Australian National University

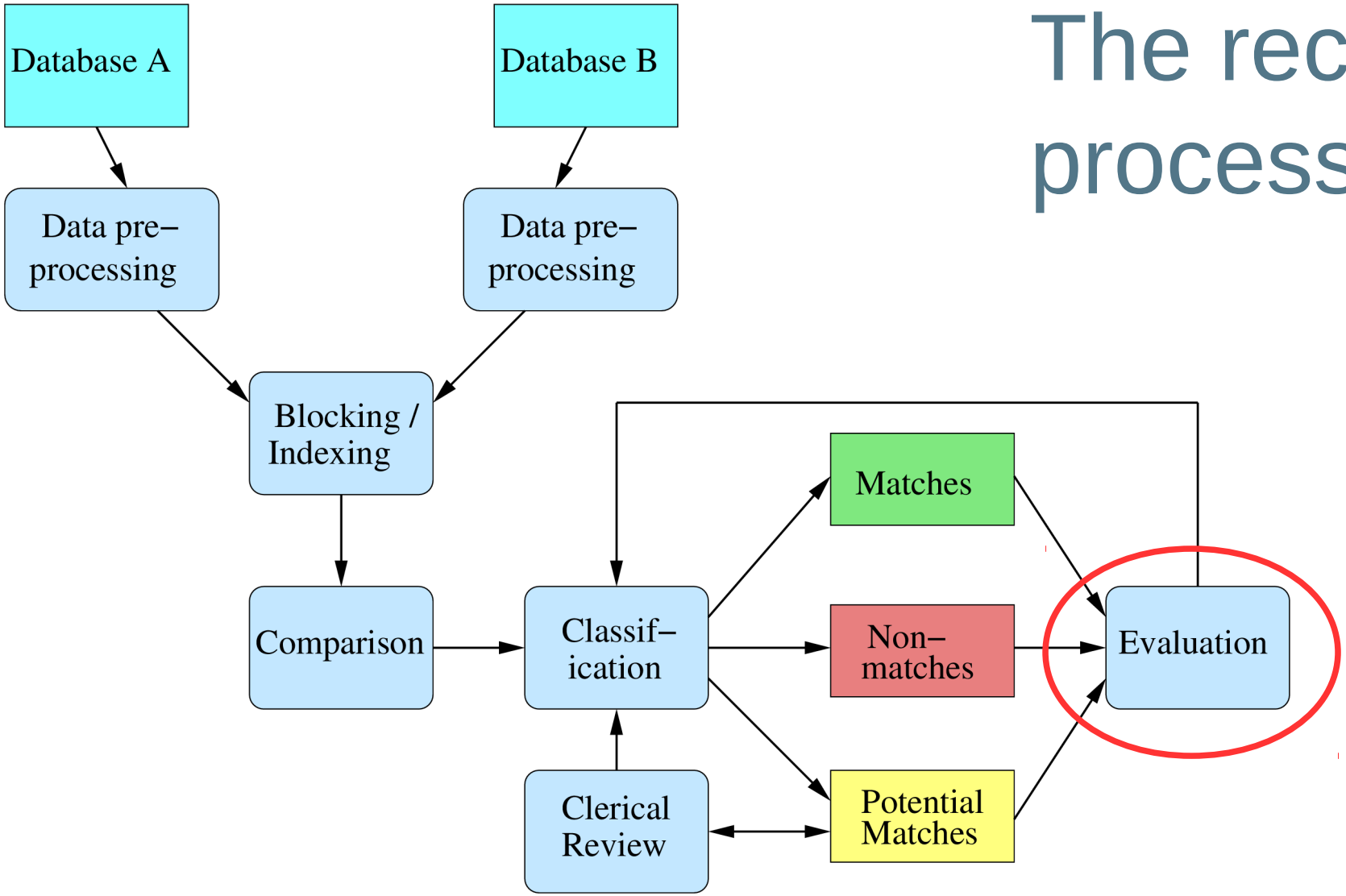
COMP3430 / COMP8430
Data wrangling

Lecture 19: Record linkage evaluation (1)
(Lecturer: Peter Christen)

Lecture outline

- Evaluating the record linkage process
- Linkage quality measures
- Linkage complexity measures

The record linkage process



Evaluating the record linkage process

- Different techniques are available for each step of the record linkage process (cleaning and standardisation, blocking, comparison, and classification)
- When employing a record linkage system, one wants to get the best possible results within operational constraints (linkage time, computational resources, minimum linkage quality, available software and human expertise, etc.)
- Measures are required to evaluate the two main aspects of record linkage
  - **Linkage quality** (effectiveness)
  - **Linkage complexity** (efficiency)

Measuring linkage quality (1)

- Achieving high linkage quality is a main goal of most record linkage projects / applications

- **Questions**: *What affects linkage quality?*
  What is required in order to be able to measure linkage quality?

