Data Visualization for Biomedical Applications

Lecture 1

BMI706 - 22 March 2018

Nils Gehlenborg, PhD

Lecture 1: Learning Goals

- What are data visualization & course about?
- What are the building blocks of data visualization?
- Preview of Plotly Library

What is data visualization?

What is data visualization?

> The use of computer-supported, interactive, visual representations of data to amplify cognition.
>
> —*Stu Card, Jock Mackinlay & Ben Shneiderman*

> Computer-based visualization systems provide visual representations of datasets intended to help people carry out some task more effectively.
>
> —*Tamara Munzner*

What is data visualization?

The use of computer-supported, interactive, visual representations of data to amplify cognition.

—*Stu Card, Jock Mackinlay & Ben Shneiderman*

Computer-based visualization systems provide visual representations of datasets intended to help **people** carry out some task **more effectively**.

—*Tamara Munzner*

What is data visualization?

Human | **Data**

Visualization

What is data visualization?

The purpose of computing is insight, not numbers.

—*Richard Hamming*

The purpose of **visualization** is insight, **not pictures**.

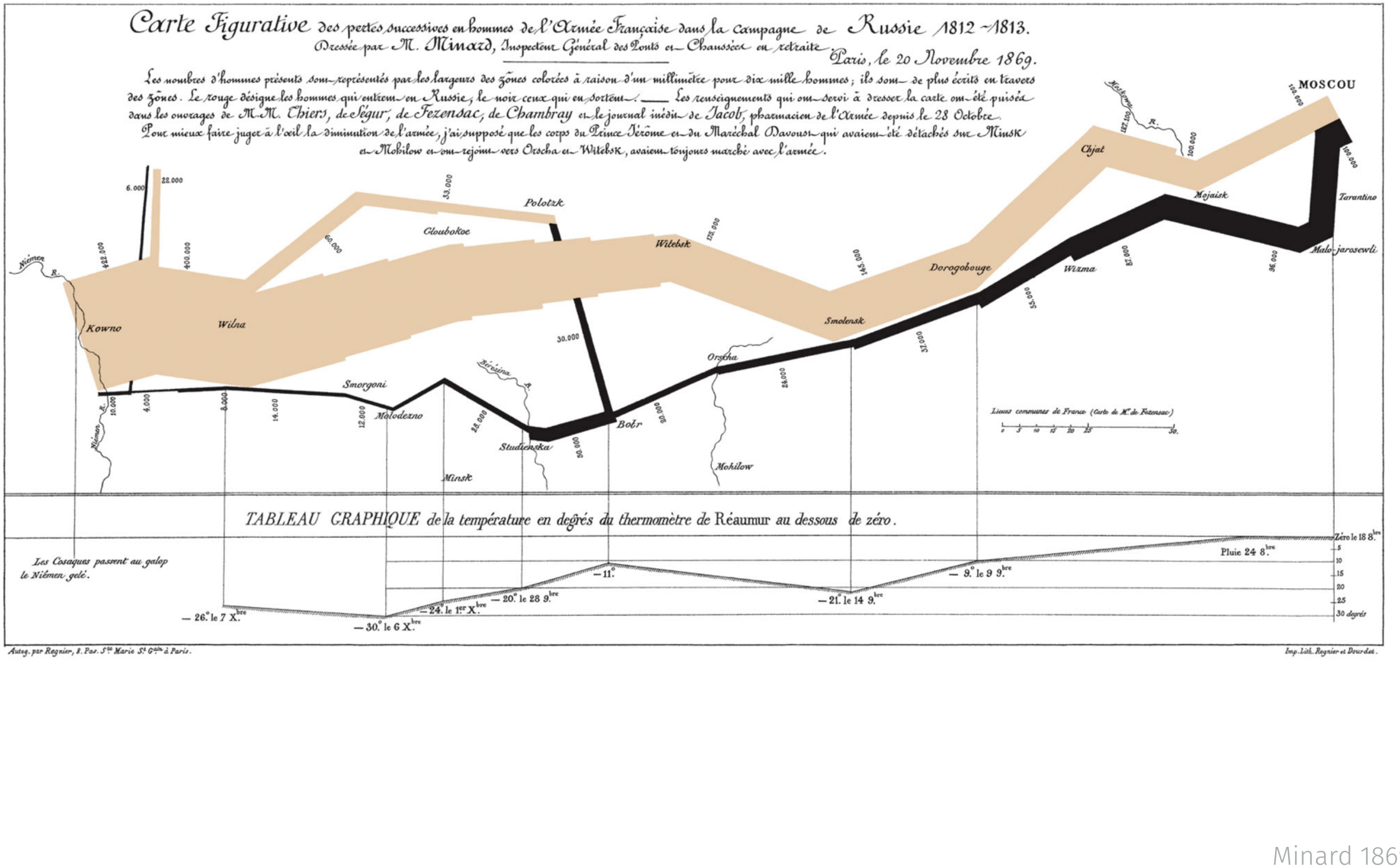
—*Stu Card, Jock Mackinlay & Ben Shneiderman*

Why do we need it?

A good sketch is better than a long speech.

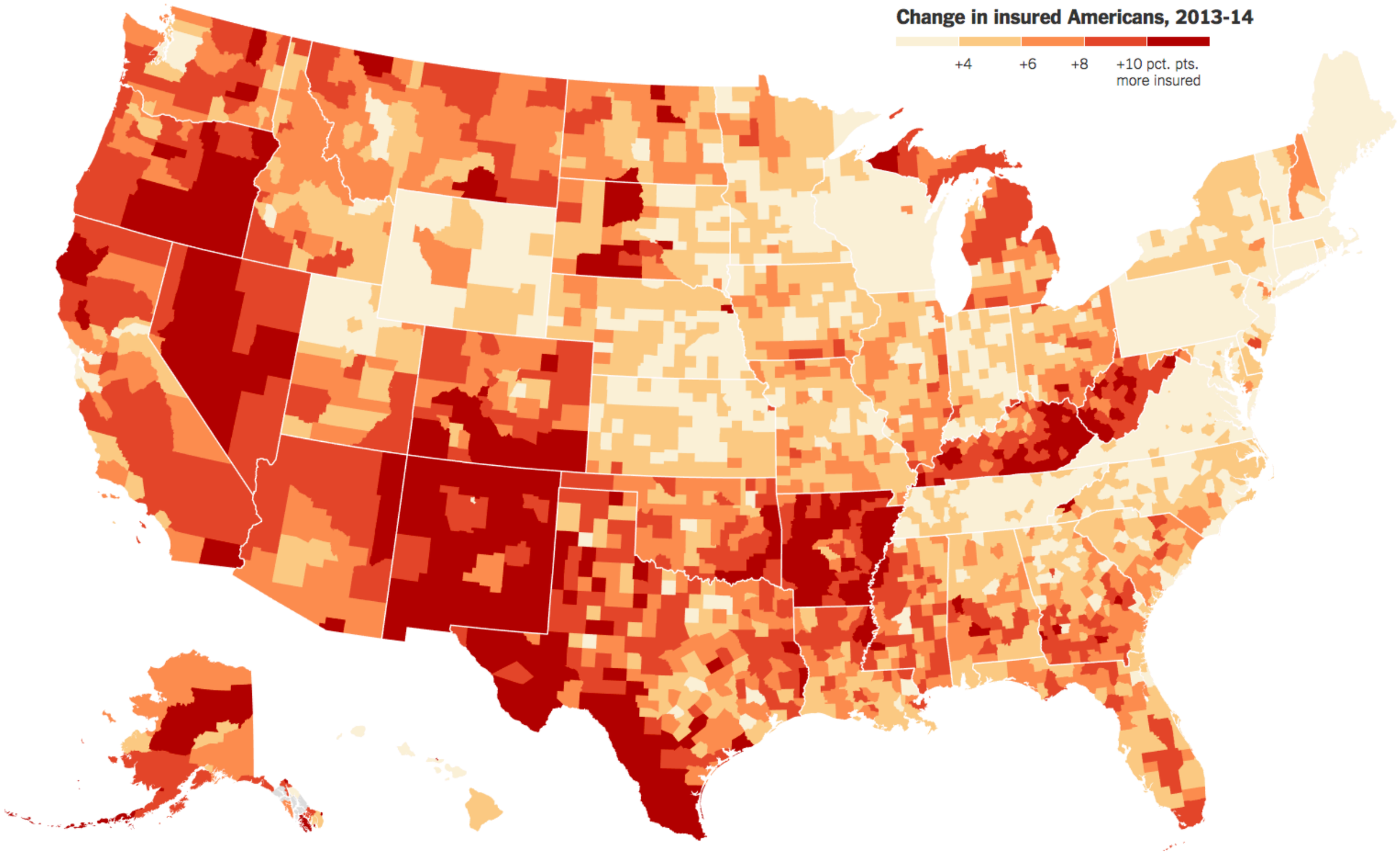
— *Napoleon Bonaparte*

Charles Minard: Napoleon's March on Moscow



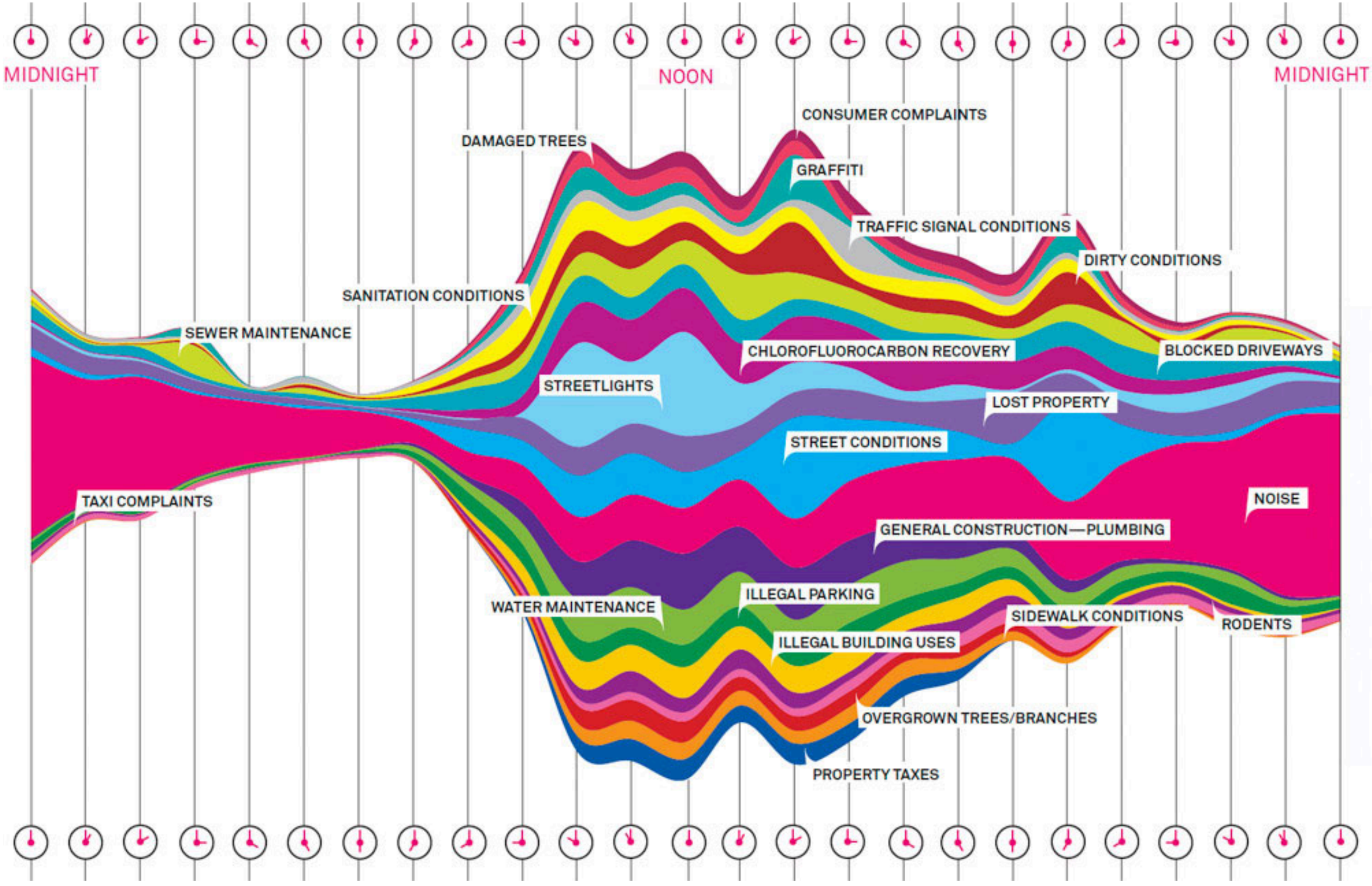
Minard 1869

New York Times: ACA Health Insurance Gains 2013 - 2014



https://www.nytimes.com/interactive/2014/10/29/upshot/obamacare-who-was-helped-most.html

Wired: 34,522 311 calls in New York City between 9/8/10 and 9/15/10



https://www.wired.com/2010/11/ff_311_new_york/

Why do we need it?

I believe it when I see it.

— *Unknown*

Table 1.1: Anscombe's Quartet (Anscombe, 1973). In each of the four data sets mean $\mu_{X_i} = 9.0$, variance $\sigma^2_{X_i} = 11.0$, $\mu_{Y_i} = 7.5$, $\sigma^2_{Y_i} = 4.12$, correlation $cor(X_i, Y_i) = 0.816$ and the linear regression line is $Y_i = 3 + 0.5X_i$ for $i \in \{1,2,3,4\}$.

