**WorkFlow**

----- Select positive patients in Sheffield PAH ---------------

* Pull out Sheffield iPAH and PAH
* Iterate 2 times on:

1. select features (i.e. on non-missing values density)
2. SMOTE algorithm to create synthetic data
3. train XGBoost on synthetic data
4. test on XGBoost real Sheffield data

* decide about the threshold to use according to test
* Apply model to the PAH full dataset to extract iPAH\_pred
* Do some comparisons between iPAH and iPAH\_pred (i.e. looks if homogeneity of each group is comparable to the homogeneity of the union)

----- Building Model------------

-Model using iPAH and negative cohort (from Sheffield) these are pure data (these datasets used for iPAH\_pred, **we cannot combine this model with models that use iPAH\_pred**).

-Model using iPAH+ iPAH\_pred + negative cohort

Real Time Estimate: 7 days without iteration, 9 days with iteration

Proposed Time Estimate: 3 weeks? (it should be less but it will be at the beginning of the project, unrelated issues might arise and it is the first time we use a synthetic data approach).

References to SMOTE

<http://contrib.scikit-learn.org/imbalanced-learn/generated/imblearn.over_sampling.SMOTE.html>

SMOTE (<http://www.inside-r.org/packages/cran/DMwR/docs/SMOTE>)

ROSE (<http://cran.r-project.org/web/packages/ROSE/ROSE.pdf>)

(paper at the bottom of the page)