Introduction to visualization

Viet-Trung Tran

Course description

- **What you will learn**
  - Fundamental principles of data visualization
  - Visual perception and cognitive principles
  - Data visualization techniques and best practices
  - Industry-standard visualization tools and libraries
  - Real-world visualization projects and case studies
- **Why it matters**
  - Essential skill in data science and analytics
  - Critical for effective data communication
  - Growing demand in industry and research
  - Foundation for advanced visualization topics

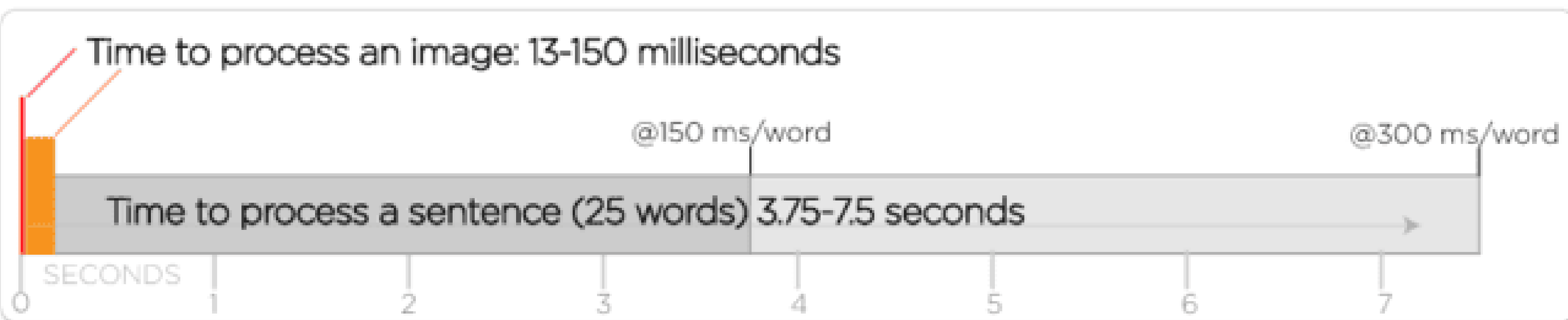
Outline

- What is visualization
- The value of visualization
- Data visualization in the big data era

What is Visualization?

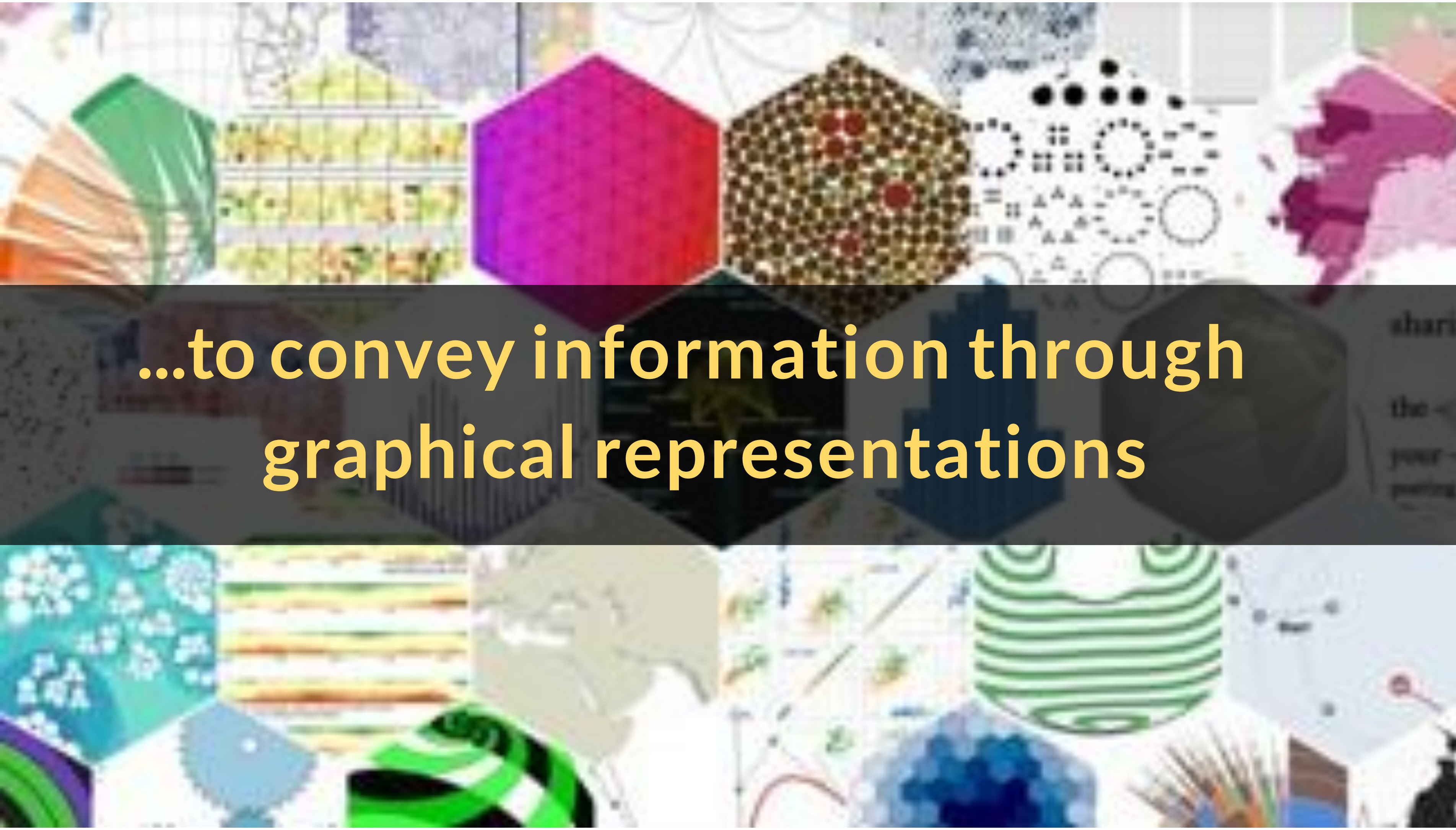
Basic definition

- Visual representation of data and information
- Uses graphical elements to convey meaning
- Helps humans understand complex information quickly
- Leverages our natural visual processing abilities
- **Why Visualization?**
    - 30% of brain cortex dedicated to visual processing
    - Process visual information 6x – 600x times faster than text
    - Can spot patterns, trends and outliers instantly
    - Enhances memory retention and recall



Academic definitions

- "Transformation of the symbolic into the geometric" — McCormick et al. 1987
  - Incorporate visual elements that translate abstract symbols into geometric forms.
- "... finding the artificial memory that best supports our natural means of perception." —Bertin 1967
  - Craft visualizations aligning with how humans naturally perceive and process information.
- "visual representations of data to amplify cognition." — Card, Mackinlay, & Shneiderman 1999
  - Utilize visualizations to enhance cognitive processes and facilitate understanding.



…to convey information through graphical representations

Anscombe's Quartet

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

Summary Statistics

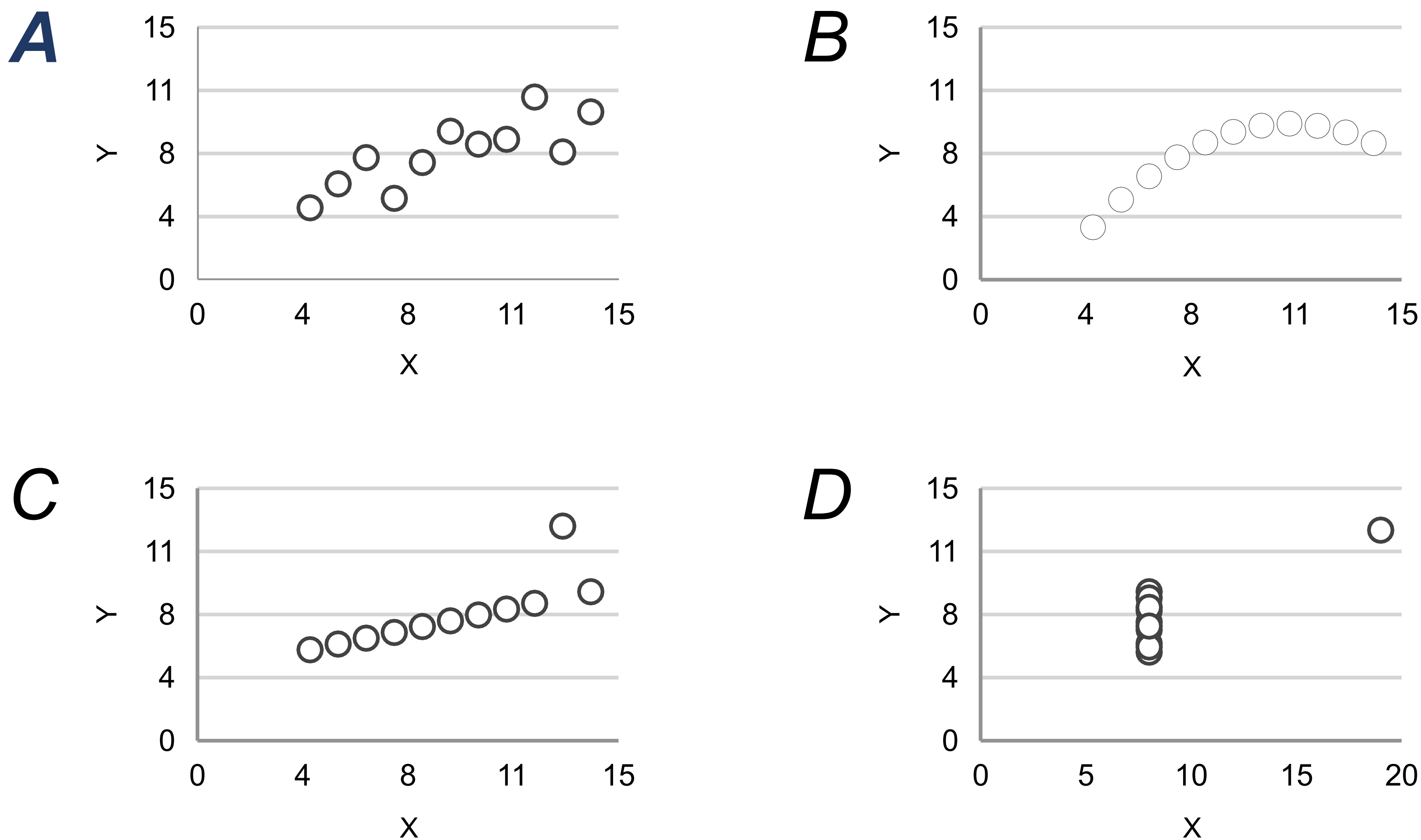
$u_X = 9.0 \quad \sigma_X = 3.317$

$u_Y = 7.5 \quad \sigma_Y = 2.03$

Linear Regression

$Y = 3 + 0.5\,X$

$R^2 = 0.67$

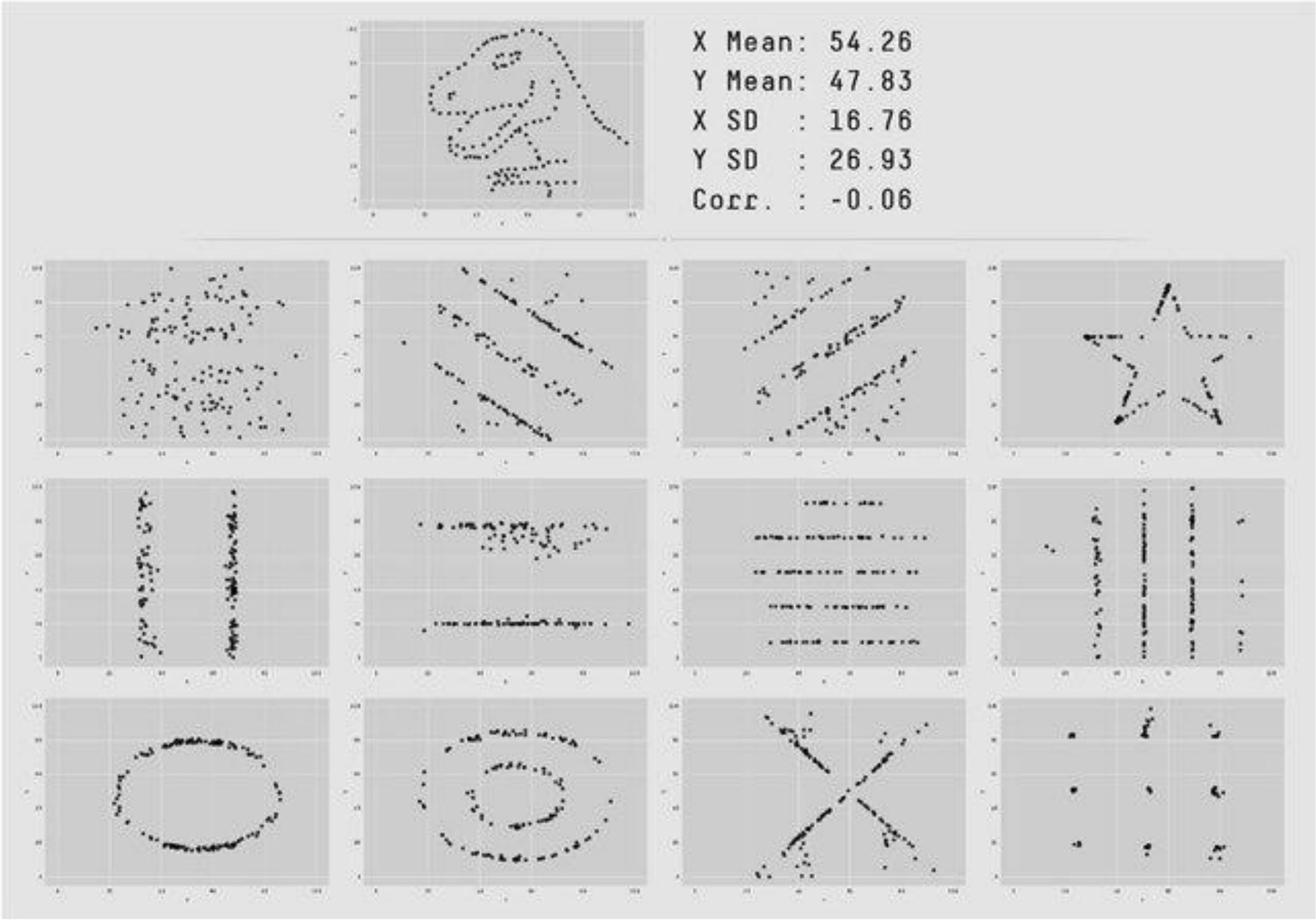

A
15
11
8
4
0
Y
0
4
8
11
15
X
B
15
11
8
4
0
Y
0
4
8
11
15
X
C
15
11
8
4
0
Y
0
4
8
11
15
X
D
15
11
8
4
0
Y
0
5
10
15
20
X

...make both calculations and graphs.

- Both sorts of output should be studied; each will contribute to understanding.

F. J. Anscombe, 1973

All distinct datasets with same statistical properties



Matejka & Fitzmaurice 2017

Data Visualization

- Is an interdisciplinary field that deals with the graphic representation of data.
- Focus
  - Represents structured data sets.
- Purpose
  - Uncover patterns, trends, and insights for decision-making.
- Examples
  - Charts, graphs, dashboards.

