



IBM Entity Analytic Solutions (EAS)

# **The Sound of One Hand Clapping Knowledge Discovery without Disclosure: A Koan for the Information Age**

**John Bliss, Privacy Strategist, Entity Analytic  
Solutions, IBM**

April 2006

© 2006 IBM Corporation

# Introduction

- Systems Research & Development founded by Jeff Jonas in 1983
- Headquarters moved to Las Vegas in early 90's
- Worked with the gaming industry to help them better understand with whom they were doing business
- Commercialized a technology which became known as NORA (Non-Obvious Relationship Awareness)
- Partially funded by In-Q-Tel in 2001
- Professional management team retained in 2002
- SRD acquired by IBM January 2005

# Agenda

- Hunting Bad Guys in Vegas
- IBM Privacy and Civil Liberties Leadership
- Introduction to Anonymous Resolution
- State of the Union
- Questions and Answers

# Hunting for Bad Guys in Vegas

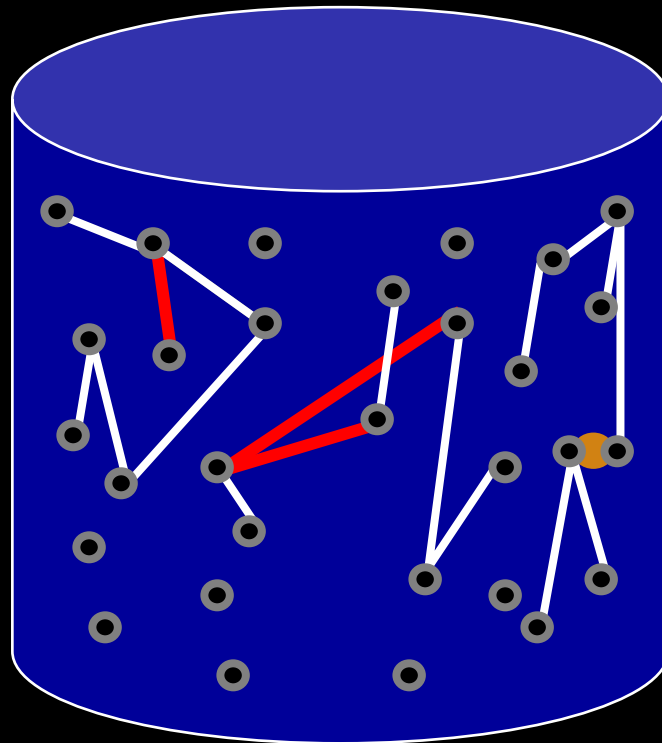
# Discovery on Data Streams

## Good Guys

- Hotel guests
- Loyalty club enrollment
- Employees
- Vendors
- Victims

## Subjects of Interest

- Specially designated nationals
- Excluded persons
- Gaming license revocations
- Known cheaters
- Interpol
- FBI Most Wanted



# Case Study: Las Vegas Casino

## Data Sources

- 20,000 plus employees
- All vendors
- All slot club & table games-related players
- In-house arrests/incidents
- Known cheaters

## Detected Relationships

- 24 active players were known cheaters
- 23 players had relationships to prior arrests/incidents
- 12 employees were themselves the player
- 192 employees had possible vendor relationships
- 7 employees were the vendor



This Became Known as ...

# Non-Obvious Relationship Awareness (NORA)

Renamed in 2005 as...

# Identity and Relationship Resolution

# Case Study: Retail

## Data Sources

- 40,000 plus employees
- 10,000 plus vendors
- 26,000 international security/arrest records (shoplifters, etc.)

## Detected Relationships

- 2 out of every 1000 employees had been arrested for shoplifting
- 8 out of every 1000 employees were related to known shoplifters
- 9 vendors on the internal security file
- 1 executive related to a vendor (a charity). Possible case of embezzlement.



# Case Study: Federal Agency

## Data Sources

- 20,000 plus employees
- 75,000 plus vendors
- 200,000 plus Type 1 security risk entities
- 200,000 plus Type 2 security risk entities

## Detected Relationships

- 140 employee relationships to vendors
- 1451 potential vendor relationships to security risks
- 253 employee relationships to security risk entities
- 2 vendors were the security risk
- "n" employees were the security risk/vendor

