DATA VISUALIZATION

KIRELL BENZI, PH.D

@KirellBenzi

www.kirellbenzi.com

About Kirell Benzi

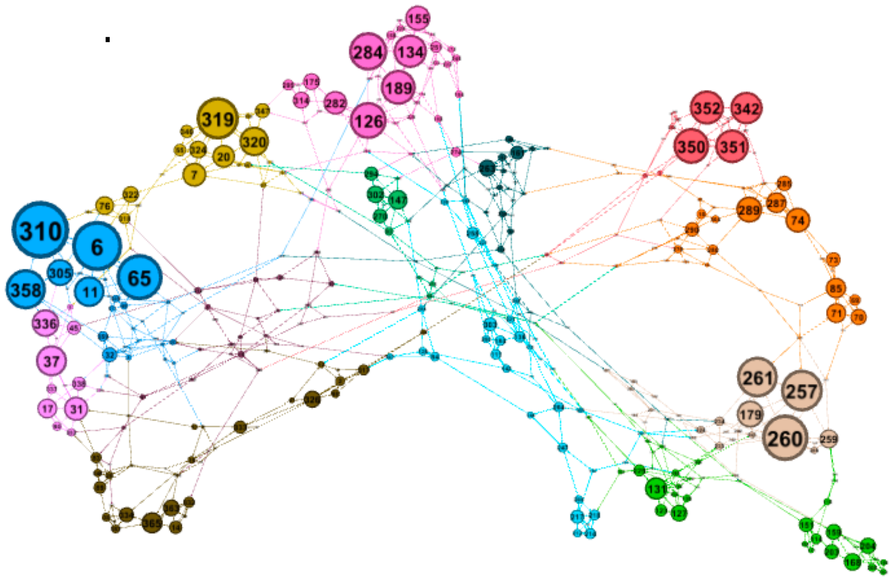
Got his Ph.D from EPFL in 2016, in network science in the LTS2 led by Prof. Vandergheynst.

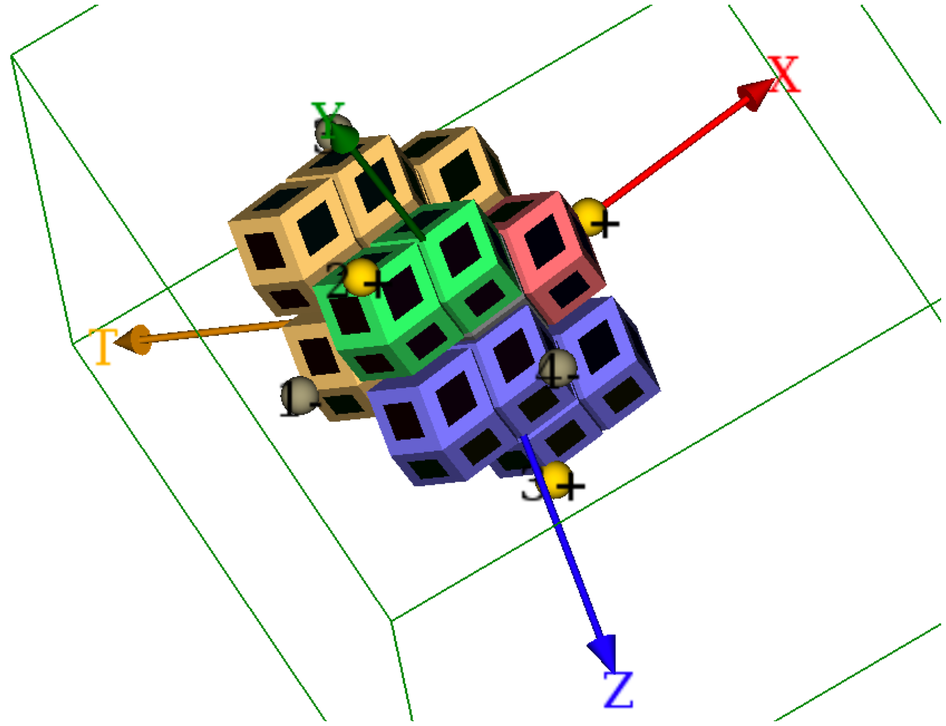
CEO, founder of Kirelion a creative data science studio

Data artist, exhibition “Singular Networks” in several countries

Laurent Vuillon

- Professor of discrete mathematics and computer science
- Graph theory, algorithms on discrete structures
- Visualization of mathematics





NLP and clustering

OPÉRA 29

DON GIOVANNI
MOZART

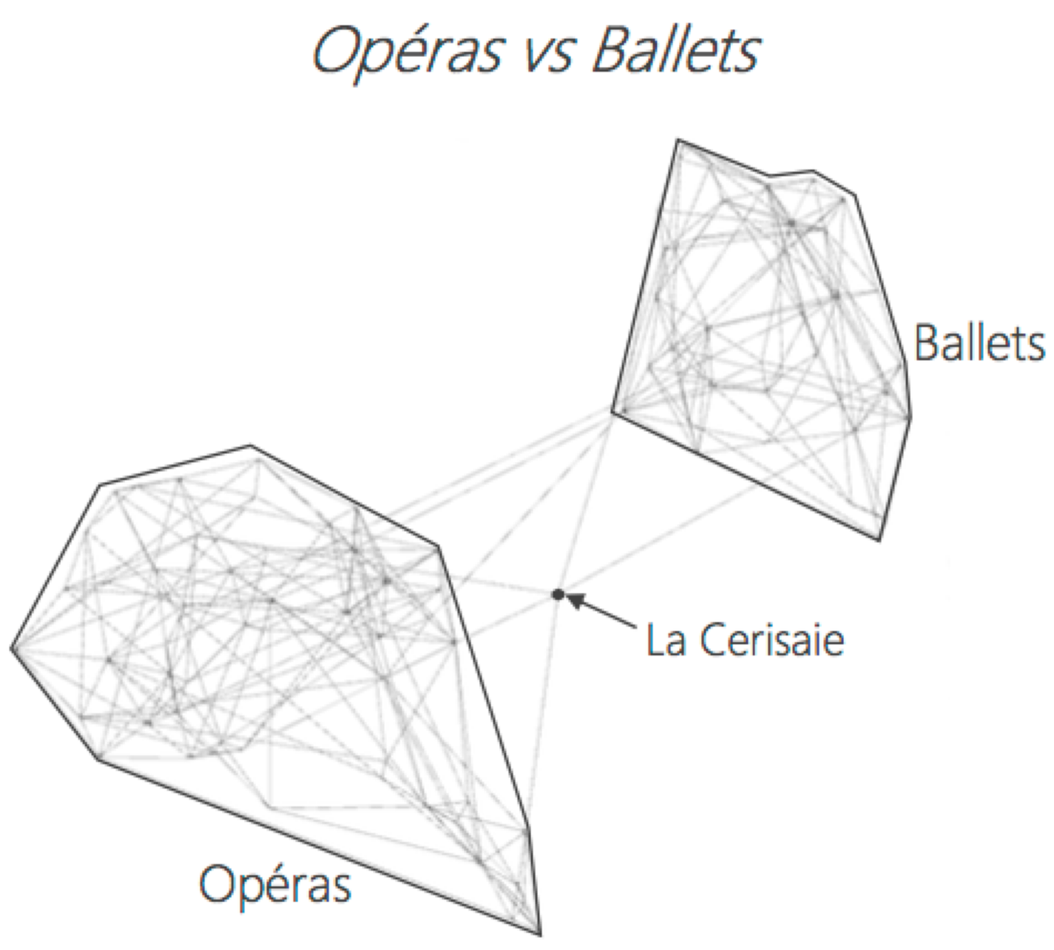
DRAMMA GIOCOSO EN DEUX ACTES *(1787)*
MUSIQUE DE WOLFGANG AMADEUS MOZART *(1756-1791)*
LIVRET DE LORENZO DA PONTE
En langue italienne

ALAIN ALTINOGLU *Direction musicale*
MICHAEL HANEKE *Mise en scène*
CHRISTOPH KANTER *Décors*
ANNETTE BEAUFAŸS *Costumes*
ANDRÉ DIOT *Lumières*

ERWIN SCHROTT *Don Giovanni*
LIAN LI *Il Commendatore*
TATIANA LISNIC *Donna Anna*
STEFAN POP *Don Ottavio*
MARIE-ADELINE HENRY *Donna Elvira*
ADRIAN SÂMPETREAN *Leporello*
ALEXANDRE DUHAMEL *Masetto*
SERENA MALFI *Zerlina*

ORCHESTRE ET CHŒUR
DE L'OPÉRA NATIONAL DE PARIS

Des trois opéras écrits avec Da Ponte, *Don Giovanni* est sans doute le plus noir, le plus désespéré. Autour du séducteur et avec lui, tous les personnages y sont hors d'haleine et hors d'eux-mêmes. Et Mozart leur a donné sa musique la plus ombrageuse, la plus haletante, la plus extrême, la plus parfaite aussi. Pierre Jean Jouve l'évoquait en ces termes : « En cet ouvrage inspiré, l'instinct est capable d'une telle Hystérie, au sens sacré du terme, d'une telle variété de comportements d'ivresse et de néant, de positif suprême et de négatif absolu, que nous devons (nous qui contenons les mêmes tendances à son image) rouler avec lui, de sphère en sphère, comme lui, sans connaître le repos. Nous poursuivons une aventure dans les éléments sombres de l'homme, sans jamais quitter le cadre infiniment doré de la parfaite beauté élucidée et devenue claire. »
Alain Altinoglu dirige la production désormais légendaire du metteur en scène et cinéaste autrichien Michael Haneke.

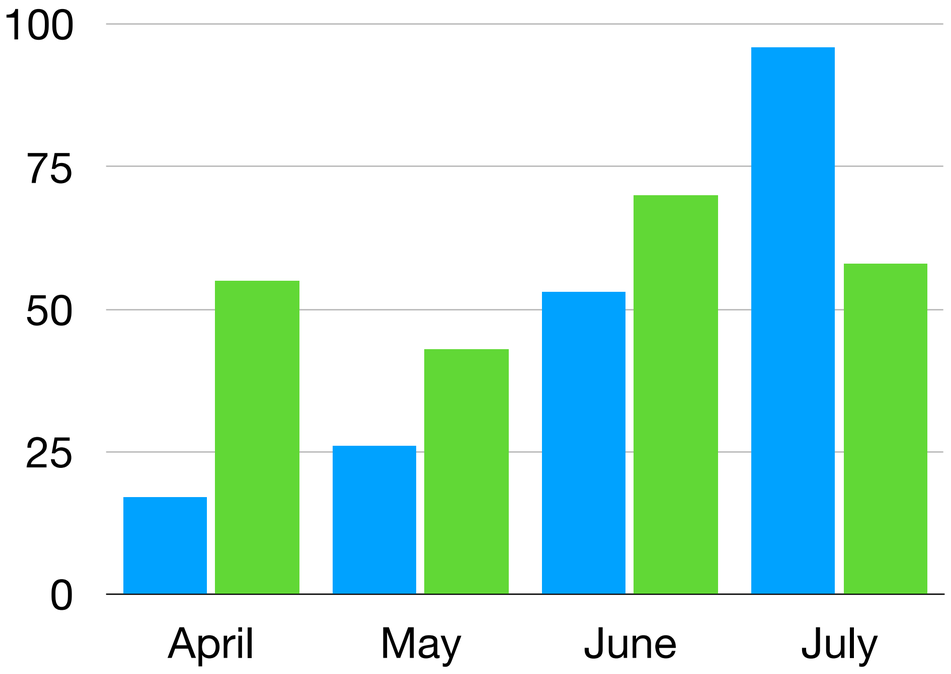


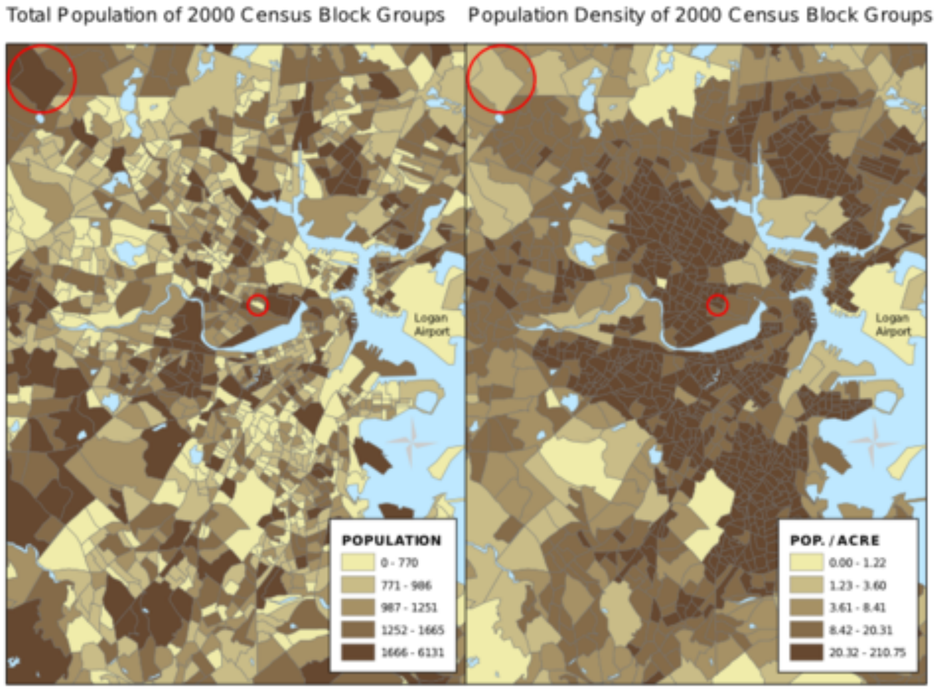
“The purpose of visualization is insight, not pictures.”

