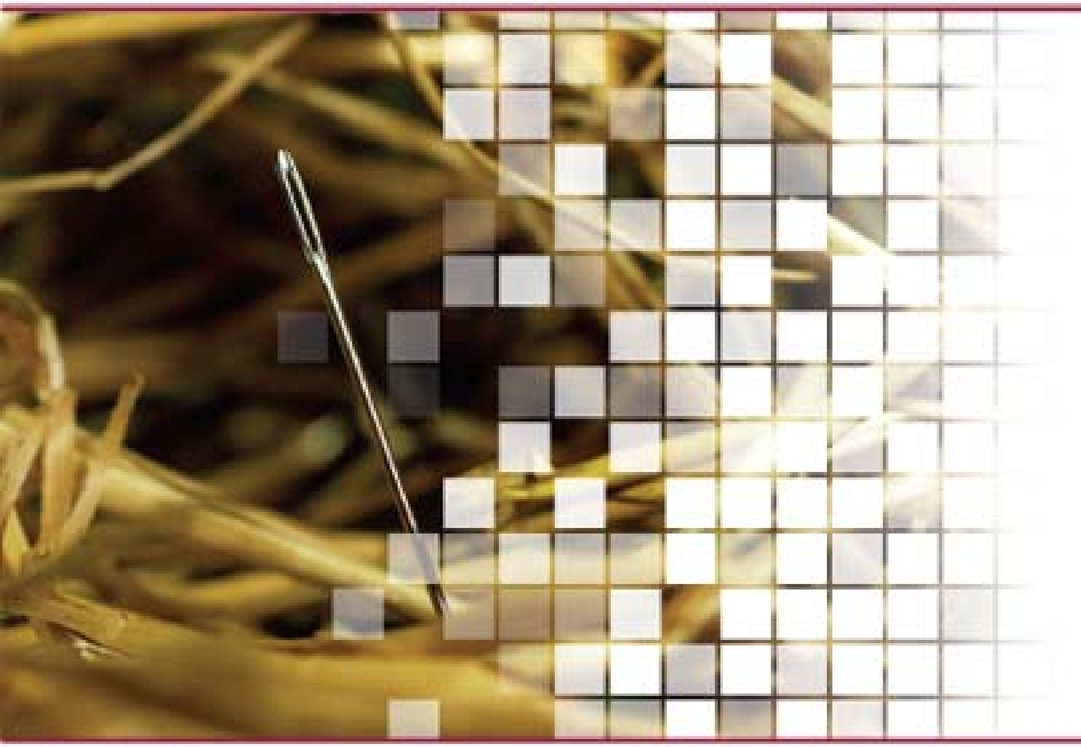
saffron TECHNOLOGY



Entity Resolution within Memory-based Analytics

Manny Aparicio
Co-founder and CEO
Saffron Technology, Inc.

Abstract

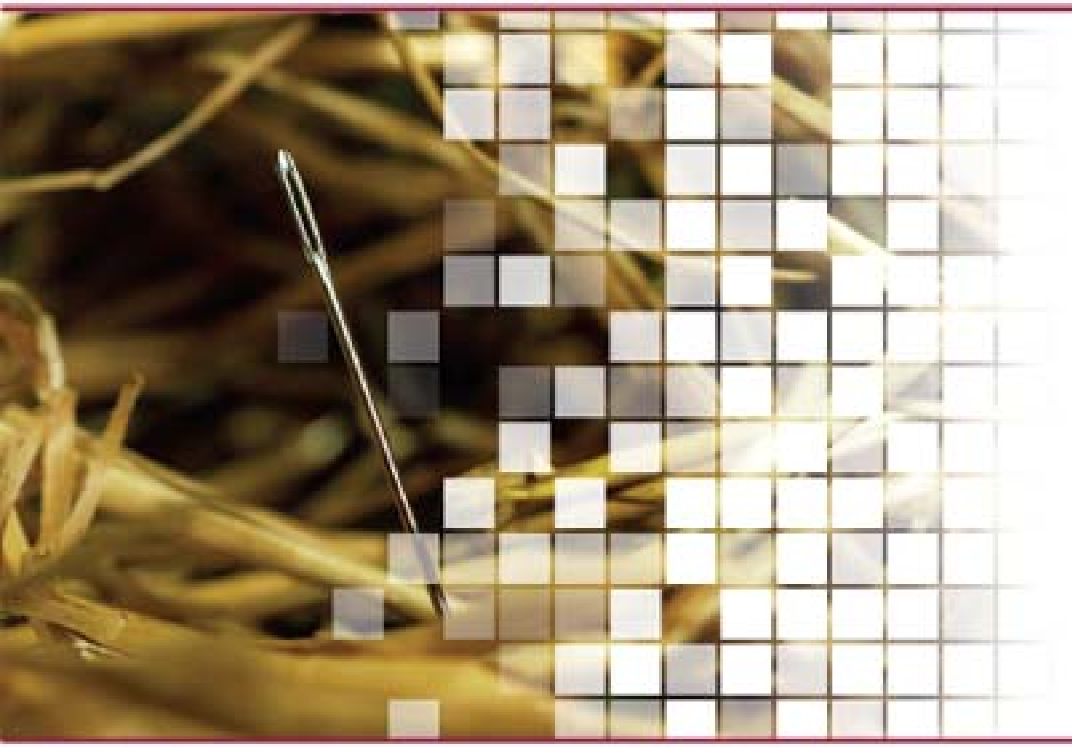
- Entity resolution has been defined for various applications and methods. A review of these definitions will include two scenarios. One scenario covers various needs, from general data cleansing to alias and group detection, when there is no particular entity of interest. A new approach, based on associative memories, will be introduced. Case-studies from national security, healthcare, and transactional integrity will be presented to prove the unprecedented accuracy of this approach for this type of entity resolution. Beyond these traditionally data-based and batch-oriented applications as one type of scenario, entity resolution in unstructured data sources, including a need for the perpetual resolution of analytically targeted entities, continues to be more challenging. Current problems and solutions to more emerging needs will be presented, A product demonstration for analytic discovery in text-based sources will focus on the quality of entity extraction and analytic reporting. Using advanced memory-based reasoning, features will include the correction of problems with text analytics as well as the more complete recall of similar and related entities from massive data stores. The first scenario addresses the integrity of the backend data store. The second scenario addresses "last foot" issues during front-end exploitation by the analyst. Given both of these scenarios for for entity resolution, the quality of data as well as the quality of reporting on such data are improved.

Biography

- Dr. Manuel Aparicio is the cofounder and CEO of Saffron Technology, the innovation leader in entity (and predictive) analytics. He leads Saffron's overall corporate vision and strategy for this disruptive technology, especially for national security in the US and allied countries. He is also growing the company to address similar critical problems in the finance and healthcare industries. Before founding Saffron, he was Chief Scientist of the IBM Knowledge Management and Intelligent Agent Center, coordinating IBM worldwide assets across all research and development labs, also working with advanced customers across several industries such as telecommunications and manufacturing, including agent applications within automotive and ship building consortia. He has over twenty years of experience in machine learning and over ten years of experience in the commercialization and industrial development of intelligent agents, including IBM’s first commercial rules-based agent in 1993 and the world's first commercial agent-based associative memory in 1997. He served on the boards of international organizations such as The Agent Society and The Foundation for Intelligent Physical Agents, in which he helped reinvigorate North America’s defense and commercial activity and established the standard now used by several defense and commercial products. He holds several patents in neural networks and knowledge management for both IBM and Saffron and has written several papers on these topics, including editorship of *Neural Networks for Knowledge Representation and Inference*. Recent publications include “Learning by Collaborative and Individual-based Recommendation Agents” in the *Journal of Consumer Psychology* and “Concepts and Practice of Personalization” in *The Practical Handbook of Internet Computing*. Interviews and positive reviews of his work have appeared in *Infoworld*, *PC IA*, *New York Times Magazine*, *PC Week*, *AI Expert*, *Contemporary Psychology*, Seybold and Butler Group industry reports, *Defense News*, AMR Research, and *MIT Sloan Management Review*. He received his doctorate in experimental psychology from the U. of South Florida, specializing in truer, biologically-based neuro-computing, now becoming a new industry for "real intelligence" and the future of data analysis.