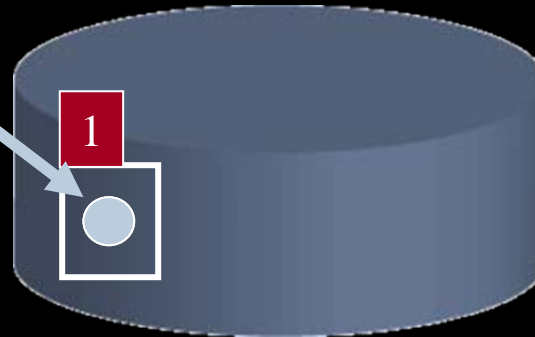


# Identity Resolution Demonstration

# In 2002 The System Observes "A" Record ...

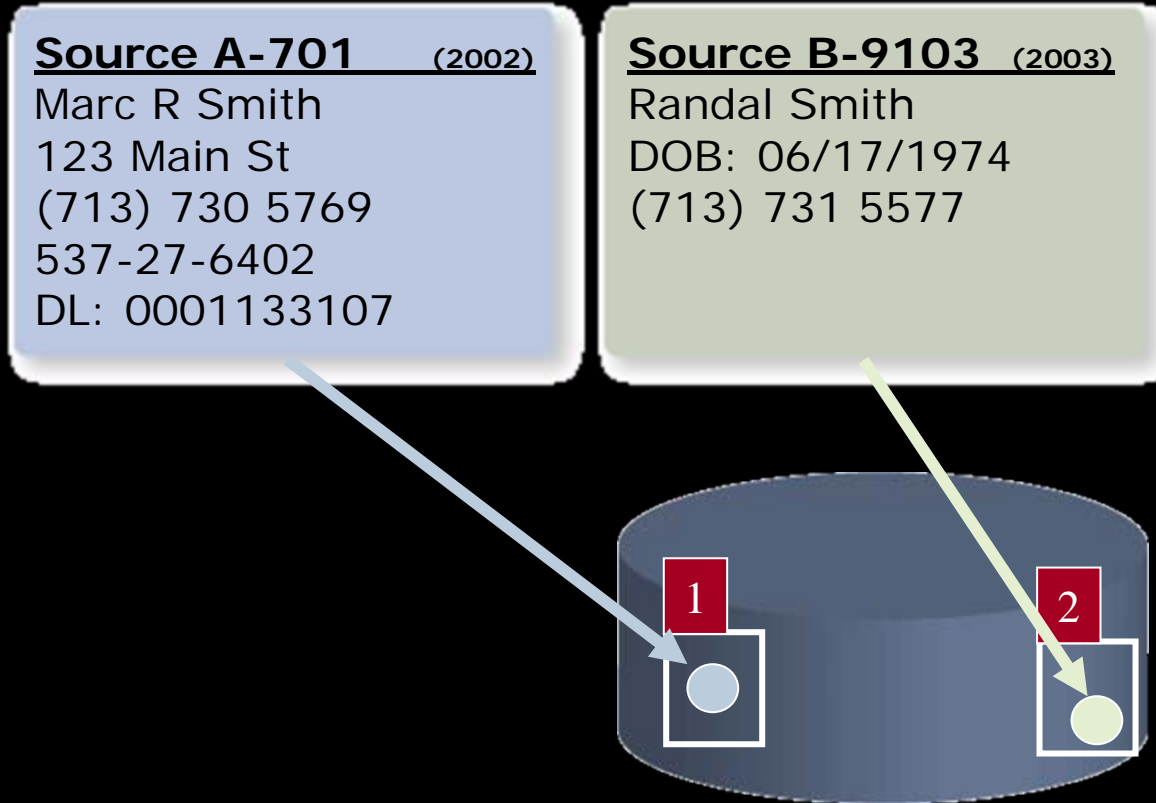
**Source A-701** (2002)

