

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 <u>cristina@icmc.usp.br</u>

Outline

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

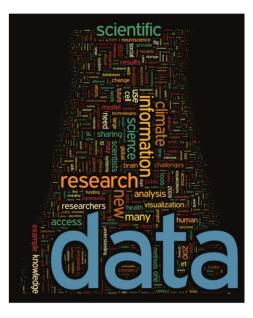
Data Science: what does it take?

- 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')

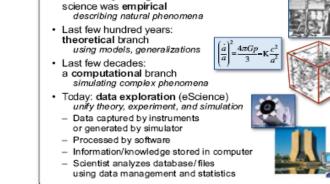
e-Science: The Fourth Paradigm





The FOURTH PARADIGM DATA-INTENSIVE SCIENTIFIC DISCOVERY

DITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

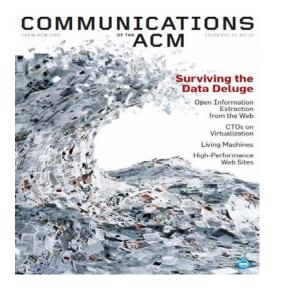


Thousand years ago:

Science Paradigms

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 crossreference 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
 Potential Productivity Gains - the power of 1%

How advantageous:

	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 <u>big data</u> (/<u>static/special-report-big-data)</u>, covering technologies that support and make sense of the growing mountains of data, and several of its applications.

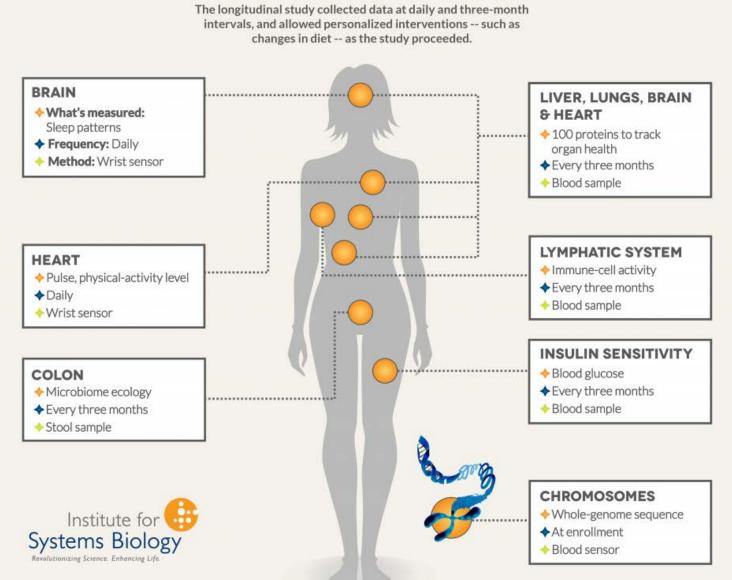
It's no surprise that keeping people healthy is costing more money. From the price of medications and the cost of hospital stays to doctors' fees and medical tests, health-care costs around the world are skyrocketing. The World Health Organization attributes much of this to wasteful spending on such 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)

