EXPLORATORY DATA ANALYSIS

CS 448B | Fall 2023

MANEESH AGRAWALA

1

2

READING RESPONSE: QUESTIONS/THOUGHTS

- I really want to know when you know your visuals are strong? ***Going from a table of data to a visual, what steps are taken to decide if one should create a novel visual or a standard graph?*** How many times do professionals usually re-do their visualizations before they realize they have gained new knowledge?
- ***Can insights into controlled and automatic processing guide the development of advanced visualization techniques?*** Exploring new, automatically processable features could lead to innovations in information visualization, enhancing our ability to convey complex data through visual representations.
- (Regarding Tufte's quote "There are right and wrong ways to show data; there are displays that reveal the truth and displays that do not") […] my initial thoughts were that I don't see how a right display of data that reveals the truth can be possible, or at least not the whole truth. In every visualization of data, choices are being made about what data to include, colors, titles, sizes, icons, and more. ***And with those choices might come bias or a perspective not relatable to everyone.***

3

TODAY

Learning Objectives

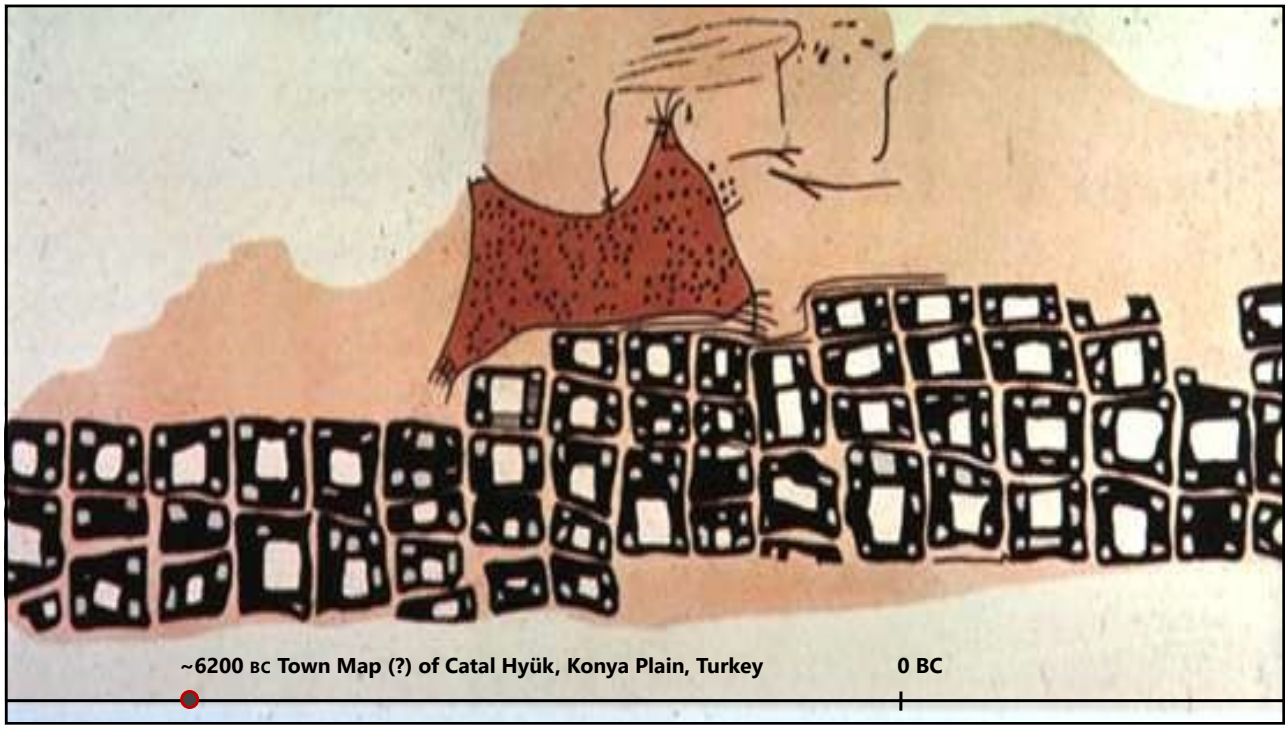
1. What is exploratory data analysis and why is it important?
2. What factors should we consider when exploring a dataset?
3. How do visualization researchers design tools to support exploratory data analysis?

4

WHAT WAS THE FIRST DATA VISUALIZATION?