Shroeder et al., 2002

Scientific Visualization

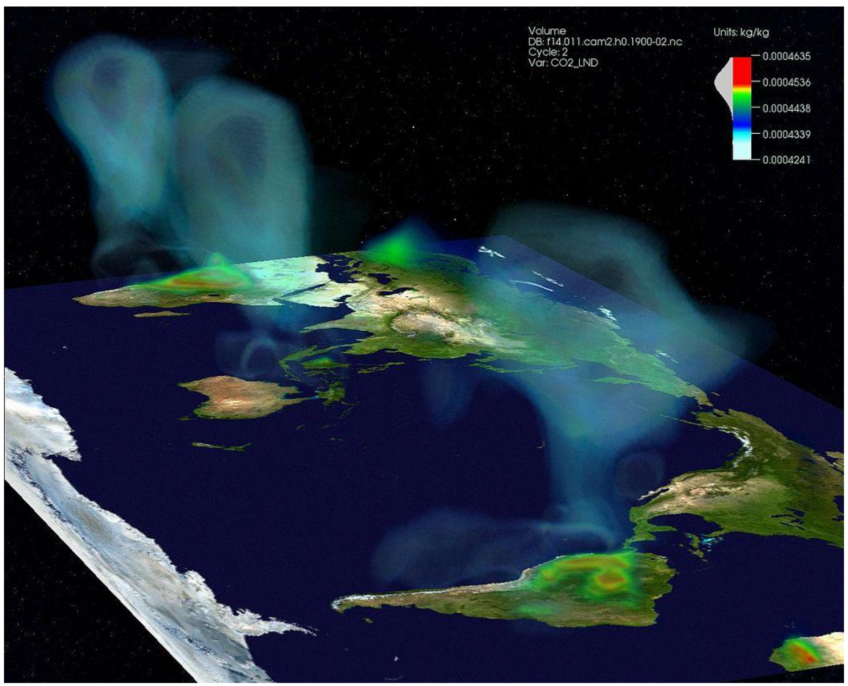
- An interdisciplinary branch of science concerned with the visualization of scientific phenomena.
- Focus
  - Illustrates complex scientific phenomena.
- Purpose
  - Aid in understanding scientific data and simulations.
- Examples
  - Molecular visualizations, simulations, 3D models.

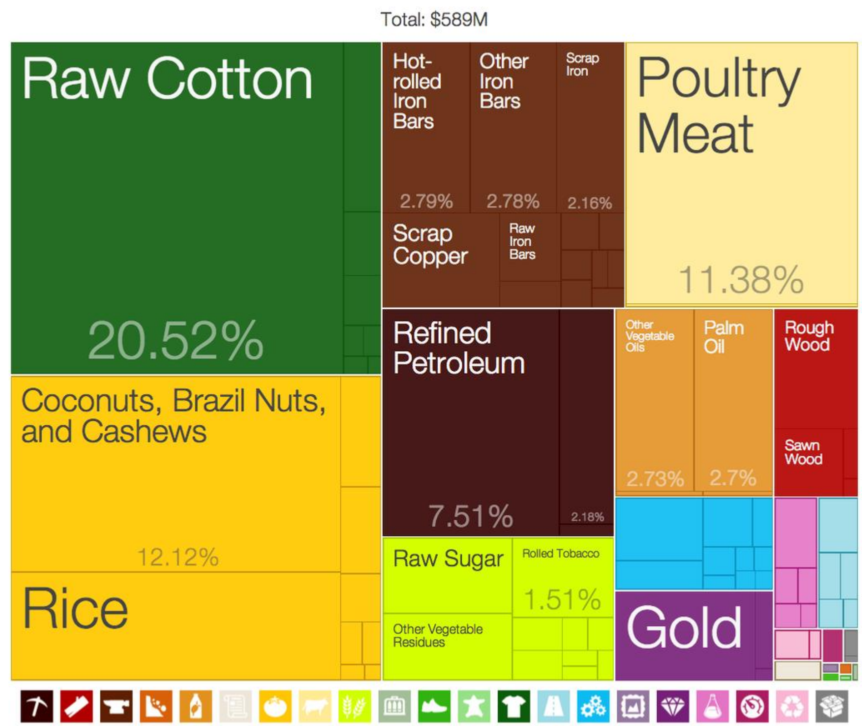
Information Visualization

- Is the study of visual representations of abstract data to reinforce human cognition.
  - Information Visualization techniques are used for searching for interesting phenomena
- Focus
  - Deals with abstract and non-numerical data.
- Purpose
  - Clarify complex information for easy understanding.
- Examples
  - Diagrams, mind maps, interactive visualizations.

Scientific vs Information visualization

- Climate visualization
- Tree Map of Benin Exports (2009) by product category.





The value of visualization

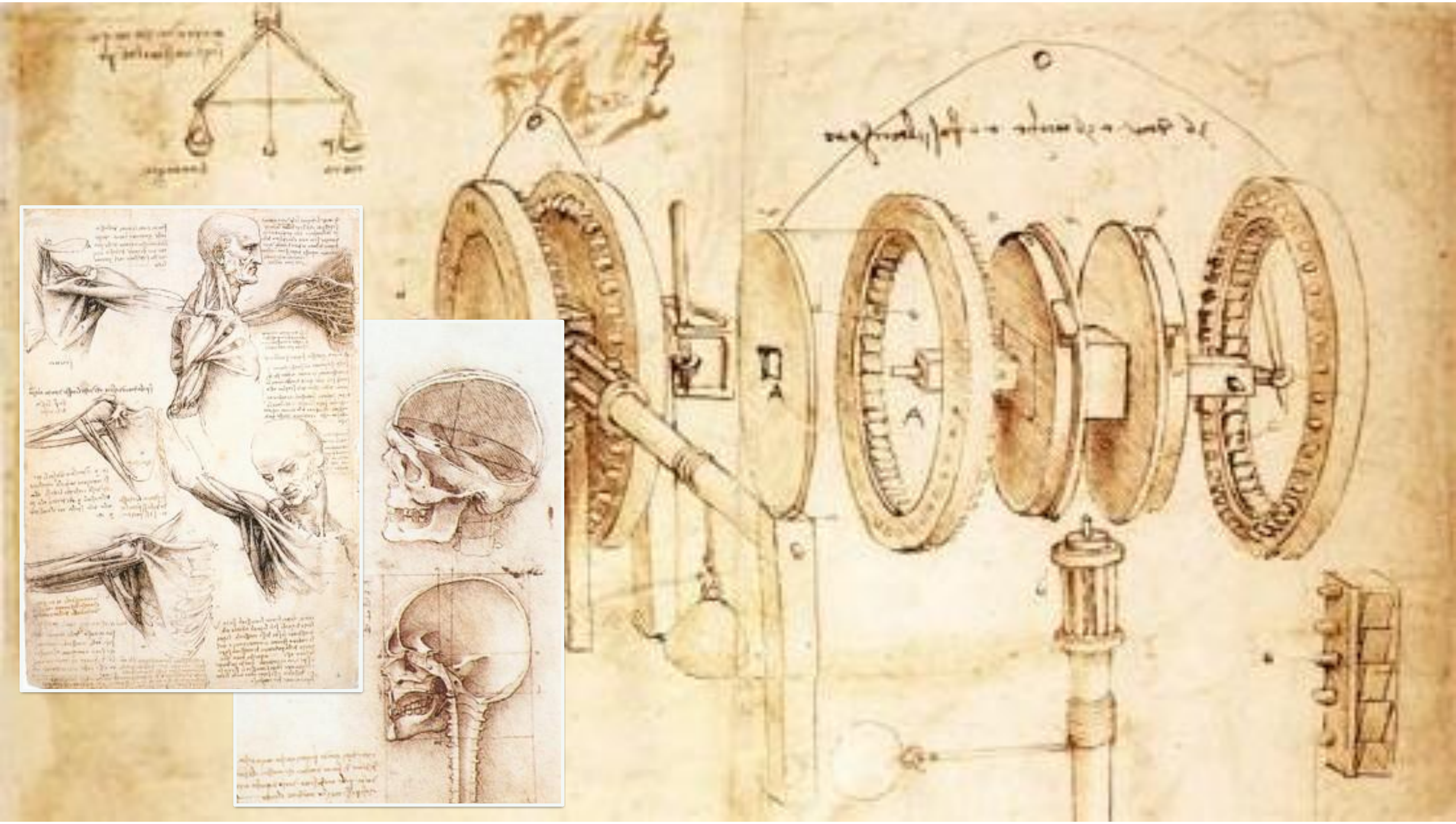
A picture is worth a thousand words, and a well-designed visualization is worth a thousand numbers

Three functions of visualization

- Record information
- Support reasoning
- Convey information to others

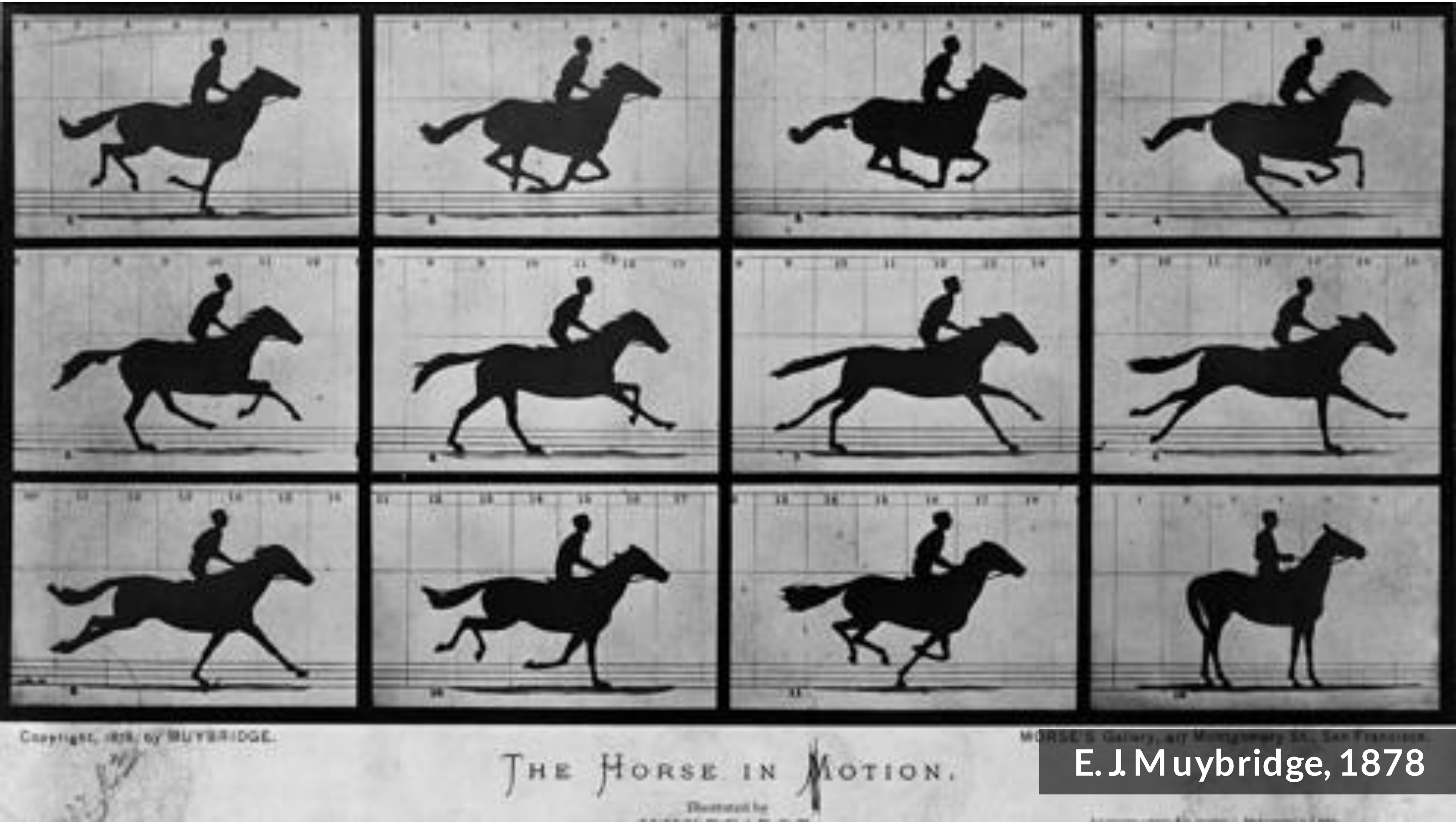
Record information: 6200 BC

Leonardo da Vinci 1485



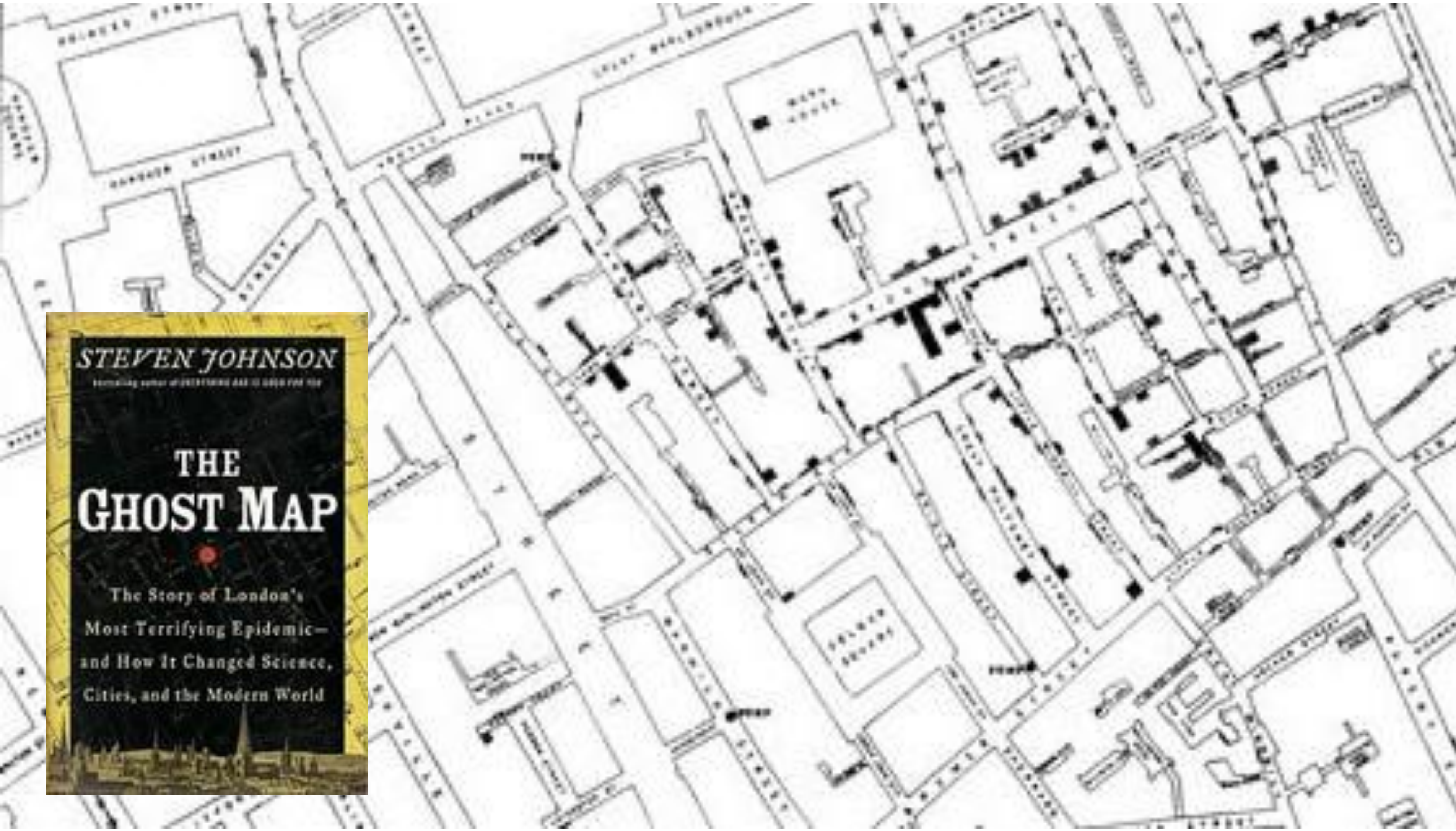
Galileo Galilei's Sketches of the Moon

(November-December 1609)



