Medical Images – Evaluation metrics

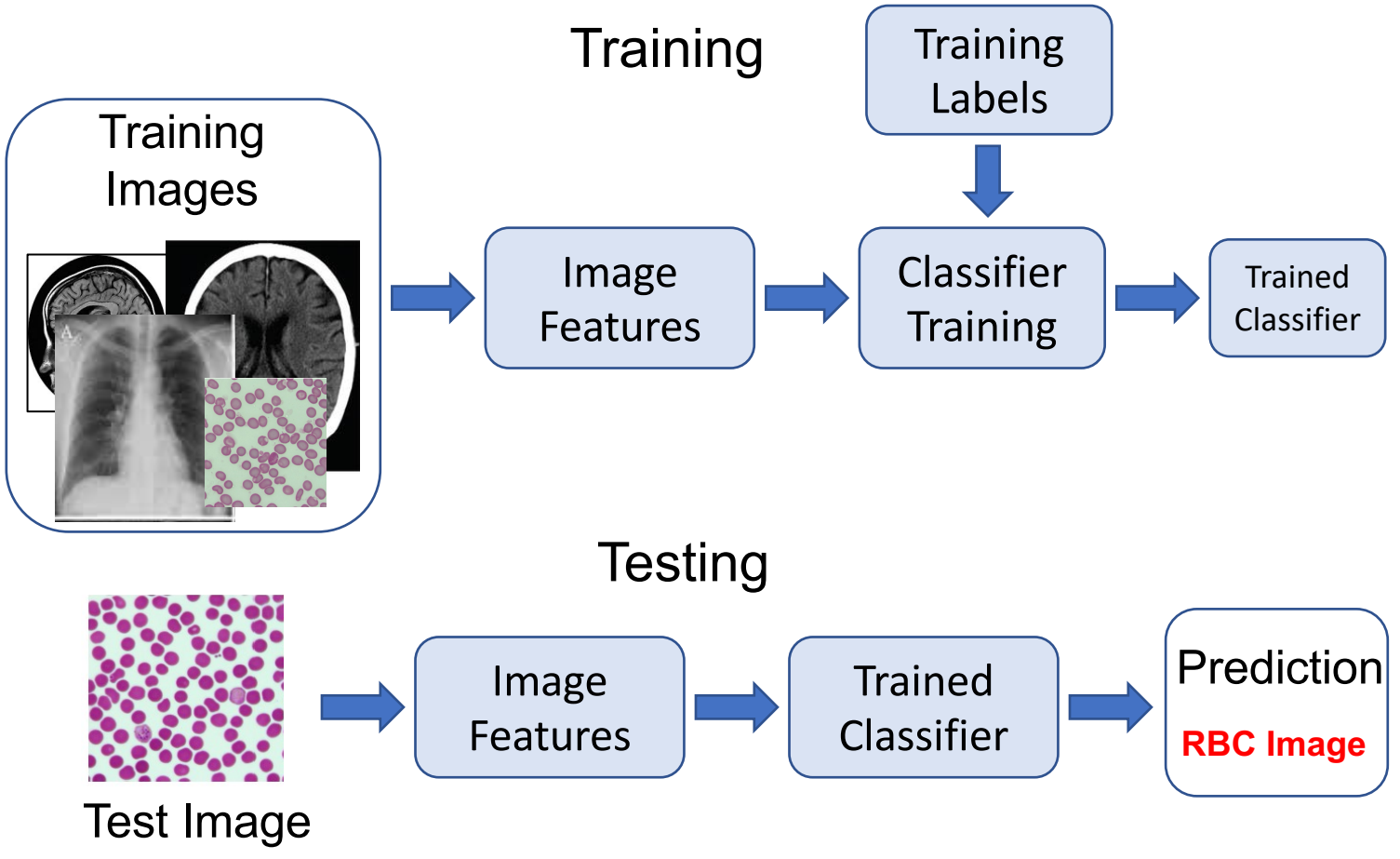
Lecture 25-Nov

Waqas Sultani
Information Technology University

Evaluations

- Image classification
- Object Detection (Classification + Localization)
- Image/Object Segmentation

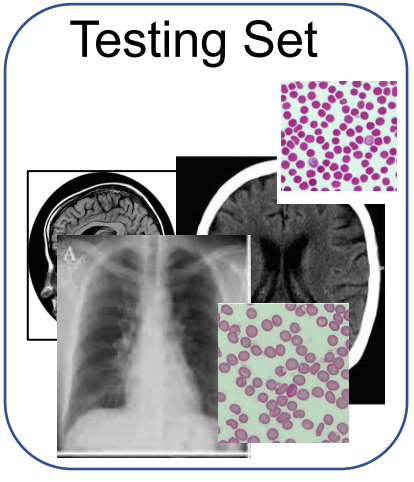
Image Categorization/Classification



Evaluations: Image Categorization/Classification

- Classifiers always return the probability (0-1)

- Multi-classifier
- Binary Classifier

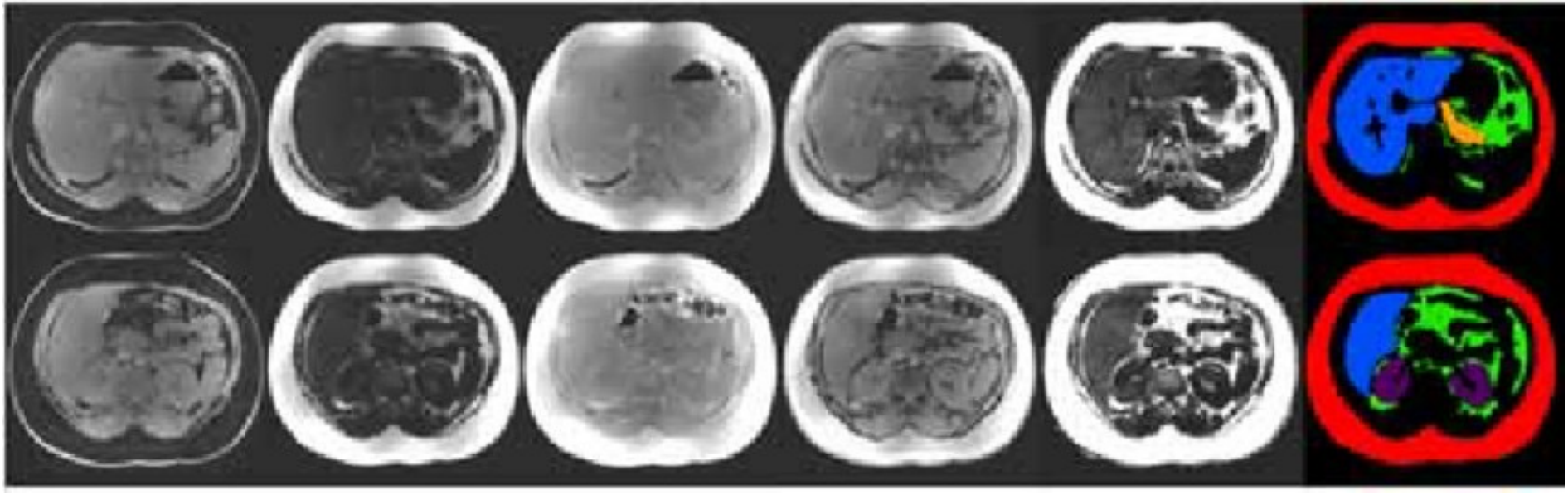


Accuracy means that how many data points are predicted correctly. It is one of the simplest form of evaluation metrics.

The accuracy score is= # of correct points / # total data points = (115) / 200 = 0.575.

Evaluations: Image Segmentation

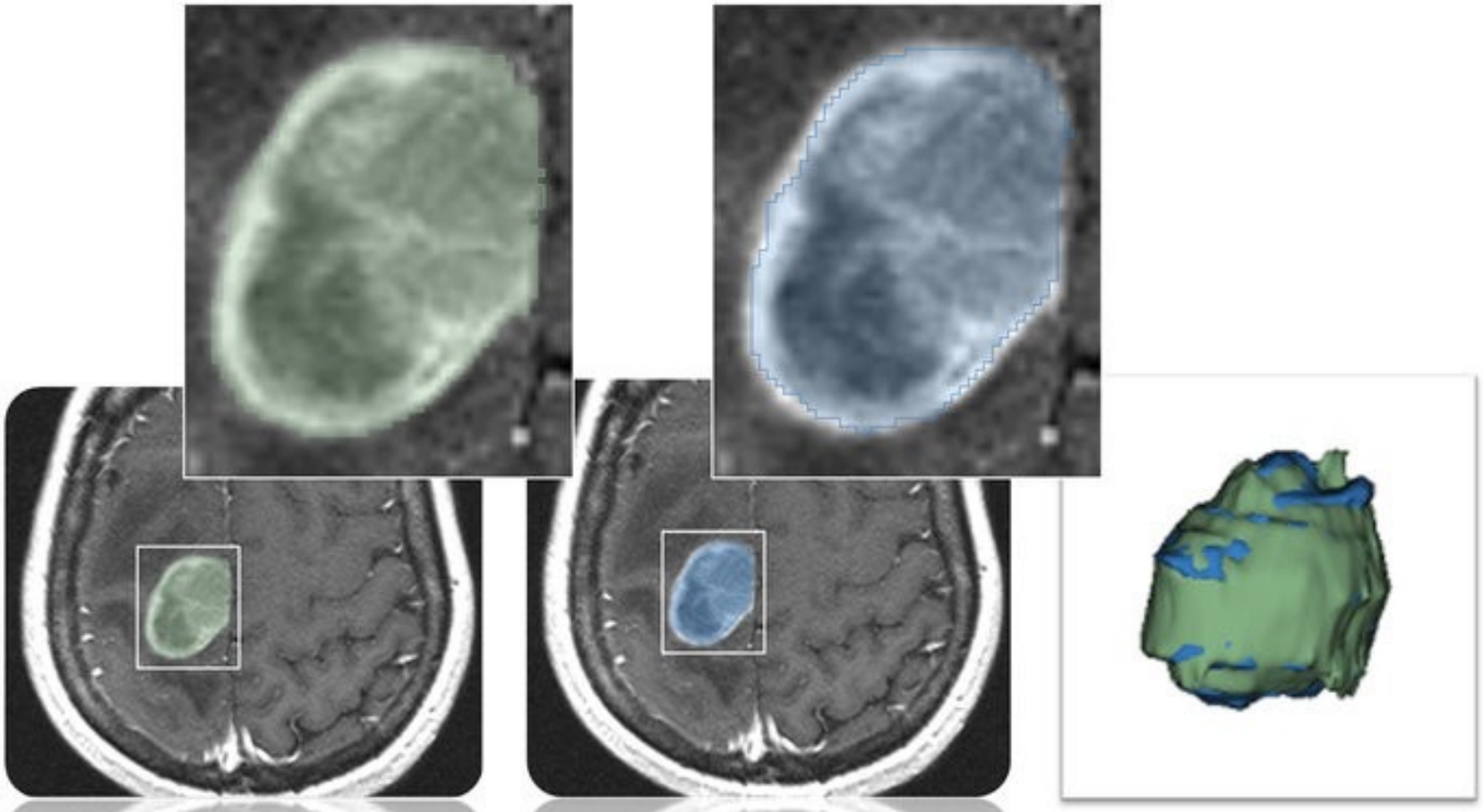
Visual Assessment



Manual image segmentation from the full spectrum of IDEAL MRI data to delineate red: SAT, green: VAT, blue: liver, yellow: pancreas, purple: kidneys. Left to right: water- only, fat-only, in-phase, out-of-phase, fat fraction, and segmented labels from ***SliceOmatic.***

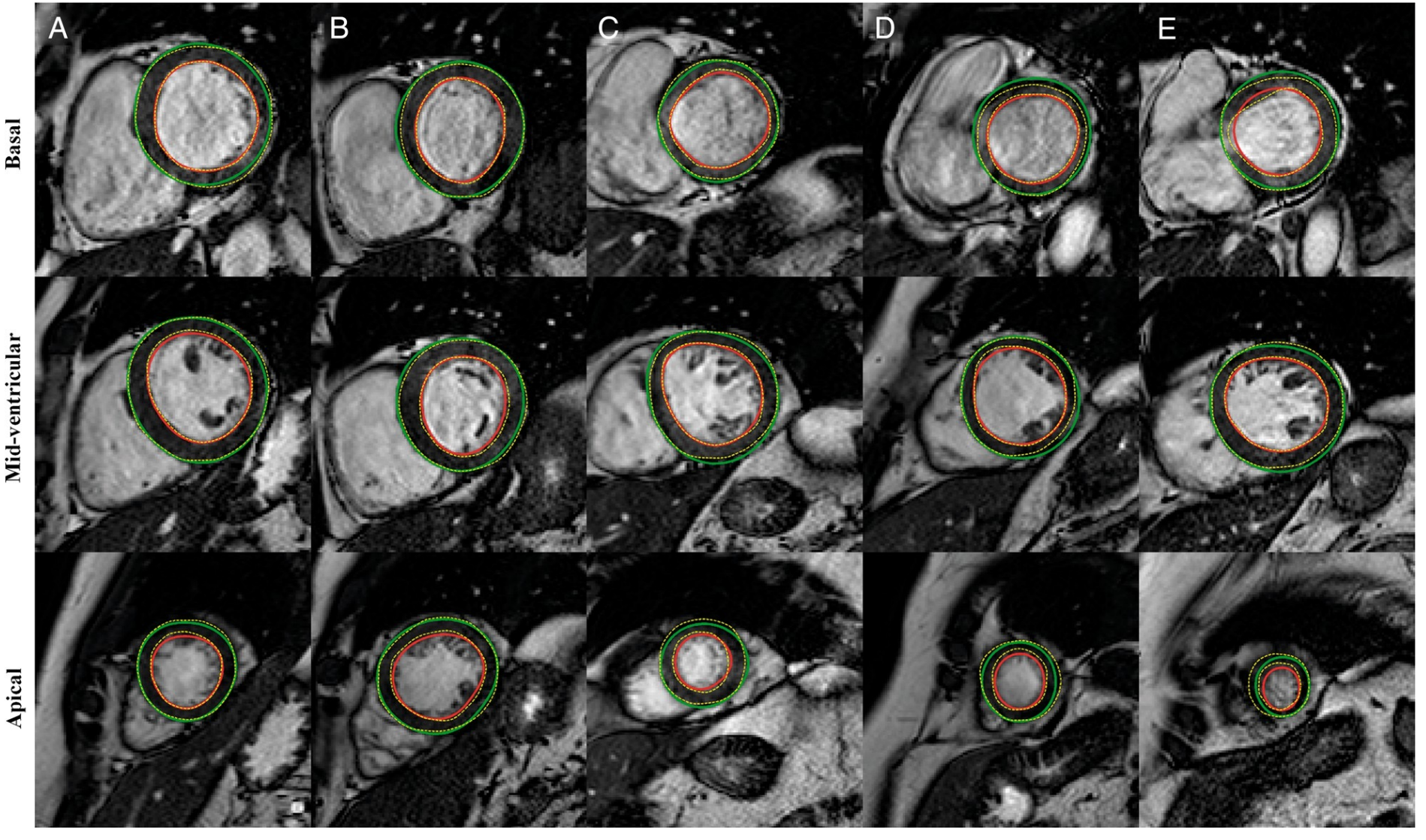
Reference: Assessment of Abdominal Adiposity and Organ Fat with Magnetic Resonance Imaging (chp11).

