

DATA VISUALIZATION

KIRELL BENZI, PH.D



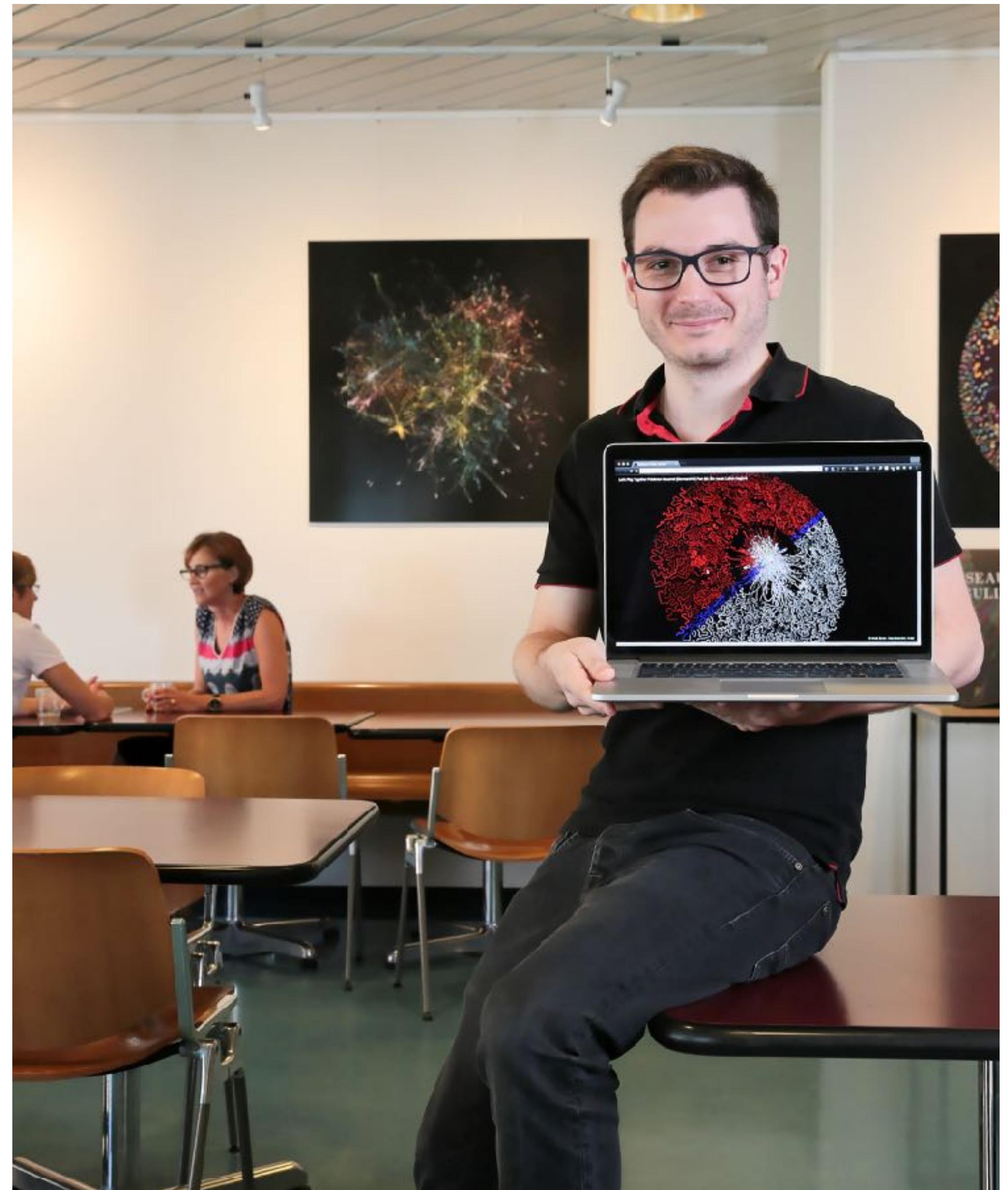
www.kirellbenzi.com

About me

Got my Ph.D from EPFL in 2016, in network science in the LTS2 led by Prof. Vandegeynst.

CEO, founder of Kirelion a creative data science studio

Data artist, exhibition “Singular Networks” in several countries

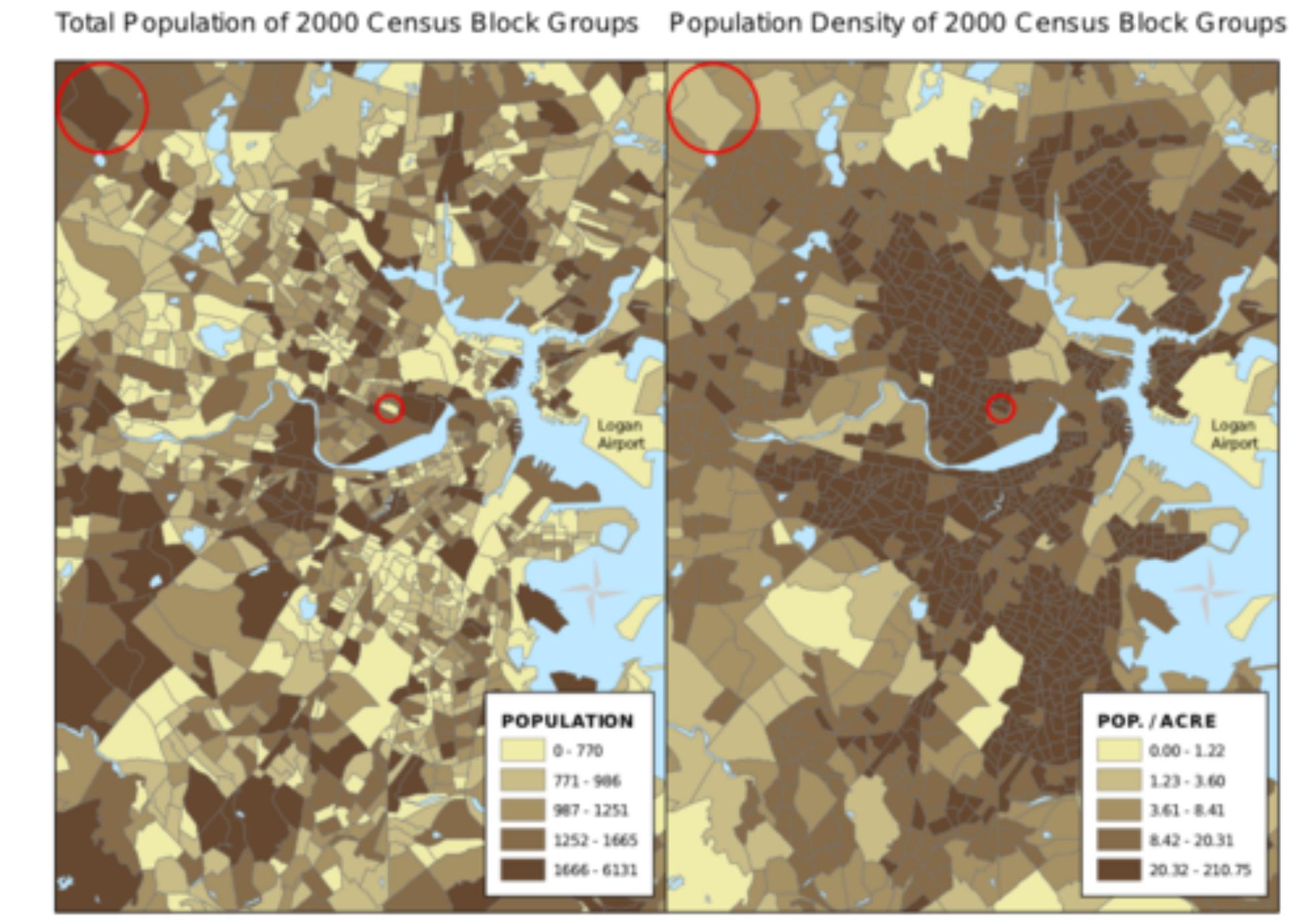
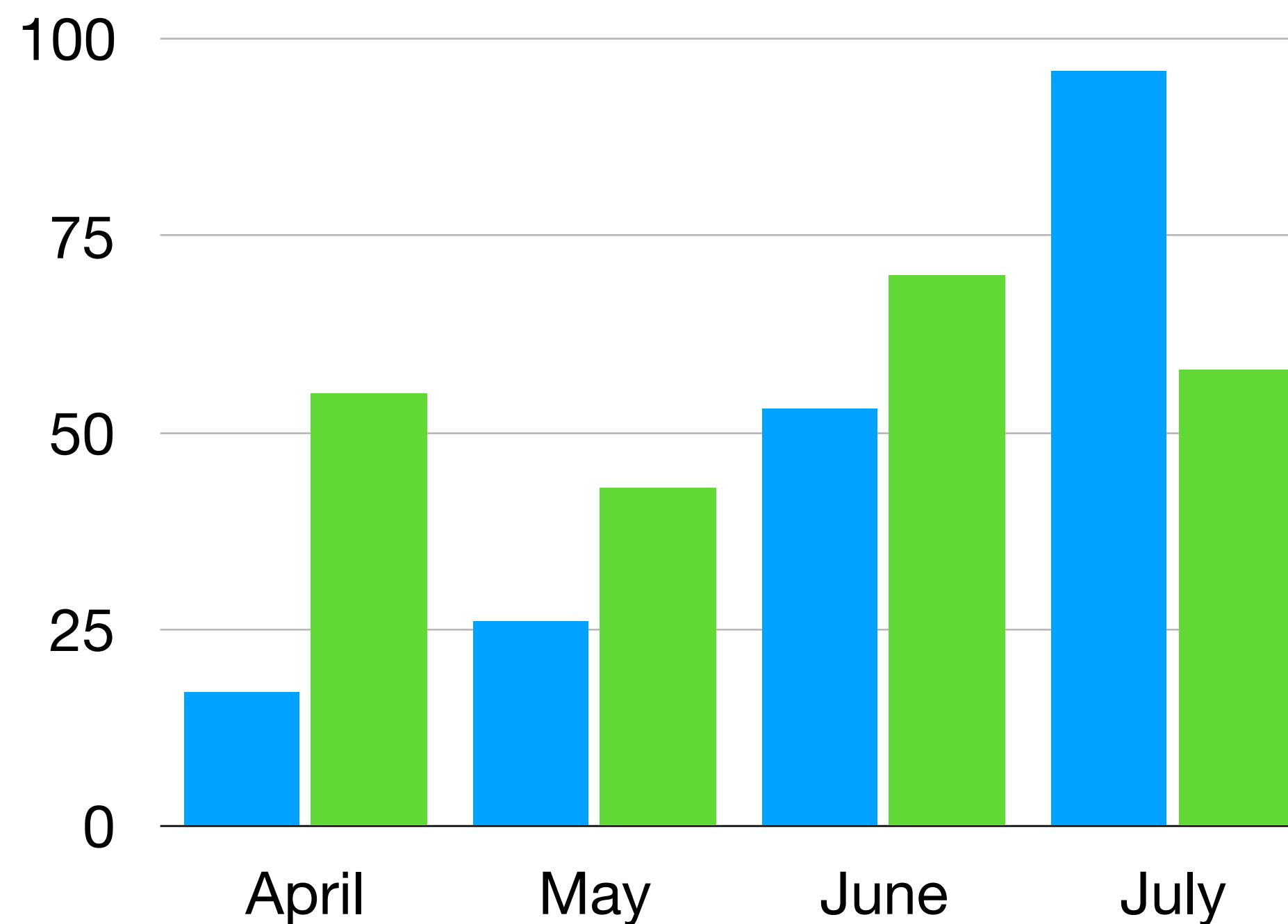


“The purpose of visualization is insight, not pictures.”

–Ben Shneiderman

Defining data visualization

Efficiently communicate information from **statistical** data using visual objects (such as bars, lines, points, etc.)



Other definitions

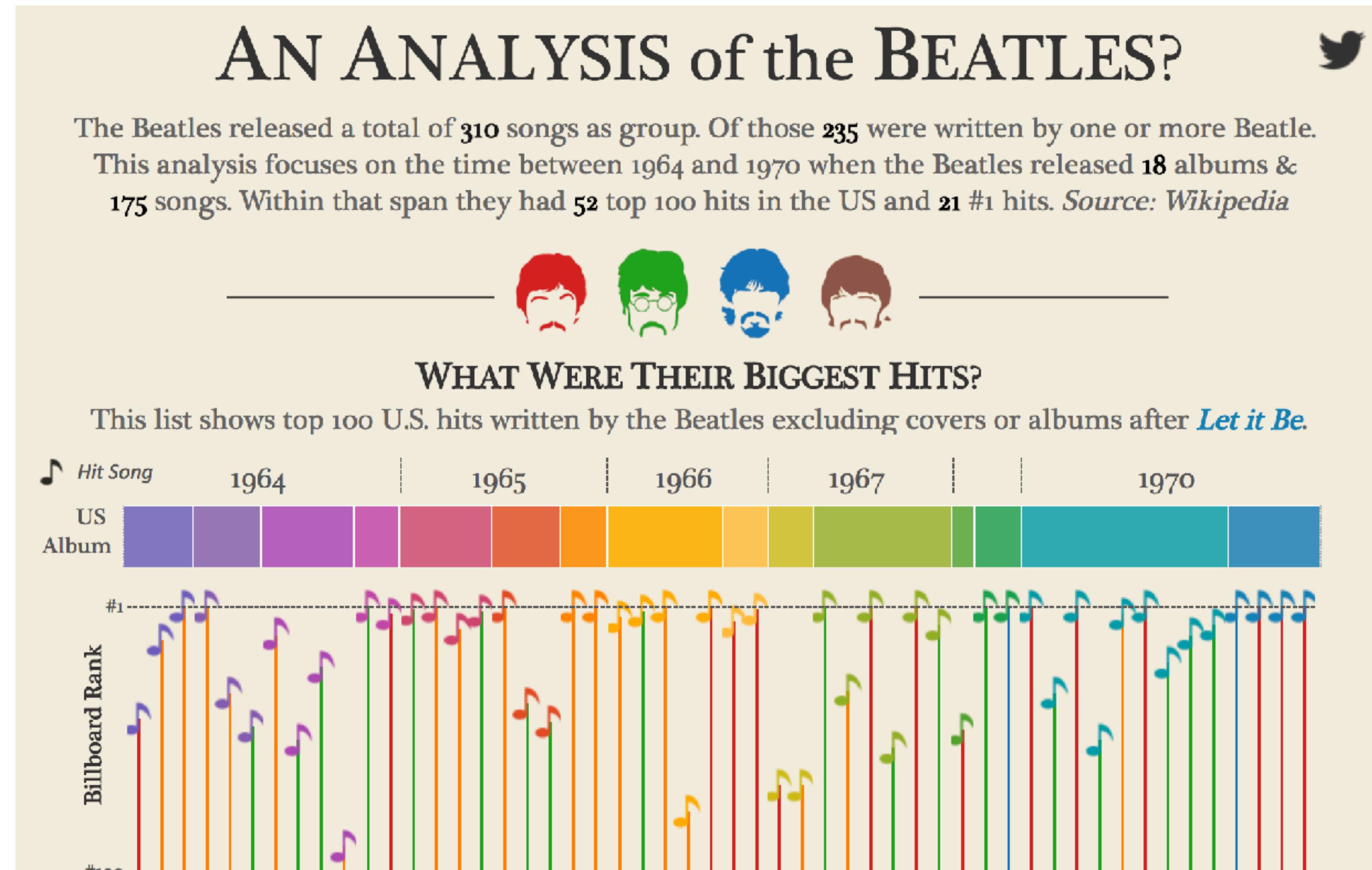
Infographics

Scientific visualization

Data art



Infographics



Scientific viz



<http://radiotzanck.com/>

Data art

When the ell... Music is Good

The Montreux Jazz Festival is one of the most famous music festival...

Secret Knowledge

Who would have thought that Wikipedia could be so structured...

Outside of the box

Virtua is a major actor in digital agency landscape of Switzerland. It relies on its collaborators, at the center of the equation, to deliver...