–Ben Shneiderman

Defining data visualization

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





Wikipedia

Other definitions

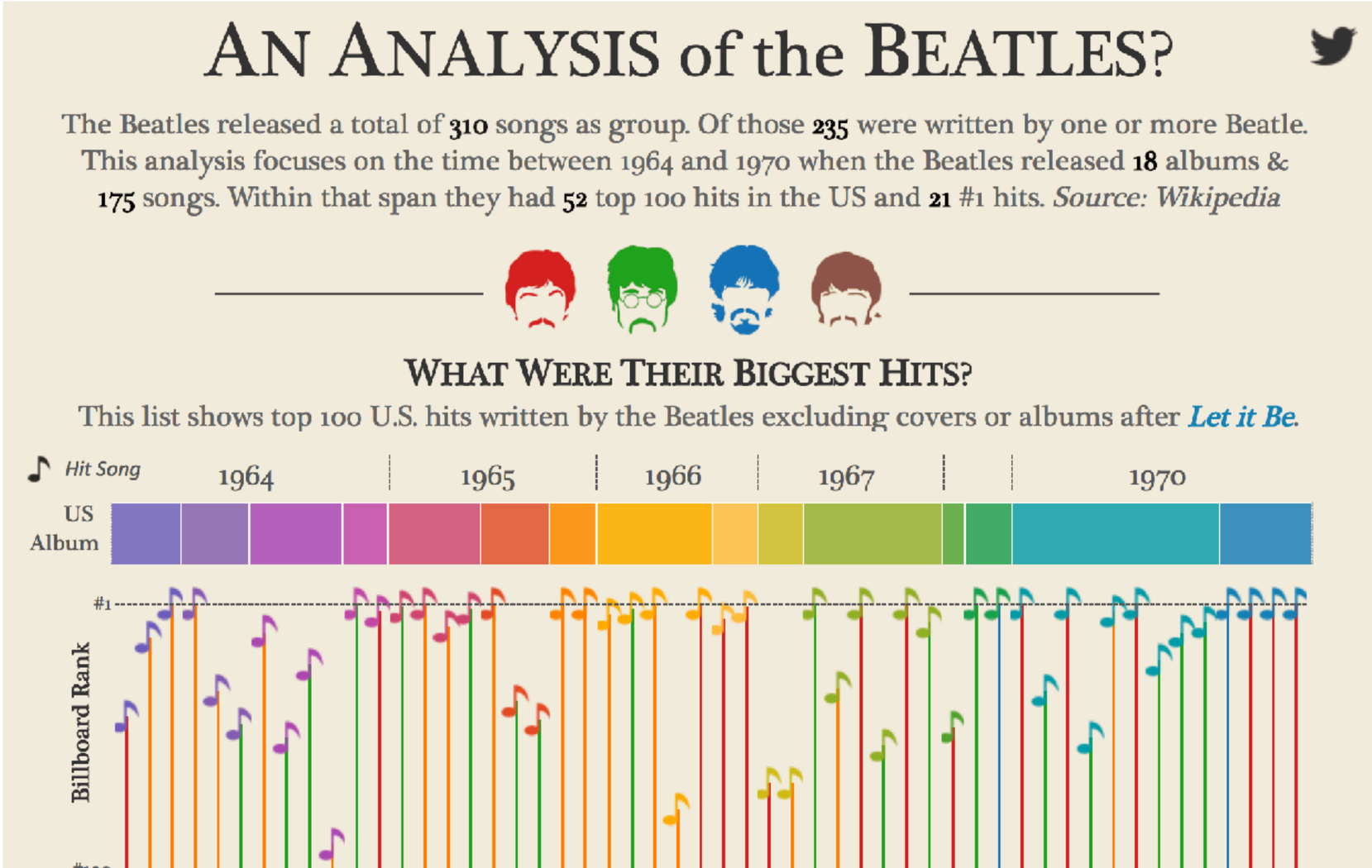
Infographics

Scientific visualization

Data art

http://duelingdata.blogspot.fr/2016/01/the-beatles.html

Infographics



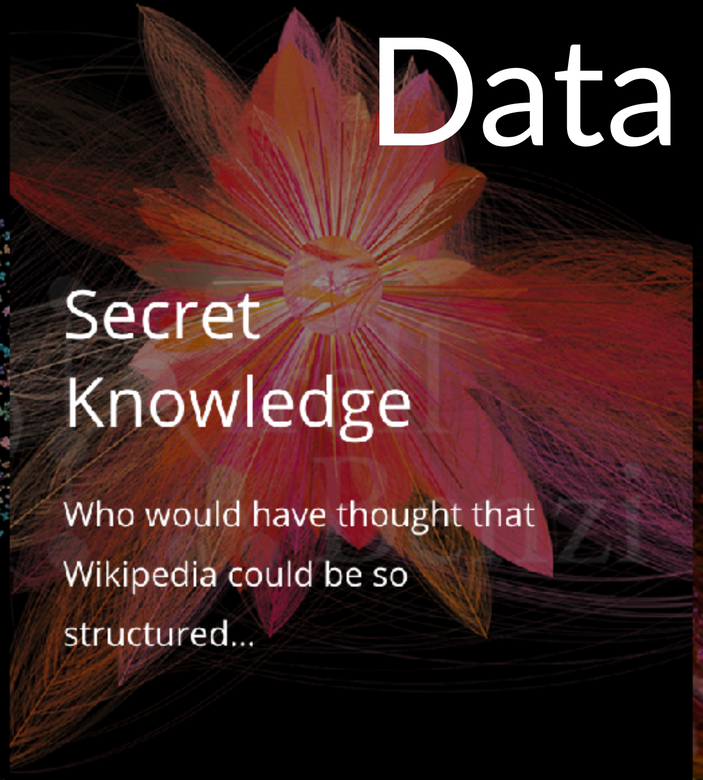
Scientific viz

http://radiotzanck.com/

Data art

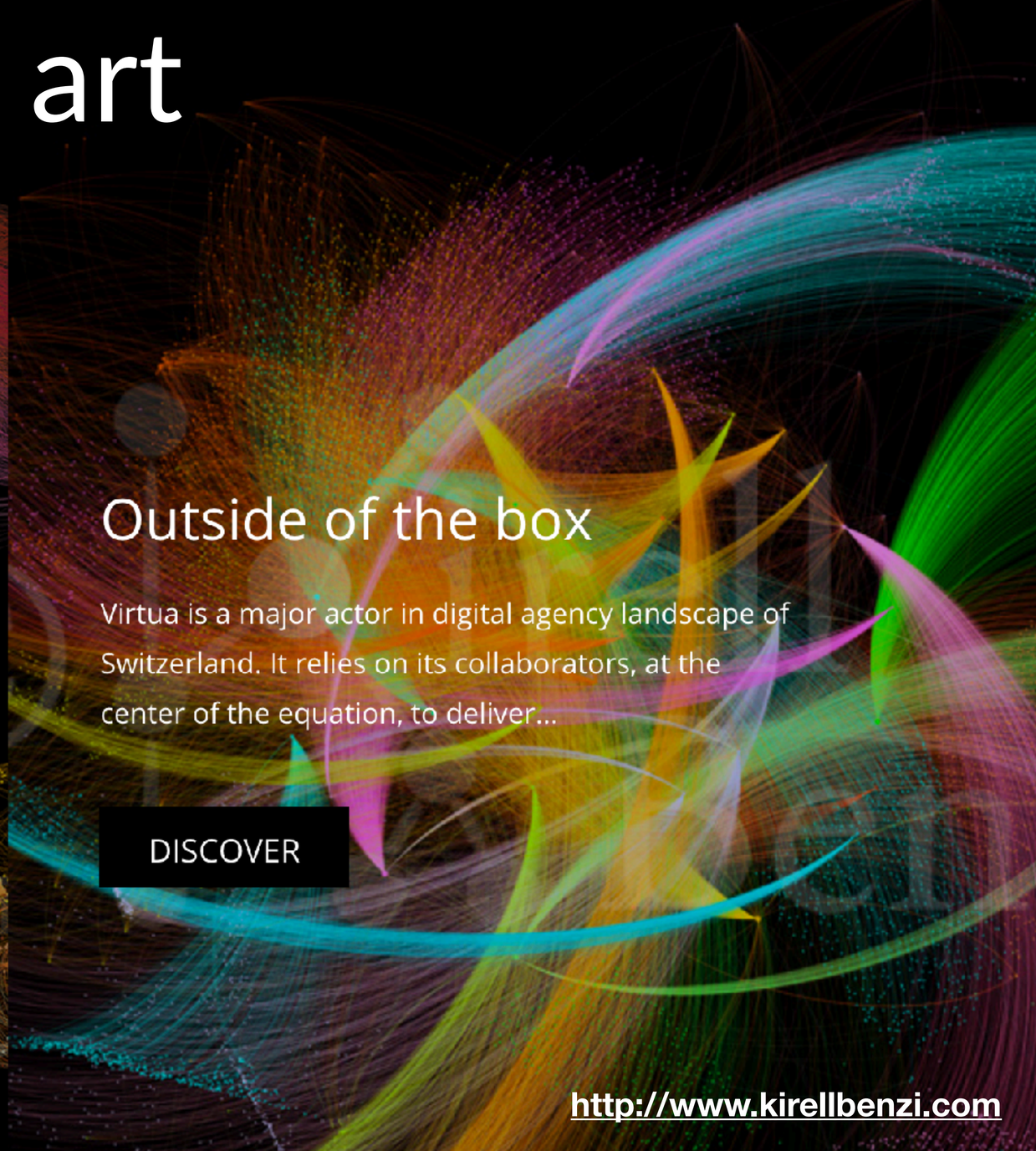
When the Music is Good

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

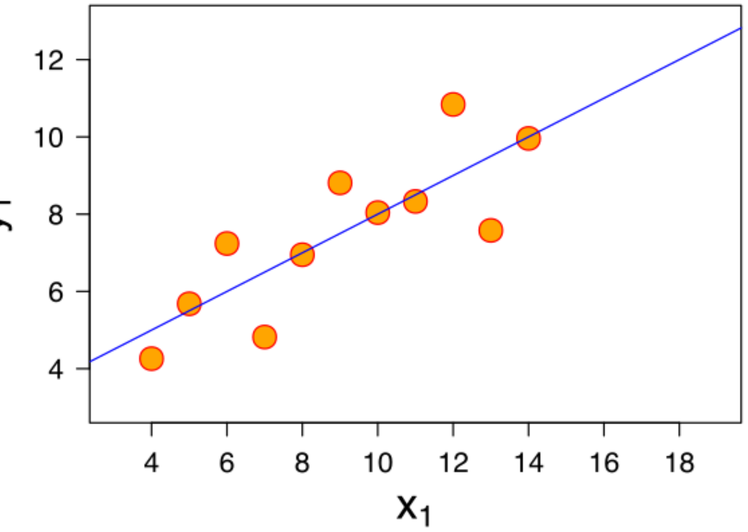
The Dark Side and the Light

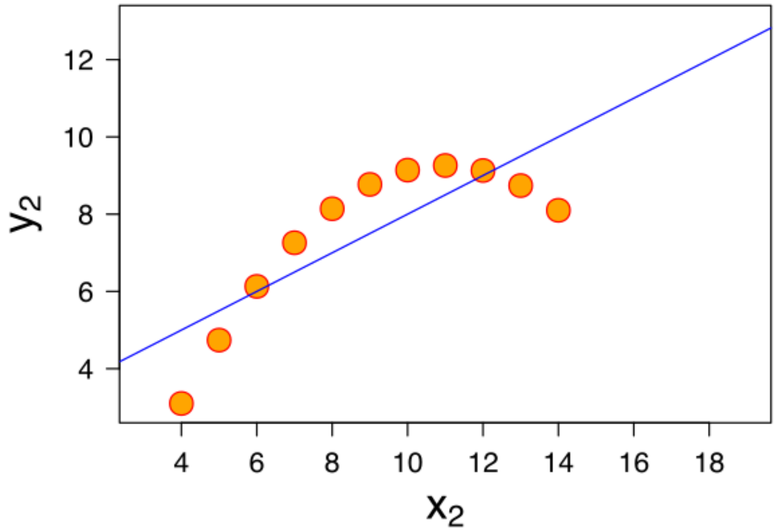
Re-invent

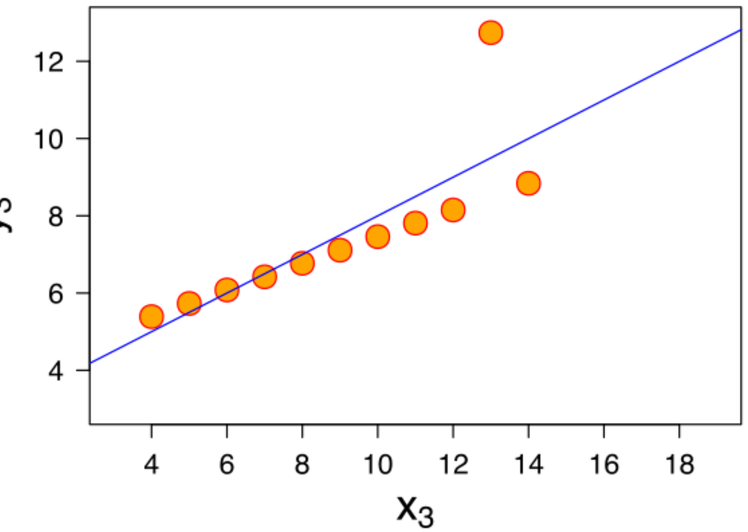
The cycle of inventions is constantly renewing itself: past

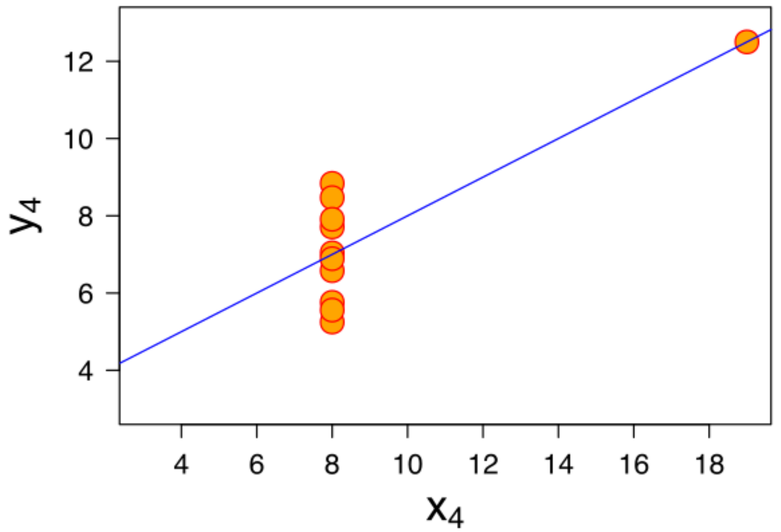
http://www.kirellbenzi.com

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