Measuring linkage quality (2)

- Ground truth data is needed to measure linkage quality
  - A set of true matching record pairs
  - A set of true non-matching record pairs
- How to obtain such ground truth data?
  - Results of a previous linkage
  - Manual clerical review (more in the next lecture) or manually classified (sampled) record pairs
  - Contact all individuals in the databases and ask them?
- How confident can one be such ground truth data is always correct?

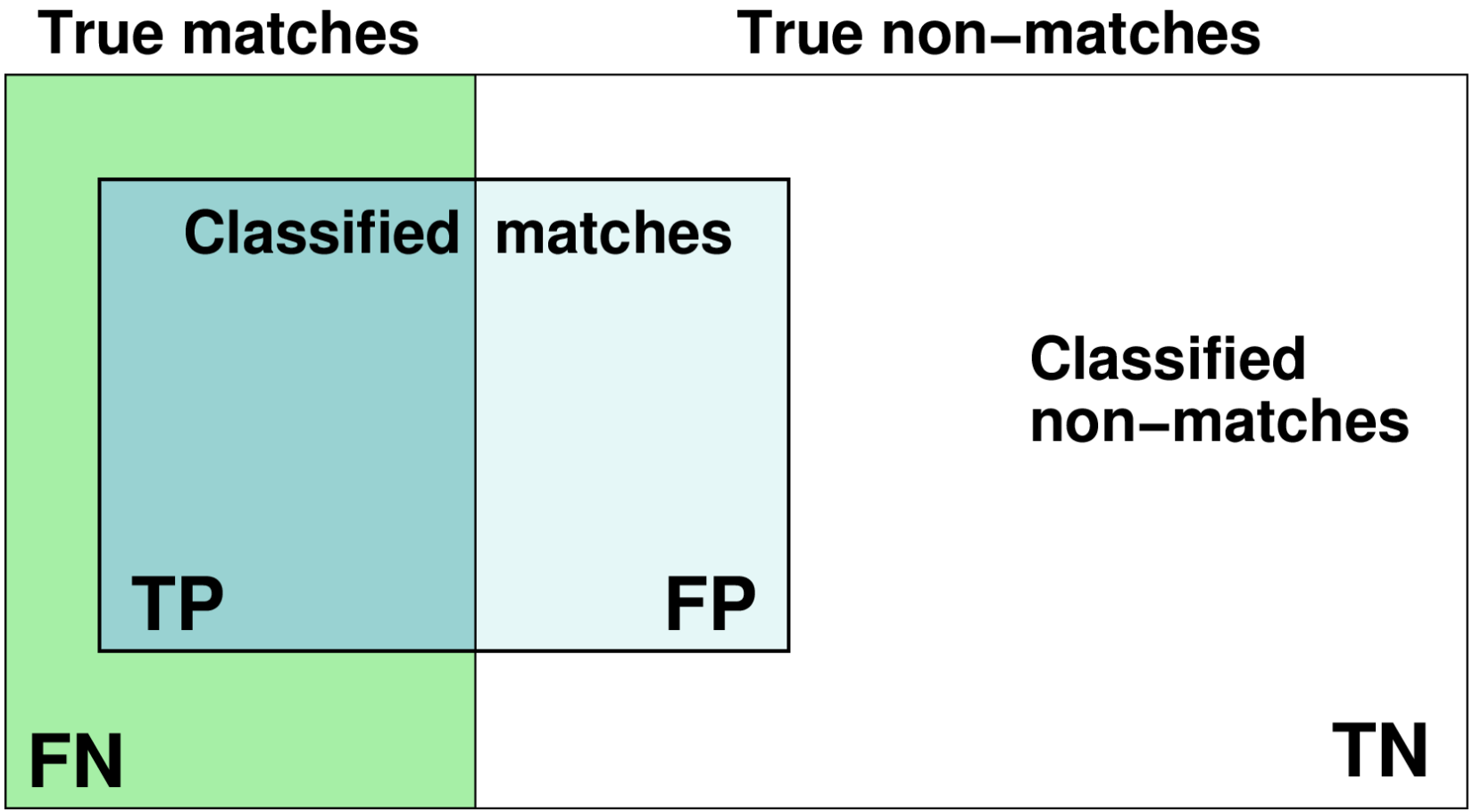
Measuring linkage quality (3)

