

CS 109: Data Science

Vis. Goals, Data Types, Statistical Graphs

Marc Streit

mstreit@seas.harvard.edu

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

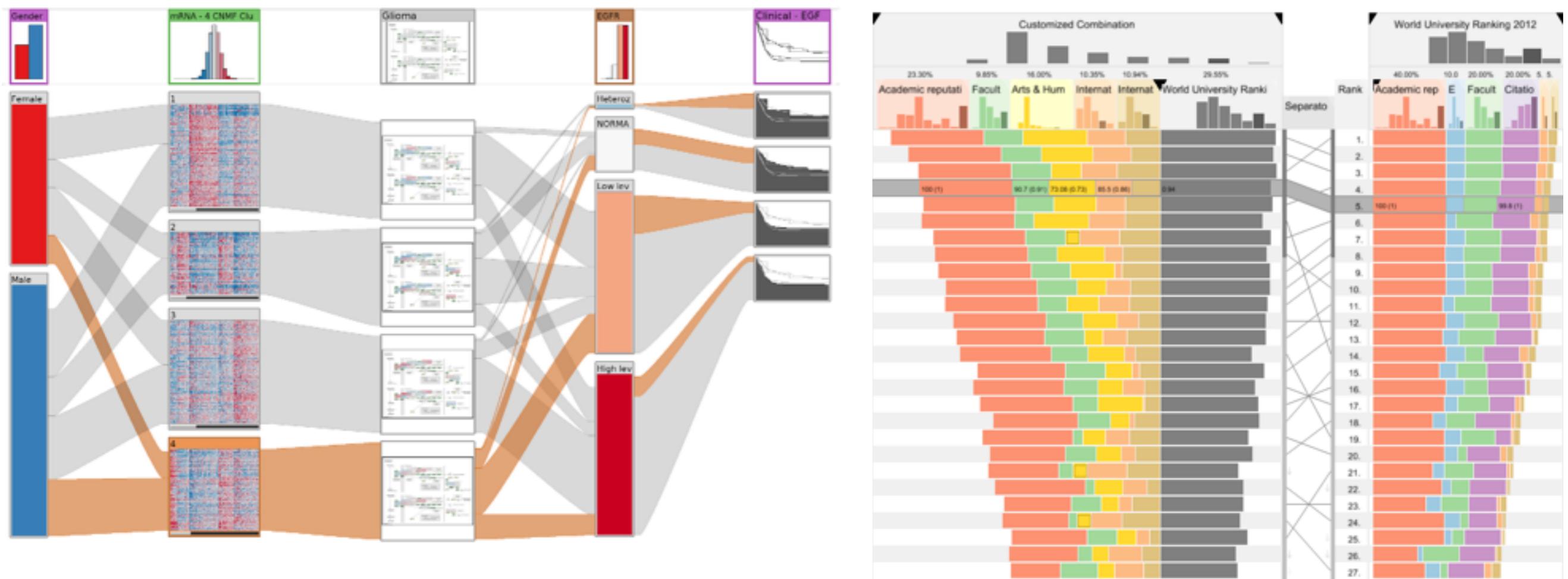
Visiting Professor @ Harvard SEAS Pfister Lab

Research Focus

Information Visualization / Visual Analytics

Biological Data Visualization

Caleydo Framework (www.caleydo.org)

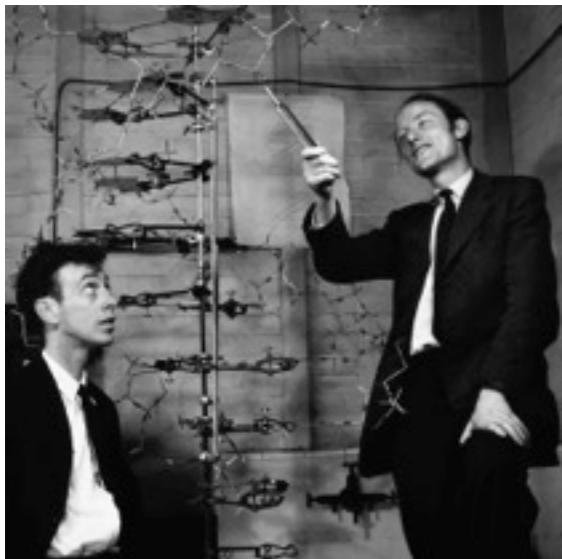


This Week

HW0 - due today Tuesday (not graded)

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

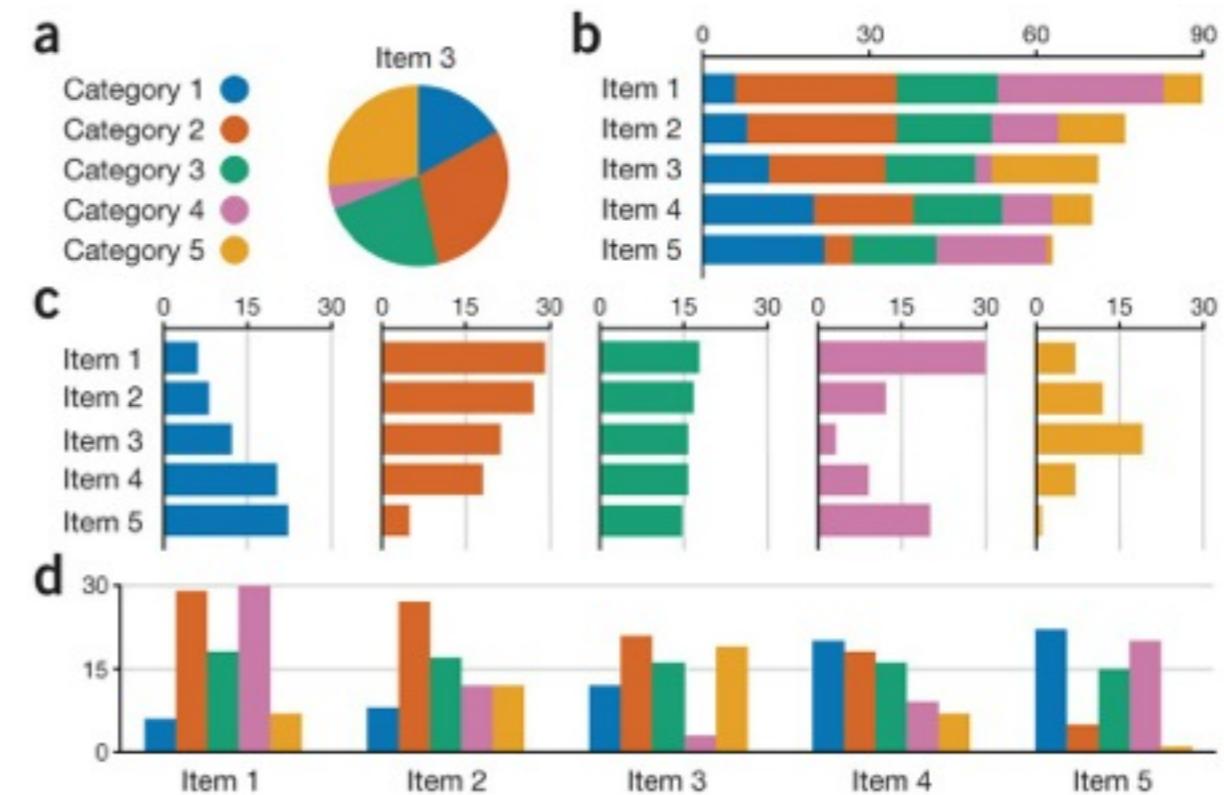
Outline



Vis Goals

Data Types
Dimensionality

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



Graph Types

Visualization Goals and Process

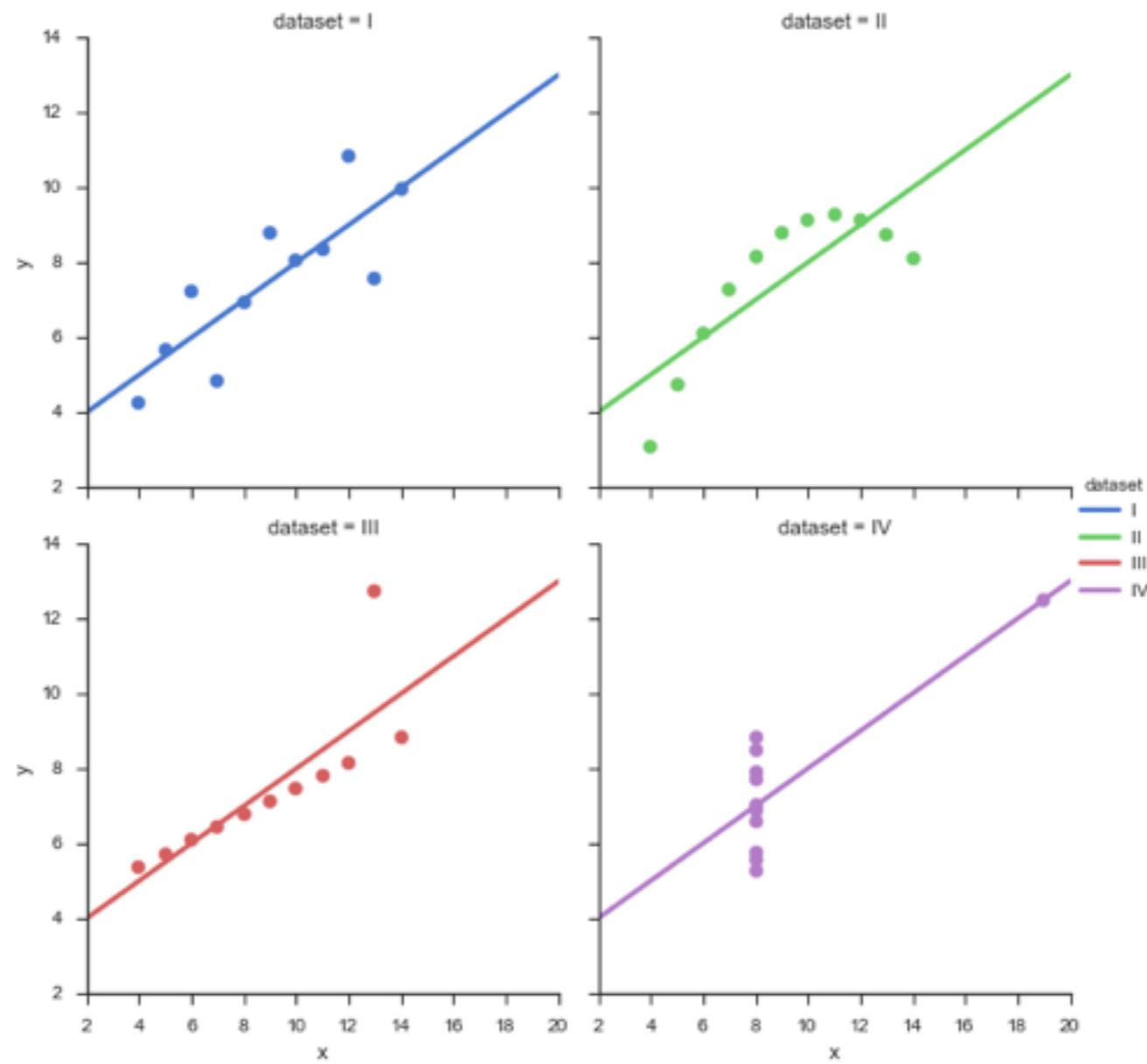
Last Week: Anscombe's Quartet

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	

Last Week: Anscombe's Quartet

Same mean, variance, correlation, and linear regression line



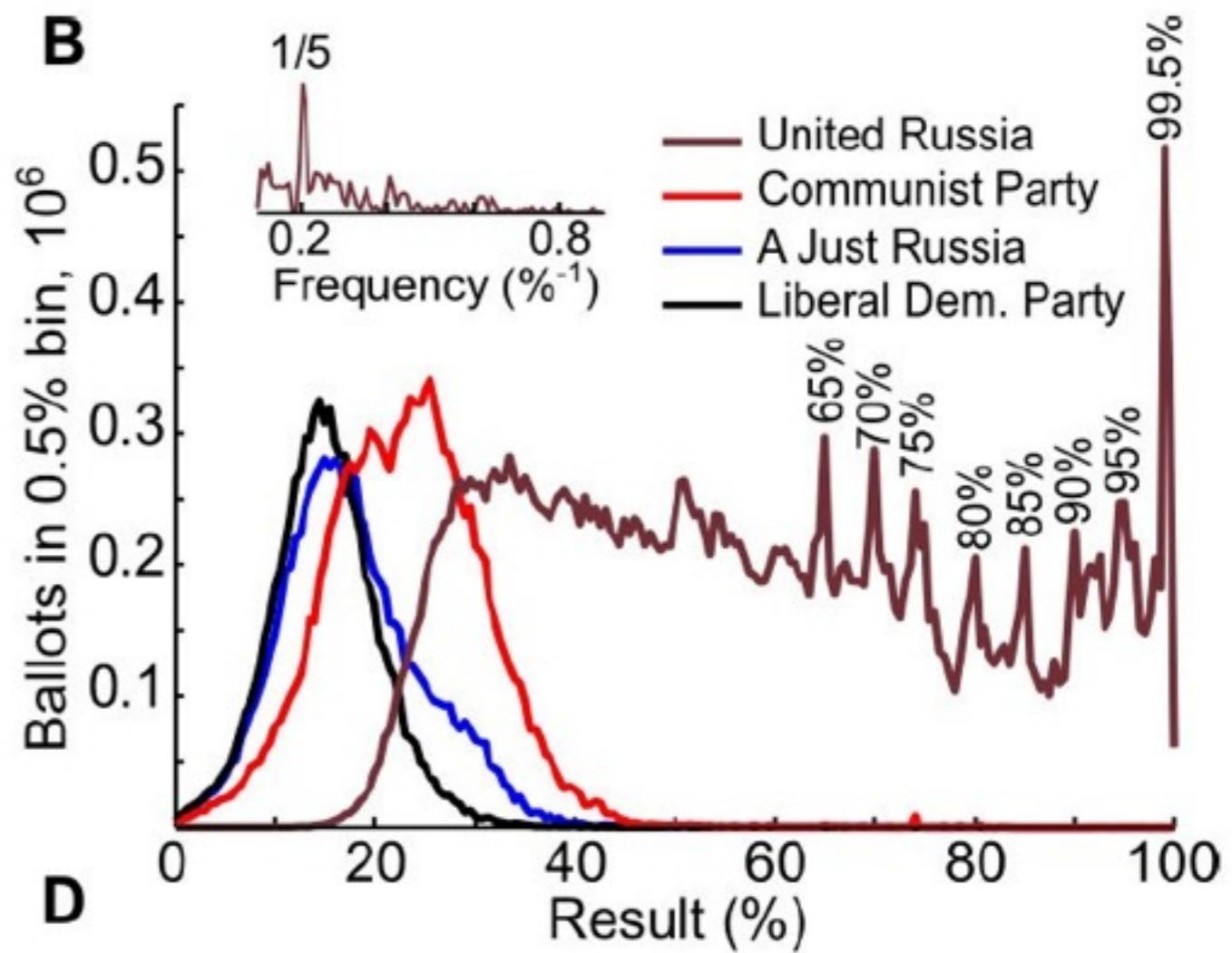
Last Week: Normal Distribution

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

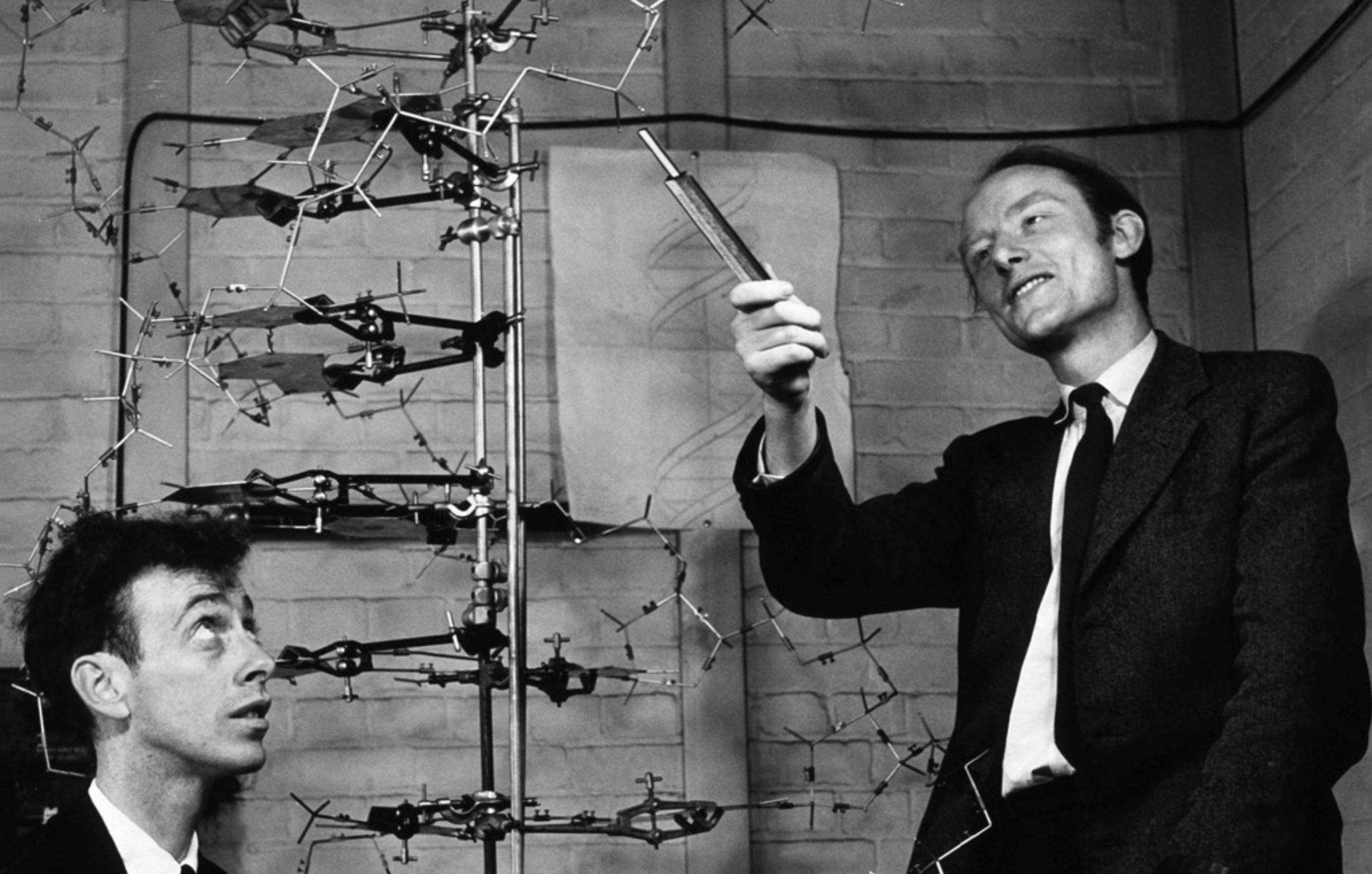
Data without hypothesis

Task: Generate hypothesis

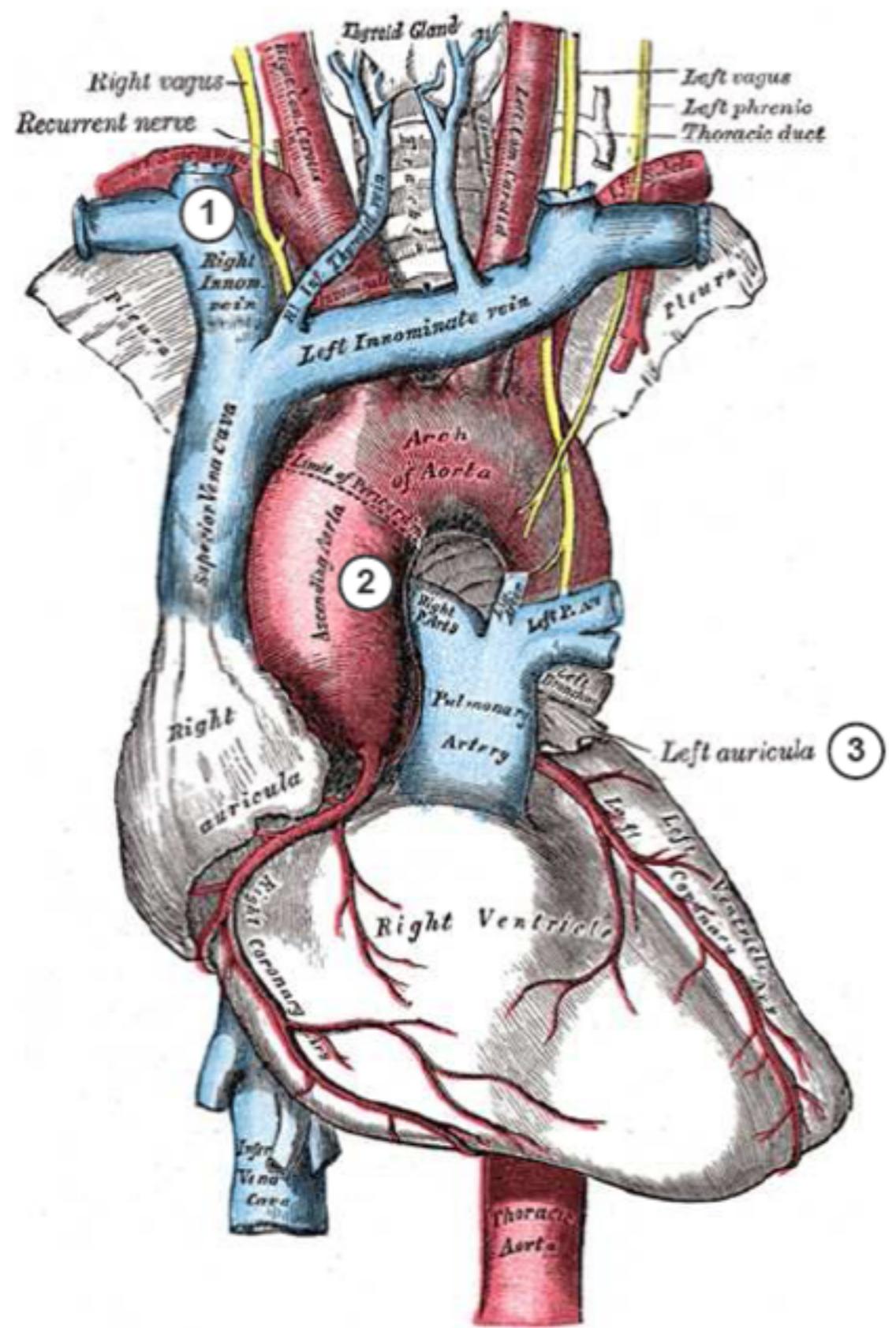
Confirmation

Hypothesis is given

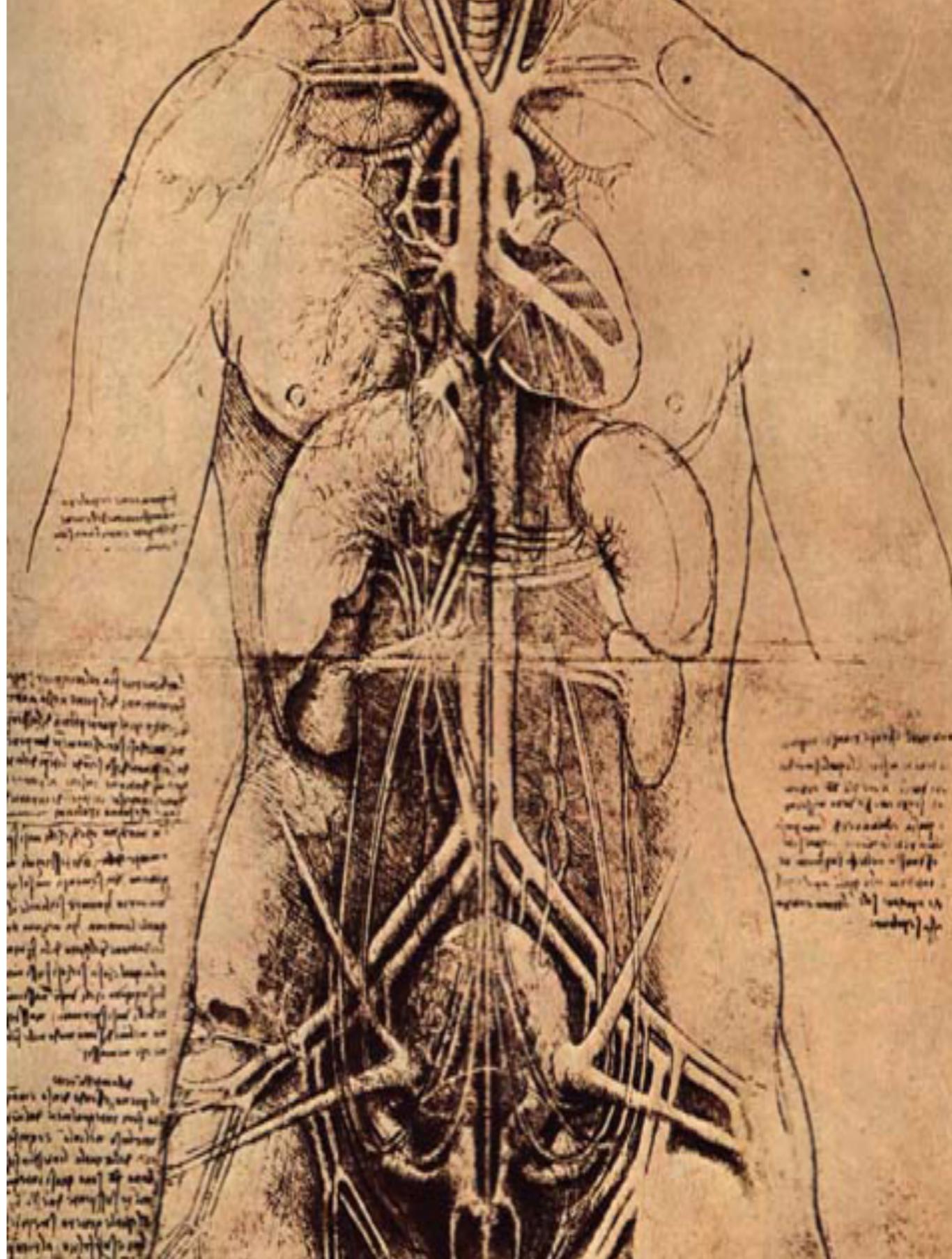
Task: Verify / falsify hypothesis



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

Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes perdus sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en termes des zones. Le rouge désigne les hommes qui ont survécu à Russie, le noir ceux qui en sont morts. — Les renseignements qui ont servi à dessiner la carte ont été puisés dans les ouvrages de M. M. Chiers, de Léger, de Fezensac; de Chambray et le journal intime de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Napoléon et du Maréchal Davout, qui avaient été détachés sur Minsk et Malibor au régime vers Orsha et Witebsk, avaient toujours marché avec l'armée.

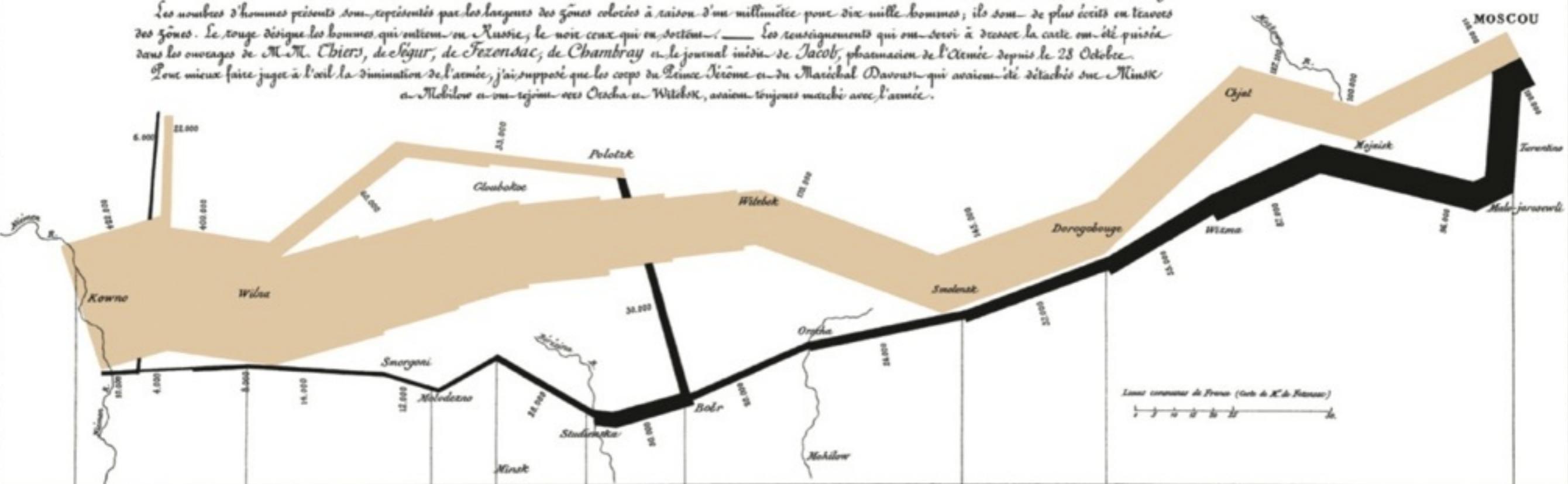


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au gelé
le Nilmen gelé.

- 26° le 7 X.^{me}

- 30° le 6 X.^{me}

- 24° le 1^{er} X.^{me}

- 20° le 28 D.^{me}

- 11°

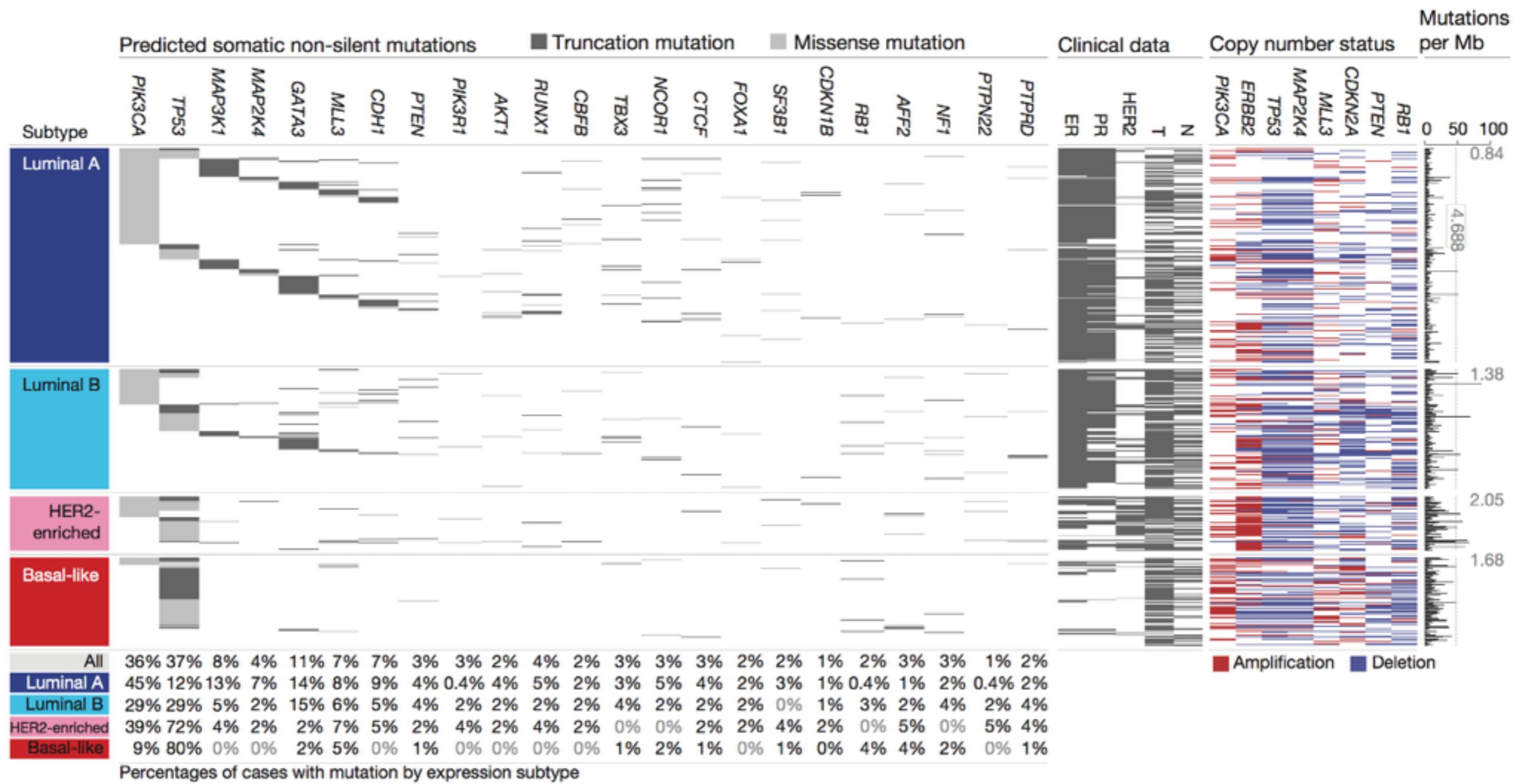
- 21° le 14 D.^{me}

- 9° le 9 D.^{me}

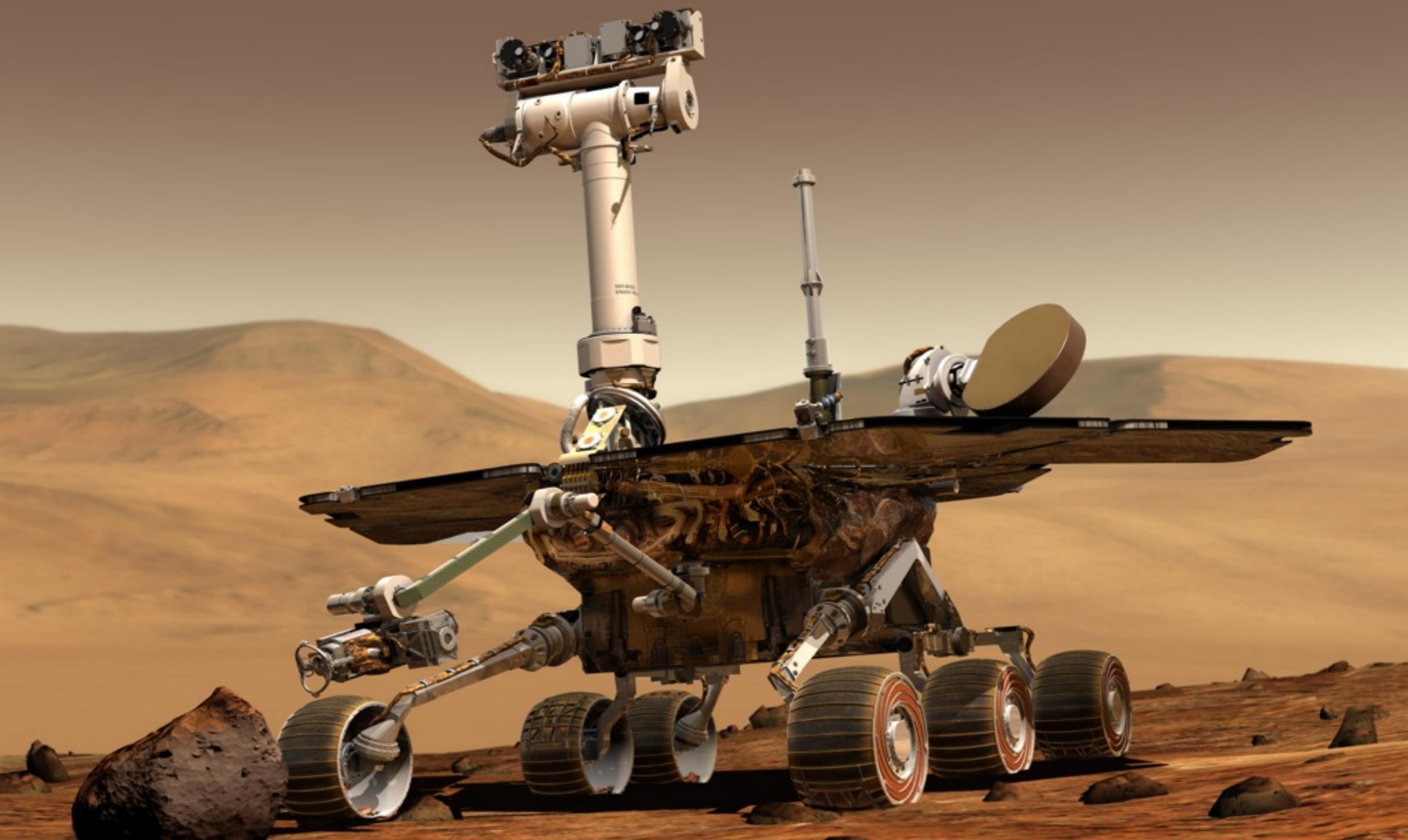
Pluie 24 D.^{me}

Zero le 18 D.^{me}

Imp. Lib. Regnier et Bourdet.



