Evaluation:

The One In Which We Evaluate

Lecture 8: Text Analytics for Big Data

Mark Keane, Insight/CSI, UCD

Selling Things

stock-markets

social media

science

news

polls

sentiment-id

sentiment-use

time-series

summaries

VSMs

Classifiers

Clustering

***

cosine

jaccard

dice

levenschtein

***

TF-IDF

LLR

PMI

Entropy

simple frequencies

pre-processed text items of some sort…

Why Evaluate?

- DOES MY SYSTEM WORK?
- Is it getting the right answers?
- Is it finding all the topics in this corpus?
- Is this a good summary of these tweets?
- Is this word a good/bad sentiment to us?
- Are my outputs the right answers?

Outline

- The Four Basic Ideas: Ground Truth, Precision, Recall, and F measures
- Other Things: ROC Curves, DET Curves, Q-Q Plots, ROUGE, BLEU…
- State of the Art & Criticisms

Evaluation

4 Basic Ideas

4 Basic Ideas

REM

- *Ground Truth (GT)*: Ask people for the right answer(s), which items are correct / relevant

- *Precision*: Fraction of output items that are correct versus incorrect, using GT

- *Recall*: Fraction of output items correct in GT

- *F-measure*: One no for precision & recall…

Evaluation

Ground Truth

4 Basic Ideas

Canfield Method V1.0

- Developed to test indexing schemes and book search in physical libraries
  - *Task*: define *user information needs* (user model)
  - *Test Set*: a sample collection of documents
  - *Judgement Methodology*: made by experts
- So, you took a set of books, and say we want to find french dictionary (define queries), find books and get experts to judge are these the ones wanted

Canfield Method V2.0

- IT systems explode in IR, topic detection, story linking, summarisation systems…
- Every aspect of the method gets stretched:
  - *Task*: now have multiple tasks, topic detection, FSD, summarisation, story linking, ad hoc retrieval
  - *Test Set*: collections explode into millions, issues of sampling for test purposes arise
  - *Judgement Methodology*: becomes very hard for experts to look at the items to make judgements

Some Large Corpora

- Ground Truths have gone from being a very small to a really tiny proportion of full corpora

- People don't look at them any more

Corpus Name	Paper	Dates	Corpus Size	Ground Truth Items	Total Jugded Items	Ratio of GT items to Corpus
TDT1		1998	72,000	100		0.001389
TDT2			"	120		
TDT3			"	"		
TDT4			"	"		
TDT5 (English)		2000-2001	278,108	126	?9300?	0.000453
Tweet11		2011	16,000,000	49	50,324	0.000003
Twitter7-SNAP		2009	467,000,000	-	-	
NYC Corpus	Becker et al (2013)	2010 Feb1-Feb28	2,600,000	675 clusters	-	0.000260
Edinburgh-I	Petrovic et al (2010)	2009 April1-Oct17	163,500,000	820 seed tweets	1000	0.000005
Edinburgh-II	Petorska et al (2012)	2011 July-Sept	50,000,000	27	3,035	0.000001
Edinburgh-III	Osborne et al. (2010)	2011 Jun20-Jul24	2M per day	235 tweets	235	
Glasgow	McMinn et al (2013)	2012 Oct10-Nov7	120,000,000	500*	50,000	0.000004
InsightCrawl	Greene (2013)	2013 Oct1-Dec5	29,000,000	50		0.000002
TREC2013	Lin & Efron (2013)	2013	240,000,000	-	-	

TDT & TRECs

- Varied systems result in TREC conferences designed to find community solution
- Large scale definition of test sets, shared GT for 100s of items; yearly competitions
- Drives field but ultimately becomes too much
- Two main innovations: pooling, crowdsourcing

Pooling

- Stop looking for Ground Truth !
- Set up task and test set; run multiple algorithms on the test set, noting outputs (items set X)
- Find outputs common to all algorithms (x in X)
- Algorithm that finds most of x is the best
- NB: really an internal validation not external

Crowdsourcing

- Get Ground Truth from Mechanical Turk
- Set up micro-tasks to get online judgements
- Need to carefully define judgement task and to find methods for defining judgements
- Issues: demographics, ethical, cultural
- One would worry about its utility

GroundTruth: External

- Finding an external source as a ground truth
  Open Calais, Previous categorisation,
  Curated source (Wikipedia)

- Compare system outputs to what was found in this external list/categorisation/etc…

Greene, D., O'Callaghan, D., & Cunningham, P. (2012). Identifying topical twitter communities via user list aggregation. arXiv preprint arXiv:1207.0017.

Egs: Traditional TDT TREC

- Topic Detecton & Tracking (TDT); Online **new** event detection and tracking
- Processing stories streaming from a newswire..new and continuations…
- Hugely formalised competition

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In 21st ACM SIGIR on Research and development in information retrieval (pp. 37-45). ACM.

Egs: TDT 1995: Not Pooling

- 15,863 Reuters and CNN; mean 400-word length after stemming and stop-word removal

- Relevance judgements for 25 events (OK bombing, earthquakes); many classes of event

- one-to-many, event-to-articles; every article judged relative to every event by two assessors (resolving third); ended with 1123 articles for evaluation

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In 21st ACM SIGIR on Research and development in information retrieval (pp. 37-45). ACM.

Results: Wha?

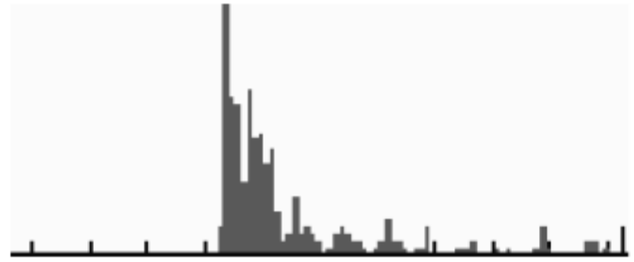


Figure 5: Number of news stories per day covering the OK city bombing event.

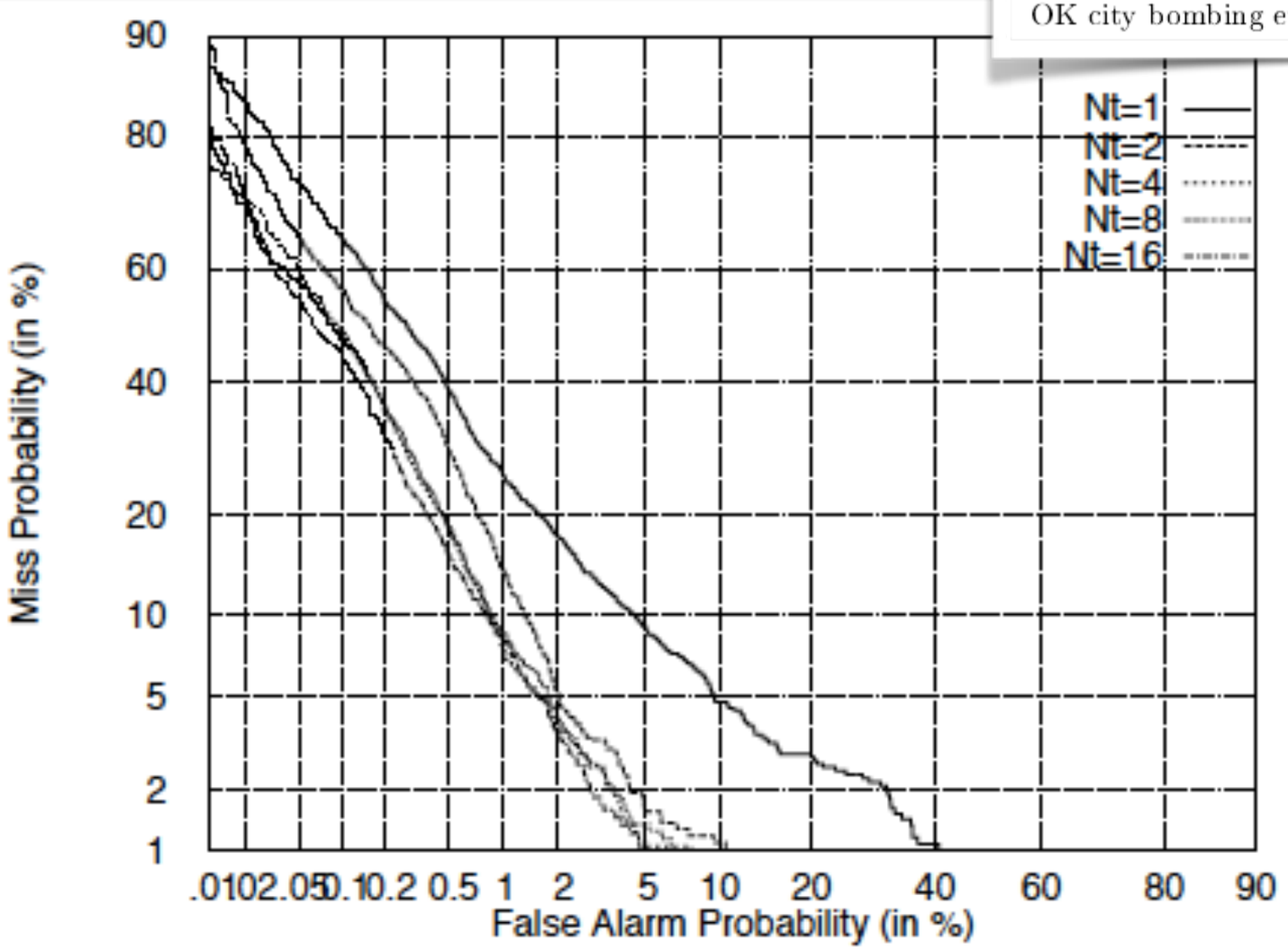


Figure 4: Comparing values of N_t. Once N_t reaches 4, adding more stories for training is only marginally helpful.

Allan, J., Papka, R., & Lavrenko, V. (1998). On-line new event detection and tracking. In 21st ACM SIGIR on Research and development in information retrieval (pp. 37-45). ACM.

Egs: SNOW 2014: Pooling

- SNOW Data Challenge; detection of news stories in twitter @ WWW Conference 2014

- Provided training sets of tweets from newshound's accounts and asked for selected tweets indicating stories at different agreed time slots

- Three assessors look at pooled outputs from systems and judged them on 4 dimensions for goodness

Papadopoulos, S., Corney, D., & Aiello, L. M. (2014). SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. In SNOW-DC@ WWW (pp. 1-8).

Egs: SNOW 2014: Insight Win

Table 6: Overview of raw scores

Team	R_{ref}	P_{ext}	R_{ext}	F_{ext}	Q	C	D	I
UKON [Pop14]	0.44	0.481	0.186	0.268	4.29	4.40	2.12	0.542
IBCN [Can14]	0.58	0.522	0.171	0.258	4.92	4.08	2.36	0.318
ITI [Pet14]	0.32	0.440	0.214	0.288	4.49	4.68	2.31	0.581
math-dyn [Bur14]	0.63	0.462	0.200	0.279	4.59	4.91	2.11	0.520
Insight [Ifr14]	**0.66**	0.560	**0.357**	**0.436**	4.74	**4.97**	2.11	0.274
FUB-TORV [Ama14]	0.39	0.267	0.029	0.052	4.18	4.78	2.00	-
PILOTS [Nut14]	0.24	0.400	0.057	0.099	**4.93**	4.83	1.92	-
RGU [Mar14]	0.60	0.388	0.243	0.299	4.71	4.22	3.27	**0.588**
UoGMIR	0.17	**0.800**	0.214	0.338	4.80	3.95	2.36	-
EURECOM	0.24	0.125	0.014	0.027	3.38	3.75	2.50	-
SNOWBITS [Bha14]	0.14	**0.800**	0.100	0.178	4.32	4.36	**3.47**	0.186

Ifrim, G., Shi, B., & Brigadir, I. (2014). Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. In SNOW-DC@ WWW (pp. 33-40).

Papadopoulos, S., Corney, D., & Aiello, L. M. (2014). SNOW 2014 Data Challenge: Assessing the Performance of News Topic Detection Methods in Social Media. In SNOW-DC@ WWW (pp. 1-8).

Ground Truth / Reliability

- note, there are number of standard methods determining inter-rater agreement: Cohen's Kappa

- also method for collating crowdsourced replies

Evaluation

Precision, Recall,F-Measure

4 Basic Ideas

Measures

REM

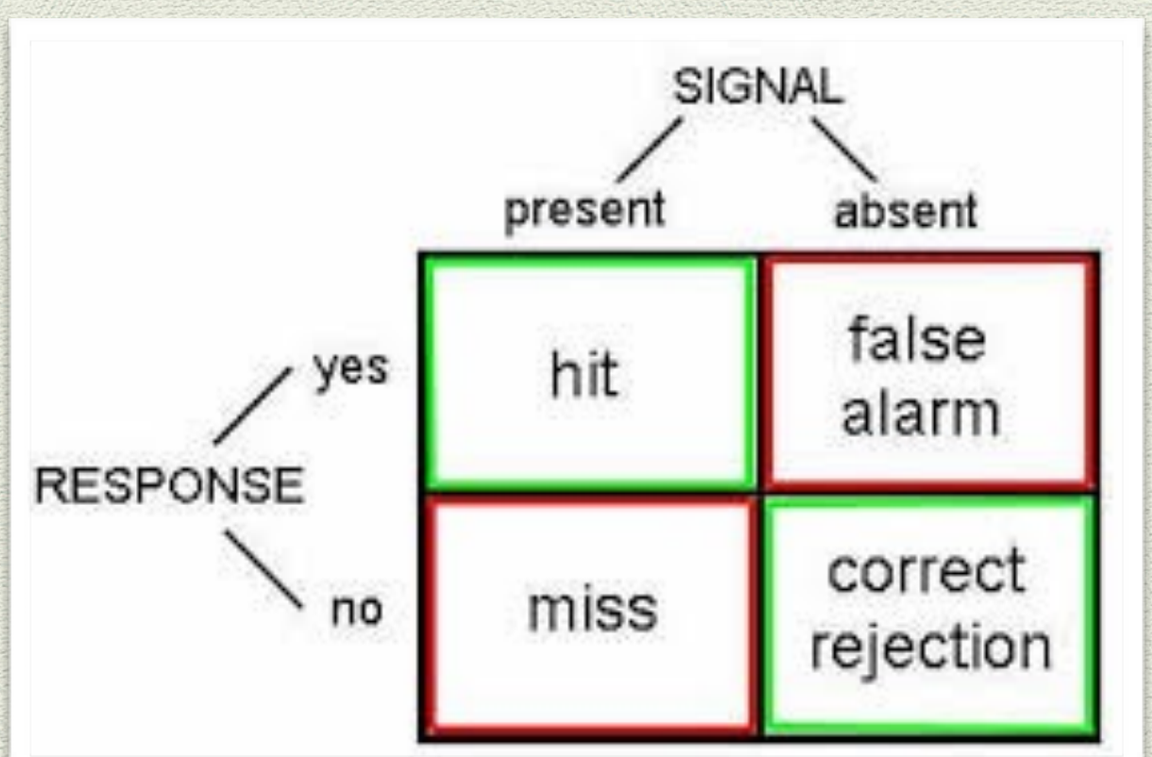
- *Ground Truth (GT)*: Ask people for the right answer(s), which items are correct / relevant

- *Precision*: Fraction of output items that are correct versus incorrect, using GT

- *Recall*: Fraction of output items correct in GT

- *F-measure*: One no for precision & recall…

Hits-Misses Matrix



4 possible outcomes
for matching an output item
and a GT item

TP-FP-FN-TN

Precise & Recall summarise
aspects of these tables

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**

Precision

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**

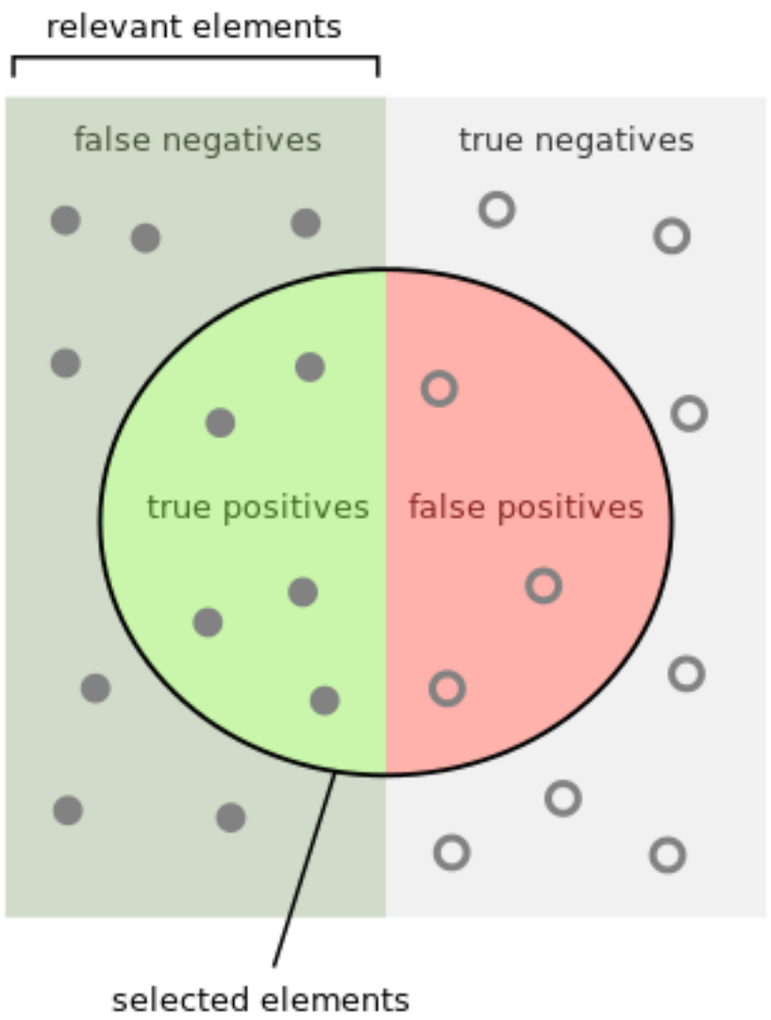
- *Precision*: Fraction of output items that are correct versus incorrect, using GT

