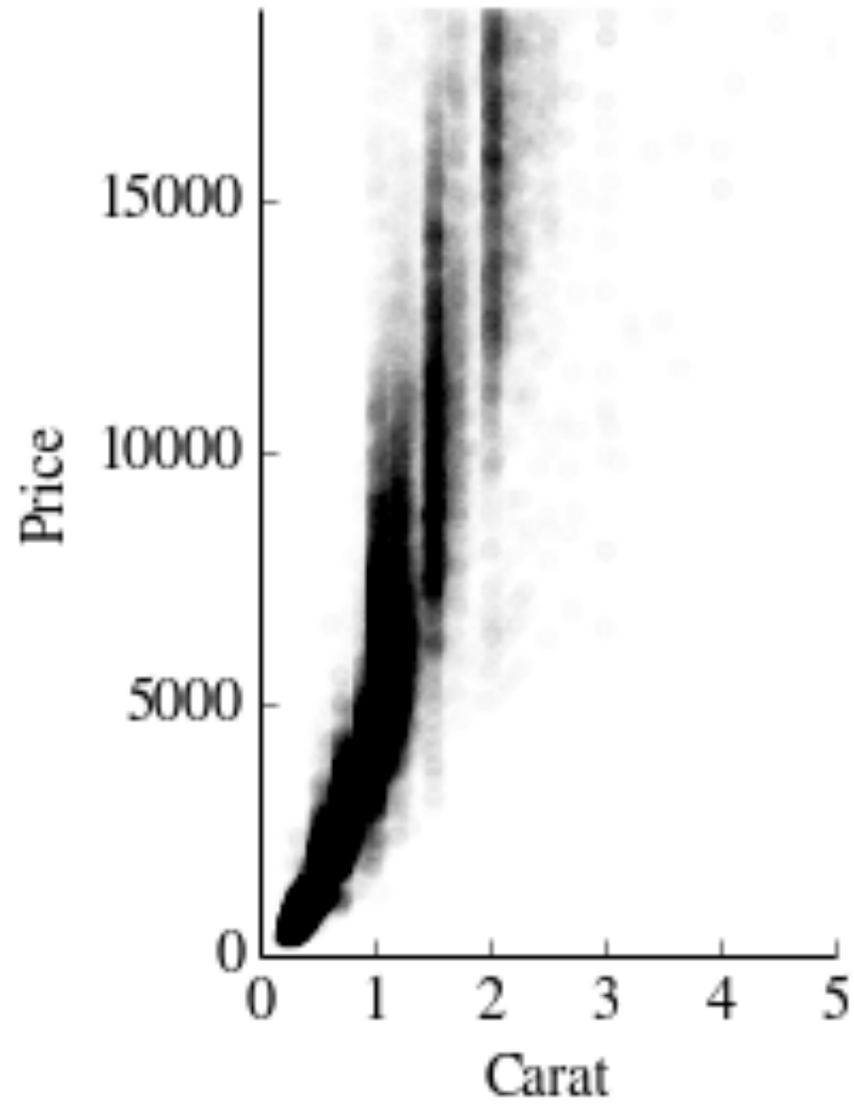


Light Grey Border

http://nbviewer.ipython.org/5357268

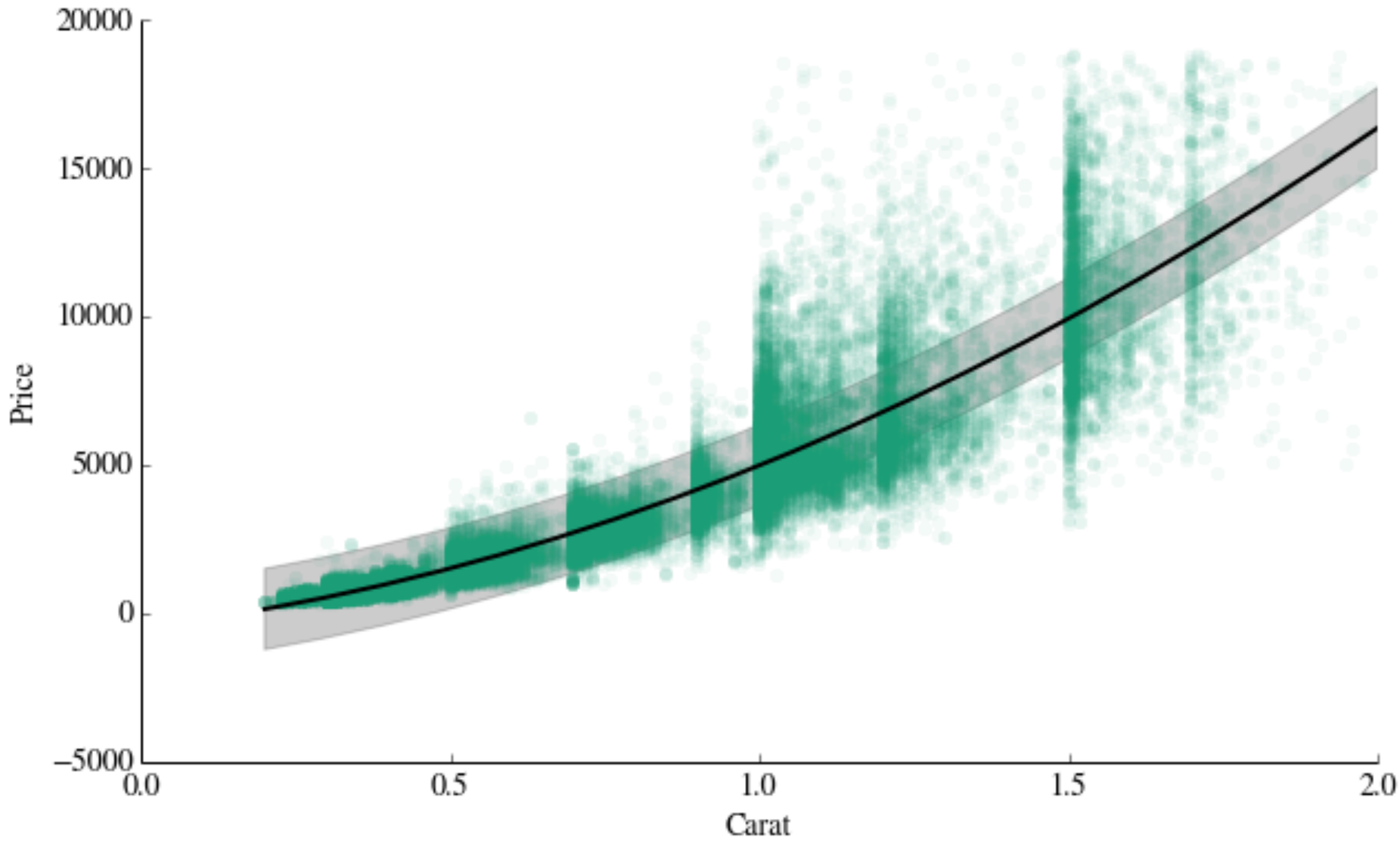
Overplotting



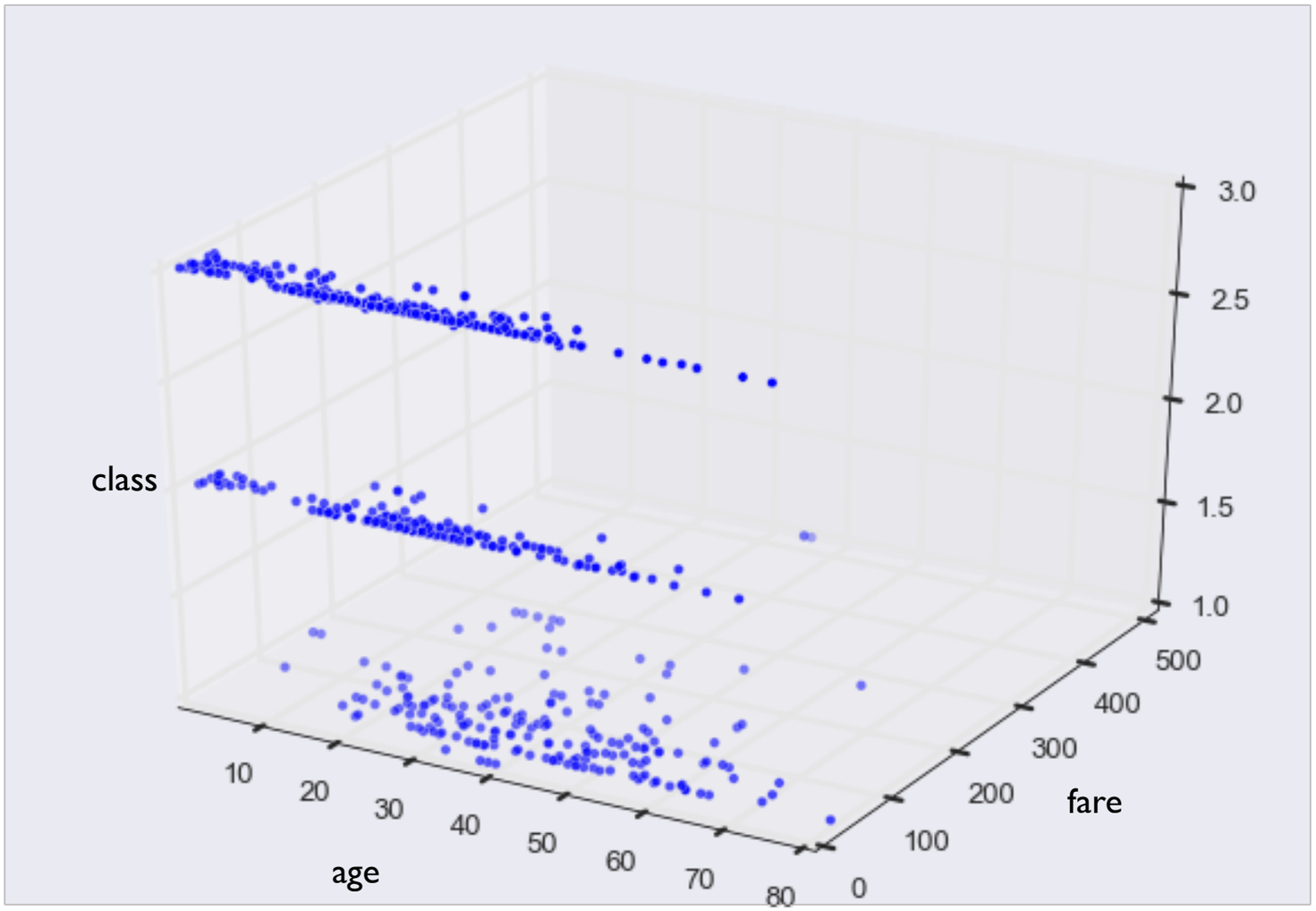


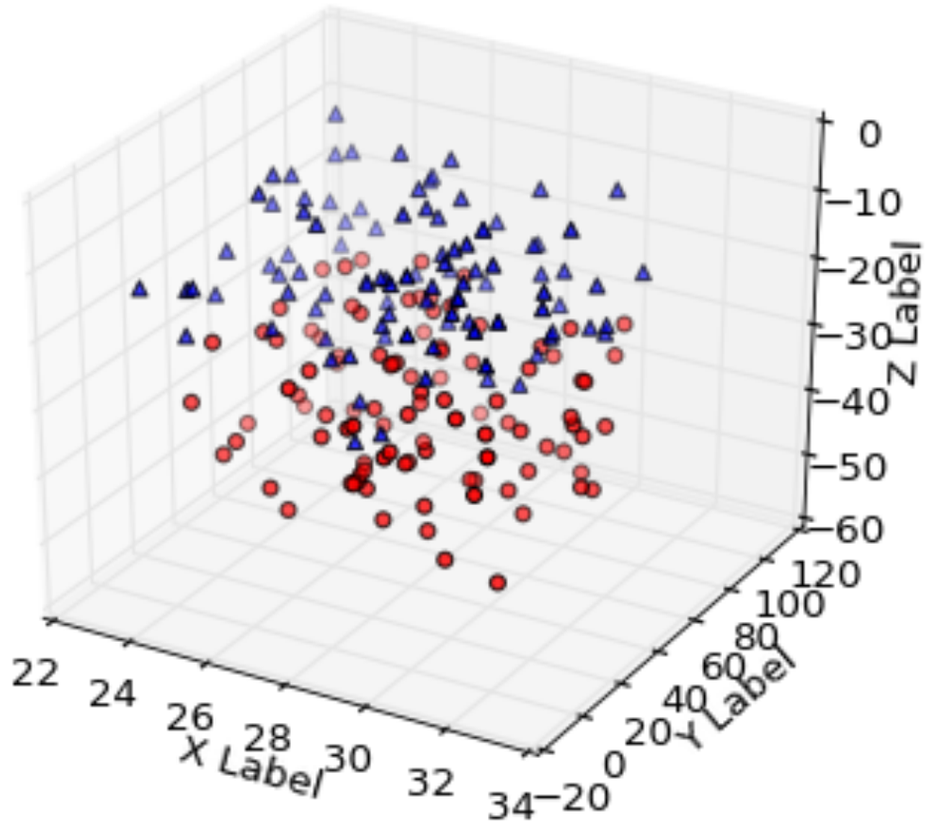
alpha = 1/100

Trend Lines



