Pre-Lecture Info

Zoom Lecture will be recorded, be aware when you participate

I know participation is hard online, try to make it interactive

Please mute yourself if you don’t speak

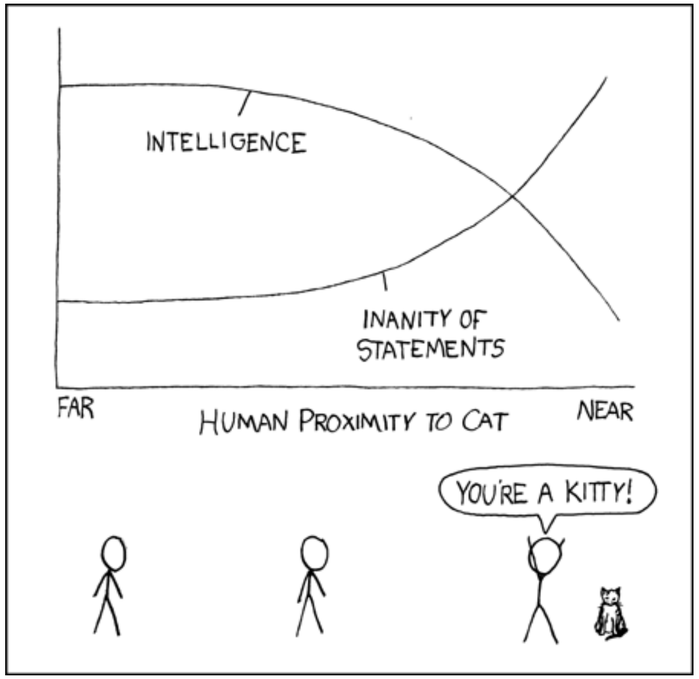
If you have questions, please raise your hand in Zoom, TA will monitor.

Alternatively, ask question on Slack, TA will ask for you

CS-5630 / CS-6630
Visualization for Data Science

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

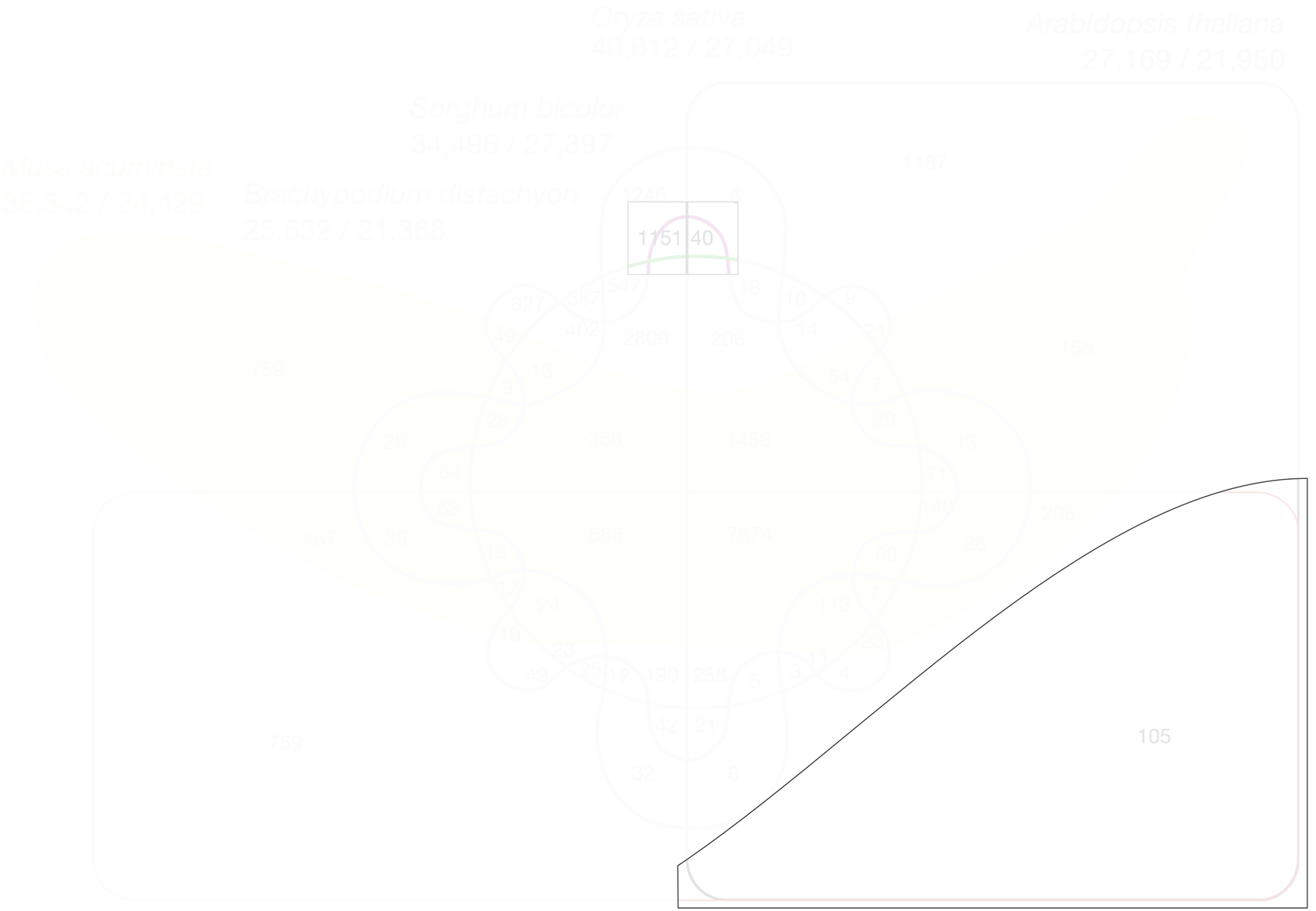
visualization *pictures*

The purpose of computing is insight, not numbers.

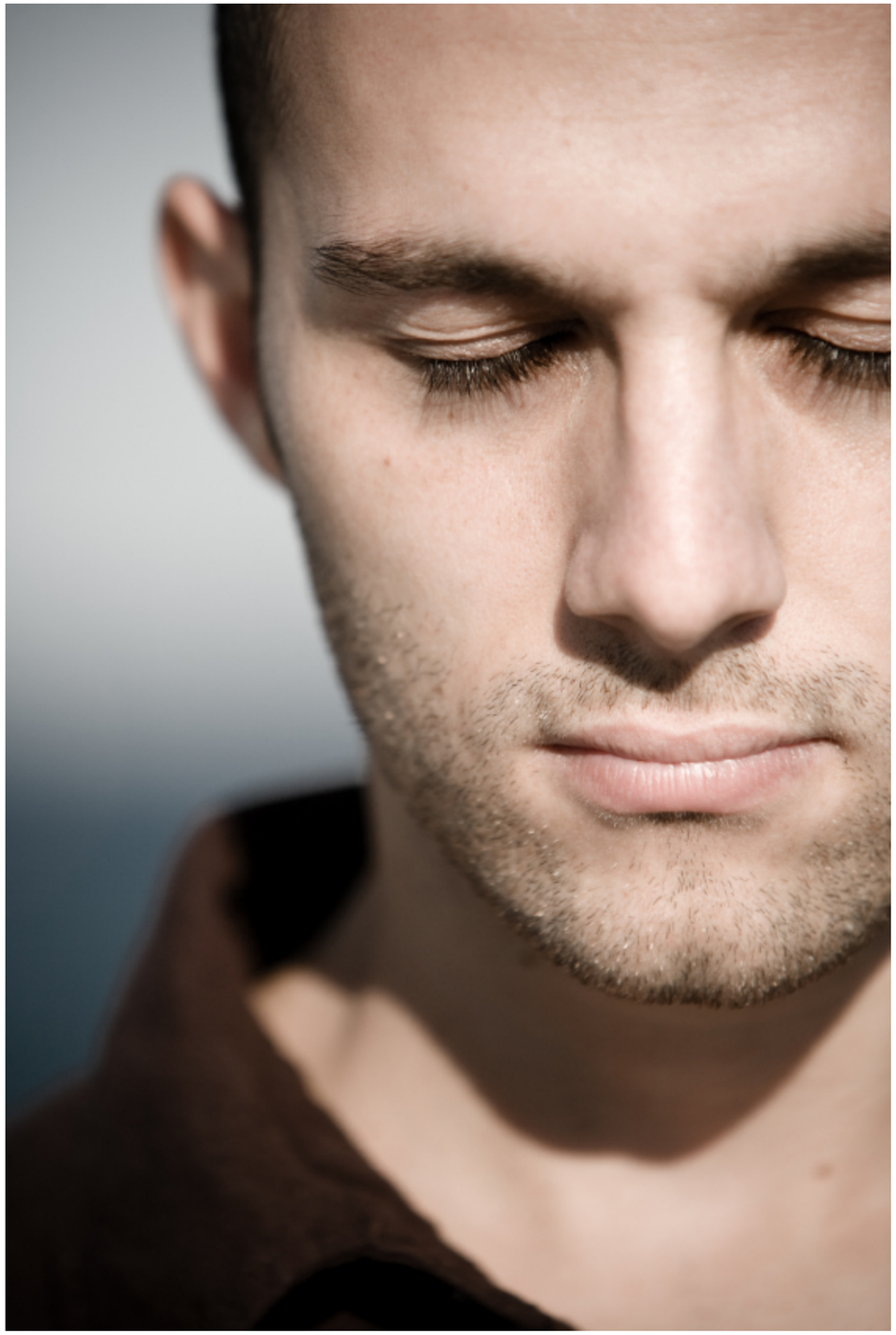
- Richard Wesley Hamming

- 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?

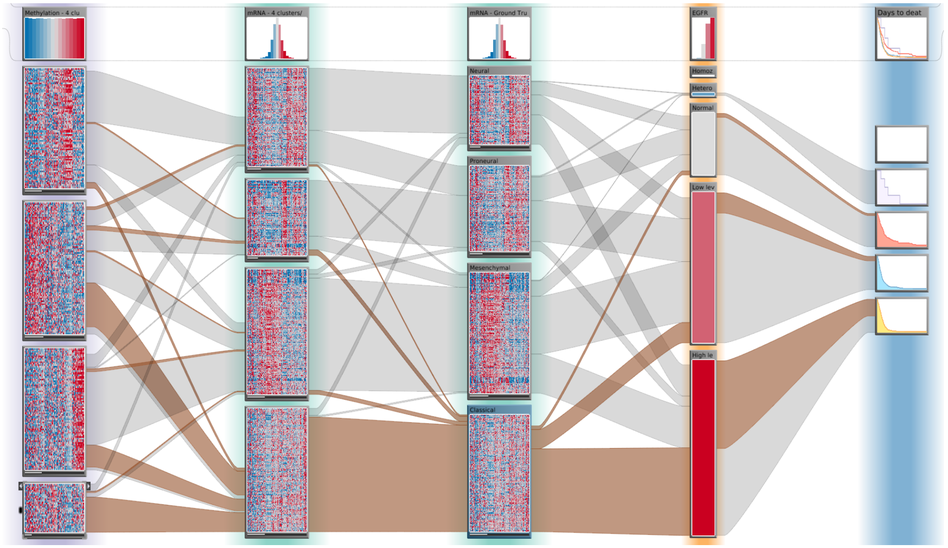
When questions are not well defined: Exploration

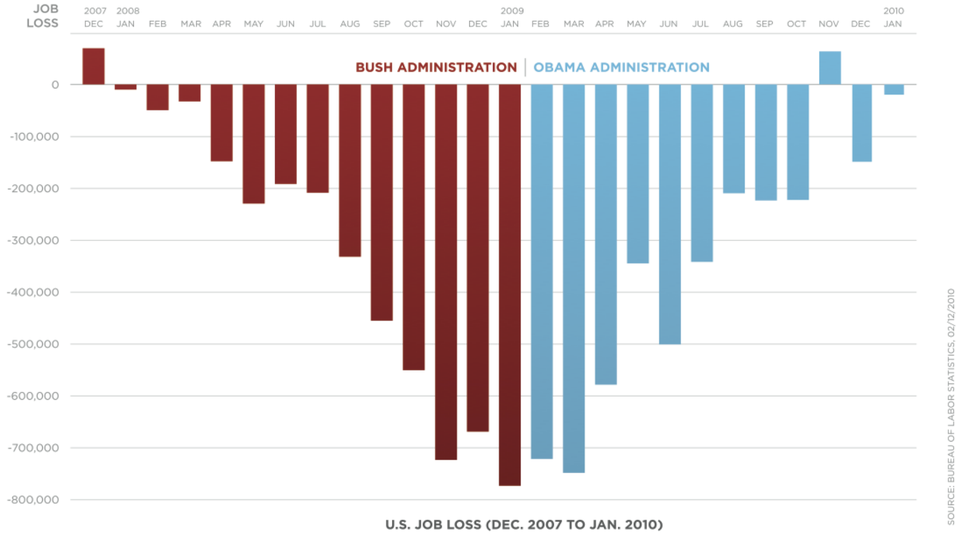
What is the structure of a terrorist network?

Which drug can help patient X?

Purpose of Visualization

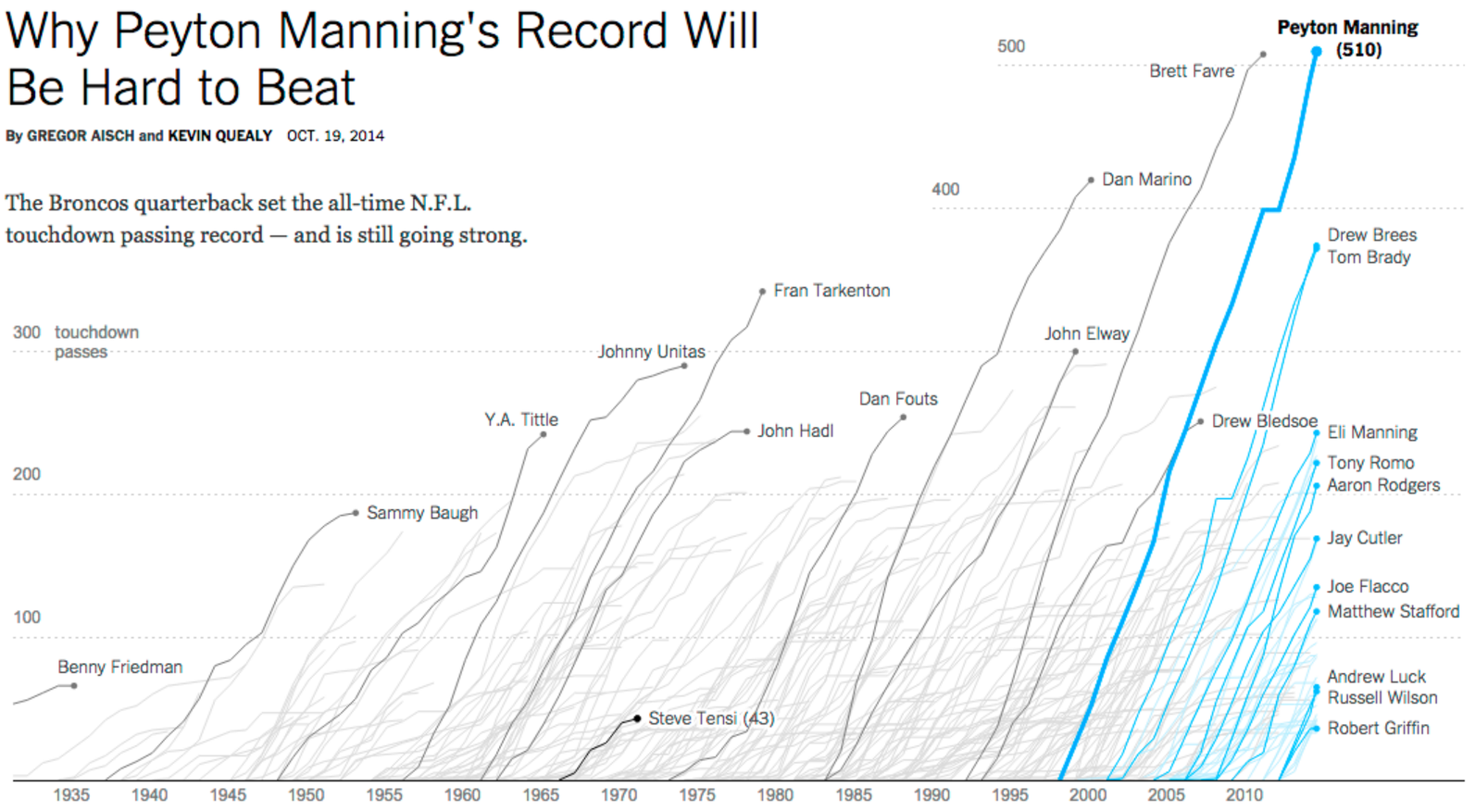
[Obama Administration]





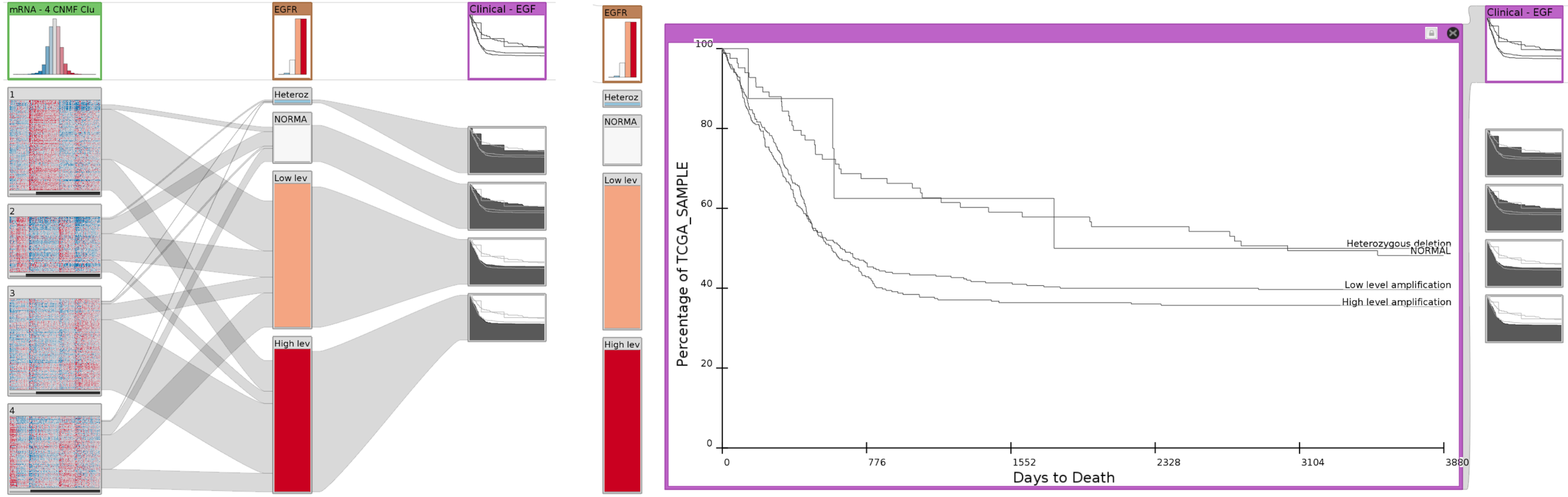
Open Exploration — Confirmation — Communication

Example Communication



[New York Times]

Example Exploration: Cancer Subtypes



[Caleydo StratomeX]

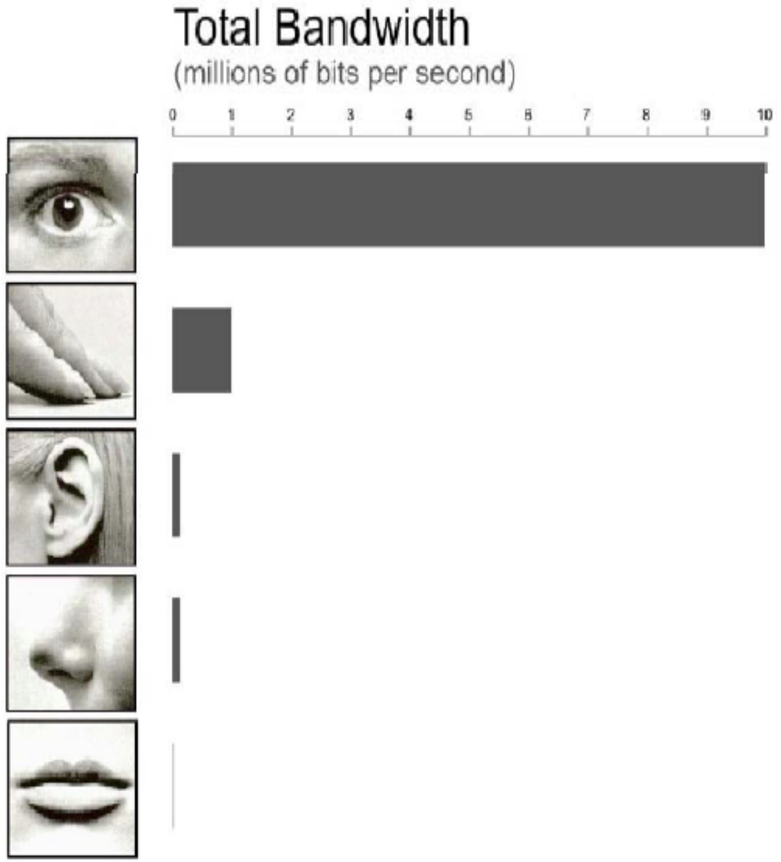
Why Graphics?