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

precision or **positive predictive value** (PPV)

$$PPV = TP/(TP + FP)$$

TP (true positive)
FP (false positive)



Precision

- System to spot Tweets about MK; I have a set of 100 tweets I have checked (60 about me, 40 not); given this set, my system ids 30 tweets as being about me; 20 are True Positives, 10 False Positives; Precision is 20/(20+10) = 0.66
- NB: P@20 or P@10; Precision for the first 20 items or 10 items using a cut-off on the set; need to rank outputs; may produce different Precision score (Google Does P@10)
- Can also be read as the probability that a (randomly selected) retrieved document is relevant

Recall

- *Recall*: Fraction output items correct in GT

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

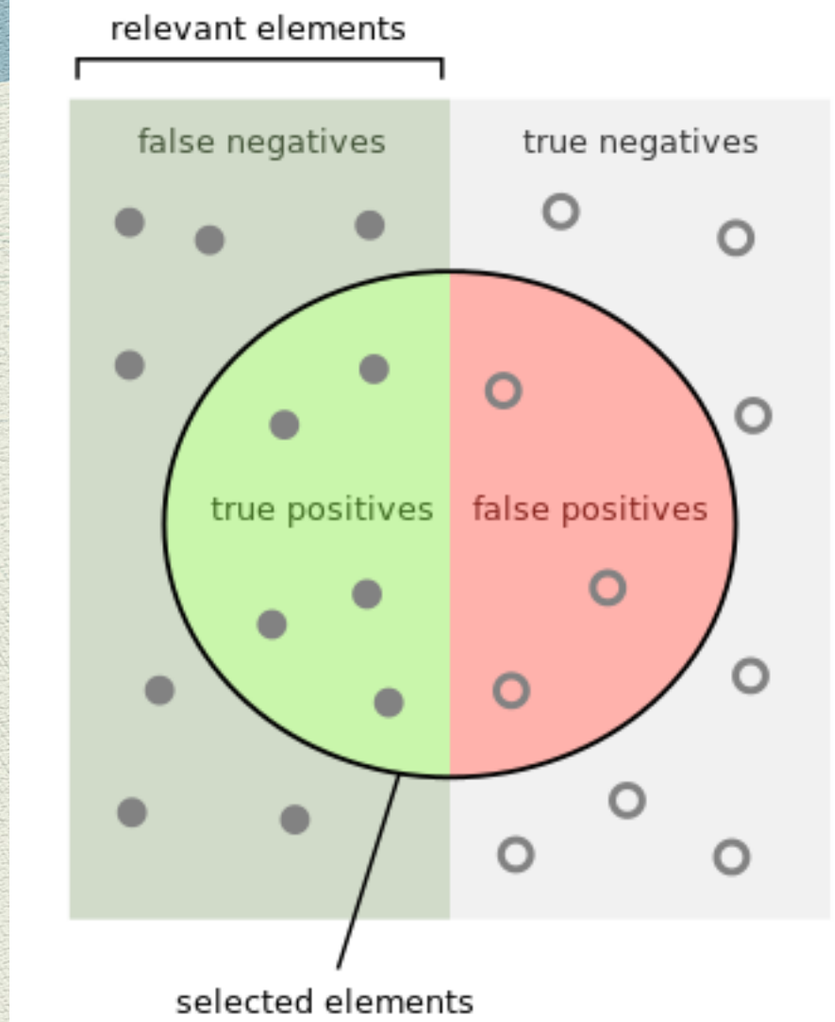
sensitivity or true positive rate (TPR)

eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

TP (true positive)
FN (false negative)
P (all positive)

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**



How many selected items are relevant?

Precision =

How many relevant items are selected?

Recall =

Recall/Sensitivity/TPR

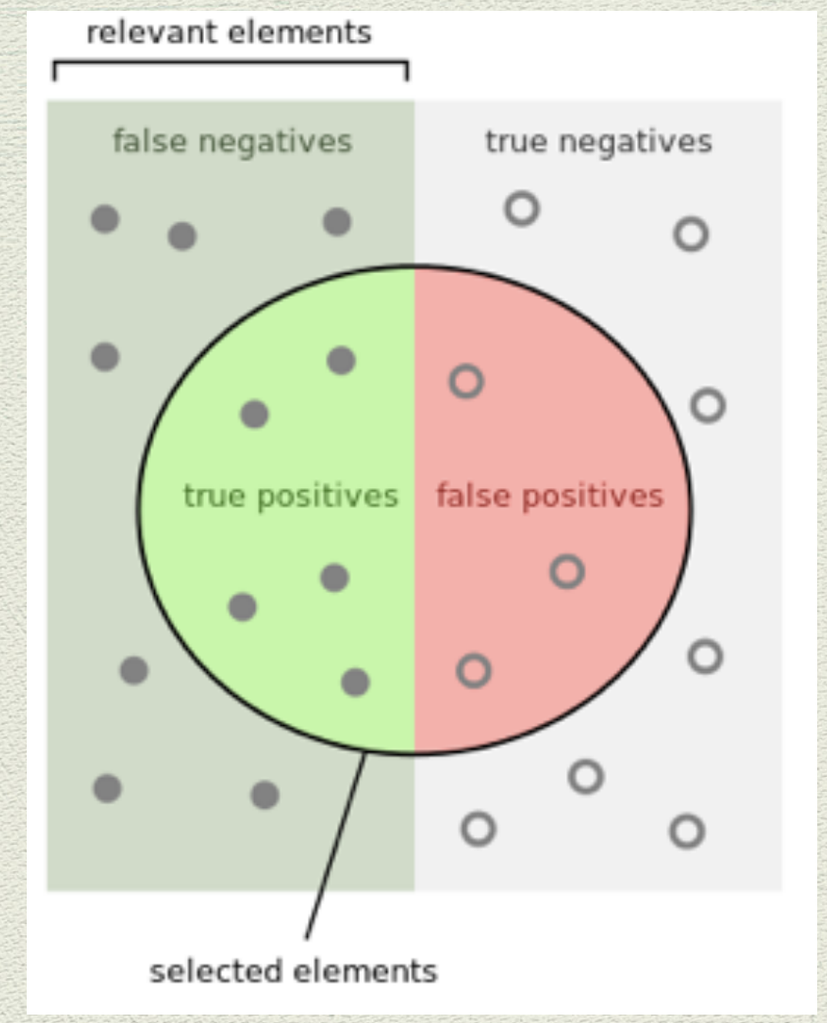
- System to spot Tweets about MK; I have a set of 100 tweets I have checked (60 about me, 40 not); given this set, my system ids 20 tweets are True Positives, 45 False Negatives; Precision is 20/(20+45) = 0.31
- NB: R@20 or P@10; Precision for the first 20 items or 10 items using a cut-off on the set; need to rank outputs; may produce different Recall score
- Can also be read as the probability that a (randomly selected) relevant document is retrieved in a search.

Accuracy

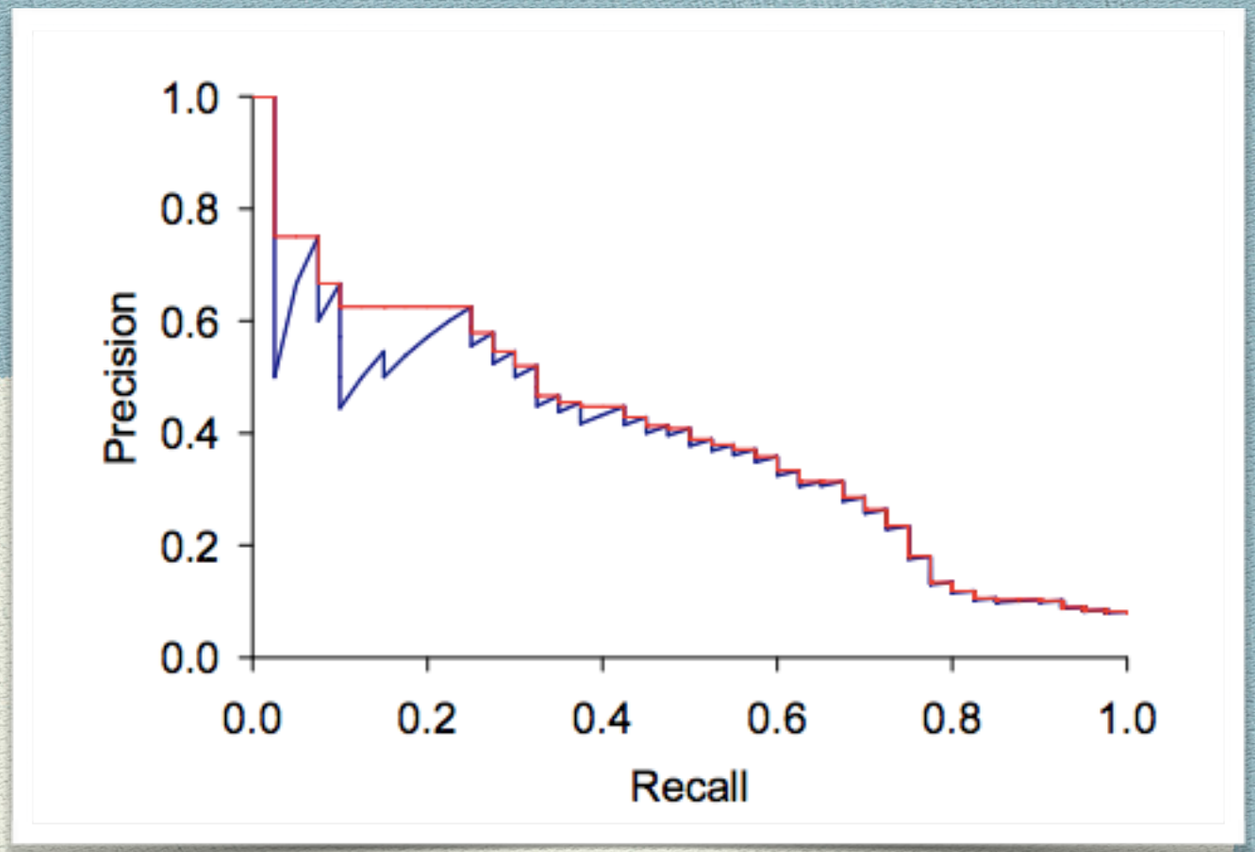
- Sort of combines all parts of the hits-misses matrix:

(TP + TN)/(TP+FP+FN+TN)

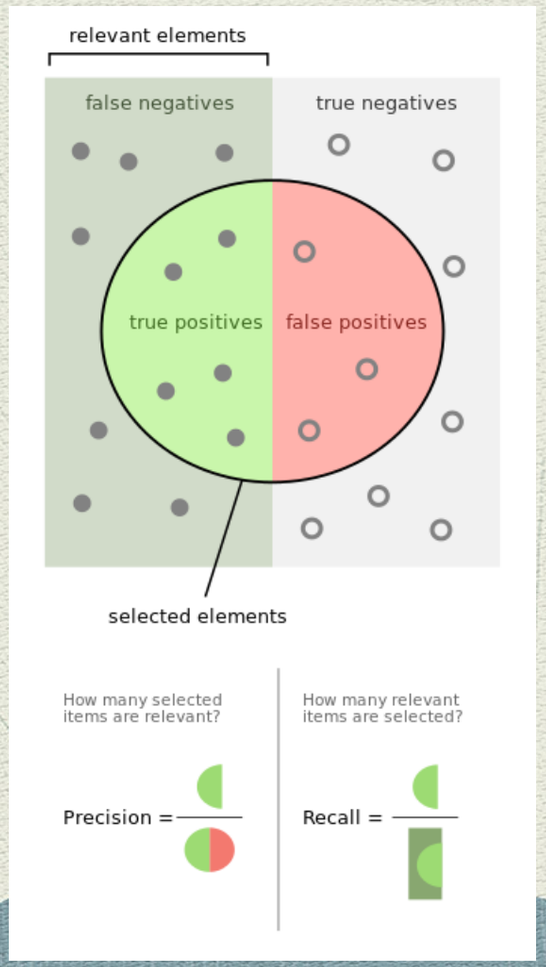
- Not greatly used, because of dominance of TNs



Precision V Recall



- Precision and Recall trade off one another; and users may want either (Web v Library)
- Examining P@20, P@25, P@30 and R@20, R@25 and R@30 you can plot tradeoff



F-measure

- *F-measure*: is the *harmonic mean* of precision and recall

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

- F_1 gives equal weighting to both; F_2 sometimes used to weight recall more, $F_{0.5}$ sometimes used to weight precision more than recall

It is a special case of the general F_β measure (for non-negative real values of β):

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

TP (true positive)
FP (false positive)
FN (false negative)

4 Basic Ideas

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**

- These are four basic ideas you will usually see mentioned in any reasonable evaluate of the text analytics system
- As we shall see, they can be elaborated in more complex ways; but you need to know them off-by-heart
- See Also:

Precision and recall are then defined as:[5]

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy.[5] True negative rate is also called specificity.

$$\text{True negative rate} = \frac{tn}{tn + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Mean Average Precision

- Just as F and F1 are used to summarise what is happening in Precision and Recall, MAP is also used as a summary score for Precision-Recall Tradeoffs (area under interpolated curve)

- Mean Average Precision (MAP): you average the precisions scores for each query; checking precision after each retrieval step; tell you how precision is unfolding (what is happening within P@10)

Mean average precision [edit]

Mean average precision for a set of queries is the mean of the average precision scores for each query.

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP(q)}}{Q}$$

where Q is the number of queries.

EG Precision & Recall

- How to find news stories occurring at different time periods
- Makes use LLR-type measure (Chi-2) to detect words, entities, phrases that stick out from the norm
- Build tracking system from TDT corpora and then tried to see if they corresponded to main news events of the year

Swan, R., & Jensen, D. (2000, August). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining* (pp. 73-80).

