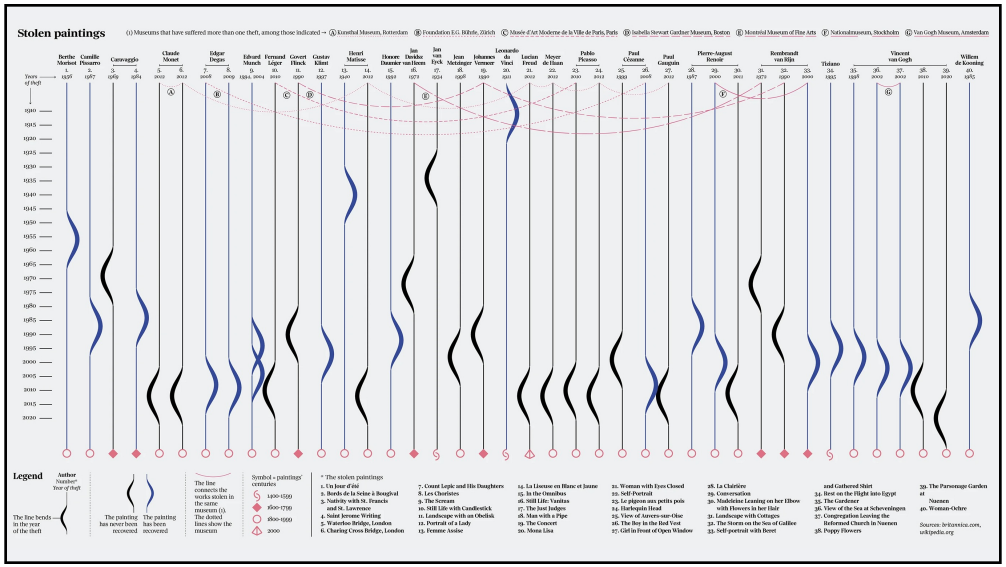
EXPLORATORY DATA ANALYSIS

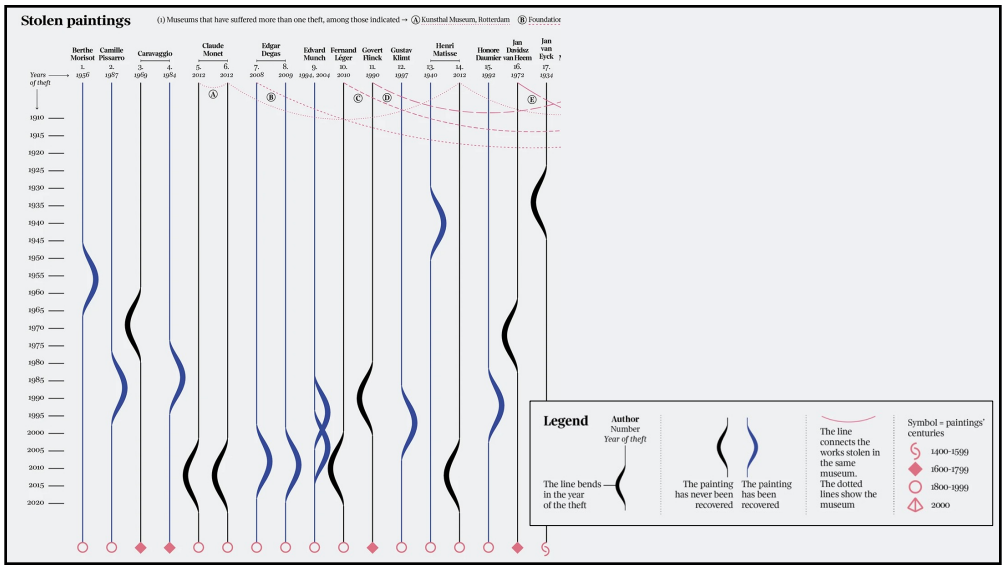
CS 448B | Fall 2025

MANEESH AGRAWALA

1



3



4

READING RESPONSE: QUESTIONS/THOUGHTS

The section **on spatial versus non-spatial data** also raised questions for me. **For information with a natural spatial mapping, like showing ozone levels on a globe, the visualization feels intuitive. But for abstract data like financial transactions, designers have to invent new mappings.** For example, mapping stock prices onto time series graphs feels "obvious" now, but it's a convention, not a natural truth. How many of the "standard" ways we visualize abstract data are shaping how we see patterns?

… **while I think this ranking is a good general principle at first glance, I am skeptical of how useful it is to draw these lines when creating visuals.** While I understand that this ordering is meant moreso as a guideline to what is traditionally effective, I do wonder how much the ordering holds in specific cases. For example, and maybe I hold bias because I am colorblind, but I see that for nominal variables hue is ranked significantly above shape in importance, which from looking at the various visualizations in the readings this week."

5

TODAY

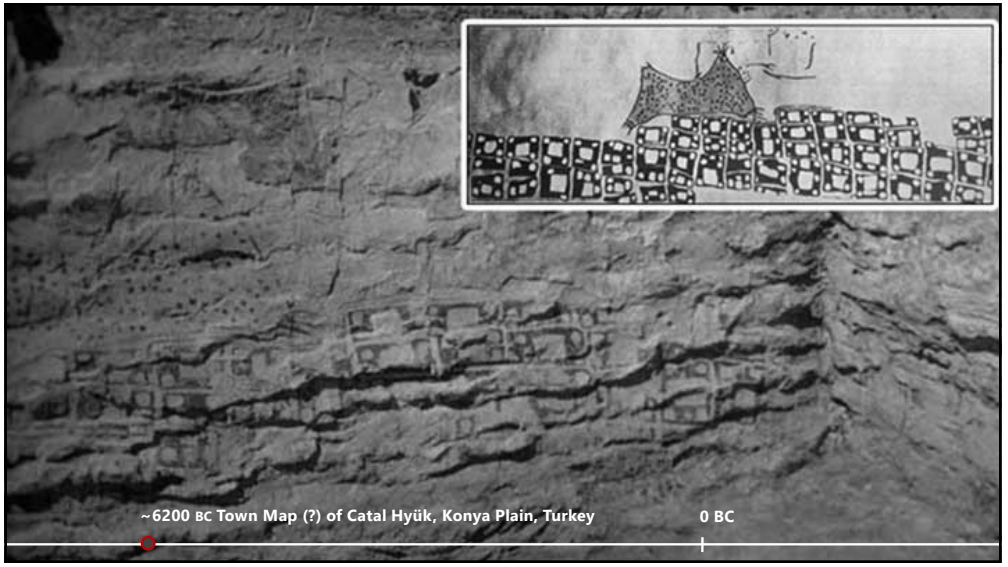
Learning Objectives

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?

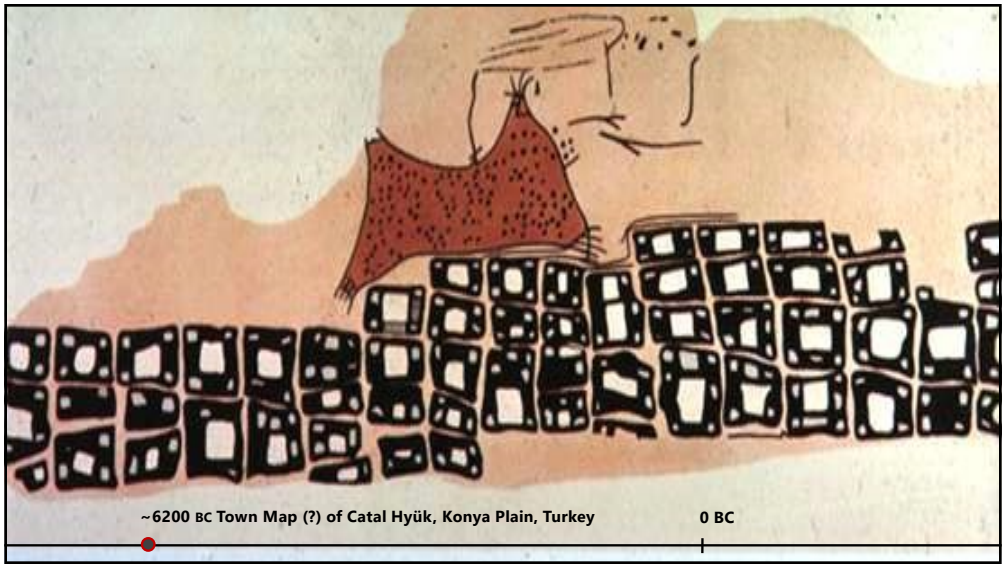
6

WHAT WAS THE FIRST DATA VISUALIZATION?

