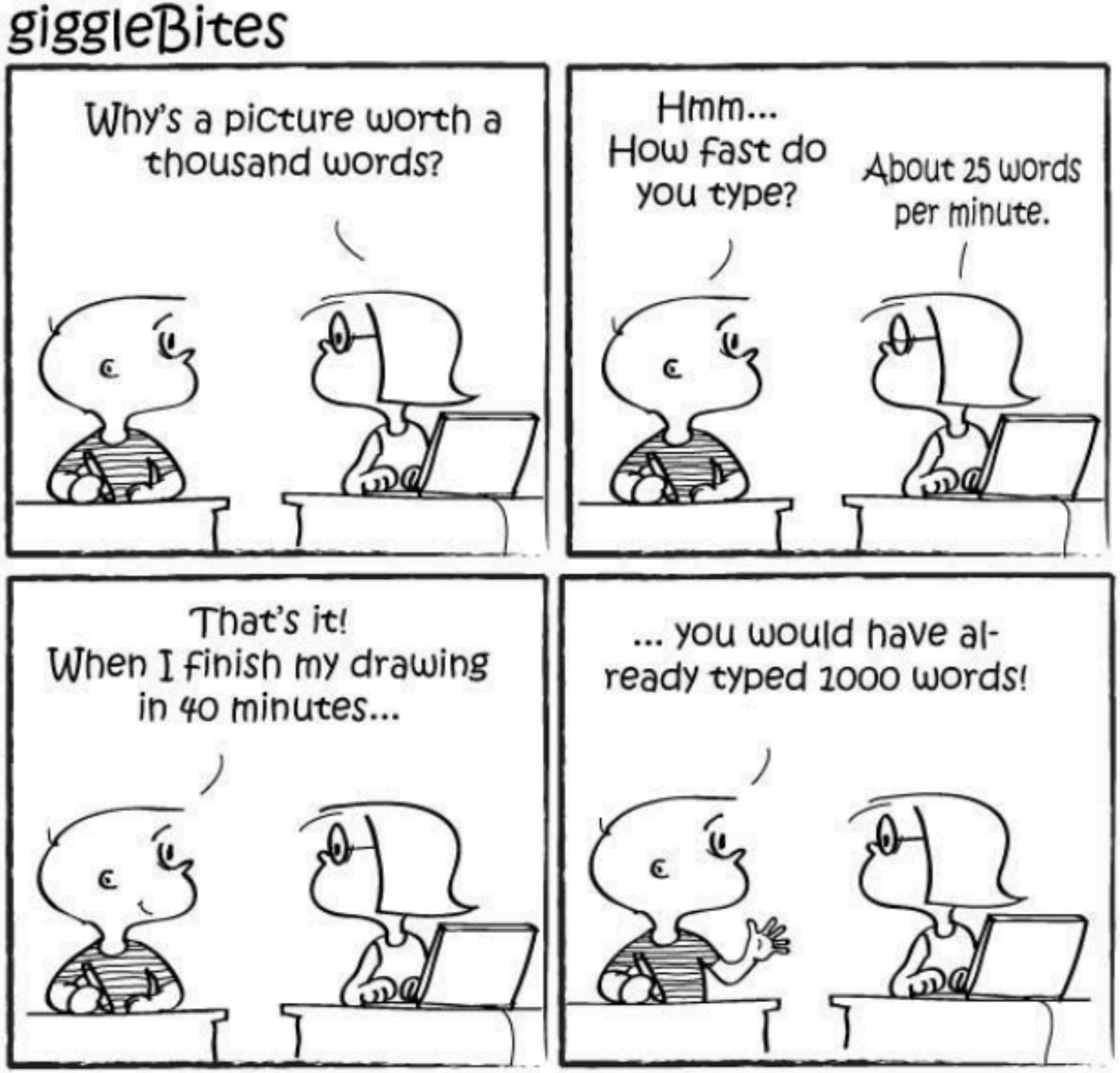
# Data Mining
## (Mining Knowledge from Data)

Data Visualization

Magda Friedjungová

# Exercises organization

- Attending the practice is compulsory.
- 4 blocks of exercises, dates on EDUX.
- 6 homework assignments, 2 of them assigned at every practice (excluding the last one), for 8 points each.
- 2 points for compulsory activity at last practice.
- The required minimum is 25 points, the maximum is 50.


giggleBites
Why's a picture worth a thousand words?
Hmm... How fast do you type?
About 25 words per minute.
That's it! When I finish my drawing in 40 minutes...
... you would have already typed 1000 words!

# Importance of Visualization

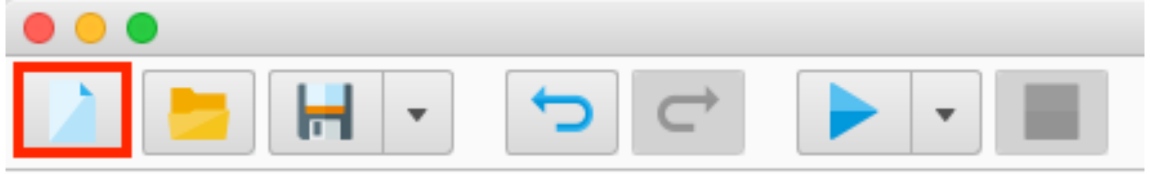
- Helps understand data
- Extremely powerful tool
- Bypass language centers, go directly to the visual cortex
- Ability to recognize patterns
- Animations, live data processing

“As the Chinese say,
1001 words is worth more
than a picture.”

