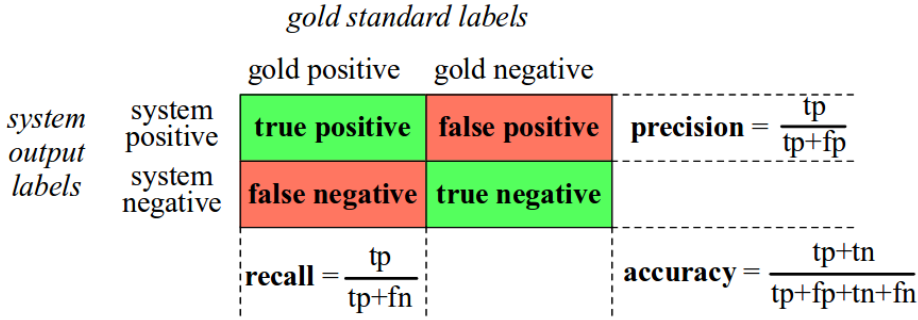
Performance Evaluation Metrics

- General
  - True positives, false positives, true negatives, false negatives
  - Percision/recall
  - F-measure, F1
  - ROC, AUC
  - Retrieval:Rank@K, Prec@K, mAP
- NLP: BLEU, ROUGE, METEOR, CIDEr
- Speech: Word Error Rate (WER)
- Signal quality: MSE, SNR, PSNR
- Image quality: Inception score
- Segmentation: Dice coefficient, Jaccard index

Performance metrics

- **Classification**
- Results of a prediction model can be summarized using a 2x2 confusion matrix depicting four possible scenarios



Performance metrics

- Classification of data samples in two categories e.g. spam detection, medical diagnosis, fire detector etc.
- We define the positive (fire) and negative (no fire) classes.
- **Confusion matrix.** Results of fire-prediction model summarized in a 2x2 confusion matrix:

True Positive Truth: a flame rose up The alarm rang Result: people have been saved	**False Positive** Truth: no fire The alarm rang Result: people made complaints
False Negative Truth: a flame rose up The alarm did not ring Result: people died in the fire	**True Negative** Truth: no fire. The alarm did not ring. Result: everyone is safe

- **True positive/negative.** The number of positive/negative samples which the model correctly predicts.
- **False positive/negative.** The number of positive/negative samples which the model incorrectly predicts.

Performance metrics

- **Precision.** it attempts to answer the following question: "what fraction of positive identification was correct?"

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

- Precision for fire-prediction model:

$$Precision = \frac{TP}{TP + FP} = \frac{1}{1 + 3} = 0.25 \tag{2}$$

True Positive=1 Truth: a flame rose up The alarm rang Result: people have been saved	**False Positive=4** Truth: no fire The alarm rang Result: people made complaints
False Negative=5 Truth: a flame rose up The alarm did not ring Result: people died in the fire	**True Negative=50** Truth: no fire. The alarm did not ring. Result: everyone is safe

Performance metrics

- **Recall.** It attempts to answer the following question: "what proportion of actual positives was identified correctly?"

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- Recall for fire-prediction model:

$$Recall = \frac{TP}{TP + FN} = \frac{1}{1 + 5} = 0.16 \tag{4}$$

True Positive=1 Truth: a flame rose up The alarm rang Result: people have been saved	**False Positive=4** Truth: no fire The alarm rang Result: people made complaints
False Negative=5 Truth: a flame rose up The alarm did not ring Result: people died in the fire	**True Negative=50** Truth: no fire. The alarm did not ring. Result: everyone is safe

Performance metrics

- **F-measure, F1.** It is the harmonic mean between precision and recall.
  - especially useful in case classes are imbalanced as it provides a weighted average of precision and recall.

$$F_1 = \left(\frac{Recall^{-1} + Precision^{-1}}{2} \right)^{-1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

