DEFRAG 2010

Enterprise Amnesia
vs. Enterprise Intelligence

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

November 18, 2010

Big Data – New Physics

- More Data: Better prediction
  - Less false positives
  - Less false negatives
- More Data: Bad data good
- More Data: Less compute effort

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)
- 2001: First technology funded by In-Q-Tel, CIA's venture capital arm
- 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

Sensemaking on Streams

1) Evaluate new information against previous information … as it arrives.

2) Determine if what is being observing is relevant.

3) Deliver this relevant, actionable insight fast enough to do something about it … as it's happening.

4) Do this with sufficient accuracy and scale to really matter.

Sensemaking Blunders

Sleep Interruption at a Five Star Hotel

After checking in at about 2am and asking for a 10am wake-up call and a 10:30am breakfast service, the maid knocks on the door at 8am to see if the room is available for cleaning.

Large Retailer Hires Known Criminals

A large retailer discovered that two out of every 1,000 people they hire have previously arrested for stealing from them … at the same store.

Bank Trying to Win Business They Have

After refinancing the house with the bank, the bank proceeds to call every week for six weeks trying to win the business … business they had already won.

National Security Sensemaking Disasters

9/11

Discovered after the fact: two known terrorists were admitted into US.

Christmas Day Bomber

Discovered after the fact: Abdulmutallab possessed a multi-entry VISA while at the same time was on the terrorist watch list.

State of the Union:
Enterprise Amnesia

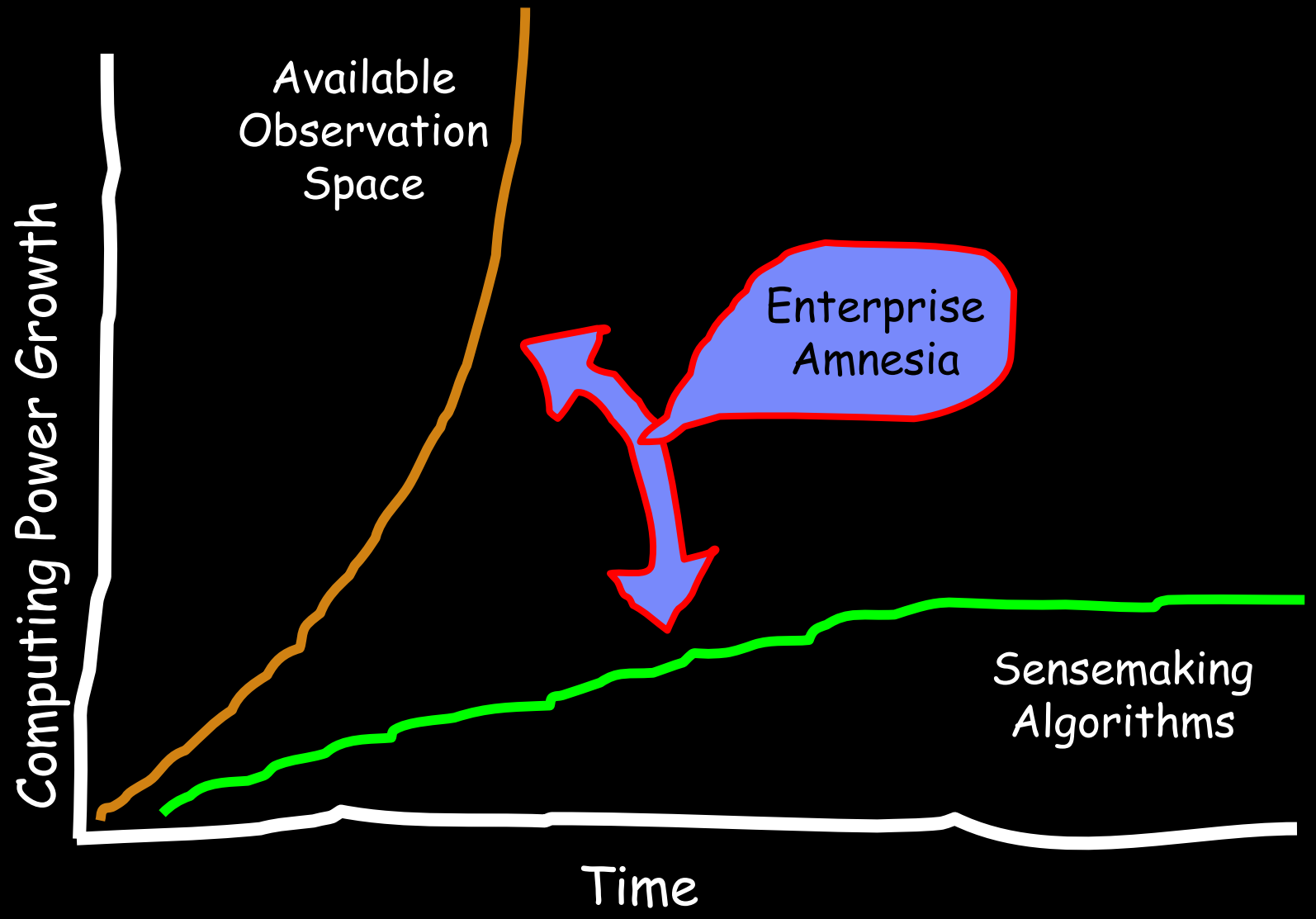
Amnesia, definition

A defect in memory, especially resulting from brain damage.

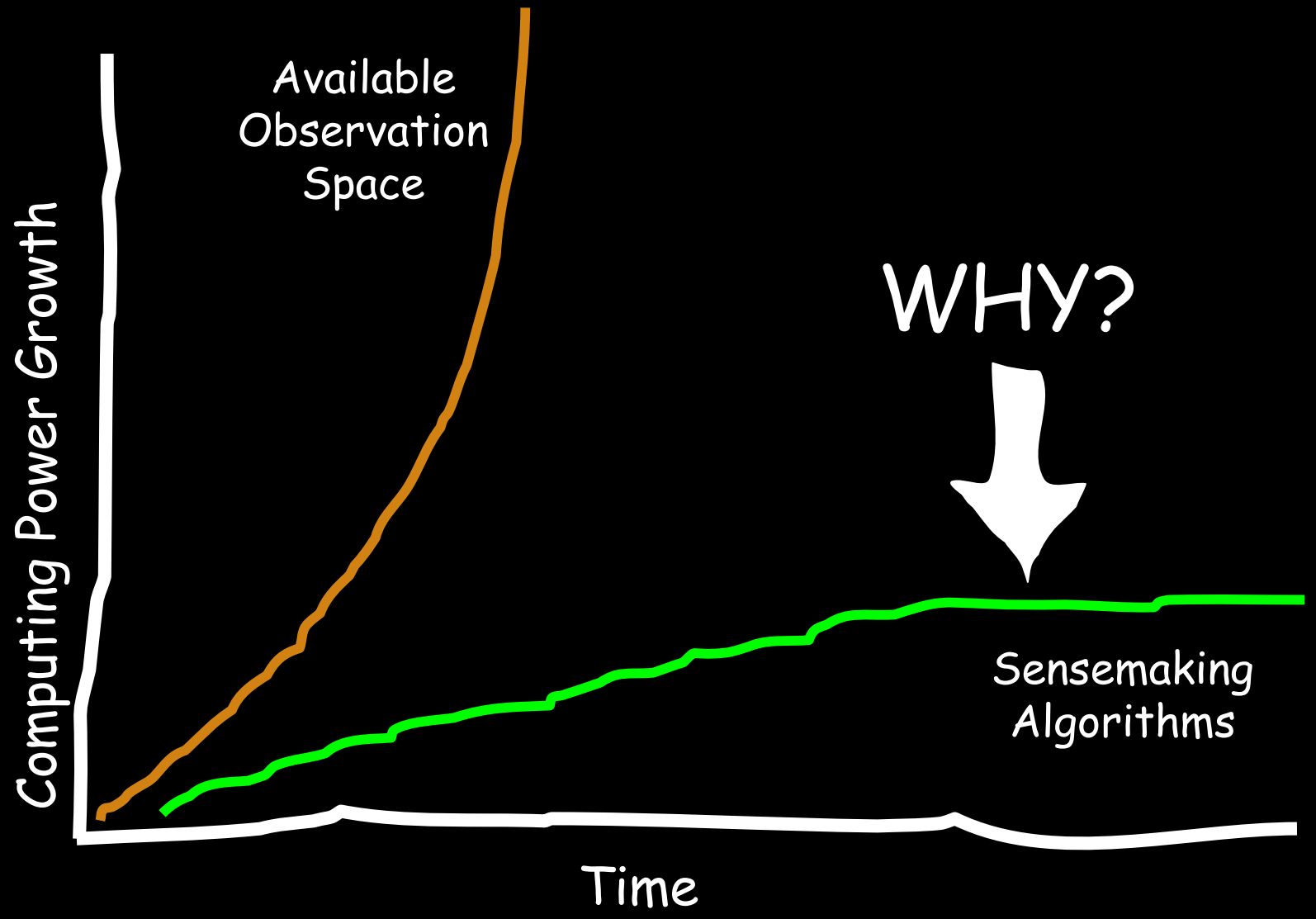
Enterprise Amnesia, definition

A defect in memory, resulting in wasted resources, poor decisions, lower revenues, unnecessary fraud losses, and more.

Trend: Organizations Are Getting Dumber



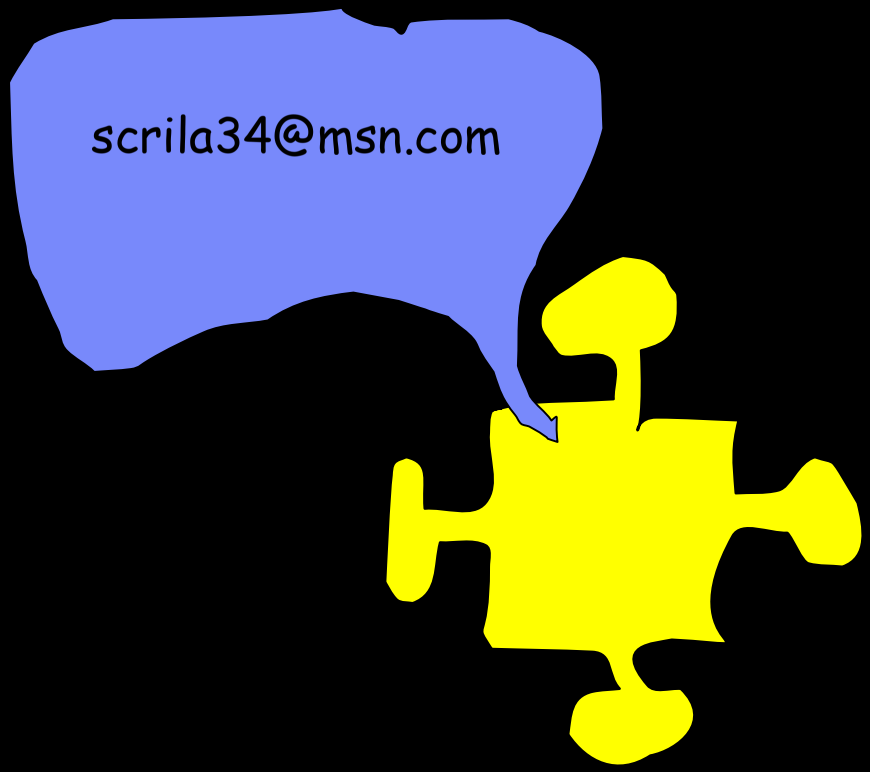
Trend: Organizations Are Getting Dumber

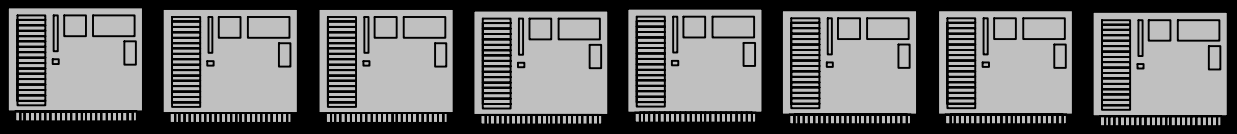


Algorithms at Dead End.