Inherent Uncertainty



Comparison of glioblastoma multiforme (GBM) segmentation results on an axial slice: semi-automatic segmentation under *Slicer* (green, left image) and pure manual segmentation (blue, middle image). *Egger et al., Nat Sci Rep., 2012.*

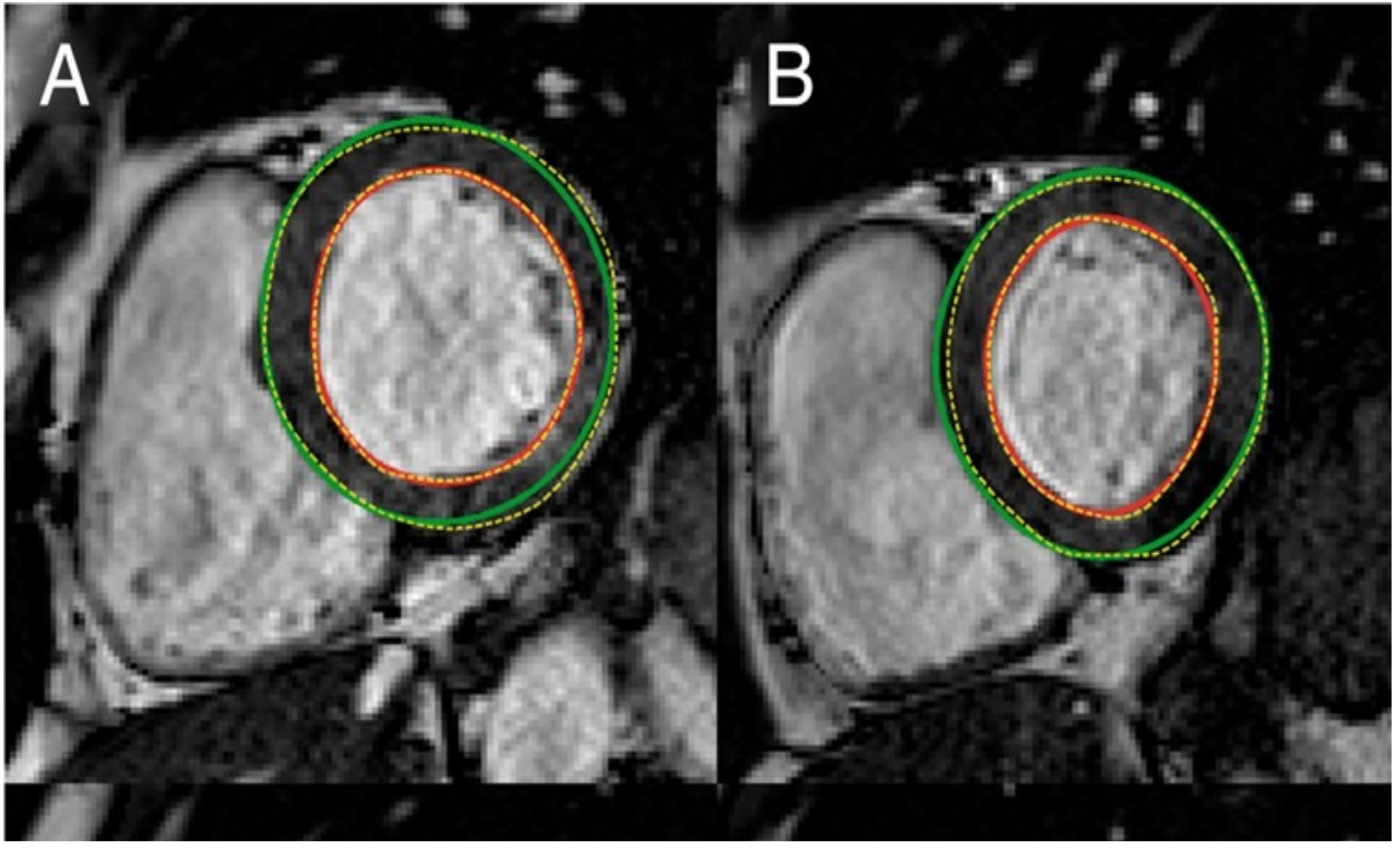
Inherent Uncertainty



red: endocardium; green: epicardium; yellow: ground truth

Queiros et al., European Heart Journal, 2016.

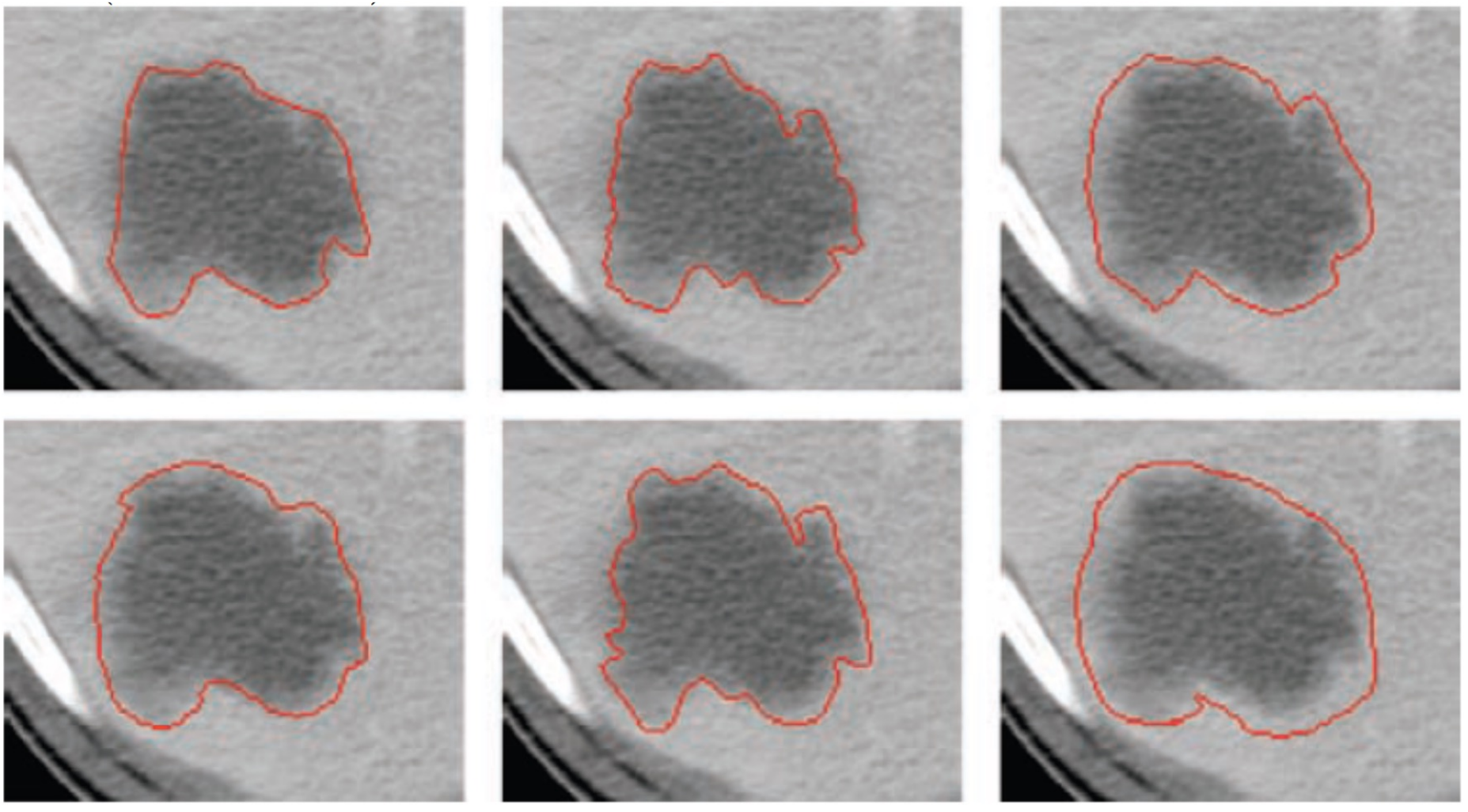
Inherent Uncertainty



red: endocardium; green: epicardium; yellow: ground truth

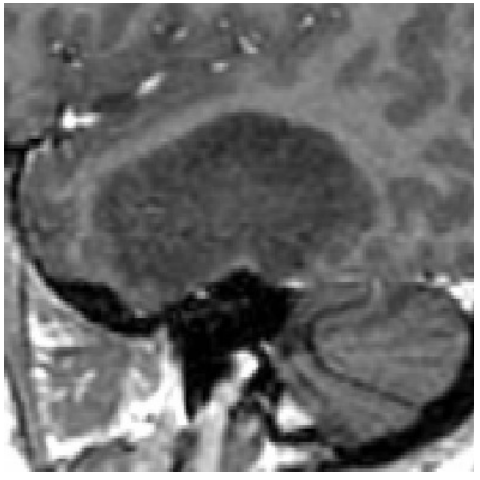
Queiros et al., European Heart Journal, 2016.

Observer Variability – Example: Liver lesion

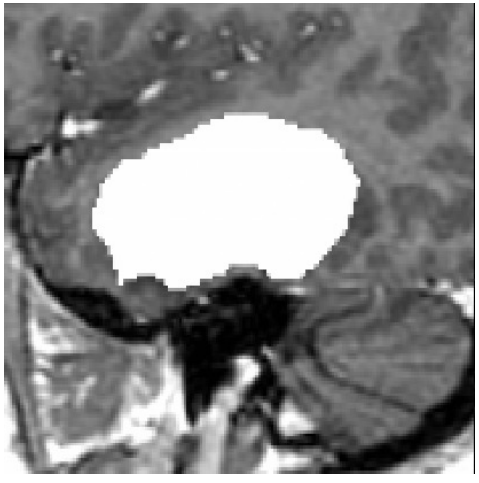


Existent Segmentation Data

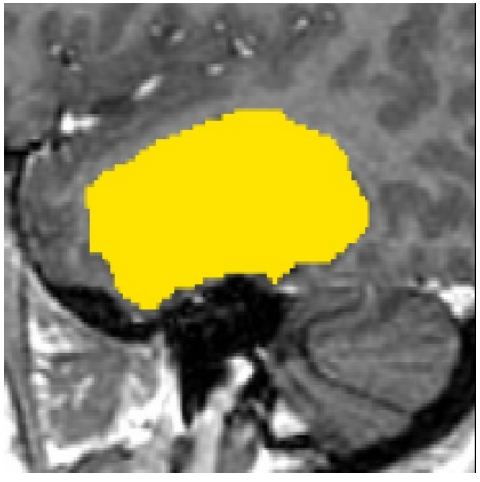
- Manual segmentation performed by 4 independent experts
- low grade glioma
- Low grade gliomas are brain tumours that come from two different types of brain cells known as astrocytes and oligodendrocytes. They are classified as a **grade 2 tumour** making them the slowest growing type of glioma in adults.



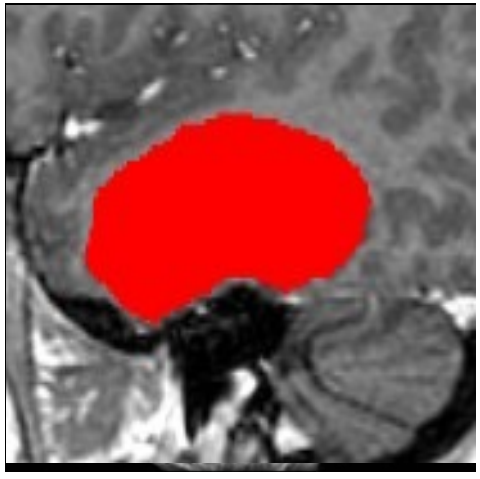
Original Image



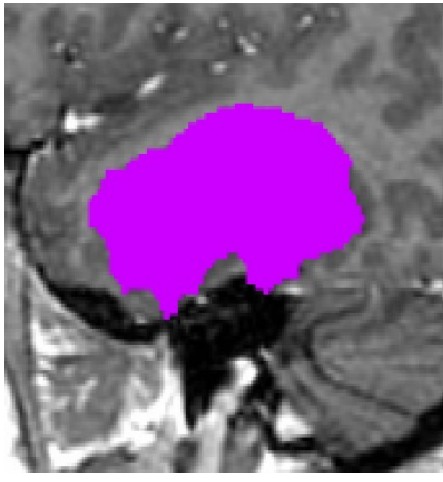
Expert 1



Expert 2



Expert 3



Expert 4

Segmentation Evaluation

Can be considered to consist of two components:

(1) Theoretical

Study mathematical equivalence among algorithms.

(2) Empirical

Study practical performance of algorithms in specific application domains.

Segmentation Evaluation: Theoretical

Fundamental challenges in segmentation evaluation:

(Ch1) Are major *pI (purely Image based)* frameworks such as active contours, level sets, graph cuts, fuzzy connectedness, watersheds, truly distinct or some level of equivalence exists among them?

Segmentation Evaluation: Theoretical

Fundamental challenges in segmentation evaluation:

(Ch1) Are major *pI (purely Image based)* frameworks such as active contours, level sets, graph cuts, fuzzy connectedness, watersheds, truly distinct or some level of equivalence exists among them?

(Ch2) How to develop truly distinct methods constituting real advance?

Segmentation Evaluation: Theoretical

Fundamental challenges in segmentation evaluation:

(Ch1) Are major *pI (purely Image based)* frameworks such as active contours, level sets, graph cuts, fuzzy connectedness, watersheds, truly distinct or some level of equivalence exists among them?

(Ch2) How to develop truly distinct methods constituting real advance?

(Ch3) How to choose a method for a given application domain?

Segmentation Evaluation: Theoretical

Fundamental challenges in segmentation evaluation:

(Ch1) Are major *pI (purely Image based)* frameworks such as active contours, level sets, graph cuts, fuzzy connectedness, watersheds, truly distinct or some level of equivalence exists among them?

(Ch2) How to develop truly distinct methods constituting real advance?

(Ch3) How to choose a method for a given application domain?

(Ch4) How to set an algorithm optimally for an application domain?

Segmentation Evaluation: Theoretical

Fundamental challenges in segmentation evaluation:

(Ch1) Are major *pI (purely Image based)* frameworks such as active contours, level sets, graph cuts, fuzzy connectedness, watersheds, truly distinct or some level of equivalence exists among them?

(Ch2) How to develop truly distinct methods constituting real advance?

(Ch3) How to choose a method for a given application domain?

(Ch4) How to set an algorithm optimally for an application domain?

Currently any method A can be shown empirically to be better than any method B, even when they are equivalent.

Segmentation Evaluation: Theoretical

Attributes commonly used by segmentation methods:

(1) Connectedness
(2) Texture
(3) Smoothness of boundary
(4) Shape information about object
(5) Noise handling
(6) Optimization employed
(7) Orientedness of boundary

Segmentation Evaluation: Empirical

T : A task - Example: Estimating the volume of brain.

B : A body region - Example: Head.

P : Imaging protocol - Example: $T2$ weighted MR imaging with a particular set of parameters.

Application domain: A particular triple $\langle T, B, P \rangle$.

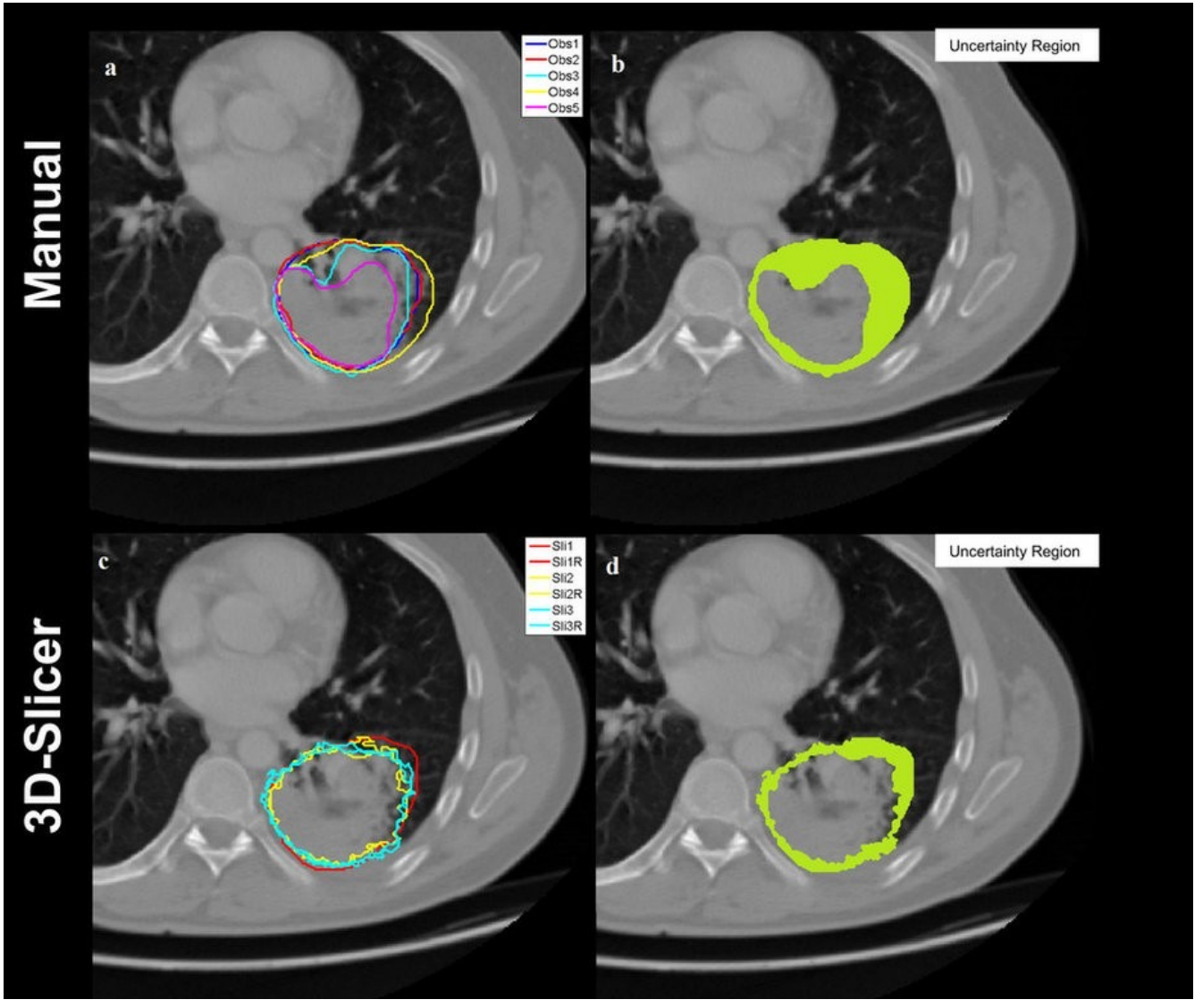
Q: A set of scenes acquired for a particular application domain $\langle T, B, P \rangle$.

