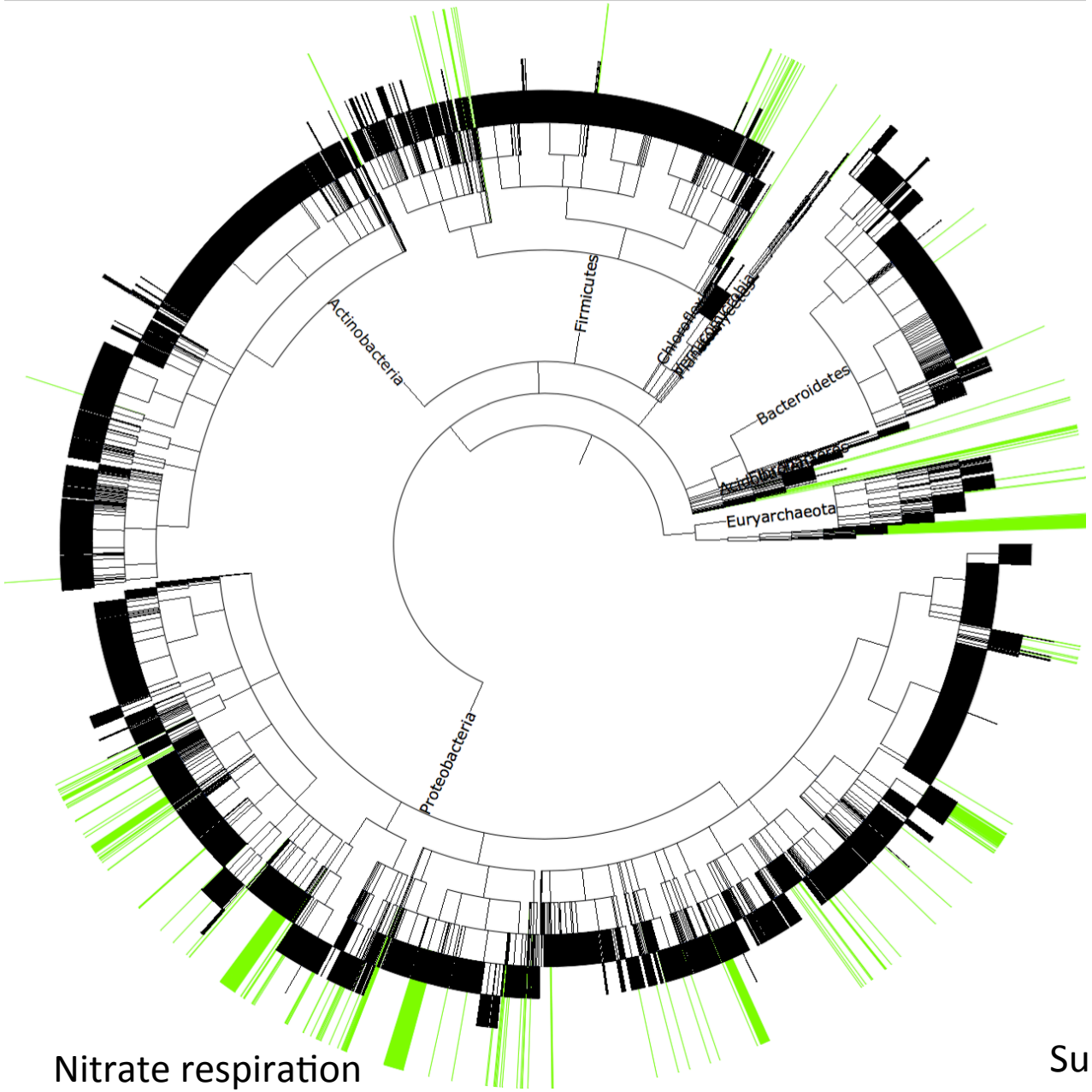
Using machine learning to infer phenotypes

