MEDD
An Approximate Matching Technology for Database Searching, Linking and De-Duplicating

Dr. Andrew Borthwick, President
Dr. Arthur Goldberg, VP Marketing and Strategy
## Same Child?

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Record</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Last name</td>
<td>Smith</td>
<td>Smith</td>
</tr>
<tr>
<td>First name</td>
<td>Emily</td>
<td>Emely</td>
</tr>
<tr>
<td>Soundex of First name</td>
<td>EML</td>
<td>EML</td>
</tr>
<tr>
<td>DOB</td>
<td>4/28/97</td>
<td>4/28/97</td>
</tr>
<tr>
<td>Street</td>
<td>4528 3rd Ave</td>
<td>4528 3rd Ave</td>
</tr>
<tr>
<td>City</td>
<td>Bronx</td>
<td>Bronx</td>
</tr>
<tr>
<td>State</td>
<td>NY</td>
<td>NY</td>
</tr>
<tr>
<td>Zip</td>
<td>10462</td>
<td>10462</td>
</tr>
<tr>
<td>Phone</td>
<td>718-123-4567</td>
<td>718-123-6789</td>
</tr>
<tr>
<td>Med Rec Number</td>
<td>11856437503</td>
<td>11856437503</td>
</tr>
<tr>
<td>Field Name</td>
<td>Record</td>
<td>Comparison</td>
</tr>
<tr>
<td>----------------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Last name</td>
<td>Lopez</td>
<td>Match</td>
</tr>
<tr>
<td>First name</td>
<td>Girl</td>
<td>Susan</td>
</tr>
<tr>
<td>Soundex of First name</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOB</td>
<td>1/11/97</td>
<td>1/2/97</td>
</tr>
<tr>
<td>Street</td>
<td>987 Cornelia</td>
<td>456 Park Pl</td>
</tr>
<tr>
<td>City</td>
<td>Brooklyn</td>
<td>Brooklyn</td>
</tr>
<tr>
<td>State</td>
<td>NY</td>
<td>NY</td>
</tr>
<tr>
<td>Zip</td>
<td>11211</td>
<td>11211</td>
</tr>
<tr>
<td>Phone</td>
<td>718-123-4567</td>
<td>718-234-5678</td>
</tr>
<tr>
<td>Med Rec Number</td>
<td>1001002</td>
<td>567435</td>
</tr>
</tbody>
</table>
## Matching Catastrophes

<table>
<thead>
<tr>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC Department of Health Child DB</td>
<td>1.4M children duplicated into 2.1M records</td>
</tr>
<tr>
<td>Removing felons from Florida’s voter roles</td>
<td>Some counties purged non-felons. Some counties did no purge because of list’s inaccuracies</td>
</tr>
<tr>
<td>Wall street business data</td>
<td>Two clerks work full time matching by hand</td>
</tr>
</tbody>
</table>
An Innovative Record Matching System

**Accurate**
- Error rate below 0.3%, rigorously tested on child database

**Customizable**
- Special comparisons easily added; machine learning tunes to the data

**Easily deployed**
- Trained with examples

**Fast**
- Match 1000s of records per second
Production Matching Basics

**Input** Search record

**Blocking**
- Find thousands of possible matches

**Match decision making**
- For each possible match
  - Evaluate many comparison functions against search record
  - Combine comparison functions by weight to produce match probability

**Output** IDs and probabilities of likely matches
Comparison Function Examples

- Do first names match?
- Do first names match approximately using "phonetic matches" such as Soundex, edit-distance, NYSIIS, or Jaro-Winkler?
- Do uncommon first names match?
- Do we have an indicator that the child is part of a multiple birth?
- Do Medicaid numbers match or mismatch?
- Do birthdays match?
Comparison Function Examples

Database of Businesses

- How many words in the name match?
- Could one name be an abbreviation for the other?
- Are the names the same after translating foreign words to English?
- Do country, phone number, or street address match?
Process for Deploying MEDD

1. Design
2. Train
3. Test
4. Accuracy Okay?
   - NO
   - YES
5. Production Matching
MEDD Production Matching

Search Record

Blocking

Many Possible Matches

Maximum Entropy Matching

Match Probabilities of Possible Matches

Non-Match

Low

Intermediate

High

Match

Match Probability

Human Review
Client Program Interacts with MEDD

Send search record to MEDD

Process returned list of possible matches

- One high probability match—link matched record to the search record
- More than one high probability matches—human review
- One or more medium probability matches—human review
- No likely matches—insert the search record
Maximum Entropy Matching Math