6

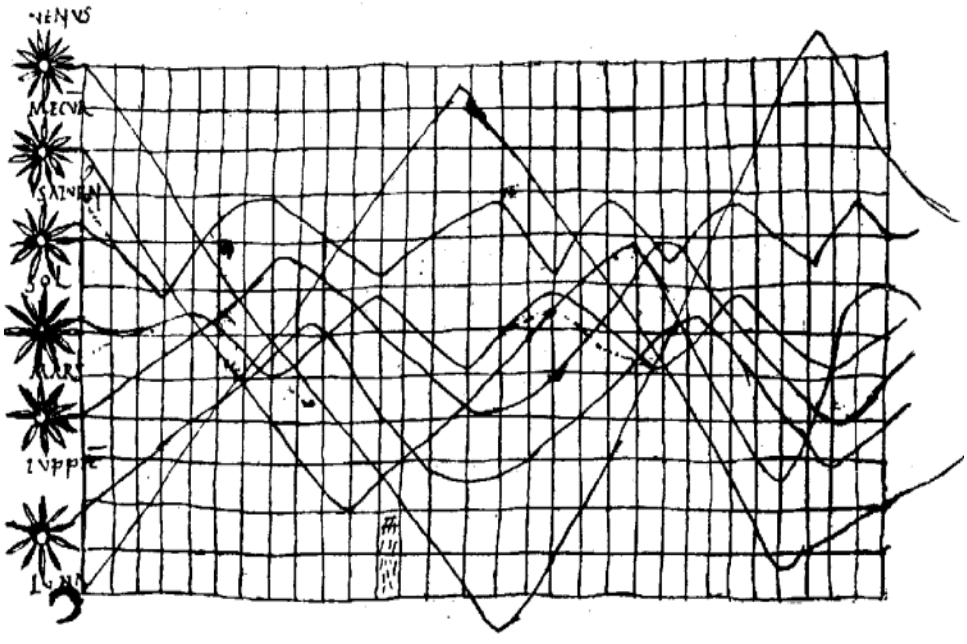
7



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

0 BC

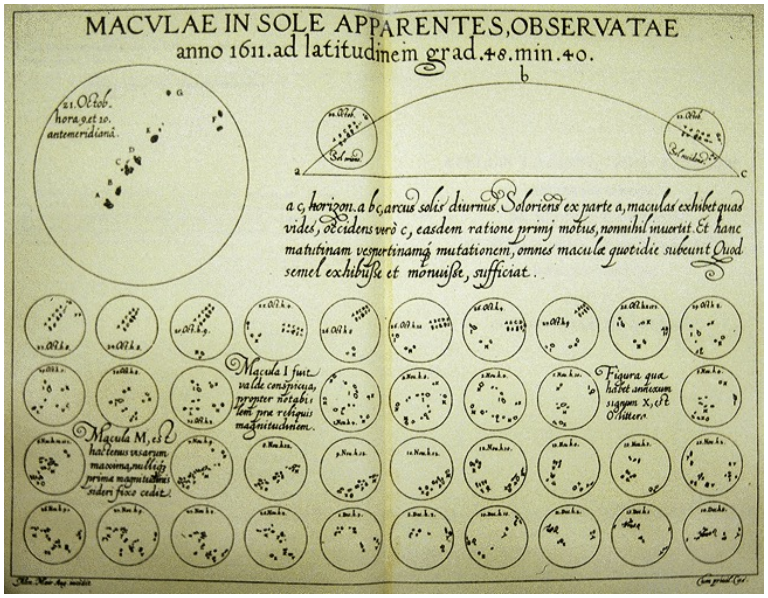
8



0 BC

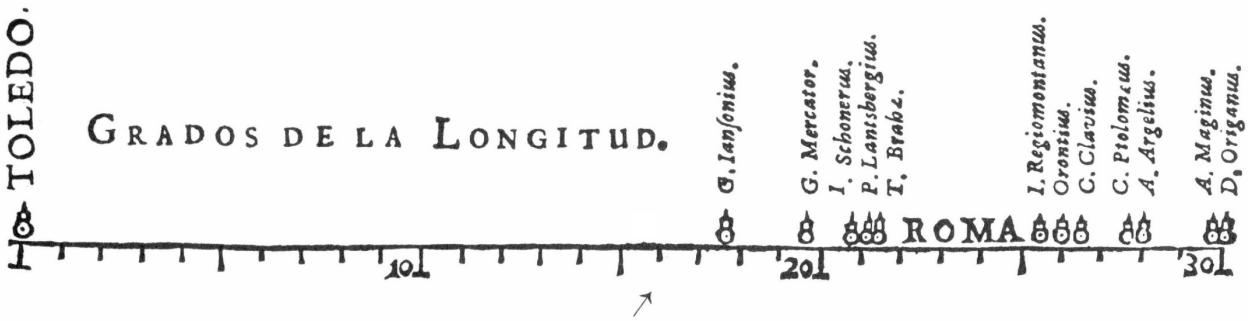
~950 AD Position of Sun, Moon and Planets

9



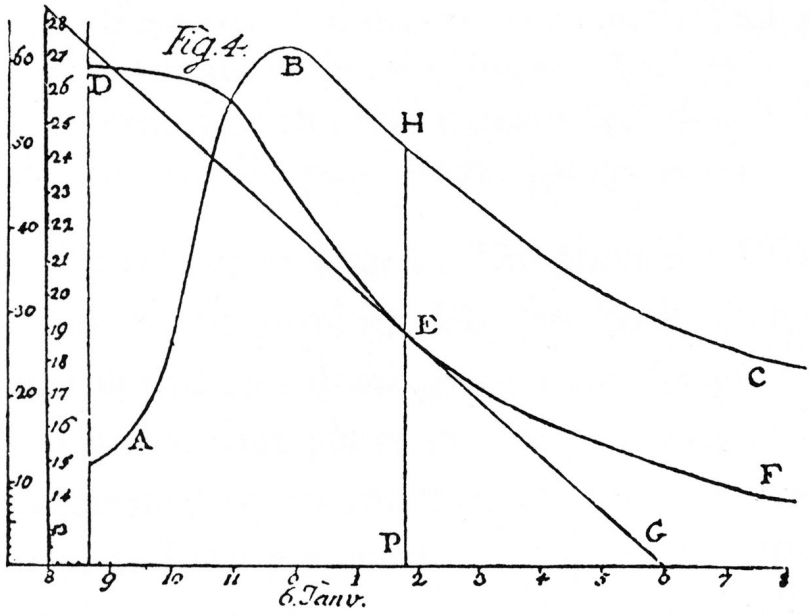
Sunspots over time, Scheiner 1626

10



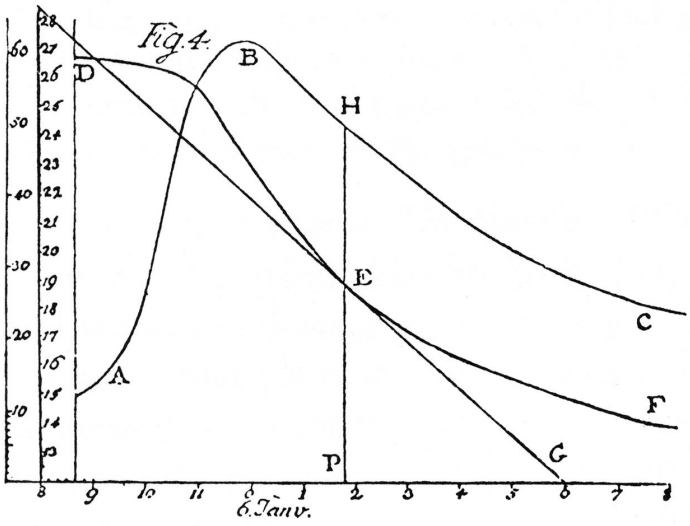
Longitudinal distance between Toledo and Rome, van Langren 1644

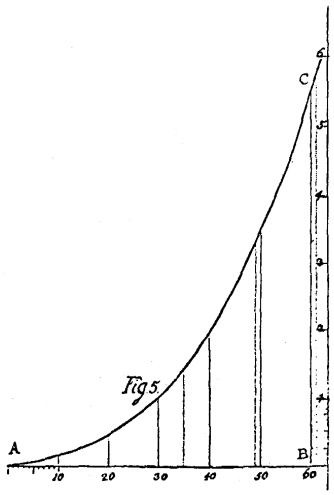
11



The Rate of Water Evaporation, Lambert 1765

12





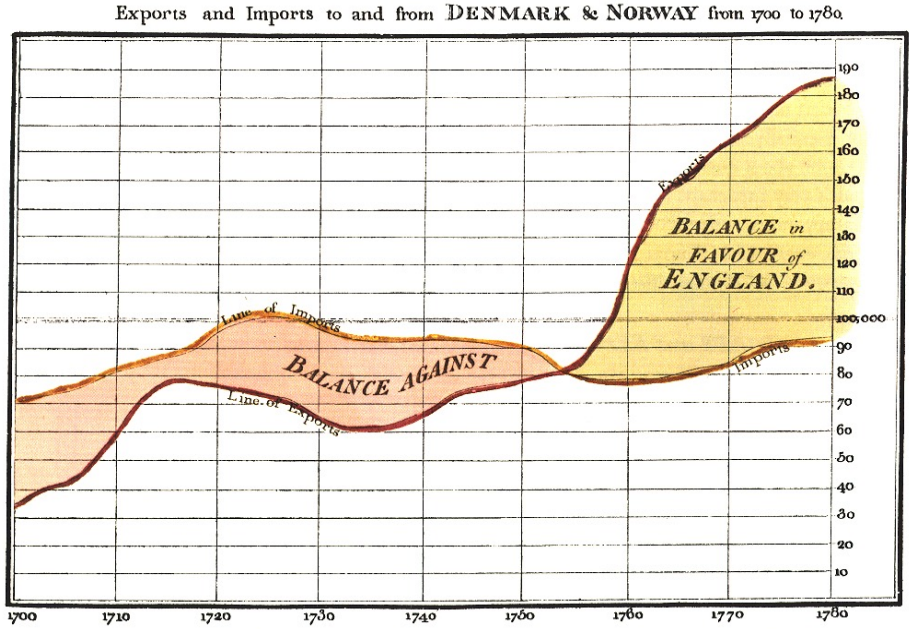
The Rate of Water Evaporation, Lambert 1765