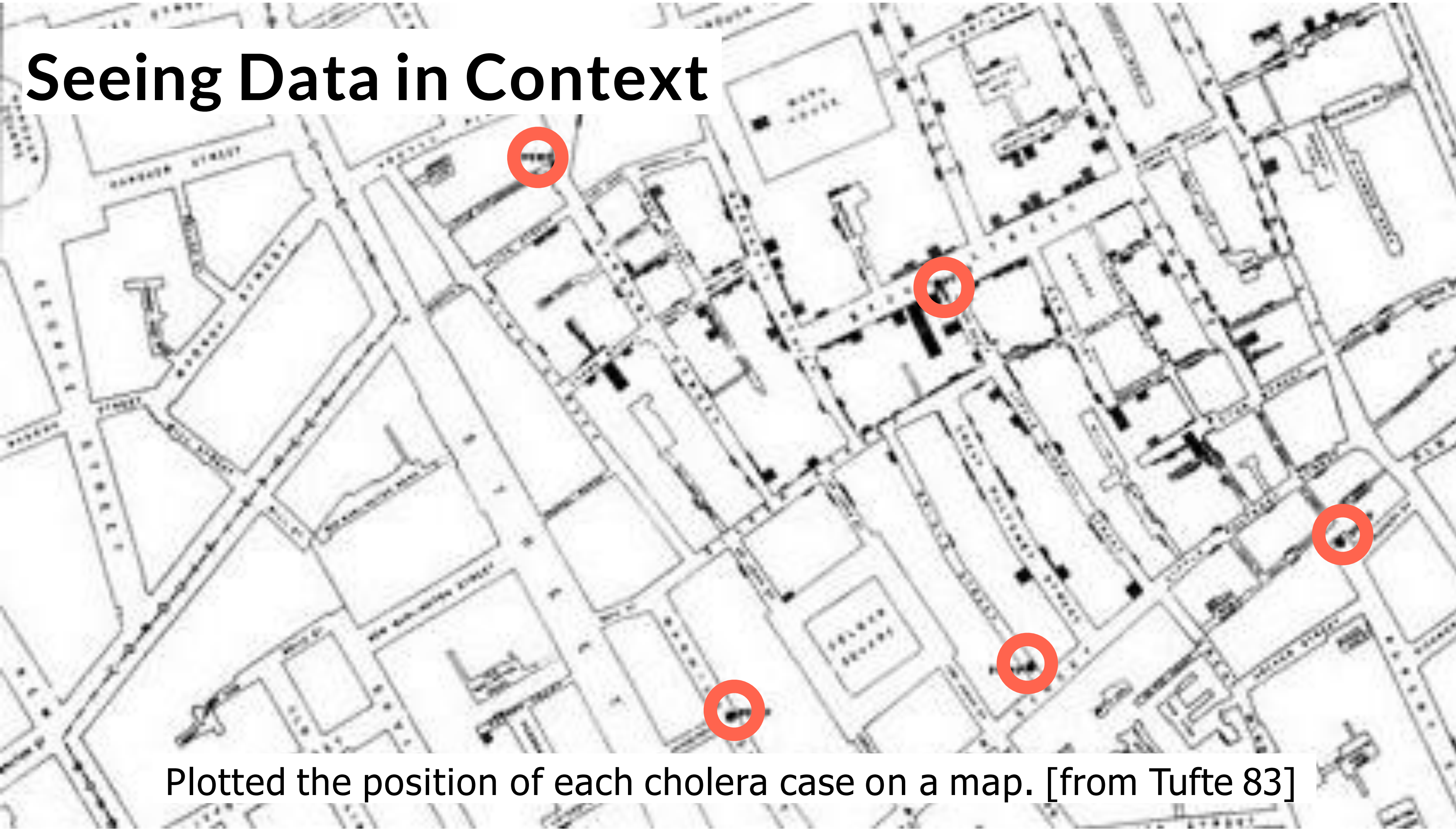
E. J. Muybridge, 1878

Support Reasoning

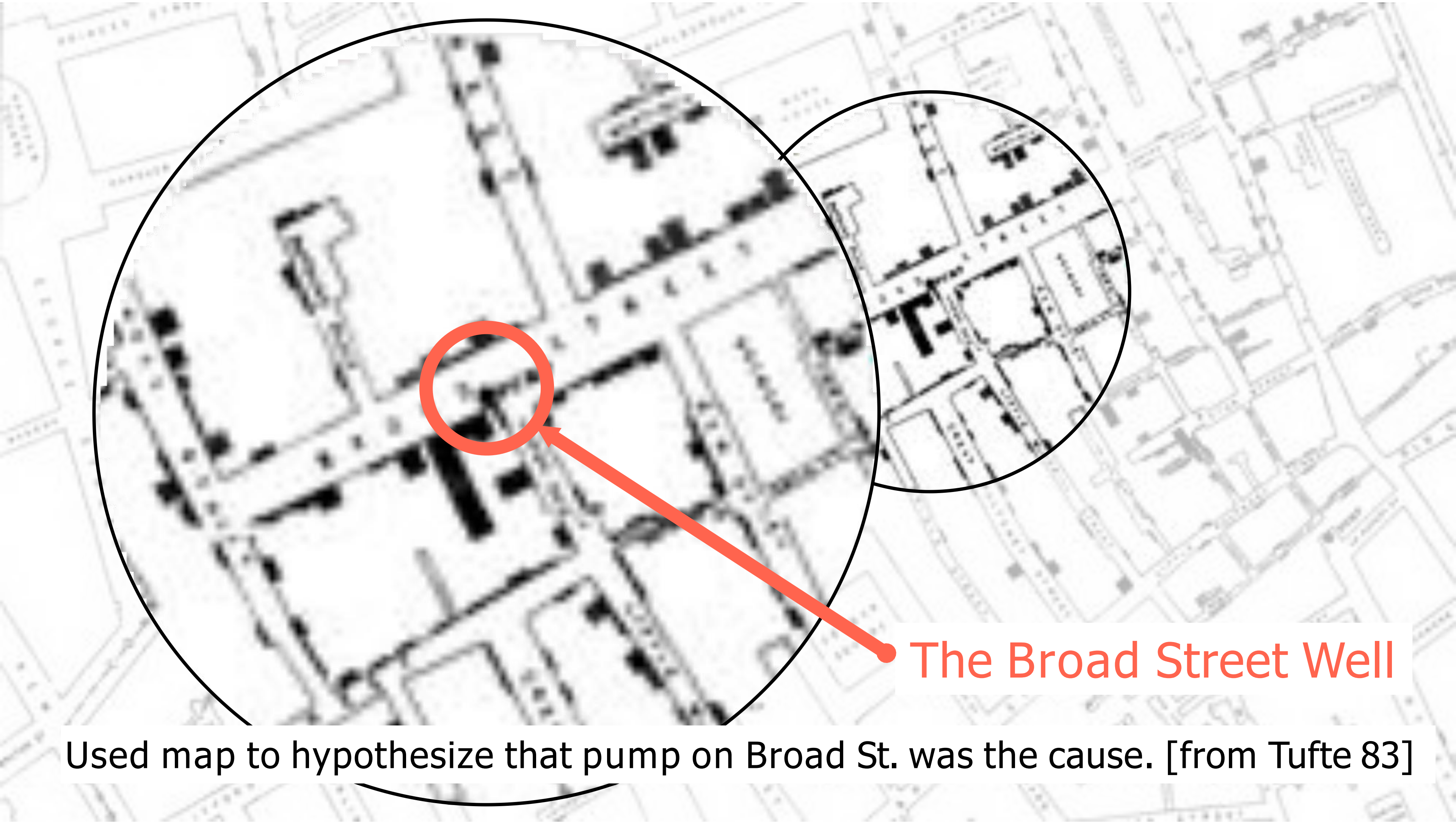


- John Snow, the Cholera Epidemic 1854

Seeing Data in Context



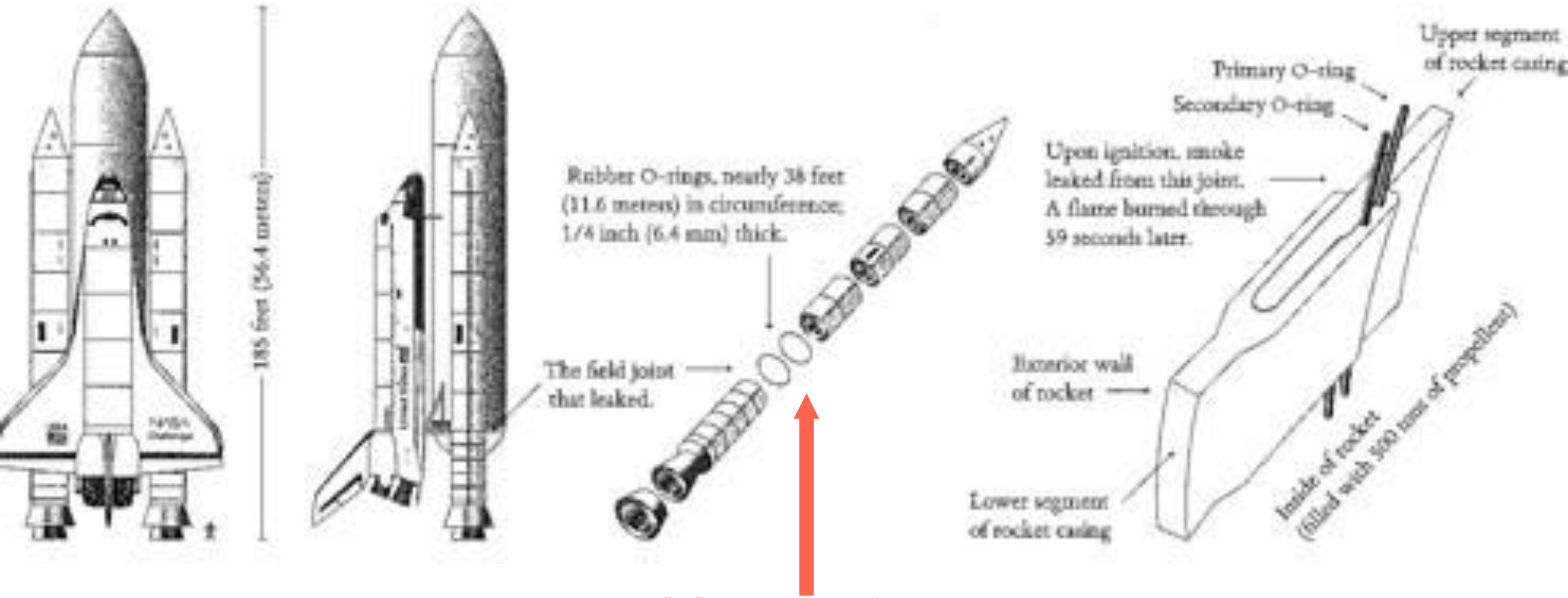
Plotted the position of each cholera case on a map. [from Tufte 83]



Used map to hypothesize that pump on Broad St. was the cause. [from Tufte 83]

Space Shuttle Challenger Disaster (1986)

approx. 73 seconds after



Rubber O-rings had problems with cold temperatures.

One of original reports sent to NASA officials before launch

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

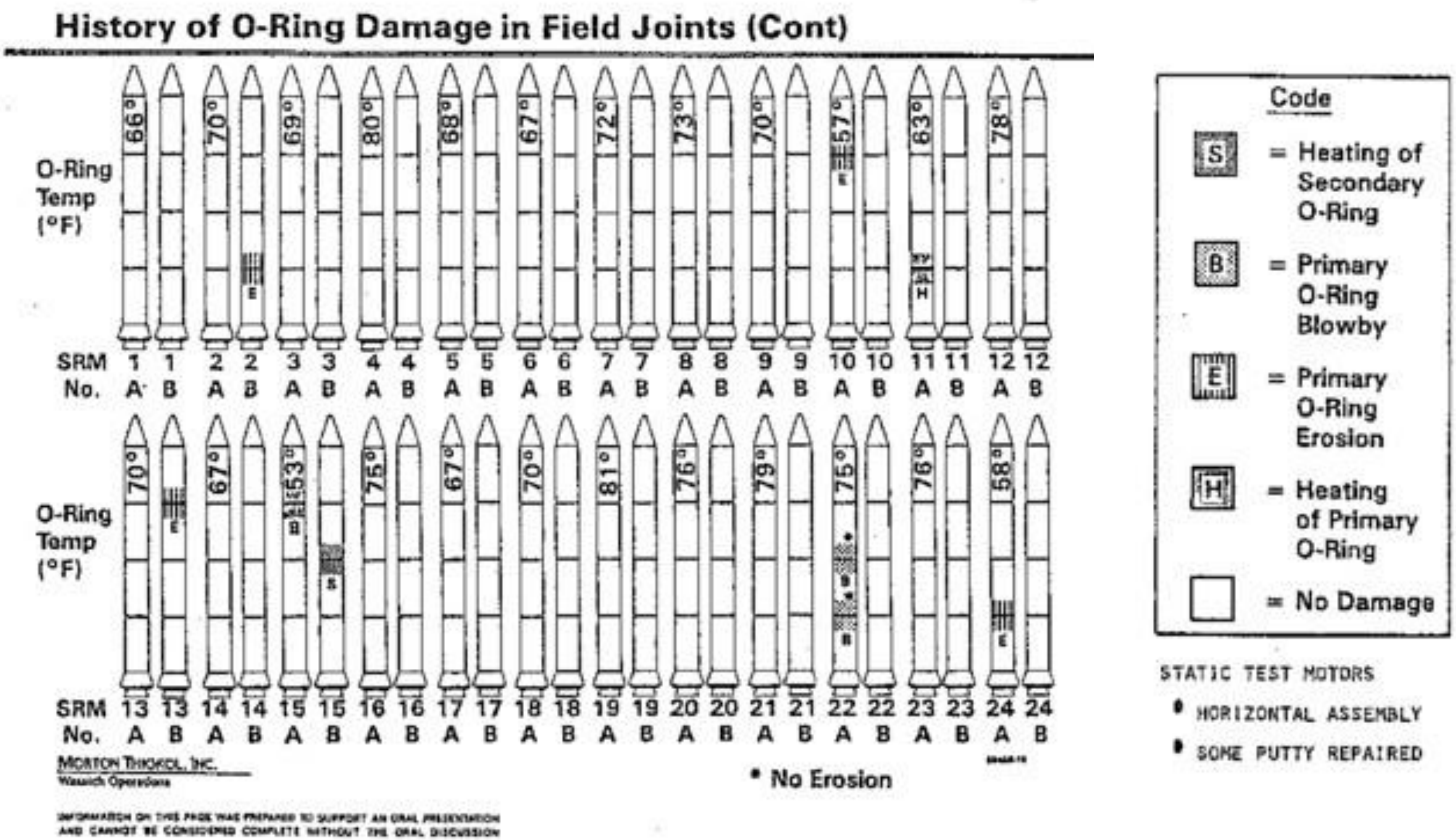
	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°-66°
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.