DISCOVER

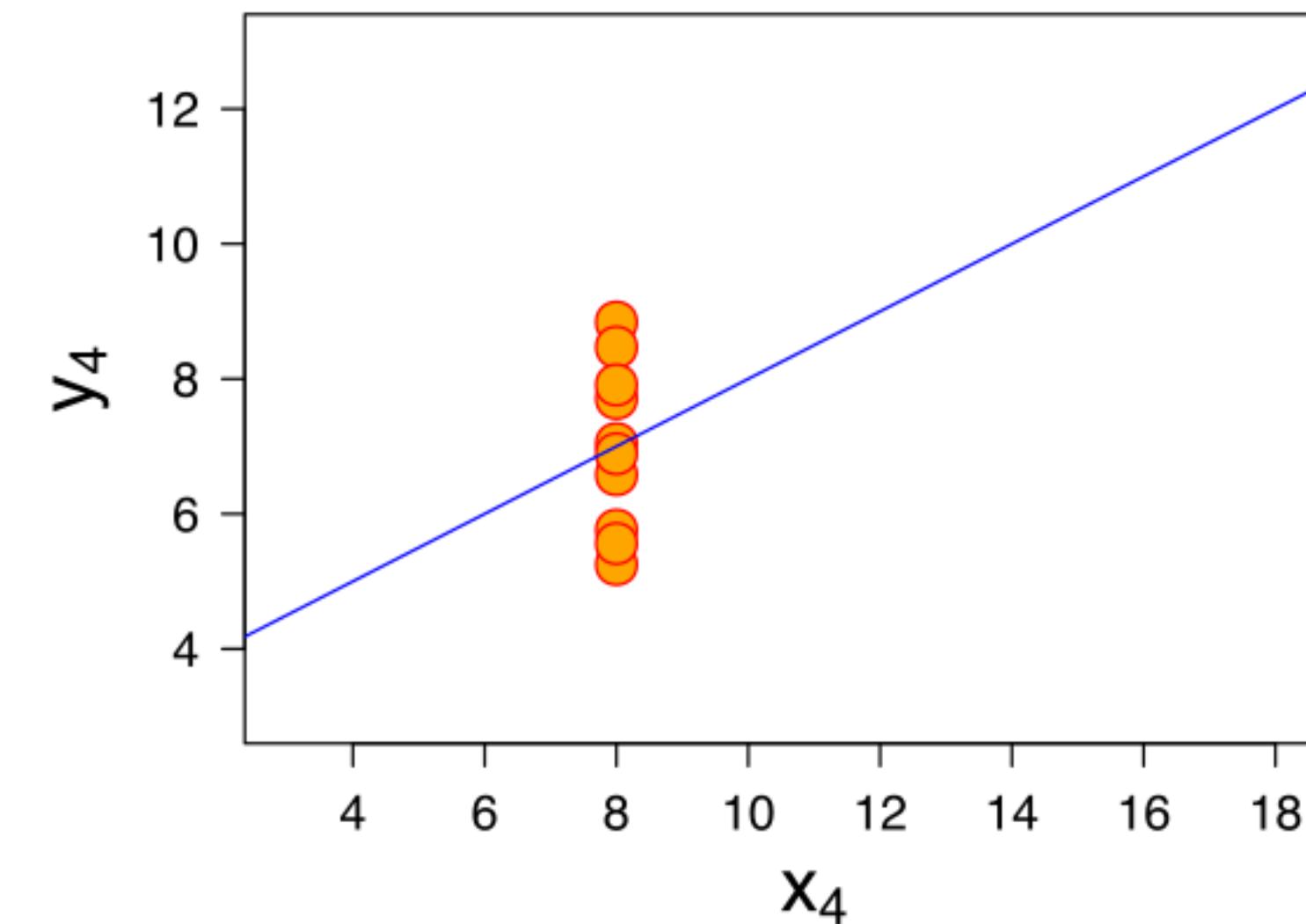
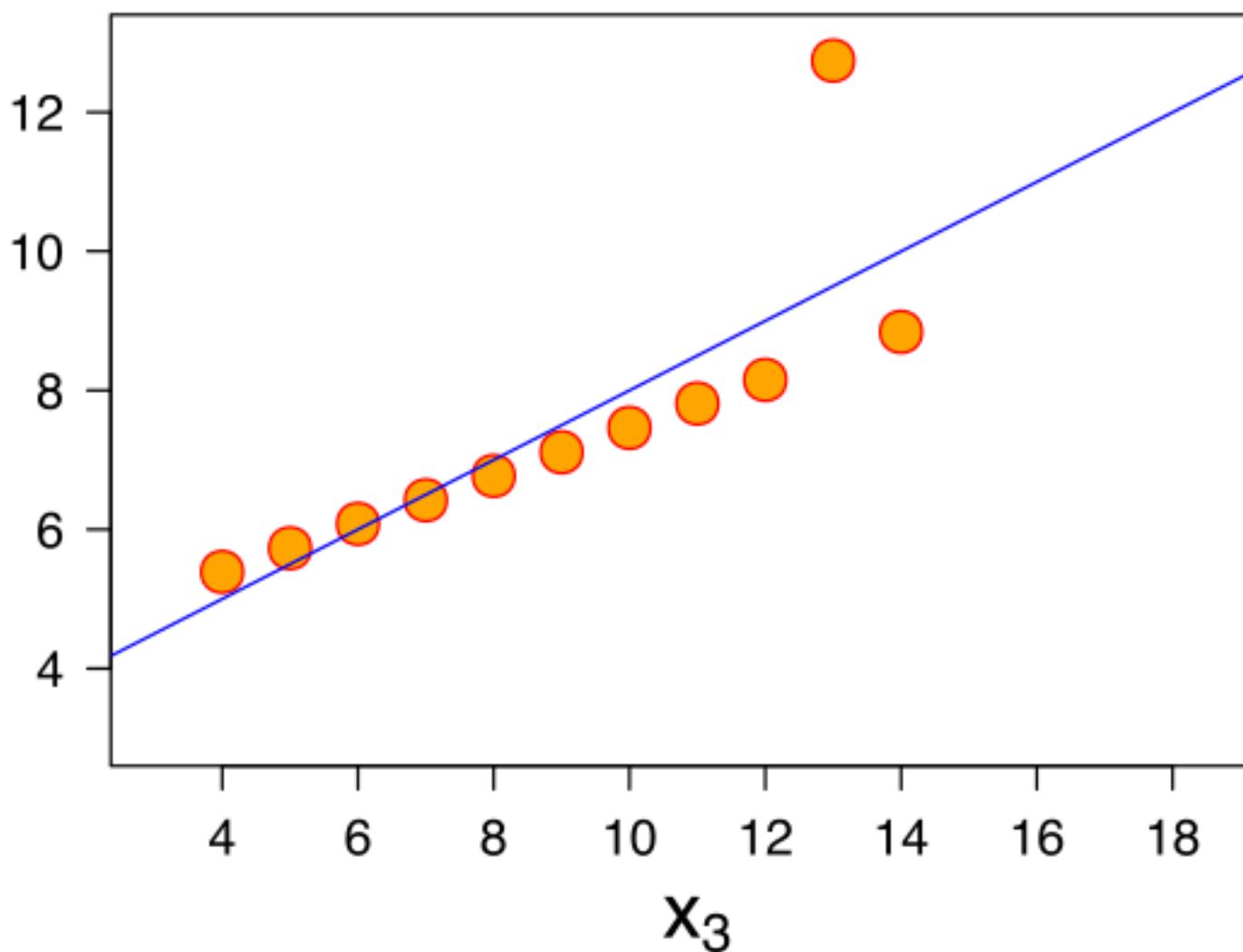
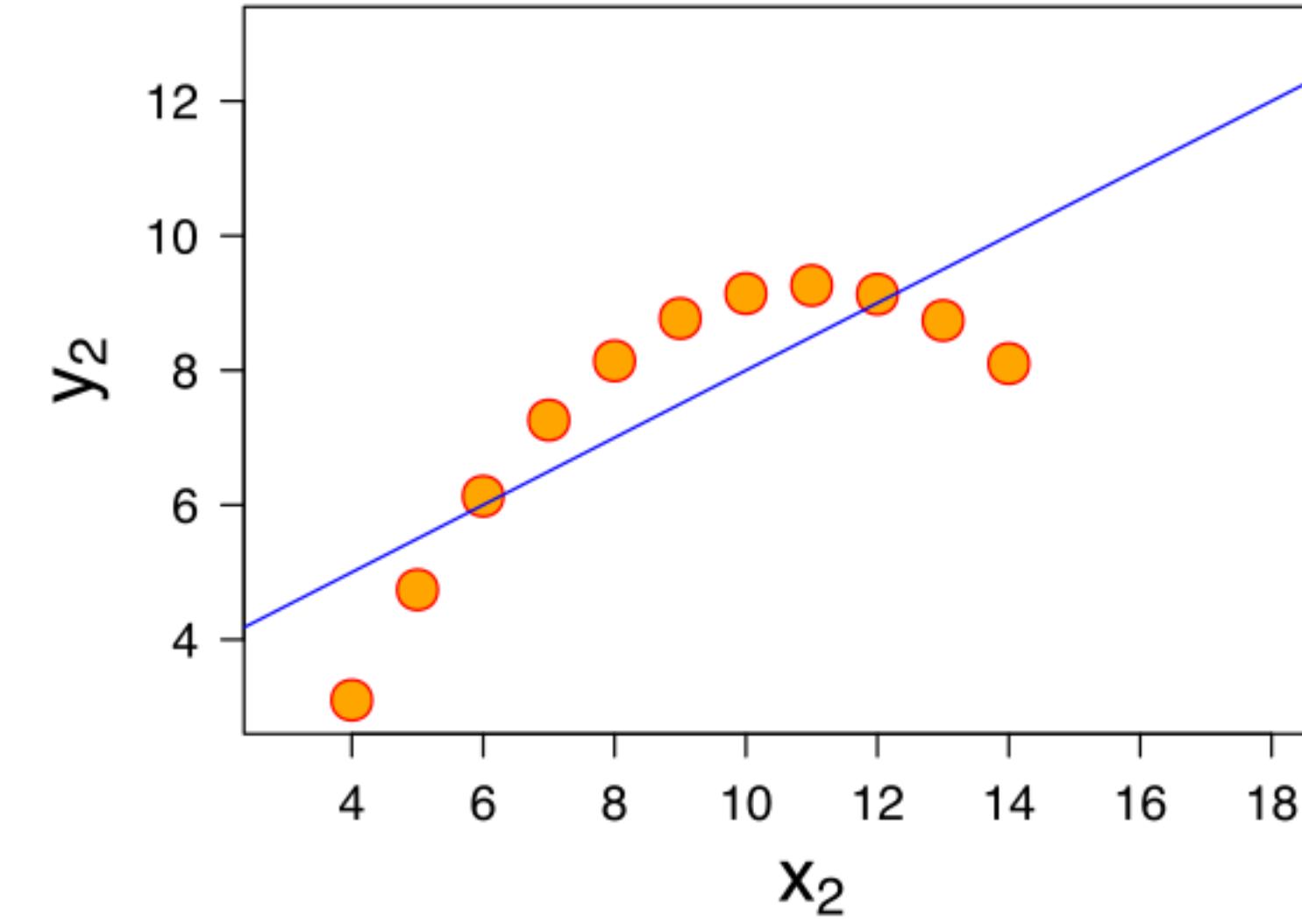
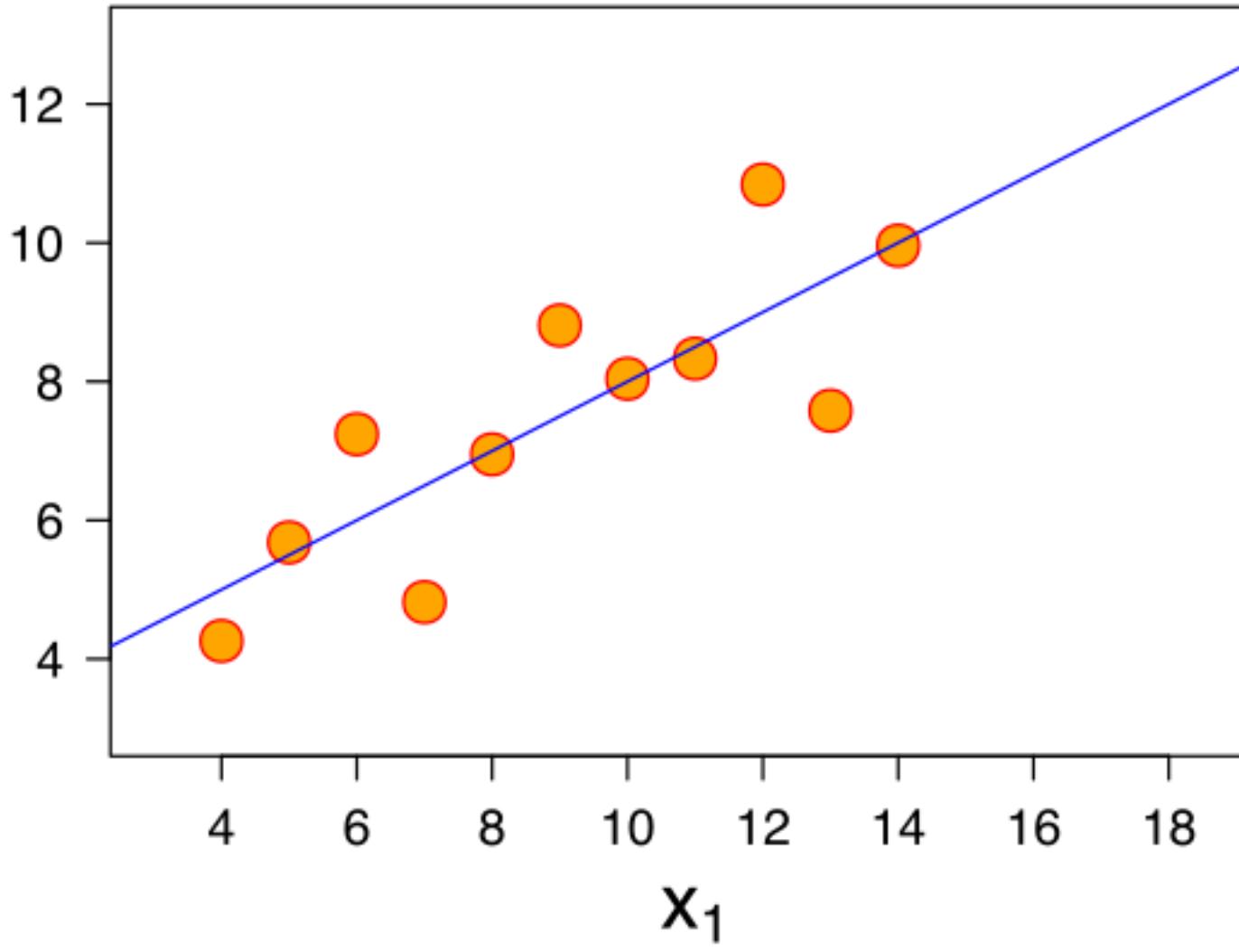
The Dark Side and the Light

Re-invent

The cycle of inventions is constantly renewing itself: past

<http://www.kirellbenzi.com>

Why not only use statistics?



Identical statistical properties

Property	
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.0 + 0.6x$
Coefficient of determination of the linear regression	0.67

[Anscombe's quartet]

Why visualize?