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)

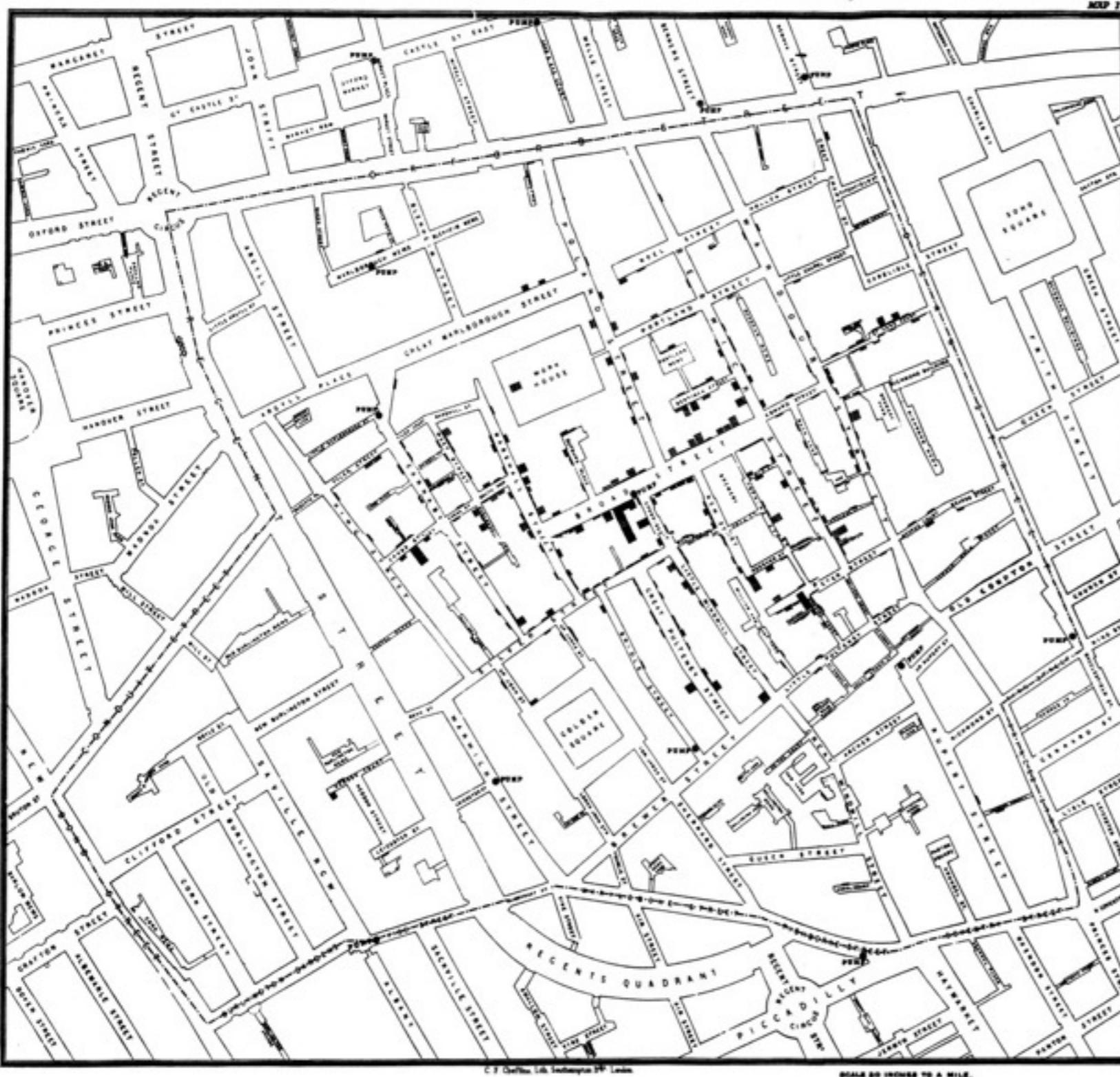


EDA Definition

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

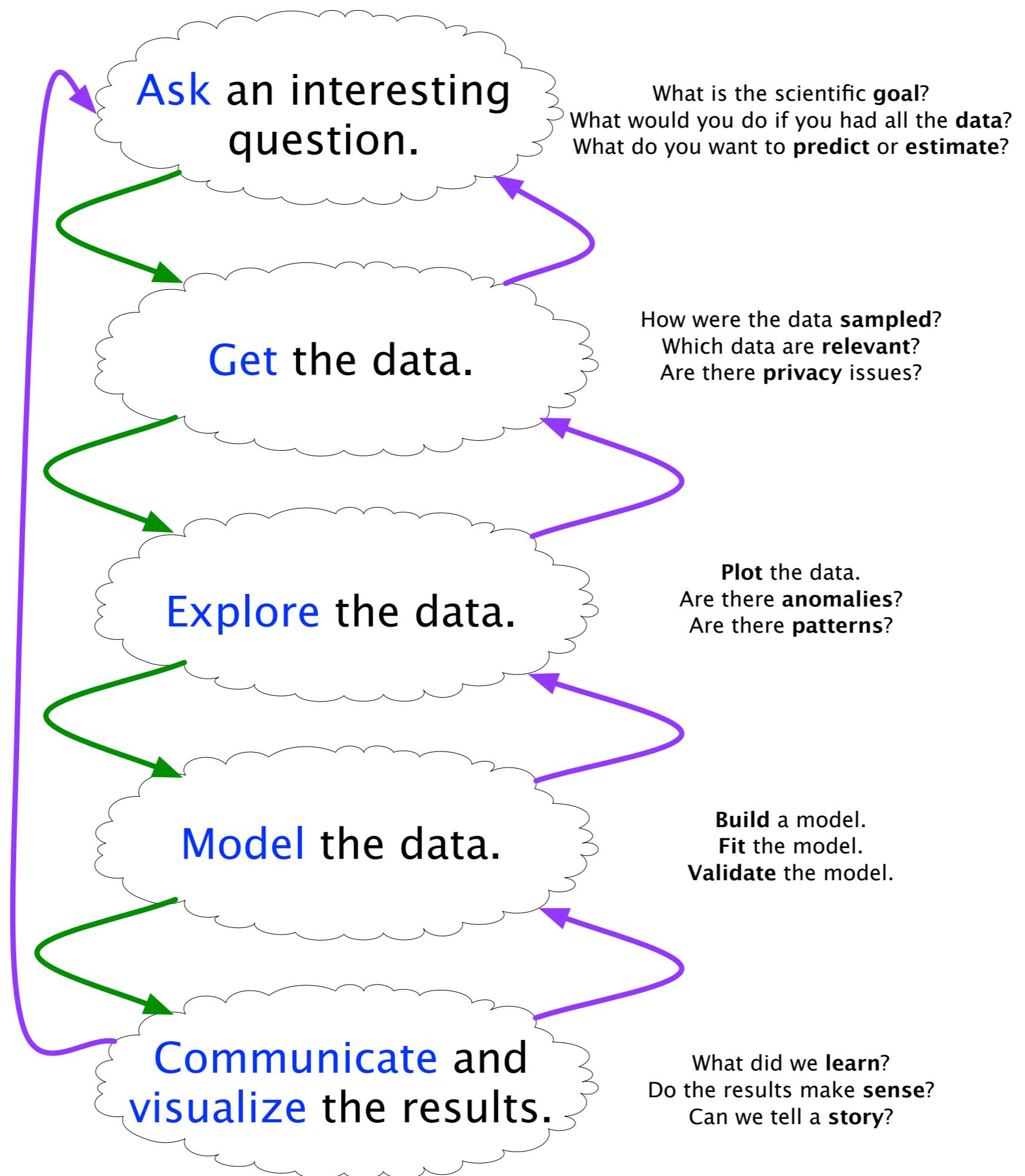


John Snow
(1813 – 1858)



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)



Example: Antibiotics
Will Burtin, 1951

Effectiveness of Antibiotics

Table 1: Burtin's data.

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

Data & Questions

What are the data types?

What are possible questions?

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

Data

Antibiotic name [string]

Gram staining? [pos/neg]

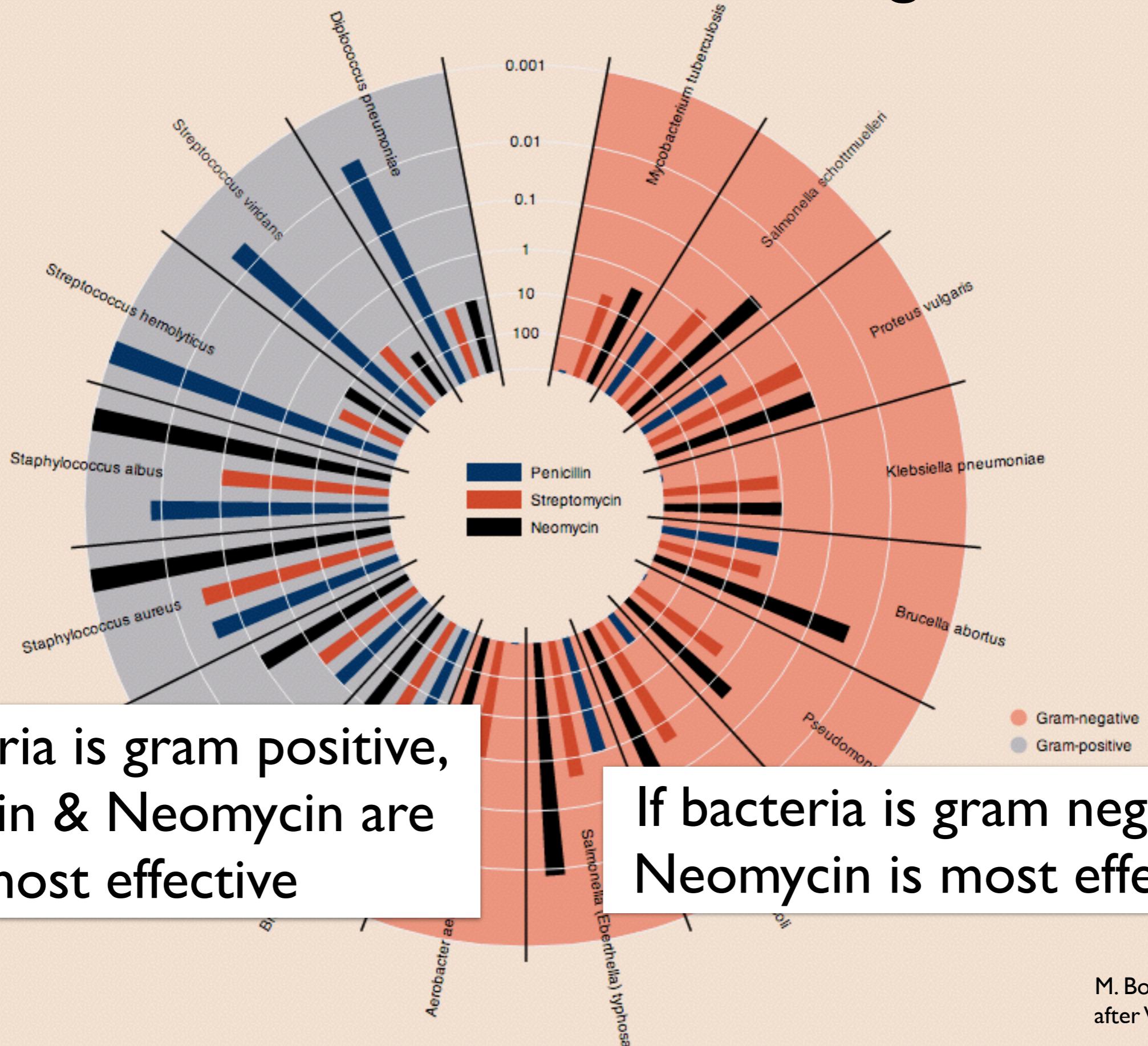
Minimum inhibitory concentration (mg/ml) [float]
(lower == more effective)

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

What Questions?

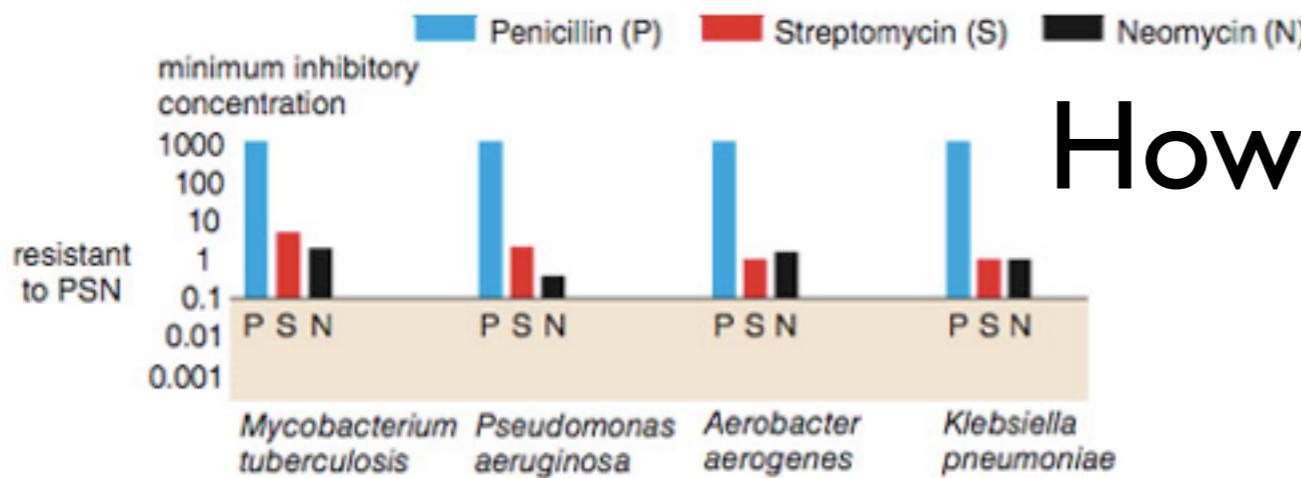
Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
<i>Aerobacter aerogenes</i>	870	1	1.6	-
<i>Brucella abortus</i>	1	2	0.02	-
<i>Bacillus anthracis</i>	0.001	0.01	0.007	+
<i>Diplococcus pneumoniae</i>	0.005	11	10	+
<i>Escherichia coli</i>	100	0.4	0.1	-
<i>Klebsiella pneumoniae</i>	850	1.2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

How effective are the drugs?

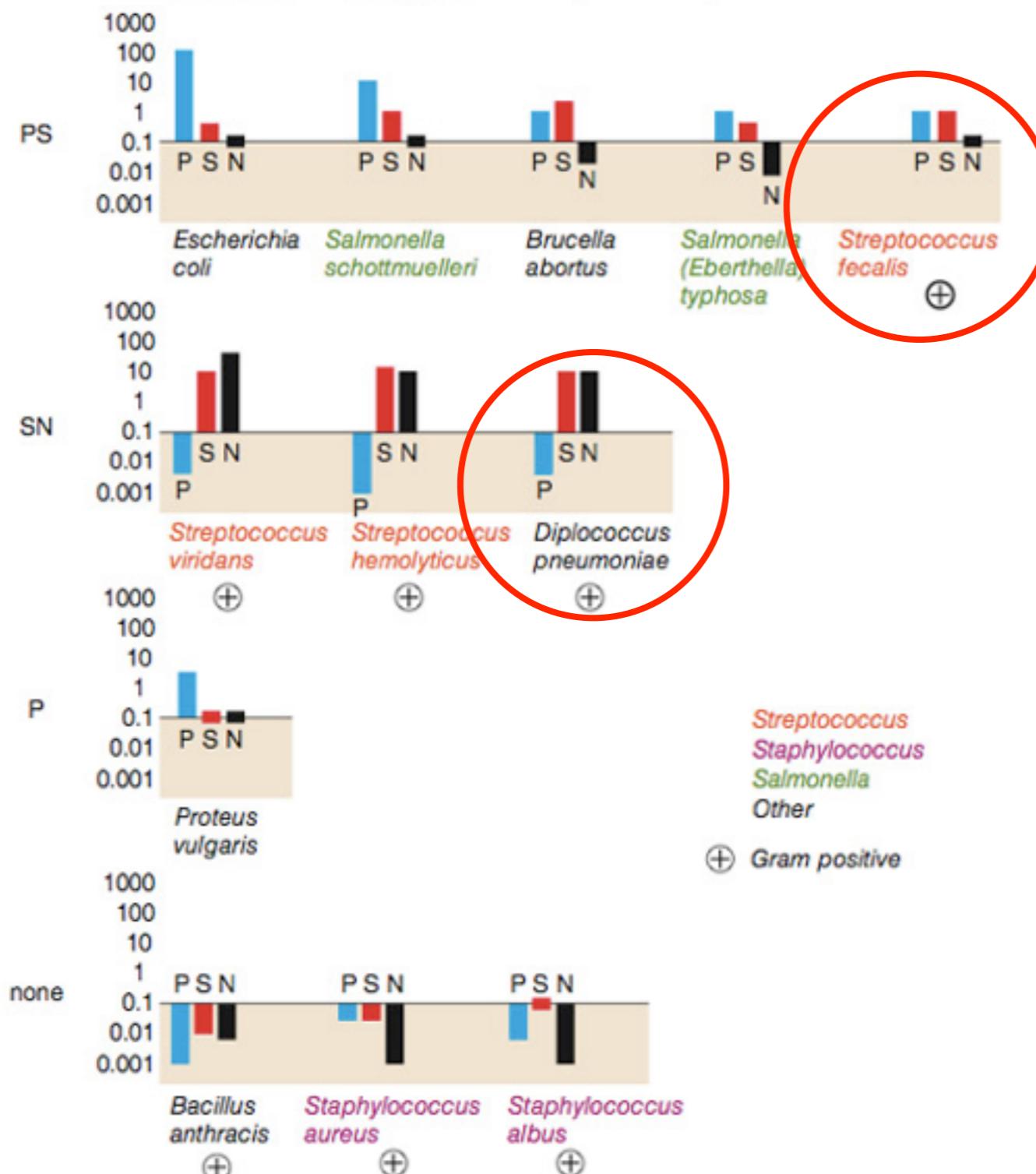


If bacteria is gram positive,
Penicillin & Neomycin are
most effective

If bacteria is gram negative,
Neomycin is most effective



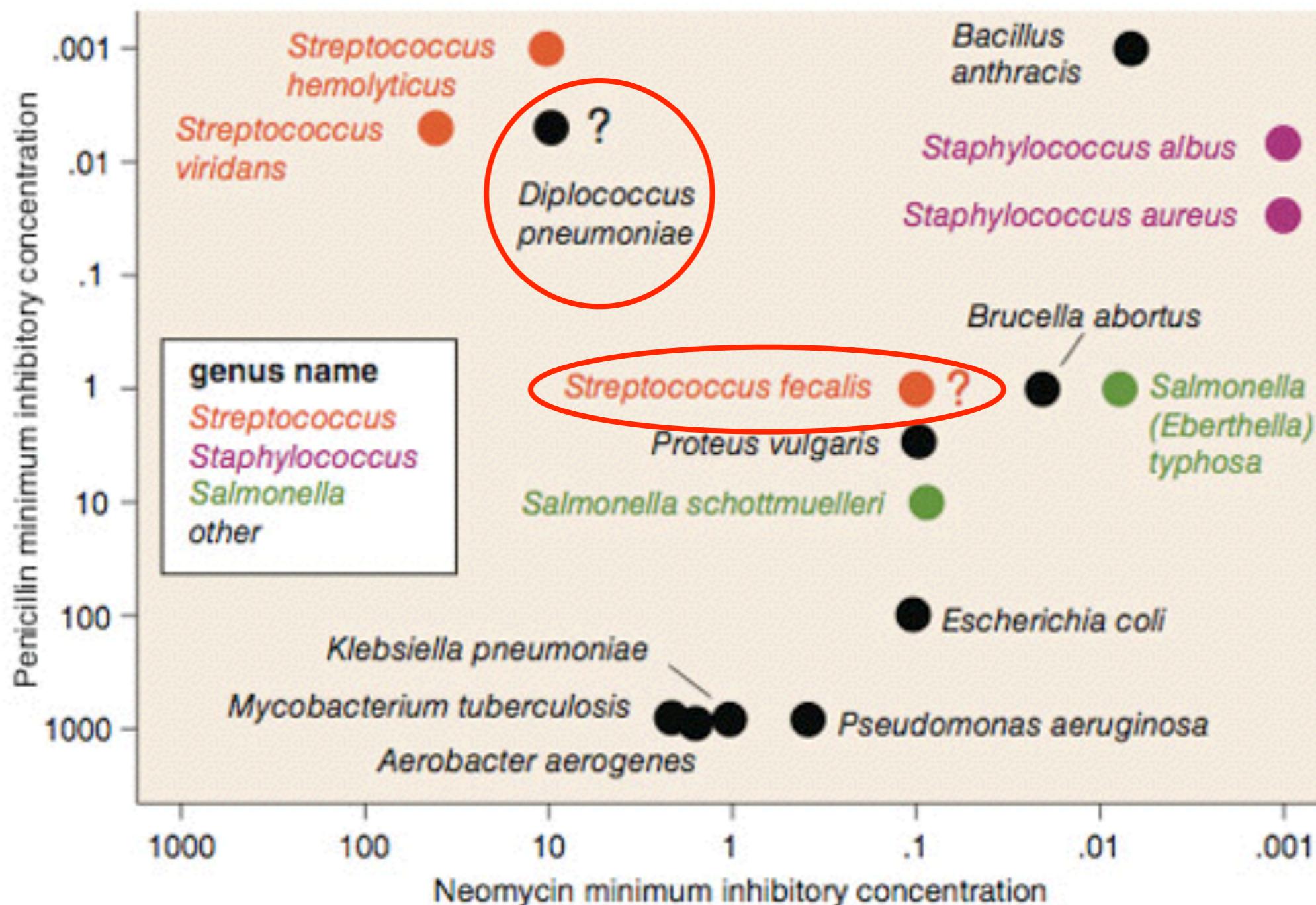
How do the bacteria compare?



Not a streptococcus!
(realized ~30 years later)

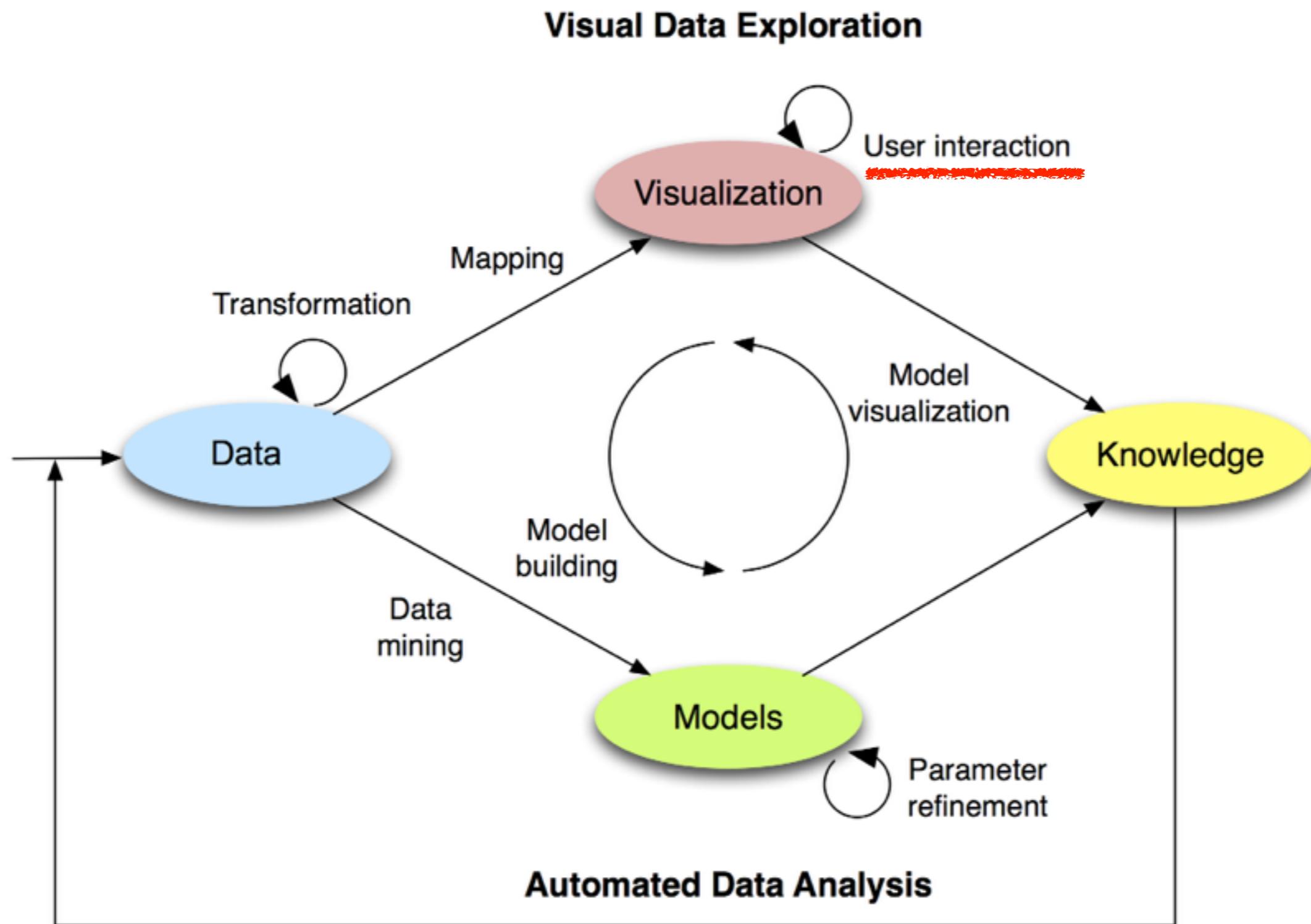
Really a streptococcus!
(realized ~20 years later)

How do the bacteria compare?



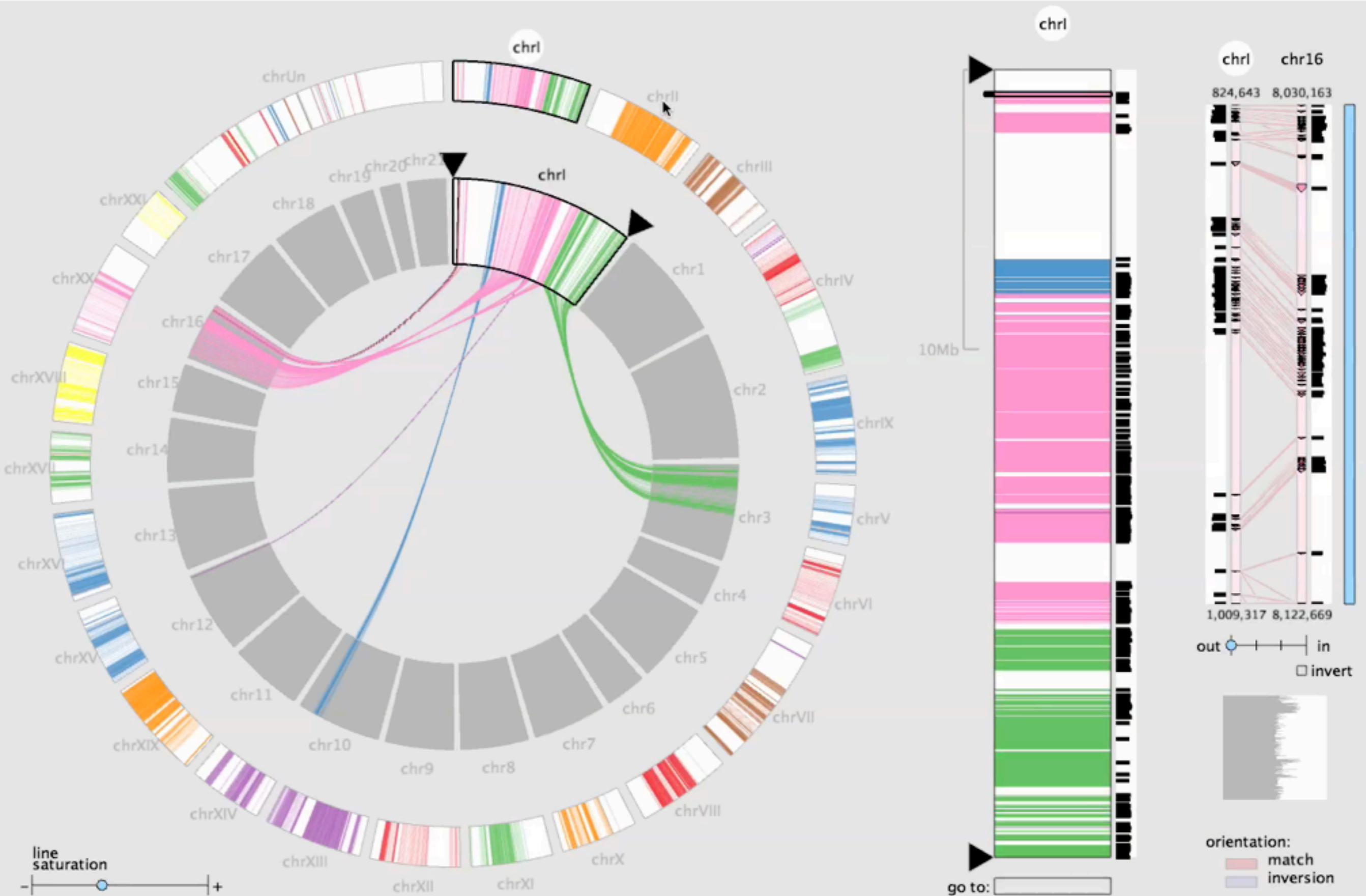
So far everything
was static...

Visual Analytics Process



Example: MizBee

[Meyer et al. 2009]



Data Types

Ben Shneiderman, 1996

1D (sequences)

Temporal

2D (maps)

3D (shaped)

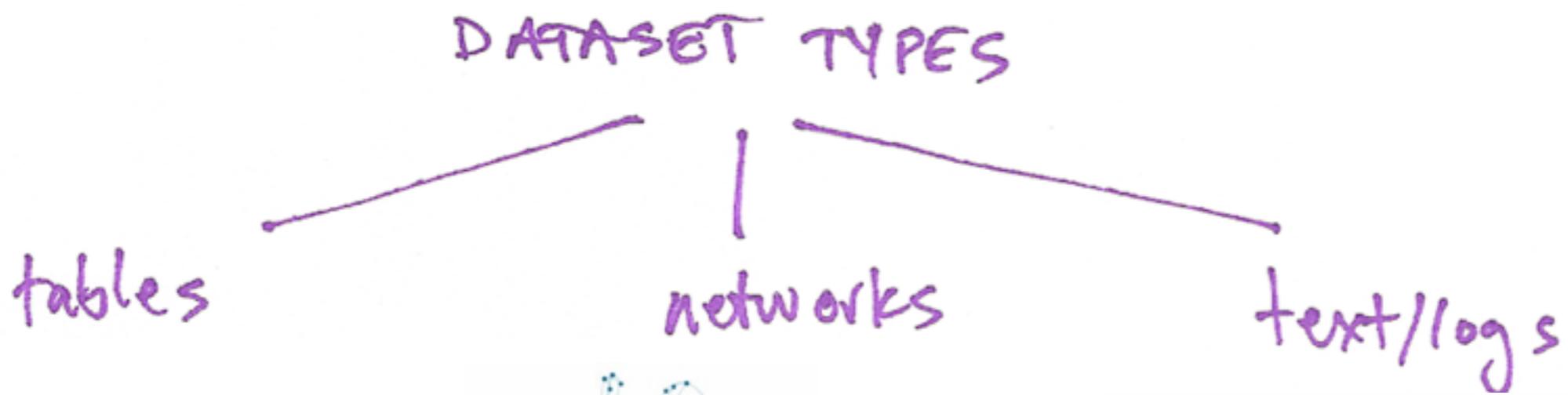
nD (relational)

Trees (hierarchical)

Networks (graphs)

Others?

Tamara Munzner, 2013

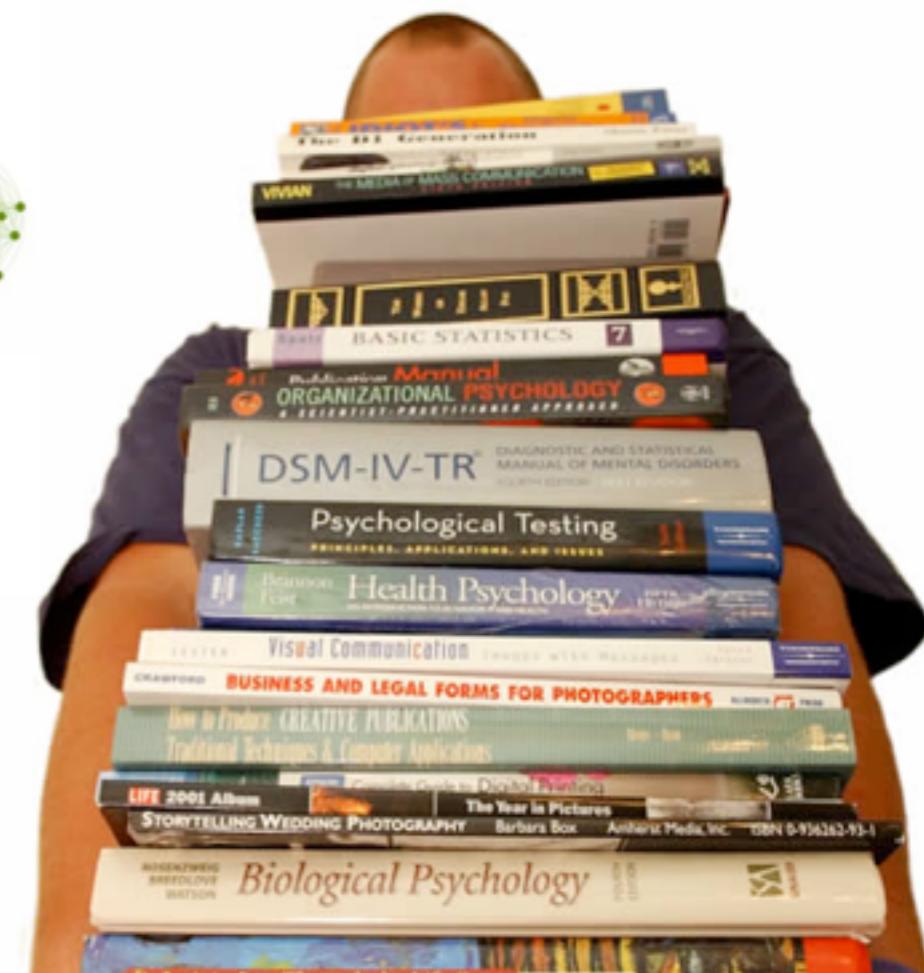
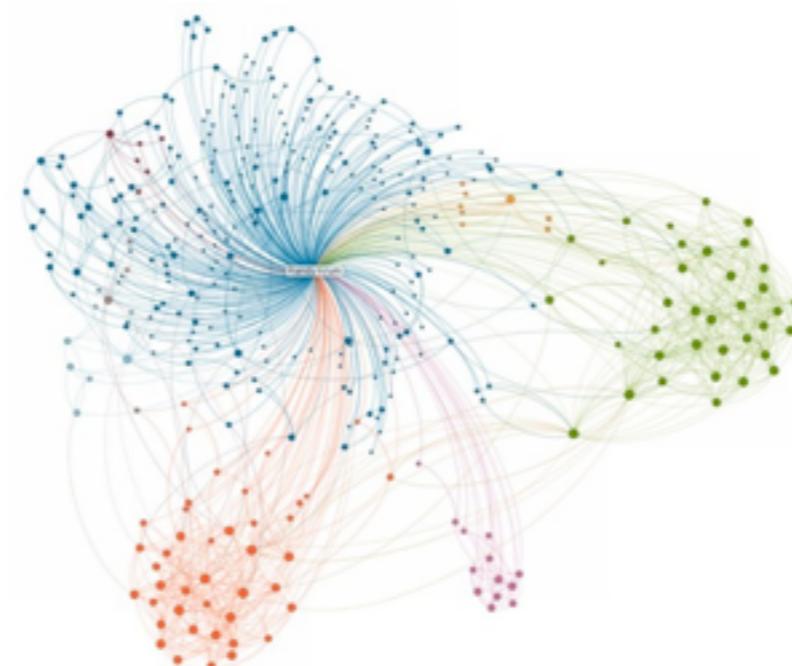


Google Docs @gmail.com | New Features | Doc Home

FriendFeed Audience

Share | Autosaved on 10:35 AM

	Site Name	Category	Compositors	Unique Use	Country	Re Page Views	Google
1	friendfeed.com	/Online Cor	160000	150000	0.1	20000000	
2	twhirl.org	/Computerz	47000	43000	0	74000	
3	twitpic.com	/Online Cor	43000	18000	0	120000	
4	chrisbrogan.com	/Online Cor	39000	29000	0	74000	
5	brightkite.com	/Telecomm	29000	68000	0	910000	
6	twitpic.com	/Home & G	24000	71000	0	340000	TRUE
7	web-strategist.c	/Online Cor	24000	32000	0	86000	
8	summize.com	/Arts & H	20000	54000	0	570000	



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	<i>Permutation group</i> $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	<i>Isotonic group</i> $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	<i>General linear group</i> $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	<i>Similarity group</i> $x' = ax$	Coefficient of variation

Data Types

Nominal (Categorical) (N)

Are = or \neq to other values

Apples, oranges, bananas, ...

