Deterministic vs. Probabilistic Record Linkage

Stat 36-491/691

September 11, 2013

Comparing Records in Two Files

- Match

  = comparison pair of records for the same unit (person)
- Non-Match

  = comparison pare of records that are for two different units (persons)
- Link

  = comparison pair of records that is accepted as being a match
- Non-Link

  = comparison pair of records that is not accepted as being a match

Matches and Links

	Matches	Non-matches
Linked	**a** (true positives)	**b** (false positives)
Unlinked	**c** (false negatives)	**d** (true negatives)

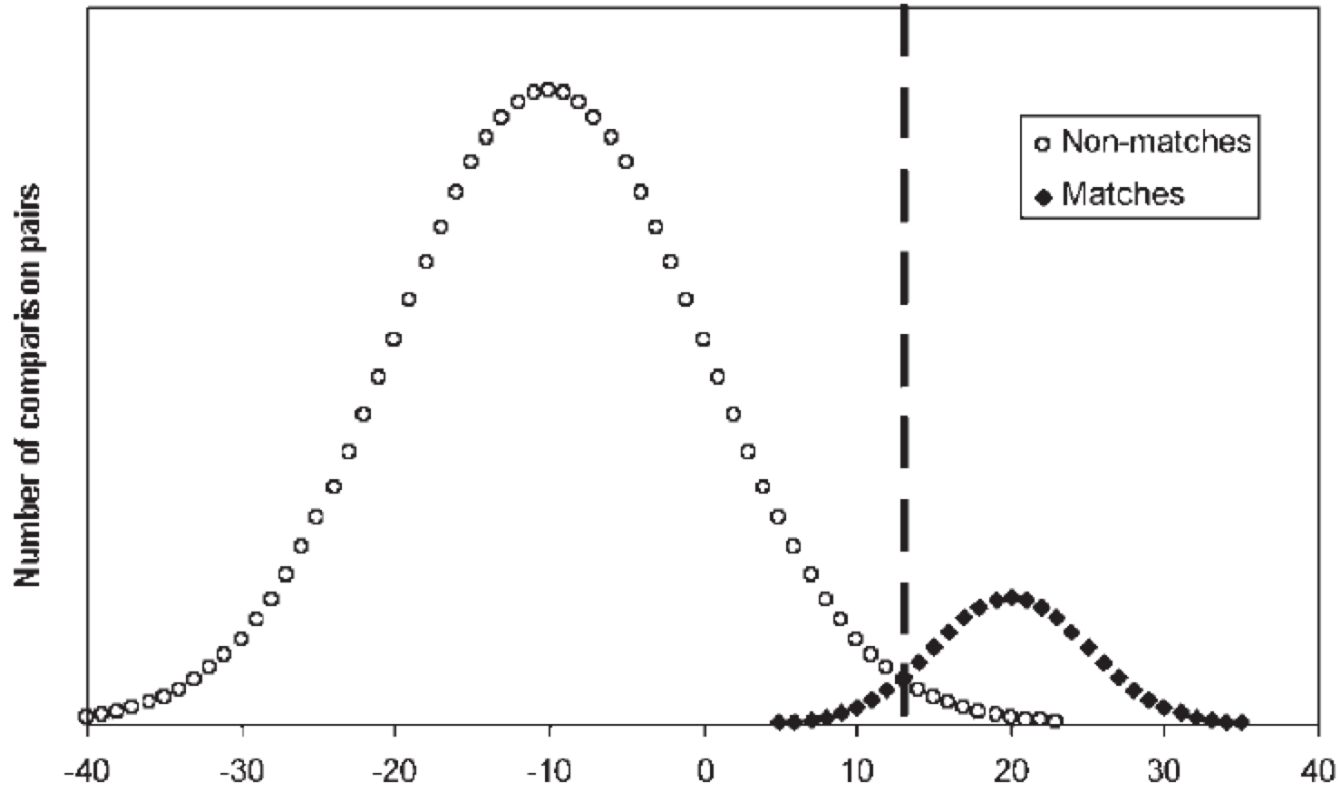
- Sensitivity = a/(a+c)
- Specificity = d/(b+d)
- Positive predictive value = a/(a+b)
- Negative predictive value = d/(c+d)

