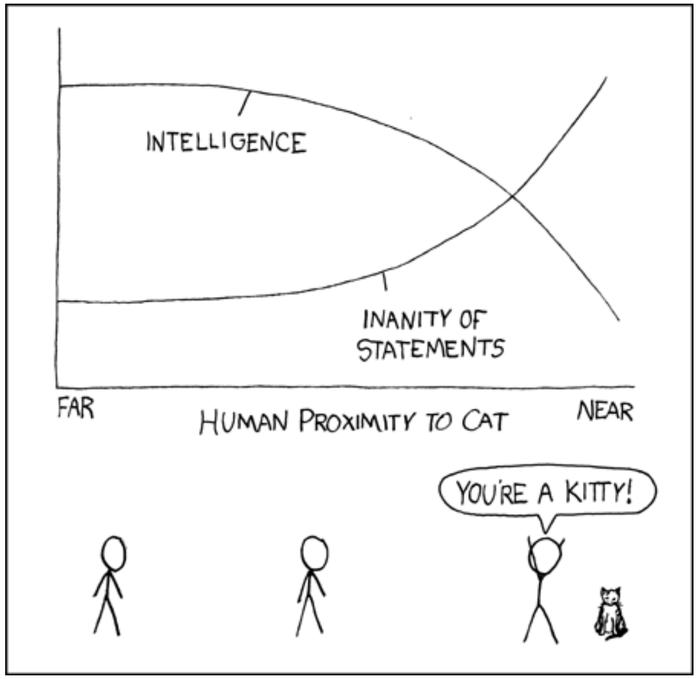
CS-5630 / CS-6630
Visualization for Data Science

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

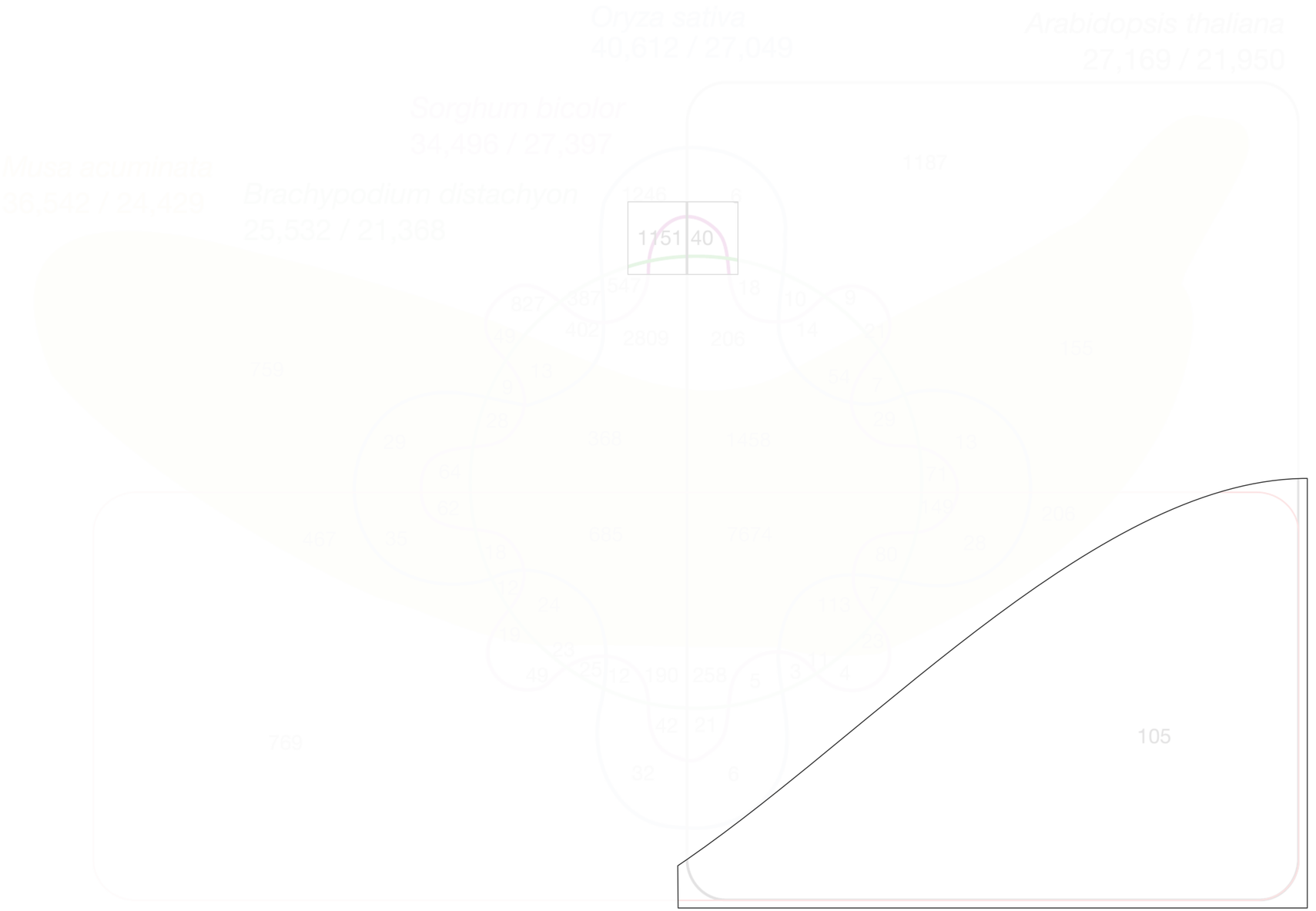
visualization *pictures*

The purpose of computing is insight, not numbers.

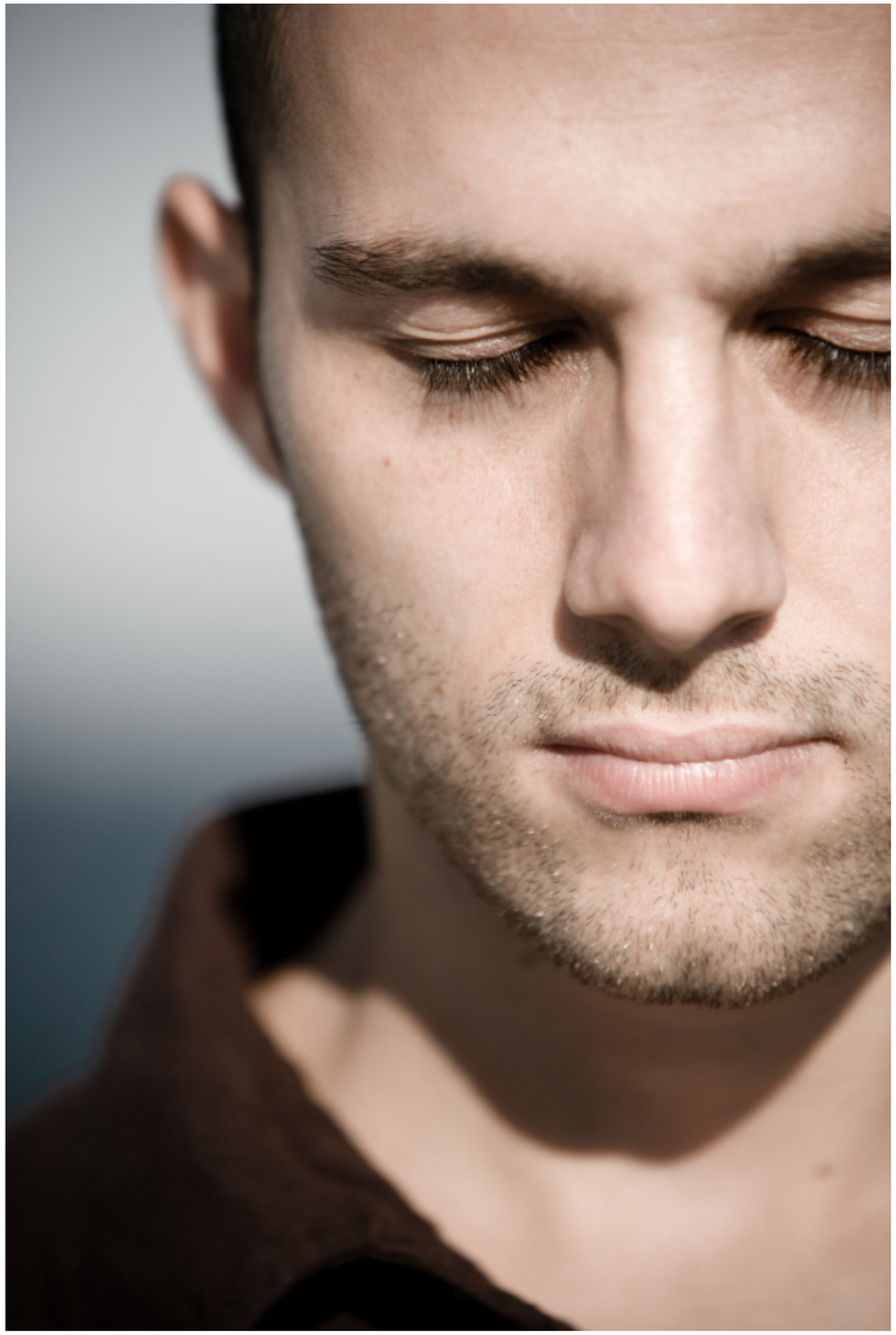
- Richard Wesley Hamming

- Card, Mackinlay, Shneiderman

Banana	***M. acuminata***
Date	***P. dactylifera***
Cress	***Arabidopsis thaliana***
Rice	***Oryza sativa***
Sorghum	***Sorghum bicolor***
Brome	***Brachypodium distachyon***



[D'Hont et al., Nature, 2012]



vi · su · al · i · za · tion

1. Formation of mental visual images
2. The act or process of interpreting in visual terms or of putting into visible form

The American Heritage Dictionary

Visualization Definition

Visualization is the process that **transform**s
(abstract) **data** into
interactive graphical representations 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

Why Visualize?

To inform humans: Communication

How is ahead in the election polls?

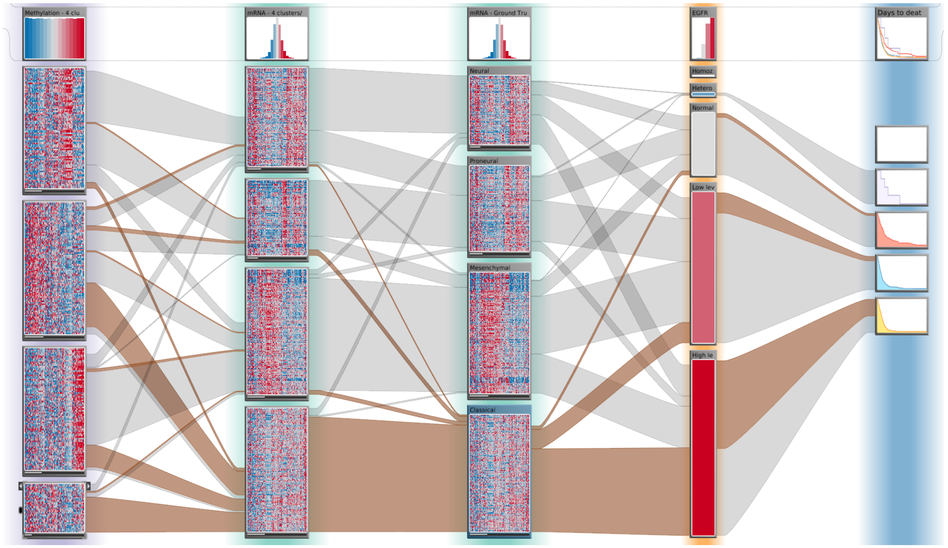
When questions are not well defined: Exploration

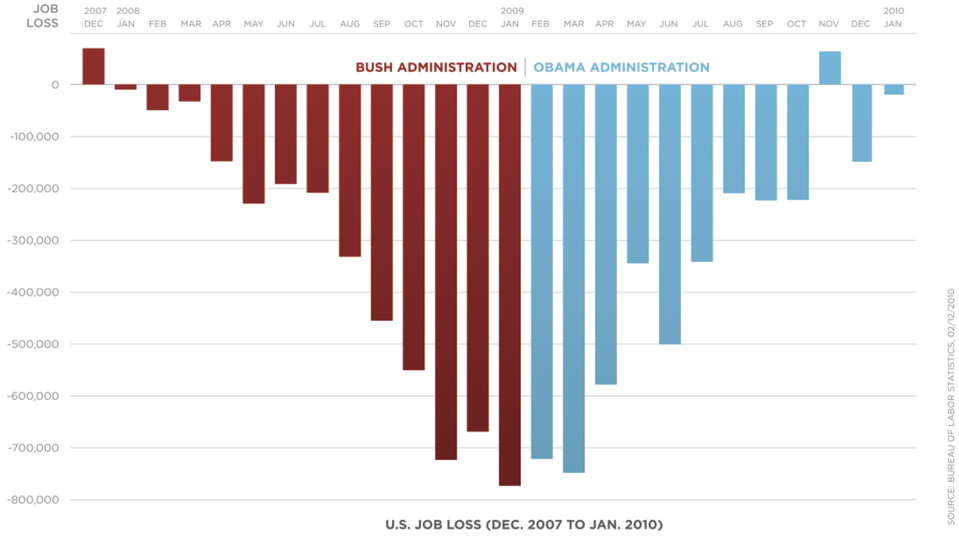
What is the structure of a terrorist network?

Which drug can help patient X?

Purpose of Visualization

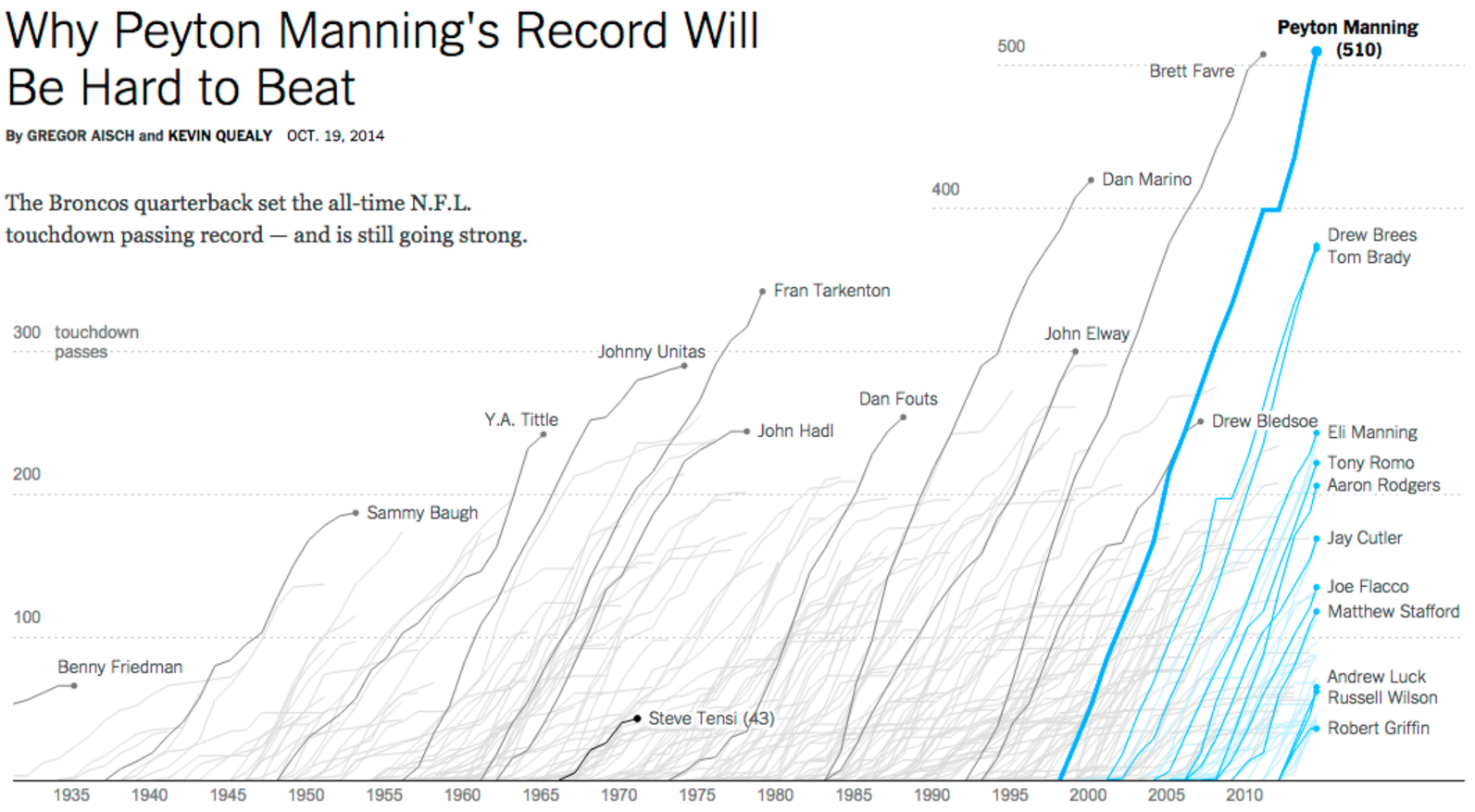
[Obama Administration]





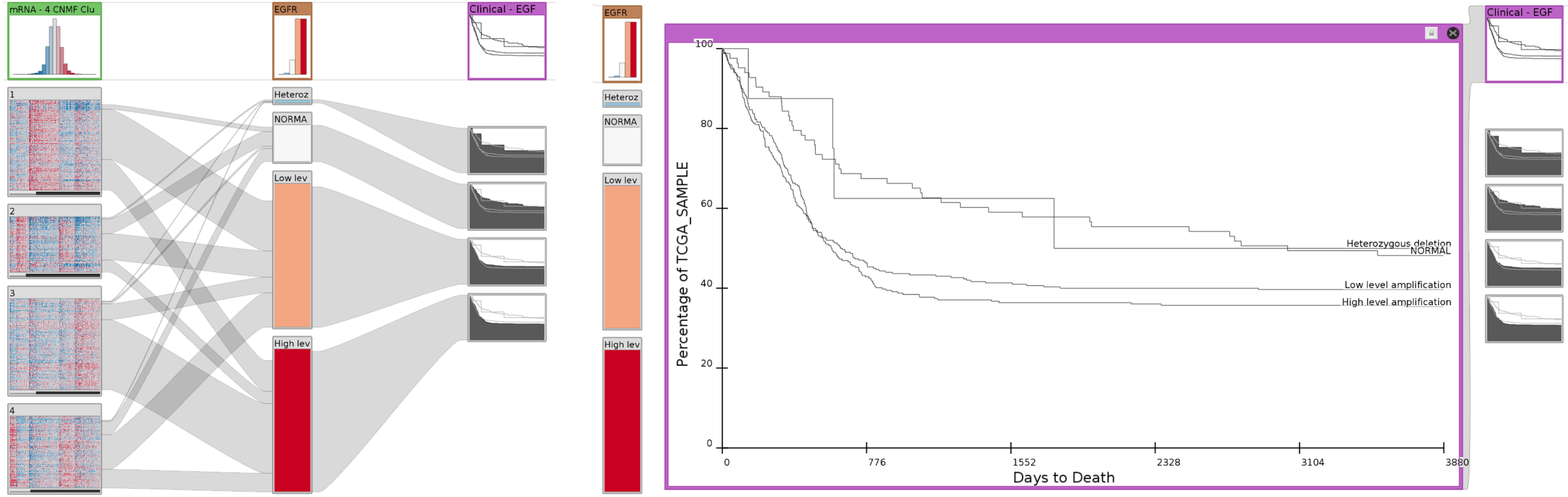
Open Exploration — Confirmation — Communication

Example Communication



[New York Times]

Example Exploration: Cancer Subtypes



[Caleydo StratomeX]

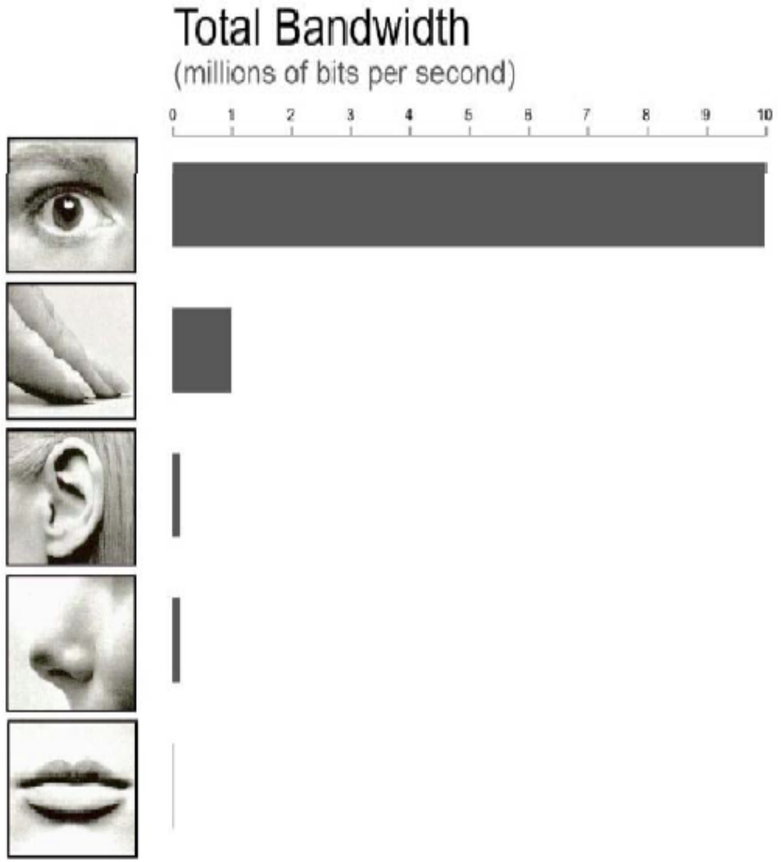
Why Graphics?

