Instituto de Ciências Matemáticas e de Computação

| Universidade de São Paulo |

VISUALIZATION

In the context of ‘data science’ & ‘big data’

Maria Cristina F. Oliveira
ICMC-USP
cristina@icmc.usp.br

Outline

- *Data Science, e-Science & Big Data*
- Data Visualization & *Visual Analytics*
- A Sample of Visualization Techniques & Applications
- Research Challenges

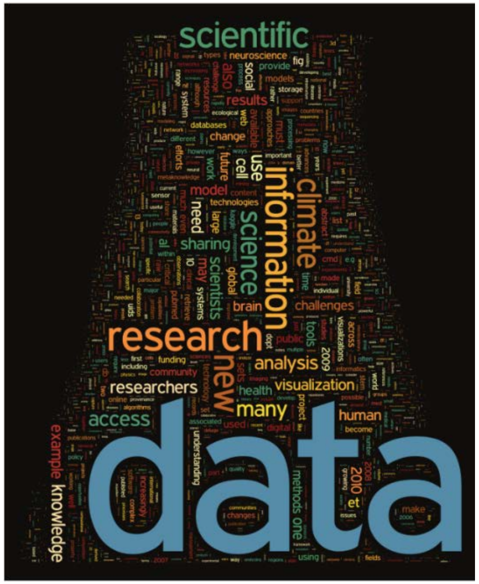
Data Science: what does it take?

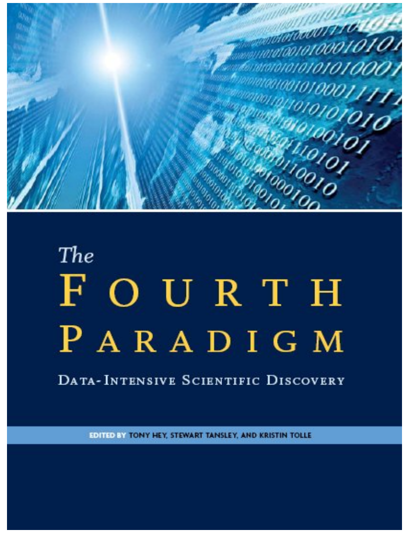
- Algorithms
- Statistics – essential
  - Alone will not do the job
- Mining – essential
  - Will not do the whole job, even with statistics
- Visualization – exploratory situations and user centric decision
- Certain skills – from complex reasoning to complete programming to innovative and daring goals. But mostly: understanding the data

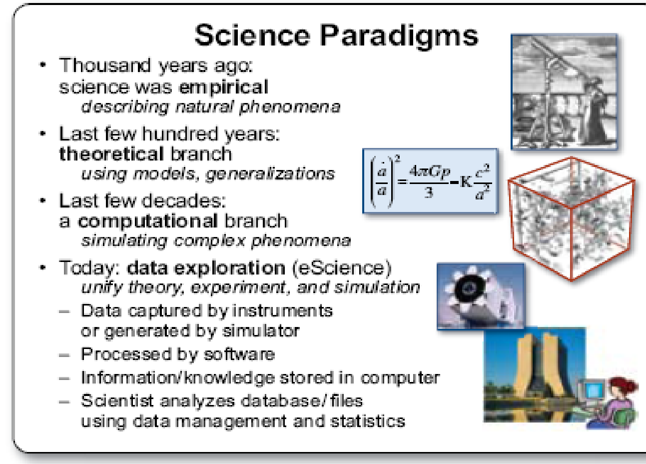
Qualification - keywords

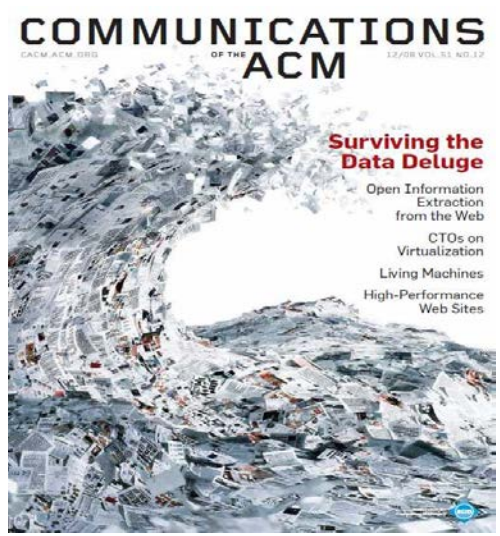
- Ex: Coursera (https://www.coursera.org/)
  - Set of (10) courses on Data Science by Johns Hopkings University
    - Intro (concepts + infra – version control and R IDE)
    - R Programming
    - Getting and cleaning data
    - Exploratory data analysis – visualization and such
      - Buzz words – visual analytics
    - Statistical Inference
    - Regression Models
    - Reproducible Research
    - Practical Machine Learning
    - Developing data products – making results usable
    - Data science capstone (‘graduation project’)

e-Science: The Fourth Paradigm









USP e-Science Research Network

http://escience.ime.usp.br/

RM Cesar et al.

On the limits of the reductionist approach!

Big Data

- “Big data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets.”

Source: Boyd & Crawford, **Critical Questions for Big Data**, *Information, Communication & Society* 15(5), 2012.

Some numbers

- In 2020: 7B people, 30 Billion Devices, 44 Zettabytes of Data
- How advantageous:

Potential Productivity Gains - the power of 1%

	Segment	Savings	15 yr. Value
Aviation	Commercial	1% fuel	$30B
Power	Gas fired generation	1% fuel	$66B
Healthcare	System wide	1% reduced inefficiency	$63B
Rail	Freight	1% reduced inefficiency	$27B
Oil & gas	Exploration & development	1% reduction in CAPEX	$90B

Better Health Care Through Data

How health analytics could contain costs and improve care

By KATHY PRETZ 8 Setembro 2014

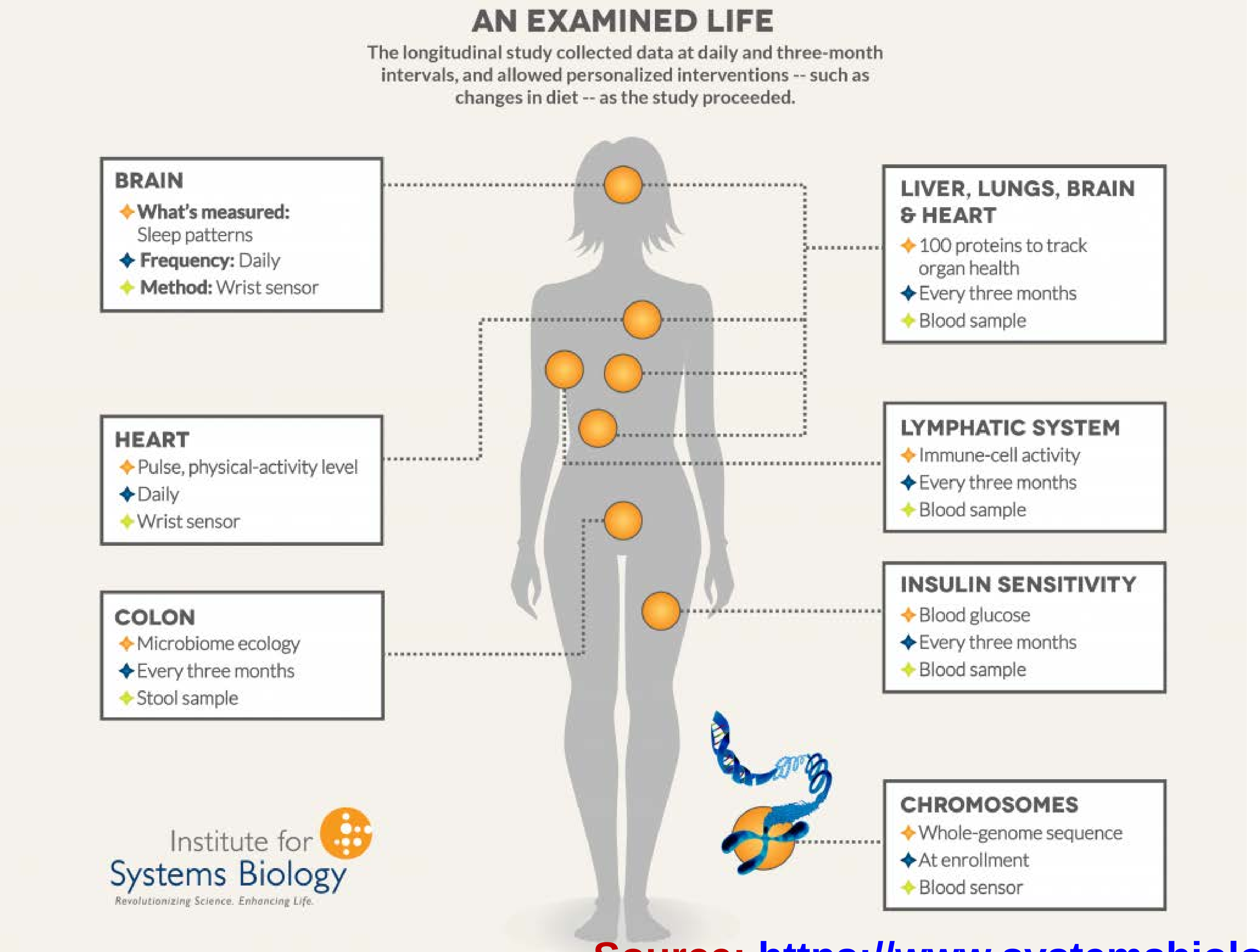
Image: iStockphoto

This article is part of our September 2014 special report on [***big data***](/static/special-report-big-data)*, covering technologies that support and make sense of the growing mountains of data, and several of its applications.*

