

EXPLORATORY DATA ANALYSIS

CS 448B | Fall 2024

MANEESH AGRAWALA

1



2

READING RESPONSE: QUESTIONS/THOUGHTS

... [John Snow's] cholera map is clear with only essential elements like dots and crosses presented in the map, and removing any irrelevant things like buildings, trees, etc.
Distinguishing the necessary and unnecessary parts and selecting between them need careful consideration for the designers.

However, ... ***users can now use tools like D3.js and Tableau*** to explore data and get interactive and dynamical experience. ..., ***users can customize the complexity of their exposure to the data***. In other words, users can interact with the data, removing irrelevant information and accessing more details in the aspect they are interested in. ... ***How to guide the users to start the interactive data visualization process while avoiding overwhelming them with too much information... ?***

3

Learning Objectives

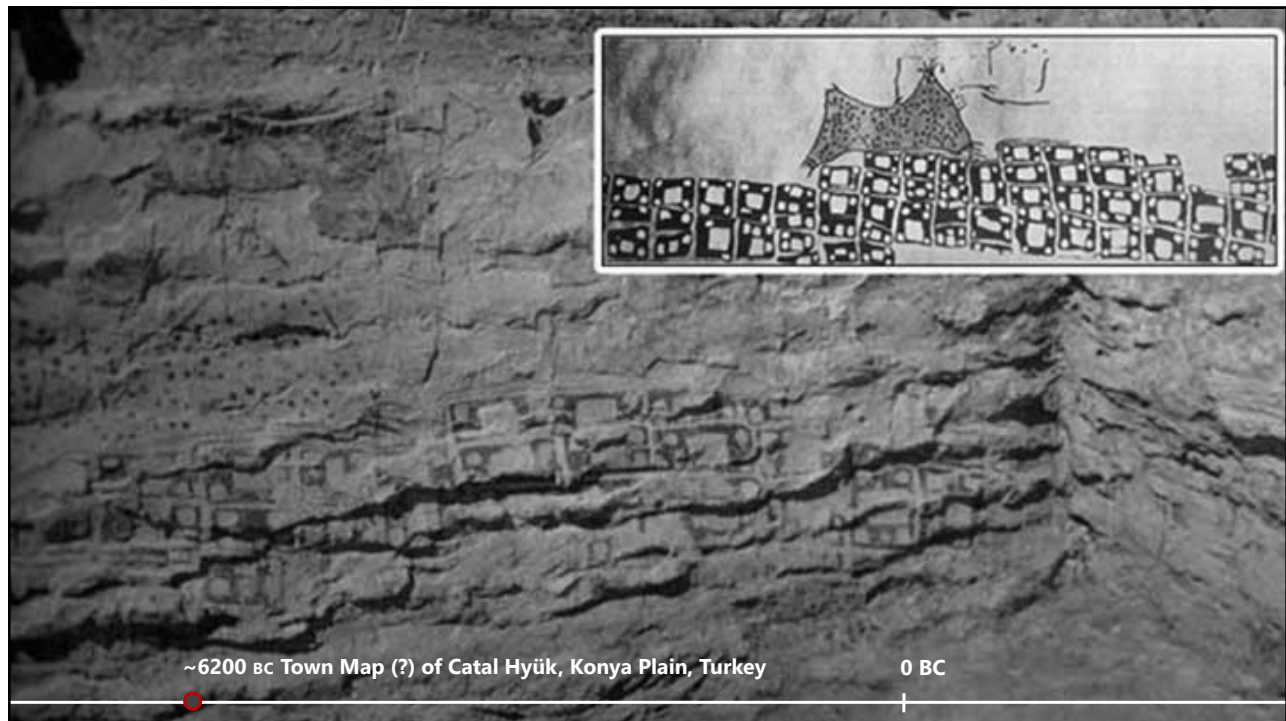
TODAY

1. What is exploratory data analysis and why is it important?
2. What factors should we consider when exploring a dataset?
3. How do visualization researchers design tools to support exploratory data analysis?

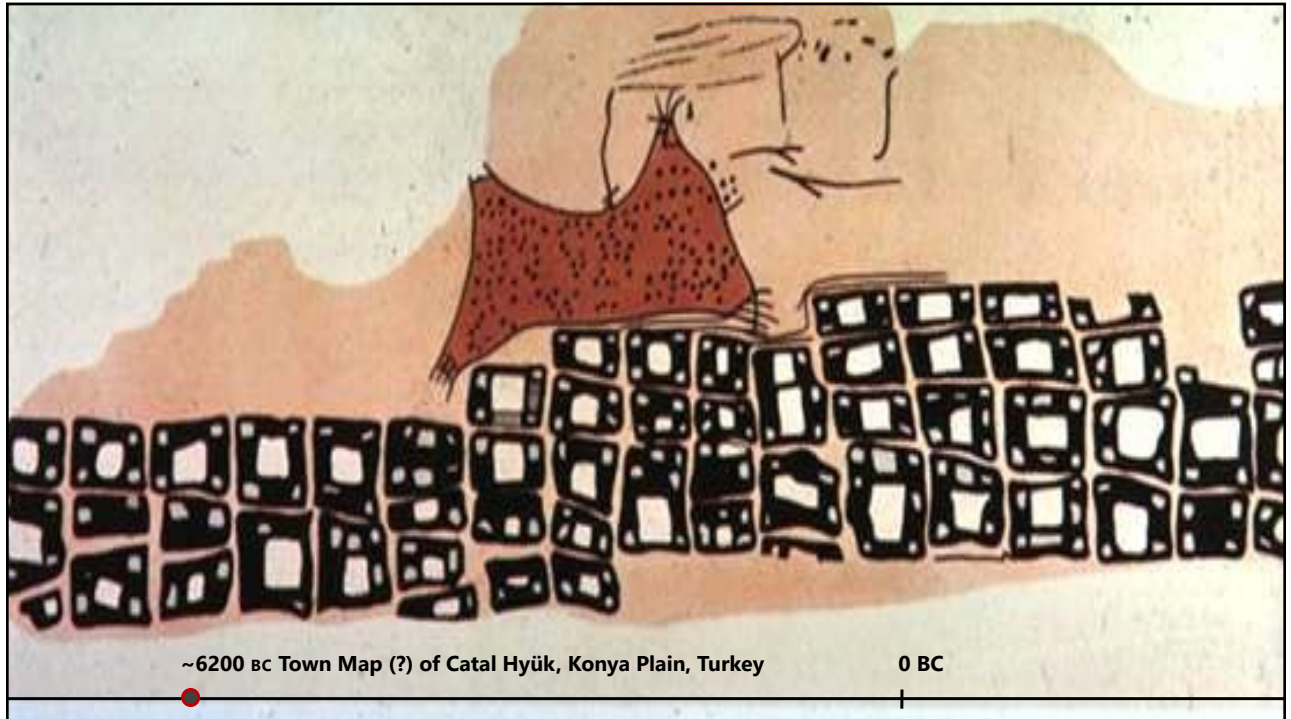
4

WHAT WAS THE FIRST DATA VISUALIZATION?