Marc R Smith  
123 Main St  
(713) 730 5769  
537-27-6402  
DL: 0001133107



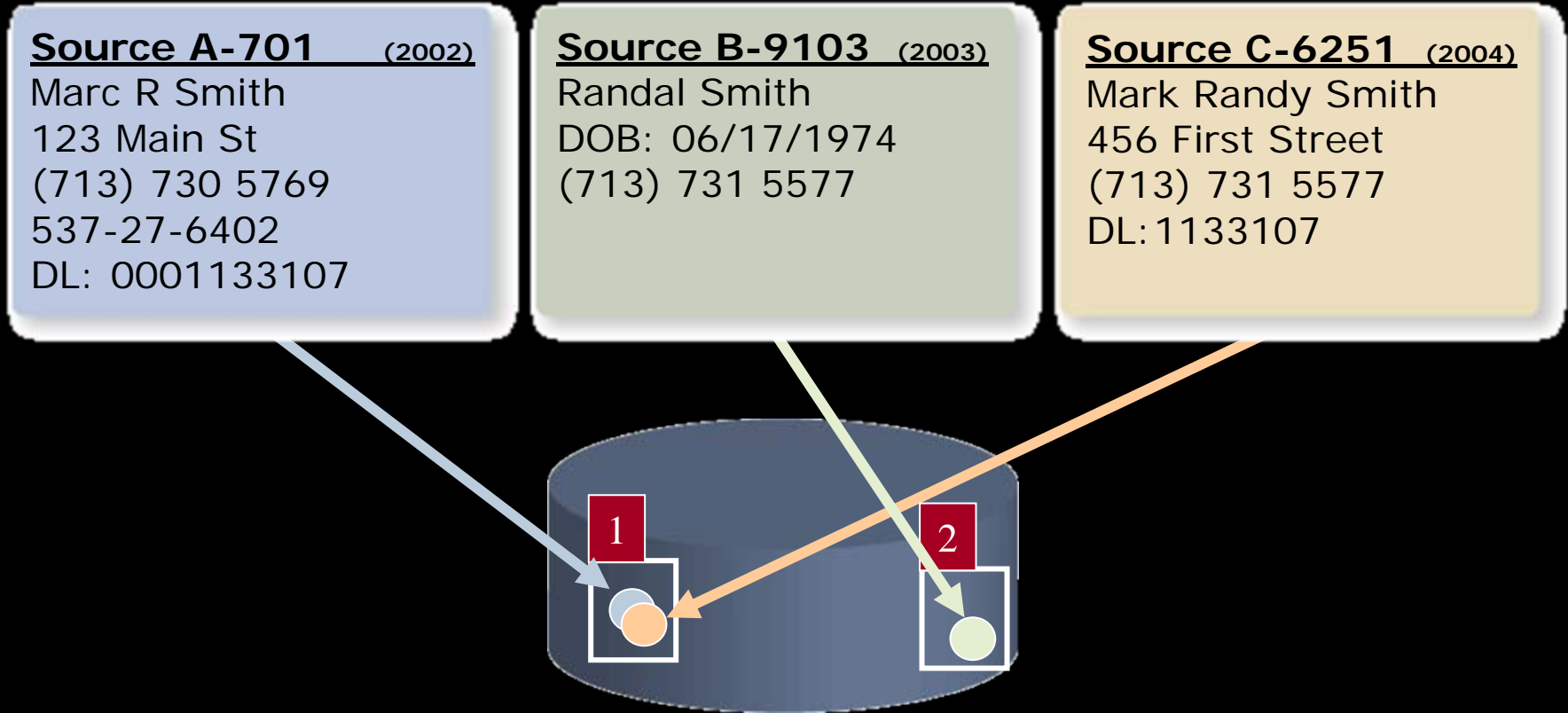
Identity is determined to be new

# In 2003 The System Observes "B" Record...



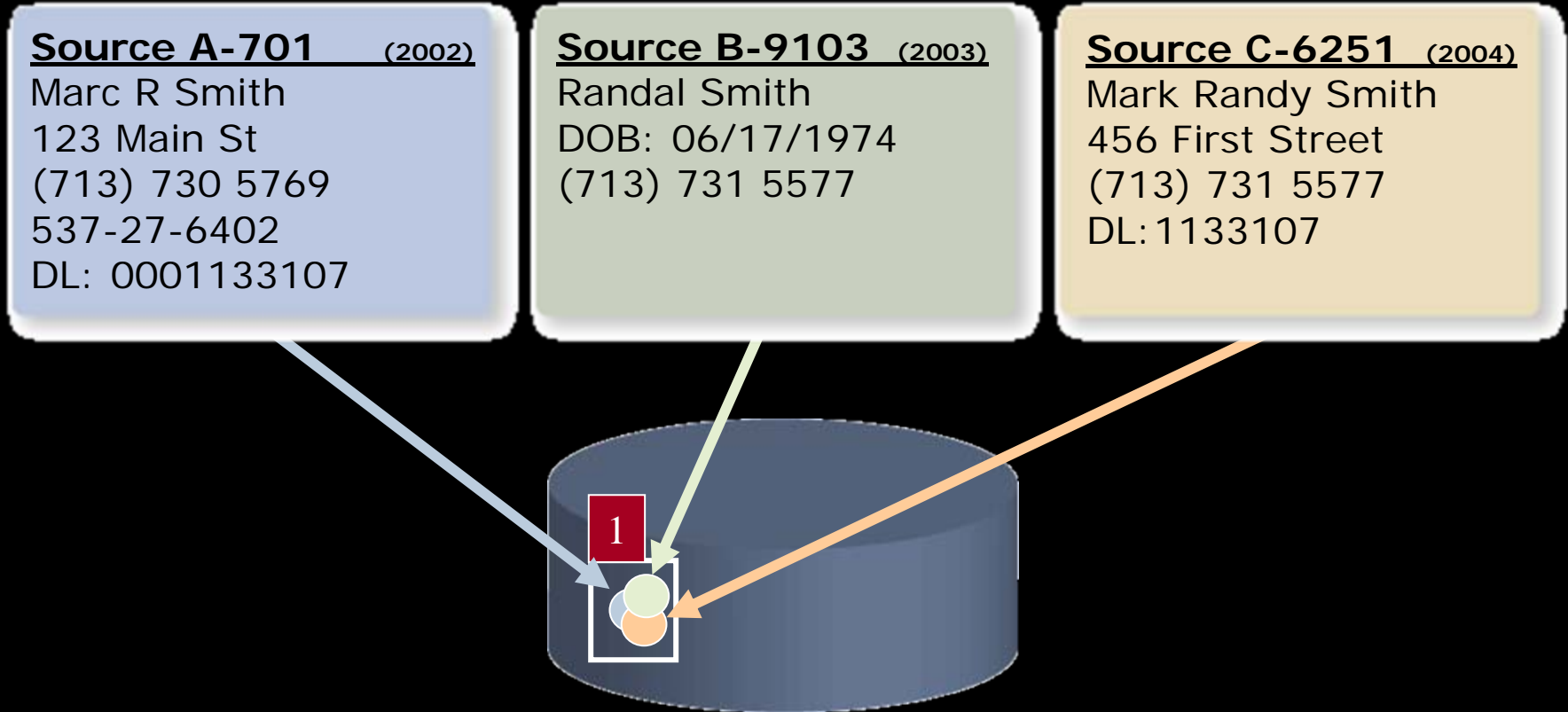
Identity is again determined to be new