13

THE GOLDEN AGE OF VISUALIZATION

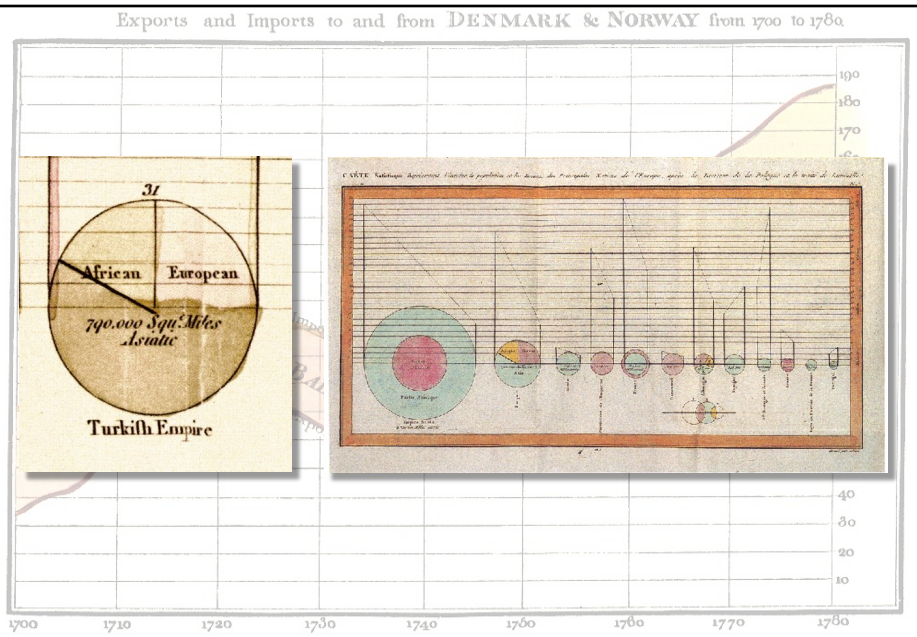
1786 1900

14



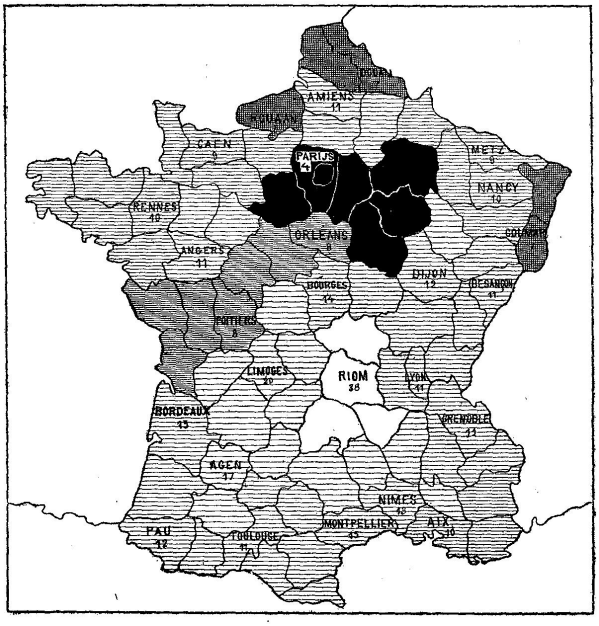
The Commercial and Political Atlas, William Playfair 1786

15



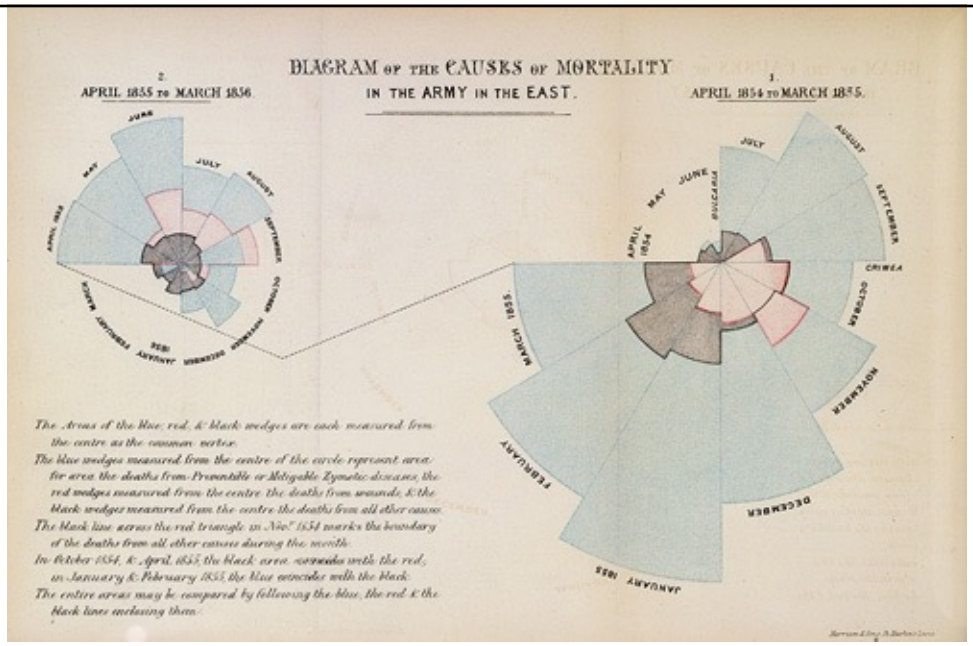
Statistical Breviary, William Playfair 1801

16



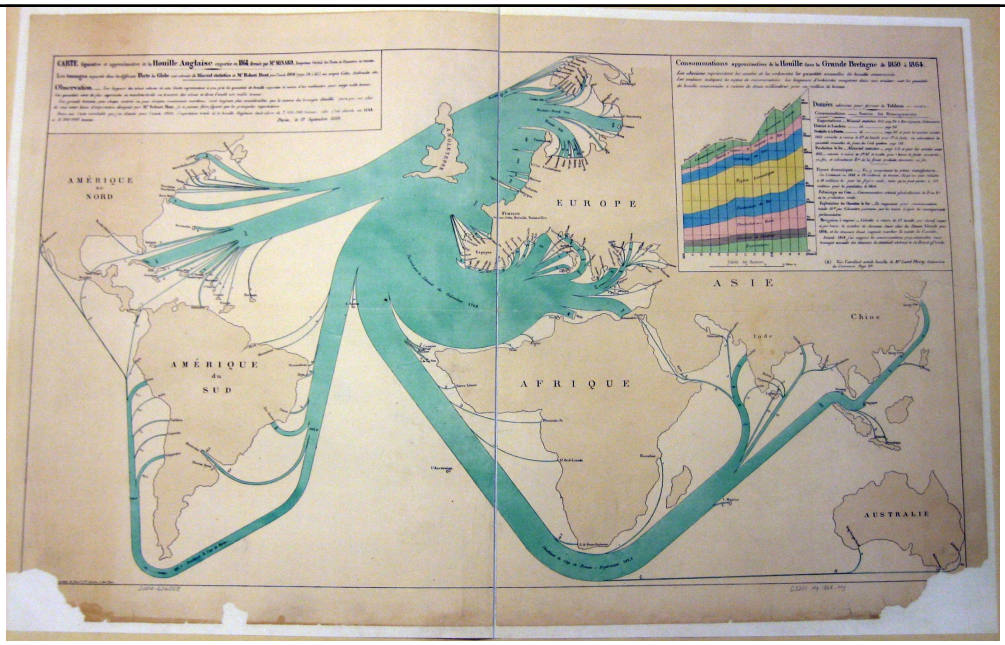
1786 1826(?) Illiteracy in France, Pierre Charles Dupin