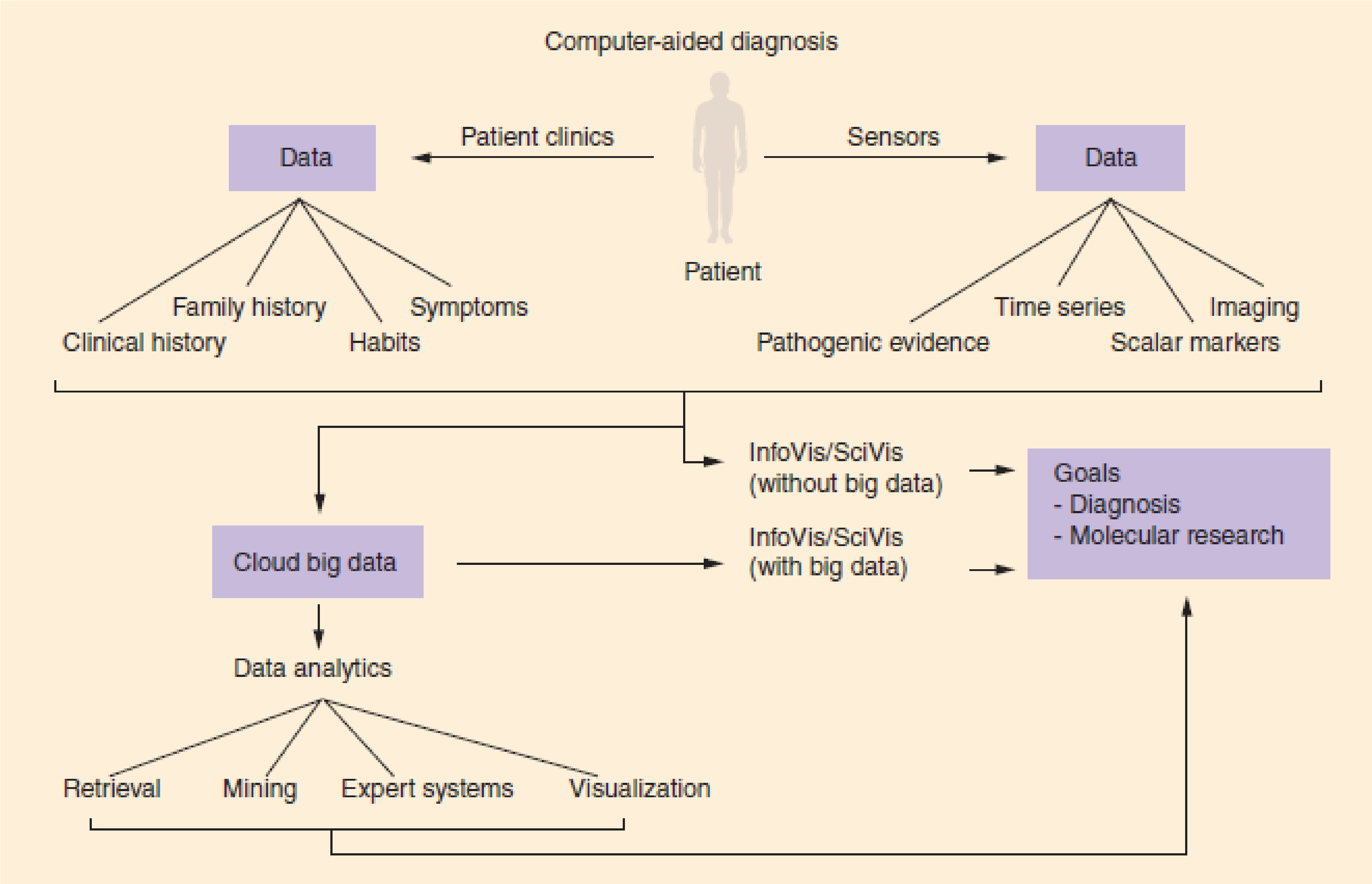
It's no surprise that keeping people healthy is costing more money. From the price of medications and the cost of hospital stays to doctors' fees and medical tests, health-care costs around the world are skyrocketing. The World Health Organization attributes much of this to wasteful spending on such things as ineffective drugs and duplicate procedures and paperwork, as well as missed disease-prevention opportunities.

Source: *IEEE Spectrum*, Sept 2014

Hundred Person Wellness Project (HPWP)



Source: https://www.systemsbiology.org/



Source: Rodrigues Jr. et al. On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis. *Nanomedicine* 2016

Data is...

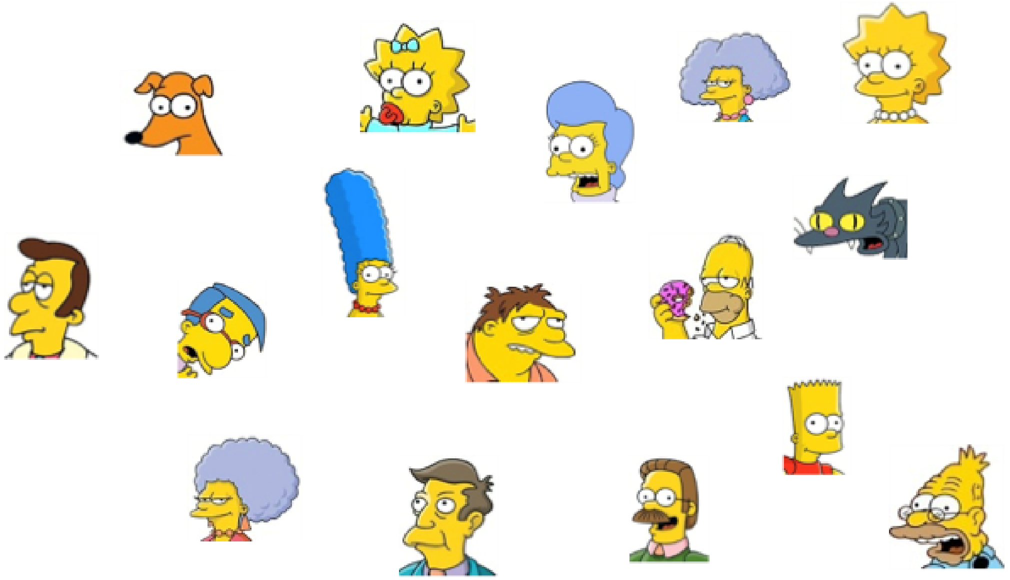
- Far too complex... (many atributes)
- Far too big... (‘easy’ to collect)
- Far too varied... (images, videos, documents, news, networks)
- Never ending... (data streams)
- Much redundancy...
- Many relationships...
- Pieces missing...

- Studying natural & artificial systems and phenomena implies in handling lots of data

Data interpretation problem

- People trying to make sense of data

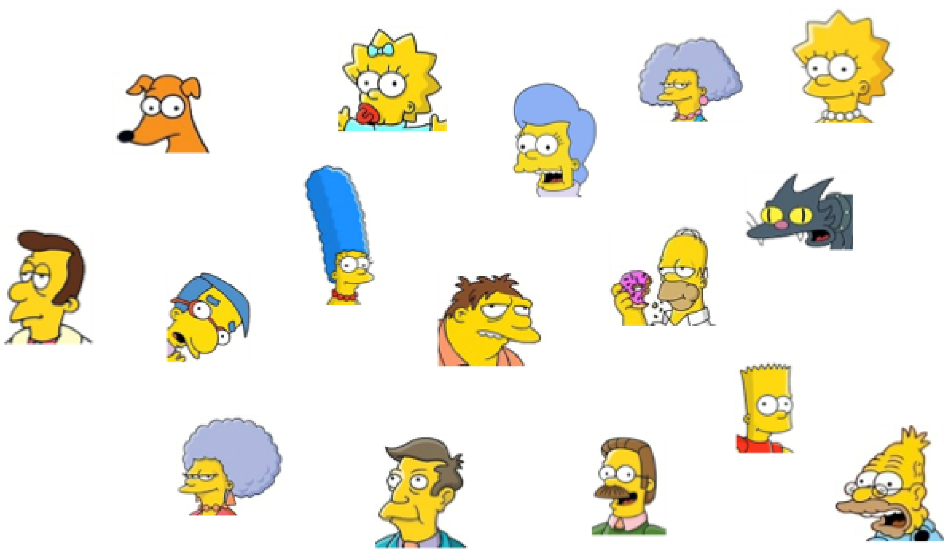
‘messy’ data

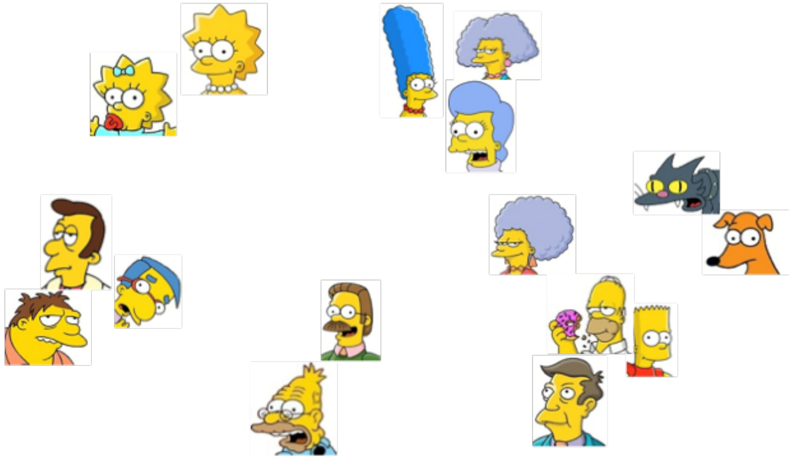


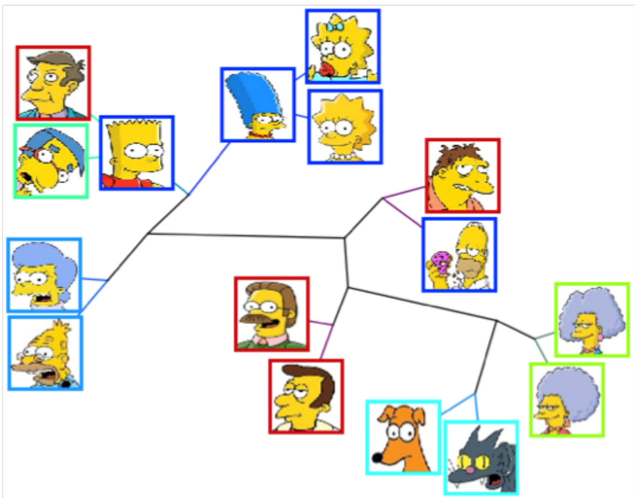
What does your data tell???

