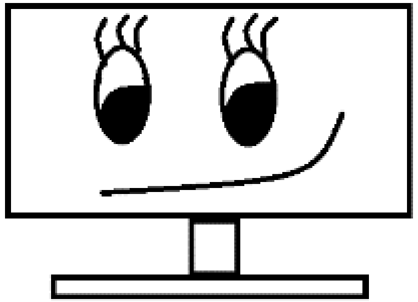


CVI620/ DPS920
Introduction to Computer Vision

Performance Evaluation

Seneca College

Vida Movahedi

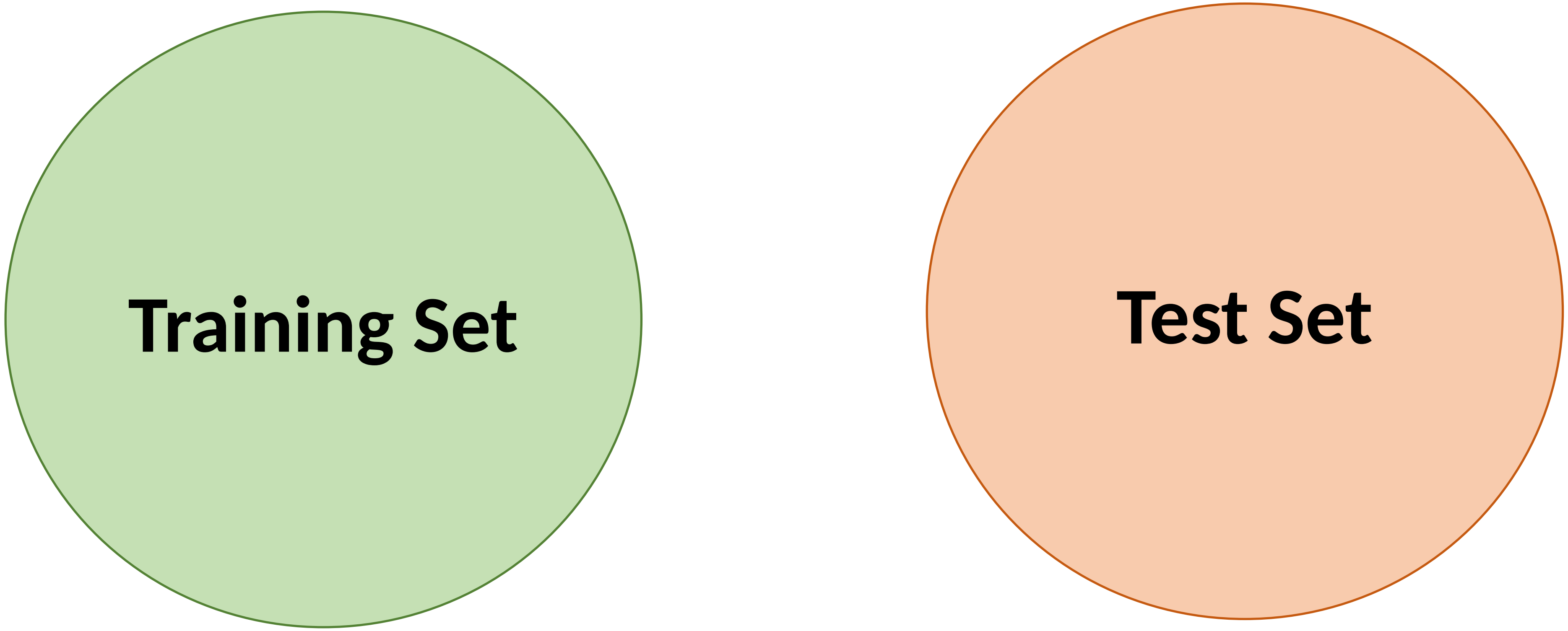
Overview

- Dataset Splits
- Overfitting vs Underfitting
- Performance Evaluation
    - Classification
    - Regression
    - ROI-based

Training / Test Sets

- Learning from training data
- Choose methods and models
- Set parameters

- Testing the trained method(s)
- Compare with competition
- Unknown until evaluation



