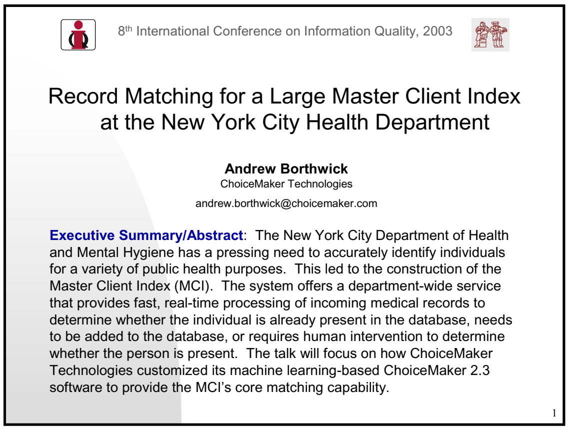


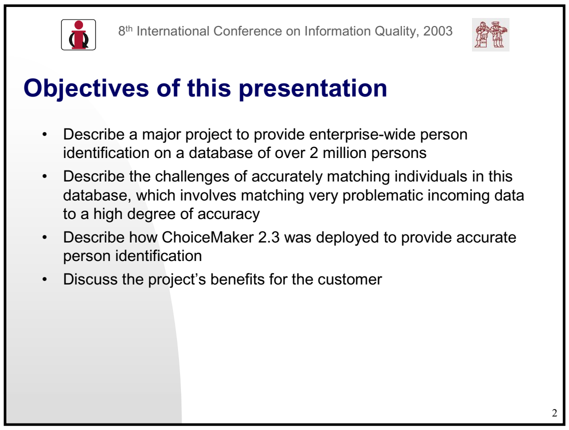
8th International Conference on Information Quality, 2003

Record Matching for a Large Master Client Index at the New York City Health Department

Andrew Borthwick
ChoiceMaker Technologies
andrew.borthwick@choicemaker.com

Executive Summary/Abstract: The New York City Department of Health and Mental Hygiene has a pressing need to accurately identify individuals for a variety of public health purposes. This led to the construction of the Master Client Index (MCI). The system offers a department-wide service that provides fast, real-time processing of incoming medical records to determine whether the individual is already present in the database, needs to be added to the database, or requires human intervention to determine whether the person is present. The talk will focus on how ChoiceMaker Technologies customized its machine learning-based ChoiceMaker 2.3 software to provide the MCI's core matching capability.

1

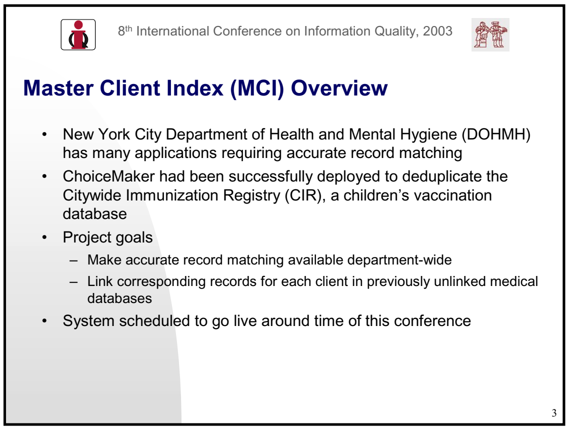


8th International Conference on Information Quality, 2003

Objectives of this presentation

- Describe a major project to provide enterprise-wide person identification on a database of over 2 million persons
- Describe the challenges of accurately matching individuals in this database, which involves matching very problematic incoming data to a high degree of accuracy
- Describe how ChoiceMaker 2.3 was deployed to provide accurate person identification
- Discuss the project's benefits for the customer

2

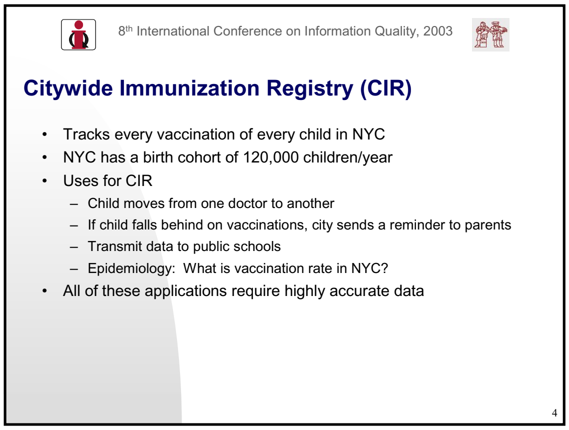


8th International Conference on Information Quality, 2003

Master Client Index (MCI) Overview

- New York City Department of Health and Mental Hygiene (DOHMH) has many applications requiring accurate record matching
- ChoiceMaker had been successfully deployed to deduplicate the Citywide Immunization Registry (CIR), a children's vaccination database
- Project goals
  - Make accurate record matching available department-wide
  - Link corresponding records for each client in previously unlinked medical databases
- System scheduled to go live around time of this conference

