Data Mining and Analysis

Data Visualization

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Ask question: What data needs to be recorded? or collected?

Why? What research question am I going to answer?

What do I want it to look like?

Real World

Humans behaving
Biology
Finance
Internet
Medicine
Sociology
Olympics

Raw Data is Collected / Recorded

email
logs
medical records
surveys
blood drawn
(microarray)
olympic records
NYT web pages

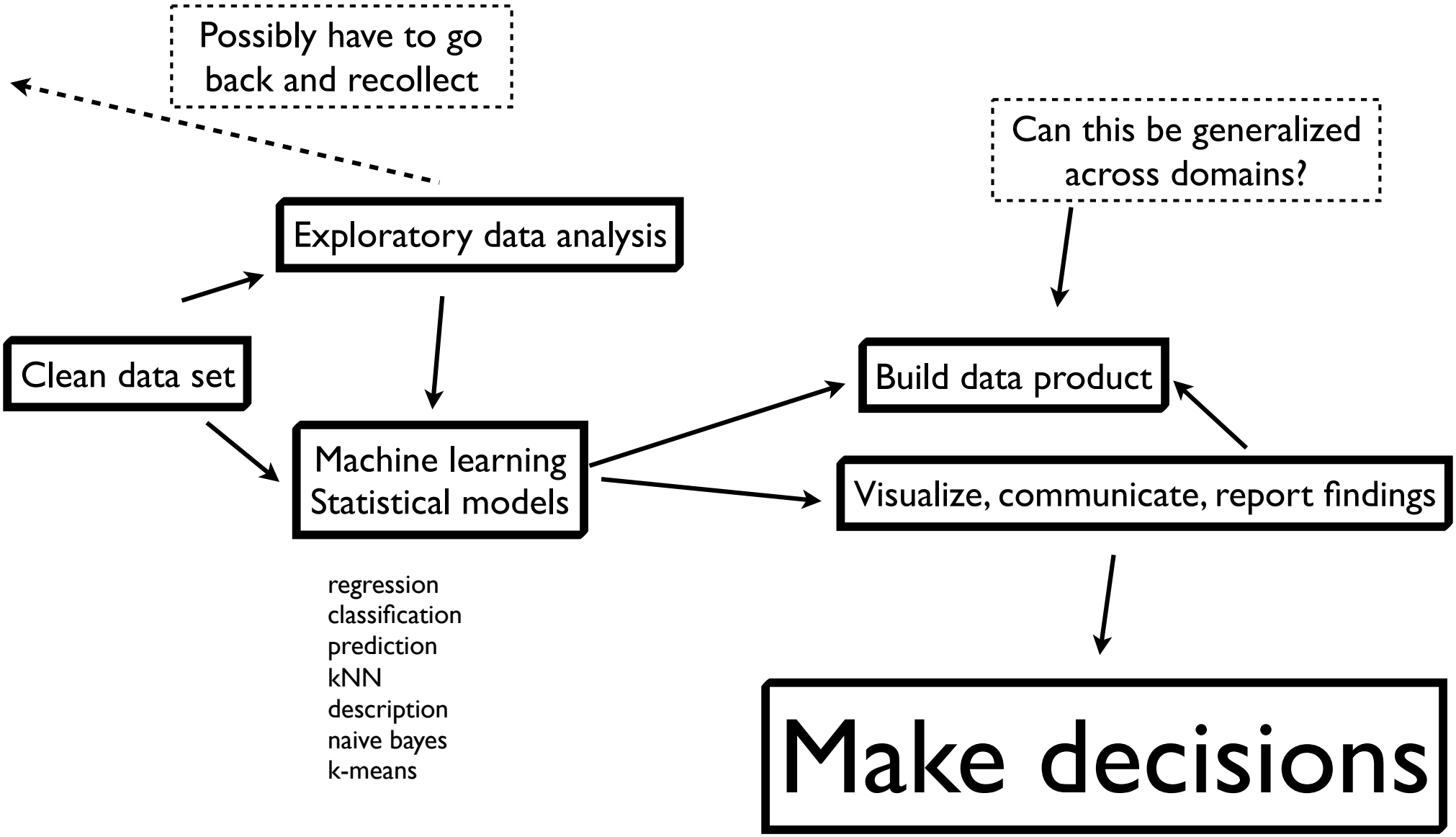
Data is Processed

pipelines
web scraping
cleaning
munging
joining
wrangling

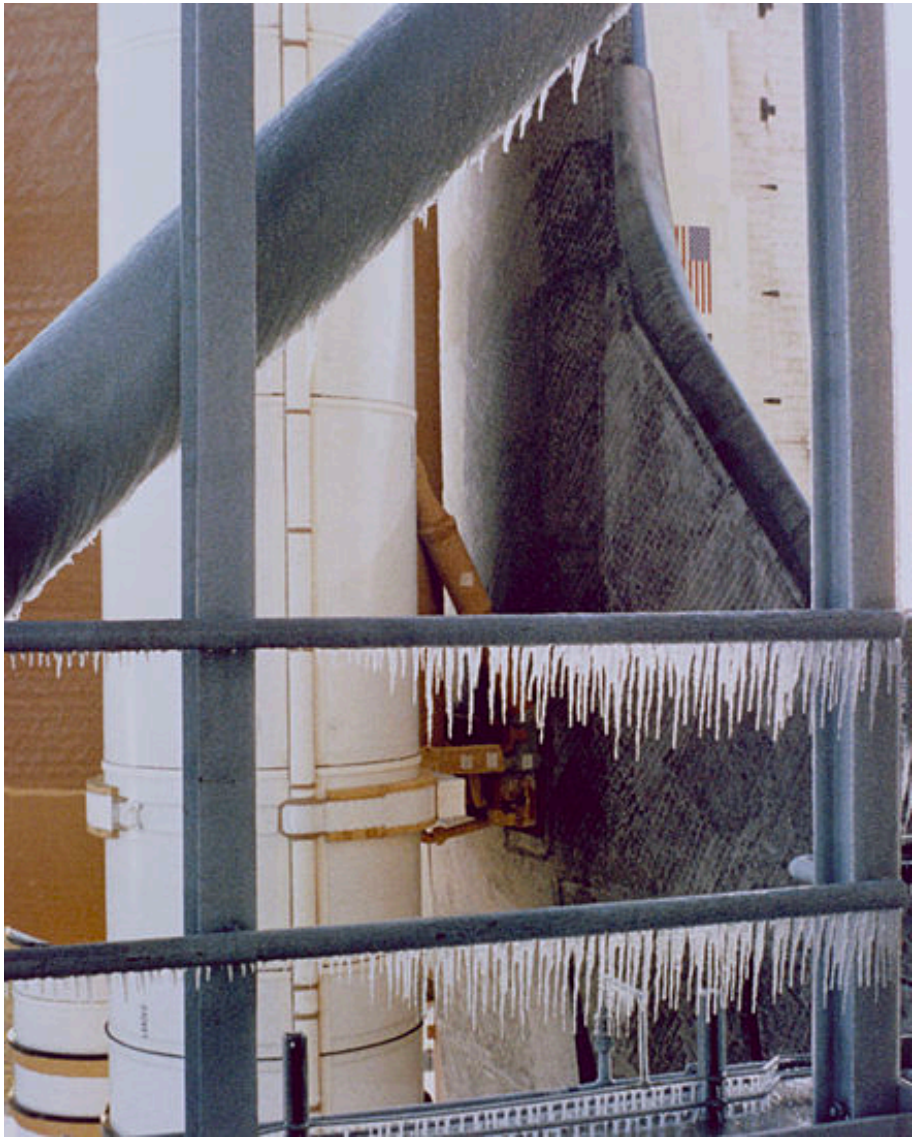
Data set

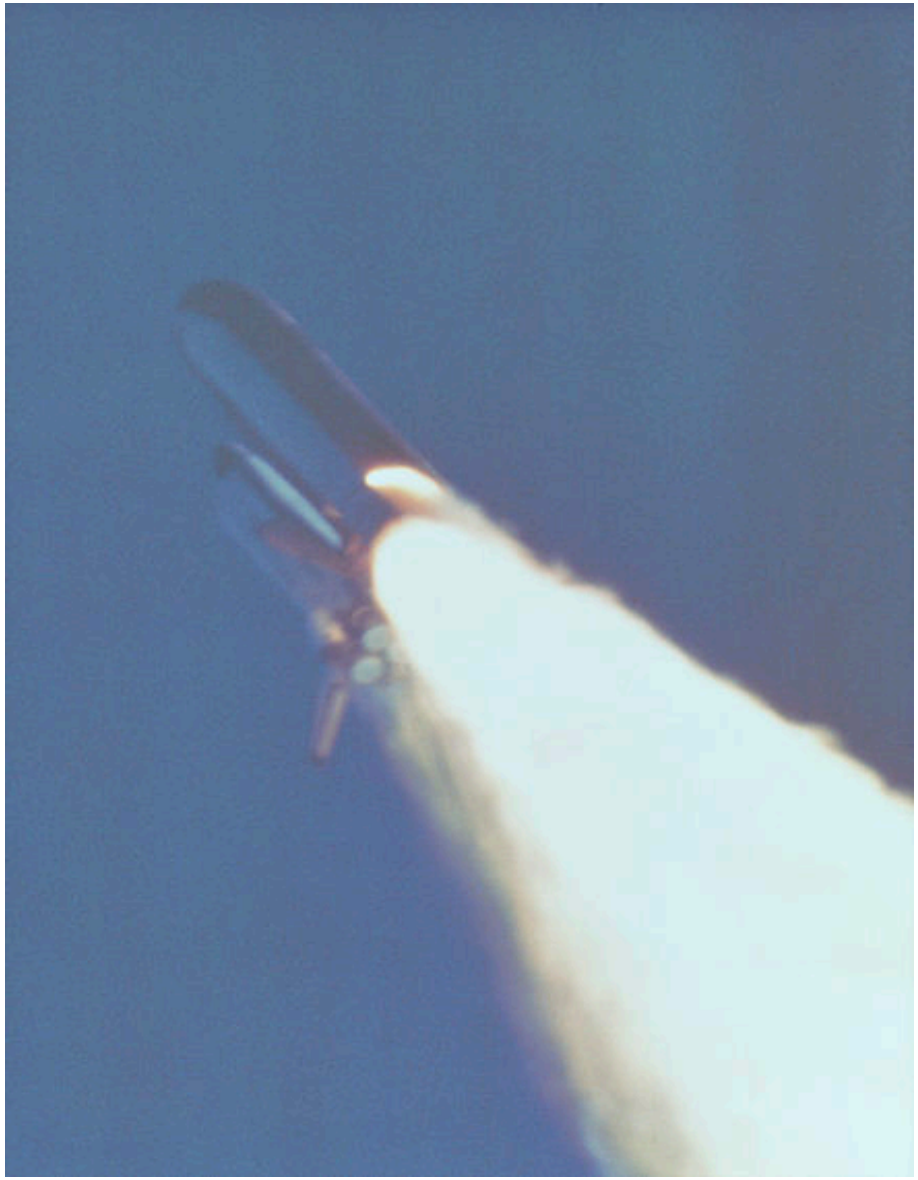
"clean"
table

http://columbiadatascience.com/2012/09/24/reflections-after-jakes-lecture/


Possibly have to go back and recollect
Can this be generalized across domains?
Exploratory data analysis
Clean data set
Build data product
Machine learning
Statistical models
Visualize, communicate, report findings
regression
classification
prediction
kNN
description
naive bayes
k-means
Make decisions

Challenger Disaster (1986)





HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36° --66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT

BLOW BY HISTORY

SRM-15 WORST BLOW-BY

- 2 CASE JOINTS (80°), (110°) ARC