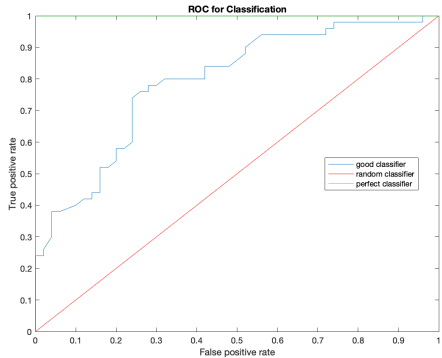
- F_1 for fire-prediction model:

$$F_1 = \frac{Precision \cdot Recall}{Precision + Recall} = \frac{0.25 \cdot 0.16}{0.25 + 0.16} = 0.0976 \tag{6}$$

True Positive=1 Truth: a flame rose up The alarm rang Result: people have been saved	**False Positive=4** Truth: no fire The alarm rang Result: people made complaints
False Negative=5 Truth: a flame rose up The alarm did not ring Result: people died in the fire	**True Negative=50** Truth: no fire. The alarm did not ring. Result: everyone is safe

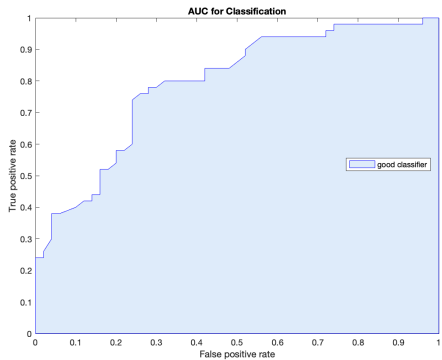
Performance metrics

- **Receiver Operating Characteristic (ROC).** A curve plotting the performance of a classification model over different classification threshold in terms of:
  - True Positive Rate $TPR = \frac{TP}{TP+FN}$
  - False Positive Rate $FPR = \frac{FP}{FP+TN}$



Performance metrics

- **Area Under the ROC Curve: (AUC).** Given a classifier C, AUC value can be interpreted as the probability a random positive sample x^+ is ranked by C higher than a random negative sample x^-, i.e. $auc(C) = P[C(x^+) > C(x^-)]$
  - $AUC \in [0, 1]$
  - $AUC = 1$ it is the perfect classifier where all positives come after all negatives
  - $AUC = 0.5$ is a random classifier



Performance metrics

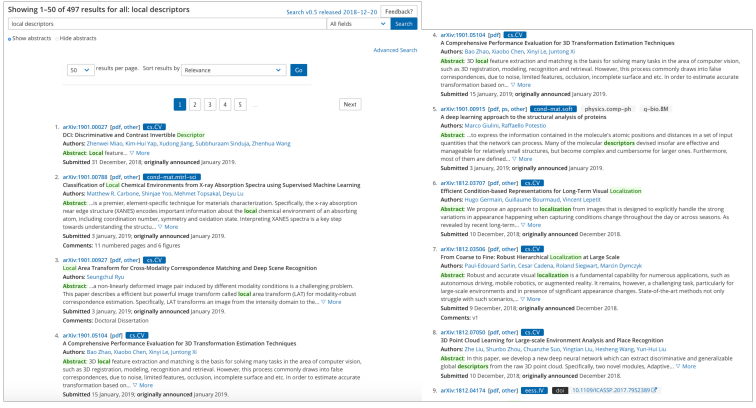
- **Recall and precision in retrieval task.**
  - Suppose you run a query in a retrieval system, e.g. arXiv.org, to search a document you need for your research. The output results can be summarized in the confusion matrix below.

True Positive Truth: the document is relevant The document is retrieved Result: you got the required document	**False Positive** Truth: the document is not relevant The document is retrieved Result: you got a useful document
False Negative Truth: the document is relevant The document is rejected. Result: you can't get the required document	**True Negative** Truth: the document is not relevant. The document is rejected. Result: you avoided useless documents

- Recall. The fraction of relevant documents that were retrieved. $Recall = \frac{RelevantRetrieved}{RelevantRetrieved + RelevantRejected}$
- Precision. The fraction of retrieved documents that is relevant. $Precision = \frac{RelevantRetrieved}{RelevantRetrieved + NotRelevantRetrieved}$

Performance metrics

- **Ranking in retrieval system.**
  - Suppose you are looking for "local descriptors in arXiv.org". Once the system processes your query, it will output an ordered list of results and you will start to examine it until you are satisfied or give up.



