

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**.

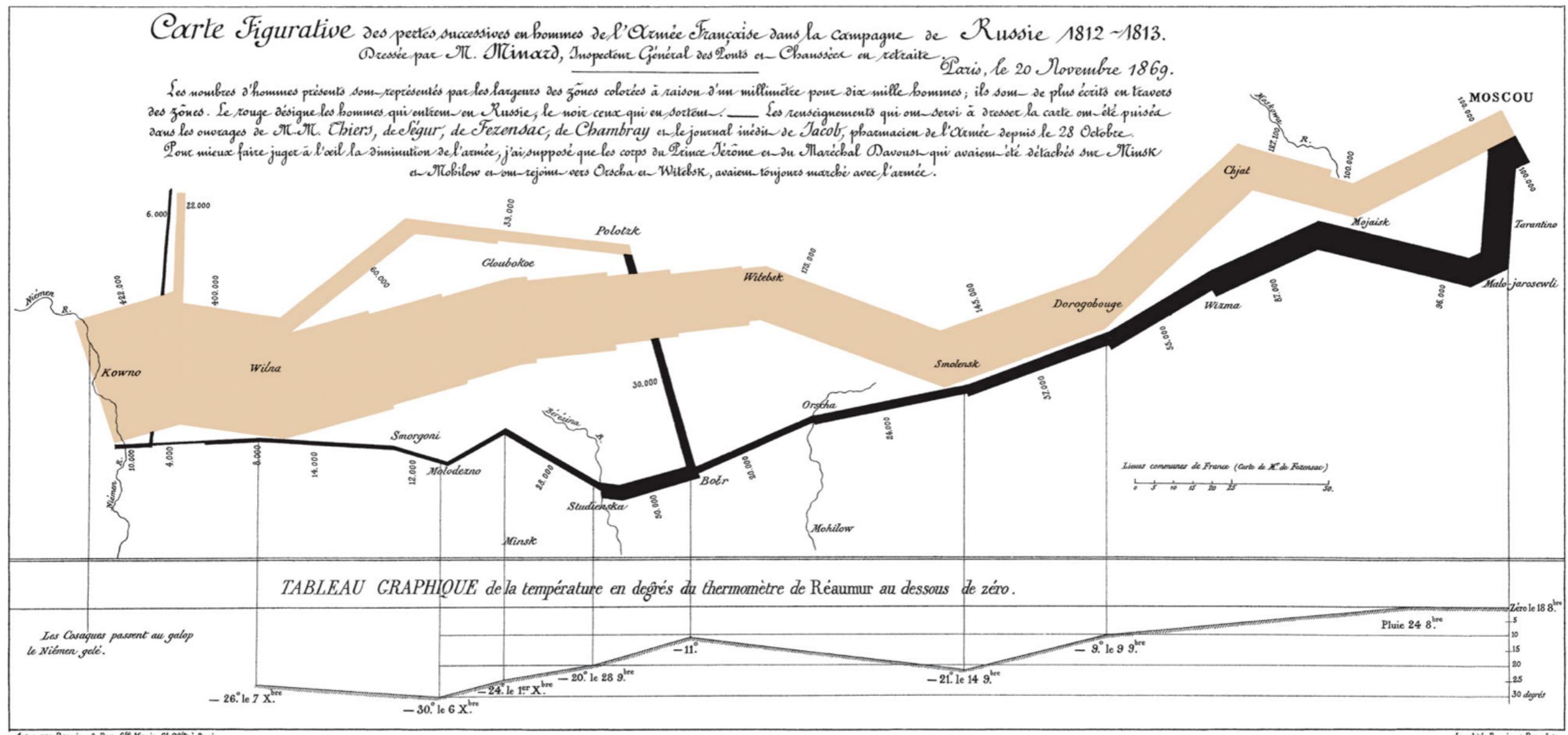
— Stu Card, Jock Mackinlay & Ben Shneiderman

Why do we need it?

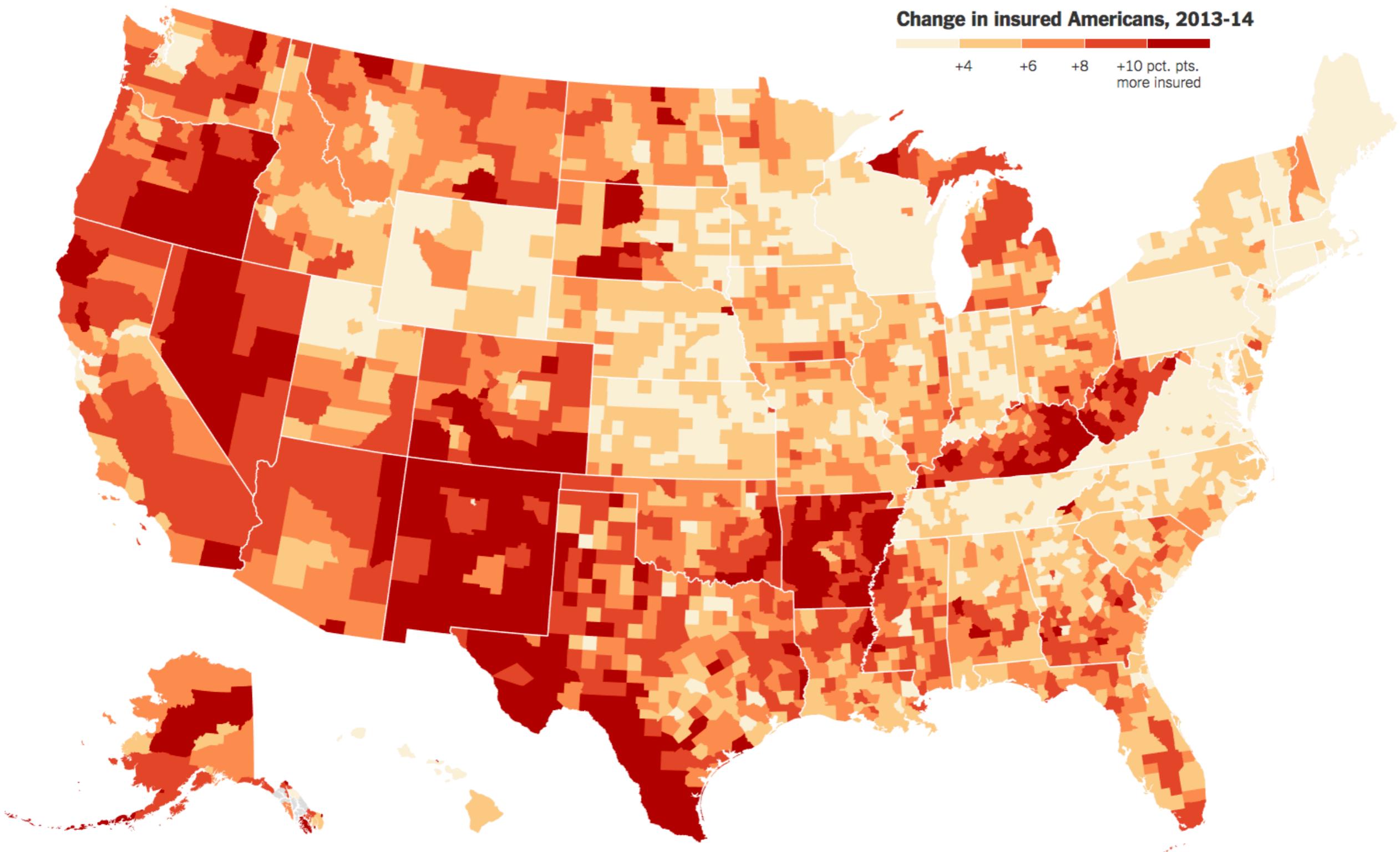
A good sketch is better than a long speech.

– Napoleon Bonaparte

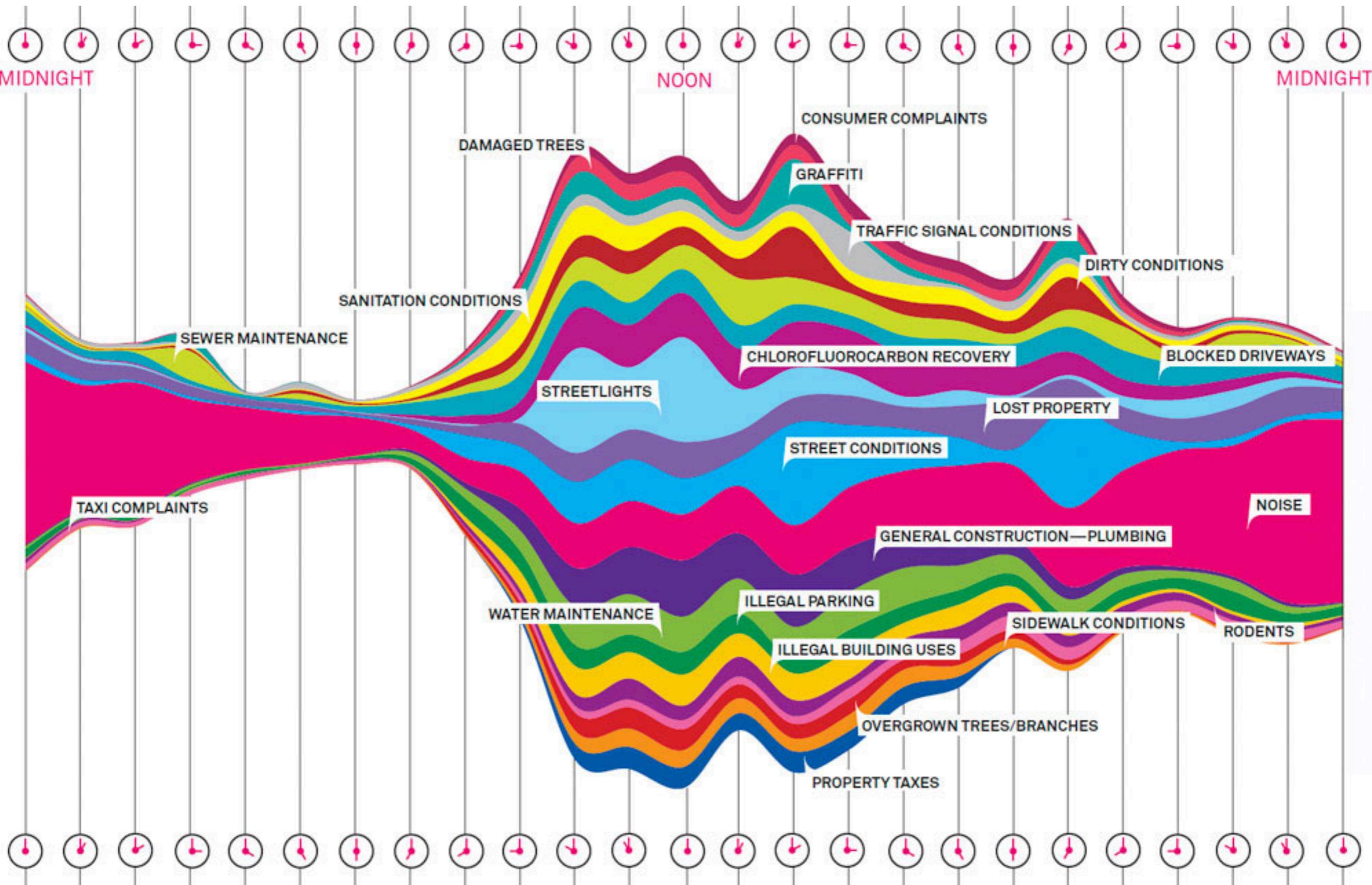
Charles Minard: Napoleon's March on Moscow



New York Times: ACA Health Insurance Gains 2013 - 2014



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



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 $\sigma_{X_i}^2 = 11.0$, $\mu_{Y_i} = 7.5$, $\sigma_{Y_i}^2 = 4.12$, correlation $\text{cor}(X_i, Y_i) = 0.816$ 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 |

$\text{mean}(X) = 9$, $\text{var}(X) = 11$, $\text{mean}(Y) = 7.5$, $\text{var}(Y) = 4.12$,
 $\text{cor}(X,Y) = 0.816$, linear regression line $Y = 3 + 0.5*X$

