

Data Finds Data

Lisa Sokol

lsokol@us.ibm.com

October 2010

Need Scalable, Flexible and Smart Systems

- Traditional data systems miss things
 - Bank loyalty – Sends out credit card notices indicating your credit limit has been raised
 - You just declared bankruptcy
 - Customer service – Book that you wanted has been published and is available for purchase
 - You find out too late – the book run is sold out
 - Security Clearances – Re-investigate every 5 years
 - You commit a felony the following month

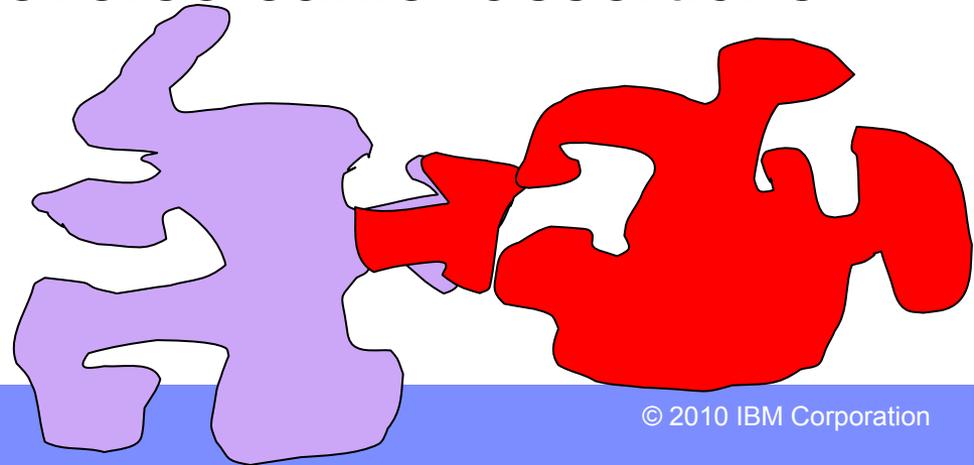
Traditional Search

- Requires the re-asking of questions
- Misses the discovery of aspects and relationships between user defined entities such as people, organizations, people, places, things, shapes, etc

“Data Finds Data” Paradigm

- Questions persist as data
 - Cumulative data creates context
- Relevance detection as data enters the system
 - Assessment of new data based on situation awareness
 - Relevance is defined by rules
 - Automatically notify users about new insights
- Real-time assertions are made when entities are the same or related
- New observations can reverse earlier assertions

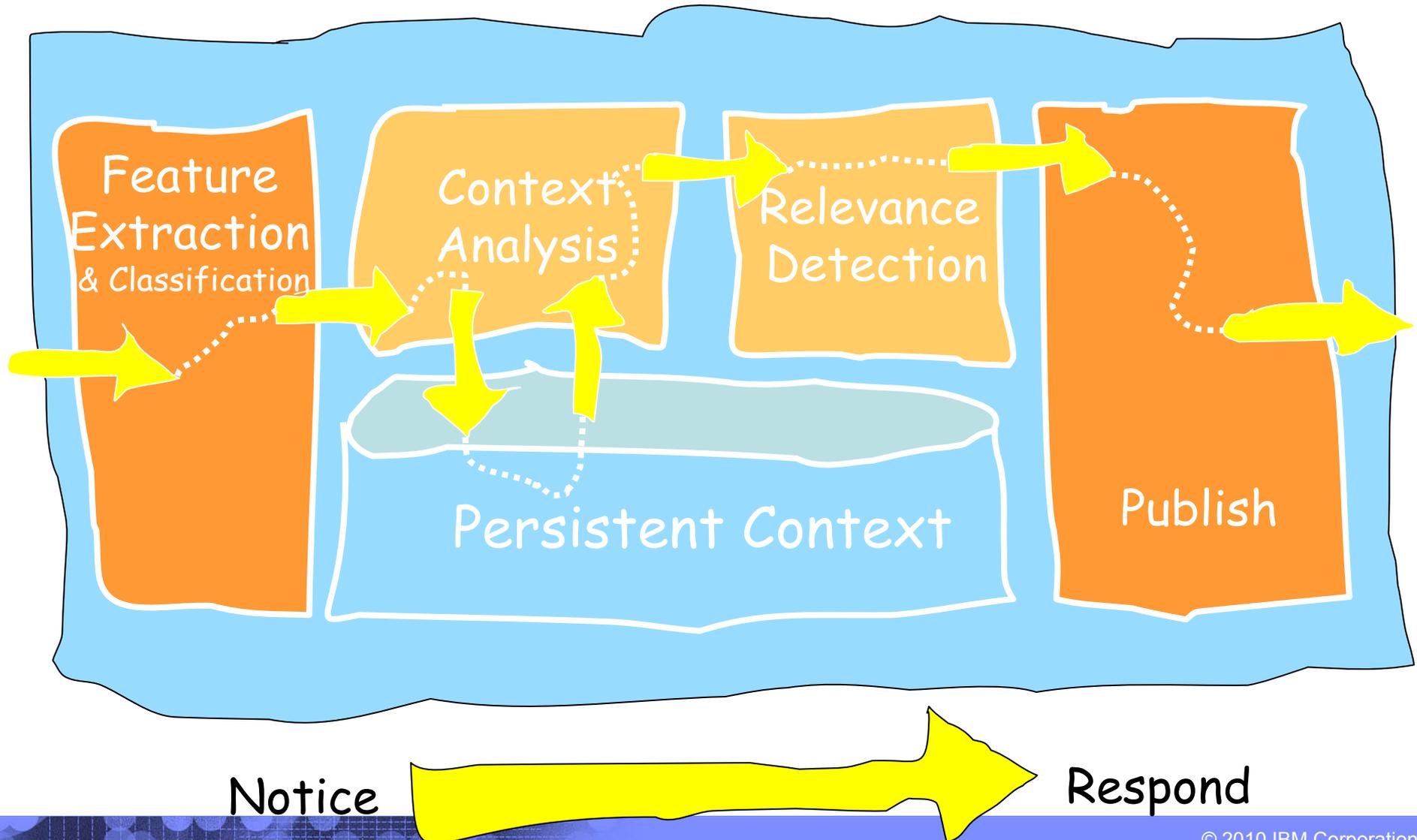
**New Puzzle
piece = New
piece of data**



