my first DM project

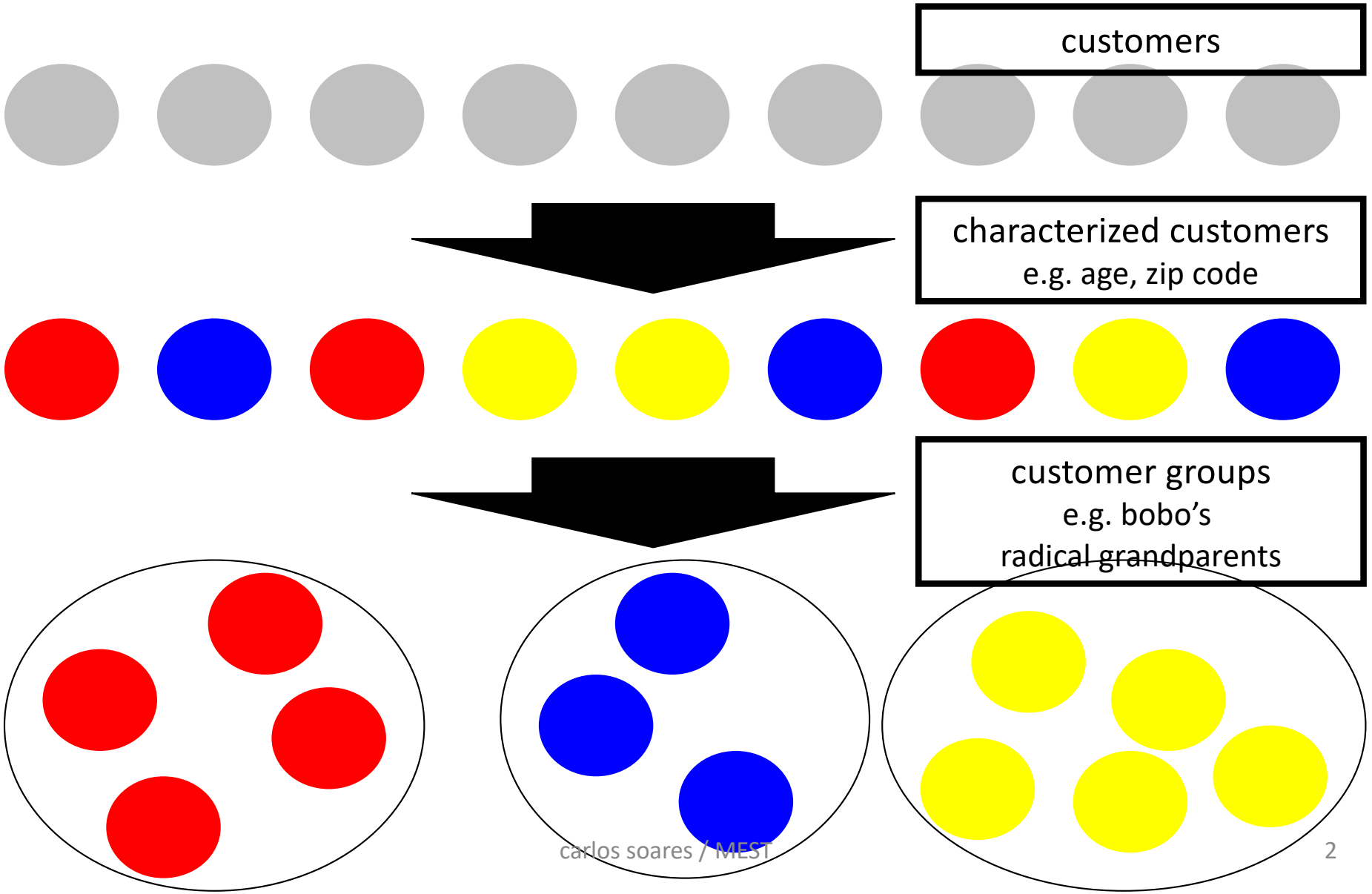
Carlos Soares

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

http://www.flickr.com/photos/upv
cfrontdoor/4642103837/

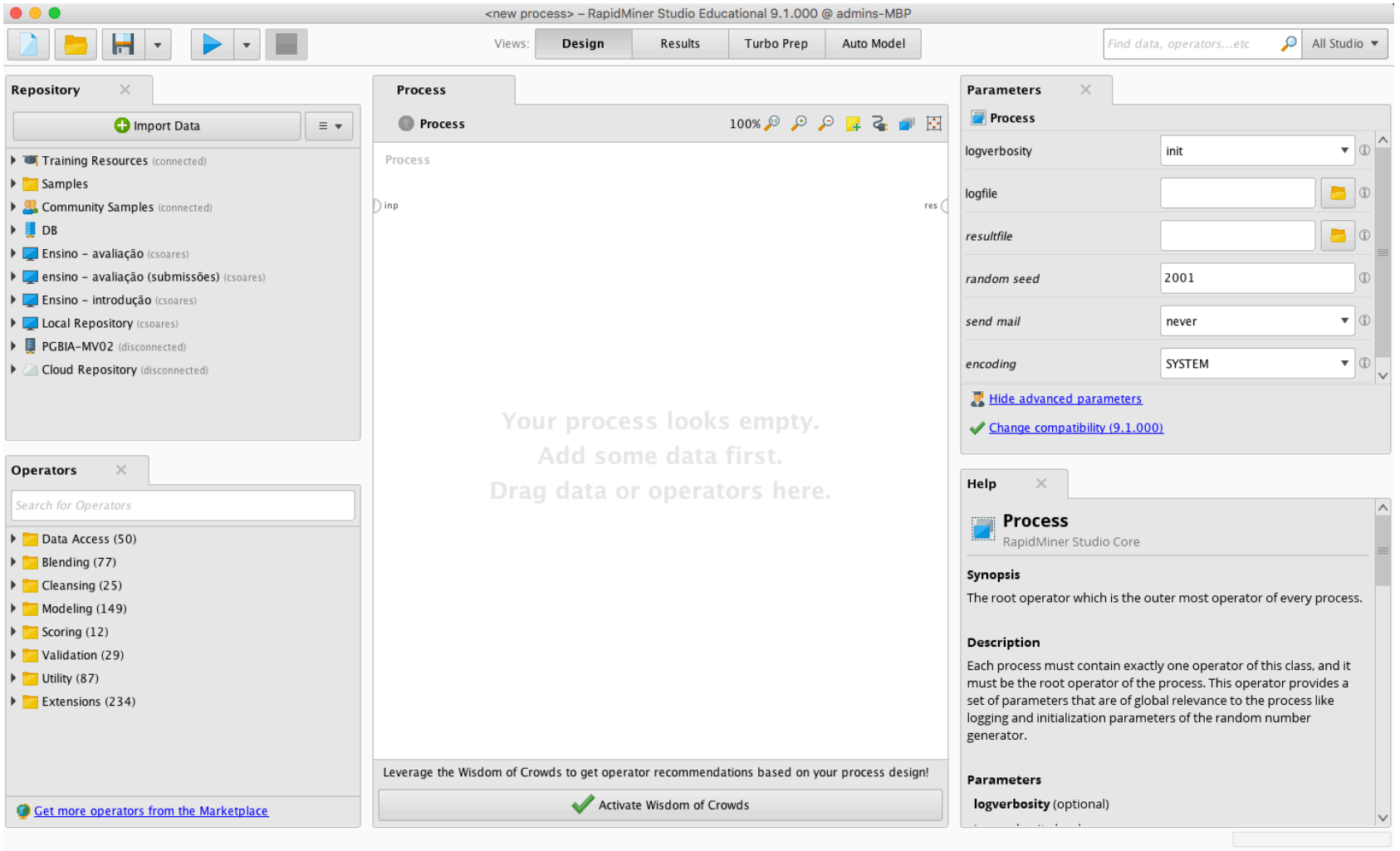
descriptive analytics: clustering for segmentation

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



but, first, introduction to rapidminer

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



hands-on

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- RapidMiner Studio
  - https://rapidminer.com/
- installation
  - http://docs.rapidminer.com/studio/installation/
- my academic license available
  - check moodle

disclaimer

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- RapidMiner UI changes too fast for me…
- … concepts are essentially the same