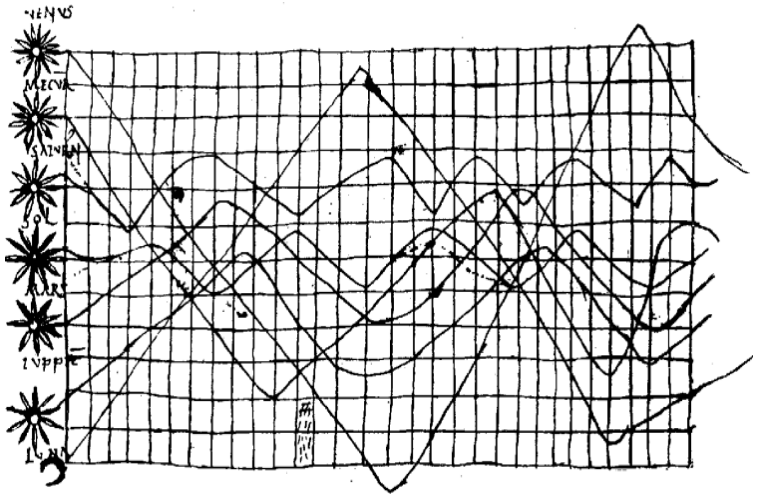
8


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

9

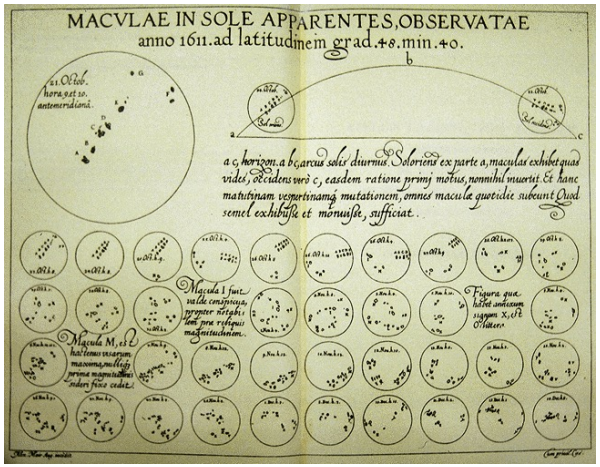

~6200 BC Town Map (?) of Catal Hyük, Konya Plain, Turkey
0 BC

10



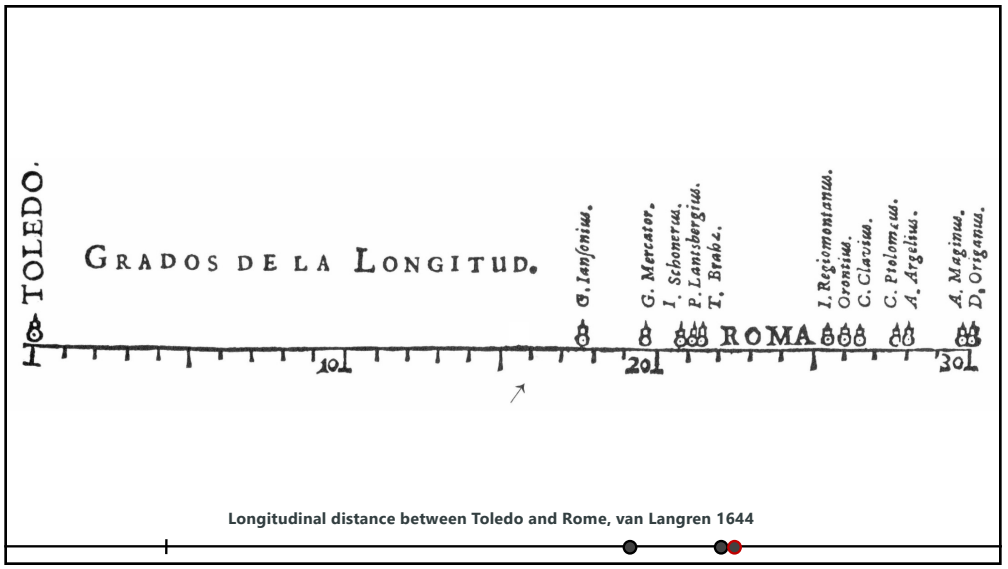
0 BC ~950 AD Position of Sun, Moon and Planets

11



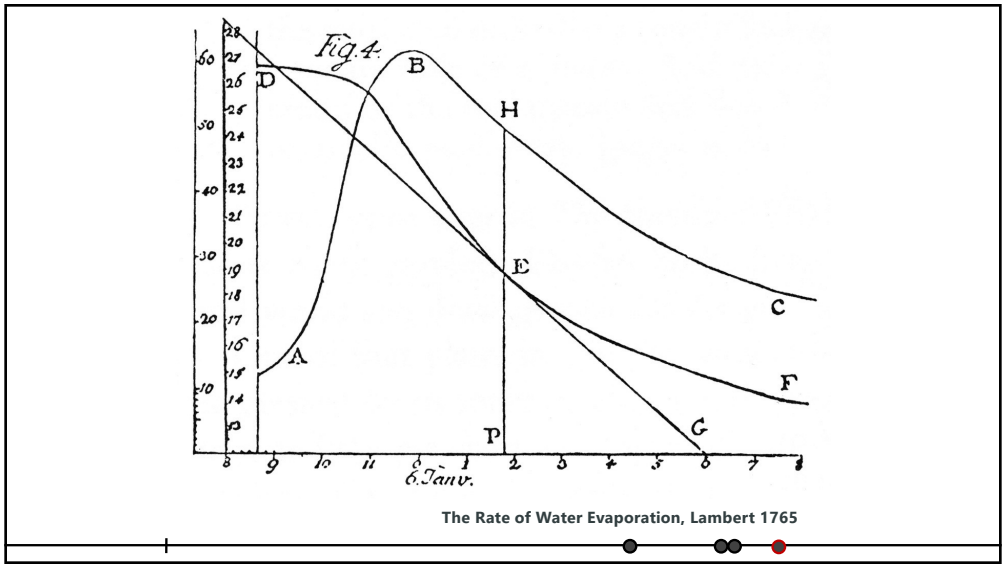
Sunspots over time, Scheiner 1626

12



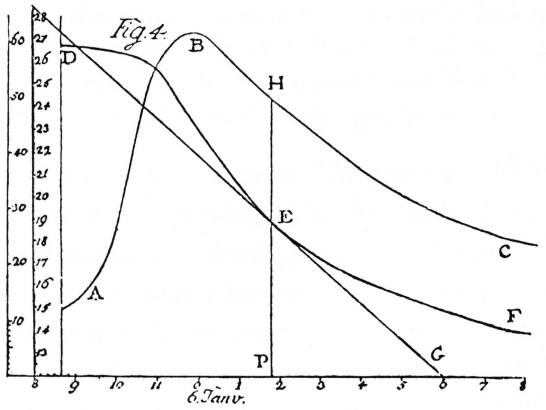
Longitudinal distance between Toledo and Rome, van Langren 1644