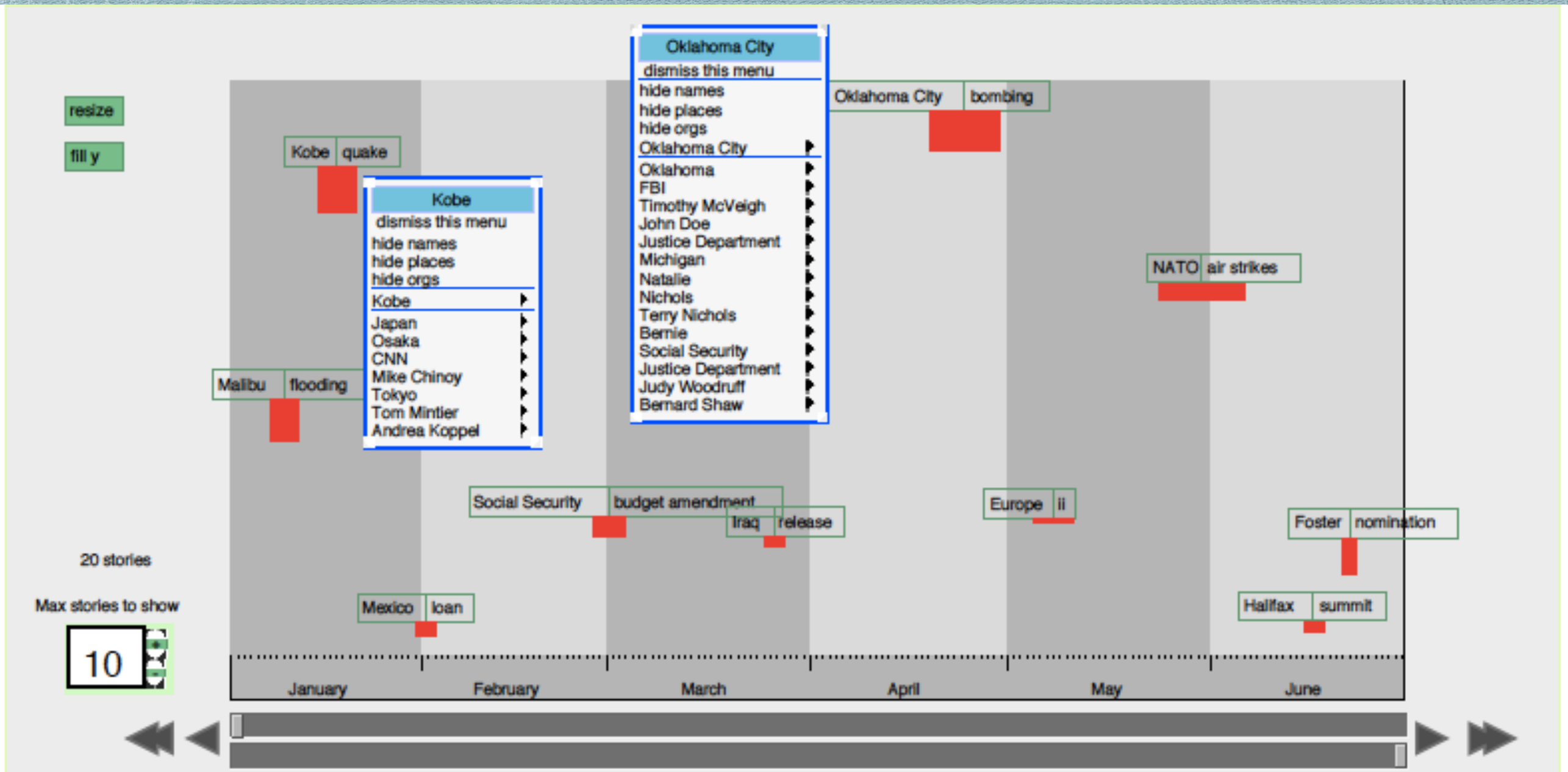


Figure 1: Overview of January - June, 1995. The topic labeled *Oklahoma City bombing* is the highest ranked topic, and the topic labeled *Kobe quake* is the second highest ranked. The pop-up on *Oklahoma City* shows significant named entities of *Oklahoma*, *FBI*, *Timothy McVeigh*, *John Doe*, *Justice Department*, etc. The other pop-up shows the terms associated with the Kobe earthquake.

Swan, R., & Jensen, D. (2000, August). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining* (pp. 73-80).

EG1 Precision & Recall

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
Earthquake in Kobe, Japan	Jan 16 - Jan 20
F-16 shot down over Bosnia	June 2 - June 5
NATO forces in Bosnia	May 25 - May 27
Flooding in California	Jan 10 - Jan 11
NATO forces in Bosnia	May 29 - May 31
Senate debates Balanced Budget	Feb 28 - Mar 2
Russia/US Summit	May 6 - May 10
Two Americans Sentenced in Iraq	Mar 25 - Mar 27
Henry Foster rejected by Senate as Surgeon General	June 21 - June 22

Table 3: Top 10 stories as calculated by named entity statistics (labels manually assigned)

Swan, R., & Jensen, D. (2000, August). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining* (pp. 73-80).

EG1 Precision & Recall

3.4 Topic Based Evaluations

Our second hypothesis states "These stories will prove comprehensible and useful to human users." Our experience confirms this, but we have not yet devised a formal evaluation. We have attempted two different evaluations but both had problems. Our first attempt was an IR-style evaluation, where we had a "truth" set that we compared our results against to measure precision and recall scores. To get a "truth" set we used the 1995 Year-In-Review section of the January 1996 *Facts on File*[1]. There were 24 stories during the time period of our TDT-1 corpus that *Facts on File* identified as major stories, and our system found 28 stories. The overlap was only seven stories, giving recall of 29% and precision of 25%. Further analysis showed that many of the stories chosen by Facts on File were either not mentioned in our corpus or were only briefly mentioned, and many of the stories chosen by TimeMines and not in Facts on File were arguably as important as the chosen stories.

Swan, R., & Jensen, D. (2000, August). Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining* (pp. 73-80).

Evaluation

Graphing Stuff

Why Bother

- Systems are so complex you need ways of looking at their behaviour
- Note, eg, a classifier will usually give a score for the likehood/goodness/strength of the classification (iris-A 0.46); and a threshold will make this a 1/0 classification, isa-iris-A or not
- So, you need to see how outputs change with these threshold changes; graphing the parameter space of the system
- You especially want to see trade-offs; does increasing the threshold increase False Positives with True Positives?

3 Common Graphing Methods…

- *ROC Curves*:
  - Receiver Operator Characteristic curves
- DET curves:
  - Detection Error Tradeoff Curves
- *Q-Q plots*:
  - Quantile-Quantile Plots

true positive rate = recall = sensitivity TP/(TP + FN)
false positive rate = fallout = FP/(FP+ TN)

ROC

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**

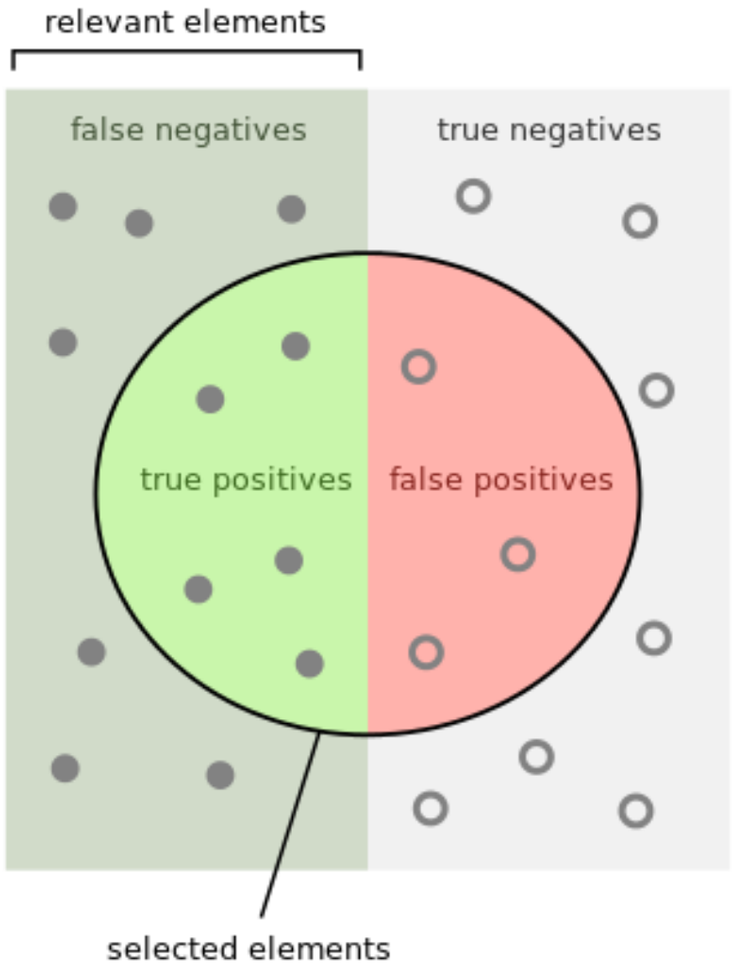
$$\text{Recall} = \frac{TP}{TP+FN}$$

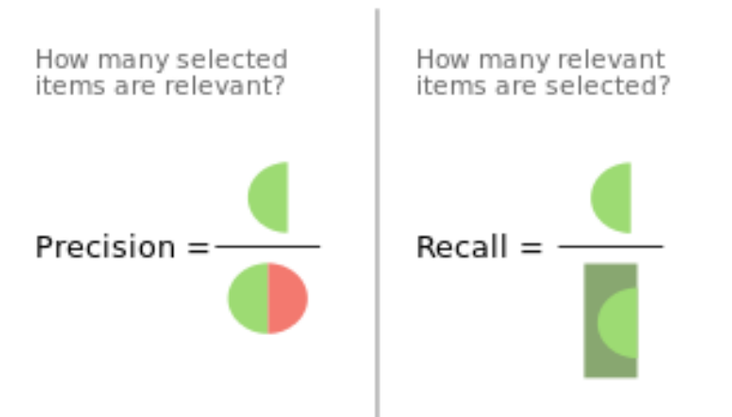
$$\text{Precision} = \frac{TP}{TP+FP}$$

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

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

- Receiver operator characteristic (ROC curve)
- Plot of true-positive-rate by false-positive-rate
- What is the cost in the false positives rate as the true positives rate increases?





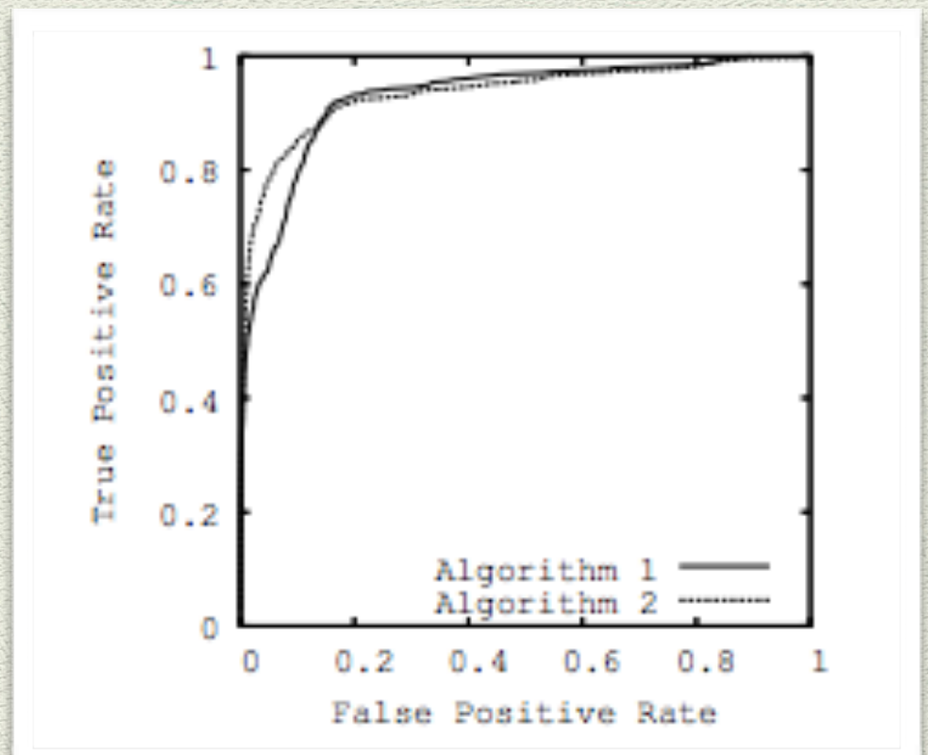
ROC

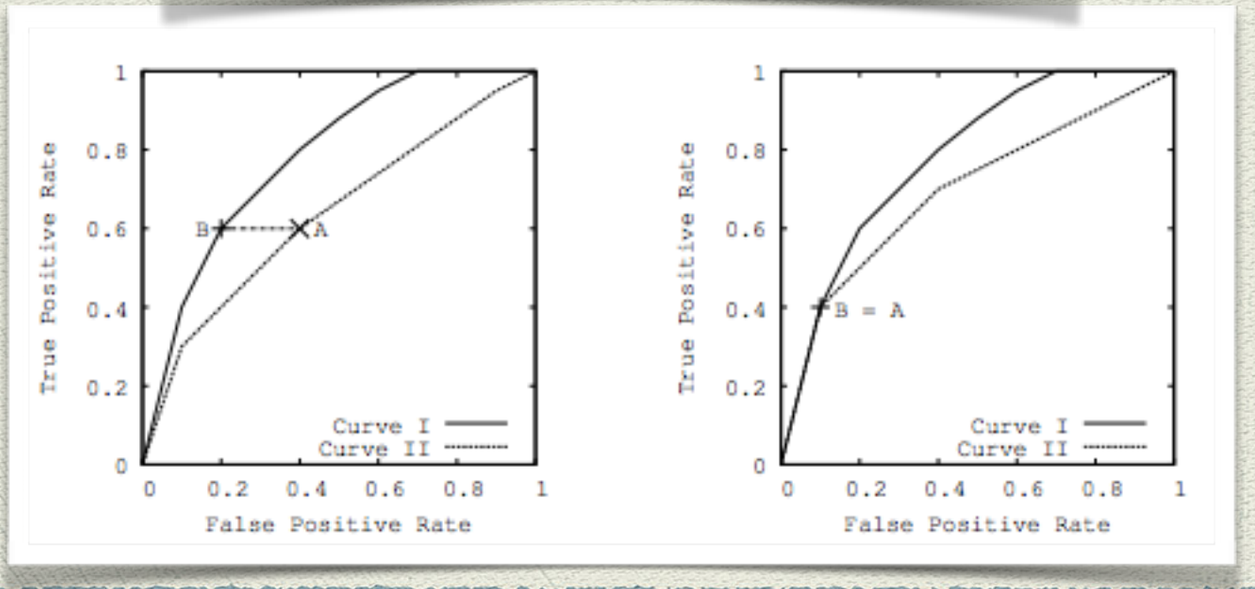
true positive rate = recall = TP/(TP + FN)

false positive rate = FP/(FP + TN)

Recall	=	$\frac{TP}{TP+FN}$
Precision	=	$\frac{TP}{TP+FP}$
True Positive Rate	=	$\frac{TP}{TP+FN}$
False Positive Rate	=	$\frac{FP}{FP+TN}$

- TPR reflects the ratio of correct found in all correct
- FPR reflects ratio of incorrect found in all incorrect



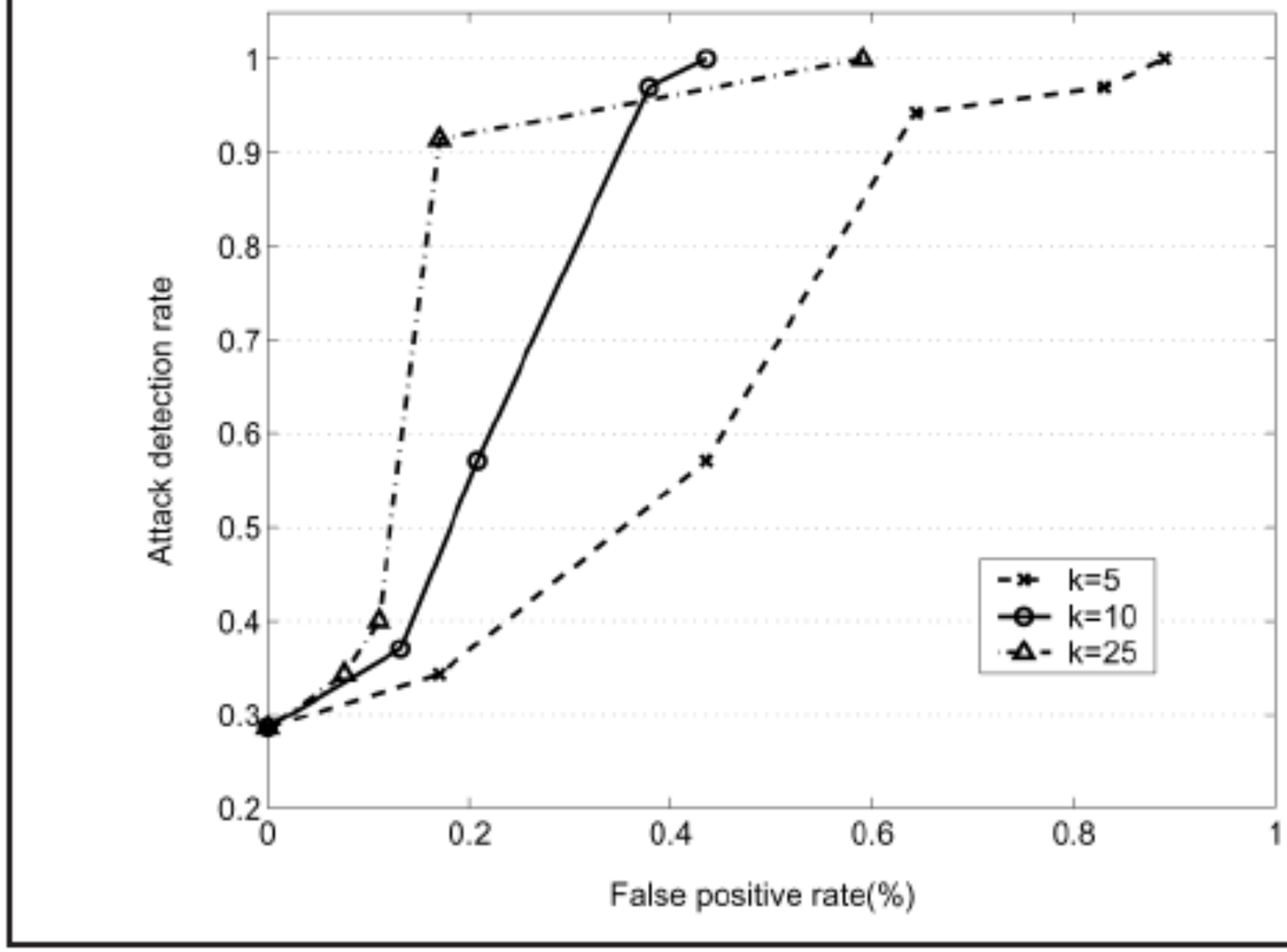


Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. ICML-06 (pp. 233-240). ACM.