- in other words, screens will be outdated
- … but you can easily find your way

a more important disclaimer

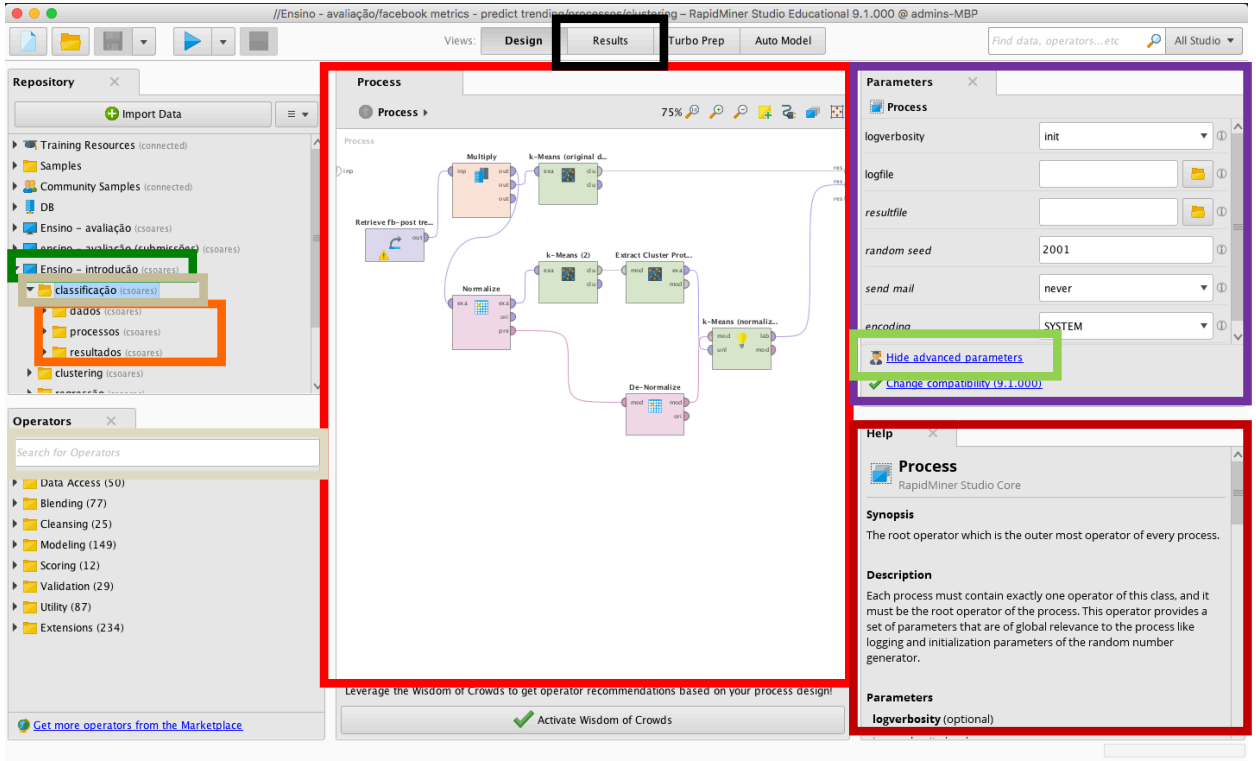
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- teaser for machine learning courses
- … meaning
  - very brief overview
  - some approaches are arguable
  - don't add "machine learning expert" to your linkedin profile just yet…
    - or "data scientist"…
- most importantly
  - understand the importance of statistics in modern business processes

rapidminer projects

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

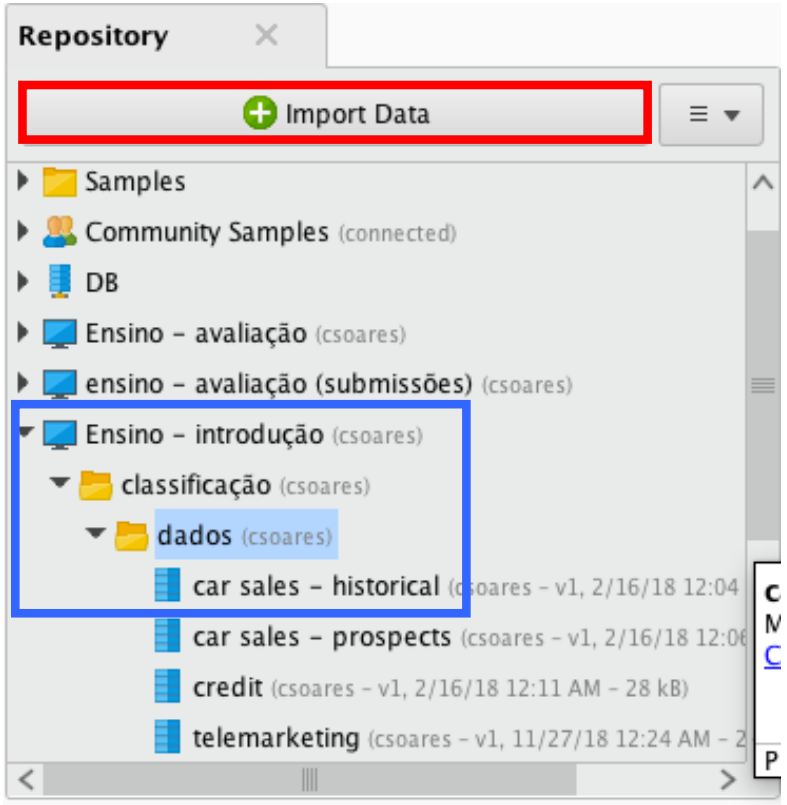
- workflow of operators
  - with parameters
    - including advanced ones
  - and help easily available
  - searchable
- repository of folders
  - e.g. repository = company
  - ... folder = project
- folders also used to organize projects
- results presented in a different view