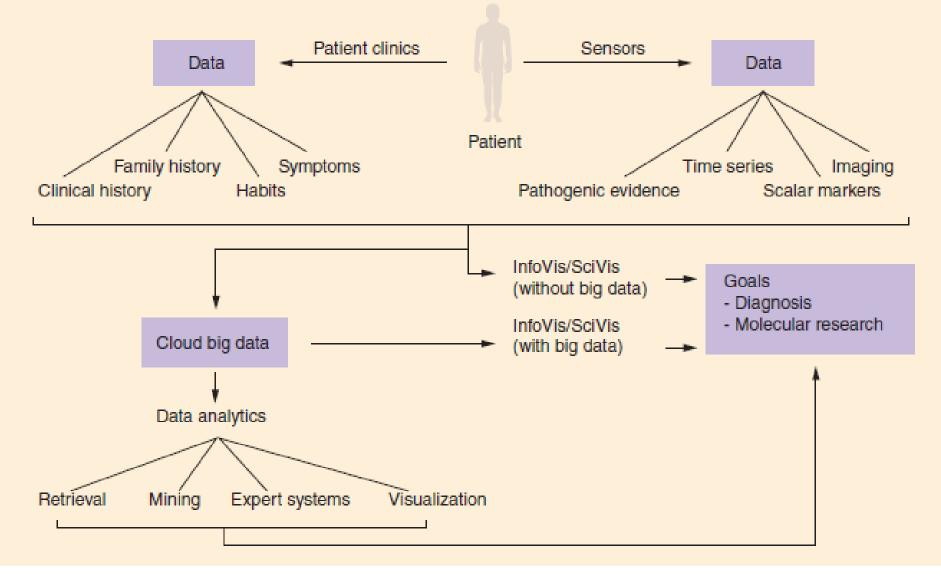
AN EXAMINED LIFE



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



Computer-aided diagnosis



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

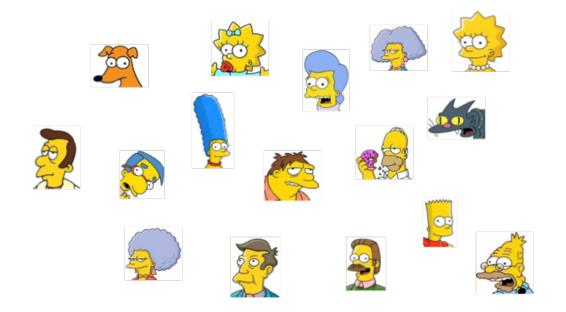
Data is...

- 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

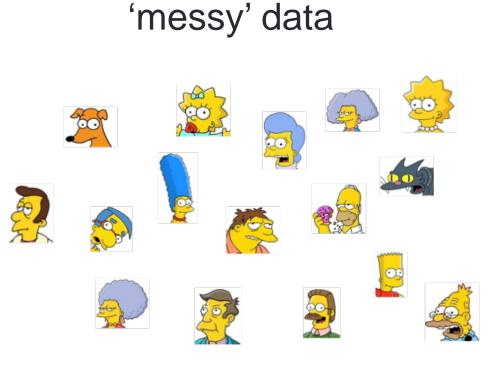
People trying to make sense of data

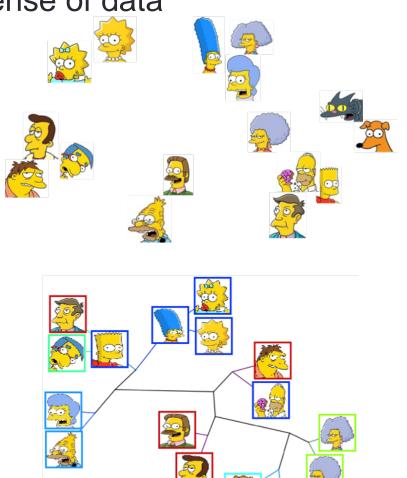
'messy' data



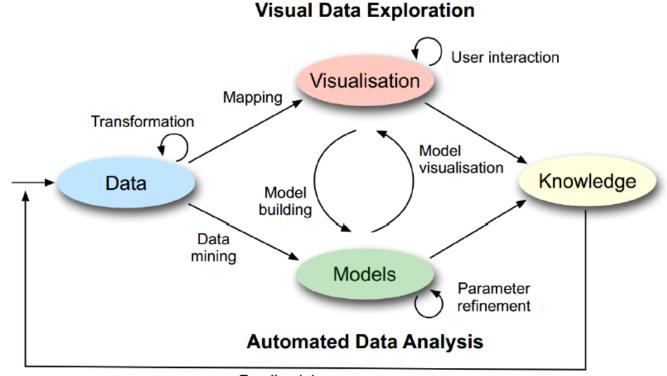
What does your data tell???

Visualization can help making sense of data





Visual Analytics process



Feedback loop

Source: Keim et al. 2010

Multidimensional data: representation

	Ö															10.000									
-	15	••														5	5	12	15	2	7	5	0	12	
	6	2	-													-	12	5	0	12	12	12	12	12	1
	15	1	2	-												-	0	1	05	10	15	12	8	12	
<u></u>	15	1	2	1	2											2	0	12	01	12	9	0	12	10	
	15	2	1	2	2												12	8	05	12	12	12	8	12	
é	15	3	10	8	15	15	<u>9</u>									ó	10	12	0	11	10	2	7	12	L
5	15	10	7	10	10	7	15									5	5	6	8	12	12	15	12	6	-
2	15	15	10	15	15	10	15	16	9							2	7	12	05	0	12	12	10	17	-
	15	15	10	10	10	10	15	16	16	5						-	2	10	05	15	12	1	12	10	
5	15	15	15	15	15	15	15	15		15	2					5	12	12	7	12	0	12	0	12	
											25	i 🚓				2	6	12	05	17	12	10	12	12	└─
	15	10	10	10	10	10	15	15	15	15	15	27				2	12	10	2	12	1	12	12	11	
<u> </u>	15	10	10	10	10	10	15	15	15	1	15	15	9			<u>a</u>	1	12	05	12	12	16	2	12	
	15	10	15	12	15	15	10	15	15	15	15	15	15	2		2	10	0	12	12	9	12	0	10	1
	15	10	10	10	10	10	15	15	15	15	15	2	8	15	-	1	0	12	1	12	12	5	1	7	1
4	15	10	10	10	10	10	15	1	15	15	15	15	15	15	15	4	8	2	11	10	7	12	5	12	1

