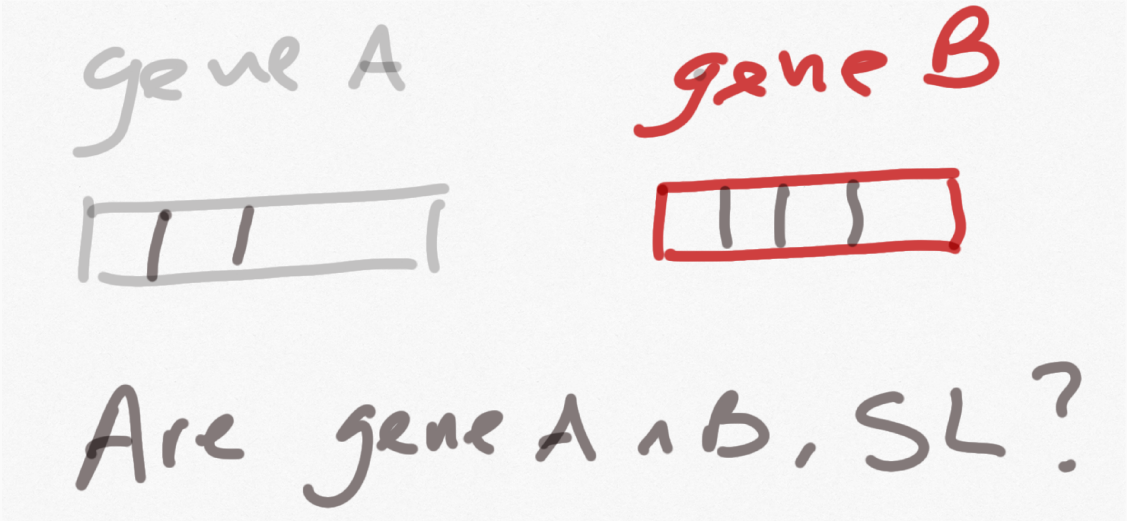
Predicting yeast synthetic lethal genetic interactions using protein domains-Code Replication

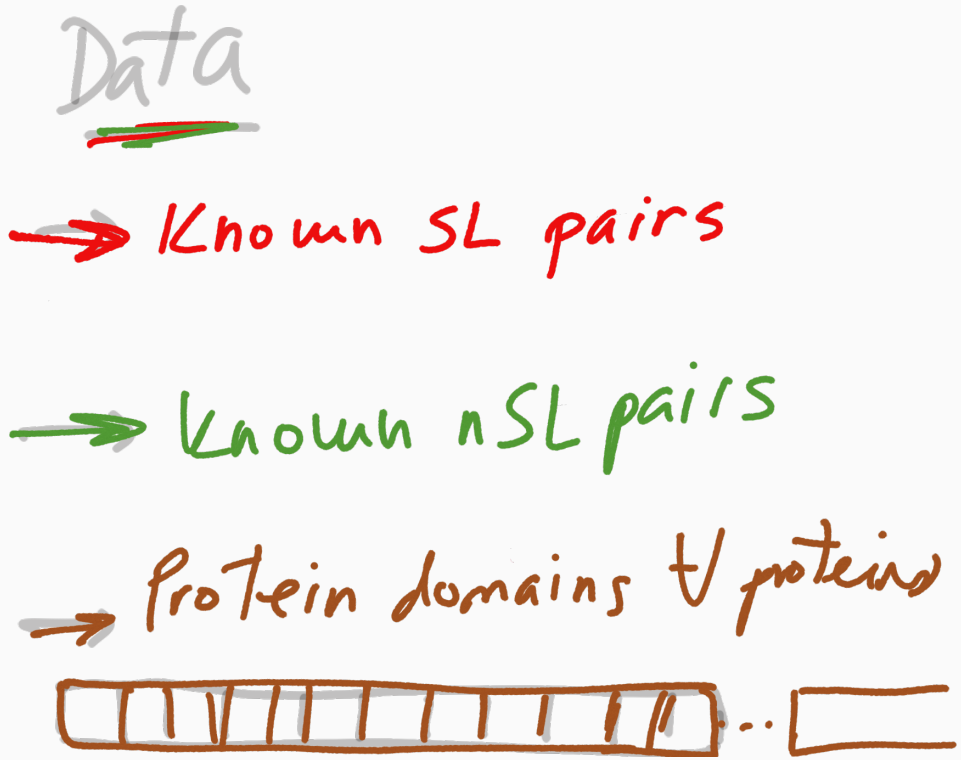
The problem :

Predicting synthetic lethality pairwise interaction upon knowledge of protein domains of the pair



What do we need first?

