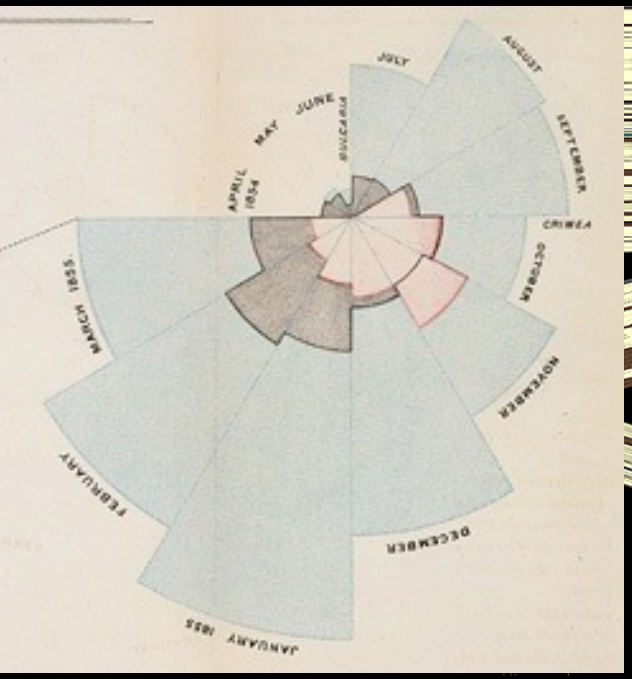
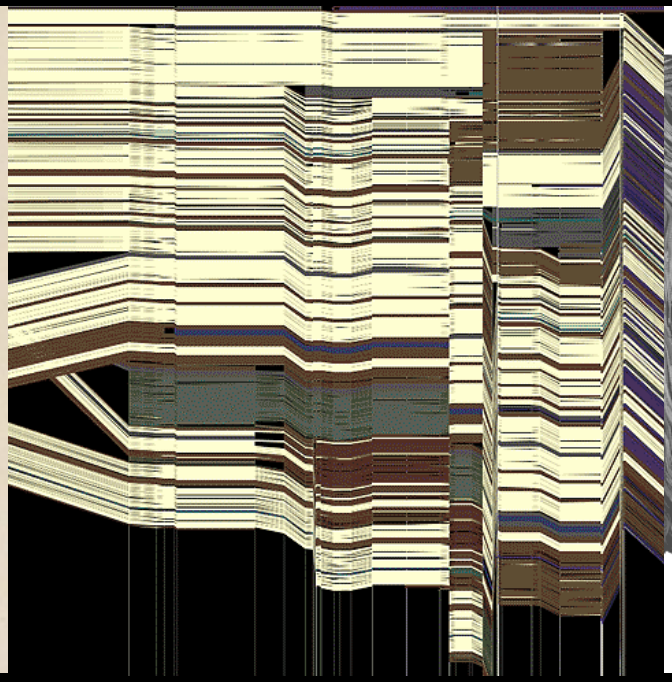
CSE 412 - Intro to Data Visualization

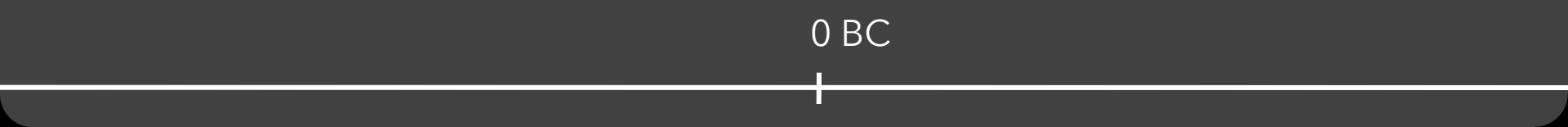
Exploratory Data Analysis





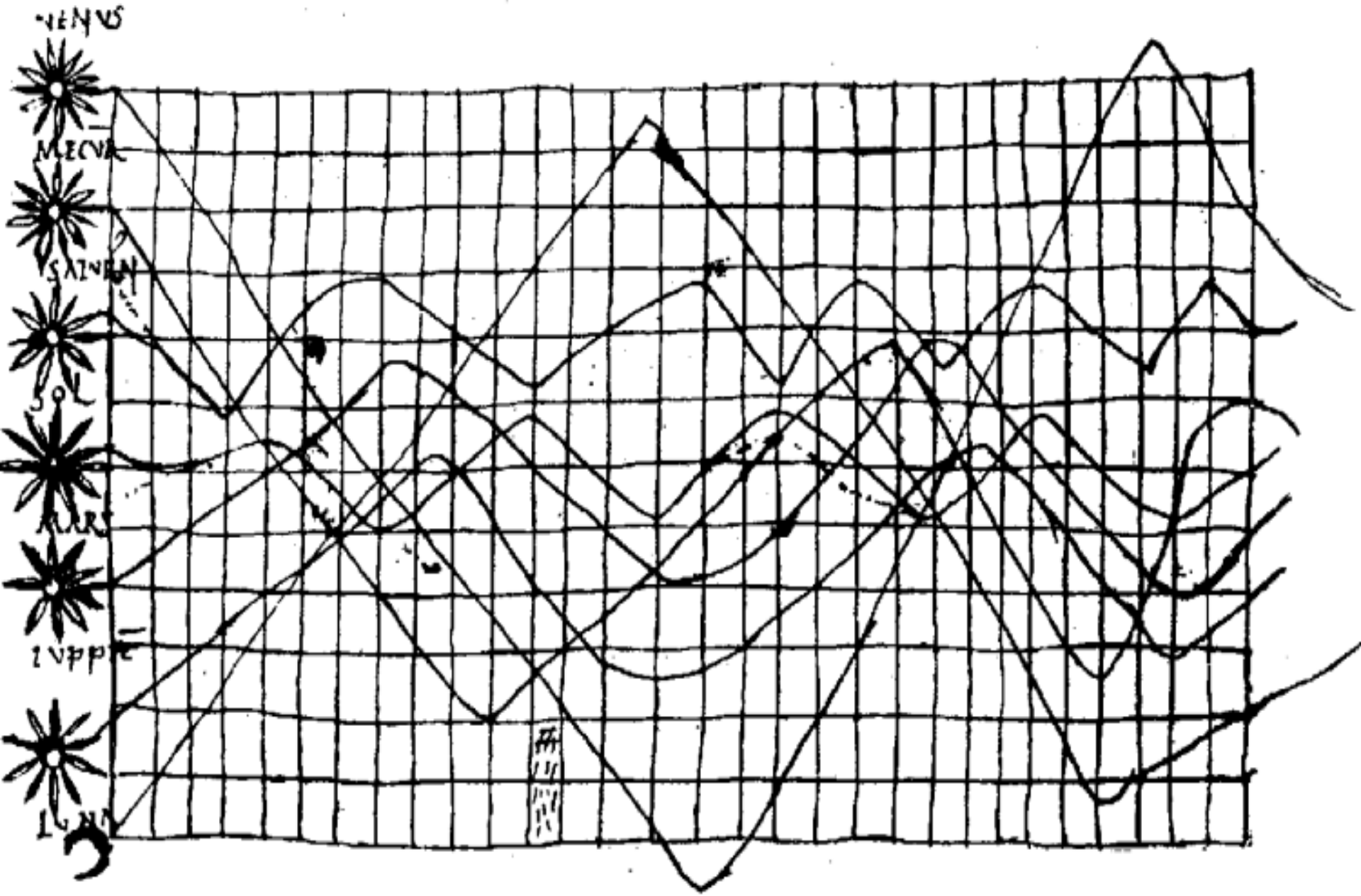
Jane Hoffswell University of Washington

What was the **first** data visualization?

