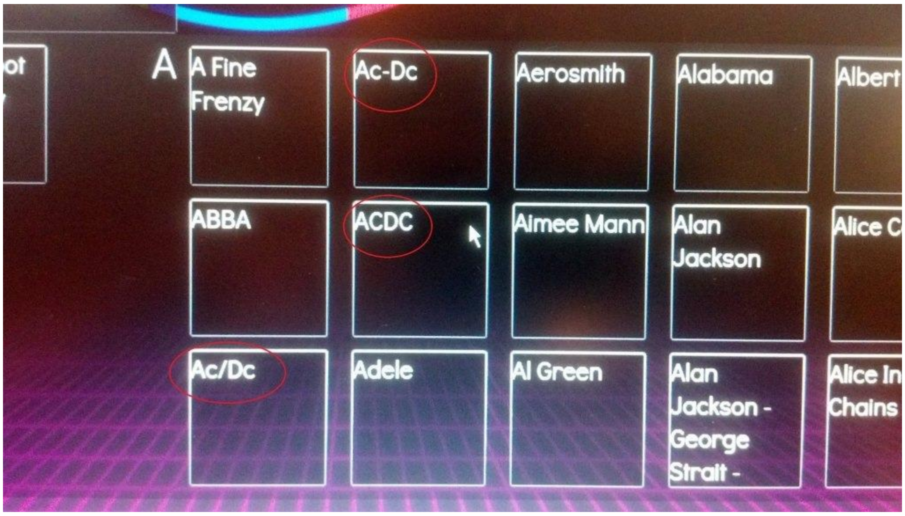
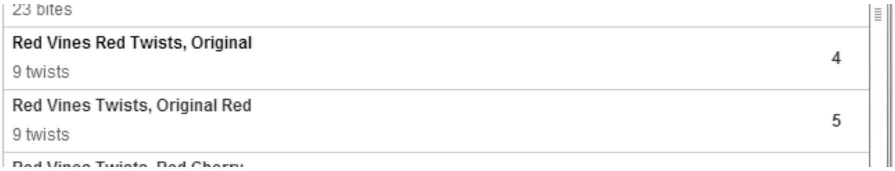
The Data Bath

Rich Dudley
http://rjdudley.com
rich@rjdudley.com
@rj_dudley

For those about to rock…



I totally ate the first ones!



23 bites

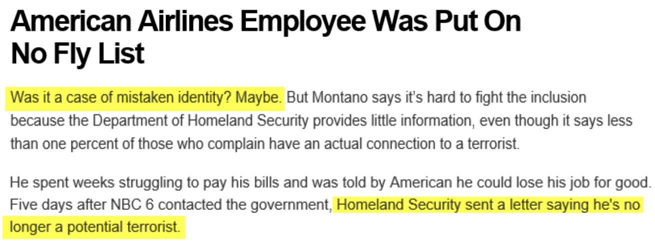
Red Vines Red Twists, Original 4
9 twists

Red Vines Twists, Original Red 5
9 twists

Red Vines Twists, Red Cherry

(from Weight Watchers mobile app)

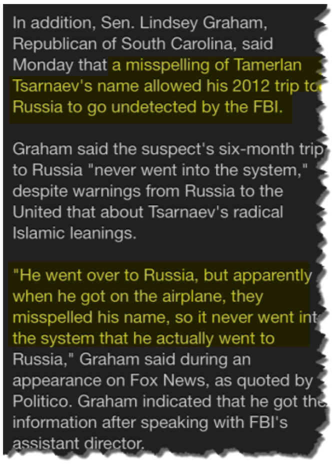
To err is a conundrum



American Airlines Employee Was Put On No Fly List

Was it a case of mistaken identity? Maybe. But Montano says it's hard to fight the inclusion because the Department of Homeland Security provides little information, even though it says less than one percent of those who complain have an actual connection to a terrorist.

He spent weeks struggling to pay his bills and was told by American he could lose his job for good. Five days after NBC 6 contacted the government, Homeland Security sent a letter saying he's no longer a potential terrorist.