13



The Rate of Water Evaporation, Lambert 1765

14



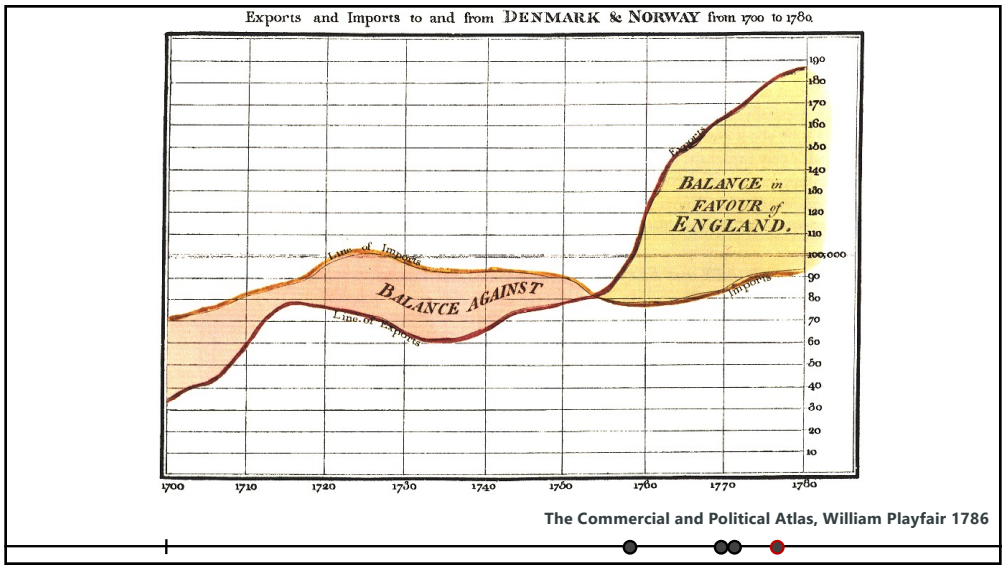
The Rate of Water Evaporation, Lambert 1765

15

THE GOLDEN AGE OF VISUALIZATION

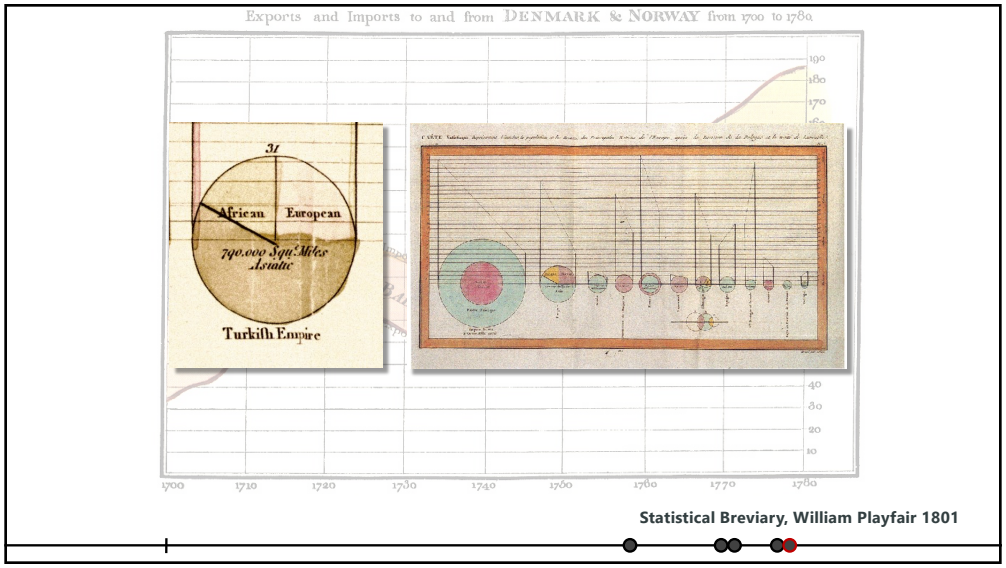
1786 1900

16



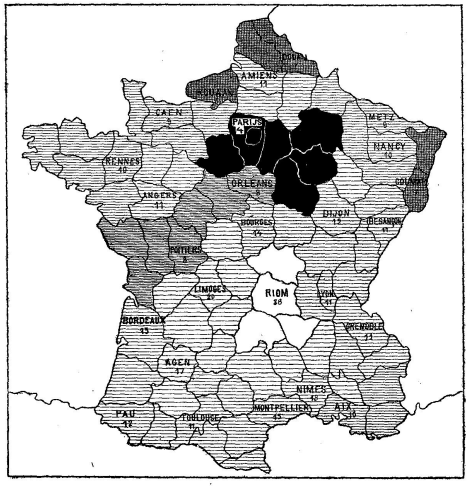
The Commercial and Political Atlas, William Playfair 1786

17



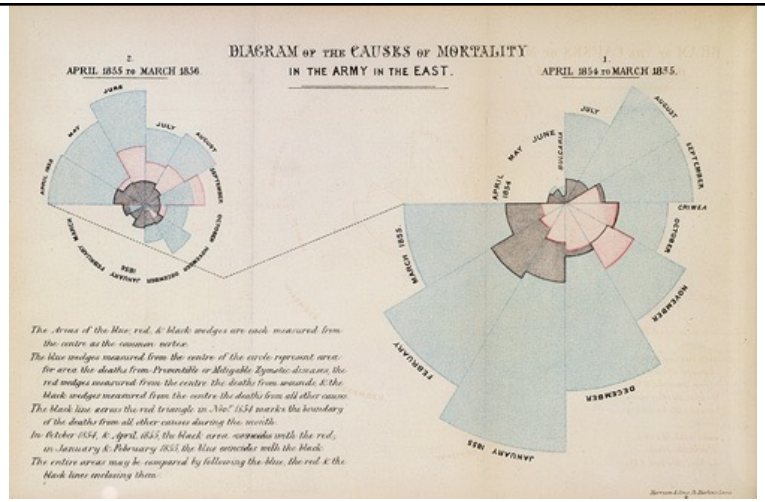
Statistical Breviary, William Playfair 1801

18



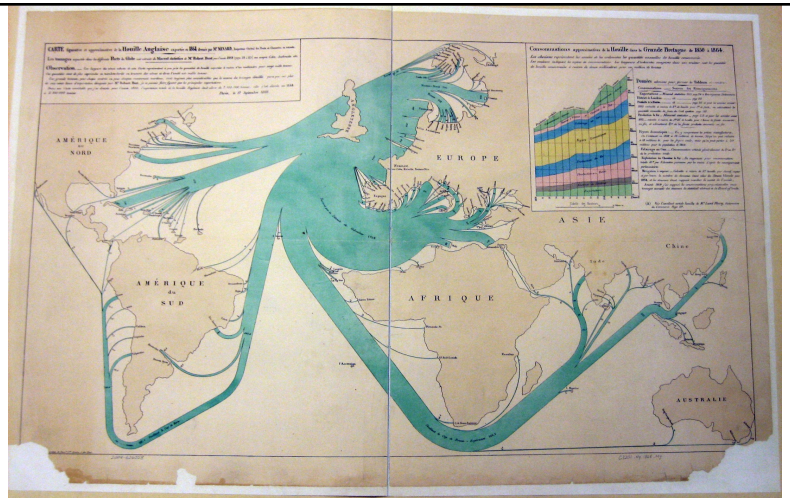
1786 **1826(?) Illiteracy in France, Pierre Charles Dupin**