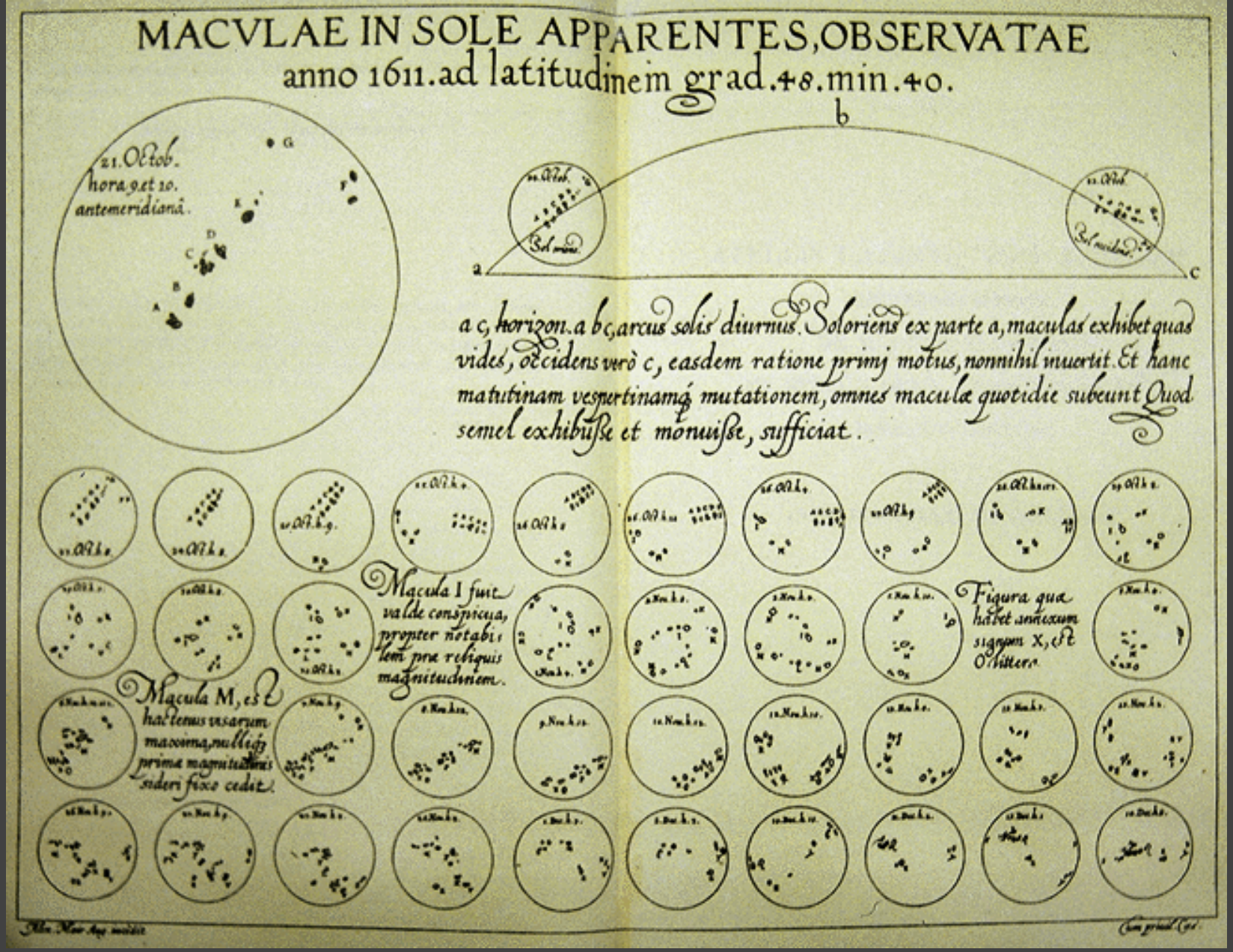


~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey

0 BC

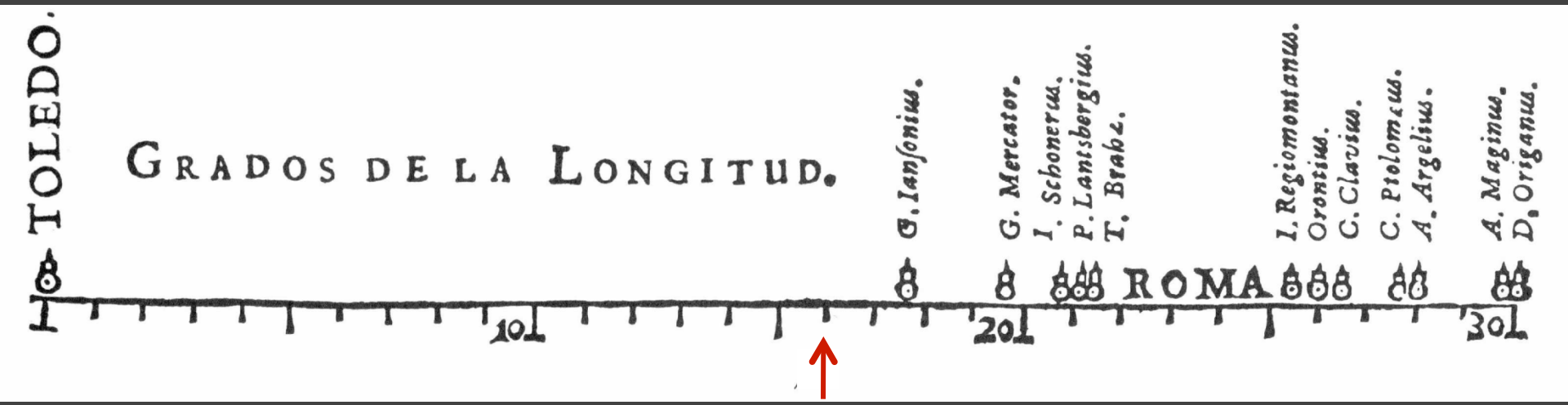


~950 AD Position of Sun, Moon and Planets

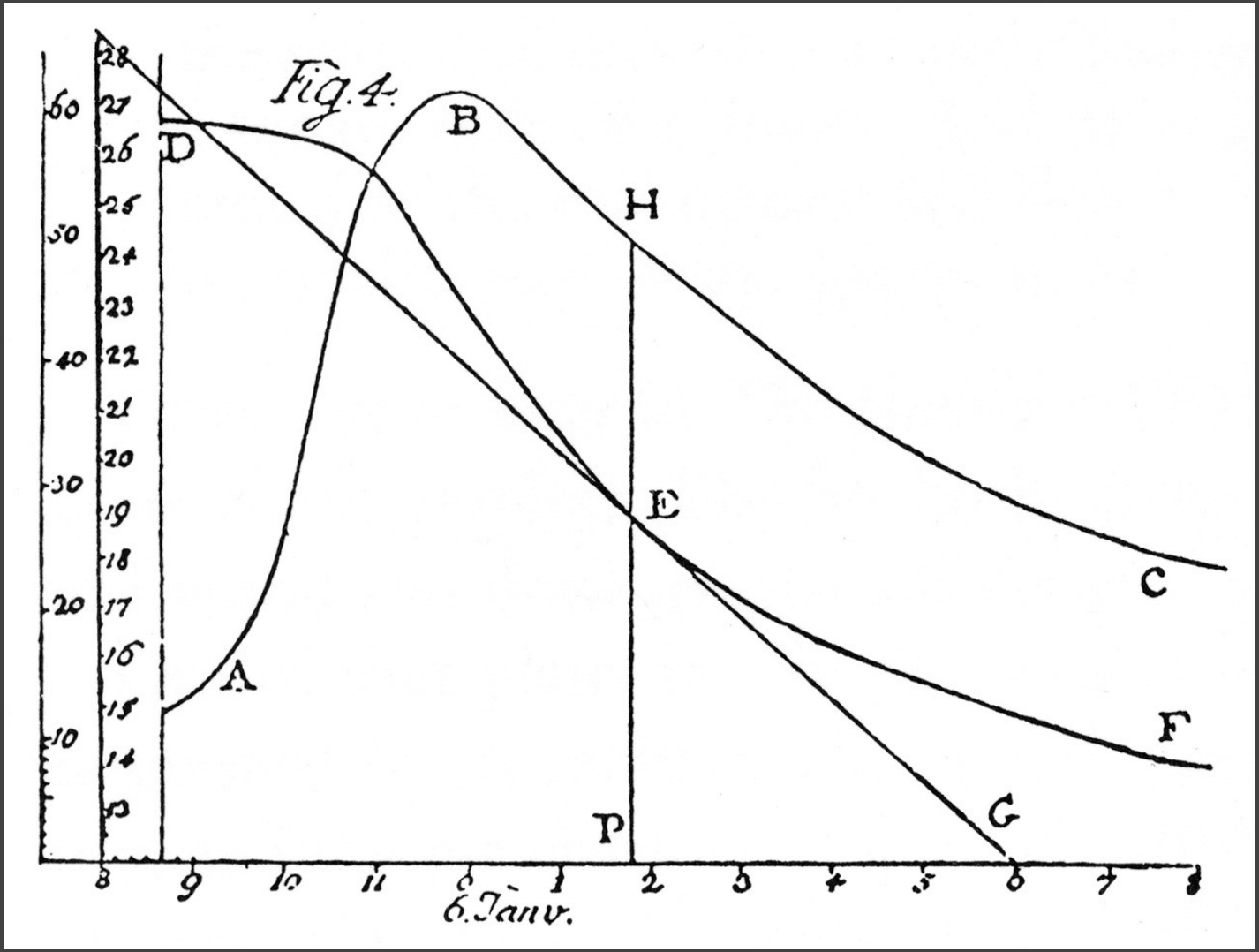


0 BC

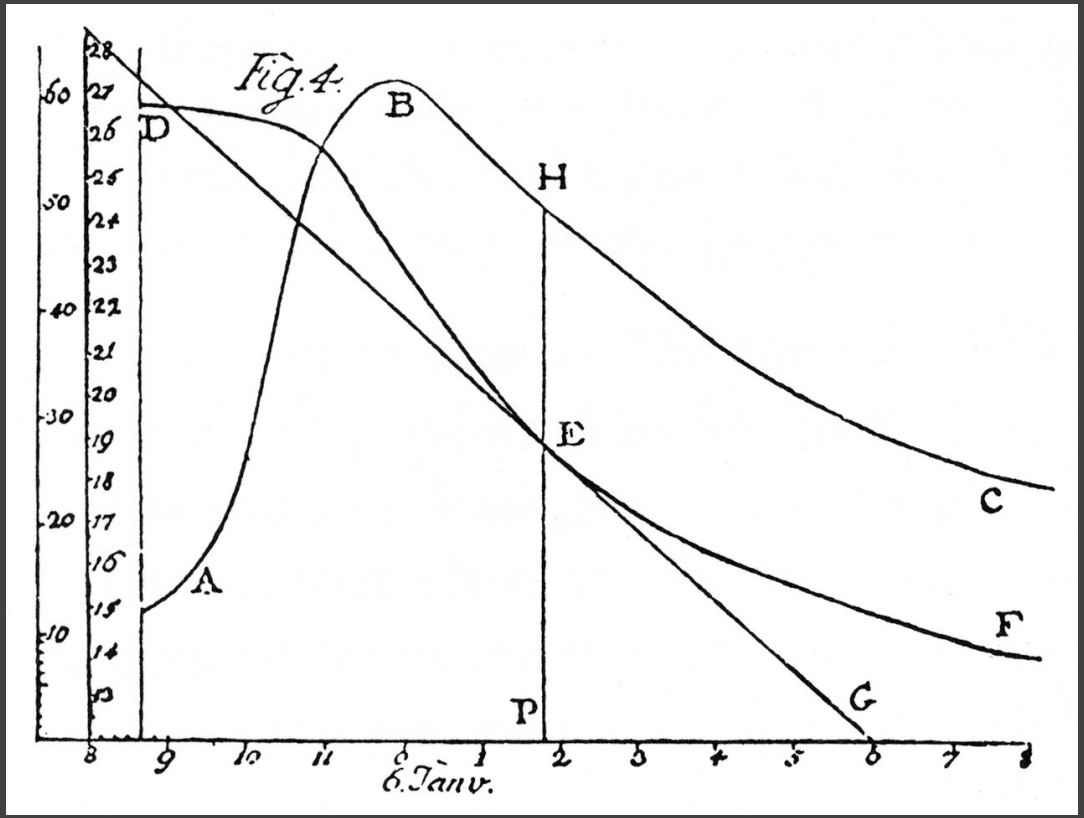
Sunspots over time, Scheiner 1626

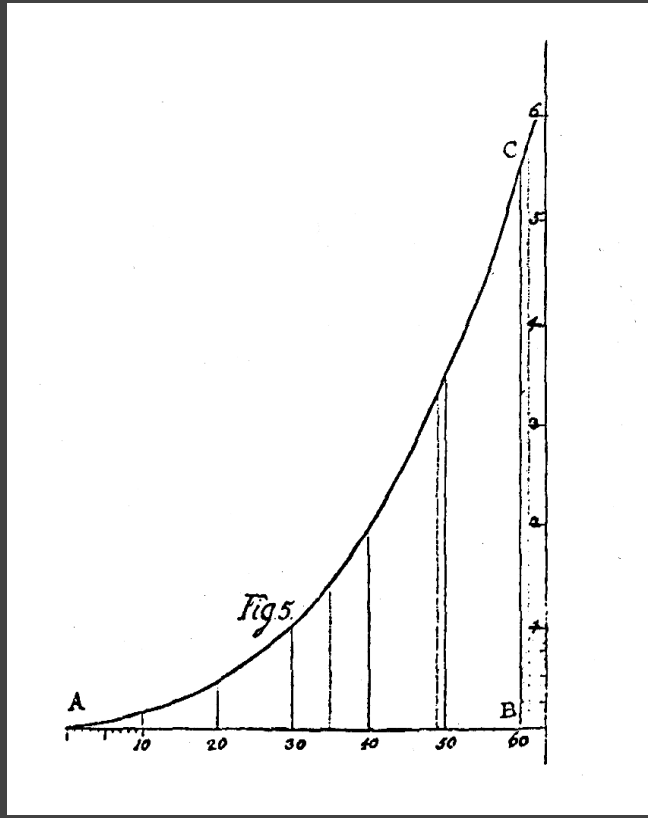


Longitudinal distance between Toledo and Rome, van Langren 1644



The Rate of Water Evaporation, Lambert 1765

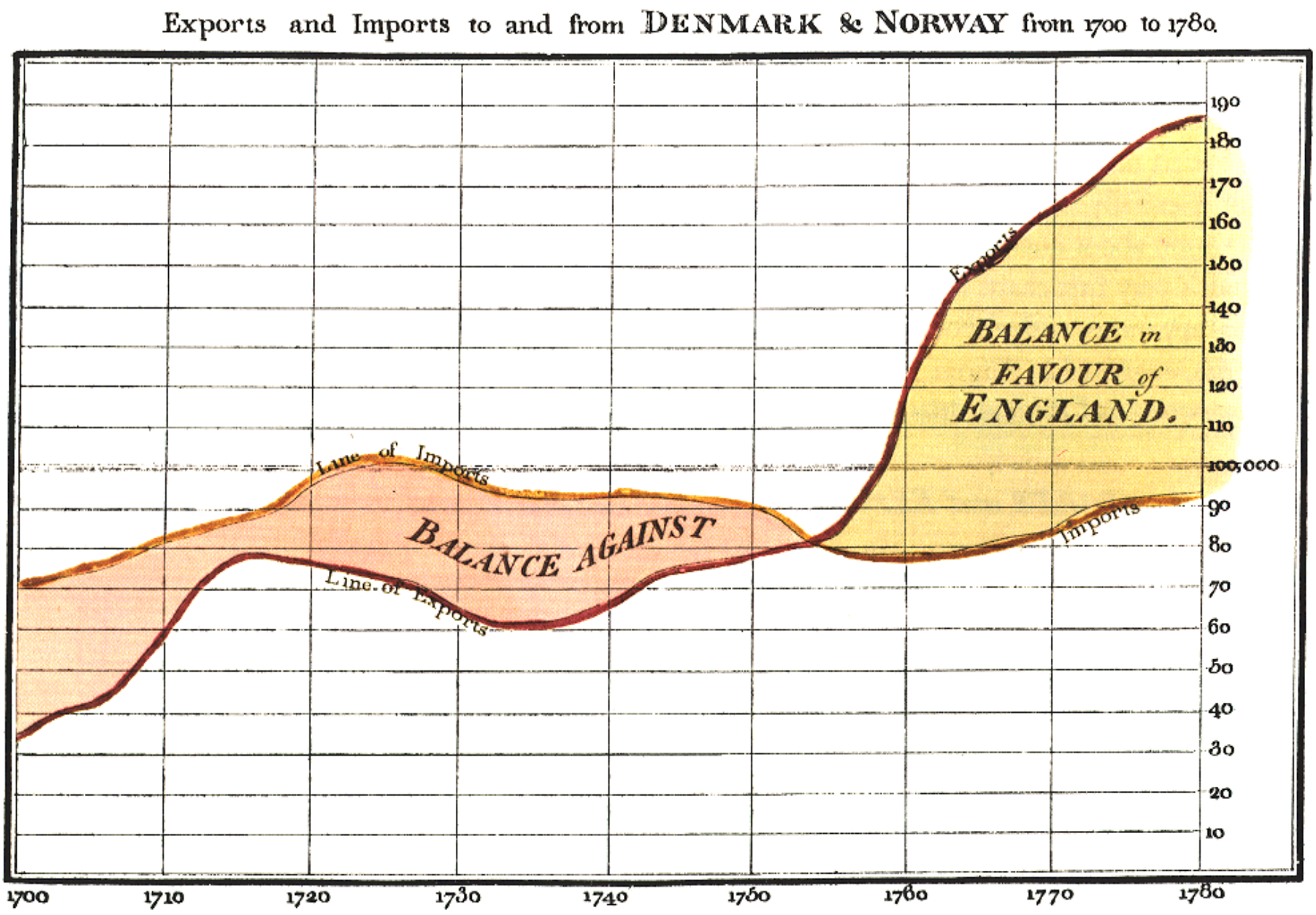




The Rate of Water Evaporation, Lambert 1765

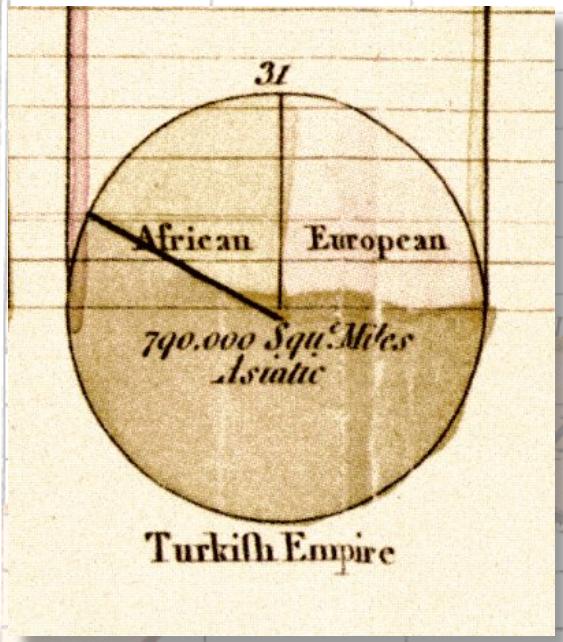
The **Golden Age** of Data Visualization