REM

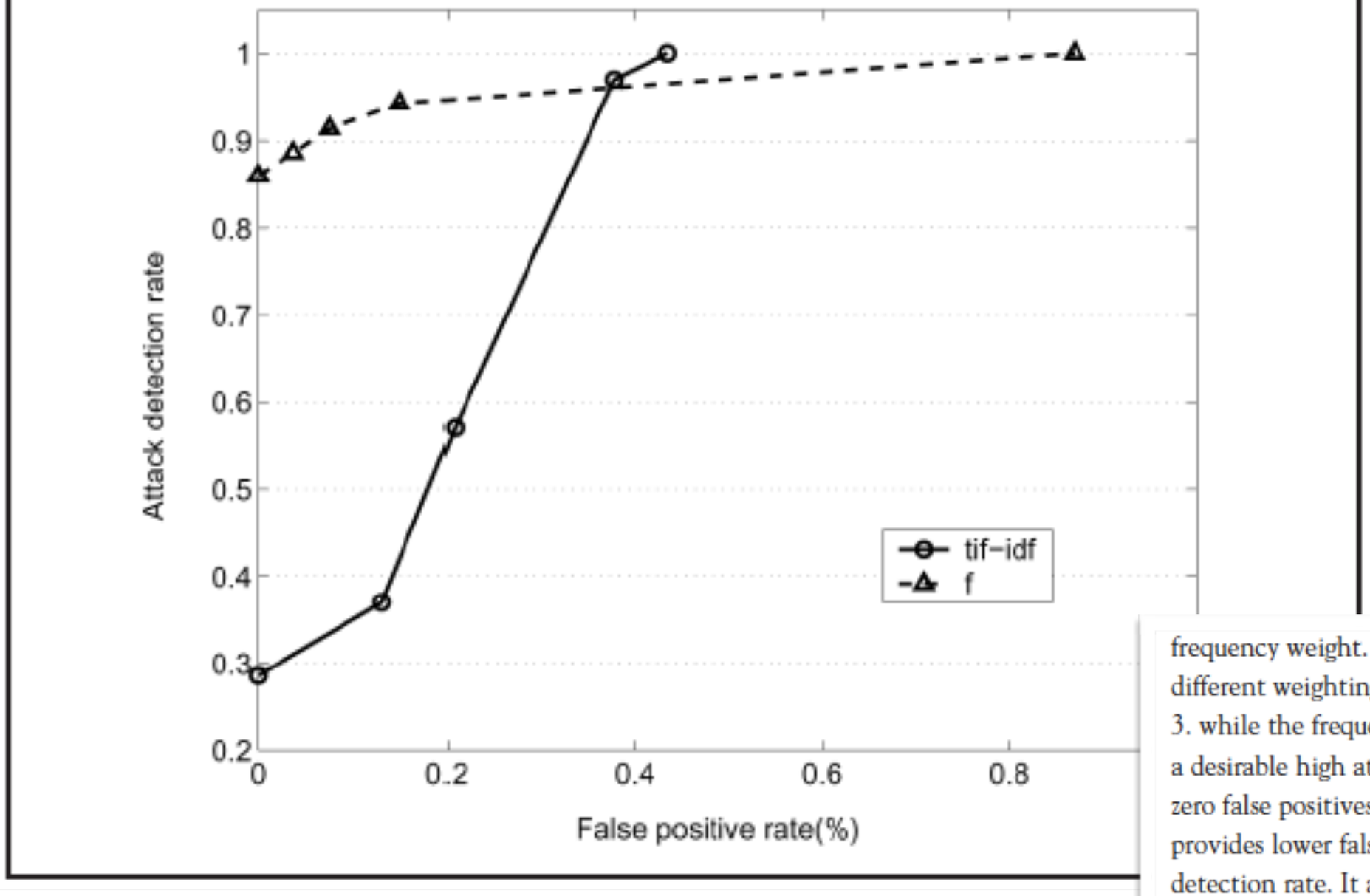
k-NN #2: Intruder Detection

Figure 2: Performance of the kNN classifier method expressed in ROC curves for the tf•idf weighting method. False positive rate vs attack detection rate for k=5, 10 and 25.



Liao, Y., & Vemuri, V. R. (2002). Use of K-nearest neighbor classifier for intrusion detection. *Computers & Security, 21(5)*, 439-448.

Figure 3: ROC curves for tf•idf weighting (k=10) and frequency weighting (k=15).

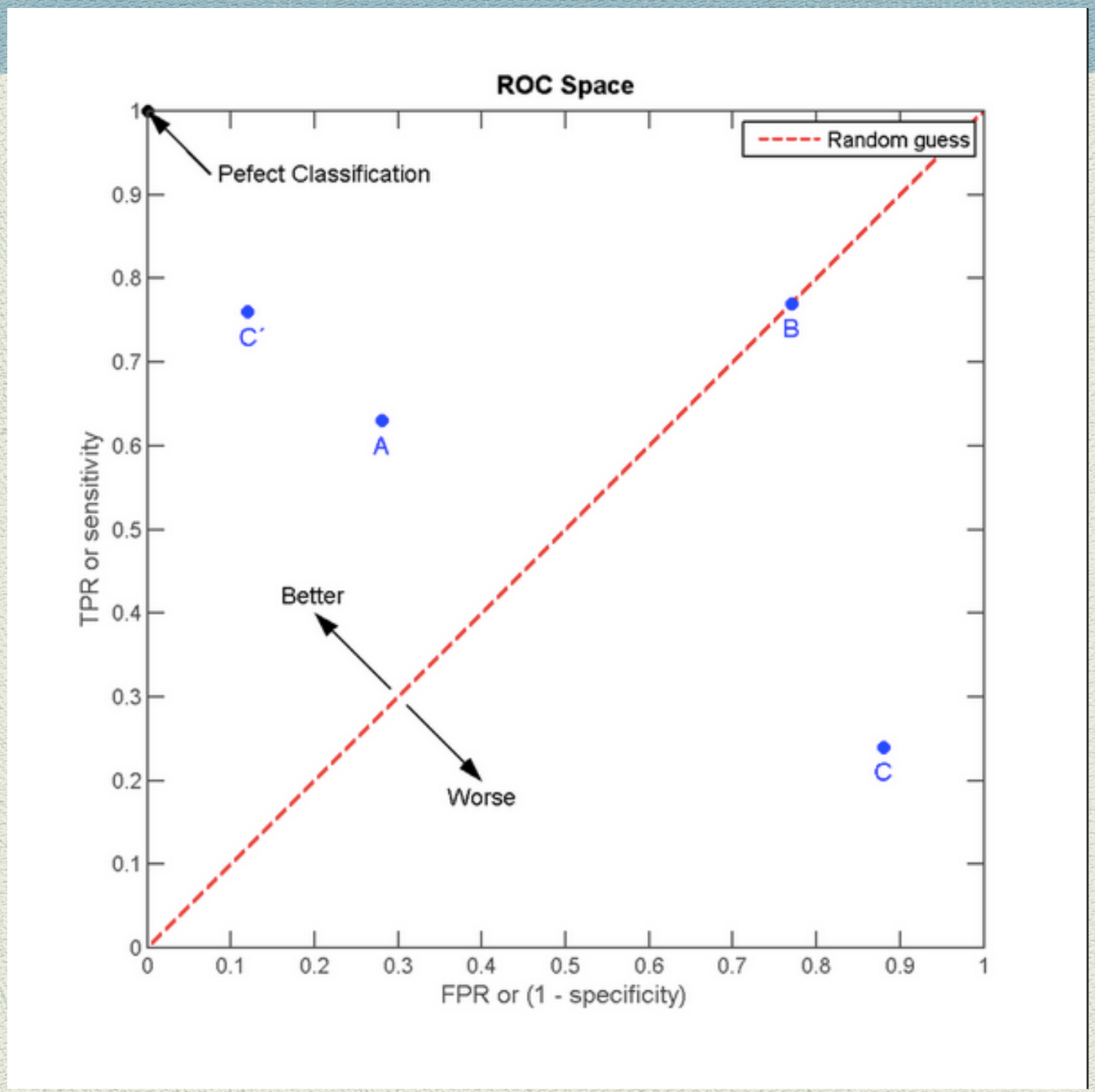


frequency weight. A comparison of two different weighting methods is shown in Figure 3. while the frequency weighting method offers a desirable high attack detection rate (86%) at zero false positives,the *tf•idf* weighting method provides lower false positive rate at 100% attack detection rate. It appears that the *tf•idf* weighting can make process vectors of two classes more distinguishable than the frequency weighting. Therefore, a lower threshold value is needed, and better false positive rate can be achieved with the *tf•idf* weighting method.

Liao, Y., & Vemuri, V. R. (2002). Use of K-nearest neighbor classifier for intrusion detection. *Computers & Security, 21(5)*, 439-448.

REM

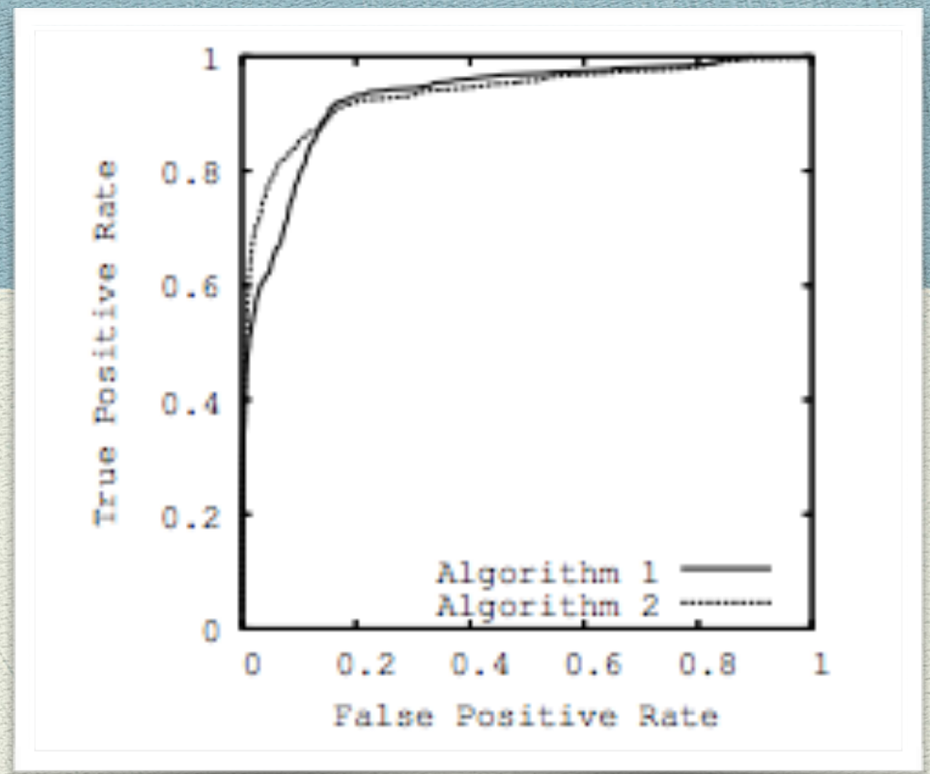
Classifiers: Some Interpretations

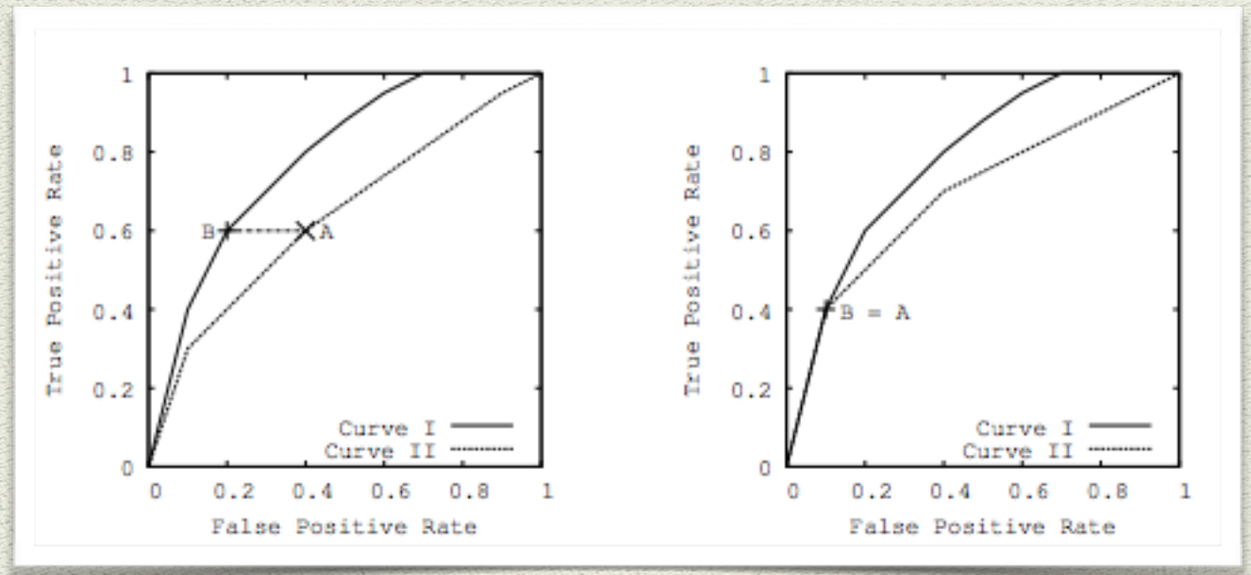


https://en.wikipedia.org/wiki/Receiver_operating_characteristic

AUC, not AUK

- Area Under Curve of an ROC plot
- When using normalized units, the area under the curve (often referred to as simply the AUC, or AUROC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than ‘negative')





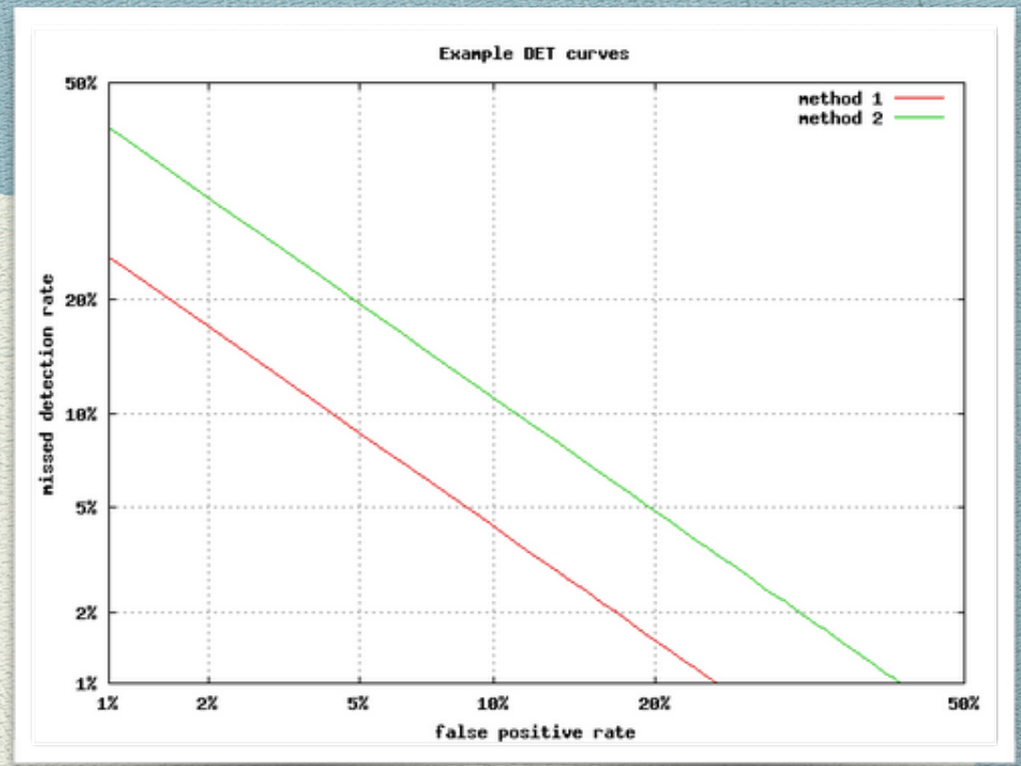
Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. ICML-06 (pp. 233-240). ACM.

