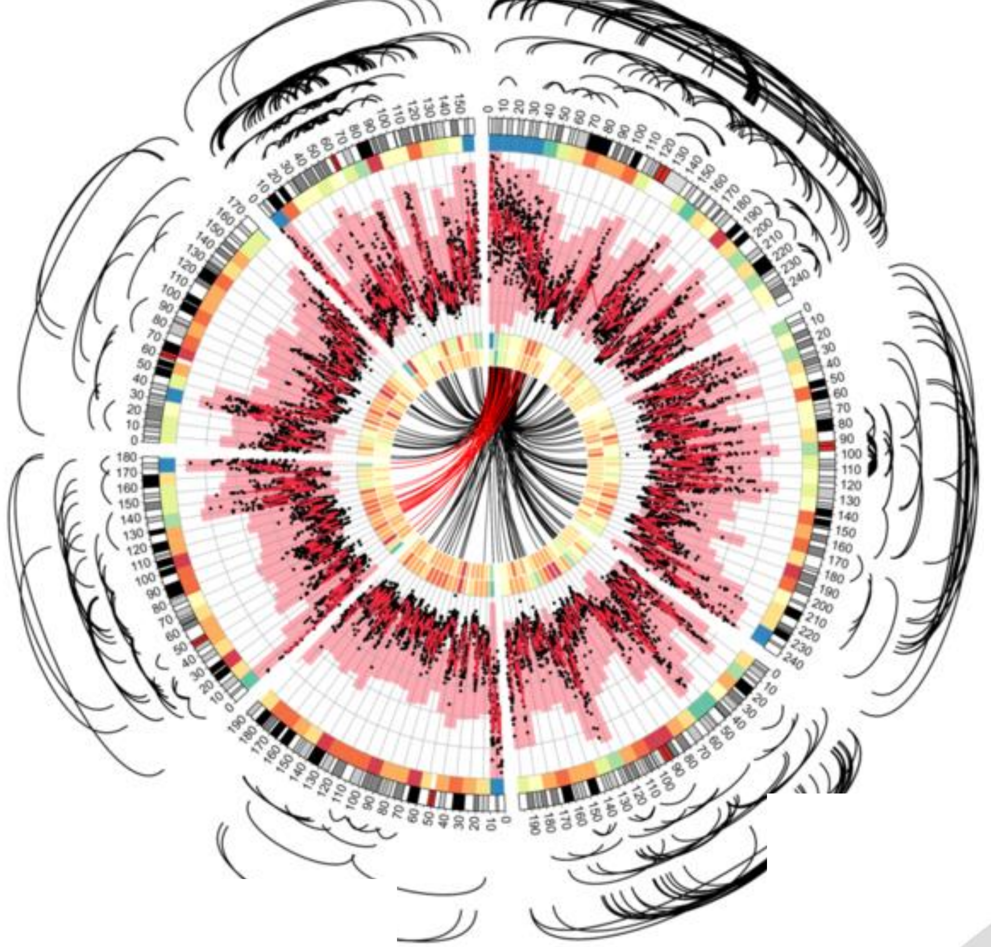
Visualization for Data Science
The Case for Visualizations

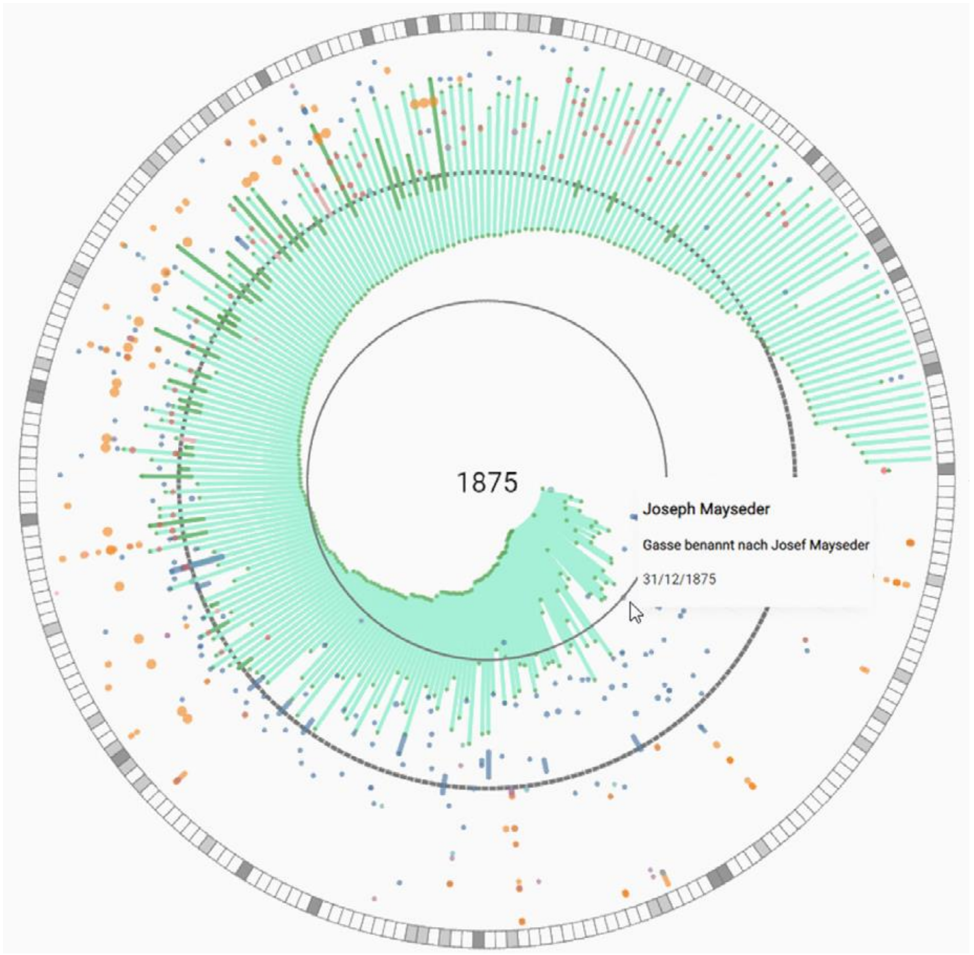
While you wait

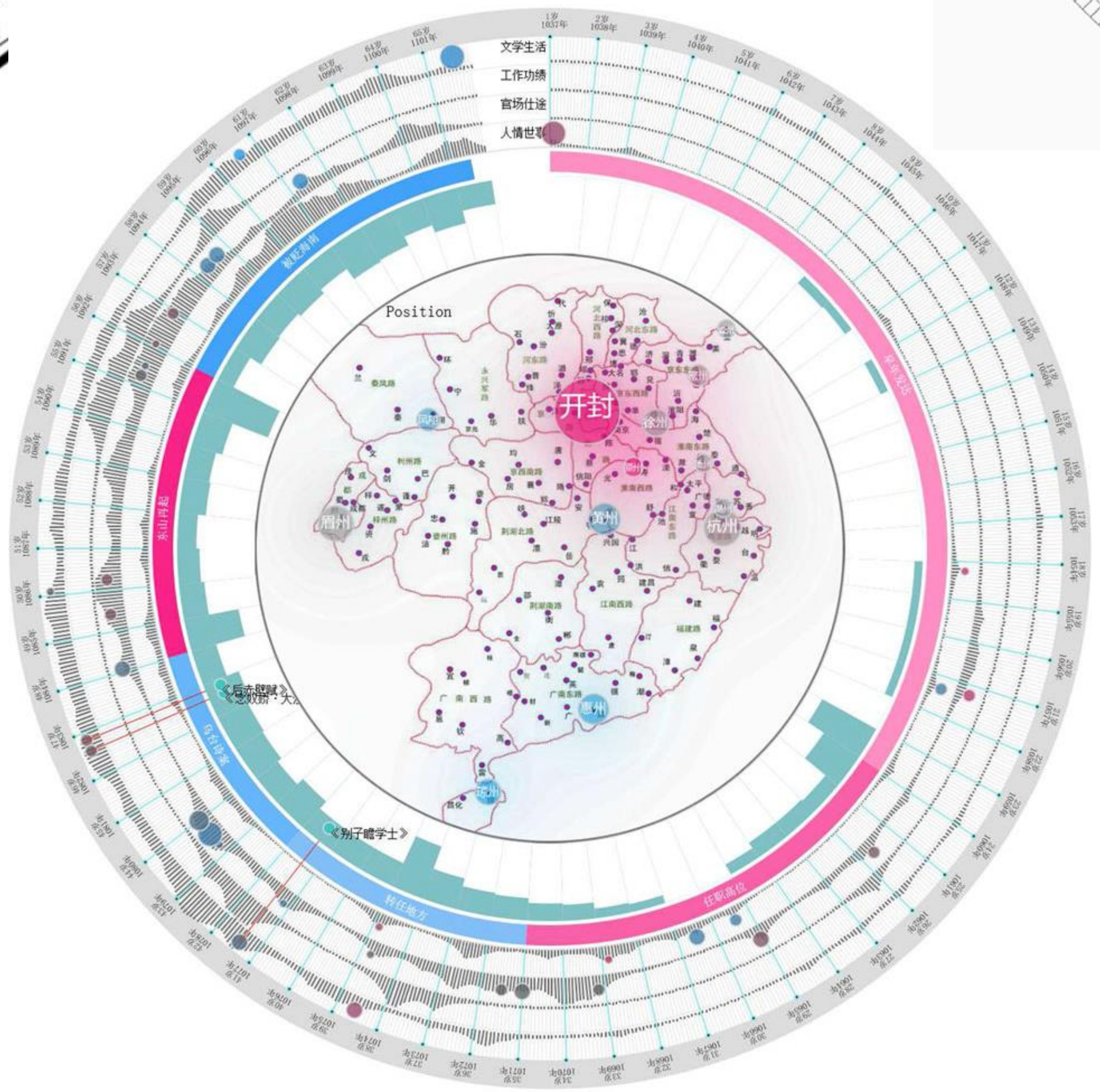
Sign up for

https://edstem.org/us/join/PENpte

with a UBC email address







https://www.ep.liu.se/ecp/034/013/ecp083413.pdf
https://doi.org/10.1101/gr.092759.109
https://doi.org/10.1007/s00371-016-1213-4
https://doi.org/10.1016/j.visinf.2021.01.001

Why should we create/use visualizations?