Validation of Image Segmentation

- Spectrum of accuracy versus realism in reference standard.
- Digital phantoms.
  - Ground truth known accurately.
  - Not so realistic.
- Acquisitions and careful segmentation.
  - Some uncertainty in ground truth.
  - More realistic.
- .
- Clinical data ?
  - Hard to know ground truth.
  - Most realistic model.

Slide Credit: N. Archip

Velazquez et al, Scientific Reports 2013.



How to get the ground truth annotations?

- Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category.

- Averaging the Annotations

- Keeping the multiple annotations as it is.

Classification

- Classifiers always return the probability (0-1)
- Apply threshold on classifiers probabilities to detect interested pixels or regions
- Usually we use threshold =0.5

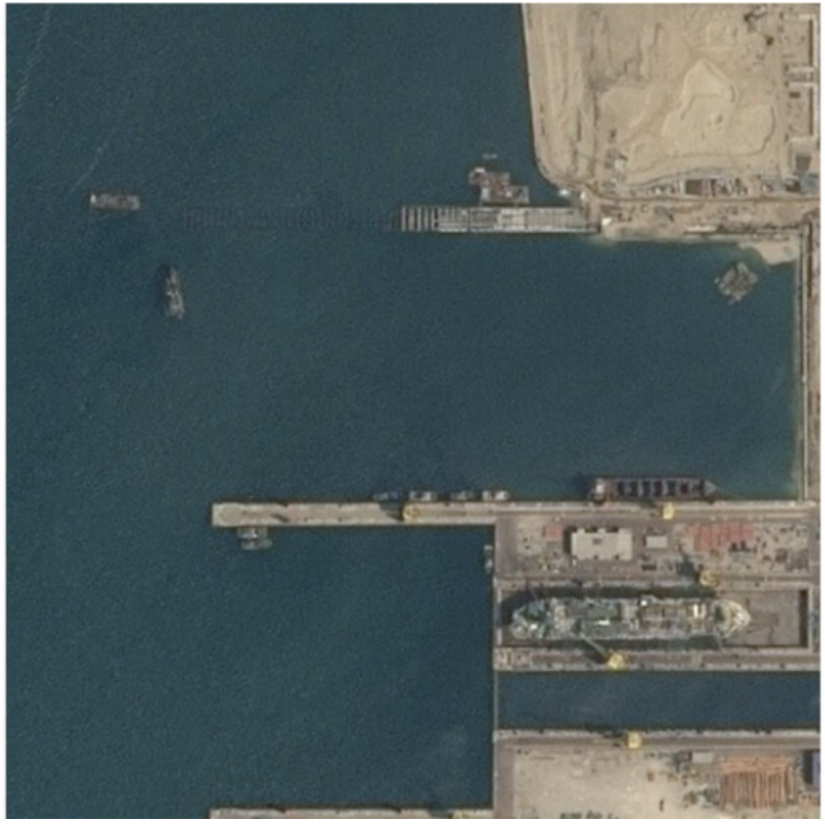
Pixel Accuracy

Pixel accuracy is perhaps the easiest to understand conceptually. *It is the percent of pixels in your image that are classified correctly.*

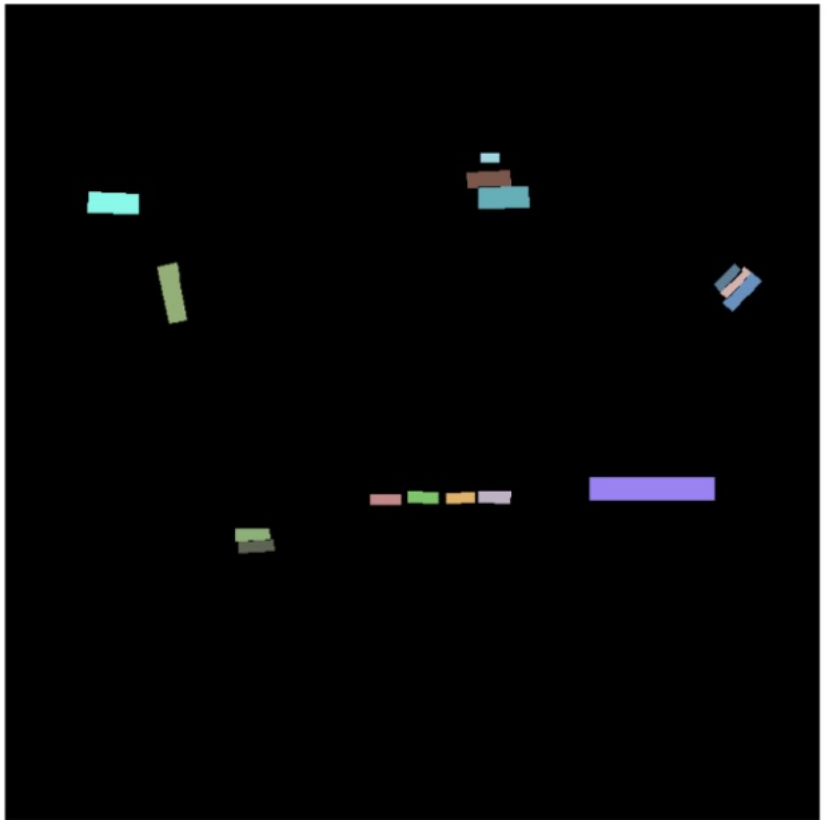
Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or No class imbalance.

Pixel Accuracy

Pixel accuracy is perhaps the easiest to understand conceptually. *It is the percent of pixels in your image that are classified correctly.*



Image

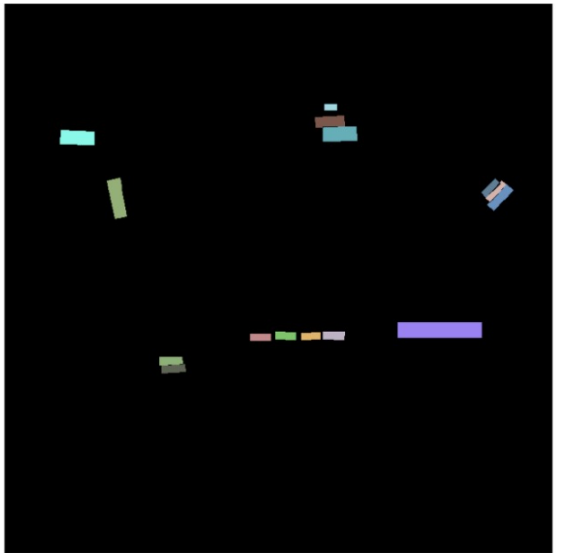


GT

Pixel Accuracy (Issue)

Pixel accuracy is perhaps the easiest to understand conceptually. *It is the percent of pixels in your image that are classified correctly.*

Prediction



GT

95% of pixels are classified accurately while the other 5% are not. As a result, although your accuracy is a whopping 95%, your model is returning a completely useless prediction. This is meant to illustrate that high pixel accuracy doesn't always imply superior segmentation ability.

This issue is called **class imbalance**.

Postive Prediction

Postive Prediction

True Postive Prediction

Postive Prediction

False Postive Prediction

Negative Prediction

Negative Prediction

True Negative Prediction

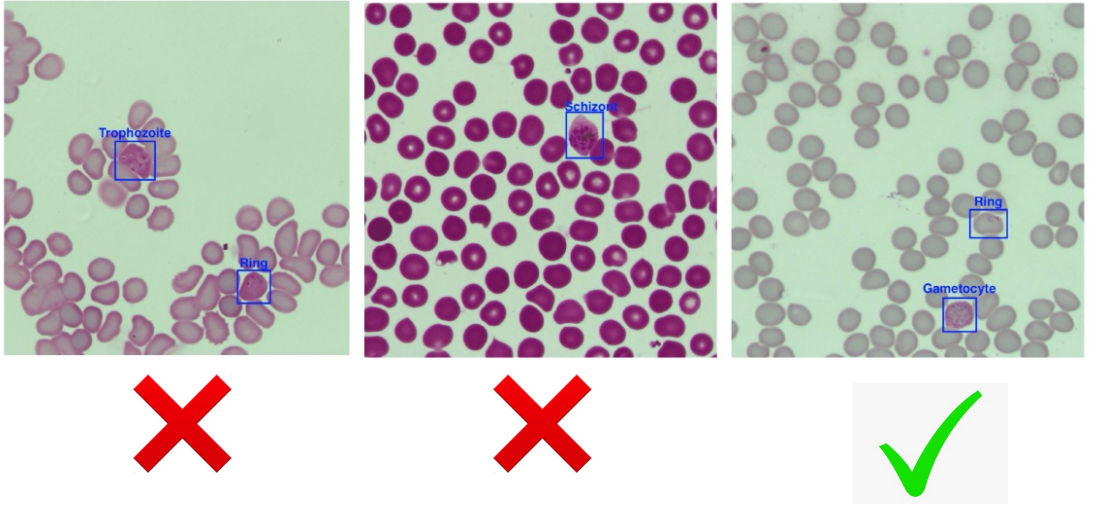
Negative Prediction

False Negative Prediction

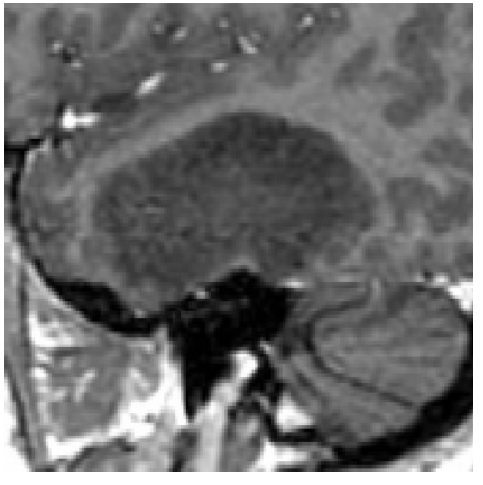
Ground Truth----Prediction

Higher False negative is acceptable

Blood sample images of the same person



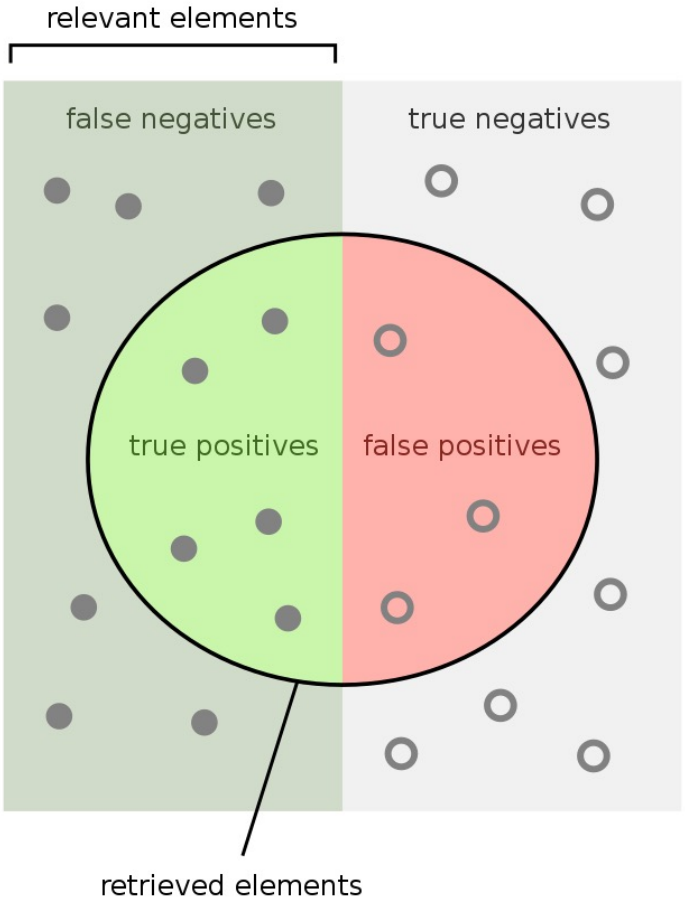
In case of tumor detection, higher False negative Not is acceptable

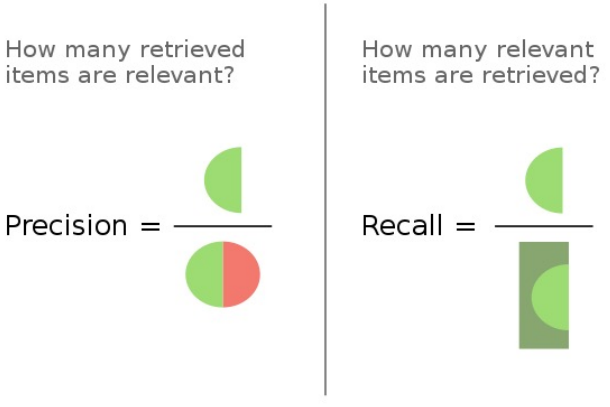


Precision

Precision is the ability of a model to identify **only** the relevant objects. It is the percentage of correct positive predictions and is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{all detections}}$$

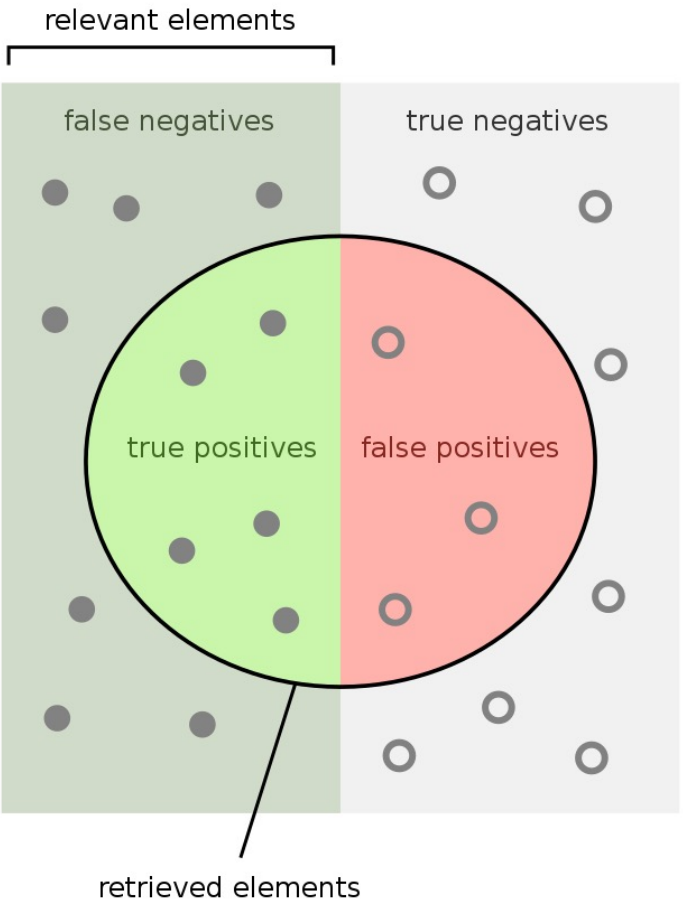


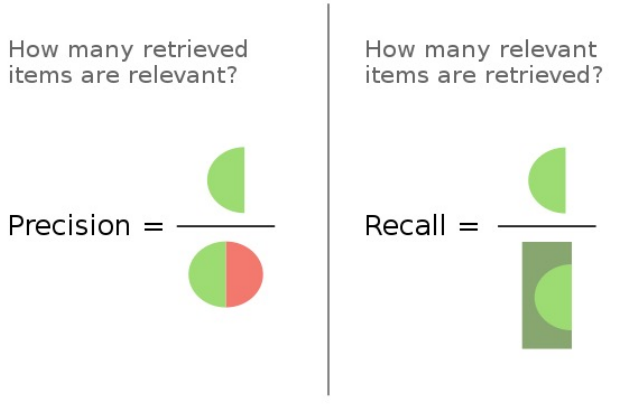


Recall

Recall is the ability of a model to find all the relevant cases. It is the percentage of true positive detected among all relevant ground truths and is given by:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{all ground truths}}$$