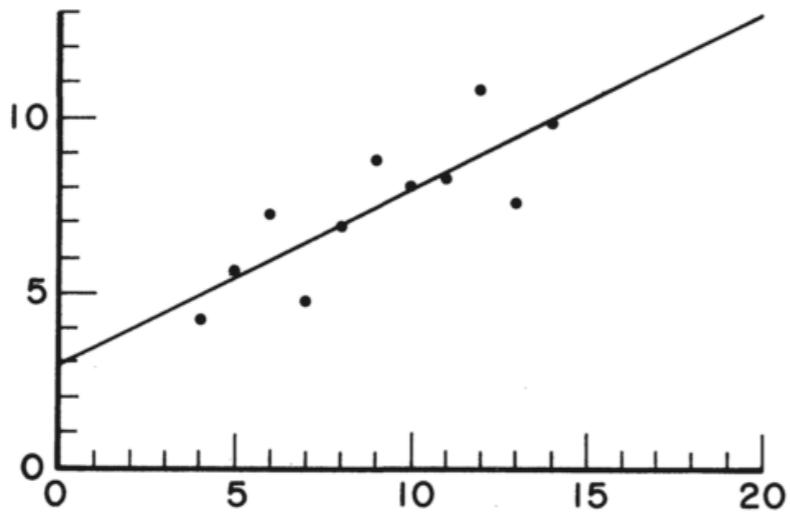


Figure 1

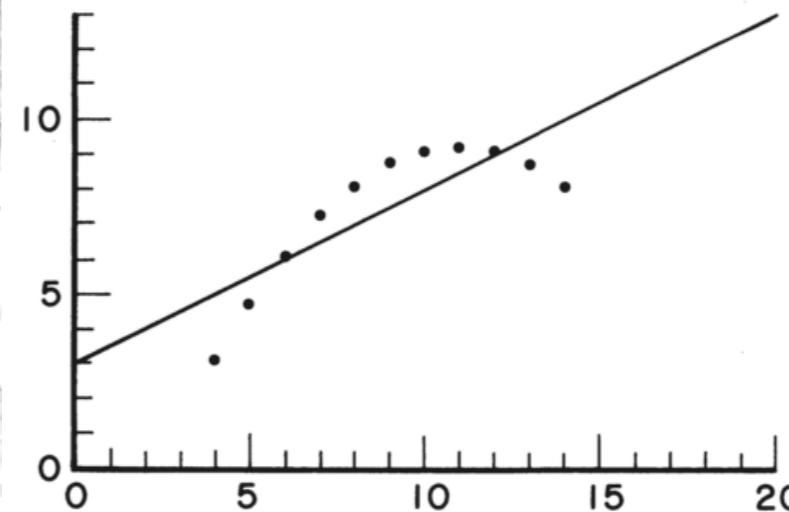


Figure 2

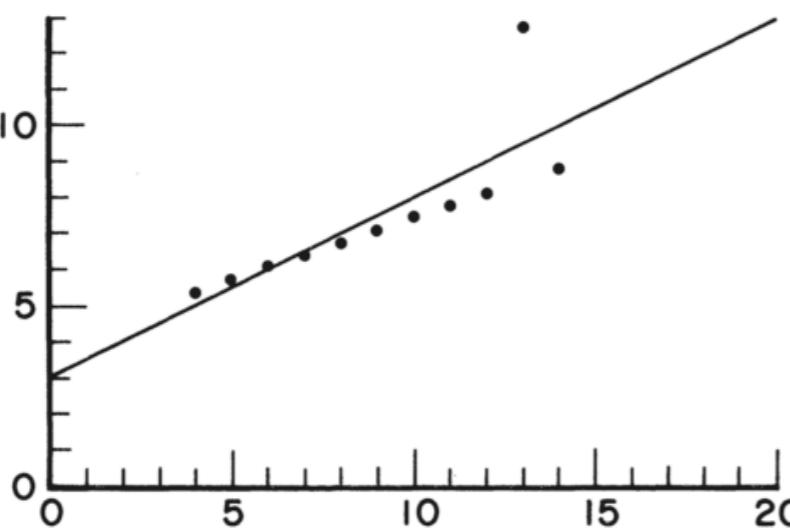


Figure 3

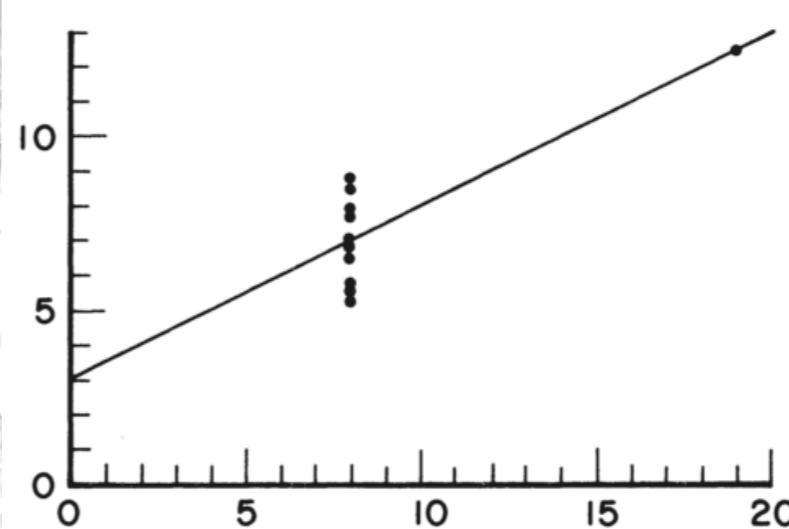


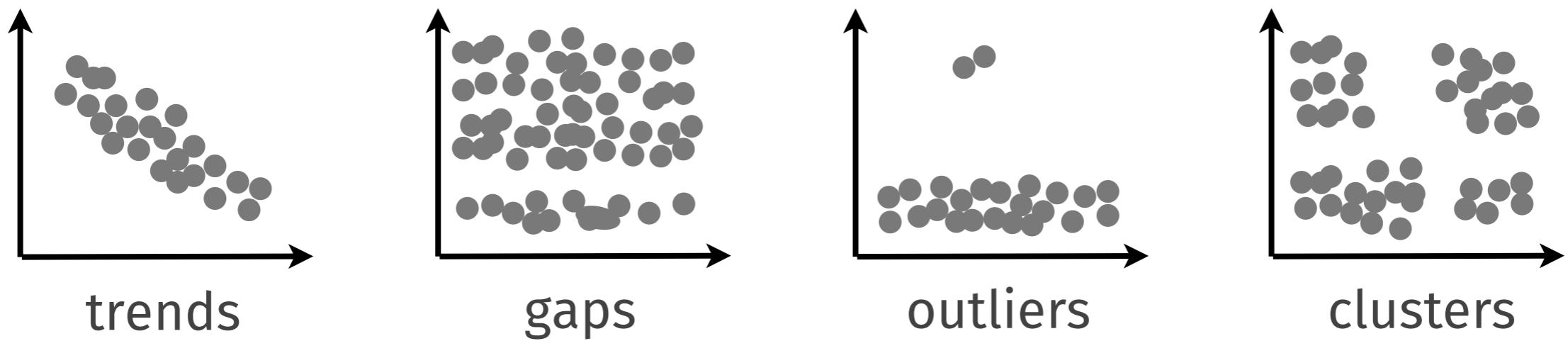
Figure 4

Why do we need it?

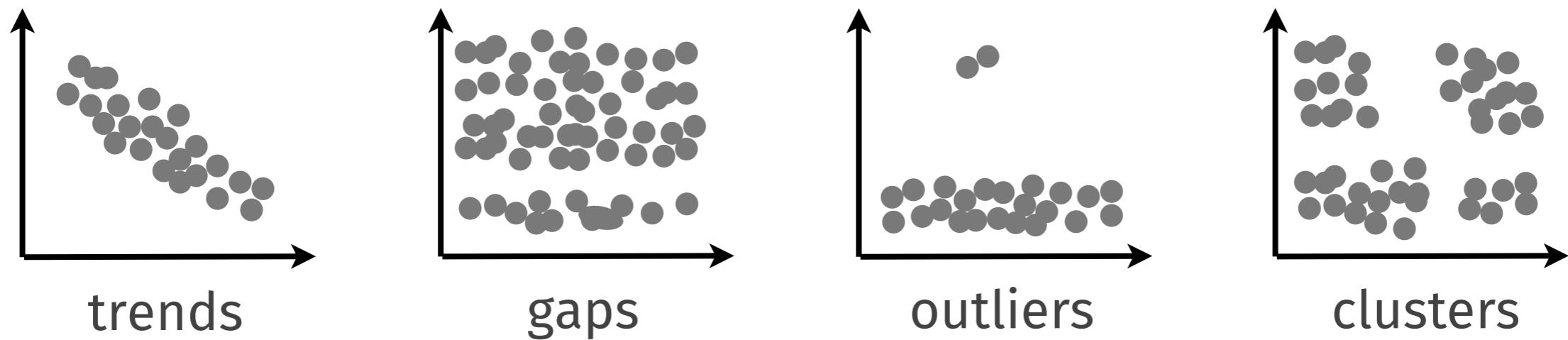
I'm wondering if there are any
interesting patterns in my data.