- Various difficulties with manually prepared ground truth data
  - It is easy to classify record pairs that have totally different attribute values as non-matches
  - It is (generally) easy to classify record pairs that are very similar as matches (but what about twins, or if not enough information is available?)
  - It is difficult to classify record pairs where some attribute values are the same/similar while others are different
  - Studies have shown that manual classification of record pairs is never 100% correct
  - Domain expertise is often required (such as knowledge about names and their origins / cultures)
  - If randomly sampled, most record pairs will be non-matches

Measuring linkage quality with ground truth (1)

- Assuming ground truth data are available, the classification of record pairs into matches and non-matches has four possible outcomes:
  - **True positives** (*TP*): True matches *correctly* classified as matches (correct matches)
  - **False negatives** (*FN*): True matches *incorrectly* classified as non-matches (false non-matches)
  - **True negatives** (*TN*): True non-matches *correctly* classified as non-matches (correct non-matches)
  - **False positives** (*FP*): True non-matches *incorrectly* classified as matches (false matches)

Measuring linkage quality with ground truth (2)



Measuring linkage quality with ground truth (3)

- Due to the quadratic comparison space, the number of true non-matches is usually much larger than the number of true matches
  - As the number of records in the databases to be linked increases, the number of true matches increases *linearly* while the number of possible record pairs increases *quadratically*
- This holds even after blocking

- **Question**: *Assuming no duplicates in two databases with 1 and 5 million records, respectively, what is the maximum number of true matches between these two databases?*

Error or confusion matrix (1)

		Predicted classes	
		Matches	Non-matches
Actual classes	Matches	True positives (true matches)	False negatives (false non-matches)
	Non-matches	False positives (false matches)	True negatives (true non-matches)

Error or confusion matrix (2)

- Based on the values in the four cells of the error/confusion matrix, different linkage quality measures can be defined
- These measures are *binary classification measures* as also used in other domains
  - Machine learning and data mining
  - Information retrieval (Web search)
  - Medical tests
  - Security (airport screening), etc.
- There is often a trade-off between the number of false positives and false negatives (as one goes down the other goes up)

Accuracy

- Widely used in machine learning and data mining
- Considers both true positives and true negatives

$$acc = (TP + TN) / (TP + FP + FN + TN)$$

- **Question**: *Is accuracy a good measure for record linkage? Why or why not?*

Precision (or positive predictive value)

- Widely used in information retrieval (Web search) to assess the quality of search results (how many documents retrieved for a query are relevant?)

- Considers only true positives

  $prec = TP / (TP + FP)$

- For record linkage, it measures how many of the classified matches are true matches

Recall (or positive predictive value)

- Widely used in information retrieval to assess the quality of search results (how many of all relevant documents have been retrieved for a query?)

- Considers only true positives

  reca = TP / (TP + FN)

- For record linkage, it measures how many of all true matches have been classified as matches

F-measure: Combining precision and recall

- Precision and recall are often combined into the F-measure (or F-score):

  *fmeas = 2 * (prec * reca) / (prec + reca)*

- It is the *harmonic mean* of precision and recall
- As precision goes up (e.g. lowering a similarity threshold), recall goes down, and vice-versa
- **But**: Recent research has shown that comparing F-measure results can be misleading
  (Hand and Christen, Statistics and Computing, 2017)