Relevance Analysis

- Ability to make assertions (same or related) about new observations
 - New entity
 - Part of a known entity and perhaps we add to the entity's identity data
 - New observation relates two or more entities together
- ☺ Assertions are used to transform data for analytics
 - If counting is off, analytics are off
 - Predictive modeling data view needs to exploit assertion data

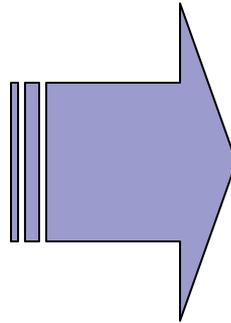
Sensemaking: Major Moving Parts



Corruption – Feature Extraction

Casino Employee

Mark R Johnson
123 Main Street,
Alexandria VA
Phone 702 111 2222
SSN 120 44 1515



Gives a Prize to

Miranda Johnson
621 Hacienda Way,
Alexandria VA
Phone 702 333 4444

**Suspicious?- Same name but
no overt relationship**

Casino Employee fills in a change of Address Form

Where have I seen that address before?

Old Data

Mark R Johnson
123 Main Street,
Alexandria VA
SSN 120 44 1515

New Address Form

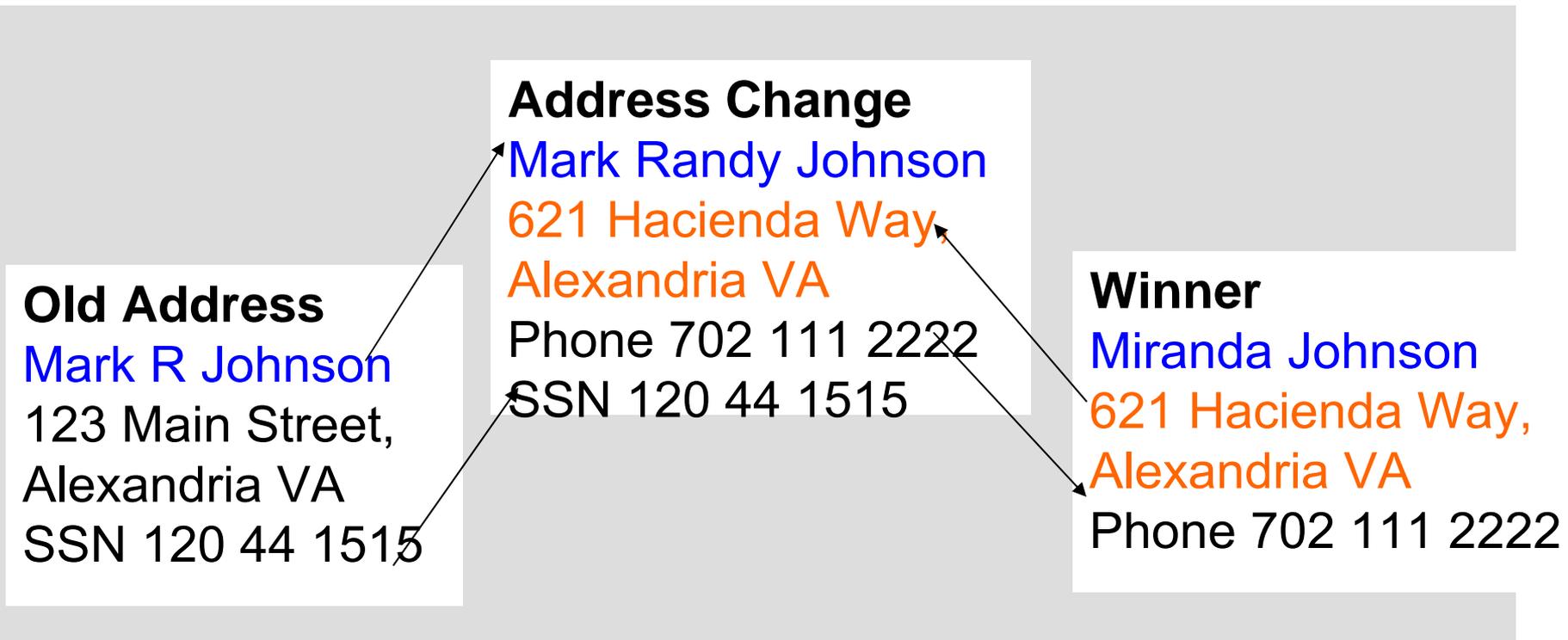
Mark Randy Johnson
621 Hacienda Way,
Alexandria VA
Phone 702 111 2222
SSN 120 44 1515