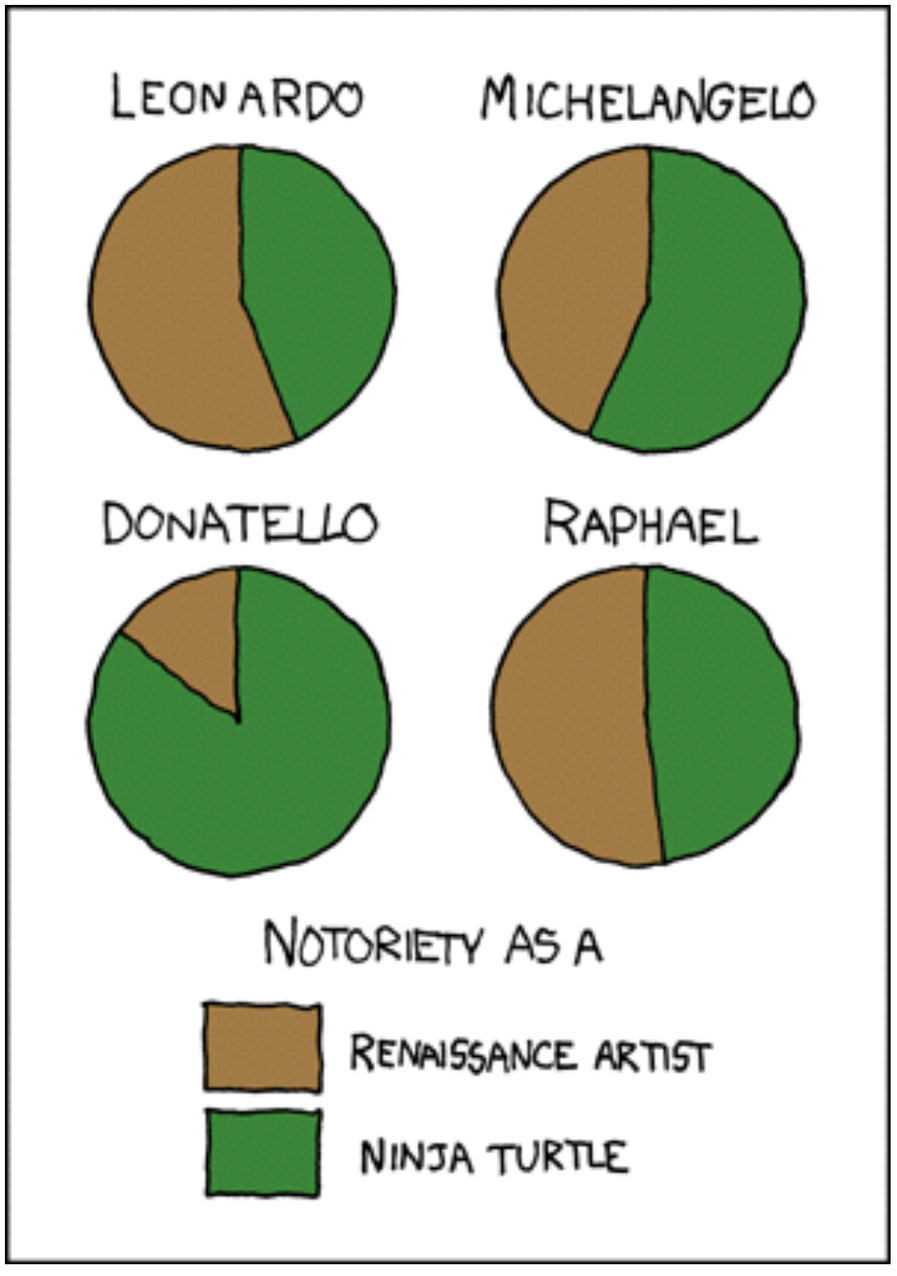
Don‘t


class
age
fare
10
20
30
40
50
60
70
80
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0

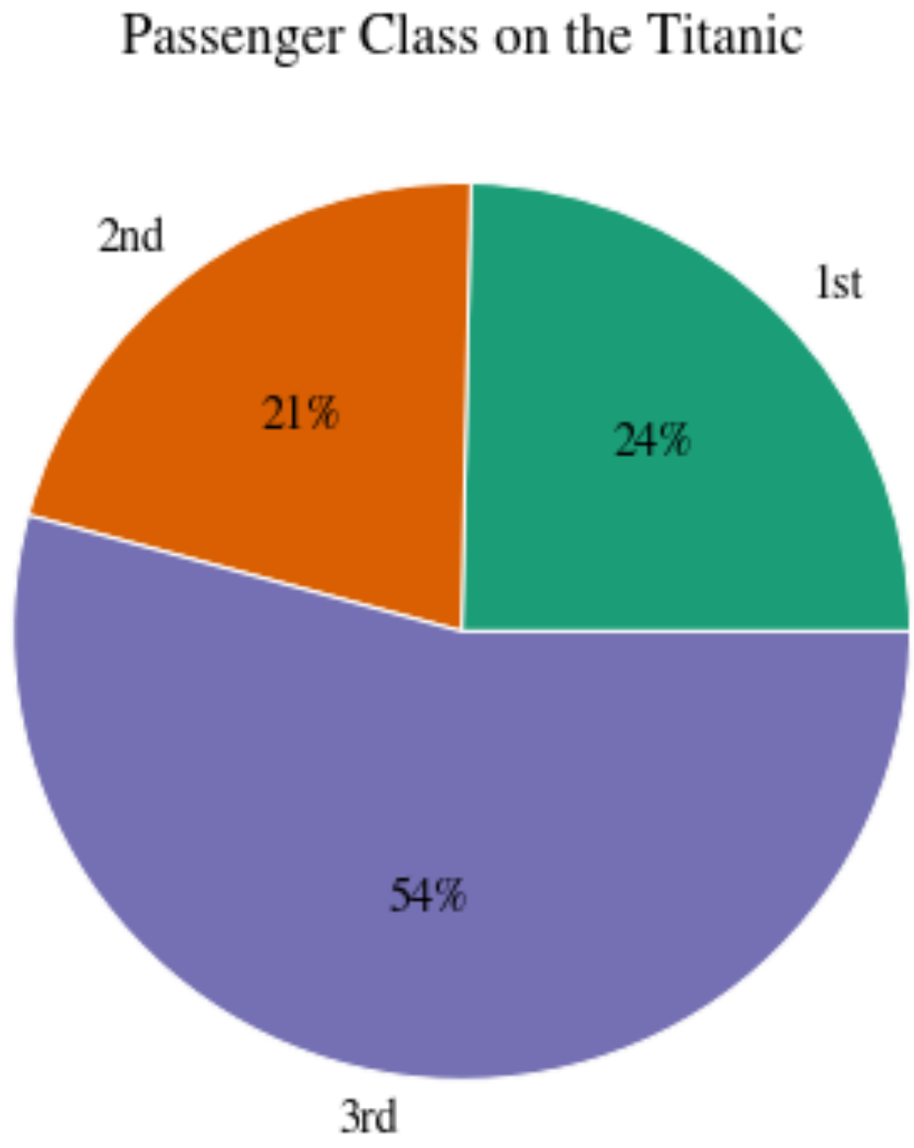

X Label
Y Label
Z Label
22
24
26
28
30
32
34
−20
0
20
40
60
80
100
120
0
−10
−20
−30
−40
−50
−60