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

mean(X) = 9, var(X) = 11, mean(Y) = 7.5, var(Y) = 4.12,

cor(X,Y) = 0.816, linear regression line Y = 3 + 0.5*X

Anscombe 1973, The American Statistician

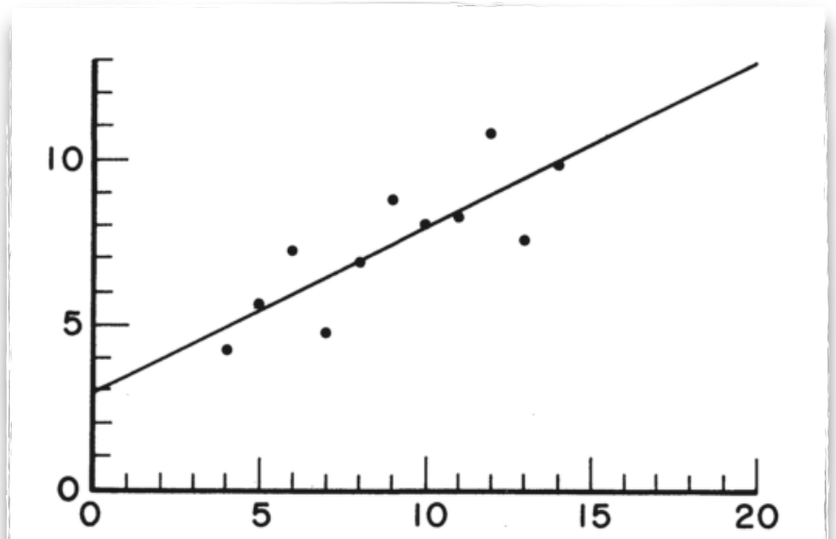


Figure 1

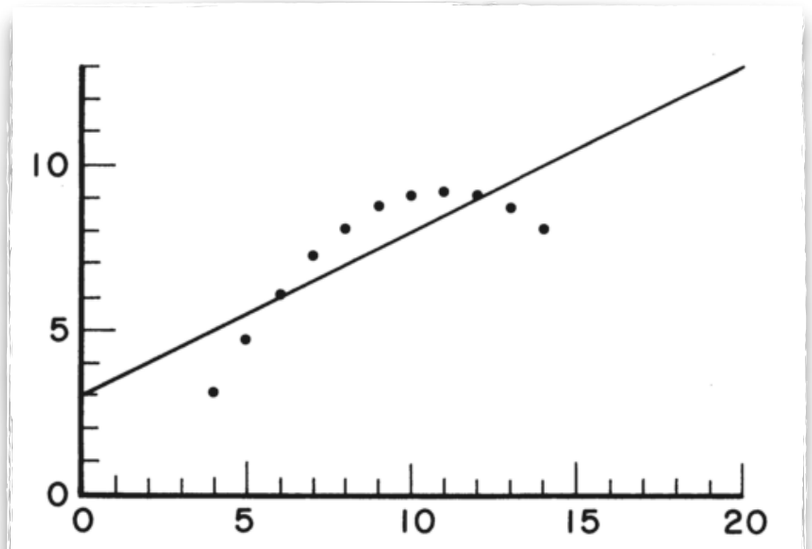


Figure 2

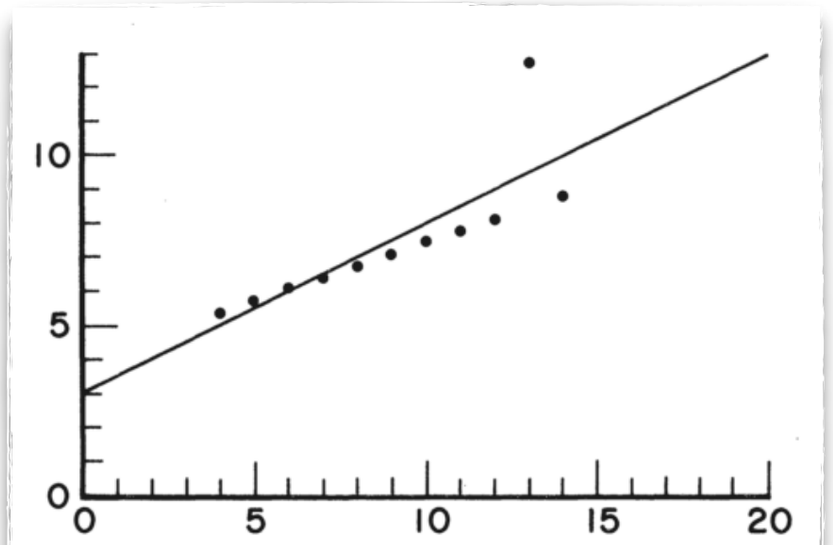


Figure 3

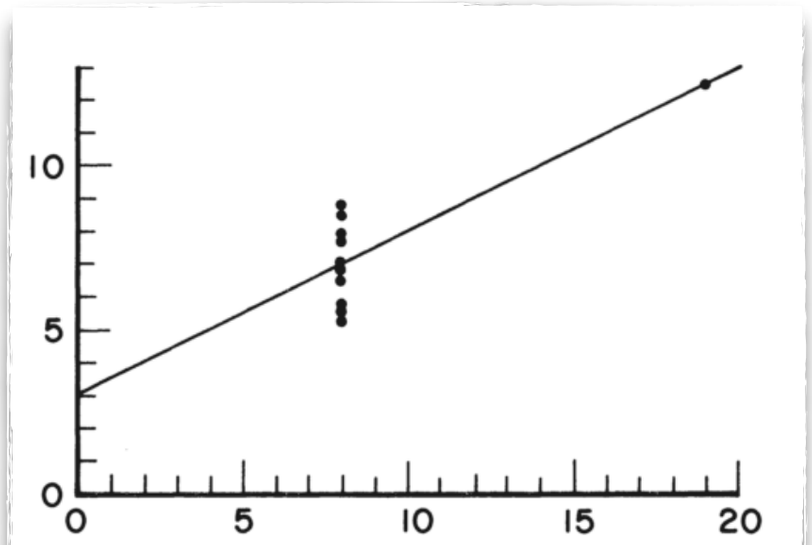


Figure 4

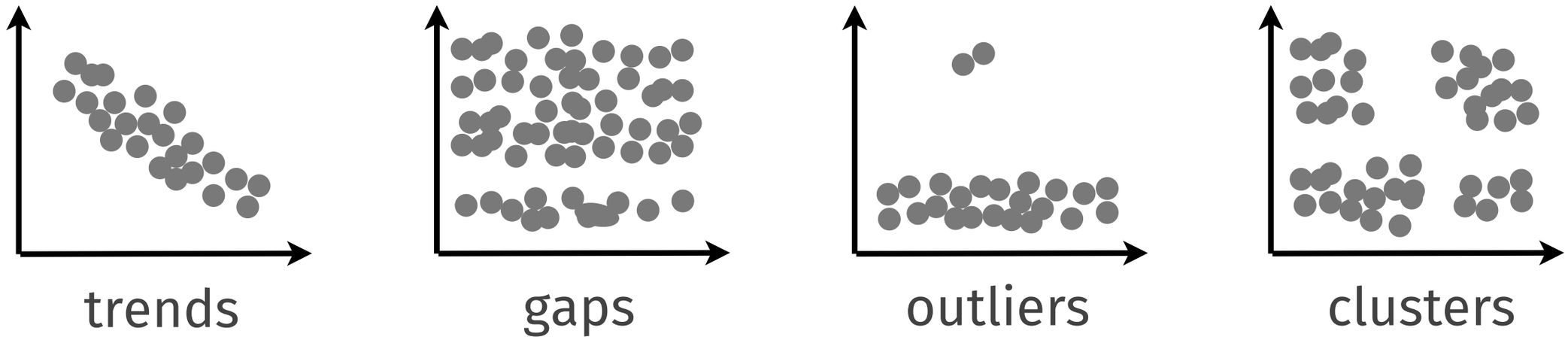
Anscombe 1973, The American Statistician

Why do we need it?

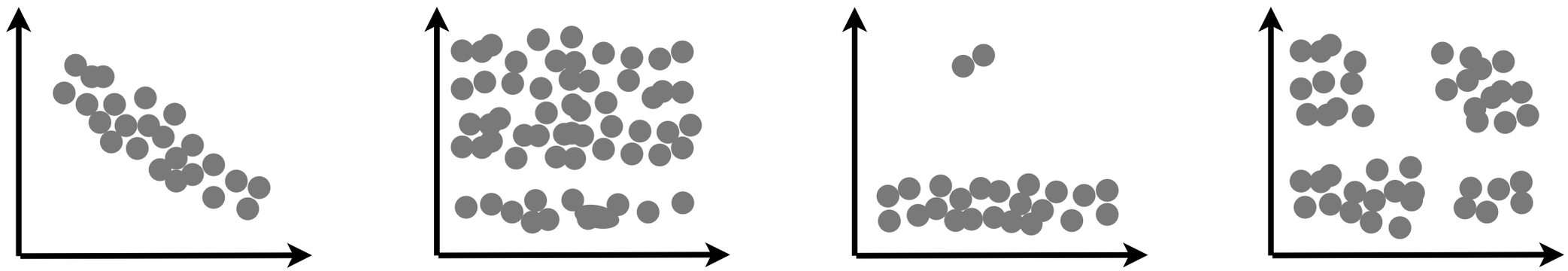
I'm wondering if there are any interesting patterns in my data.

—*Almost Everyone*

Exploration: Hypothesis Generation



Exploration: Hypothesis Generation



Why? Generate hypotheses that can be tested with statistical methods or follow-up experiments.

How? Visualization is employed to perform pattern detection using the human visual system.

Visualization Use Cases

Exploration

Confirmation

Communication

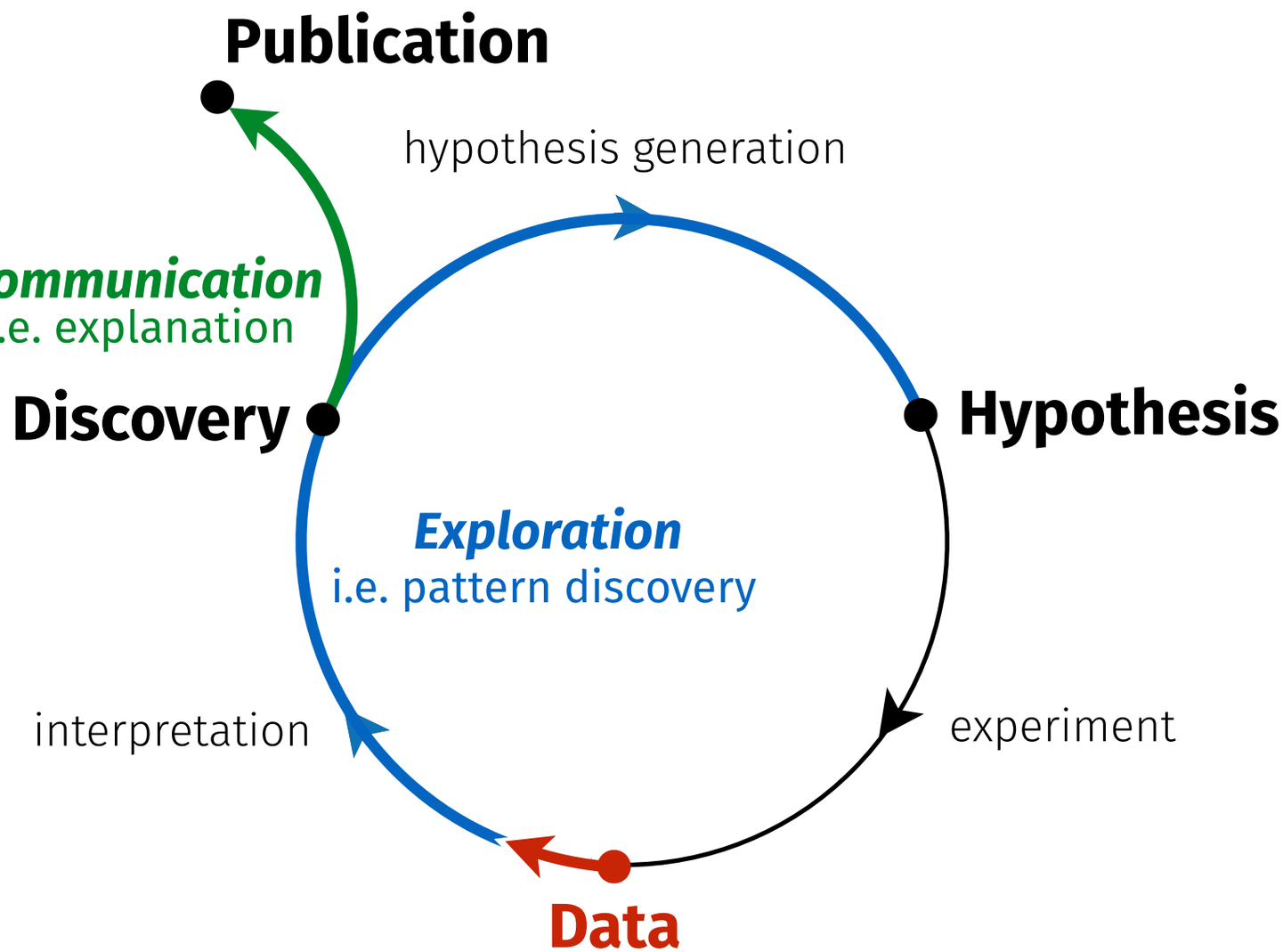
Visualization Use Cases

Exploration

Confirmation

Communication

Discovery Process

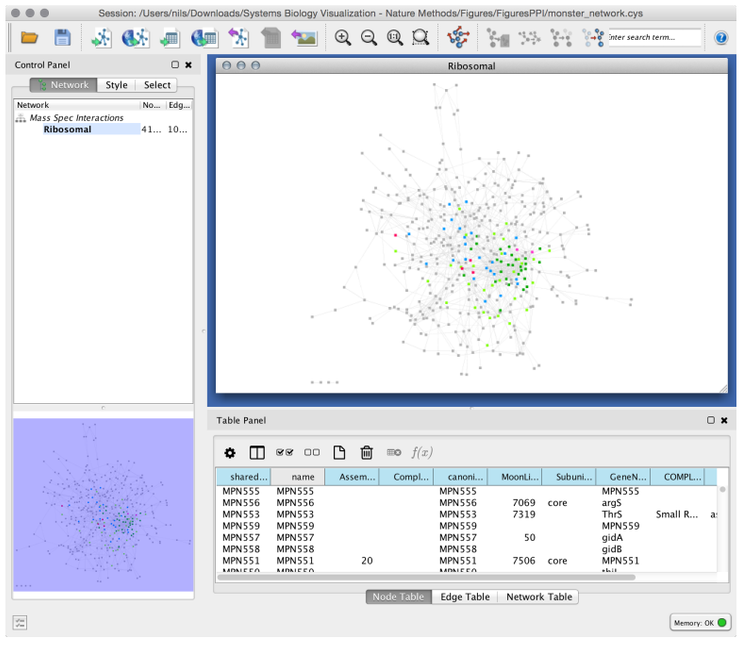


Discovery Process

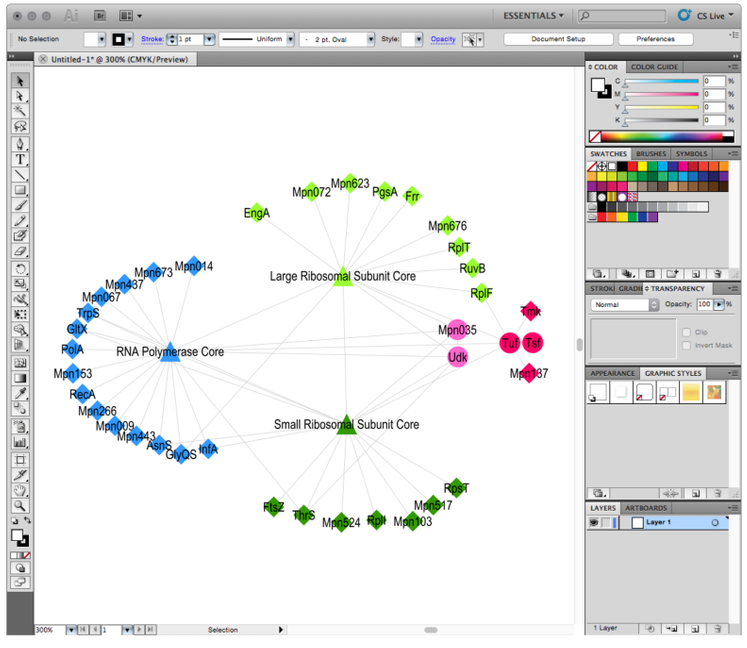
Discovery Process

Exploration ***Communication***

Insight



Cytoscape



Adobe Illustrator

Discovery Process

Exploration

Communication

Insight

Scientists

Scientists
Educators
Policy Makers
Citizens
...

BMI 706

Exploration

Communication

Insight

Syllabus

- Session 1: Introduction to Data Visualization

- Session 2: Design Process, Evaluation, and Interaction

- Session 3: High-Dimensional Data

- Session 4: Networks

- Session 5: Genomes and Epigenomes

- Session 6: Time and Event Sequences

- Session 7: *Project Presentations* and Review

Syllabus

- Session 1: Introduction to Data Visualization

- Session 2: Design Process, Evaluation, and Interaction

- Session 3: High-Dimensional Data

- Session 4: Networks

- Session 5: Genomes and Epigenomes

- Session 6: Time and Event Sequences

- Session 7: *Project Presentations* and Review

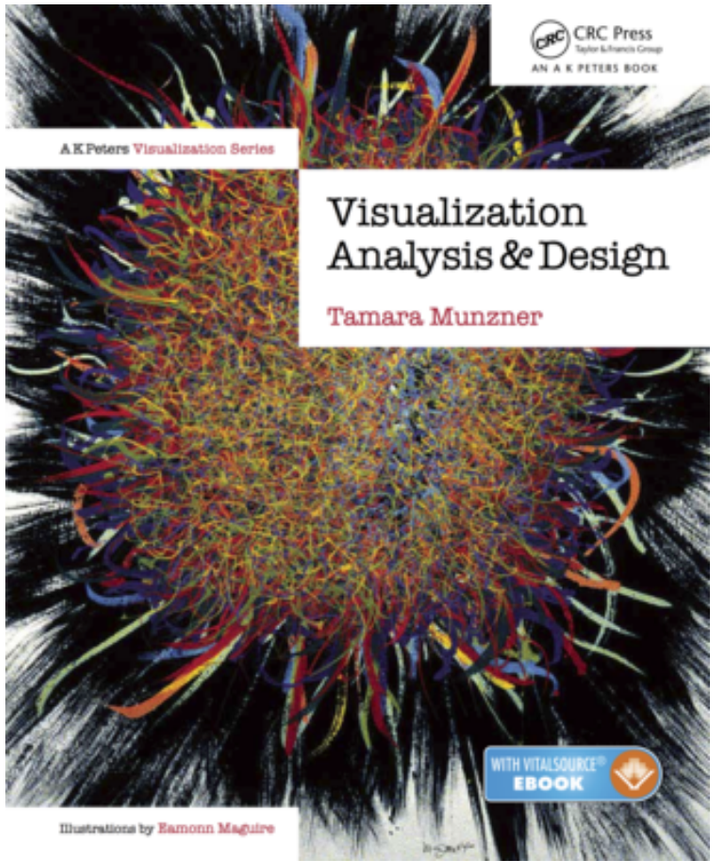
Syllabus

Piazza

https://piazza.com/harvard/spring2018/bmi706/home

Reading

Required Textbook



Tamara Munzer, Visualization Analysis & Design, CRC Press, 2014

Reading

Papers

Selected papers from biomedical informatics literature, data visualization literature, and fringe areas such as design and usability.