- Visualization can help making sense of data

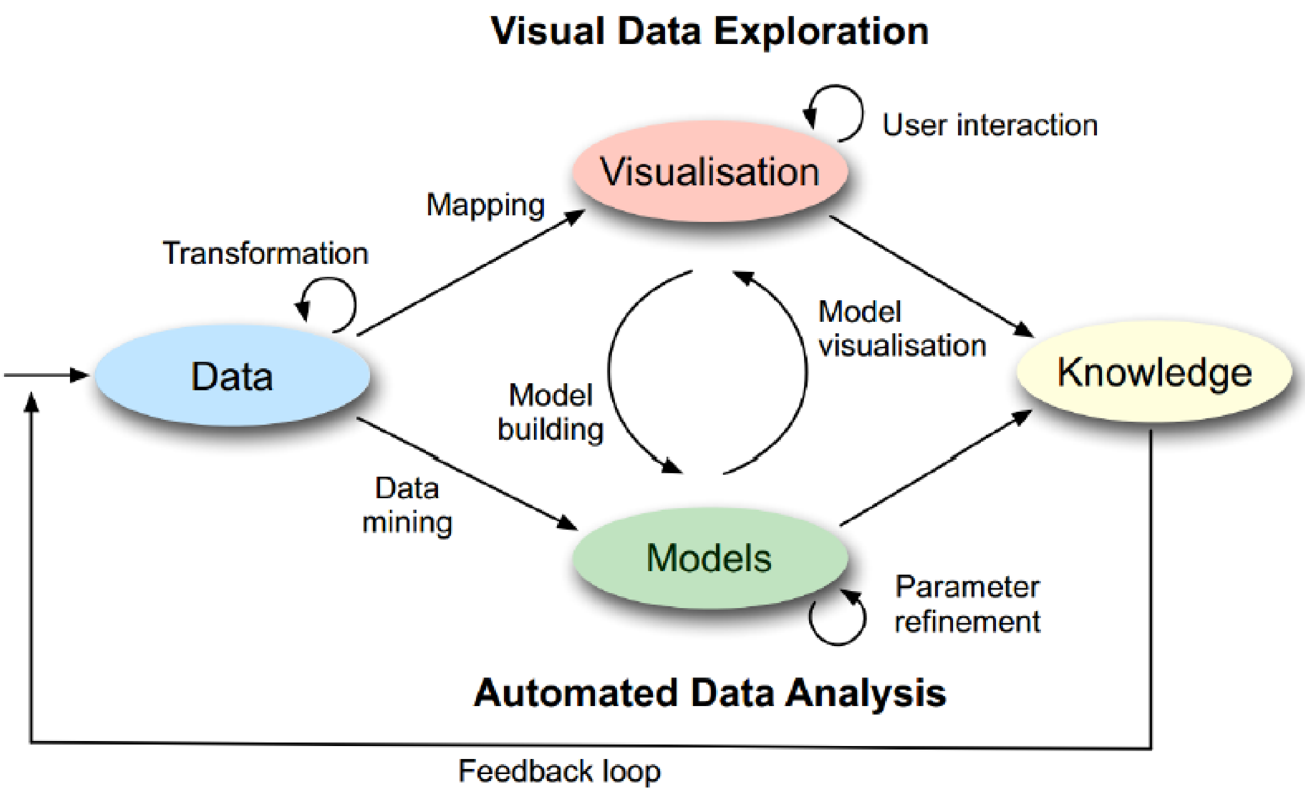
‘messy’ data







Visual Analytics process



Source: Keim et al. 2010

Multidimensional data: representation

15														
6	2													
15	1	2												
15	1	2	1											
15	2	1	2	2										
15	3	10	8	15	15									
15	10	7	10	10	7	15								
15	15	10	15	15	10	15	16							
15	15	10	10	10	10	15	16	16						
15	15	15	15	15	15	15	15	3	15					
15	10	10	10	10	10	15	15	15	15	15				
15	10	10	10	10	10	15	15	15	1	15	15			
15	10	15	12	15	15	10	15	15	15	15	15	15		
15	10	10	10	10	10	15	15	15	15	15	2	8	15	
15	10	10	10	10	10	15	1	15	15	15	15	15	15	15

pairwise distances

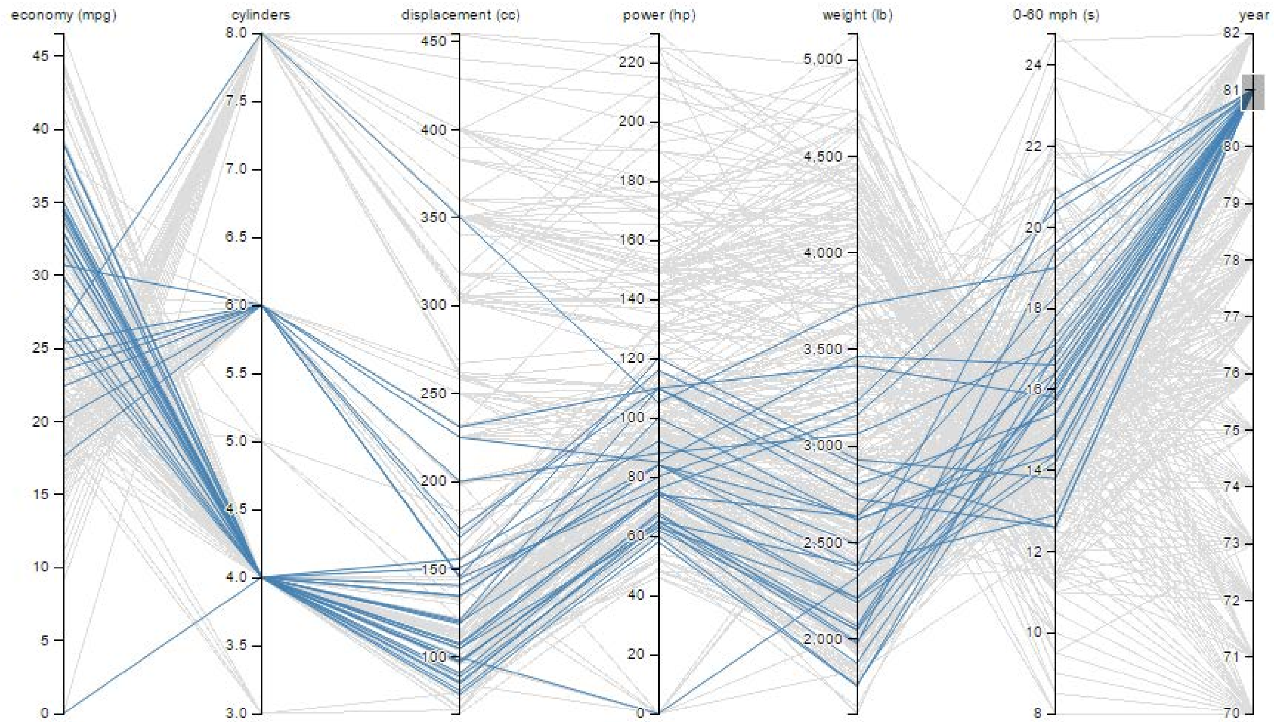
and/or

5	12	15	2	7	5	0	12	9	0	8
12	5	0	12	12	12	12	12	18	12	12
0	1	05	10	15	12	8	12	9	11	5
0	12	01	12	9	0	12	10	5	5	12
12	8	05	12	12	12	8	12	9	12	12
10	12	0	11	10	2	7	12	2	16	7
5	6	8	12	12	15	12	6	9	17	0
7	12	05	0	12	12	10	17	9	12	12
2	10	05	15	12	1	12	10	9	8	2
12	12	7	12	0	12	0	12	10	12	12
6	12	05	17	12	10	12	12	9	12	8
12	10	2	12	1	12	12	11	6	0	12
1	12	05	12	12	16	2	12	9	12	0
10	0	12	12	9	12	0	10	12	12	8
0	12	1	12	12	5	1	7	11	12	12
8	2	11	10	7	12	5	12	15	10	0

dimensional embedding
(feature space)