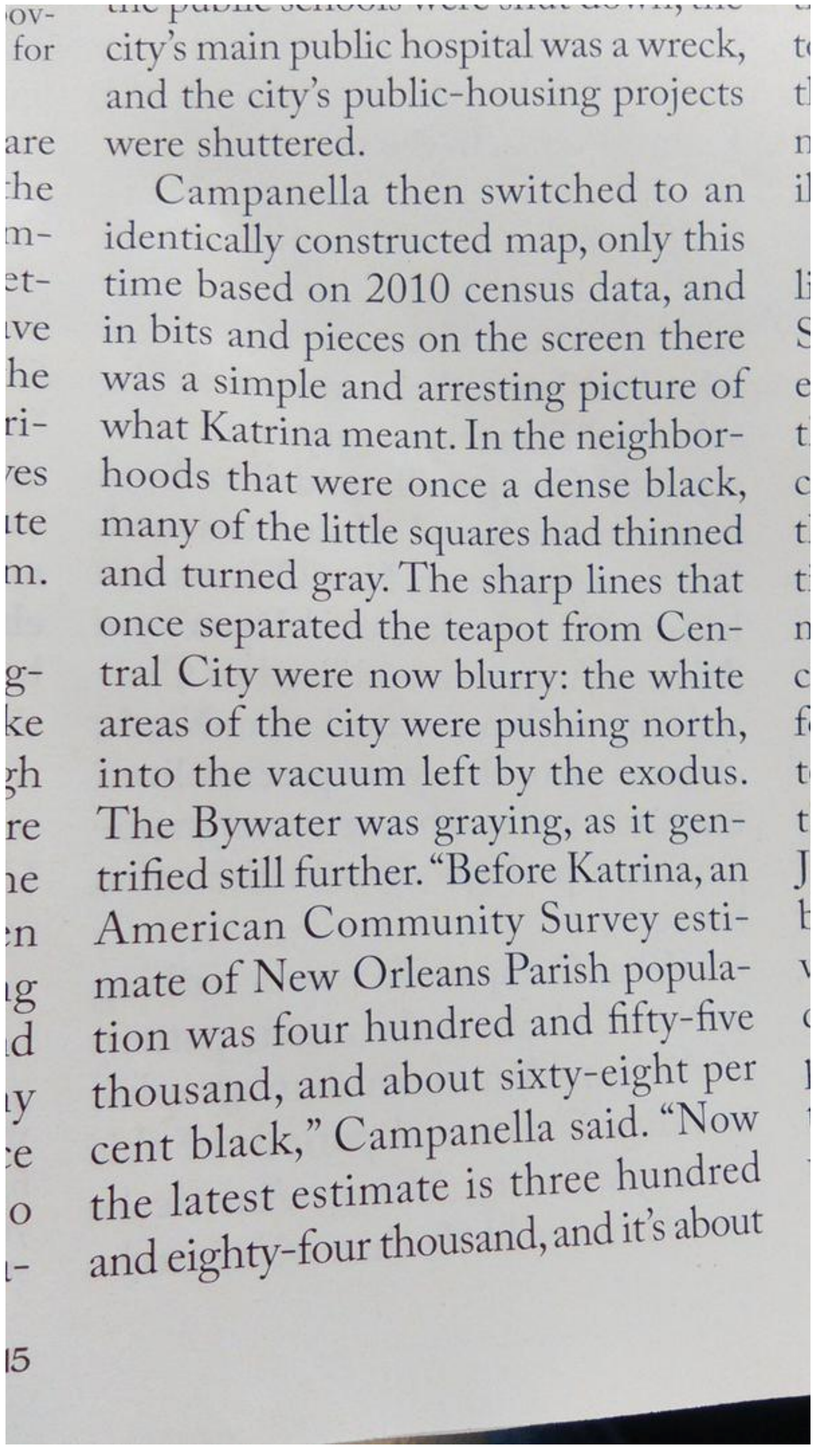
Figures are richer; provide more information with less clutter and in less space.

Figures provide the *gestalt* effect: they give an overview; make structure more visible.

Figures are more accessible, easier to understand, faster to grasp, more comprehensible, more memorable, more fun, and less formal.

list adapted from: [Stasko et al. 1998]





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

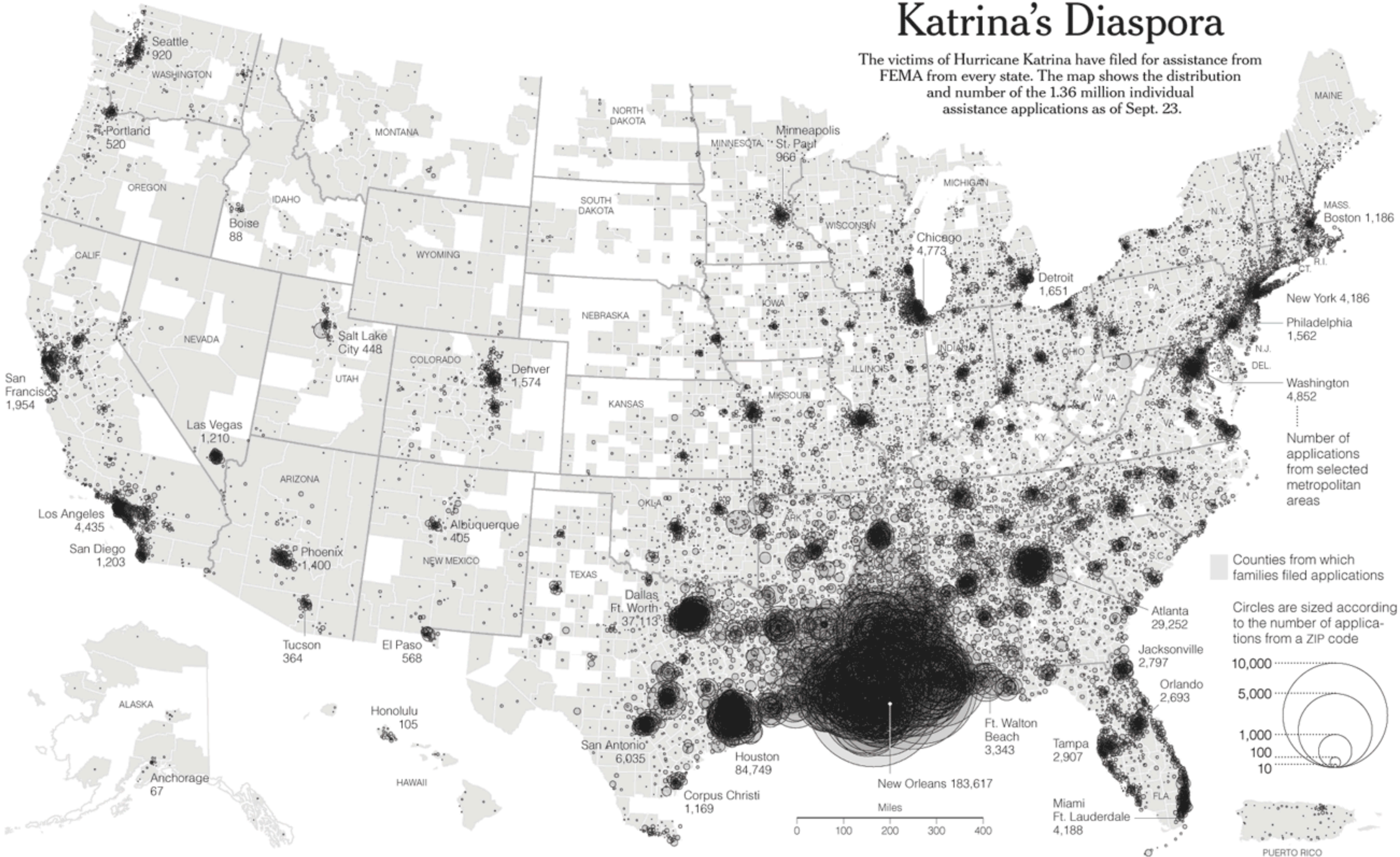
Textual description of a map of the effects of hurricane Katrina on New Orleans.
New Yorker, posted by Alberto Cairo

The New York Times

October 2, 2005

Katrina's Diaspora

The victims of Hurricane Katrina have filed for assistance from FEMA from every state. The map shows the distribution and number of the 1.36 million individual assistance applications as of Sept. 23.



When not to visualize?
When to automate?

Well defined question on well-defined dataset

Which gene is most frequently mutated in this set of patients?

What is the current unemployment rate?

No human intervention possible/necessary

Decisions needed in minimal time

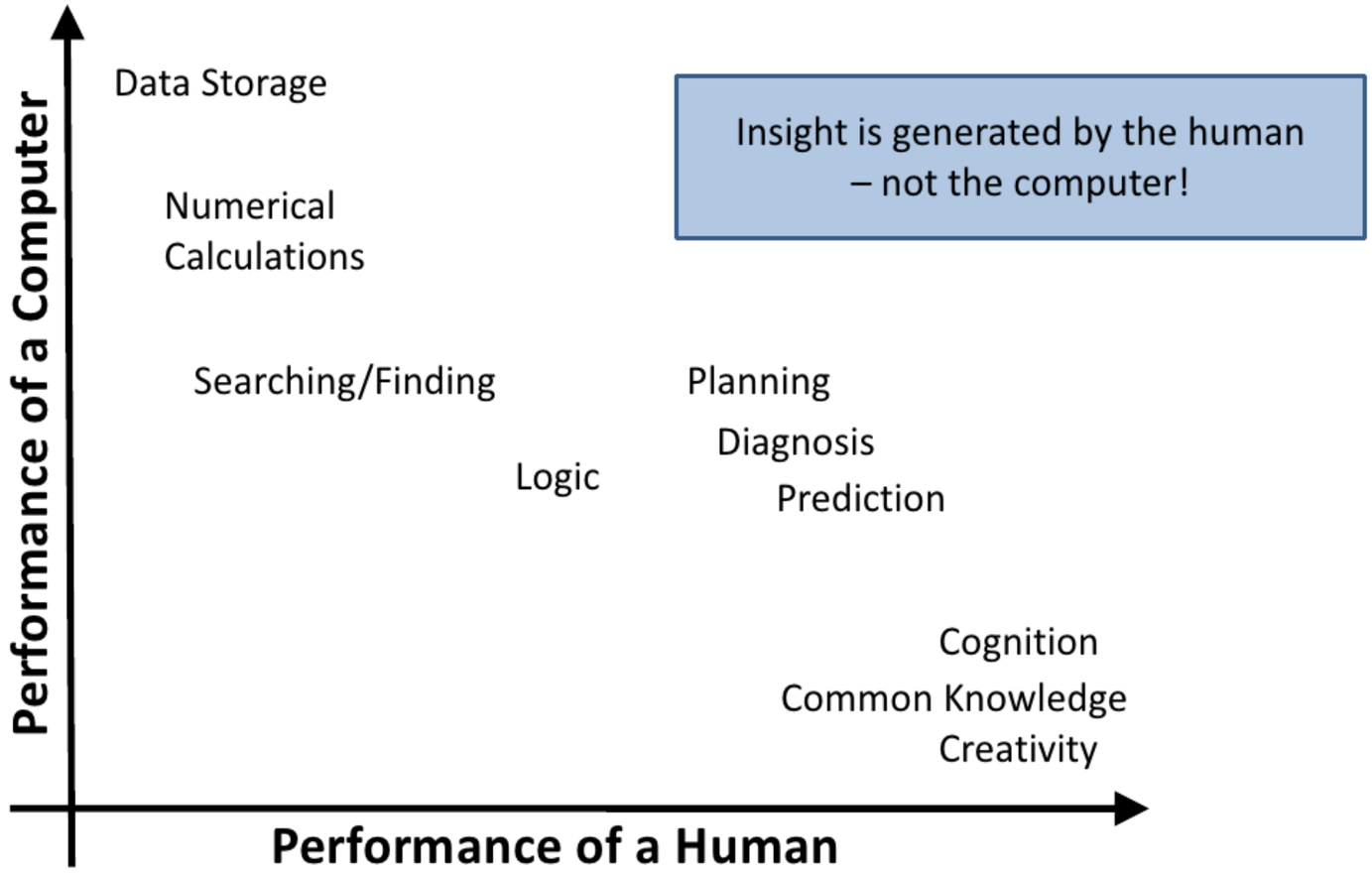
High frequency stock market trading: which stock to buy/sell?

Manufacturing: is bottle broken?

Impractical for human to be involved

Automatic data products

The Ability Matrix



Why Use Computers?

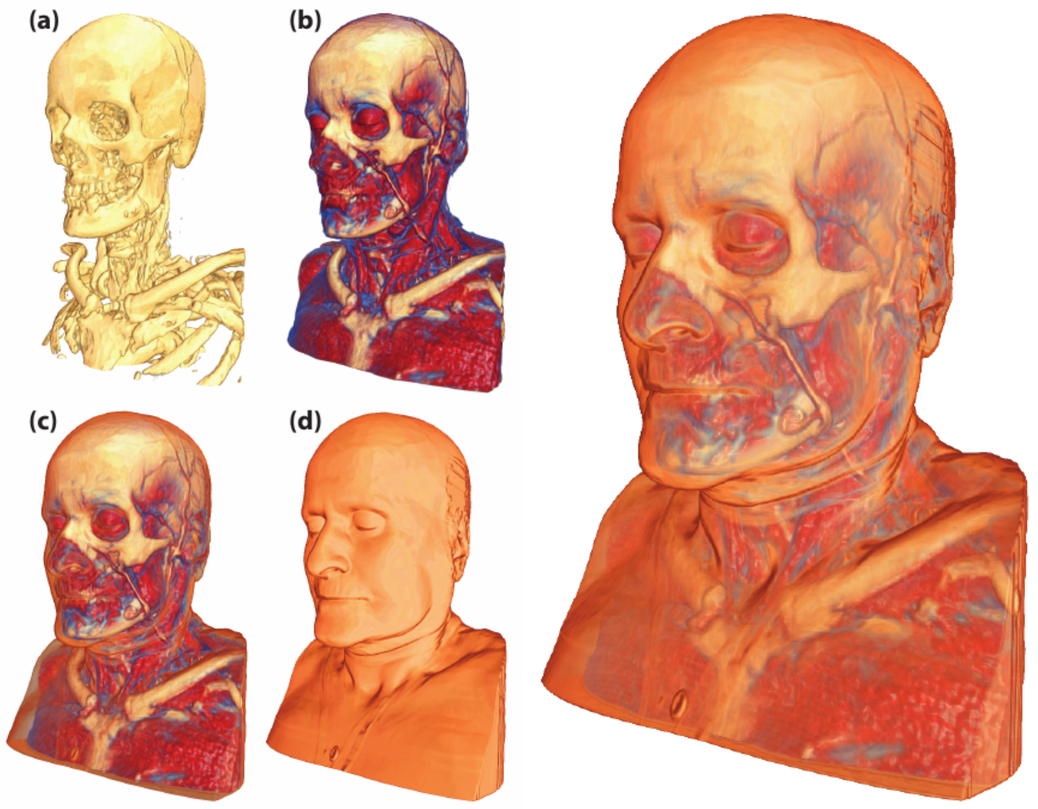
Scale

Drawing by hand (or Illustrator)

infeasible

inflexible (updates!)

How to draw an MRI scan?



[Bruckner 2007]

Why Use Computers?

Interaction

Interaction allows to “drill down” into data

Integration

Integration with algorithms

Make visualization part of a data analysis pipeline

[Sunburst by John Stasko, Implementation in Caleydo by Christian Partl]

Why User Computers?

Efficiency

Re-use charts / methods for
different datasets

Quality

Precise data driven rendering

Storytelling

Use time

Tell Stories

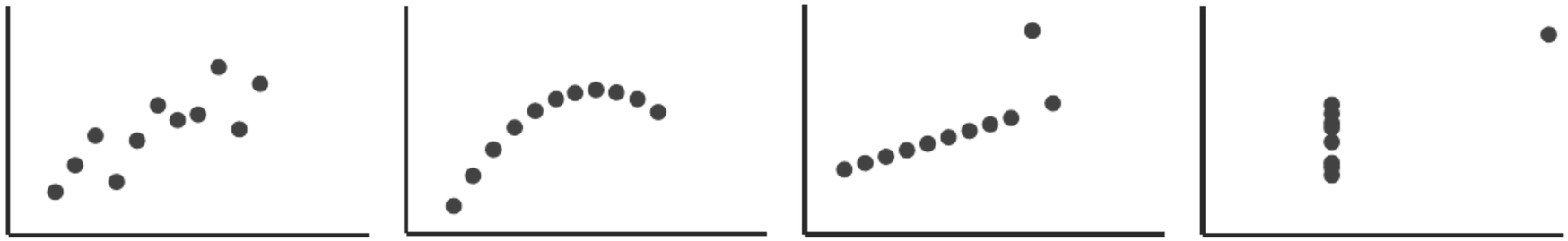
[New York Times]

Why not just use Statistics?

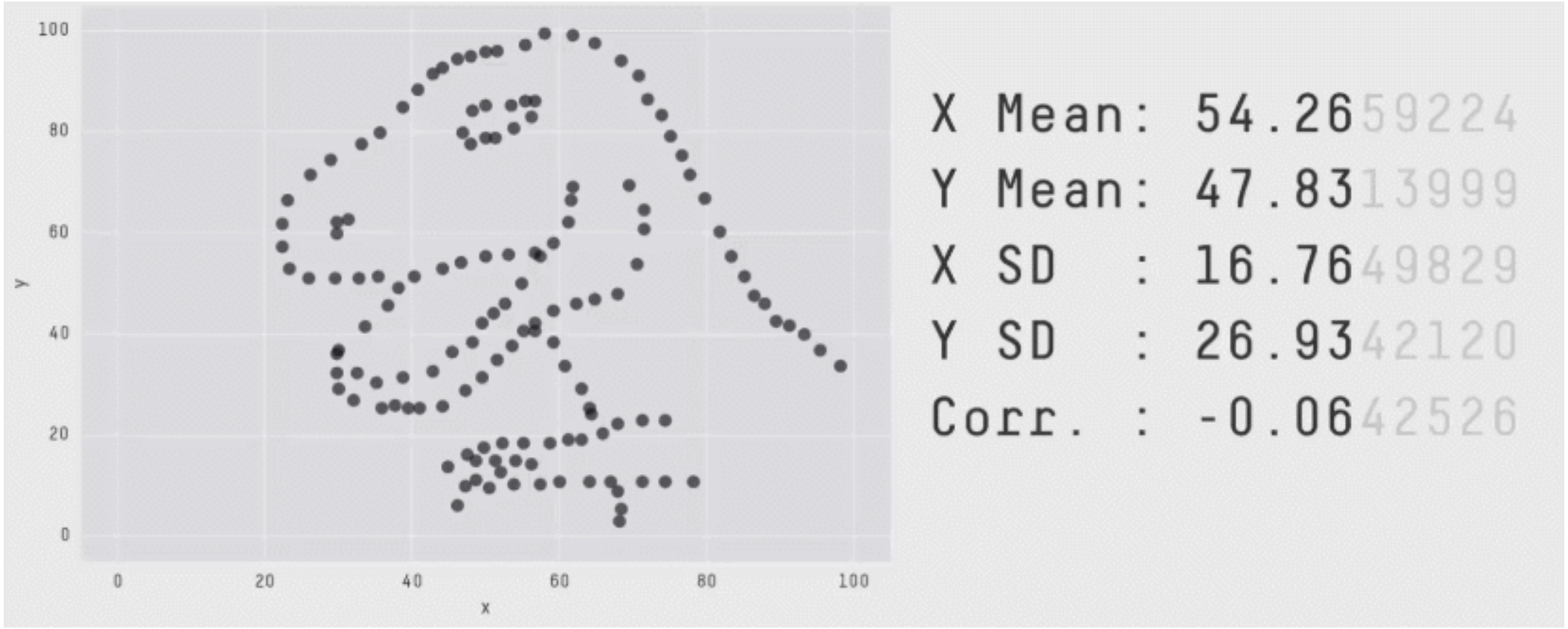
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.4	8	7.9
5	5.						6.8