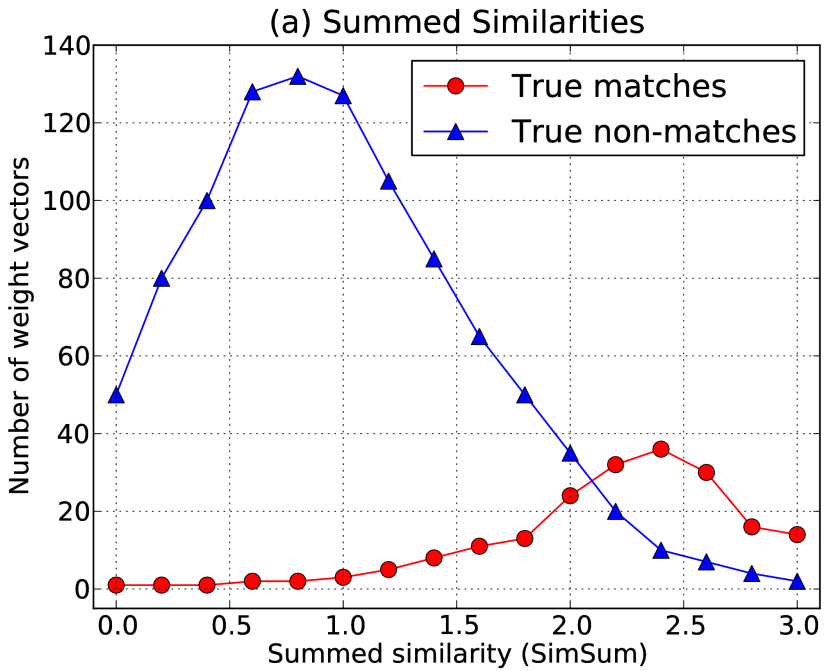
Visualising linkage quality results (1)

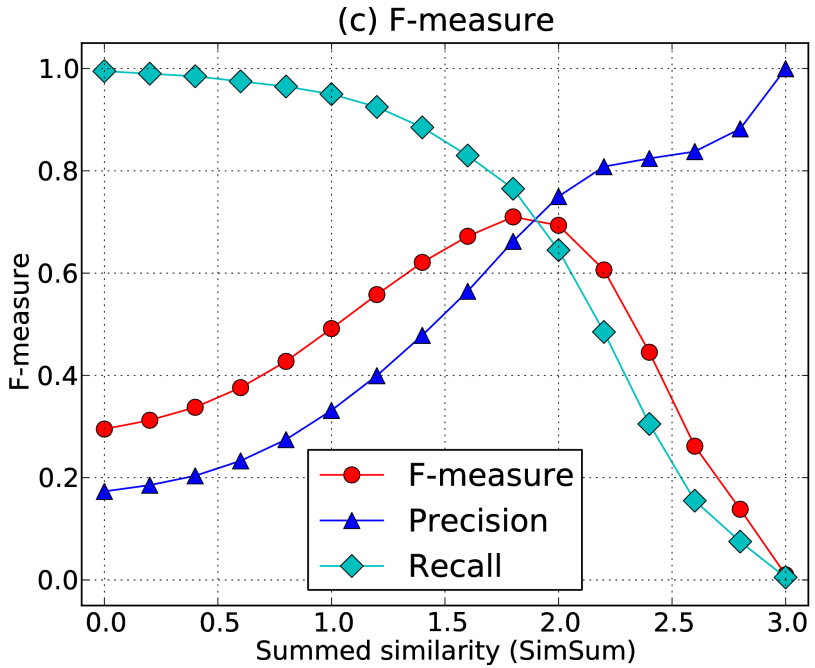
- Each of the presented measures is calculated based on a specific error / confusion matrix
- Each classifier, and each change in a classifier parameter, will produce a different error matrix
  - Lowering a classification threshold, *t*, will usually increase the numbers of TP but also FP, and lower the numbers of TN and FN
  - Raising a classification threshold leads to the opposite

- To better understand classifiers and to compare them, plots are useful tools (for example for different classification thresholds)

Visualising linkage quality results (2)