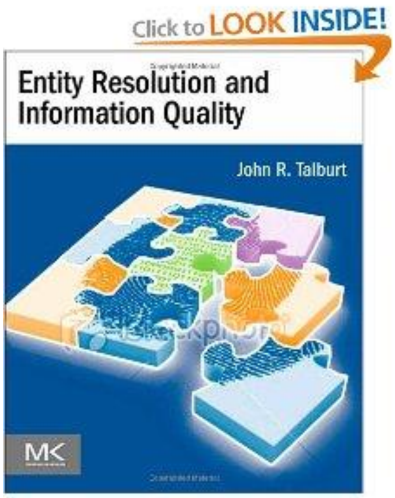


In addition, Sen. Lindsey Graham, Republican of South Carolina, said Monday that a misspelling of Tamerlan Tsarnaev's name allowed his 2012 trip to Russia to go undetected by the FBI.

Graham said the suspect's six-month trip to Russia "never went into the system," despite warnings from Russia to the United that about Tsarnaev's radical Islamic leanings.

"He went over to Russia, but apparently when he got on the airplane, they misspelled his name, so it never went into the system that he actually went to Russia," Graham said during an appearance on Fox News, as quoted by Politico. Graham indicated that he got the information after speaking with FBI's assistant director.

The Talburt Book

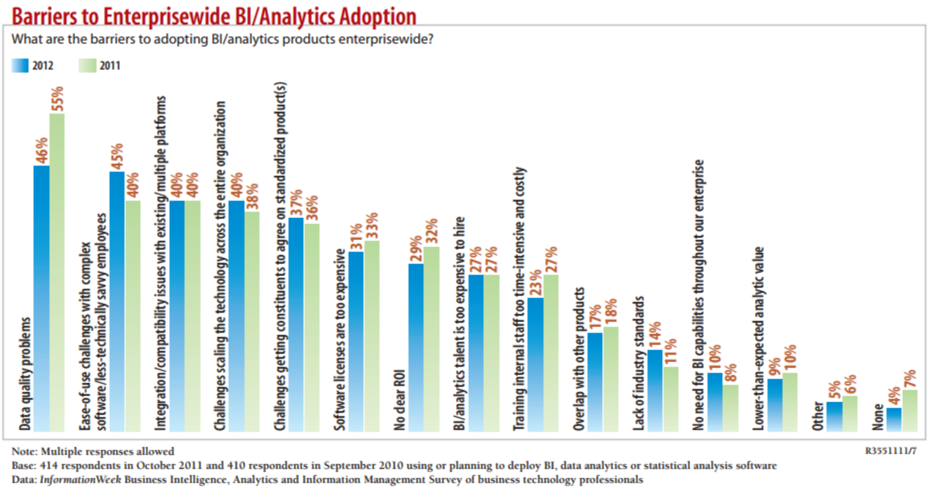


- University of Arkansas Little Rock Center for Advanced Research in Entity Resolution and Information Quality (ERIQ)
- http://pewee.ws/6q

Why so important today?

- 1993: Name and mailing address
- 2013: Facebook Username, Twitter Id, Email, multiple phone #, multiple mailing addresses
  - EMS renaming/numbering streets
- If the service is free, you're the product!