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x

Anscombe’s Quartett



Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x



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

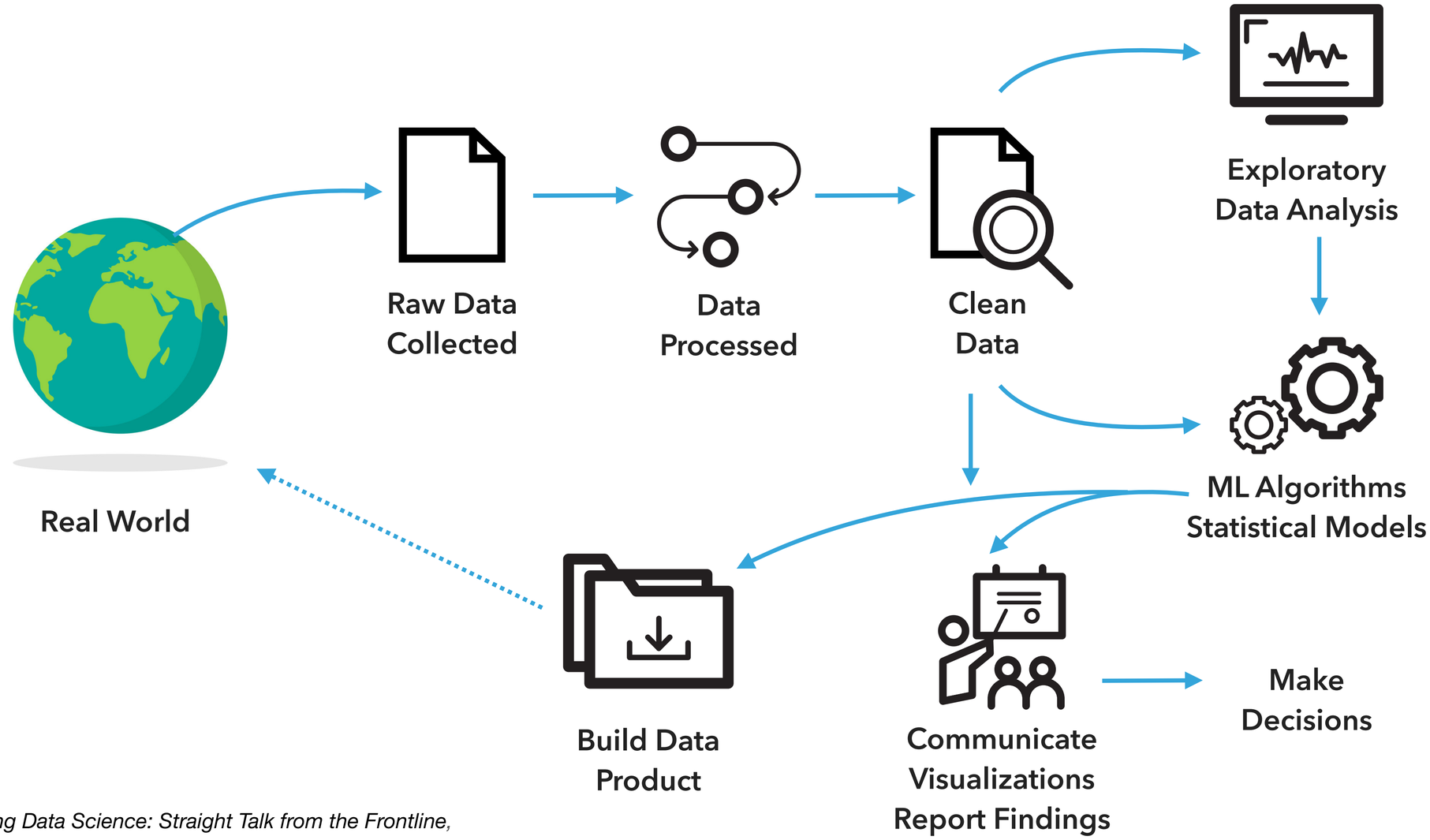
Visualization =

Human Data Interaction

Data

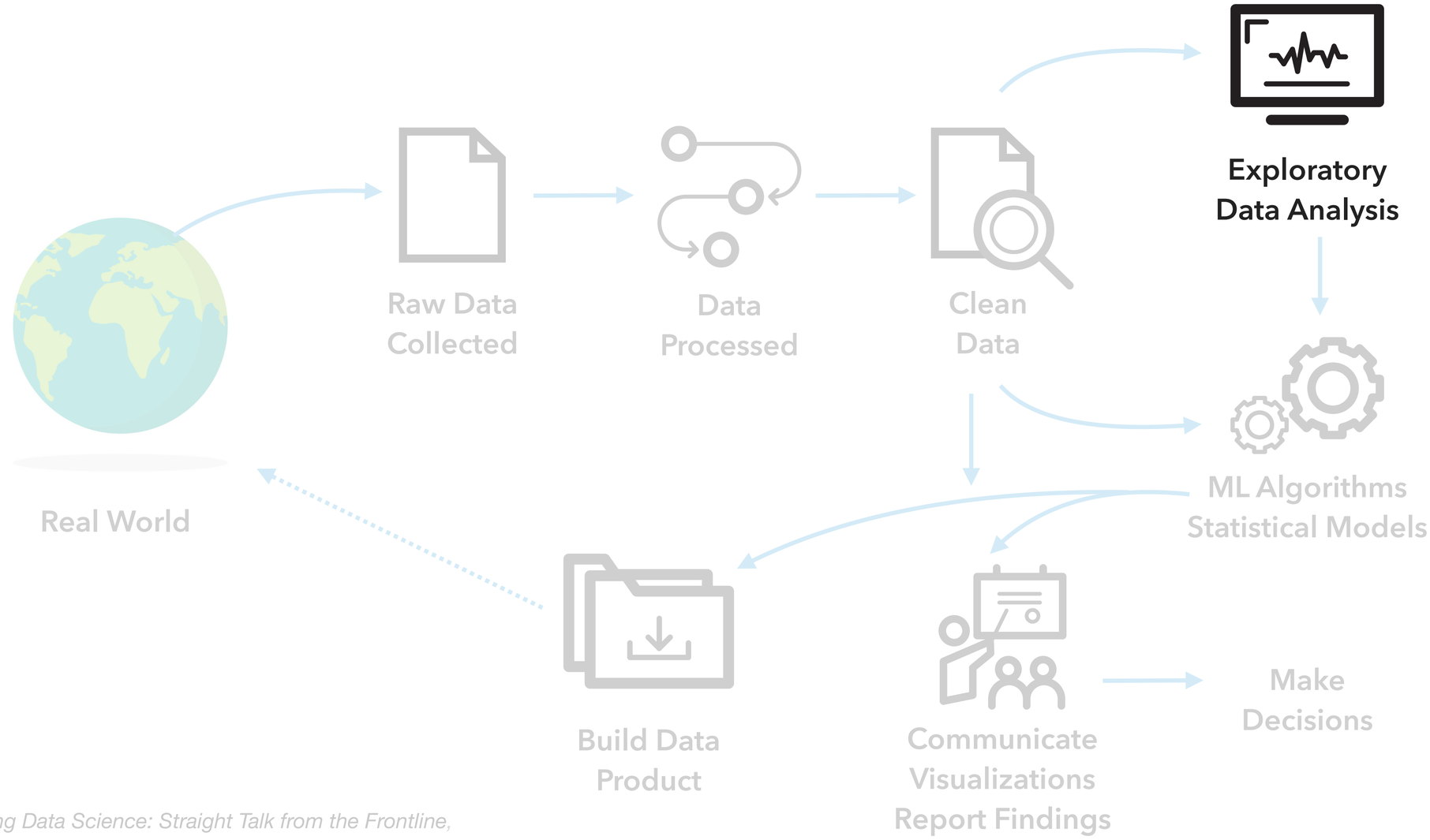
Human-Data Interaction

What's the Role of Visualization in Data Science?



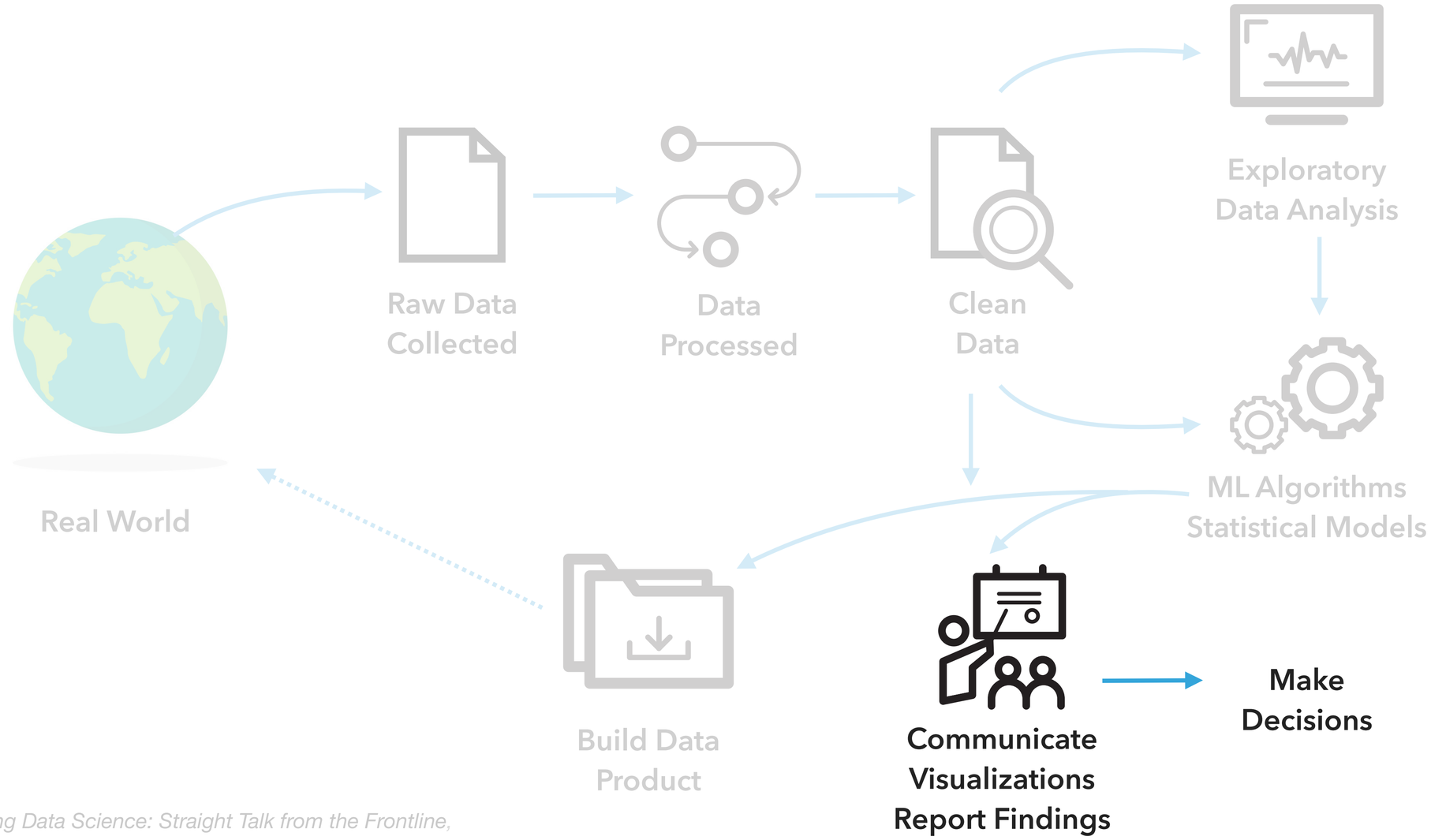
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

What's the Role of Visualization in Data Science?



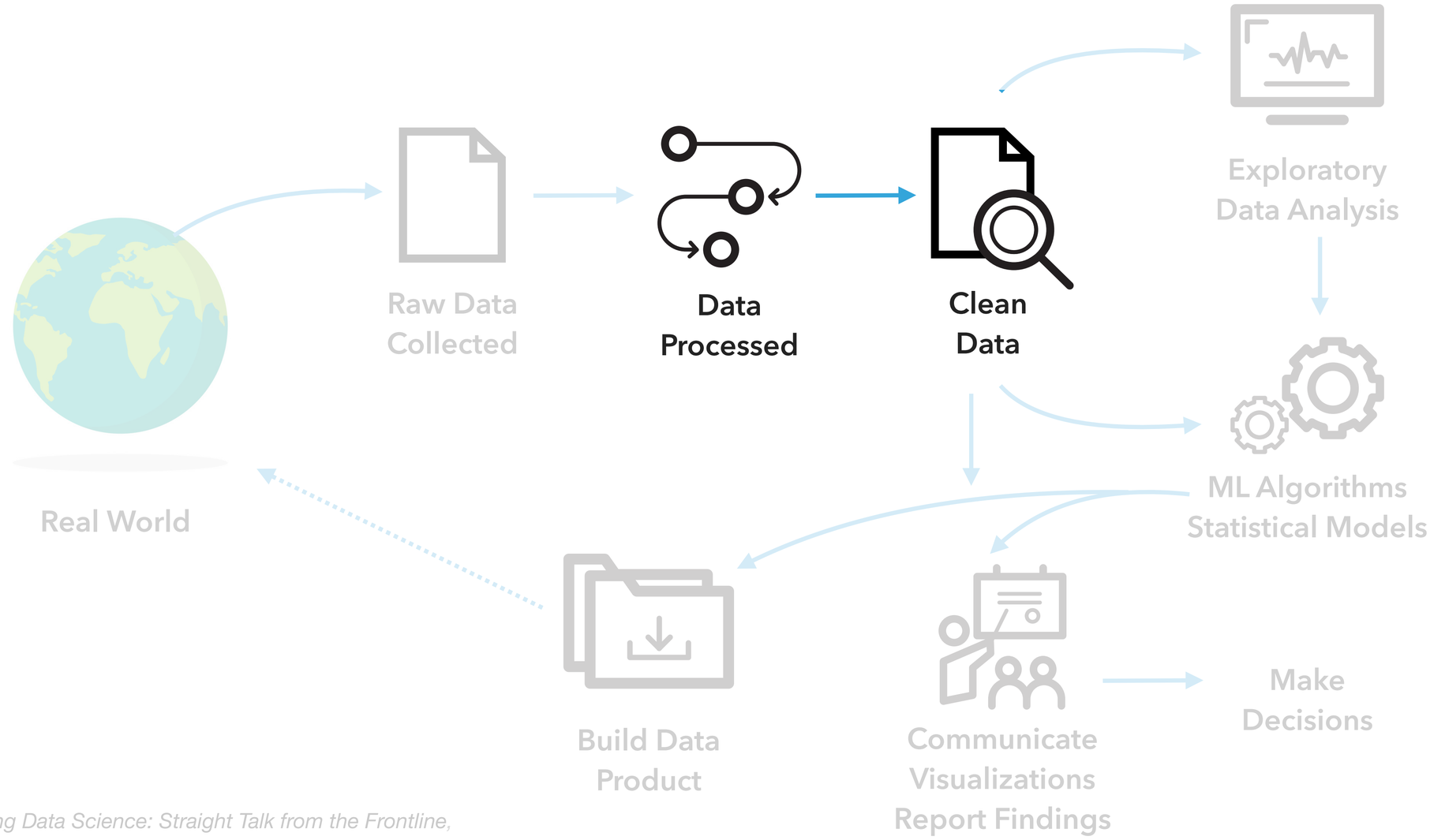
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

What's the Role of Visualization in Data Science?



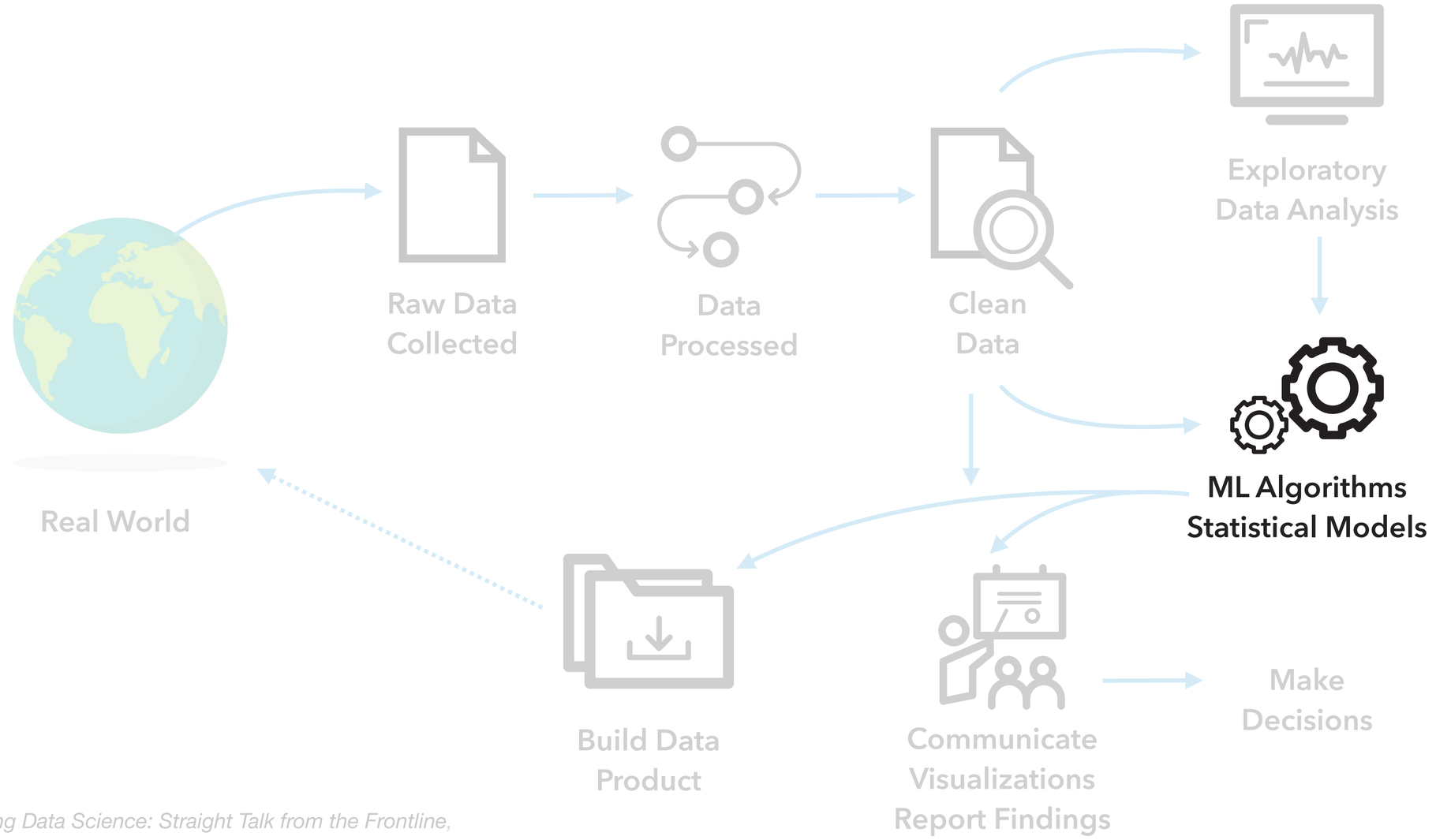
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

What's the Role of Visualization in Data Science?



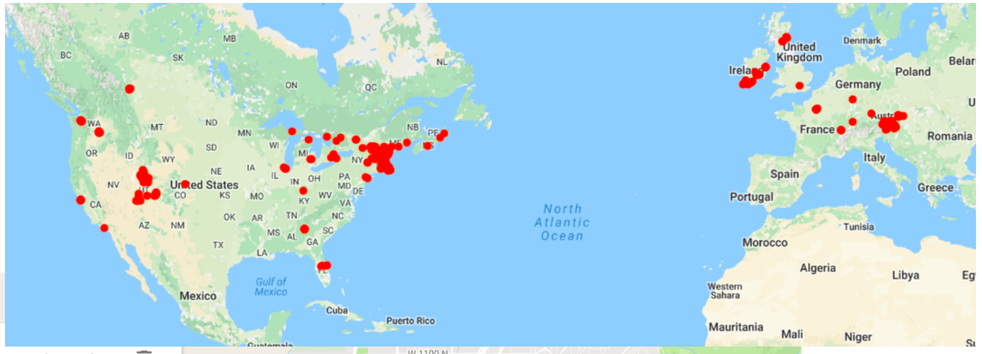
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

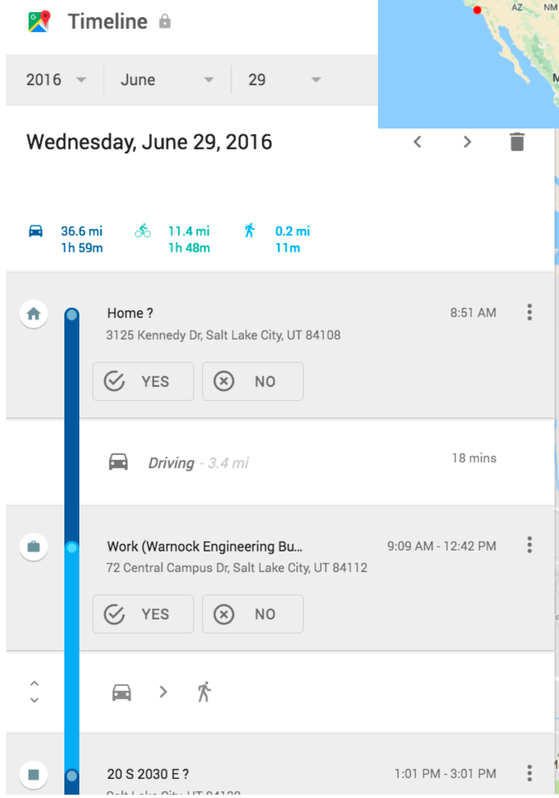
What's the Role of Visualization in Data Science?

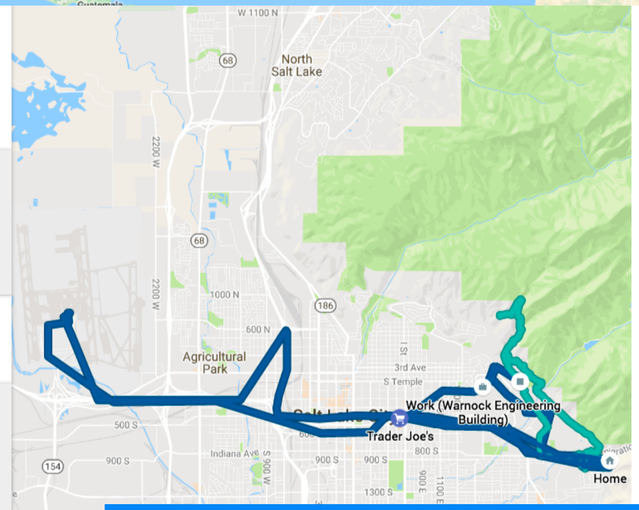


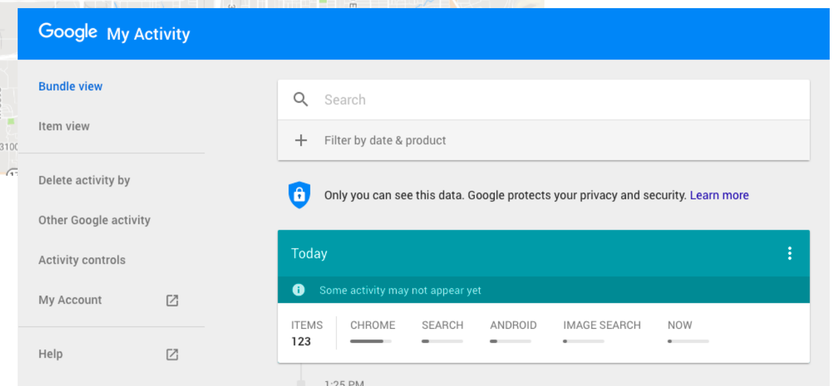
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