Communication

Transmit information to others

Exploration

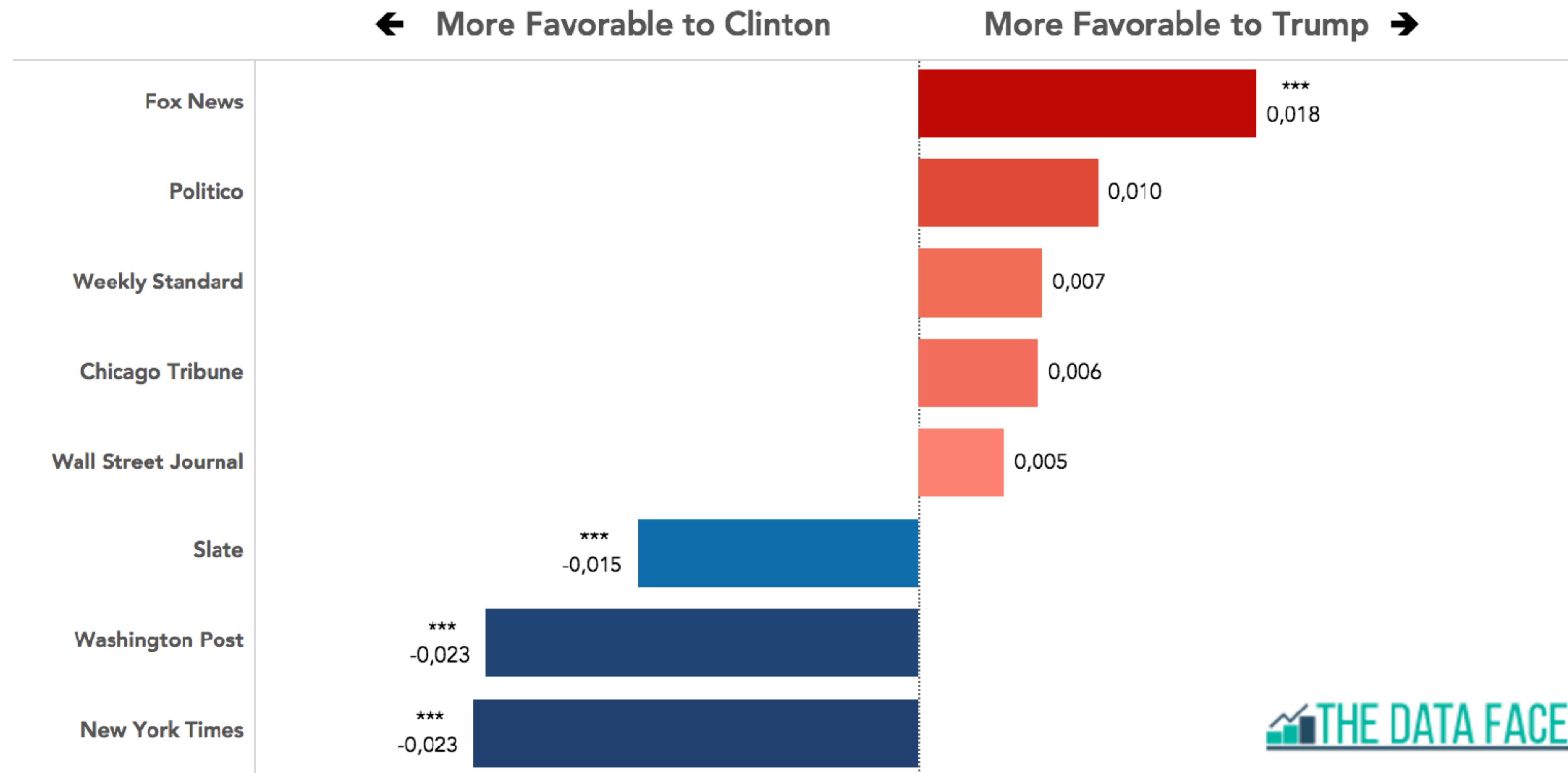
Reveal previously hidden insights from the data

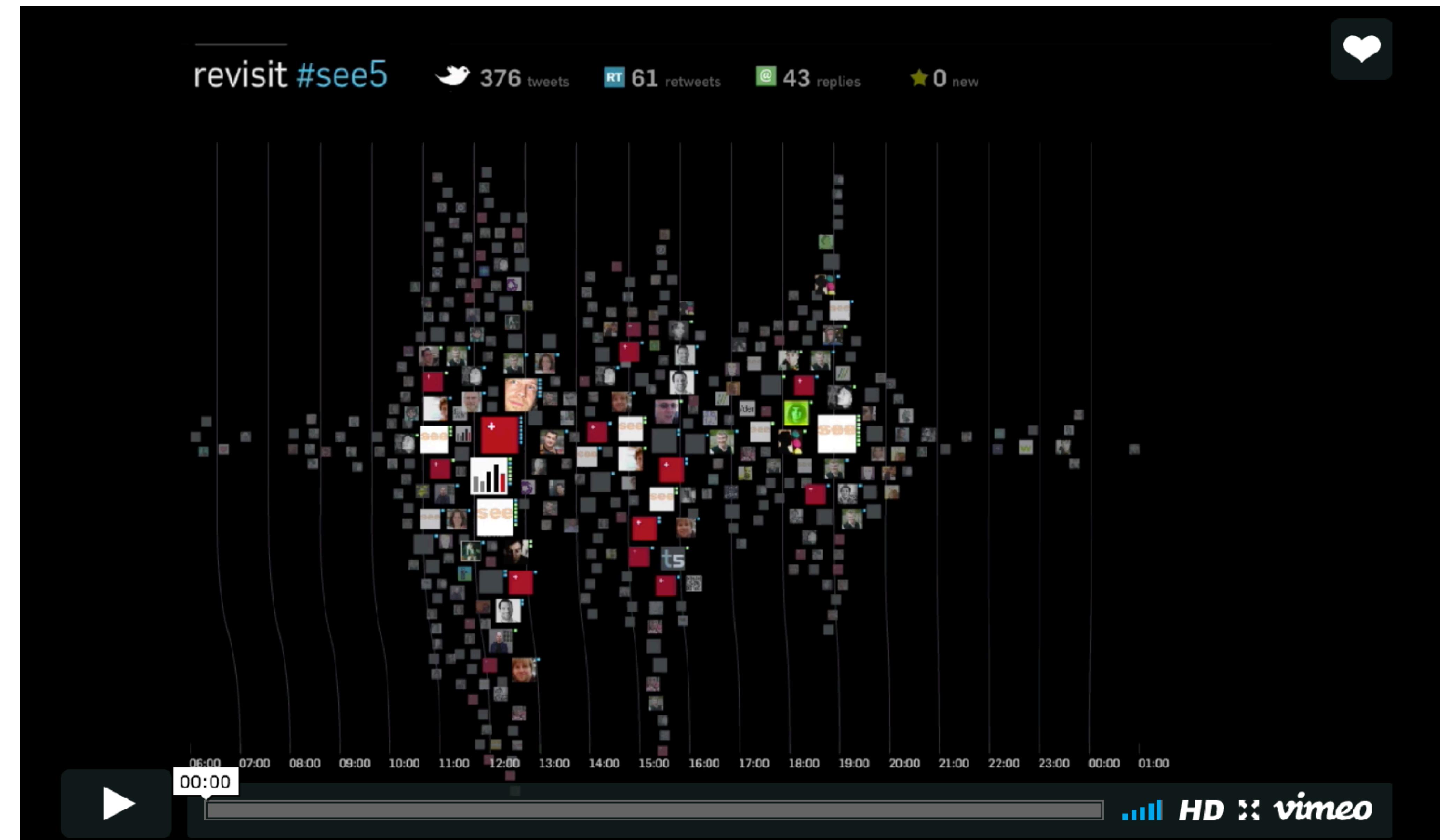
Cognition

Understand and reason about the data

How Favorably Does Each Media Outlet Treat The 2016 Presidential Candidates?

Shown below are the differences in median sentiment scores assessed to opinion articles about Trump vs. Clinton for each outlet in our sample†. Significant differences are denoted by an asterisk (according to a Mann-Whitney Test).



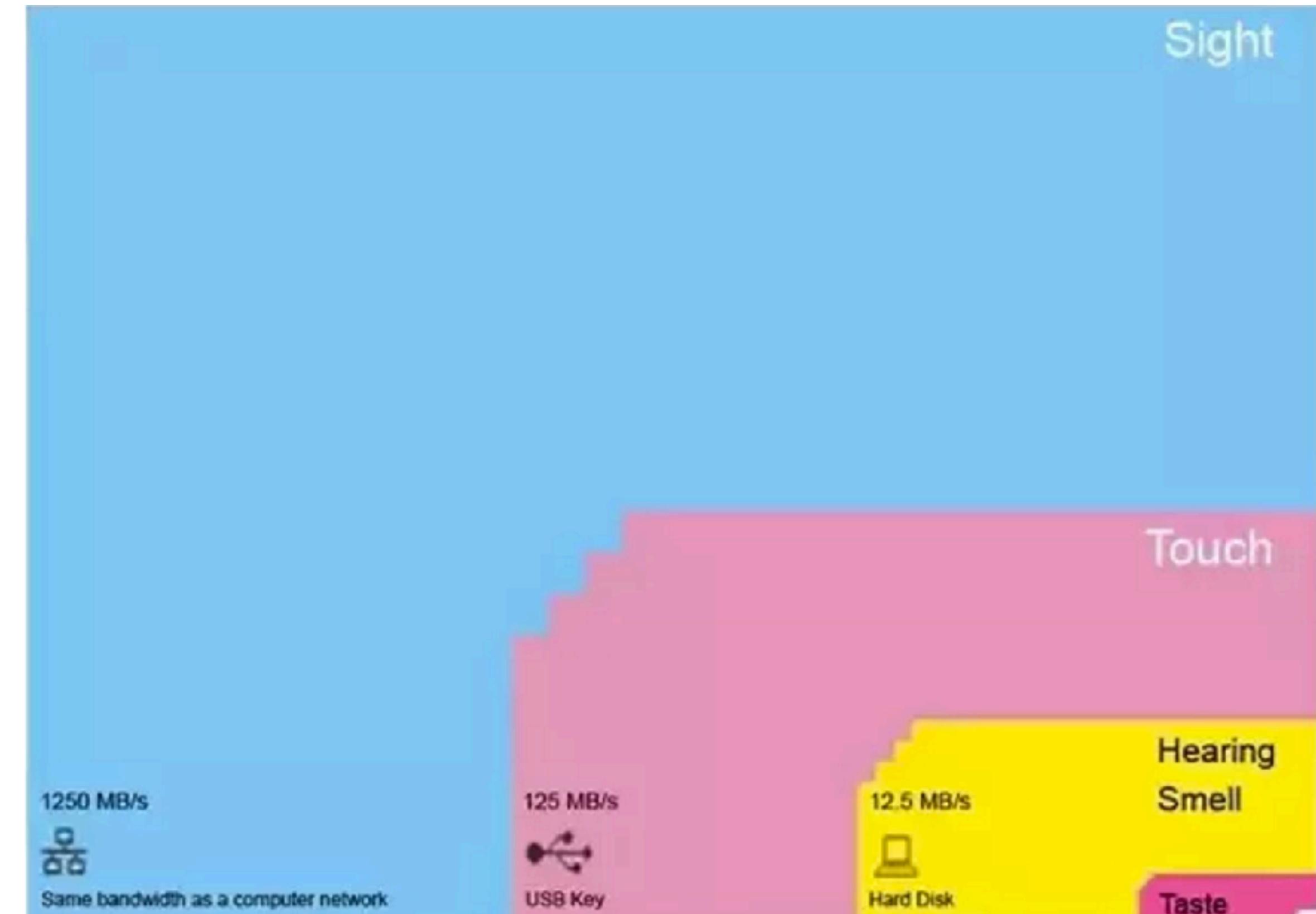


Using a graphical representation

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.

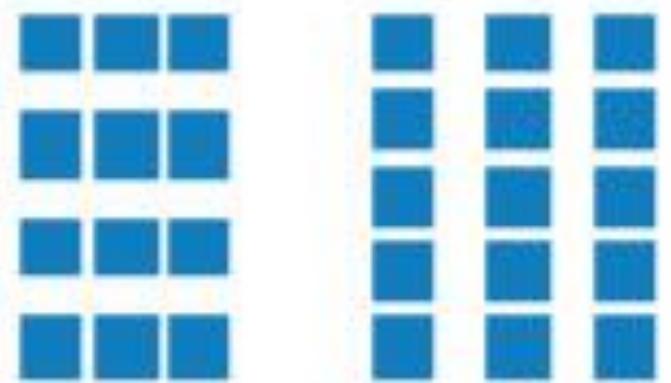


Tor Nørretranders

Alexander Lex

Gestalt effect

PROXIMITY



Most people see rows on the left and columns on the right.

SIMILARITY



Most people see a larger group containing a sub-group.

FIGURE GROUND



Most people first see a vase or a face then flip between the two.

CONTINUATION



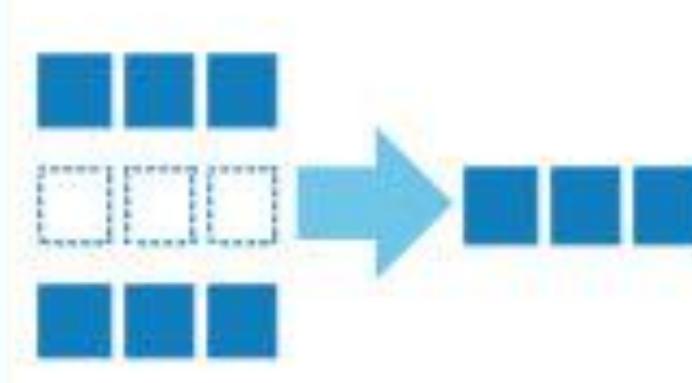
Most people see 2 rows crossing rather than for lines meeting at a single point.

CLOSURE



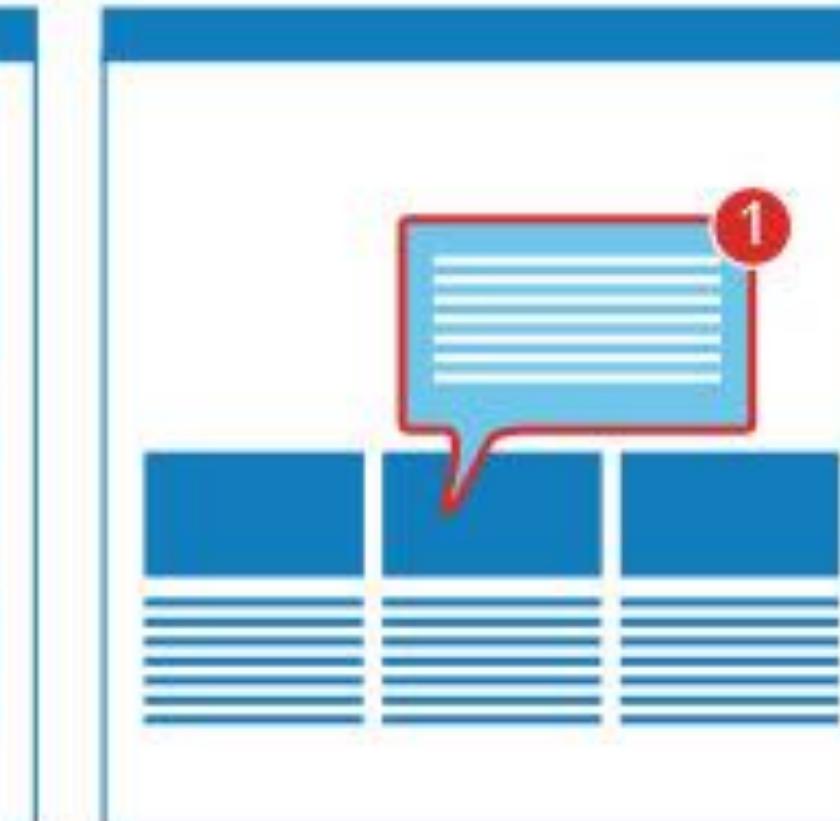
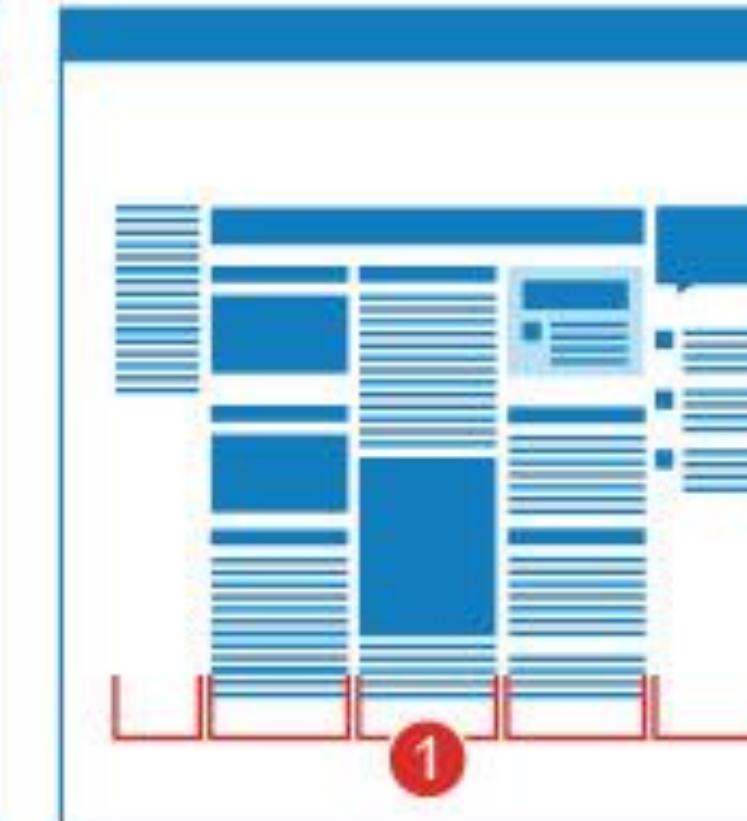
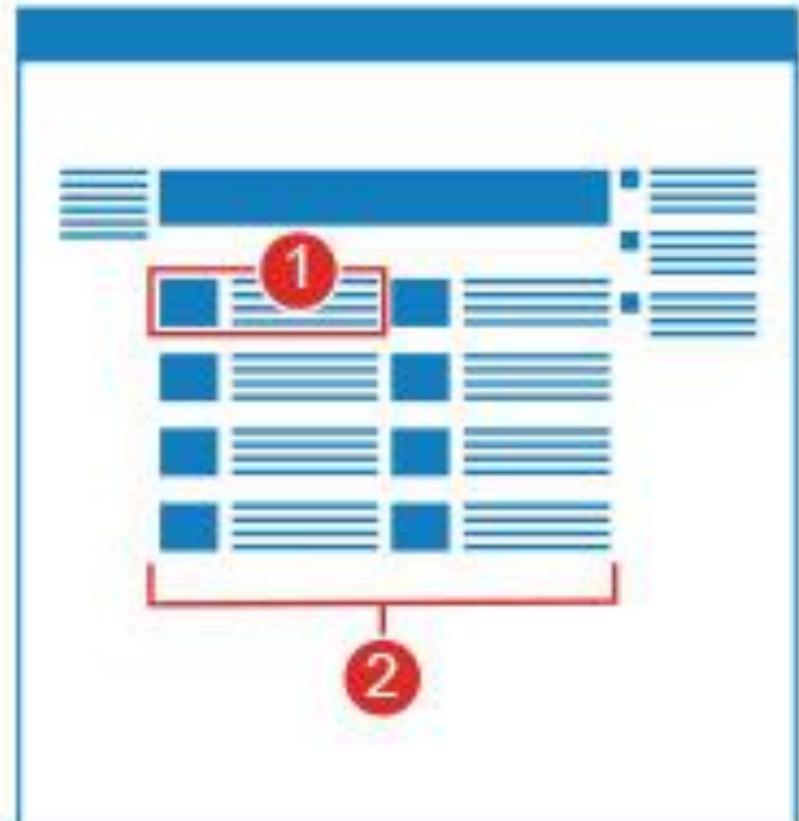
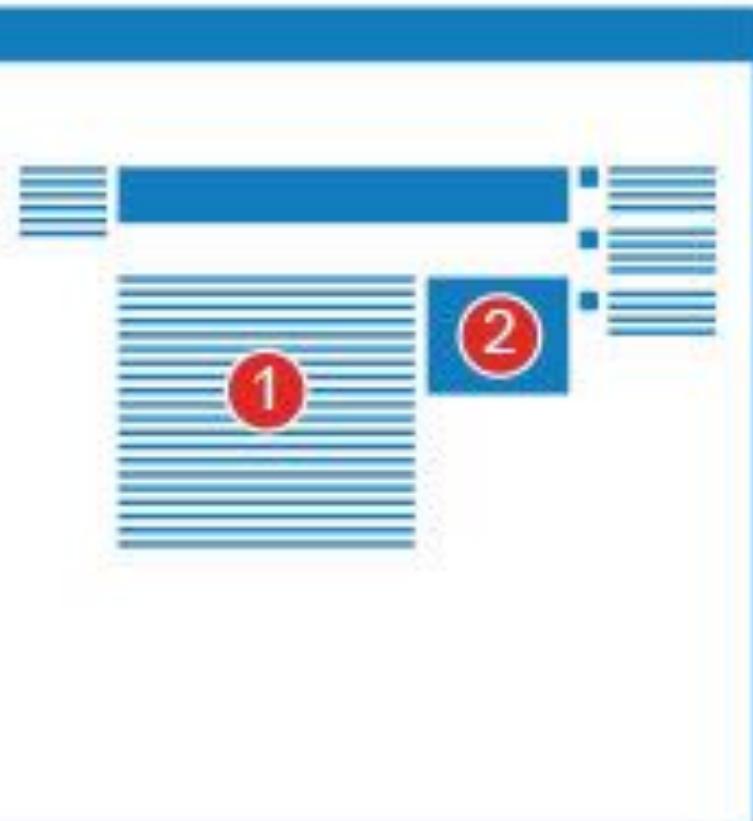
Most people see the white triangle not the 3 circles with segments missing.

