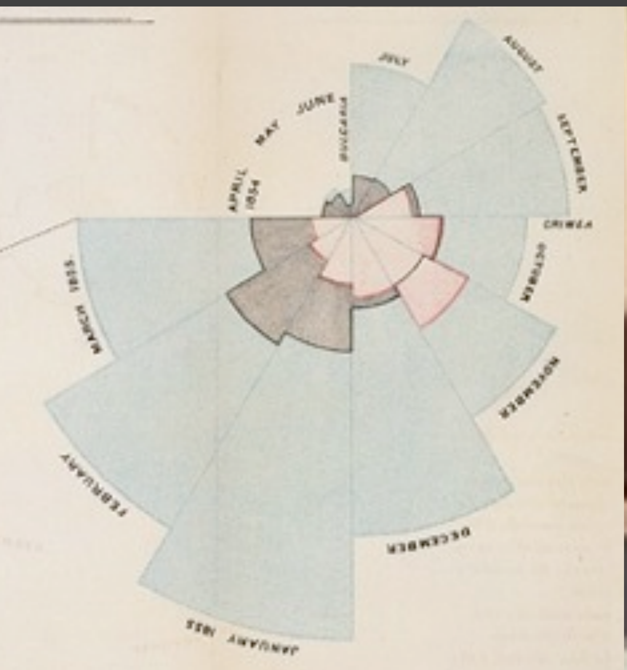
CSE512 :: 16 Jan 2014

Exploratory Data Analysis



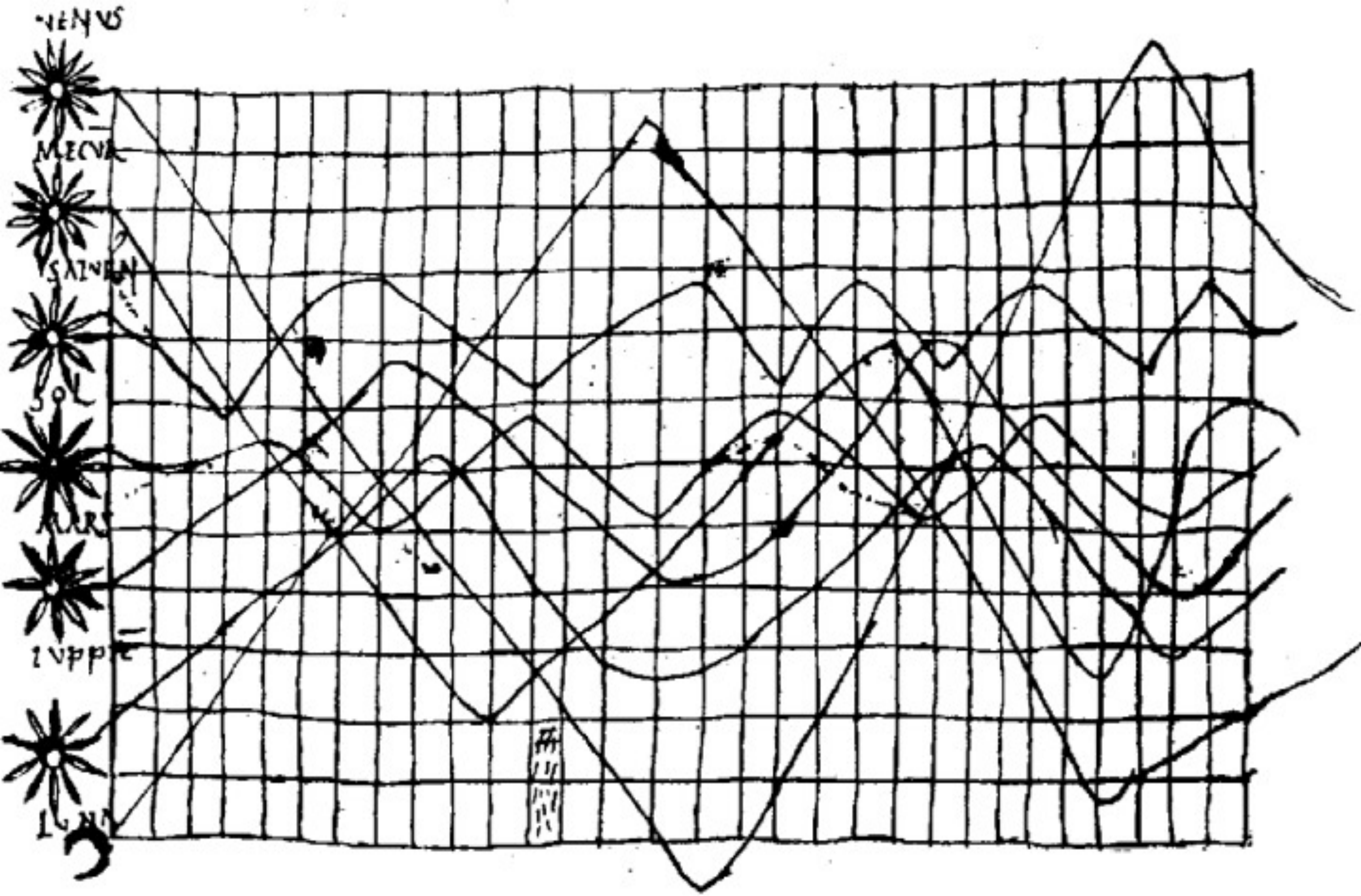
Jeffrey Heer University of Washington

What was the **first** data visualization?

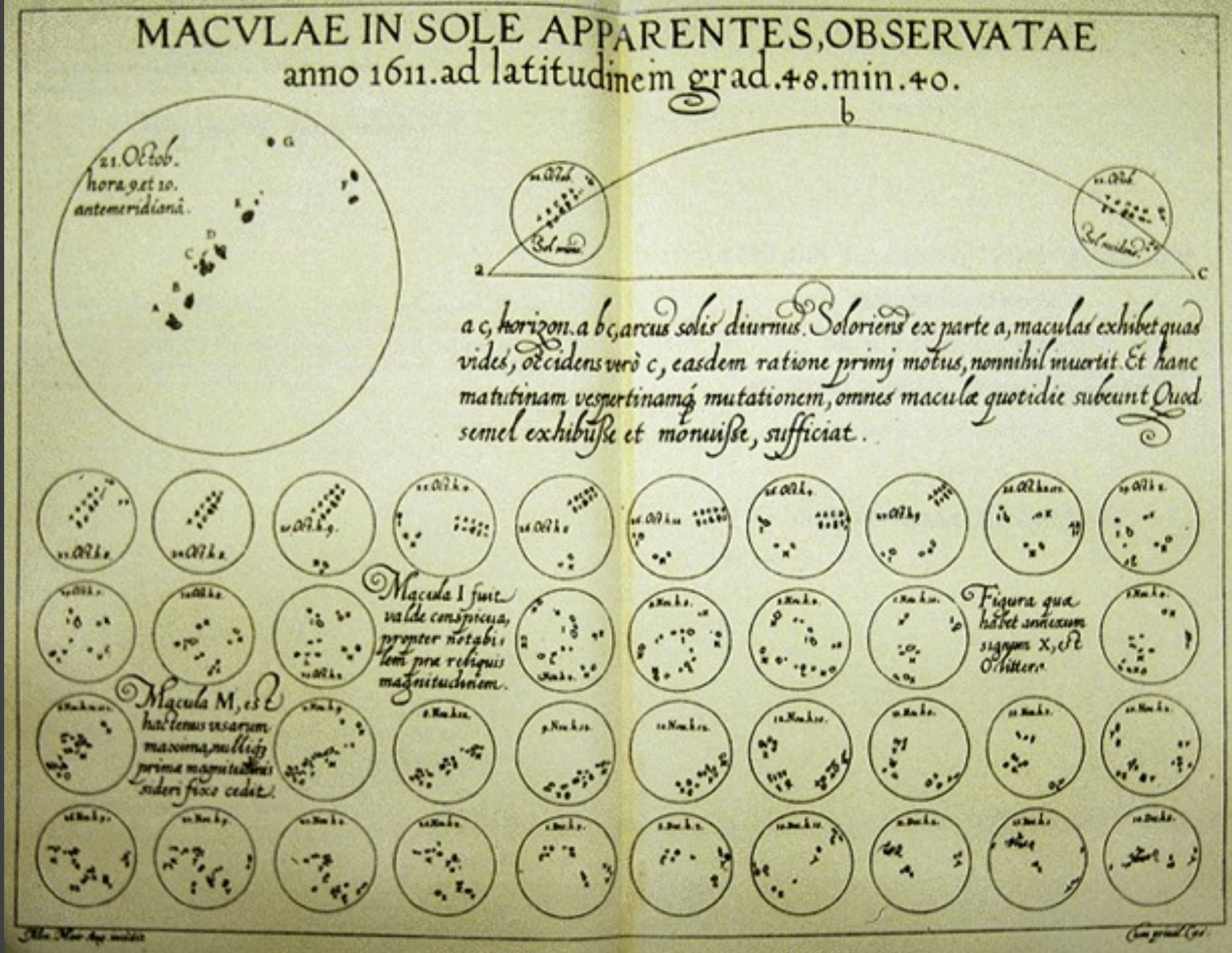
0 BC

~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey

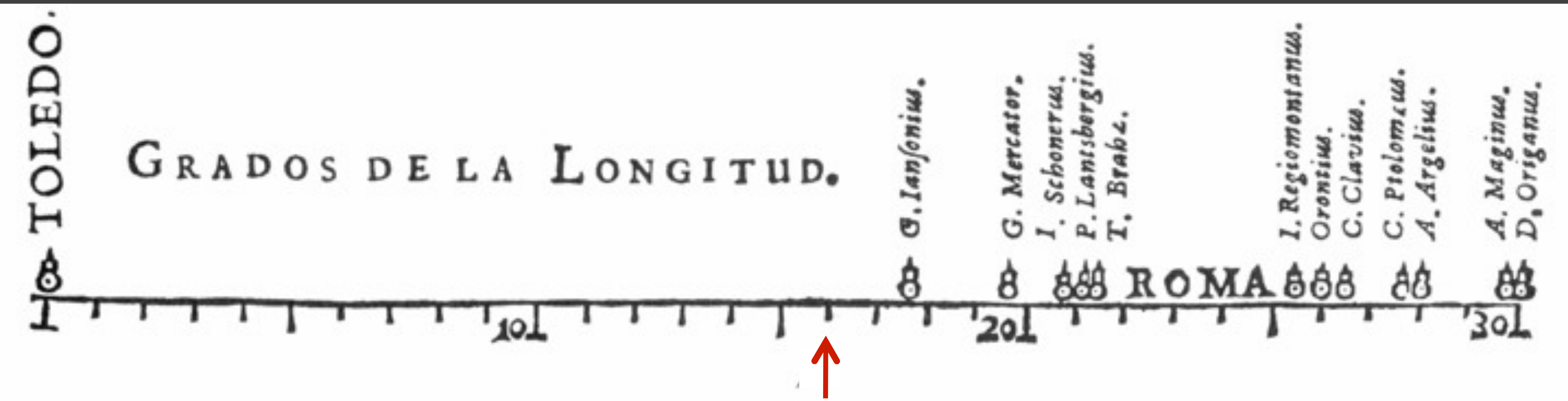
0 BC



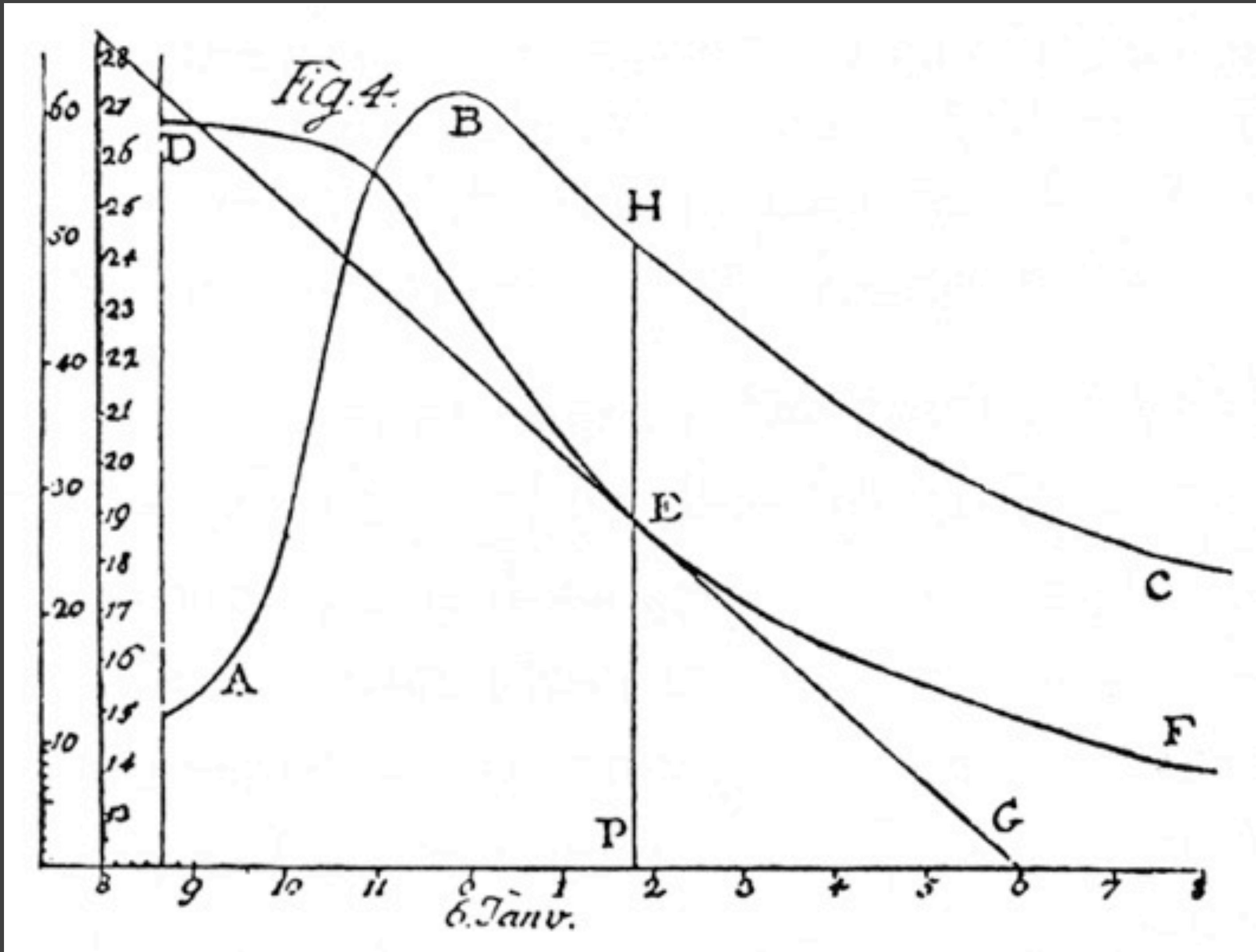
~950 AD Position of Sun, Moon and Planets