- P@K. Proportion of retrieved top-k papers which are relevant. $P@K = \frac{RelevantRetrievedTopK}{RelevantRetrievedTopK + NotRelevantRetrievedTopK}$
- R@K. Proportion of relevant papers which are retrieved in top-k. $R@K = \frac{RelevantRetrievedTopK}{RelevantRetrievedTopK + RelevantRejected}$

Performance metrics

- **Average precision.**
  - Given a query, analyze the ranking list at each relevant entry at a time, and measure the precision for the relevant entry $P@K = \frac{RelevantRetrievedTopK}{RelevantRetrievedTopK + NotRelevantRetrievedTopK}$.
  - Average precision is the average of all $P@K$ values.

Title	Relevant	R@k	P@k
Paper 1		0.10	1.00
Paper 2		0.10	0.50
Paper 3		0.20	0.67
Paper 4		0.30	0.76
Paper 5		0.40	0.80
Paper 6		0.50	0.83
Paper 7		0.60	0.86
Paper 8		0.60	0.75
Paper 9		0.70	0.78
Paper 10		0.70	0.70
Paper 11		0.80	0.73
Paper 12		0.80	0.67
Paper 13		0.80	0.62
Paper 14		0.90	0.64
Paper 15		0.90	0.60
Paper 16		0.90	0.56
Paper 17		0.90	0.53
Paper 18		0.90	0.50
Paper 19		0.90	0.47
Paper 20		1.00	0.50
		average-precision	**0.76**

- R@K. Proportion of retrieved top-k papers which are relevant. $R@K = \frac{RelevantRetrievedTopK}{TotalRelevant}$

Performance metrics

- **Mean average precision.**
  - Given a set of queries, mean average precision is the average of all average precision values.

Title	Relevant	R@k	P@k	Title	Relevant	R@k	P@k	Title	Relevant	R@k	P@k
Paper 1		0.10	1.00	Paper 1		0.10	1.00	Paper 1		0.00	0.00
Paper 2		0.10	0.50	Paper 2		0.20	1.00	Paper 2		0.00	0.00
Paper 3		0.20	0.67	Paper 3		0.30	1.00	Paper 3		0.00	0.00
Paper 4		0.30	0.76	Paper 4		0.40	1.00	Paper 4		0.00	0.00
Paper 5		0.40	0.80	Paper 5		0.50	1.00	Paper 5		0.00	0.00
Paper 6		0.50	0.83	Paper 6		0.60	1.00	Paper 6		0.00	0.00
Paper 7		0.60	0.86	Paper 7		0.70	1.00	Paper 7		0.00	0.00
Paper 8		0.60	0.75	Paper 8		0.80	1.00	Paper 8		0.00	0.00
Paper 9		0.70	0.78	Paper 9		0.90	1.00	Paper 9		0.00	0.00
Paper 10		0.70	0.70	Paper 10		1.00	1.00	Paper 10		0.00	0.00
Paper 11		0.80	0.73	Paper 11		1.00	0.91	Paper 11		0.10	0.09
Paper 12		0.80	0.67	Paper 12		1.00	0.83	Paper 12		0.20	0.17
Paper 13		0.80	0.62	Paper 13		1.00	0.77	Paper 13		0.30	0.23
Paper 14		0.90	0.64	Paper 14		1.00	0.71	Paper 14		0.40	0.29
Paper 15		0.90	0.60	Paper 15		1.00	0.67	Paper 15		0.50	0.33
Paper 16		0.90	0.56	Paper 16		1.00	0.63	Paper 16		0.60	0.38
Paper 17		0.90	0.53	Paper 17		1.00	0.59	Paper 17		0.70	0.41
Paper 18		0.90	0.50	Paper 18		1.00	0.56	Paper 18		0.80	0.44
Paper 19		0.90	0.47	Paper 19		1.00	0.53	Paper 19		0.90	0.47
Paper 20		1.00	0.50	Paper 20		1.00	0.50	Paper 20		1.00	0.50
		average-precision	0.76			average-precision	1.00			average-precision	0.33
						mean-average-precision	**0.69**				

