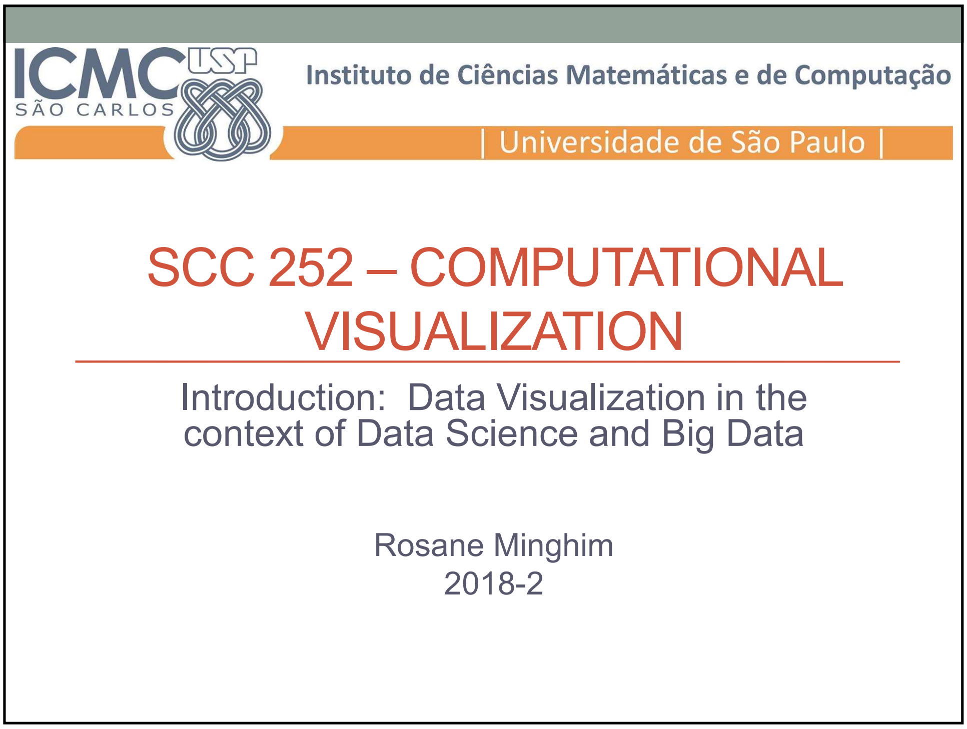


Instituto de Ciências Matemáticas e de Computação

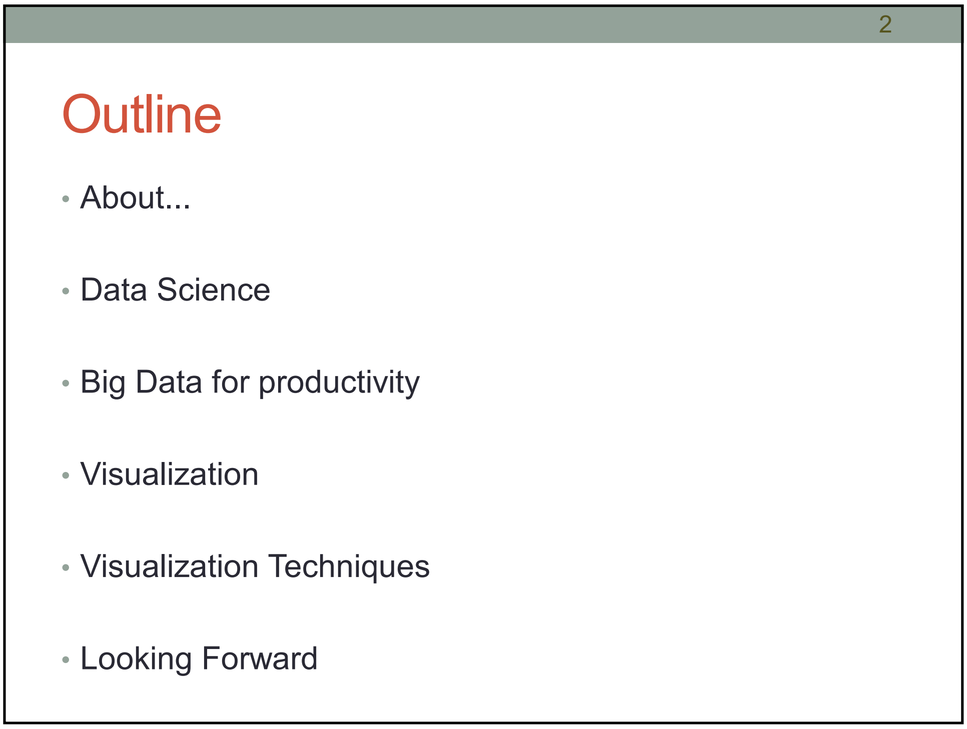
| Universidade de São Paulo |

# SCC 252 – COMPUTATIONAL VISUALIZATION

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

Rosane Minghim

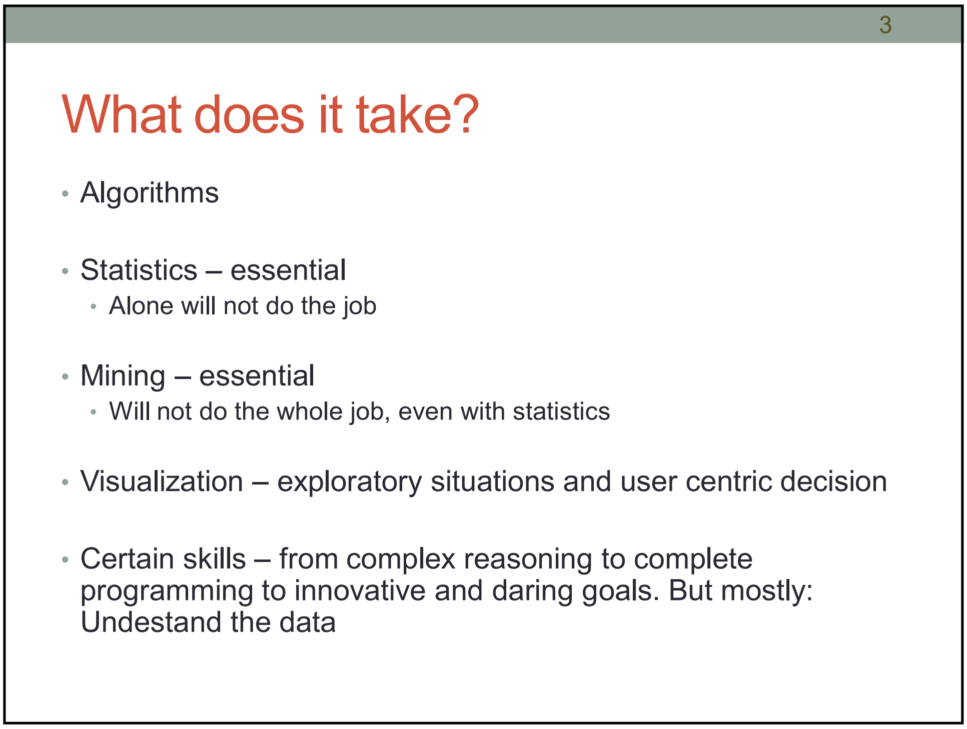
2018-2



2

## Outline

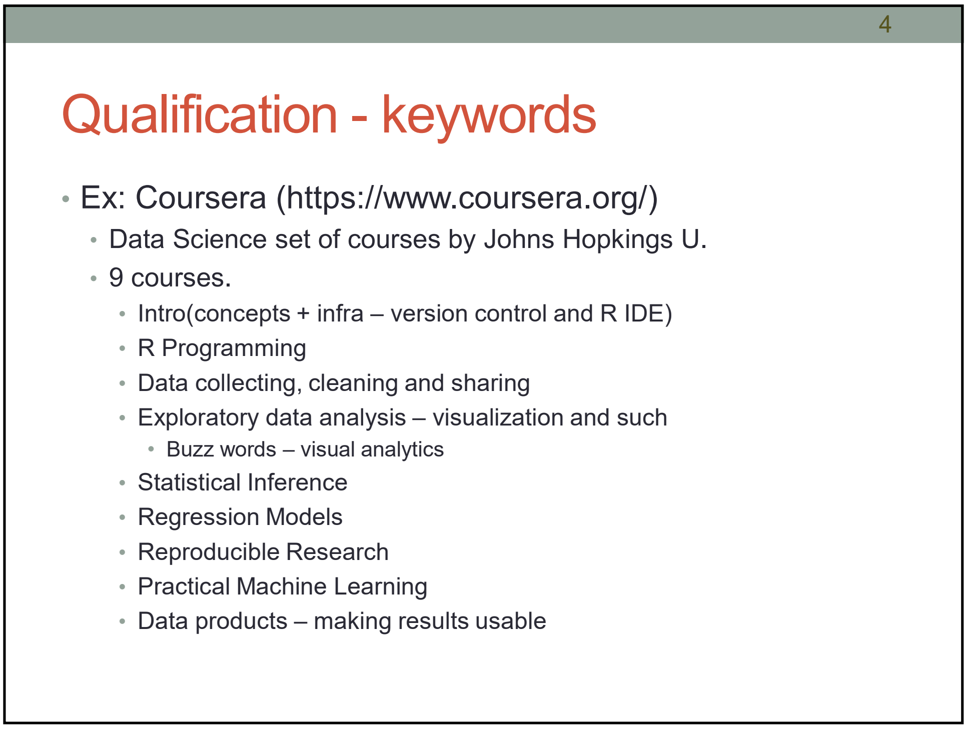
- About...
- Data Science
- Big Data for productivity
- Visualization
- Visualization Techniques
- Looking Forward



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 !

BIG DATA FOR PRODUCTIVITY CONGRESS 2015
EVERYTHING CHANGES
THE ONLY CONFERENCE IN 2015 TO FOCUS ON THE BUSINESS SIDE OF BIG DATA
OCTOBER 19 TO 21, 2015 / HALIFAX, NOVA SCOTIA
digital nova scotia
DALHOUSIE UNIVERSITY Inspiring Minds
WORLD CONFEDERATION OF PRODUCTIVITY SCIENCE
