



ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Introduction to visualization

Viet-Trung Tran

Credit: Prof. Nam Wook Kim

Outline

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

What is Visualization?

What is visualization

“Transformation of the **symbolic** into the **geometric**”

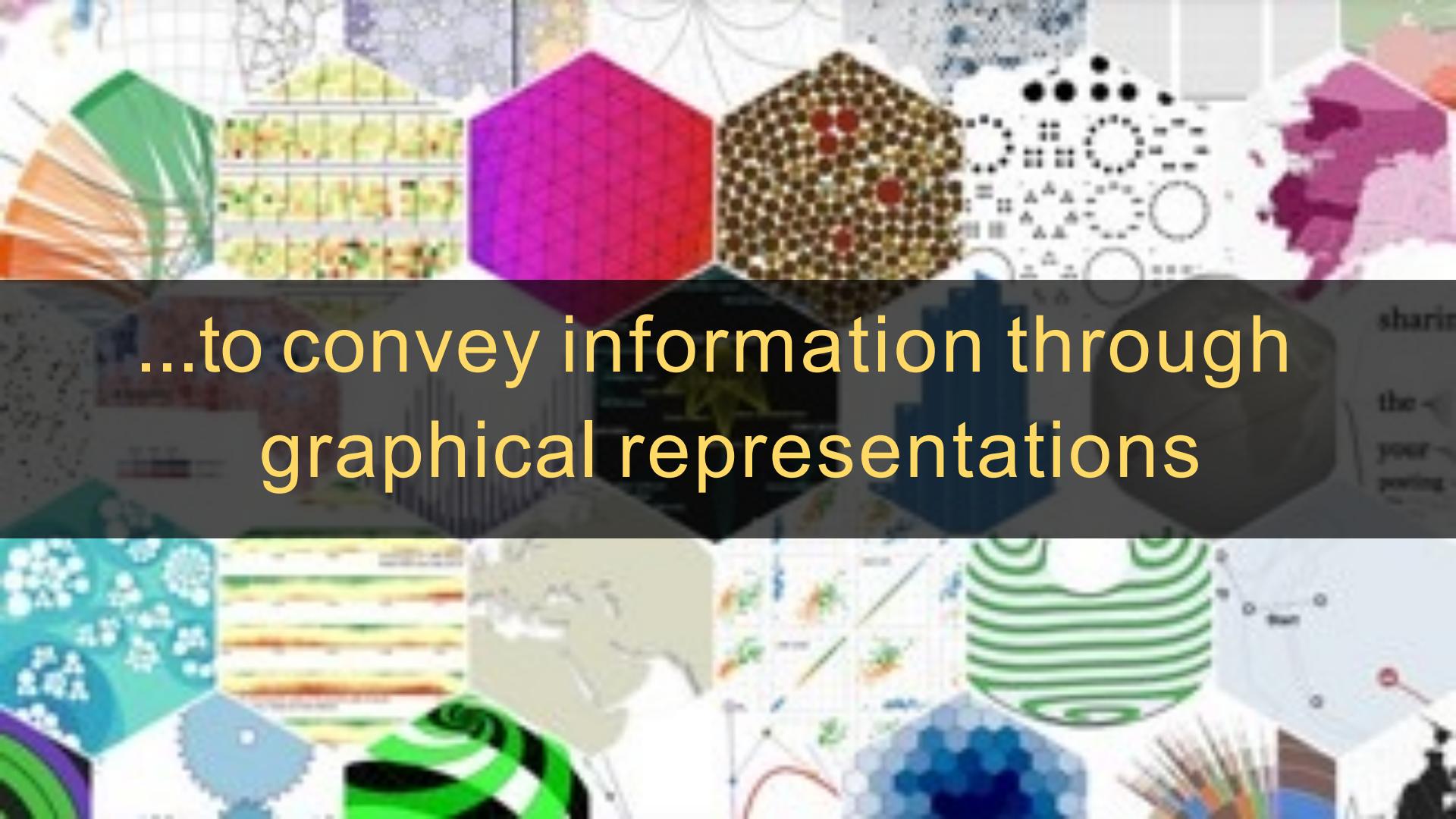
—McCormick et al. 1987

“... finding the **artificial memory** that best supports our natural means of perception.”

—Bertin 1967

“visual representations of data to **amplify cognition**.”

—Card, Mackinlay, & Shneiderman 1999



...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

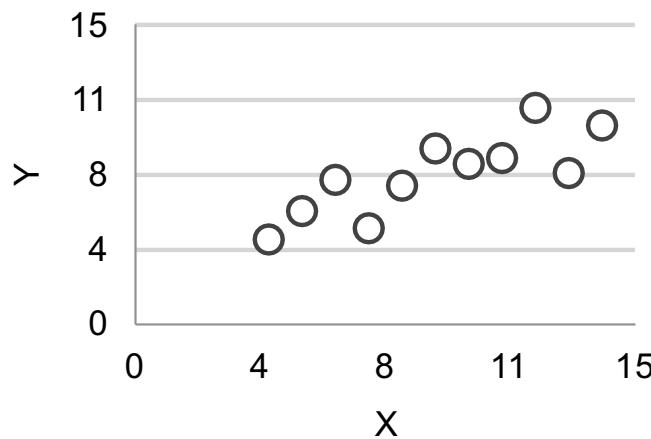
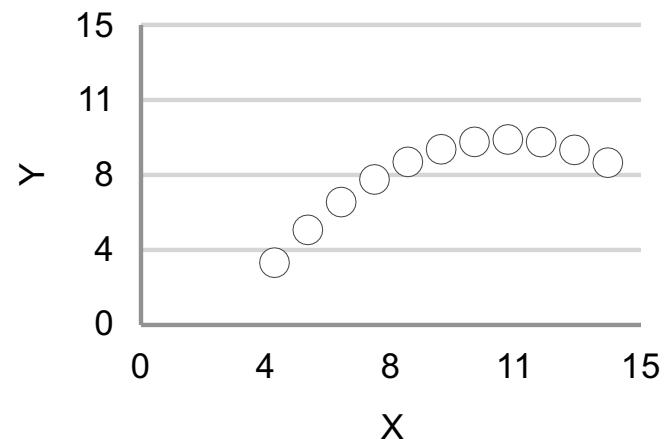
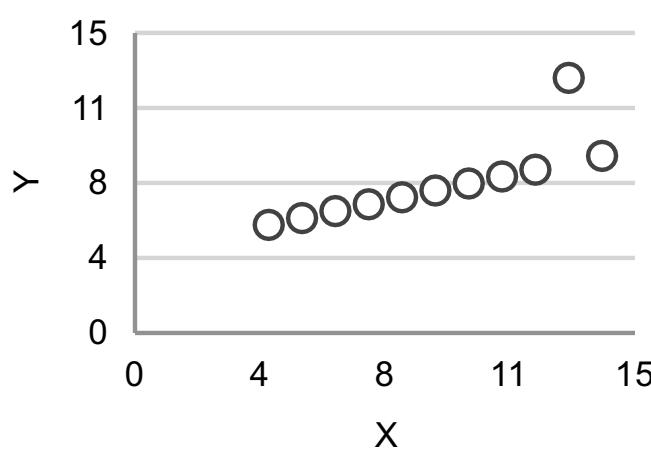
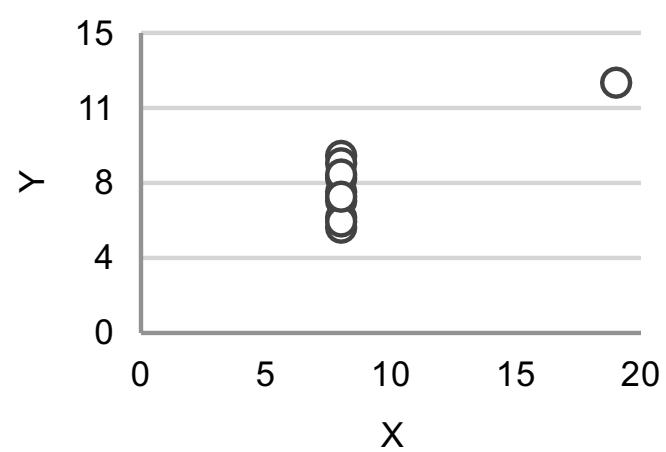
$$\bar{x} = 9.0 \quad \sigma_x = 3.317$$

$$\bar{y} = 7.5 \quad \sigma_y = 2.03$$

Linear Regression

$$Y = 3 + 0.5 X$$

$$R^2 = 0.67$$

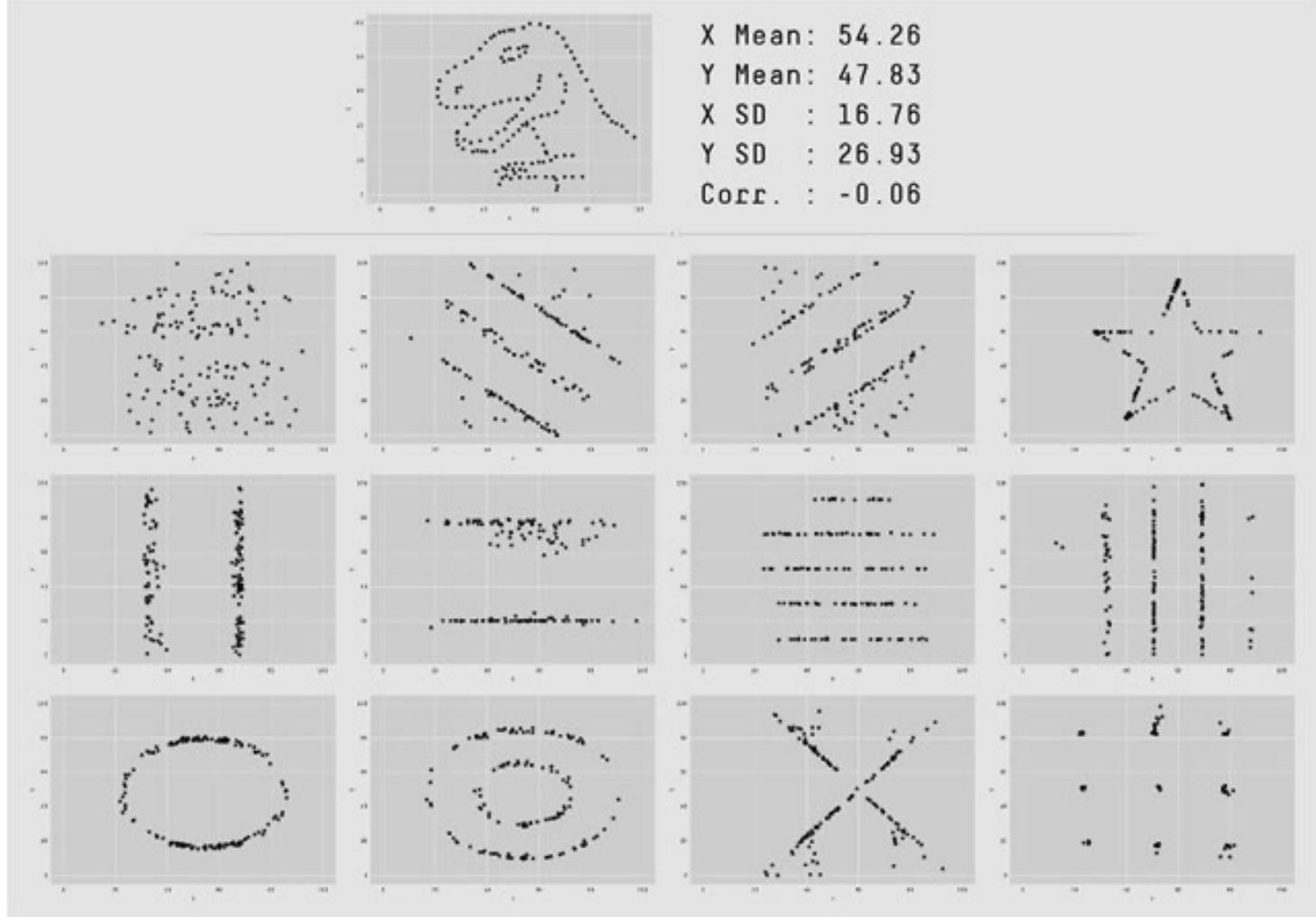
A**B****C****D**

...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

Thuật ngữ

- **Scientific Visualization**

- Chuyên ngành trong khoa học máy tính bao gồm: giao diện người sử dụng, biểu diễn dữ liệu, giải thuật xử lý, kỹ thuật hiển thị,... (McCormick, 1987)

- **Data Visualization**

- Rộng hơn Scientific visualization, trong đó bao hàm cả các phương pháp thống kê và phân tích dữ liệu (Rosenblum, 1994)

- **Information Visualization**

- Shneiderman:

- ~ “sử dụng biểu diễn đồ họa tương tác trên dữ liệu để chắt lọc thông tin và tri thức” (Ware, 2008; Card et al., 1999)

The value of visualization

Why Create Visualizations?