3

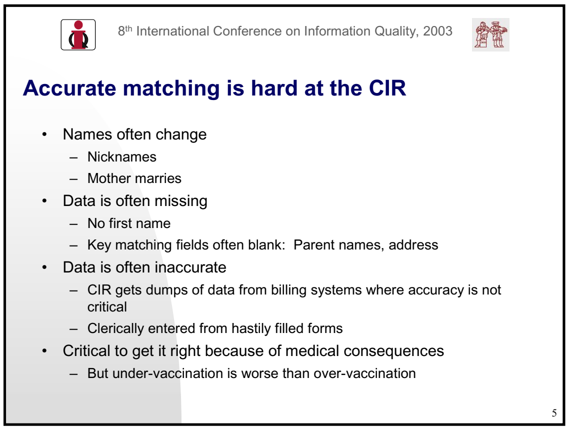


8th International Conference on Information Quality, 2003

Citywide Immunization Registry (CIR)

- Tracks every vaccination of every child in NYC
- NYC has a birth cohort of 120,000 children/year
- Uses for CIR
  - Child moves from one doctor to another
  - If child falls behind on vaccinations, city sends a reminder to parents
  - Transmit data to public schools
  - Epidemiology: What is vaccination rate in NYC?
- All of these applications require highly accurate data

4

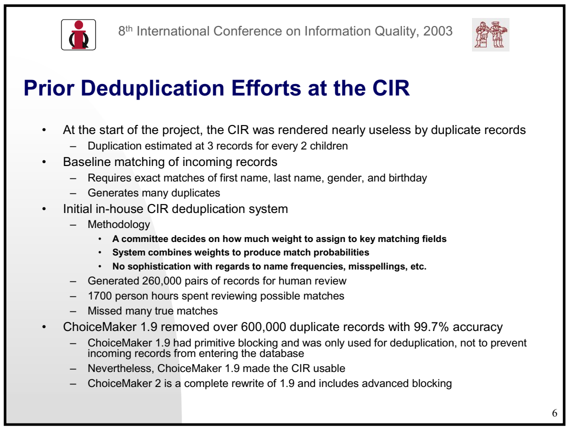


8th International Conference on Information Quality, 2003

Accurate matching is hard at the CIR

- Names often change
  - Nicknames
  - Mother marries
- Data is often missing
  - No first name
  - Key matching fields often blank: Parent names, address
- Data is often inaccurate
  - CIR gets dumps of data from billing systems where accuracy is not critical
  - Clerically entered from hastily filled forms
- Critical to get it right because of medical consequences
  - But under-vaccination is worse than over-vaccination

5

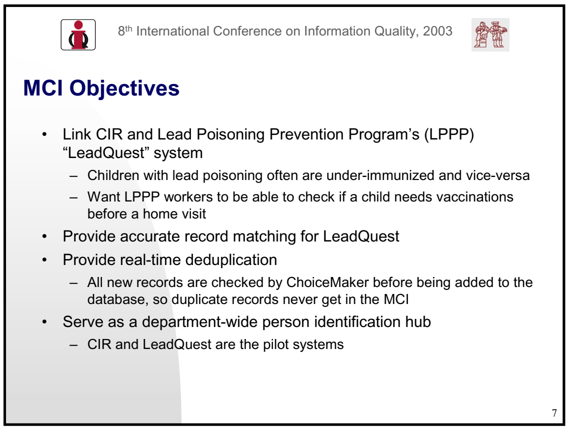


8th International Conference on Information Quality, 2003

Prior Deduplication Efforts at the CIR

- At the start of the project, the CIR was rendered nearly useless by duplicate records
  - Duplication estimated at 3 records for every 2 children
- Baseline matching of incoming records
  - Requires exact matches of first name, last name, gender, and birthday
  - Generates many duplicates
- Initial in-house CIR deduplication system
  - Methodology
    - **A committee decides on how much weight to assign to key matching fields**
    - **System combines weights to produce match probabilities**
    - **No sophistication with regards to name frequencies, misspellings, etc.**
  - Generated 260,000 pairs of records for human review
  - 1700 person hours spent reviewing possible matches
  - Missed many true matches
- ChoiceMaker 1.9 removed over 600,000 duplicate records with 99.7% accuracy
  - ChoiceMaker 1.9 had primitive blocking and was only used for deduplication, not to prevent incoming records from entering the database
  - Nevertheless, ChoiceMaker 1.9 made the CIR usable
  - ChoiceMaker 2 is a complete rewrite of 1.9 and includes advanced blocking