—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

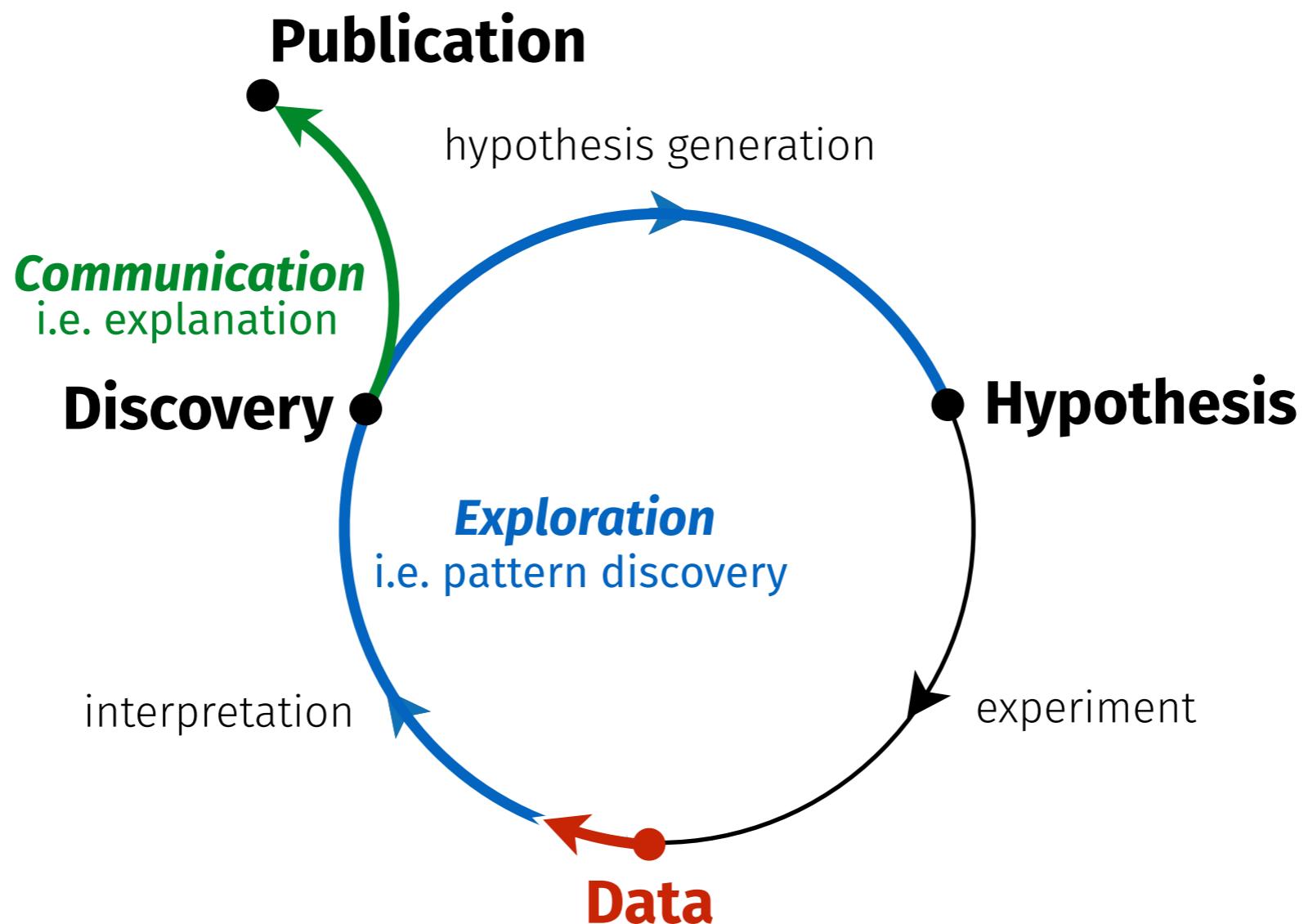
Visualization Use Cases

Exploration

Confirmation

Communication

Discovery Process



Discovery Process

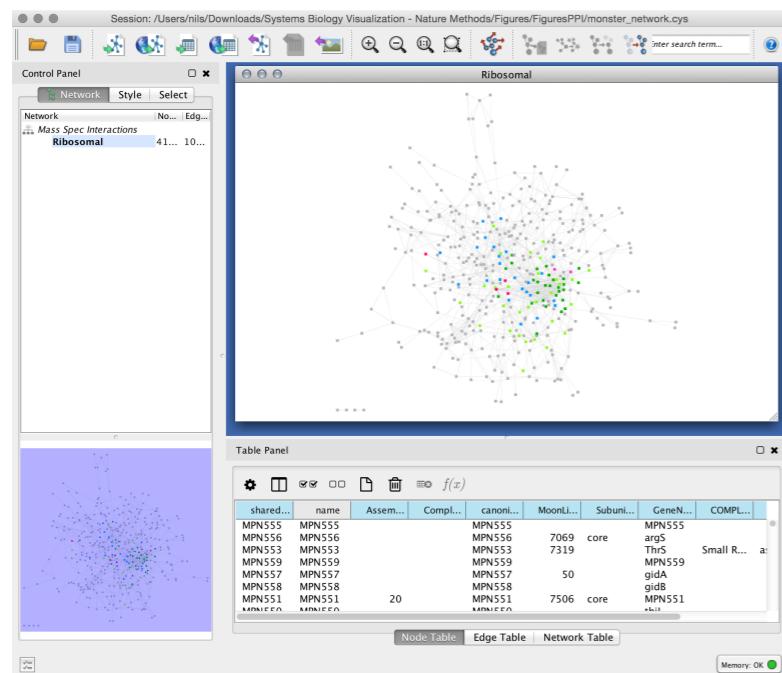


Discovery Process

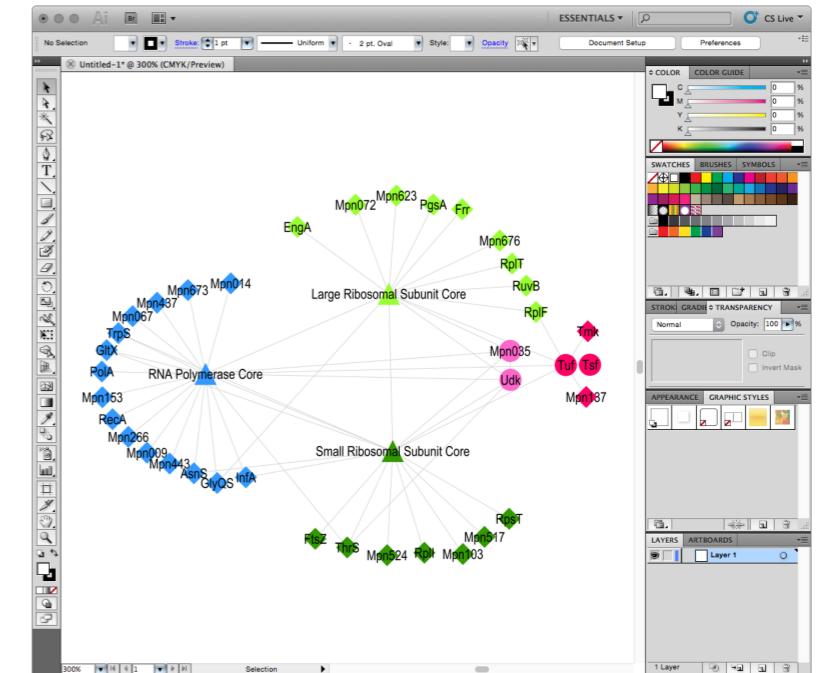
Exploration

Communication

Insight

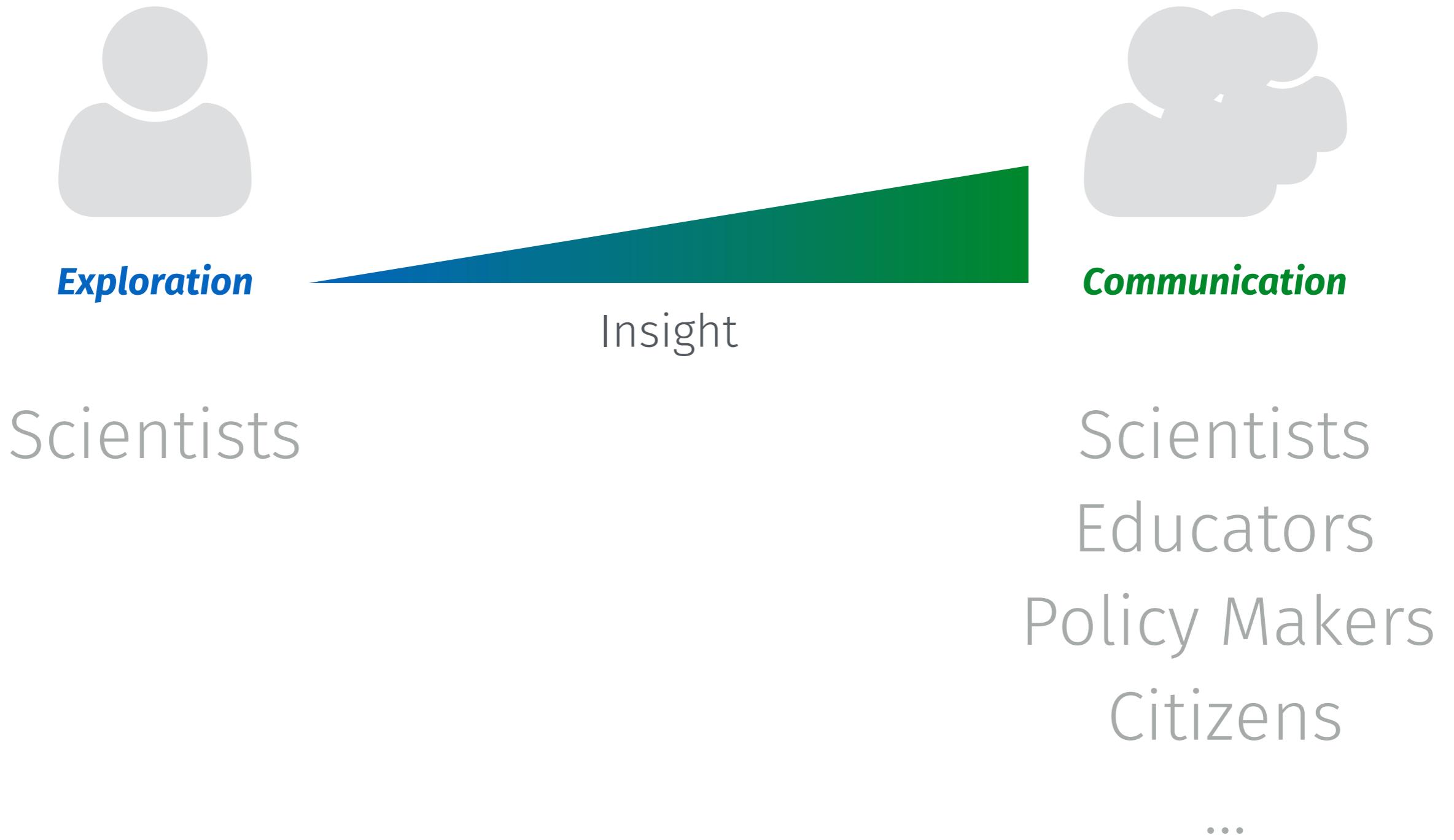


Cytoscape

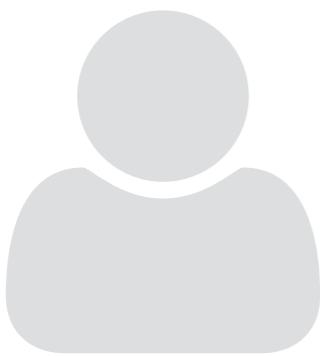


Adobe Illustrator

Discovery Process



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

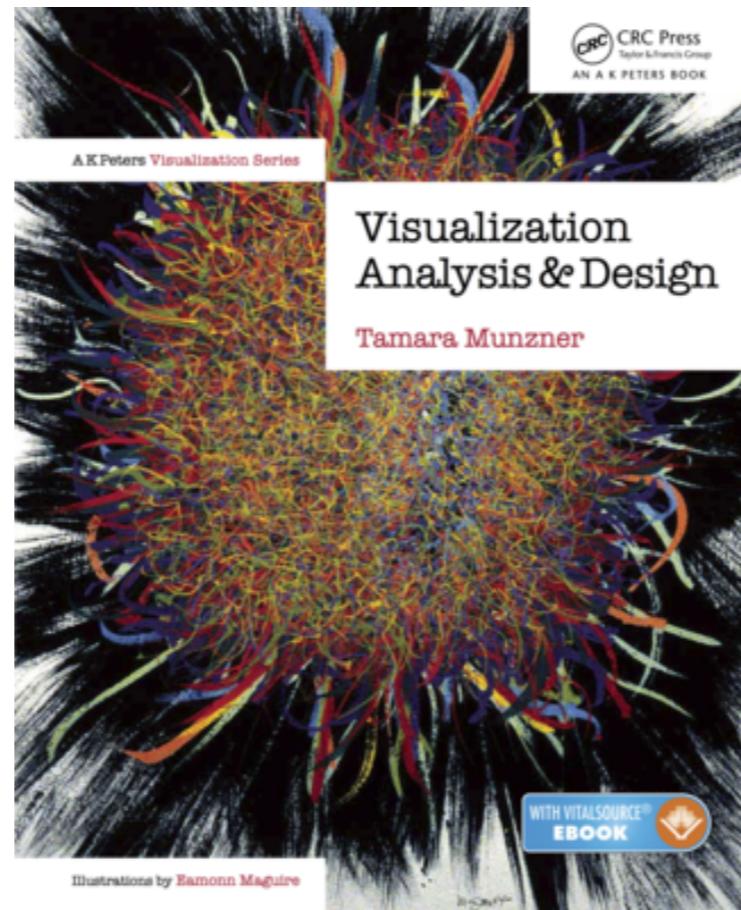
Syllabus

Piazza

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

Reading

Required Textbook



Tamara Munzner, 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 Kerpeljiev, 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