- The probability a pair of records match

\[
\frac{\text{MatchProduct}}{\text{MatchProduct} + \text{No-MatchProduct}}
\]

\text{MatchProduct} = \text{product of weights of all comparison functions predicting \textbf{Match} for the pair}

\text{No-MatchProduct} = \text{product of weights of all comparison functions predicting \textbf{No-Match} for the pair}
### MEDD Decides Match
**99.5% Confidence**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Record</th>
<th>Match?</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Last name</strong></td>
<td>Smith</td>
<td>Smith</td>
<td>Match</td>
</tr>
<tr>
<td><strong>First name</strong></td>
<td>Emily</td>
<td>Emely</td>
<td>No-match</td>
</tr>
<tr>
<td><strong>Soundex First name</strong></td>
<td>EML</td>
<td>EML</td>
<td>Match</td>
</tr>
<tr>
<td><strong>DOB</strong></td>
<td>4/28/97</td>
<td>4/28/97</td>
<td>Match</td>
</tr>
<tr>
<td><strong>Street</strong></td>
<td>4528 3rd Ave</td>
<td>4528 3rd Ave</td>
<td>Match</td>
</tr>
<tr>
<td><strong>City</strong></td>
<td>Bronx</td>
<td>Bronx</td>
<td>Match</td>
</tr>
<tr>
<td><strong>State</strong></td>
<td>NY</td>
<td>NY</td>
<td></td>
</tr>
<tr>
<td><strong>Zip</strong></td>
<td>10462</td>
<td>10462</td>
<td>Match</td>
</tr>
<tr>
<td><strong>Phone</strong></td>
<td>718-123-4567</td>
<td>718-123-6789</td>
<td>No-match</td>
</tr>
<tr>
<td><strong>Med Rec Number</strong></td>
<td>11856437503</td>
<td>11856437503</td>
<td>Match</td>
</tr>
</tbody>
</table>

\[
\text{Match product} = 587.2 \\
\text{No-match product} = 2.9 \\
\frac{587.2}{587.2 + 2.9} = 0.995
\]
**MEDD Decides No-match**

**97.9% Confidence**

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Record</th>
<th>Comparison</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Last name</td>
<td>Lopez</td>
<td>Lopez</td>
<td>Match</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.153</td>
</tr>
<tr>
<td>First name</td>
<td>Girl</td>
<td>Susan</td>
<td>No data</td>
</tr>
<tr>
<td>Soundex First name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOB</td>
<td>1/11/97</td>
<td>1/2/97</td>
<td>No-match</td>
</tr>
<tr>
<td>Street</td>
<td>987 Cornelia</td>
<td>456 Park</td>
<td>No-match</td>
</tr>
<tr>
<td>City</td>
<td>Brooklyn</td>
<td>Brooklyn</td>
<td>Match</td>
</tr>
<tr>
<td>State</td>
<td>NY</td>
<td>NY</td>
<td>1.103</td>
</tr>
<tr>
<td>Zip</td>
<td>11211</td>
<td>11211</td>
<td>Match</td>
</tr>
<tr>
<td>Phone</td>
<td>718-123-4567</td>
<td>718-234-5678</td>
<td>No-match</td>
</tr>
<tr>
<td>Med Rec Number</td>
<td>1001002</td>
<td>567435</td>
<td></td>
</tr>
</tbody>
</table>

MatchProduct = 3.8

No-matchProduct = 181.1

\[
\frac{3.8}{181.1 + 3.8} = 0.021
\]
Principles of Maximum Entropy

How are weights determined?

- Input record pairs marked **Match** or **No-match**
- Weights selected so model predicts average probability of match for each comparison function equal to average probability for that comparison function in training data

Name

Name matches and Records match
Name matches and Records differ

Probability records match given that name matches = 2/3

Phone

Phone matches and Records match
Phone matches and Records differ

Probability records match given that Phone matches = 7/9
Mathematics of Max. Ent. (cont.)