17



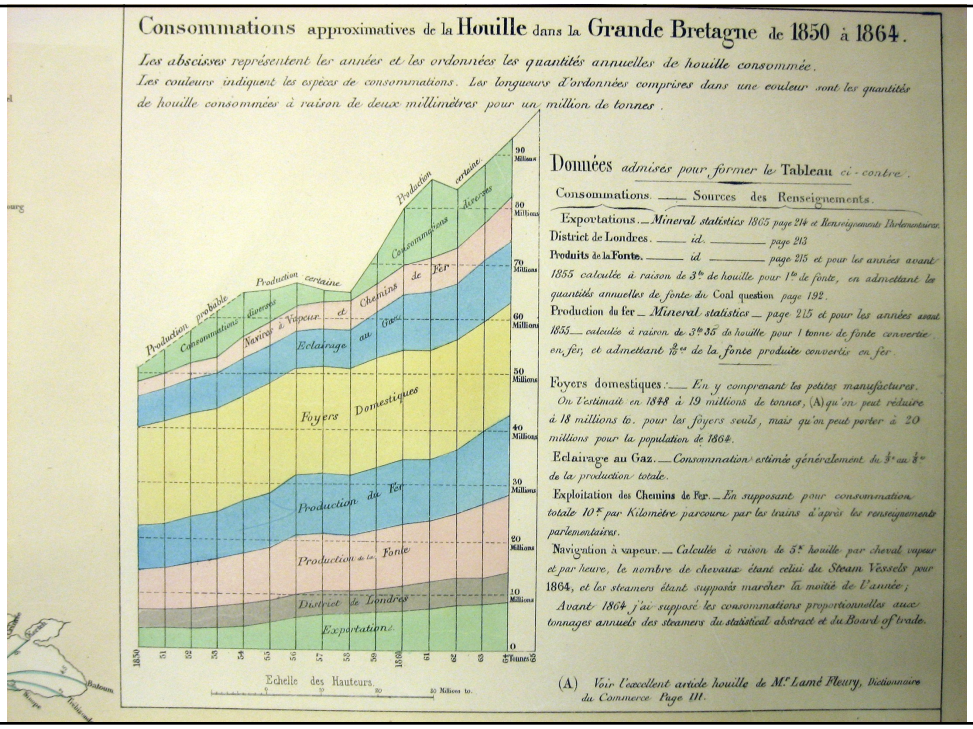
1786 **1856 "Coxcomb" of Crimean War Deaths, Florence Nightingale**

18

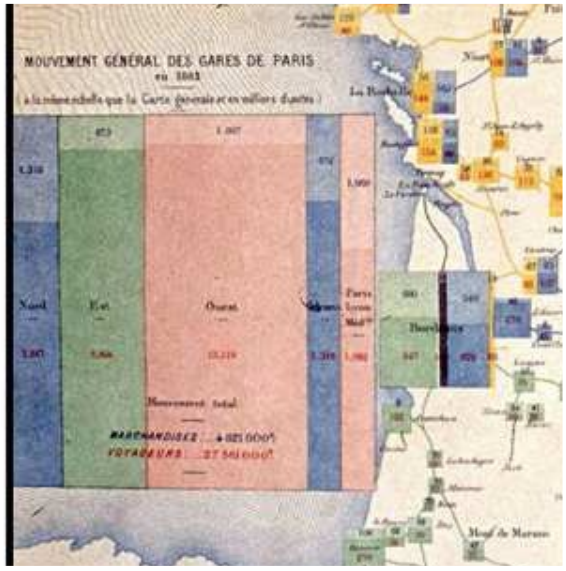


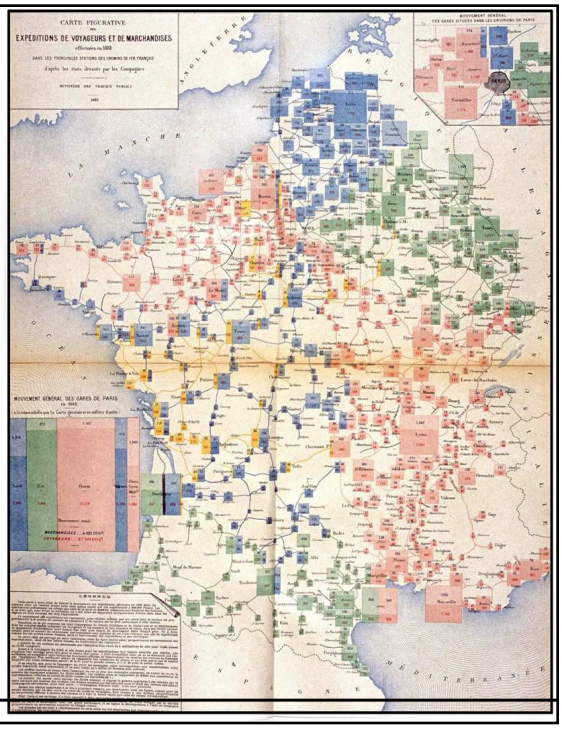
1786 **1864 British Coal Exports, Charles Minard**