# In 2004 The System Observes "C" Record...



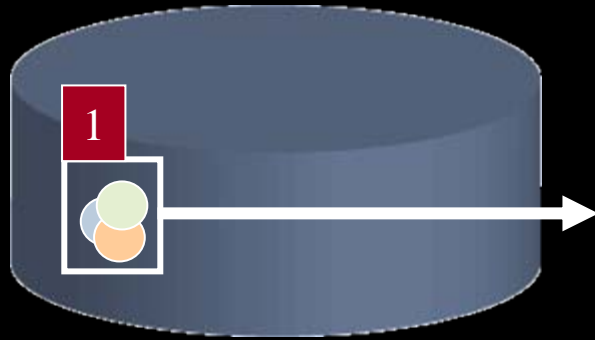
Identity is determined to be known

# Instantly The System Discovers “A is B is C”



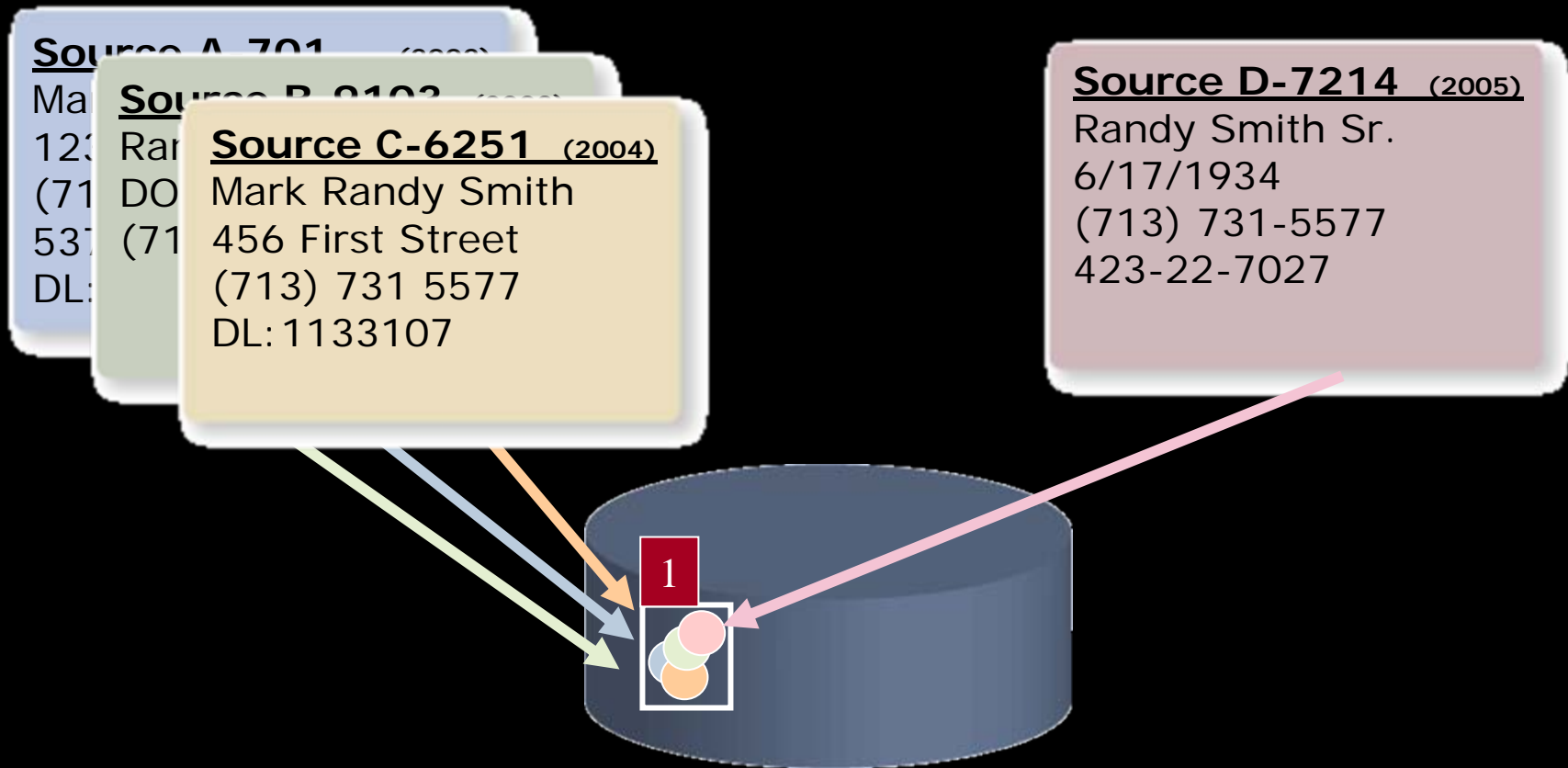
Sequence neutrality rules cause the two identities to “collapse”