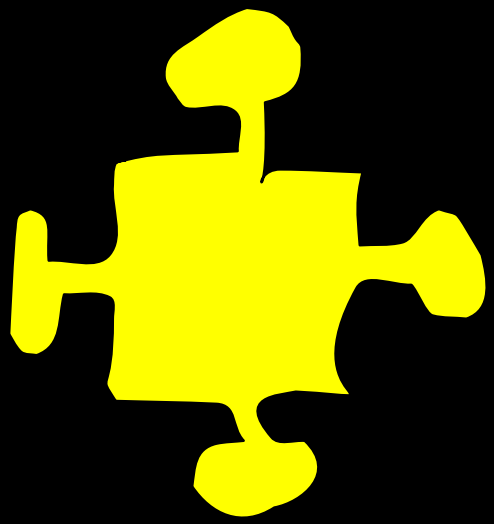
You Can't Squeeze Knowledge Out of a Pixel.

No Context





Information without context



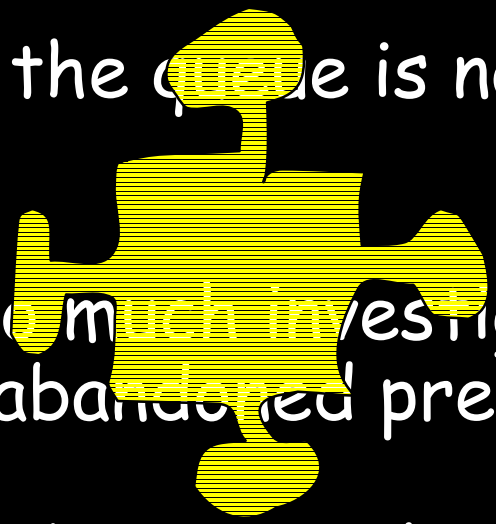
is hardly actionable.

Context, definition

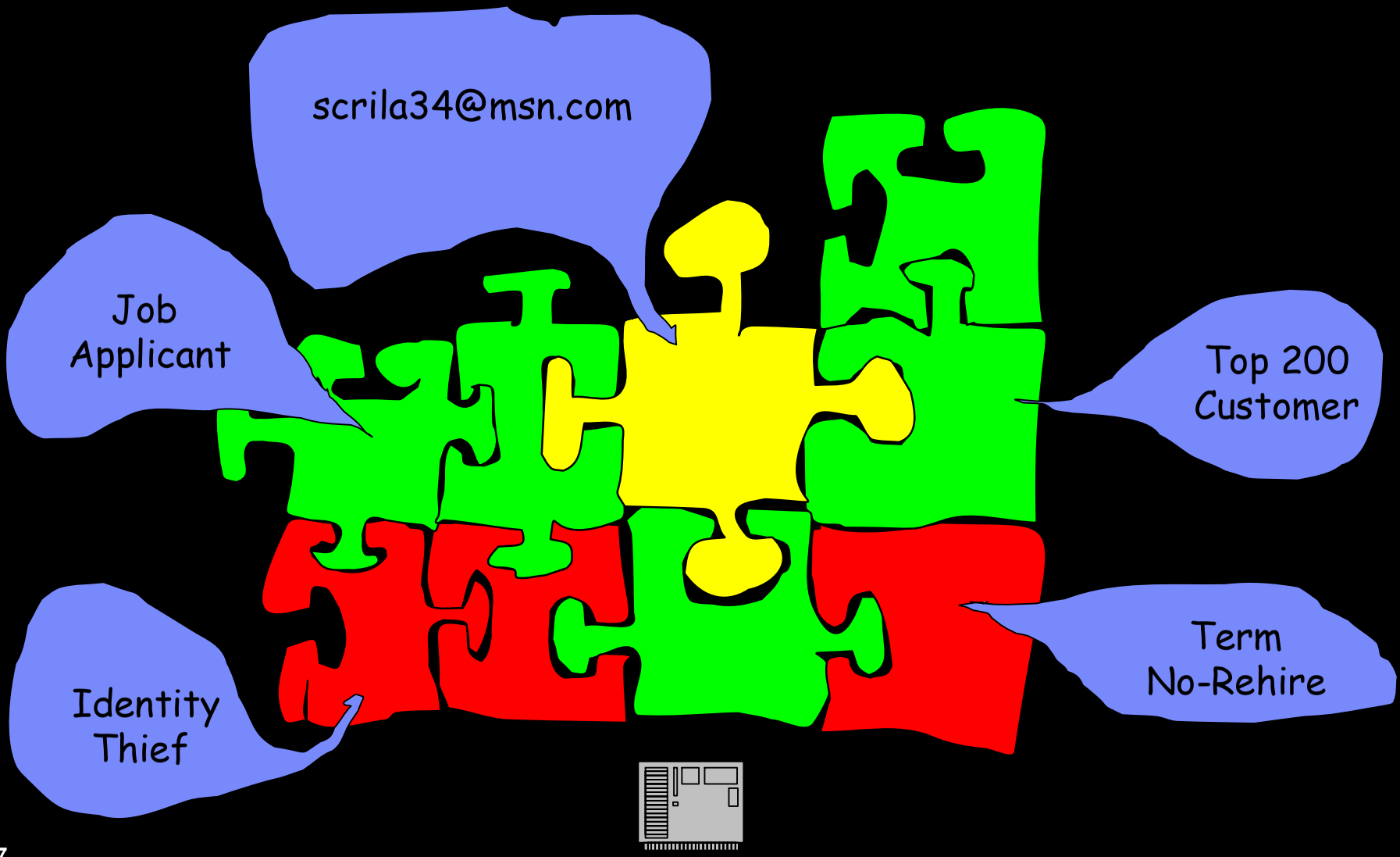
Better understanding something by taking into account the things around it.

Lack of Context – Consequences

- Alert queues filled with false positives and growing faster than humans can address
- The top item in the queue is not the most relevant item
- Items require so much investigative effort – they are often abandoned prematurely
- Risk assessment becomes the risk

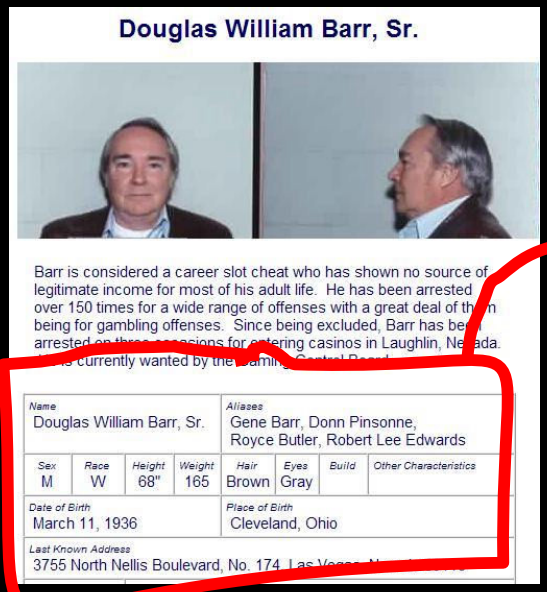


Information in Context … and Accumulating



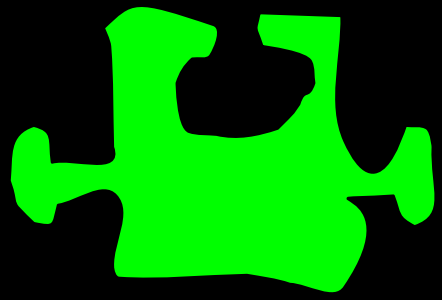
Context Accumulation Requires Feature Extraction

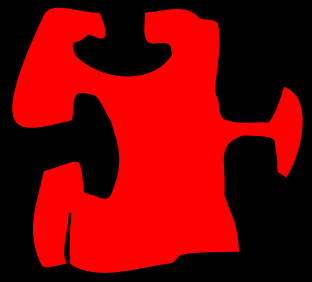
LP#: "Not Cop"



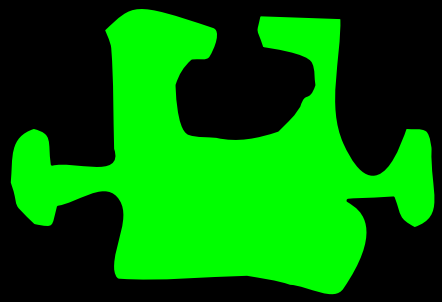
Douglas William Barr, Sr.
Gene barr, Donn Pinsonne
Royce Butler, Robert Lee Edwards
DOB: 11 Mar 1936
POB: Cleveland, Ohio
Add: 3755 N. Nellis Blvd

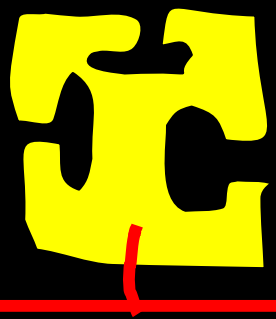
Some Pieces Just Don't Relate … (yet)

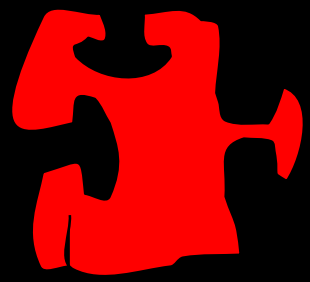




Although … Observations Add Up







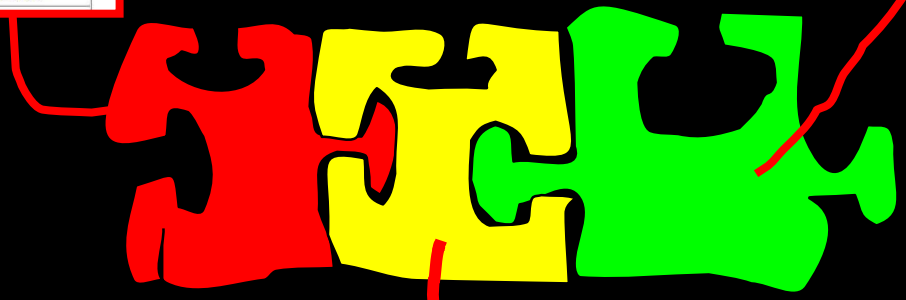
"Not Cop"
Doug Barr, Sr.
DOB: 11 Mar 1936
Add: Las Vegas

Observations Add Up

Douglas William Barr, Sr.