Ordinal (O)

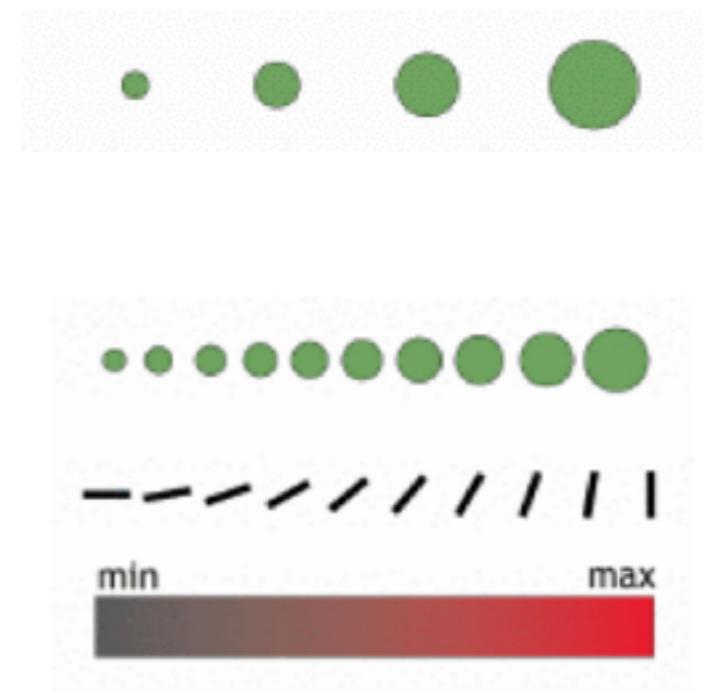
Obey a $<$ relationship

Small, medium, large

Quantitative (Q)

Can do arithmetic on them

10 inches, 23 inches, etc.



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

Data Types

N - Nominal (labels)

Operations: $=, \neq$

O - Ordinal (ordered)

Operations: $=, \neq, >, <$

Q - Interval (location of zero arbitrary)

Operations: $=, \neq, >, <, +, -$ (distance)

Q - Ratio (zero fixed)

Operations: $=, \neq, >, <, +, -, \times, \div$ (proportions)

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

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	Second	woman	False	C	Southampton	yes	False
0	3	male	35.0	0	0	8.05	S	Second	man	True		Southampton	no	True
0	3	male		0	0	8.4583	Q	Second	man	True		Queenstown	no	True
0	1	male	54.0	0	0	51.8625	S	Second	man	True	E	Southampton	no	True
0	3	male	2.0	3	1	21.075	S	Second	man	True		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					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

I = Quantitative
2 = Nominal
3 = Ordinal



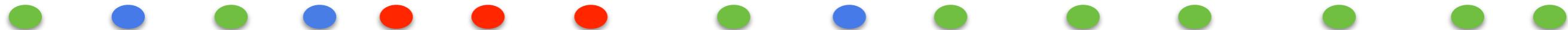
survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	3	male	22.0	1	0	7.25	S	Third	man	True		Southampton	no	False
1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
1	3	female	26.0	0	0	7.925	S	Third	woman	False		Southampton	yes	True
1	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									Southampton	yes	False
1	1	female	58.0									Southampton	yes	True
0	3	male	20.0									Southampton	no	True
0	3	male	39.0									Southampton	no	False
0	3	female	14.0									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

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

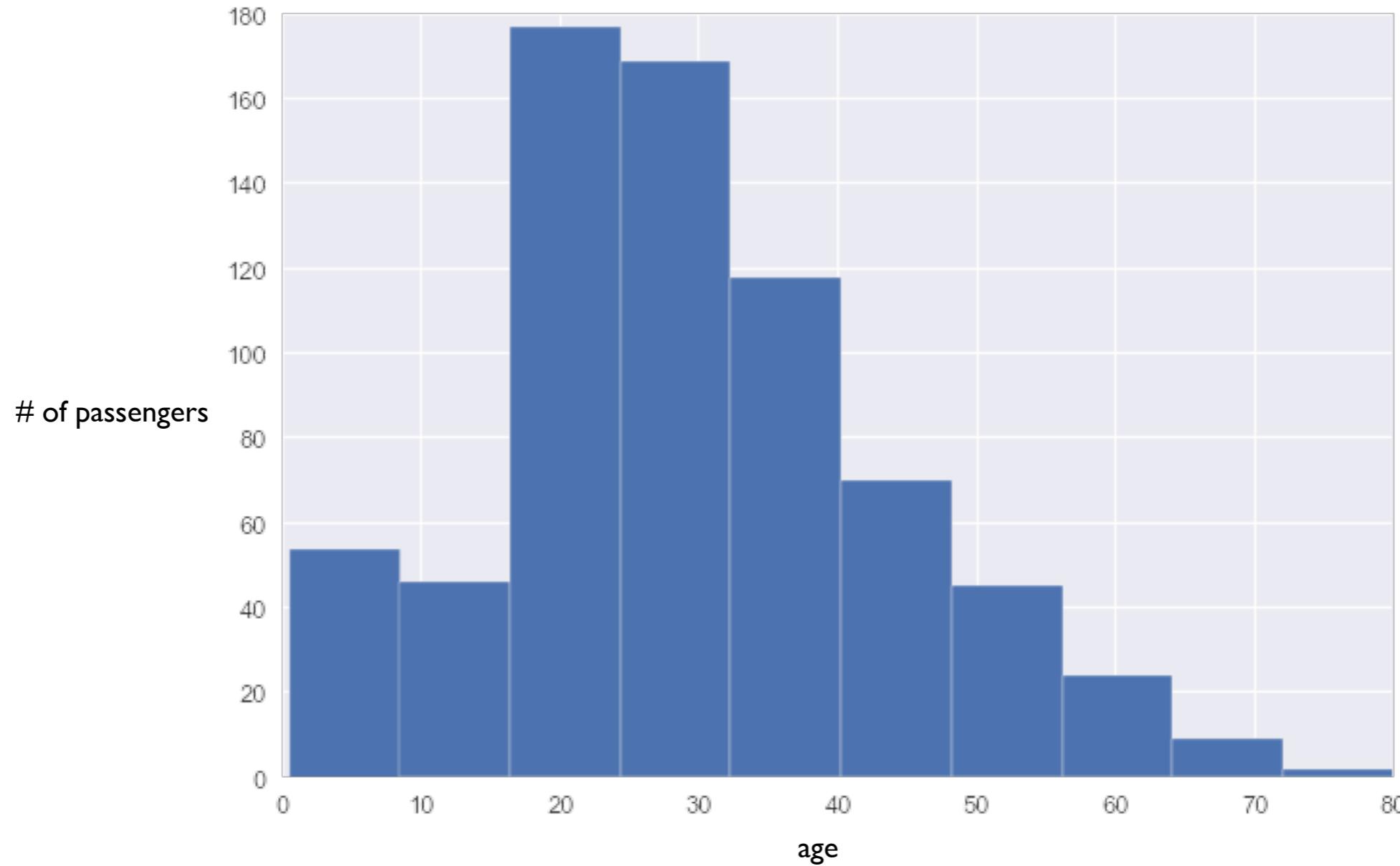
Continuous to 4 significant figures (Q)

Hot, warm, cold (O)

Burned vs. Not burned (N)

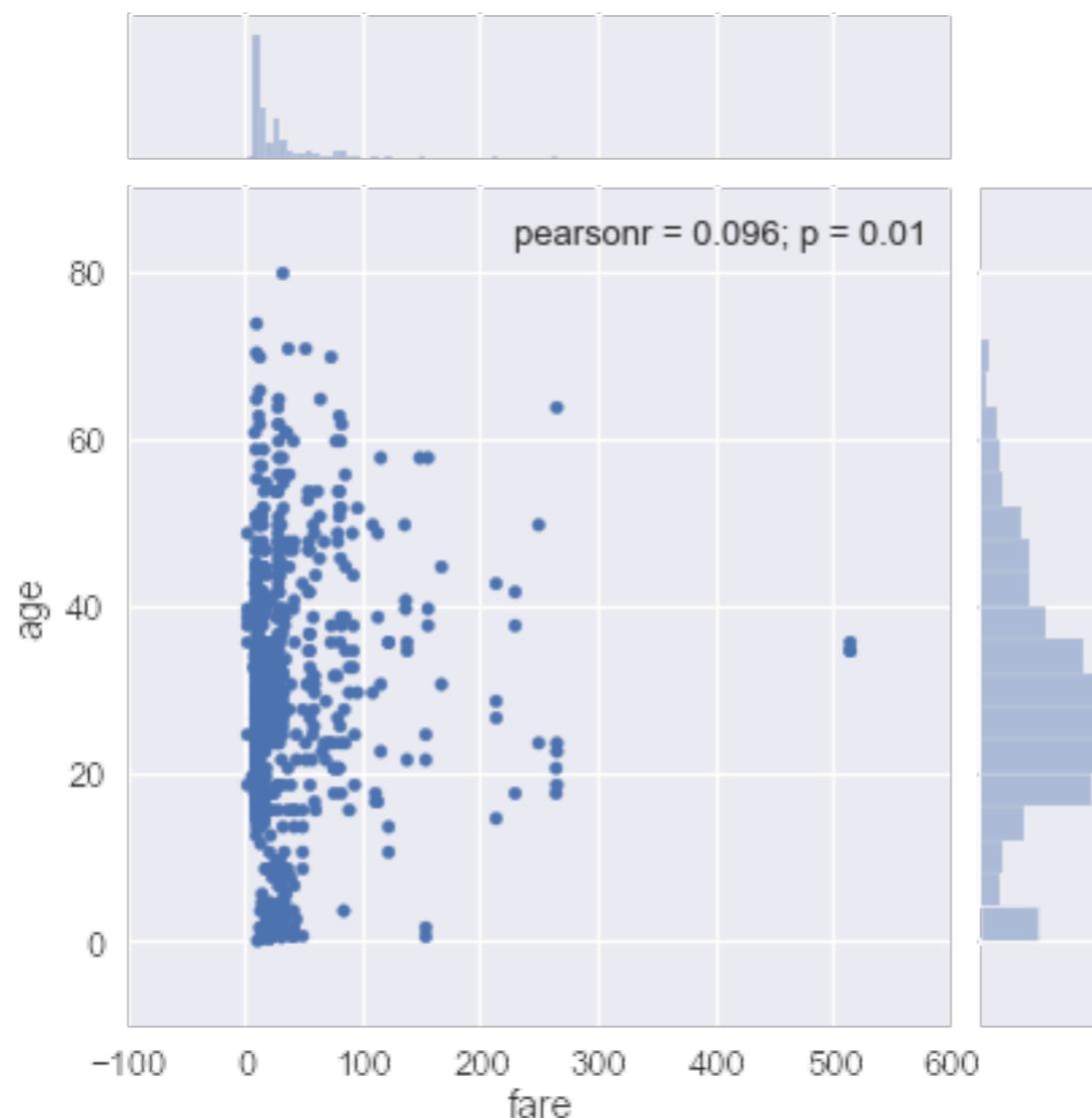
Data Dimensions

Univariate Data



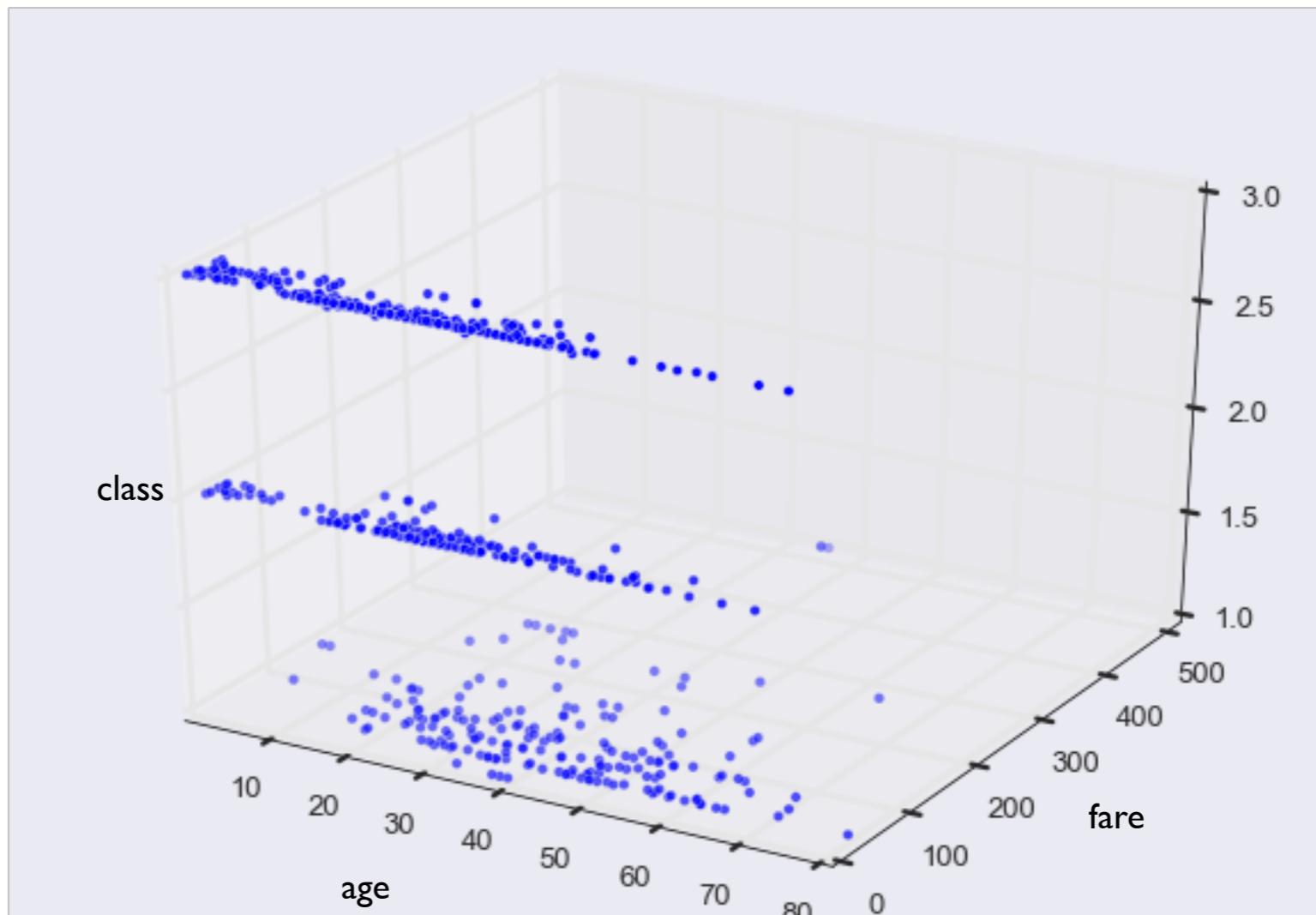
age	fare
22.0	7.25
38.0	71.2833
26.0	7.925
35.0	53.1
35.0	8.05
	8.4583
54.0	51.8625
2.0	21.075
27.0	11.1333
14.0	30.0708
4.0	16.7
58.0	26.55
20.0	8.05
39.0	31.275
14.0	7.8542
55.0	16.0
2.0	29.125
	13.0
31.0	18.0
	7.225
35.0	26.0
34.0	13.0
15.0	8.0292

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

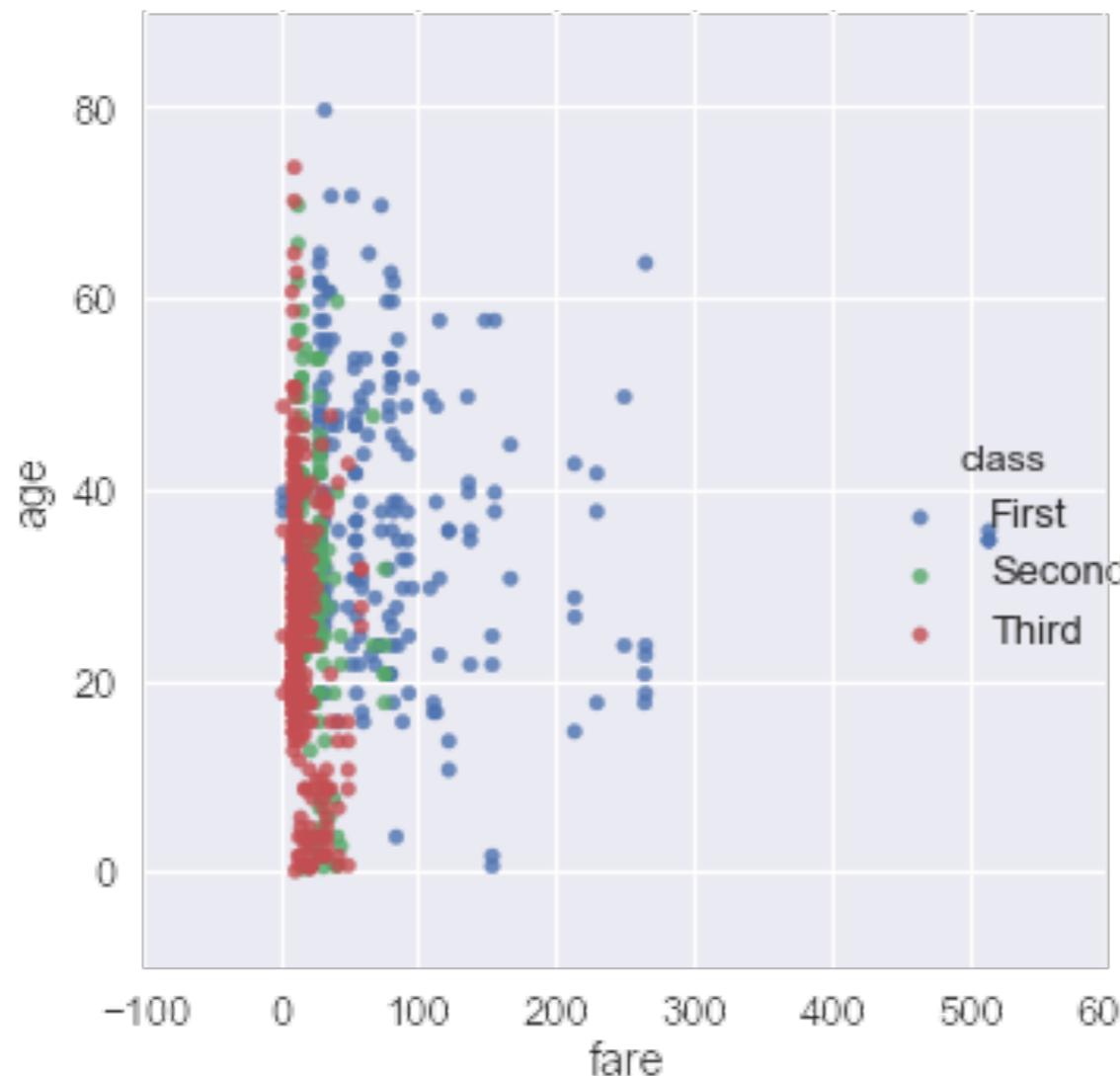
Trivariate Data



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

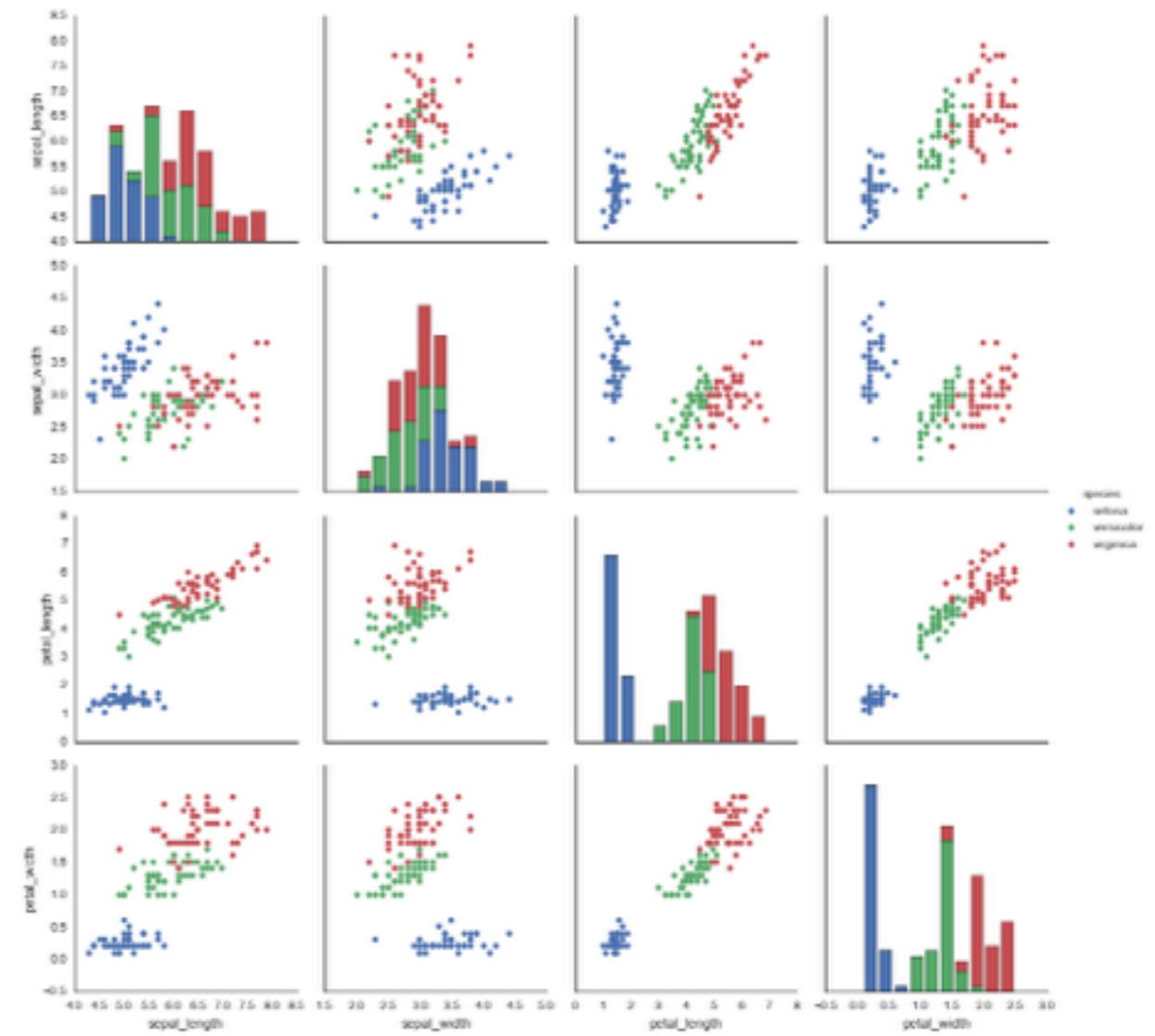
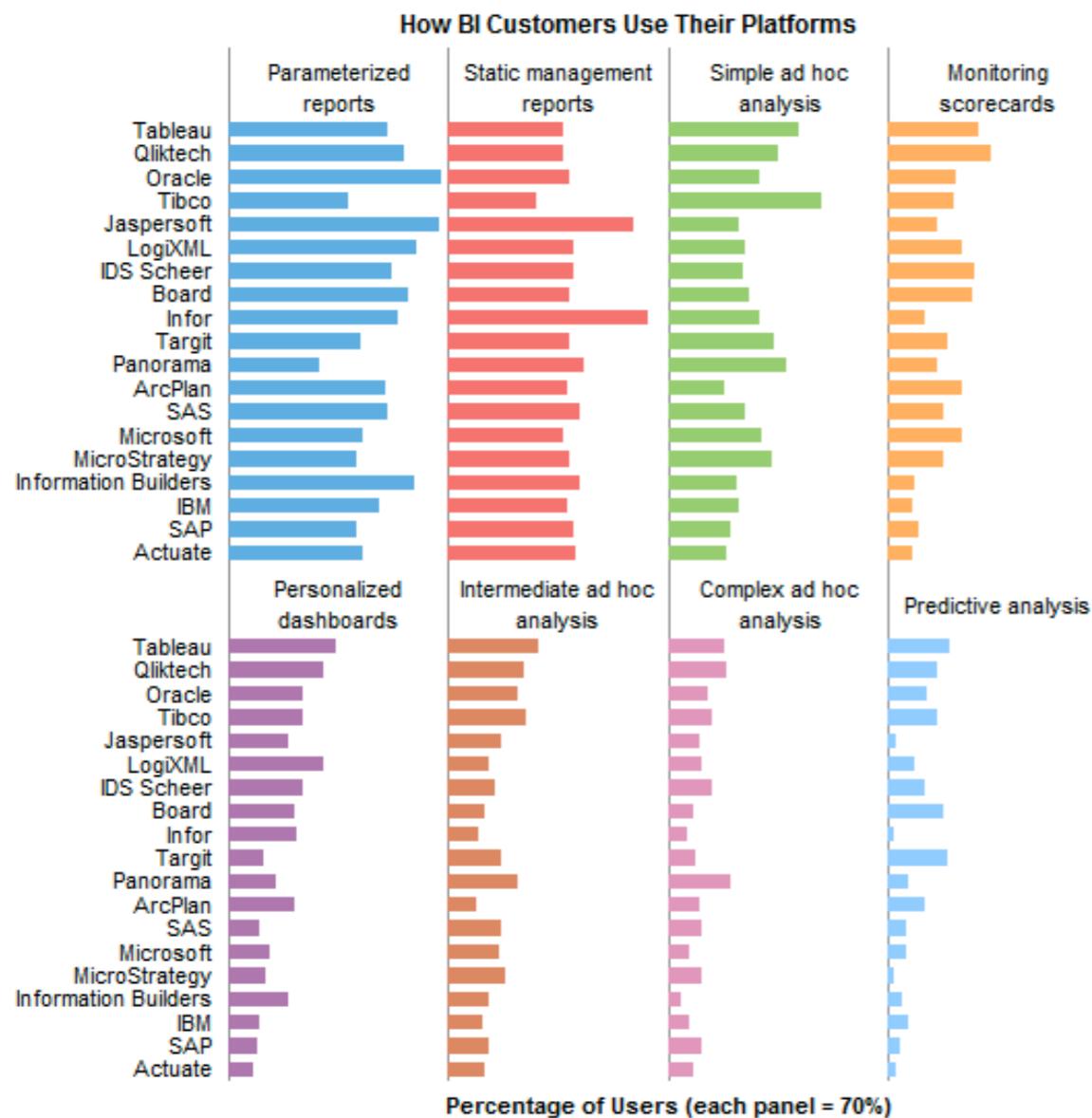
Trivariate Data



Map the third dimension to some other visual attribute

Multivariate Data

Give each attribute its own display (small multiples)

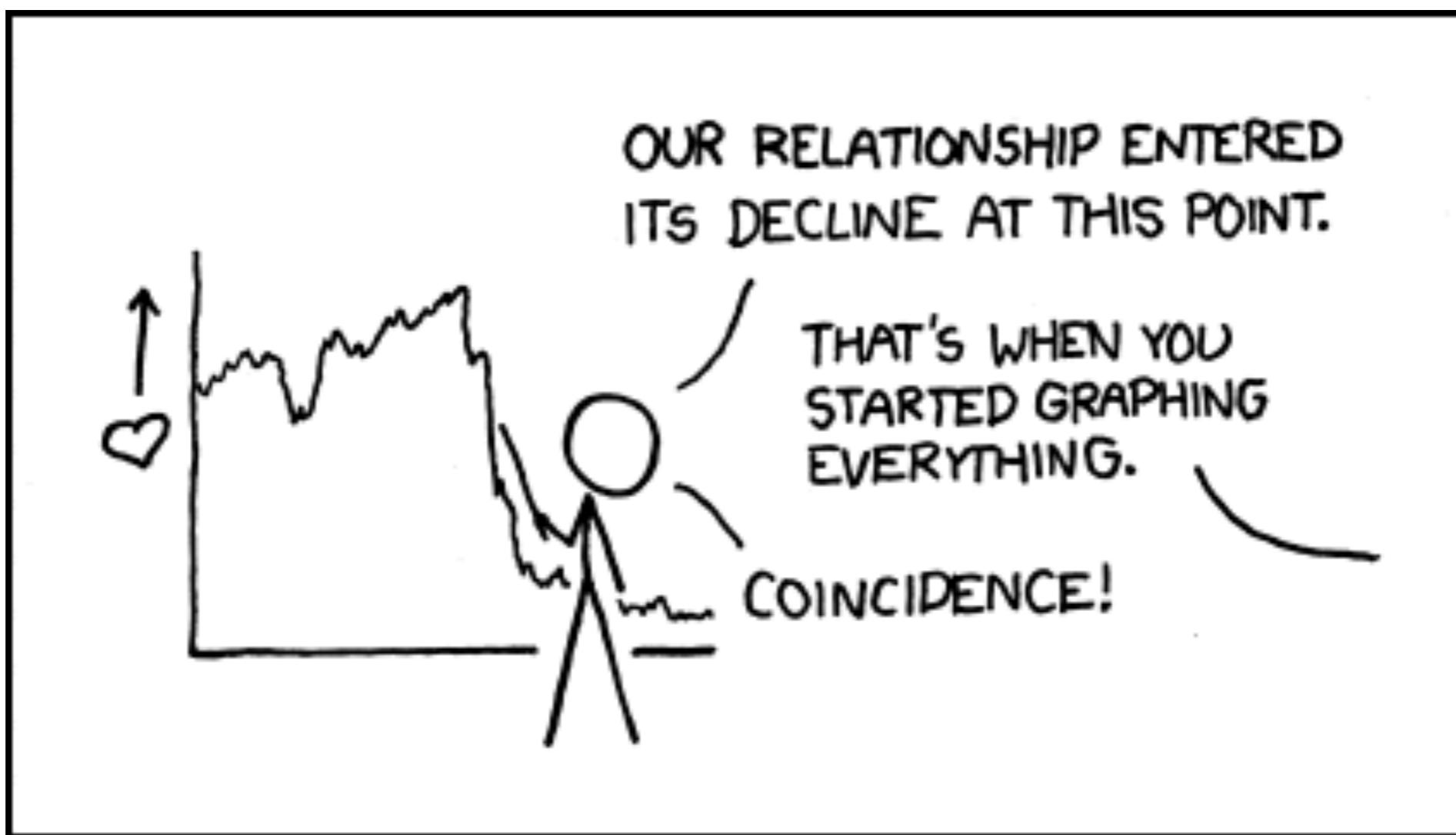


More next Tuesday!