- MUCH WORSE VISUALLY THAN SRM-22

SRM 22 BLOW-BY

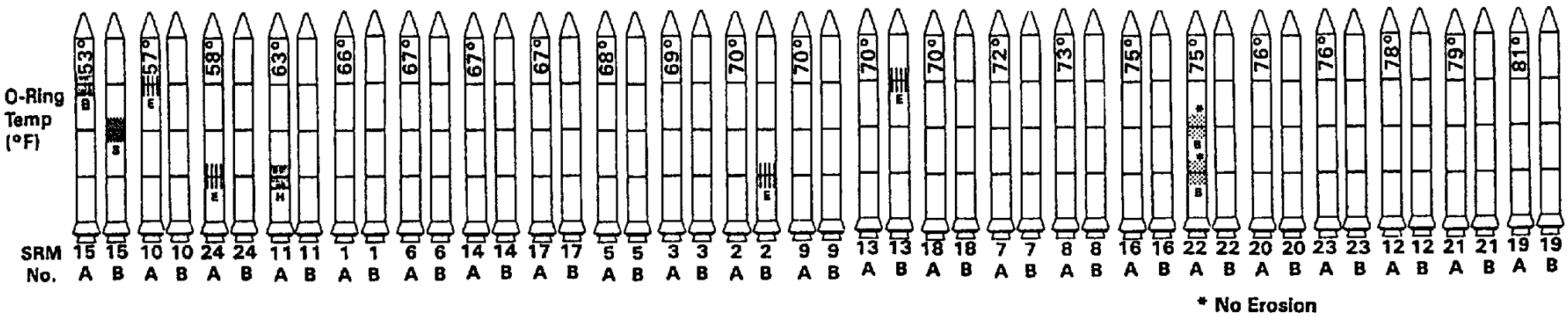
- 2 CASE JOINTS (30-40°)

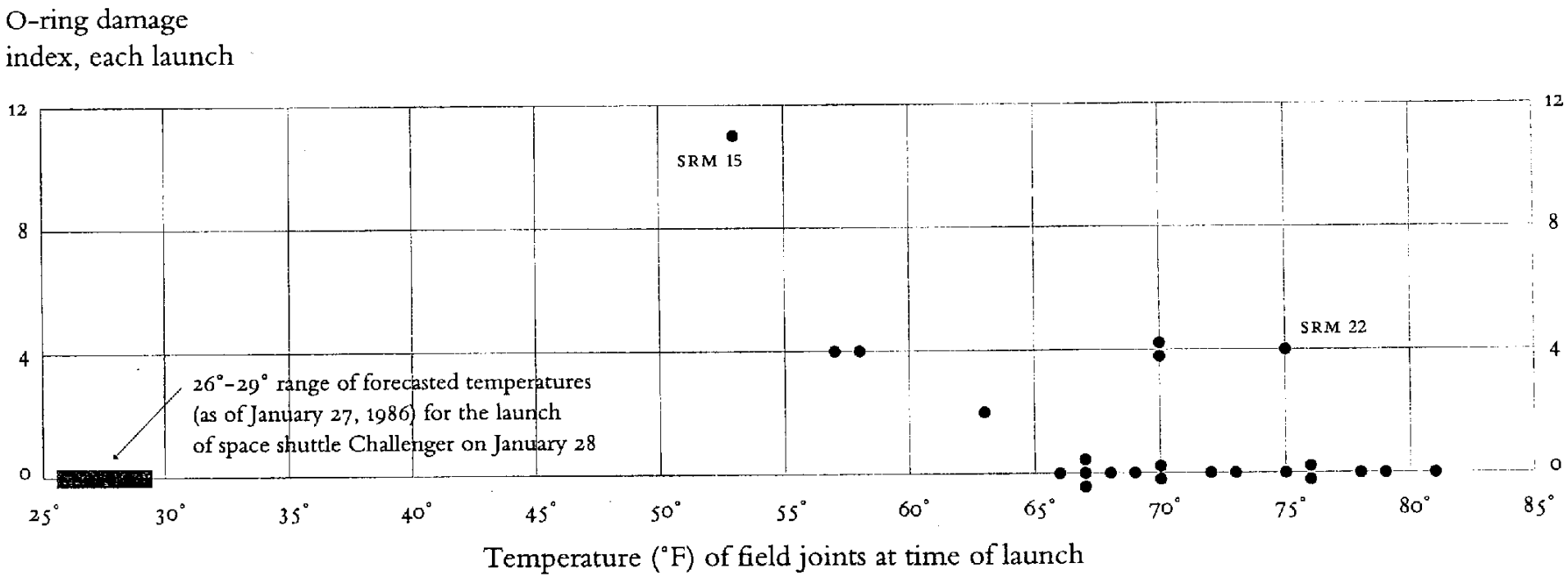
SRM-13A, 15, 16A, 18, 23A 24A

- NOZZLE BLOW-BY

HISTORY OF O-RING TEMPERATURES (DEGREES - F)

MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH


O-Ring Temp (°F)
53° 57° 58° 63° 66° 67° 67° 67° 68° 69° 70° 70° 70° 70° 72° 73° 75° 75° 76° 76° 78° 79° 81°
SRM No.
15 A 15 B 10 A 10 B 24 A 24 B 11 A 11 B 1 A 1 B 6 A 6 B 14 A 14 B 17 A 17 B 5 A 5 B 3 A 3 B 2 A 2 B 9 A 9 B 13 A 13 B 18 A 18 B 7 A 7 B 8 A 8 B 16 A 16 B 22 A 22 B 20 A 20 B 23 A 23 B 12 A 12 B 21 A 21 B 19 A 19 B
* No Erosion


O-ring damage
index, each launch
12
8
4
0
SRM 15
SRM 22
26°–29° range of forecasted temperatures
(as of January 27, 1986) for the launch
of space shuttle Challenger on January 28
25°
30°
35°
40°
45°
50°
55°
60°
65°
70°
75°
80°
85°
Temperature (°F) of field joints at time of launch

Stepping back …

Why do we create visualizations?

Answer questions (or discover them)

Make decisions

See data in context

Expand memory

Support graphical calculation

Find patterns

Present argument or tell a story

Inspire

Jeff Heer

The Value of Visualization

Record information

- Blueprints, photographs, seismographs, ...