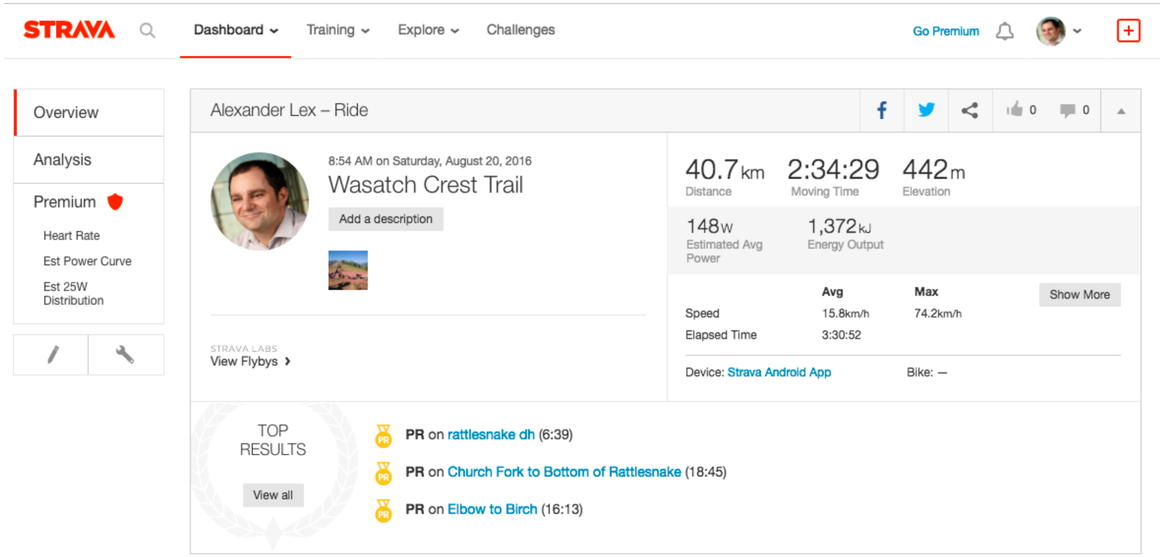
Example: Personal Data

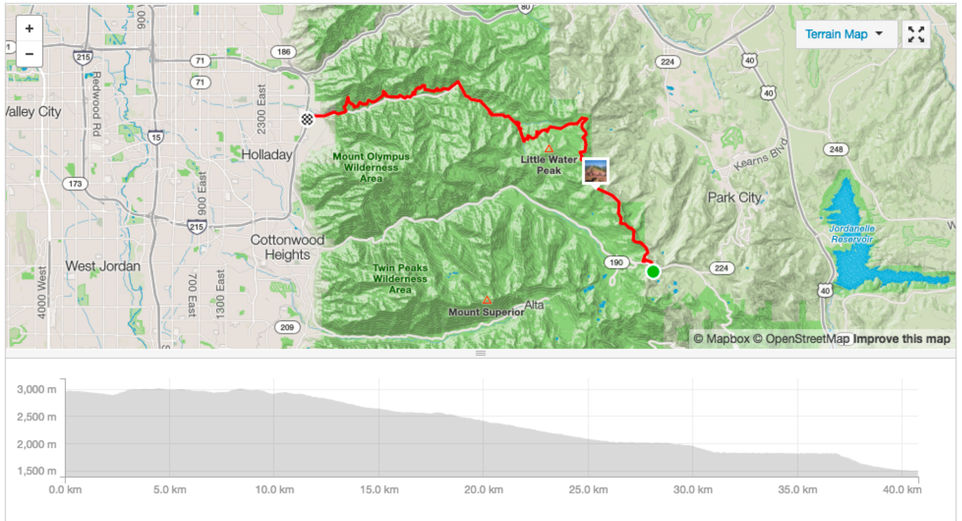












Big Data in Science and Engineering

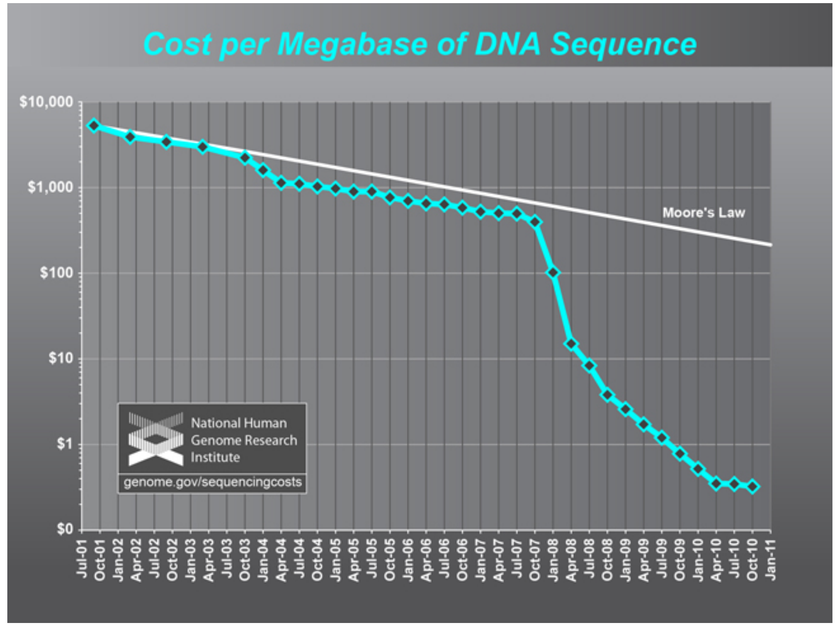
“Big Data” has transformed science and engineering.

Cheap sensors (e.g. imaging) have changed everything

Examples:

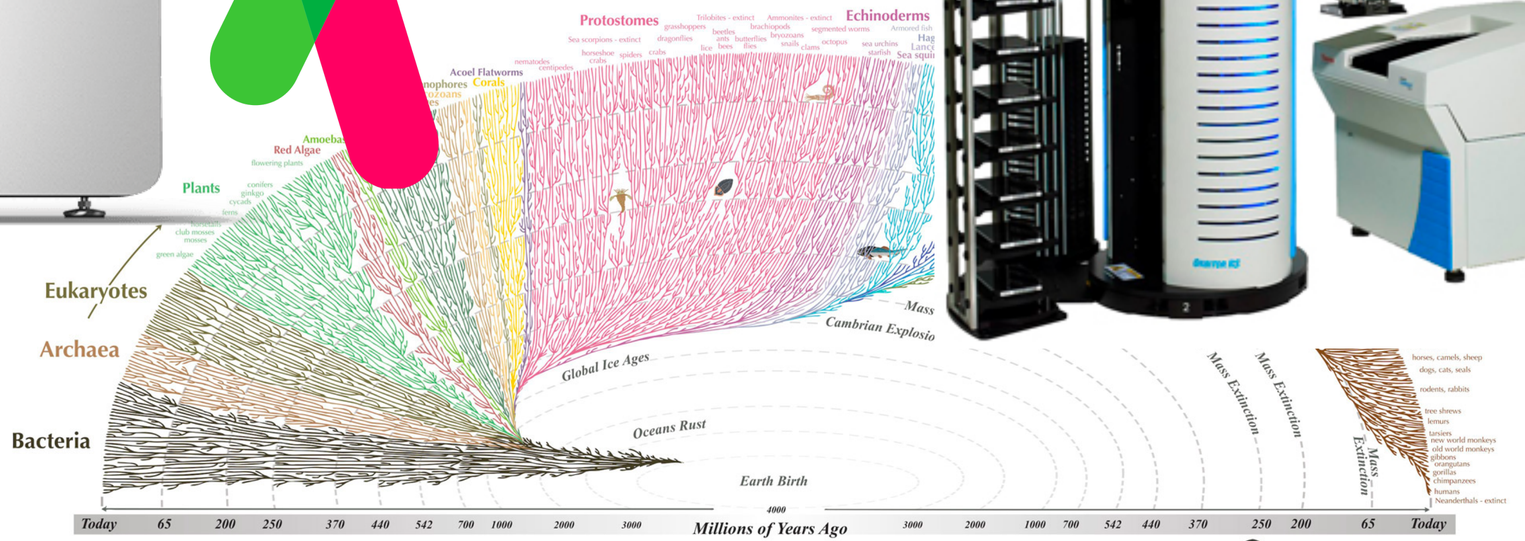
- Large physics experiments and observations
- Cheaper and automated genome sequencing
- Smart buildings / cities (blyncsy)

Controversy: Hypothesis or data driven methods



Example: Genomics

23andMe



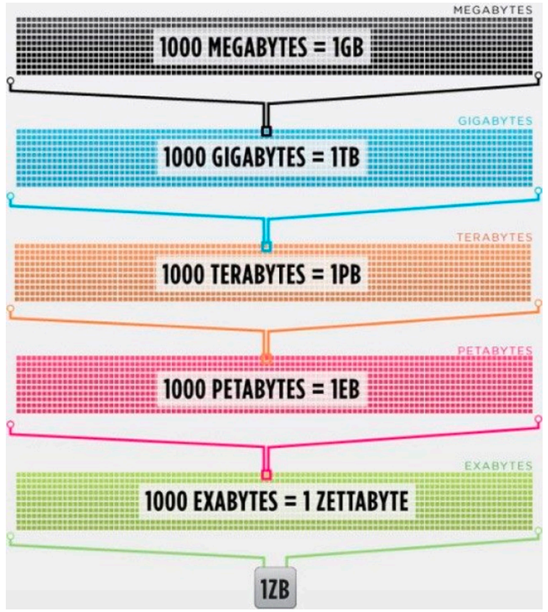
Example TCGA: 1 Petabyte

NSA Utah Data Center (Bluffdale, Utah)

Storage Capacity?

estimates vary,

NPR/NG estimate five zettabytes of data



“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

Humans!

Human Data Interaction

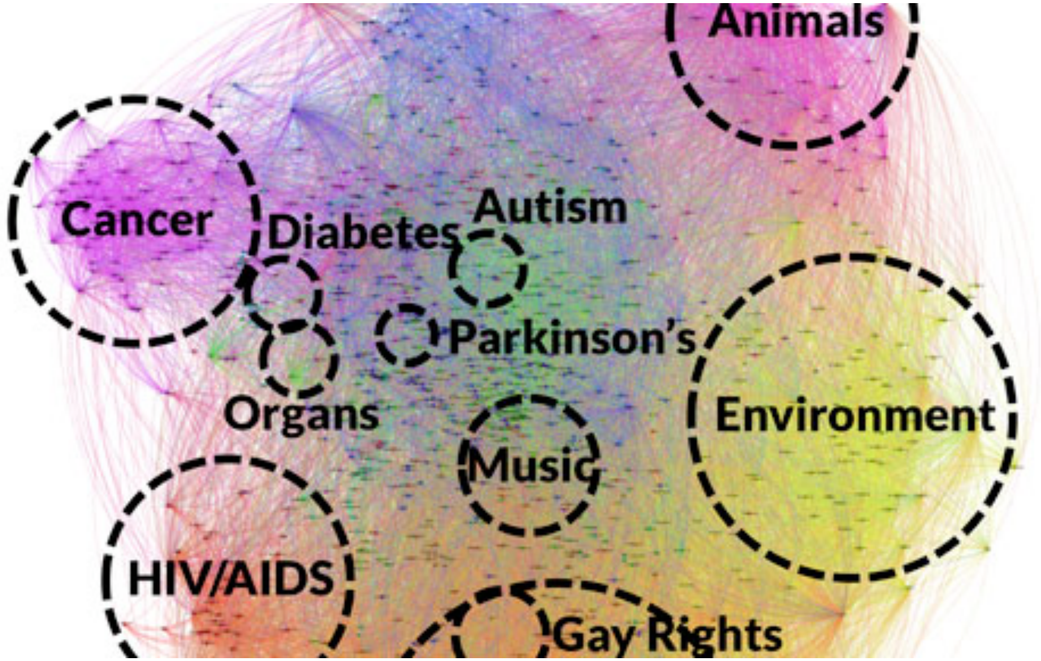
Why Humans?

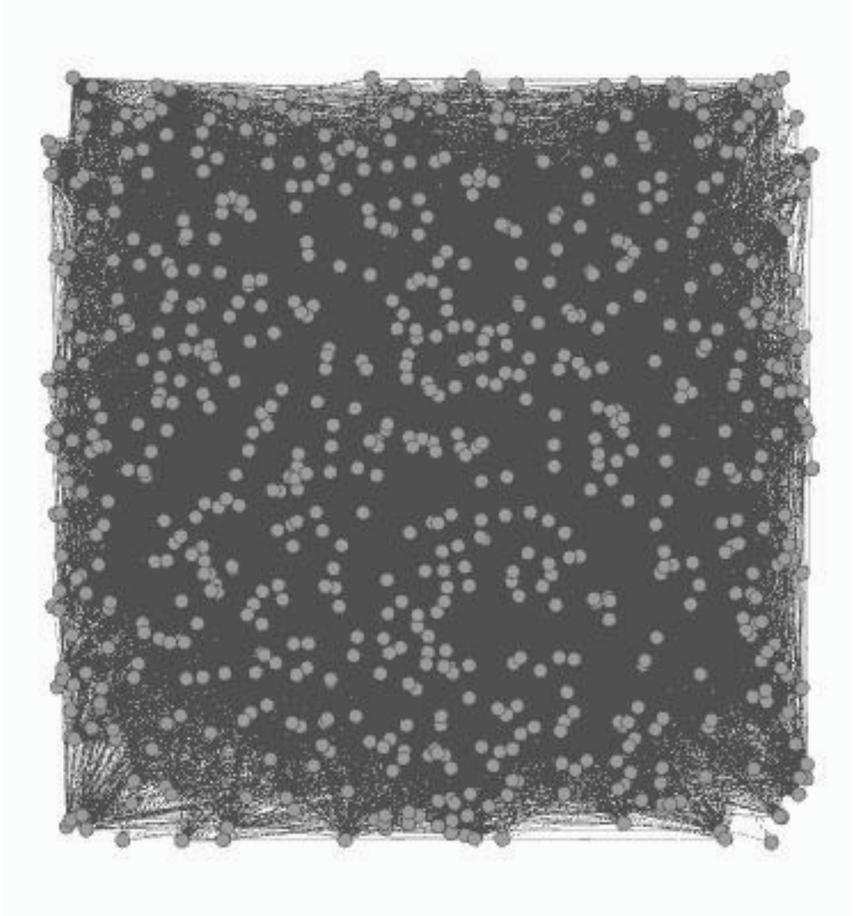
Leveraging human capabilities

- **Pattern Discovery**: clusters, outliers, trends
- **Contextual Knowledge**: expectations for dataset, explanations for patterns
- **Action**: humans learn and take action

But: we also have to **design for Humans and their limitations**

Not everything that can be drawn can be read!





Limits of Cognition

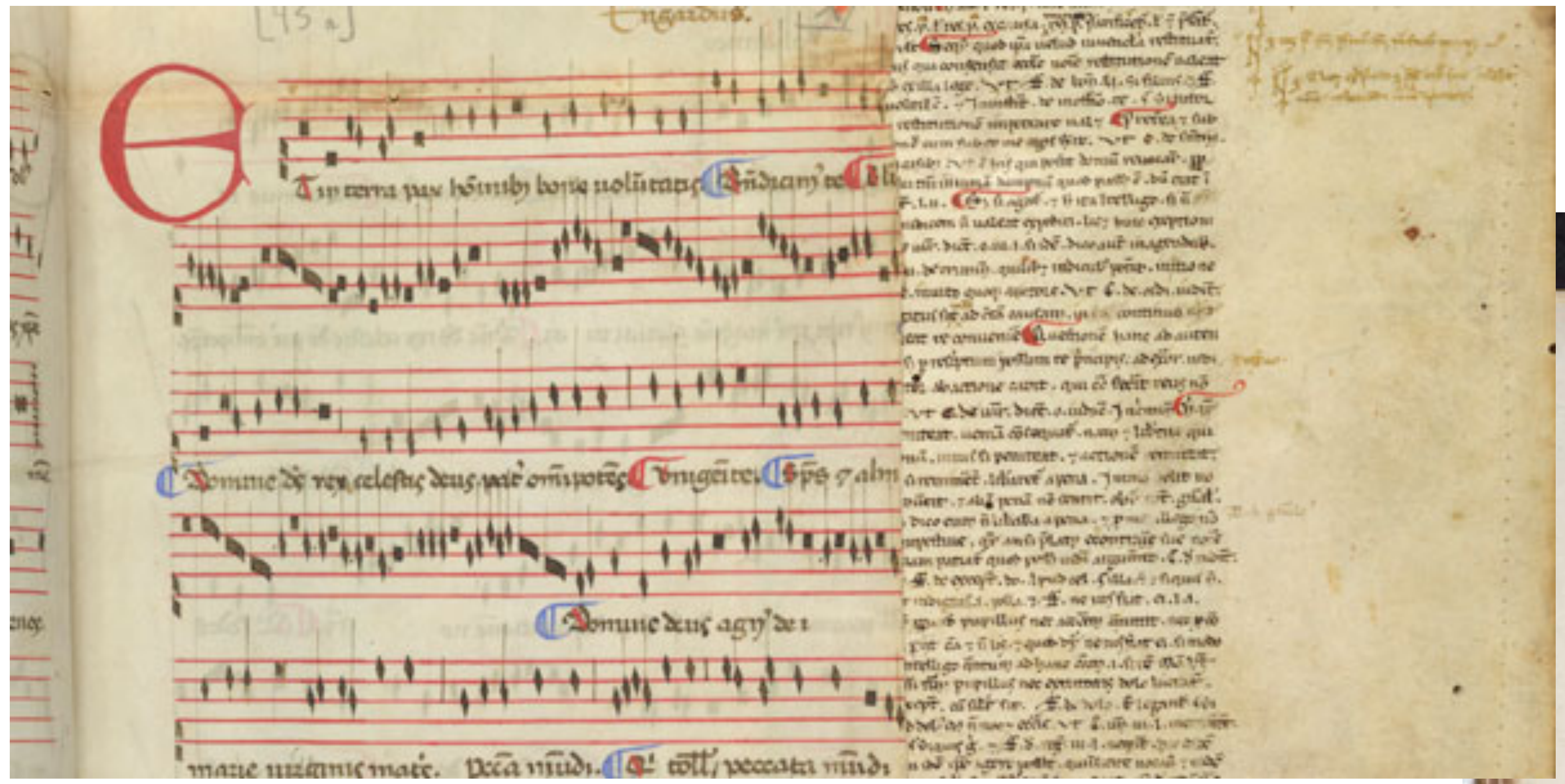
Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

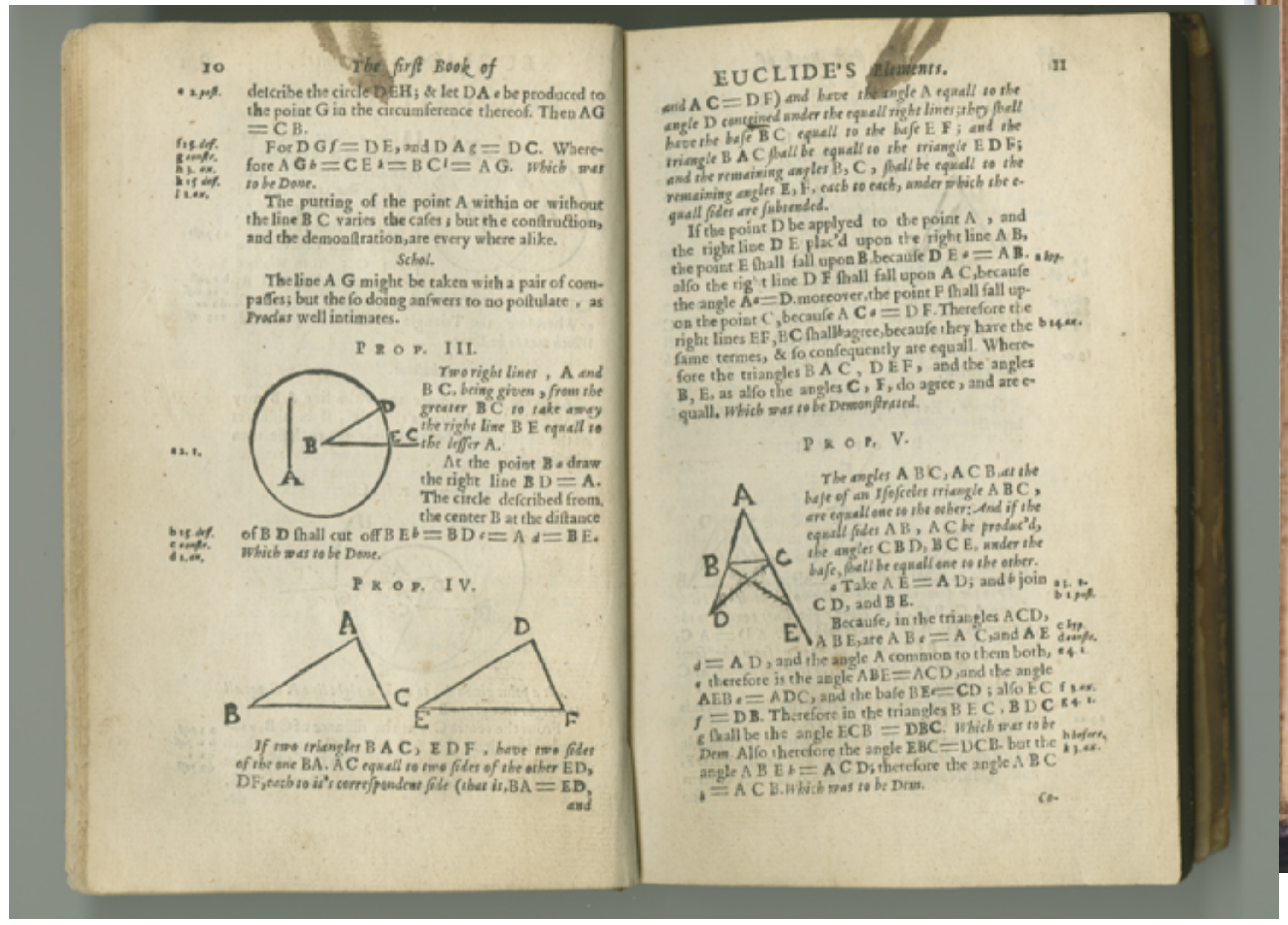
How did we get here?

A bit of history

“It is things that make us smart”

Donald A. Norman





10 *The first Book of*

describe the circle DEH; & let DA e be produced to the point G in the circumference thereof. Then AG = CB.

For DG f = DE, and DA g = DC. Wherefore AG h = CE k = BC l = AG. *Which was to be Done.*

The putting of the point A within or without the line BC varies the cases; but the construction, and the demonstration, are every where alike.

Schol.

The line AG might be taken with a pair of compasses; but the so doing answers to no postulate, as *Proclus* well intimates.

PROP. III.

Two right lines, A and BC being given, from the greater BC to take away the right line BE equall to the lesser A.

At the point B a draw the right line BD = A. The circle described from the center B at the distance of BD shall cut off BE b = BD c = A d = BE. *Which was to be Done.*

PROP. IV.

If two triangles BAC, EDF, *have two sides of the one* BA. AC *equall to two sides of the other* ED, DF, *each to it's correspondent side (that is,* BA = ED, *and*

EUCLIDE'S *Elements.* 11

and AC = DF) *and have the angle A equall to the angle D contained under the equall right lines; they shall have the base* BC *equall to the base* EF; *and the triangle* BAC *shall be equall to the triangle* EDF; *and the remaining angles* B, C, *shall be equall to the remaining angles* E, F, *each to each, under which the equall sides are subtended.*

If the point D be applyed to the point A, and the right line DE plac'd upon the right line AB, the point E shall fall upon B, because DE a = AB. also the right line DF shall fall upon AC, because the angle A = D. moreover, the point F shall fall upon the point C, because AC = DF. Therefore the right lines EF, BC shall agree, because they have the same termes, & so consequently are equall. Wherefore the triangles BAC, DEF, and the angles B, E, as also the angles C, F, do agree, and are equall. *Which was to be Demonstrated.*

PROP. V.