Training Set

Test Set

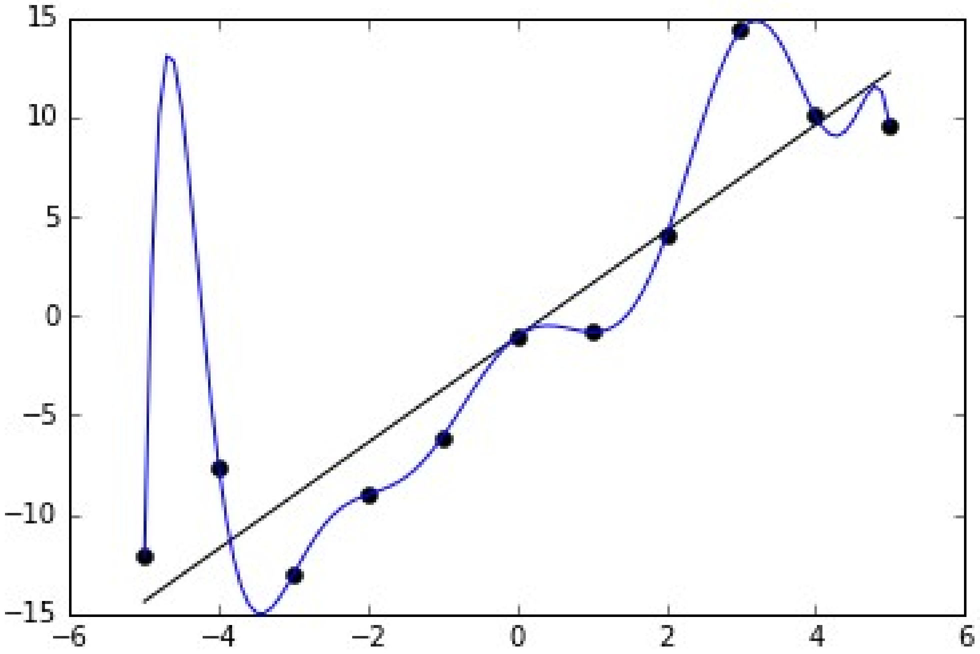
Low Performance

- Measuring Performance
- Error

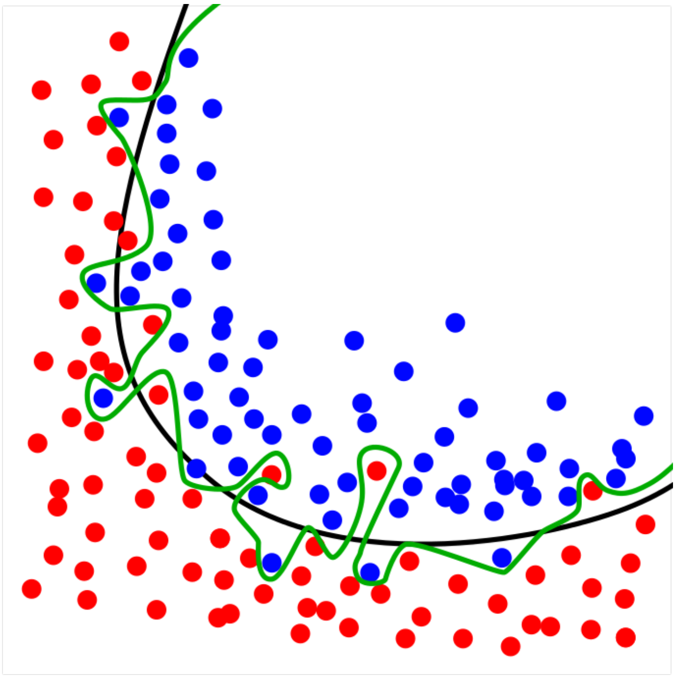
- Possible reasons for low performance?
  - Not enough data, or useful data (not separable when classifying?)
  - Not a suitable learning algorithm
  - Not enough computing to run the algorithm longer (need for more iterations)
  - **Overfitting / Underfitting**

Overfitting

- Corresponding too closely or exactly to the training data
- Not generalizing ☾ Poor performance when seeing new data / test data