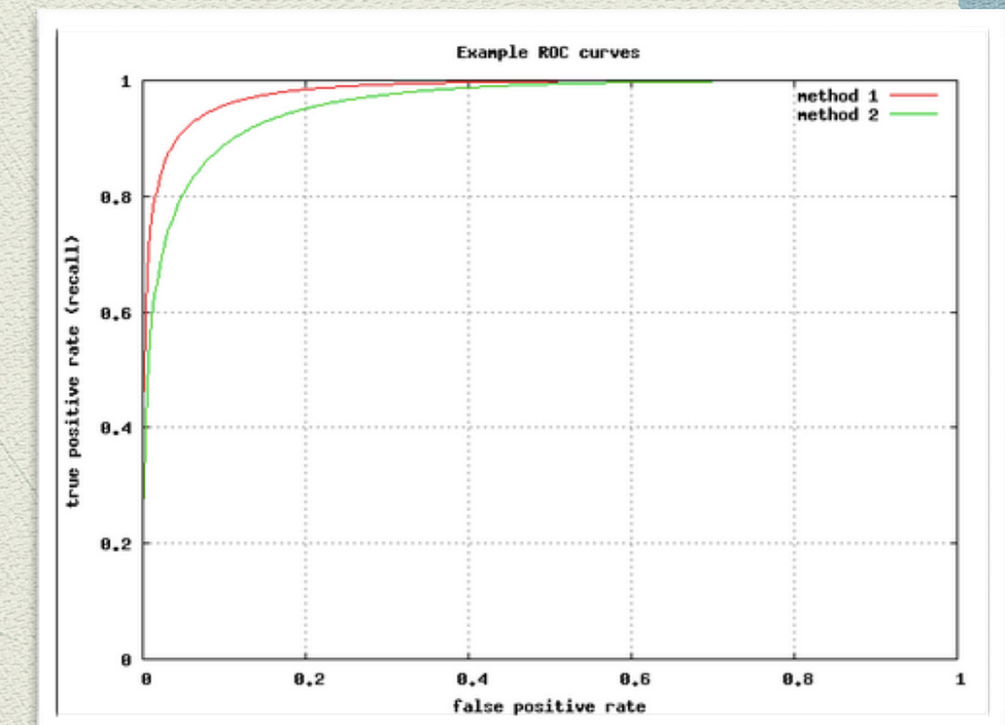
3 Common Graphing Methods…

- *ROC Curves*:
  - Receiver Operator Characteristic curves
- DET curves:
  - Detection Error Tradeoff Curves
- *Q-Q plots*:
  - Quantile-Quantile Plots

Detection Error Tradeoff (DET) Curves

- Another way plotting what system is doing
- DET curves focus on errors more and “zoom in” to key parts of ROC curve by using log axes
- Often, you will see ROC and DET curves produced for system comparisons



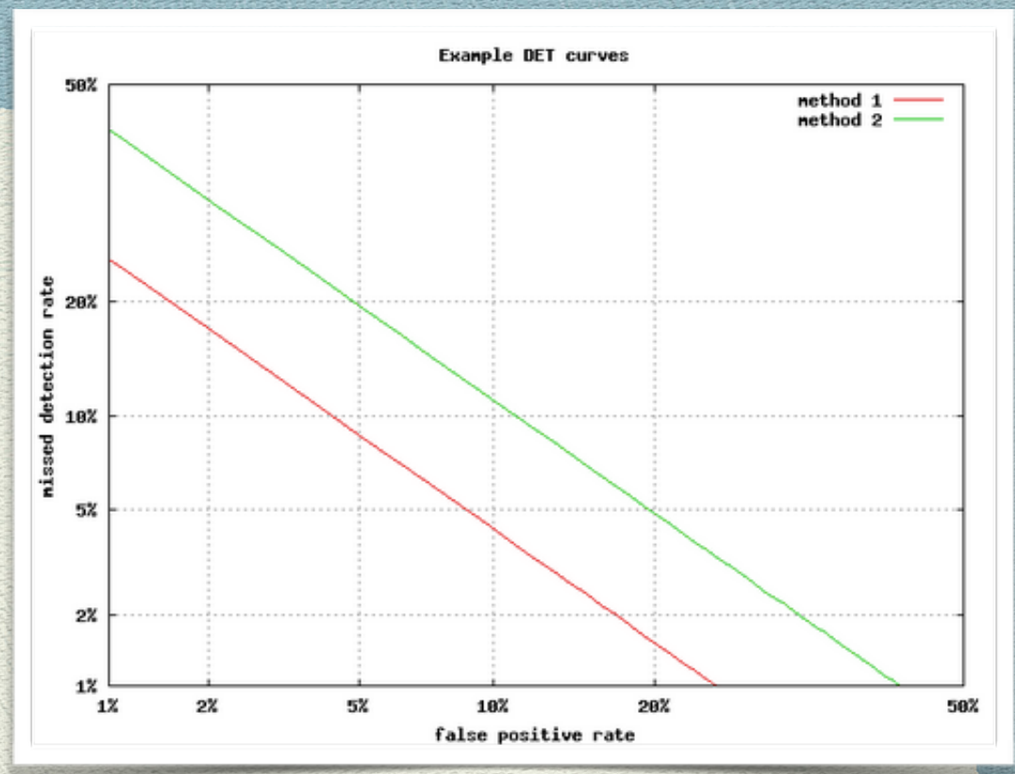


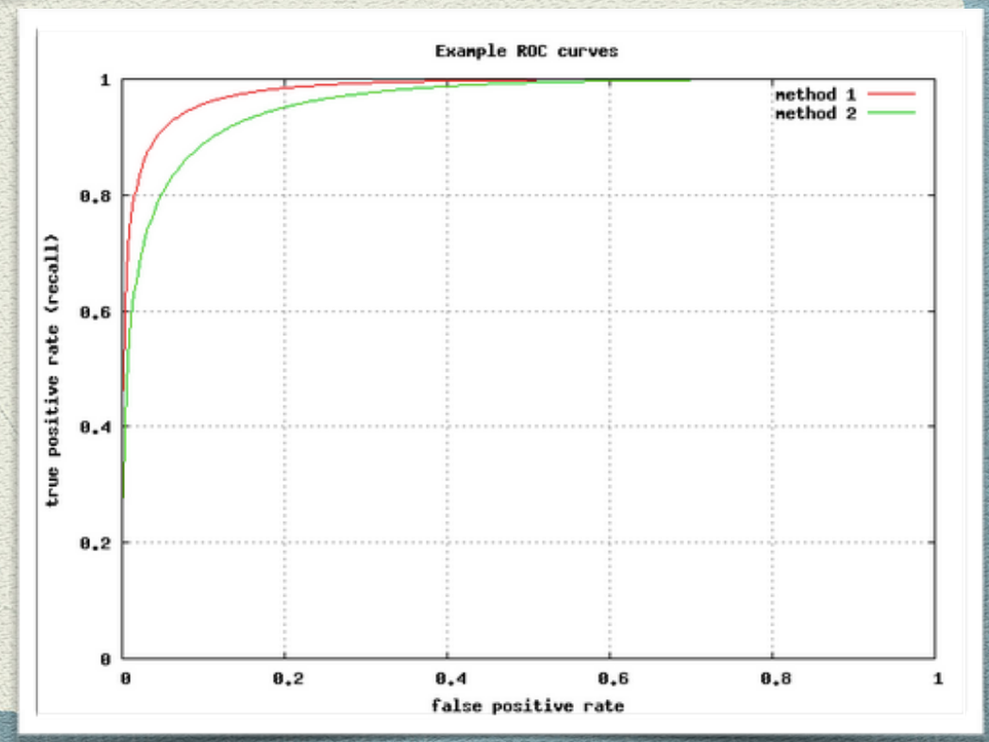
ROC -> DET Curves: 3 Changes

1. Vertical axes is flipped using miss rate (false negative rate) instead of hit rate (true positive rate)

2. (small) Rates are expressed as probabilities (or % probability)

3. Axes are logged to zoom in on a key area of ROC





ROC -> DET Curves: 3 Changes

1. Vertical axes is flipped using miss rate (false negative rate) instead of hit rate (true positive rate)

2. Rates are expressed as probabilities (or % probability)

3. Axes are logged to zoom in on a key area of ROC

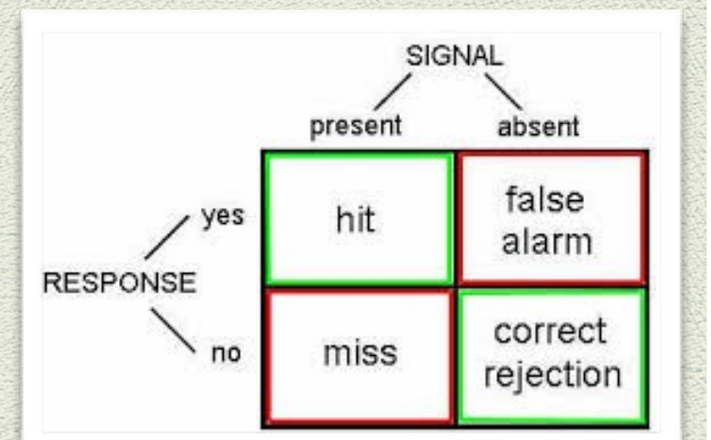
The ROC Curve traditionally has been used for this purpose. Here ROC has been taken to denote either the Receiver Operating Characteristic [2,3,4] or alternatively, the Relative Operating Characteristic [1]. Generally, false alarm rate is plotted on the horizontal axis, while correct detection rate is plotted on the vertical.

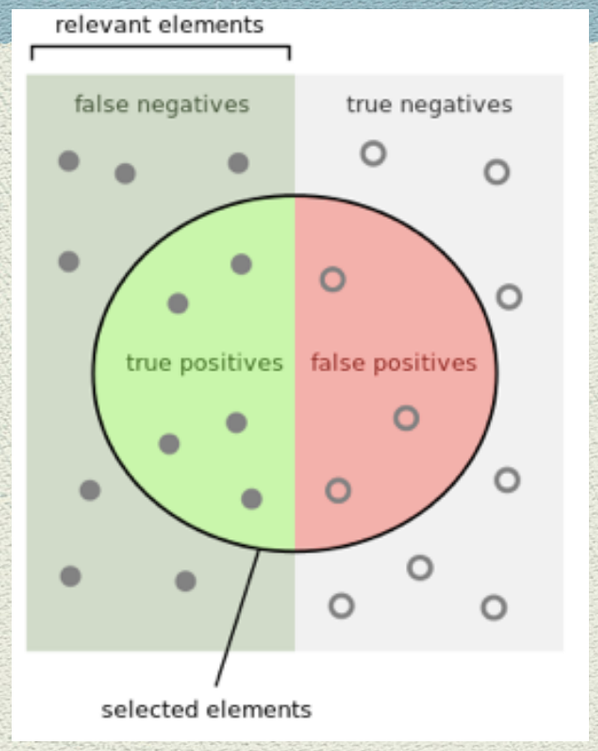
We have found it useful in speech applications to use a variant of this which we call the DET (Detection Error Tradeoff) Curve, described below. In the DET curve we plot error rates on both axes, giving uniform treatment to both types of error, and use a scale for both axes which spreads out the plot and better distinguishes different well performing systems and usually produces plots that are close to linear.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. NIST. GAITHERSBURG MD.

Axes

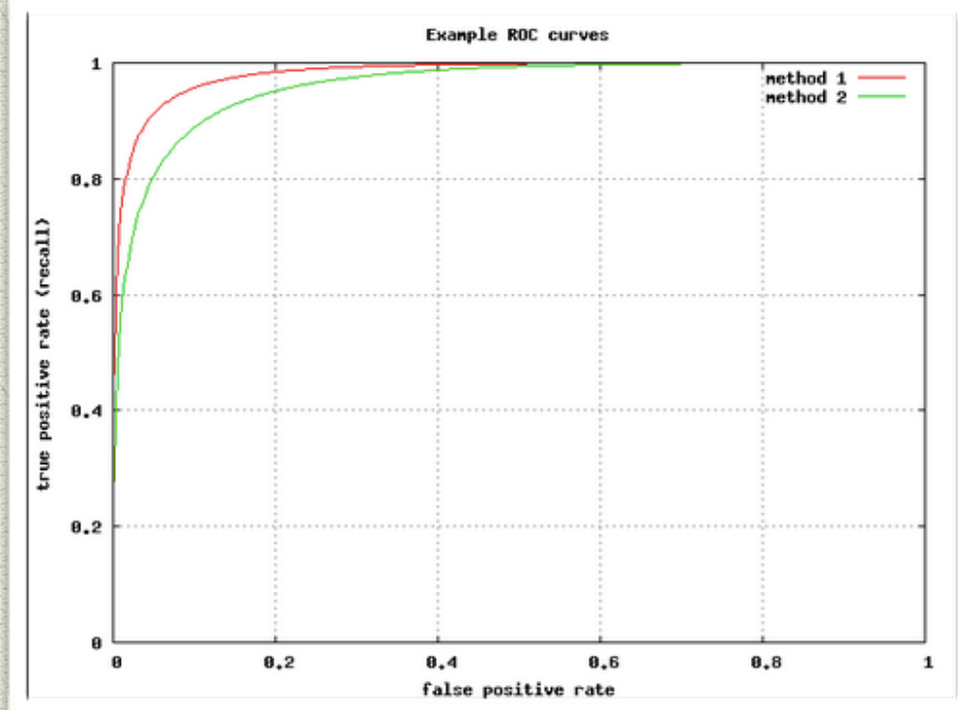
false alarm rate = false positive rate = FP/(FP+TN)
miss rate = false negative rate = FN/(TP + FN) = 1 - TPR





plots errors

		Condition (as determined by "Gold standard")	
	Total population	Condition positive	Condition negative
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)
	Test outcome negative	**False negative** (Type II error)	**True negative**



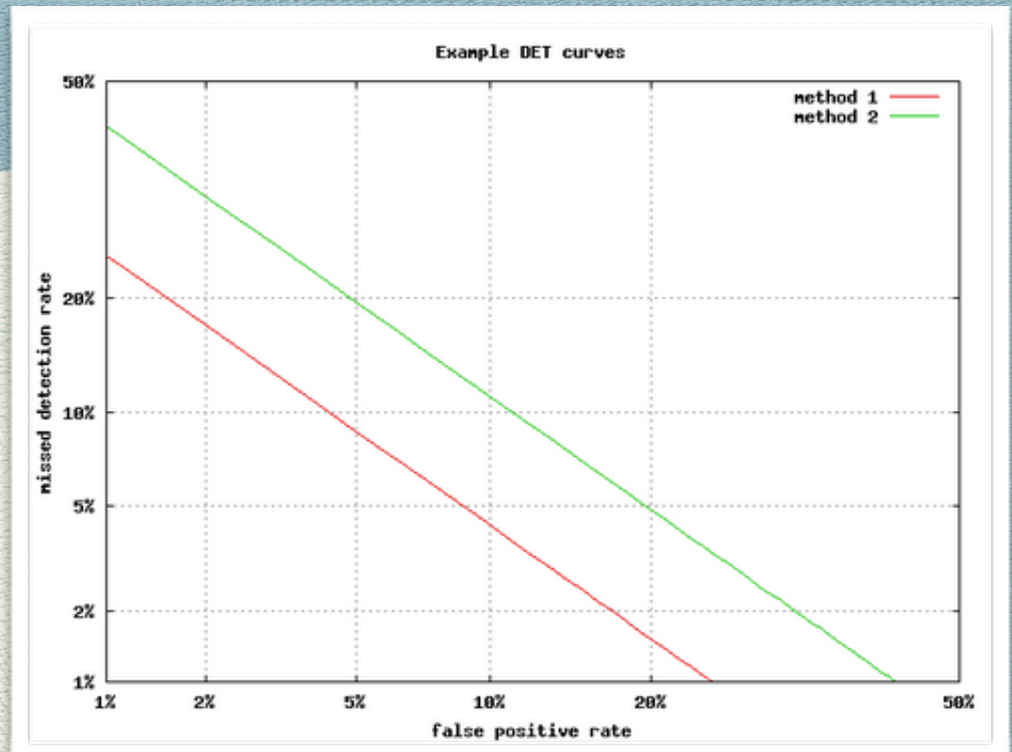
Changes y axis from True Positive Rate (TPR) to Miss Rate or False Negative Rate (FNR)