Discriminating Between Two Groups

- Same ideas of
  - False positives
  - False negatives
  - PPV
  - NPV

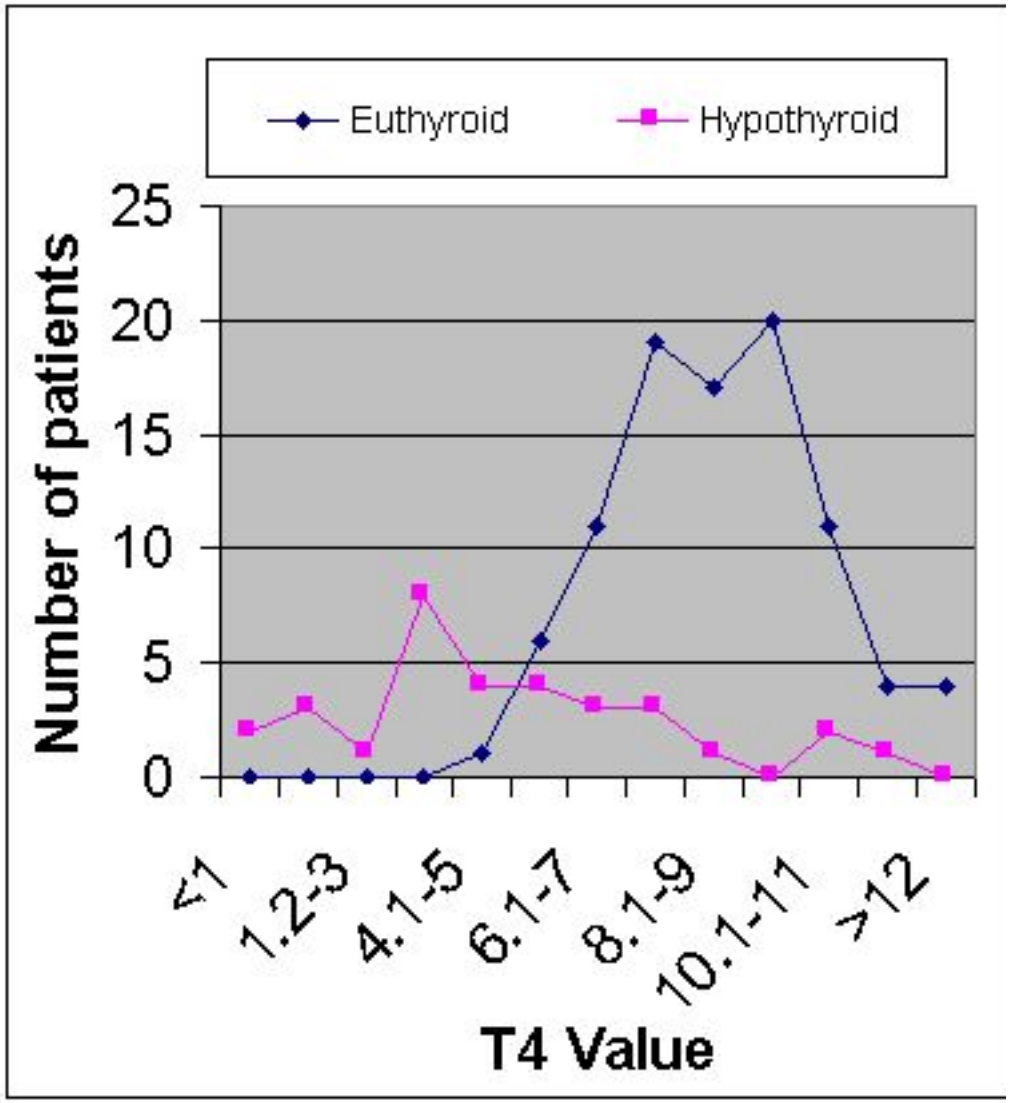
ROC Curves

- Trace out the tradeoff between the two kinds of error as we shift the cut-off:



Example: Hyperthyroidism

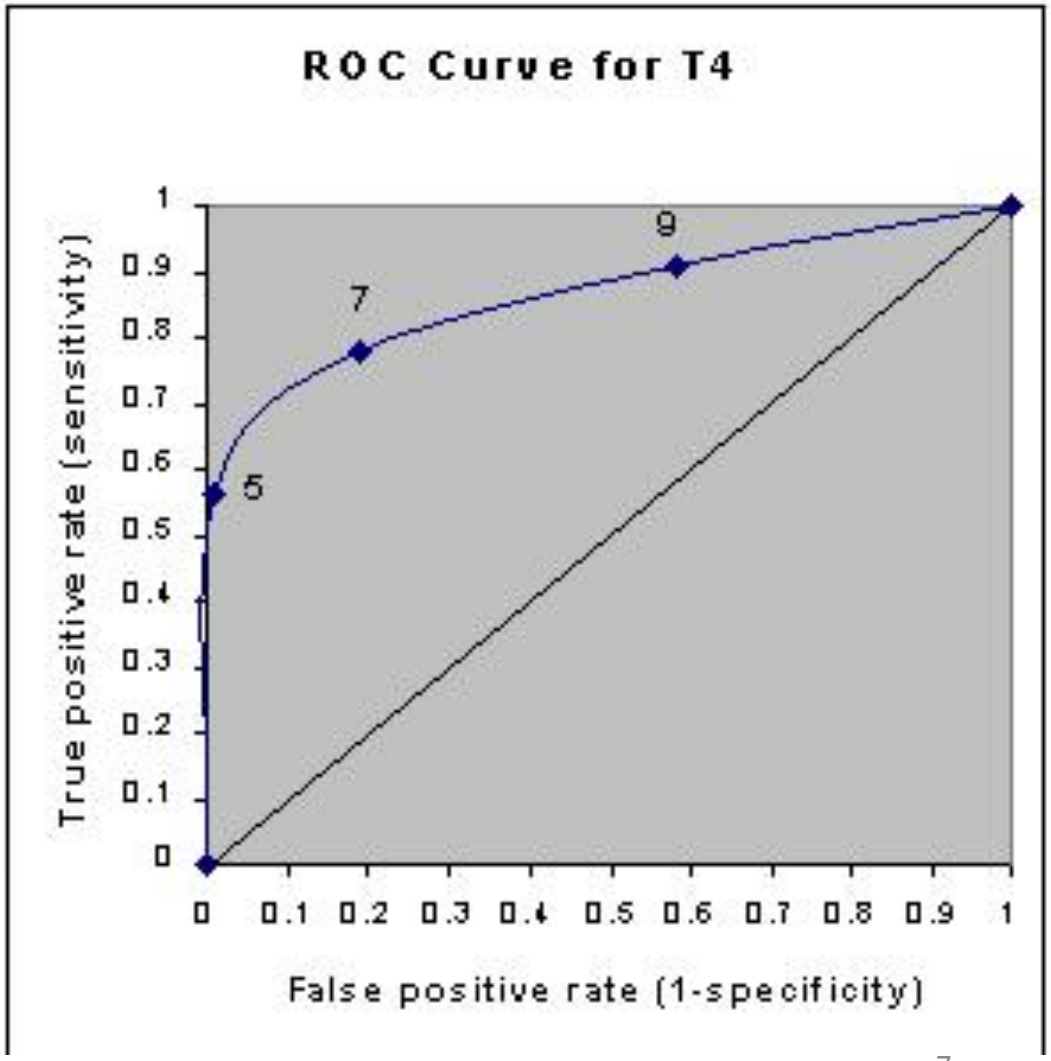
T4 value	Hypothyroid	Euthyroid
5 or less	18	1
5.1 - 7	7	17
7.1 - 9	4	36
9 or more	3	39
Totals:	32	93



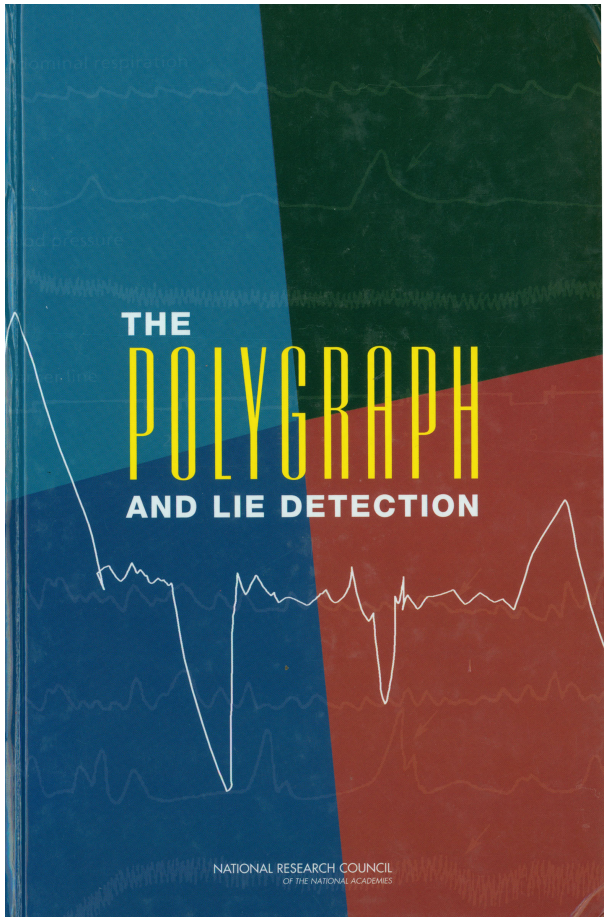
Example (cont.)