COMMON FATE



Objects moving in a similar direction are perceived as belonging together.

Examples:



Elements that are closer together are perceived as being more related compared to those spaced further apart. In this

Web parts that look similar are perceived as being grouped together or related.

We automatically perceive objects as being in the foreground, (figure) or in the background (ground). Anything

We tend to perceive contours as objects. In this way we perceive lines continue in an established direction (even when they don't).

When we perceive a pattern the gaps (1) between the objects (negative or white space) are just as important as the objects

When objects move in the same direction we perceive them to be related and moving on an invisible path. Even if objects

Why use computers for viz?

*Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.*

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Segur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk en Malibow et qui rejoignirent Oroscha en Witebsk, avaient toujours marché avec l'armée.

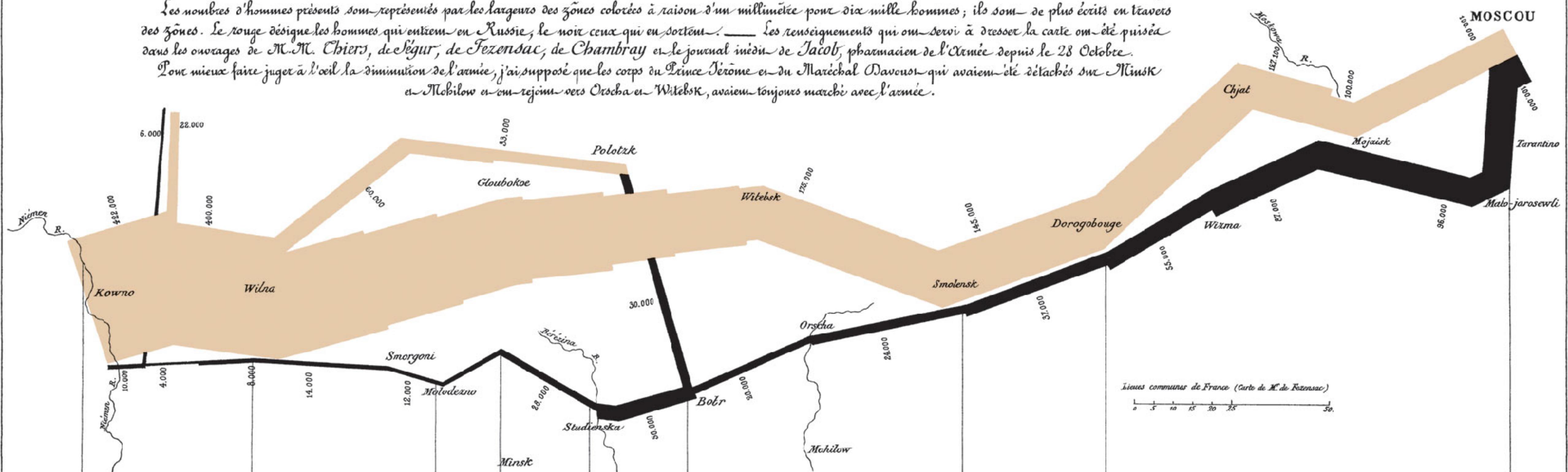
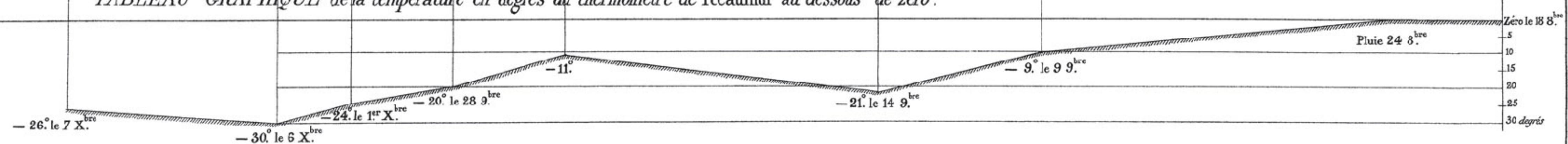


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop
le Niemen gelé.



Why use computers for viz?