NTT Data
EY
Nova Scotia Business Inc.
CISCO
Google
BOEING
NOVA SCOTIA
IBM
HITACHI Inspire the Next

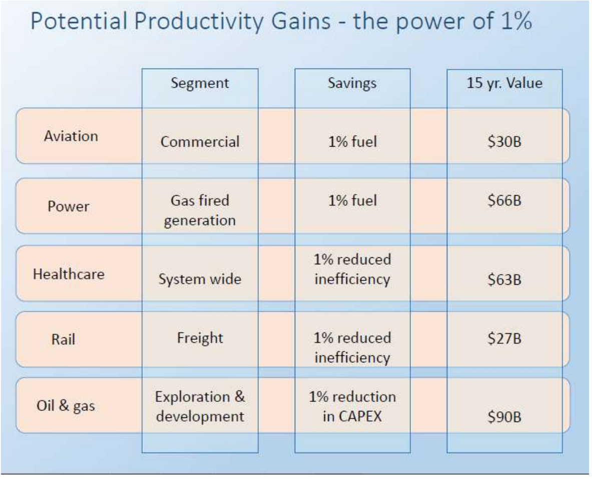
6

## 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.

# 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 |

# 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,

# How big – just a sample

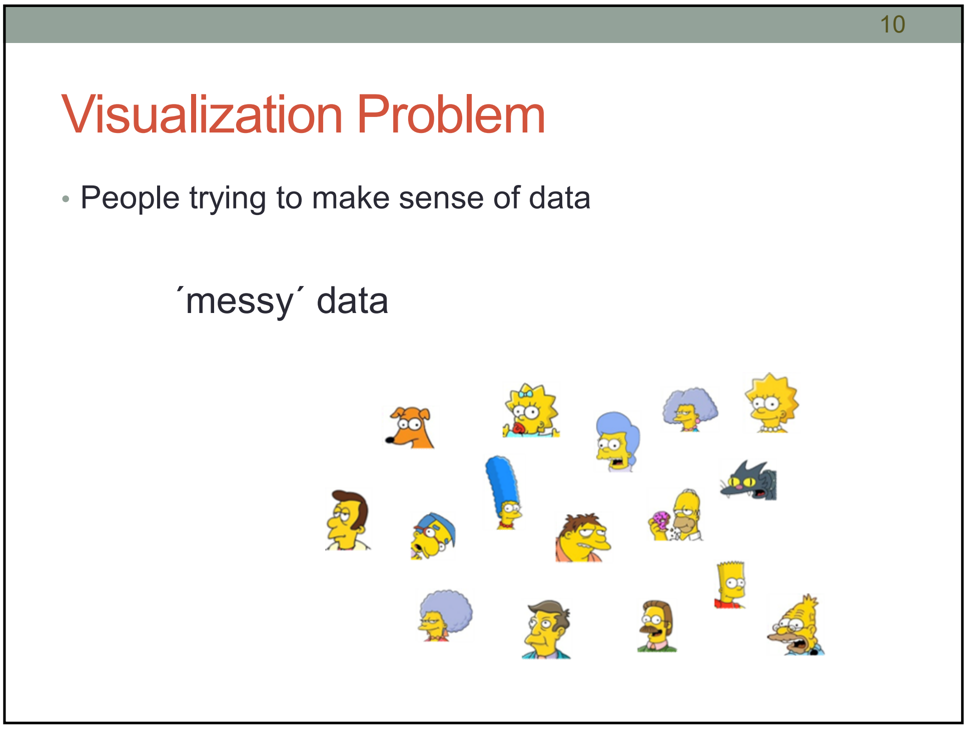
## From IoT

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



# Visualization Problem

- People trying to make sense of data

´messy´ data

# Data is...

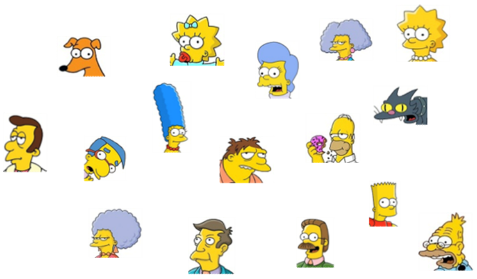
- Far too complex... (many dimensions, many types)
- Far too big... (´easy´ to collect)
- Far too varied... (images, videos, documents, news, networks)
- Never ending... (data streams)
- Much redundancy...
- Many relationships...
- Pieces missing...