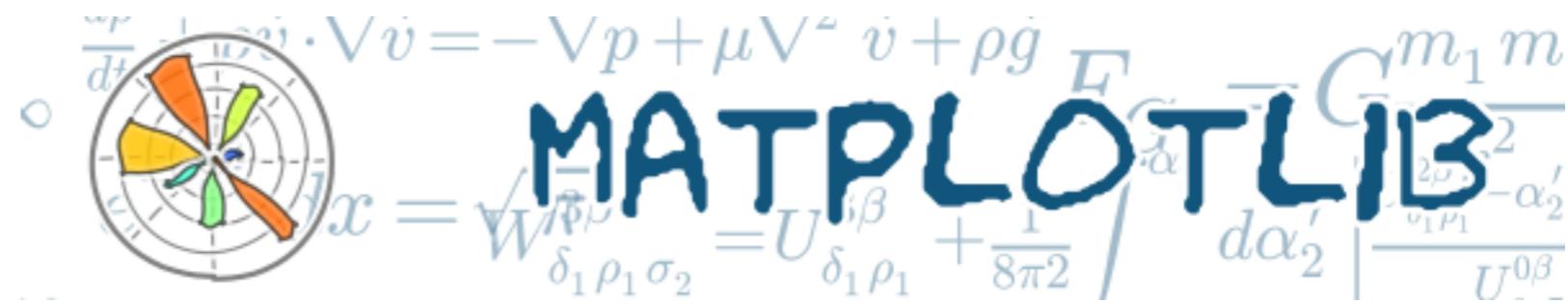
Data Reduction

- 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

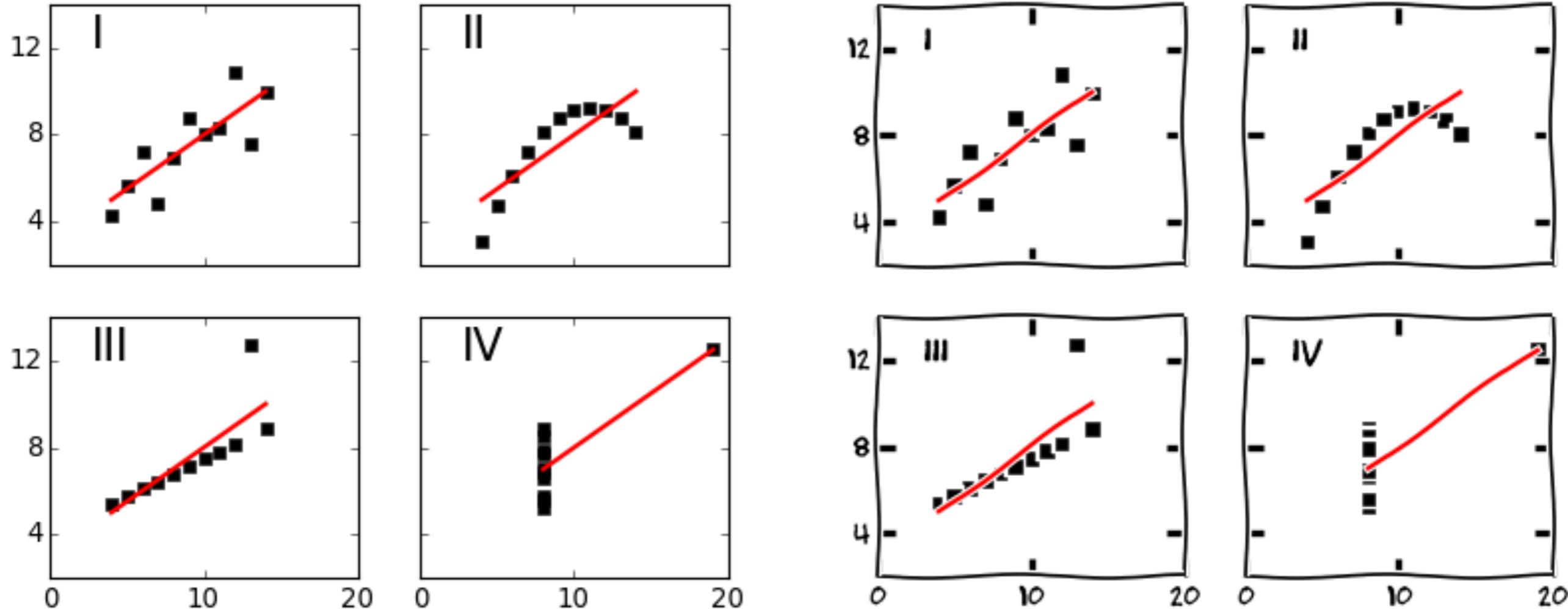
Statistical Graph Types



Side Note

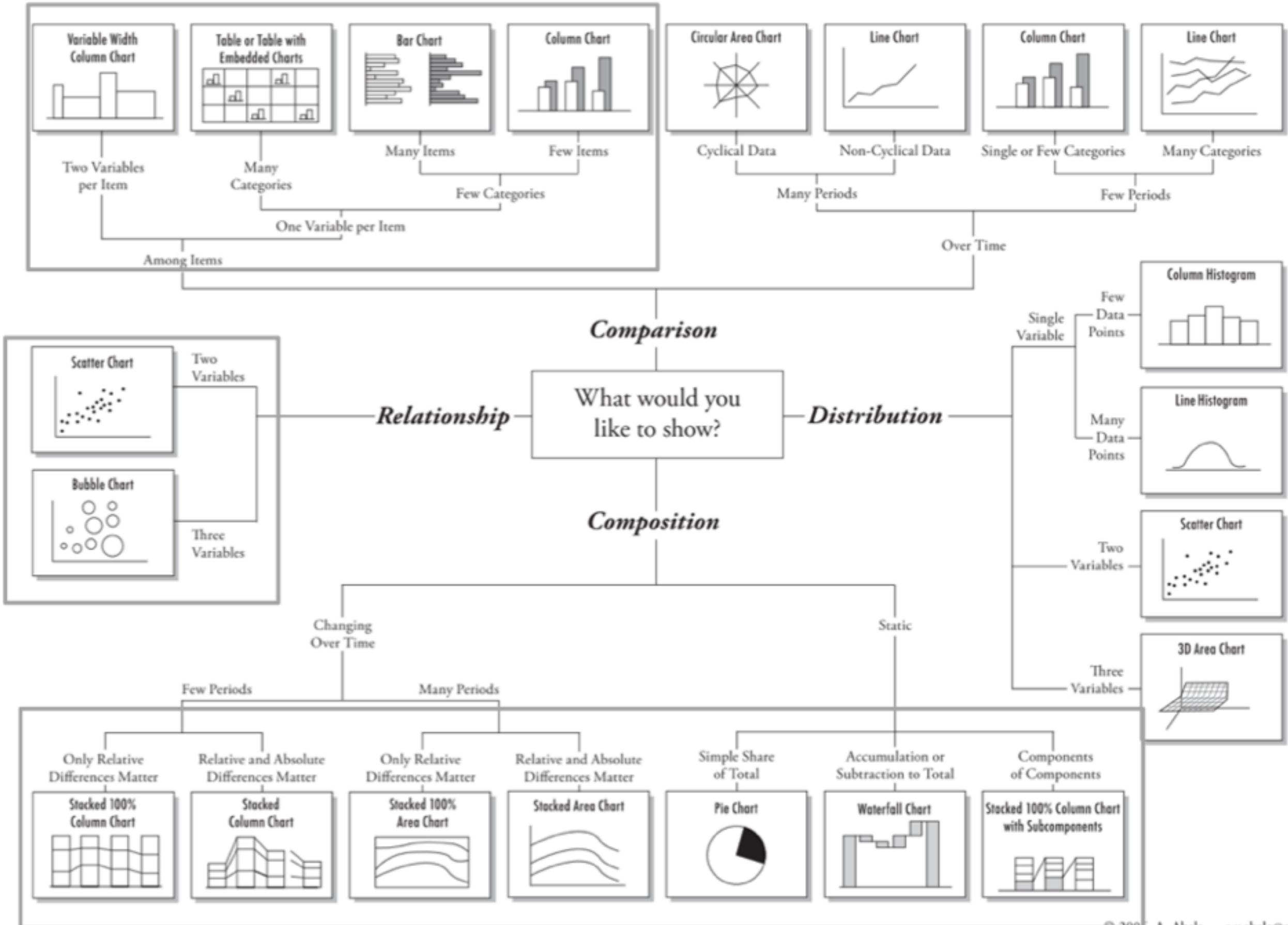


XKCD-ify your plot



<http://matplotlib.org/xkcd>

Chart Suggestions—A Thought-Starter

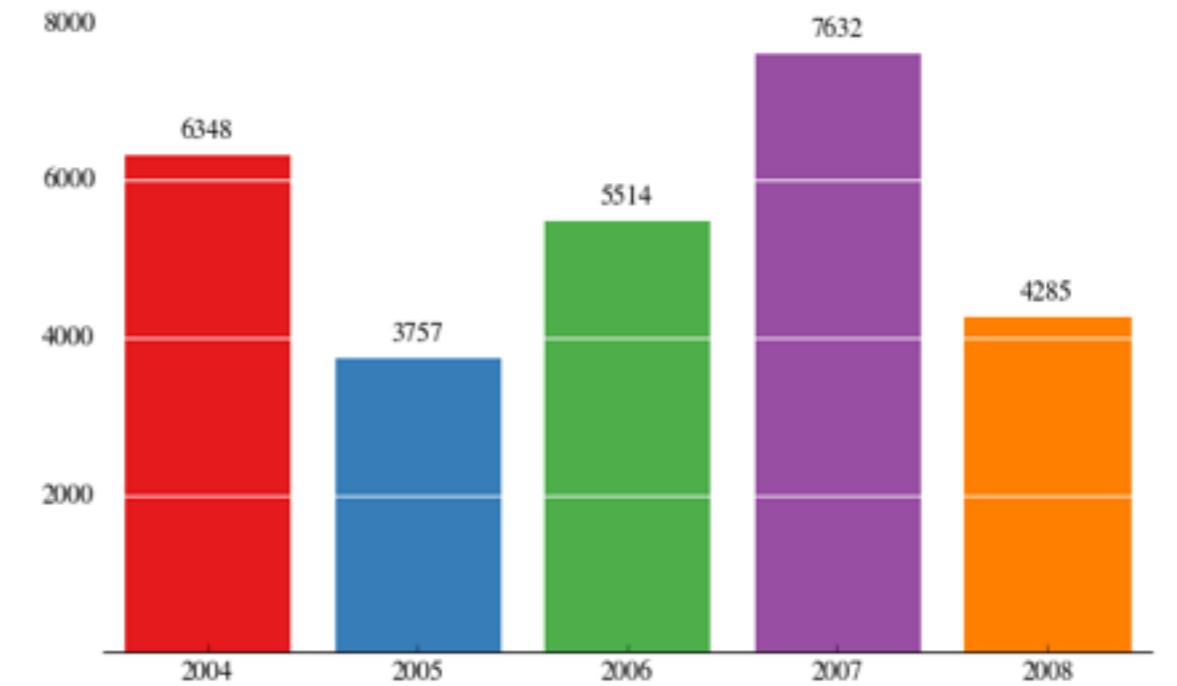
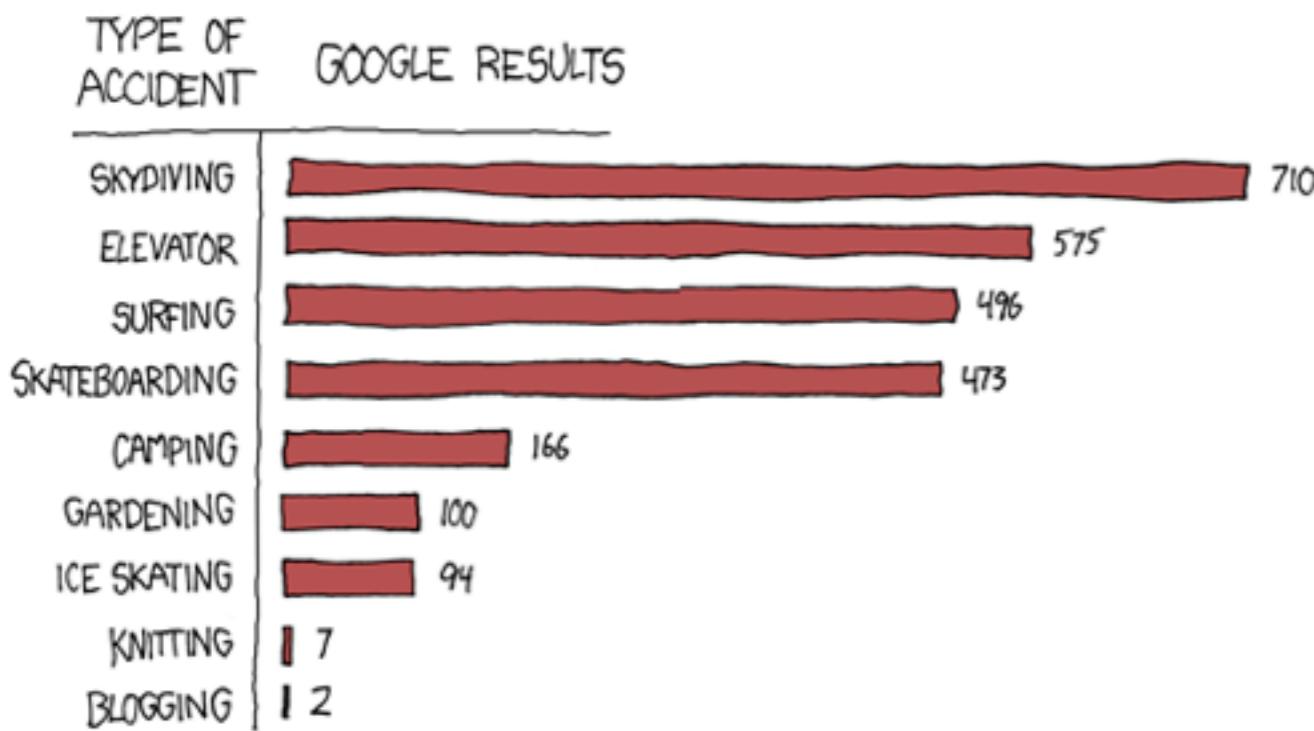


Comparisons

Bar Chart

DANGERS

INDEXED BY THE NUMBER OF GOOGLE RESULTS FOR
"DIED IN A _____ ACCIDENT"

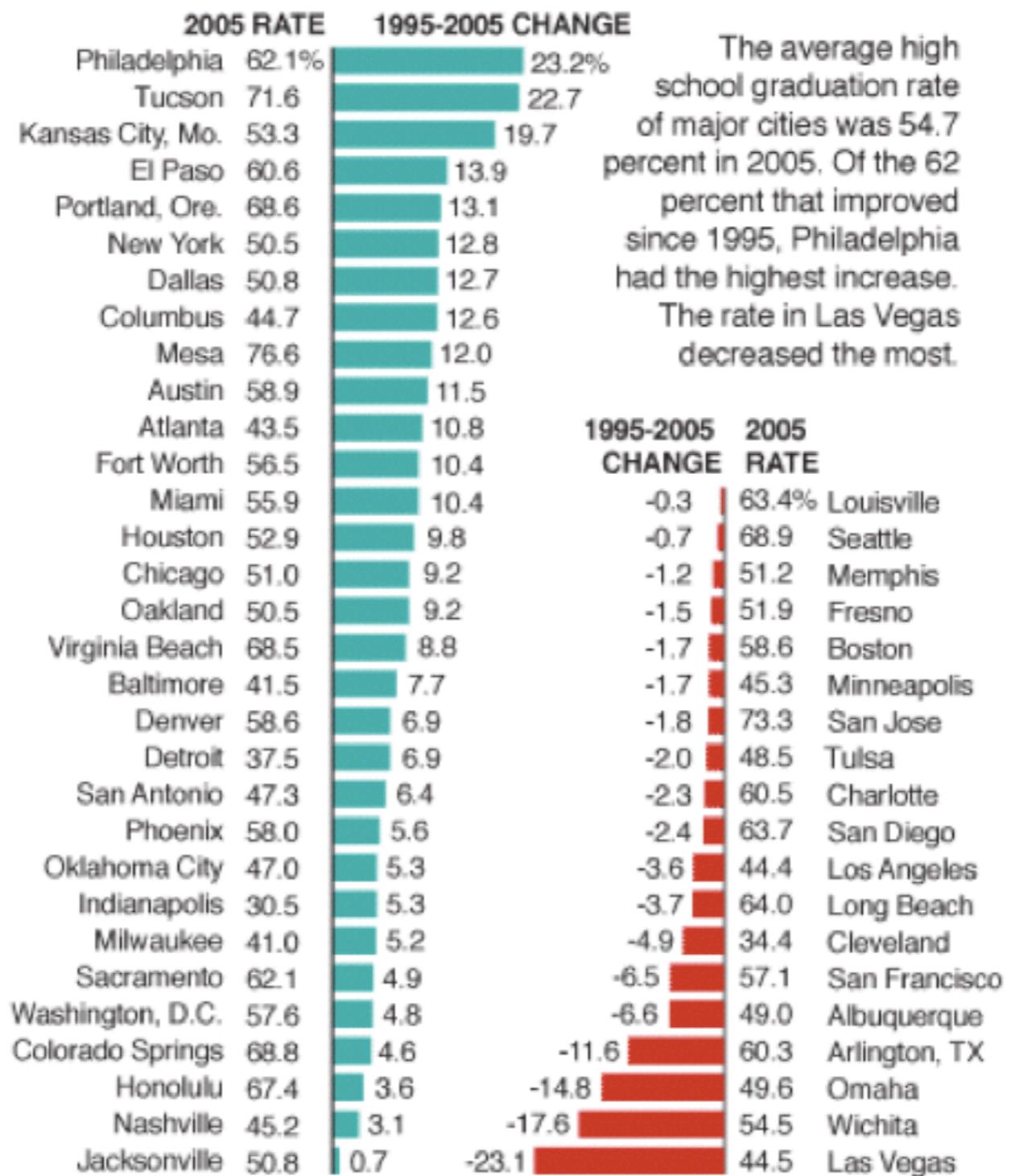


<http://nbviewer.ipython.org/gist/olgabot/5357268>

Direction

Graduation rates up in most cities

Graduation rate for principal school district of the largest cities



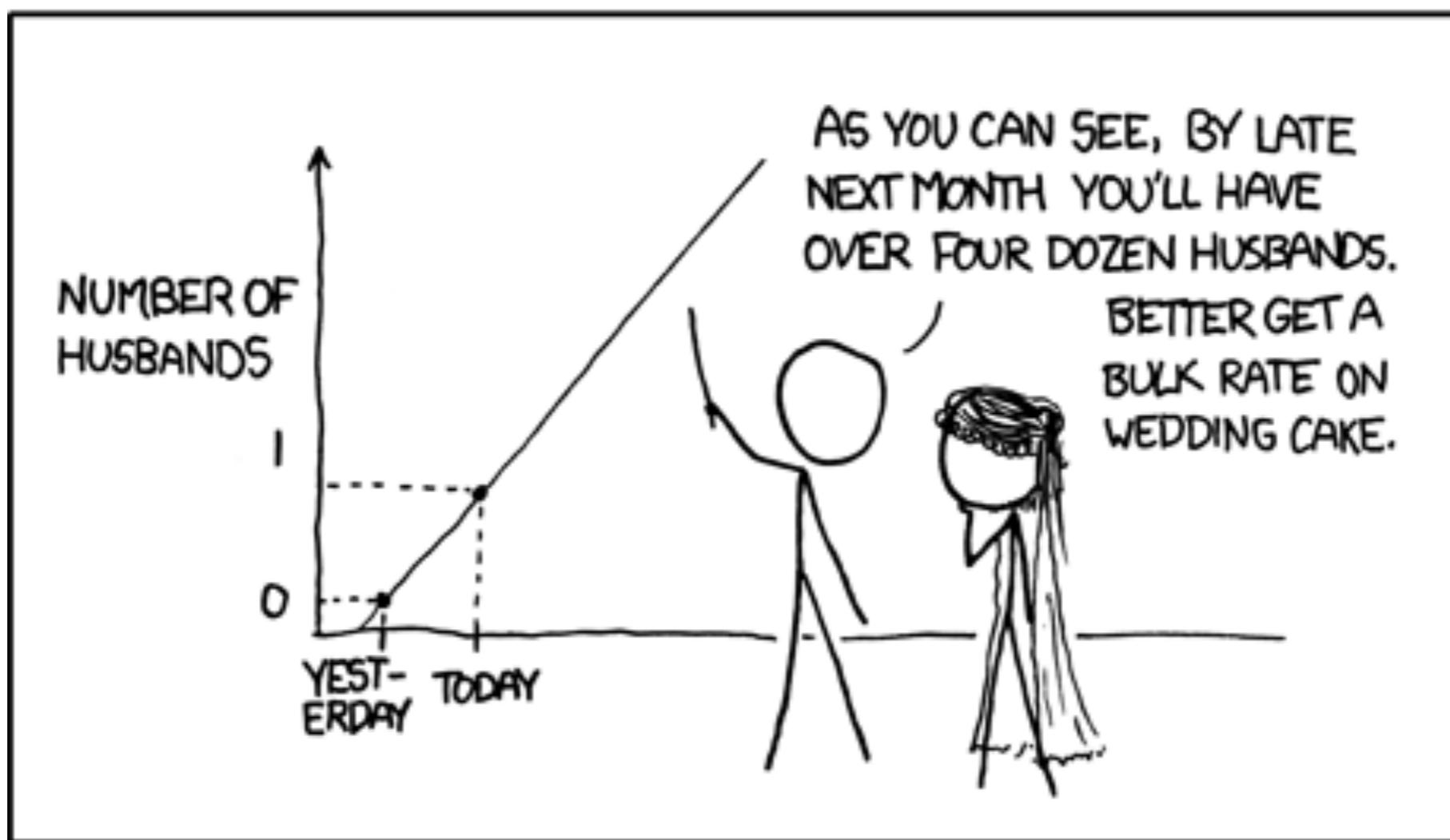
SOURCE: EPE Research Center

AP

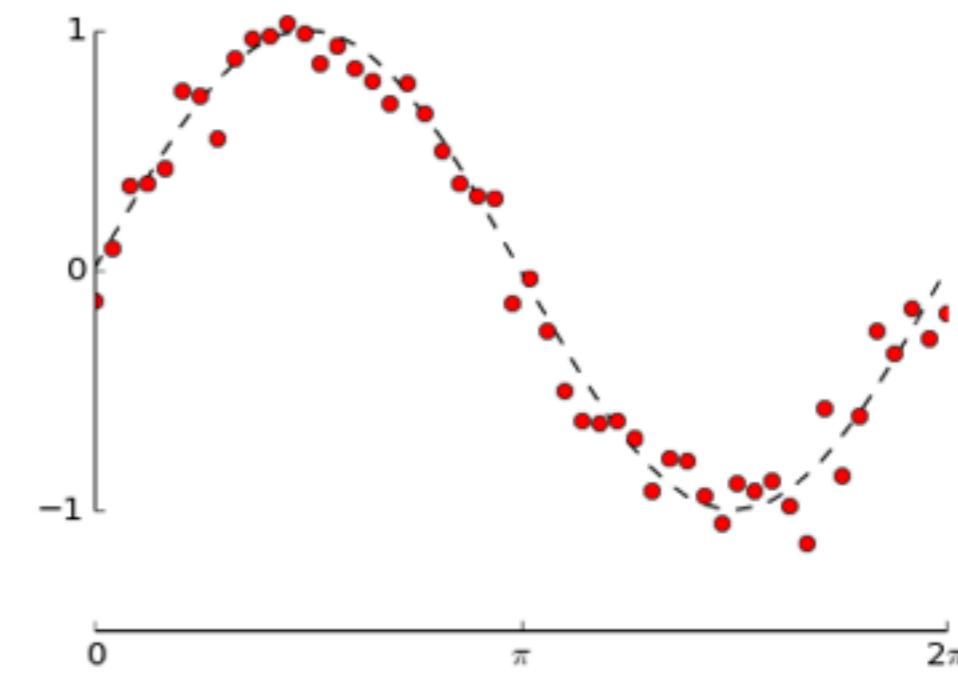
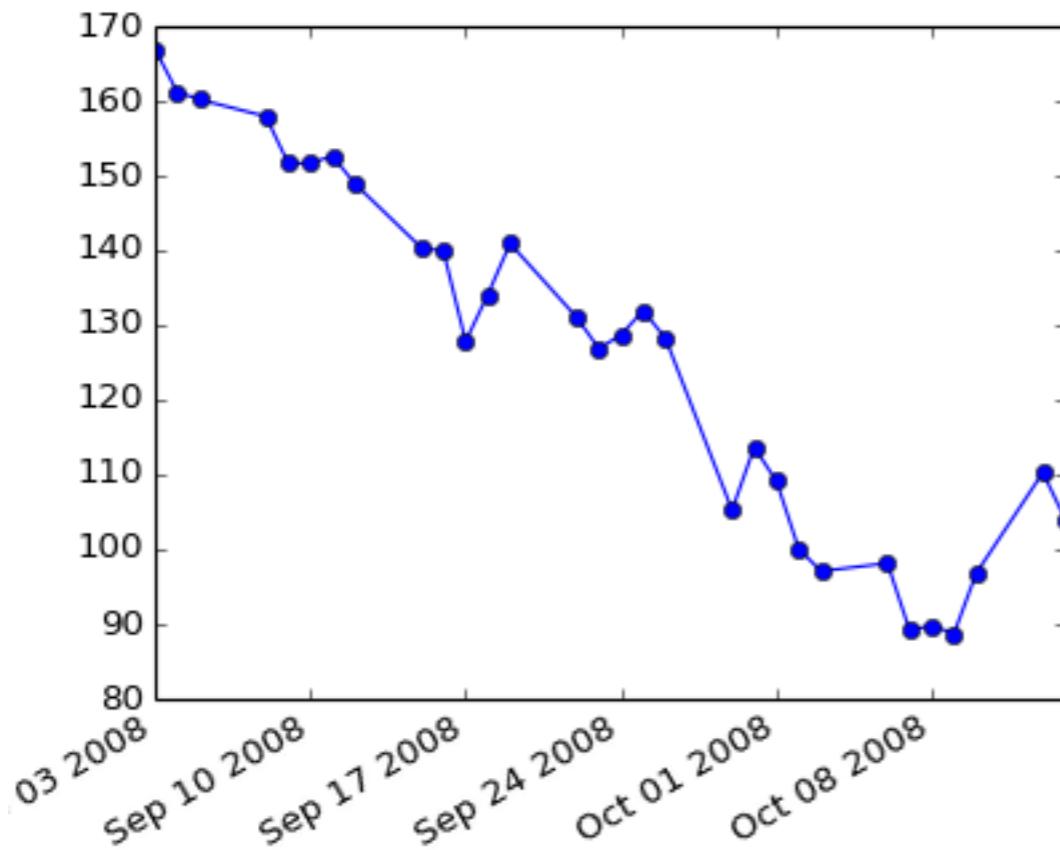
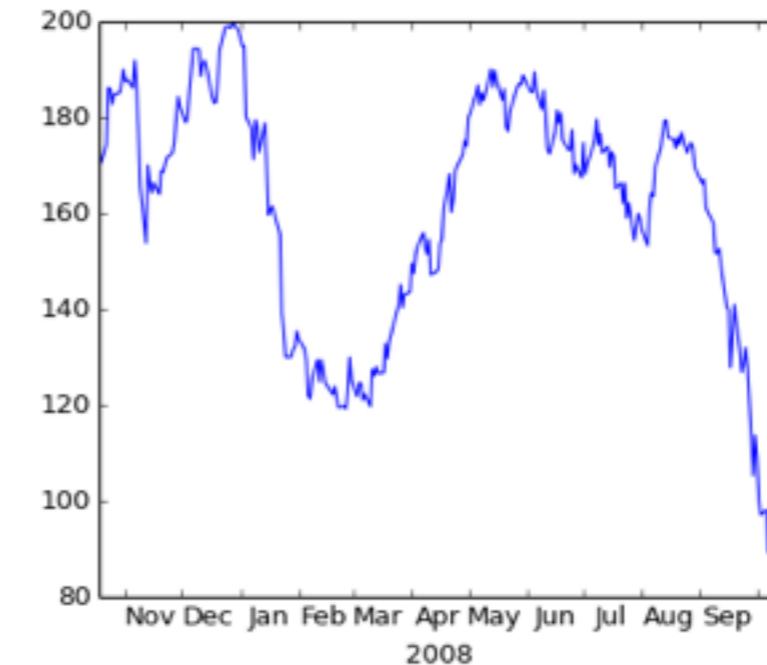
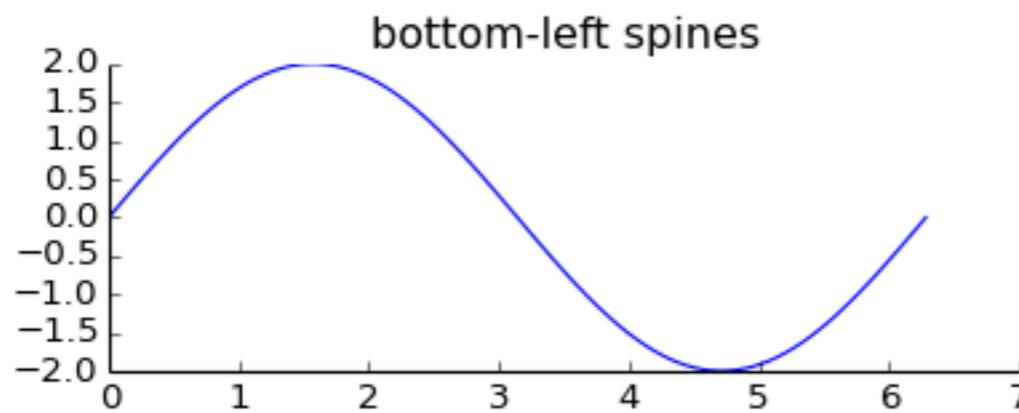
Nicolas Rapp

Trends Over Time

MY HOBBY: EXTRAPOLATING



Line Charts



Linear vs. Logarithmic Scale

May 1990: AAPL 1.4732

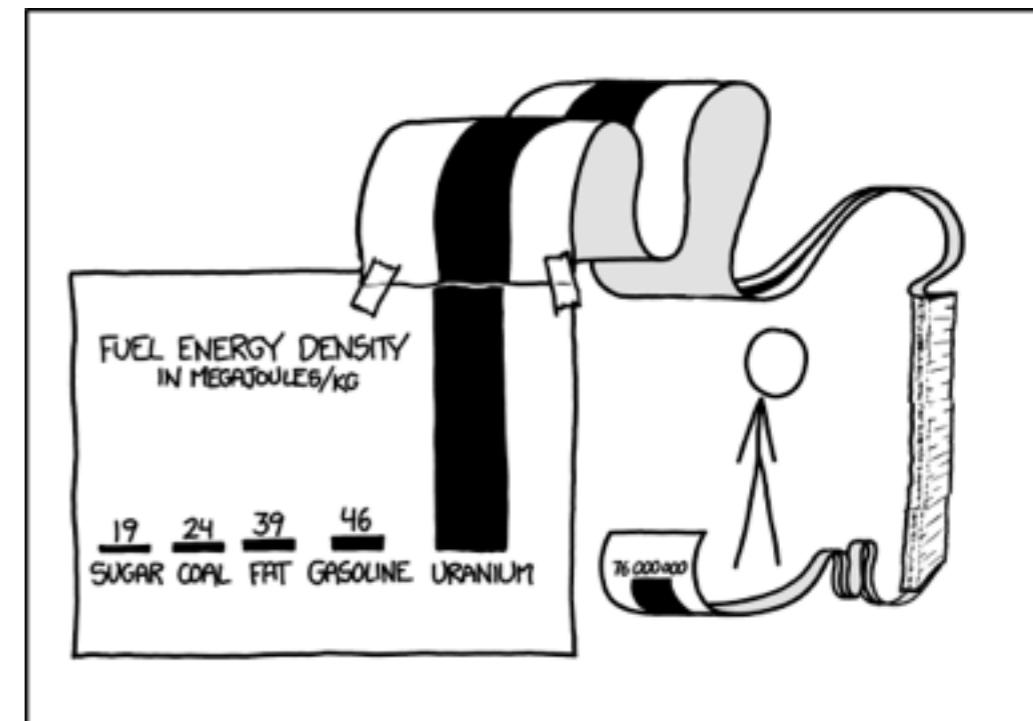
Linear Scale

May 1990: AAPL 1.4732

Log Scale

Apple Stock Price

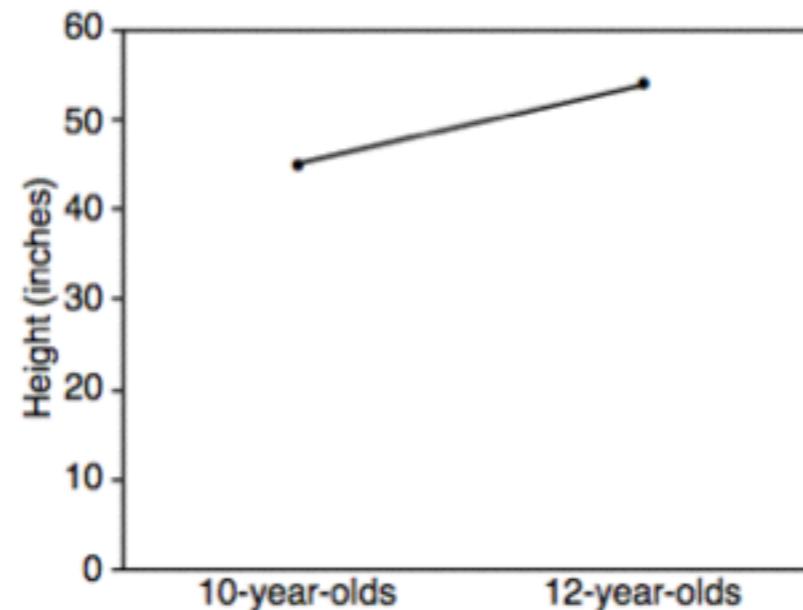
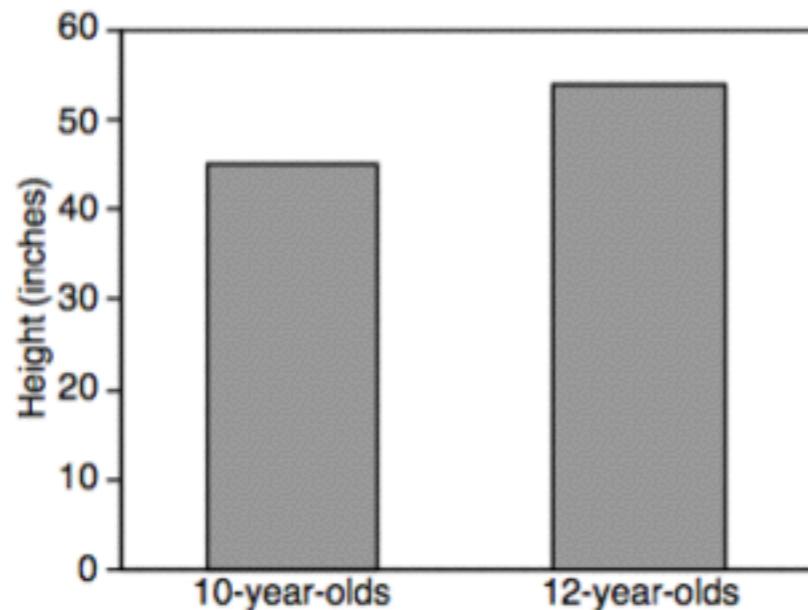
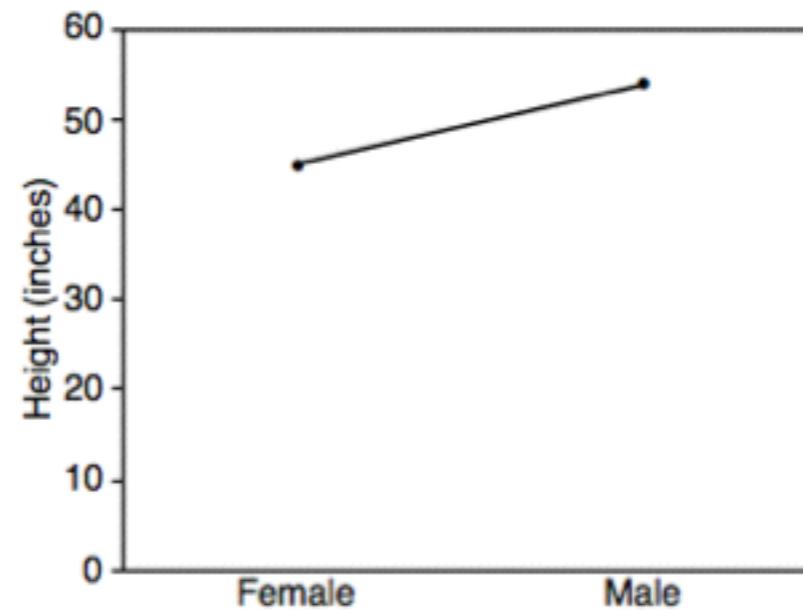
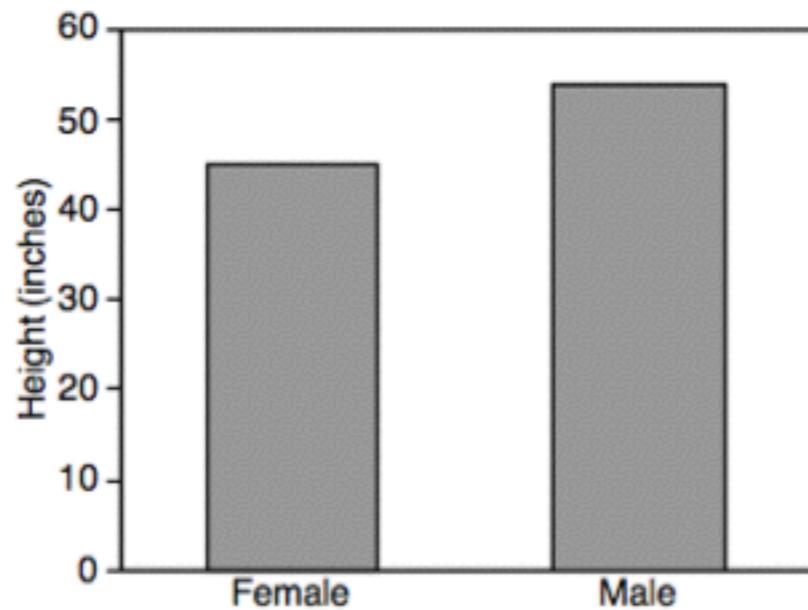
<http://finance.yahoo.com/echarts?s=AAPL>



