ICMC USP SÃO CARLOS

Instituto de Ciências Matemáticas e de Computação

| Universidade de São Paulo |

SCC 0252/0652 – COMPUTATIONAL VISUALIZATION

Introduction: Data Visualization in the context of Data Science and Big Data

Rosane Minghim
2019-2

Outline

- About...
- Data Science
- Big Data for productivity
- Visualization
- Visualization Techniques
- Looking Forward

3

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: Undestand the data

4

Qualification - keywords

- Ex: Coursera (https://www.coursera.org/)
  - Data Science set of courses by Johns Hopkings U.
  - 9 courses.
    - Intro(concepts + infra – version control and R IDE)
    - R Programming
    - Data collecting, cleaning and sharing
    - Exploratory data analysis – visualization and such
      - Buzz words – visual analytics
    - Statistical Inference
    - Regression Models
    - Reproducible Research
    - Practical Machine Learning
    - Data products – making results usable

5

Big Data
is this a real thing?

YES it is !

6

Some numbers

- Big Data: Growing 40X to $32.4 billion by 2017
- 4300% increase in data by 2020
- Internet of Things growth 1-b to 26-billion units by 2020
- 2014: Increase of 125% in companies with data driven projects.
- 69% of unstructured data never makes it to decision making.

7

Some more numbers

- In 2020: 7B people, 30Billion 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

8

Applications (Data Analytics – large scale)

- Cities: Transportation/Integration, Crime Prevention, Citizen Information, Currency, Energy, Utilities, Waste, Parking, Hospitality

(Open Data)

- Health: Health Manager, Cost Optimization, Death prevention
- Internet-of-things: Customer, Devices, Sensors, Robots. Ex. Environment monitoring sensors, factories, phones, energy.
- Aerospace & flying: Reports: structural changes ($$$) and customer needs (on-time flights & changes in bagage handling: 80 million US)
- Commerce: marketing wrong, social network analytics
- Agriculture:
- Government: Costs, Well being, Logistics, Tax,

9

How big – just a sample
From IoT

46 million smart meters in the U.S alone
1.1 billion data points (.5TB) per day

A single consumer packaged good manufacturing machine generates 13B data samples per day

A large offshore field produces 0.75TB of data weekly
A large refinery generates 1TB of raw data per day

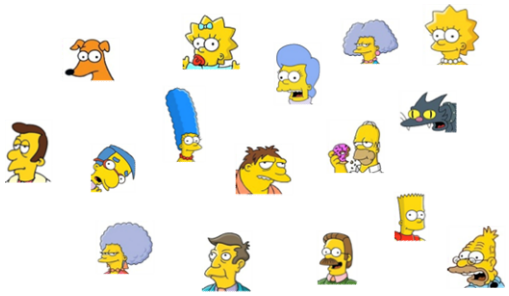
10TB of data for every 30 minutes of flight
With >25,000 flights per day, petabytes daily

10

Visualization Problem

- People trying to make sense of data

´messy´ data



11

Data is...

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

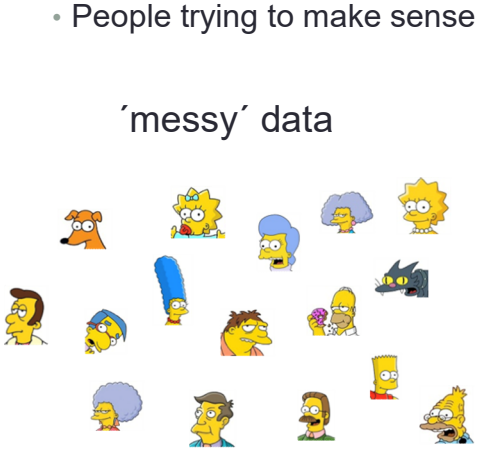
- Studying natural & artificial systems and phenomena implies in handling lots of data...

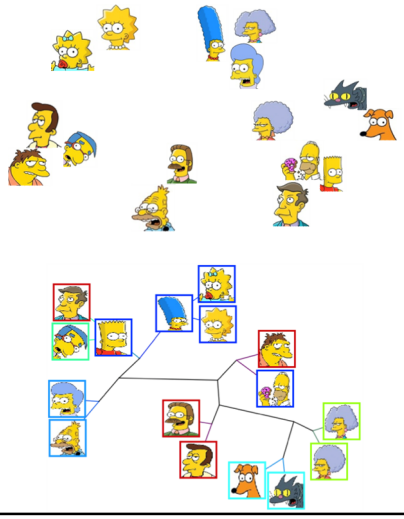
12

What does your data tell???