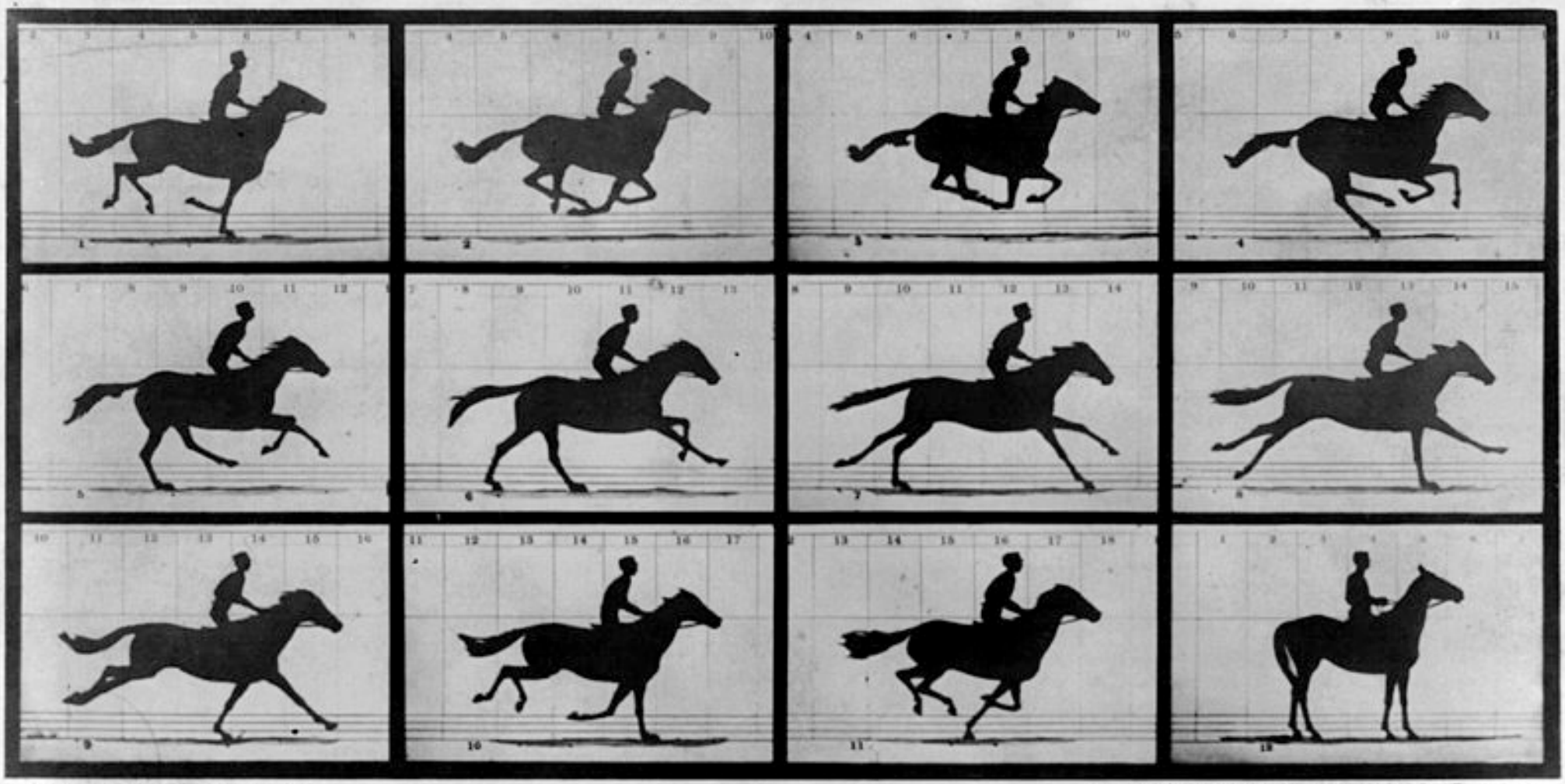
Analyze data to support reasoning

- Develop and assess hypotheses
- Discover errors in data
- Expand memory
- Find patterns

Communicate information to others

- Share and persuade
- Collaborate and revise

Record Information



Copyright, 1878, by MUYBRIDGE.

MORSE'S Gallery, 417 Montgomery St., San Francisco.

THE HORSE IN MOTION.

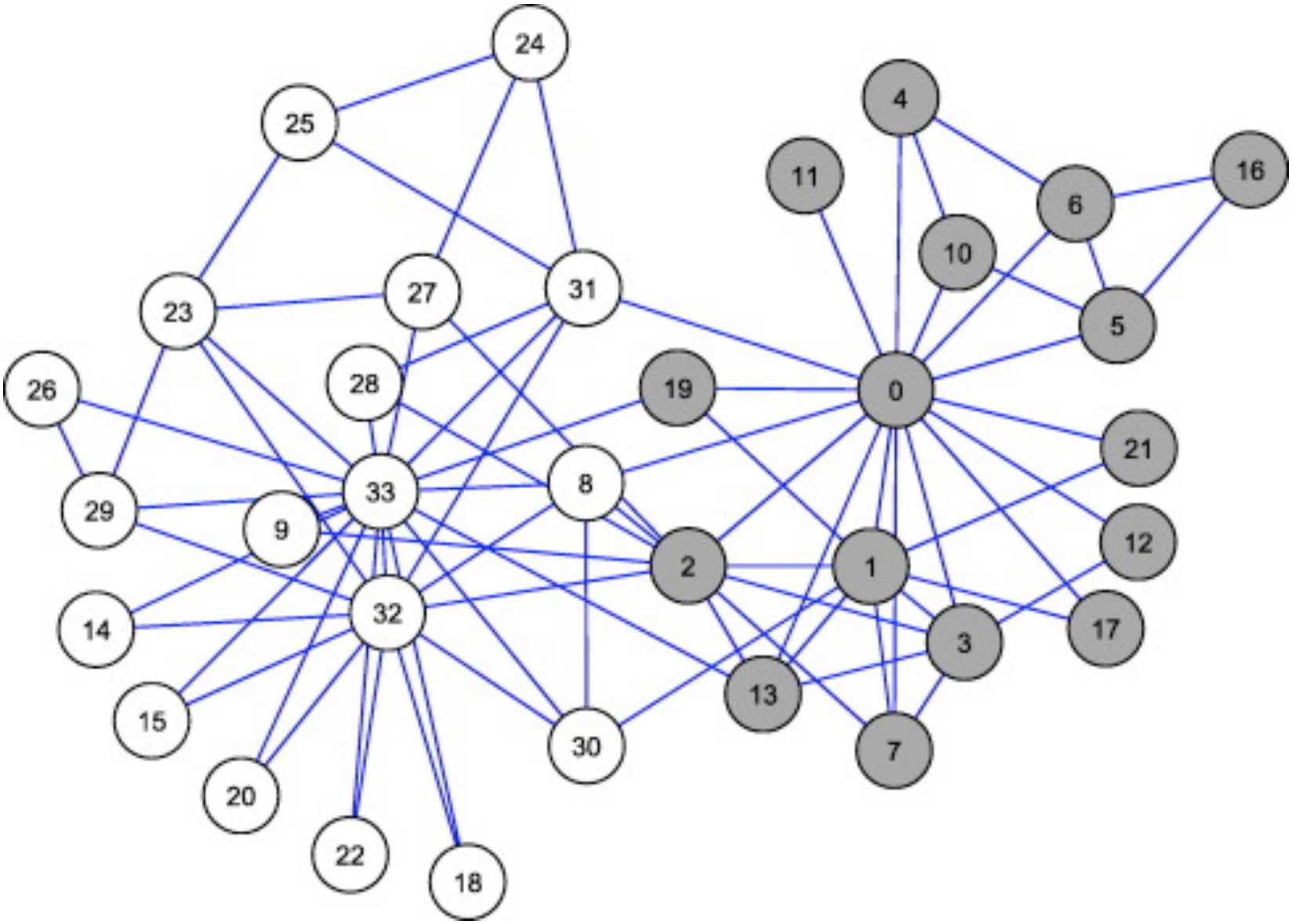
Illustrated by

MUYBRIDGE.

AUTOMATIC ELECTRO-PHOTOGRAPH.

"SALLIE GARDNER," owned by LELAND STANFORD; running at a 1.40 gait over the Palo Alto track, 19th June, 1878.

The negatives of these photographs were made at intervals of twenty-seven inches of distance, and about the twenty-fifth part of a second of time; they illustrate consecutive positions assumed in each twenty-seven inches of progress during a single stride of the mare. The vertical lines were twenty-seven inches apart; the horizontal lines represent elevations of four inches each. The exposure of each negative was less than the two-thousandth part of a second.