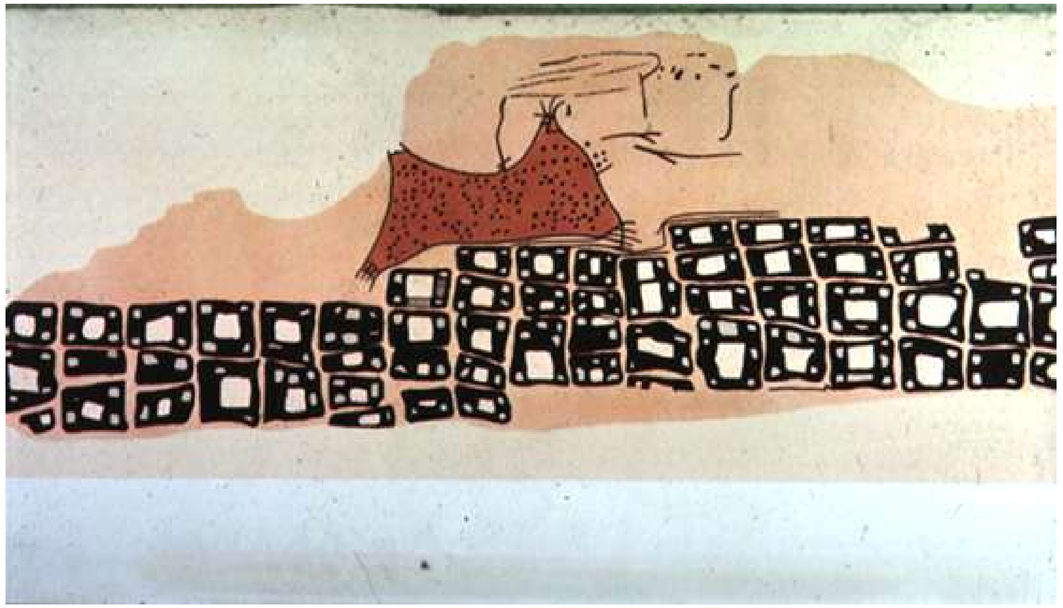
The angles ABC, ACB *at the base of an Isosceles triangle* ABC, *are equall one to the other: And if the equall sides* AB, AC *be produc'd, the angles* CBD, BCE, *under the base, shall be equall one to the other.*

a Take AE = AD; and b join CD, and BE.

Because, in the triangles ACD, ABE, are AB c = AC, and AE d = AD, and the angle A common to them both, e therefore is the angle ABE = ACD, and the angle AEB e = ADC, and the base BE = CD; also EC f = DB. Therefore in the triangles BEC, BDC g shall be the angle ECB = DBC. *Which was to be Dem.* Also therefore the angle EBC = DCB. but the angle ABE h = ACD; therefore the angle ABC i = ACB. *Which was to be Dem.*

Co-

Record



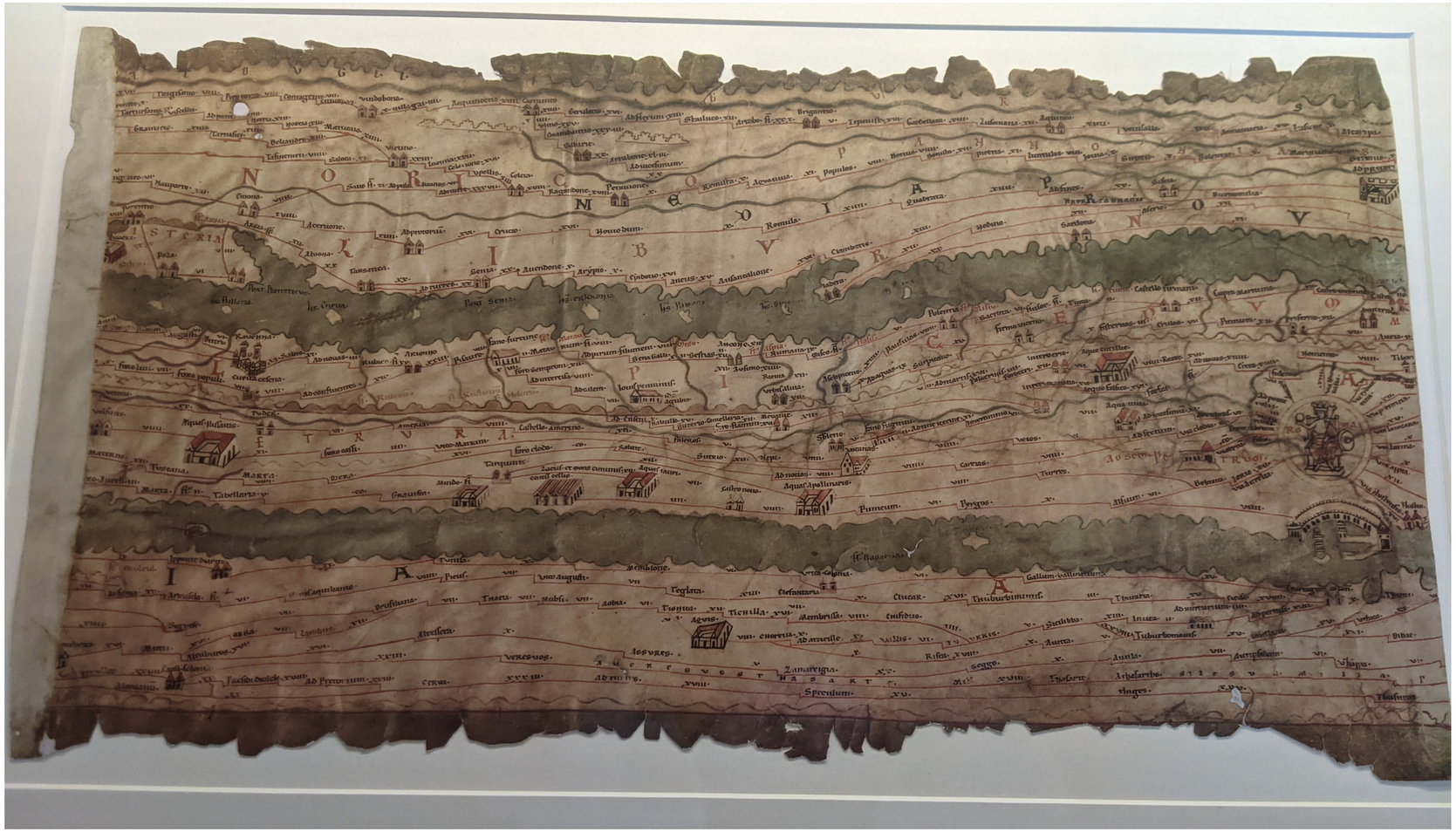
Konya town map, Turkey, c. 6200 BC

Anaximader's Map of the World

Anaximander of Miletus, c. 550 BC

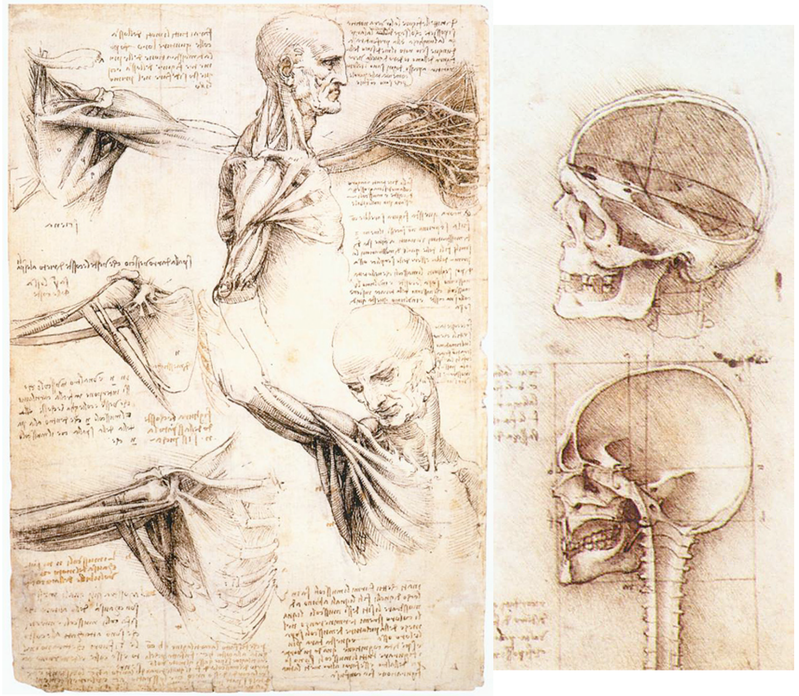
Milestones Project

Tabula Peutingeriana – Roman Road Map

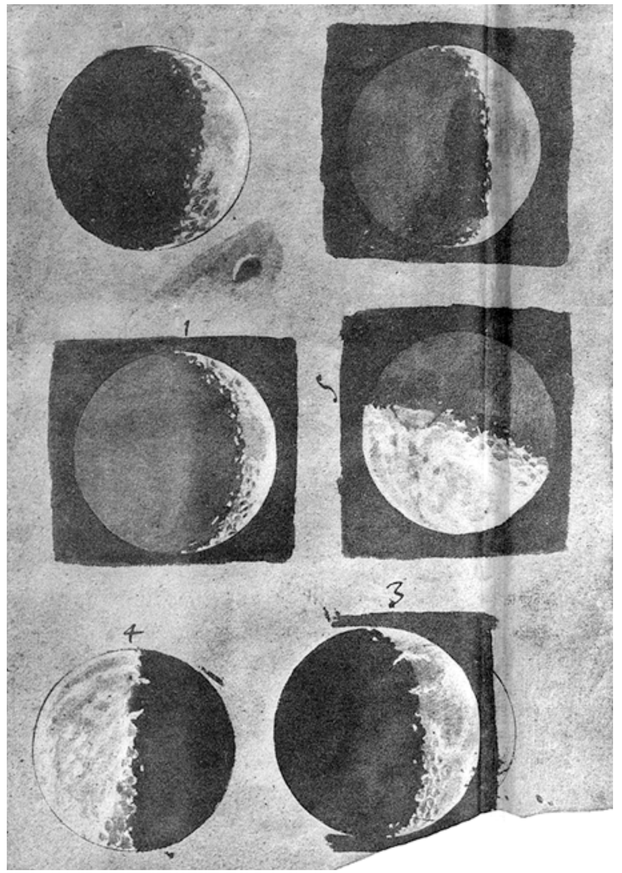


https://en.wikipedia.org/wiki/Tabula_Peutingeriana

Record



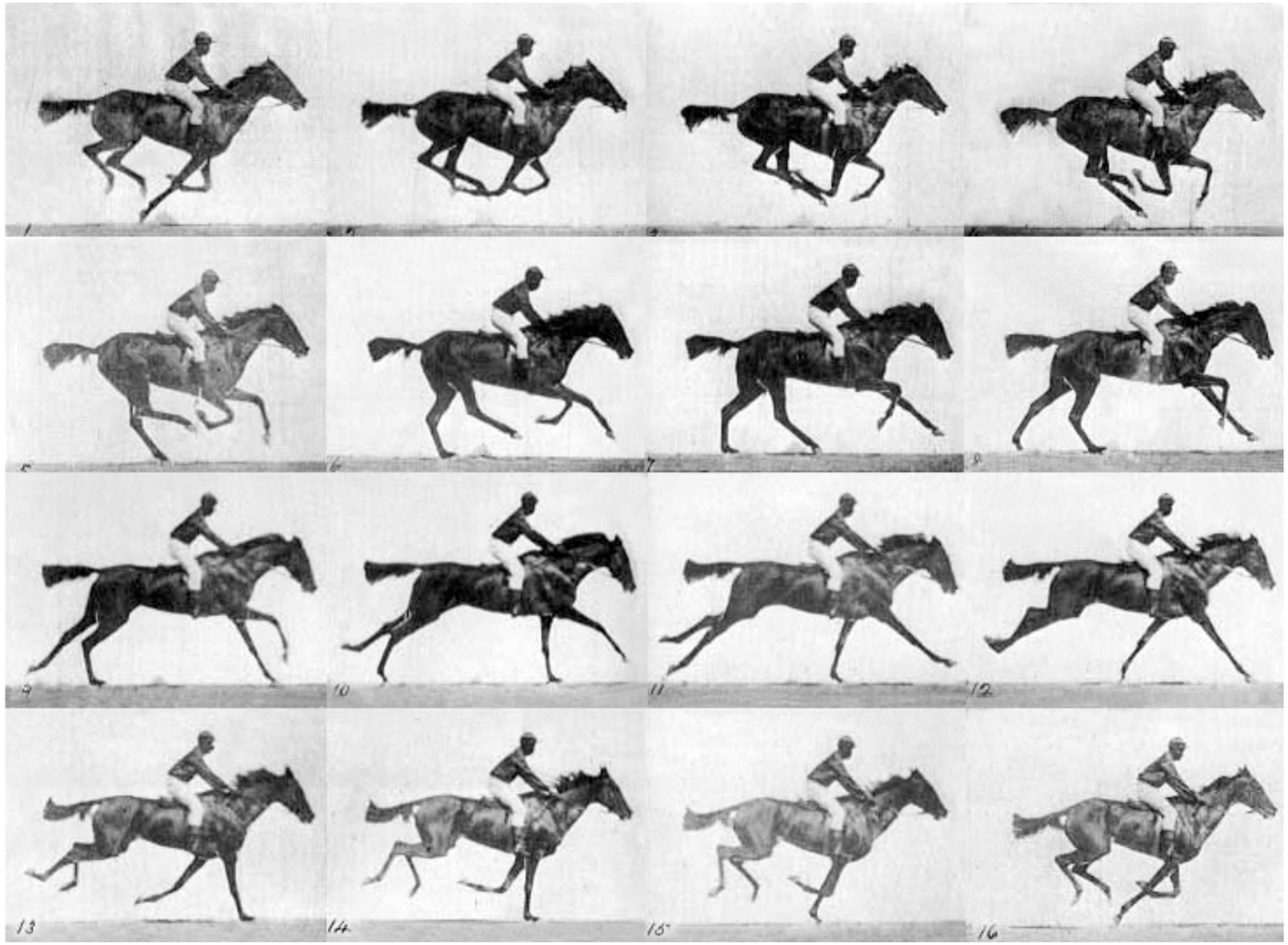
Leonardo Da Vinci, ca. 1500



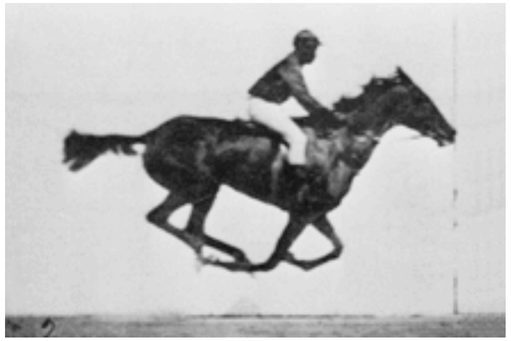
Galileo Galilei, 1616

William Curtis (1746-1799)