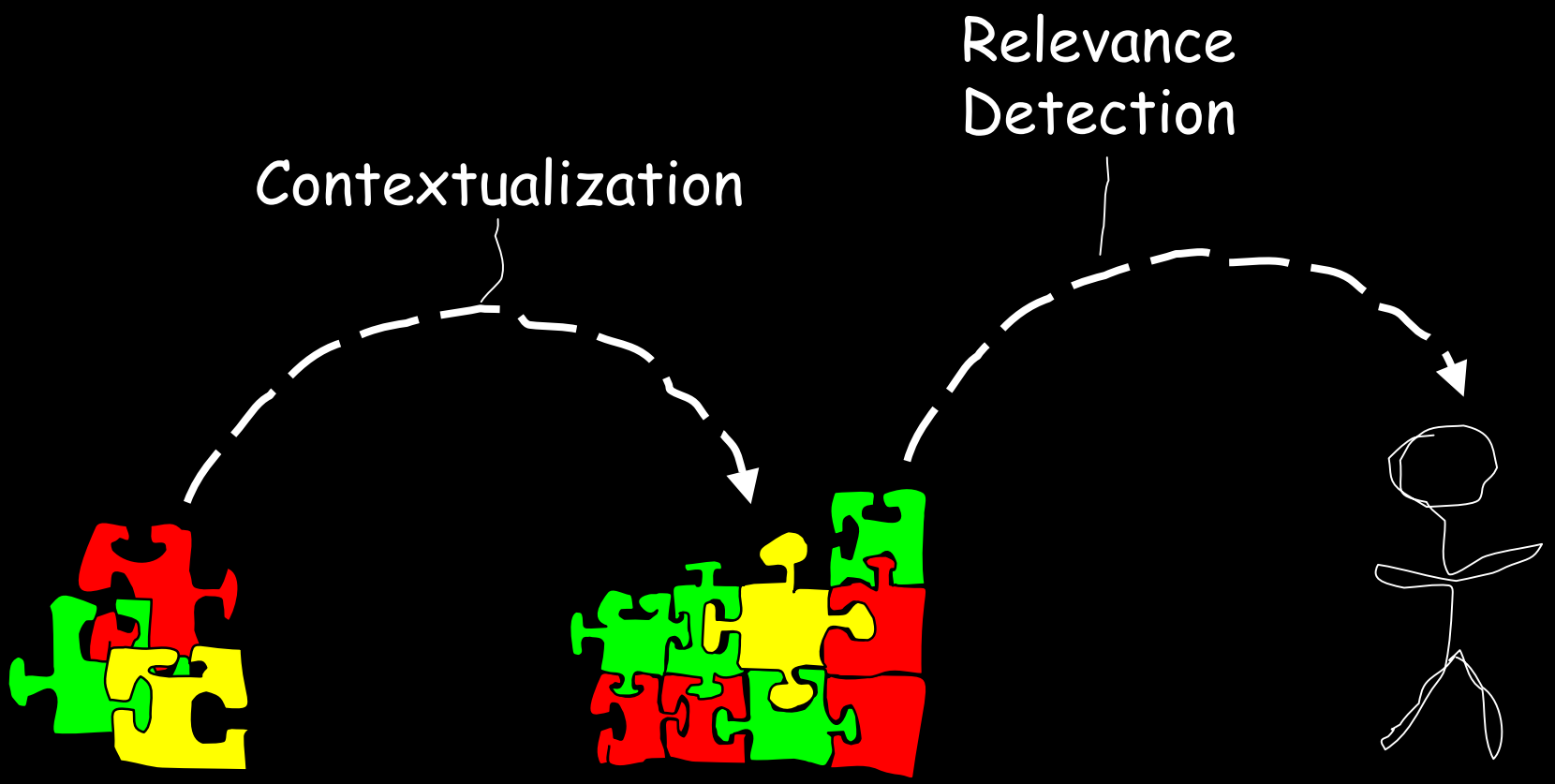
Barr is considered a career slot cheat who has shown no source of legitimate income for most of his adult life. He has been arrested over 150 times for a wide range of offenses with a great deal of them being for gambling offenses. Since being excluded, Barr has been arrested on three occasions for entering casinos in Laughlin, Nevada. He is currently wanted by the Gaming Control Board.

Name				Aliases			
Douglas William Barr, Sr.				Gene Barr, Donn Pinsonne, Royce Butler, Robert Lee Edwards			
Sex	Race	Height	Weight	Hair	Eyes	Build	Other Characteristics
M	W	68"	165	Brown	Gray		
Date of Birth				Place of Birth			
March 11, 1936				Cleveland, Ohio			
Last Known Address							
3755 North Nellis Boulevard, No. 174, Las Vegas, Nevada 89115							

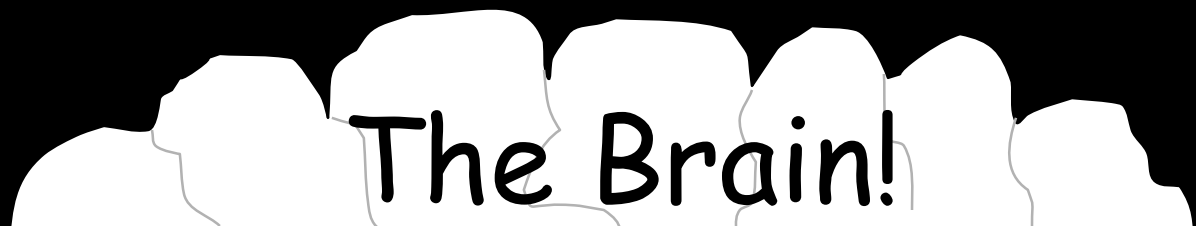


"Not Cop"
Doug Barr, Sr.
DOB: 11 Mar 1936
Add: Las Vegas

From Pixels to Pictures to Insight



Enterprise intelligence
requires
context accumulation
and persistence.



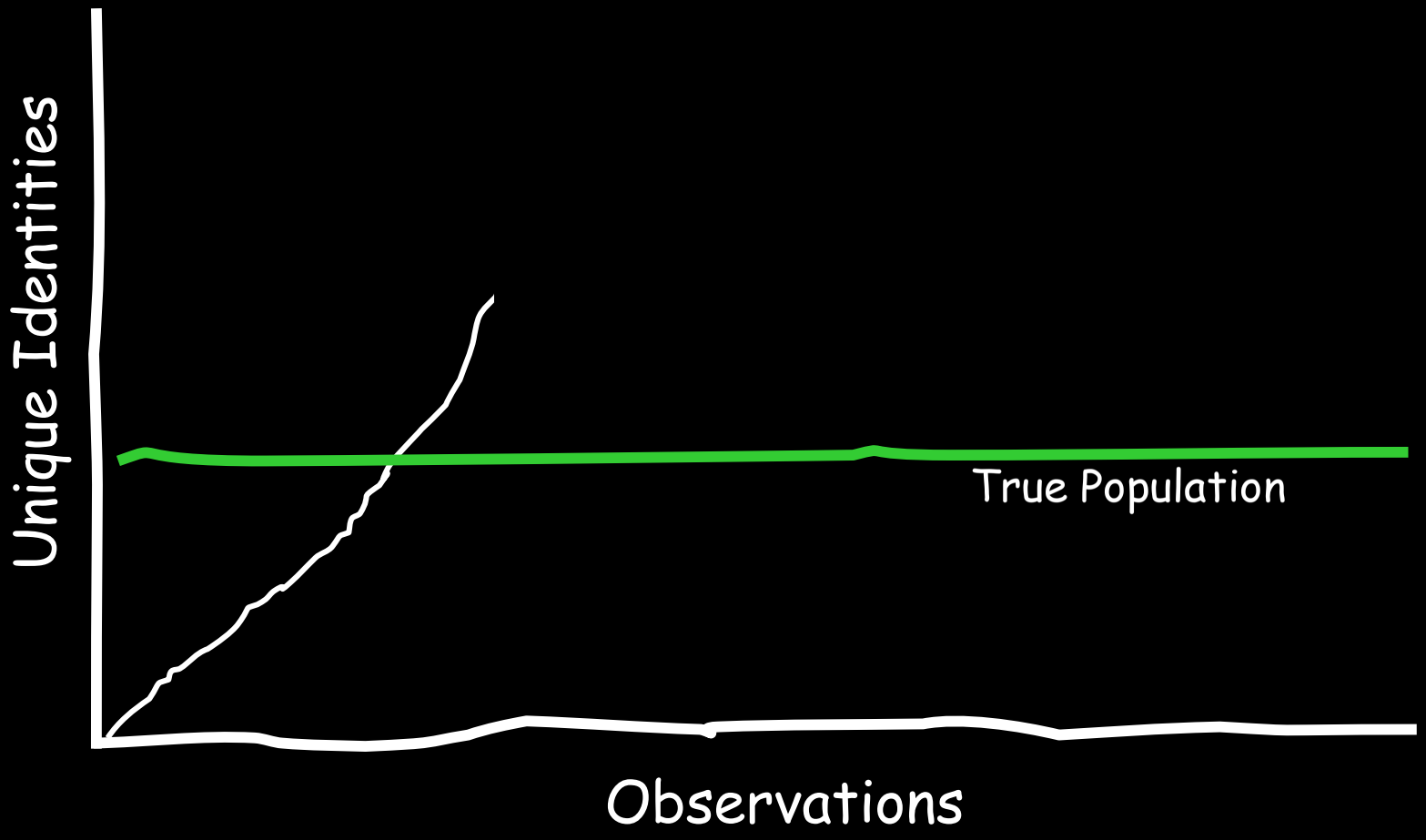
The Puzzle Metaphor

- Imagine an ever-growing pile of puzzle pieces of varying sizes, shapes and colors
- What it represents is unknown (there is no picture on hand)
- Is it one puzzle, 15 puzzles, or 1,500 different puzzles?
- Some pieces are duplicates, missing, incomplete, low quality, or have been misinterpreted
- Some pieces may even be professionally fabricated lies
- Point being: Until you take the pieces to the table and attempt assembly, you don't know what you are dealing with

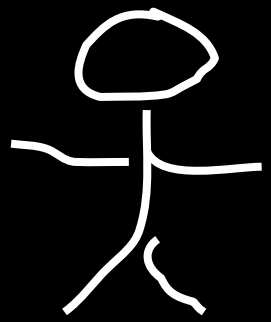
How Context Accumulates

- With each new observation … one of three assertions are made: 1) Un-associated; 2) placed near like neighbors; or 3) connected
- Must favor the false negative
- New observations sometimes reverse earlier assertions
- Some observations produce novel discovery
- The emerging picture helps focus collection interests
- As the working space expands, computational effort increases
- Given sufficient observations, there can come a tipping point
- Thereafter, confidence begins to improve while computational effort decreases!

