Big Data Flows
vs. Wicked Leaks

Jeff Jonas, IBM Distinguished Engineer
Chief Scientist, IBM Entity Analytics
JeffJonas@us.ibm.com

December 1, 2010

Background

- Early 80's: Founded Systems Research & Development (SRD), a custom software consultancy
- 1989 – 2003: Built numerous systems for Las Vegas casinos including a technology known as Non-Obvious Relationship Awareness (NORA)
- 2005: IBM acquires SRD, now chief scientist of IBM Entity Analytics
- Personally designed and deployed +/- 100 systems, a number of which contained multi-billions of transactions describing 100's of millions of entities
- Today: My focus is in the area of 'sensemaking on streams' with special attention towards privacy and civil liberties protections
  - Markle Foundation, Member, Task Force on National Security
  - EPIC, Member, Advisory Board

Data Volumes Exploding

"Every two days now we create as much information as we did from the dawn of civilization up until 2003."

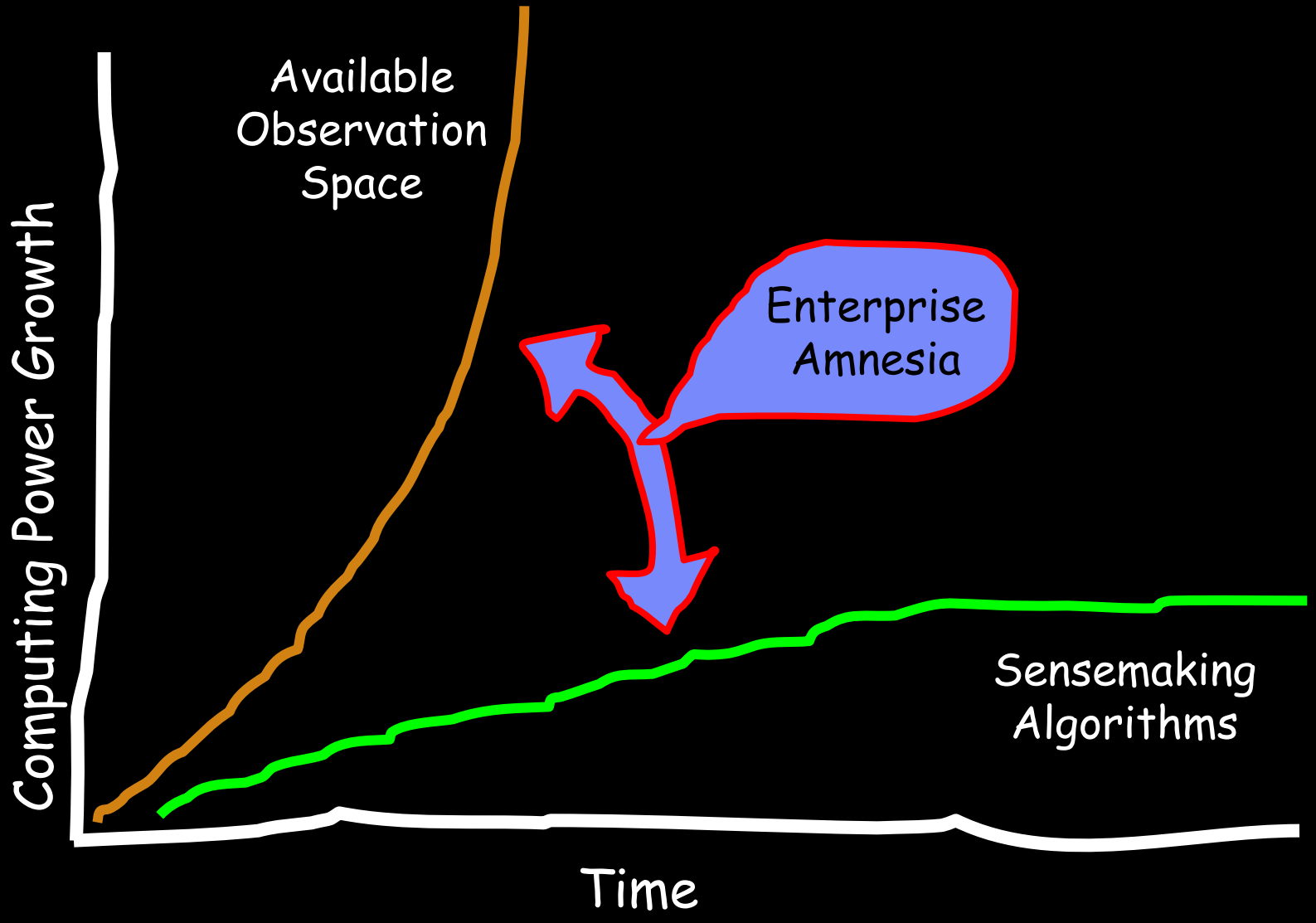
-Eric Schmidt, CEO Google

Big Data Flows: How Many Copies?

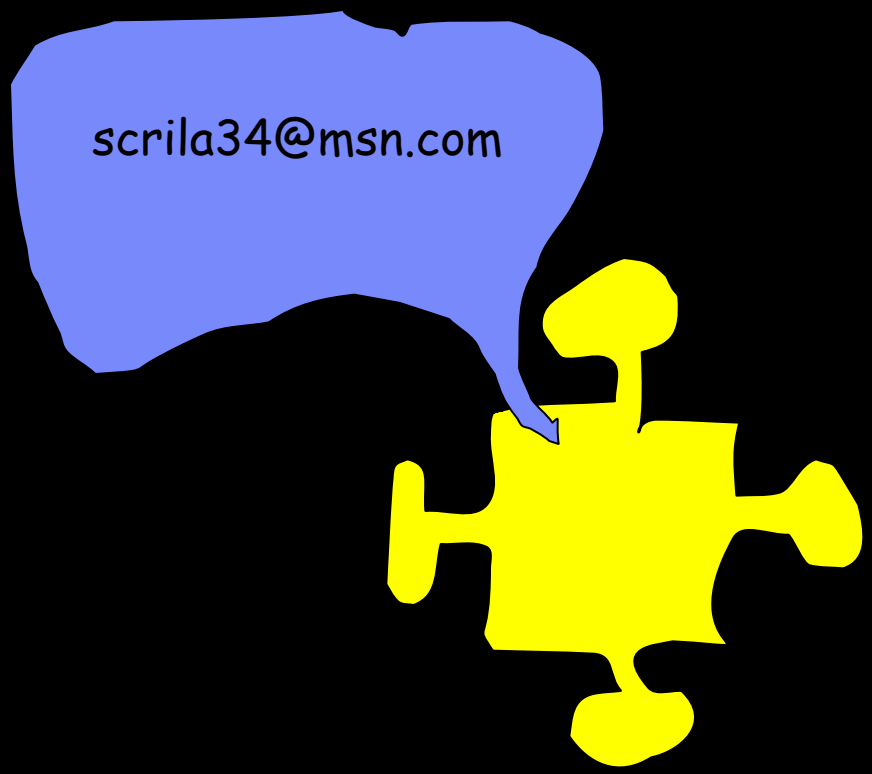
- Blog Post: How Many Copies of Your Data? Is Somewhat Like Asking: How Many Licks to the Center of the Tootsie Pop?
- Often, at minimum, 144 copies
  - Backups
  - Internal transfers
    - Other operational systems
    - Operational data stores
    - Data warehouses
    - Data marts
    - Testing systems
    - Training systems
    - Their backups
  - External transfers (information sharing partners)
    - And then their entire ecosystem (from warehouses to backups)
    - And their information sharing partners
- Sometimes 10,000's of copies

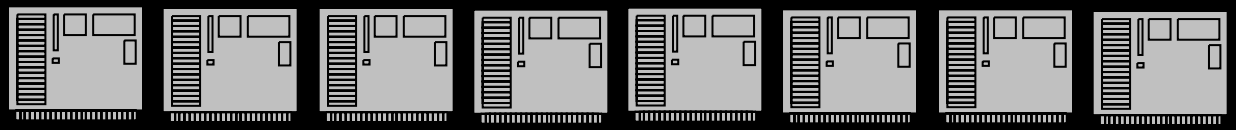
What's driving information sharing and big data?

