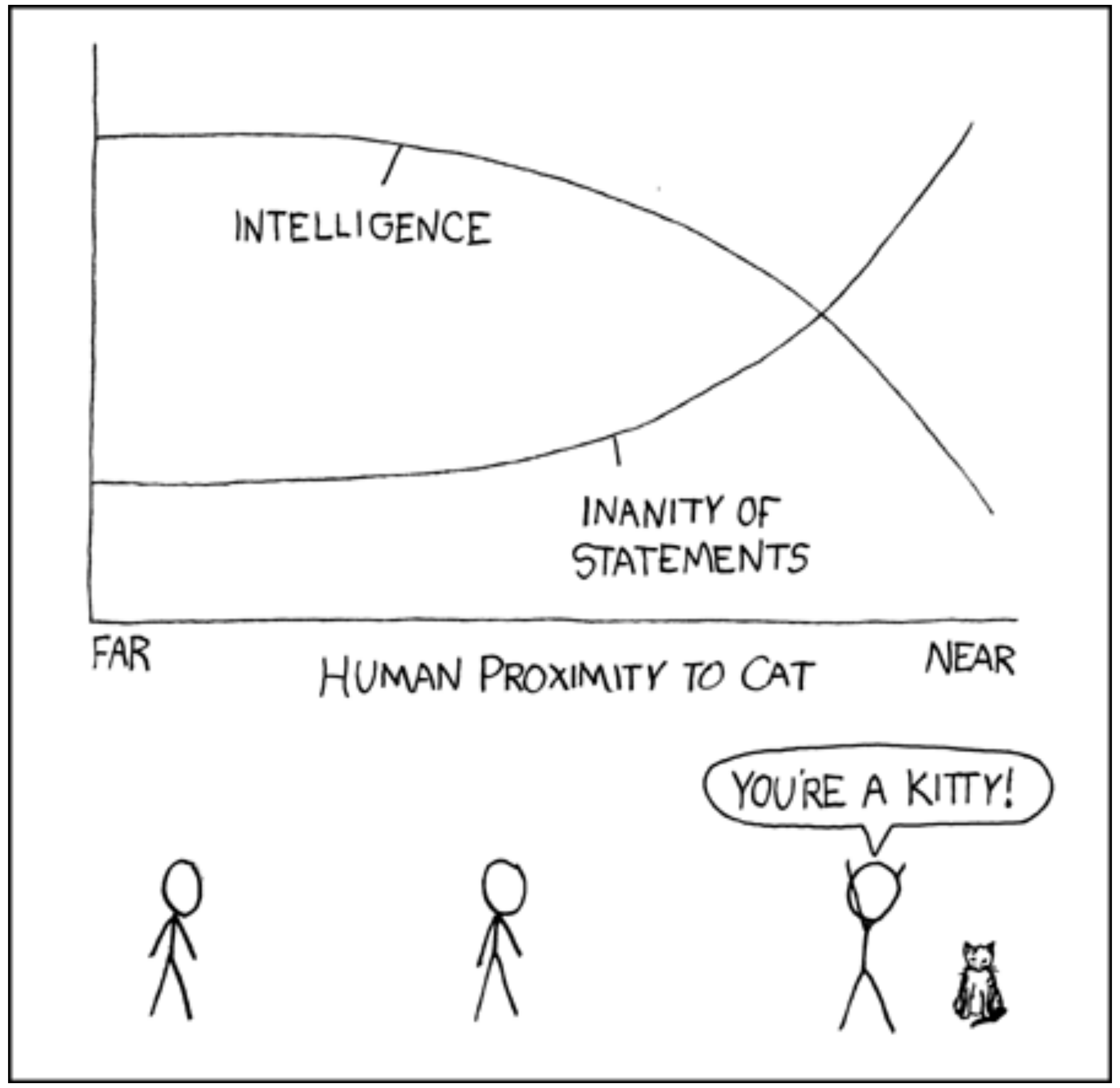
CS-5630 / CS-6630 Visualization for Data Science

Alexander Lex
alex@sci.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

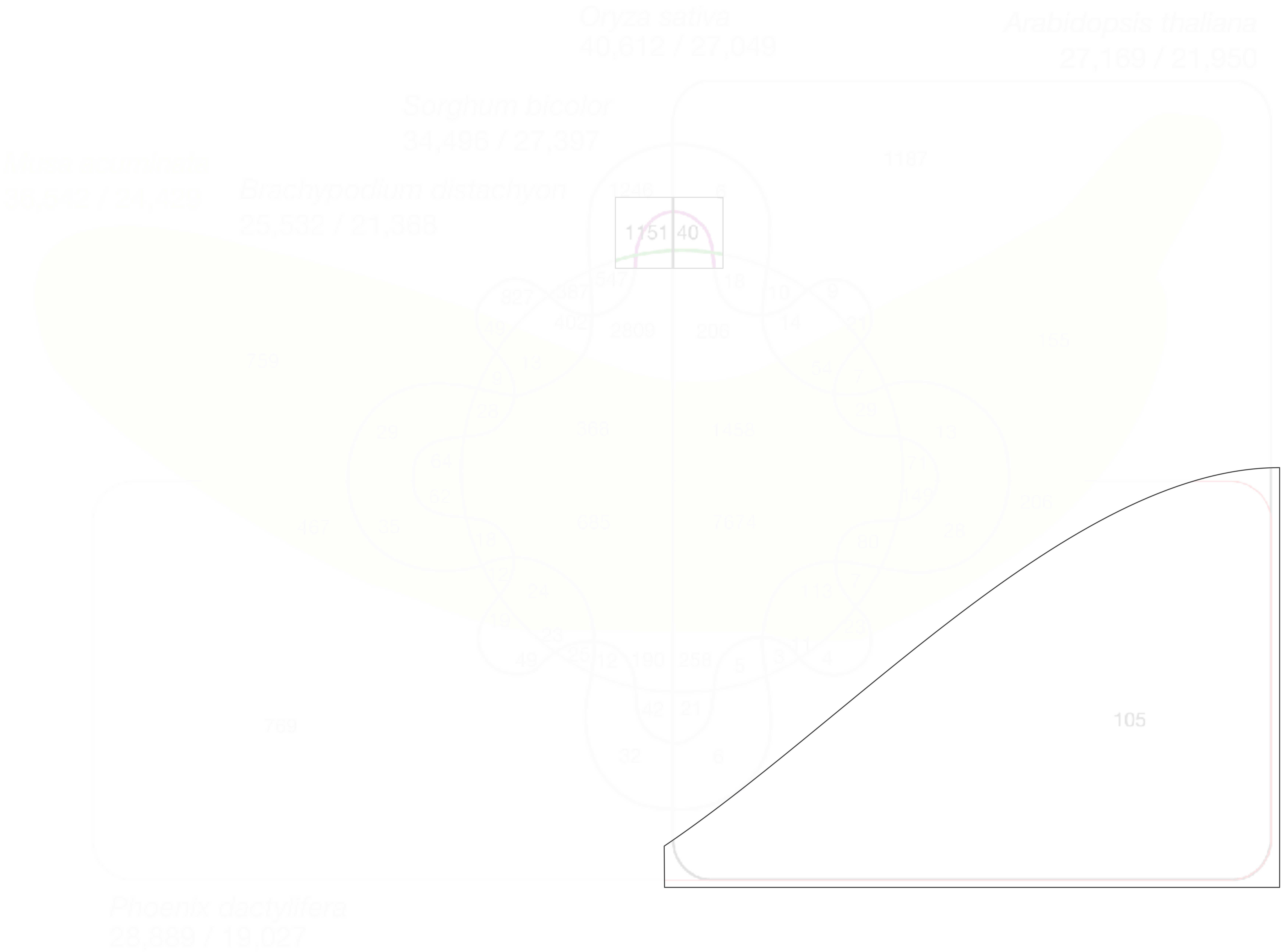
visualization *pictures*

The purpose of computing is insight, not numbers.

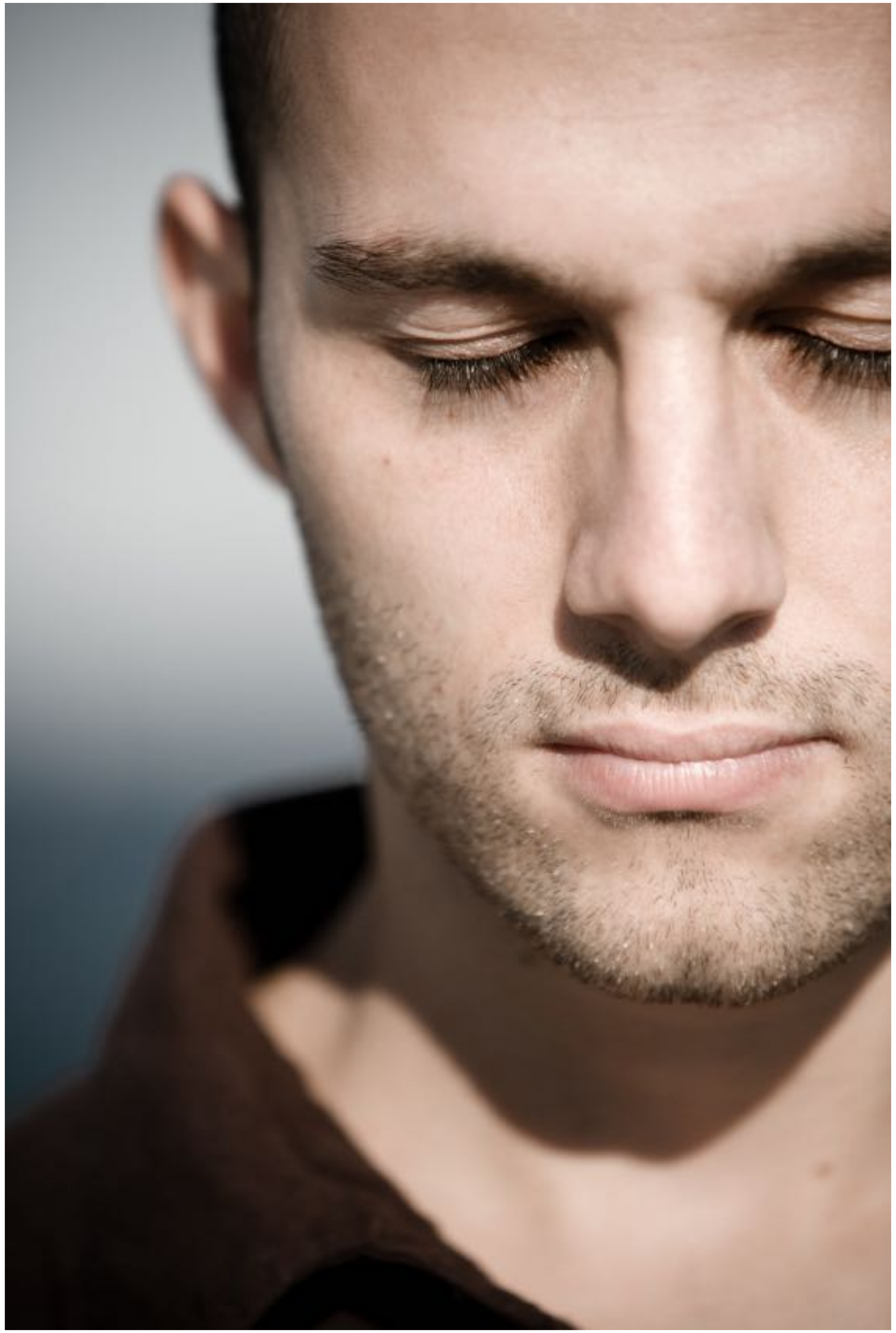
- Richard Wesley Hamming

- 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 **transforms** (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?

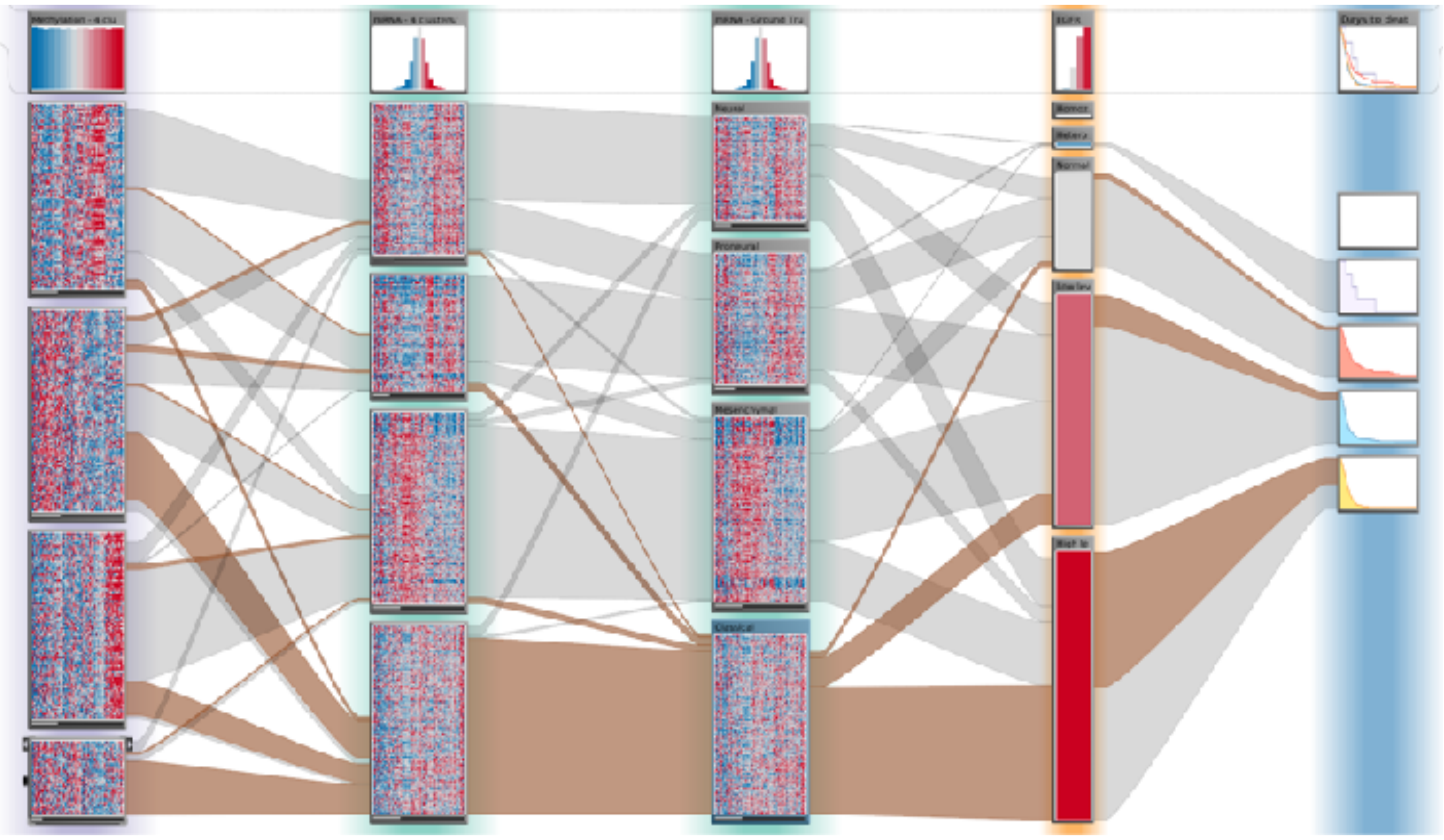
When questions are not well defined: Exploration

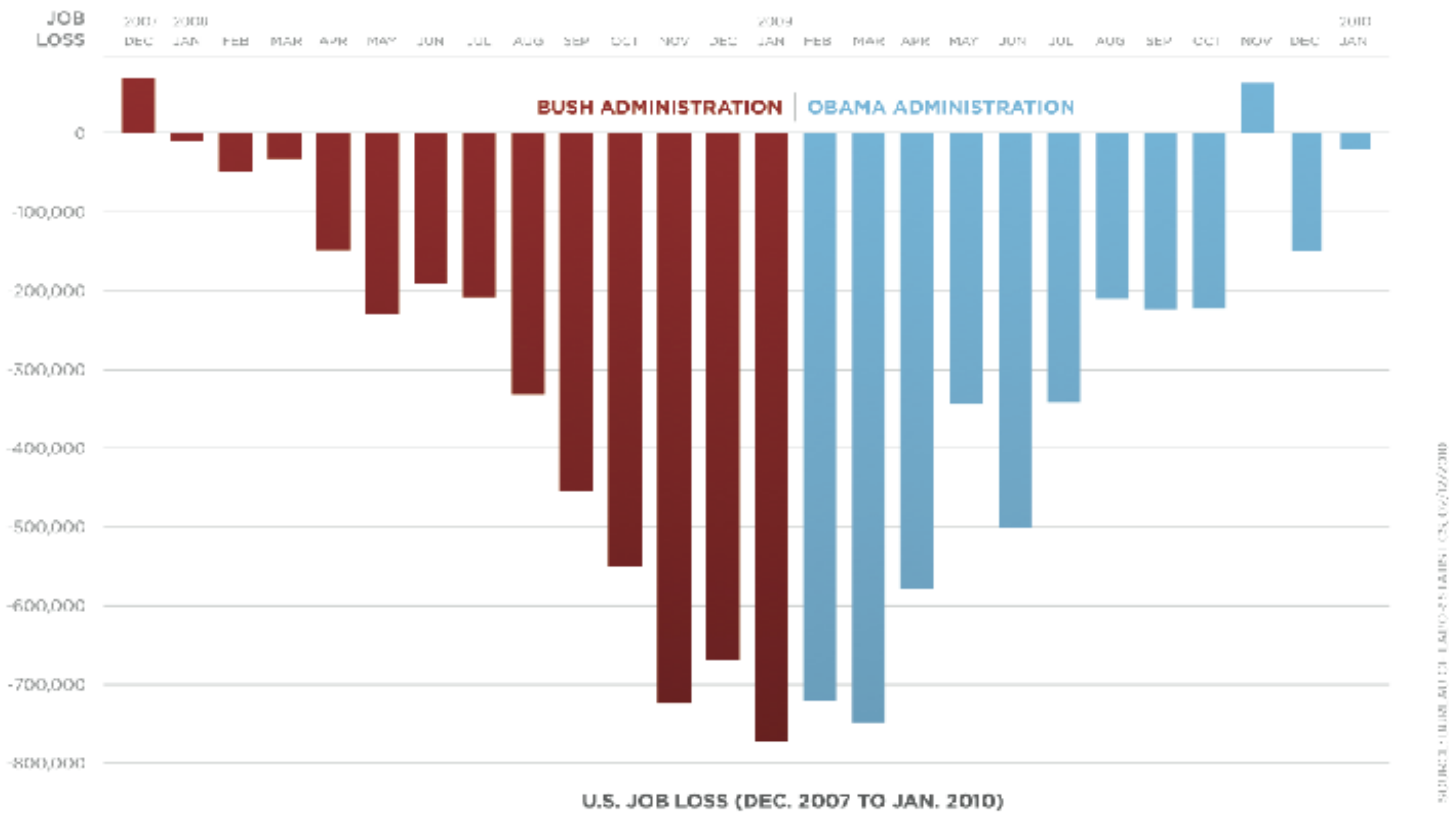
What is the structure of a terrorist network?

Which drug can help patient X?

Purpose of Visualization

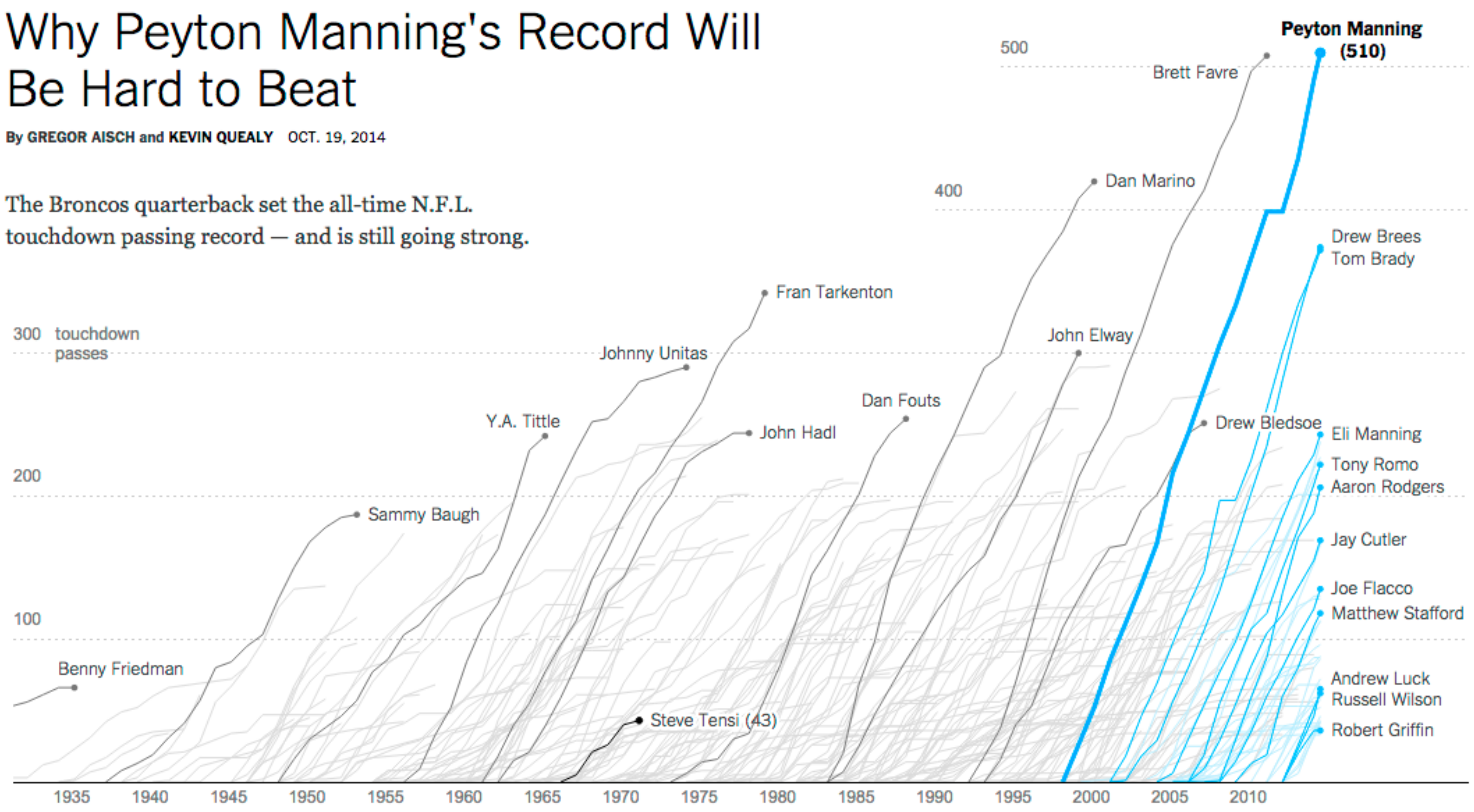
[Obama Administration]