19



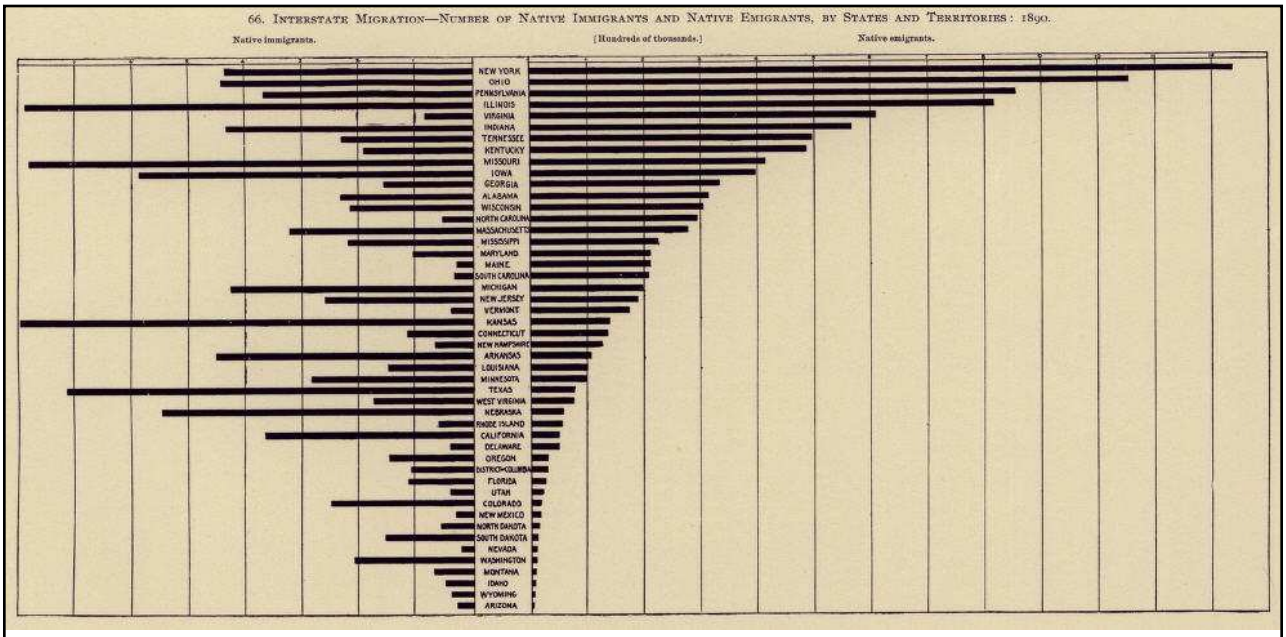
20





1786 — 1884 Rail Passengers and Freight from Paris

21



1786 — 1890 Statistical Atlas of the Eleventh U.S. Census

22

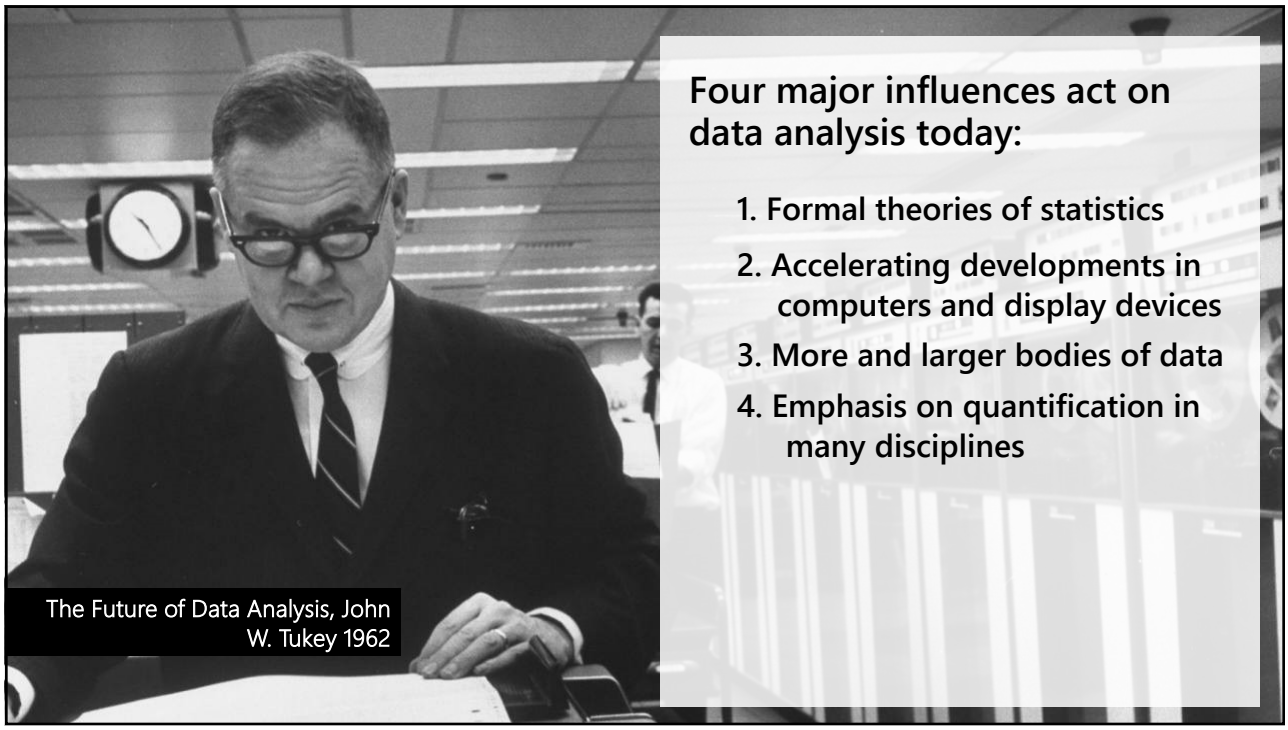
THE RISE OF STATISTICS

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

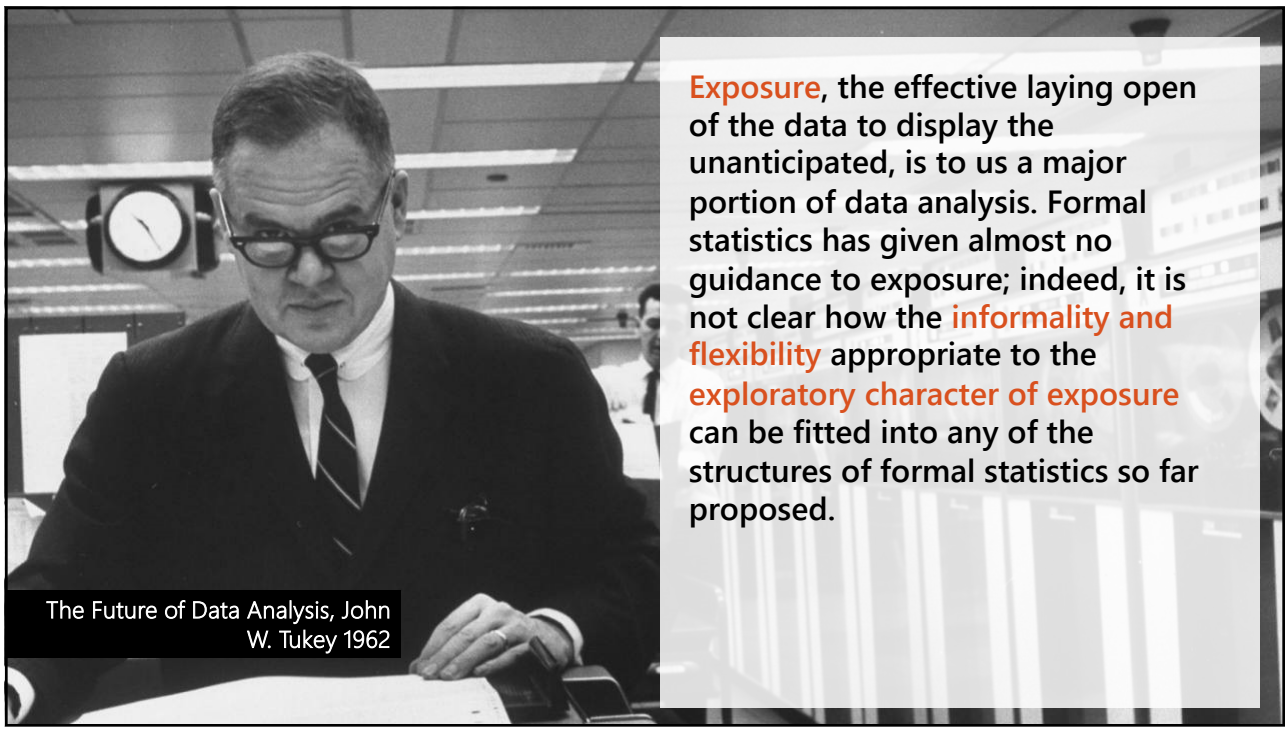
24



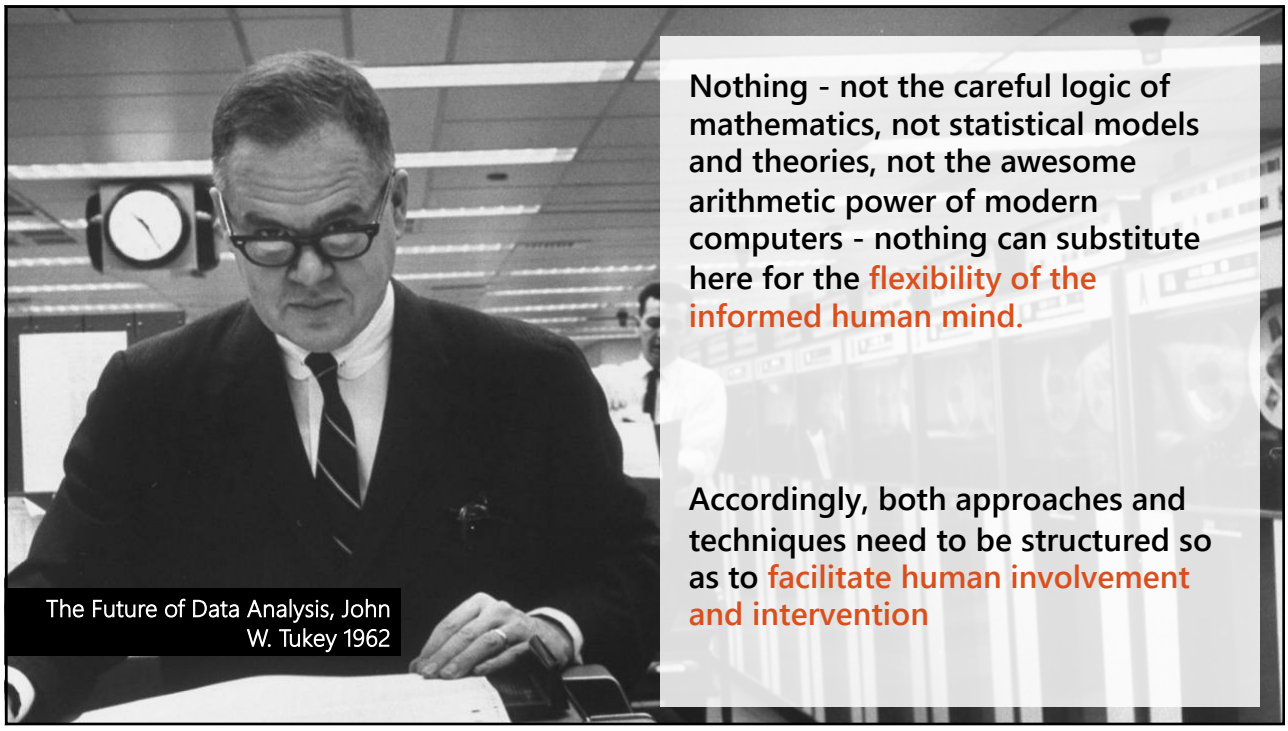
Four major influences act on data analysis today:

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

25


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.
The Future of Data Analysis, John W. Tukey 1962

28


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
The Future of Data Analysis, John W. Tukey 1962