6

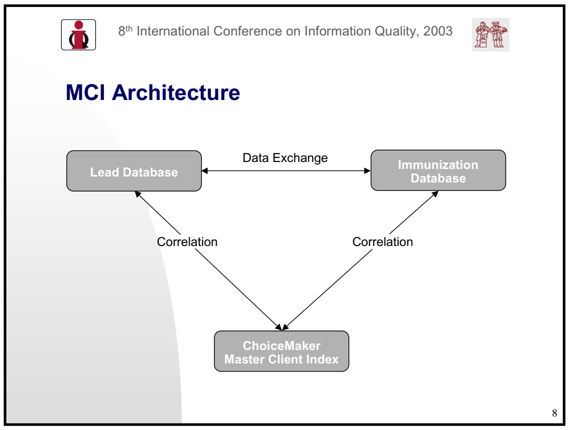


8th International Conference on Information Quality, 2003

MCI Objectives

- Link CIR and Lead Poisoning Prevention Program's (LPPP) "LeadQuest" system
  - Children with lead poisoning often are under-immunized and vice-versa
  - Want LPPP workers to be able to check if a child needs vaccinations before a home visit
- Provide accurate record matching for LeadQuest
- Provide real-time deduplication
  - All new records are checked by ChoiceMaker before being added to the database, so duplicate records never get in the MCI
- Serve as a department-wide person identification hub
  - CIR and LeadQuest are the pilot systems

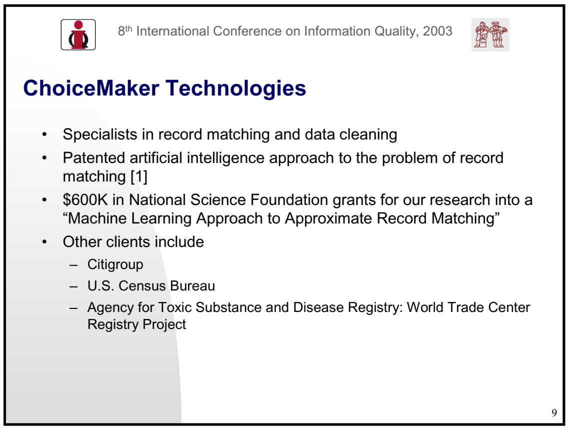
7



8th International Conference on Information Quality, 2003

MCI Architecture

8

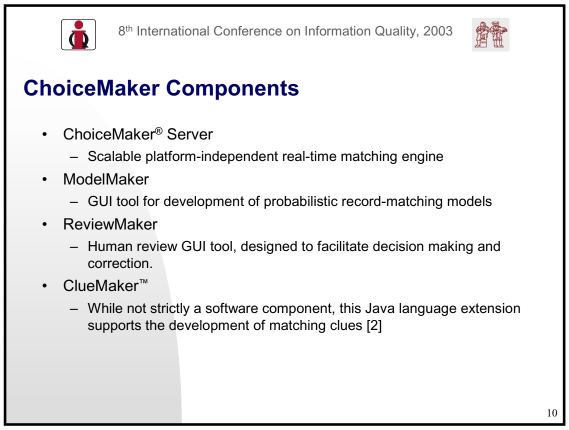


8th International Conference on Information Quality, 2003

ChoiceMaker Technologies

- Specialists in record matching and data cleaning
- Patented artificial intelligence approach to the problem of record matching [1]
- $600K in National Science Foundation grants for our research into a "Machine Learning Approach to Approximate Record Matching"
- Other clients include
  - Citigroup
  - U.S. Census Bureau
  - Agency for Toxic Substance and Disease Registry: World Trade Center Registry Project

9

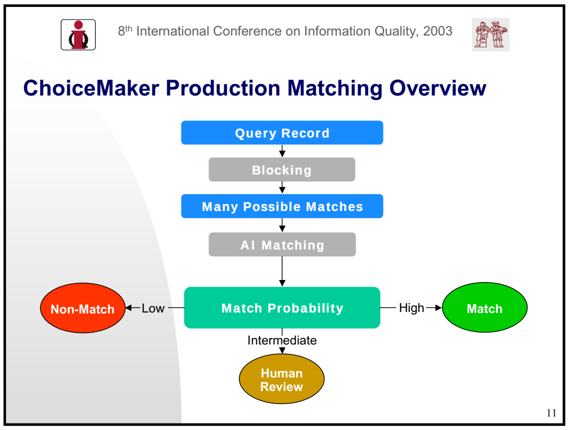


8th International Conference on Information Quality, 2003

ChoiceMaker Components

- ChoiceMaker® Server
  - Scalable platform-independent real-time matching engine
- ModelMaker
  - GUI tool for development of probabilistic record-matching models
- ReviewMaker
  - Human review GUI tool, designed to facilitate decision making and correction.
- ClueMaker™
  - While not strictly a software component, this Java language extension supports the development of matching clues [2]