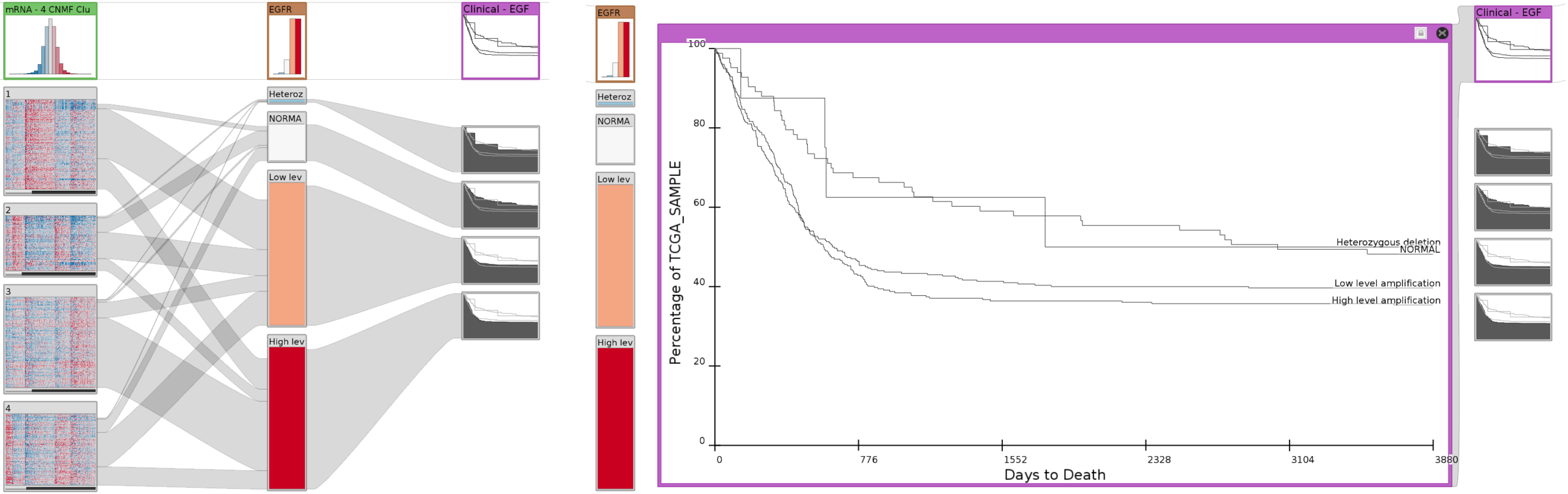
Open Exploration
Confirmation
Communication

Example Communication



[New York Times]

Example Exploration: Cancer Subtypes



[Caleydo StratomeX]

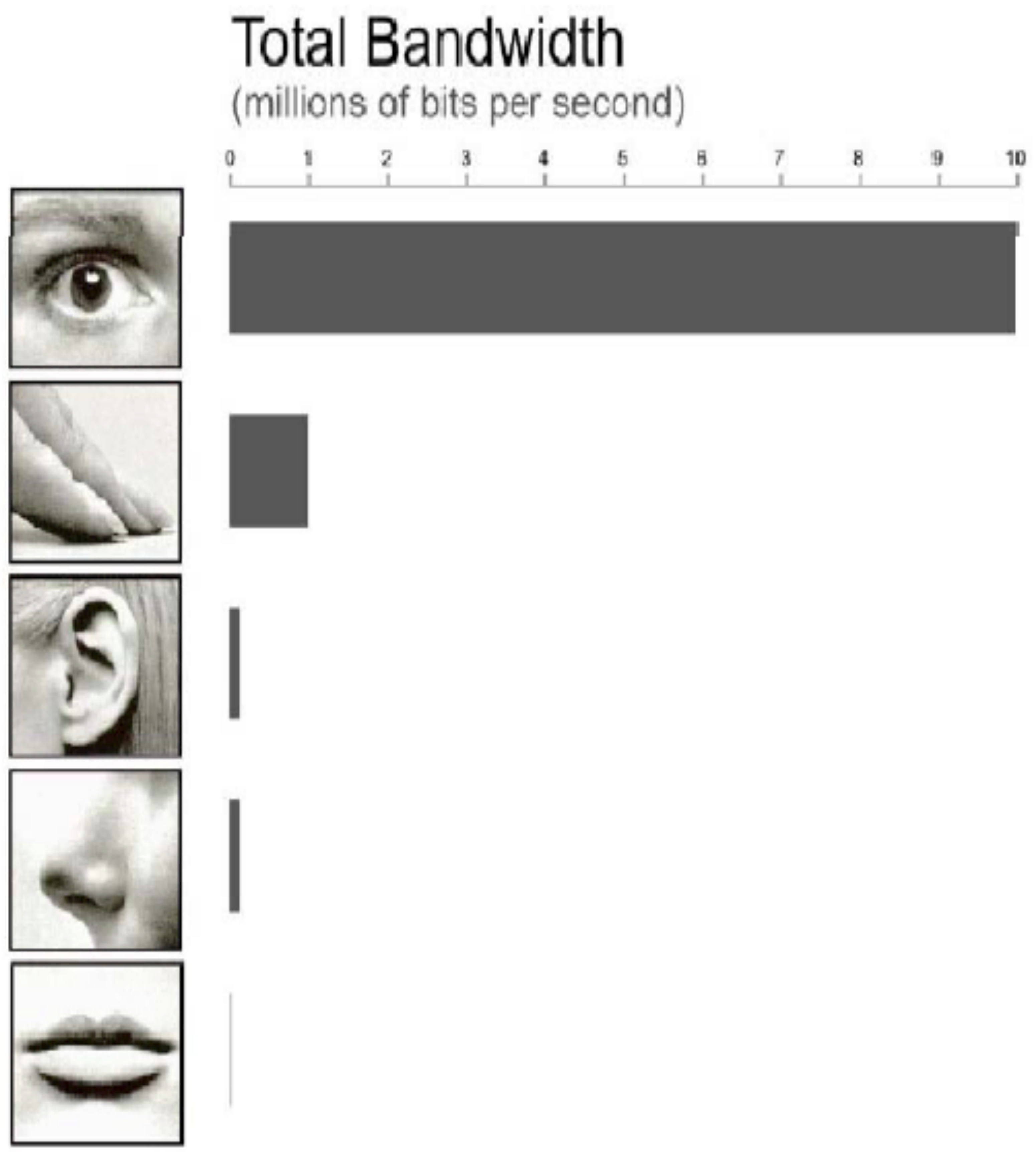
Why Graphics?

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

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?

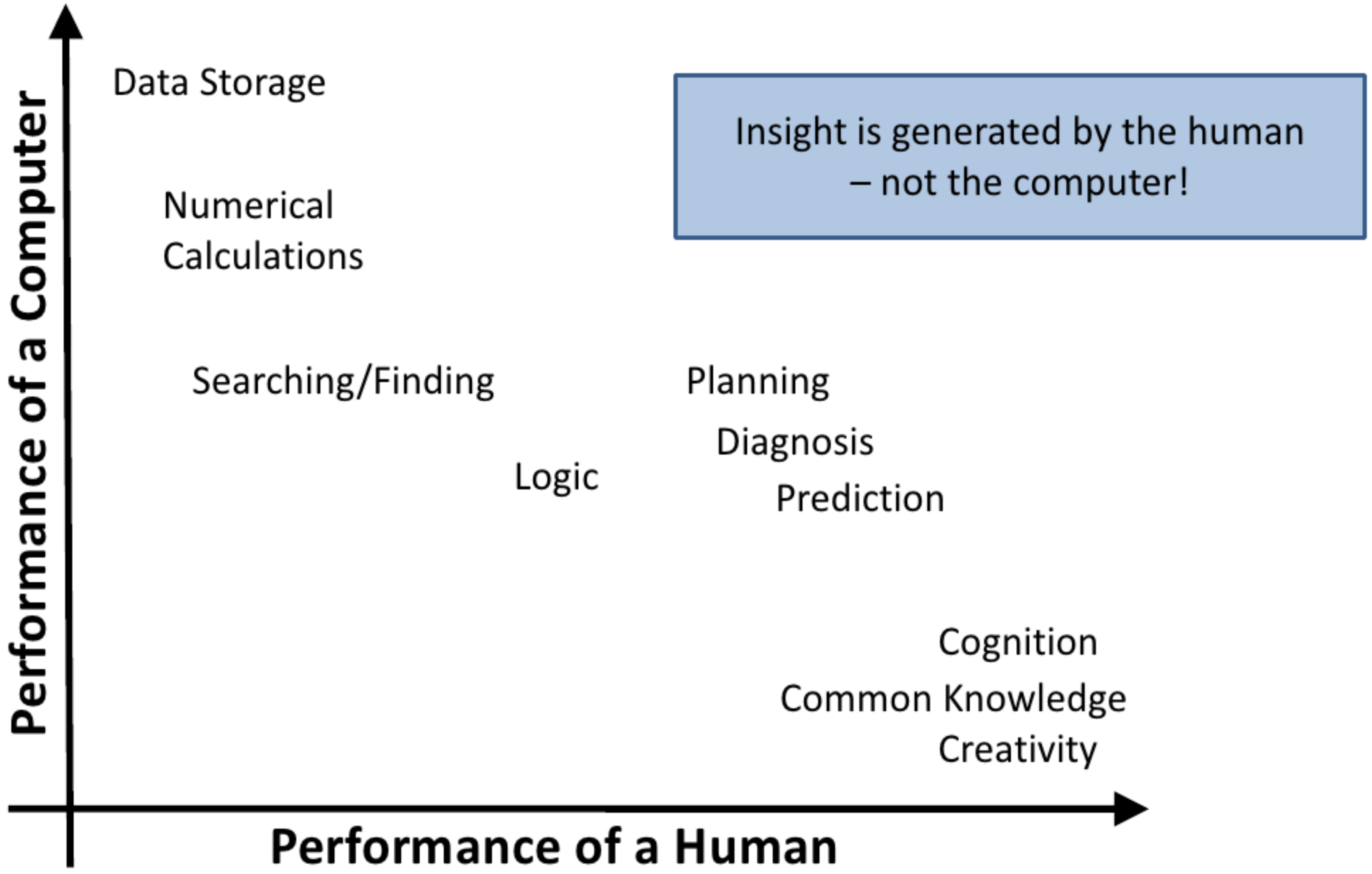
What is the current unemployment rate?

Decisions needed in minimal time

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

Manufacturing: is bottle broken?

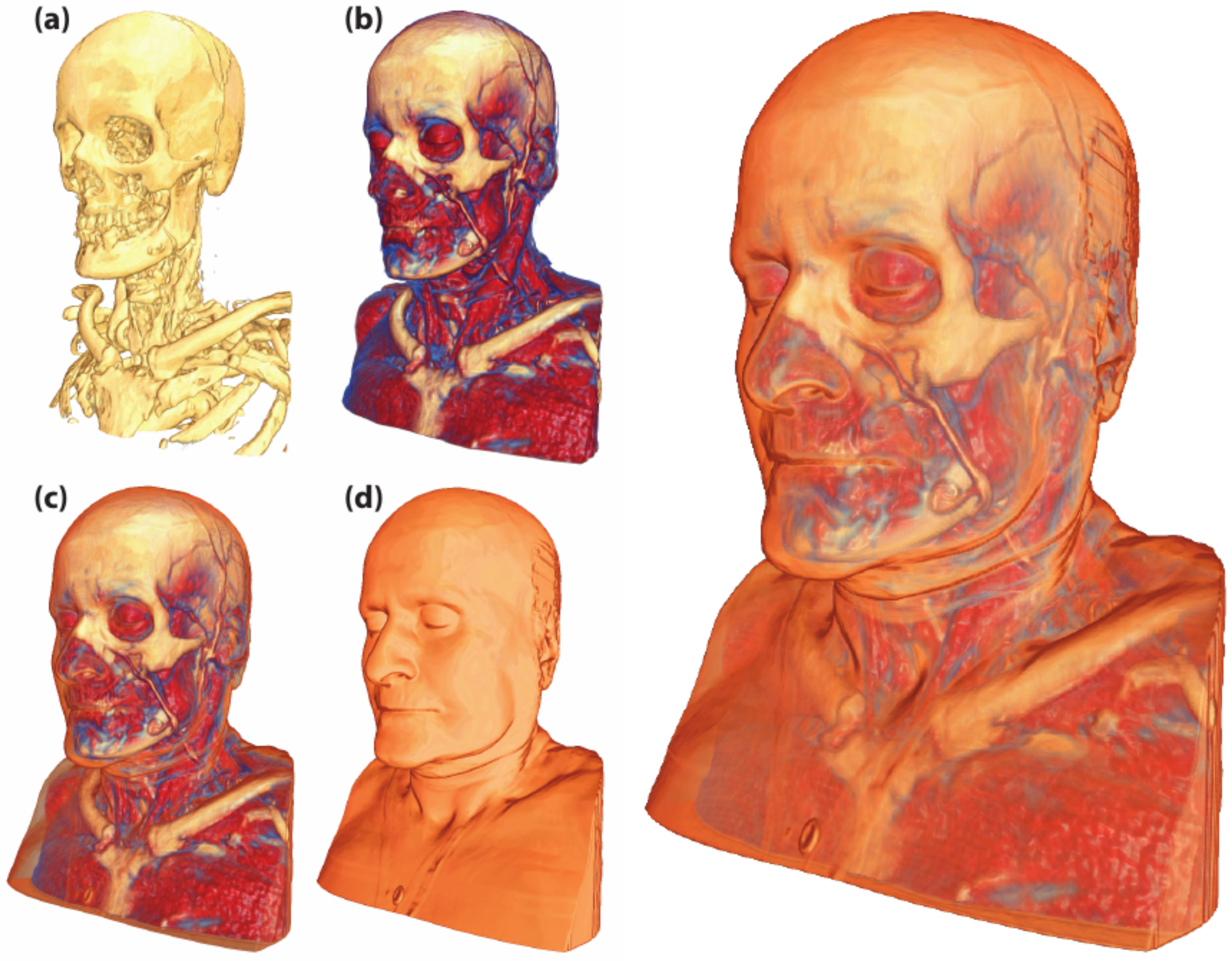
The Ability Matrix



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

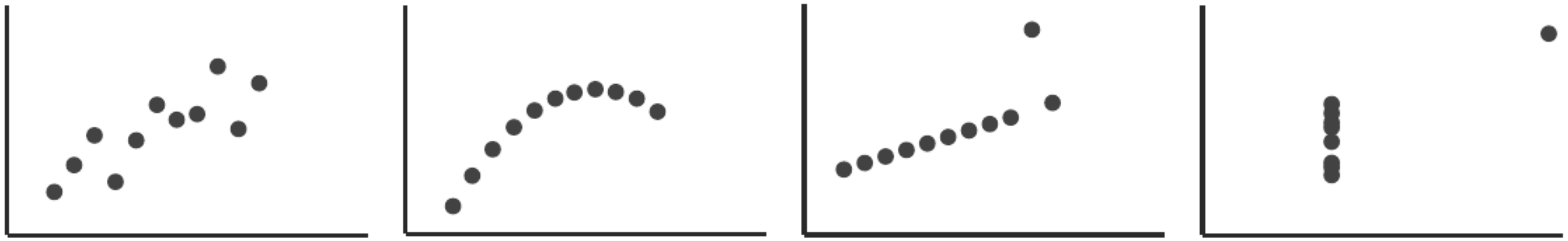
[New York Times]

Why not just use Statistics?

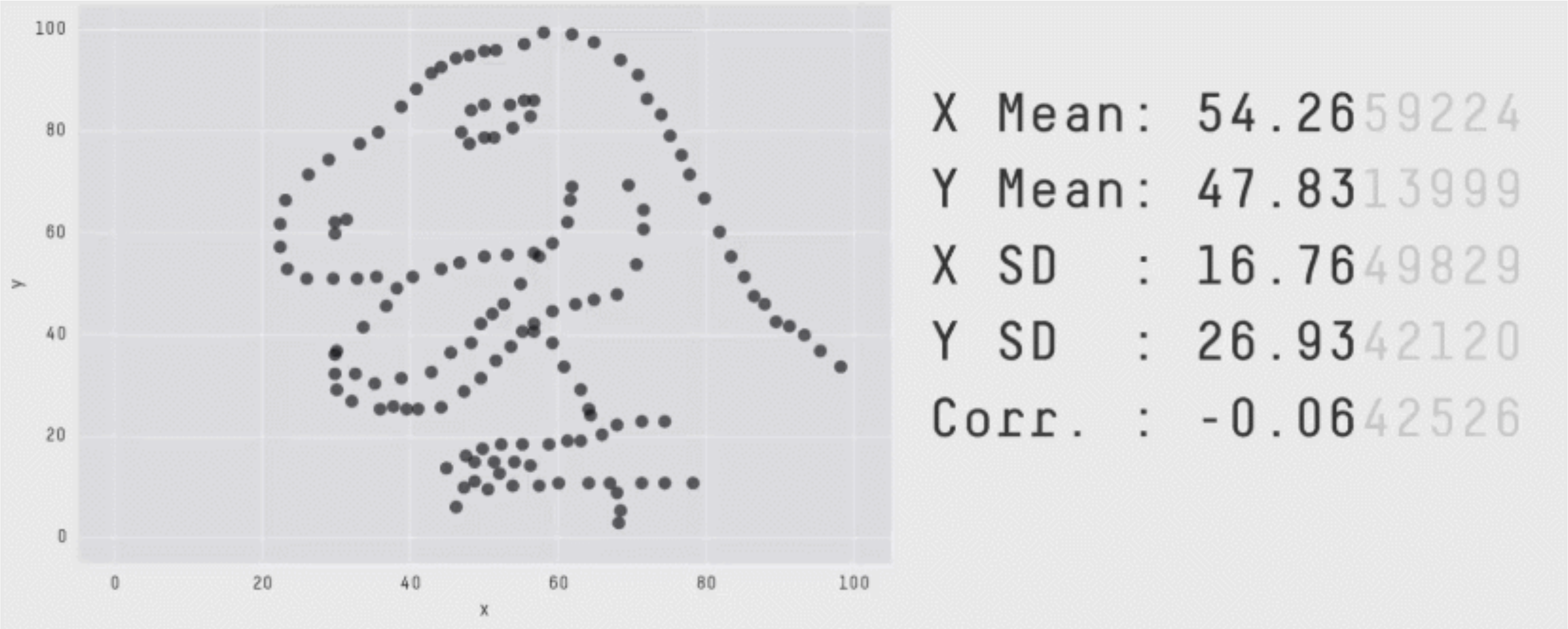
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.4	8	7.9
5	5.						6.8

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

Visualization in the Data Science Process

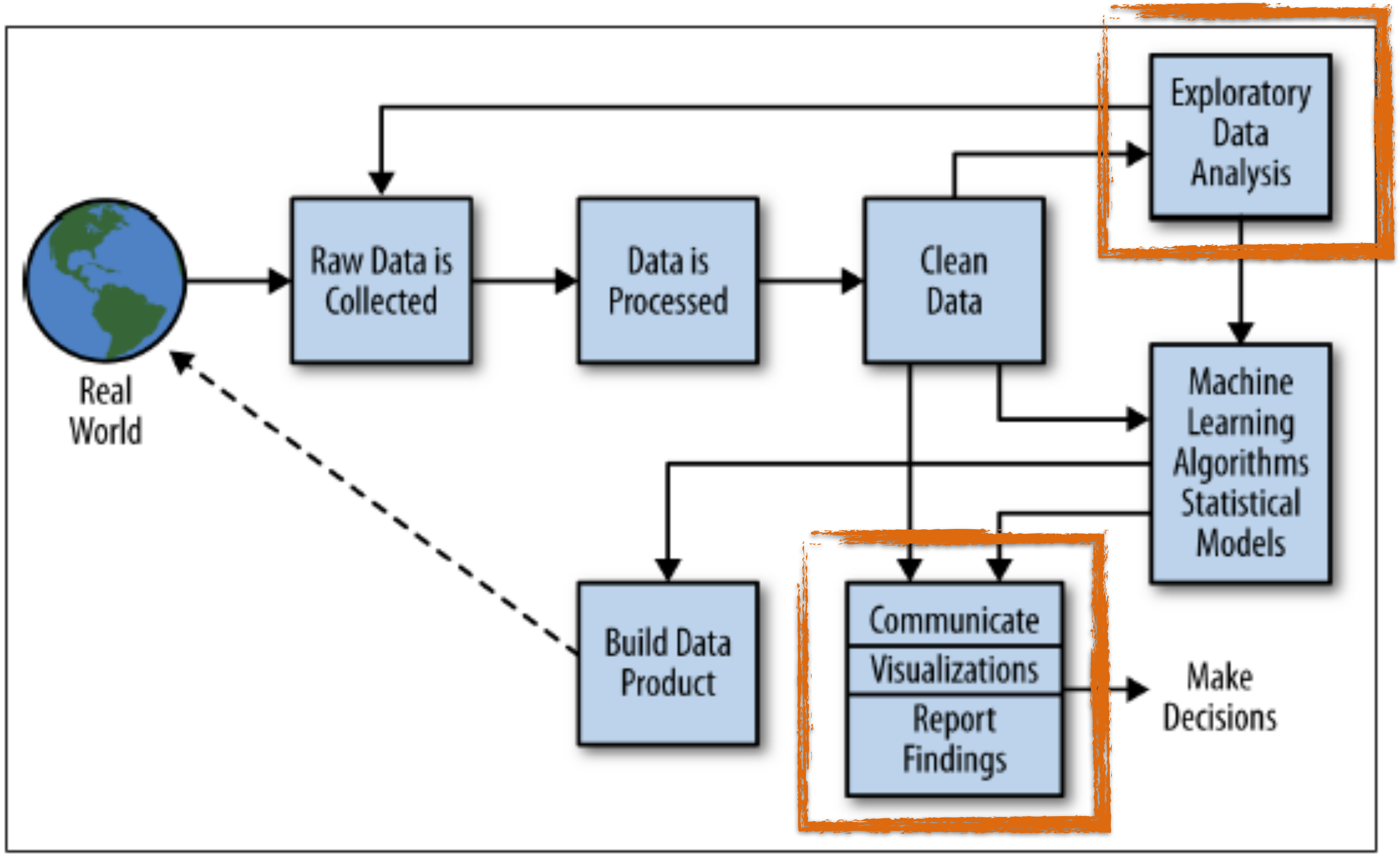


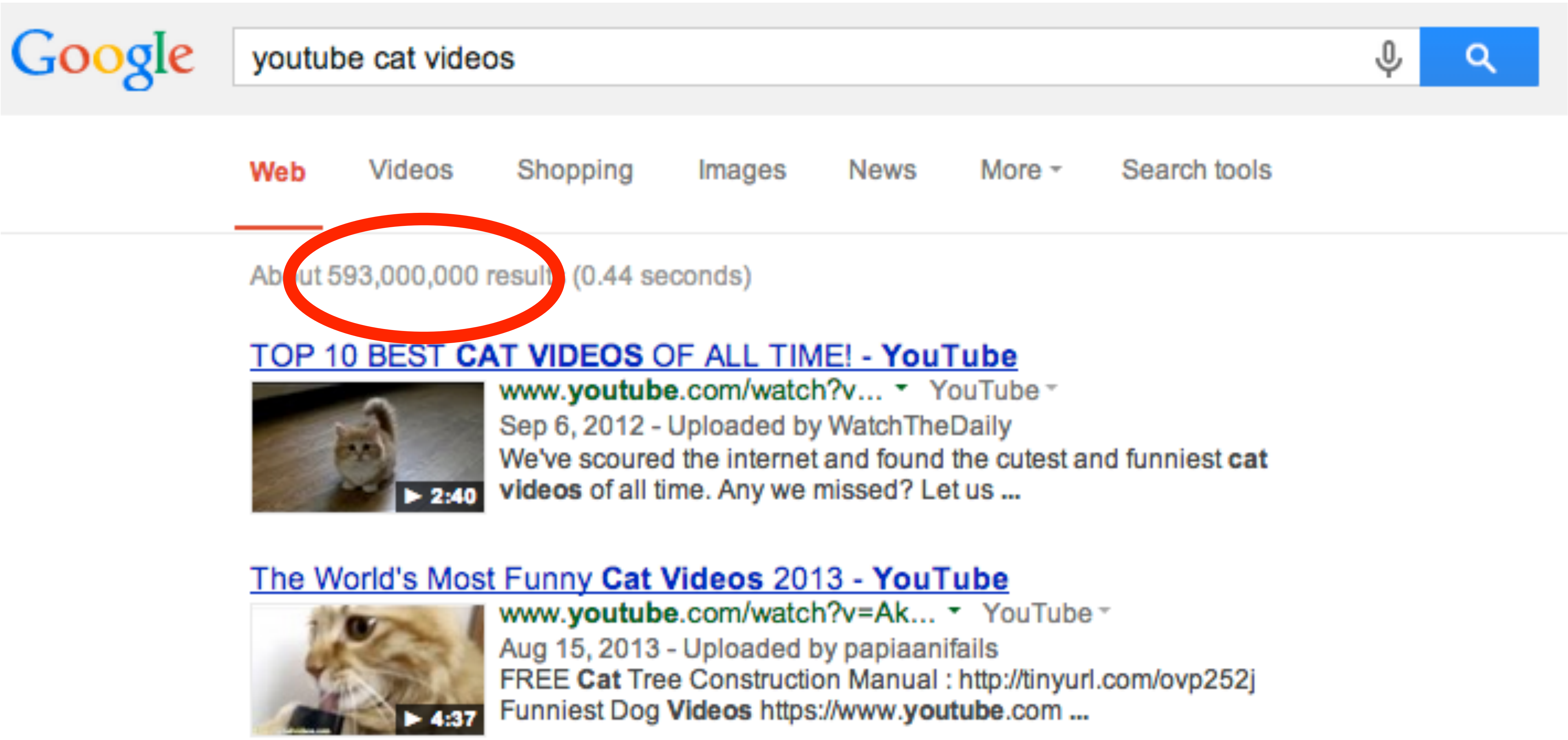
Figure 2-2. The data science process

Big Data

2010: 1,200 exabytes, largely unstructured

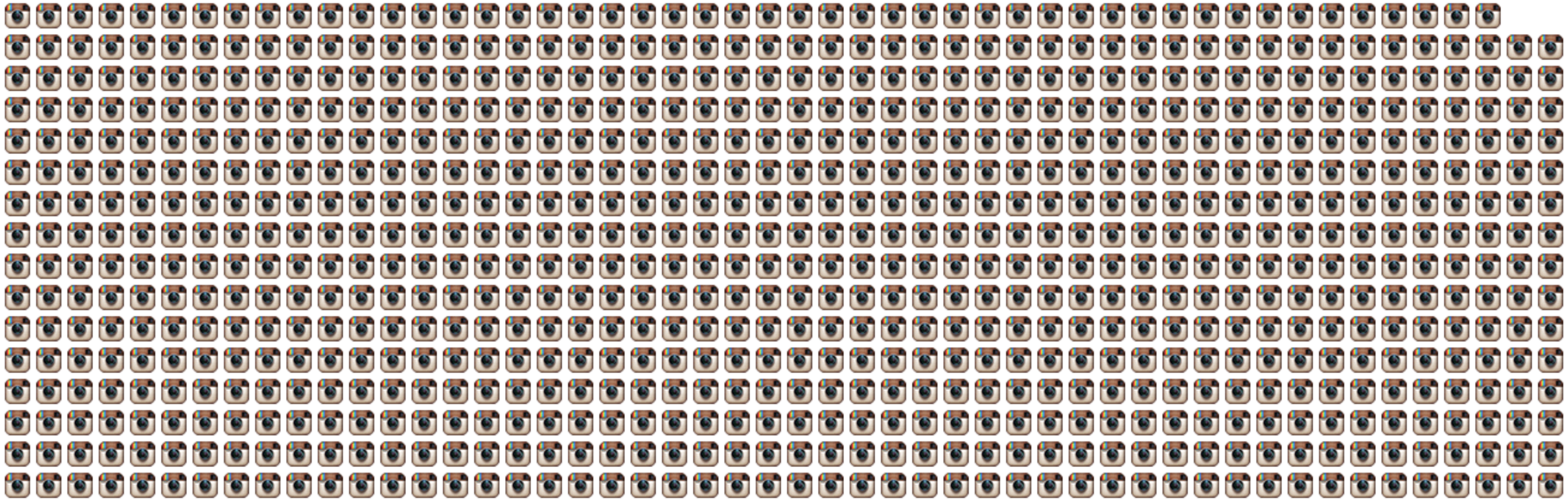
Google stores ~10 exabytes (2013)