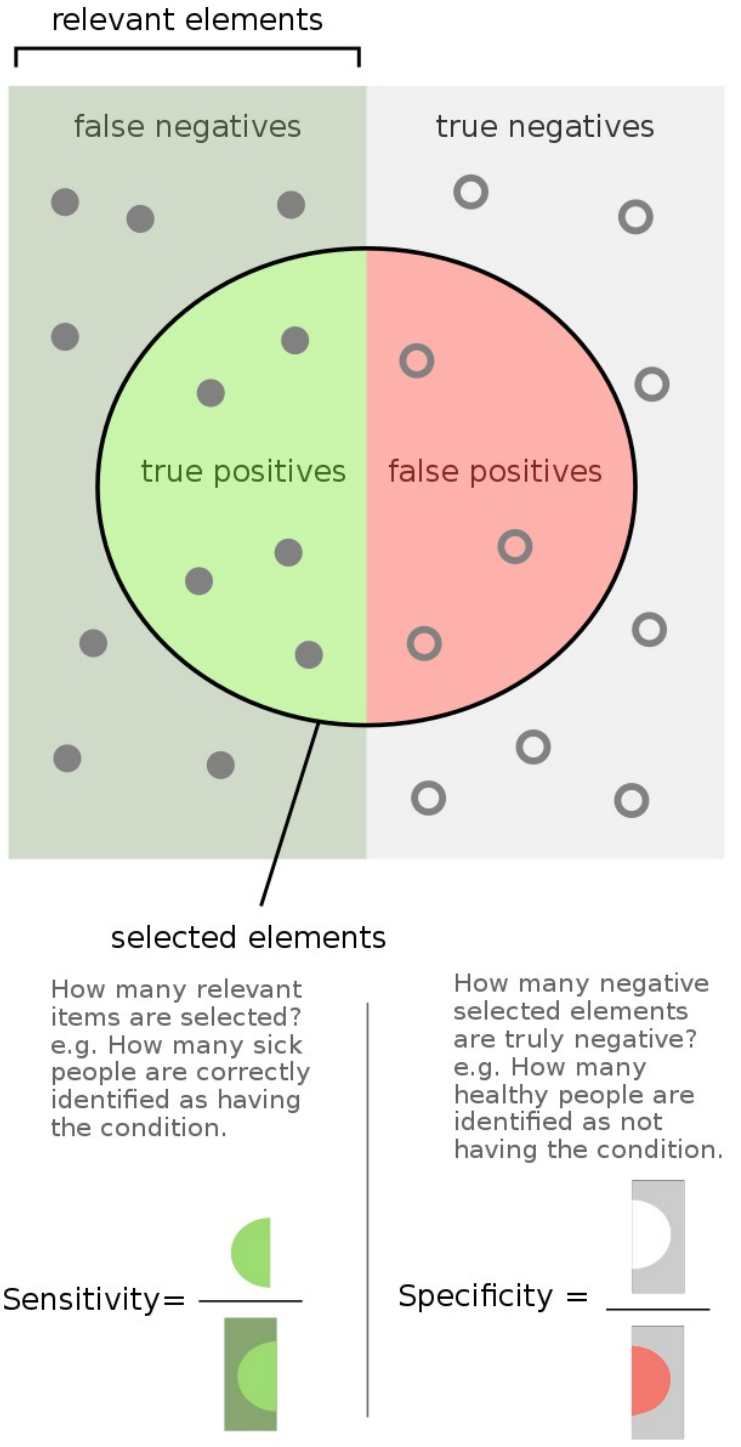




Sensitivity=True Positive Rate =Recall

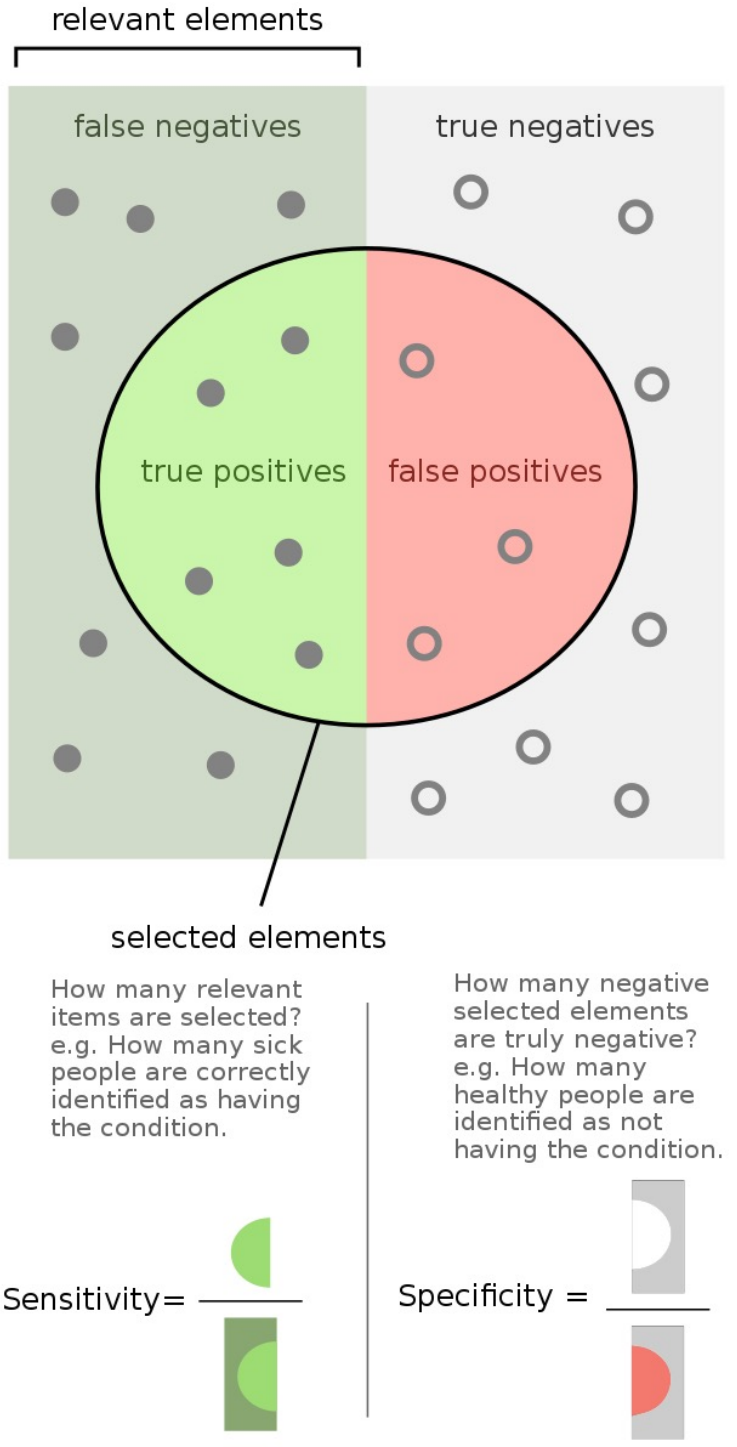
refers to the proportion of those who received a positive result on this test out of those who actually have the condition：

$$\text{Sensitivity= TPR=}\frac{TP}{TP+FN}$$



Specificity=True Negative Rate

$$\text{Specificity= TNR}=\frac{TN}{TN+FP}$$



Sensitivity and specificity

- In a diagnostic test, sensitivity is a measure of how well a test can identify **true positives** and specificity is a measure of how well a test can identify **true negatives**.

- For all testing, both diagnostic and screening, there is usually a trade-off between sensitivity and specificity, such that **higher sensitivities will mean lower specificities (due to more false positive)** and vice versa.

- **If the goal of the test is to identify everyone who has a condition, the number of false negatives should be low (you don't want to miss anyone)**, which requires high sensitivity. That is, people who have the condition should be highly likely to be identified as such by the test. This is especially important when the consequence of failing to treat the condition are serious and/or the treatment is very effective and has minimal side effects.

- **If the goal of the test is to accurately identify people who do not have the condition, the number of false positives should be very low (you don't want to select a wrong person)**, which requires a high specificity. That is, people who do not have the condition should be highly likely to be excluded by the test. This is especially important when people who are identified as having a condition may be subjected to more testing, expense, stigma, anxiety, etc.

How to combine Precision and Recall

- Arithmetic Mean
- Geometric Mean
- Harmonic   Mean

How to combine Precision and Recall

- Arithmetic Mean

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

- Geometric mean

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- Harmonic mean

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1}.$$

How to combine Precision and Recall

- Arithmetic Mean

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

- Geometric mean

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- Harmonic mean

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n}\right)^{-1}.$$

$$H = \frac{2x_1 x_2}{x_1 + x_2}$$

How to combine Precision and Recall

- Arithmetic Mean

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$\frac{1}{2}(\text{Precision} + \text{Recall}) = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP}+\text{FP}} + \frac{\text{TP}}{\text{TP}+\text{FN}}\right)$$

- Geometric mean

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

$$(\text{Precision} \times \text{Recall})^{\frac{1}{2}} = \left(\frac{\text{TP}}{\text{TP}+\text{FP}} \times \frac{\text{TP}}{\text{TP}+\text{FN}}\right)^{\frac{1}{2}}$$

- Harmonic mean

$$H = \frac{2x_1 x_2}{x_1 + x_2}$$

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

How to combine Precision and Recall

* True Positive Rate (TPR), Recall, or Sensitivity

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

* Precision

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

* F-1 Score

$$\text{F1} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

F-Score: Combination of Precision and Recall

F-scores (usually) follow an F-distribution, which is said to be named after Ronald Fisher

Fbeta-Measure

$$Fbeta = ((1 + beta^2) * Precision * Recall) / (beta^2 * Precision + Recall)$$

Beta =1, We call F-1 Score

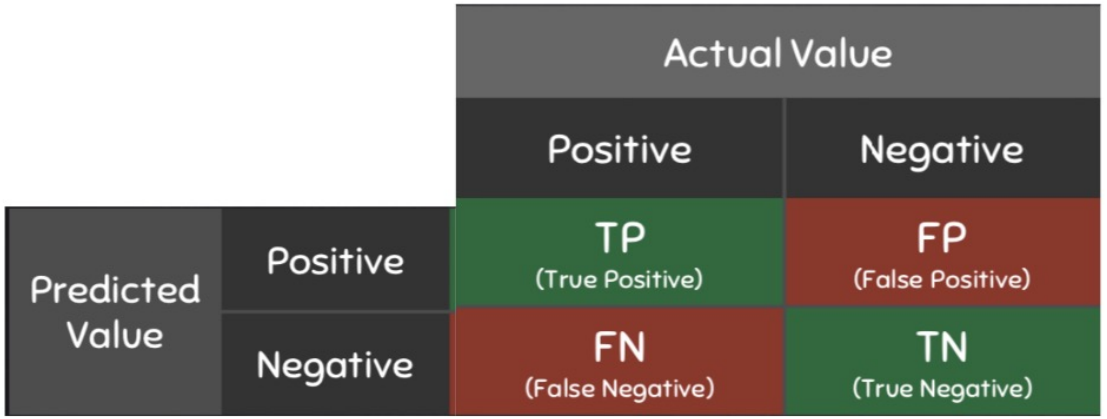
Beta =2, We call F-2 Score

Three common values for the beta parameter are as follows:

- **F0.5-Measure** (beta=0.5): More weight on precision, less weight on recall.
- **F1-Measure** (beta=1.0): Balance the weight on precision and recall.
- **F2-Measure** (beta=2.0): Less weight on precision, more weight on recall

Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.



- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Precision Recall Curve

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{TP}}{\text{all detections}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{all ground truths}}$$

Imbalance Data

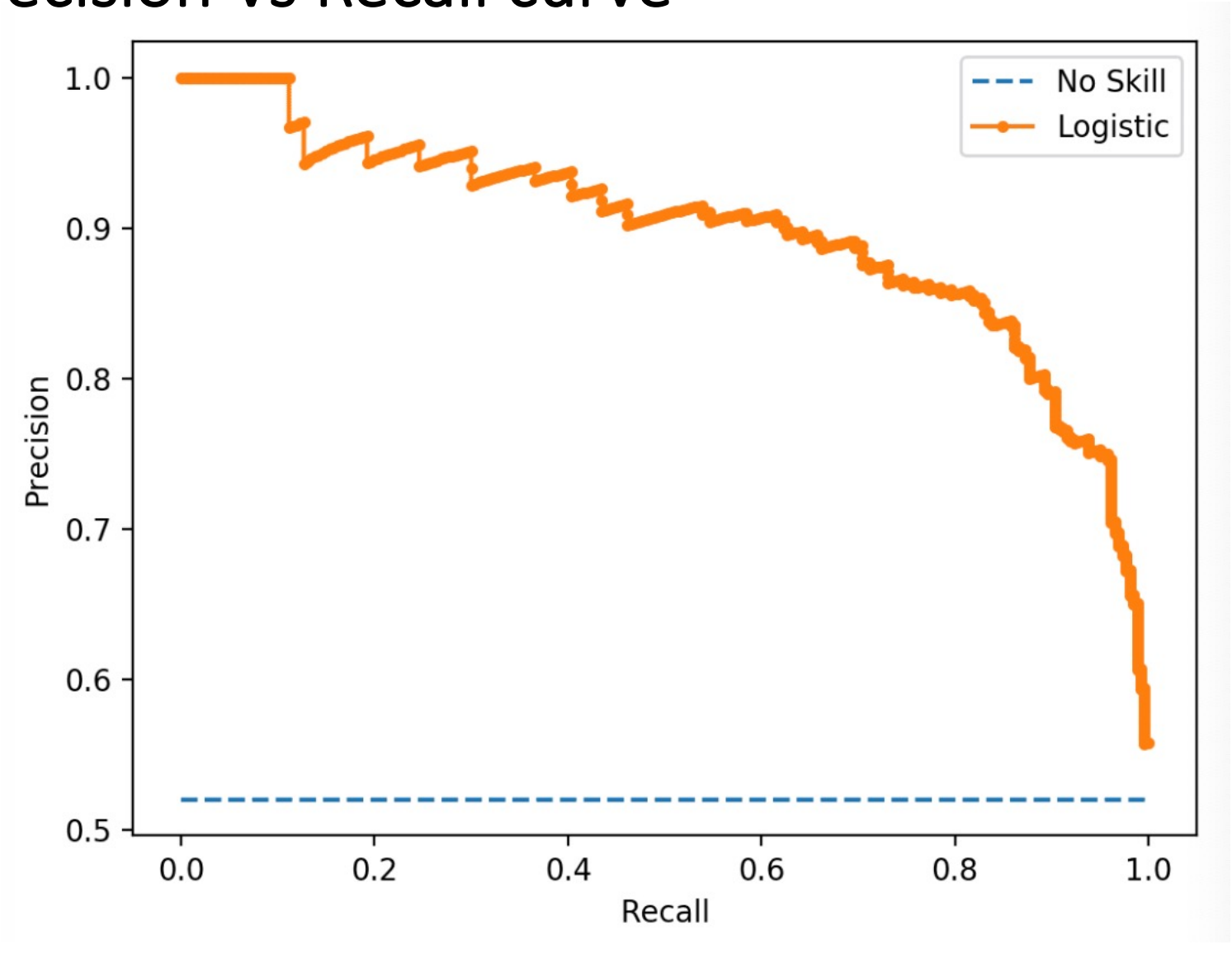
Receiver Operatoring Characteristics (ROC) Curve