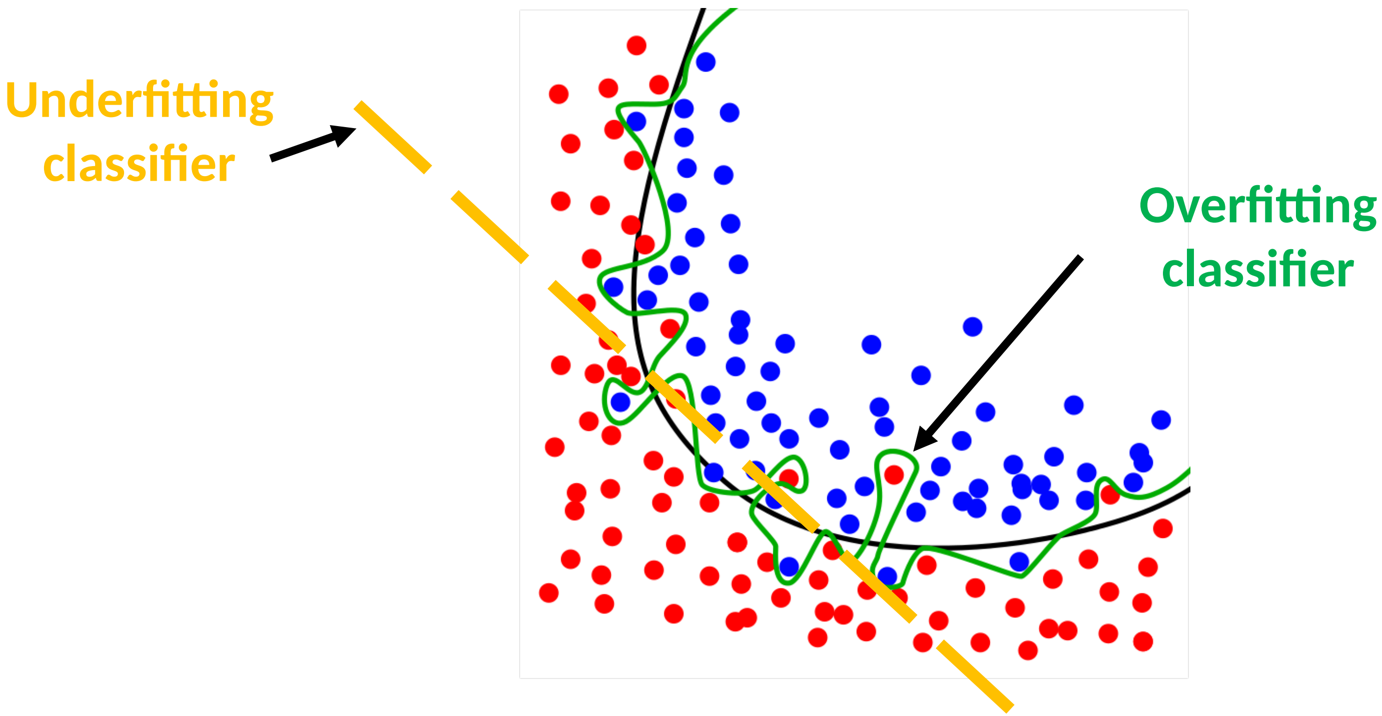
[Wikipedia] Overfitting in regression



[Wikipedia] Overfitting in classification

Underfitting

- The learned model, being too simple/ too general
- Not learning the details

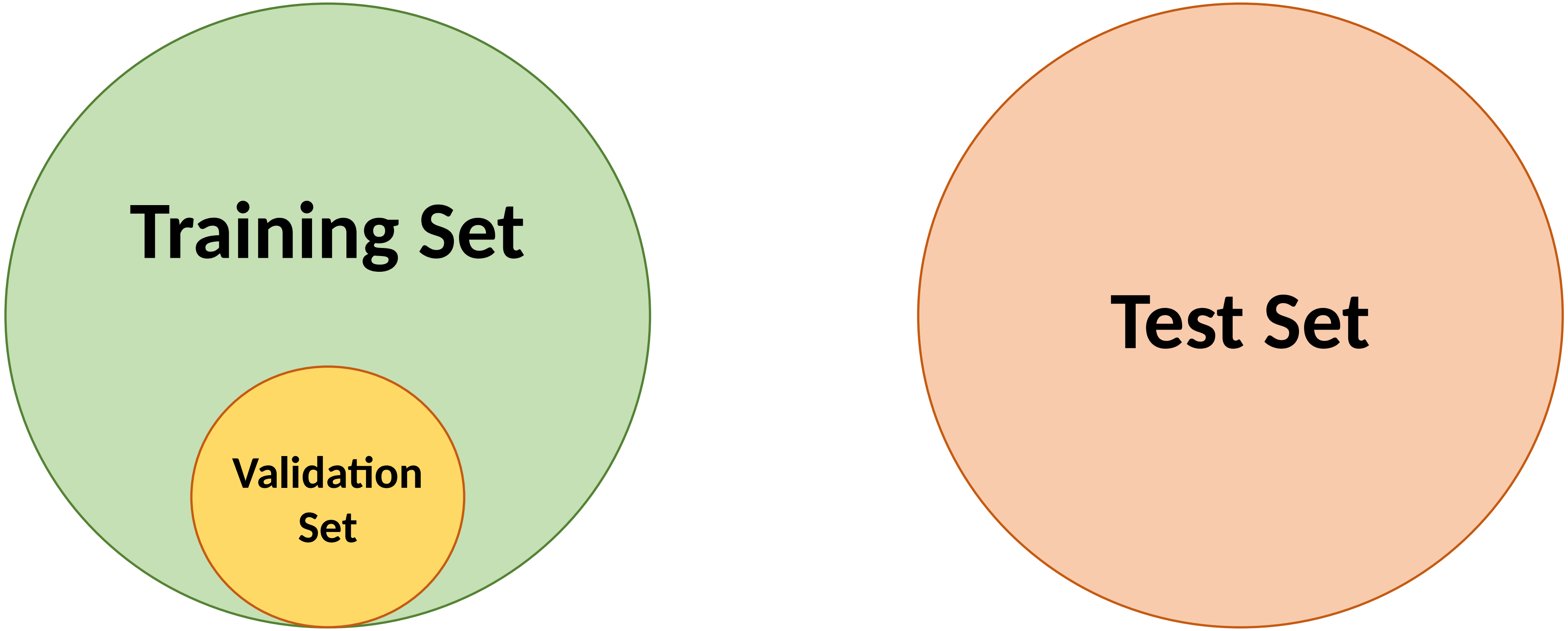


Underfitting classifier

Overfitting classifier

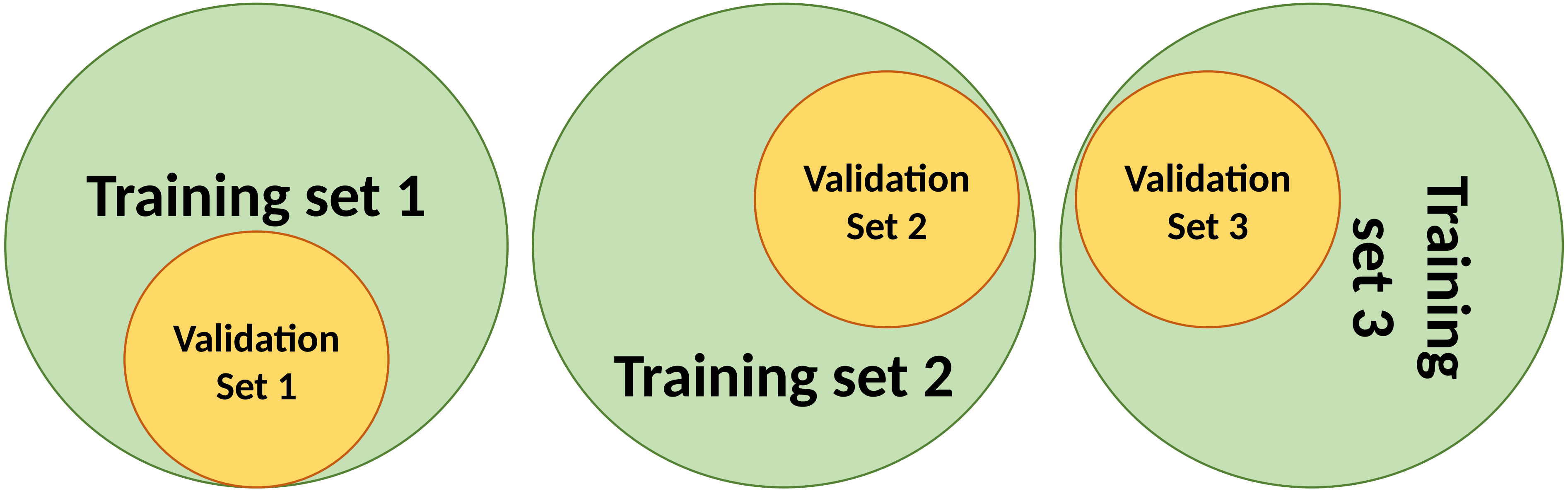
Validation Set

- A subset of training set
- Avoid overfitting by choosing methods / models / parameters that perform best on validation set



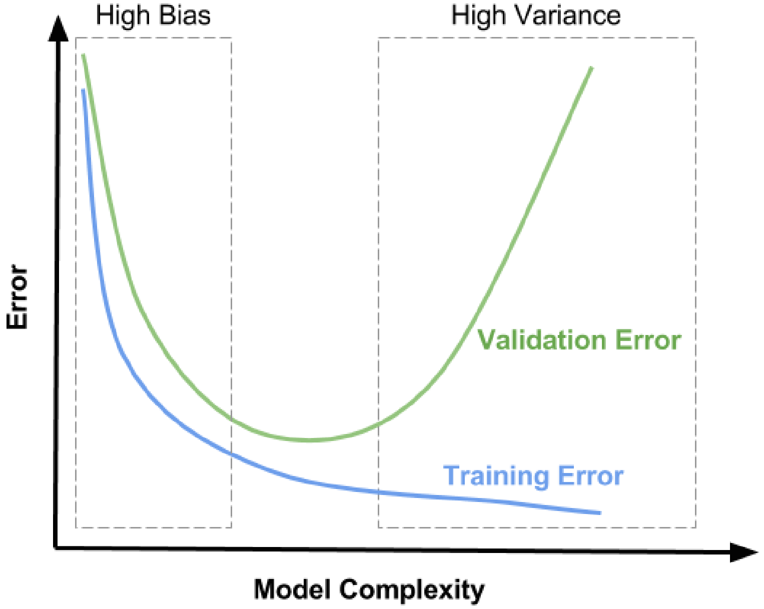
Cross Validation

- Choose different **subsets** of training set as validation sets, and train on remaining data
- Avoid overfitting by choosing methods / models / parameters that on average perform best on various validation sets