DATA!!!

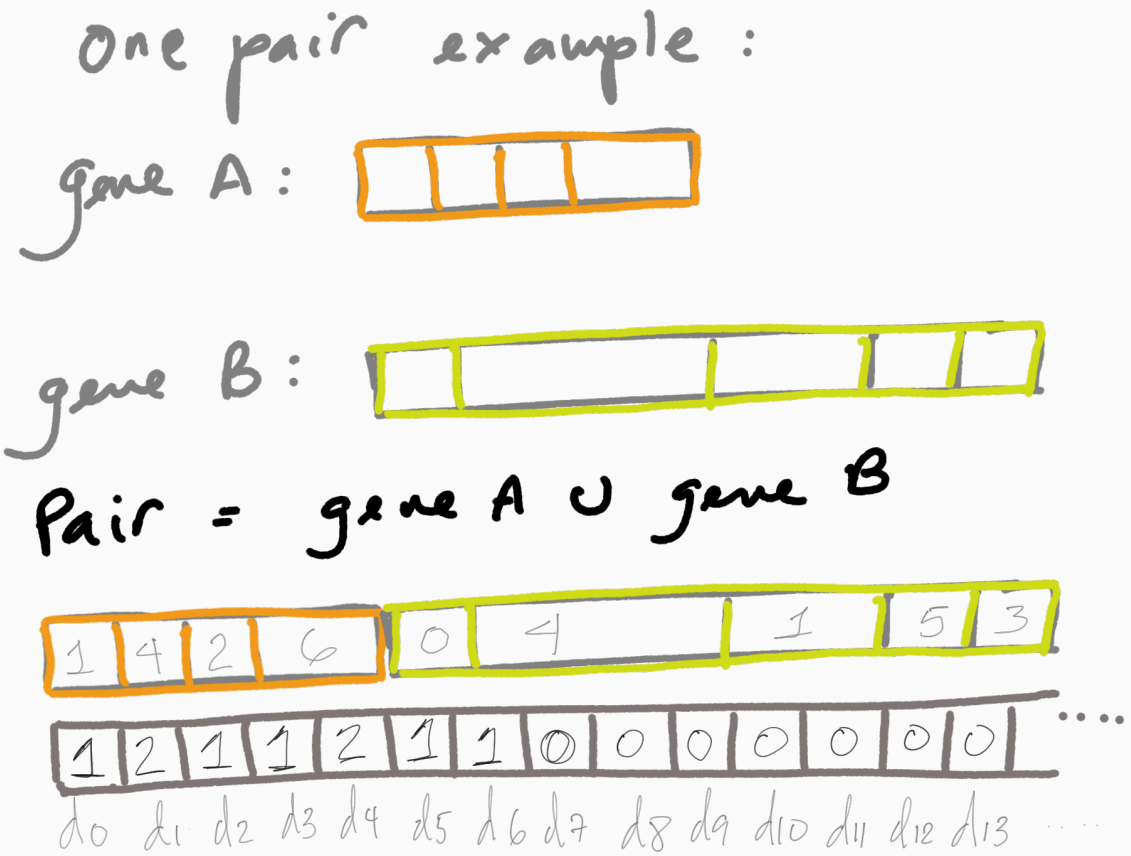

Data
Known SL pairs
Known nSL pairs
Protein domains & proteins

- All the current knowledge on yeast genetic interaction is in **BioGrid**

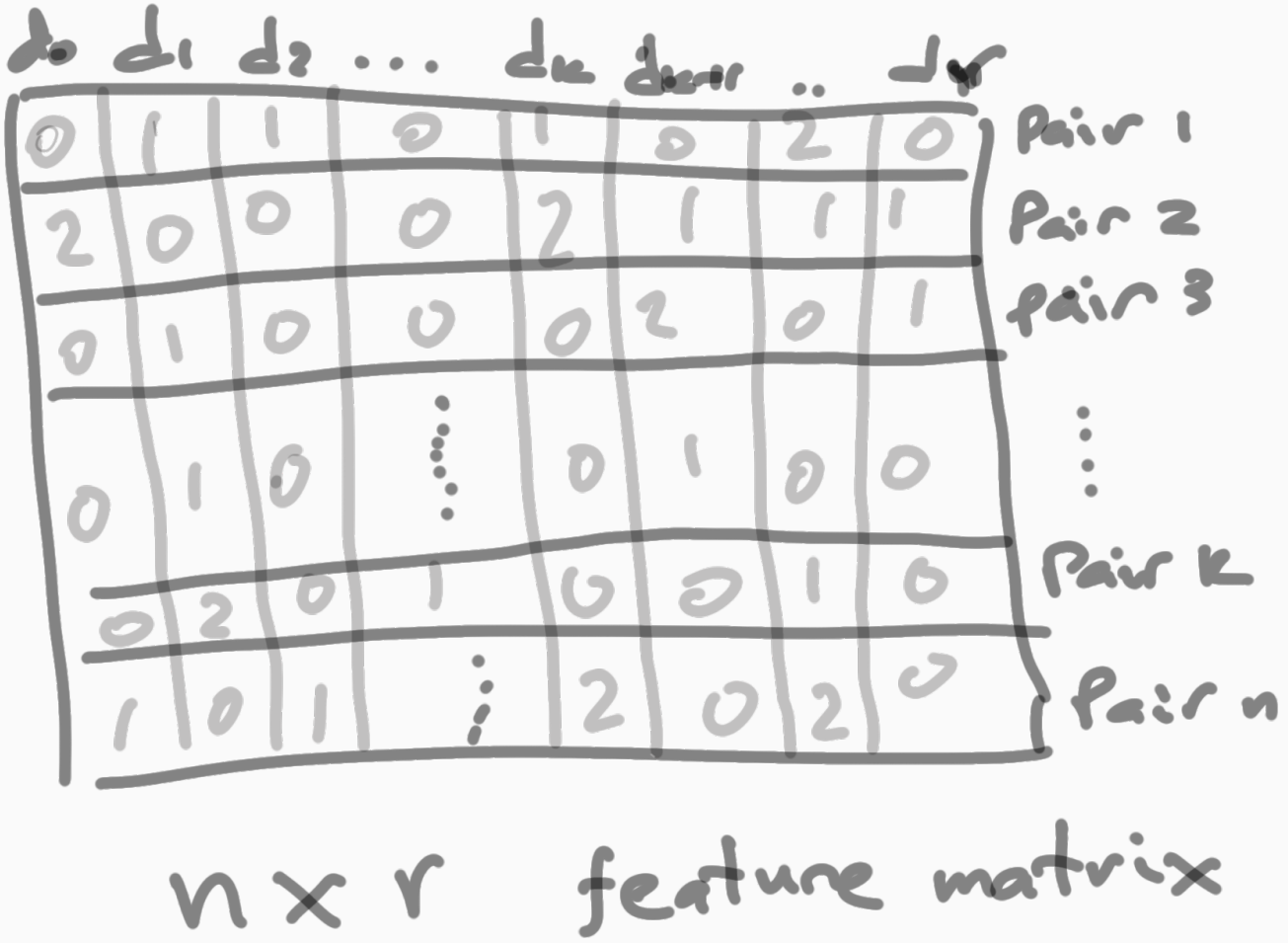
What do we need 2nd?