Sensitivity= True Positive Rate=$\frac{TP}{TP+FN}$

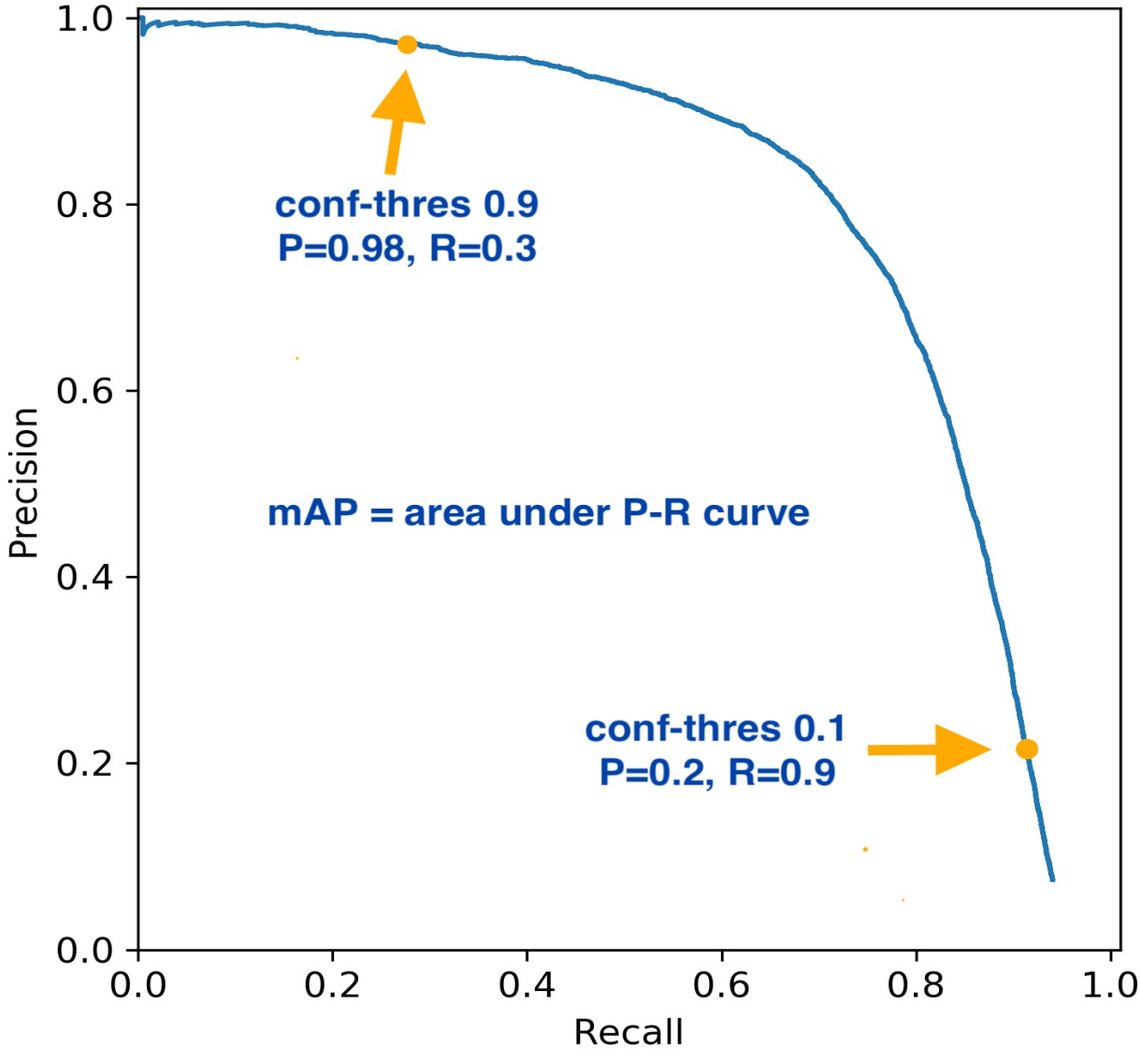
False Positive Rate= 1-Specificity=$\frac{FP}{FP+TN}$

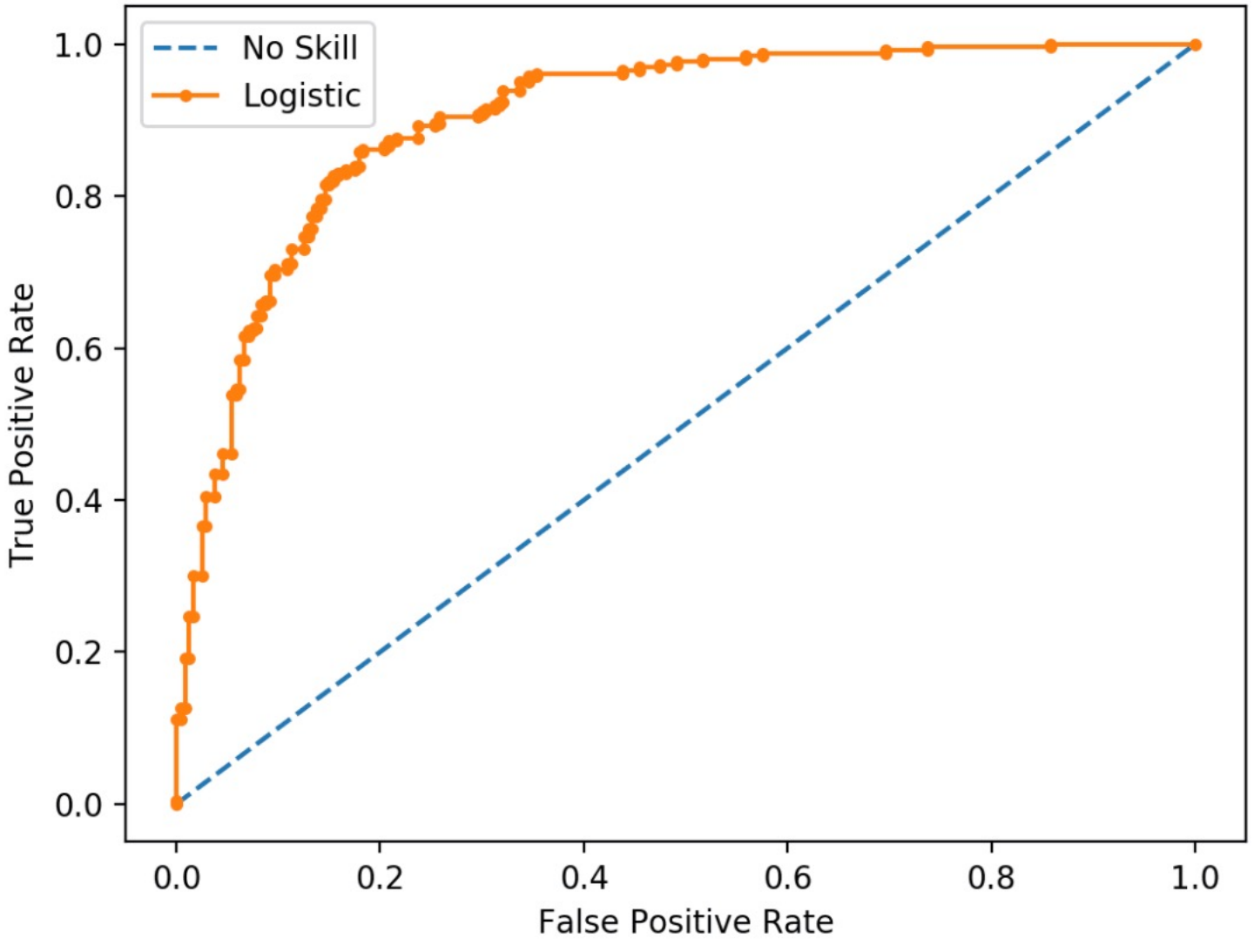
Balance Data

Precision Vs Recall curve

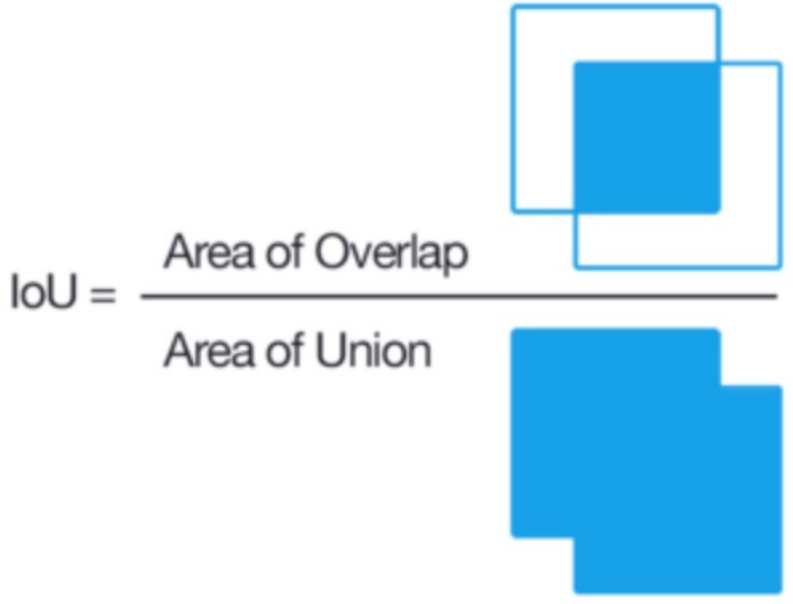


Precision Vs Recall curve




No Skill
Logistic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Intersection-over-Union (IOU, Jaccard Index)



This metric ranges from 0–1 (0–100%) with 0 signifying no overlap and 1 signifying perfectly overlapping segmentation

For **binary** (two classes) or **multi-class segmentation**, the mean IoU of the image is calculated by **taking the IoU of each class and averaging them**


IoU: 0.4034
IoU: 0.7330
IoU: 0.9264
Poor
Good
Excellent

Intersection-over-Union (IOU, Jaccard Index)

True Positive, False Positive, False Negative and True Negative
Some basic concepts used by the metrics:
•**True Positive (TP)**: A correct detection. Detection with IOU ≥ *threshold*

•**False Positive (FP)**: A wrong detection. Detection with IOU < *threshold*

•**False Negative (FN)**: A ground truth not detected

•**True Negative (TN)**: Does not apply. It would represent a corrected misdetection. In the object detection task there are many possible bounding boxes that should not be detected within an image. Thus, TN would be all possible bounding boxes that were correctly not detected (so many possible boxes within an image). That's why it is not used by the metrics.

threshold: depending on the metric, it is usually set to 50%, 75% or 95%.

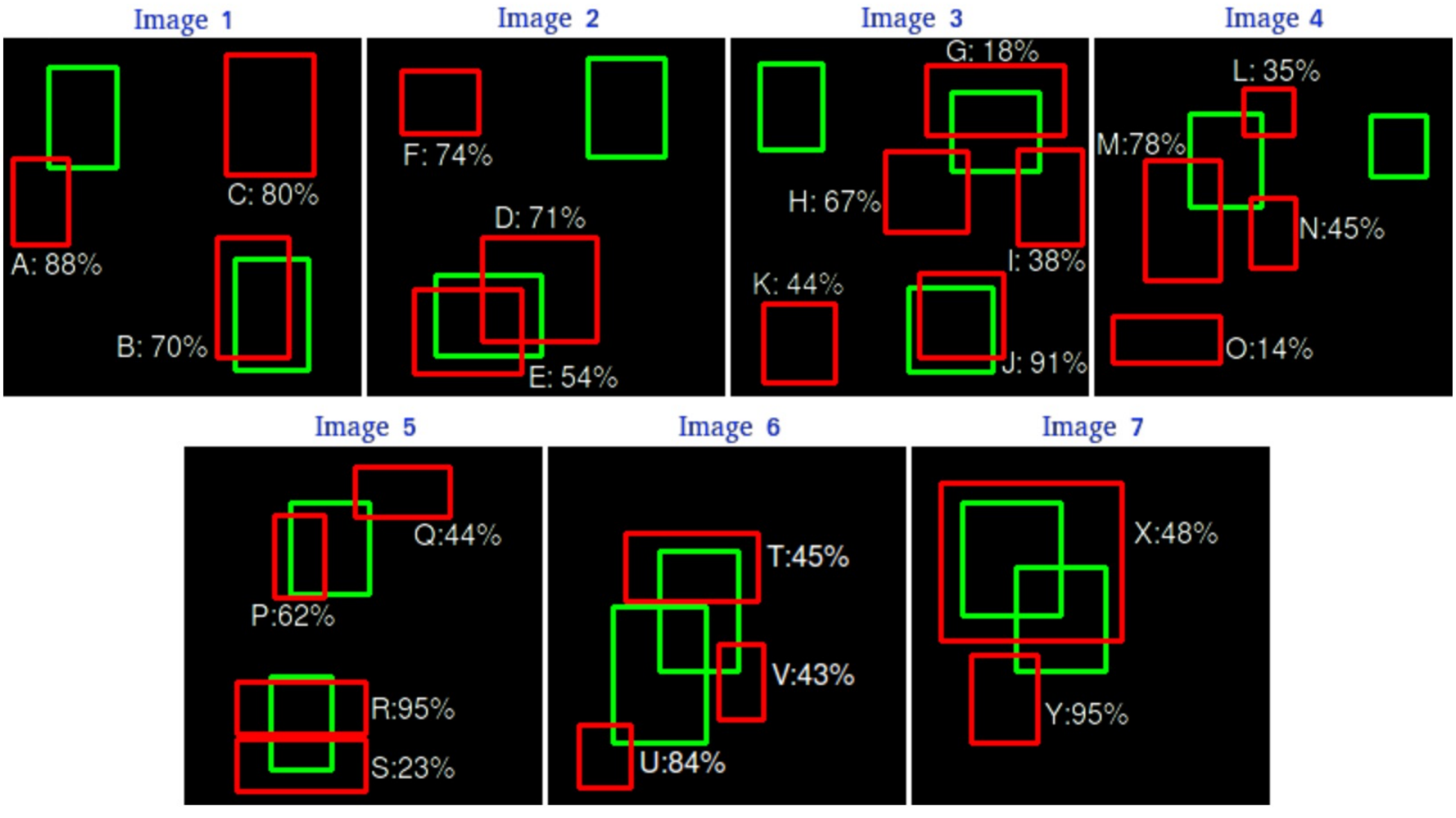
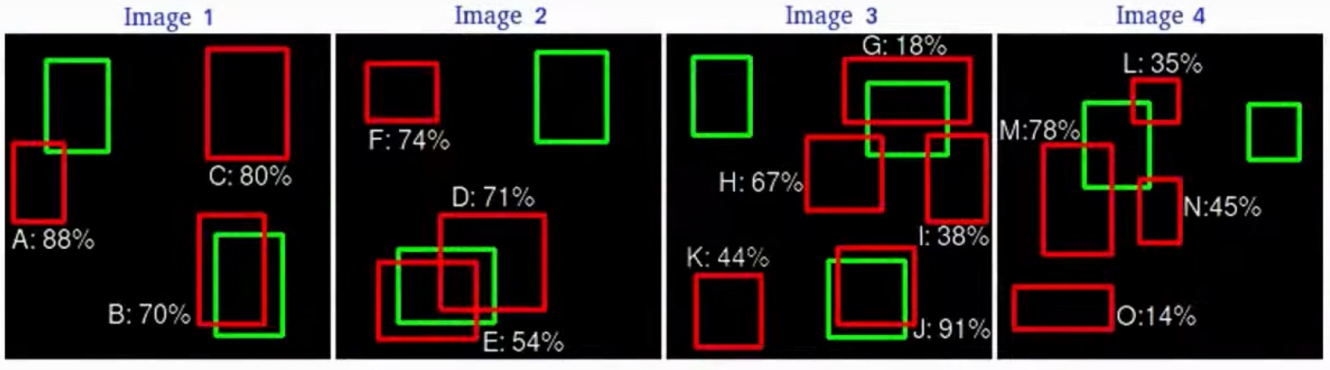
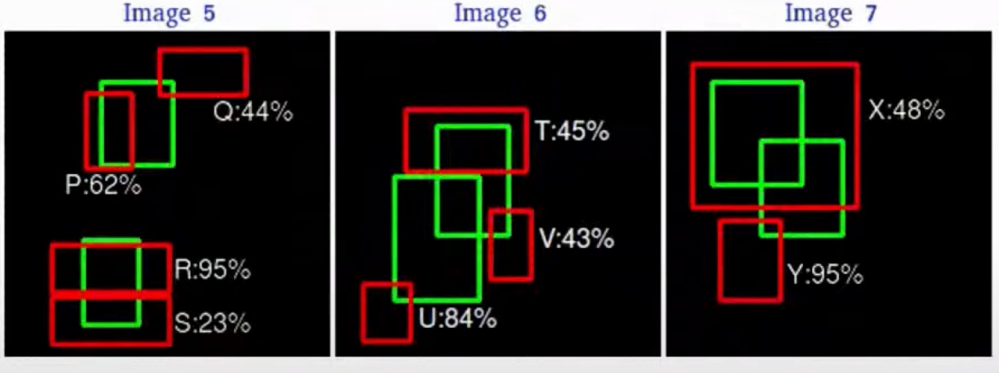


Image 1
A: 88%
B: 70%
C: 80%
Image 2
F: 74%
D: 71%
E: 54%
Image 3
G: 18%
H: 67%
I: 38%
K: 44%
J: 91%
Image 4
L: 35%
M:78%
N:45%
O:14%
Image 5
Q:44%
P:62%
R:95%
S:23%
Image 6
T:45%
V:43%
U:84%
Image 7
X:48%
Y:95%

mAP





Hat Tip to Rafael: https://github.com/rafaelpadilla/Object-Detection-Metrics

15 Ground Truth

24 Predictions

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

Precision Recall Curve

An object detector of a particular class is considered good if its precision stays high as recall increases, which means that if you vary the confidence threshold, the precision and recall will still be high. Another way to identify a good object detector is to look for a detector that can identify only relevant objects (0 False Positives = high precision), finding all ground truth objects (0 False Negatives = high recall).