19



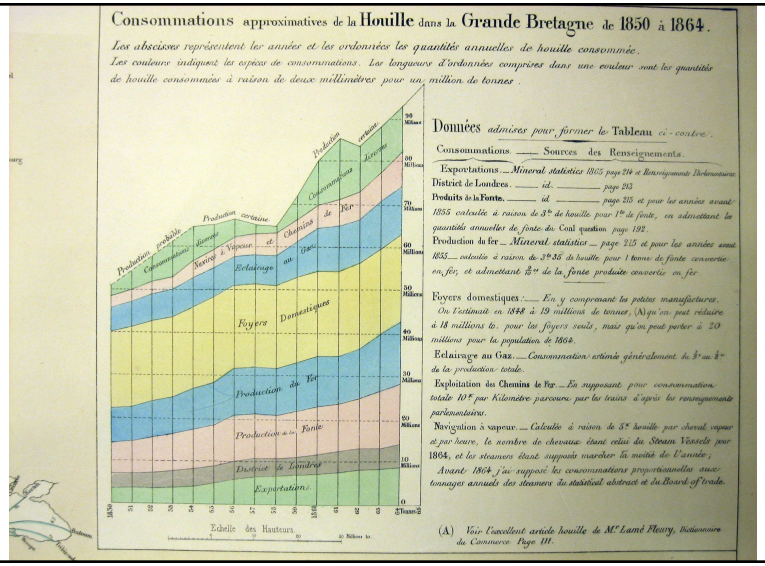
1786 **1856 “Coxcomb” of Crimean War Deaths, Florence Nightingale**

20

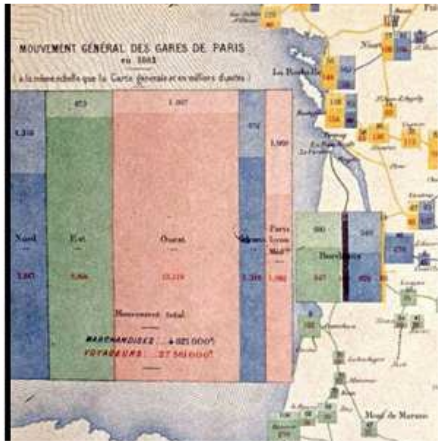


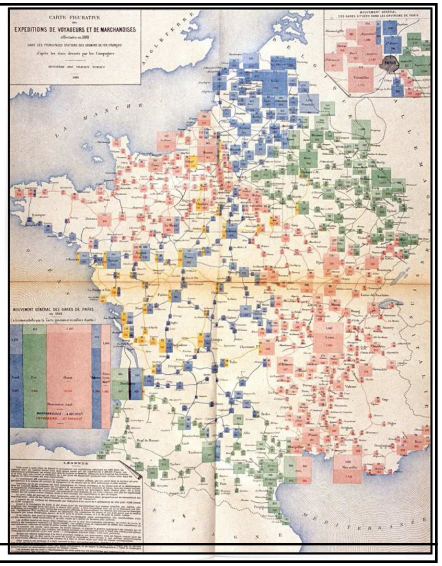
1786 **1864 British Coal Exports, Charles Minard**

