DATA SCIENCE AND VISUALIZATION

Trần Thanh Tùng
tttung@hcmiu.edu.vn

MOTIVATION

The purpose of computing is insight, not numbers.

…. of **visualization** is insight, not **pictures**

DEFINITION

Visualization

1. Formation of mental visual images

2. The act or process of interpreting in visual terms or of putting into visible form

DEFINITION

Visualization is the process that

Transforms (abstract) **data** into

interactive graphical representation for the purpose of

exploration, confirmation, or presentation.

GOOD DATA VISUALIZATION

Makes data **accessible**

Combines strengths of **humans** and **computers**

Enables **insight**

communicates

VISUALIZATION

"Visualization is really about external cognition, that is, how resources outside the mind can be used to boost the cognitive capabilities of the mind."

Stuart Card

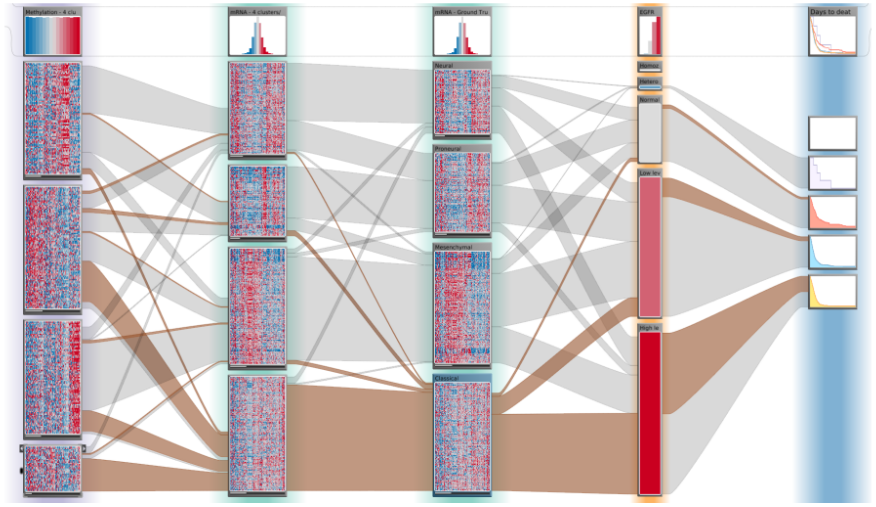
one of the pioneers of applying human factors in human–computer interaction

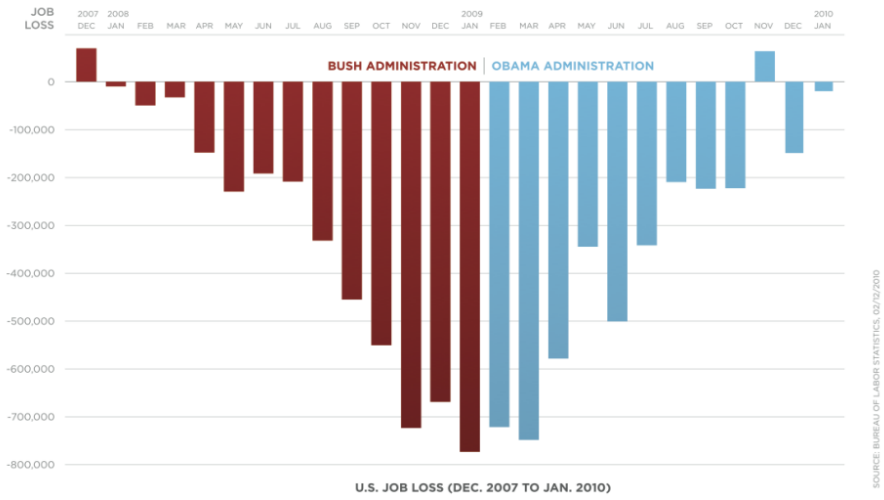
WHY VISUALIZE?

- To inform humans: communication

- When questions are not well defined: exploration
  - *What is the structure of positive comments in social network?*
  - *What is the movement pattern of customer in a supermarket?*
  - *Which drug can help patient X ?*

PURPOSE OF VISUALIZATION





Open Exploration — Confirmation — Communication

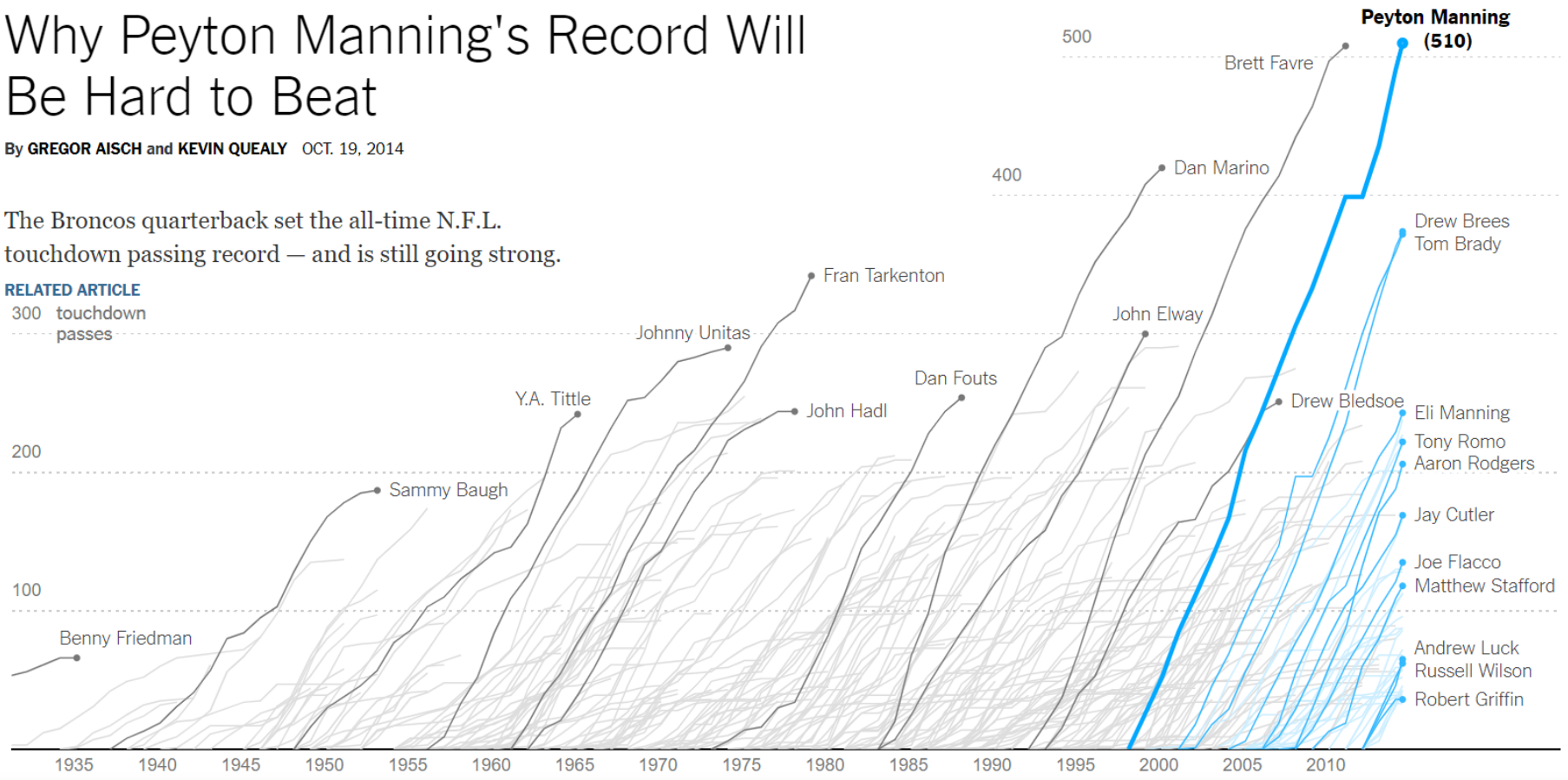
COMMUNICATION

Why Peyton Manning's Record Will Be Hard to Beat

By **GREGOR AISCH** and **KEVIN QUEALY** OCT. 19, 2014

The Broncos quarterback set the all-time N.F.L. touchdown passing record — and is still going strong.

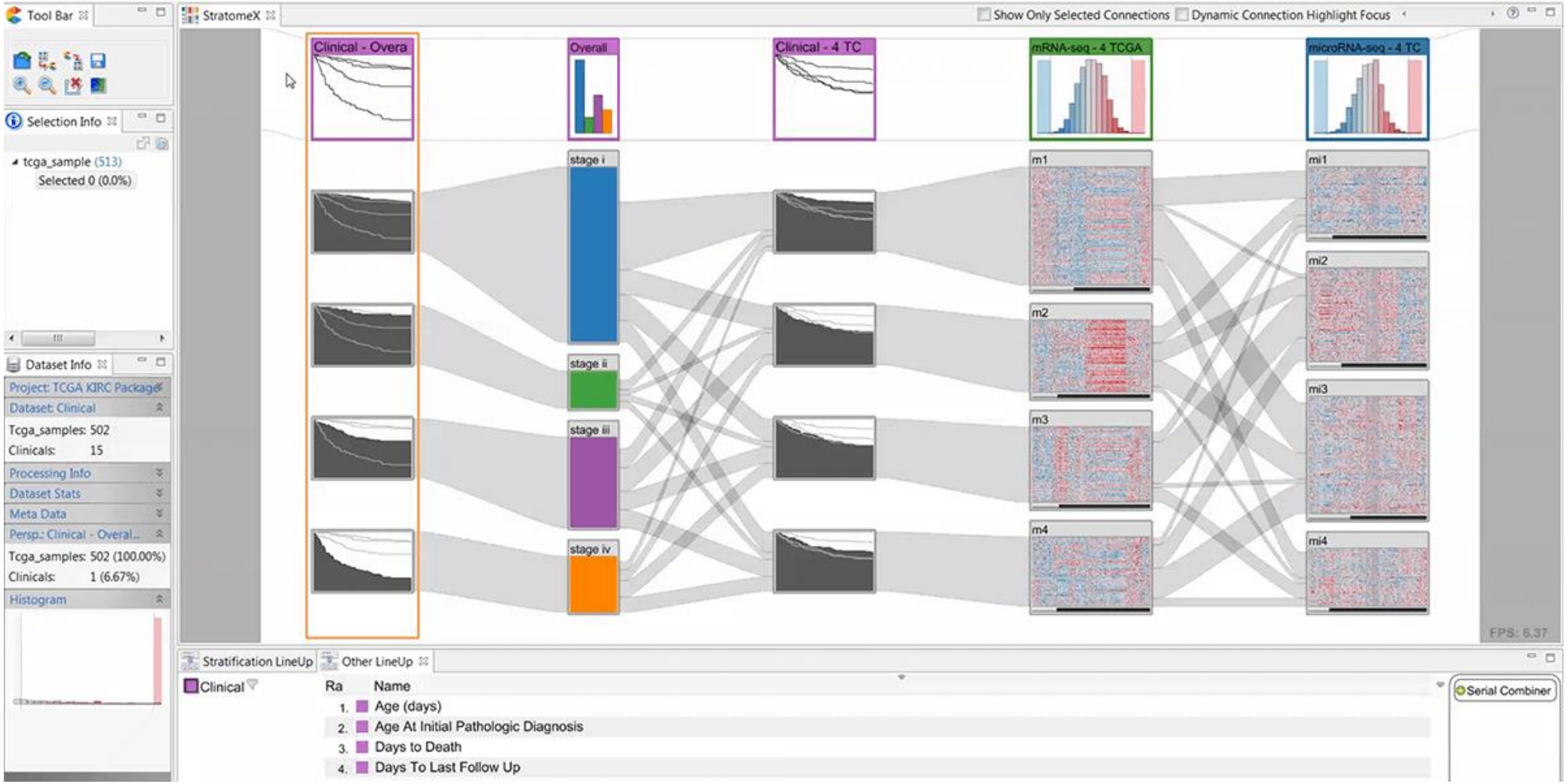
RELATED ARTICLE



https://www.nytimes.com/interactive/2014/10/19/upshot/peyton-manning-breaks-touchdown-passing-record.html?abt=0002&abg=1&smid=tw-upshotnyt

EXAMPLE OF EXPLORATION

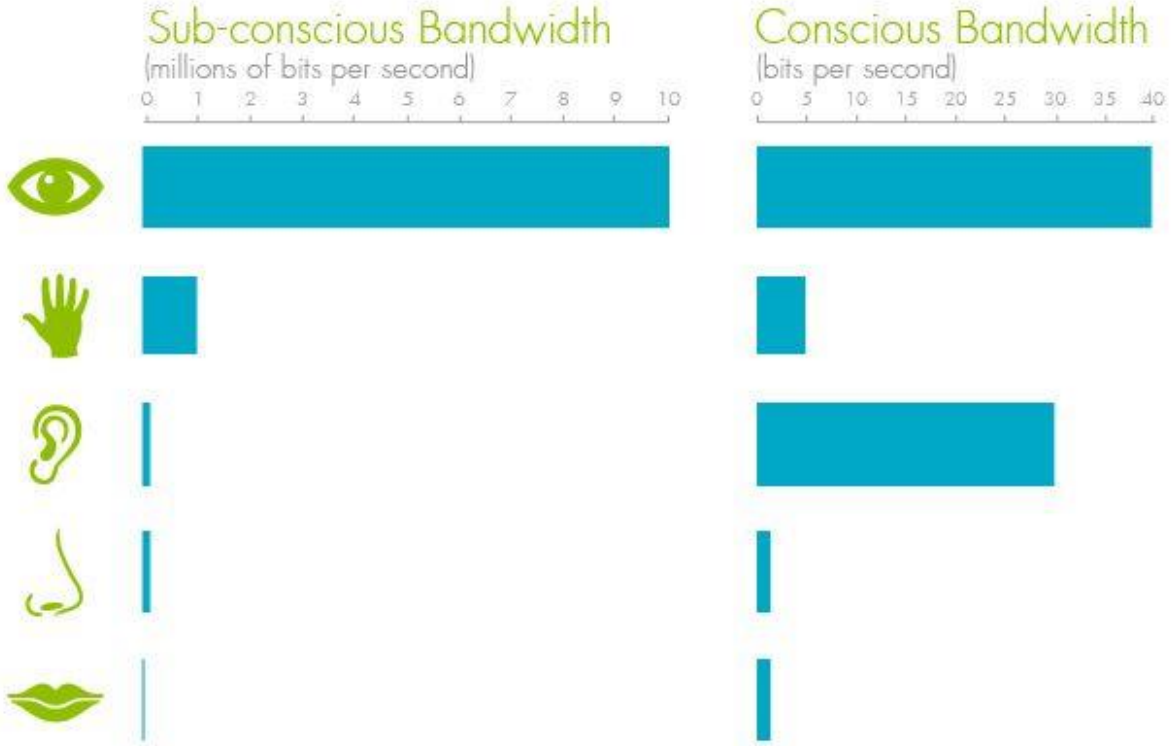
Tool to explore Genomic Stratifications in Cancer



WHY GRAPHICS?