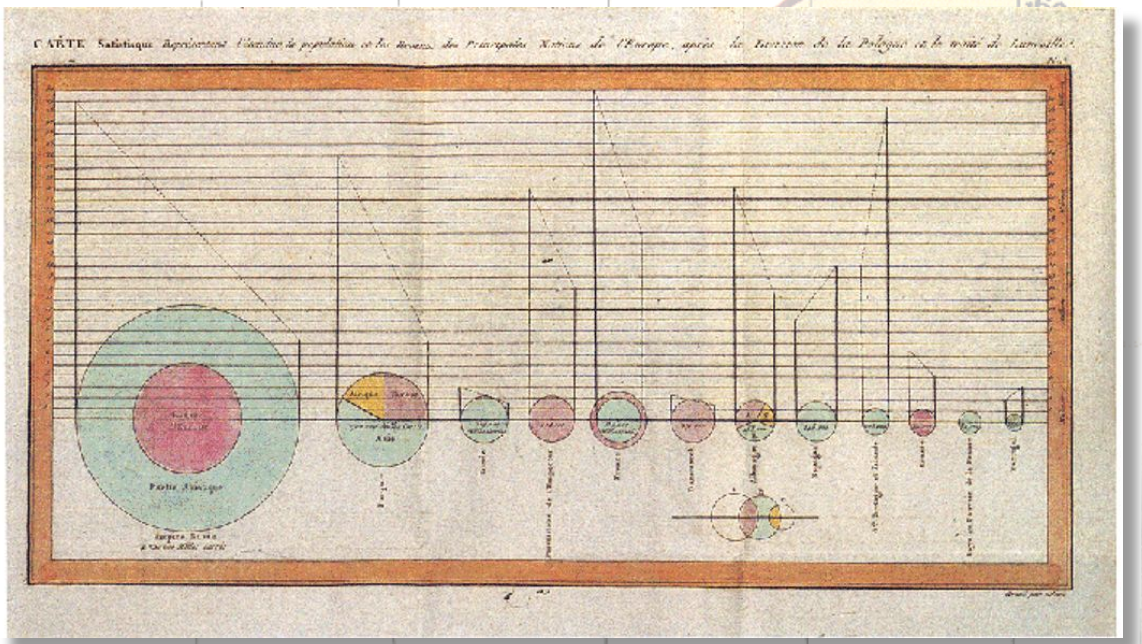
1786 1900



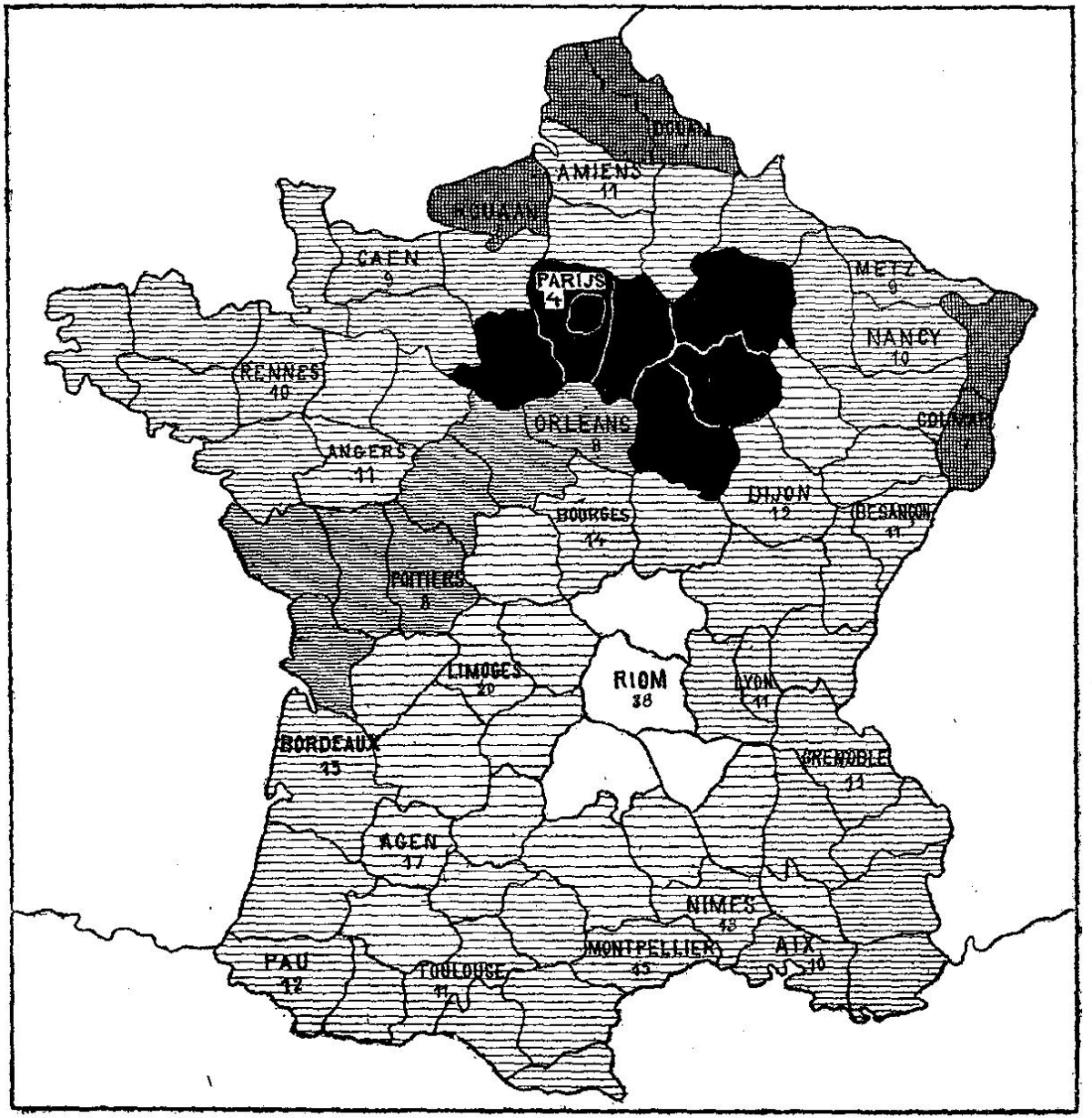
The Commercial and Political Atlas, William Playfair 1786