Organizations Are Getting Dumber

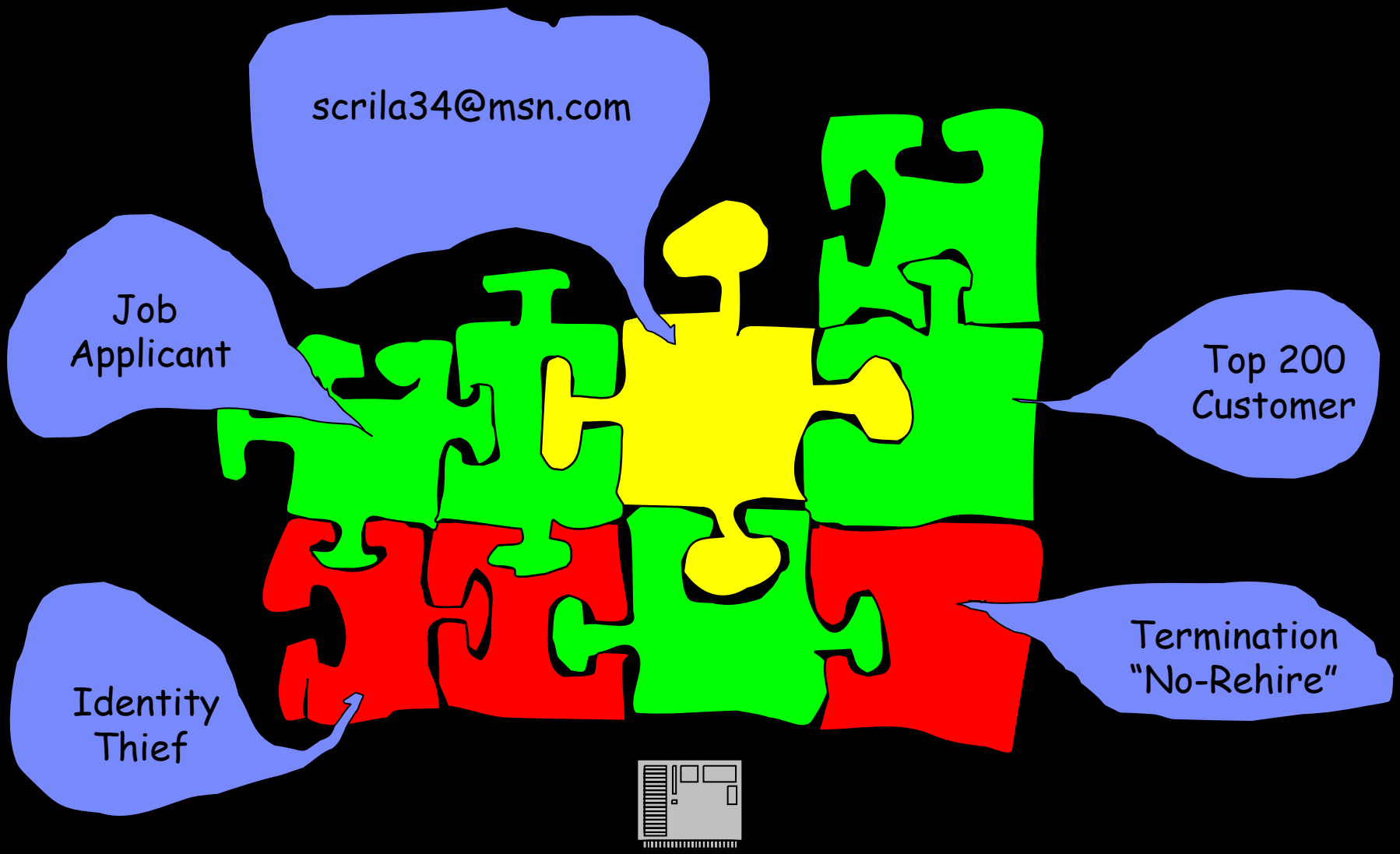


No Context





Information in Context … and Accumulating



Demonstration

Is This Voter Deceased?