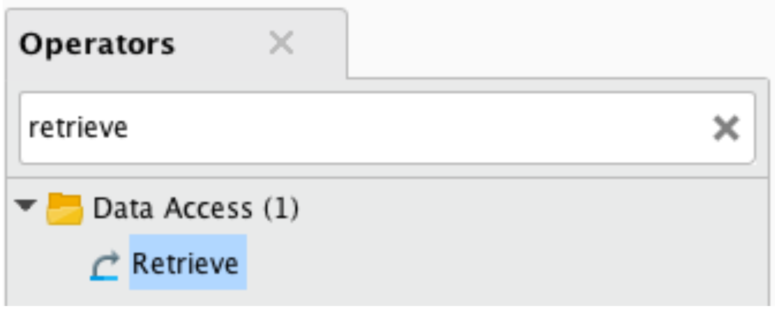
Wikipedia

# RapidMiner Studio

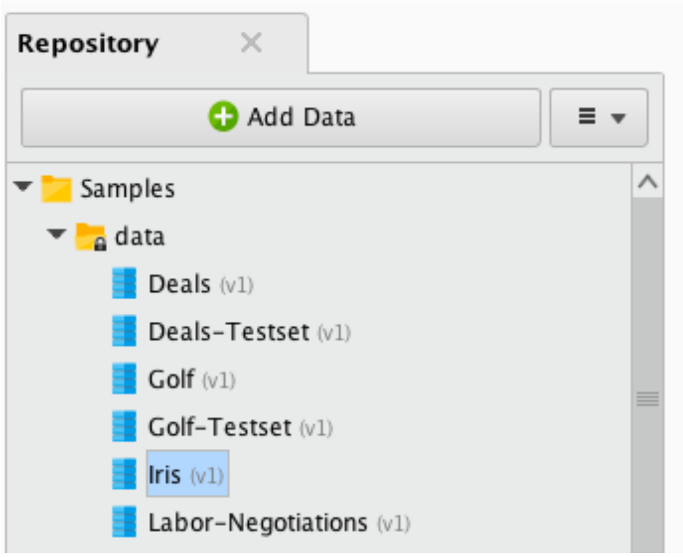
- Start RapidMiner Studio.
- Create new blank process.

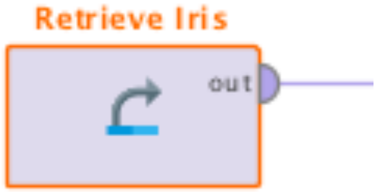


- Find the block “Retrieve” and drag to your process (at MacOS you can skip this step).

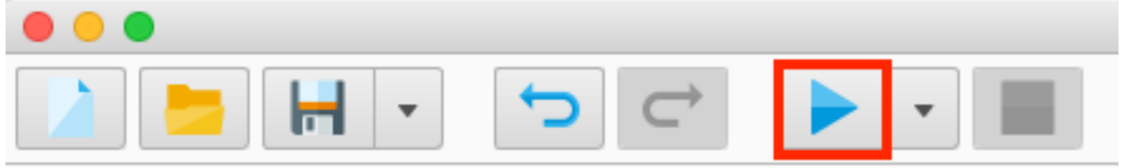


- Find the Iris dataset in Samples and set it to the Retrieve block.

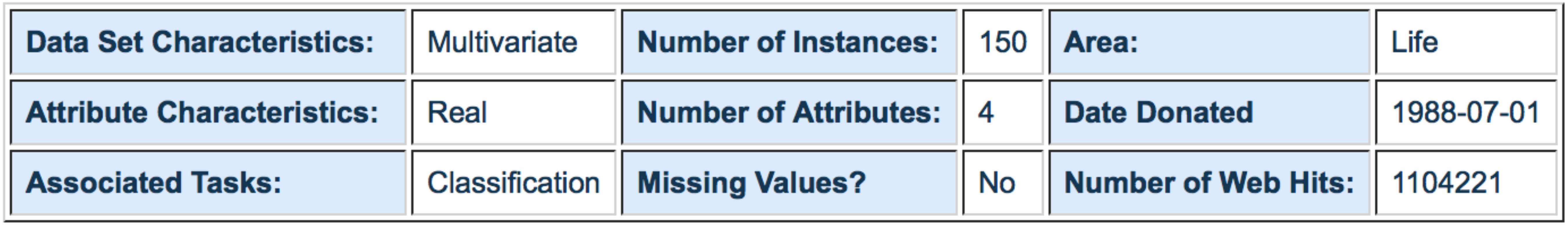




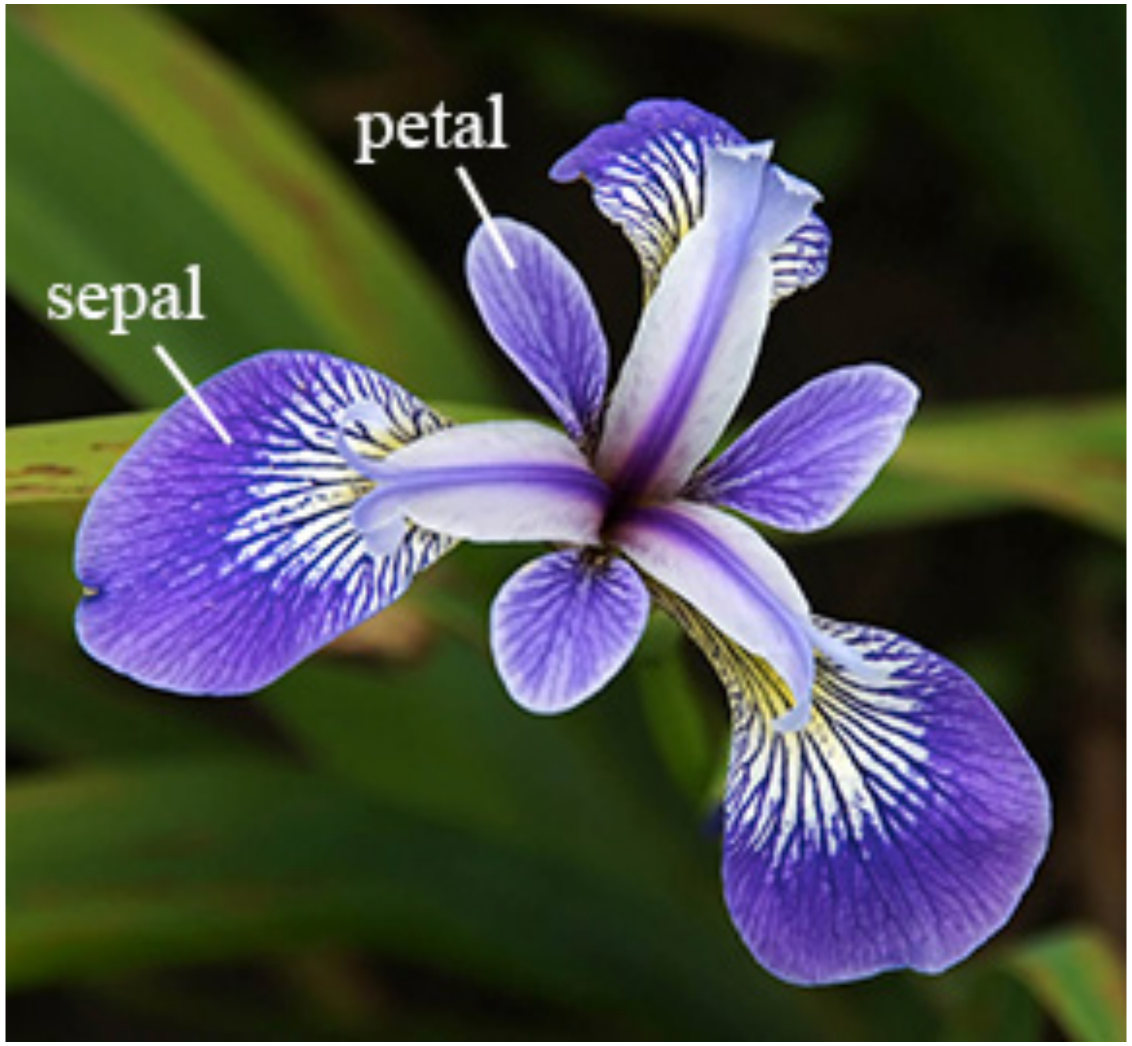
- Run the model.