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



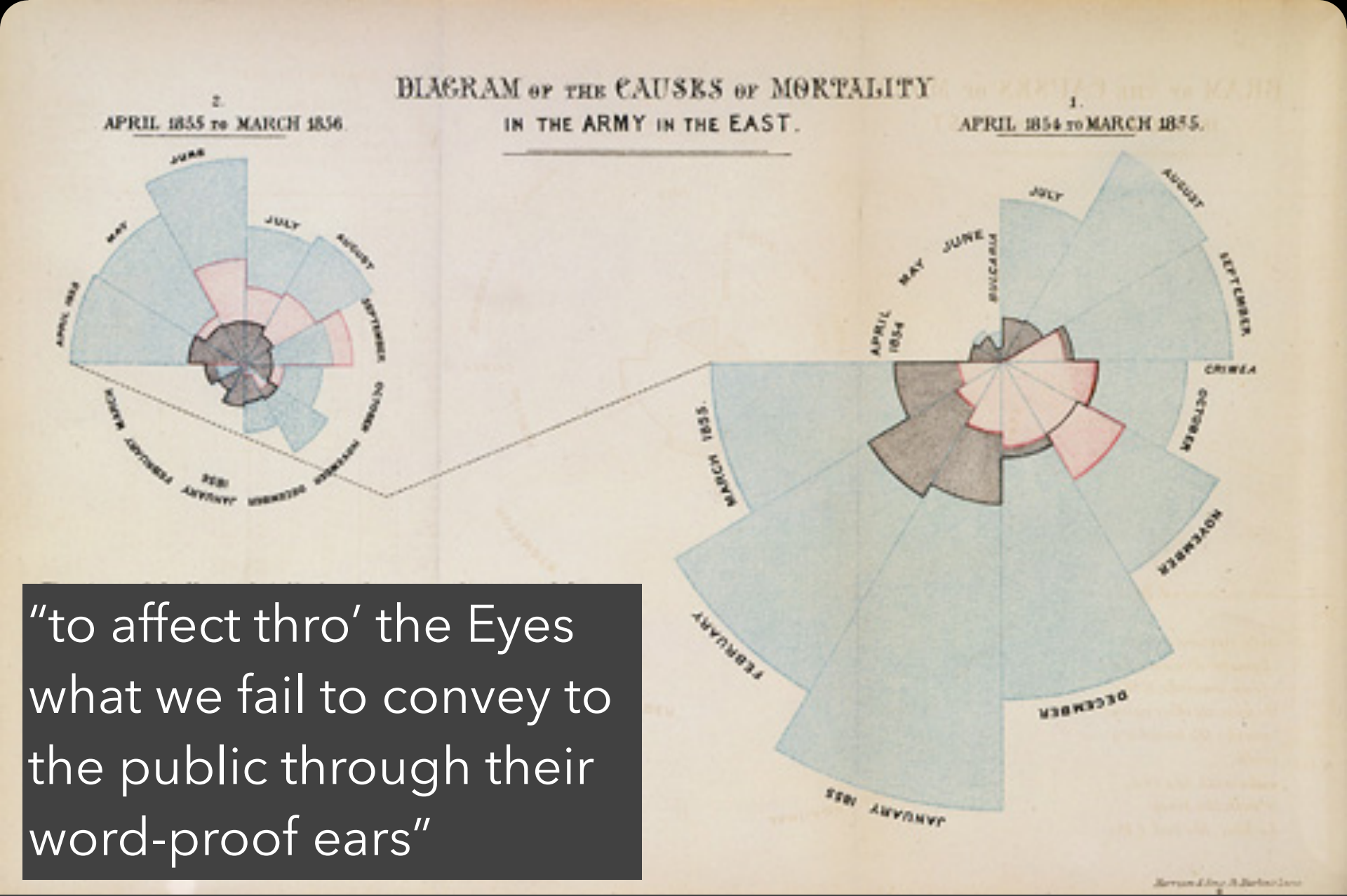


Statistical Breviary, William Playfair 1801

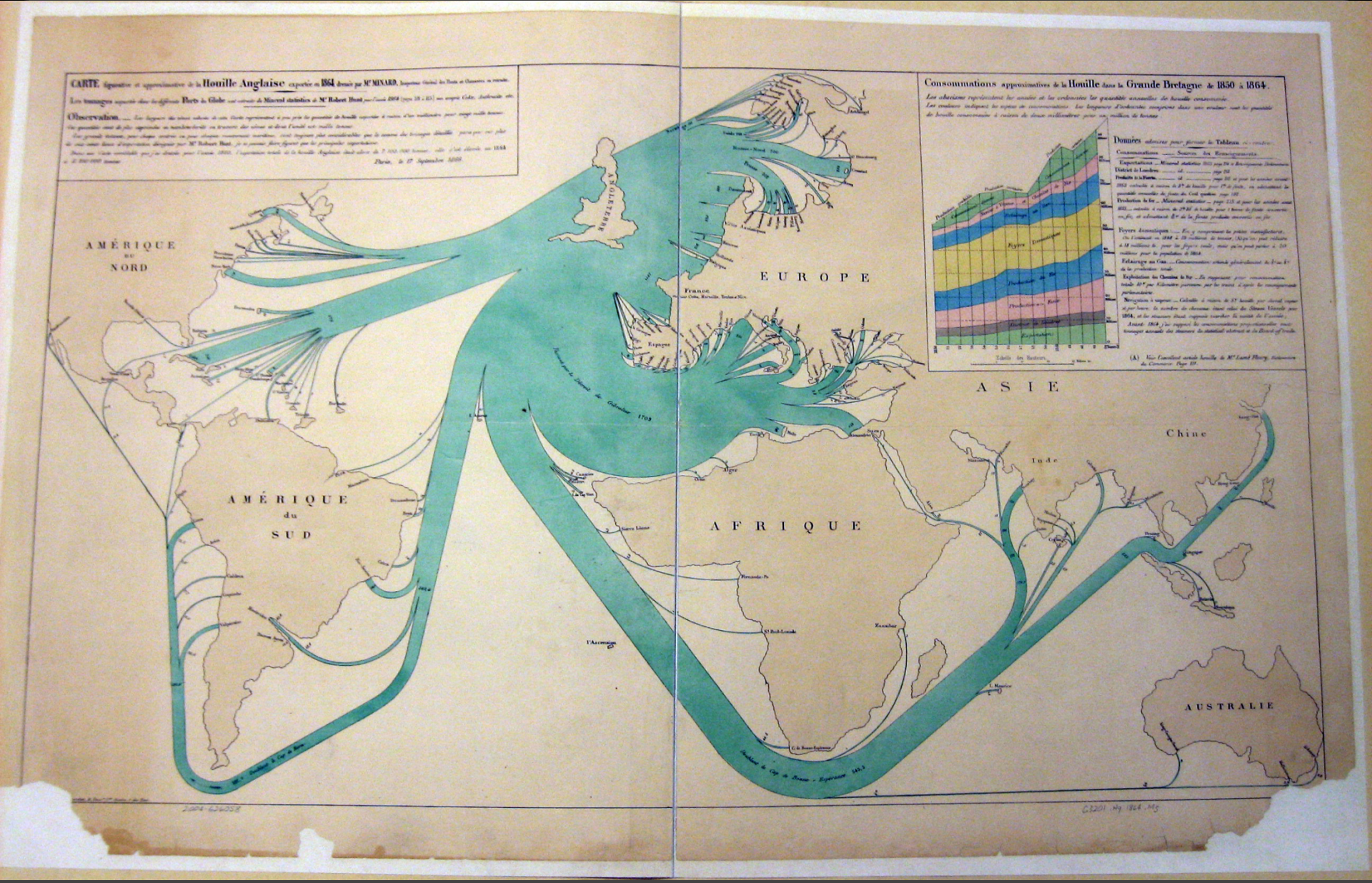


1786

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"

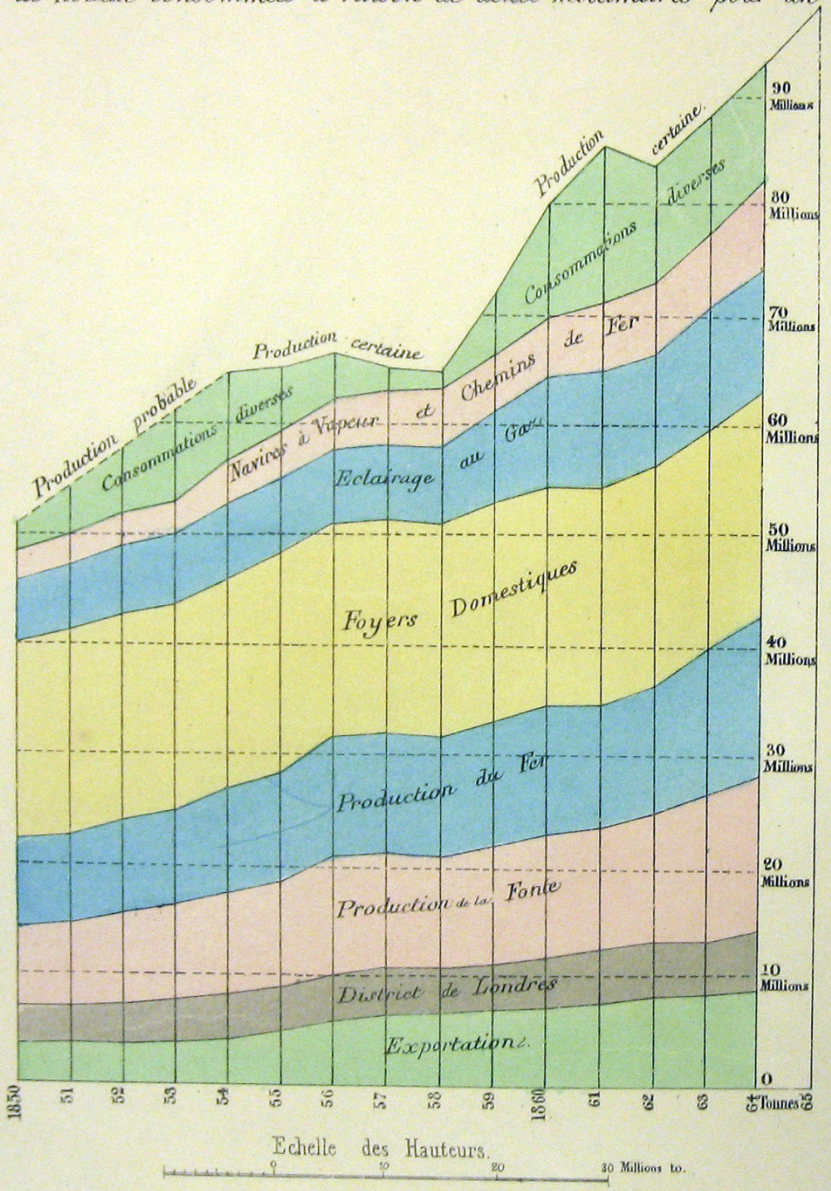


1786

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 to de houille pour 1 to 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 to 35 de houille pour 1 tonne de fonte convertie en fer, et admettant $\frac{2}{10}$ es de la fonte produite convertis en fer.*

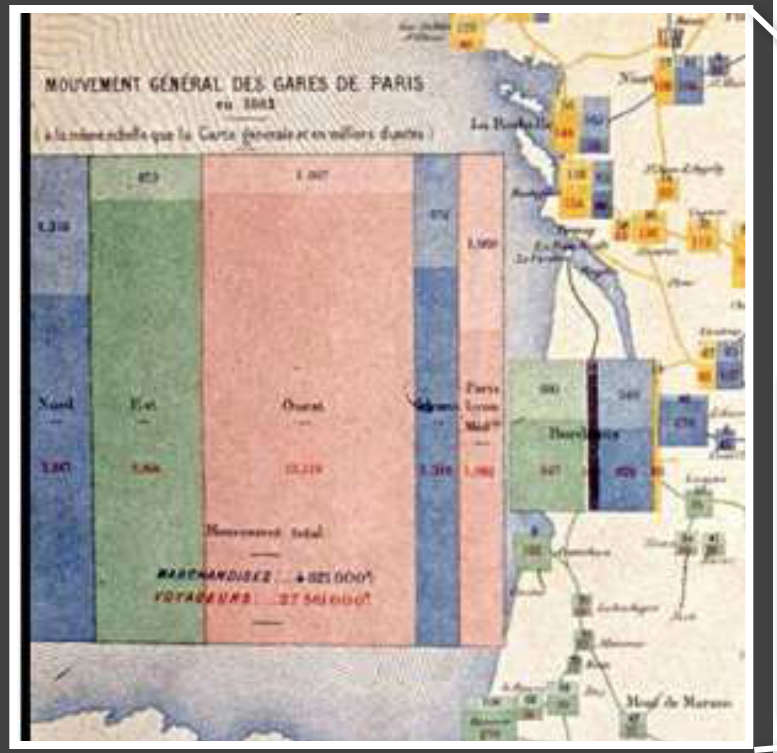
Foyers domestiques. — *En y comprenant les petites manufactures. On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions to. pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.*

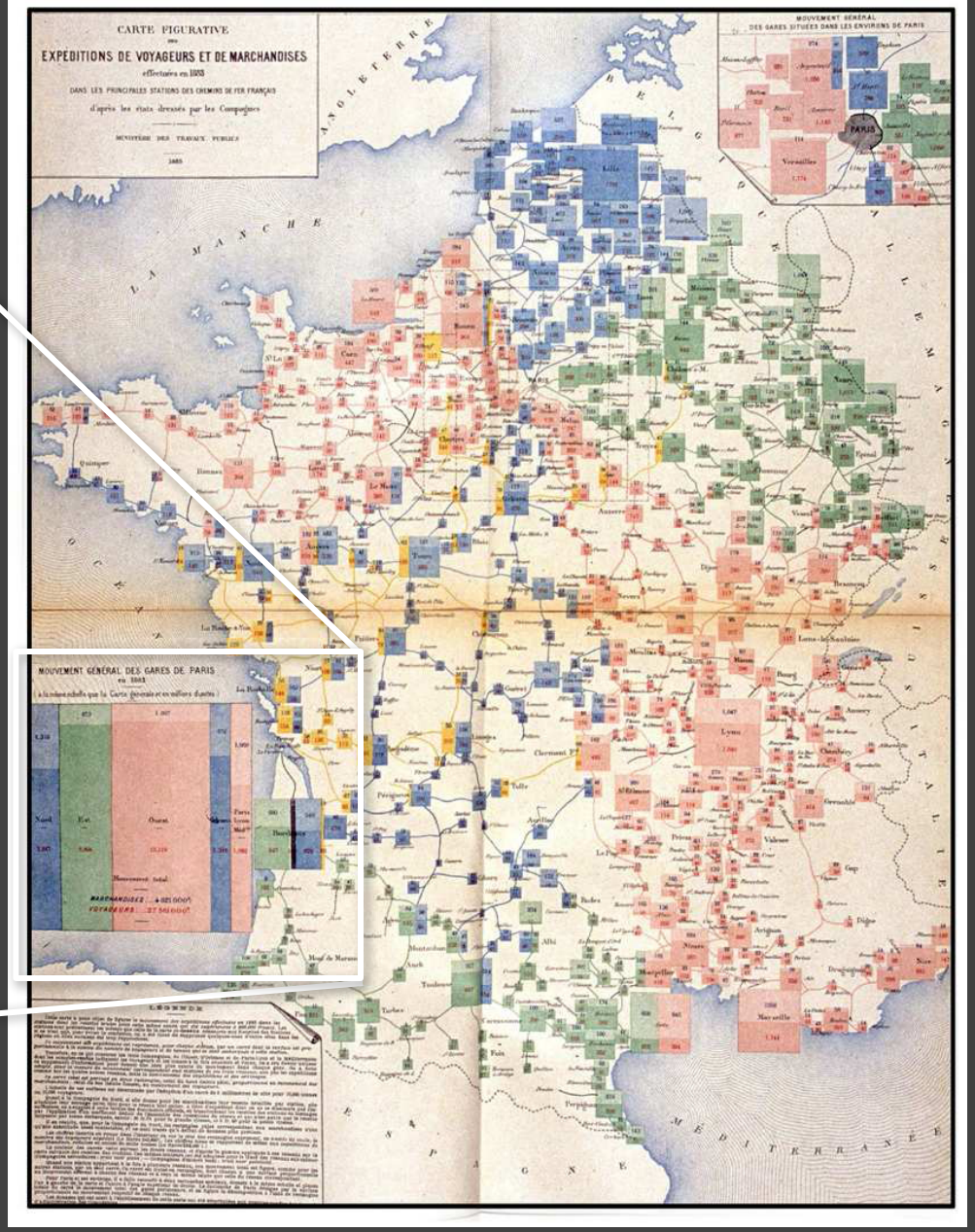
Eclairage au Gaz. — *Consommation estimée généralement du $\frac{1}{3}$e au $\frac{1}{8}$e de la production totale.*