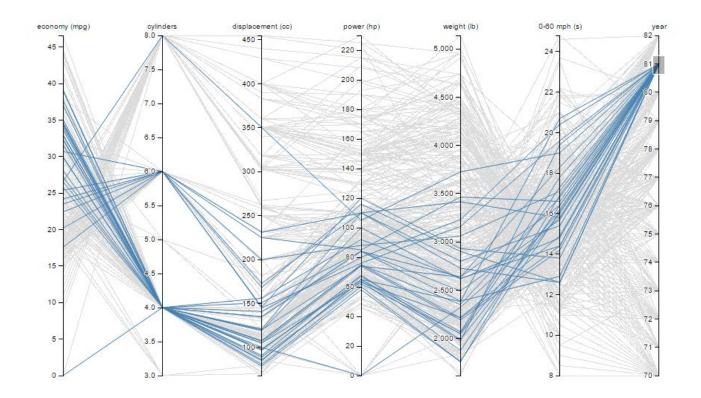
pairwise distances

and/or

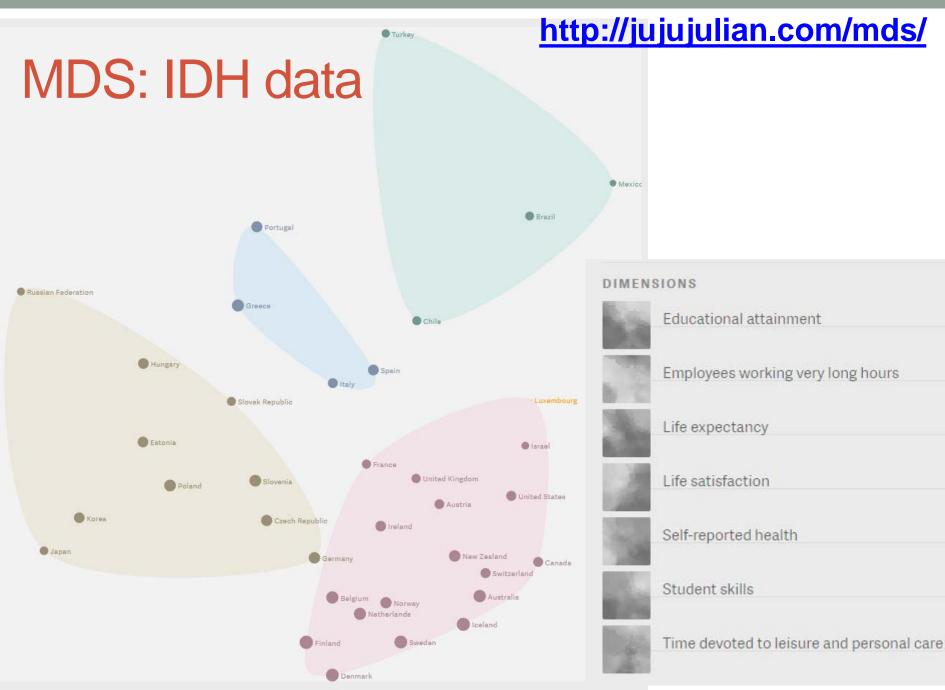
dimensional embedding (feature space)