Coefficient of determination of the linear regression	0.67

[Anscombe's quartet]

Why visualize?

Communication

Transmit information to others

Exploration

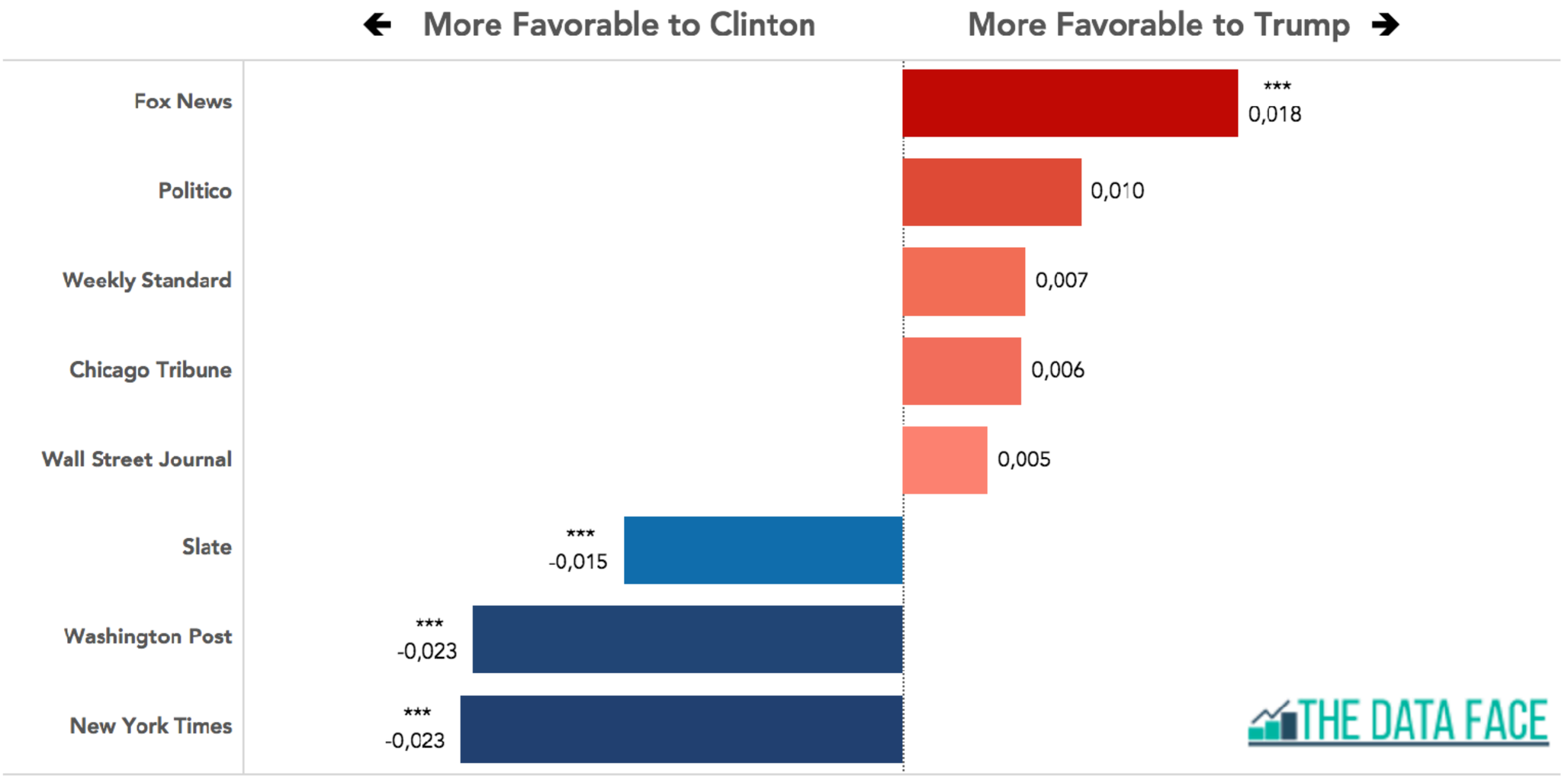
Reveal previously hidden insights from the data

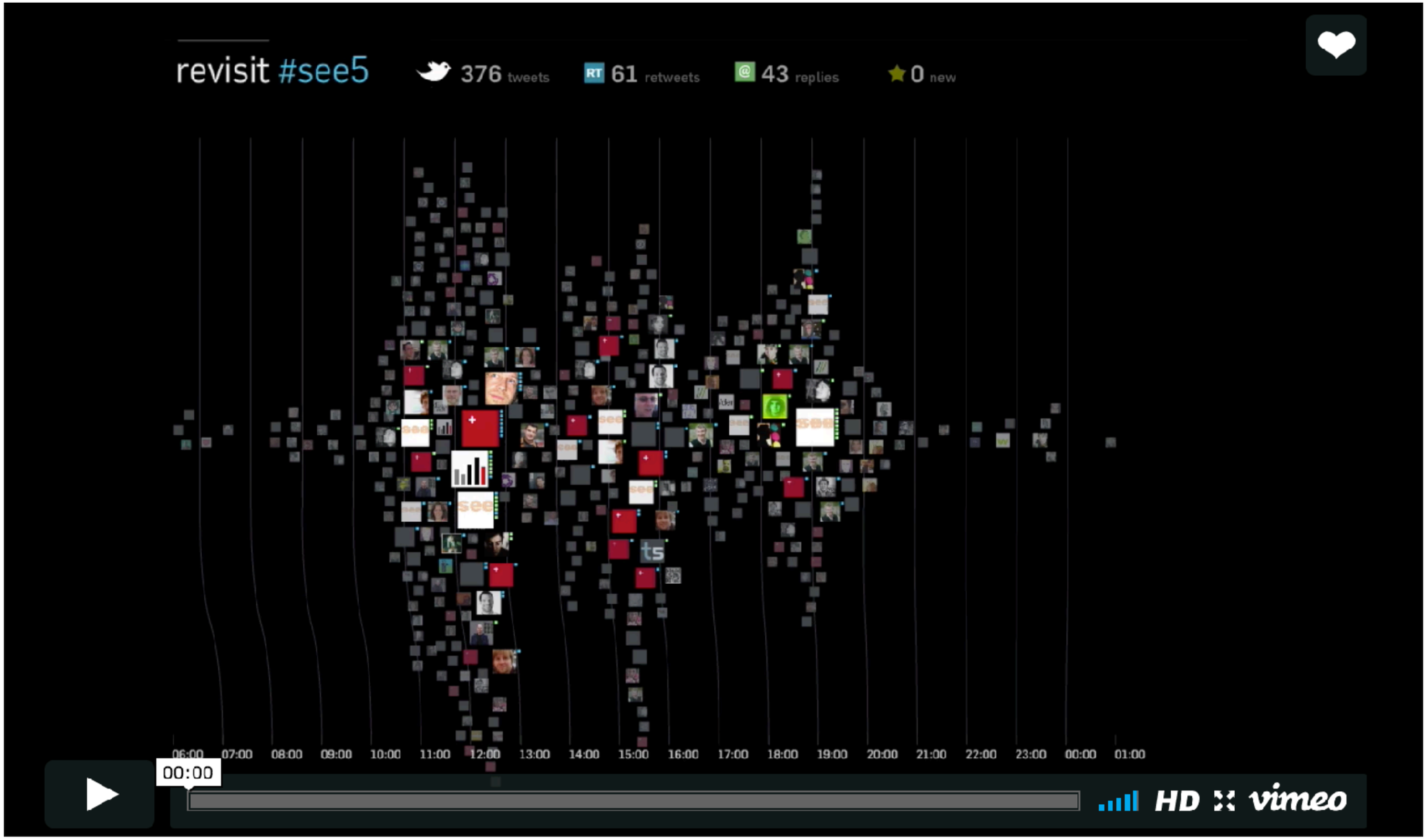
Cognition

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).





http://moritz.stefaner.eu/projects/revisit/

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.

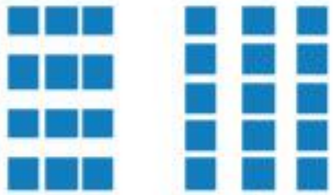
Alexander Lex

Tor Nørretranders

https://atrowpoole.wordpress.com/2014/07/15/gestalt-principles-of-visual-perception/

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.

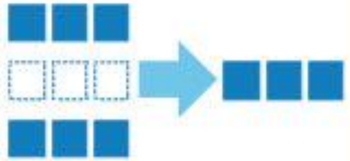
CONTINUATION

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

CLOSURE

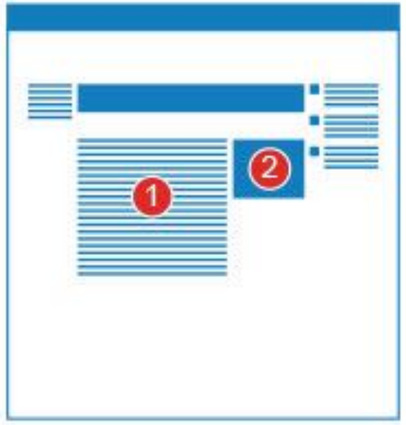
Most people see the white triangle not the 3 circles with segments missing.