A poor object detector needs to increase the number of detected objects (increasing False Positives = lower precision) in order to retrieve all ground truth objects (high recall). That's why the Precision x Recall curve usually starts with high precision values, decreasing as recall increases.

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

Images	Detections	Confidences
Image 5	R	95%
Image 7	Y	95%
Image 3	J	91%
Image 1	A	88%
Image 6	U	84%
Image 1	C	80%
Image 4	M	78%
Image 2	F	74%
Image 2	D	71%
Image 1	B	70%
Image 3	H	67%
Image 5	P	62%
Image 2	E	54%
Image 7	X	48%
Image 4	N	45%
Image 6	T	45%
Image 3	K	44%
Image 5	Q	44%
Image 6	V	43%
Image 3	I	38%
Image 4	L	35%
Image 5	S	23%
Image 3	G	18%
Image 4	O	14%

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

Images	Detections	Confidences	TP	FP
Image 5	R	95%	1	0
Image 7	Y	95%	0	1
Image 3	J	91%	1	0
Image 1	A	88%	0	1
Image 6	U	84%	0	1
Image 1	C	80%	0	1
Image 4	M	78%	0	1
Image 2	F	74%	0	1
Image 2	D	71%	0	1
Image 1	B	70%	1	0
Image 3	H	67%	0	1
Image 5	P	62%	1	0
Image 2	E	54%	1	0
Image 7	X	48%	1	0
Image 4	N	45%	0	1
Image 6	T	45%	0	1
Image 3	K	44%	0	1
Image 5	Q	44%	0	1
Image 6	V	43%	0	1
Image 3	I	38%	0	1
Image 4	L	35%	0	1
Image 5	S	23%	0	1
Image 3	G	18%	1	0
Image 4	O	14%	0	1

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

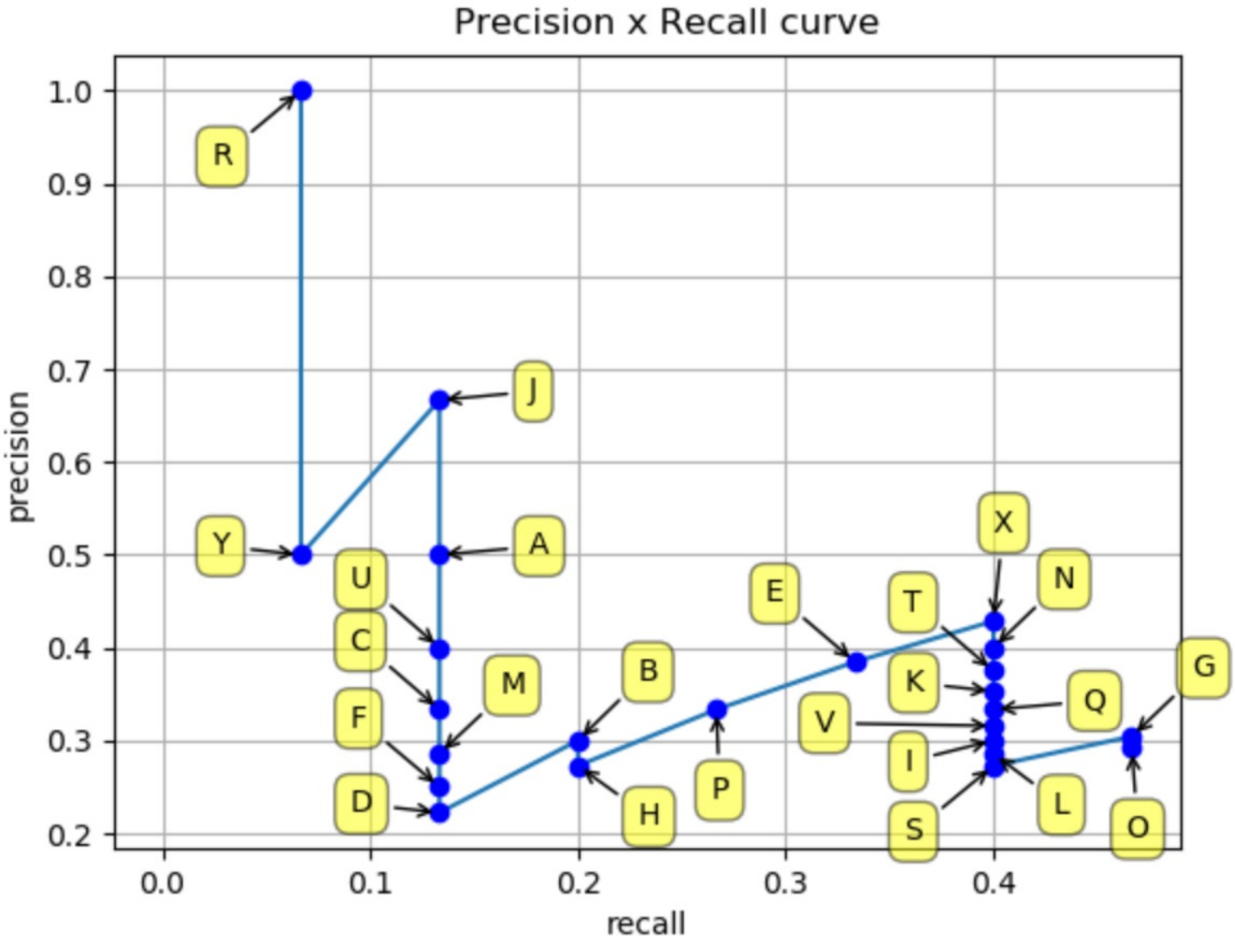
Images	Detections	Confidences	TP	FP	Acc TP
Image 5	R	95%	1	0	1
Image 7	Y	95%	0	1	1
Image 3	J	91%	1	0	2
Image 1	A	88%	0	1	2
Image 6	U	84%	0	1	2
Image 1	C	80%	0	1	2
Image 4	M	78%	0	1	2
Image 2	F	74%	0	1	2
Image 2	D	71%	0	1	2
Image 1	B	70%	1	0	3
Image 3	H	67%	0	1	3
Image 5	P	62%	1	0	4
Image 2	E	54%	1	0	5
Image 7	X	48%	1	0	6
Image 4	N	45%	0	1	6
Image 6	T	45%	0	1	6
Image 3	K	44%	0	1	6
Image 5	Q	44%	0	1	6
Image 6	V	43%	0	1	6
Image 3	I	38%	0	1	6
Image 4	L	35%	0	1	6
Image 5	S	23%	0	1	6
Image 3	G	18%	1	0	7
Image 4	O	14%	0	1	7

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

Images	Detections	Confidences	TP	FP	Acc TP	Acc FP
Image 5	R	95%	1	0	1	0
Image 7	Y	95%	0	1	1	1
Image 3	J	91%	1	0	2	1
Image 1	A	88%	0	1	2	2
Image 6	U	84%	0	1	2	3
Image 1	C	80%	0	1	2	4
Image 4	M	78%	0	1	2	5
Image 2	F	74%	0	1	2	6
Image 2	D	71%	0	1	2	7
Image 1	B	70%	1	0	3	7
Image 3	H	67%	0	1	3	8
Image 5	P	62%	1	0	4	8
Image 2	E	54%	1	0	5	8
Image 7	X	48%	1	0	6	8
Image 4	N	45%	0	1	6	9
Image 6	T	45%	0	1	6	10
Image 3	K	44%	0	1	6	11
Image 5	Q	44%	0	1	6	12
Image 6	V	43%	0	1	6	13
Image 3	I	38%	0	1	6	14
Image 4	L	35%	0	1	6	15
Image 5	S	23%	0	1	6	16
Image 3	G	18%	1	0	7	16
Image 4	O	14%	0	1	7	17

Images	Detections	Confidences	TP or FP
Image 1	A	88%	FP
Image 1	B	70%	TP
Image 1	C	80%	FP
Image 2	D	71%	FP
Image 2	E	54%	TP
Image 2	F	74%	FP
Image 3	G	18%	TP
Image 3	H	67%	FP
Image 3	I	38%	FP
Image 3	J	91%	TP
Image 3	K	44%	FP
Image 4	L	35%	FP
Image 4	M	78%	FP
Image 4	N	45%	FP
Image 4	O	14%	FP
Image 5	P	62%	TP
Image 5	Q	44%	FP
Image 5	R	95%	TP
Image 5	S	23%	FP
Image 6	T	45%	FP
Image 6	U	84%	FP
Image 6	V	43%	FP
Image 7	X	48%	TP
Image 7	Y	95%	FP

Images	Detections	Confidences	TP	FP	Acc TP	Acc FP	Precision	Recall
Image 5	R	95%	1	0	1	0	1	0.0666
Image 7	Y	95%	0	1	1	1	0.5	0.0666
Image 3	J	91%	1	0	2	1	0.6666	0.1333
Image 1	A	88%	0	1	2	2	0.5	0.1333
Image 6	U	84%	0	1	2	3	0.4	0.1333
Image 1	C	80%	0	1	2	4	0.3333	0.1333
Image 4	M	78%	0	1	2	5	0.2857	0.1333
Image 2	F	74%	0	1	2	6	0.25	0.1333
Image 2	D	71%	0	1	2	7	0.2222	0.1333
Image 1	B	70%	1	0	3	7	0.3	0.2
Image 3	H	67%	0	1	3	8	0.2727	0.2
Image 5	P	62%	1	0	4	8	0.3333	0.2666
Image 2	E	54%	1	0	5	8	0.3846	0.3333
Image 7	X	48%	1	0	6	8	0.4285	0.4
Image 4	N	45%	0	1	6	9	0.4	0.4
Image 6	T	45%	0	1	6	10	0.375	0.4
Image 3	K	44%	0	1	6	11	0.3529	0.4
Image 5	Q	44%	0	1	6	12	0.3333	0.4
Image 6	V	43%	0	1	6	13	0.3157	0.4
Image 3	I	38%	0	1	6	14	0.3	0.4
Image 4	L	35%	0	1	6	15	0.2857	0.4
Image 5	S	23%	0	1	6	16	0.2727	0.4
Image 3	G	18%	1	0	7	16	0.3043	0.4666
Image 4	O	14%	0	1	7	17	0.2916	0.4666

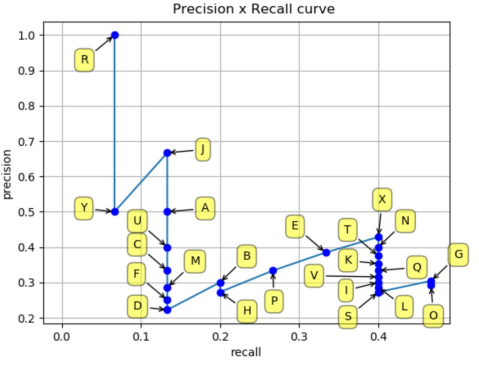

Precision x Recall curve
precision
recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.0
0.1
0.2
0.3
0.4
R
J
Y
A
U
C
M
F
D
B
H
P
E
X
T
N
K
Q
V
G
I
L
S
O

Average Precision

Another way to compare the performance of object detectors is to calculate the area under the curve (AUC) of the Precision x Recall curve. As AP curves are often zigzag curves going up and down, comparing different curves (different detectors) in the same plot usually is not an easy task - because the curves tend to cross each other much frequently. That's why Average Precision (AP), a numerical metric, can also help us compare different detectors. In practice AP is the precision averaged across all recall values between 0 and 1.

- **11-point interpolation**

- **Interpolating all points**



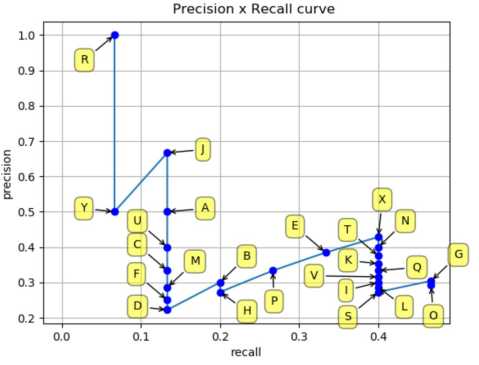
Precision x Recall curve

Precision

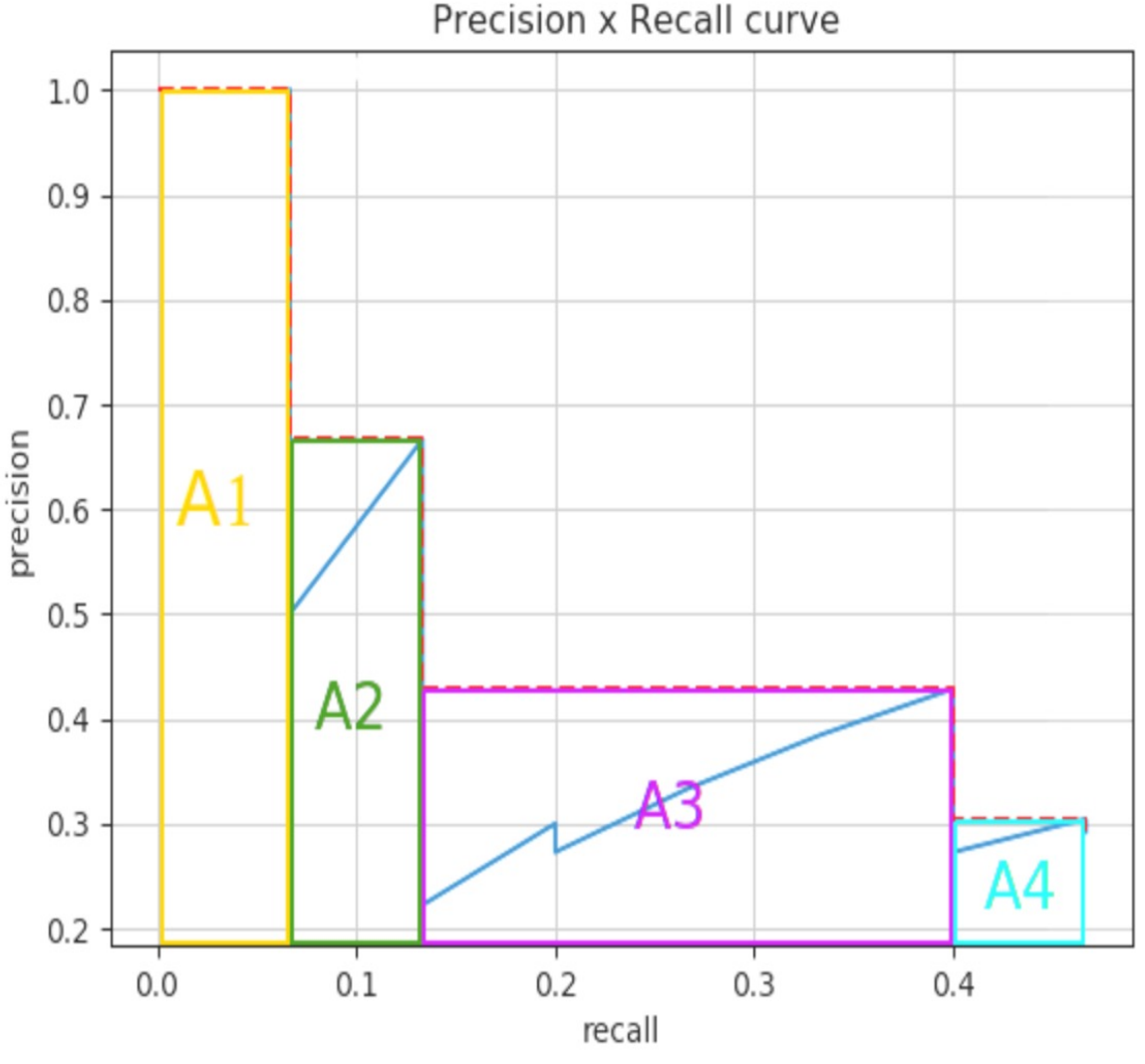
Interpolated precision

precision

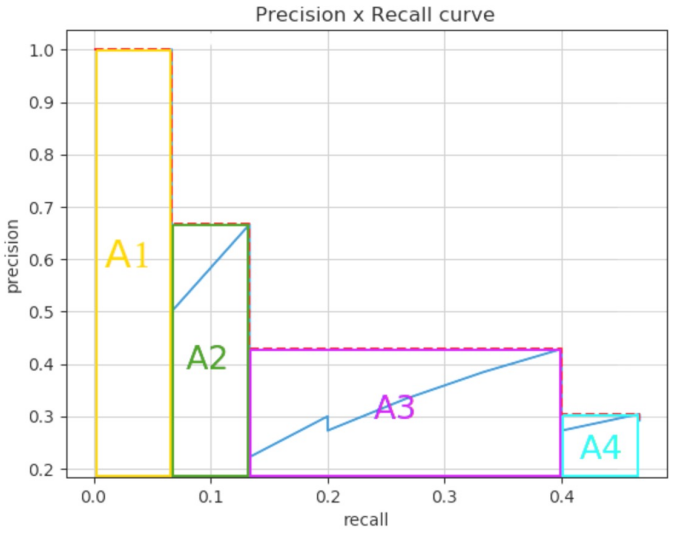
recall



Looking at the plot above, we can divide the AUC into 4 areas (A1, A2, A3 and A4):