Three functions of visualization

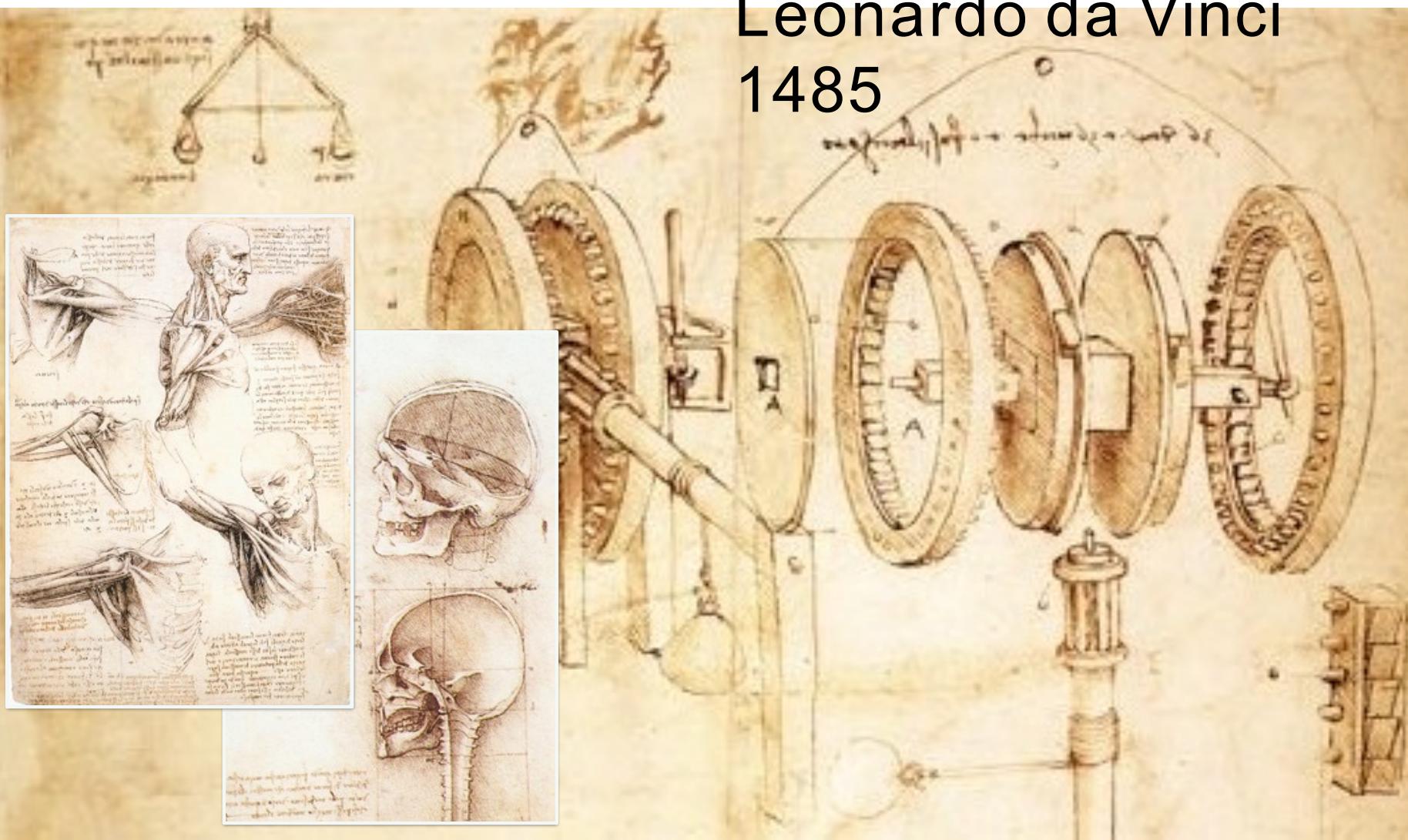
- Record information
- Support reasoning
- Convey Information to Others

Record information: 6200 BC



Leonardo da Vinci

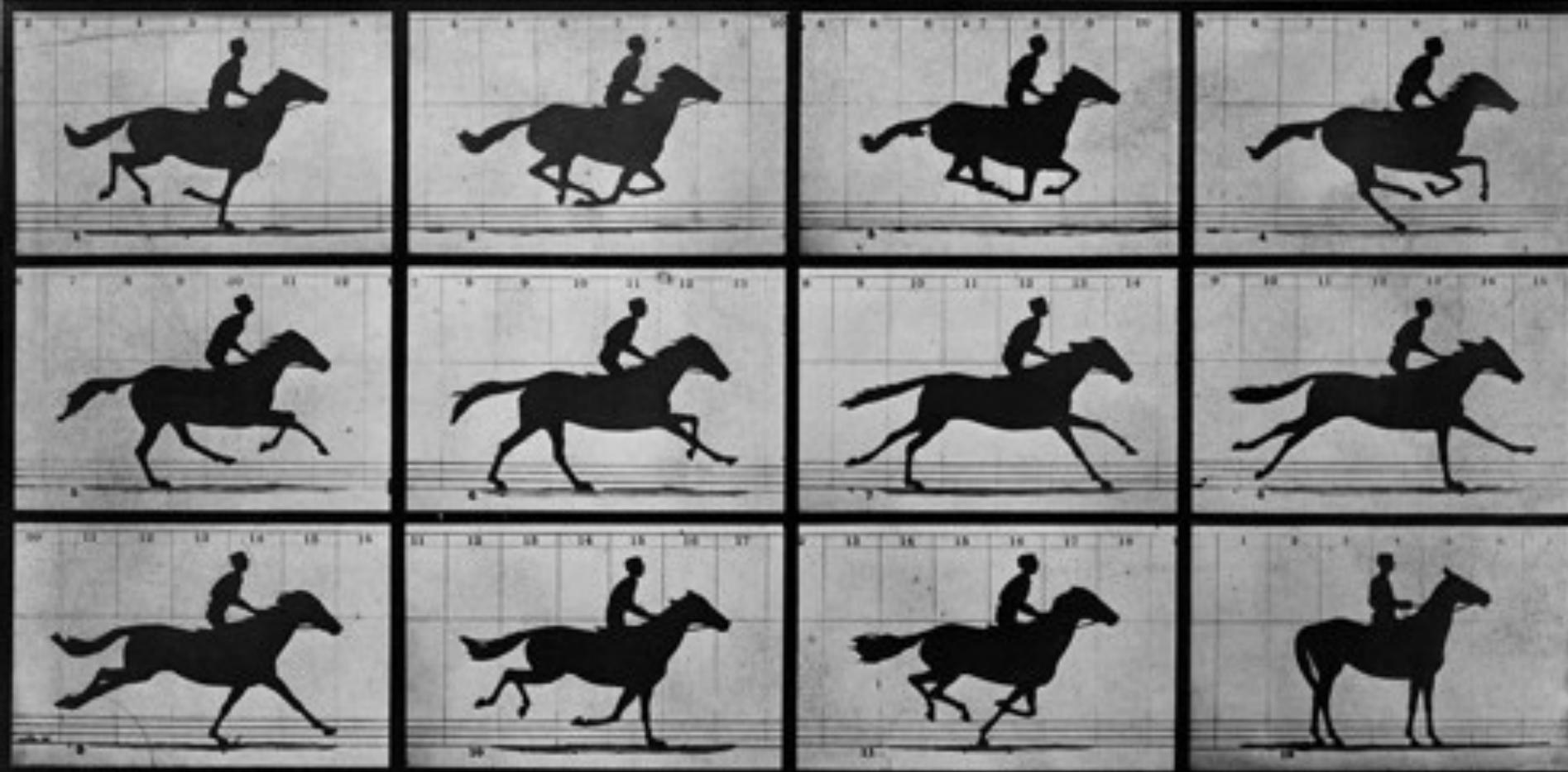
1485



Galileo Galilei's Sketches of the Moon

(November-December 1609)





Copyright, 1878, by MUYBRIDGE.

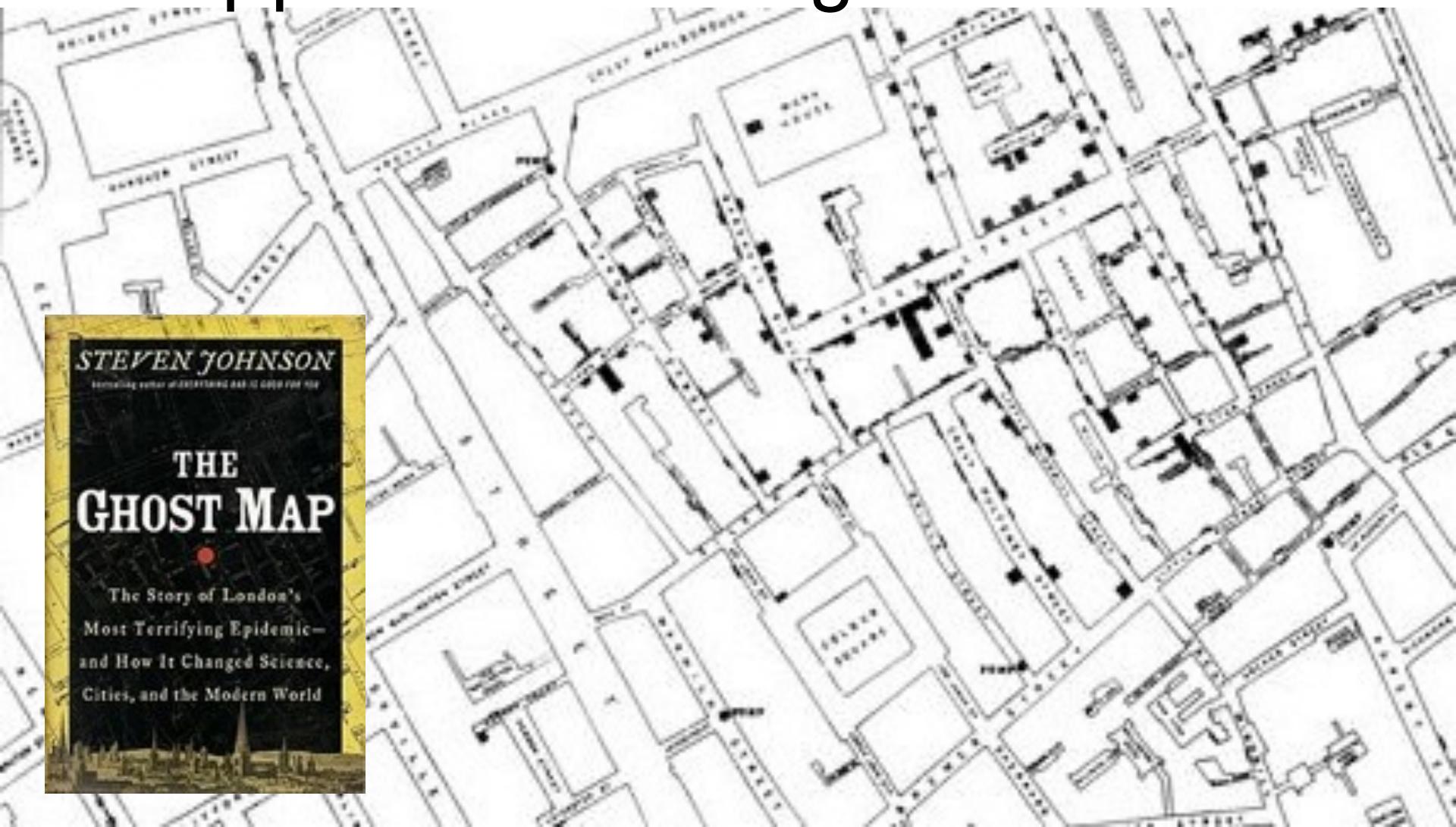
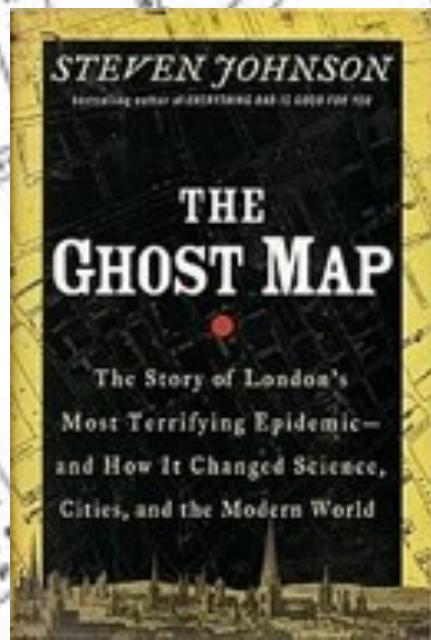
THE HORSE IN MOTION.

Illustrated by

MUYBRIDGE, San Francisco.

