Data Visualisation

Patrick De Mazière, MScEng, PhD

Bio Patrick De Mazière °1973

- MSc Eng Computer Sciences, Programming (KU Leuven, 1998);
- PhD Biomedical Sciences (KU Leuven, 2007)
- Masterclass High-Tech Entrepreneurship (KU Leuven-LRD, 2010)
- Postdoc, Neurophysiology/Fac. Medicine (KU Leuven, 2007 - 2014)
  - Member Steering Commitee KU Leuven HPC (2005 - 2010)
  - Brain Research using Mathematical Models (~ HPC + Data Mining => ANN, Deeplearning, ...)
  - Statistics, Data/Text Mining (~ HPC)
  - Educational Projects
- Teamleader of KIC ITech & Lector Applied Informatics (UCLL, 2012 – 2017; #15 heads)
- Head of Research & Expertise Center Digital Solutions (UCLL, 2018 - ; #35 heads)

Data Visualisation

Why ?

- Get to know your data
- Get insights, trends
- Not machine learning, AI or other hocus pocus, albeit that understanding your data is the basis for ML, AI, …

Coming 50 minutes…

- Depending on the kind of data
  - Relationships / correlations between items
  - Hi-Dim, multi-parameter datasets
  - …

I will give you some tips on how to reveal the hidden trends, info, … using simple, easy software apps/tricks to meet your data

Steps to become familiar with your data

1. Discover nature of your data
2. Chose corresponding tricks aka algorithms aka software
3. Analyse/Interpret results

Data nature discovery

- Not ordinal, categorical, … albeit interesting as well
- But by nature
  - Samples identified by lots of parameters with or w/o classification

    Hi-Dim data
  - Relationships between items

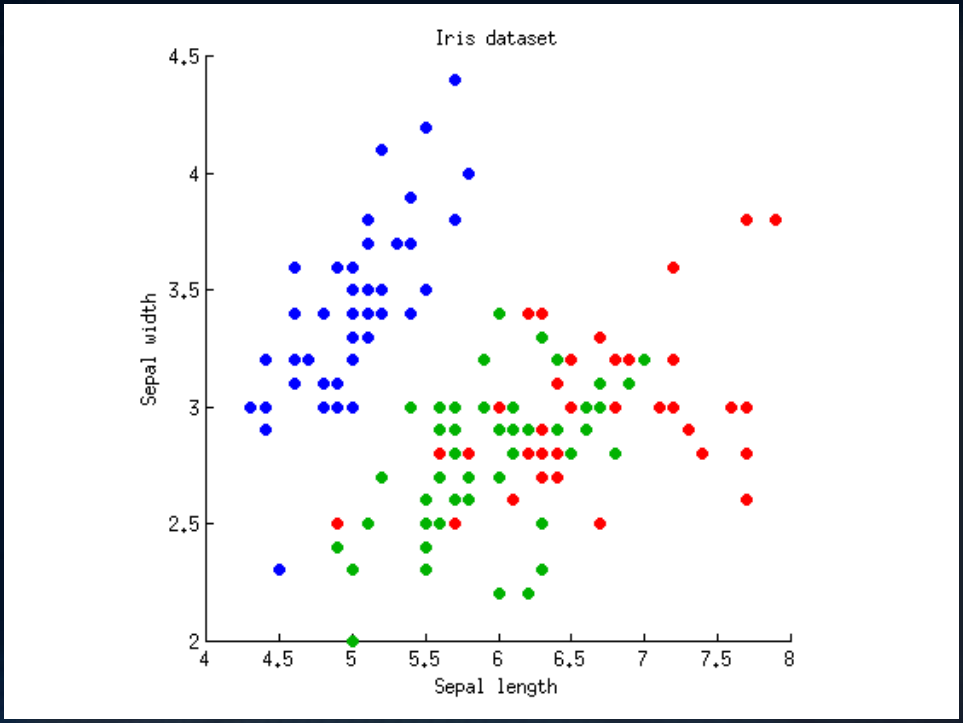
    Highly correlated data

Excel can be of help to discover that nature..