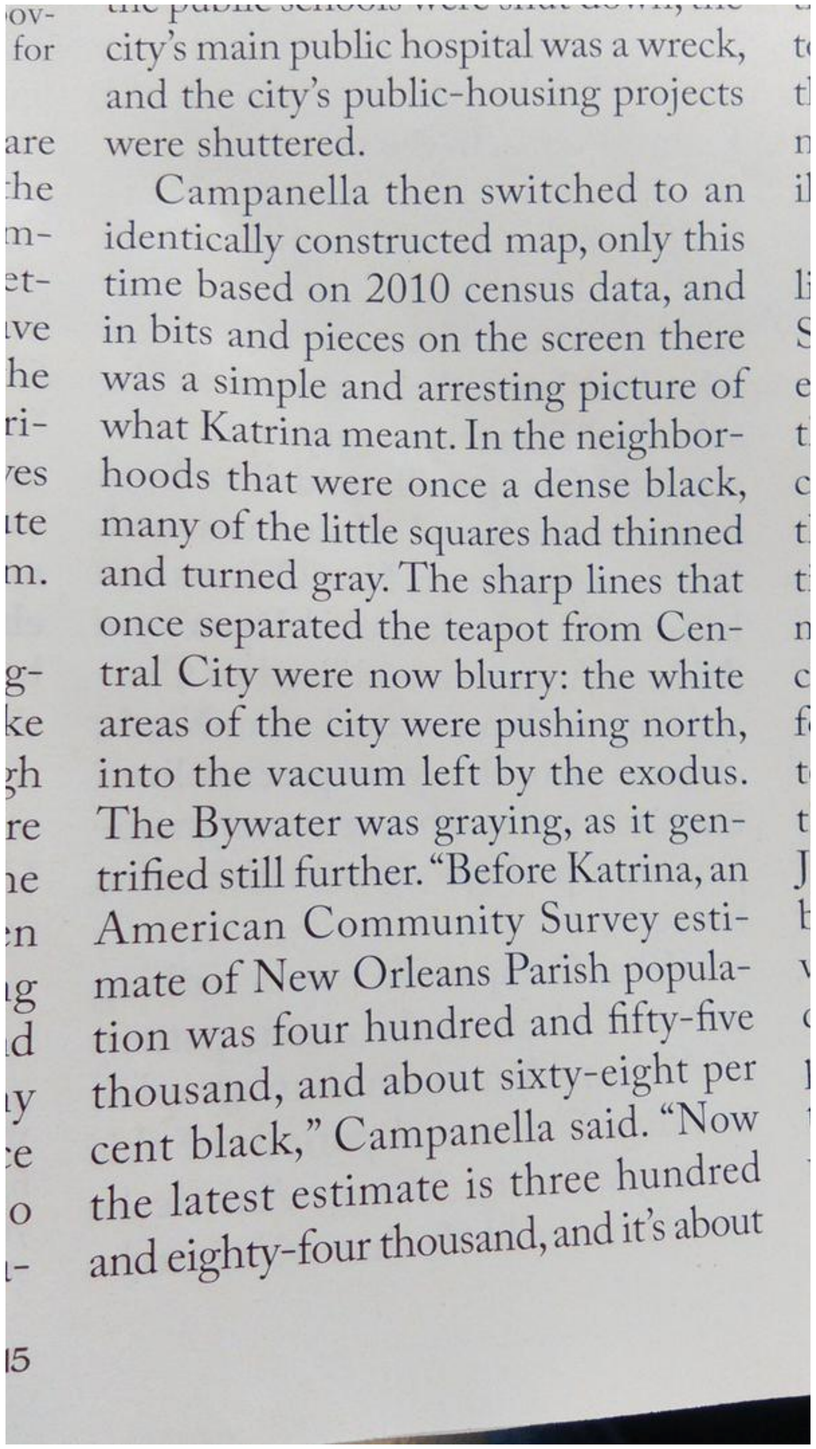
Figures are richer; provide more information with less clutter and in less space.

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

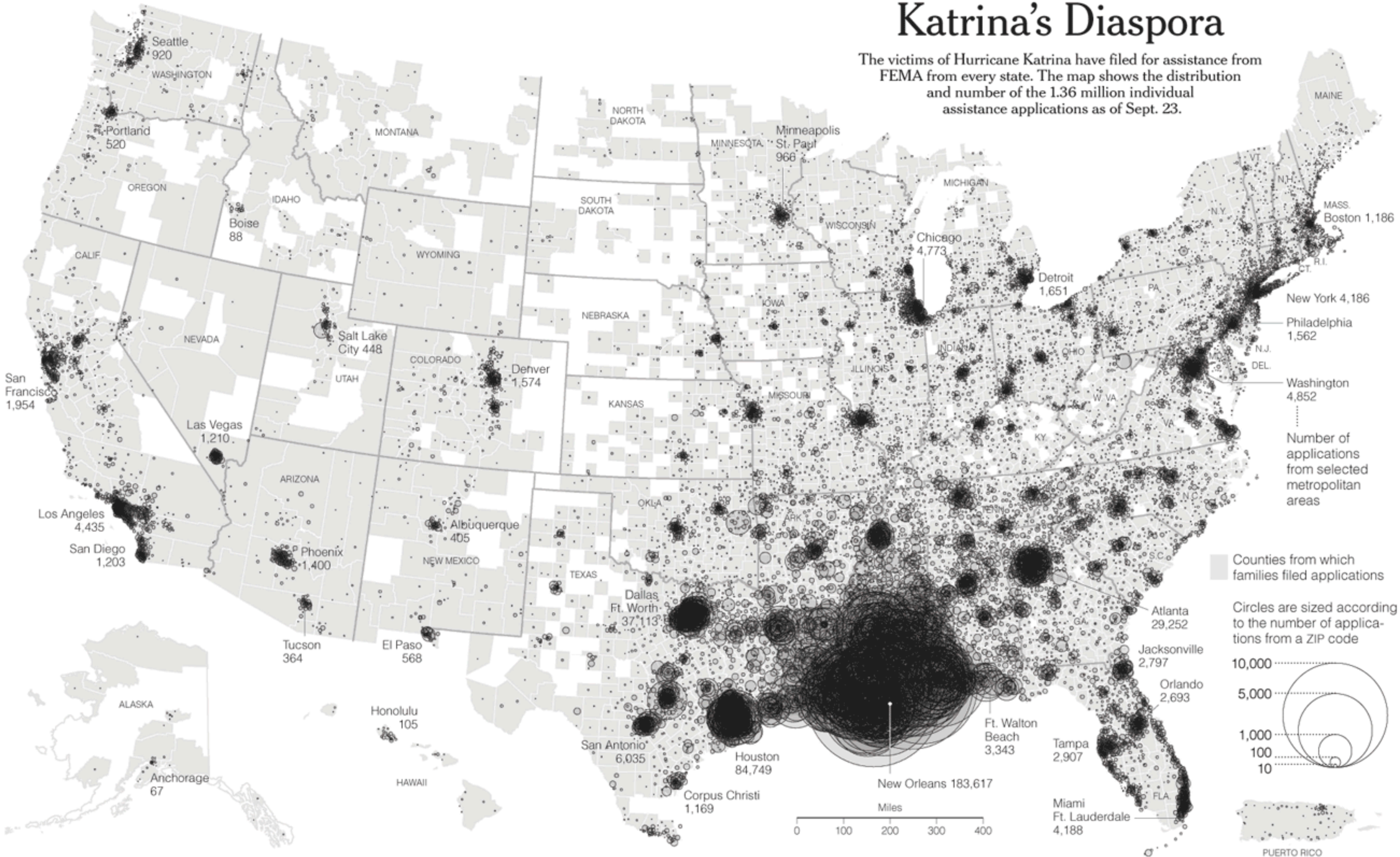
Textual description of a map of the effects of hurricane Katrina on New Orleans.
New Yorker, posted by Alberto Cairo

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

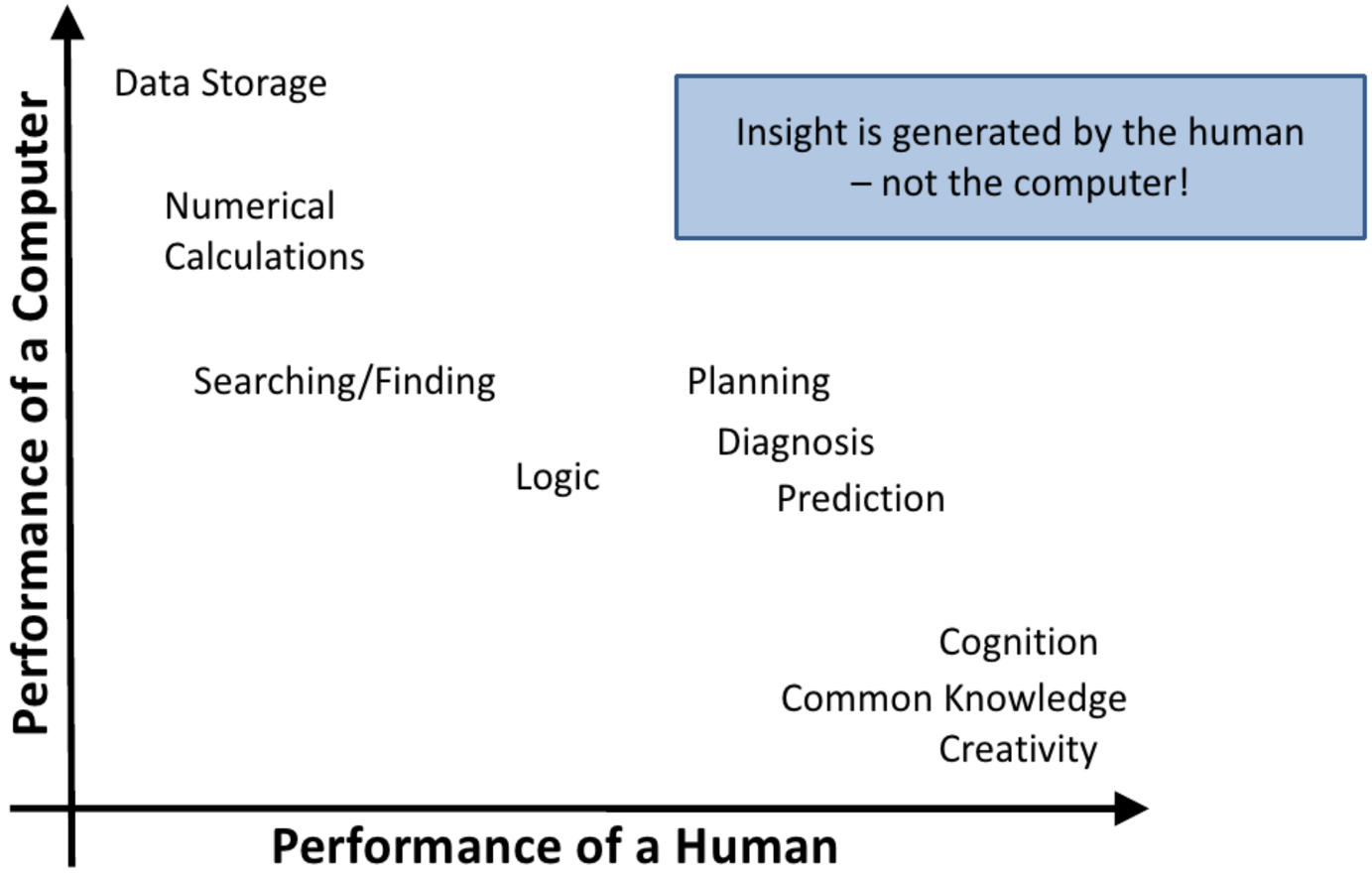
High frequency stock market trading: which stock to buy/sell?

Manufacturing: is bottle broken?

Impractical for human to be involved

Automatic data products

The Ability Matrix



Why Use Computers?

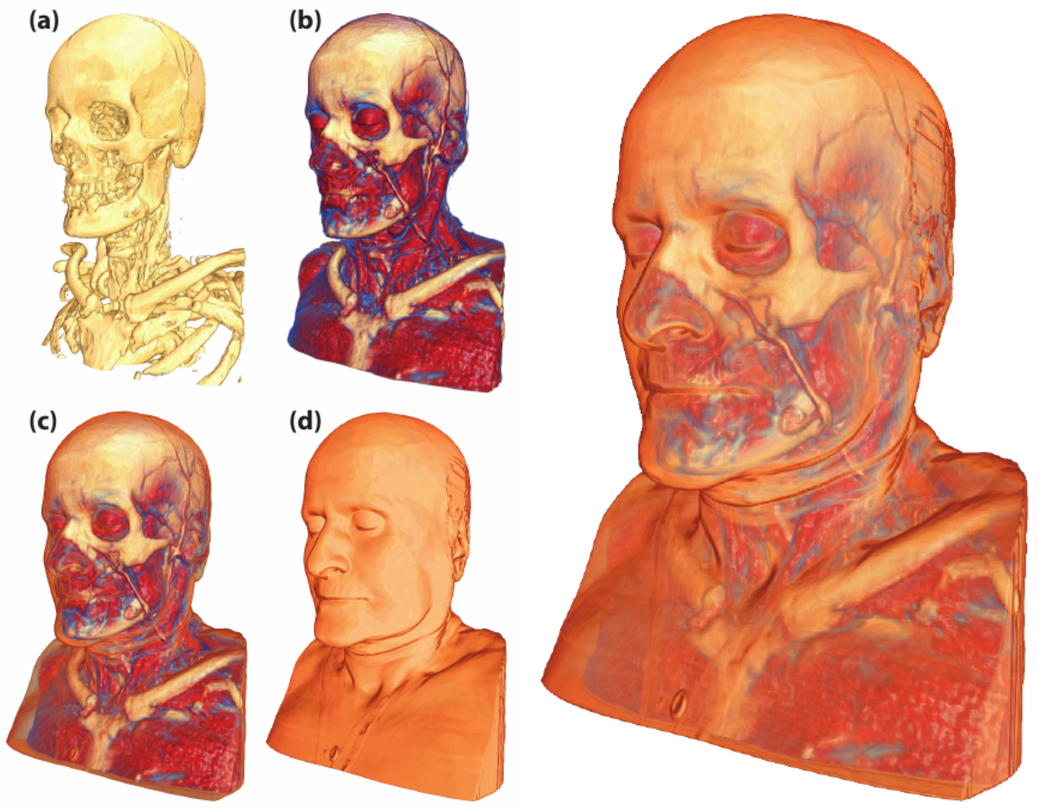
Scale

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

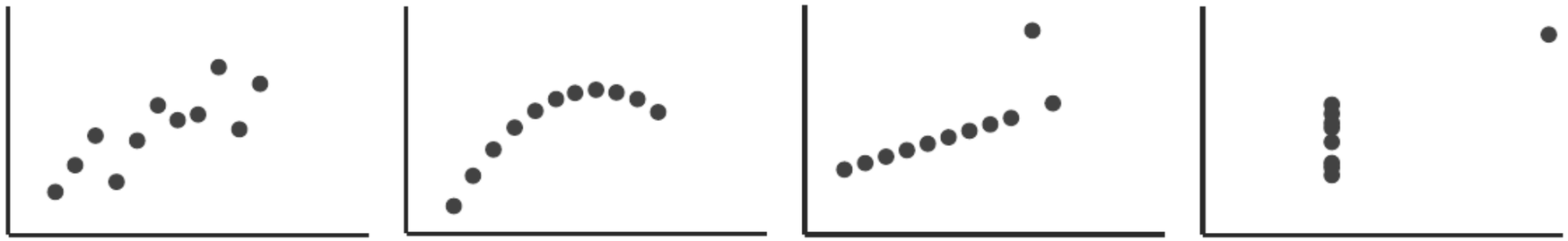
[New York Times]

Why not just use Statistics?

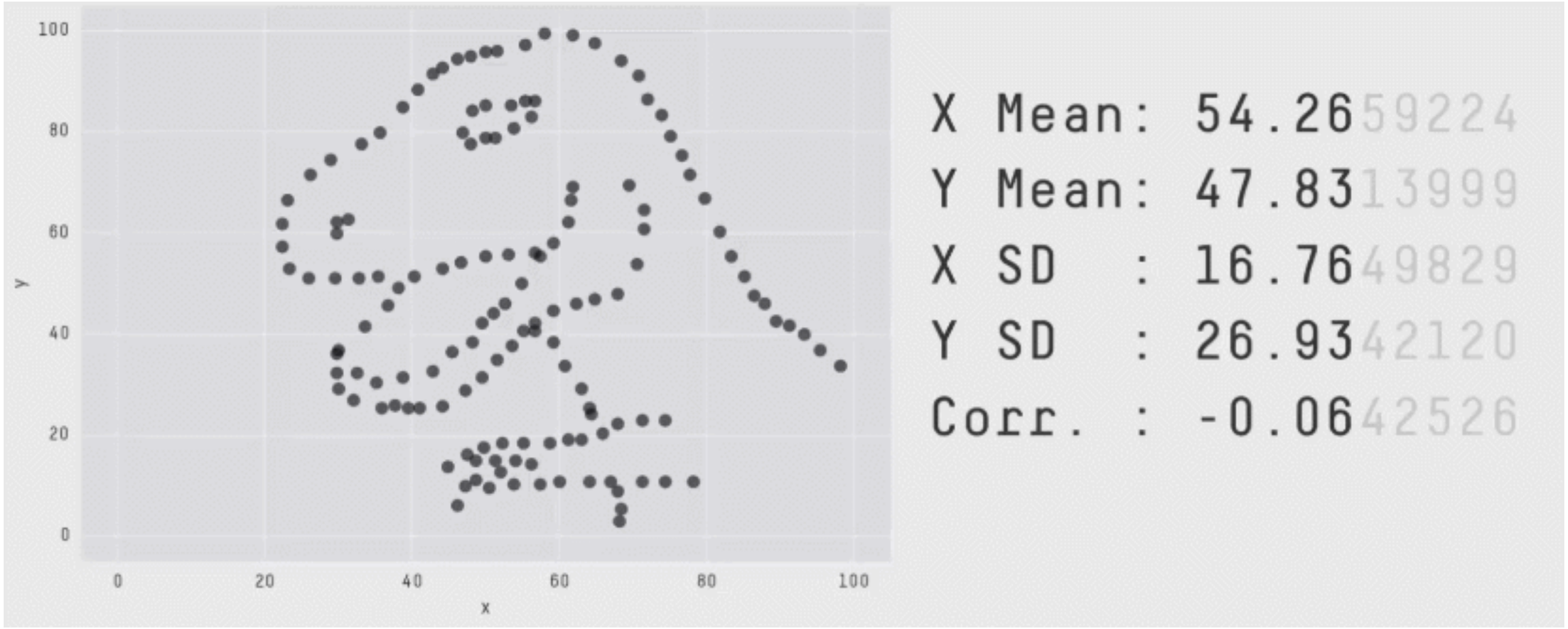
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.4	8	7.9
5	5.						6.8