Hard disk industry ships ~8 exabytes/year



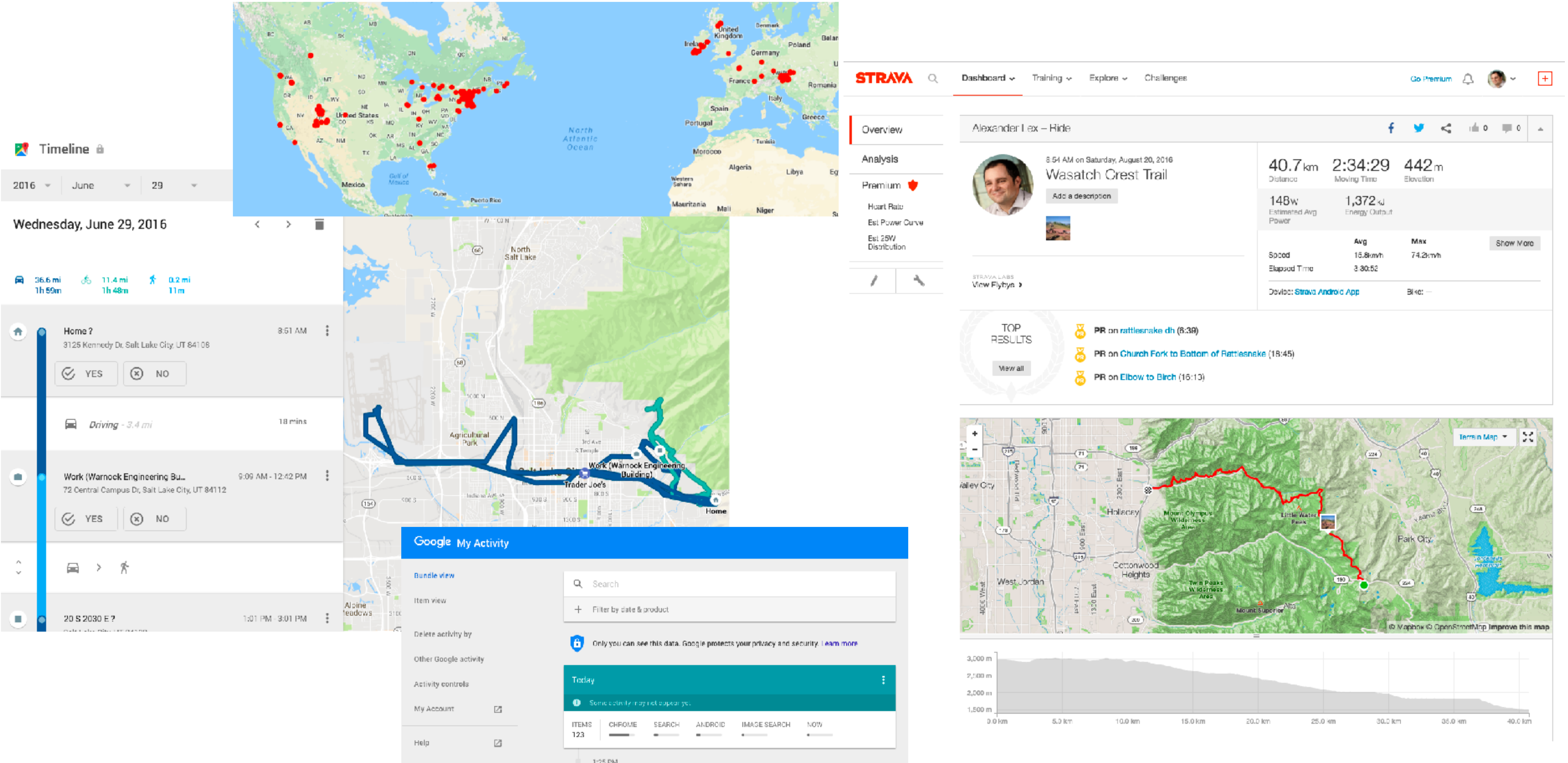
15 Exabytes in Punch Cards:
4.5 km over New England

798 Instagram photos uploaded in 1 second



1,277 Tumblr posts in 1 second

Example: Personal Data



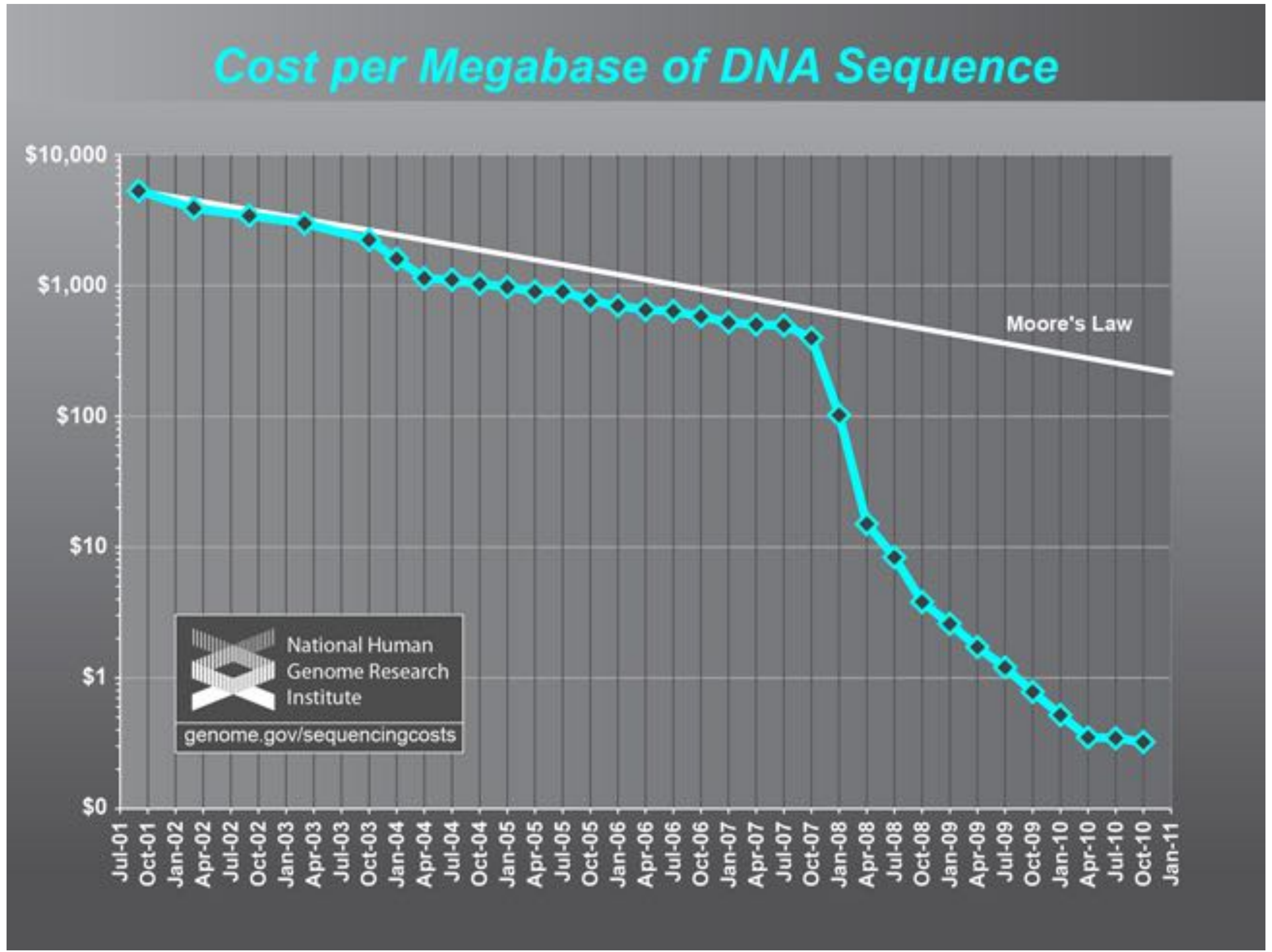
Big Data in Science and Engineering

"Big Data" hasn't just transformed industry!

It's also transformed science and engineering. Cheap sensors (e.g. imaging) have changed the way science and engineering are done.

Examples:

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



Controversy: Hypothesis or data driven methods

Example: CERN Large Hadron Collider Data

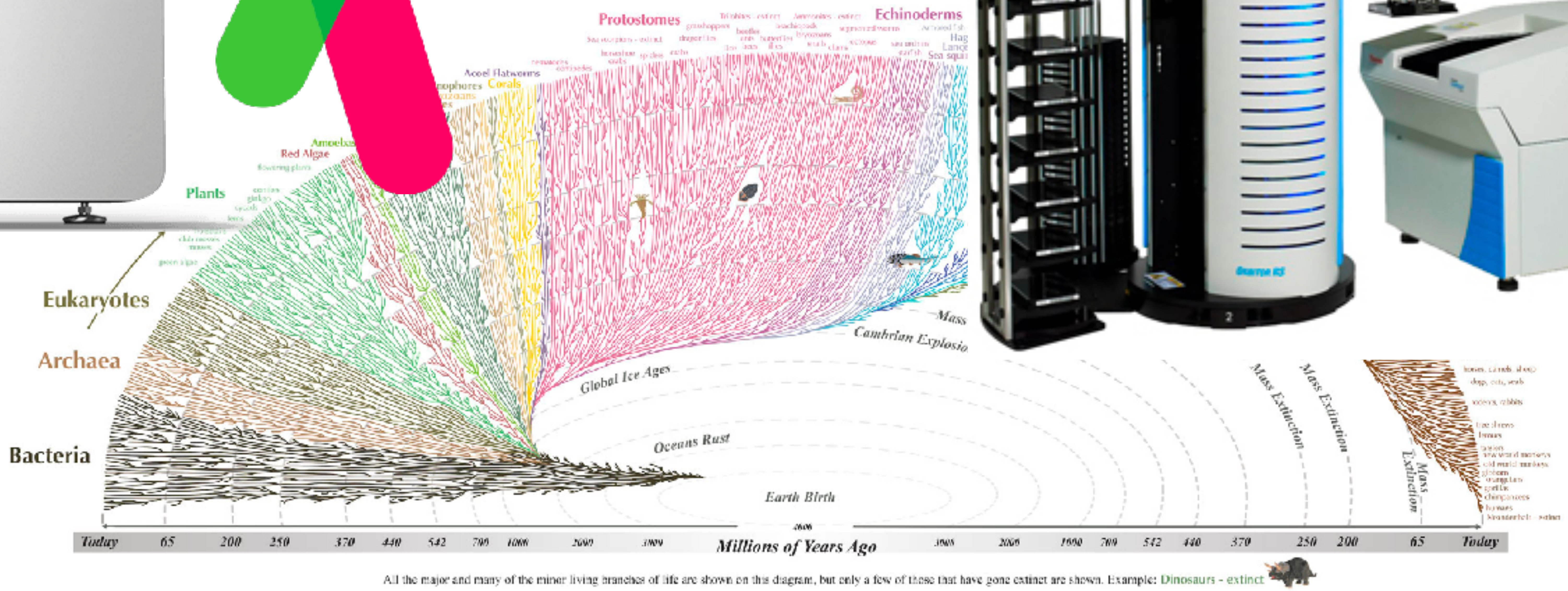
CERN has publicly released over 300TB of data: CERN Open Data Portal

How much is that?

- At 15 GB of storage a piece, you'd need **20,000 Gmail** accounts to store the whole shebang. If you wanted to send that much data at the max attachment size of 25 MB, it would take you **12 million emails**.
- A **DVD-R** holds 4.7 GB. You'd need **63,830** of them to hold 300 TB.
- Your Blu-ray collection wouldn't need to expand quite so much. **6,000** discs ought to hold it.
- 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)

Example: Genomics



Example TCGA: 1 Petabyte

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

Humans!

Human Data Interaction

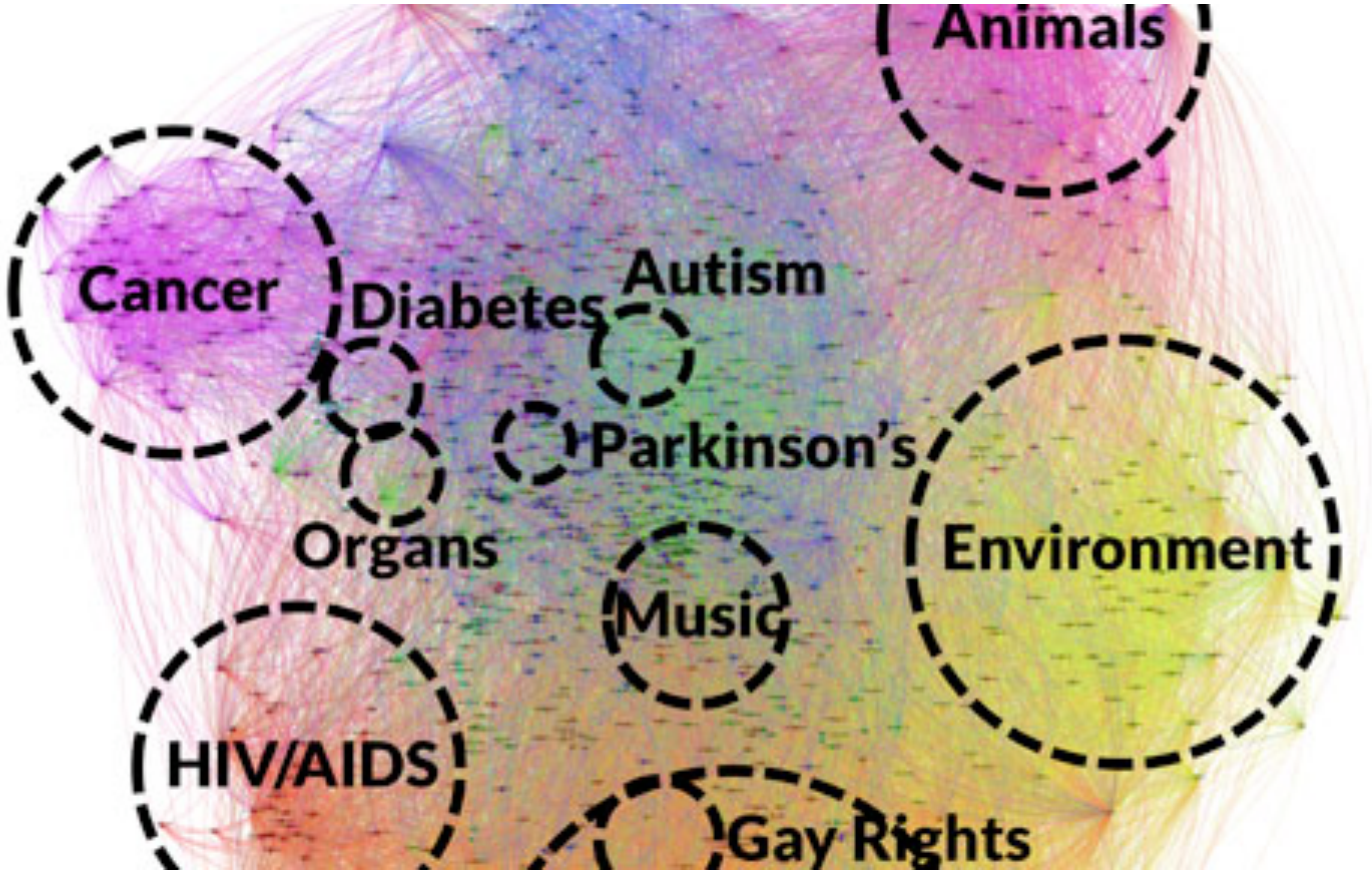
Why Humans?

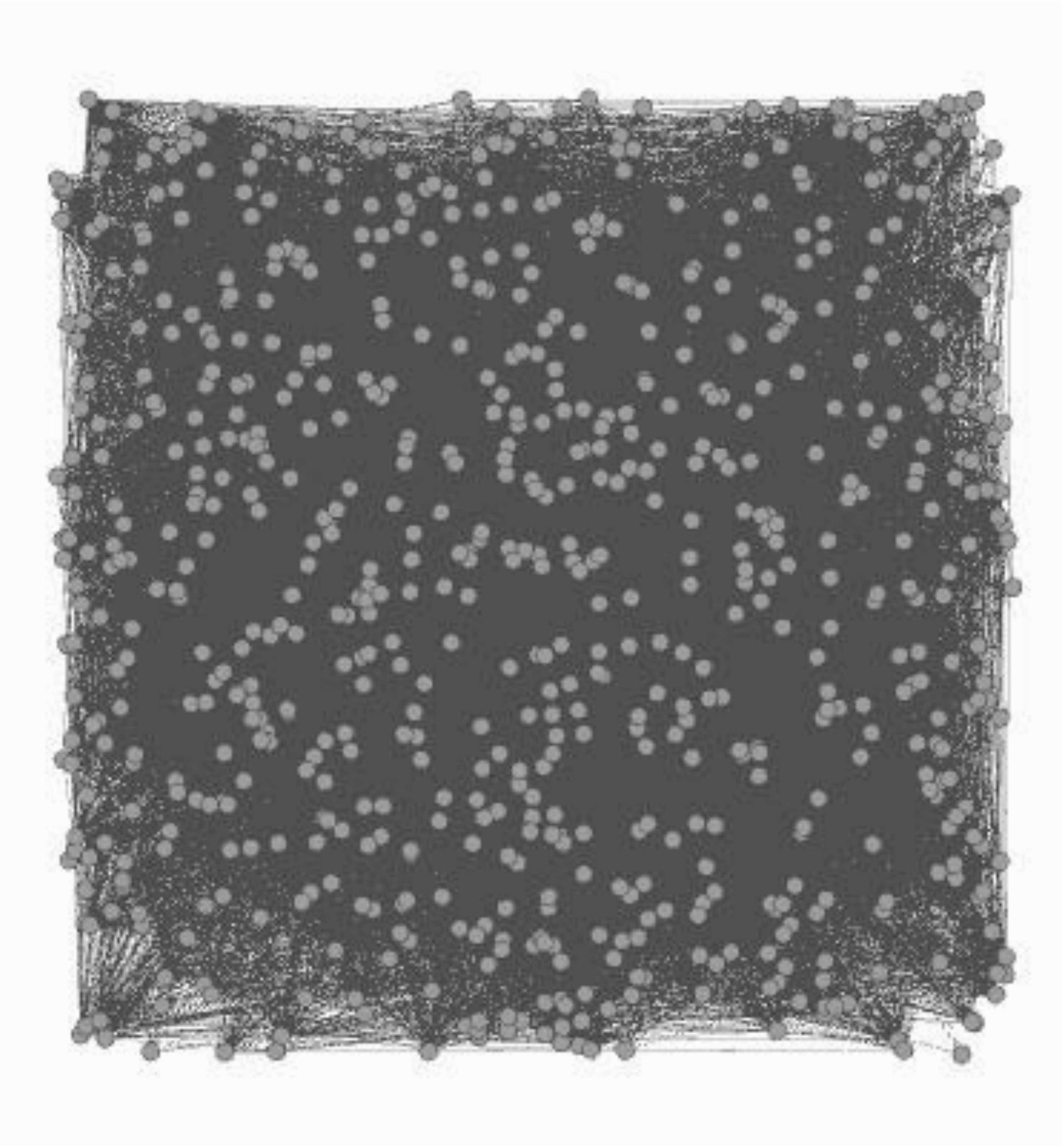
Leveraging human capabilities

- **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

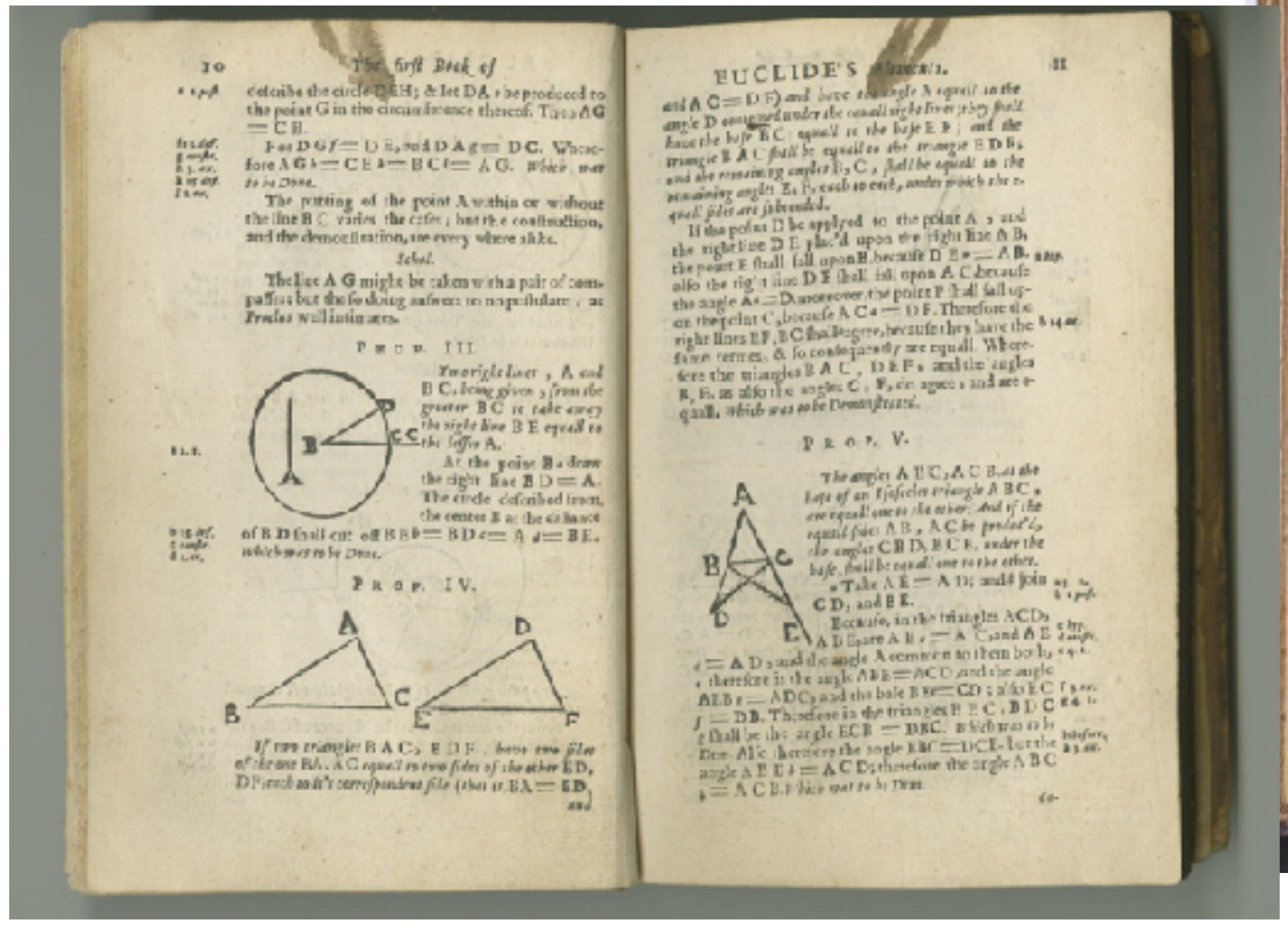
Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

How did we get here?