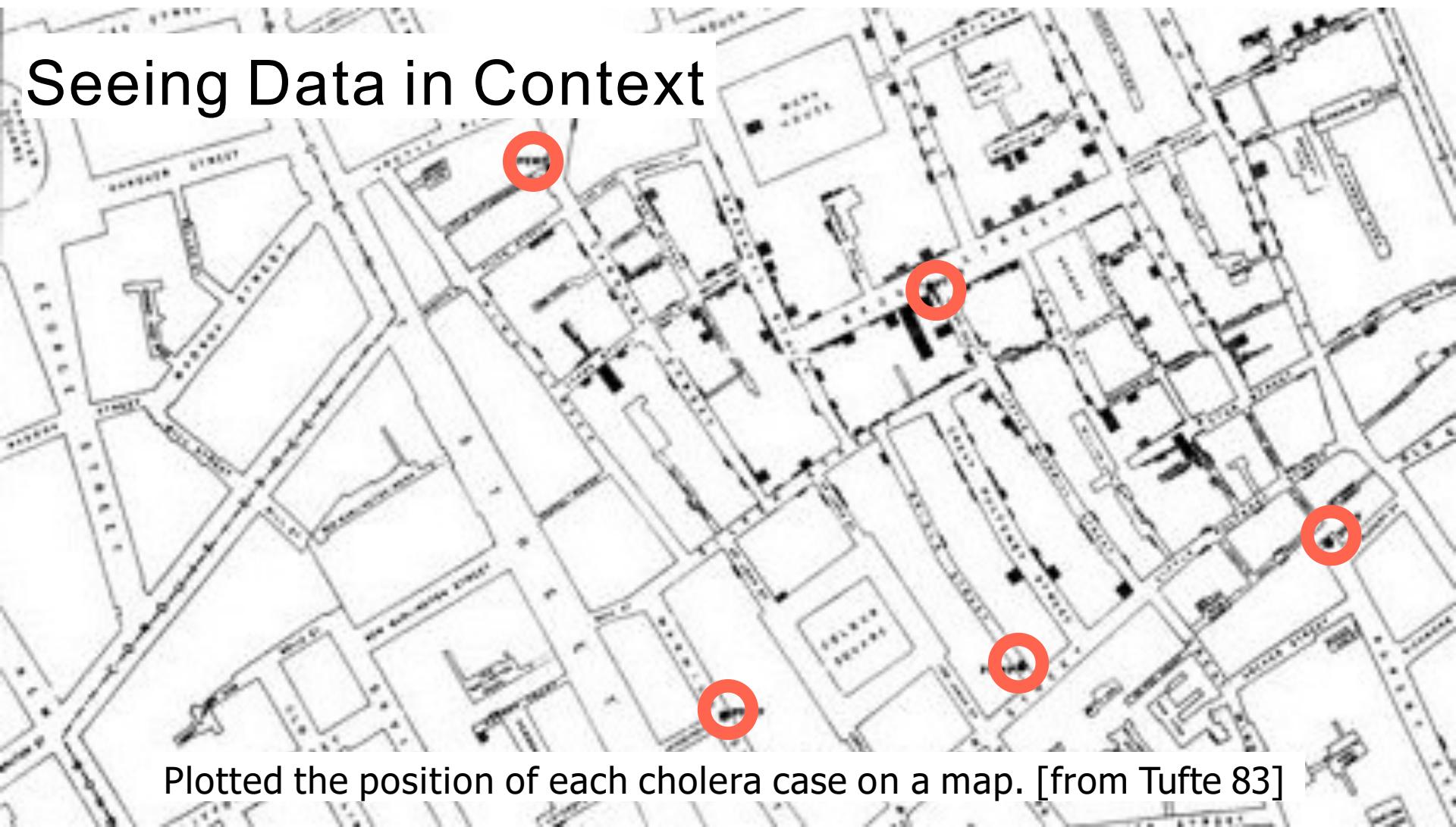
E. J. Muybridge,
1878

Support Reasoning

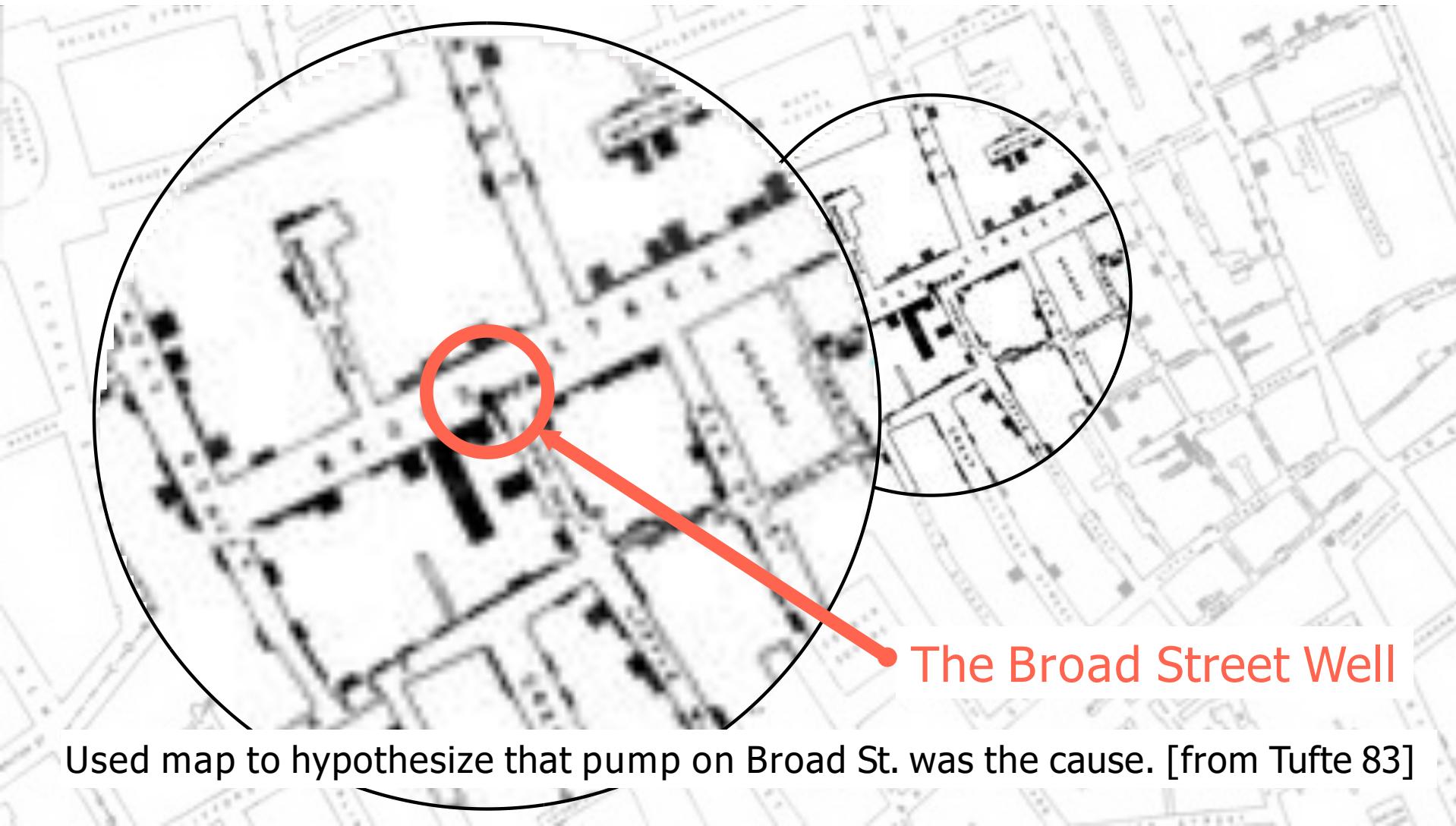


- John Snow, the Cholera Epidemic 1854

Seeing Data in Context

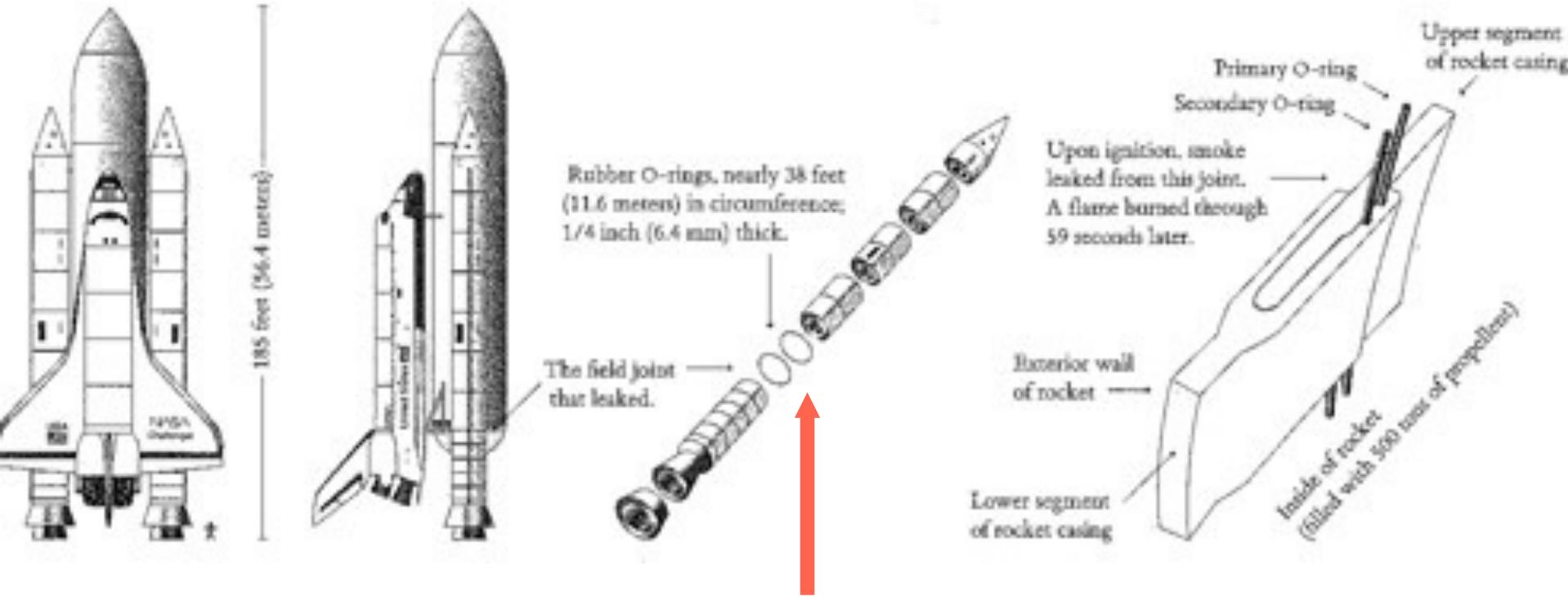


Plotted the position of each cholera case on a map. [from Tufte 83]



Space Shuttle Challenger Disaster (1986)





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			Top View		Clocking Location (deg)
		Erosion Depth (in.)	Perimeter Affected (deg)	Nominal Dia. (in.)	Length Of Max Erosion (in.)	Total Heat Affected Length (in.)	
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.

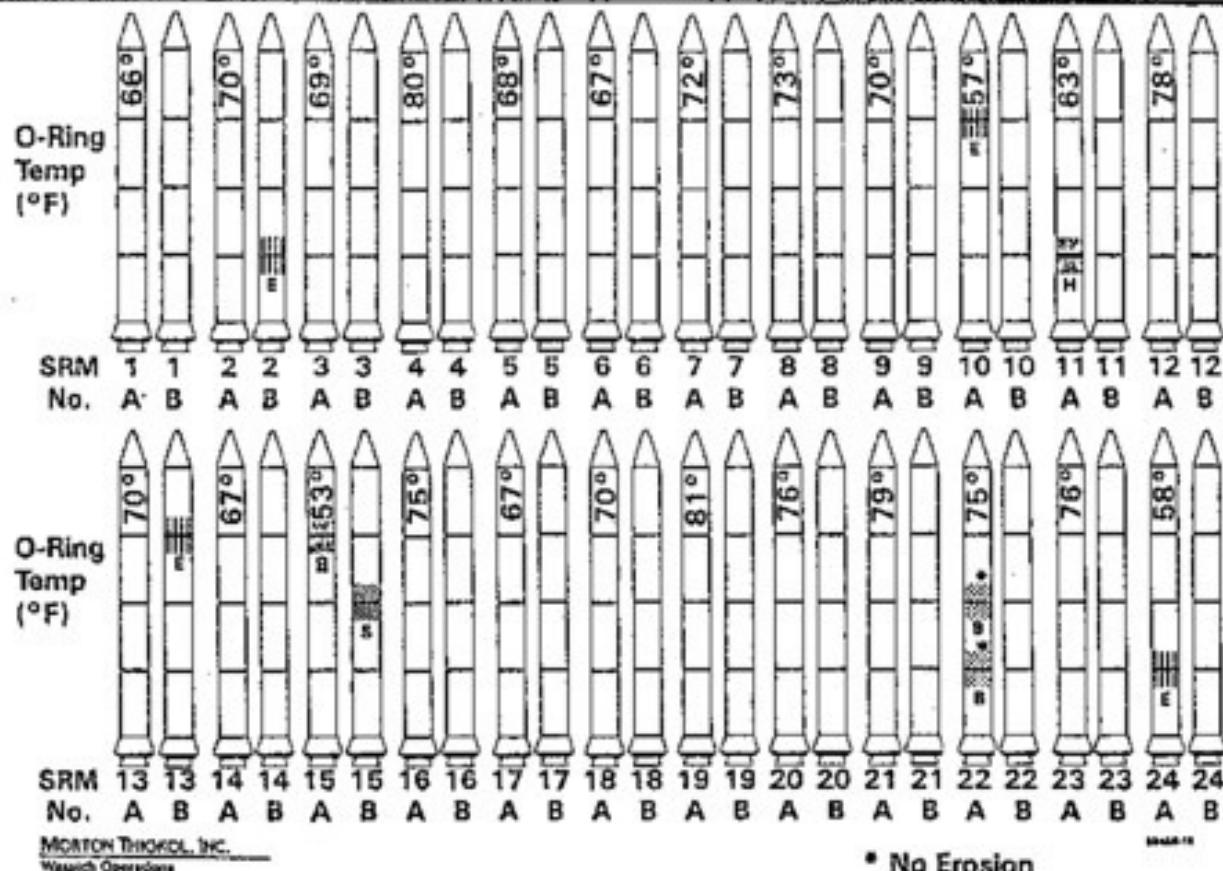
***Soot behind primary O-ring, heat affected secondary O-ring.