Agenda of this presentation

- Entity resolution. Various definitions and approaches
- Memory-based analysis. New approach for real intelligence
- Entity resolution examples. Proof in the pudding.
- Emerging directions. Quality targeting in unstructured sources
- SaffronWeb demonstration. Quality in text analytic discovery



Entity Resolution

Various definitions and approaches

Scenarios of Entity Resolution

- Resolution of dups, errors, name variants, and intentional aliasing
  - Reason by attributes, relationships, or transactional behavior
  - Also relevant to group detection such as terror cells and drug cartels
- Two use cases
  - No a priori target: unusual similarity is “signal” to detect
    - “Boil” the database and suggest similarities to investigate
    - Cross-database integration to merge community knowledge
  - Specific target of interest: recall other variants, aliases, or type
    - Analyst has a given target or watch list (looking for other identities)
    - Uncertainty about person at point of analysis or transaction (border entry, police stop)
- Compliment problem of identity separation
  - Expansion of one overlapping name into separate identities
- Real world challenges in National Security
  - In very sparse foreign intelligence data (“sparse” is too kind a word!)
  - At large (1M entities) to massive data scale (200M+ entities)

Review of Approaches

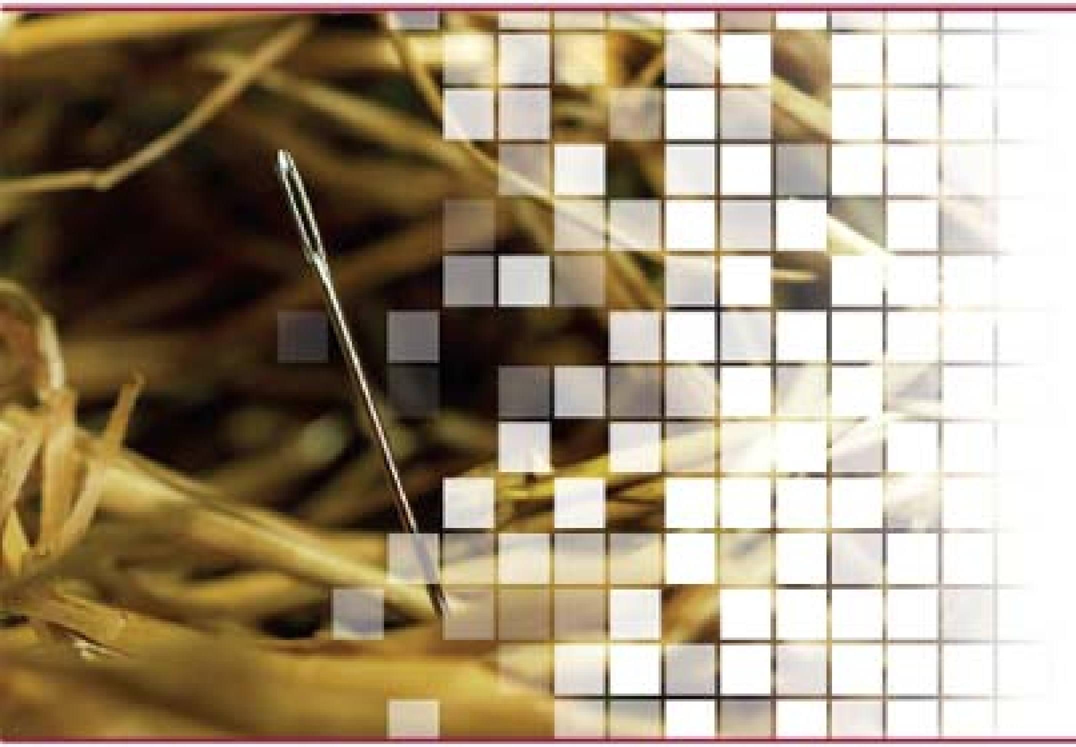
- Manual search with SQL queries
  - Labor-intensive with few discoveries
- Automated data cleansing
  - Rules of 2-4 primary, most informative attributes
- Lexical similarity
  - Rational and useful, but can generate many false alarms
- Document similarity
  - Penalty functions are inappropriate for uncertainties in intelligence
- Network “diffusion” mathematics
  - Requires symmetry and other matrix properties, only on abstract graph
- Feature and relationship statistics
  - Yes, but doesn’t address non-linear effects of transactional behaviors

- Accuracy is dependent on sparseness, size, the nature of data, and whether the task is to resolve dups, errors, variants, or aliases

Associative Memory Approach

- Information-based similarity
  - Absence of evidence is not evidence of absence
  - Entropy measure of unusual attribute-values
  - “Analogical entropy” to distinguish each candidate pair
- Unusual grouping operator
  - Computation of unusual “reciprocal coherence”
  - Rare occurrence when points are closest to each other
  - Non-centroid, non-globular crème-de-la-crème groups

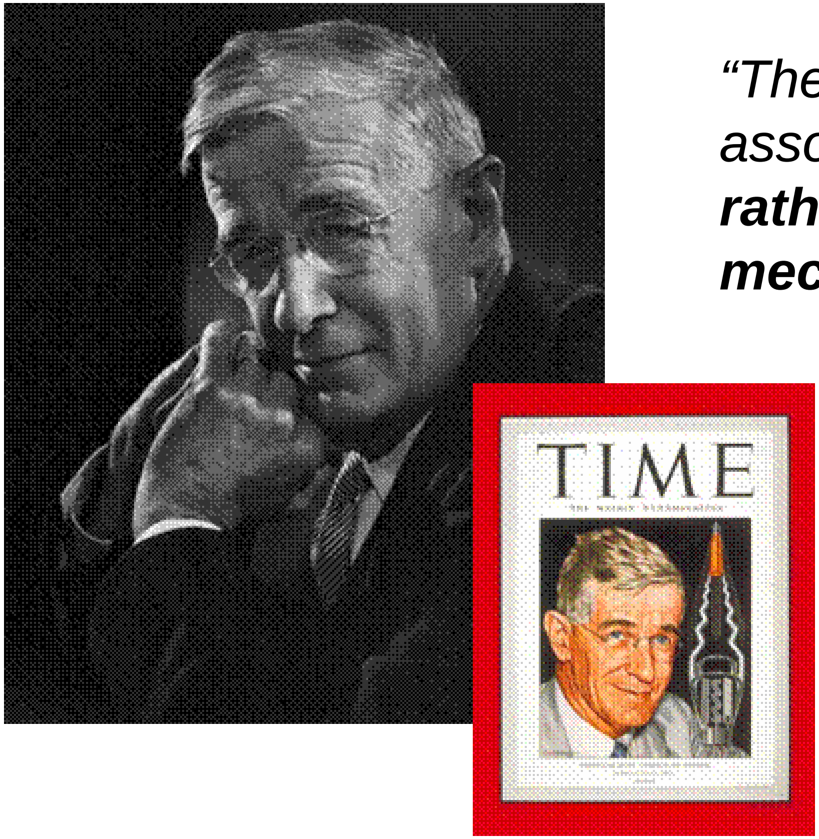
- No threshold of similarity could filter population noise
- Success only when using both computations together



Memory-based Analysis

New approach for real intelligence

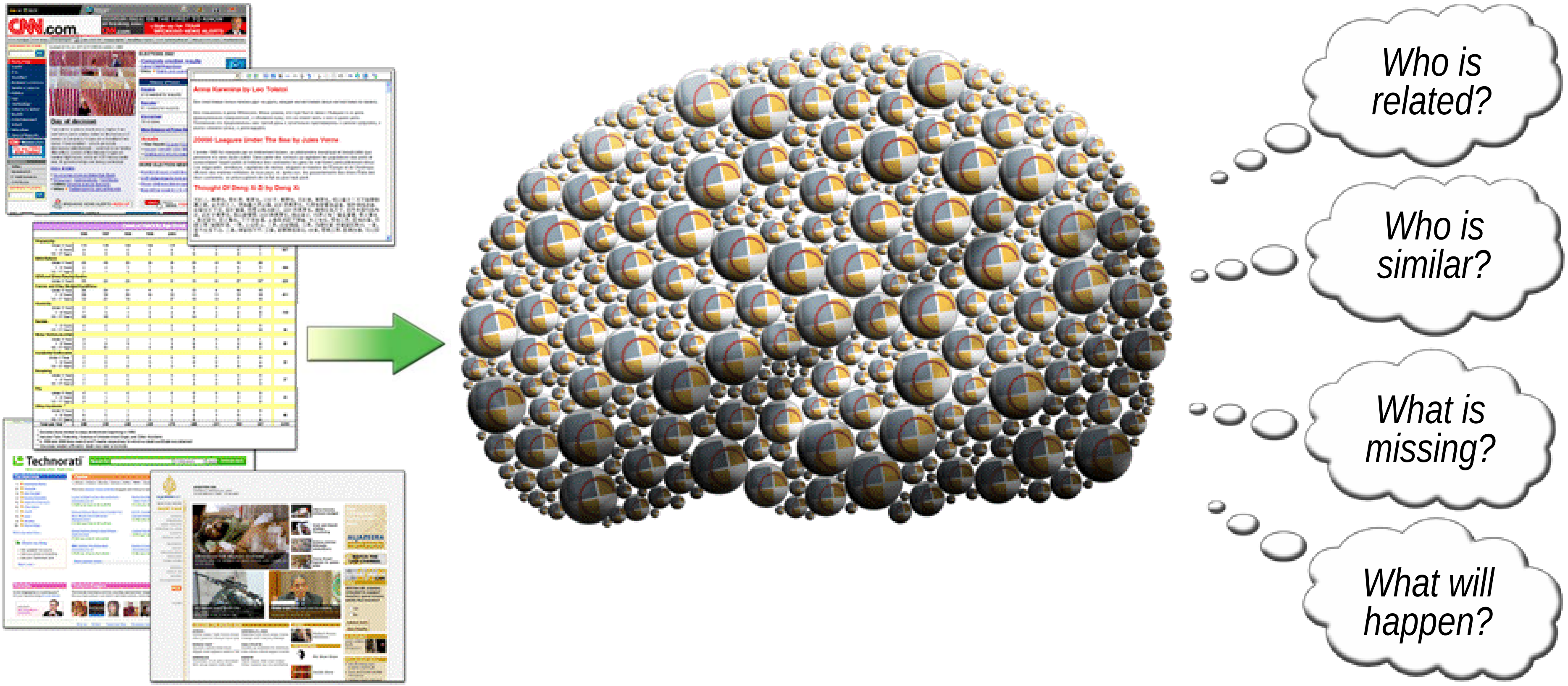
Can We Make Smarter, Not Just Stronger, Machines?