Looking at the plot above, we can divide the AUC into 4 areas (A1, A2, A3 and A4):



Calculating the total area, we have the AP:

$$AP = A1 + A2 + A3 + A4$$

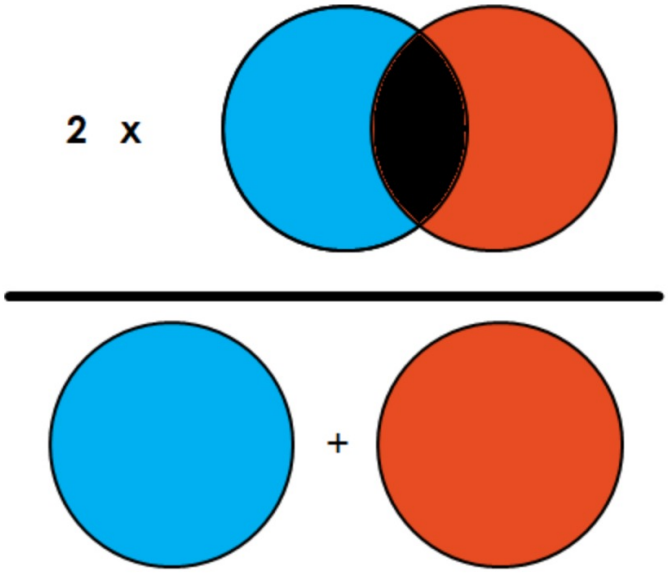
with:

$$A1 = (0.0666 - 0) \times 1 = \mathbf{0.0666}$$
$$A2 = (0.1333 - 0.0666) \times 0.6666 = \mathbf{0.04446222}$$
$$A3 = (0.4 - 0.1333) \times 0.4285 = \mathbf{0.11428095}$$
$$A4 = (0.4666 - 0.4) \times 0.3043 = \mathbf{0.02026638}$$

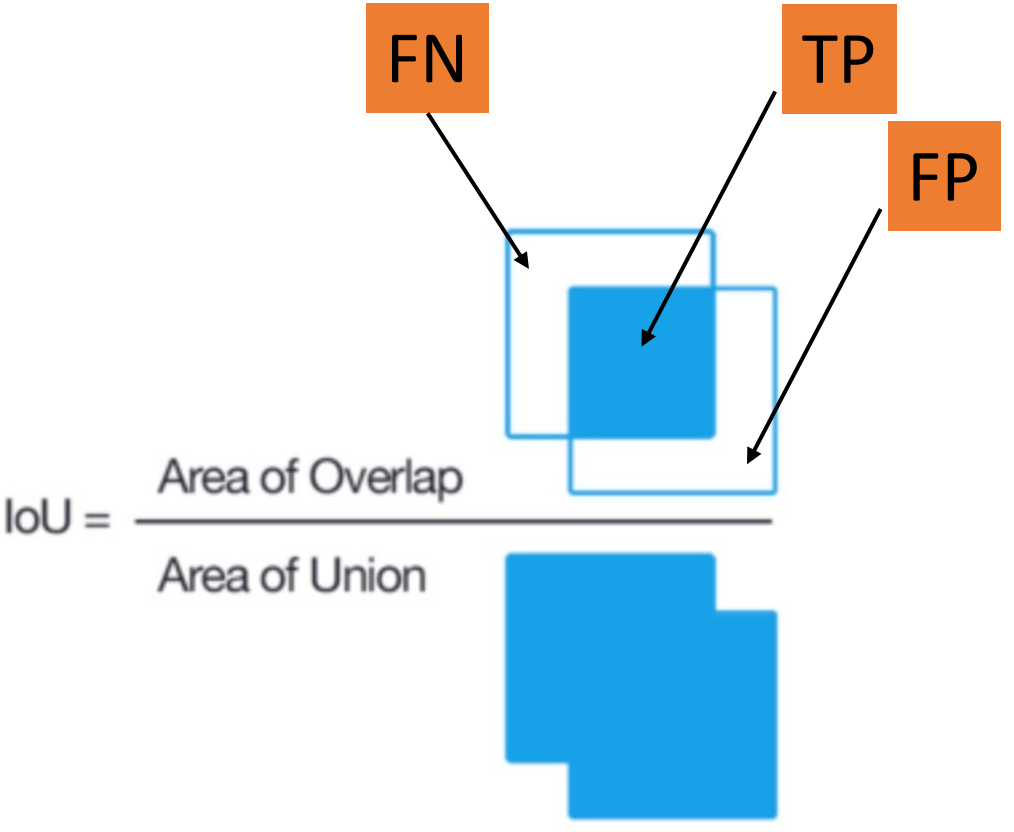
$$AP = 0.0666 + 0.04446222 + 0.11428095 + 0.02026638$$
$$AP = 0.24560955$$
$$AP = \mathbf{24.56}\%$$

Dice Co-efficient (F1-score)

Dice Coefficient is 2 * the Area of Overlap divided by the total number of pixels in both images



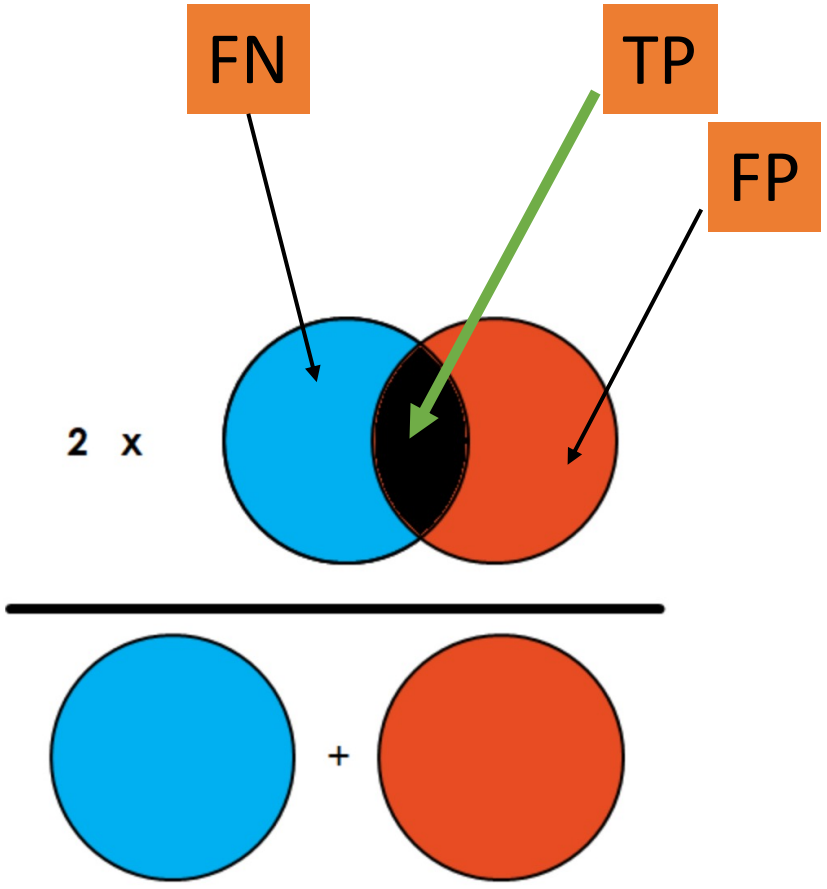
Intersection-over-Union (IOU, Jaccard Index)



Jaccard = $|A \cap B| / |A \cup B| = TP / (TP + FN+FP)$

Dice Co-efficient (F1-score)

Dice Coefficient is 2 * the Area of Overlap divided by the total number of pixels in both images



Dice = 2 |A∩B| / (|A|+|B|) = 2 TP / (2 TP + FP + FN)