PDFs will be provided.

Reading Assignments

- chapters from VAD book by Tamara Munzner

- 3 papers per week, 1 to be presented and discussed (please sign up)

- assignments will be posted on Piazza

Documents & Announcements

Piazza

https://piazza.com/harvard/spring2018/bmi706/home

Project Assignments

Piazza

https://piazza.com/harvard/spring2018/bmi706/home

Instructors

Nils Gehlenborg, PhD

Course Director

nils@hms.harvard.edu

Countway 308

Peter Kerpedjiev, PhD

Teaching Assistant

pkerp@hms.harvard.edu

Sabrina Nusrat, PhD

Teaching Assistant

sabrina_nusrat@hms.harvard.edu

Office Hours

?

Gehlenborg Lab @ DBMI

Methodology

Data Visualization

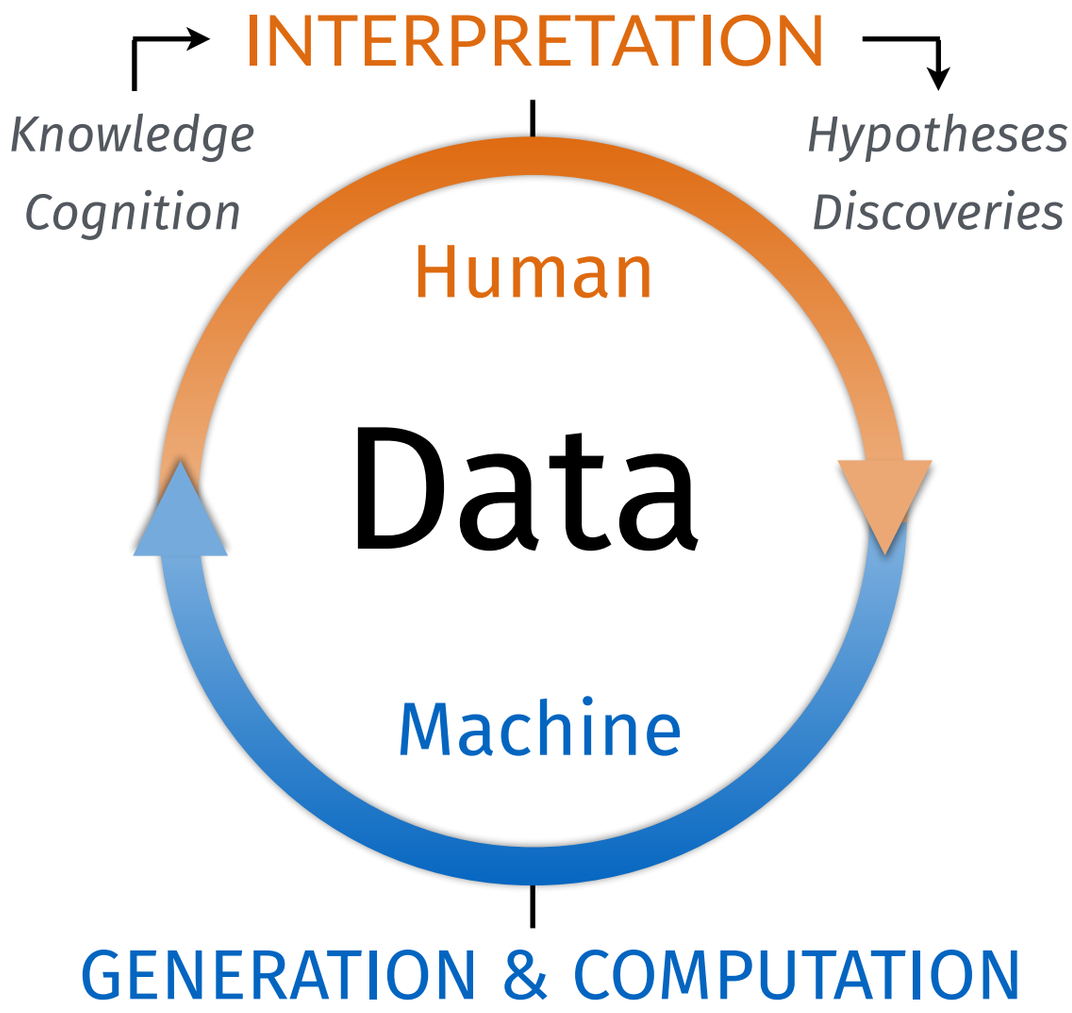
Reproducible Research

Applications

Cancer Genomics

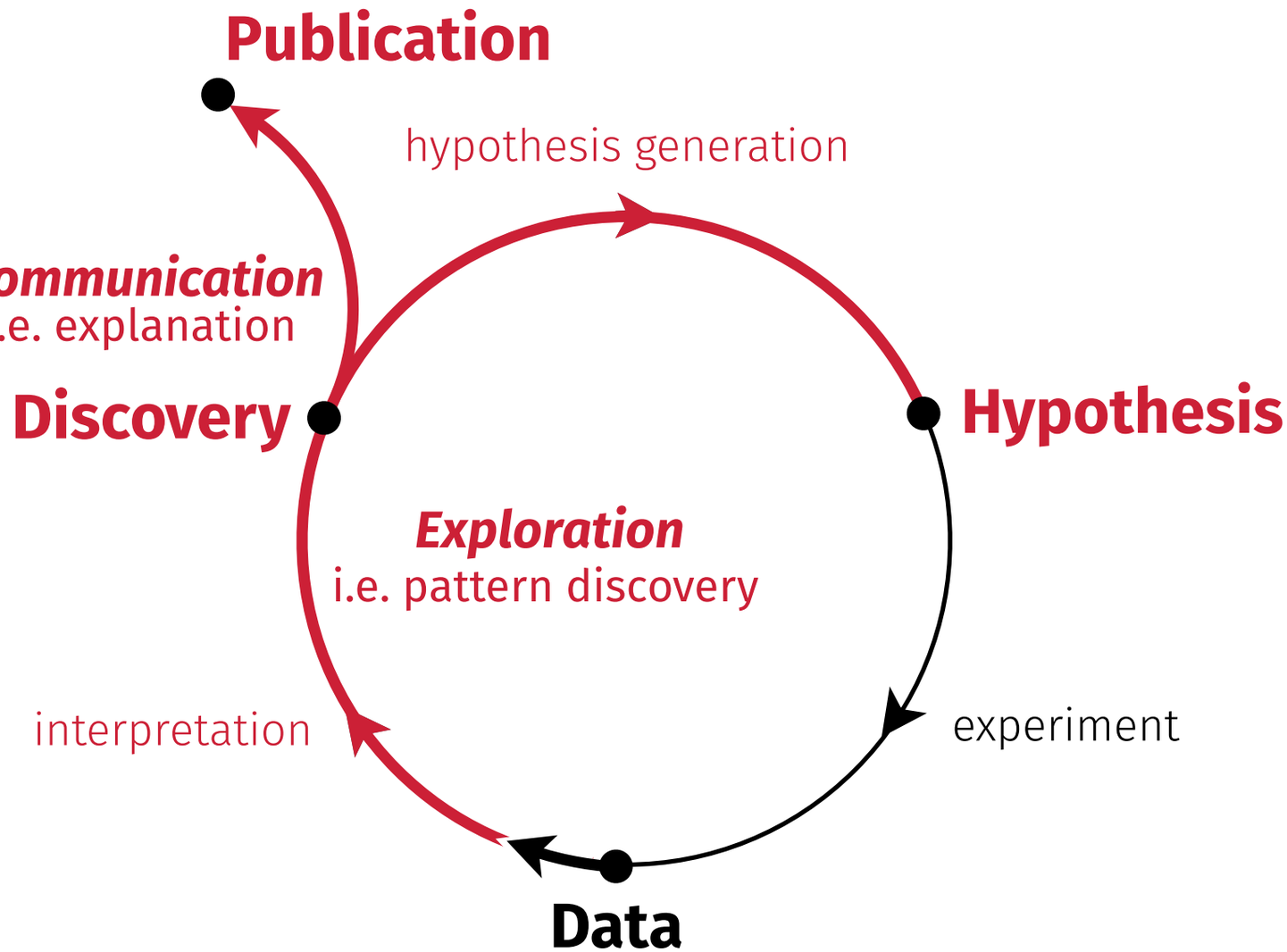
3D Genome Structure