Barriers to BI Adoption


Barriers to Enterprisewide BI/Analytics Adoption

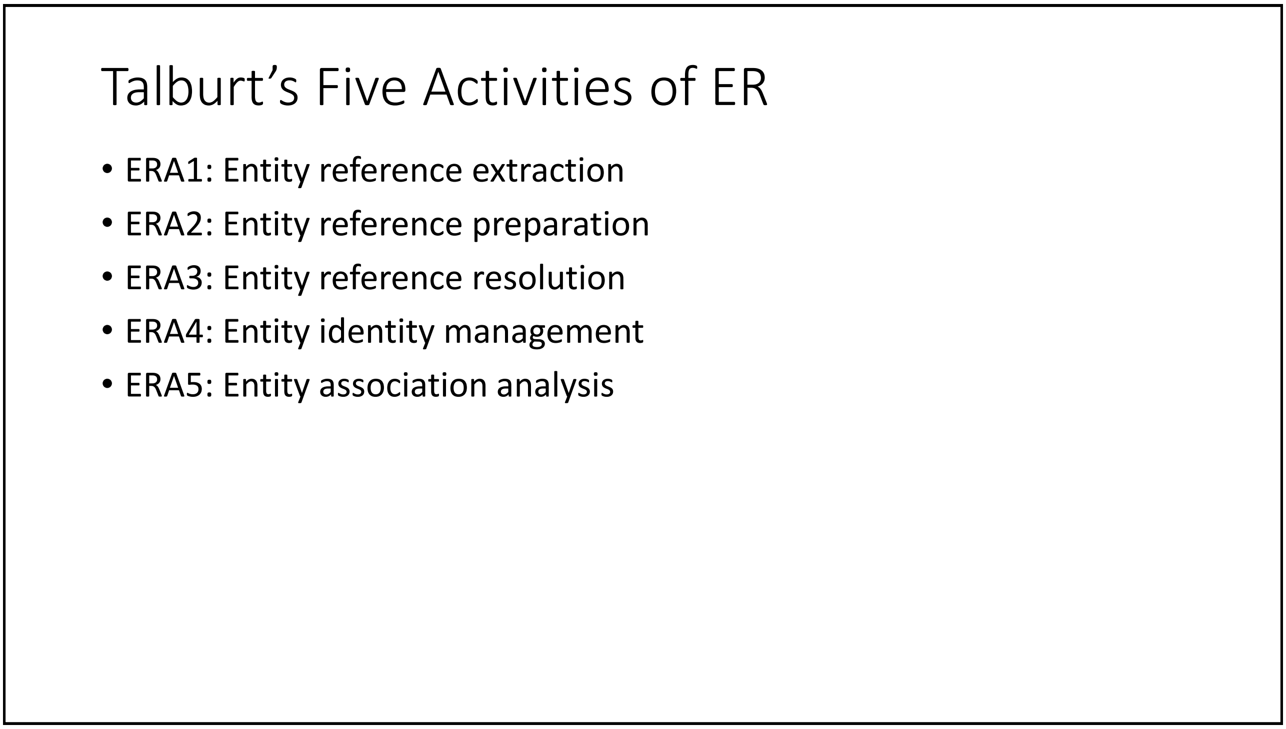
What are the barriers to adopting BI/analytics products enterprisewide?

Note: Multiple responses allowed

Base: 414 respondents in October 2011 and 410 respondents in September 2010 using or planning to deploy BI, data analytics or statistical analysis software