FRED FARRELL

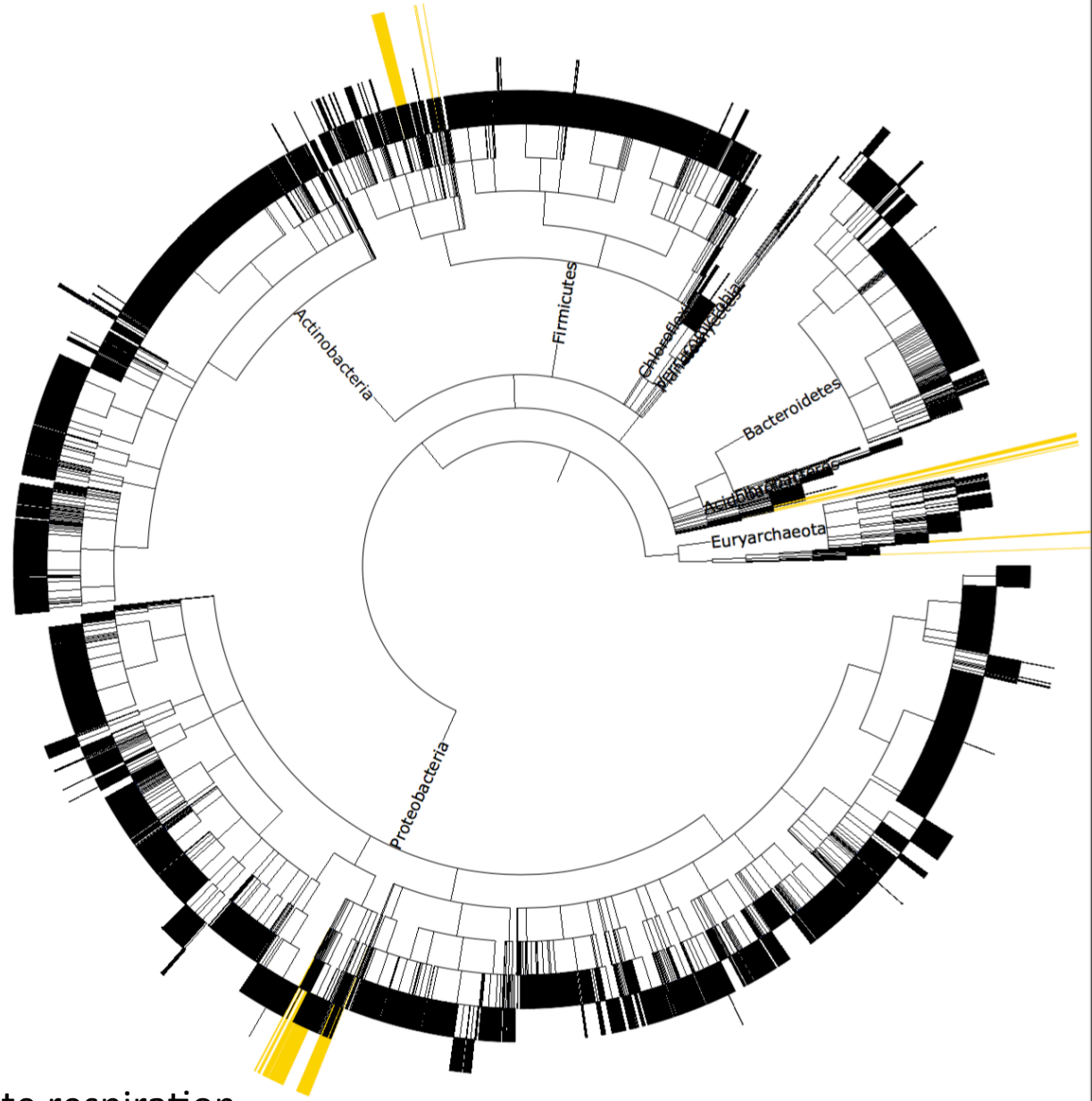
27/09/17

The problem

- We have MAGs and can assign genes to them
- Given a large collection of MAGs from different samples, difficult to characterize their ecological roles
- Can we infer phenotypes (e.g. metabolic capabilities) from genomes in an automated fashion?