Methods for Data Visualization and Exploration

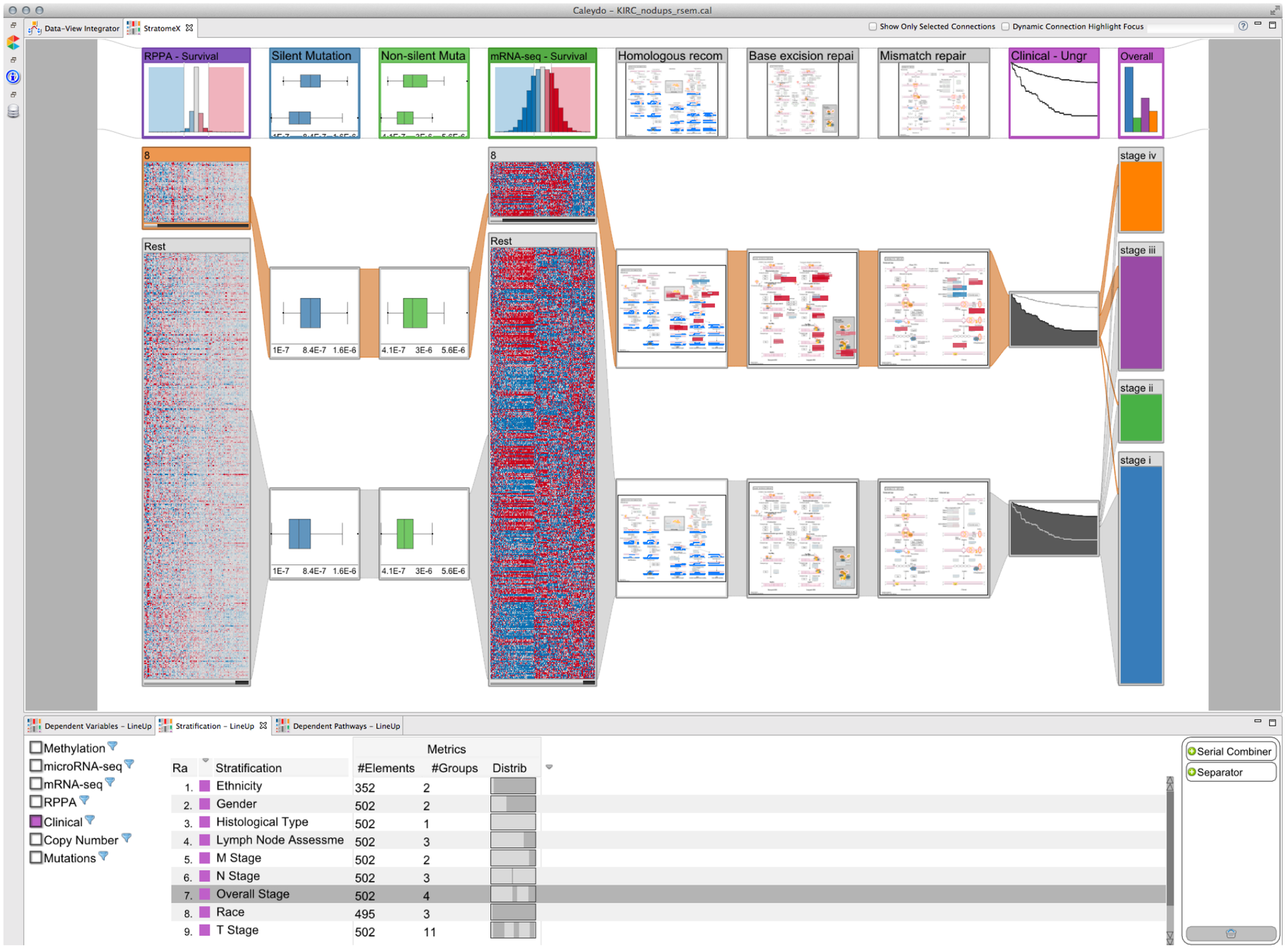


Tools for Reproducible Research

Spectrum of Projects

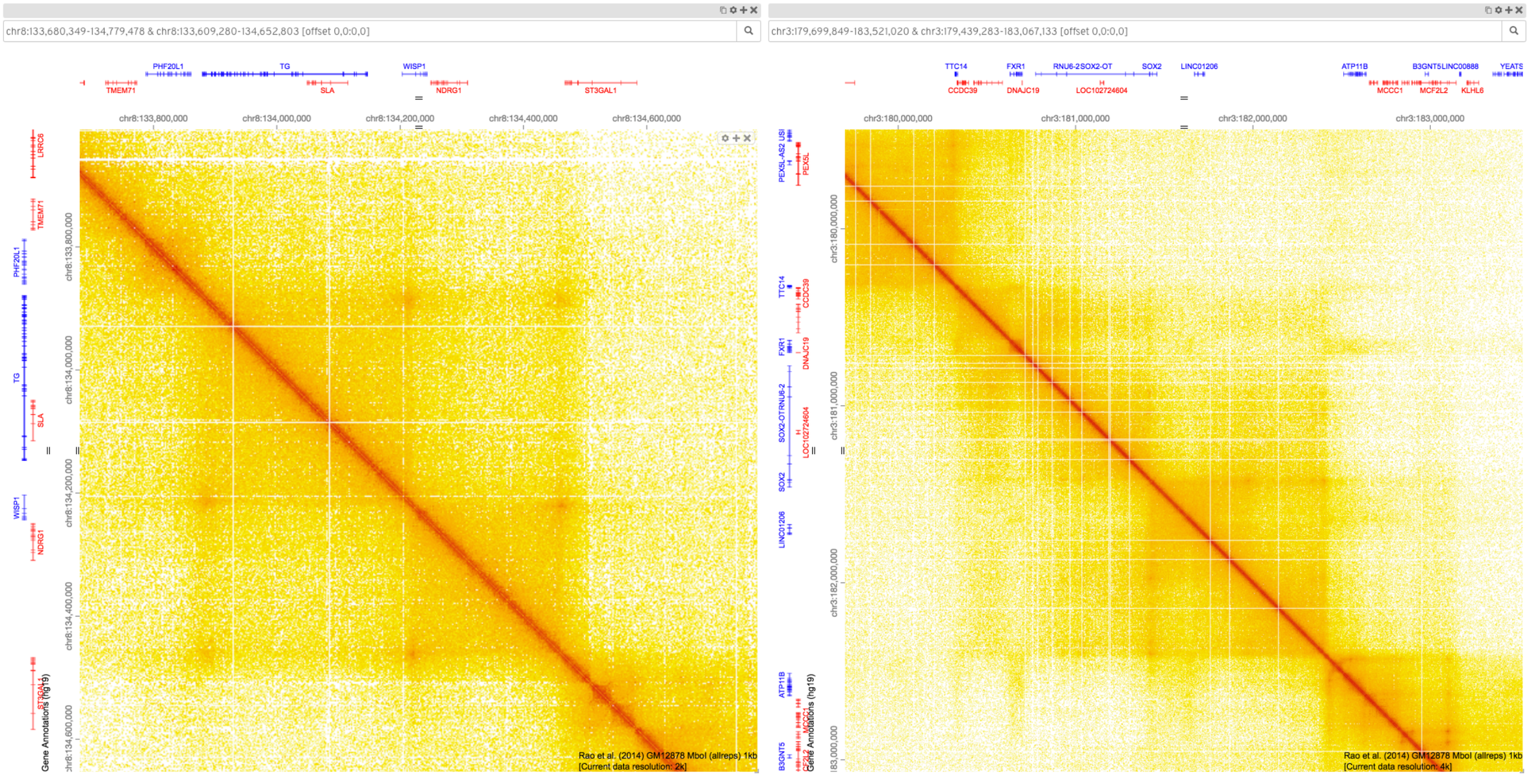


Methods for Data Visualization and Exploration



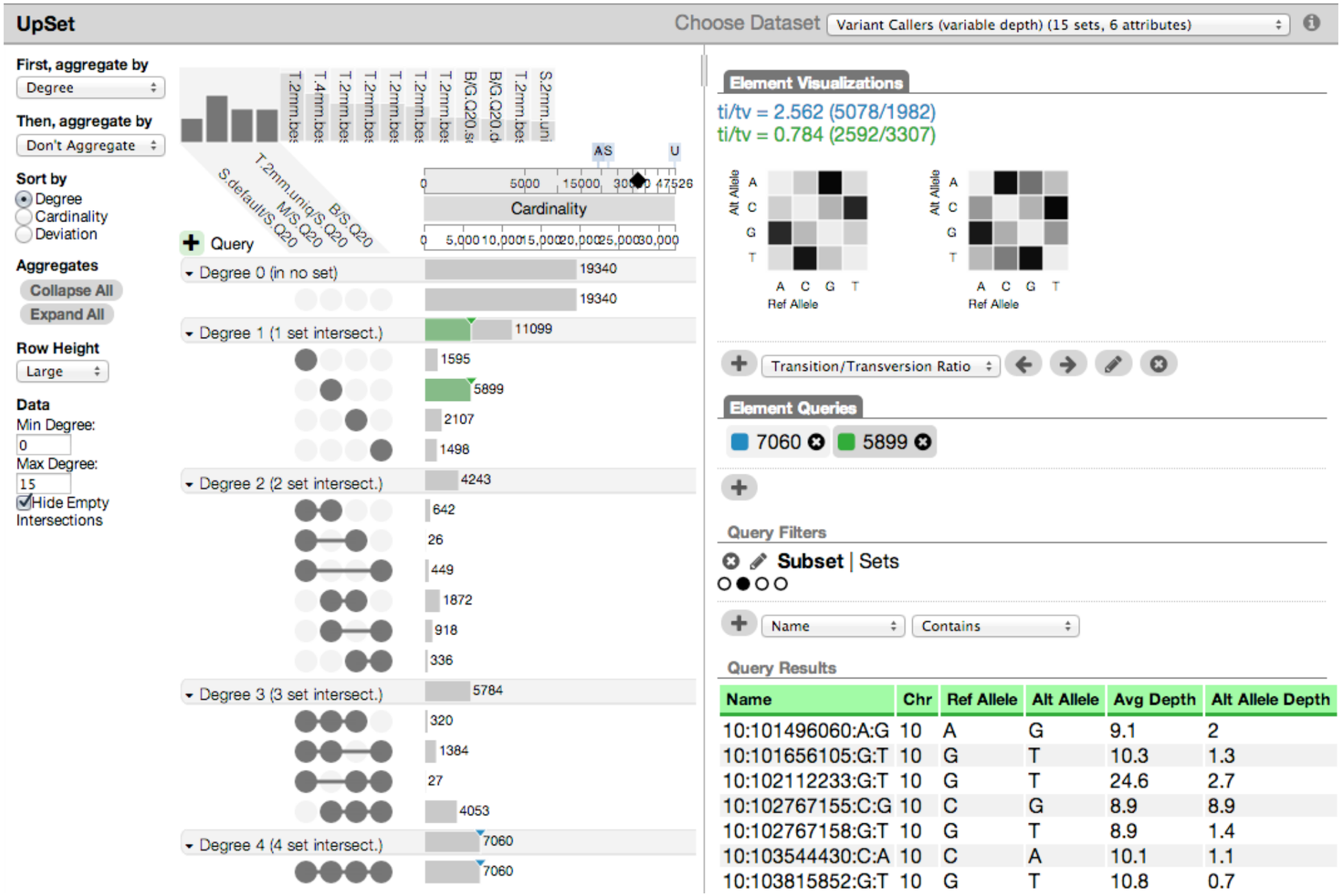
StratomeX - Cancer Subtypes: http://stratomex.caleydo.org

Methods for Data Visualization and Exploration



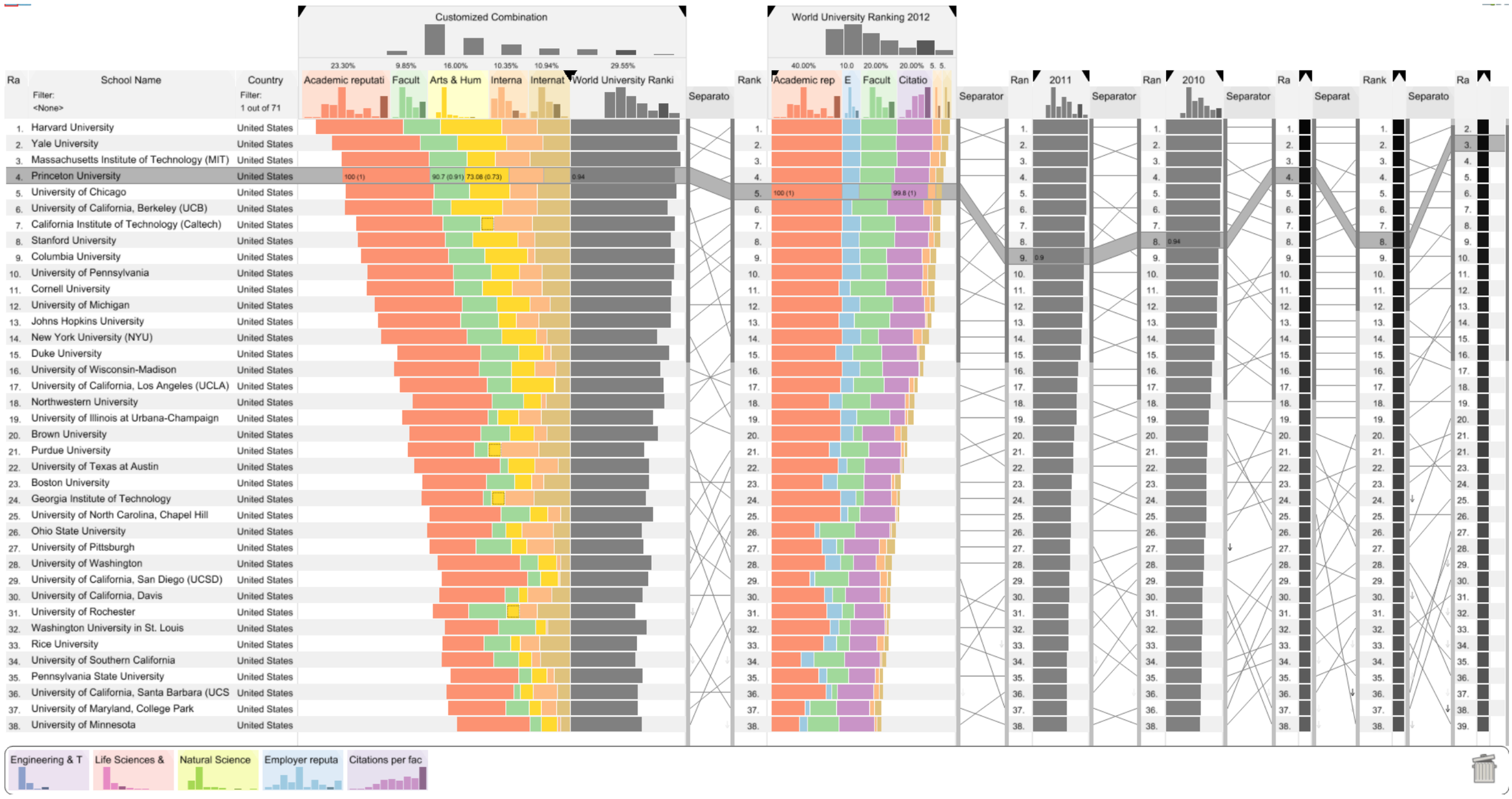
HiGlass - Visualization of Chromatin Interaction Maps: http://higlass.io

Methods for Data Visualization and Exploration



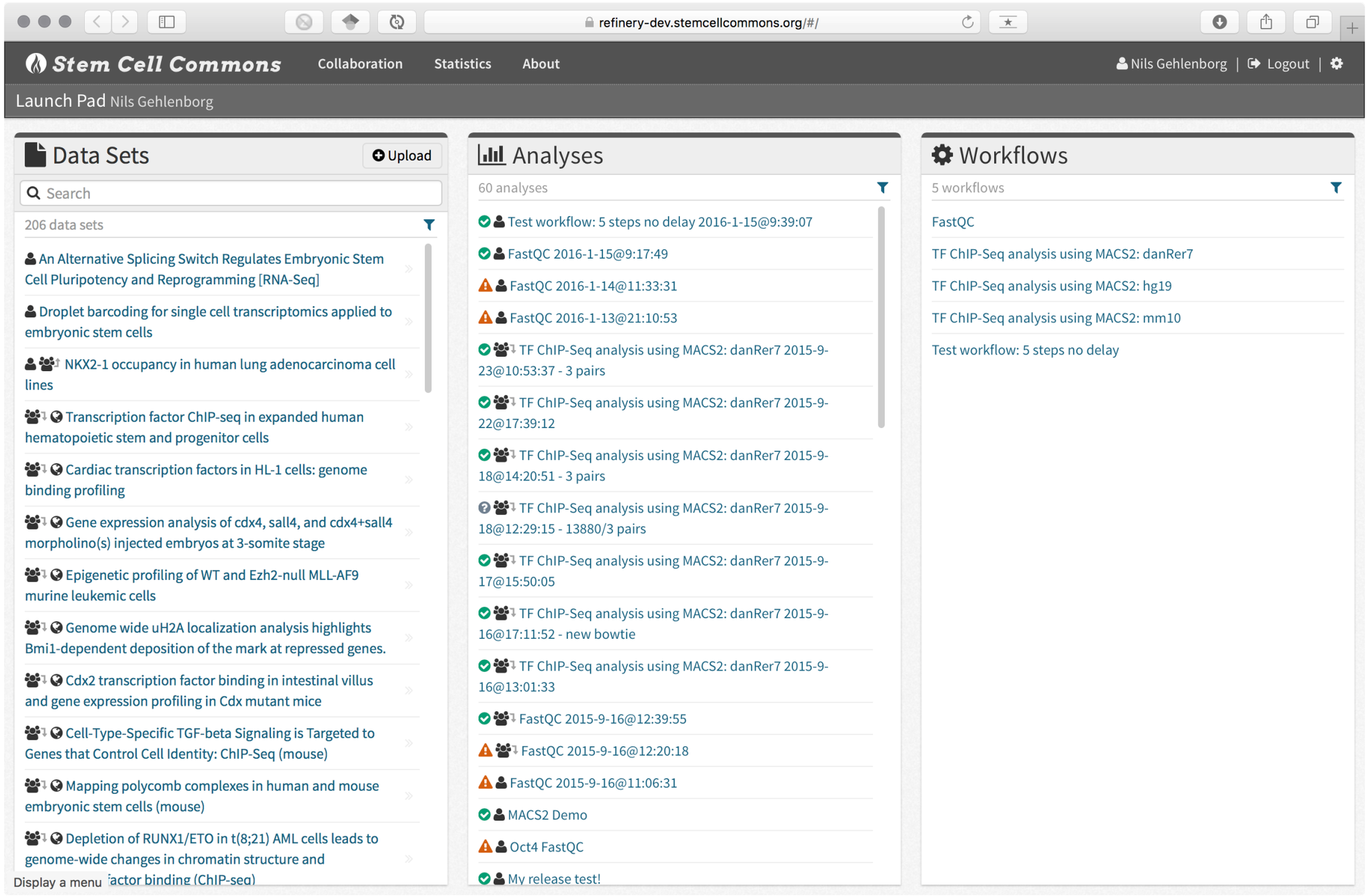
Upset & UpSetR - Visualization of Set Intersections: http://upset.caleydo.org

Methods for Data Visualization and Exploration



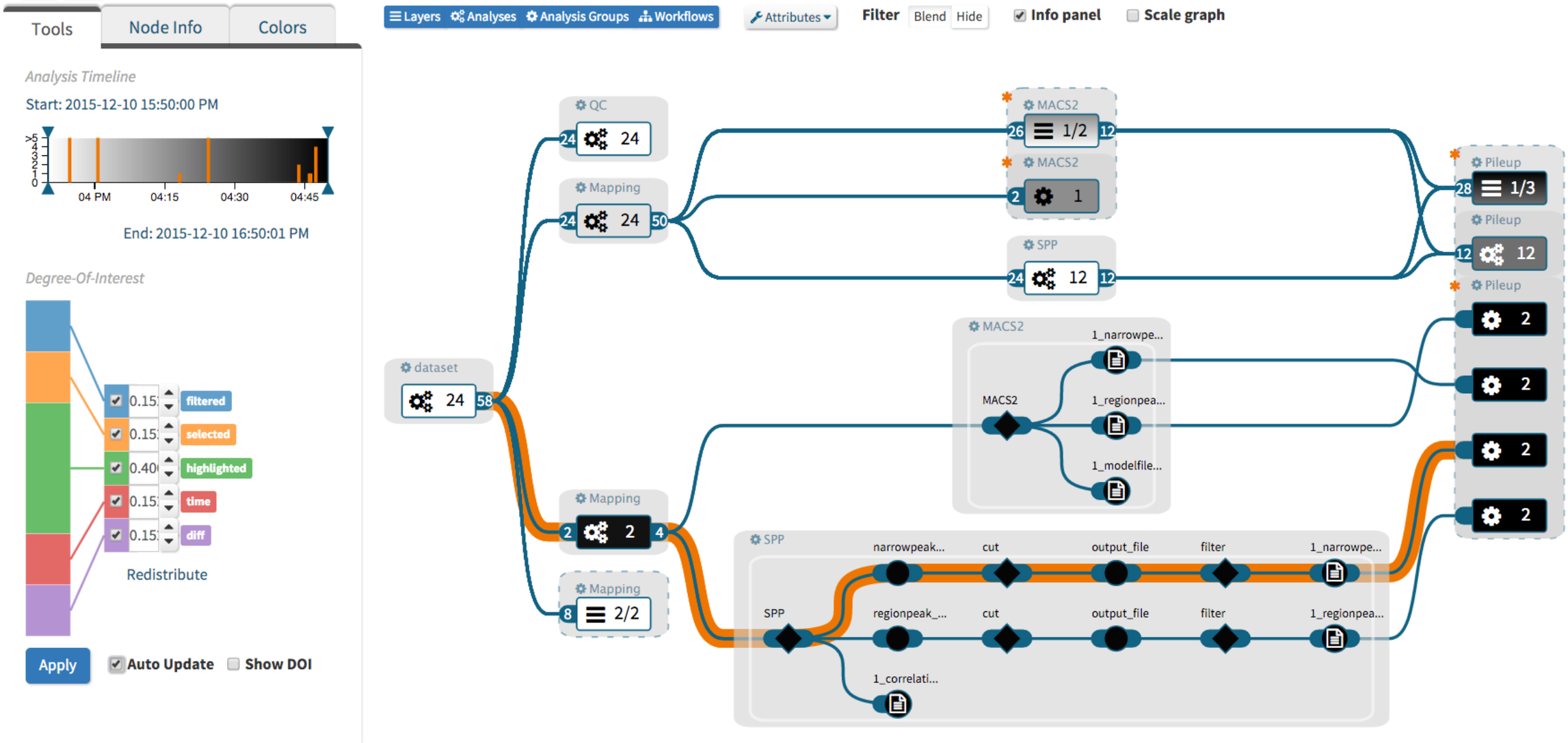
LineUp - Visualization of Rankings: http://lineup.caleydo.org