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.*

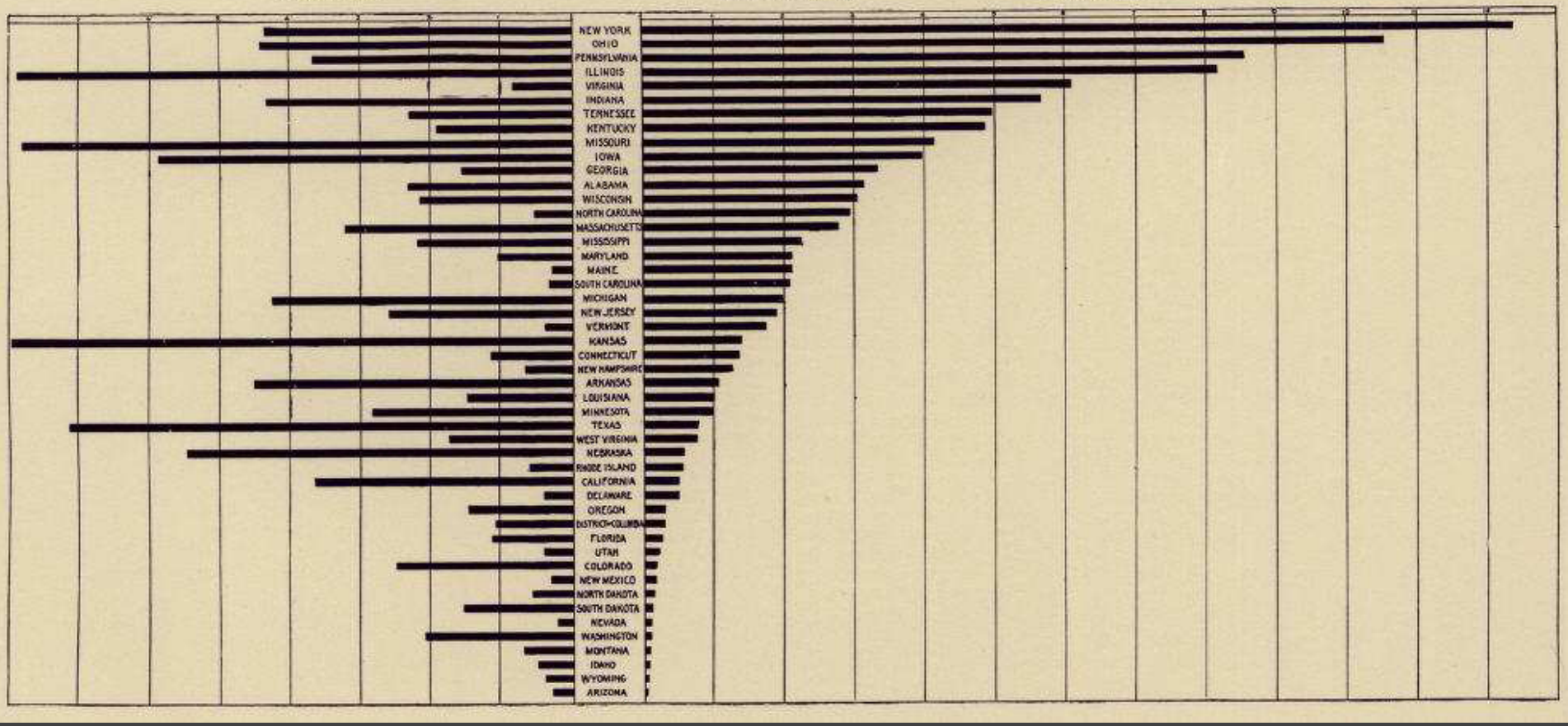




1786

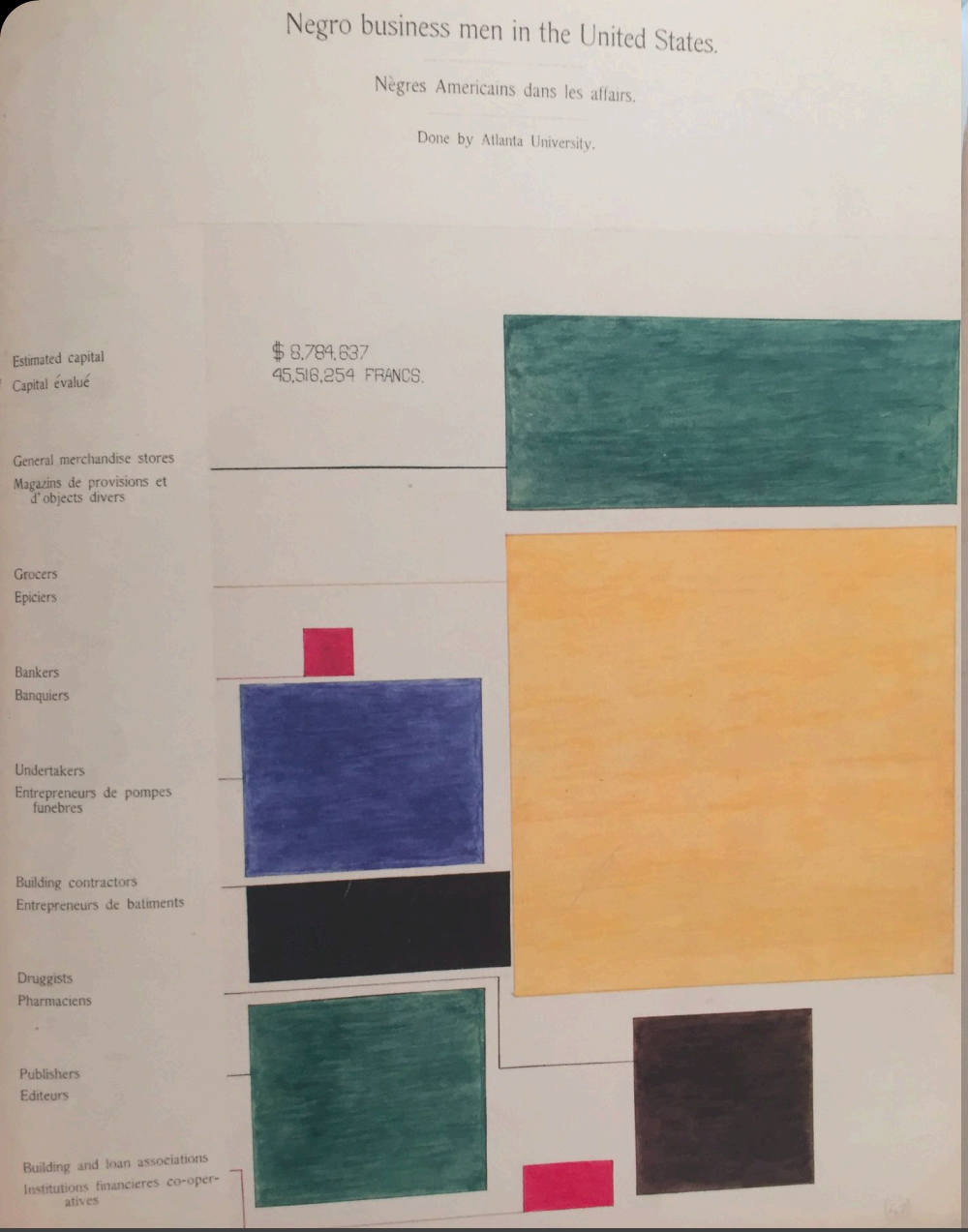
1884 Rail Passengers and Freight from Paris

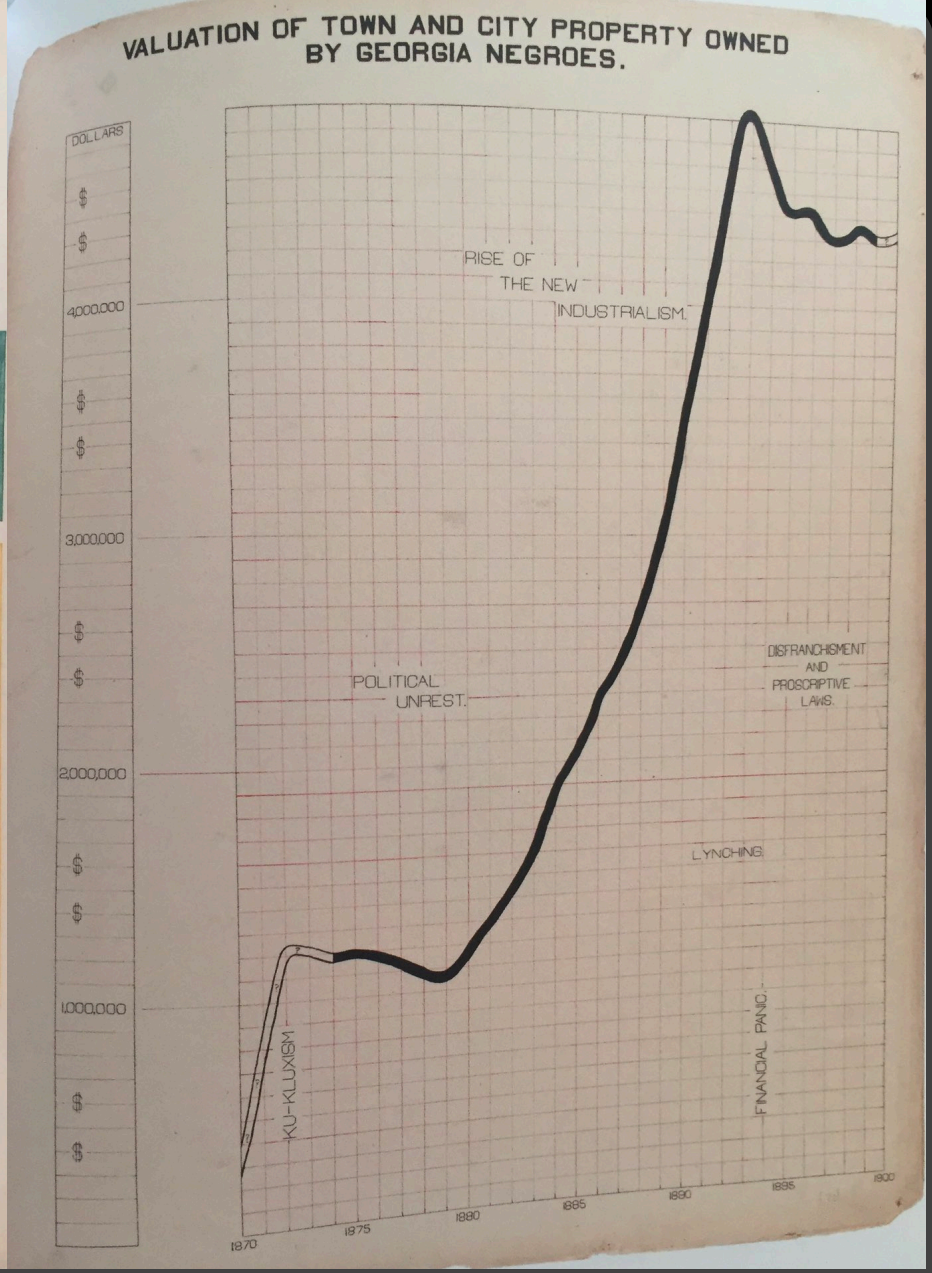
66. Interstate Migration—Number of Native Immigrants and Native Emigrants, by States and Territories: 1890.



1786

1890 Statistical Atlas of the Eleventh U.S. Census





1786

1900 Visualizing Black America , W. E. B. DuBois et al.

The Rise of Statistics

1786 1900 1950

Rise of **formal statistical methods** in the physical and social sciences

Little innovation in graphical methods

A period of **application and popularization**

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


LIFE

Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.

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.

Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind**.

Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention**.

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

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

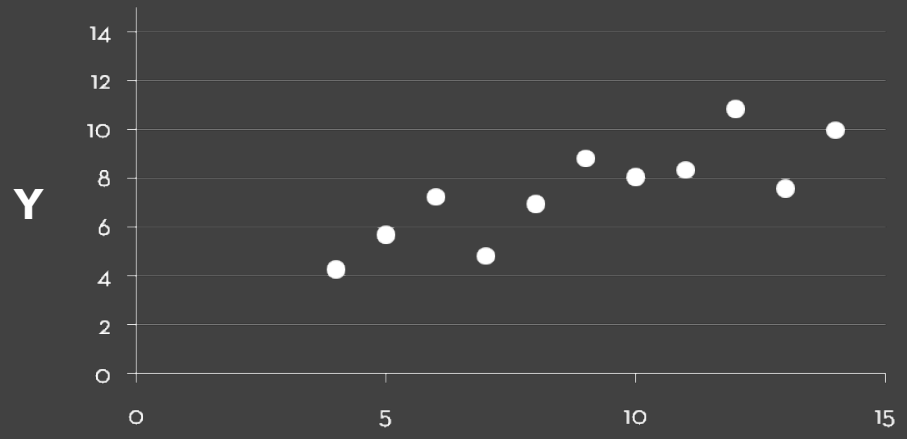
Linear Regression

$Y = 3 + 0.5\ X$

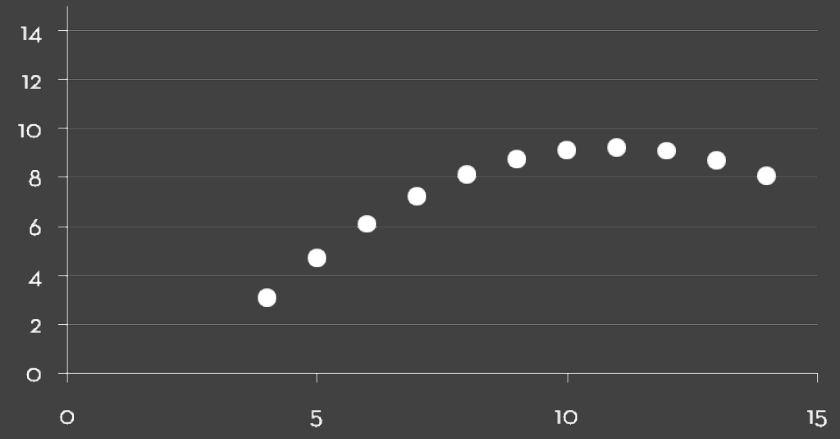
$R^2 = 0.67$

[Anscombe 1973]