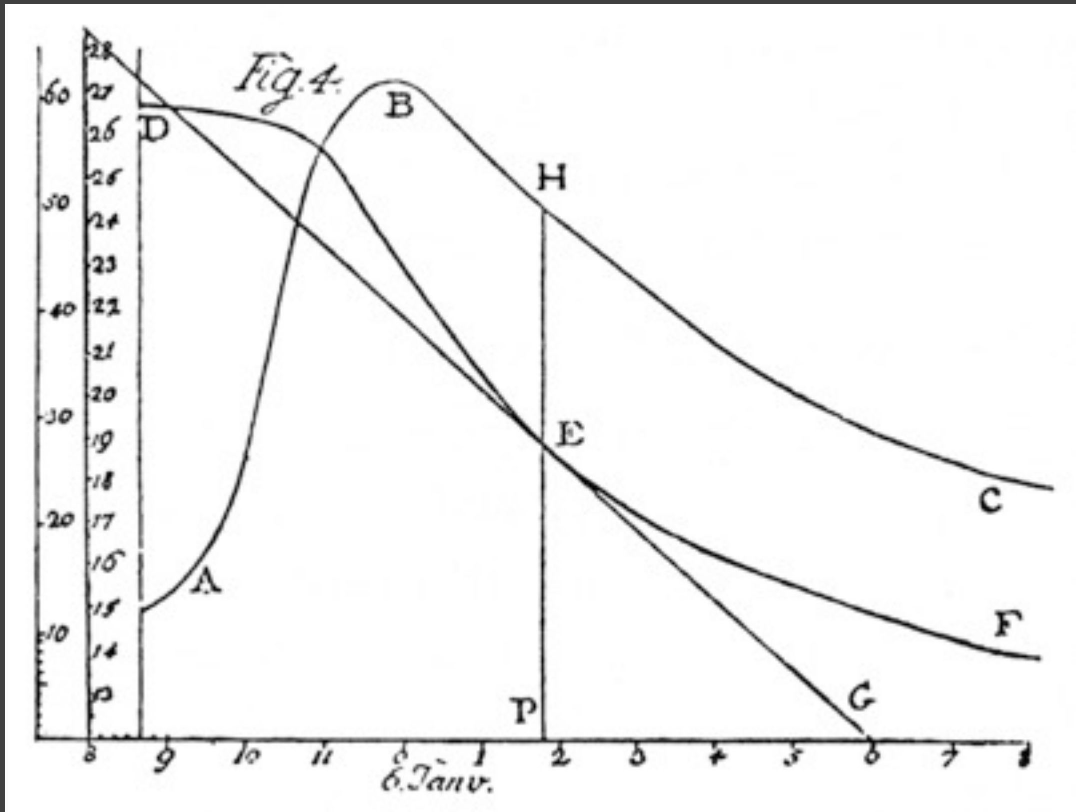
Sunspots over time, Scheiner 1626

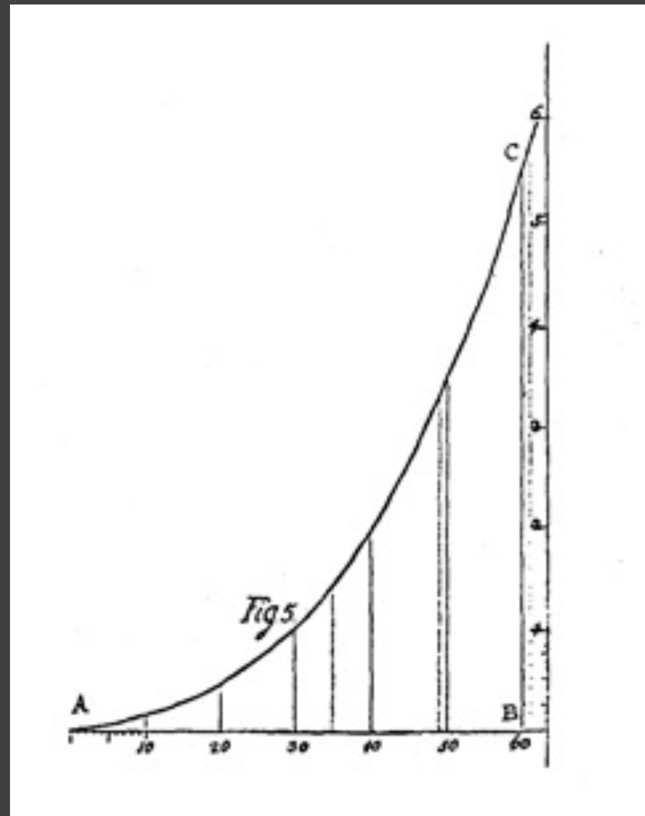


Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765

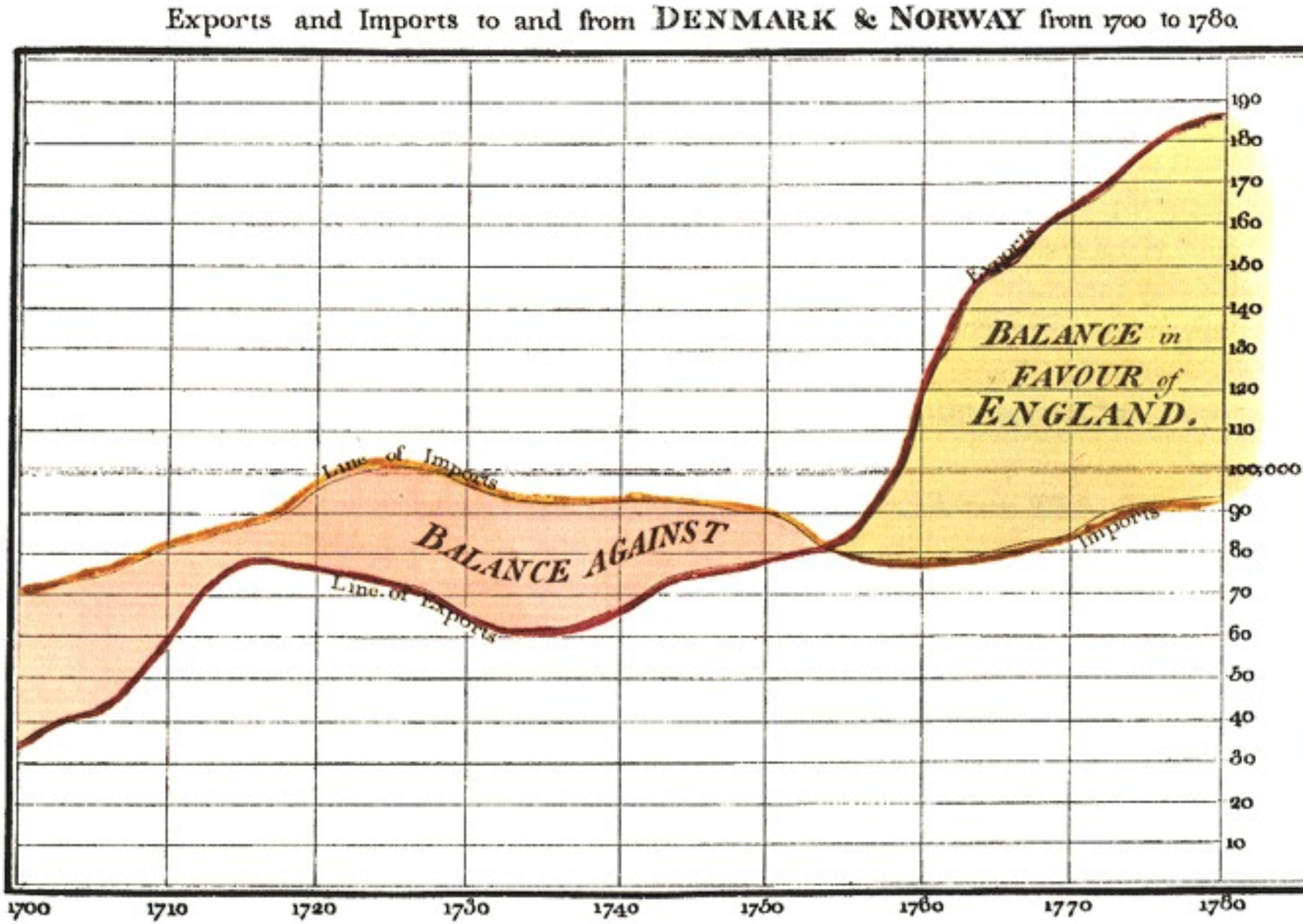




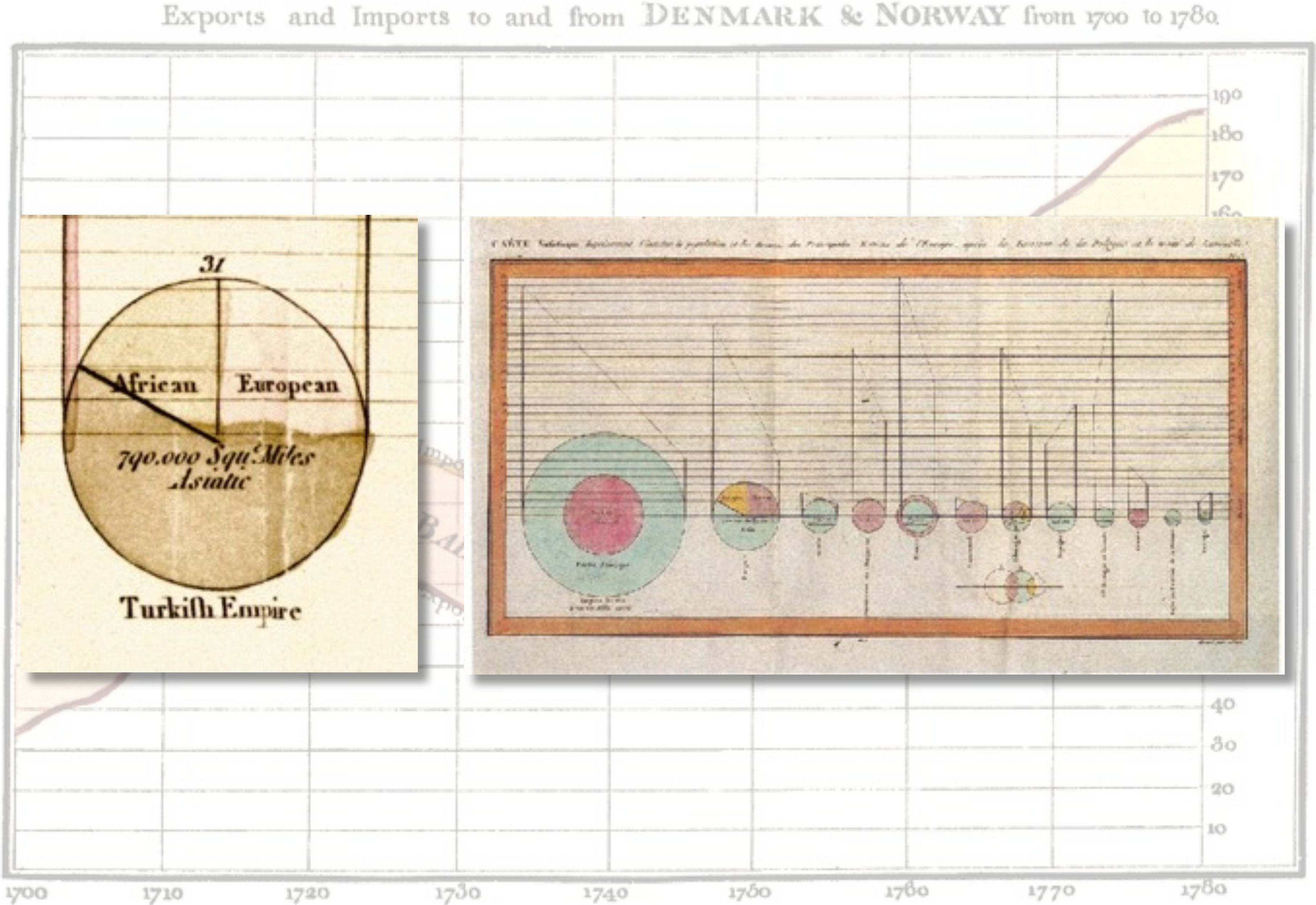
The Rate of Water Evaporation, Lambert 1765

The **Golden Age** of Data Visualization

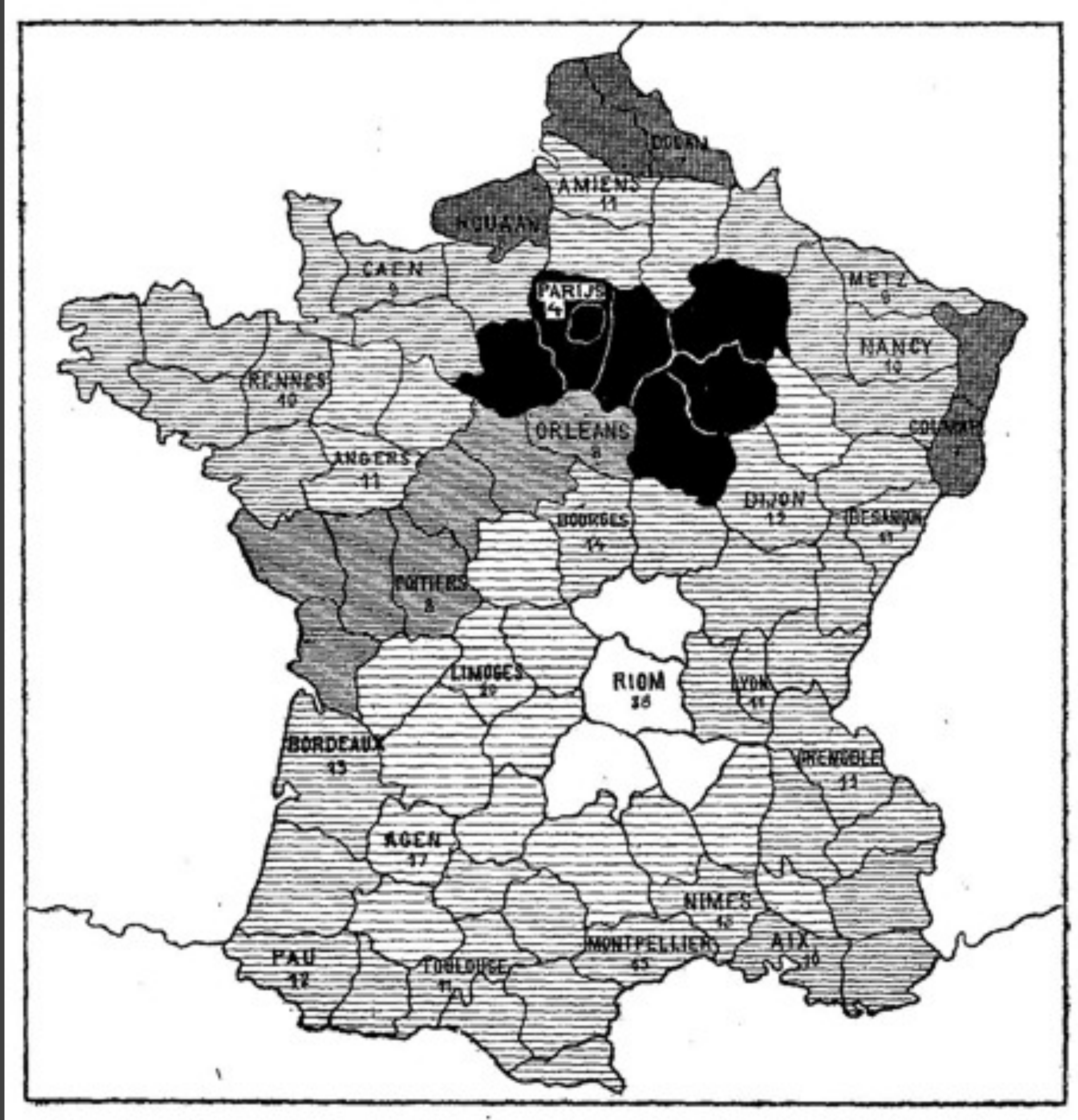
1786 1900



The Commercial and Political Atlas, William Playfair 1786

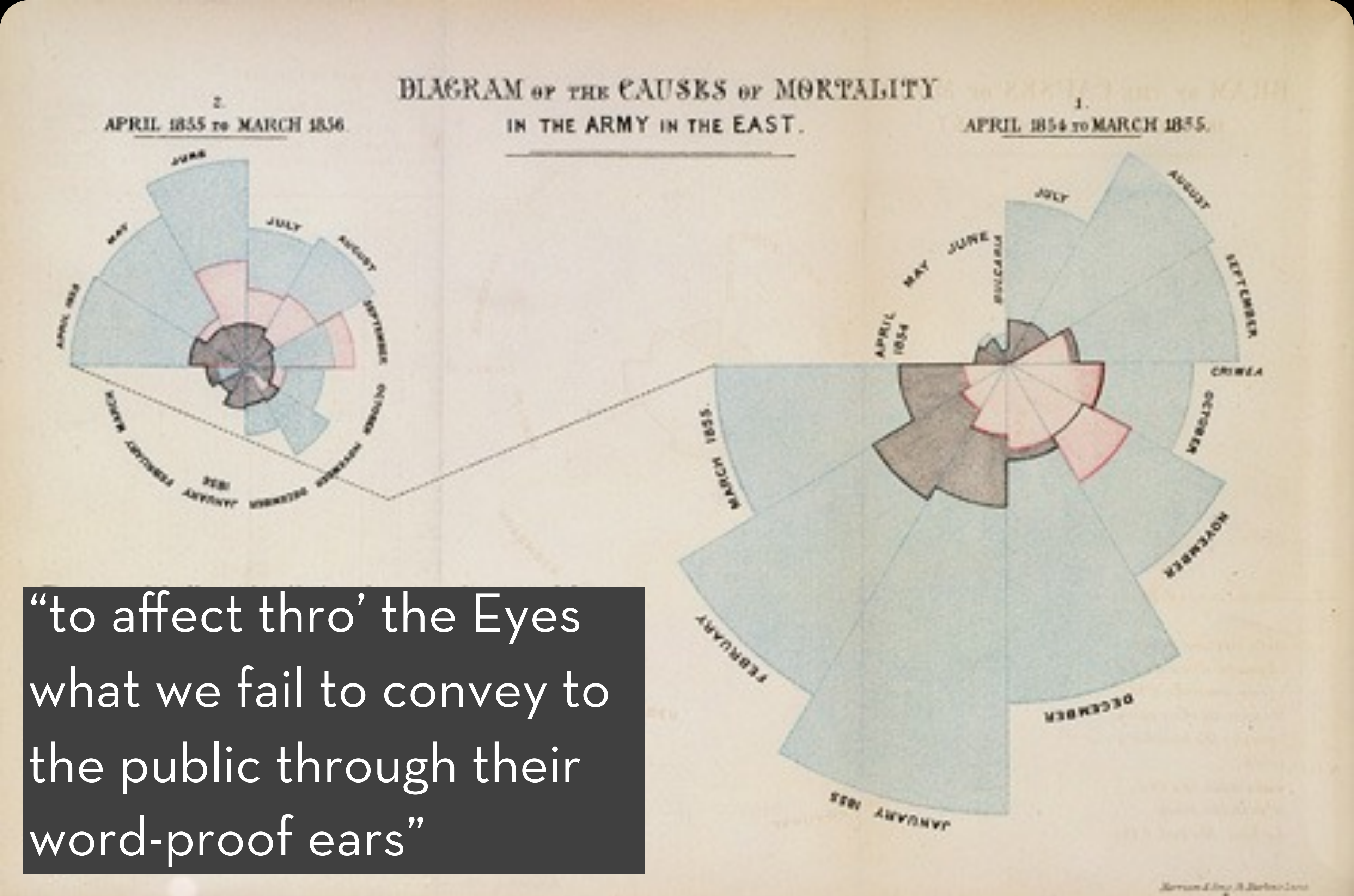


Statistical Breviary, William Playfair 1801



1786

1826(?) Illiteracy in France, Pierre Charles Dupin



“to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears”

1786

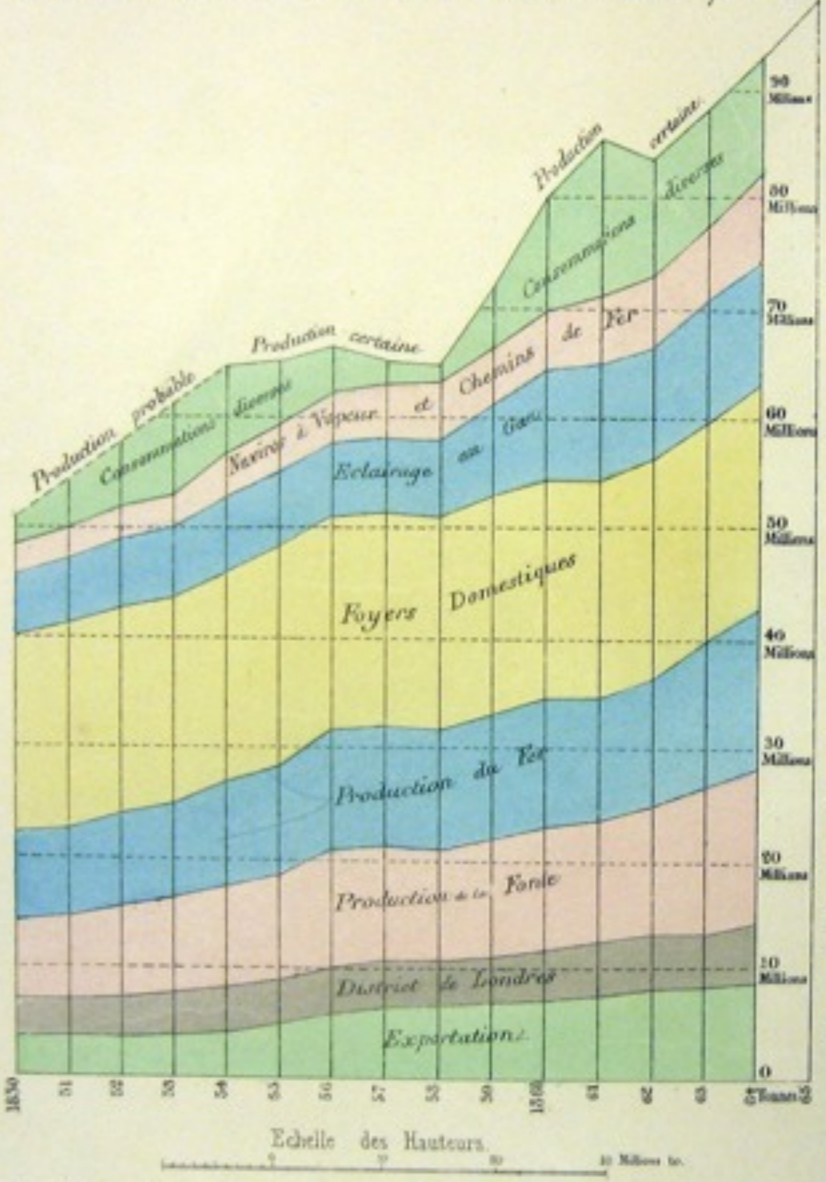
1856 “Coxcomb” of Crimean War Deaths, Florence Nightingale

1786

1864 British Coal Exports, Charles Minard

Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.

Les abscisses représentent les années et les ordonnées les quantités annuelles de houille consommée.
Les couleurs indiquent les espèces de consommations. Les longueurs d'ordonnées comprises dans une couleur sont les quantités de houille consommées à raison de deux millimètres pour un million de tonnes.



Données *admises pour former le* Tableau *ci-contre.*

Consommations. —— Sources des Renseignements.

Exportations. — *Mineral statistics 1865 page 214 et Renseignements Parlementaires.*

District de Londres. —— *id.* ———— *page 213*

Produits de la Fonte. —— *id* ———— *page 215 et pour les années avant 1855 calculée à raison de 3$^{t.}$ de houille pour 1$^{t.}$ de fonte, en admettant les quantités annuelles de fonte du* Coal question *page 192.*

Production du fer — *Mineral statistics — page 215 et pour les années avant 1855 — calculée à raison de 3$^{t.}$35 de houille pour 1 tonne de fonte convertie en fer, et admettant $\frac{8}{10}$^{es} de la fonte produite convertis en fer*

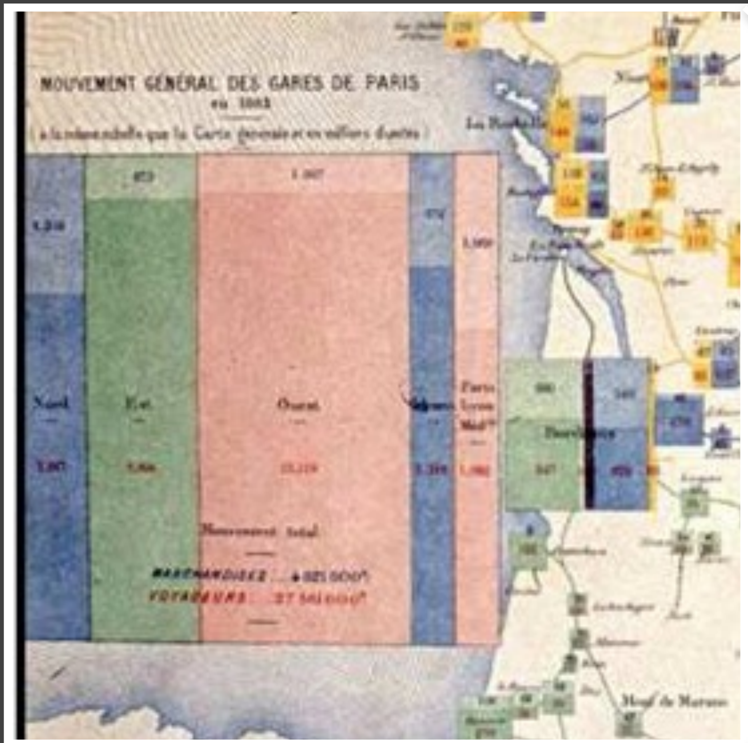
Foyers domestiques : — *En y comprenant les petites manufactures.*
On l'estimait en 1848 à 19 millions de tonnes, (A) *qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.*

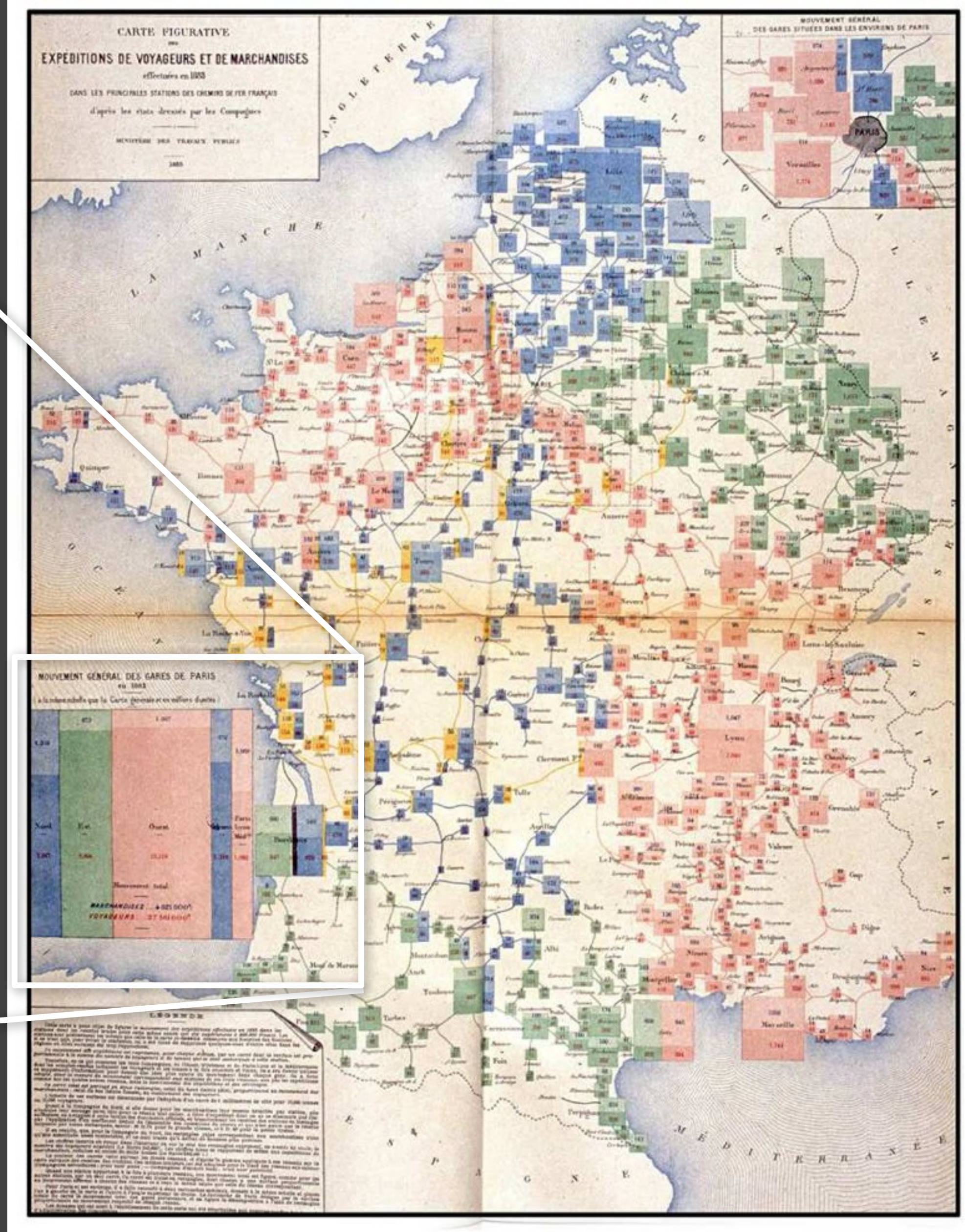
Eclairage au Gaz. — *Consommation estimée généralement du $\frac{1}{9}$^{e} au $\frac{1}{8}$^{e} de la production totale.*

Exploitation des Chemins de Fer. — *En supposant pour consommation totale 10^{K} par Kilomètre parcouru par les trains d'après les renseignements parlementaires.*

Navigation à vapeur. — *Calculée à raison de 5^{K} houille par cheval vapeur et par heure, le nombre de chevaux étant celui du Steam Vessels pour* 1864, *et les steamers étant supposés marcher la moitié de l'année;*
Avant 1864 j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of trade.