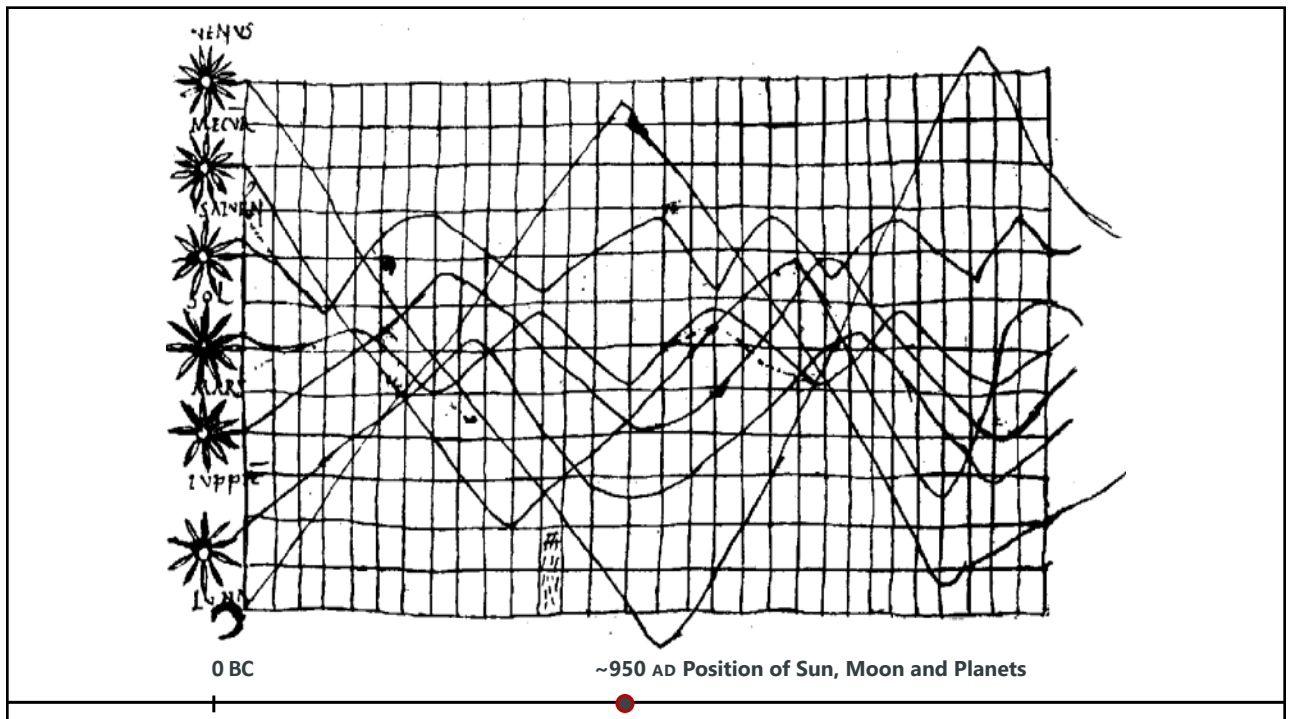
6



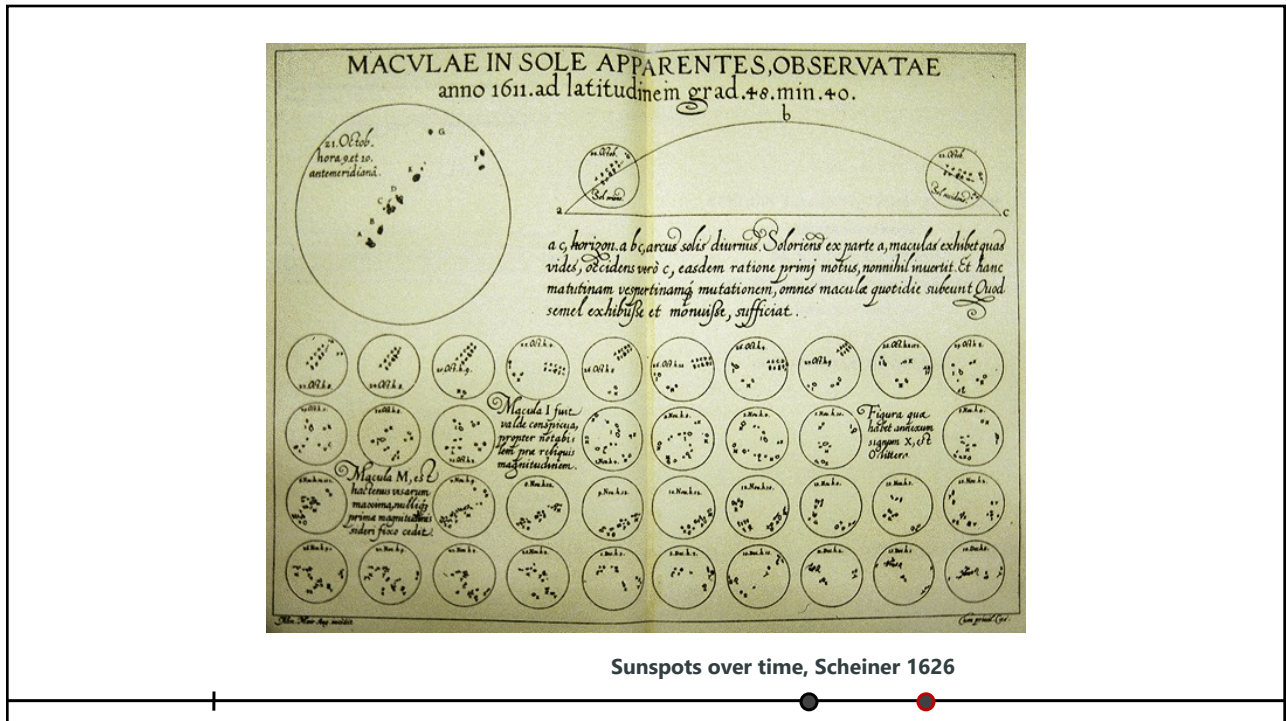
7



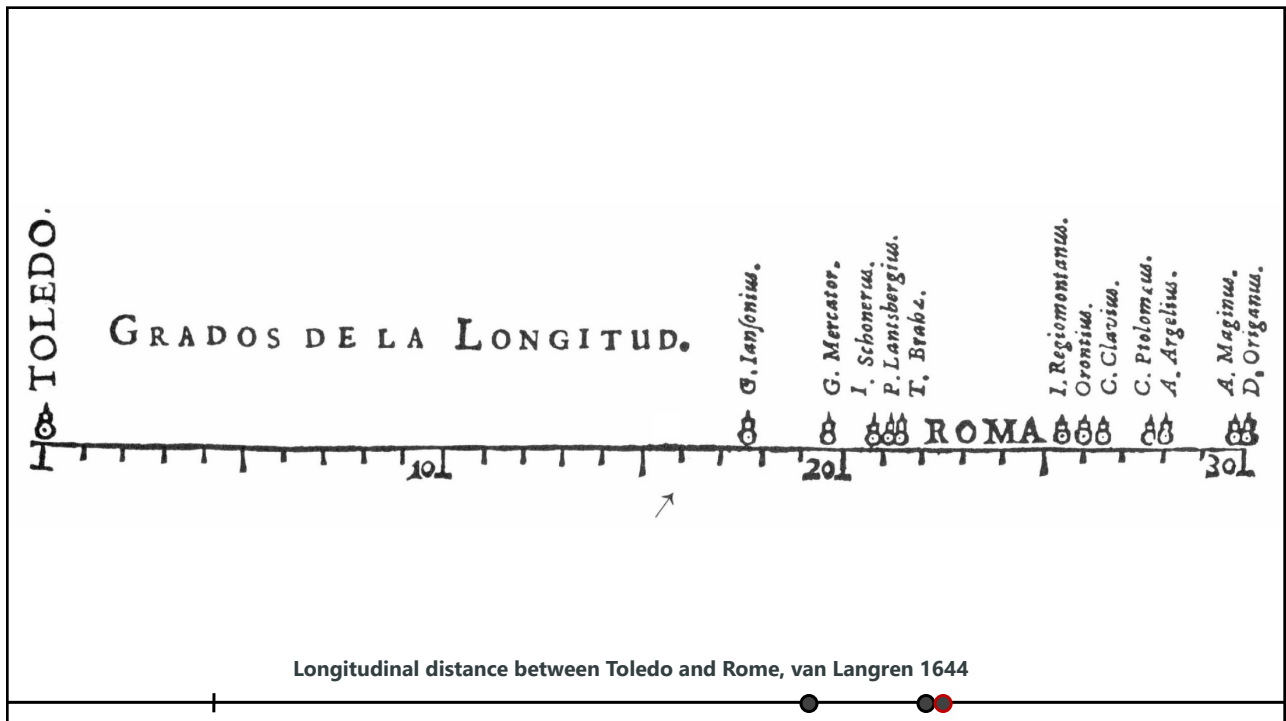
8



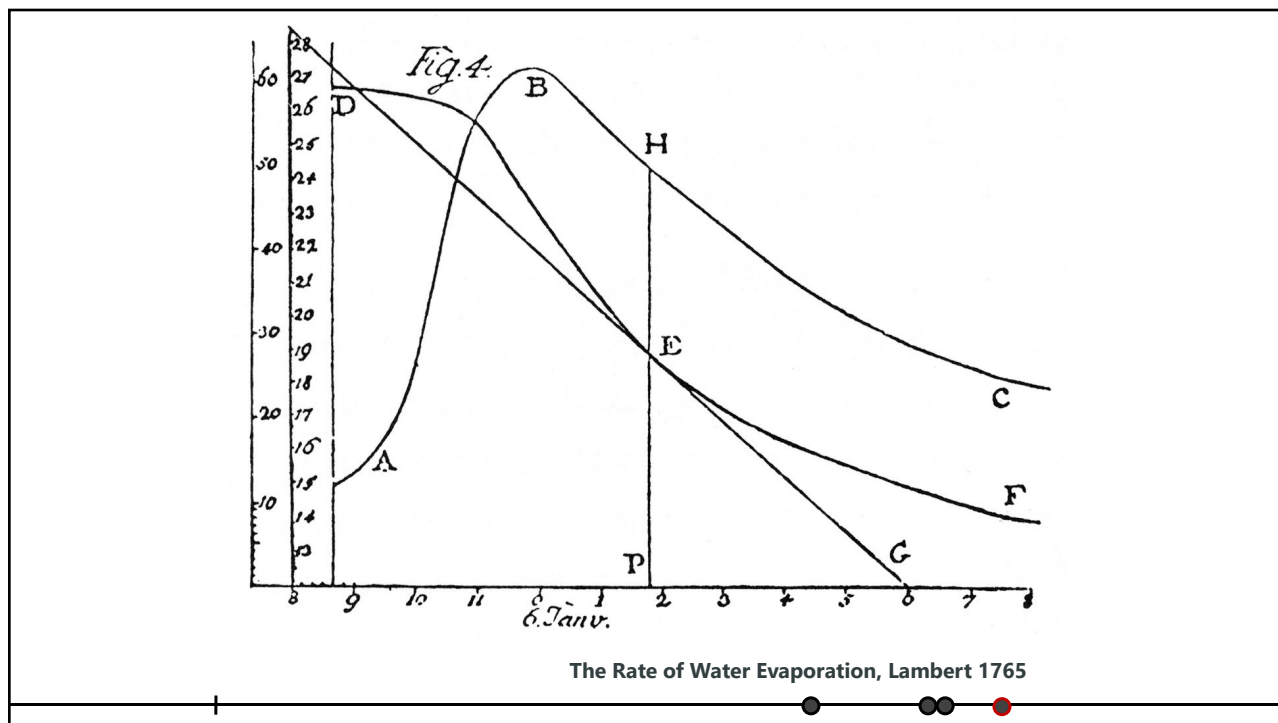
9



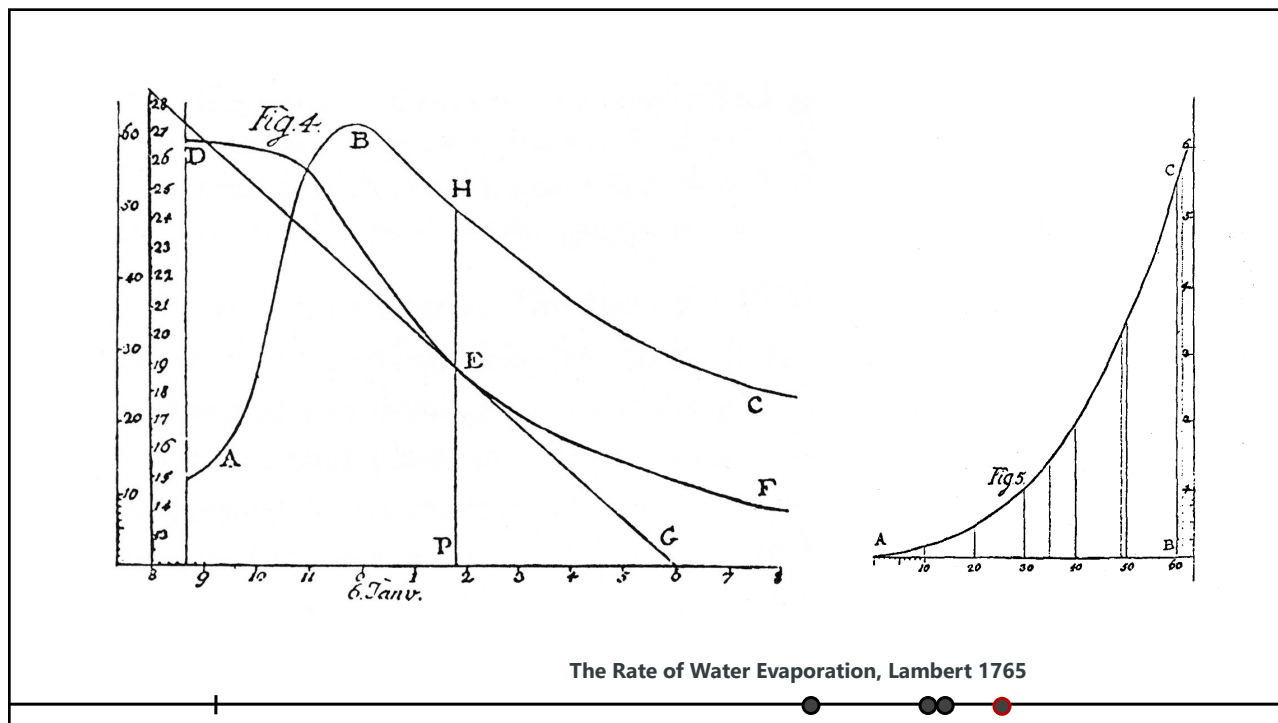