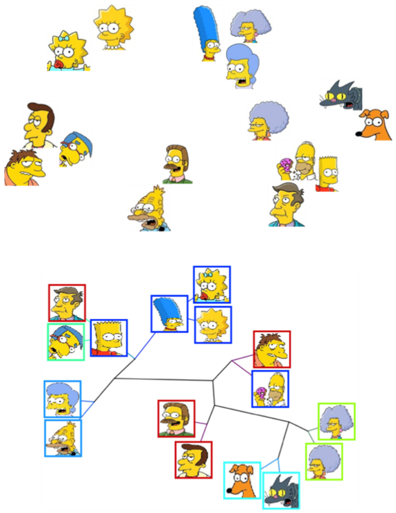
- Studying natural & artificial systems and phenomena implies in handling lots of data...

# What does your data tell???

- People trying to make sense of data

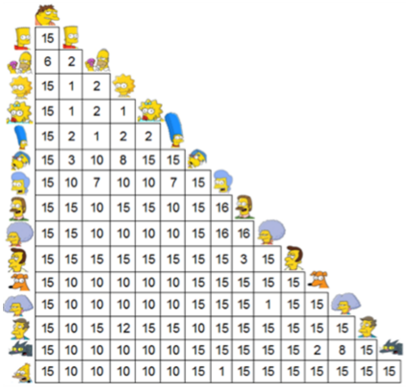
´messy´ data





# Techniques

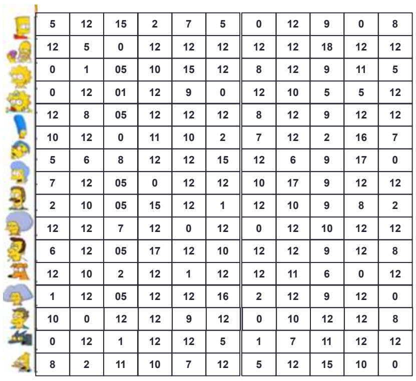
- Multidimensional visualization: data organization



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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)

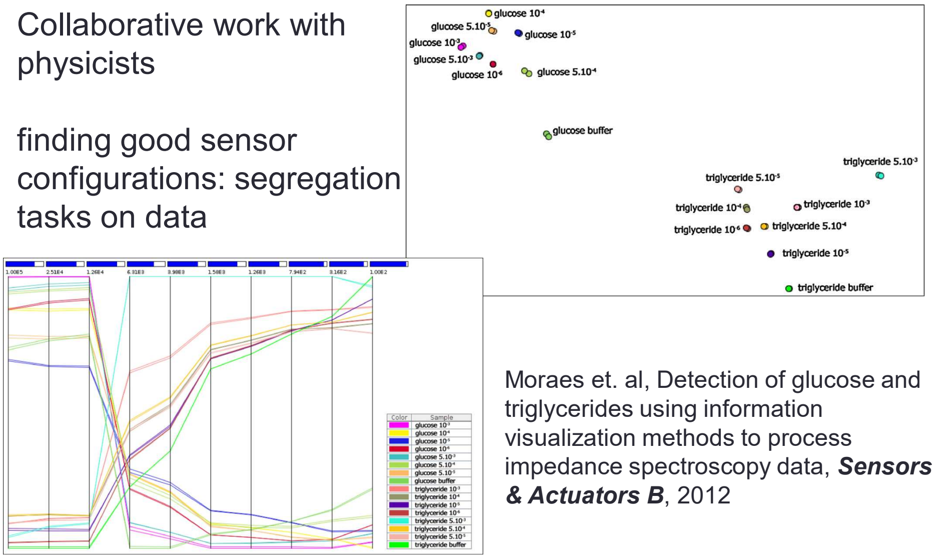
# Example: On studies on ecology and environment

- 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

# Example: Data from nanotech sensors & biosensors

Collaborative work with physicists

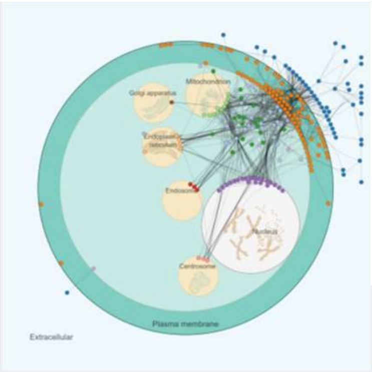
finding good sensor configurations: segregation tasks on data

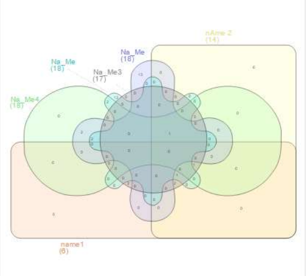


Moraes et. al, Detection of glucose and triglycerides using information visualization methods to process impedance spectroscopy data, ***Sensors & Actuators B***, 2012

# 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.

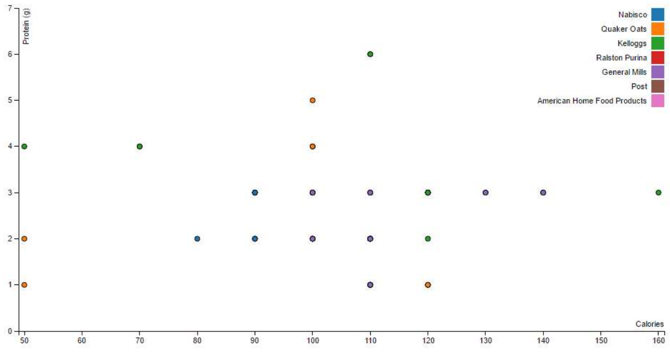
# Links to sources of data visualization tools and data

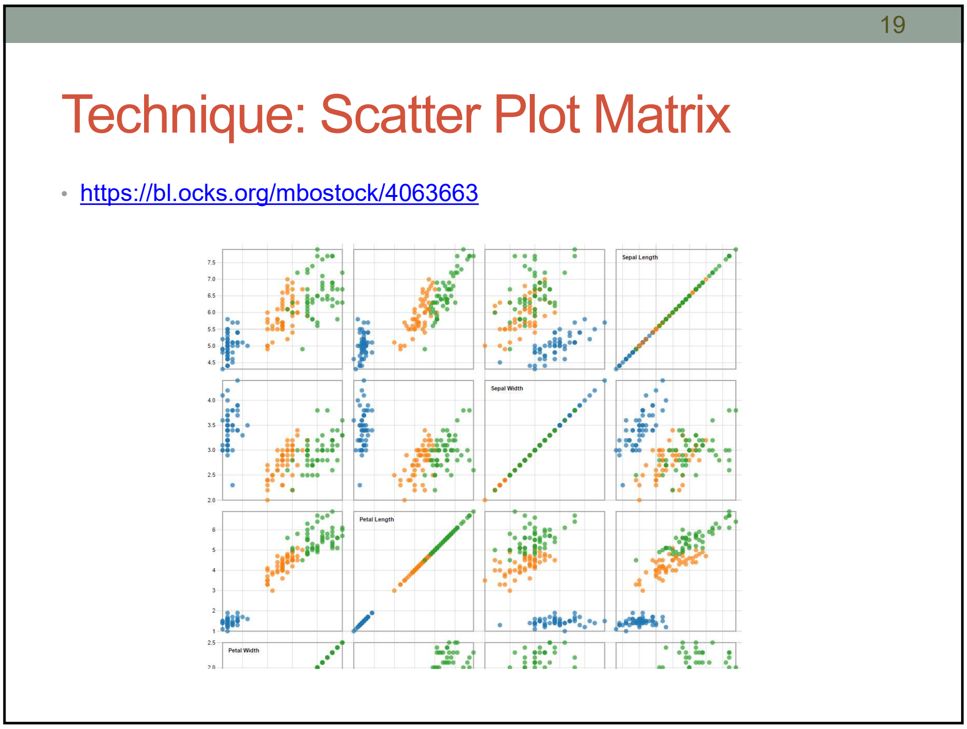
- HDR (ONU):
  - (data) http://hdr.undp.org/en/composite/GII
  - (vis) http://hdr.undp.org/en/data-explorer/