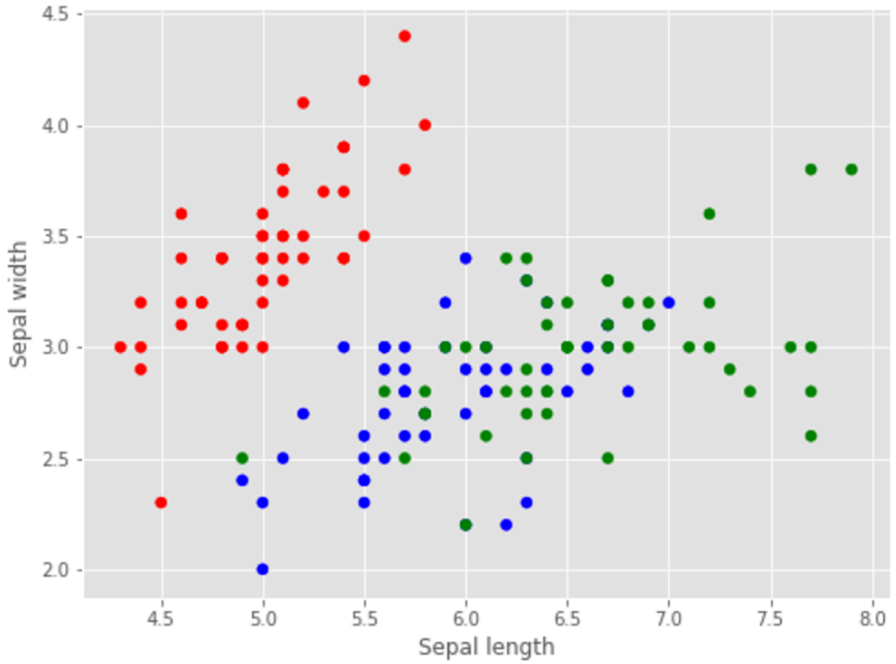


Nitrate respiration



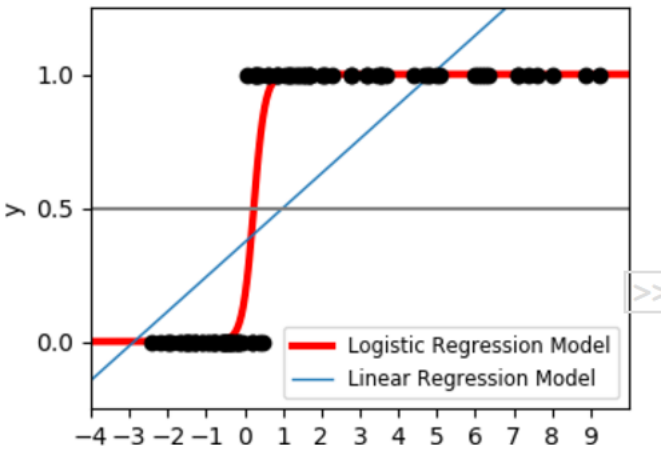
Sulfate respiration

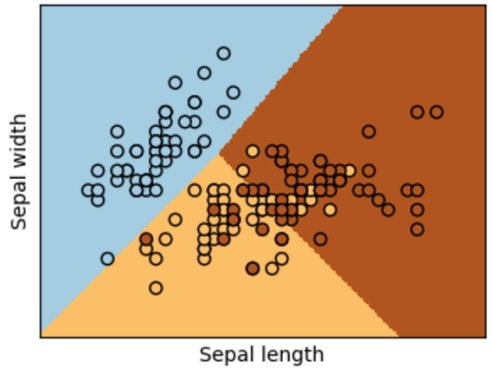
Classification problems



- Problem: given some features X, predict the class y of a data point, e.g. predict iris species from properties such as sepal length and width
- In reality, dimensionality of X will be much larger than 2 (in our case many thousands!)