Faster

Reproducible

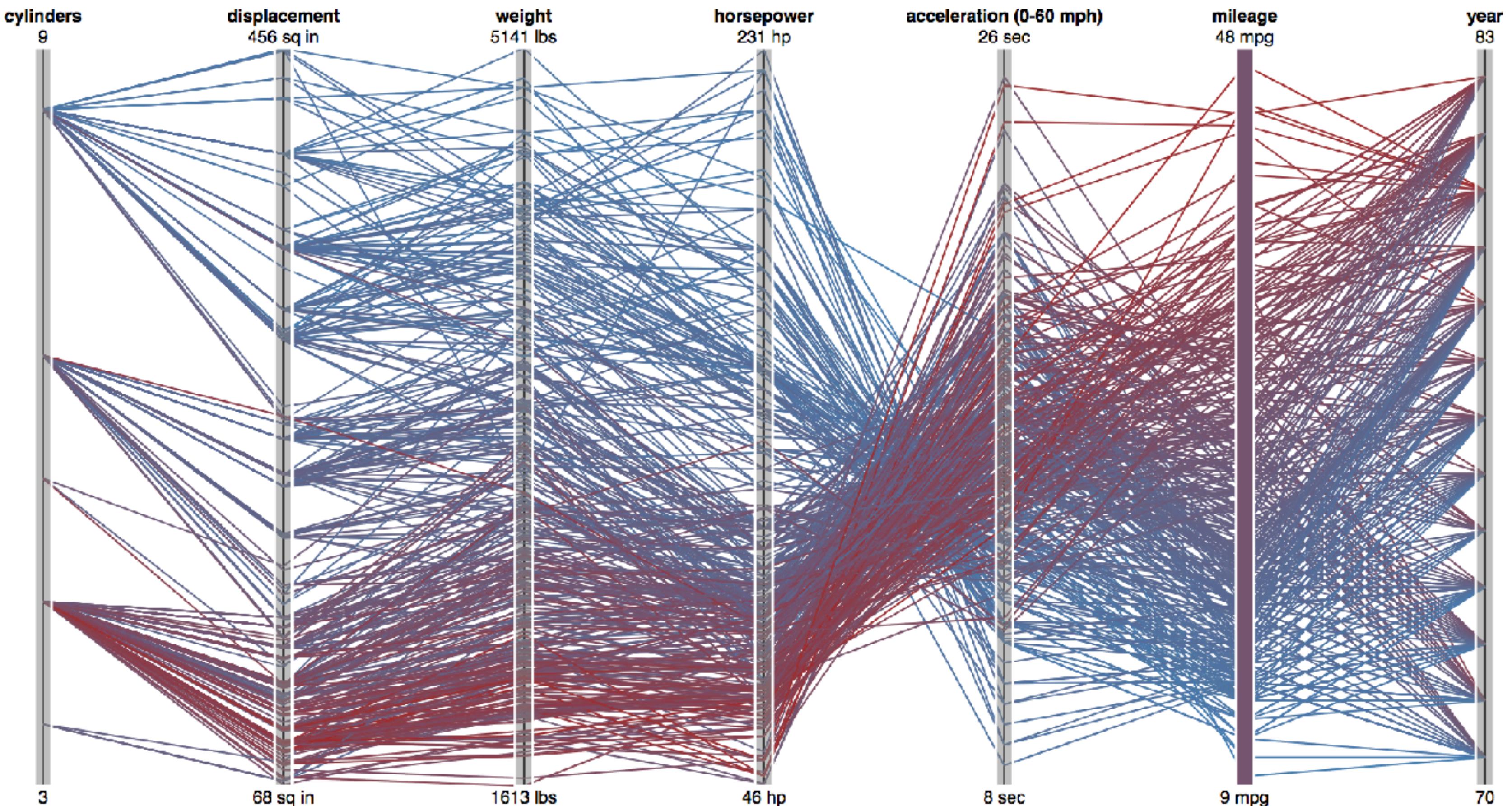
Efficient

Precise

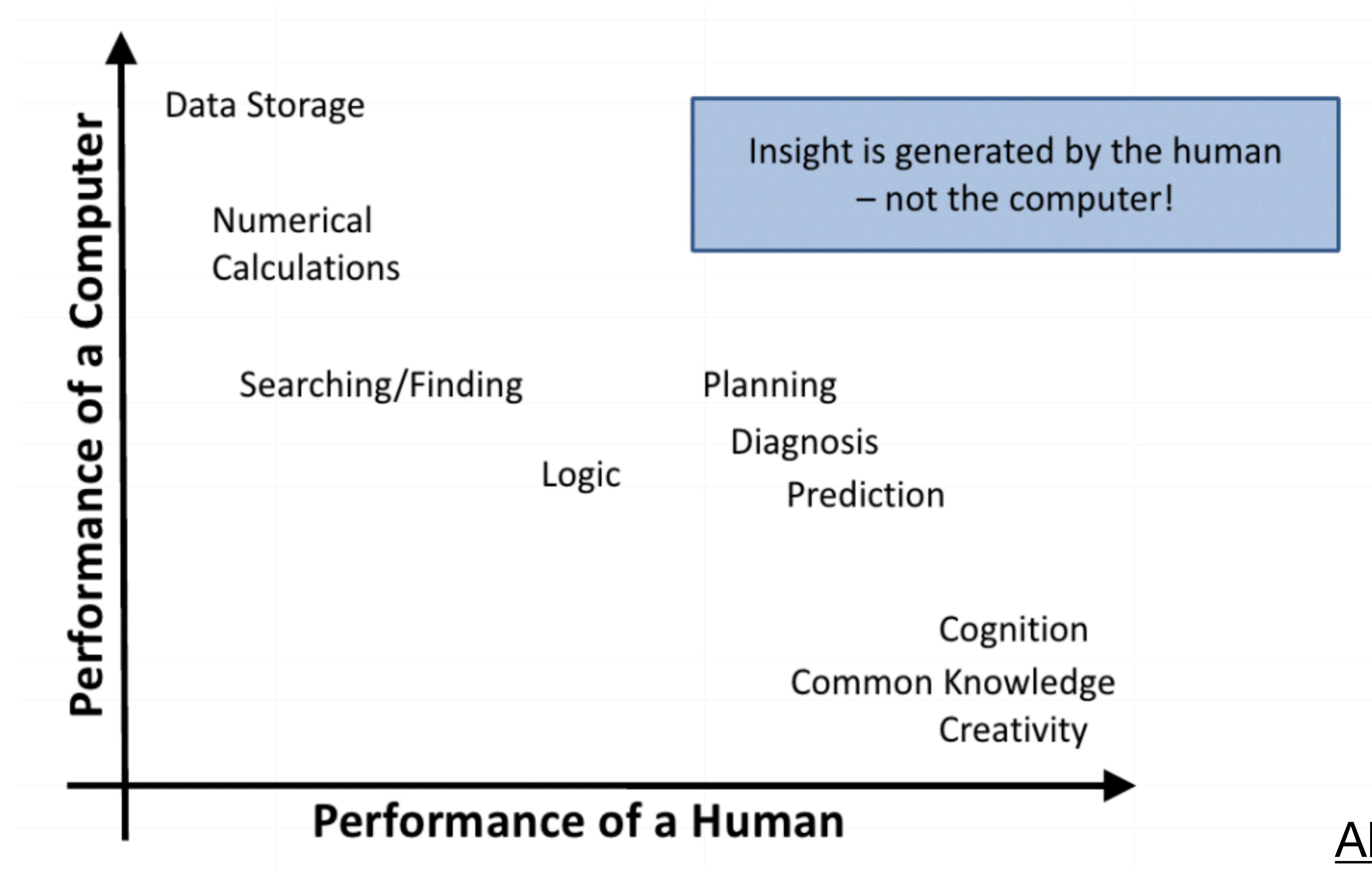
Interactive

Animated

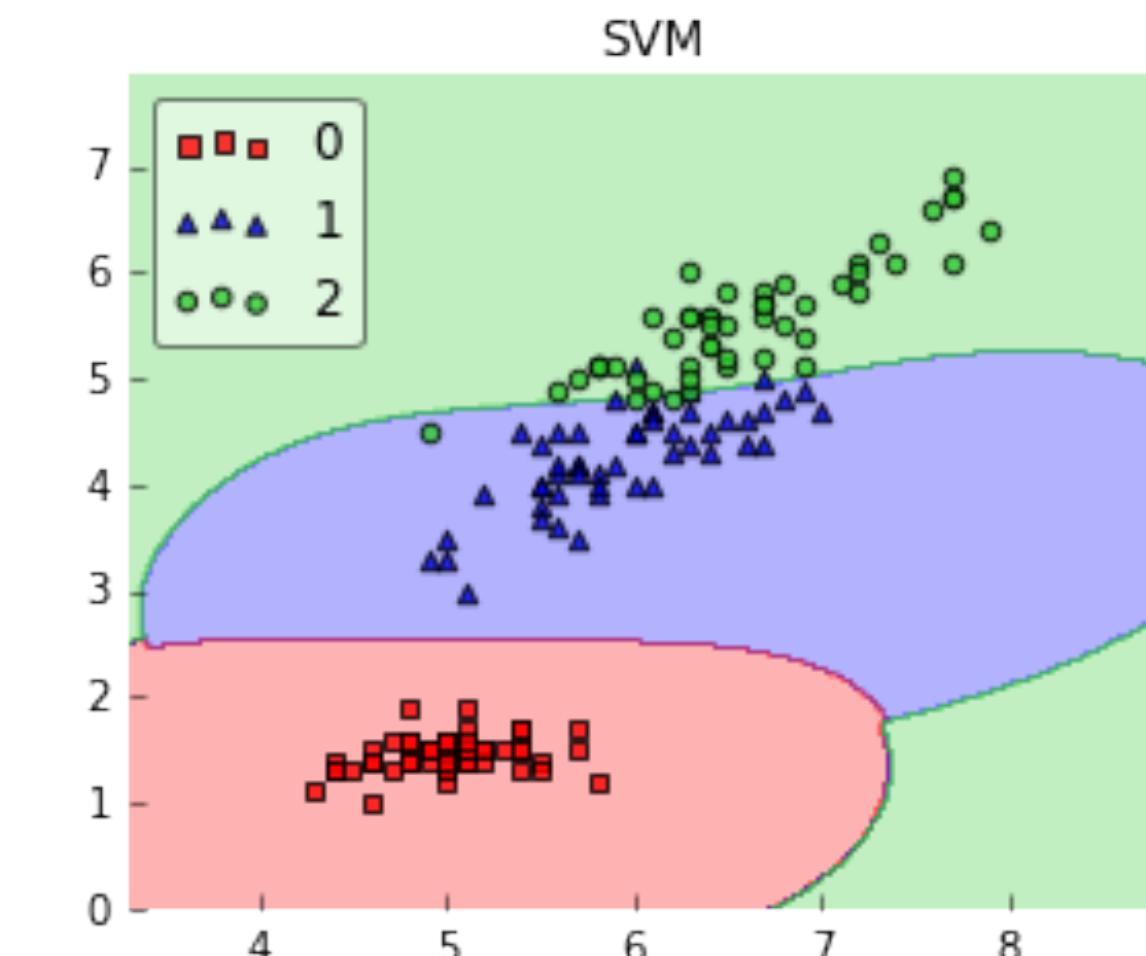
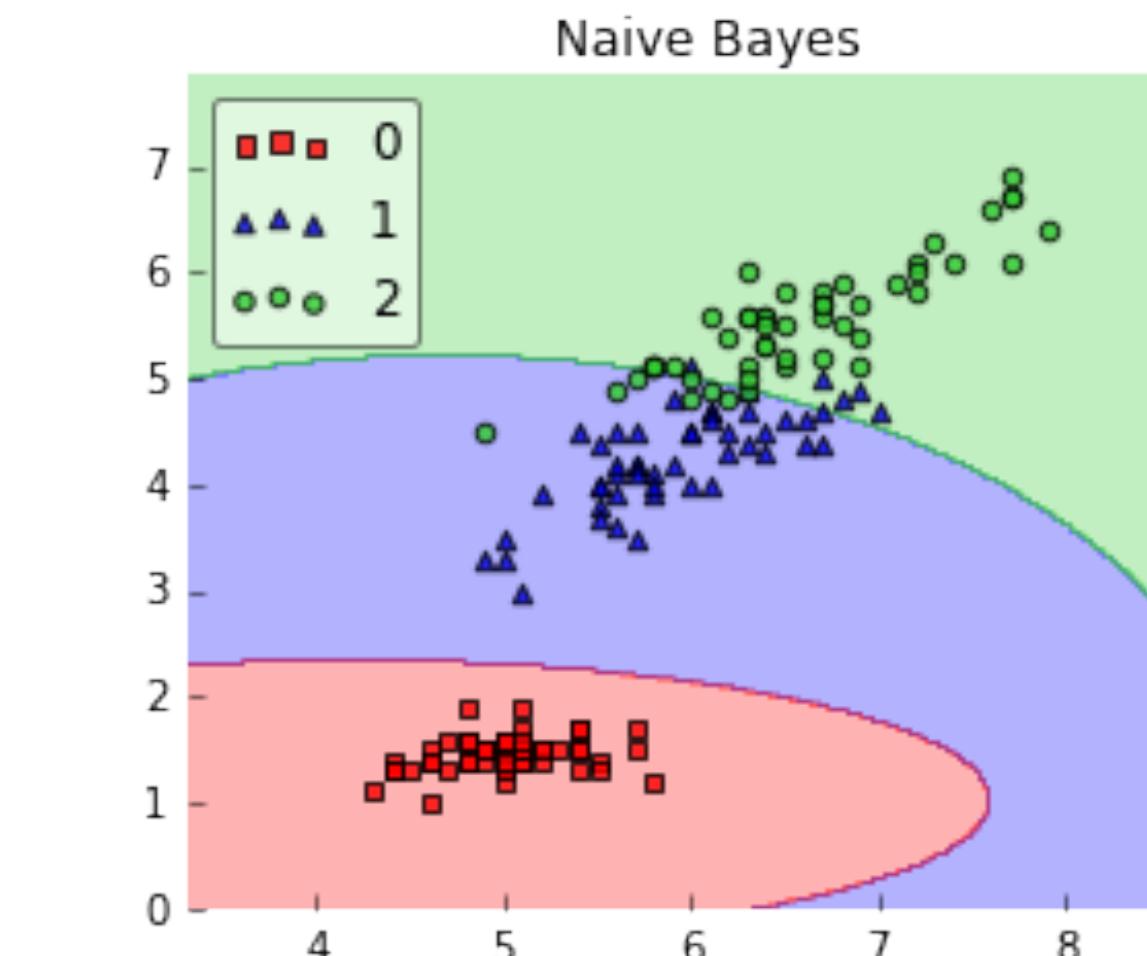
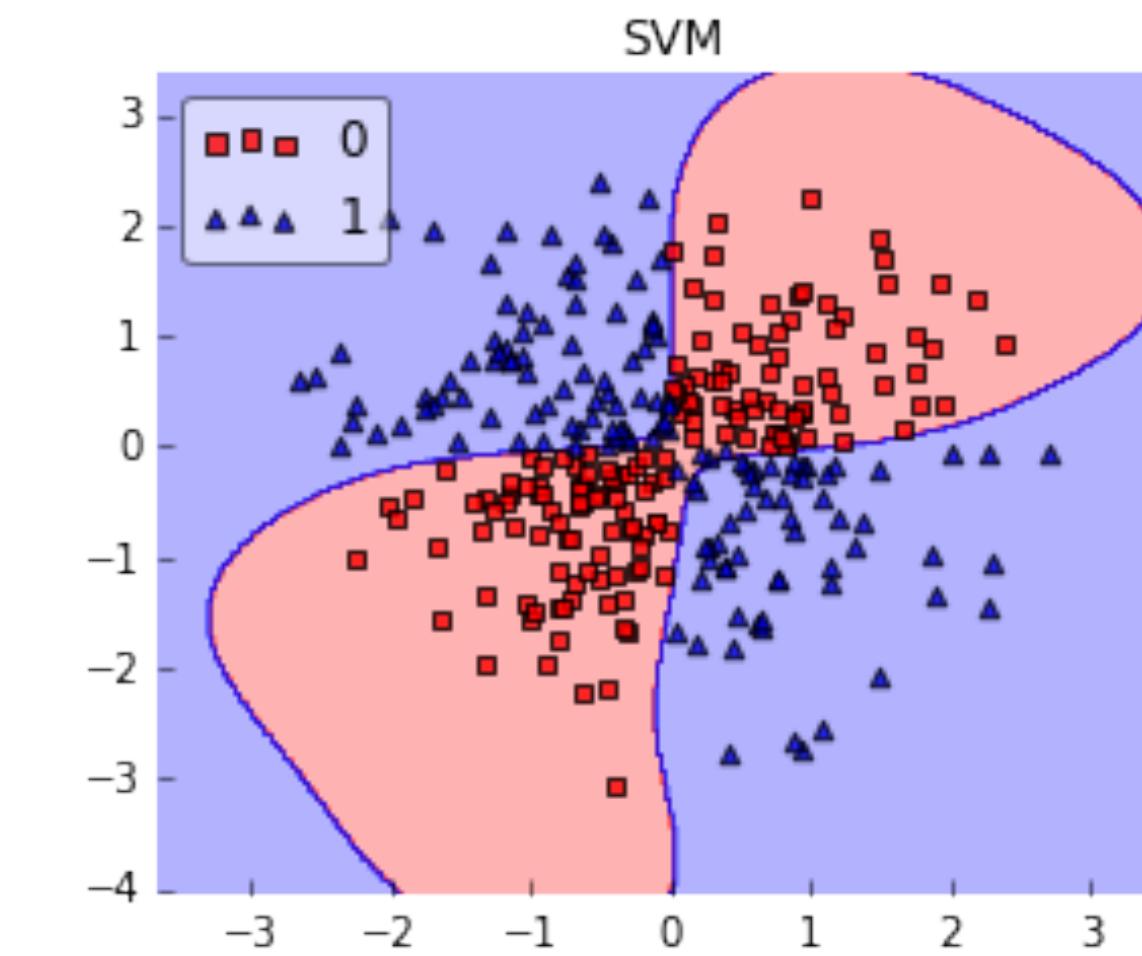
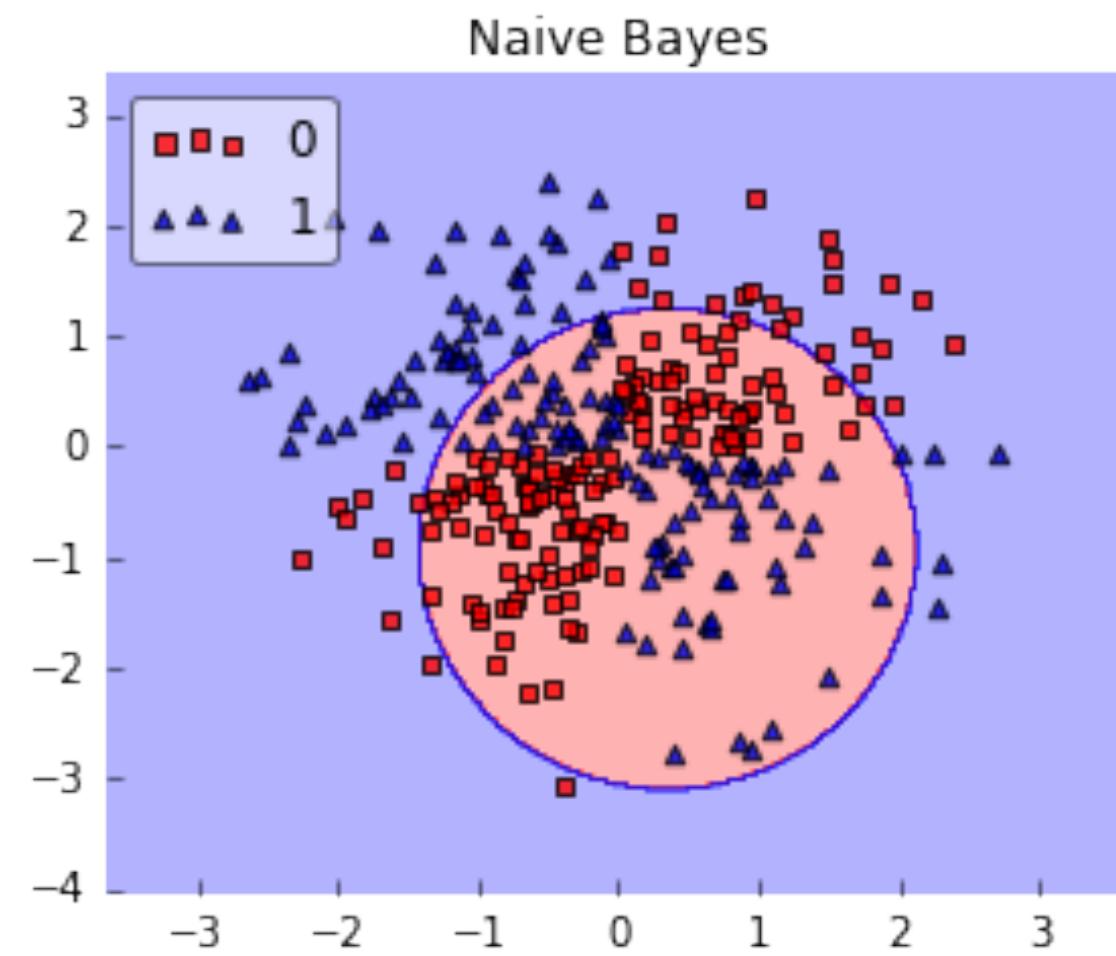
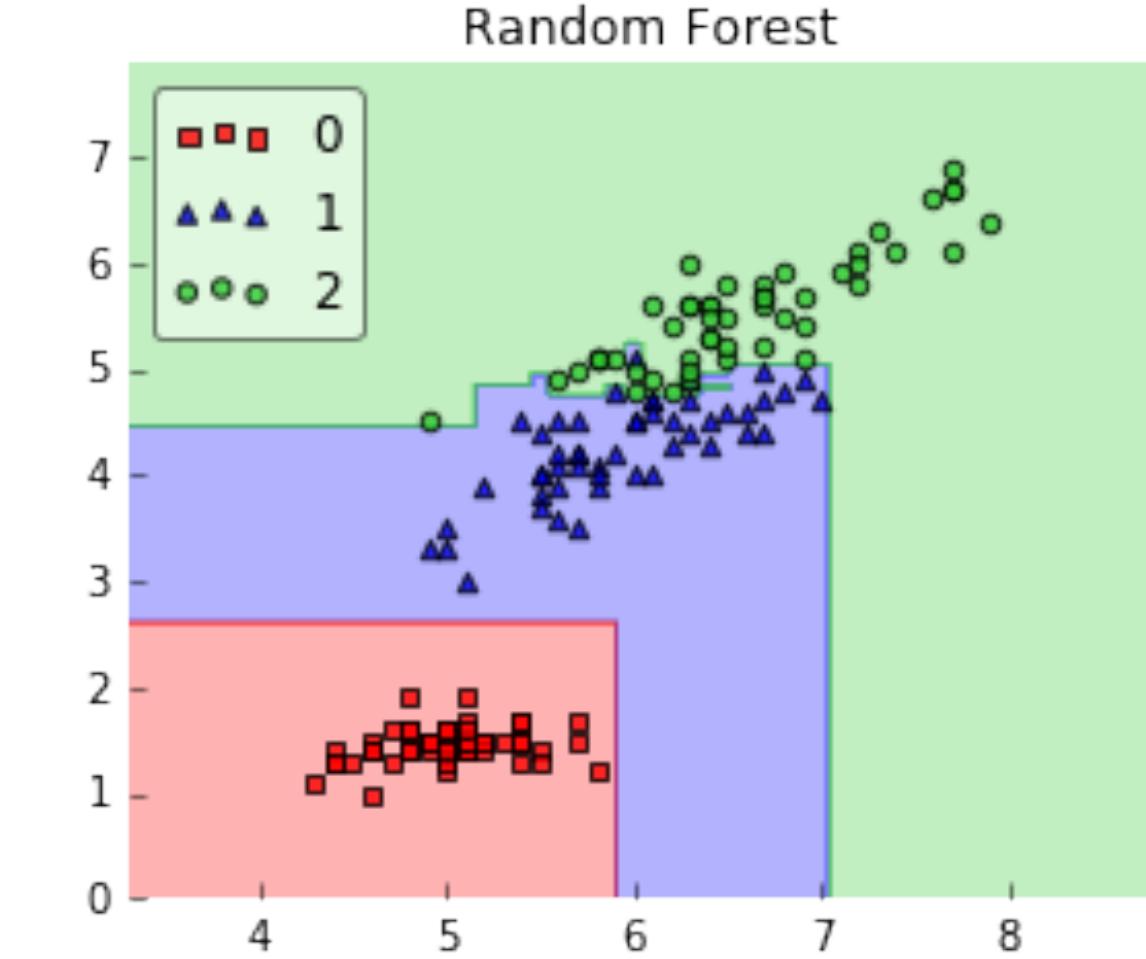
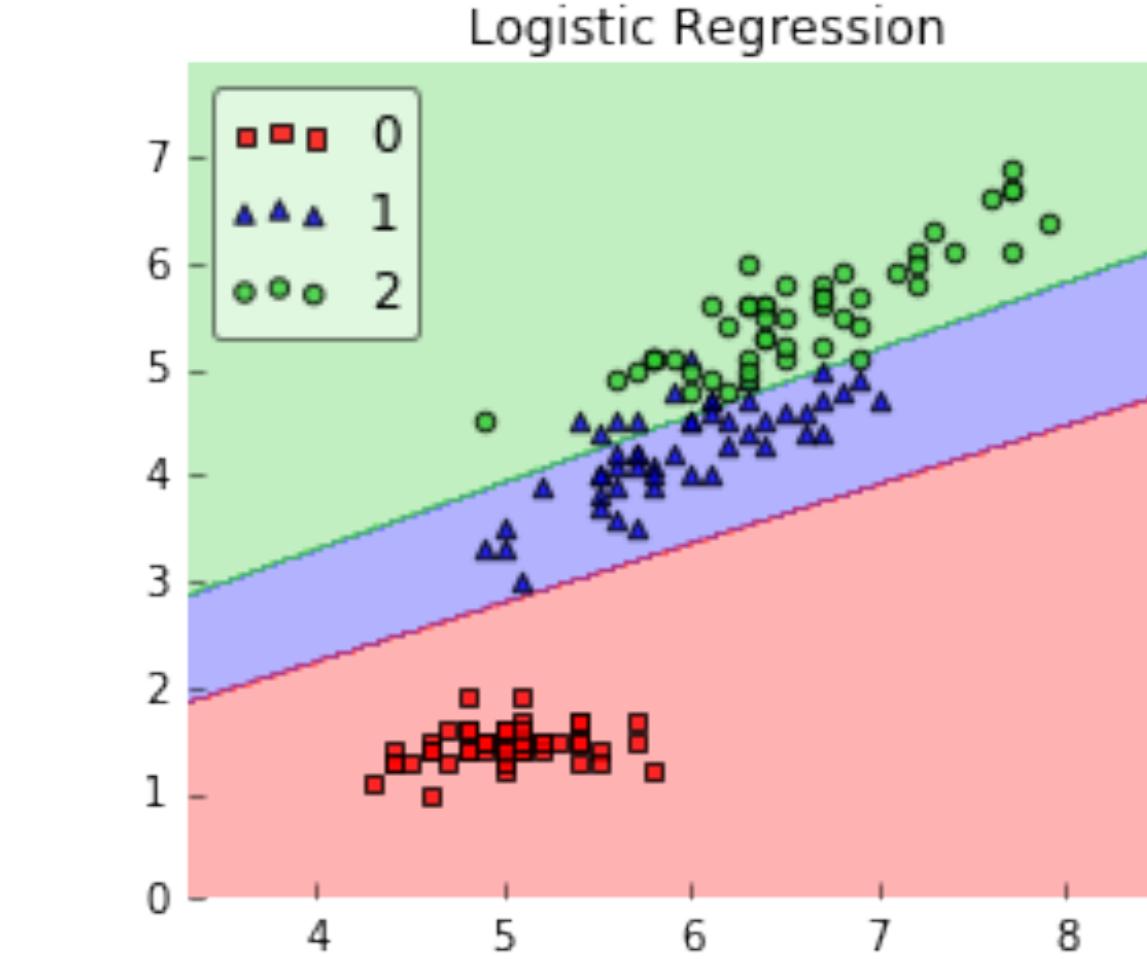
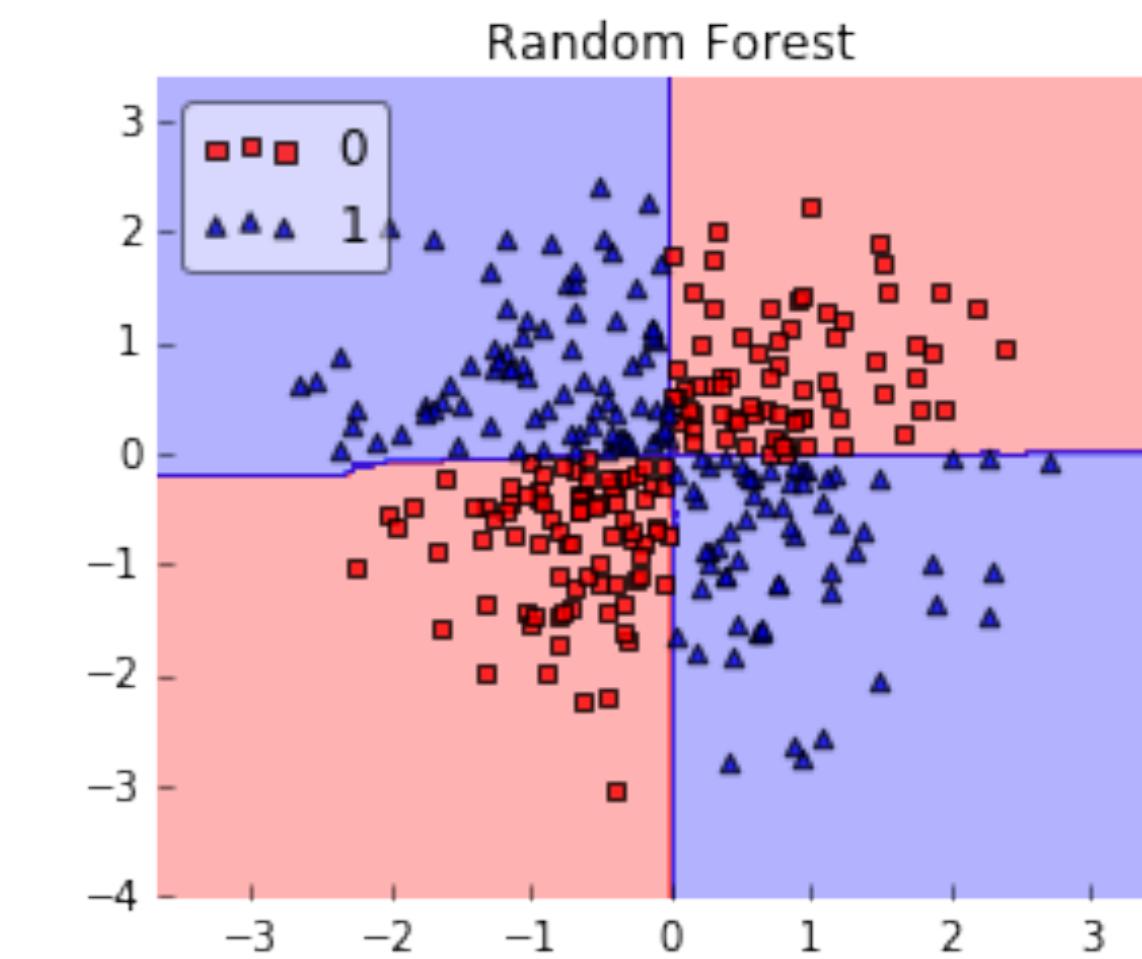
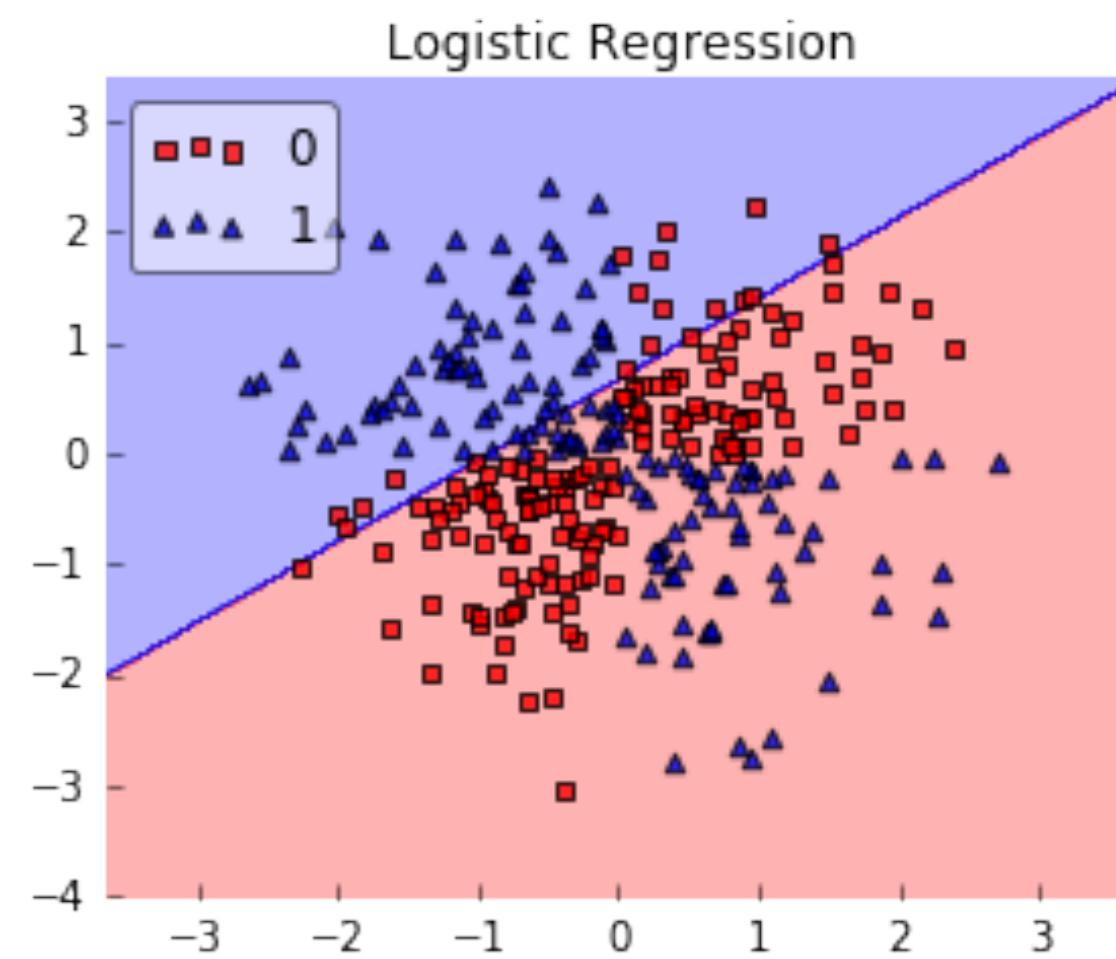
Parallel Coordinates