VOTER
George F Balston
YOB: 1951 D/L: 4801
13070 SW Karen Blvd Apt 7
Beaverton, OR 97005
Last voted: 2008

DECEASED PERSON
George Balston
YOB: 1951 SSN: 5598
DOD: 1995

When it comes to best practices in voter matching, if only a name and year of birth match, this is insufficient proof of a match. Many different people in the U.S. share a name and year of birth.

Human review is required.

Unfortunately, there are thousands and thousands of cases just like this and state election offices don't have the staff (or budget) to manually review such volumes.

Now Consider This Tertiary DMV Record

VOTER
George F Balston
YOB: 1951 D/L: 4801
13070 SW Karen Blvd Apt 7
Beaverton, OR 97005
Last voted: 2008

DECEASED PERSON
George Balston
YOB: 1951 SSN: 5598
DOD: 1995

DMV
George F Balston
YOB: 1951 SSN: 5598 D/L: 4801
3043 SW Clementine Blvd Apt 210
Beaverton, OR 97005

The DMV record contains enough features to match both the voter (name, year of birth and driver's license) and/or the deceased persons record (name, year of birth and SSN). For the sake of argument, let's say it matches the voter best.

Is This Voter/DMV Person Deceased?

VOTER
George F Balston
YOB: 1951 D/L: 4801
13070 SW Karen Blvd Apt 7
Beaverton, OR 97005
Last voted: 2008

DMV
George F Balston
YOB: 1951 SSN: 5598 D/L: 4801
3043 SW Clementine Blvd Apt 210
Beaverton, OR 97005

DECEASED PERSON
George Balston
YOB: 1951 SSN: 5598
DOD: 1995

The voter/DMV record now shares a name, year of birth and SSN with the deceased person record. In voter matching best practices, this evidence would be sufficient to make a determination that this voter is in fact deceased. This case no longer needs human review.

Context Accumulates

VOTER
George F Balston
YOB: 1951 D/L: 4801
13070 SW Karen Blvd Apt 7
Beaverton, OR 97005
Last voted: 2008

DMV
George F Balston
YOB: 1951 SSN: 5598 D/L: 4801
3043 SW Clementine Blvd Apt 210
Beaverton, OR 97005

DECEASED PERSON
George Balston
YOB: 1951 SSN: 5598
DOD: 1995

As features accumulate it becomes easier to match future identity records.

As events and transactions accumulate – detection of relevance improves.

Here we can see George who died in 1995 voted in 2008.

Flows vs. Leaks

Flows vs. Leaks

- Most flows are by design
  - Better context
  - Better prediction
- Unintended disclosure flows
  - External: e.g., Cyber
  - Internal: e.g., Insider threats
- Recent WikiLeaks
  - Devastating consequences
  - Could have been worse. What if the leaks were not made public rather selectively and quietly passed around to others?

Wicked Leaks, Prediction

- Sudden change of leak cadence – so much, so fast, no considered intolerable
  - May result in new harsh laws and prosecutions
- Receiving stolen property (e.g., data) and benefiting from it … met with severe pursuit and prosecution
  - Stealing party
  - The information exchange points (e.g., Wikileaks)
  - Those publishing the contents (e.g., the media)
- What if? Could result in less transparency, less accountability

Protecting Big Data from Wicked Leaks

- Central indexes
  - fewer copies of the data moved; easier to control and audit usage
- Analytics in the anonymized data space
  - data anonymization before transfer, reducing the risk of unintended disclosure
- Immutable (tamper resistant) audit logs
  - to help prove the system is being used within policy and law
- Real-time active audits
  - to evaluate the actions of authorized users

Big Data Flows
vs. Wicked Leaks

Jeff Jonas, IBM Distinguished Engineer
Chief Scientist, IBM Entity Analytics
JeffJonas@us.ibm.com

December 1, 2010