NAHDO

Linking Workshop

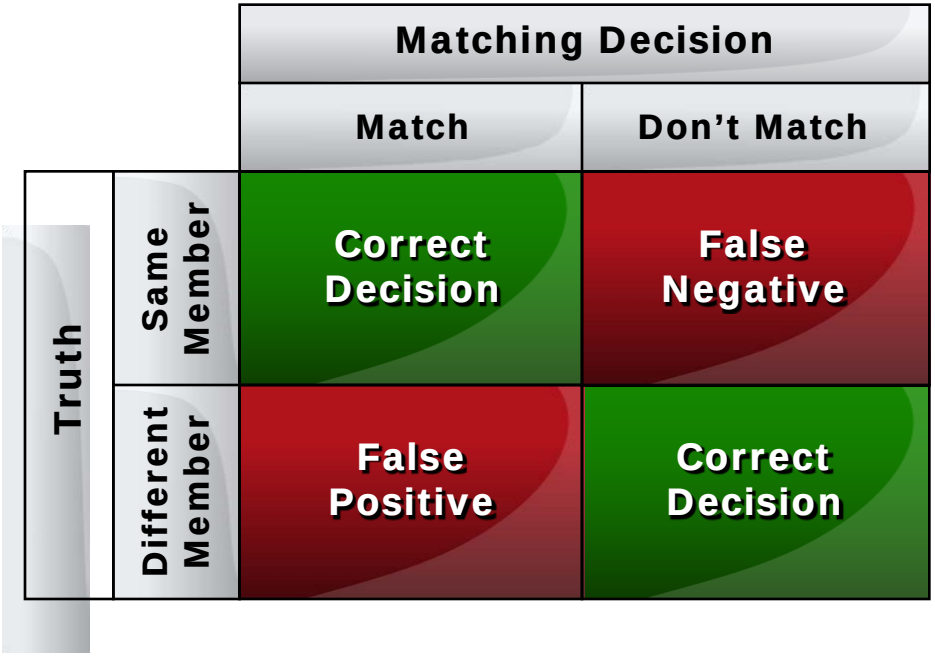
11 May 2011

Workshop Outline

- Overview
  - The linking problem
  - Classes of approaches
  - Variables in the linking problem and their effect
- Discussion
  - Linking examples from the audience
    - Approach used
    - Classification by variables
    - Experience
  - Best practices
    - What can we abstract from the above cases

The linking problem

- Determine whether two records in a database refer to the same person or not
- Decision based upon applying an algorithm to demographic data
  - Don’t assume the existence of a unique identifier
  - Standard attributes are name (first, middle, last, suffix), birth date, gender
  - Can include identifiers such as SSN, driver license number, passport number
  - Also may use location information such as address and phone numbers
- The standard type I and type II errors are referred to as false positives and false negatives