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

# Technique: Scatter Plot

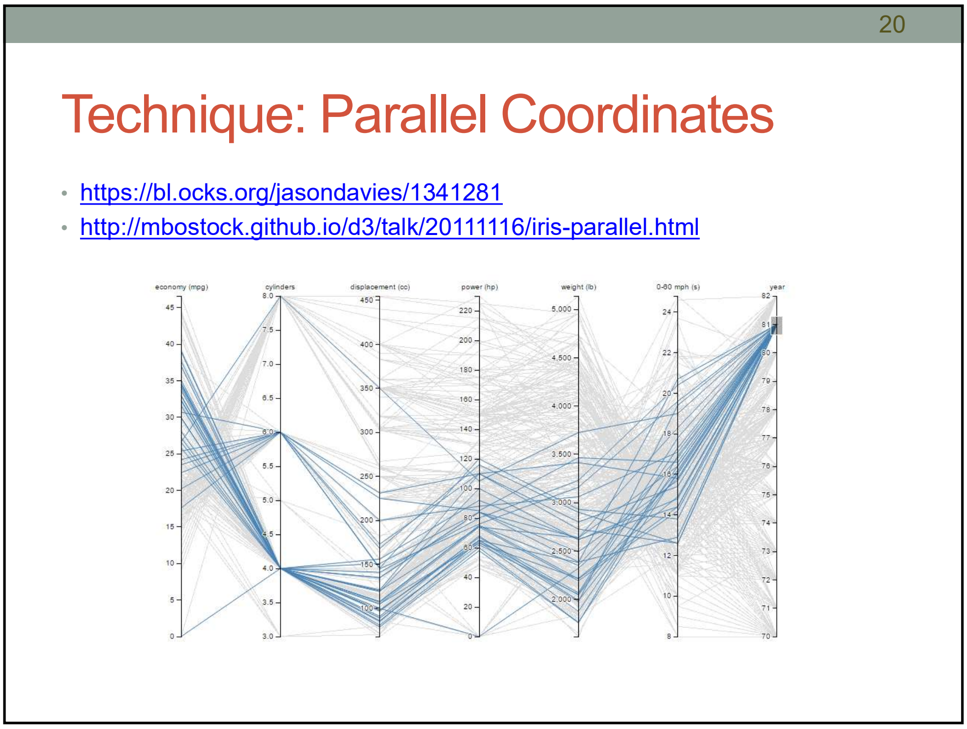
- http://bl.ocks.org/weiglemc/6185069





# Technique: Scatter Plot Matrix

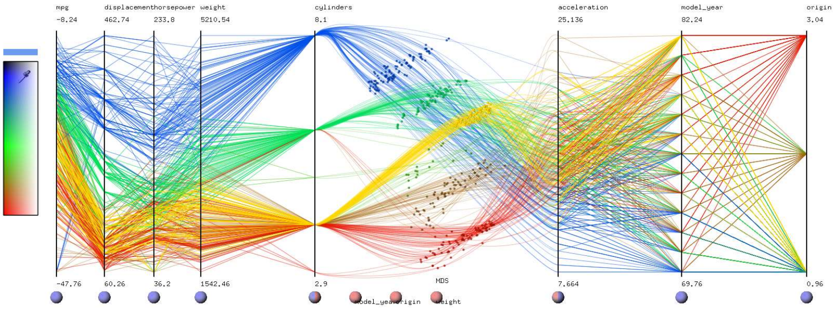
- https://bl.ocks.org/mbostock/4063663



# Technique: Parallel Coordinates

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

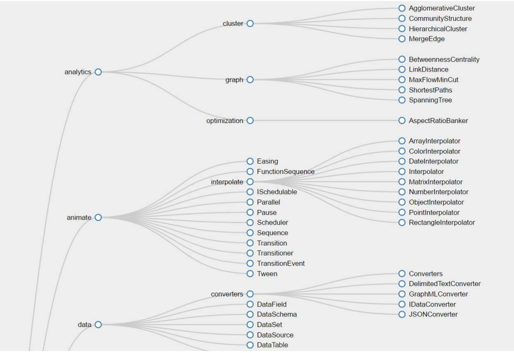
# Technique: Scattering Points in Parallel Coordinates



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

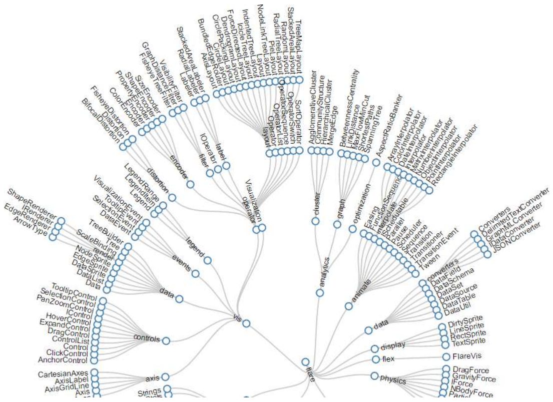
# Technique: Tree Visualization

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



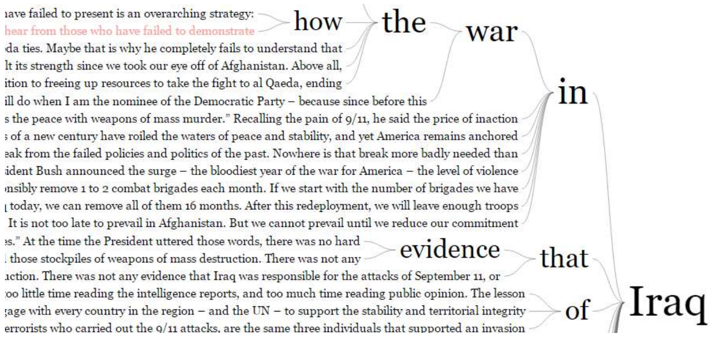
# Technique: Radial Reingold–Tilford Tree

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



# Technique: Word Tree

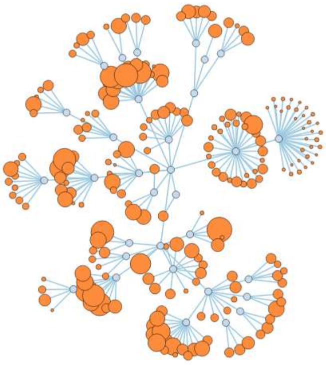
- https://www.jasondavies.com/wordtree/?source=obama-war-speech.txt&prefix=Iraq&reverse=1