Clockwise 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)

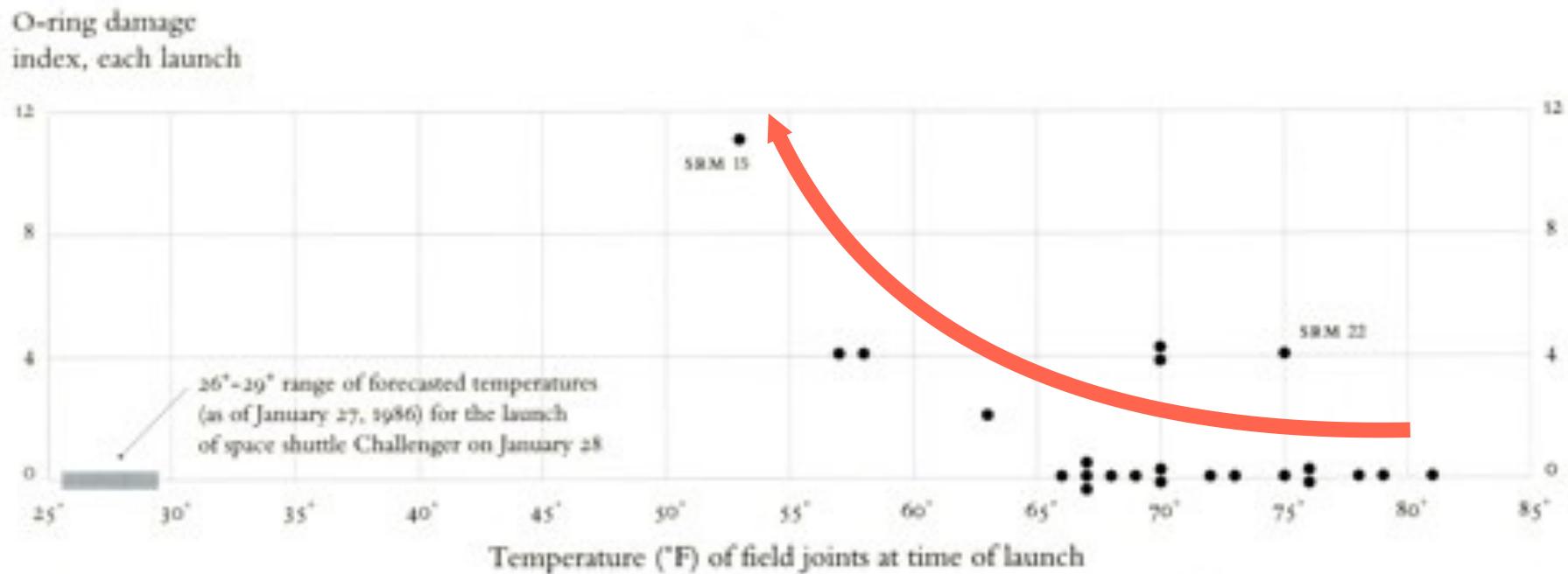


Code	
S	= Heating of Secondary O-Ring
B	= Primary O-Ring Blowby
E	= Primary O-Ring Erosion
H	= Heating of Primary O-Ring
□	= No Damage

STATIC TEST MOTORS

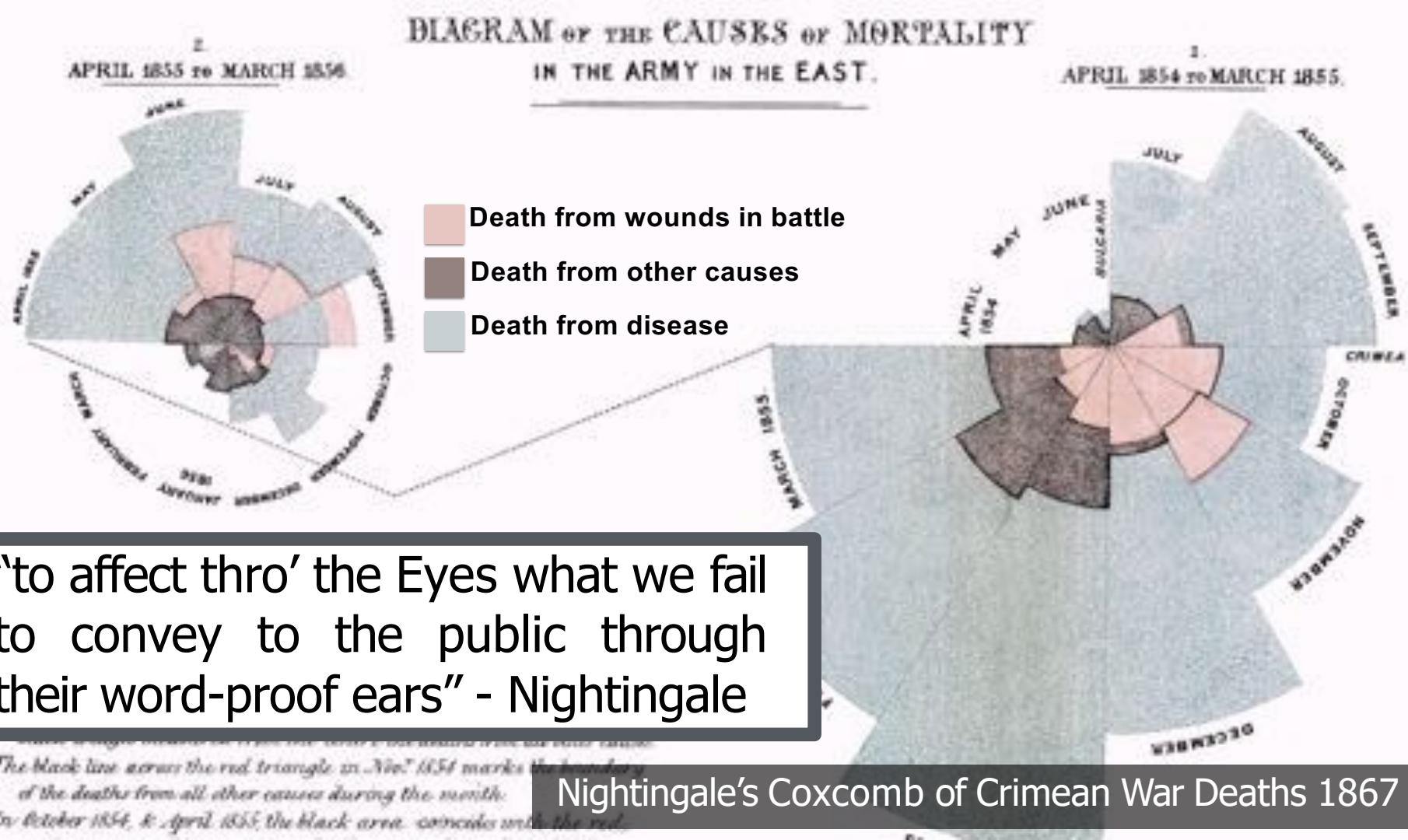
- HORIZONTAL ASSEMBLY
- SOME PUTTY REPAIRED