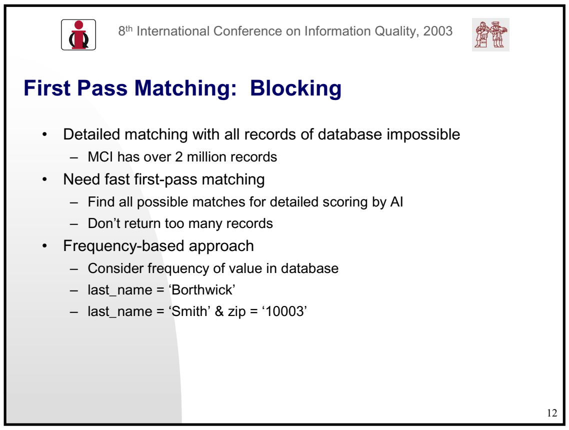
10



8th International Conference on Information Quality, 2003

ChoiceMaker Production Matching Overview

11

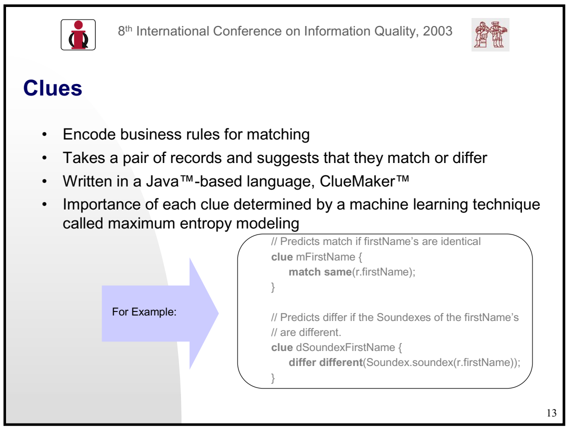


8th International Conference on Information Quality, 2003

First Pass Matching: Blocking

- Detailed matching with all records of database impossible
  - MCI has over 2 million records
- Need fast first-pass matching
  - Find all possible matches for detailed scoring by AI
  - Don't return too many records
- Frequency-based approach
  - Consider frequency of value in database
  - last_name = 'Borthwick'
  - last_name = 'Smith' & zip = '10003'

12



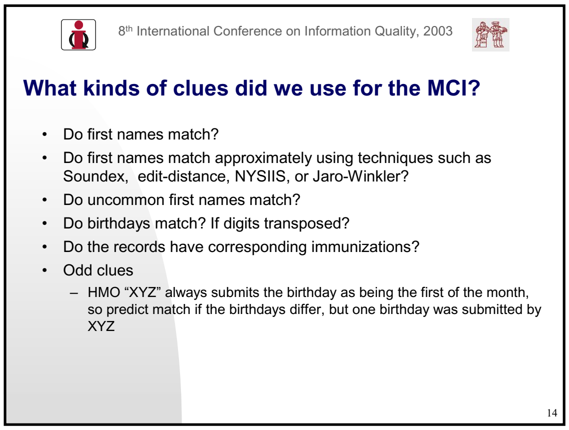
Clues

- Encode business rules for matching
- Takes a pair of records and suggests that they match or differ
- Written in a Java™-based language, ClueMaker™
- Importance of each clue determined by a machine learning technique called maximum entropy modeling

For Example:

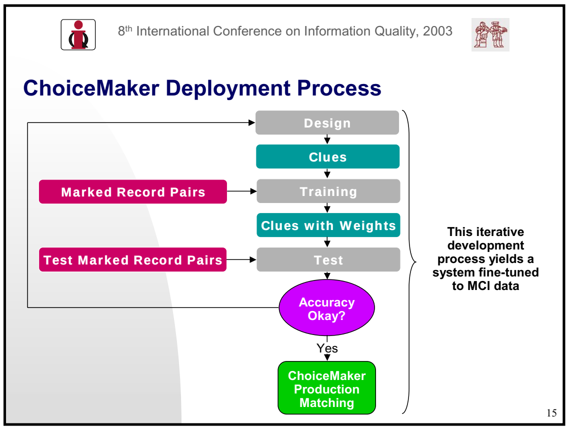
```
// Predicts match if firstName's are identical
clue mFirstName {
    match same(r.firstName);
}

// Predicts differ if the Soundexes of the firstName's
// are different.
clue dSoundexFirstName {
    differ different(Soundex.soundex(r.firstName));
}
```



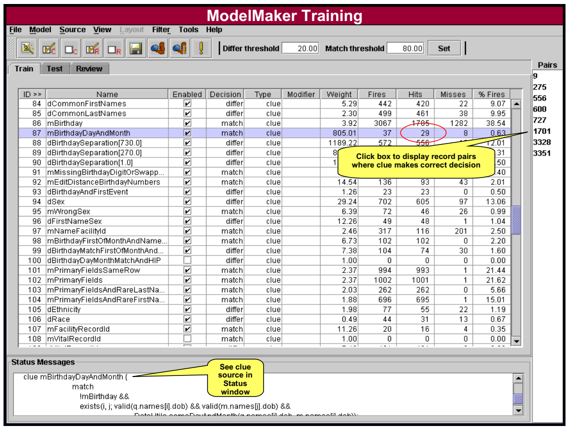
What kinds of clues did we use for the MCI?