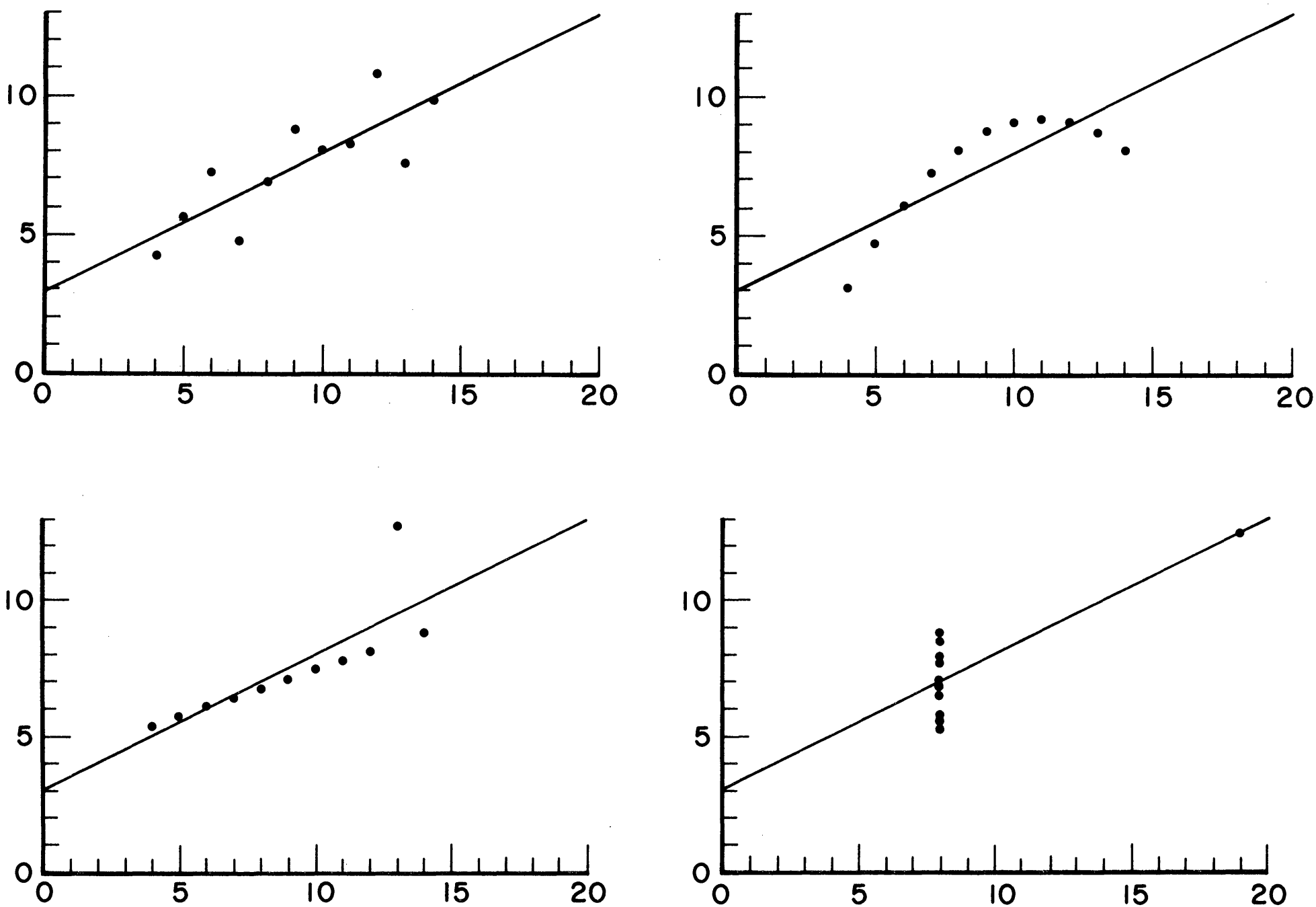

24
25
4
11
16
6
10
23
27
31
5
26
28
19
0
21
29
33
8
9
12
2
1
14
32
17
3
13
15
30
7
20
22
18

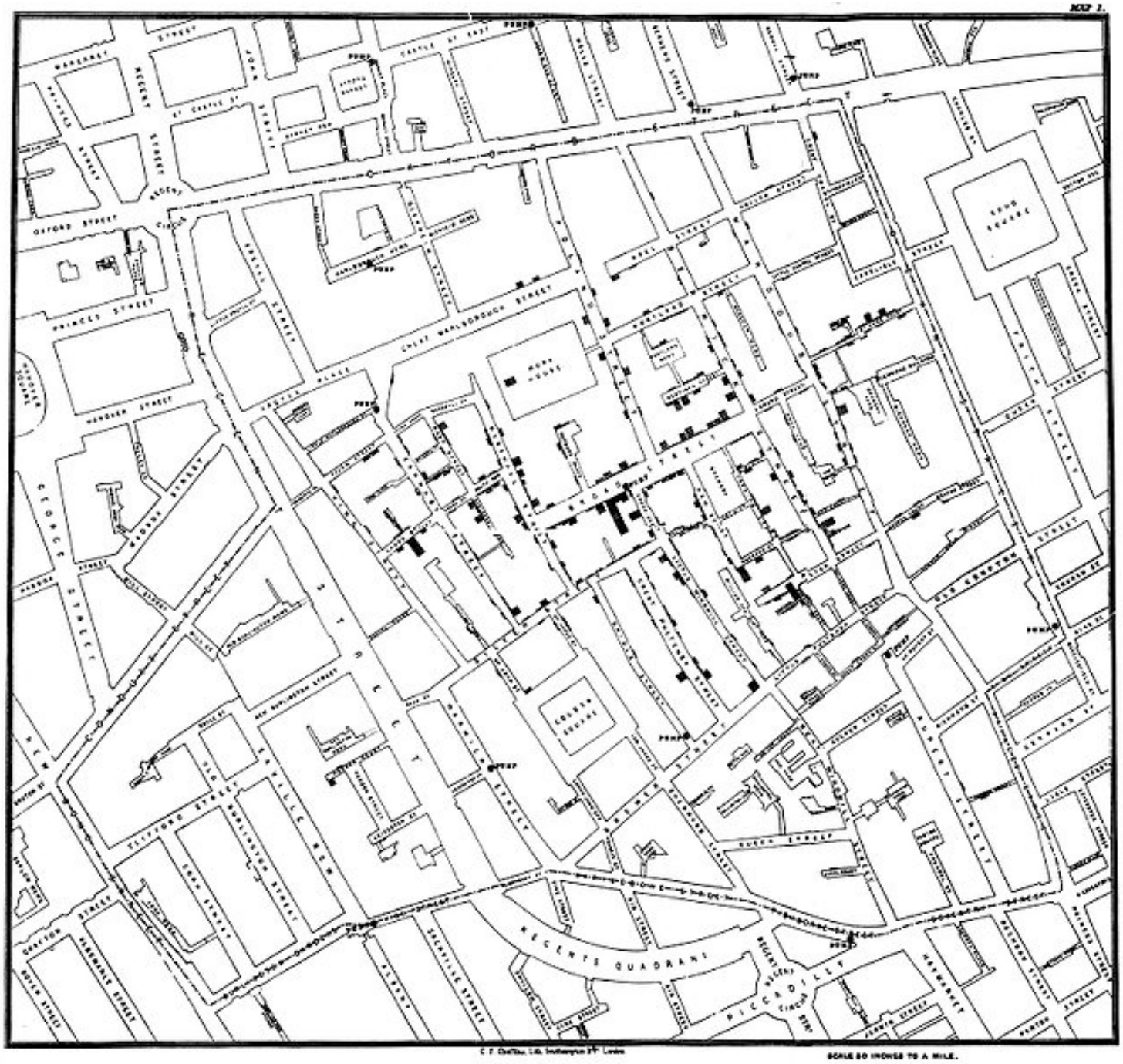
Analyze

Data set	1-3	1	2	3	4	4
Variable	x	y	y	y	x	y
Obs. no. 1 :	10.0	8.04	9.14	7.46 :	8.0	6.58
2 :	8.0	6.95	8.14	6.77 :	8.0	5.76
3 :	13.0	7.58	8.74	12.74 :	8.0	7.71
4 :	9.0	8.81	8.77	7.11 :	8.0	8.84
5 :	11.0	8.33	9.26	7.81 :	8.0	8.47
6 :	14.0	9.96	8.10	8.84 :	8.0	7.04
7 :	6.0	7.24	6.13	6.08 :	8.0	5.25
8 :	4.0	4.26	3.10	5.39 :	19.0	12.50
9 :	12.0	10.84	9.13	8.15 :	8.0	5.56
10 :	7.0	4.82	7.26	6.42 :	8.0	7.91
11 :	5.0	5.68	4.74	5.73 :	8.0	6.89