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

Big Data Borat
@BigDataBorat

Following

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

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

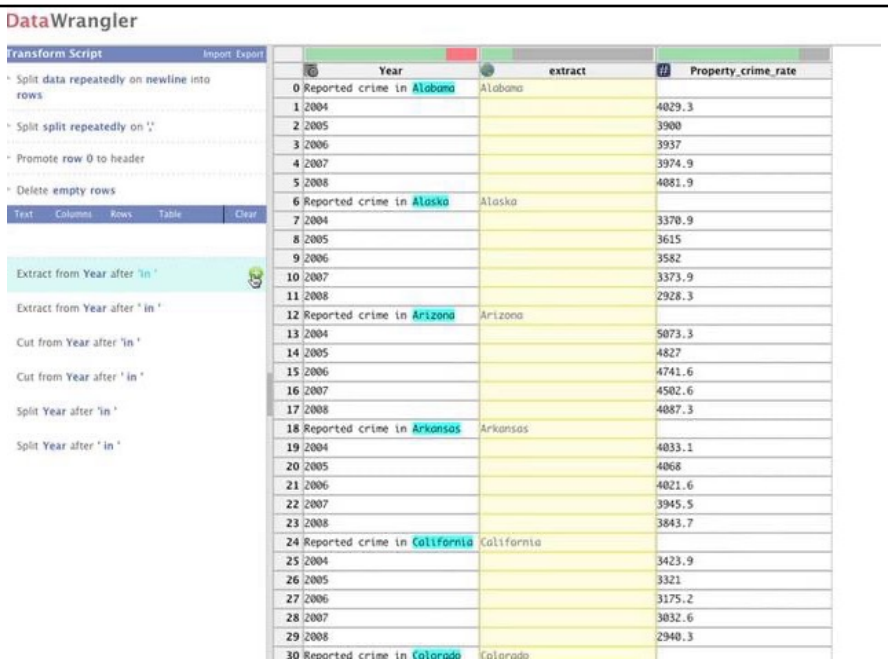
ARQUERO

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

```
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



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

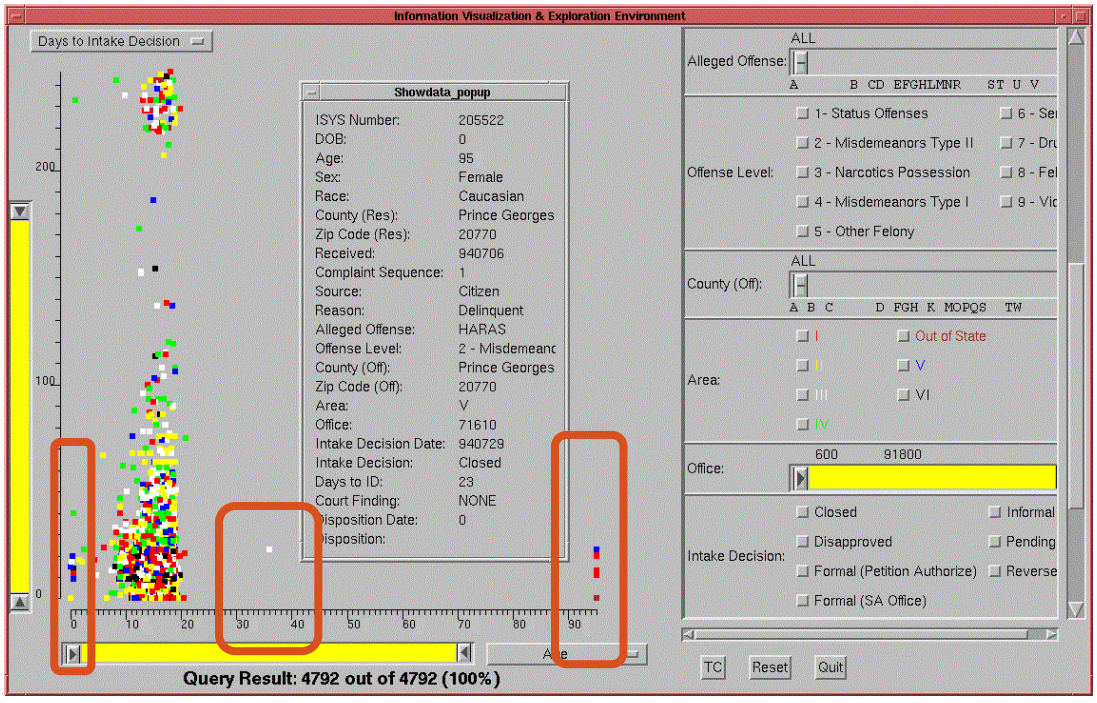
39

"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

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

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

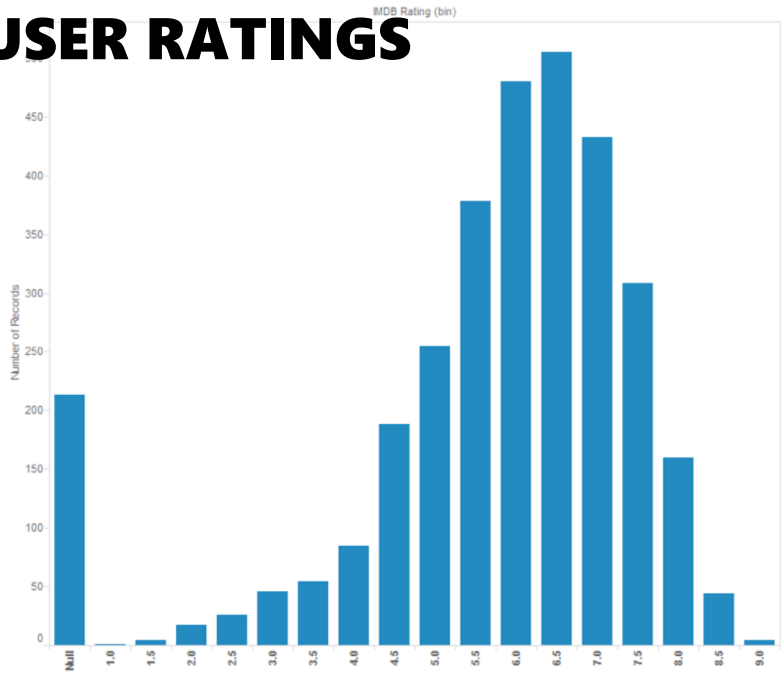
48

MOTION PICTURES DATA TYPES

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

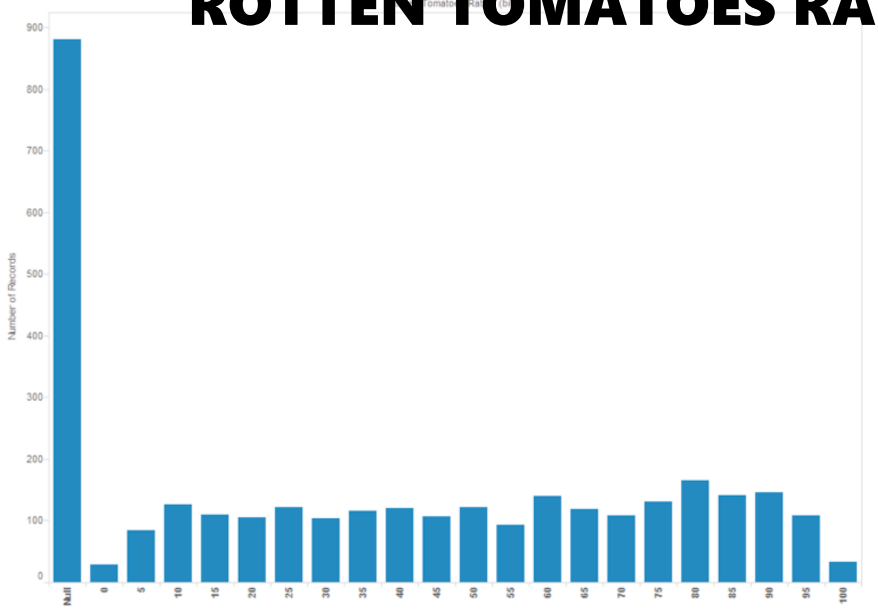
49

IMDB USER RATINGS

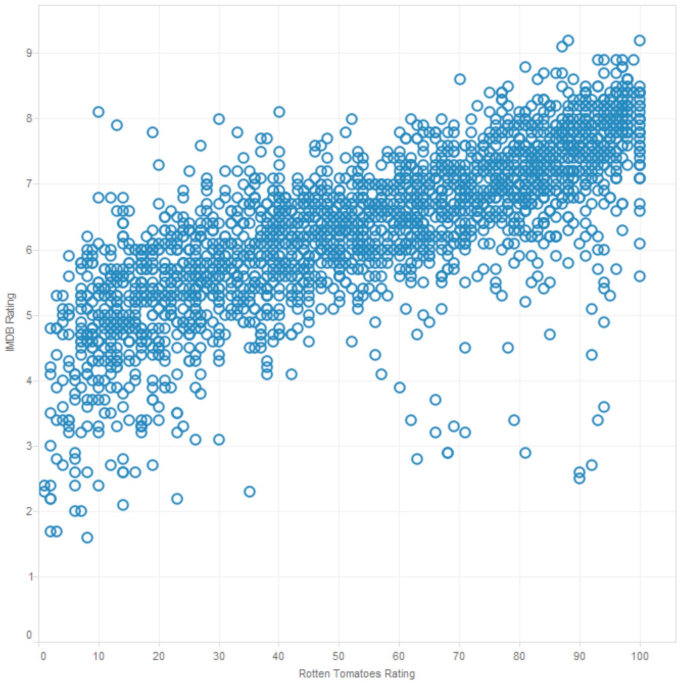


50

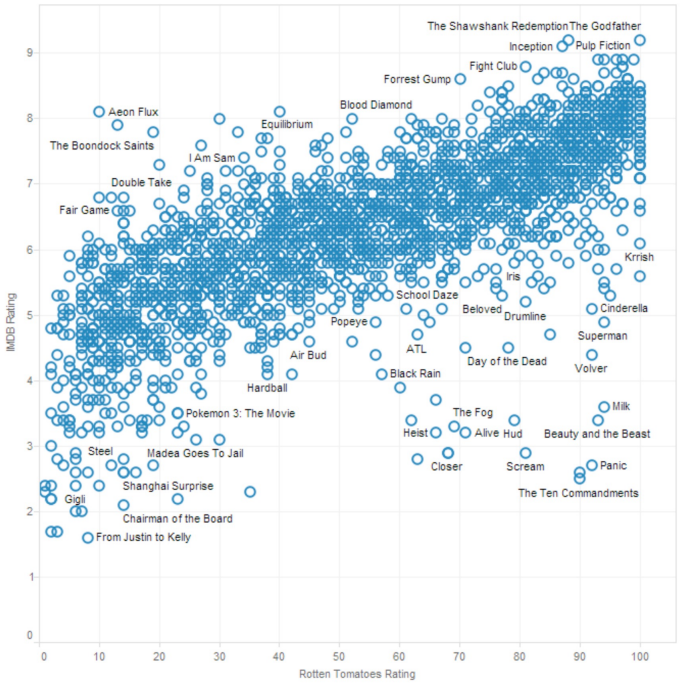
ROTTEN TOMATOES RATINGS