Tools for Reproducible Research



Refinery Platform: http://refinery-platform.org

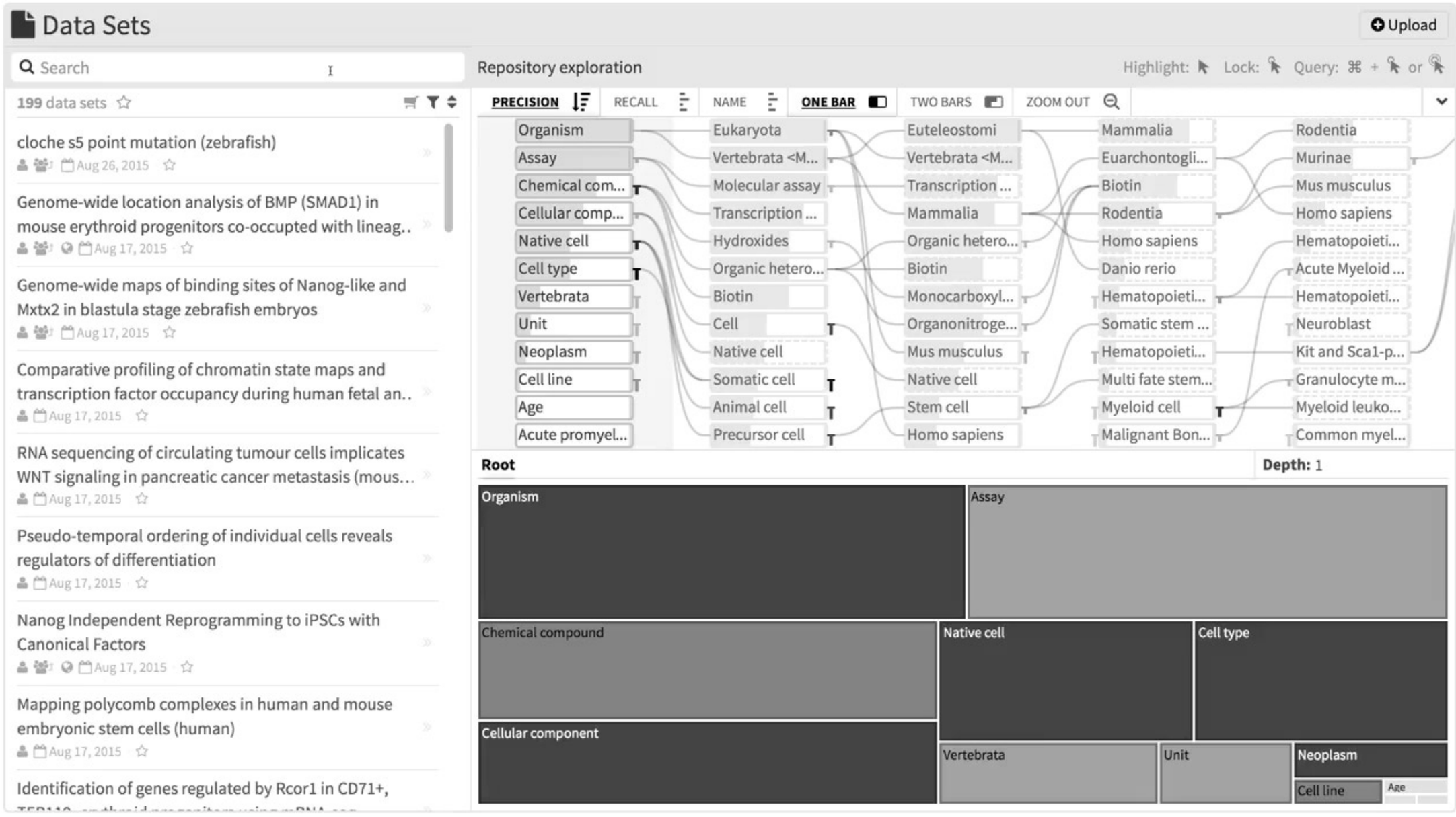
Tools for Reproducible Research



AVOCADO Provenance Visualization

Refinery Platform: http://refinery-platform.org

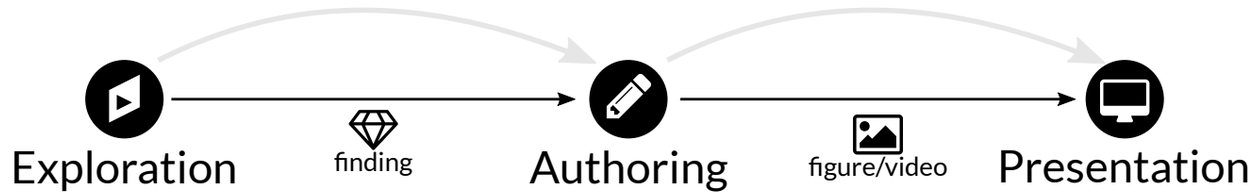
Tools for Reproducible Research

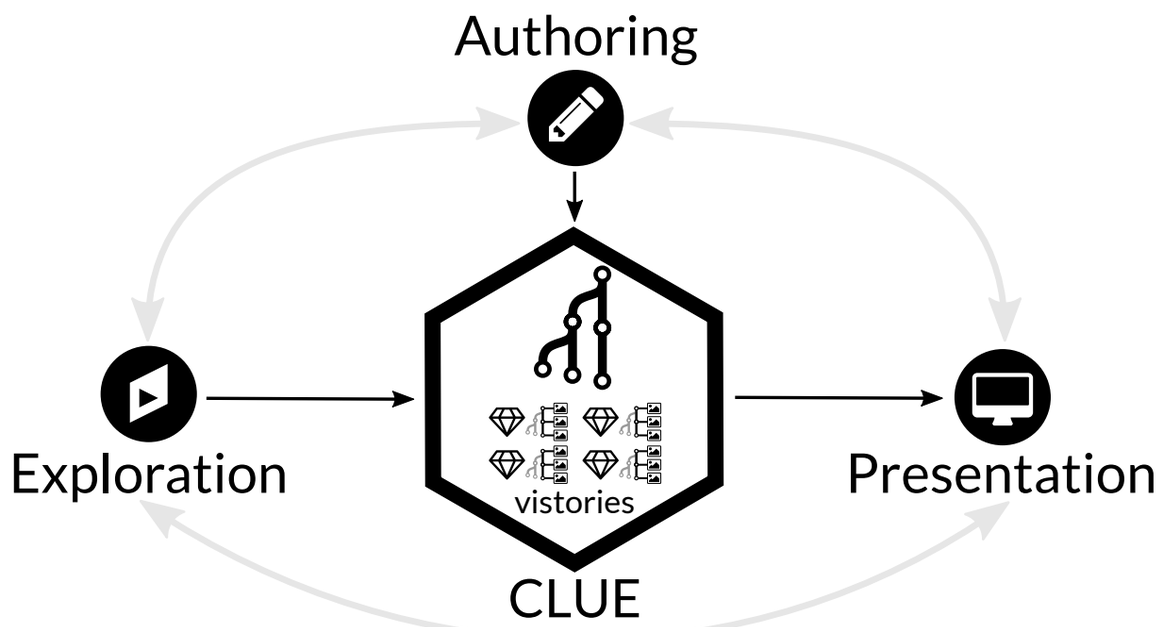


SATORI Ontology-Guided Repository Exploration

Refinery Platform: http://satori.refinery-platform.org

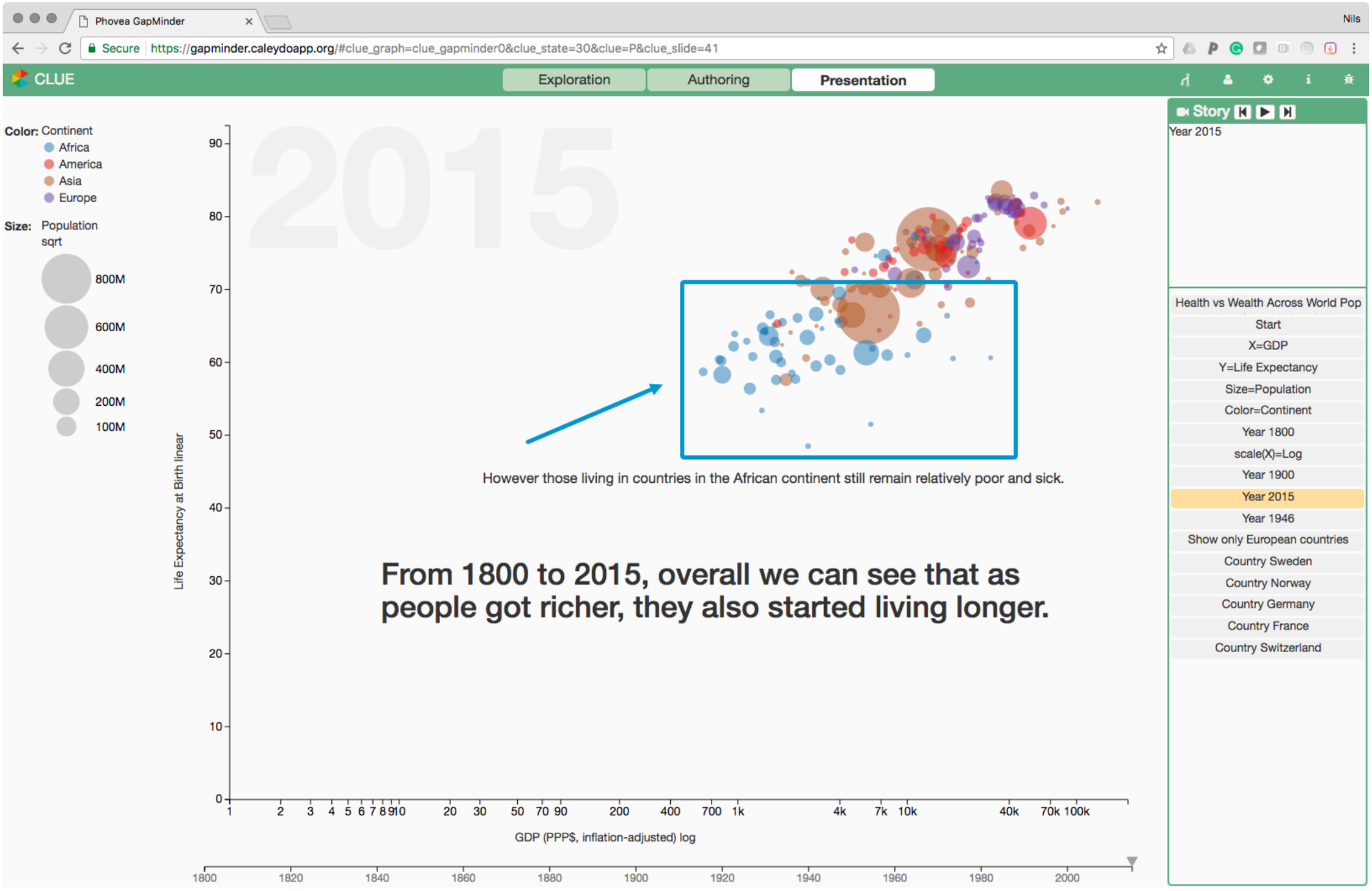
Tools for Reproducible Research





Vistories & CLUE: http://vistories.org

Tools for Reproducible Research



Vistories in GapMinder: http://vistories.org/v/gapminder

How does it work?

> Visualization is really about external cognition, that is, how resources outside the mind can be used to boost the cognitive capabilities of the mind.
>
> — *Stu Card*

How does it work?

Why do we use the visual system and not other sensory systems?

Information bandwidth of visual system is much higher than of all other sensory system.

How does it work?

Visualization uses perception to free up cognition.

How does visualization work?

MALWMRLLPLLALLALWGPDPA

AAFVNQHLCGSHLVEALYLVCG

ERGFFYTPKTRREAEDLQVGQV

ELGGGPGAGSLQPLALEGSLQK

RGIVEQCCTSICSLYQLENYCN

http://www.uniprot.org/uniprot/P01308

How does visualization work?

MALWMRLLPLLALLALWGPDPA

AAFVNQHLCGSHLVEALYLVCG

ERGFFYTPKTRREAEDLQVGQV

ELGGGPGAGSLQPLALEGSLQK

RGIVEQCCTSICSLYQLENYCN

http://www.uniprot.org/uniprot/P01308

How does visualization work?

Visualization uses perception to free up cognition.

Visualization is an external cognitive aid and augments working memory.

How does visualization work?

453 x 862 = ?

How does visualization work?

$$\begin{array}{r} \mathbf{453 \times 862 = ?} \\ \hline 906 \\ +\ 27{,}180 \\ +\ 362{,}400 \\ \hline 390{,}486 \end{array}$$

How does visualization work?

	f_1		f_2	
	x	y	x	y
1	10	9.14	10	7.46
2	8	8.14	8	6.77
3	13	8.74	13	12.74
4	9	8.77	9	7.11
5	11	9.26	11	7.81
6	14	8.10	14	8.84
7	6	6.13	6	6.08
8	4	3.10	4	5.39
9	12	9.13	12	8.15
10	7	7.26	7	6.42
11	5	4.74	5	5.73

Tasks

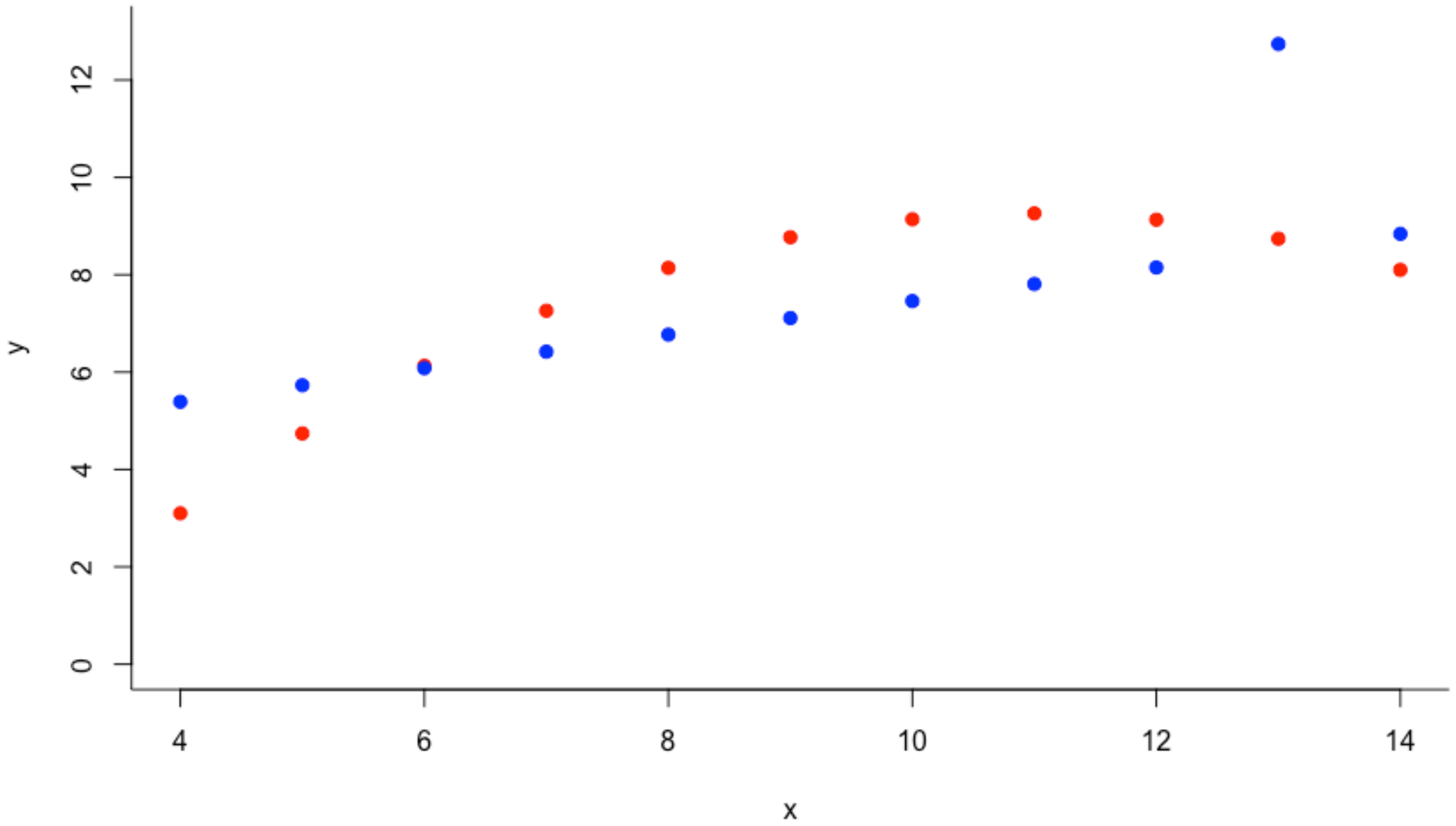
What is the shape of f_1? Of f_2?