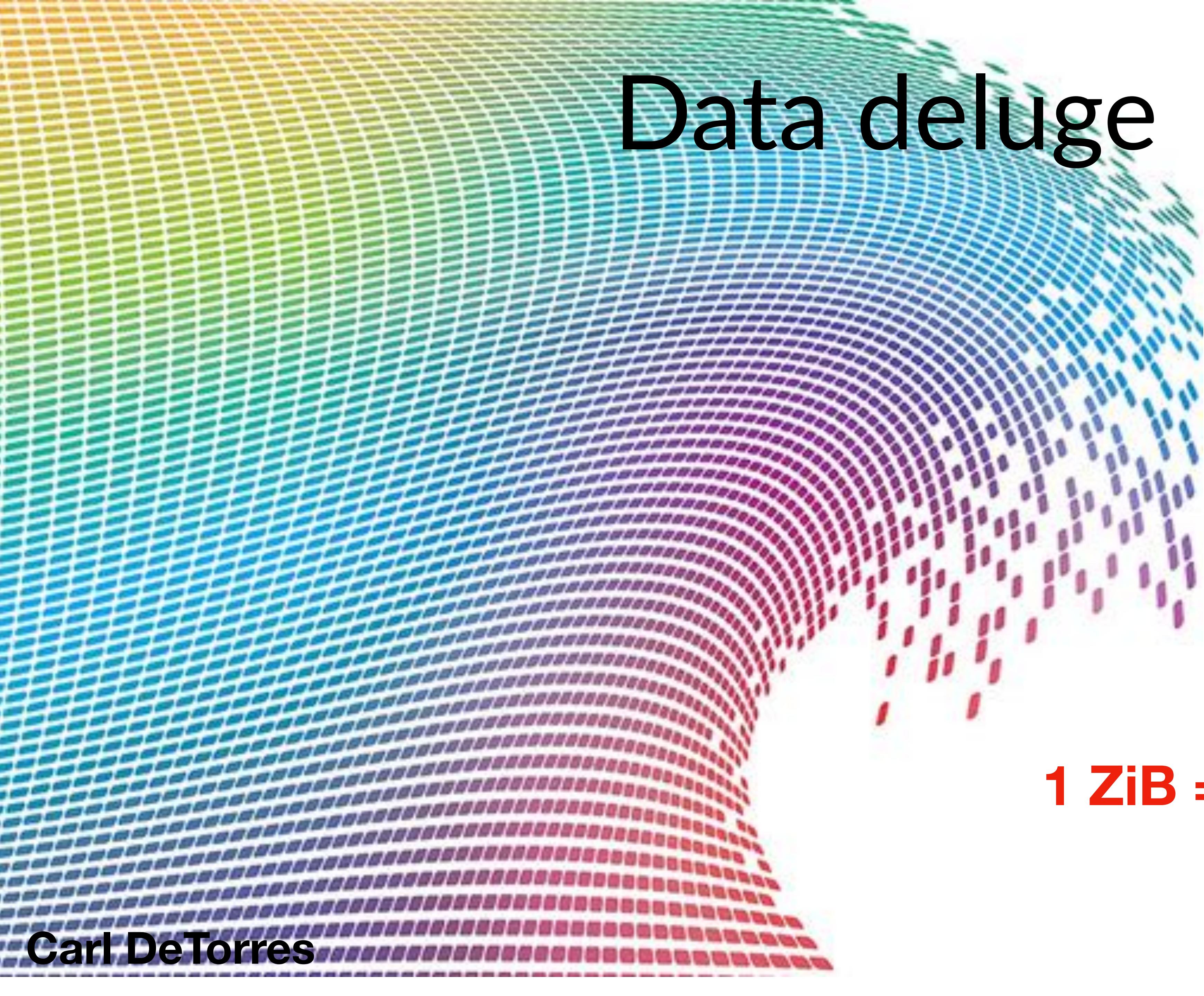


Performance of computers / humans



Human vs computer





Data deluge

2007: 281 Exabytes

2010: Zettabytes barrier

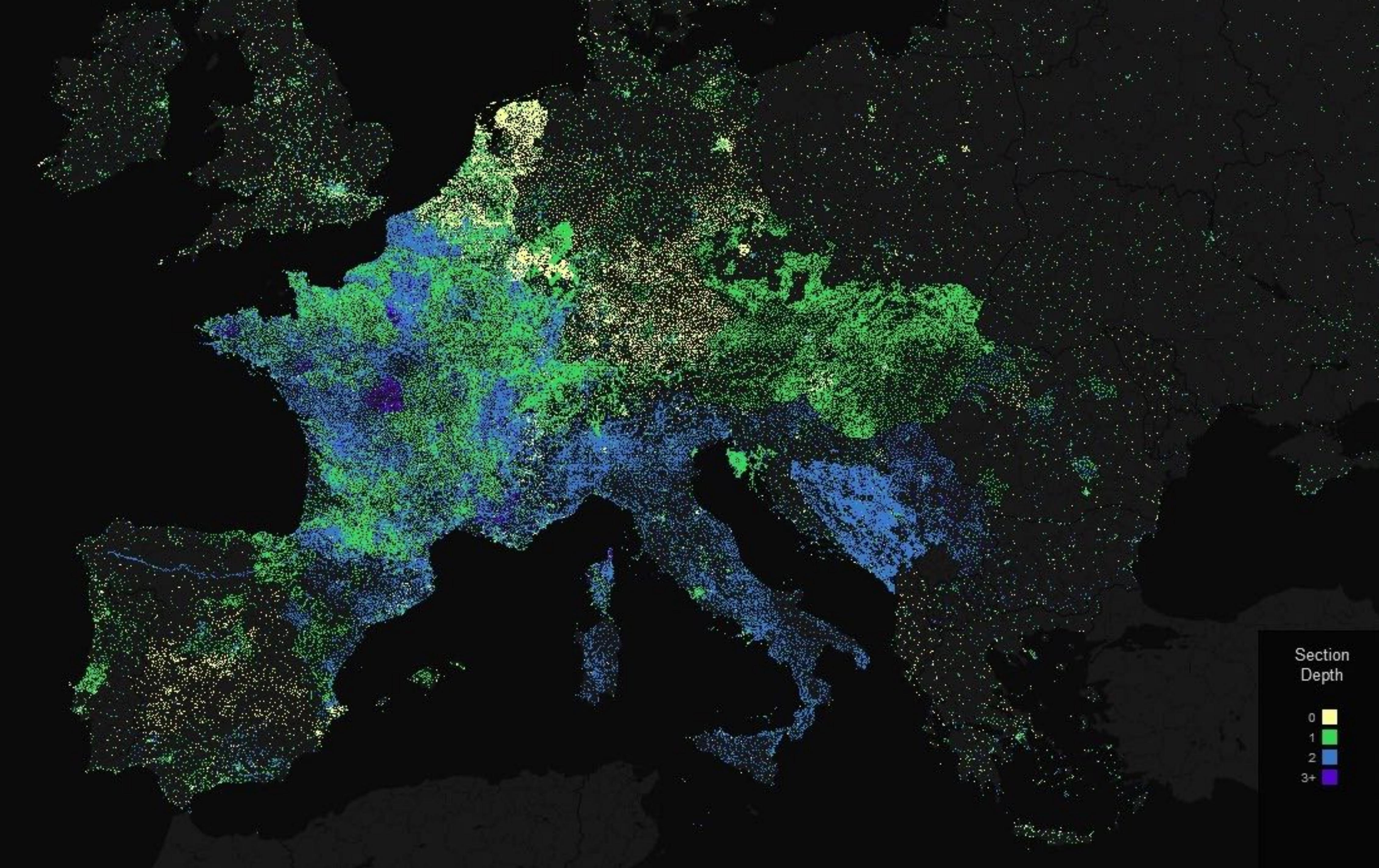
2011: 1.8 Zettabytes

2020: 44 Zettabytes

1 ZiB = 10^{21} bytes

Section
Depth

- 0
- 1
- 2
- 3+





facebook

December 2010

Data science questions

How do we effectively access data?