Traditional Federated Searches Can't Deliver Enterprise Discoverability

- Relevance discoverability requires that new observations be evaluated against ALL prior observations
 - Recursion is required and is computationally expensive
 - Interestingness of the discovery is context dependent
 - Who to notify is mission dependent

Our Casino Employee fills in a change of Address Form - Context

Familiar address— but not in the HR database



Context Analysis

- The ability to recognize the formation of relevance/insight
- Did the system learn something that matters?
 - Detect pre-specified patterns of interest, e.g., Good guys know bad guys

Good Guy - 1

Mark R Johnson

123 Main Street,
Alexandria VA

Phone 702 111 2222

SSN 120 44 1515

Single Identity

Record 3

Mark R Johnson

345 Broad St
Alexandria Va

Phone 702 111 2222

SSN 120 44 1515

KNOWS

Bad Guy – Record 2

Miranda Johnson

345 Broad St
Alexandria Va

Phone 702 111 2222

Passport USA 123456

Persistent Context

- The ability to accumulate and persist this asserted context
 - New knowledge embedded in system
 - Incremental learning

Record 1

Mark R Johnson

123 Main Street,
Alexandria VA

Phone 702 111 2222

SSN 120 44 1515

Record 2

Mark Randy Johnson

123 Main Street

Alexandria VA

Phone 702 111 2222

Passport USA 123456

Record 3

Mark R Johnson SR

123 Main Street

Alexandria VA

Passport USA 123456

SSN 345 44 13456

Reversal- Relevance Analysis Correction

- The ability to recognize when new observations reverse earlier assertions

Single Identity

Record 1

Mark R Johnson

123 Main Street,
Alexandria VA

Phone 702 111 2222

SSN 120 44 1515

Record 2

Mark Randy Johnson

123 Main Street

Alexandria VA

Phone 702 111 2222

Passport USA 123456

OOPS

Record 3

Mark R Johnson SR

123 Main Street

Alexandria VA

Passport USA 123456

SSN 345 44 13456

Next Generation – Batch and Streams



Feature & Entity
Extraction

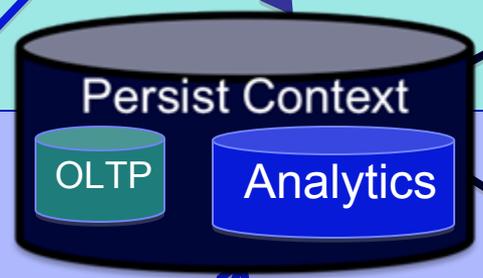
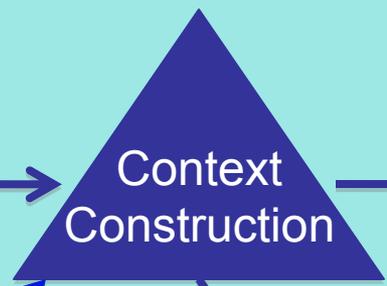
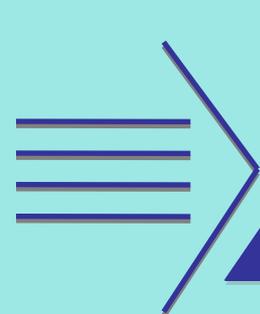
Context
Accumulation

Situational
Awareness

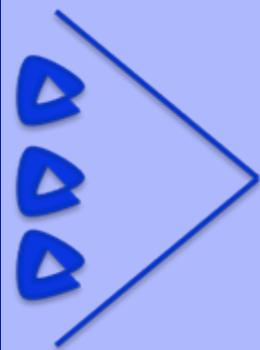
Real-Time Streams

Insight!

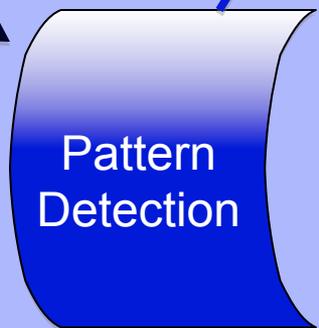
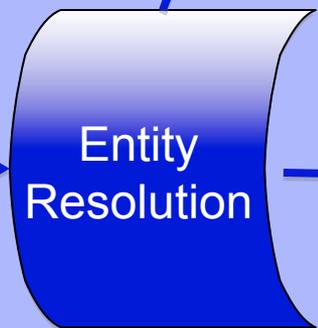
- Person
- Events
- System
- Sensor
- Mart
- Etc.



Batch



Overflow



Stuff to
watch
for

Puzzle Pieces

Puzzles