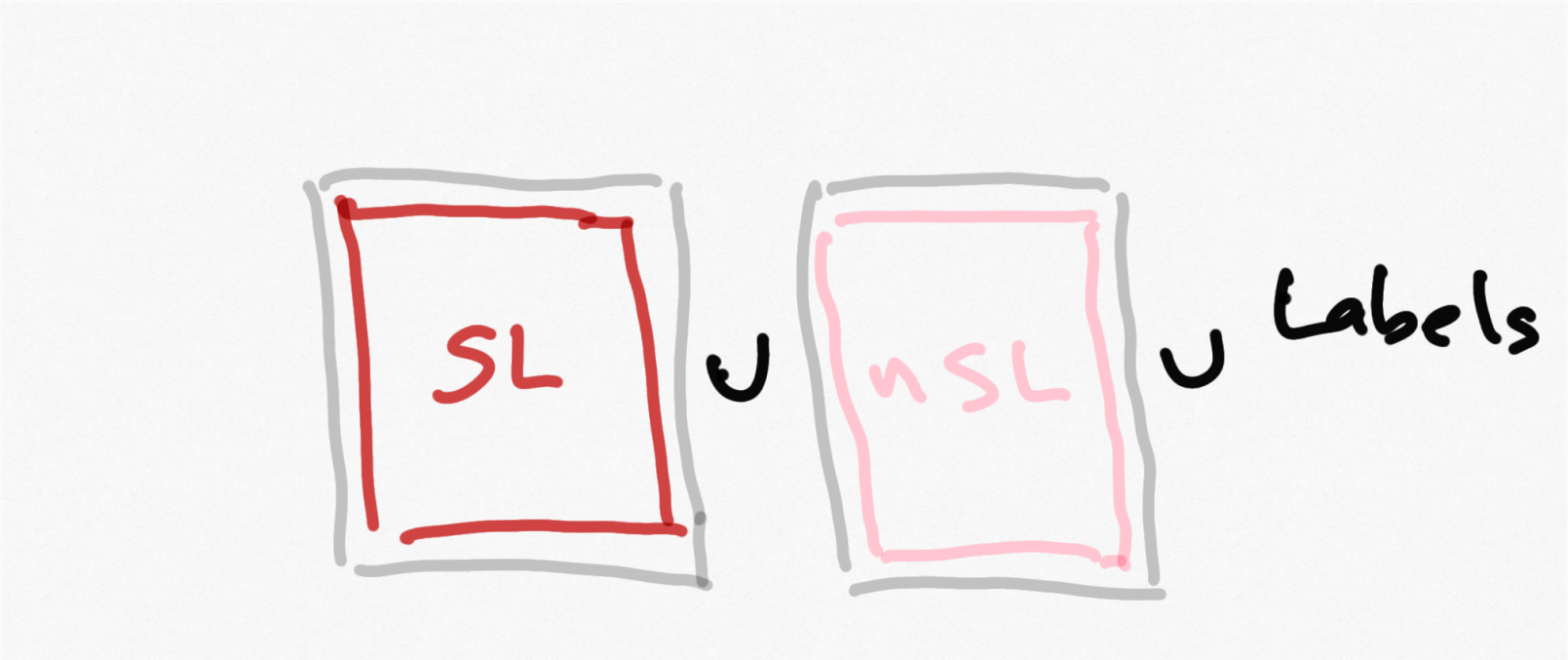
- to implement the features of the problem in order for the method to “learn” from