… need data

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- data are imported to rapidminer
  - many different formats accepted
  - … including databases
- … wizards available
  - use carefully!
- … and stored in a folder



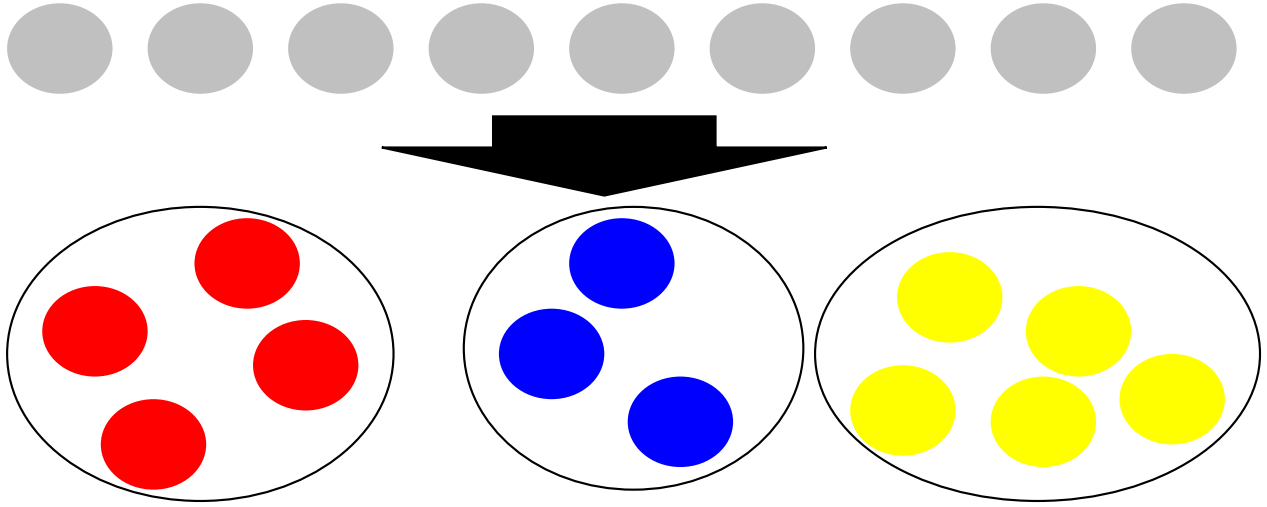
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

CLUSTERING

definition

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- partition set of objects
- ... elements within subgroups are similar among themselves
- ... according to one (or more) relevant characteristics



my first data mining project!

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

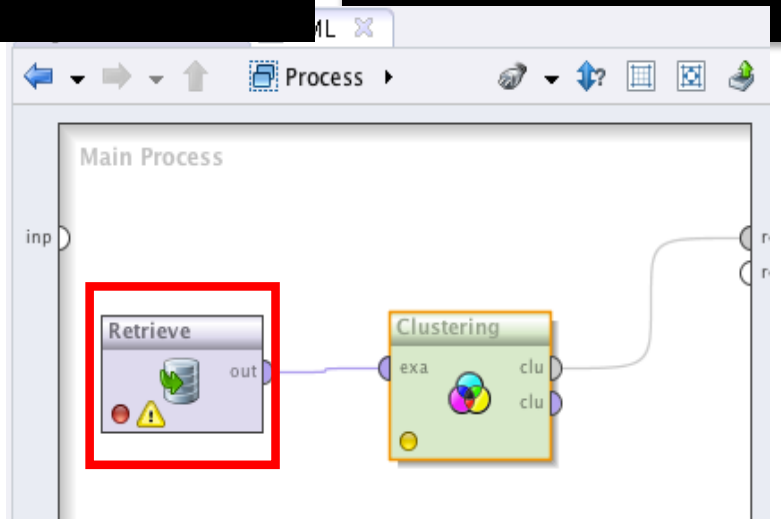
segment customers of a catalog sales company