Figures

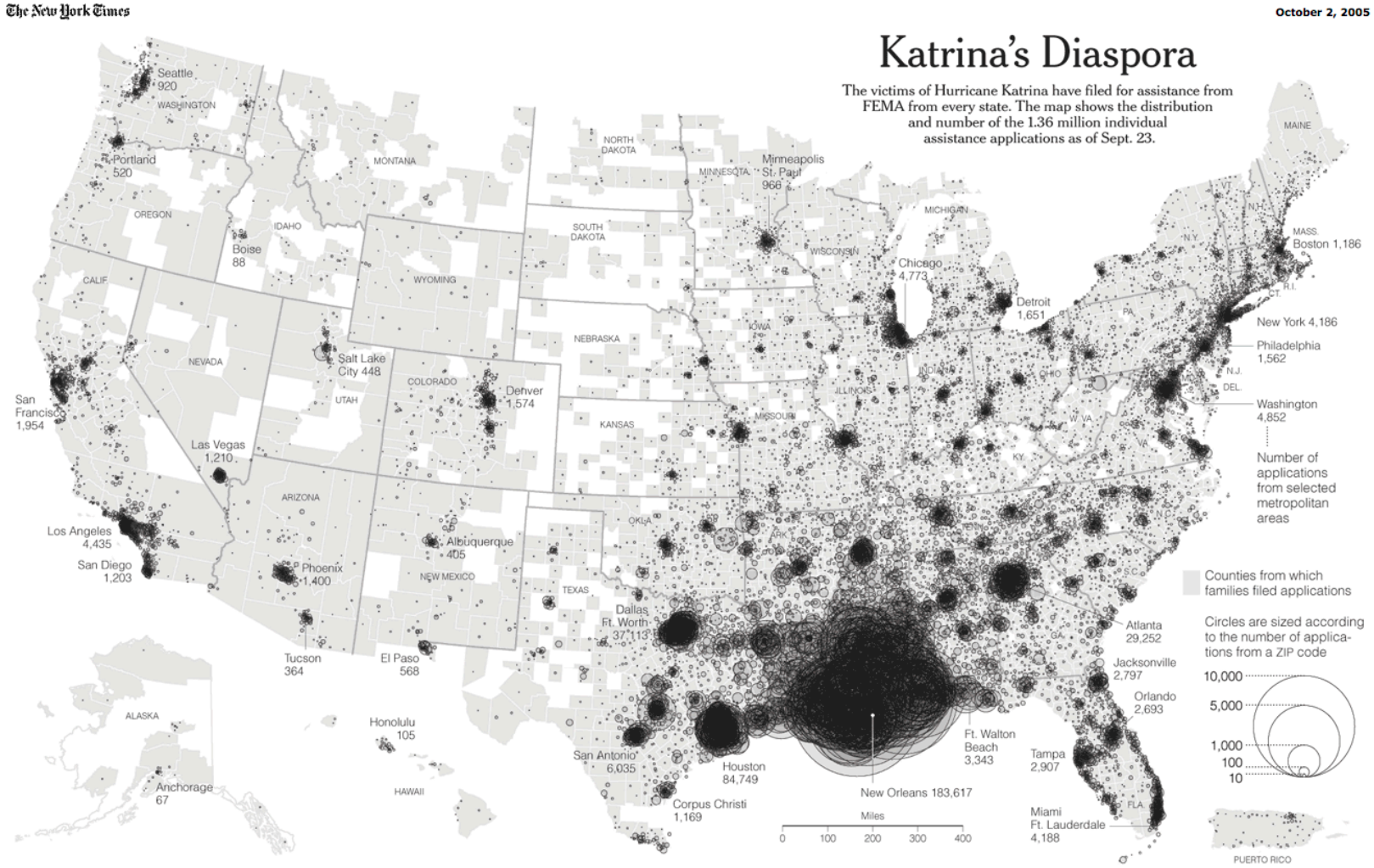
- Are **richer**; provide more info with less clutter and in less space
- Provide the *gestalt effect*: give an overview, **make structure more visible**
- Are more **accessible**, easier to understand, faster to grasp, more comprehensible, more memorable, more fun, and less formal



city's main public hospital was a wreck, and the city's public-housing projects were shuttered.

Campanella then switched to an identically constructed map, only this time based on 2010 census data, and in bits and pieces on the screen there was a simple and arresting picture of what Katrina meant. In the neighborhoods that were once a dense black, many of the little squares had thinned and turned gray. The sharp lines that once separated the teapot from Central City were now blurry: the white areas of the city were pushing north, into the vacuum left by the exodus. The Bywater was graying, as it gentrified still further. "Before Katrina, an American Community Survey estimate of New Orleans Parish population was four hundred and fifty-five thousand, and about sixty-eight per cent black," Campanella said. "Now the latest estimate is three hundred and eighty-four thousand, and it's about

TEXTUAL VS IMAGE



WHEN NOT TO VISUALIZE? WHEN TO AUTOMATE?

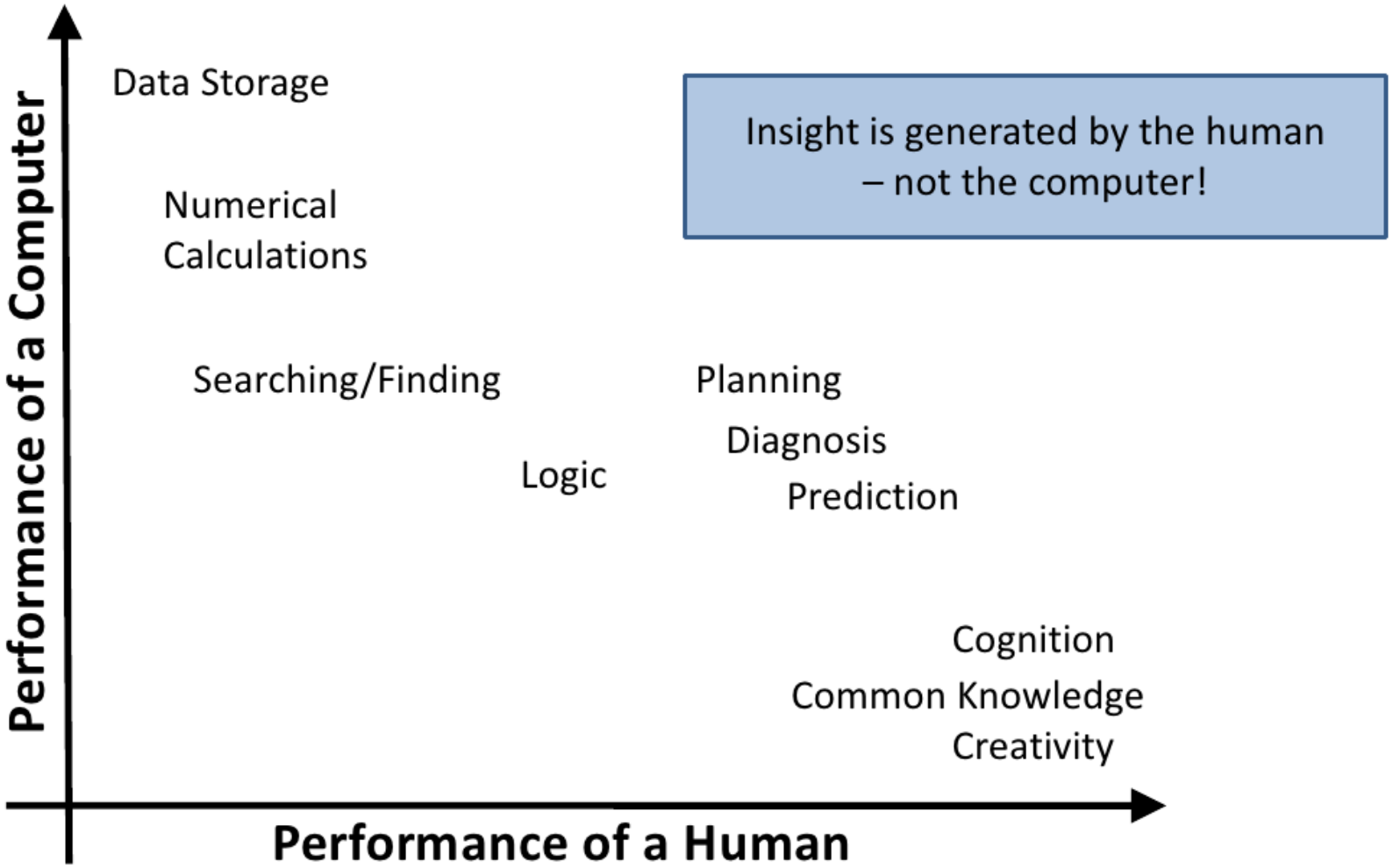
Well defined question on well-defined dataset

- *What is the current unemployment rate?*
- *Which city has the biggest contribution to the GDP of a country?*

No human intervention possible/necessary

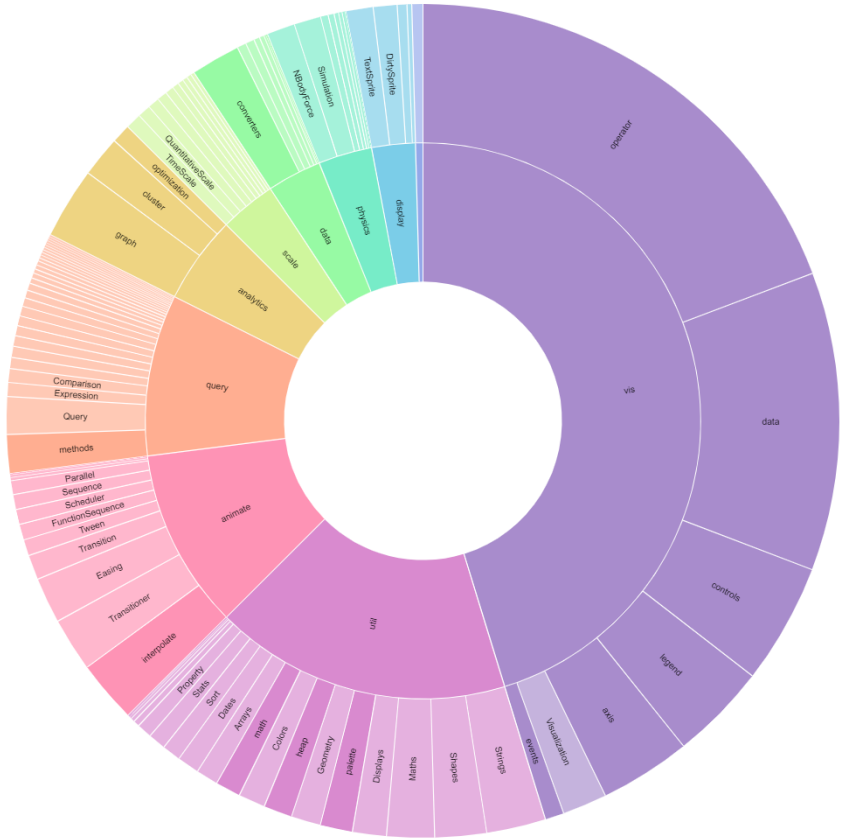
- Decision needed in minimal time
  - *High frequency stock market trading: which stock to buy/sell?*
  - *Manufacturing: is bottle broken?*
- Impractical for human to be involved
  - *Automatic data products*

COMPUTER VS HUMAN



WHY USE COMPUTER?

- Interaction
  - "drill down" into data
- Integration
  - Integrate with algorithms
  - Make visualization part of a data analysis pipeline



[Sunburst by John Stasko, Implementation in Caleydo by Christian Partl]
https://observablehq.com/@d3/zoomable-sunburst

WHY USE COMPUTER?