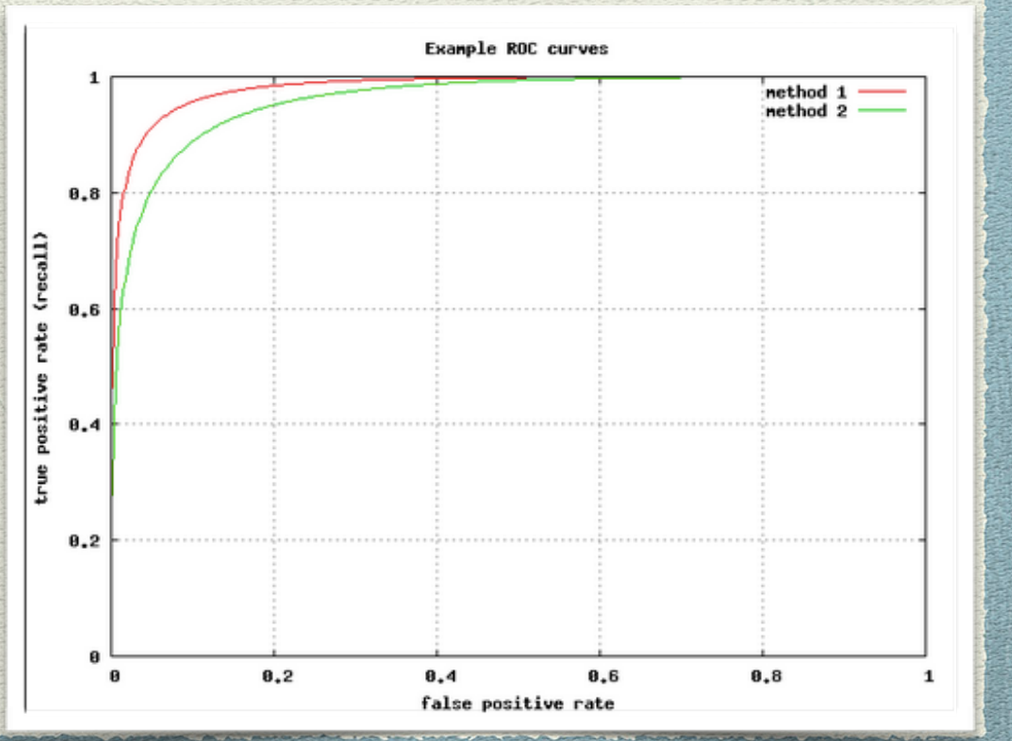
Log Scale Used

1. Vertical axes is flipped using miss rate (false negative rate) instead of hit rate (true positive rate)

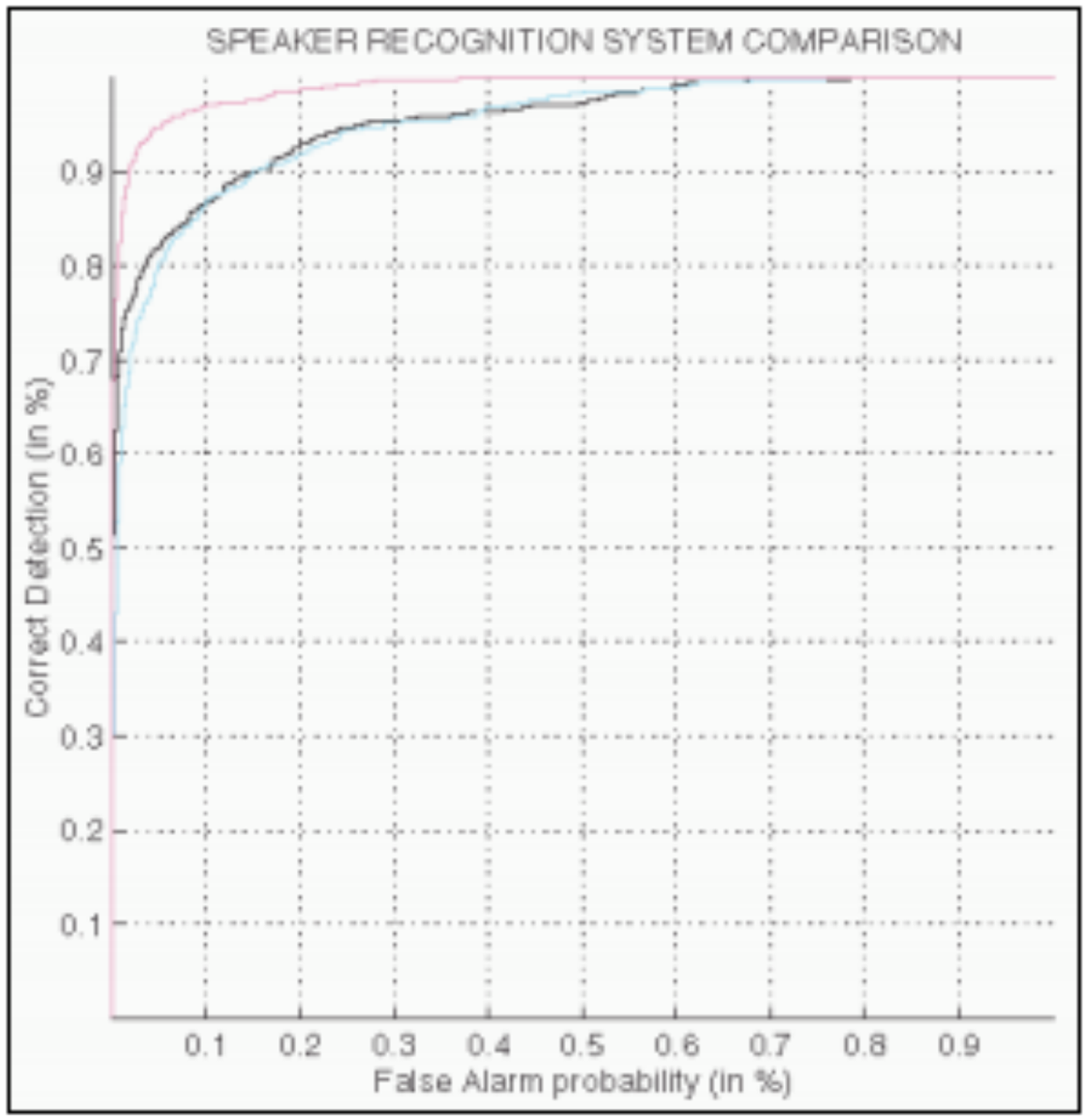
2. Rates are expressed as probabilities (or % probability)

3. Axes use logarithmic scale to zoom in on a key area of ROC (also standard deviations)





DET Curve Eg



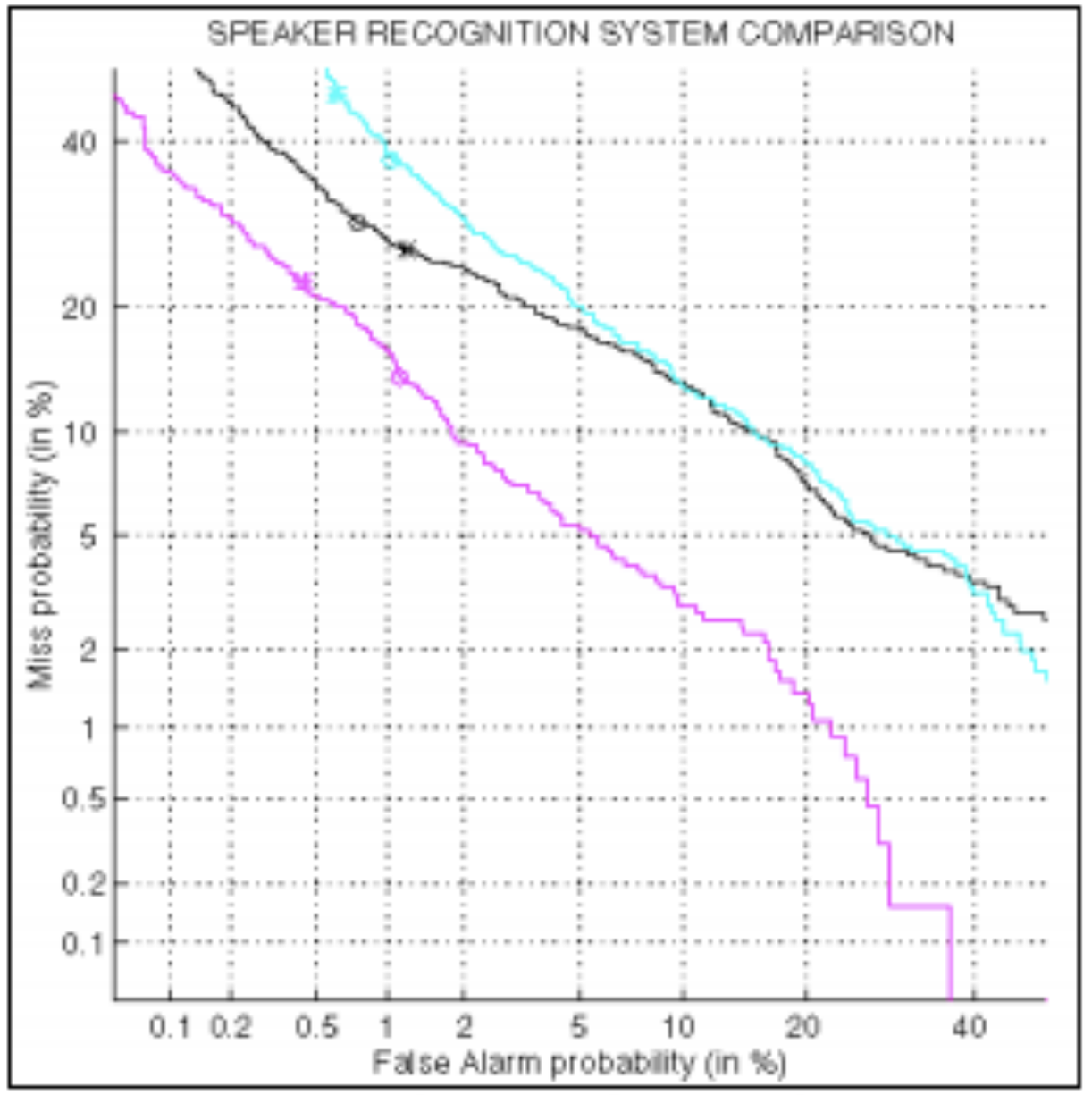


Figure 1: Plot of DET Curves for a speaker recognition evaluation.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. NIST. GAITHERSBURG MD.

DET Curve Interpretation

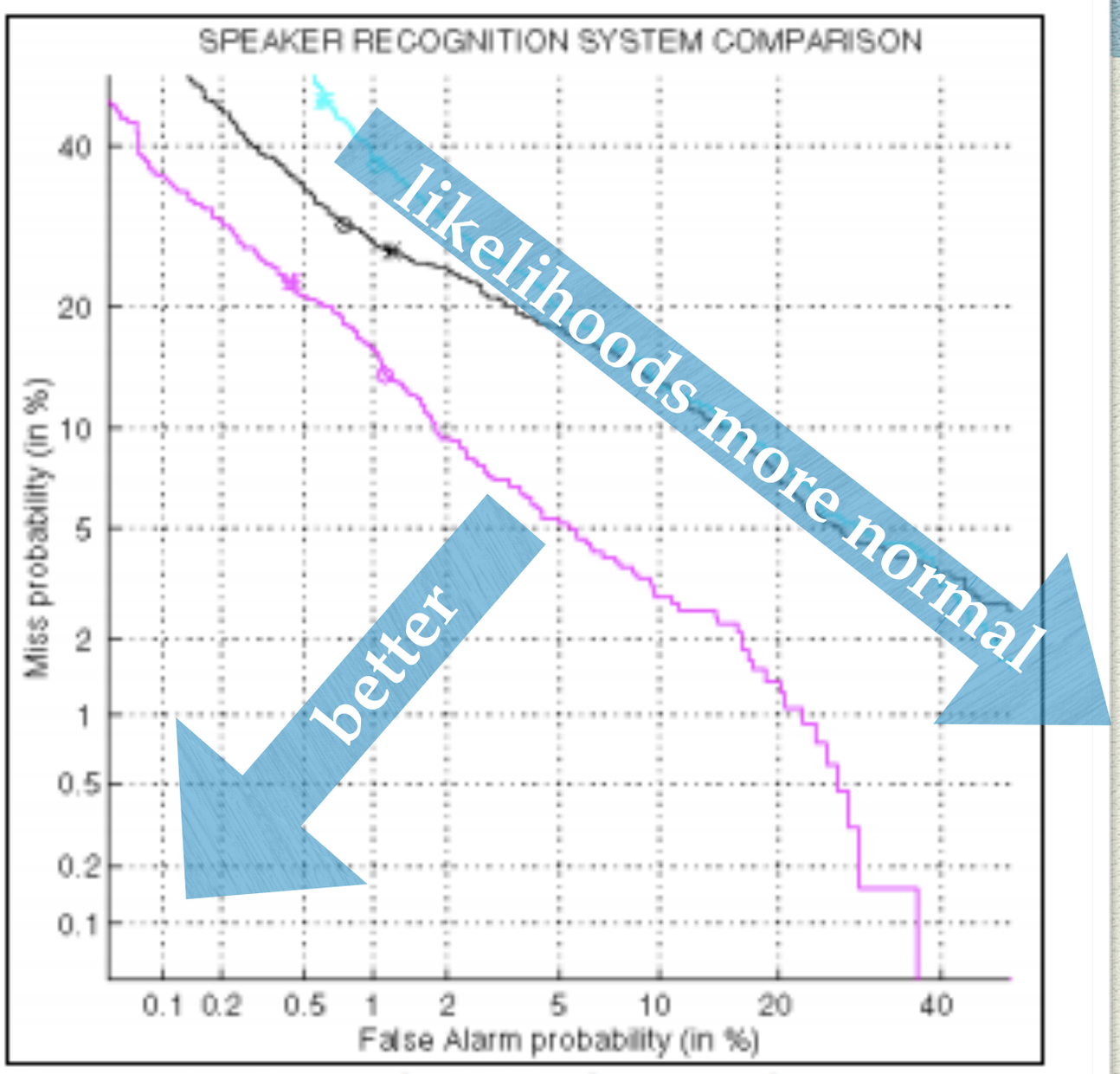


Figure 1: Plot of DET Curves for a speaker recognition evaluation.

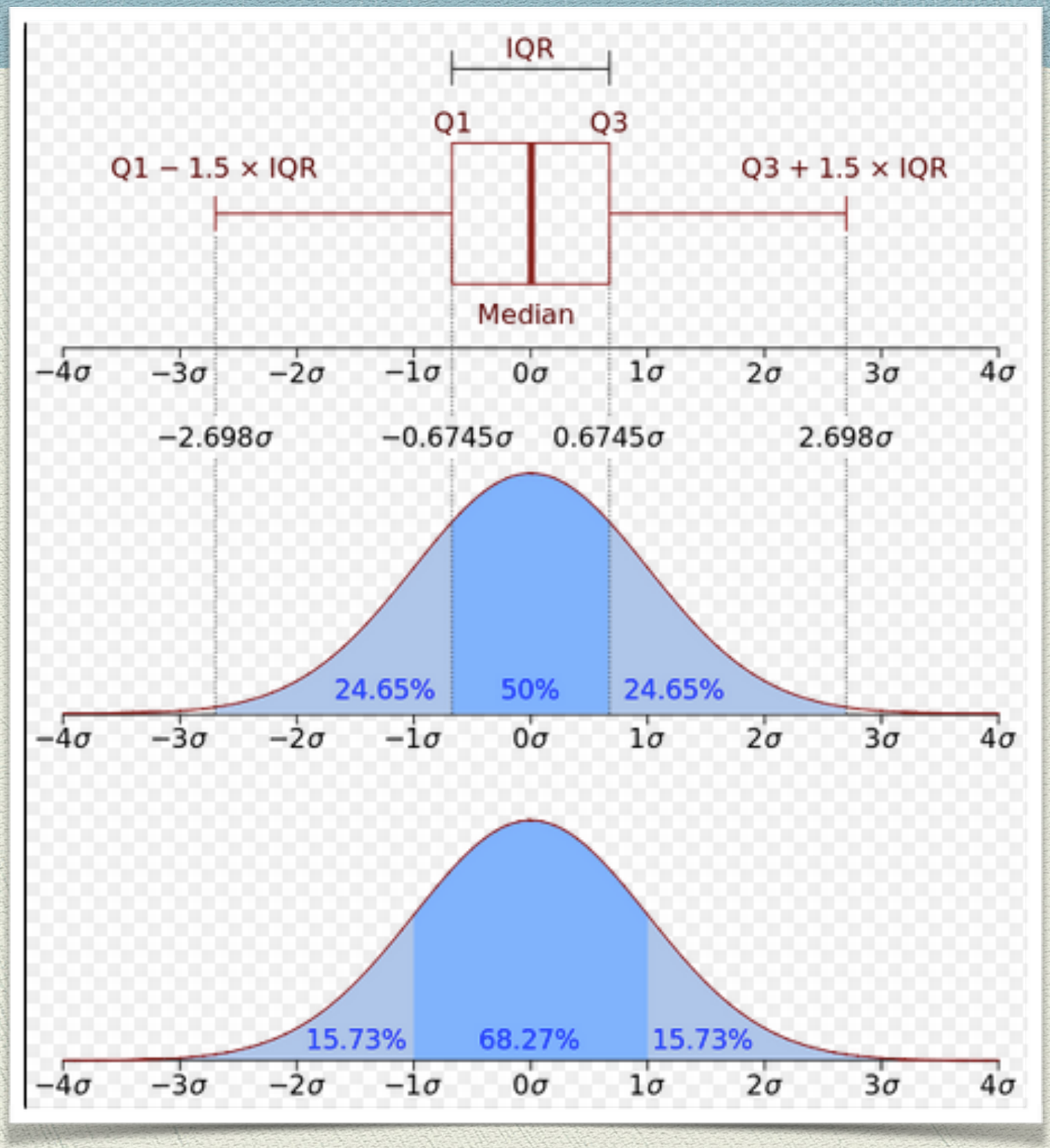
3 Common Graphing Methods…

- *ROC Curves*:
  - Receiver Operator Characteristic curves
- DET curves:
  - Detection Error Tradeoff Curves
- *Q-Q plots*:
  - Quantile-Quantile Plots

Q-Q Plots: To Answer

- Does the distribution for my system conform to some know distribution
- Are these two systems essentially the same in their behaviour (ie their distributions)
- What distribution best describes my system's outputs

Remember Normal Dist.