- Load your data into Excel (CSV format is often used)
- Use filter and sort options per column to have a look at the data
- Create additional columns to *check* data's nature:
  - Use functions: isnumber(), istext(), islogical(), … and filter on column => if only true comes out, you're OK ☺
- Use scatterplot and/or histograms to find out data content of individual columns
- …

Visualise Hi-Dim data

- In 2D we use a scatterplot with each axis a dimension/parameter
- E.g. well-known Iris flower dataset (3 types of flowers, more info on https://en.wikipedia.org/wiki/Iris_flower_data_set)



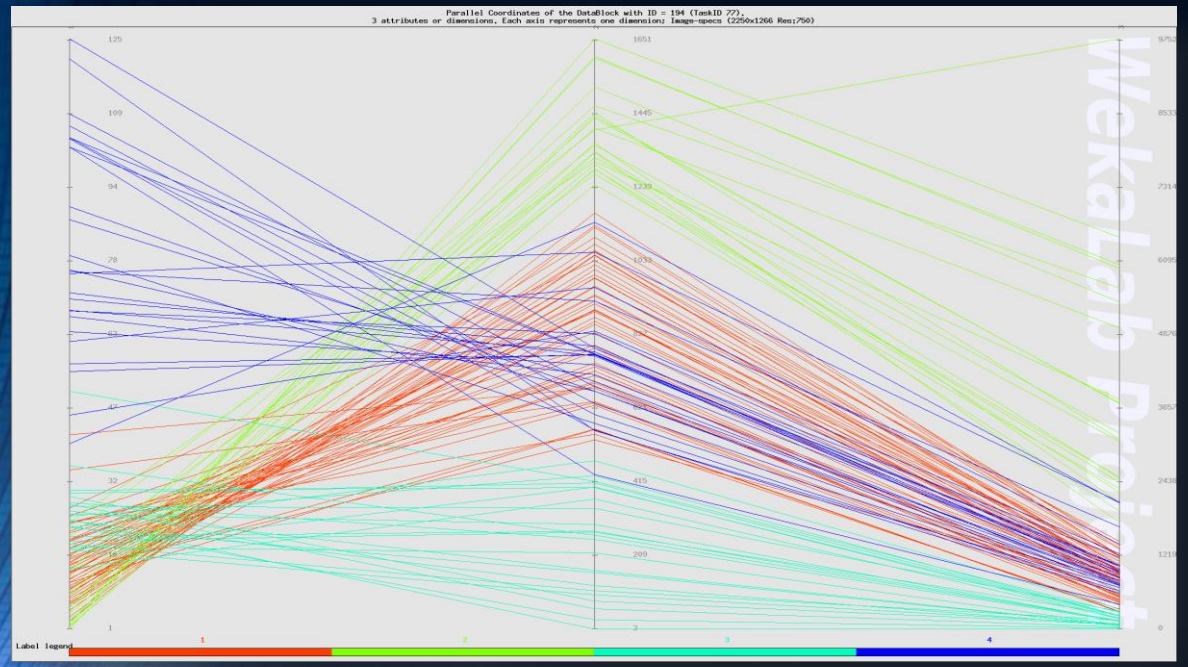
Visualise Hi-Dim data

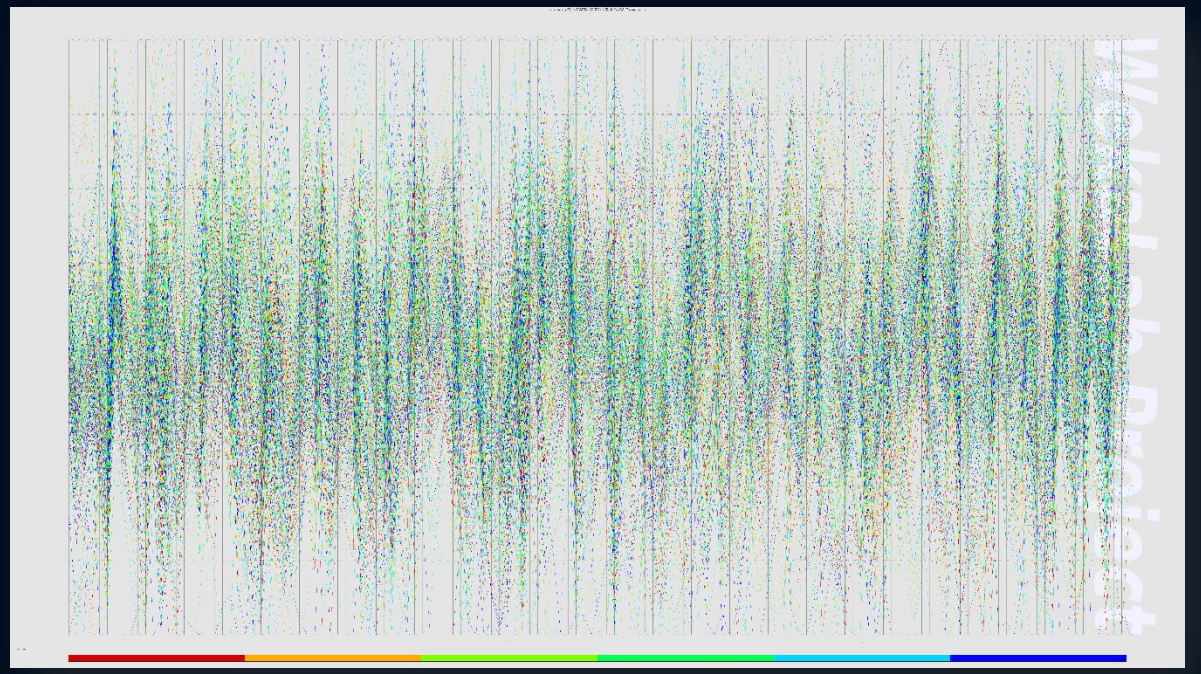
But if it's not 2D, but xD with x >> 10 or even x >> 100

- 2D scatterplot is too limited,
- xD visualisation not possible on paper/screen if we think of xD as the next step after 2D, 3D, …
- So here comes parallel coordinates to the rescue. PC recipe is easy:
  - Think of a spreadsheet of data: rows are samples, columns are parameters (dims)
  - Scale values per column to interval [-1, 1] (*normalisation*)
  - Consider each column as a vertical axis and draw the samples as lines connecting the different axes