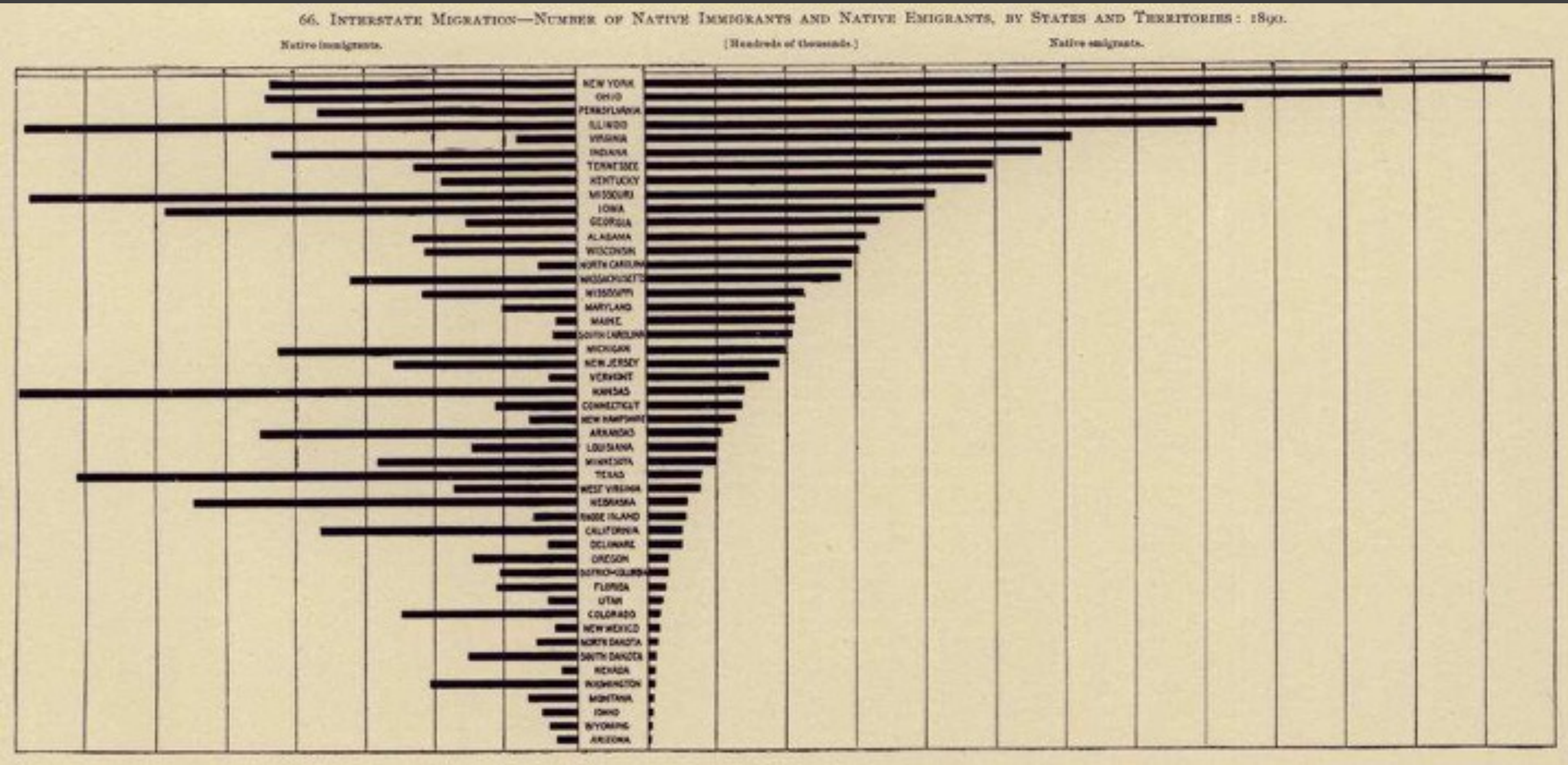
(A) *Voir l'excellent article houille de M.r Lamé Fleury, Dictionnaire du Commerce Page 111.*





1786

1884 Rail Passengers and Freight from Paris



1786

1890 Statistical Atlas of the Eleventh U.S. Census

The Rise of Statistics

Rise of **formal methods** in statistics and social science — Fisher, Pearson, ...

Little innovation in graphical methods

A period of **application and popularization**

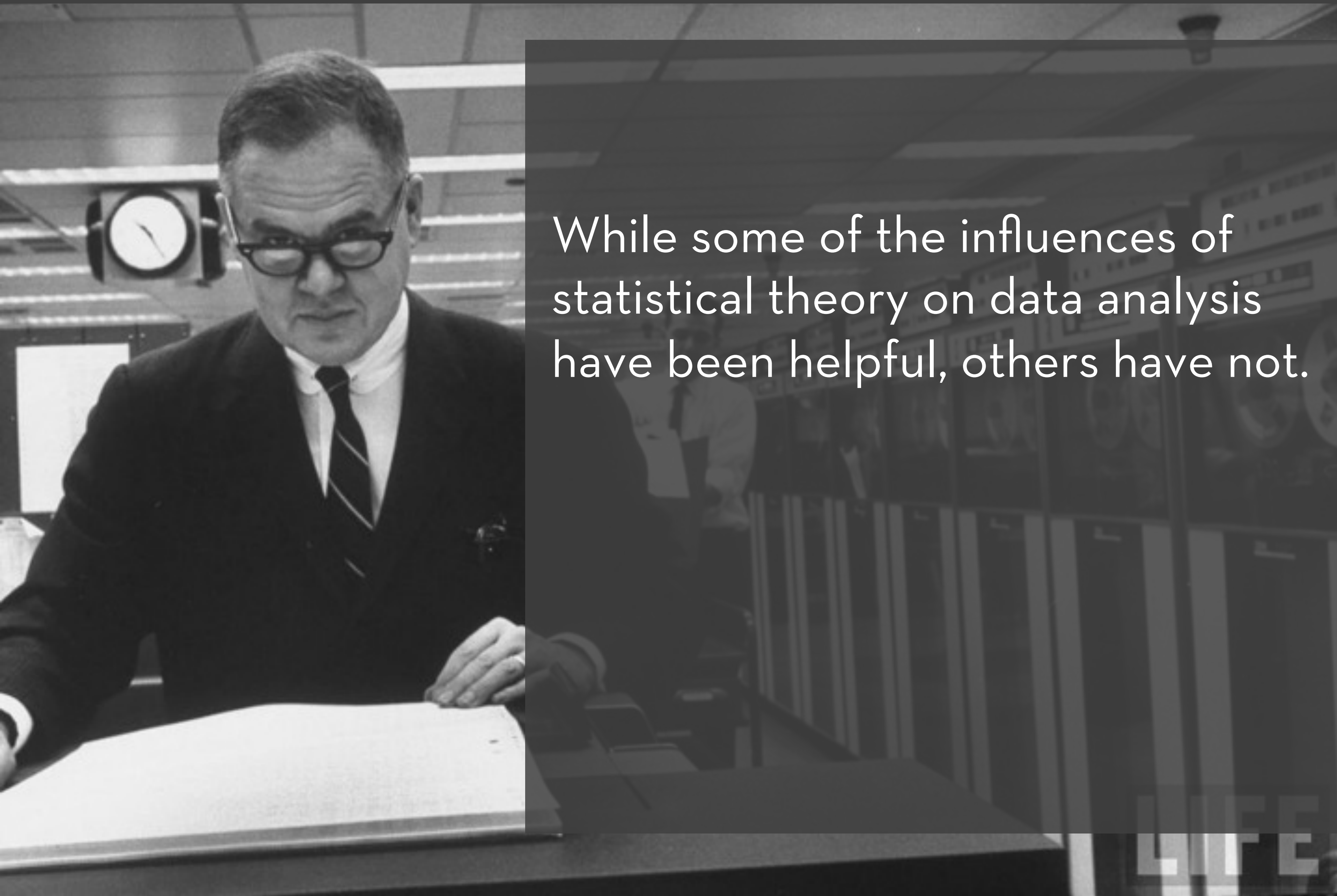
Graphical methods enter textbooks, curricula, and **mainstream use**

1786 1900 1950

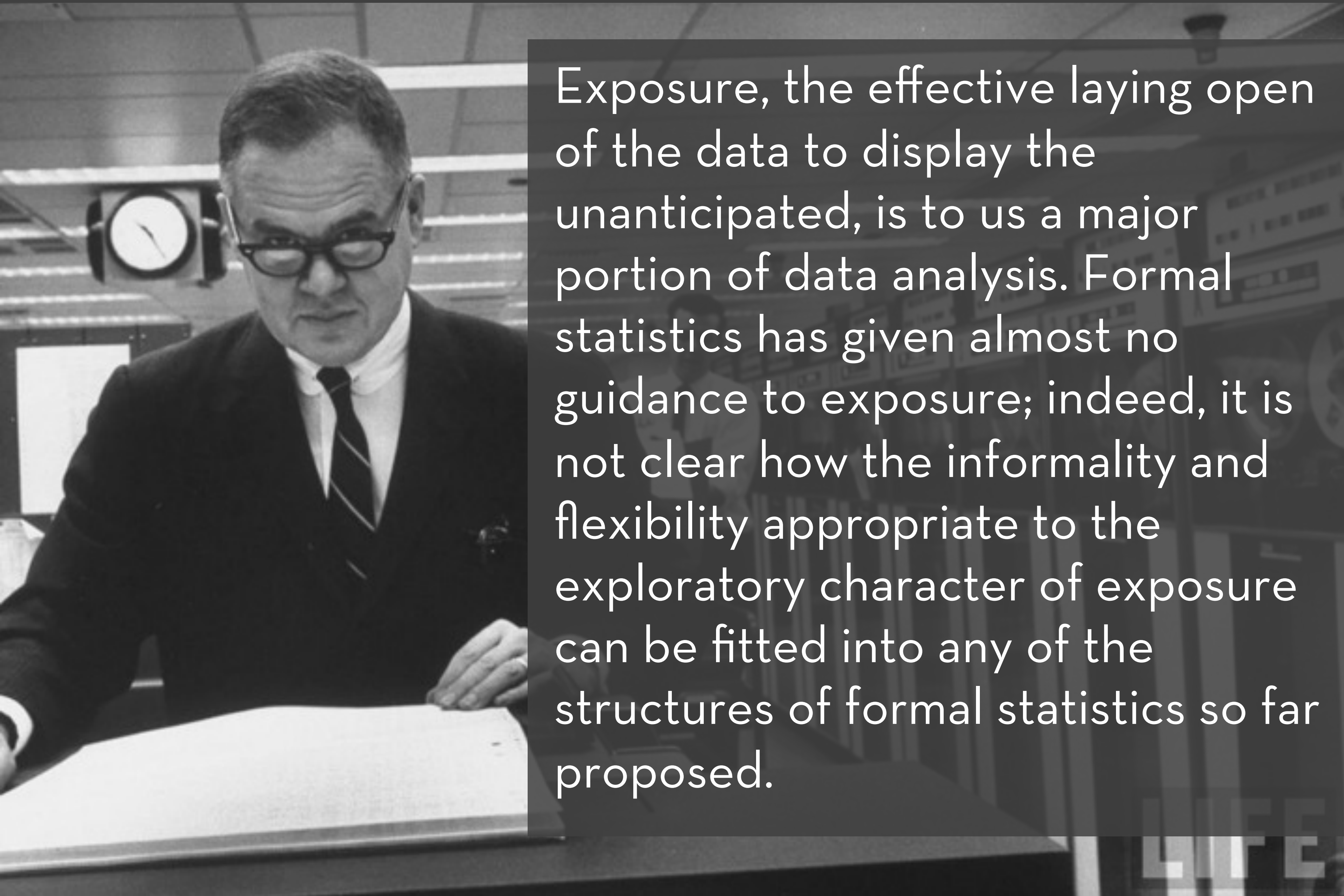
1786

Data Analysis & Statistics, Tukey 1962

The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.



While some of the influences of statistical theory on data analysis have been helpful, others have not.


Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

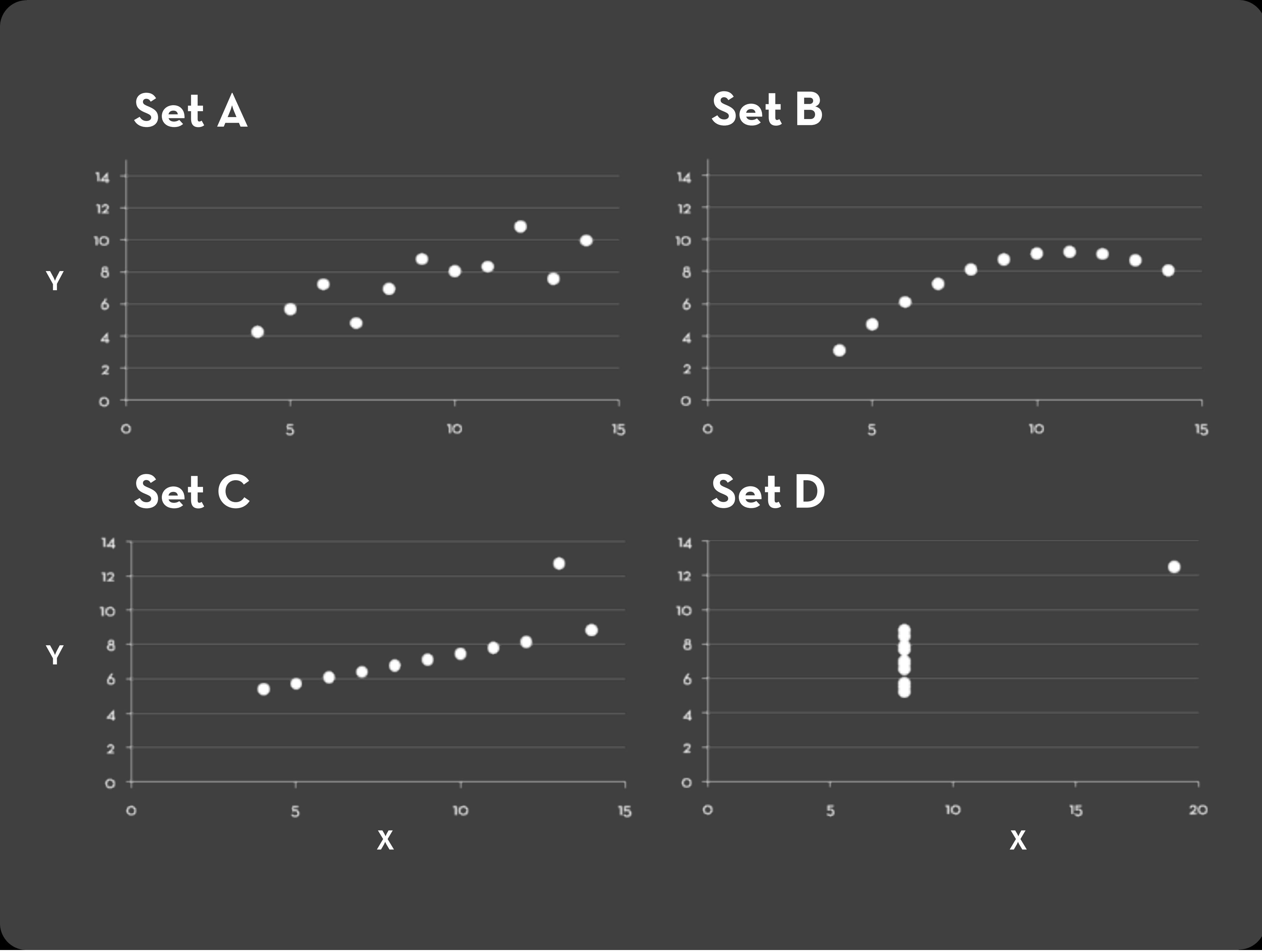
$u_Y = 7.5$ $\sigma_Y = 2.03$

Linear Regression

$Y = 3 + 0.5 X$

$R^2 = 0.67$

Anscombe 1973


Set A
Set B
Set C
Set D
Y
Y
X
X
0 2 4 6 8 10 12 14
0 5 10 15
0 5 10 15 20

Topics

- Exploratory Data Analysis
  - Data Diagnostics
  - Graphical Methods
  - Data Transformation
- Incorporating Statistical Models
  - Statistical Hypothesis Testing
- Using Graphics and Models in Tandem

Data Diagnostics

Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4525375	4029.3	987	2732.4	309.9
2005	4548327	3900	955.8	2656	289
2006	4599030	3937	968.9	2645.1	322.9
2007	4627851	3974.9	980.2	2687	307.7
2008	4661900	4081.9	1080.7	2712.6	288.6

Reported crime in Alaska

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	657755	3370.9	573.6	2456.7	340.6
2005	663253	3615	622.8	2601	391
2006	670053	3582	615.2	2588.5	378.3
2007	683478	3373.9	538.9	2480	355.1
2008	686293	2928.3	470.9	2219.9	237.5

Reported crime in Arizona

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	5739879	5073.3	991	3118.7	963.5
2005	5953007	4827	946.2	2958	922
2006	6166318	4741.6	953	2874.1	914.4
2007	6338755	4502.6	935.4	2780.5	786.7
2008	6500180	4087.3	894.2	2605.3	587.8

Reported crime in Arkansas

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	2750000	4033.1	1096.4	2699.7	237
2005	2775708	4068	1085.1	2720	262
2006	2810872	4021.6	1154.4	2596.7	270.4
2007	2834797	3945.5	1124.4	2574.6	246.5
2008	2855390	3843.7	1182.7	2433.4	227.6

Reported crime in California

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	35842038	3423.9	686.1	2033.1	704.8
2005	36154147	3321	692.9	1915	712
2006	36457549	3175.2	676.9	1831.5	666.8
2007	36553215	3032.6	648.4	1784.1	600.2
2008	36756666	2940.3	646.8	1769.8	523.8

Reported crime in Colorado

Year	Population	Property crime rate	Burglary rate	Larceny-theft rate	Motor vehicle theft rate
2004	4601821	3918.5	717.3	2679.5	521.6

Data "Wrangling"

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration.

Some approaches include:

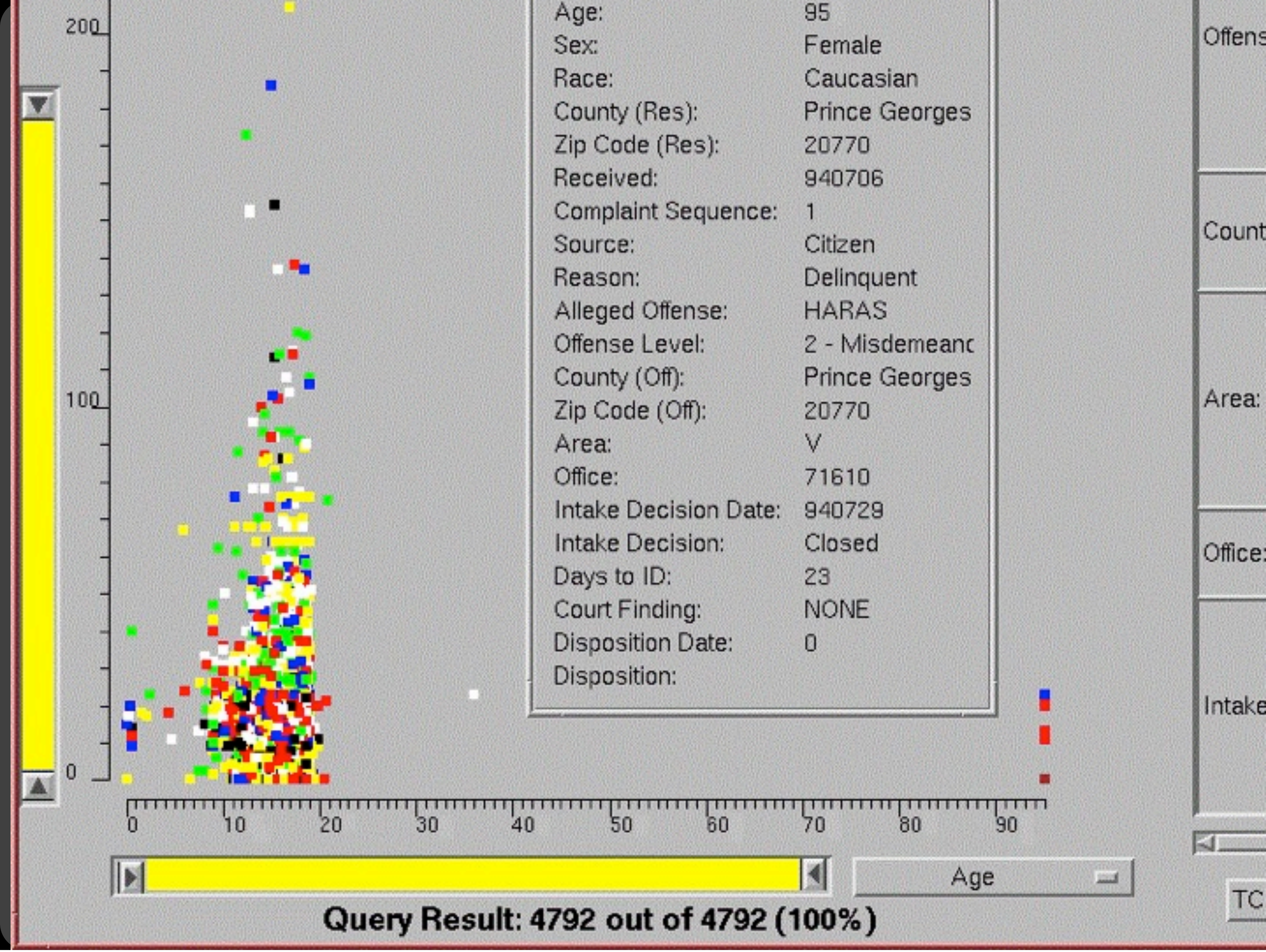
- Writing custom scripts
- Manual manipulation in spreadsheets
- Data Wrangler: http://vis.stanford.edu/wrangler
- Google Refine: http://code.google.com/p/google-refine

How to gauge the quality of a visualization?