Q-Q Plots: What Are They?

- Q-Q Plots: is a quantile-quantile plot; quantile is the fraction of points in a distribution below the given value
- So, 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value; quantile-2 in the median of the distribution (0.5 quantile)
- Q-Q Plots: way of determining whether two data sets come from populations with the same distribution
- The comparison of the two set of quantiles is done on a 45-degree reference line; they fit this line if the same distribution

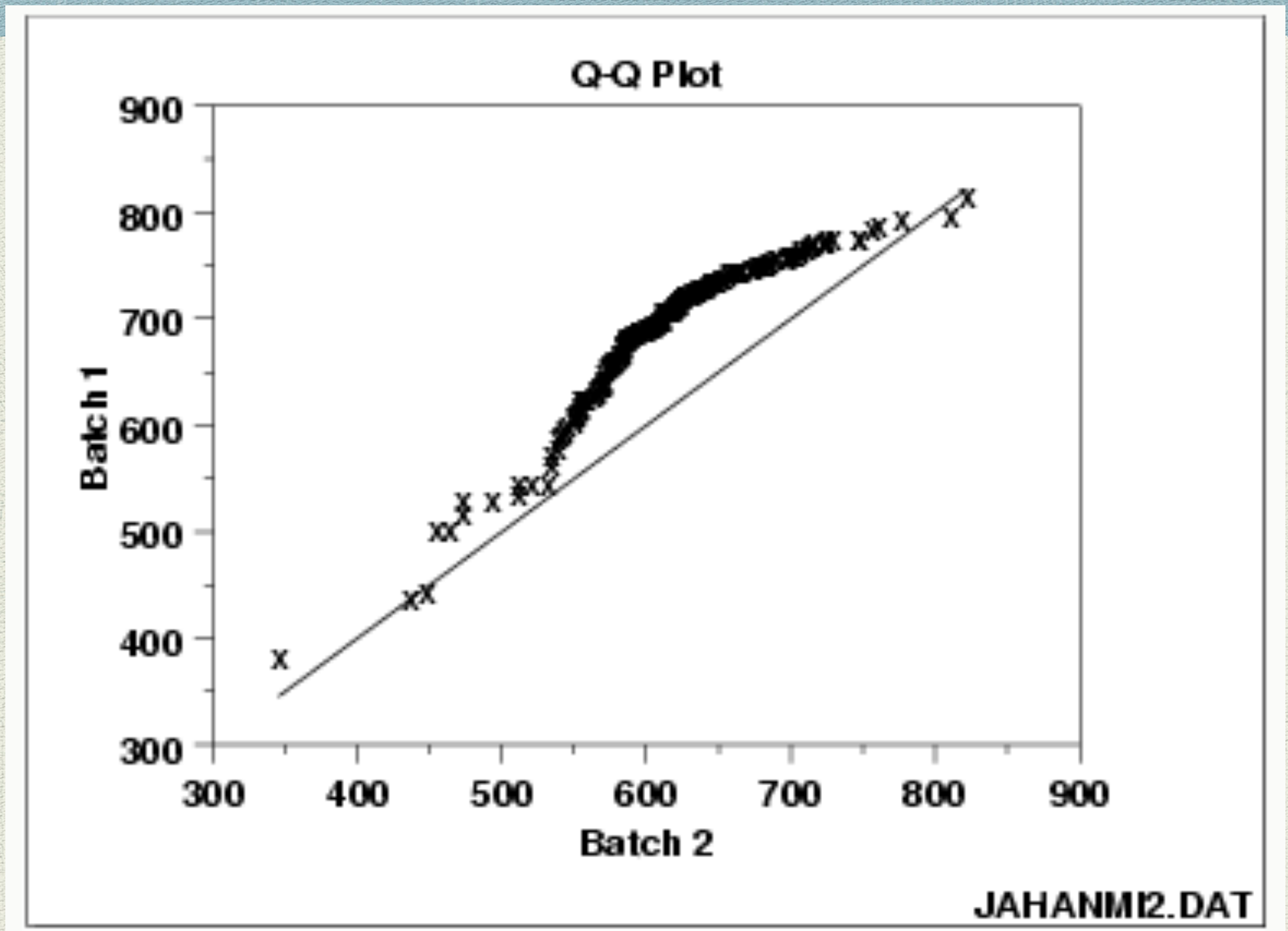
Q-Q Plots: Usages I

- Several methods for estimating quantiles from raw data; e.g. see excel PERCENTILE

- So, you can compare two data-sets by plotting the data points in each that have the same quantile

- If you get a 45-degree straight line; then they are the same; if not then they aren't

Q-Q Plots: Usages I



https://intellinexus.wordpress.com/2010/11/29/creating-a-q-q-plot/

Q-Q Plots: Usages II

- Several methods for estimating quantiles from raw data and in-theory
- So, you can compare an observed data-set with the quantiles of a theoretical plot
- Again, you get a 45-degree straight line; then they are the same; if not then they aren't

Q-Q Plots: Usages II

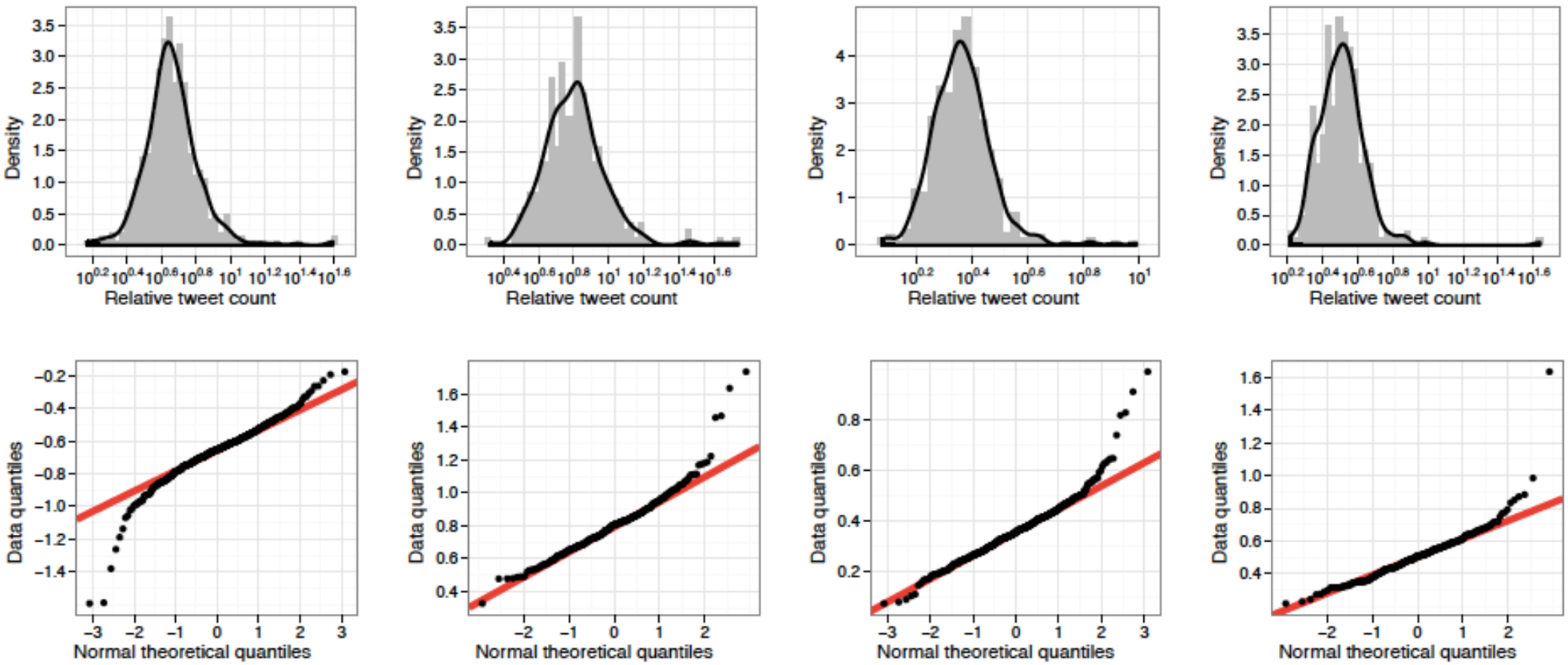
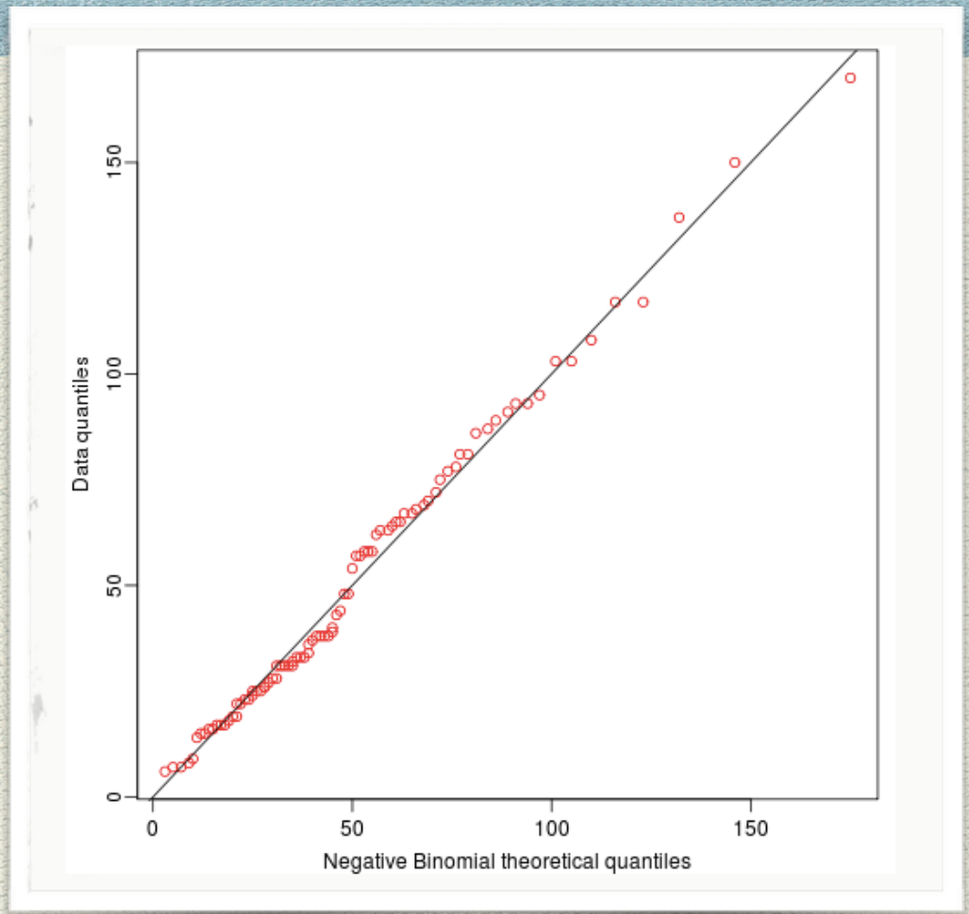
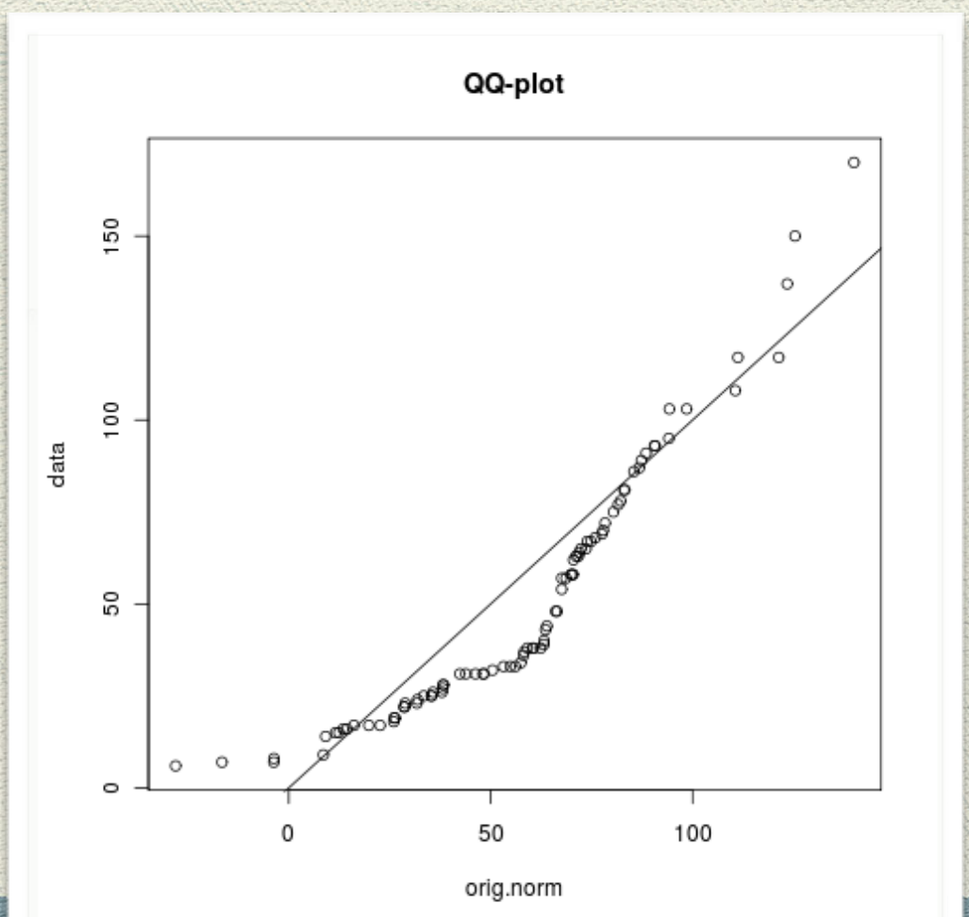


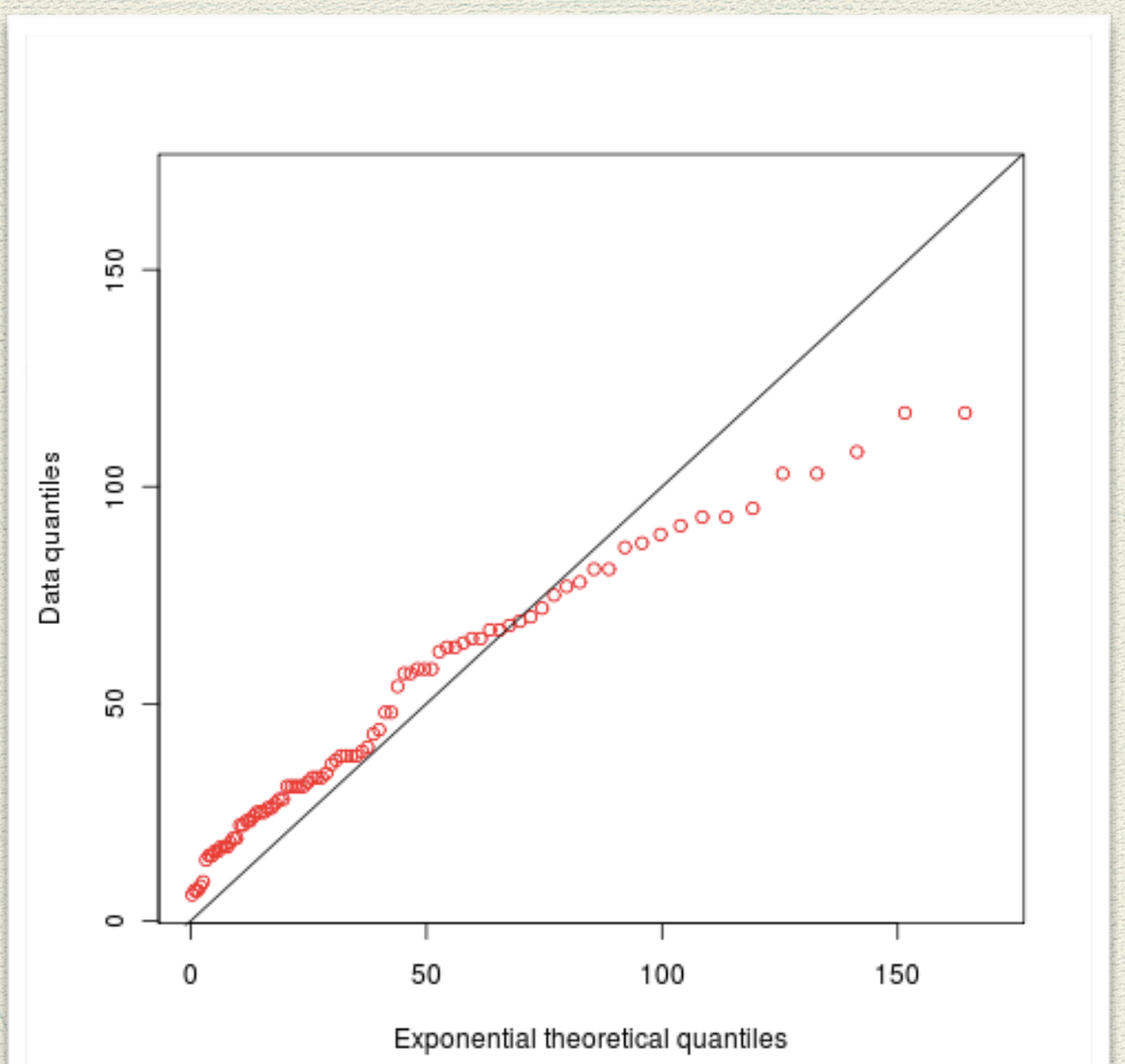
Figure 1: **(a) The densities of the ratios between cumulative tweet counts measured in two respective time frames. From left to right in the figure, the indices of the time frames between which the ratios were taken are: (2, 10), (2, 14), (4, 10), and (4, 14), respectively. The horizontal axis has been rescaled logarithmically, and the solid line in the plots shows the density estimates using a kernel smoother.** ***(b)*** **The Q-Q plots of the cumulative tweet distributions with respect to normal distributions. If the random variables of the data were a linear transformation of normal variates, the points would line up on the straight lines shown in the plots. The tails of the empirical distributions are apparently heavier than in the normal case.**

Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011). Trends in social media: Persistence and decay. Available at SSRN 1755748.

