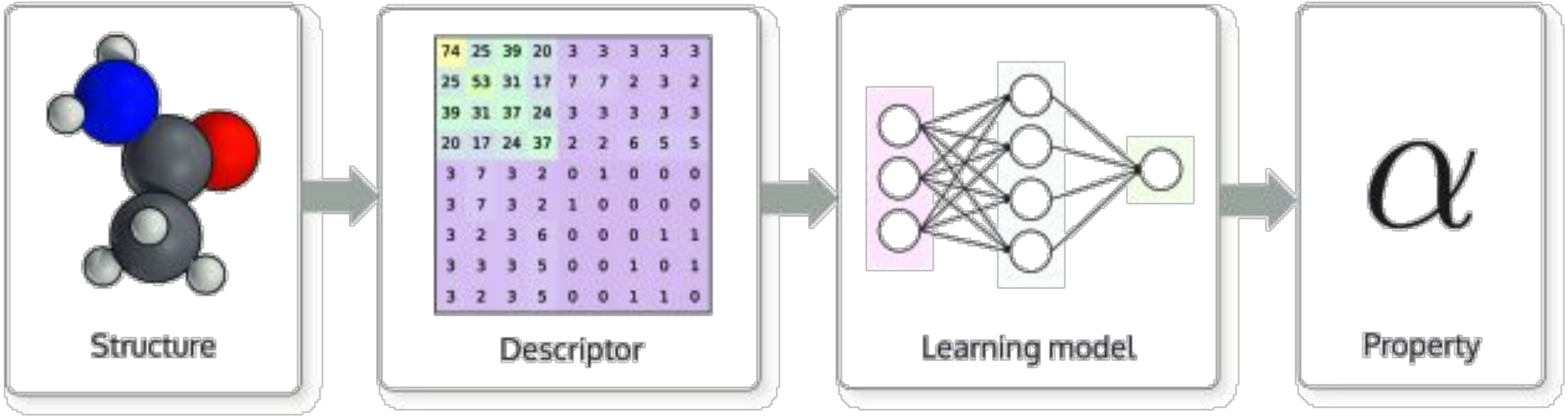
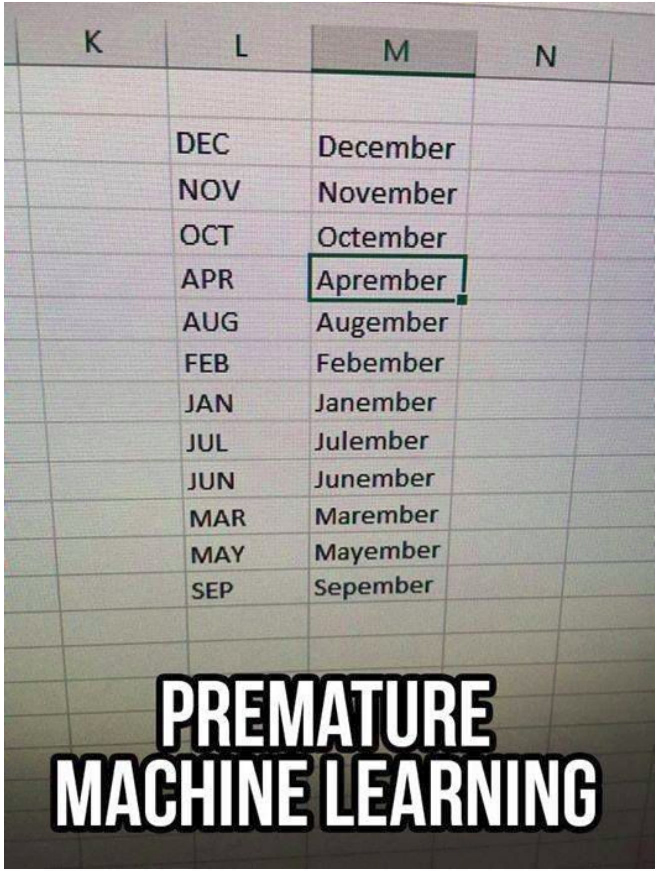
A tool for descriptor calculation in R

Ronaldo C. Prati (CMCC/UFABC)
ronaldo.prati@ufabc.edu.br

The machine learning pipeline



74	25	39	20	3	3	3	3	3
25	53	31	17	7	7	2	3	2
39	31	37	24	3	3	3	3	3
20	17	24	37	2	2	6	5	5
3	7	3	2	0	1	0	0	0
3	7	3	2	1	0	0	0	0
3	2	3	6	0	0	0	1	1
3	3	3	5	0	0	1	0	1
3	2	3	5	0	0	1	1	0



K	L	M	N
	DEC	December	
	NOV	November	
	OCT	Octember	
	APR	Aprember	
	AUG	Augember	
	FEB	Febember	
	JAN	Janember	
	JUL	Julember	
	JUN	Junember	
	MAR	Marember	
	MAY	Mayember	
	SEP	Sepember	