# Behind the Scenes: Context is Accumulating



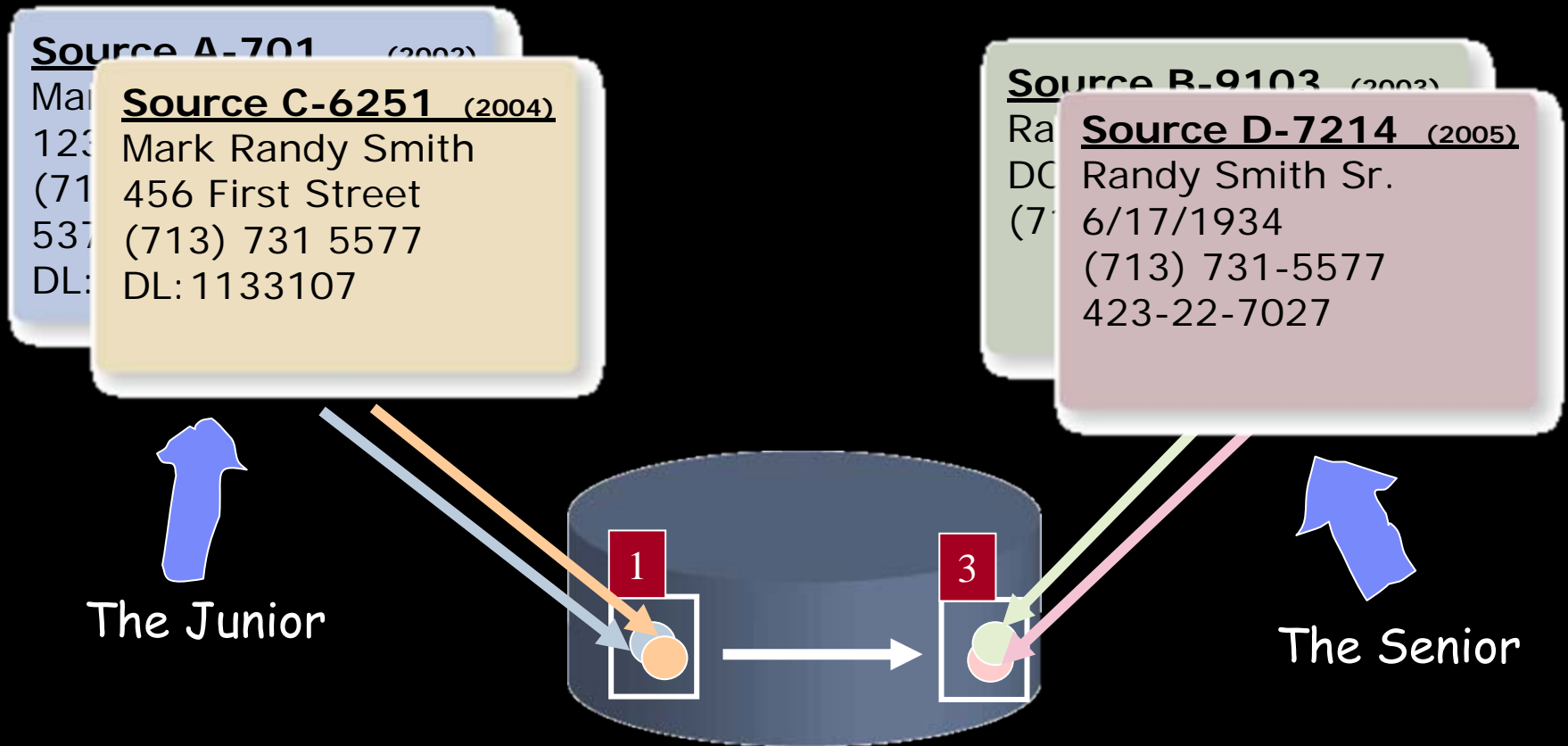
ELEMENT	VALUE (FEATURE)	ATTRIBUTION
<b>Names</b>	Marc R Smith	A-701
	Randal Smith	B-9102
	Mark Randy Smith	C-6251
<b>Address</b>	123 Main St.	A-701
	456 First Street	C-6251
<b>Phones</b>	(713) 730-5769	A-701
	(713) 731-5577	B-9102
	(713) 731-5577	C-6251
<b>SSN</b>	537-27-6402	A-701
<b>DL</b>	0001133107	A-701
	1133107	C-6251
<b>DOB</b>	06/17/1974	B-9103