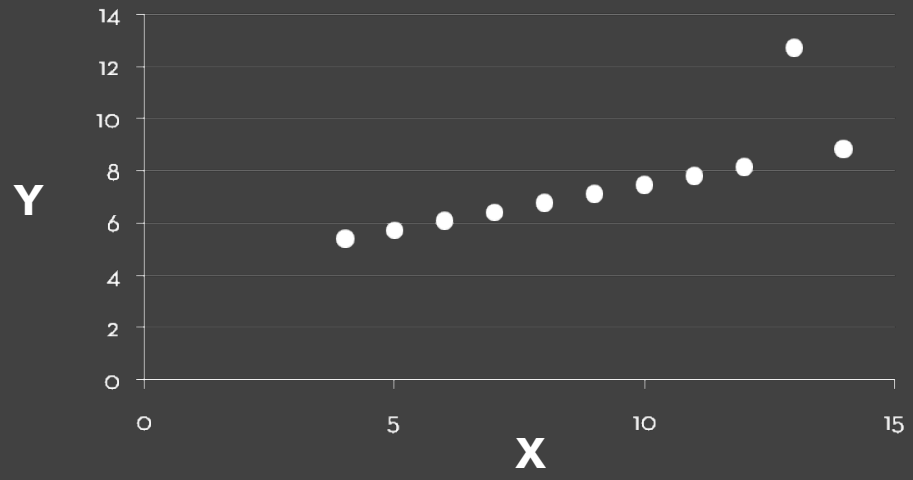
Set A



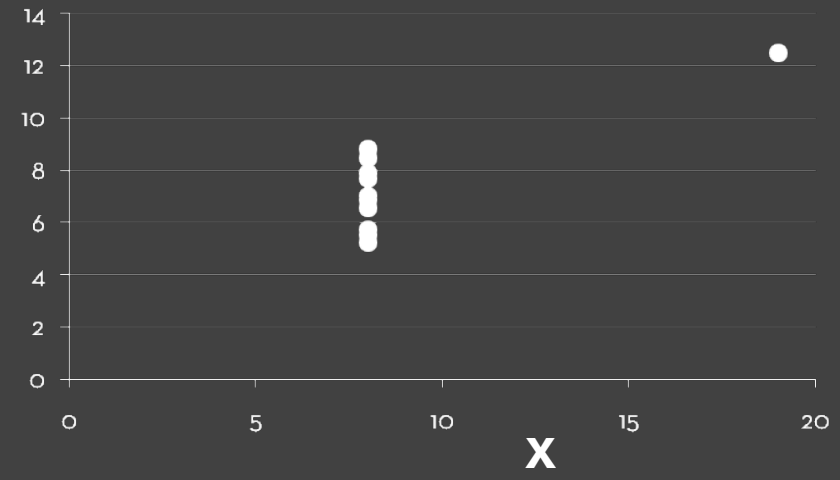
Set B



Set C



Set D



[Anscombe 1973]

Administrivia

A1: Visualization Design

Pick a **guiding question**, use it to title your vis.
Design a **static visualization** for that question.
You are free to **use any tools** (inc. pen & paper).

Deliverables (upload via Canvas; see A1 page)
Image of your visualization (PNG or JPG format)
Short description + design rationale (≤ 4 paragraphs)

Due **tonight** by **11:59 pm PST, Monday Jan. 11**.

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

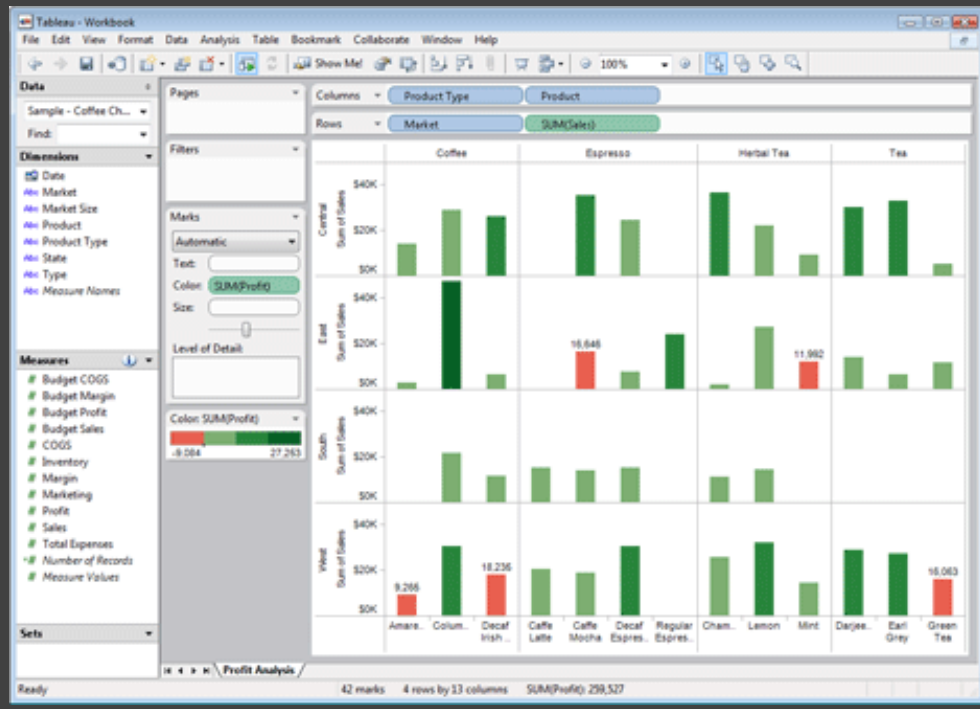
First steps:

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

Author a report

Screenshots of most insightful views *(8+)*

Include titles and captions for each view

Due by 11:59pm

Monday, Jan 25

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

Step 1: Pick domain & data
You can analyze any dataset of your choice
Consider exploring a dataset for your Final Project

Step 2: Pose questions
Write down three initial questions

Step 3: Profile the data
Assess data quality, create visualizations, refine questions

Assignment description on course website
Example report as Observable document

Course Participation & Ed

Week 1 Discussion & Quiz due **today** by 11:59pm

Week 2 Discussion & Quiz posted this afternoon, due Monday Jan 18th, by 11:59pm PST

Final project team selection thread now open on Ed, post project ideas or see ideas from classmates

Data Wrangling

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any "analysis" at all.

Anonymous Data Scientist
[Kandel et al. '12]

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

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

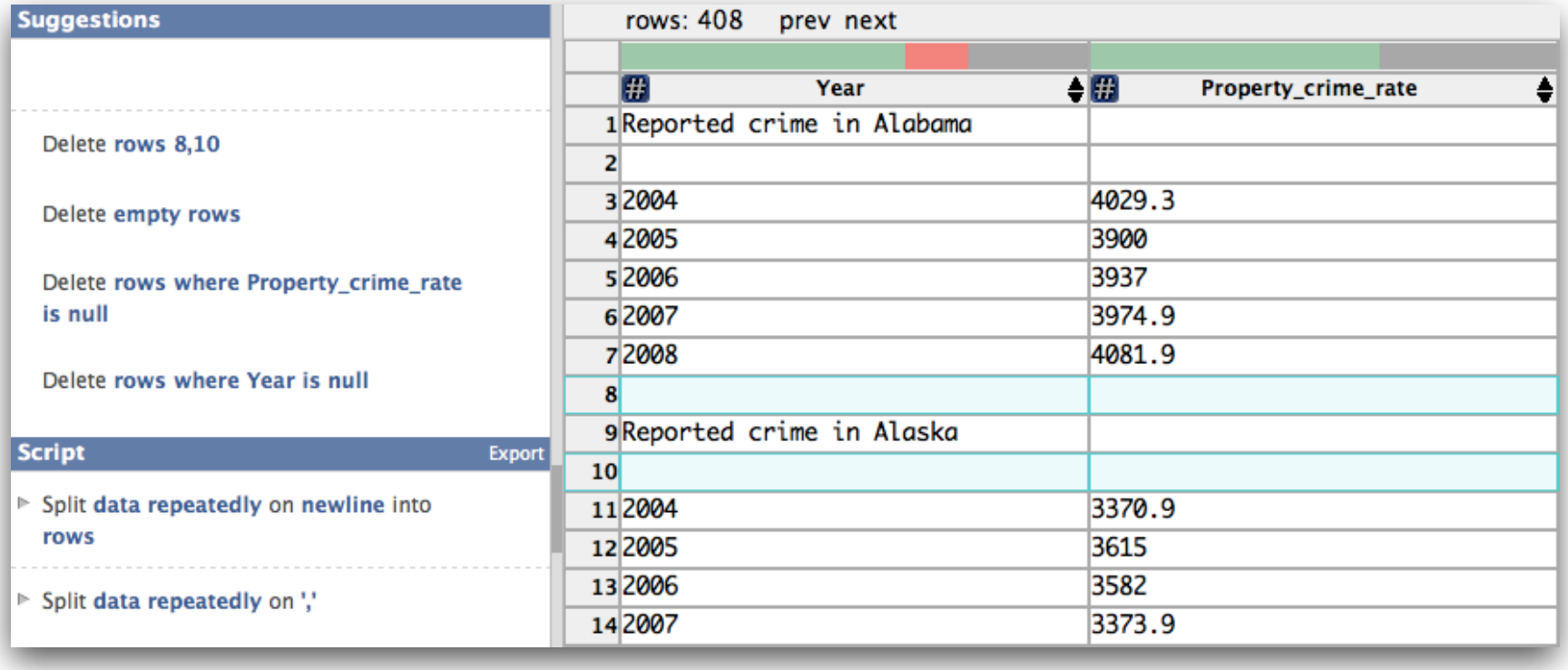
Reported crime in California

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

DataWrangler



Wrangler: Interactive Visual Specification of Data Transformation Scripts

Sean Kandel et al. *CHI'11*

Data Wrangling

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

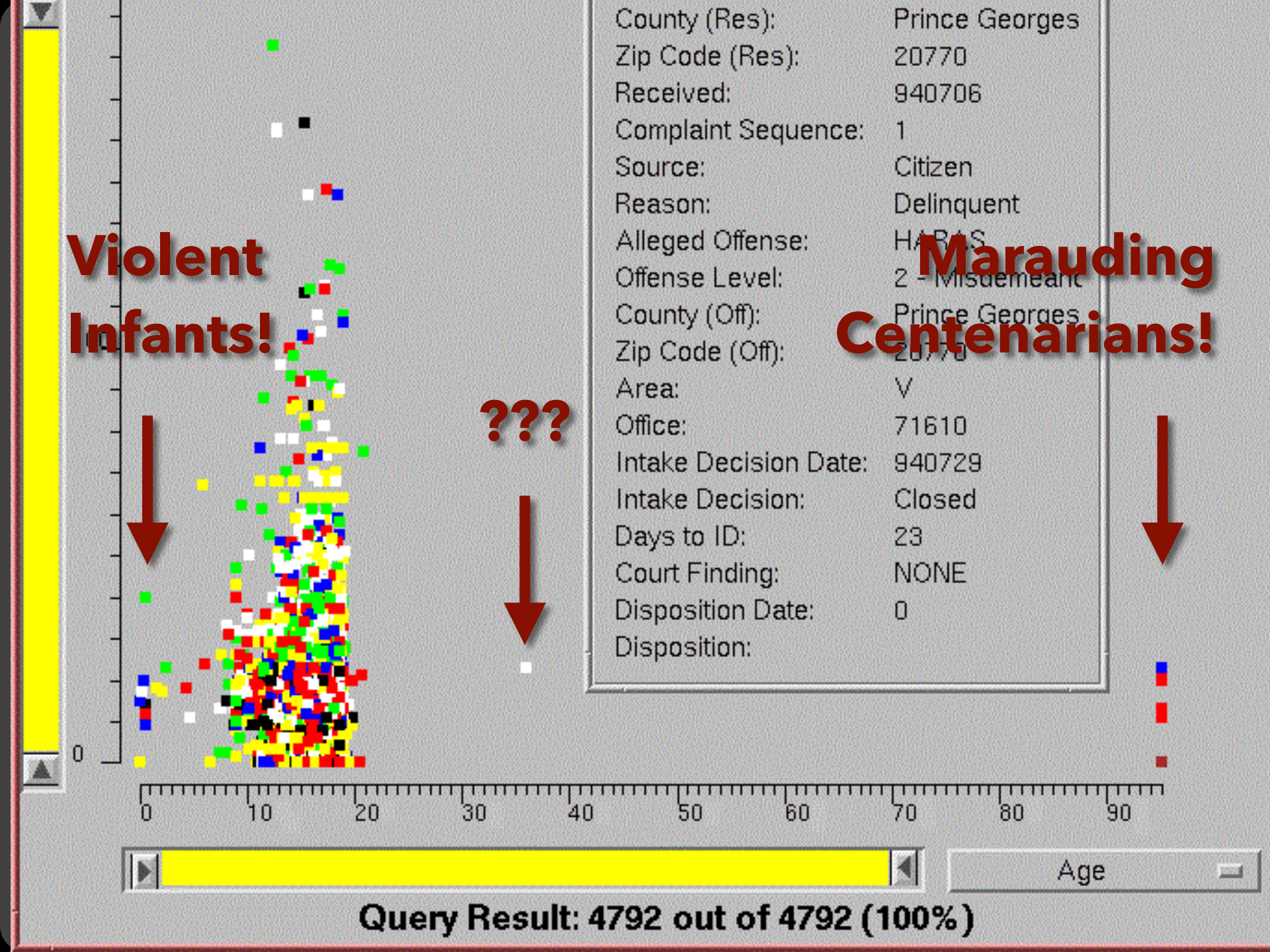
Approaches include:
Manual manipulation in spreadsheets
Code: arquero (JS), dplyr (R), pandas (Python)
Trifacta Wrangler http://www.trifacta.com/products/wrangler/
Open Refine http://openrefine.org/