Performance metrics in Natural Language Processing

- **Evaluation metrics for text generation**, e.g, translation from one language to another language, speech synthesis:
  - BLEU (Bilingual Evaluation Understudy Score)
  - METEOR (Metric for Evaluation of Translation with Explicit Ordering)
  - WER (Word Error Rate)
- **Evaluation metrics for text summarizing**, e.g, automatic paper summary:
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
- **Evaluation metrics for image description**, e.g, describing one image with a sentence:
  - CIDEr (Consensus-based Image Description Evaluation)

Performance metrics in NLP

- **BLEU (Bilingual Evaluation Understudy Score)** the main idea is "the closer a machine translation is to a professional human translation, the better it is" *1
  - It requires two inputs: 1) a machine translation; 2) a set of human translations.
    - ★ Candidate: the the the the the the the.
    - ★ Reference 1. The cat is on the mat.
    - ★ Reference 2. There is a cat on the mat.
  - **n-gram precision** $P = m/w_t$, where m is number of words (n-gram) in candidate translation which occur in any reference translations and w_t is the total number of words (n-gram) in the candidate translation. Thus the precision of the candidate text is $P = 7/7 = 1$
  - **modified n-gram precision** $P = m_{max}/w_t$, where m_{max} is maximum total count of unigrams in any reference translation $P = 2/7$
  - **modified n-gram precision** on full blocks of text. $p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{Count(n-gram')}$

*1. Papineni, K., et al. BLEU: a method for automatic evaluation of machine translation (PDF) in ACL 2002:

Performance metrics

- **METEOR (Metric for Evaluation of Translation with Explicit Ordering)**[*2]
  - The metric is an extension of BLEU metric providing stemming and synonymy matching. It's based on the harmonic mean of n-gram precision and recall with recall weighted higher than precision.
  - **n-gram recall** $R = m/w_r$, where m is number of words (n-gram) in candidate translation which occur in any reference translations and w_r is the number of words (n-gram) in the reference translation.
  - **harmonic mean** $F_{mean} = \frac{10PR}{R+9P}$, precision and recall are combined using a type of harmonic mean
  - **chunk** is a set of unigrams that are adjacent in the hypothesis and in the reference. The longer the adjacent mappings between the candidate and the reference, the fewer chunks there are.
  - **penality** $0.5 \cdot \left(\frac{c}{u_m}\right)^3$, where c is the number of chunks, and u_m is the number of mapped n-grams.
  - **meteor** $M = F_{mean}(1 - P)$, the final meteor metric where the penalty has the effect of reducing the F_{mean} by up to 50%

[*2]. Banerjee, S. and Lavie, A.: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments in ACL 2005.

Performance metrics

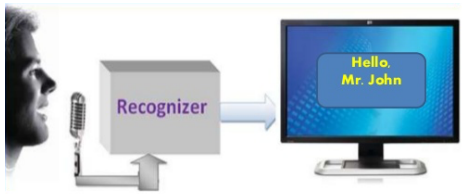
- **Evaluation metrics for text summarizing**,
- **Rouge (Recall-Oriented Understudy for Gisting Evaluation)**[*3]
  - The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.
- **Evaluation metrics for image description**,
- **CIDEr (Consensus-based Image Description Evaluation)**[*4]
  - The metrics evaluate for image how well a candidate caption matches the consensus of a set of image descriptions.

*3. Lin, Chin-Yew. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. In WAS 2004.