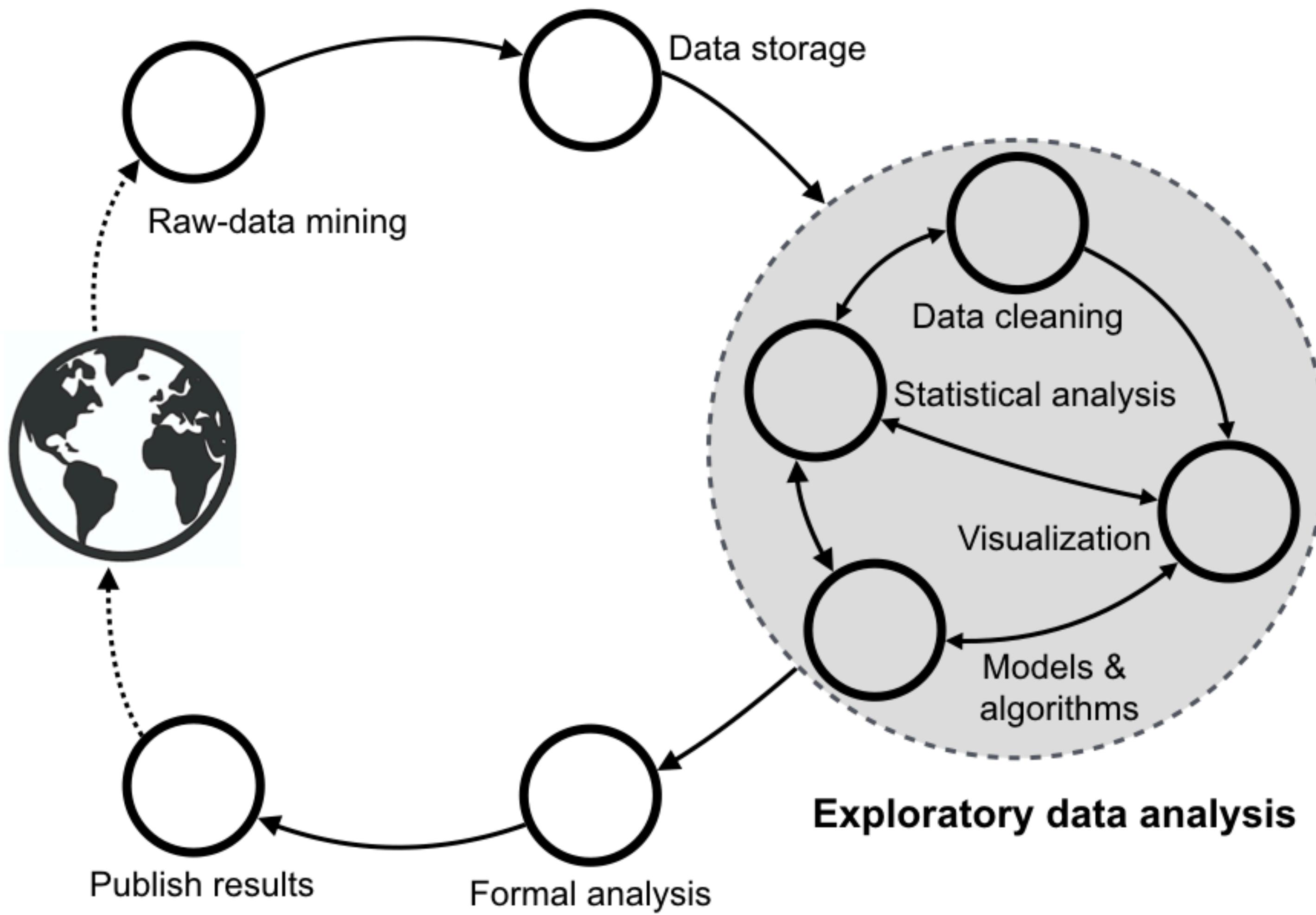
Understand their structure?

Compare, reason?

Gain insights?



Data science pipeline



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

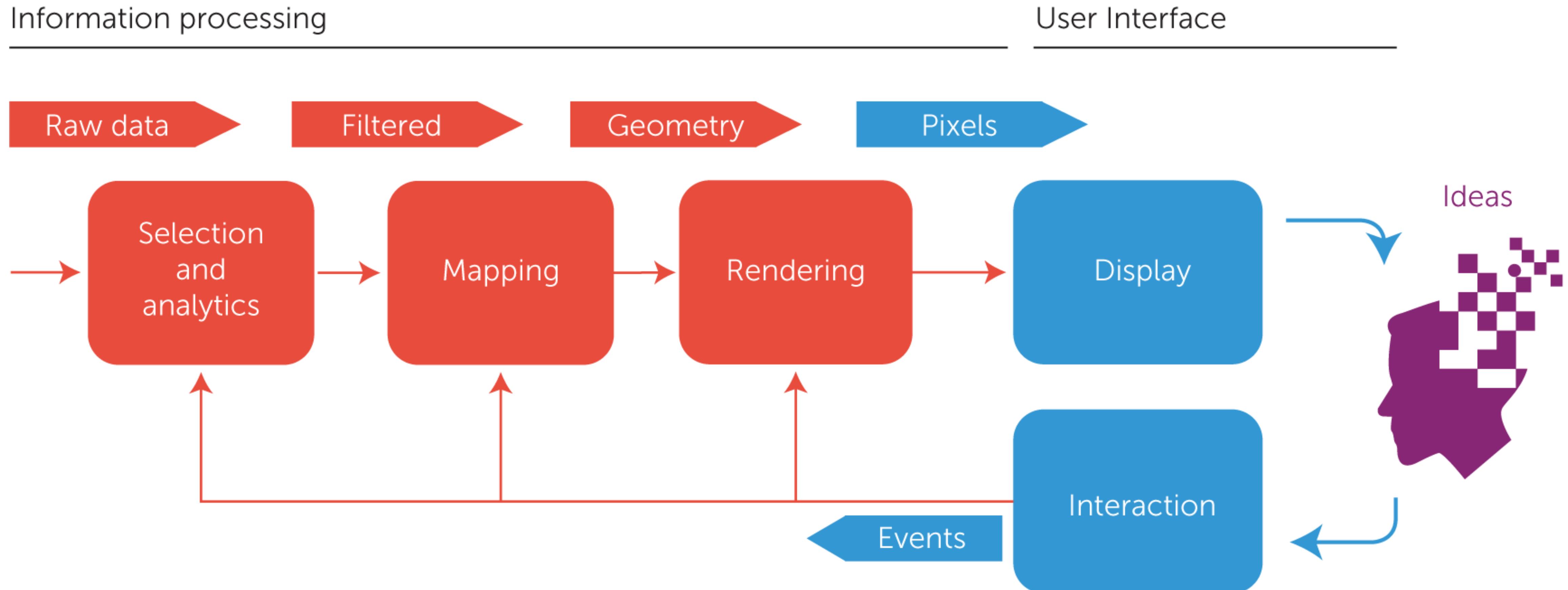
COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau





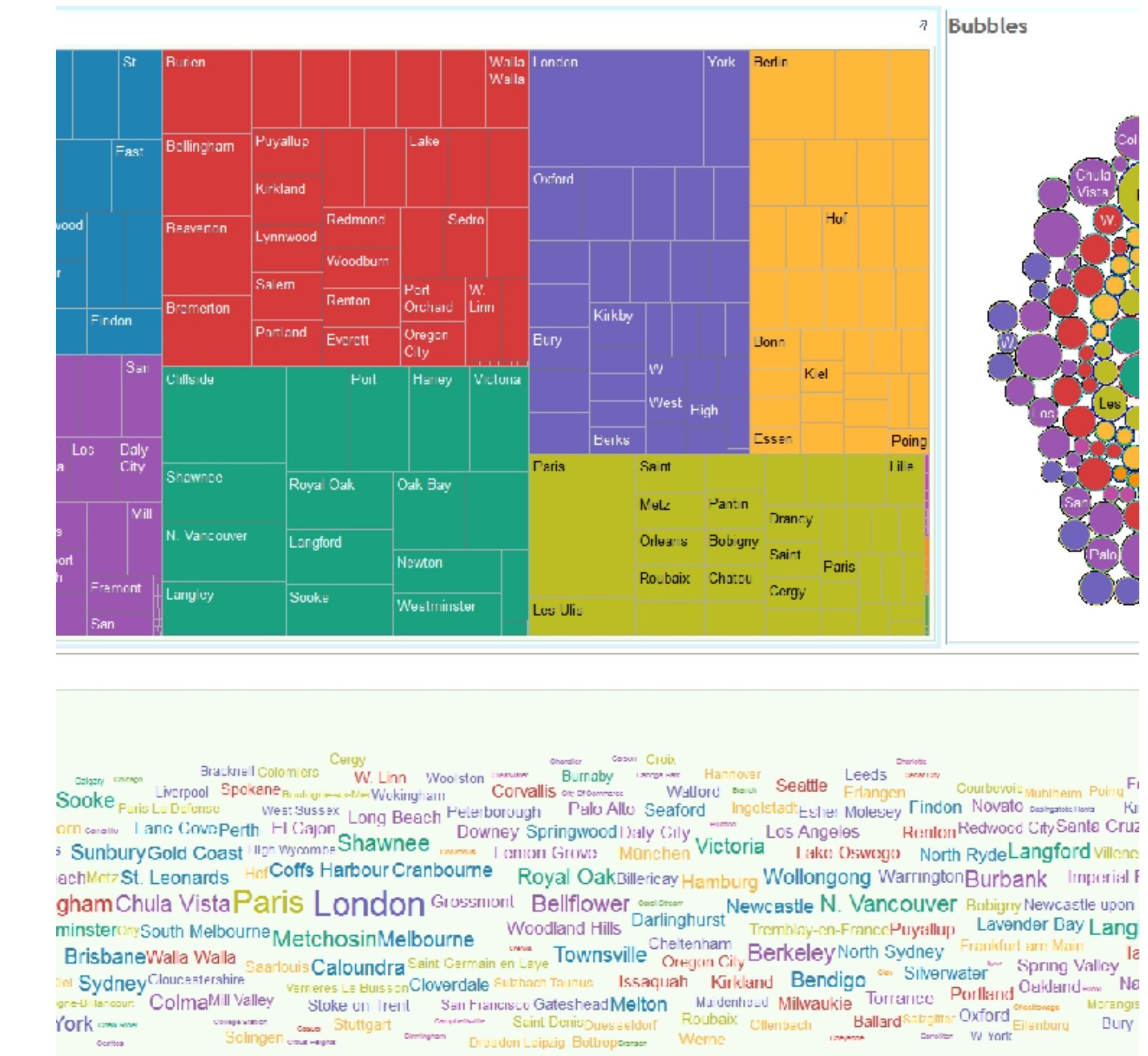
Visualization Pipeline



How to create a data-viz?

Use a commercial software: Tableau,
Qlik, PowerBI, etc.

**Take advantage of the Web stack
to create complex dataviz using
open-source libraries and tools**



p3

Tech stack



Syllabus

1. The Web “stack”

1. HTML, CSS, DOM

2. SVG, Canvas, WebGL

3. ECMA 2015+
(javascript)

4. D3.js

The screenshot shows a code editor interface with three panels:

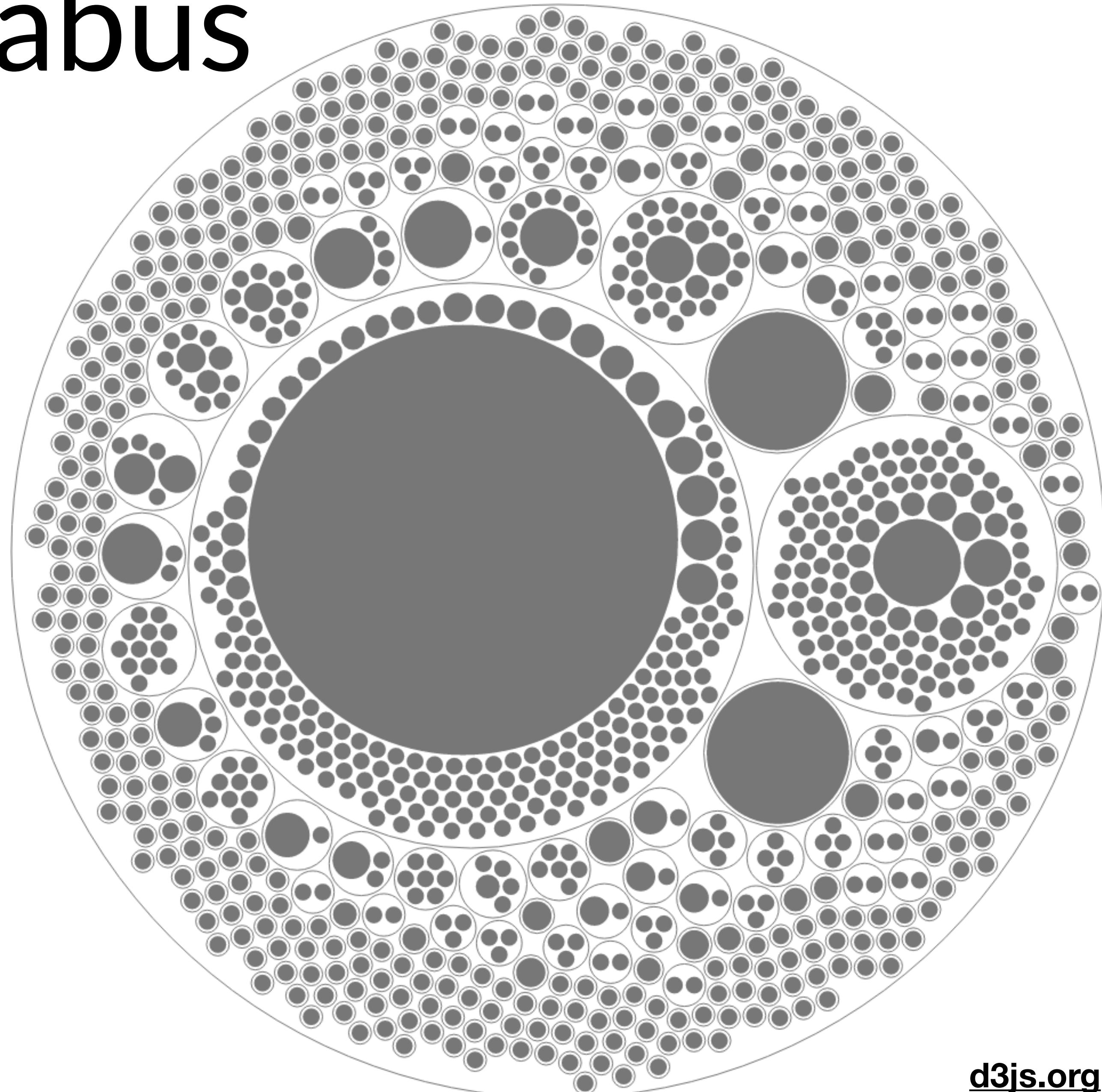
- HTML** panel: Contains the code `<div id="viz"></div>`.
- CSS** panel: Contains the code

```
1 #myrect {  
2   fill: red;  
3   opacity: 0.6;  
4 }
```
- JS (Babel)** panel: Contains the code

```
1 "use strict";  
2 // Test ES6 and d3  
3 const data = [1, 2, 3];  
4 const [x, y] = [70, 10];  
5  
6 let helpers = {  
7   translate: function(x, y) {  
8     this.attr("transform",  
9       `translate(${x}, ${y})`);  
10    return this;  
11  }  
12}  
13
```