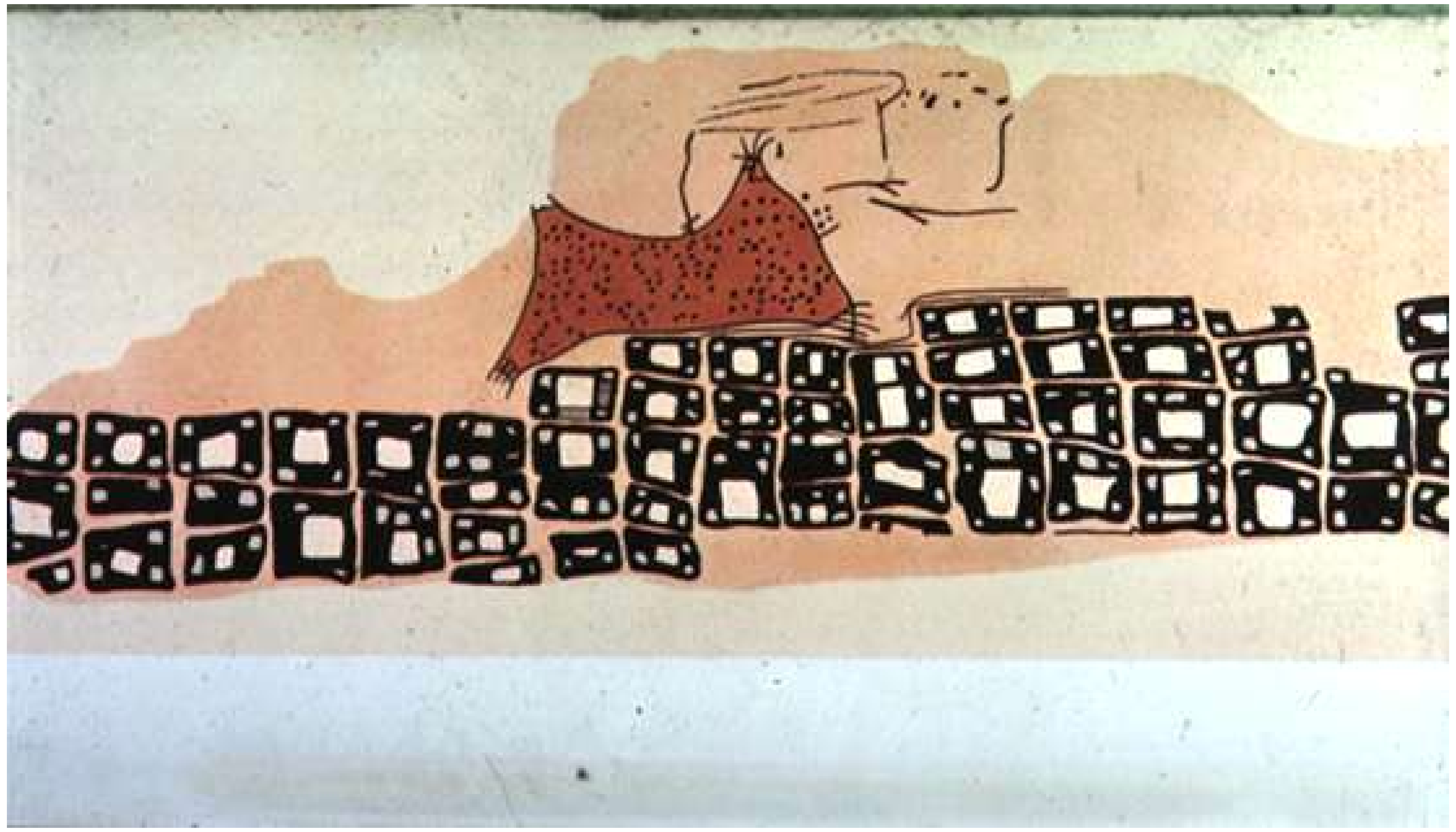
A bit of history

“It is things that make us smart”

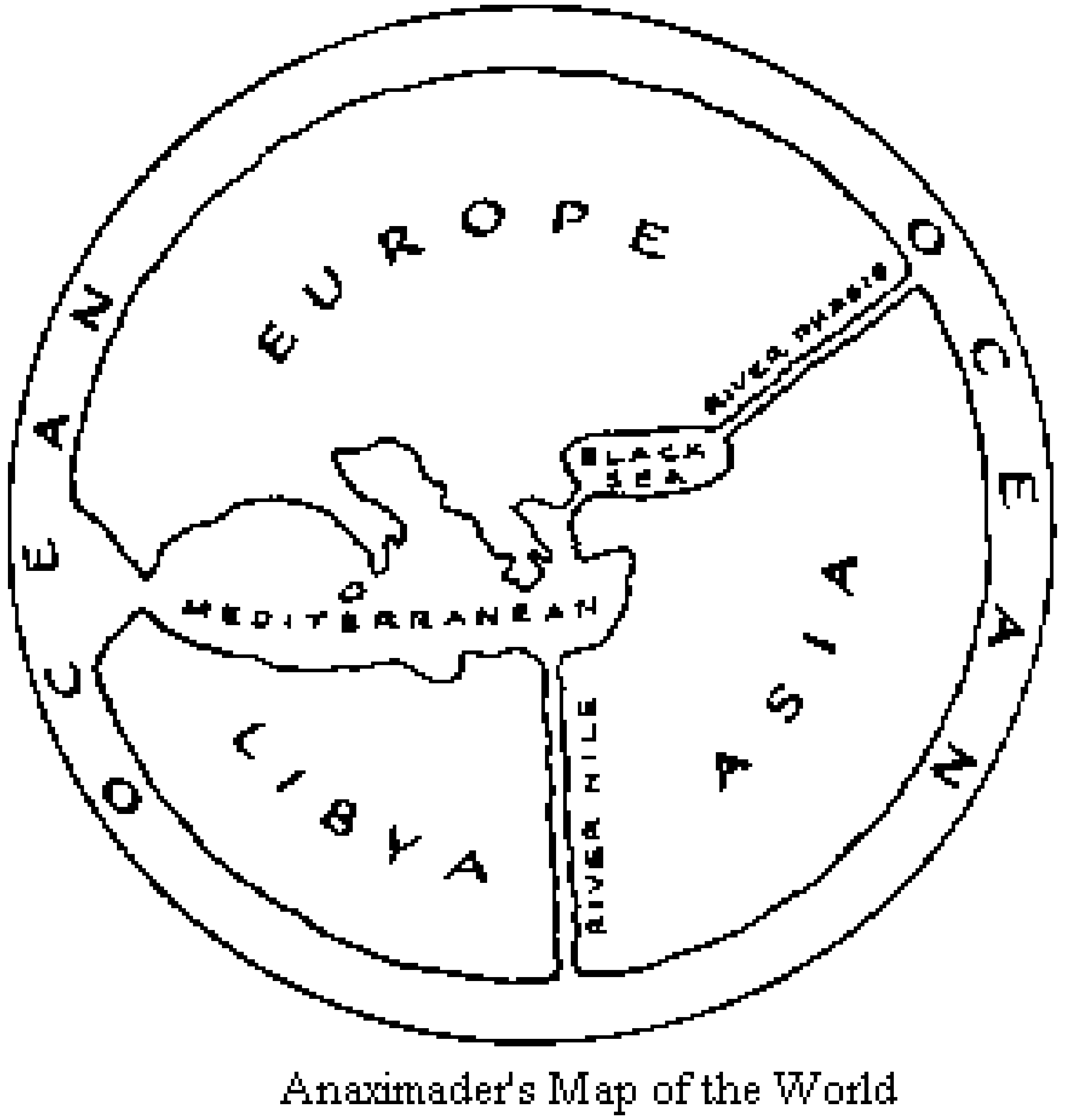
Donald A. Norman



Record



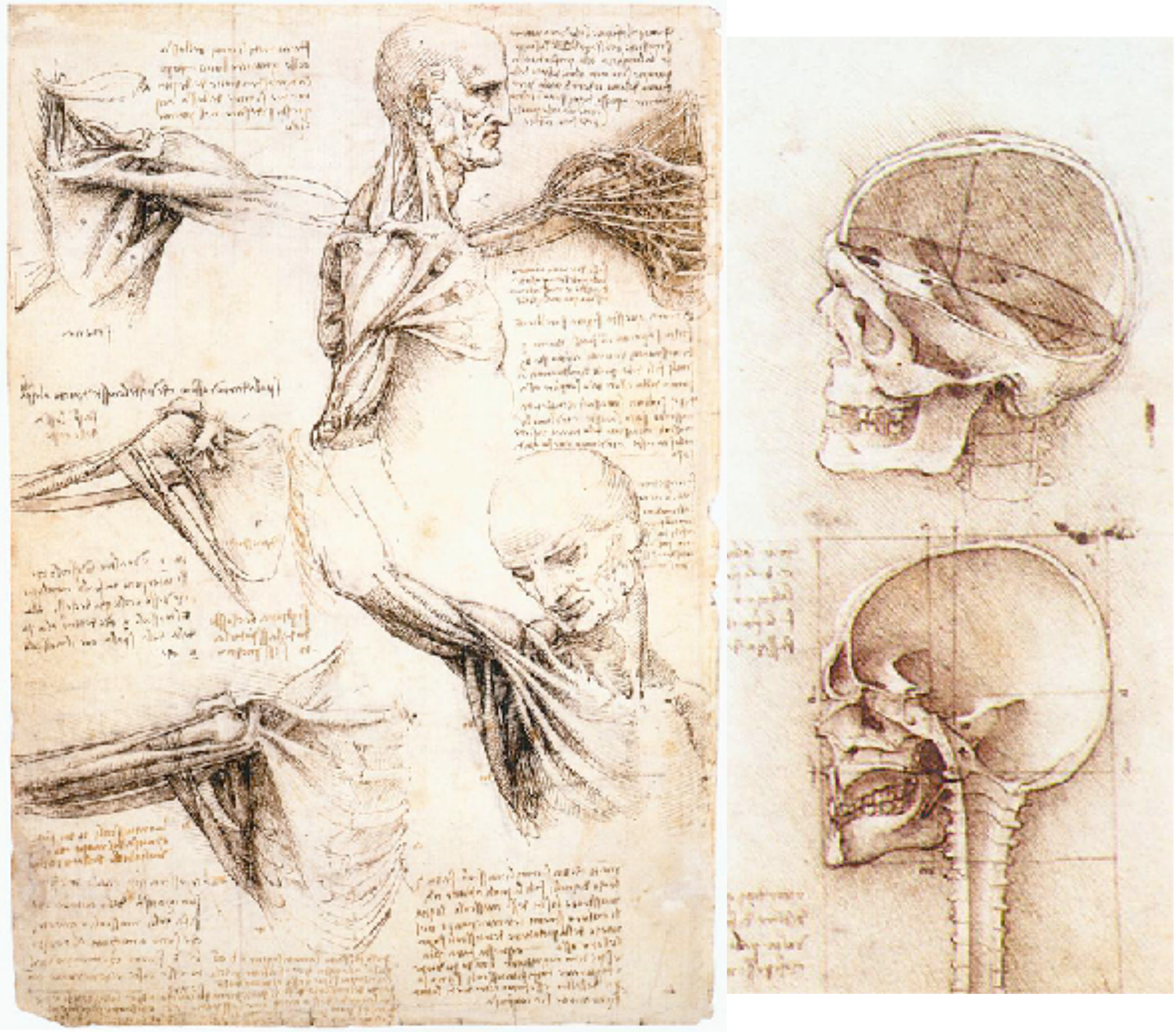
Konya town map, Turkey, c. 6200 BC



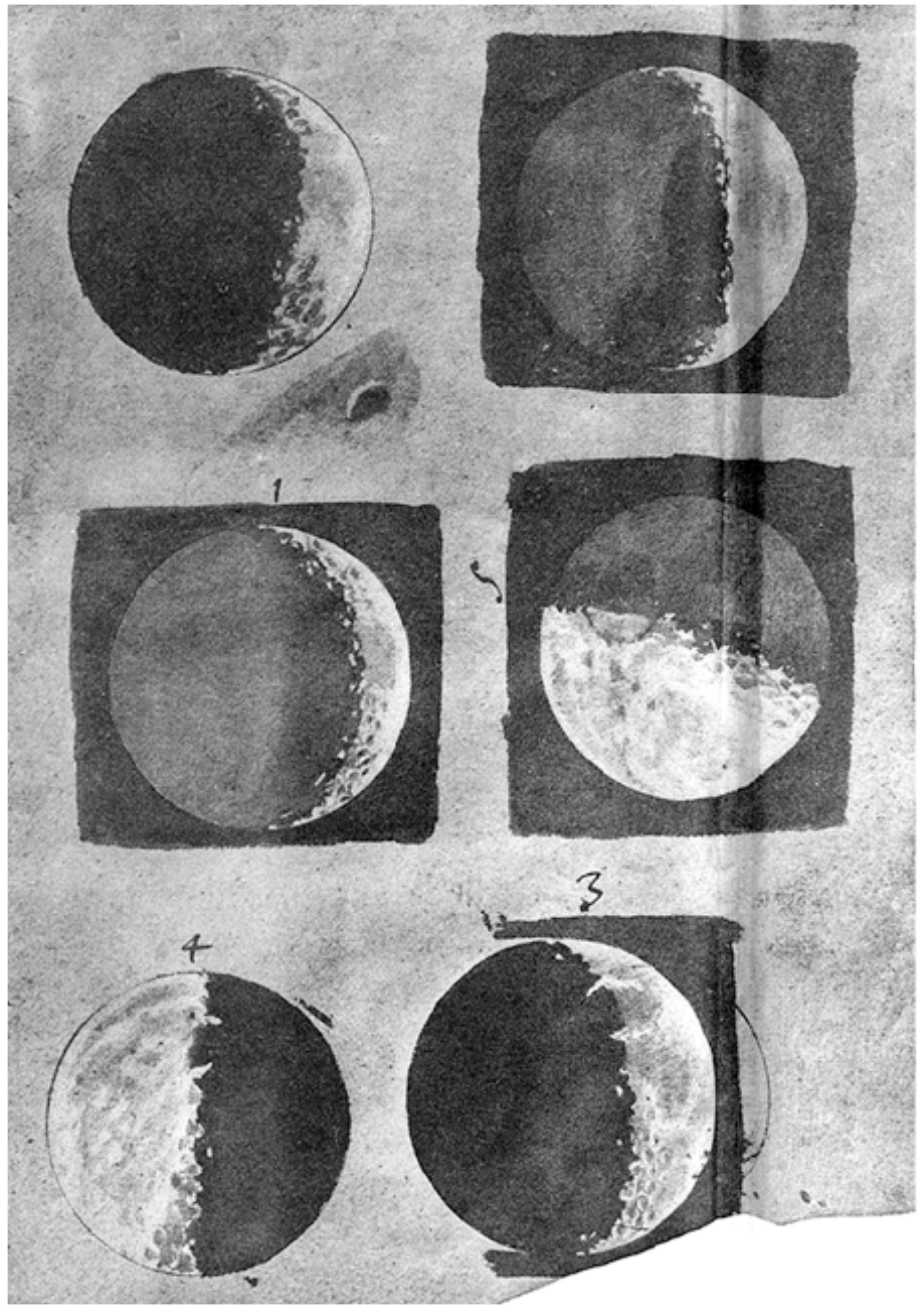
Anaximander of Miletus, c. 550 BC

Milestones Project

Record



Leonardo Da Vinci, ca. 1500



Galileo Galilei, 1616

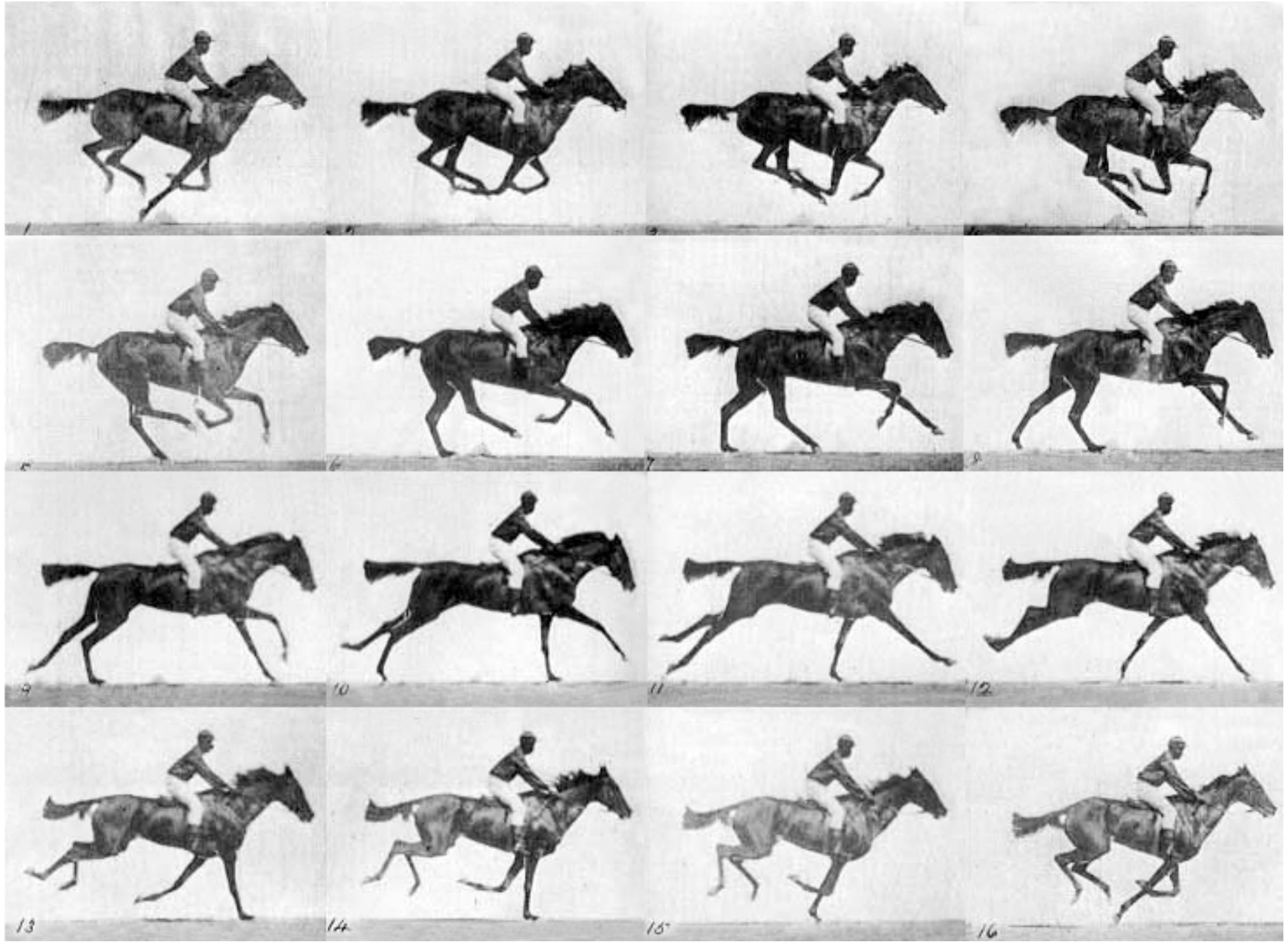
William Curtis (1746-1799)