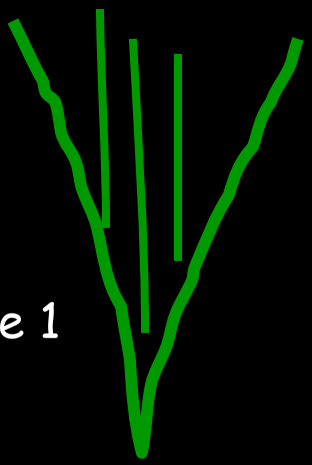
Overstated Population



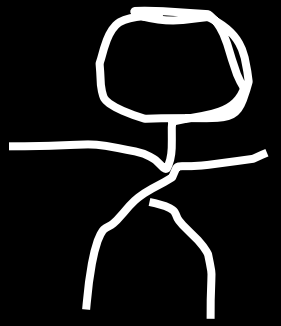
Counting Is Difficult



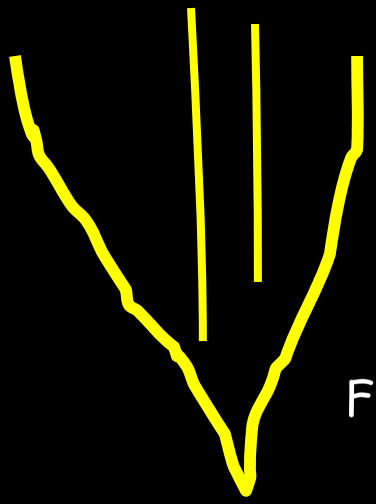
Mark Smith
6/12/1978
443-43-0000



File 1

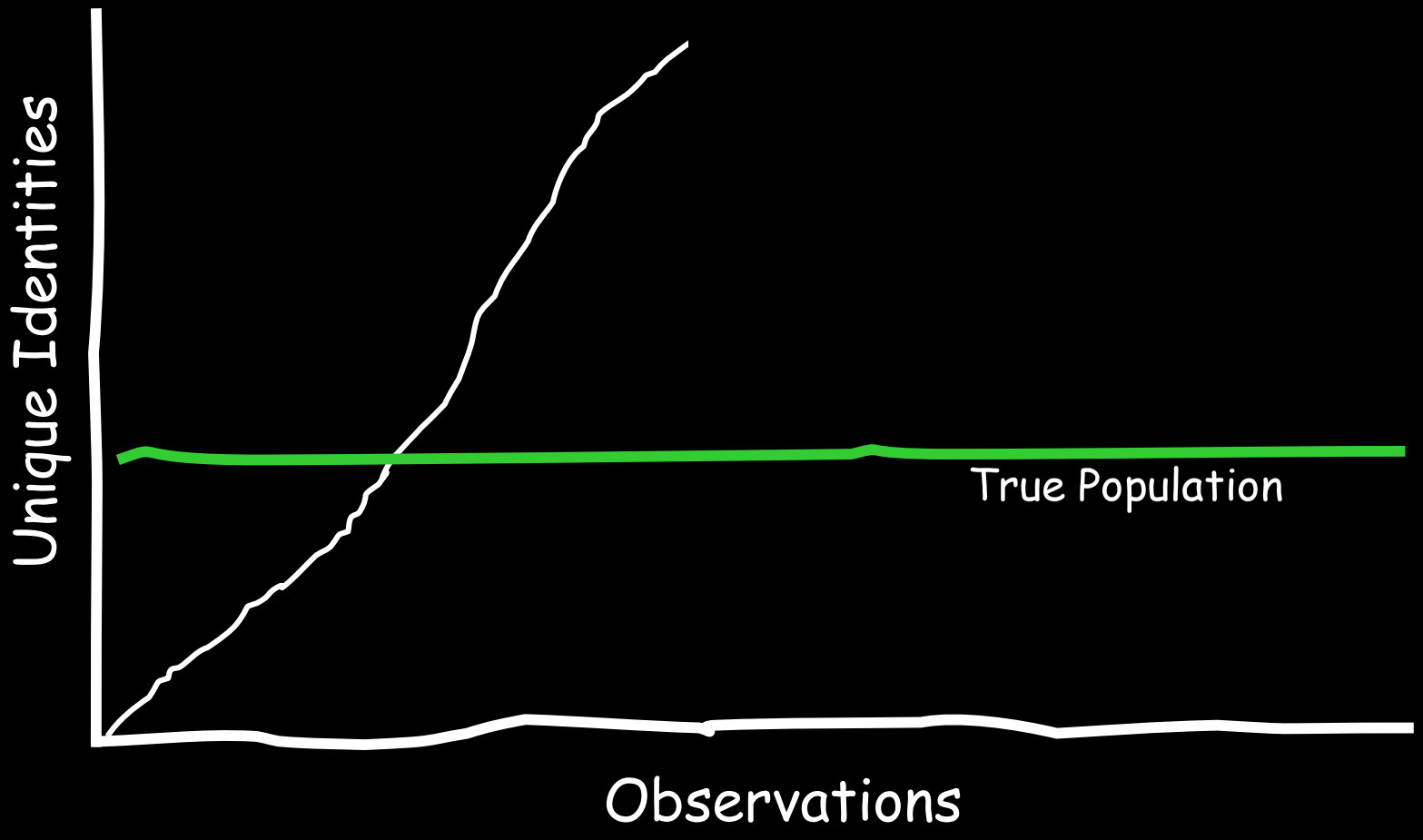


Mark R Smith
(707) 433-0000
DL: 00001234

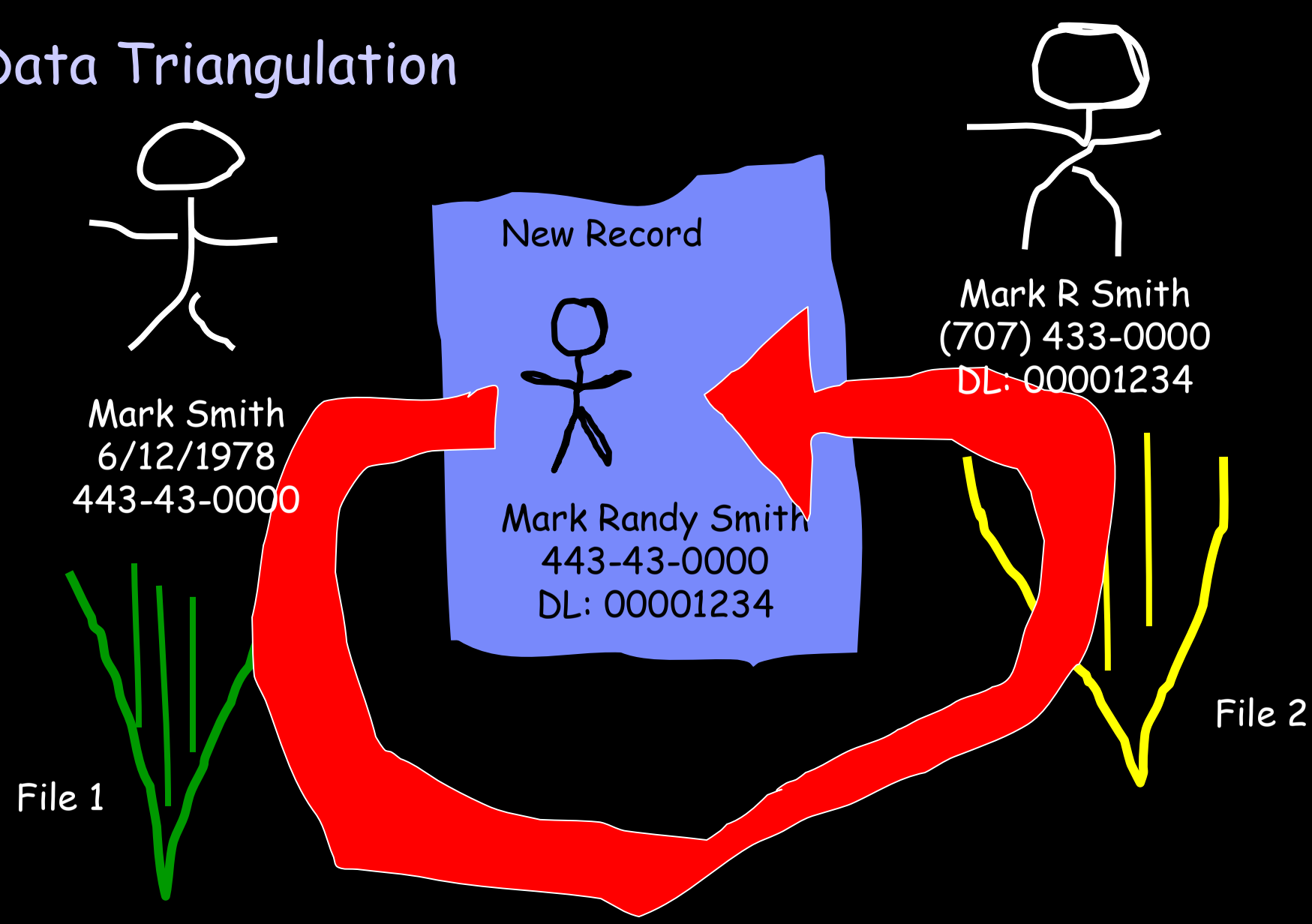


File 2

The Rise and Fall of a Population



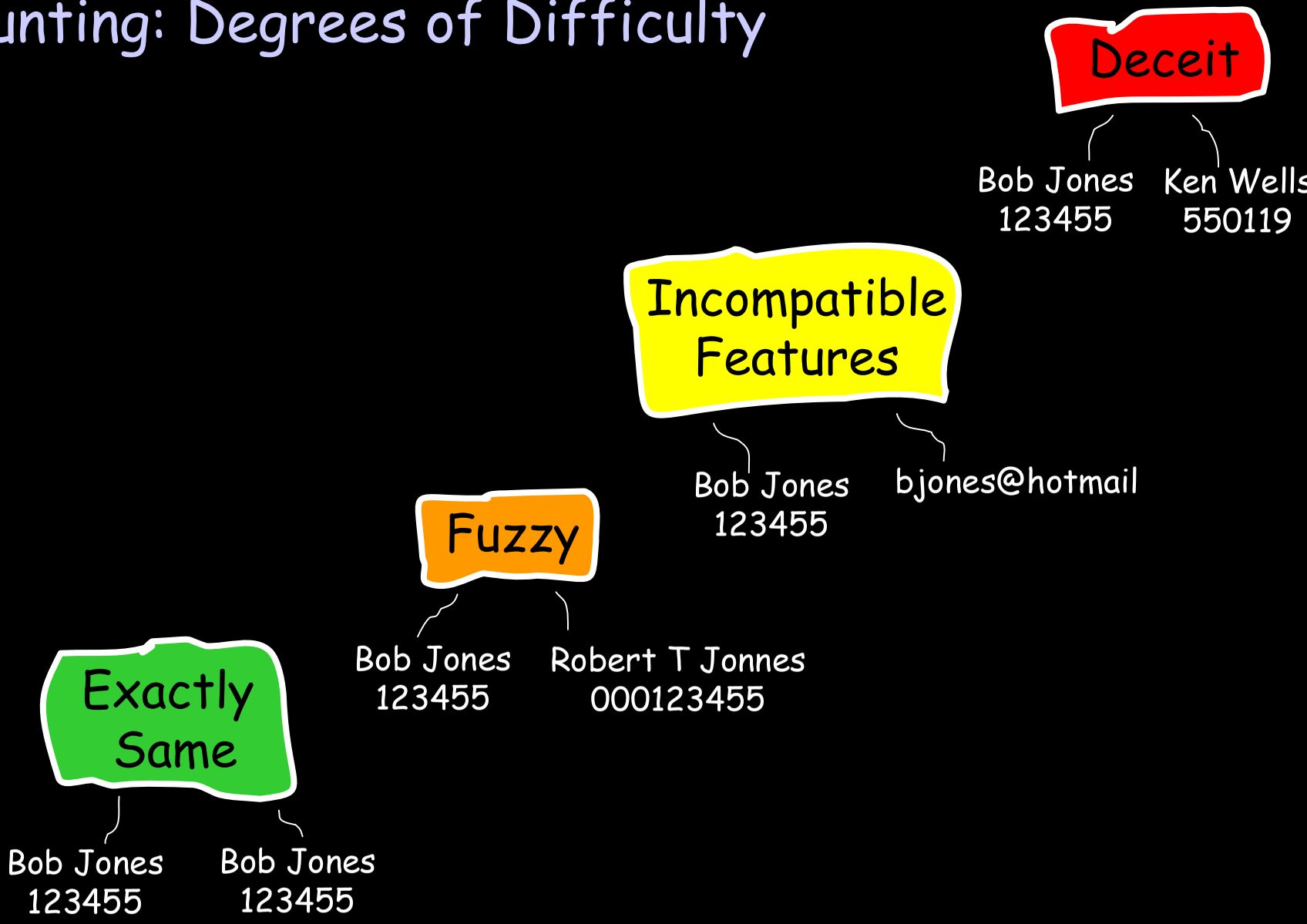
Data Triangulation



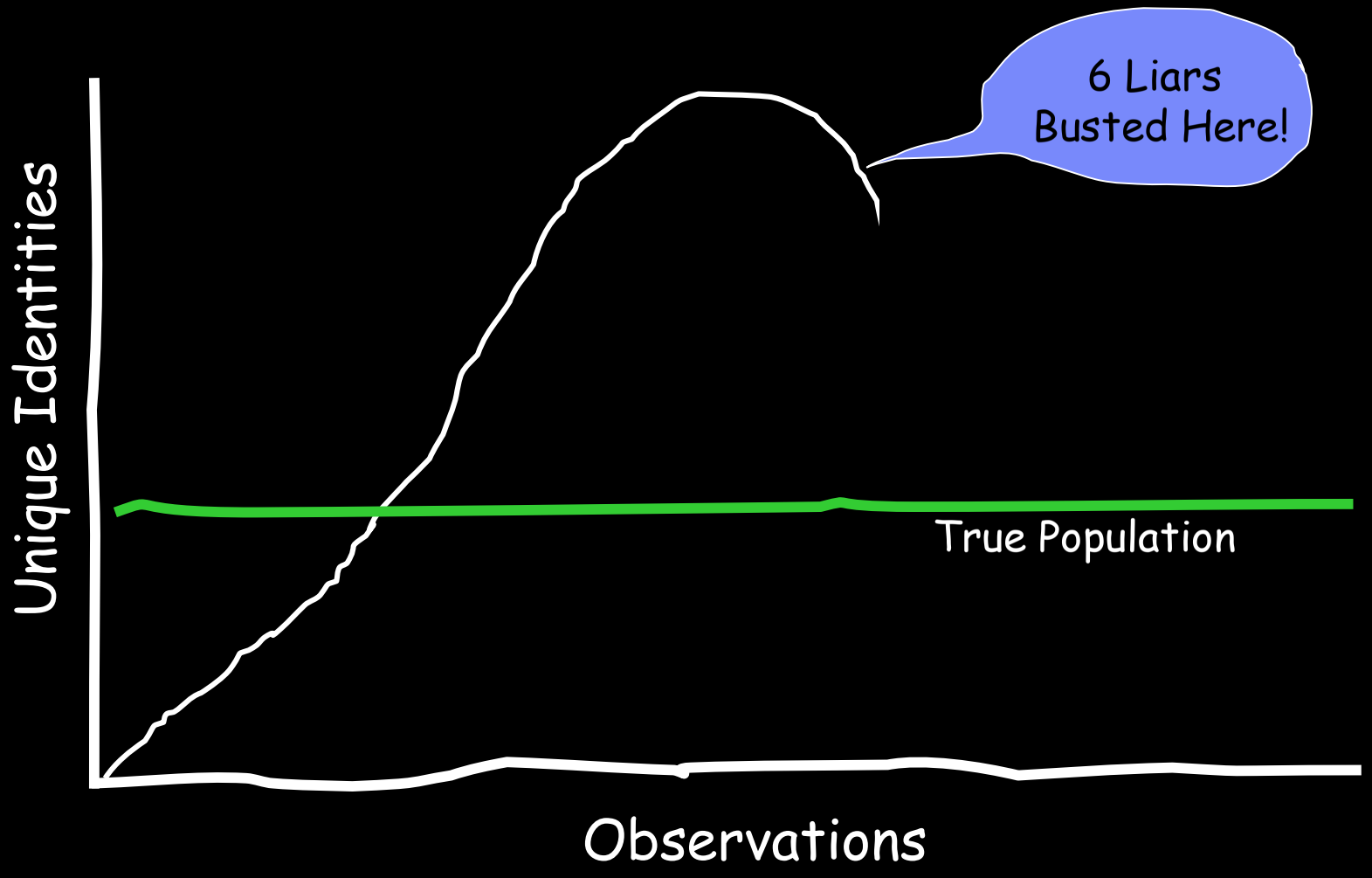
Counting is Essential to Prediction