10



11



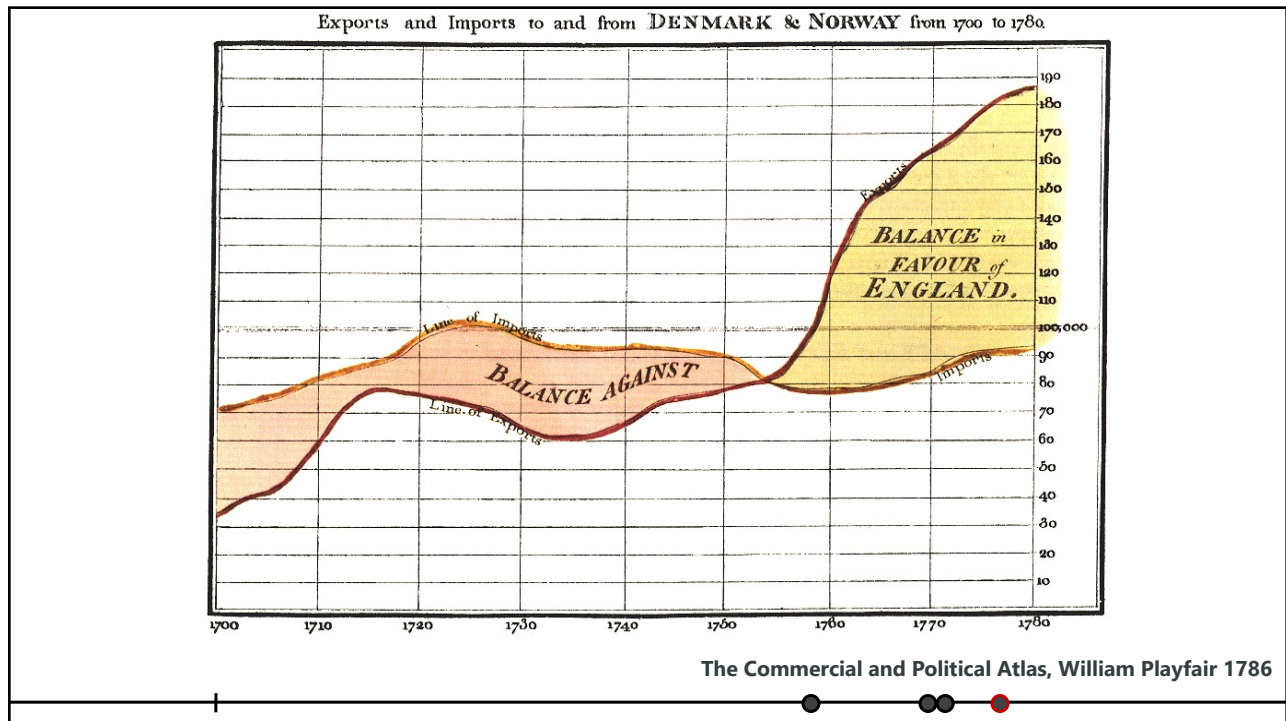
12



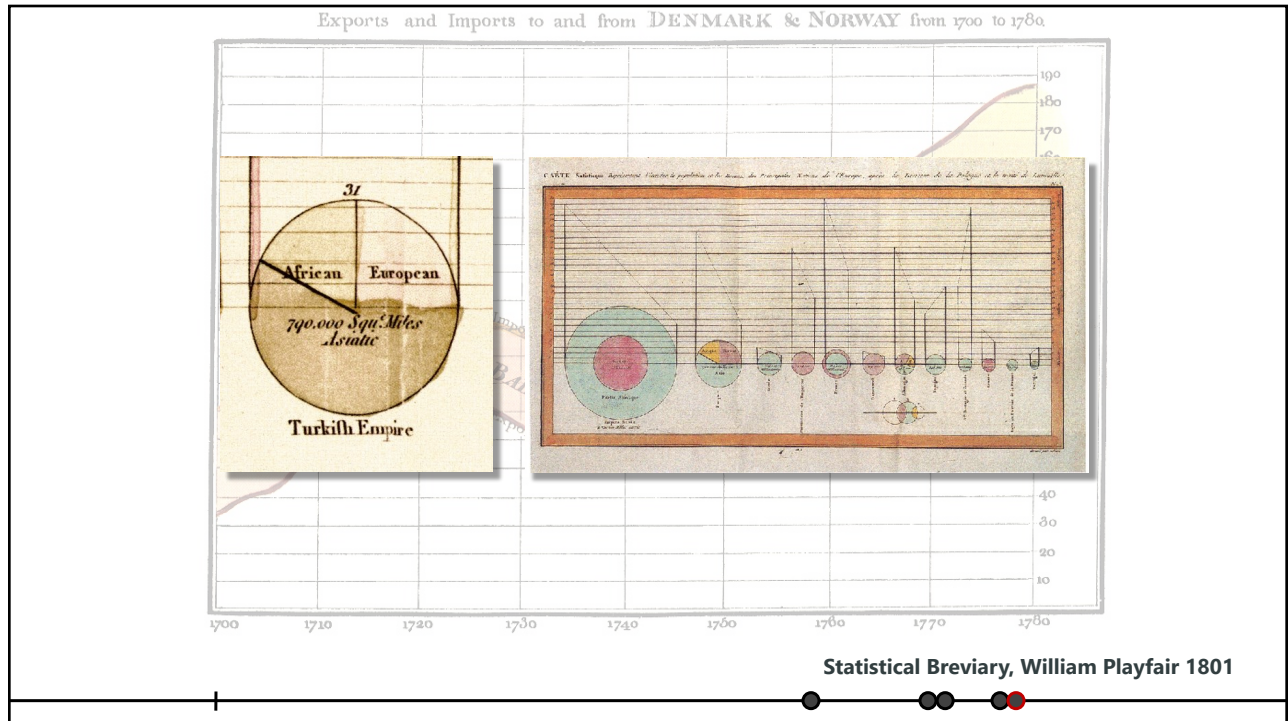
13

THE GOLDEN AGE OF VISUALIZATION

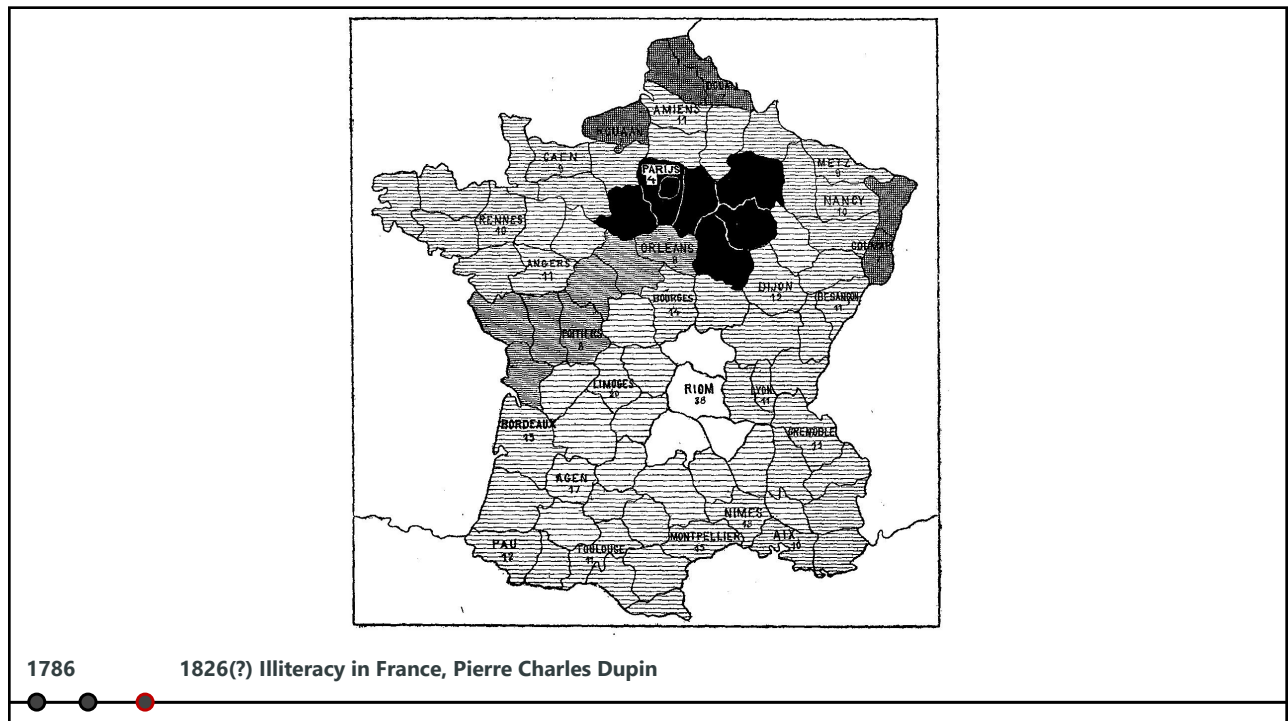
14