Validation Set

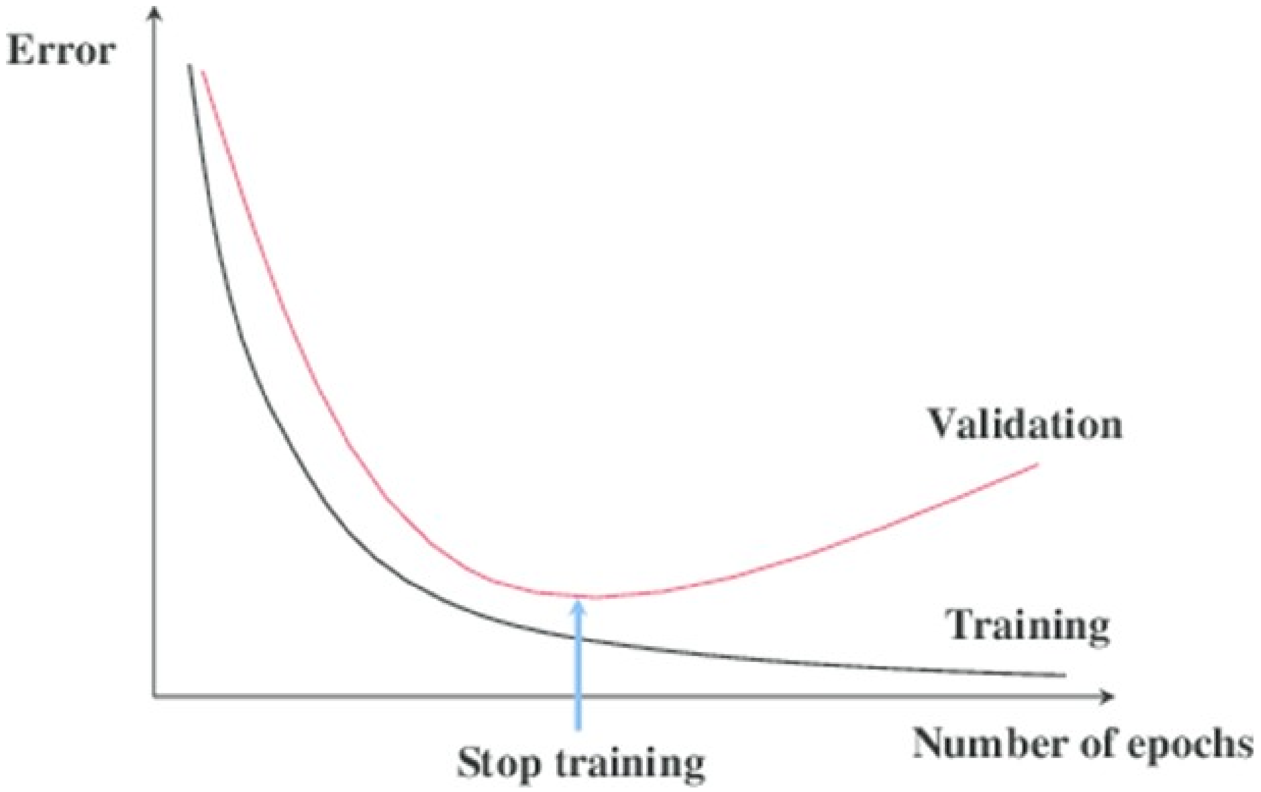
- Optimize parameters based on data not used in training
- Validation set
  - A randomly selected subset of the training set
  - To avoid under or overfitting when optimizing classifiers or regressors
- Cross validation
  - Leave-1-out
  - Leave-p-out
  - **Repeat and take average**



https://learnopencv.com/bias-variance-tradeoff-in-machine-learning/

Validation Sets for Neural Network Training

- Neural networks must be trained over many iterations (epochs). To ensure an optimum number of epochs, the performance error is measured on the training and validation sets.

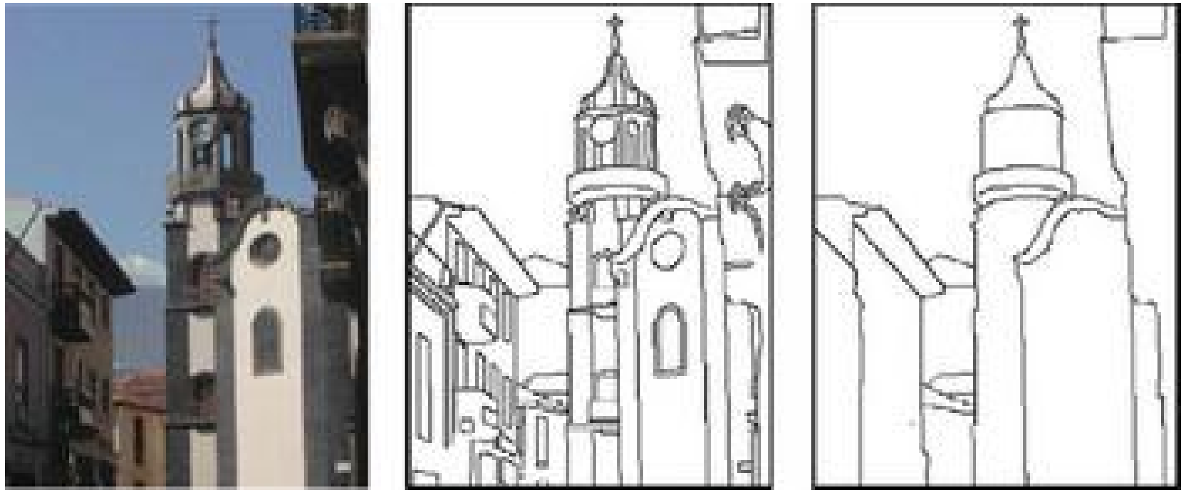


[Ref: http://fouryears.eu/tags/theory/]

Ground Truth [3]

- In simple words, ground truth is the perfect answer
- Sometimes not possible to calculate/ measure
- Often annotated/ provided / labelled by multiple human subjects (e.g. Amazon Mechanical Turk)
- The more human subjects, the less it will be considered as ‘subjective’

Example:
Berkeley segmentation and boundary detection dataset
30 human subjects
https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/



Performance Evaluation

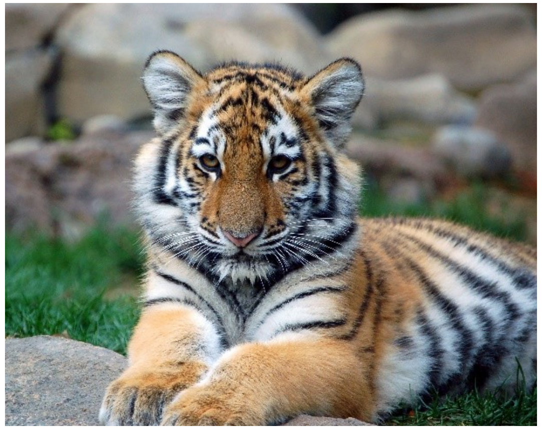
How well is the algorithm working?

- After implementing an algorithm, we want to know
  - How well it is working?
  - How does it compare with other algorithms
- **Qualitative** Performance Evaluation
  - Show the performance of the algorithm (by examples)
  - Show cases where it works well (or better than competition)
  - Show cases where it fails
- **Quantitative** Performance Evaluation
  - Not subjective, based on a test set and an evaluation measure
  - Fair to all methods being compared

Assessing Classification Models

Example: (House) Cat-Finder!

Is there a cat in the image?