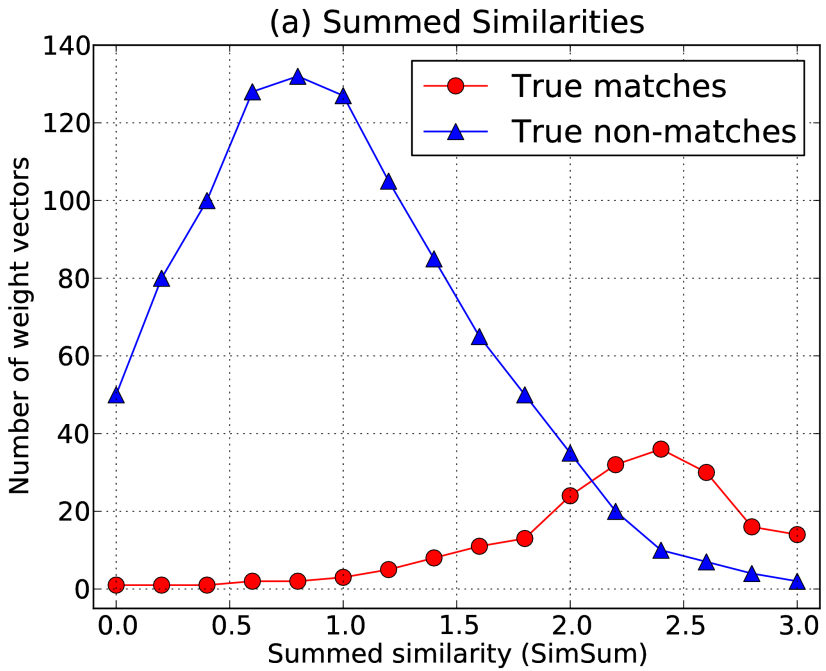
- F-measure graph shows precision, recall and f-measure for different classifier thresholds

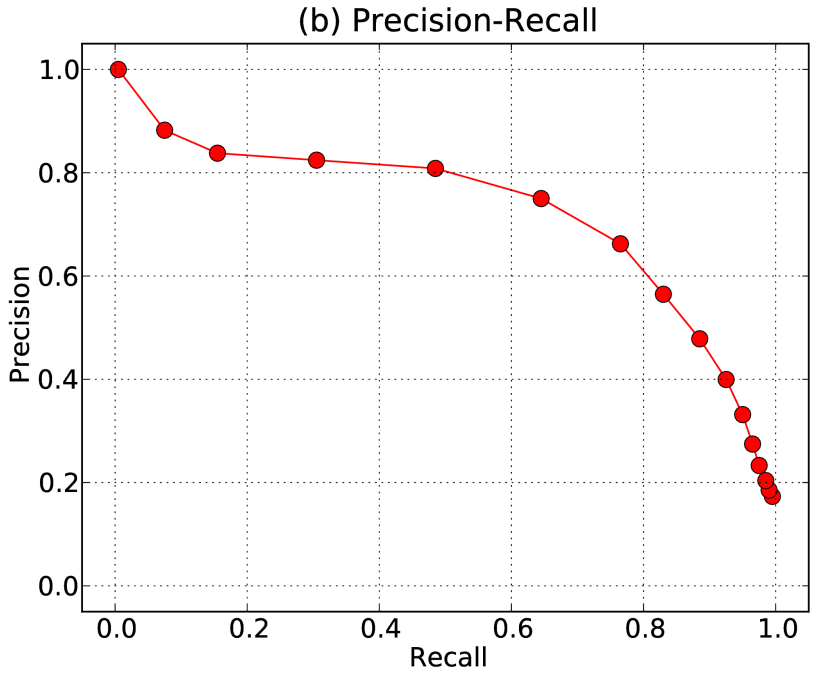




Visualising linkage quality results (3)

