**Pivigo Challenges Scope**

In this document we set the priorities for each of the challenges that we will propose to the community as an opportunity to help in the training and testing of the skills most in use in Data Science roles. We intend to bring together a testing/training + rewarding system as a ‘badges’ system, that would be used as a verification of the leverage of the different skills. There are many training/testing badges systems available online. We are going to be focusing our attention in building an integrated profile for a Data Scientist. This has to be reflected in each of the challenges that will be linked and very much DS ‘ flavoured’. This is the difference we will offer with respect to the rest of the challenges.

In the following we explain the scope of the skills/tools separated in broad areas as (by no means this is an extensive list):

1-Programming skills- Preferred languages: Python, R, SQL

2-Statistical analysis

3-Machine Learning tools

4-Soft skills (this could include contributions to the community/volunteer work/networking/English proficiency/presentation skills, although it’s difficult to test the results)

**1-Programming skills**

**Training level**

**Python:**

The basics of programming in python should include:

**-understanding of data types and structures (lists, tuples, sets, dictionaries, strings, etc)**

Examples in ipython--

**types of data: numbers/strings/lists** http://www2.mpia-hd.mpg.de/~robitaille/PY4SCI\_SS\_2015/\_static/04.%20Numbers%2C%20String%2C%20and%20Lists.html

**booleans/tuples/dictionaries:** (http://www2.mpia-hd.mpg.de/~robitaille/PY4SCI\_SS\_2015/\_static/05.%20Booleans%2C%20Tuples%2C%20and%2C%20Dictionaries.html)

**-being able to create simple functions:**

examples:

* + Write a function to return the area of a square given a side length

-Write a function to compute metres, given feet and inches (covering I/O, string manipulation)

-Fizzbuzz

-use of already defined functions as len, sum, int, etc

**-control flow- if, for j in [..]., while, break, scope, etc**

**-reading/writing data.** Example**:** populate a random list and write an algorithm to sort it

**-slicing data**

**-ipython, jupyter notebook**

**Advanced Python:**

-being able to create a dataframe

-classes, person\_class, objects

-understanding the different use of libraries as pandas, scipy, numpy, matplotlib.

<http://pandas.pydata.org/pandas-docs/stable/10min.html> (use as example of exercise)

-basic plotting with the libraries .-ggplot

**R:**

The Rstudio and the R books links to the statistics, can be combined exercises for R programmers.

**Basic SQL:**

* Select all data
* Select specific columns
* Select data Where a specific column meets a condition
* Inner join
* Orderby

**Advanced SQL**

* Outer join
* Groupby
* Nested queries

**2-Statistical Analysis**

**Training Level**

-understanding of variables and probabilities, binomial, poisson, gaussian . distributions,mean, variance, covariance.

-statistical modelling : linear and multiple regression

-statistical tests (Fisher, Student, Kolmogorov?)

-p-values

-montecarlo methods

-MCMC

-Bayes (or in ML)

-Maximum Entropy/ Information theory

Examples-

**3-Machine Learning Tools**

**- Training Level**

**http://www.cs.cmu.edu/%7Etom/mlbook.html**

**supervised/unsupervised algorithms**

Basic-

Linear regression

K-means classification

Decision trees

Advanced-

Regularisation

Polynomial feature engineering

Neural-networks

Random forest

-regression

-classification

-confusion matrix

What do you consider basic from this list?

http://www.cs.cmu.edu/%7Etom/10701\_sp11/hws.shtml

-https://www.coursera.org/learn/machine-learning#syllabus