"The human mind … operates by association. ***Selection by association, rather than indexing, may yet be mechanized."***

As We May Think, 1945
Vannevar Bush

An Army of Personal Assistant Memories



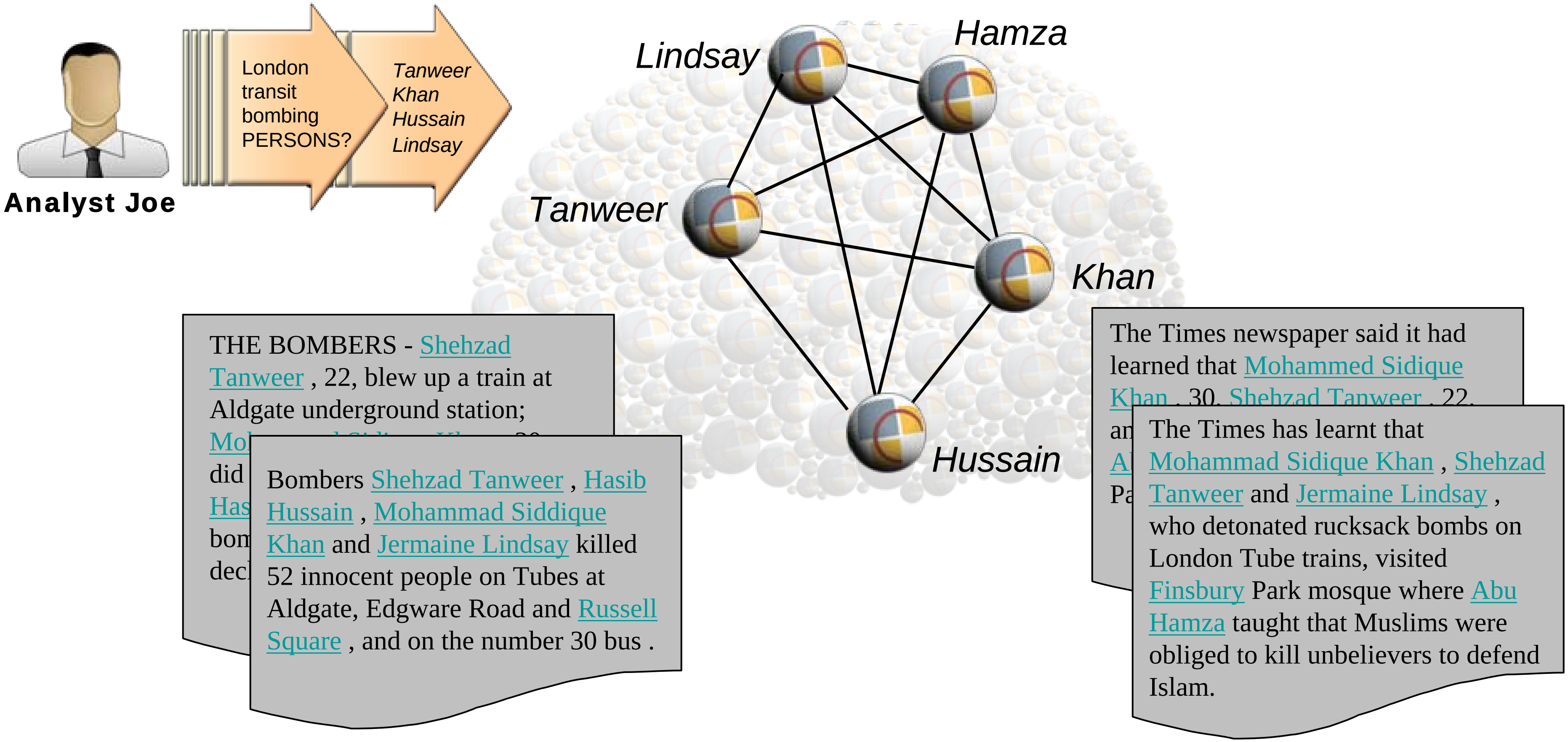
Who is related?

Who is similar?

What is missing?

What will happen?

Context Filtering of Entity Networks



London transit bombing PERSONS?

Tanweer
Khan
Hussain
Lindsay

THE BOMBERS - Shehzad Tanweer , 22, blew up a train at Aldgate underground station;
Mo
did
Has
bom
decl

Bombers Shehzad Tanweer , Hasib Hussain , Mohammad Siddique Khan and Jermaine Lindsay killed 52 innocent people on Tubes at Aldgate, Edgware Road and Russell Square , and on the number 30 bus .

The Times newspaper said it had learned that Mohammed Sidique Khan , 30, Shehzad Tanweer , 22,
an
Al
Pa

The Times has learnt that Mohammad Sidique Khan , Shehzad Tanweer and Jermaine Lindsay , who detonated rucksack bombs on London Tube trains, visited Finsbury Park mosque where Abu Hamza taught that Muslims were obliged to kill unbelievers to defend Islam.

Memories as Matrices: Between Data and Models

- ARE "Lazy" learners, such as:
  - Memory-based
  - Instance-based
  - Exemplar-based
  - Case-based
  - Experience-based
  - Nearest neighbor
- ARE NOT compiled functions:
  - Rules
  - Clustering
  - Regression
  - "Eager" neural networks

- Properties:
  - Incremental. Start from zero, learns case-by-case
  - Non-parametric. No knob-tweaking to build
  - Malleable. Adapt on the fly to new features
  - No over-training. Don't get worse as more data is seen!
  - Anomaly detection. Knows what it doesn't know
  - Unified representation. Various inferences can be computed at query-time

Recall Similar Objects from Data

animals

animal	blood	birth	legs	hair	scales	fins
horse	warm	livebearer	4	y	n	n
dog	warm	livebearer	4	y	n	n
dolphin	warm	livebearer	0	y	n	y
platypus	**warm**	**eggbearer**	**4**	**y**	**n**	**n**
trout	cold	eggbearer	0	n	y	y
thresher shark	warm	livebearer	0	n	n	y
tiger shark	cold	eggbearer	0	n	n	y
alligator	cold	eggbearer	4	n	y	n

Output

```
+---------------+------------+
|    animal     | similarity |
+---------------+------------+
| platypus      |      6     |
| dog           |      5     |
| horse         |      5     |
| dolphin       |      3     |
| alligator     |      3     |
| thresher shark|      2     |
| tiger shark   |      2     |
| trout         |      1     |
+---------------+------------+
```