- There are an infinite number of models which satisfy these training-data conformance criteria.
- In M.E., we choose the model which has maximum entropy (greatest uncertainty).
  - “Account for all that is known, carefully avoid assuming anything that is not known.”
- It can be shown that the M.E. model will be of the form shown on the previous slide.
- Model of that form which assigns highest probability to training corpus is the M.E. model.
- The search space for M.E. model parameters is convex (no local minima).
MEDD Deployment Details

1. Design
2. Data Comparison Functions
3. Train
4. Data Comparison Functions with Weights
5. Test
6. Test Marked Record Pairs
7. NO
8. Accuracy Okay?
9. YES
10. MEDD Production Matching
11. Marked Record Pairs
Training/Testing Data Selection

Run blocking procedure with which the model will be deployed

Randomly select the desired number of record-pairs (at least 3,000) from the blocking output
- But oversample difficult cases (twins)

Hand-score the record-pairs
- Discard record-pairs human-marked as “unsure”

Divide the data into three equal parts
- Training data
- Development test data
- Held-back test data
Testing Procedure

Run system on record-pairs held back by the client, outputting a probability of match for each pair

A testing program searches for the point $c_1$ which divides the “match” records from the “unsure” records and the point $c_2$ which divides the “no-match” and “unsure” records

$c_1$ and $c_2$ are selected such that the error rate among the records declared to be match and no-match are within the user’s error tolerance $t_1$ and $t_2$ respectively

We report the human review percentage (number of records between $c_1$ and $c_2$ /Total number of recs) at the given error tolerance
Matching with MEDD

Threshold Probabilities
Tradeoff Human Review Against Accuracy

Records Needing Review

Match Accuracy

%
Accuracy On NYC Children

932 test record pairs

Number record pairs

MEDD's Probability of Match

MATCHES
DIFFERS
MEDD’s Ongoing Training and Design

Production Matching

Match Probability

Intermediate

Human Review

New Production Matching

Old Marked Pairs

Design (optional)

Train

Test

Accuracy Okay?

YES

NO

More Marked Record Pairs
Designing a MEDD Deployment

- Study database
- Select matching fields
- Select derived field functions
- Extend or replicate DB with derived fields
- Add indices on derived and matching fields
- Select and/or design comparison functions
Complex Comparison Functions

Adapt to database quirks

Child medical database example

HMO XYZ sends Day of Birth = “1”
Birthday = July 1, 1998 not July 15, 1998

A special comparison function
IF Provider = “HMO XYZ”
AND Day of Birth = 1
AND dates differs only on day of birth
THEN Match
Technical Information

Platforms
- Win32, Linux, Solaris, and other UNIX

Modes of operations
- Online as a CORBA Module
- Batch mode with a flat file input

Written in ANSI C++

Available for Oracle, other DB’s to follow
Database Interaction Alternatives

**MEDD Server**
- Replicate matching data inside MEDD
- Replication frequency determined by need for current data in matching
- Simpler deployment; more expensive computationally and financially

**MEDD Library**
- Existing database modified with additional matching fields and indices
- MEDD interacts with modified database
- More complex deployment; saves disk space and probably money
# ChoiceMaker Clients

## Solved and In Progress

<table>
<thead>
<tr>
<th>Client</th>
<th>Problem</th>
<th>Number of Records</th>
<th>Number of Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC Dept. of Health</td>
<td>Duplicates</td>
<td>2.1 M</td>
<td>50%</td>
</tr>
<tr>
<td>NYC Dept. of Health</td>
<td>Match 2 medical databases</td>
<td>1.7 M</td>
<td></td>
</tr>
<tr>
<td>Wall St. data provider</td>
<td>Match 50 databases</td>
<td>150K</td>
<td>About 30K</td>
</tr>
</tbody>
</table>
ChoiceMaker

Management

- Andrew Borthwick, President
  - Designed and implemented MEDD
  - NYU CS PhD 1999
  - Expert on maximum entropy modeling
- Arthur Goldberg, VP Strategy and Marketing
  - NYU CS Professor, co-director MSIS graduate program
  - Expert on network performance
  - Five years at IBM Research

Funding

- NSF Small Business Innovation Research Grant
- Investment from CCS, a $120M annual revenue Japanese software firm
Advantages of MEDD vs. Fillegi-Sunter

- MEDD outputs probabilities
  - Easier to interpret than log-likelihood ratio
- Probabilities are derived from human intuition
- No assumption of statistical independence
- Arbitrary features can be cleanly integrated into the model
Questions

Arthur.Goldberg@choicemaker.com
Andrew.Borthwick@choicemaker.com
212 905-6031
ChoiceMaker Technologies, Inc.
41 East 11th Street, 11th Floor
New York, NY 10003