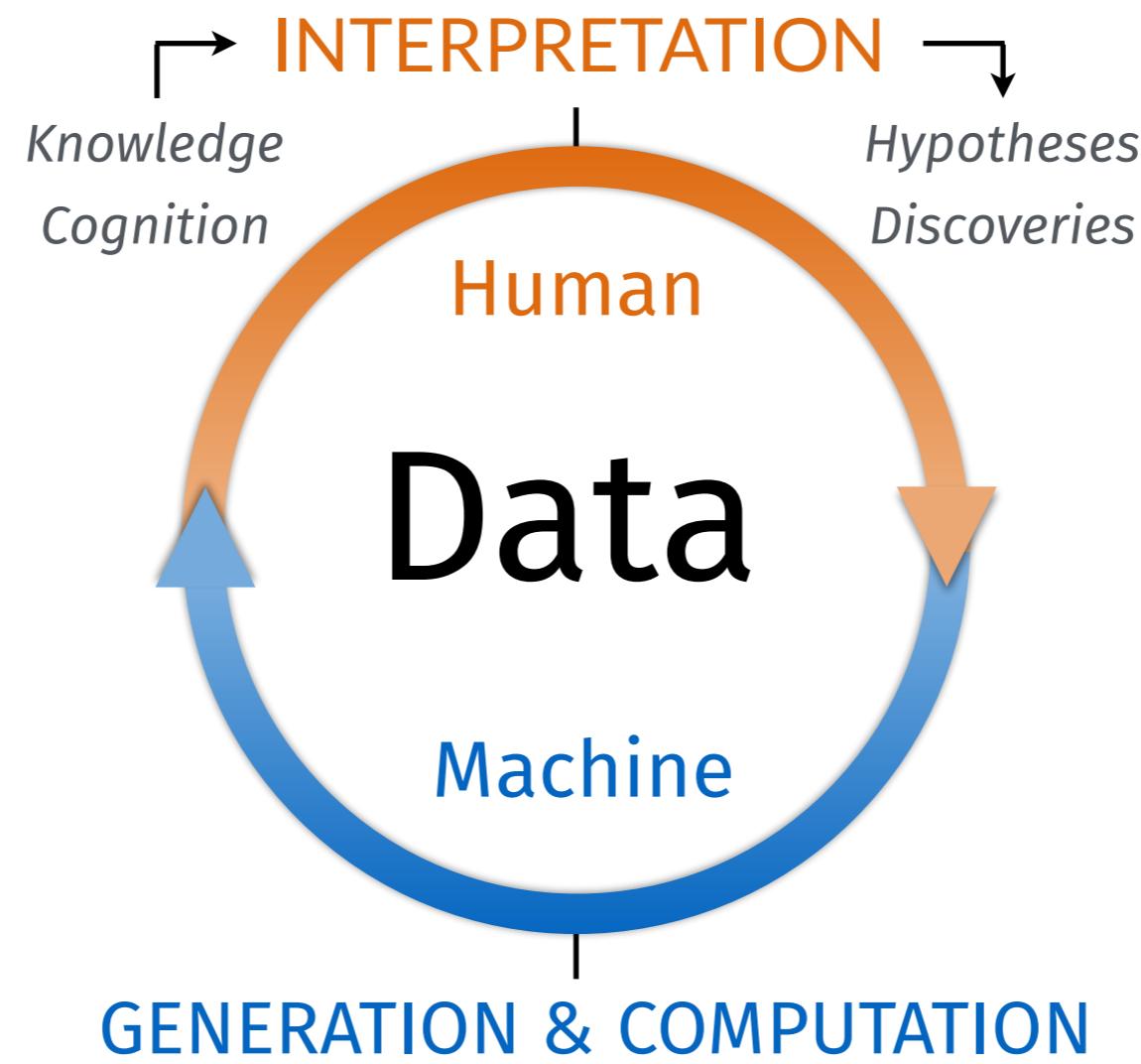
Reproducible Research

Applications

Cancer Genomics

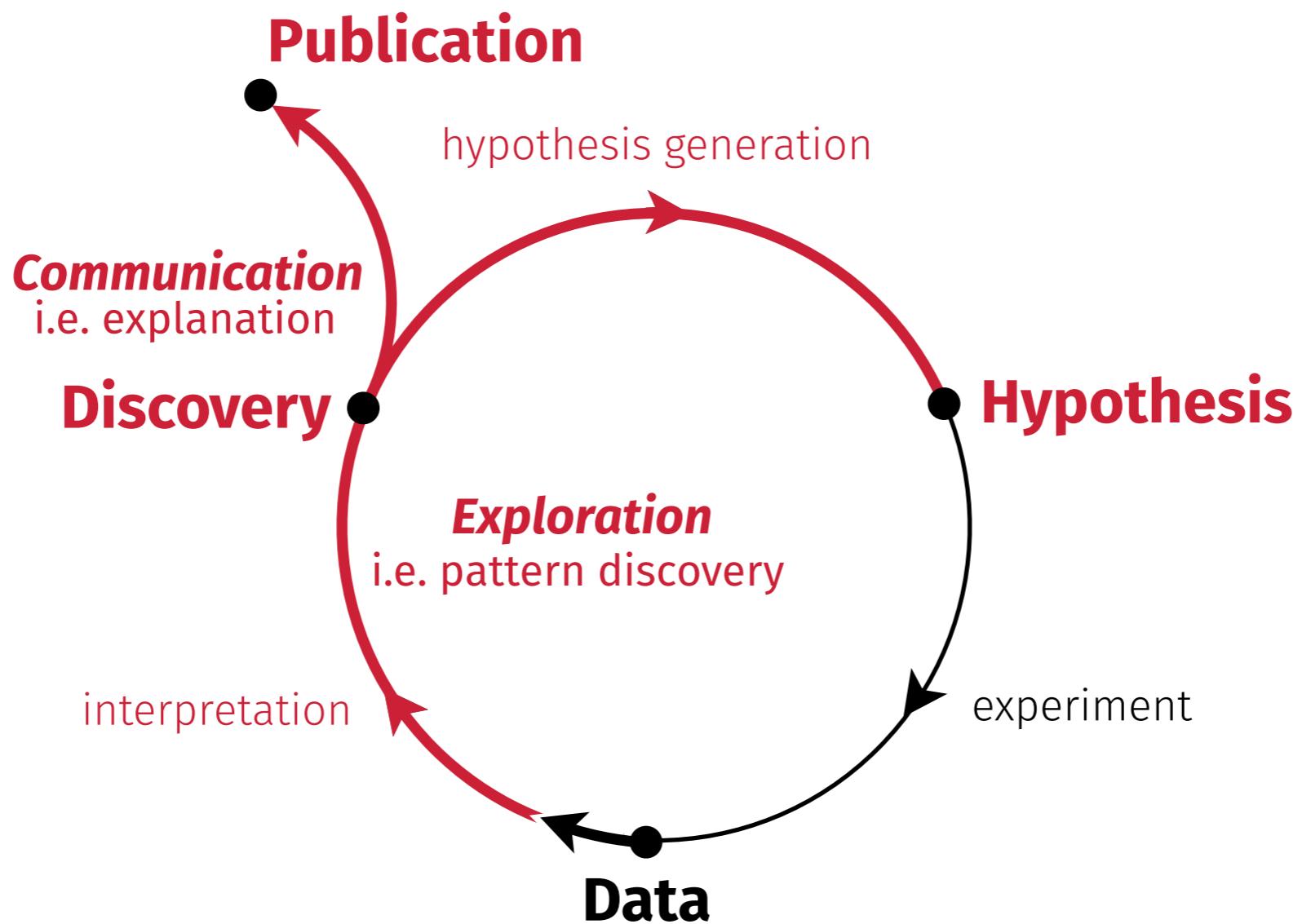
3D Genome Structure

Methods for Data Visualization and Exploration

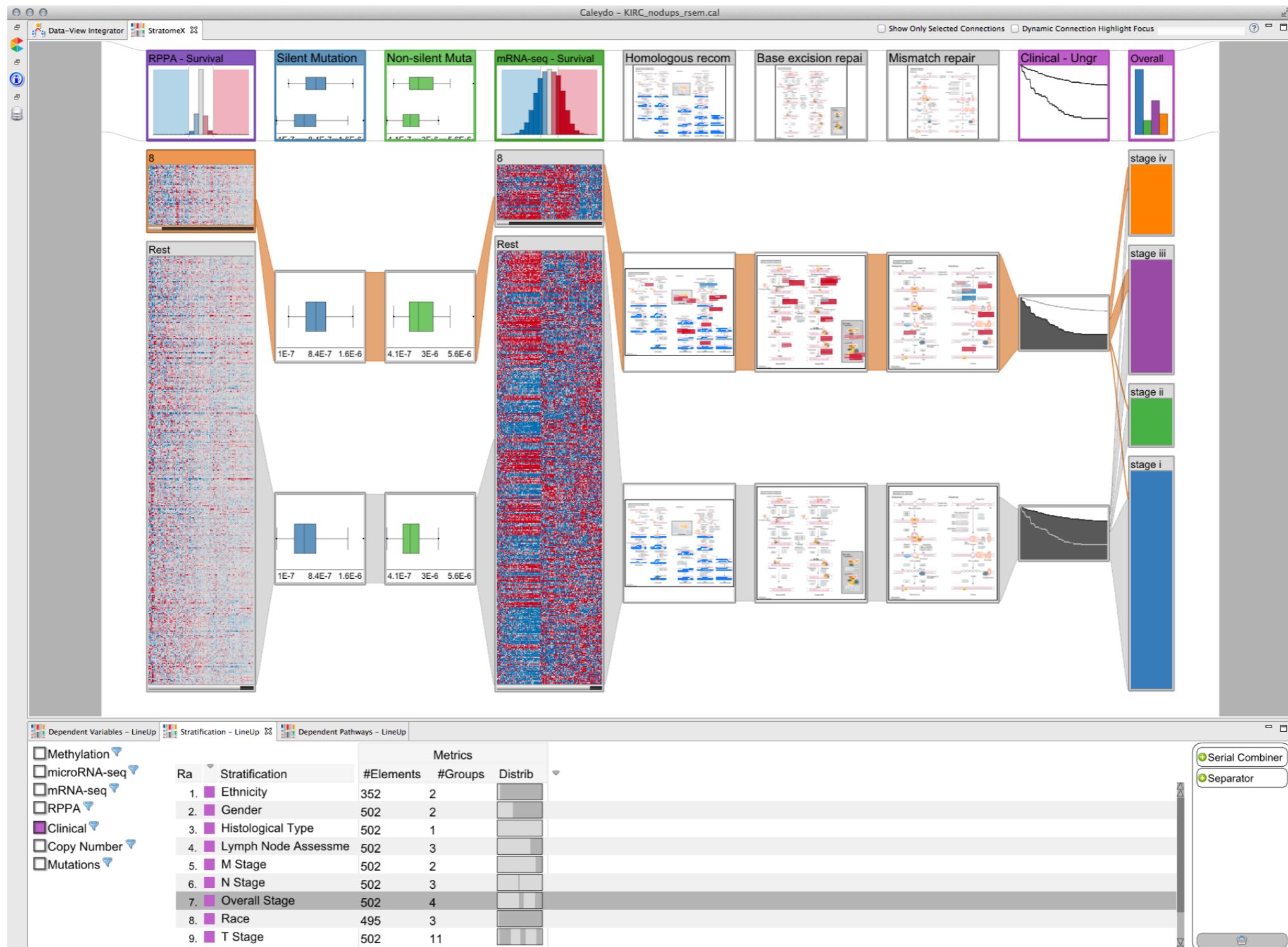


Tools for Reproducible Research

Spectrum of Projects

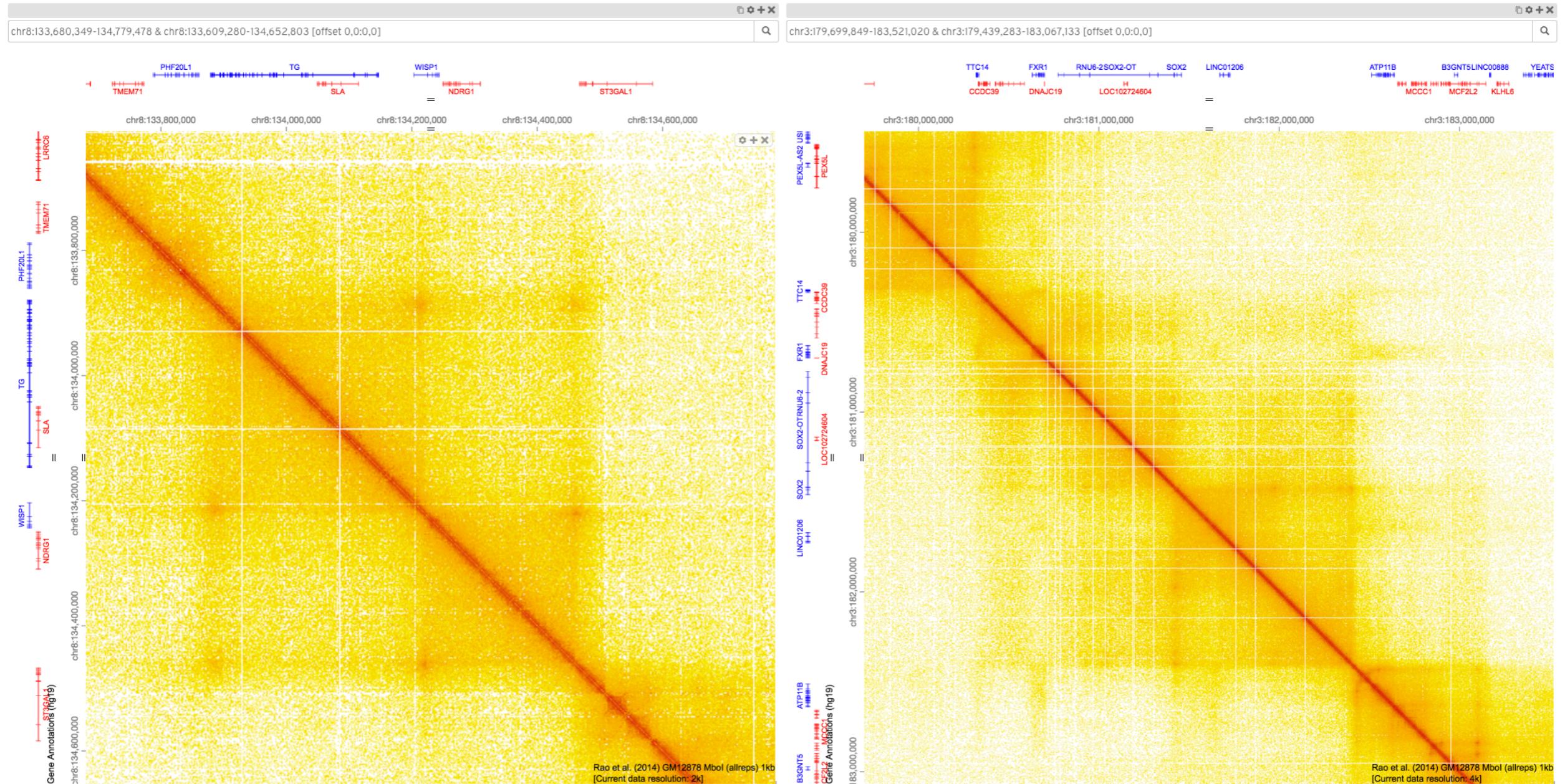


Methods for Data Visualization and Exploration



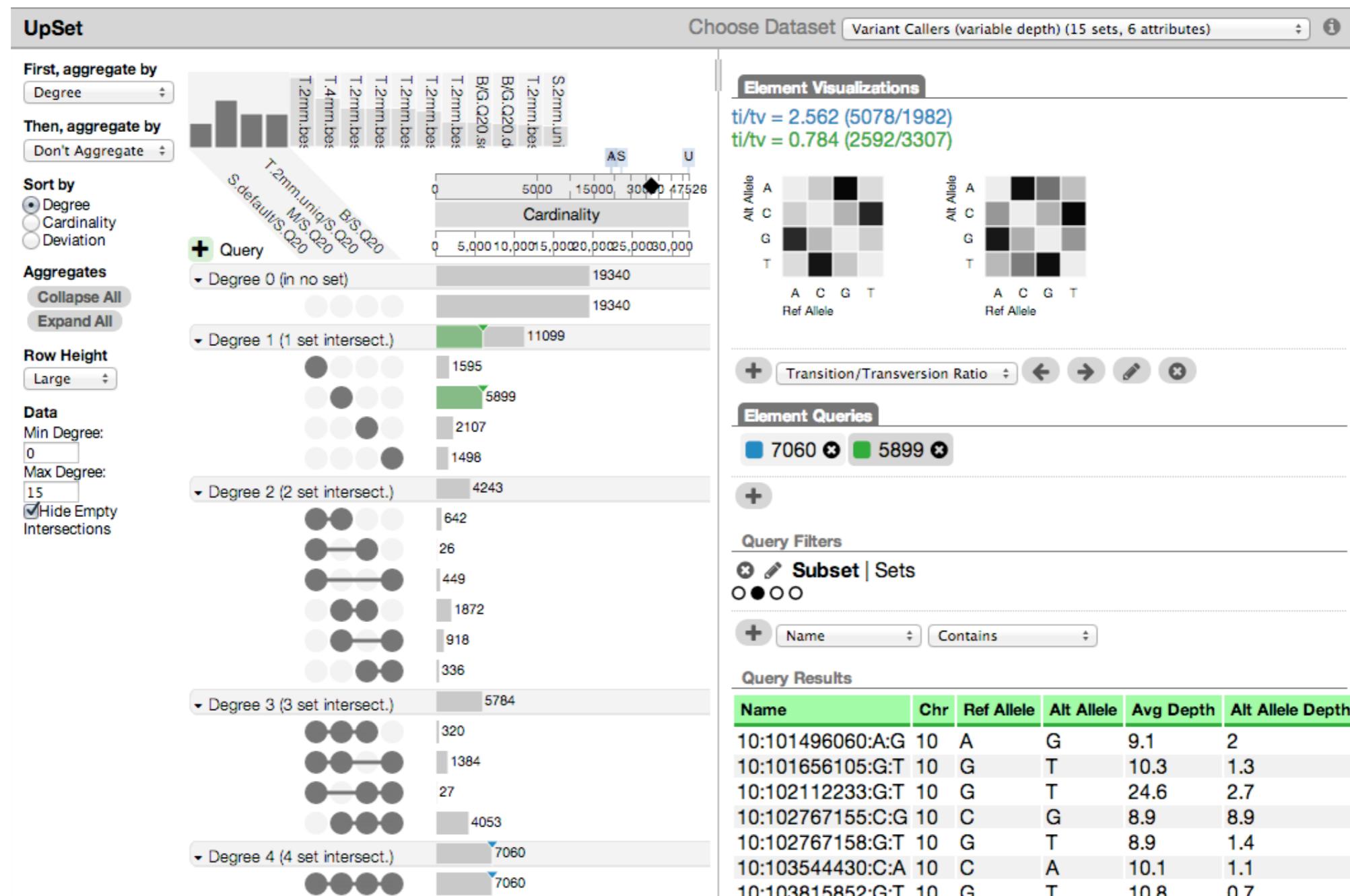
StratomeX - Cancer Subtypes: <http://stratomex.caleydo.org>

Methods for Data Visualization and Exploration



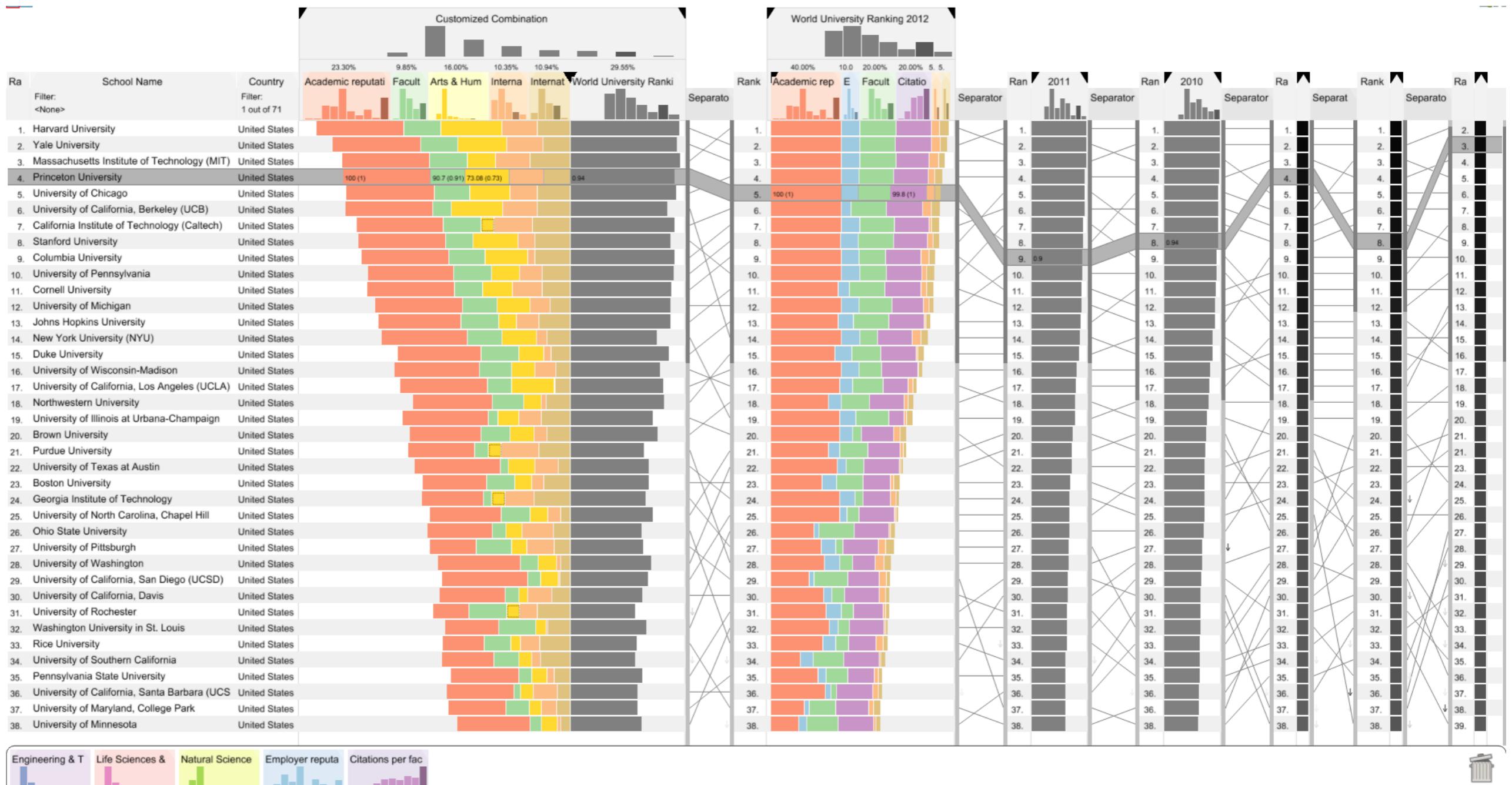
HiGlass - Visualization of Chromatin Interaction Maps: <http://higlass.io>

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

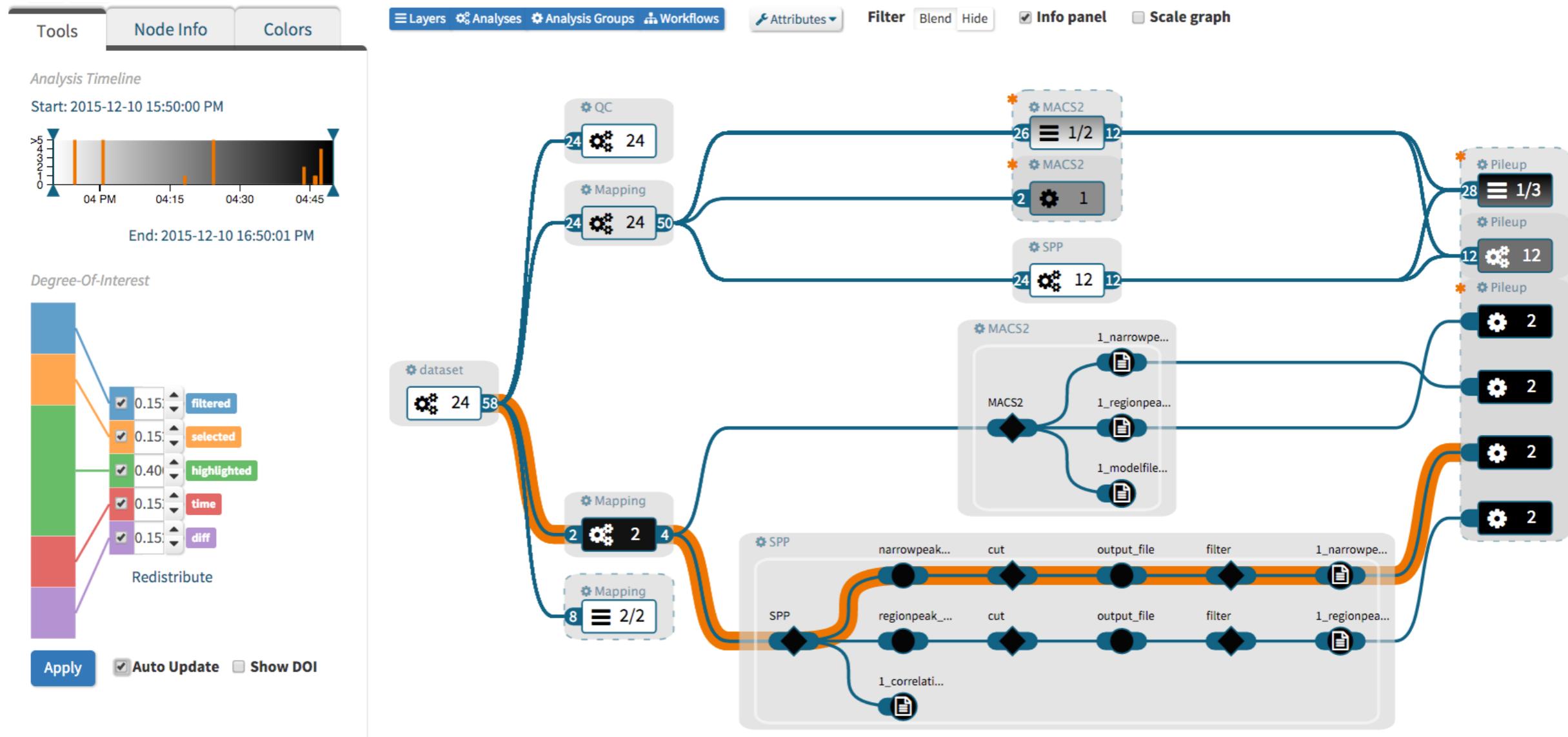
The screenshot shows the Stem Cell Commons Refinery Platform interface. At the top, there is a navigation bar with links for Stem Cell Commons, Collaboration, Statistics, About, and user authentication (Nils Gehlenborg | Logout). Below the navigation bar, the page title is "Launch Pad Nils Gehlenborg".