The purpose of computing is insight, not numbers

Richard Hamming (1973)

The purpose of visualization is insight, not pictures

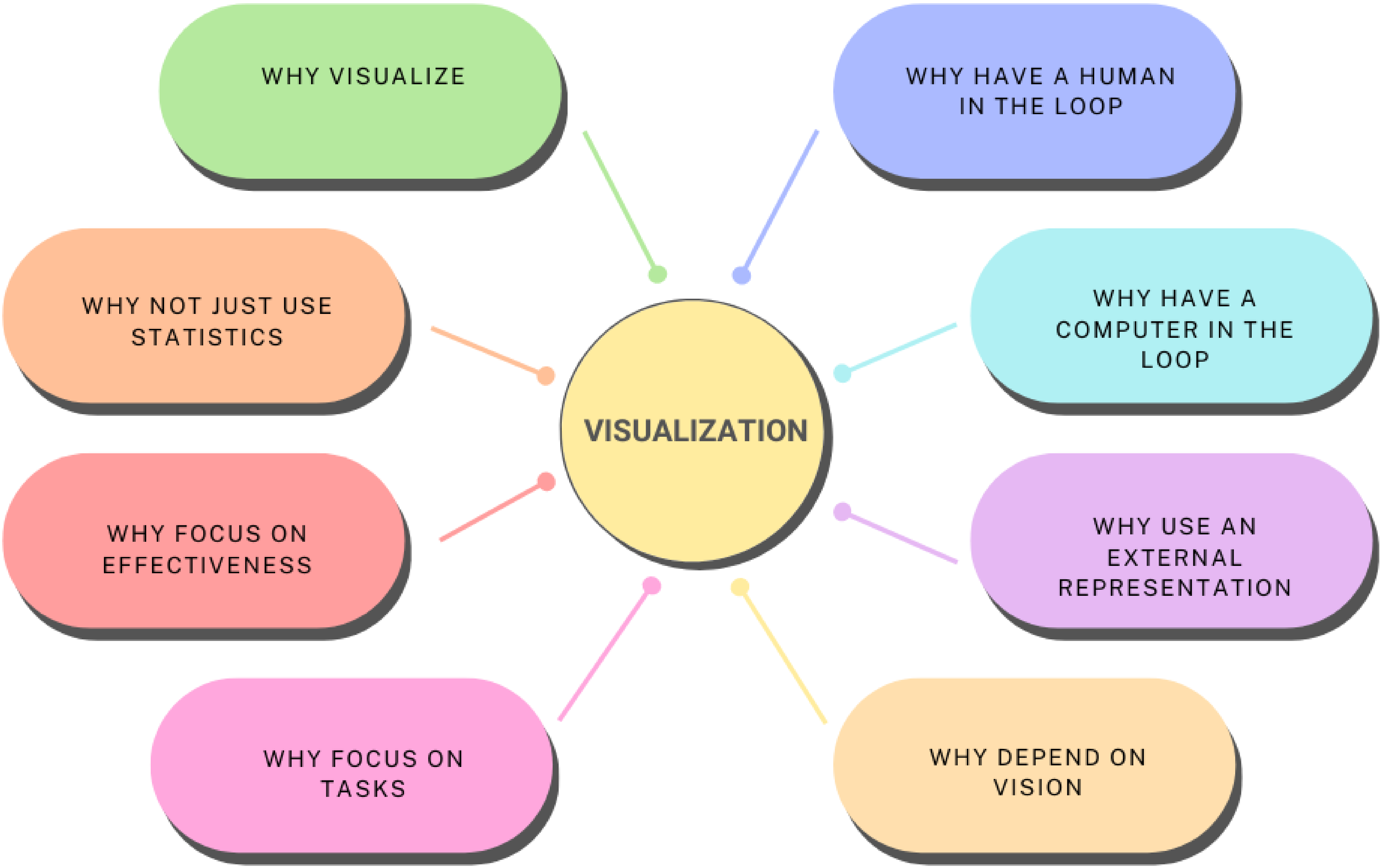
Card, Mackinlay, Shneiderman (1999)

Defining visualization (vis)

Visualization is the process that transforms (abstract) data into interactive graphical representations for the purpose of exploration, confirmation, or presentation. - Alex Lex

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

The visual representation (static or interactive) of data (abstract or spatial) to reinforce human cognition (sensemaking, decision-making, communication, analysis)


VISUALIZATION
WHY VISUALIZE
WHY HAVE A HUMAN IN THE LOOP
WHY NOT JUST USE STATISTICS
WHY HAVE A COMPUTER IN THE LOOP
WHY FOCUS ON EFFECTIVENESS
WHY USE AN EXTERNAL REPRESENTATION
WHY FOCUS ON TASKS
WHY DEPEND ON VISION

Why visualize data?

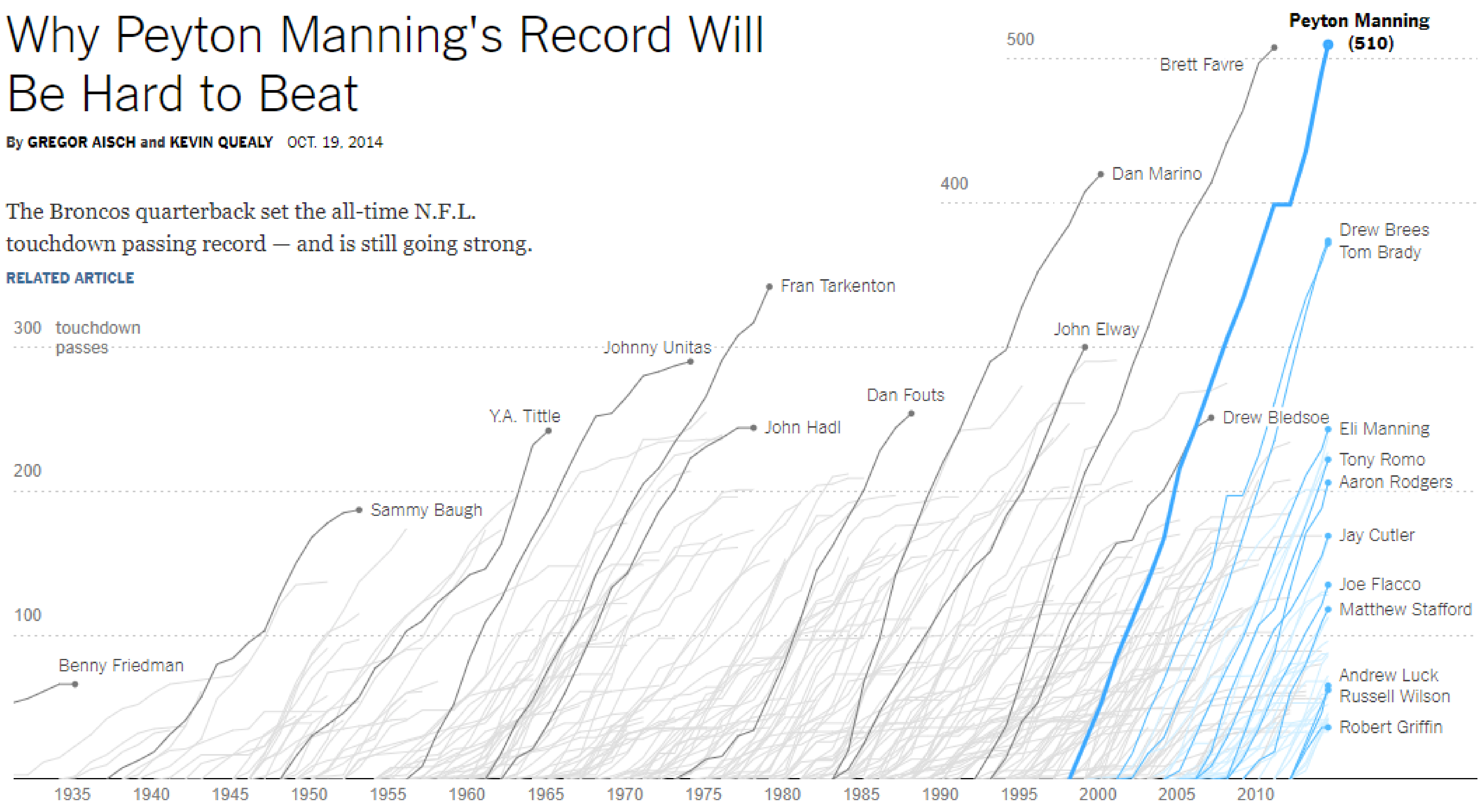
- Inspire
- Communicate
- Clarify
- Record
- Answer questions
- Expand memory
- Generate hypotheses
- Make decisions
- Find patterns