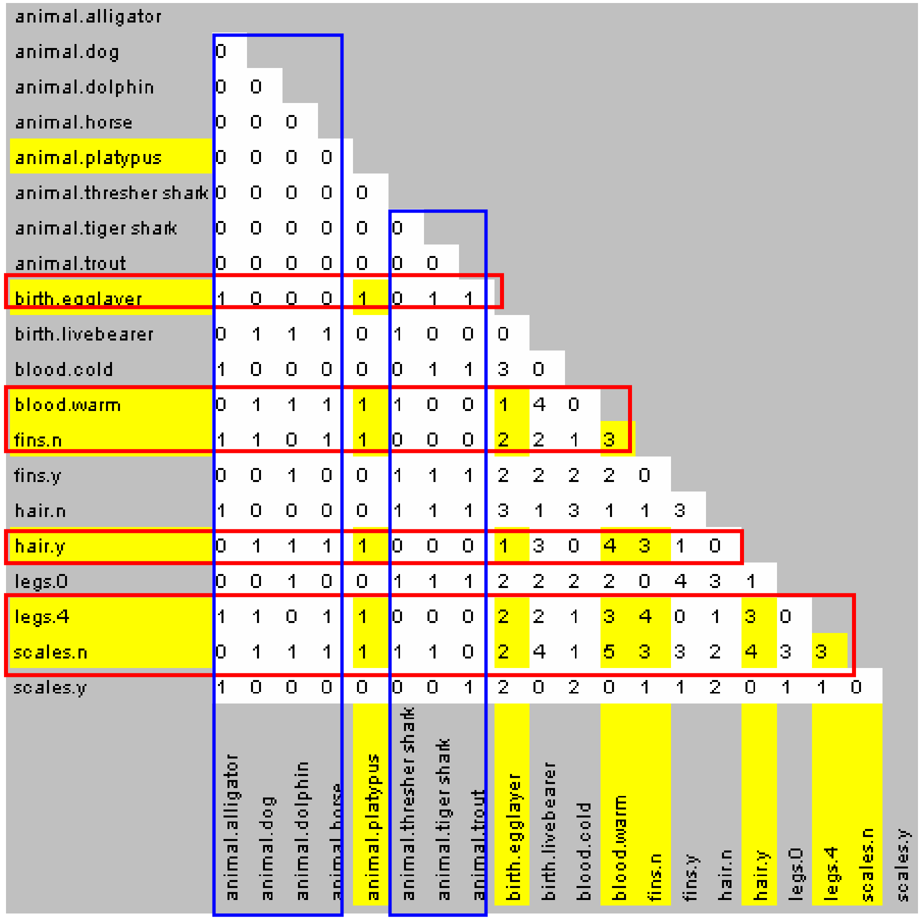
SQL

```
SELECT a.animal,
 ((CASE WHEN a.blood = b.blood THEN 1 ELSE 0 END) +
  (CASE WHEN a.birth = b.birth THEN 1 ELSE 0 END) +
  (CASE WHEN a.legs = b.legs THEN 1 ELSE 0 END) +
  (CASE WHEN a.hair = b.hair THEN 1 ELSE 0 END) +
  (CASE WHEN a.scales = b.scales THEN 1 ELSE 0 END) +
  (CASE WHEN a.fins = b.fins THEN 1 ELSE 0 END)) AS similarity
FROM animals a, animals b
WHERE b.animal = 'platypus'
ORDER BY similarity DESC;
```

Adding up number of shared values in different columns is tricky.

Recall Similar Objects from Memory

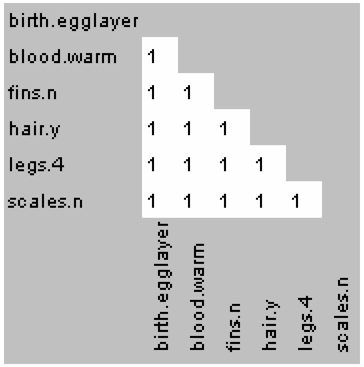
entity.animal



	animal.alligator	animal.dog	animal.dolphin	animal.horse	animal.platypus	animal.thresher shark	animal.tiger shark	animal.trout	birth.egglayer	birth.livebearer	blood.cold	blood.warm	fins.n	fins.y	hair.n	hair.y	legs.0	legs.4	scales.n	scales.y
animal.alligator																				
animal.dog	0																			
animal.dolphin	0	0																		
animal.horse	0	0	0																	
animal.platypus	0	0	0	0																
animal.thresher shark	0	0	0	0	0															
animal.tiger shark	0	0	0	0	0	0														
animal.trout	0	0	0	0	0	0	0													
birth.egglayer	1	0	0	0	1	0	1	1												
birth.livebearer	0	1	1	1	0	1	0	0	0											
blood.cold	1	0	0	0	0	0	1	1	3	0										
blood.warm	0	1	1	1	1	1	0	0	1	4	0									
fins.n	1	1	0	1	1	0	0	0	2	2	1	3								
fins.y	0	0	1	0	0	1	1	1	2	2	2	2	0							
hair.n	1	0	0	0	0	1	1	1	3	1	3	1	1	3						
hair.y	0	1	1	1	1	0	0	0	1	3	0	4	3	1	0					
legs.0	0	0	1	0	0	1	1	1	2	2	2	2	0	4	3	1				
legs.4	1	1	0	1	1	0	0	0	2	2	1	3	4	0	1	3	0			
scales.n	0	1	1	1	1	1	1	0	2	4	1	5	3	3	2	4	3	3		
scales.y	1	0	0	0	0	0	0	1	2	0	2	0	1	1	2	0	1	1	0	

3 5 3 5 2 2 1

entity.animal.platypus



	birth.egglayer	blood.warm	fins.n	hair.y	legs.4	scales.n
birth.egglayer						
blood.warm	1					
fins.n	1	1				
hair.y	1	1	1			
legs.4	1	1	1	1		
scales.n	1	1	1	1	1	

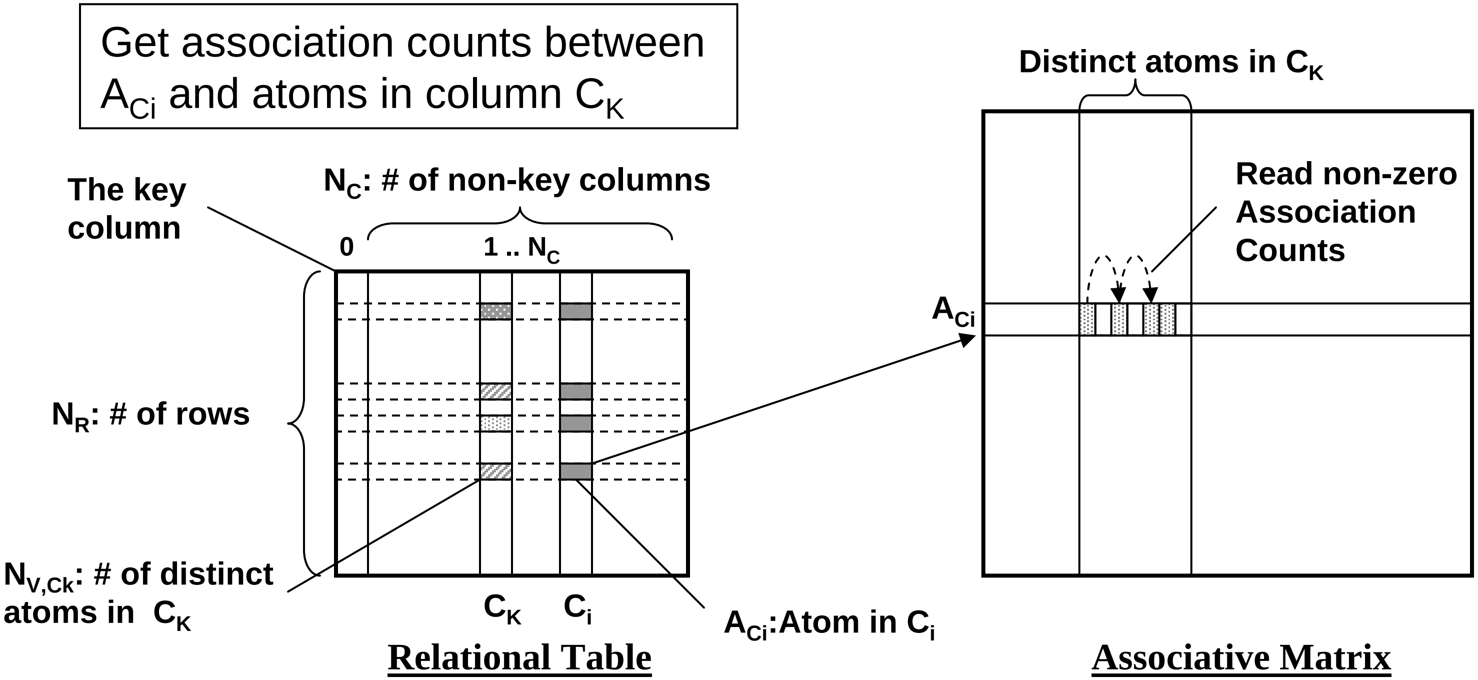
Saffron SQL

```
SELECT animal
FROM entity.animal
ASSOCIATED WITH
  (SELECT * FROM entity.animal.platypus)
```

Output

```
+----------+----------------+--------+
| category | value          | metric |
+----------+----------------+--------+
| animal   | platypus       | 6.000  |
| animal   | horse          | 5.000  |
| animal   | dog            | 5.000  |
| animal   | dolphin        | 3.000  |
| animal   | alligator      | 3.000  |
| animal   | thresher shark | 2.000  |
| animal   | tiger shark    | 2.000  |
| animal   | trout          | 1.000  |
+----------+----------------+--------+
```