Correct Answer (ground truth):

Yes Yes

No No

Algorithm’s output:

Yes No Yes No

TP FN FP TN

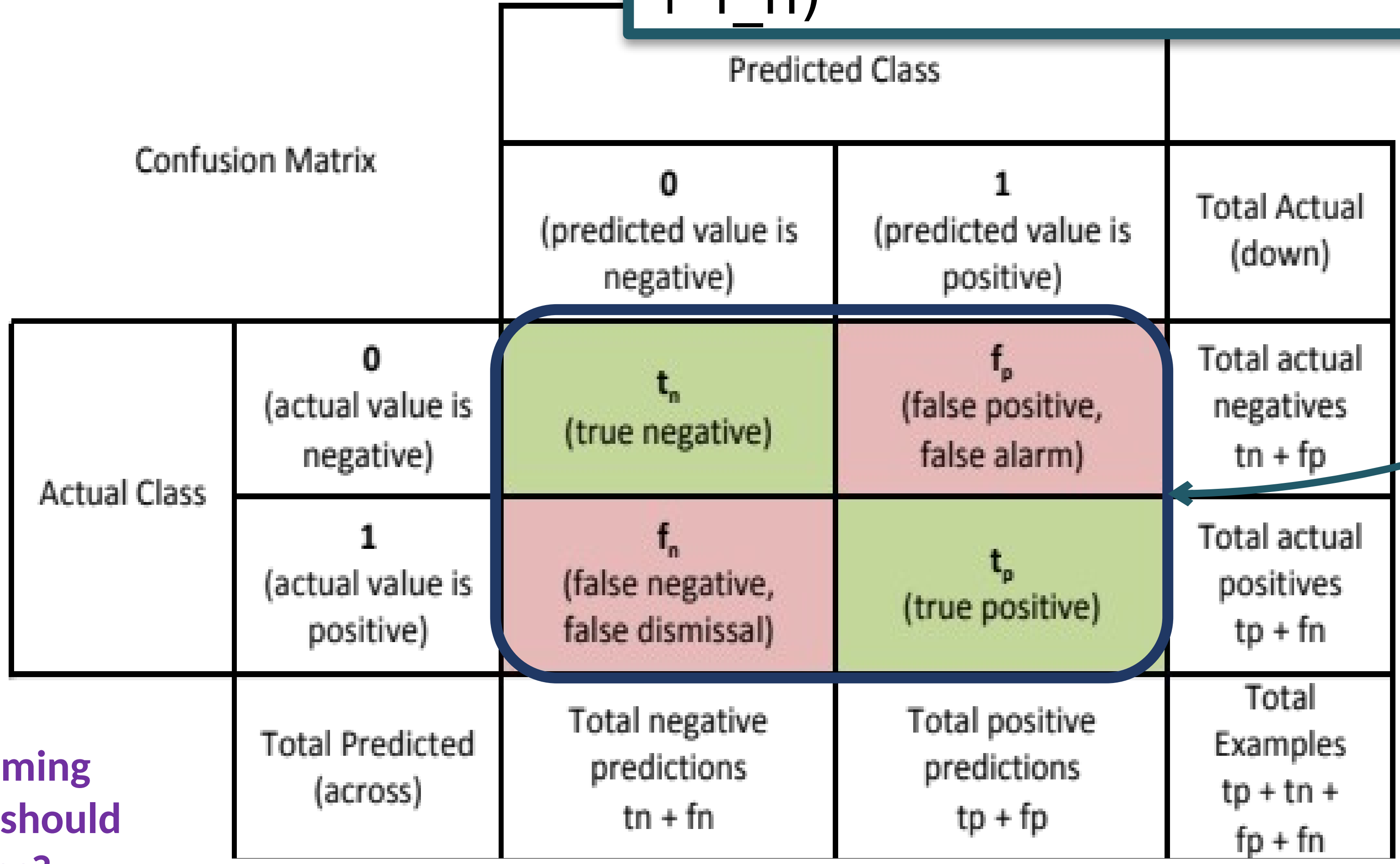
Metrics for Recognition/ Classification

- TP (True Positives):
  - Number of samples identified correctly as belonging to a class / category
- FP (False Positive): (False Alarms)
  - Number of samples identified incorrectly as belonging to a class / category
- TN (True Negative):
  - Number of samples identified correctly as NOT belonging to a class / category
- FN (False Negatives): (False Dismissal)
  - Number of samples identified incorrectly as NOT belonging to a class / category
- **Total number of samples = TP + FP + TN + FN**

Binary Classification Accuracy Metrics

Percent Correct Classification (PCC)

PCC = (t_n + t_p)/(t_p + t_n + f_p + f_n)



		Predicted Class		
Confusion Matrix		0 (predicted value is negative)	1 (predicted value is positive)	Total Actual (down)
Actual Class	0 (actual value is negative)	t_n (true negative)	f_p (false positive, false alarm)	Total actual negatives tn + fp
	1 (actual value is positive)	f_n (false negative, false dismissal)	t_p (true positive)	Total actual positives tp + fn
	Total Predicted (across)	Total negative predictions tn + fn	Total positive predictions tp + fp	Total Examples tp + tn + fp + fn

For a highly- performing model, which cells should contain larger values?