- Do first names match?
- Do first names match approximately using techniques such as Soundex, edit-distance, NYSIIS, or Jaro-Winkler?
- Do uncommon first names match?
- Do birthdays match? If digits transposed?
- Do the records have corresponding immunizations?
- Odd clues
  - HMO "XYZ" always submits the birthday as being the first of the month, so predict match if the birthdays differ, but one birthday was submitted by XYZ



ChoiceMaker Deployment Process

This iterative development process yields a system fine-tuned to MCI data



ModelMaker Training

Differ threshold 20.00 Match threshold 80.00

ID >>	Name	Enabled	Decision	Type	Modifier	Weight	Fires	Hits	Misses	% Fires
84	dCommonFirstNames	☑	differ	clue		5.29	442	420	22	9.07
85	dCommonLastNames	☑	differ	clue		2.30	499	461	38	9.95
86	mBirthday	☑	match	clue		3.92	3067	1785	1282	38.54
87	mBirthdayDayAndMonth	☑	match	clue		805.01	37	29	8	0.63
88	dBirthdaySeparation[730.0]	☑	differ	clue		1189.22	572	556	[illegible]	12.01
89	dBirthdaySeparation[270.0]	☑	differ	clue		8[illegible]	[illegible]	[illegible]	[illegible]	[illegible].31
90	dBirthdaySeparation[1.0]	☑	differ	clue		1[illegible]	[illegible]	[illegible]	[illegible]	[illegible].50
91	mMissingBirthdayDigitOrSwapp...	☑	match	clue		[illegible]	[illegible]	[illegible]	[illegible]	[illegible].40
92	mEditDistanceBirthdayNumbers	☑	match	clue		14.54	136	93	43	2.01
93	dBirthdayAndFirstEvent	☑	differ	clue		1.26	23	23	0	0.50
94	dSex	☑	differ	clue		29.24	702	605	97	13.06
95	mWrongSex	☑	match	clue		6.39	72	46	26	0.99
96	dFirstNameSex	☑	differ	clue		12.26	49	48	1	1.04
97	mNameFacilityId	☑	match	clue		2.46	317	116	201	2.50
98	mBirthdayFirstOfMonthAndName...	☑	match	clue		6.73	102	102	0	2.20
99	dBirthdayMatchFirstOfMonthAnd...	☑	differ	clue		7.38	104	74	30	1.60
100	dBirthdayDayMonthMatchAndHIP	☐	differ	clue		1.00	0	0	0	0.00
101	mPrimaryFieldsSameRow	☑	match	clue		2.37	994	993	1	21.44
102	mPrimaryFields	☑	match	clue		2.37	1002	1001	1	21.62
103	mPrimaryFieldsAndRareLastNa...	☑	match	clue		2.03	262	262	0	5.66
104	mPrimaryFieldsAndRareFirstNa...	☑	match	clue		1.88	696	695	1	15.01
105	dEthnicity	☑	differ	clue		1.98	77	55	22	1.19
106	dRace	☑	differ	clue		0.49	44	31	13	0.67
107	mFacilityRecordId	☑	match	clue		11.26	20	16	4	0.35
108	mVitalRecordId	☐	match	clue		1.00	0	0	0	0.00

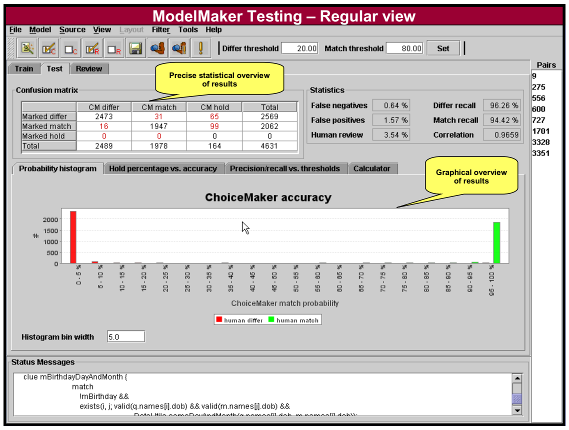
Pairs: 9, 275, 556, 600, 727, 1701, 3328, 3351

Click box to display record pairs where clue makes correct decision

Status Messages

```
clue mBirthdayDayAndMonth {
        match
          !mBirthday &&
          exists(i, j; valid(q.names[i].dob) && valid(m.names[j].dob) &&
                 DateUtils.sameDayAndMonth(q.names[i].dob, m.names[j].dob));
```

