# Introduction to data analysis with Python and R in Kaggle

Statistical Learning – Part II

Alberto Castellini
University of Verona

# *Kaggle, Python and R*

# *Kaggle*

**References:**

- Kaggle https://www.kaggle.com/
- What is Kaggle: https://www.kaggle.com/getting-started/44916
- Kernels: https://www.youtube.com/watch?v=FloMHMOU5Bs
- Learn: https://www.kaggle.com/learn/overview
  - Python: https://www.kaggle.com/learn/python
  - R: https://www.kaggle.com/learn/r
- A first data analysis case study: Titanic: Machine Learning for Disaster: https://www.kaggle.com/c/titanic

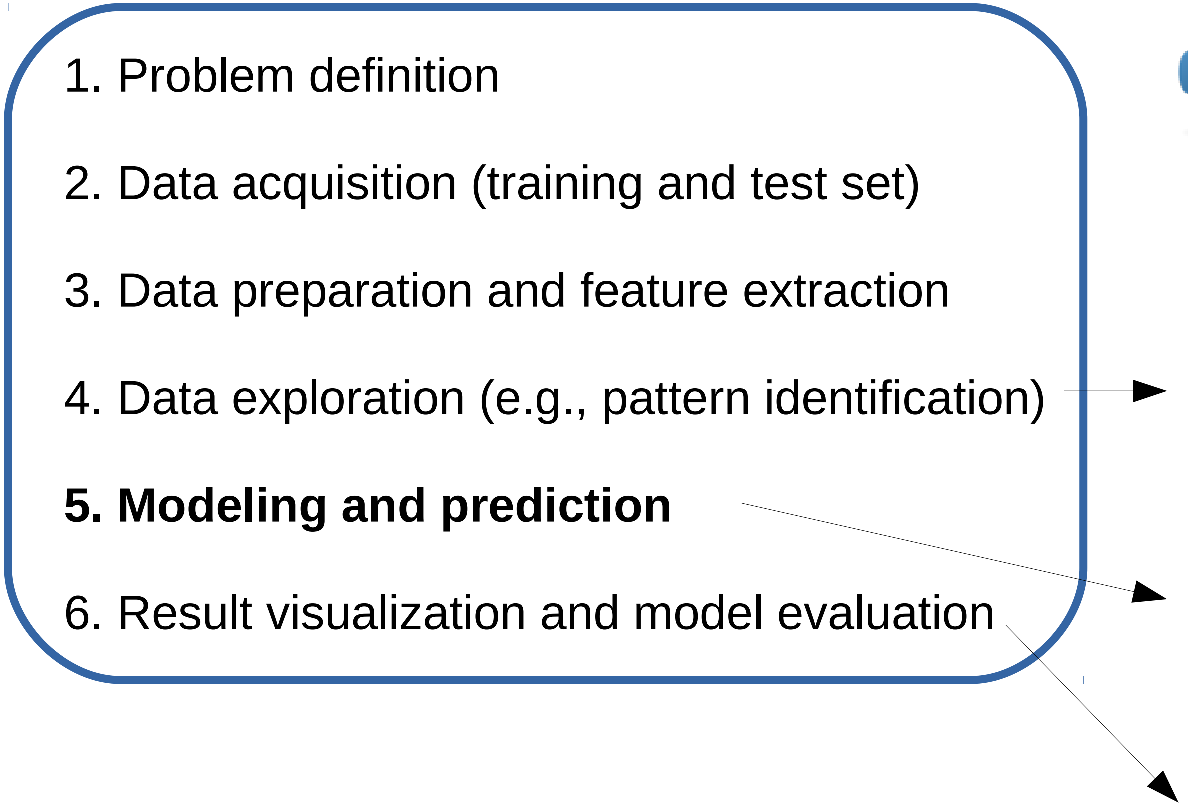
# Python

**References:**

- Kaggle’s tutorial for Python:https://www.kaggle.com/learn/python
- Material of Prof. Farinelli course (see section “Course Material” for slides and book references): http://profs.sci.univr.it/~farinelli/courses/python/python.html
- Python official site: https://www.python.org/
- Python documentation: https://docs.python.org/3/
- Python tutorial (pdf): http://www.cse.unsw.edu.au/~en1811/python-docs/python-3.6.4-docs-pdf/tutorial.pdf
- Spyder IDE: https://www.spyder-ide.org/
- NumPy (scientific computing): http://www.numpy.org/
- Pandas (data analysis): https://pandas.pydata.org/
- Seaborn (visualization): https://seaborn.pydata.org/
- Matplotlib (plotting): https://matplotlib.org/
- Scikit-learn (machine learning): http://scikit-learn.org/stable/