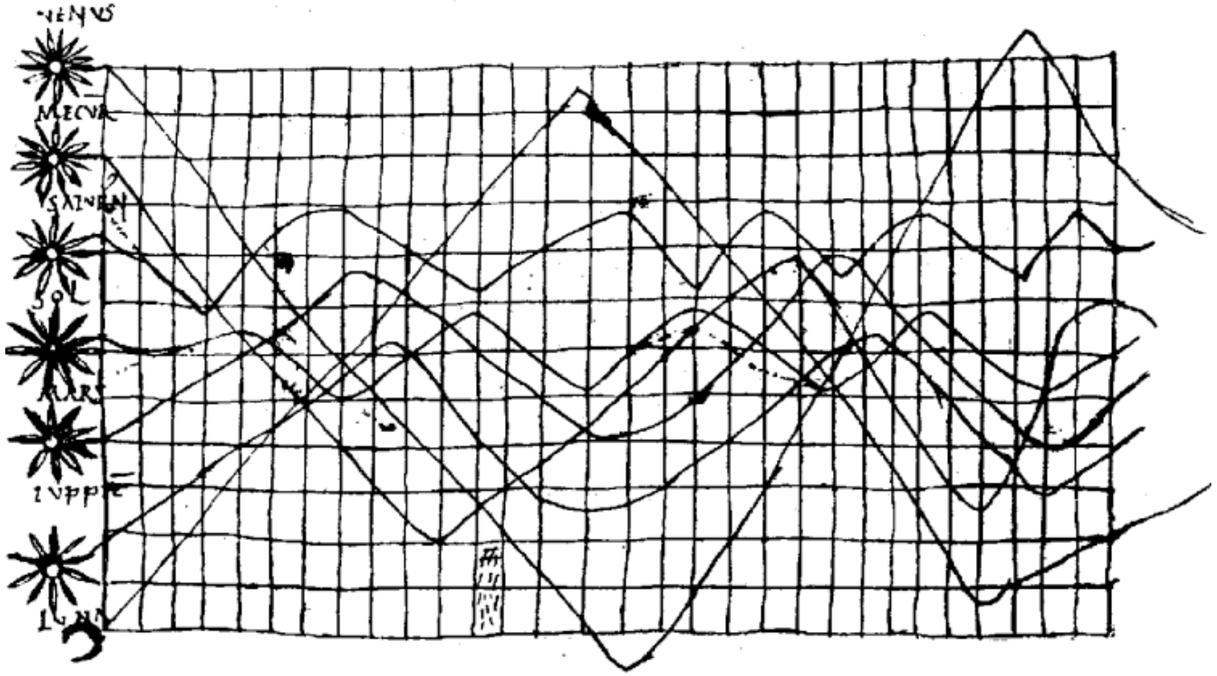
Record



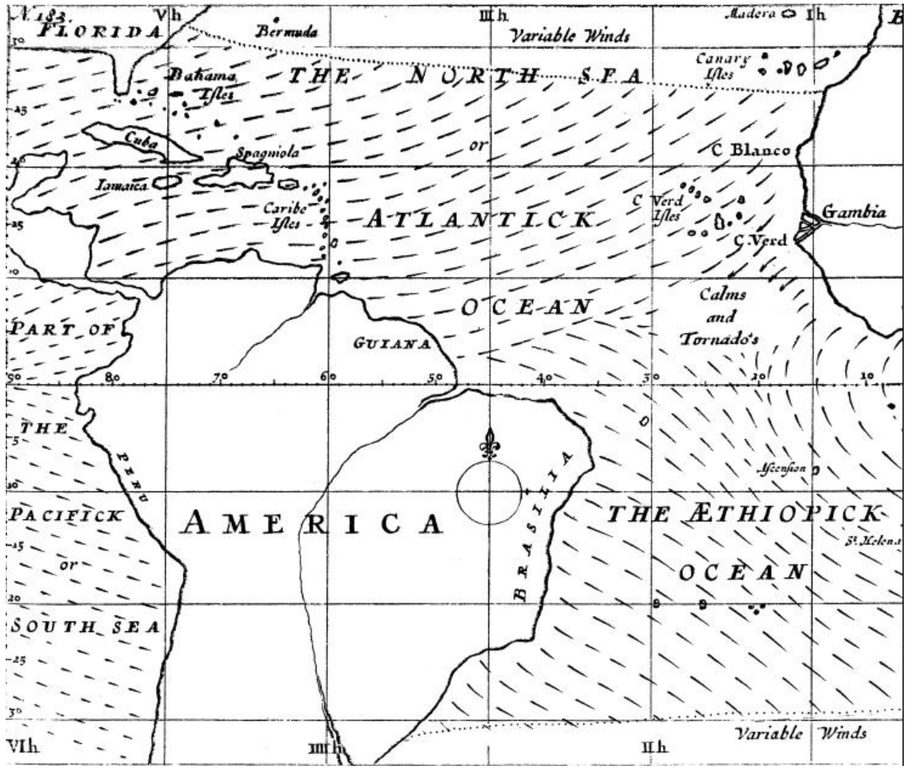
Eadweard J. Muybridge, 1878



Analyze

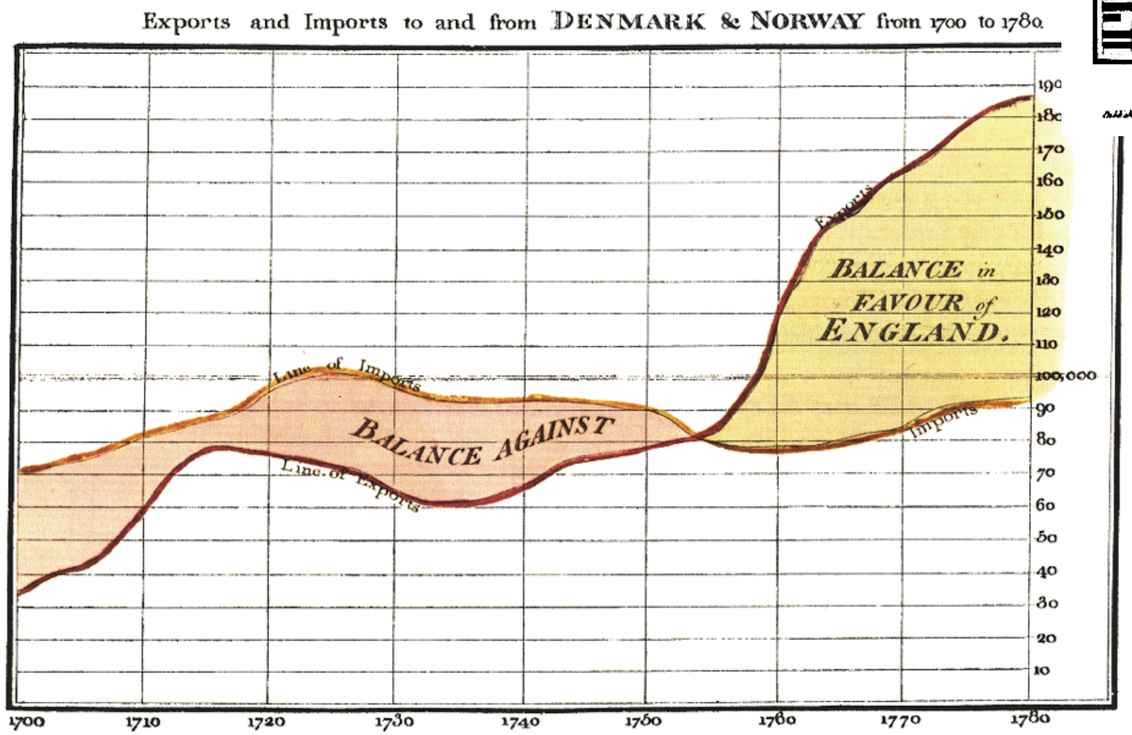


Planetary Movement Diagram, c. 950

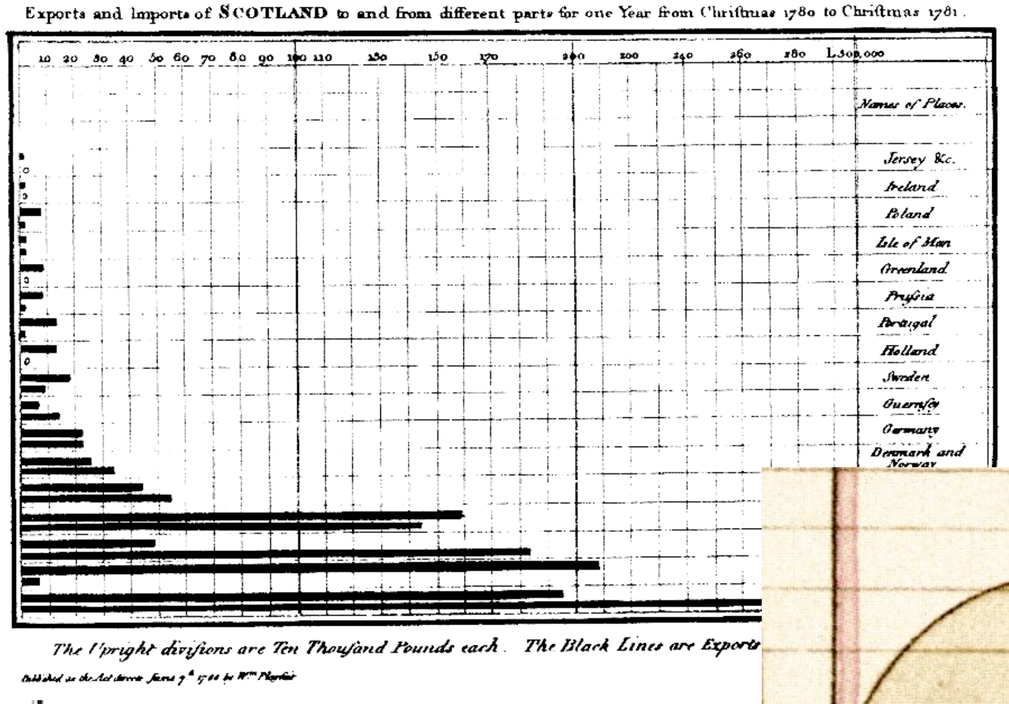


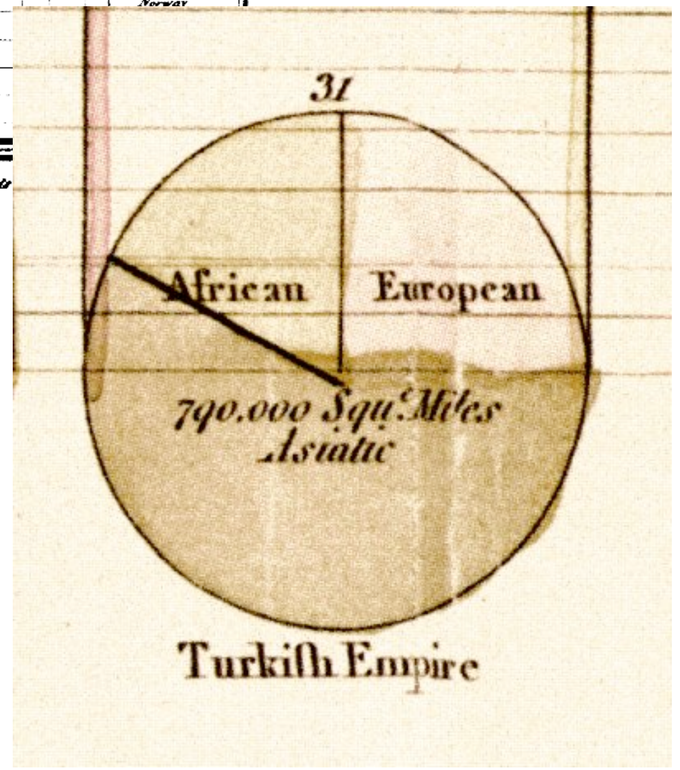
Halley's Wind Map, 1686

Analyze



W. Playfair, 1786



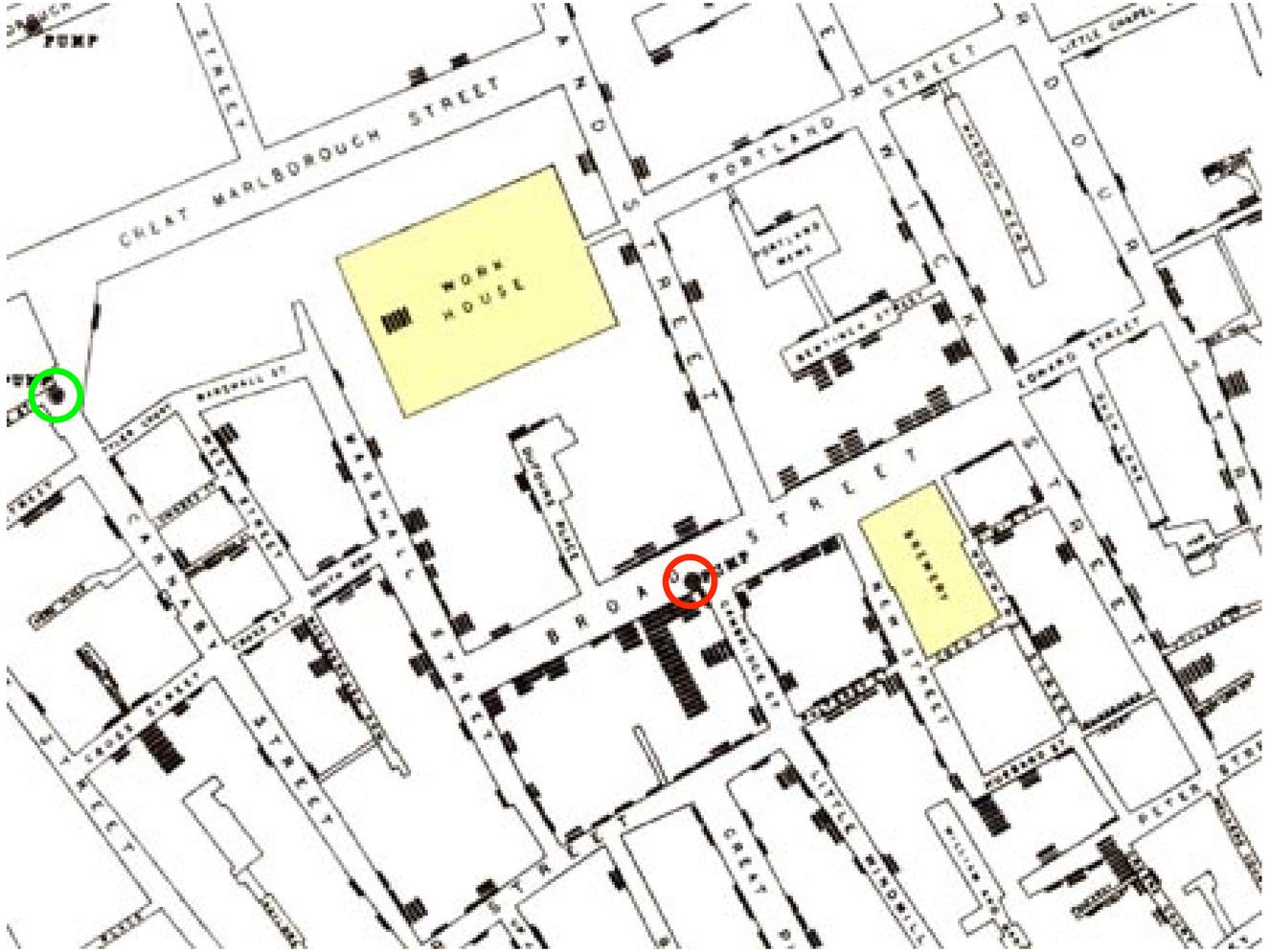


proportions of the Turkish Empire located in Asia, Europe and Africa before 1789

wikipedia.org

W. Playfair, 1801

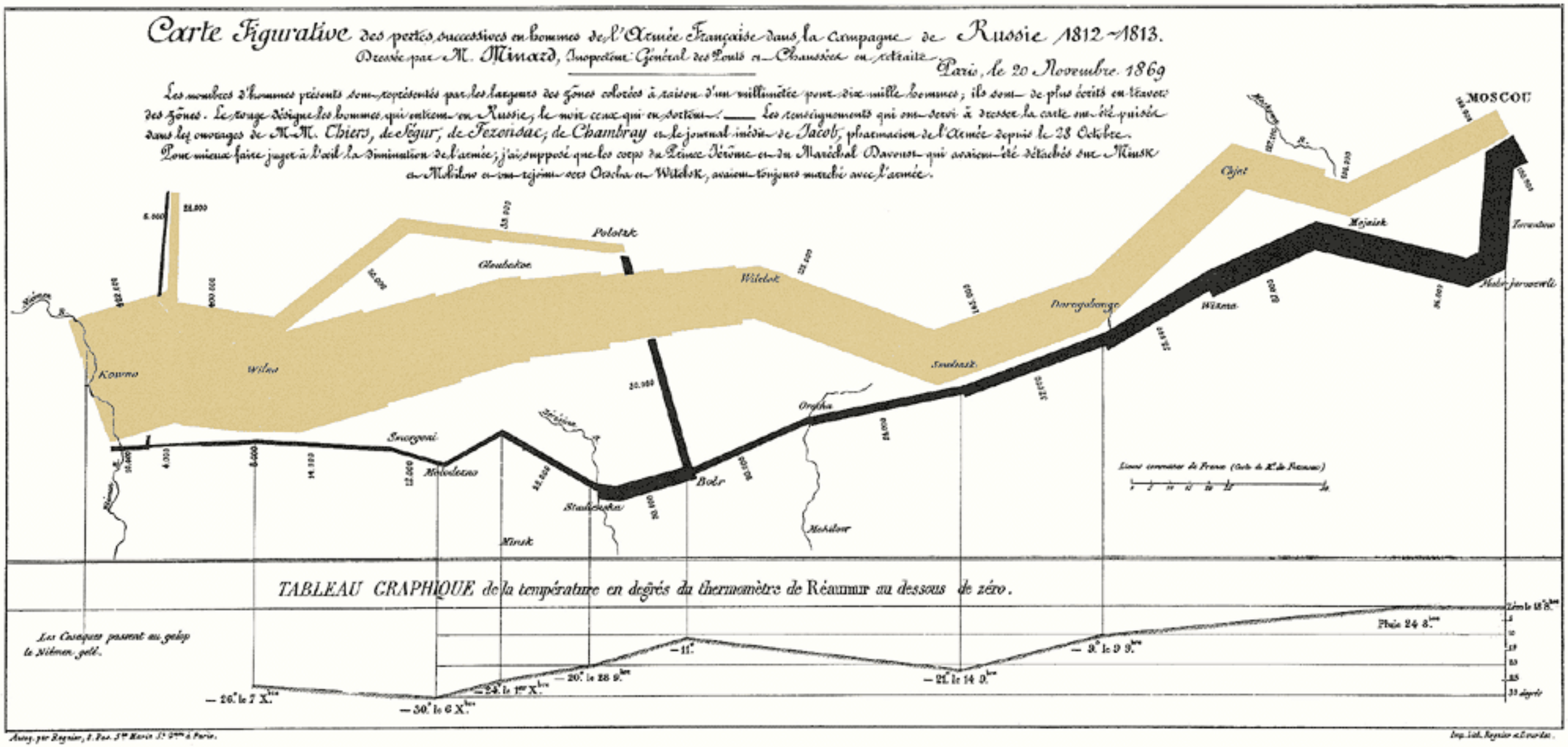
Find Patterns



John Snow, 1854

E. Tufte, Visual Explanations, 1997

Communicate

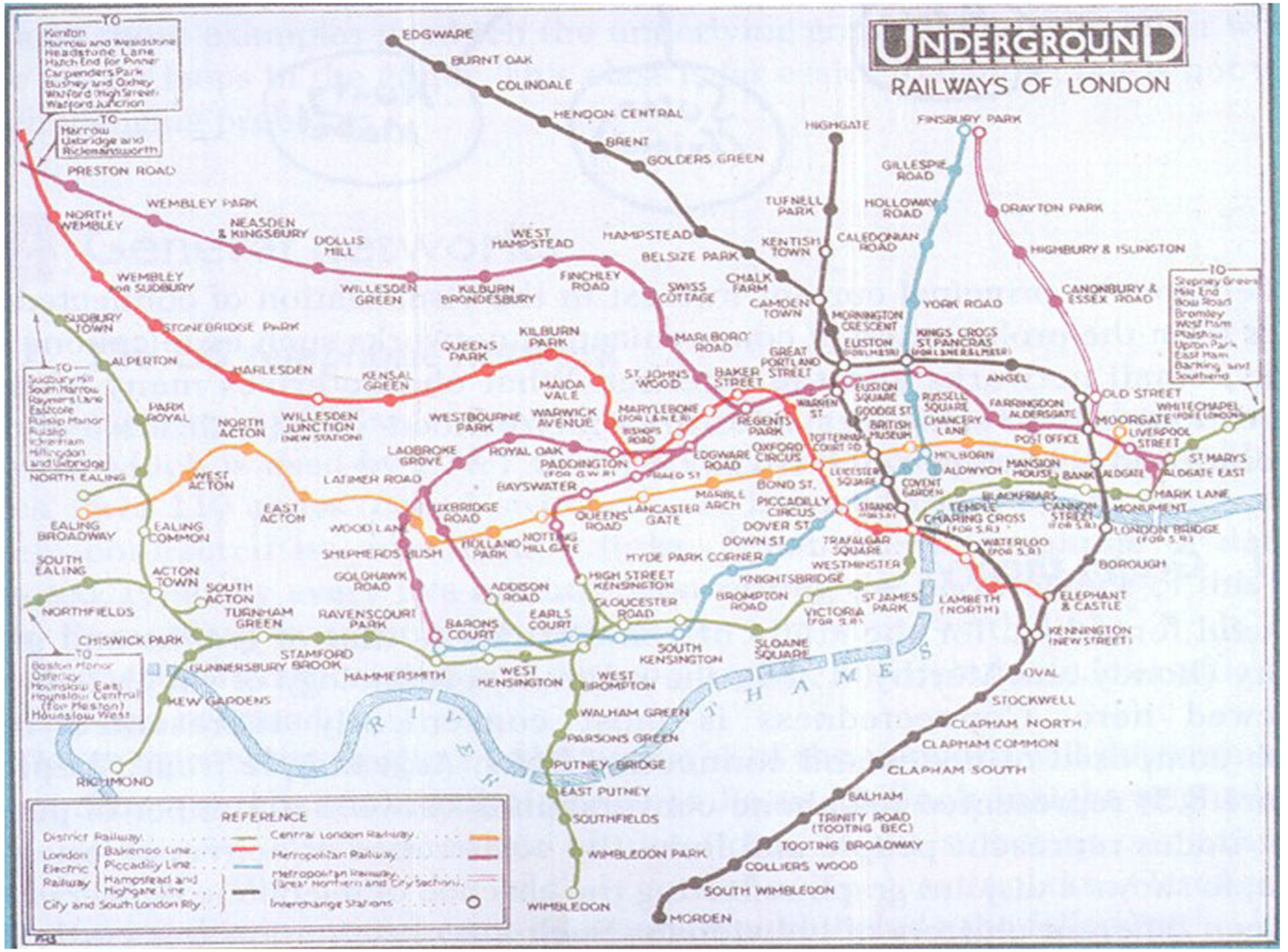


C.J. Minard, 1869

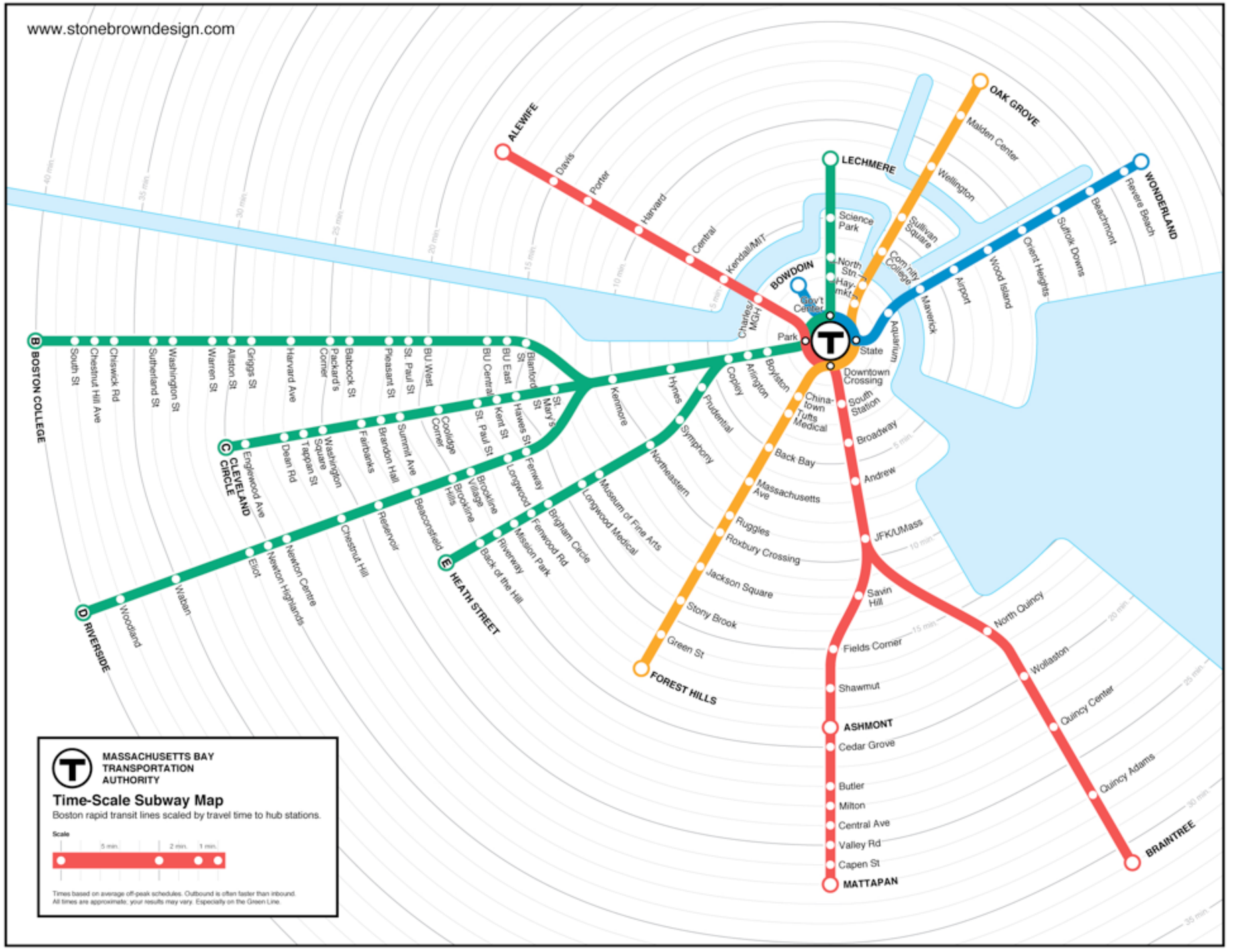
E. Tufte, Writings, Artworks, News

https://history.infowetrust.com/

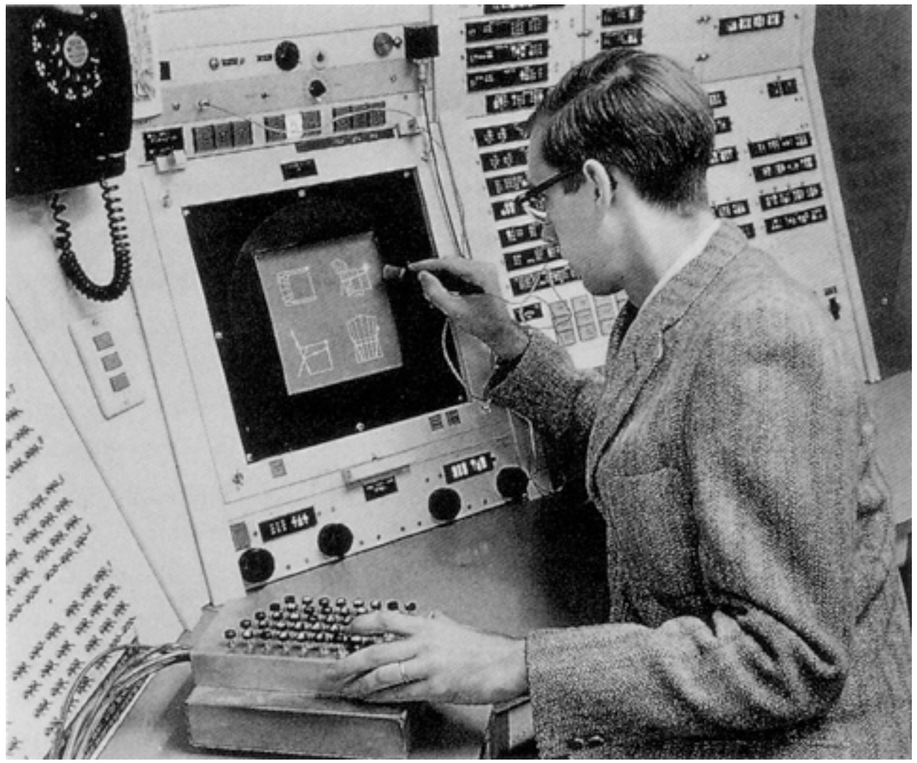
Communicate



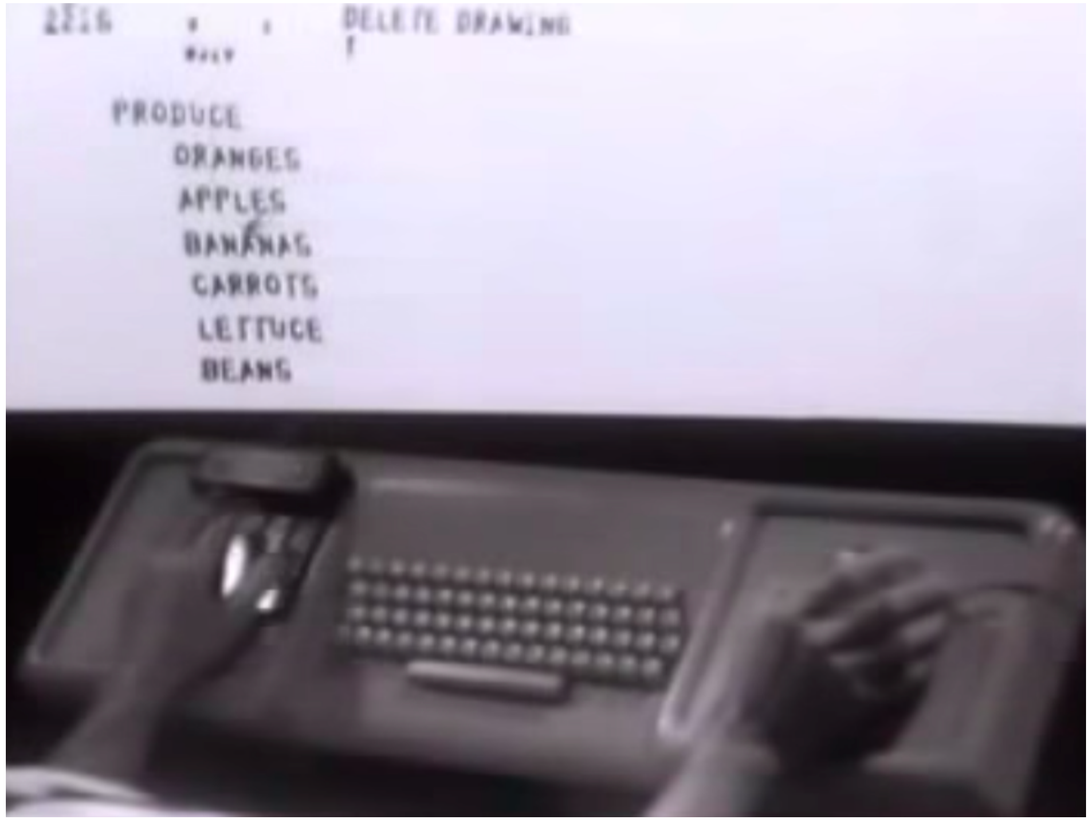
London Subway Map, 1927


www.stonebrowndesign.com
ALEWIFE
Davis
Porter
Harvard
Central
Kendall/MIT
LECHMERE
Science Park
North Stn
Haymarket
Gov't Center
BOWDOIN
Charles/MGH
Park
State
Downtown Crossing
OAK GROVE
Malden Center
Wellington
Sullivan Square
Comm'ty College
Maverick
Airport
Wood Island
Orient Heights
Suffolk Downs
Beachmont
Revere Beach
WONDERLAND
Aquarium
Boylston
Arlington
Copley
Prudential
Hynes
Kenmore
BOSTON COLLEGE
South St
Chestnut Hill Ave
Chiswick Rd
Sutherland St
Washington St
Warren St
Allston St
Griggs St
Harvard Ave
Packard's Corner
Babcock St
Pleasant St
St. Paul St
BU West
BU Central
BU East
Blanford St
CLEVELAND CIRCLE
Englewood Ave
Dean Rd
Tappan St
Washington Square
Fairbanks
Brandon Hall