Logistic regression





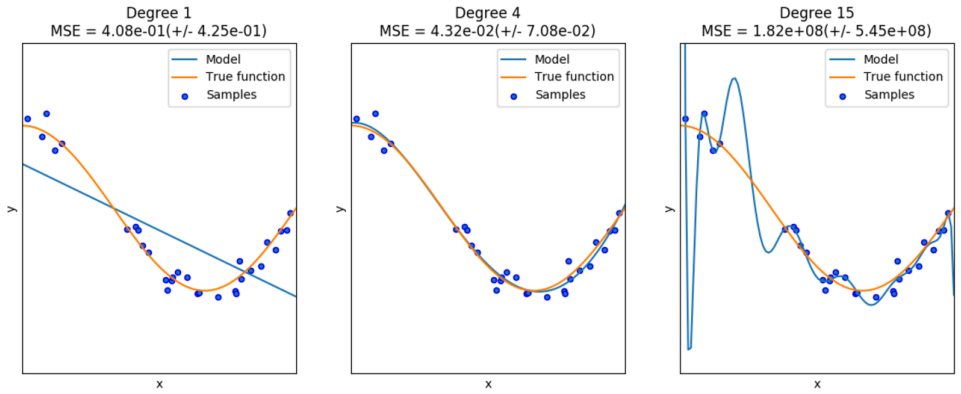
- Logistic regression: predict class based on a function

$$\mathrm{P}(\mathrm{y}=1)=f(\alpha\downarrow 1\ X\downarrow 1+\alpha\downarrow 2\ X\downarrow 2+\alpha\downarrow 3\ X\downarrow 3+\ldots)$$

where P is the (probabilistic) prediction and X_i are the features. We need to find $\alpha\downarrow i$.

- f is a function which saturates at 0 at 1 (red line on the left)
- For more than 2 classes, train an LR model for each class and predict the one with the highest probability