Compositions

Pie Charts

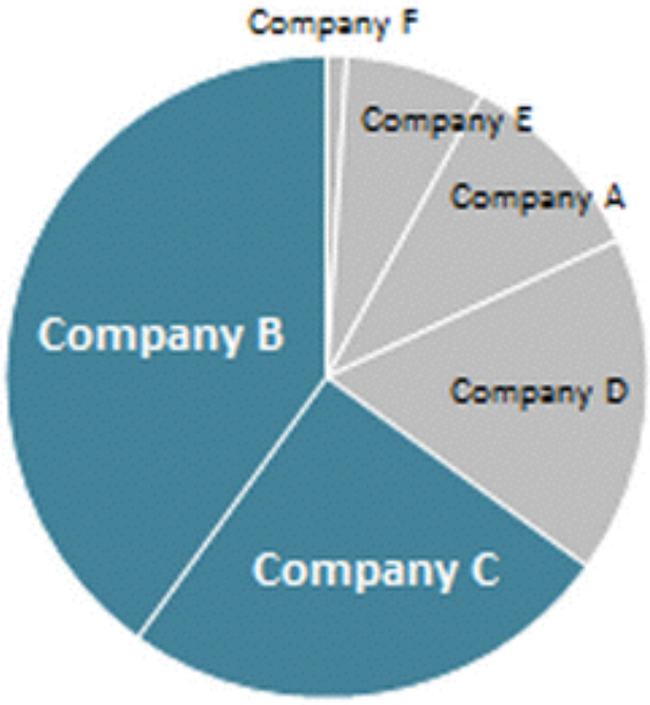


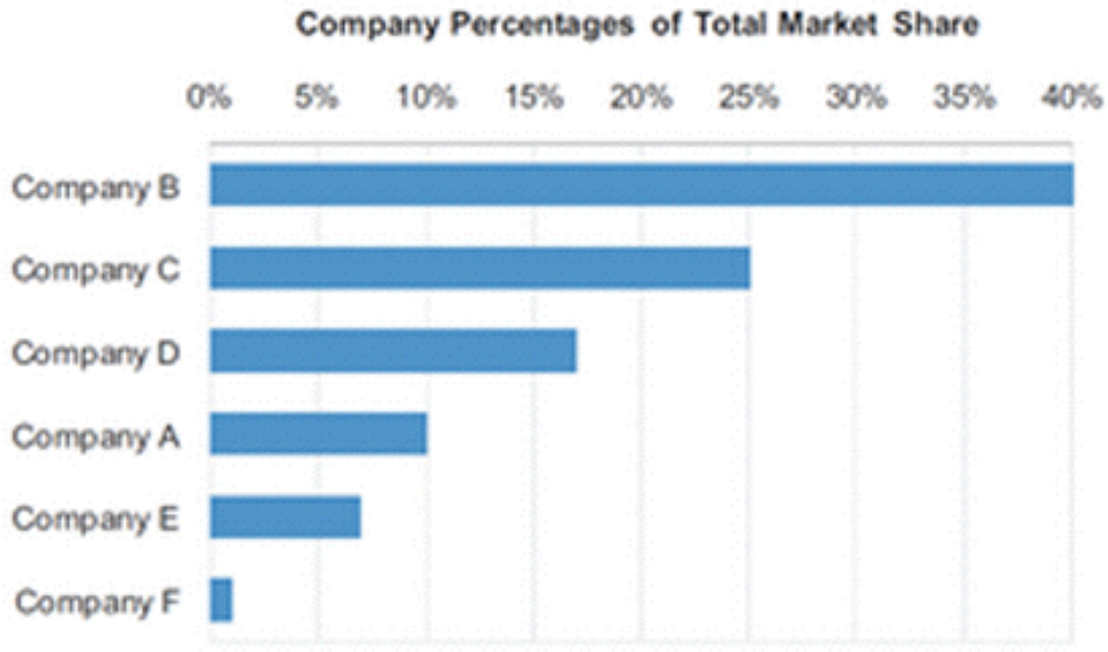
http://xkcd.com/197/



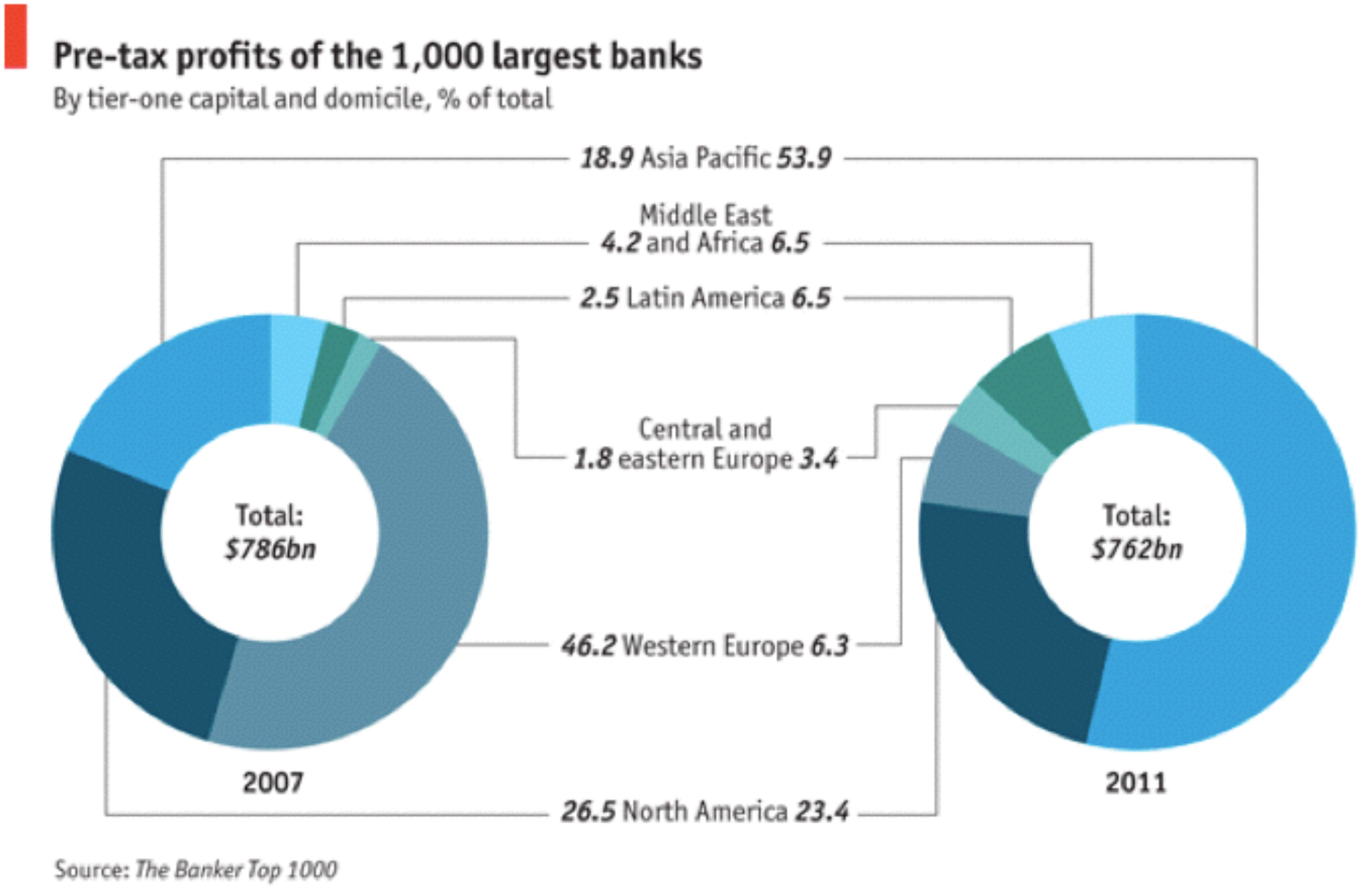
Pie vs. Bar Charts

65% of the market is controlled by companies B and C

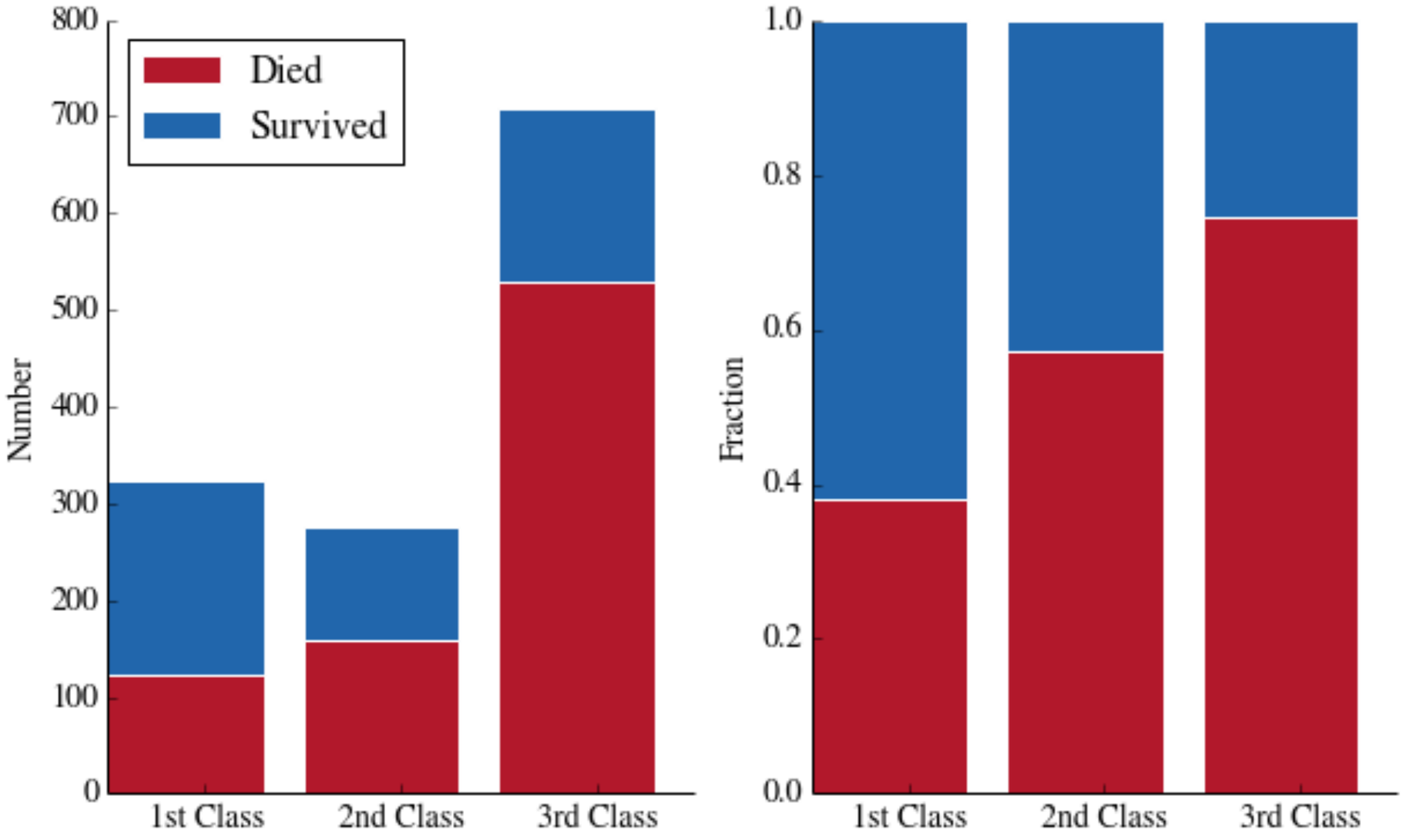




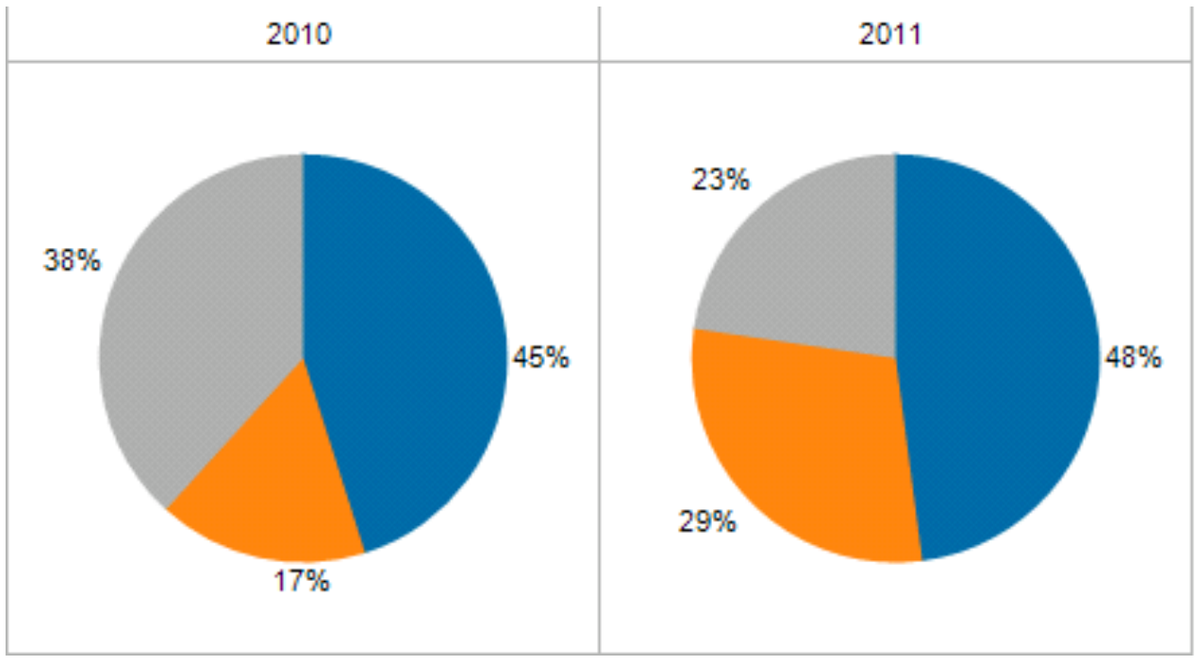
Donut Chart


Pre-tax profits of the 1,000 largest banks
By tier-one capital and domicile, % of total
18.9 Asia Pacific 53.9
Middle East
4.2 and Africa 6.5
2.5 Latin America 6.5
Central and
1.8 eastern Europe 3.4
46.2 Western Europe 6.3
26.5 North America 23.4
Total:
$786bn
2007
Total:
$762bn
2011
Source: The Banker Top 1000

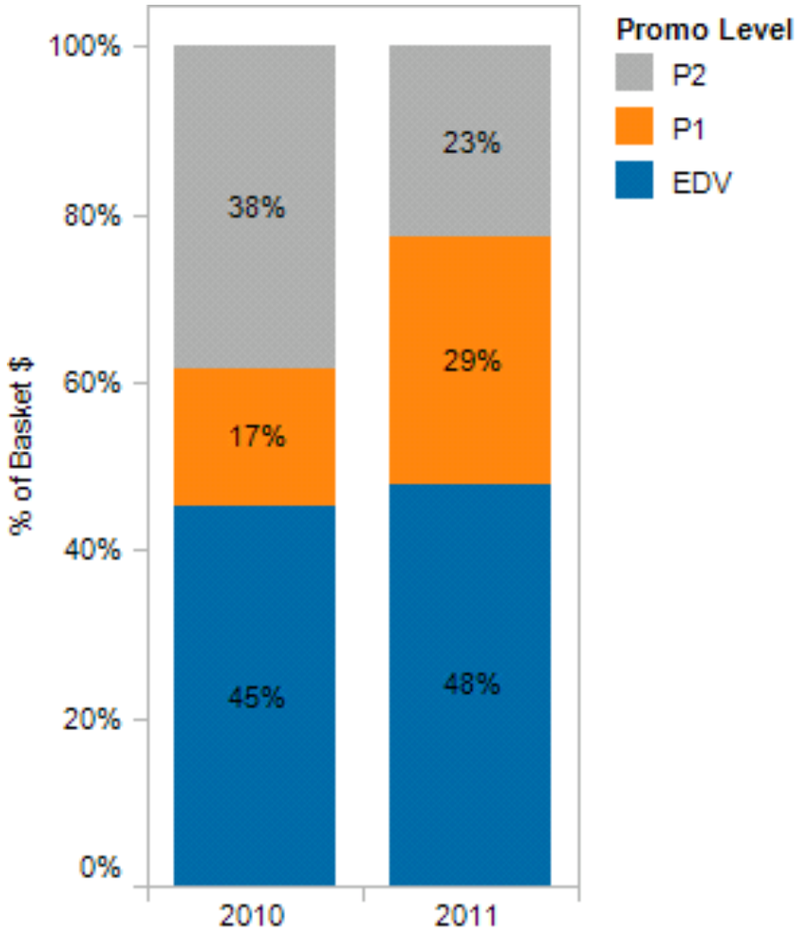
Stacked Bar Chart



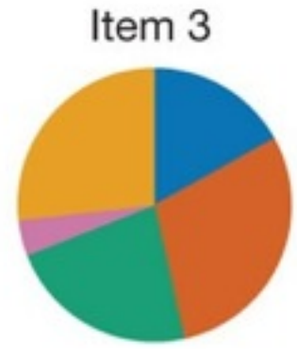
Stacked Bar Chart



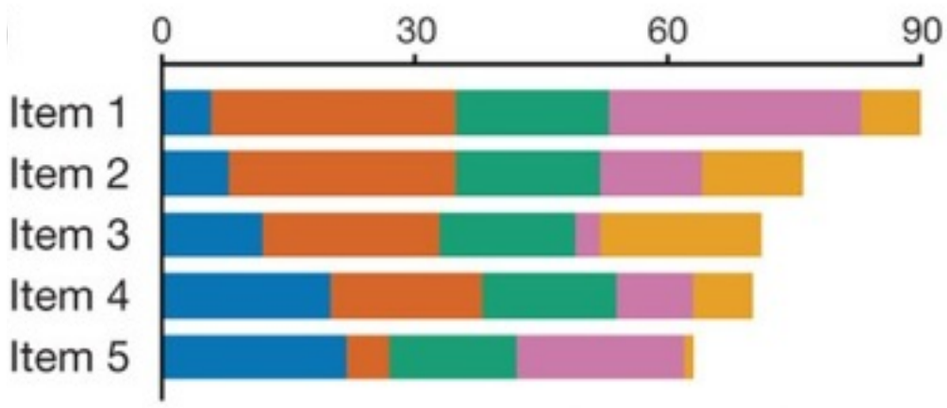
vs.