# In 2005 The System Observes "D" Record...



Identity is determined to be known

# Instantly The System Discovers “A is C” and “B is D”

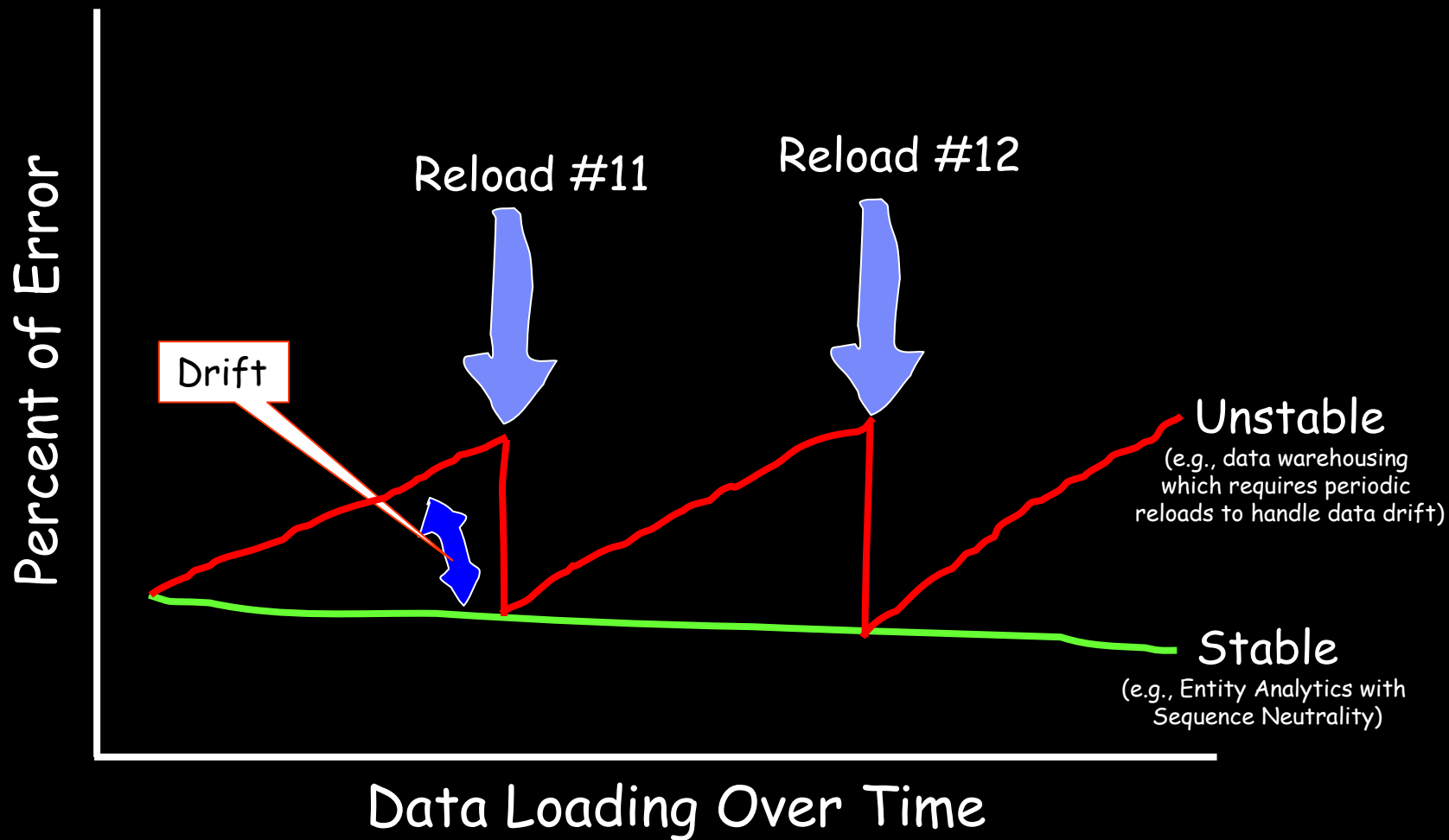


Sequence neutrality rules cause the identity to “split”

# Identity Resolution Requires ...

- **Persistent Context**
  - Received data is reconciled to historical holdings and persisted (versus context-on-the-fly)
  
- **Tethering to Source Systems**
  - Processing of adds, changes and deletes from source systems
  
- **Full Attribution**
  - Every row retains its pedigree (no data survivorship processing)
  
- **Sequence Neutrality**
  - New data corrects previous outcomes improving accuracy over time
  - The database end-state is the same despite the arrival order or timing of the data

# Sequence Neutrality is Critical for Context Stability



# IBM Leadership in Privacy and Civil Liberty Protections

# Listening to the Privacy Community ... and support

- Knowledge discovery programs with additional degrees of transparency, oversight and accountability
- Use of tamper-resistant audits logs (e.g., Immutable Audit Logs) especially in non-transparent systems
- Where there are watch lists, adoption of strong oversight and redress policies
- Opposition of data mining to predict terrorists based on anomalies/behavior
- Limiting transfer of sensitive data
- Concerns about use of “privacy enhancing technologies” as a fig leaf masking intrusive programs
- If there is to be a data transfer (or knowledge discovery event), anonymize when possible

# Introducing ... Anonymous Resolution

# Introducing Anonymous Resolution

A technique that allows  $n$  data holders to share anonymized identity-based data ...

whereby all identities are managed and correlated while in their cryptographic form ...

resulting in a more secure and privacy-enhanced way to resolve identities and discover relationships of particular interest.