Visualise Hi-Dim Data

- Two PC Examples of labeled (supervised) data: 3D en 139D



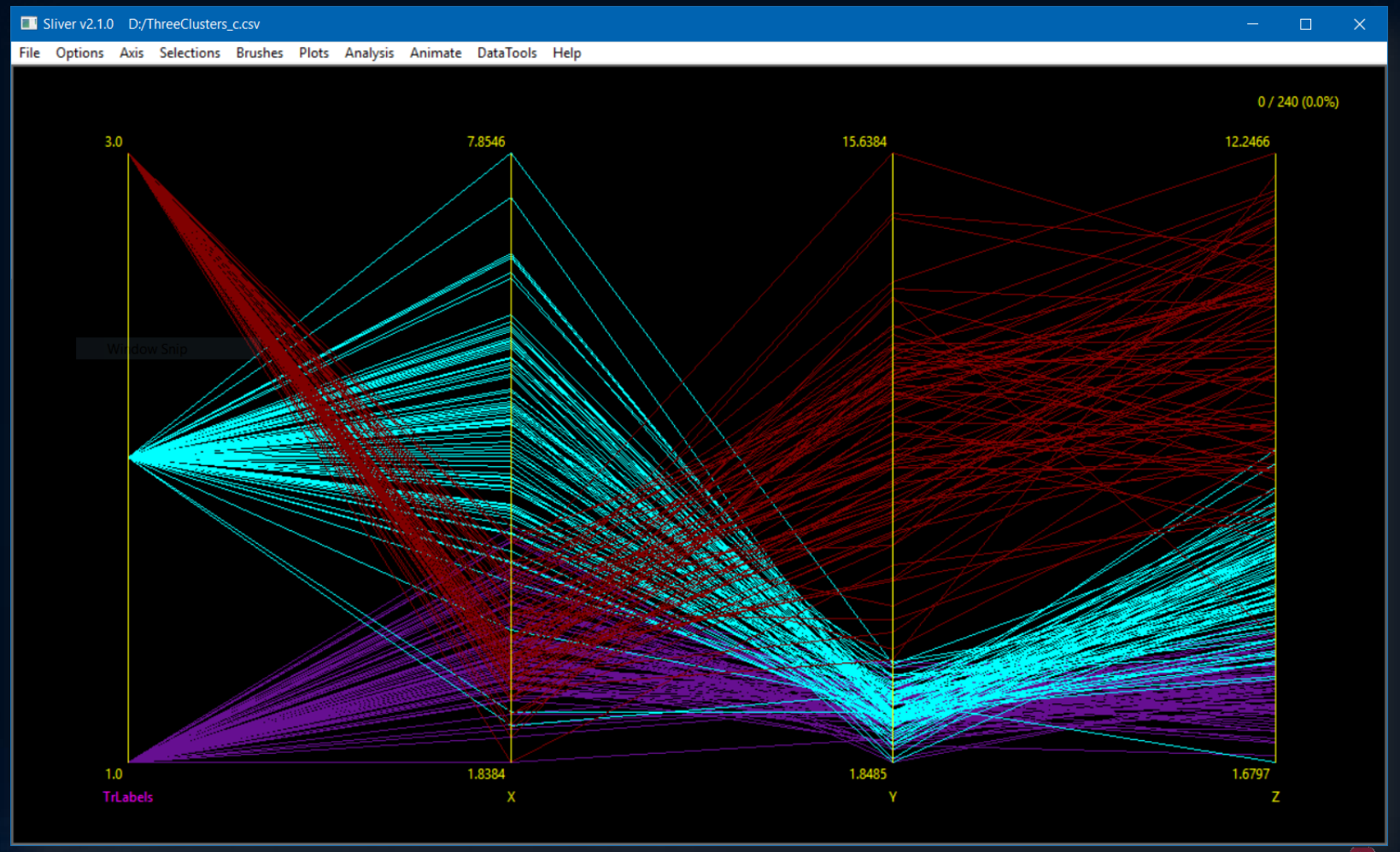


- Recommended tool: Sliver software http://www.sliversoftware.com/download.htm (standalone, no installation needed)

Visualise Hi-Dim Data with Sliver

Recipe

1. Load CSV file
   File > Read File > All Var
2. Choose Axis
   Axis > Select Axis by Name
3. Panel opens, choose most discriminating Axis, ie TrLabels
4. Press r key to colourise according to chose axis