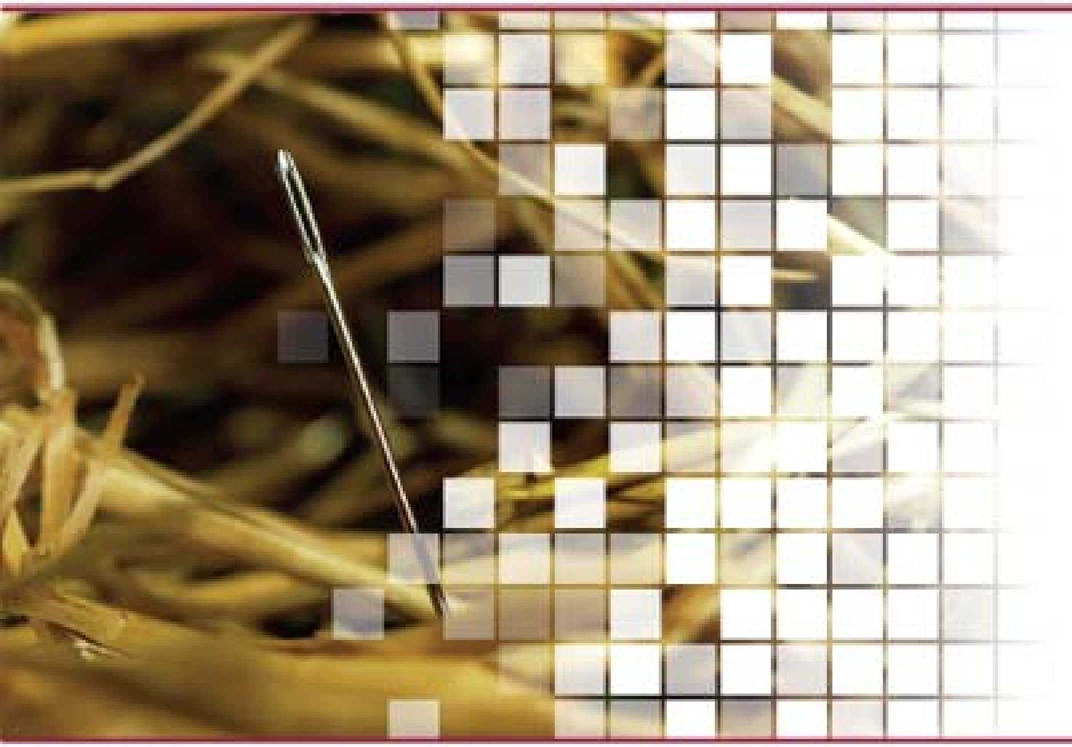
Memory Performance Advantage


Get association counts between A_{Ci} and atoms in column C_K
The key column
N_C: # of non-key columns
0
1 .. N_C
N_R: # of rows
$N_{V,Ck}$: # of distinct atoms in C_K
C_K
C_i
A_{Ci}:Atom in C_i
Relational Table
Distinct atoms in C_K
Read non-zero Association Counts
A_{Ci}
Associative Matrix