Summit Ave
Coolidge Corner
St. Mary's St
Hawes St
Kent St
St. Paul St
Fenway
Longwood
Brookline Village
Brookline Hills
Beaconsfield
RIVERSIDE
Woodland
Waban
Eliot
Newton Highlands
Newton Centre
Chestnut Hill
Reservoir
HEATH STREET
Riverway
Back of the Hill
Mission Park
Fenwood Rd
Brigham Circle
Longwood Medical
Museum of Fine Arts
Northeastern
Symphony
FOREST HILLS
Green St
Stony Brook
Jackson Square
Roxbury Crossing
Ruggles
Massachusetts Ave
Back Bay
China-town
Tufts Medical
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
40 min
35 min
30 min
25 min
20 min
15 min
10 min
5 min
MASSACHUSETTS BAY TRANSPORTATION AUTHORITY
Time-Scale Subway Map
Boston rapid transit lines scaled by travel time to hub stations.
Scale
5 min.
2 min.
1 min.
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



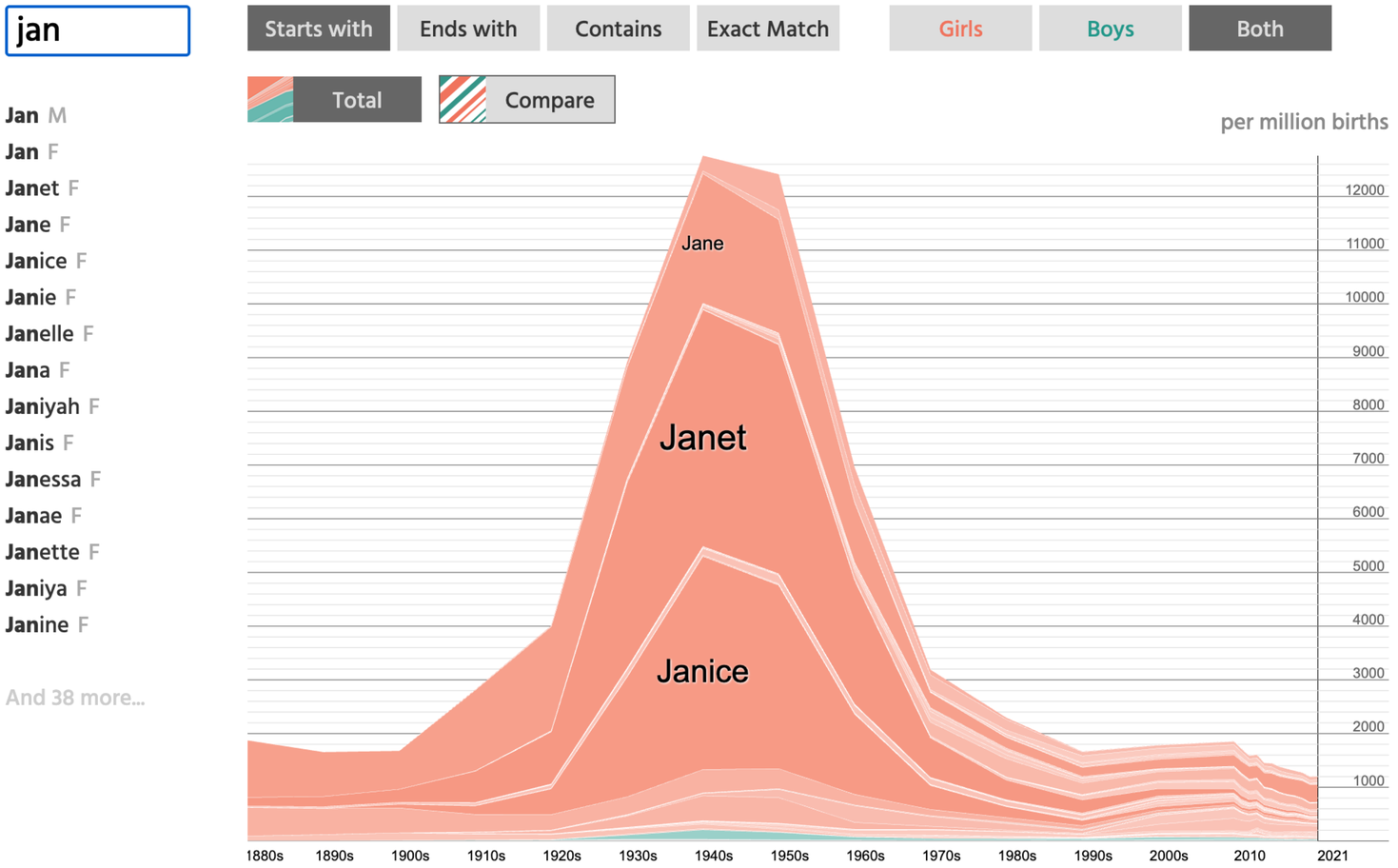
Ivan Sutherland, Sketchpad, 1963



Doug Engelbart, 1968

Modern Examples

Analyze



M. Wattenberg, 2005

Communicate

Hans Rosling, TED 2006

Who is CS-5630 / CS-6630?

Course Staff

Haihan Lin

Teaching Assistant

Kiran Gadhave

Teaching Assistant

Pranav Rajan

Teaching Assistant

Tripti Agarwal

Teaching Mentee

Alexander Lex

@alexander_lex
http://alexander-lex.net

Associate Professor, Computer Science

Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology

visualization
design lab

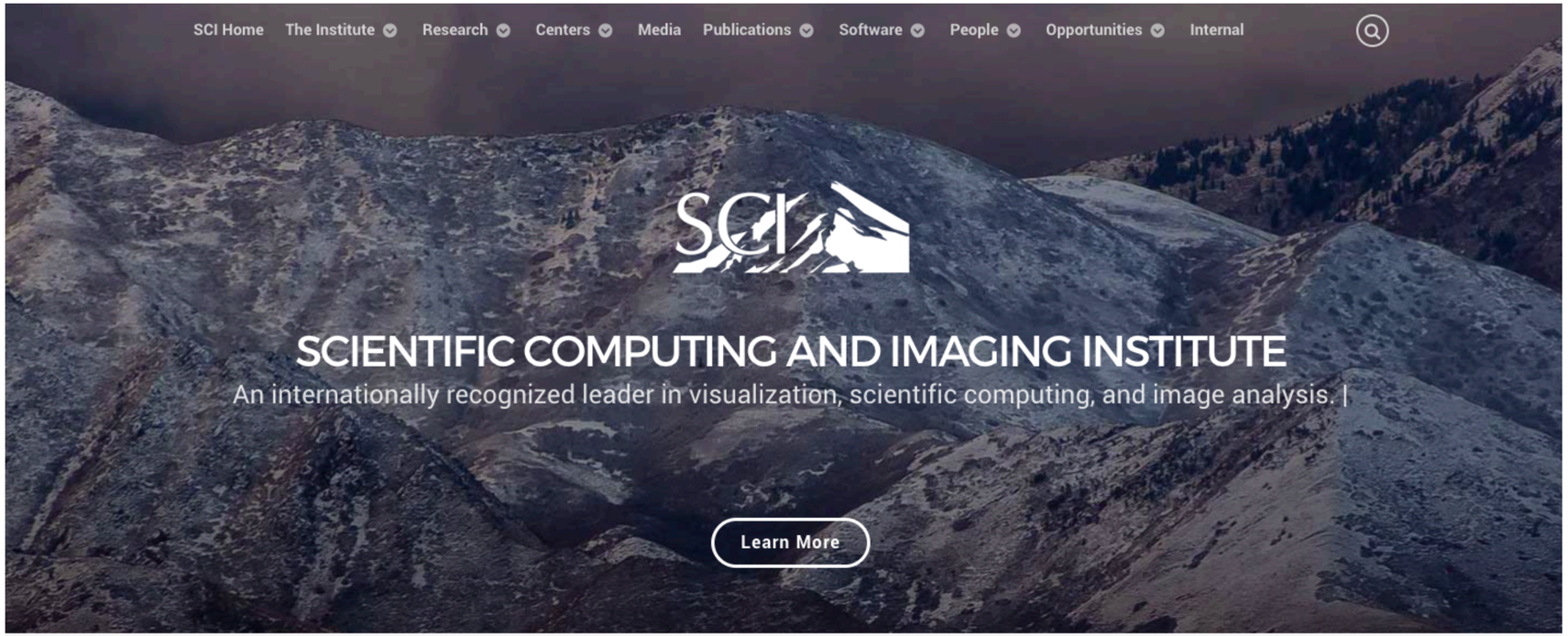
http://vdl.sci.utah.edu/

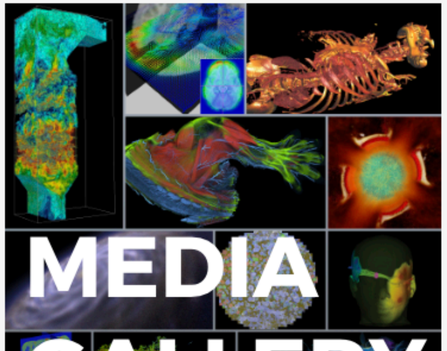
SCI Institute

Scientific Computing and Imaging Institute

- Scientific Computing
- Biomedical Computing
- Scientific Visualization
- Information Visualization
- Image Analysis