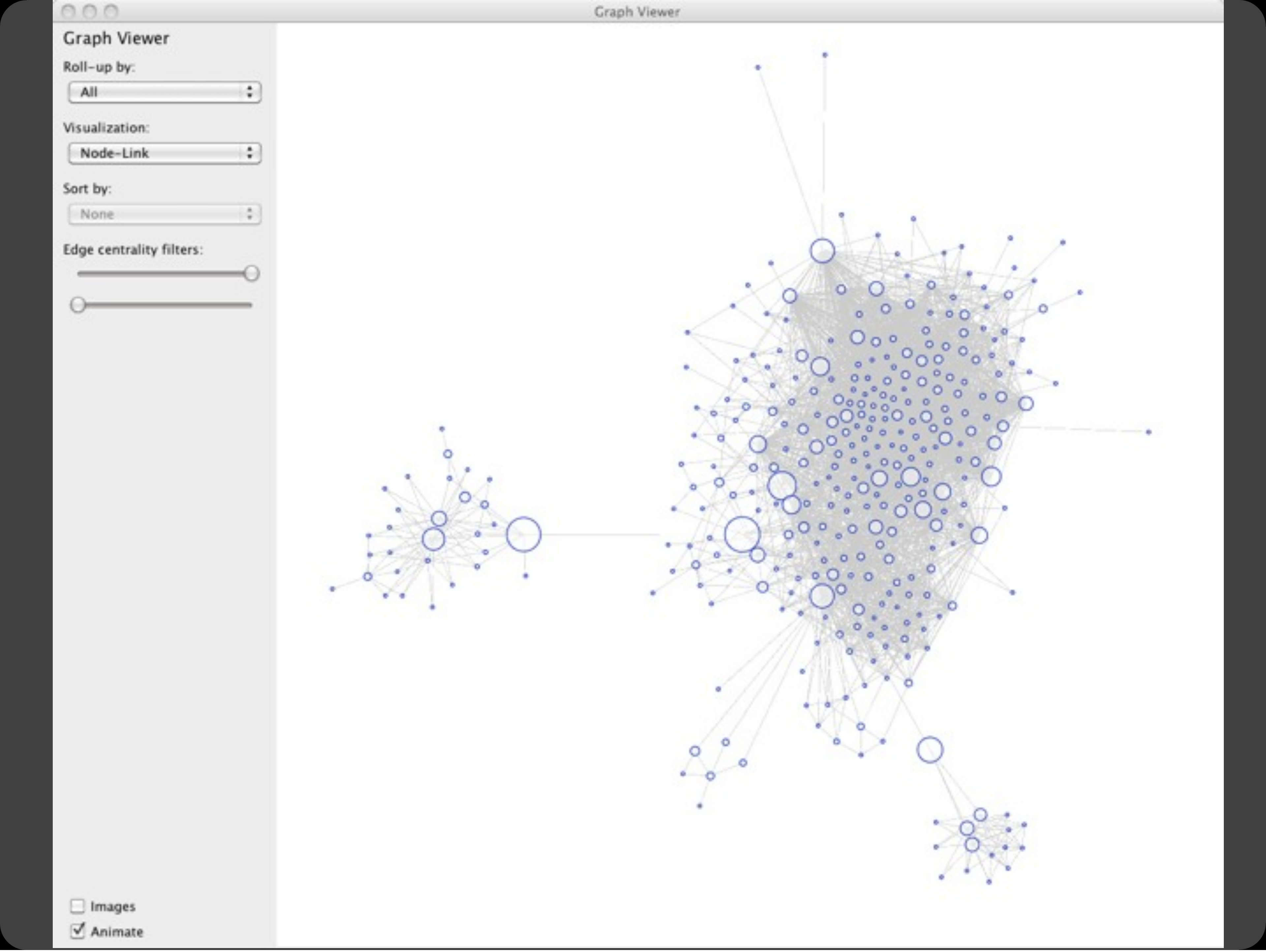
“The first sign that a visualization is good is that it shows you a problem in your data...

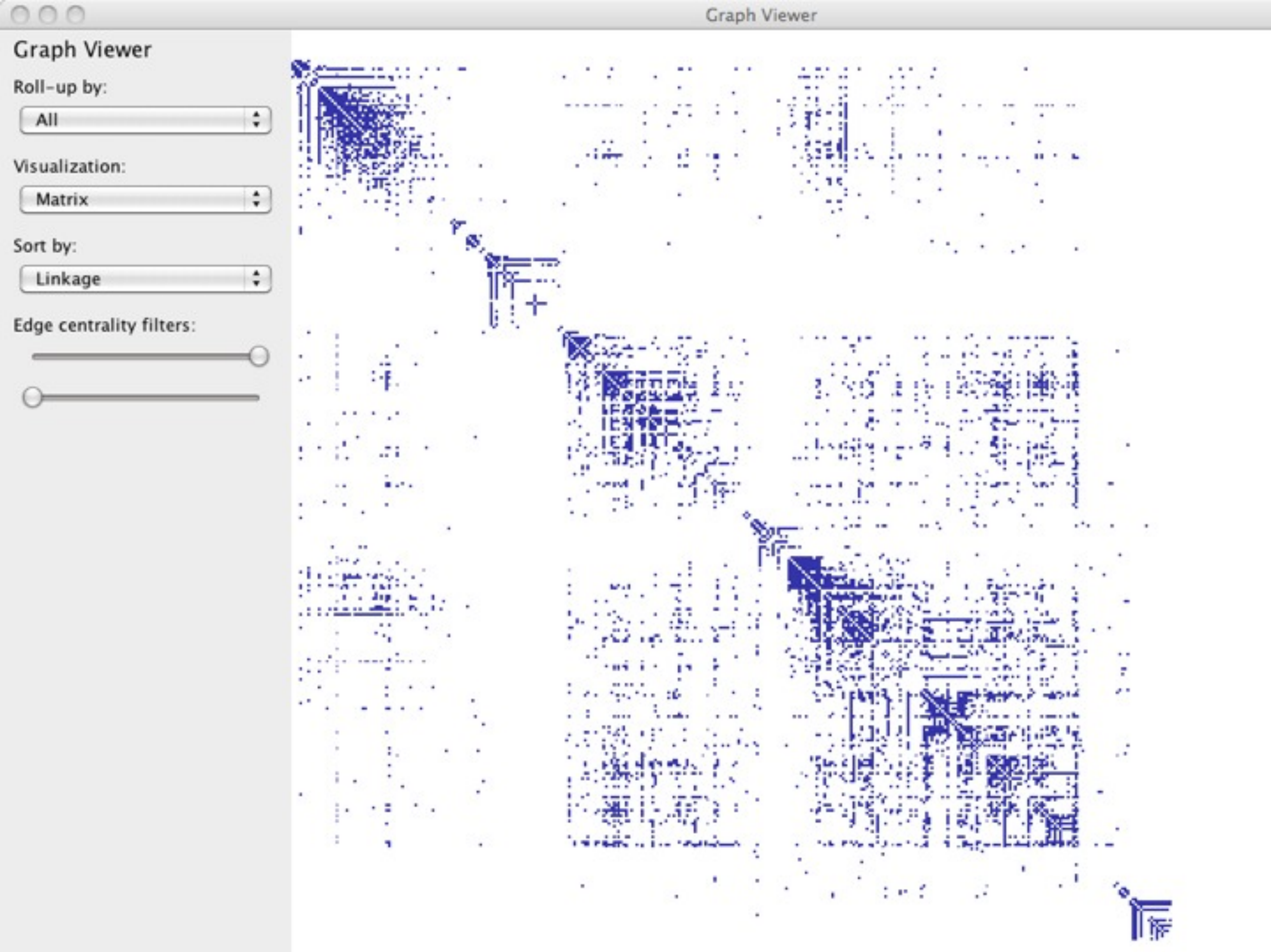
...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

- Martin Wattenberg

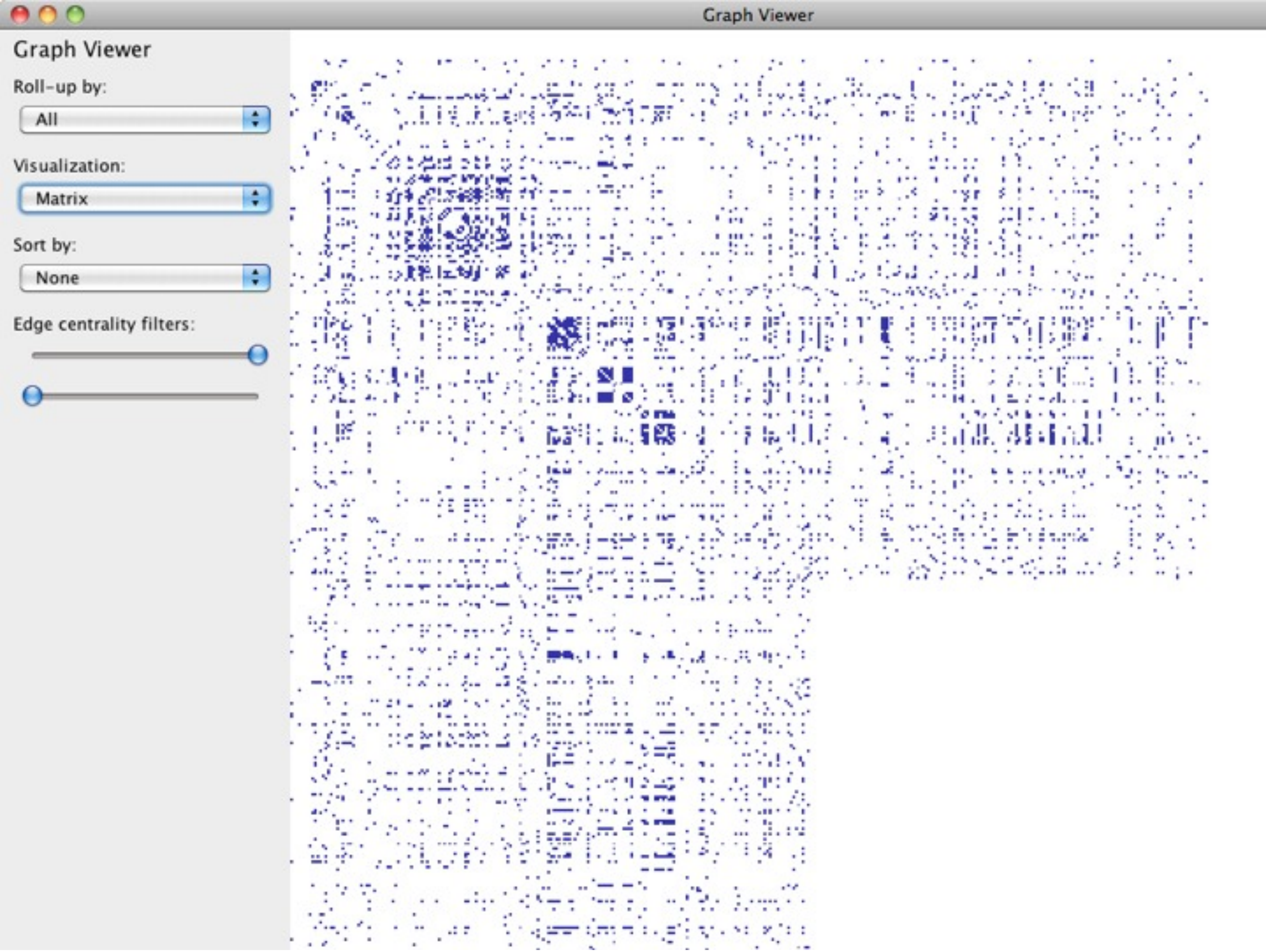


Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
County (Off): Prince Georges
Zip Code (Off): 20770
Area: V
Office: 71610
Intake Decision Date: 940729
Intake Decision: Closed
Days to ID: 23
Court Finding: NONE
Disposition Date: 0
Disposition:
200
100
0
0 10 20 30 40 50 60 70 80 90
Age
Query Result: 4792 out of 4792 (100%)
Offens
Count
Area:
Office:
Intake
TC


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate



Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:



Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

Visualize Friends by School?

Berkeley	|||||||||||||||||||||||||||||||
Cornell	||||
Harvard	|||||||||
Harvard University	|||||||
Stanford	||||||||||||||||||||
Stanford University	||||||||||
UC Berkeley	|||||||||||||||||||||
UC Davis	||||||||||
University of California at Berkeley	|||||||||||||||
University of California, Berkeley	||||||||||||||||||
University of California, Davis	|||

Data Quality & Usability Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

LESSON: Anticipate problems with your data.

Many research problems around these issues!

Exploratory Analysis:
Effectiveness of Antibiotics

The Data Set

Genus of Bacteria	String
Species of Bacteria	String
Antibiotic Applied	String
Gram-Staining?	Pos / Neg
Min. Inhibitory Concent. (g)	Number

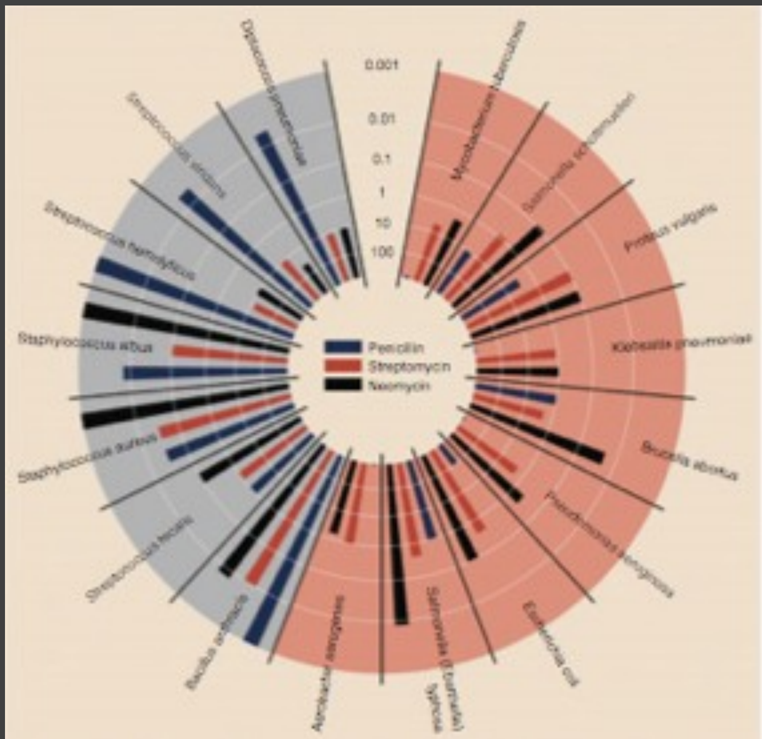
Collected prior to 1951.

What questions might we ask?

Table 1: Burtin's data.

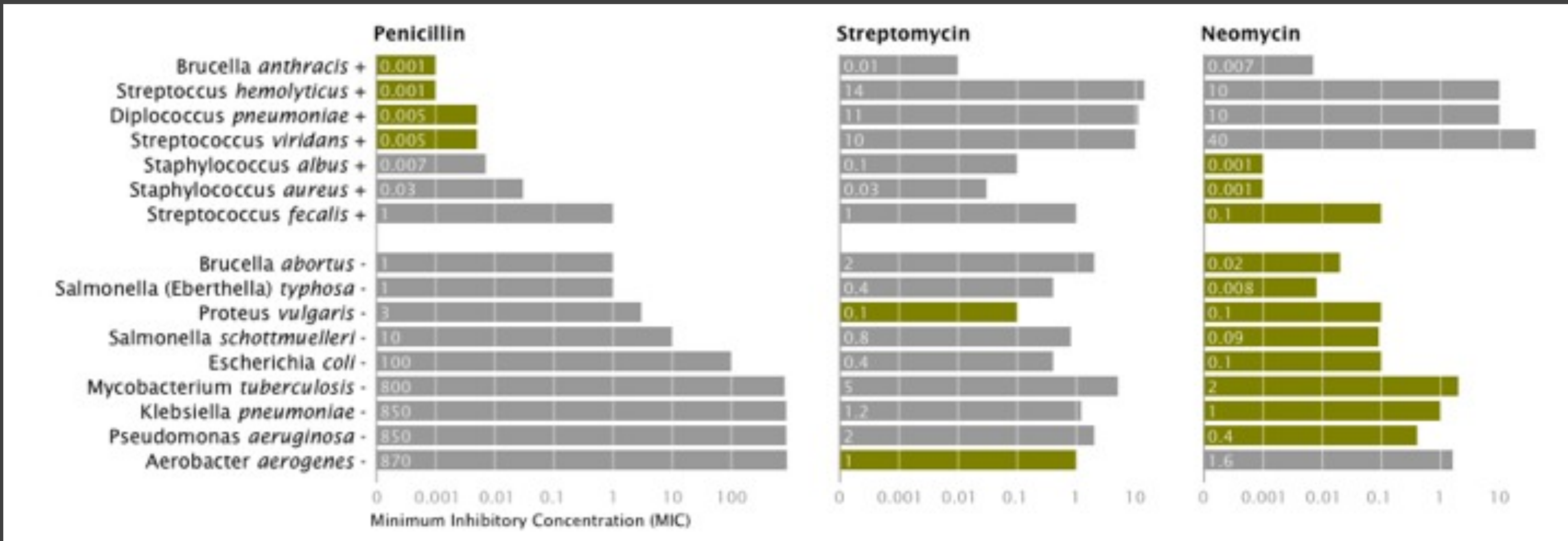
Bacteria	Antibiotic: Penicillin	Streptomycin	Neomycin	Gram Staining
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	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

Will Burtin, 1951

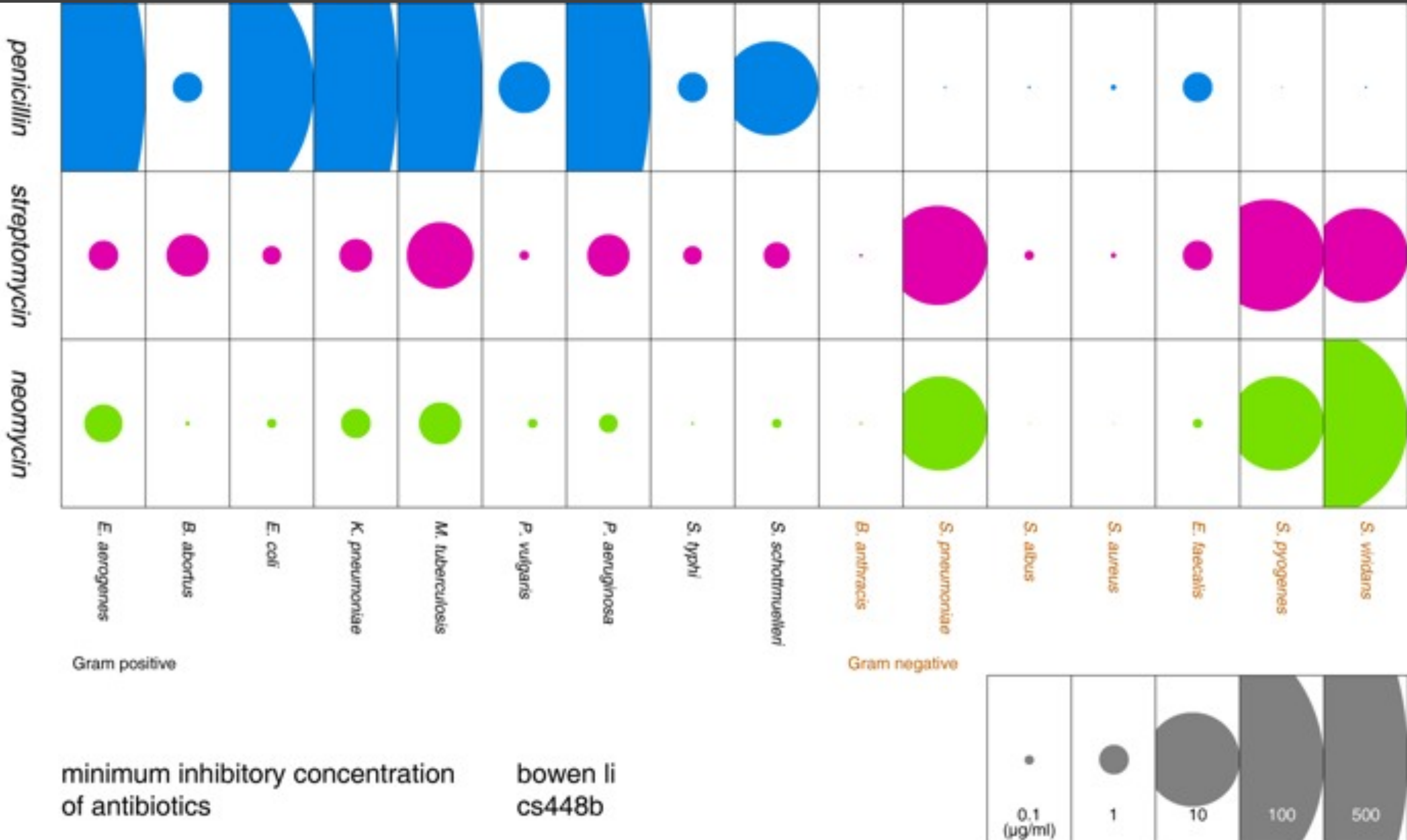


Bacteria	Antibiotic			Gram stain
	Penicillin	Streptomycin	Neomycin	
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	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	+

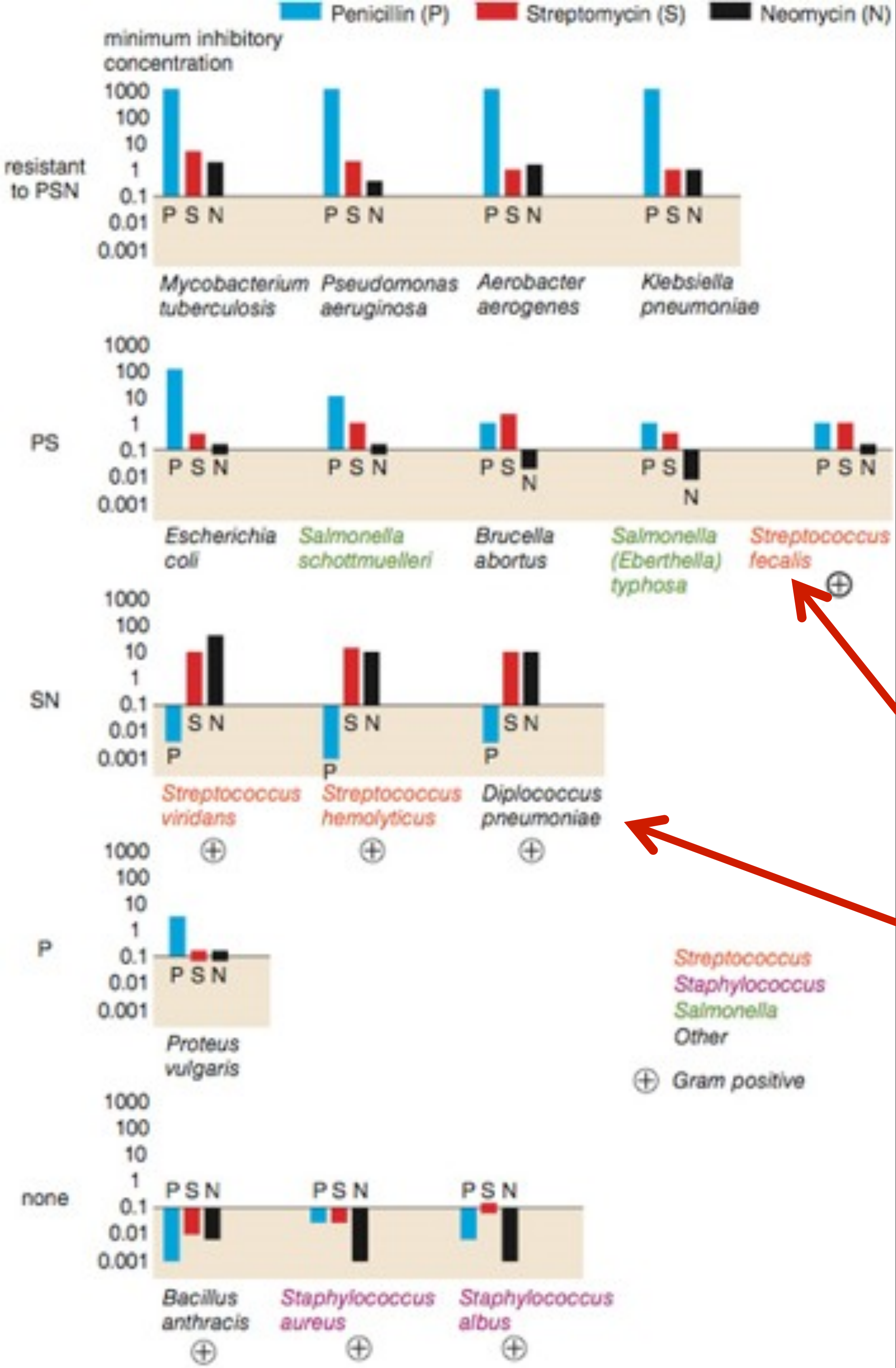
How do the drugs compare?