Parallel Coordinates

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







Multidimensional projection

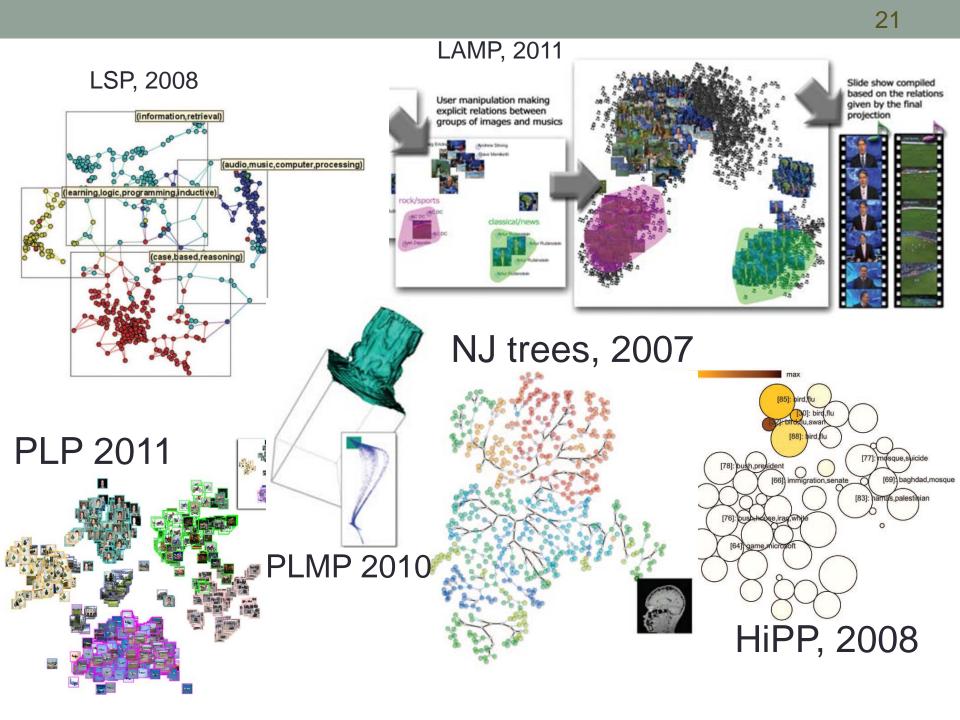
$$X \in \mathbb{R}^{m} \quad f \quad Y \in \mathbb{R}^{k=\{1,2,3\}}$$

$$\begin{split} \bullet & \delta : x_i, x_j \to R, x_i, x_j \in X \\ \bullet & d : y_i, y_j \to R, y_i, y_j \in Y \\ \bullet & f : X \to Y, |\delta(x_i, x_j) - d(f(x_i), f(x_j))| \approx 0, \ \forall \ x_i, x_j \in X \end{split}$$

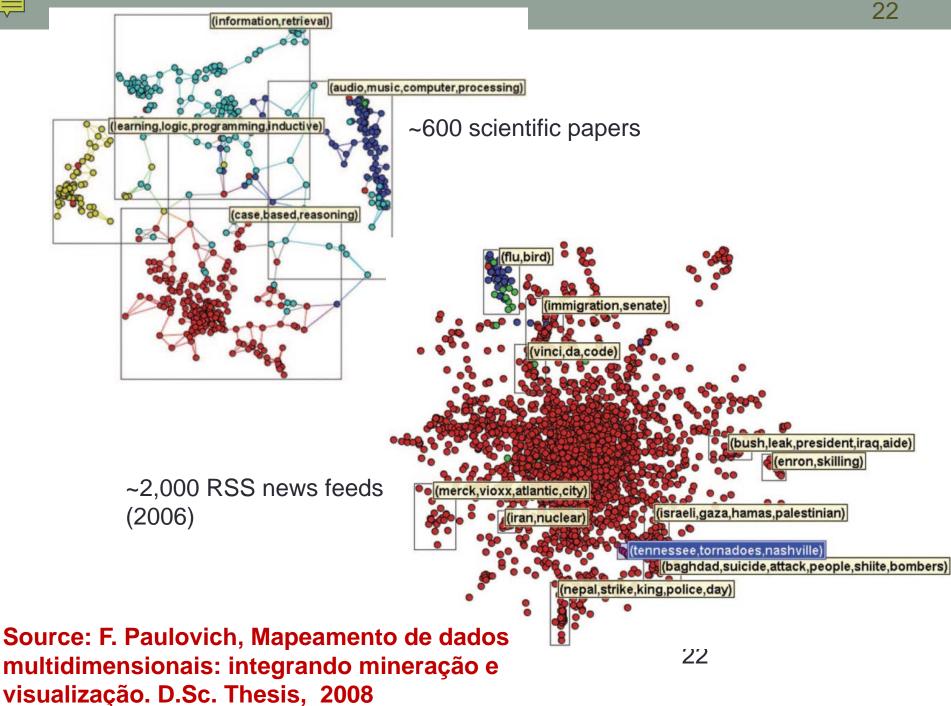
$$E = \frac{\sum_{ij} (\delta(x_{i}, x_{j}) - d(y_{i}, y_{j}))^{2}}{\sum_{ij} \delta(x_{i}, x_{j})^{2}}$$
18