Comparison of bar chart types

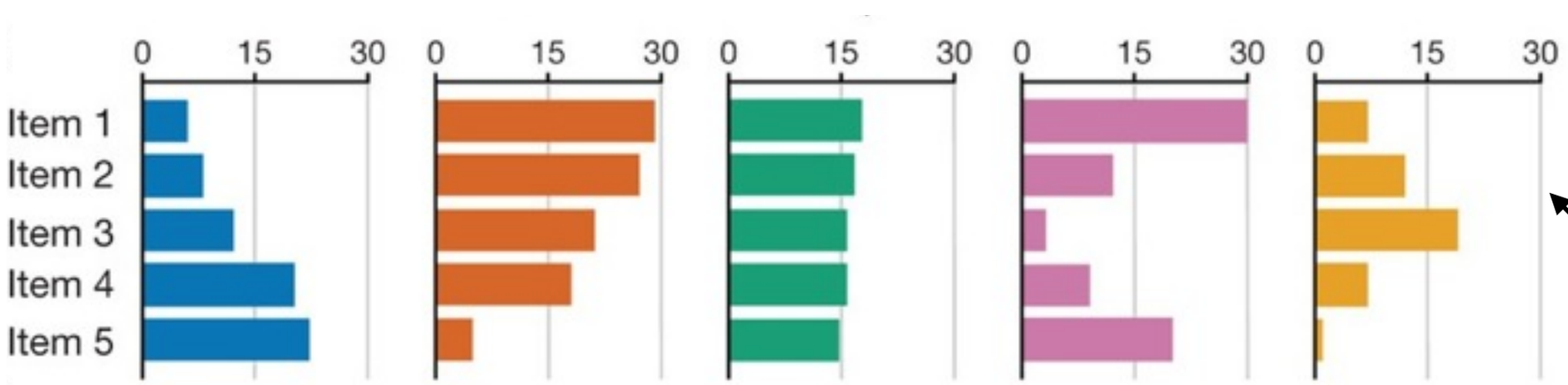


Pie Chart



Stacked bar chart

Layered Bar Chart

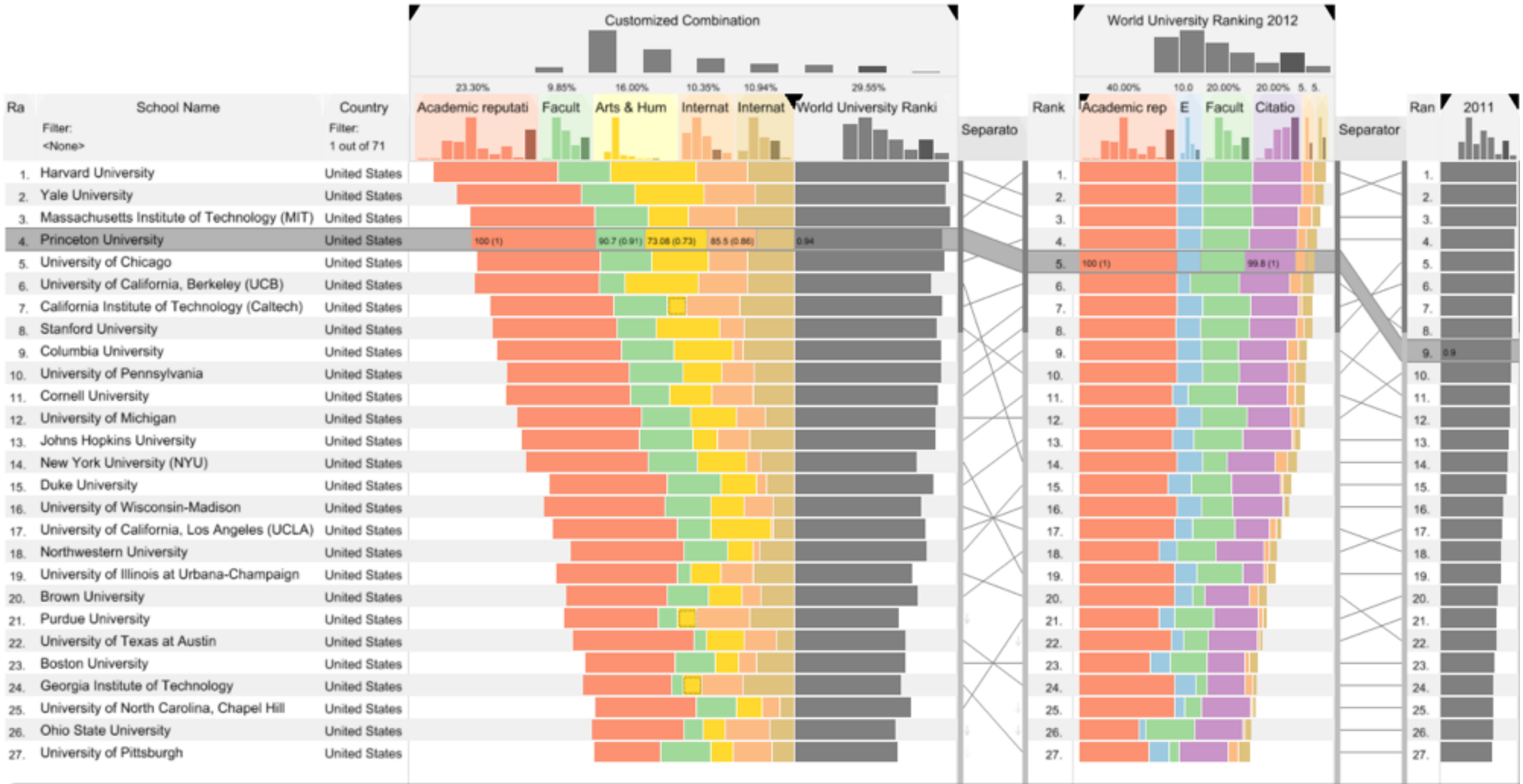


Grouped Bar Chart

Small Multiples

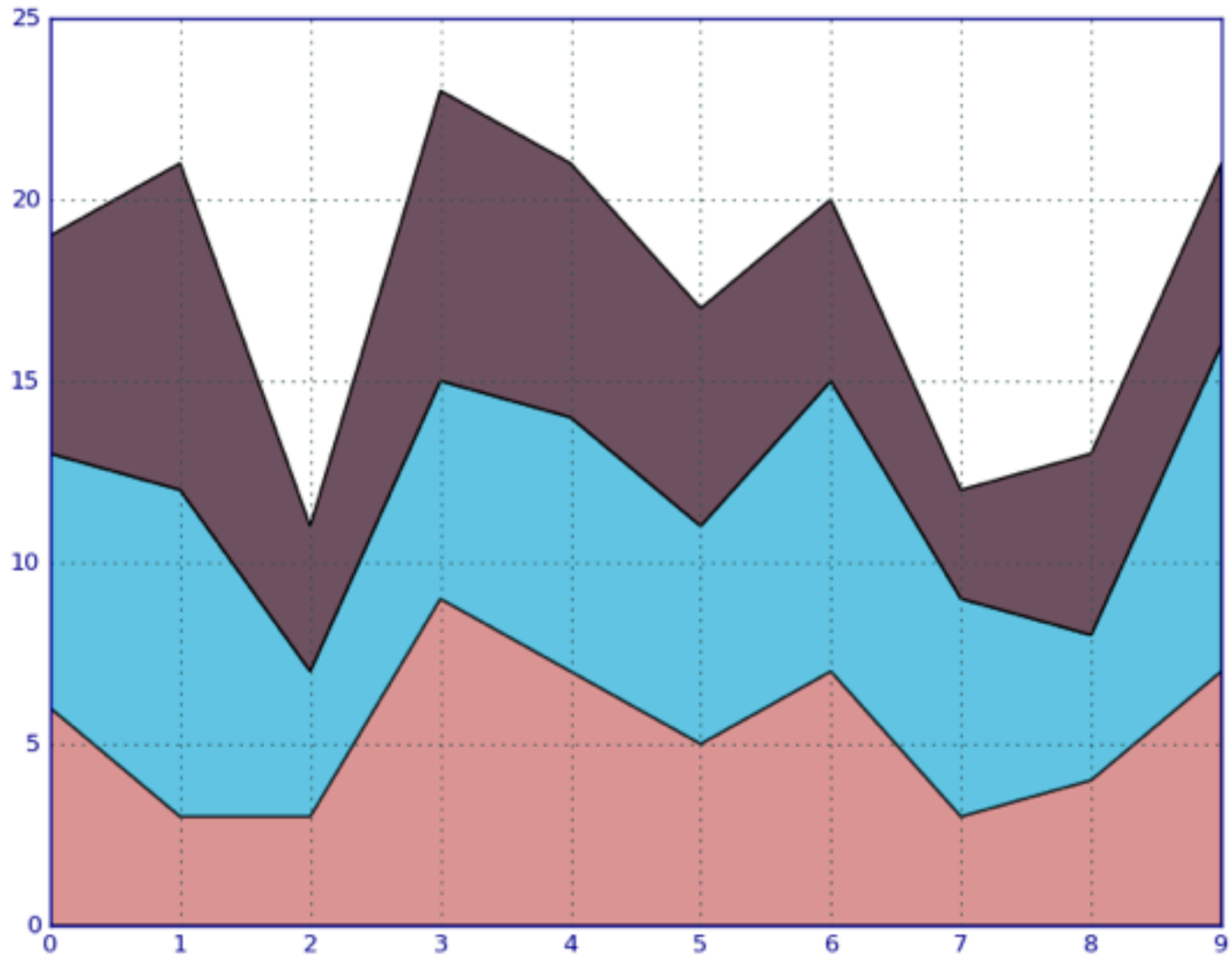
Streit & Gehlenborg, PoV, Nature Methods, 2014