Why visualize?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

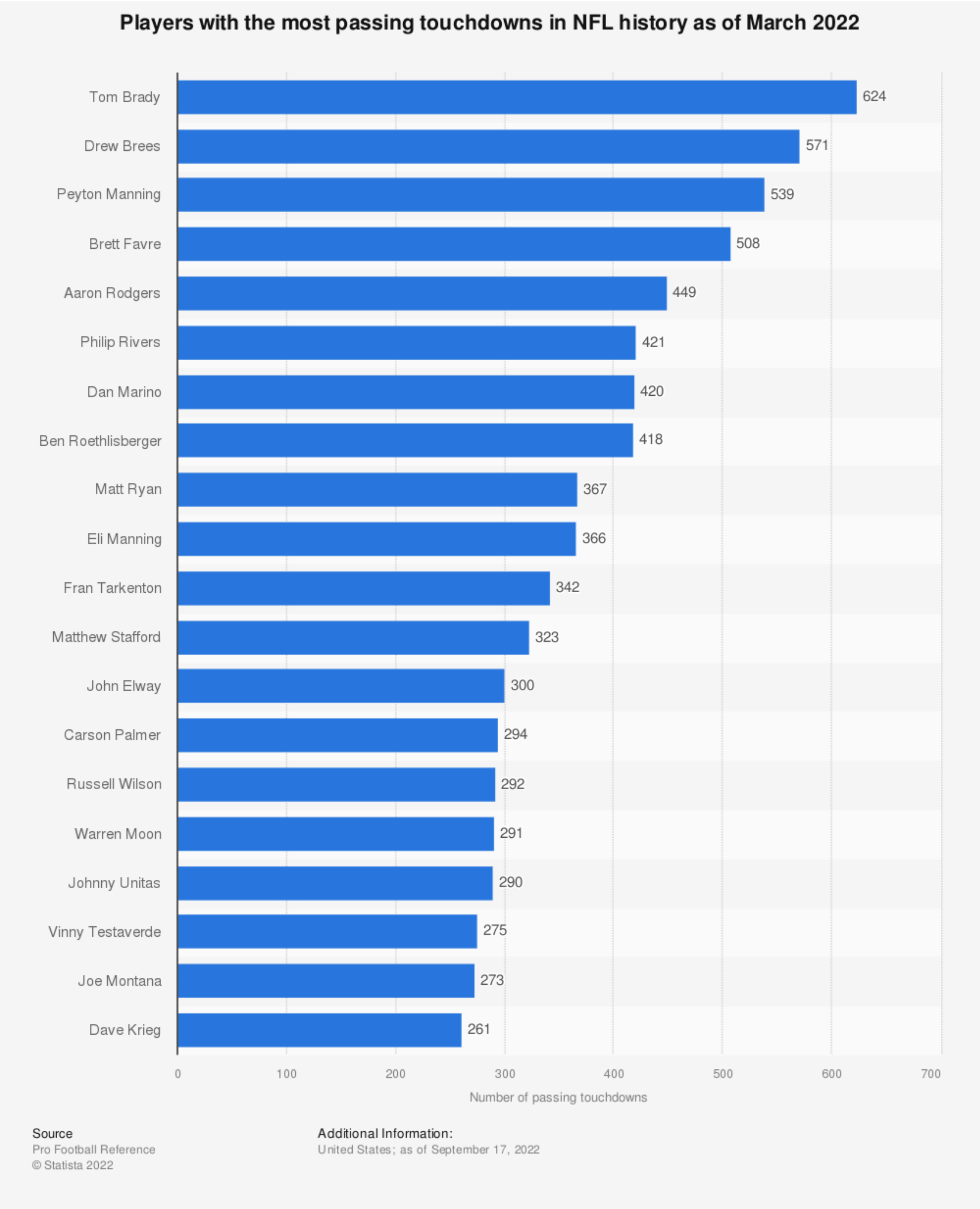
- Communication - To inform humans
- Exploration - When questions are not well defined

Communication Example



https://www.nytimes.com/interactive/2014/10/19/upshot/peyton-manning-breaks-touchdown-passing-record.html

Communication Example



https://www.statista.com/statistics/784370/players-with-the-most-passing-touchdowns-in-nfl-history/

Communication Example

McCarthy Elected House Speaker After 15 Ballots

https://www.nytimes.com/interactive/2023/01/04/us/politics/house-speaker-vote-tally.html

How Every Representative Voted

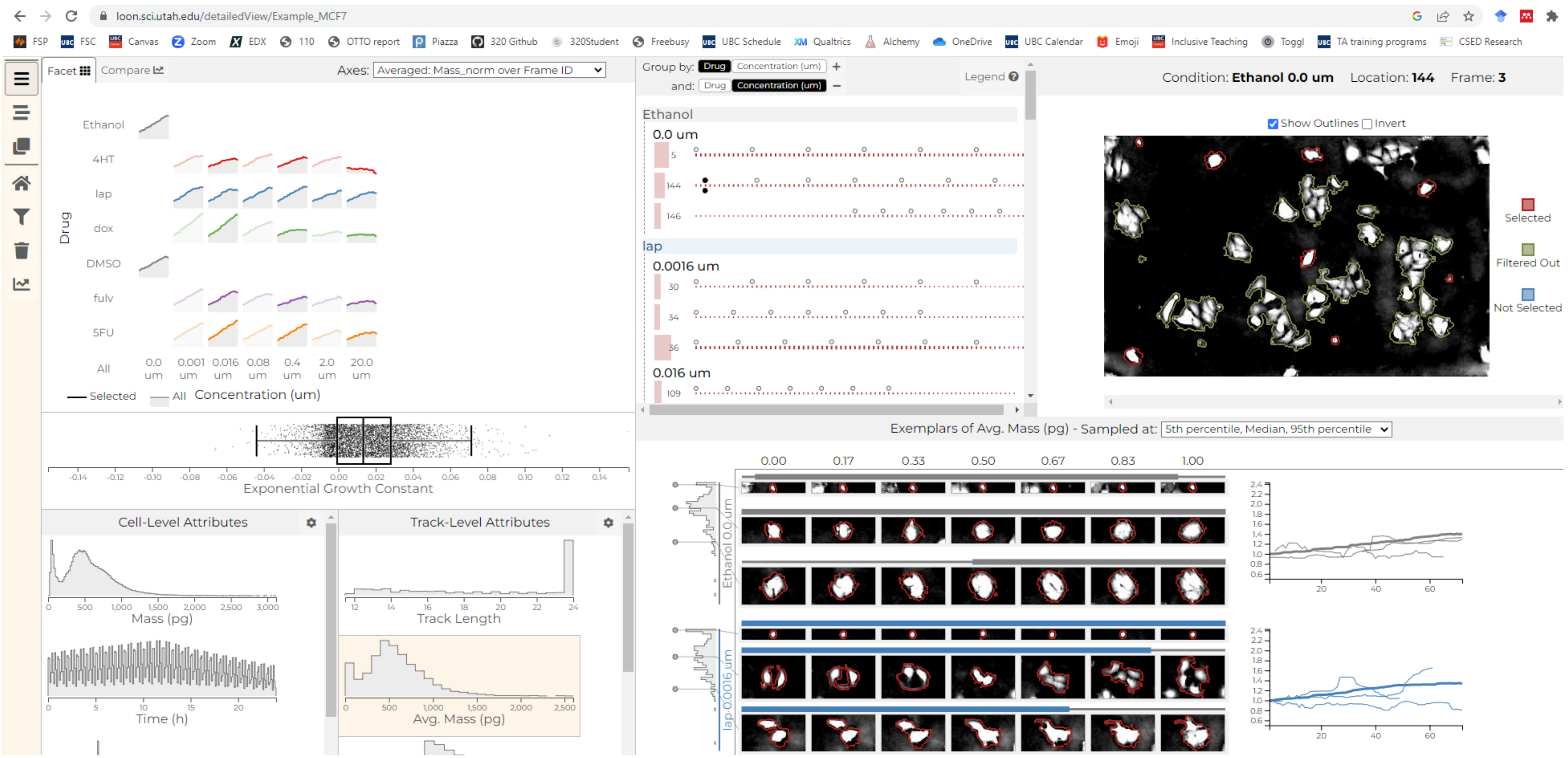
Members voting differently from the majority of their party are highlighted.

Vote key

Biggs · Donalds · Jordan · Trump · Hern · "Present" · Other · McCarthy or Jeffries

Republicans	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Andy Biggs, Ariz. 5th	B	J	J	D	D	D	D	D	D	H	H	J	J	J	P
Lauren Boebert, Colo. 3rd	J	J	J	D	D	D	D	H	H	H	H	H	J	P	P
Eli Crane, Ariz. 2nd	B	J	J	D	D	D	D	D	D	H	H	H	J	B	P
Matt Gaetz, Fla. 1st	B	J	J	D	D	D	T	T	H	H	T	J	J	P	P
Bob Good, Va. 5th	B	J	J	D	D	D	D	D	D	D	H	J	J	J	P
Matt Rosendale, Mont. 2nd	B	J	J	D	D	D	D	D	D	H	H	H	J	B	P
Dan Bishop, N.C. 8th	B	J	J	D	D	D	D	D	D	D	D	M	M	M	M
Josh Brecheen, Okla. 2nd	O	J	J	D	D	D	D	H	H	H	H	M	M	M	M
Michael Cloud, Texas 27th	J	J	J	D	D	D	D	D	D	D	D	M	M	M	M
Andrew Clyde, Ga. 9th	B	J	J	D	D	D	D	D	D	D	D	M	M	M	M
Byron Donalds, Fla. 19th	M	M	J	D	D	D	D	D	D	D	D	M	M	M	M
Paul Gosar, Ariz. 9th	B	J	J	D	D	D	D	D	D	D	D	M	M	M	M
Andy Harris, Md. 1st	O	J	J	D	D	D	D	D	D	H	H	J	M	M	M
Anna Paulina Luna, Fla. 13th	J	J	J	D	D	D	D	D	D	D	D	M	M	M	M