http://sci.utah.edu





TECHNICAL CONTRIBUTIONS

Novel Visualization Techniques

Visualization Process Innovations

Data Wrangling Methods

DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

EMPIRICAL & THEORETICAL WORK

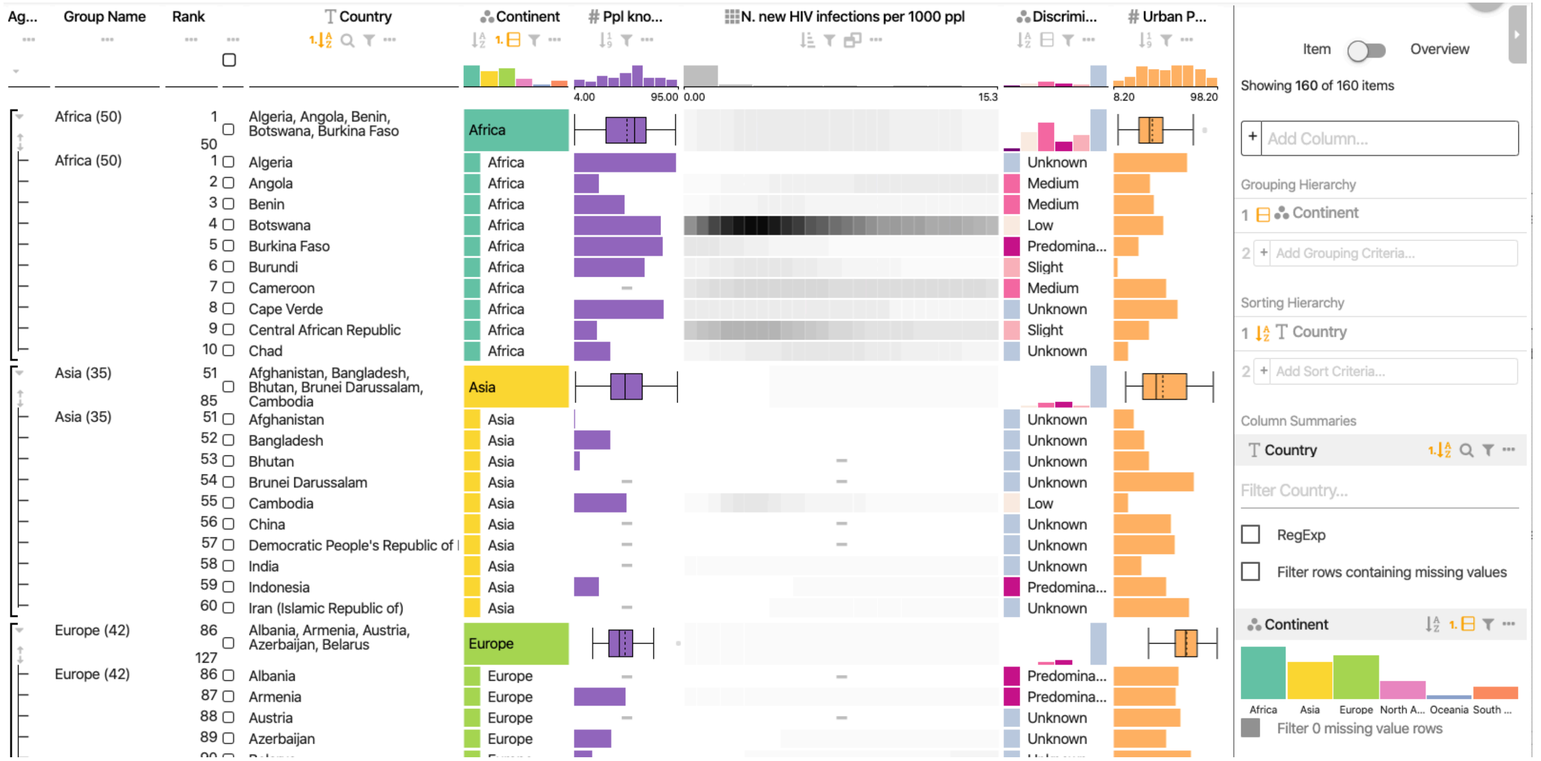
Evaluation Methodology

Design Spaces / Taxonomies

TECHNICAL CONTRIBUTIONS

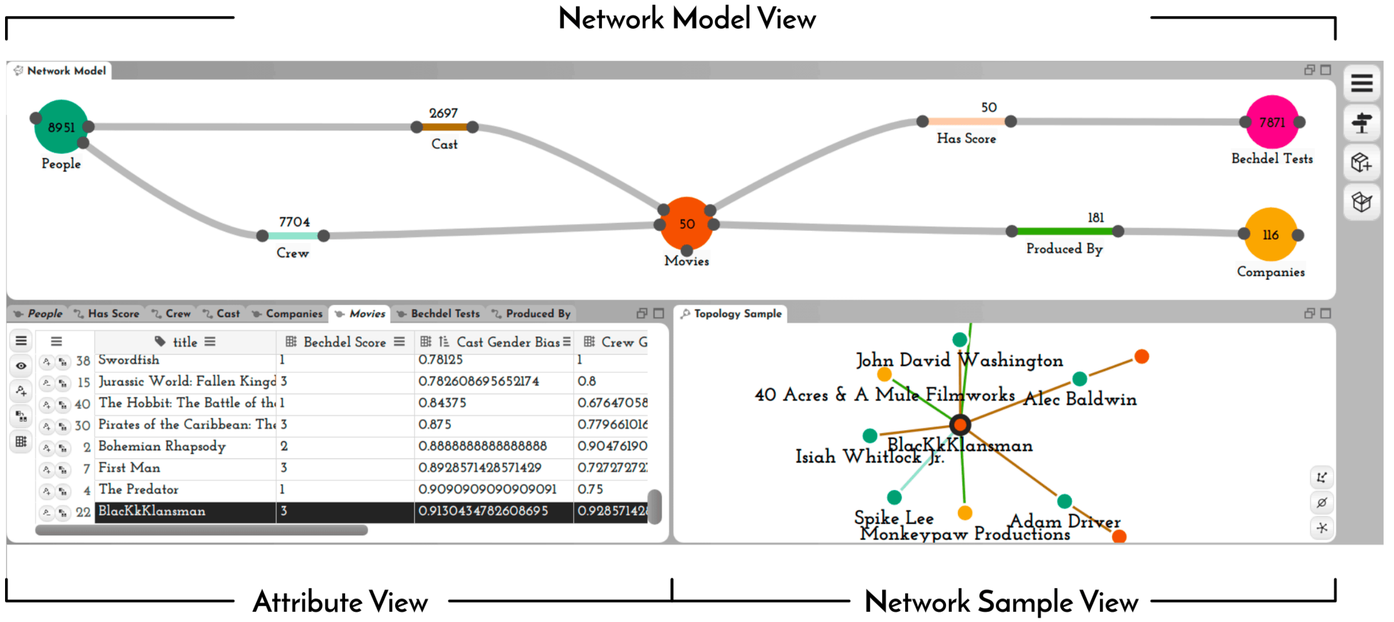
Tabular Data

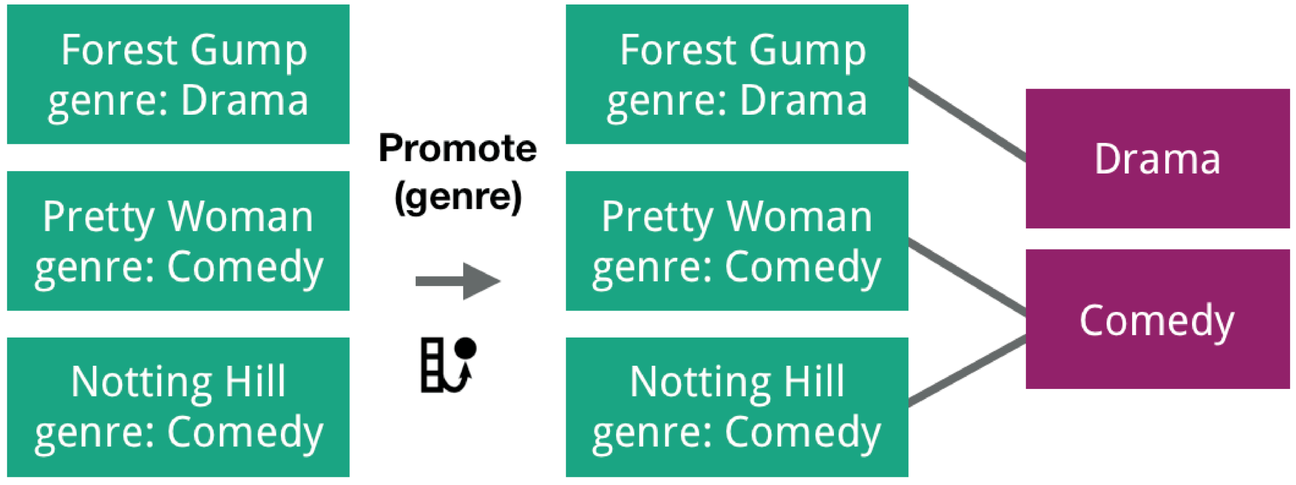
Novel Visualization Techniques



Reshaping Networks

Data Wrangling Methods





TECHNICAL CONTRIBUTIONS

Data Wrangling Methods

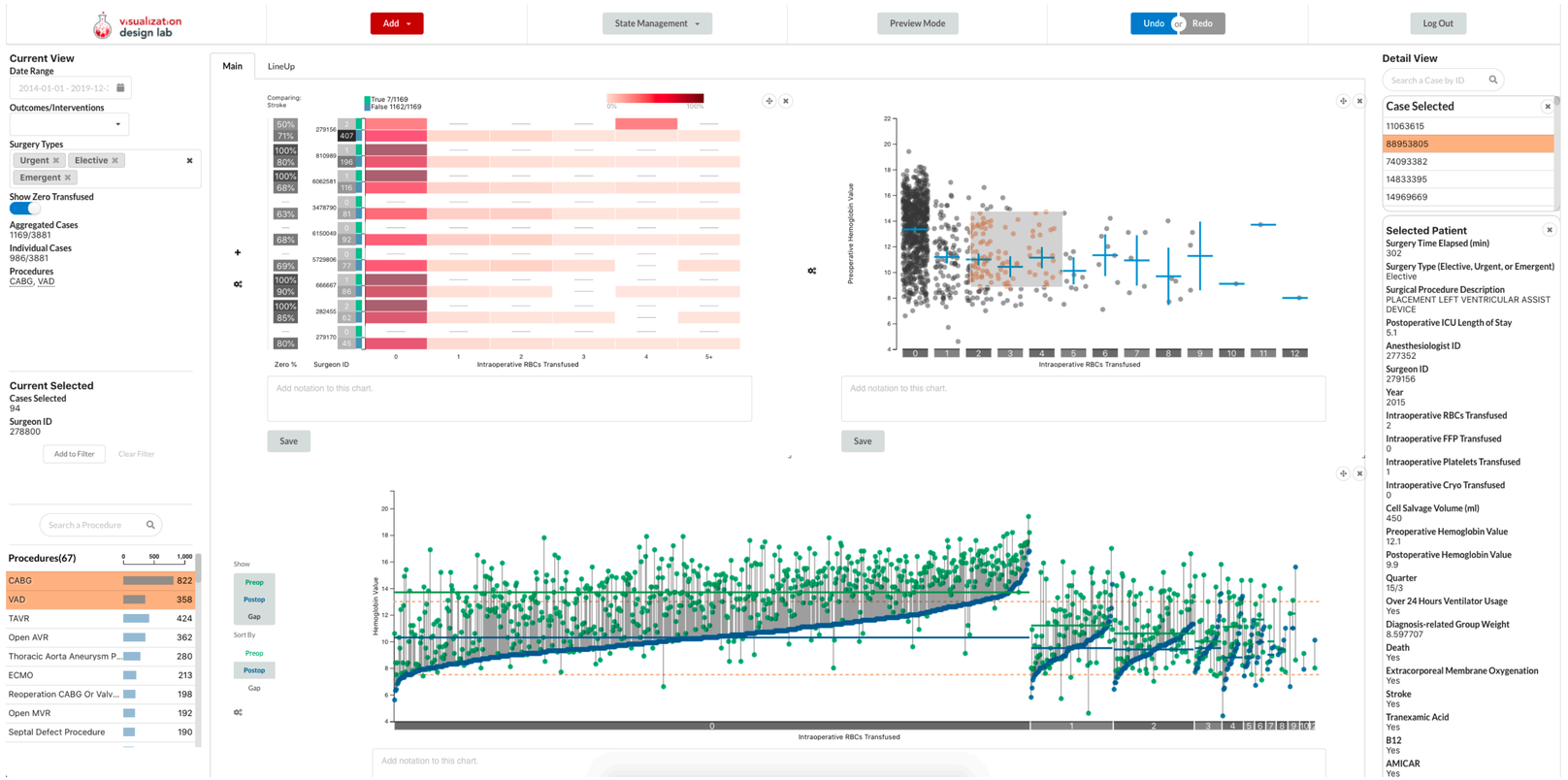
DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

DOMAIN DRIVEN TECHNIQUES

Electronic Health Records

Tailored Methods and Systems for High Impact Problems



DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

EMPIRICAL & THEORETICAL WORK

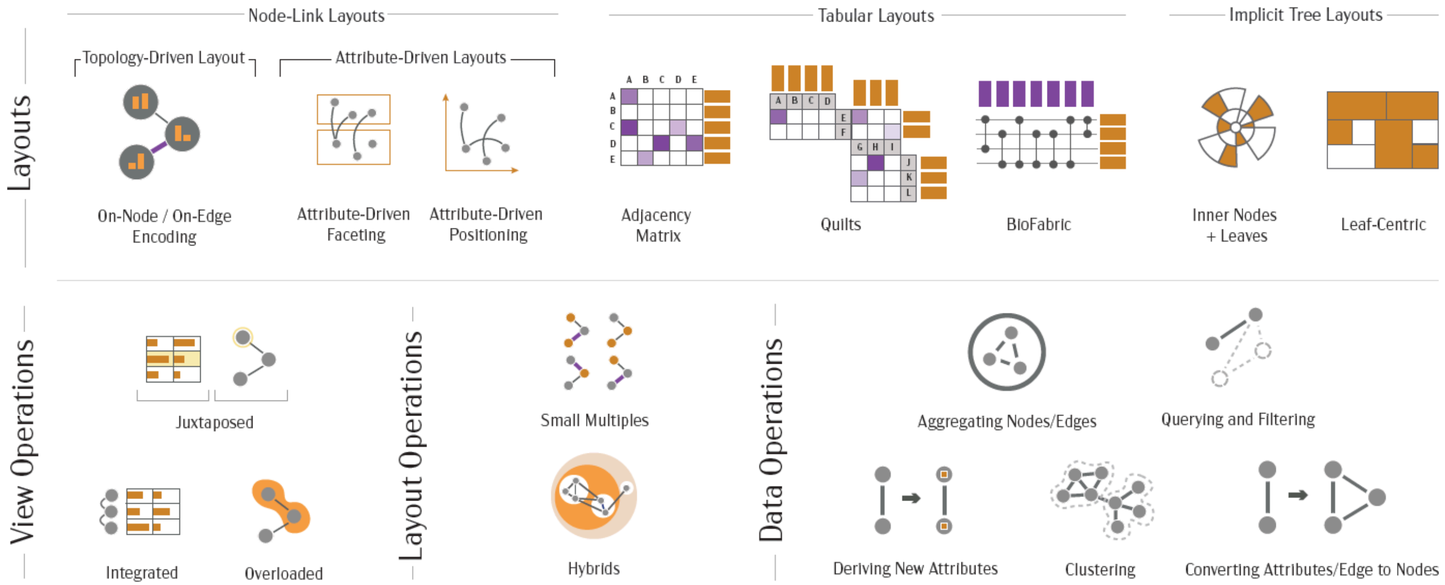
Evaluation Methodology

Design Spaces / Taxonomies

EMPIRICAL & THEORETICAL WORK

Evaluation Methodology

Design Spaces / Taxonomies



Structure & Goals

Course Goals. You will learn:

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

Course Components

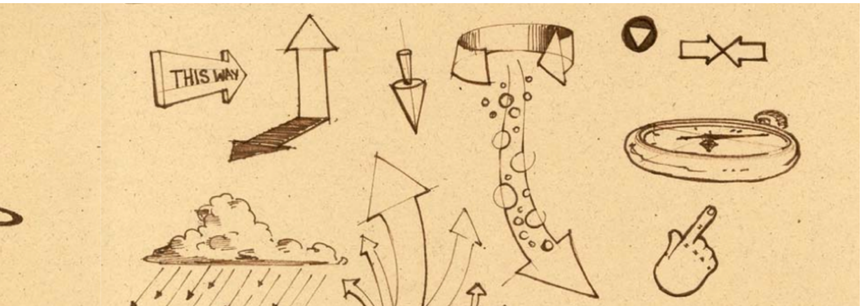
Lecture
Reading
Discussion

Theory

Design Critiques
Redesigns

Labs
D3 reading
Self-study
Office hours

Design Skills ⟷ Coding Skills




```
<!DOCTYPE html>
<meta charset="utf-8">
<style>

text {
  font: 10px sans-serif;
}

</style>
<body>
<script src="http://d3js.org/d3.v3.min.js"></script>
<script>
```

Course Components

Lectures: introduce theory

Design Critiques / Redesigns: develop “an eye” for vis design, critique, learn by example;

Will happen in breakout groups,

Submit 5 critiques, within 2 days after posting

Labs: short coding tutorials, examples

Based on a published script on website

Strongly related to homework assignments

Homeworks help practice specific skills

Final Project gives you a chance to go through a complete vis project

Schedule

Lectures: Tuesday and Thursday 2:00-3:20 pm

Labs: Mondays, 5:00-6:00 pm, L110 (scheduled on demand)

Review Lectures:
YouTube Channel

Three Parts:

I. Technical Foundations
HTML, Javascript, D3

II. Visualization Fundamentals
Perception, Visual encodings, Design Guidelines, Tasks..

III. Abstract Data Visualization
Tables, Graphs, Maps

Schedule

CS 5630/6630

Today | September 2022 | Print | Week | Month | Agenda

Mon	Tue	Wed	Thu	Fri	Sat	Sun
29 Week 2	30 10:00 Tripti's Office H 14:00 Vis for Data Sc 16:00 Alexander Lex	31 09:30 Kiran Office Ho	Sep 1 10:00 Pranav Rajan (14:00 Vis for Data Sc	2 Drop Deadline HW1 Due 10:00 Haihan's Office	3	4
5 Week 3	6 10:00 Tripti's Office H 14:00 Vis for Data Sc 16:00 Alexander Lex	7 09:30 Kiran Office Ho	8 10:00 Pranav Rajan (14:00 Vis for Data Sc	9 HW2 Due 10:30 Haihan's Office	10	11
12 Week 4	13 10:00 Tripti's Office H 14:00 Vis for Data Sc 16:00 Alexander Lex	14 09:30 Kiran Office Ho	15 10:00 Pranav Rajan (14:00 Vis for Data Sc	16 HW3 Due 10:00 Haihan's Office	17	18
19 Week 5	20 10:00 Tripti's Office H 14:00 Vis for Data Sc	21 09:30 Kiran Office Ho	22 10:00 Pranav Rajan (14:00 Vis for Data Sc	23 10:00 Haihan's Office	24	25
26 Week 6	27 10:00 Tripti's Office H 14:00 Vis for Data Sc 16:00 Alexander Lex	28 09:30 Kiran Office Ho	29 10:00 Pranav Rajan (14:00 Vis for Data Sc	30 HW4 Due 10:00 Haihan's Office	Oct 1	2

Events shown in time zone: Mountain Time - Denver

Google Calendar

Subject to change

Week 1

Introduction — Tuesday

What is visualization? Why is it important? Who are we? Course overview.
Download slides

Recommended reading

- A Tour through the Visualization Zoo. Jeffrey Heer, Michael Bostock, Vadim Ogievetsky. Communications of the ACM, 53(6), pp. 59-67, Jun 2010.

Information http://dataviscourse.net

Visualization for Data Science
CS-5630 / CS-6630

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Project Resources Fame

UpSet visualizing intersecting sets | *Wind map* | *How states have shifted*

The amount and complexity of information produced in science, engineering, business, and everyday human activity is increasing at staggering rates. The goal of this course is to expose you to visual representation methods and techniques that increase the understanding of complex data. Visualization for data discovery and communication is an important part of the data science pipeline. Good visualizations not only present a visual interpretation of data, but do so by improving comprehension, communication, and decision making.

In this course you will learn about the fundamentals of perception, the theory of visualization, good design practices for visualization, and how to develop your own web-based visualizations using HTML5, CSS, JavaScript, SVG, and D3.

Communicate

Slack

https://visfordatascience2022.slack.com/

Please use slack for all general questions - code, concepts, etc.

Please don’t use personal messages to me or TAs

Only use e-mail for personal inquiries

Canvas

https://utah.instructure.com/courses/803252

Homework submissions, Grades

Office Hours

Alex: Tuesdays, 4-5pm

TAs: starting next week

E-Mail

alex@sci.utah.edu

New U policy: must use utah.edu e-mail address to communicate (FERPA)

Lecture Recordings / Stream

In-Person Class

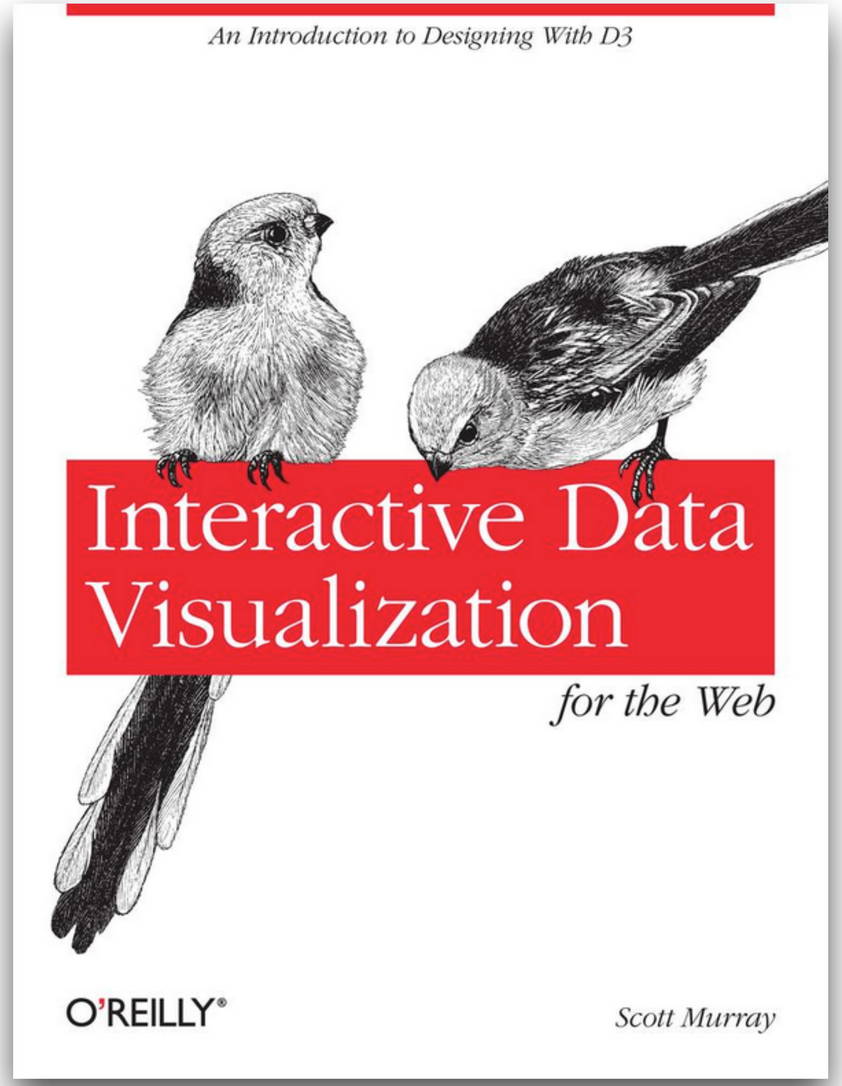
Lot’s of discussion and in class activities; handouts, etc.

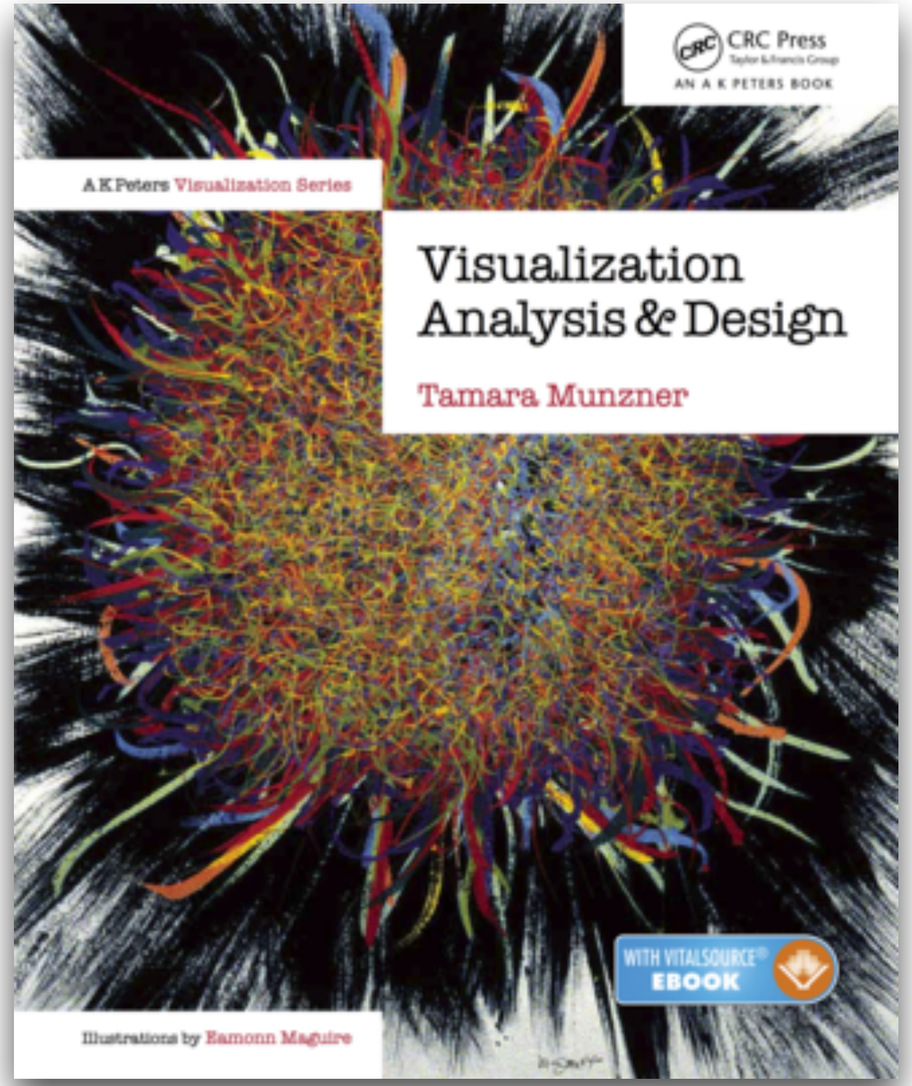
Recording as a back-up:

- If you’re sick or have emergency travel
- To review lecture content

https://www.youtube.com/channel/UCDUS80bdunpmvWVPyFRPqFQ/live

Required Books





Programming

HTML

JS

Data-Driven Documents

Prerequisites

Programming experience

- C, C++, Java, Python, etc.

Willingness to think about user-centered design

- This is not your average CS course! We care about the human in the loop!

Willingness to learn new software & tools

- This can be time consuming

You will need to build skills by yourself!

Formalities

How are you graded?

5-6 Homework Assignments: 35%

- Varying value, 2%-10%, depending on length/difficult
- Start early! Will take long if you don’t know JS/D3 yet
- Due on Fridays, late days: -10% per day, up to two days.

Final Project: 40%

- Teams, proposal and two milestones

Exams: 20%

- Two exams: last class before fall break and end of term

Group Activities: 5%

Code of Conduct

- We are committed to providing an inclusive and harassment-free environment in all interactions regardless of gender, sexual orientation, disability, physical appearance, race, or religion.
- We do not tolerate harassment in any form.
- Please report any harassment to me or the appropriate university office, which you can find at https://safeu.utah.edu/
- Please review the syllabus on these issues and the student code of conduct at https://regulations.utah.edu/academics/6-400.php

Cheating

You are welcome to **discuss** the course's ideas, material, and homework with others in order to better understand it, but **the work you turn in must be your own** (or for the project, yours and your teammate's). For example, you must **write your own code**, design your own visualizations, and critically evaluate the results in your own words.

You **may not submit the same or similar work** to this course that you have submitted or will submit to another. **Nor may you provide or make available solutions to homeworks to individuals** who take or may take this course in the future.

See also the SoC Academic Misconduct Policy:
https://www.cs.utah.edu/docs/misc/cheating_policy.pdf

You will fail the class if you cheat.

A "strike" will be recorded.

We will **automatically check for plagiarism** in all your submissions.

Cheating

Cheating is easy to catch

Do a cost-benefit analysis:

- loosing points on a homework vs failing a class (and loosing tuition)

Tools like MOSS make it easy to catch cheating

- If you have copied code in the past but have not been caught, it's likely no one checked, or they didn't want to bother with the hassle

This Week

HW0, including course survey

Lecture on Perception

Readings

- D3 Book, Chapters 1-3
- VDA Book, Chapter 1

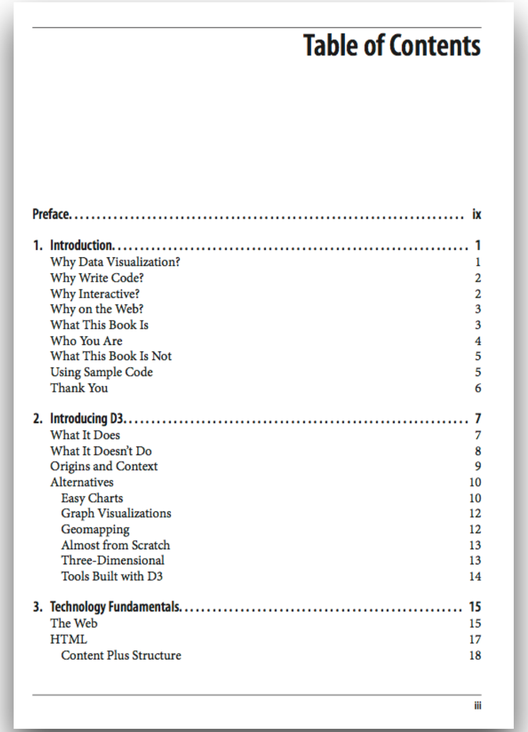


Table of Contents

Preface. ix

1. Introduction. 1
   Why Data Visualization? 1
   Why Write Code? 2
   Why Interactive? 2
   Why on the Web? 3
   What This Book Is 3
   Who You Are 4
   What This Book Is Not 5
   Using Sample Code 5
   Thank You 6

2. Introducing D3. 7
   What It Does 7
   What It Doesn't Do 8
   Origins and Context 9
   Alternatives 10
      Easy Charts 10
      Graph Visualizations 12
      Geomapping 12
      Almost from Scratch 13
      Three-Dimensional 13
      Tools Built with D3 14

3. Technology Fundamentals. 15
   The Web 15
   HTML 17
      Content Plus Structure 18

Next Week

HW1 due

Introduction to Git, HTML, CSS

Office hours start!

https://github.com/dataviscourse/2022-homework