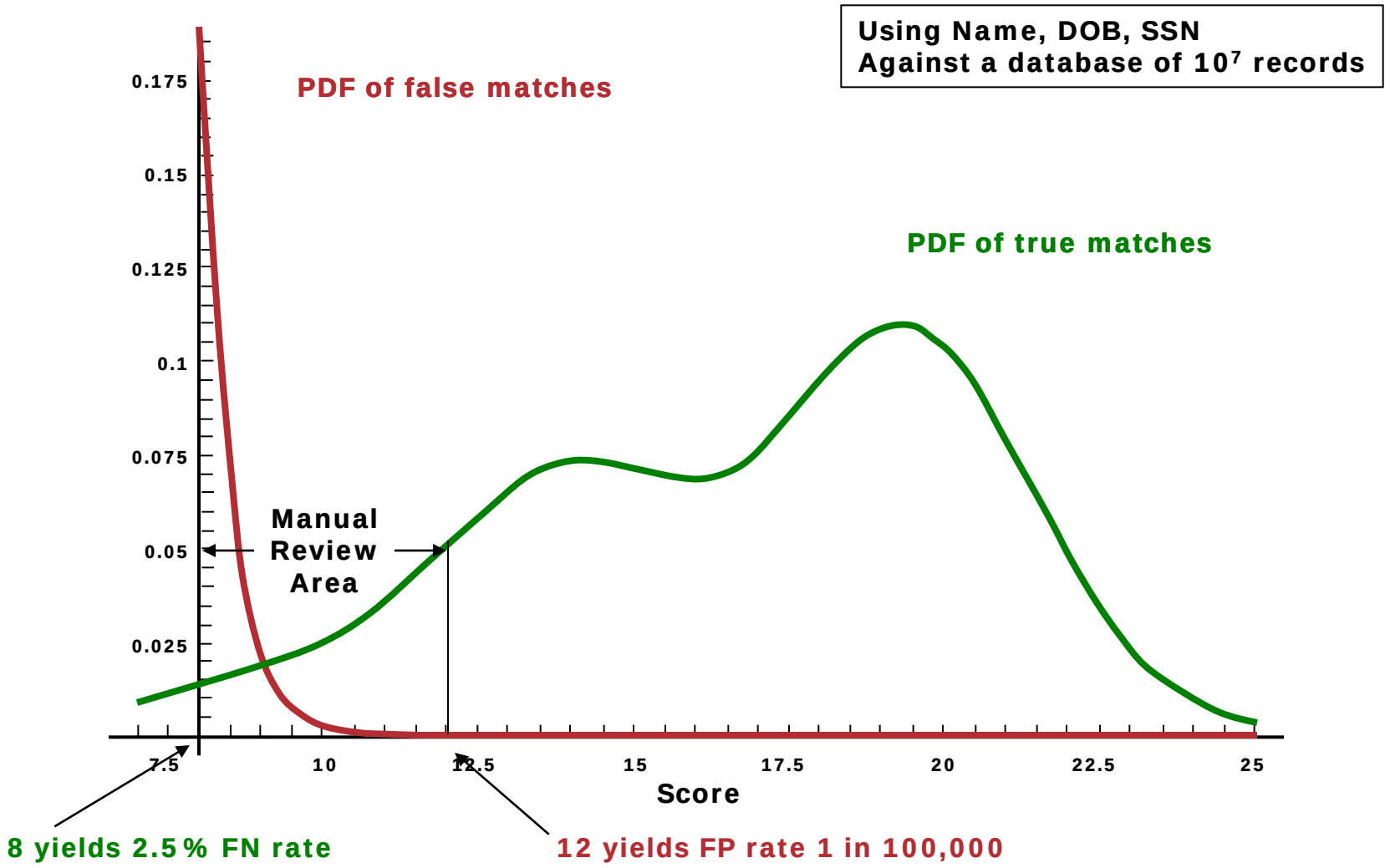
Classification of approaches

- Algorithm distinction is popularly referred to as probabilistic versus deterministic (which is bad nomenclature but we can't fix it here)
  - Their fundamental distinction is how they make the decision
  - Both approaches can incorporate nicknames, phonetic codes, and typographical distances
- Deterministic matching best illustrated by a truth-table approach
  - E.G. If the last names match, the first names match (either exactly or nickname), and the DOB is the same, then link … unless the SSNs are different
  - A variant of this is an ad-hoc weighting approach – a match on the last name is worth 70, an exact match on first name is 15 while a nickname match is 10 etc.

Classification of approaches (continued)

- Probabilistic matching uses a likelihood ratio based decision
  - If the last names match and the name is SMITH assign a 2, if the last names match and the name is EINSTEIN assign a 5, if the last names don't match, assign -2
  - These weights are based upon calculations of conditional probabilities
    - E.G. - Probability that last names agree on SMITH when the records refer to the same member divided by the probability that the last names agree on SMITH when the records don't refer to the same person
    - For historical consistency, most use the log-base 10 of this ratio
  - This is calculated for each attribute and then summed
  - The result is compared to a threshold (or thresholds) which determine the final decision