Query: $S_R/S_A = (N_R/N_{V,Ck})*(N_C+1) \propto N_R/N_{V,Ck}$

Insertion: $I_R/I_A = ((N_C+1)*N_R) / ((N_C+1)* N_C*N_R *2/2) = 1 / N_C$

- The associative matrix has increasingly better query performance



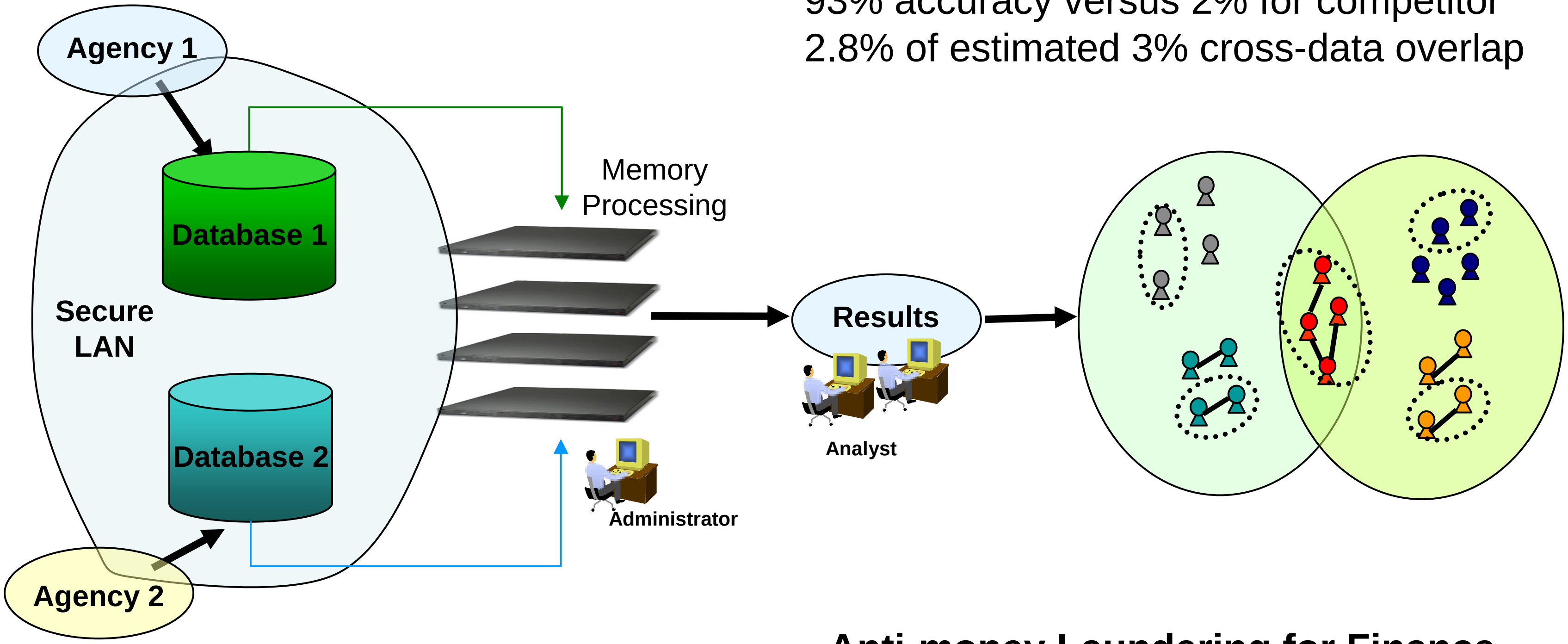
Entity Resolution Examples

Proof in the pudding

Alias Detection for Foreign Intelligence

93% accuracy versus 2% for competitor

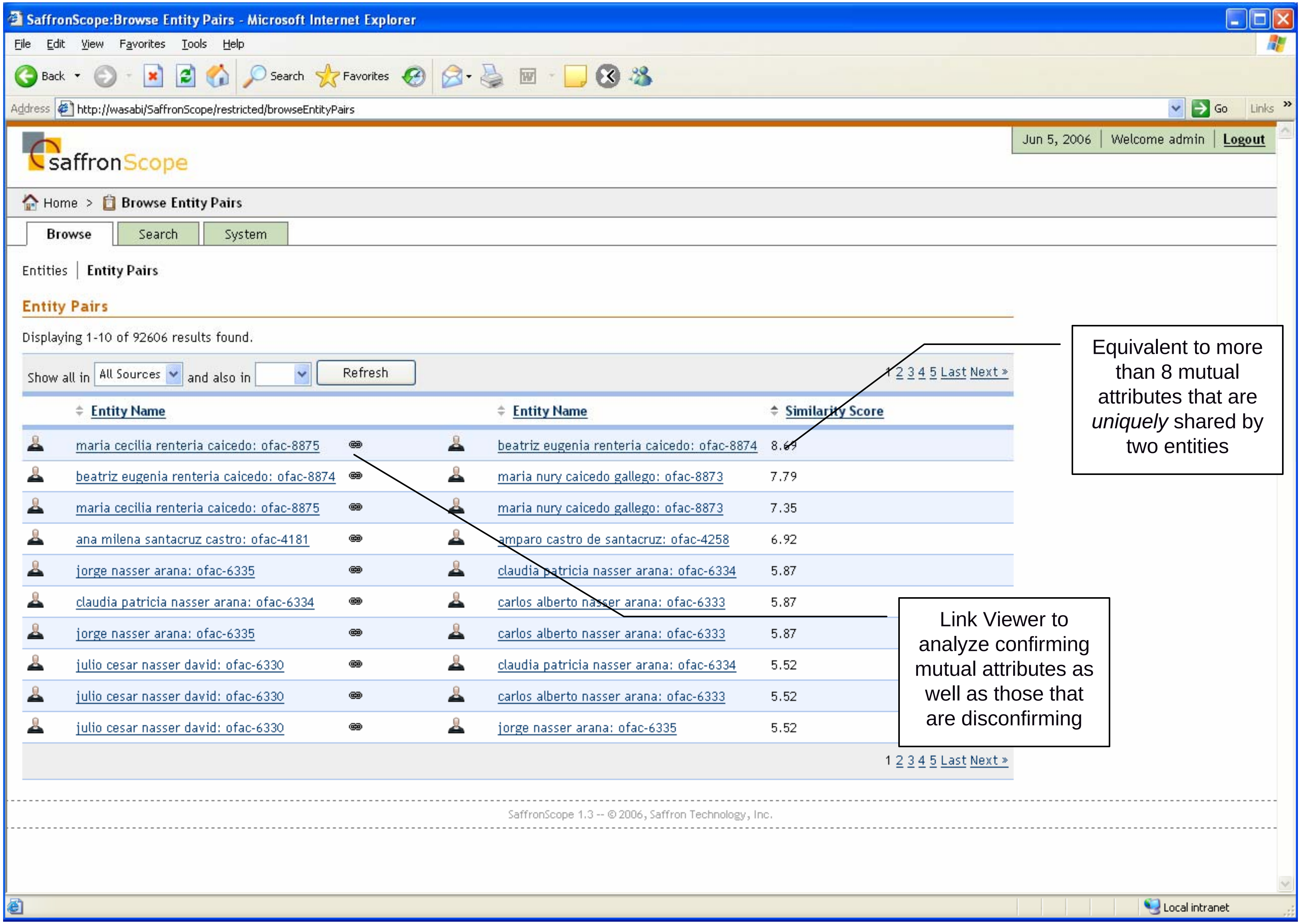
2.8% of estimated 3% cross-data overlap



Anti-money Laundering for Finance

100% accuracy joining PFA and OFAC

Identification of drug cartel grouping


SaffronScope:Browse Entity Pairs - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
Go
Links
Local intranet

http://wasabi/SaffronScope/restricted/browseEntityPairs

saffronScope

Jun 5, 2006 | Welcome admin | Logout

Home > Browse Entity Pairs

Browse | Search | System

Entities | Entity Pairs

Entity Pairs

Displaying 1-10 of 92606 results found.

Show all in All Sources and also in [Refresh]

1 2 3 4 5 Last Next »

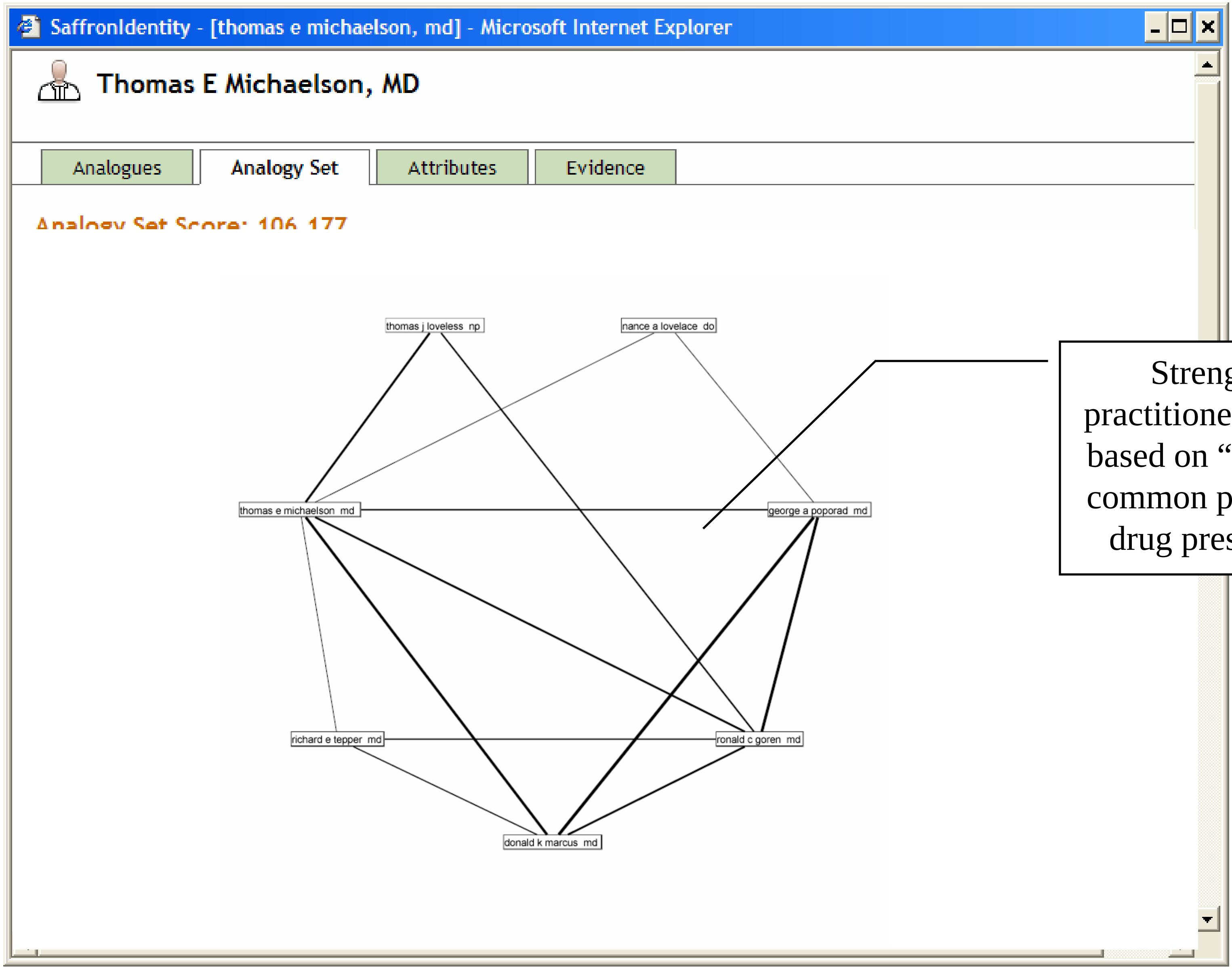
Entity Name	Entity Name	Similarity Score
maria cecilia renteria caicedo: ofac-8875	beatriz eugenia renteria caicedo: ofac-8874	8.69
beatriz eugenia renteria caicedo: ofac-8874	maria nury caicedo gallego: ofac-8873	7.79
maria cecilia renteria caicedo: ofac-8875	maria nury caicedo gallego: ofac-8873	7.35
ana milena santacruz castro: ofac-4181	amparo castro de santacruz: ofac-4258	6.92
jorge nasser arana: ofac-6335	claudia patricia nasser arana: ofac-6334	5.87
claudia patricia nasser arana: ofac-6334	carlos alberto nasser arana: ofac-6333	5.87
jorge nasser arana: ofac-6335	carlos alberto nasser arana: ofac-6333	5.87
julio cesar nasser david: ofac-6330	claudia patricia nasser arana: ofac-6334	5.52
julio cesar nasser david: ofac-6330	carlos alberto nasser arana: ofac-6333	5.52
julio cesar nasser david: ofac-6330	jorge nasser arana: ofac-6335	5.52

1 2 3 4 5 Last Next »

SaffronScope 1.3 -- © 2006, Saffron Technology, Inc.

Equivalent to more than 8 mutual attributes that are *uniquely* shared by two entities

Link Viewer to analyze confirming mutual attributes as well as those that are disconfirming

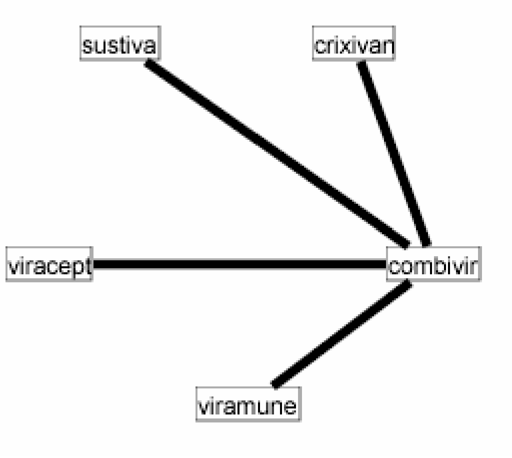


SaffronIdentity - [thomas e michaelson, md] - Microsoft Internet Explorer

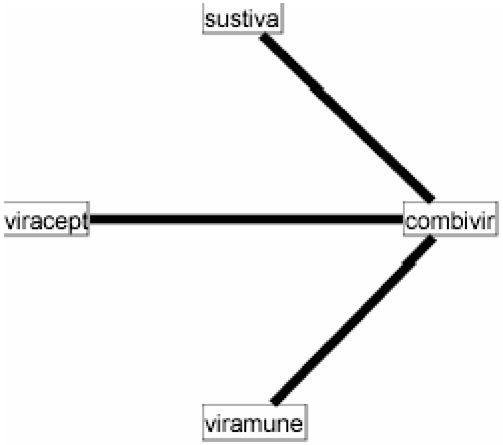
Thomas E Michaelson, MD

Analogy Set Score: 106.177

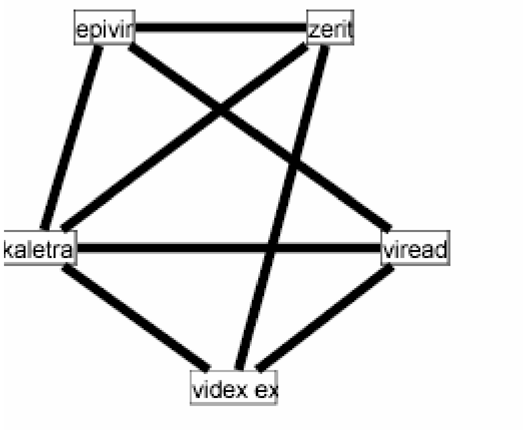
Strengths of practitioner similarity based on “unusually” common patients and drug prescriptions