Multidimensional projection

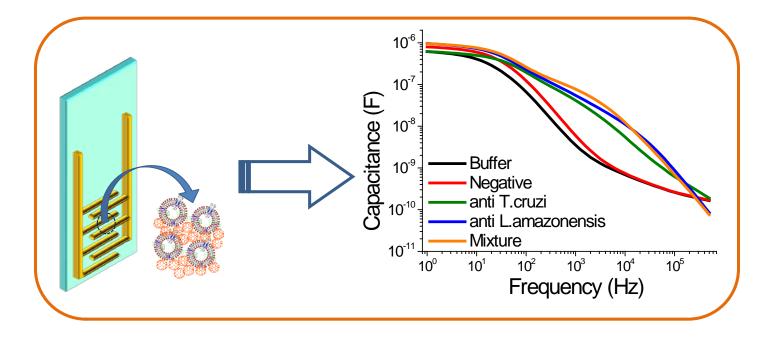
- old idea, old & new techniques...
- current techniques must comply with requirements imposed by interactive applications:
 - speed (low computation cost)
 - capability to handle very large & massive data
 - interactivity (allow user intervention)







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

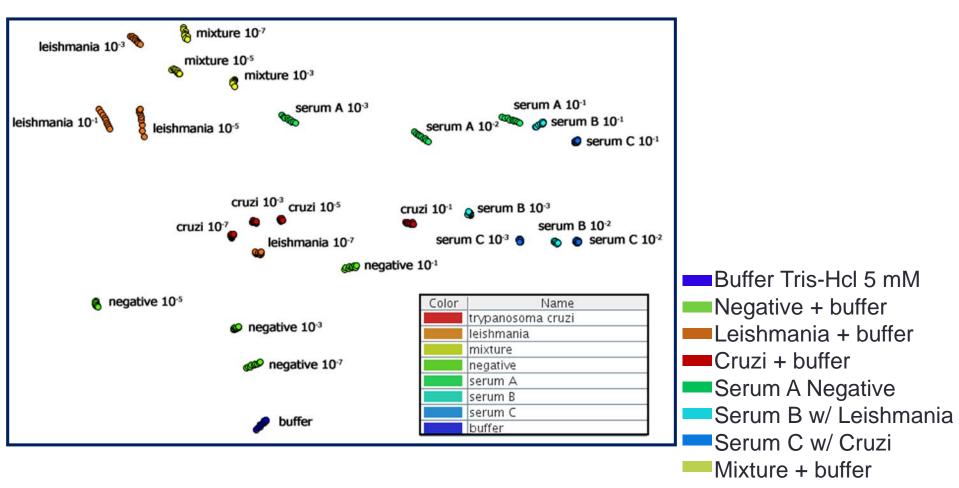
 understanding/explaining why it is optimal

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

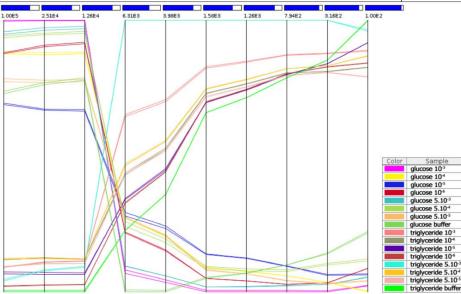


Perinotto et al., *Anal. Chem.* 2010 Paulovich et al., *Anal. Bioanal. Chem.* 2011

Biosensor data analysis

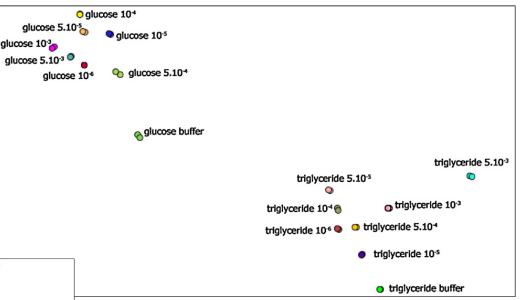
Collaborative work with material scientists