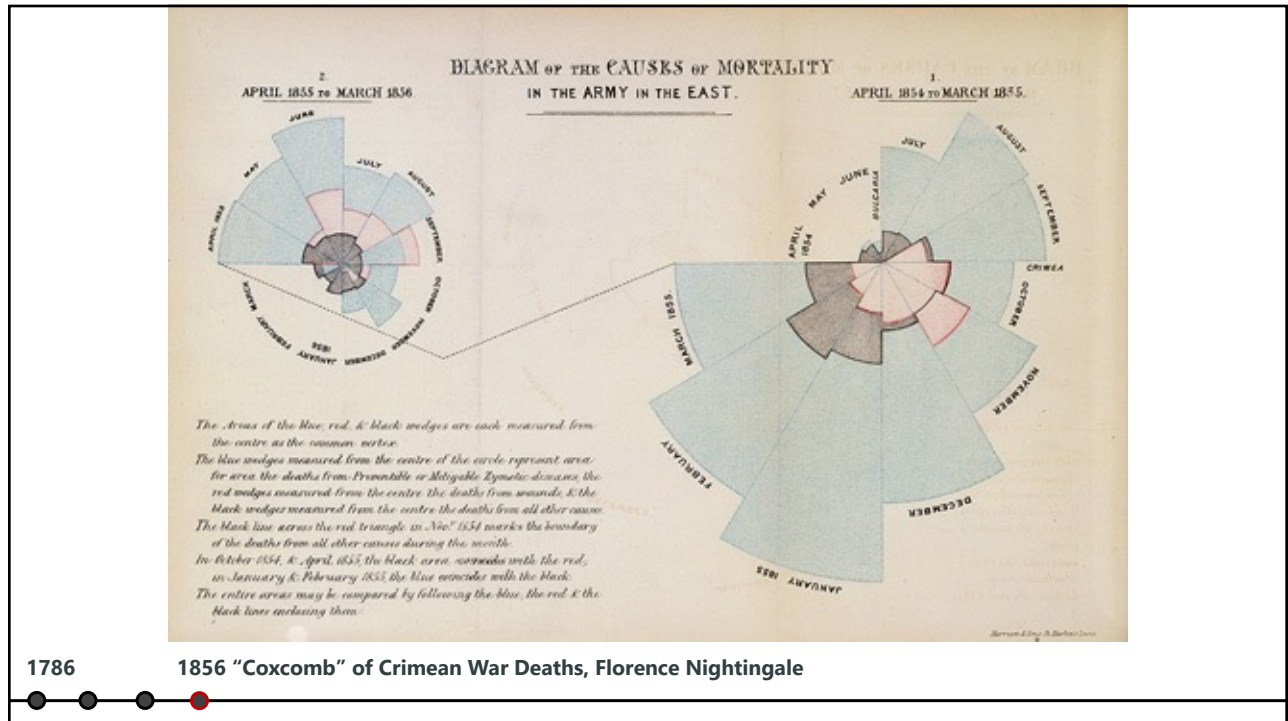
15



16

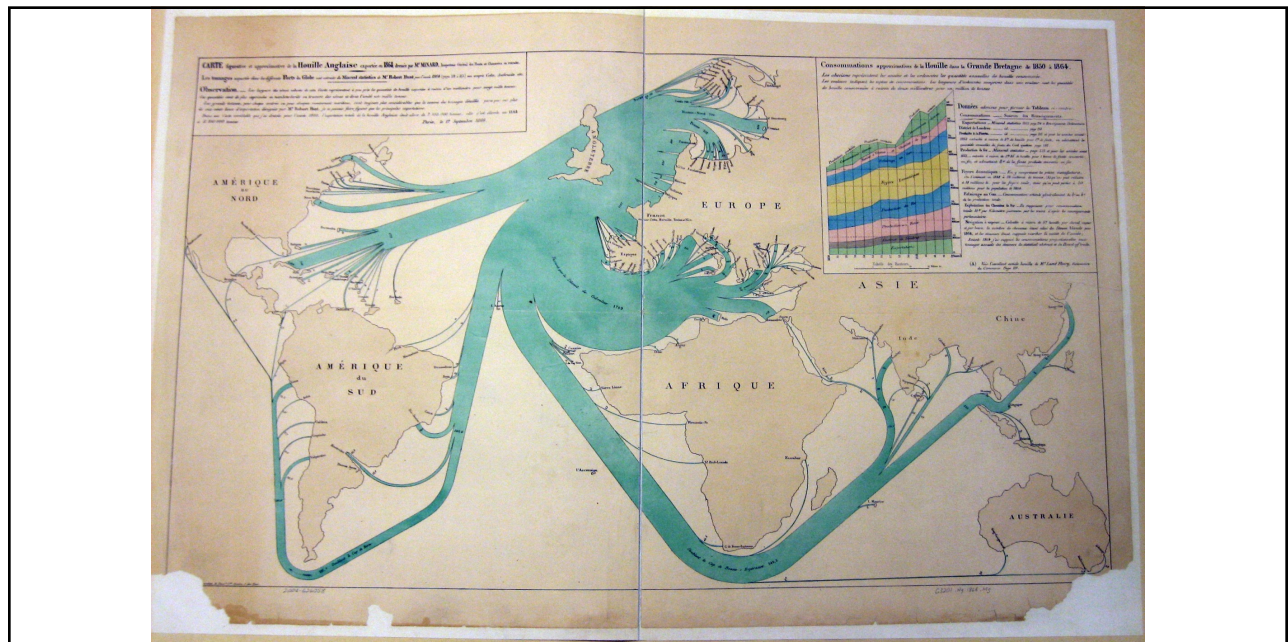


17



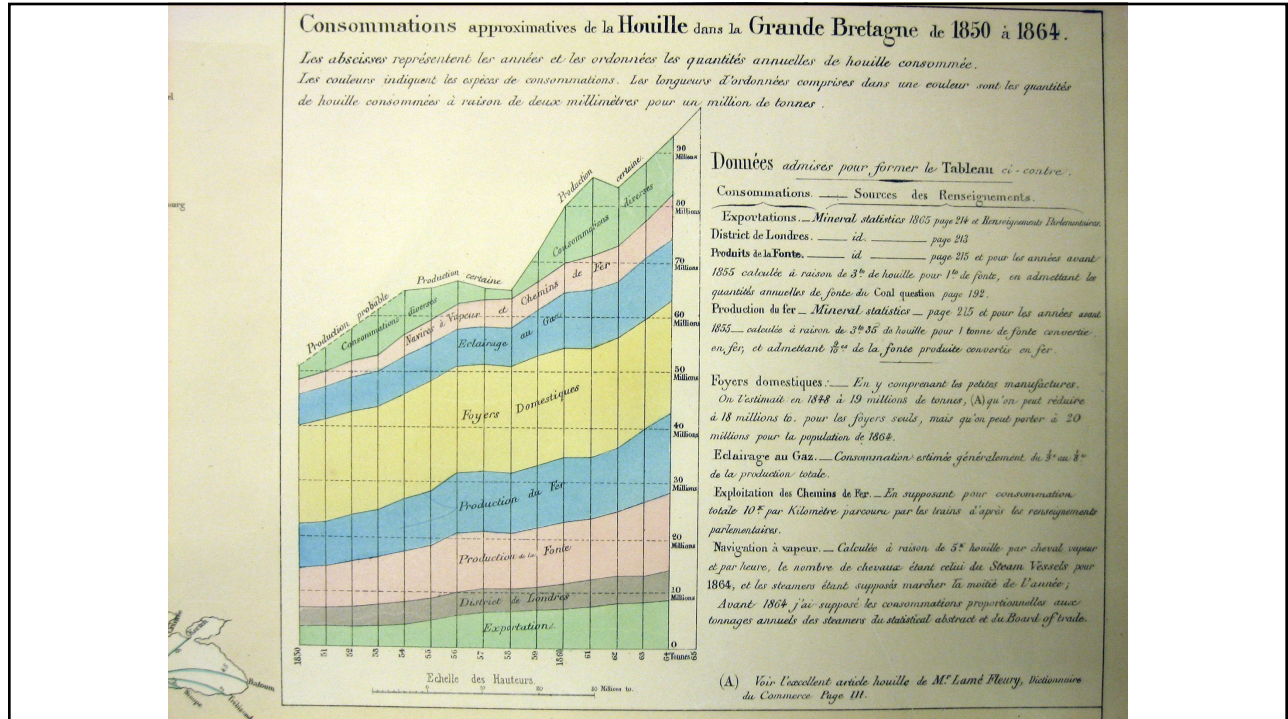
1786 1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale

18

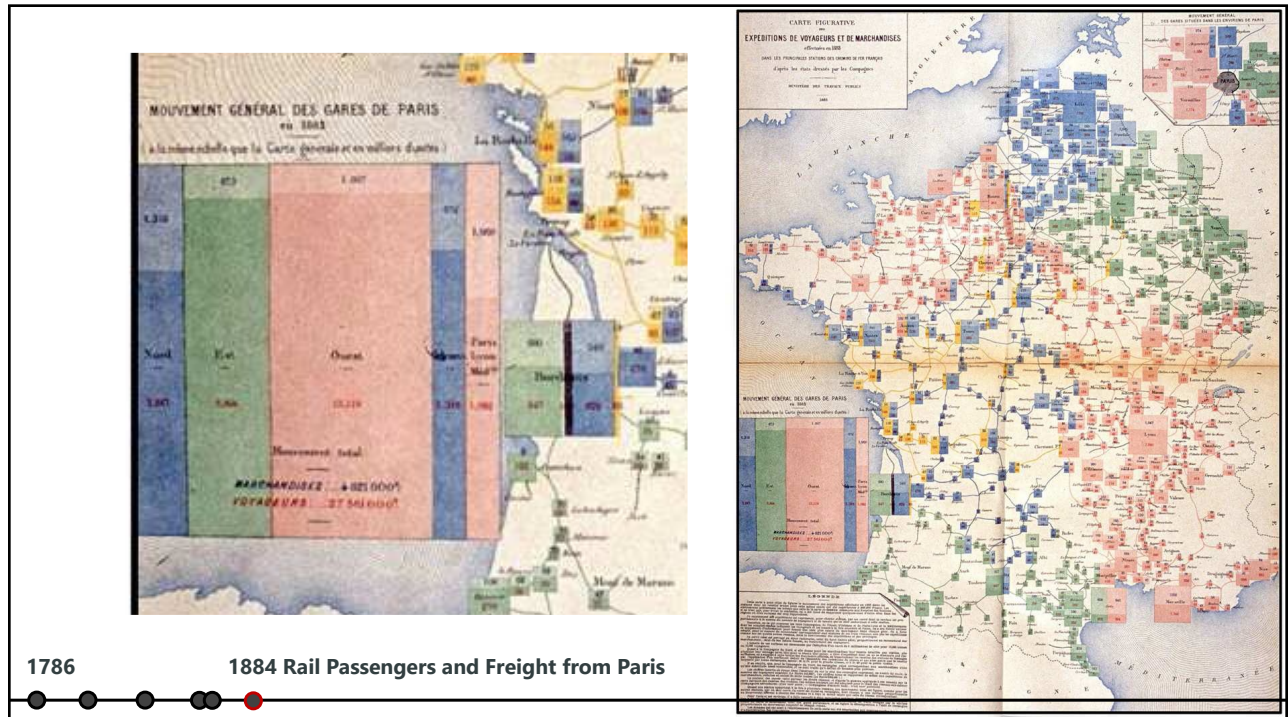


1786 1864 British Coal Exports, Charles Minard

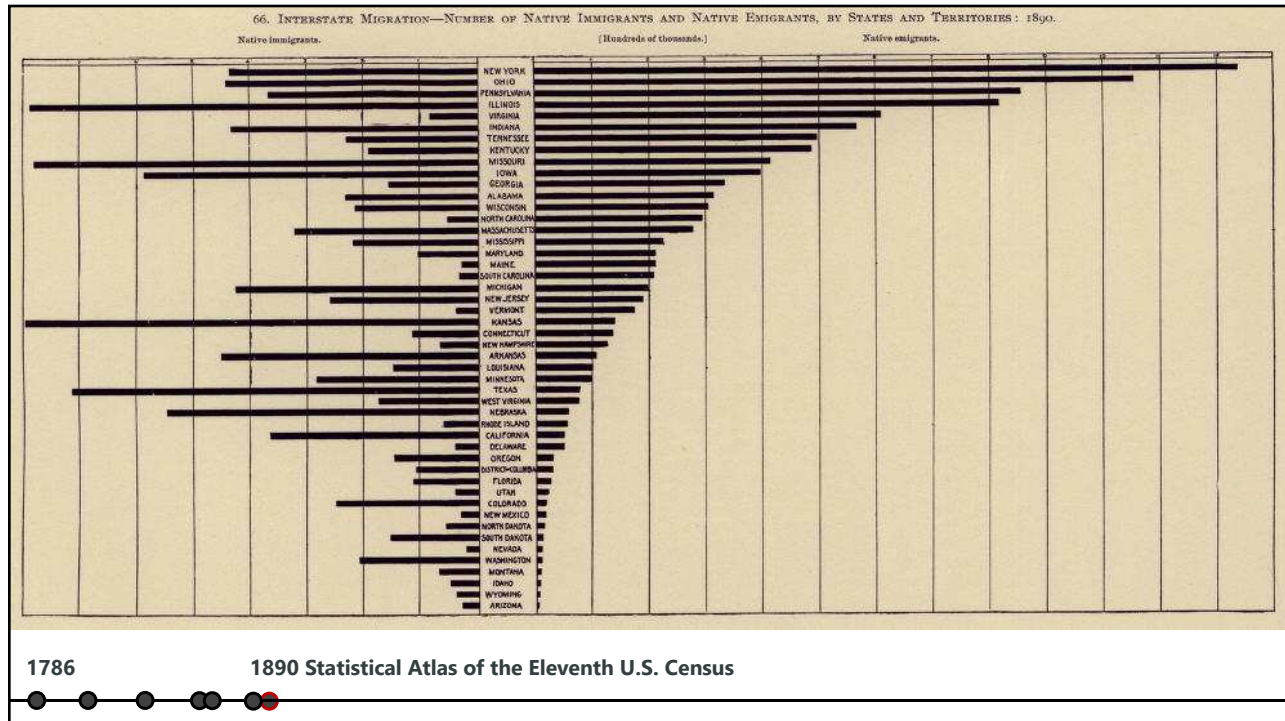
19



20



21



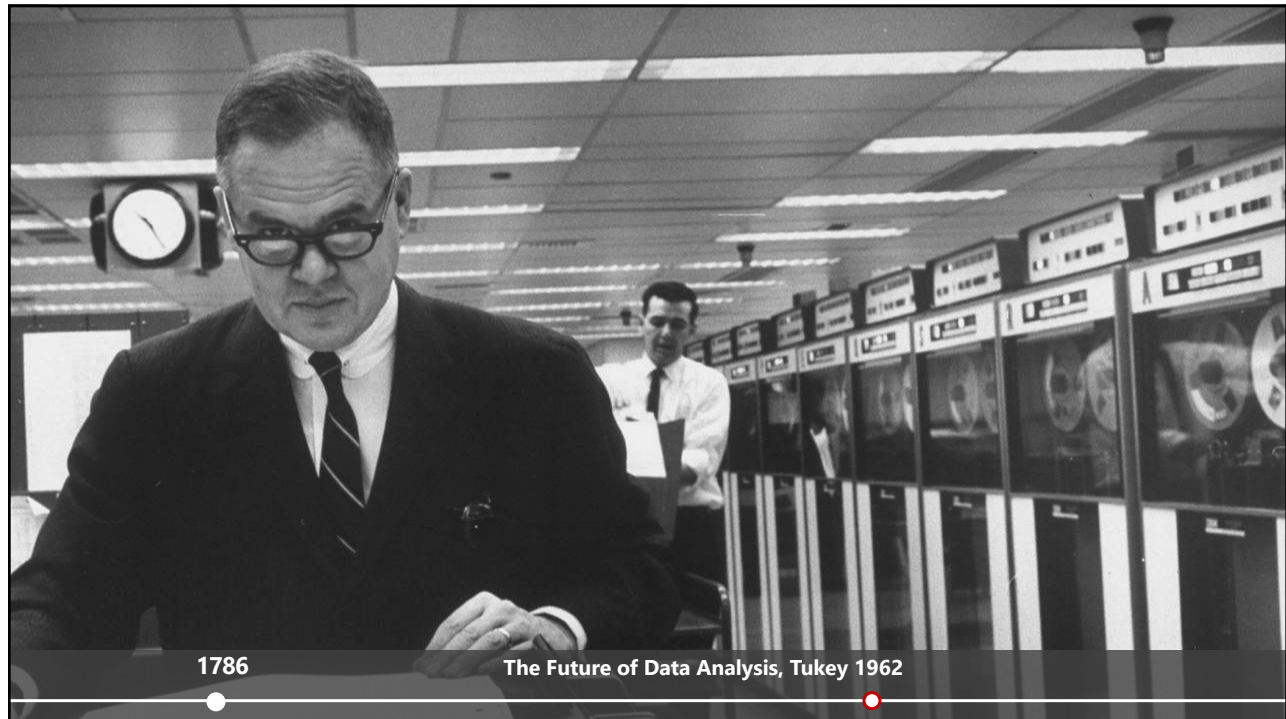
22

THE RISE OF STATISTICS

1. Use of formal methods from statistics and social science
2. Little innovation in graphical methods
3. A period of popularization and application
4. Graphical methods enter textbooks, curricula and mainstream use

1786 1900 1950

23




24

Four major influences act on data analysis today:

1. Formal theories of statistics
2. Accelerating developments in computers and display devices
3. More and larger bodies of data
4. Emphasis on quantification in many disciplines

The Future of Data Analysis, John W. Tukey 1962


25



The Future of Data Analysis, John W. Tukey 1962

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.

28



The Future of Data Analysis, John W. Tukey 1962

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**

29

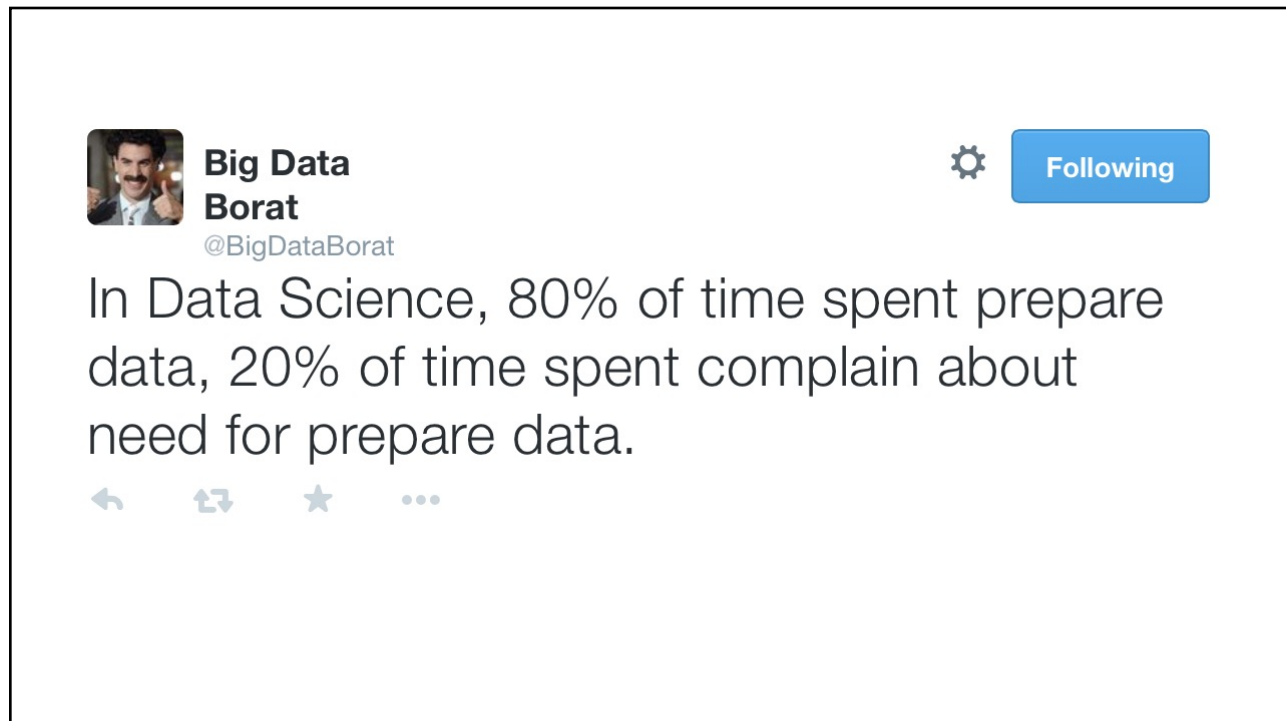
DATA WRANGLING

32

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 2012]

33



The image shows a screenshot of a tweet from a user named 'Big Data Borat' with the handle '@BigDataBorat'. The user's profile picture is a man with a mustache. To the right of the name is a gear icon and a blue button that says 'Following'. The tweet text reads: 'In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.' Below the text are icons for reply, retweet, favorite, and a menu.

34

TIDY DATA [Wickham 2014]

How do rows and columns, match up with data fields, and observations?