- Efficiency
  Re-use charts / methods for different datasets
- Quality
  Precise data driven rendering
- Storytelling
  - Use time

TELL STORIES

WHY NOT JUST USE STATISTICS

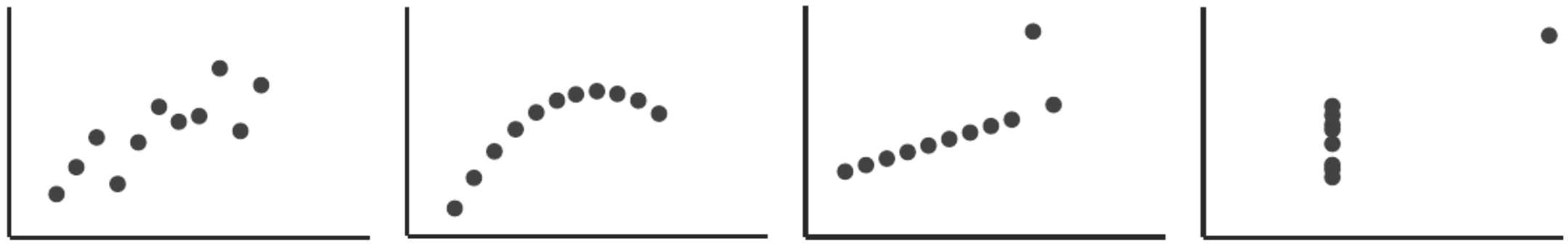
Anscombe's quartet

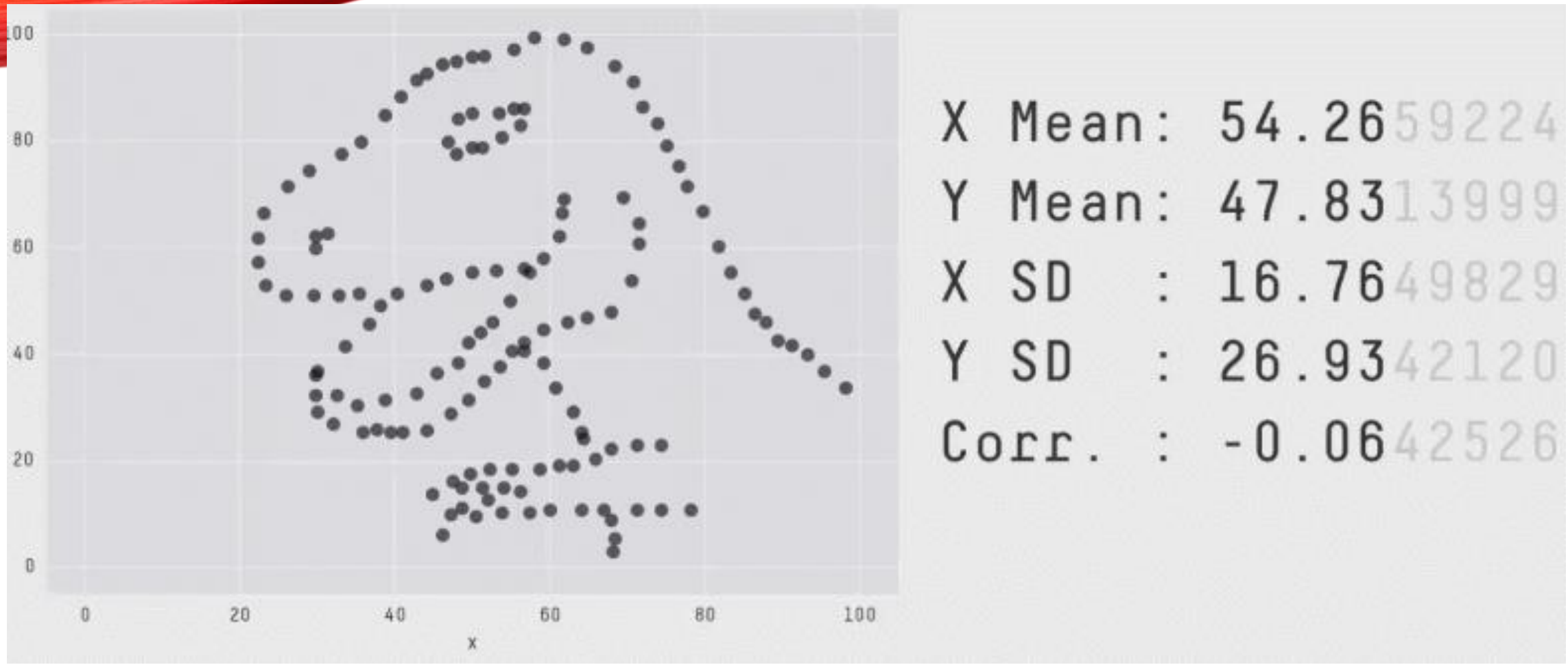
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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	6.89

For all four datasets

Property	Value
Mean of *x*	9
Sample variance of *x* : σ^2	11
Mean of *y*	7.50
Sample variance of *y* : σ^2	4.125
Correlation between *x* and *y*	0.816
Linear regression line	*y* = 3.00 + 0.500*x*
Coefficient of determination of the linear regression : R^2	0.67

ANSCOMBE'S QUARTET





Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, CHI 2017, Justin Matejka, George Fitzmaurice

VISUALIZATION = **HUMAN** **DATA** **INTERACTION**

DATA

VISUALIZATION IN DATA SCIENCE

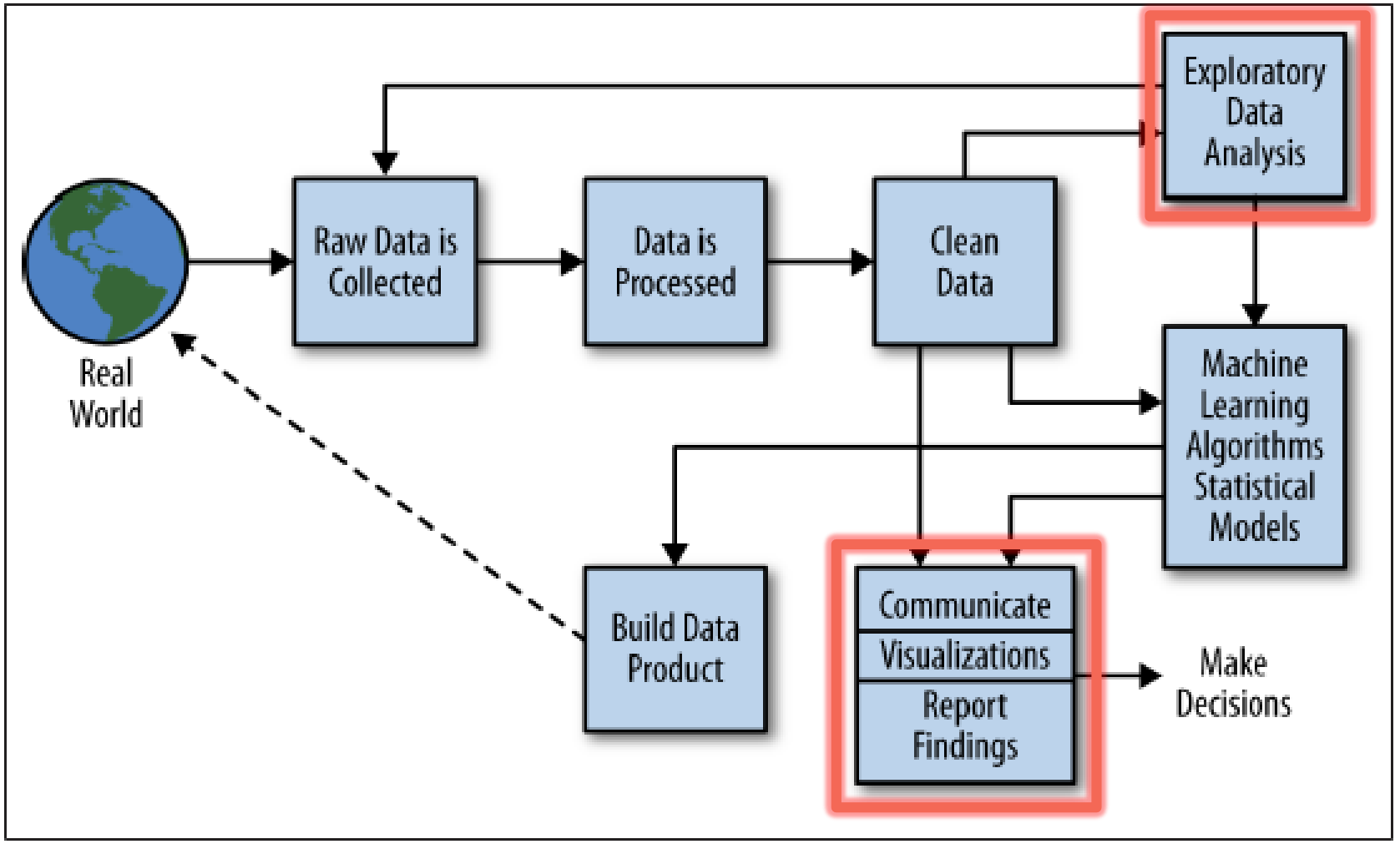


Figure 2-2. The data science process

BIG DATA

- Big Data

  2017: 2.5 exabytes (quintillion bytes)of data per day, largely unstructured

  2020: expected to reach 44 zettabytes

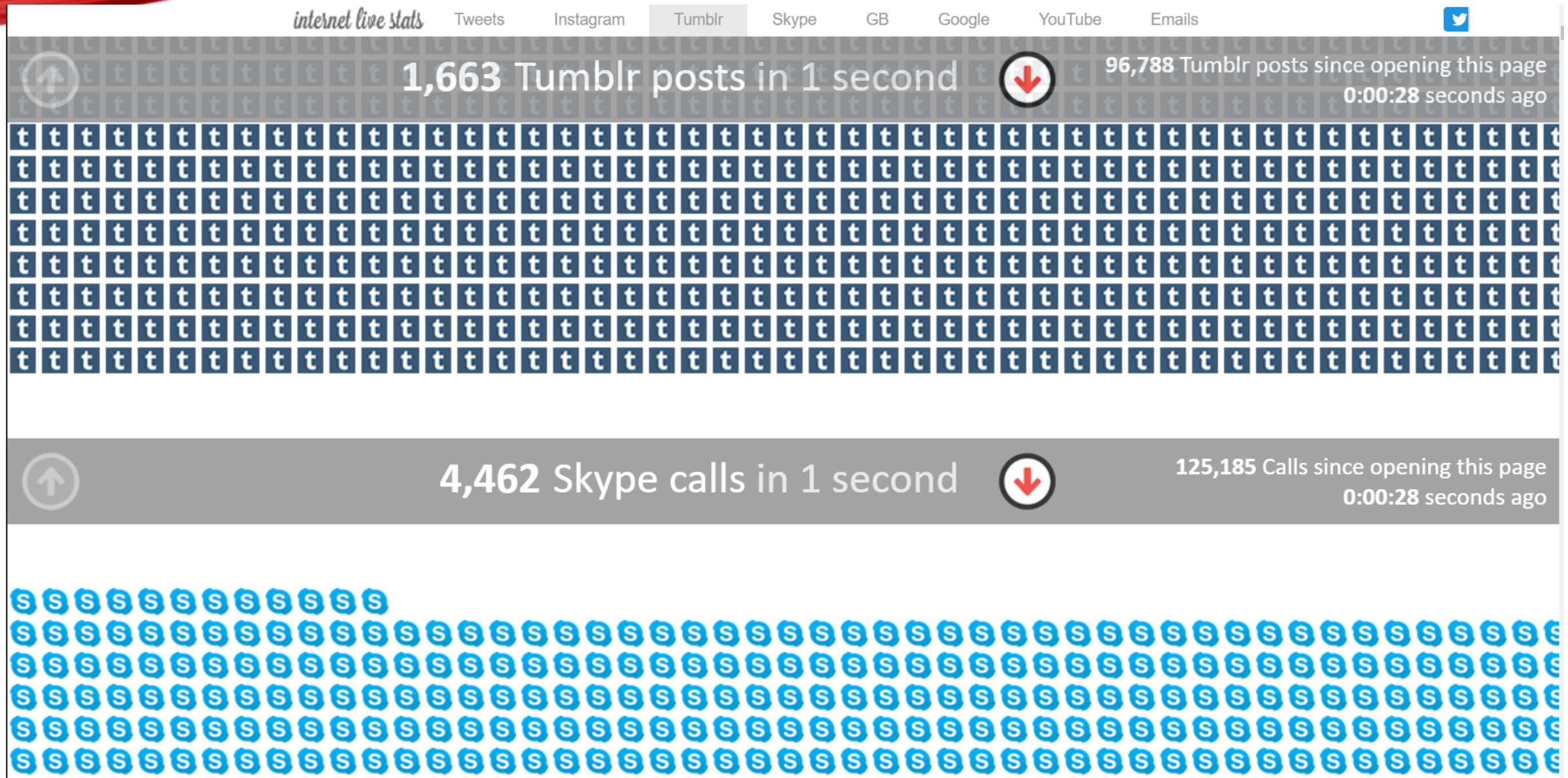
- 90% of the data created in last two years

Abbreviation	Unit	Value	Size (in bytes)
b	bit	0 or 1	1/8 of a byte
B	bytes	8 bits	1 byte
KB	kilobytes	1,000 bytes	1,000 bytes
MB	megabyte	$1{,}000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1{,}000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1{,}000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1{,}000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1{,}000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1{,}000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1{,}000^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