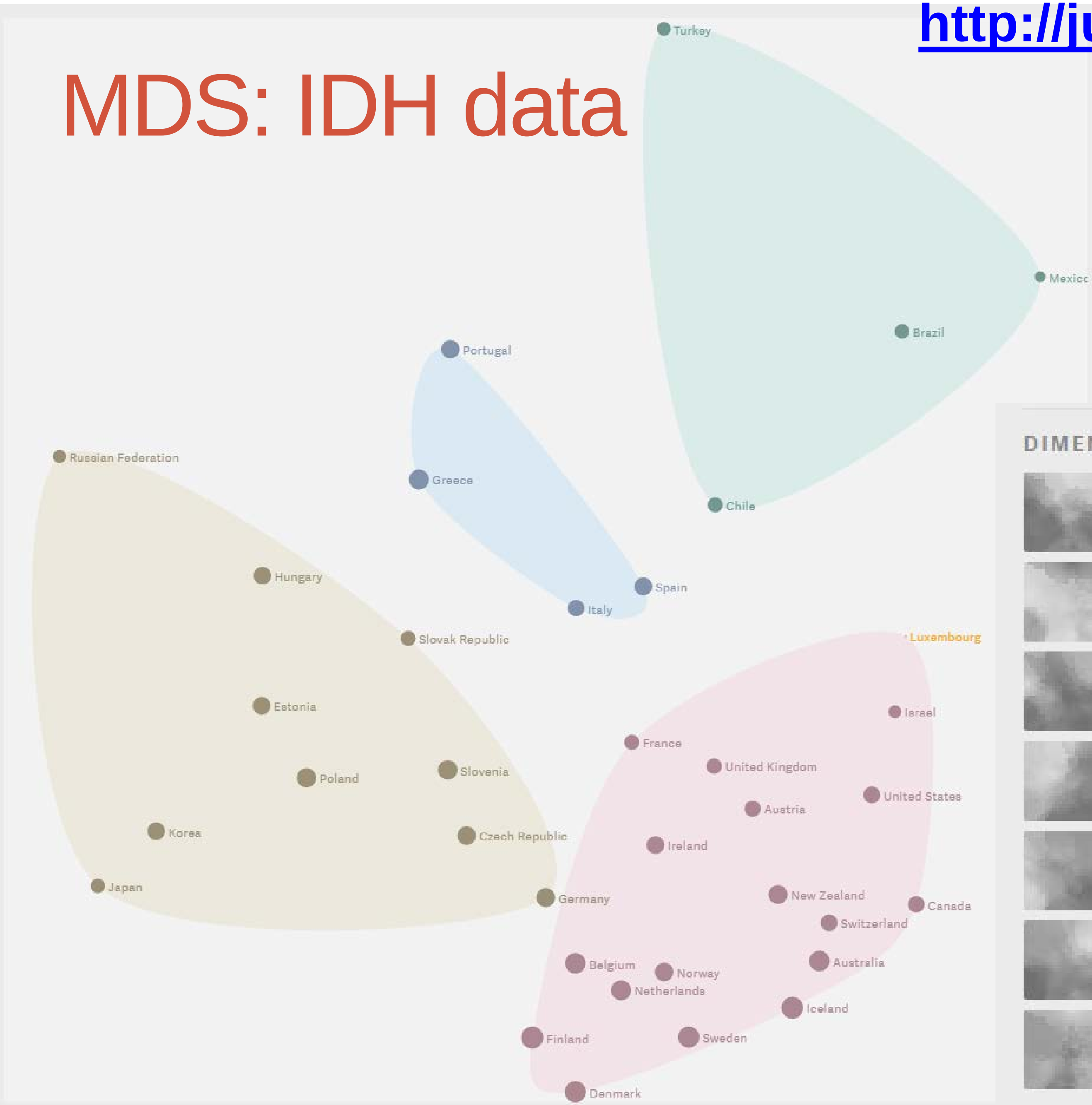
Parallel Coordinates

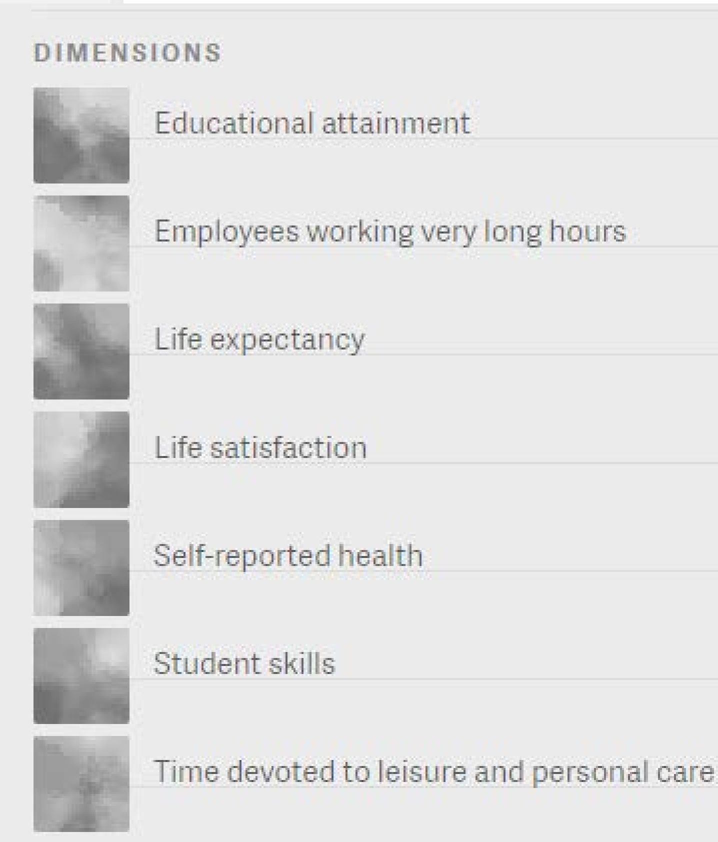
- https://bl.ocks.org/jasondavies/1341281
- http://mbostock.github.io/d3/talk/20111116/iris-parallel.html



http://jujujulian.com/mds/

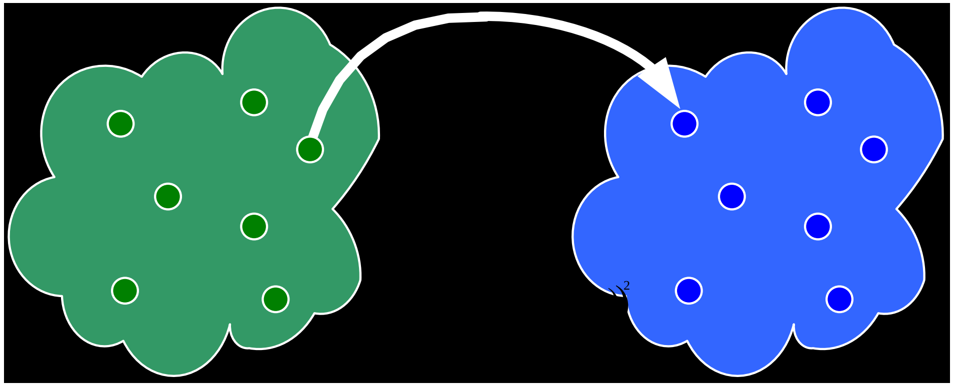
MDS: IDH data





Multidimensional projection

$$X \in R^m \quad f \quad Y \in R^{k=\{1,2,3\}}$$



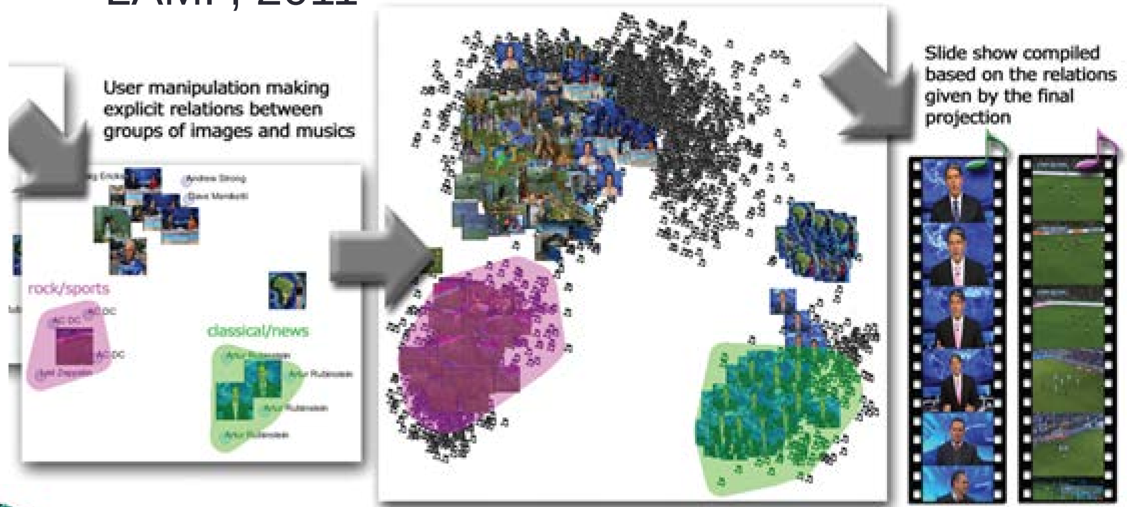
- δ: x_i, $x_j \rightarrow R$, $x_i, x_j \in X$
- d: y_i, $y_j \rightarrow R$, $y_i, y_j \in Y$
- f: $X \rightarrow Y, |\delta(x_i,x_j) - d(f(x_i), f(x_j))| \approx 0, \forall\, x_i,x_j \in X$

$$E = \frac{\sum_{ij}(\delta(x_i,x_j) - d(y_i,y_j))^2}{\sum_{ij}\delta(x_i,x_j)^2}$$

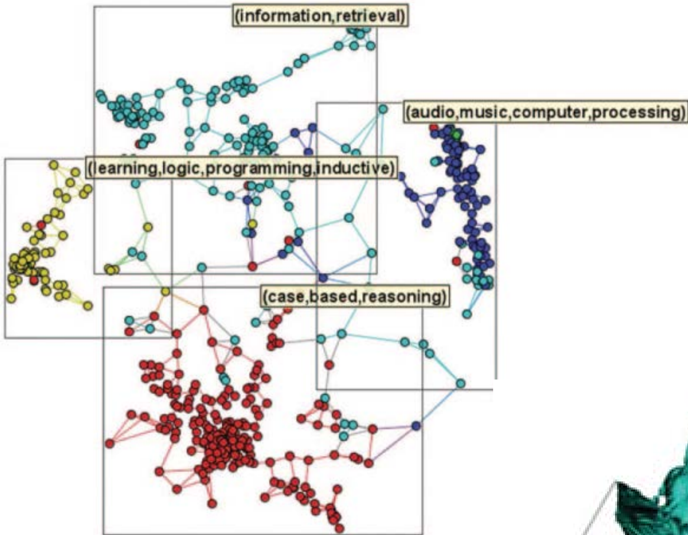
Multidimensional projection

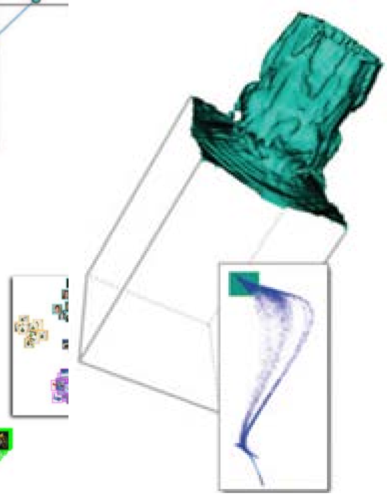
- old idea, old & new techniques…
- current techniques must comply with requirements imposed by interactive applications:
  - speed (low computation cost)
  - capability to handle very large & massive data
  - interactivity (allow user intervention)