Use a right visualization to make a right decision

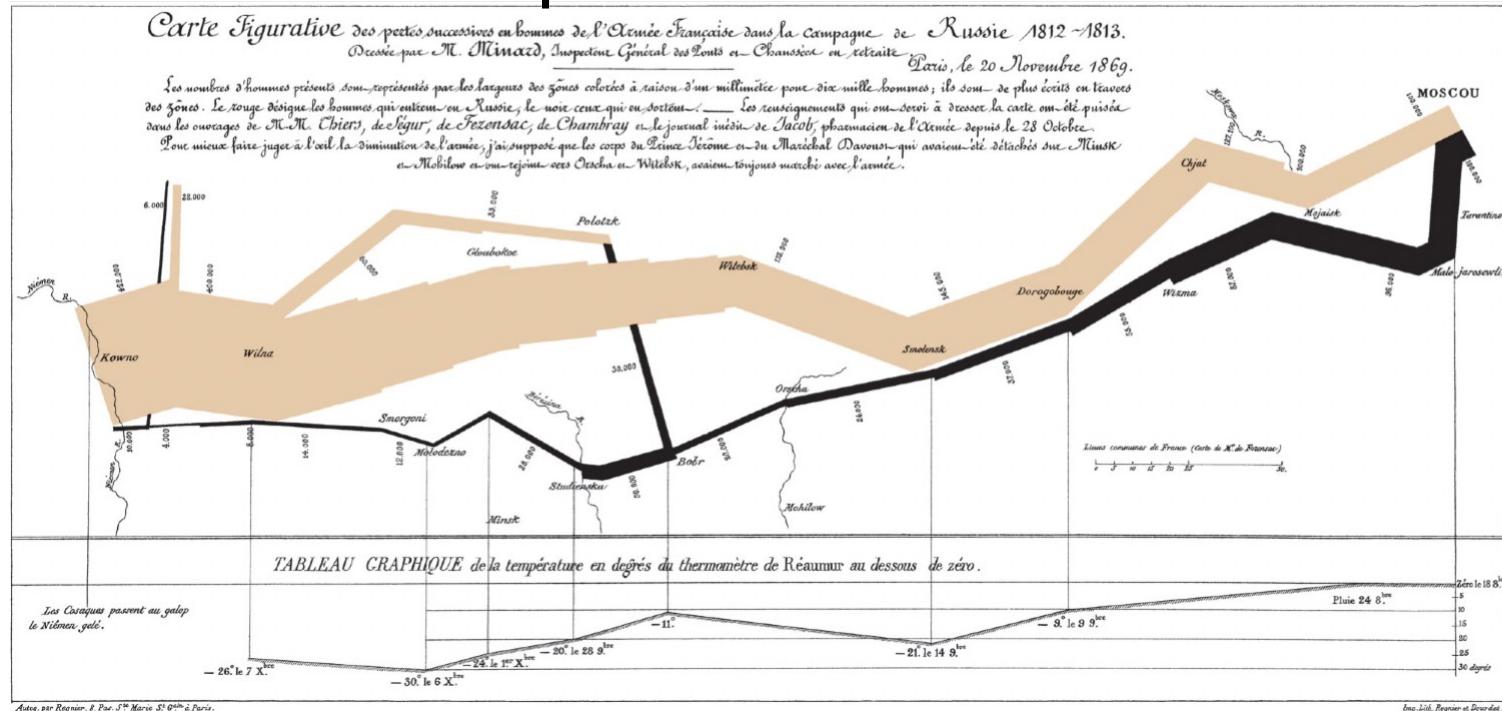


Convey Information to Others

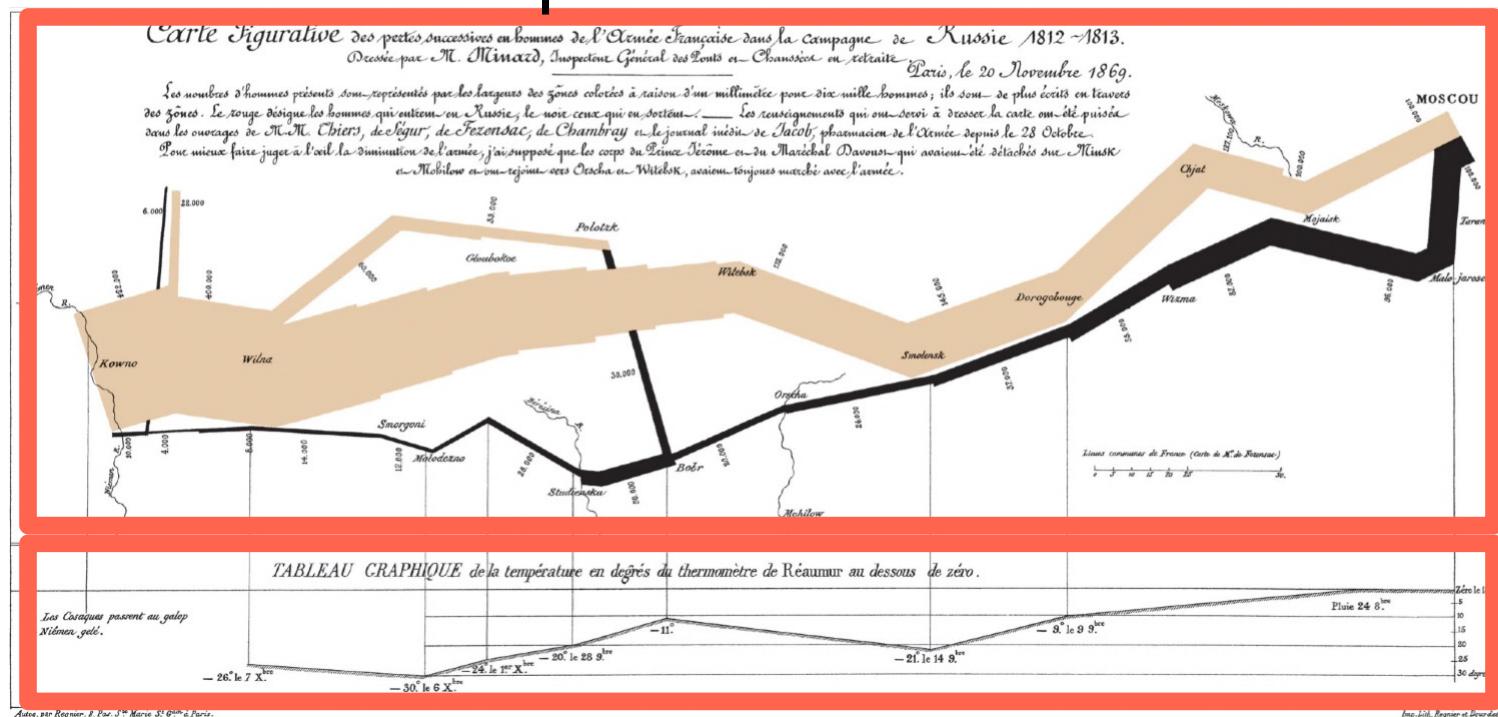
Convey Information to Others



Minard 1869: Napoleon's March

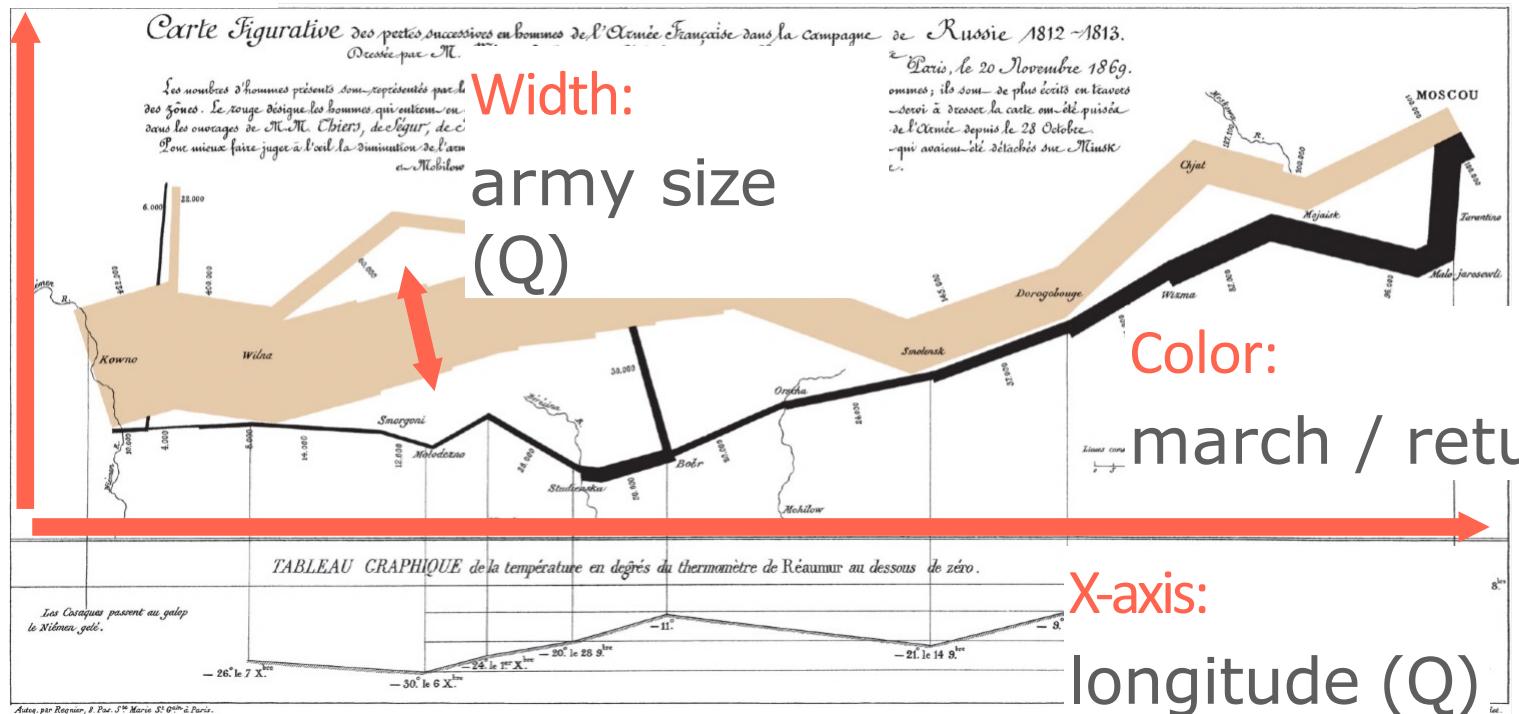


Minard 1869: Napoleon's March



Y-axis:

latitude (Q)



Width:

army size
(Q)

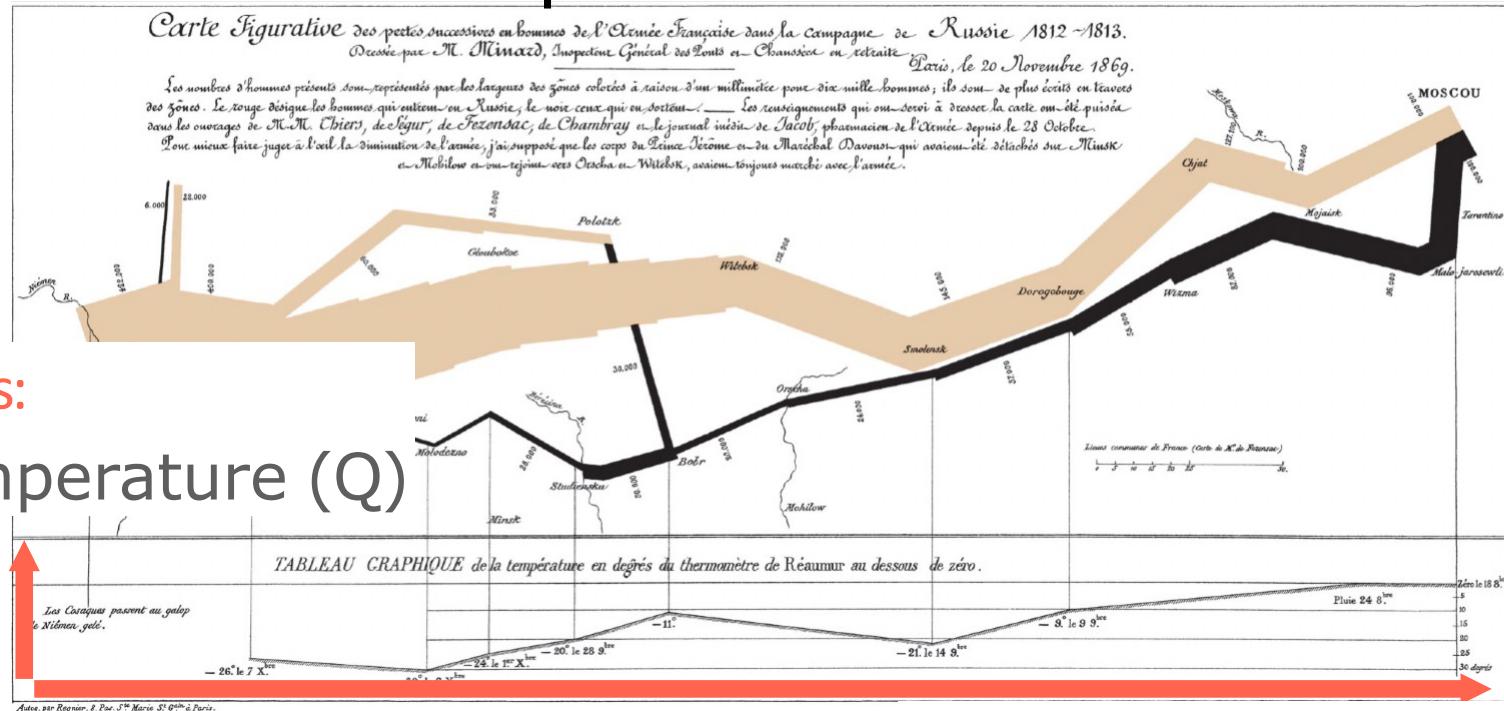
Color:

march / return

X-axis:

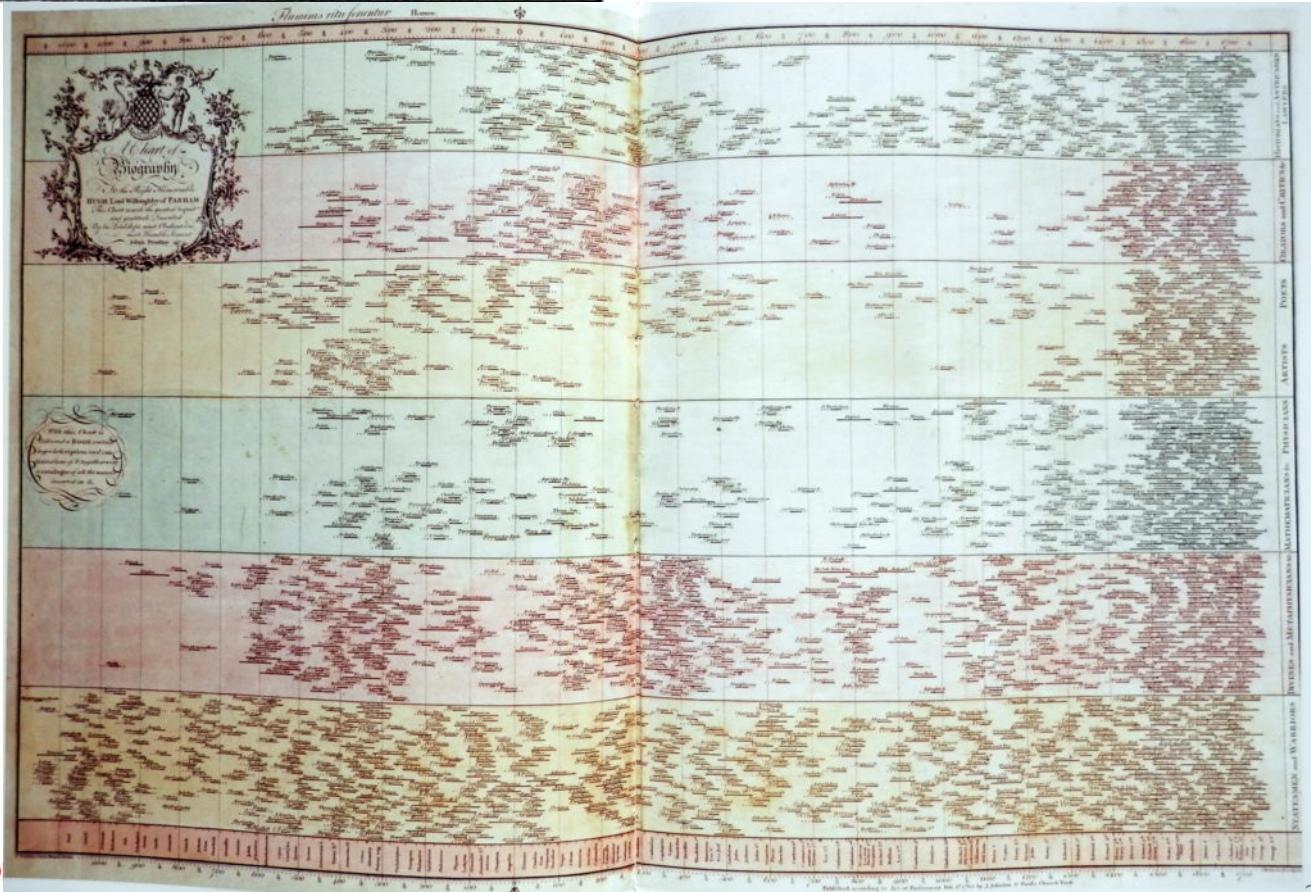
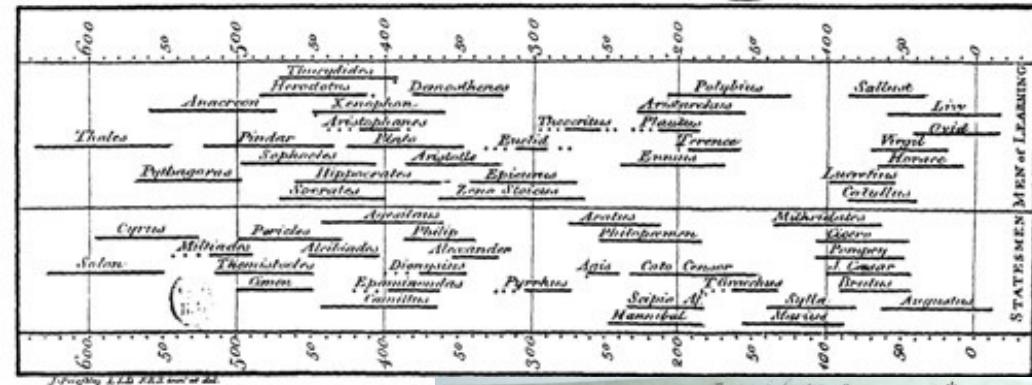
longitude (Q)

