YarcData
Getting to **Eureka!** faster™

Data Analytics:

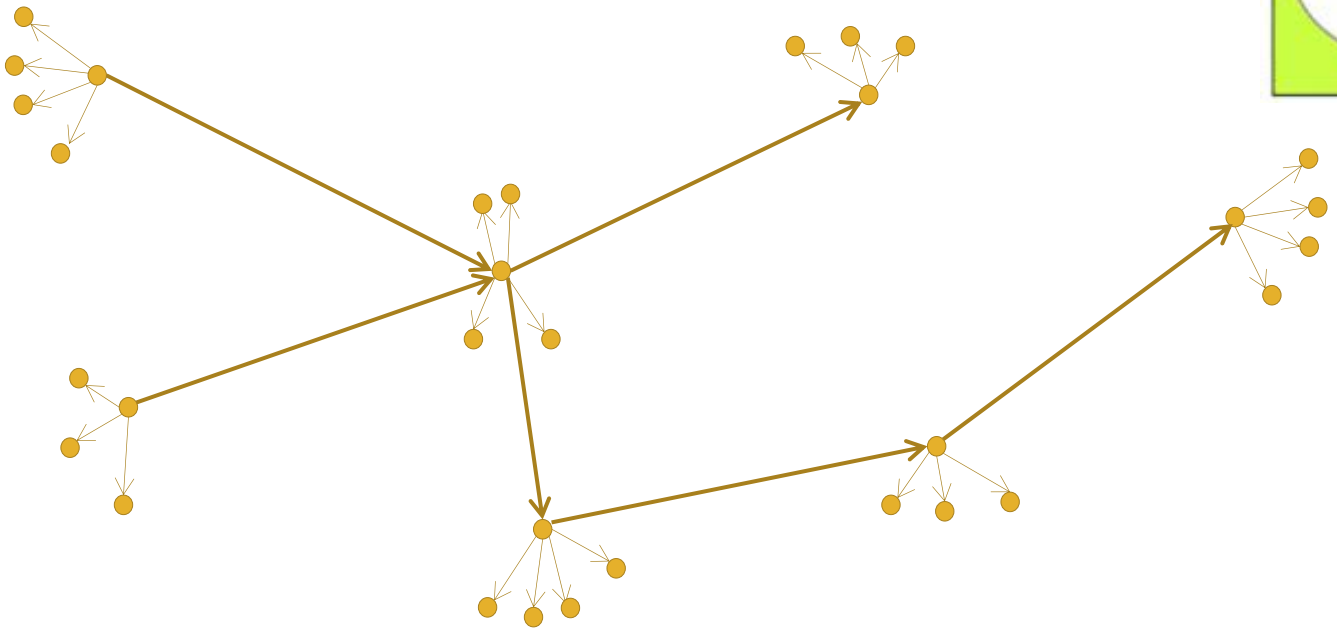
Supercomputers & Graph Analysis

Tim White

tim@yarcdata.com

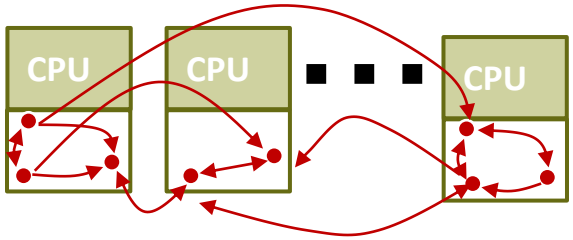
What is a Graph?

- Not a chart
- A fundamental data structure
- A collection of vertices (nodes) and edges (links, relationships, connections)