Exploration Example



https://loon.sci.utah.edu/detailedView/Example_MCF7

Communication

Confirmation

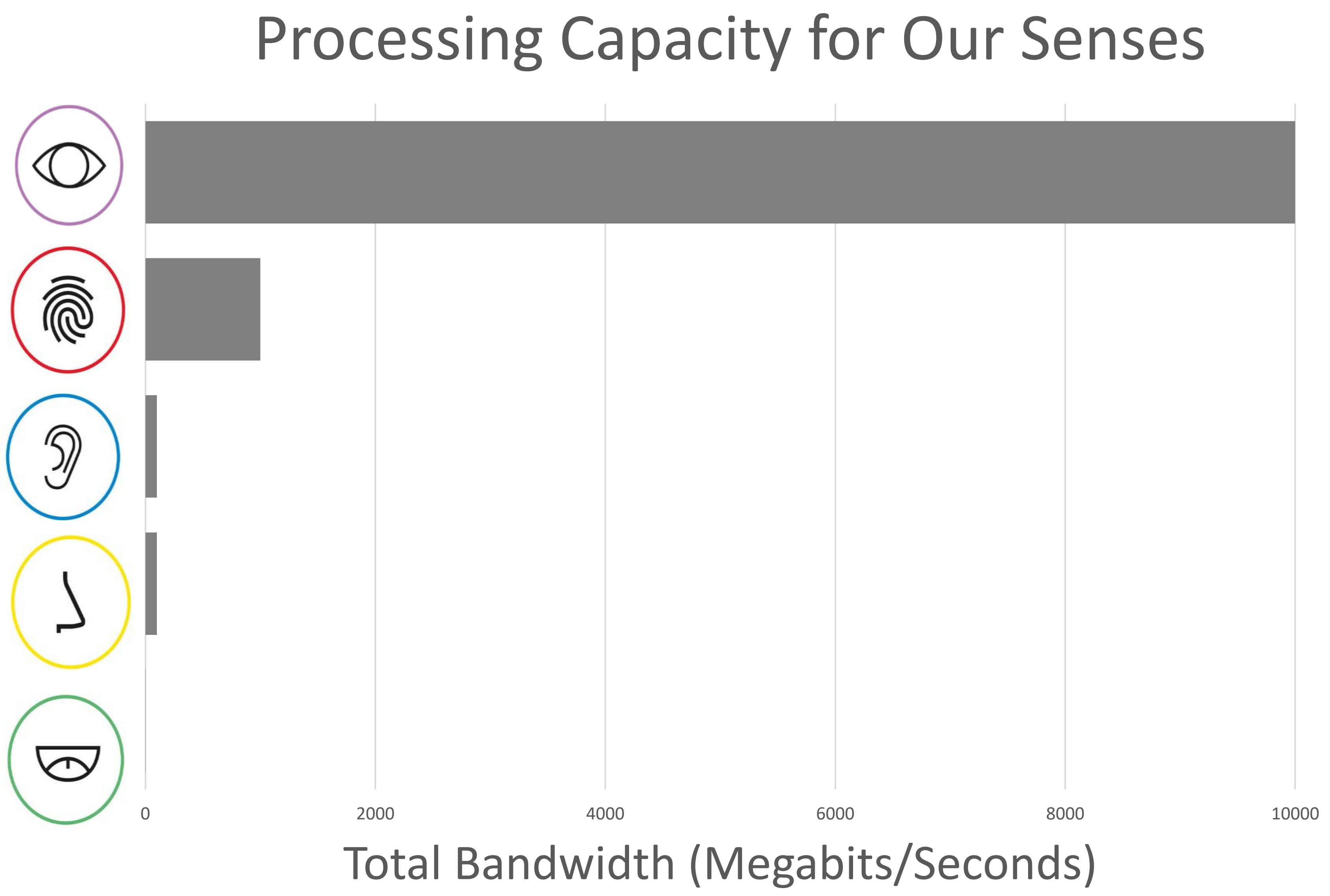
Exploration

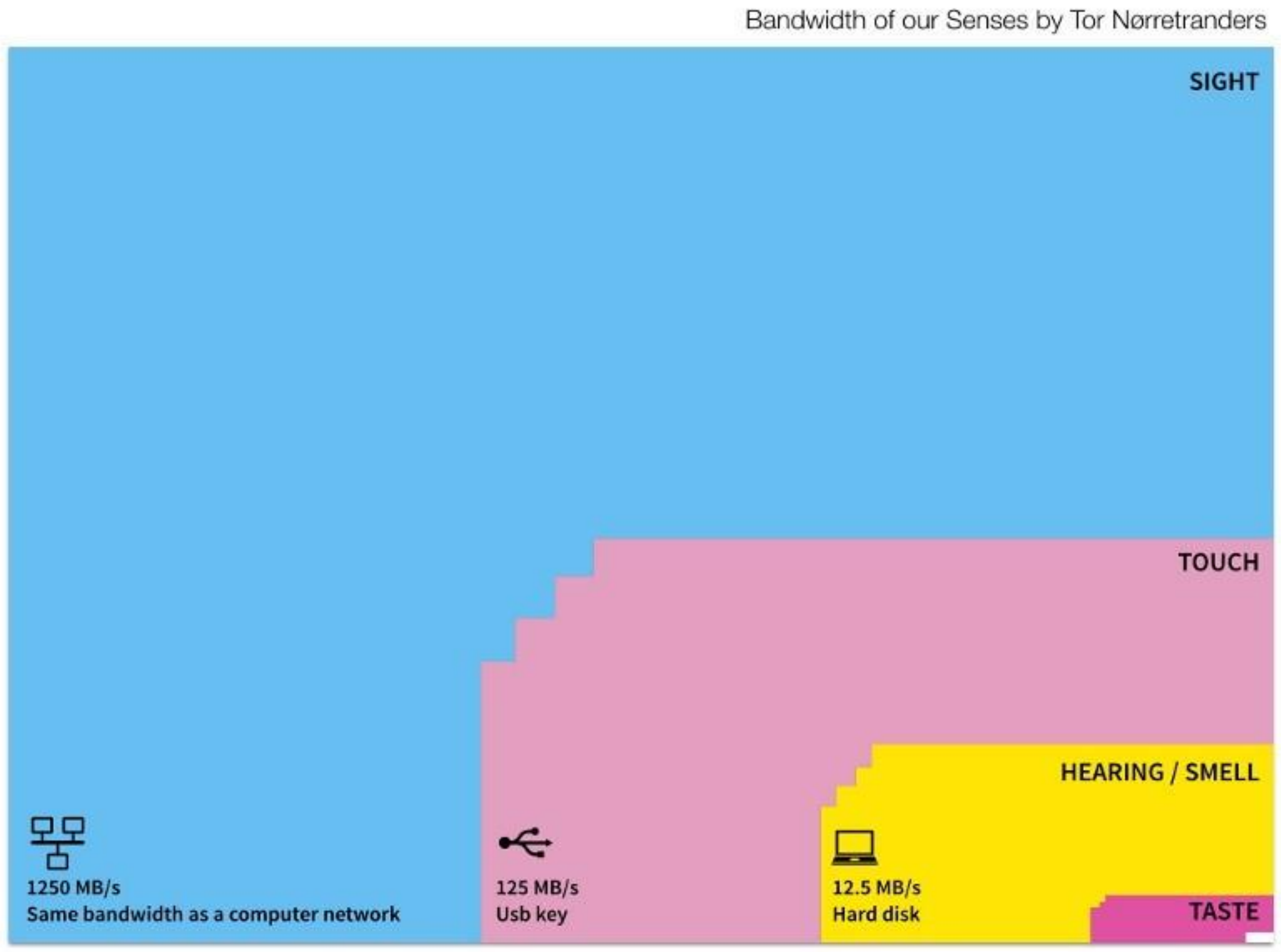
Why depend on vision?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- human visual system is high-bandwidth channel to brain
  - overview possible due to background processing
    - subjective experience of seeing everything simultaneously
    - significant processing occurs in parallel and pre-attentively
- sound: lower bandwidth and different semantics
  - overview not supported
    - subjective experience of sequential stream
- touch/haptics: impoverished record/replay capacity
  - only very low-bandwidth communication thus far
- taste, smell: no viable record/replay devices

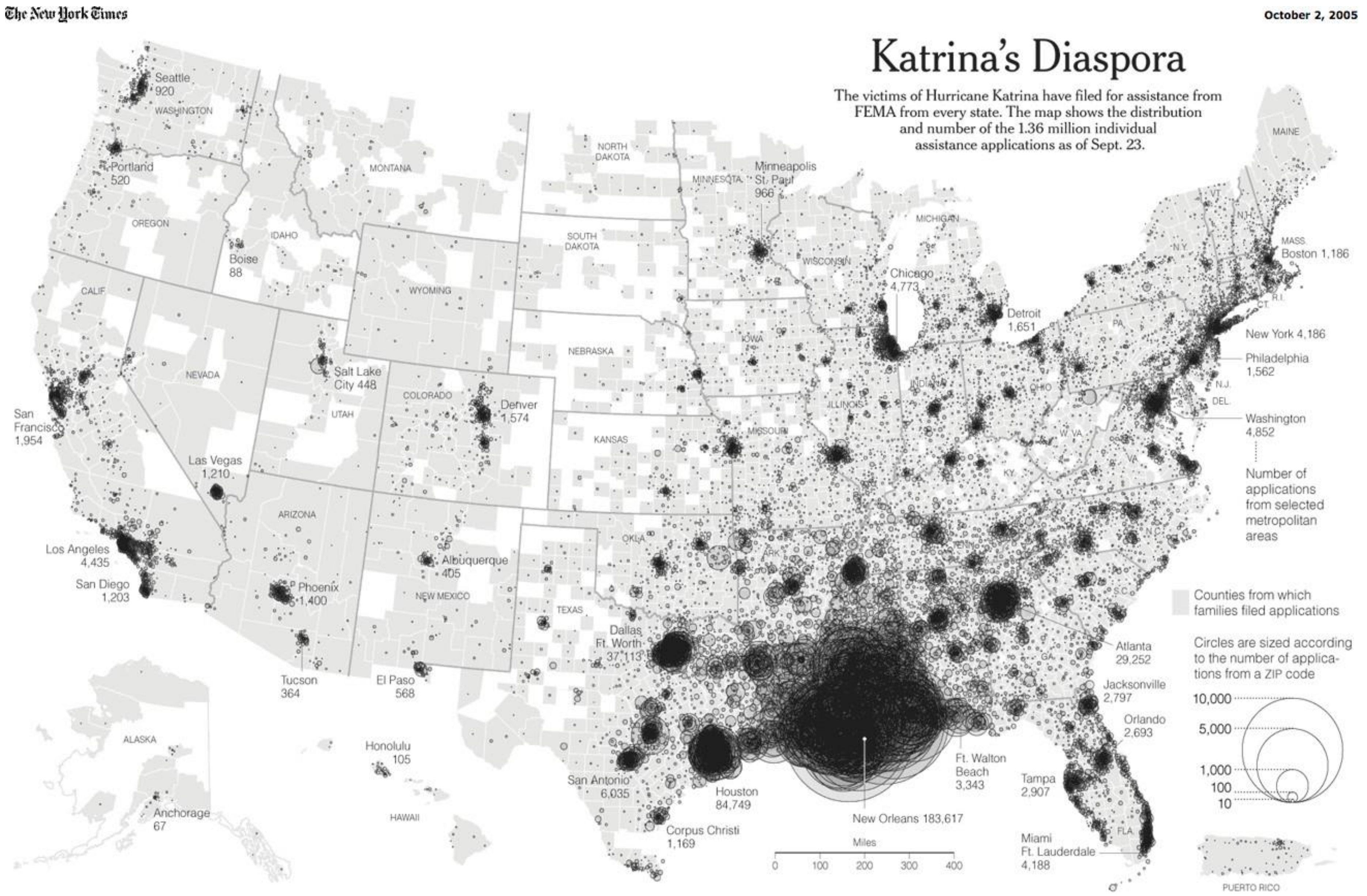
Why visual?