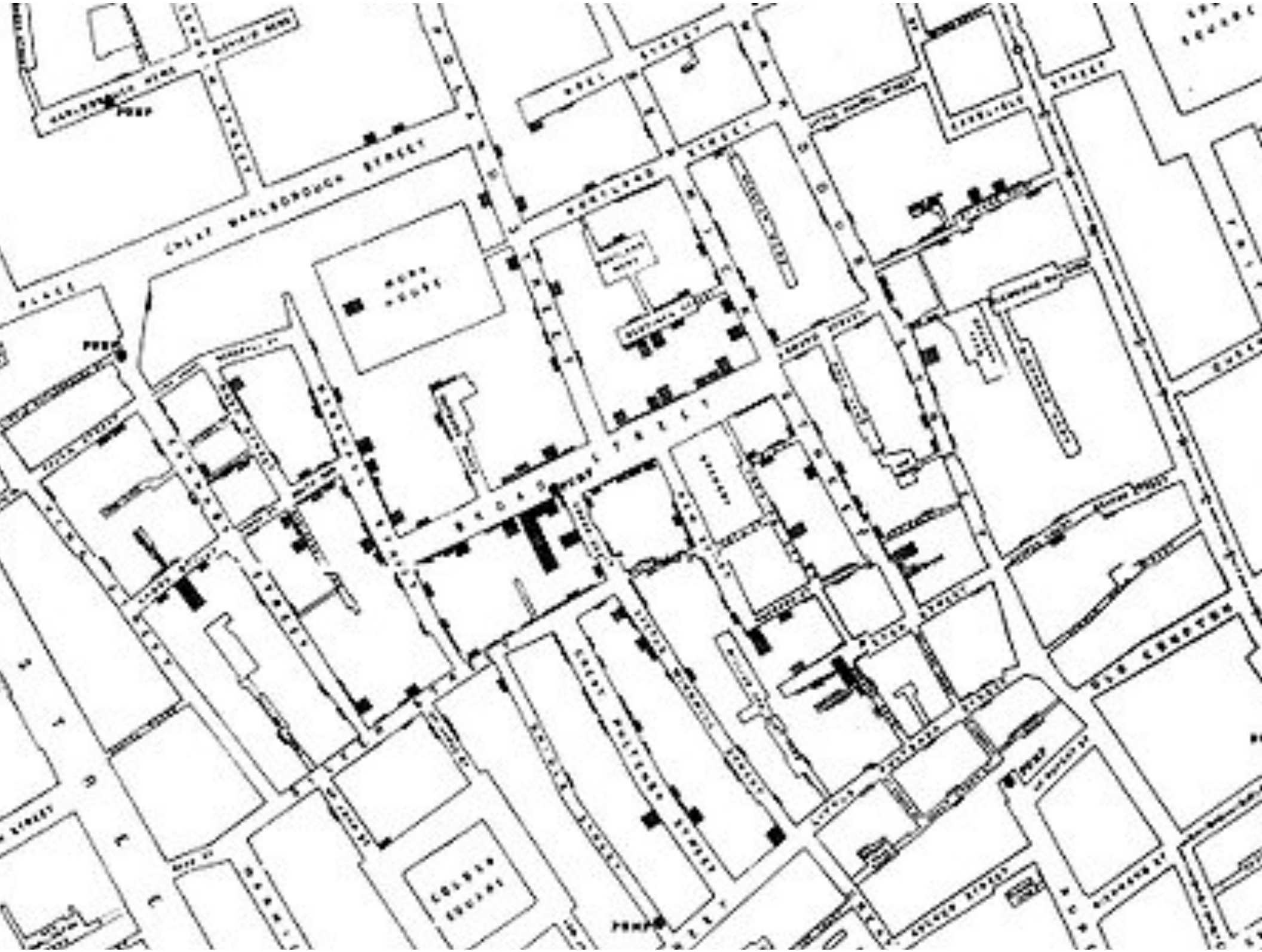
Anscombe “Graphs in Statistical Analysis” 1973

Number of observations (n) = 11
Mean of the x's ($\bar{x}$) = 9.0
Mean of the y's ($\bar{y}$) = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667


0
5
10
0
5
10
15
20
0
5
10
0
5
10
15
20
0
5
10
0
5
10
15
20
0
5
10
0
5
10
15
20

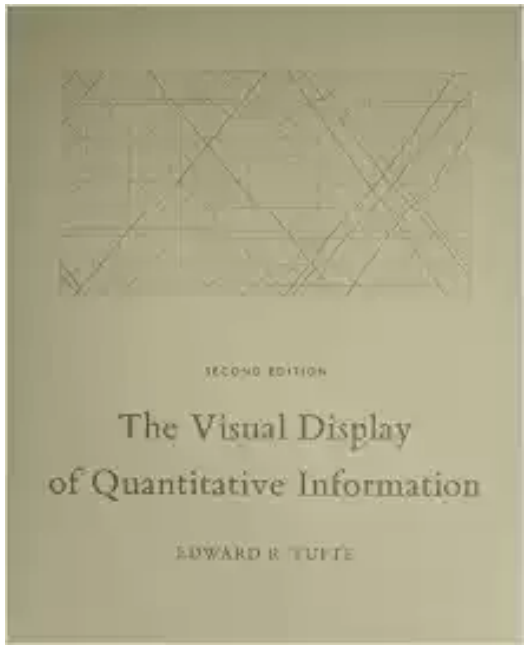


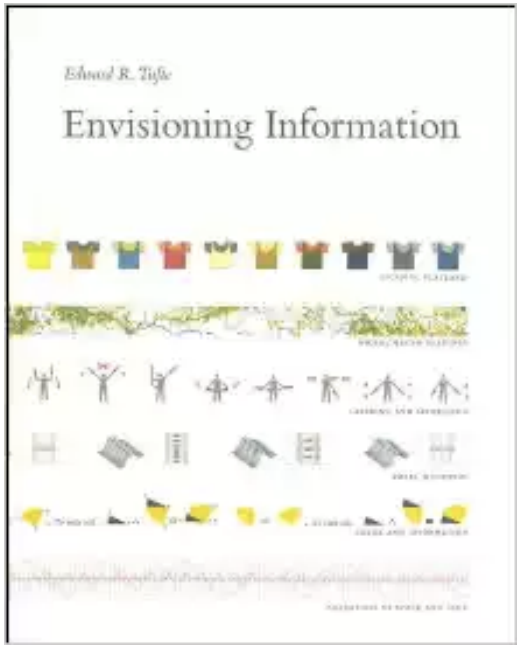
http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