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

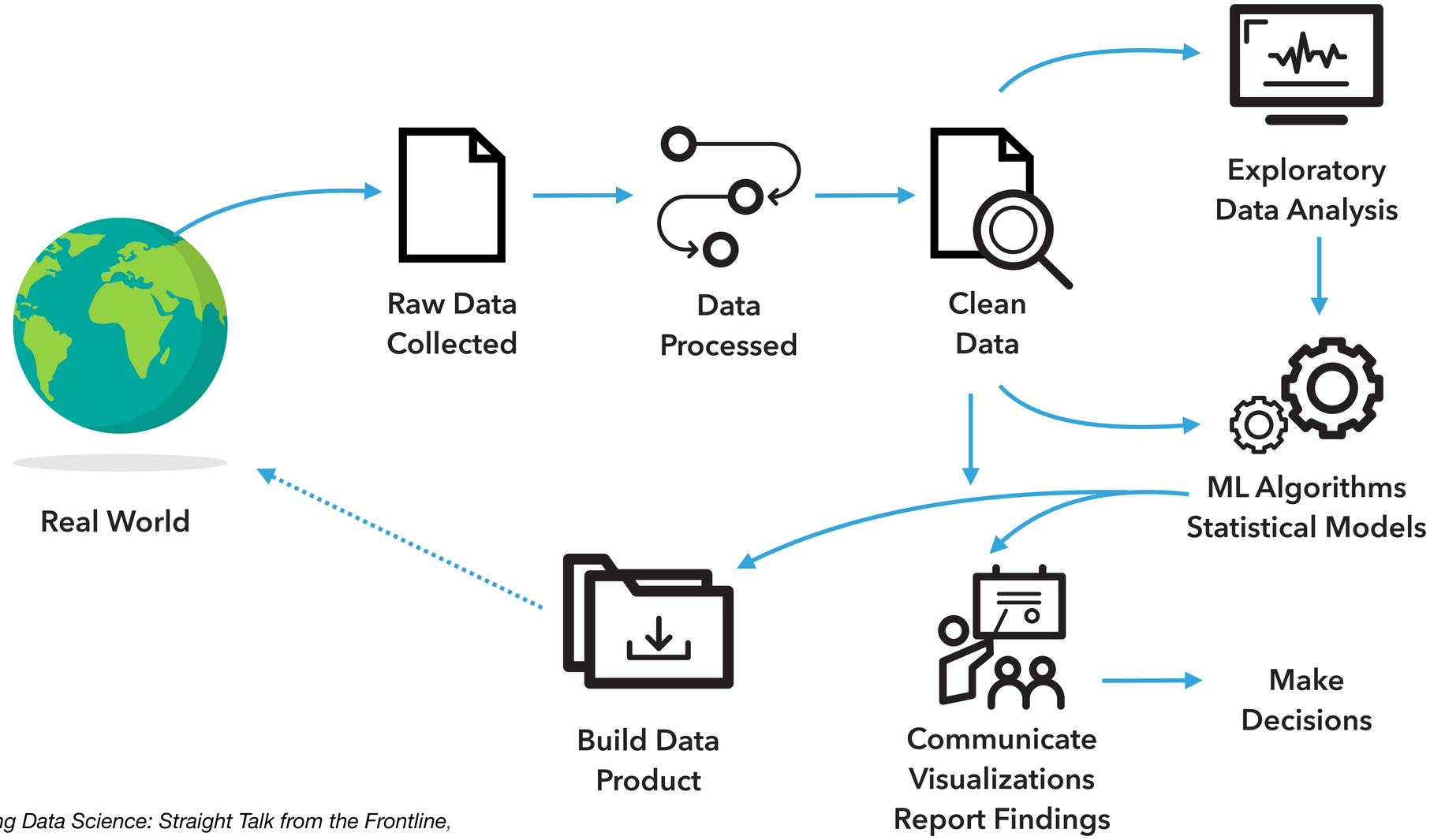
Visualization =

Human Data Interaction

Data

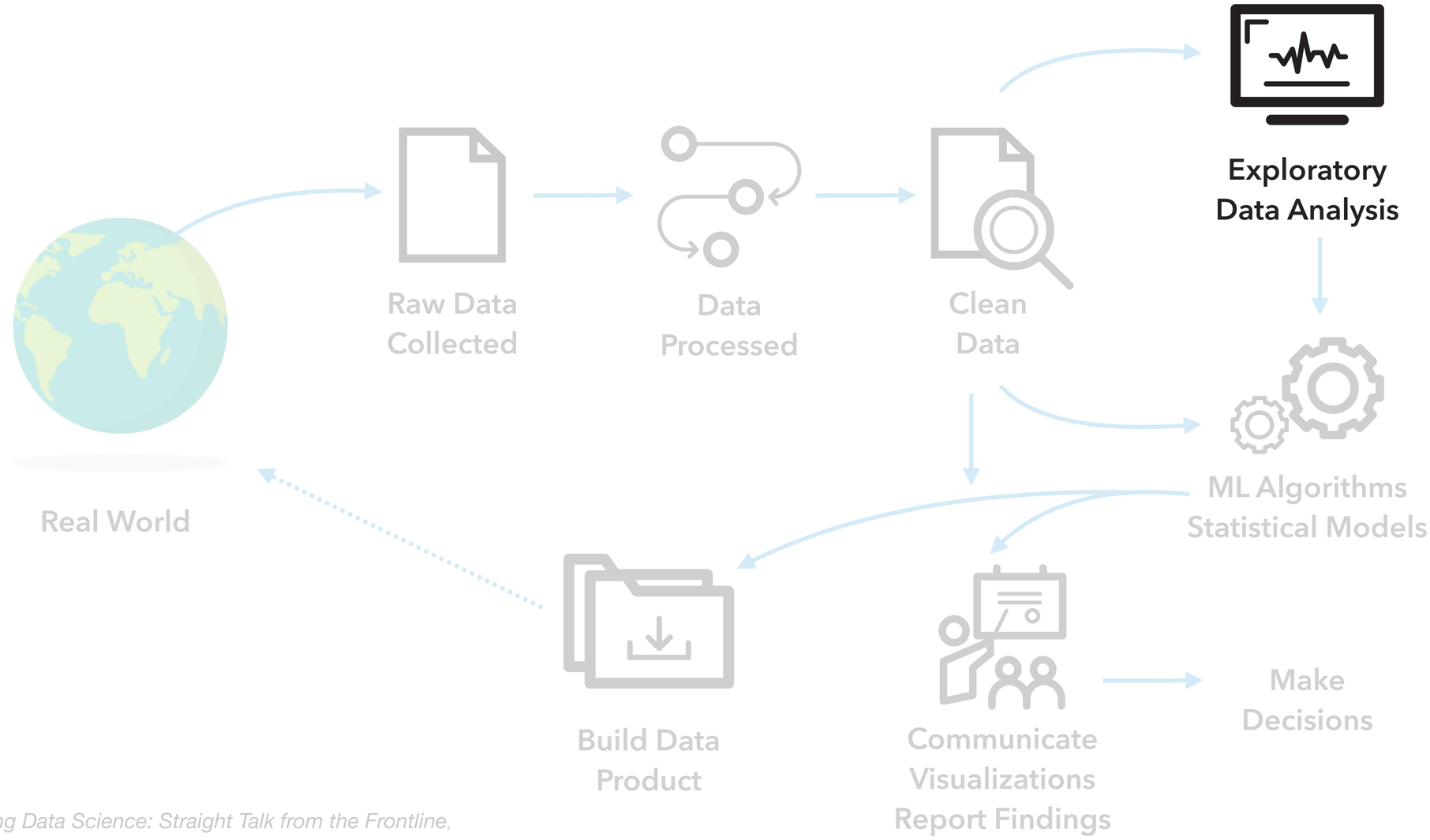
Human-Data Interaction

What's the Role of Visualization in Data Science?



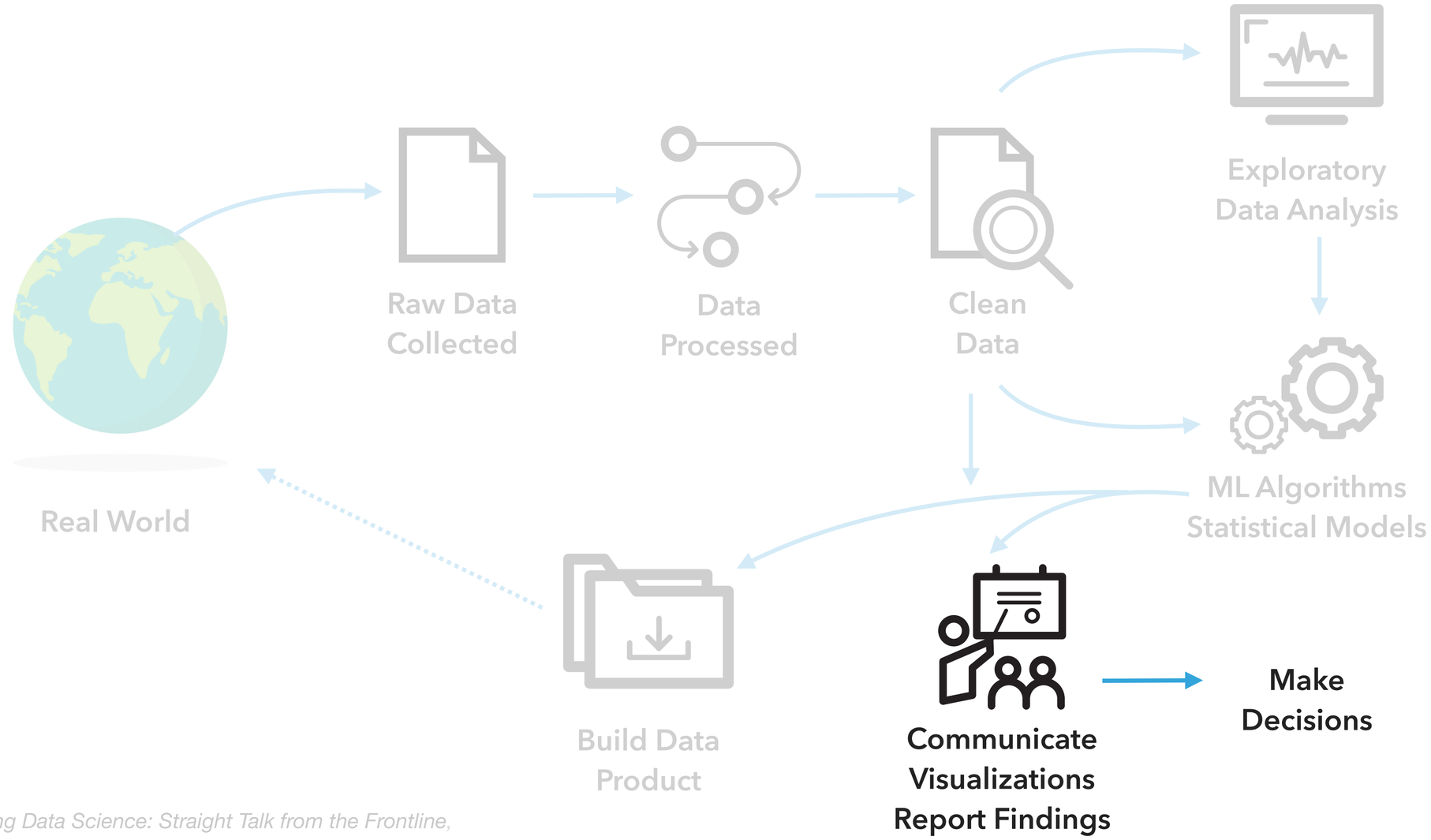
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

What's the Role of Visualization in Data Science?



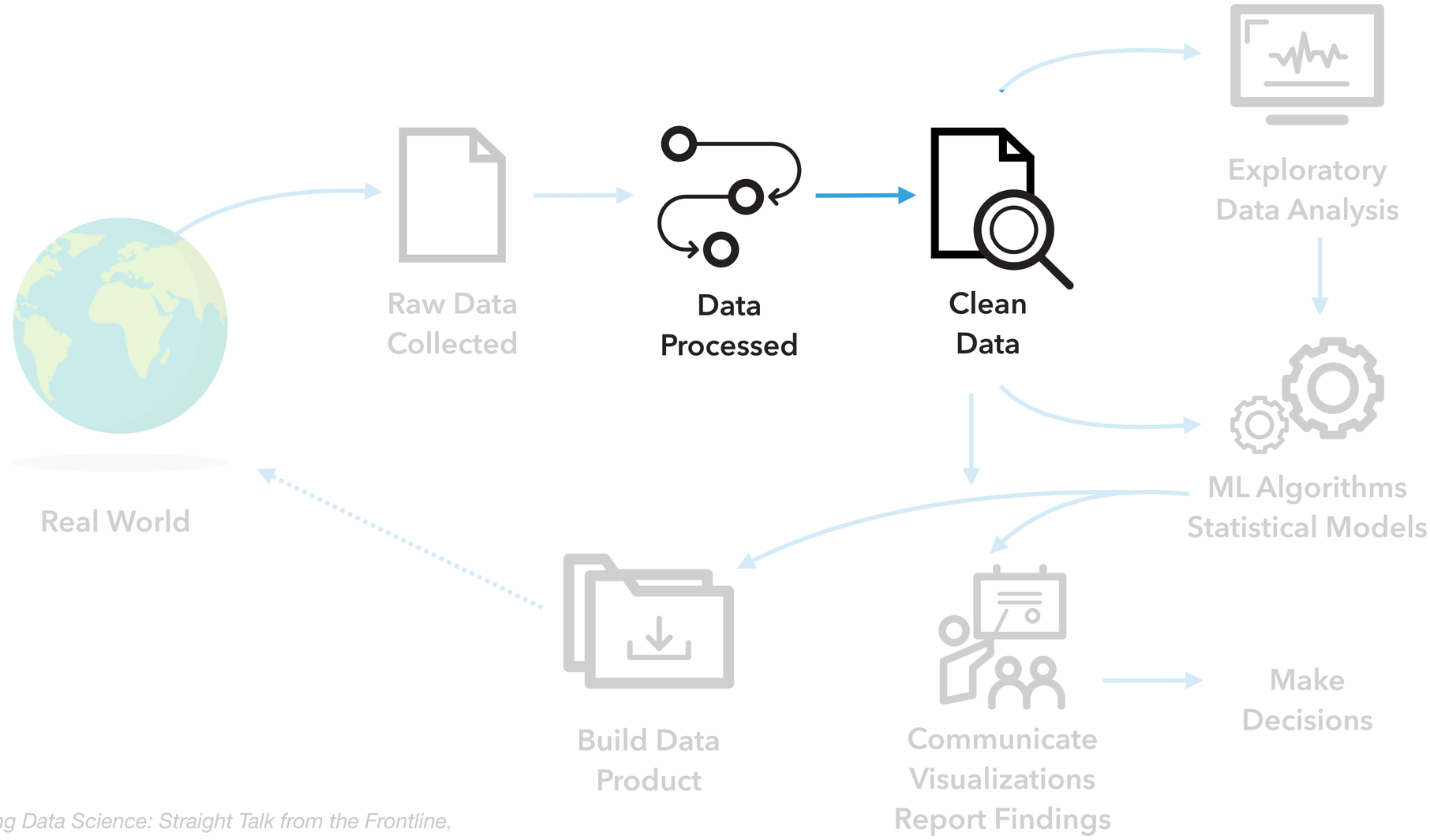
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

What's the Role of Visualization in Data Science?



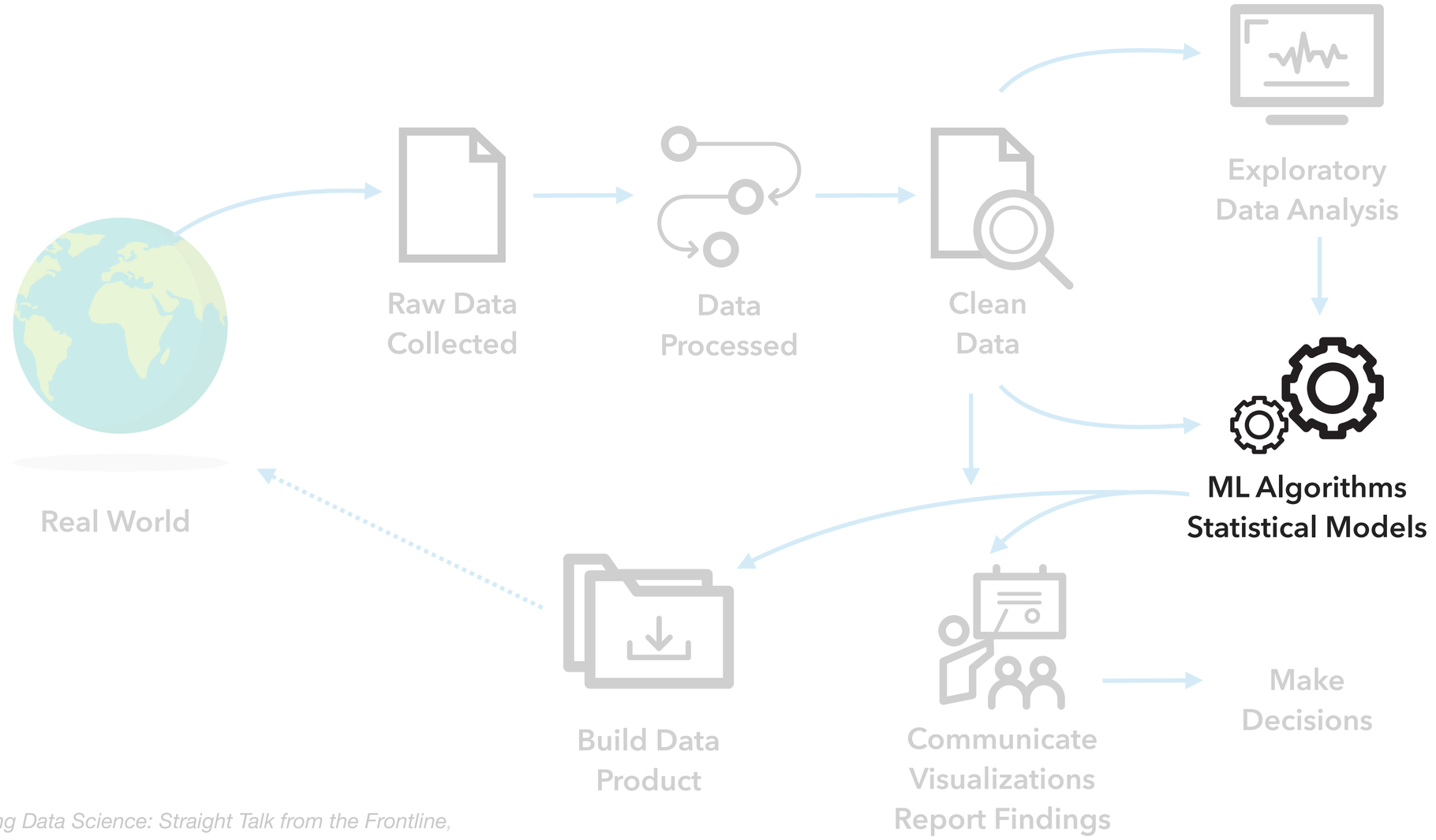
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

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?



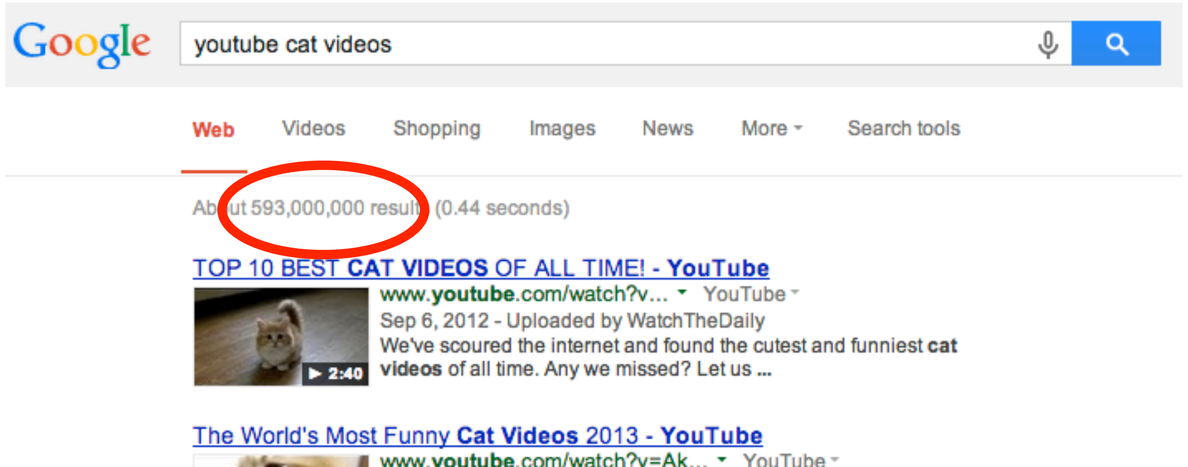
Adapted from Doing Data Science: Straight Talk from the Frontline, [O'Neil and Schutt, 2013]

Big Data

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

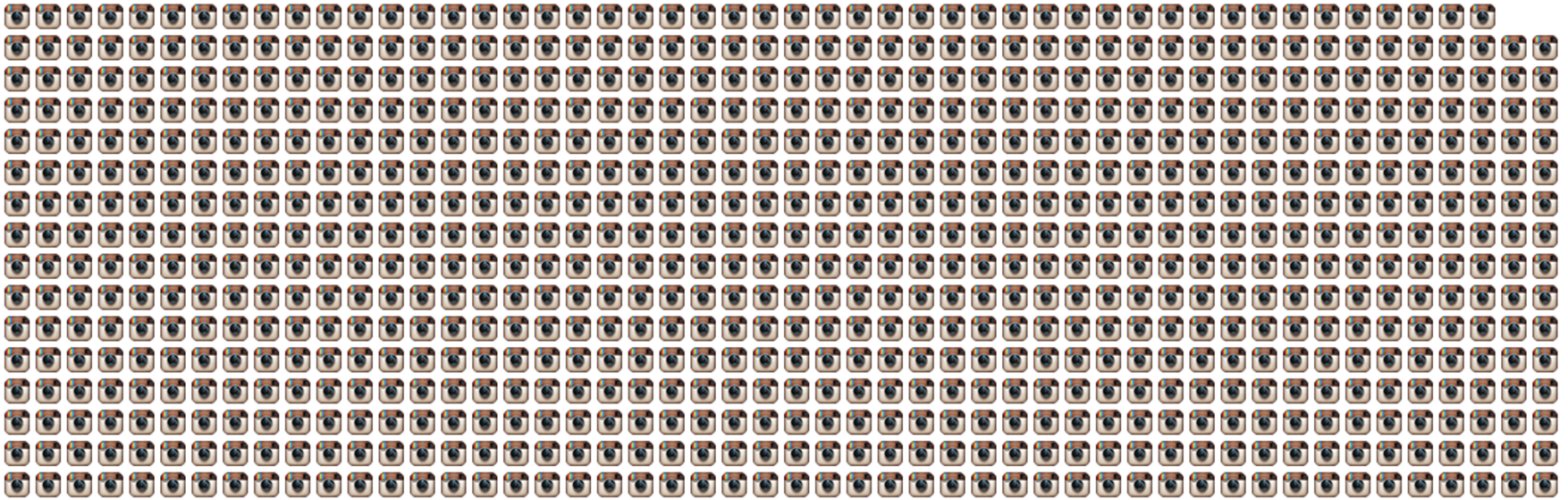
Source: IBM

90% of the data created in last two years



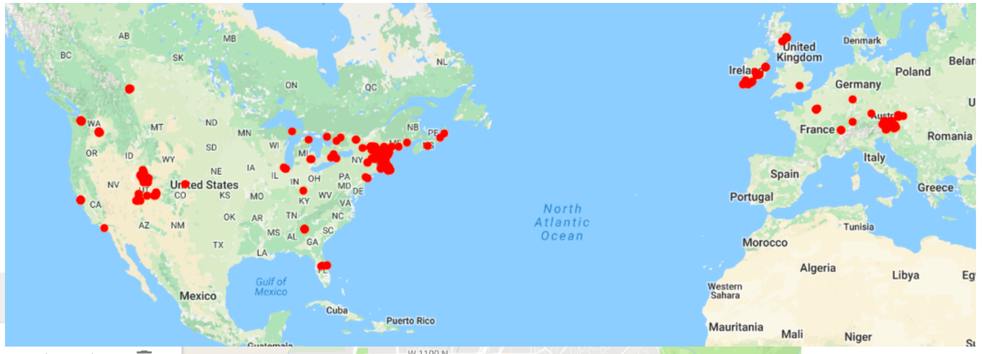
15 Exabytes in Punch Cards:
4.5 km over New England