*4. Vedantam et al. CIDEr: Consensus-based image description evaluation. In: CVPR. (2015)

Performance metrics

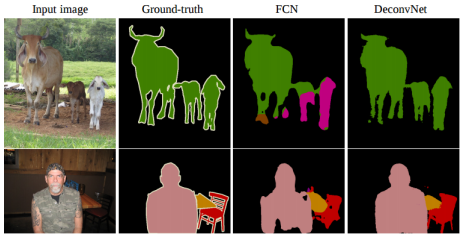
- **Evaluation measure for Speech recognition system.** WER (Word Error Rate)*6
  - $WER = (S + D + I)/N$, where S is number of substitutions, D number of deletions and I insertions, N number of words in the reference.
  - **Deletion** input: I read interesting papers → result: I read papers
  - **Insertion** input: I read papers → result: I read interesting papers
  - **Substitution** input: I read interesting papers → result: I read relevant papers



*6. Klakow Dietrich and Jochen Peters. Testing the correlation of word error rate and perplexity. In Speech Communication 2013.

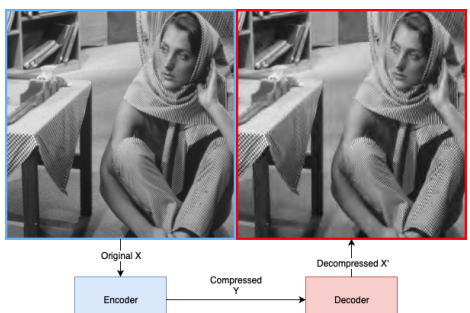
Performance metrics

- **Evaluation measure for image segmentation**. Given A the automatic segmented image and G the Grand Truth.
- **Dice coefficient.** It measures of the extent of spatial overlap between two binary images. Typically used for training.
  - $D = \frac{2(A \cap G)}{(A \cap G + A \cup G)}$
- **Jaccard index.** Also referred to as the intersection over union (IoU) metric. It measures the number of pixels common between the grand truth and prediction masks divided by the total number of pixels present across both masks.
  - $IoU = \frac{G \cap A}{G \cup A}$



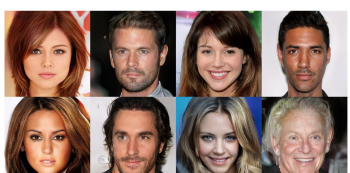
Performance metrics

- **Evaluation measure for signal quality**. Suppose you have an original signal X and reconstructed one X' with size e.g. for image $M \times N$
  - $MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} ||X(i,j) - X'(i,j)||$
  - $PNR = 20 * log_{10} \frac{MAX(X)}{\sqrt{MSE}}$
  - $SNR = 20 * log_{10} \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} X(i,j)^2}{\sqrt{MSE}}$



Performance metrics

- **Evaluation measure for quality of images**.
- **Inception score.**[*5]. it encodes two desiderable properties:
  - the images generated should contain clear objects
  - it should output a high diversity of images from all the different classes in ImageNet.
- $ISG(G) = exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|k)||p(y)))$
- $x \sim p_g$ indicates that x is an image sampled from p_g, which is the distribution encoded by the generative model G.
- D_{KL} is the KL-divergence between the distributions p and q, where is the distribution of real images.
- $p(y|x)$ is the conditional class distribution indicating the probability that the Inception v3 Network assigns the class y to the sample x.
- $p(y) = \int_x p(y|x) * p_g(x)$ is the marginal class distribution.



[*5]. Salimans et al. Improved techniques for training GANs. In NIPS 2016.

Performance Metrics

- General
  - True positives, false positives, true negatives, false negatives
  - Percision/recall
  - F-measure, F1
  - ROC, AUC
  - Retrieval:Rank@K, Prec@K, mAP
- NLP: BLEU, ROUGE, METEOR, CIDEr
- Speech: Word Error Rate (WER)
- Signal quality: MSE, SNR, PSNR
- Image quality: Inception score
- Segmentation: Dice coefficient, Jaccard index