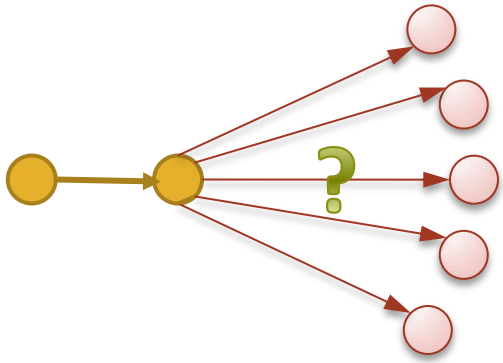


Graphs and Traditional Technologies

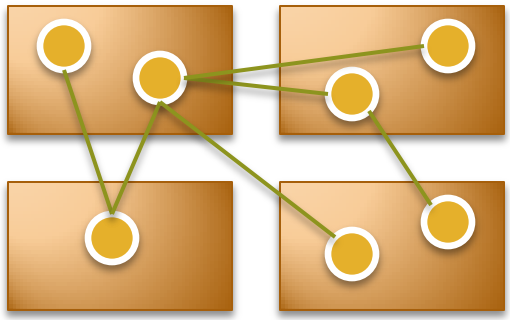
- **Square peg, round hole:**
  - Current technology does not support efficient representation, storage, and interaction with complex graph structures
  - Traditional relational models only add the an already complex structure
  - Traditional hardware approaches do not support efficient access to highly interconnected graphs



- **You don't know what you don't know:**
  - Efficient relational schemas require prior knowledge of the relationships between database fields
  - Updating and modifying schemas frequently introduces delays and errors



- **Problems in partitioning the problem:**
  - Distributed computing solutions are good…If your problem can be easily partitioned
  - Graphs are not predictable; accessing graph nodes across large clusters can be unwieldy at best and does not work at scale

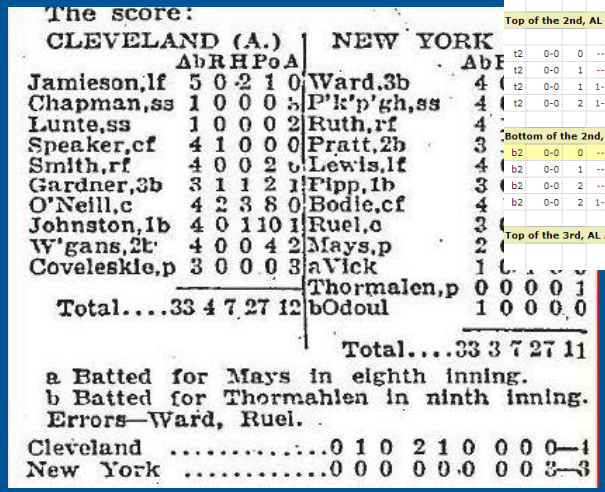


Approach to Baseball Analytics

-A Use Case

The Data World Has Changed

Box Score

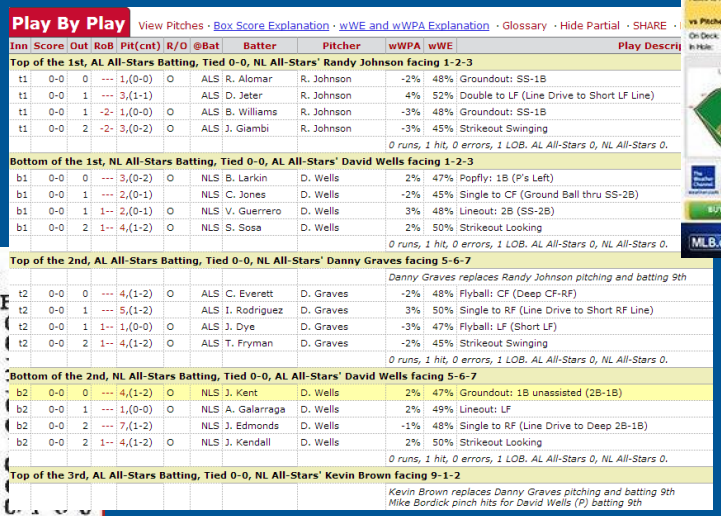


The score:

CLEVELAND (A.)	AB	R	H	PO	A	NEW YORK	AB	R	H	PO	A
Jamieson,lf	5	0	2	1	0	Ward,3b	4				
Chapman,ss	1	0	0	0	5	P'k'p'gh,ss	4				
Lunte,ss	1	0	0	0	2	Ruth,rf	4				
Speaker,cf	4	1	0	0	0	Pratt,2b	3				
Smith,rf	4	0	0	2	0	Lewis,lf	4				
Gardner,3b	3	1	1	2	1	Pipp,1b	3				
O'Neill,c	4	2	3	8	0	Bodie,cf	4				
Johnston,1b	4	0	1	10	1	Ruel,c	3				
W'gans,2b	4	0	0	4	2	Mays,p	2				
Coveleskie,p	3	0	0	0	3	aVick	1				
						Thormalen,p	0	0	0	0	1
Total	33	4	7	27	12	bOdoul	1	0	0	0	0
						Total	33	3	7	27	11

a Batted for Mays in eighth inning.
b Batted for Thormahlen in ninth inning.
Errors—Ward, Ruel.