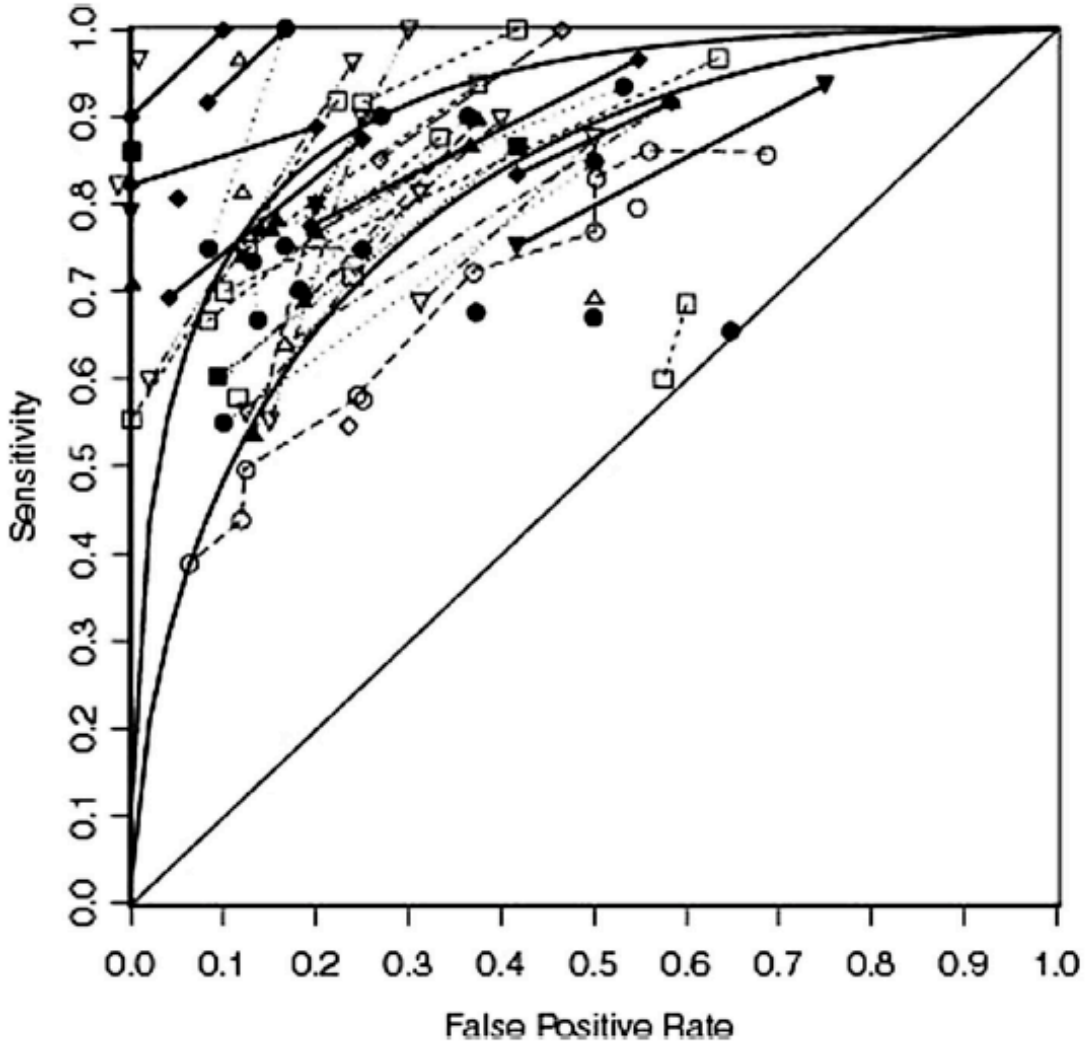
Cutpoint	True Positives	False Positives
5	0.56	0.01
7	0.78	0.19
9	0.91	0.58

Cutpoint	Sensitivity	Specificity
5	0.56	0.99
7	0.78	0.81
9	0.91	0.42



Example: Polygraph Accuracy





Knowing “Ground Truth”

- Can only validate accuracy and estimates of both kinds of errors if we have ground truth.
- This will be an issue when we come to apply ideas to RL.