See clue source in Status window



ModelMaker Testing – Regular view

Differ threshold 20.00 Match threshold 80.00

Precise statistical overview of results

Confusion matrix

	CM differ	CM match	CM hold	Total
Marked differ	2473	31	65	2569
Marked match	16	1947	99	2062
Marked hold	0	0	0	0
Total	2489	1978	164	4631

Statistics

False negatives	0.64 %	Differ recall	96.26 %
False positives	1.57 %	Match recall	94.42 %
Human review	3.54 %	Correlation	0.9659

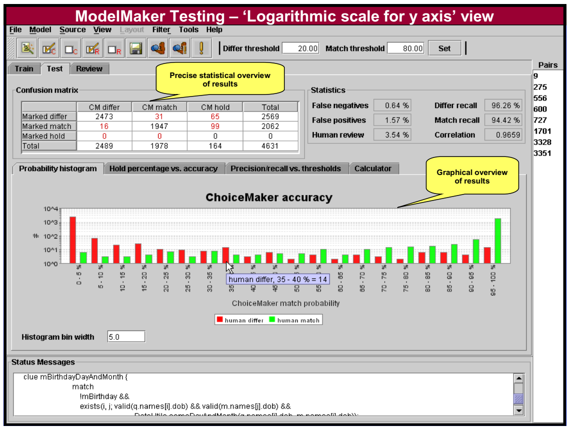
Pairs: 9, 275, 556, 600, 727, 1701, 3328, 3351

Graphical overview of results

Histogram bin width 5.0

Status Messages

```
clue mBirthdayDayAndMonth {
        match
          !mBirthday &&
          exists(i, j; valid(q.names[i].dob) && valid(m.names[j].dob) &&
                 DateUtils.sameDayAndMonth(q.names[i].dob, m.names[j].dob));
```



ModelMaker Testing – 'Logarithmic scale for y axis' view

Differ threshold 20.00 Match threshold 80.00

Precise statistical overview of results

Confusion matrix

	CM differ	CM match	CM hold	Total
Marked differ	2473	31	65	2569
Marked match	16	1947	99	2062
Marked hold	0	0	0	0
Total	2489	1978	164	4631

Statistics

False negatives	0.64 %	Differ recall	96.26 %
False positives	1.57 %	Match recall	94.42 %
Human review	3.54 %	Correlation	0.9659

Pairs: 9, 275, 556, 600, 727, 1701, 3328, 3351

Graphical overview of results

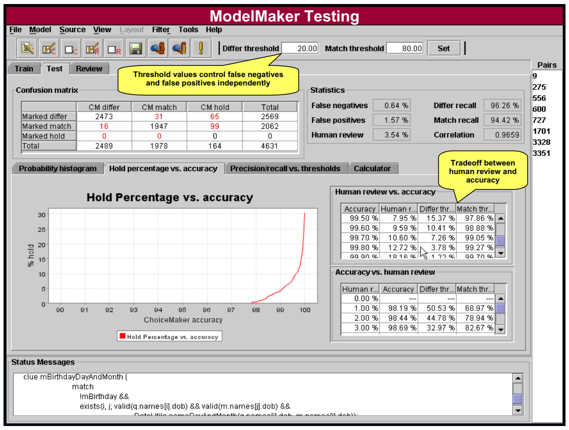
Histogram bin width 5.0

Status Messages

```
clue mBirthdayDayAndMonth {
        match
          !mBirthday &&
          exists(i, j; valid(q.names[i].dob) && valid(m.names[j].dob) &&
                 DateUtils.sameDayAndMonth(q.names[i].dob, m.names[j].dob));
```

Hold Percentage vs. Accuracy

- Tradeoff between work and accuracy is critical
- Judge a model by the degree of tradeoff required
- Graph shows hold percentage (work) as a function of required accuracy
- Human review vs. accuracy chart is a critical measure of success
  - ChoiceMaker often judges the quality of a model based on this chart
  - E.g. if Human Review Percentage at 99.5% accuracy goes down, the model is better



ModelMaker Testing

Differ threshold 20.00 Match threshold 80.00

Threshold values control false negatives and false positives independently

Confusion matrix

	CM differ	CM match	CM hold	Total
Marked differ	2473	31	65	2569
Marked match	16	1947	99	2062
Marked hold	0	0	0	0
Total	2489	1978	164	4631

Statistics

False negatives	0.64 %	Differ recall	96.26 %
False positives	1.57 %	Match recall	94.42 %
Human review	3.54 %	Correlation	0.9659