Comparison to 3 Distributions







https://intellinexus.wordpress.com/2010/11/29/creating-a-q-q-plot/

Q-Q Plots

- Note, the pattern of the deviation from the 45-degree line is telling you about the nature of the deviation from a standard distribution (fat tails etc…)

- See excel example of computing from a sample: http://www.excel-easy.com/examples/percentiles-quartiles.html

Evaluation

Automated Measures

A Few Measures More…

- *ROUGE*: Recall-based ngram co-occurrence
- *BLEU*: Precision-based ngram co-occurrence
- *WER (word error rate)*: Simple error measure

https://en.wikipedia.org/wiki/F1_score

ROUGE…

- **Recall-Oriented Understood for Gisting Evaluation**: automated summary evaluation: intrinsic rather than extrinsic (using the summary for sthm else)

- Used in DUC conferences; can be extend to tasks beyond summarisation ones (eg translation)

- Tasks: (i) single document summary: 100 word summary by H, (ii) multiple documents on same topic: 4 generic summaries of diff. lengths

https://en.wikipedia.org/wiki/ROUGE_(metric)

ROUGE…Guts

Ngram co-occurrence statistics for a *system's summary* compared to a set of *reference summaries* by Humans

Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

ROUGE-N

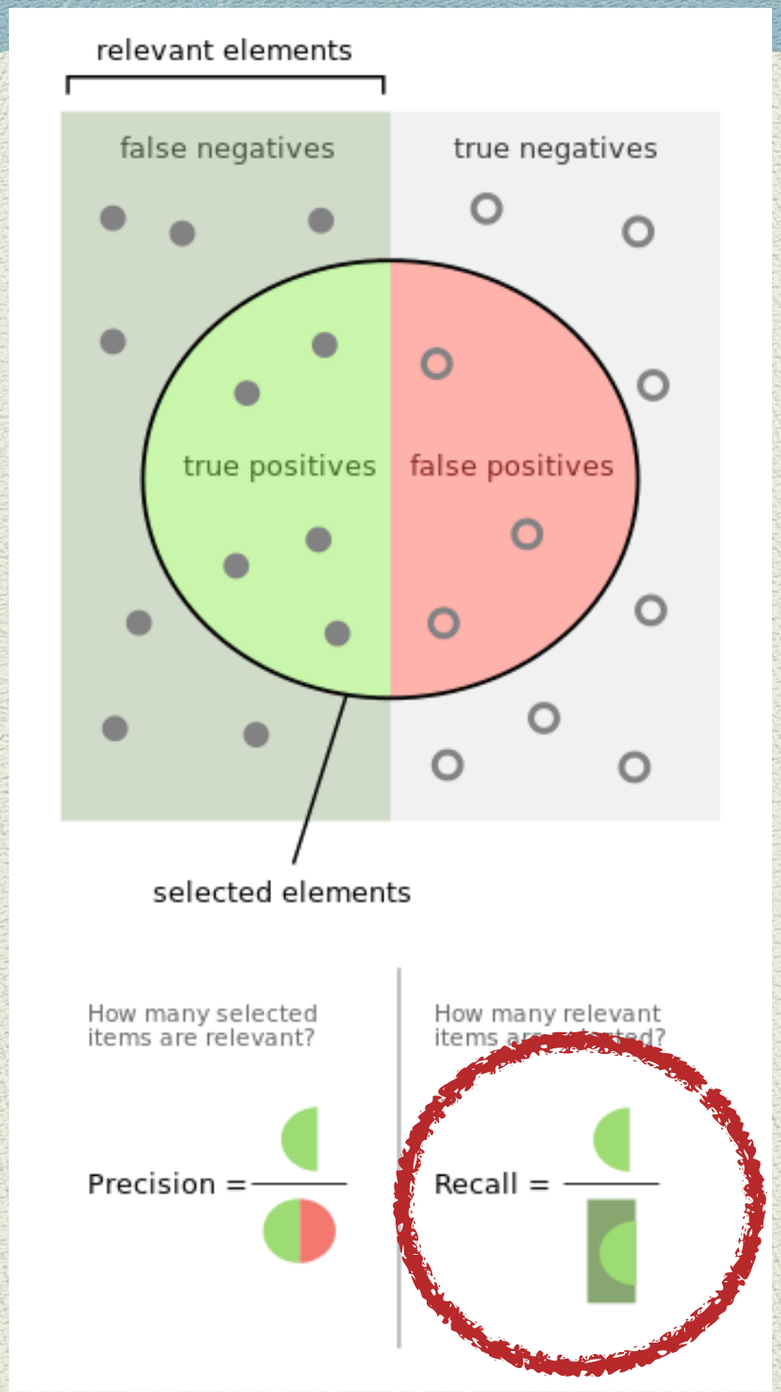
$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

Where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

It is clear that ROUGE-N is a recall-related measure because the denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: ACL-04 workshop (Vol. 8).

ROUGE…Guts



Formally, ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

Where n stands for the length of the n-gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

It is clear that ROUGE-N is a recall-related measure because the denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: ACL-04 workshop (Vol. 8).

ROUGE…Variants

Metrics [edit]

The following five evaluation metrics[2] are available.

- ROUGE-N: N-gram[3] based co-occurrence statistics.
- ROUGE-L: Longest Common Subsequence (LCS)[4] based statistics. Longest common subsequence problem takes into account sentence level structure similarity naturally and identifies longest co-occurring in sequence n-grams automatically.
- ROUGE-W: Weighted LCS-based statistics that favors consecutive LCSes .
- ROUGE-S: Skip-bigram[5] based co-occurrence statistics. Skip-bigram is any pair of words in their sentence order.
- ROUGE-SU: Skip-bigram plus unigram-based co-occurrence statistics.

ROUGE can be downloaded from berouge download link.

http://www.berouge.com/Pages/DownloadROUGE.aspx

BLEU…

- **Bilingual Evaluation Understudy**: automated translation evaluation; precision-based approach
- Also used in DUC conferences; mainly in translation but usable for summarisation etc.
- To compare candidate document (from machine translation) with 1-or-more reference translations by humans…
- Generally, sentence-by-sentence scoring then averaged over whole document (geometric mean)

https://en.wikipedia.org/wiki/BLEU

BLEU…Guts

Ngram counts but with clip on the count for repeated terms

The cornerstone of our metric is the familiar *precision* measure. To compute precision, one simply counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation.

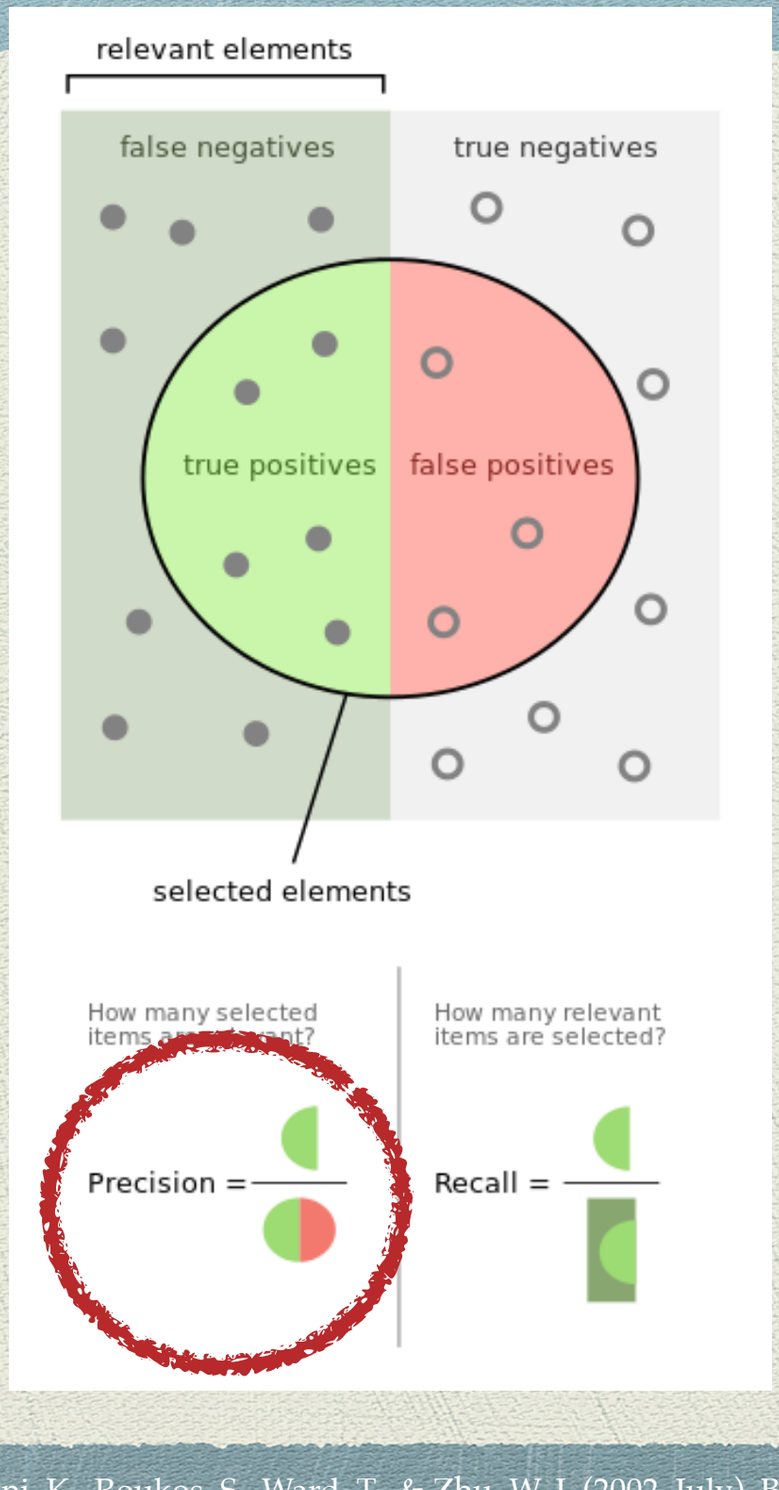
We first compute the *n*-gram matches sentence by sentence. Next, we add the clipped *n*-gram counts for all the candidate sentences and divide by the number of candidate *n*-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

[2] $Count_{clip} = \min(Count, Max\_Ref\_Count)$. In other words, one truncates each word's count, if necessary, to not exceed the largest count observed in any single reference for that word.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings 40th ACL (pp. 311-318). ACL.

BLEU…Guts



The cornerstone of our metric is the familiar *precision* measure. To compute precision, one simply counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation.

We first compute the *n*-gram matches sentence by sentence. Next, we add the clipped *n*-gram counts for all the candidate sentences and divide by the number of candidate *n*-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

[2] $Count_{clip} = \min(Count, Max\_Ref\_Count)$. In other words, one truncates each word's count, if necessary, to not exceed the largest count observed in any single reference for that word.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings 40th ACL (pp. 311-318). ACL.

BLEU…Guts

The cornerstone of our metric is the familiar *precision* measure. To compute precision, one simply counts up the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation.

We first compute the n-gram matches sentence by sentence. Next, we add the clipped n-gram counts for all the candidate sentences and divide by the number of candidate n-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

$Count_{clip} = \min(Count, Max\_Ref\_Count)$. In other words, one truncates each word's count, if necessary, to not exceed the largest count observed in any single reference for that word.

unigram precision. To compute this, one first counts the maximum number of times a word occurs in any single reference translation. Next, one clips the total count of each candidate word by its maximum reference count,[2] adds these clipped counts up, and divides by the total (unclipped) number of candidate words.

Example 2.

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.

Modified Unigram Precision = 2/7.[3]

Similarly, the modified unigram precision in Example 2 is 2/7, even though its standard unigram precision is 7/7.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings 40th ACL (pp. 311-318). ACL.

BLEU…
Guts of the Guts

Computing over all the sentences in a candidate documents and set of reference translations

2.3 BLEU details

We take the geometric mean of the test corpus' modified precision scores and then multiply the result by an exponential brevity penalty factor. Currently, case folding is the only text normalization performed before computing the precision.

We first compute the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one.

Next, let c be the length of the candidate translation and r be the effective reference corpus length. We compute the brevity penalty BP,

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

Then,

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

The ranking behavior is more immediately apparent in the log domain,

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings 40th ACL (pp. 311-318). ACL.

WER…

- **Word Error Rate:** also called length normalised edit distance
- Used in speech recognition and translation
- Really Levenshtein Distance for Words
- So, given two sequences of words how many edits do you need to change one to there other (every step reflects and error)

https://en.wikipedia.org/wiki/Word_error_rate

WER…Guts

After aligning sequences that should be identical…

Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N}$$

or

$$WER = \frac{S + D + I}{S + D + C}$$

where

- *S* is the number of substitutions,
- *D* is the number of deletions,
- *I* is the number of insertions,
- *C* is the number of the corrects,
- *N* is the number of words in the reference (N=S+D+C)

https://en.wikipedia.org/wiki/Word_error_rate

WER…Guts

After aligning sequences that should be identical…

Word error rate can then be computed as:

$$WER = \frac{S + D + I}{N}$$

or

$$WER = \frac{S + D + I}{S + D + C}$$

where

- *S* is the number of substitutions,
- *D* is the number of deletions,
- *I* is the number of insertions,
- *C* is the number of the corrects,
- *N* is the number of words in the reference (N=S+D+C)

https://en.wikipedia.org/wiki/Word_error_rate

WER…Oh…And…

When reporting the performance of a speech recognition system, sometimes *word accuracy (WAcc)* is used instead:

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

where

- *H* is N-(S+D), the number of correctly recognized words.

IF I=0 then WAcc will be equivalent to Recall (information retrieval) a ratio of correctly recognized words 'H' to Total number of words in reference 'N'.

https://en.wikipedia.org/wiki/Word_error_rate

Evaluation

Criticism & Issues

Overall

- Evaluation is clearly important to the whole enterprise; without it you cannot be sure whether your system is any good
- However, one would worry about the solipsism of some of the TREC/SNOW assessments
- Evaluation methods become overwrought; a whole world unto themselves

Take Pooling…Please take it!

- On the face of it, looks reasonable
- But, if you are only looking at the thing most system's find and not checking are these representative of the best answers…
- May propagate groupthink in system development and pass over best systems producing good outliers

Take Ground Truths…

- None of these methods are perfect for establishing a gold standard set of correct / relevant items
- The move to big-data has screwed an already shaky enterprise
- No obvious best solution…

Plots…

- Whenever you plot something, you throw something away (DET, ROC, Q-Q)

- Then, the conversation moves to a discussion of the plots alone…forgetting what was thrown away

- Need to always consider the original data, not some graphic abstraction of that data

Automated measures…

- ROUGE and BLEU are great innovations for performing large scale assessments on the outputs of many systems
- But, need to keep human grounding in mind; ROUGE shown not to correspond well to human judgement, tho' BLEU better
- Yet, many challenges persist in using ROUGE because it is an "objective" standard (of what?)

Corpora

- Assuming you have a ground truth corpus that you are sharing…it is fine as long as it is current

- Recently, several problems have arise with Tweet-corpora: sharing tweets and deletion

- Interesting, evaluation as service model but future looks rocky (companies holding onto data)