Highly Correlated Data

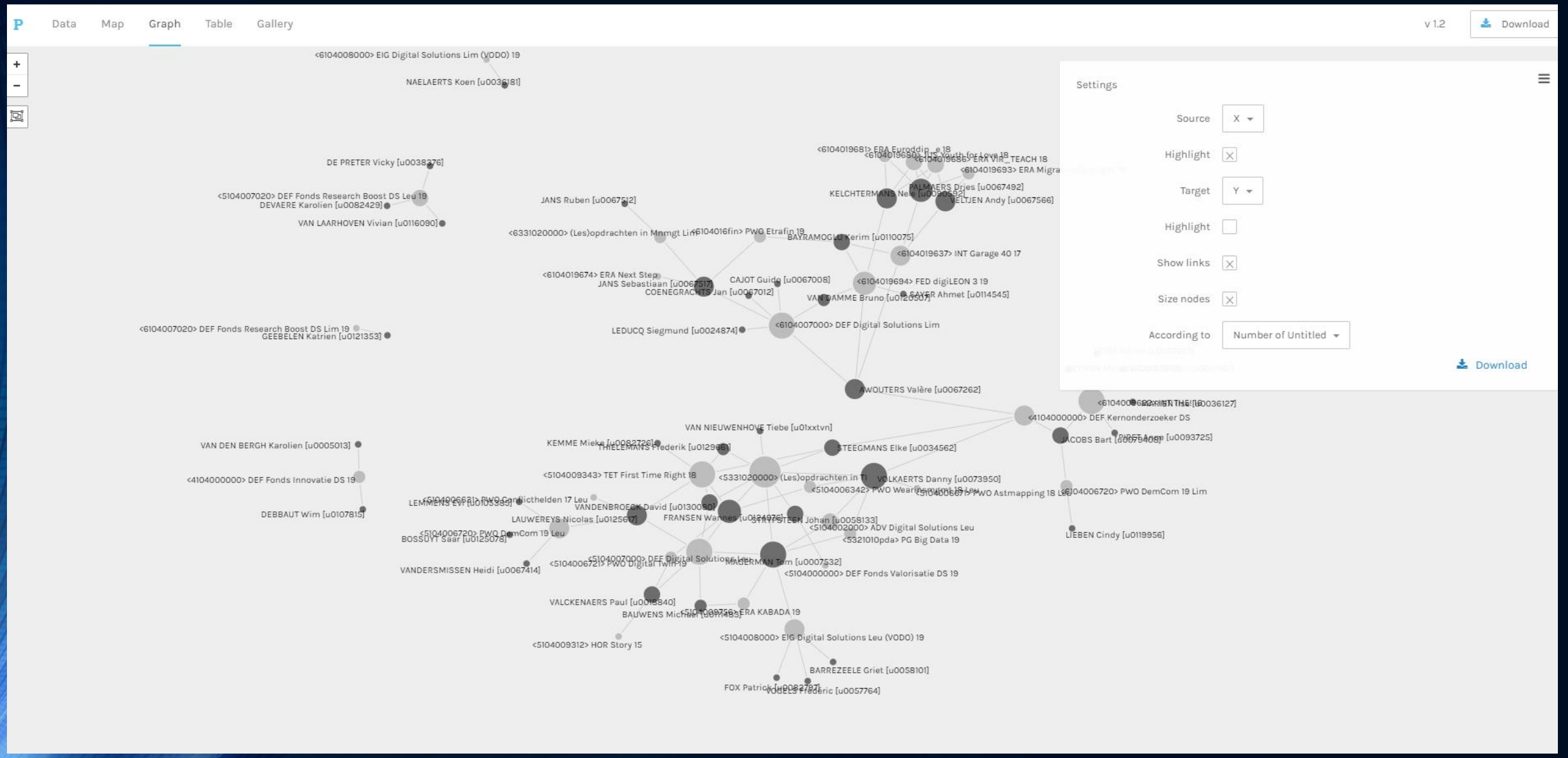
- Mathematically, almost all data is correlated
- Here we deal with data with structural/intended correlations:
  - Genealogy/Family trees
  - LinkedIn relations
  - Data with a geographical link
  - eg.: Two-columned Excel with column one containing projects and column two the people assigned to it

Highly Correlated Data

- Several, powerful free and sometimes online tools exist:
  - Palladio: https://hdlab.stanford.edu/palladio
    - Online
    - For relationships and/or geographically linked data
  - Gehpi: https://gephi.org
    - to be downloaded & installed
    - general purpose
- If you are a programmer, definitely look at
  - d3.js, cola.js, … which ae performant Javascript libs that turn the data avaliable to your static web page (even offline ones) into powerful visuals