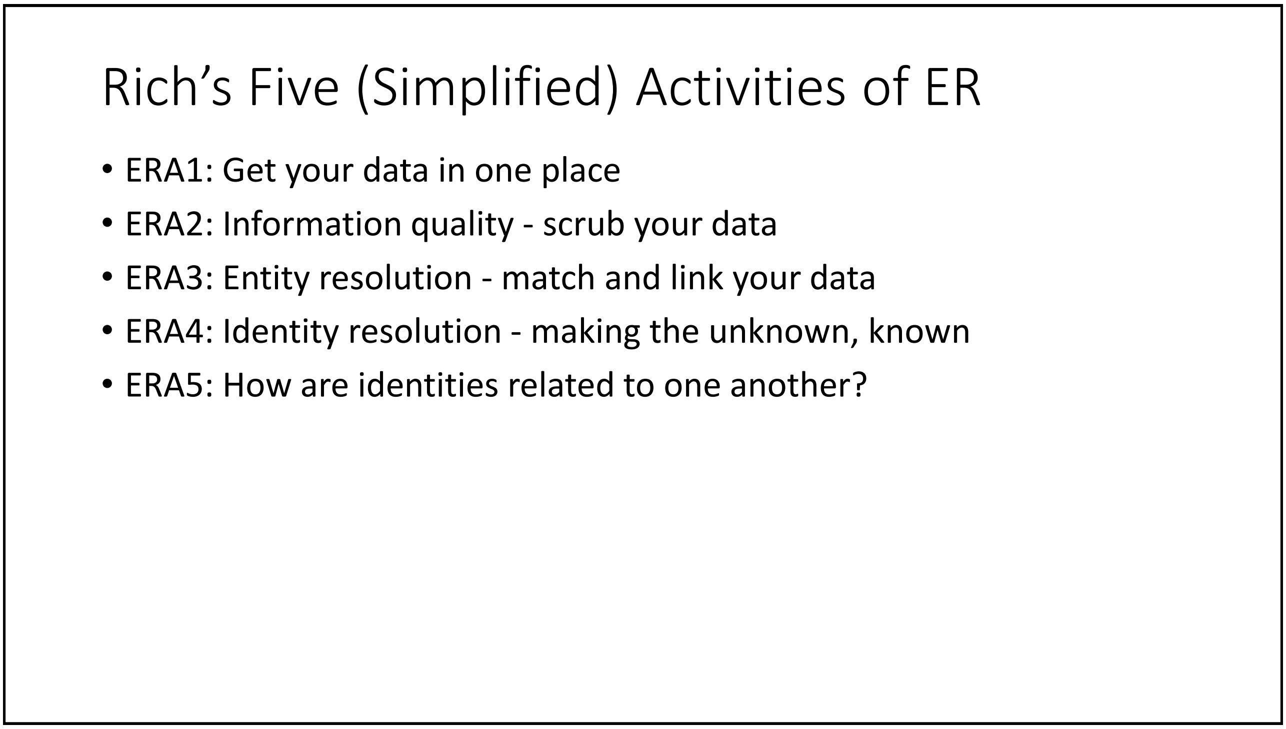
Data: *InformationWeek* Business Intelligence, Analytics and Information Management Survey of business technology professionals

"Research: 2012 BI and Information Management", Information Week, http://pewee.ws/6y



Talburt’s Five Activities of ER

- ERA1: Entity reference extraction
- ERA2: Entity reference preparation
- ERA3: Entity reference resolution
- ERA4: Entity identity management
- ERA5: Entity association analysis



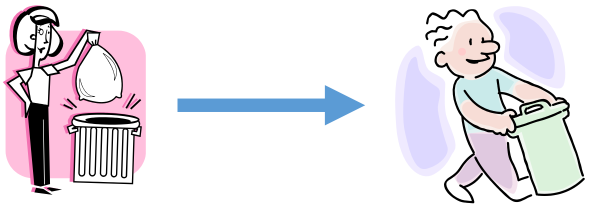
Rich’s Five (Simplified) Activities of ER

- ERA1: Get your data in one place
- ERA2: Information quality - scrub your data
- ERA3: Entity resolution - match and link your data
- ERA4: Identity resolution - making the unknown, known
- ERA5: How are identities related to one another?

ERA2: Information Quality

“The degree of completeness, accuracy, timeliness, believability, consistency, accessibility and other aspects of reference data can affect the operation of ER processes and produce better or worse outcomes.”