Mike Bostock, CS448B Winter 2009



Bowen Li, CS448B Fall 2009

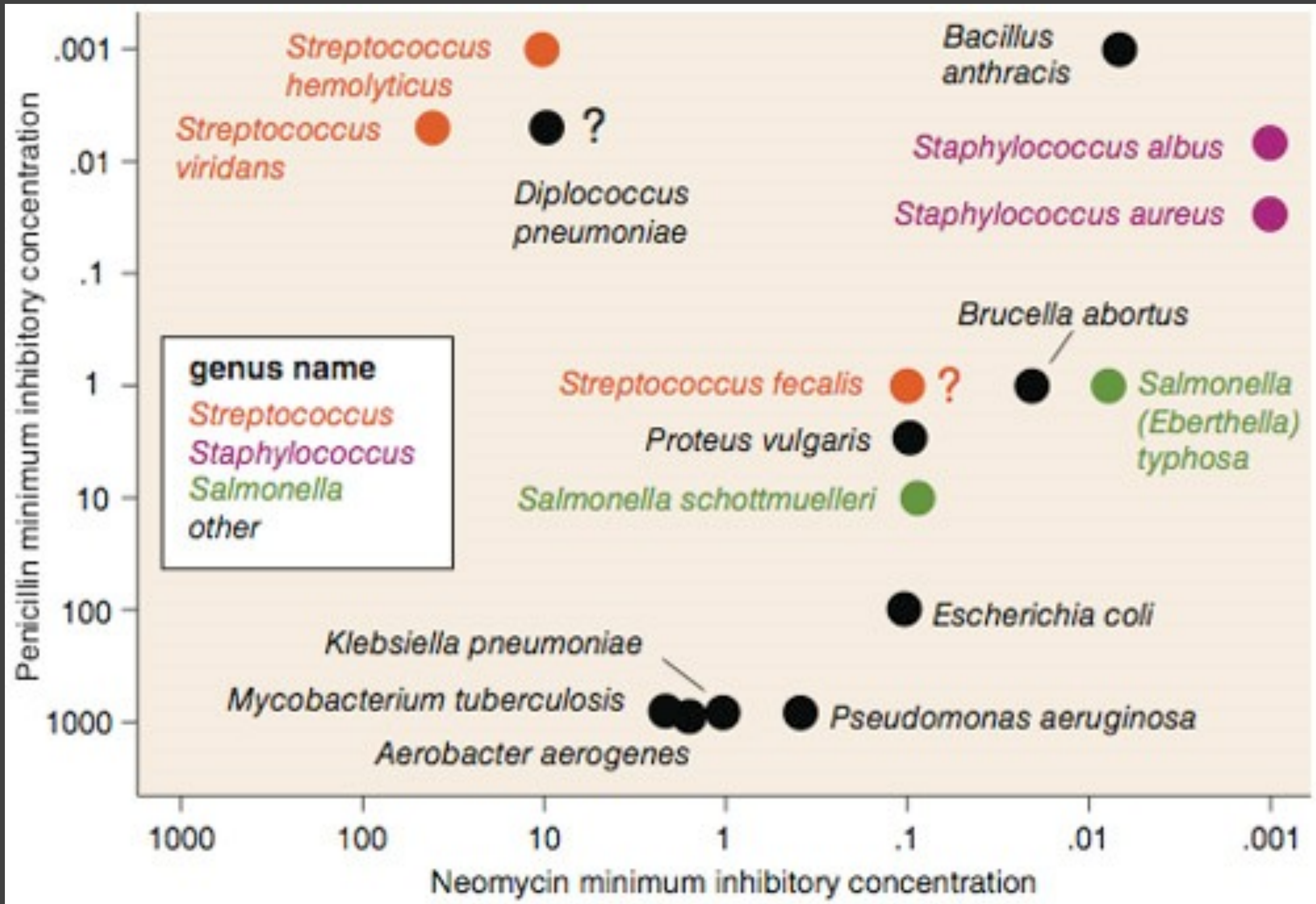


How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009



How do the bacteria group w.r.t. resistance?

Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009

Lesson: Iterative Exploration

Exploratory Process

1. Construct graphics to address questions
2. Inspect “answer” and assess new questions
3. Repeat!

Transform the data appropriately (e.g., invert, log)

“Show data variation, not design variation” -Tufte

Common Data Transformations

Normalize	$y_i \,/\, \Sigma_i\, y_i$	(among others)
Log	$\log y$	
Power	$y^{1/k}$	
Box-Cox Transform	$(y^\lambda - 1) \,/\, \lambda$	if $\lambda \neq 0$
	$\log y$	if $\lambda = 0$
Binning	e.g., histograms	
Grouping	e.g., merge categories	

Often performed to aid comparison (% or scale difference) or better approx. normal distribution

Exploratory Analysis: Participation on Amazon's Mechanical Turk

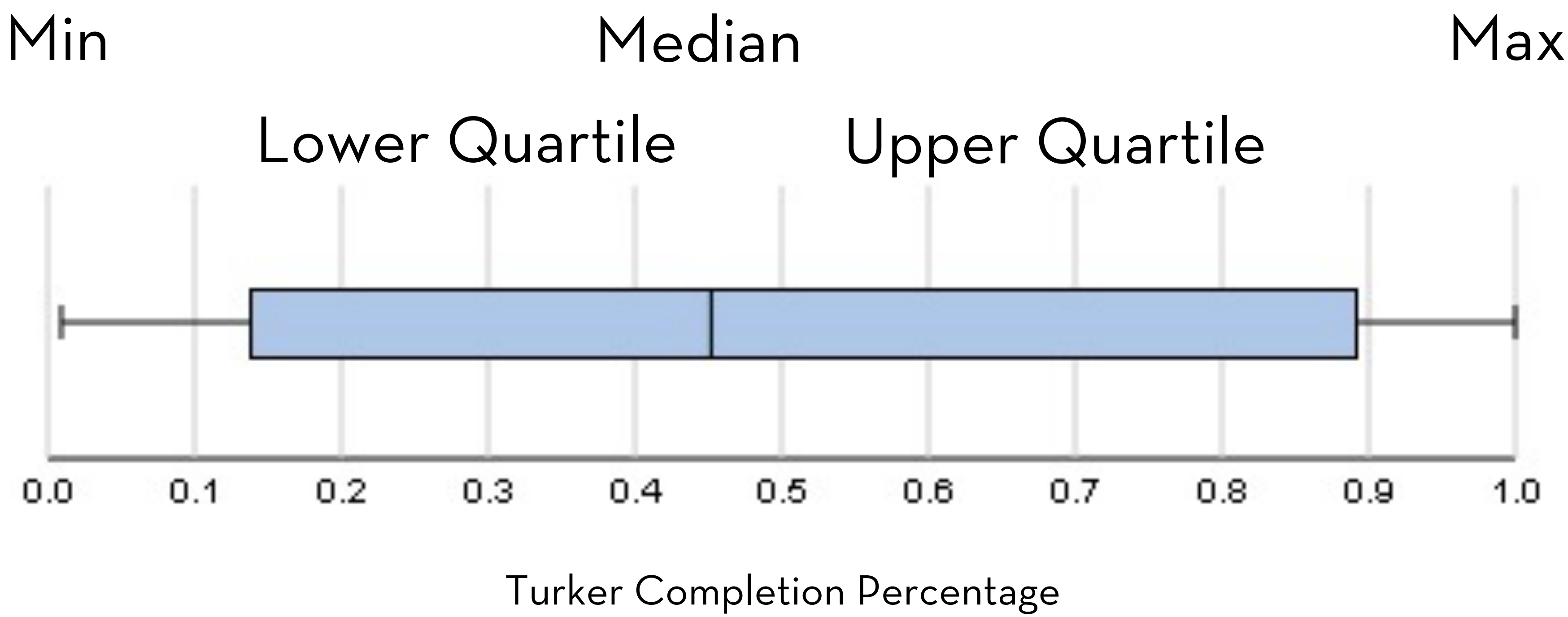
The Data Set (~200 rows)

Turker ID	String
Avg. Completion Rate	Number [0,1]

Collected in 2009 by Heer & Bostock.

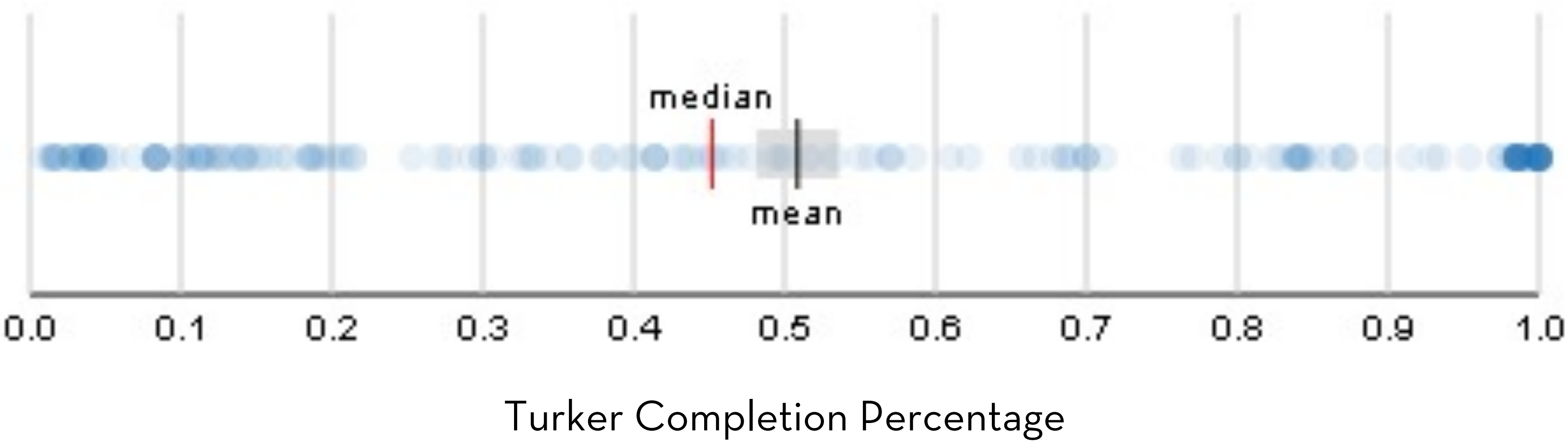
What questions might we ask of the data?

What charts might provide insight?

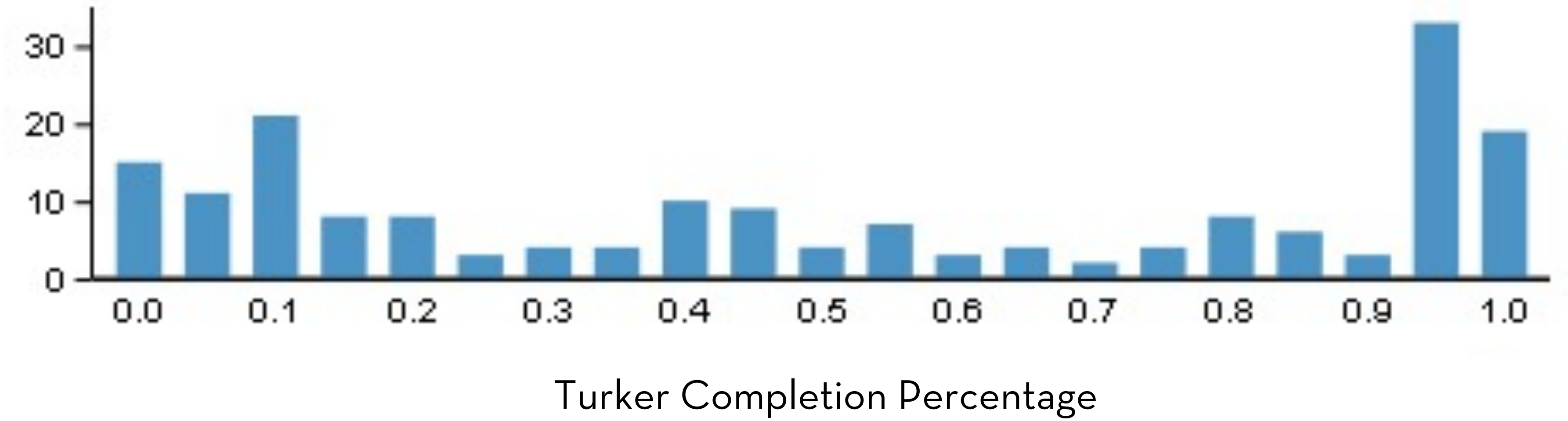


Box (and Whiskers) Plot

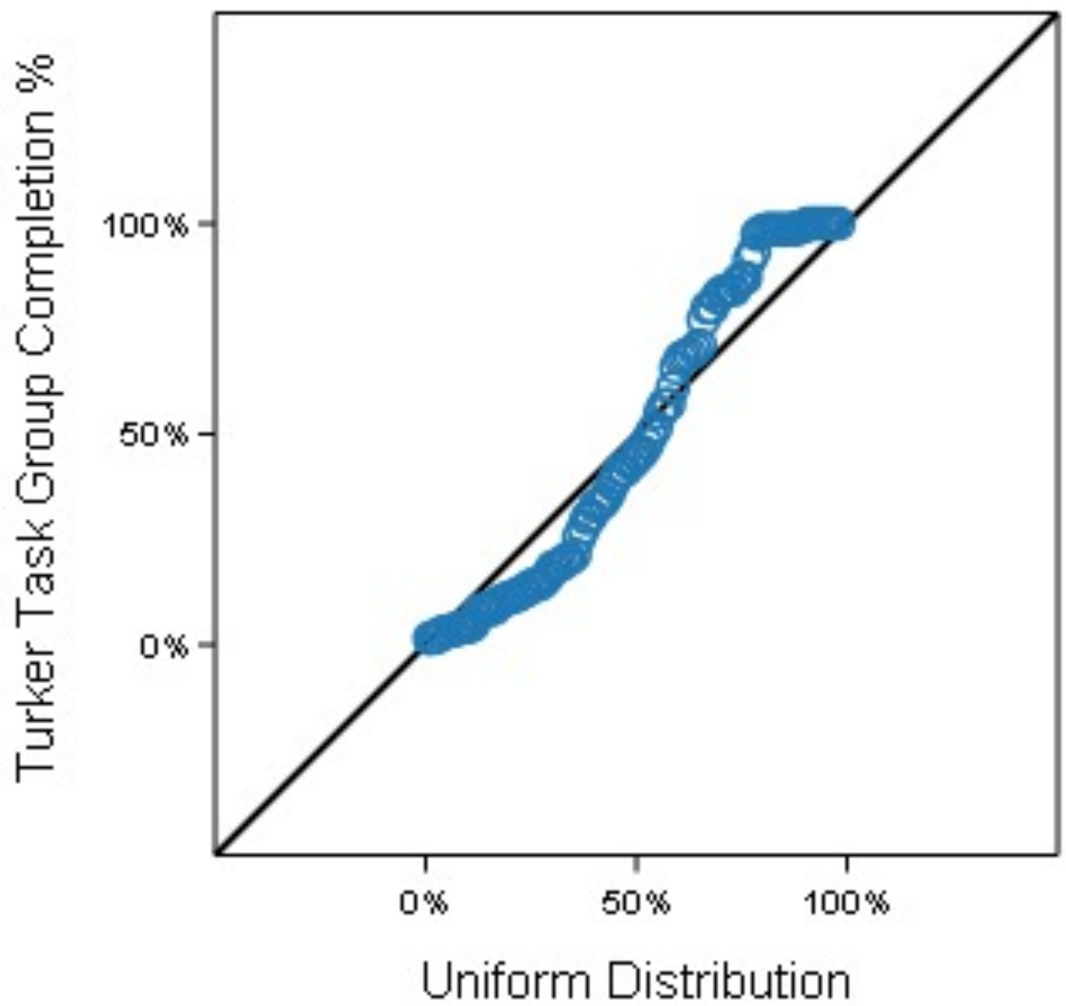
Dot Plot (with transparency to indicate overlap)



Dot Plot w/ Reference Lines



Histogram (binned counts)

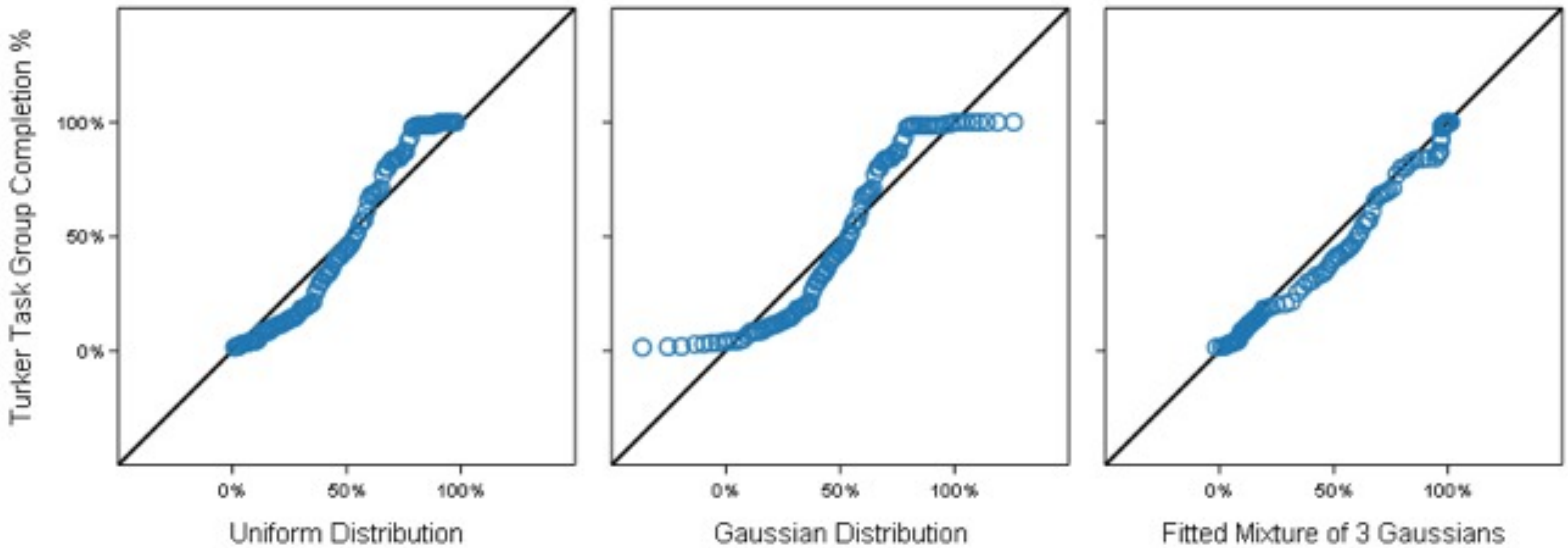


Quantile-Quantile Plot

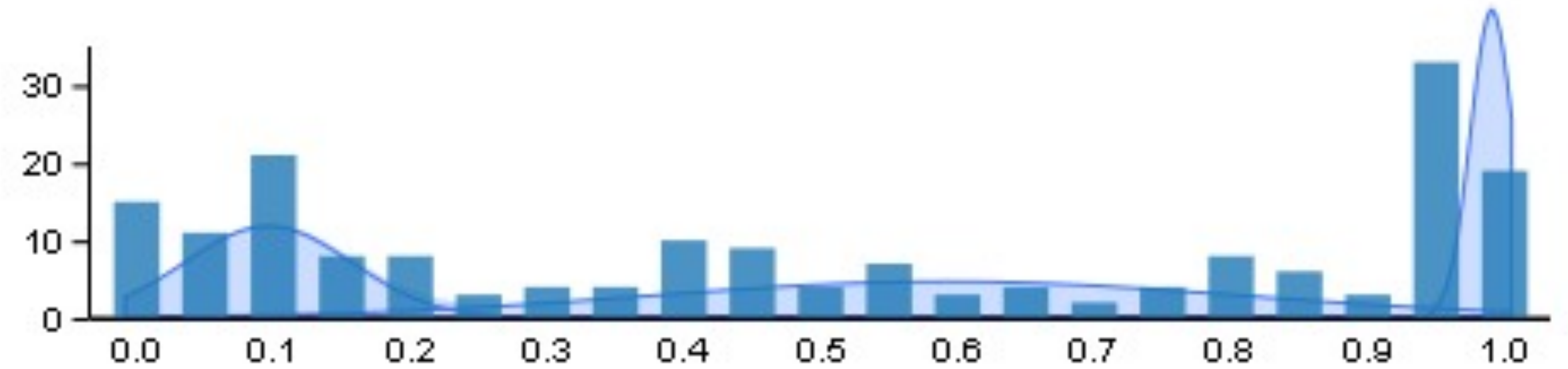
Used to compare two distributions; in this case, one actual and one theoretical.

Plots the quantiles (here, the percentile values) against each other.

Similar distributions lie along the diagonal. If linearly related, values will lie along a line, but with potentially varying slope and intercept.



Quantile-Quantile Plots



Turker Completion Percentage

Histogram + Fitted Mixture of 3 Gaussians

Lessons

Even for “simple” data, a variety of graphics might provide insight. Again, tailor the choice of graphic to the questions being asked, but be open to surprises.

