Evaluation

**Evaluation**

* How do we know:
  + If our results are any good?
  + Which system is better?
  + If a change is for the better?
* Learning task T is any process by which a system improves performance **measured by P** from experience E.
* Example:
  + **T:** predict housing price
  + **E:** houses with known prices and independent variables
  + **P:** ?
* Measure the performance for a system:
  + Wrt to the task for which the system was built
  + Using a **TEST SET** for that task
  + Against a known **GROUND TRUTH**
  + Using **METRICS** that measures performance

**Test Set**

* Using a **TEST SET** for that task
  + Test cases:
    - Houses with known features
    - Images of cars
  + Against known **GROUND TRUTH** labels
    - House prices
    - Car brand/Driving direction/Right of way
* We can use a test set to evaluate how close to the truth we are
* **Never train on your test data!**
* Goal:
  + Learn a model for **unseen** (future) cases
  + Having learned to predict the test cases you can no longer use them to evaluate the performance on unseen cases
* **Split your data** in Training and Test set

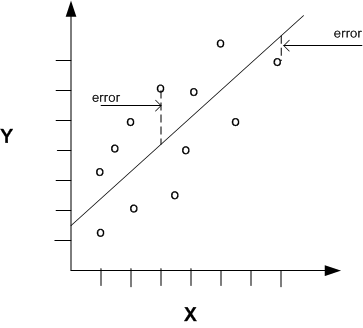
****

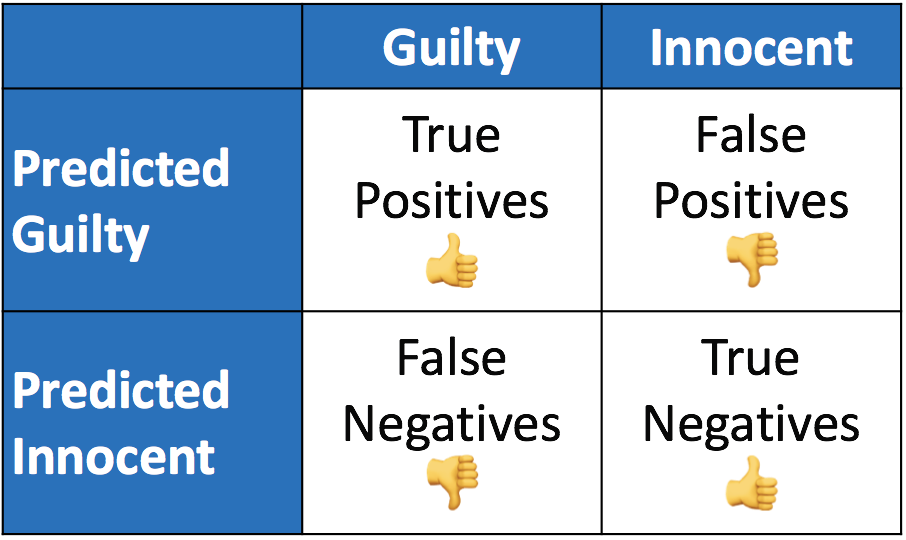
* Example:
  + 80%-20% split
  + The test set size corresponds to confidence in the evaluation
* **Cross validation:**
  + Commonly used
    - N-fold cross validation (small data sets)



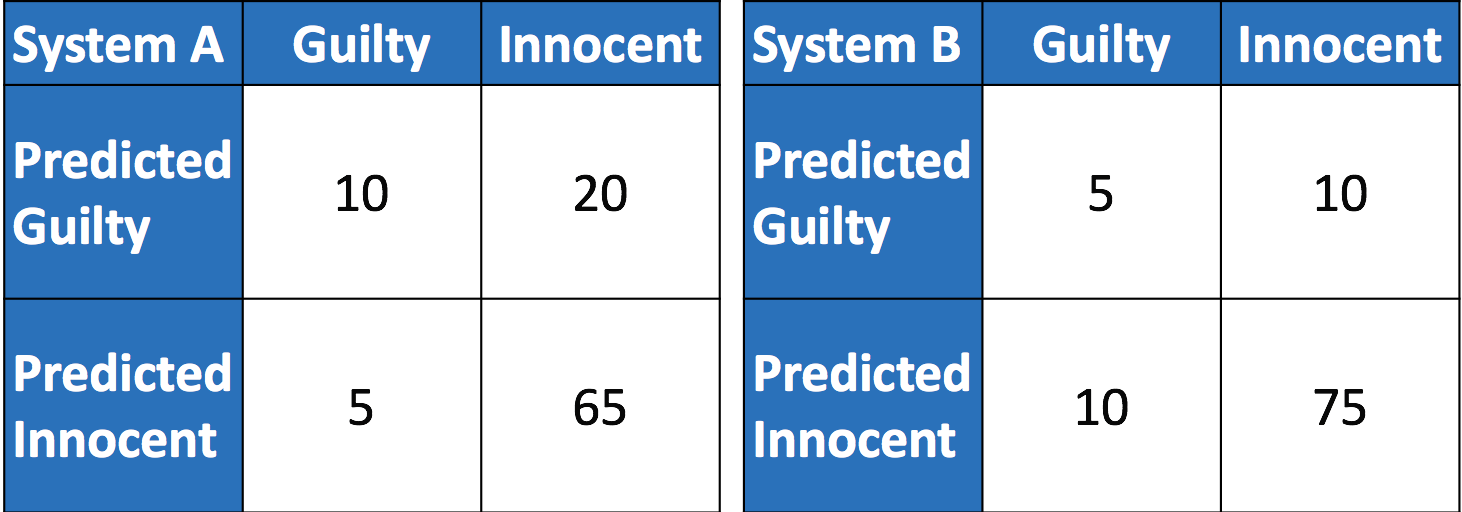
* + - Leave-one-out (for very small data sets)