In *tidy data*

1. Each field forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Flexible starting point for analysis, transformation, and visualization

35

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

36

ARQUERO

<https://observablehq.com/@uwdata/tidy-data-in-javascript>

```

state      year      rate
Alabama    2004    4029.3
Alabama    2005    3900.0
Alabama    2006    3937.0
Alabama    2007    3974.9
Alabama    2008    4081.9
Alaska     2004    3370.9
Alaska     2005    3615.0
Alaska     2006    3582.0
Alaska     2007    3373.9
Alaska     2008    2928.3
Arizona    2004    5073.3
Arizona    2005    4827.0
Arizona    2006    4741.6
Arizona    2007    4502.6

```

```

({) aq.fromCSV(crime_csv(), { header: false, names: ['year', 'rate'] })
  .filter(d => d.year != null || d.rate != null)
  .derive({
    state: d => op.fill_down(op.match(d.year, /Reported crime in (.*)/, 1)) // <- extract state name
  }, { before: 0 })
  .filter(d => d.rate != null) // <-- or, we could delete when year column starts with "Reported crime in"
  .view(100)

```

37

DataWrangler

Transform Script Import Export

- Split data repeatedly on newline into rows
- Split split repeatedly on "
- Promote row 0 to header
- Delete empty rows

Text Columns Rows Table Clear

Extract from Year after " in "

Extract from Year after " in "

Cut from Year after " in "

Cut from Year after " in "

Split Year after " in "

Split Year after " in "

Year	extract	#	Property_crime_rate
0	Reported crime in Alabama	Alabama	
1	2004		4029.3
2	2005		3900
3	2006		3937
4	2007		3974.9
5	2008		4081.9
6	Reported crime in Alaska	Alaska	
7	2004		3370.9
8	2005		3615
9	2006		3582
10	2007		3373.9
11	2008		2928.3
12	Reported crime in Arizona	Arizona	
13	2004		5073.3
14	2005		4827
15	2006		4741.6
16	2007		4502.6
17	2008		4087.3
18	Reported crime in Arkansas	Arkansas	
19	2004		4033.1
20	2005		4068
21	2006		4021.6
22	2007		3945.5
23	2008		3843.7
24	Reported crime in California	California	
25	2004		3423.9
26	2005		3321
27	2006		3175.2
28	2007		3032.6
29	2008		2940.3
30	Reported crime in Colorado	Colorado	

Wrangler: Interactive Visual Specification of Data Transformation Scripts [Kandel 2011]

38

WRANGLING DATA

One often needs to reformat, clean, quality assess, and integrate data prior to analysis

Some approaches:

Code: [arquero](#) (Javascript), [dplyr](#) (R), [pandas](#) (python)

Manual manipulation in spreadsheets

[Open Refine](#)

[Tableau](#)

Data wrangler [Kandel 2011] became Trifacta Wrangler but was recently bought by [Alteryx](#) and is a little harder to use now

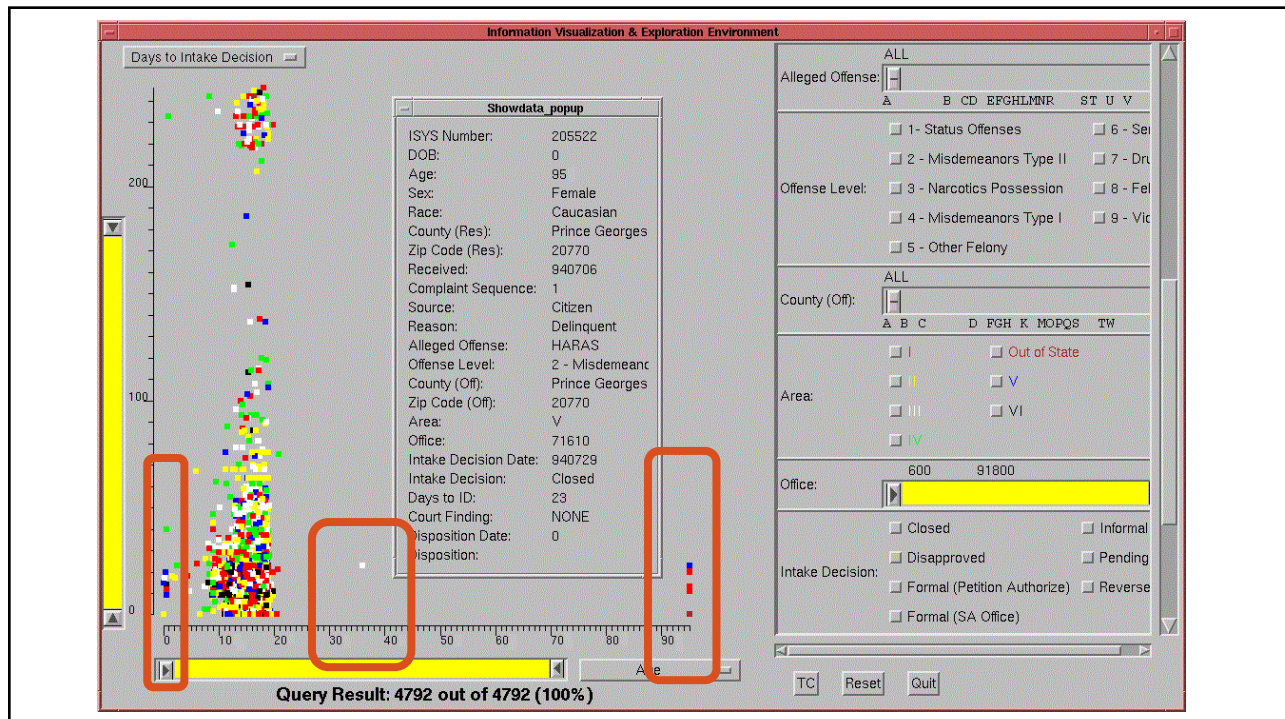
39

“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.”

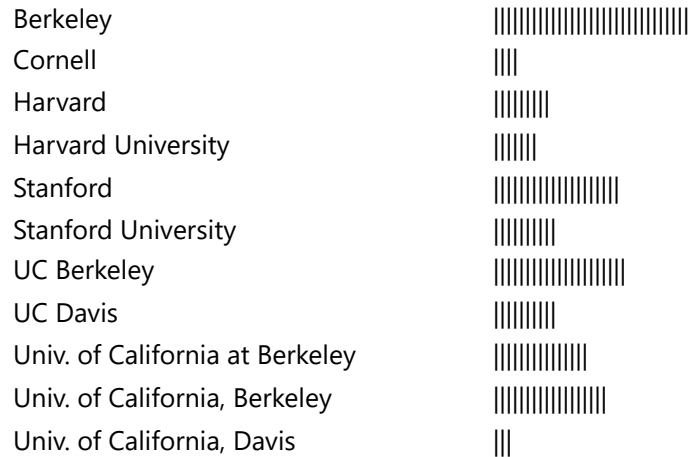
- Martin Wattenberg

40



41

VISUALIZE FRIENDS BY SCHOOL



46

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!