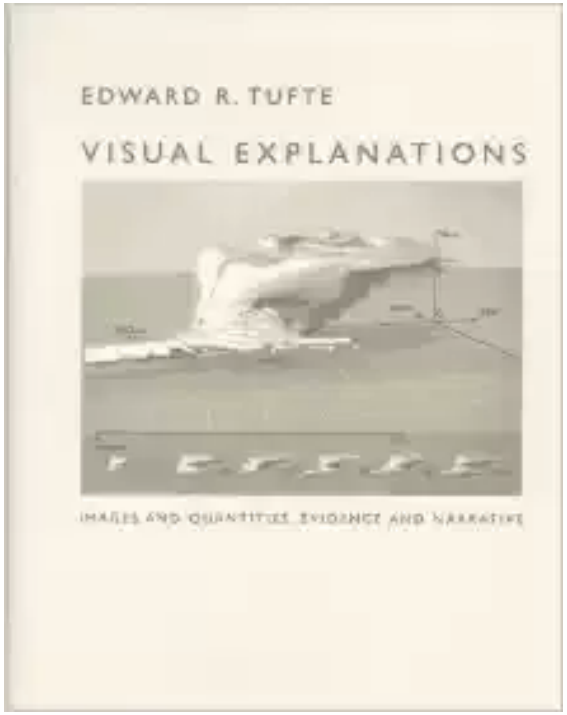

CHEAP MARLBOROUGH STREET
GREAT PULTENEY STREET
GOLDEN SQUARE

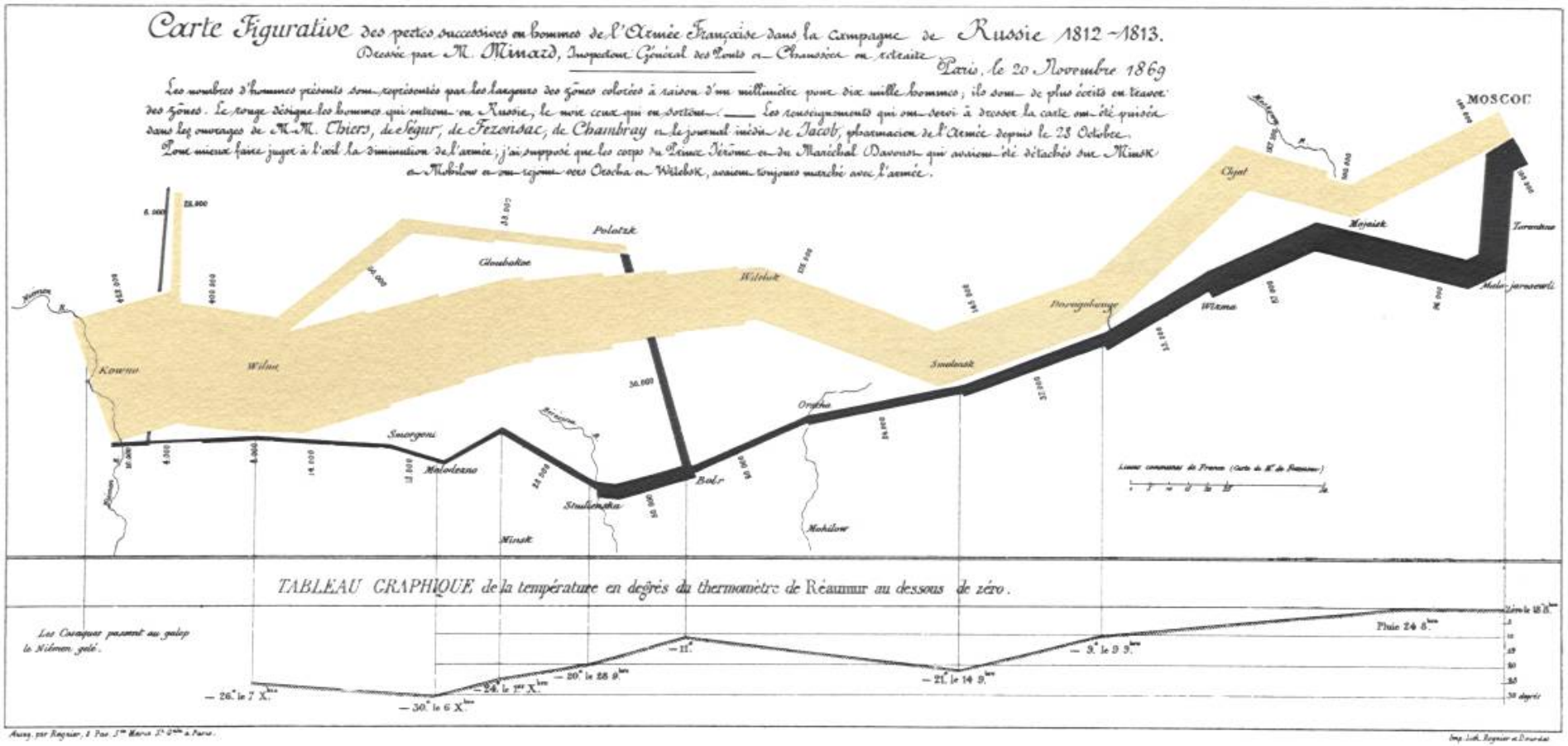
Communicate Information to Others

Tufte

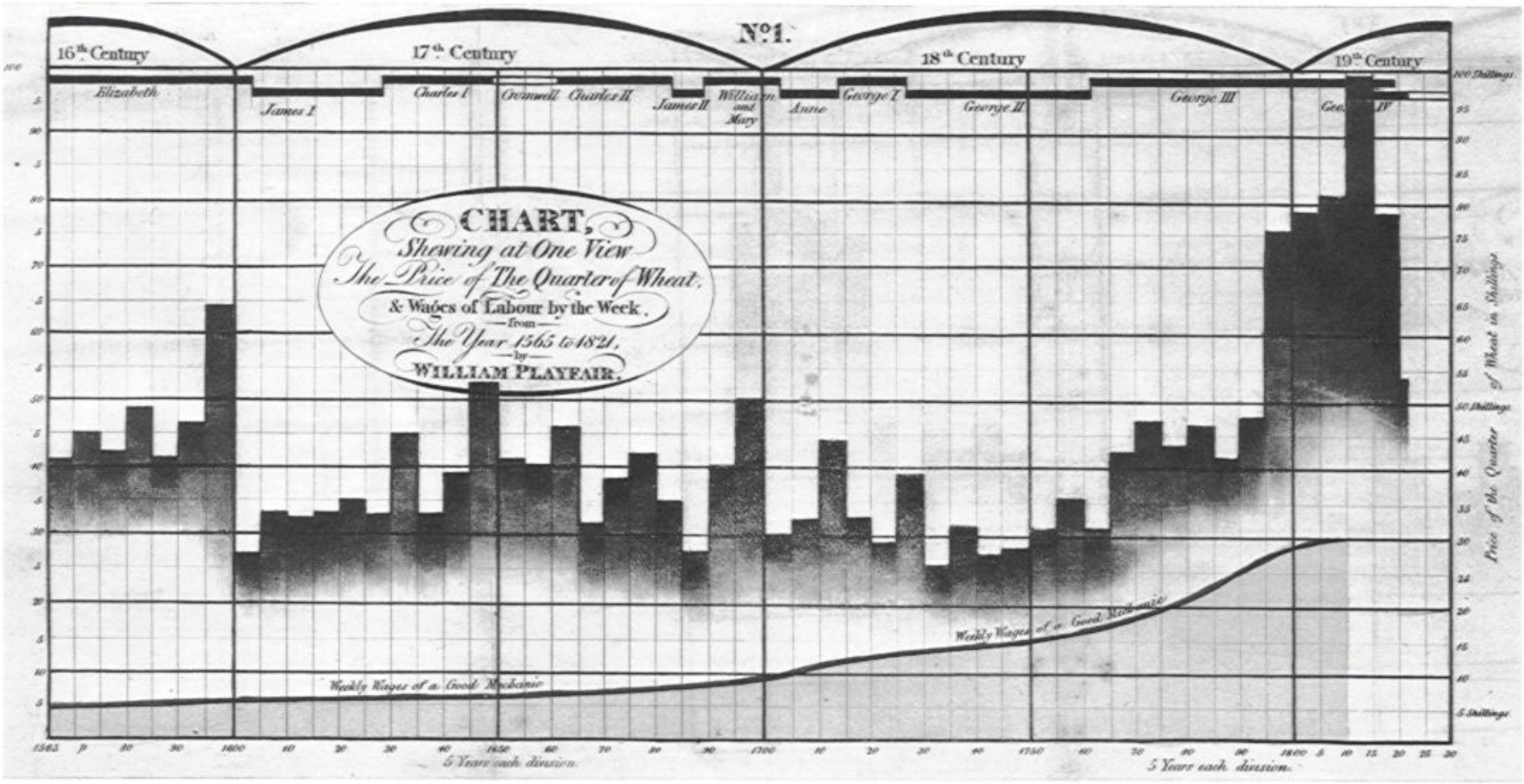

SECOND EDITION
The Visual Display
of Quantitative Information
EDWARD R. TUFTE


Edward R. Tufte
Envisioning Information

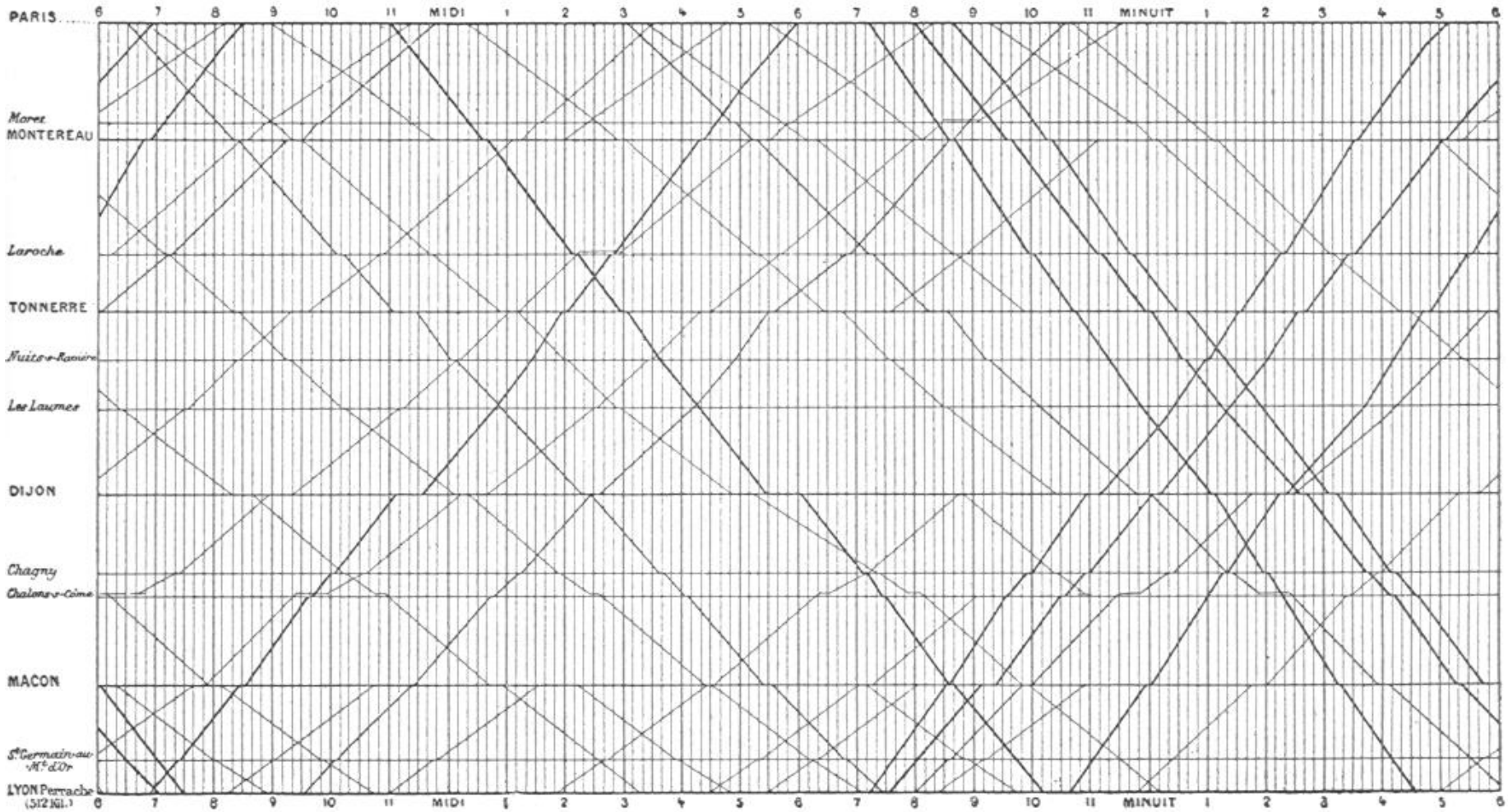

EDWARD R. TUFTE
VISUAL EXPLANATIONS



Napolean's invasion of Russia, as drawn by Charles Joseph Minard (1781-1870)



The price of wheat compared to labour wages, William Playfair (1759-1823)



French train schedule, as drawn by E.J. Marey (1830-1904)

Tufte: Principles of Graphical Excellence

Substance, Statistics, Design

“Complex ideas communicated with clarity, precision, and efficiency”

Tufte's Principles

Graphical Integrity

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.

Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

Show data variation, not design variation.

In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

Graphics must not quote the data out of context.

Theory of Data Graphics

Above all else, show the data.

Maximize the data-ink ratio

Erase non-data-ink

Erase redundant data-ink

Revise and edit

Other comments

Graphical elegance is often found in simplicity of design and complexity of data

Data graphics are paragraphs about data and should be treated as such

Let’s take a look …

Check out Piazza

1) Drugs

https://www.good.is/infographics/transparency-americas-problem-drugs

2) Energy

https://www.good.is/infographics/vampire-energy-2

3) Bankruptcy

https://www.good.is/infographics/transparency-the-largest-bankruptcies-in-history

Design Critique Questions

- What is the purpose?
- Does it serve this purpose well?
  - Does it convey the data honestly?
  - Does it show the appropriate amount of data?
- Does it address an important topic?
- Is it innovative?

Our Favorite Drugs

The war on drugs keeps law enforcement busy—13 percent of all arrests made in 2007 were drug related—but the kinds of battles police are fighting vary widely across the country, from meth labs in California to cocaine dealers in Florida. This is a look at what drugs local law enforcement officials said were posing the greatest dangers to their communities, when asked by the Department of Justice.

WEST

WEST/ISLANDS
ALASKA, AMERICAN SAMOA, CENTRAL CALIFORNIA, GUAM, HAWAII, IDAHO, NEVADA, NORTHERN CALIFORNIA, NORTHERN MARIANA ISLANDS, OREGON, WASHINGTON

NORTH/MIDWEST
COLORADO, IOWA, KANSAS, MISSOURI, MONTANA, NEBRASKA, NORTH DAKOTA, SOUTH DAKOTA, SOUTHERN ILLINOIS, UTAH, WYOMING

MIDWEST
INDIANA, KENTUCKY, MICHIGAN, MINNESOTA, NORTHERN ILLINOIS, OHIO, WISCONSIN

SOUTHWEST
ARIZONA, NEW MEXICO, OKLAHOMA, SOUTHERN CALIFORNIA, TEXAS

SOUTHEAST
ALABAMA, ARKANSAS, GEORGIA, LOUISIANA, MISSISSIPPI, NORTH CAROLINA SOUTH CAROLINA, TENNESSEE

FLORIDA/ISLANDS
FLORIDA, PUERTO RICO, THE U.S. VIRGIN ISLANDS

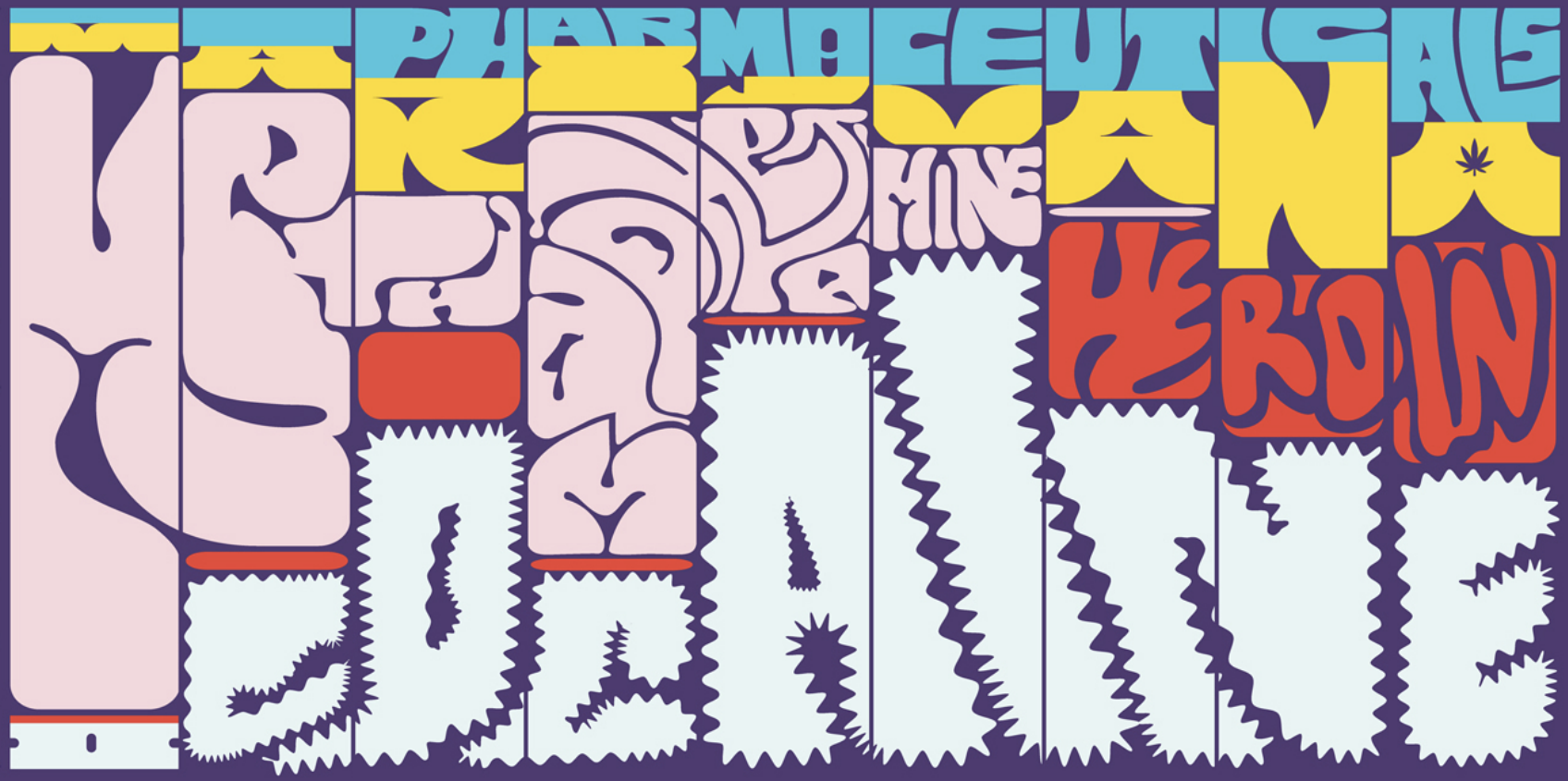
MID-ATLANTIC
DELAWARE, MARYLAND, PENNSYLVANIA, VIRGINIA, WASHINGTON, D.C., WEST VIRGINIA