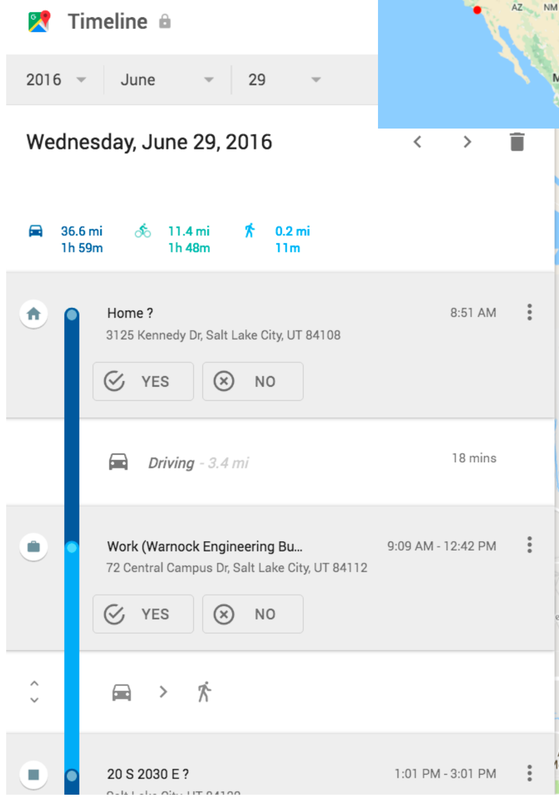
798 Instagram photos uploaded in 1 second

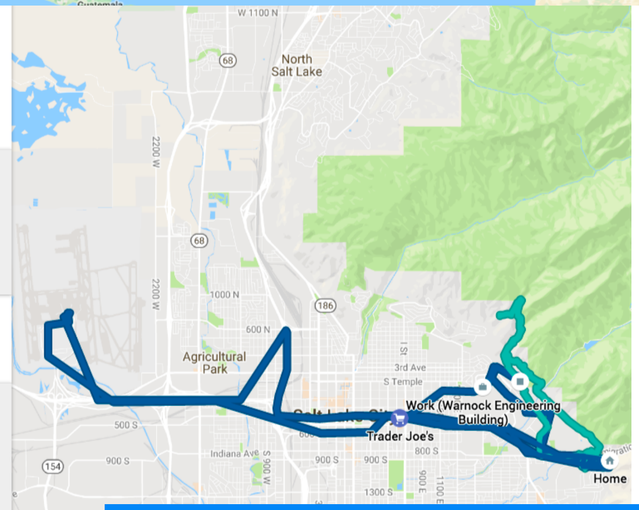


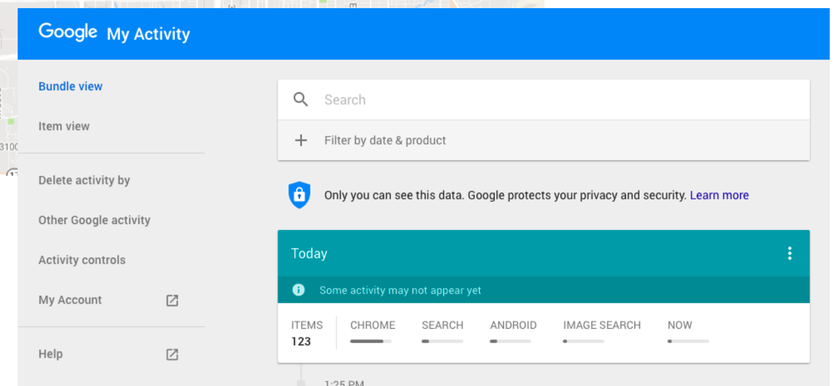
1,277 Tumblr posts in 1 second

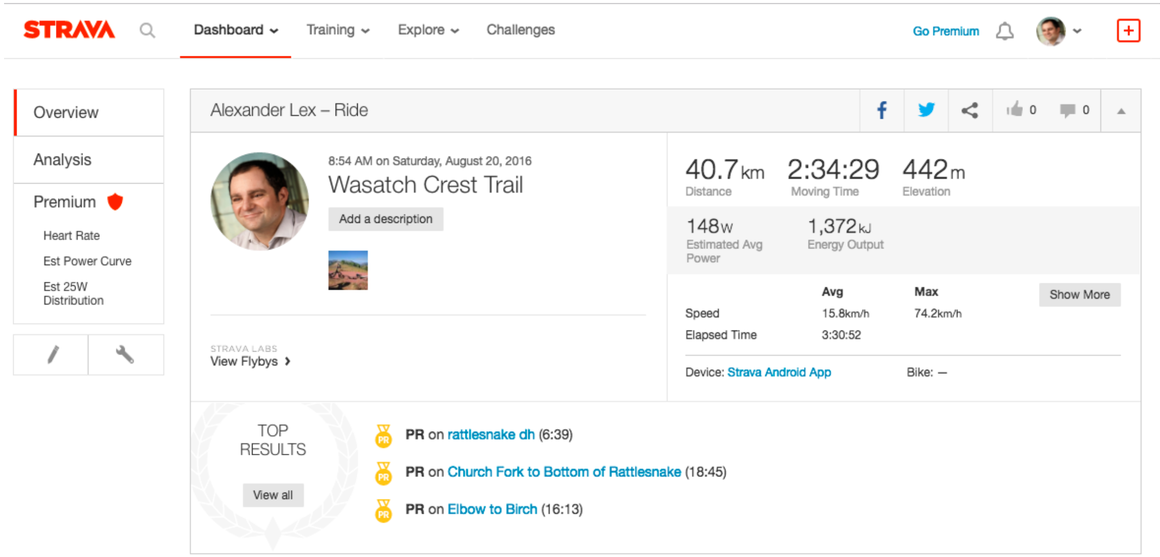
Example: Personal Data

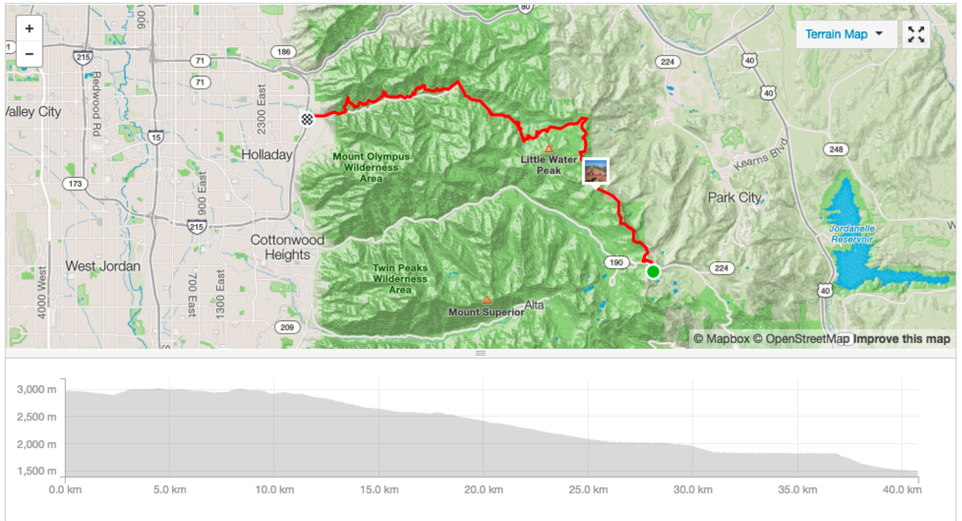












Big Data in Science and Engineering

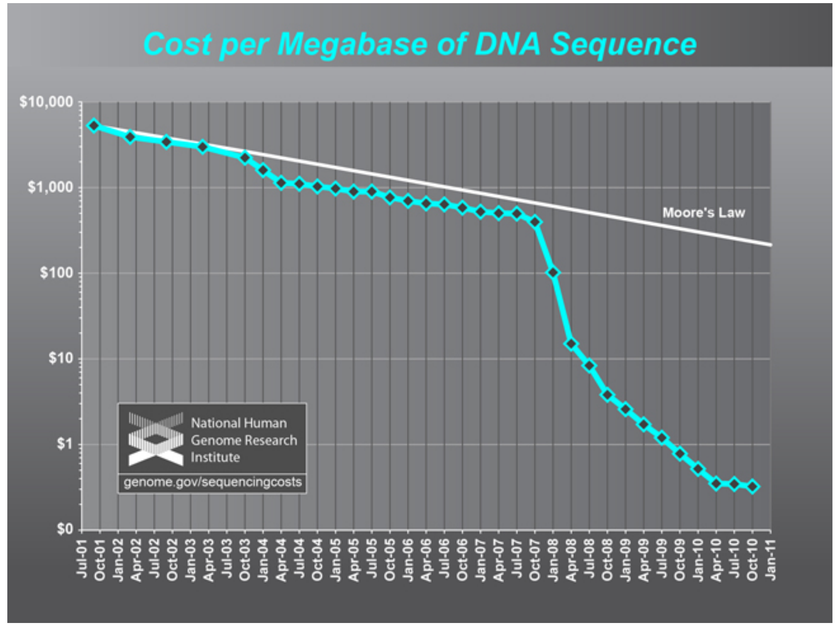
“Big Data” has transformed science and engineering.

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

Examples:

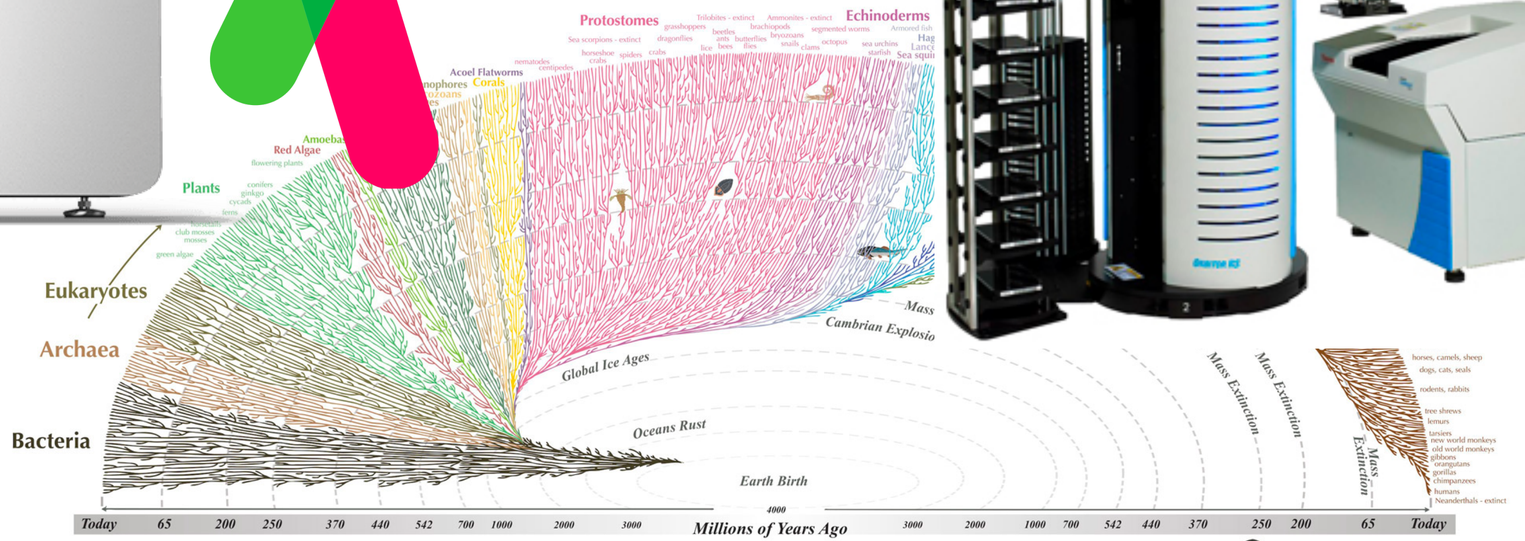
- Large physics experiments and observations
- Cheaper and automated genome sequencing
- Smart buildings / cities (blyncsy)

Controversy: Hypothesis or data driven methods



Example: Genomics

23andMe



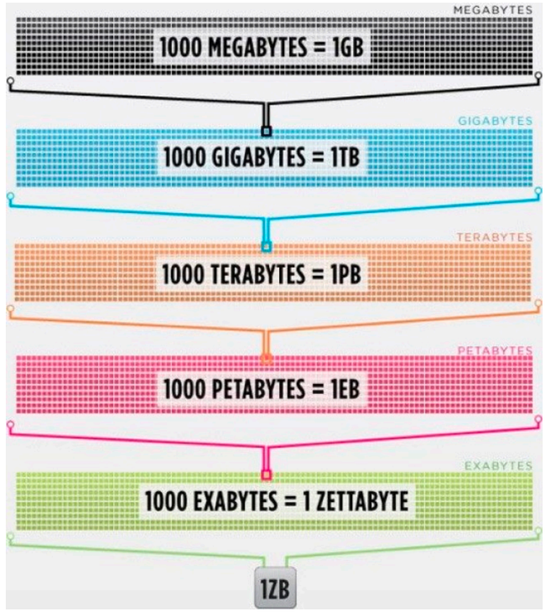
Example TCGA: 1 Petabyte

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

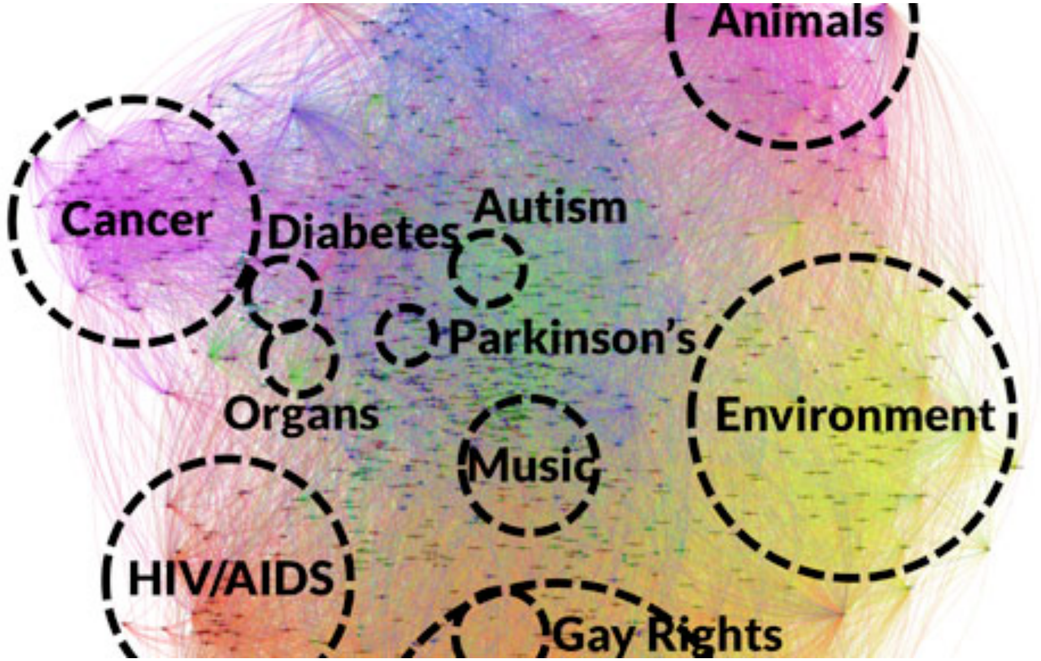
Why Humans?

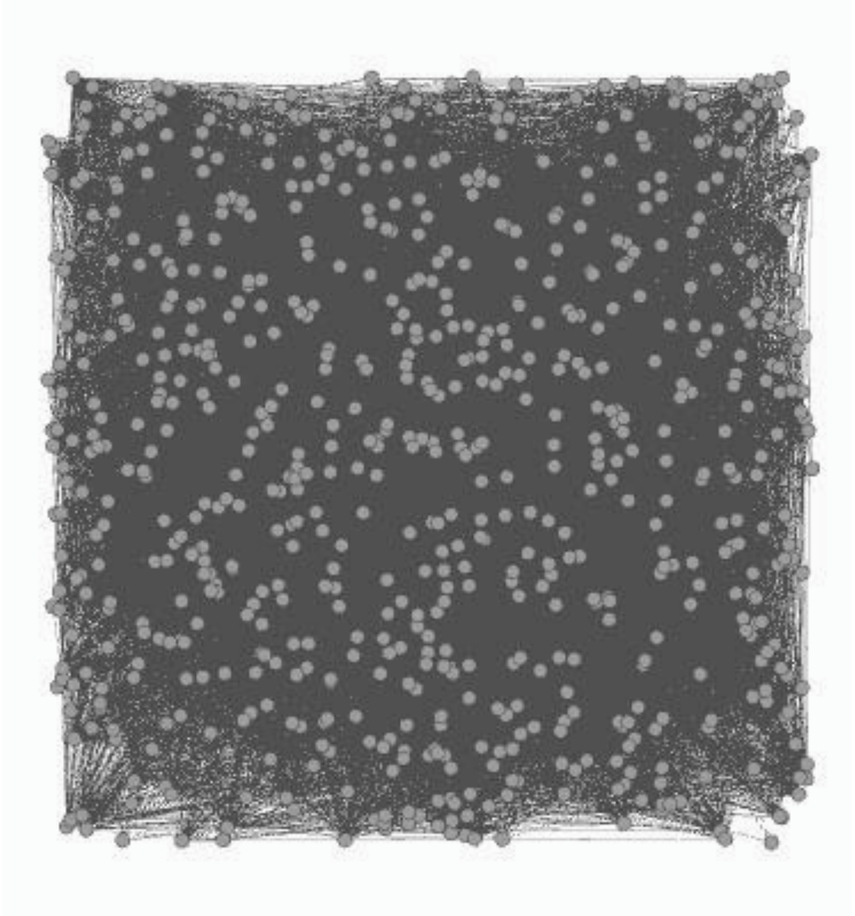
Leveraging human capabilities

- **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

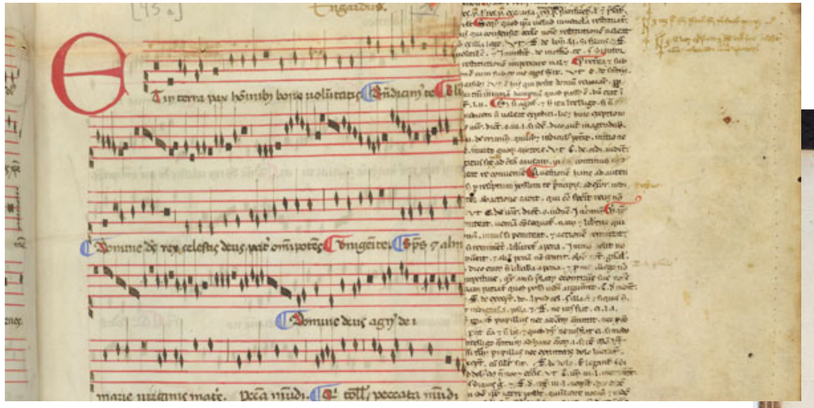
Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

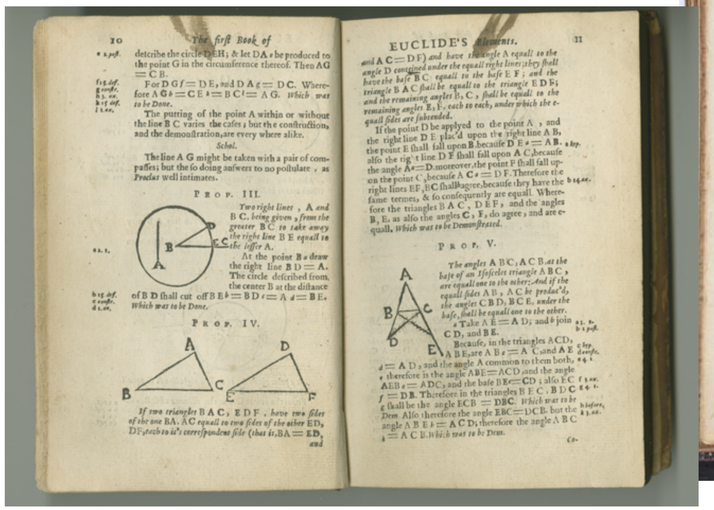
How did we get here?