47

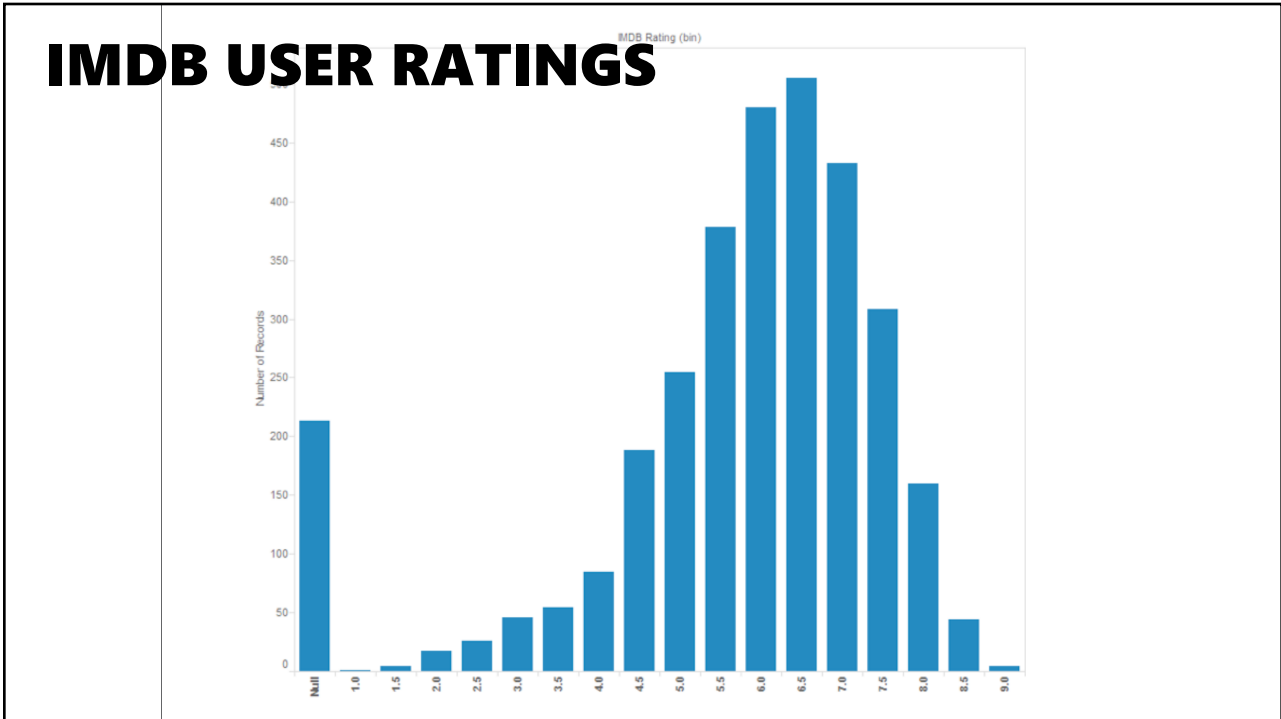
ANALYSIS EXAMPLE: MOTION PICTURES DATA

48

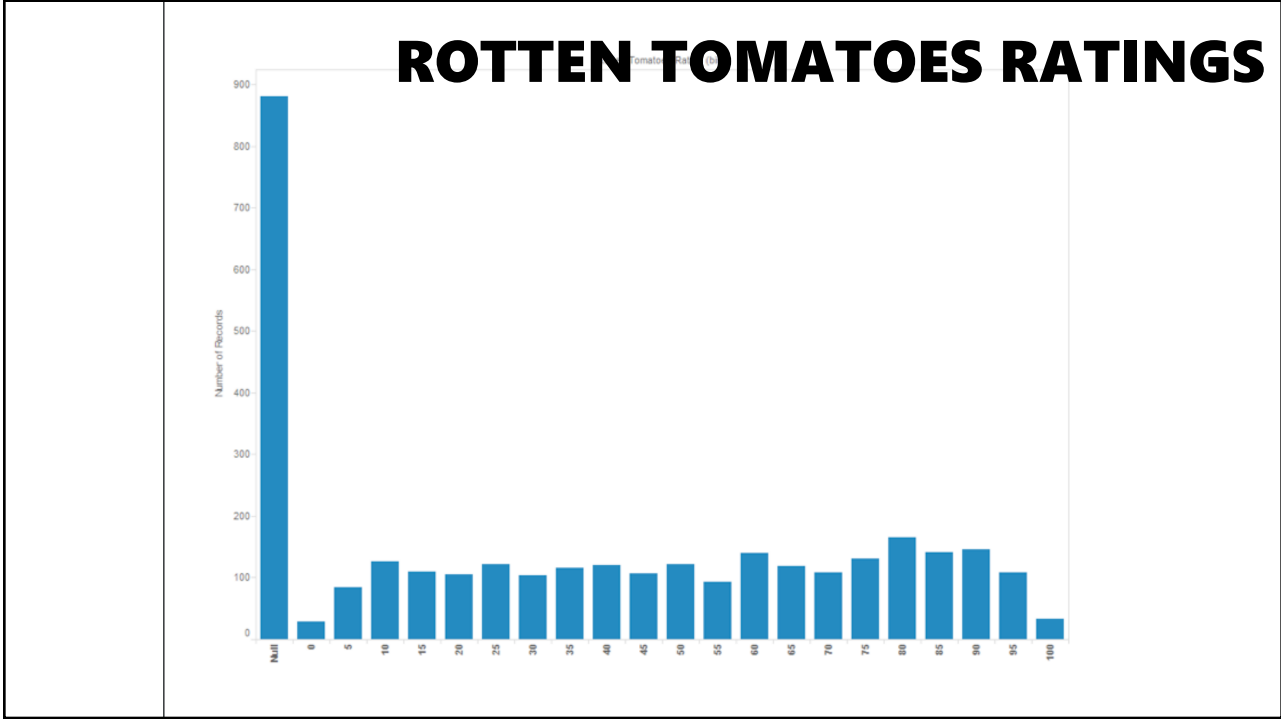
MOTION PICTURES DATA TYPES

Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)

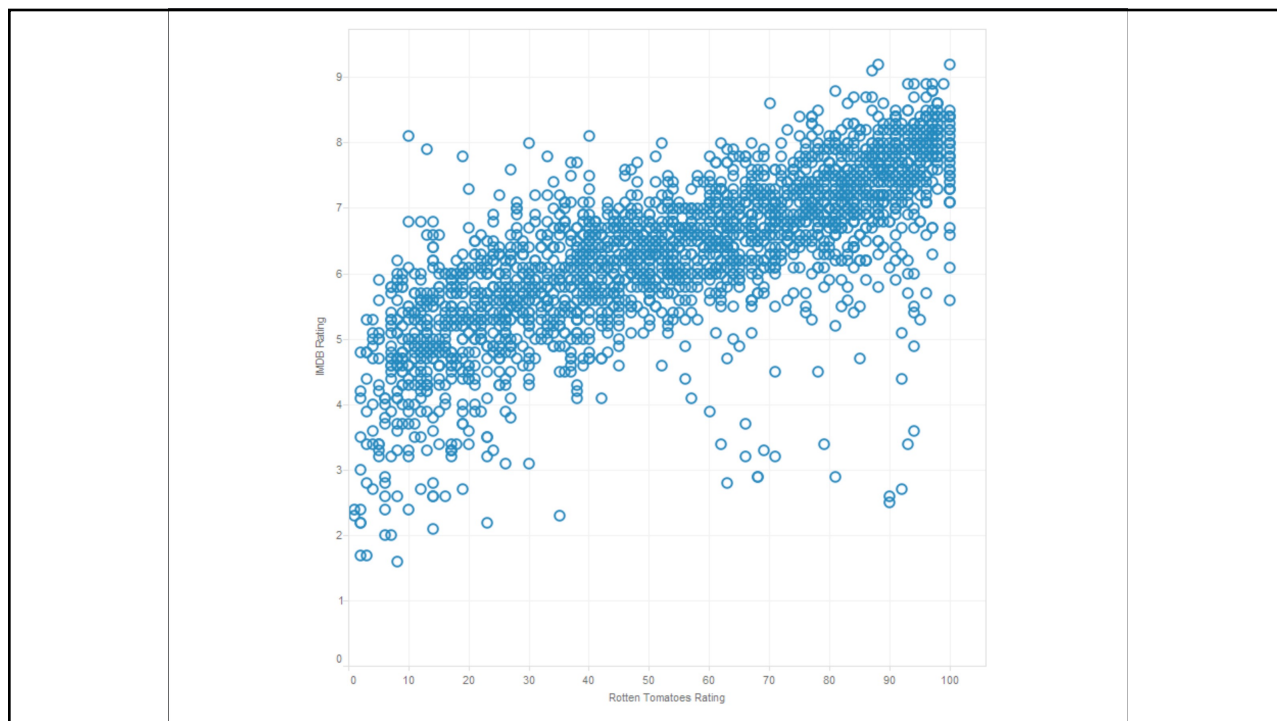
49



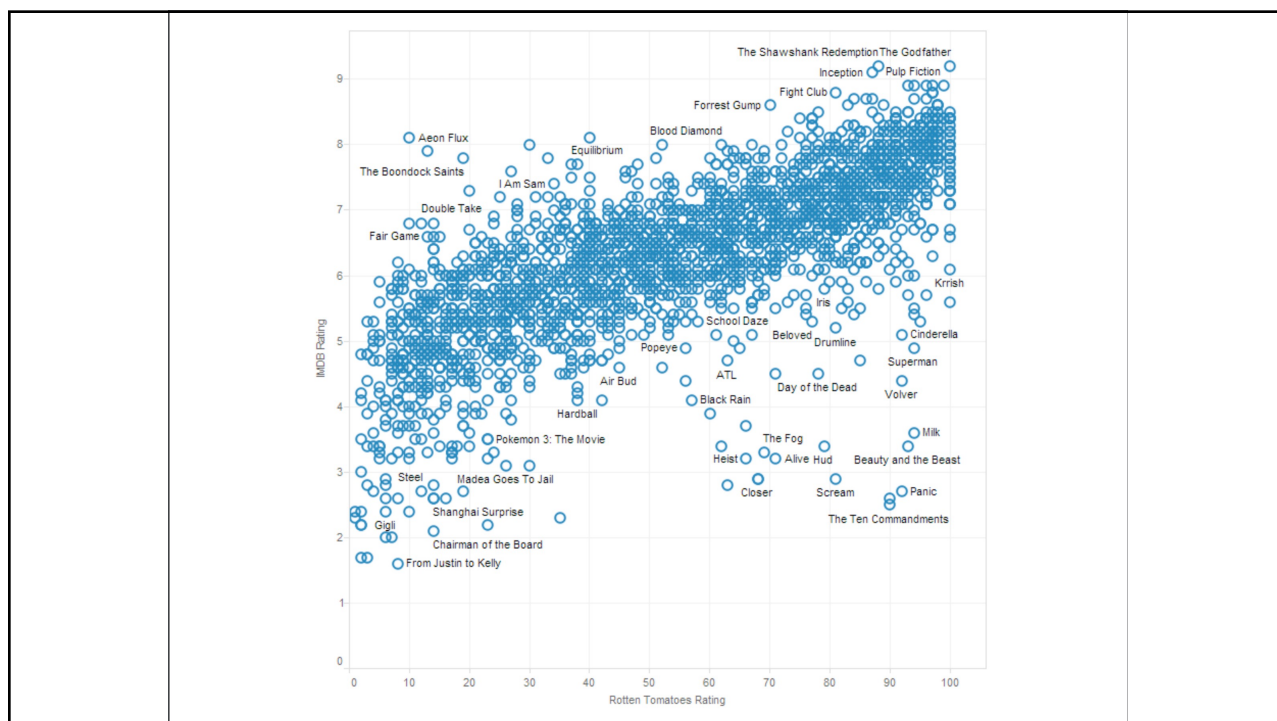
50



51



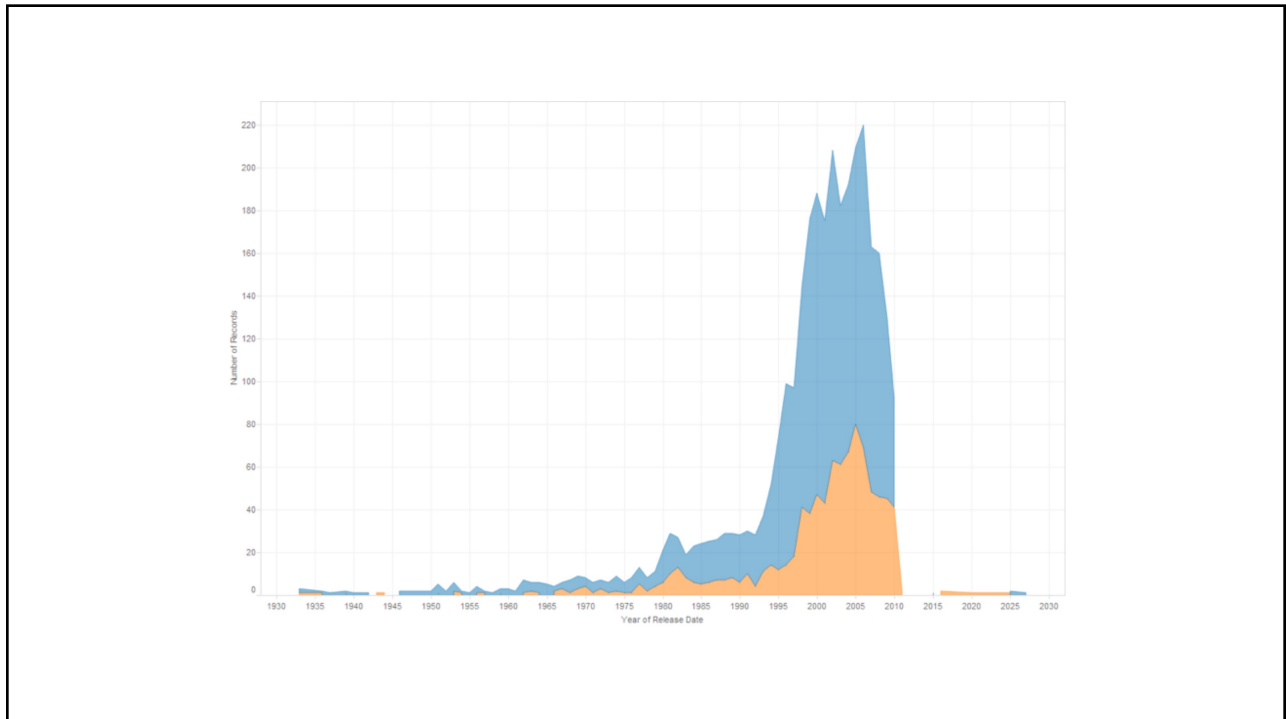
52



53



54



55

LESSON: EXERCISE SKEPTICISM

Check **data quality** and your **assumptions**

Start with **univariate summaries**, then consider **relationships between variables**

Avoid premature fixation!