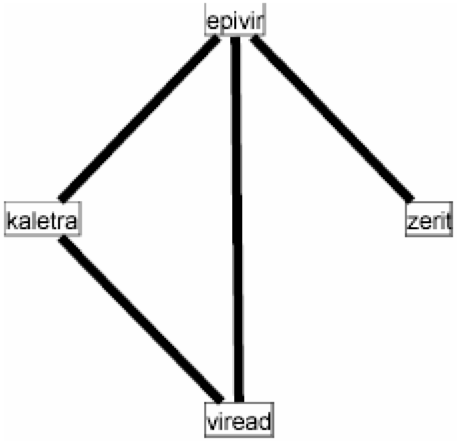
May 2002 - Jun 2002



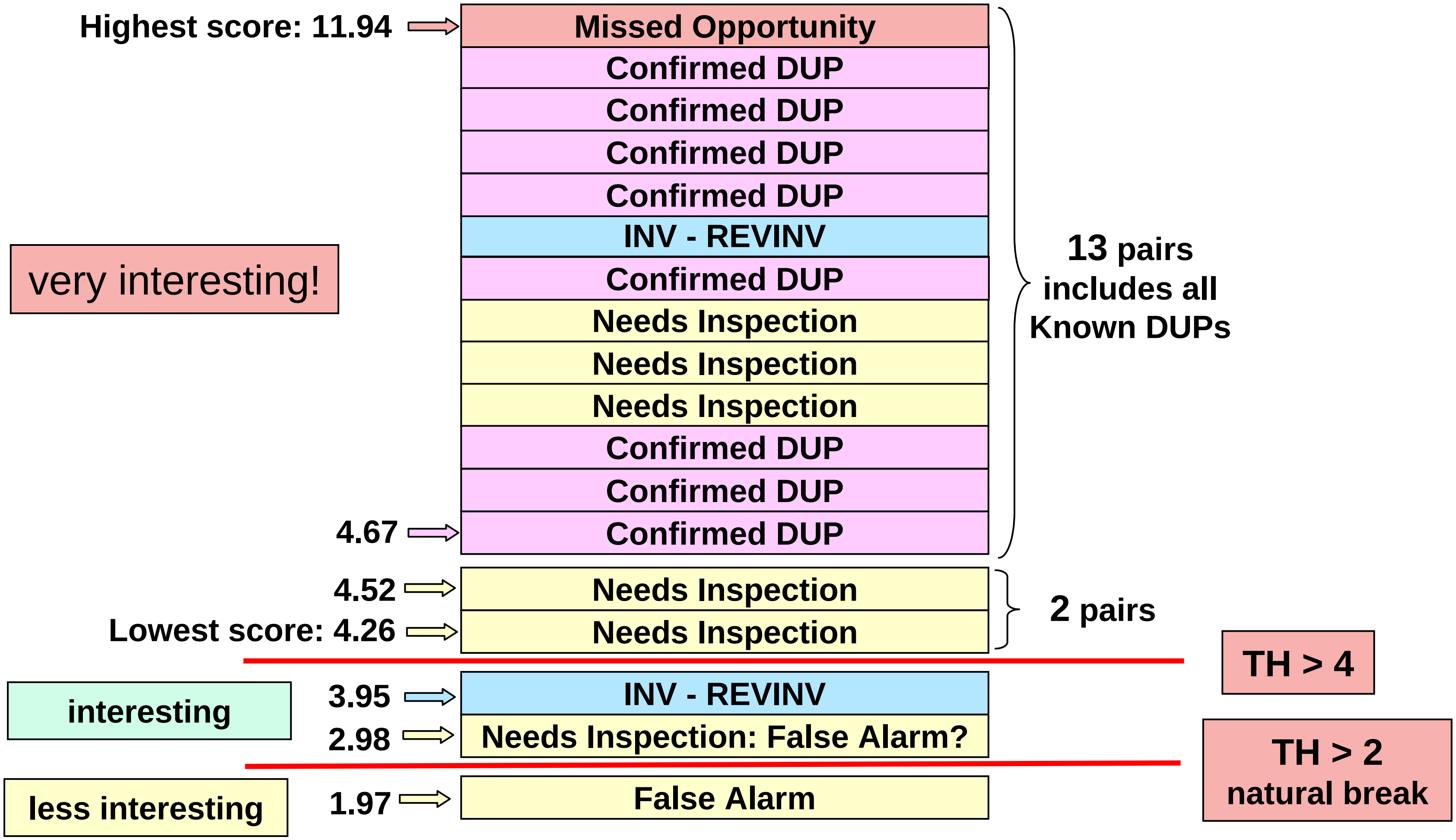
Jul 2002 - Mar 2003



Apr 2003 - Oct 2003



Nov 2003 – Jan 2004



Highest score: 11.94

Missed Opportunity

Confirmed DUP

Confirmed DUP

Confirmed DUP

Confirmed DUP

INV - REVINV

Confirmed DUP

Needs Inspection

Needs Inspection

Needs Inspection

Confirmed DUP

Confirmed DUP

4.67

Confirmed DUP

13 pairs includes all Known DUPs

very interesting!

4.52

Needs Inspection

Lowest score: 4.26

Needs Inspection

2 pairs

TH > 4

interesting

3.95

INV - REVINV

2.98

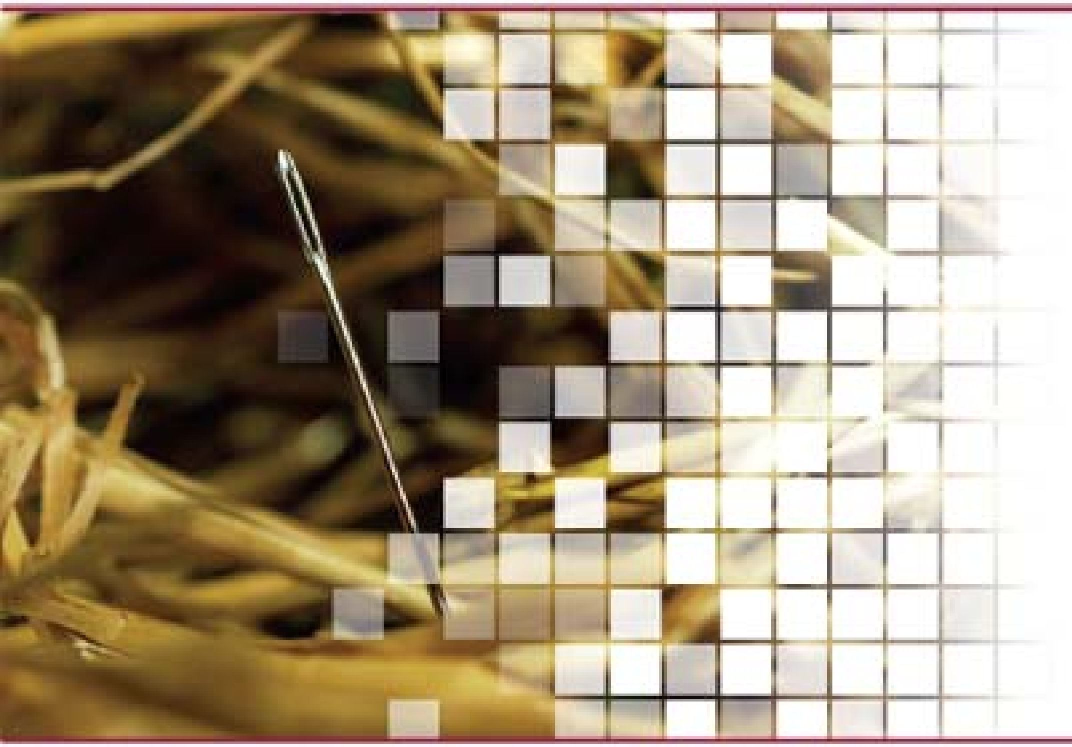
Needs Inspection: False Alarm?