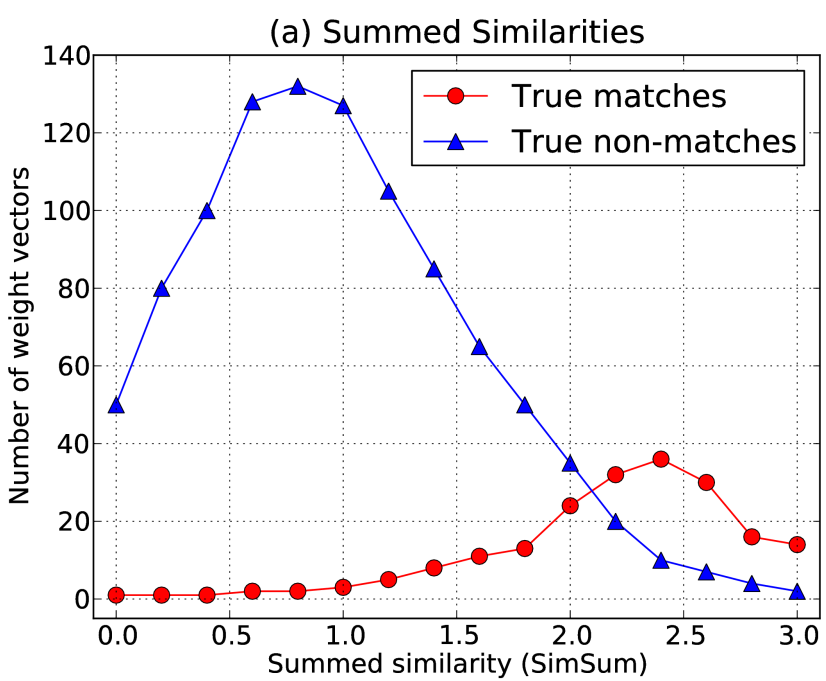
- Precision-recall graph shows precision versus recall for different classifier thresholds

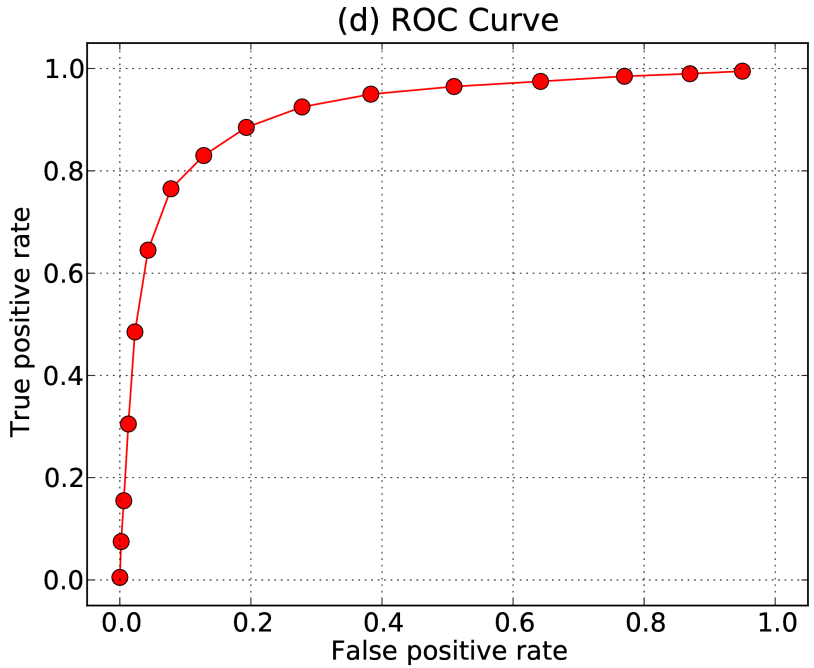




Visualising linkage quality results (4)

- ROC (receiver operating curve) graph shows true positive rate versus false positive rate for different classifier thresholds





Measuring linkage complexity

- We can easily measure run-time and memory consumption of a linkage program / system (but is this meaningful)?

- Generally, platform independent measure are better
  - Allows the performance of systems to be compared even when not run on the same computing platform (but same data sets and same parameter settings)
- Linkage complexity is generally measured by the number of record pairs that need to be compared
  - The number of candidate record pairs generated by blocking

Reduction ratio

- Measures by how much a blocking technique is able to reduce the comparison space
  – Compared to the full pair-wise comparison of all record pairs

$$rr = 1 - (s_M + s_N) / (n_M + n_N)$$

where:

– s_M and s_N are the number of true matching and non-matching candidate record pairs generated by a blocking technique

– n_M and n_N are the total number of true matching and non-matching record pairs (in the pair-wise comparison space)

Pairs completeness

- Measures how many true matches 'pass' through a blocking process
- It corresponds to the *recall* of blocking

$$pc = s_M / n_M$$

- It requires the truth match status of all record pairs
  (as with the linkage quality measures)

Pairs quality

- Measures how many candidate record pairs generated by blocking are true matches
- It corresponds to the *precision* of blocking

$$pq = s_M / (s_M + s_N)$$

- It requires the truth match status of all record pairs
  (as with the linkage quality measures)