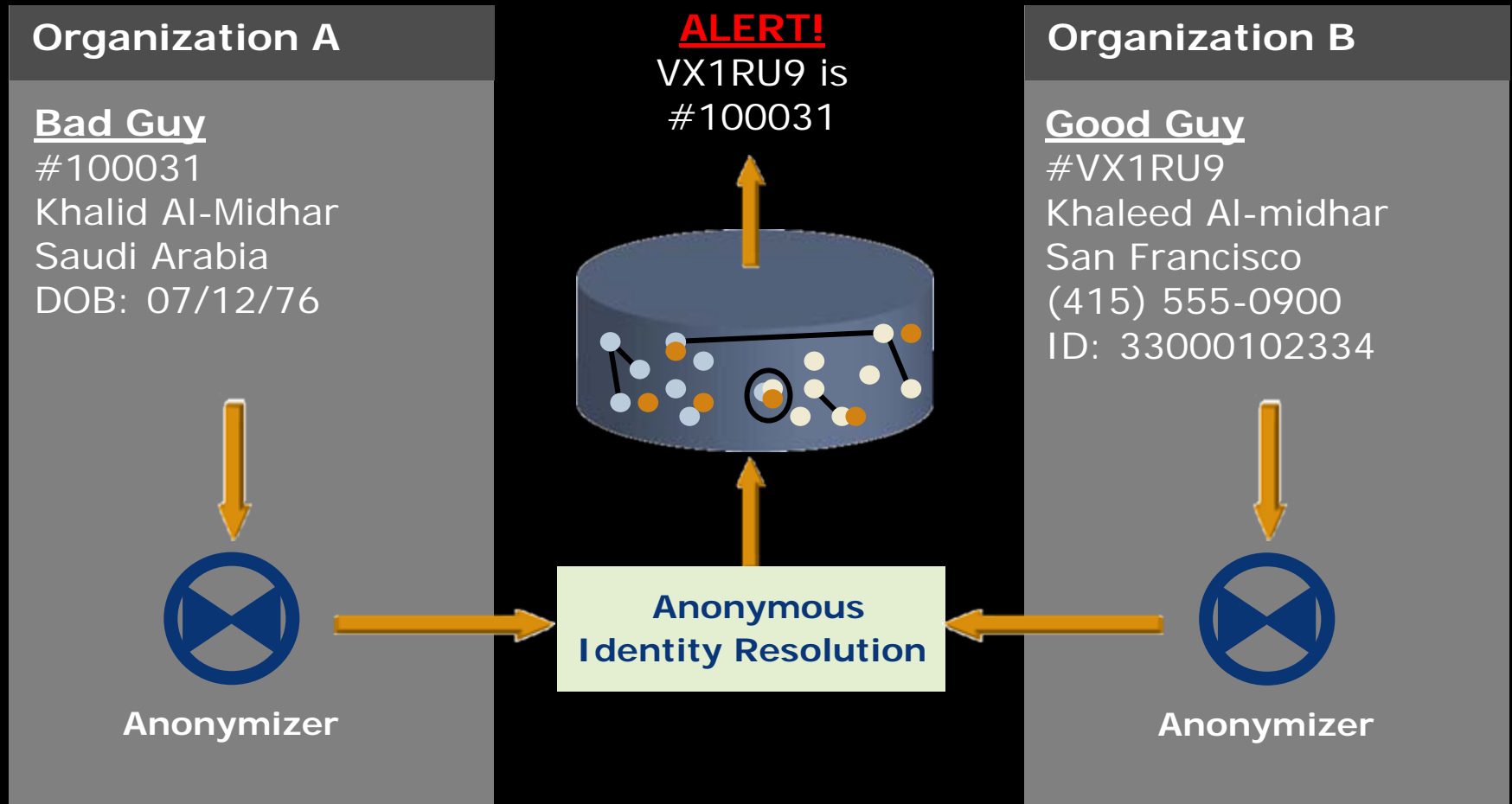
# The Anonymizer

**Name:** Rob Smith  
**DOB:** 6/7/1972

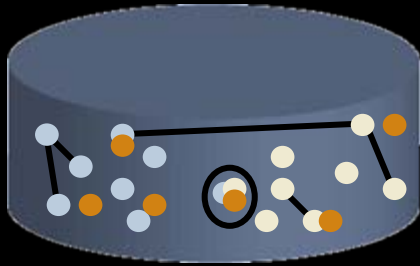


cd5dced41028cb7ea51d52a888089d73  
00c9782a552a2d09b1b85e0d0db52ef3  
7f2b6e48ea7d042bbe85e46ef2107da4  
0d06b31faa7c44682d770706640465d2  
B5e341a4b0cdf0e8de7b6f957818d746  
bd0ec72f2424729efa7baac9a636970a