Image Segmetation

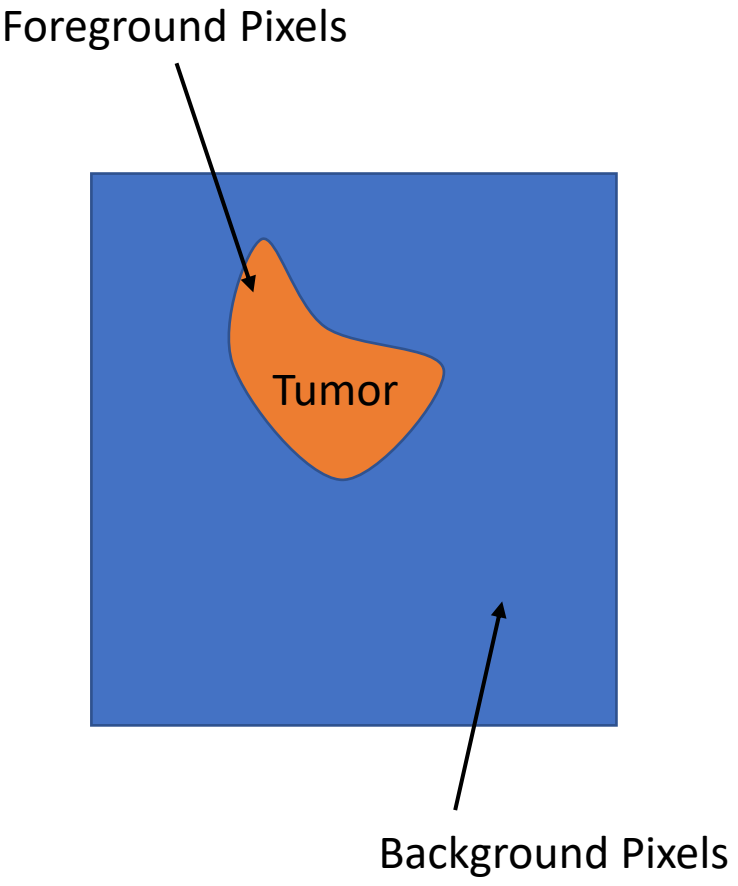


Image Segmetation

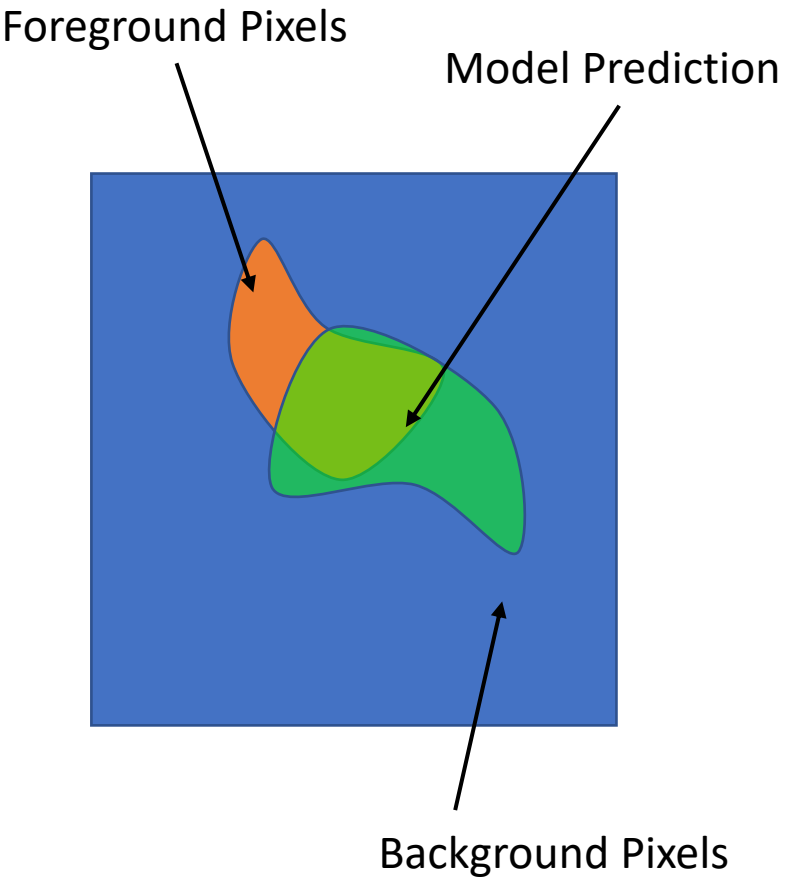


Image Segmetation

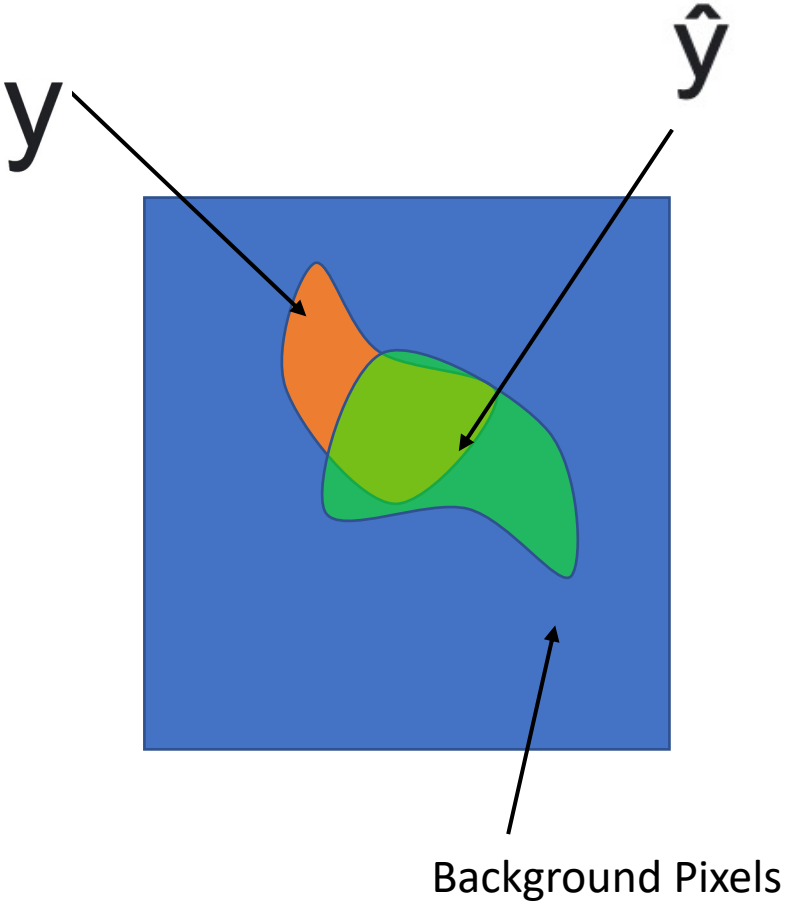


Image Segmetation

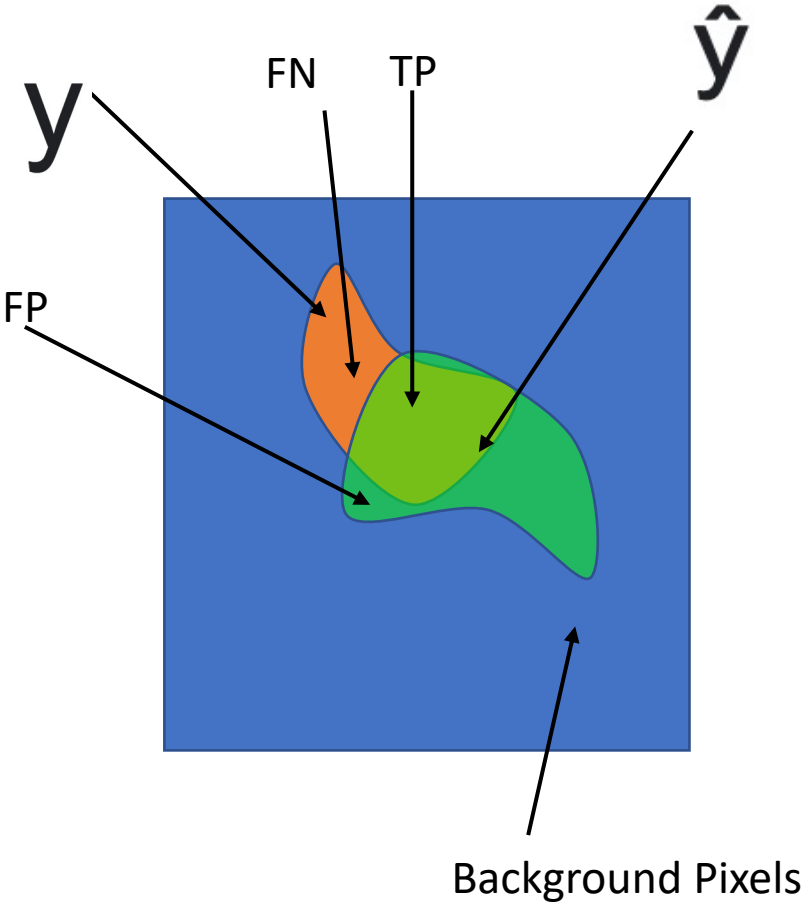
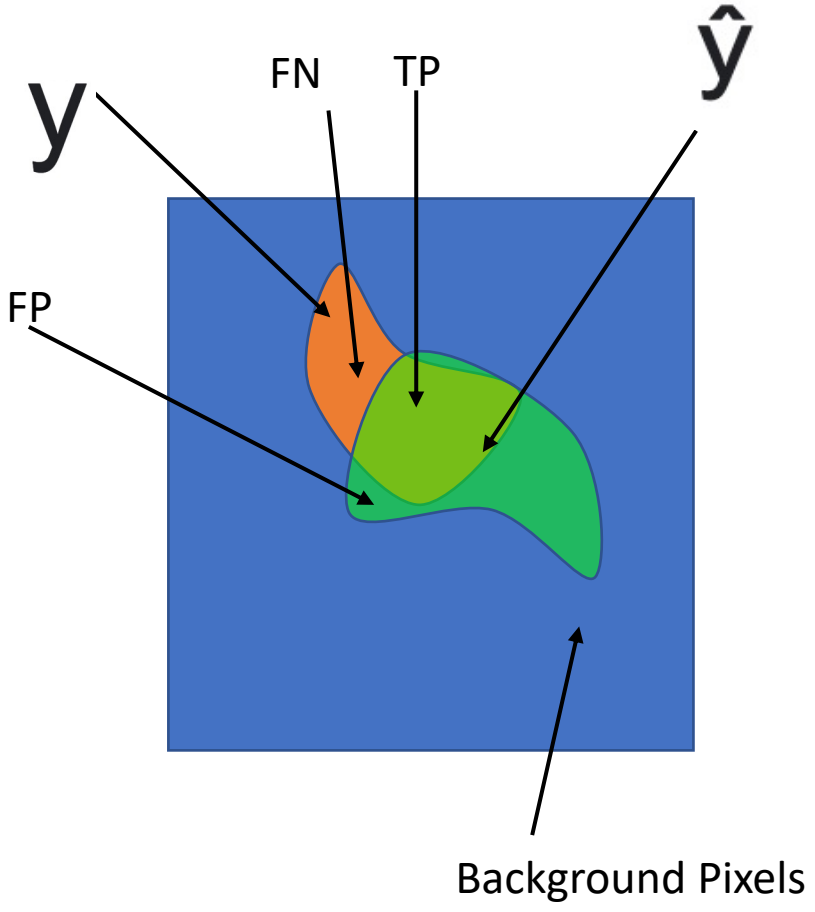


Image Segmetation



$$\text{IOU} = \frac{TP}{TP+FN+FP}$$

$$\text{IOU\_max} = \frac{TP}{TP}=1$$

$$\text{IOU\_min} = \frac{0}{0+FN+FP} = 0$$

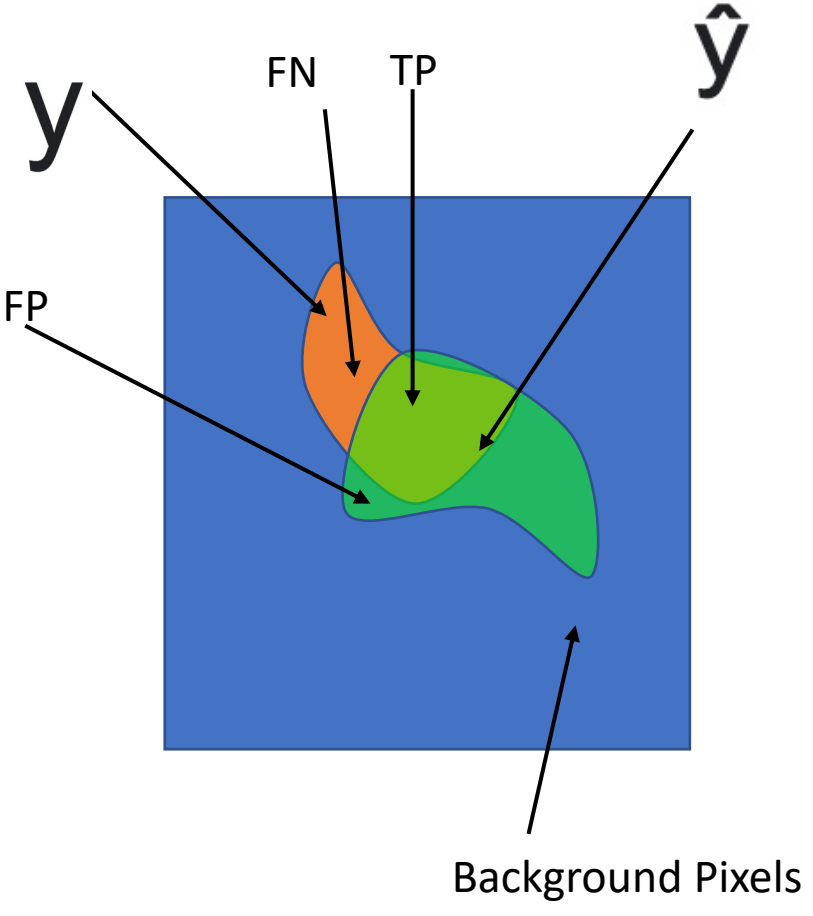
Loss fuction for IOU and Dice Co-efficient

$$\text{IOU} = \frac{\sum y\hat{}y}{\sum y\hat{}+y-y\hat{}y} \qquad \text{IOU} = \frac{TP}{TP+FN+FP}$$

$$\text{Dice} = \frac{2*\sum y\hat{}y}{\sum y\hat{}+y}$$

$$\text{Dice\_Loss} = 1 - \frac{2*\sum y\hat{}y}{\sum y\hat{}+y}$$

$$\text{IOU\_Loss} = 1 - \frac{\sum y\hat{}y}{\sum y\hat{}+y-y\hat{}y}$$



Assignement

https://github.com/rafaelpadilla/Object-Detection-Metrics

Regression

Mean Absolute Error

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don't gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data. Mathematically, it is represented as :

$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

Regression

Mean Squared Error

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

Image Comparision

MSE (Mean Square error)

The mean-square error (MSE) are used to compare image compression quality. The MSE represents the cumulative squared error between the compressed and the original image on the bases of pixels of the target and input images. It is a full reference metric the lower the value of MSE, the lower the error. The MSE introduces the Root-Mean-Square Error (RMSE) or Root-Mean-Square Deviation (RMSD) and often referred to as standard deviation of the variance.

Image Comparision

Peak Signal to Noise Ratio:

$$
\begin{aligned}
PSNR &= 10 \cdot \log_{10} \left(\frac{MAX_I^2}{MSE} \right) \\
&= 20 \cdot \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right) \\
&= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE)
\end{aligned}
$$

MAX_I =256

Image Enhancement

BRISQUE (Blind/Reference-less Image Spatial Quality Evaluator)

- Brisque calculates the no-reference image quality score for an image using the Blind/Reference-less Image Spatial Quality Evaluator (BRISQUE).
- The only input the algorithm gets is the image whose quality you want to measure. This is thus called, No-Reference or Objective-Blind.
- BRISQUE score is computed using Support Vector Regression (SVR) trained on an image with corresponding mean opinion score
- The database contains images with known distortion such as compression artifacts, blurring and noise.
- Ideally, the image to be scored must have at least one of the distortions for which the model was trained.

https://www.mathworks.com/help/images/ref/brisque.html

Image Enhancement

BRISQUE (Blind/Reference-less Image Spatial Quality Evaluator)

BRISQUE score for original image is 20.6586.
BRISQUE score for noisy image is 52.6074.
BRISQUE score for blurry image is 47.7553.

Image Enhancement

Original paper:
Sharpness Estimation for Document and Scene Images
by Jayant Kumar , Francine Chen , David Doermann

Uses difference of differences in grayscale values of a median-filtered image (ΔDoM) as an indicator of edge sharpness.

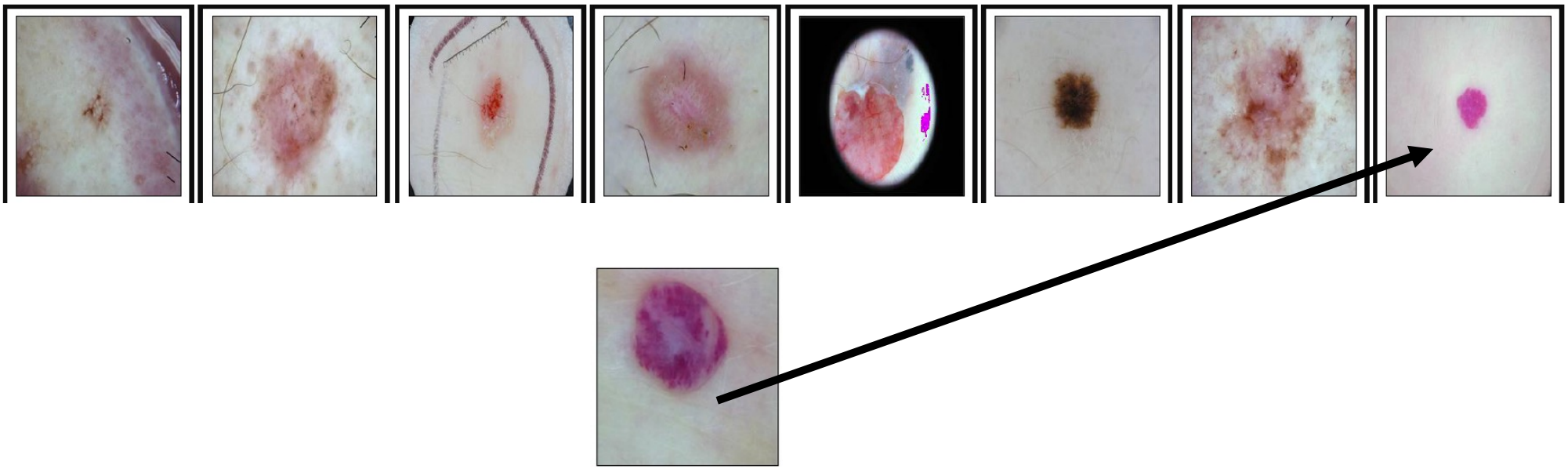
Image/Region/Cell Retrieval

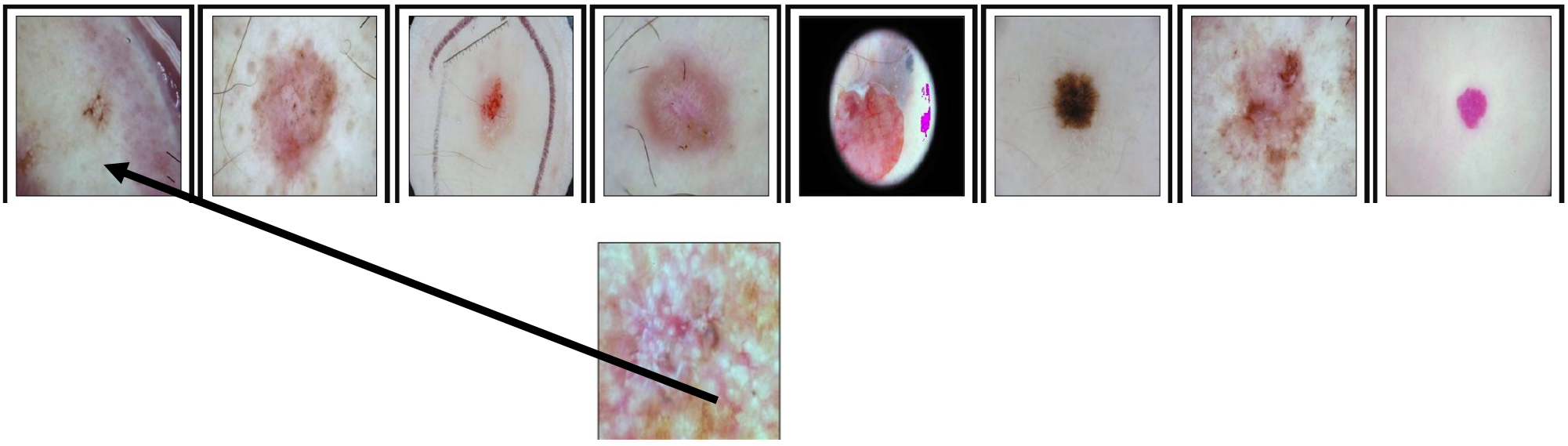
R@K

R@1	R@5	R@10	R@1%
0.54%	2.52%	4.48%	18.55%
0.68%	2.92%	5.06%	21.81%
0.63%	2.83%	5.03%	21.51%
1.39%	**6.50%**	**10.45%**	32.42%
1.80%	6.45%	10.36%	**34.38%**

- Query Image
- Reference Database

- Given the query image, if the ground truth image ranks in the first `K' most similar images, it is considered to be a correct query





Fake Data Generations

https://openaccess.thecvf.com/content_cvpr_2018/papers/Regmi_Cross-View_Image_Synthesis_CVPR_2018_paper.pdfs

Useful Links

F-Score:

- https://machinelearningmastery.com/fbeta-measure-for-machine-learning/
- https://www.youtube.com/watch?v=_OCYto4zK0g
- https://felipepenha.github.io/data-science-bits/performance_metrics/F1_score_unbalanced_granular.html
- https://en.wikipedia.org/wiki/Harmonic_mean
- https://stats.stackexchange.com/questions/300975/why-is-f-score-called-f-score

Useful Links

IOU

- https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/
- https://www.youtube.com/watch?v=EKm6cjnKaN8

- https://blog.paperspace.com/mean-average-precision/

- https://www.youtube.com/watch?v=QdWidmgLwbw
- https://www.youtube.com/watch?v=t98TA2RYQvw

Useful Links

Segmentation

- https://www.youtube.com/watch?v=AZr64OxshLo
- https://towardsdatascience.com/metrics-to-evaluate-your-semantic-segmentation-model-6bcb99639aa2

- Jayaram K. Udupa, MIPG of University of Pennsylvania, PA.
- P. Suetens, Fundamentals of Medical Imaging, Cambridge Univ. Press.
- N. Bryan, Intro. to the science of medical imaging, Cambridge Univ. Press.
- CAP 5415 Computer Vision (Fall 2016) Lecture Presentations
- Computer Vision (Lecture Presentations) by Dr. Mohsen Ali