COMMON FATE

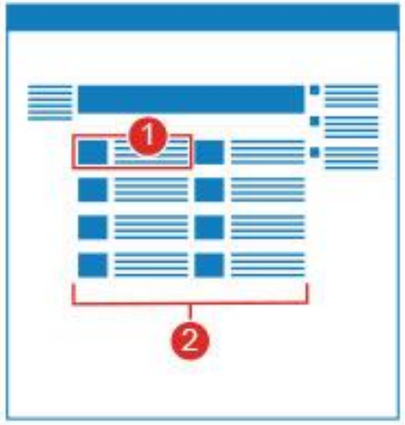


Objects moving in a similar direction are percieved as belonging together.

Examples:

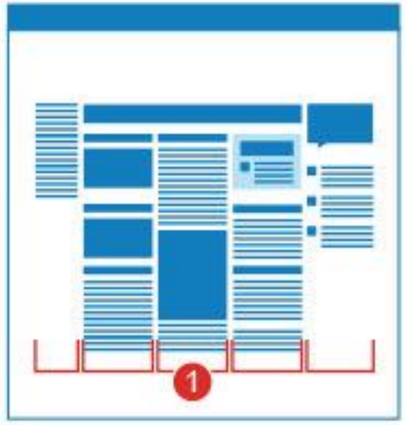


Elements that are closer together are perceived as being more related compared to those spaced further apart. In this



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

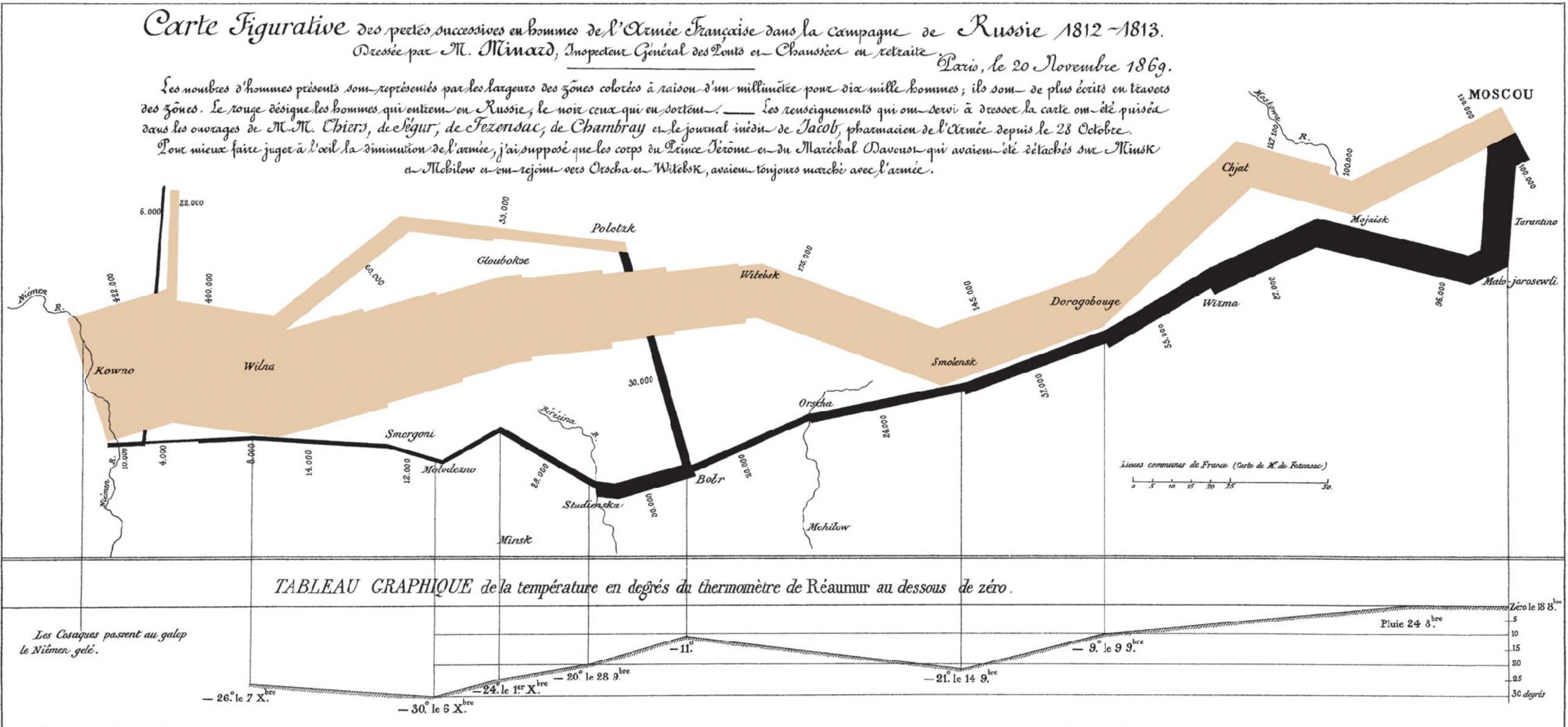


We tend to perceive contours as objects. In this way we perceive lines continue in an established direction (even when they don't) .

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 to be related and moving on an invisible path. Even if objects

Why use computers for viz?



Why use computers for viz?