Back to Record Linkage

- Units of interest are record pairs:
  - Now I will use match to denote the labeling from a record linkage method (previously "link")
  - Get distribution of matches and distribution of non-matches
  - Choose a cut-off and look at errors

- All RL methods make errors
- We want ones with high sensitivity and high specificity
- The choice of cut-offs determines the trade-off

Deterministic Matching

- Record linkage of two or more files based on exact agreement of matching variables. (Sort/ Merge)
  - Works best when there is a single unique identifier (key), e.g., SS#
  - Can also use with multiple matching variables that are not unique to each person
  - **What is the implication of a coding error?**
  - **What happens if there are duplicate files?**
    - **How can this happen?**

Probabilistic Matching

- Record linkage of two or more files that utilizes the probability of agreement or disagreement on a range of matching variables, or estimates of probabilities
  - Can use more variables and allow for error of measurement or coding error
  - Need a metric to apply to pairs that combines "similarities" on different variables
    - weights

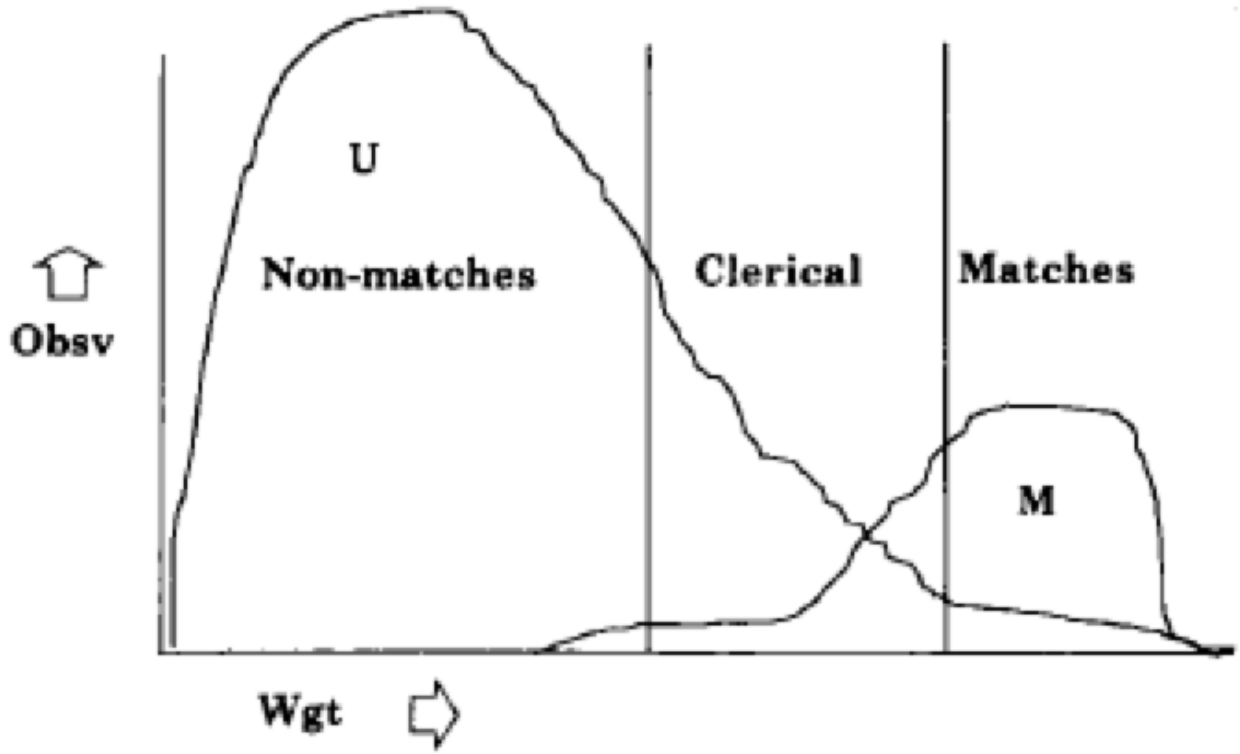
Are String Metrics Distances?

- What are the properties of distances?
- Which string metrics satisfy them?

- Do you know a metric that isn't a distance?