Syllabus

- 2. Visualization fundamentals
 - 1. Data
 - 2. Interaction, filtering, aggregation
 - 3. Perception, cognition, color
 - 4. Marks and channels
 - 5. Designing visualizations



Syllabus

3. Techniques and algorithms

1. Maps

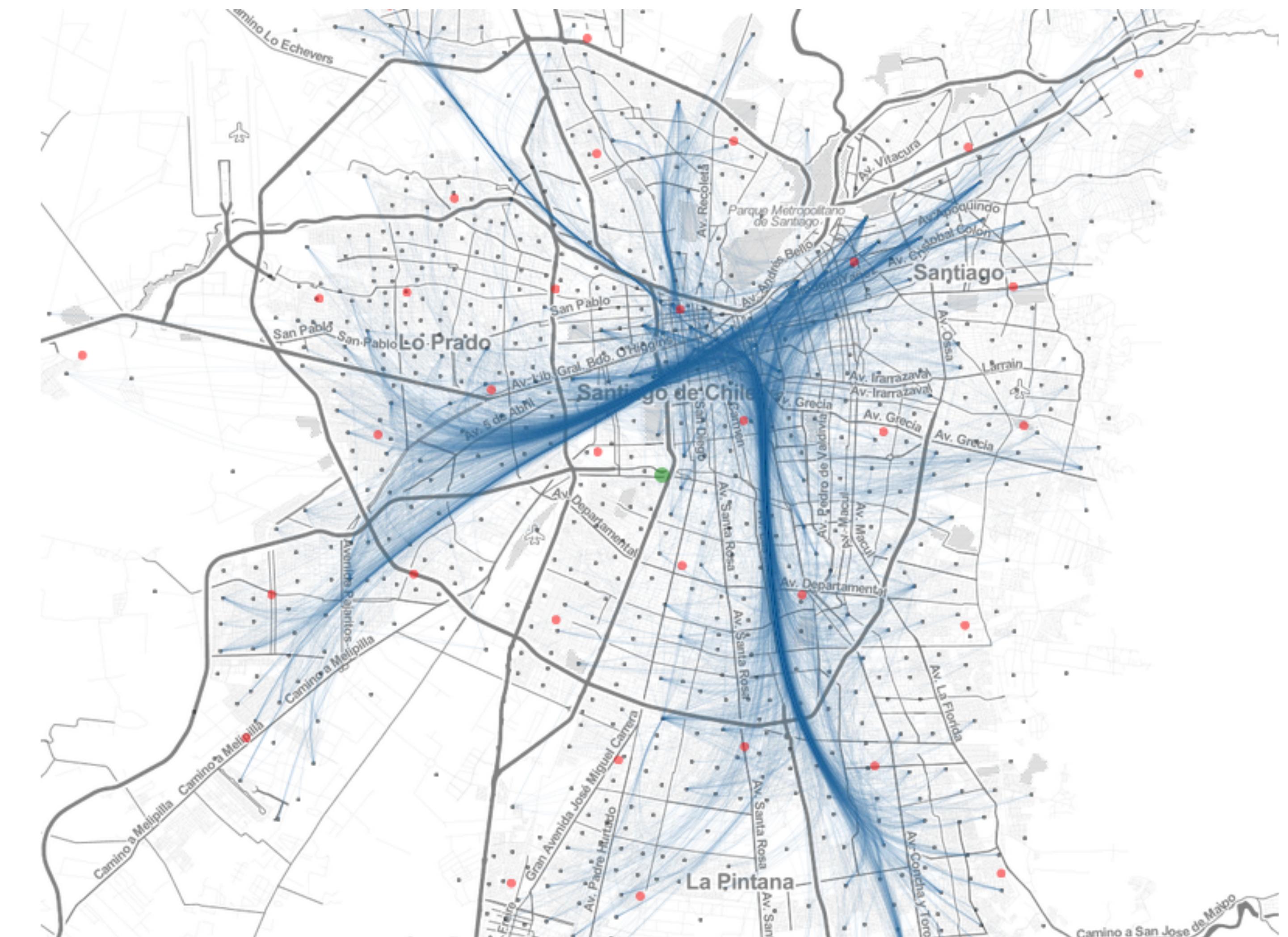
2. Text

3. Trees, graphs

4. Tabular data

5. Sound

6. Volumes



Structure and goals

COM-480

Learning objectives

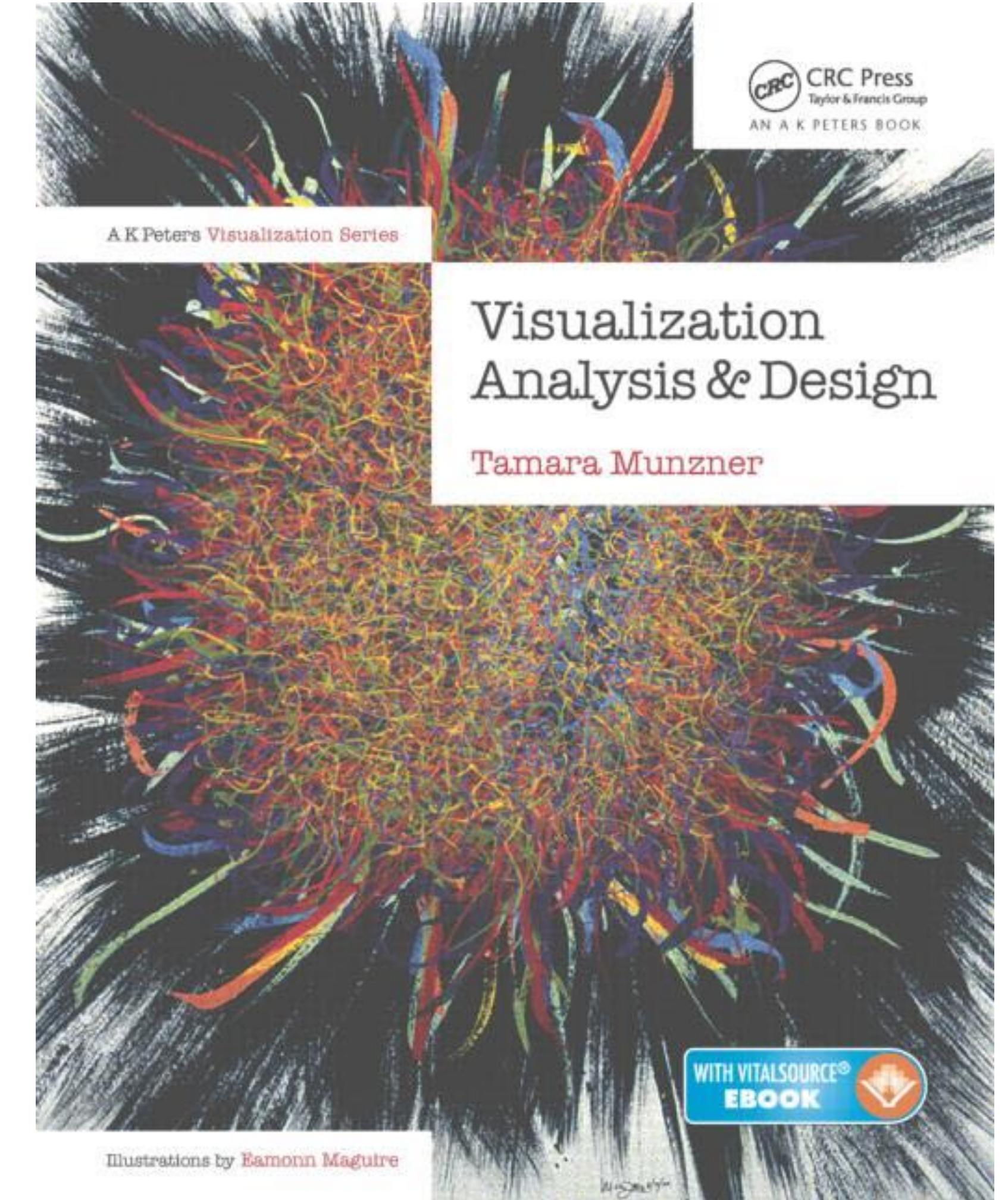
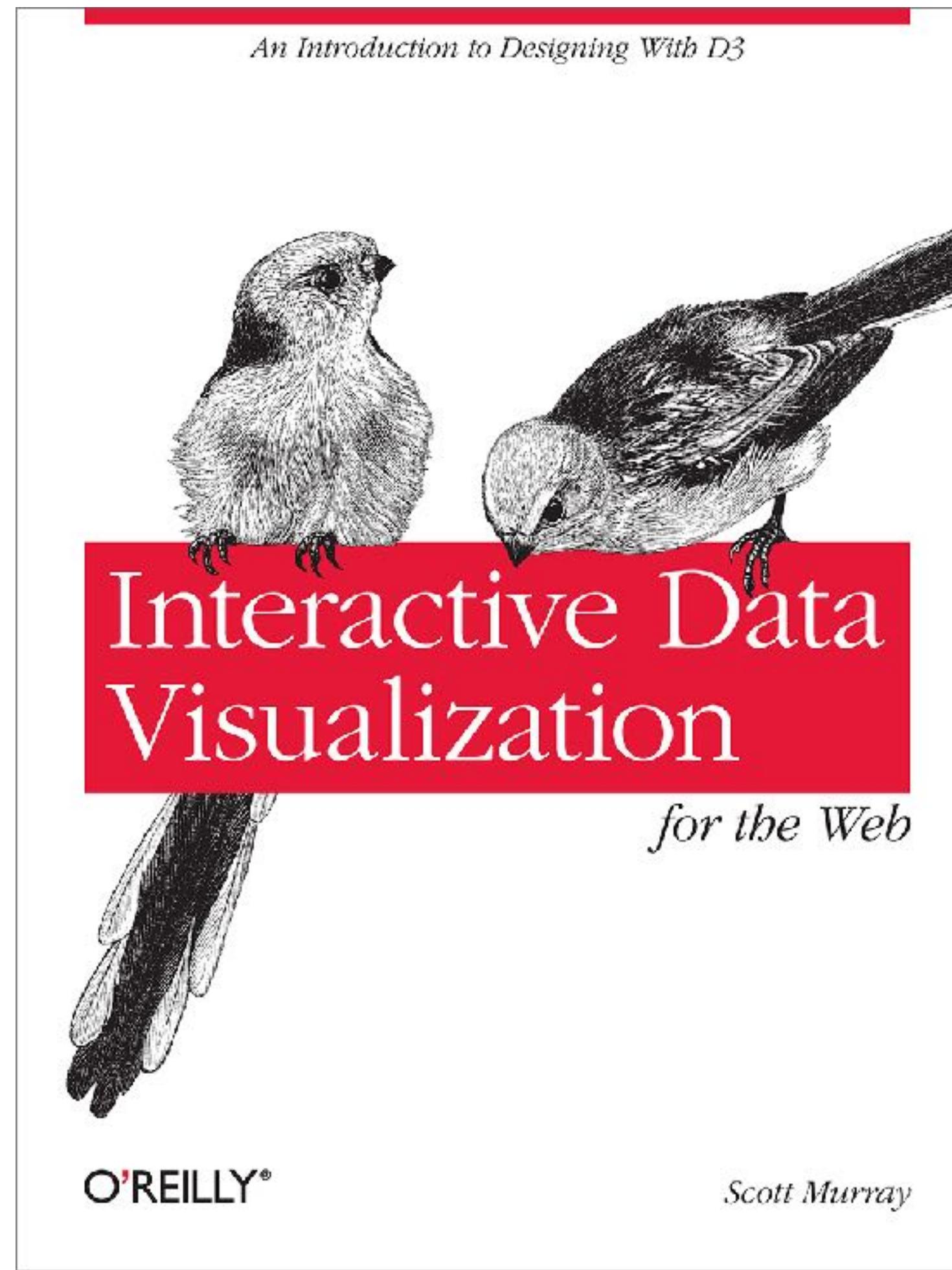
Efficiently visualize data

Design visual data analysis solutions

Implement interactive data visualizations

Acquire real web development skills for data science

Required textbooks



<http://www.crcnetbase.com/isbn/9781466508910>

<http://chimera.labs.oreilly.com/books/1230000000345>

Prerequisites

**Good programmer (C++,
Java, Python, etc.)**

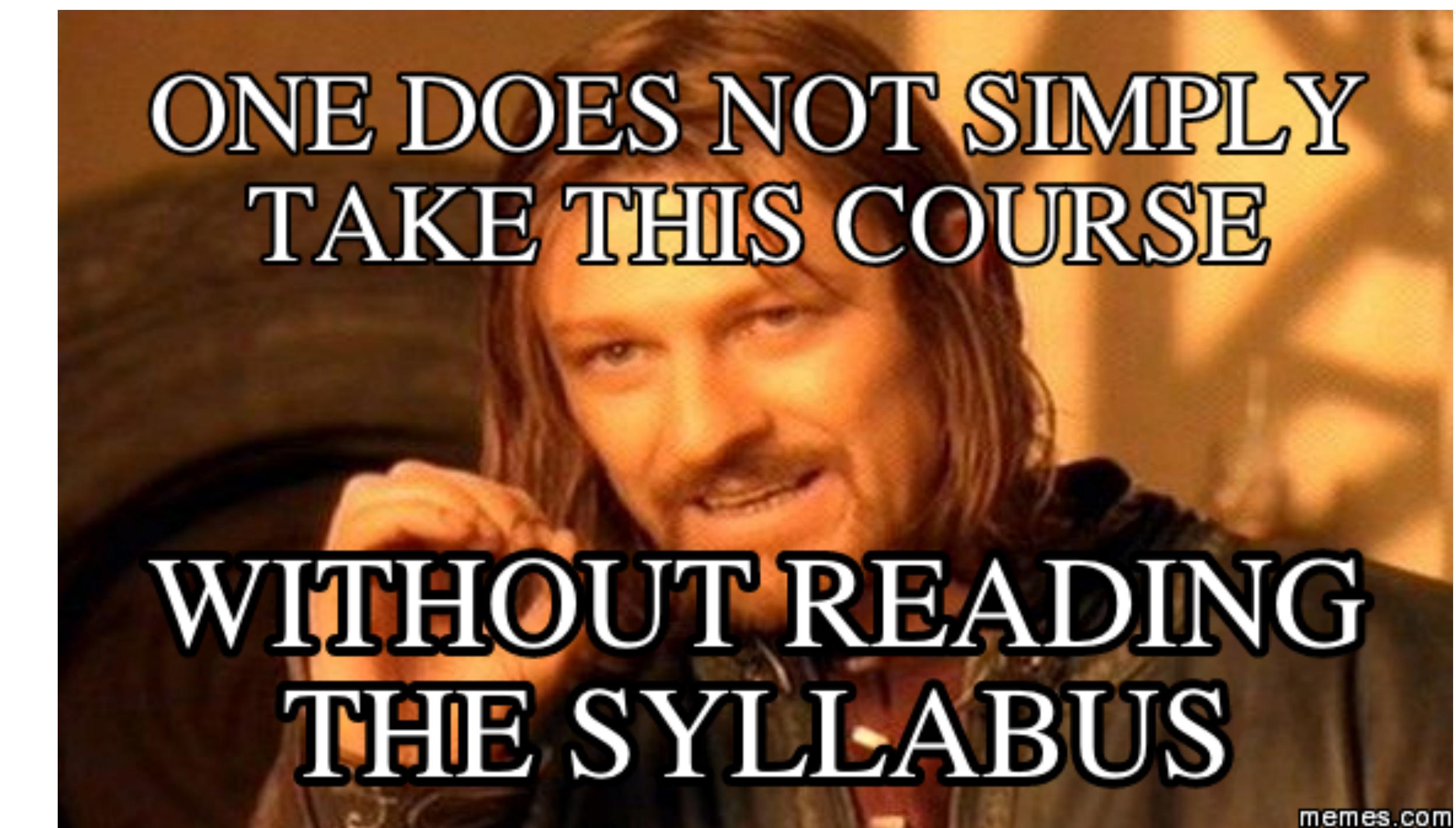
Autonomy

Willingness to learn

CS-305 Software engineering (BA)

CS-250 Algorithms (BA)

CS-401 Applied data analysis (MA)



Grading for group project

Visualization (35%)

Technical Implementation (15%)

Website, presentation, screencast (15%)

Process book (35%)

Cheating (don't)

Write your own code

Design your own visualizations

Critically evaluate the results in your own words.

All the projects will be automatically checked for plagiarism!!

About you?