The main content area is divided into three sections:

- Data Sets:** This section lists 206 data sets. Some examples include:
 - An Alternative Splicing Switch Regulates Embryonic Stem Cell Pluripotency and Reprogramming [RNA-Seq]
 - Droplet barcoding for single cell transcriptomics applied to embryonic stem cells
 - NKX2-1 occupancy in human lung adenocarcinoma cell lines
 - Transcription factor ChIP-seq in expanded human hematopoietic stem and progenitor cells
 - Cardiac transcription factors in HL-1 cells: genome binding profiling
 - Gene expression analysis of cdx4, sall4, and cdx4+sall4 morpholino(s) injected embryos at 3-somite stage
 - Epigenetic profiling of WT and Ezh2-null MLL-AF9 murine leukemic cells
 - Genome wide uH2A localization analysis highlights Bmi1-dependent deposition of the mark at repressed genes.
 - Cdx2 transcription factor binding in intestinal villus and gene expression profiling in Cdx mutant mice
 - Cell-Type-Specific TGF-beta Signaling is Targeted to Genes that Control Cell Identity: ChIP-Seq (mouse)
 - Mapping polycomb complexes in human and mouse embryonic stem cells (mouse)
 - Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and factor binding (ChIP-seq)
- Analyses:** This section lists 60 analyses. Examples include:
 - Test workflow: 5 steps no delay 2016-1-15@9:39:07
 - FastQC 2016-1-15@9:17:49
 - FastQC 2016-1-14@11:33:31
 - FastQC 2016-1-13@21:10:53
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-23@10:53:37 - 3 pairs
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-22@17:39:12
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-18@14:20:51 - 3 pairs
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-18@12:29:15 - 13880/3 pairs
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-17@15:50:05
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-16@17:11:52 - new bowtie
 - TF ChIP-Seq analysis using MACS2: danRer7 2015-9-16@13:01:33
 - FastQC 2015-9-16@12:39:55
 - FastQC 2015-9-16@12:20:18
 - FastQC 2015-9-16@11:06:31
 - MACS2 Demo
 - Oct4 FastQC
 - My release test!
- Workflows:** This section lists 5 workflows. Examples include:
 - FastQC
 - TF ChIP-Seq analysis using MACS2: danRer7
 - TF ChIP-Seq analysis using MACS2: hg19
 - TF ChIP-Seq analysis using MACS2: mm10
 - Test workflow: 5 steps no delay

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

Tools for Reproducible Research



AVOCADO Provenance Visualization
Refinery Platform: <http://refinery-platform.org>

Tools for Reproducible Research

Data Sets

Search

199 data sets

- cloche s5 point mutation (zebrafish)
- Genome-wide location analysis of BMP (SMAD1) in mouse erythroid progenitors co-occupied with lineage..
- Genome-wide maps of binding sites of Nanog-like and Mtx2 in blastula stage zebrafish embryos
- Comparative profiling of chromatin state maps and transcription factor occupancy during human fetal an..
- RNA sequencing of circulating tumour cells implicates WNT signaling in pancreatic cancer metastasis (mous..)
- Pseudo-temporal ordering of individual cells reveals regulators of differentiation
- Nanog Independent Reprogramming to iPSCs with Canonical Factors
- Mapping polycomb complexes in human and mouse embryonic stem cells (human)
- Identification of genes regulated by Rcor1 in CD71+,

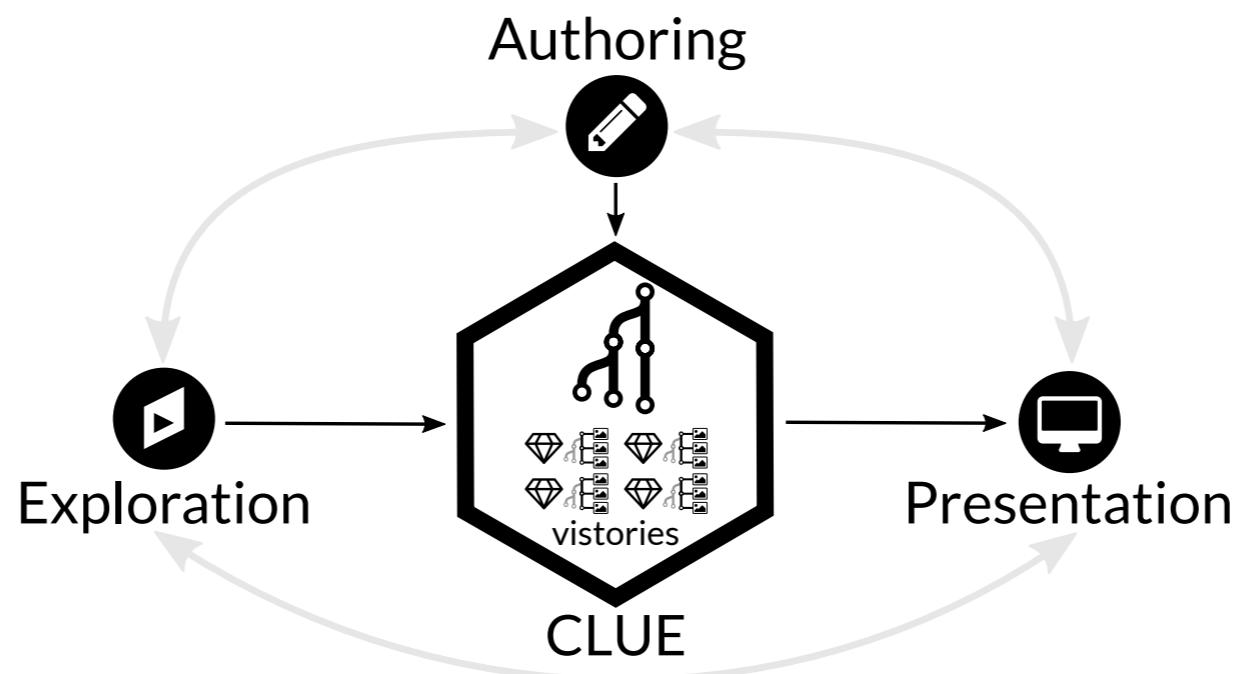
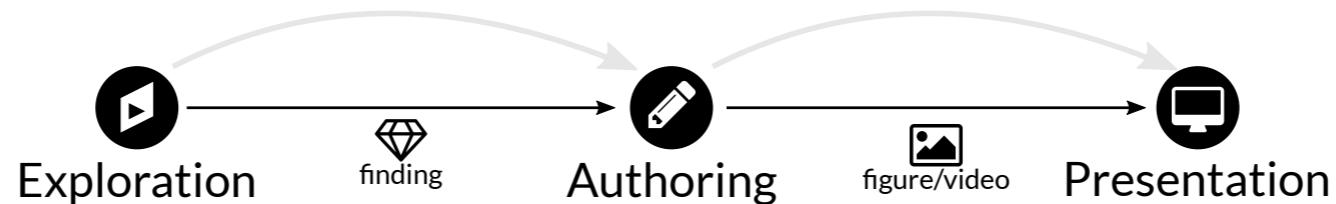
Repository exploration

PRECISION RECALL NAME ONE BAR TWO BARS ZOOM OUT Highlight: Lock: Query: + or

The interface displays a semantic network graph on the right side, where nodes represent biological concepts like 'Organism', 'Assay', 'Chemical compound', etc., and edges show their relationships. Below the graph is a heatmap with three main columns: 'Root' (Organism, Assay), 'Chemical compound', and 'Cellular component'. The 'Root' column has two rows: 'Organism' and 'Assay'. The 'Chemical compound' column has one row: 'Native cell'. The 'Cellular component' column has one row: 'Vertebrata'. Underneath these are smaller sub-categories: 'Unit', 'Neoplasm', 'Cell line', 'Age', 'Organism', 'Assay', 'Chemical compound', 'Cellular component', 'Native cell', 'Vertebrata', 'Unit', 'Neoplasm', 'Cell line', and 'Age'. The entire interface is titled 'SATORI Ontology-Guided Repository Exploration'.

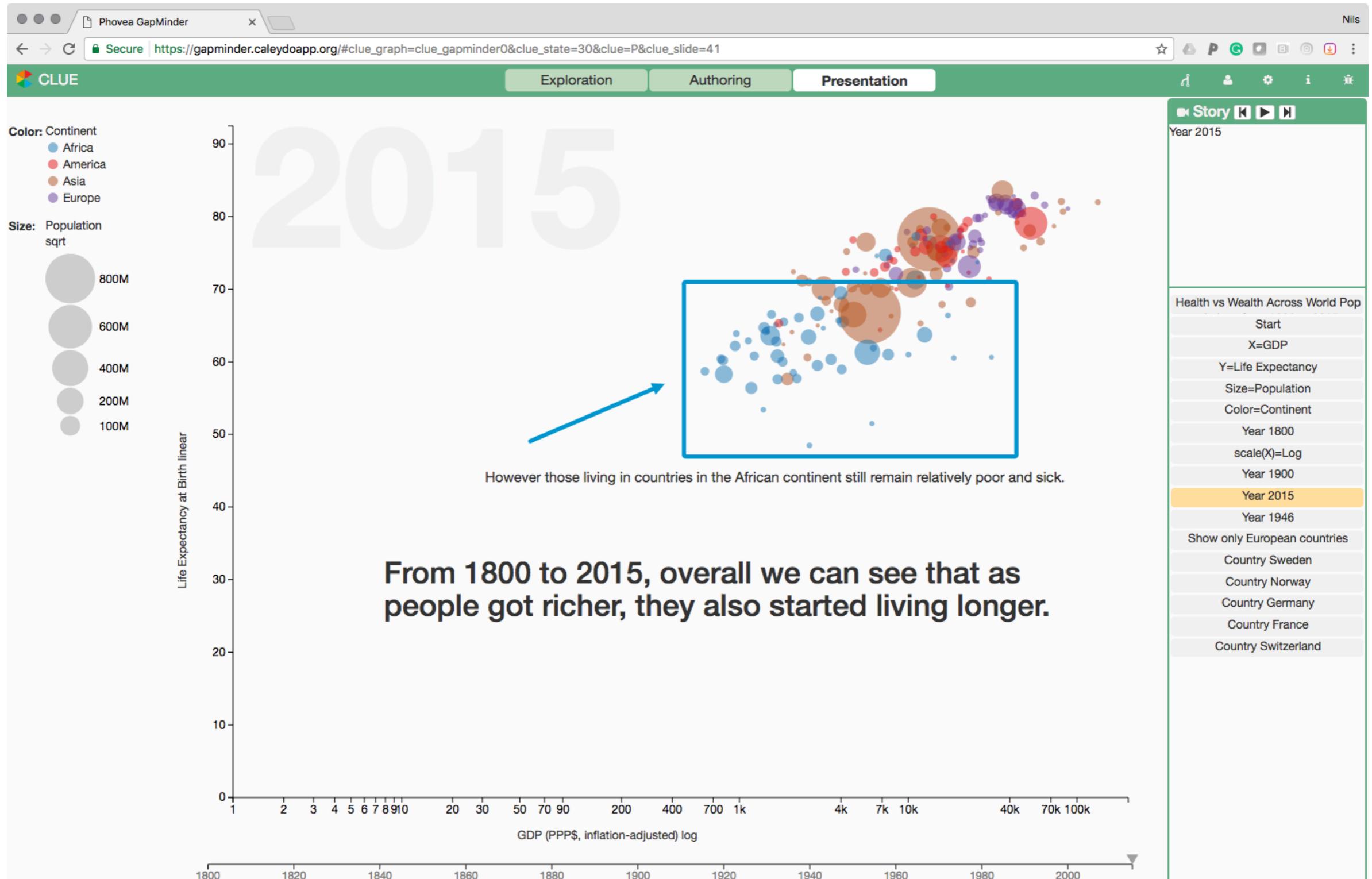
SATORI Ontology-Guided Repository Exploration
Refinery Platform: <http://satori.refinery-platform.org>

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?

MALWMRLLPLALLALWGPDPA
AAFVNQHLCGSHLVEALYLVCG
ERGFFFYTPKTRREAEDLQVGQV
ELGGPGAGSLQPLALEGSLQK
RGIVEQCCTSICSLYQLENYCN

How does visualization work?

MALWMRLLPLLLALLALWGDPA
AAFVNQHLCGSHLVEALYLVCG
ERGFFFYTPKTRREAEDLQVGQV
ELGGGPAGSLQPLALEGSLQK
RGIVEQCCTSICSLYQLENYCN

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?

$$453 \times 862 = ?$$

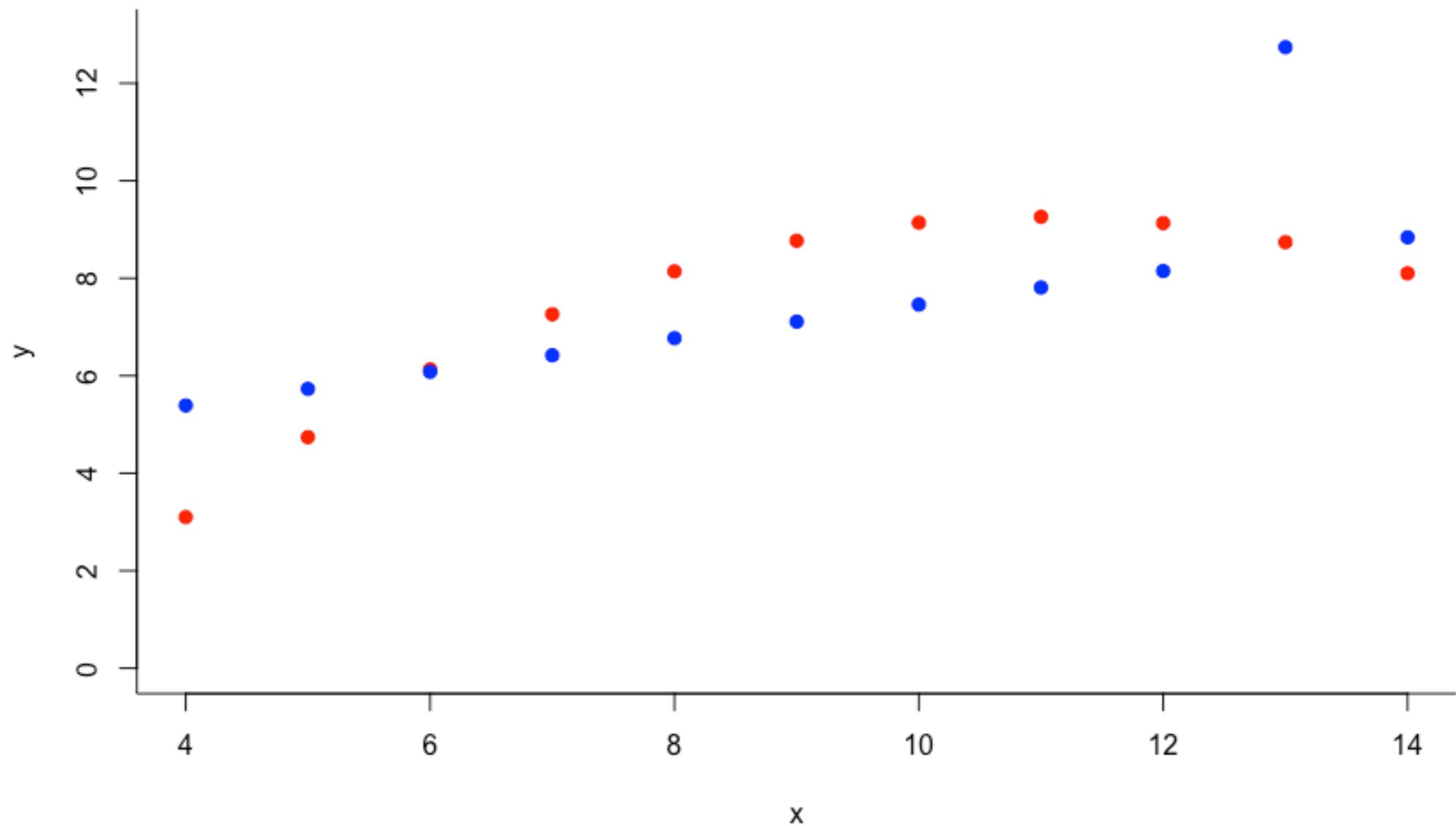
How does visualization work?

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

How does visualization work?

