April 3rd, 2015

PyData

Paris 2015

Reaching your DREAMs with Python

Chloé-Agathe Azencott

Systematic

Paris Region Systems & ICT Cluster

Num FOCUS

Open Code, Better Science

Centre for Computational Biology

MINES ParisTech

Inserm
Institut national
de la santé et de la recherche médicale

institutCurie

Machine learning for computational biology & cancer research

Precision medicine

- Treatment is **adapted to the (genetic) specificities** of the patient.
- **data-based** biology/medicine:

  identify similarities between patients that exhibit similar susceptibilities / prognoses / responses to treatment.

DREAM Challenges

http://dreamchallenges.org

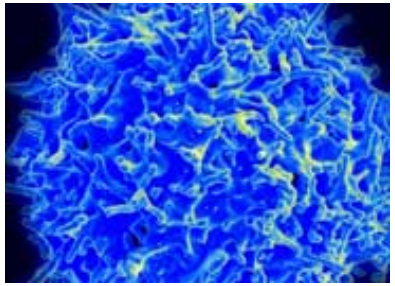
Dialogue for **R**everse **E**ngineering **A**ssessments and **M**ethods

Open science, crowdsourcing challenges examining questions in biology and medicine.

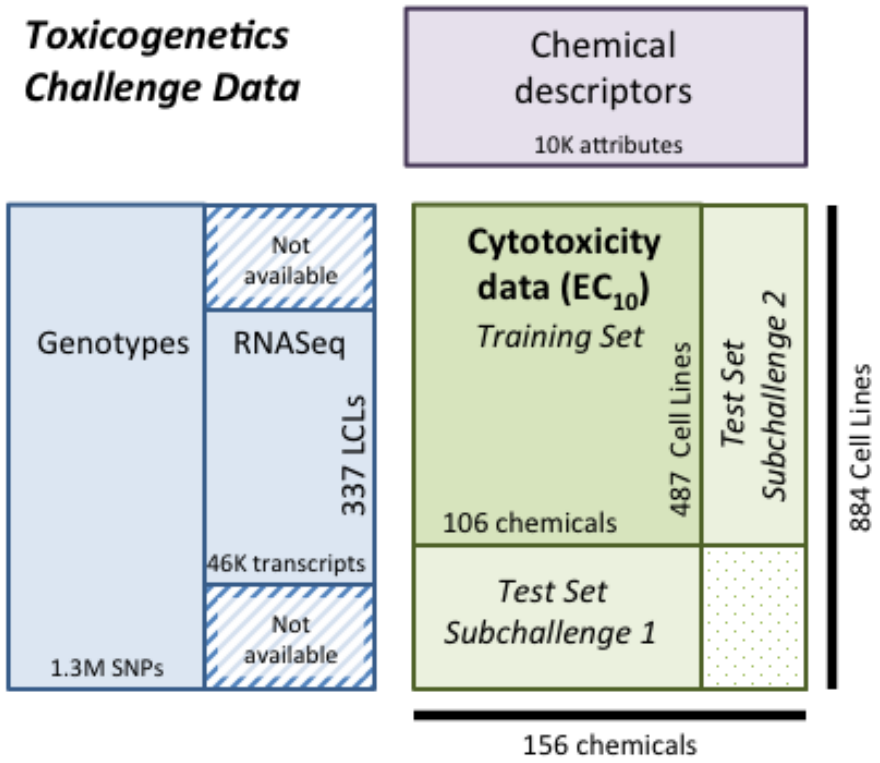
DREAM Challenges

"As the **volume and complexity of data** continues to increase, it is critical to develop **new methods** to use data to address fundamental questions to better understand and improve **biological sciences** and **human health**."

DREAM 8 (2013)



The NIEHS-NCATS-UNC Toxicogenetics challenge

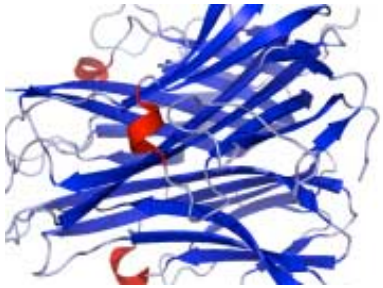


Subchallenge 1:
Predict the toxicity of known compounds on new cell lines.

Subchallenge 2:
Predict the toxicity of new compounds on known cell lines.

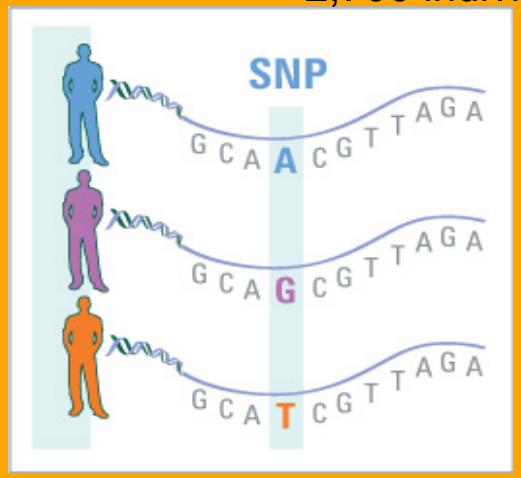
Team MLCB placed 2nd on SC2

DREAM 8.5 (2014)



The Rheumatoid Arthritis responder challenge

Genotype Data 2,5 million SNPS
2,706 individuals



Clinical Data 6 treatments

- Age
- Sex
- DAS before treatment

Outcome Data

- ΔDAS
- Responder or not

Goal: Predict response to treatment.