Pairs
9
275
556
600
727
1701
3328
3351

Tradeoff between human review and accuracy

Human review vs. accuracy

Accuracy	Human r...	Differ thr...	Match thr...
99.50 %	7.95 %	15.37 %	97.86 %
99.60 %	9.59 %	10.41 %	98.88 %
99.70 %	10.60 %	7.26 %	99.05 %
99.80 %	12.72 %	3.78 %	99.27 %
99.90 %	18.16 %	1.22 %	99.70 %

Accuracy vs. human review

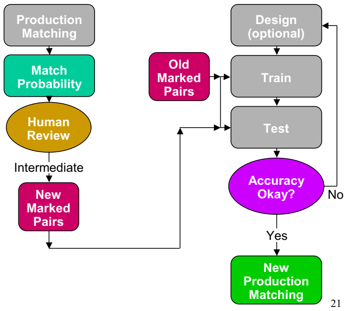
Human r...	Accuracy	Differ thr...	Match thr...
0.00 %	---	---	---
1.00 %	98.19 %	50.53 %	68.97 %
2.00 %	98.44 %	44.78 %	78.94 %
3.00 %	98.69 %	32.97 %	82.67 %

Status Messages

```
clue mBirthdayDayAndMonth {
            match
              !mBirthday &&
              exists(i, j; valid(q.names[i].dob) && valid(m.names[j].dob) &&
                          DateUtils.sameDayAndMonth(q.names[i].dob, m.names[j].dob));
```

Ongoing Training and Design Process

- •MCI will get more accurate over time
- •Records reviewed by DOHMH in production are fed back into ChoiceMaker training process
- •ChoiceMaker Technologies retrains the system with the augmented training data and possibly with additional clues



MCI Deployment Process

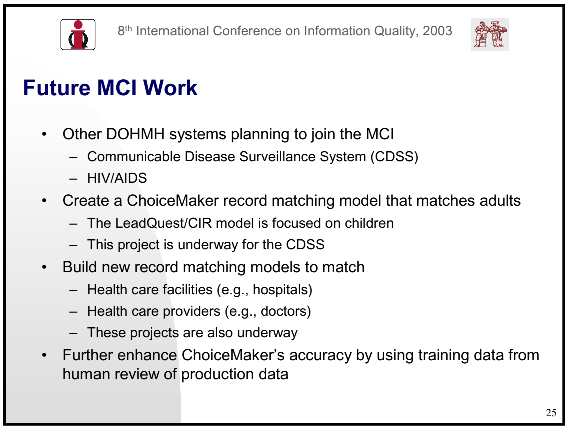
- Records from CIR are loaded into the MCI one at a time
  - MCI detects duplicates in CIR during this process
- LeadQuest records then loaded one at a time into MCI
  - MCI detects duplicates and links with CIR
- Client systems are synced with MCI by merging records detected as duplicate so that no two records from a single client system have the same MCI ID
- Records marked hold by ChoiceMaker are queued for human review

Problem Resolution Example

- Test load of the system revealed the need for the DOHMH to revise and clarify its definition of a “hold” record
  - Initial concern that ChoiceMaker might erroneously label records “match” that should be “hold” led to thresholds being set too low and too many records sent to human review
- Action plan
  - DOH labels additional records
  - ChoiceMaker deploys new advanced address parser, writes some new clues
  - Model retrained on new data and new clues
  - Final results to be presented at the conference
- Record matching is an iterative process. Tools must be able to diagnose problems and permit speedy problem resolution

MCI Expected Results

- DOHMH will have a strong data integrity hub
  - For instance, in a test run ChoiceMaker identified over 35,000 duplicate records from a birth cohort (of about 120,000) that had not been previously deduplicated
  - Tests show ChoiceMaker maintaining over 99.8% accuracy on records labeled “match” or “differ”
- Improved data integrity will enable better data sharing of immunizations with NYC schools and physicians

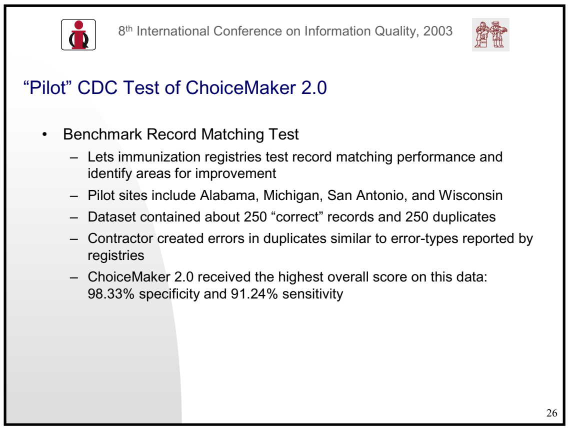


8th International Conference on Information Quality, 2003

Future MCI Work

- Other DOHMH systems planning to join the MCI
  - Communicable Disease Surveillance System (CDSS)
  - HIV/AIDS
- Create a ChoiceMaker record matching model that matches adults
  - The LeadQuest/CIR model is focused on children
  - This project is underway for the CDSS
- Build new record matching models to match
  - Health care facilities (e.g., hospitals)
  - Health care providers (e.g., doctors)
  - These projects are also underway
- Further enhance ChoiceMaker's accuracy by using training data from human review of production data