Donald Norman

The History of Visual Communication

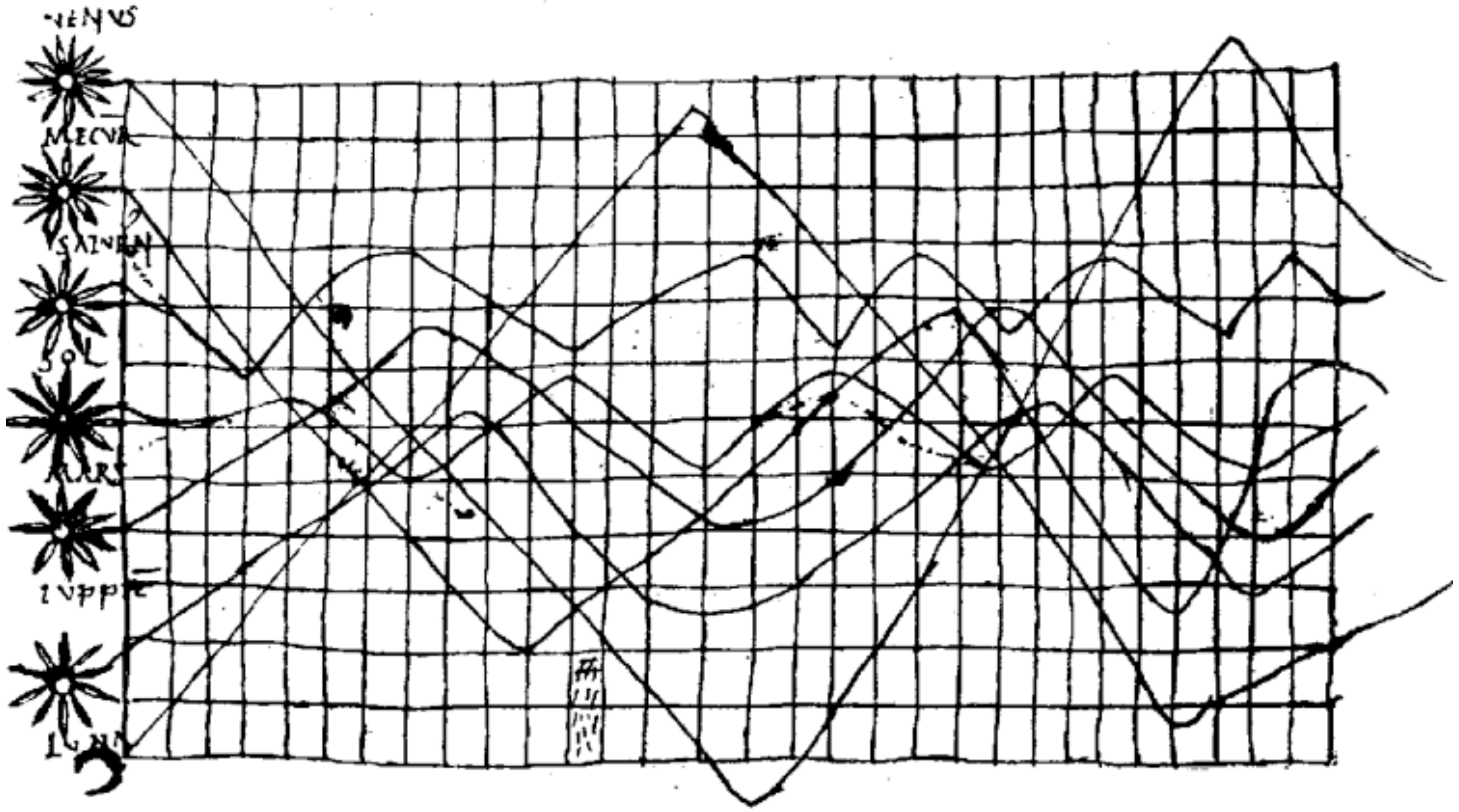
The Galileo Project, Rice University

Record

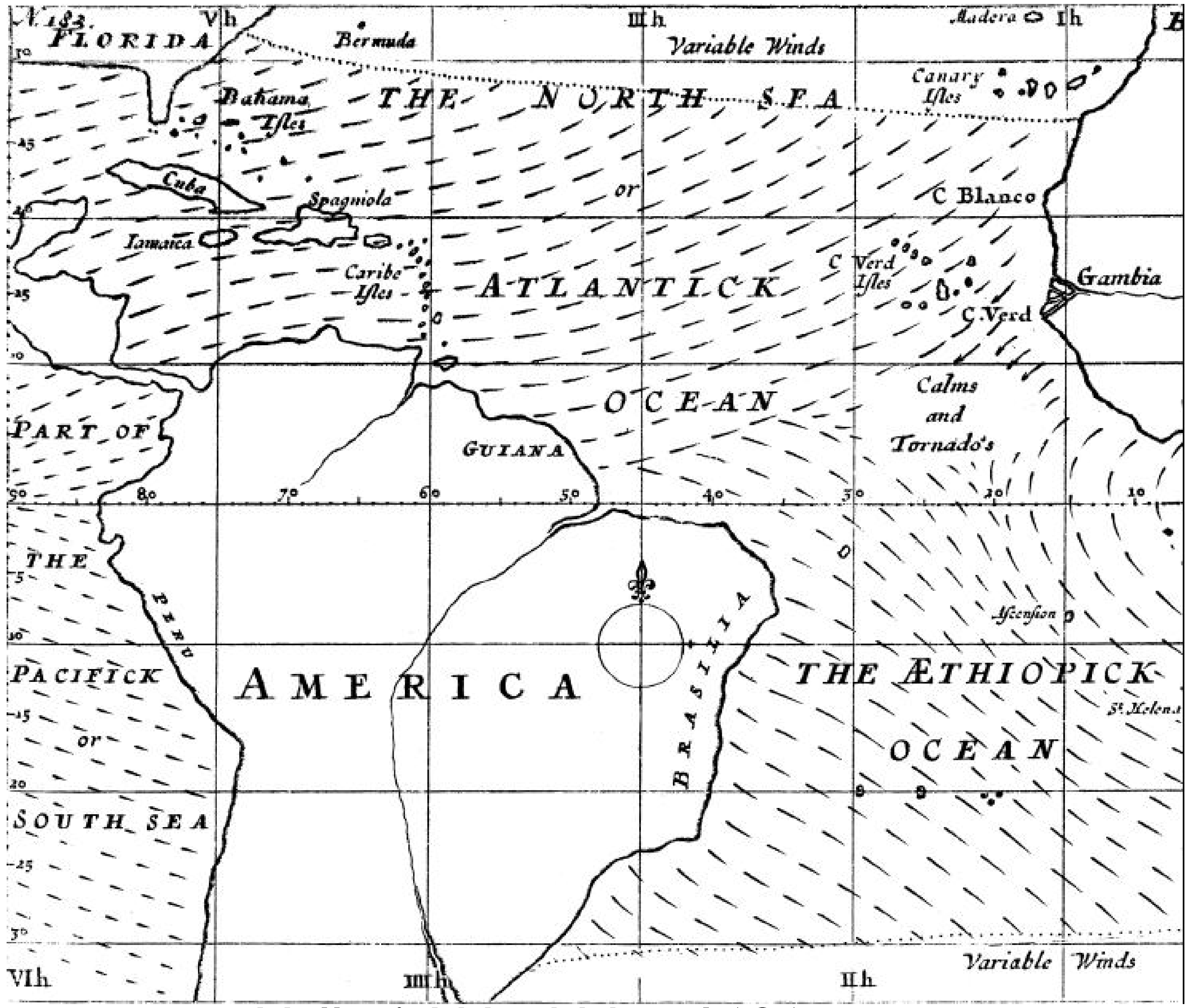


Eadweard J. Muybridge, 1878

Analyze

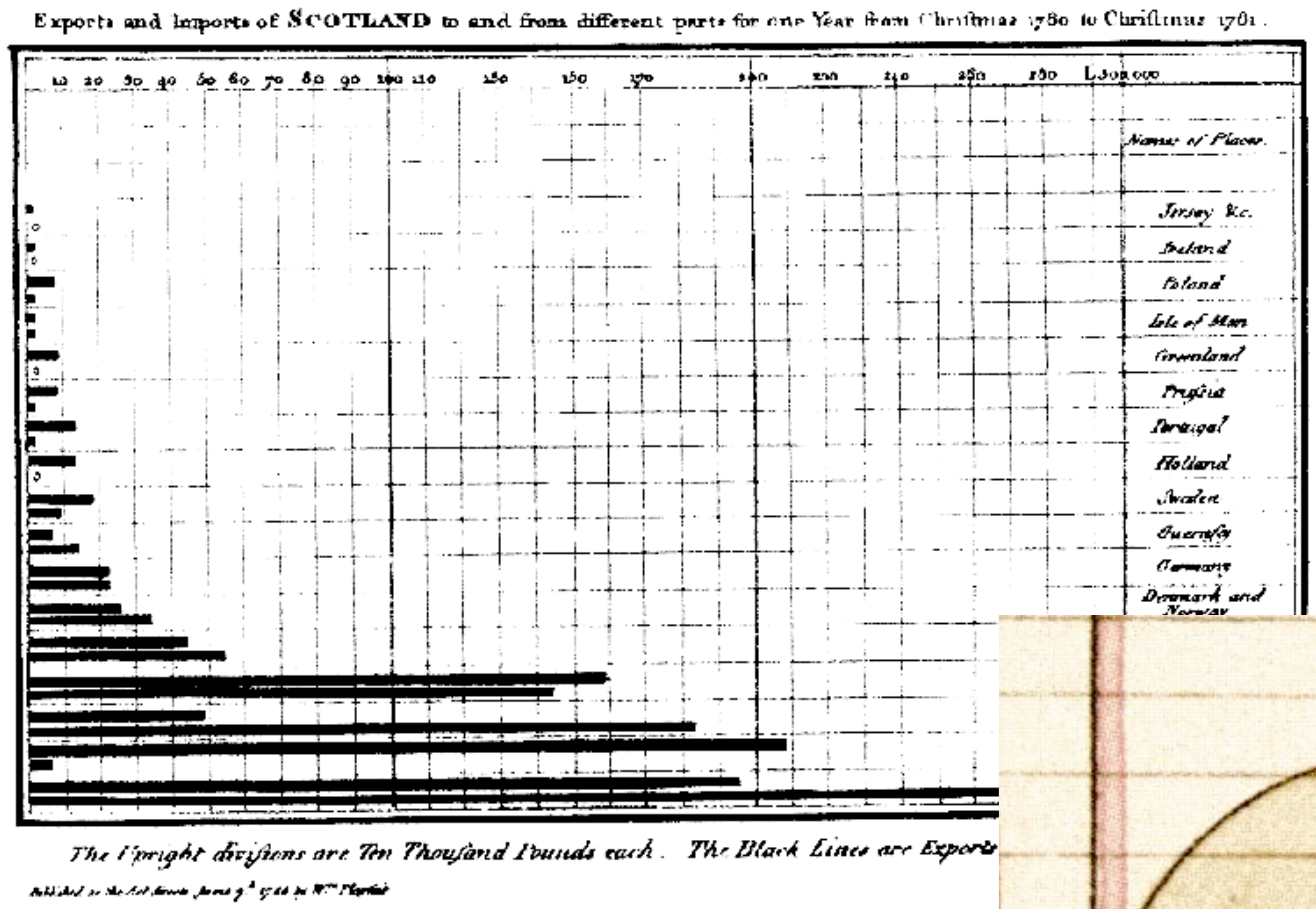


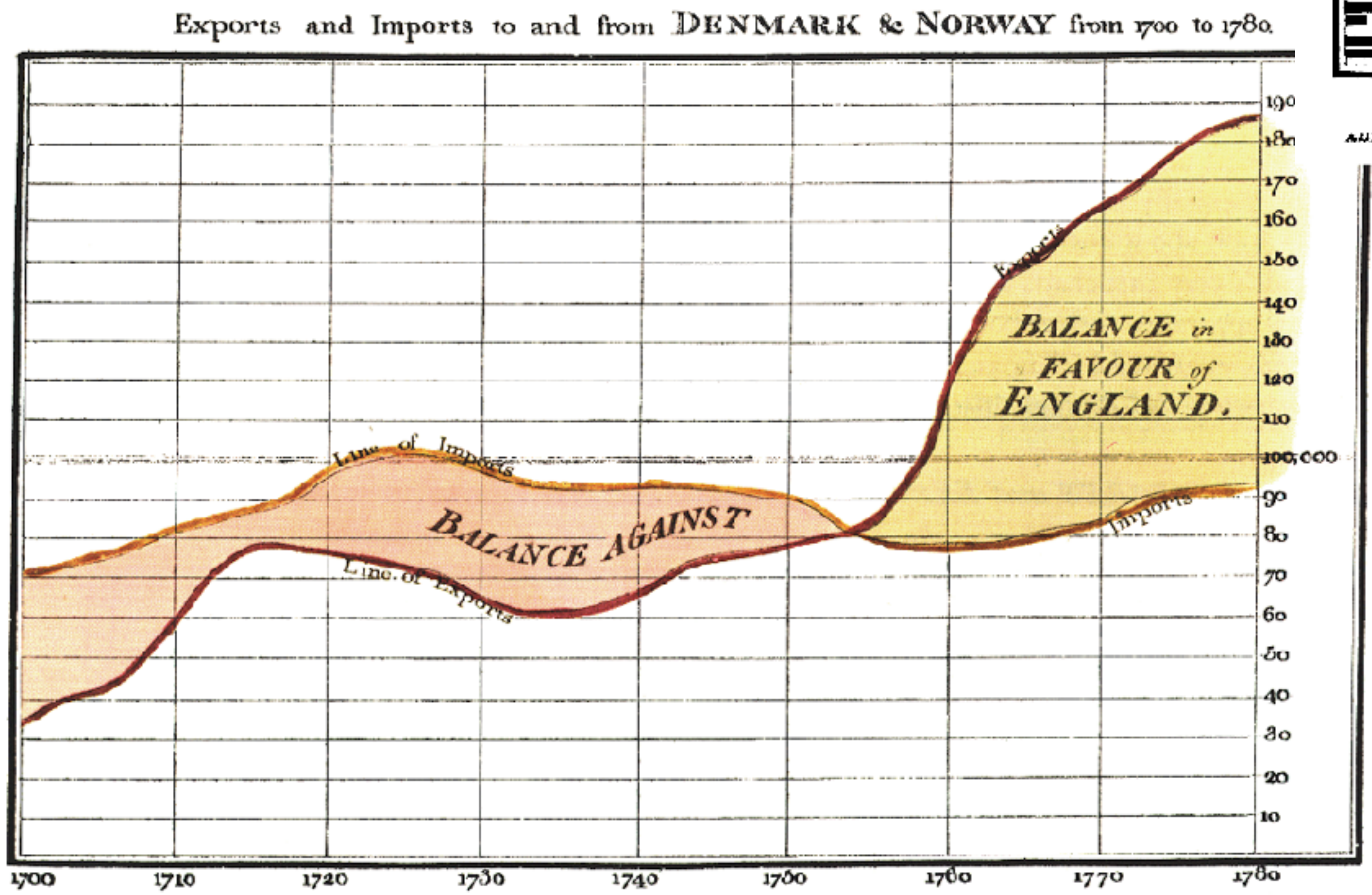
Planetary Movement Diagram, c. 950



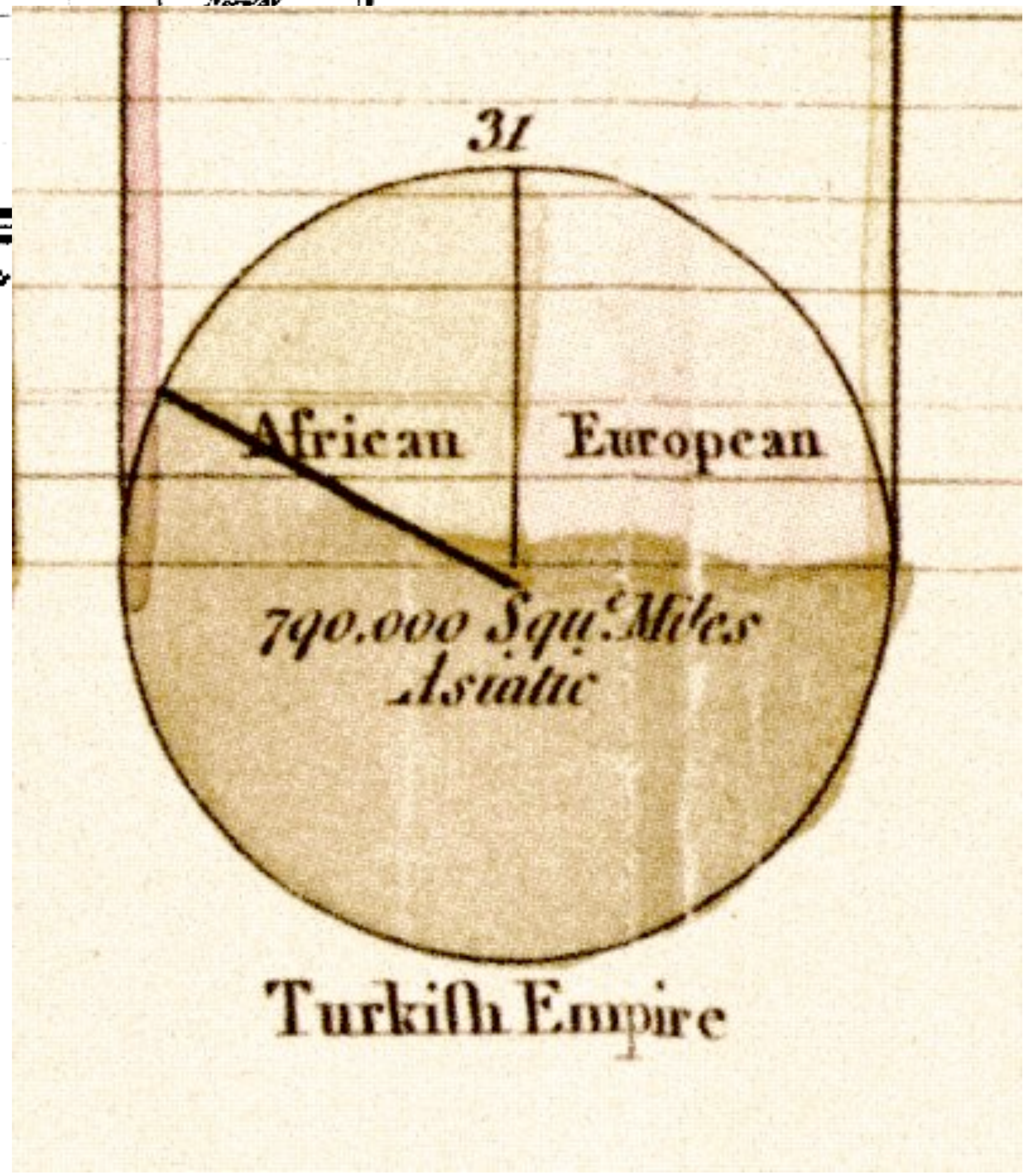
Halley's Wind Map, 1686

Analyze





W. Playfair, 1786



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

wikipedia.org

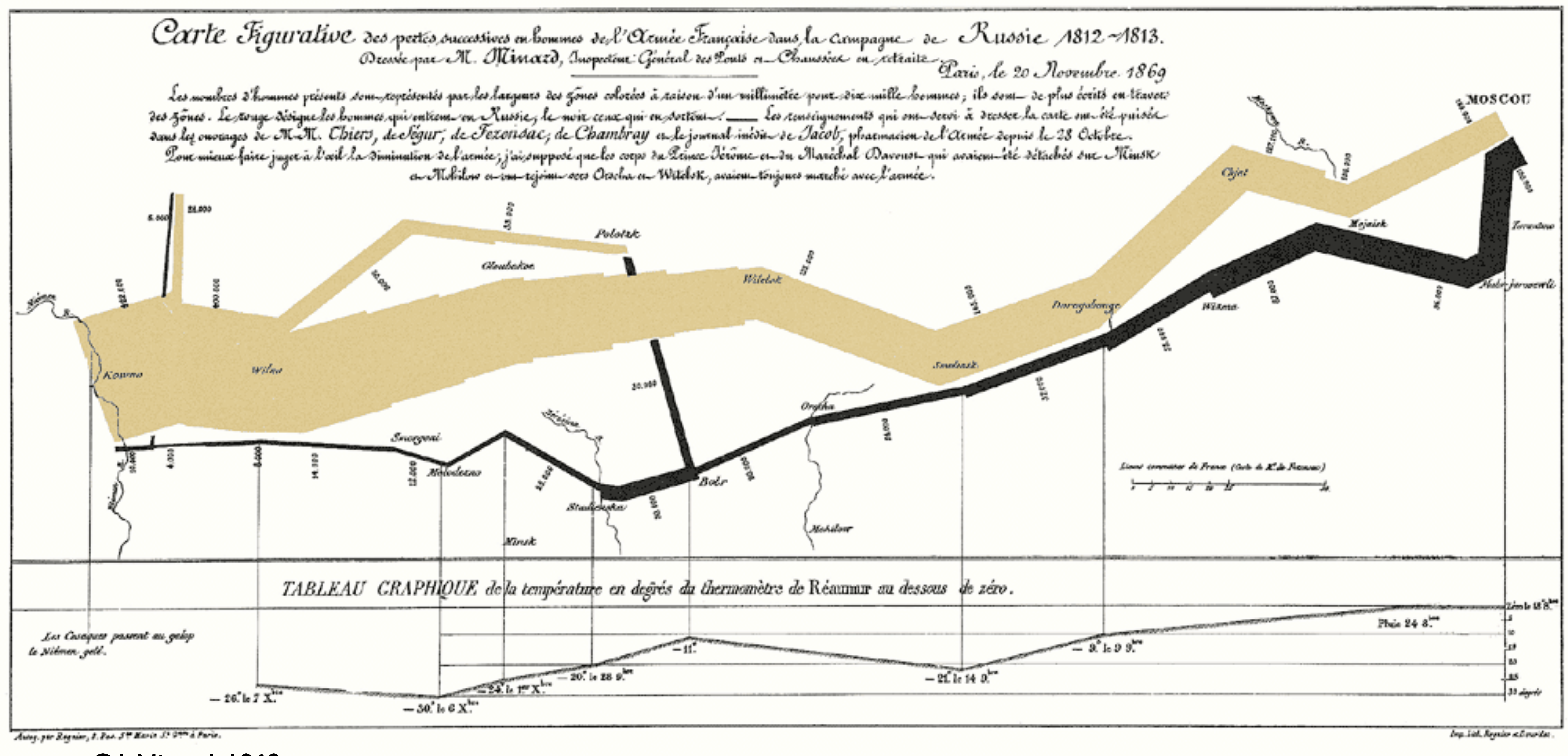
W. Playfair, 1801

Find Patterns

John Snow, 1854

E. Tufte, Visual Explanations, 1997

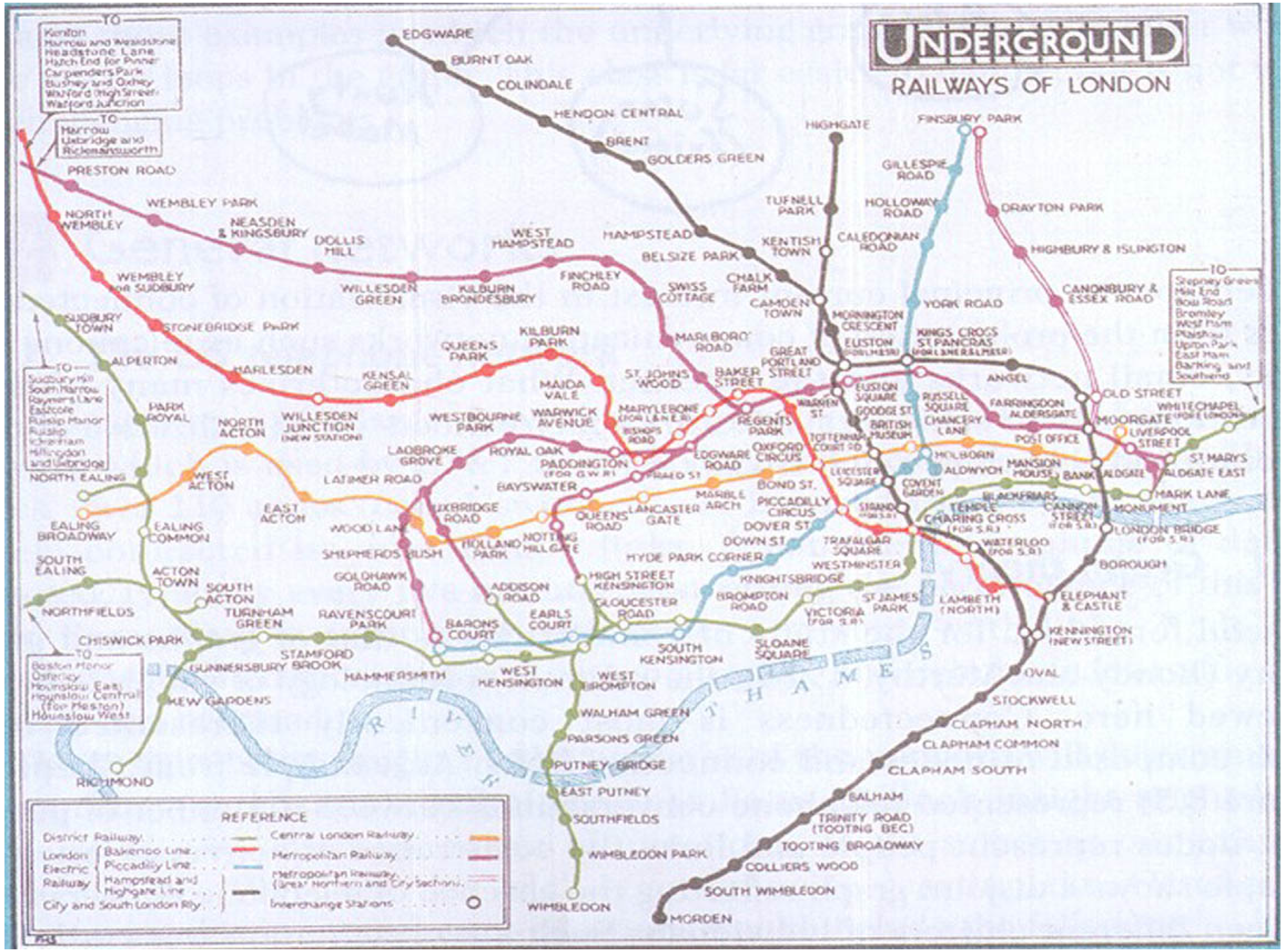
Communicate



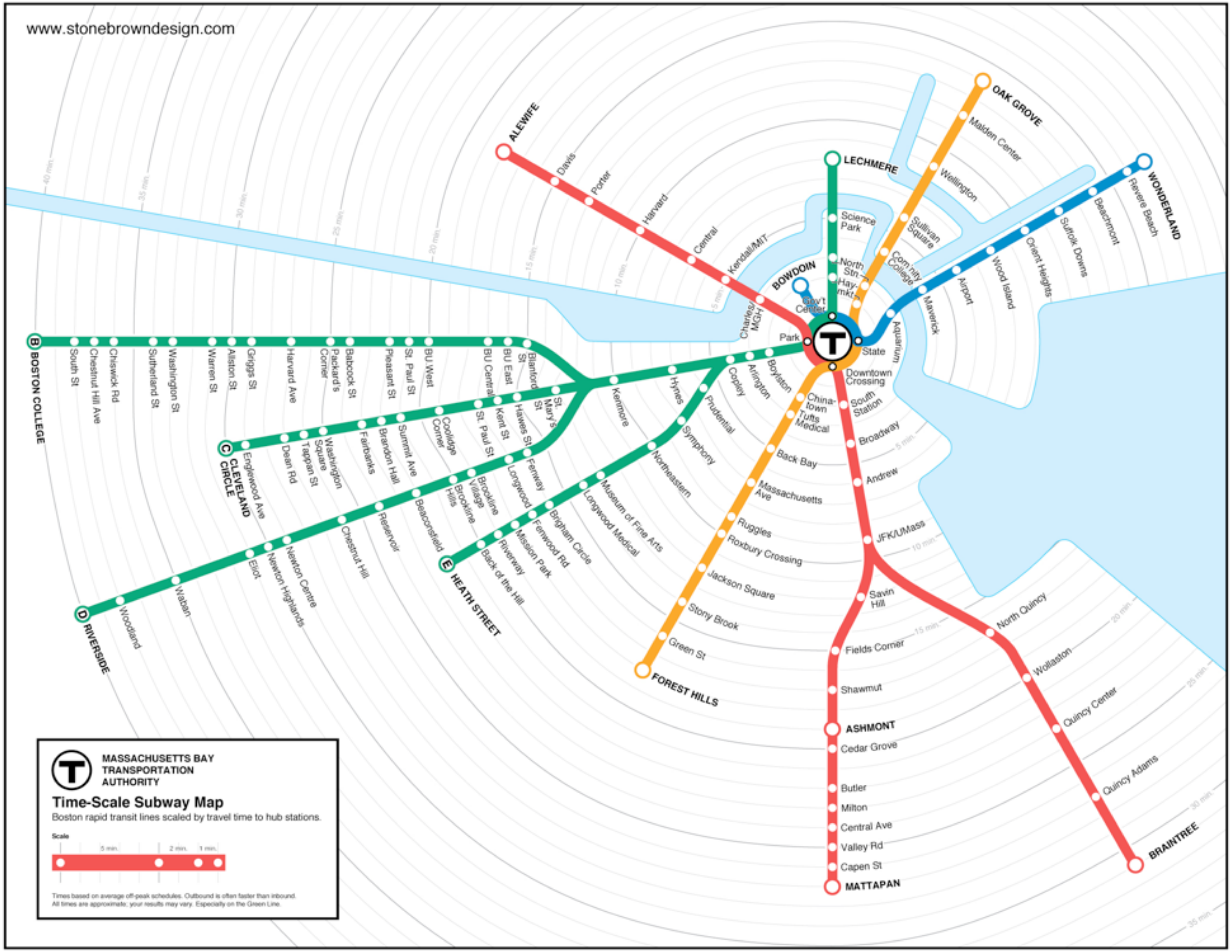
C.J. Minard, 1869

E. Tufte, Writings, Artworks, News

Communicate



London Subway Map, 1927



www.stonebrowndesign.com
ALEWIFE
Davis
Porter
Harvard
Central
Kendall/MIT
Charles/MGH
Park
LECHMERE
Science Park
North Stn
Haymkt
BOWDOIN
Gov't Center
OAK GROVE
Malden Center
Wellington
Sullivan Square
Com'nity College
WONDERLAND
Revere Beach
Beachmont
Suffolk Downs
Orient Heights
Wood Island
Airport
Maverick
Aquarium
State
Downtown Crossing
Chinatown
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
Boylston