1 million TB →

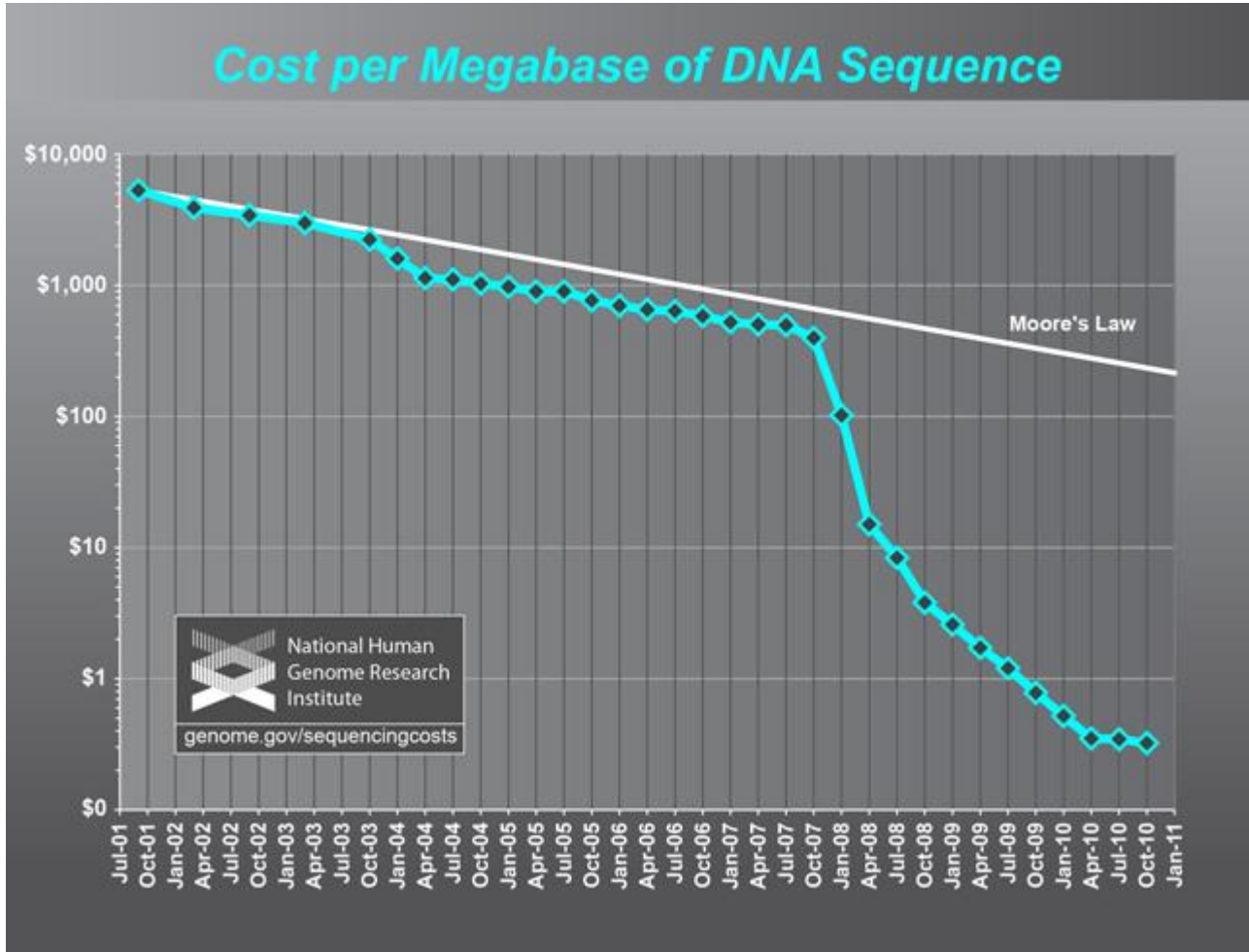
1000 billion TB →

SOURCE OF DATA



BIG DATA IN SCIENCE

- Big data is transforming science and engineering
  - Internet of things
  - Sensor
  - Camera
- In chemistry
- In medicine
- In smart building/smart city



GET A SENSE OF BIG DATA

- Example: CERN Large Hadron Collider Data
  CERN has publicly released over 300TB of data: CERN Open Data Portal

How much is that?

- A **DVD-R** holds 4.7 GB. You'd need **63,830** of them to hold 300 TB.
- It takes Pandora about a day and a half to burn through a gig of mobile data. So if the CERN data was an album, you could **stream it in just over 1,230 years**.
- At 350 MB per hour for 4K video streaming, so if the CERN data was a 4K movie it'd probably be about 857,142 hours, or about **98 years** long.
- But it ain't no thing compared to what the National Security Agency works with. Going by 2013 figures the agency released, the NSA's various activities **"touch" 300 TB of data every 15 minutes** or so
(Popular Mechanics Article

BIG DATA IN DATA CENTER

- NSA Utah Data Center (Bluffdale, Utah)
  Storage Capacity?
  estimates vary, but Forbes magazine estimates 12 exabytes (12,000 petabytes or 12 **million** terabytes)

"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, … because now we really do have **essentially free and ubiquitous data**."

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

HUMAN

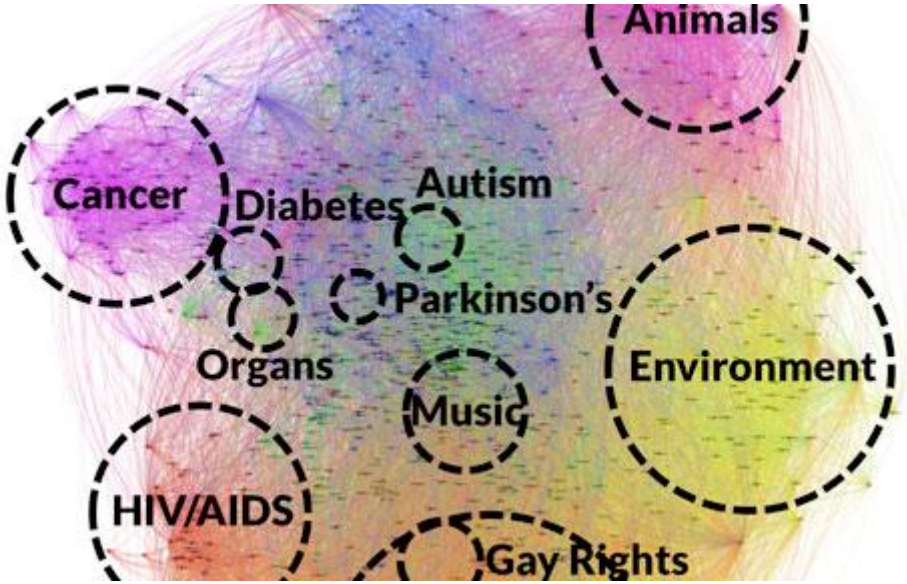
In human-data interaction

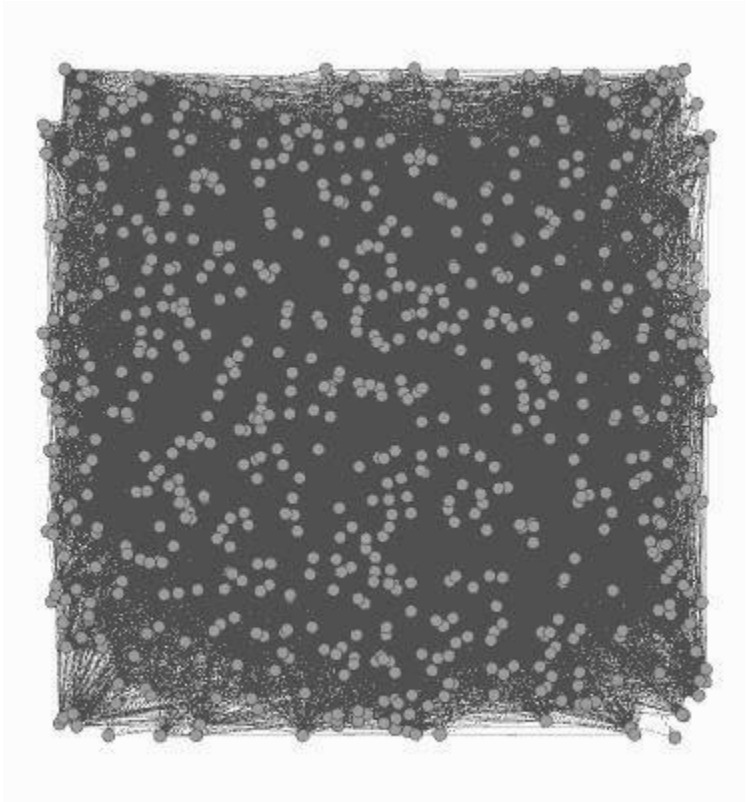
WHY HUMAN?

- Leveraging human capabilities
  Pattern discovery: clusters, outliers, trends
  Contextual knowledge: expectations for datasets, explanations for patterns
  Action: humans learn and take action

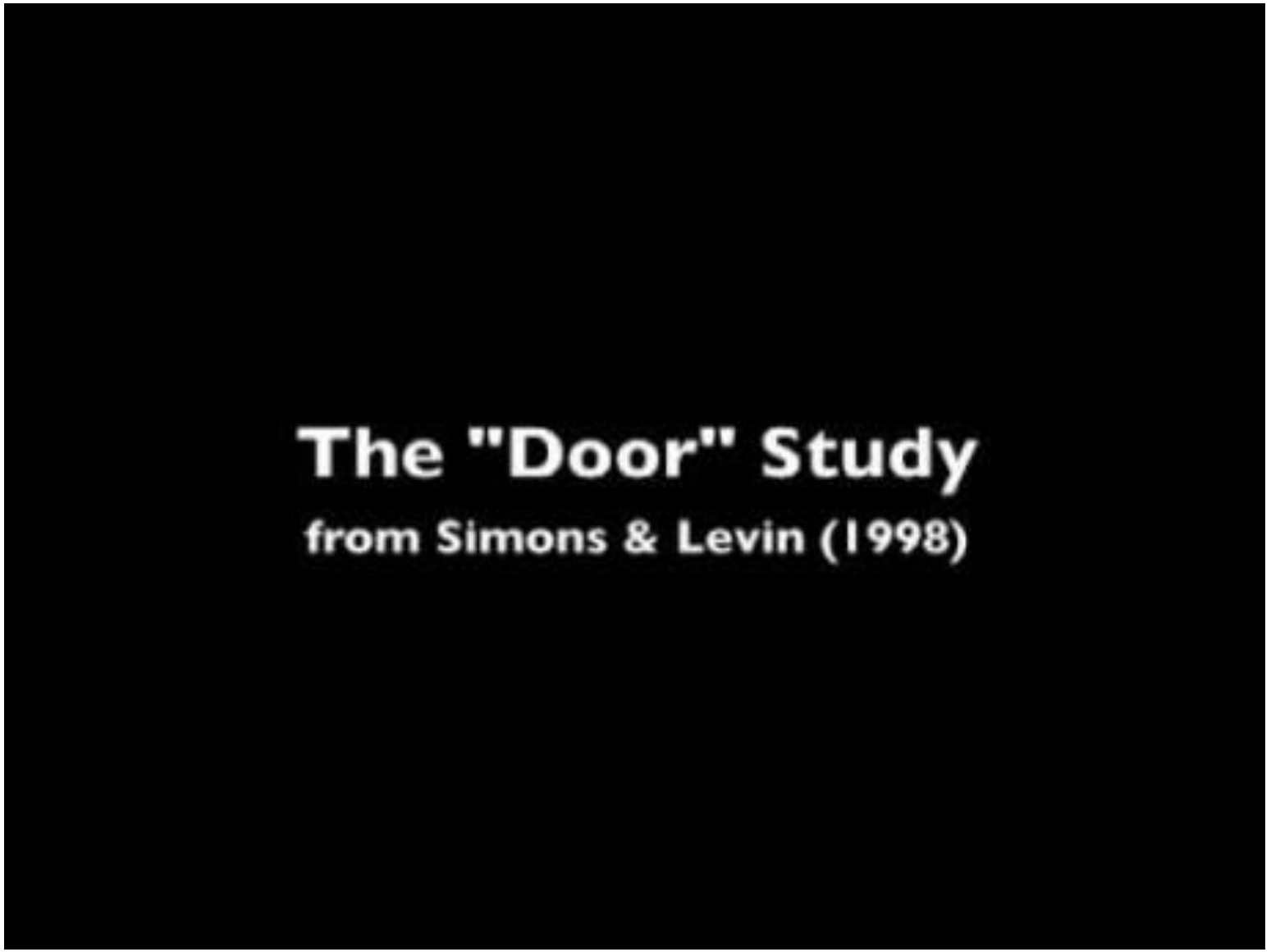
But humans have **limitation**, we have to design for them.

CAN BE DRAWN BUT CAN WE READ?