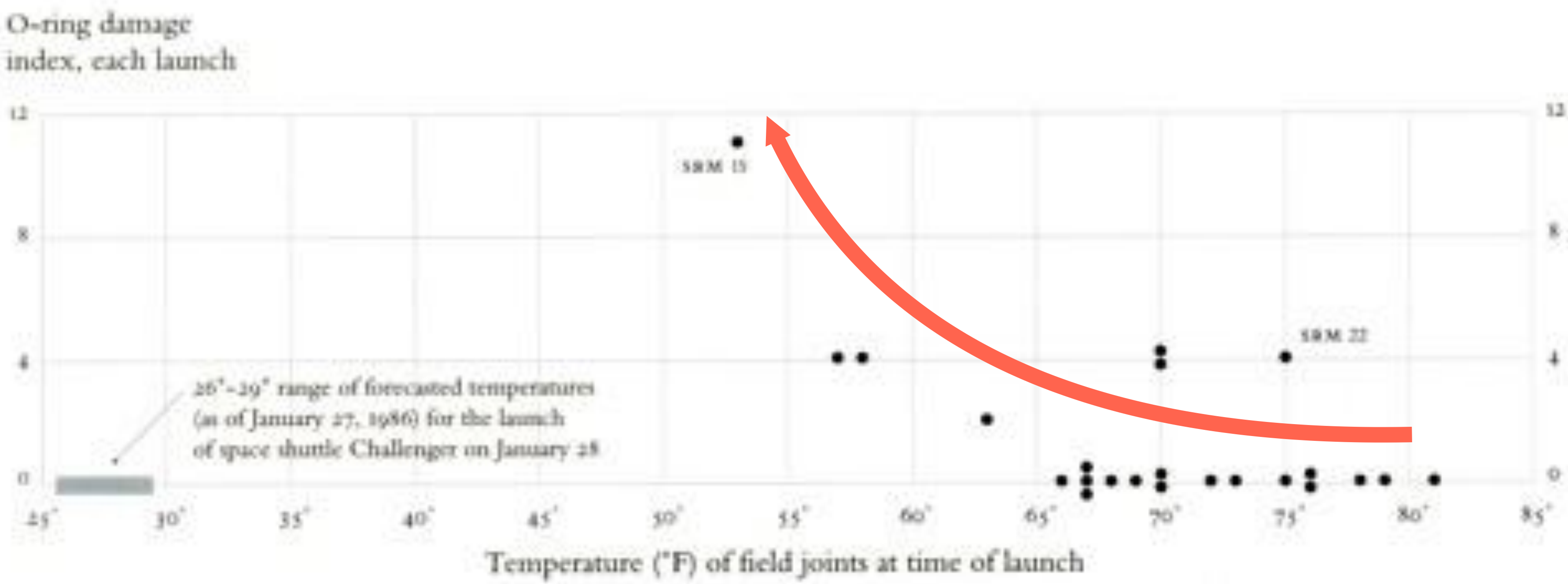
Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

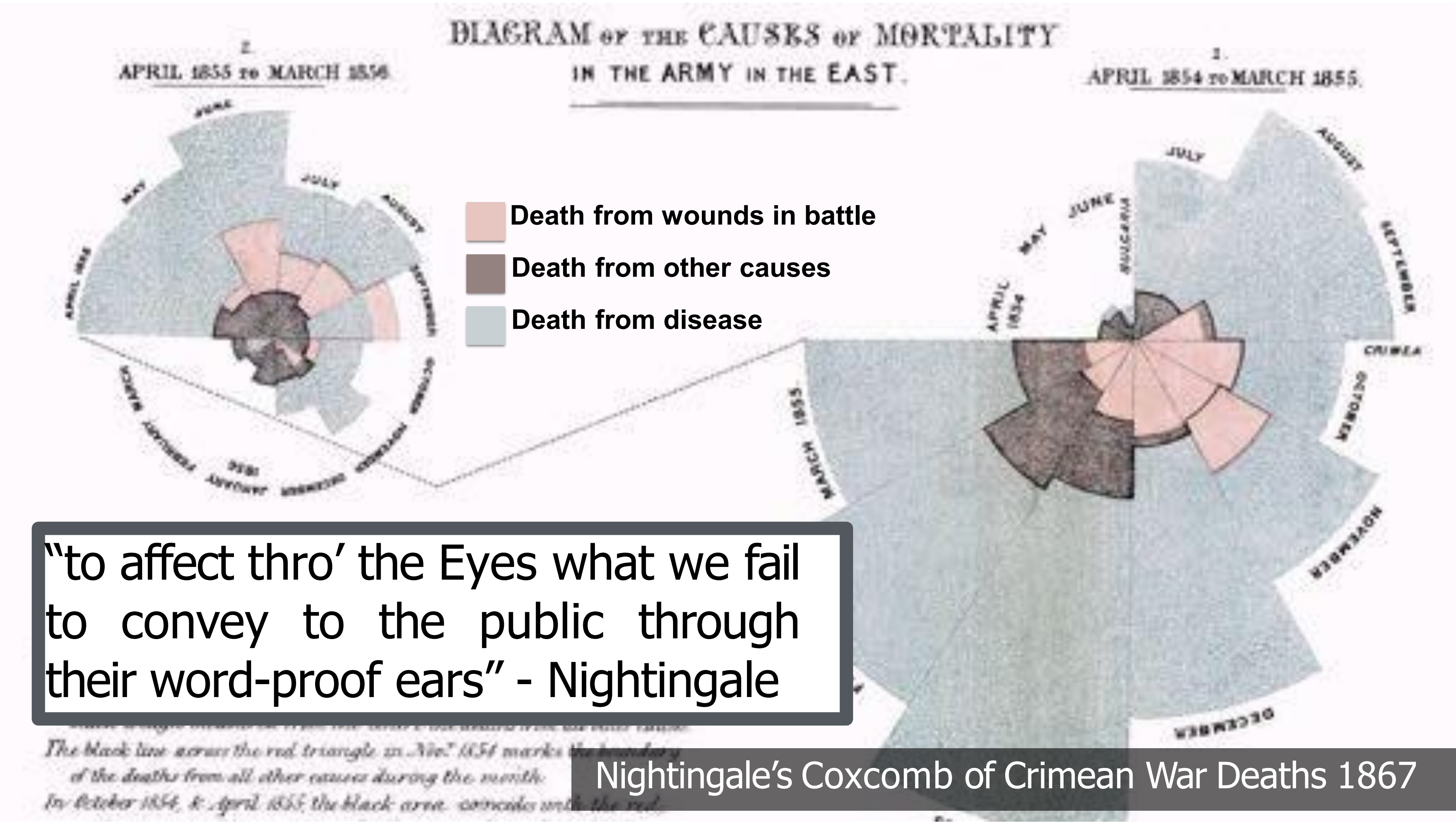

History of O-Ring Damage in Field Joints (Cont)
O-Ring Temp (°F)
66° 70° 69° 80° 68° 67° 72° 73° 70° 57° 63° 78°
SRM No. 1A 1B 2A 2B 3A 3B 4A 4B 5A 5B 6A 6B 7A 7B 8A 8B 9A 9B 10A 10B 11A 11B 12A 12B
O-Ring Temp (°F)
70° 67° 53° 75° 67° 70° 81° 76° 79° 75° 76° 58°
SRM No. 13A 13B 14A 14B 15A 15B 16A 16B 17A 17B 18A 18B 19A 19B 20A 20B 21A 21B 22A 22B 23A 23B 24A 24B
Morton Thiokol, Inc.
Wasatch Operations
* No Erosion
Code
S = Heating of Secondary O-Ring
B = Primary O-Ring Blowby
E = Primary O-Ring Erosion
H = Heating of Primary O-Ring
= No Damage
STATIC TEST MOTORS
• HORIZONTAL ASSEMBLY
• SOME PUTTY REPAIRED

Use a right visualization to make a right decision



Convey Information to Others

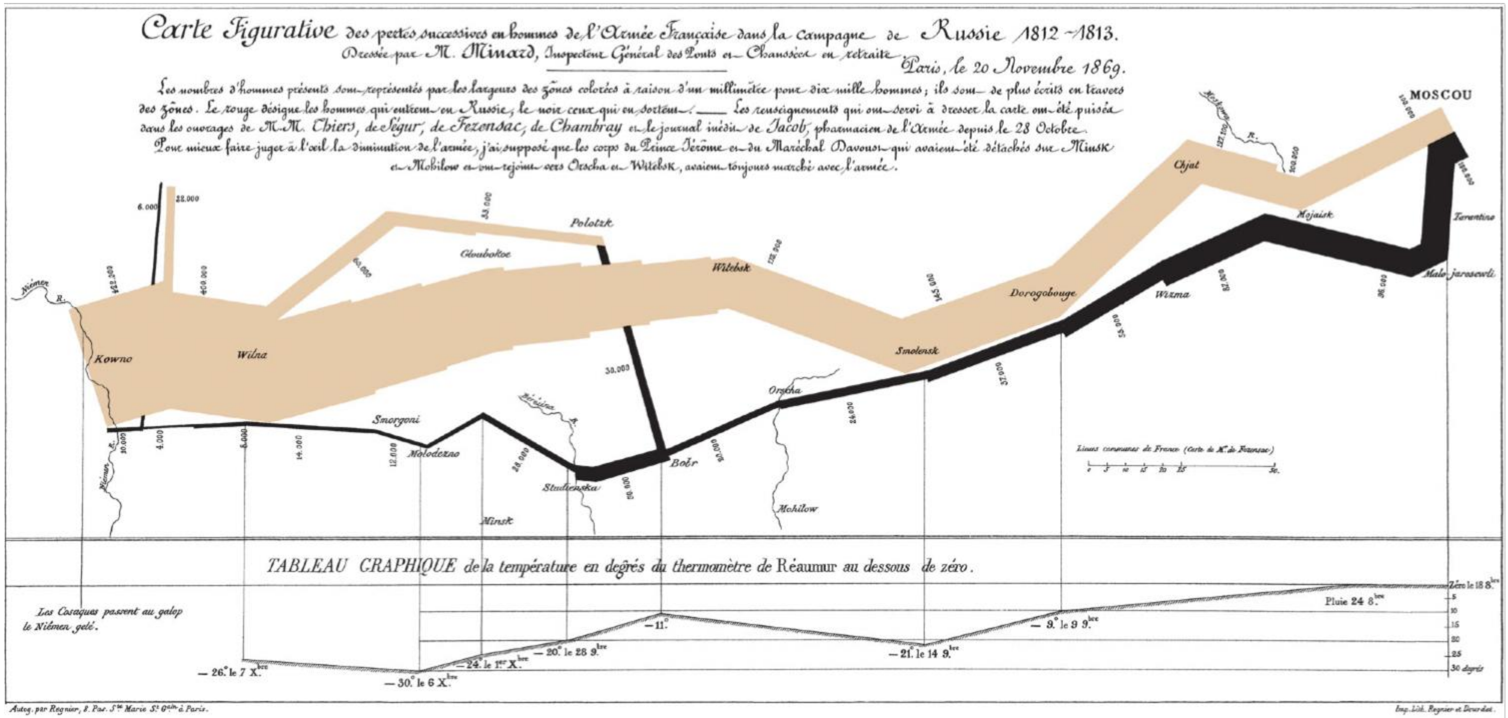
Convey Information to Others



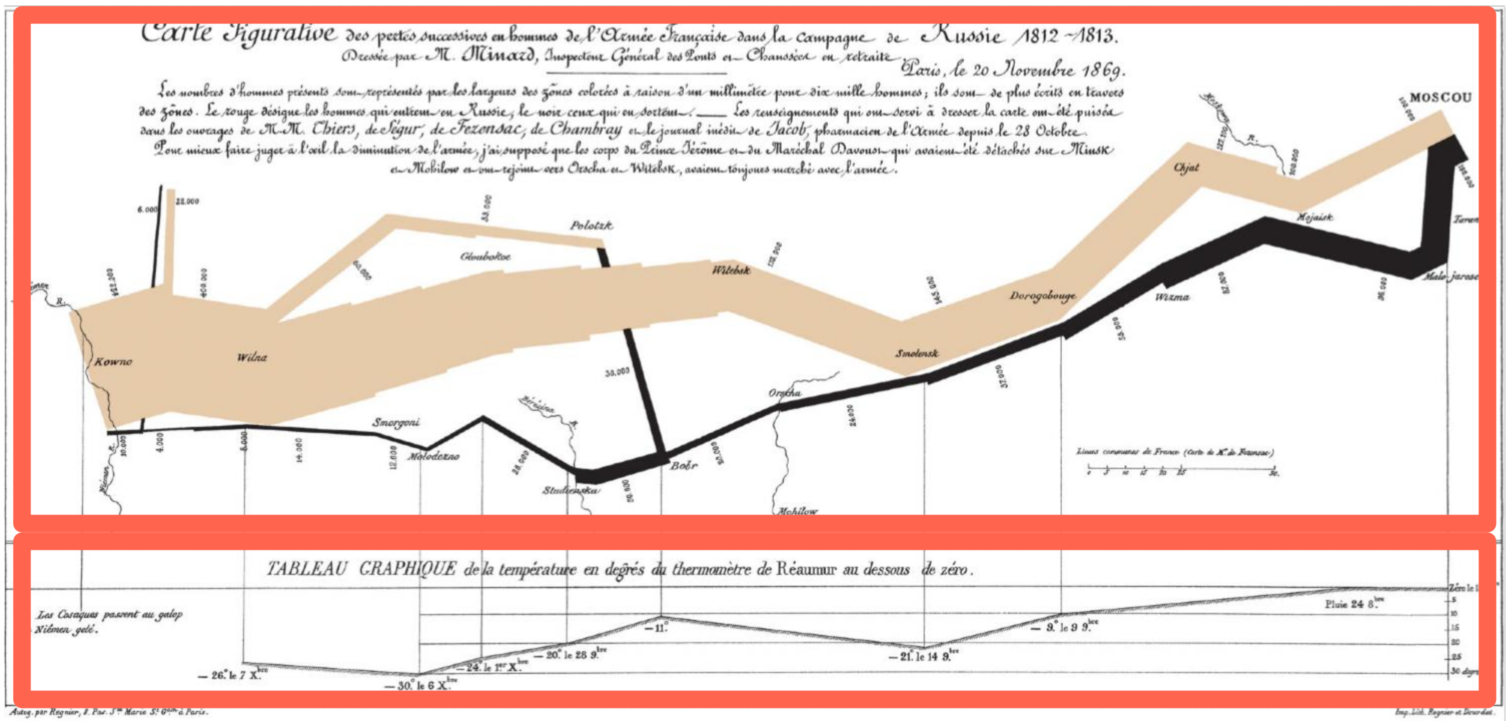
"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears" - Nightingale