21



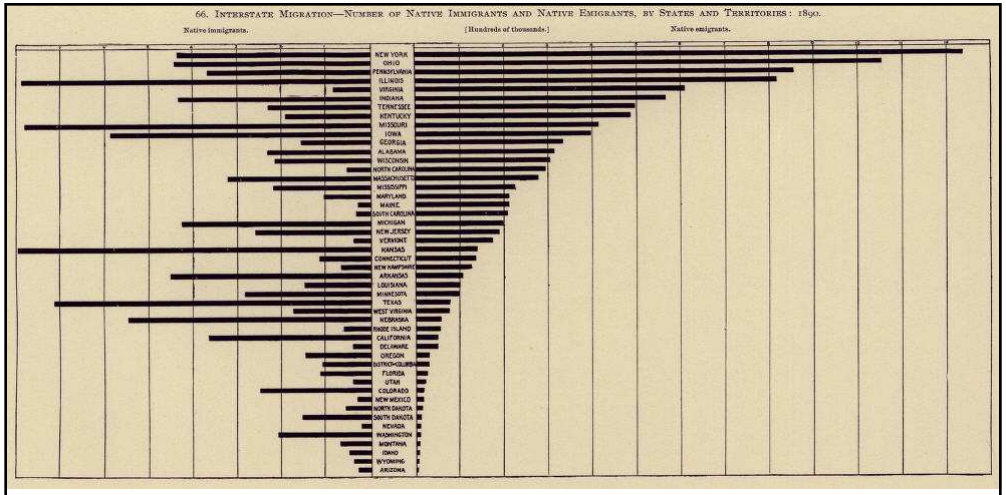
22





1786 — 1884 Rail Passengers and Freight from Paris

23



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

24

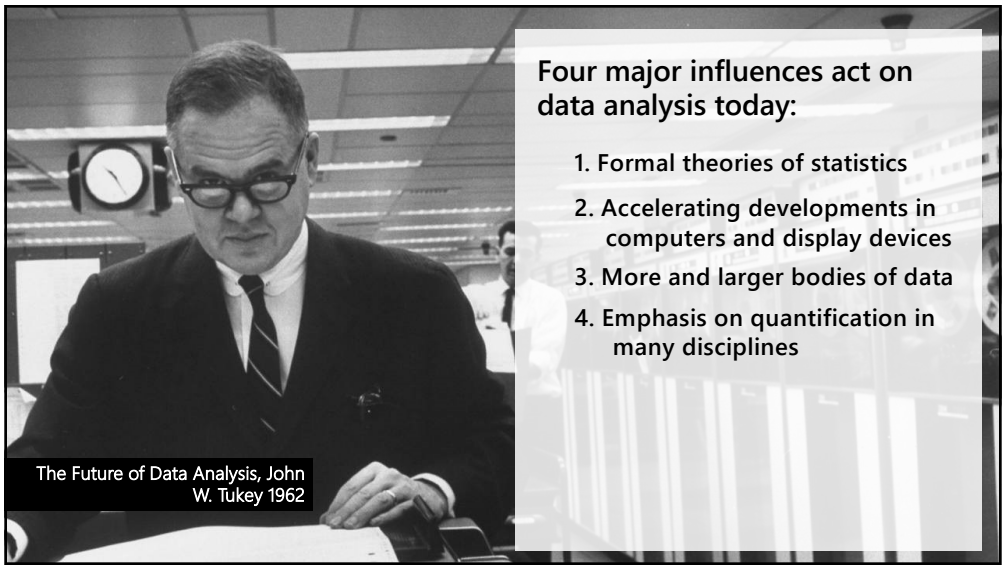
THE RISE OF STATISTICS

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

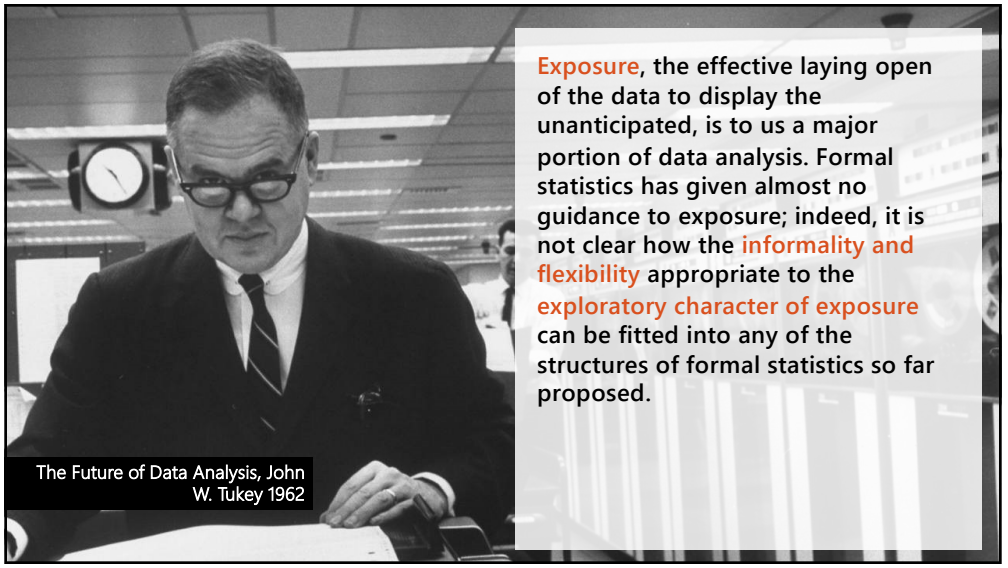
1786 1900 1950

25

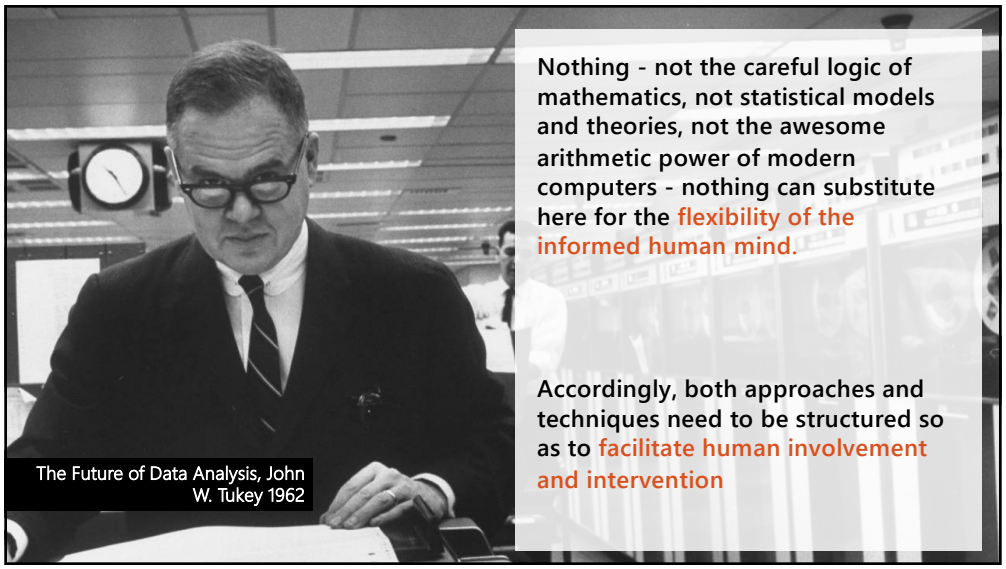
26


The Future of Data Analysis, John W. Tukey 1962
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

27


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.

30



31

DATA WRANGLING

34

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]

35

Big Data Borat
@BigDataBorat

Following

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

36

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

37

Dataset

Data Field

Halloween Candy Power Ranking Dataset

	competitorname	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent	winpercent
2	100 Grand	1	0	1	0	0	1	0	1	0	.73199999	.86000001	66.971725
3	3 Musketeers	1	0	0	0	1	0	0	1	0	.60399997	.51099998	67.602936
4	One dime	0	0	0	0	0	0	0	0	0	.011	.116	32.261086
5	One quarter	0	0	0	0	0	0	0	0	0	.011	.51099998	46.116505
6	Air Heads	0	1	0	0	0	0	0	0	0	.90600002	.51099998	52.341465
7	Almond Joy	1	0	0	1	0	0	0	1	0	.465	.76700002	50.347546
8	Baby Ruth	1	0	1	1	1	0	0	1	0	.60399997	.76700002	56.914547
9	Boston Baked Beans	0	0	0	1	0	0	0	0	1	.31299999	.51099998	23.417824
10	Candy Corn	0	0	0	0	0	0	0	0	1	.90600002	.32499999	38.010963
11	Caramel Apple Pops	0	1	1	0	0	0	0	0	0	.60399997	.32499999	34.517681
12	Charleston Chew	1	0	0	0	1	0	0	1	0	.60399997	.51099998	38.975037
13	Chewey Lemonhead Fruit Mix	0	1	0	0	0	0	0	0	1	.73199999	.51099998	36.017628