- People trying to make sense of data

´messy´ data

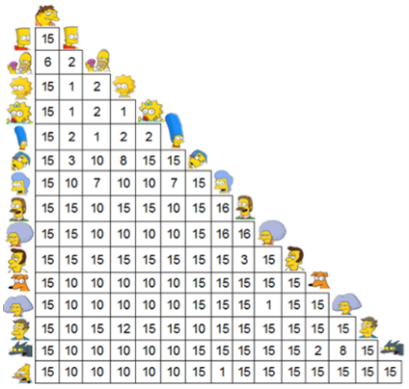




13

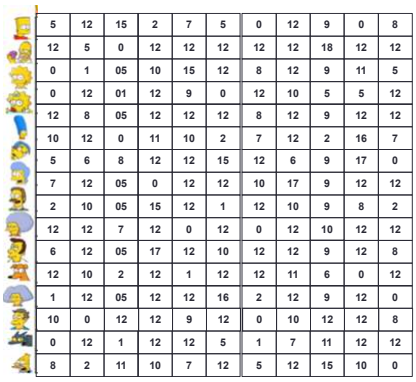
Techniques

- Multidimensional visualization: data organization



pairwise distances

and/or



dimensional embedding (feature space)

14

Example: On studies on ecology and environment

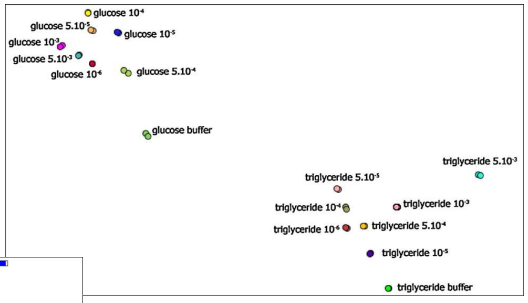
- Collaborative work with biologists
- D.Sc. project: Visual exploration of feature spaces to support green algae taxonomic classification
- Classification based on features from images & other sources
- Time-varying images, feature extraction, representation and analysis

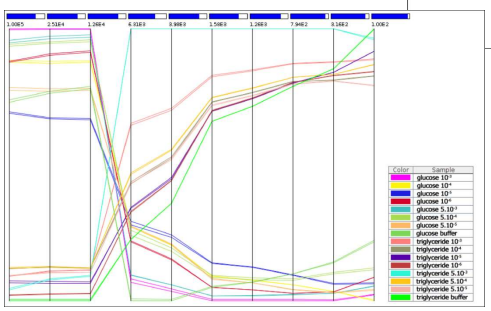
15

Example: Data from nanotech sensors & biosensors

Collaborative work with physicists

finding good sensor configurations: segregation tasks on data



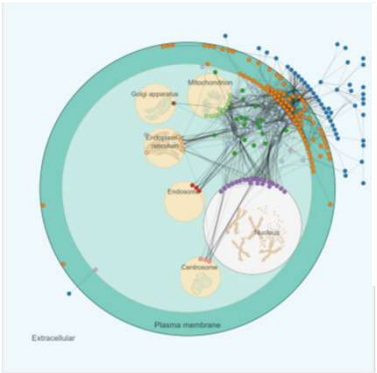


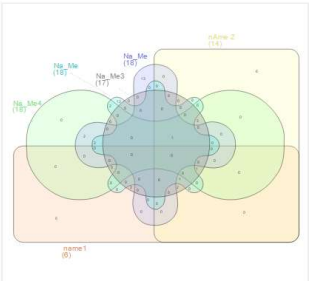
Moraes et. al, Detection of glucose and triglycerides using information visualization methods to process impedance spectroscopy data, ***Sensors & Actuators B***, 2012

16

Example: Proteomics and Cancer

- Masters/PhD project: Visual comparison of protein candidates.





Heberle, H. ; Meirelles, G. V. ; Silva, F. R. ; Telles, G. P. ; MINGHIM, R. . InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. **BMC Bioinformatics**, v. 16, p. (169), 2015.