Precision / Recall

- Recall: Percentage of objects successfully identified

$$Recall = \frac{TP}{TP + FN}$$

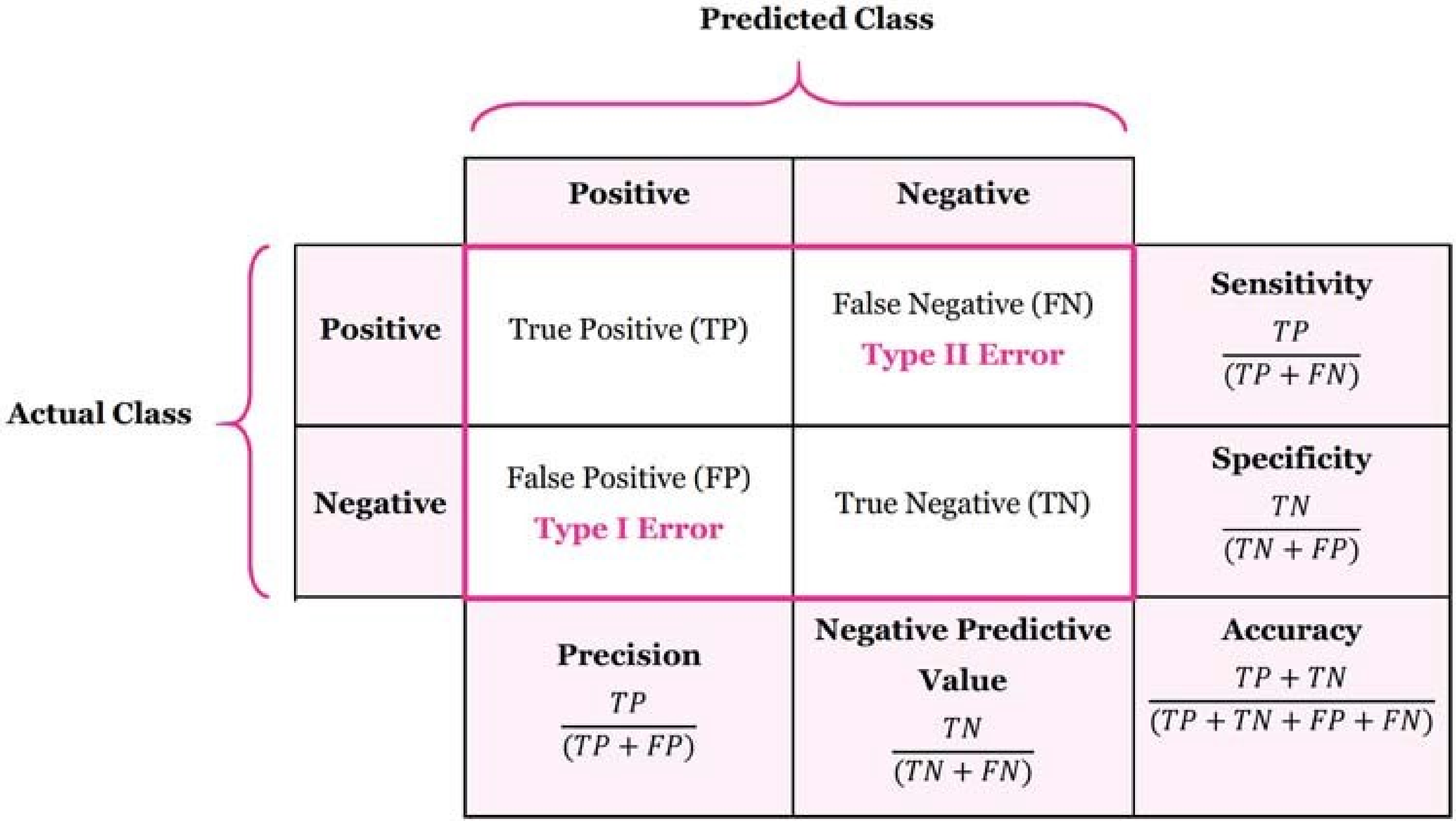
- Precision: Percentage of identified objects which are actually correct (not false alarms)

$$Precision = \frac{TP}{TP + FP}$$

- F- measure:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

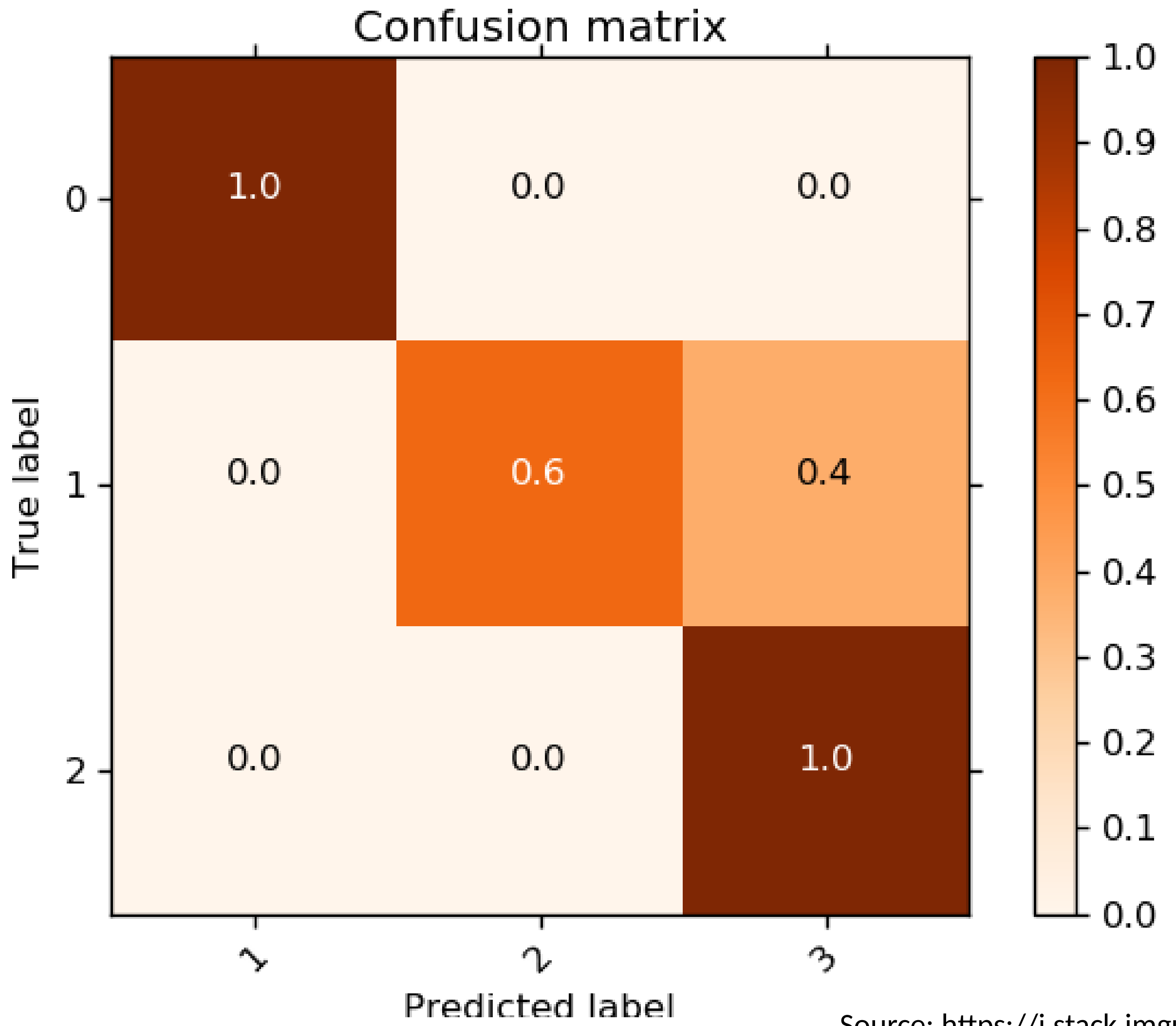
F_1 (β=1) is the most commonly used



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

https://stats.stackexchange.com/questions/122225/what-is-the-best-way-to-remember-the-difference-between-sensitivity-specificity

Confusion Matrix for Multi-Class Classification



Source: https://i.stack.imgur.com/9JbVM.png

Assessing Regression Models

Residuals, Errors

Actual Target Value ()	Predicted Target Value ()	Residual or Error ()
5.2	5.8	-0.6
6.5	6.2	0.3
8.2	8	0.2
10.5	10.2	0.3
4.8	5.8	-1
2.7	8	-5.3
3.4	6.5	-3.1
5.8	6	-0.2
9.7	8.2	1.5
6.3	6.1	**0.2**

Mean Squared Error (MSE)

- Low MSE ☾ good prediction
- High MSE ☾ bad prediction

R-Squared Measure

- 1 minus the ratio of the **variance of the residuals** and the **variance of the target variable** (total variance explained by model / total variance)
- The proportion of the variance in the dependent variable that is predictable (explained) from the independent variable(s).