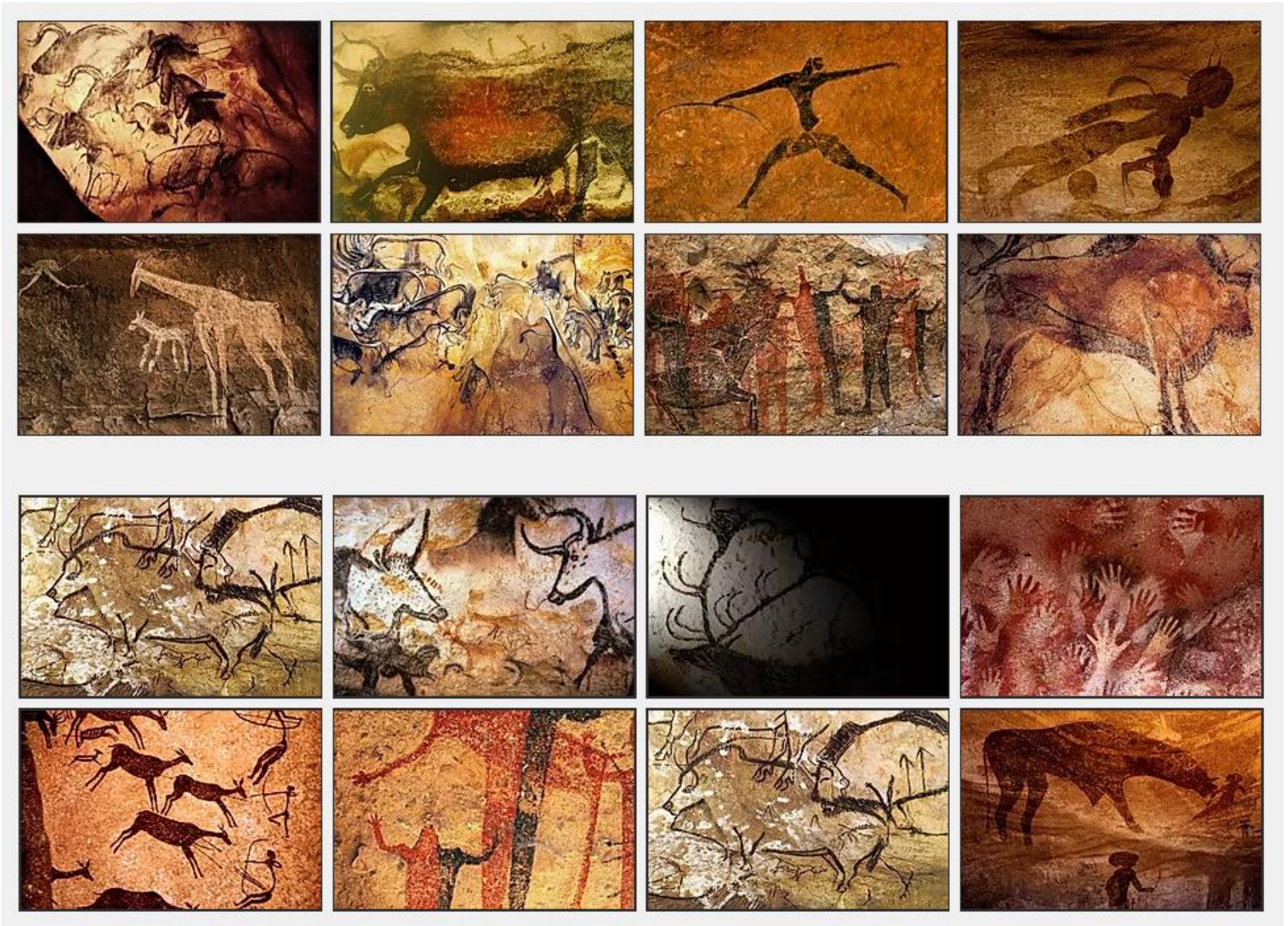


LIMITS OF COGNITION



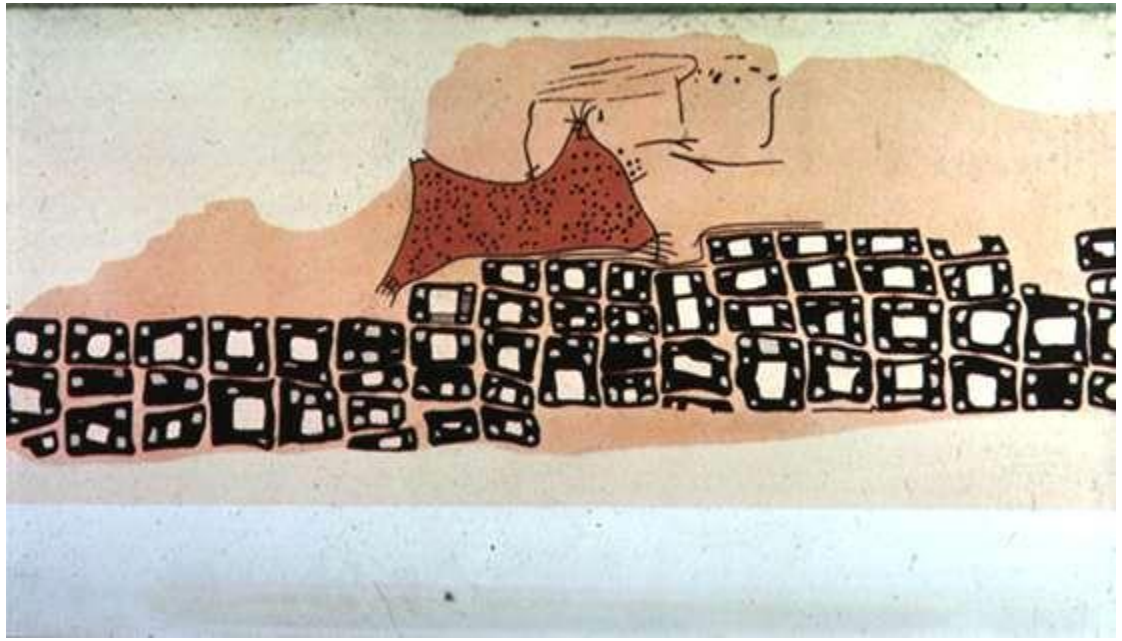
HISTORY OF VISUALIZATION

ROCK AND CAVE



The History of Visual Communication

RECORD

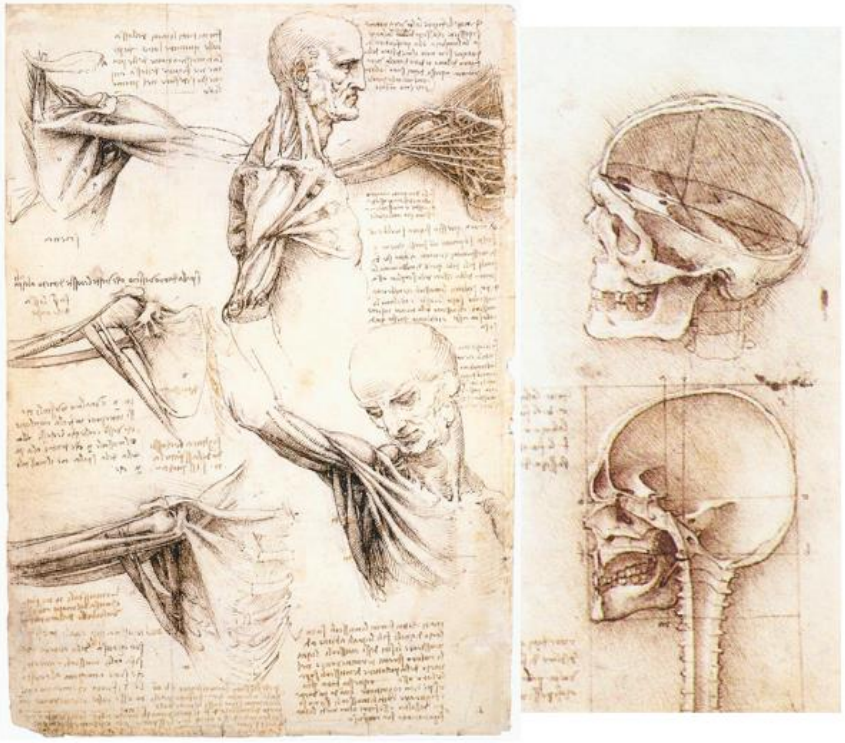


Konya town map, Turkey, c. 6200 BC

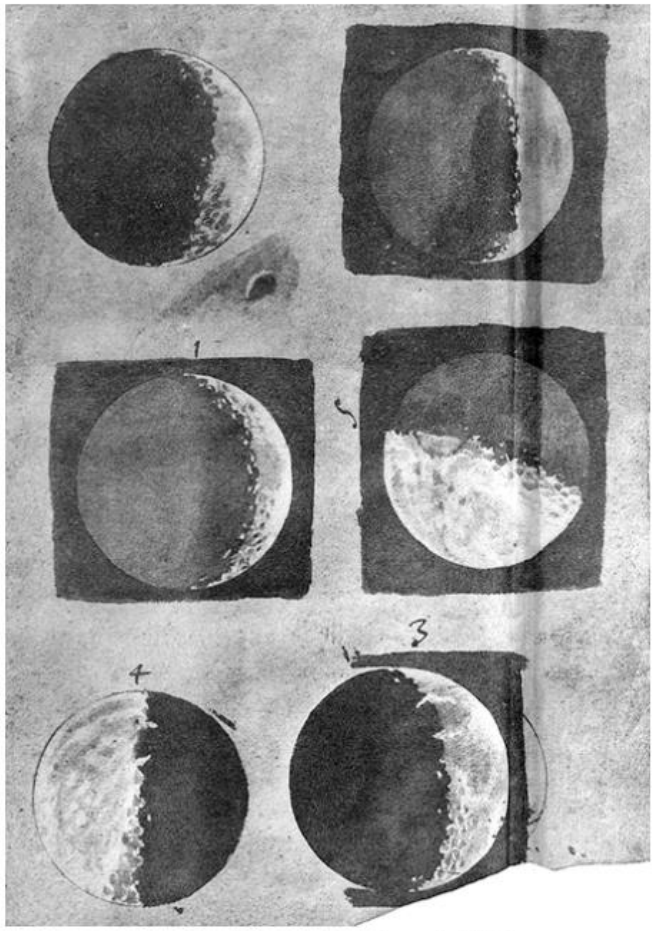
Anaximander's Map of the World

Anaximander of Miletus, c. 550 BC

RECORD



Leonardo Da Vinci, ca. 1500



Galileo Galilei, 1616

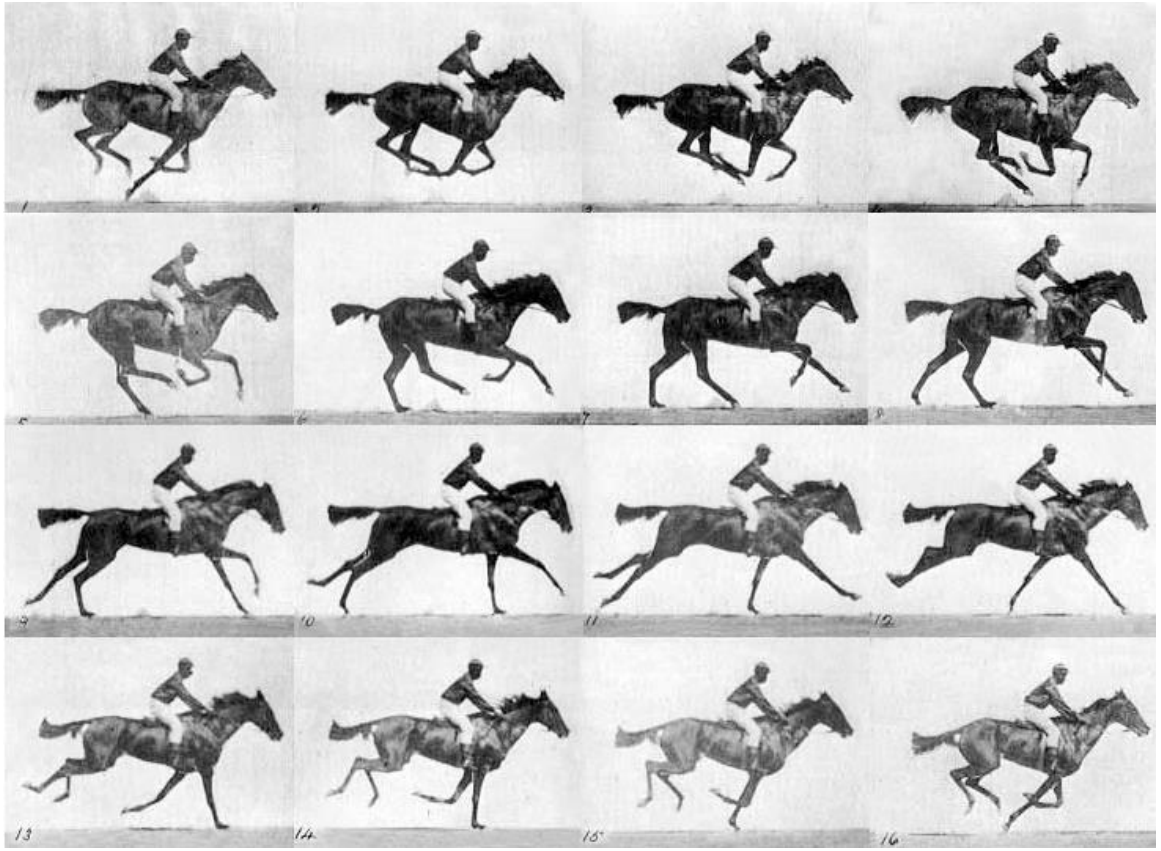
William Curtis (1746-1799)