Arlington
Copley
Hynes
Kenmore
Prudential
Symphony
Northeastern
Museum of Fine Arts
Longwood Medical
Brigham Circle
Fenwood Rd
Mission Park
Riverway
Back of the Hill
HEATH STREET
B 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
Blandford 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
D RIVERSIDE
Woodland
Waban
Eliot
Newton Highlands
Newton Centre
Chestnut Hill
Reservoir
Beaconsfield
Brookline Hills
Brookline Village
Longwood
Fenway
5 min.
10 min.
15 min.
20 min.
25 min.
30 min.
35 min.
40 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.

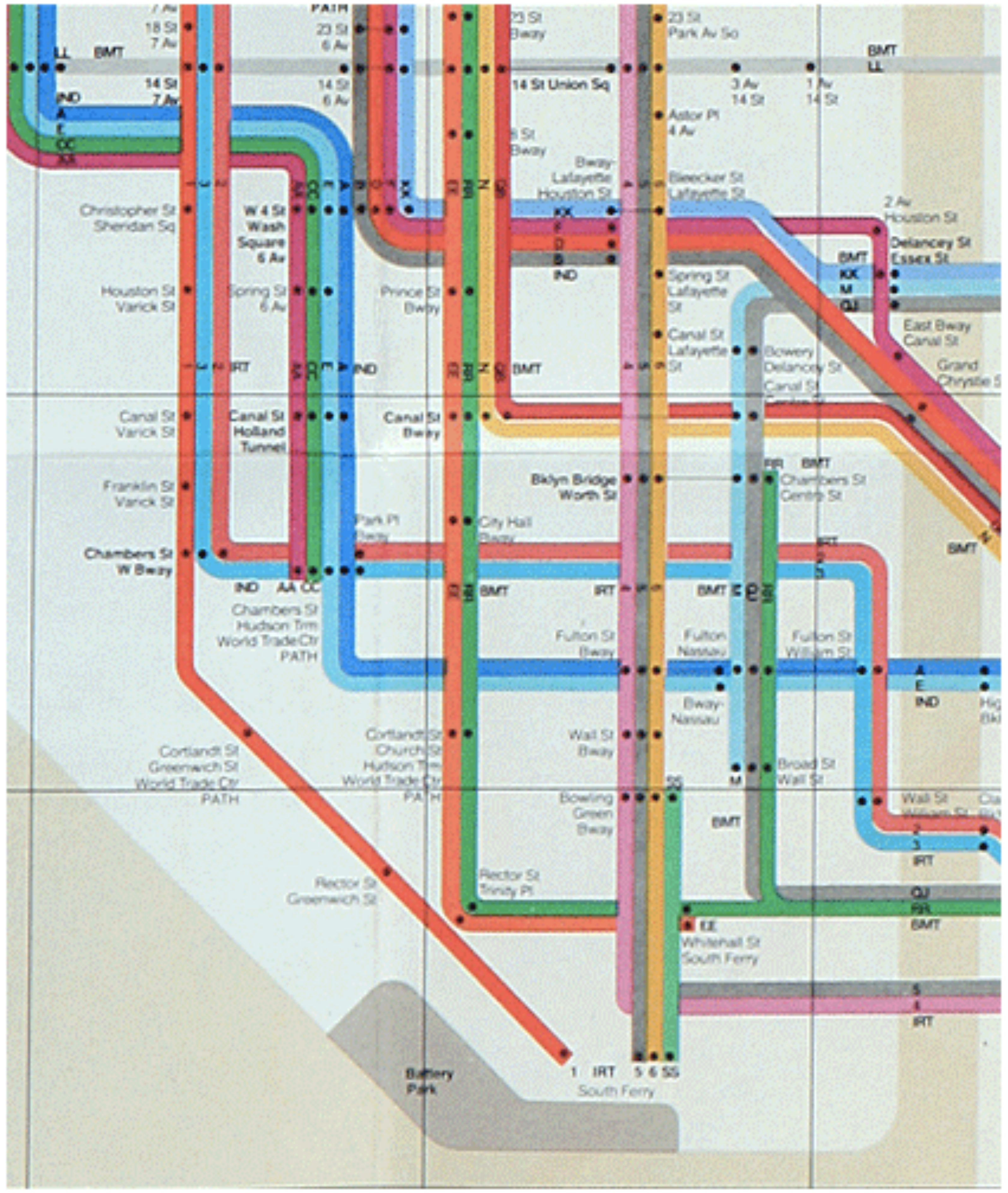
An Overhaul of an Underground Icon

Next month, the Metropolitan Transportation Authority will unveil a resized, recolored and simplified edition of the well-known map, its first overhaul in more than a decade. Related Article »

- FULL MAP
- BRONX
- BROOKLYN
- MANHATTAN
- QUEENS

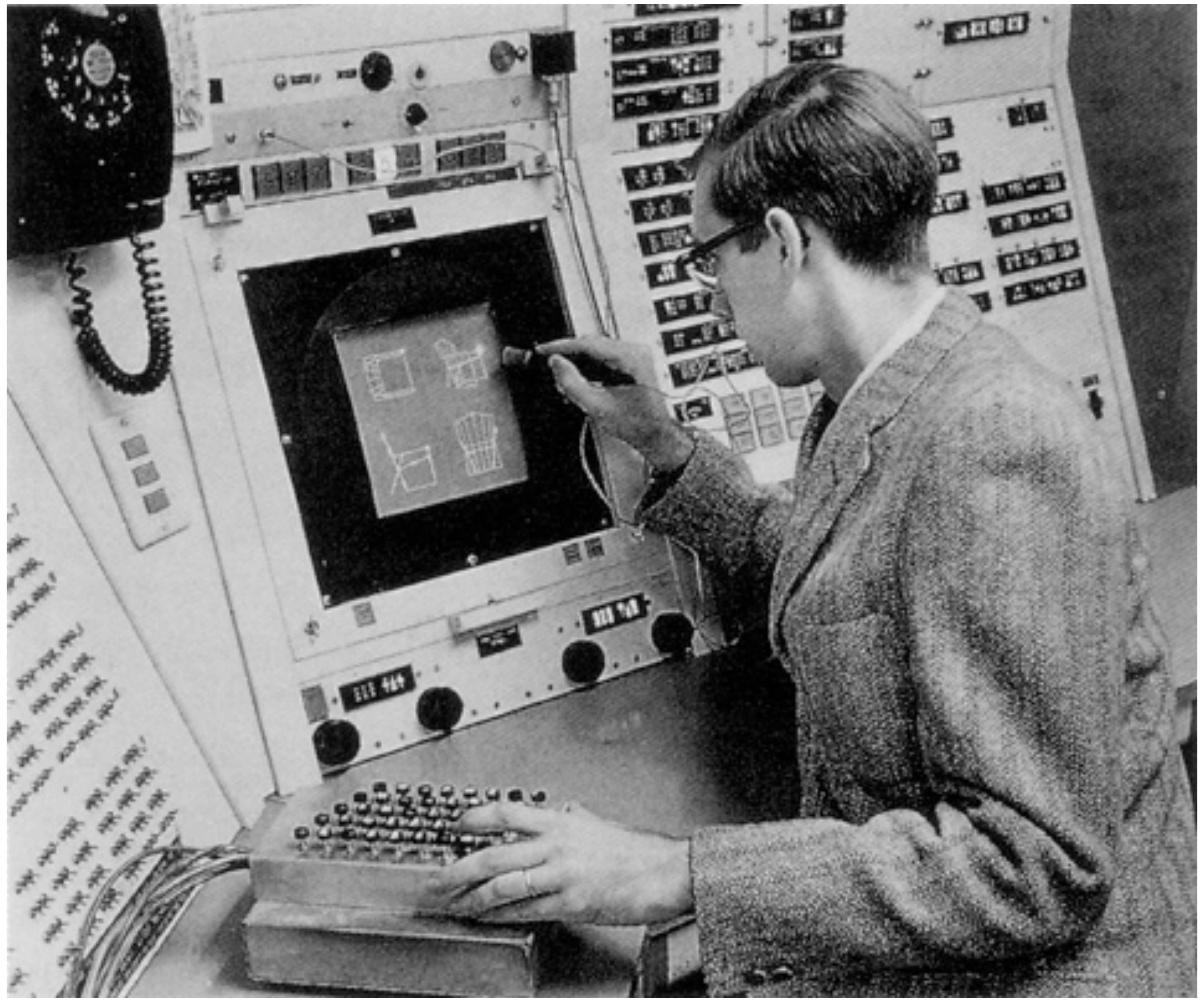
The New Map

1972: Vignelli's Classic

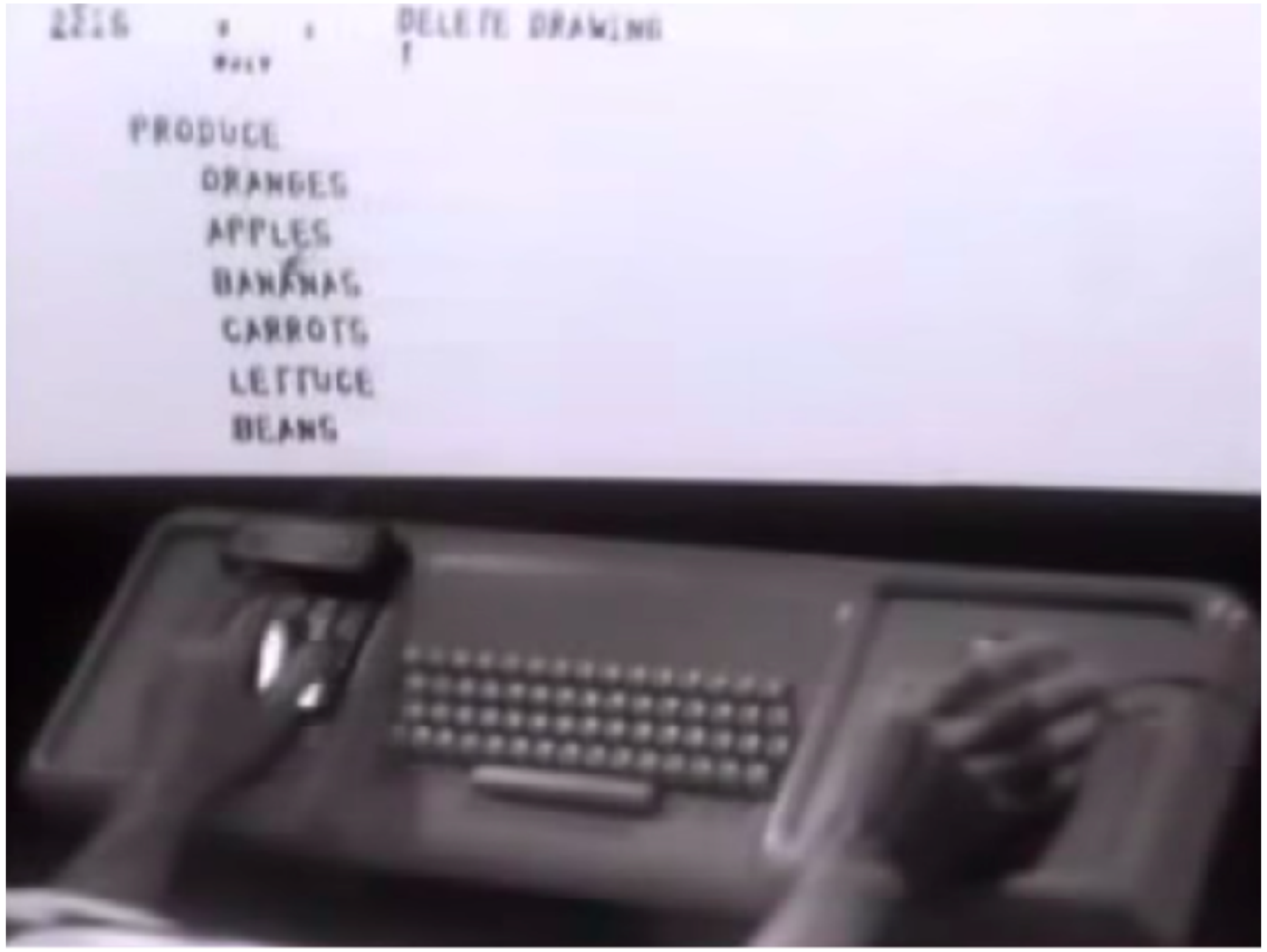


New York Times, 2010

Interact



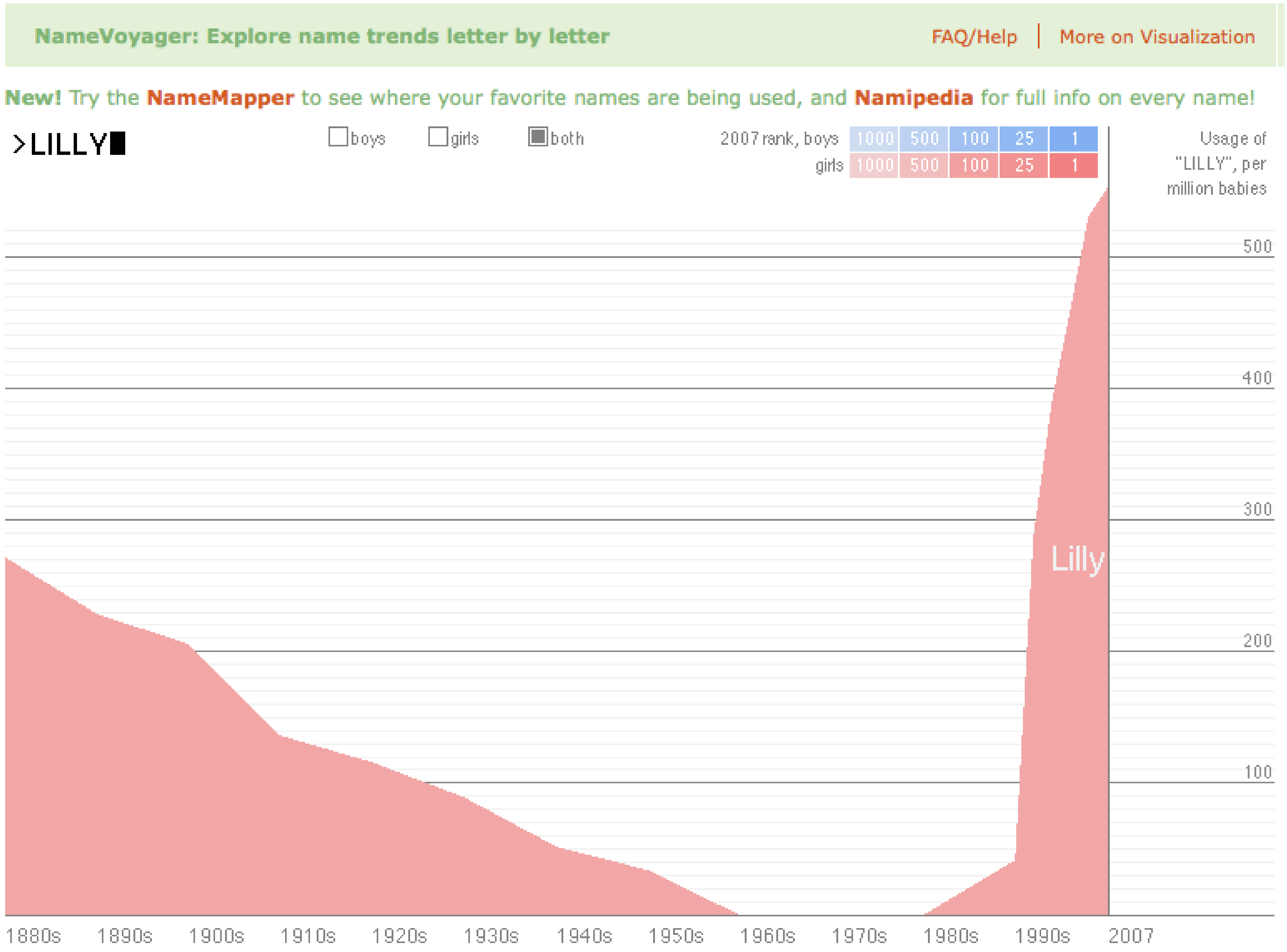
Ivan Sutherland, Sketchpad, 1963



Doug Engelbart, 1968

Modern Examples

Analyze



M. Wattenberg, 2005

Communicate

Hans Rosling, TED 2006

Who is CS-5630 / CS-6630?