LAMP, 2011

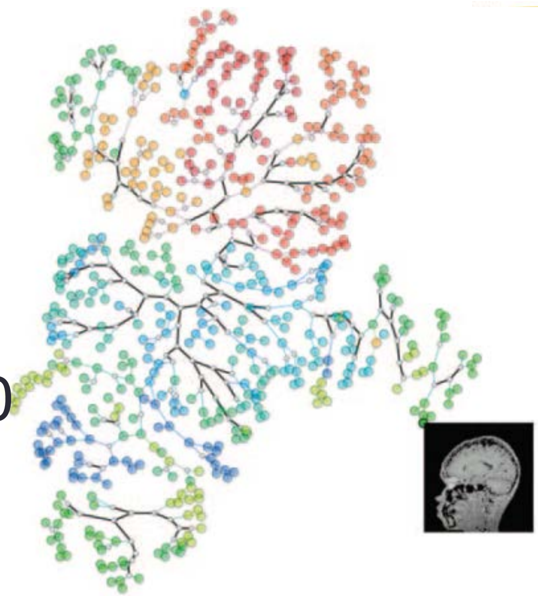


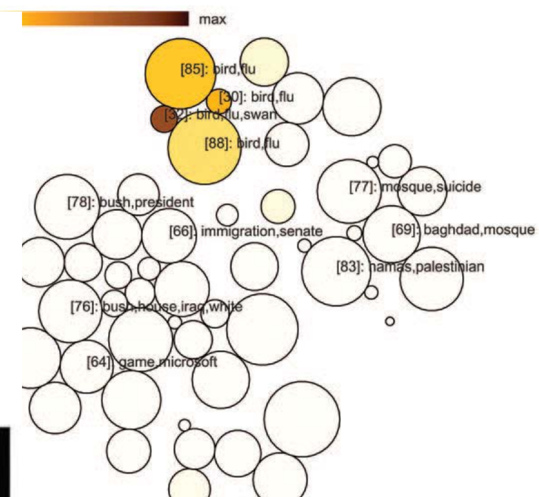
LSP, 2008





NJ trees, 2007

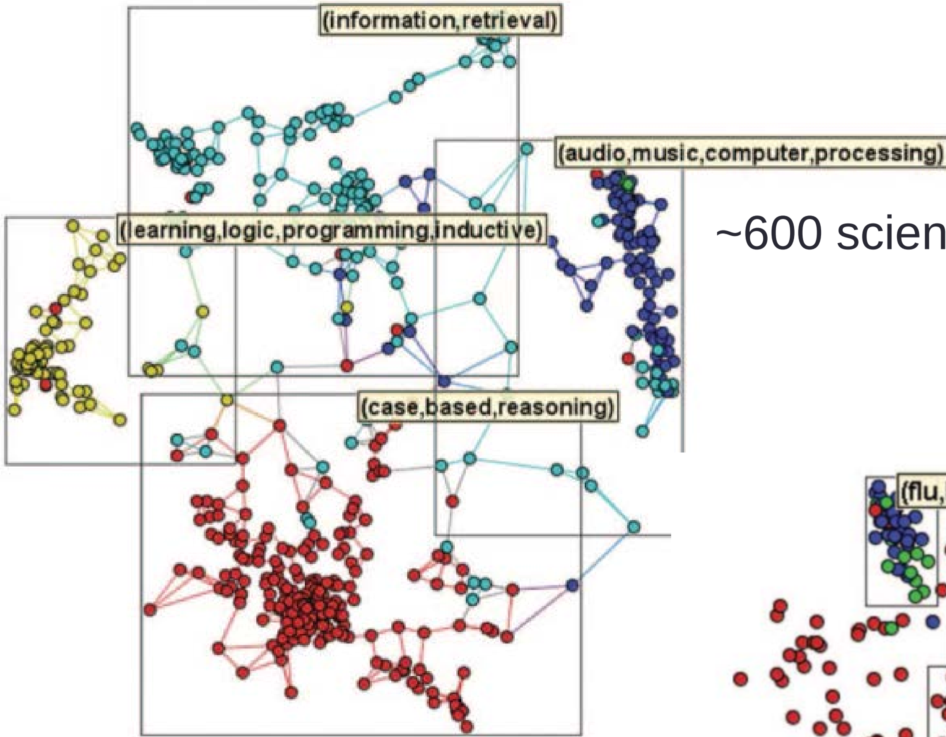




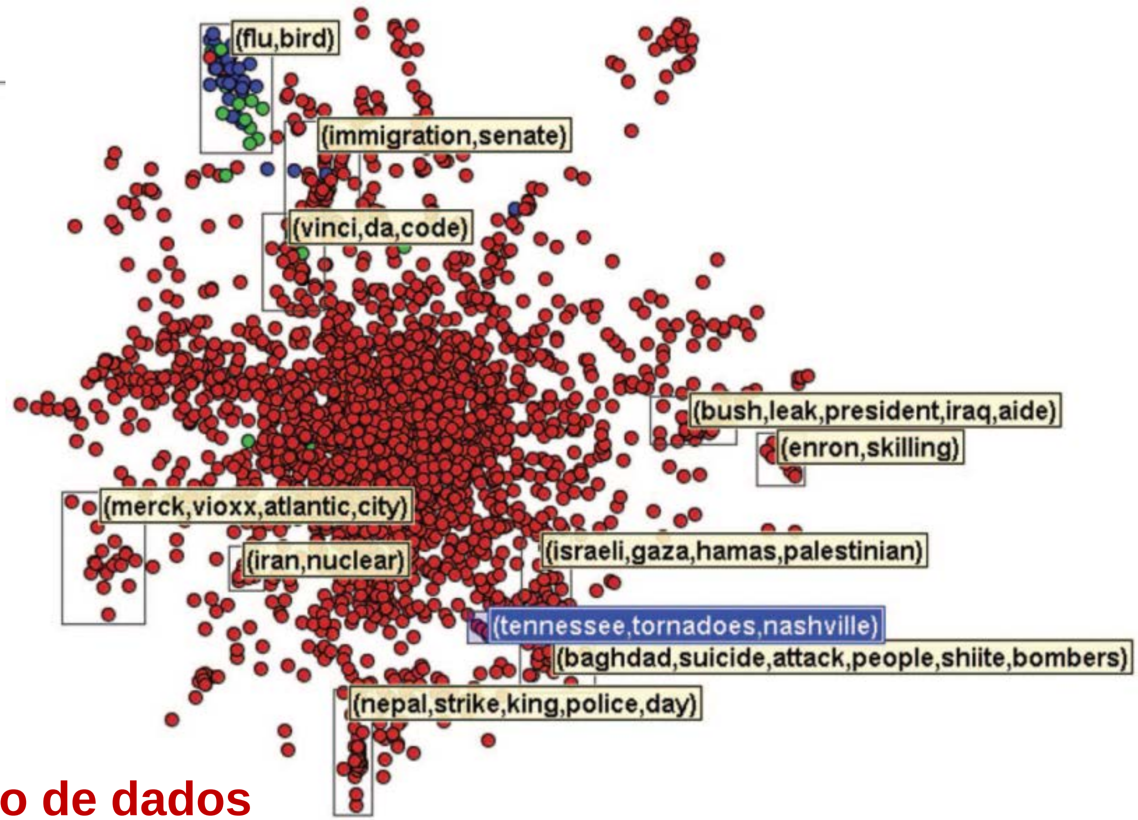
PLP 2011

PLMP 2010

HiPP, 2008



~600 scientific papers

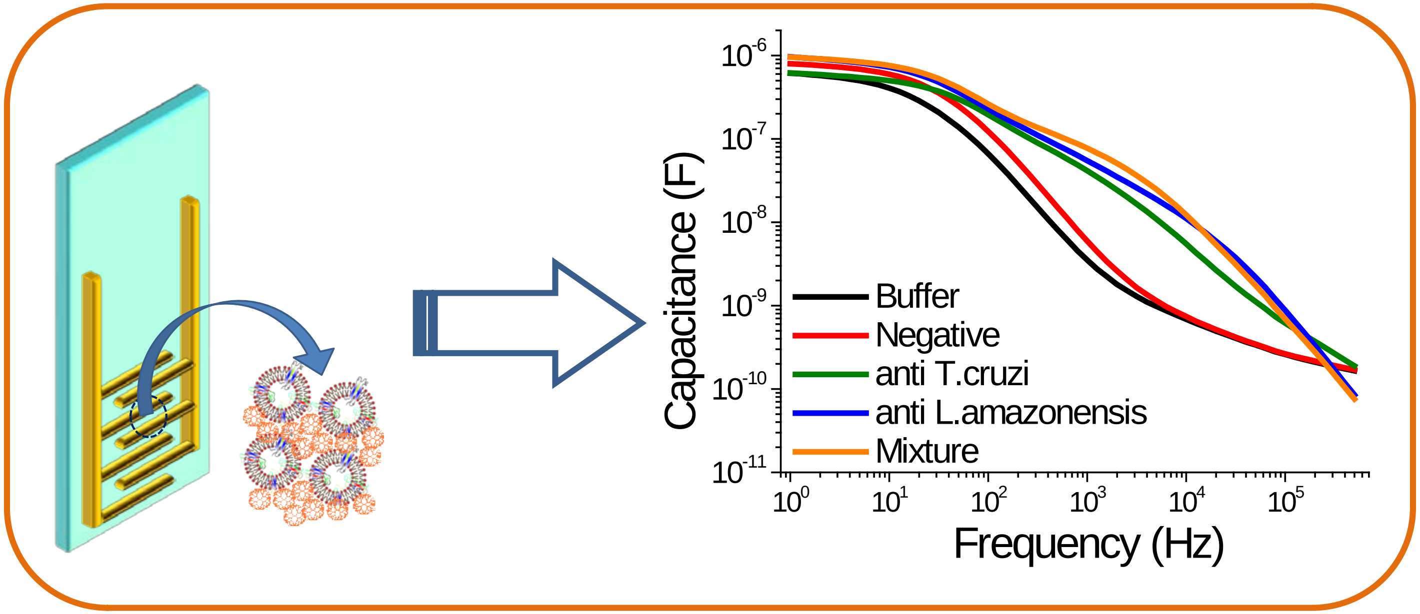


~2,000 RSS news feeds (2006)

Source: F. Paulovich, Mapeamento de dados multidimensionais: integrando mineração e visualização. D.Sc. Thesis, 2008

22

A real scenario



- molecular interaction between different materials produce electrical responses that can be measured, e.g., with impedance spectroscopy

Source: Osvaldo N Oliveira Jr., IFSC-USP

Biosensor data analysis

- sensor to detect the presence of antibodies for Chagas' Disease (caused by Tripanosoma Cruzi) or Leishmaniasis in blood samples
- sensors to detect glucose and triglycerides at very low concentrations, electronic 'tongues', …
- test a wide variety of sensor configurations to obtain optimal selectivity and sensitivity: lots of measurements, very dynamic scenario