Data Item/Observation

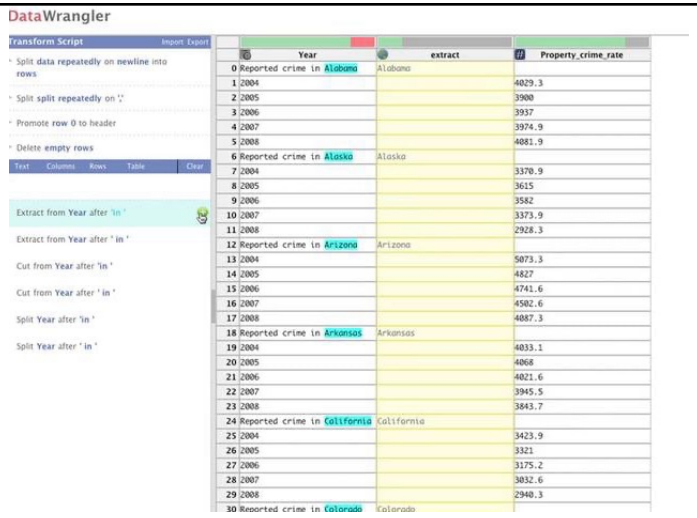
Cell Value

https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/

38

```
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
```

39



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

40

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)
```

41

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

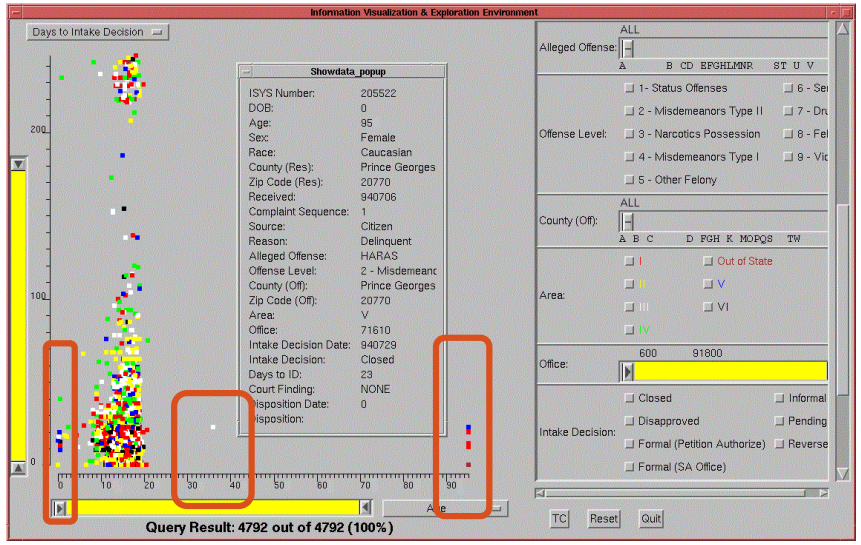
42

"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

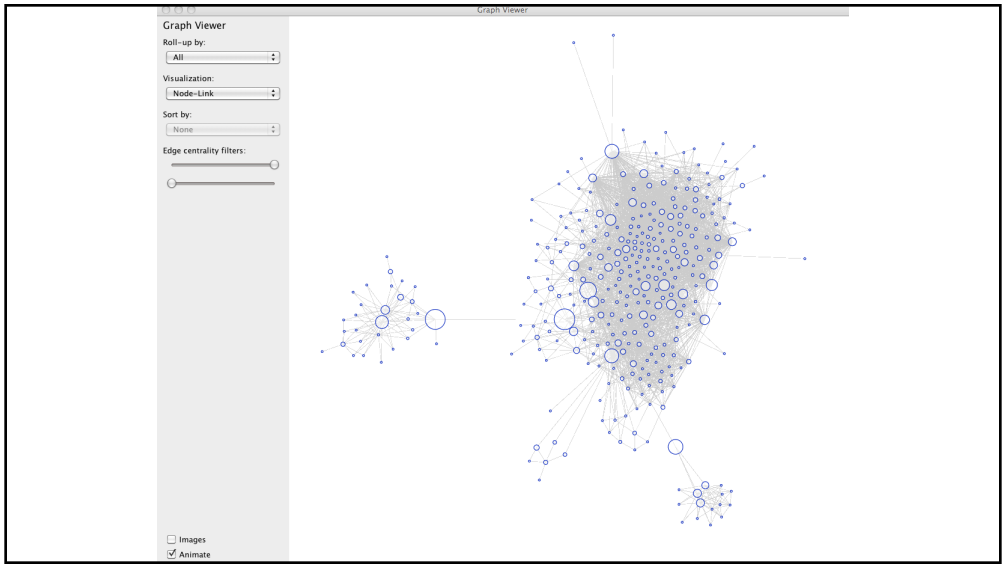
43



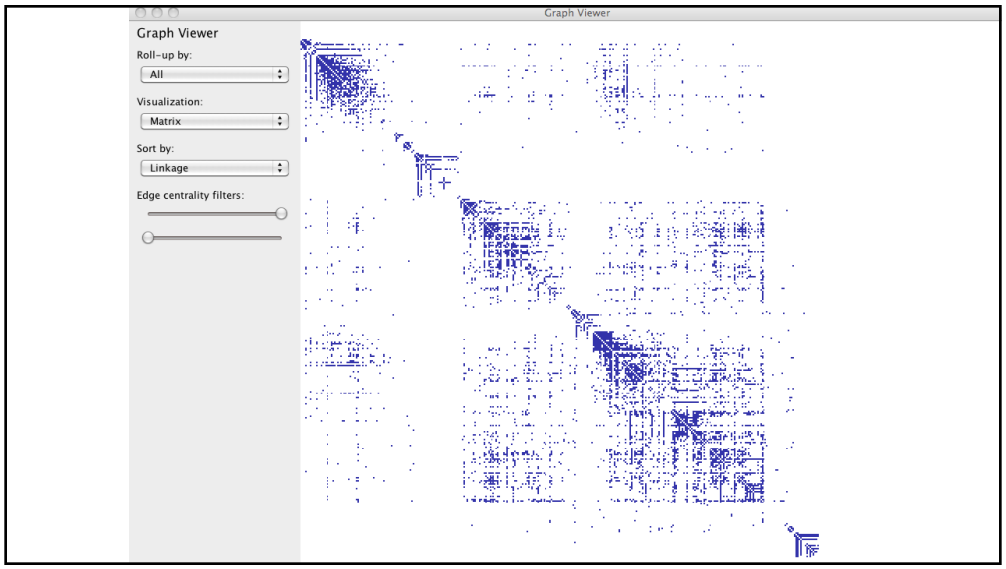
44


facebook
Profile edit
Friends
Networks
Inbox
home account privacy logout
Search
Applications edit
Photos
Groups
Events
Marketplace
The New York Times News Quiz
more
Send Bill a Gift
Send Bill a Message
Poke Bill!
Friends
See All
Bill Gates
is glad he finally joined facebook and hopes you will too!!! :)
Updated 6 minutes ago
Mini-Feed
Displaying 15 stories
See All
Update: Bill has posted a note:
Friends, I have finally caved and joined facebook, America's fastest-growing social-networking web site! At first I didn't join because you needed a college alumni address, and I never quite got one... Then when the place started opening up to high schools and corporations, with everyone and his grandmother joining, I wanted in. But by then I was mad I didn't have any shares in this $15 billion baby. ... So just now I decided to plunk down $240 million to buy 1.6% of the company from cool kid CEO Mark Zuckerberg. Sure I saw the potential for ad revenue right away -- but this is wild. I've never had my own Facebook page before!
Don't have a lot of friends yet but I've been running into people... Seeing their status updates... Wow, it's a great place to check up on my employees and my kids! They keep saying what their weekend plans are in their status bars. And of course I love how you can add all these little software applications to the page. Or write your very own!