Minard 1869: Napoleon's March



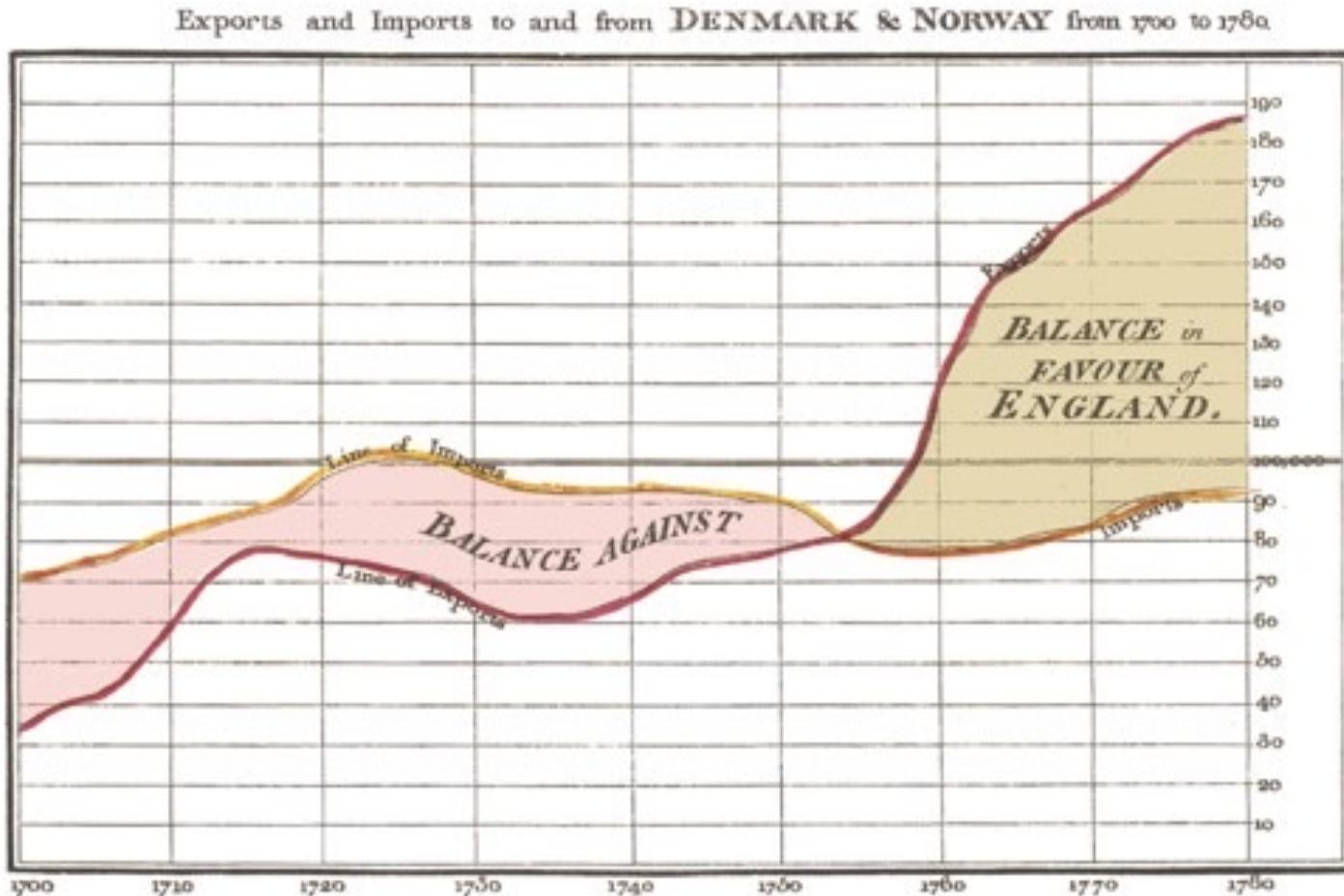
A Specimen of a Chart of Biography.

Joseph Priestley 1765



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

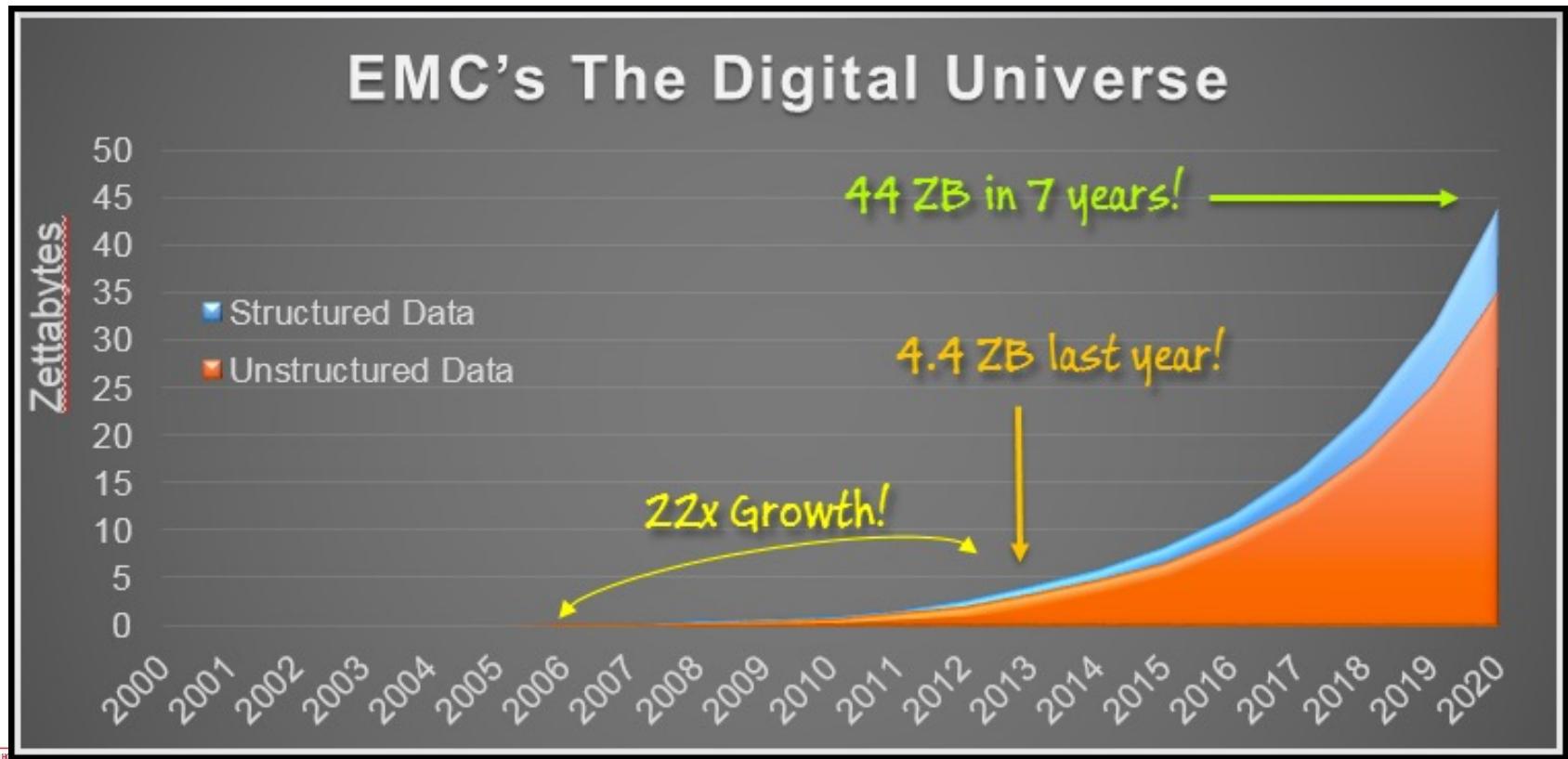
Goals of visualization research

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

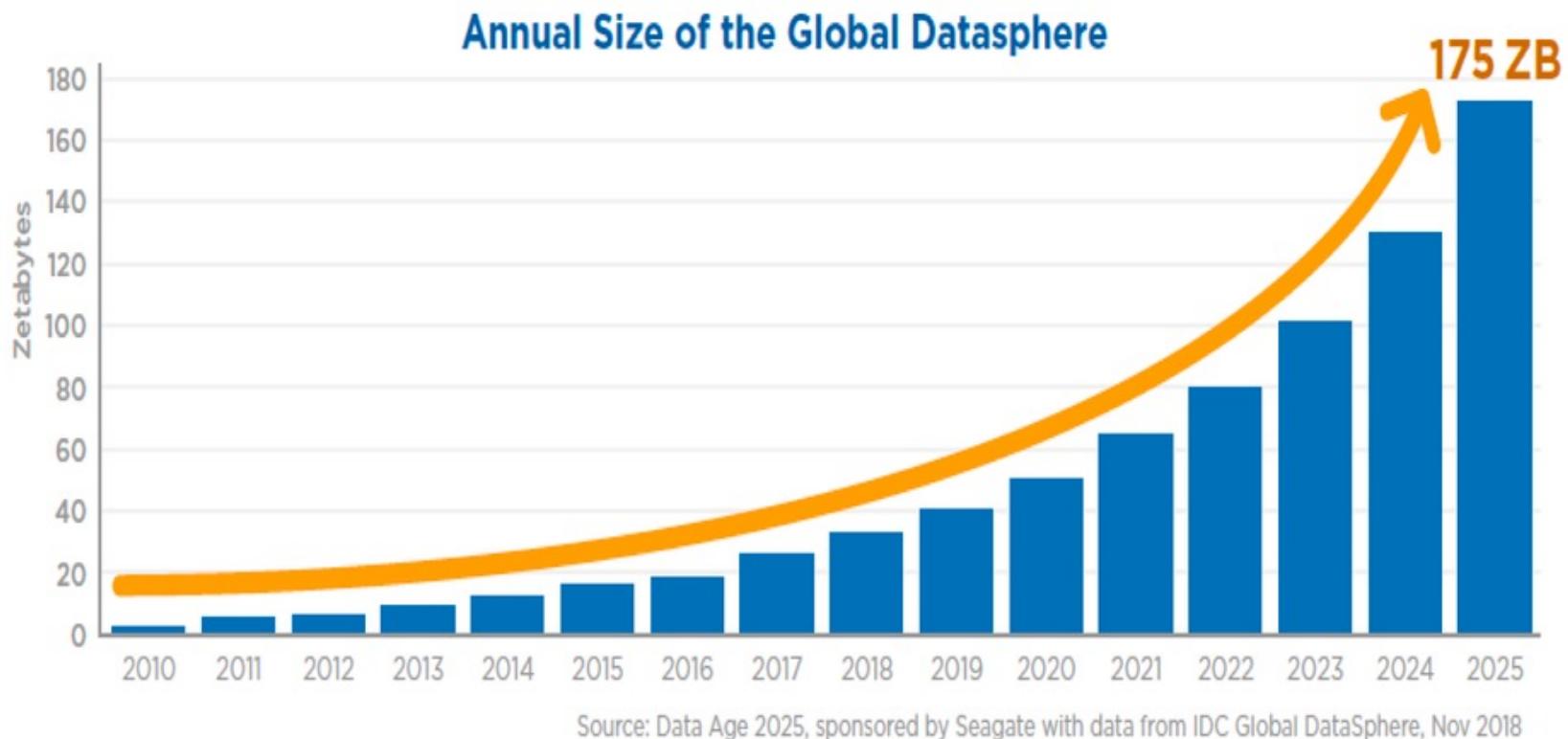
Data visualization in the big data era

The industrial revolution of data

Big data growth



Big data growth to 2025



How big is big data?



Data is everywhere

Coronavirus Company Message

SAS Coronavirus Report - Powered by SAS® Viya®

Back to Summary

Global Status

Location Analysis

Epidemiological Analysis

Trend Analysis

Spread over time

Collective Insights



2019-20 Novel Coronavirus Outbreak

Global Cases and Analysis of SARS-CoV-2

Total Confirmed Cases

6,166,946

Total Deaths

372,035

Case Fatality Rate

6.03%

Mortality Rate

0.005%

More details by each location available [here](#)

Daily % change in confirmed cases (global)



Sas

Case Fatality Rate (CFR) is a crude indicator defined as the ratio of number of deaths to the number of reported confirmed cases for SARS-CoV-2. Mortality Rate is the ratio of number of deaths to the total country population (2019). World population at 7.72B. Prevalence - Total confirmed cases / country population (2019), represented per 100,000 people

Cases as of:

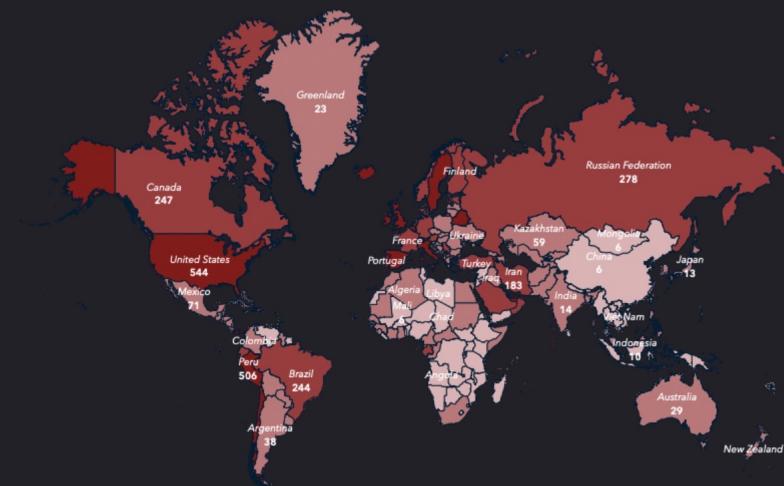
May 31, 2020



SAS Viya

In early December 2019, a new coronavirus, designated **SARS-CoV-2**, was identified in Wuhan, China. The illness from the outbreak, termed COVID-19, on March 11, 2020, has now been declared as a global pandemic by [World Health Organization](#).