ıave failed to present is an overarching strategy: how the war in Iraq
hear from those who have failed to demonstrate how the war in Iraq
·da ties. Maybe that is why he completely fails to understand that the war in Iraq
lt its strength since we took our eye off of Afghanistan. Above all, the war in Iraq
ition to freeing up resources to take the fight to al Qaeda, ending the war in Iraq
ill do when I am the nominee of the Democratic Party – because since before this war in Iraq
s the peace with weapons of mass murder." Recalling the pain of 9/11, he said the price of inaction in Iraq
; of a new century have roiled the waters of peace and stability, and yet America remains anchored in Iraq
eak from the failed policies and politics of the past. Nowhere is that break more badly needed than in Iraq
ident Bush announced the surge – the bloodiest year of the war for America – the level of violence in Iraq
ınsibly remove 1 to 2 combat brigades each month. If we start with the number of brigades we have in Iraq
ı today, we can remove all of them 16 months. After this redeployment, we will leave enough troops in Iraq
It is not too late to prevail in Afghanistan. But we cannot prevail until we reduce our commitment in Iraq
es." At the time the President uttered those words, there was no hard evidence that Iraq
l those stockpiles of weapons of mass destruction. There was not any evidence that Iraq
ıction. There was not any evidence that Iraq was responsible for the attacks of September 11, or that Iraq
oo little time reading the intelligence reports, and too much time reading public opinion. The lesson of Iraq
;age with every country in the region – and the UN – to support the stability and territorial integrity of Iraq
errorists who carried out the 9/11 attacks, are the same three individuals that supported an invasion of Iraq

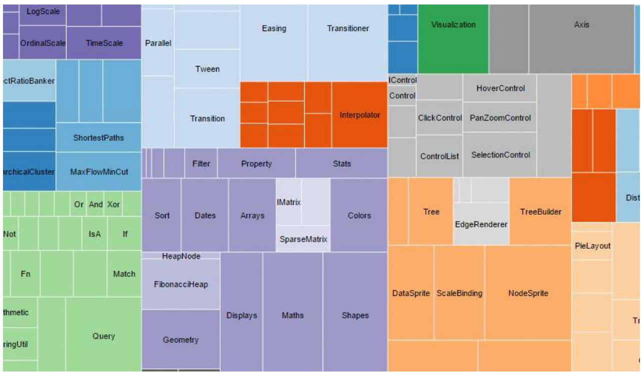
# Technique: Tree - Force Layout

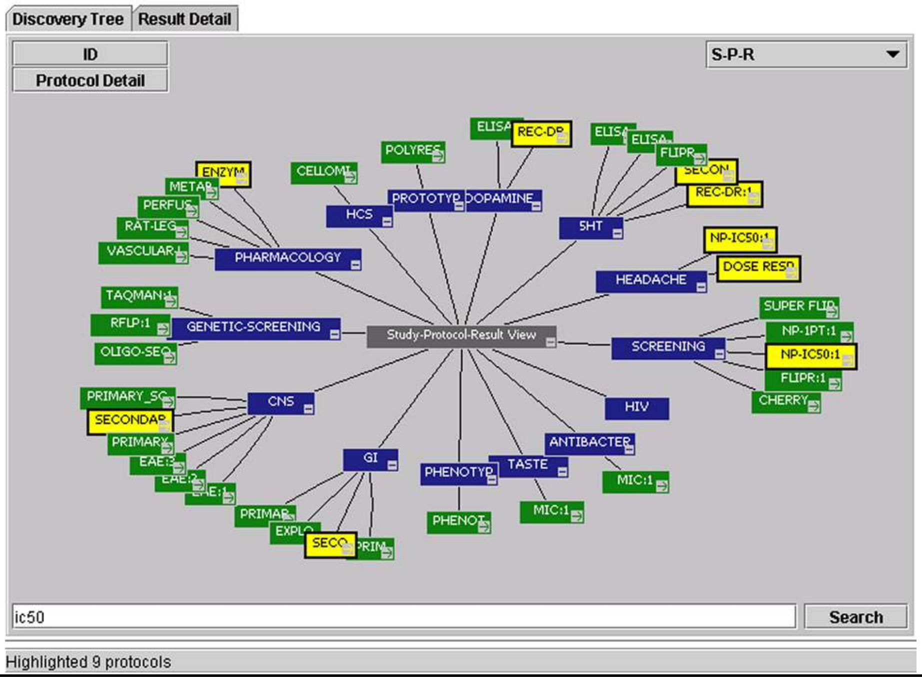
- http://mbostock.github.io/d3/talk/20111116/force-collapsible.html