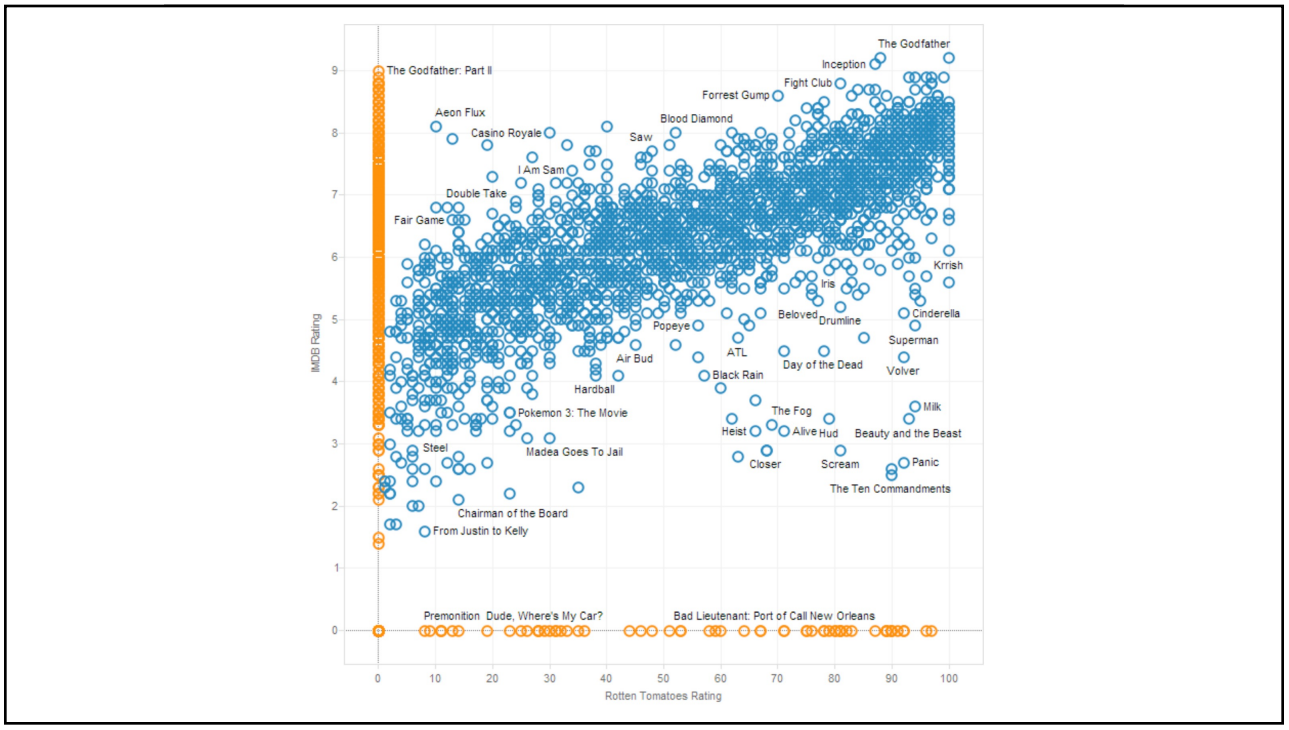
51


IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100

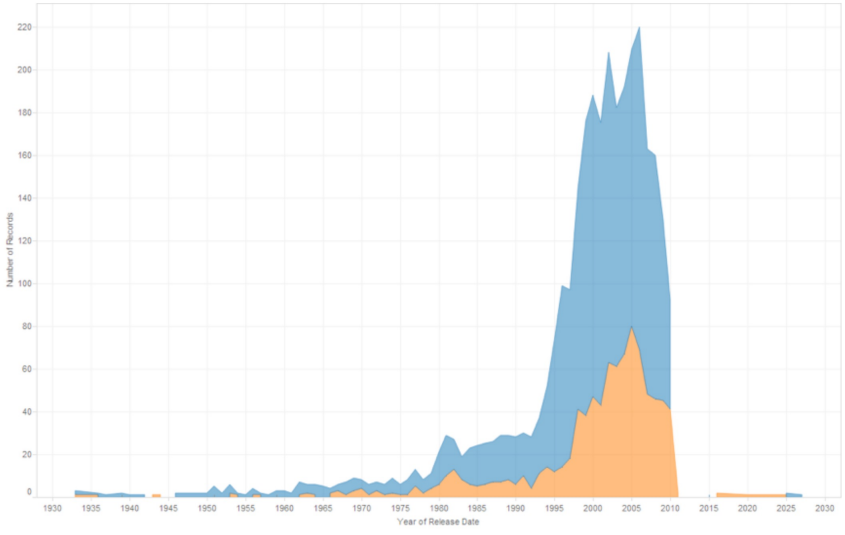
52


The Shawshank Redemption
The Godfather
Inception
Pulp Fiction
Fight Club
Forrest Gump
Blood Diamond
Aeon Flux
Equilibrium
The Boondock Saints
I Am Sam
Double Take
Fair Game
Krrish
Iris
School Daze
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Milk
The Fog
Pokemon 3: The Movie
Heist
Alive Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
Shanghai Surprise
The Ten Commandments
Gigli
Chairman of the Board
From Justin to Kelly
IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100

53


The Godfather
Inception
Fight Club
Forrest Gump
The Godfather: Part II
Aeon Flux
Casino Royale
Blood Diamond
Saw
I Am Sam
Double Take
Fair Game
Krrish
Iris
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Milk
The Fog
Pokemon 3: The Movie
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