- Is it 5 people each with 1 account … or is it 1 person with 5 accounts?
- Is it 20 cases of H1N1 in 20 cities … or one case reported 20 times?
- If one cannot count … one cannot estimate vector or velocity (direction and speed).
- Without vector and velocity … prediction is nearly impossible.

Counting: Degrees of Difficulty



And Deceit Revealed



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

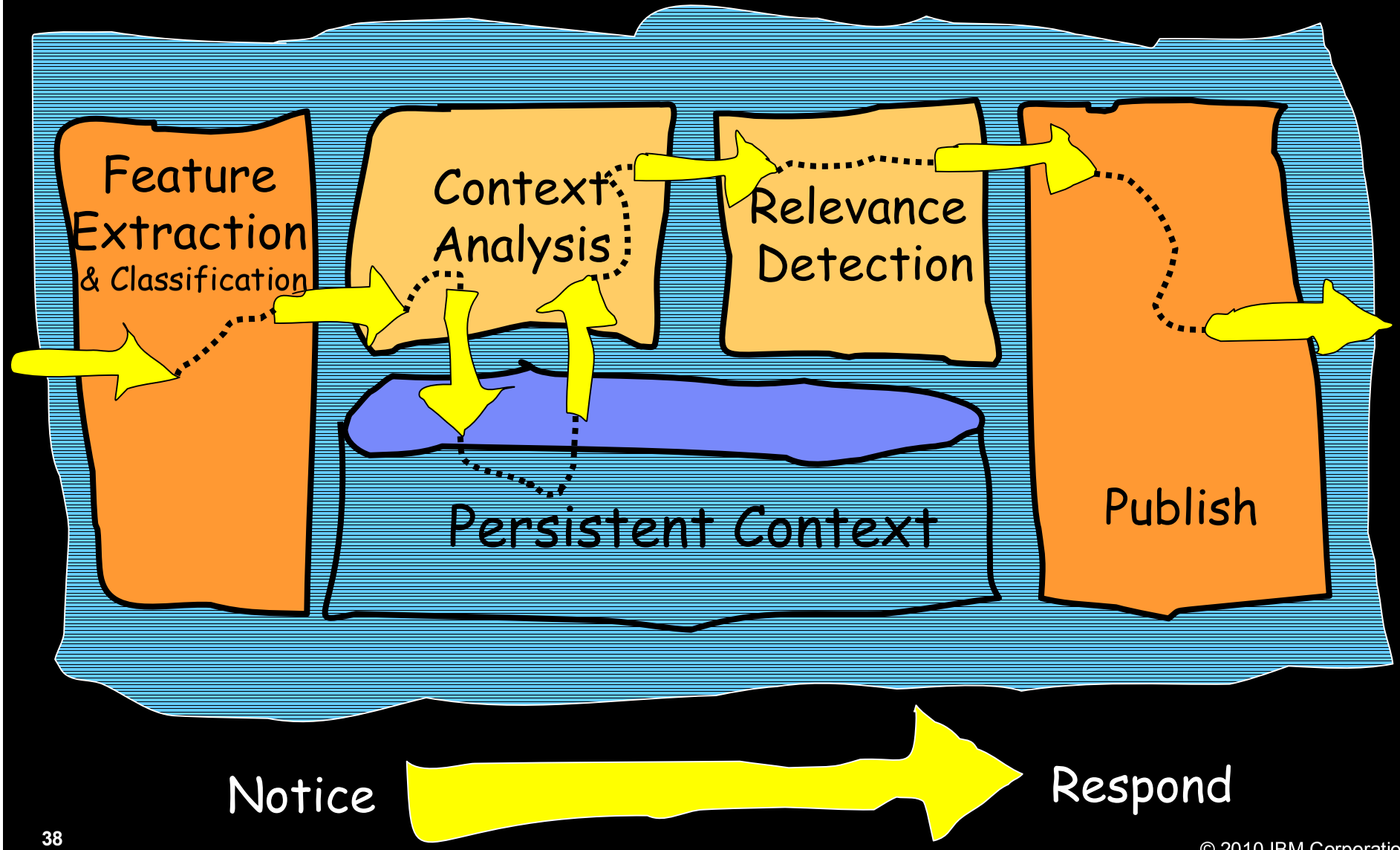
DECEASED PERSON
George Balston
YOB: 1951 SSN: 5598
DOD: 1995

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.