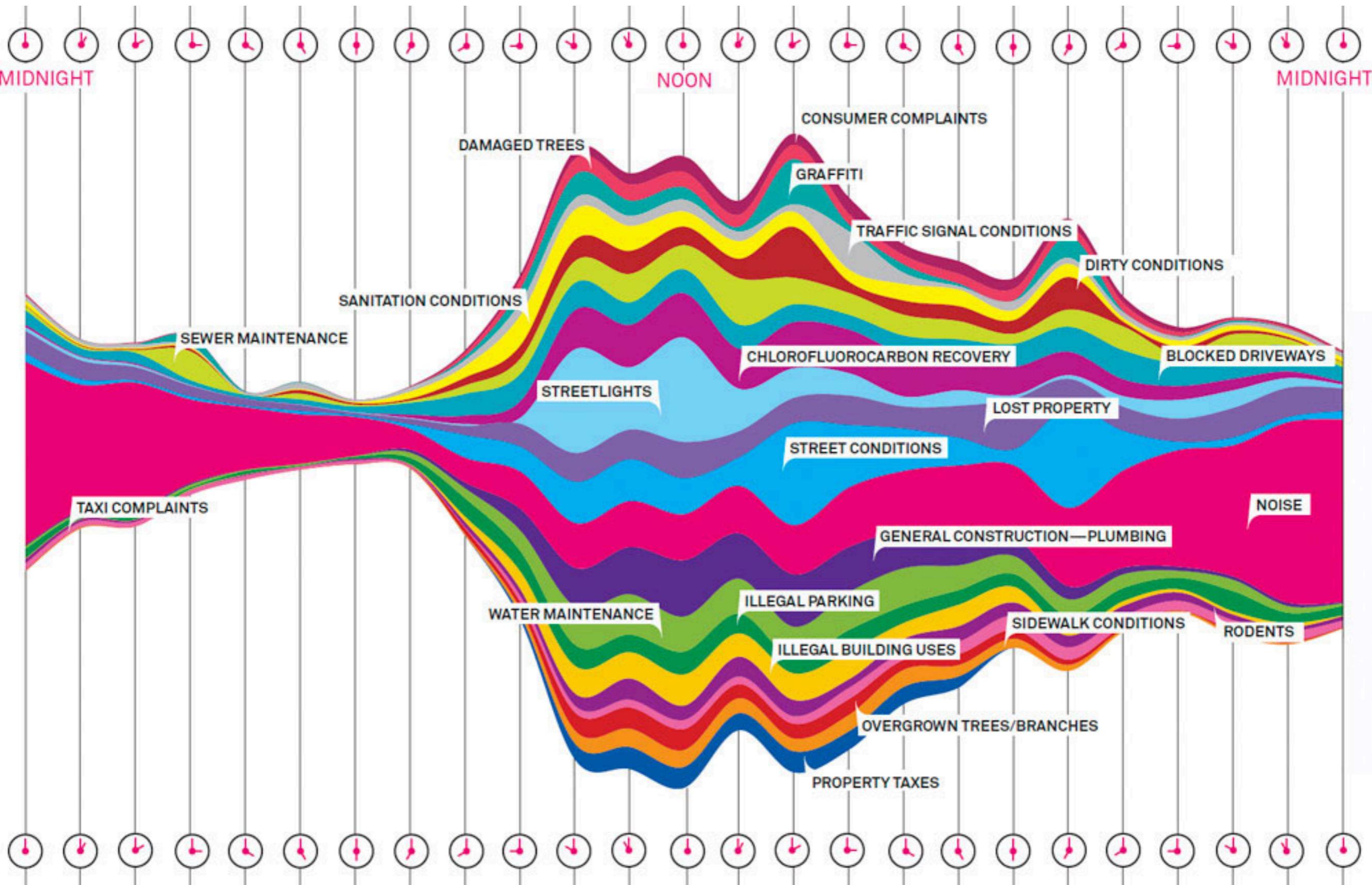
| | f_1 | | f_2 | | Tasks |
|----|-------|------|-------|-------|--|
| | x | y | x | y | |
| 1 | 10 | 9.14 | 10 | 7.46 | What is the shape of f_1 ? Of f_2 ? |
| 2 | 8 | 8.14 | 8 | 6.77 | |
| 3 | 13 | 8.74 | 13 | 12.74 | How many times do f_1 and f_2 intersect? |
| 4 | 9 | 8.77 | 9 | 7.11 | Do they cross 0? |
| 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 | |

How does visualization work?



Design Critique

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



Visual Encoding of Data

data

apples
oranges
bananas

small
medium
large

10 inches
13 inches
18.5 inches

trees
networks

*intrinsic
position*

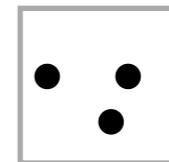
Visual Channels: Rankings

Categorical

What? Where?



position*
planar



color hue



shape

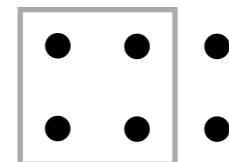


Relational

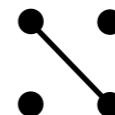
With whom?



containment



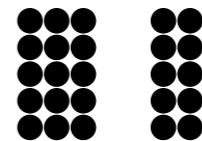
connection



similarity

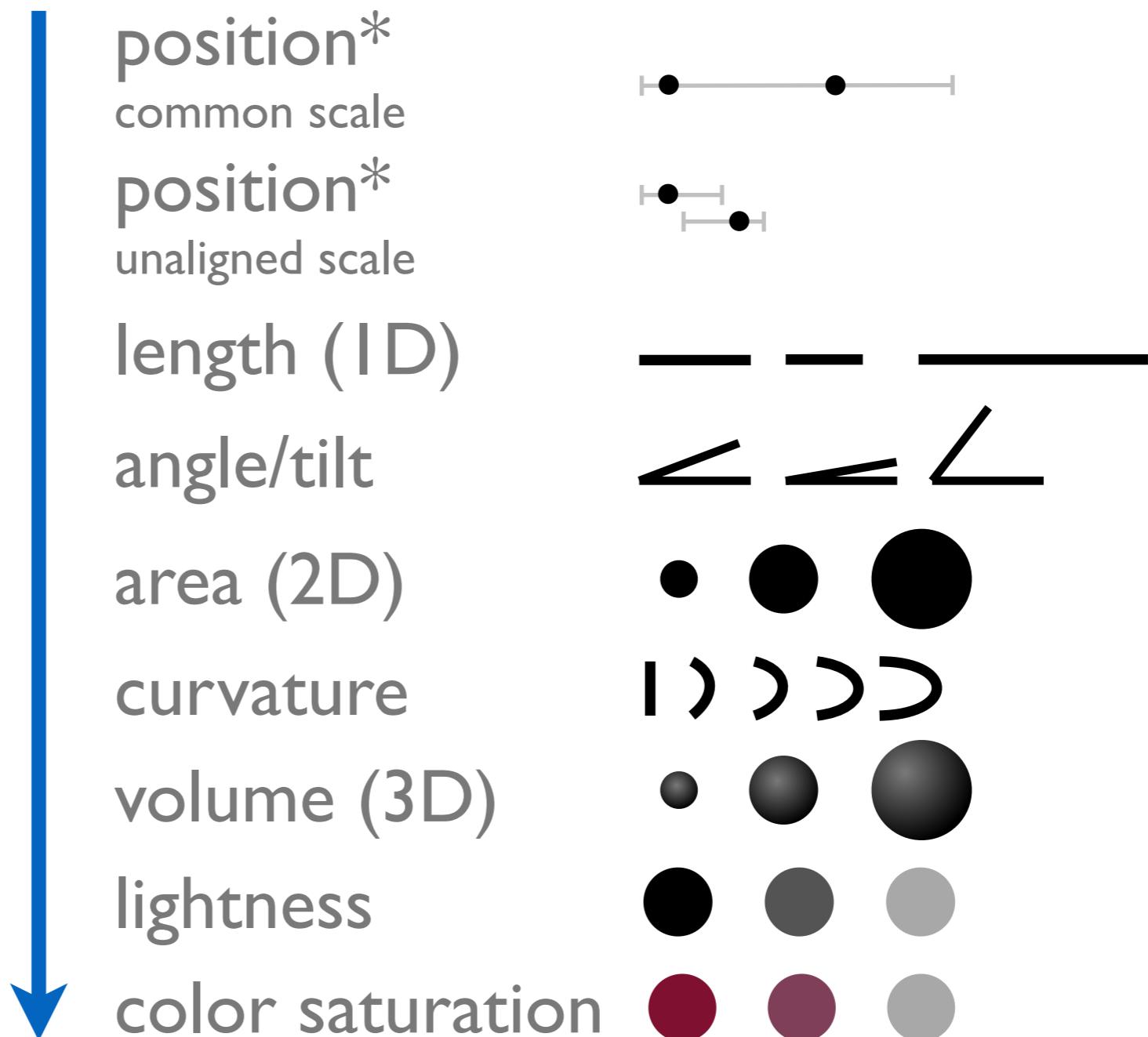


position*
proximity

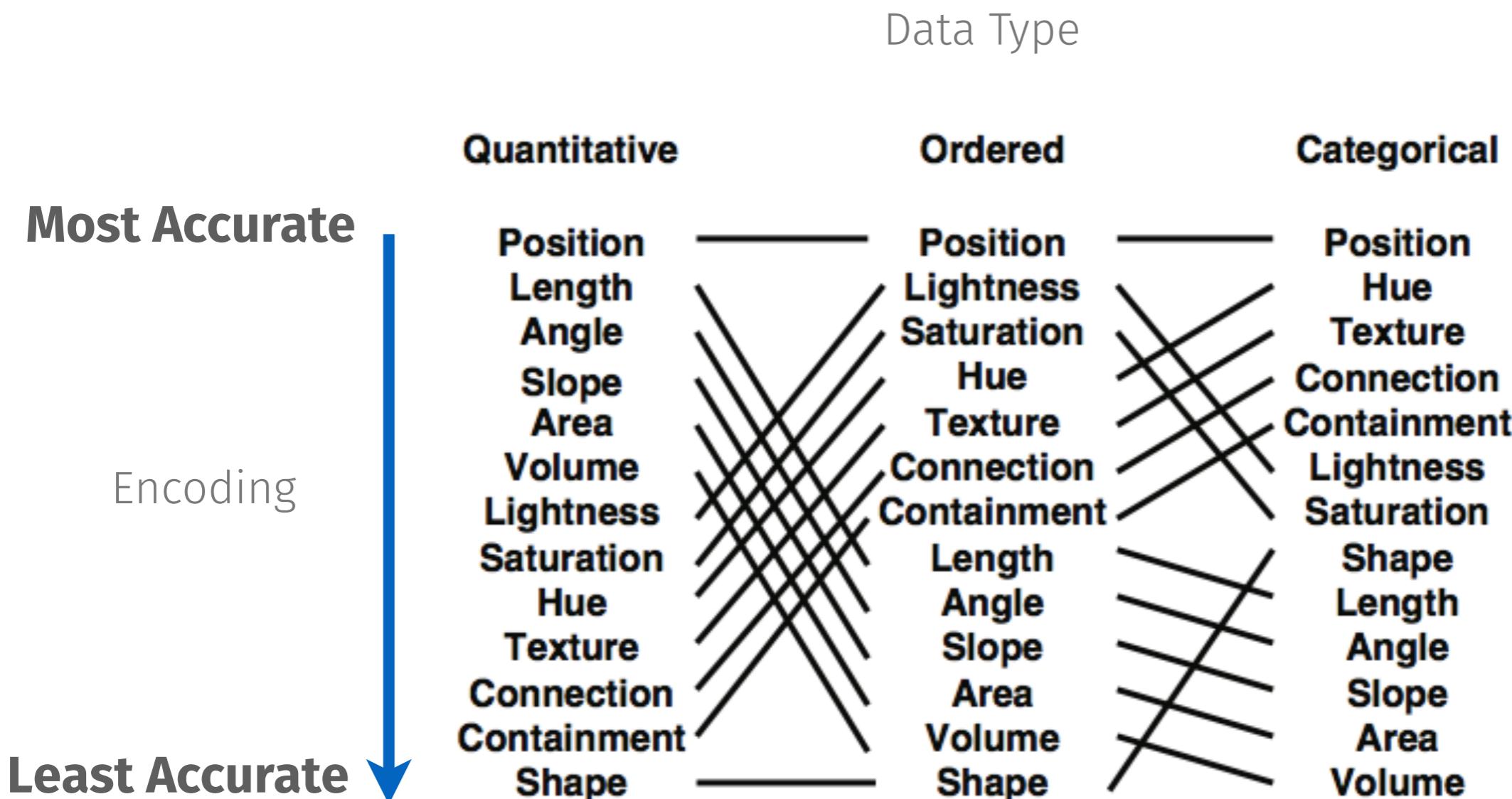


Visual Channels: Rankings

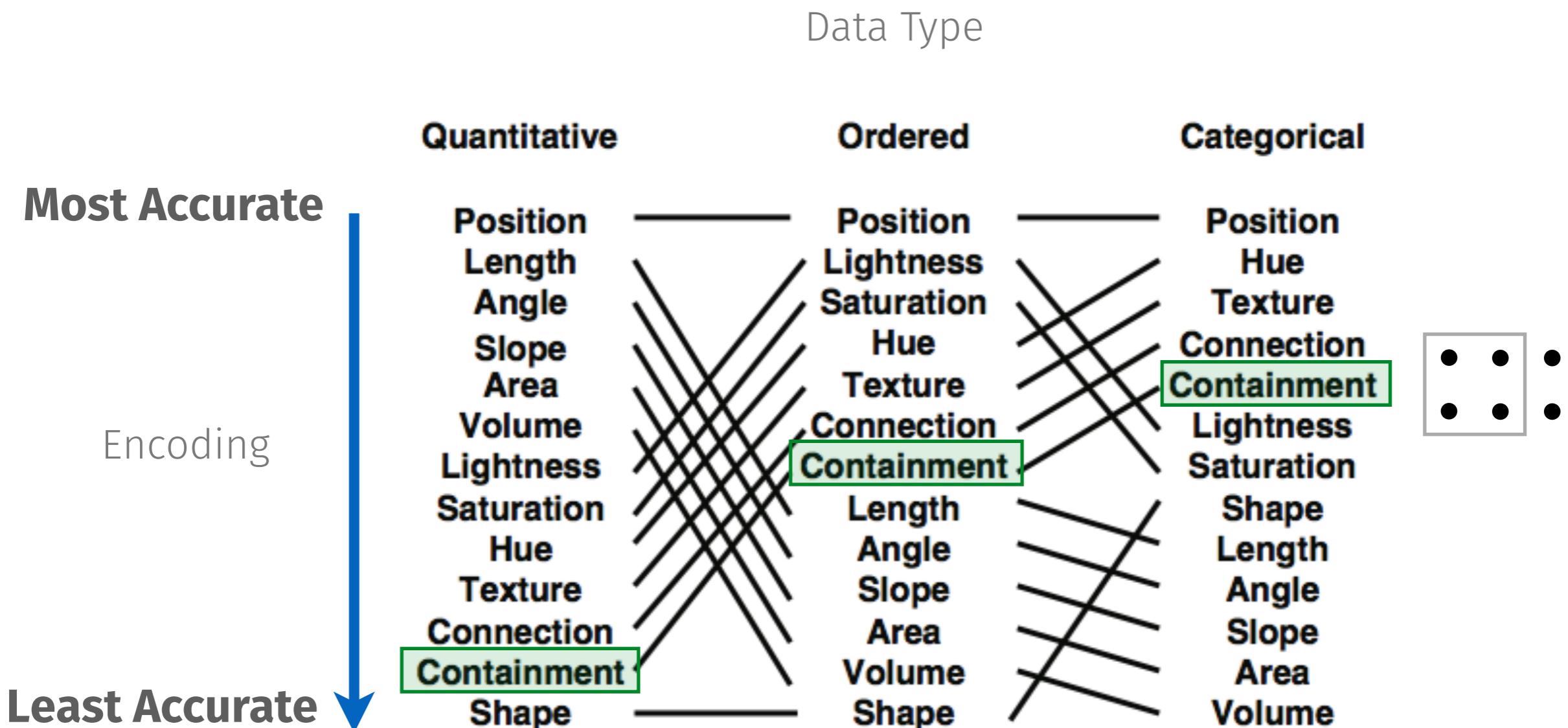
Ordinal &
Quantitative
How much?