It varies from 0% to 100%

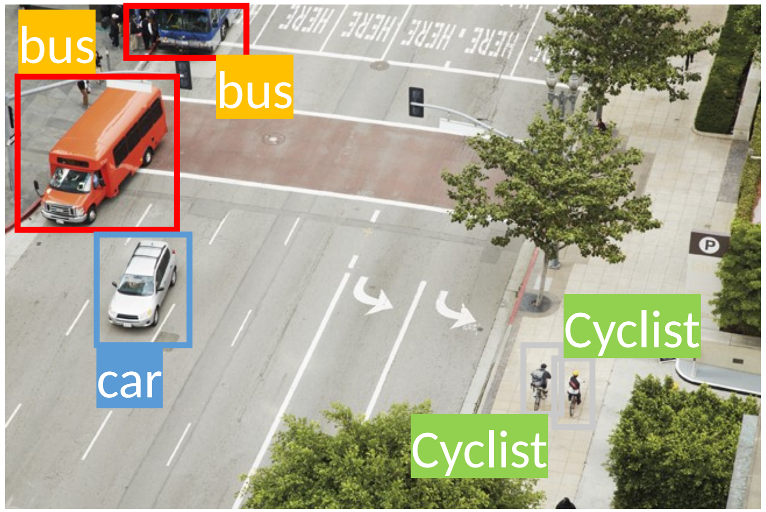
- High R^2 indicates a clear trend ☾ good prediction
- Low R^2 (close to zero) indicates randomness ☾ bad prediction

Assessing ROI-Based Methods

Region of Interest (ROI)

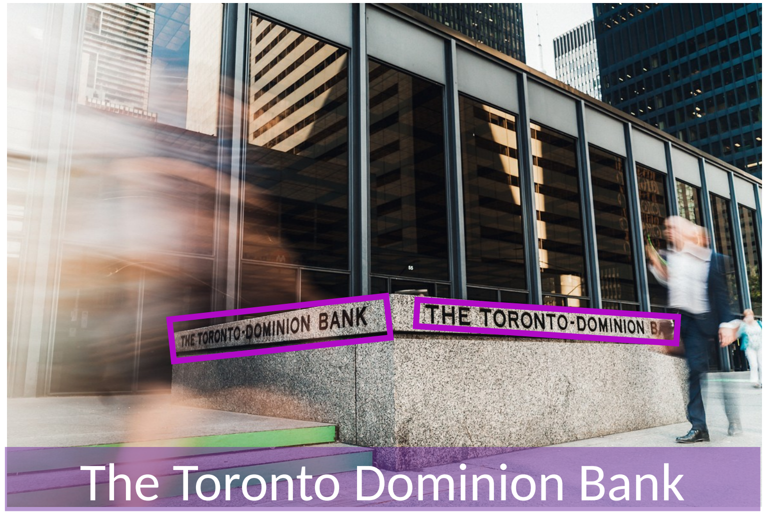
- A region in the image, often specified by a rectangle