PREMATURE
MACHINE LEARNING

Descriptor Engineering

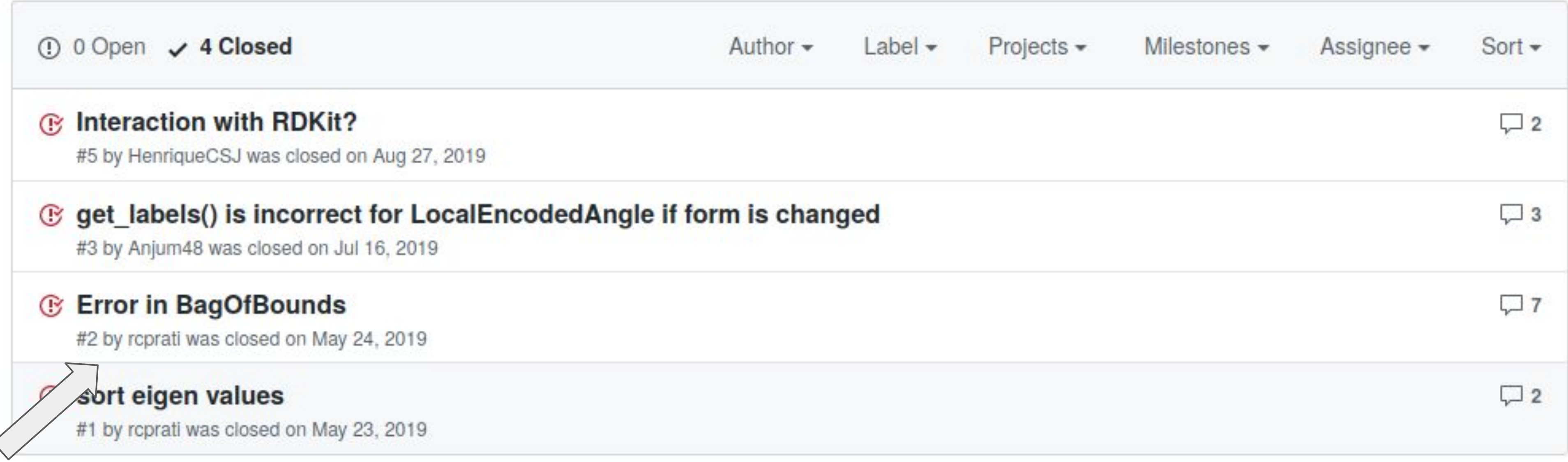
- Very hot topic of research in the application of machine learning to materials science
- Many research groups around the world are developing tools for descriptor engineering
- Good descriptors and quality and sufficient data allows us to build good predictive models (after an experimental research for a good algorithm)

MolDescriptoR

- A package build using R programming language for computing descriptor from molecules
    - Flexibility
    - Extensibility
    - Based on built in R functions

Why another package?

- It helps in getting a deeper understanding on how these methods work
- It helps in playing around with new ideas
- It opens the opportunity to use different set of tools outside the python ecosystem (and R is very fruitifull for that)
- It provides flexibility. You may try to combine different techniques easily with a consistent interface



Interaction with RDKit? 💬 2
#5 by HenriqueCSJ was closed on Aug 27, 2019

get_labels() is incorrect for LocalEncodedAngle if form is changed 💬 3
#3 by Anjum48 was closed on Jul 16, 2019

Error in BagOfBounds 💬 7
#2 by rcprati was closed on May 24, 2019

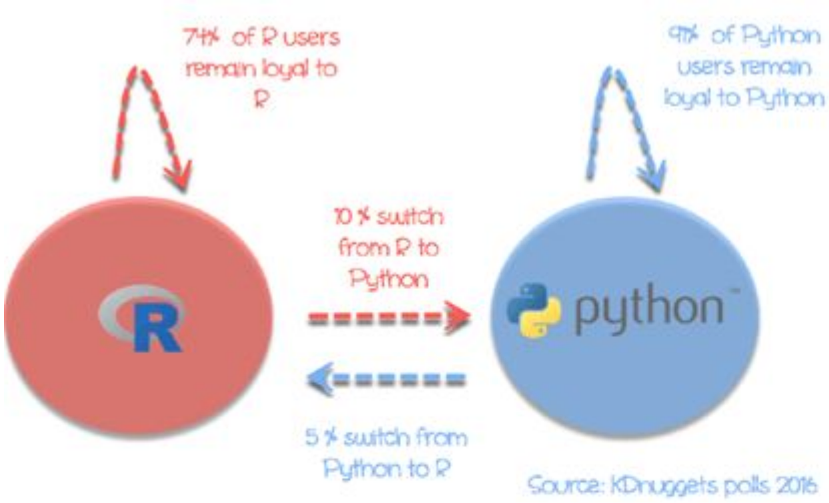
sort eigen values 💬 2
#1 by rcprati was closed on May 23, 2019

What is R?

- Programming Language and an open-source environment for statistical computing
- Widely used among statisticians and data scientists for developing software and data analysis
- Runs in many OS and on cloud
- A package system with 15371 contributed packages
- Publishing quality graphical capabilities

Why not Python?

- There are many tools available for python already (even though different groups around the world are developing their own)
- In some sense, they have complementar tools (many packages are available in only one environment)



74% of R users remain loyal to R
91% of Python users remain loyal to Python
10 % switch from R to Python
5 % switch from Python to R
R
python

Source: KDnuggets polls 2016

The package

- Input of data uses CDK format (thanks to rcdk package)
- It uses flexible functions for computing norms and density estimation, allowing the selection of several parameters and options not available in other packages
- Extensible

Implemented methods (so far)

State-of-the art methods implemented so far

- Coulomb-matrix (with random, norm-sorting, and eigenvalues sorting)
- Bag-Of-Bonds
- Many-body Tensor Representation (up to 4-body terms, feature not available in many packages)
- Smooth Overlap of Atomic Positions (under development)

Planned Features

- Automatic cut-off radii determination by using Voronoi and Delaunay cells (not available in other packages)
- Merging of different descriptors using multi-view learning
- Release as an open-source package