Data Quality

“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


Violent Infants!
???
Marauding Centenarians!
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanr
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:
0
0 10 20 30 40 50 60 70 80 90
Age
Query Result: 4792 out of 4792 (100%)

Graph Viewer

Graph Viewer

Roll-up by:

All

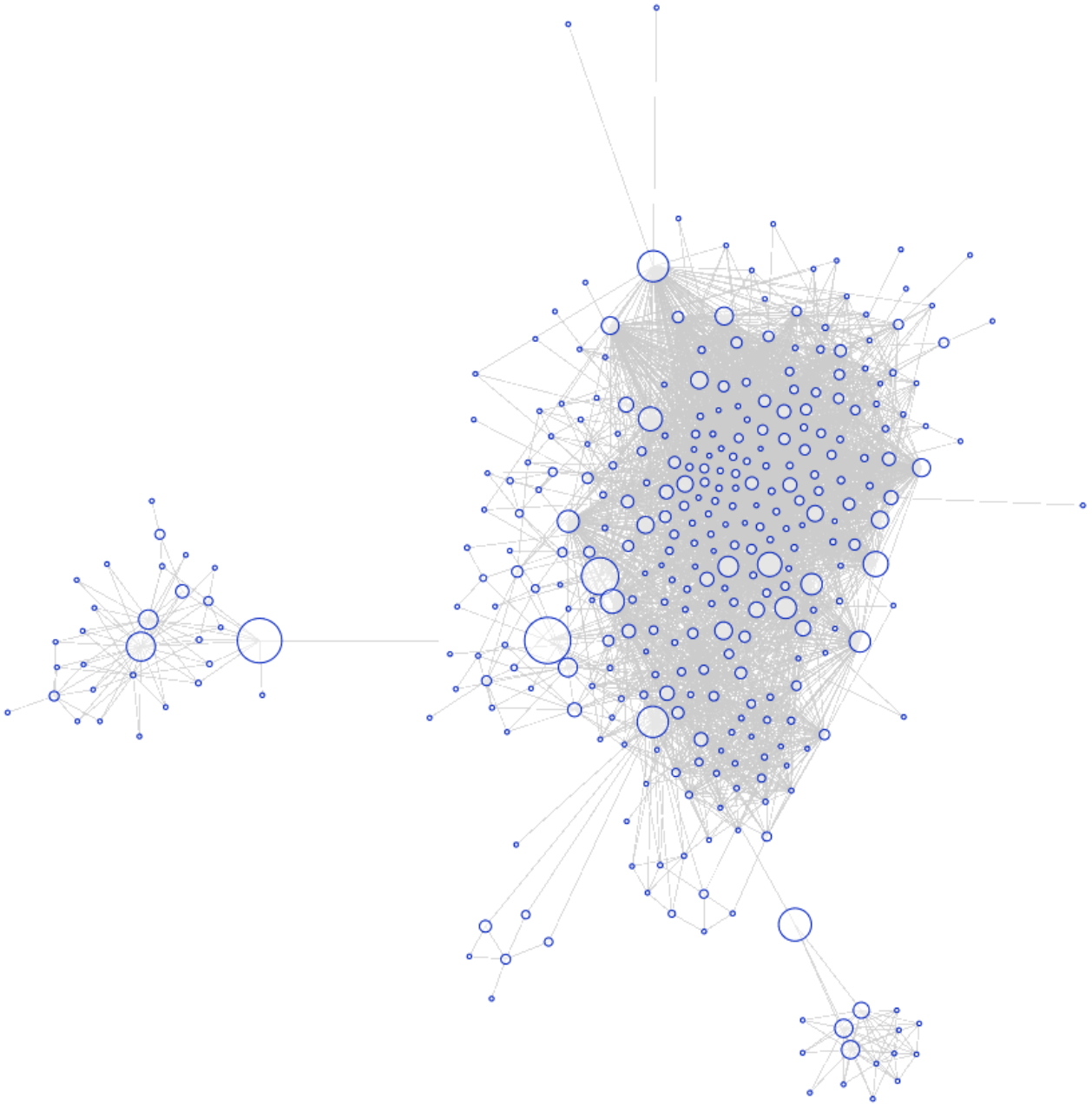
Visualization:

Node-Link

Sort by:

None

Edge centrality filters:



Images

Animate

Graph Viewer

Graph Viewer

Roll-up by:

All

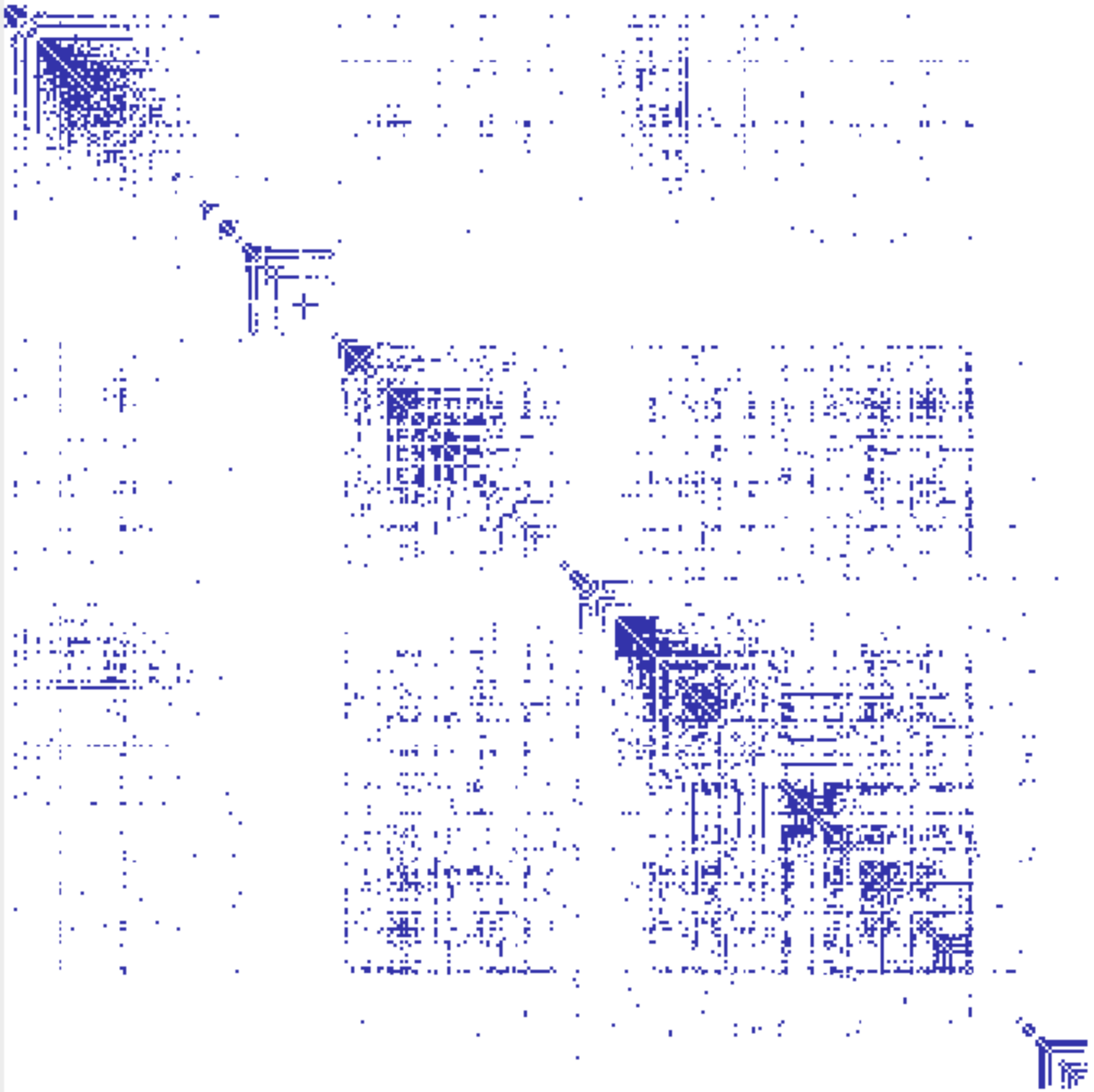
Visualization:

Matrix

Sort by:

Linkage

Edge centrality filters:



Graph Viewer

Graph Viewer

Roll-up by:

All

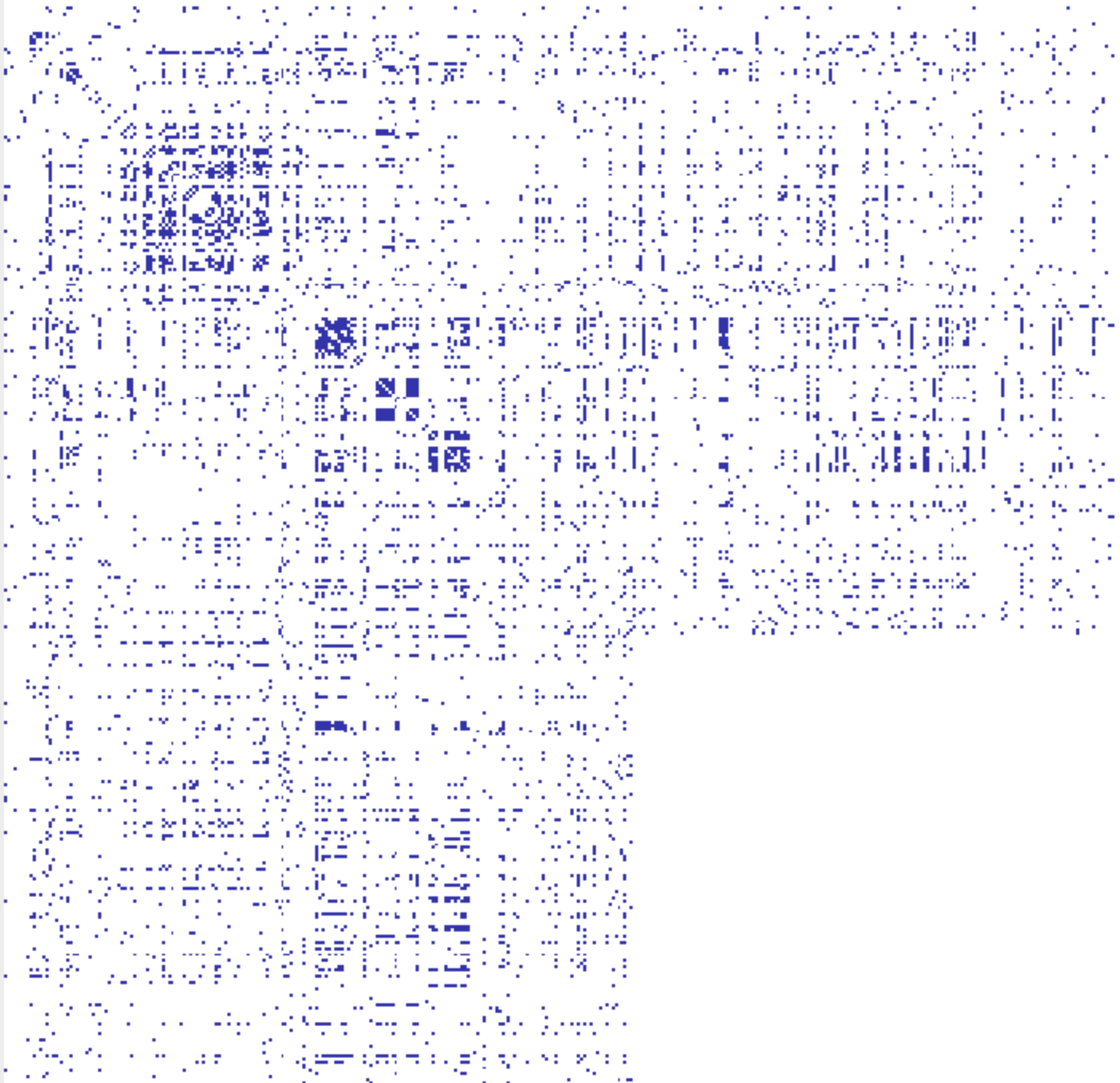
Visualization:

Matrix

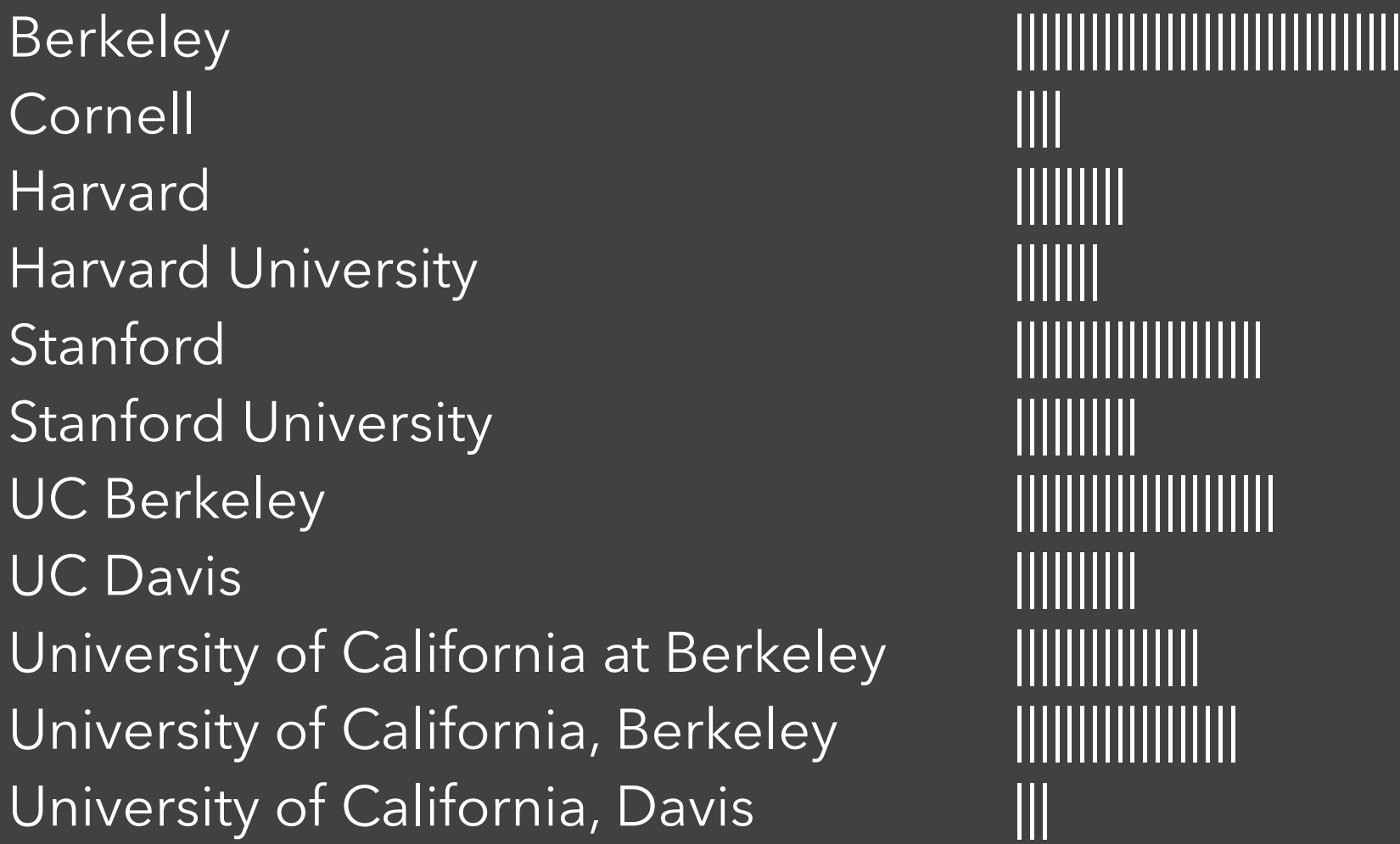
Sort by:

None

Edge centrality filters:



Visualize Friends by School?



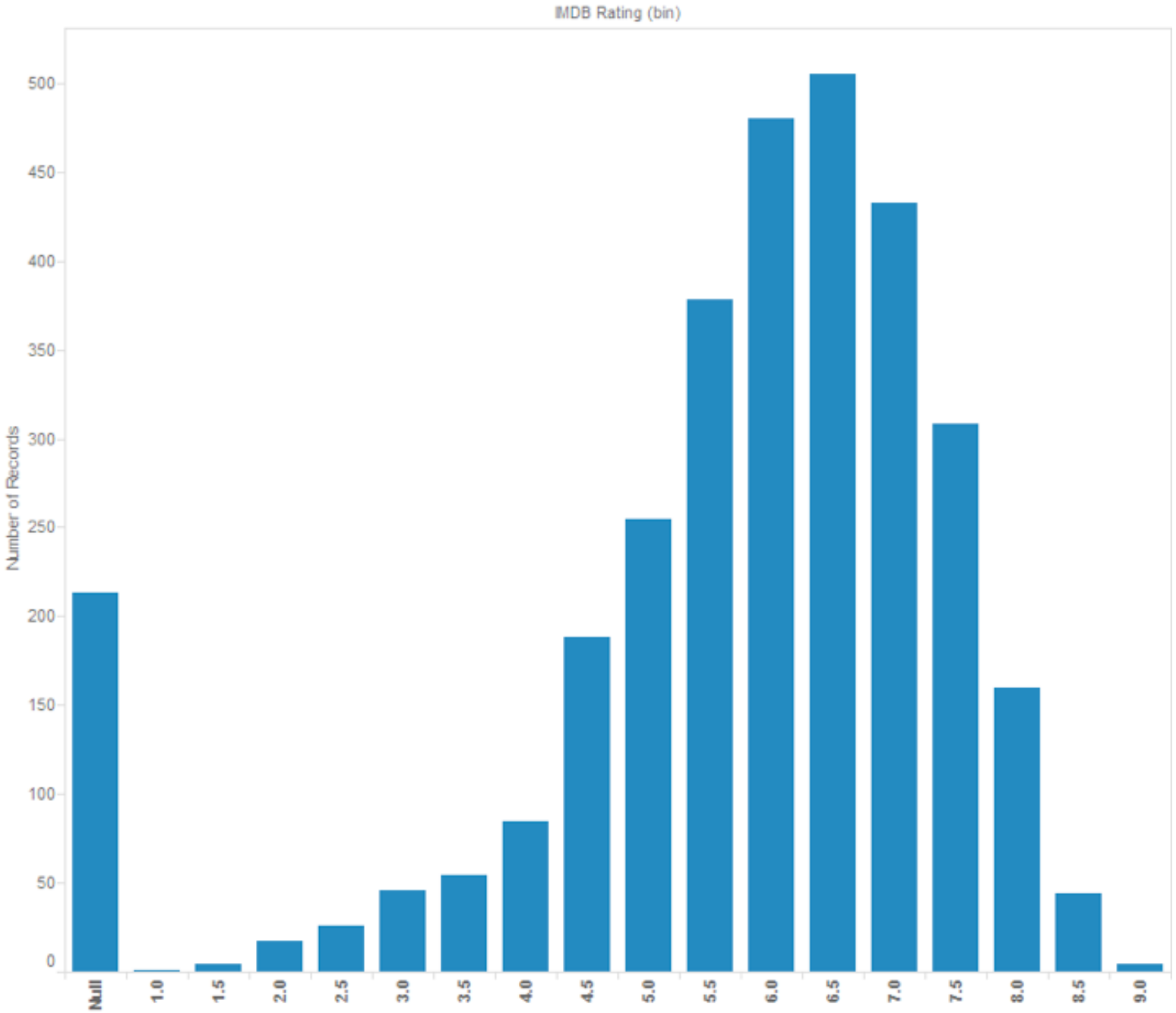
Data Quality Hurdles

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!

Sneek Peek

Analysis Example: Motion Pictures Data


IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0

Analysis Example: Antibiotic Effectiveness

What questions might we ask?

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

Tableau / Polaris

Tableau

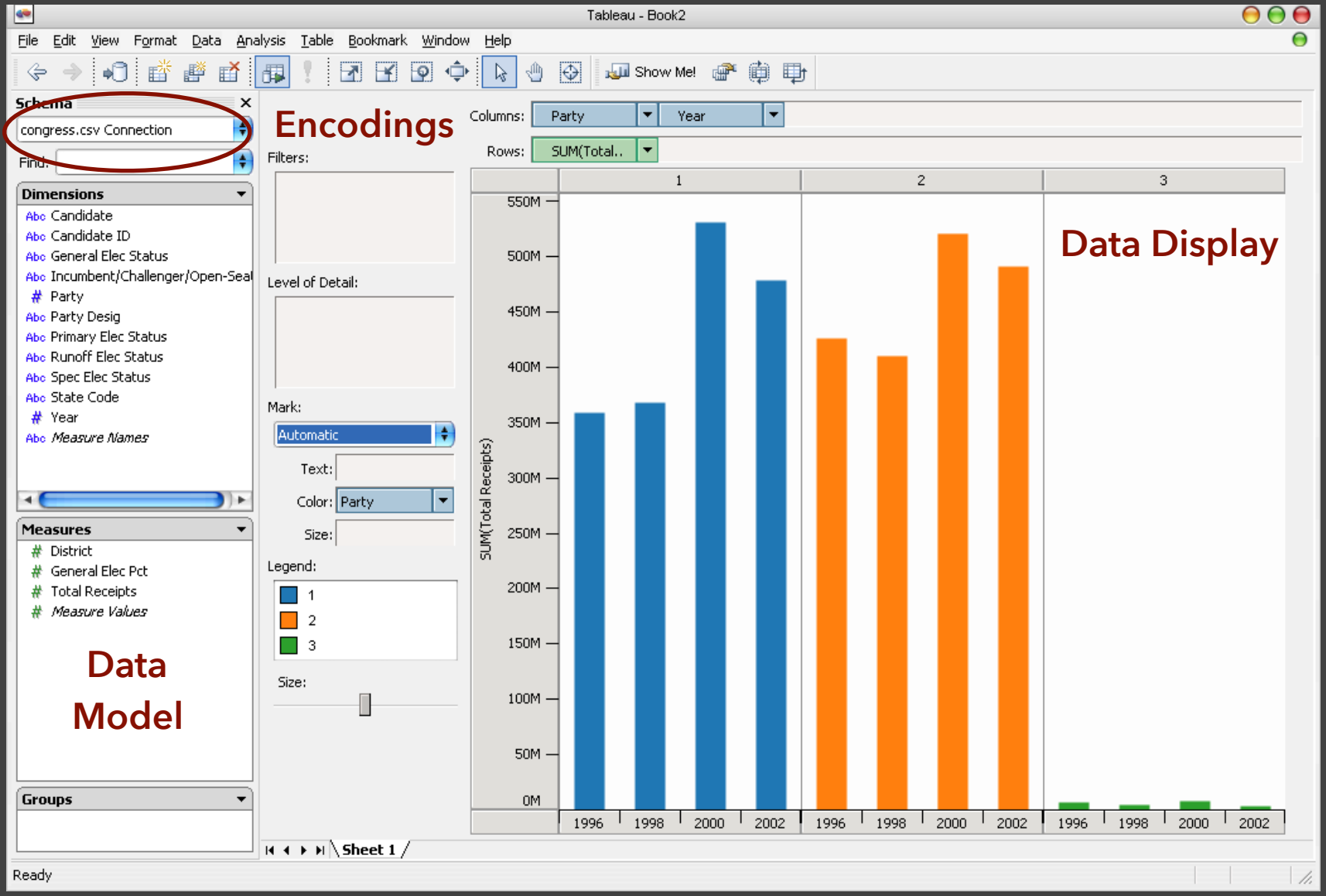


Tableau - Book2
File Edit View Format Data Analysis Table Bookmark Window Help
Show Me!
Schema
congress.csv Connection
Find:
Dimensions
Abc Candidate
Abc Candidate ID
Abc General Elec Status
Abc Incumbent/Challenger/Open-Seat
Party
Abc Party Desig
Abc Primary Elec Status
Abc Runoff Elec Status
Abc Spec Elec Status
Abc State Code
Year
Abc Measure Names
Measures
District
General Elec Pct
Total Receipts
Measure Values
Data Model
Groups
Encodings
Filters:
Level of Detail:
Mark:
Automatic
Text:
Color: Party
Size:
Legend:
1
2
3
Size:
Columns: Party Year
Rows: SUM(Total..
Data Display
SUM(Total Receipts)
550M
500M
450M
400M
350M
300M
250M
200M
150M
100M
50M
0M
1996 1998 2000 2002 1996 1998 2000 2002 1996 1998 2000 2002
Sheet 1
Ready