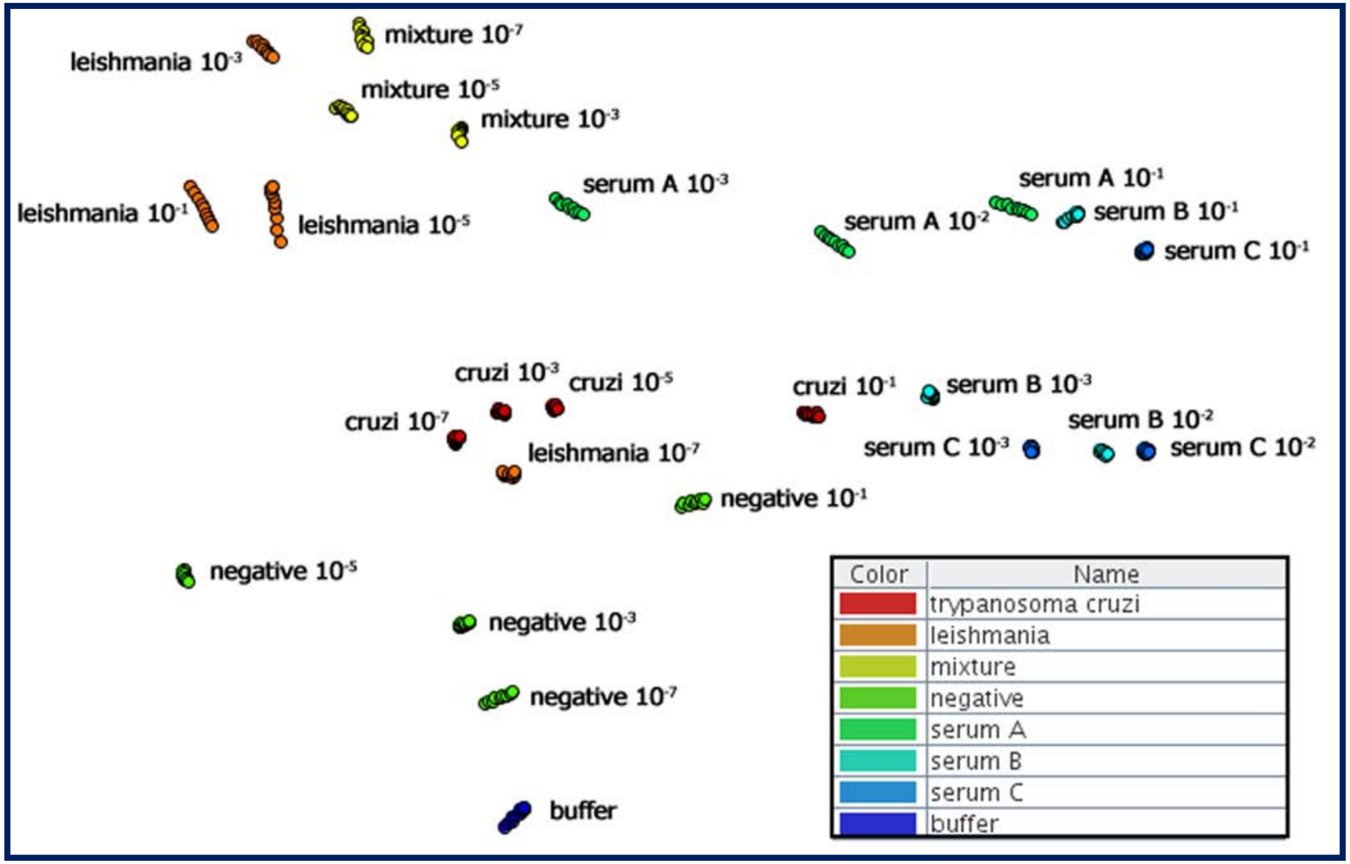
Biosensor data analysis

- Goals
  - finding an optimal sensor (thin film architecture) or optimizing performance of existing sensor
    - sensitivity & selectivity
  - understanding/explaining why it is optimal

Example: T. Cruzi x Leishmania

- 8 types of analytes
  - 25 different substances (some analytes at different concentrations), 9 samples each: 25 x 9 = 225 samples
- Configuration with 4 sensors
  - bare electrode, PAMAM/antigen Leish electrode, PAMAM/antigen T. Cruzi electrode, PAMAM/PVS electrode
  - capacitance spectrum on 58 frequencies, 2 each (real & imaginary): 116 data attributes for each sensor
  - 464 attributes in total describing each sample
  - data normalization: 0 average, 1 standard deviation

Sammon's Mapping: four sensors



- Buffer Tris-Hcl 5 mM
- Negative + buffer
- Leishmania + buffer
- Cruzi + buffer
- Serum A Negative
- Serum B w/ Leishmania
- Serum C w/ Cruzi
- Mixture + buffer

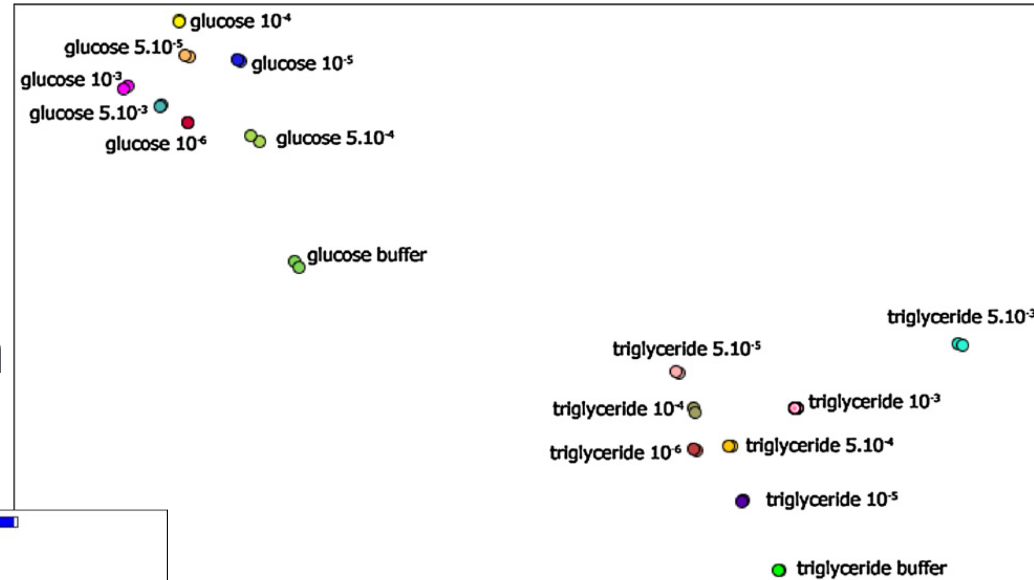
Perinotto et al., *Anal. Chem.* 2010

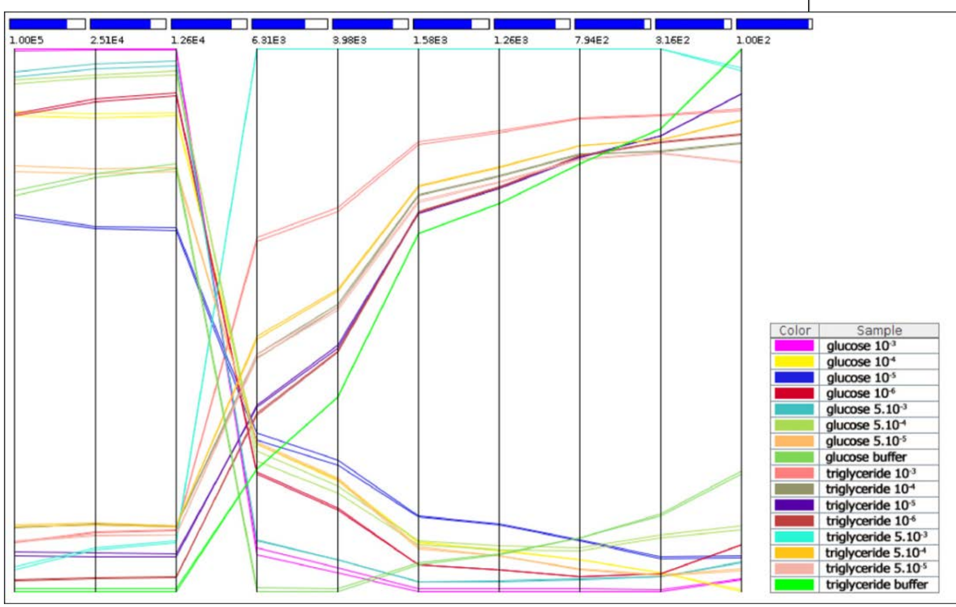
Paulovich et al., *Anal. Bioanal. Chem. 2011*

Biosensor data analysis

Collaborative work with material scientists

finding good sensor configurations: segregation tasks on data





Moraes et. al. Detection of glucose and triglycerides using information visualization methods to process impedance spectroscopy data, ***Sensors & Actuators B***, 2012

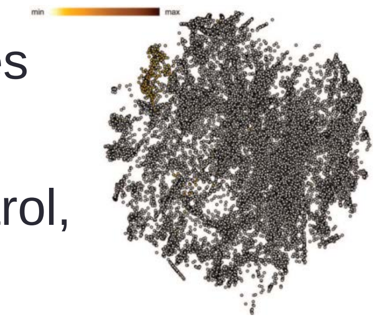
Data analysis: why visualization

- Exploratory scenario
- Flexibility
- Rapid feedback
- User knowledge input

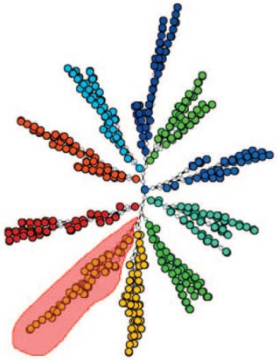
- Multidisciplinary & applied
- Lots of room for novel contributions, both in applications and in fundamental aspects of CS

Similarity based Techniques

- Projections
  - variations on MDS, dimension reduction, or other approaches
  - data mapped to low-dimensional visual space
  - preserving distances vs neighborhoods, global vs. local control, segregation