# Analytics in the Anonymized Data Space



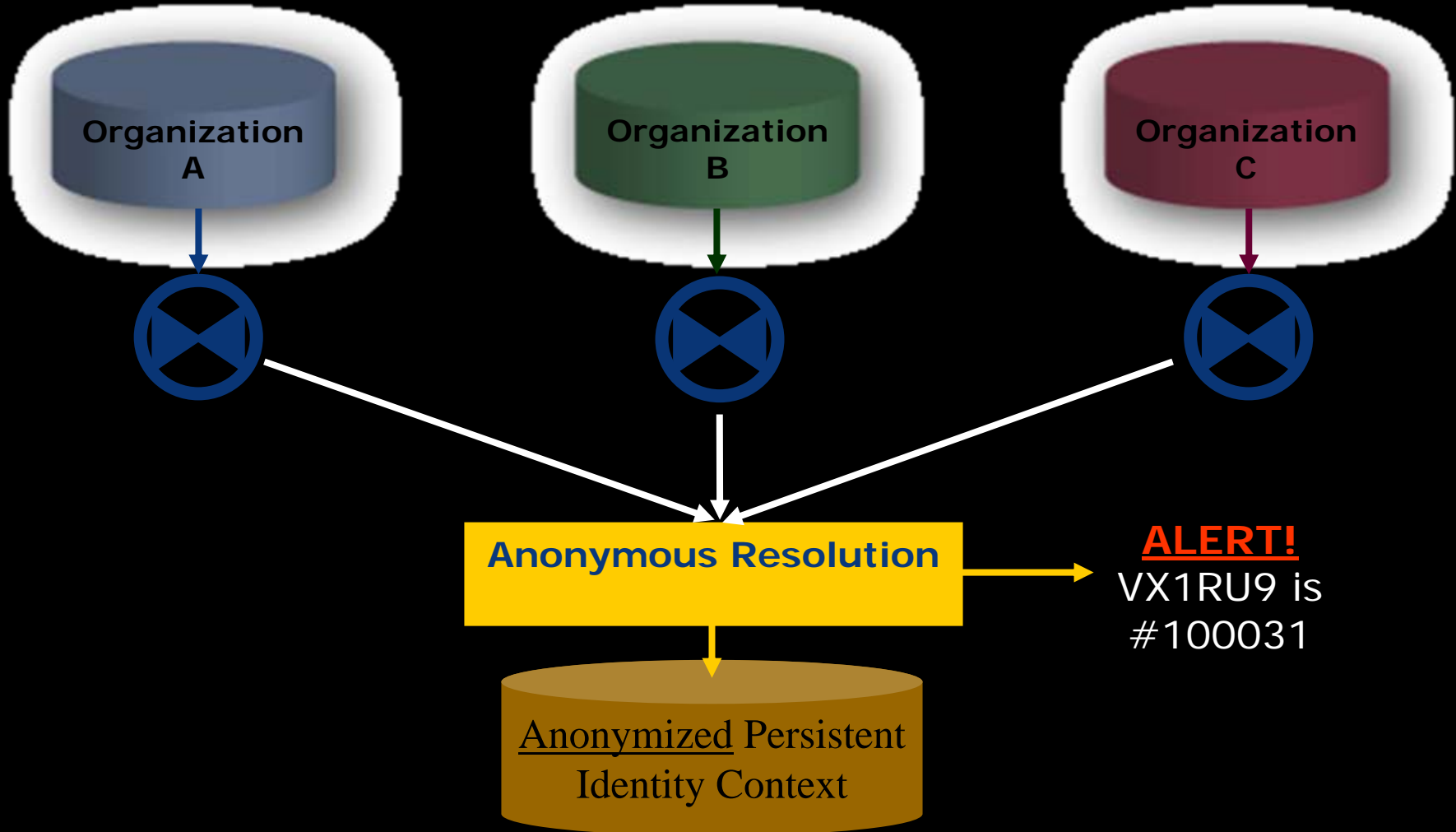
# Risk of Unintended Disclosure is Reduced



**What's in this Database?**

A-1031	Name	Normalized	cbd034409c22929518fa494f99dc9964
		Variation 1	9269bb3bc60366245144cbd5e960cfd8
	Citizenship	Normalized	b835b521c29f399c78124c4b59341691
	Date of Birth	Normalized	799709b2e5f26f796078fd815bebf724
		Variation 1	40ddba83c22acc2acaddff12c66d7adf
		Variation 2	e4310b75f2fa9595f8154411924b19b1

# Basic Concept of Operations



# Differentiators

- After de-identification, matching can still occur
- Necessary to count unique records within and across data sets
- Ability to re-identify
  - Not decrypted
  - Based on the record's attribution
  - Enables data holder to control clear text record disclosure
- Limited data repurposing, engineered in

## Attacks – Requiring Special Handling

- Dictionary attacks
- Chosen text attacks
- Statistical re-identification attacks
- Traffic analysis attacks

**CLAIM:** If you are already sharing clear text data, this new technique reduces the risk of unintended disclosure

## Different Missions, Different Remedies

- Example 1: Discovery across internal silos, with basic objective of reducing unintended disclosure
- Example 2: Two governments want to learn what entities they have in common without revealing sensitive information

# External Reference Materials

## **Heritage Foundation and Center for Democracy and Technology**

*Technologies That Can Protect Privacy as Information Is Shared to Combat Terrorism*

<http://www.heritage.org/RESEARCH/homelanddefense/lm11.cfm>

<http://www.cdt.org/security/usapatriot/20040526technologies.pdf>

## **Step toe & Johnson**

*Anonymization, Data-Matching and Privacy: A Case Study*

<http://www.step toe.com/publications/279d.pdf>

## **Wall Street Journal**

*Entrepreneur Offers Solution for Security-Privacy Clash*

<http://webreprints.djreprints.com/946631285938.html>

## **Digital ID World**

*Finding Identity in the Noise*

<http://magazine.digitalidworld.com/Mar04/Page20.pdf>

## **Markle Foundation – National Security in the Information Age Task Force**

*Second Report: Creating a Trusted Network for Homeland Security*

[http://www.markletaskforce.org/Report2\\_Full\\_Report.pdf](http://www.markletaskforce.org/Report2_Full_Report.pdf)

## Markle Foundation

“Data should be anonymized when possible; that is, the personally identifiable information should be removed, but analysts should maintain the ability to perform link analysis, queries, and entity resolution”

Report Two  
Creating a Trusted Network for Homeland Security

# State of the Union

## Anonymous Resolution In Practice

- Has achieved materially similar matching results when compared to IBM's market-leading Identity Resolution technology
- Example Customer: A non-US government classified program
  - A cross-compartment exploitation problem
- Two other companies have announced they will offer similar capabilities

## A Plausible Future

“If information can be shared in an anonymized form whereby a materially similar result can be achieved ...

why would an organization share information any other way?”

# Questions and Answers?



IBM Entity Analytic Solutions (EAS)

# **The Sound of One Hand Clapping Knowledge Discovery without Disclosure: A Koan for the Information Age**

**John Bliss, Privacy Strategist, EAS, IBM  
jblisslv@us.ibm.com**

April 2006

© 2006 IBM Corporation