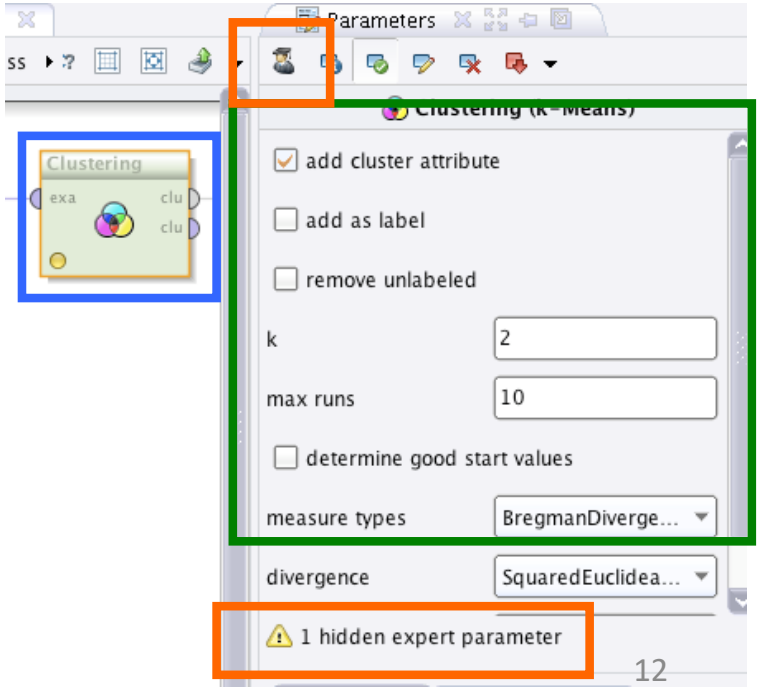
- goal
  - know customers better
- data available in file L01-first project.xlsx
  - (adapted from file provided with XLMiner)
  - worksheet “catalog”
- available customer characteristics (i.e. variable)
  - (worksheet catalog-dictionary)
  - ID, Total LTD Orders, Total 24 Month Orders, Number of Divisions with Purchase, Number of Credit Cards Used, Gender, Different Day/Night Phone, Dwelling Type Indicator, Overall RFM Points, First Purch Mail Order
- tool
  - rapidminer

clustering with rapidminer

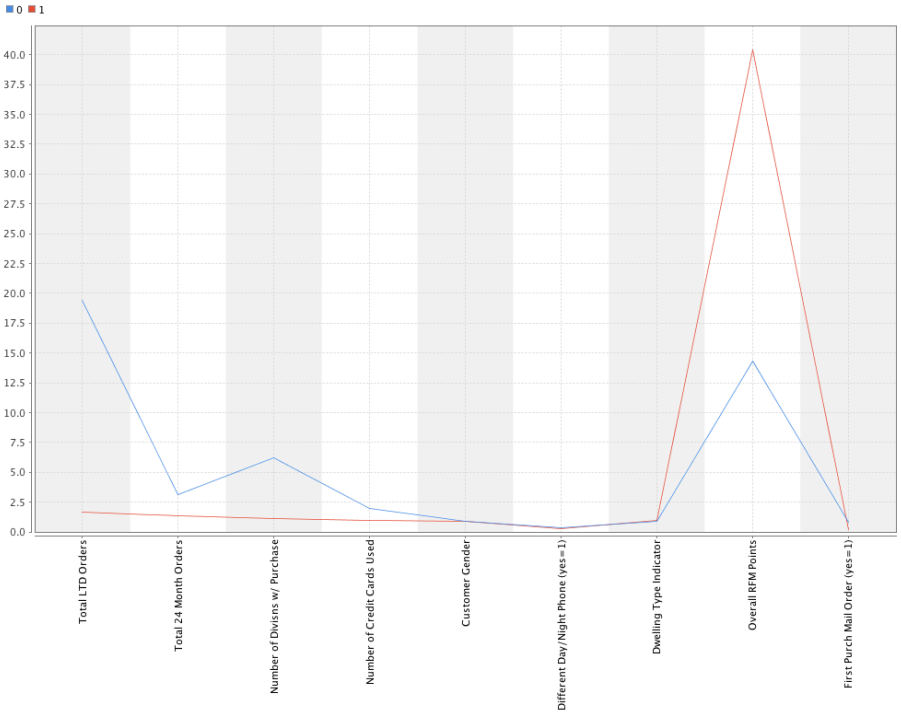
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

1. load data
   - drag DB from repository into workspace
2. insert kmeans operator
   - operators ➔ modeling ➔ clustering and segmentation ➔ k-means
3. set parameters
   - in expert mode
   - suitable for application
     - how?
4. run
5. interpret results





clustering results



- which variables distinguish the customers of the 2 segments?
  - what does the rest of this course tell us about this?
- describe each cluster in a short sentence