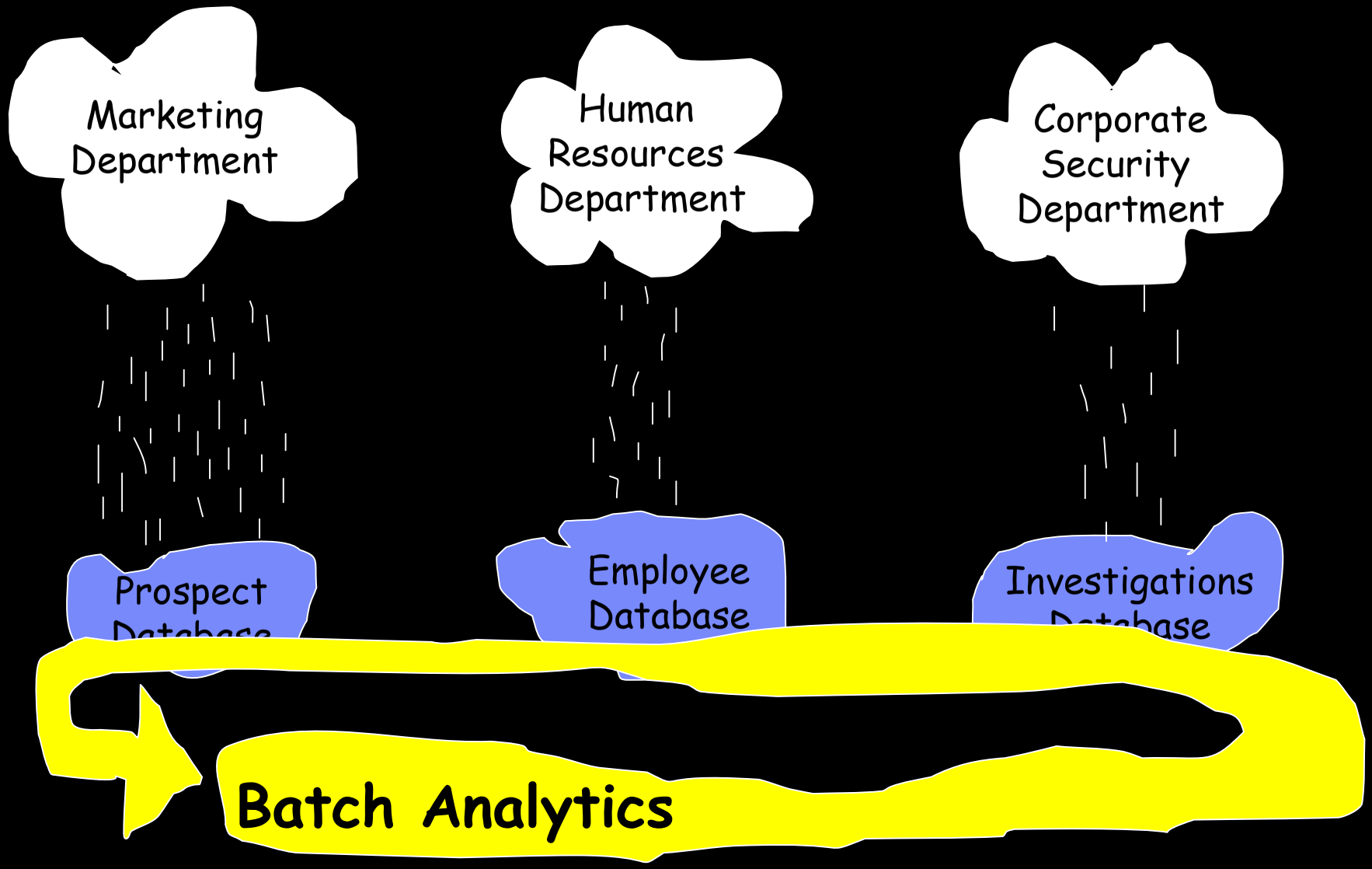
Major Moving Parts



1st principle

If you do not process every new piece of key data (perception) first like a query … then you will not know if it matters … until someone asks.

"The Data is a Query" **Beats** "Boil the Ocean"



2nd principle

Treat queries like data to avoid having to ask every question every day.

New Think: Data and Query Equality

3rd principle

Enterprise awareness is computationally most efficient when performed at the moment the observation is perceived.

Big Data – New Physics

Context Accumulation + Big Data = New Physics

- More Data: Better prediction
  - Less false positives
  - Less false negatives
- More Data: Bad data good
- More Data: Less compute effort
- More Data: Better sense of where to place one's attention
  - Tolerance for ambiguity
  - Select what is worth being curious about

"G2"

My Skunk Works Effort

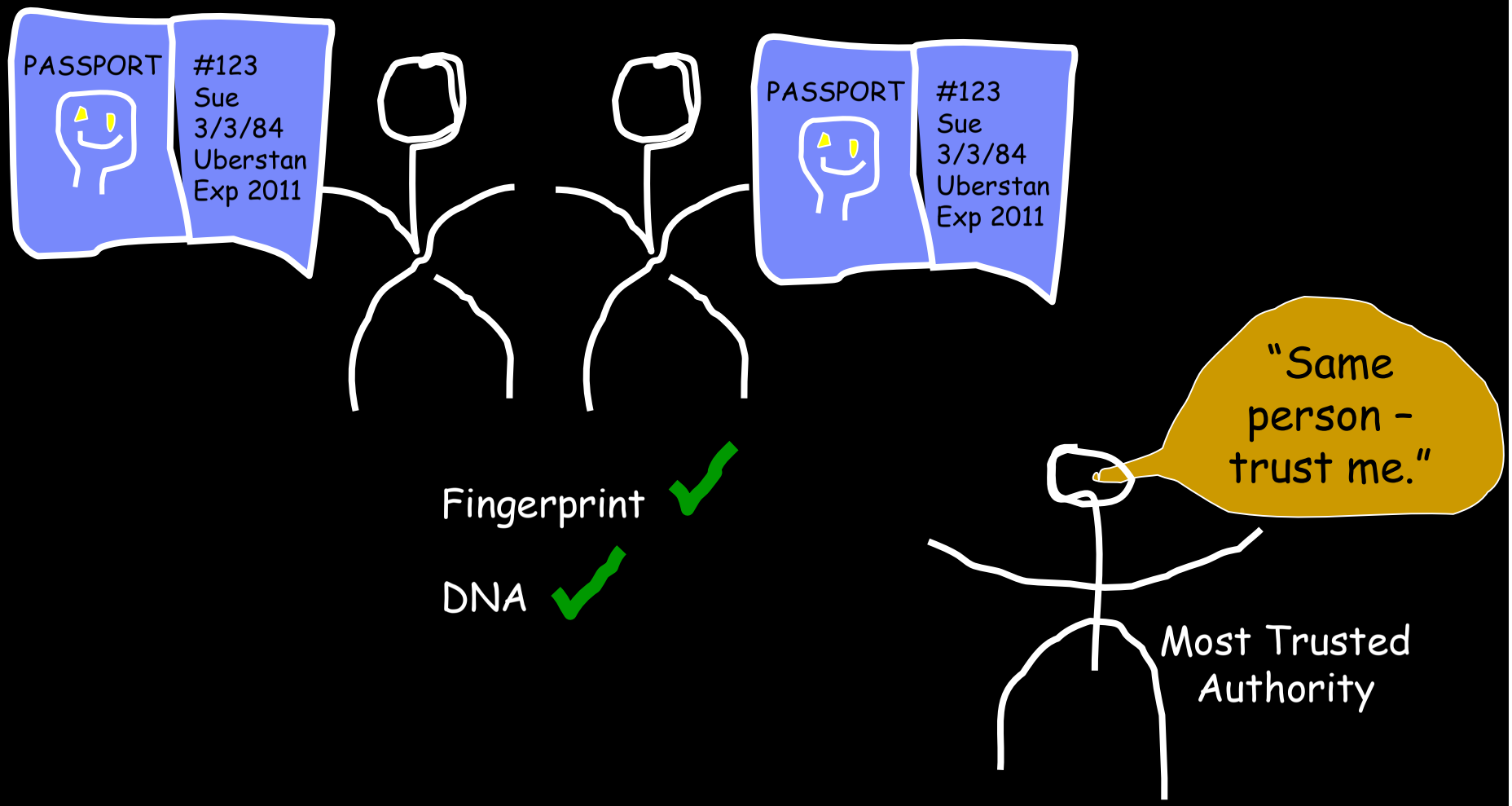
My G2 Effort

- Exploiting the big data new physics
- Generalized entity domain (actors, events, things, locations, and user defined) – adding new data sources, new entity classes, and new features on the fly
- Structured, unstructured, biographic, biometric, geospatial, reference data, anonymized features and user queries … simultaneous contextualized and persisted in the same data space
- Designed for massively distributed, share nothing, cloud compute environments
- Privacy by design (ranging from immutable audit logs to analytics over anonymized data)
- Deep space/time awareness

"Key Features" Enable Expert Counting

People	Cars	Router
Name	Make	Device ID
Address	Model	Make
Date of Birth	Year	Model
Phone	License Plate No.	Firmware Vers.
Passport	VIN	Asset ID
Nationality	Owner	Etc.
Biometric	Etc.	
Etc.		

Consider Lying Identical Twins



- The same thing cannot be in two places … at the same time.
- Two different things cannot occupy the same space … at the same time.

Space & Time Enables Absolute Disambiguation

People	Cars	Router
Name	Make	Device ID
Address	Model	Make
Date of Birth	Year	Model
Phone	License Plate No.	Firmware Vers.
Passport	VIN	Asset ID
Nationality	Owner	Etc.
Biometric	Etc.	
Etc.		