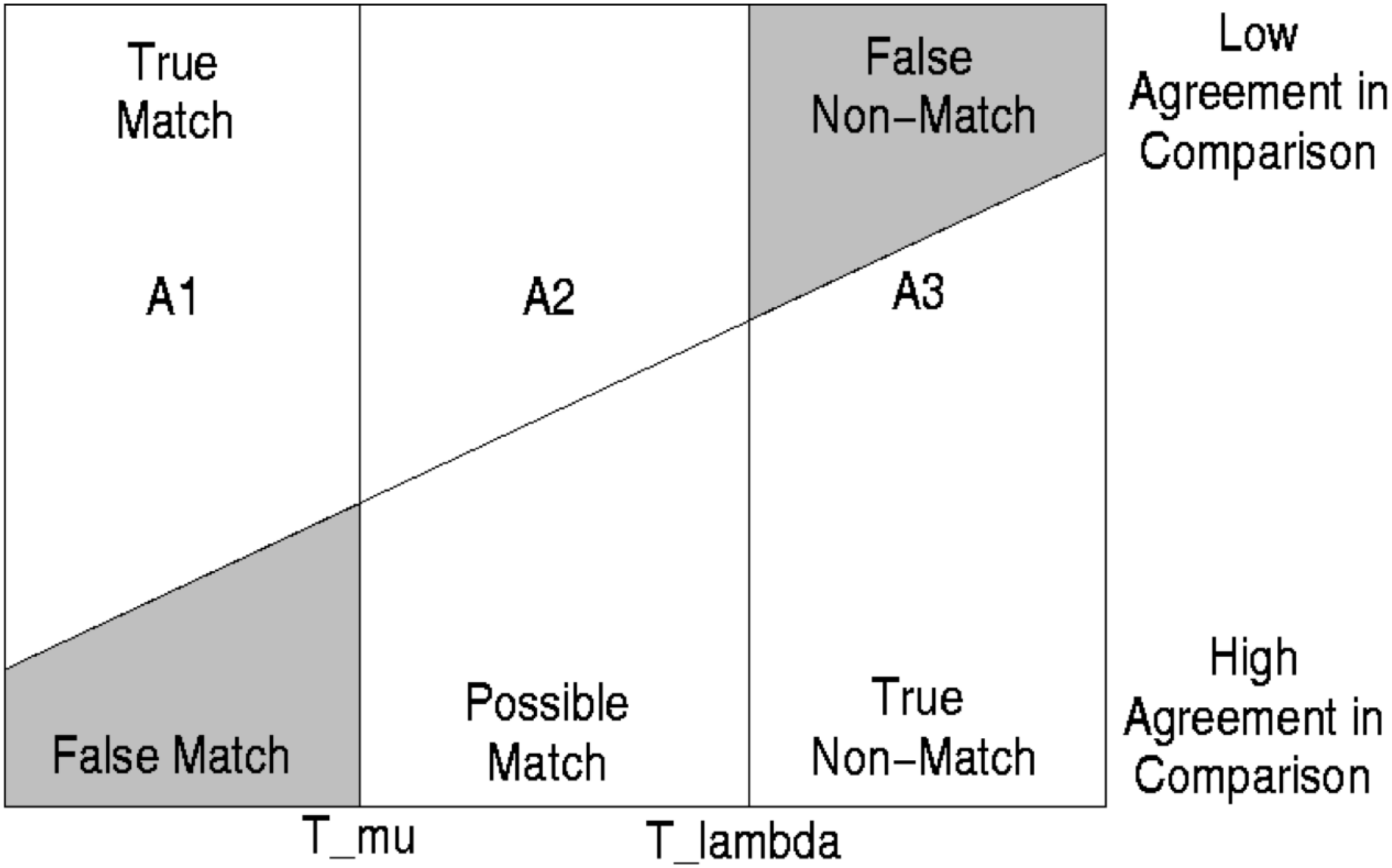
Record Linkage With Two Cutoffs

- Use one cut-off for links, and another for non-links



Choosing Cut-offs

- Supervised versus unsupervised learning
  - For supervised learning we need ground truth for a set of cases in both files
  1. We can choose cut-offs to distinguish between the the matches and non-matches in the known cases.
  2. Then apply to everything.
- Otherwise we need to estimate from the data or choose in some other way.


True
Match
A1
False
Non–Match
A3
A2
Low
Agreement in
Comparison
False Match
Possible
Match
True
Non–Match
High
Agreement in
Comparison
T_mu
T_lambda

Next Week

- Fellegi-Sunter method for record linkage
- Homeworks are due on Wednesdays
- I am away M through Th (not in office hours on Tuesday)
- Special lecture on Friday:
  - Patrick Ball (HRDAG) talking about linking files from NGOs on casualaties in the Syrian Civil War