Kawahara, R., Meirelles, G., Heberle, H., Domingues, R., Granato, D., Yokoo, S., Canevarolo, R., Winck, F., Ribeiro, A. C., Brand~ao, T. B., Filgueiras, P., Cruz, K., Barbuto, J. A., Poppi, R., Minghim, R., Telles, G., Fonseca, F. P., Fox, J., Santos-Silva, A., Coletta, R., Sherman, N., and Leme, A. P. Integrative analysis to select cancer candidate biomarkers to targeted validation. **Oncotarget** 6, 41 (2015), 43635-43652.

17

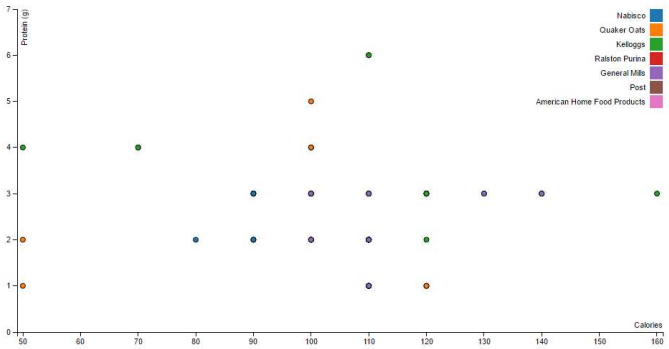
Links to sources of data visualization tools and data

- HDR (ONU):
  - (data) http://hdr.undp.org/en/composite/GII
  - (vis) http://hdr.undp.org/en/data-explorer/
- OECD (Health) : https://www.oecd.org/els/health-systems/health-data.htm
- Google (health public): https://www.google.com/publicdata/directory#!
- D3:
  - https://d3js.org/
  - (gallery) https://github.com/mbostock/d3/wiki/Gallery/