A bit of history

“It is things that make us smart”

Donald A. Norman





10 *The first Book of*

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

For DG (= DE,) and DA (= DC. Wherefore A G (= CE & = BC (= AG. *Which was to be Done.*

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

Schol.

The line A G 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* B C, *being given, from the greater* B C *to take away the right line* B E *equall to the other lesse* A.

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

PROP. IV.

If two triangles B A C, E D F, *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 A C = DF) *and have the angle* A *equall to the angle* D *contained under the equall right lines; they shall have the base* B C *equall to the base* E F; *and the triangle* B A C *shall be equall to the triangle* E D F; *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 D E laid upon the right line A B, the point E shall fall upon B, because D E = AB. Also the right line D F shall fall upon A C, because the angle A = D. moreover, the point F shall fall upon the point C, because A C (= DF. Therefore the right lines EF, B C shall agree, because they have the same termes, & so consequently are equall. Wherefore the triangles B A C, D E F, and the angles B, E, as also the angles C, F, do agree, and are equall, *Which was to be Demonstrated.*

PROP. V.

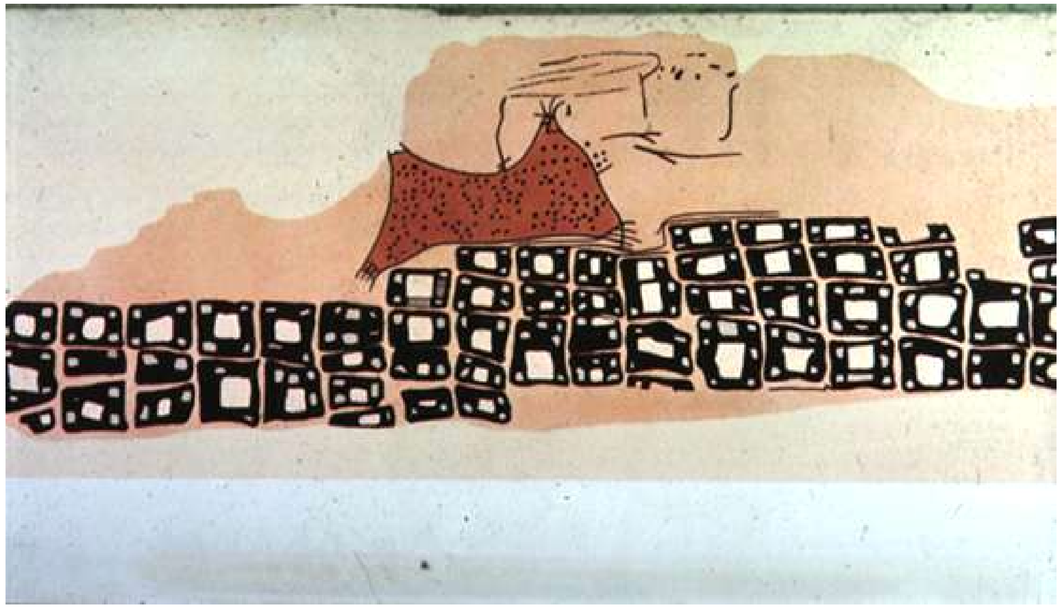
The angles A B C, A C B, *at the base of an Isosceles triangle* A B C, *are equall one to the other; And if the equall sides* A B, A C *be produc'd, the angles* C B D, B C E, *under the base, shall be equall one to the other.*

Take A E = A D; and join C D, and B E.

Because, in the triangles A C D, A B E, are A B (= A C, and A E = A D, and the angle A common to them both, therefore is the angle A B E = A C D and the angle A E B (= A D C, and the base B E = C D; also E C (= D B. Therefore in the triangles B E C, B D C shall be the angle E C B = D B C. *Which was to be Shewn.* Also therefore the angle E B C = D C B. but the angle A B E (= A C D; therefore the angle A B C = A C B. *Which was to be Shewn.*

Co-

Record



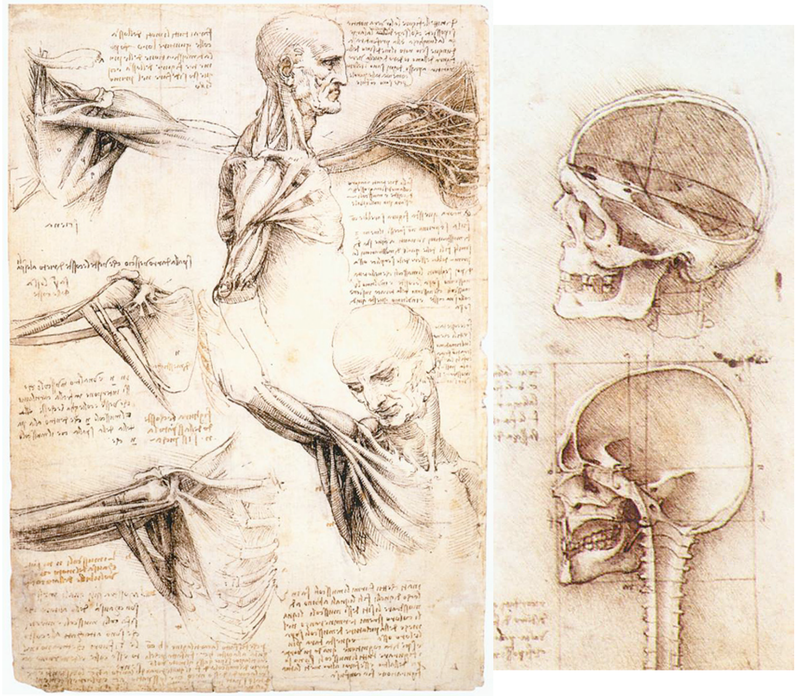
Konya town map, Turkey, c. 6200 BC

Anaximader's Map of the World

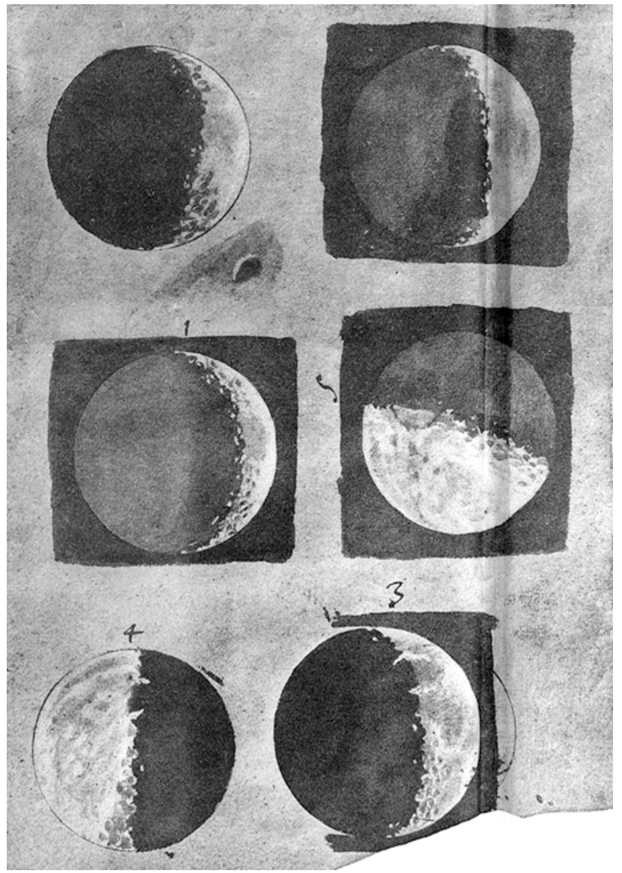
Anaximander of Miletus, c. 550 BC

Milestones Project

Record



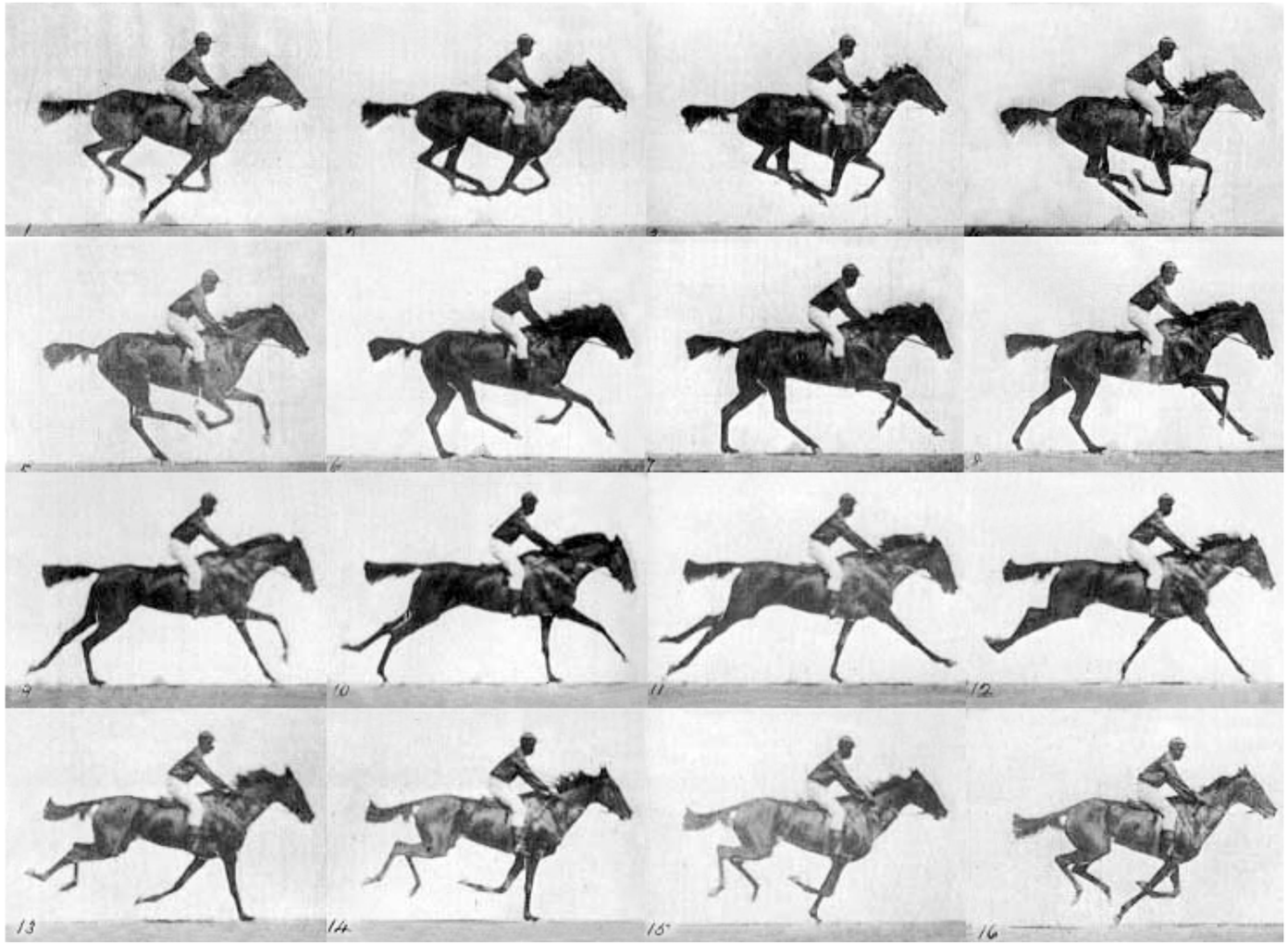
Leonardo Da Vinci, ca. 1500



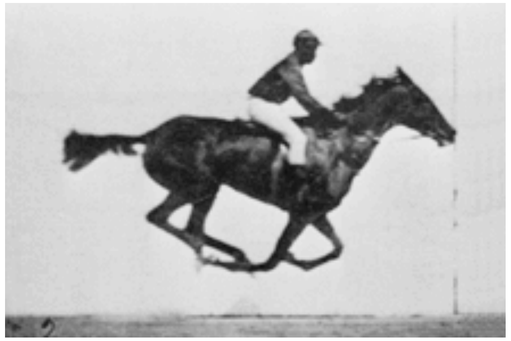
Galileo Galilei, 1616

William Curtis (1746-1799)