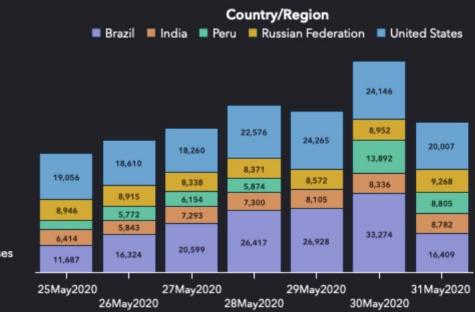
Total confirmed cases of COVID-19 per 100k population by country



Overall outbreak by country (Top 10)

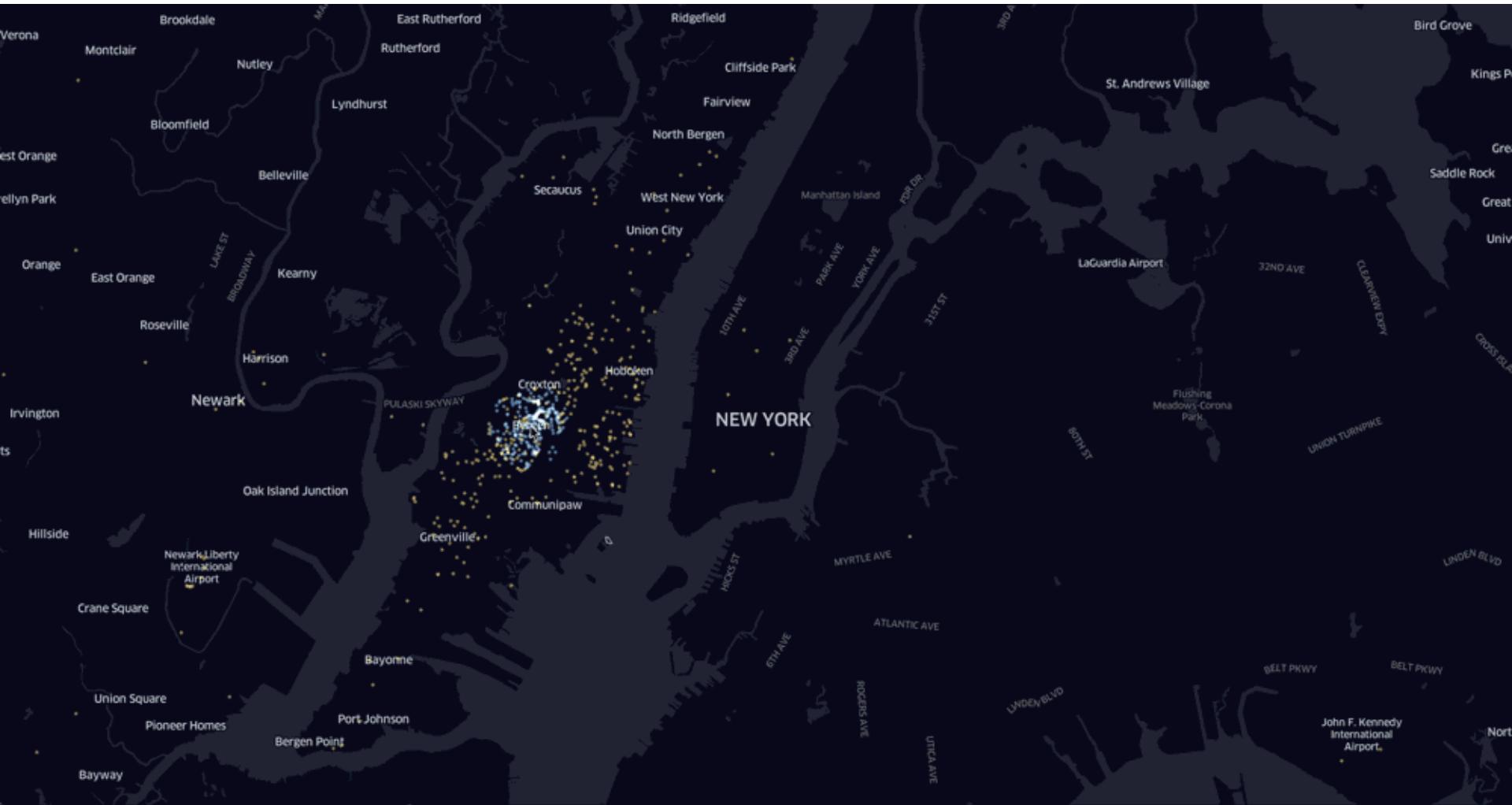
Country/Region	Total Confirmed Cases	Prevalence (/100k)	Total Deaths
United States	1.8M	544	104K
Brazil	515K	244	29K
Russian Federation	406K	278	4.7K
United Kingdom	276K	409	39K
Spain	239K	512	27K
Italy	233K	385	33K
India	191K	14	5.4K
France	189K	290	29K
Germany	183K	220	8.5K
Peru	164K	506	4.5K
All Other	2M	21,587	87K

New Cases - Top 5 impacted countries (last 7 days)



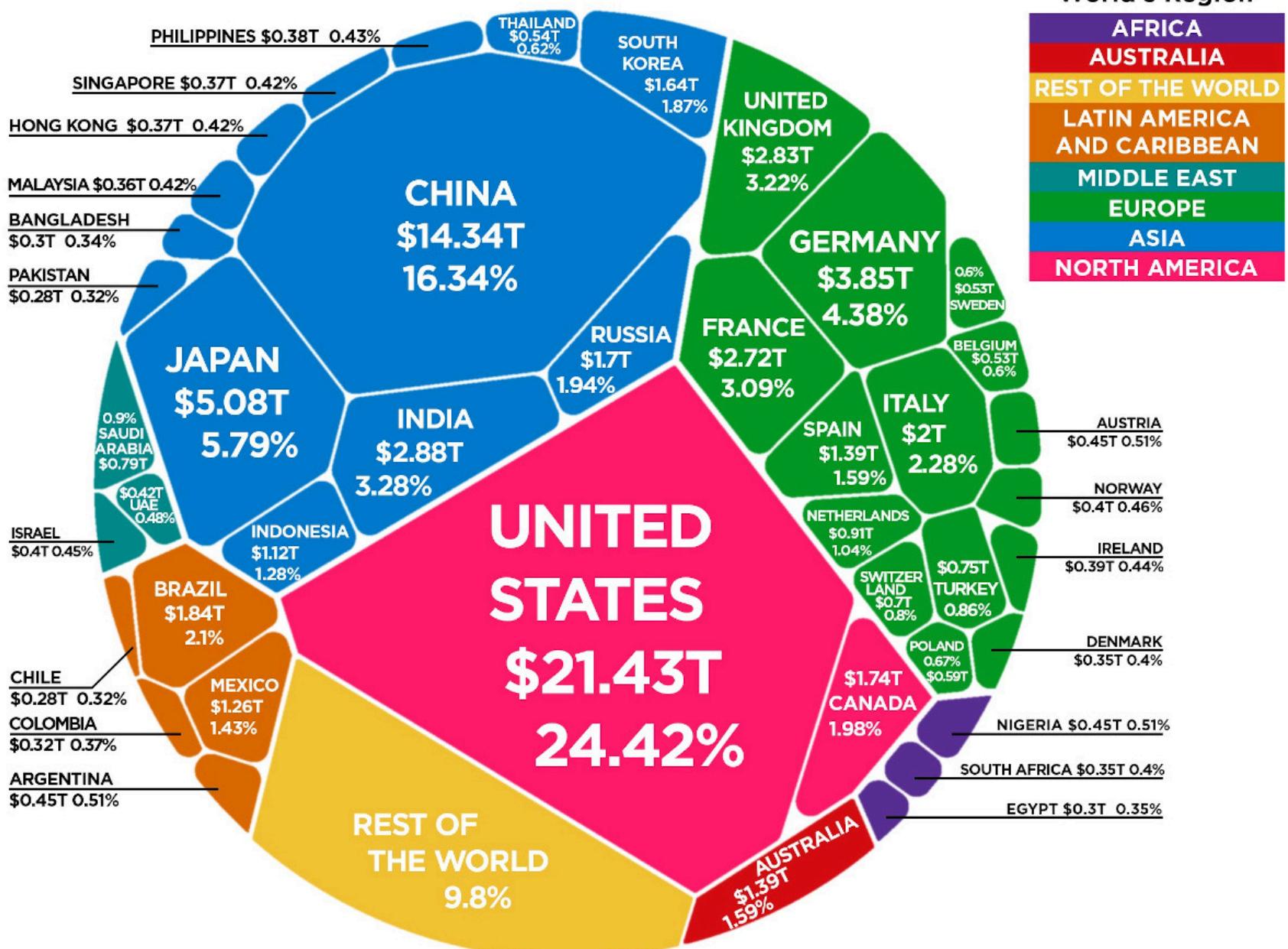
Source, Disclaimer and Data Information

Uber traffic visualization







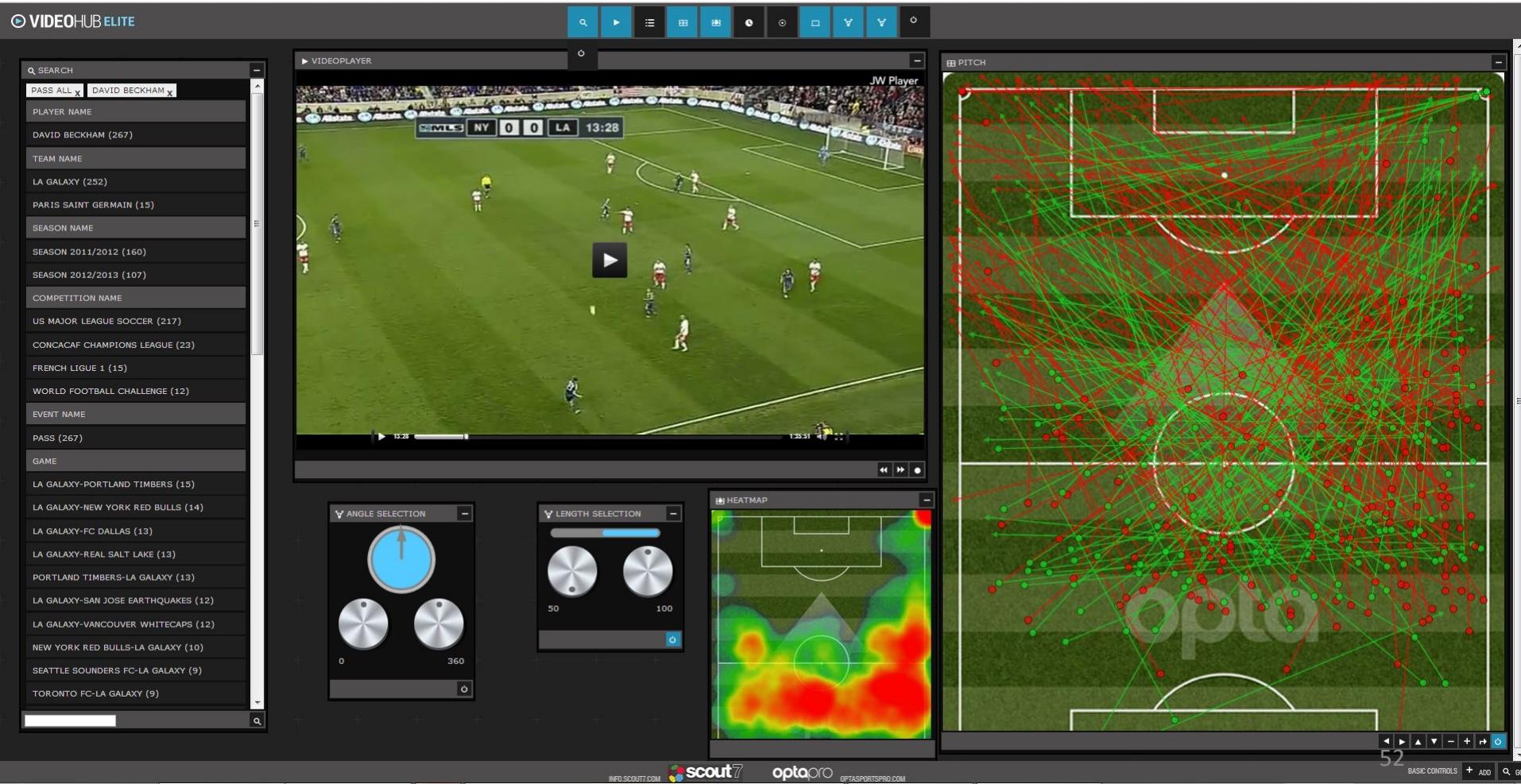


Article & Sources:

<https://howmuch.net/articles/the-world-economy-2019>
<https://databank.worldbank.org>

Football performance analysis

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



THE AIR WE BREATHE

Healthy - Less than 12.1 ug/m³

Moderate - 12.1 ug/m³ to 35.4 ug/m³

Unhealthy - Greater than 35.4 ug/m³

This visualization examines the global prevalence 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.