NORTHEAST
NEW JERSEY, NEW YORK

NEW ENGLAND
CONNECTICUT, MAINE, MASSACHUSETTS, NEW HAMPSHIRE, RHODE ISLAND, VERMONT

EAST

- PHARMACEUTICALS
- MARIJUANA
- METHAMPHETAMINE
- HEROIN
- COCAINE

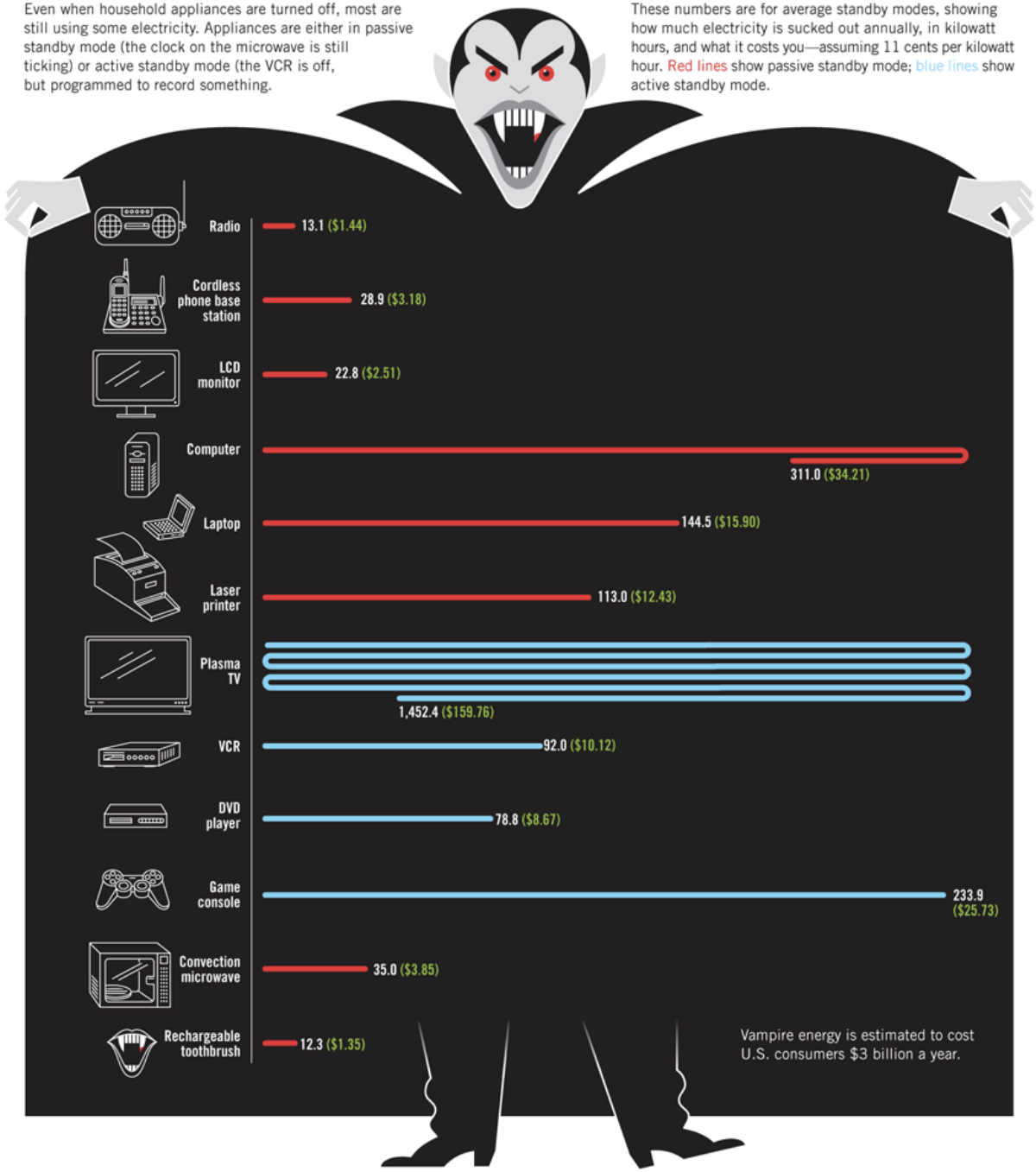


SOURCE: Department of Justice

Vampire Energy

Even when household appliances are turned off, most are still using some electricity. Appliances are either in passive standby mode (the clock on the microwave is still ticking) or active standby mode (the VCR is off, but programmed to record something.

These numbers are for average standby modes, showing how much electricity is sucked out annually, in kilowatt hours, and what it costs you—assuming 11 cents per kilowatt hour. Red lines show passive standby mode; blue lines show active standby mode.



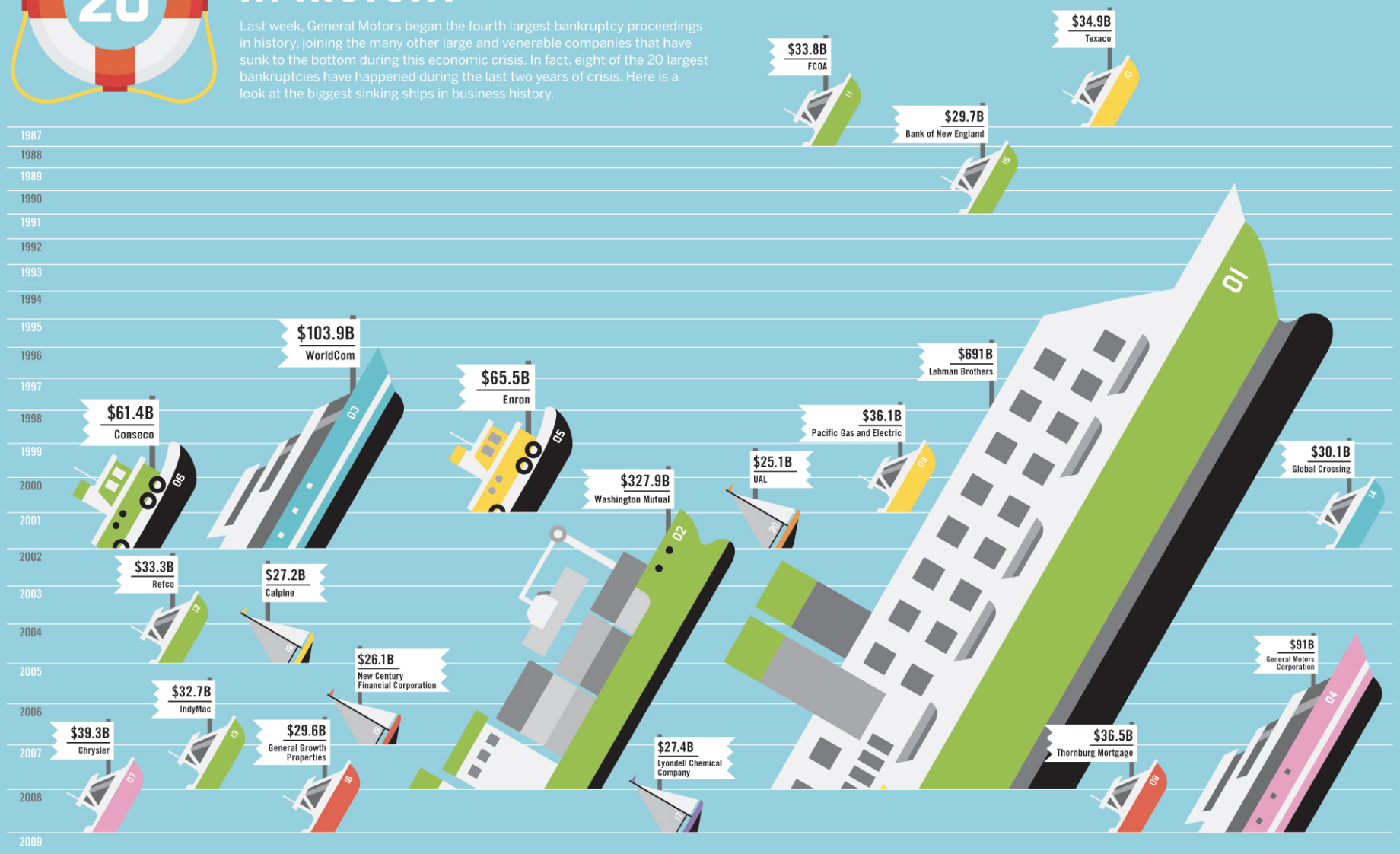
Vampire energy is estimated to cost U.S. consumers $3 billion a year.

SOURCES 2005 Intrusive Residential Standby Service Report;

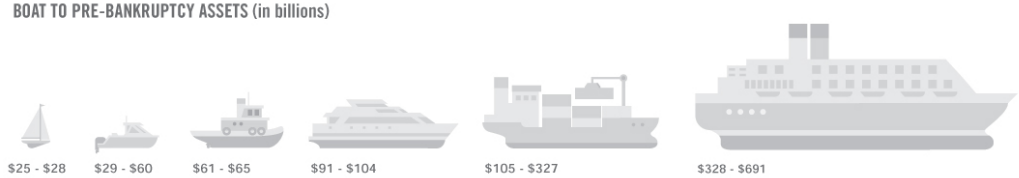
LARGEST BANKRUPTCIES IN HISTORY

20

Last week, General Motors began the fourth largest bankruptcy proceedings in history, joining the many other large and venerable companies that have sunk to the bottom during this economic crisis. In fact, eight of the 20 largest bankruptcies have happened during the last two years of crisis. Here is a look at the biggest sinking ships in business history.



BOAT TO PRE-BANKRUPTCY ASSETS (in billions)



SECTOR

- Finance
- Energy
- Air Travel
- Real Estate
- Automotive
- Chemicals
- Telecommunications

PRE-BANKRUPTCY ASSETS (in billions) / Company

$XXX
Xxxxxxxxx

SOURCE BankruptcyData.com

Class Design Principles: Thoughts?

1. Don’t get too cute; focus on the data (focus on data variation, not design variation)

2. Have a good, explanatory, clear title

3. Data integrity; maintain scale, right axes (no assumptions: be clear with units, etc.)

4. Know your goal! Know your audience.

5. Balance high-level with detail

6. Consider and respect spatial logic

7. Don’t be overconfident

8. Color and design are important