<http://xkcd.com/1162/>

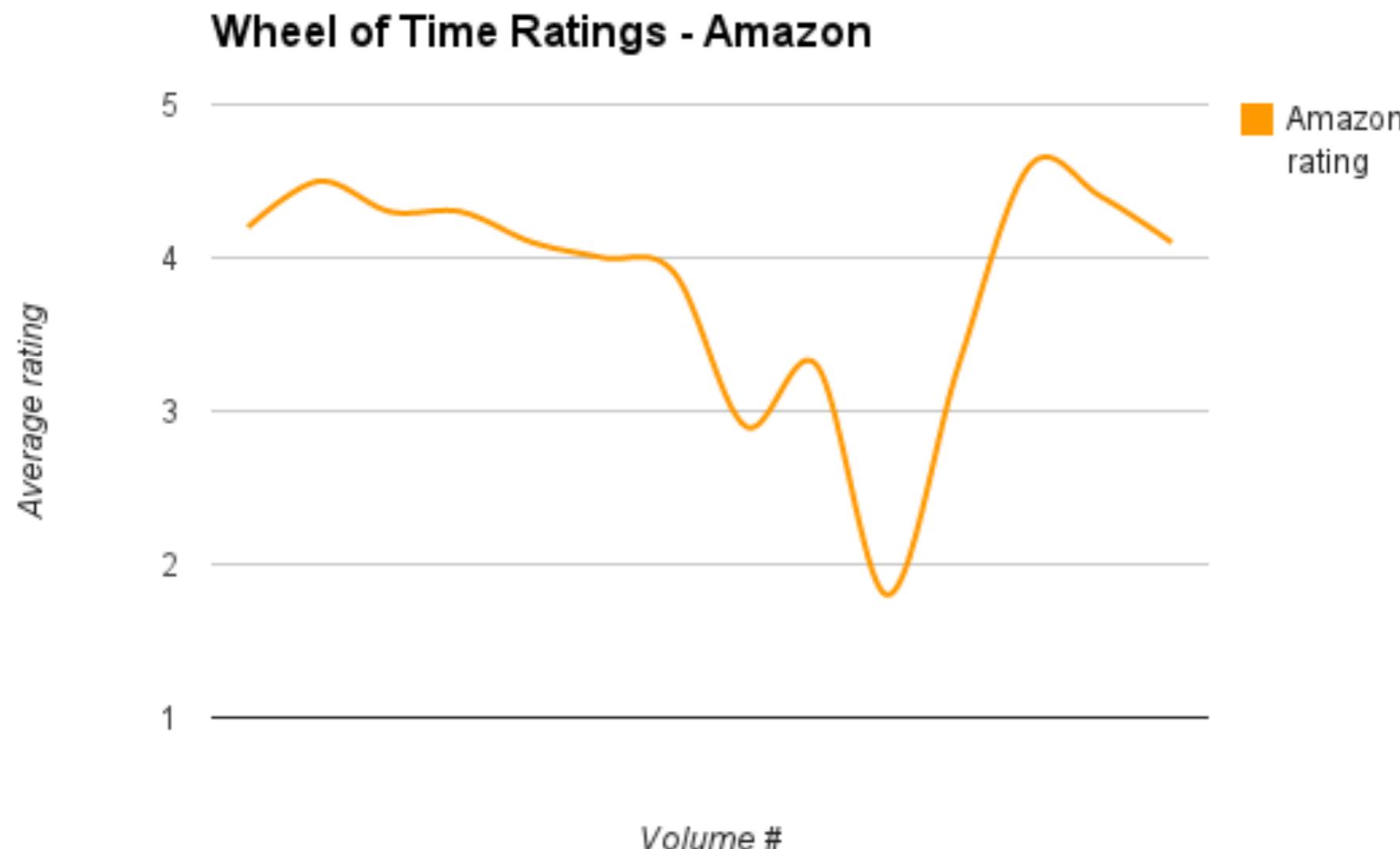
Bars vs. Lines

Lines imply connections - do not use for categorical data



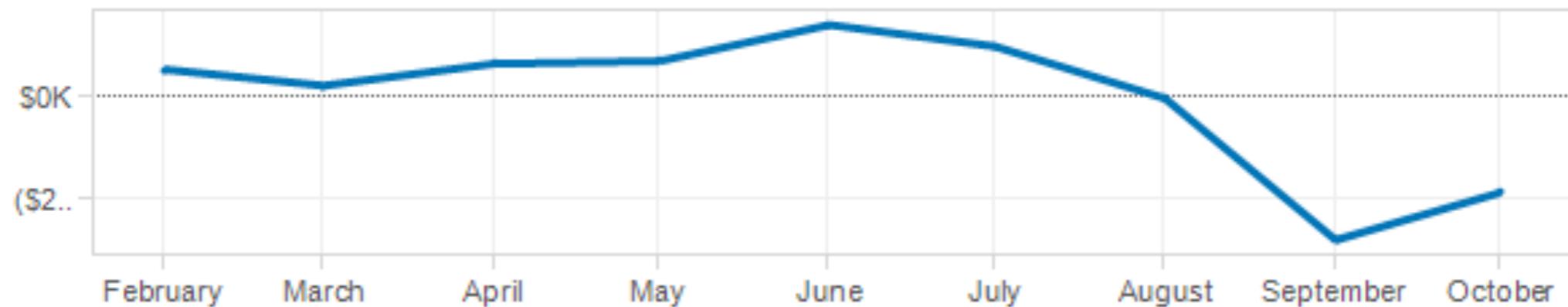
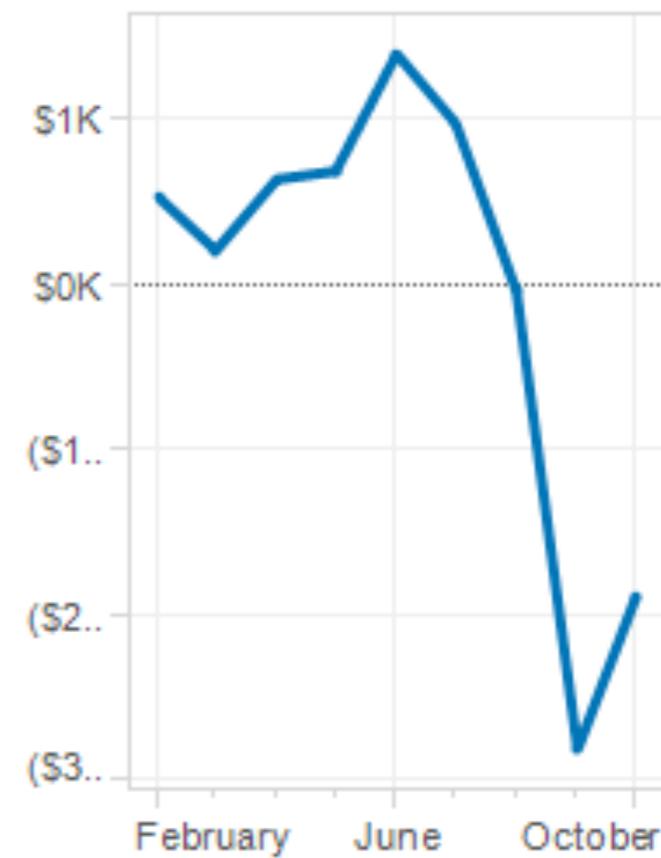
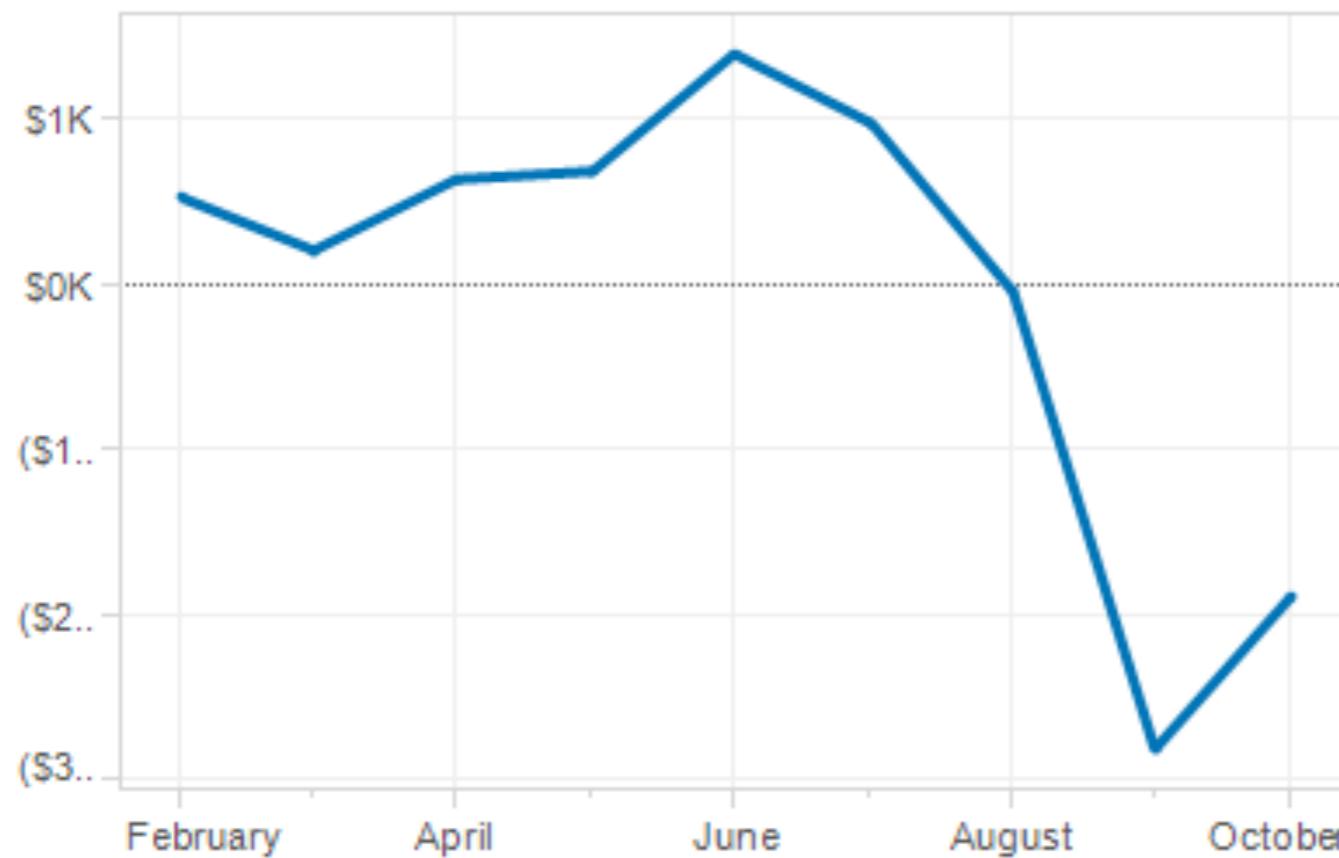
Don't

Use bar charts to compare book ratings



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

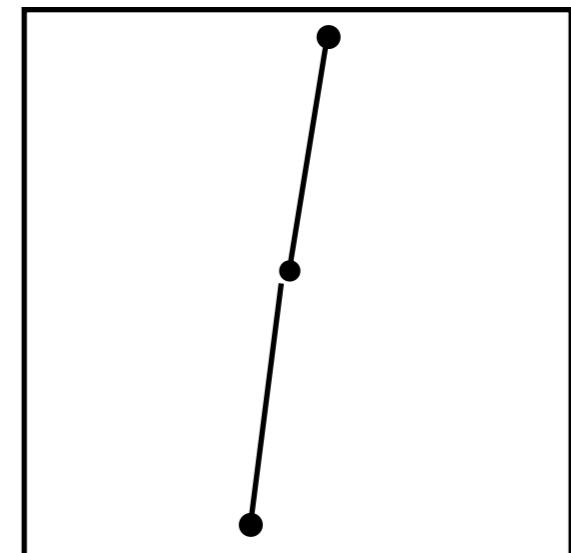
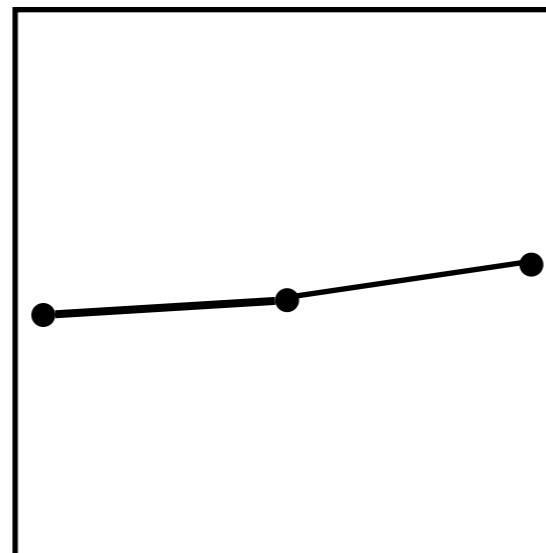
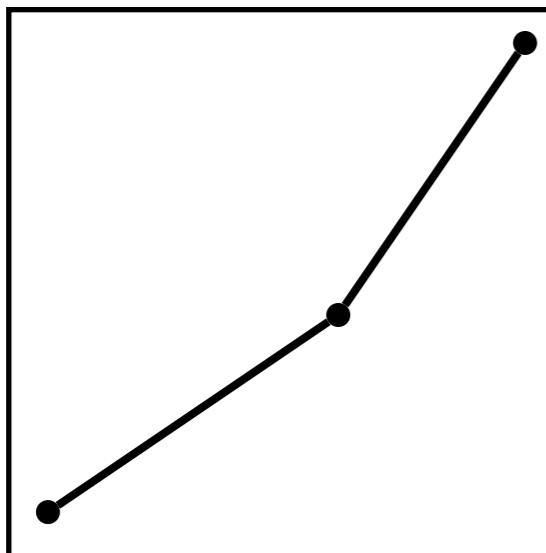
Aspect Ratios



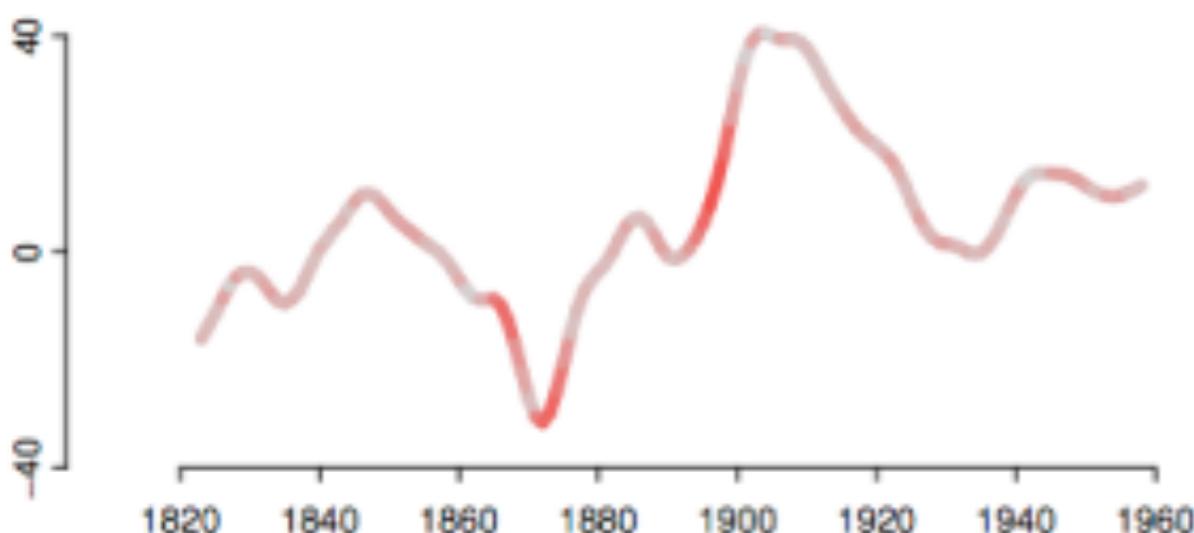
Banking to 45°

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

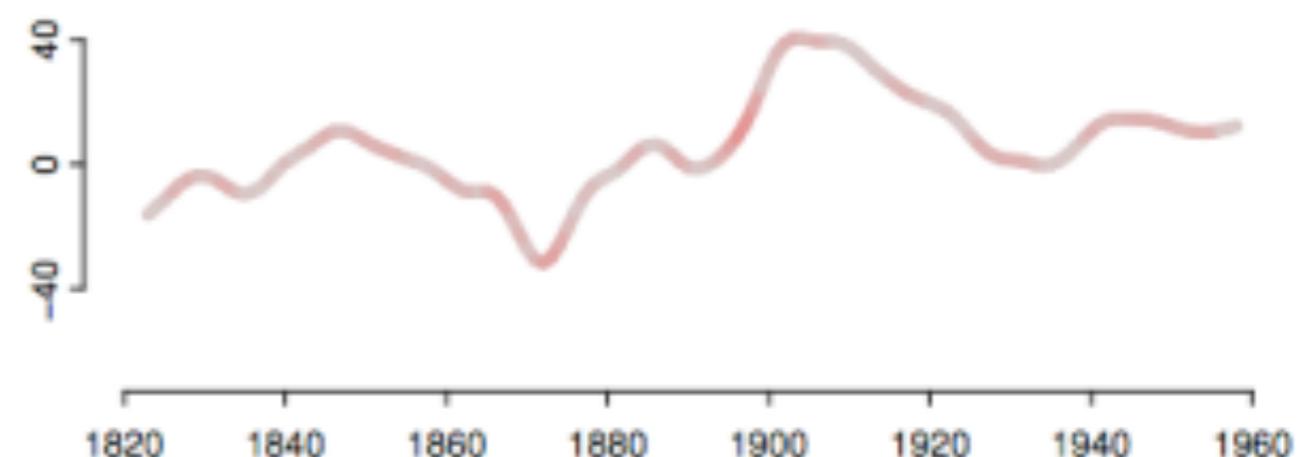
W. Cleveland



Banking to 45°



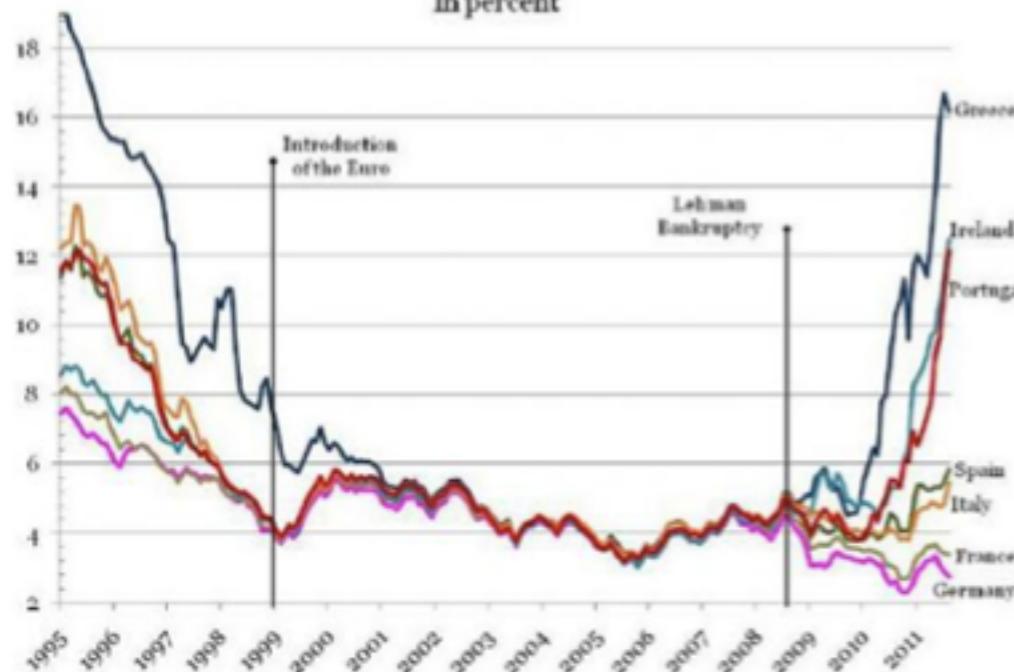
Error Prone



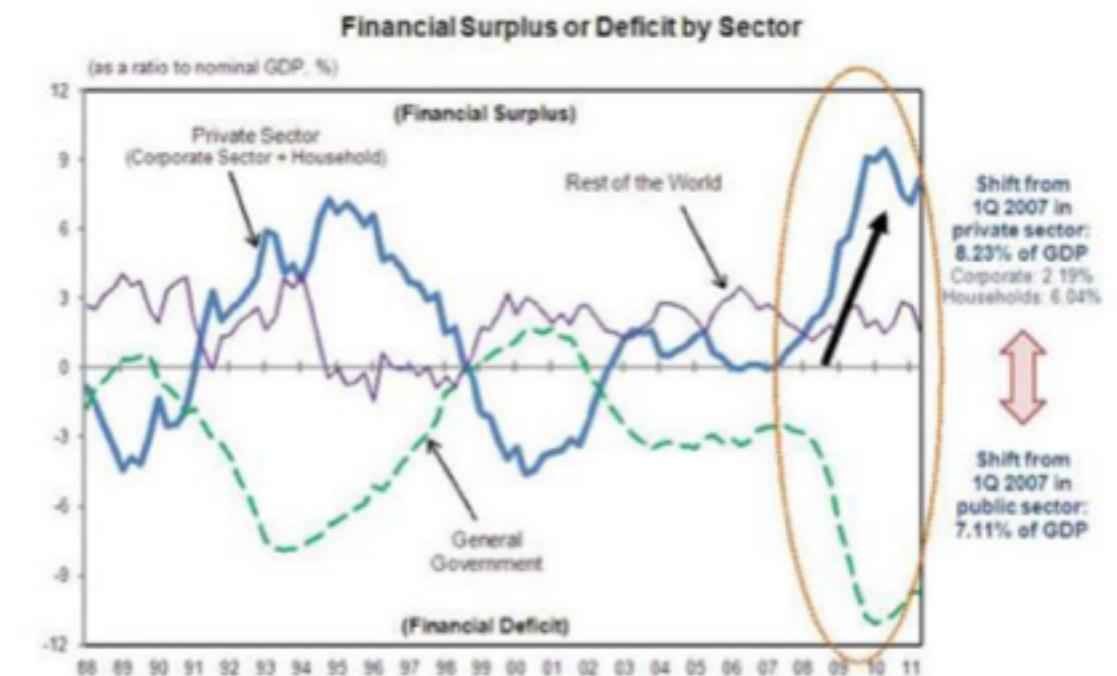
Optimal Aspect Ratio

Don't

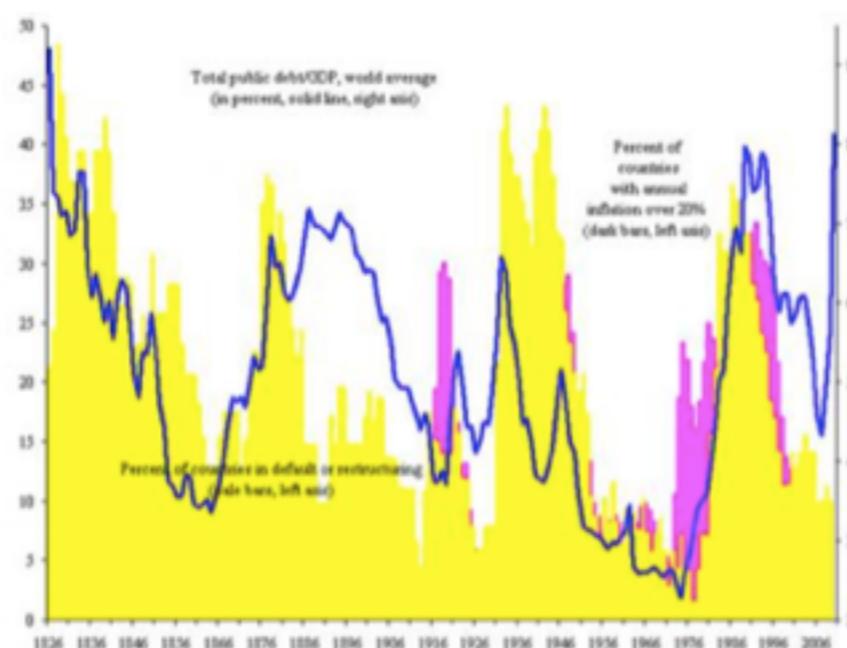
Interest Rates on 10-Year Government Bonds
In percent



**UK in Balance Sheet Recession: UK Private Sector Increased Savings
Massively after the Bubble**

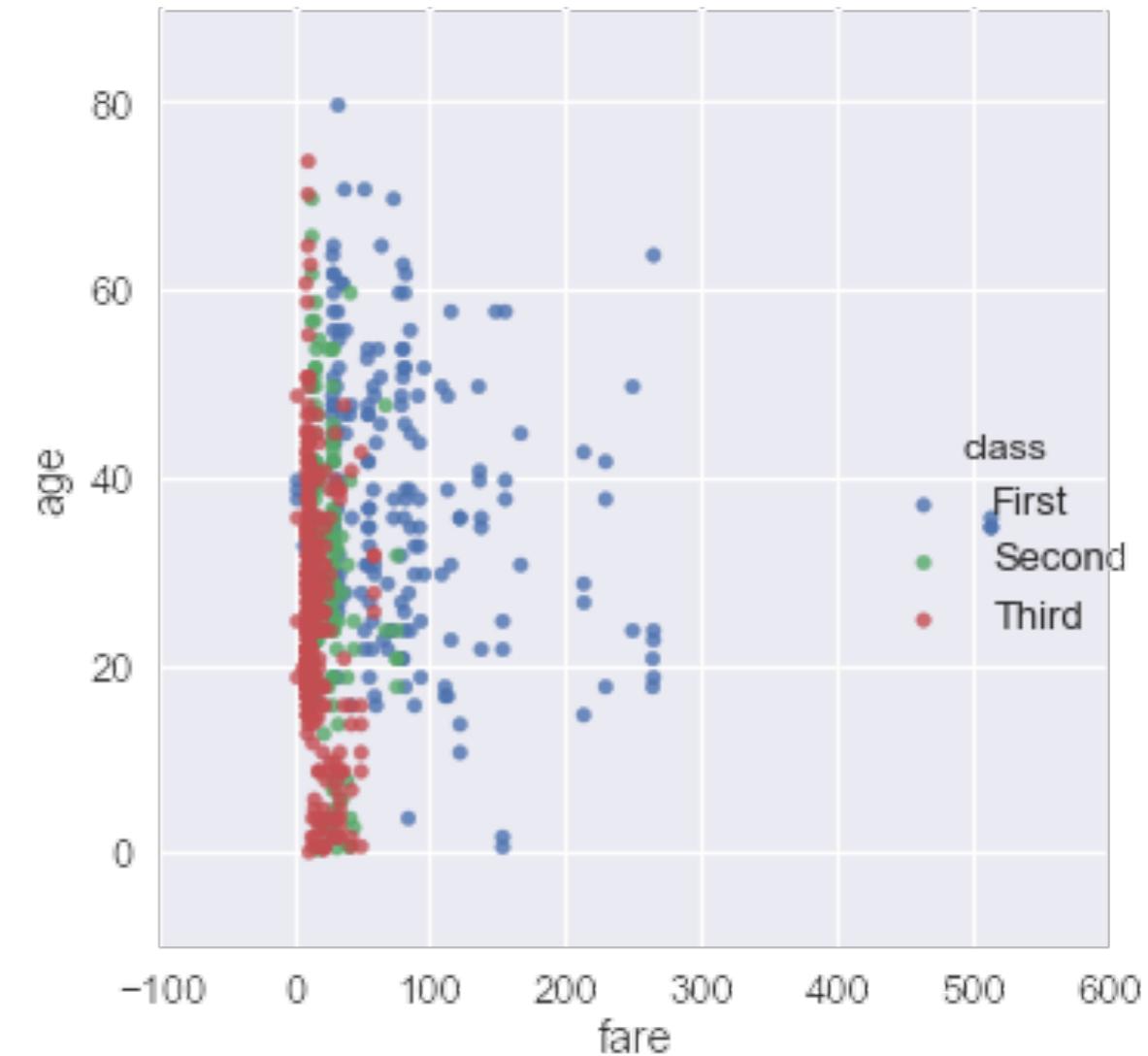
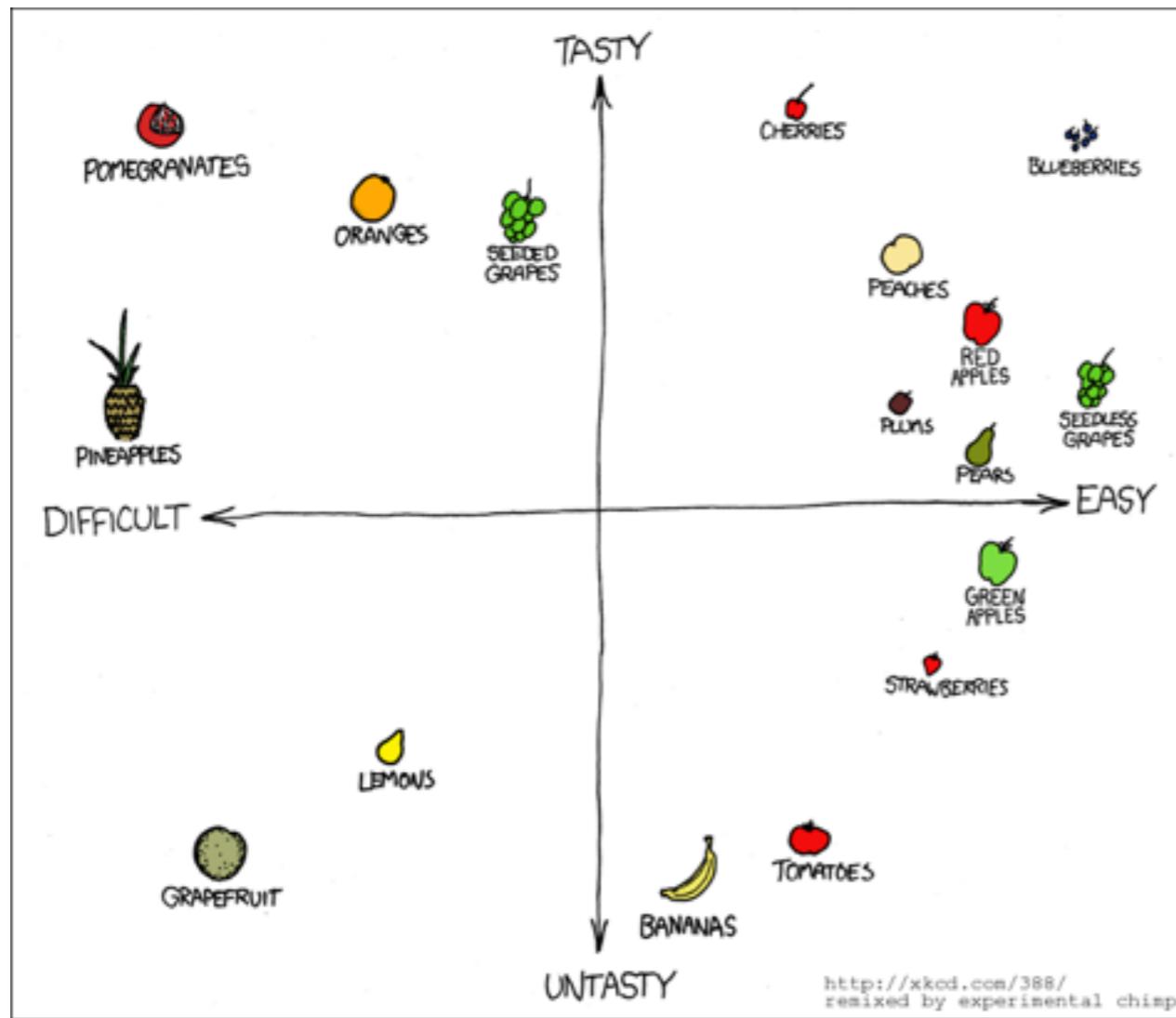


Note: For the latest figures, 4 quarter averages ending with 2Q/11 are used.
Source: Office for National Statistics, UK

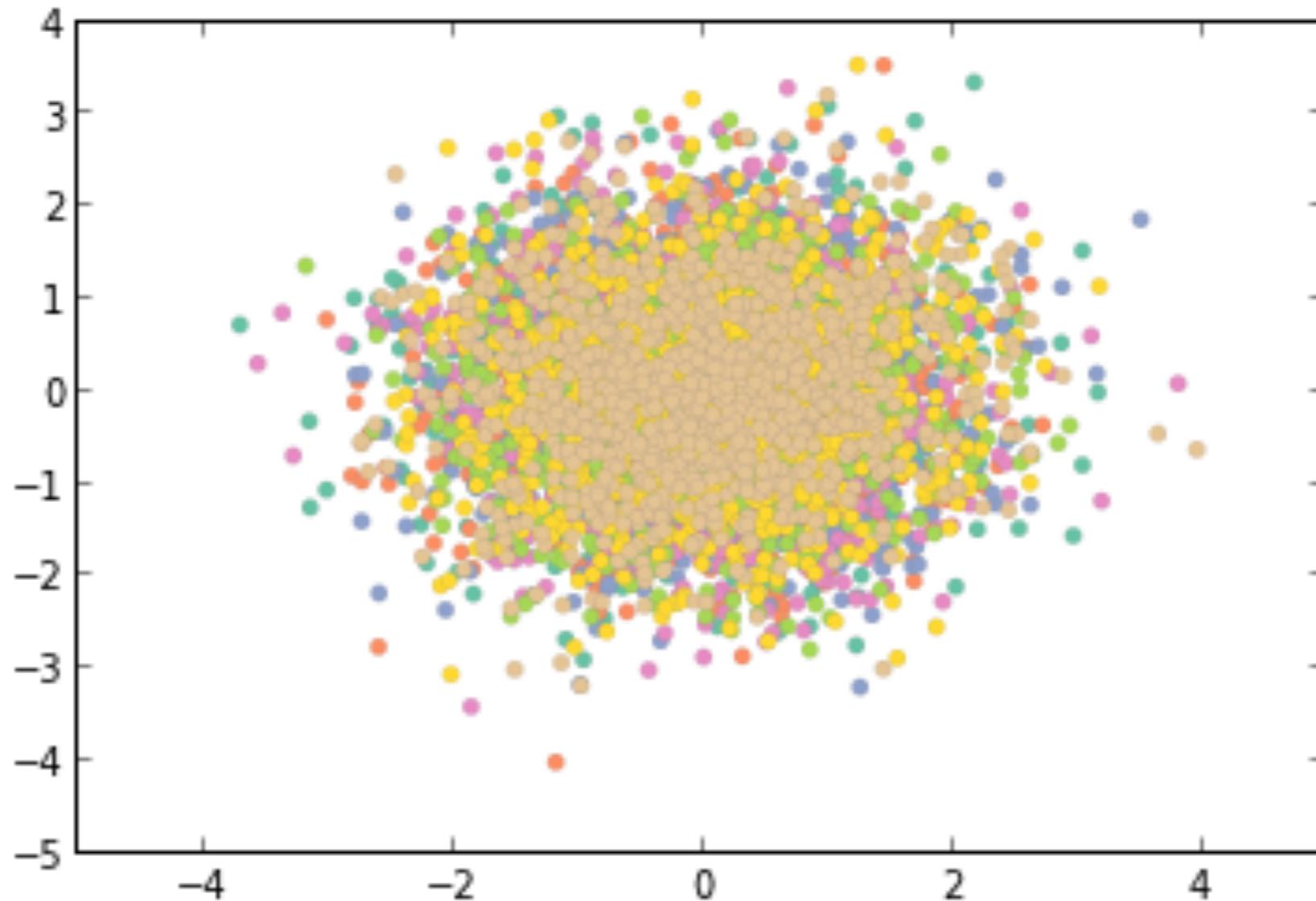


Correlations

Scatterplots

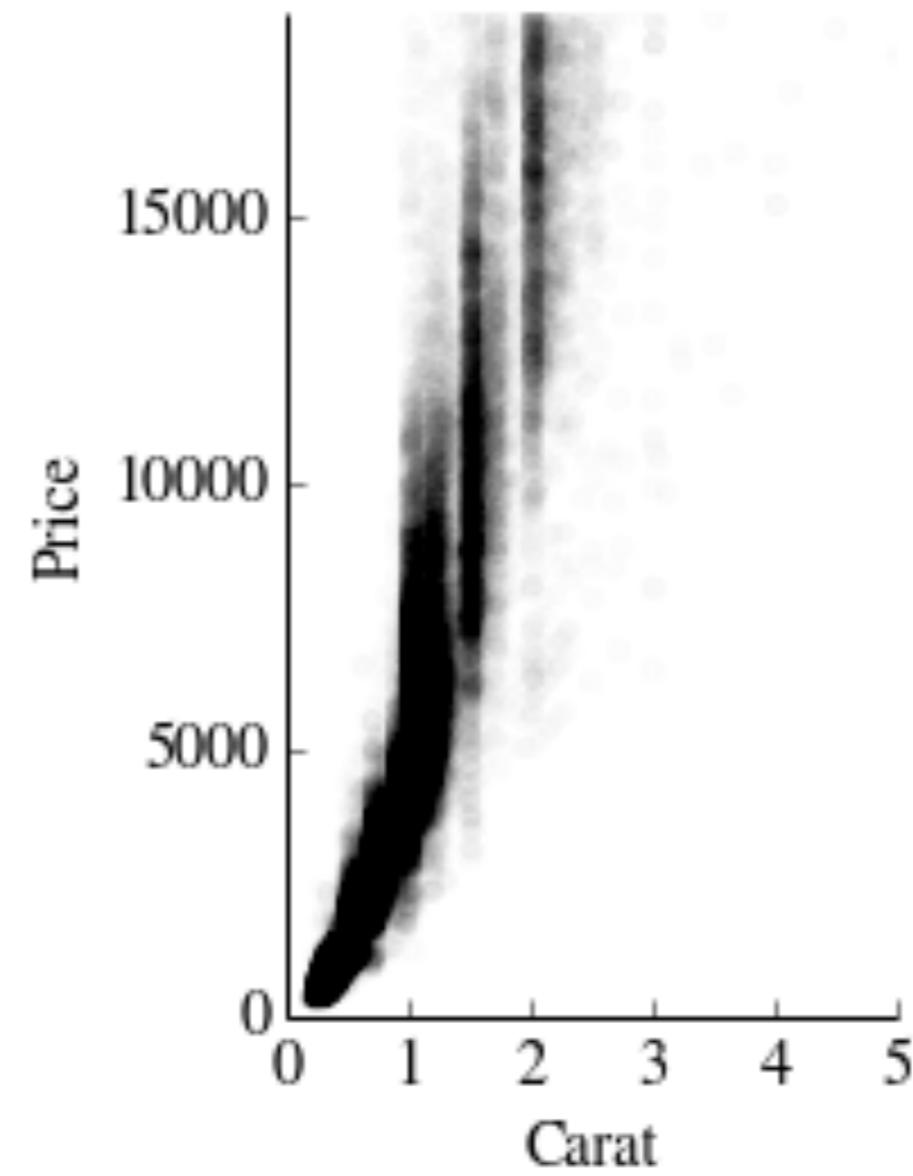
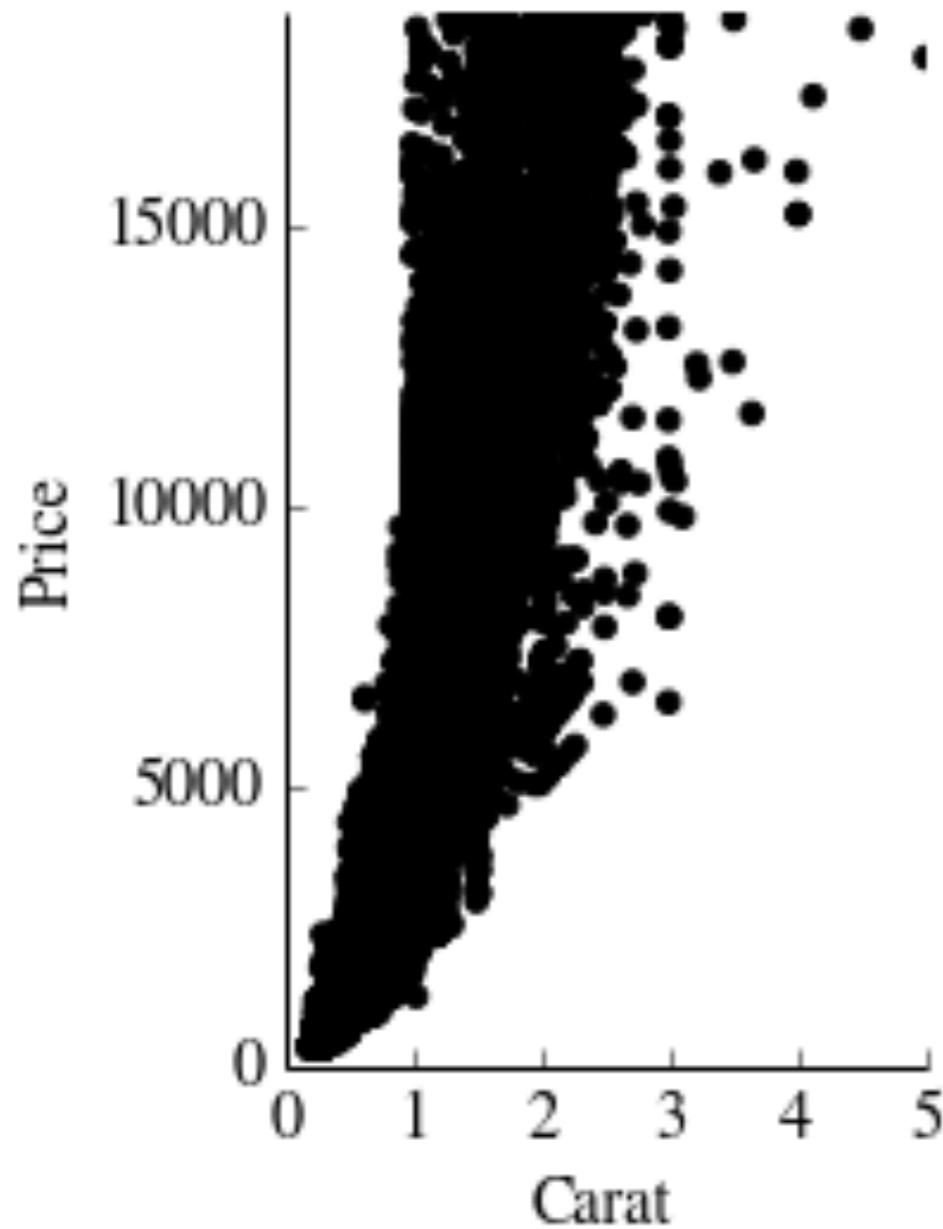