# Iris Dataset



| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 150 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 4 | **Date Donated** | 1988-07-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 1104221 |

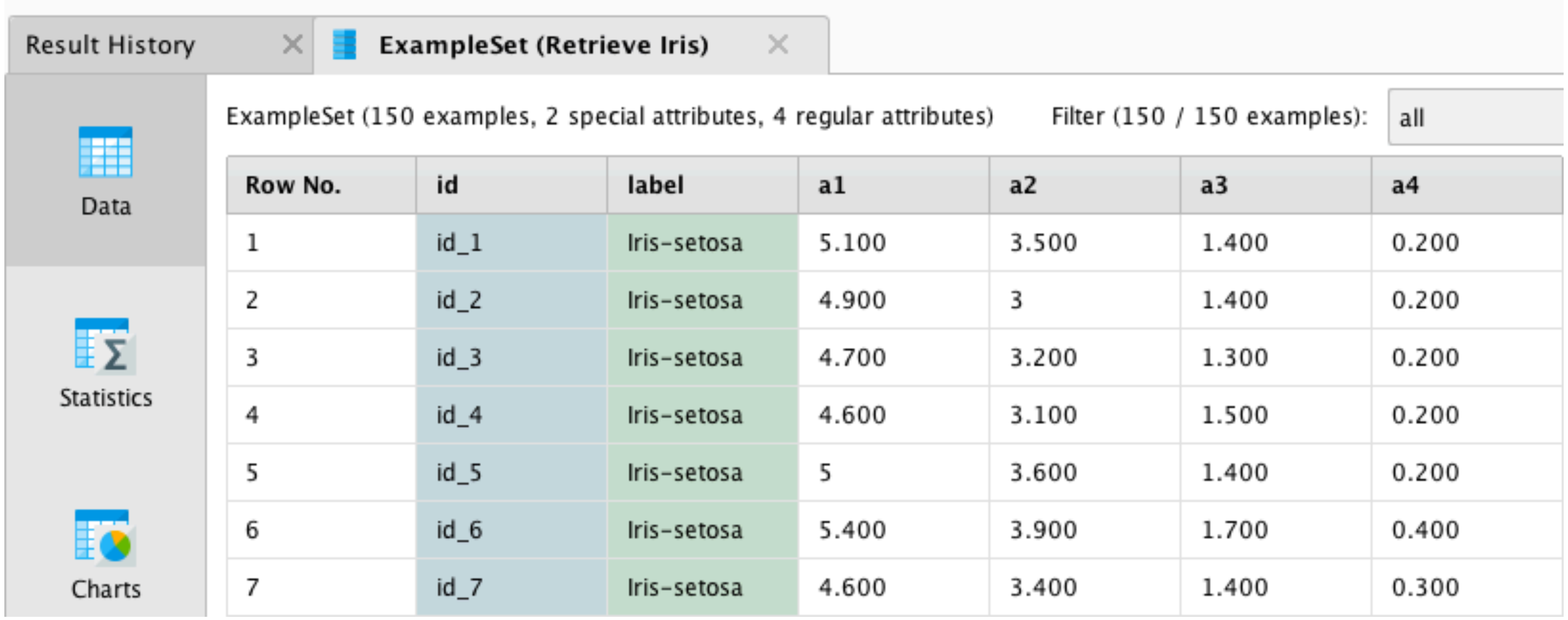


# Questions 1/2

- Switch to Data.
  - How many attributes are in the dataset?
  - How many samples are in the dataset?
  - How many classes are in the dataset?
  - How many samples are in each class?

# Answers 1/2

- There are 4 attributes in the dataset.
- The dataset has 150 examples.
- There are 3 classes in the dataset.
- Each class has 50 samples.