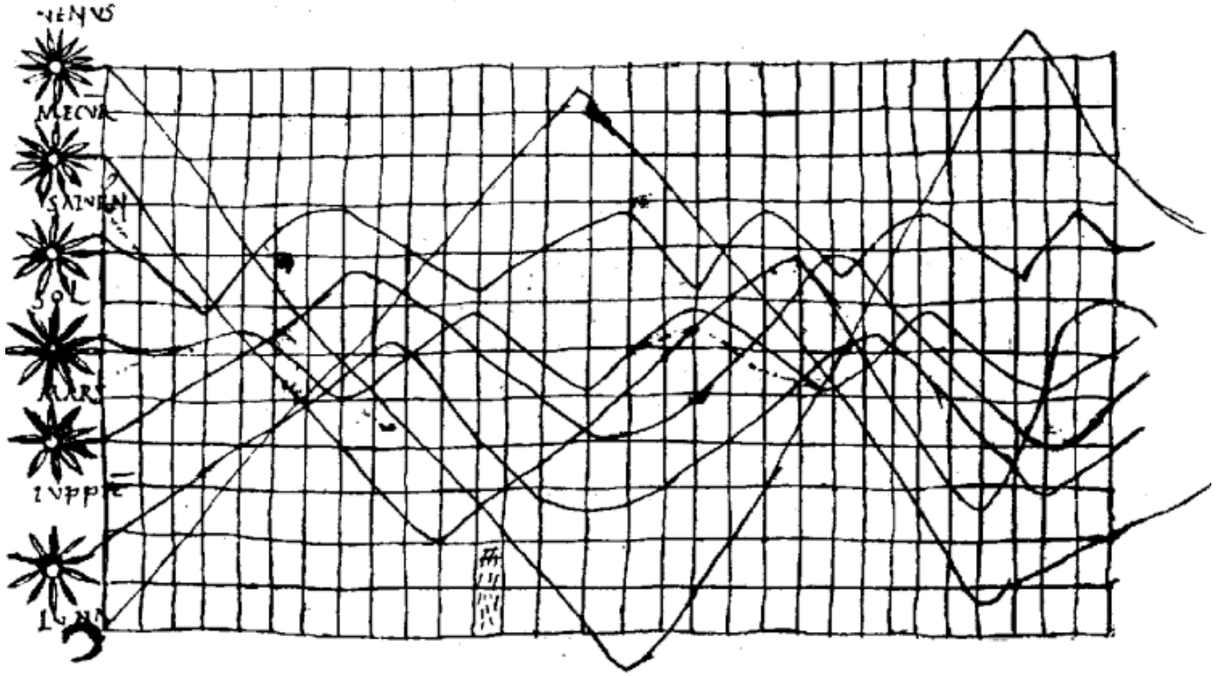
Record



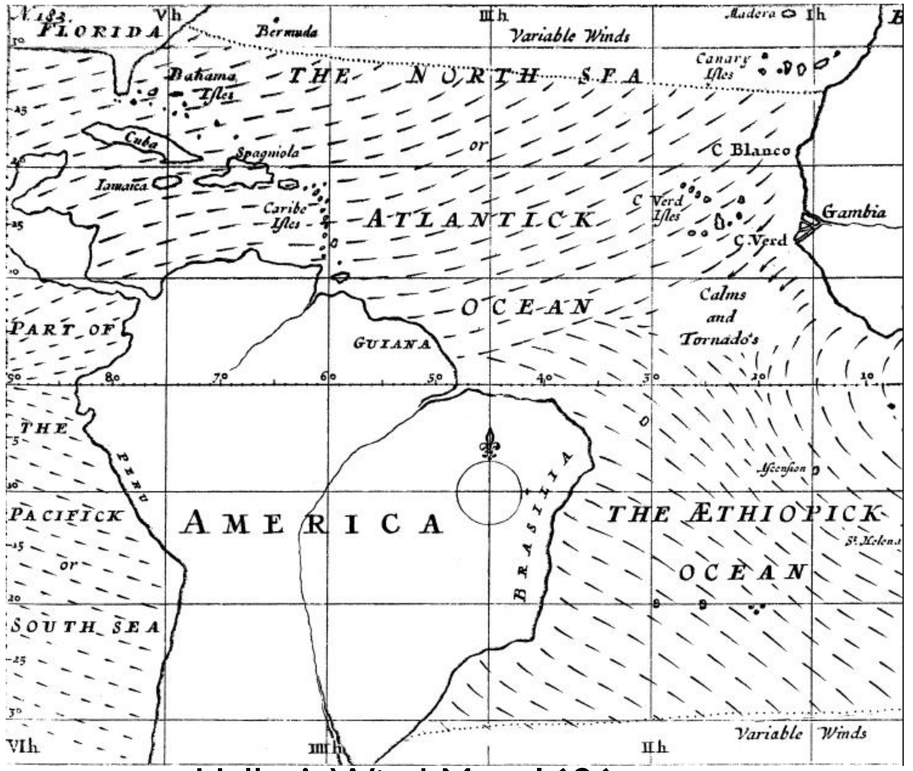
Eadweard J. Muybridge, 1878



Analyze

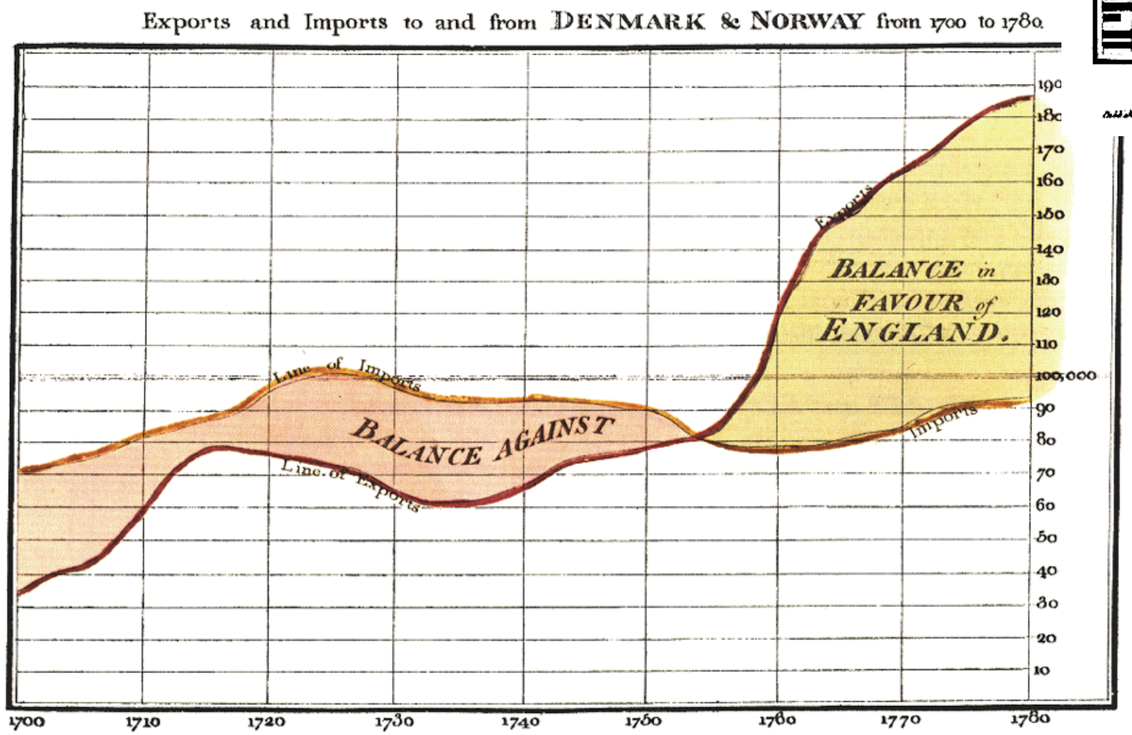


Planetary Movement Diagram, c. 950

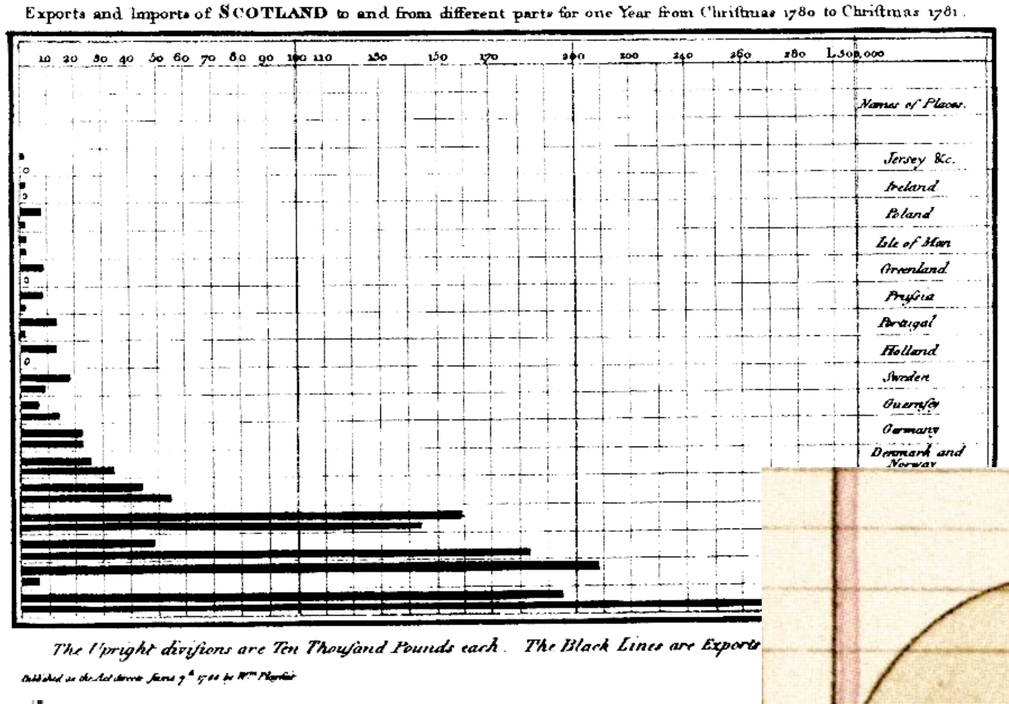


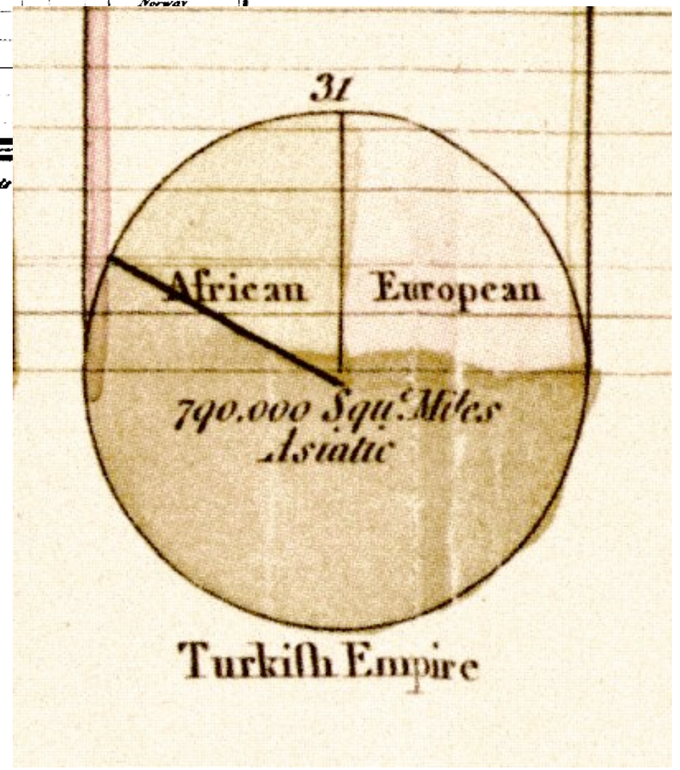
Halley's Wind Map, 1686

Analyze



W. Playfair, 1786



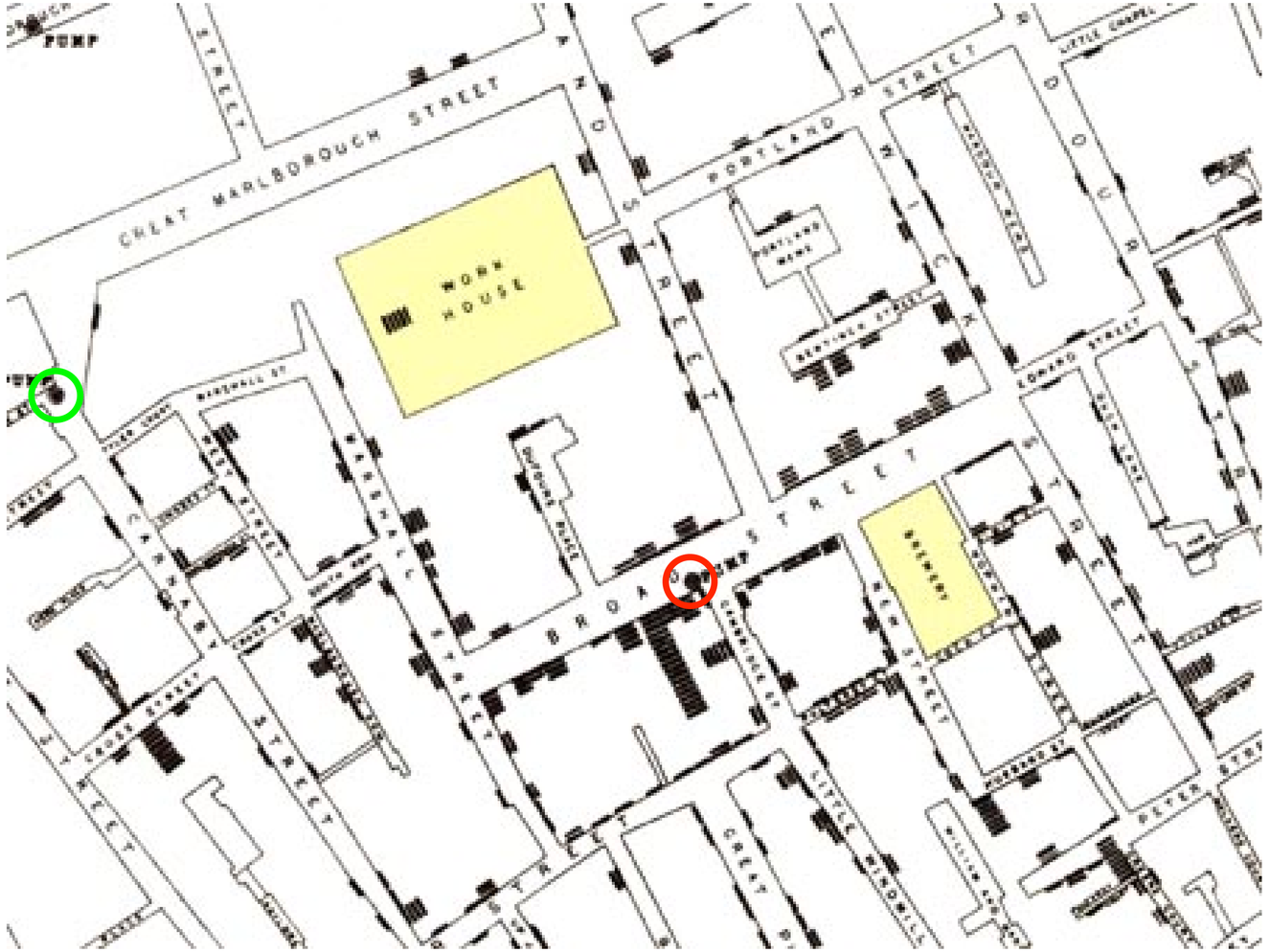


proportions of the Turkish Empire located in Asia, Europe and Africa before 1789

wikipedia.org

W. Playfair, 1801

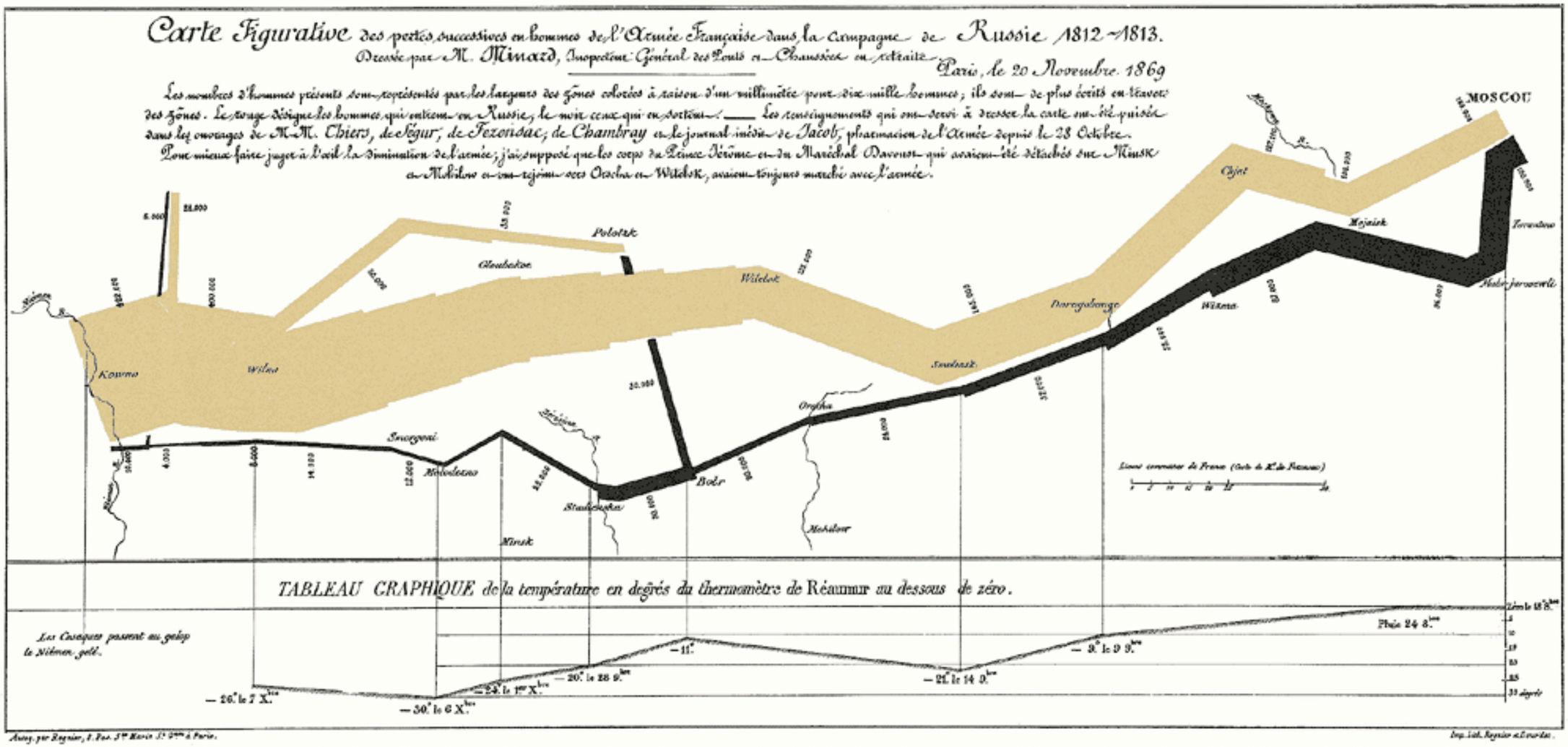
Find Patterns



John Snow, 1854

E. Tufte, Visual Explanations, 1997

Communicate

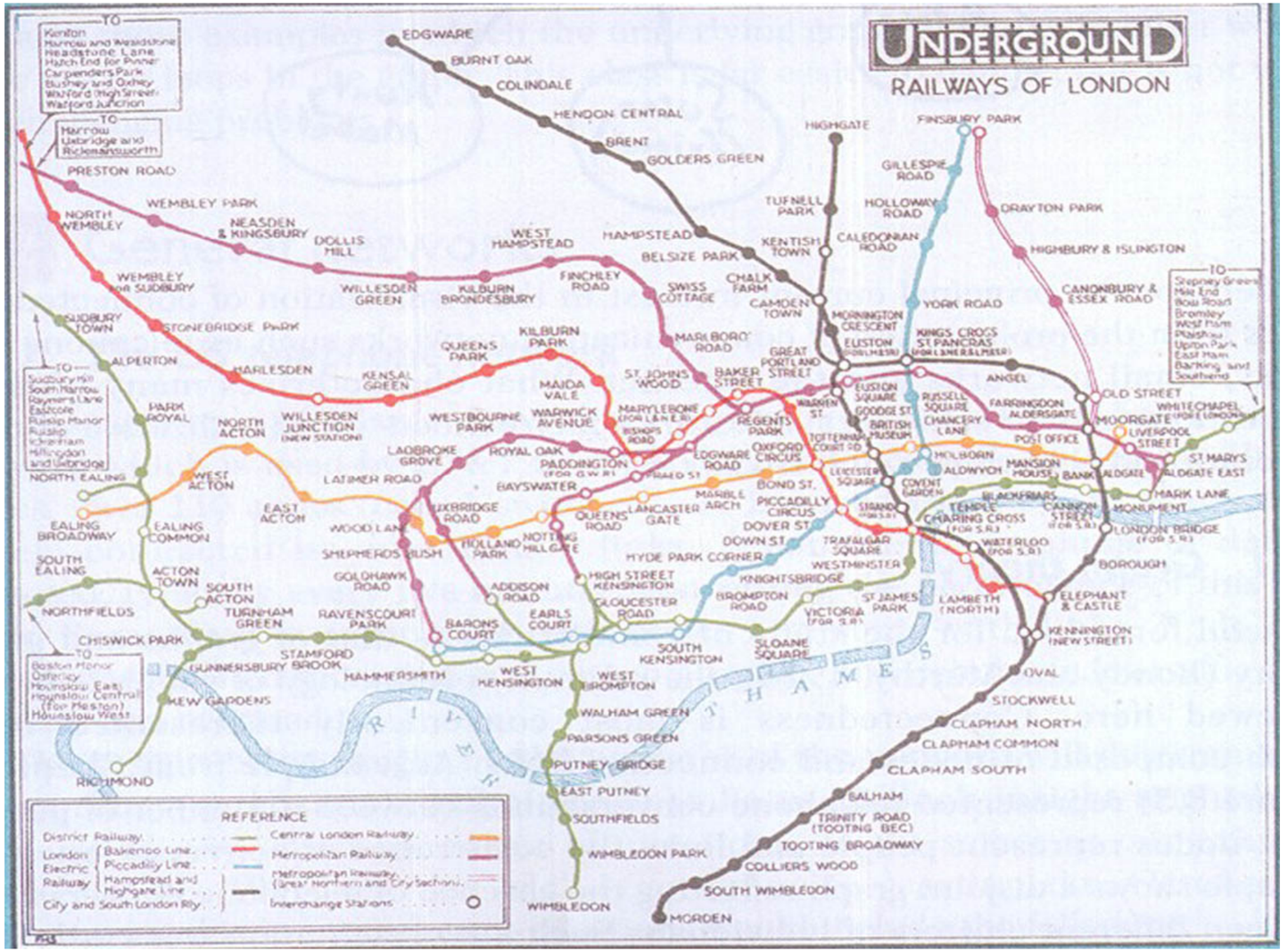


C.J. Minard, 1869

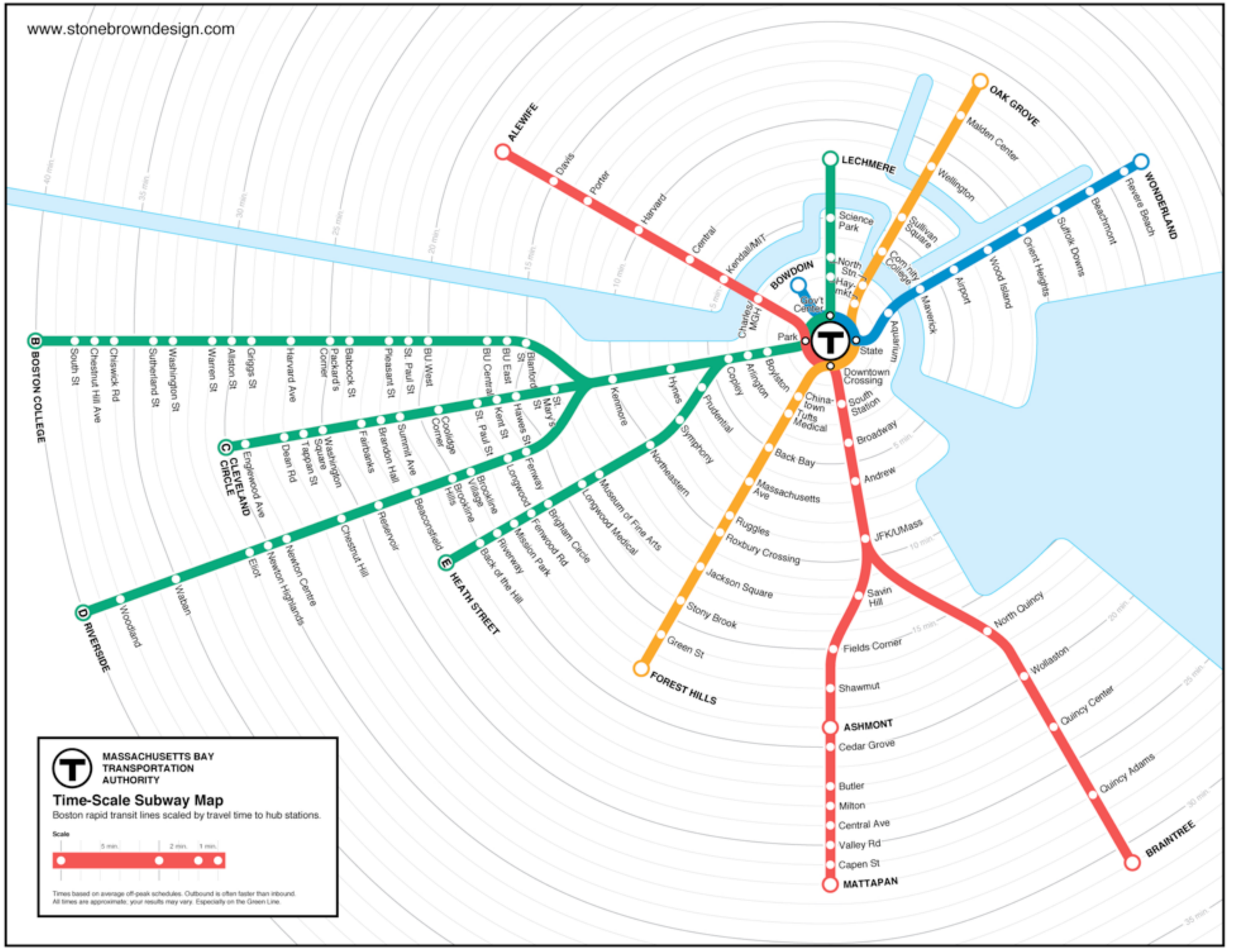
E. Tufte, Writings, Artworks, News

http://infowetrust.com/scroll/

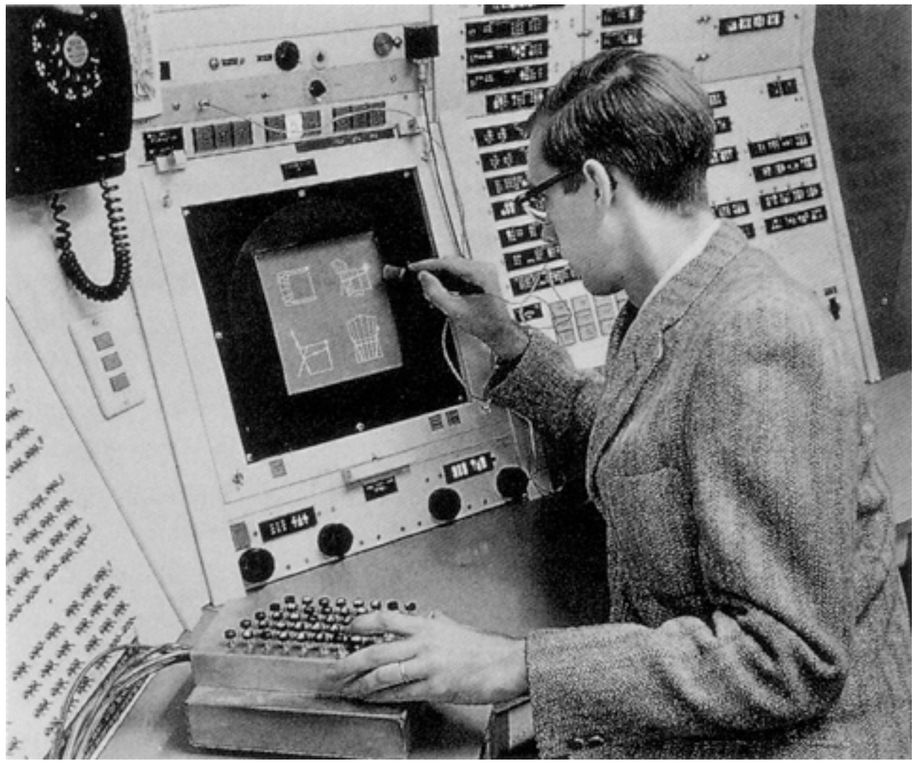
Communicate



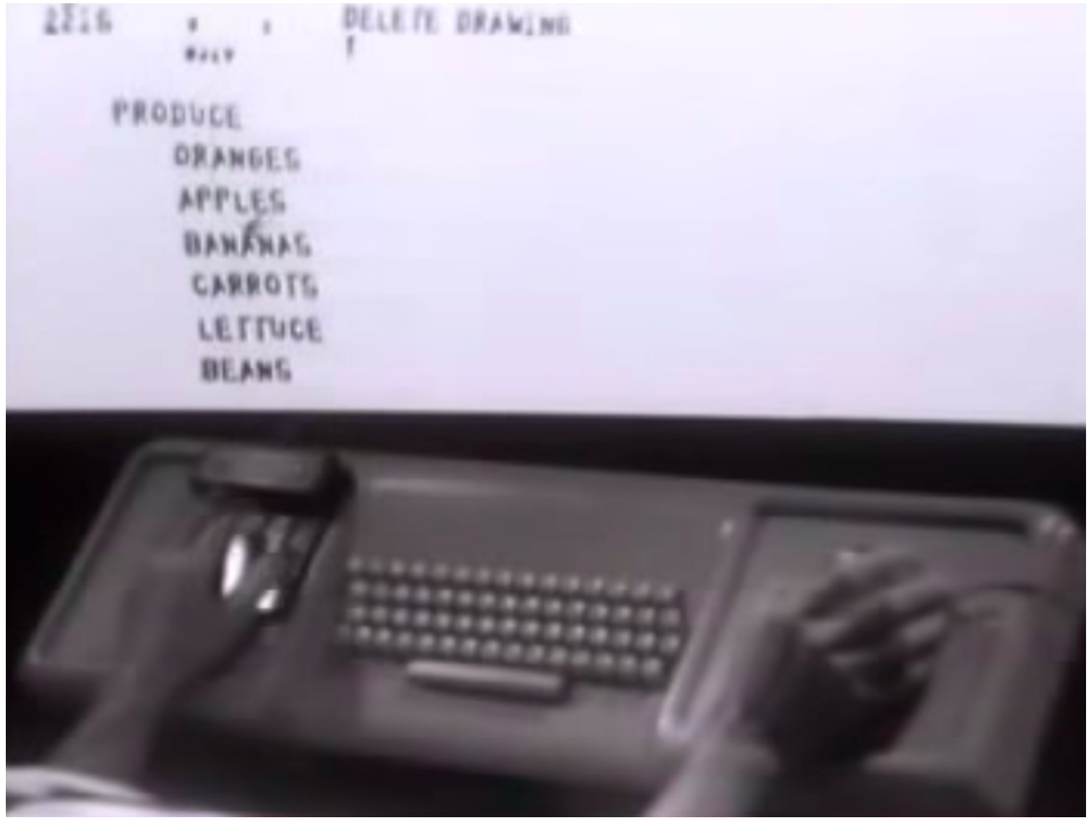
London Subway Map, 1927


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



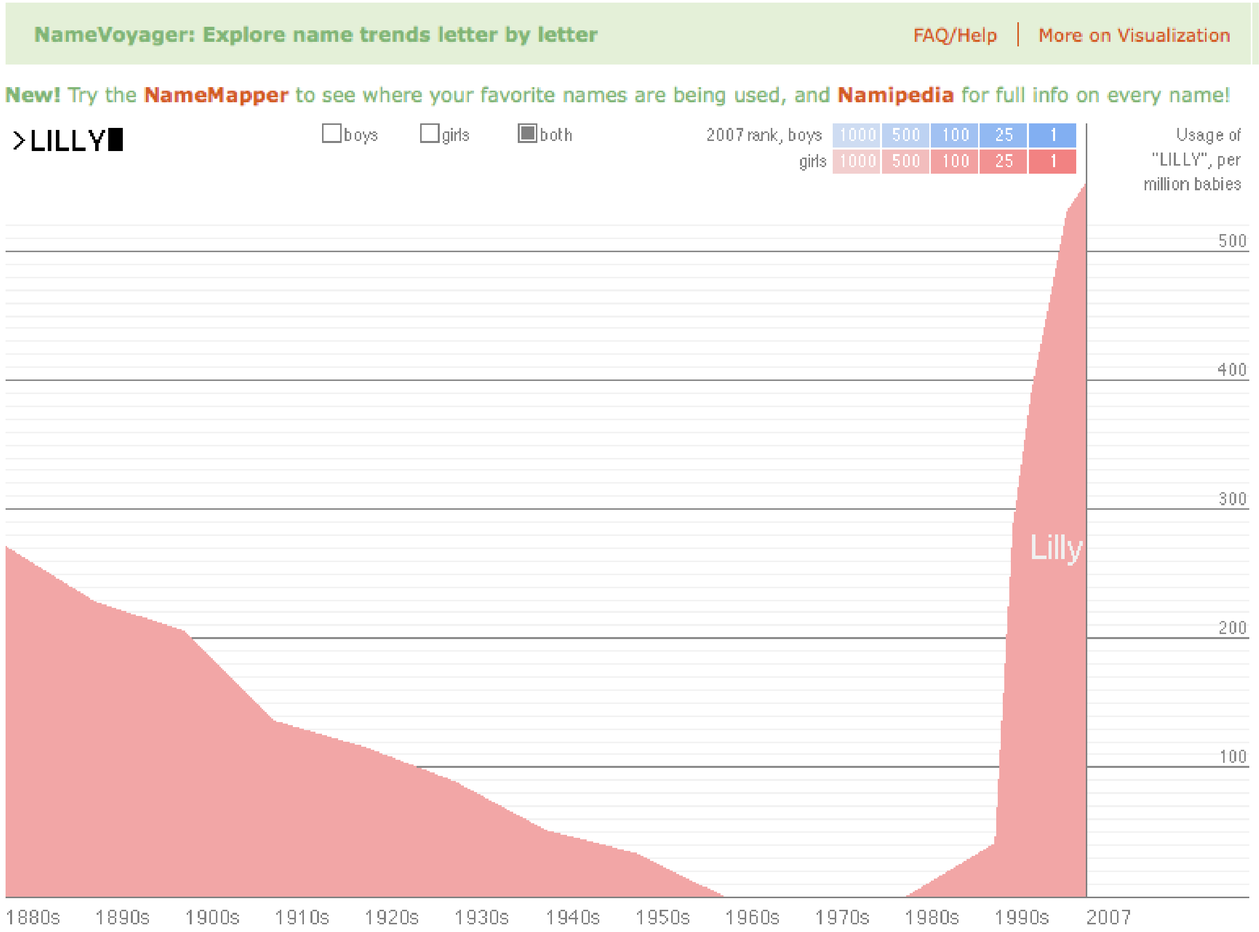
Ivan Sutherland, Sketchpad, 1963



Doug Engelbart, 1968

Modern Examples

Analyze



M. Wattenberg, 2005

Communicate

Hans Rosling, TED 2006

Who is CS-5630 / CS-6630?