# Technique: Tree - TreeMap

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

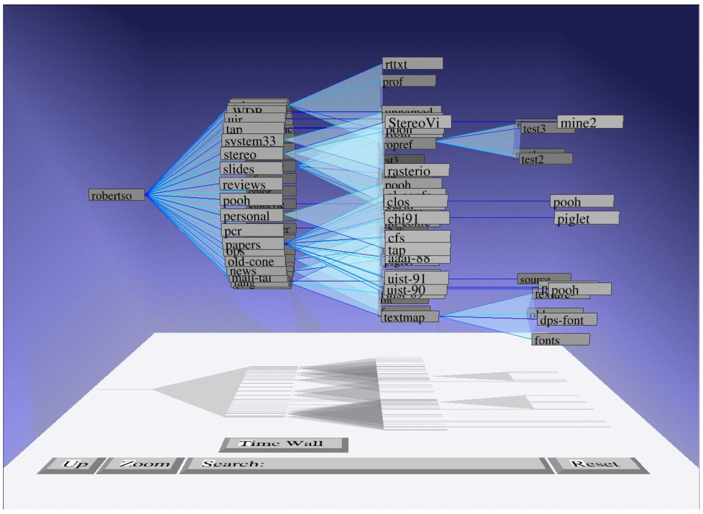




ic50

Highlighted 9 protocols

# Cone Tree

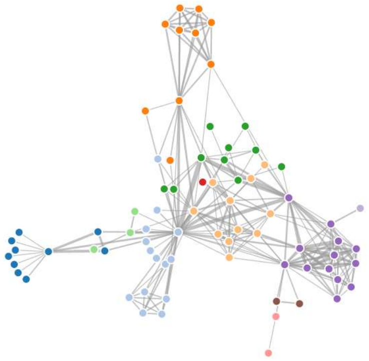


File system structure visualized as a cone tree

Animated 3D visualization of hierarchical data

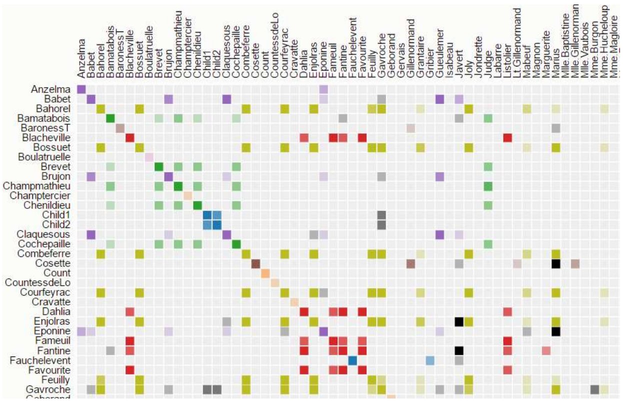
# Force-directed Graph Layout

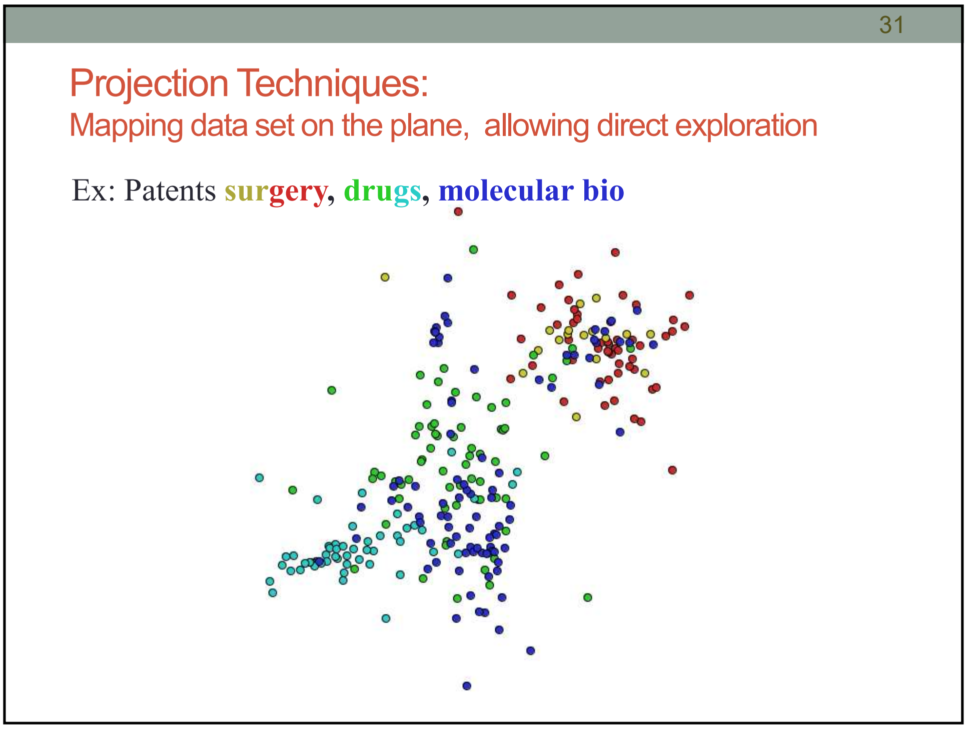
- http://bl.ocks.org/mbostock/4062045



# Adjacency Matrix Graph Layout

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



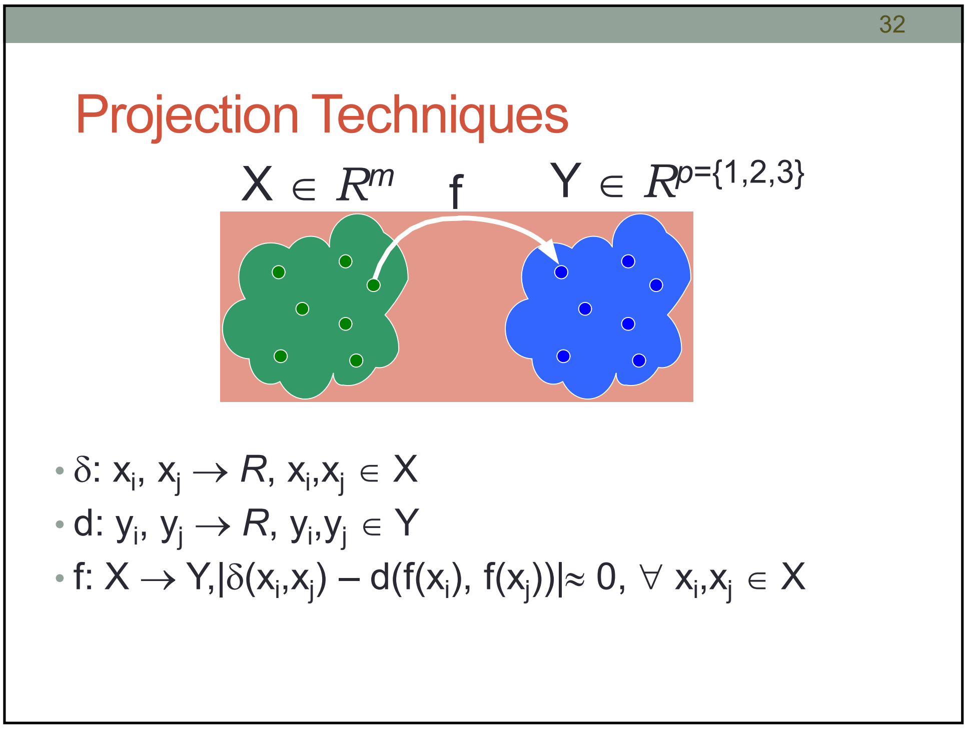


31

## Projection Techniques:

Mapping data set on the plane,  allowing direct exploration

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

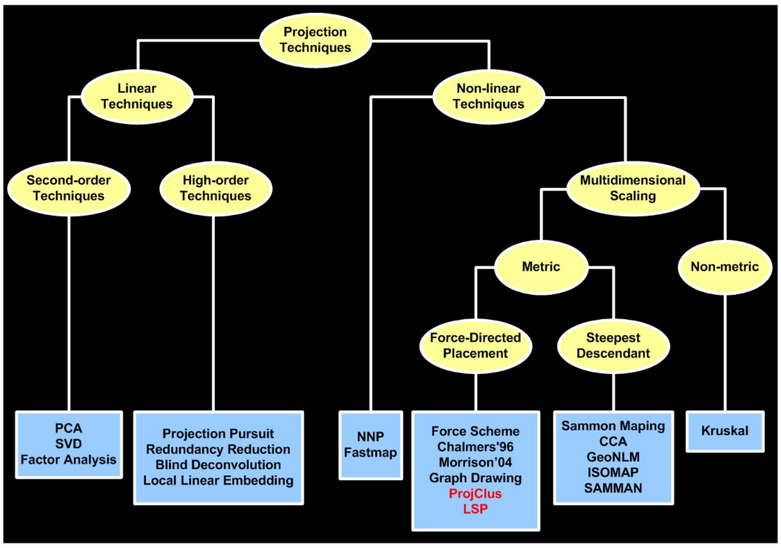


32

## Projection Techniques

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

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



Projection Techniques

Linear Techniques

Non-linear Techniques

Second-order Techniques

High-order Techniques

Multidimensional Scaling

Metric

Non-metric

Force-Directed Placement

Steepest Descendant

PCA<br>SVD<br>Factor Analysis

Projection Pursuit<br>Redundancy Reduction<br>Blind Deconvolution<br>Local Linear Embedding

NNP<br>Fastmap

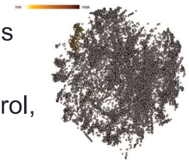
Force Scheme<br>Chalmers'96<br>Morrison'04<br>Graph Drawing<br>ProjClus<br>LSP