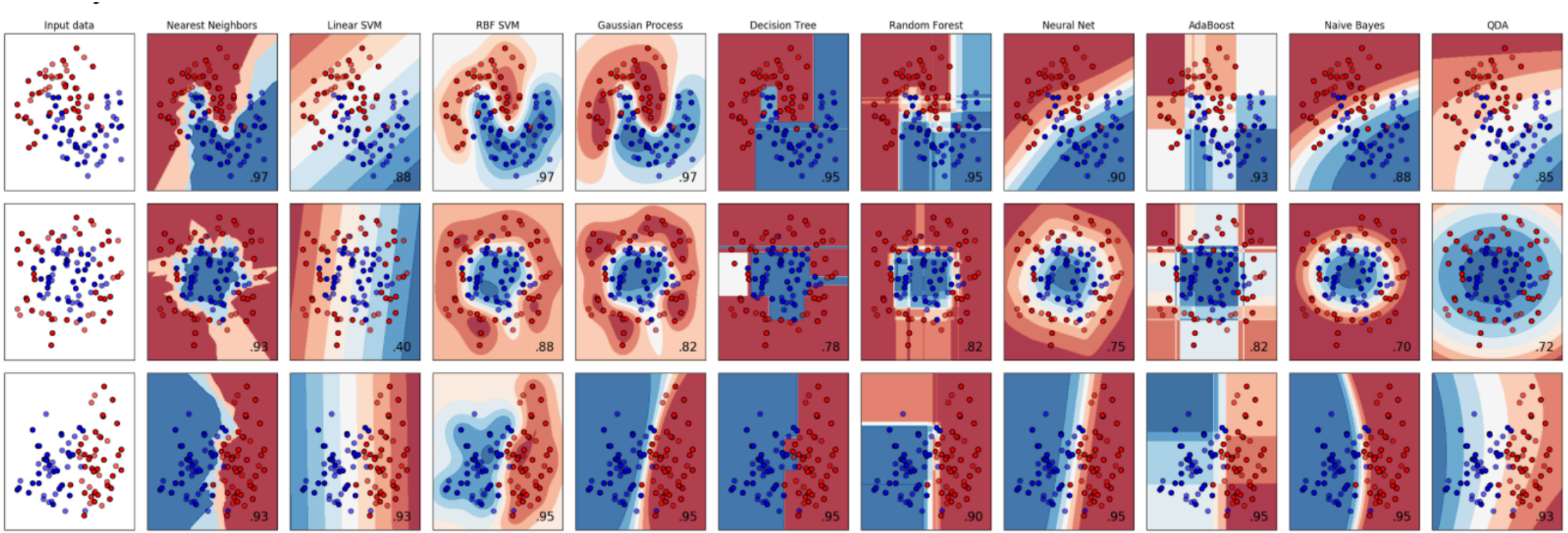
Training and testing a classifier



Examples of under- and overfitting

- To make sure the classifier works on new data, we hold back some data – a 'test set'
- We learn model parameters based only on the training set, and measure the model's performance on the test set
- Want to avoid 'overfitting', where the model learns very detailed features of the training data which don't generalize to new data

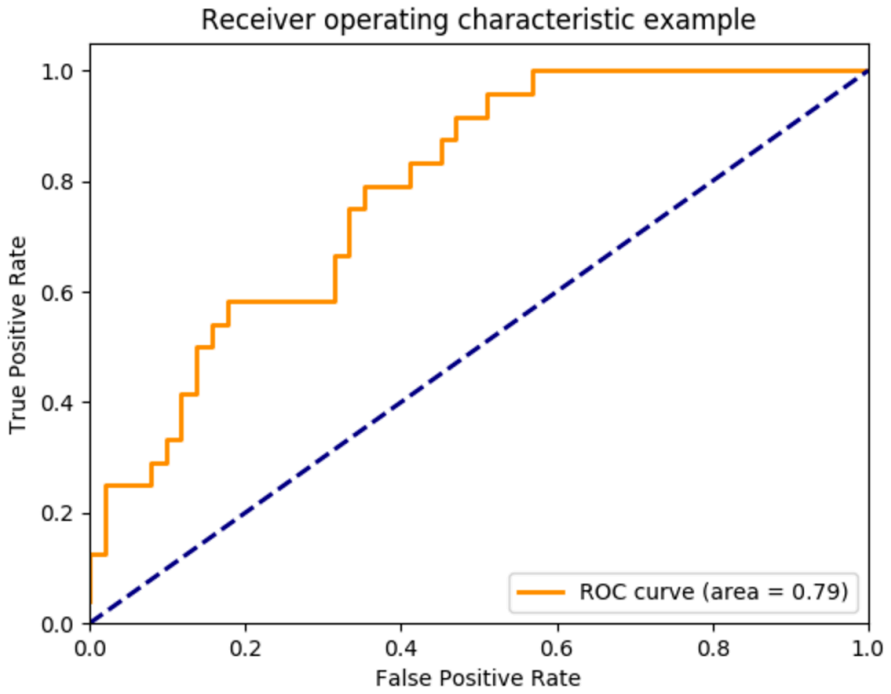
Other classification algorithms



Class imbalance and accuracy

- Often, one of the classes will have many more examples than another
- For example, if you're trying to find a rare disease based on a blood test the number of positive examples may be << 1 in 1000
- Simple accuracy (% correct) is then not a good measure of classifier performance, as you would do very well (>99.9% accuracy) by assigning everything to the negative class
- Can use other metrics – look at true positive, true negative, false positive and false negative rates. Sensitivity and specificity.

The Receiver Operating Characteristic (ROC)



- A good way to measure performance in highly unbalanced datasets is the so-called 'receiver operating characteristic'.
- LR outputs a probability. Can shift the 'cutoff' for making a positive prediction
- ROC is a plot of false positive rate against true positive rate as you do this
- Want the area under the curve as large as possible (perfect classifier=1).

Application to MAGs

- We'll now go through an example of applying logistic regression to predicting traits based on MAGS

Setting up Jupyter Notebook

- Log on to the virtual machine. Launch a Jupyter Notebook instance by typing:

```
jupyter notebook --no-browser --port=8889 --ip=127.0.0.1
```

- Then go to a local terminal and type:

```
ssh -N -f -L localhost:8887:localhost:8889 ubuntu@137.205.69.42
```

- This allows you to access the notebook server from your local machine
- Finally, open a web browser window and type ‘localhost:8887’ into the URL bar
- There, start a new notebook by going to New -> Python3