Course Staff

Devin Lange

Teaching Mentee

Haihan Lin

Teaching Mentee

Youjia Zhou

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

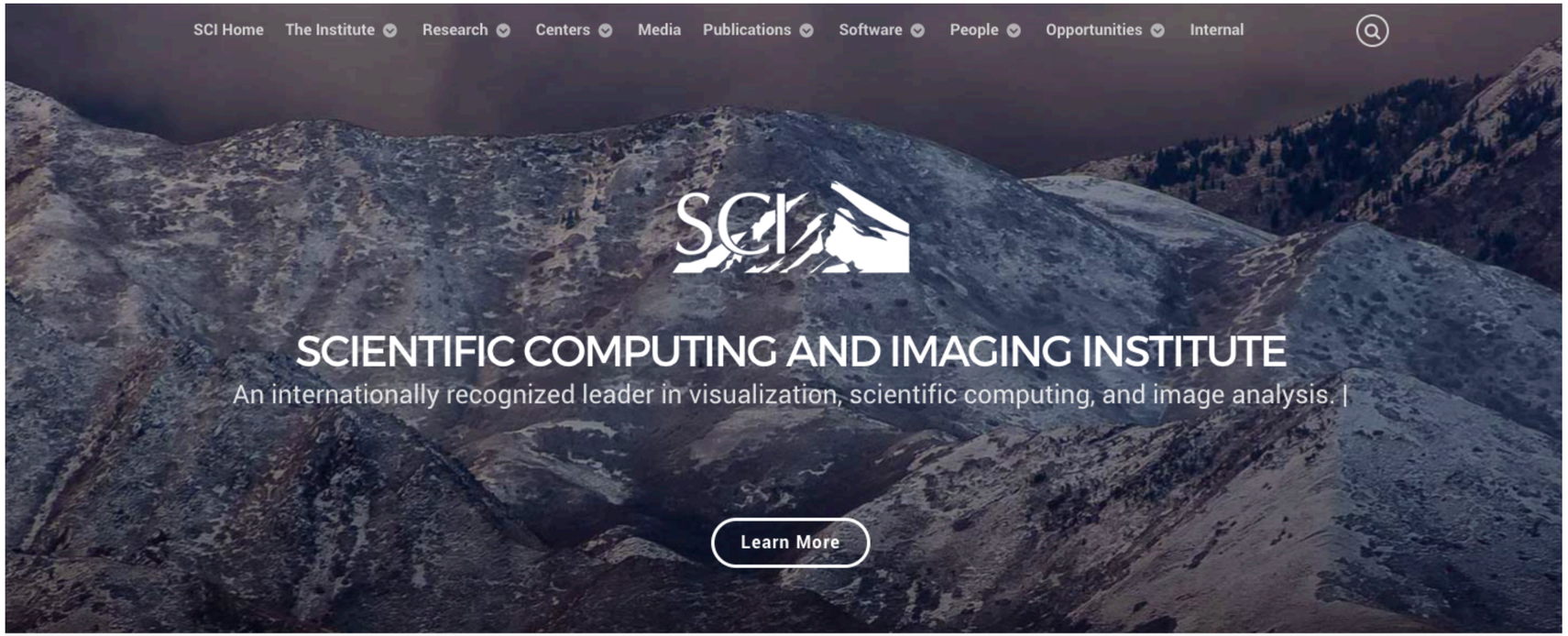
http://vdl.sci.utah.edu/

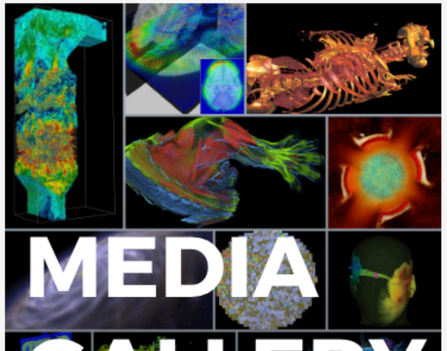
SCI Institute

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

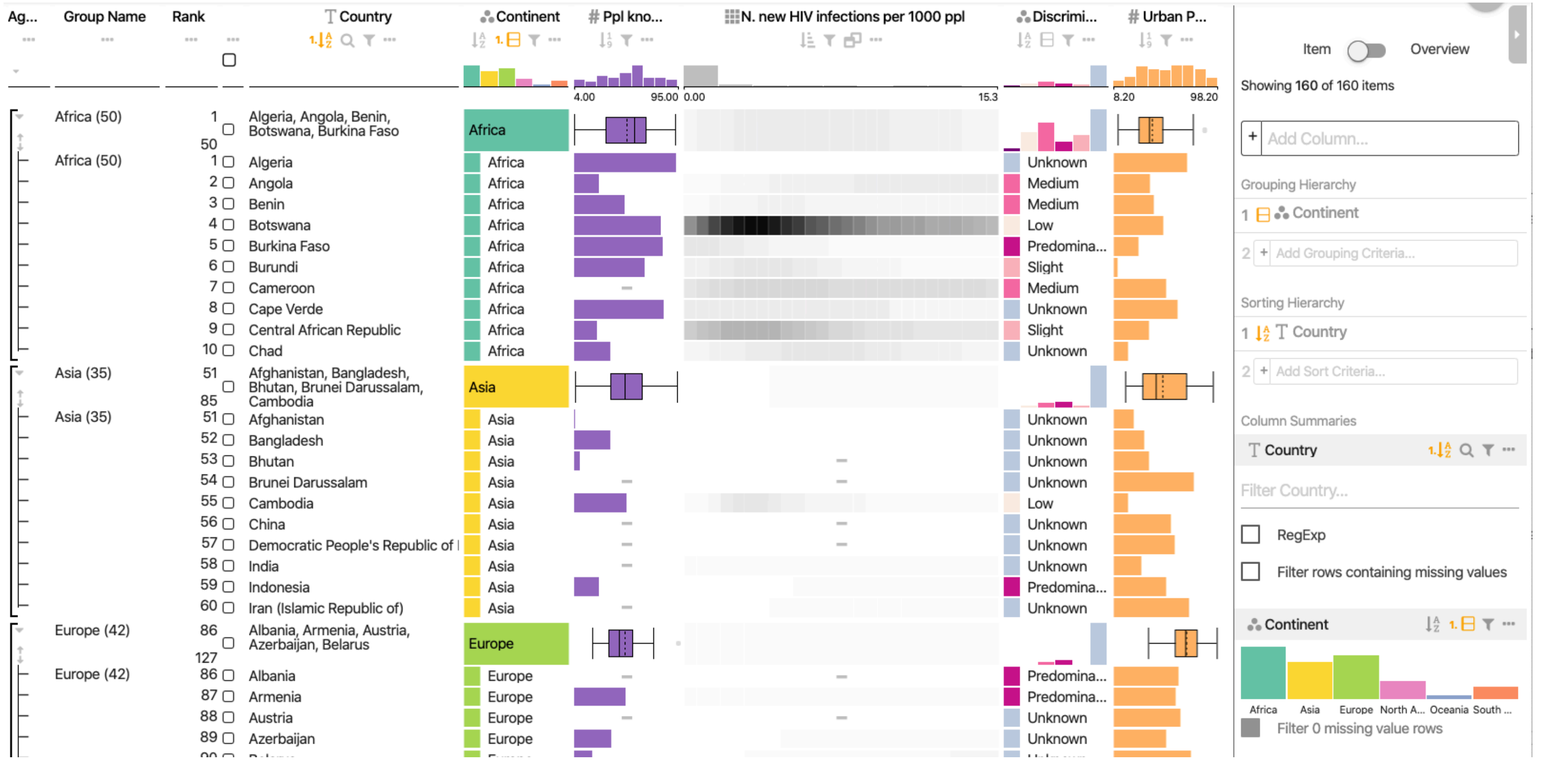
Evaluation Methodology

Design Spaces / Taxonomies

TECHNICAL CONTRIBUTIONS

Tabular Data

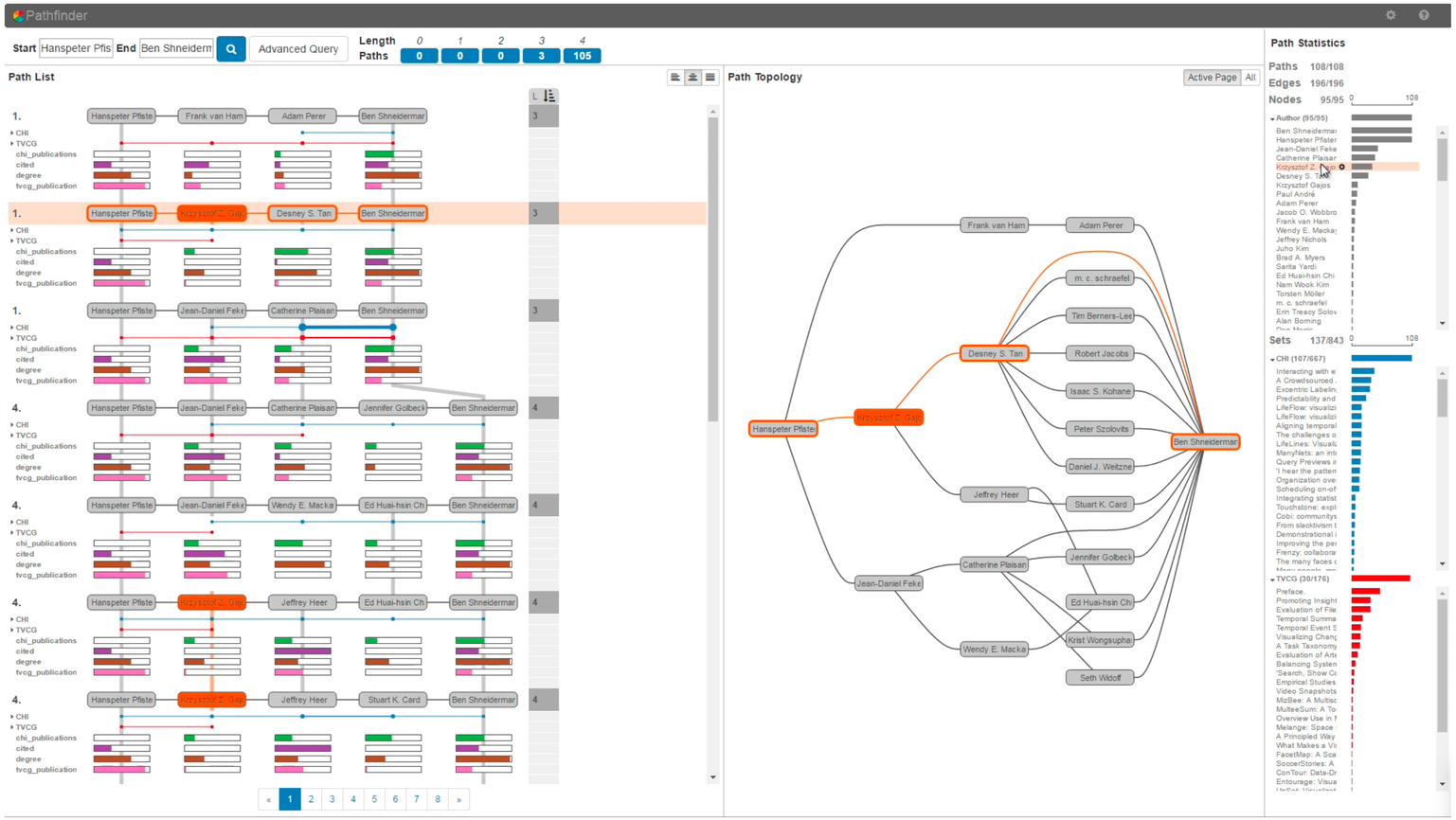
Novel Visualization Techniques



TECHNICAL CONTRIBUTIONS

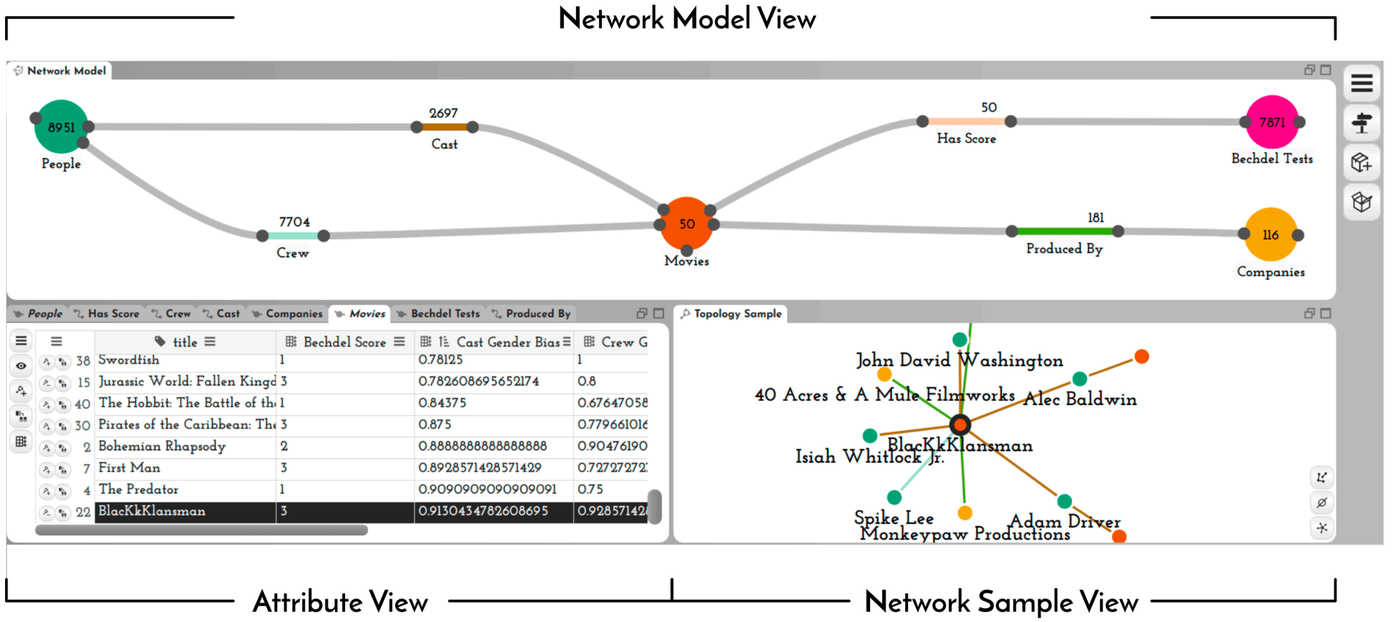
Multivariate Networks

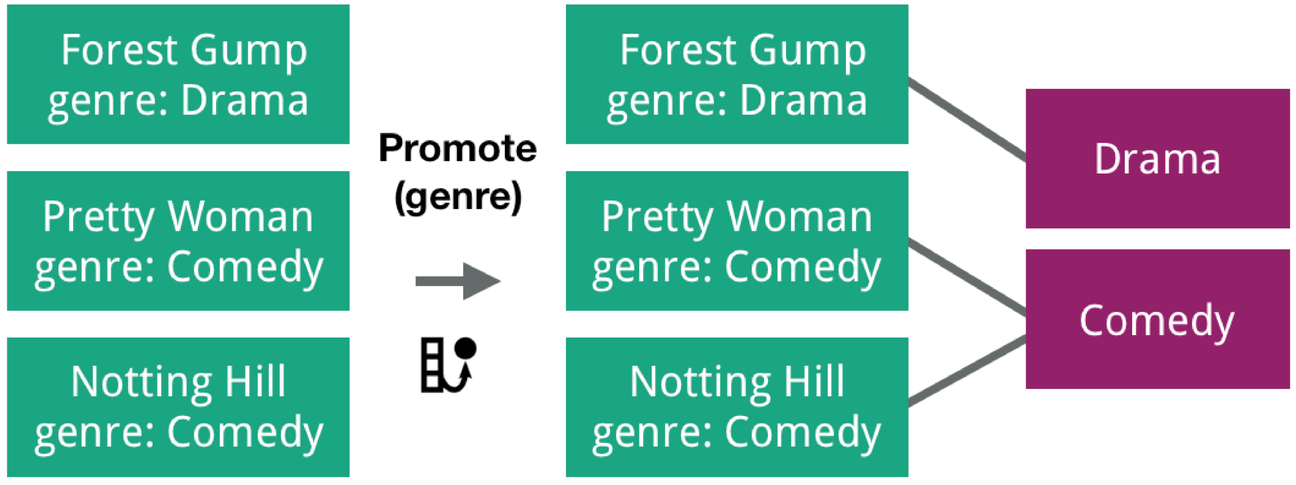
Novel Visualization Techniques



Reshaping Networks

Data Wrangling Methods





TECHNICAL CONTRIBUTIONS

Data Wrangling Methods

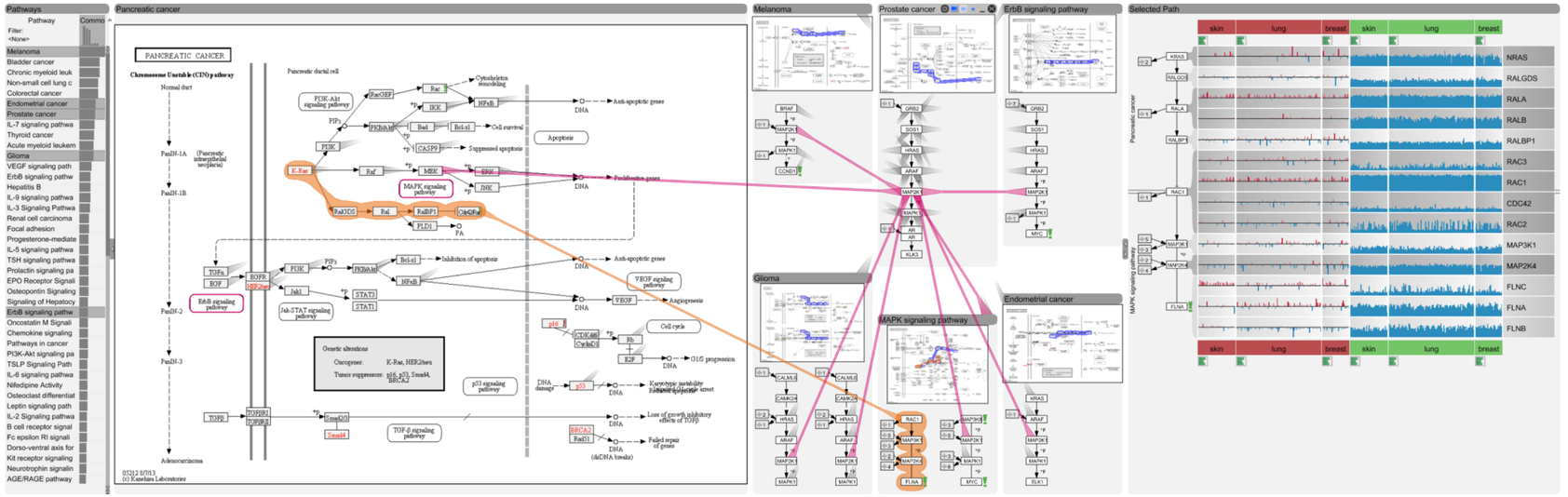
DOMAIN DRIVEN TECHNIQUES

Tailored Methods and Systems for High Impact Science Problems

DOMAIN DRIVEN TECHNIQUES

Pathways and Omics Data

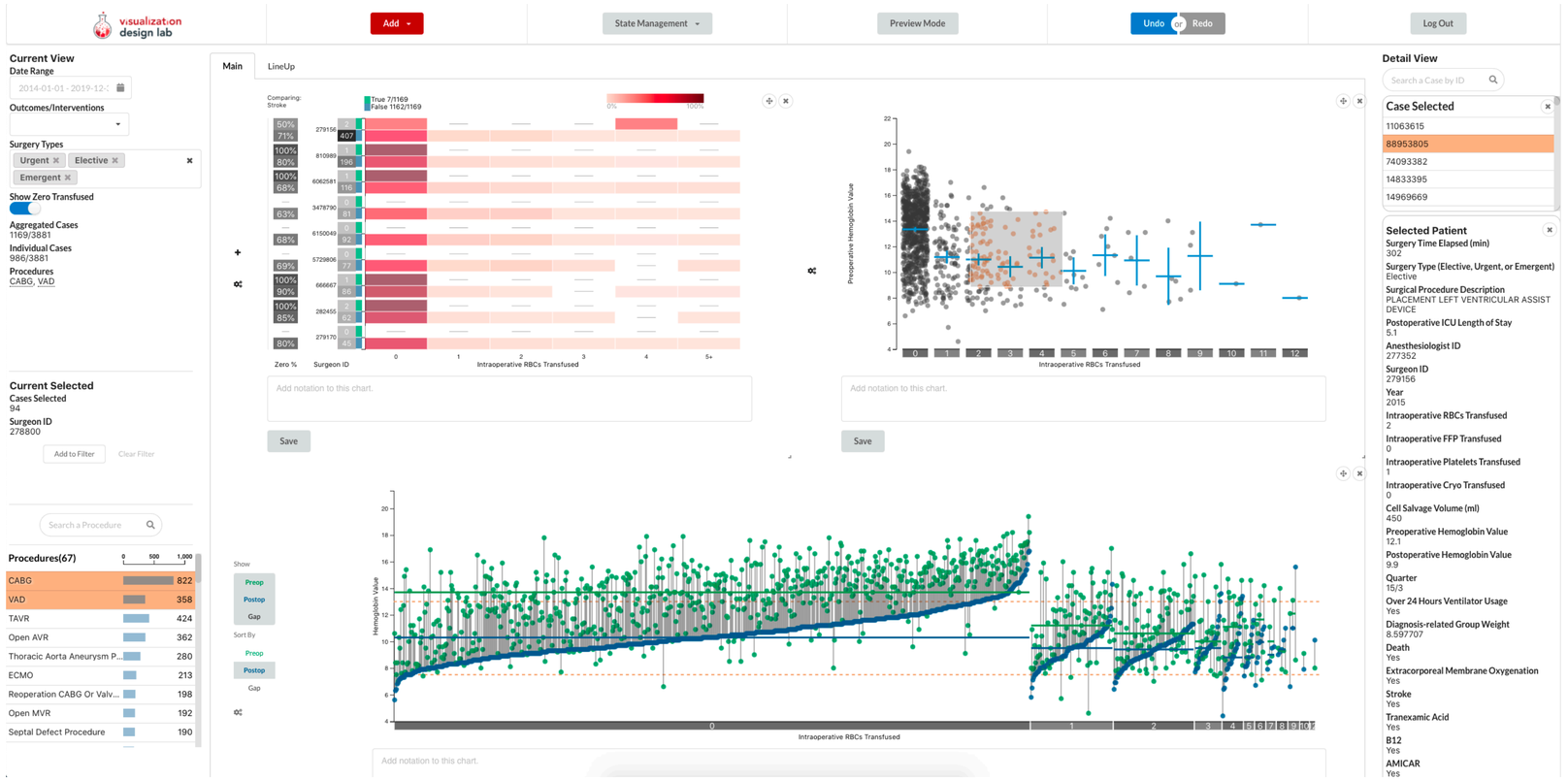
Tailored Methods and Systems for High Impact Science Problems



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

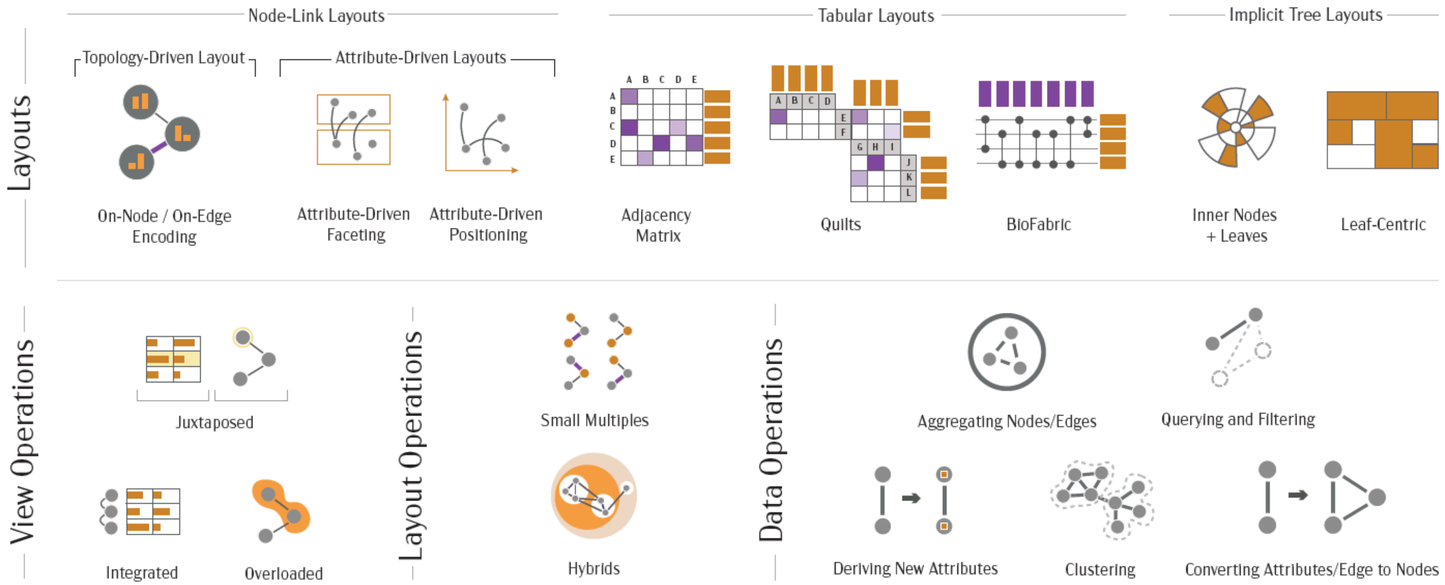
Evaluation Methodology

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

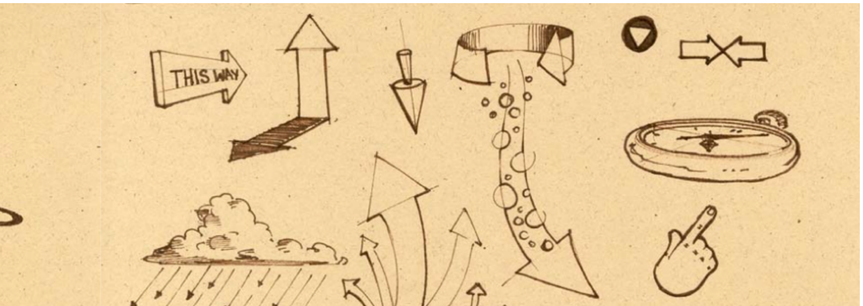
Lecture
Reading
Discussion

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; *RECORDED*

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

Will happen in breakout groups, *NOT RECORDED*

Submit 5 critiques, within 2 days after posting

Labs: short coding tutorials, examples; *RECORDED*

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 12:25-1:45 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 2020 | Print | Week | Month | Agenda

Mon	Tue	Wed	Thu	Fri	Sat	Sun
31 Week 2 17:00 Vis for Data Sc	Sep 1 12:25 Vis for Data Sc 14:00 Alex Lex Office	2 10:00 Youjia's Office	3 12:25 Vis for Data Sc 14:00 Haihan's Office	4 Drop Deadline HW1 Due 10:00 Devin's Office	5	6
7 Week 3 17:00 Vis for Data Sc	8 12:25 Vis for Data Sc 14:00 Alex Lex Office	9 10:00 Youjia's Office	10 12:25 Vis for Data Sc 14:00 Haihan's Office	11 HW 2 Due 10:00 Devin's Office	12	13
14 Week 4 17:00 Vis for Data Sc	15 12:25 Vis for Data Sc 14:00 Alex Lex Office	16 10:00 Youjia's Office	17 12:25 Vis for Data Sc 14:00 Haihan's Office	18 HW 3 Due 10:00 Devin's Office	19	20