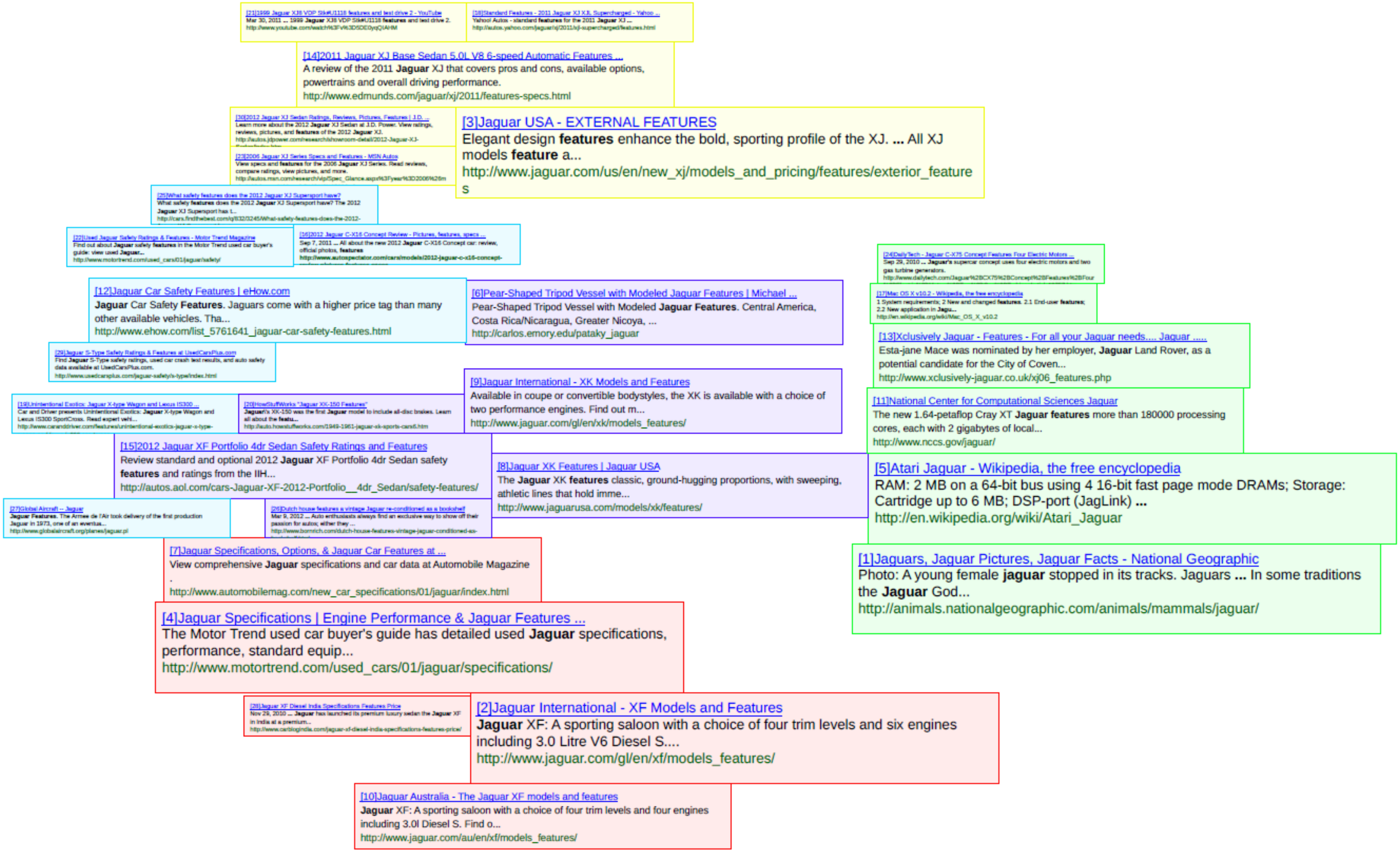


- fully interactive manipulation, dynamically adapting to user feedback
- massive data, sparse high-dimensional data, streaming data



- Tree-based
  - hierarchy of similarity relations
  - variations on tree layouts

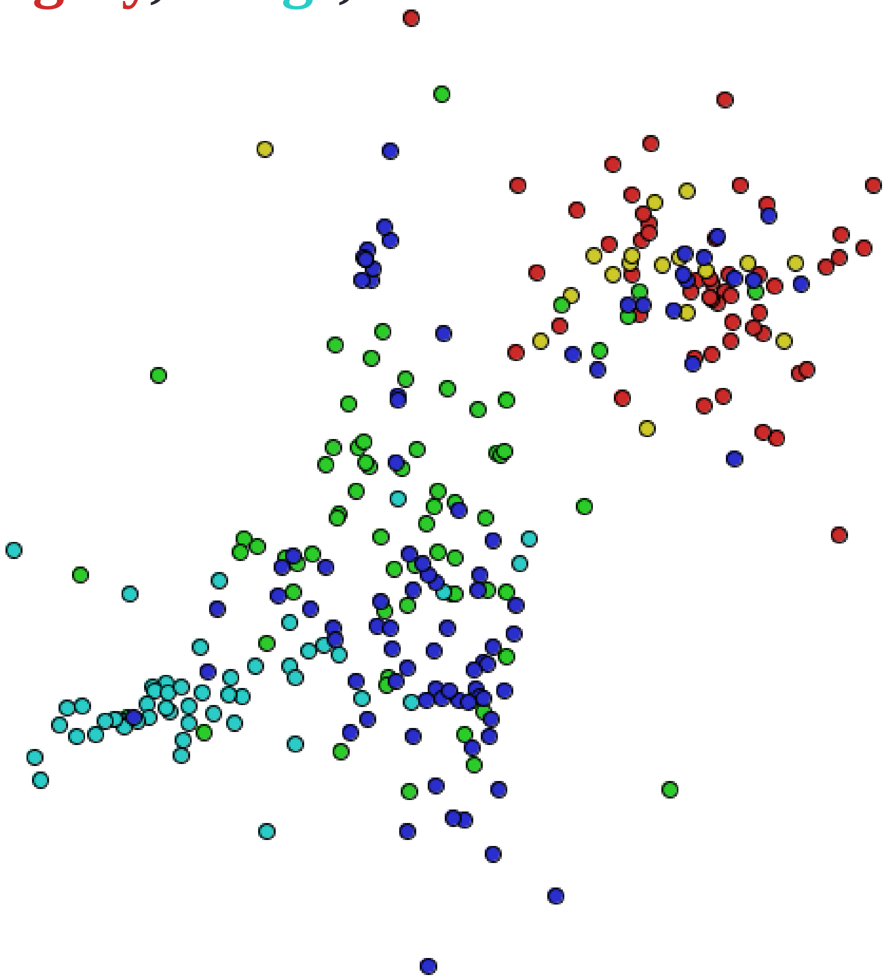
Application: text, web search



Gomez-Nieto et al. Similarity Preserving Snippet-Based Visualization of Web Search Results. *IEEE Trans. Visualization &Computer Graphics, 2014*

Application: text, web search

Ex: Patents surgery, drugs, molecular bio



Soriano et al. 2016 A Visual Analytics System for Time-varying Multidimensional Ionospheric Scintillation Data

- Ionospheric scintillation: phenomenon affects GPS measurements
- Regions in Brazil located around the magnetic equator are severely affected: applications that rely on GPS technology and require full availability and good accuracy face significant and potentially damaging issues
- Collaboration FCT-UNESP PP

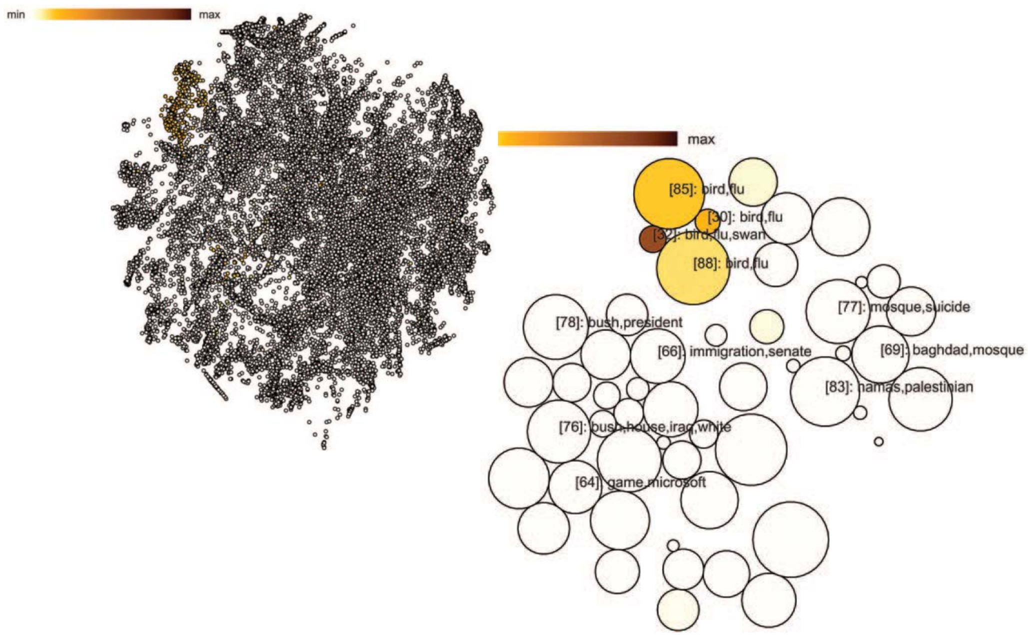
Challenges

- Data Issues
  - Sheer volume
  - Data transformation/formatting/structuring
  - Diversity of data types
  - Spurious correlations
  - Data ownership, ethical issues