Donald Norman

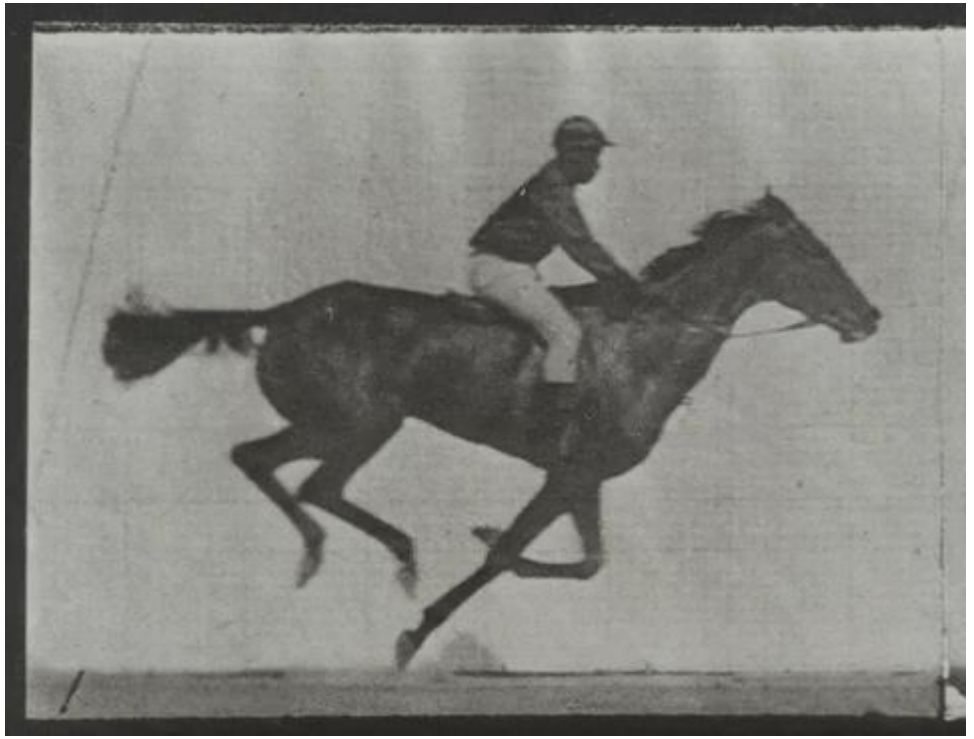
The History of Visual Communication

The Galileo Project, Rice University

RECORD



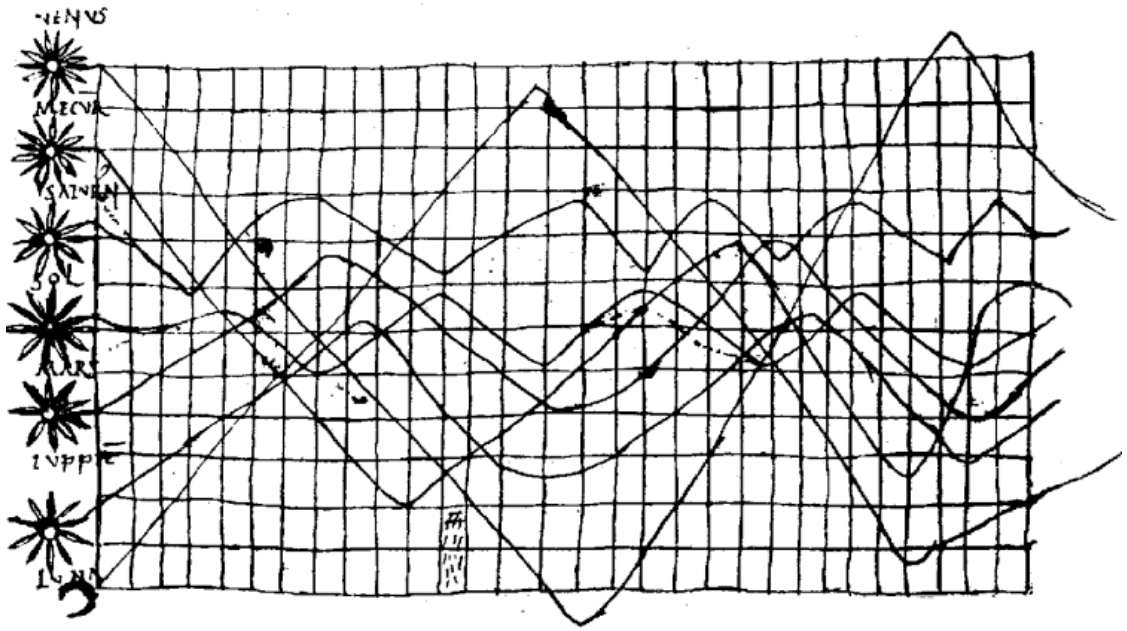
Eadweard J. Muybridge, 1878.
A Galloping horse



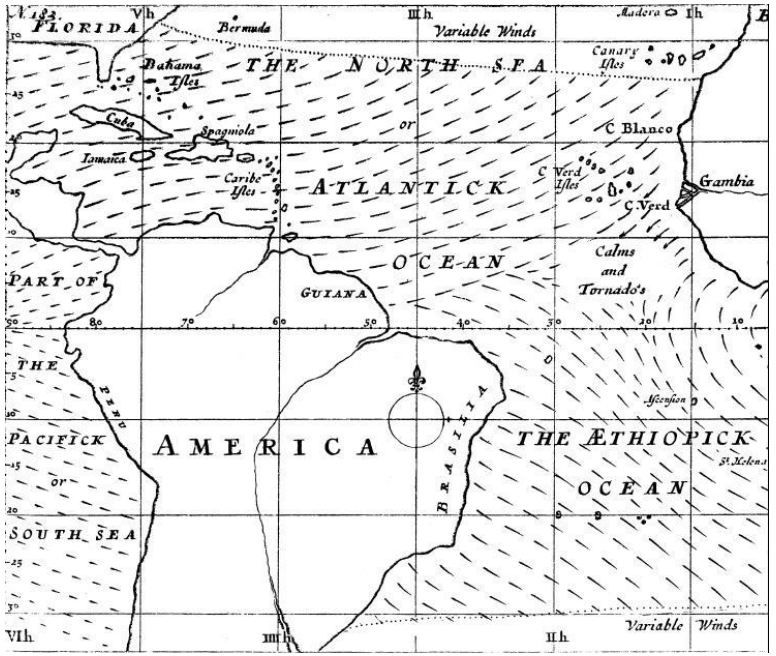
https://en.wikipedia.org/wiki/Eadweard_Muybridge

ANALYZE – TIME SERIES

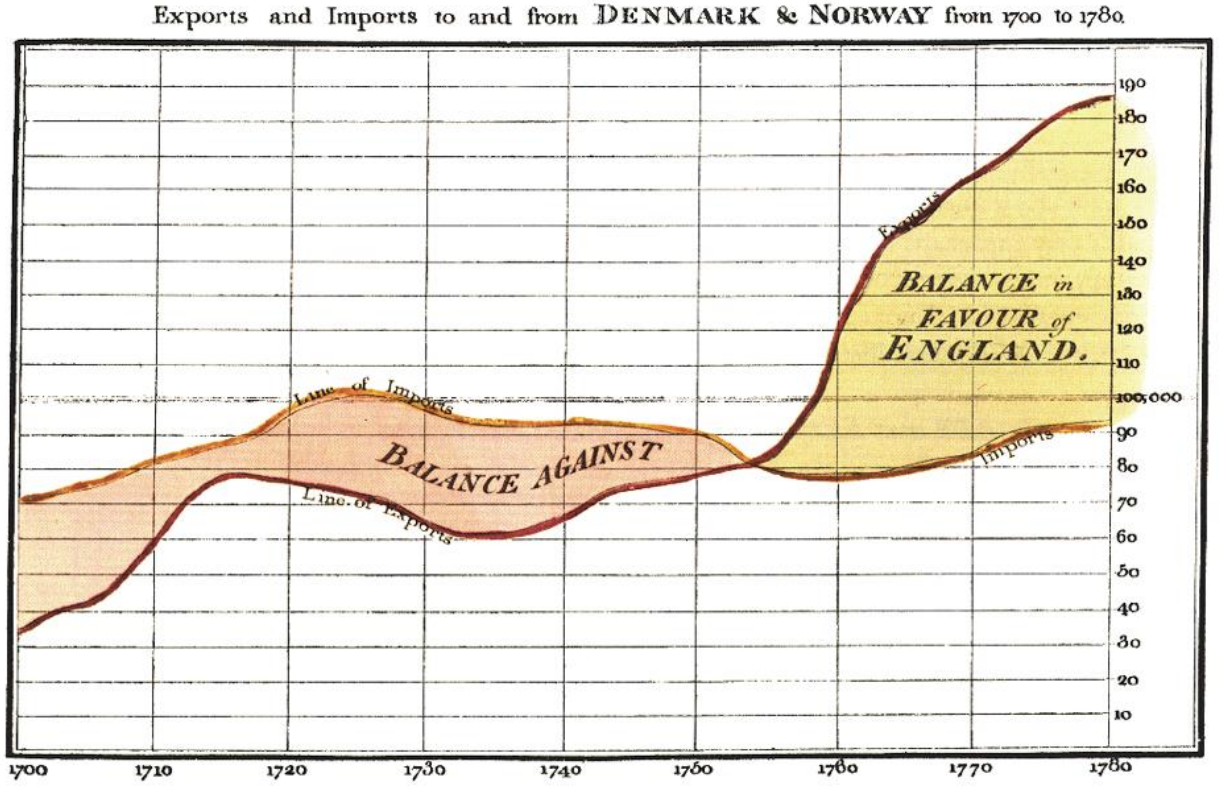
Planetary Movement Diagram, c. 950

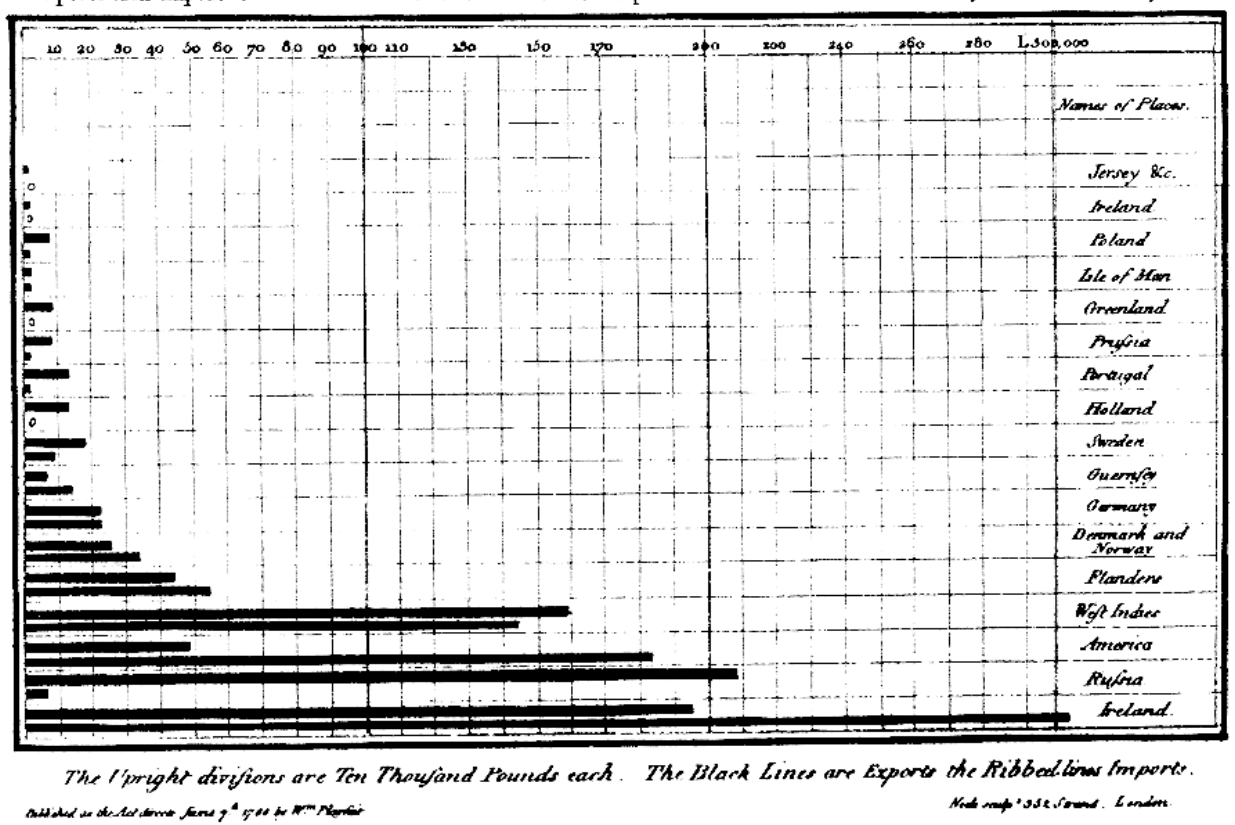


Halley's Wind Map, 1686



ANALYZE


Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780
BALANCE in FAVOUR of ENGLAND.
BALANCE AGAINST
Line of Imports
Line of Exports
Exports
Imports
190
180
170
160
150
140
130
120
110
100,000
90
80
70
60
50
40
30
20
10
1700
1710
1720
1730
1740
1750
1760
1770
1780