# *Data analysis process*

**Main steps of the data analysis process**



1. Problem definition

2. Data acquisition (training and test set)

3. Data preparation and feature extraction

4. Data exploration (e.g., pattern identification)

**5. Modeling and prediction**

6. Result visualization and model evaluation

**Main programming languages**

**Main focus of this course**

- k-means,
- PCA
- Spectral clustering

- Linear regression models
- Regularized linear regression
- **Logistic regression**

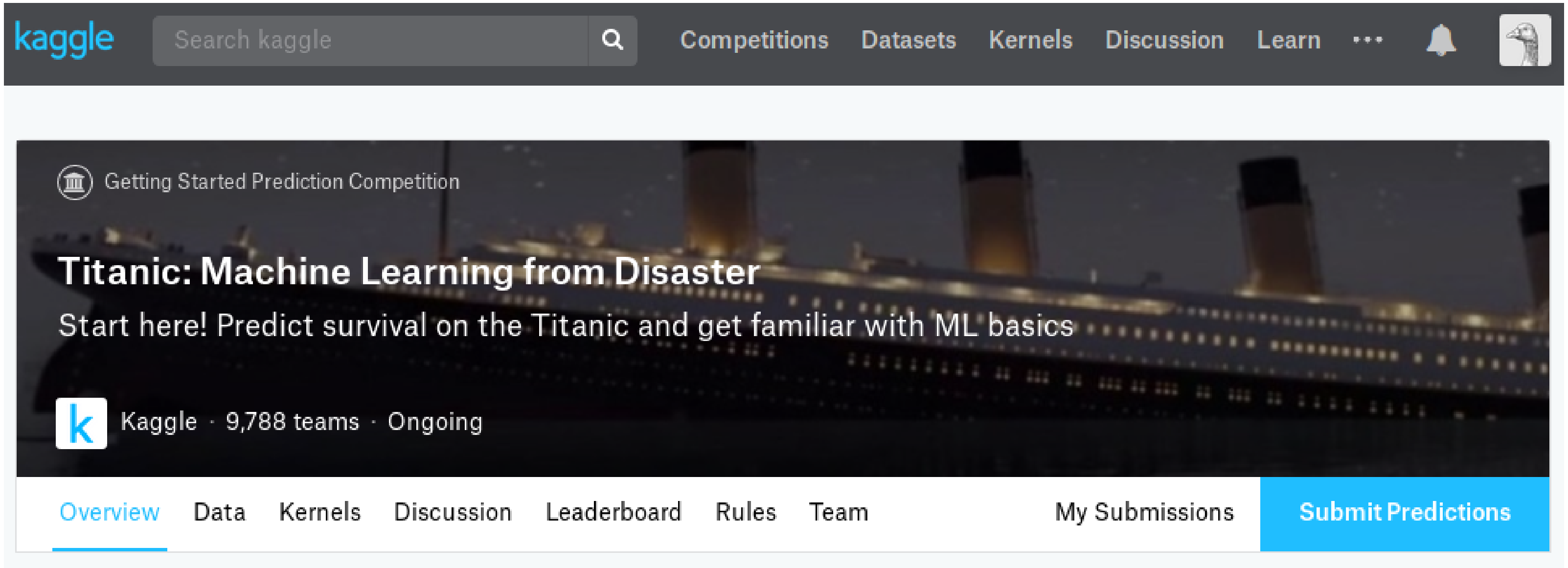
- Cross-validation
- Bootstrap

# *Intro to Kaggle and Python, analysis of the Titanic dataset*

**Introduction to Kaggle**
- Competitions
- Datasets
- Kernels
- Learn

**Introduction to Python**
- Tutorials
- References
- Practice

## A first example of data analysis project (in python)



Getting Started Prediction Competition

Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

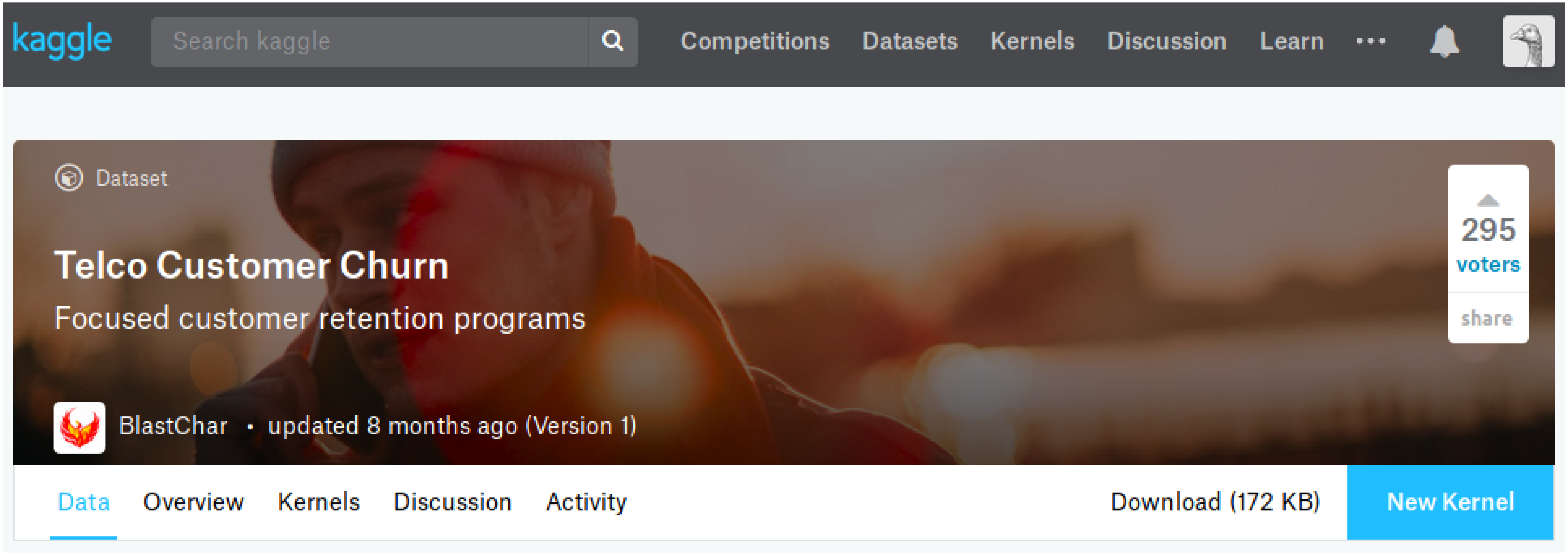
Kaggle · 9,788 teams · Ongoing

- Csv files, training/test sets
- Main libraries
  - Pandas
  - Numpy
  - Seaborn
  - Matplotlib
  - Sklearn)
- Data acquisition - read_csv()
- Dataframe (attributes and methods)
  - Shape, size
  - head(), tail()
  - info()
  - describe()

# Exercise: analysis of the Telco Customer Churn dataset

- We conclude the analysis of the Titanic dataset project together
- It's your turn… You will generate your first data analysis project

**Telecommunications Customer Churn analysis (in python)**



Telco Customer Churn

Focused customer retention programs

https://www.kaggle.com/blastchar/telco-customer-churn

See Exercise SL2019-20_L3_Exercise_Part1.pdf
in the E-learning (lecture 3)

# R programming language

**References:**

- R project for statistical computing (official website): https://www.r-project.org/
- R Studio IDE: https://www.rstudio.com/
- R documentation: https://cran.r-project.org/manuals.html
- R introduction manual: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Kaggle’s R tutorial (complete): https://www.kaggle.com/rtatman/getting-started-in-r-first-steps/
- Exploring the Titanic dataset in R https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic
- Other resources in Kaggle: https://www.kaggle.com/learn/r