LineUp



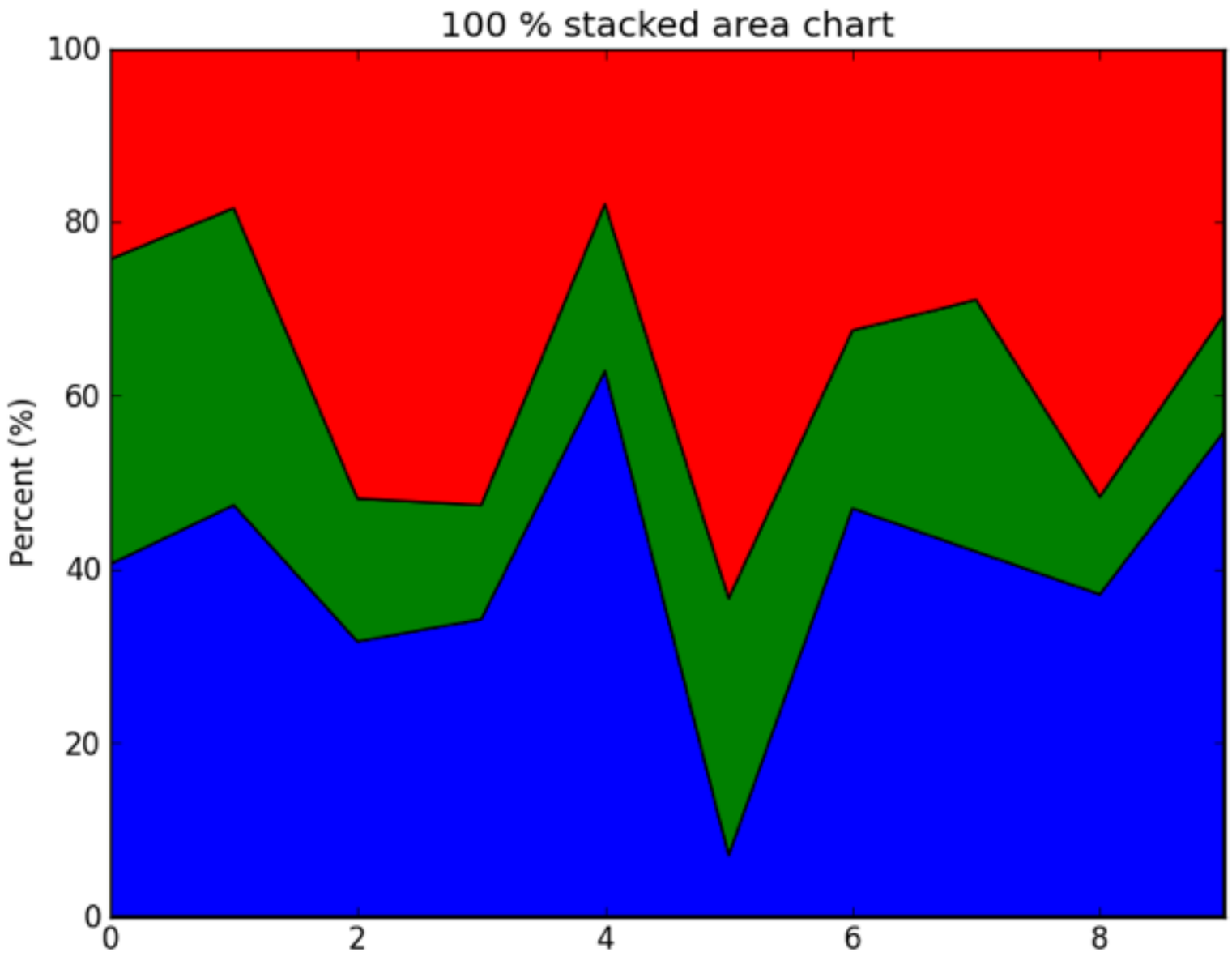
Video at http://lineup.caleydo.org

Stacked Area Chart



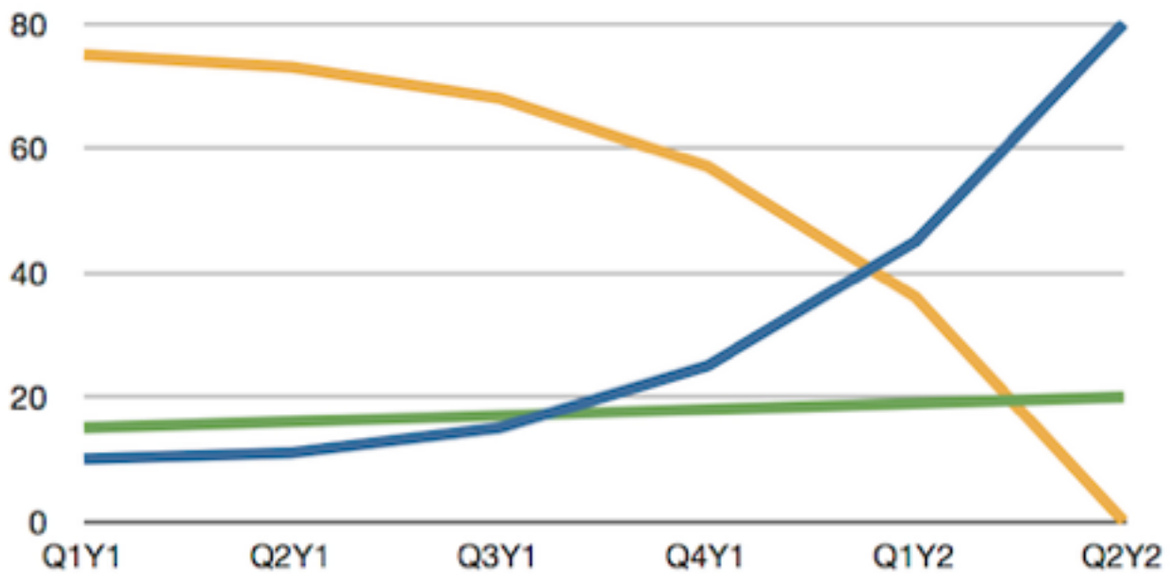
http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

100% Stacked Area Chart

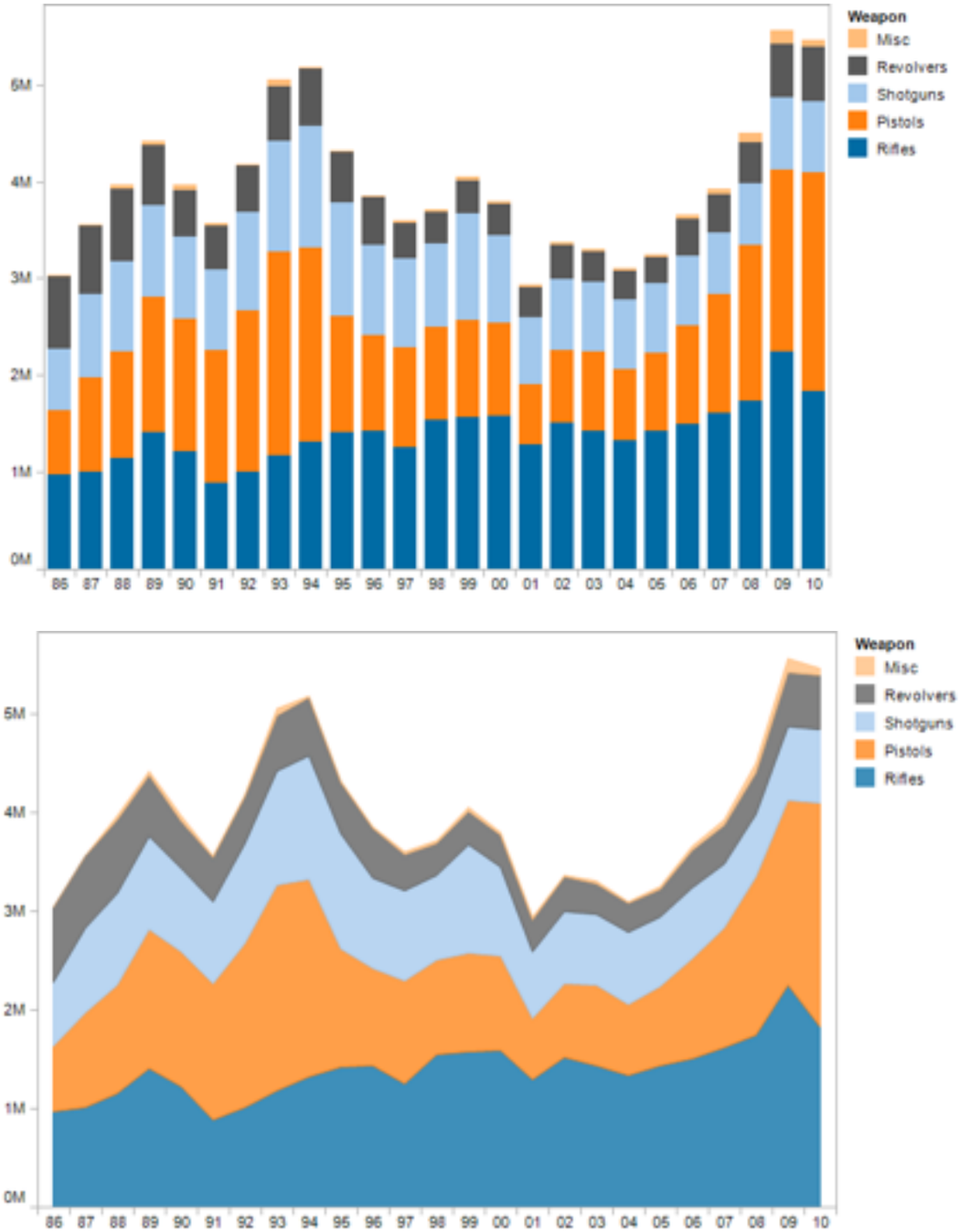


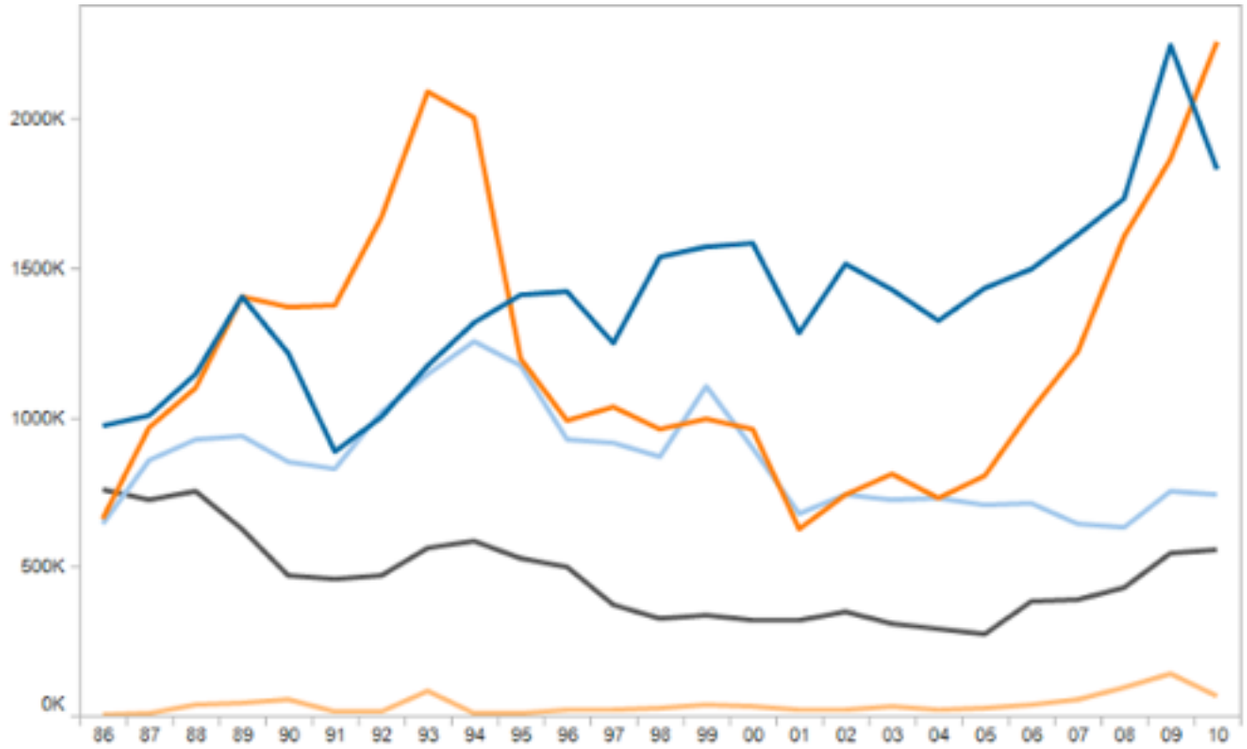
http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