Garbage In, Garbage Out



Information Quality

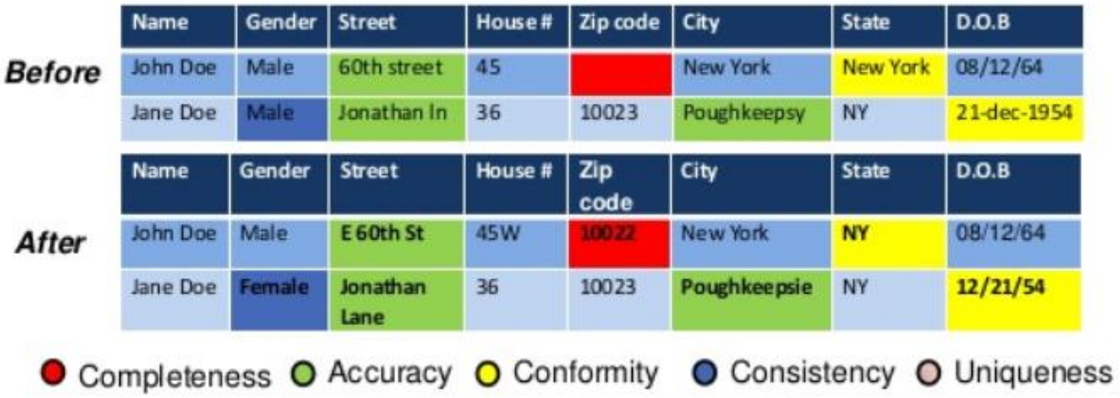
- Data quality
- Master data management
- Data governance

- What’s sauce for the goose may not be sauce for the gander
- IQ management is not just the responsibility of the IT department
- Success depends on demonstrating business value and having support at the top

Scrub-a-dub-dub

- Encoding
  - UTF-8 to ASCII
- Conversion
  - varchar to nvarchar
- Standardization
  - “123 Main Street” to “123 Main ST.”
- Correction
  - Confirming ZIP codes and area codes
- Bucketing
  - Age 25-24 = Cohort A
  - Age 35-44 = Cohort B
- Bursting
  - “123 Main Street” into “123”, “Main”, “ST.”
- Validation (are the data rational)
  - Patient’s body temp was not 350F
- Enhancement (adding additional information)
  - Geocoding addresses

Common Data Quality Issues



Before

Name	Gender	Street	House #	Zip code	City	State	D.O.B
John Doe	Male	60th street	45	🟥	New York	New York	08/12/64
Jane Doe	Male	Jonathan ln	36	10023	Poughkeepsy	NY	21-dec-1954

After

Name	Gender	Street	House #	Zip code	City	State	D.O.B
John Doe	Male	E 60th St	45W	10022	New York	NY	08/12/64
Jane Doe	Female	Jonathan Lane	36	10023	Poughkeepsie	NY	12/21/54

🔴 Completeness 🟢 Accuracy 🟡 Conformity 🔵 Consistency ⚪ Uniqueness

“Master Data and Data Quality Management in SQL Server 2012”, Mark Gschwind, http://pewee.ws/6z

ERA3: Entity Resolution

“Entity resolution is the process of probabilistically identifying some real thing based on a set of possibly ambiguous clues”

Baseline

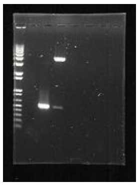
- An *entity* is a real world object (person, place or thing)
- *Resolution* is deciding if two references point to the same or different entities
- An *identity* is a known entity
- *Identity attributes* are a set of defining characteristics that when taken together distinguish one entity from another
  - Cars – same make, model, color, VIN differentiates

What Distinguishes Identities

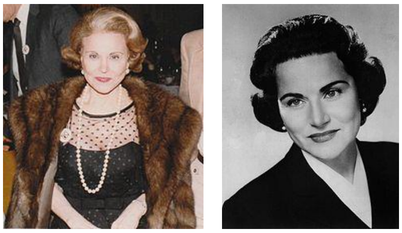
Rich Dudley

September 28, 1967

DNA is the ultimate identity attribute…



… or is it?



Demo: R-Swoosh

Using direct matching

R-Swoosh

D

ID	FirstName	LastName	Address	City	State	Zip	Source
1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB
2	Han	Solo	1600 Pennsylvania Ave	Springfield	RI	12345	LI
3	Jenny	Jani	867 E Rt 5309	Springfield	UT	76543	Raffle
4	H	Solo	123 West Main Street	Springfield	RI	12345	Local
5	Jenny	Jani	867 East Rte 5309	Springfield	UT	76543	FB
6	Arthur	Dent	42 Vogon Bypass	Springfield	IL	60543	FB
7	Jenni	Jani	867 East Rte 5309	Springfield	UT	76543	Local