ExampleSet (150 examples, 2 special attributes, 4 regular attributes)    Filter (150 / 150 examples):

| Row No. | id | label | a1 | a2 | a3 | a4 |
|---|---|---|---|---|---|---|
| 1 | id_1 | Iris-setosa | 5.100 | 3.500 | 1.400 | 0.200 |
| 2 | id_2 | Iris-setosa | 4.900 | 3 | 1.400 | 0.200 |
| 3 | id_3 | Iris-setosa | 4.700 | 3.200 | 1.300 | 0.200 |
| 4 | id_4 | Iris-setosa | 4.600 | 3.100 | 1.500 | 0.200 |
| 5 | id_5 | Iris-setosa | 5 | 3.600 | 1.400 | 0.200 |
| 6 | id_6 | Iris-setosa | 5.400 | 3.900 | 1.700 | 0.400 |
| 7 | id_7 | Iris-setosa | 4.600 | 3.400 | 1.400 | 0.300 |

# Questions 2/2

- Switch to Charts -> Plot View.
    - Is there any attribute which differentiates among the Irises?
    - Which attribute is the best one for differentiating among the classes?

# Answers 2/2

- The best attribute which differentiates among classes does not exist (versicolor mixes with virginica).

- The best attribute for separation is a3 or a4.

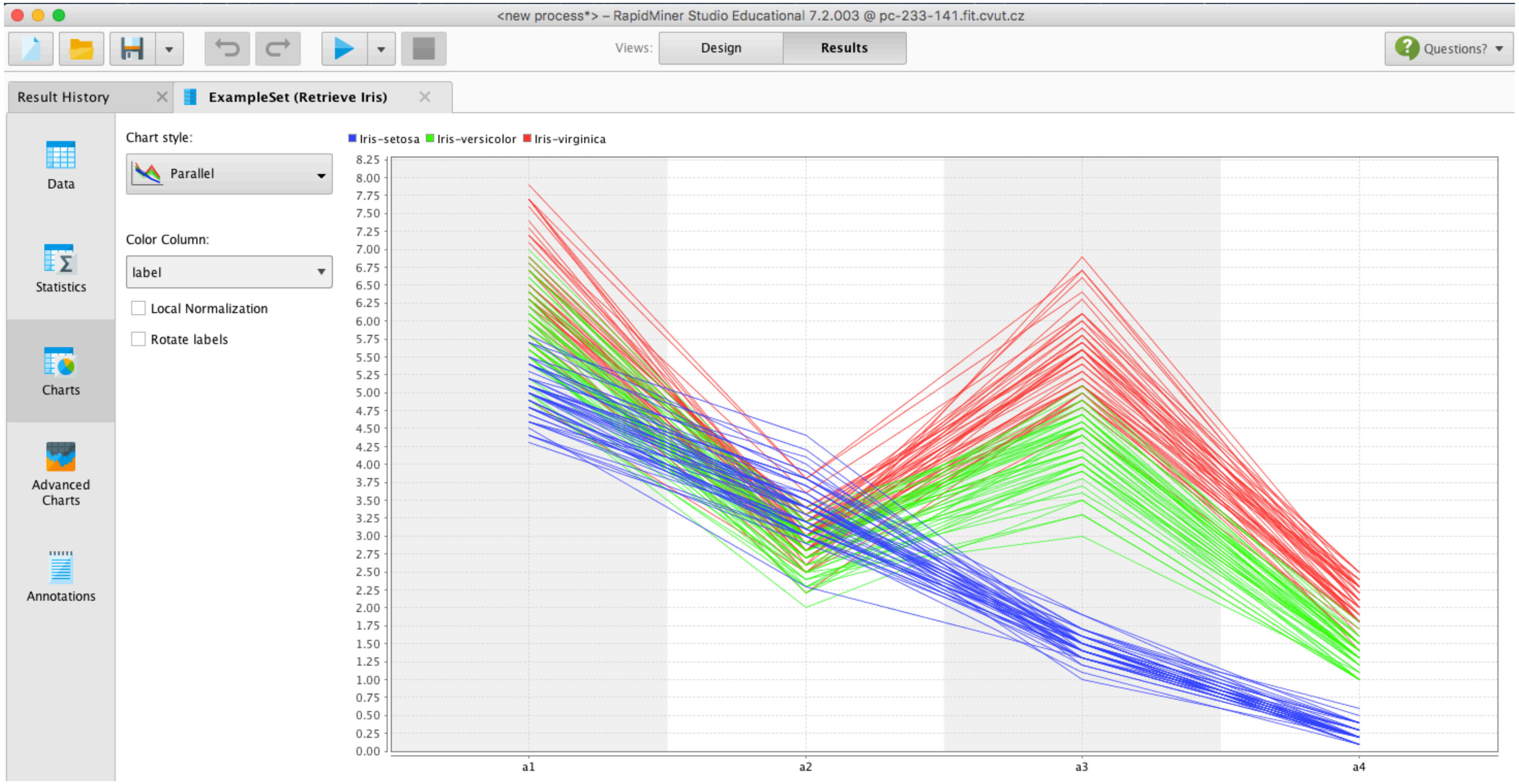


Chart style: Parallel

Color Column: label

☐ Local Normalization

☐ Rotate labels

# Some Popular Tools

- Tableau Software (student license for free)
  http://www.tableau.com

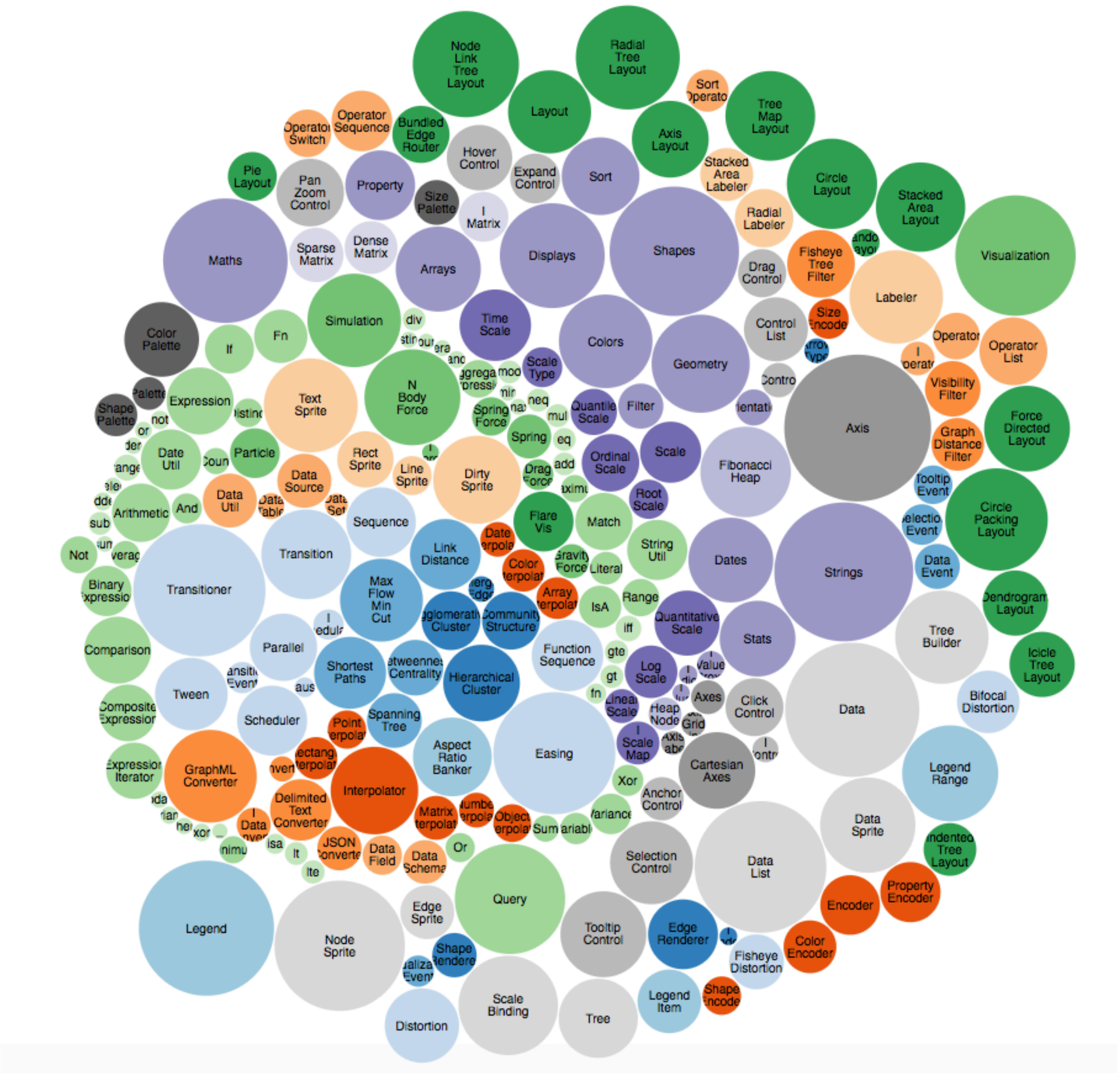
- Google Data Studio (beta version)
  https://www.google.com/analytics/data-studio/

- Chart.js
  http://www.chartjs.org

- D3.js
  https://d3js.org

- Visualize Free
  http://visualizefree.com

- Gephi
  https://gephi.org



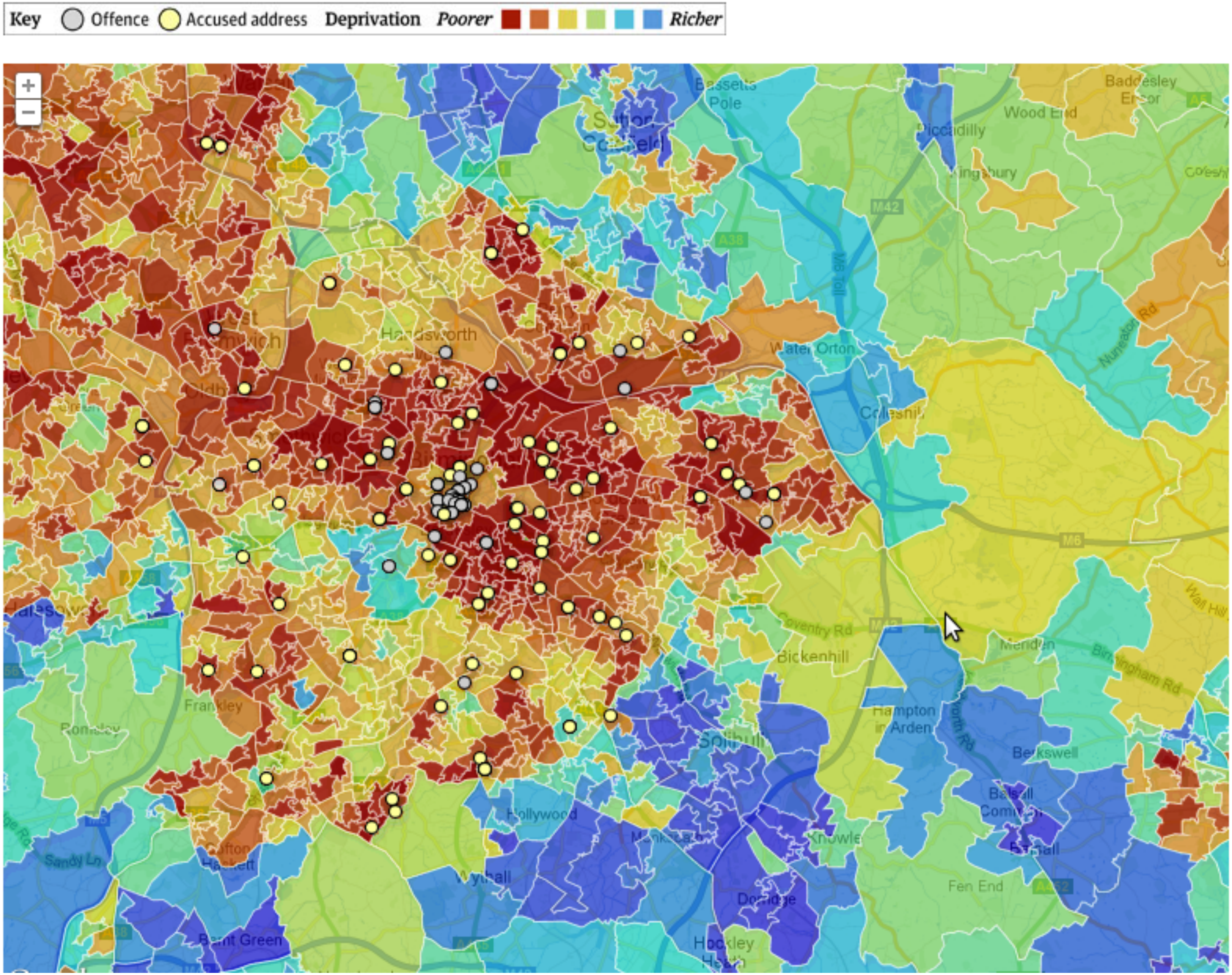
- Microsoft Power BI (Cloud version for free, register with another domain than google, yahoo etc.)
  https://powerbi.microsoft.com/en-us/