Ranking of Encodings

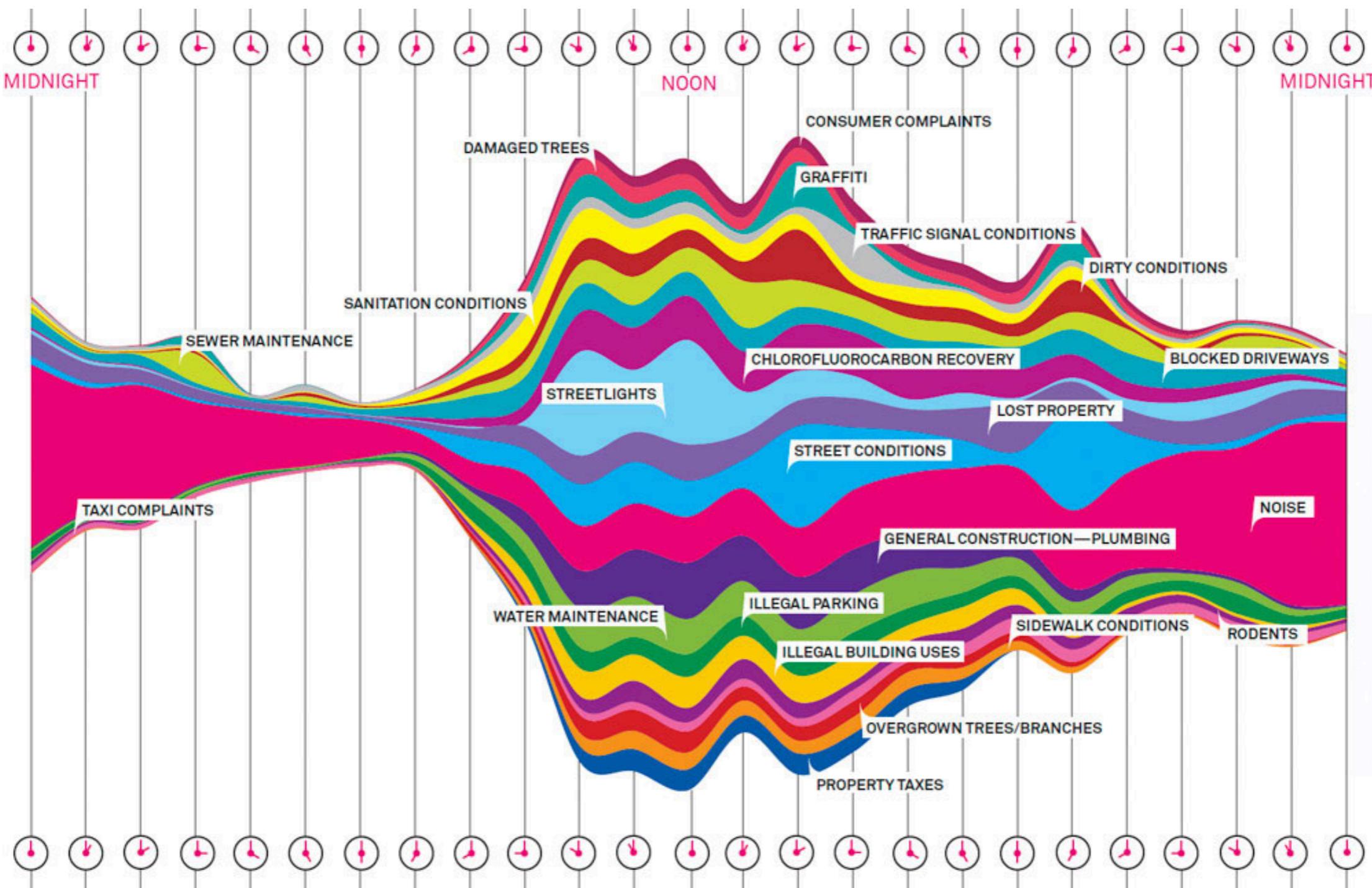


Ranking of Encodings



Back to our Design Critique ...

Exercise: Design Critique



Ranking of Encodings

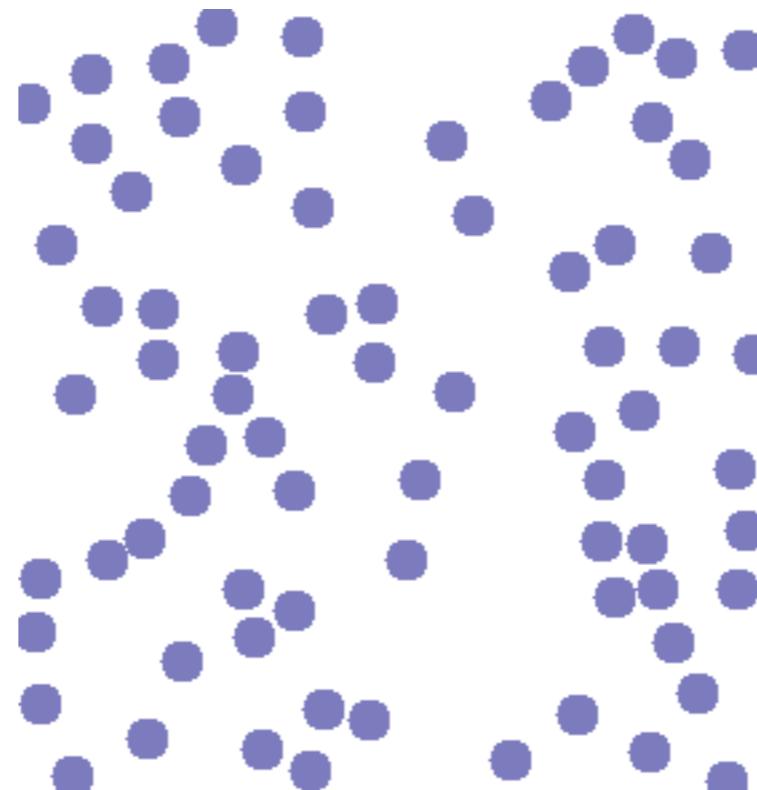
Principle of Importance Ordering (Mackinlay 1986):

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?

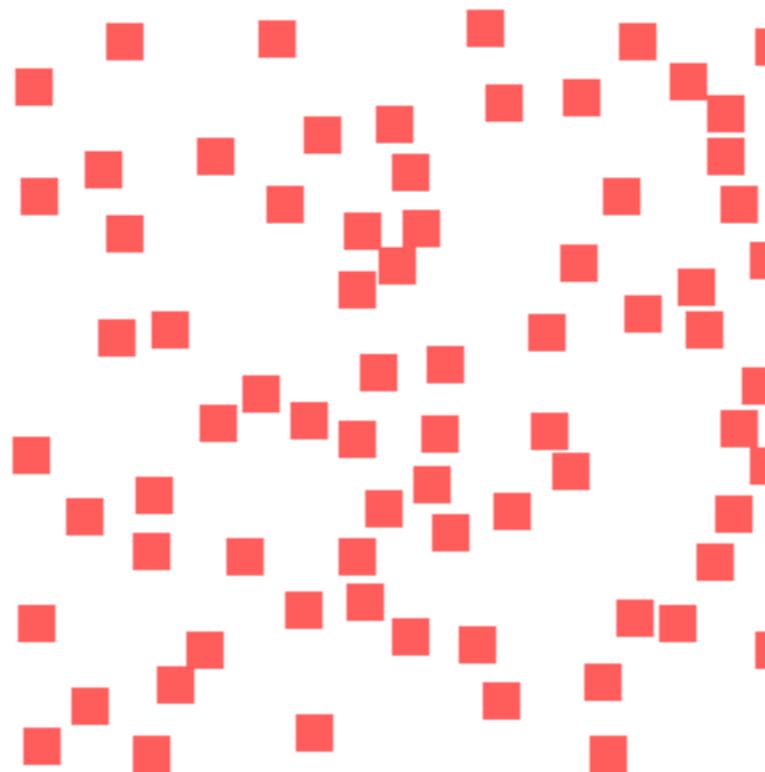
Preattentive Processing: Color

Can you spot this: ● ?



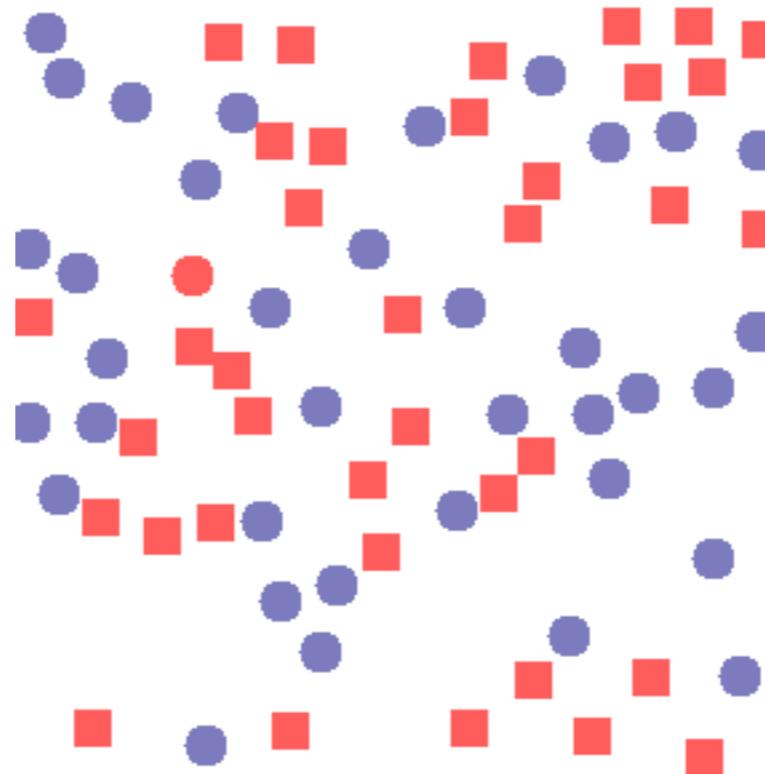
Preattentive Processing: Shape

Can you spot this: ● ?

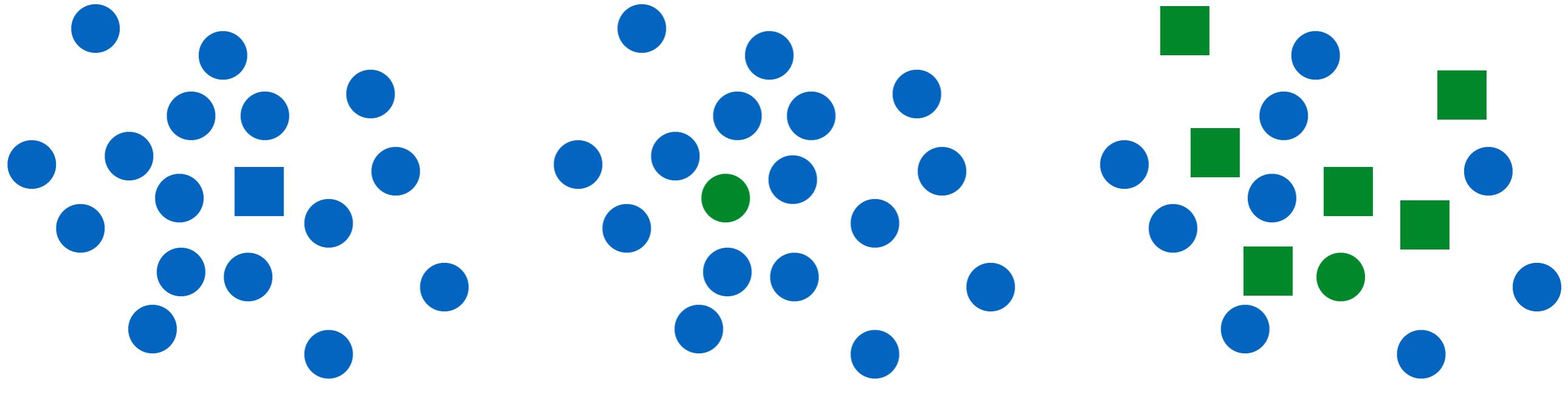


Preattentive Processing: Shape & Color

Can you spot this: ● ?