ER(R)

ID	FirstName	LastName	Address	City	State	Zip	Source
1							

R-Swoosh: Match Functions

X	1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB
Y	1							

- Is M(x,y) == true?
    - M(x.FirstName === y.FirstName)
    - M(x.LastName === y.LastName)
    - …

X	1							
Y	1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB

R-Swoosh: Lather, Rinse, Repeat

X	2	Han	Solo	1600 Pennsylvania Ave	Springfield	RI	12345	LI
Y	1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB

X	3	Jenny	Jani	867 E Rt 5309	Springfield	UT	76543	Raffle
Y	1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB
	2	Han	Solo	1600 Pennsylvania Ave	Springfield	RI	12345	LI

X	4	H	Solo	123 West Main Street	Springfield	RI	12345	Local
Y	1	Hope	Solo	123 W. Main St.	Springfield	RI	12345	FB
	2	Han	Solo	1600 Pennsylvania Ave	Springfield	RI	12345	LI
	3	Jenny	Jani	867 E Rt 5309	Springfield	UT	76543	Raffle

Determining equivalence

- Direct matching
- Transitive equivalence
- Association analysis
- Asserted equivalence

Direct Matching

- Exact match (deterministic matching)
  - John Doe != JOHN DOE
  - Success depends on ERA2
- Numerical difference
  - How close can the numbers be? Server time differences
- Approximate syntactic match
  - Approximate String Matching (ASM)
  - SimMetrics library
- Approximate semantic
  - Synonym tables
  - “Towing and Recovery” = “Wrecker service”
- Derived match codes
  - Hash codes
  - Soundex
  - Blocking

Demo: SimMetrics Library

18 algorithms

Java, C# and SQL Server

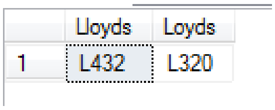
Results are *p* values

Demo: Derived Match Codes

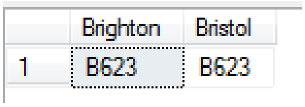
Soundex

- Replace consonants with digits as follows (but do not change the first letter):
  - b, f, p, v => 1
  - c, g, j, k, q, s, x, z => 2
  - d, t => 3
  - l => 4
  - m, n => 5
  - r => 6
- Collapse *adjacent* identical digits into a single digit of that value.
- Remove all non-digits after the first letter.
- Return the starting letter and the first three remaining digits. If needed, append zeroes to make it a letter and three digits.

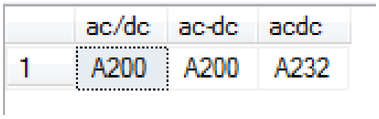
```
select SOUNDEX('Lloyds') as "Lloyds",
       SOUNDEX('Loyds') as "Loyds"
```



	Lloyds	Loyds
1	L432	L320



	Brighton	Bristol
1	B623	B623



	ac/dc	ac-dc	acdc
1	A200	A200	A232

Blocking Strategy

- Blocking = partial match codes, using small blocks of the overall data

- Real life example
  - ZIP code = 16001
  - Previous house owner = Kathryn “Kathy” Davis
  - Current: Kathleen “Kathy” Dudley

  - Blocking Code = KathD16001

  - Result: After 10 years, we still get all kinds of solicitations for Kathryn Davis

- Photo of blocked junk mail address label?

Transitive Equivalence

- Relies on intermediate entities, so several passes may be necessary

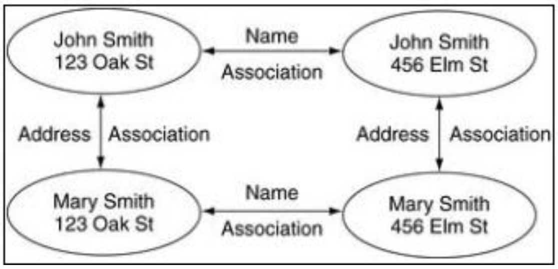
1. Mary Smith, 123 Main St, 555-1234
2. Mary Smith, 456 Elm St, 555-1234
3. Mary Smith, 456 Elm St, null