Faster

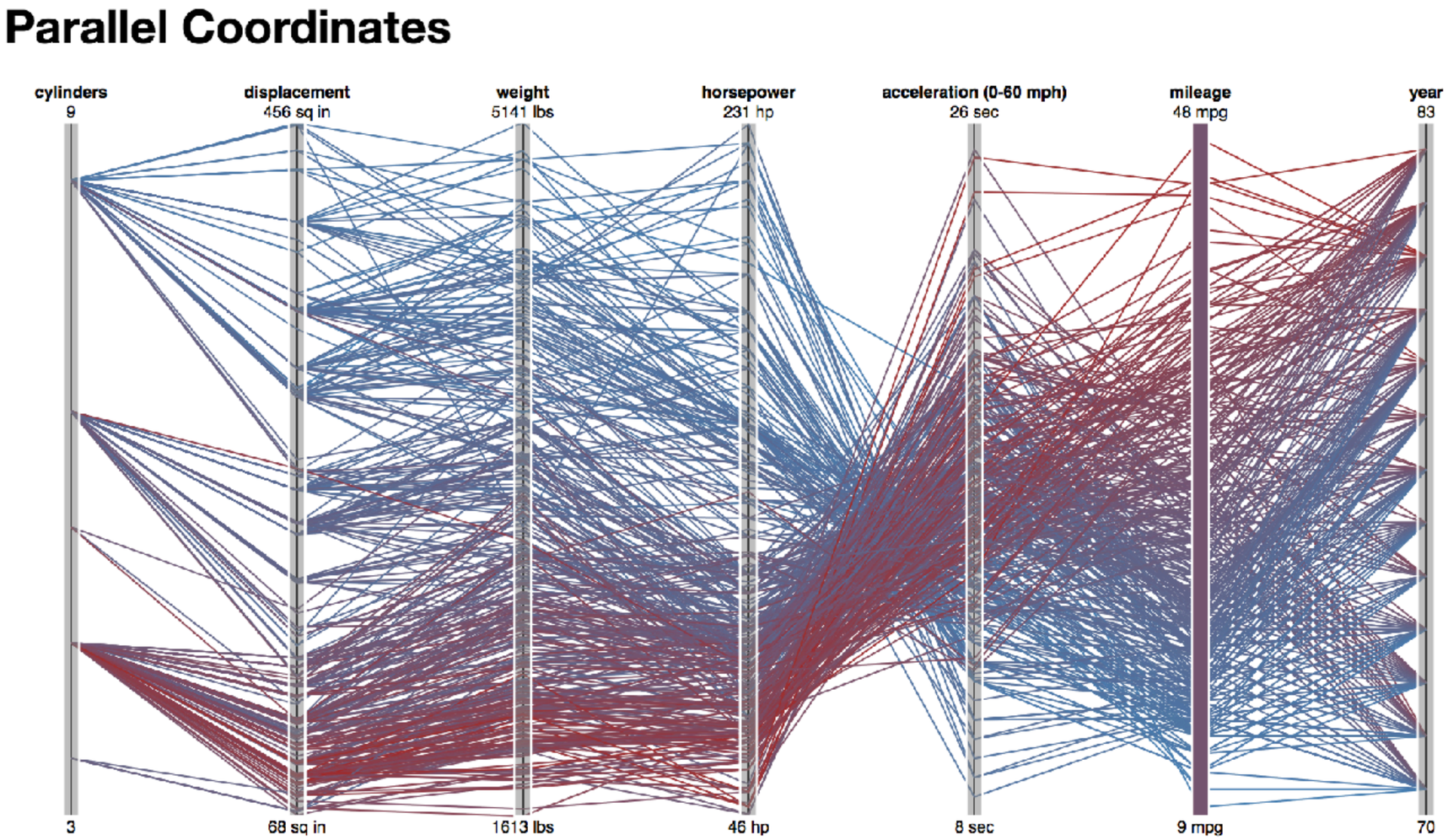
Reproducible

Efficient

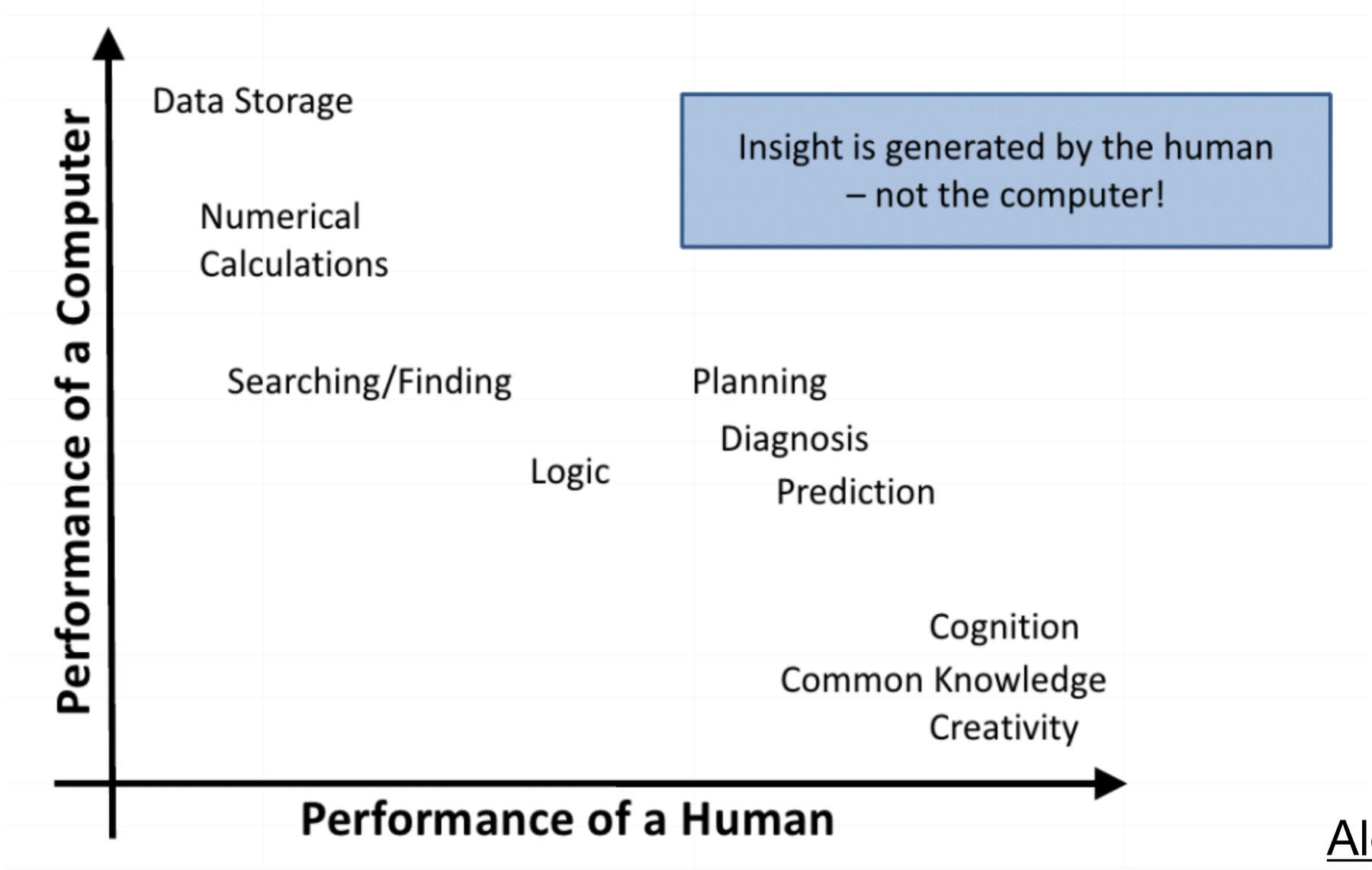
Precise

Interactive

Animated

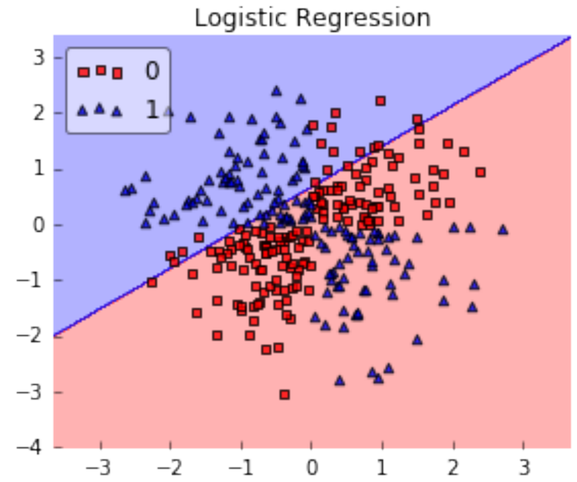


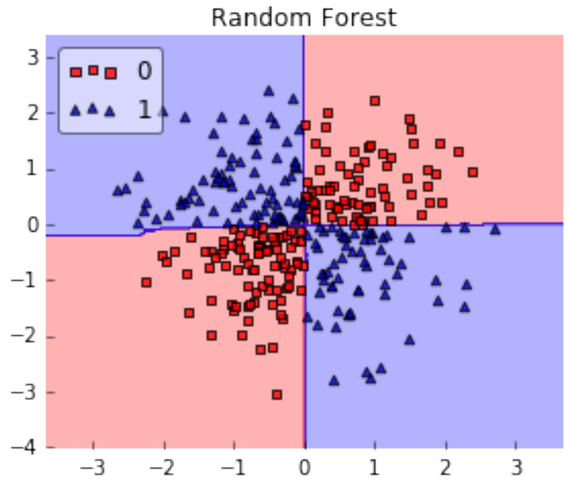
Performance of computers / humans

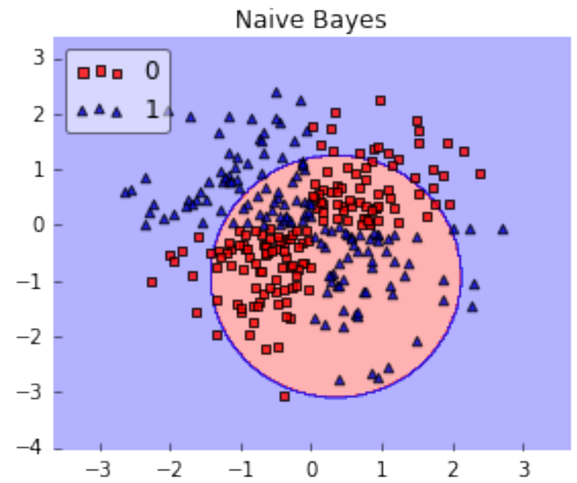


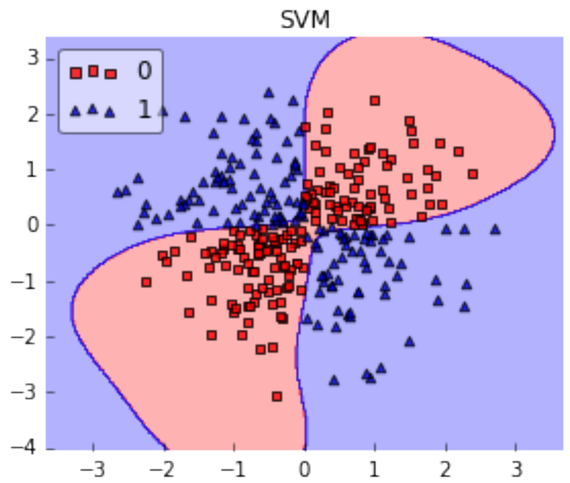
Alexander Lex

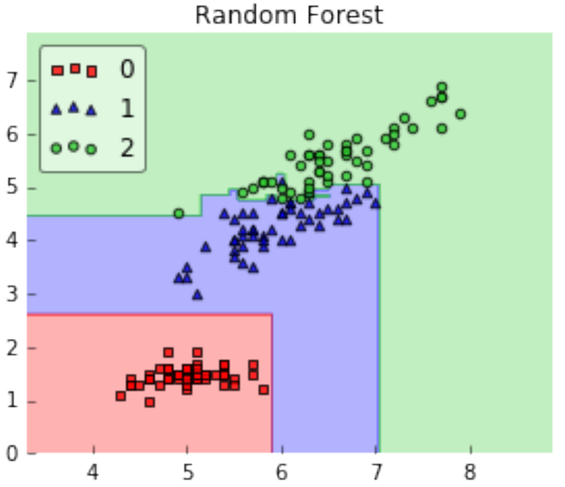
Human vs computer

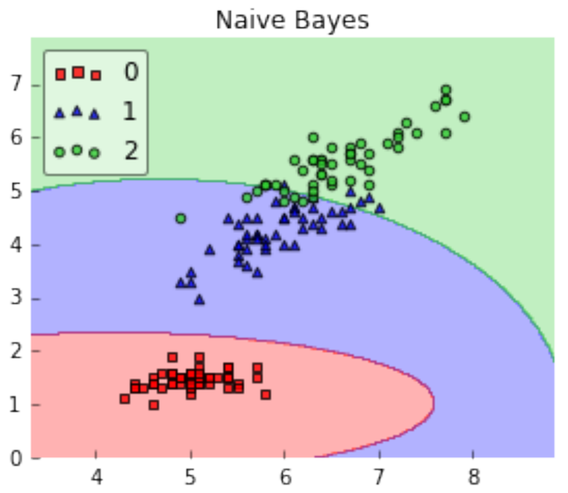


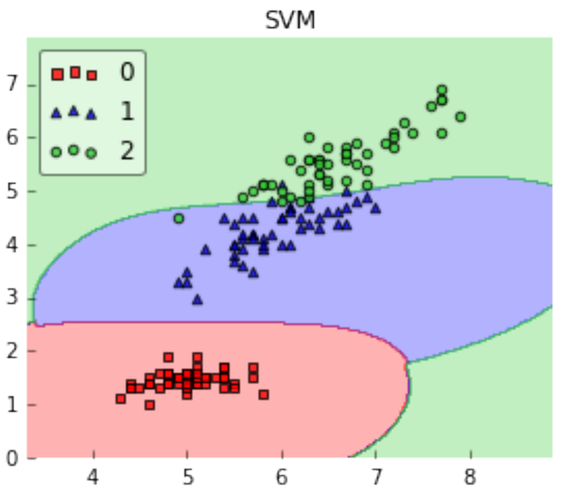




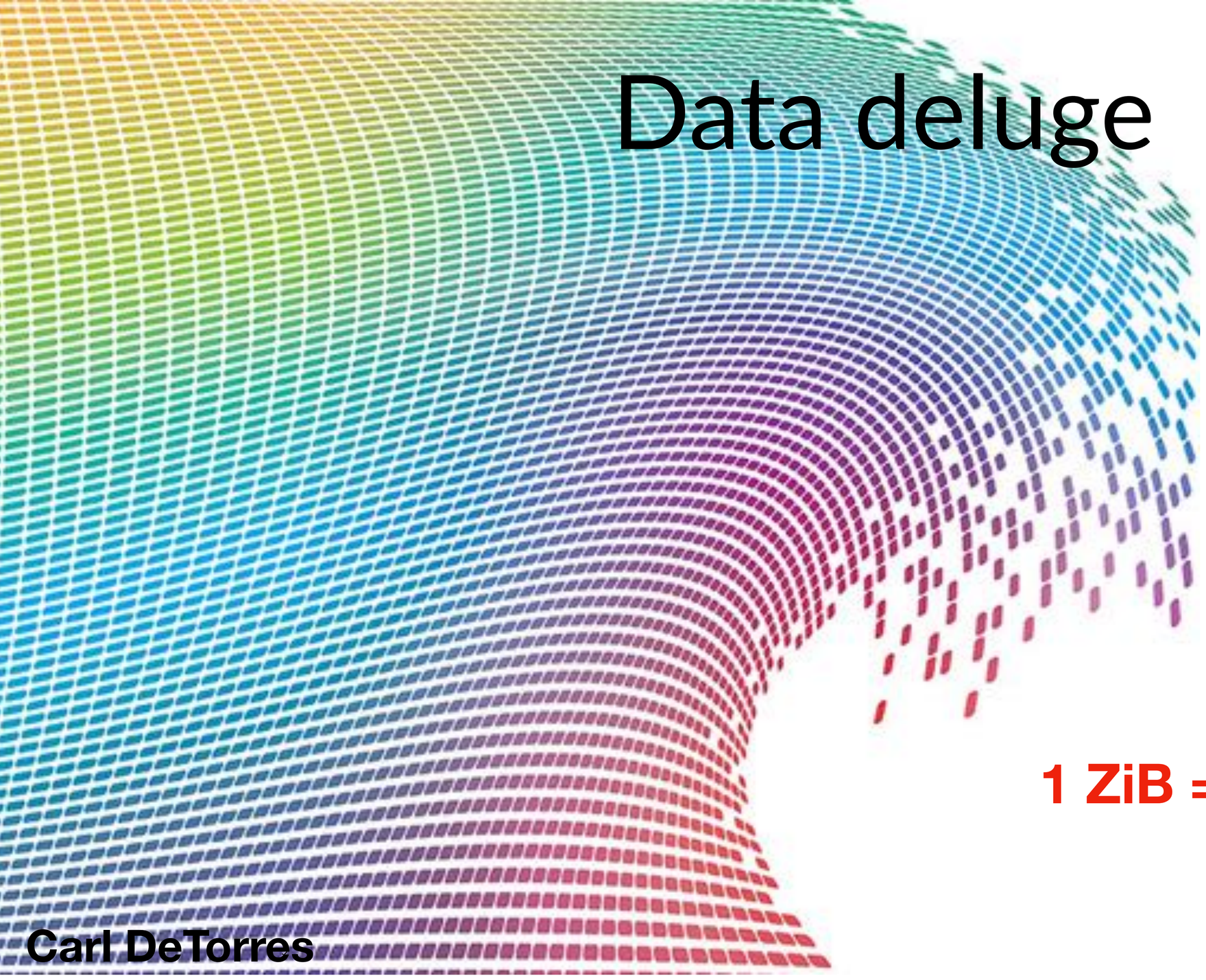








Source: Quora



Data deluge

2007: 281 Exabytes