18

Technique: Scatter Plot

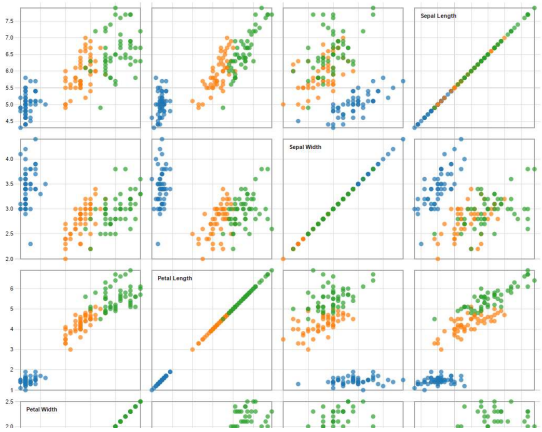
- http://bl.ocks.org/weiglemc/6185069



19

Technique: Scatter Plot Matrix

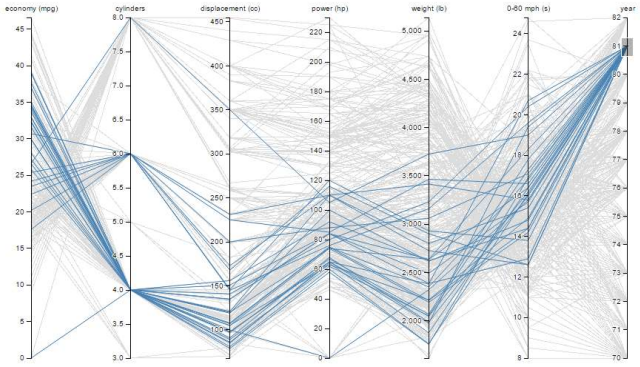
- https://bl.ocks.org/mbostock/4063663



20

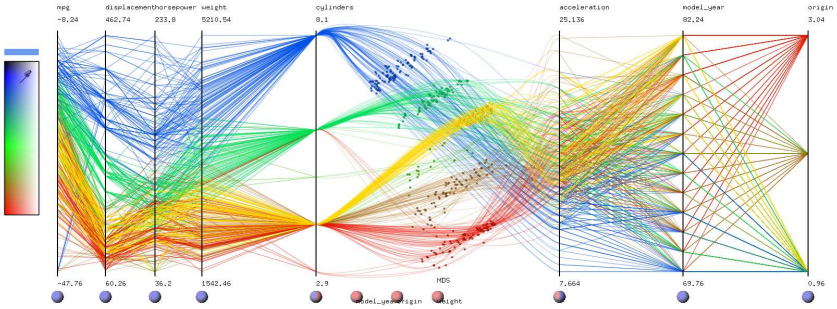
Technique: Parallel Coordinates

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



21

Technique: Scattering Points in Parallel Coordinates

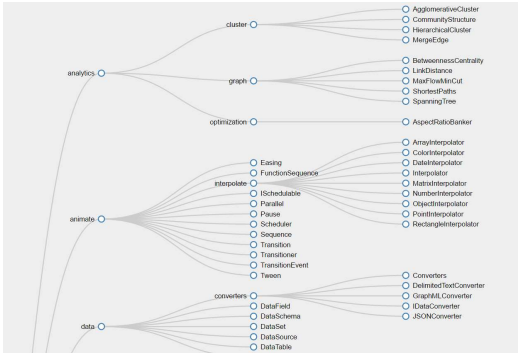


Project: http://vis.pku.edu.cn/wiki/project/hdvis

22

Technique: Tree Visualization

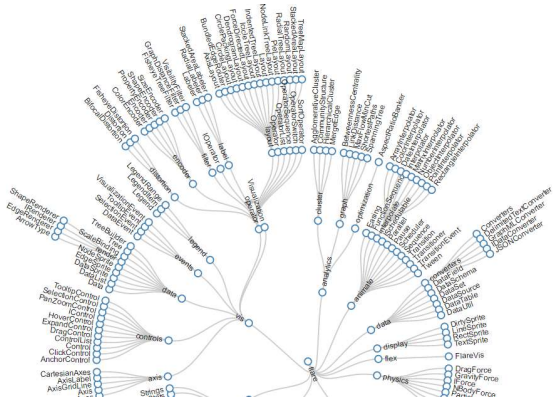
- http://bl.ocks.org/robschmuecker/raw/7880033/
- Drag and Drop, Zoomable, Panning, Collapsible Tree with auto-sizing



23

Technique: Radial Reingold–Tilford Tree

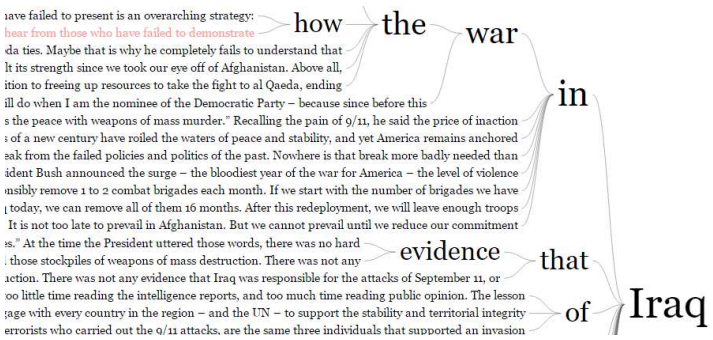
- http://bl.ocks.org/mbostock/4063550/



24

Technique: Word Tree

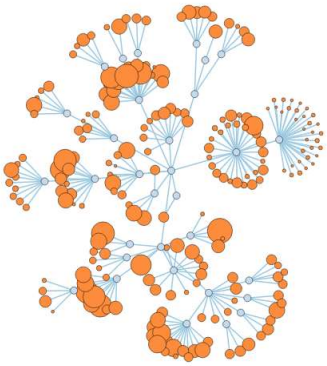
- https://www.jasondavies.com/wordtree/?source=obama-war-speech.txt&prefix=Iraq&reverse=1



25

Technique: Tree - Force Layout

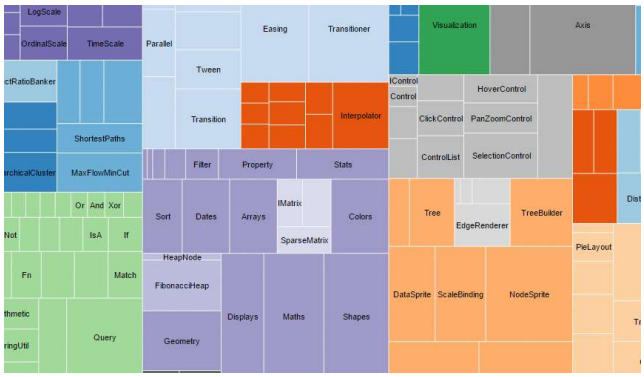
- http://mbostock.github.io/d3/talk/20111116/force-collapsible.html