Stacked Area vs. Line Graphs



leancrew.com &
Practically Efficient



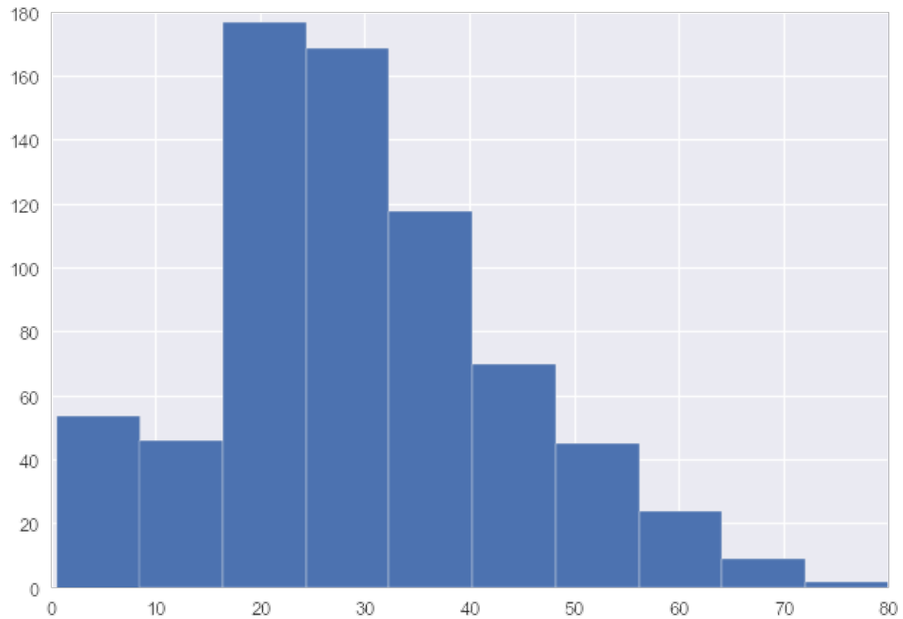


VizWiz, A. Kriebel

Distributions

Histogram

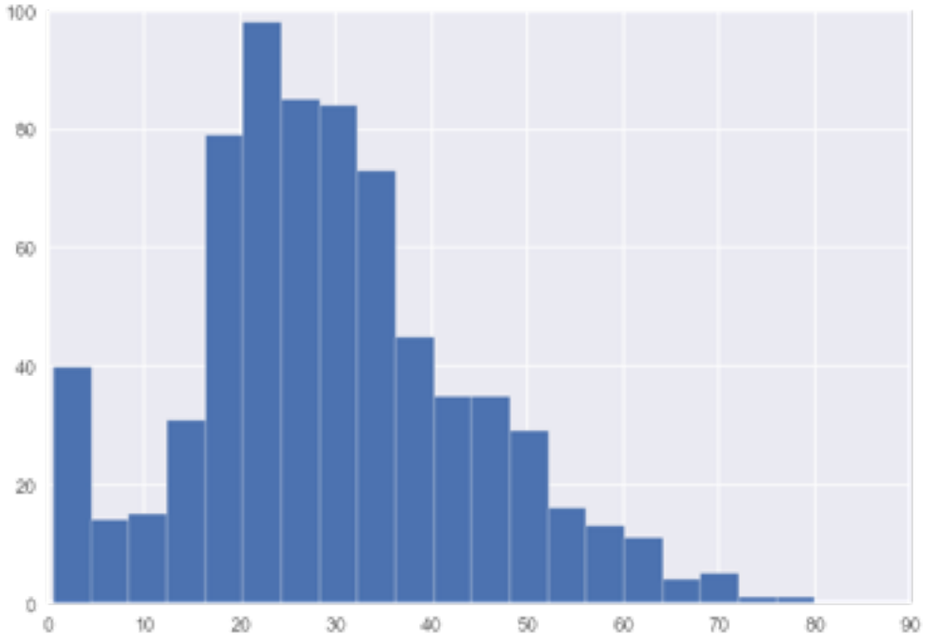
passengers



age

10 Bins

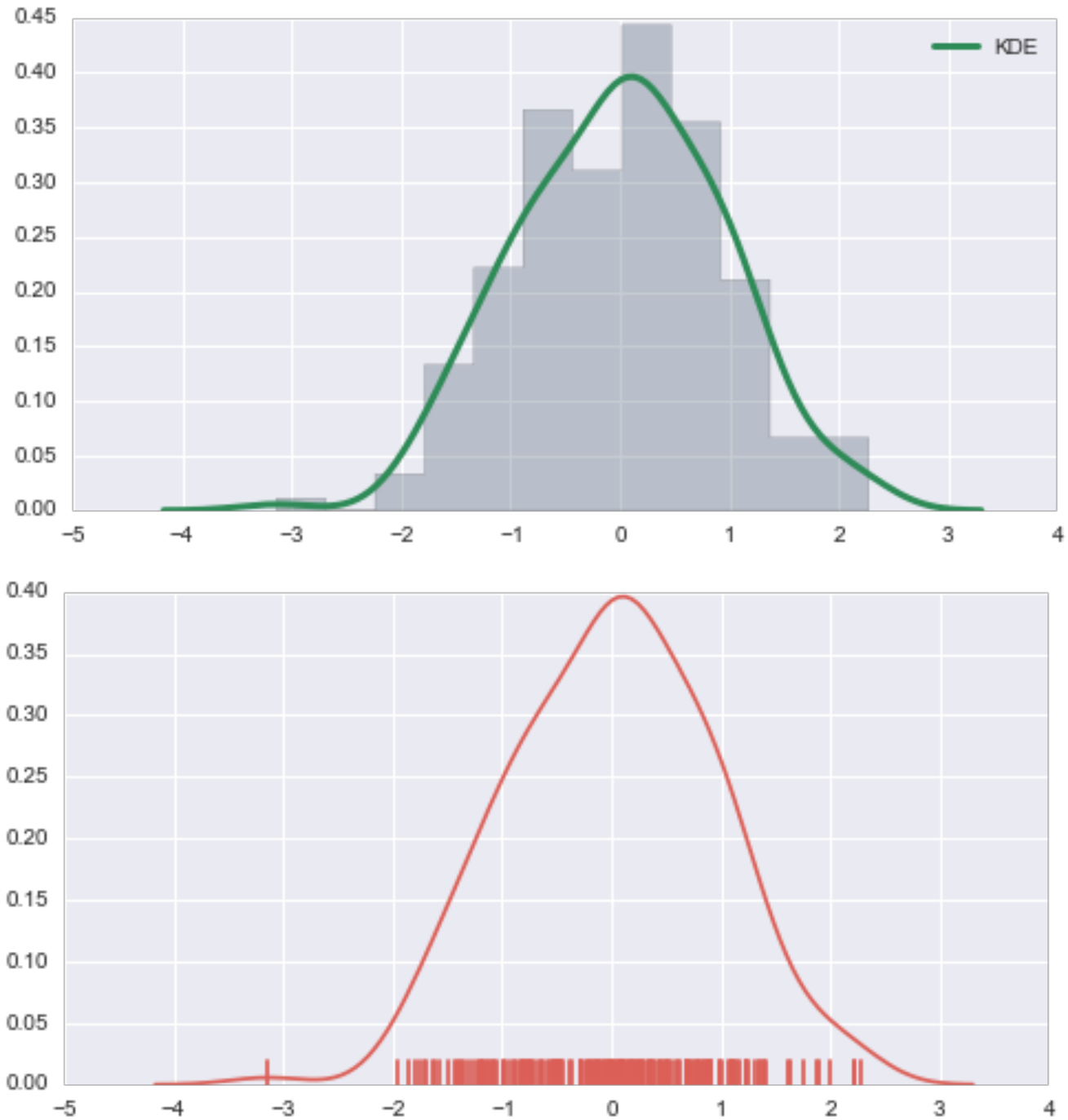
passengers



age

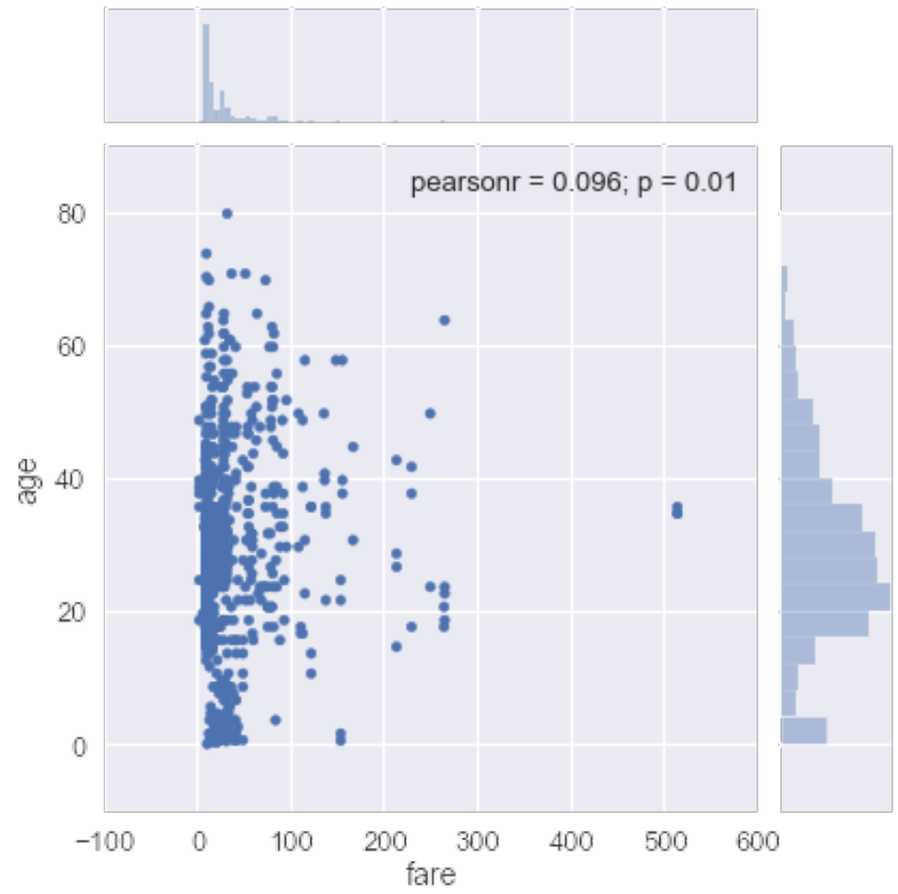
20 Bins

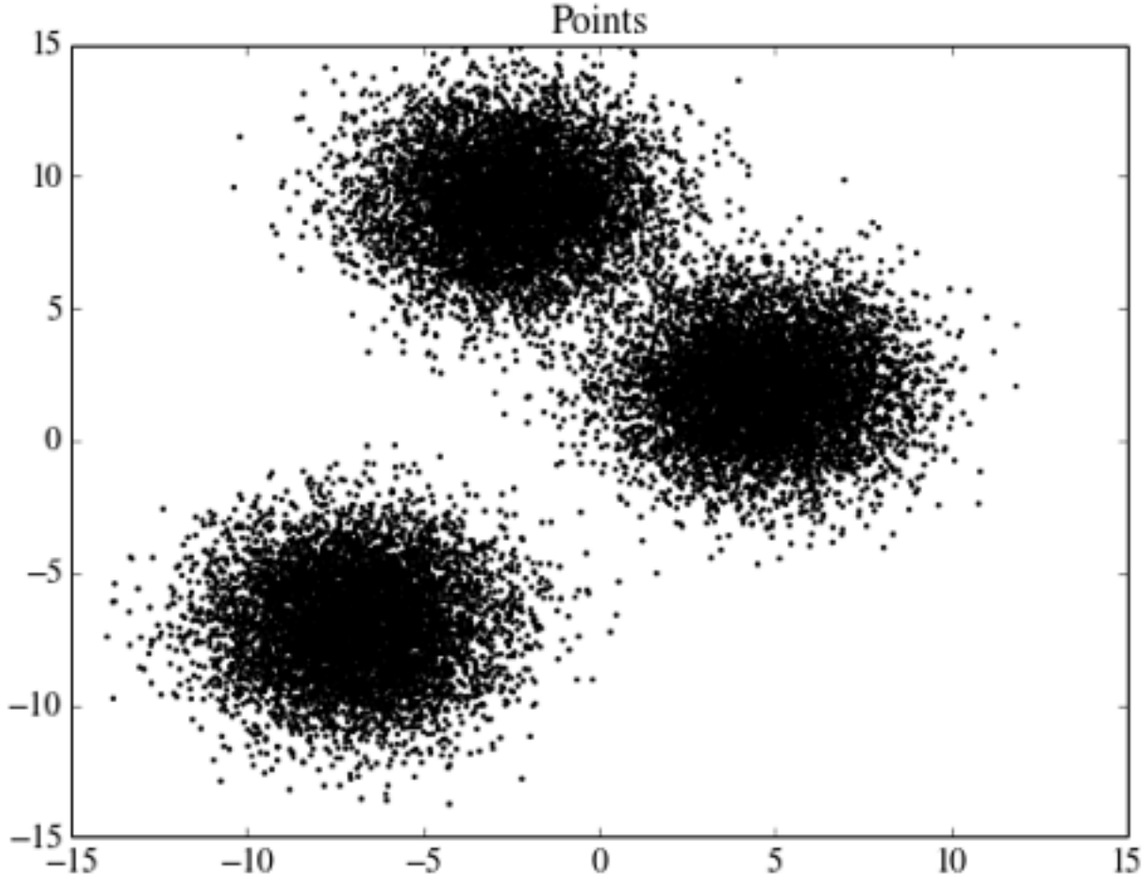
Density Plots

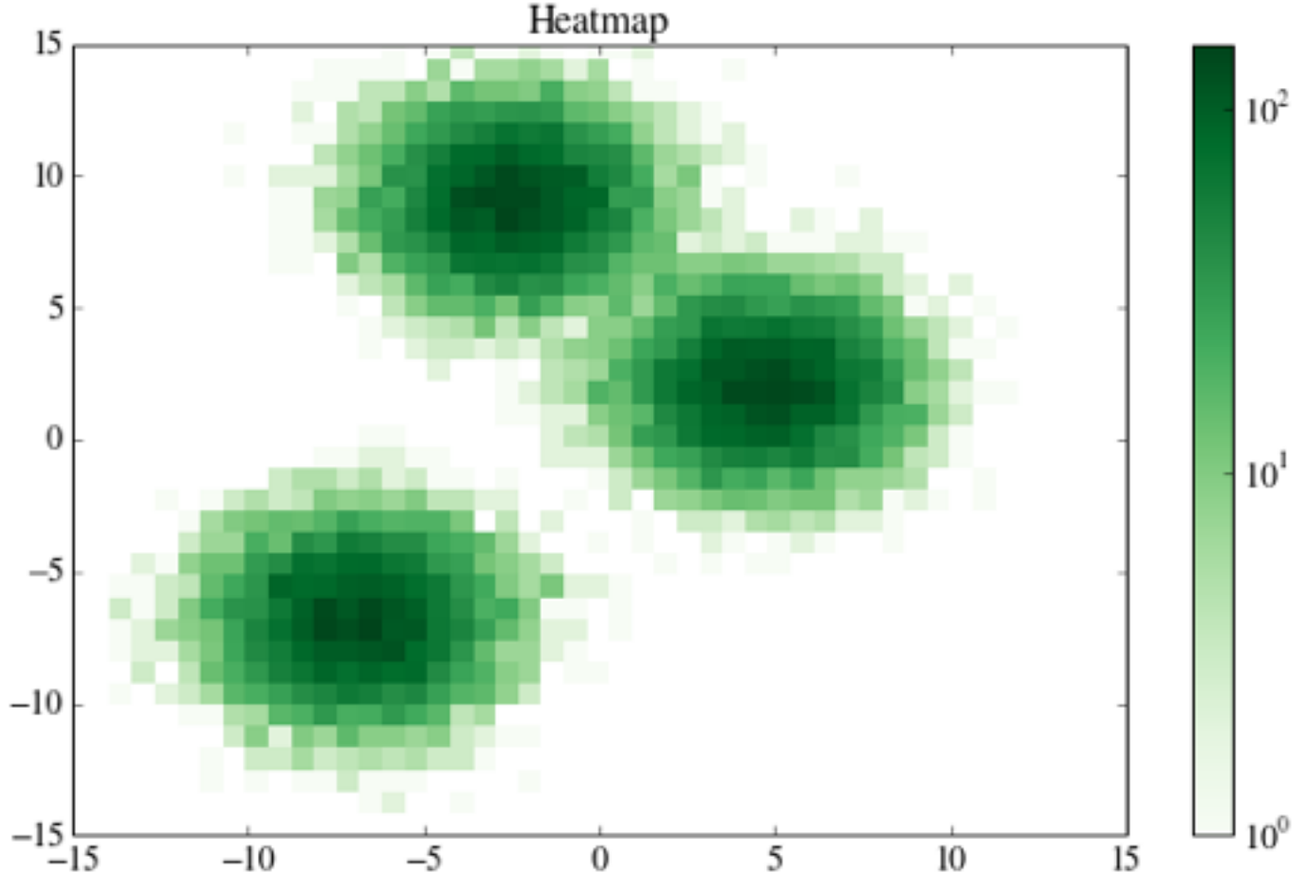