Nightingale's Coxcomb of Crimean War Deaths 1867

Minard 1869: Napoleon’s March

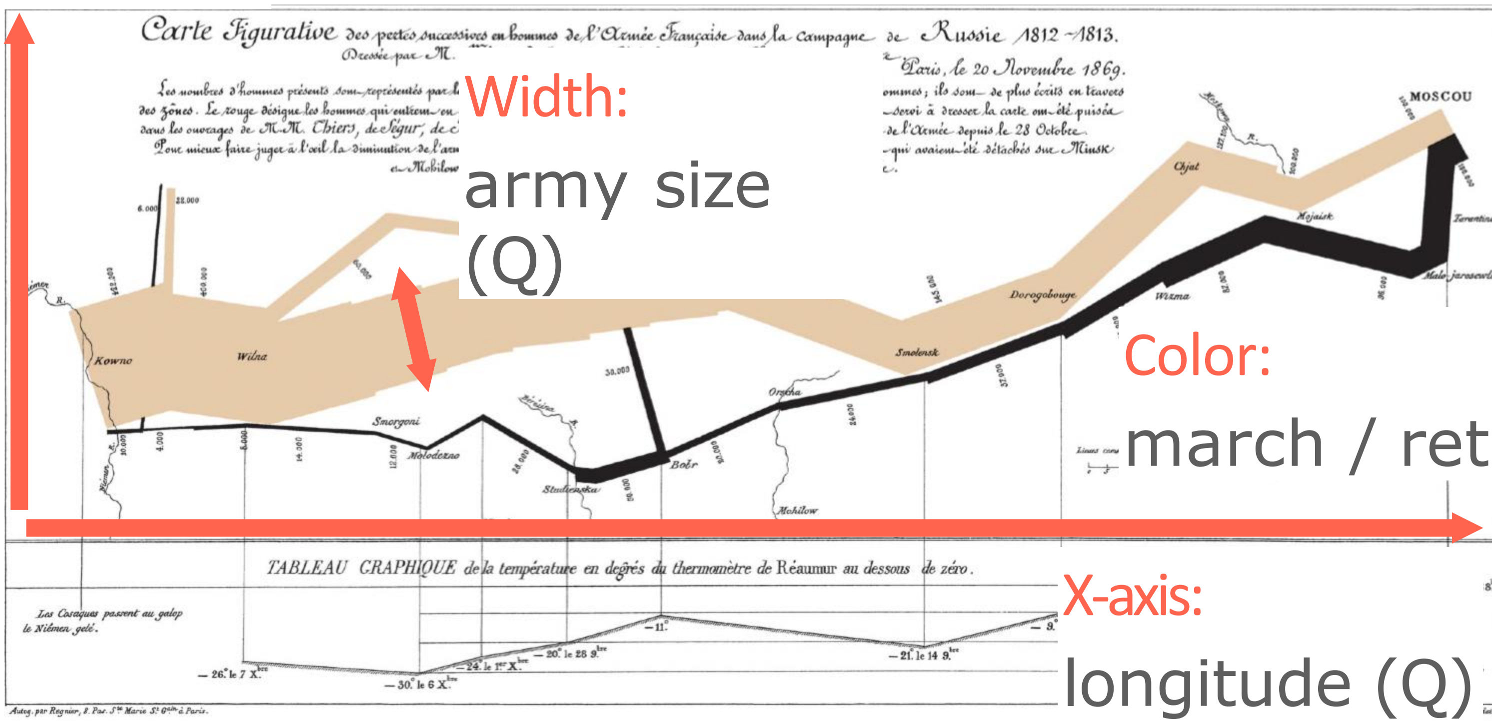


Minard 1869: Napoleon’s March



Y-axis:

latitude (Q)



Width:

army size (Q)

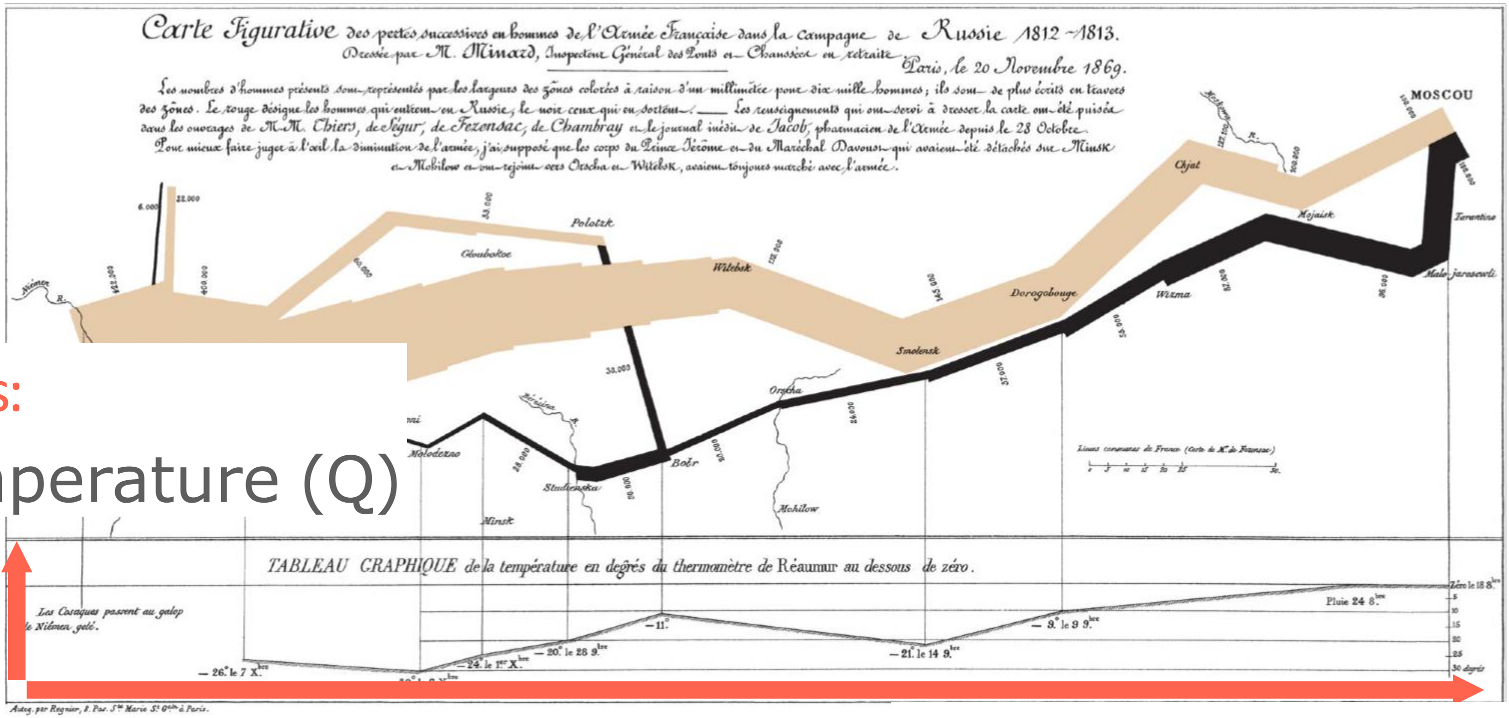
Color:

march / return

X-axis:

longitude (Q)

Minard 1869: Napoleon's March



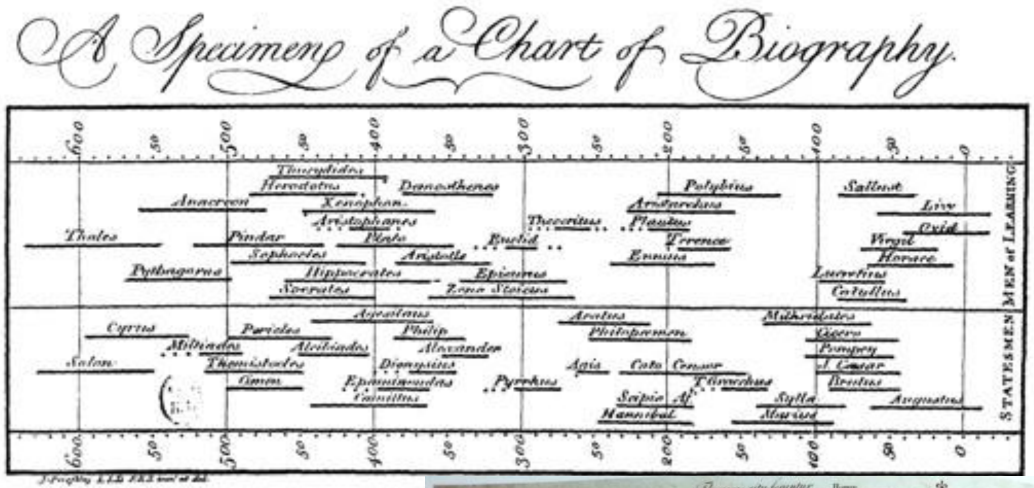
Y-axis:

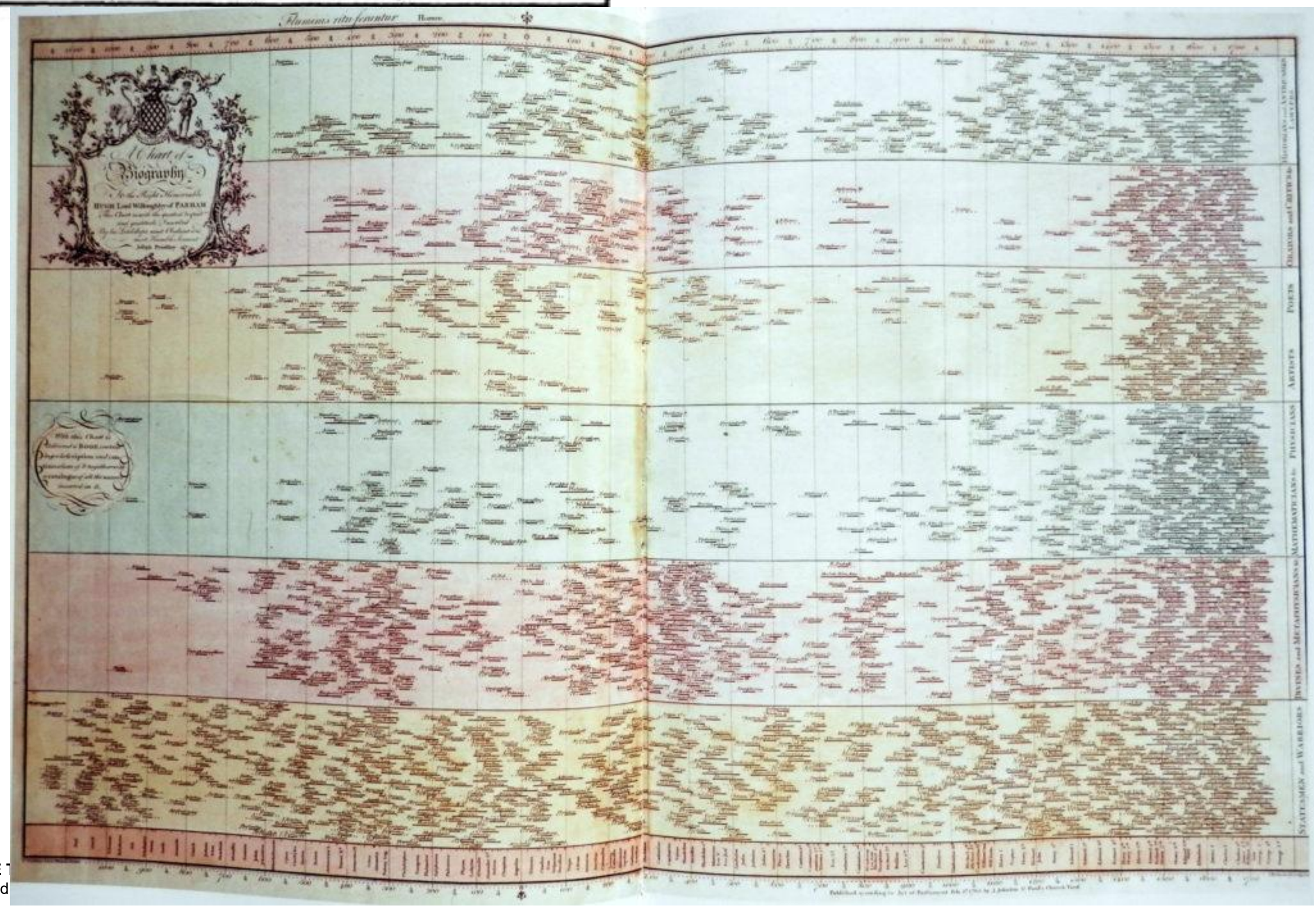
temperature (Q)

X-axis:

longitude (Q) / time (O)

Joseph Priestley 1765

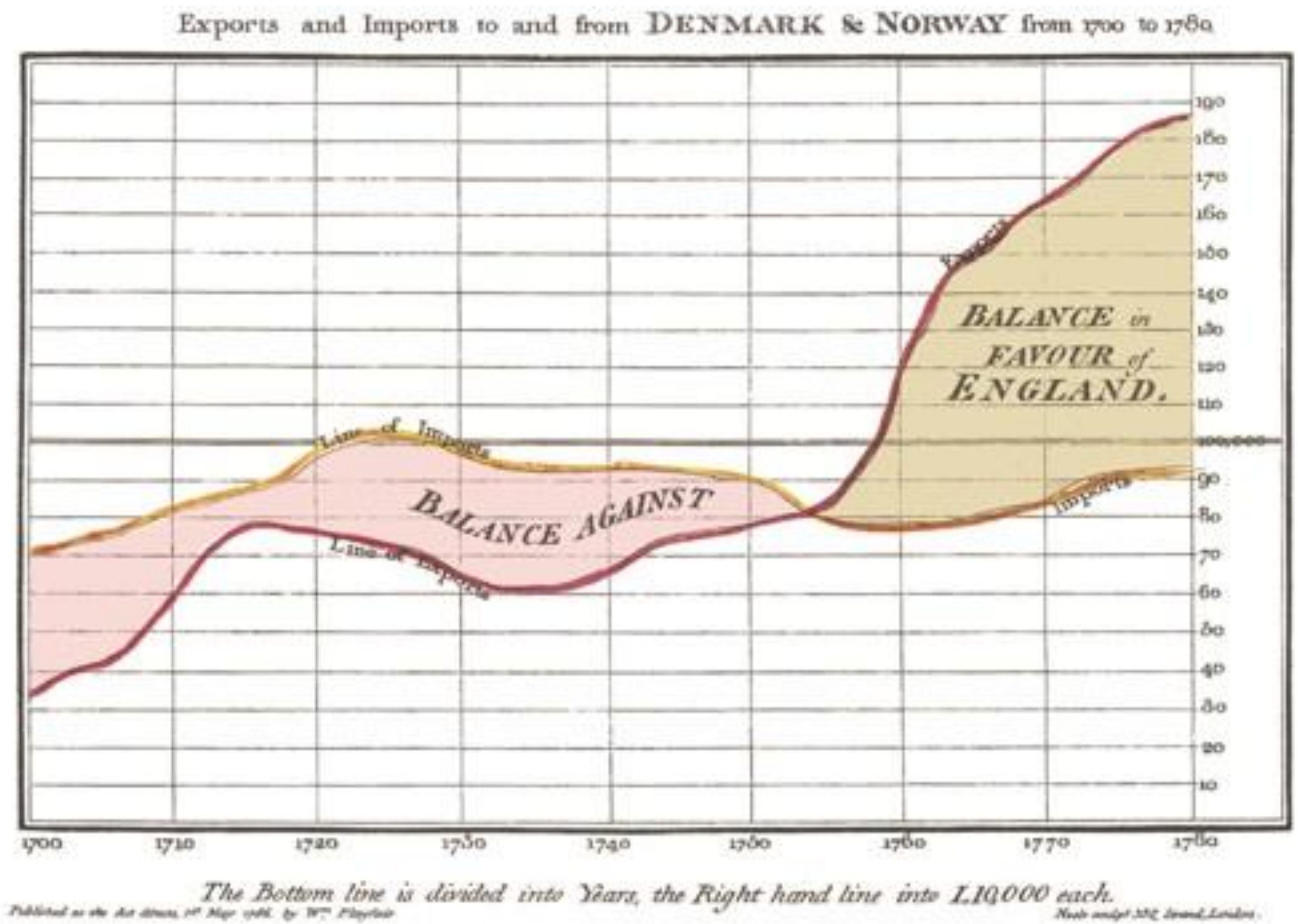




These were the first timeline charts in which individual bars were used to visualize the life span of a person

William Playfair 1786

The founder of graphical methods of statistics,[2] Playfair invented several types of diagrams: in 1786 the line, area and bar chart of economic data, and in 1801 the pie chart and circle graph, used to show part-whole relations.[3]