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?

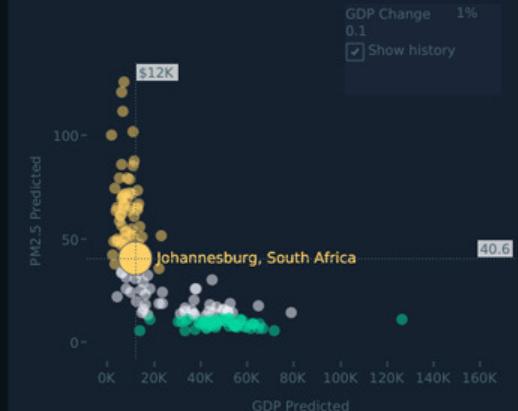
Poor Air Quality is a multi-dimensional problem, and there are many factors that play a part in the state of a localities air quality, however one of the major factors is GDP Per Capita.

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%.



Go To Sources

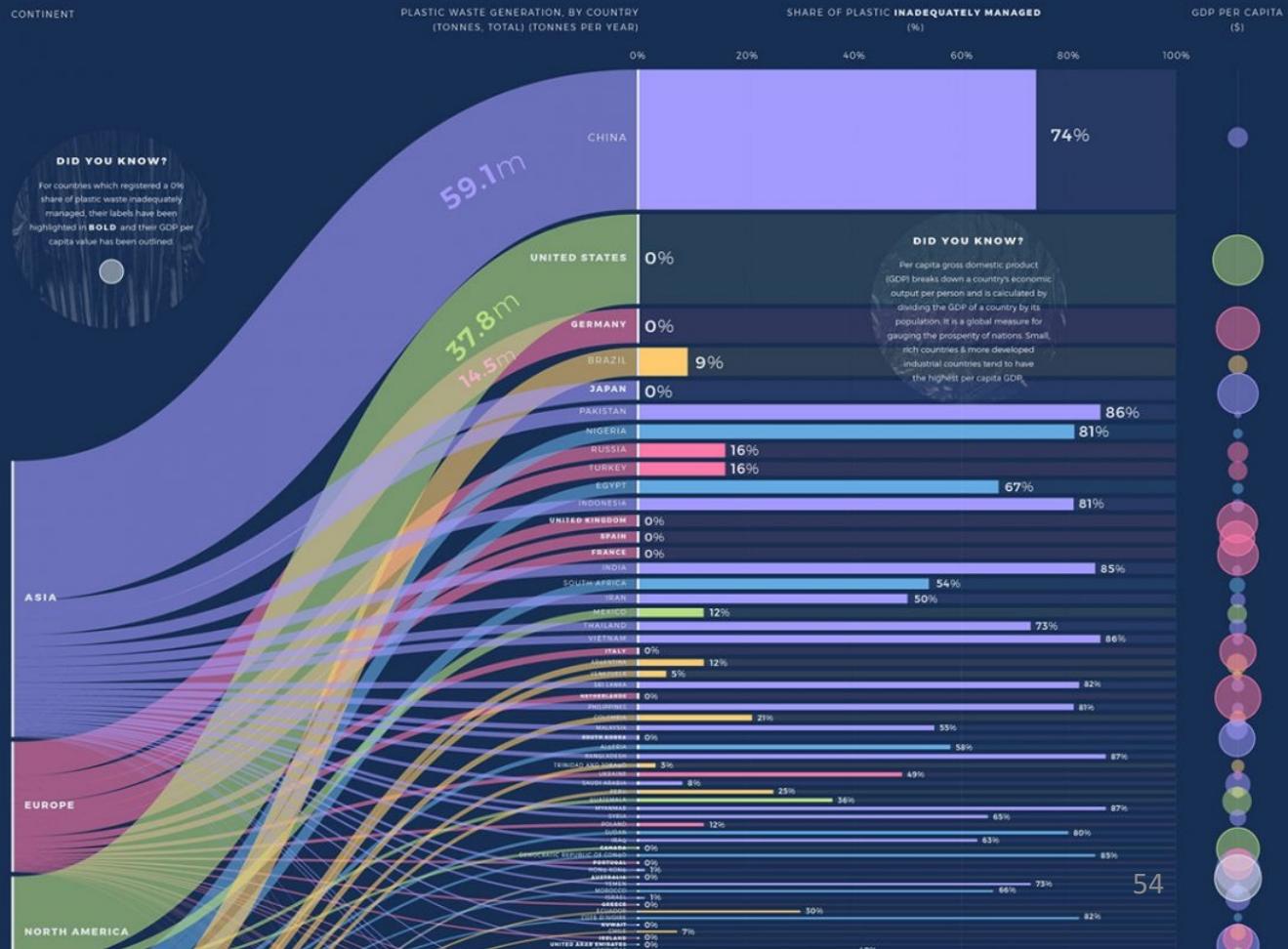
WHO IS BOTTLING PLASTIC WASTE POLLUTION?

This visualisation explores which countries in 2010 produce the largest amount of plastic waste and what percentage of this is inadequately managed, but more importantly how does this management correlate to the nation's GDP per capita?

On the left side, we see the distribution of total plastic waste generation by continent leading into segregation by country, measured in ascending order of tonnes per year. This takes account of per capita waste generation and population size.

This leads into the estimated total percentage of this waste that will be inadequately disposed. This includes disposal in dumps or open, uncontrolled landfills; this means the material is not fully contained and can be lost to the surrounding environment. This makes it at risk of leakage and transport to the natural environment and oceans via waterways, winds and tides.

Finally, these findings are presented with the caveat of the individual country's GDP per capita, which provides a basic insight into national wealth and infrastructure, both of which could have a significant bearing on waste management.



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

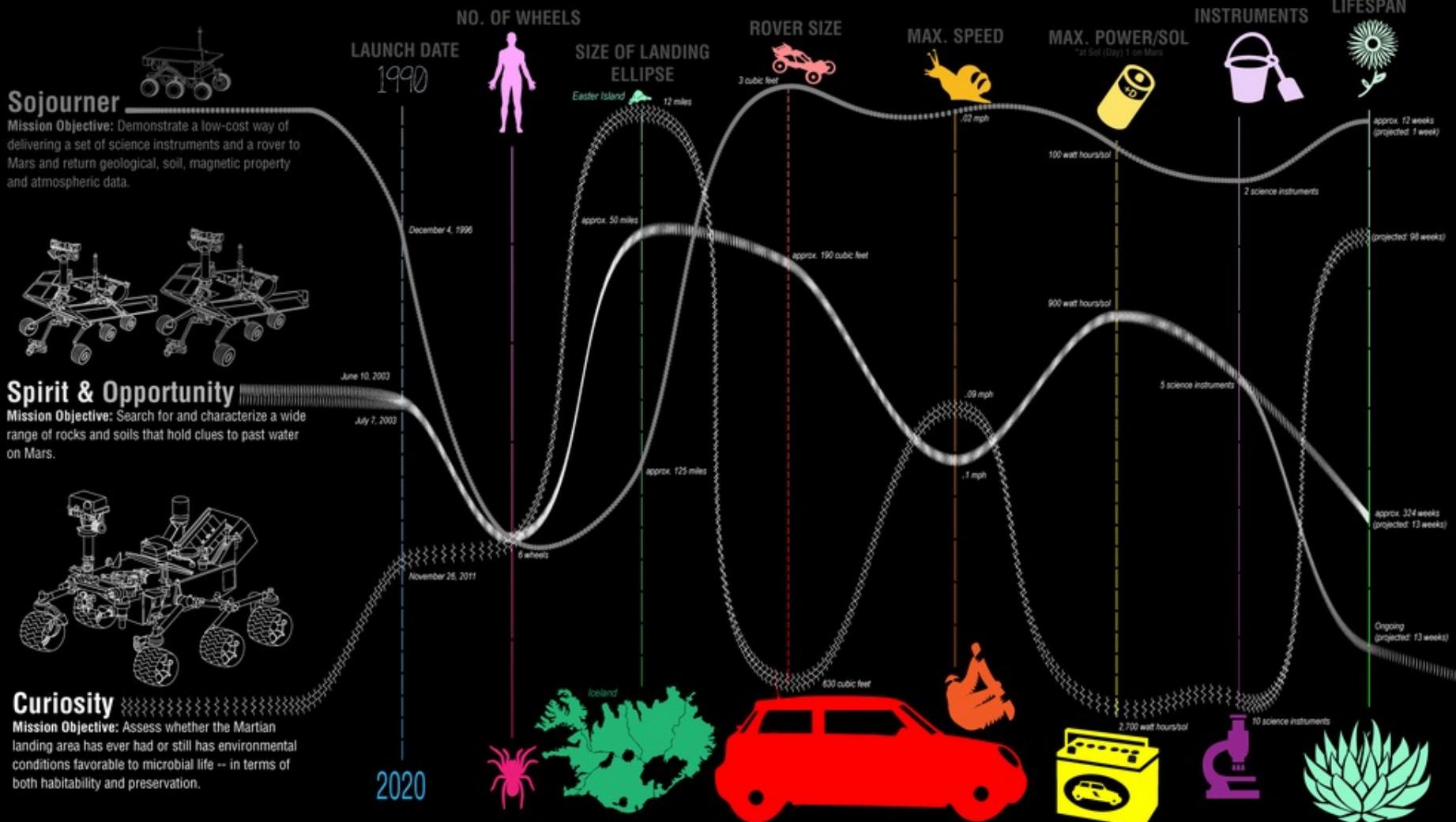


VIỆN CÔNG NGHỆ

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-worldly machines.



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

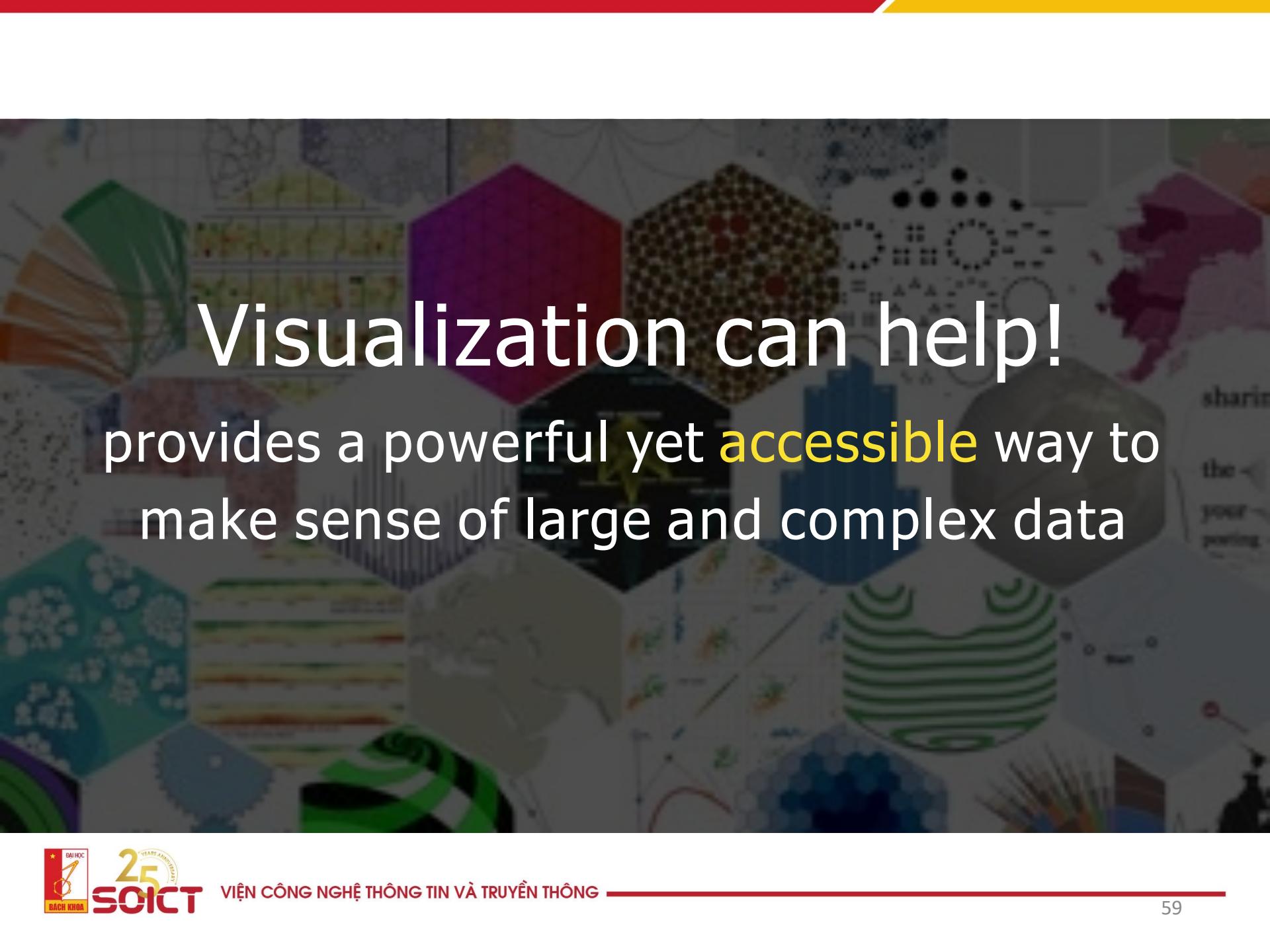


A Poverty of Attention

“...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