2010: Zettabytes barrier

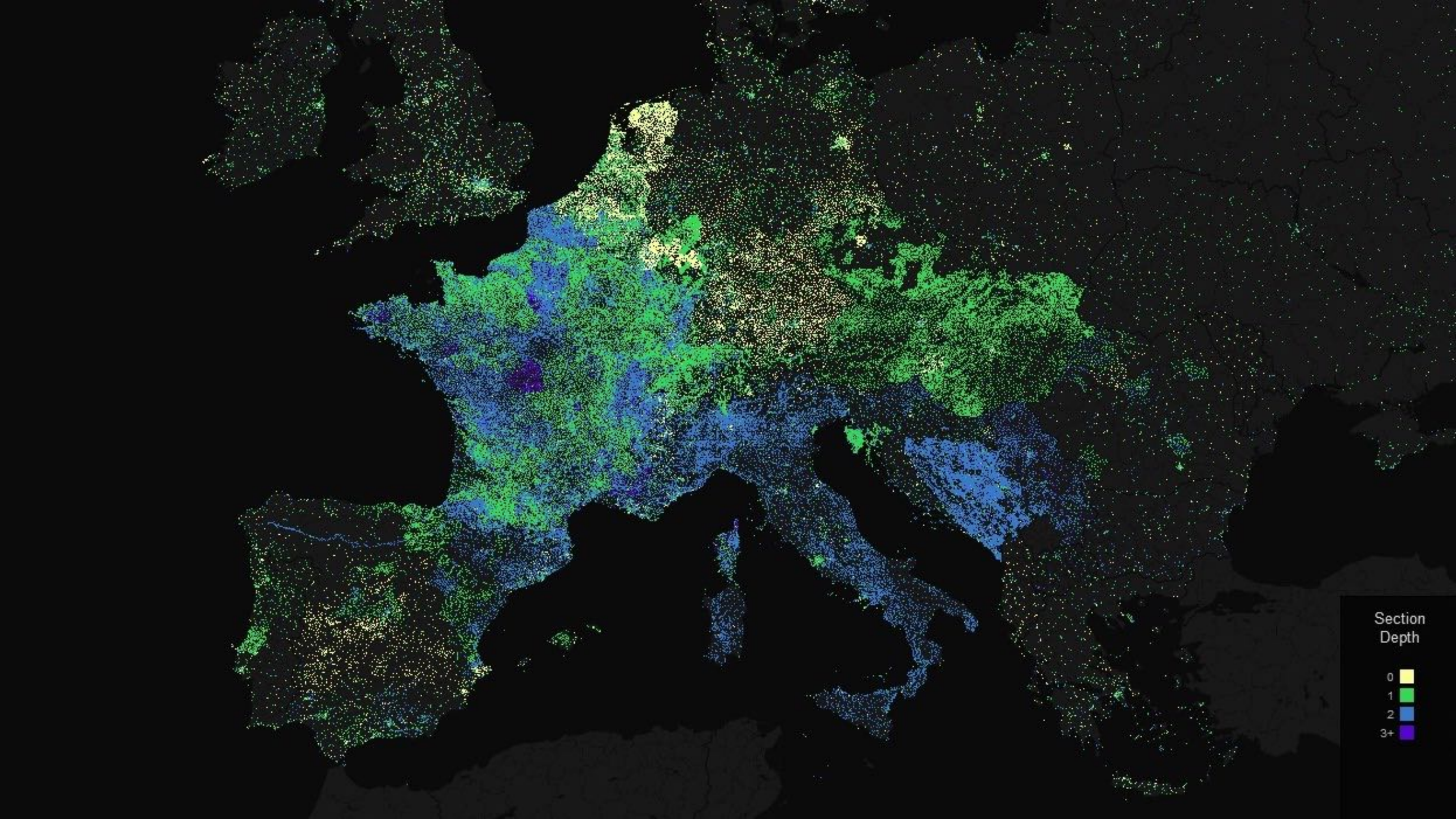
2011: 1.8 Zettabytes

2020: 44 Zettabytes

1 ZiB = 10^21 bytes

Carl DeTorres

Source: EMC


Section
Depth
0
1
2
3+


facebook
December 2010

Data science questions

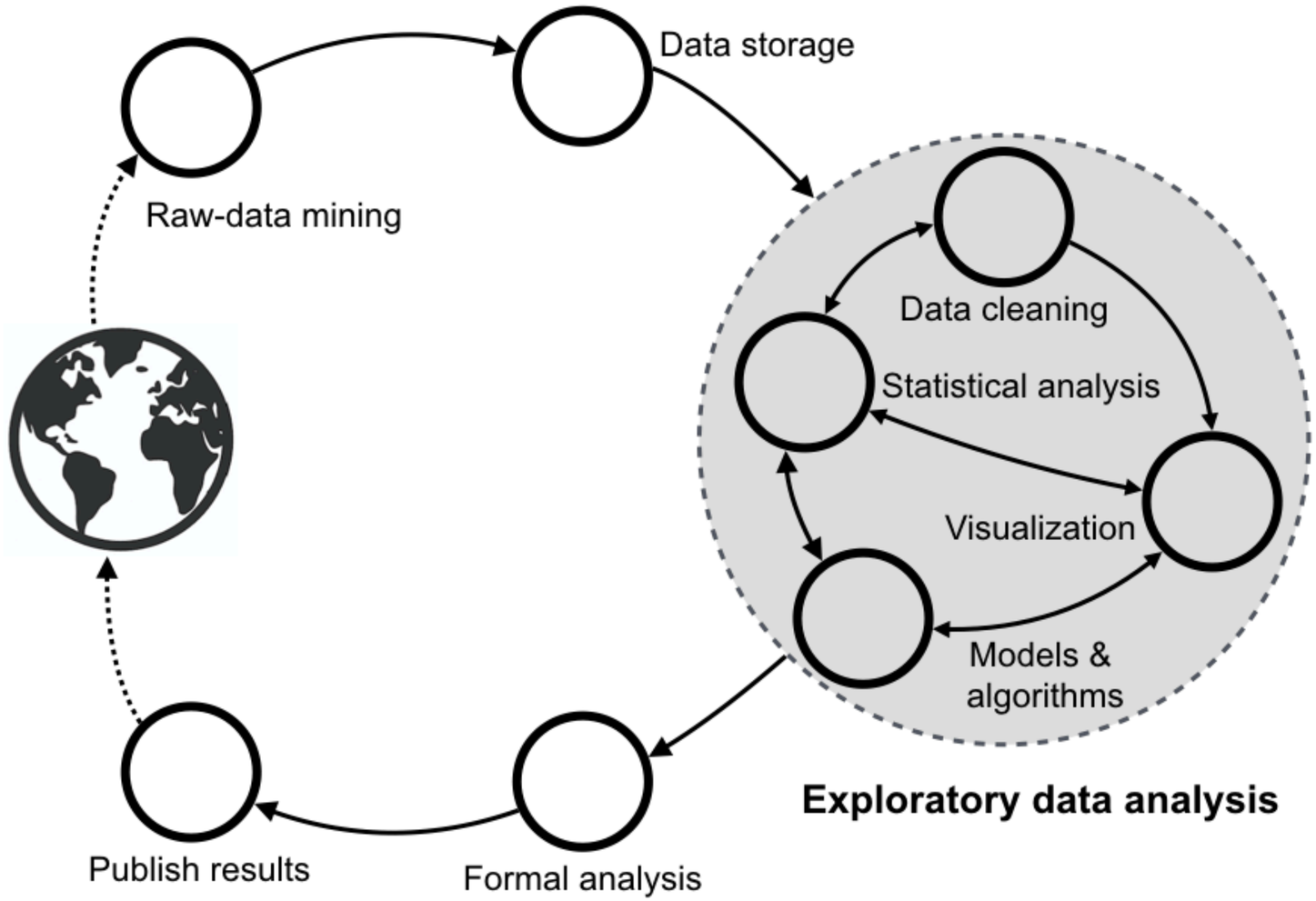
How do we effectively access data?

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

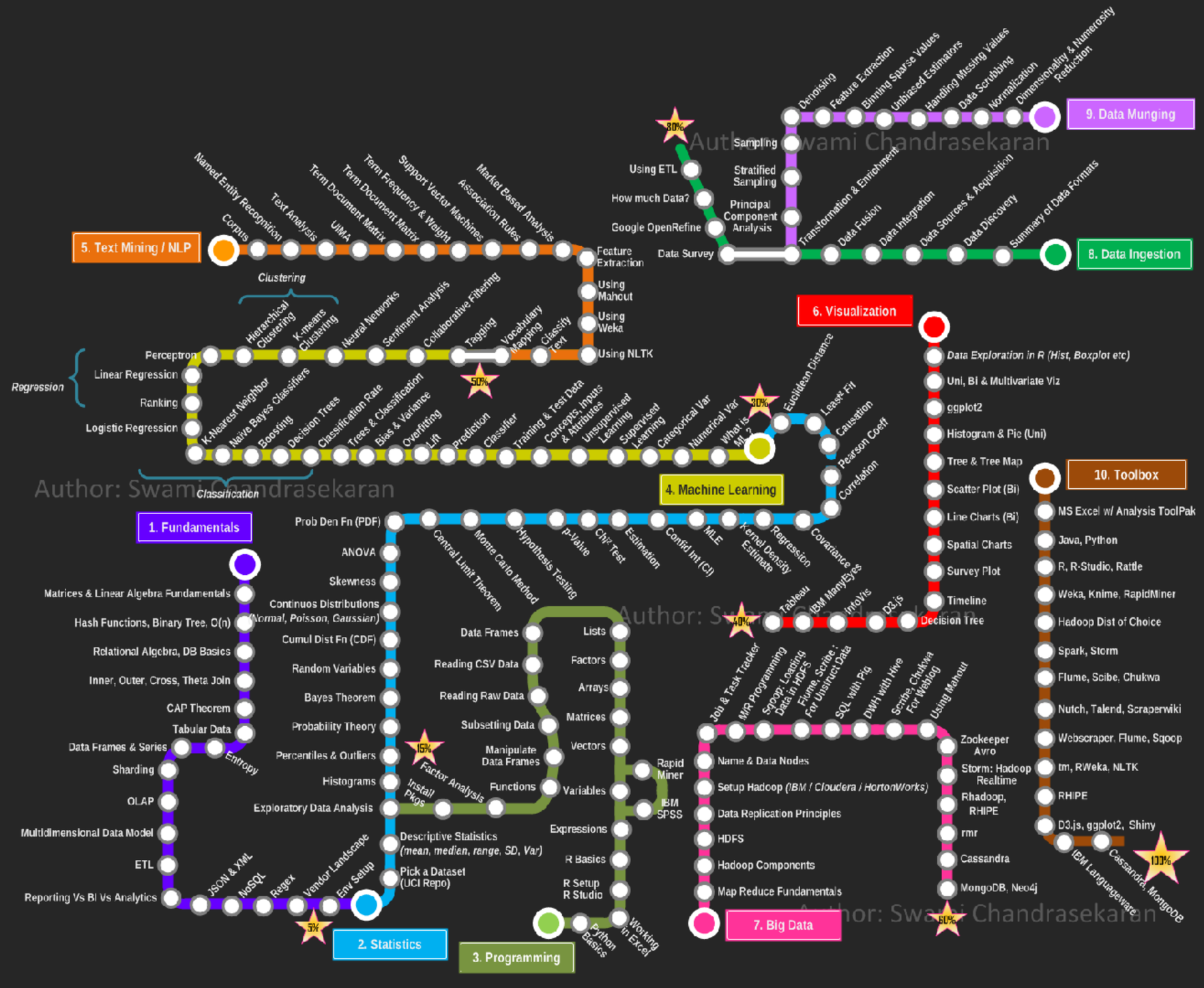
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

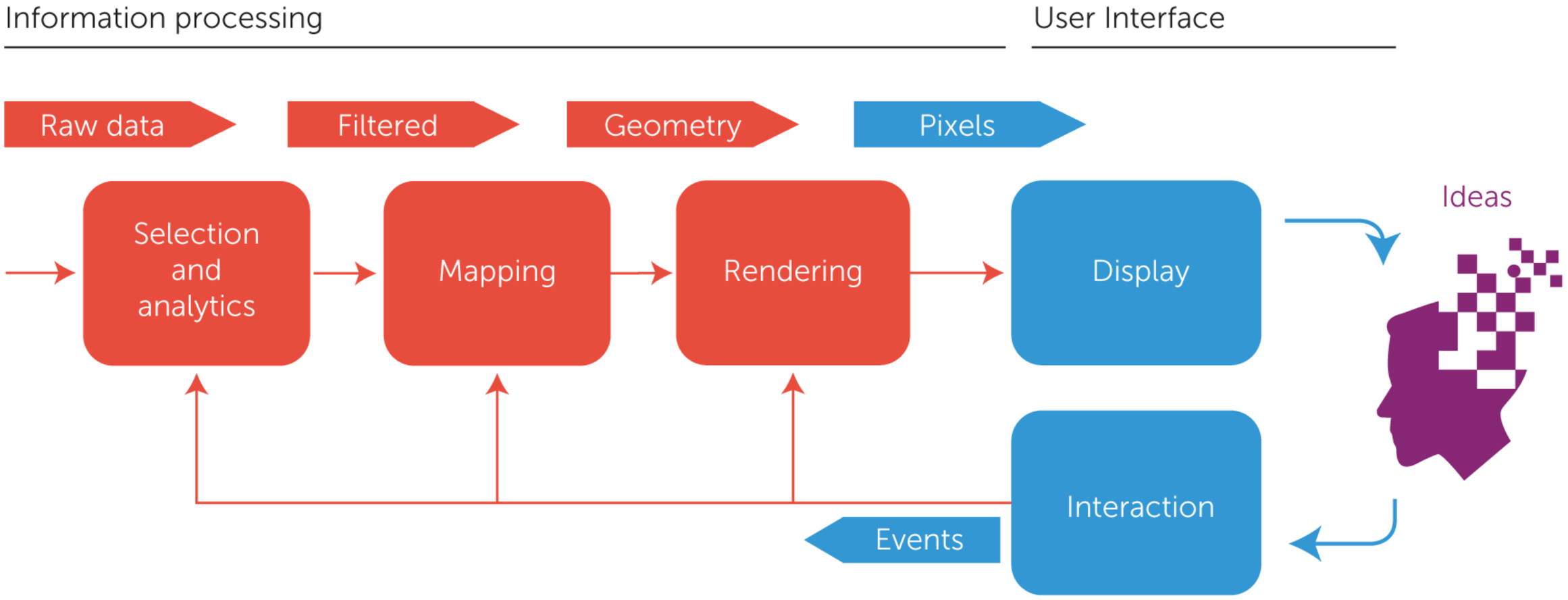
- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of ary of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY


Author: Swami Chandrasekaran
1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions
(Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics
(mean, median, range, SD, Var)
Pick a Dataset
(UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup
R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Regression
Perceptron
Linear Regression
Ranking
Logistic Regression
Classification
K-Nearest Neighbor
Naive Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
30%
Clustering
Hierarchical Clustering
K-means Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
Euclidean Distance
Least Fn
Causation
Pearson Coeff
Correlation
5. Text Mining / NLP
Named Entity Recognition
Text Analysis
Corpus
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper
Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Data Survey
Using ETL
How much Data?
Google OpenRefine
Sampling
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
30%
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%

Visualization Pipeline

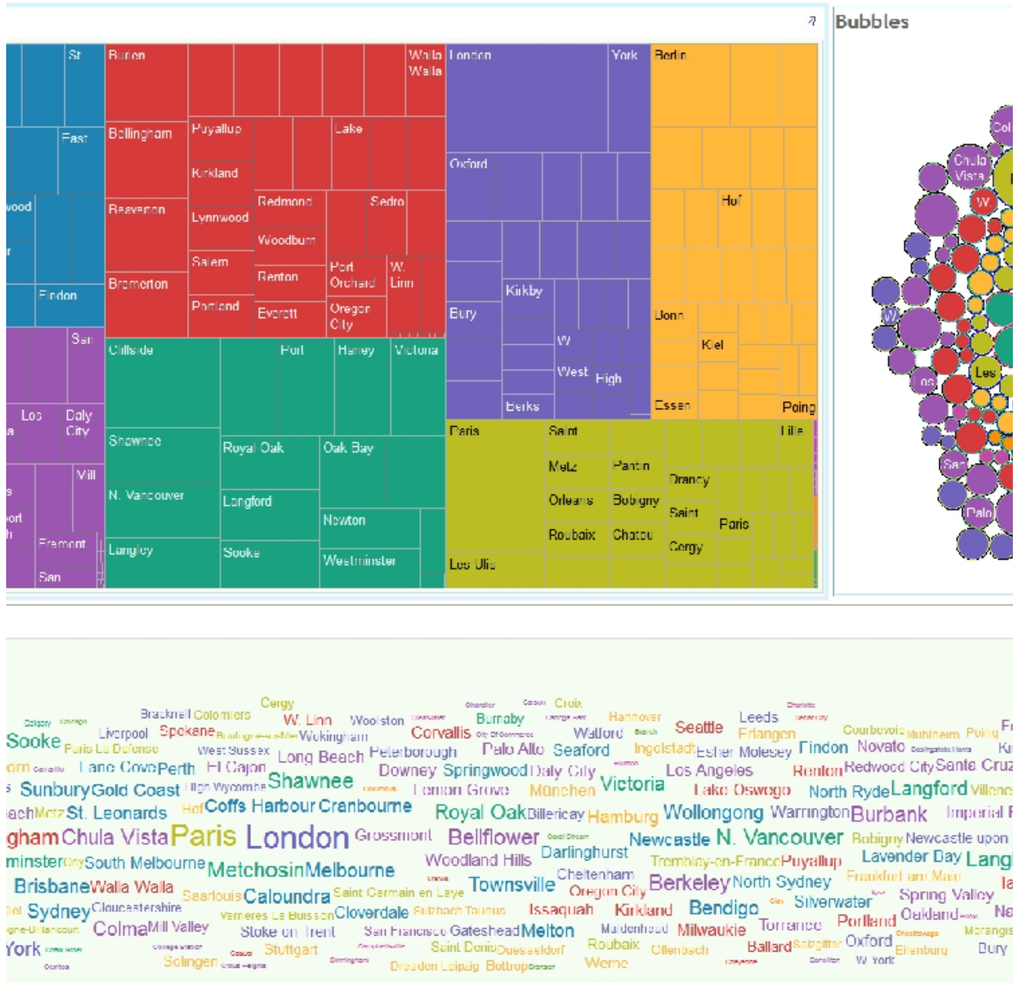


[Holliman 2015]

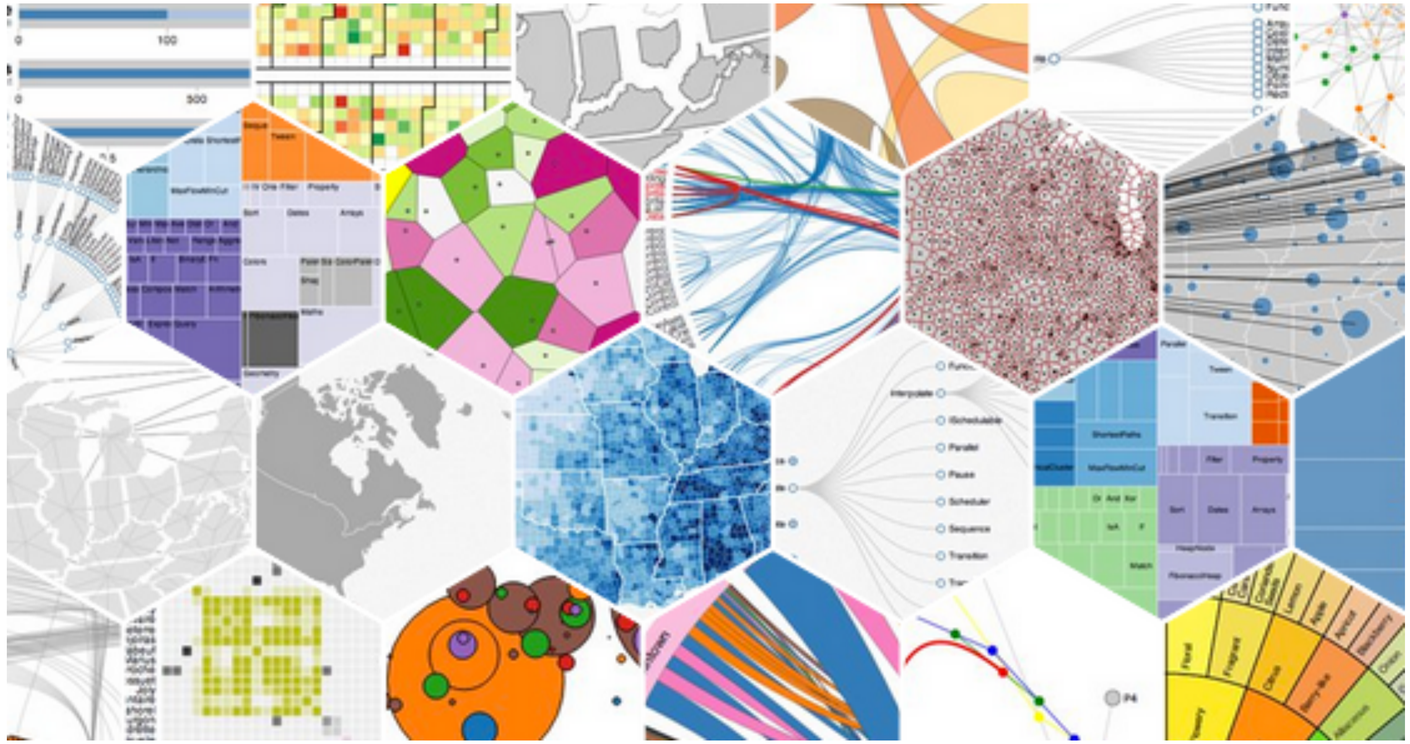
How to create a data-viz?