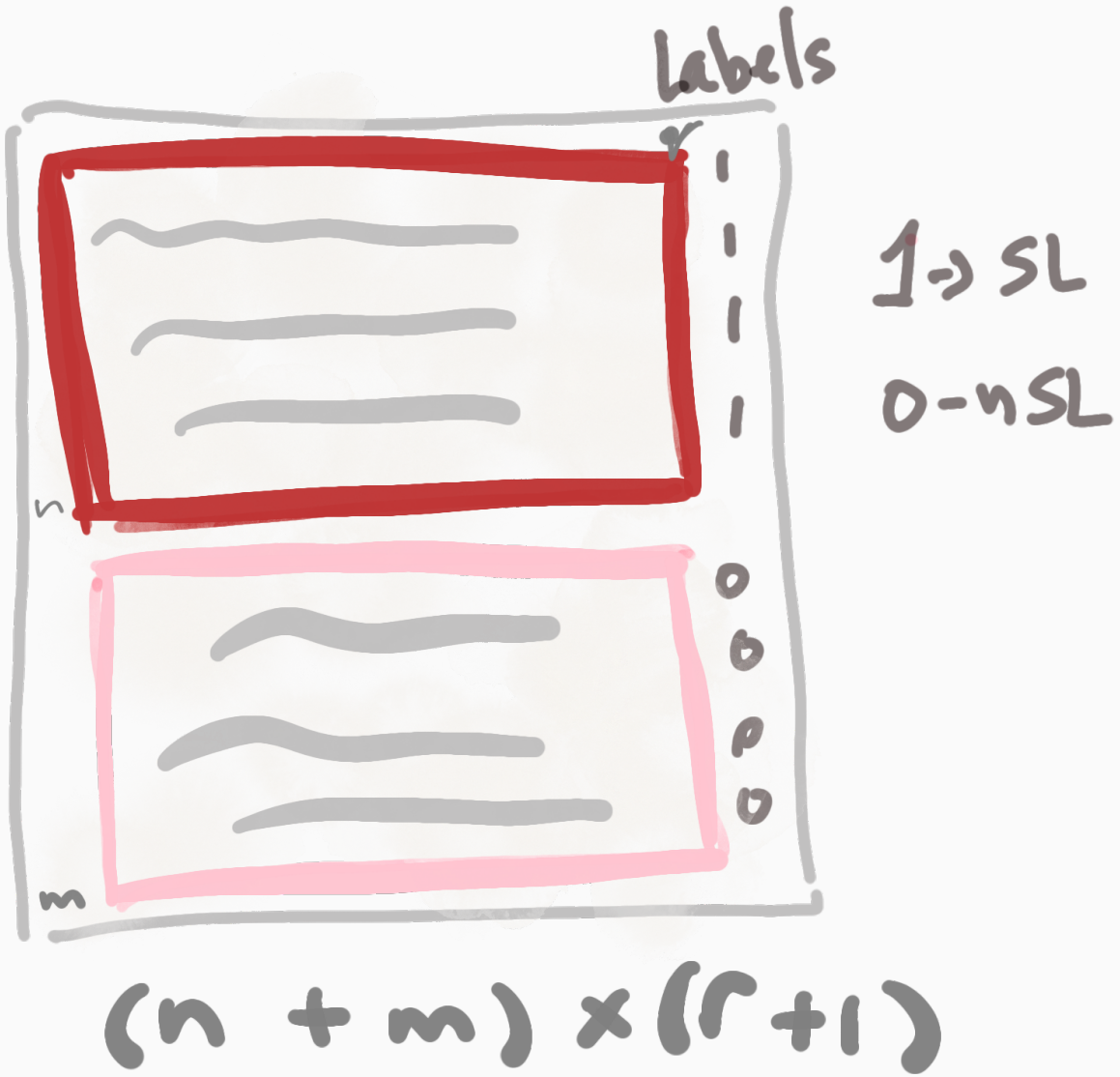
Features of the paper



Feature matrix




SL
U
nSL
U
Labels


labels
1
1
1
1
n
0
0
0
0
m
1 → SL
0 - n SL
(n + m) x (r + 1)

Everything starts! :)

- Splitting the data[1] for training and testing

[1] The data was a random sample of 10000 pairs taken from the given

Everything starts! :)

- Splitting the data[1] for training and testing
- Train a classfier with the training data

[1] The data was a random sample of 10000 pairs taken from the given

Everything starts! :)

- Splitting the data[1] for training and testing
- Train a classfier with the training data
- Make some predictions

[1] The data was a random sample of 10000 pairs taken from the given

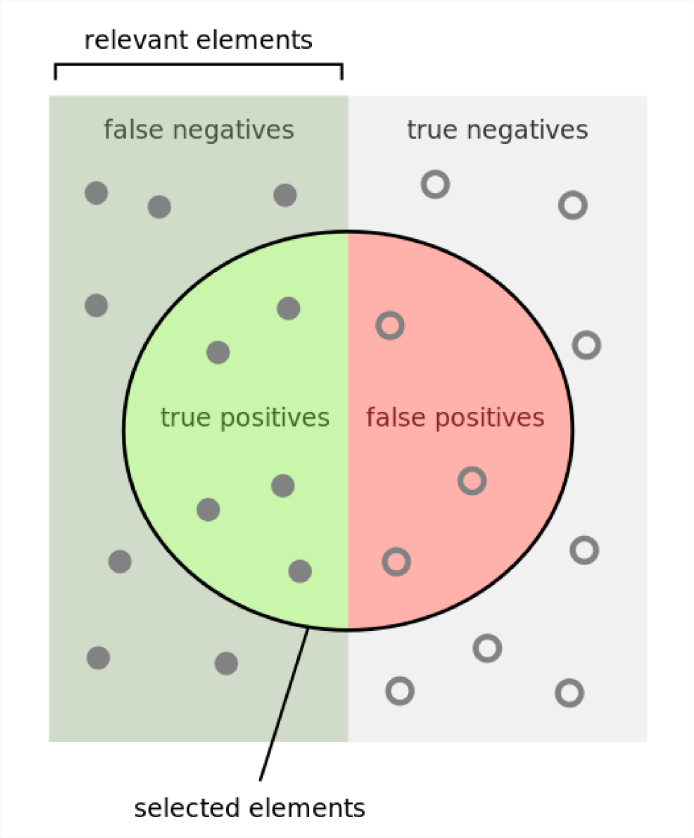
Everything starts! :)

- Splitting the data[1] for training and testing
- Train a classfier with the training data
- Make some predictions
- Evaluate the predictions based on the testing data.

[1] The data was a random sample of 10000 pairs taken from the given

Evaluating the method based on its prediction

- Accuracy : The percentage of in how many cases the correct class was predicted.
- Sensitivity : Percentage of correctly predicted SL interactions over the total number of SL interactions in the test dataset.
- Specificity : Percentage of correctly identified negative data over the total number of negative data.
- Precision : Percentage of correctly predicted positive data over the total number of predicted positive data.



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

From Wikipedia

Results of the evaluation

	Paper	My replication
Accuracy	0.84	0.89
Specificity (Recall of the negative class)	0.83	0.89
Sensitivity (Recall of the positive class)	0.85	0.89
Precision	0.83	0.89
F-score	0.84	0.89
AUC	0.927	0.9

5- fold Cross validation study