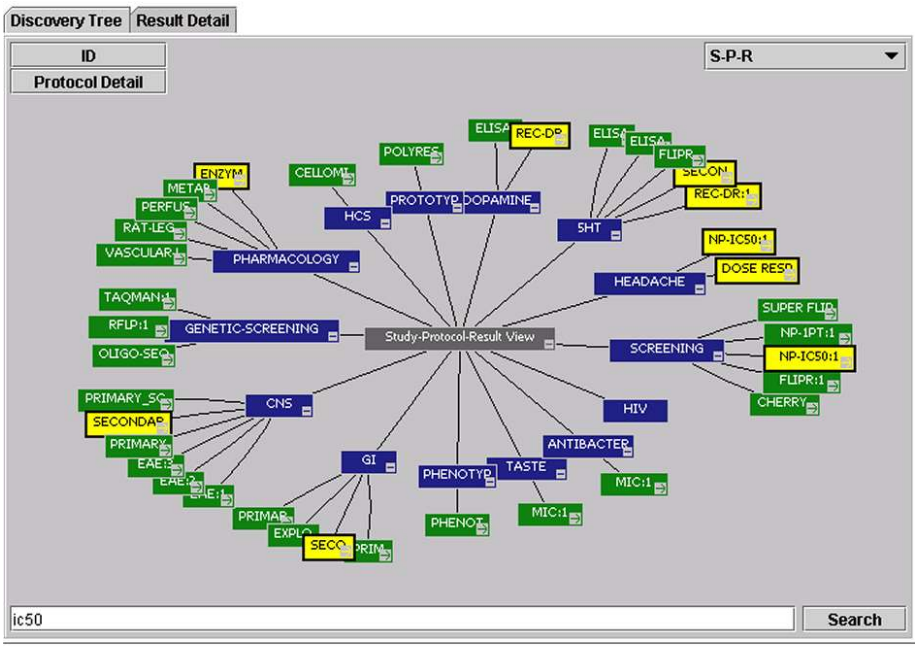
26

Technique: Tree - TreeMap

- http://mbostock.github.io/d3/talk/20111018/treemap.html

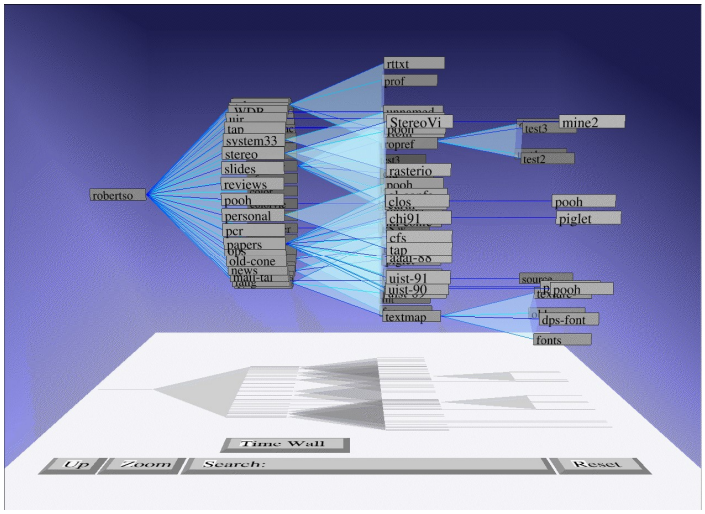


27



28

Cone Tree



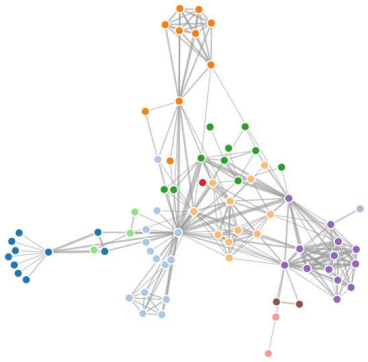
File system structure visualized as a cone tree