How many times do f_1 and f_2 intersect?

Do they cross 0?

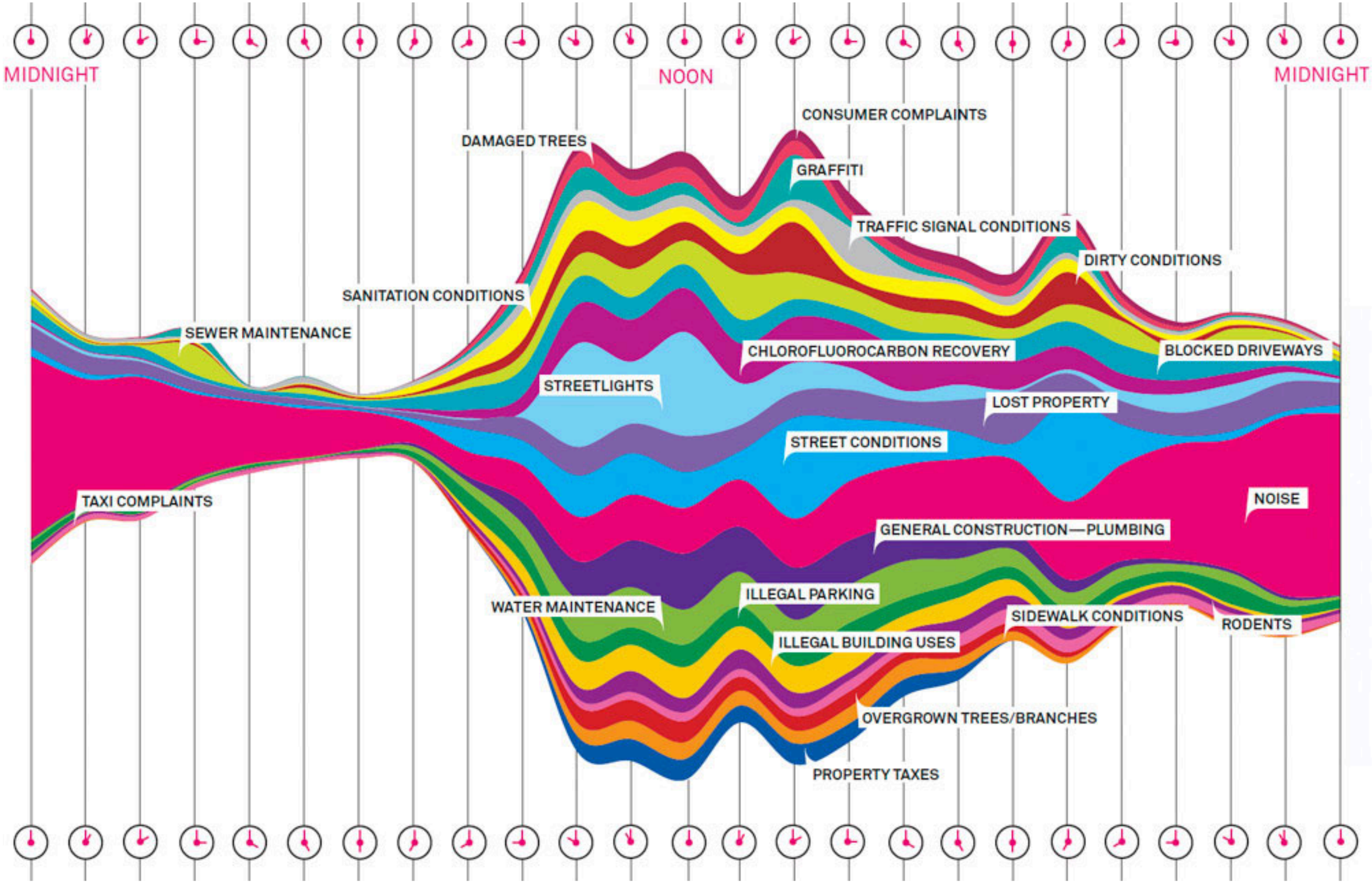
…

How does visualization work?



Design Critique

Wired: 34,522 311 calls in New York City between 9/8/10 and 9/15/10



https://www.wired.com/2010/11/ff_311_new_york/

Visual Encoding of Data

data

apples	*small*	*10 inches*	*trees*	*intrinsic*
oranges	*medium*	*13 inches*	*networks*	*position*
bananas	*large*	*18.5 inches*		

slide adapted from Munzner 2011, Visualization Principles

Visual Channels: Rankings

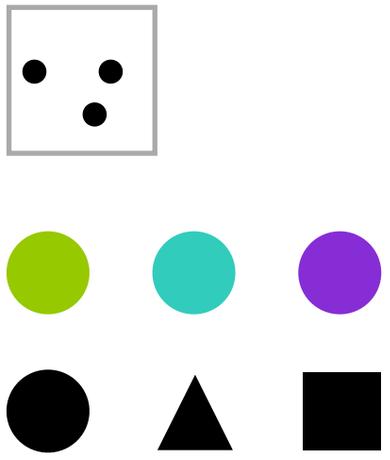
Categorical

What? Where?

position*
planar

color hue

shape



Relational

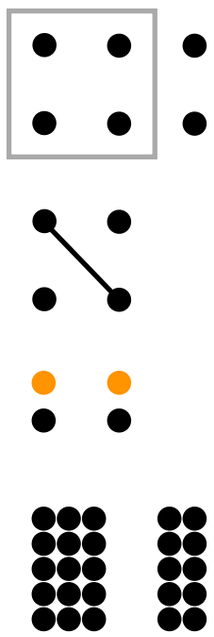
With whom?

containment

connection

similarity

position*
proximity



slide adapted from Munzner 2011, Visualization Principles

Visual Channels: Rankings

Ordinal & Quantitative

How much?

position*
common scale

position*
unaligned scale

length (1D)

angle/tilt

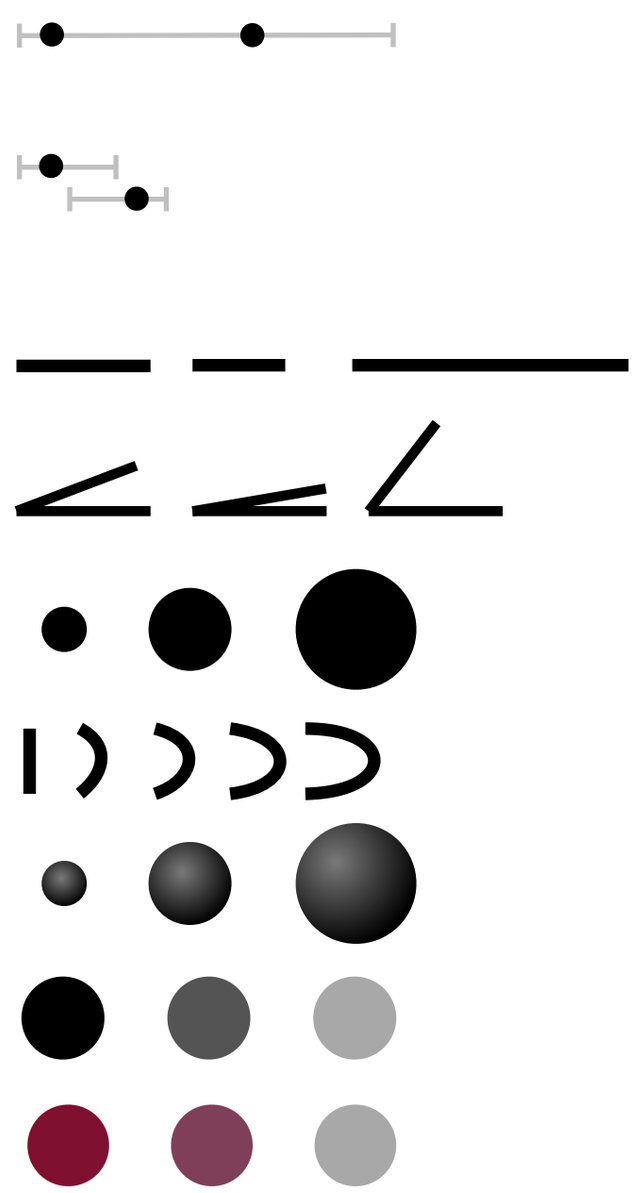
area (2D)

curvature

volume (3D)

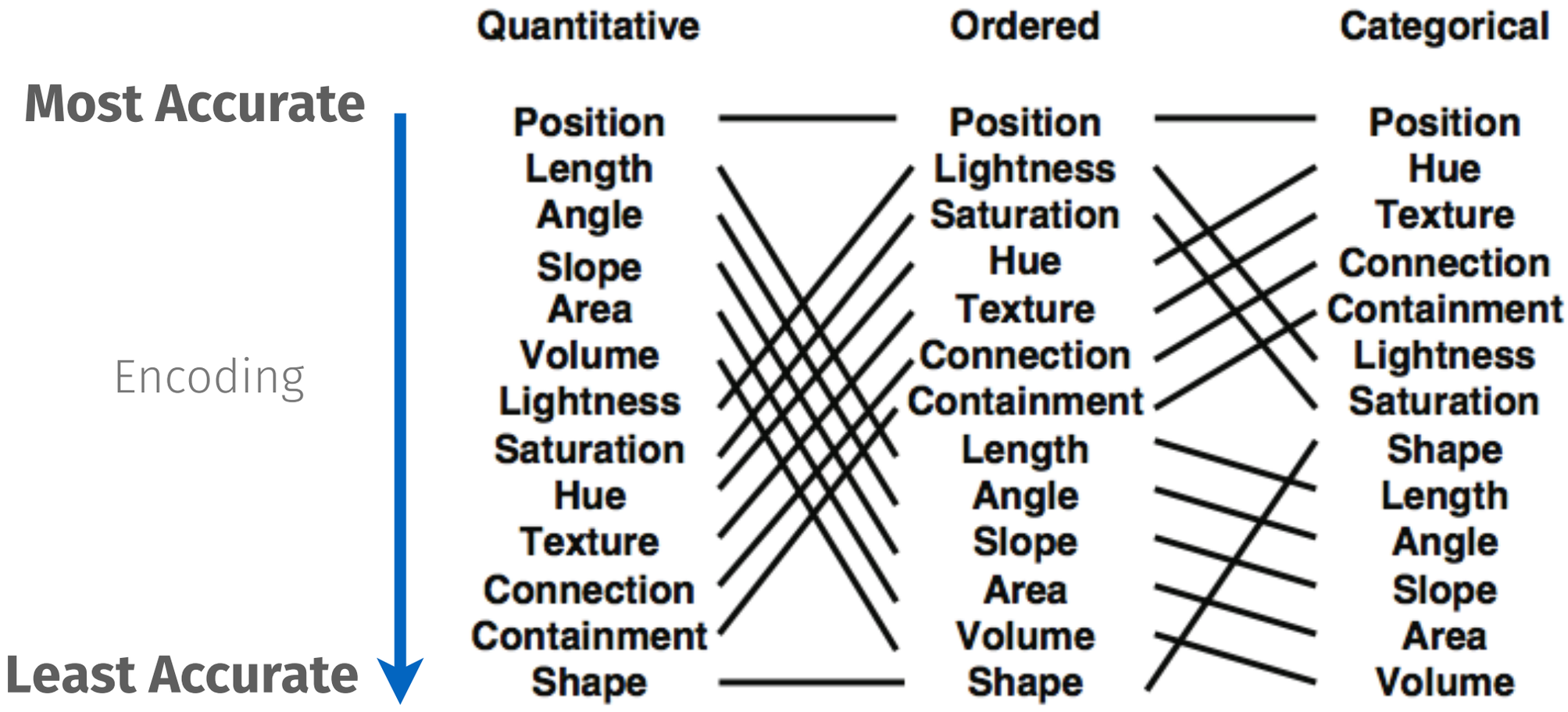
lightness

color saturation



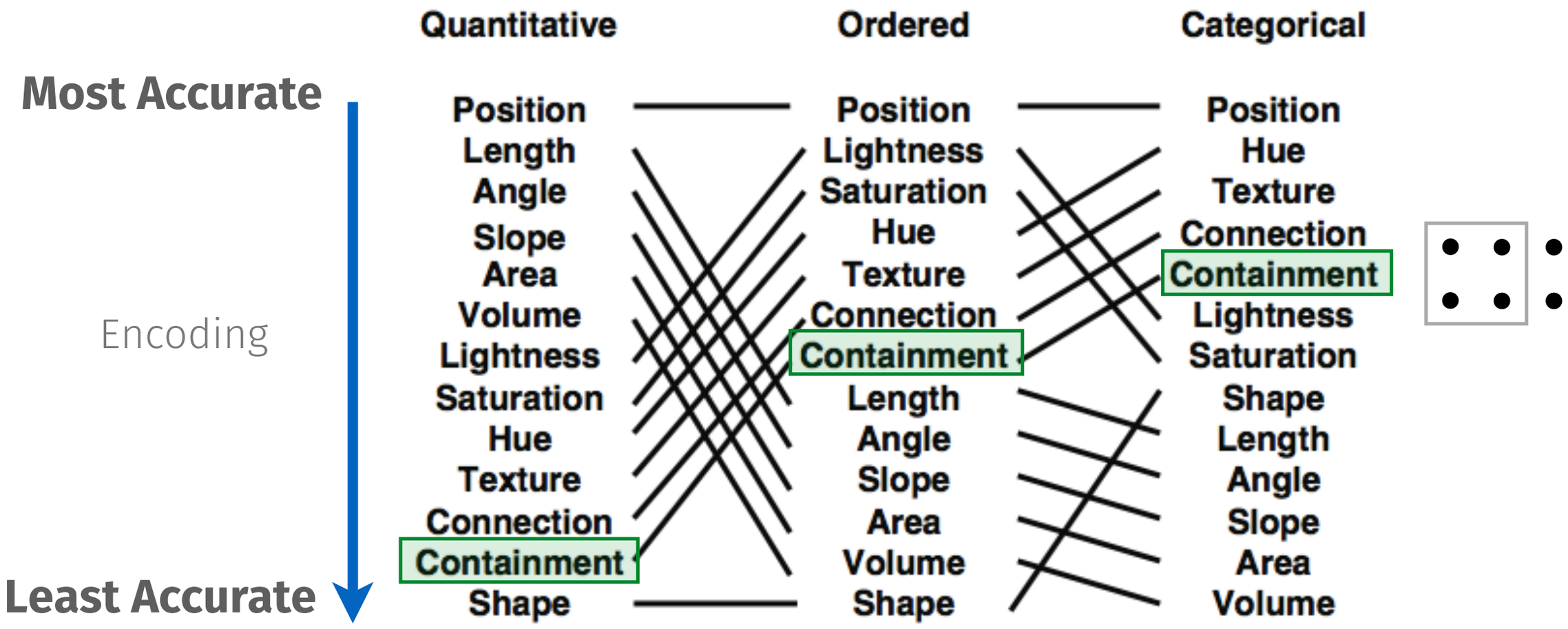
slide adapted from Munzner 2011, Visualization Principles