- 1 != 3, but 1 == 2 and 2 == 3 :: 1 == 3

Association Analysis

- Considering multiple relationships and making multiple decisions at the same time

Name	Address
Mary Smith	123 Oak St
Mary Smith	456 Elm St
John Smith	123 Oak St
John Smith	456 Elm St



Asserted Equivalence

- Knowledge-based linking, “we already know these are the same”

- Cassius Clay, 1960 Gold Medal, Boxing (Light Heavyweight)
- Muhammad Ali, 1964 World Champion, Boxing (Heavyweight)

- MS Carriers, purchased by Swift Transportation

- We don’t want to recalculate historical data, but we need a link to the current identity

Putting it all together

- Show match functions of different natures
  - Bursted + standardized address
  - Blocked name+zip – SimMetrics?

X	4	H	Solo	123	West	Main	St	Springfield	RI	12345	HSolo12345
Y	1	Hope	Solo	123	West	Main	St	Springfield	RI	12345	HopeSolo12345

Summary

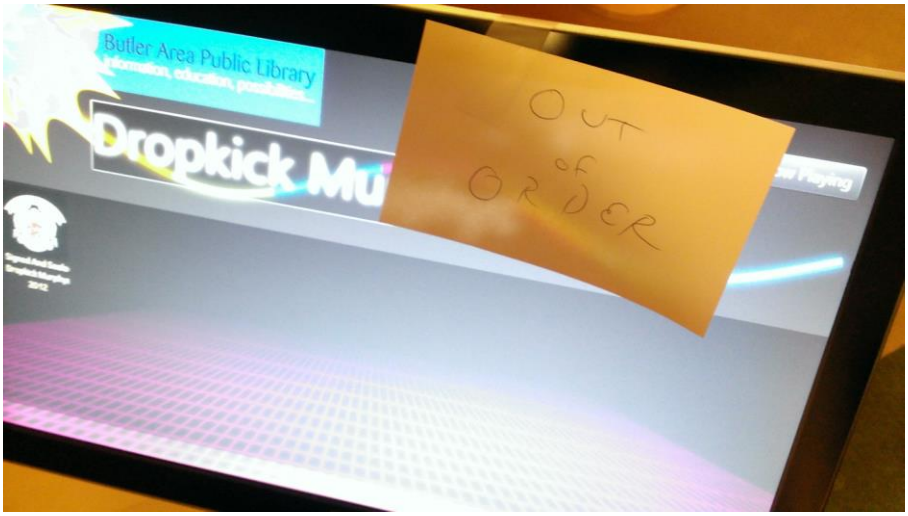
- Information quality and entity resolution go hand-in-hand
- Both are growing in importance
- Identity attributes and level of resolution vary from need to need
- R-Swoosh + direct matching techniques are fairly universal

References

- Entity Resolution and Information Quality, John R. Talburt, Amazon link: http://pewee.ws/6q
- How to solve common Data Quality Problems using Data Quality Services (Part 1), Paras Doshi, http://pewee.ws/6t
- “Research: 2012 BI and Information Management”, Douglas Henschen, Information Week, http://pewee.ws/6y
- “Master Data and Data Quality Management in SQL Server 2012”, Mark Gschwind, http://pewee.ws/6z

SimMetrics References

- Java and C# (some ported to F#)
  - http://sourceforge.net/projects/simmetrics/files/
- Installing into SQL Server
  - http://anastasiosyal.com/POST/2009/01/11/18.ASPX
- Originator's current page:
  - http://www.aktors.org/technologies/simmetrics/
- Archives of original project pages:
  - http://web.archive.org/web/20081230184321/http://www.dcs.shef.ac.uk/~sam/simmetrics.html
  - http://web.archive.org/web/20081224234350/http://www.dcs.shef.ac.uk/~sam/stringmetrics.html


Butler Area Public Library
information, education, possibilities...
Dropkick Mu
Now Playing
OUT
of
ORDER