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



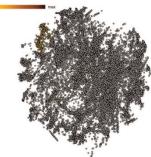


Data analysis: why visualization

- 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



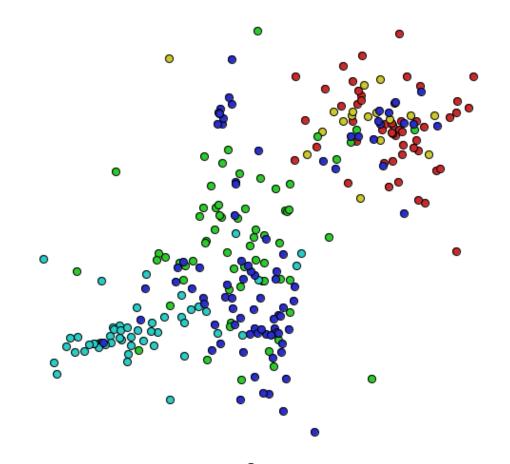
Application: text, web search

			(21)1999 Mar 30, 3 http://www.	Jaguer XJS VDP SkeU1115 feat 2011 -, 1999 Jaguar XJS VDP St wcyoulube.com/waich/U3FV63D	ures and lest drive 2 - YouTube #U1118 features and test drive 2. DEDyqQLAHM	Yahool Autos - standard	2011 Jaguer XJ XJ, Supercharged - Vehoo features for the 2013 Jaguer XJ jeguer/sj2011/sij-supercharged Nestures.html								
				A review of t powertrains	uar XJ Base Sedan 5.0 he 2011 Jaguar XJ that and overall driving perfo dmunds.com/jaguar/xj/2	covers pros and ormance.	l cons, available options,								
	12	S)What safety features does	reviews, pict http://wutos.j [22]2005 Jag View specs a compare rati http://wutos.m	gain XI Secton Ratings, Reviews, down the 2012 Seguer XO Sector trues, and teamso of the 2012 A depower combineses that was and and teatowa for the 2006 Seguer and Celativas for the 2006 Seguer man.combineses.th/dpDpec_Glan trues XI Seguergort have? and XI Seguergort have?	guer XJ. setall/2012-Jaguer-X3- s - MSN Autos XJ Sertika: Bead neviews, ce.expr/M3PywarM3D2000%25m	Elegant des models feat	SA - EXTERNAL FEATURES ign features enhance the bold ure a aguar.com/us/en/new_xj/mode								
Find out about guide: view up	Ja Internet Society Ratings A out Jaguar safety feat used Jaguar	aguar XJ Supersport has t	L. gR32/3245/Wh Magazine	hal-safety-features-does-the-2012 (16)2012 Jaguer C-X1 Sep 7, 2011 — All abo official photos. feature	5 Concept Review - Pickares, features, 4 the new 2012 Jeguar C-X15 Concep	t car: review,				2420ah/turl-1_bijaar C-2723 Concret Finitizes Tata Electric Mitters Tag 25, 7200 - Segarah separat aliceget asse floor deally model and ten get before generatizes					
Jag oth http	[12]Jaguar Car Safety Features eHow.com Jaguar Car Safety Features. Jaguars come with a higher price tag than many other available vehicles. Tha http://www.ehow.com/list_5761641_jaguar-car-safety-features.html [20bgar 5-the Setty Hetge 4 Features (UNIDerPhase)					Pear-Shaped Costa Rica/Nic	d Tripod Vessel with Modeled Jaguar Featu Tripod Vessel with Modeled Jaguar Featur caragua, Greater Nicoya, nory.edu/pataky_jaguar			III/Mac 00 X vi02- Watperfs, the five encyclopedia 15 phore mediaments: 2 Www.udl disreget features: 21 Drd-user fleatures: 12 the englisher in Sadem: 13 the englisher in Sadem: 14 the the englisher in Sadem: 15 the in an Mace was nominated by her englisher in Sadem: 14 the the the englisher in Sadem:					
Find Jaguar S-Type data available at Use	e safety ratings, used	I car crash lest results, an	d auto safety			[9].]aguar Inter	national - XK Models and Features			Detential candidate for the City of Coven http://www.xclusively-jaguar.co.uk/xj06_features.php					
Car and Driver presents Unintentic Lexus (\$300 SportCross, Read ex	Institut Tarlino Januar X. (Jap Wages and Lease 1020) Chine person Uniter the Control Exception Appart Appa Negen and Source person Uniter the Control Appart Appa Negen and Source and Negen and Appa Negen Appa Negen and Source and Negen and Appa Negen Appa Negen and Source and Negen Appa Negen Negen Appa Neg Negen Appa Negen Negen Appa Negen Appa Negen Nege				eguar-sk-sports-cars6.htm	Available in co two performan	upe or convertible bodystyles, the XK is av ce engines. Find out m uar.com/gl/en/xk/models_features/	ailable with a choice of	J11National Center for Computational Sciences Jaquar The new 1.64-petaflop Cray XT Jaguar features more than 180000 processing cores, each with 2 gigabytes of local http://www.nccs.gov/gaguar/						
	152012 Jaguar XF Portfolio 4dr Sedan Safety Ratings and Features Review standard and optional 2012 Jaguar XF Portfolio 4dr Sedan safety features and ratings from the IIH http://autos.aol.com/cars-Jaguar-XF-2012-Portfolio4dr_Sedan/safety-feat				lio 4dr Sedan safety	res/ The Jag athletic li	r XK Features Jaquar USA uar XK features classic, ground-hugging p ines that hold imme	roportions, with sweeping,	[5]Atari Jaguar - Wikipedia, the free encyclopedia RAM: 2 MB on a 64-bit bus using 4 16-bit fast page mode DRAMs; Storage: Cartridge up to 6 MB; DSP-port (JagLink) http://en.wikipedia.org/wiki/Atari_Jaguar						
27)Global Aircraft Jaguar Jaguar Features. The Armee de l'Air Jaguar in 1973, one of an eventua http://www.globalaircraft.org/planes/ja		(20D)utch house leafures a virtage Jaguar re-conditioned as a hookoh Mer 8, 2022				f their	w.jaguarusa.com/models/xk/features/								
	[7]Jaguar Specifications, Options, & Jaguar Car Features at View comprehensive Jaguar specifications and car data at Autor http://www.automobilemag.com/new_car_specifications/01/jaguar [4]Jaguar Specifications Engine Performance								[1]Jaguars, Jaguar Pictures, Jaguar Facts - National Geographic Photo: A young female jaguar stopped in its tracks. Jaguars In some traditions the Jaguar God http://animals.nationalgeographic.com/animals/mammals/jaguar/						
	1	The Motor performanc	Trend ce, star	used car buye ndard equip		ailed used J	aguar specifications,								
2013 Jaguer 2017 Device Finds - Specifications Findures Price 1014 Jaguer 2017 Device Finds - Specifications Findures Price 1016 State 4 appendixm. 1017 Prices carbinging and - Specifications - Features prices 1017 Prices carbinging and - Specifications - Features prices						Jaguar X including	International - XF Models and F: A sporting saloon with a cho 3.0 Litre V6 Diesel S wjaguar.com/gl/en/xf/models_f	ice of four trim level	s and	d six engines					
10Jaguar Australia - Jaguar XF: A sporting including 3.01 Diesel http://www.jaguar.com						saloon with a cho Find o	pice of four trim levels and four engines								