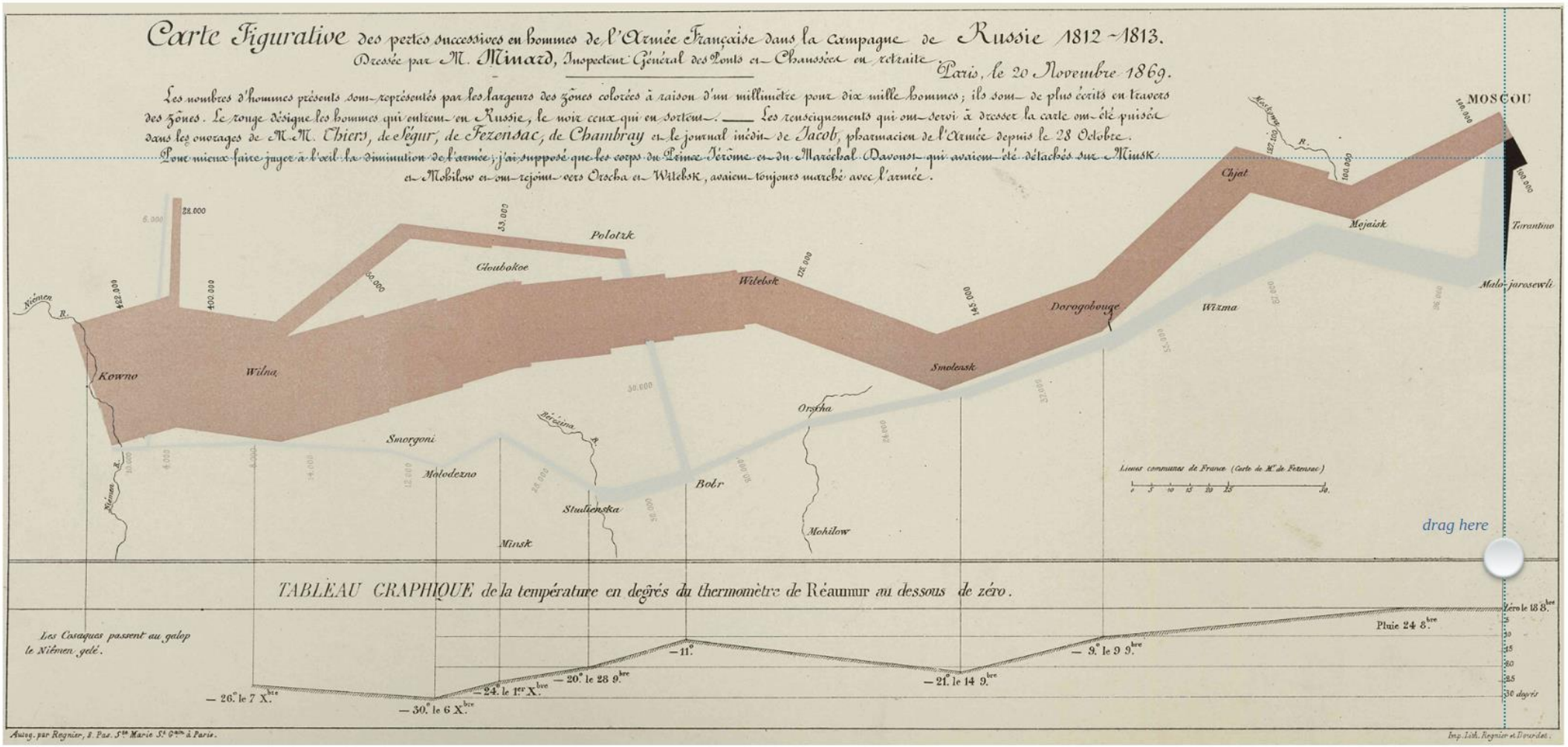

Exports and Imports of SCOTLAND to and from different parts for one Year from Christmas 1780 to Christmas 1781.
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 L.300,000
Names of Places.
Jersey &c.
Ireland
Poland
Isle of Man
Greenland
Prussia
Portugal
Holland
Sweden
Guernsey
Germany
Denmark and Norway
Flanders
West Indies
America
Russia
Ireland
The Upright divisions are Ten Thousand Pounds each. The Black Lines are Exports the Ribbed lines Imports.