Cleveland .....0 1 0 2 1 0 0 0 0—4
New York .....0 0 0 0 0 0 0 0 3—3



Play By Play View Pitches · Box Score Explanation · wWE and wWPA Explanation · Glossary · Hide Partial · SHARE

Inn	Score	Out	RoB	Pit(cnt)	R/O	@Bat	Batter	Pitcher	wWPA	wWE	Play Descri
Top of the 1st, AL All-Stars Batting, Tied 0-0, NL All-Stars' Randy Johnson facing 1-2-3											
t1	0-0	0	---	1,(0-0)	O	ALS	R. Alomar	R. Johnson	-2%	48%	Groundout: SS-1B
t1	0-0	1	---	3,(1-1)		ALS	D. Jeter	R. Johnson	4%	52%	Double to LF (Line Drive to Short LF Line)
t1	0-0	1	-2-	1,(0-0)	O	ALS	B. Williams	R. Johnson	-3%	48%	Groundout: SS-1B
t1	0-0	2	-2-	3,(0-2)	O	ALS	J. Giambi	R. Johnson	-3%	45%	Strikeout Swinging
											0 runs, 1 hit, 0 errors, 1 LOB. AL All-Stars 0, NL All-Stars 0.
Bottom of the 1st, NL All-Stars Batting, Tied 0-0, AL All-Stars' David Wells facing 1-2-3											
b1	0-0	0	---	3,(0-2)	O	NLS	B. Larkin	D. Wells	2%	47%	Popfly: 1B (P's Left)
b1	0-0	1	---	2,(0-1)		NLS	C. Jones	D. Wells	-2%	45%	Single to CF (Ground Ball thru SS-2B)
b1	0-0	1	1--	2,(0-1)	O	NLS	V. Guerrero	D. Wells	3%	48%	Lineout: 2B (SS-2B)
b1	0-0	2	1--	4,(1-2)	O	NLS	S. Sosa	D. Wells	2%	50%	Strikeout Looking
											0 runs, 1 hit, 0 errors, 1 LOB. AL All-Stars 0, NL All-Stars 0.
Top of the 2nd, AL All-Stars Batting, Tied 0-0, NL All-Stars' Danny Graves facing 5-6-7											
											Danny Graves replaces Randy Johnson pitching and batting 9th
t2	0-0	0	---	4,(1-2)	O	ALS	C. Everett	D. Graves	-2%	48%	Flyball: CF (Deep CF-RF)
t2	0-0	1	---	5,(1-2)		ALS	I. Rodriguez	D. Graves	3%	50%	Single to RF (Line Drive to Short RF Line)
t2	0-0	1	1--	1,(0-0)	O	ALS	J. Dye	D. Graves	-3%	47%	Flyball: LF (Short LF)
t2	0-0	2	1--	4,(1-2)	O	ALS	T. Fryman	D. Graves	-2%	45%	Strikeout Swinging
											0 runs, 1 hit, 0 errors, 1 LOB. AL All-Stars 0, NL All-Stars 0.
Bottom of the 2nd, NL All-Stars Batting, Tied 0-0, AL All-Stars' David Wells facing 5-6-7											
b2	0-0	0	---	4,(1-2)	O	NLS	J. Kent	D. Wells	2%	47%	Groundout: 1B unassisted (2B-1B)
b2	0-0	1	---	1,(0-0)	O	NLS	A. Galarraga	D. Wells	2%	49%	Lineout: LF
b2	0-0	2	---	7,(1-2)		NLS	J. Edmonds	D. Wells	-1%	48%	Single to RF (Line Drive to Deep 2B-1B)
b2	0-0	2	1--	4,(1-2)	O	NLS	J. Kendall	D. Wells	2%	50%	Strikeout Looking
											0 runs, 1 hit, 0 errors, 1 LOB. AL All-Stars 0, NL All-Stars 0.
Top of the 3rd, AL All-Stars Batting, Tied 0-0, NL All-Stars' Kevin Brown facing 9-1-2											
											Kevin Brown replaces Danny Graves pitching and batting 9th *Mike Bordick pinch hits for David Wells (P) batting 9th*