The Ten Commandments
Chairman of the Board
From Justin to Kelly
Premonition
Dude, Where's My Car?
Bad Lieutenant: Port of Call New Orleans
IMDB Rating
Rotten Tomatoes Rating

54


Number of Records
Year of Release Date

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

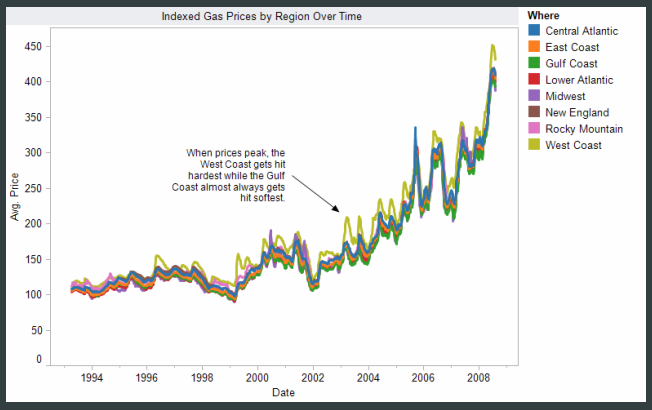
Due 10/16 11:30am

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

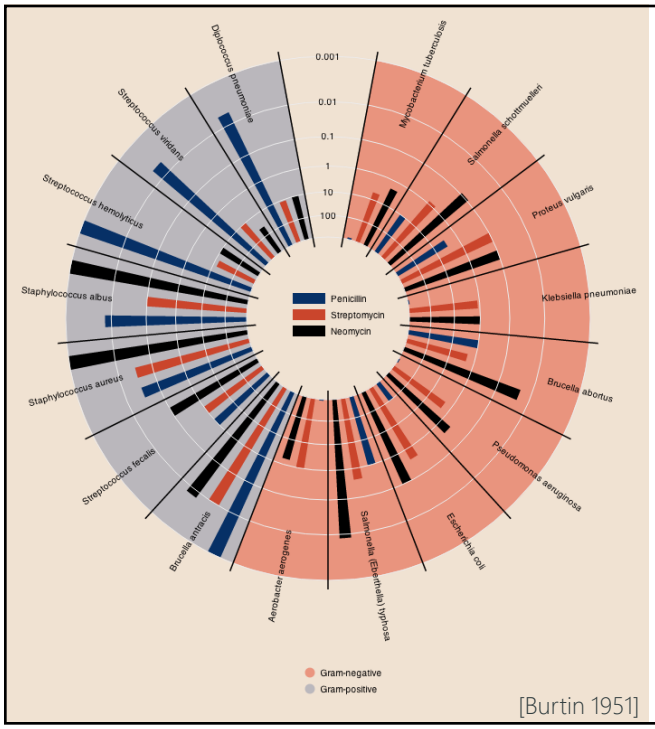
60

WHAT QUESTIONS MIGHT WE ASK?

Table 1: Burtin's data.

Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
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

61

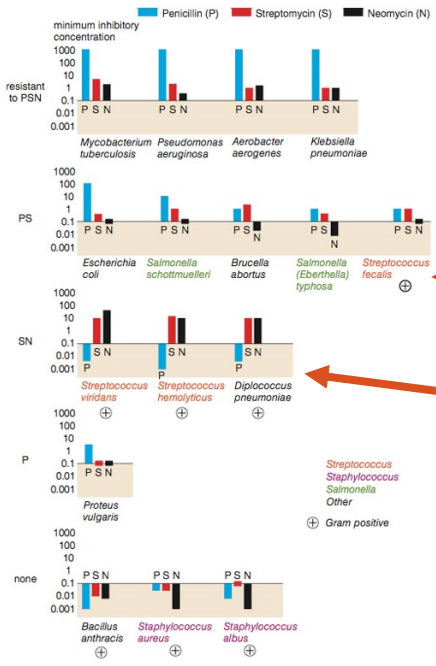


[Burtin 1951]

HOW DO DRUGS COMPARE?

Radius: 1/log(MIC)
Bar Color: Antibiotic
Background Color: Gram Staining

62



HOW DO BACTERIA GROUP W.R.T ANTIBIOTIC RESISTANCE?

Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

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