The value of visualization

- **Record** information
  - Blueprints, photographs, seismographs, …
- **Analyze data** to support reasoning
  - Develop and assess hypotheses
  - Explore patterns and discover the unknown
  - Expand memory
- **Communicate** information to others
  - Explain and persuade
  - Share and inspire

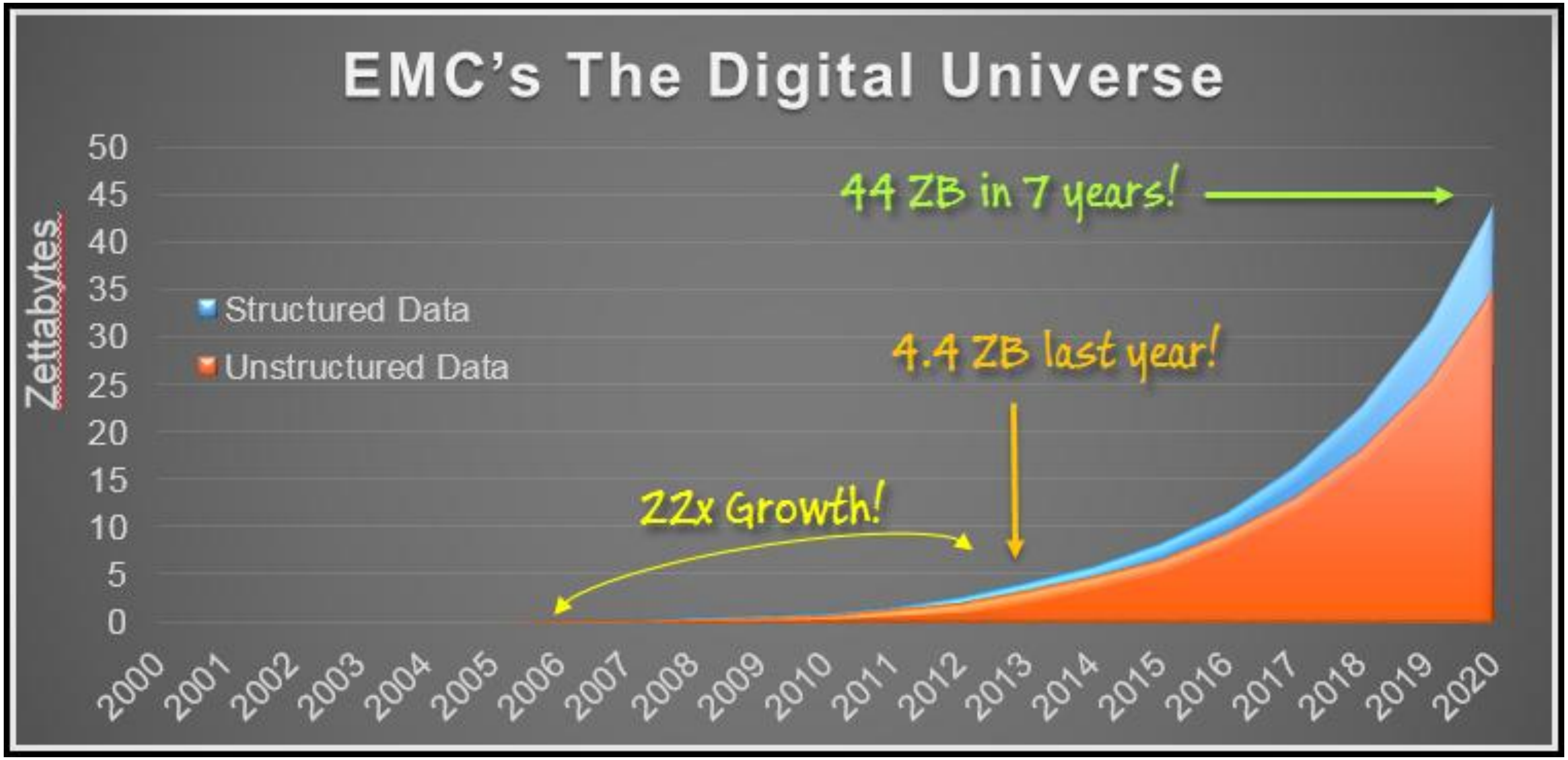
Goals of visualization research

- Understand how people perceive/comprehend visualizations
- Develop principles and techniques for effective visualizations