56

ANNOUNCEMENTS

57

ASSIGNMENT 2: EXP. DATA ANALYSIS

Due 10/14 10:30am

Use **Tableau** or **Vega-Lite** to formulate & answer data questions

First steps

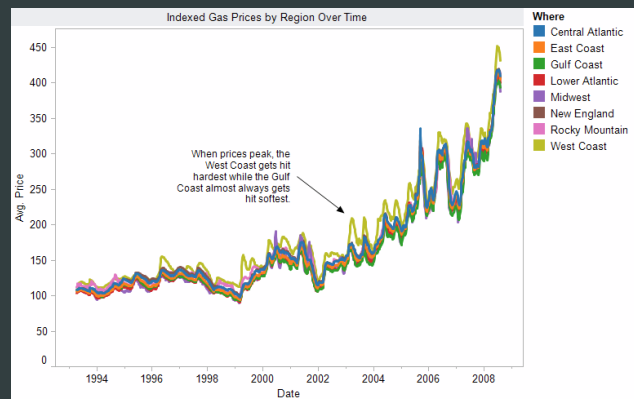
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile data
- Iterate as needed

Create visualizations

- See different views of data
- Refine questions

Author a report

- Screenshots of most insightful views (8+)
- Include titles and captions for each view



58

ANALYSIS EXAMPLE: ANTIBIOTIC EFFECTIVENESS

59

ANTIBIOTIC EFFECTIVENESS DATA TYPES

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

Collected prior to 1951

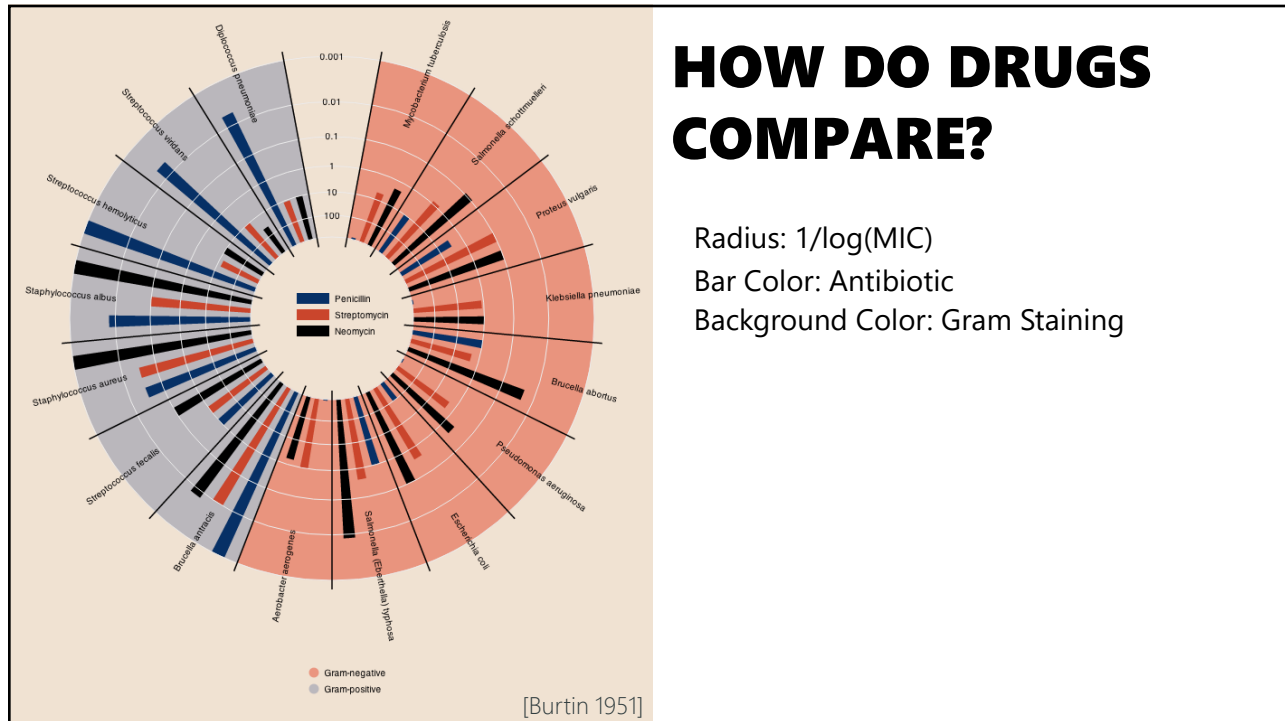
60

WHAT QUESTIONS MIGHT WE ASK?

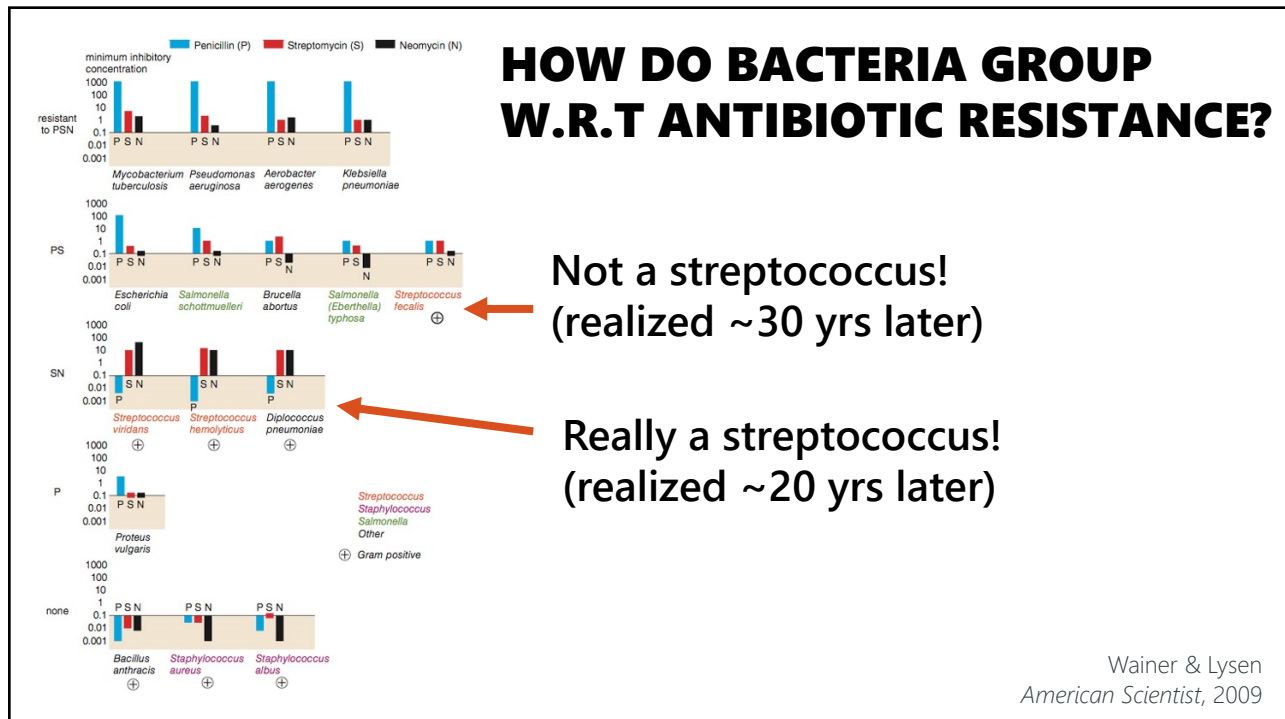
Table 1: Burtin's data.

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 fecalis</i>	1	1	0.1	positive
<i>Streptococcus hemolyticus</i>	0.001	14	10	positive
<i>Streptococcus viridans</i>	0.005	10	40	positive

61



62



63

LESSON: EDA IS AN ITERATIVE 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

65

TABLEAU/POLARIS

100

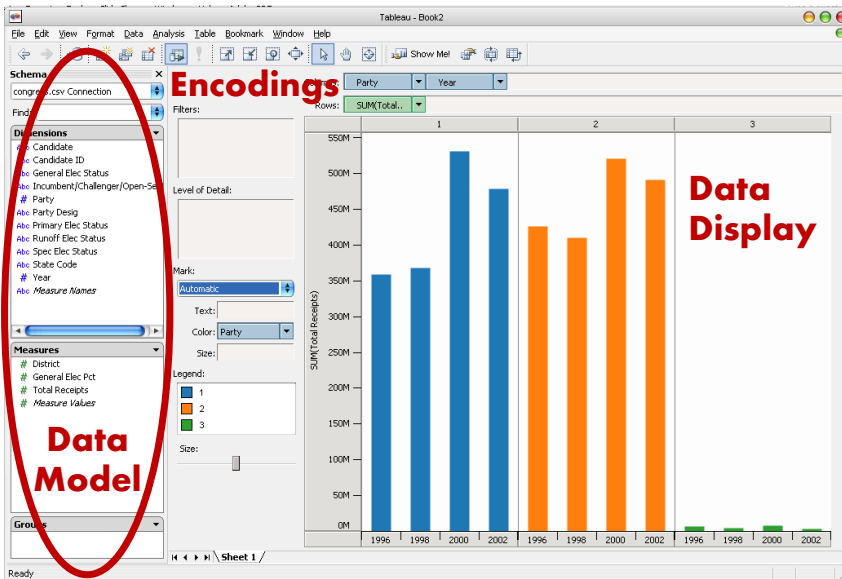
POLARIS [Stolte 2002]



Started as a Stanford research project by C. Stolte, D. Tang & P. Hanrahan

101

TABLEAU



102

POLARIS/TABLEAU APPROACH

Insight: simultaneously specify both database queries & visualization

Choose data, then visualization, not vice versa

Use **smart defaults** for visual encodings (Like APT)

Can also suggest more encodings upon request (ShowMe)

103

TABLEAU DEMO

Dataset:

Federal Elections Commission Receipts
Every Congressional Candidate from 1996 to 2002
4 Election Cycles
9216 Candidacies

104

DATA TYPES

Year (Qi)
Candidate Code (N)
Candidate Name (N)
Incumbent / Challenger / Open-Seat (N)
Party Code (N) [1=Dem,2=Rep,3=Other]
Party Name (N)
Total Receipts (Qr)
State (N)
District (N)

This is a subset of the larger data set available from the FEC, but should be sufficient for the demo

105

HYPOTHESES

What might we learn from this data?

Have receipts increased over time?
Do democrats or republicans spend more?
Candidates from which state spend the most money?

106



107