- Tag Cloud
  - Create Tag Cloud for an arbitrary document.
  - http://www.wordle.net
  - Meaningless word (a, the, for, etc.) can be removed by right-clicking.

UFO
AIR
PROJECT
FORCE
FLYING
OBJECTS
SIGHTINGS
REPORT
REPORTS
GOVERNMENT
CASES
REPORTED
INVESTIGATION
BOOK
STUDIES
OBJECT
STUDY
UNIDENTIFIED
RETRIEVED
AIRCRAFT
MILITARY
EXTRATERRESTRIAL
CONDON
SAUCER
PHENOMENON
PHENOMENA
SCIENTIFIC
BLUE
USAF
SEE
TWO
NEW
MORE
NATIONAL
INCIDENT
SECRET
ONE
CIA
NEAR
ARMY
HYNEK
PRESS
EVIDENCE
PHYSICAL
FBI
ORG
CASE
LATER
MADE
WWW
COM
TIME
PANEL
OVER
MANY
MARCH
DOCUMENTS
NEWS
PUBLIC
AERIAL
CRAFT
VARIOUS
AGENCY
SIGN
DEAD
UNITED
SPACE
DEFENCE
UNEXPLAINED
INTELLIGENCE
RELEASED
HISTORY
LIST
BOOKS
PAGE
ARTICLES
TERM
ALIENS
RESEARCH
FILES
HIGH
TOP
FRENCH
WAR
LINK
MAIN
YEARS
CONSPIRACY
COMMITTEE
OFFICIAL
FIRST
UFOLOGY
KNOWN
ARTICLE
STATED
LIFE
ANOTHER
THREE
CNES
CONCLUSION
POLL
FOUND
ALIEN
STATES
FURTHER
WIKIPEDIA
ASSOCIATED
PLOT
INVESTIGATIONS
REAL
HOAXES
DIRECTOR
HILL

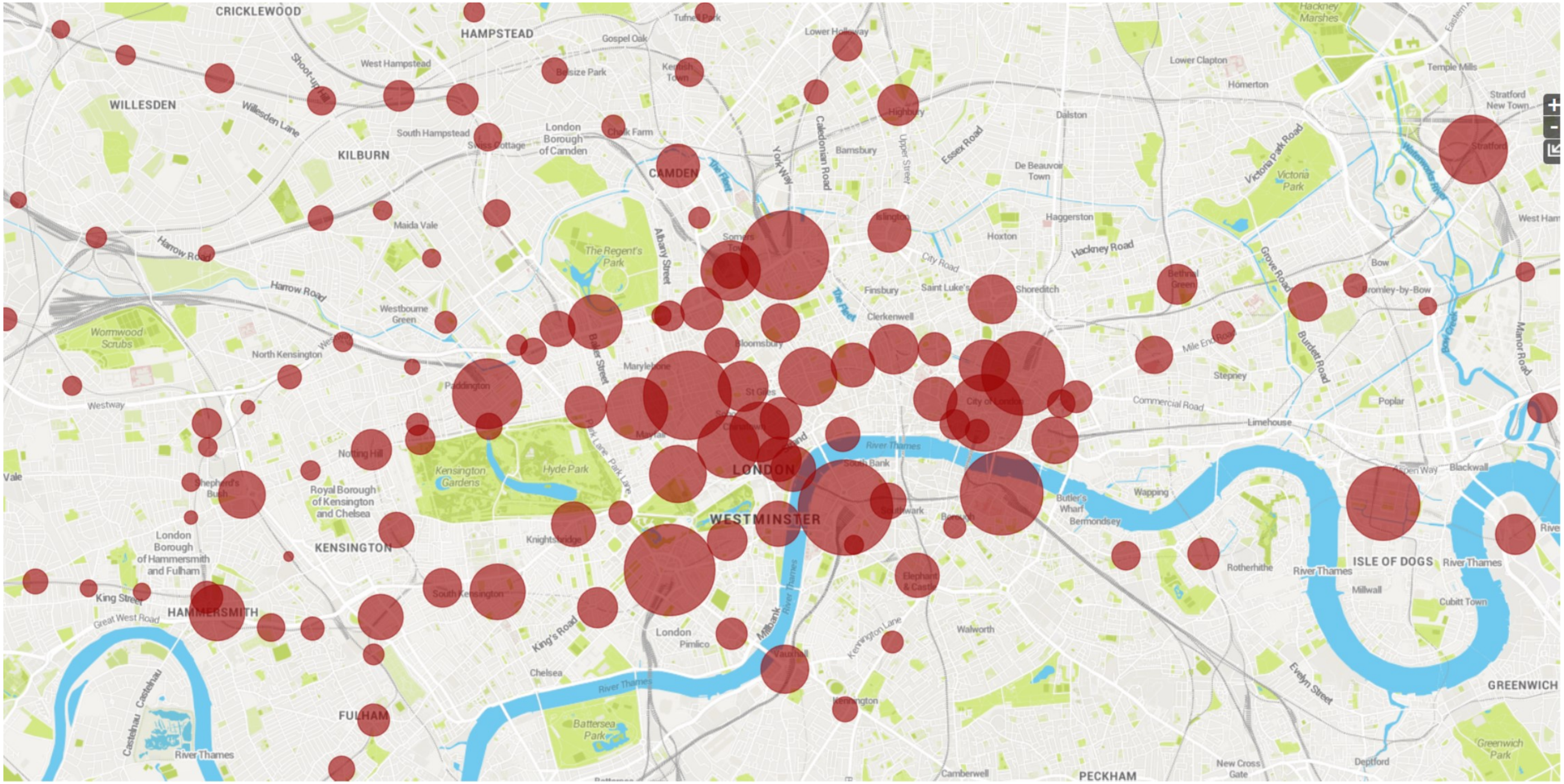
BIG
DATA
INFORMATION
STORAGE
LARGE
SOFTWARE
SAN
NAS
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
HAND
CAPACITY
CHALLENGES
LIMITS
COMPLEX
ANNOUNCED
PAST
MILLION
SETS
INCLUDE
NETWORKS
HYPOTHESIS
SCIENCE
END
MOST
NEARLY
MAPREDUCE
FUTURE
CRITIQUES
GOVERNMENT
FACTOR
SYSTEMS
MANAGE
ONLINE
DAY
WORLDWIDE
PROCESS
MAKING
INTEGRATION
VOLUME
ONE
OUT
DEVELOPMENT
TECHNOLOGY
MANY
APPROACH
BILLION
NEW
PARALLEL
STATE
CRITIQUE
USE
PEOPLE
TRAFFIC
CURRENTLY
TECHNOLOGIES
MASSIVE
DEFINITION
USING
FRAMEWORK
BETWEEN
STATISTICS
BUSINESS
EFFECTIVE
VISUALIZATION
DATABASES
SIZE
ANALYTICS
PROCESSING
SIMULATIONS
MANAGEMENT
EXABYTES
DETERMINE
WELL
SET
DEPARTMENT
QUERIES
HIGH
USERS
HIGHER
ALGORITHMS
COMPANIES
PETABYTES
AMOUNT
RATE
REAL
STRUCTURE
WORK
APPLICATIONS
INFRASTRUCTURE
ANALYSIS
TERMS
PRIVATE
RESEARCH
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LEARNING
LESS
HUGE
INTERNET
EVERY
WORLD
COMPARED
VARIOUS
BASED
TIMES
SHARED
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
INTELLIGENCE
FLOW
APPROACHES
PROCESSES
SEARCH
ARCHITECTURE
PARADIGM
FUNDING
CENTER
RECORDS
COST
INSIGHT
TOOLS
INITIATIVE
ECONOMIC
PROJECT
DISTRIBUTED
PER
RELEVANT
TERABYTES
MARKET
VARIETY
ANNUAL
USED
RELATED
TIME
SENSOR
WOULD
NOW