21 Week 5 17:00 Vis for Data Sc	22 12:25 Vis for Data Sc 14:00 Alex Lex Office	23 10:00 Youjia's Office	24 12:25 Vis for Data Sc 14:00 Haihan's Office	25 HW 4 Due 10:00 Devin's Office	26	27
28 Week 6 17:00 Vis for Data Sc	29 12:25 Vis for Data Sc 14:00 Alex Lex Office	30 10:00 Youjia's Office	Oct 1 12:25 Vis for Data Sc 14:00 Haihan's Office	2 10:00 Devin's Office	3	4

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.

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.
- The Value of Visualization.Jarke van Wijk. Proceedings of the IEEE Visualization Conference, pp. 79-86, 2005.

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.

The course begins by bootstrapping your web development skills, moves on to fundamentals of perception, introduces data types you will encounter, and then focuses on visualization techniques and methods for a broad range of data types. An integral component of the course are regular design critiques and redesigns that will hone your skills in understanding, critiquing and developing visualization techniques.

The course is offered in the fall term 2020 at the University of Utah in three variants: CS-5630 and DS-4630 for undergraduate computer science and data science students; and CS-6630 for graduate students, with a special section of CS-6630 (002) designated for the data science certificate students. Classes start on Tuesday, August 25.

Due to COVID-19, this class is taught in the "IVC" (Interactive Video Conferencing) form, which means we will use live-video conferencing via Zoom. We will also use Slack for asynchronous communication, and Canvas for submission of assignments. All classes and sections are also archived online.

Communicate

Slack

https://datavis-2020.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/574005

Homework submissions, Grades

Office Hours

Alex: Tuesdays after Class

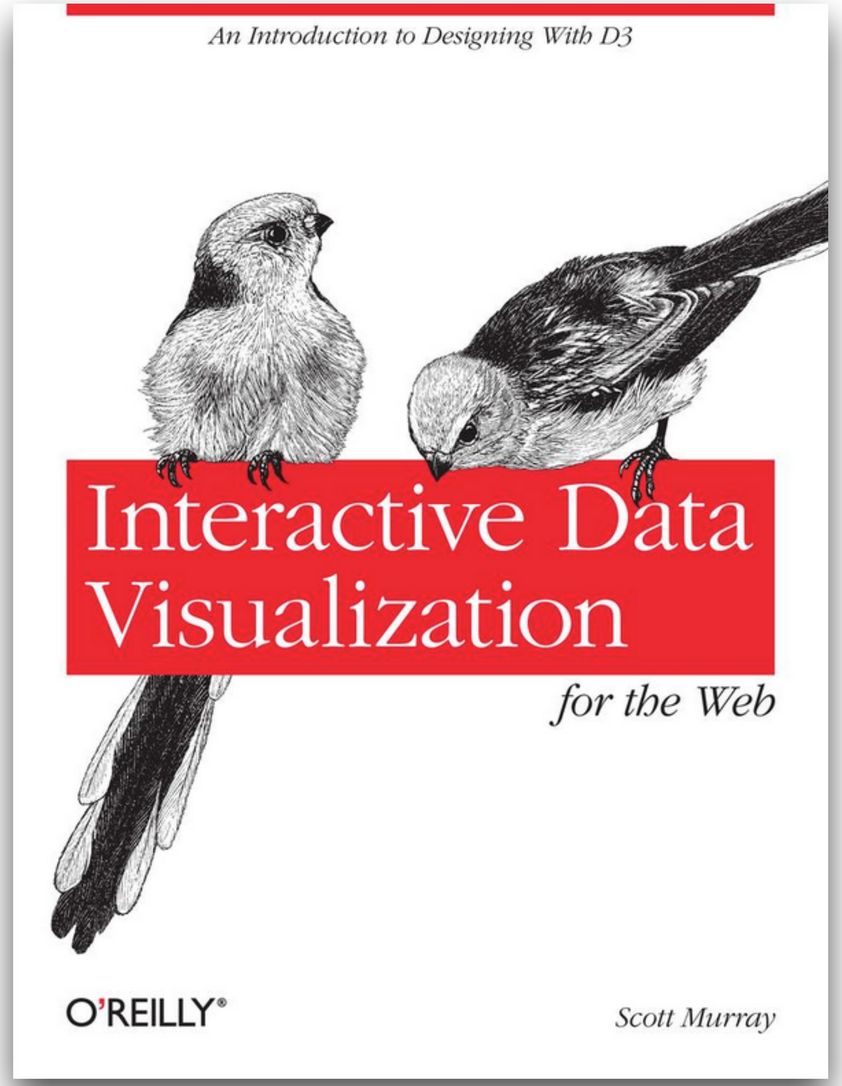
TAs: starting next week

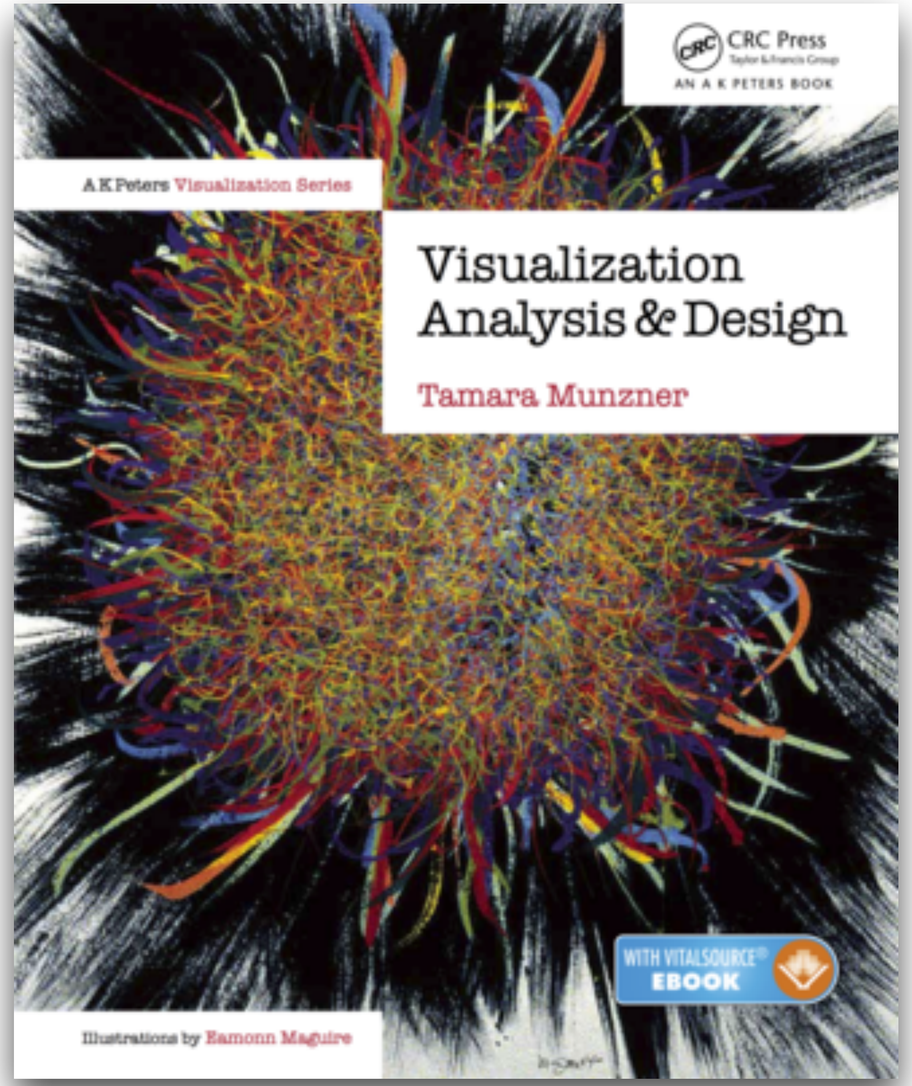
E-Mail

alex@sci.utah.edu

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

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?

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:
http://www.cs.utah.edu/wp-content/uploads/2014/12/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

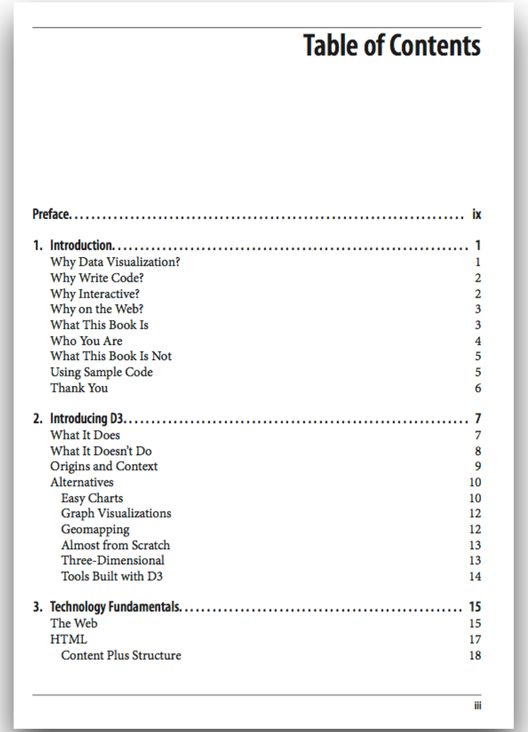


Table of Contents

Next Week

HW1 due

Introduction to Git, HTML, CSS

Office hours start!

https://github.com/dataviscourse/2020_dataviscourse_homework