Processing Capacity for Our Senses
0
2000
4000
6000
8000
10000
Total Bandwidth (Megabits/Seconds)


Bandwidth of our Senses by Tor Nørretranders
SIGHT
TOUCH
HEARING / SMELL
TASTE
1250 MB/s
Same bandwidth as a computer network
125 MB/s
Usb key
12.5 MB/s
Hard disk

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

Humans vs. Computers

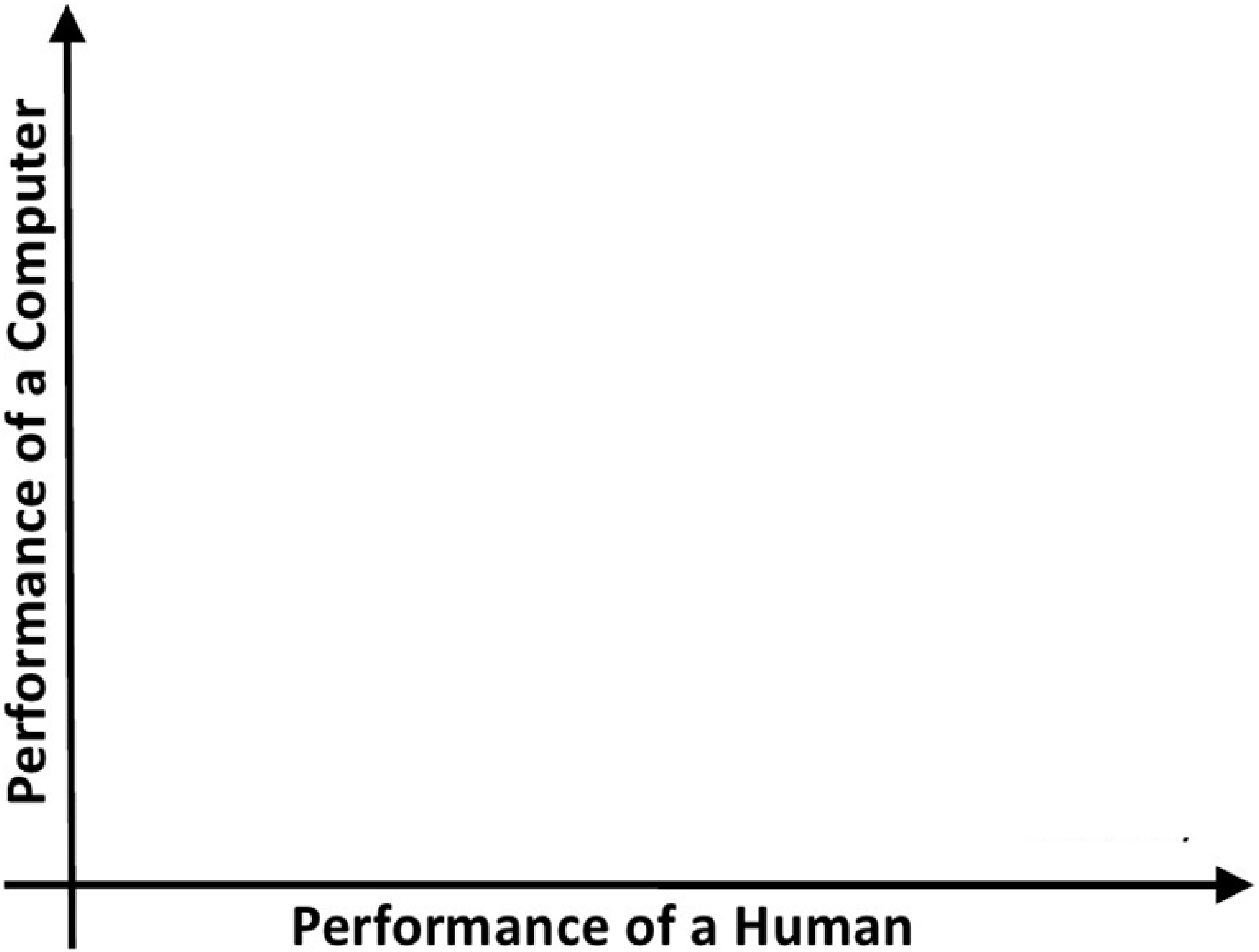
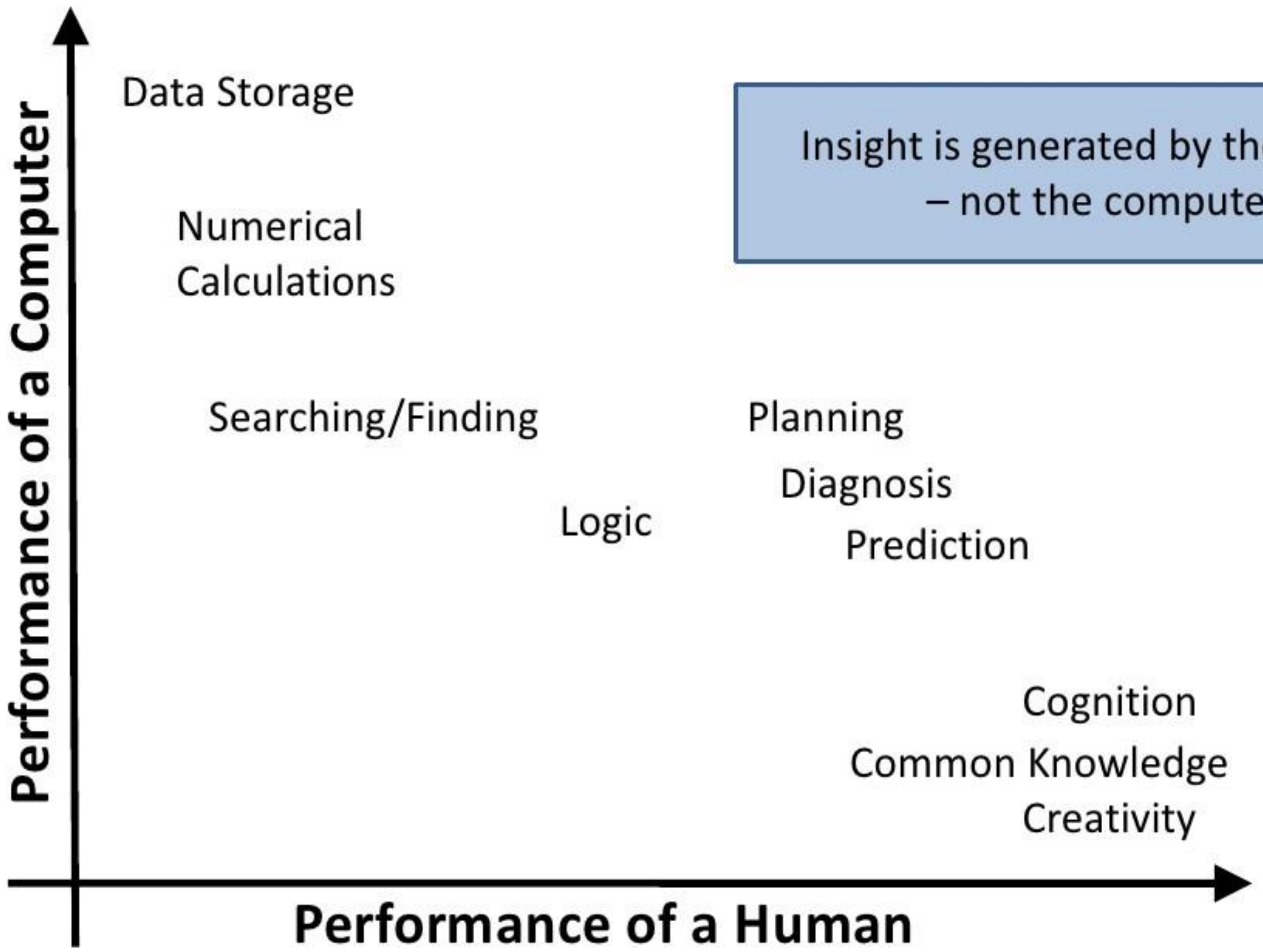


Image from “Superhuman Robots With Artificial Muscles: A Reality”

Humans vs. Computers



Insight is generated by the human
– not the computer!

Image from "Superhuman Robots With Artificial Muscles: A Reality"

When to automate? (when not to visualize?)

Don't need vis when fully automatic solution exists and is trusted

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*

Why have a human in the loop?

Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.