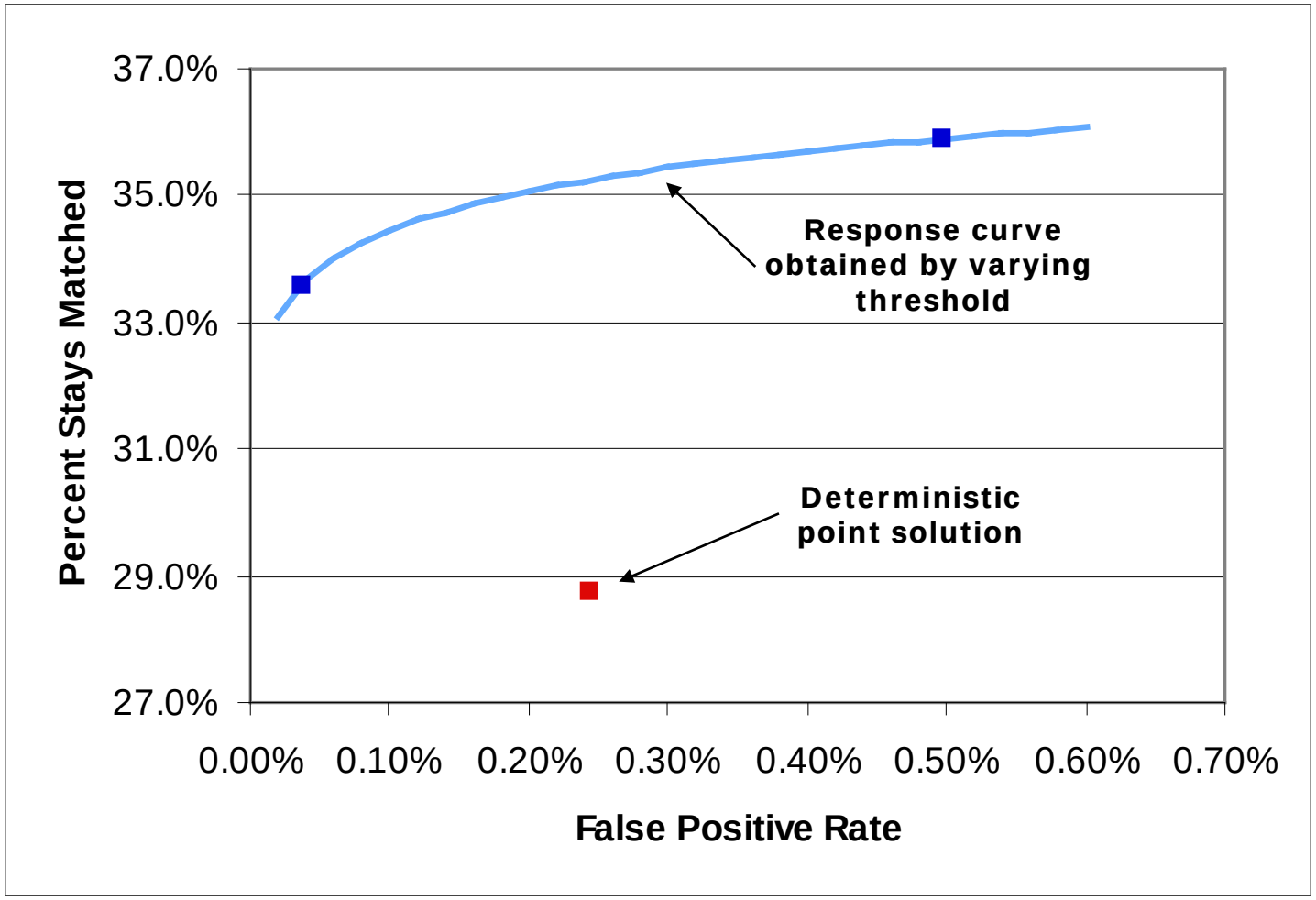
Dual threshold



Approach comparison

- Deterministic
  - Much simpler to implement, particularly if you are building it yourself
    - Although some of the ad-hoc weighting schemes can get complex
    - Complexity grows with the number of attributes and number of partial matches (e.g. nickname) allowed
  - Best suited to binary decision problems
  - Easier to explain results
  - Difficult to scale
- Probabilistic
  - Higher accuracy – theoretically sound decision method
    - Supports scaling as well
  - More complicated to build and instantiate
    - The utility which estimates the weights for each of the attributes is a non-trivial undertaking
  - Well suited to problems with many attributes and multiple partial match features
  - Adjustable threshold provides decision flexibility

Sample comparison



Variables – key variables to consider in selecting a linking approach

- Required false-positive rate – 1 in 10 thousand or 1 in 10 million
  - Smaller rates favor probabilistic
- Number of records - < 1 million or > 50 million
  - Higher volumes favor probabilistic for accuracy but favor deterministic for simplicity
- Number of attributes – 5 or 10
  - More of a complexity issue
- Partial matching
  - Combining multiple partial match techniques favors probabilistic
- Families or related records
  - High percentage of families can impede probabilistic
- Real-time or batch
  - Real-time requirements are easier to meet with deterministic
- Sparseness
  - Sparse data favor probabilistic approach

Example – impact of validity

- Analytical simulation of matching performance
  - Single threshold – low false-positive rate
  - Search against 10 million member database
- Four attributes - name, DOB, Zip, SSN
- Vary data validity
  - Fraction of the time an attribute is available
  - Full SSN or only the last 4-digits
- Simulate false-negative rate

Name	DOB	Zip	SSN	False-negative rate
100%	100%	100%	0%	6%
100%	90%	90%	0%	22%
100%	90%	90%	70%	7%
100%	90%	90%	70% (4 digits)	8%

Example - RxHub

Extremely low FP rate – 1 in 100 million	Probabilistic
Large record volume – 150 million	Probabilistic
Small number of attributes – 4	Deterministic
Partial matching on names, DOB, and ZIP	Probabilistic
Significant family population	Deterministic
Real-time	Deterministic

- Selected probabilistic
  - Actually coded up both for comparison
  - “Paid by the match” business model gave high weight to incorporating partial match logic on all attributes
  - Needed to demonstrate that the last two were tractable

Audience Cases