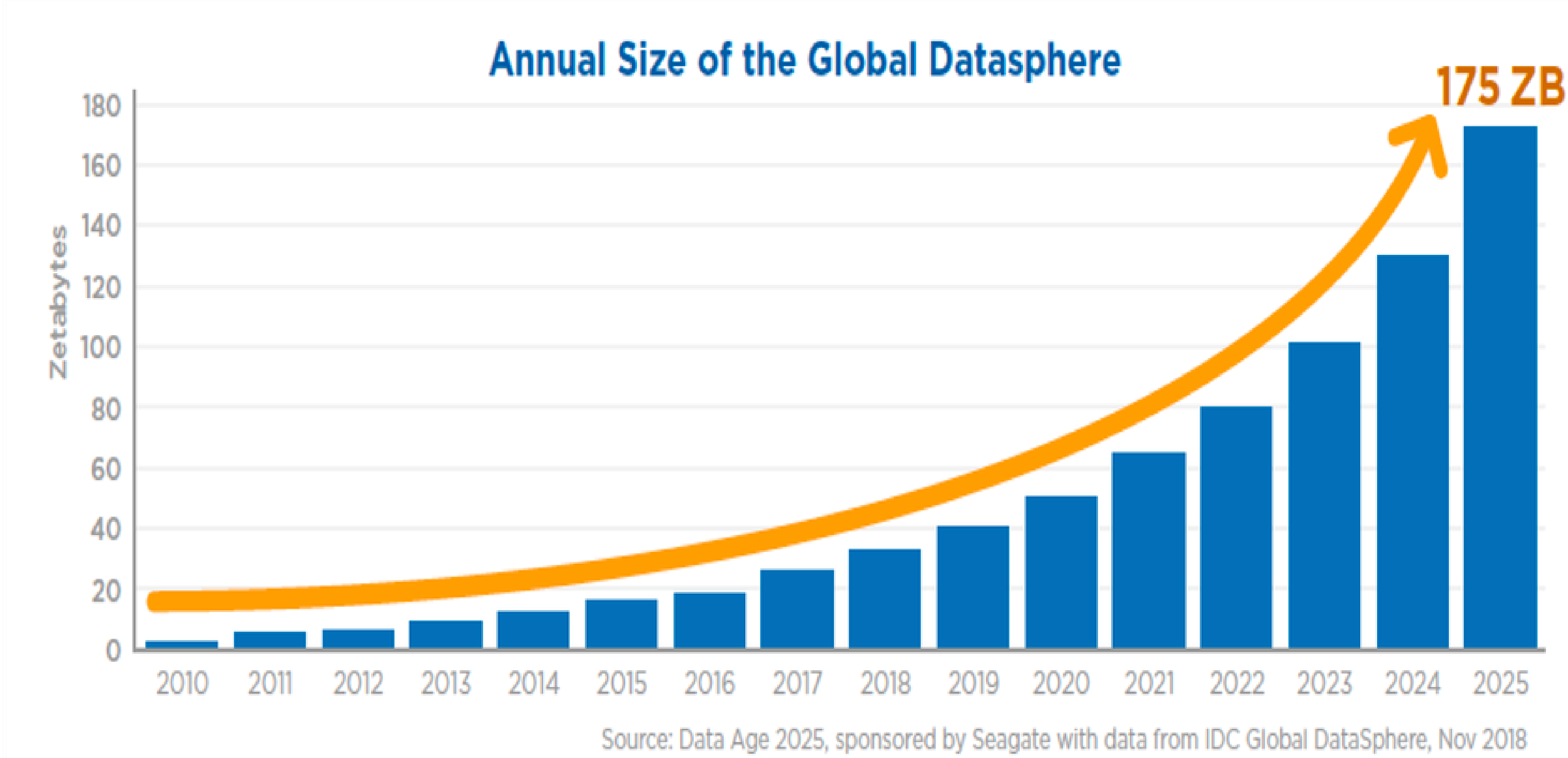
Data visualization in the big data era

The industrial revolution of data

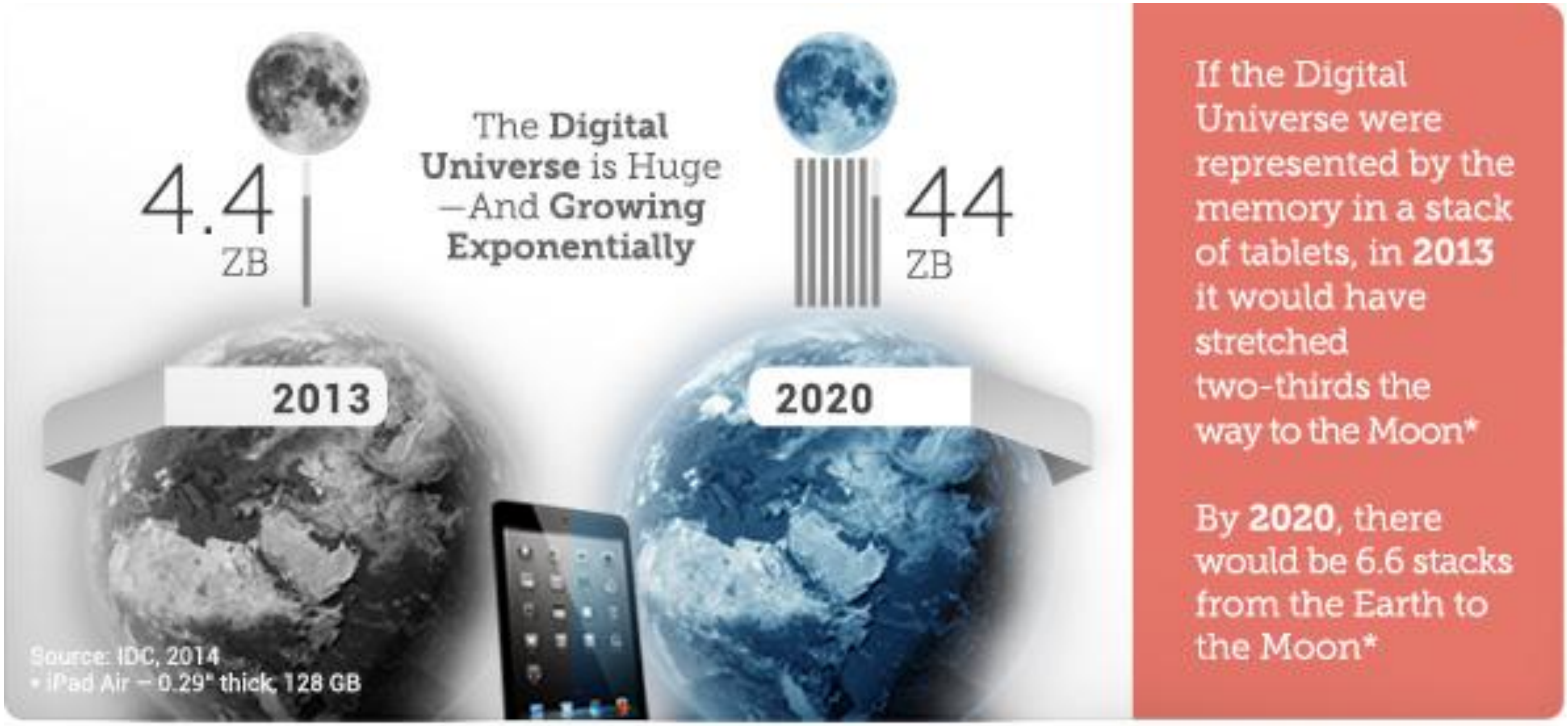
Big data growth



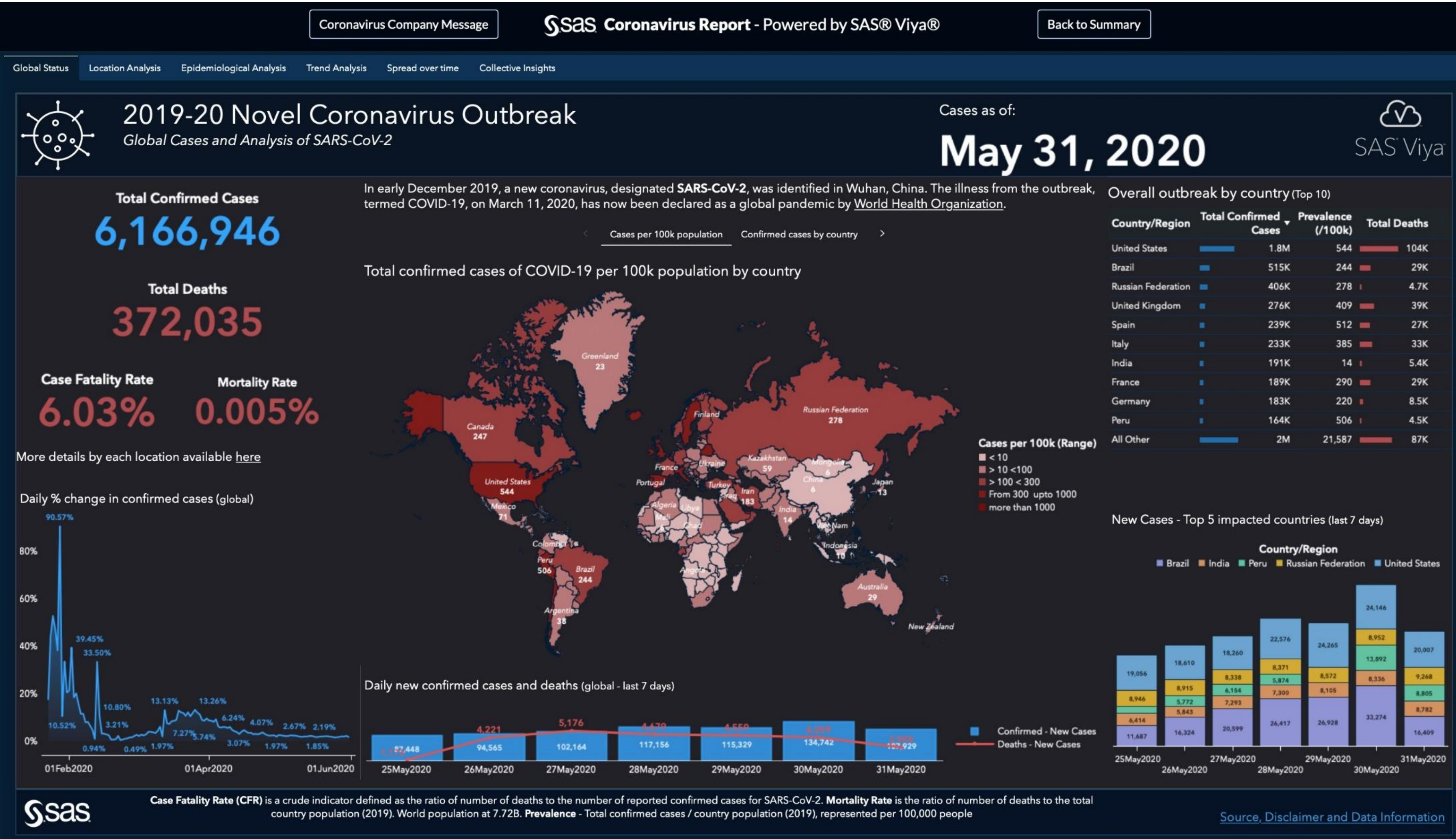
Big data growth to 2025



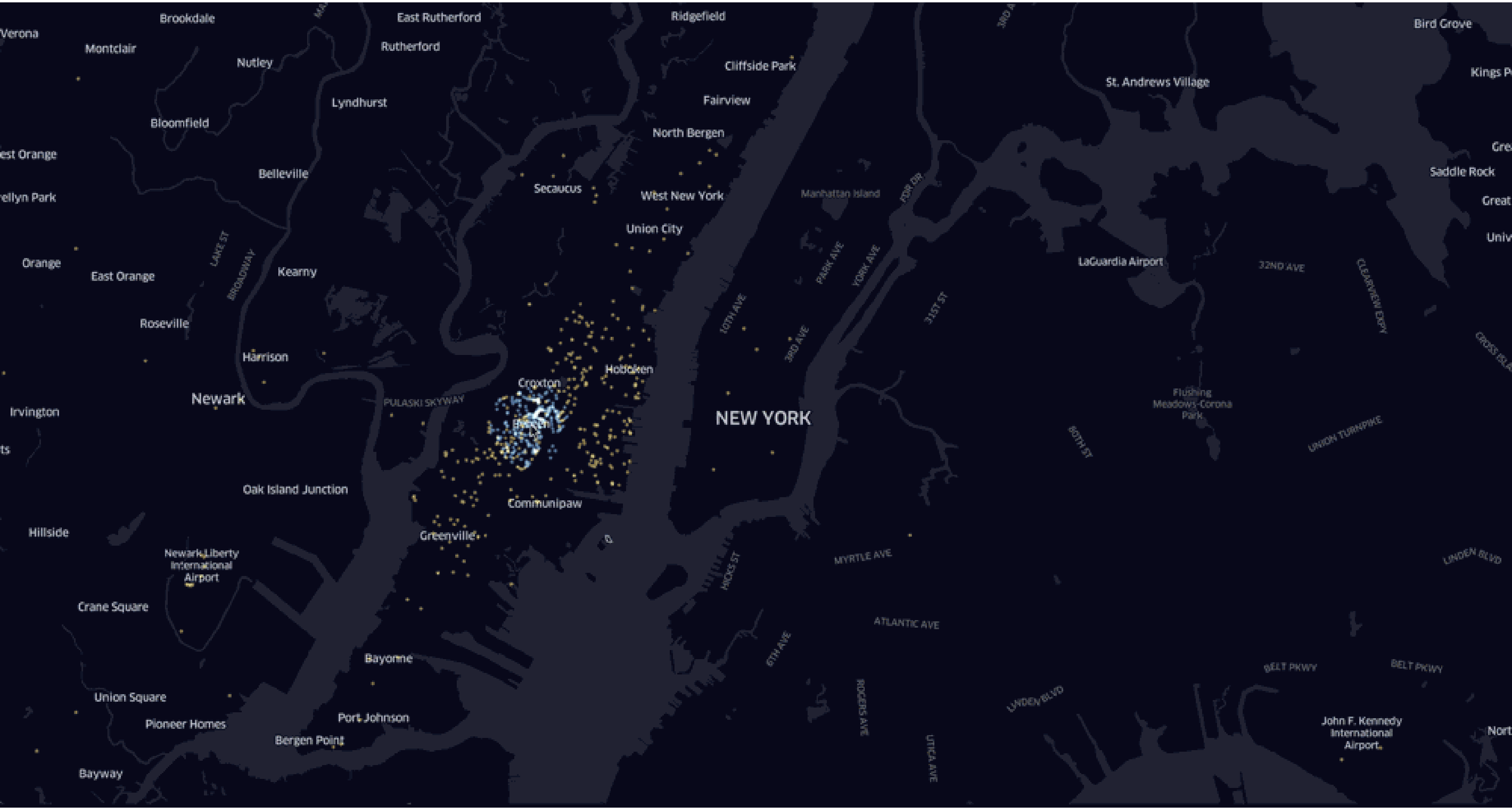
How big is big data?

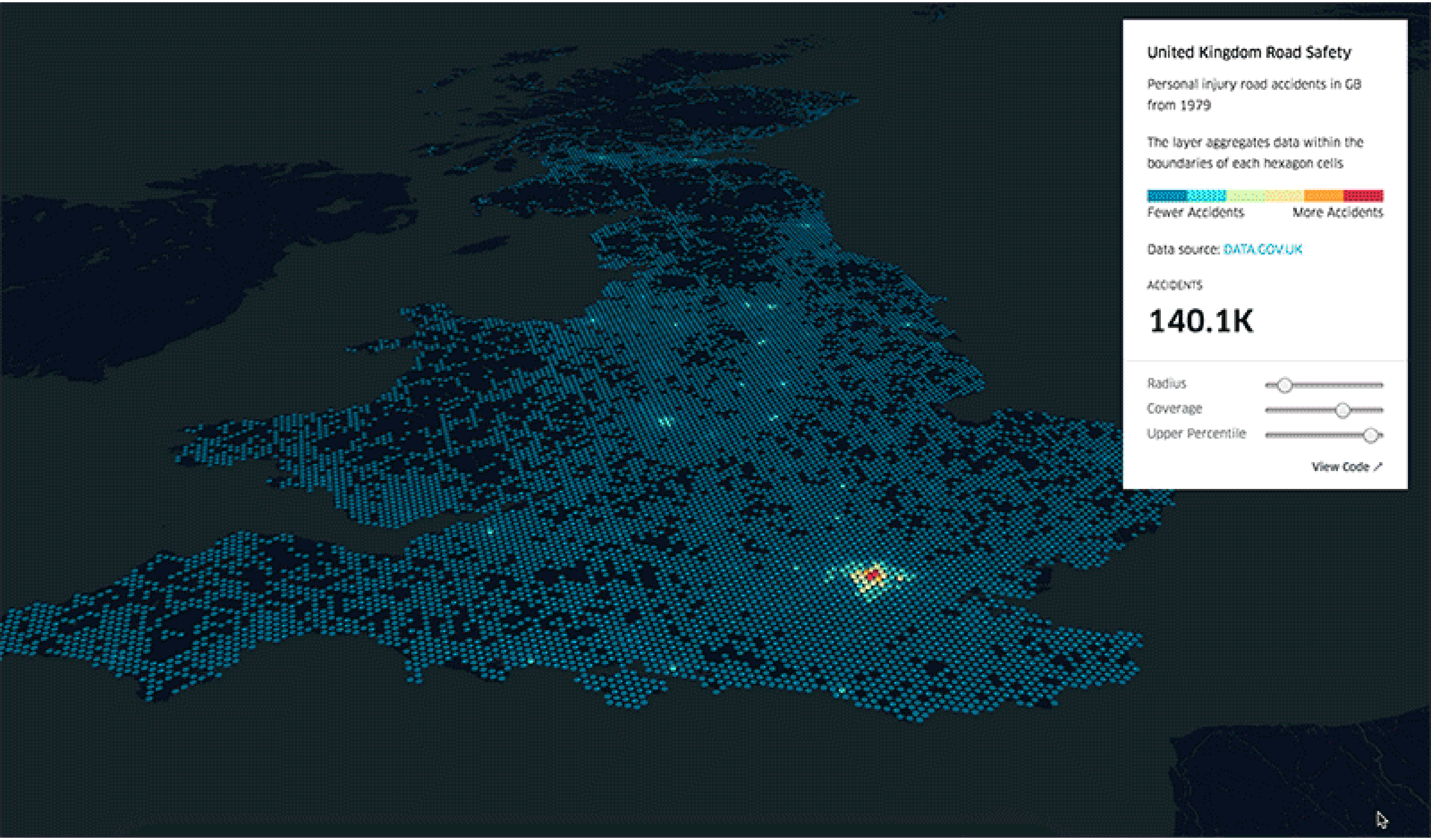


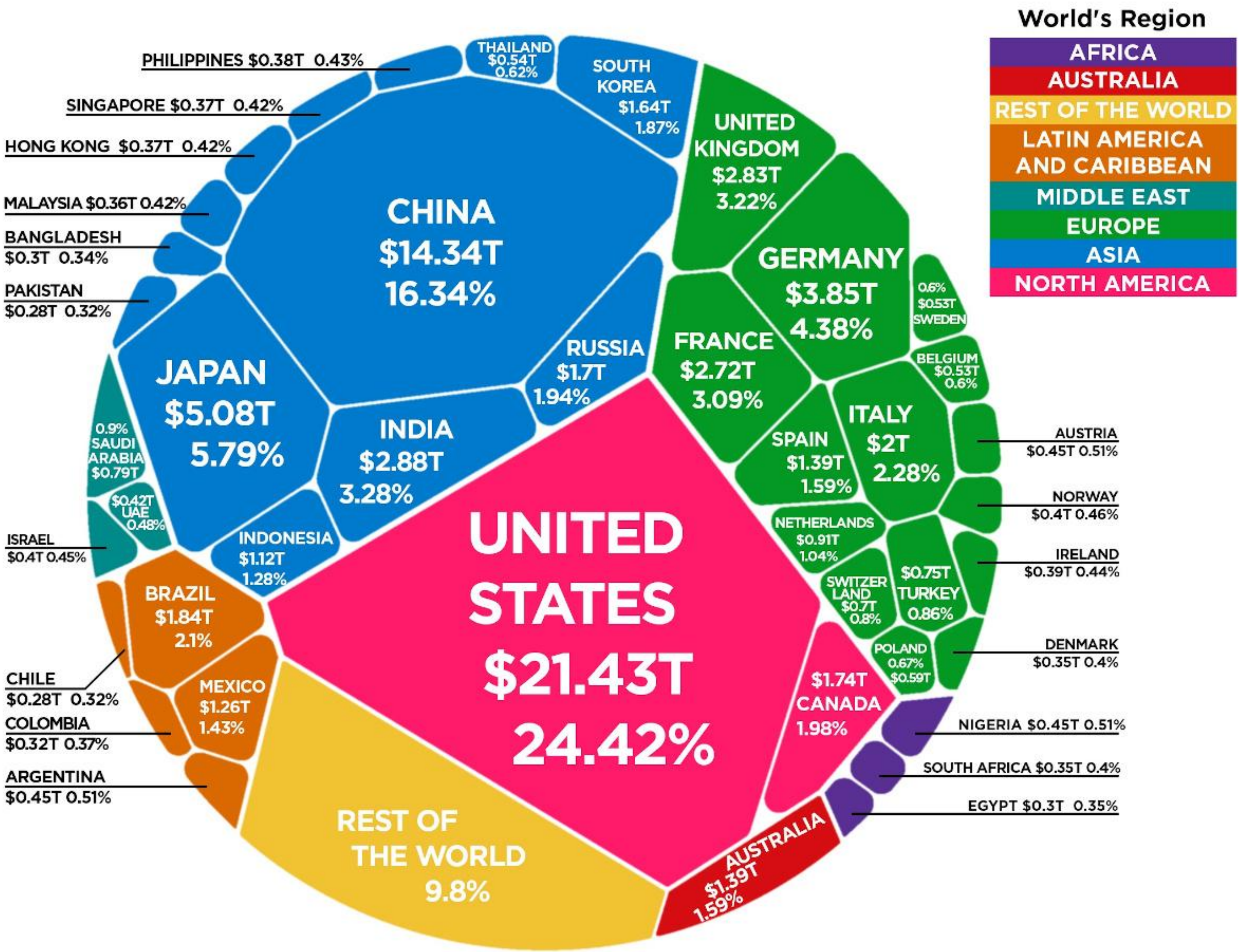
Data is everywhere



Uber traffic visualization



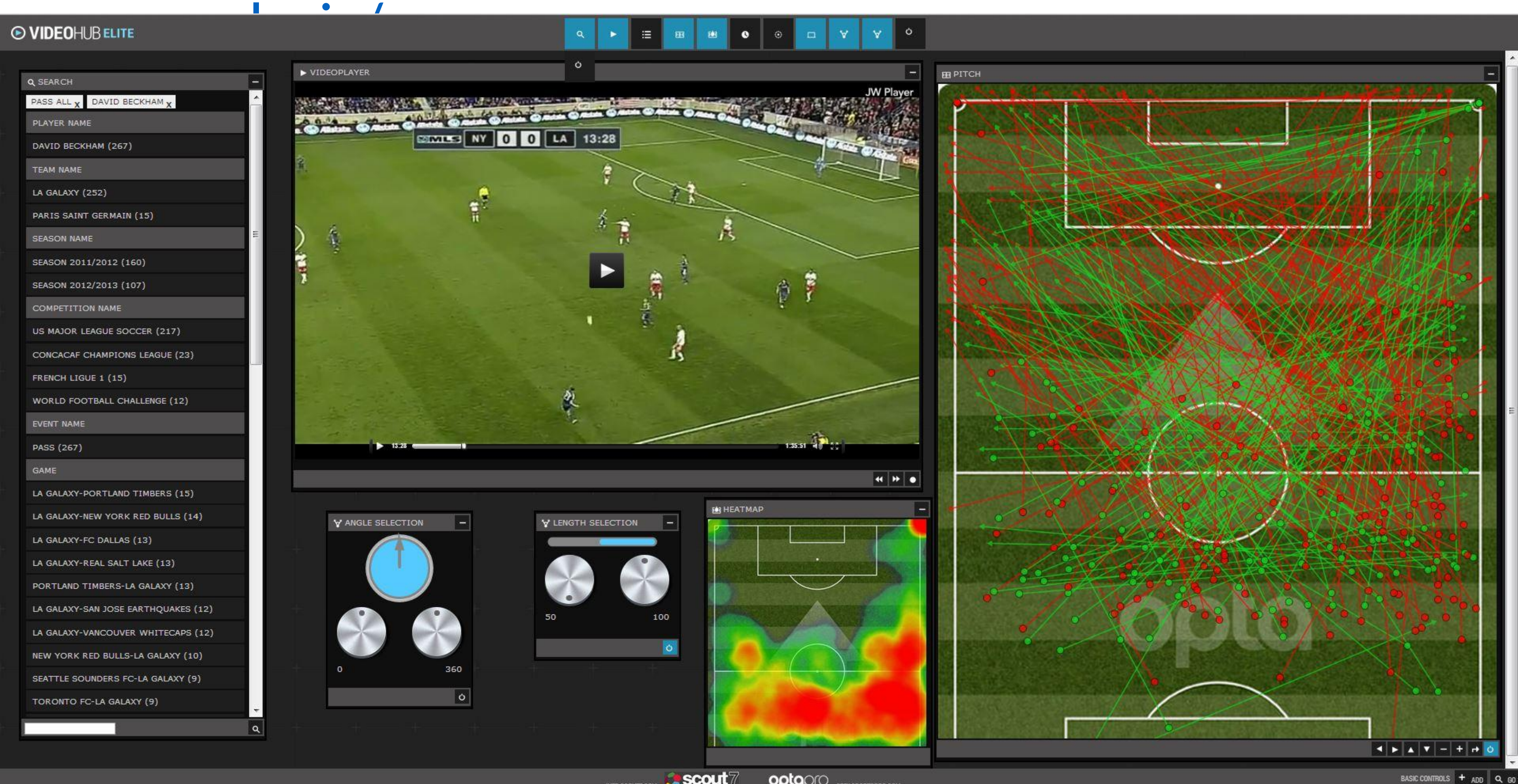

United Kingdom Road Safety
Personal injury road accidents in GB from 1979
The layer aggregates data within the boundaries of each hexagon cells
Fewer Accidents
More Accidents
Data source: DATA.GOV.UK
ACCIDENTS
140.1K
Radius
Coverage
Upper Percentile
View Code