Alexander Lex

@alexander_lex
http://alexander-lex.net

Assistant Professor, Computer Science

Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology

visualization design lab


Miriah Meyer


Alexander Lex

SCI
www.sci.utah.edu

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


Pascal Goffin


Ethan Kerzner


Alex Bigelow


Nina McCurdy


Jennifer Rogers


Sam Quinan


Jimmy Moore


Cameron Waller


Carolina Nobre

We're looking for PhD Students!

visualization design lab

Miriah Meyer

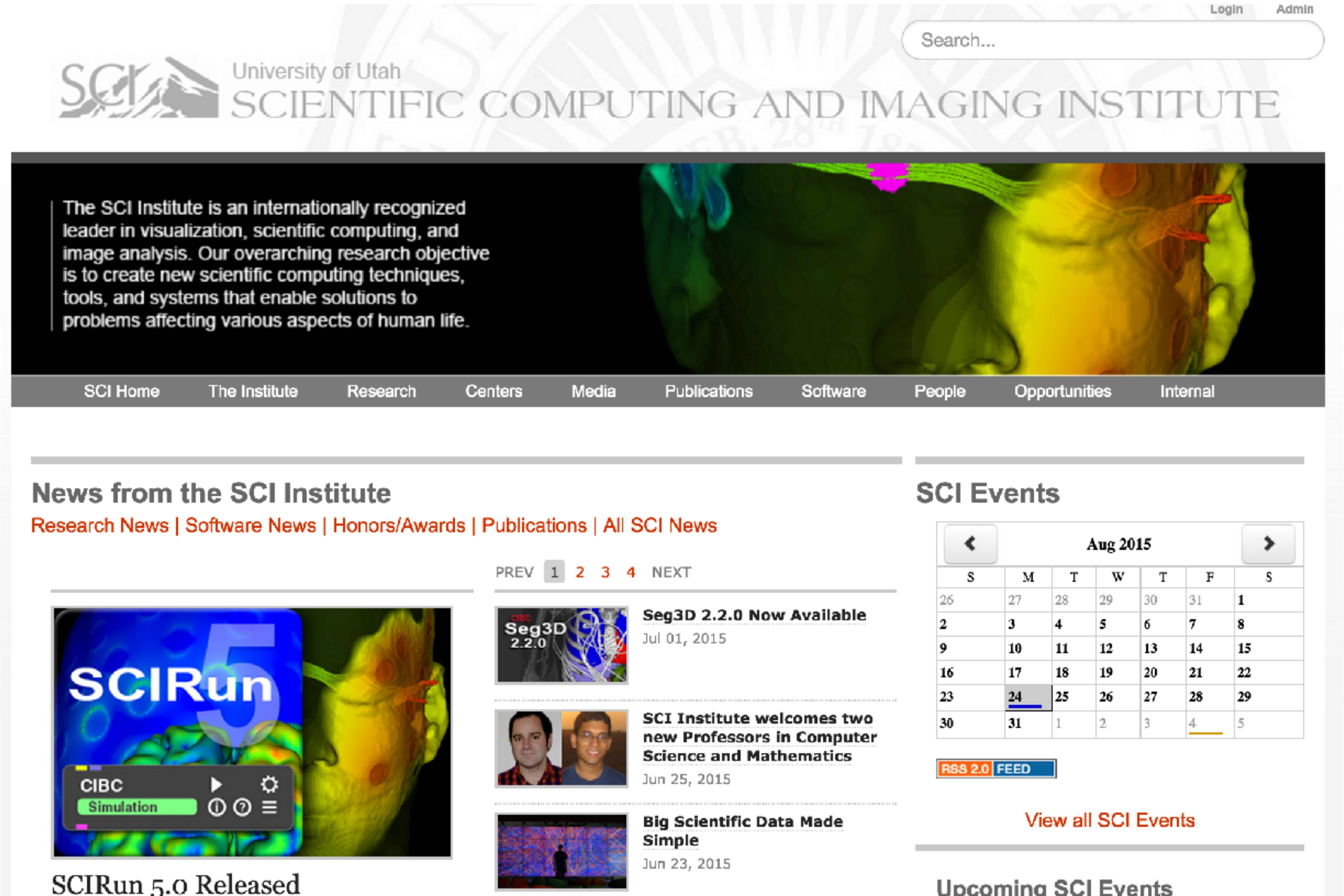
Alexander Lex

SCI Institute

Scientific Computing and Imaging Institute

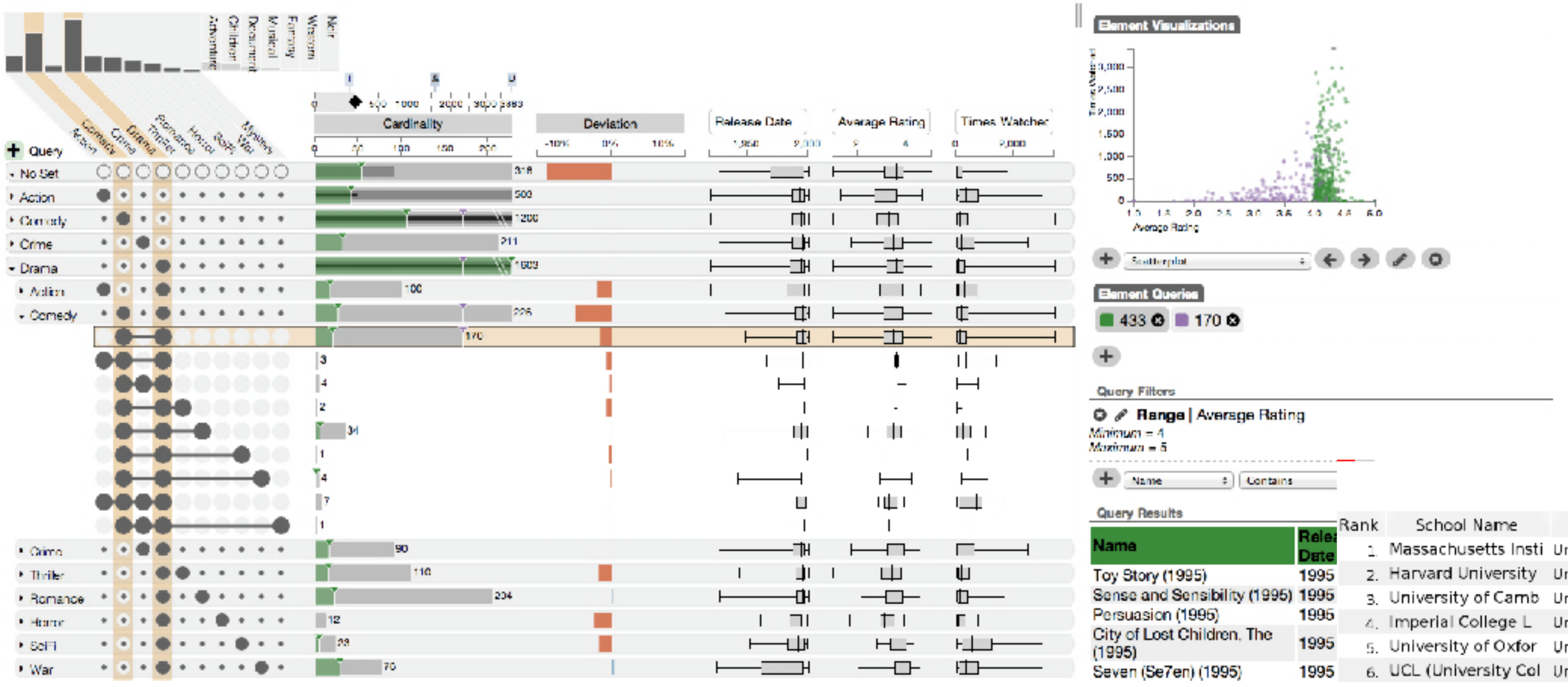
- Scientific Computing
- Biomedical Computing
- Scientific Visualization
- Information Visualization
- Image Analysis

http://sci.utah.edu



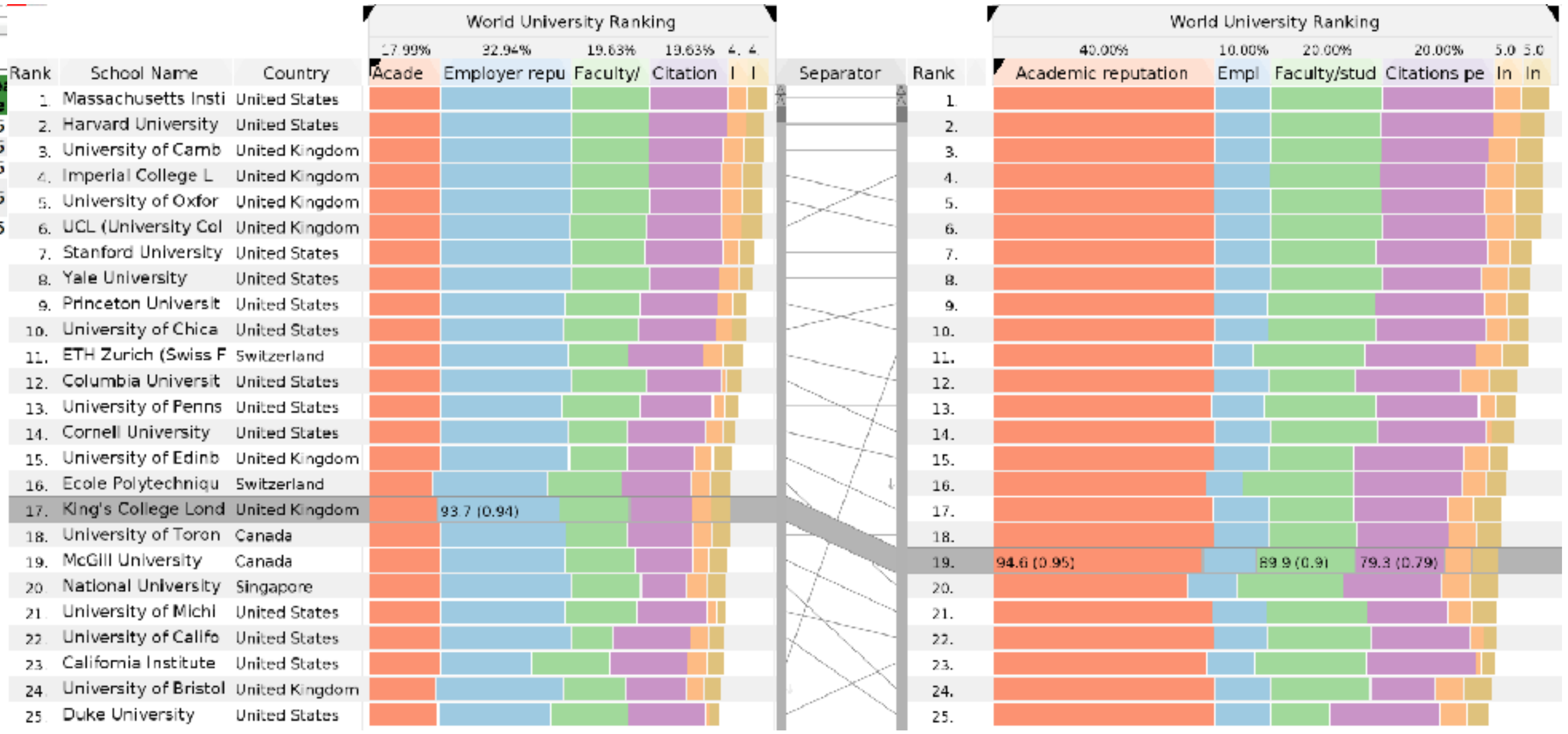
Large, Multivariate (Biological) Networks

Multidimensional Data

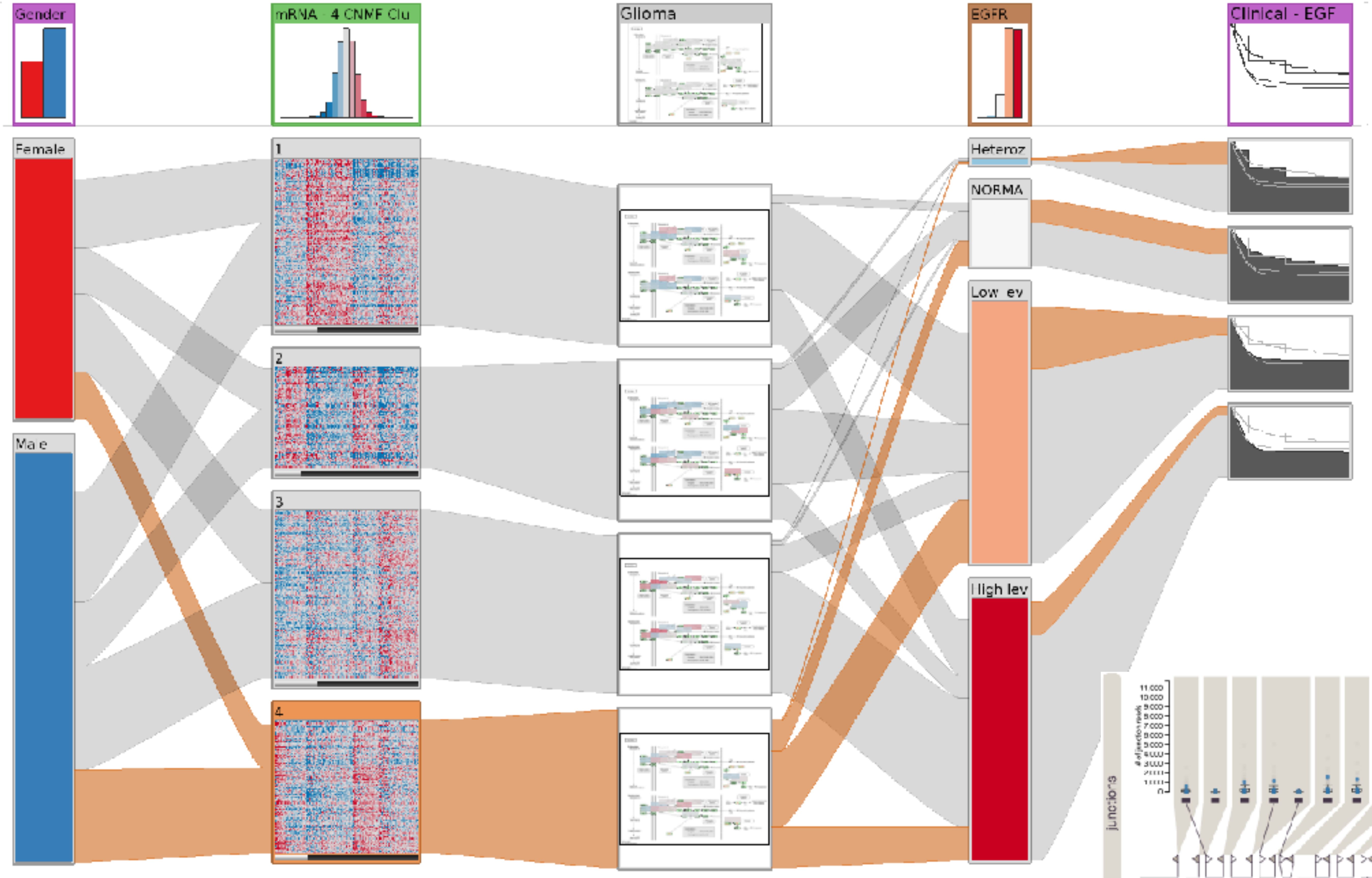


Set Visualization

Multivariate Rankings

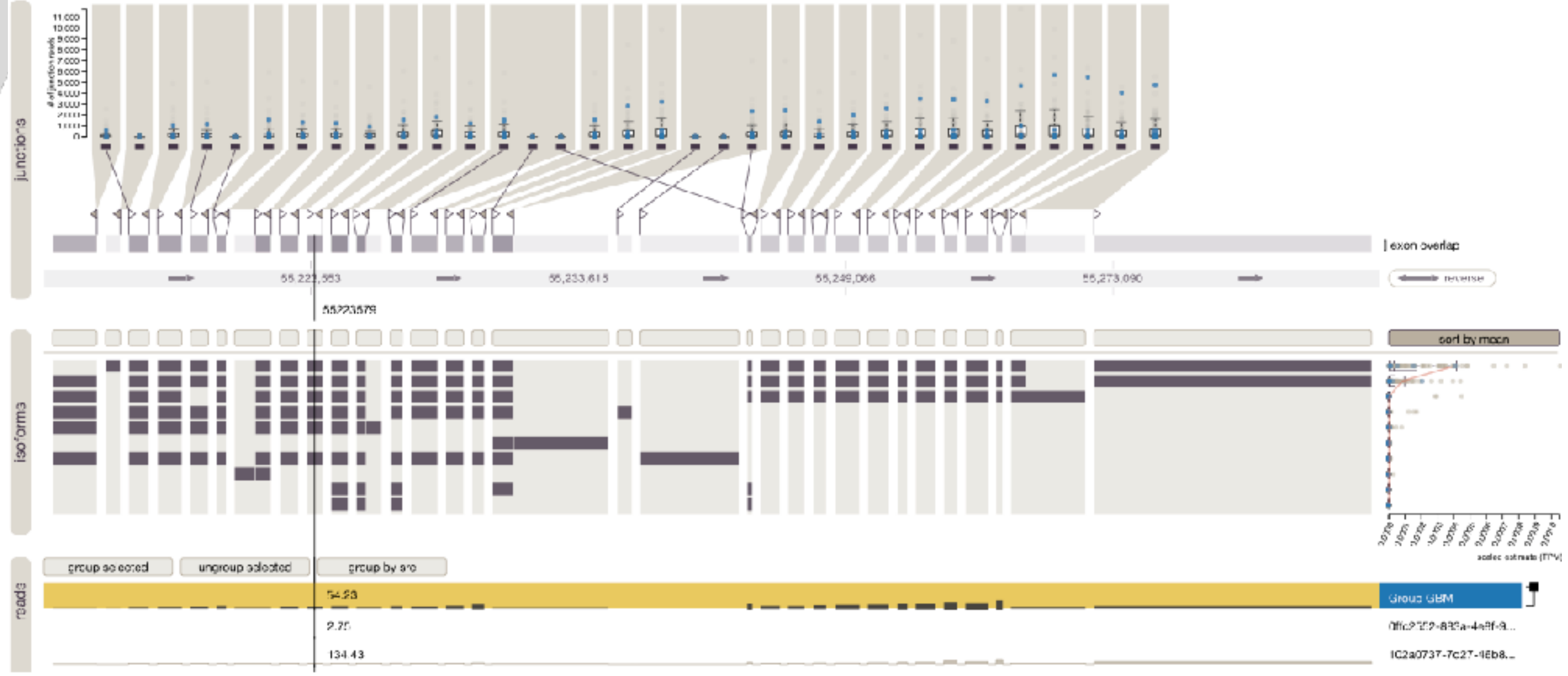


Genomic Data

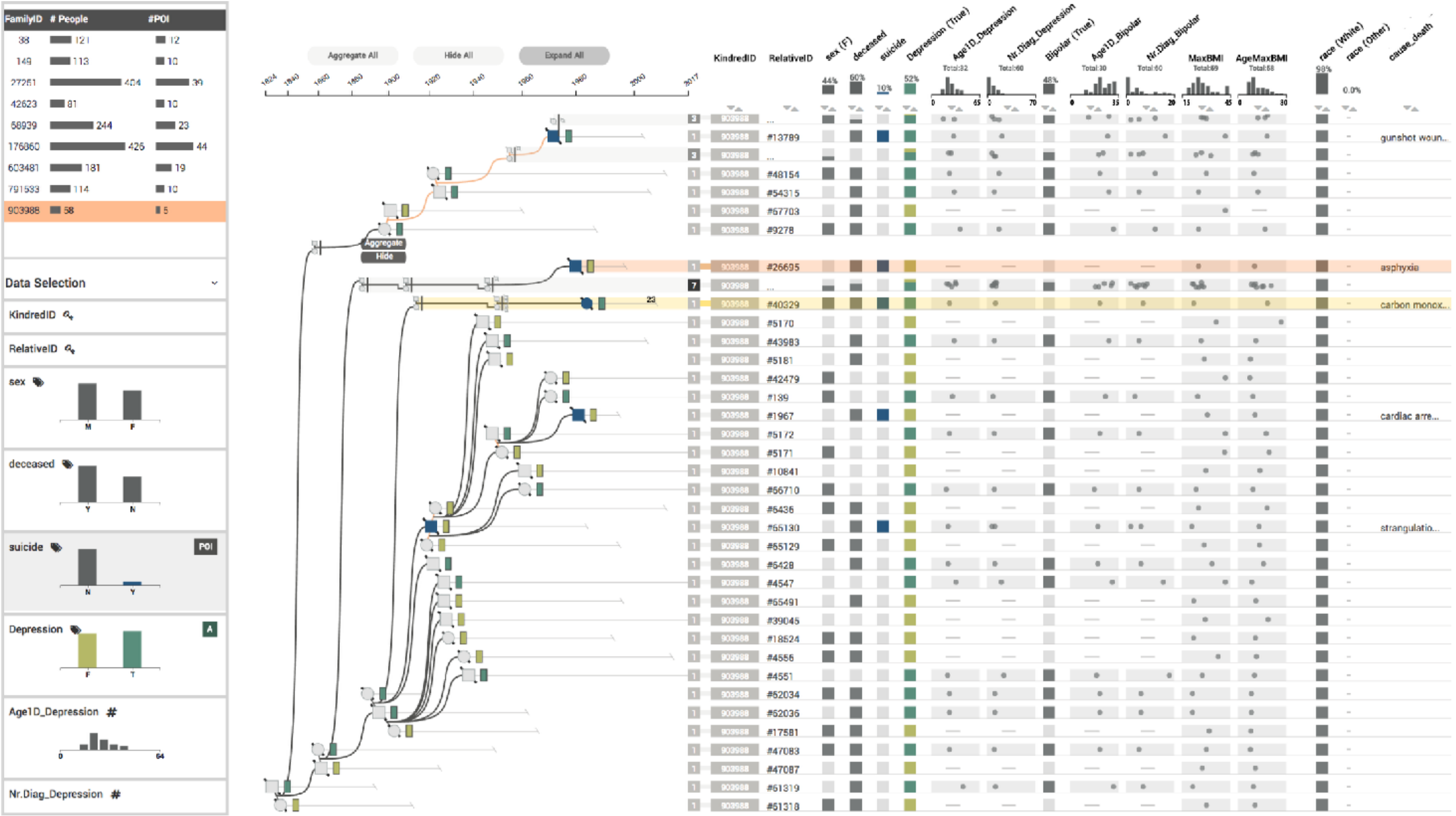


Cancer Subtypes / Omics Clustering and Stratification

Alternative Splicing / mRNA-seq



Genealogies & Clinical Data



Course Staff

TBA

Teaching Assistant

Carolina Nobre

Teaching Mentee

Pranav Dommata

Teaching Assistant

About You

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

Lectures: introduce theory

- **Design Critiques:** develop “an eye” for vis design, critique, learn by example