I'm still getting the hang of the whole thing... It seems there are a variety of forums where I can vent to others and display details of my life? O.K.: It wasn't easy being so much smarter than everyone else, pretending to be a grown-up over the telephone so I could get grown-up jobs programming these new things called computers when I was still a child.
At college I led the anti-social group. Never led a social group, or had a social

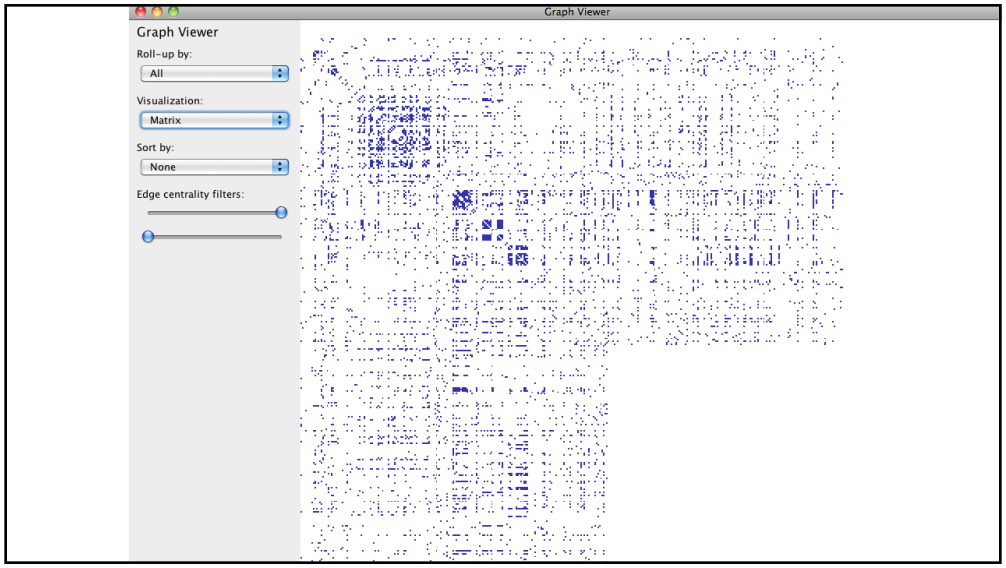
45


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

46


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

47


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

48

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	|||

49

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!

50

ANALYSIS EXAMPLE: MOTION PICTURES DATA

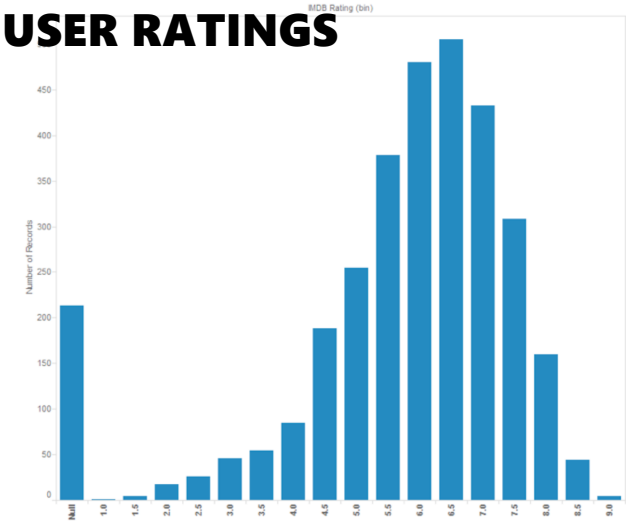
51

MOTION PICTURES DATA TYPES

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

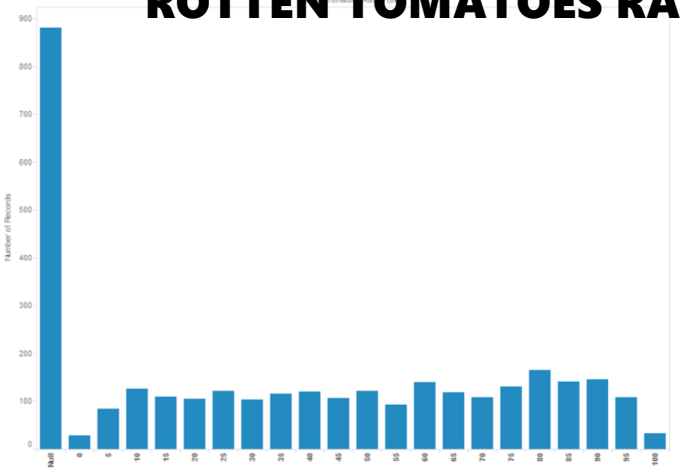
52

IMDB USER RATINGS

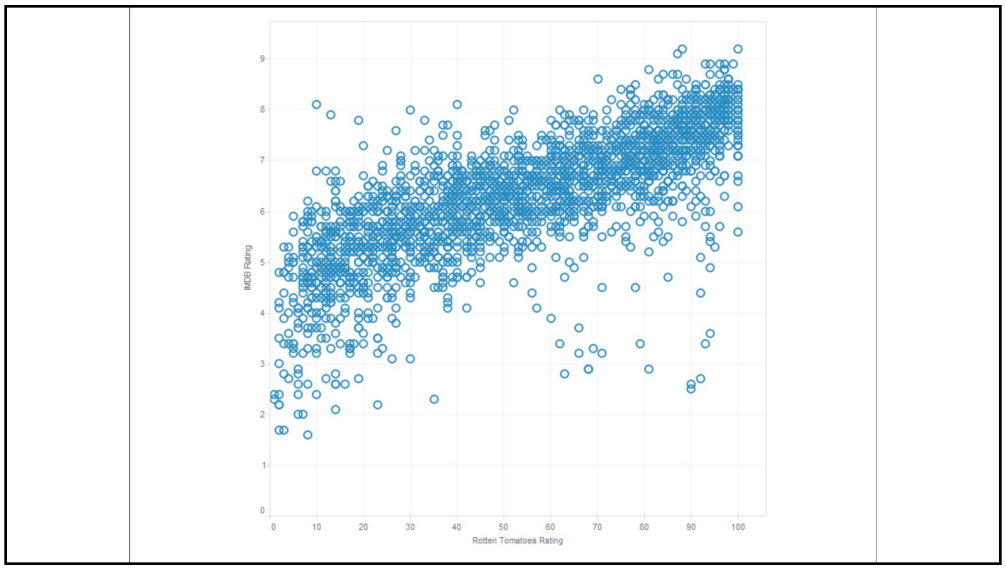


53

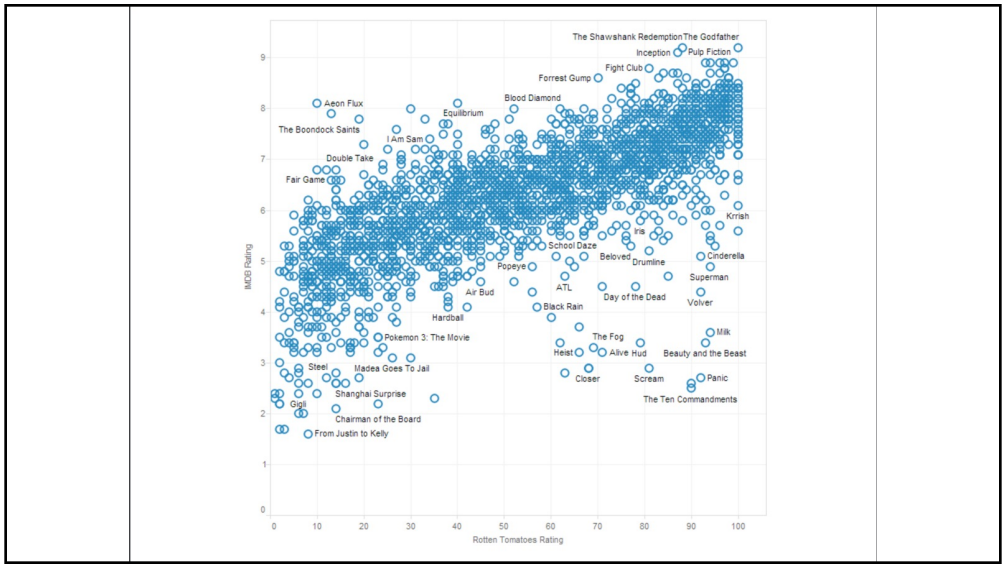
ROTTEN TOMATOES RATINGS



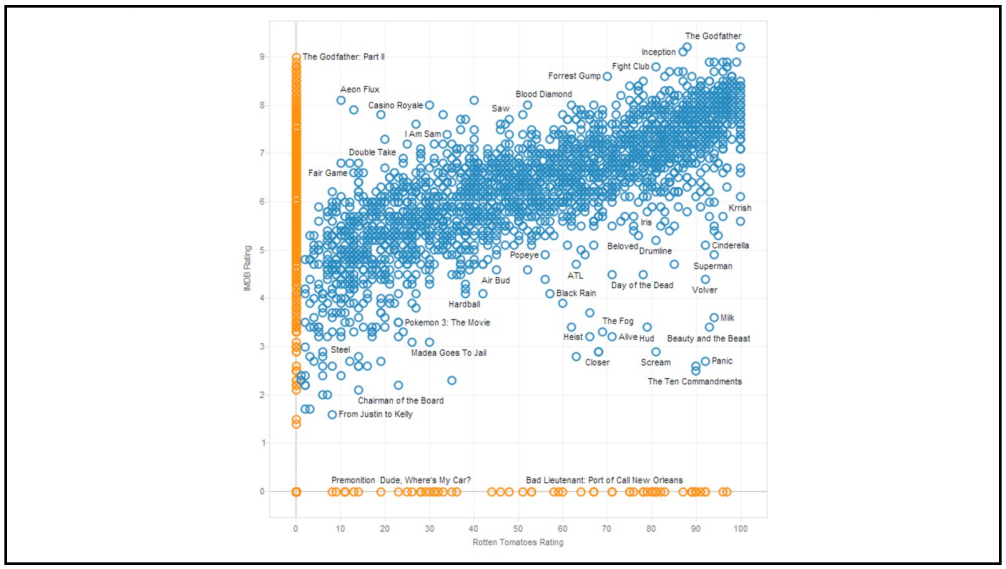
54


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

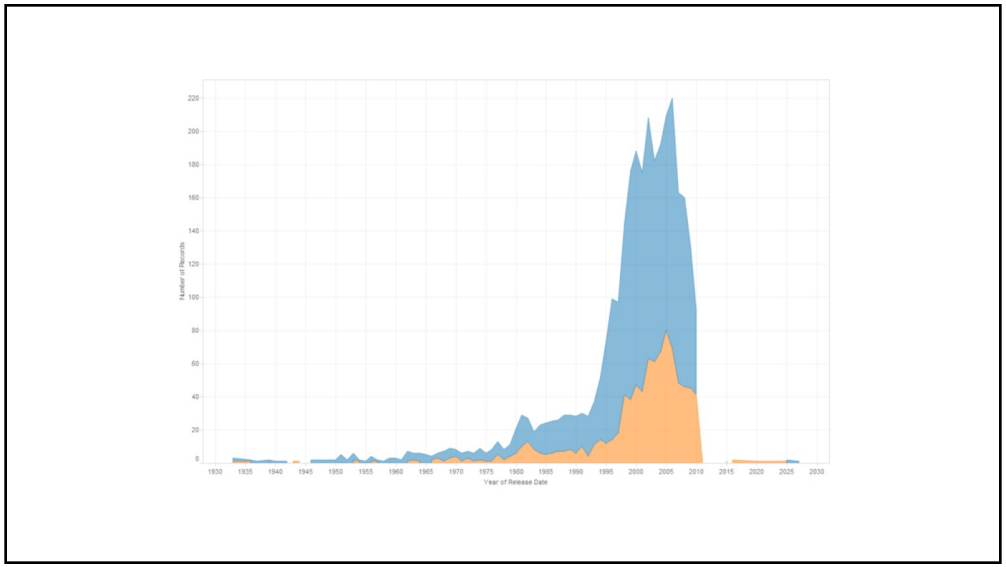
55


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

56


The Godfather
Inception
Fight Club
Forrest Gump
Blood Diamond
Saw
Aeon Flux
Casino Royale
I Am Sam
Double Take
Fair Game
The Godfather: Part II
Krrish
Cinderella
Beloved
Drumline
Superman
Popeye
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Pokemon 3: The Movie
The Fog
Milk
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

57


Number of Records
Year of Release Date

58