http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Heat Maps



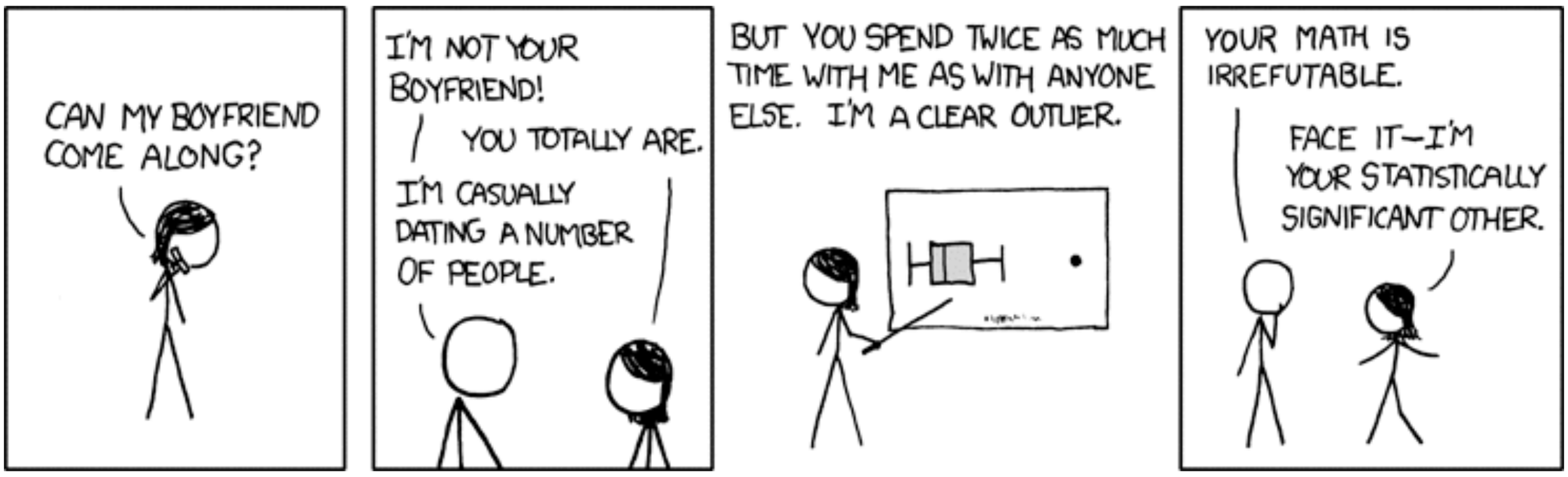




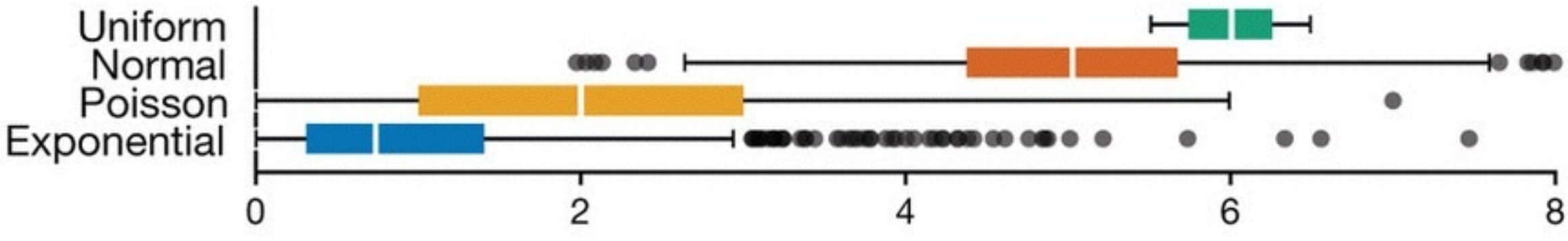
2D Density Plots

Box Plots

aka Box-and-Whisker Plot

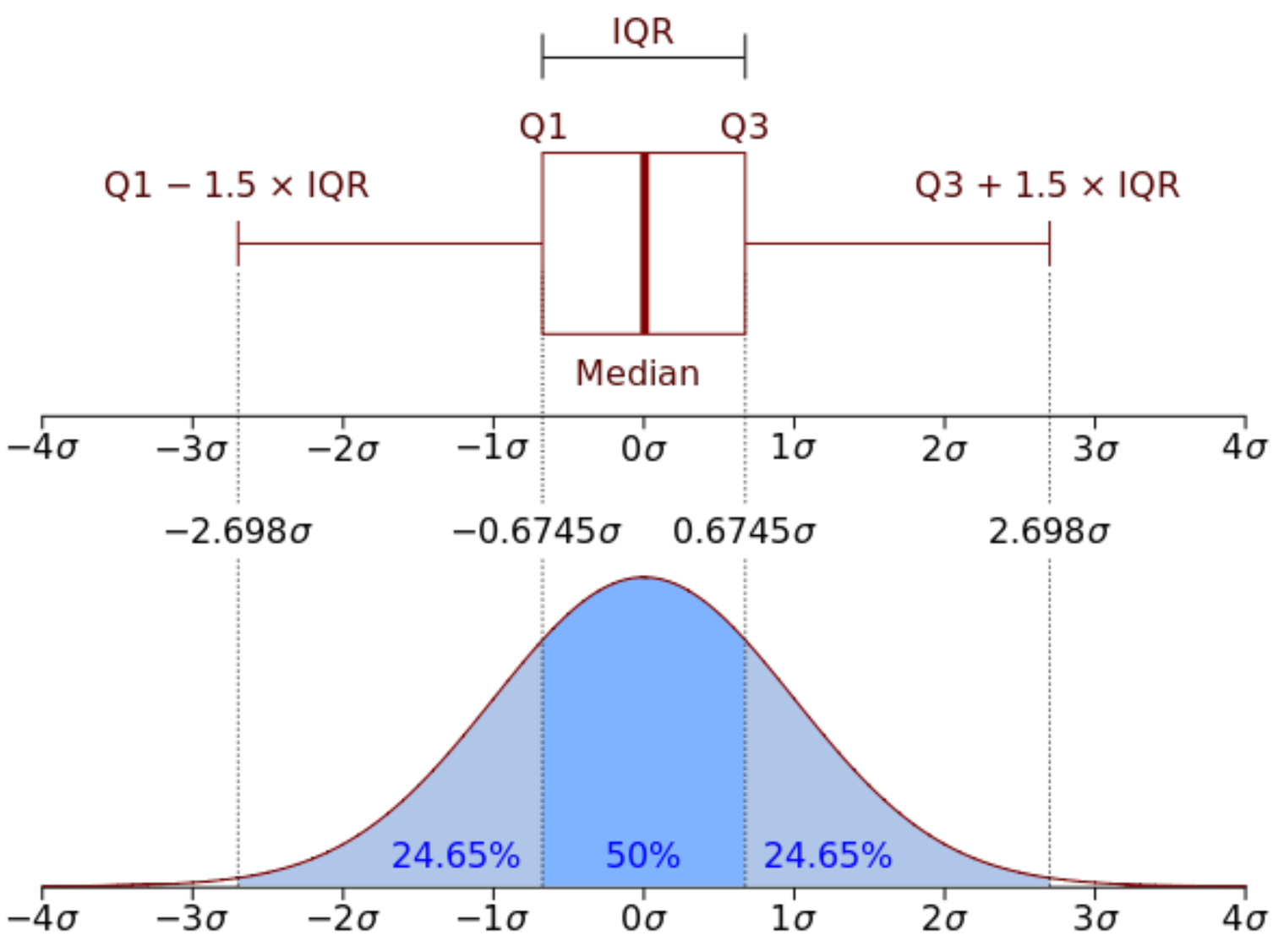


http://xkcd.com/539/



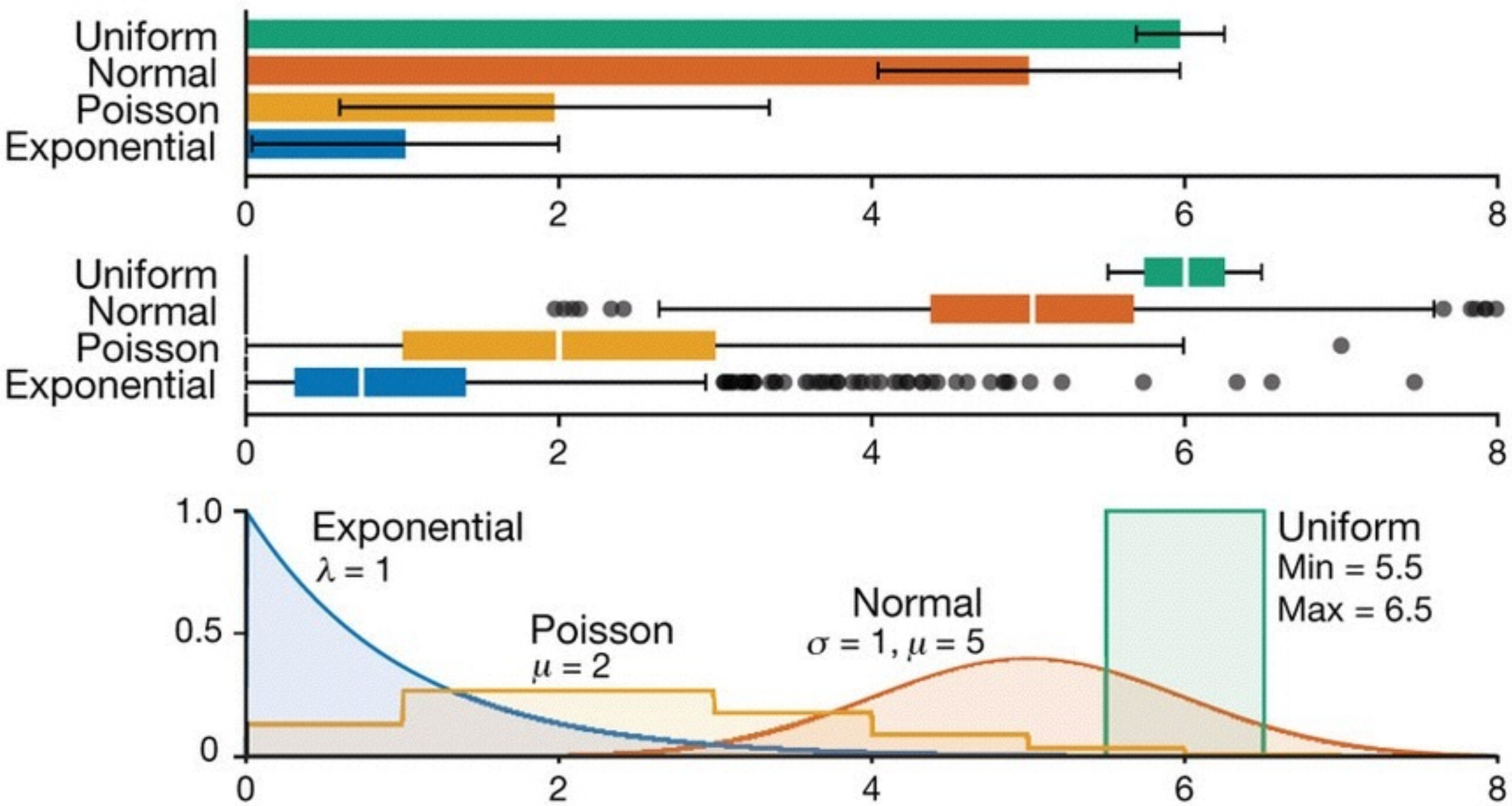
Box Plots

aka Box-and-Whisker Plot



Wikipedia

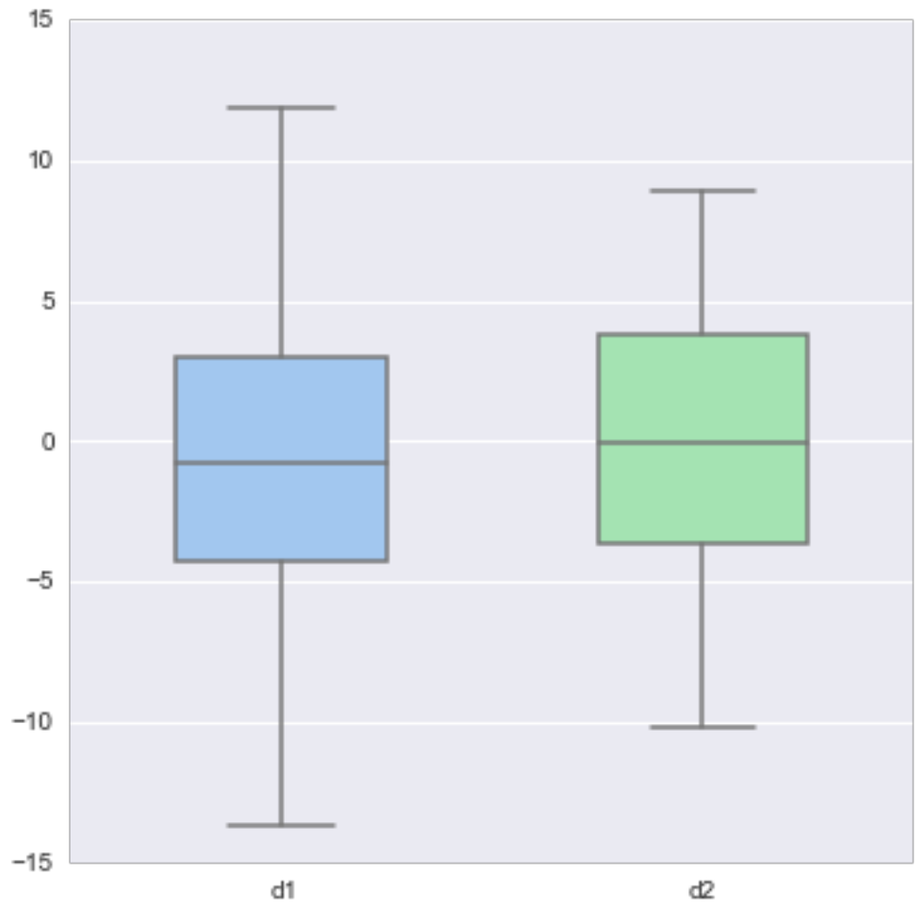
Comparison

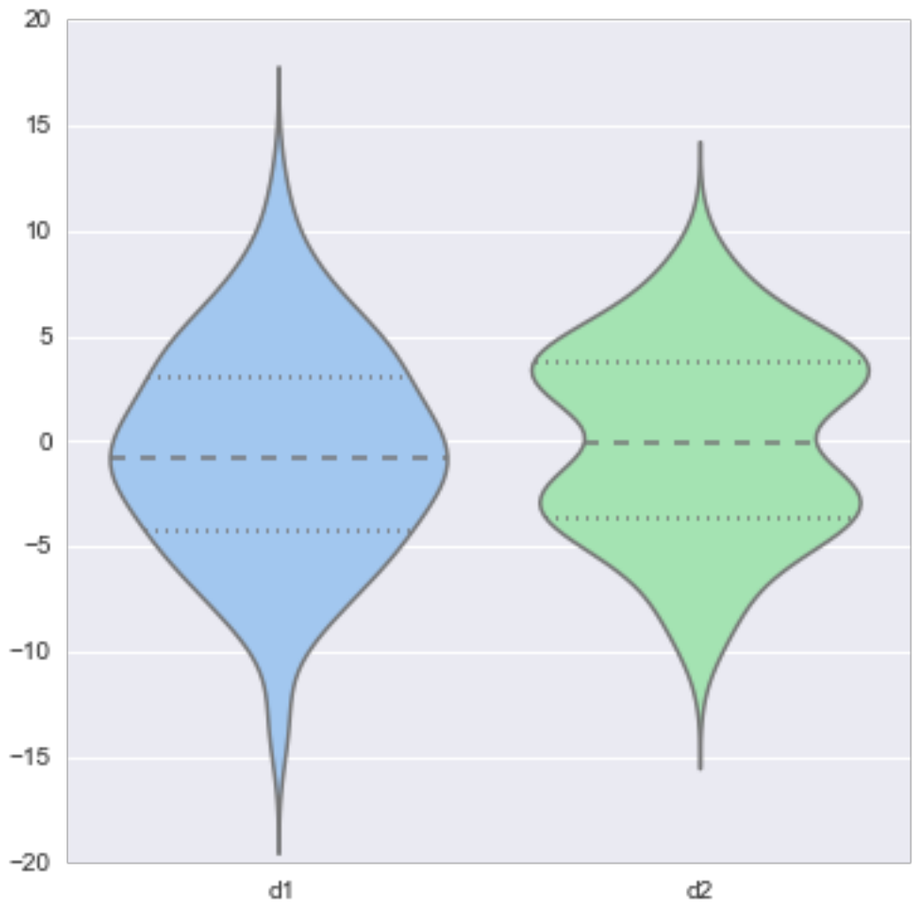