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

Course Components

Lecture
Reading
Discussion

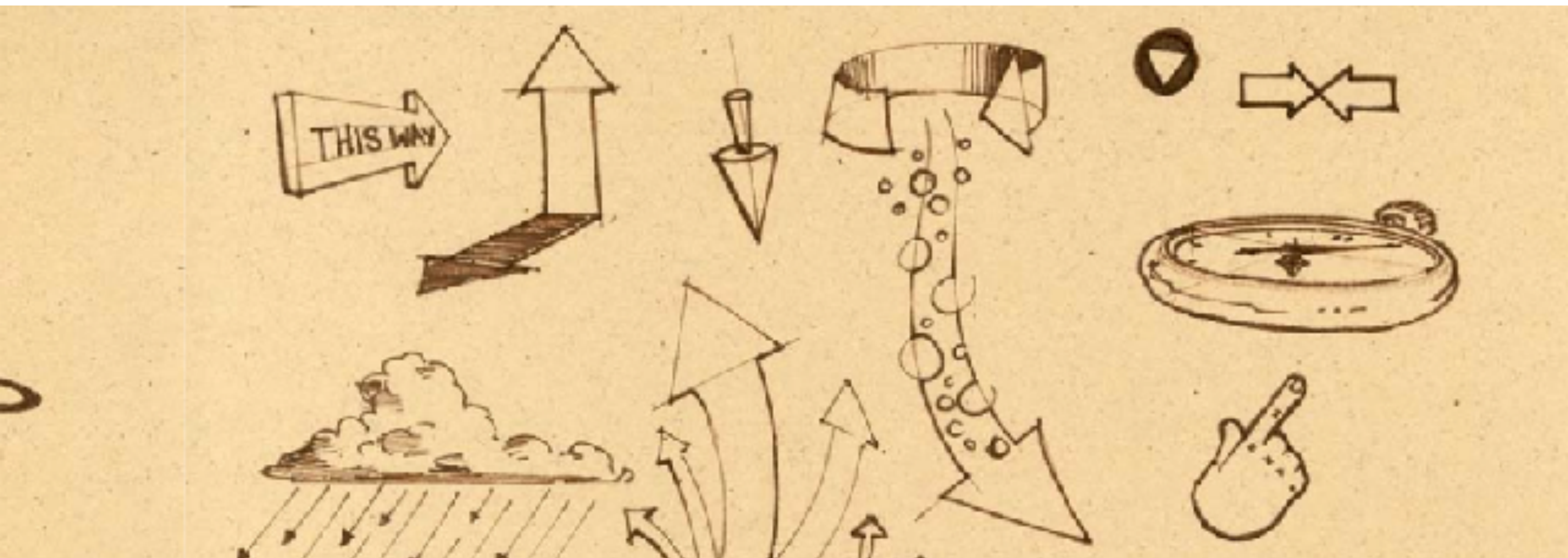
Theory

Design Lecture
Design Studios

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>
```

Schedule

Lectures: Tuesday and Thursday 2:00-3:20 pm, L101 WEB
Labs: Wednesday, 6:00-7:30 pm, L110 WEB (scheduled on demand)

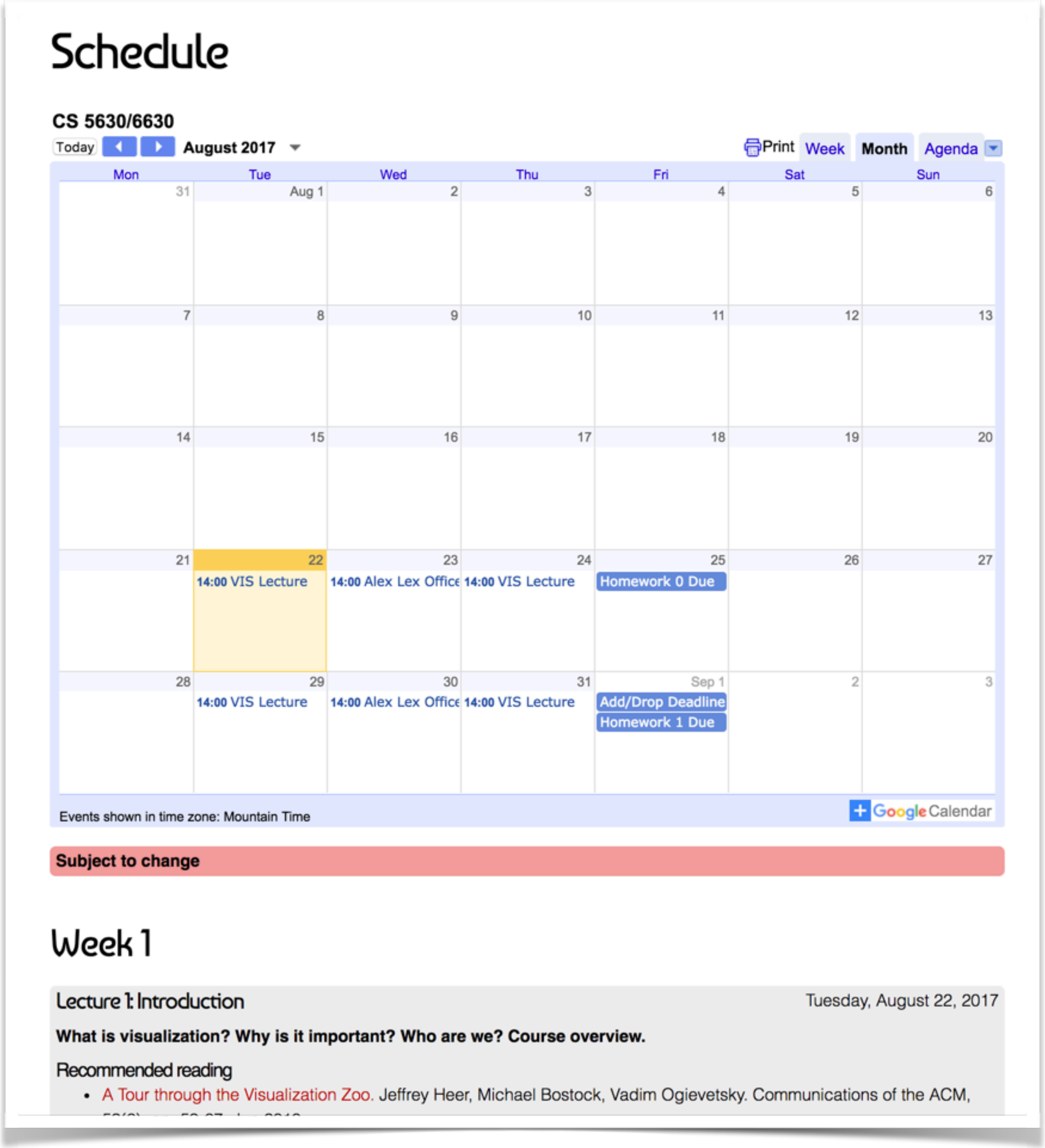
Online Students:
YouTube Channel

Three Parts:

I. Technical Foundations
HTML, Javascript, D3

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

III. Abstract Data Visualization
Tables, Graphs, Maps



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.

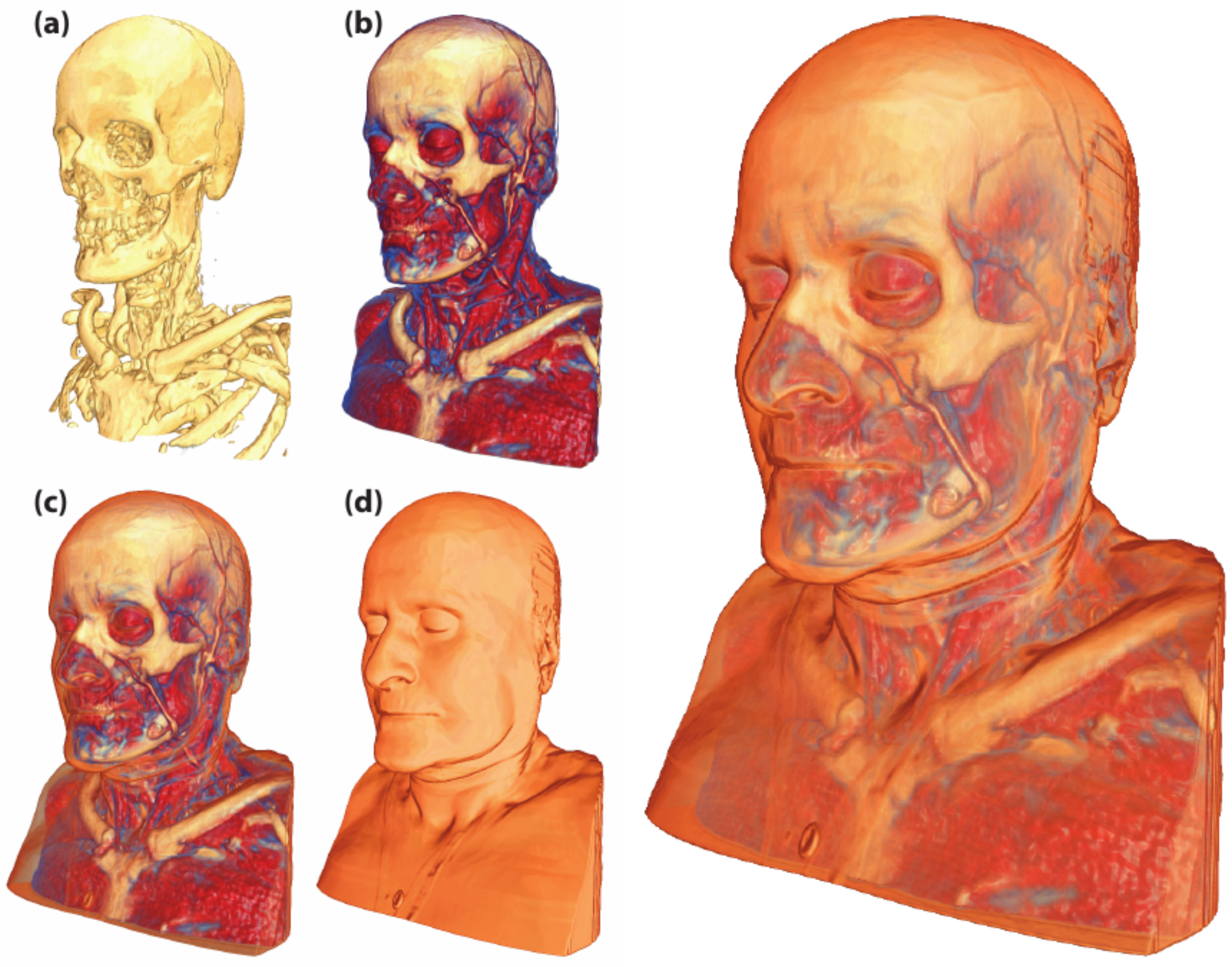
The course is offered in the fall term 2017 at the University of Utah in two variants: CS-5630 for undergraduates and CS-6630 for graduate students, with a special section of CS-6630 (002) designated for data certificate students. Classes start on Tuesday,

Companion Course: Visualization for Scientific Data

CS 5635 / CS 6635

Chris Johnson

Spring 2018



Communicate

Slack

- **http://dataviscourse2017.slack.com/**
- **Please use slack for all general questions - code, concepts, etc.**
- **Only use e-mail for personal inquiries**

Canvas

- **https://utah.instructure.com/courses/448047**
- **Homework submissions, Grades**

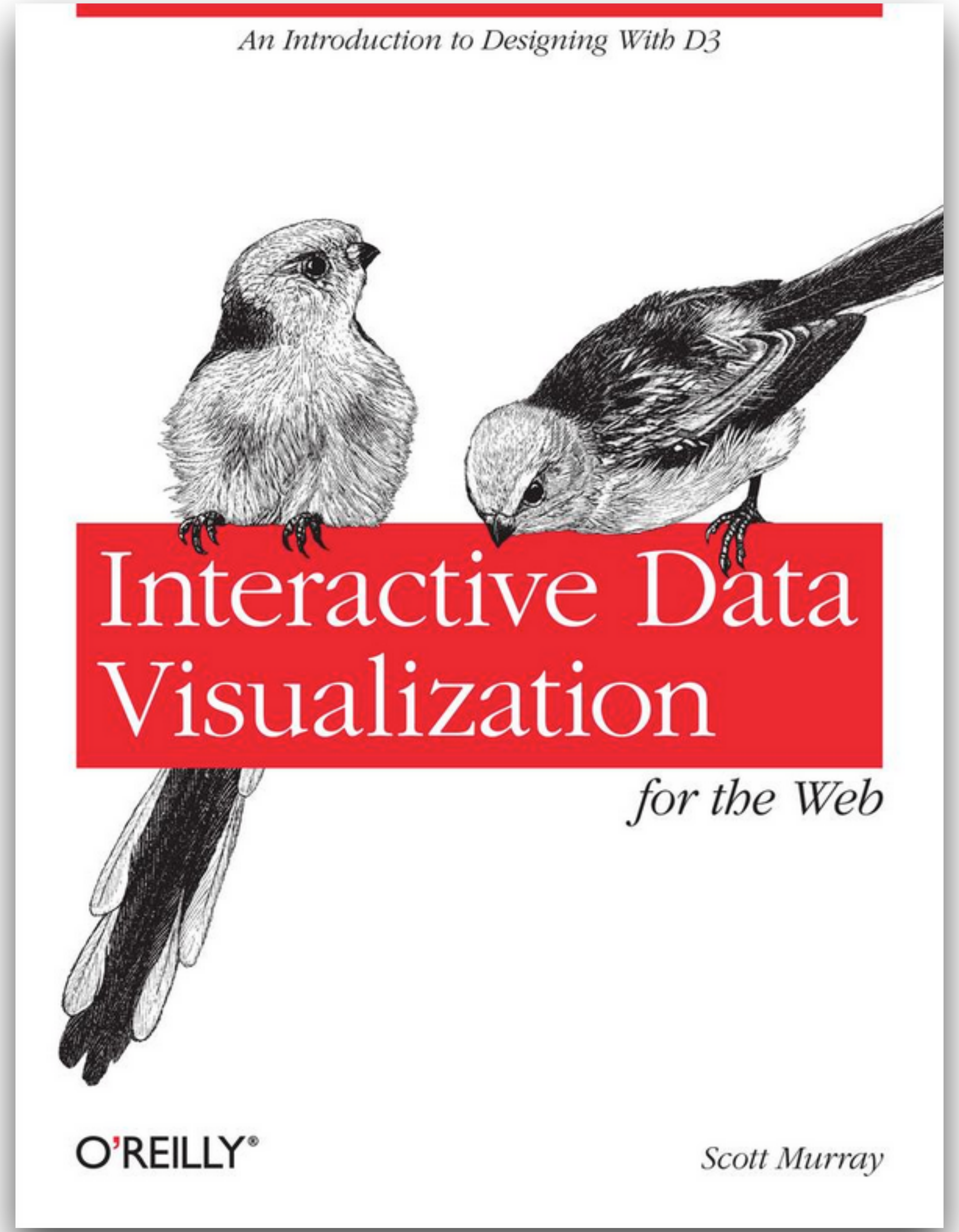
Office Hours

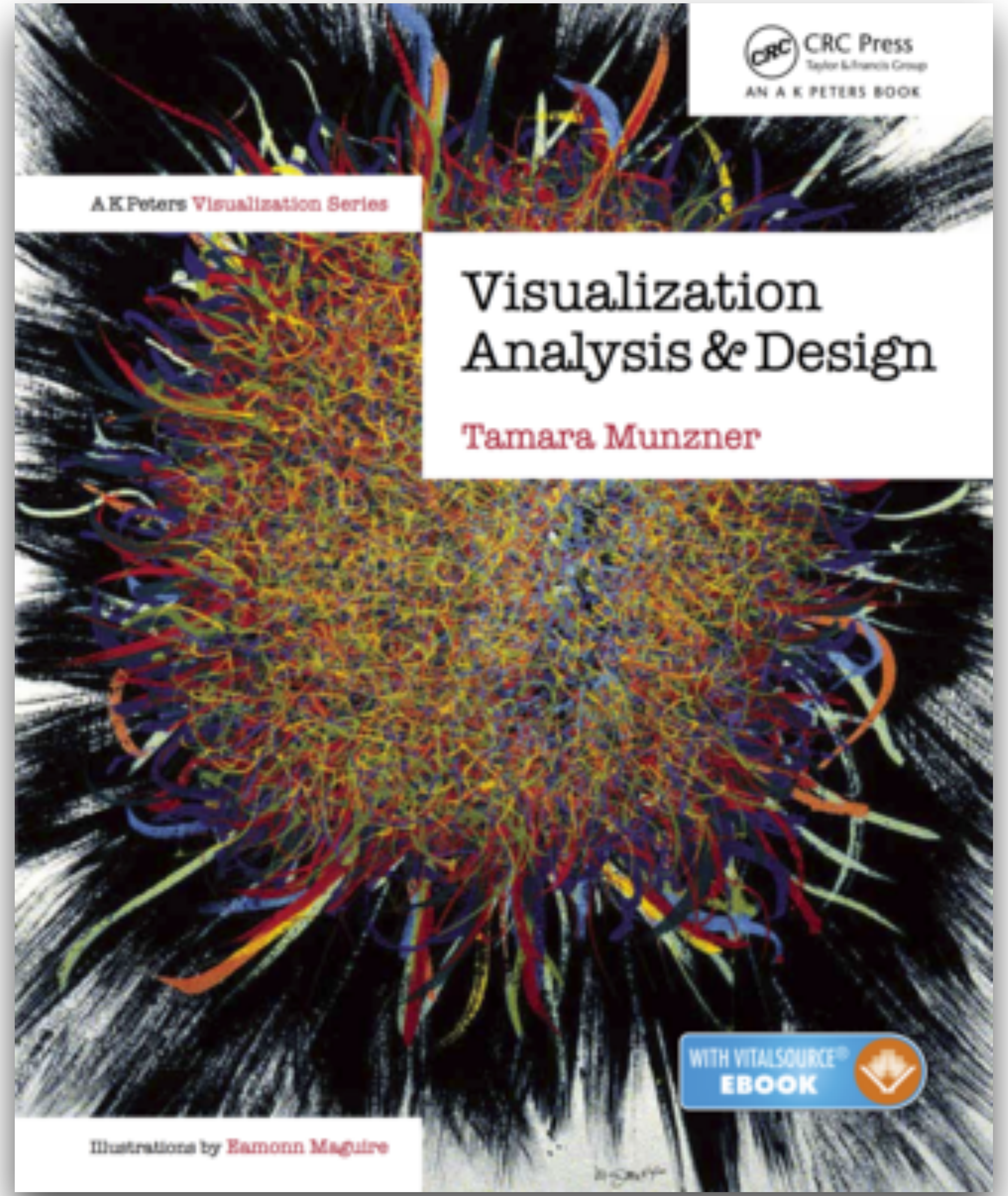
- **Alex: Wednesdays at 2pm, WEB 3887**
- **TAs: starting next week**

E-Mail

- **alex@sci.utah.edu**

Required Books





Programming

HTML

JS

Data-Driven Documents

Is this course for me ???

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!
  - Engineering vs Computer Science

Formalities

How are you graded?

6 Homework Assignments: 40%

- 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

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.

Will **automatically check for plagiarism** in all your submissions

No Device Policy

No Computers, Tablets, Phones in lecture hall

- except when used for exercises

Switch off, mute, flight mode

Why?

- It’s better to take note by hand
- Notifications are designed to grab your attention

Applies to Theory lectures, coding along in technical lectures encouraged

This Week

HW0, including course survey

Lecture on Perception

Readings

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

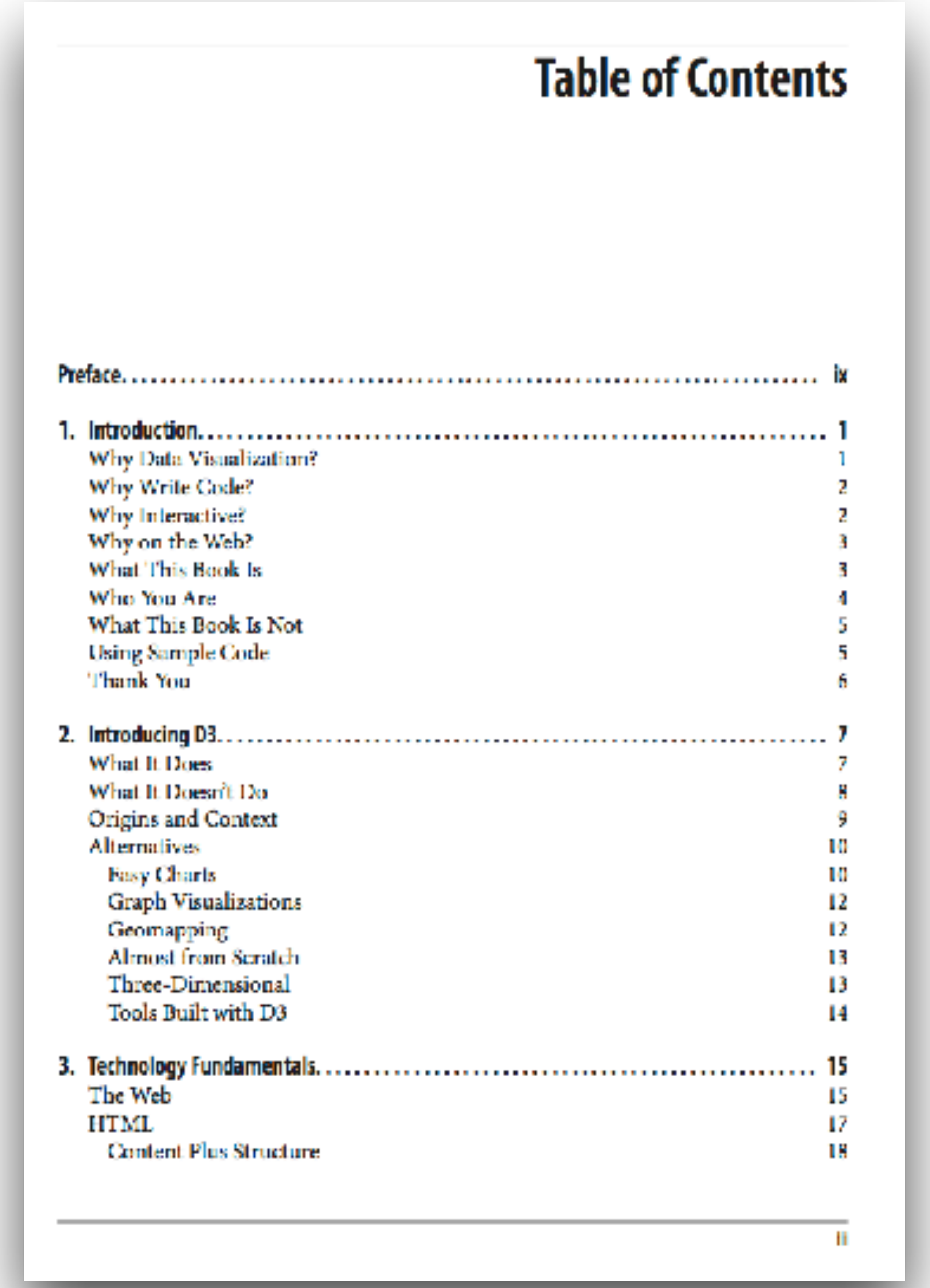


Table of Contents

iii

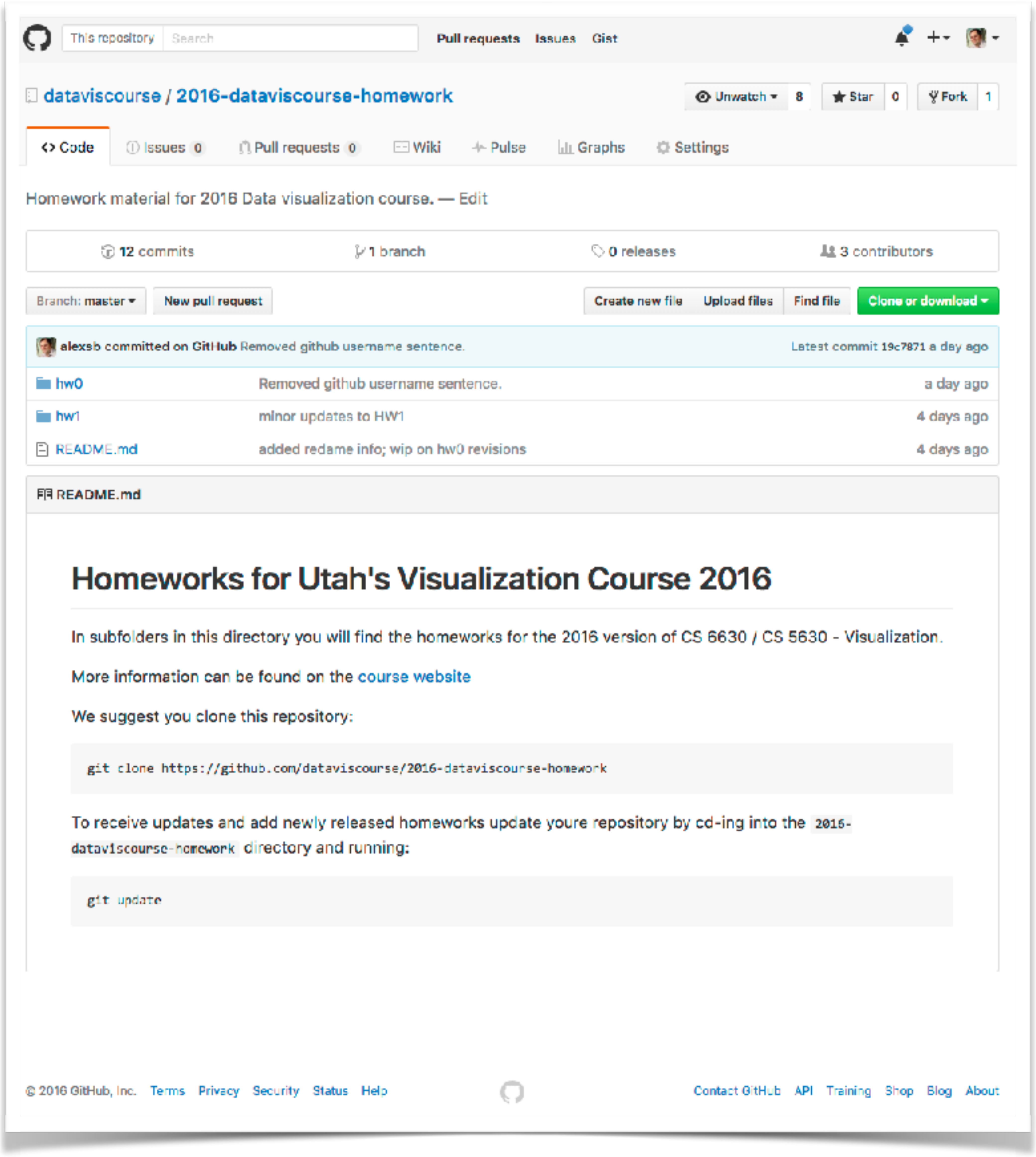
Next Week

HW1 due

Introduction to Git, HTML, CSS

Office hours start!

https://github.com/dataviscourse/2017-dataviscourse-homework/



New Track: Human Centered Computing

REQUIRED COURSES

CS 6540 - HCI (humans + interfaces)

CS 6xxx - Advanced HCI (humans + things)

CS 6630 - Visualization for Data Science (humans + data)

ED PSY 6010: Introduction to Stats and Research Design (methods)

ELECTIVES

Pre-approved course list from within CS and across campus

Up to 3 electives can be taken from outside CS

NON-CS COURSES

Design

- DES 5320 - Typographic Communication
- DES 5370 - Digital Fabrication
- DES 5710 - Product Design and Development

Ed Psych

- ED PSY 6030 - Introduction to Research Design

Psych

- PSY 6120 - Advanced Human Cognition
- PSY 6140 - Cognitive Neuroscience Approaches to Research
- PSY 6420 - Methods in Social Psychology
- PSY 6700 - Neuropsychology

Anthropology

- ANTH 6169 - Ethnographic Methods

Sociology

- SOC 6110 - Methods of Social Research

EAE

- EAE 6900 - Games User Research
- EAE 6900 - A.I. For Games