LESSON: EXERCISE SKEPTICISM

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

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

Avoid premature fixation!

59

ANNOUNCEMENTS

60

ASSIGNMENT 2: EXP. DATA ANALYSIS

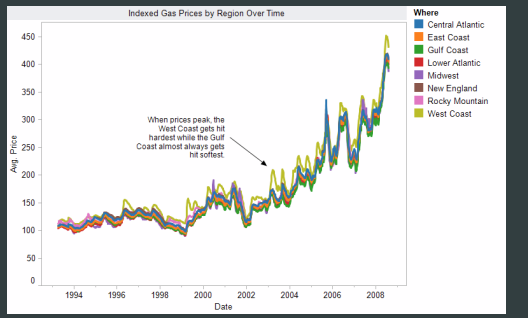
Due 10/13 10:30am

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

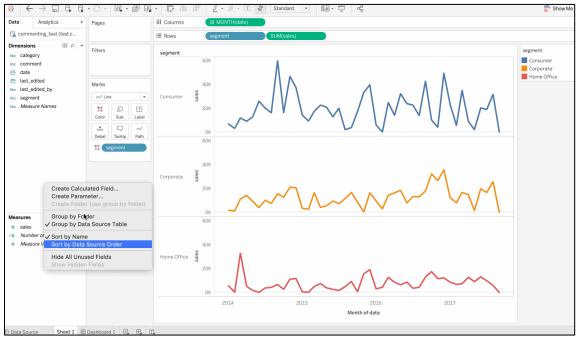
Create visualizations
- See different views of data
- Refine questions

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



61

Tableau Tutorial Session



This Fri 10/3 10:30-11:30 We will walk through the basics of using Tableau. Should be helpful for A2

62

ANALYSIS EXAMPLE: ANTIBIOTIC EFFECTIVENESS

63

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

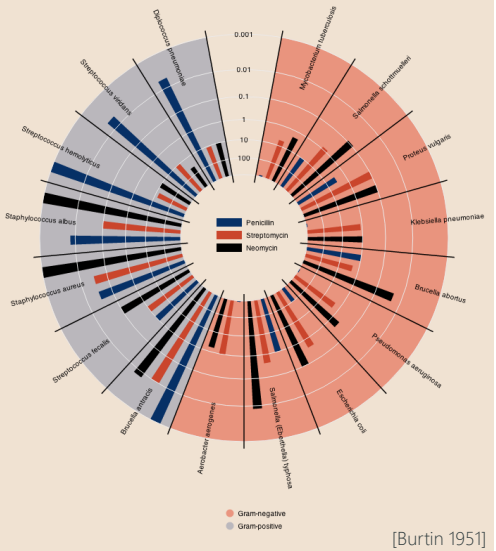
64

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

65

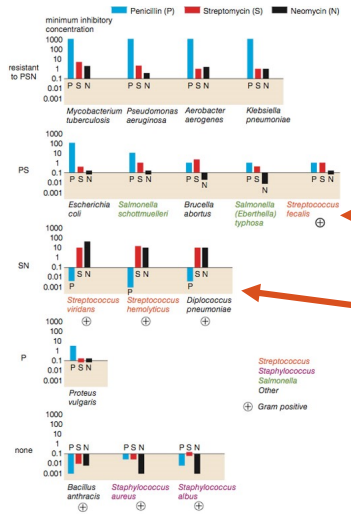


[Burtin 1951]

HOW DO DRUGS COMPARE?

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

66



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

67

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)

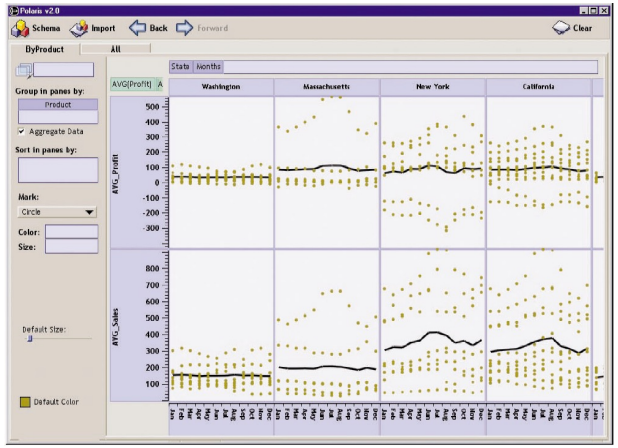
"Show data variation, not design variation" -Tufte

69

TABLEAU/POLARIS

104

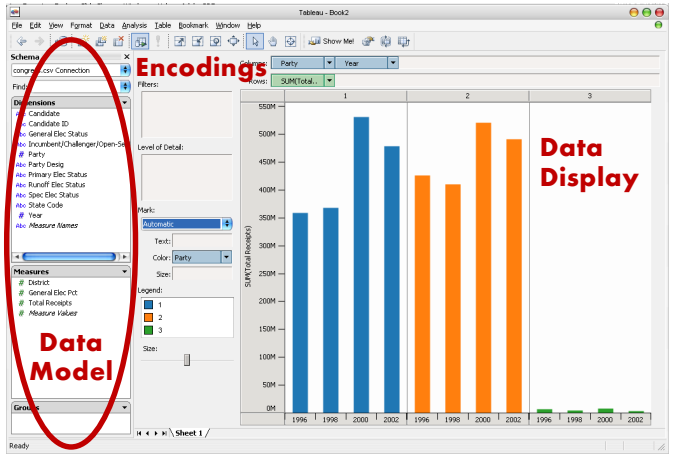
POLARIS [Stolte 2002]



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

105

TABLEAU



106

TABLEAU DEMO

Dataset:

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

107

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

108

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?

109

TABLEAU DEMO

110

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)

111

SUMMARY

Exploratory analysis combines data analysis and graphical methods

Use questions to uncover more questions

Interaction is essential for exploring large multidimensional datasets

129