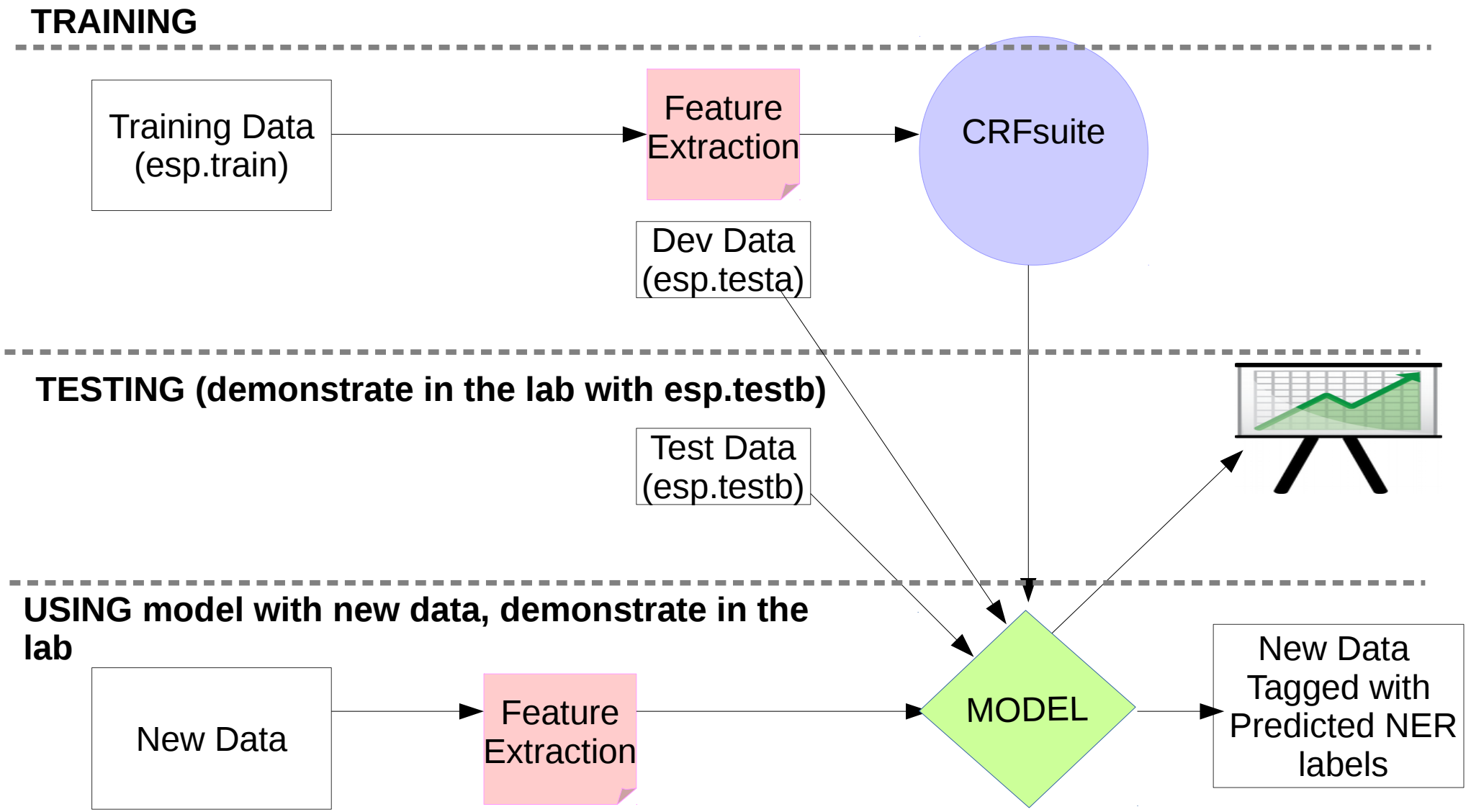
IE assignment

Build an Spanish Named Entity Recognition (NER) using the pycrfsuite

➔ CoNLL2002	NER data sets for training and testing from Conference on Natural Language Learning
➔ CRF	Python pycrfsuite
➔ NLP library	Python code + NLTK
➔ Evaluation tool	sklearn.metrics


TRAINING
Training Data
(esp.train)
Feature
Extraction
CRFsuite
Dev Data
(esp.testa)
TESTING (demonstrate in the lab with esp.testb)
Test Data
(esp.testb)
USING model with new data, demonstrate in the lab
New Data
Feature
Extraction
MODEL
New Data
Tagged with
Predicted NER
labels

CRF regularization parameter

Regularization is technique to **prevent over-fitting**,

- improve the generalizability of a learned model
- prevents the coefficients to fit so perfectly to over-fit
- typically chosen to impose a penalty on the complexity

L1 is the sum of the weights

L2 is the sum of the square of the weights

HINT Generally speaking, L2 performs slightly better than L1

http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/

What is a baseline?

- Information that is used as a starting point by which to compare other information
- Benchmark
- Something you want to beat

How a baseline looks like?

- Simple heuristics
- Simple Machine Learning techniques
- Simple feature sets
- The system I want to beat!

POS-Tagging baseline example

- **Simple heuristics**: all the words ending with “ing” in English are “gerunds”.
- **Simple Machine Learning techniques**: for example, Naive Bayes classifier.
- **Simple feature sets**: only use the words as features.
- **The system I want to beat!** The state of the art in POS-tagging have an accuracy of 97.24% (Toutanova et al. 2003), If you want to beat the best system, that system is your baseline.