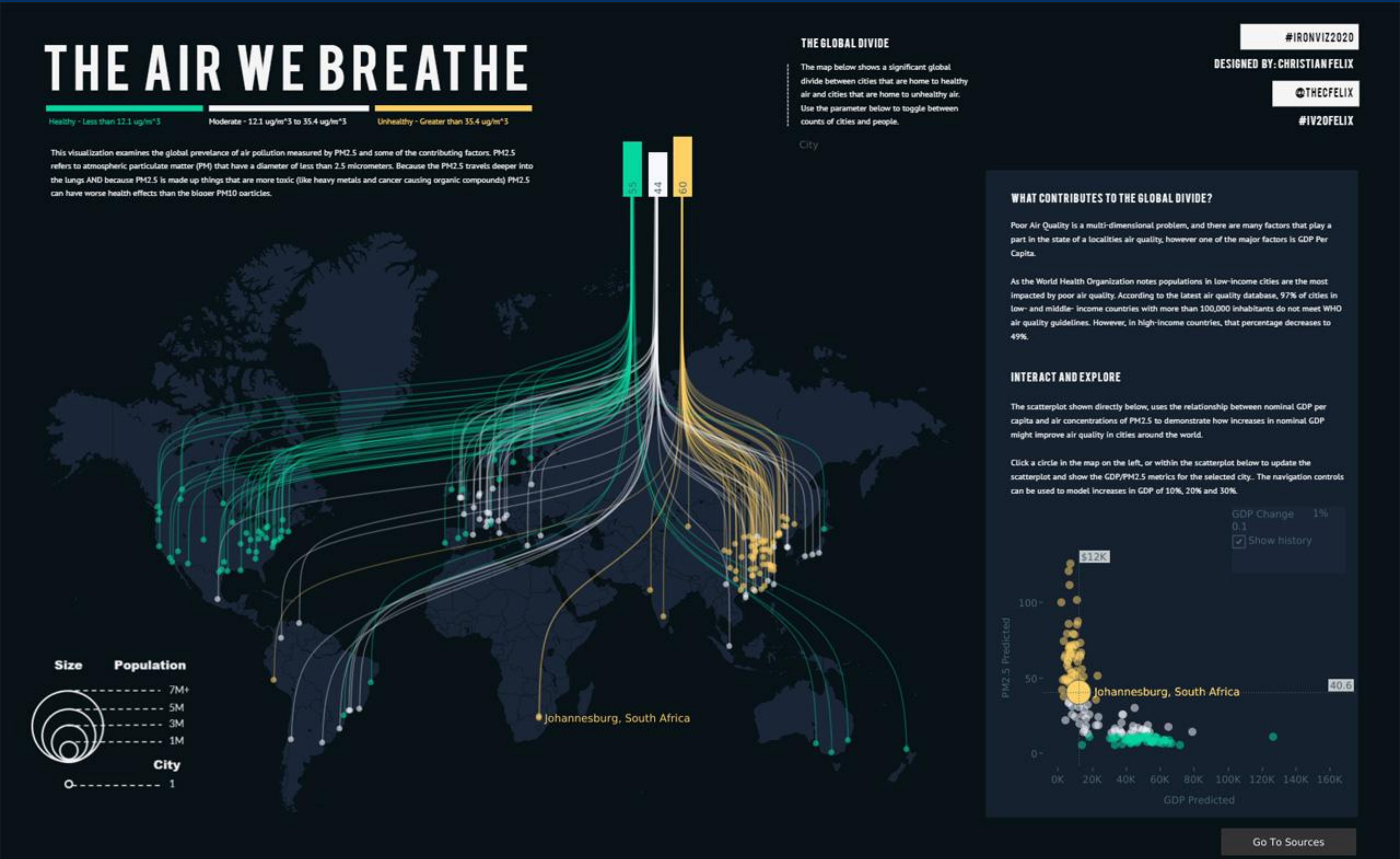


Article & Sources:
https://howmuch.net/articles/the-world-economy-2019
https://databank.worldbank.org

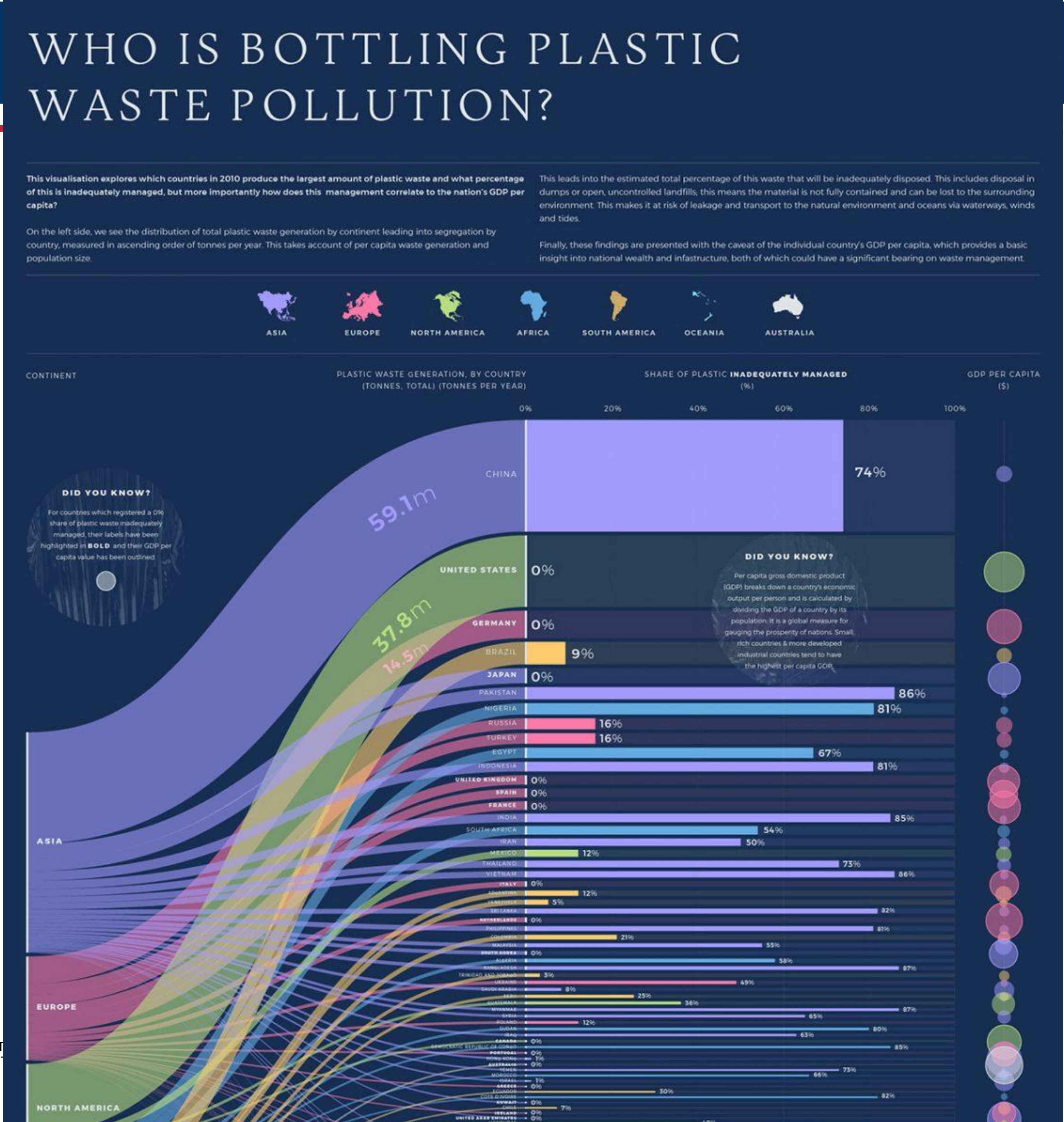
Football performance analysis

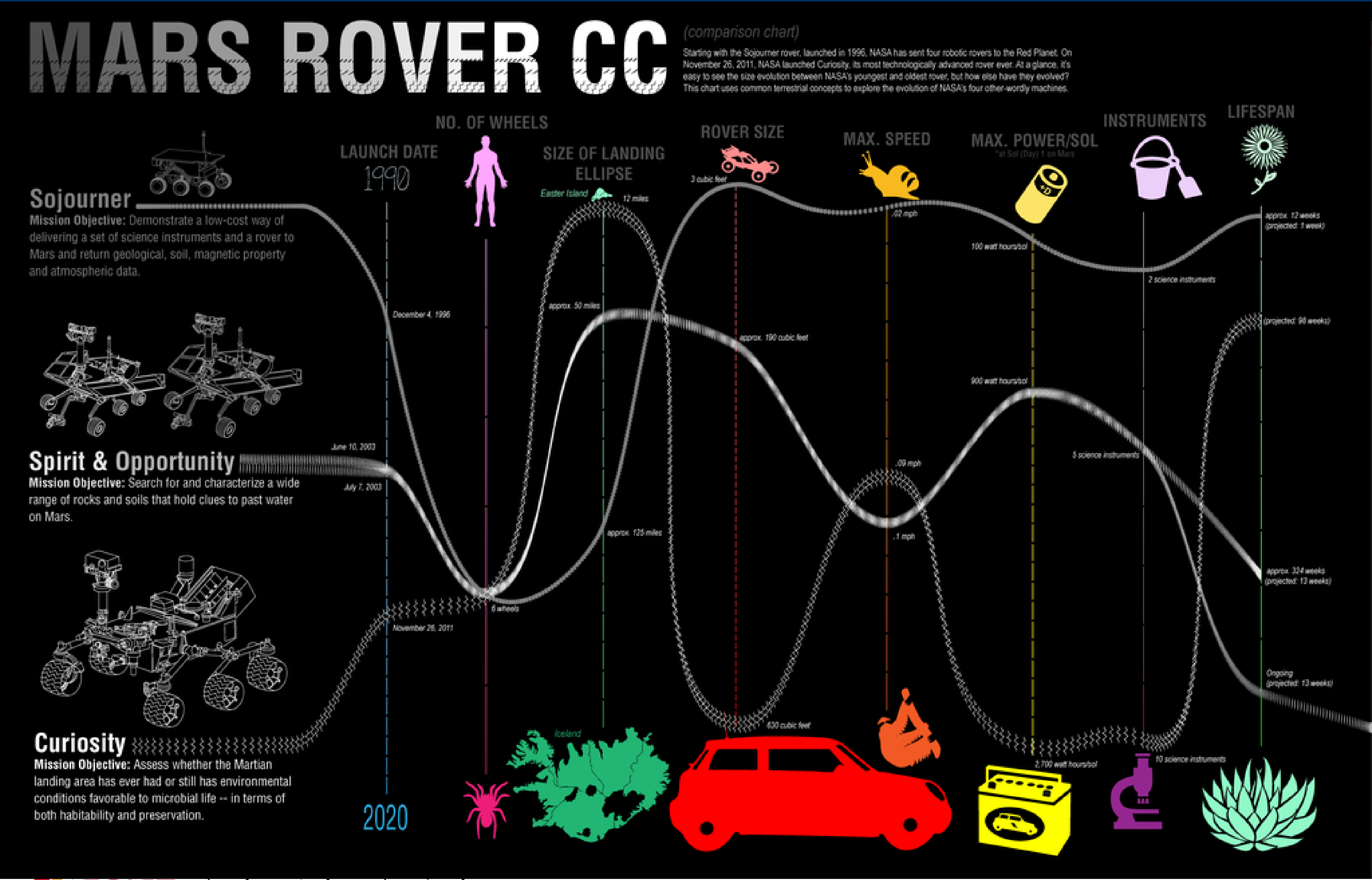
- https://www.statsperform.com/team-performance/football-performance/edge-




THE AIR WE BREATHE
Healthy - Less than 12.1 ug/m^3
Moderate - 12.1 ug/m^3 to 35.4 ug/m^3
Unhealthy - Greater than 35.4 ug/m^3
This visualization examines the global prevelance of air pollution measured by PM2.5 and some of the contributing factors. PM2.5 refers to atmospheric particulate matter (PM) that have a diameter of less than 2.5 micrometers. Because the PM2.5 travels deeper into the lungs AND because PM2.5 is made up things that are more toxic (like heavy metals and cancer causing organic compounds) PM2.5 can have worse health effects than the bigger PM10 particles.
55
44
60
THE GLOBAL DIVIDE
The map below shows a significant global divide between cities that are home to healthy air and cities that are home to unhealthy air. Use the parameter below to toggle between counts of cities and people.
City
#IRONVIZ2020
DESIGNED BY: CHRISTIAN FELIX
@THECFELIX
#IV20FELIX
WHAT CONTRIBUTES TO THE GLOBAL DIVIDE?
Poor Air Quality is a multi-dimensional problem, and there are many factors that play a part in the state of a localities air quality, however one of the major factors is GDP Per Capita.
As the World Health Organization notes populations in low-income cities are the most impacted by poor air quality. According to the latest air quality database, 97% of cities in low- and middle- income countries with more than 100,000 inhabitants do not meet WHO air quality guidelines. However, in high-income countries, that percentage decreases to 49%.
INTERACT AND EXPLORE
The scatterplot shown directly below, uses the relationship between nominal GDP per capita and air concentrations of PM2.5 to demonstrate how increases in nominal GDP might improve air quality in cities around the world.
Click a circle in the map on the left, or within the scatterplot below to update the scatterplot and show the GDP/PM2.5 metrics for the selected city. The navigation controls can be used to model increases in GDP of 10%, 20% and 30%.
GDP Change 1%
0.1
Show history
$12K
Johannesburg, South Africa
40.6
PM2.5 Predicted
100
50
0
0K 20K 40K 60K 80K 100K 120K 140K 160K
GDP Predicted
Johannesburg, South Africa
Size Population
7M+
5M
3M
1M
City
1
Go To Sources

https://www.behance.net/gallery/106936329/Plastic-Waste-Pollution-data-visualisation




MARS ROVER CC
(comparison chart)
Starting with the Sojourner rover, launched in 1996, NASA has sent four robotic rovers to the Red Planet. On November 26, 2011, NASA launched Curiosity, its most technologically advanced rover ever. At a glance, it's easy to see the size evolution between NASA's youngest and oldest rover, but how else have they evolved? This chart uses common terrestrial concepts to explore the evolution of NASA's four other-wordly machines.
Sojourner
Mission Objective: Demonstrate a low-cost way of delivering a set of science instruments and a rover to Mars and return geological, soil, magnetic property and atmospheric data.
Spirit & Opportunity
Mission Objective: Search for and characterize a wide range of rocks and soils that hold clues to past water on Mars.
Curiosity
Mission Objective: Assess whether the Martian landing area has ever had or still has environmental conditions favorable to microbial life -- in terms of both habitability and preservation.
LAUNCH DATE
1990
December 4, 1996
June 10, 2003
July 7, 2003
November 26, 2011
2020
NO. OF WHEELS
6 wheels
SIZE OF LANDING ELLIPSE
Easter Island
12 miles
approx. 50 miles
approx. 125 miles
Iceland
ROVER SIZE
3 cubic feet
approx. 190 cubic feet
630 cubic feet
MAX. SPEED
.02 mph
.09 mph
.1 mph
MAX. POWER/SOL
*at Sol (Day) 1 on Mars
100 watt hours/sol
900 watt hours/sol
2,700 watt hours/sol
INSTRUMENTS
2 science instruments
5 science instruments
10 science instruments
LIFESPAN
approx. 12 weeks
(projected: 1 week)
(projected: 90 weeks)
approx. 324 weeks
(projected: 13 weeks)
Ongoing
(projected: 13 weeks)

Data literacy

"The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades …"

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, January 2009

A poverty of attention

"…Information consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, a need to allocate that attention efficiently among the overabundance of information sources that might consume it."

Herbert A. Simon
Economist & Psychologist

Visualization can help!

provides a powerful yet accessible way to make sense of large and complex data