TH > 2 natural break

less interesting

1.97

False Alarm



Emerging Directions

Quality targeting in unstructured sources

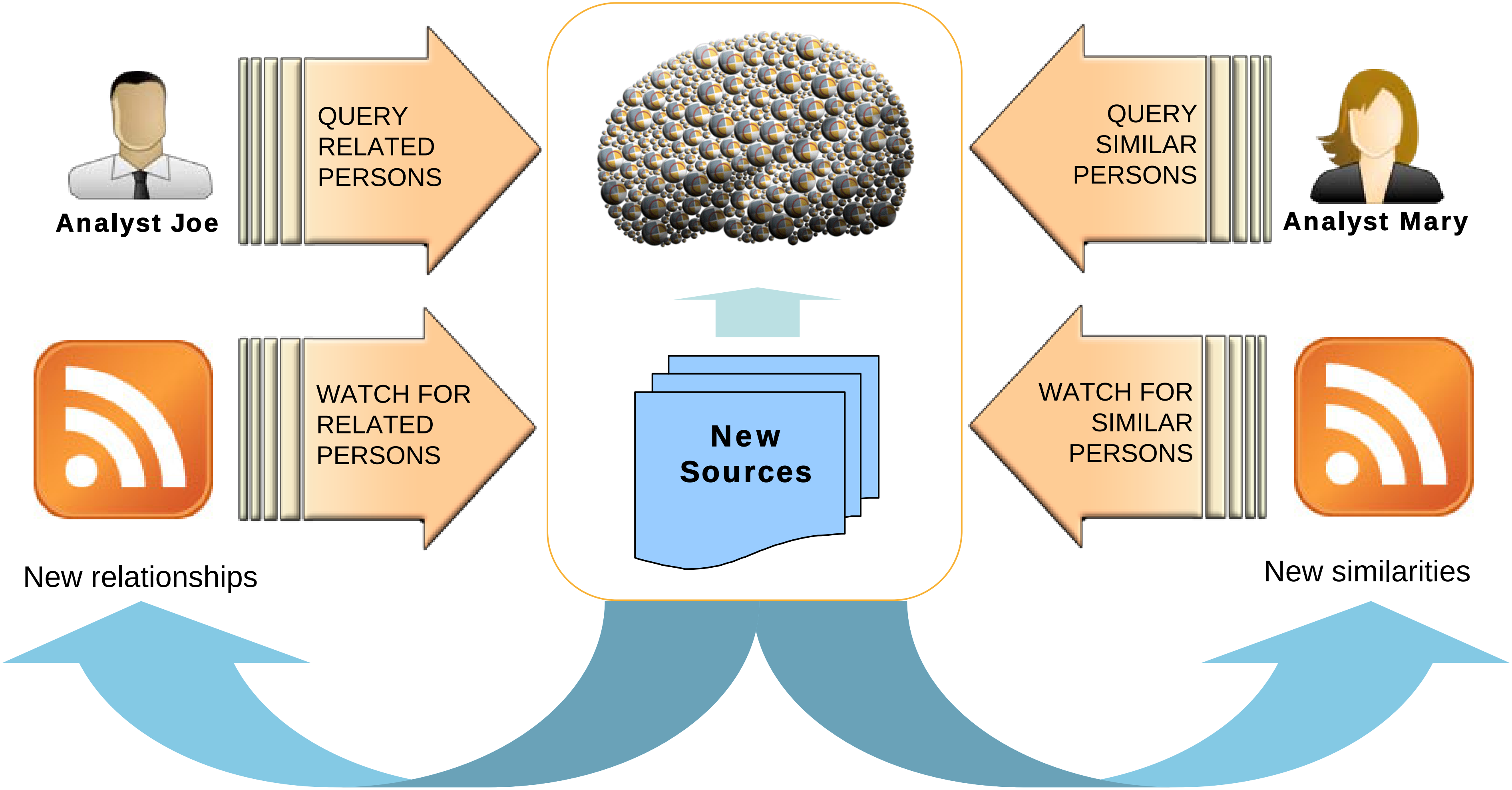
Entity Extraction from Unstructured Sources

- Industry is mature but not yet at high quality
  - Operational solutions require name list authoring to make things right
  - Mixed results for resolution methods intended for structured sources

- Continuing quality problems
  - Misclassification of entity type (Mr. Saab said, "…")
  - Name variants and aliasing of each identity (Mohammed, IBM, etc)
  - No disambiguation of different entities (John Smith #33)

- Remaining needs for higher accuracy
  - Real-time machine learning for specific corpus and continuous change
  - Systems solution (not a single product algorithm) -- including users

Perpetual Associative Targeting by Analytic Exploitation



Analyst Joe

QUERY RELATED PERSONS

QUERY SIMILAR PERSONS

Analyst Mary

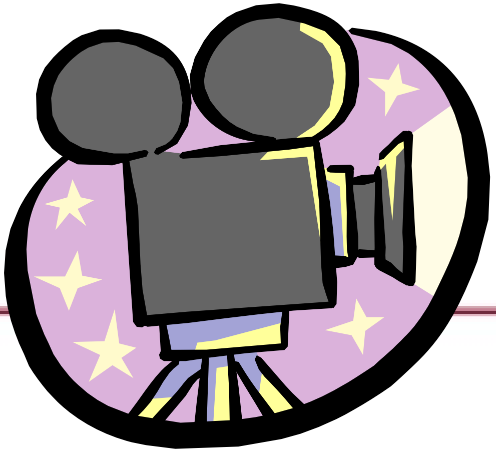
WATCH FOR RELATED PERSONS

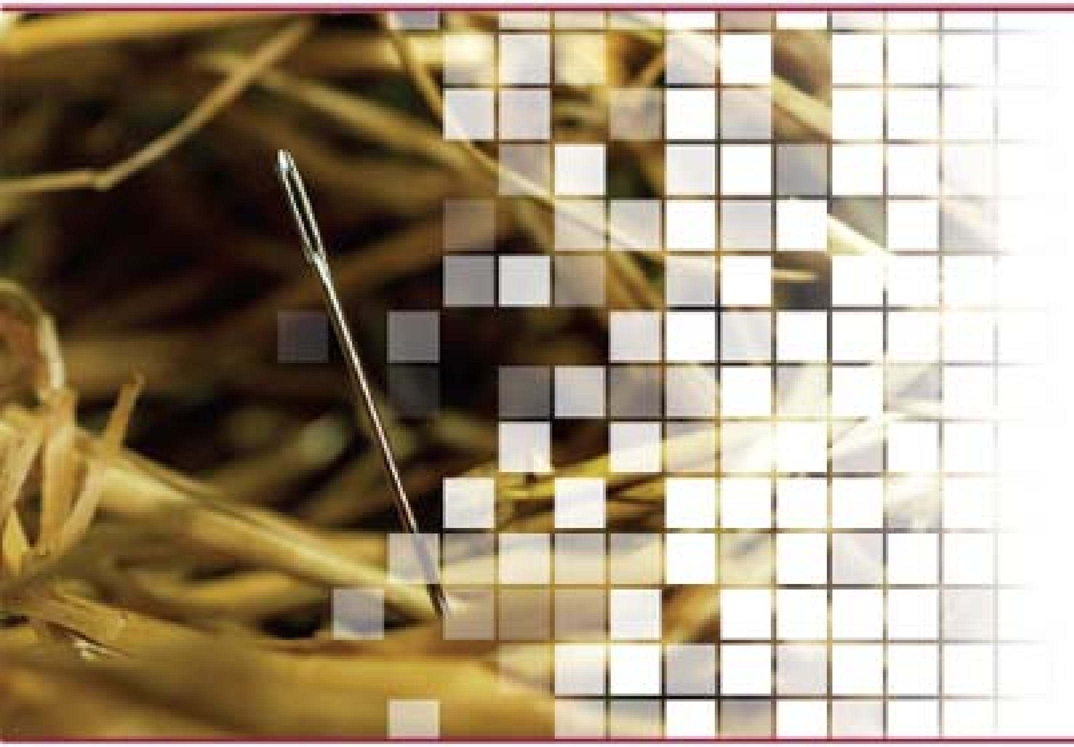
New Sources

WATCH FOR SIMILAR PERSONS

New relationships

New similarities





SaffronWeb Demonstration

Quality in text analytic discovery

Product Demonstration Review

- Analyst corrections of extraction
  - Reclassification. Correct any instances when collecting snippets
  - Grouping. Manage personal/collective variant and alias lists
- Advanced memory-based recall
  - Entities like this. Similarity-based query to better cover identity
  - Tag dipping. Relationship-based query to better complete report
- Allows analyst to clean up extraction problems before reporting
- Greater completeness in recall of similarities and relationships

Thanks to MIT IQ and to You

Manny Aparicio

WEB:	*www.saffrontech.com*
BLOG:	*www.manuelaparicio.com*
EMAIL:	*maparicio@saffrontech.com*

Saffron and all other Saffron Technology Inc. product or service names are registered trademarks or trademarks of Saffron Technology Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

This presentation is the confidential and proprietary property of Saffron Technology Inc. It may contain approaches, techniques, and other information proprietary to Saffron, and shall not be disclosed in whole or in part to third parties without the prior written consent of Saffron.