- many analysis problems ill-specified
  - don’t know exactly what questions to ask in advance
- possibilities
  - long-term use for end users (ex: exploratory analysis of scientific data)
  - presentation of known results (ex: New York Times Upshot)
  - stepping stone to assess requirements before developing models
  - help automatic solution developers refine & determine parameters
  - help end users of automatic solutions verify, build trust

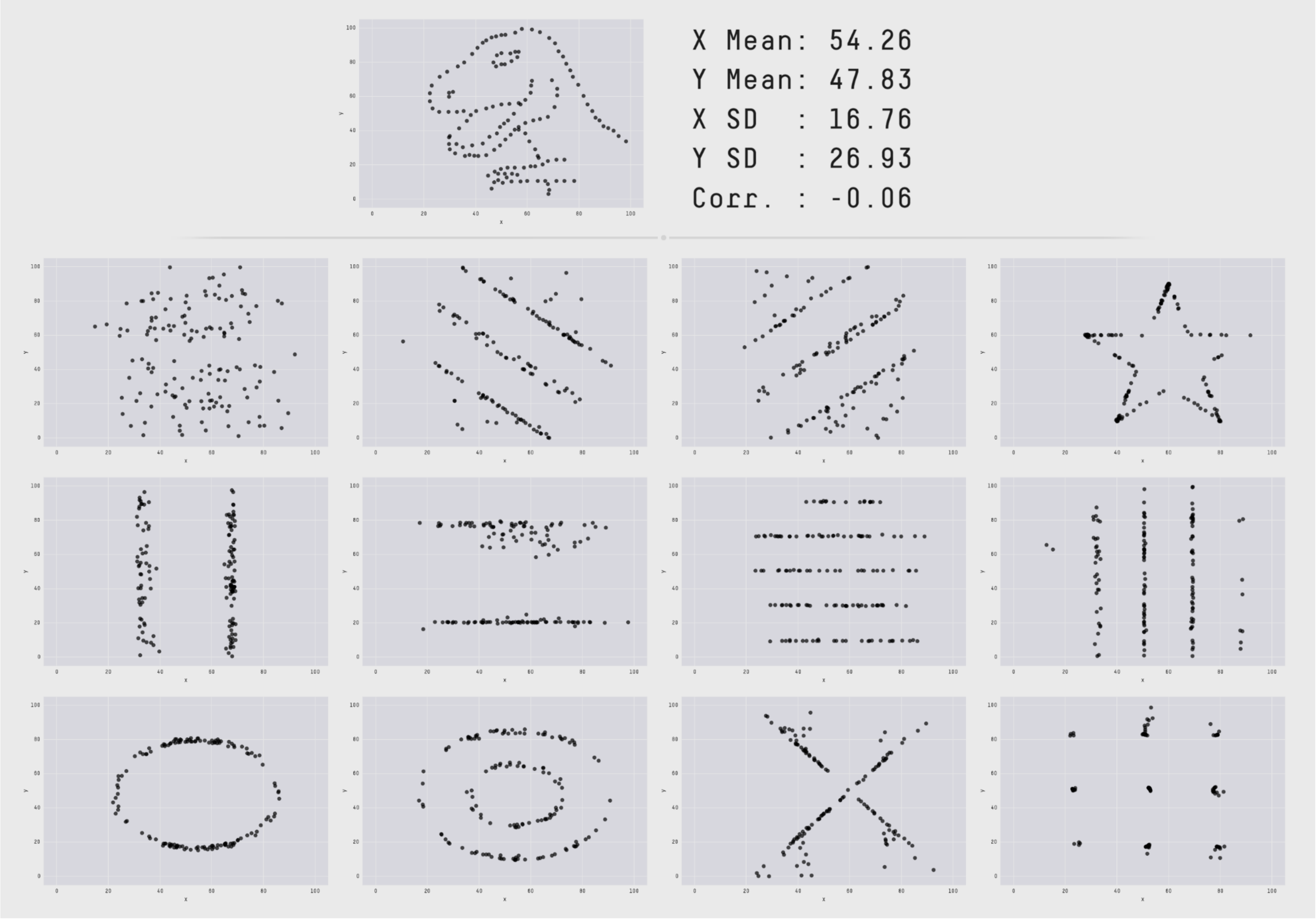
Why use computerized external representation?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

- External representation: replace cognition with perception
- Interaction – allows us to interact with the data
- Integration – integrate with algorithms
- Make visualization part of a data analysis pipeline
- Efficiency- re-use charts/methods for different datasets
- Quality- precise data driven rendering
- Storytelling – use time

Why not just use statistics?

Anscombe's Quartet



https://www.youtube.com/watch?v=DbJyPELmhJc

https://www.autodesk.com/research/publications/same-stats-different-graphs

https://www.youtube.com/watch?v=RbHCeANCbW0

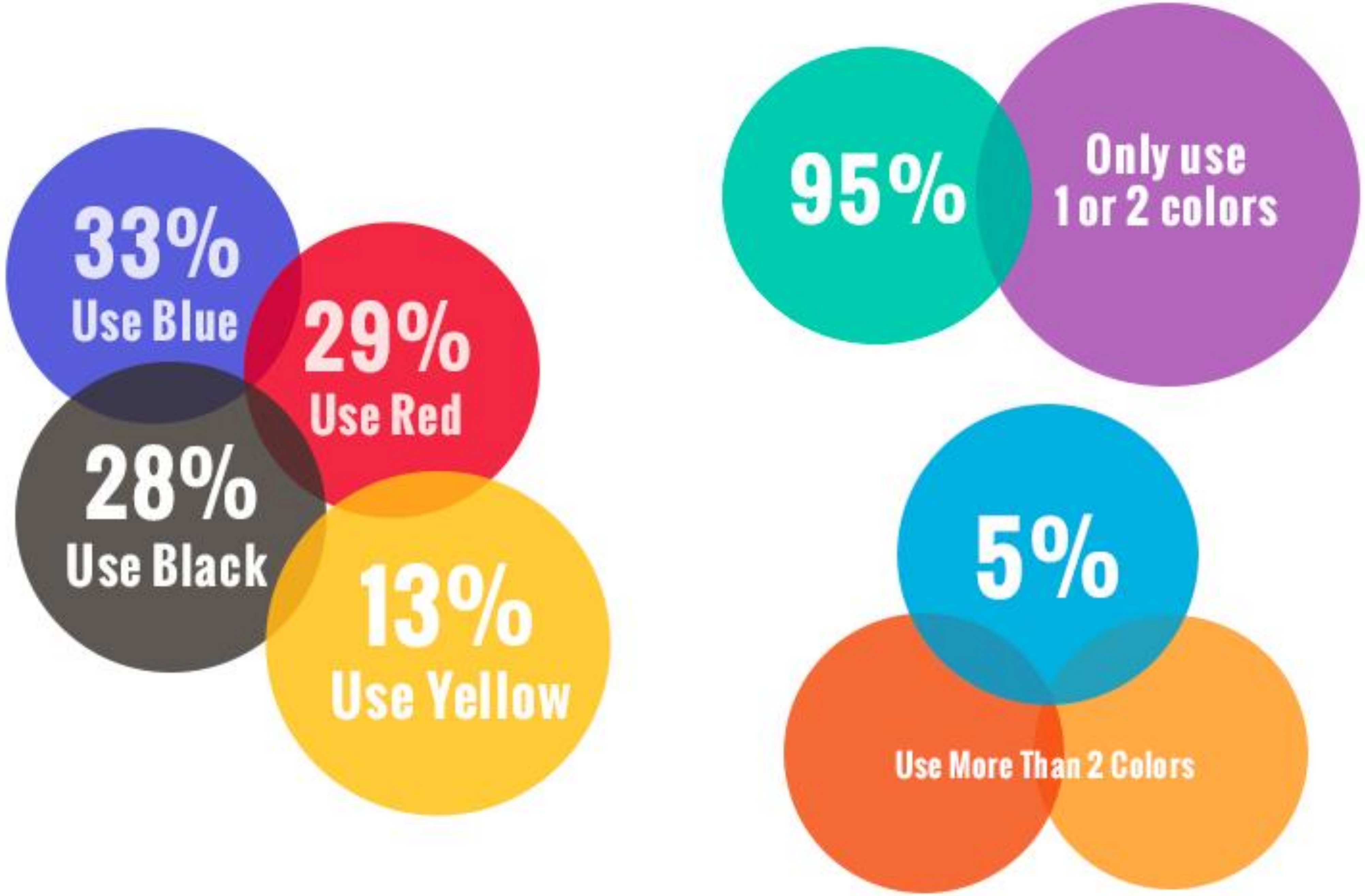
Why Focus on Tasks?

Human-computer interaction is a sub-domain of visualization that focuses on the design, evaluation and implementation of computing systems for human use.

- Users
- Tasks
- Contexts

COLOR TRENDS

THE MOST-USED COLORS OF THE WORLD'S TOP BRANDS



Why Focus on Effectiveness?

- Correctness
- Accuracy

What resource limitations are we faced with?

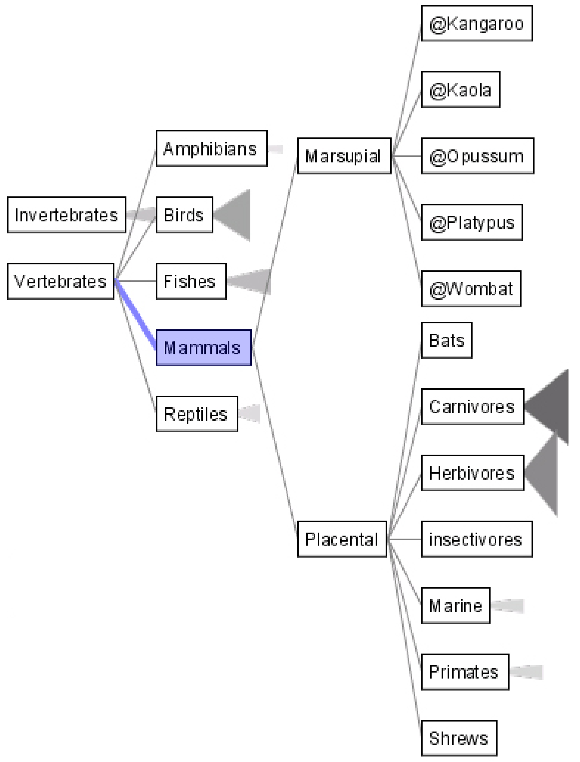
Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays.

- computational limits
  - computation time, system memory
- display limits
  - pixels are precious & most constrained resource
  - **information density**: ratio of space used to encode info vs unused whitespace
    - tradeoff between clutter and wasting space
    - find sweet spot between dense and sparse
- human limits
  - human time, human memory, human attention

Why analyze?