Graphics can be used to understand and help assess the quality of statistical models.

Premature commitment to a model and lack of verification can lead an analysis astray.

Administrivia

Assignment 2: Exploratory Data Analysis

Use visualization software to form & answer questions

First steps:

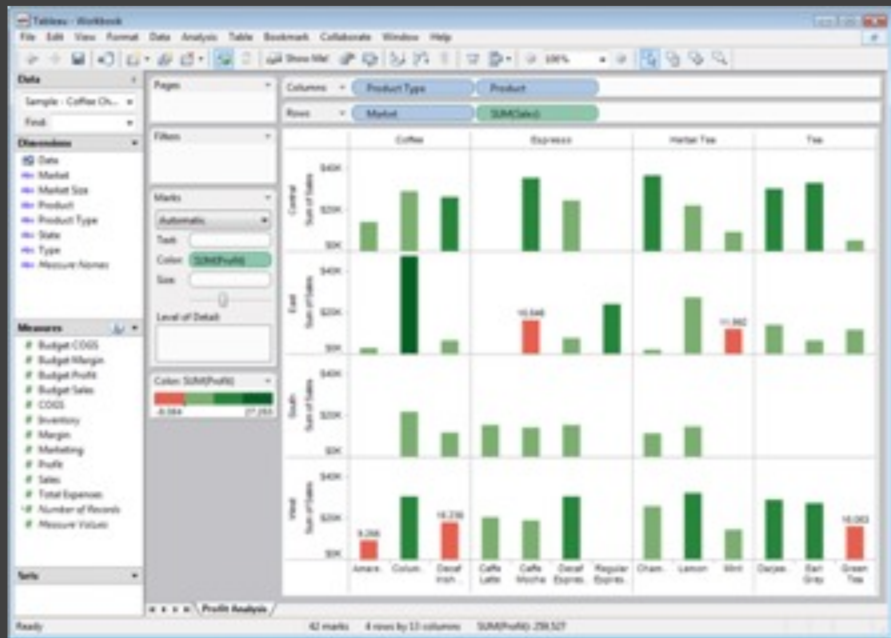
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile the data
- Iterate as needed

Create visualizations

- Interact with data
- Refine your questions

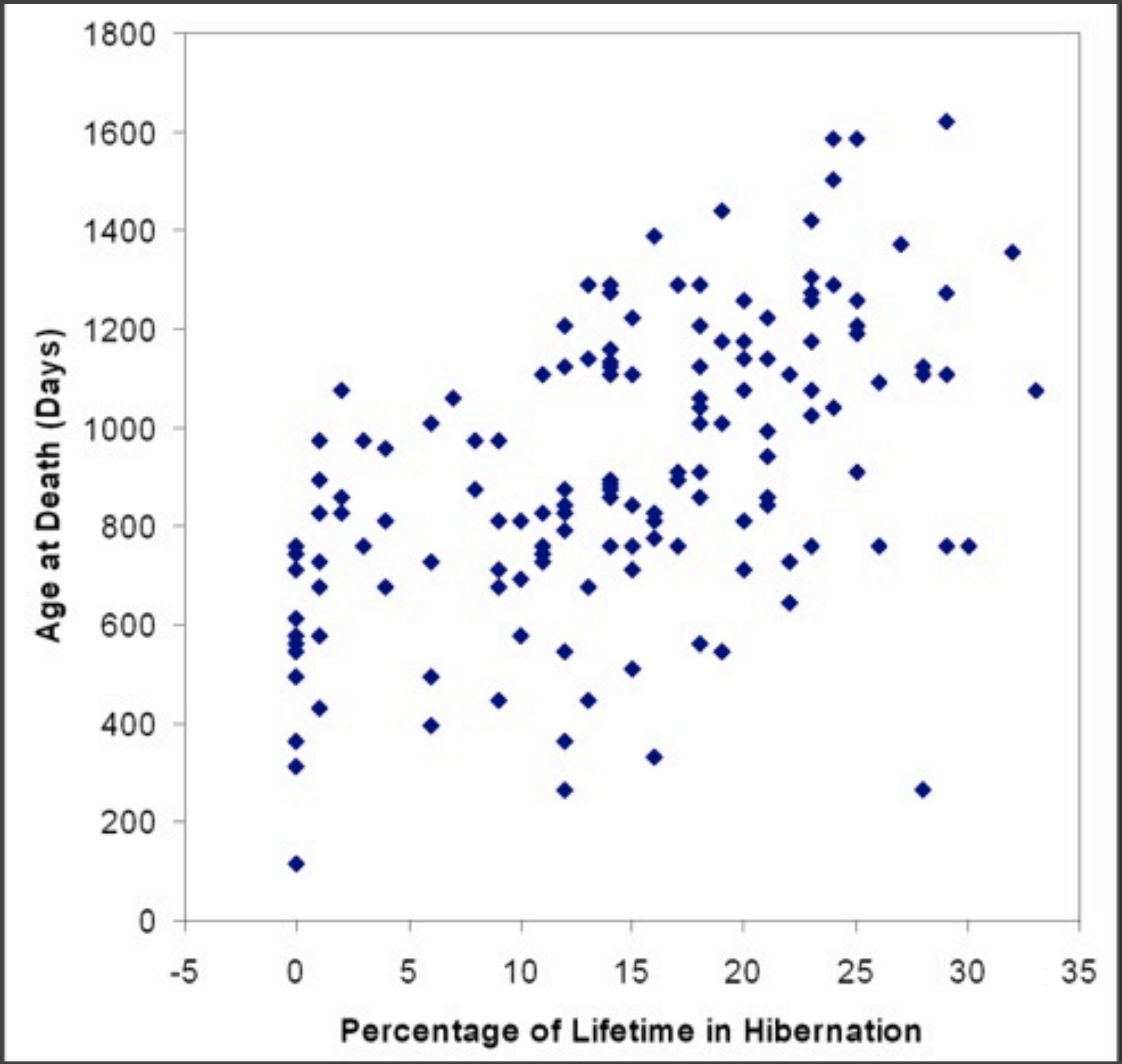
Make wiki notebook

- Keep record of your analysis
- Prepare a final graphic and caption

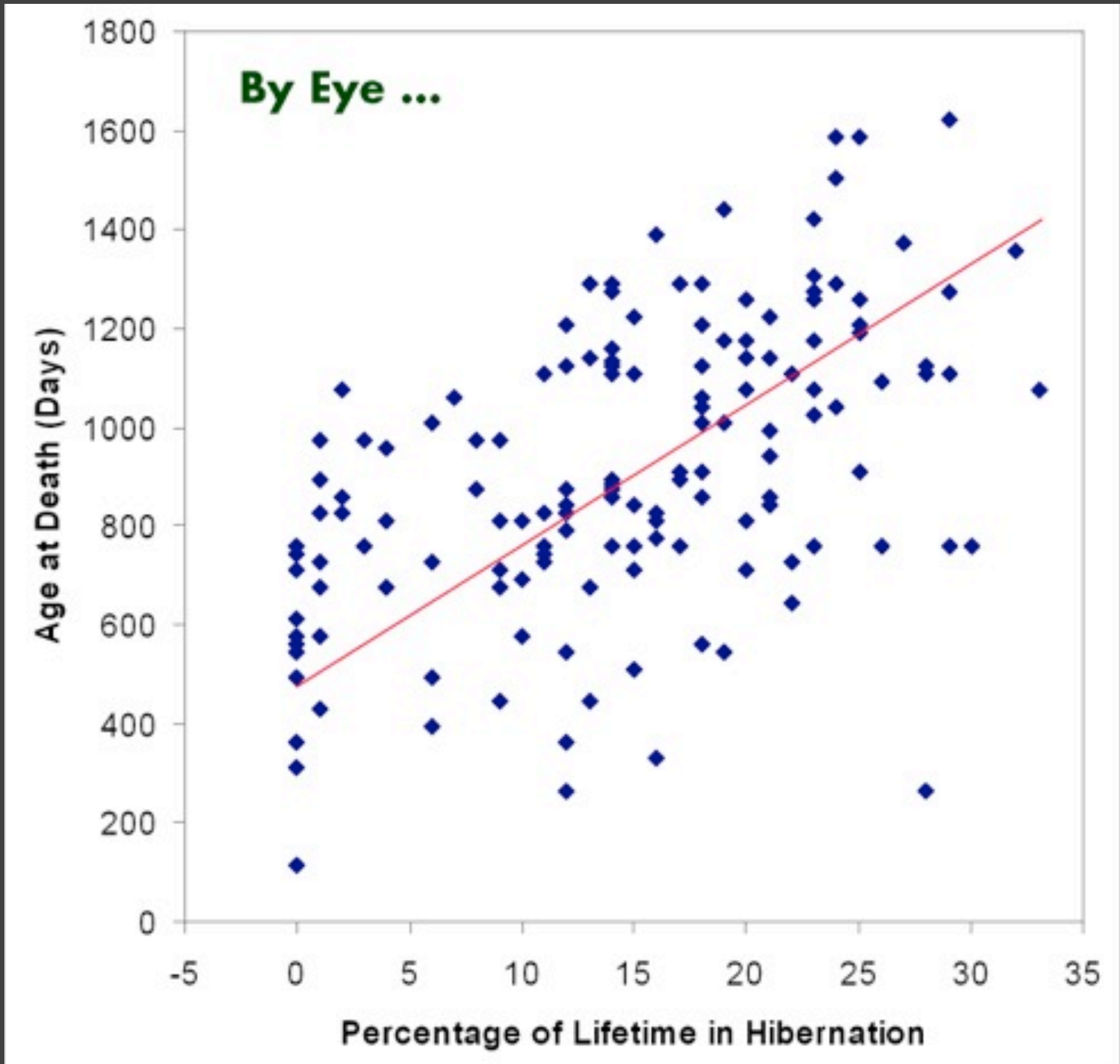


Due by 5:00pm
Thursday, Jan 23

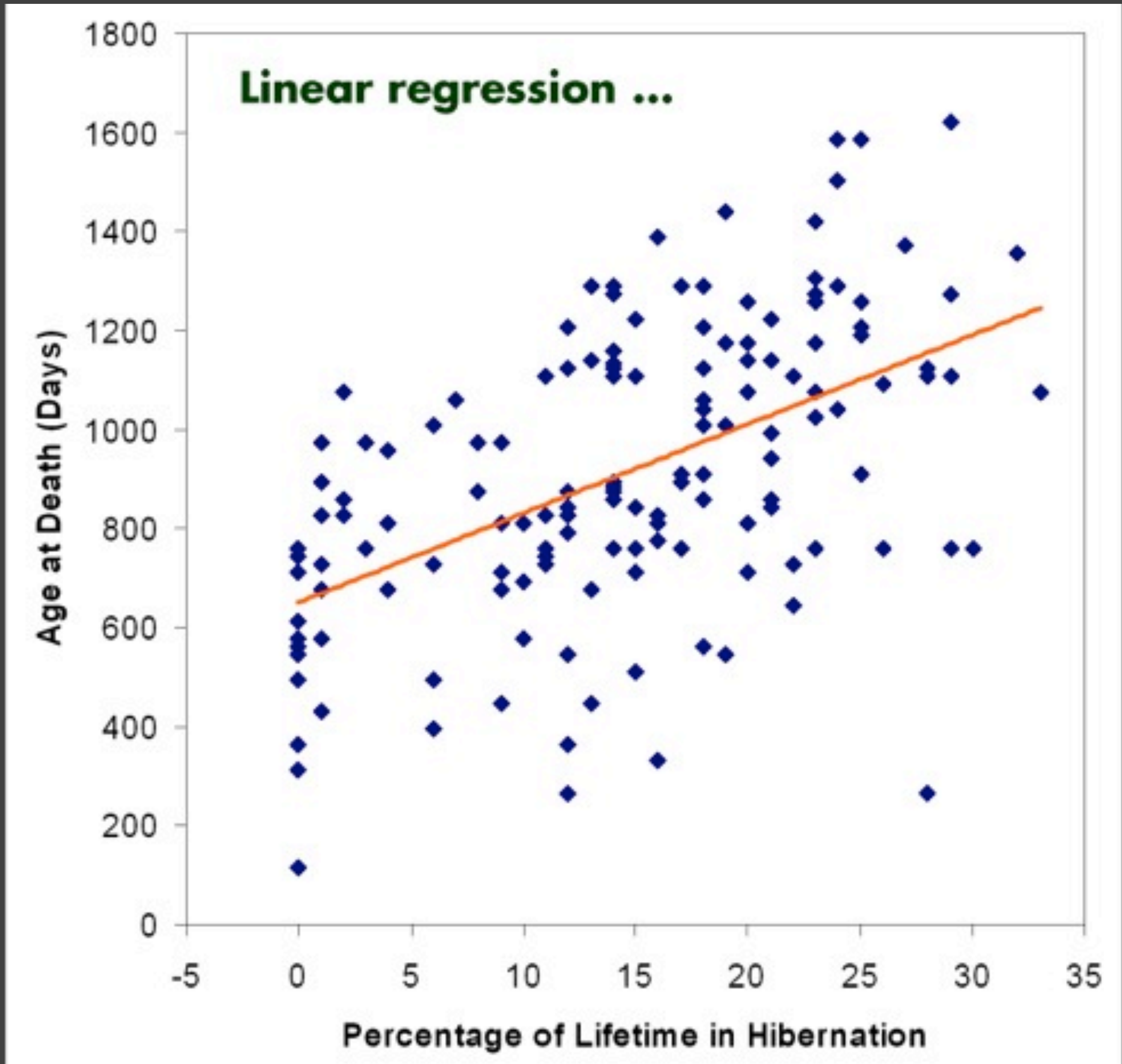
Using Visualization and Statistics Together



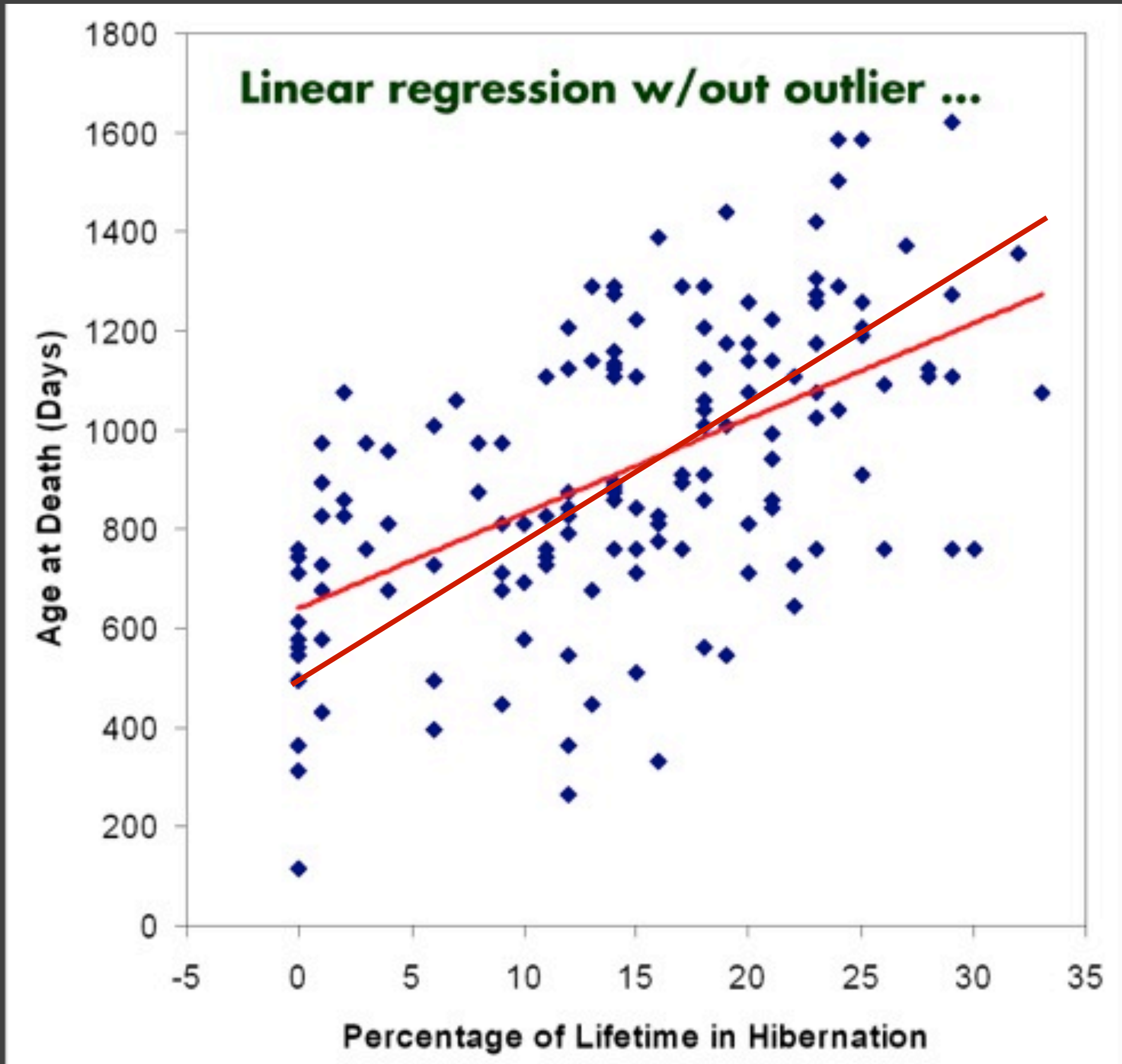
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]



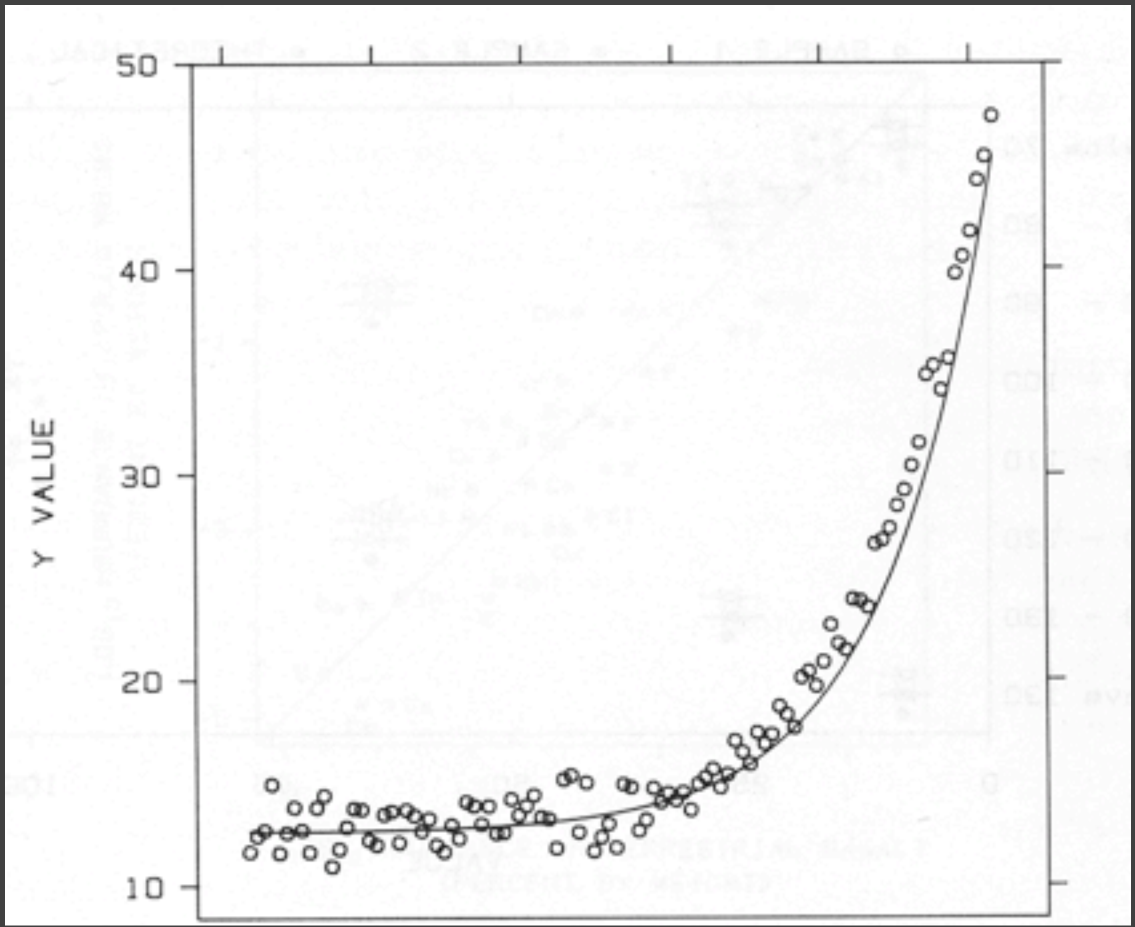
[The Elements of Graphing Data. Cleveland 94]



[The Elements of Graphing Data. Cleveland 94]

Transforming data

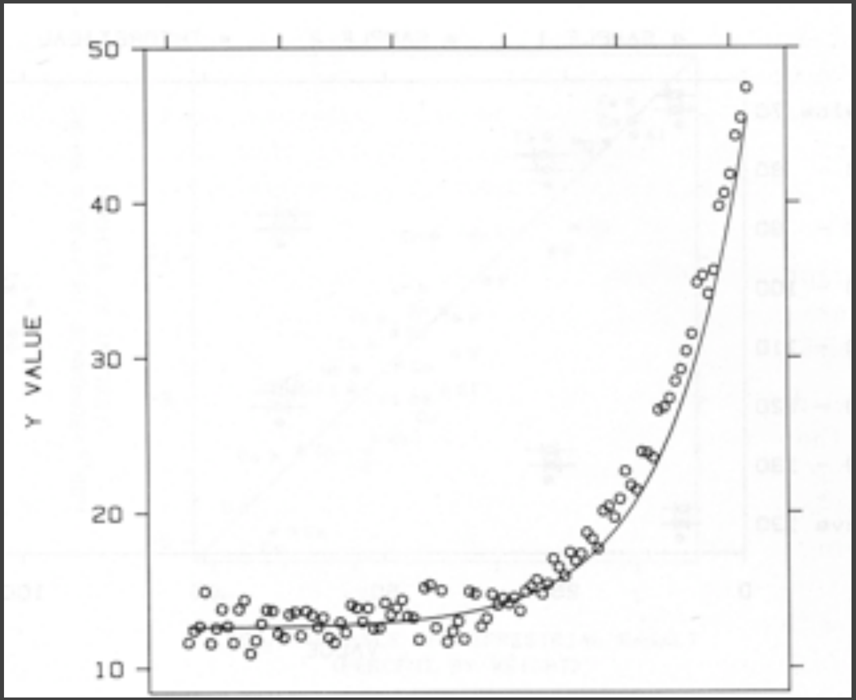
How well does the curve fit data?