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

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



Tech stack



ECMAScript 6

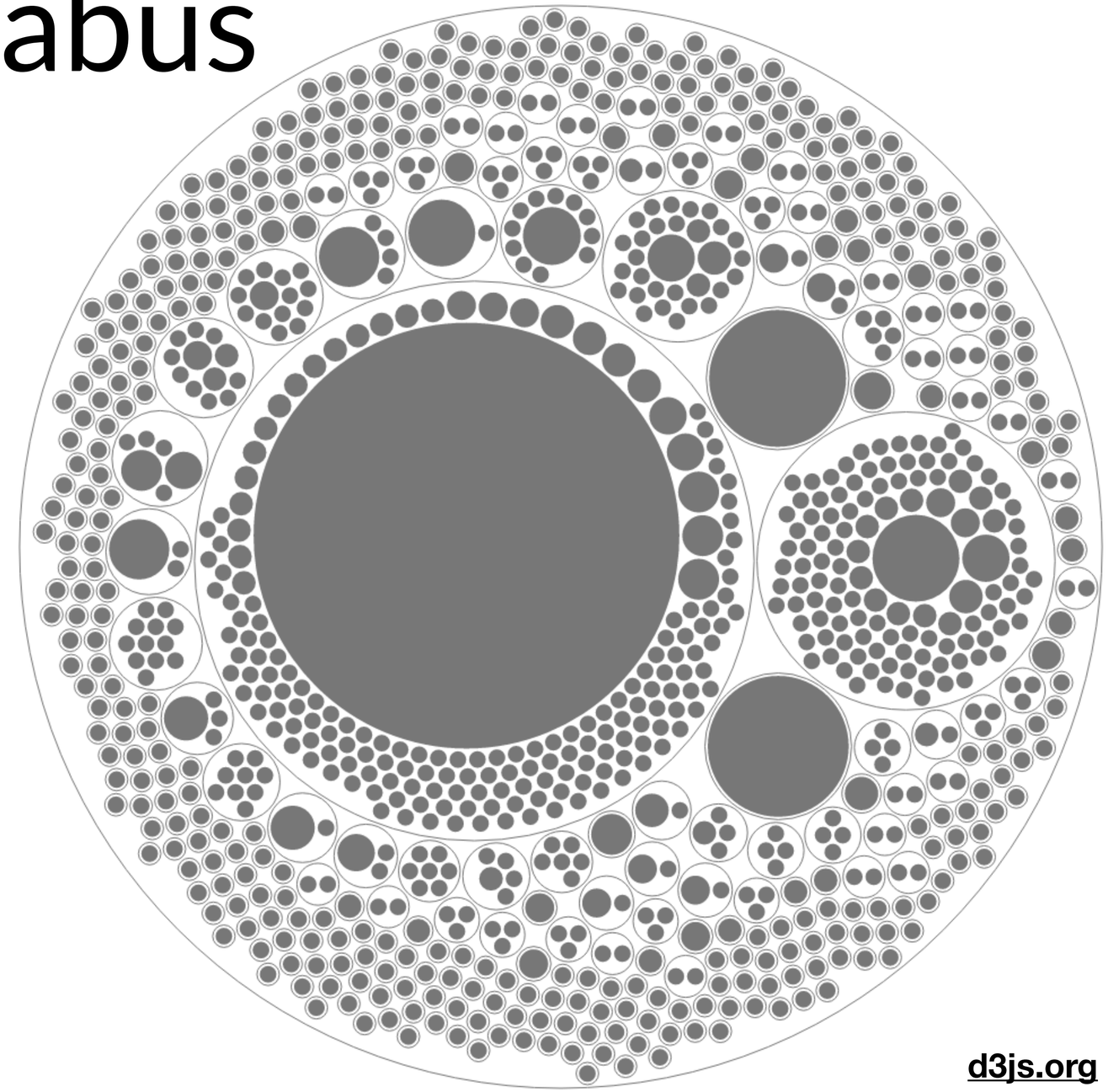
JS

Syllabus

1. The Web “stack”
    1. HTML, CSS, DOM
    2. SVG, Canvas, WebGL
    3. ECMA 2015+ (javascript)
    4. D3.js

Syllabus

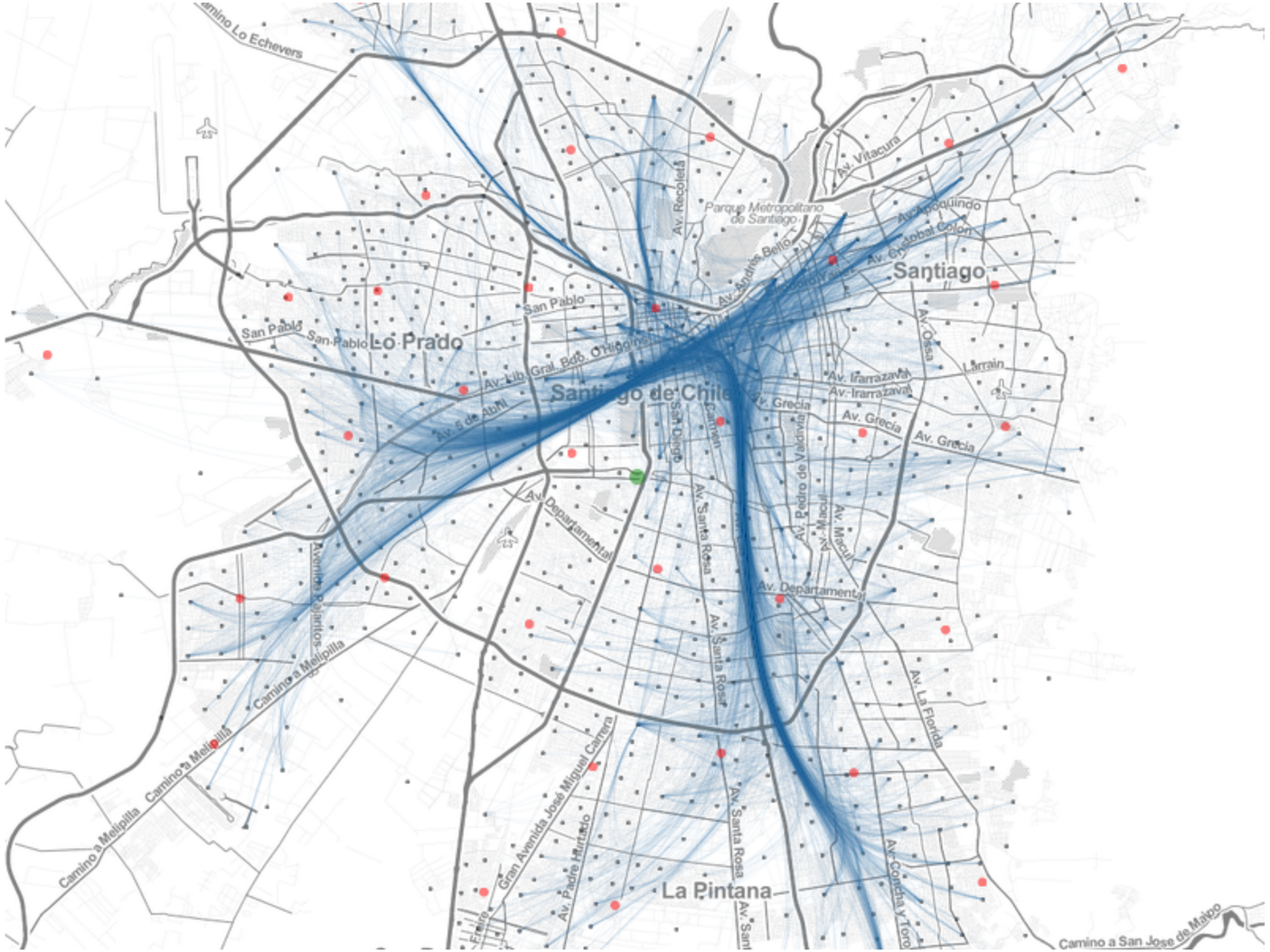
2. Visualization fundamentals
    1. Data
    2. Interaction, filtering, aggregation
    3. Perception, cognition, color
    4. Marks and channels
    5. Designing visualizations



d3js.org

Syllabus

3. Techniques and algorithms
    1. Maps
    2. Text
    3. Trees, graphs
    4. Tabular data
    5. Sound
    6. Volumes



http://carnby.github.io/

Structure and goals
COM-480

Learning objectives

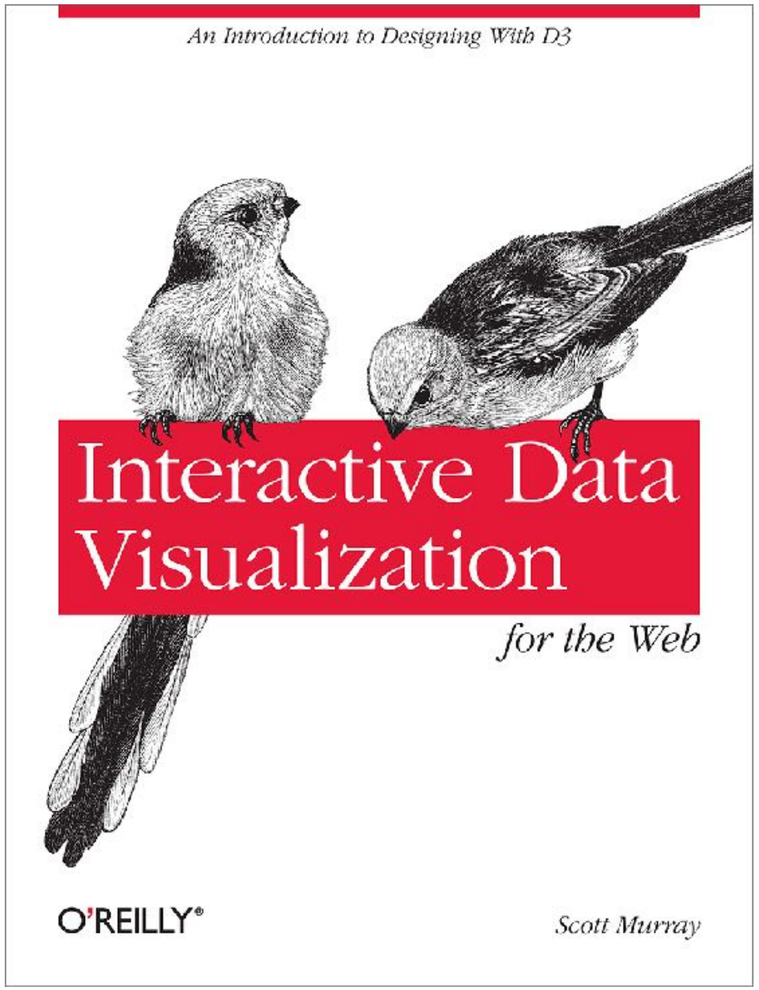
Efficiently visualize data

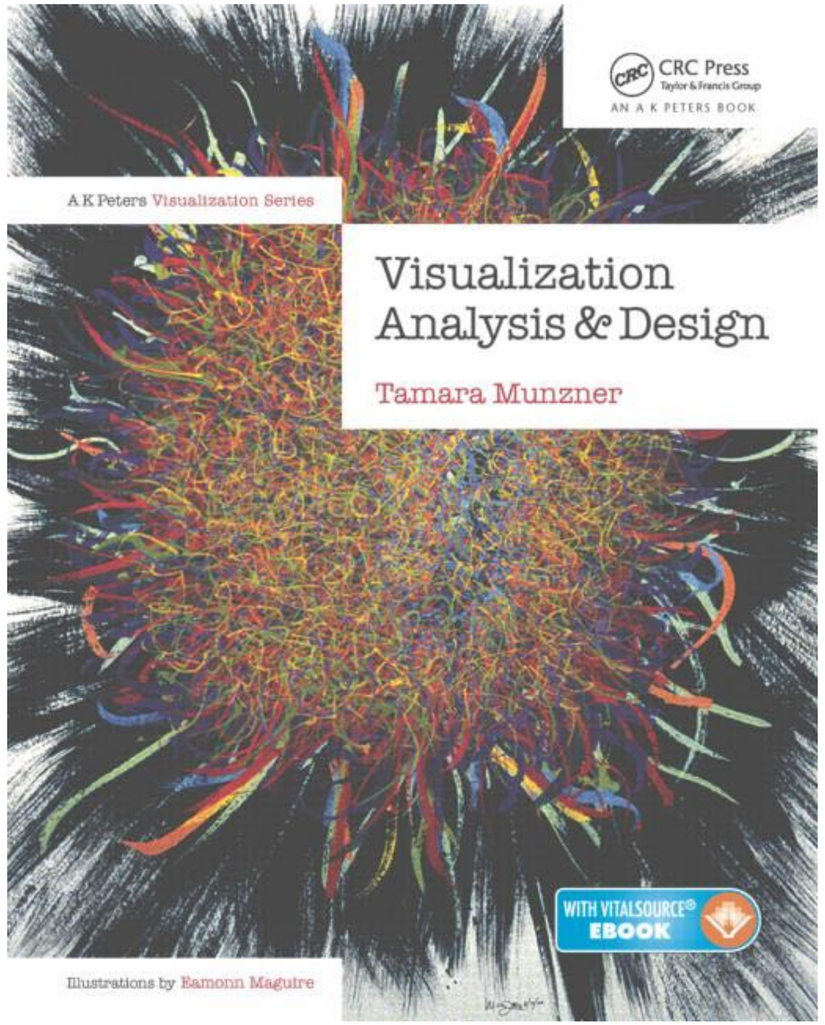
Design visual data analysis solutions

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

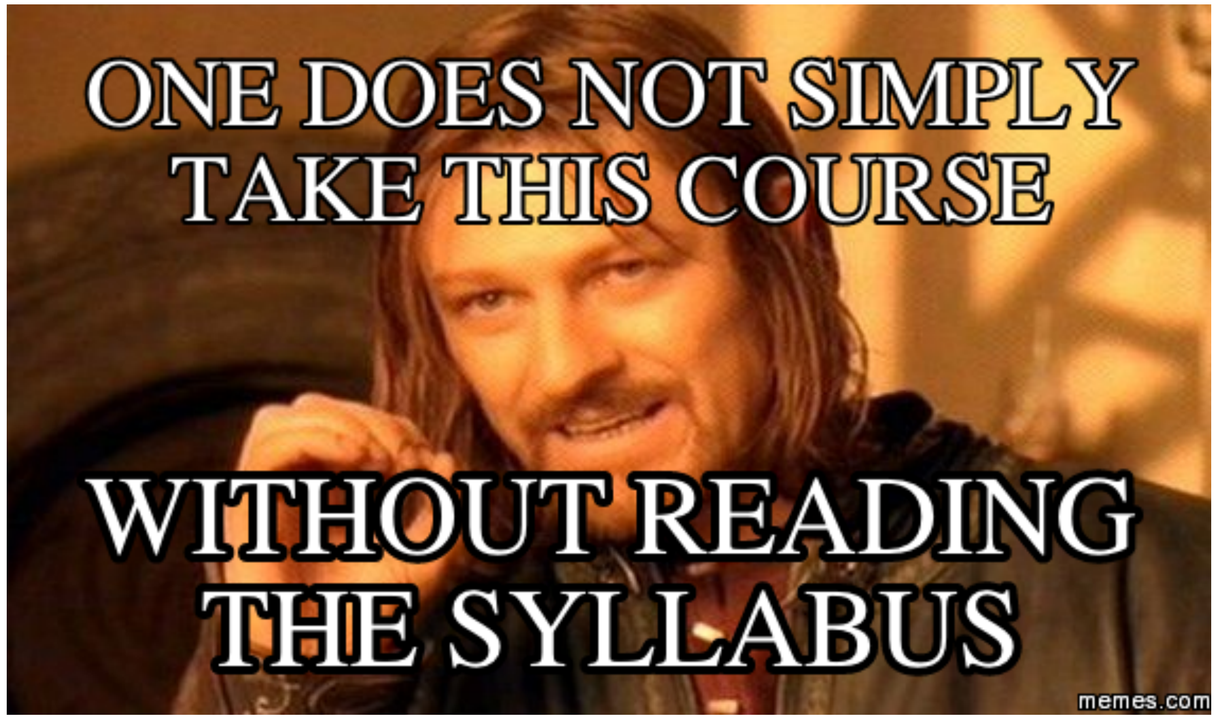
Prerequisites

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

Group of 3 students for the end of March

Visualization (35%)

Technical Implementation (15%)

Website, presentation, screencast (25%)

Process book (25%)

Milestone 1 (7th April 5pm) Data Set, Problematic, EDA, Related work

Milestone 2 (5th May 5pm) project goal, functional project prototype

Milestone 3 (2nd June 5pm) Web site, GitHub, Screencast, Process book

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?