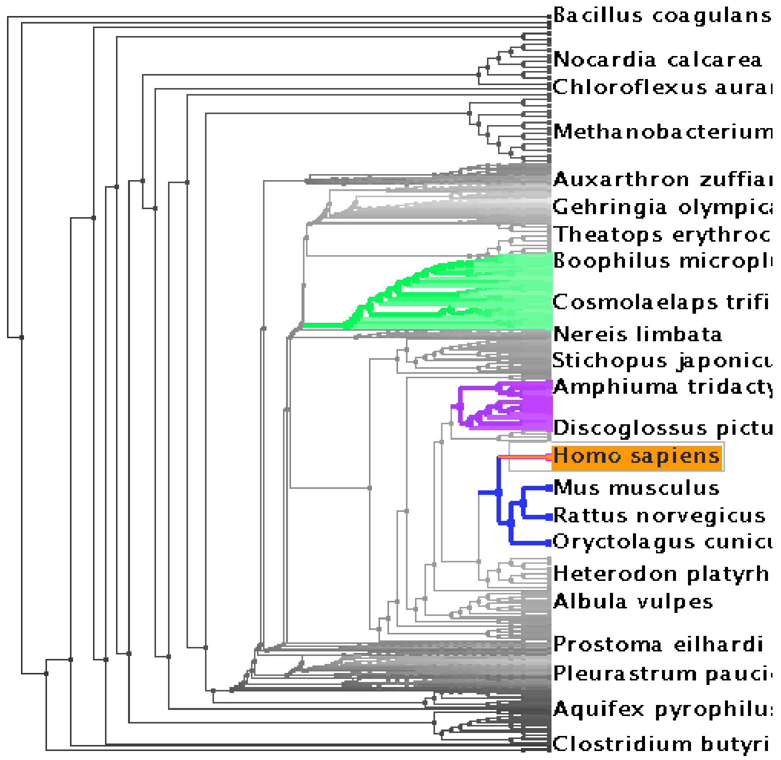
- imposes structure on huge design space
  - scaffold to help you think systematically about choices
  - analyzing existing as stepping stone to designing new
  - most possibilities ineffective for particular task/data combination

SpaceTree

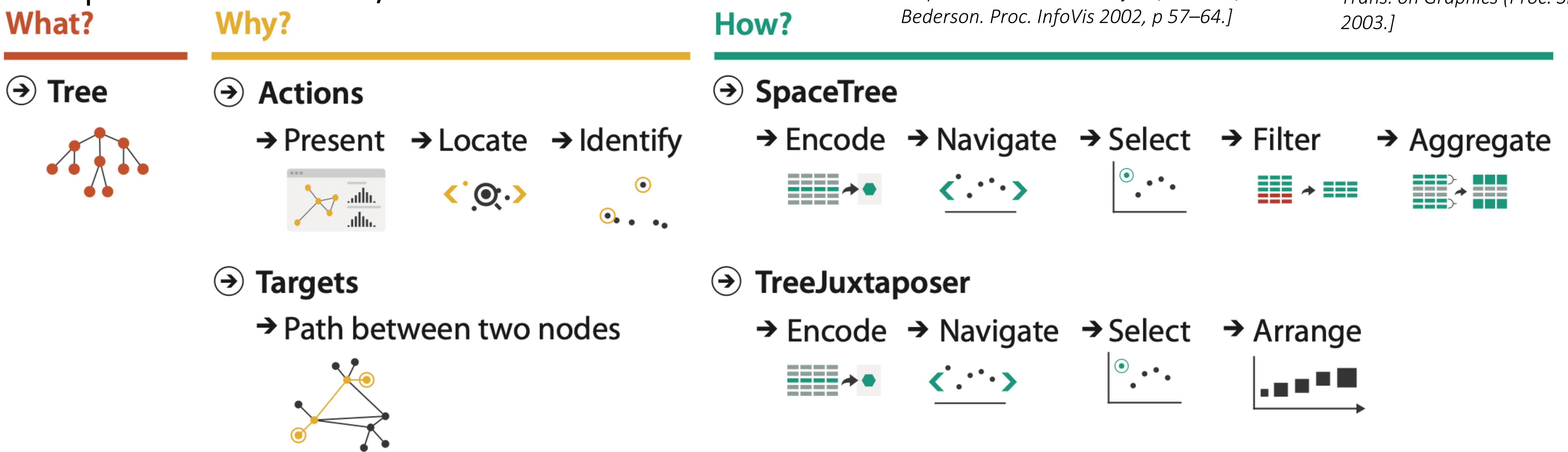


[SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation. Grosjean, Plaisant, and Bederson. Proc. InfoVis 2002, p 57–64.]

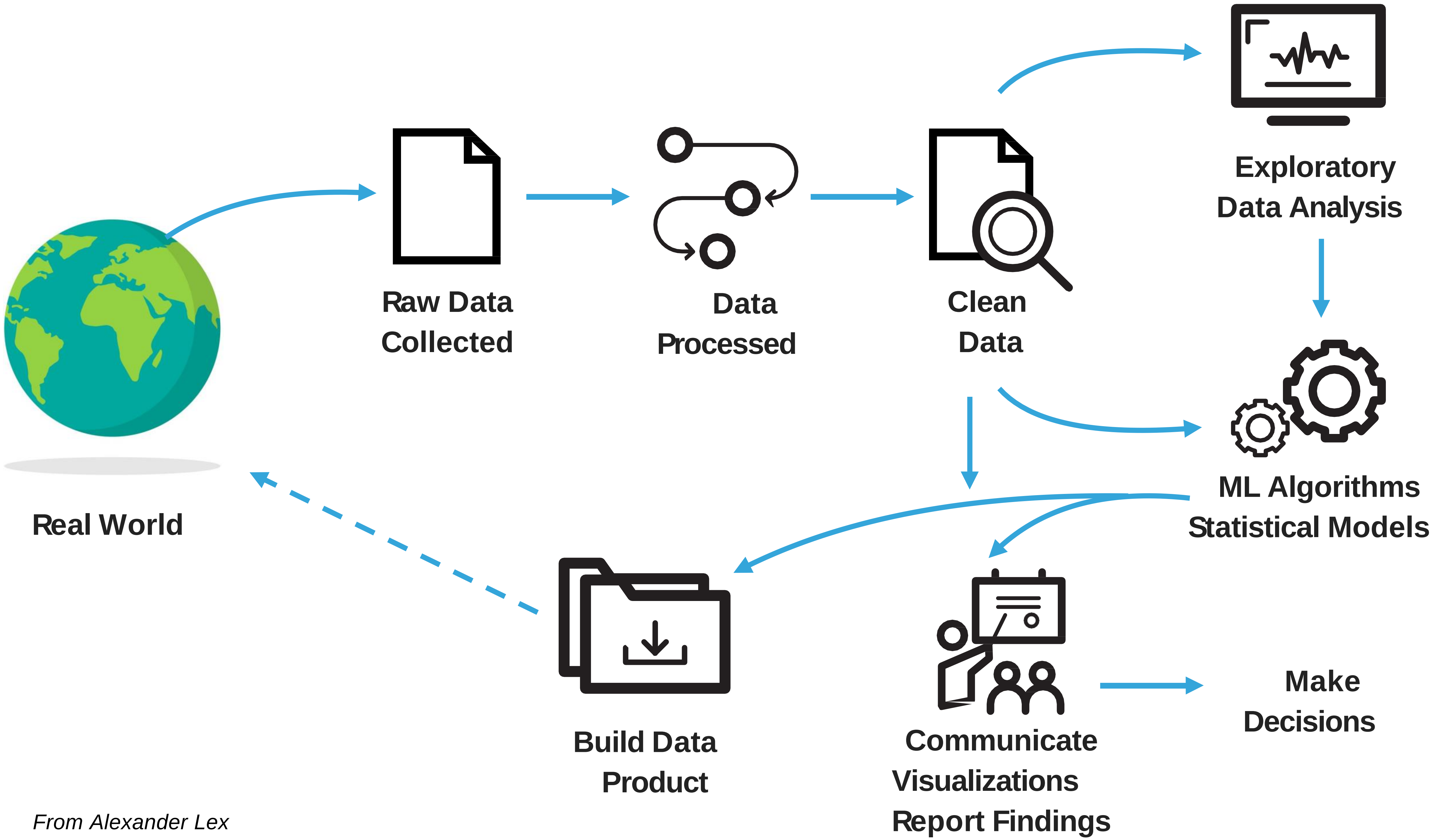
TreeJuxtaposer



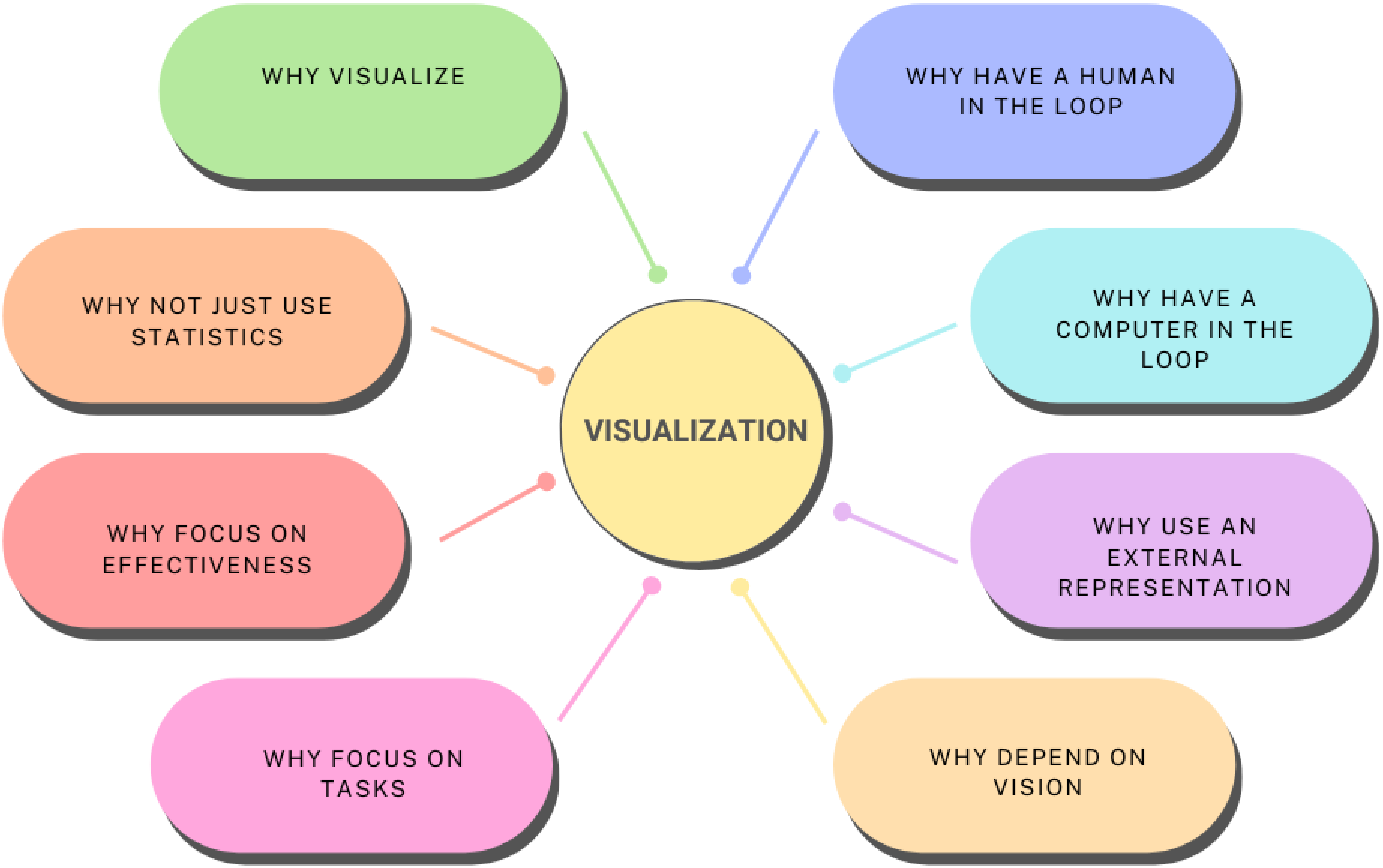
[TreeJuxtaposer: Scalable Tree Comparison Using Focus+Context With Guaranteed Visibility. ACM Trans. on Graphics (Proc. SIGGRAPH) 22:453– 462, 2003.]