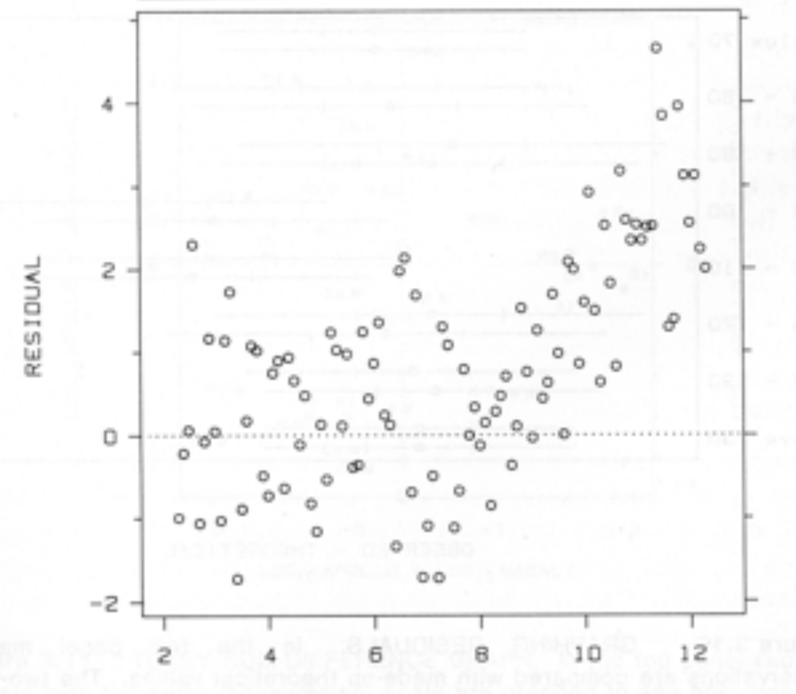


[Cleveland 85]

Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

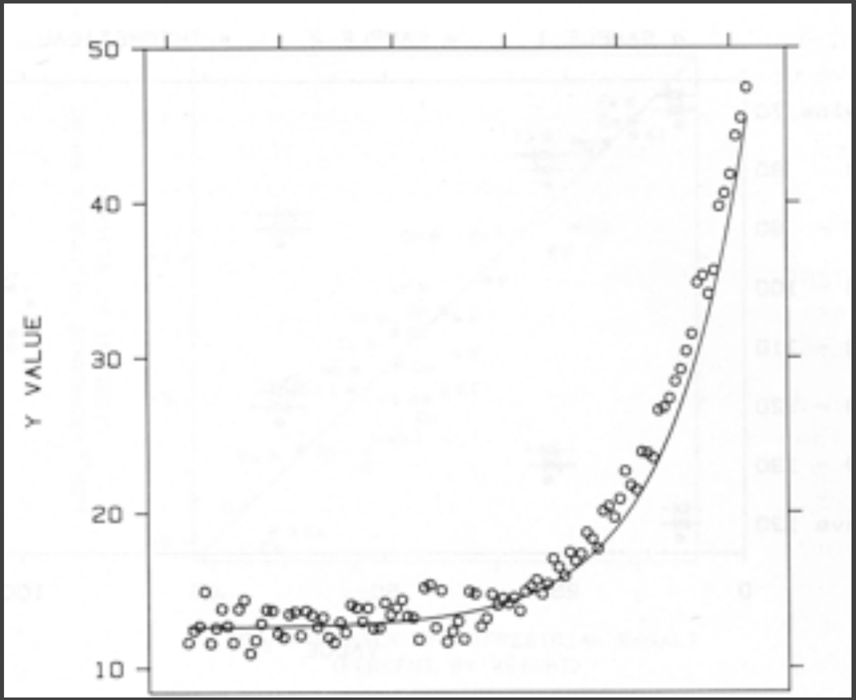




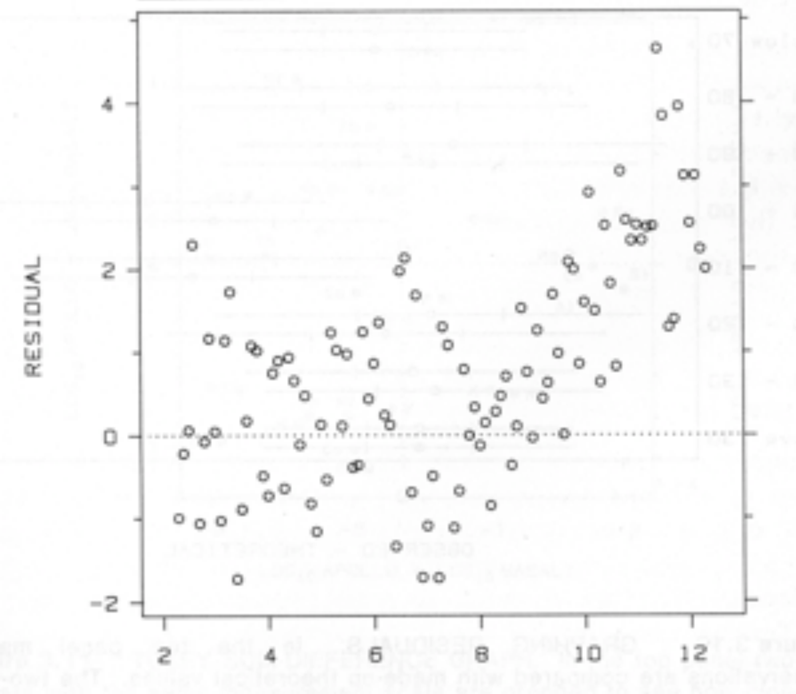
[Cleveland 85]

Multiple Plotting Options

Plot model in data space



Plot data in model space



[Cleveland 85]

Confirmatory Analysis

Incorporating Models

Hypothesis testing: What is the probability that the pattern might have arisen by chance?

Prediction: How well do one (or more) data variables predict another?

Abstract description: With what parameters does the data best fit a given function? What is the goodness of fit?

Scientific theory: Which model explains reality?

Example: Heights by Gender

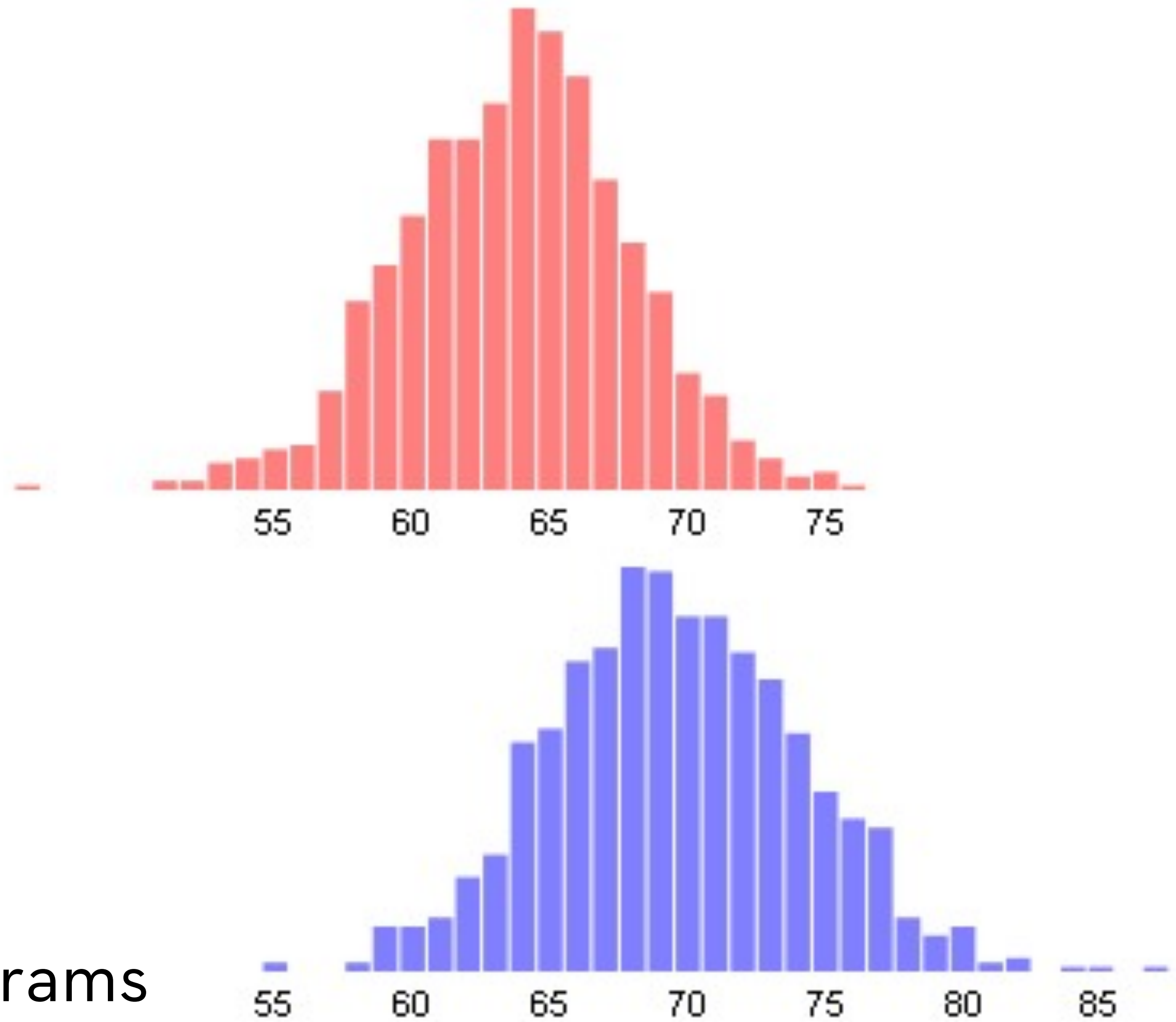
Gender	Male / Female
Height (in)	Number

$\mu_m = 69.4 \quad \sigma_m = 4.69 \quad N_m = 1000$

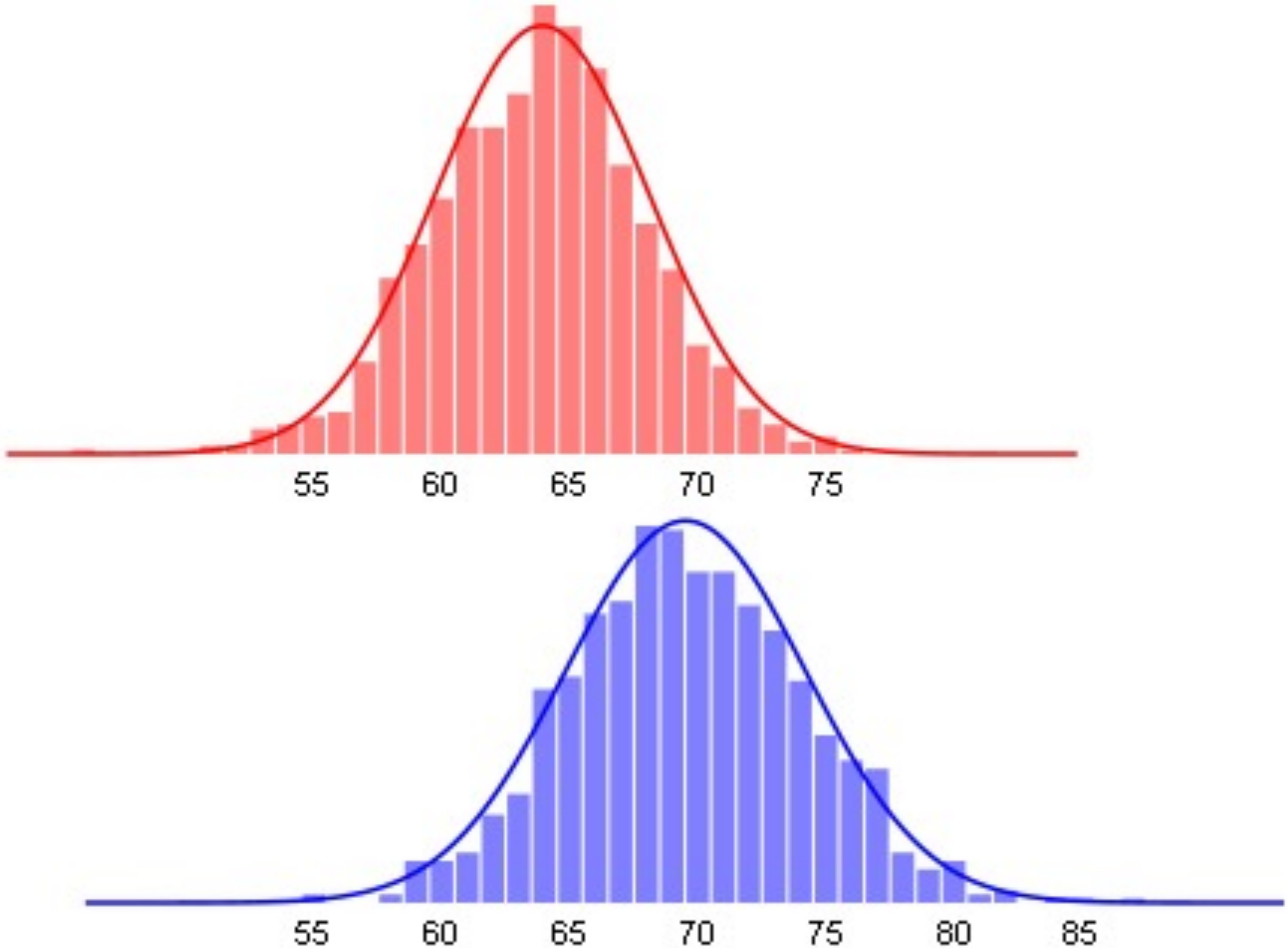
$\mu_f = 63.8 \quad \sigma_f = 4.18 \quad N_f = 1000$

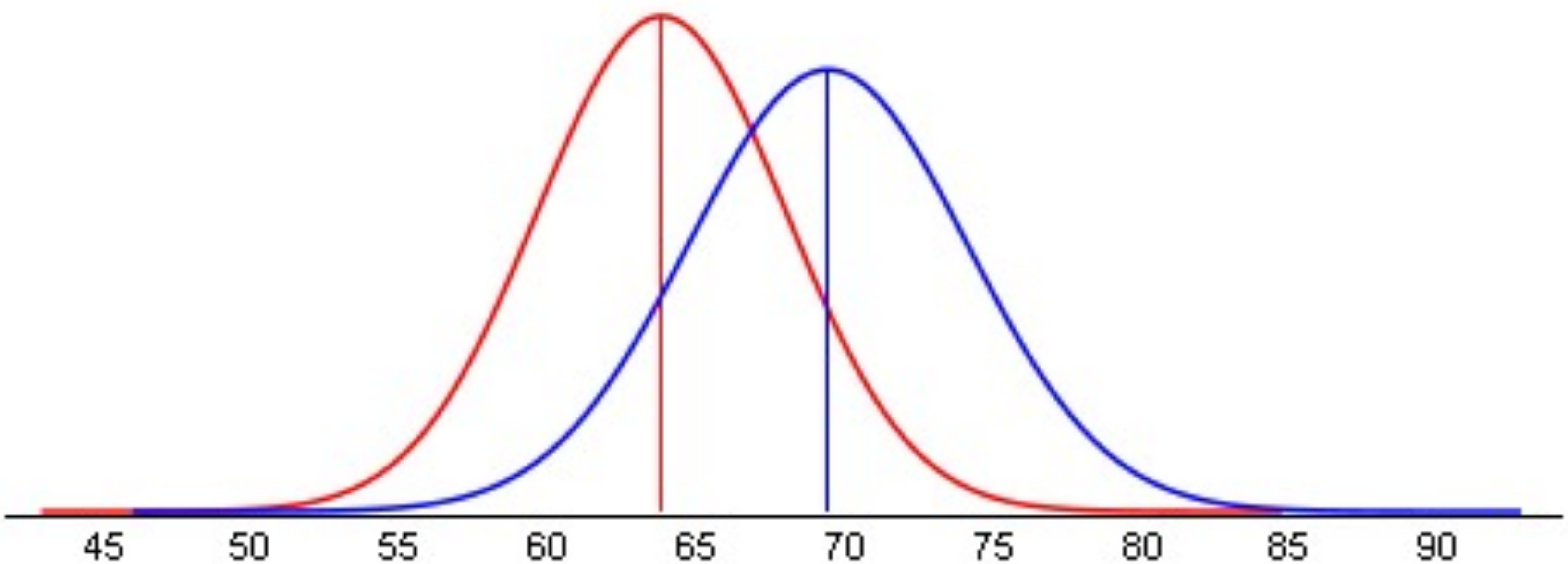
Is this difference in heights significant?

In other words: assuming no true difference, what is the prob. that our data is due to chance?



Histograms


55 60 65 70 75
55 60 65 70 75 80 85


45
50
55
60
65
70
75
80
85
90

Formulating a Hypothesis

Null Hypothesis (**H$_o$**): $\mu_m = \mu_f$ (population)

Alternate Hypothesis (**H$_a$**): $\mu_m \neq \mu_f$ (population)

A **statistical hypothesis test** assesses the likelihood of the null hypothesis.

What is the probability of sampling the observed data assuming population means are equal?

This is called the ***p* value**.

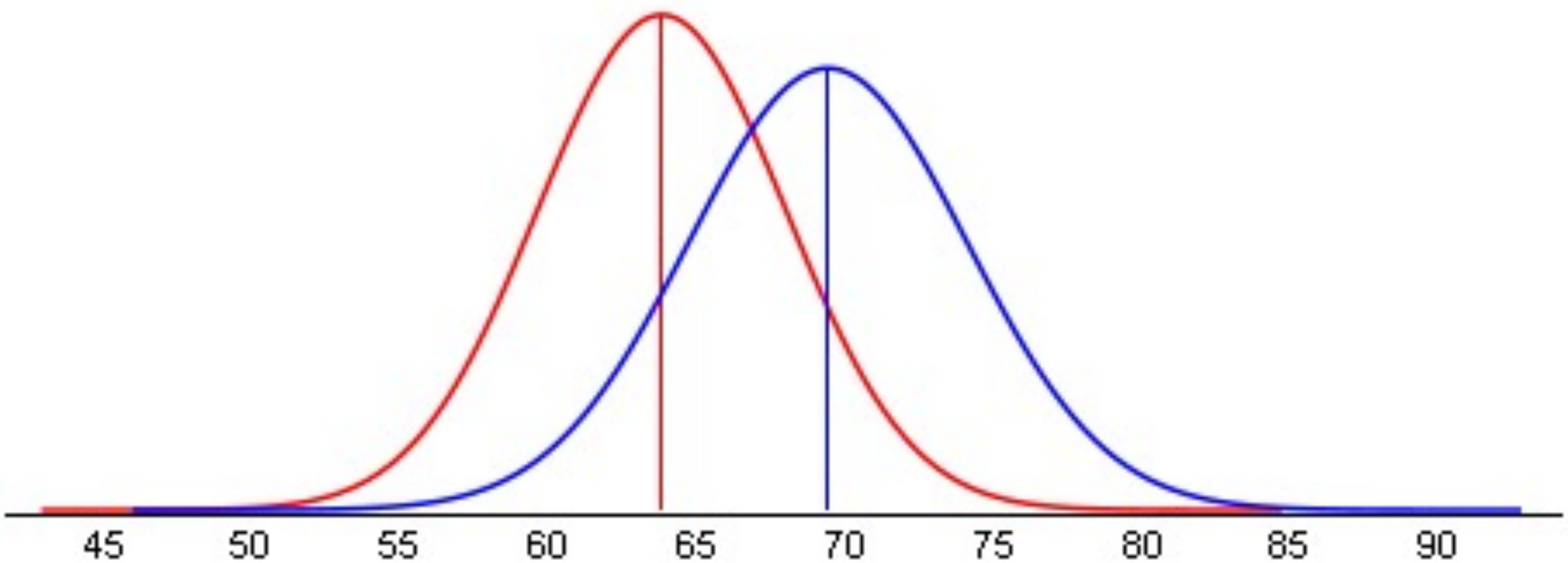
Testing Procedure

Compute a **test statistic**. This is a number that in essence summarizes the difference.

Compute test statistic

$$Z = \frac{\mu_m - \mu_f}{\sqrt{\sigma^2_m / N_m + \sigma^2_f / N_f}}$$

$\mu_m - \mu_f = 5.6$



Testing Procedure

Compute a **test statistic**. This is a number that in essence summarizes the difference.

The possible values of this statistic come from a **known probability distribution**.

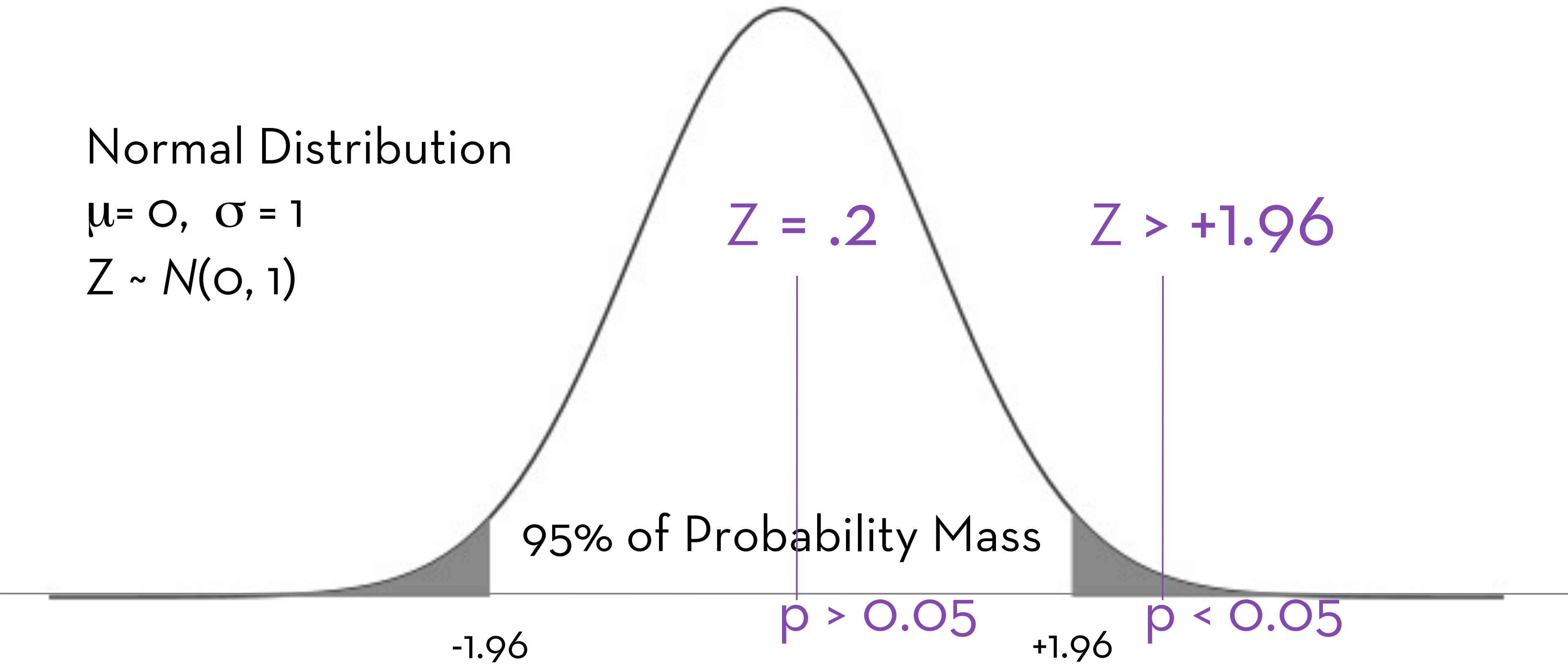
According to this distribution, determine the probability of seeing a value meeting or exceeding the test statistic. This is the ***p* value**.