**Metrics**

* Think of metrics as:
  + An (imperfect) indication of improvement
* Many different metrics
  + **Effectiveness:**
    - Accuracy
  + **Efficiency:**
    - Speed, memory usage
* ****Type of task:
  + Regression
* Metric: **RMSE**
  + Root Mean Squared Error
* Weakness:
  + Outliers!
* Type of task:
  + Classification
  + **Confusion matrix**



* **Accuracy:**
  + The fraction of correctly classified cases
* Accuracy = (TP + TN) / total
* System A = (10 + 65) / 100 = 0.75
* System B = (5 + 75) / 100 = 0.80

****

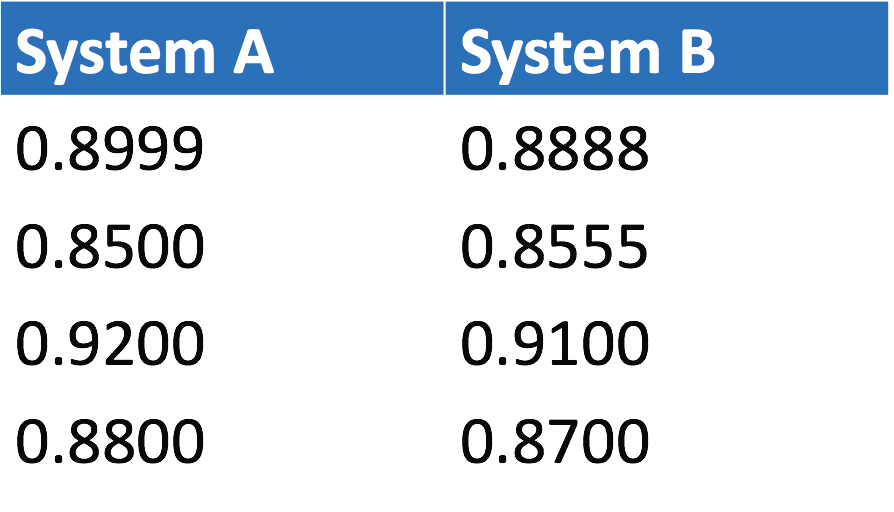
* **Recall:**
  + The fraction of positives that is identified
* Recall = TP / (TP + FN)
* System A = 10 / (10 + 5) = 0.67
* System B = 5 / (5 + 10) = 0.33
* **Precision:**
  + The fraction of the identified cases that is indeed positive
* Precision = TP / (TP + FP)
* System A = 10 / (10 + 20) = 0.33
* System B = 5 / (5 + 10) = 0.33
* Classic **tradeoff between metrics:**
  + The ‘same’ system can only improve one metric by sacrificing another

****

* There is a classic **tradeoff between metrics**
  + The ‘same’ system can improve on metric by sacrificing another
    - If you wish to convict more truly guilty persons you will also convict more innocent (higher recall 🡪 lower precision)
    - If you wish faster results you sacrifice accuracy
* True improvement is when you improve one metric without sacrificing another
* What metric to choose for our project?
  + Use what is commonly used in literature
  + The metric is beyond debate, reproducible, your results comparable to other work
* What if we think of a better metric?
  + That is one of the most common pitfalls.

**Confidence**

* How confident are we that B is better than A?
  + System A has RMSE = 0.8999
  + System B has RMSE = 0.8888
* Lower RMSE is better!

****

* We can statistically test whether B is **significantly** better than A:
  + Hypothesis 0: B is not better than A
  + Hypothesis 1: B is better than A
  + Based on the estimated probability of rejecting H0 when it is true.