Sammon Maping<br>CCA<br>GeoNLM<br>ISOMAP<br>SAMMAN

Kruskal

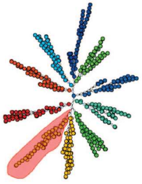
# Similarity based Techniques

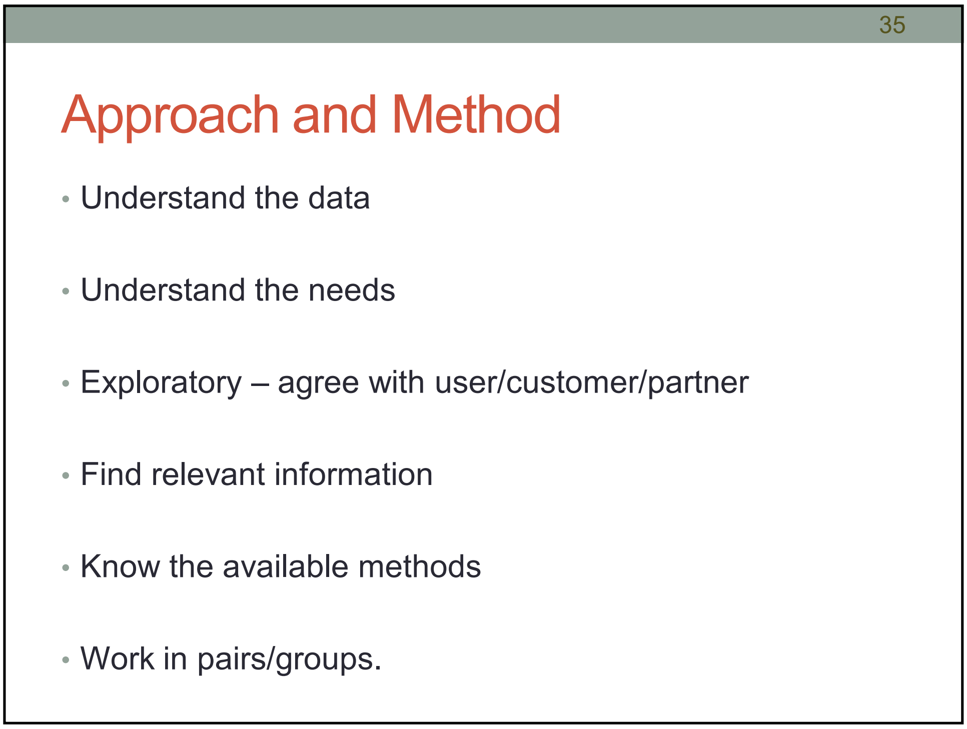
- Projections
  - variations on MDS, dimension reduction, or other approaches
  - data mapped to low-dimensional visual space
  - preserving distances vs neighborhoods, global vs. local control, segregation



- 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

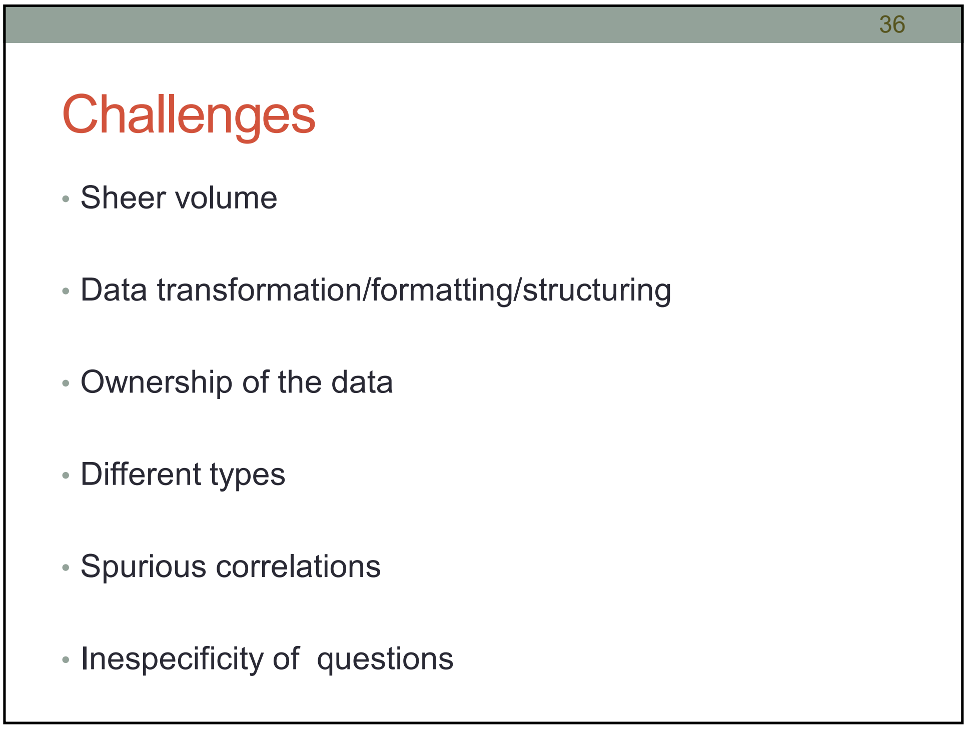




35

## 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.

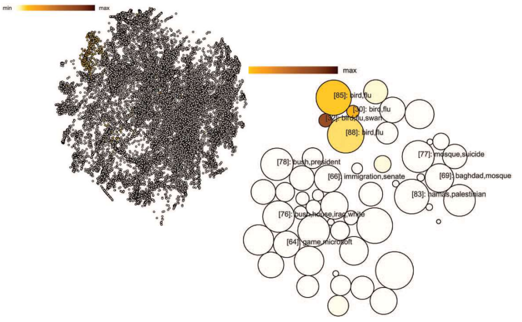


36

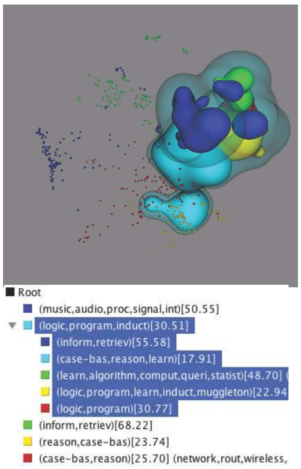
## Challenges

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

# Visualization examples: clutter

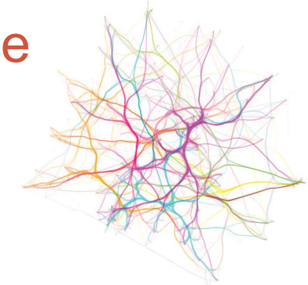


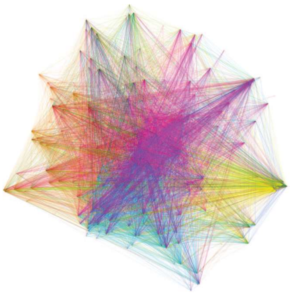
Paulovich and  Minghim, HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections, ***IEEE Trans. Visualization & Computer Graphics***, *2008*


Root
(music,audio,proc,signal,int)[50.55]
(logic,program,induct)[30.51]
(inform,retriev)[55.58]
(case-bas,reason,learn)[17.91]
(learn,algorithm,comput,queri,statist)[48.70]
(logic,program,learn,induct,muggleton)[22.94
(logic,program)[30.77]
(inform,retriev)[68.22]
(reason,case-bas)[23.74]
(case-bas,reason)[25.70] (network,rout,wireless,

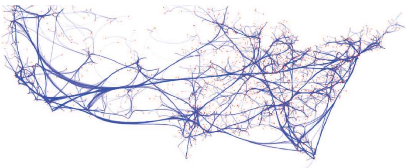
Poco; Etedmapour, Paulovich, Long, Rosenthal, Oliveira, Linsen, Minghim. A framework for exploring multidimensional data with 3D projections, ***Computer Graphics Forum***, Eurovis 2011.