Scatterplots



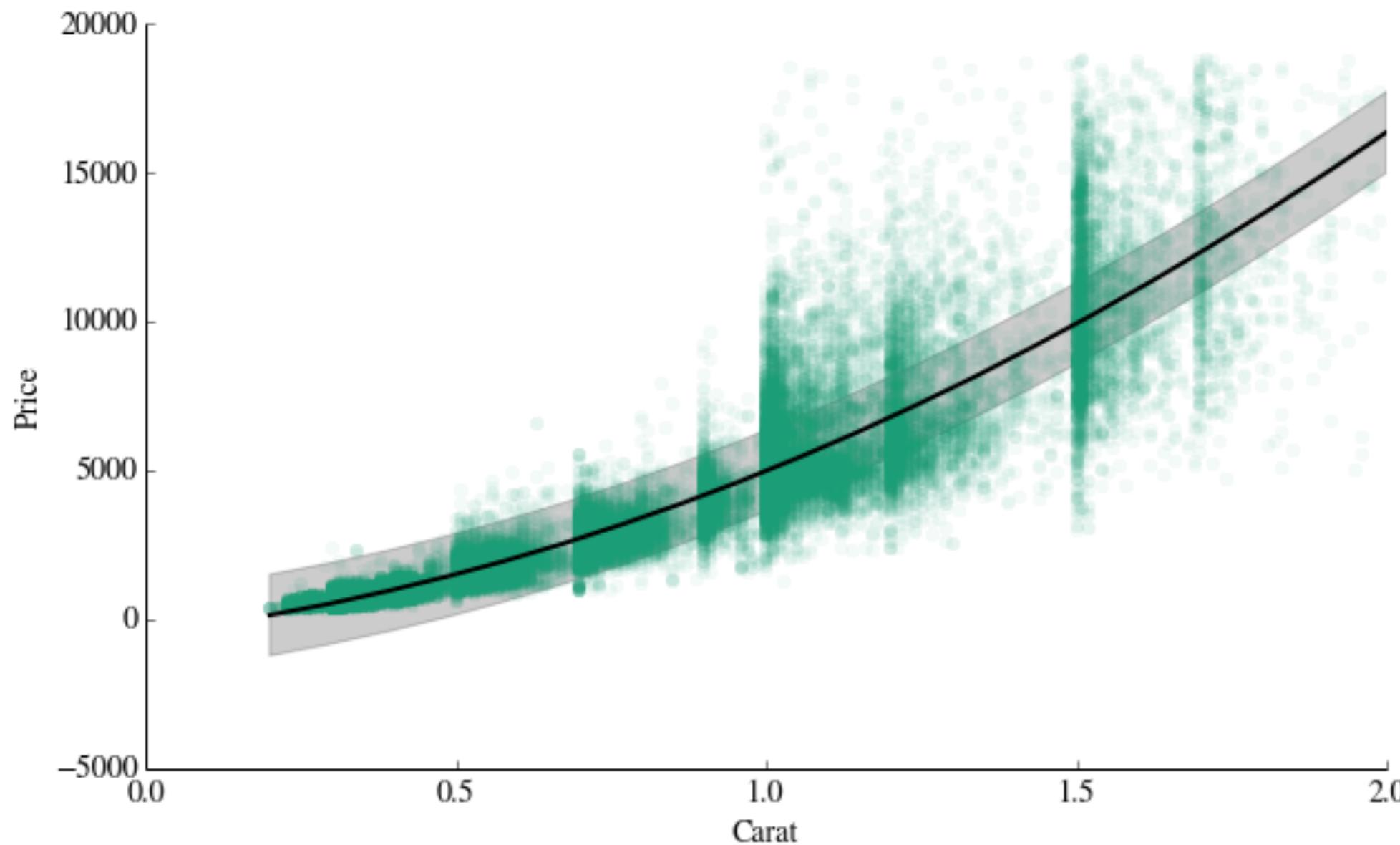
Light Grey Border

Overplotting

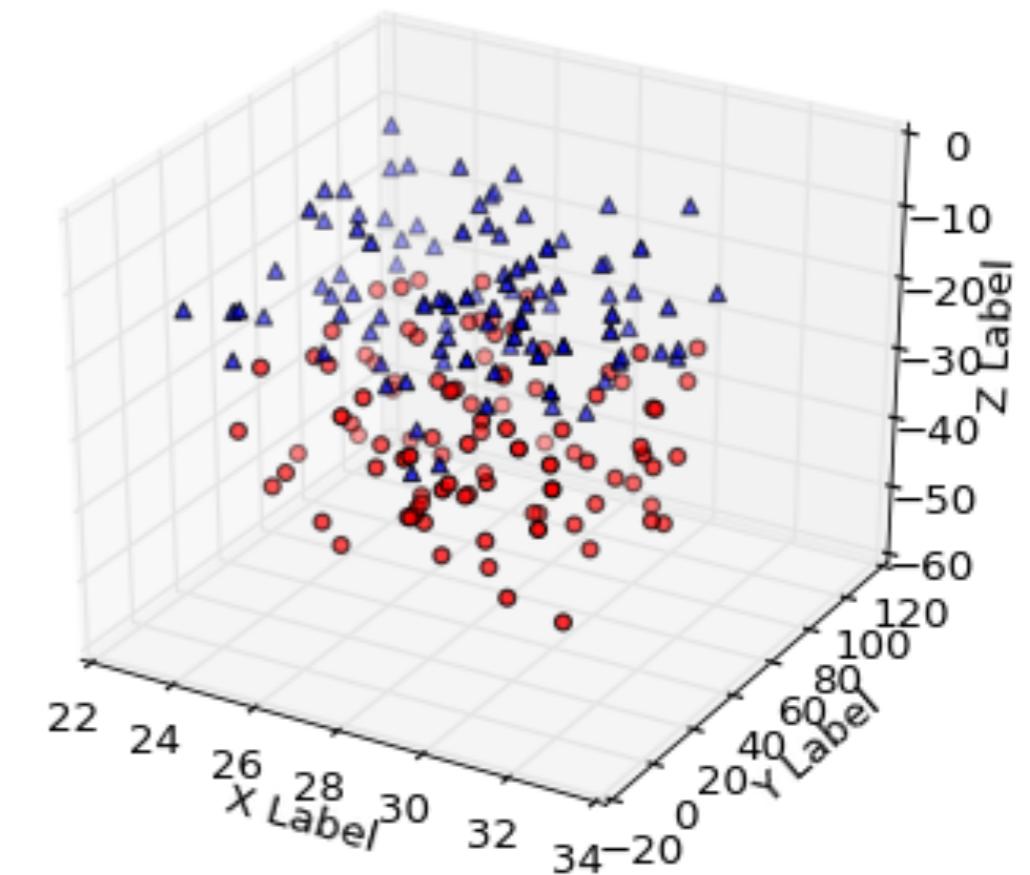
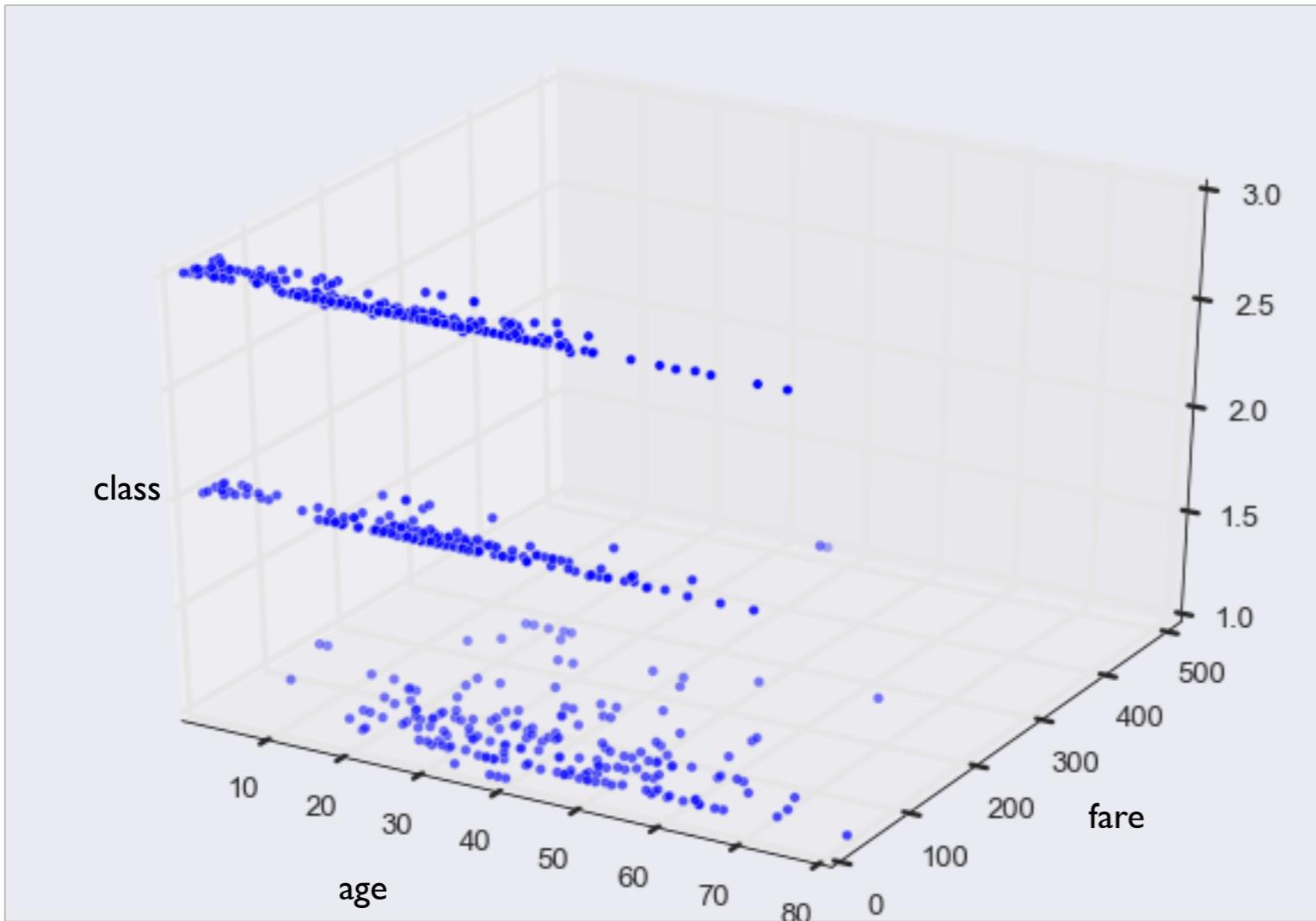


$\alpha = 1/100$

Trend Lines

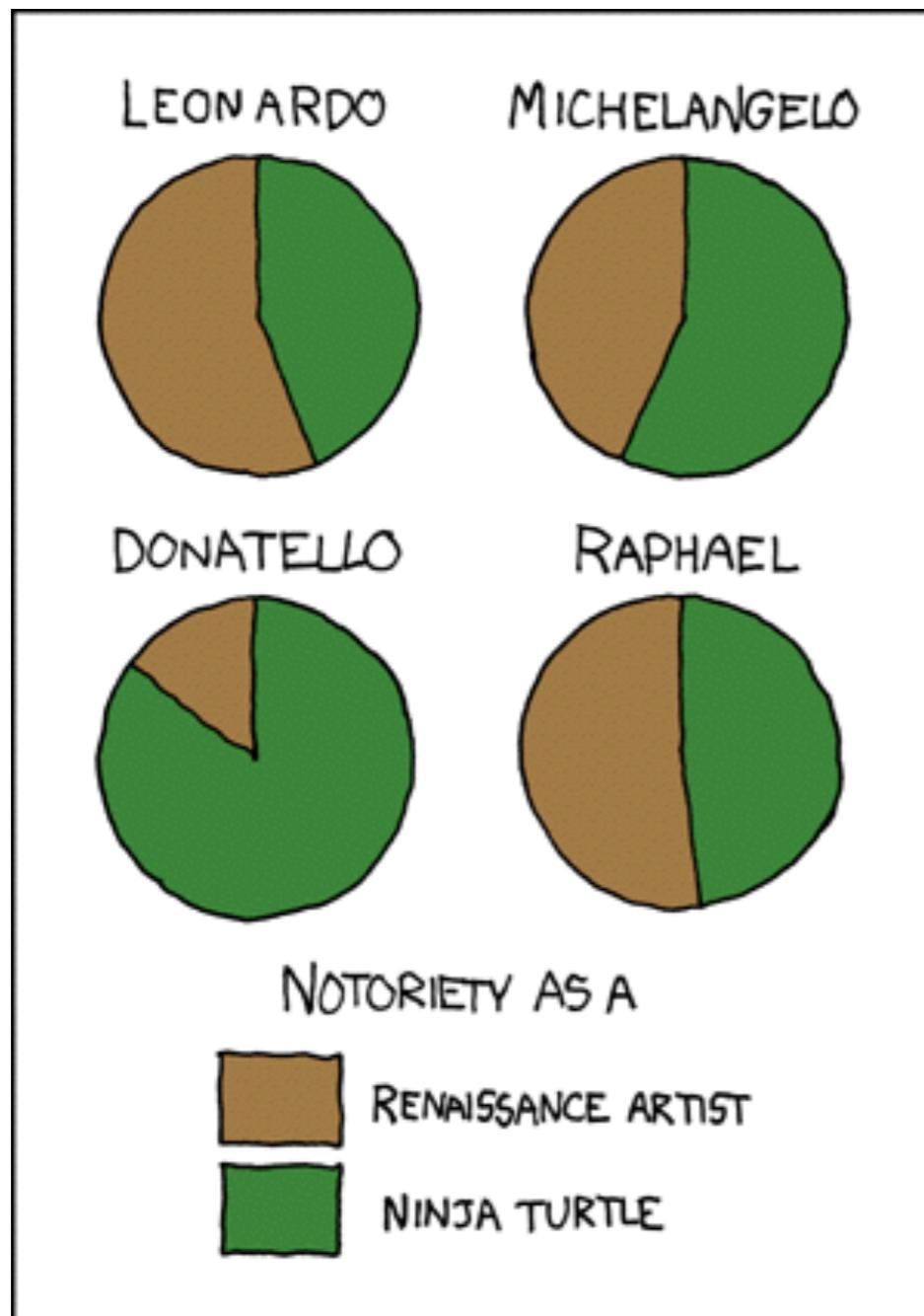


Don't

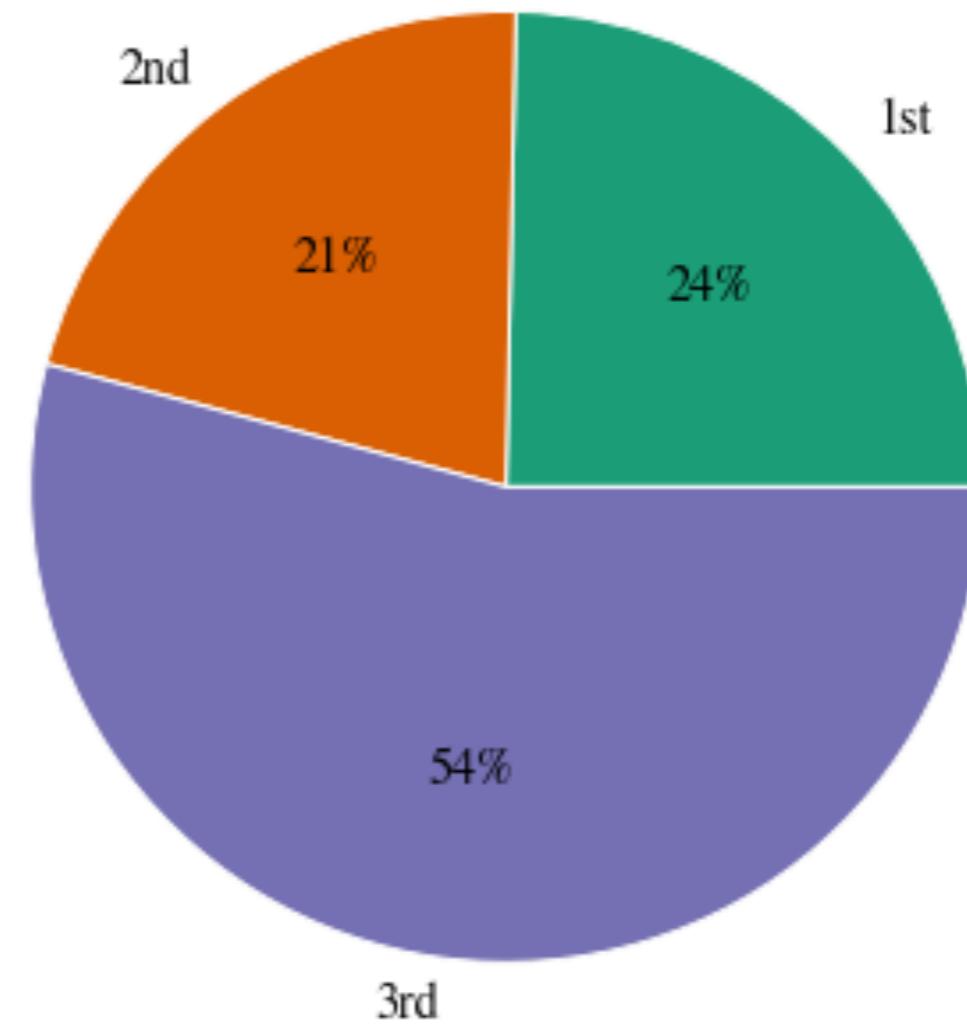


Compositions

Pie Charts

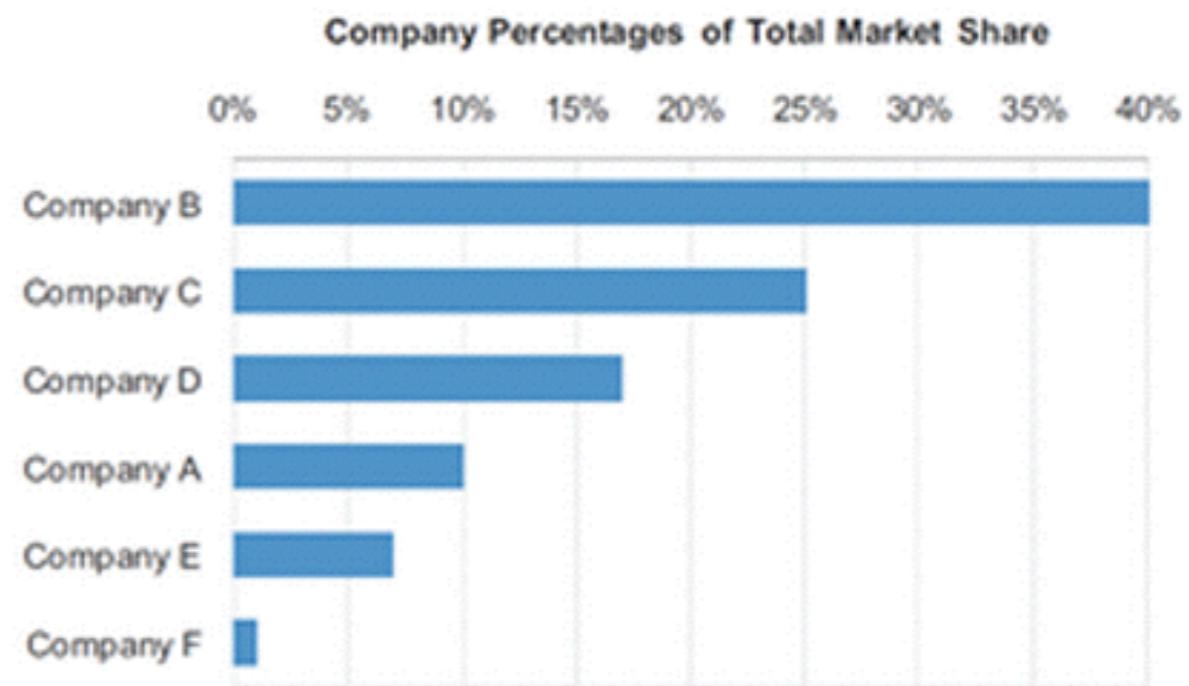
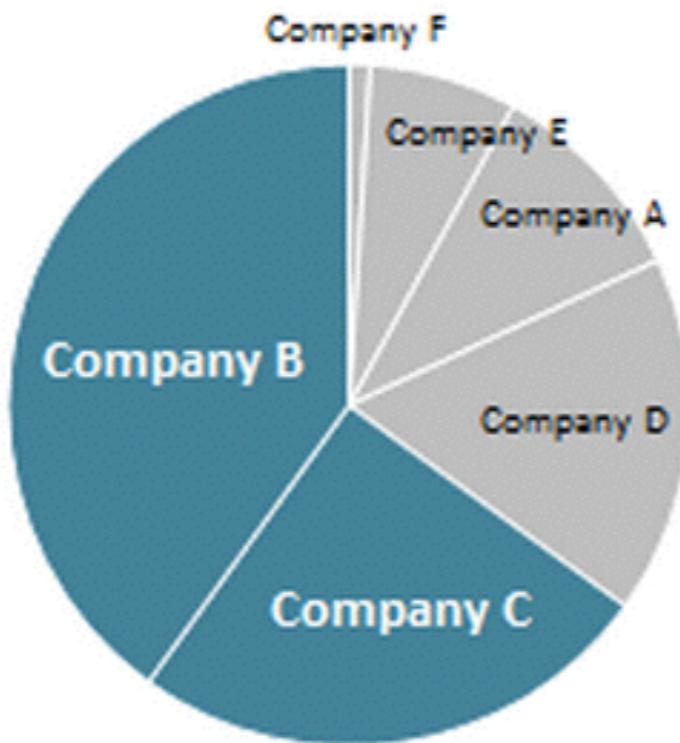


Passenger Class on the Titanic



Pie vs. Bar Charts

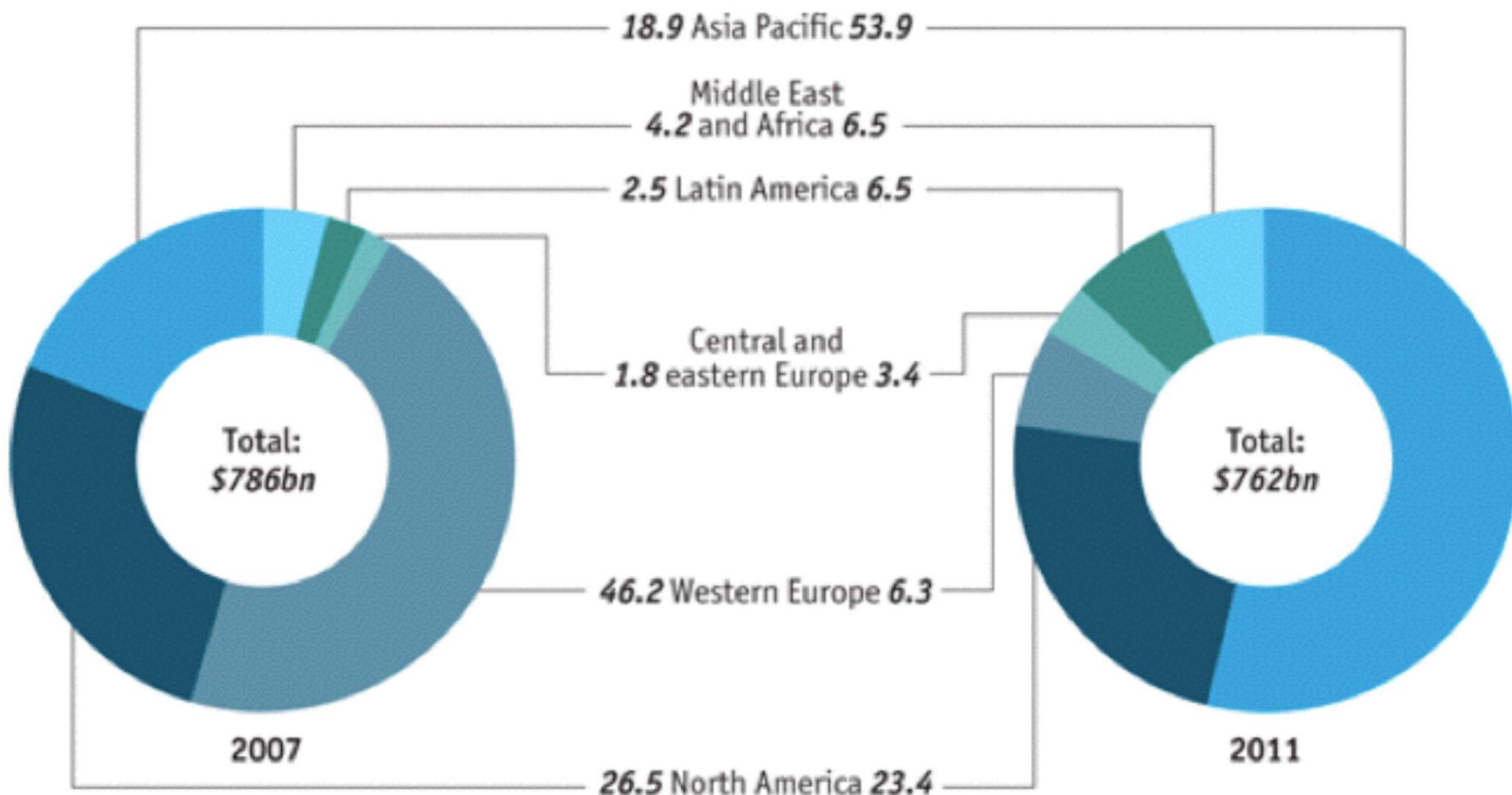
65% of the market is controlled by companies B and C



Donut Chart

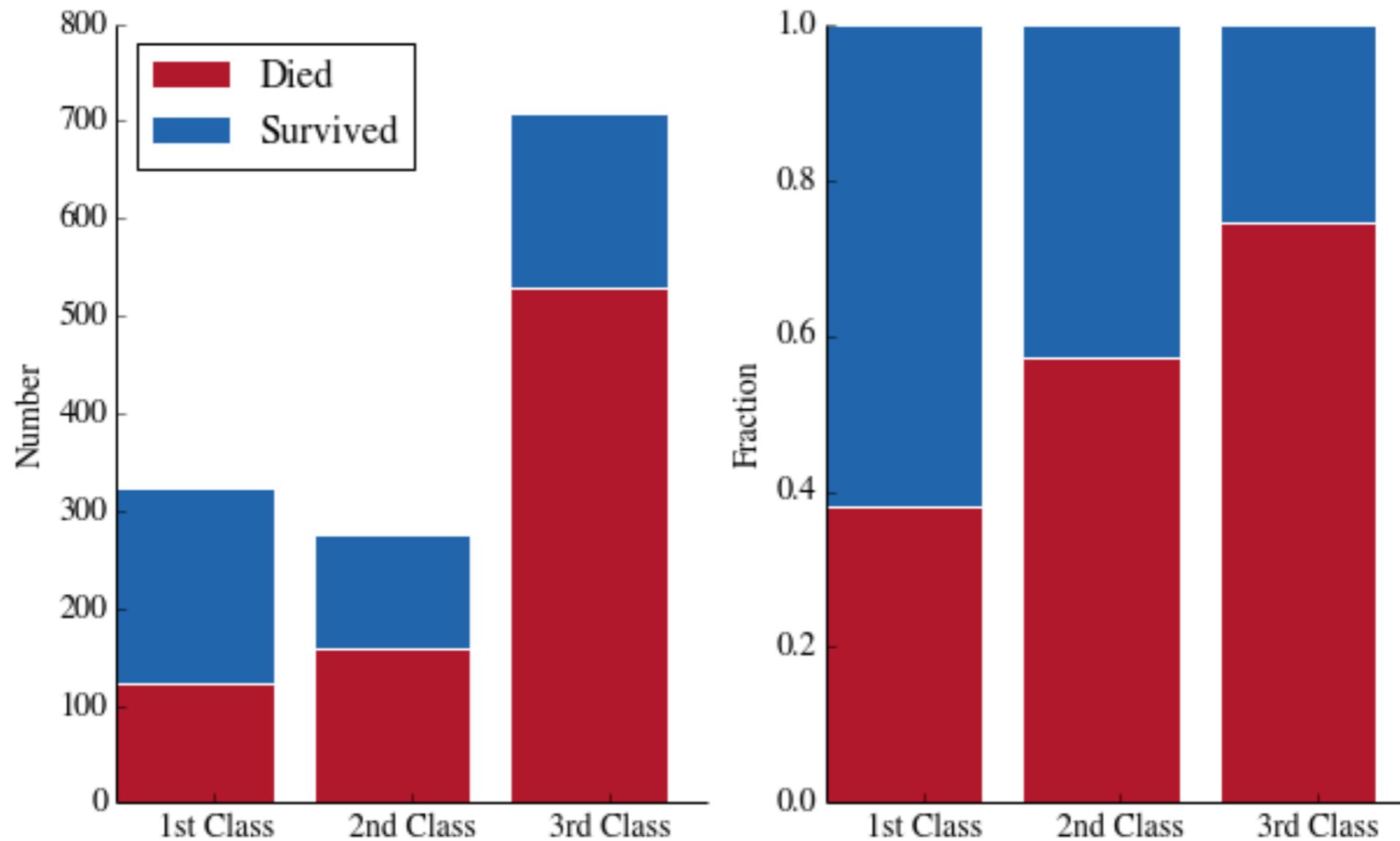
Pre-tax profits of the 1,000 largest banks

By tier-one capital and domicile, % of total

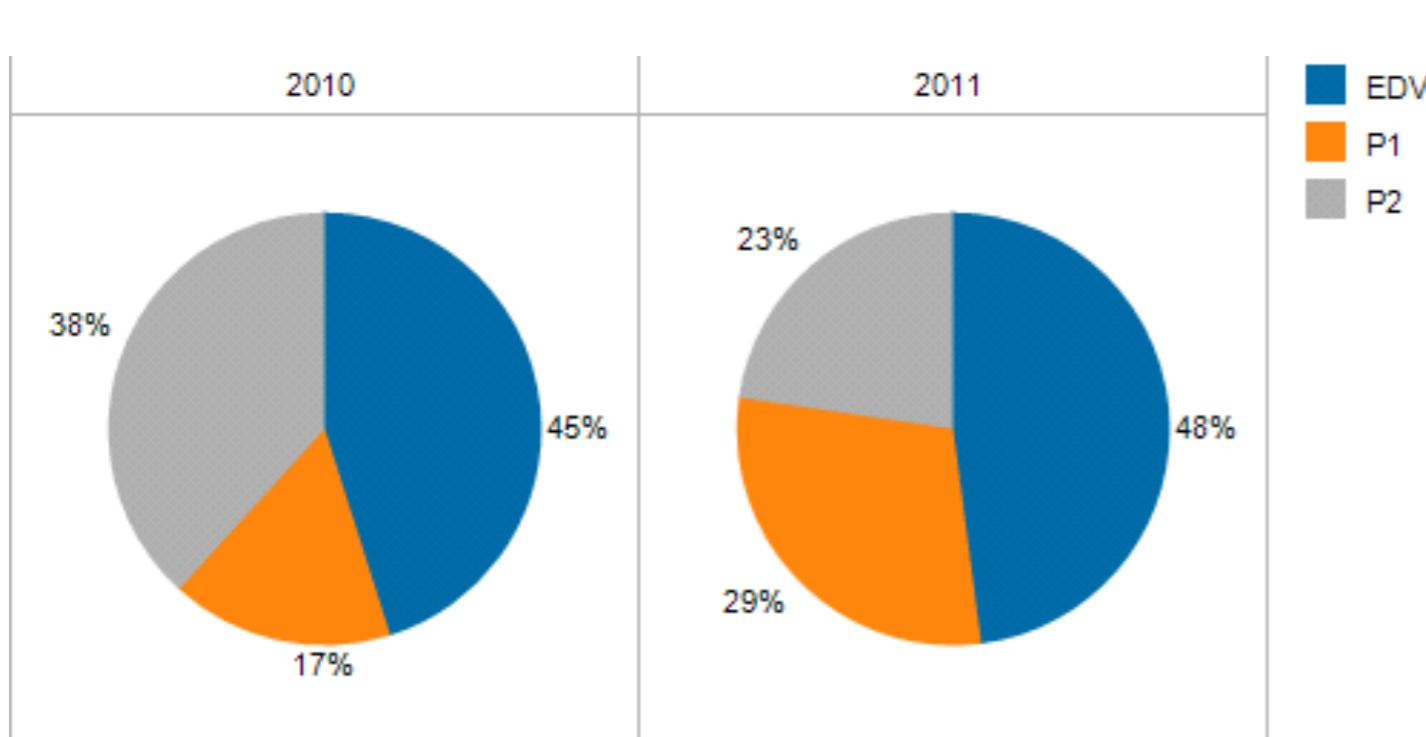


Source: *The Banker Top 1000*

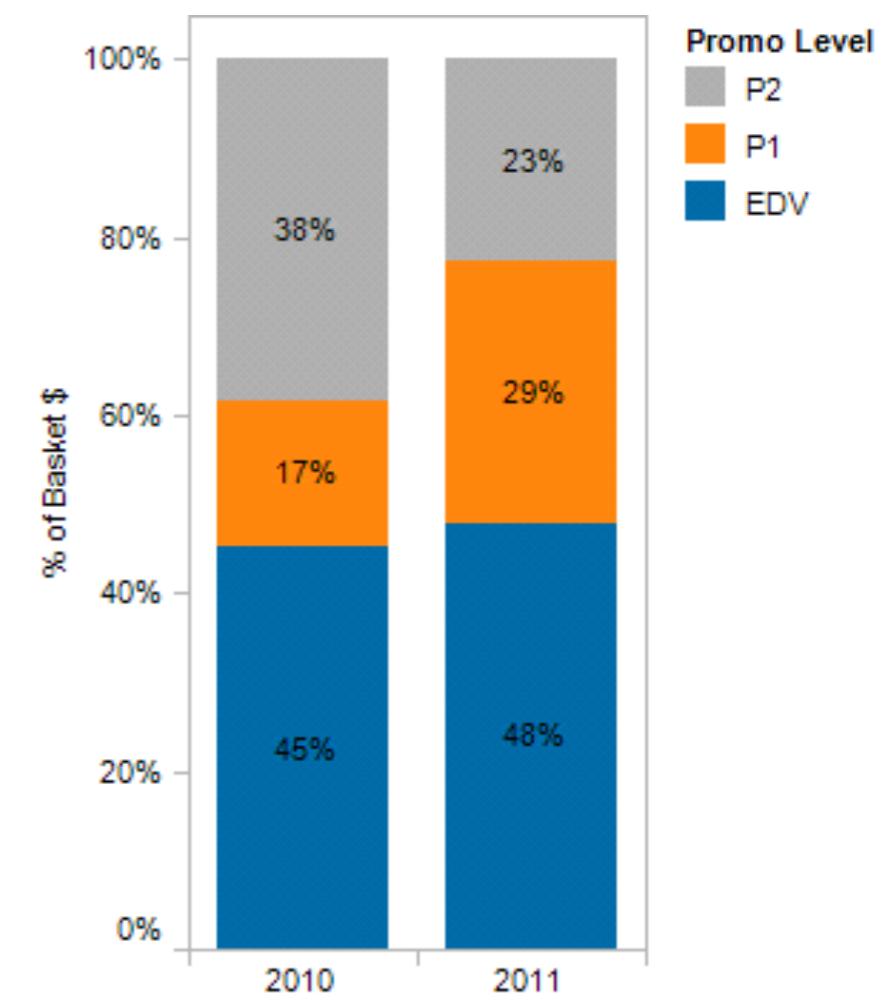
Stacked Bar Chart



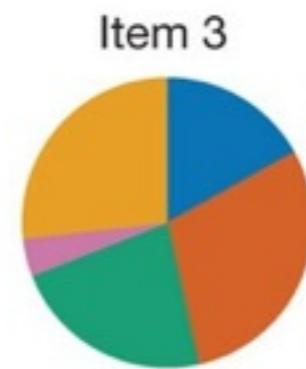
Stacked Bar Chart



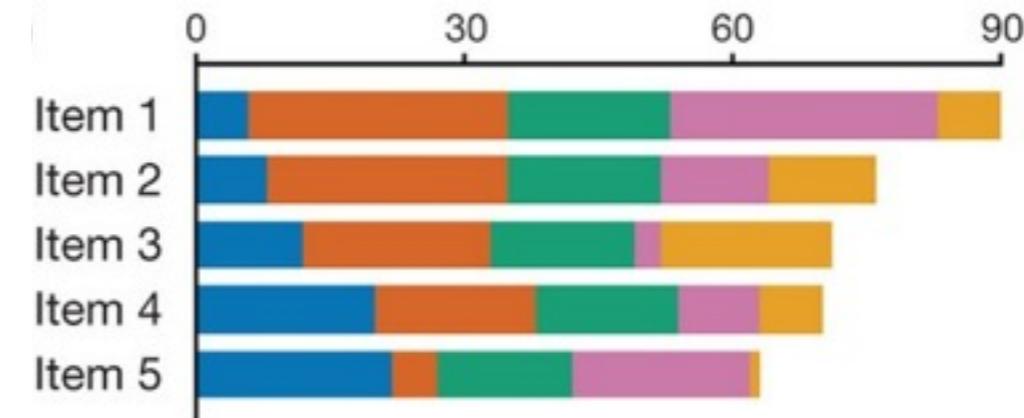
VS.