25

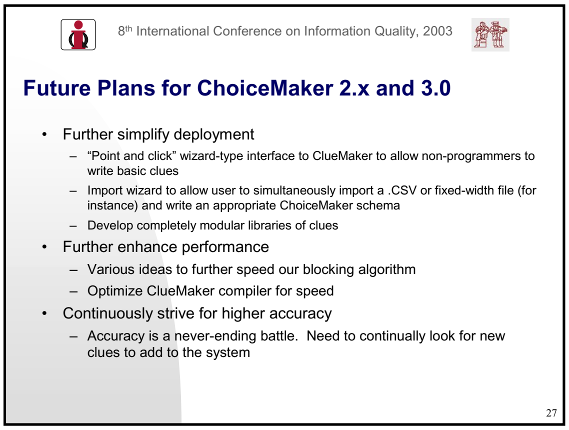


8th International Conference on Information Quality, 2003

"Pilot" CDC Test of ChoiceMaker 2.0

- Benchmark Record Matching Test
  - Lets immunization registries test record matching performance and identify areas for improvement
  - Pilot sites include Alabama, Michigan, San Antonio, and Wisconsin
  - Dataset contained about 250 "correct" records and 250 duplicates
  - Contractor created errors in duplicates similar to error-types reported by registries
  - ChoiceMaker 2.0 received the highest overall score on this data: 98.33% specificity and 91.24% sensitivity

26

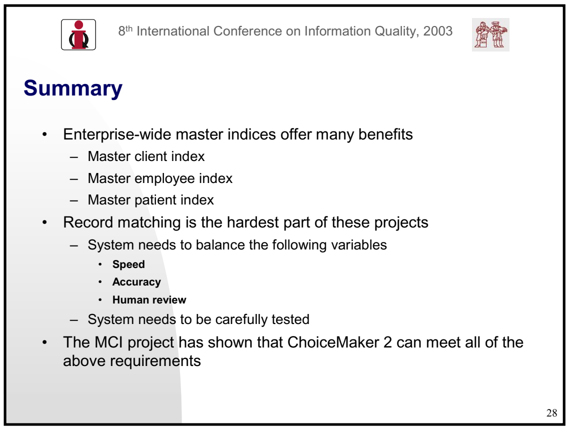


8th International Conference on Information Quality, 2003

Future Plans for ChoiceMaker 2.x and 3.0

- Further simplify deployment
  - "Point and click" wizard-type interface to ClueMaker to allow non-programmers to write basic clues
  - Import wizard to allow user to simultaneously import a .CSV or fixed-width file (for instance) and write an appropriate ChoiceMaker schema
  - Develop completely modular libraries of clues
- Further enhance performance
  - Various ideas to further speed our blocking algorithm
  - Optimize ClueMaker compiler for speed
- Continuously strive for higher accuracy
  - Accuracy is a never-ending battle. Need to continually look for new clues to add to the system

27

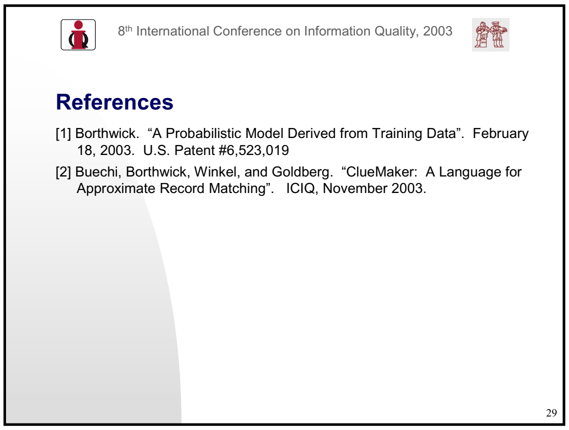


8th International Conference on Information Quality, 2003

Summary

- Enterprise-wide master indices offer many benefits
  - Master client index
  - Master employee index
  - Master patient index
- Record matching is the hardest part of these projects
  - System needs to balance the following variables
    - **Speed**
    - **Accuracy**
    - **Human review**
  - System needs to be carefully tested
- The MCI project has shown that ChoiceMaker 2 can meet all of the above requirements

28



8th International Conference on Information Quality, 2003

References

[1] Borthwick. "A Probabilistic Model Derived from Training Data". February 18, 2003. U.S. Patent #6,523,019

[2] Buechi, Borthwick, Winkel, and Goldberg. "ClueMaker: A Language for Approximate Record Matching". ICIQ, November 2003.

29