Life Arcs Are Also Telling

Bill Smith
4/13/67
Salem, OR

Bill Smith
4/13/67
Seattle, WA

Address History	
Tampa, FL	2008-2008
Biloxi, MS	2005-2008
NY, NY	1996-2005
Tampa, FL	1984-1996

Address History	
San Diego, CA	2005-2009
San Fran, CA	2005-2005
Phoenix, AZ	1990-2005
San Jose, CA	1982-1990

Space-Time-Travel

- Cell phones are generating a staggering amount of geo-locational data (600B transactions a day being created in the US)
- This data is being "de-identified" and shared with third parties – in volume and in real-time
- Your movement quickly reveals where you spend your time (e.g., evenings, working hours) and who you spend time with (e.g., friends)
- Re-identification (figuring who is who) is somewhat trivial

Consequences

- Space-time-travel makes for absolute identification and disambiguation
- It is the ultimate biometric
- It will help reshape other very tough problems like image classification and identification
- It will unravel ones secrets
- It will prove to be enable enormous opportunity and will challenge existing notions of privacy
- It's here now and more to come

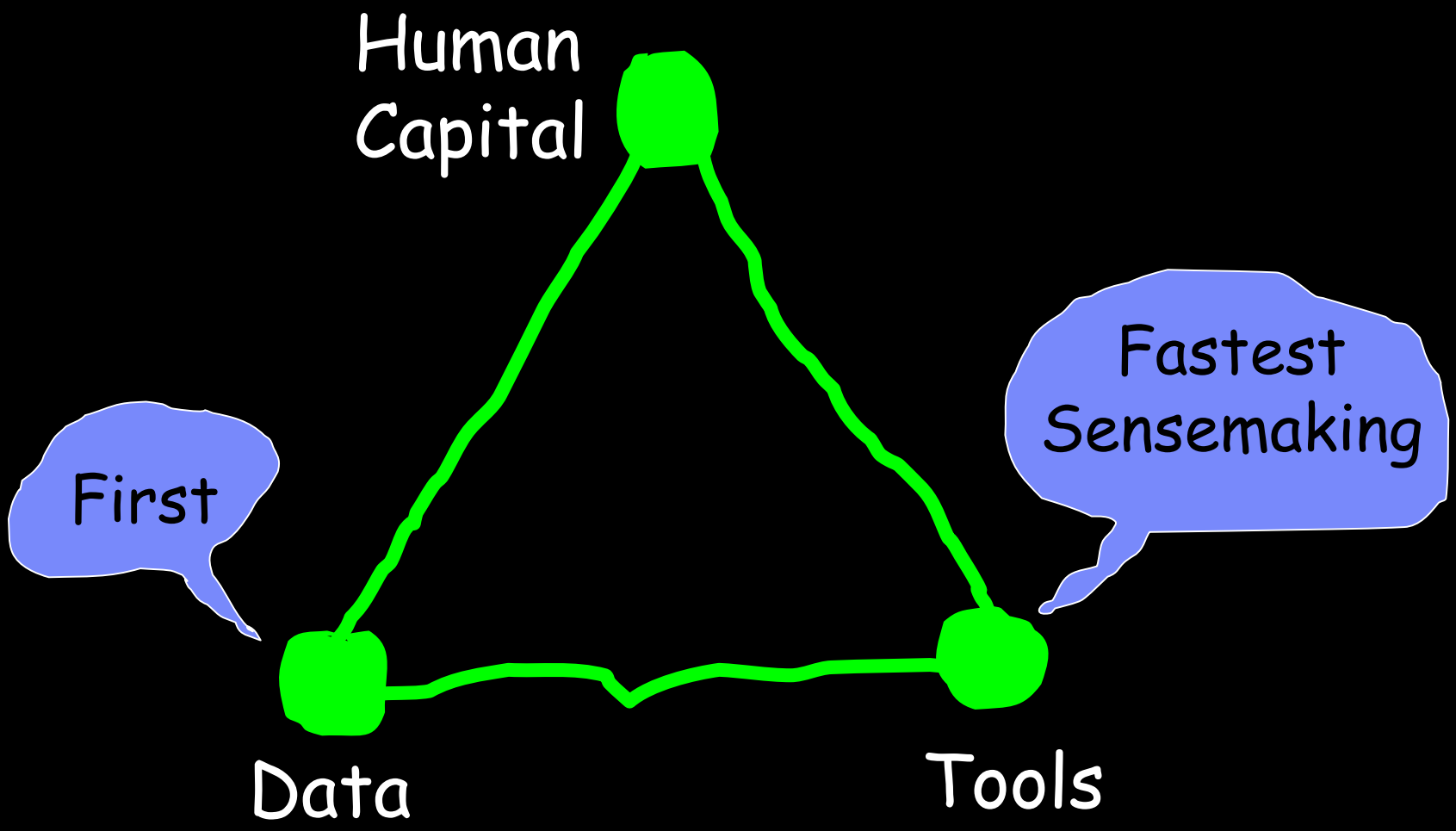
Like Magic …

- Prediction with 87% certainty where you will be next Thursday at 5:35pm
- Names of the top 10 people you co-locate with, not at home and not at work
- High quality traffic-avoiding predictions, pushed to you just in time
- Transactional activity not consistent with your pattern of life used to reduce credit card theft by 90%
- Uberstan preempts the next mass protest in real-time
- A political opponent is crushed and resigns two days after announcing their candidacy

Closing Thoughts

It's all about competition.

To Beat the Competition …



"Every millisecond gained in our program trading applications is worth $100 million a year."

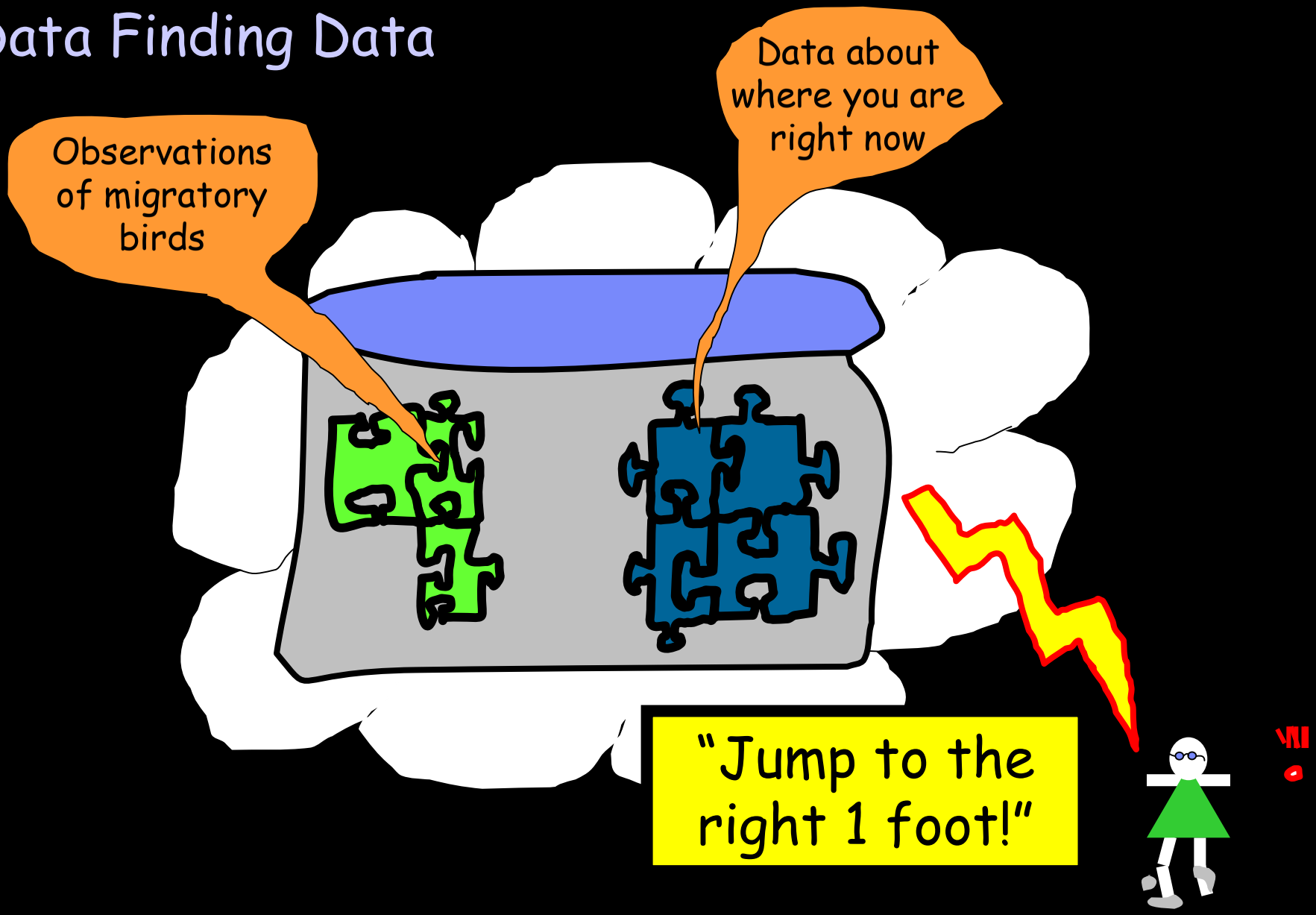
Goldman Sachs, 2007 * Source Automated Trader Magazine 2007

Key Points

- Information in context is necessary to deliver next generations smart, sensemaking systems
- Without context … data visualization, data mining, and pattern discovery systems ain't all that
- Enterprise intelligence is not a compartmental activity … it comes from enterprise-wide context accumulation
- Because mankind cannot dream up and ask every smart question every day …

"The data will find the data … and the relevance will find you."