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

Application: text, web search

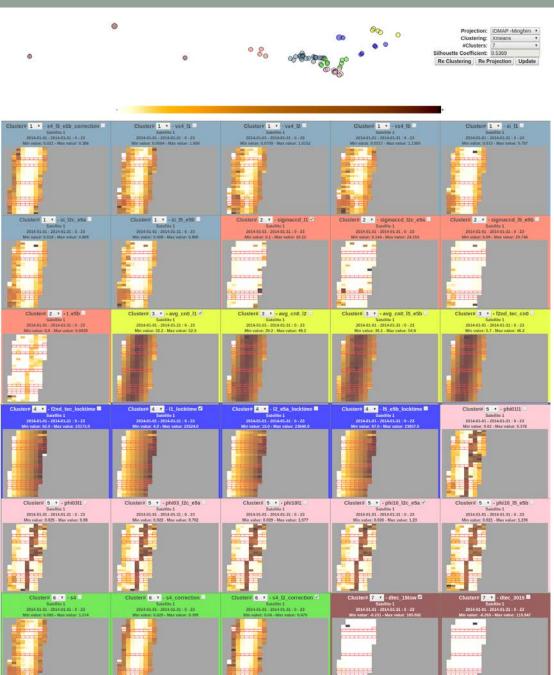
Ex: Patents surgery, drugs, molecular bio



35

Soriano et al. 2016 A Visual Analytics System for Timevarying Multidimensional Ionospheric Scintillation Data

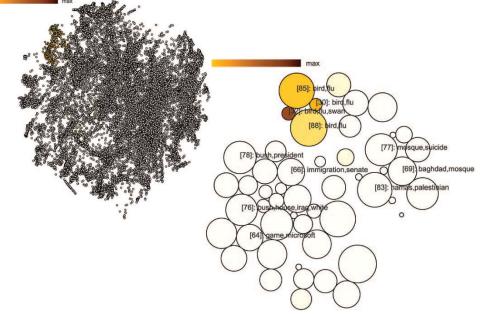
- Ionospheric scintillation: phenomenon affects GPS measurements
- Regions in Brazil located around the magnetic equator are severely affected: applications that rely on GPS technology and require full availability and good accuracy face significant and potentially damaging issues
- Collaboration FCT-UNESP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



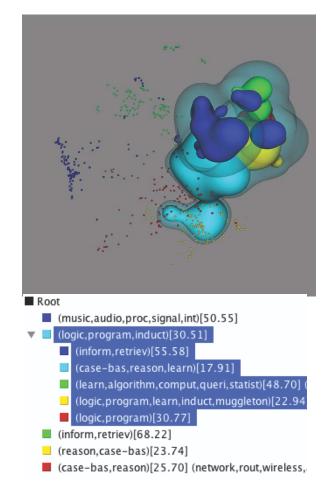
Challenges

- 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.