- Google Fusion Tables
  http://www.google.com/fusiontables
  - Search for interesting data.
  - Visualize an example, using the “Map”.

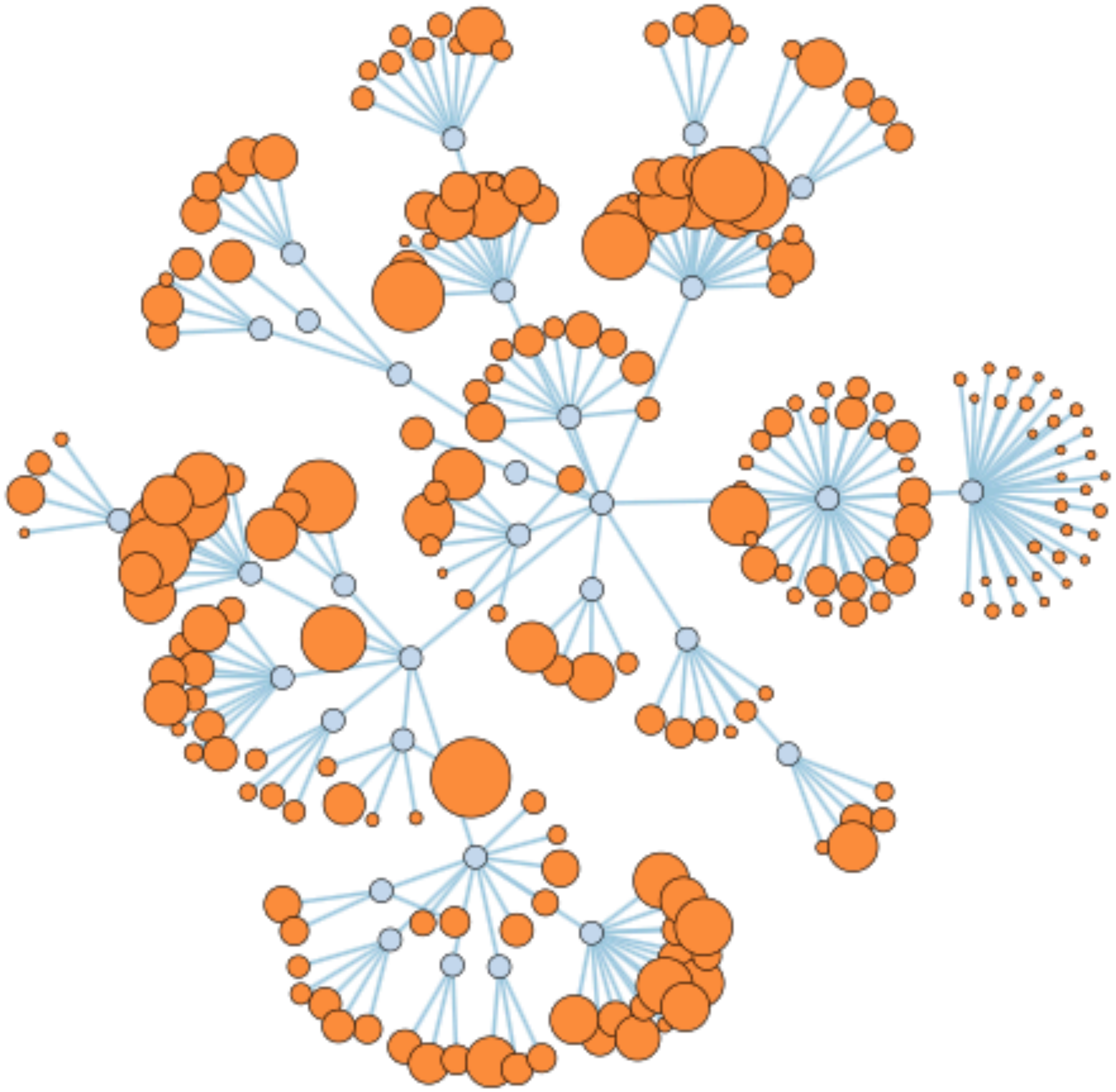

Key
○ Offence
○ Accused address
Deprivation
Poorer 🟥 🟧 🟨 🟩 🟦 🟦 Richer
Bassetts Pole
Sutton Coldfield
Baddesley Ensor
Wood End
Piccadilly
Kingsbury
Coleshill
A5
A4041
M42
A38
M6 Toll
West Bromwich
Handsworth
Water Orton
Oldbury
Birmingham
Nuneaton Rd
M6
Halesowen
Coventry Rd
Bickenhill
M42
Meriden
Birmingham Rd
Wall Hill
Hampton in Arden
Frankley
Romsley
Solihull
Berkswell
Balsall Common
Balsall
Hollywood