Ranking of Encodings



Munzner 2009, in Fundamentals of Computer Graphics (redrawn from Mackinlay 1986)

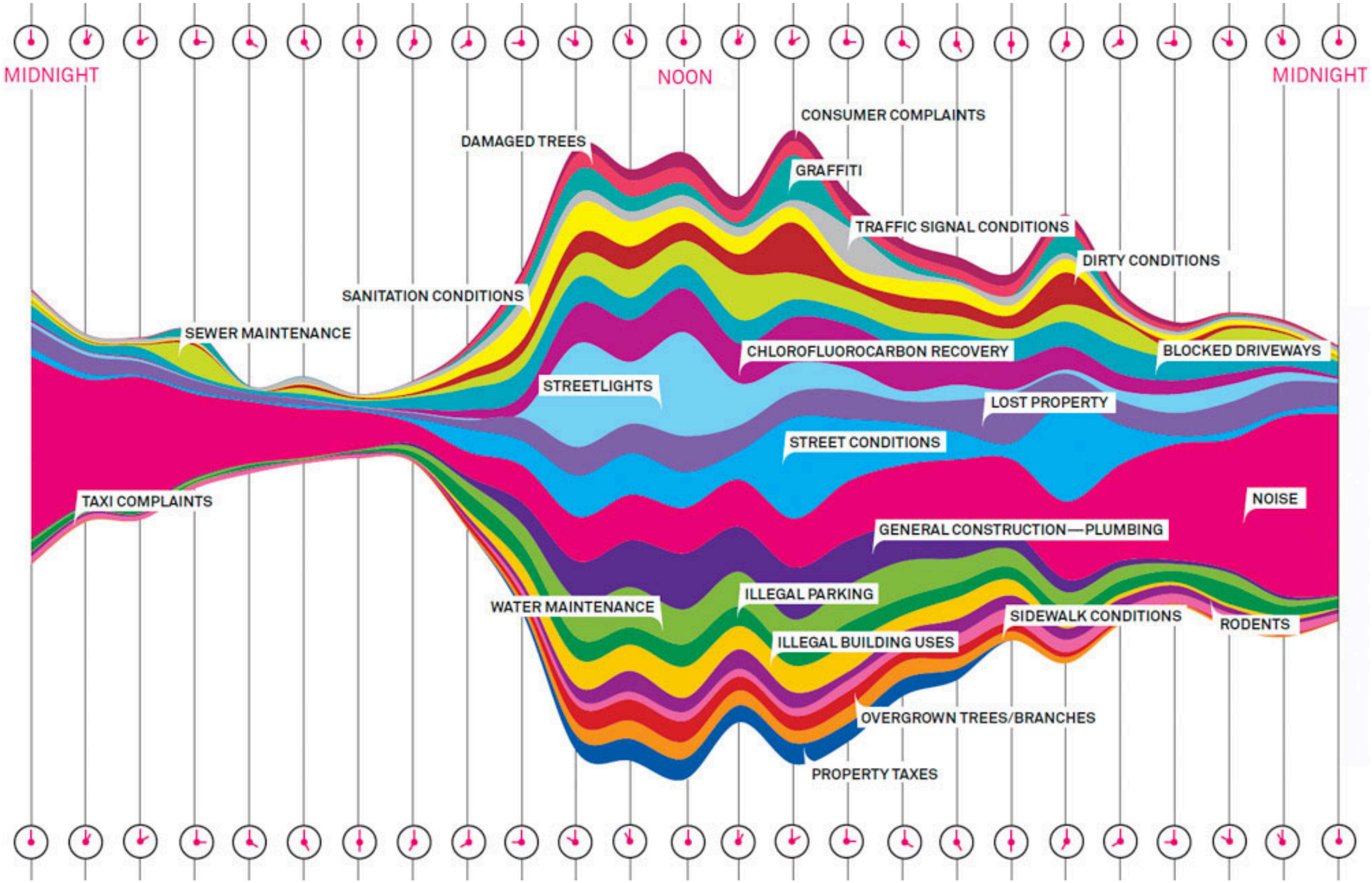
Ranking of Encodings



Munzner 2009, in Fundamentals of Computer Graphics (redrawn from Mackinlay 1986)

Back to our Design Critique …

Exercise: Design Critique



F Xu et al., Nature Communications 3, 1258 doi:10.1038/ncomms2256

Ranking of Encodings

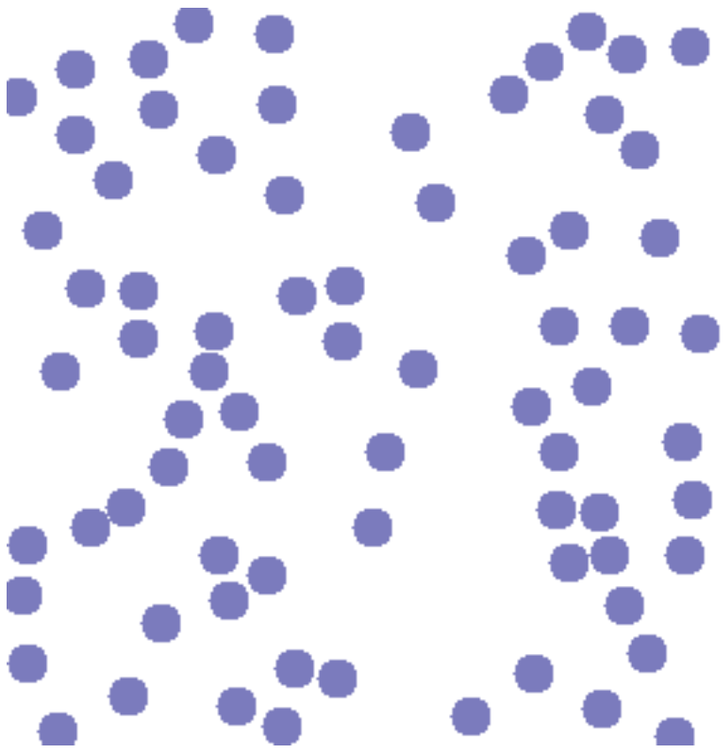
Principle of Importance Ordering (Mackinaly 1986):

Encode more important information more effectively.

- How accurately can the data be read from the visualization?
- How many classes can be distinguished?
- Can the channels be separated from each other?
- Which channels are processed preattentively?

Preattentive Processing: Color

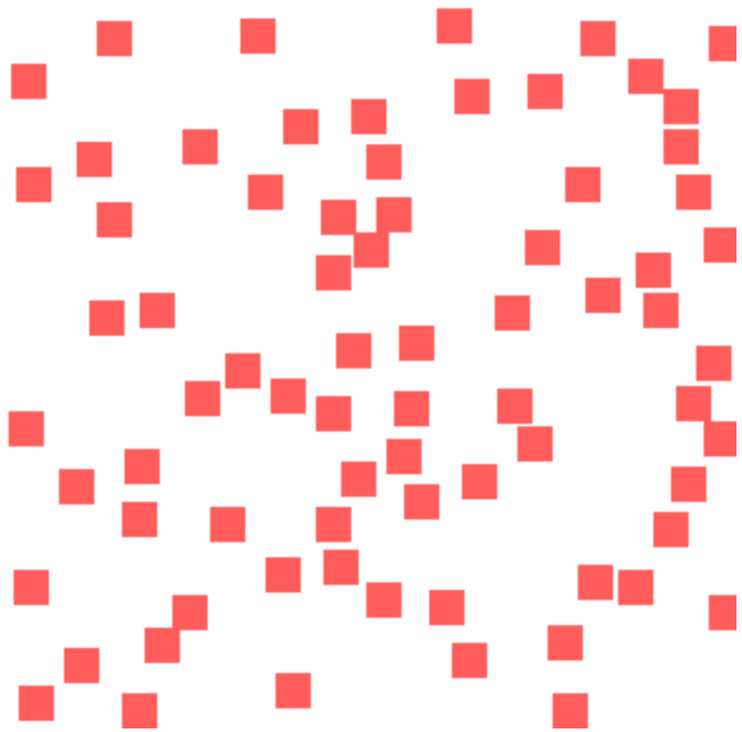
Can you spot this: ●?



Healey, http://www.csc.ncsu.edu/faculty/healey/PP/

Preattentive Processing: Shape

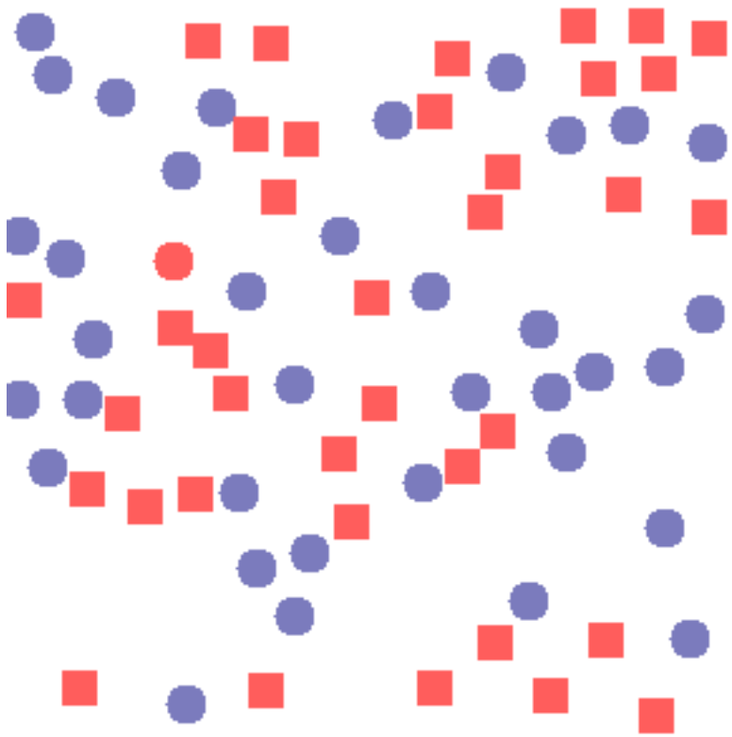
Can you spot this: ● ?



Healey, http://www.csc.ncsu.edu/faculty/healey/PP/

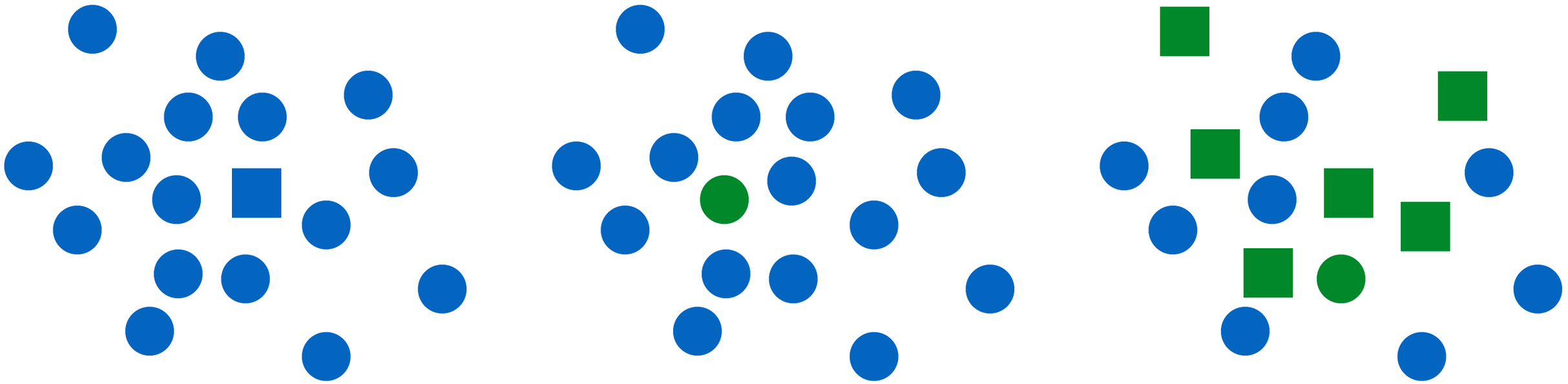
Preattentive Processing: Shape & Color

Can you spot this: ●?



Healey, http://www.csc.ncsu.edu/faculty/healey/PP/

Preattentive Processing



Shape: Yes!

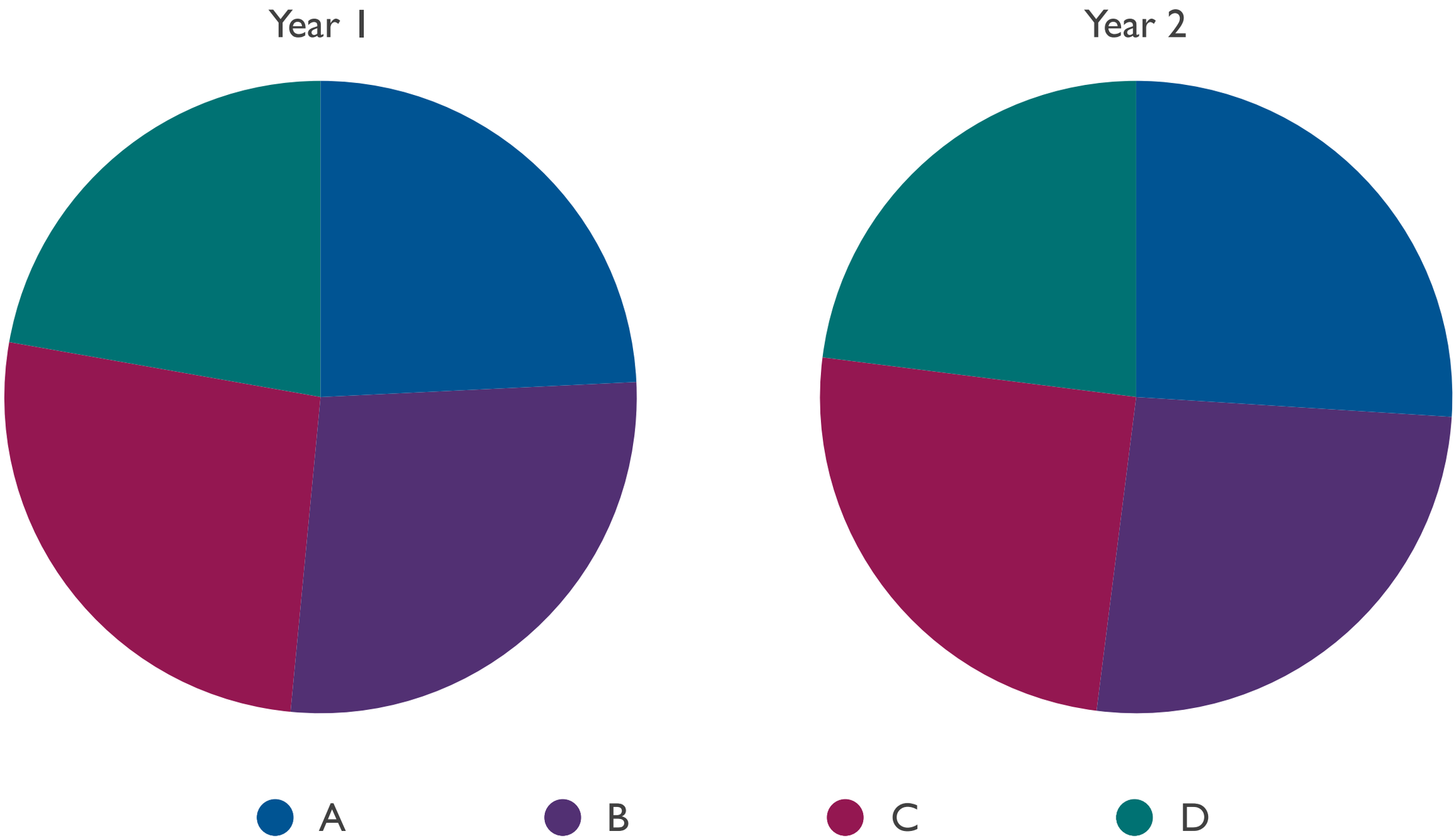
Color: Yes!