# Networks: even worse



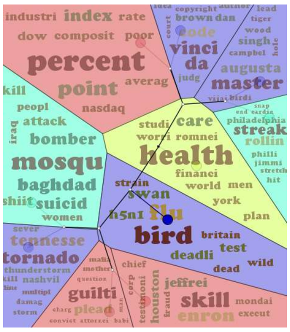


Ersoy, Hurter, Paulovich, Cantareira, Telea, Skeleton-based edge bundling for graph visualization. ***IEEE Trans. Visualization and Computer Graphics,*** Infovis 2011



## More Data – Summarization

- Wordclouds
- Representative Images

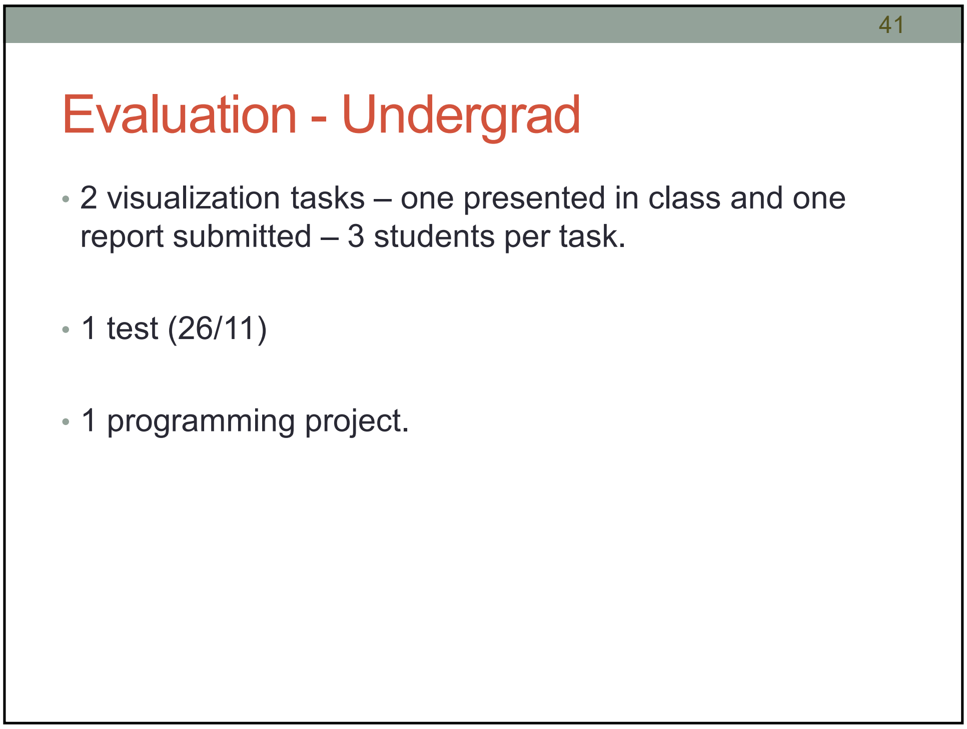


Multi-level text

Multi-level images

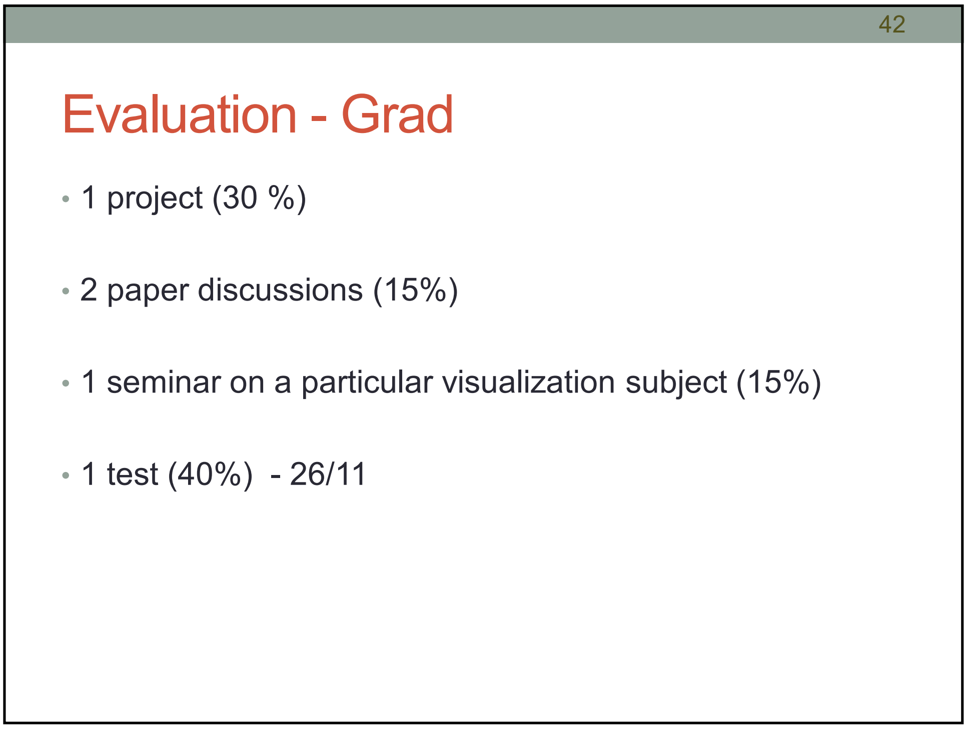
## Homework

- Explore HDR variables and indices (see slide 17 – links)
  - 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?
- Choose 2 different visualizations programmed in D3 (see slide 17 – links)
  - Run each one of these with 2 different data sets of your choice.



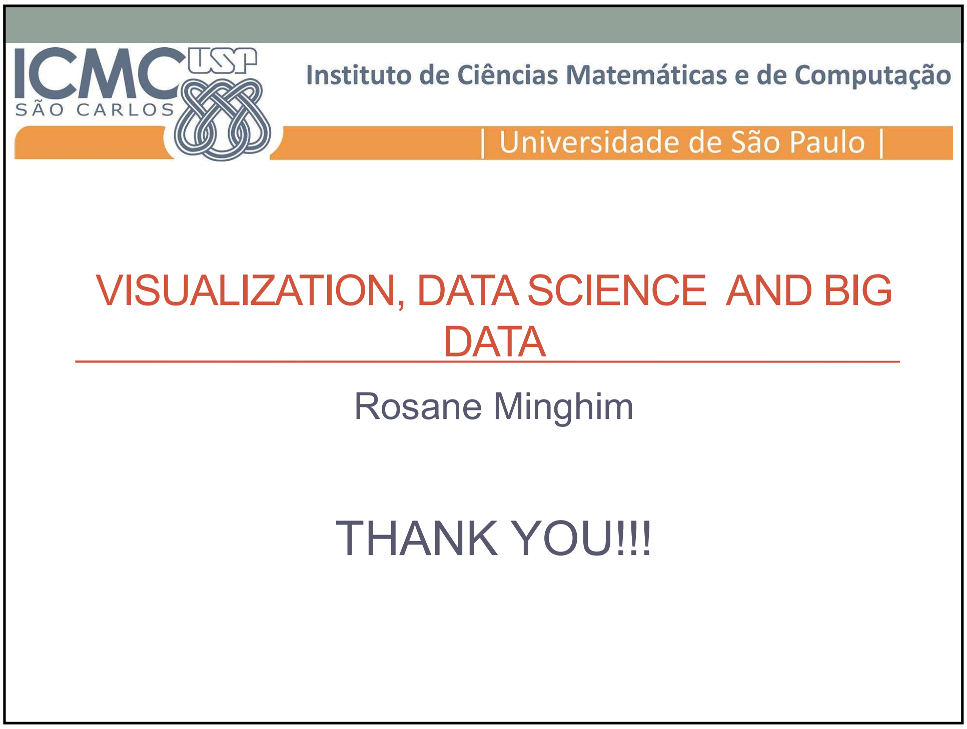
## Evaluation - Undergrad

- 2 visualization tasks – one presented in class and one report submitted – 3 students per task.
- 1 test (26/11)
- 1 programming project.



## Evaluation - Grad

- 1 project (30 %)
- 2 paper discussions (15%)
- 1 seminar on a particular visualization subject (15%)
- 1 test (40%)  - 26/11



Instituto de Ciências Matemáticas e de Computação

| Universidade de São Paulo |

# VISUALIZATION, DATA SCIENCE AND BIG DATA

Rosane Minghim

THANK YOU!!!