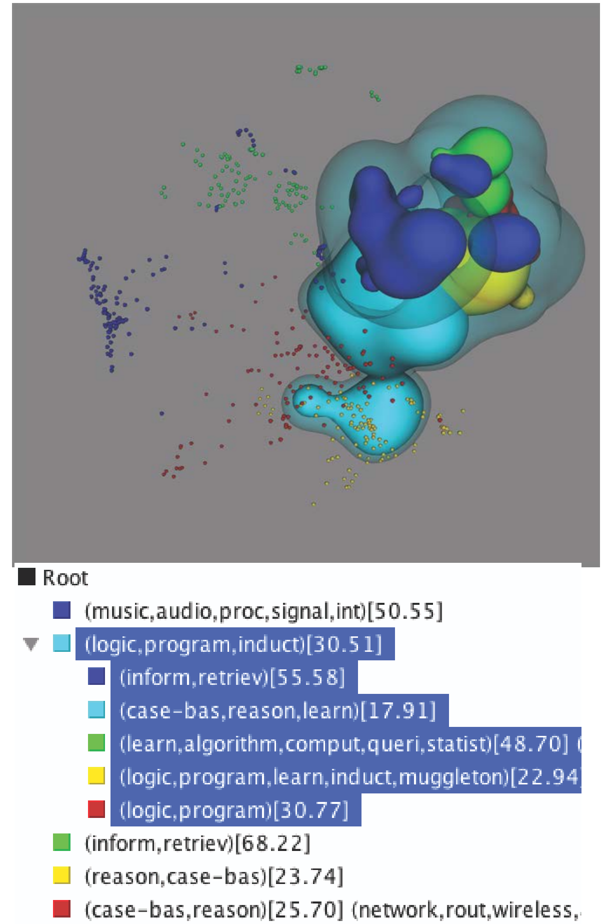
- User issues
  - Inespecificity of questions
  - Interpretation, training

- Visual mapping issues
  - Choice of representation
  - Mapping errors & model-vis correspondence
  - Interactivity & user interface
  - Evaluation (quality & effectivness)

Challenges: clutter, interaction



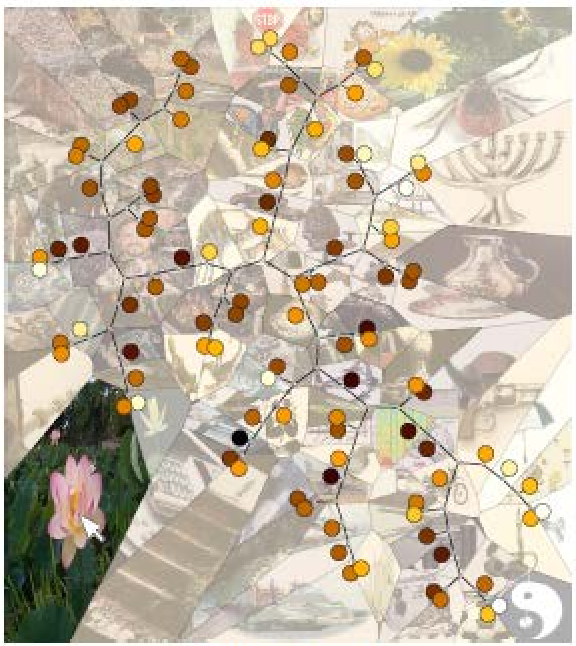
Paulovich and Minghim, HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections, ***IEEE Trans. Visualization & Computer Graphics***, *2008*



Poco; Etedmapour, Paulovich, Long, Rosenthal, Oliveira, Linsen, Minghim. A framework for exploring multidimensional data with 3D projections, ***Computer Graphics Forum***, Eurovis 2011.

Challenges: multiscale

- Caltech data set: 9,144 images, 121 attributes, 101 classes

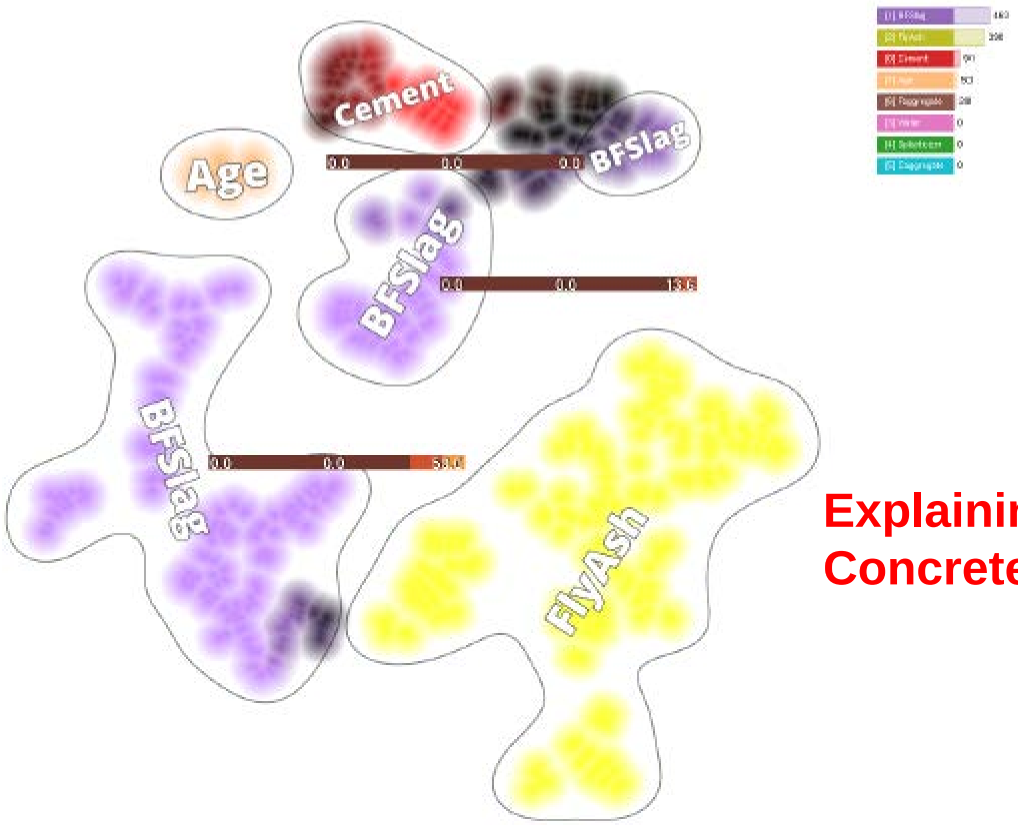


(a) Global view.

(b) Group flowers.

Source: RRO Silva, Visualizing Multidimensional Data Similarities Improvements and Applications. PhD Thesis, USP/University of Gröeningen, 2016

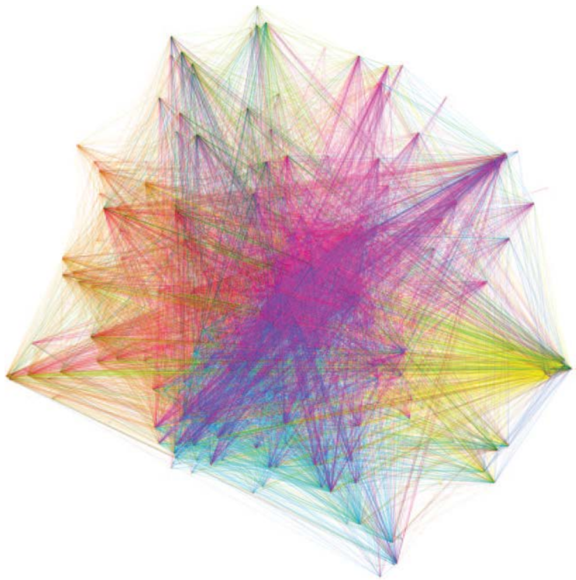
Challenges: interpretation



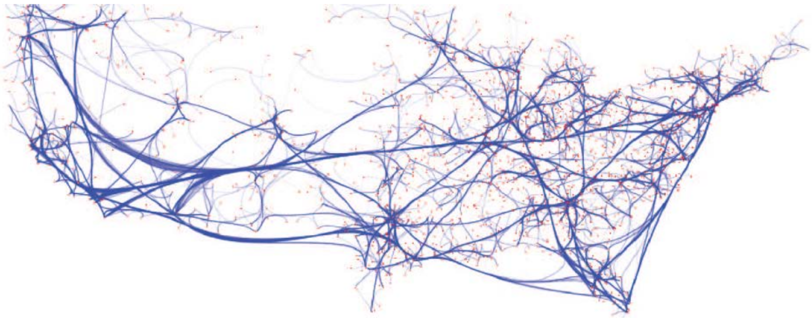
Explaining groups in similarity maps
Concrete data set: 1,030 samples

Source: RRO Silva, Visualizing Multidimensional Data Similarities Improvements and Applications. PhD Thesis, USP & U. Gröeningen 2016.

Networks: even worse



Ersoy, Hurter, Paulovich, Cantareira, Telea, Skeleton-based edge bundling for graph visualization. *IEEE Trans. Visualization and Computer Graphics,* Infovis 2011



Links to sources of data visualization tools & data

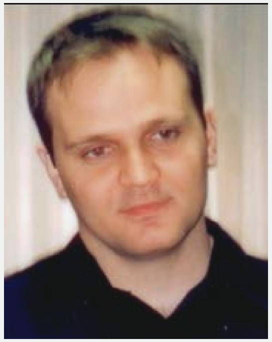
- HDR (ONU):
  - (data) http://hdr.undp.org/en/composite/GII
  - (vis) http://hdr.undp.org/en/data-explorer/

- D3:
  - https://d3js.org/
  - (gallery) https://github.com/mbostock/d3/wiki/Gallery/

VICG - Visualization & Imaging faculty

http://vicg.icmc.usp.br

Fernando Paulovich



Maria Cristina

Luis Gustavo Nonato

Rosane Minghim

João E. S. Batista

Moacir Ponti

Further Readings

- Oliveira, MCF & Levkowitz, H. From visualization to visual data mining. *IEEE Computer Graphics & Applications* 9(3), 378-394, 2003.

- Keim, DA et al. Mastering the information age: solving problems with visual analytics. 2010. http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf

- Alencar, AB; Oliveira, MCF**;** Paulovich, FV. Seeing beyond reading: a survey on visual text analytics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 476-492, 2012.

- Rodrigues, JF; Paulovich, FV; Oliveira, MCF; Oliveira, ON. On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis. *Nanomedicine*, 11, p. 959-982, 2016.

Instituto de Ciências Matemáticas e de Computação

| Universidade de São Paulo |

Thanks!

(some slides by O.N. Oliveira Jr. & Rosane Minghim)

Maria Cristina F. Oliveira
ICMC-USP
cristina@icmc.usp.br
http://vicg.icmc.usp.br