Team Lucia placed 2nd in the 1st phase

Image courtesy of Víctor Bellon.

Tasks

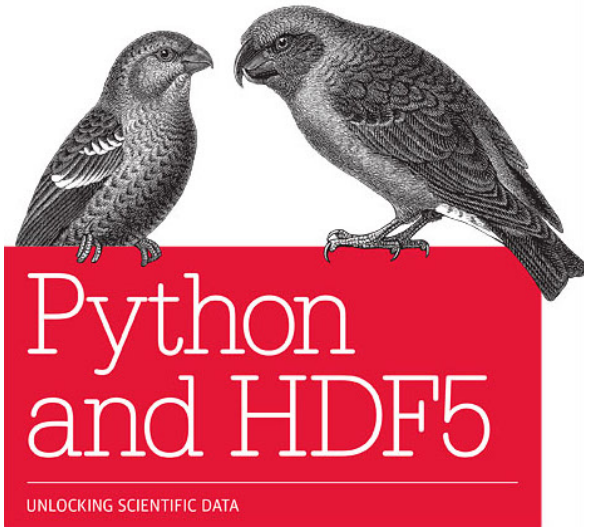
- Data **pre-processing**;
- **Feature selection** algorithms, both **classical** and **in-house**;
- **Classification/regression** algorithms, both **classical** and **in-house**;
- **Plotting** data & results.

Data pre-processing

- **hdf5 format**
  - allow team members to work with **diverse platforms** and **languages**;
  - **multiple views** of the data.

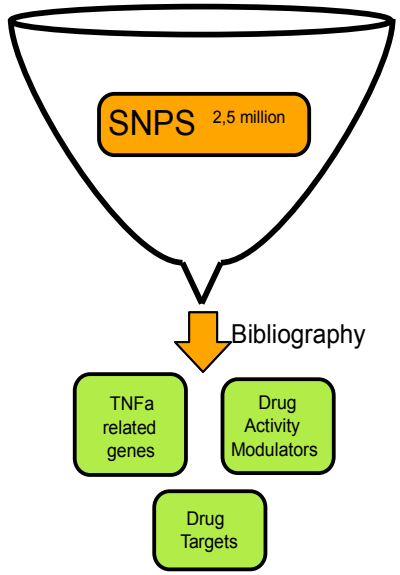
```
import h5py
import numpy as np
import tables
```

- (**normalization**);
- missing data.

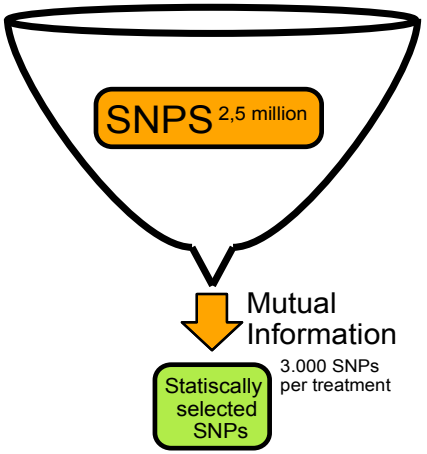


Feature selection

- **Knowledge-based**
  - **ad-hoc scripts** to extract e.g. features within a certain distance of genes of interest;
  - re-use of scripts we had written before.



- **Statistical**
  - `from sklearn import feature_selection`
  - Mutual information.



Images courtesy of Víctor Bellon.

Classification & Regression

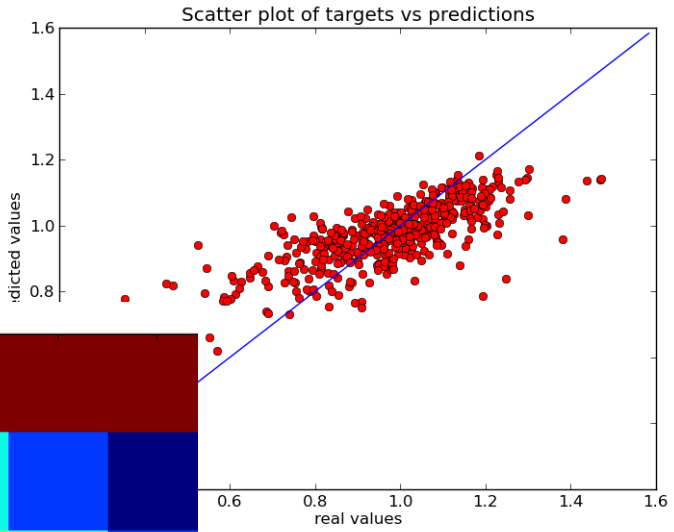
- `scikit-learn`
  - random forests, lasso approaches, nearest neighbors, SVMs and more;
  - custom kernels/distances (e.g. Tanimoto, Kronecker products);
  - `from sklearn import cross_validation, metrics`