Data Finding Data



When this technology serves …

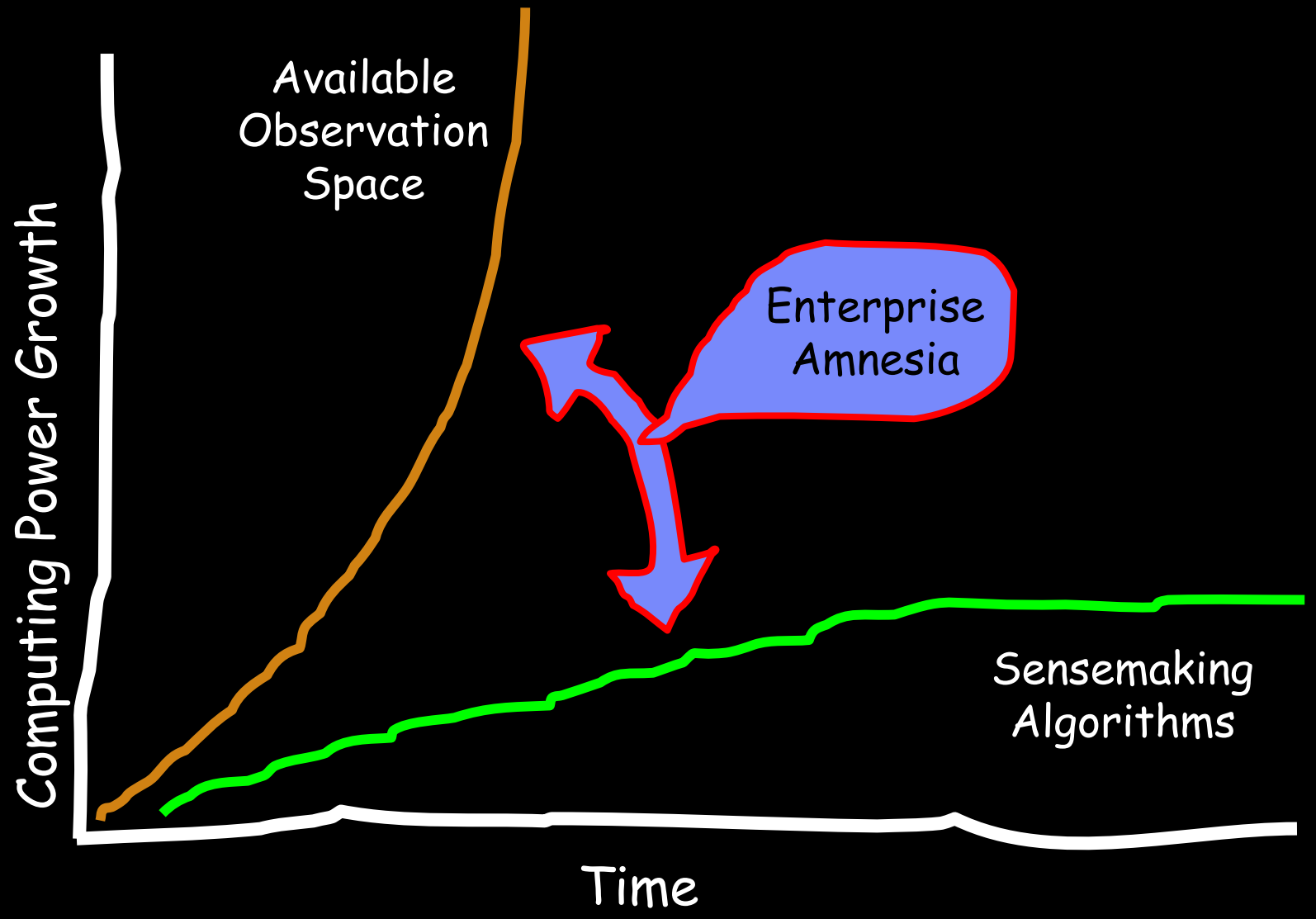
… you and your doctor …

LOVE!

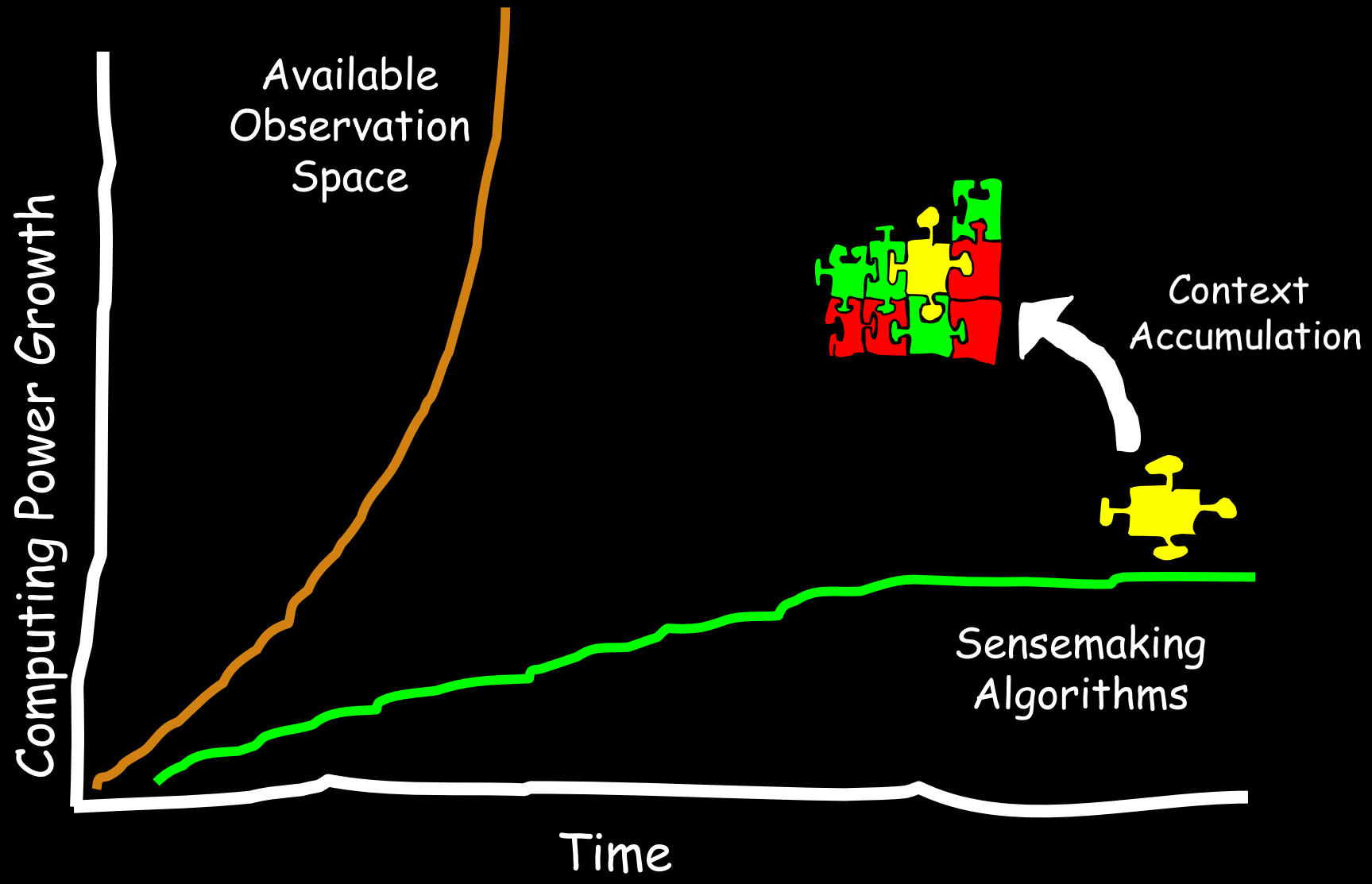
HATE!

… the police looking at you …

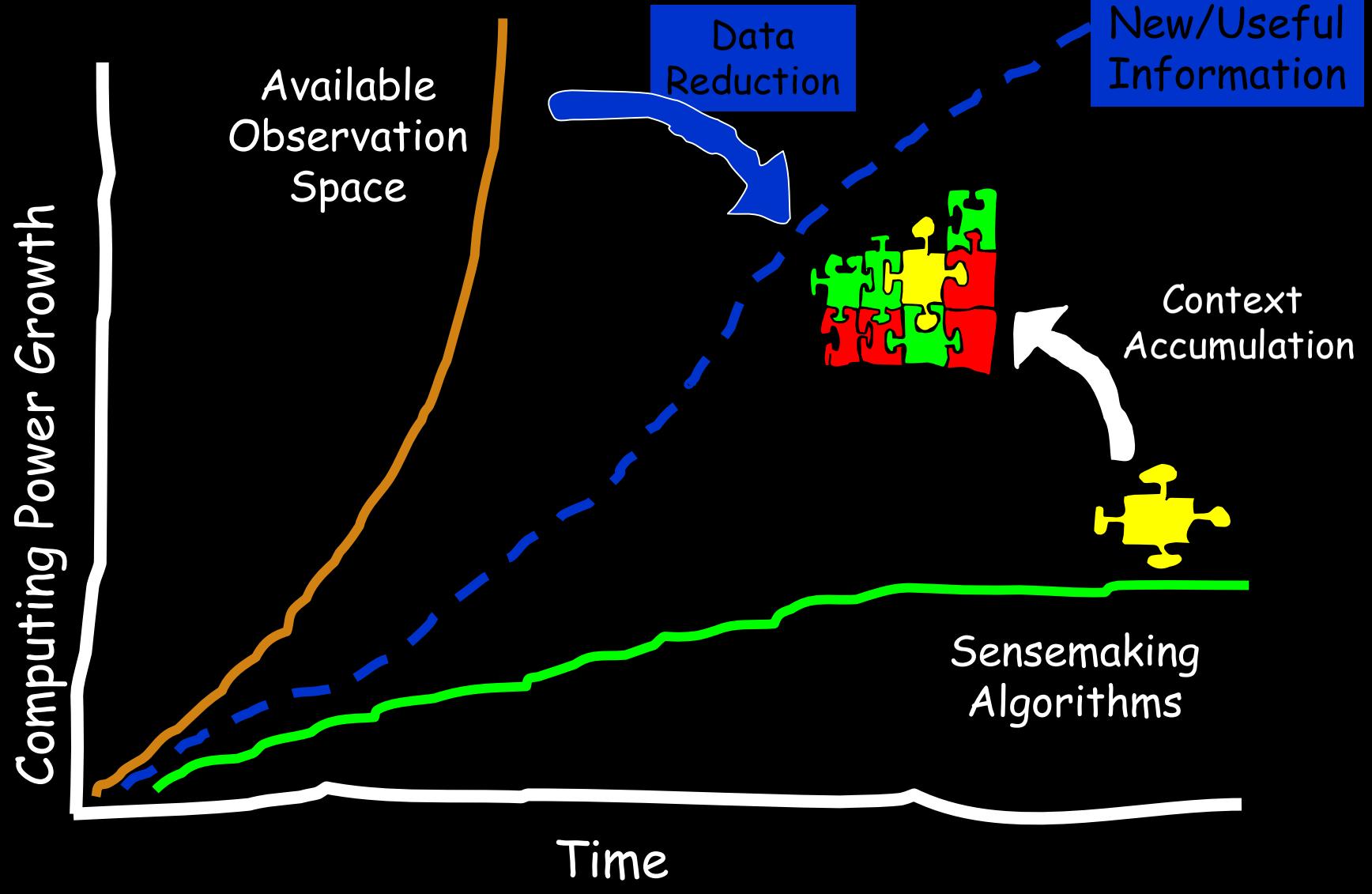
Wish This On The Enemy



Enterprise Intelligence: The Way Forward



Better Prediction to Discard



Related Blog Posts

Algorithms At Dead-End: Cannot Squeeze Knowledge Out Of A Pixel

Smart Sensemaking Systems, First and Foremost, Must be Expert Counting Systems

Data Finds Data

Puzzling: How Observations Are Accumulated Into Context

Your Movements Speak for Themselves: Space-Time Travel Data is Analytic Super-Food!

"Macro Trends: The Privacy and Civil Liberties Consequences … and Comments on Responsible Innovation" – My DHS DPIAC Testimony, September 2008

IBM

Blogging At:

www.JeffJonas.TypePad.com

Information Management
Privacy
National Security

and Triathlons

DEFRAG 2010

Enterprise Amnesia
vs. Enterprise Intelligence

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

November 18, 2010