Lookup probability of test statistic

Normal Distribution

$\mu = 0, \ \sigma = 1$

$Z \sim N(0, 1)$



Statistical Significance

The threshold at which we consider it safe (or reasonable?) to *reject the null hypothesis*.

If $p < 0.05$, we typically say that the observed effect or difference is **statistically significant**.

This means that there is a less than 5% chance that the observed data is due to chance.

Note that the choice of 0.05 is a somewhat arbitrary threshold (chosen by R. A. Fisher)

Common Statistical Methods

Question	Data Type	Parametric	Non-Parametric

Assumes a particular distribution for the data -- usually normal, a.k.a. Gaussian.

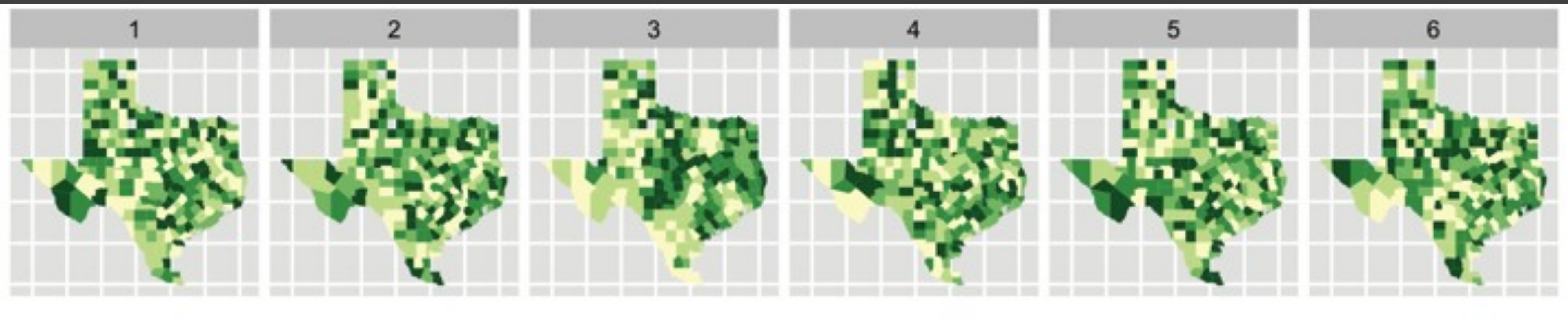
Does not assume a distribution. Typically works on rank orders.

Common Statistical Methods

Question	Data Type	Parametric	Non-Parametric
Do data distributions have different "centers"? (aka "location" tests)	2 uni. dists > 2 uni. dists > 2 multi. dists	t-Test ANOVA MANOVA	Mann-Whitney U Kruskal-Wallis Median Test
Are observed counts significantly different?	Counts in categories		χ^2 (chi-squared)
Are two vars related?	2 variables	Pearson coeff.	Rank correl.
Do 1 (or more) variables predict another?	Continuous Binary	Linear regression Logistic regression	

Graphical Inference

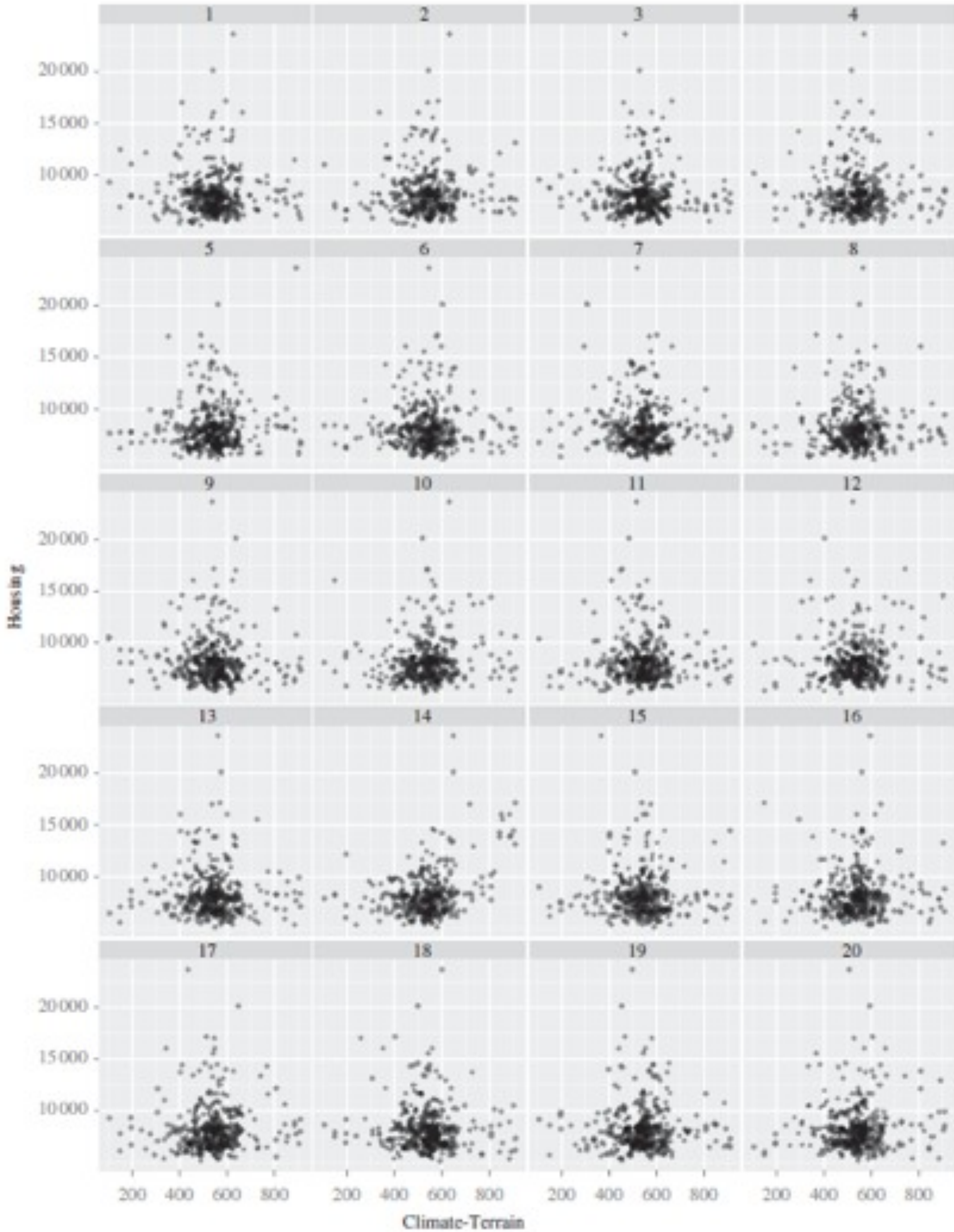
(Buja, Cook, Hofmann, Wickham, et al.)



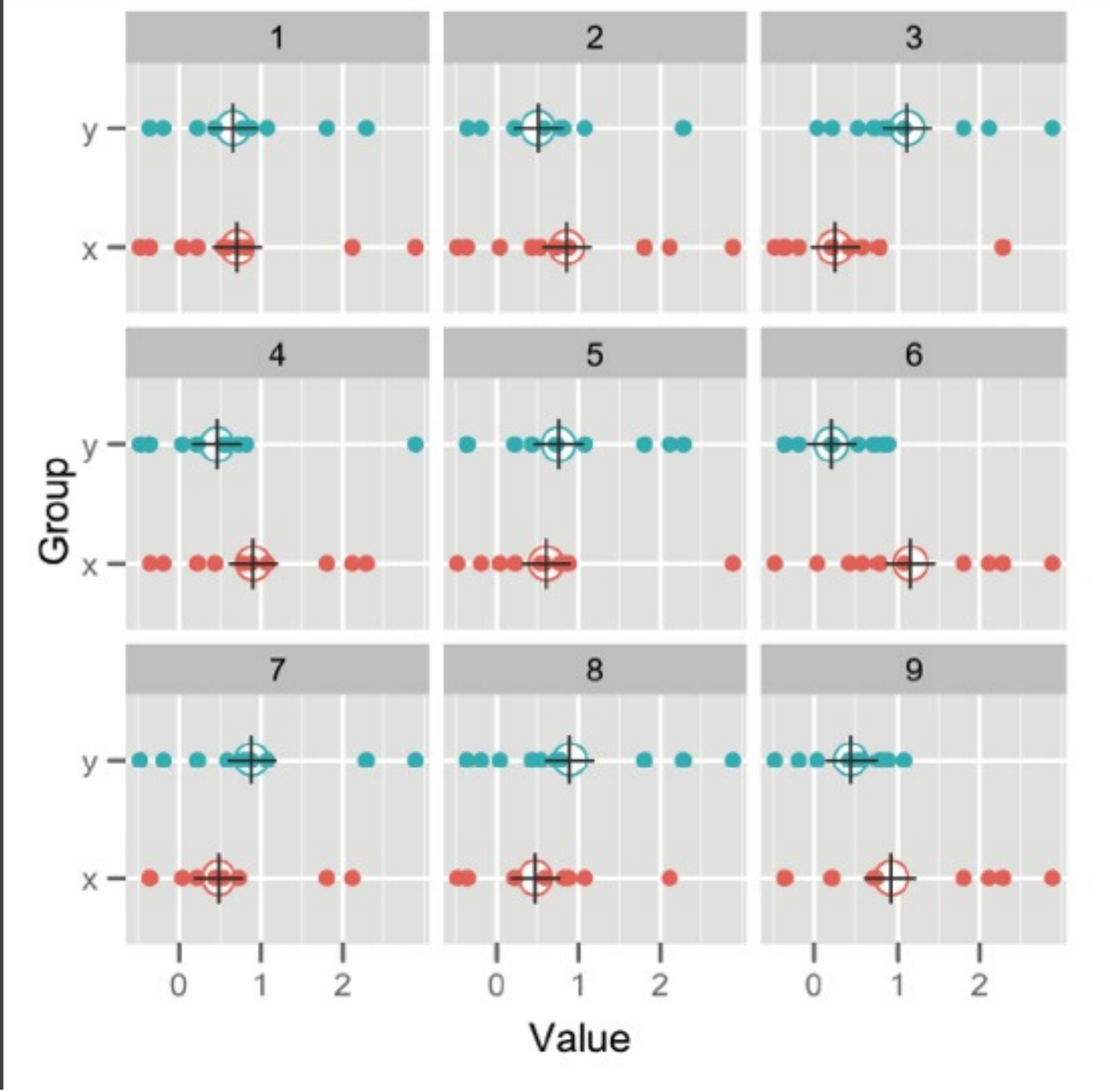
Choropleth maps of cancer deaths in Texas.

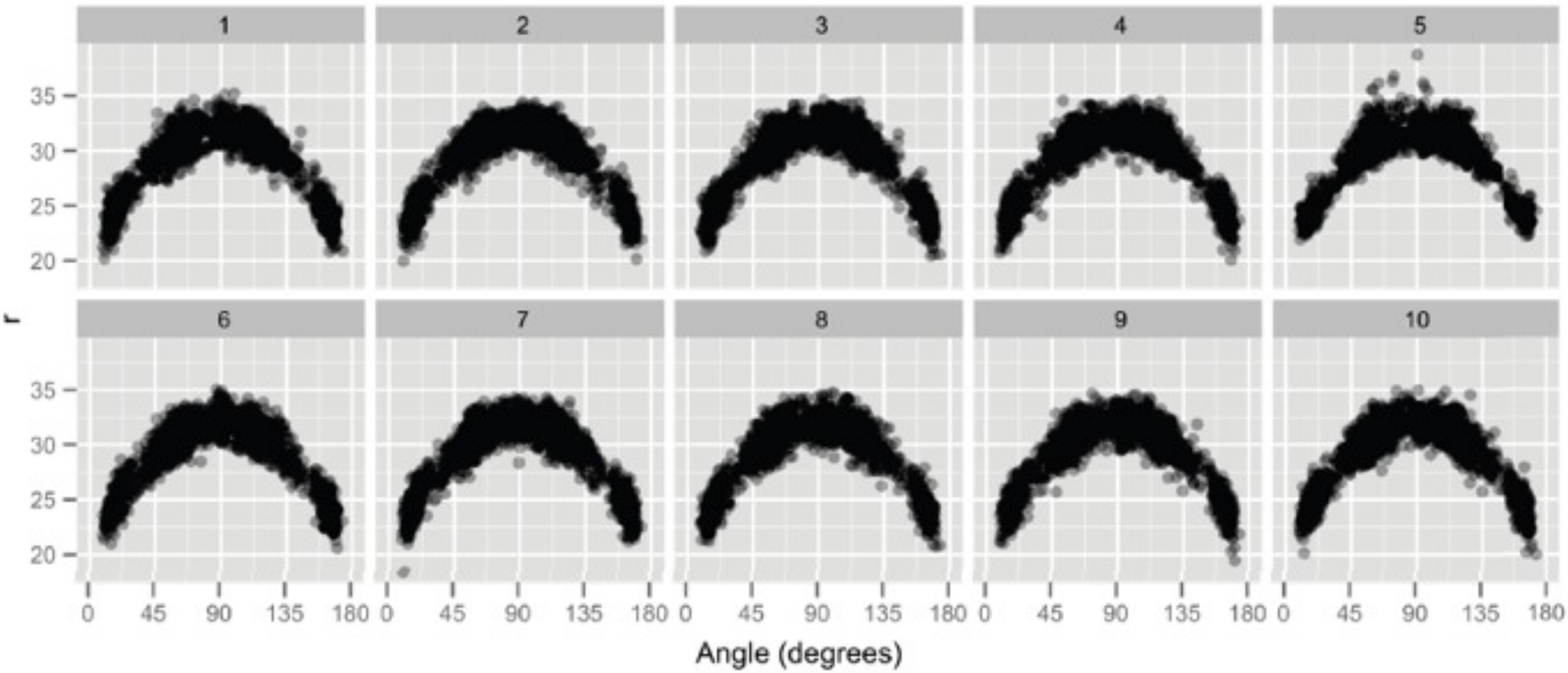
One plot shows a real data set. The others are simulated under the null hypothesis of spatial independence.

Can you spot the real data? If so, you have some evidence of spatial dependence in the data.



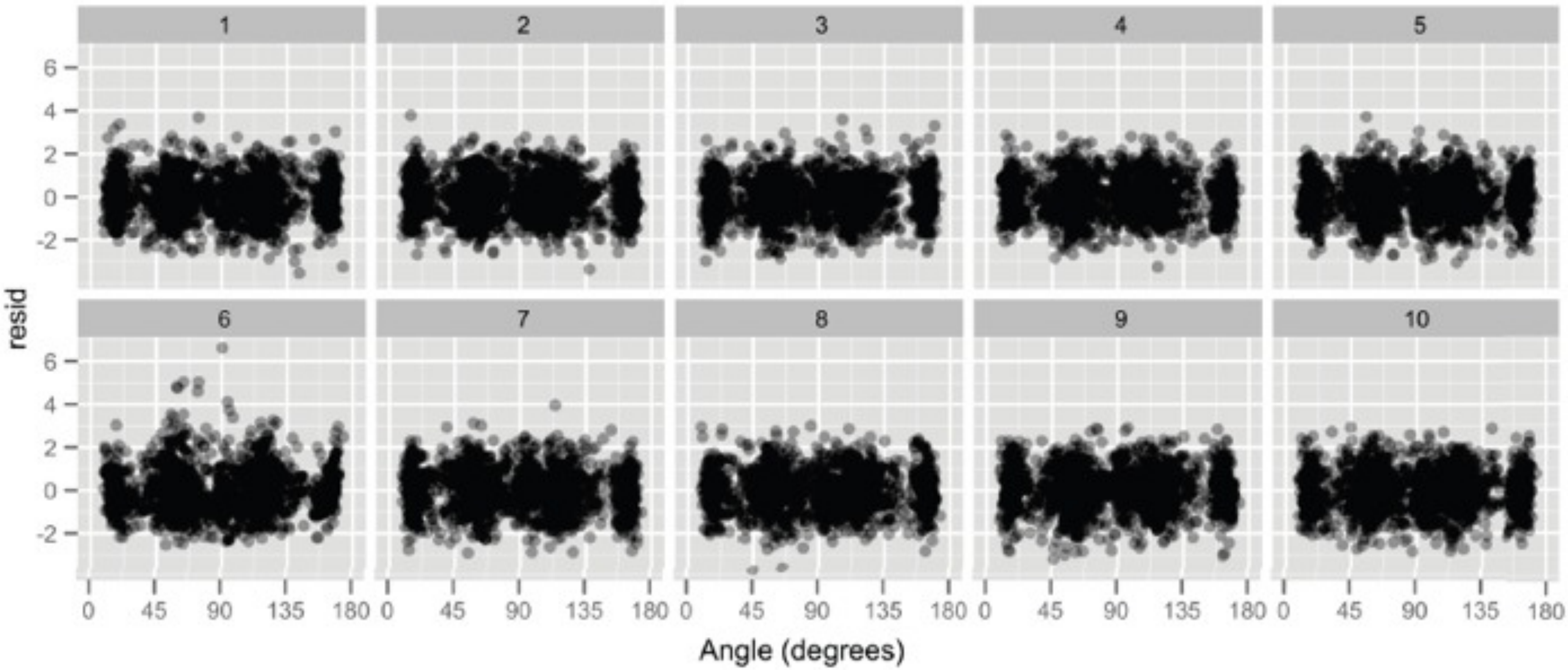
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
20000
15000
10000
Housing
200 400 600 800
Climate-Terrain


1
2
3
4
5
6
7
8
9
y
x
Group
0
1
2
Value



Distance vs. angle for 3 point shots by the LA Lakers.

One plot is the real data. The others are generated according to a null hypothesis of quadratic relationship.



Residual distance vs. angle for 3 point shots.

One plot is the real data. The others are generated using an assumption of normally distributed residuals.

Summary

Exploratory analysis may combine graphical methods, data transformations, and statistics.

Use questions to uncover more questions.

Formal methods may be used to confirm, sometimes on held-out or new data.

Visualization can further aid assessment of fitted statistical models.

Extra Material

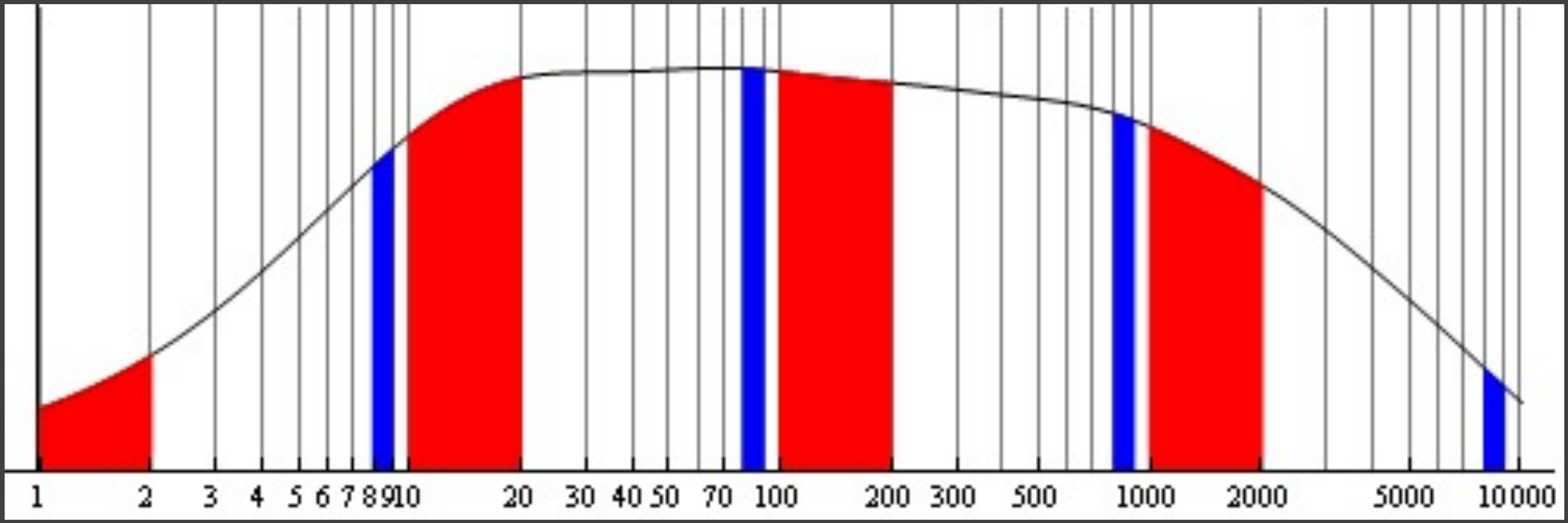
A Detective Story

You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

Firm A		***Firm B***	**LIARS!**
283.08	25.23	283.08	75.23
153.86	385.62	353.86	185.25
1448.97	12371.32	5322.79	9971.42
18595.91	1280.76	8795.64	4802.43
21.33	257.64	61.33	57.64
Amt. Paid:	$34823.72	Amt. Rec'd:	$29908.67

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.



Hence the leading digit **1** has a ~30% likelihood. Larger digits are increasingly less likely.

Benford's Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Holds for many (but certainly not all) real-life data sets: Addresses, Bank accounts, Building heights, ...

Data must span multiple orders of magnitude.

Evidence that records do not follow Benford's Law is admissible in a court of law!

Model-Driven Data Validation

Deviations from the model *may* represent errors

Find Statistical Outliers

- # std dev, Mahalanobis dist, nearest-neighbor, non-parametric methods, time-series models
- *Robust statistics* to combat noise, masking

Data Entry Errors

- Product codes: PZV, PZV, PZR, PZC, PZV
- Which of the above is most likely in error?

Opportunity: combine with visualization methods