Monkspath
Knowle
Cofton Hackett
Sandy Ln
Wythall
Fen End
A452
Dorridge
Barnt Green
Hockley Heath

- mapsdata
  http://www.mapsdata.co.uk/

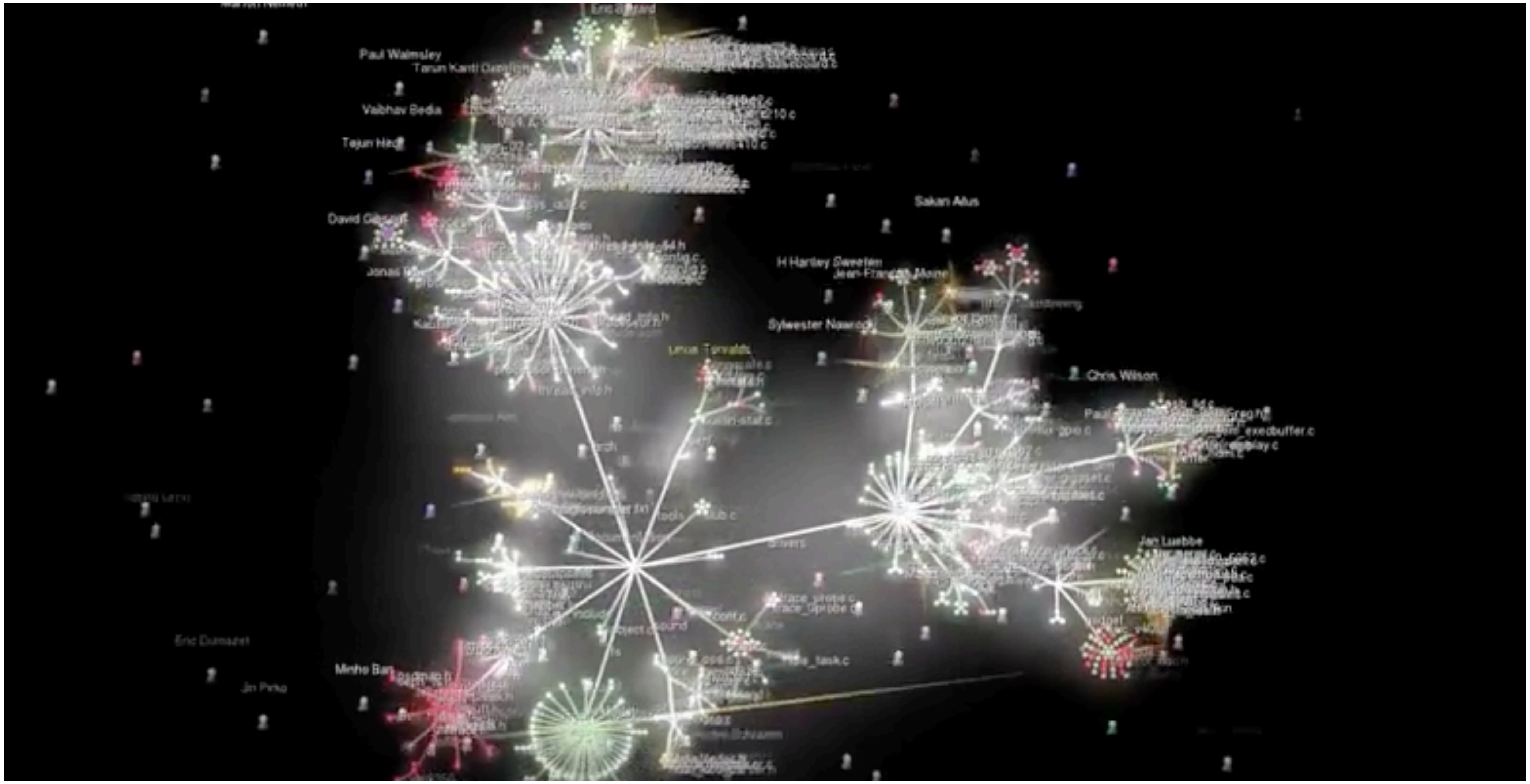
  - The Busiest Stations on the London Underground



- D3.js http://mbostock.github.io/d3/talk/20111116/force-collapsible.html



- Linux Kernel Development Visualization https://www.youtube.com/watch?v=P_02QGsHzEQ


Paul Walmsley
Vaibhav Bedia
David Gibson
Sakari Ailus
H Hartley Sweeten
Sylwester Nawrocki
Linus Torvalds
Chris Wilson
Jan Luebbe
Minho Ban
Eric Dumazet
Jiri Pirko
drivers
trace_probe.c
execbuffer.c