Projects for the course Predictive Modeling and Analytics

Prof Dr Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science

Predictive Modeling and Analytics
ESIGELEC, Rouen, November 2018

Project asignment

- Find a suitable problem, present it as a predictive modeling task and analyze it. Present findings.

The procedure

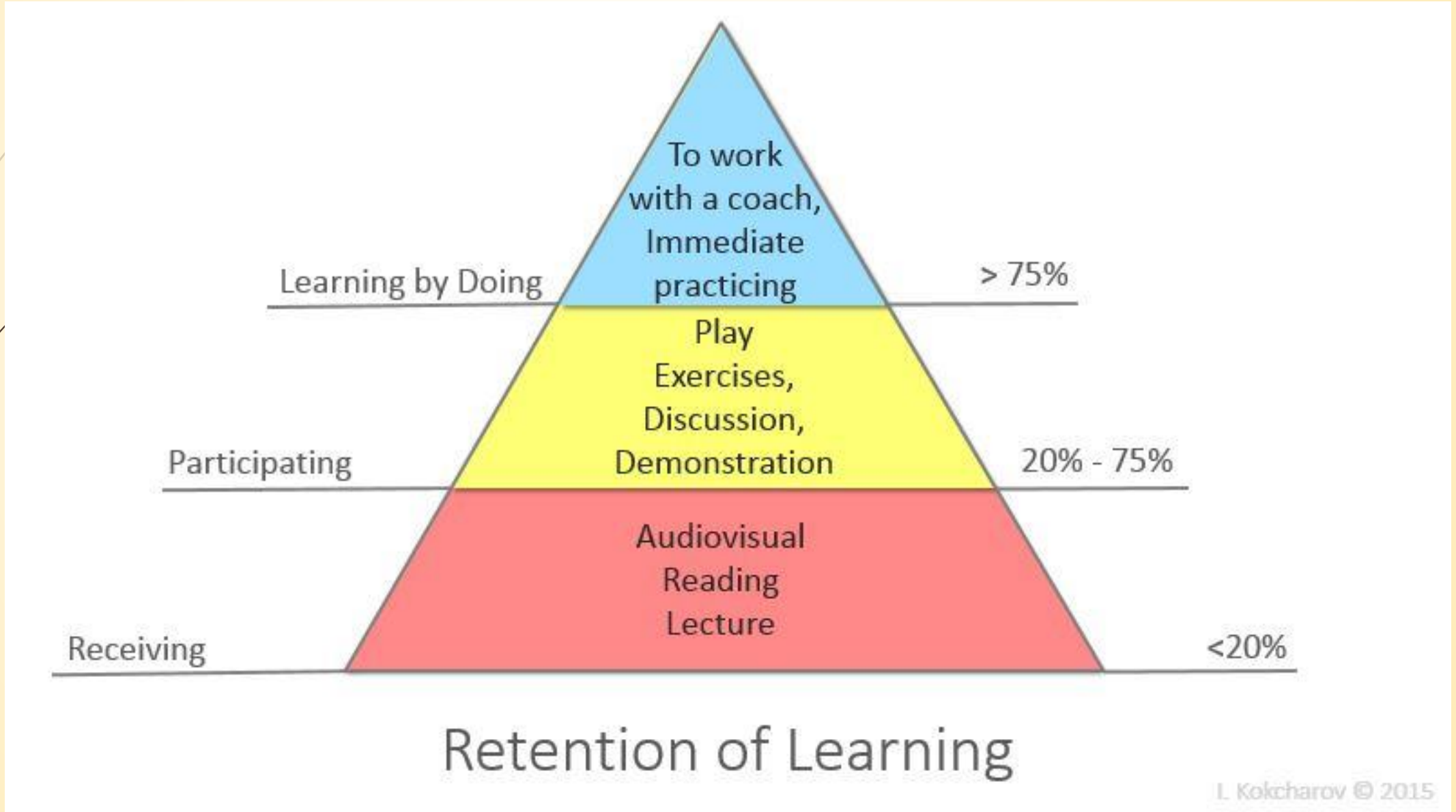
- form groups of three students
- each group finds its own interesting problem illustrated with an appropriate data set
- a few examples are presented on the third day
- by the fourth day each group prepares its own project proposal
- the project proposals are discussed individually with the instructor and have to be approved

- projects (problem and proposed solutions) will be publicly presented before the class (10% of points) by **all members** of the group on **Friday, Nov 9th, 10:30-12:30**
- written reports (4-6 pages, scientific paper format) together with source code (fully reproducible results) and presentation slides are handed in (20% points) by **Sunday, Dec 16th, 23:59**

Grading

Obligation	% of total	subject to
Daily assignment, 3 times x 5%	15%	≥ 7,5%
Project public presentation	10%	≥ 5%
Project report	25%	≥ 12.5%
Written exam	50%	≥ 25%

BTW: retention of learning



Aim of projects: build and evaluate models in R

The project shall clearly demonstrate that you can do the following tasks:

- visualize the data set and created models
- prepare data into a suitable form suitable for modeling algorithms
- select an appropriate modeling technique
- apply classification and/or regression models to solve a prediction task with a given data set
- estimate error of models using statistically valid approaches
- select models and tune their parameters using cross-validation and bootstrapping
- visualize models and explain their predictions

Format of the report

- 4-6 pages of A4 format, 11pt font size
  1. abstract: 200 words (background, motivation, results)
  2. introduction: background, clear statement of the problem, literature review, overview of the approach (approx. 1 page)
  3. description of the data set and its basic visualization (approx. 1 page)
  4. methods: explanation of the models and how they are applied to solve the problem (approx. 1 page)
  5. results: clear presentation of findings (approx. 2 pages)
  6. conclusion of the most important findings, directions for further work (approx. 1/2 page)
  7. references

- Grading: quality of the analysis, originality of the approach, complexity of the problem, readability, quality of the code (efficiency, legibility)

Format of the presentation

- 7 minutes for each group, all members do the presentation, slides are obligatory
- introduction of the group, declaration who was doing what
- problem description
- data set description
- description of the intended approach
- analysis of possible obstacles
- 3 minutes discussion, everybody takes part with questions, comments, and advise

Project ideas

- The best possible project is based on data science problem you are working on in practice/company/as a hobby...
- ... or one you recently met
- and have access to the data!

Other sources of interesting problems and useful data sets

- UCI ML repository http://archive.ics.uci.edu/ml/
- UCI KDD repository http://kdd.ics.uci.edu/
- Kdnuggets http://www.kdnuggets.com/datasets/index.html
- Kaggle https://www.kaggle.com/
- Awesome public data sets https://github.com/caesar0301/awesome-public-datasets
- Quandl https://www.quandl.com/browse
- Google dataset search https://toolbox.google.com/datasetsearch