Animated 3D visualization of hierarchical data

29

Force-directed Graph Layout

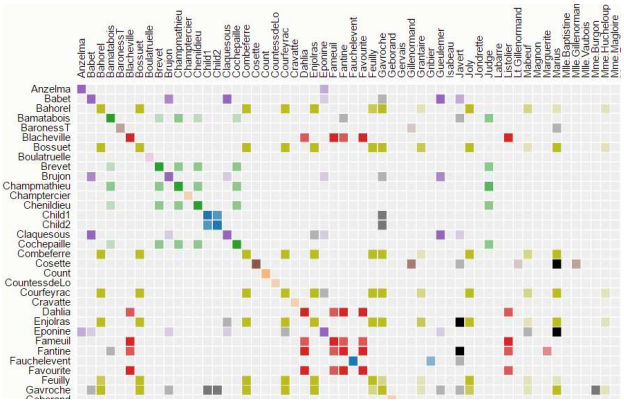
- http://bl.ocks.org/mbostock/4062045



30

Adjacency Matrix Graph Layout

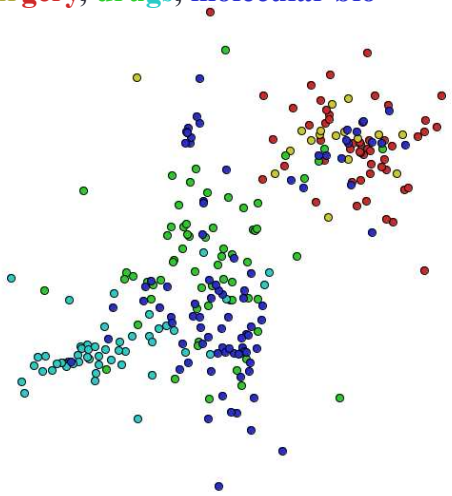
- https://bost.ocks.org/mike/miserables/



31

Projection Techniques:
Mapping data set on the plane, allowing direct exploration

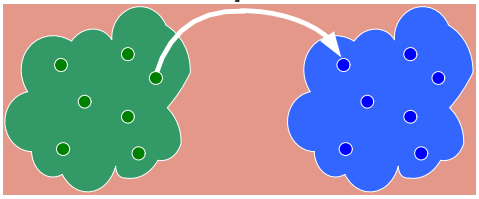
Ex: Patents **surgery**, **drugs**, **molecular bio**



32

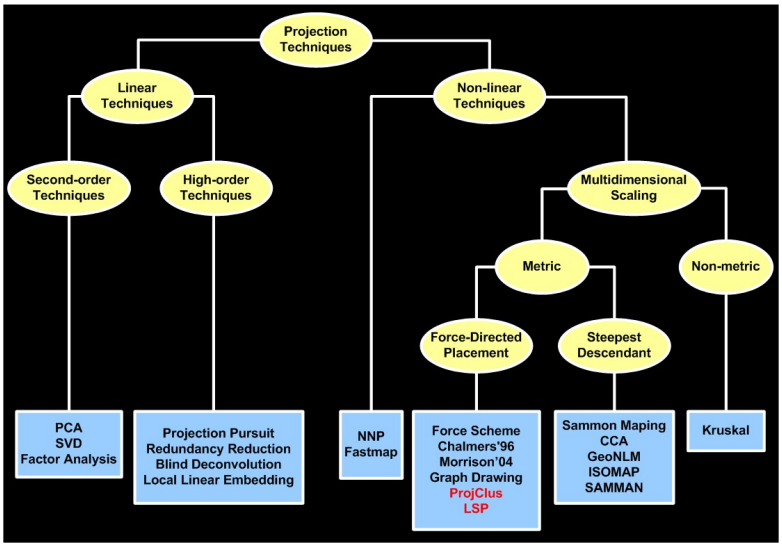
Projection Techniques

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



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

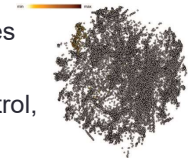
33



34

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

35

Approach and Method

- Understand the data
- Understand the needs
- Exploratory – agree with user/customer/partner
- Find relevant information
- Know the available methods
- Work in pairs/groups.

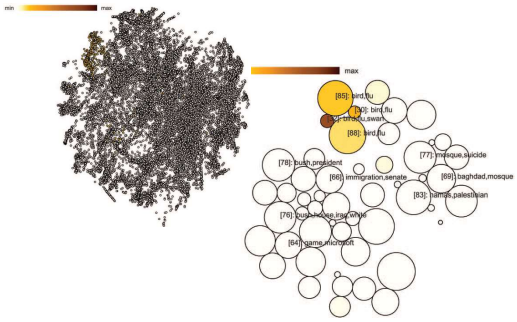
36

Challenges

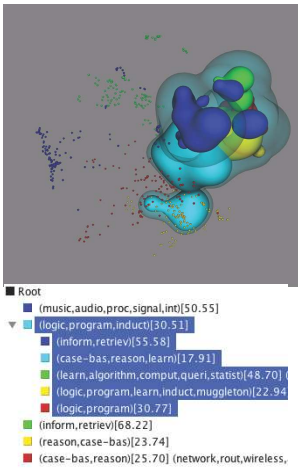
- Sheer volume
- Data transformation/formatting/structuring
- Ownership of the data
- Different types
- Spurious correlations
- Inespecificity of questions