Comparison of bar chart types

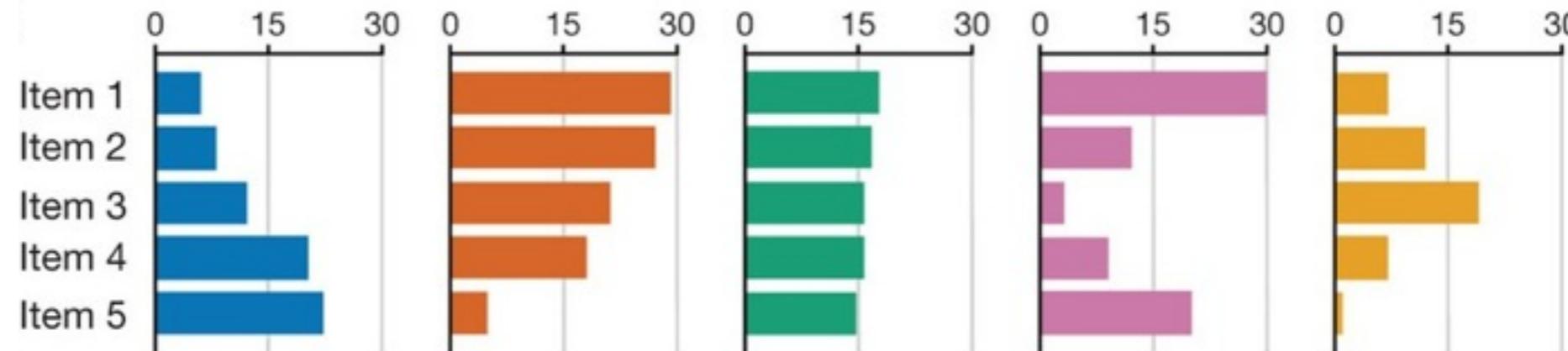


Pie Chart

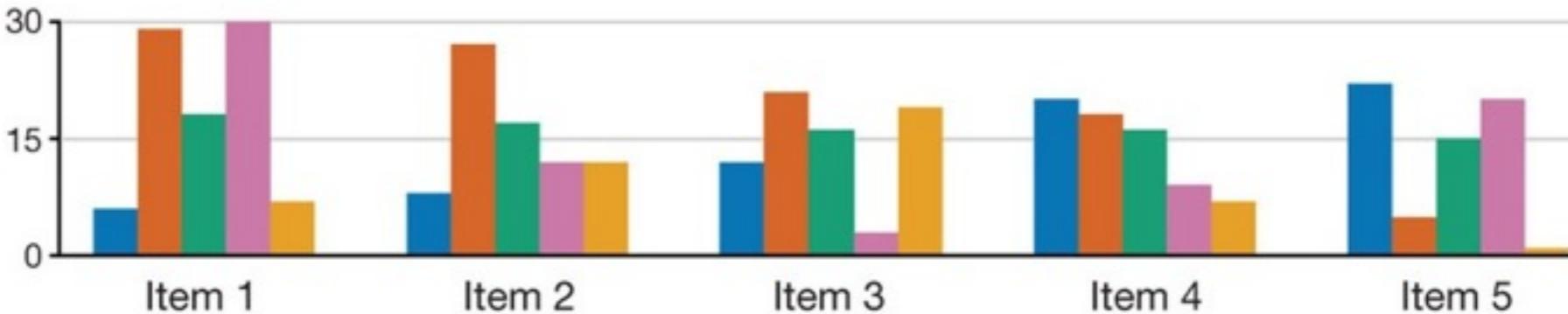


Stacked bar chart

Layered
Bar
Chart

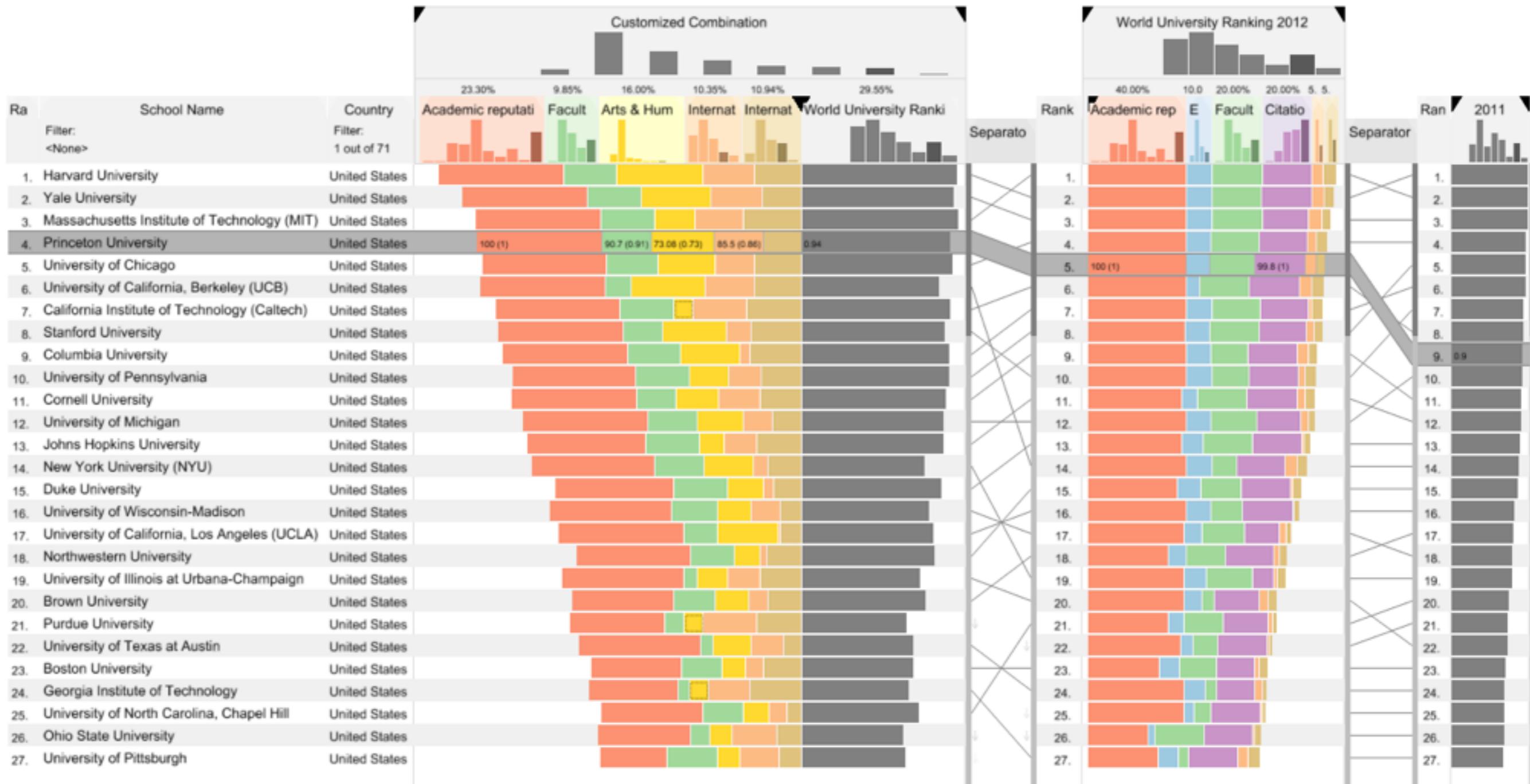


Grouped
Bar
Chart



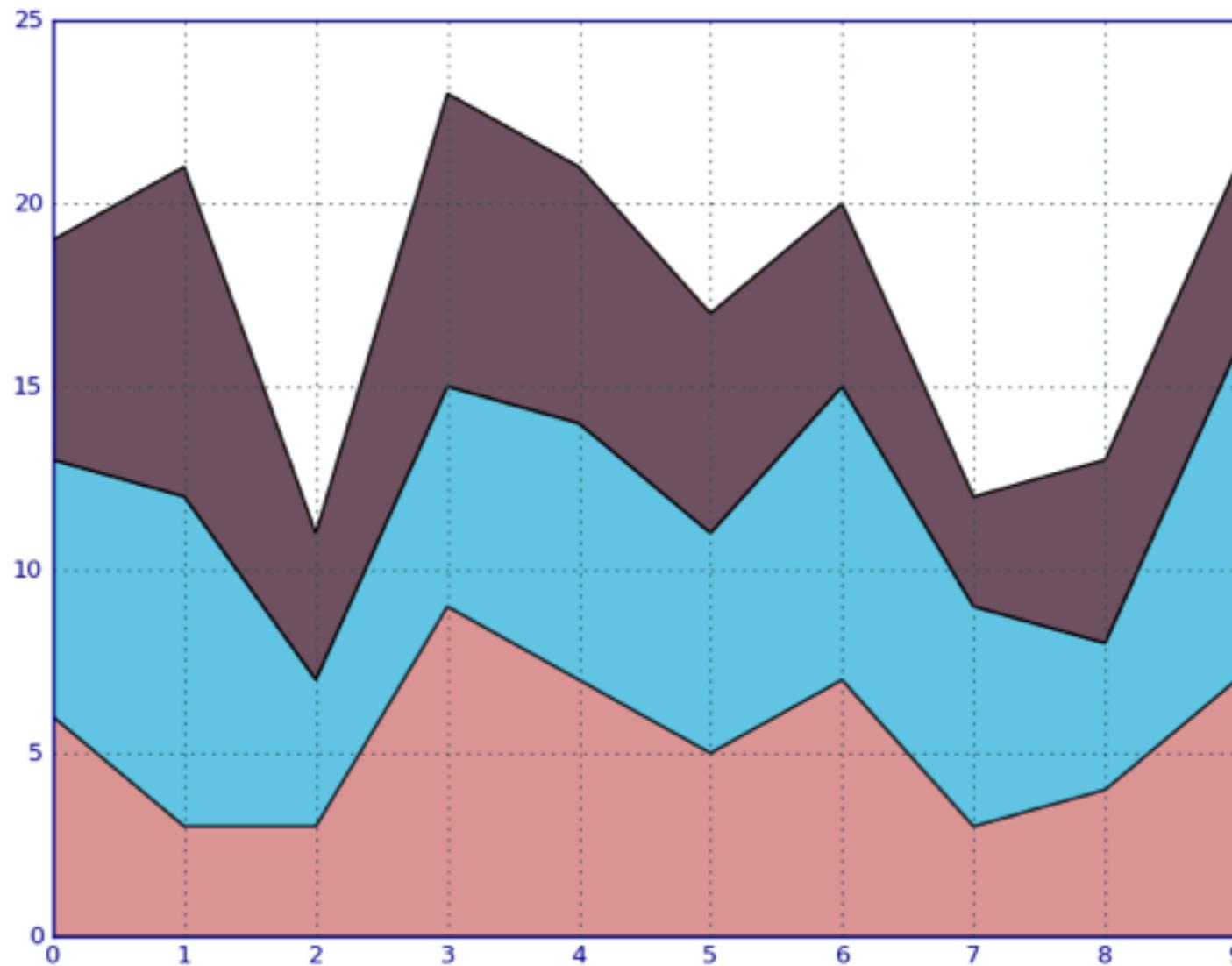
Small
Multiples

LineUp



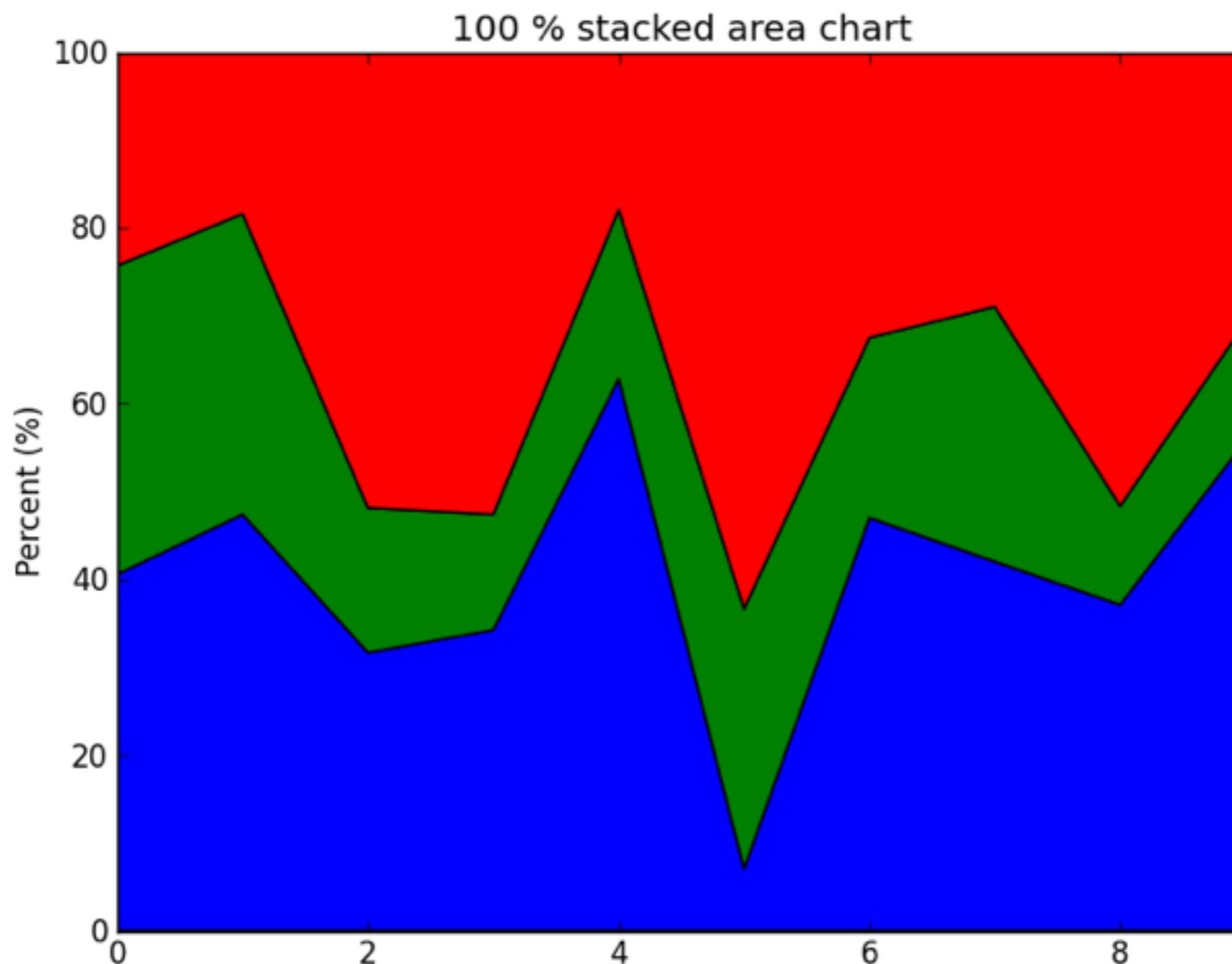
Video at <http://lineup.caleydo.org>

Stacked Area Chart

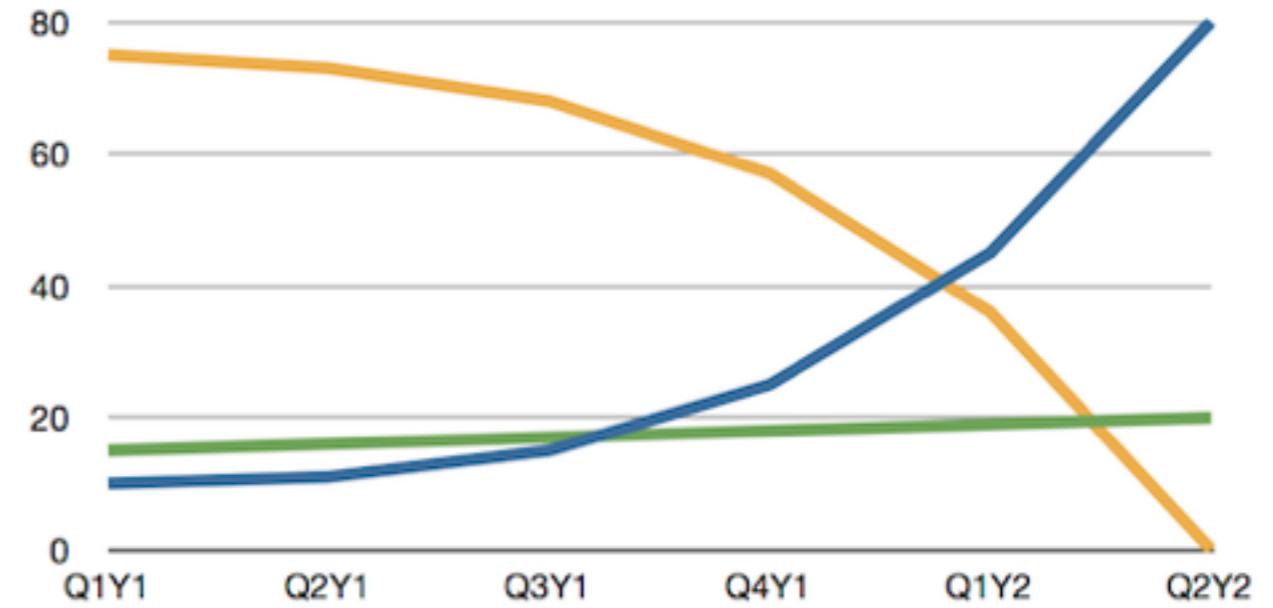
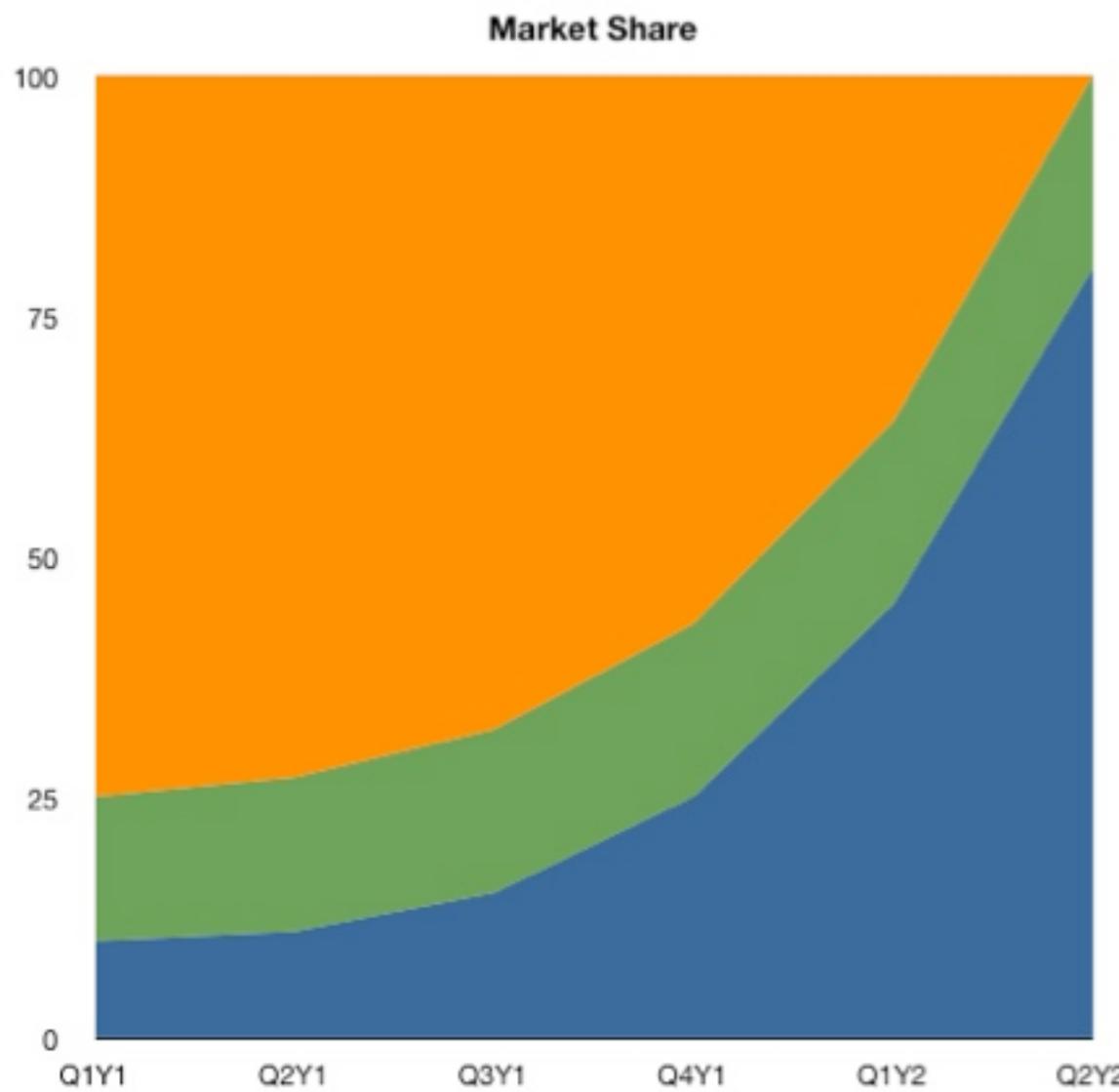


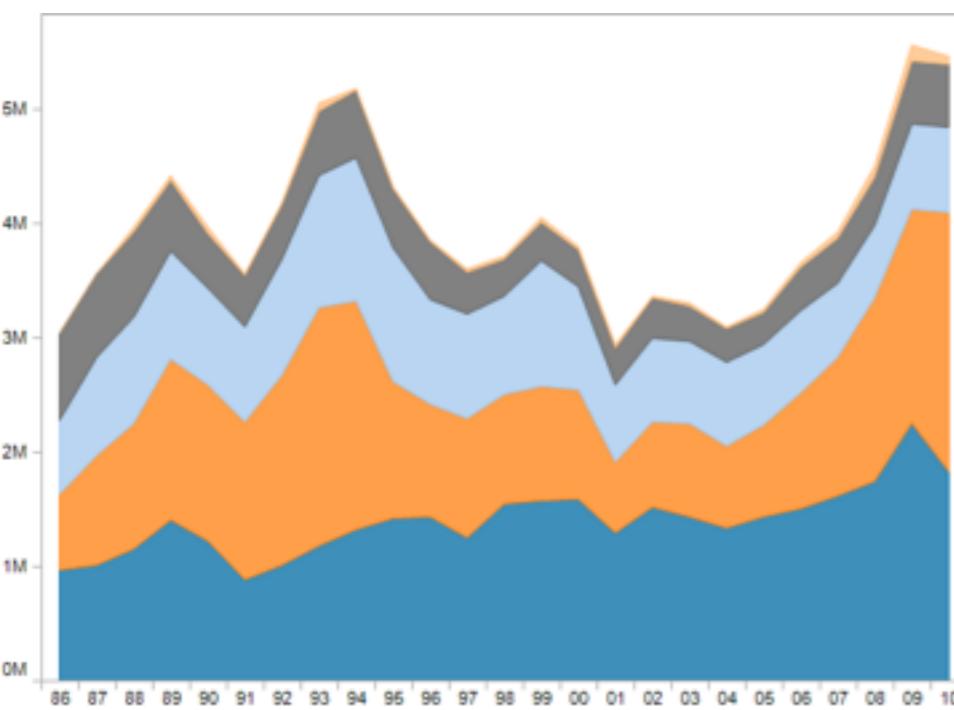
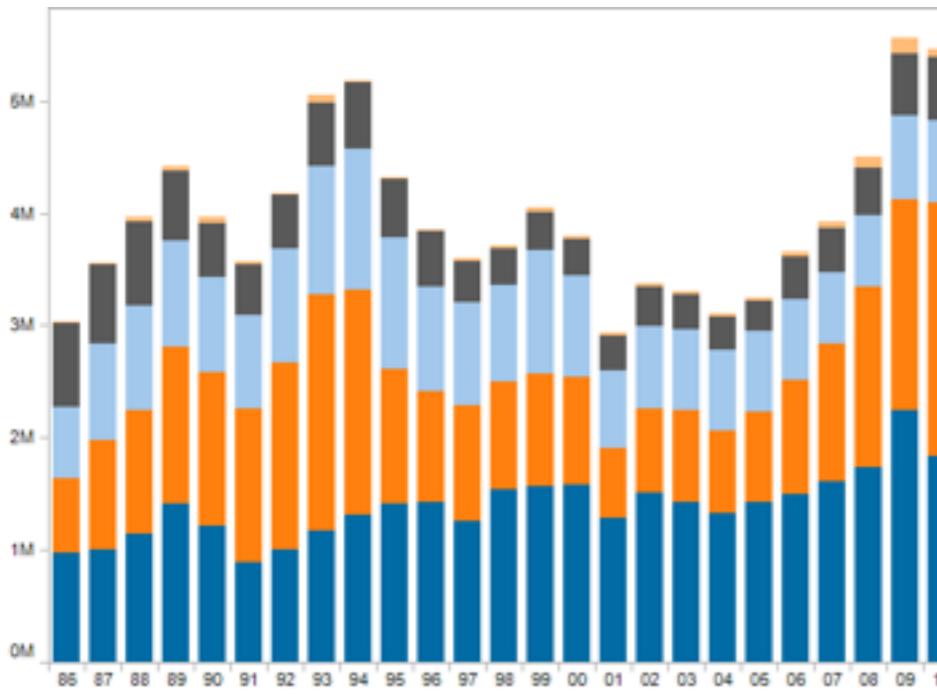
<http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib>

100% Stacked Area Chart



Stacked Area vs. Line Graphs



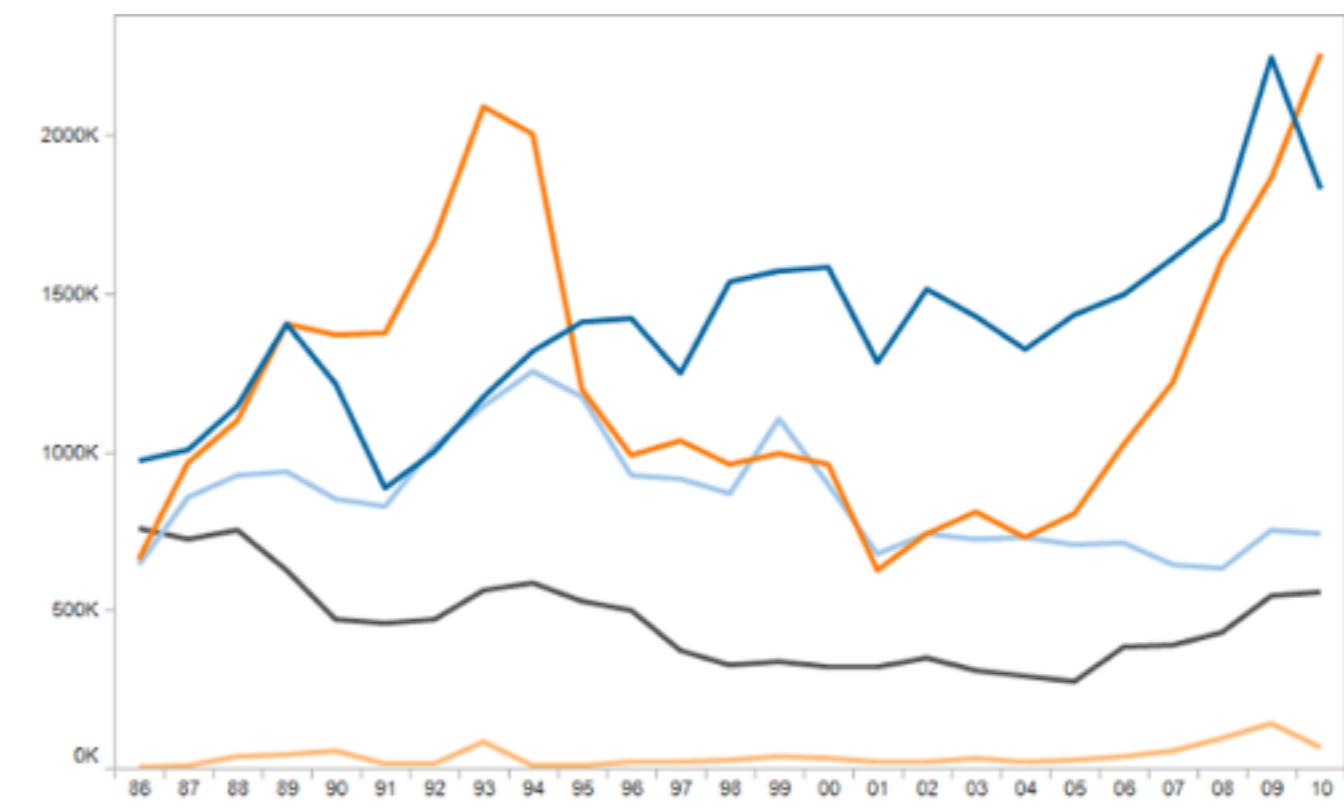


Weapon

- Misc
- Revolvers
- Shotguns
- Pistols
- Rifles

Weapon

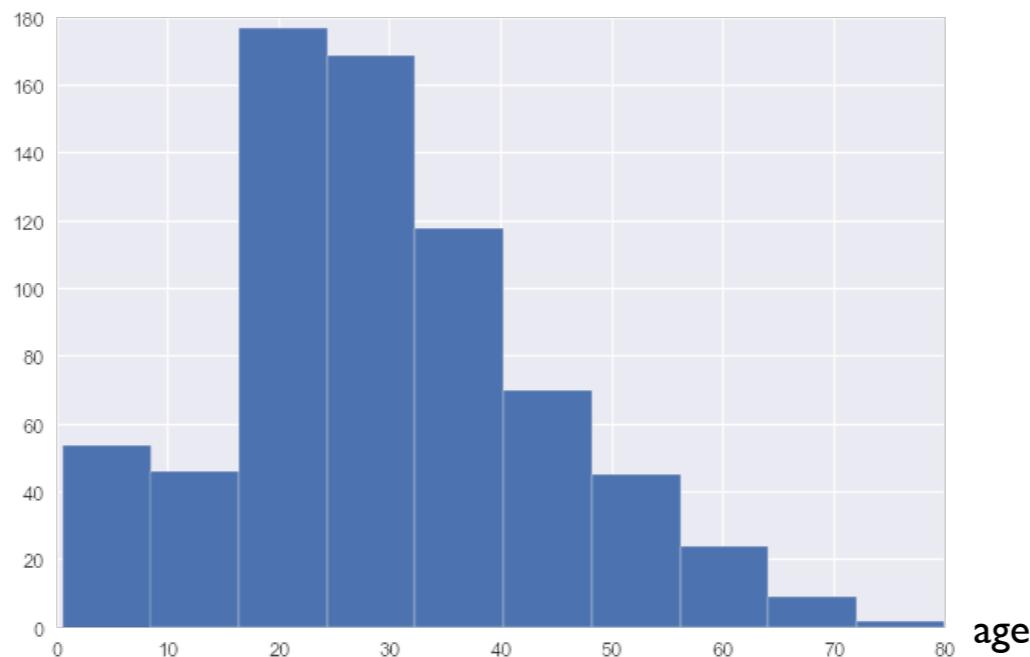
- Misc
- Revolvers
- Shotguns
- Pistols
- Rifles



Distributions

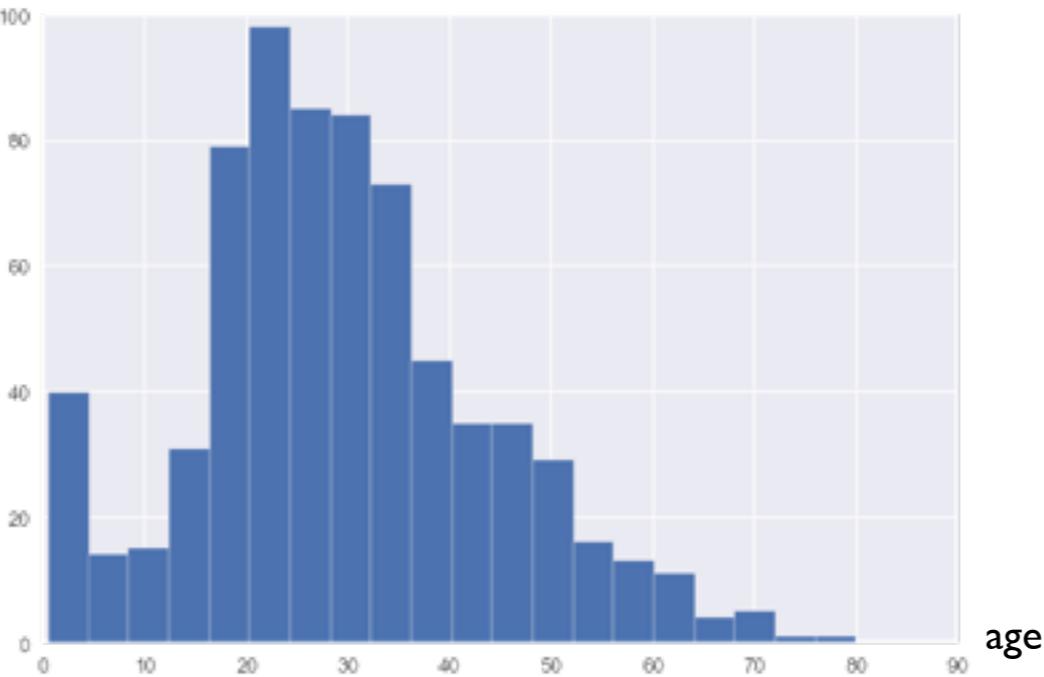
Histogram

passengers



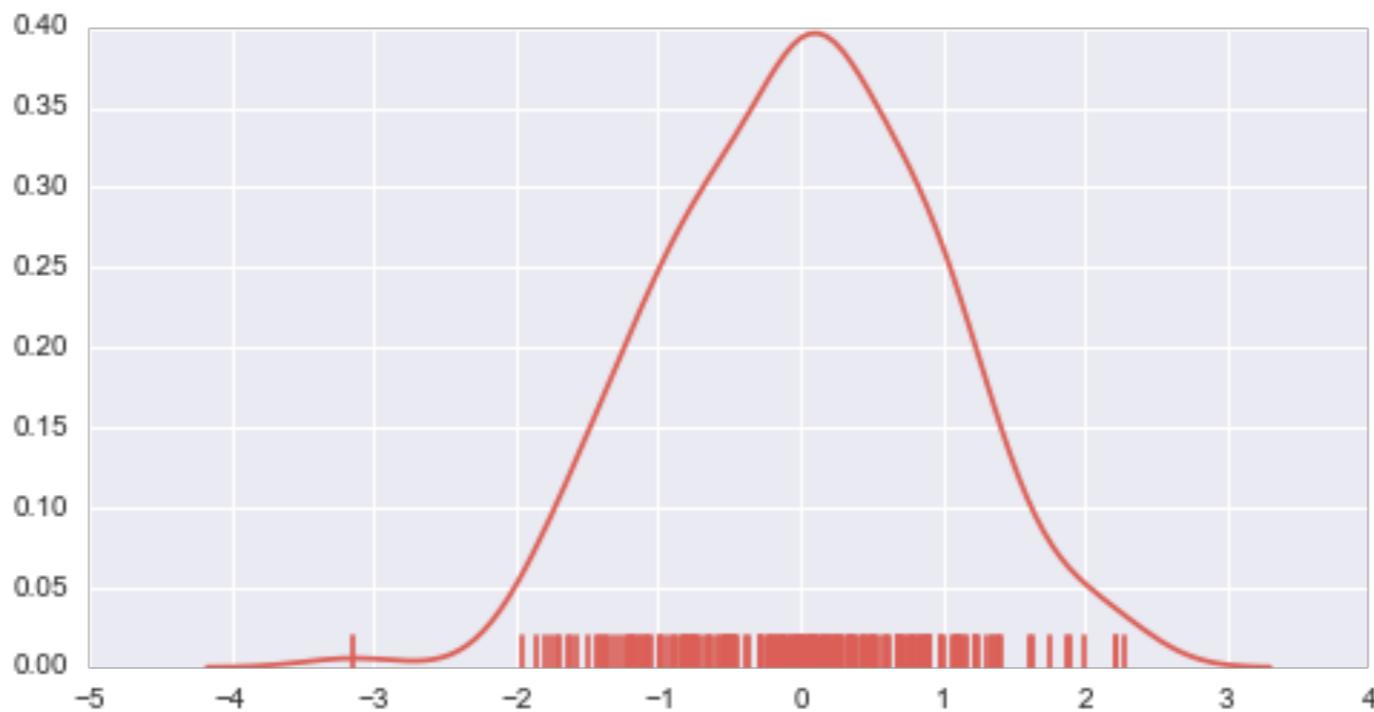
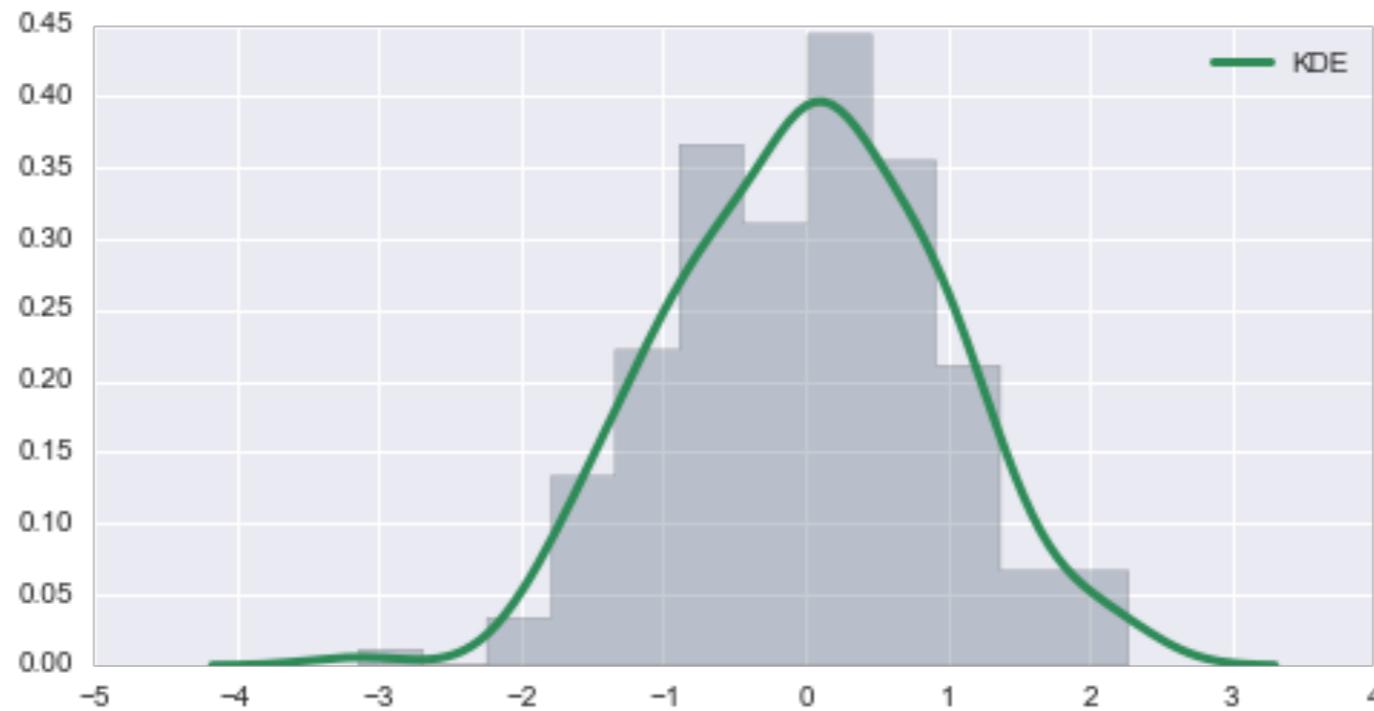
10 Bins