Streit & Gehlenborg, PoV, Nature Methods, 2014

Violin Plot

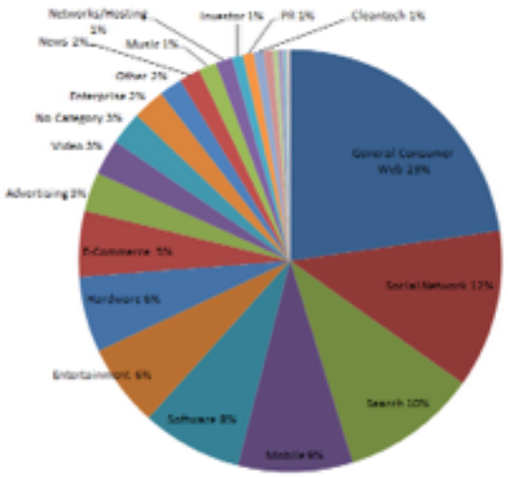
= Box Plot + Probability Density Function



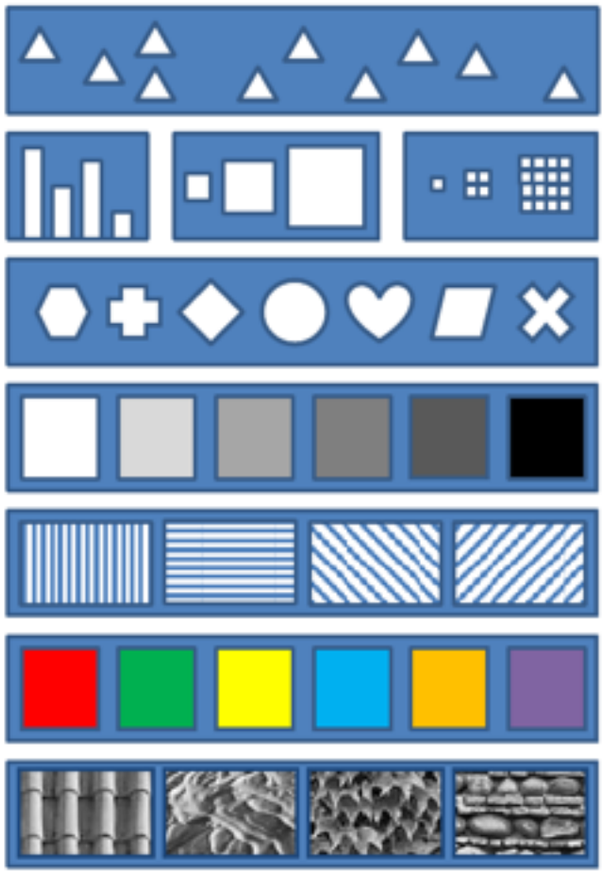


http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

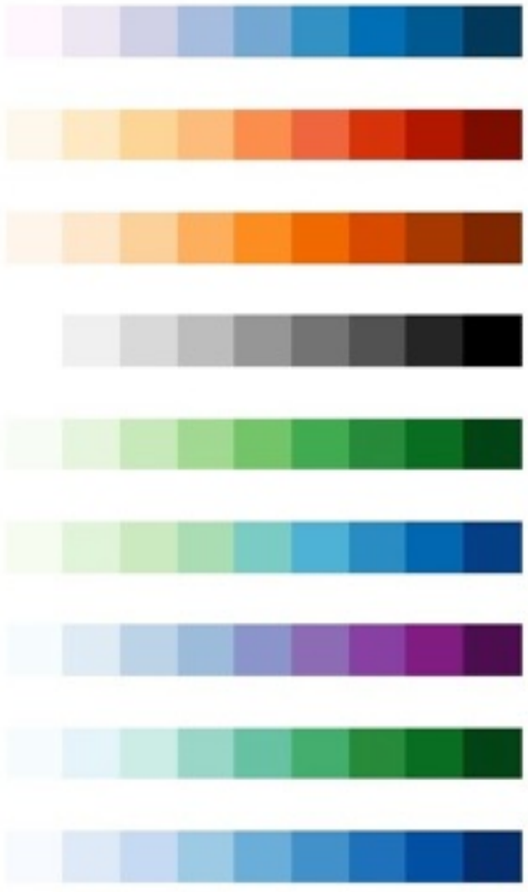
On Thursday…



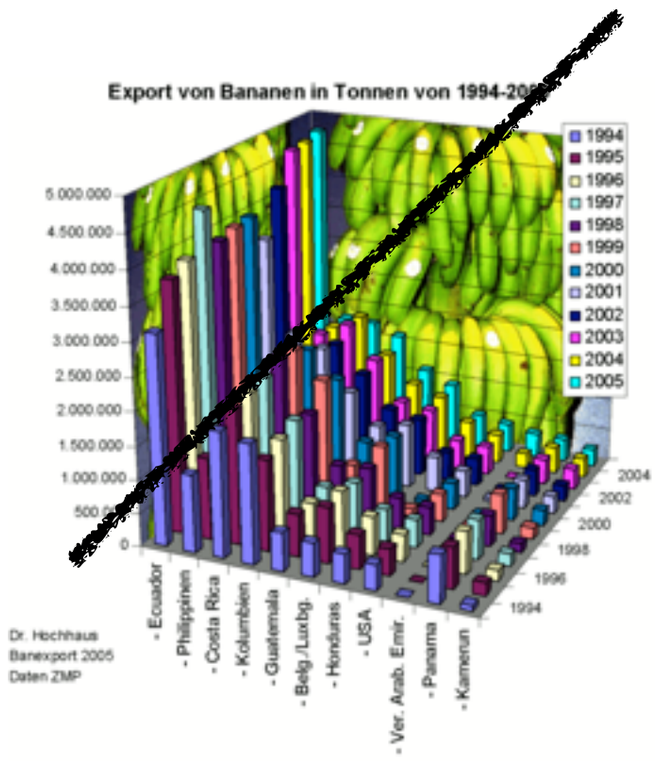
Effective vs. Non Effective Visualizations



Visual Attributes



Color



Design Principles

Questions