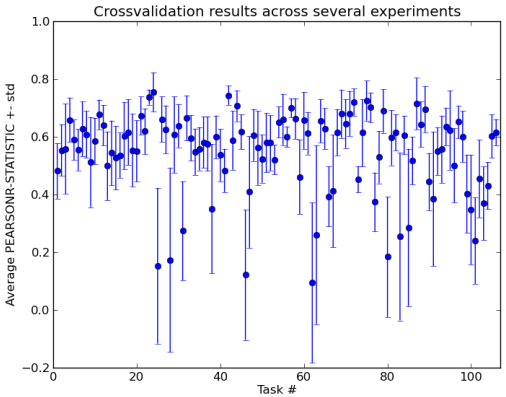
- `GPy`
  - Gaussian Process Regression;
  - custom multi-task formulation.

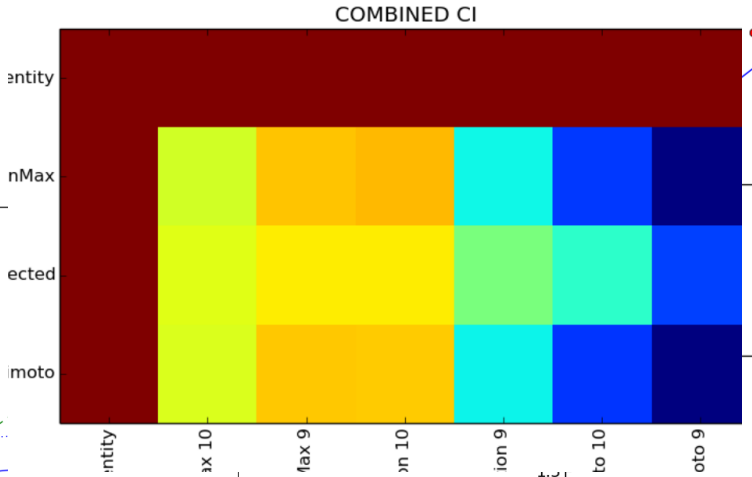
Plotting data & results

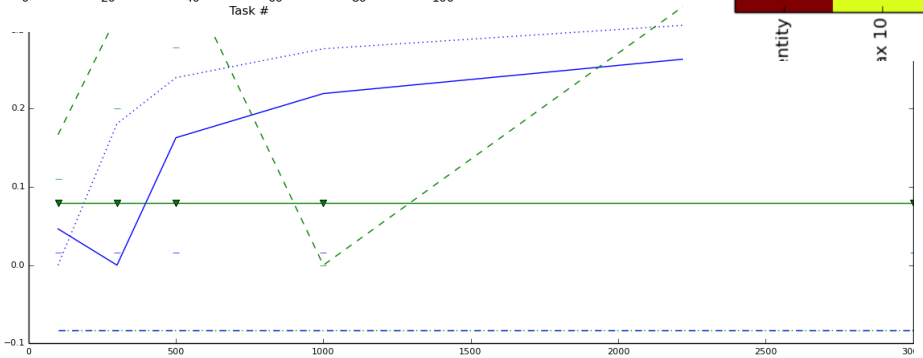
`matplotlib`

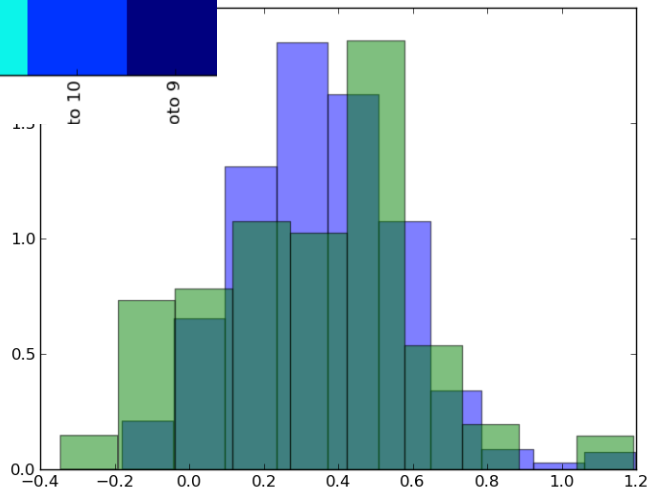
- **Understand** the data;
- Choose the **best algorithm(s)**.











Images courtesy of Víctor Bellon & Felipe Llinares López.

Thanks to...

Felipe Llinares López and **Dominik Grimm**

(Team MLCB)

Víctor Bellon (Team Lucia)

Image source: woodleywonderworks on Flickr.