passengers

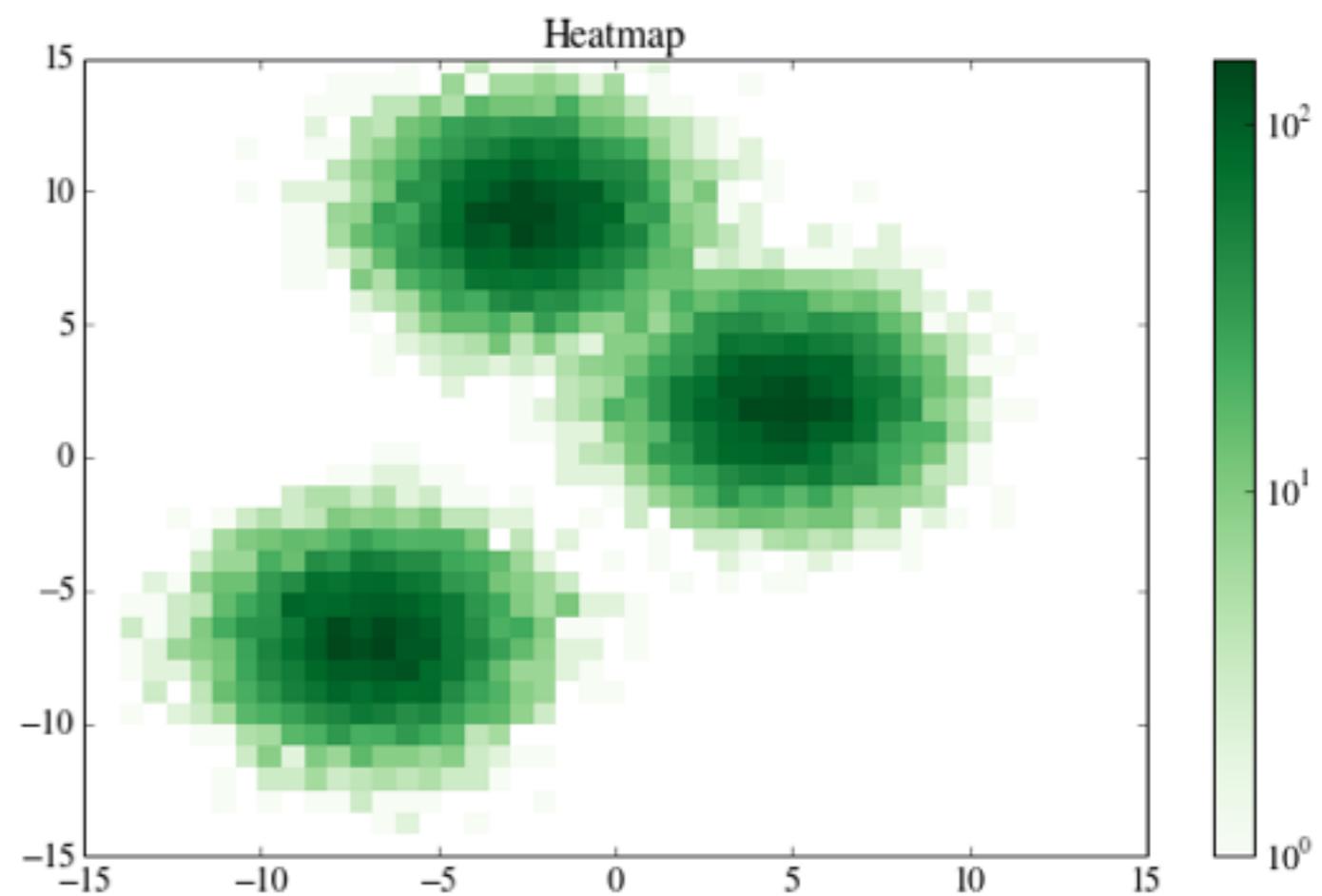
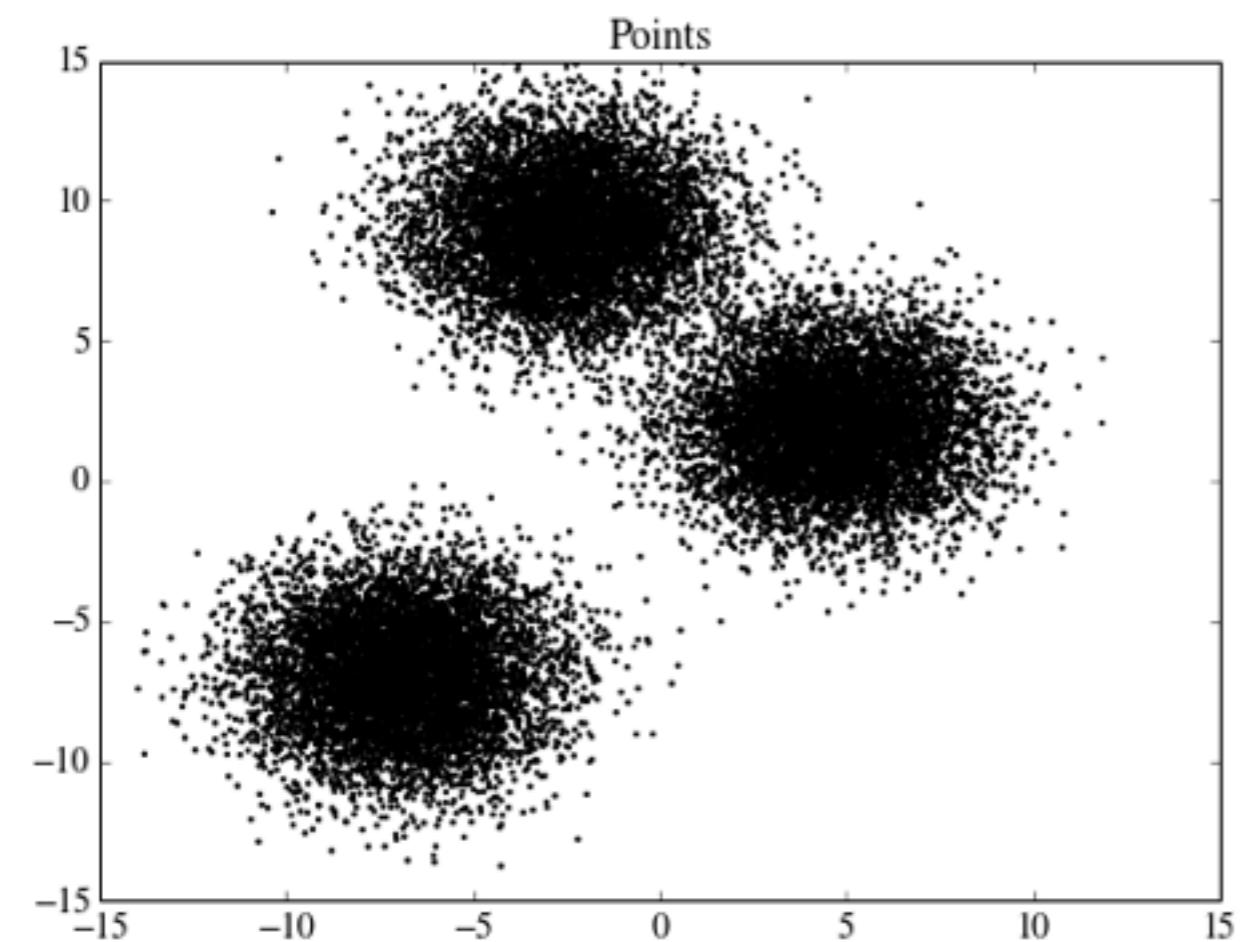
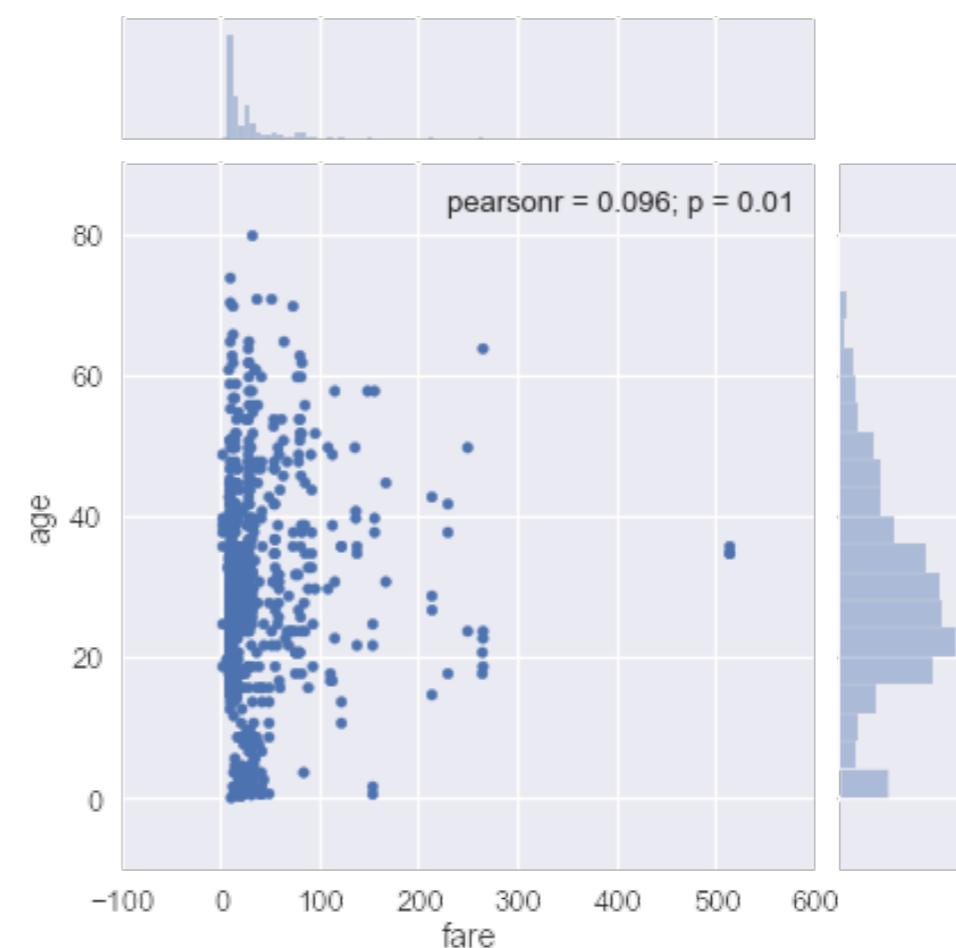


20 Bins

Density Plots



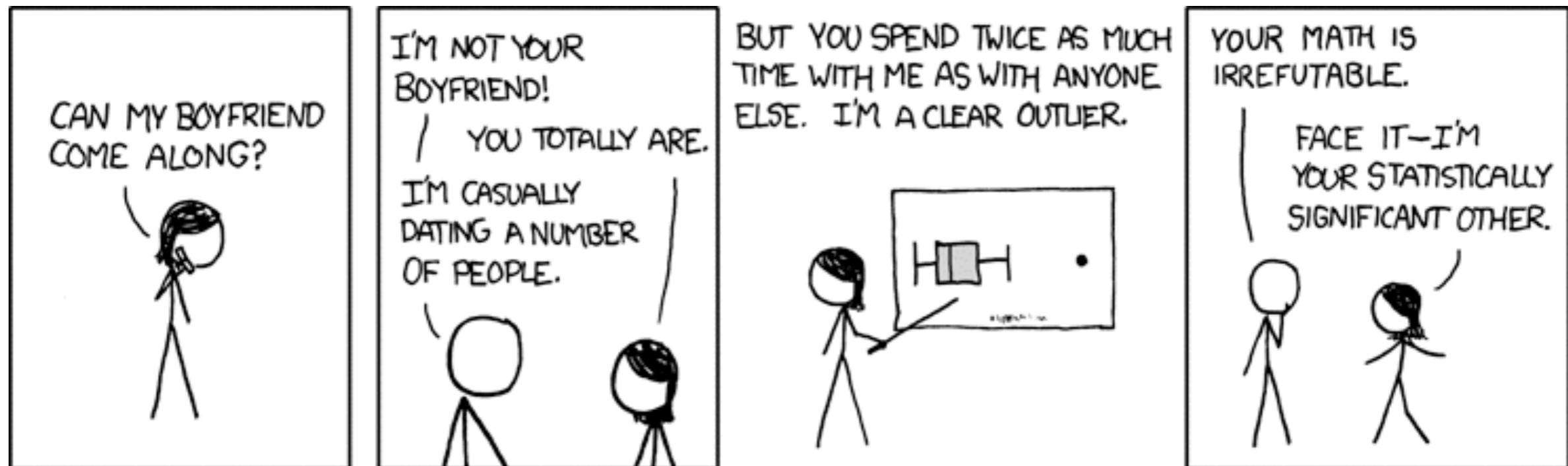
Heat Maps



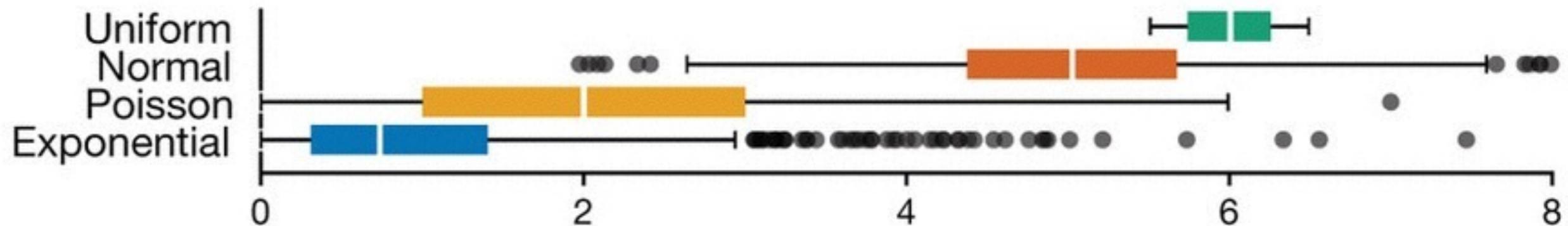
2D Density Plots

Box Plots

aka Box-and-Whisker Plot

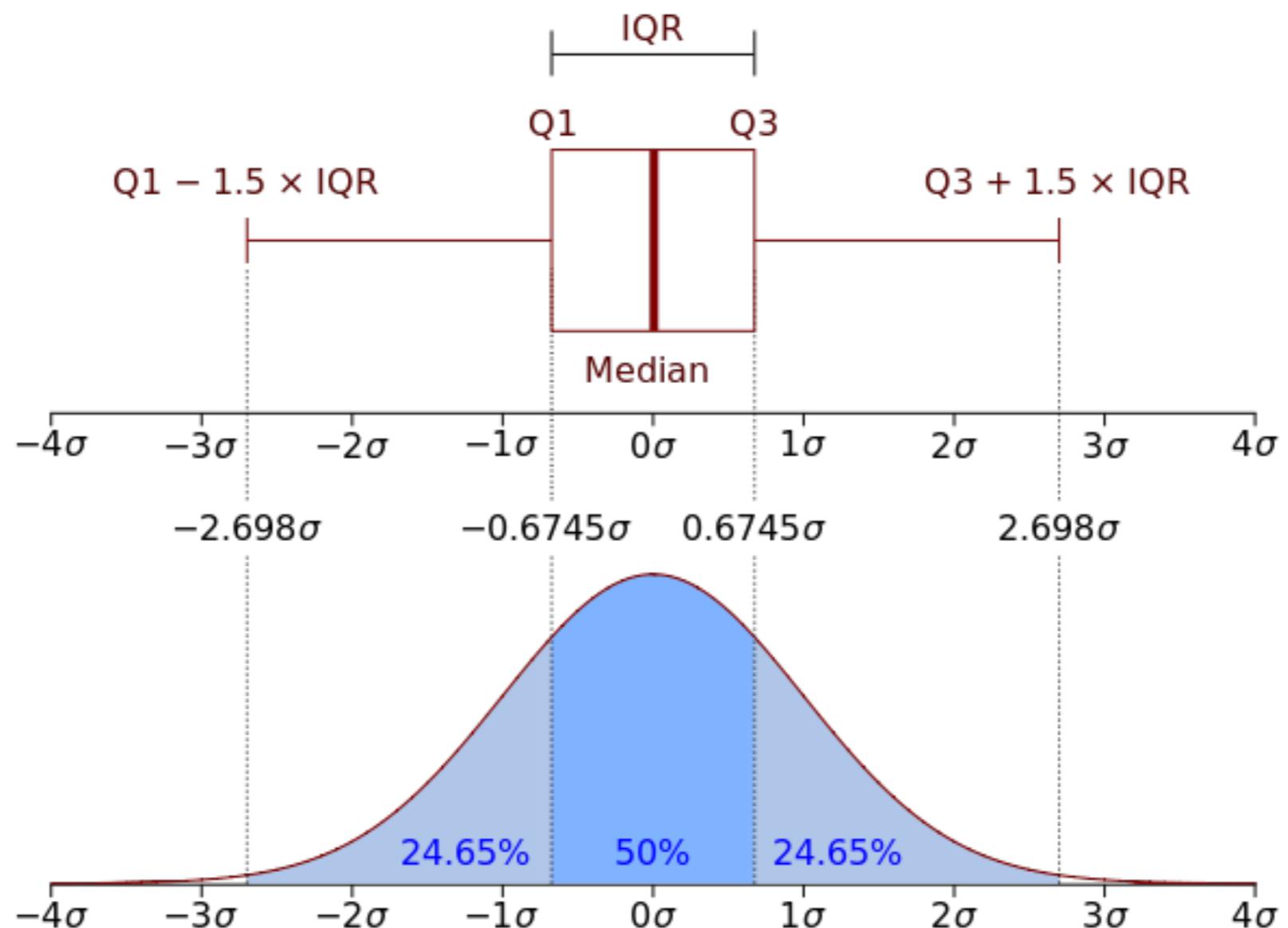


<http://xkcd.com/539/>



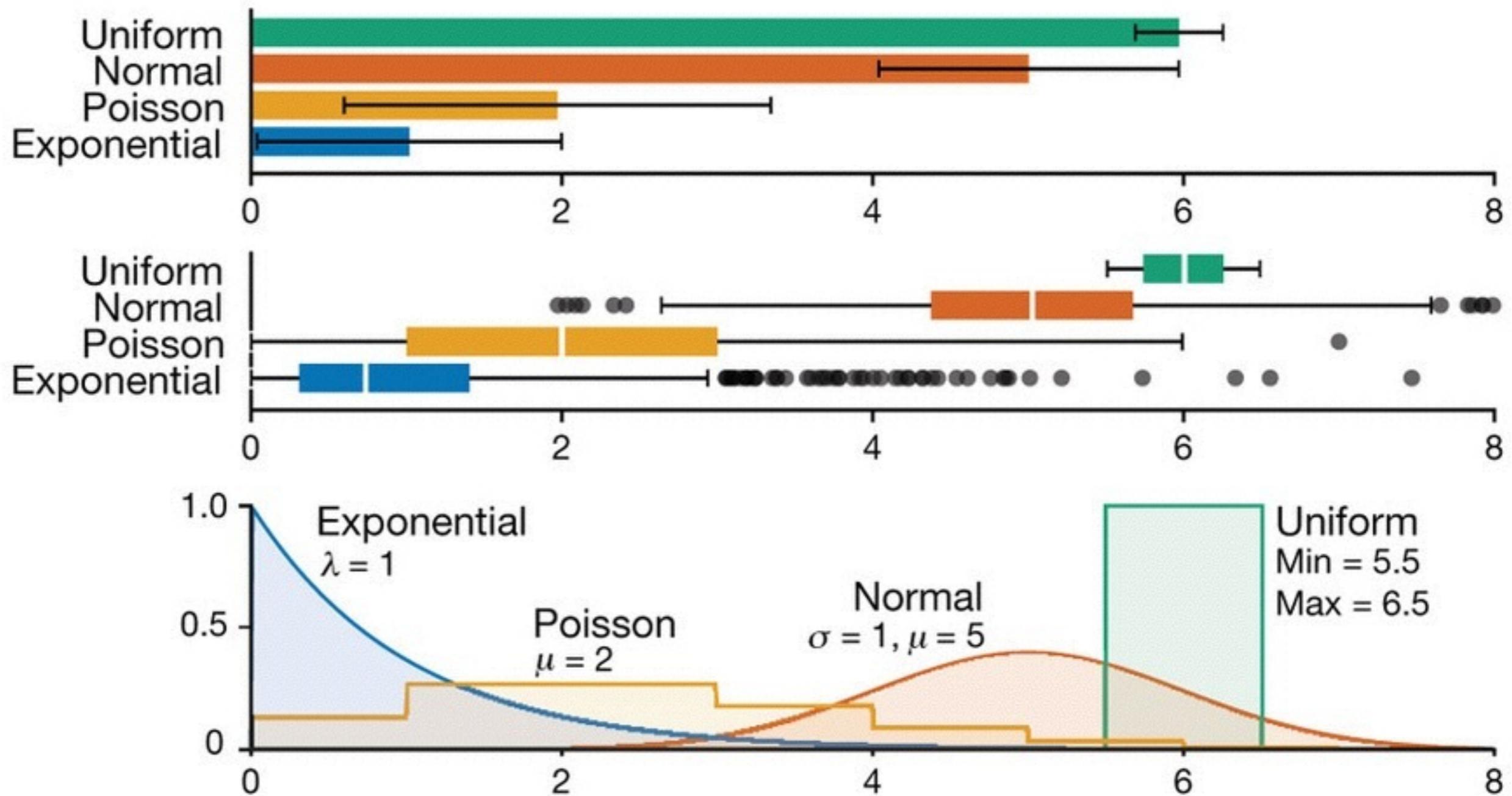
Box Plots

aka Box-and-Whisker Plot



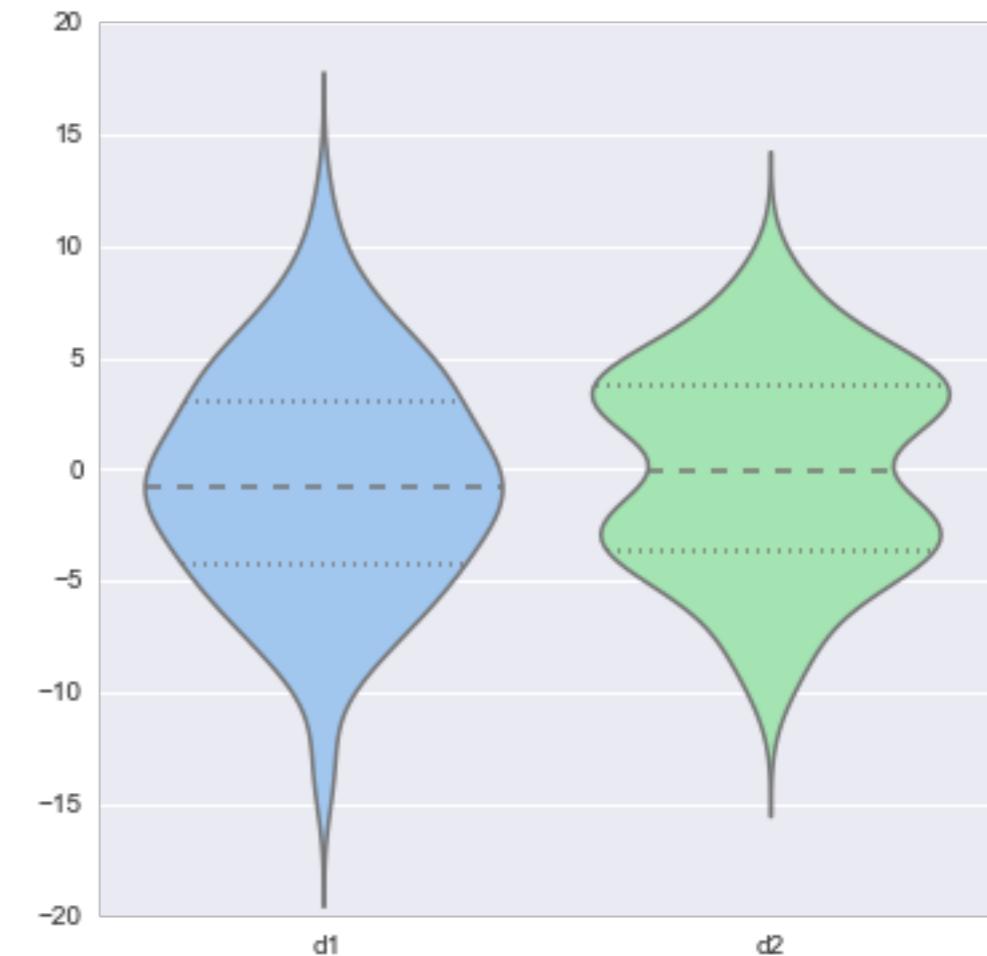
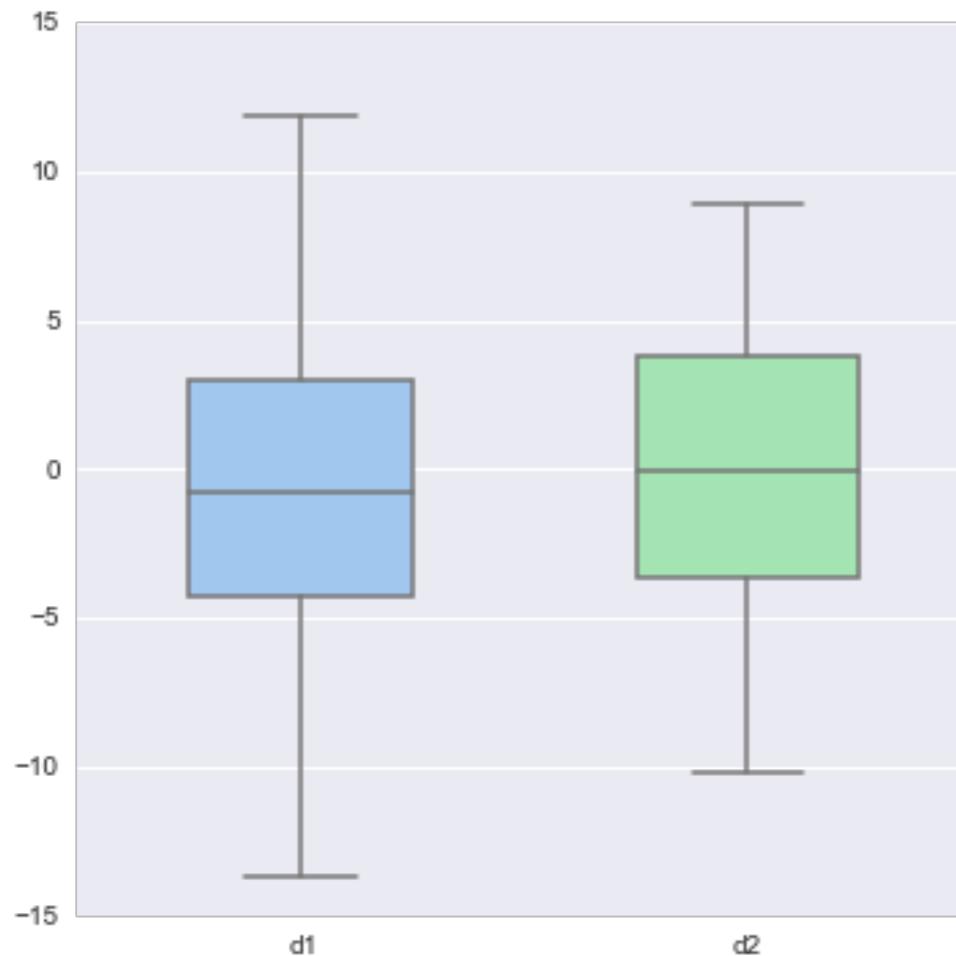
Wikipedia

Comparison

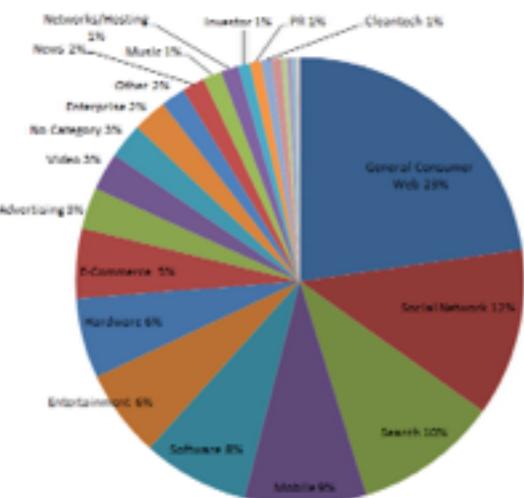


Violin Plot

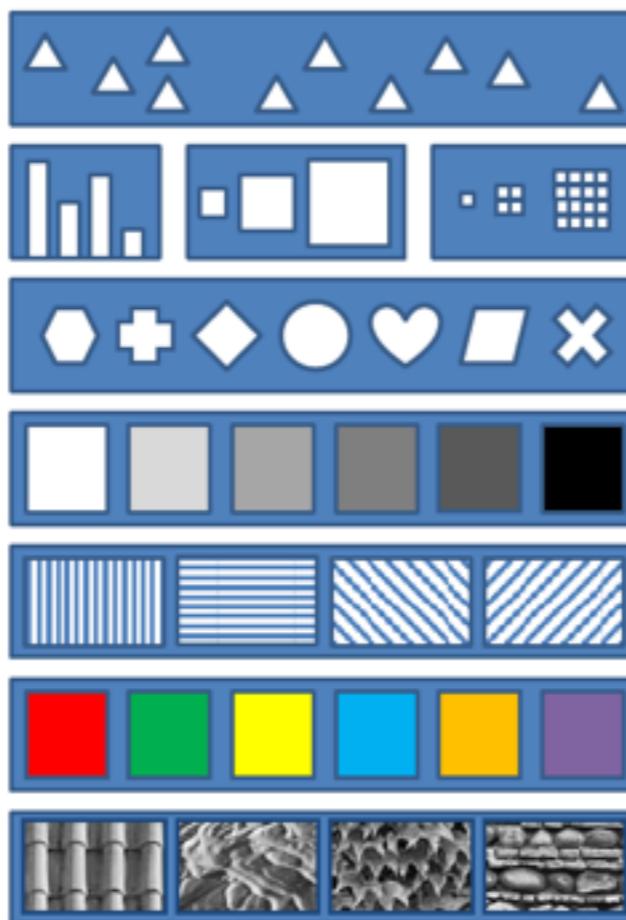
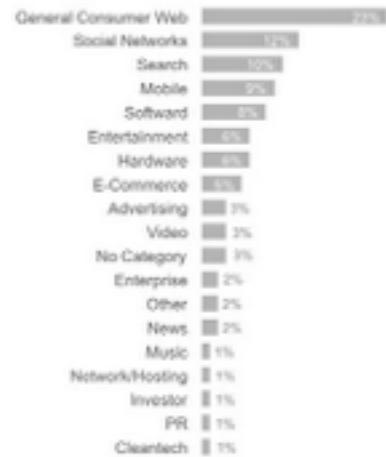
= Box Plot + Probability Density Function



On Thursday...

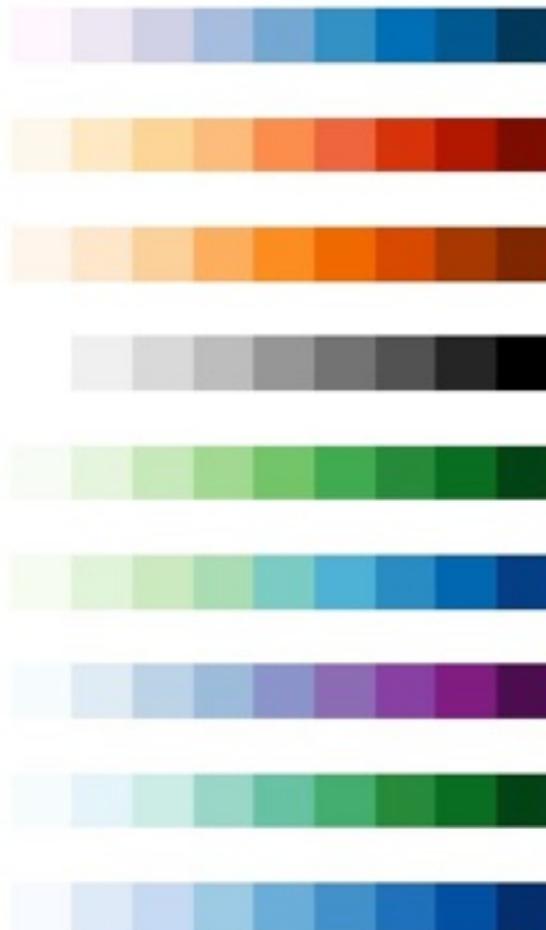


TechCrunch Coverage: 2005 - 2011
Bars are best!

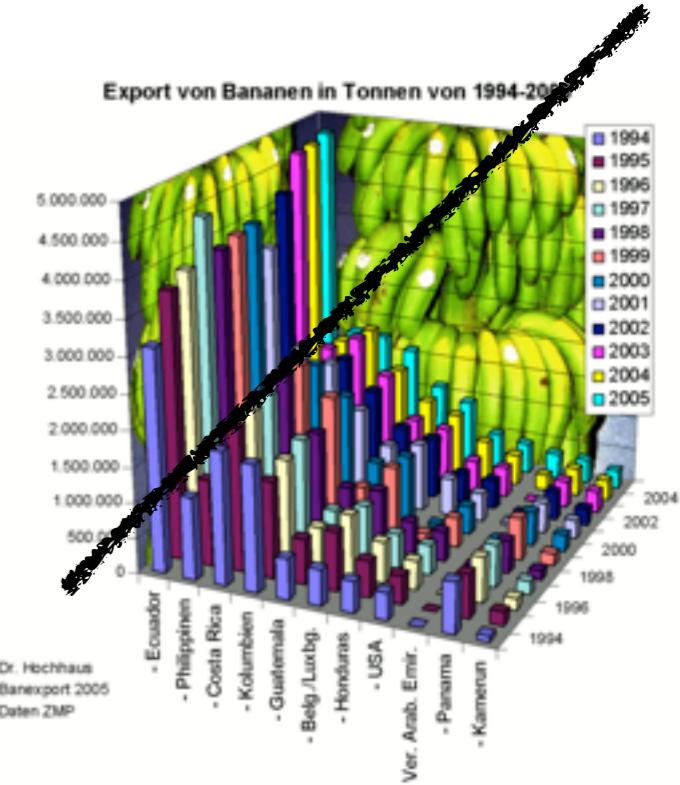


Effective vs.
Non Effective
Visualizations

Visual
Attributes



Color



Design
Principles

Questions