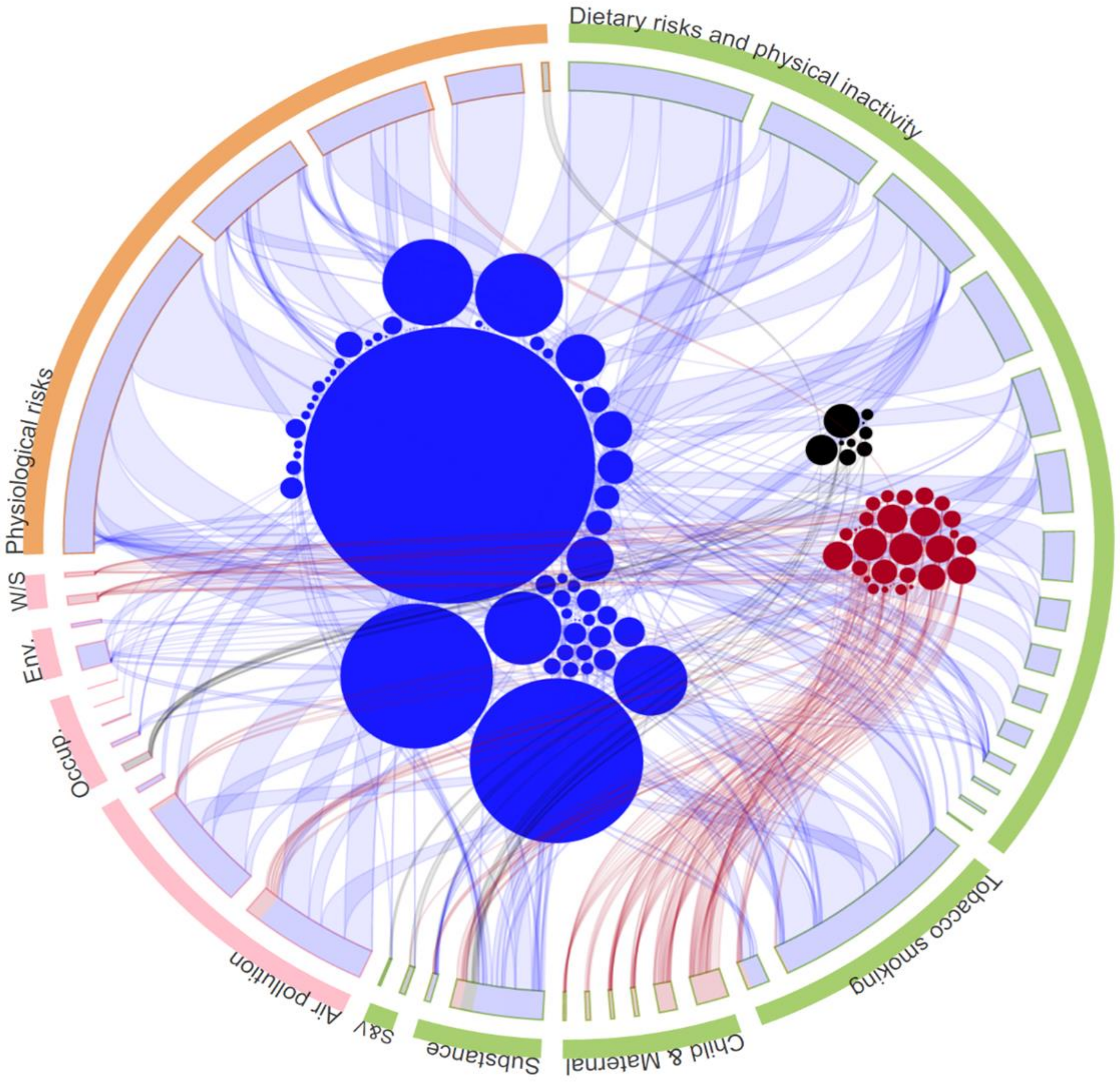
What's the Role of Visualization in Data Science?



From Alexander Lex


WHY VISUALIZE
WHY HAVE A HUMAN IN THE LOOP
WHY NOT JUST USE STATISTICS
VISUALIZATION
WHY HAVE A COMPUTER IN THE LOOP
WHY FOCUS ON EFFECTIVENESS
WHY USE AN EXTERNAL REPRESENTATION
WHY FOCUS ON TASKS
WHY DEPEND ON VISION

Dr. Oluwakemi Ola (she/her)   [Dr. K.]
Assistant Professor of Teaching
Computer Science Department



Teaching
- Computational Thinking (100)
- Computation, Programs & Programming (110)
- Introduction to Computer Systems (213)
- Visualization for Data Science (DSCI 320)

Research Focus
- Socio-emotional Learning
- CS Pedagogy
- Information Visualization

Objectives

- Use the Vega-lite visualization grammar to categorize, describe, design, and create existing and novel visualizations.
- Use Python to create reproducible visualizations.
- Select and tailor the appropriate plot for the given dataset, audience, and data analytical question at hand.
- Compare methods for visualizing data and explain the tradeoffs of different approaches.
- Evaluate visualizations based on principles of effective design.
- Using a structured approach, design and implement effective visualizations for exploration and communication.

Components

- **Lectures** will be a mix of instructional periods and hands-on activities that allow students to confront their misconceptions about visual representations; practice skills they have been exposed to in pre-lecture videos/readings, and develop new knowledge about visualizations. Attendance is required.

- **Labs** support programming assignments, focus on technical content, and allow students to engage in activities to build their programming skills. You will attend a mandatory in-person programming lab every Thursday. There are no solutions to the exercises completed during lab, so your active participation is critical to your success. Attendance is required.

- **Unsupervised & Asynchronous Learning:** You will engage in unsupervised learning activities each week in preparation for class. This includes re-lecture readings/videos, and programming tutorials. You will also engage in non-contact, low stakes engagement activities (i.e., active course engagement) outside of class.

Tentative Schedule

Wk	Monday	Tuesday	Wednesday	Thursday	Assignments
9-Jan	Case for Viz	UNLAB	Data	LAB – Jumpstart and T1	Jan 15 - A1 - Catching Pythons & Finding Pandas
16-Jan	Visual Marks and Channels II	UNLAB - Work through T2 & T3	Visual Marks and Channels II	Lab 1 - Scatter Charts	
23-Jan	Perception & Cognition I	UNLAB - Catch up T1 - T4	Bar Galore (T5)	Lab 2 - Bar Charts	Jan 29: A2 - Into the Critique of It
30-Jan	Visualizing Temporal Data (T6)		Perception & Cognition II	Lab 3 – line chart	
6-Feb	Design Principles		T7 –Data Transformation	Slack	Feb 12: A3 - Visualizing Data Breaches
13-Feb	EDA & Task Abstraction		T8 – Distributions	Lab 4 – Distributions	Feb 26 A4 Design in Cycles
Feb 20- Break					

Weekly Rhythm

- **Sunday (1hr)** Prior to lecture: Read the provided materials to get an initial exposure to the theoretical concepts that will be emphasized that week.
- **Monday (1.5hr)** Lecture
- **Tuesday (1hr) UNLAB -** Work through the tutorial in the Viz Curriculum modules to be exposed to the technical aspects of visualization development.
- **Wednesday (1.5hr)** Lecture
- **Thursday (1.5hr)** Lab
- **Unscheduled Activities (~4.5hrs)** In addition to the 6.5 hours itemized above, we expect students to spend an additional 4.5 hours working on assigned deliverables.

First week will be more

Assessment

Activity	Weight	Policy
Assignments	30%	5 deliverables (weights increase with difficulty)
Labs	12%	Due at the end of the lab time
Project	18%	3 deliverables, must work with a partner
Engagement	10%	Active course engagement (lecture, discussion forum, survey, etc)
Final Exam	30%	Covers all aspects of the course. Date determined by the University

In order to pass the course, you must achieve a passing grade in the mini-project and final exam components of the course.

Resources

Visualization analysis and design: Prof. Munzner textbook, can be downloaded through the UBC online library

Important Links

Lecture Participation: iClicker Cloud

Discussion Forum: EdSTEM

Grades & Quizzes: Canvas

Most Submissions: Gradescope

Deinde (Coming up next)

- Read through the Syllabus
- Buy the required supplies
  - At least five different colors of pens—e.g., Pilot G2 packs 31128 or pencils
  - Plenty of white or near-white drawing paper
- Work through the JumpStart
  - Setup
  - Introduction to JupyterLab
  - Python Crash Course