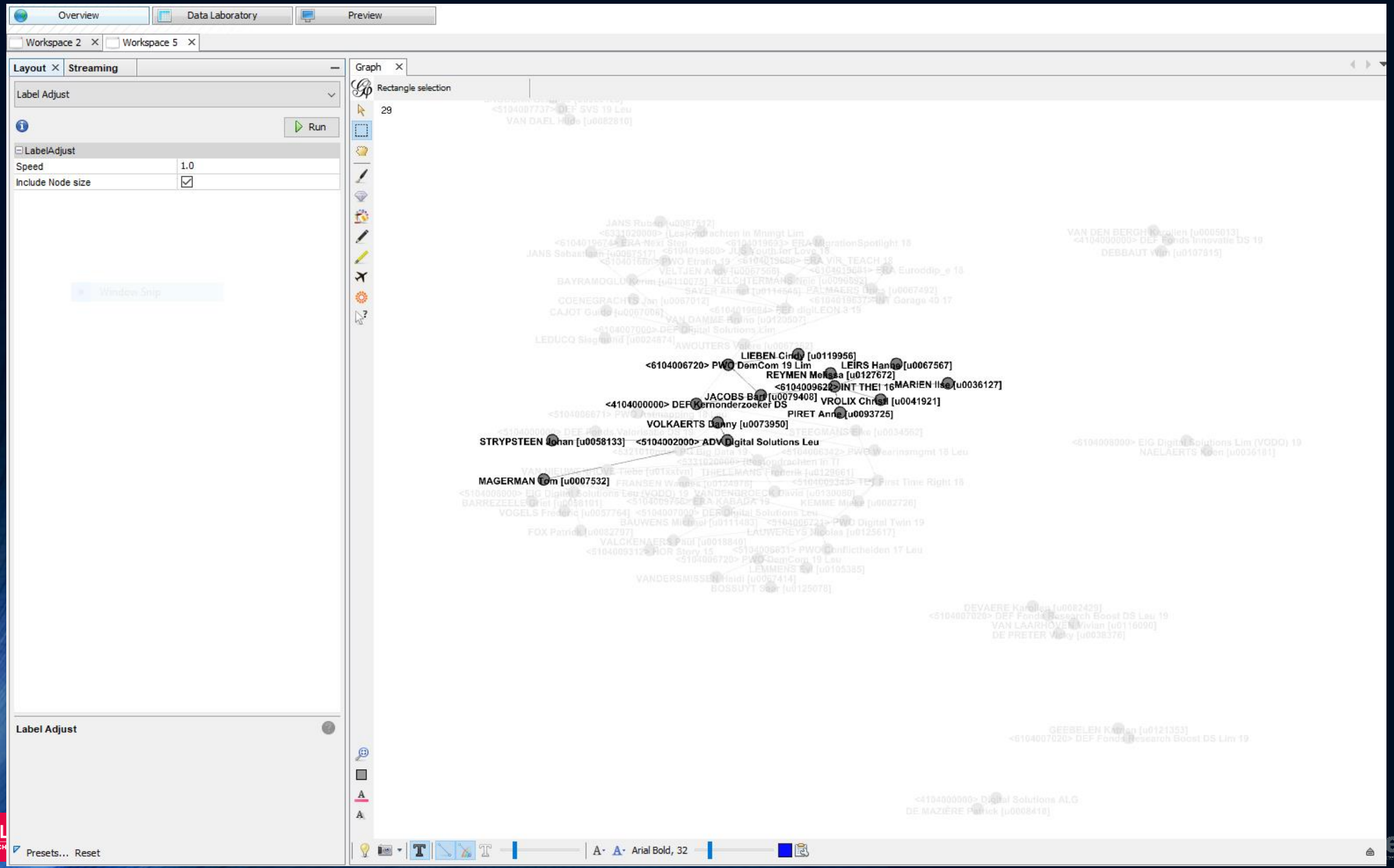
Highly Correlated Data with Palladio

- Easy to use, just copy paste your two-columned CSV data into the application and start
- For the example: Select Source & Target columns, choose Graph and you get your network
- Simple setup but also little tuning possibilities



Highly Correlated Data with Gephi

- More complicated, more possibilities as well
- Just load your XLS, with two sheets: one with all the labels and their IDs (*nodes*) and one with all the links represented as ID-tuples (*edges*).
- Check that "show labels of nodes" is active
- Select a visualisation algorithm, make it run and off you go


Overview
Data Laboratory
Preview
Workspace 2
Workspace 5
Layout
Streaming
Label Adjust
Run
LabelAdjust
Speed
1.0
Include Node size
Label Adjust
Presets... Reset
Graph
Rectangle selection
29
LIEBEN Cindy [u0119956]
<6104006720> PWO DamCom 19 Lim
LEIRS Hanne [u0067567]
REYMEN Melissa [u0127672]
<6104009622> INT THE! 16
MARIEN Ilse [u0036127]
JACOBS Bart [u0079408]
<4104000000> DEF Kernonderzoeker DS
VROLIX Chris [u0041921]
PIRET Anne [u0093725]
VOLKAERTS Danny [u0073950]
STRYPSTEEN Johan [u0058133]
<5104002000> ADV Digital Solutions Leu
MAGERMAN Tom [u0007532]
Arial Bold, 32

Advanced

You have to deal with…

- Complex correlated data
- Preprocesing and/or analysis before visualisation
- The whole thing to be run frequently upon arrival of new data
- …

- Contact Patrick.DeMaziere@ucll.be as UCLL R&E possesses a data analysis & visualisation platform that uses state-of-the-art software & hardware (TensorFlow, Keras, R, Python running on nVidia workstations) that can handle also machine learning or predictions