Object Detection



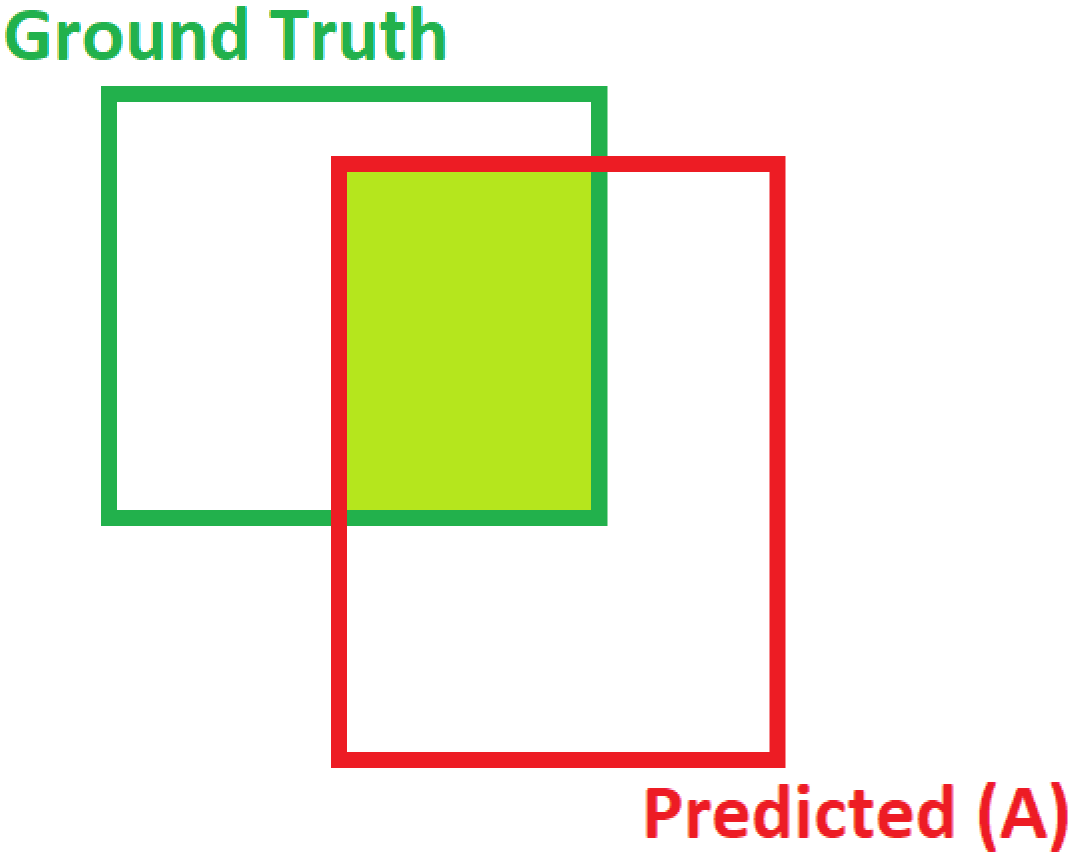
Face Detection & Recognition

Optical Character Recognition



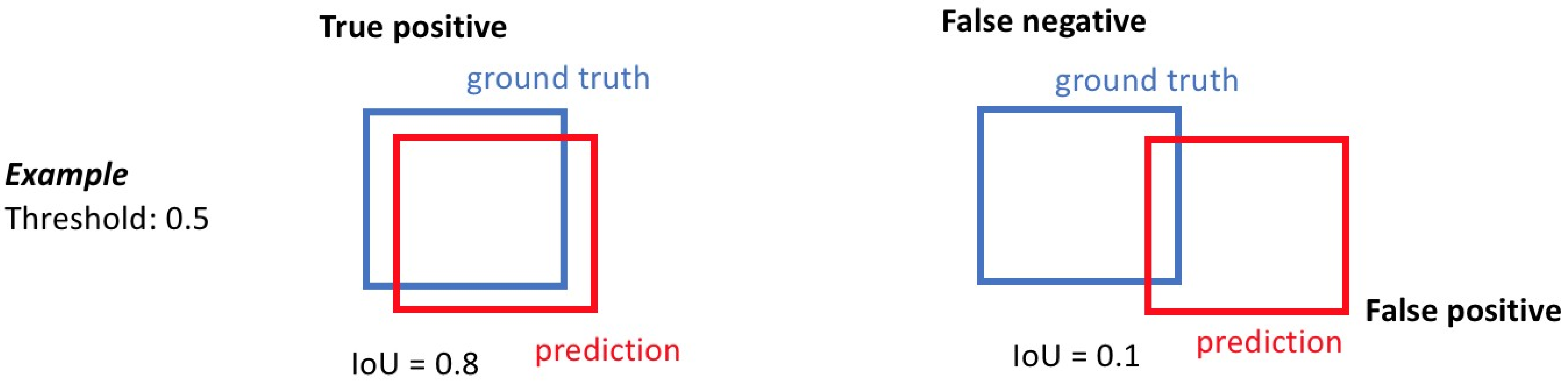
Intersection-over-Union (IoU)

- G: Ground Truth Region
- A: Predicted Region by Algorithm



From IoU to Precision & Recall

- By choosing a threshold, IoU can be translated to a hit (True Positive) or miss (False Positive)
- https://www.jeremyjordan.me/evaluating-image-segmentation-models/



Precision / Recall

- Recall: Percentage of objects successfully identified

$$Recall = \frac{TP}{TP + FN}$$

- Precision: Percentage of identified objects which are actually correct (not false alarms)

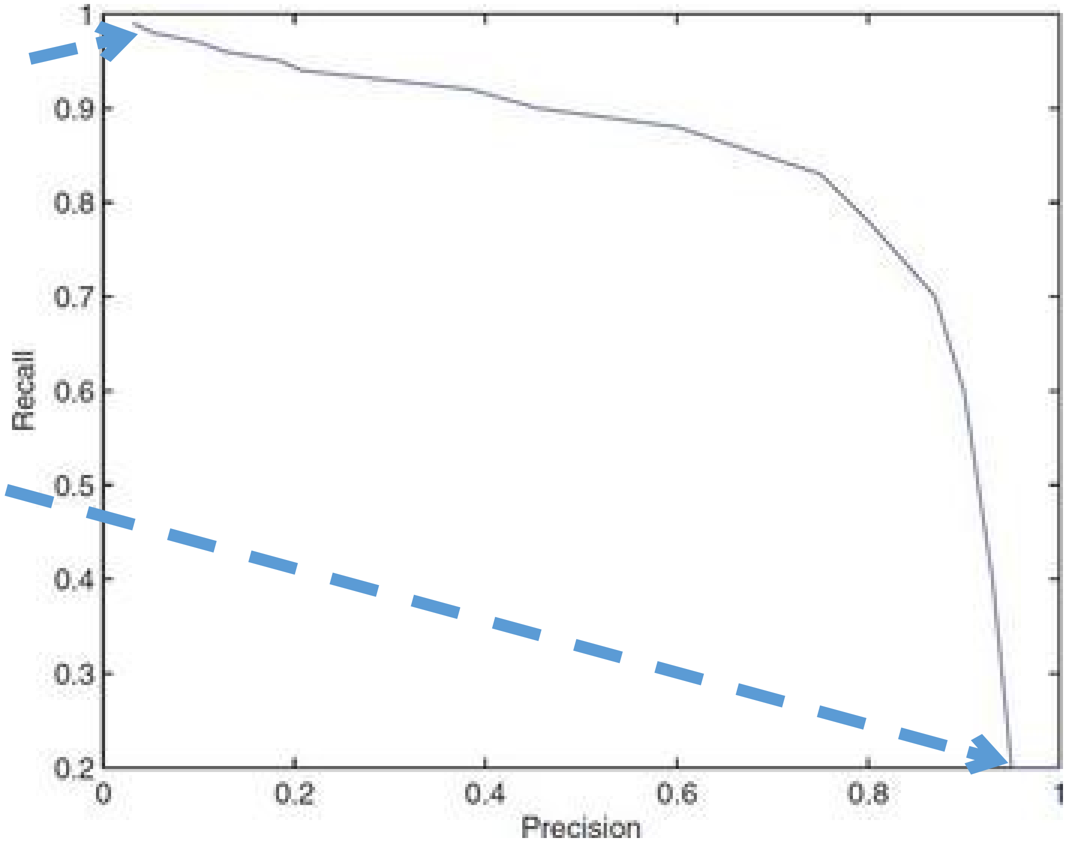
$$Precision = \frac{TP}{TP + FP}$$

Precision-Recall (PR) Curves

- Low threshold (close to 0) – high recall, low precision

- High threshold (close to 1) – low recall, high precision

- A common performance measure: Area under the PR curve (or ROC curve)



Summary

- In order to obtain models that perform well on future samples, datasets are often split into training, test and validation sets. Only the training dataset is used for training models.
- In addition to evaluating the performance of an algorithm **qualitatively** and **subjectively** (by looking at sample output), it is better to use **quantitative** and **objective** performance measures.
- There are general performance measures for classification and regression. Vision-based measures (e.g. ROI-based IoU) are also used.

References

[1] Computer Vision: Algorithms and Applications, R. Szeliski (http://szeliski.org/Book)

[2] Learning OpenCV 3, A. Kaehler & G. Bradski

- Available online through Safari Books, Seneca libraries
- https://senecacollege-primo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=01SENC_ALMA51532449200003226&context=L&vid=01SENC&search_scope=default_scope&tab=default_tab&lang=en_US

[3] Practical introduction to Computer Vision with OpenCV, Kenneth Dawson-Howe

- Available through Seneca libraries
- https://senecacollege-primo.hosted.exlibrisgroup.com/primo-explore/fulldisplay?docid=01SENC_ALMA51428109500003226&context=L&vid=01SENC&search_scope=default_scope&tab=default_tab&lang=en_US

[4] Applied Predictive Analytics: Principles and Techniques for the Professional Data