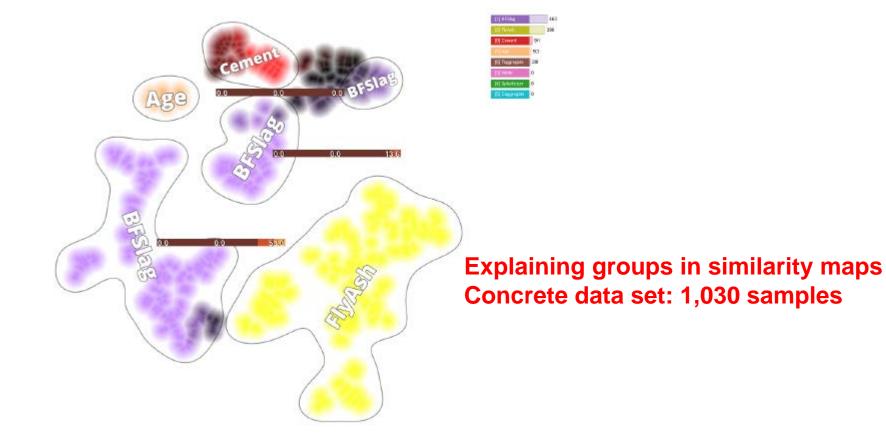


(b) Group flowers.

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

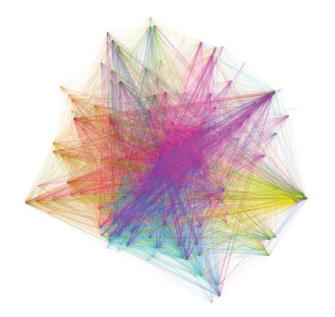


Challenges: interpretation



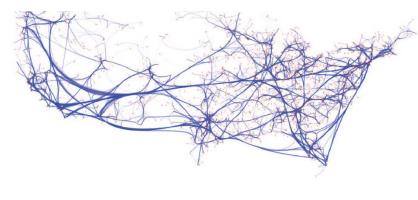
Source: RRO Silva, Visualizing Multidimensional Data Similarities Improvements and Applications. PhD Thesis, USP & U. Gröeningen 2016.

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

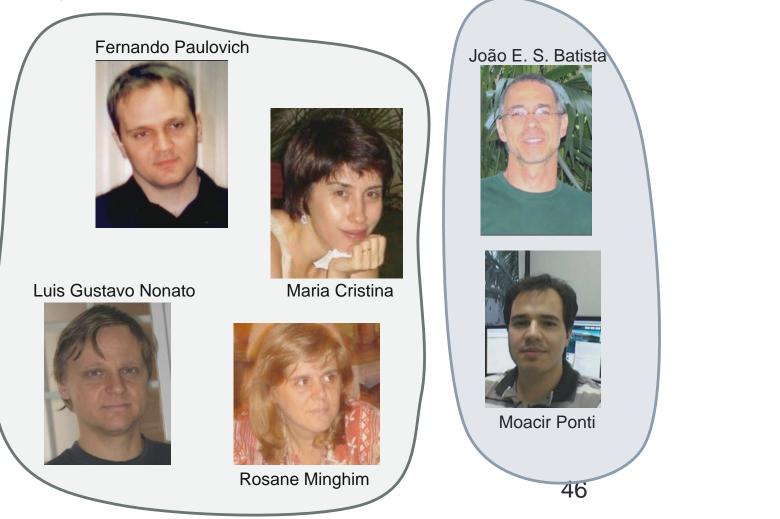
• HDR (ONU):

- (data) http://hdr.undp.org/en/composite/GII
- (vis) <u>http://hdr.undp.org/en/data-explorer/</u>

• <u>D3:</u>

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

VICG - Visualization & Imaging faculty <u>http://vicg.icmc.usp.br</u>



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/wpcontent/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 computeraided 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 <u>cristina@icmc.usp.br</u> <u>http://vicg.icmc.usp.br</u>