37

Visualization examples: clutter



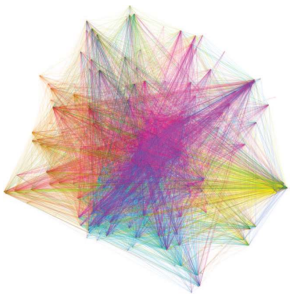
Paulovich and Minghim, HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections, ***IEEE Trans. Visualization & Computer Graphics***, *2008*



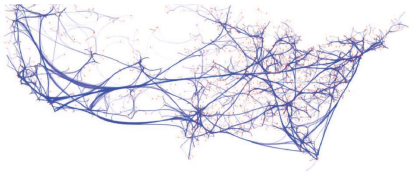
Poco; Etedmapour, Paulovich, Long, Rosenthal, Oliveira, Linsen, Minghim. A framework for exploring multidimensional data with 3D projections, ***Computer Graphics Forum***, Eurovis 2011.

38

Networks: even worse



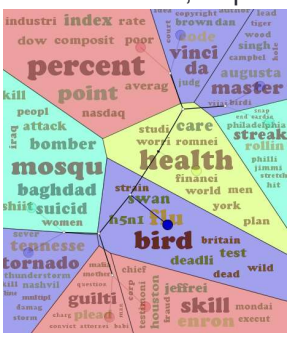
Ersoy, Hurter, Paulovich, Cantareira, Telea, Skeleton-based edge bundling for graph visualization. ***IEEE Trans. Visualization and Computer Graphics,*** Infovis 2011



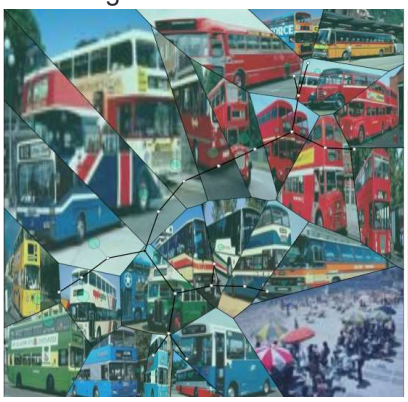
39

More Data – Scalability and Summarization

- Wordclouds, Representative Images



Multi-level text



Multi-level images

Silva, Paiva, Telles, Zampieri, Rolli, Minghim. The Visual SuperTree: Similarity-based Multi-scale Visualization, ***The Visual Computer (CGI'19)***.

40

Content

- Point-based techniques
- Attribute based techniques
- An overview of Scientific Visualization
- Tools
  - Existing Public and Commercial tools
  - Python for Vis
  - R for vis
  - Orange
  - Tableau Public
- Case Studies

MOODLE USP

41

Homework

- (in preparation for project 1) Explore HDR variables and indices (see slide 17 – links). Download 2018 indices and analyze them.
  - Mention 5 interesting patterns found, 3 expected, 2 somewhat surprising
  - How does Brazil relate with other countries with similar HDI, both in a positive and in a negative way?
  - How does Brazil relate directly (if at all) with countries in a different range of HDI?
- (in preparation for project 2) Download 2018 indices from HDR. Download a Health data set of your choice. Choose a visualizer on the internet and analyze them. Report your findings and the features of the tool of your choice.
- (in preparation for project 4) Choose 2 different visualizations programmed in D3 (see slide 17 – links)
  - Run each one of these with 2 different data sets of your choice.

42

Evaluation - Undergrad

- 4 projects - 3 visualization tasks, one programing task (70%)
- 1 test (31/10)

Evaluation - Grad

- 1 project (30 %)
- 2 paper discussions (15%)
- 1 seminar on a particular visualization subject (15%)
- 1 test (40%) - 31/10

ICMC USP SÃO CARLOS

Instituto de Ciências Matemáticas e de Computação

| Universidade de São Paulo |

VISUALIZATION, DATA SCIENCE AND BIG DATA

Rosane Minghim

THANK YOU!!!