Shape + Color: No!

- visual properties that can be perceived in less than 250 ms
- no sequential scanning of the image required, unlike text or numbers
- examples for other visual properties that can be processed preattentively: orientation, curvature, direction of motion, size and others

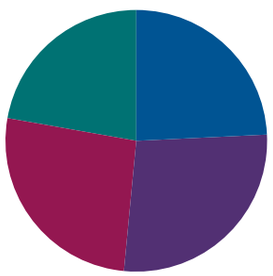
after Healey et al. 1993, Healey and Enns 1999

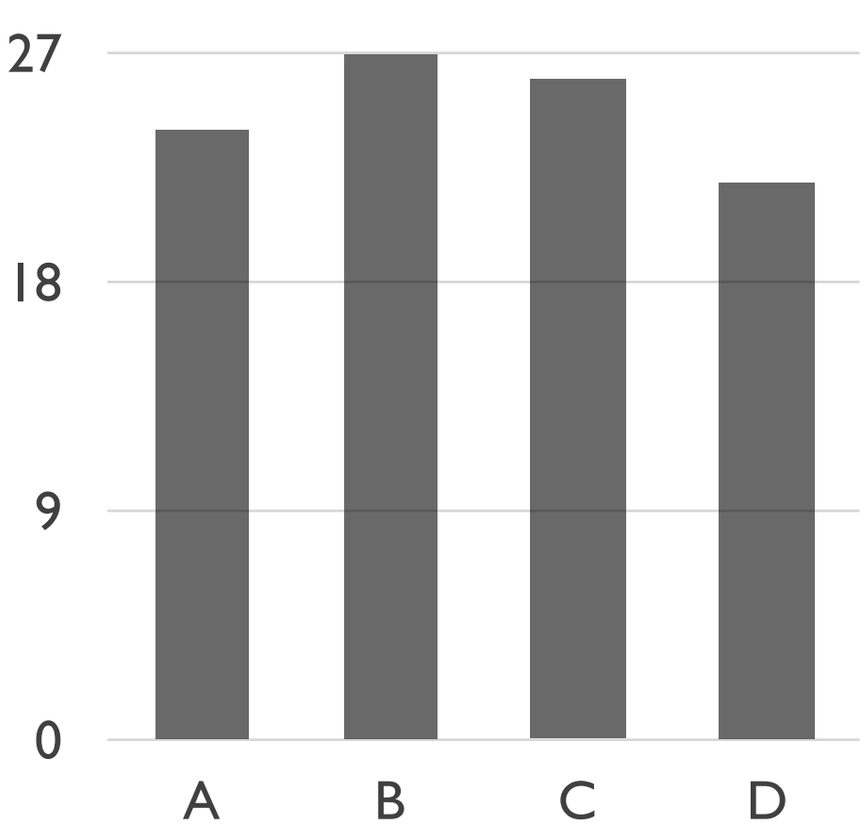
Using Rankings



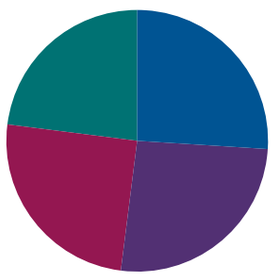
Using Rankings

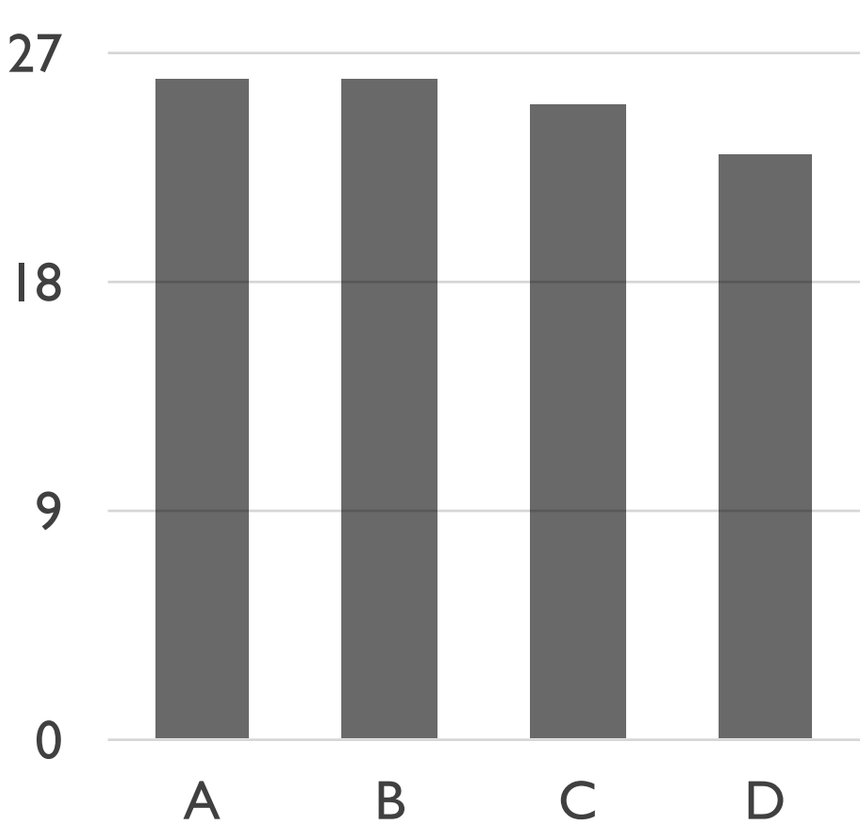
Year 1





Year 2





Bar Charts for Items & Categories

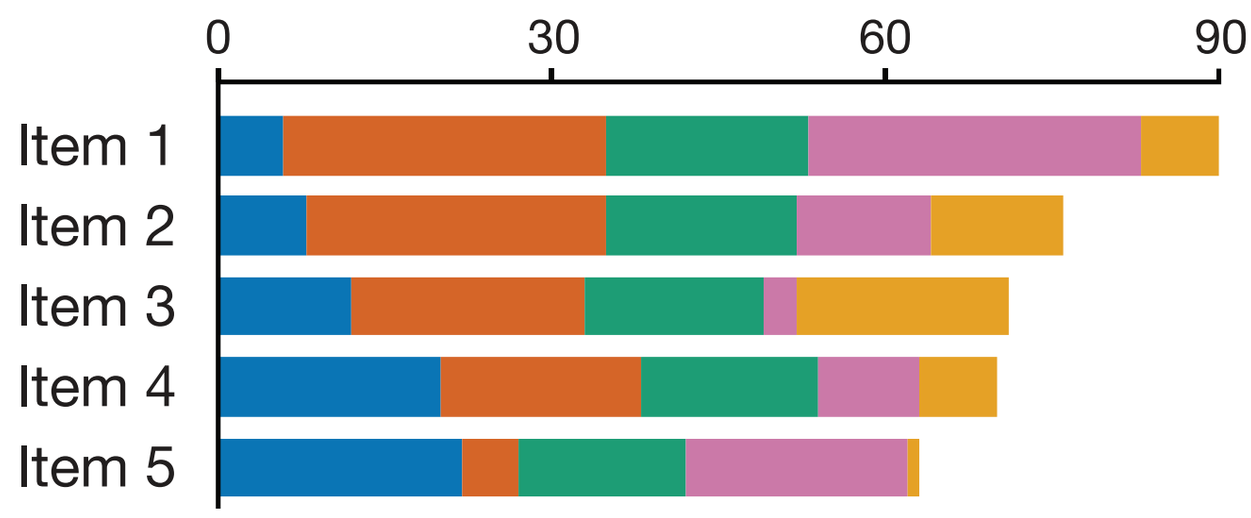
	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

Streit and Gehlenborg, Nature Methods, 2014

Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

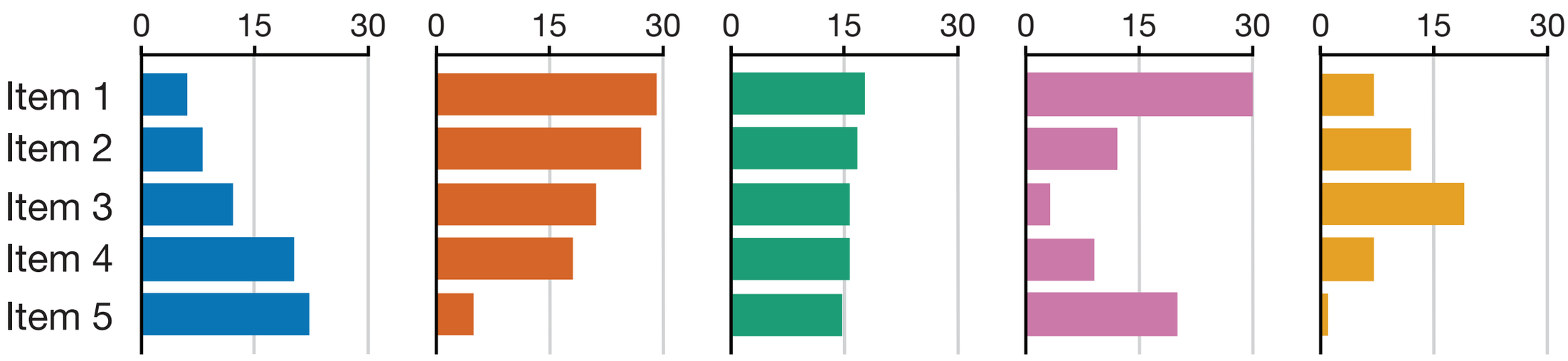
Stacked Bar Chart



Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

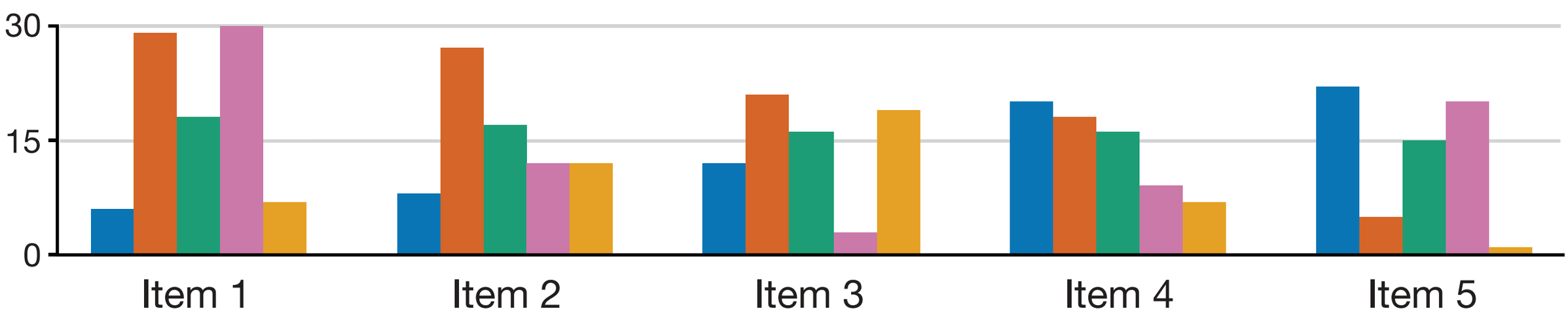
Layered Bar Chart



Bar Charts for Items & Categories

	1	2	3	4	5
Item	6	29	18	30	7
Item	8	27	17	12	12
Item	12	21	16	3	19
Item	20	18	16	9	7
Item	22	5	15	20	1

Grouped Bar Chart



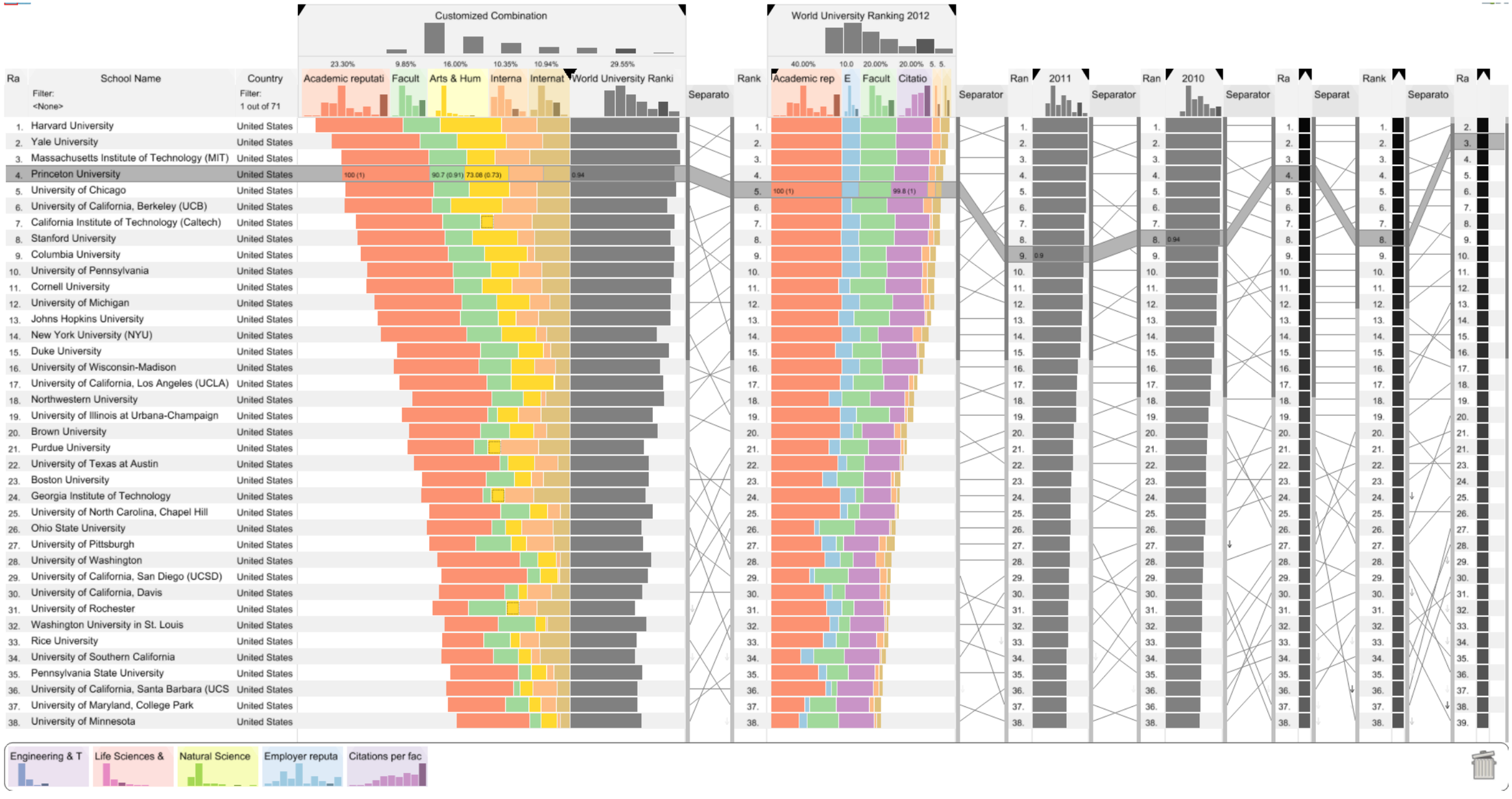
Streit and Gehlenborg, Nature Methods, 2014

Bar Charts for Items & Categories

- **Stacked Bar Chart**
  - if focus is on comparing the overall quantities across items but also need to illustrate contributions of each category to the total
- **Layered Bar Chart**
  - if focus is on distribution of values in each category across all items
  - comparisons within each category are more accurate than in stacked bar charts due to common baseline for the values in each category
- **Grouped Bar Chart**
  - if focus is on comparison of values across categories within each item while still enabling comparisons across items
  - if quantities add up to the same total for each item, then a grouped bar chart is equivalent to multiple pie charts, yet a grouped bar chart affords more accurate readings of values and comparisons

Streit and Gehlenborg, Nature Methods, 2014

LineUp: Ranking Visualization



http://lineup.caleydo.org, Gratzl, Lex, Gehlenborg, Pfister and Streit, IEEE Trans. Comput. Graph. Vis. (InfoVis 2013), 2013