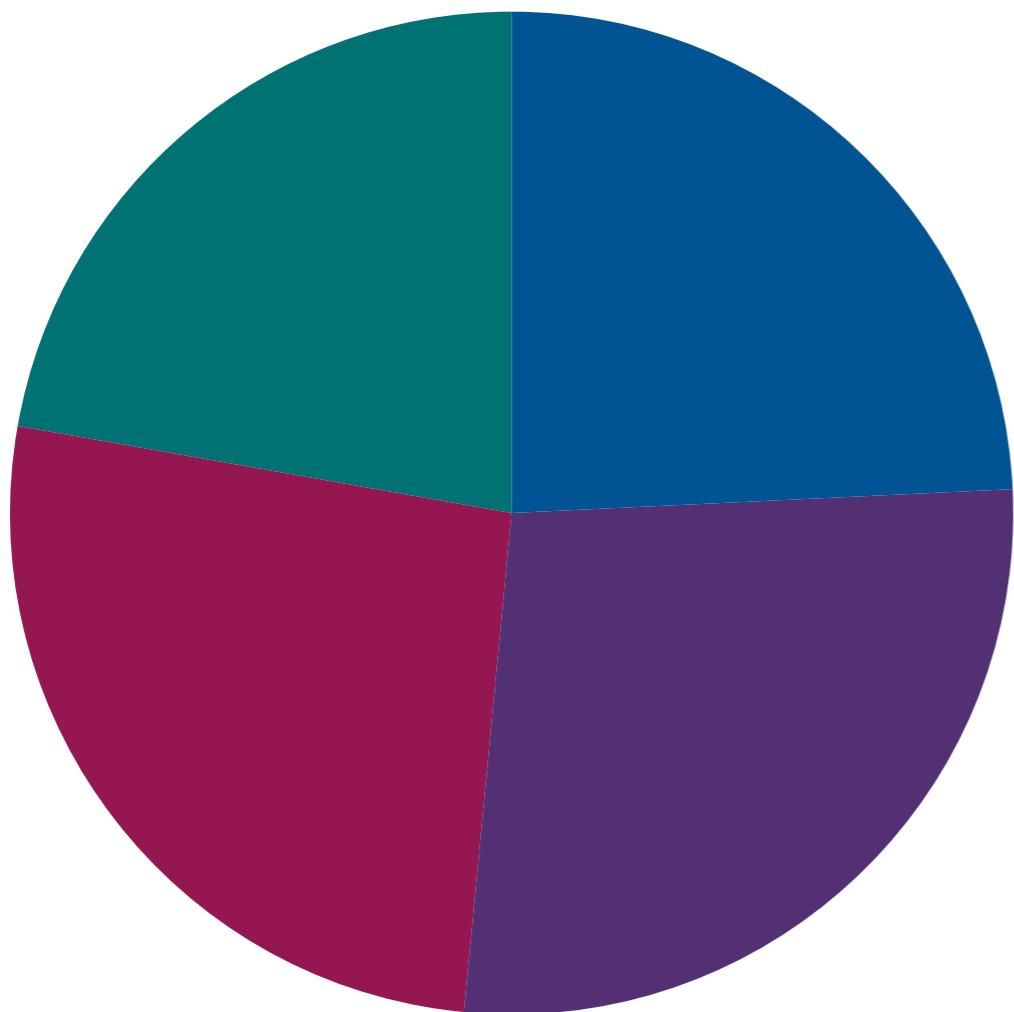
Preattentive Processing



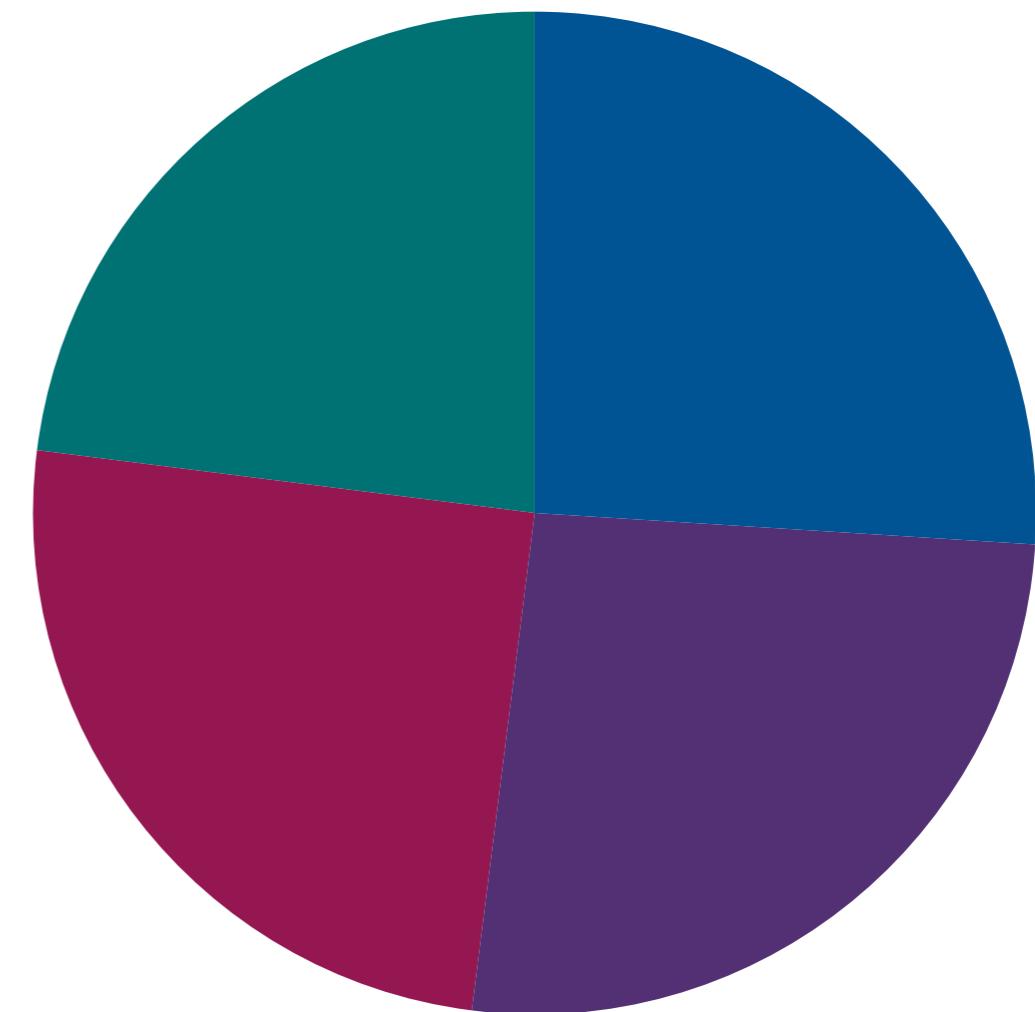
- visual properties that can be perceived in less than 250 ms
- no sequential scanning of the image required, unlike text or numbers
- examples for other visual properties that can be processed preattentively: orientation, curvature, direction of motion, size and others

Using Rankings

Year 1



Year 2



A

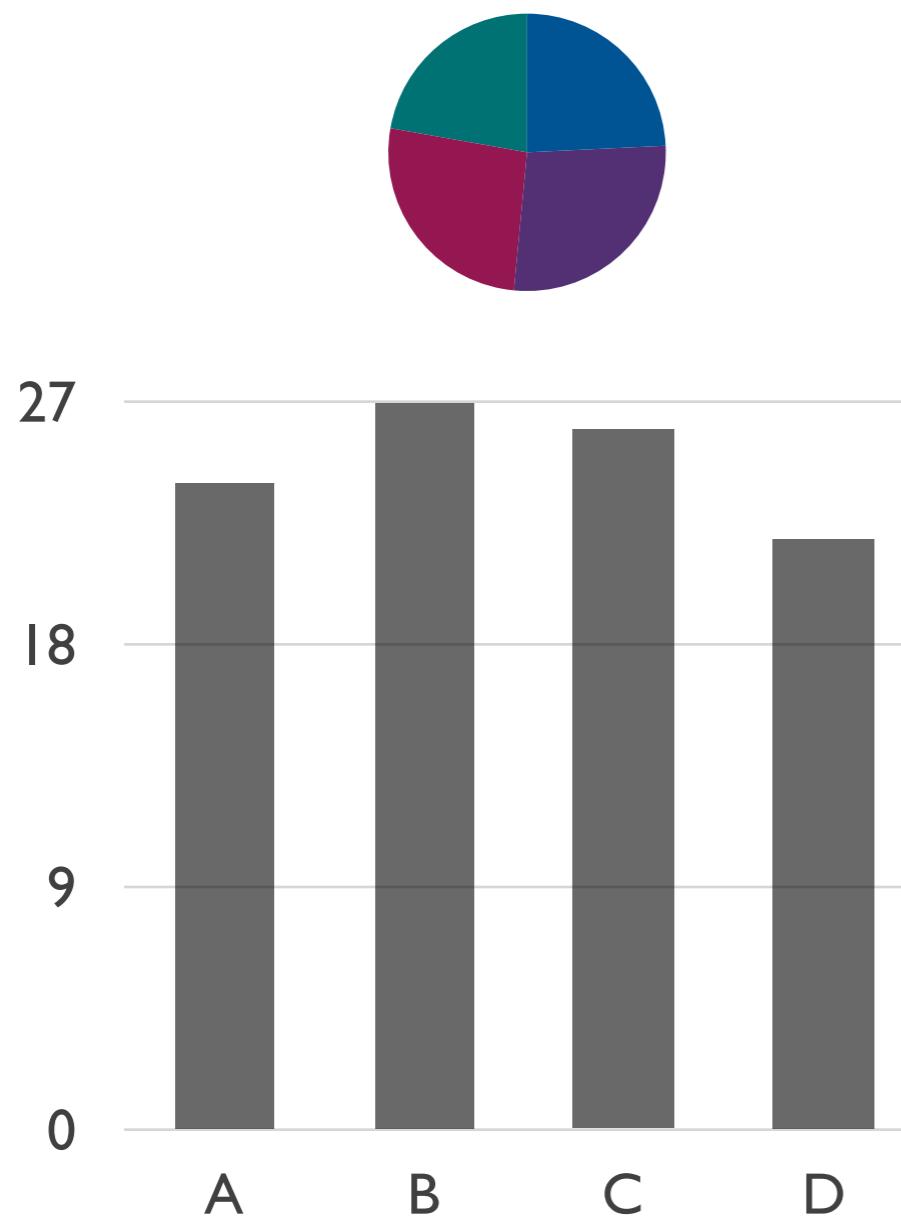
B

C

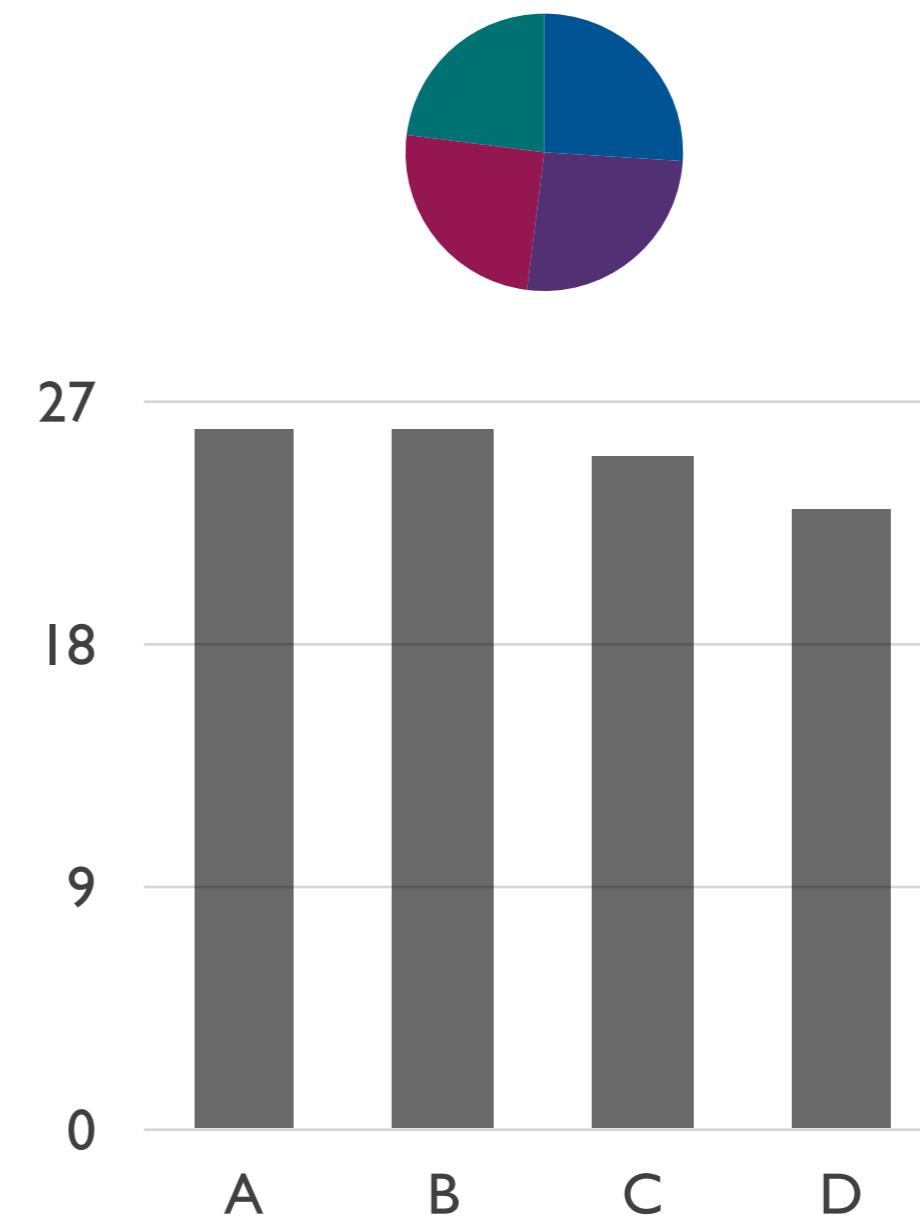
D

Using Rankings

Year 1



Year 2



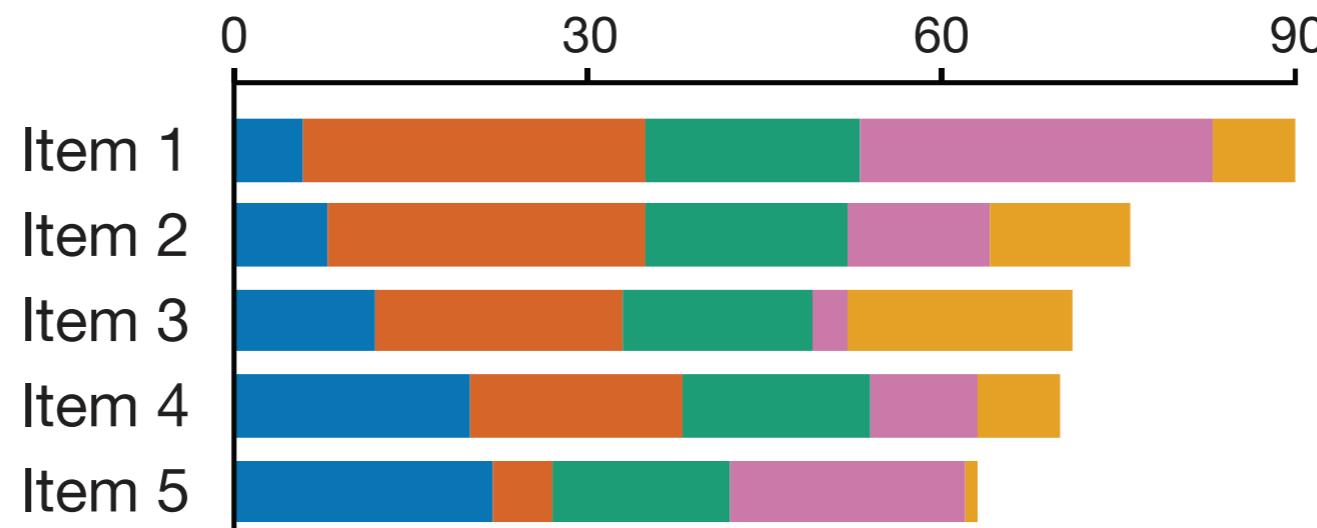
Bar Charts for Items & Categories

| | 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 |

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 |

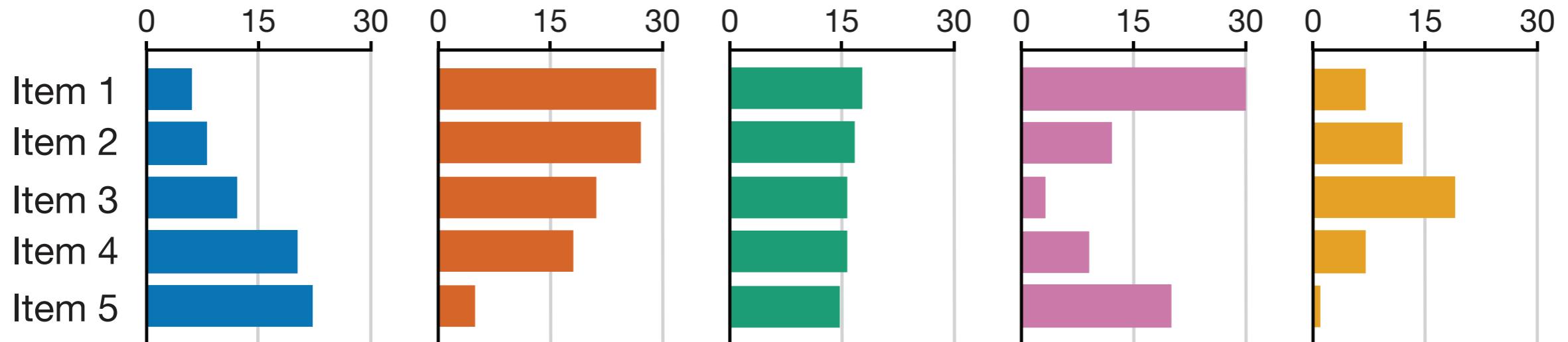
Stacked Bar Chart



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 |

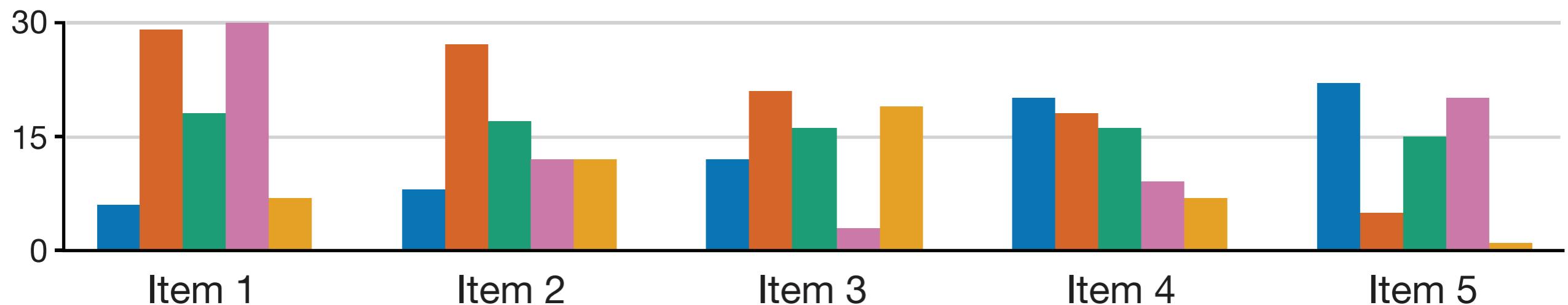
Layered Bar Chart



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



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

LineUp: Ranking Visualization