FIND PATTERN

John Snow, 1854
Cholera outbreak

Spread by airbone ?
Map to show it is waterbone

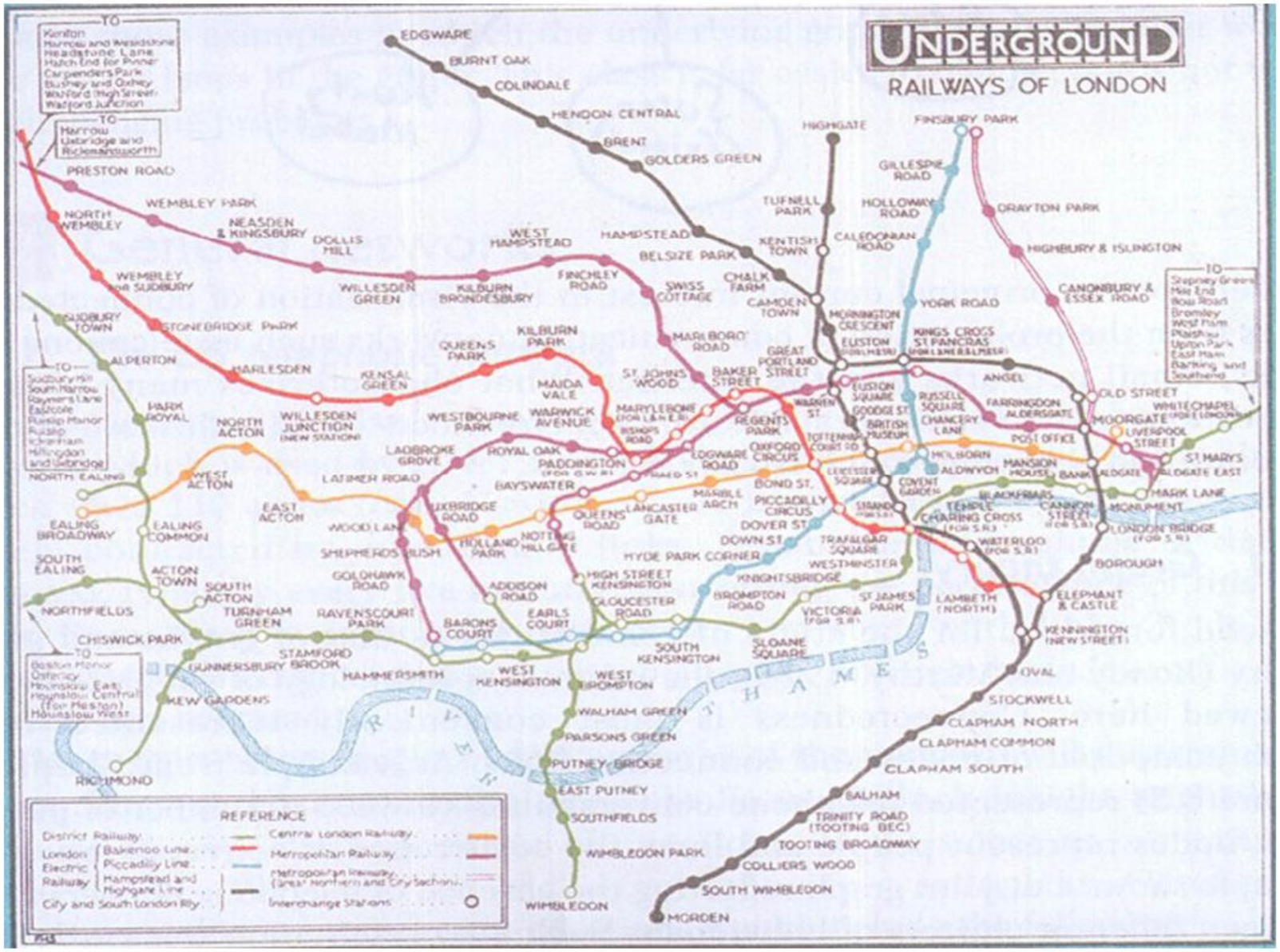
COMMUNICATE



https://www.masswerk.at/minard/ - Napoleon march - C.J. Minard, 1869

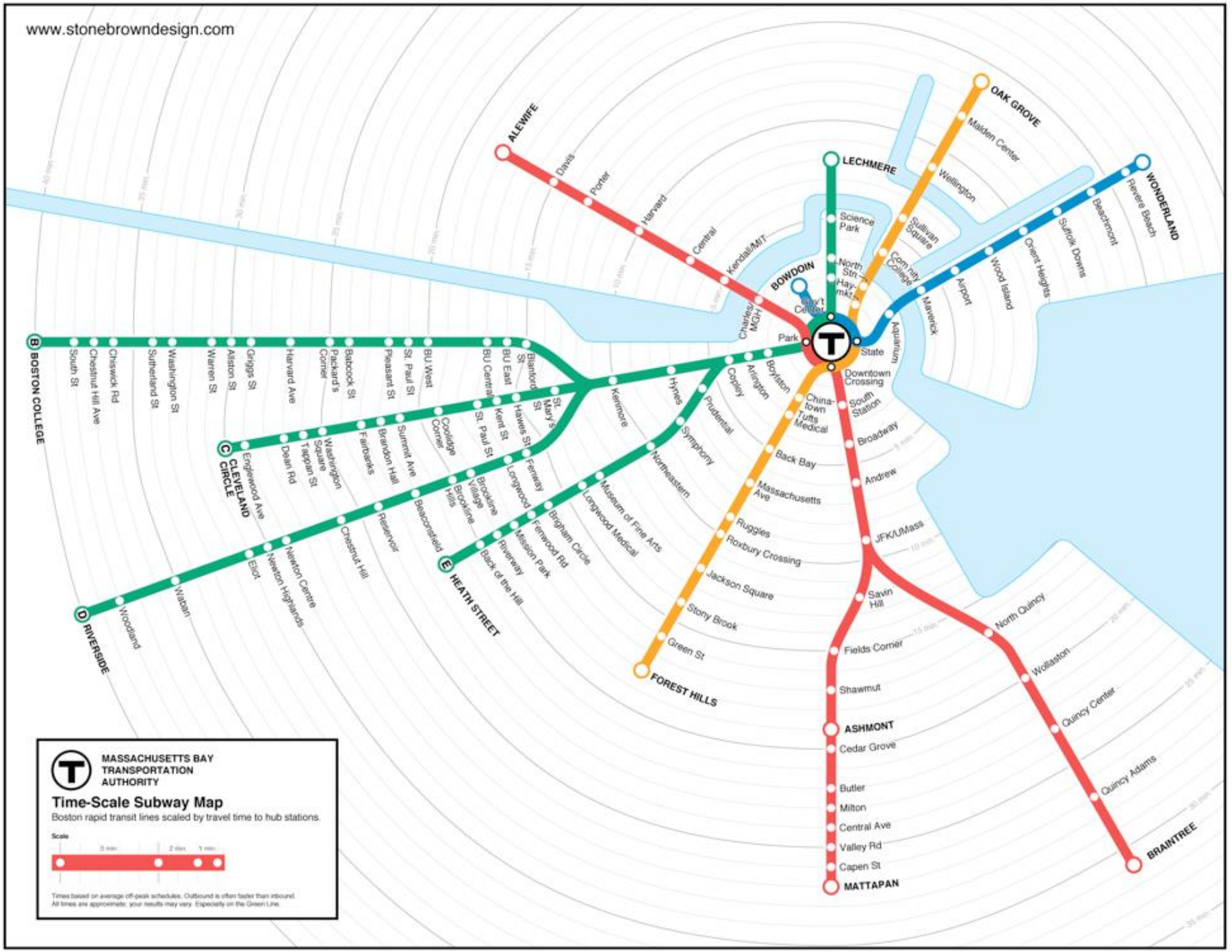
http://infowetrust.com/scroll/

COMMUNICATE

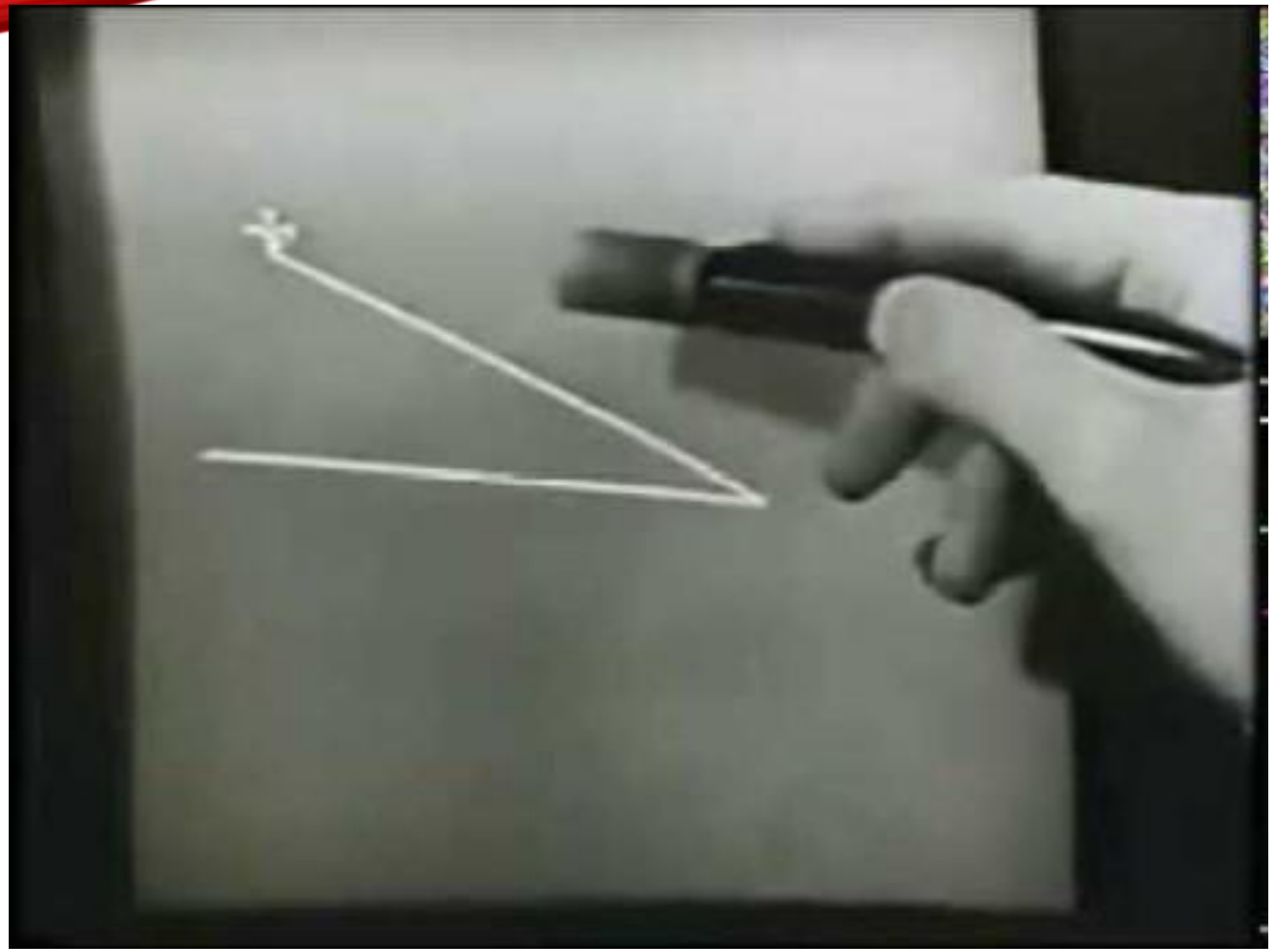


London Subway Map, 1927

SUBWAY MAP

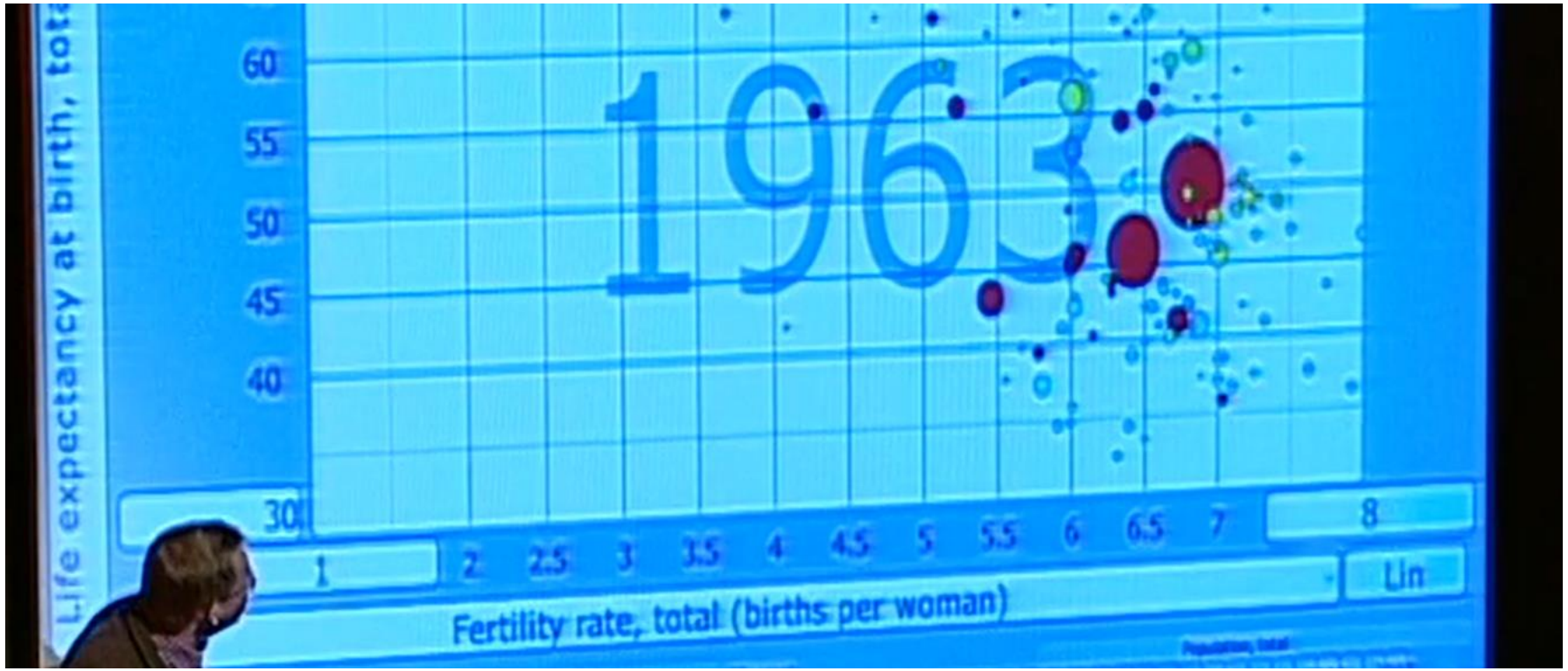

www.stonebrowndesign.com
ALEWIFE
Davis
Porter
Harvard
Central
Kendall/MIT
LECHMERE
Science Park
North Stn
Hay-mkt
Charles/MGH
BOWDOIN
Gov't Center
Park
State
Downtown Crossing
OAK GROVE
Malden Center
Wellington
Sullivan Square
Comm'nty College
Maverick
Airport
Wood Island
Orient Heights
Suffolk Downs
Beachmont
Revere Beach
WONDERLAND
Aquarium
Boylston
Arlington
Copley
Hynes
Kenmore
B BOSTON COLLEGE
South St
Chestnut Hill Ave
Chiswick Rd
Sutherland St
Washington St
Warren St
Alston St
Griggs St
Harvard Ave
Packard's Corner
Babcock St
Pleasant St
St. Paul St
BU West
BU Central
BU East
Blanford St
C CLEVELAND CIRCLE
Englewood Ave
Dean Rd
Tappan St
Washington Square
Fairbanks
Brandon Hall
Summit Ave
Coolidge Corner
St. Paul St
Kent St
Hawes St
St. Mary's St
Fenway
Longwood
Brookline Village
Brookline Hills
Beaconsfield
D RIVERSIDE
Woodland
Waban
Eliot
Newton Highlands
Newton Centre
Chestnut Hill
Reservoir
E HEATH STREET
Back of the Hill
Riverway
Mission Park
Fenwood Rd
Longwood Medical
Museum of Fine Arts
Northeastern
Symphony
Prudential
China-town
Tufts Medical
Back Bay
Massachusetts Ave
Ruggles
Roxbury Crossing
Jackson Square
Stony Brook
Green St
FOREST HILLS
South Station
Broadway
Andrew
JFK/UMass
Savin Hill
Fields Corner
Shawmut
ASHMONT
Cedar Grove
Butler
Milton
Central Ave
Valley Rd
Capen St
MATTAPAN
North Quincy
Wollaston
Quincy Center
Quincy Adams
BRAINTREE
MASSACHUSETTS BAY TRANSPORTATION AUTHORITY
Time-Scale Subway Map
Boston rapid transit lines scaled by travel time to hub stations.
Scale
Times based on average off-peak schedules. Outbound is often faster than inbound.
All times are approximate; your results may vary. Especially on the Green Line.

INTERACT



Ivan Sutherland, 1963

MODERN EXAMPLES



Hans Rosling, TED 2006

https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen

COURSE OBJECTIVES

How to **efficiently visualize data**

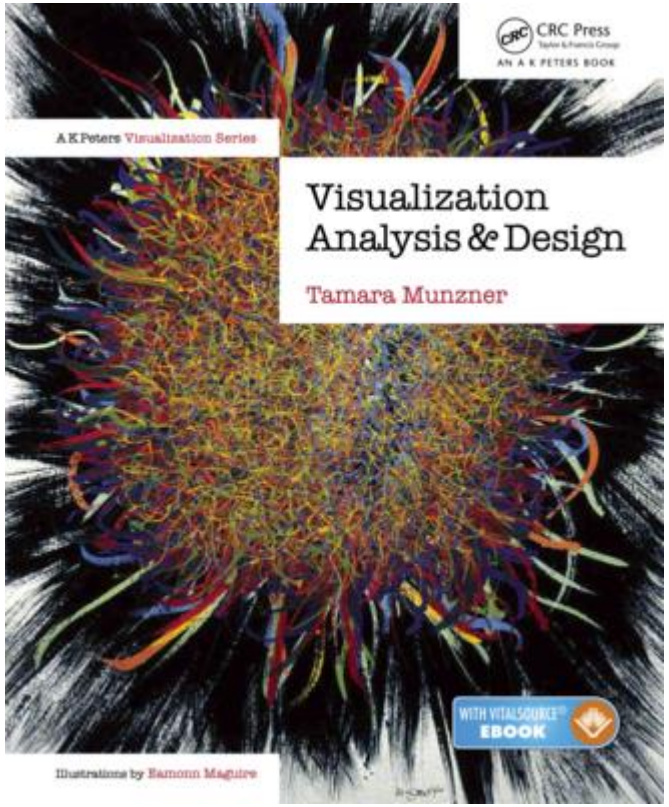
Evaluate and **critique** visualization designs

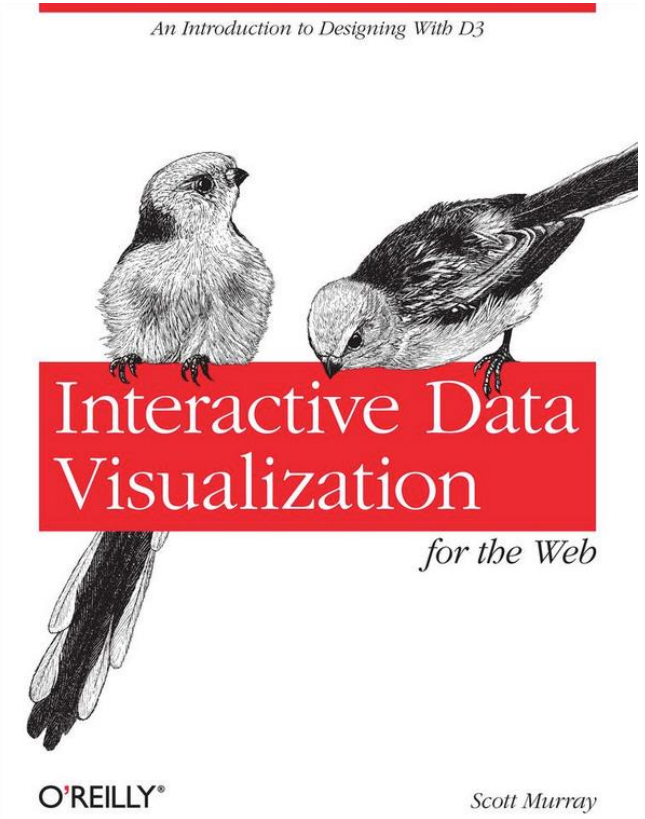
Apply fundamental principles & techniques

Design visual data analysis solutions

Implement interactive data visualizations (Web development skills)

TEXTBOOK


CRC Press
AN A K PETERS BOOK
A K Peters Visualization Series
Visualization Analysis & Design
Tamara Munzner
WITH VITALSOURCE EBOOK
Illustrations by Eamonn Maguire


An Introduction to Designing With D3
Interactive Data Visualization
for the Web
O'REILLY
Scott Murray

COURSE SCHEDULE

1. Introduction
2. Perception
3. Data abstraction, data types
4. Marks and channels
5. Design guidelines
6. Visualizing Interaction
7. Views, focus and context
8. Tabular Data
9. Storytelling with visualization
10. Visualizing networks and trees
11. Visualizing multivariate networks
12. Tasks Analysis, Designing and Evaluating Visualizations
13. Maps
14. Text visualization
15. Filtering and aggregation

GRADING

- Lab/Assignments/Quiz: 40%
- Midterm exam: 30%
- Final exam: 30%

PROJECT

- Design and implement a web-based **interactive** visualization to answer questions you have about some topic of your own choosing
- Task

1. Acquire data
2. Design a visualization
3. Implement it
4. Evaluate the result

- Team: 3-4 students
- To SUBMIT

1. A project proposal (by 16/03/2020)
2. A functional project prototype (The week after the midterm exam)
3. A final project submission (The week before the final exam)

PROJECT PROPOSAL

- Project title, link to github repo
- Background and motivation
- Objectives: the primary questions to answer
- Data: where and how to collect
- Data processing:
- Visualization design: 3 alternative prototype designs
- Must-have features
- Optional features
- Project schedule

TO SUBMIT:

3-4 pages of text, 5-6 pages of sketches of the prototype design