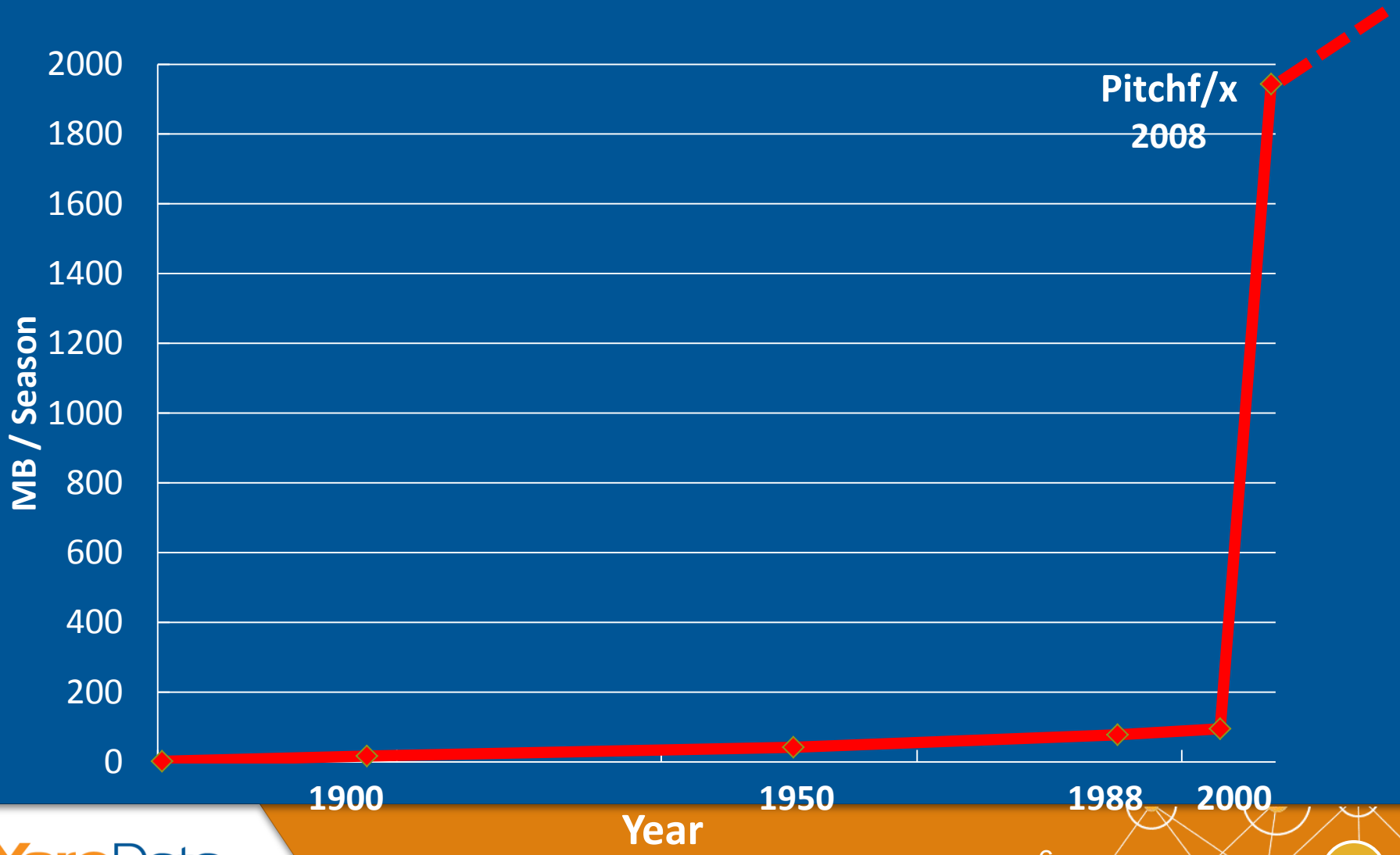
Play-by-Play

Pitchf/x

Source: MLB.com and Baseball-Reference

5

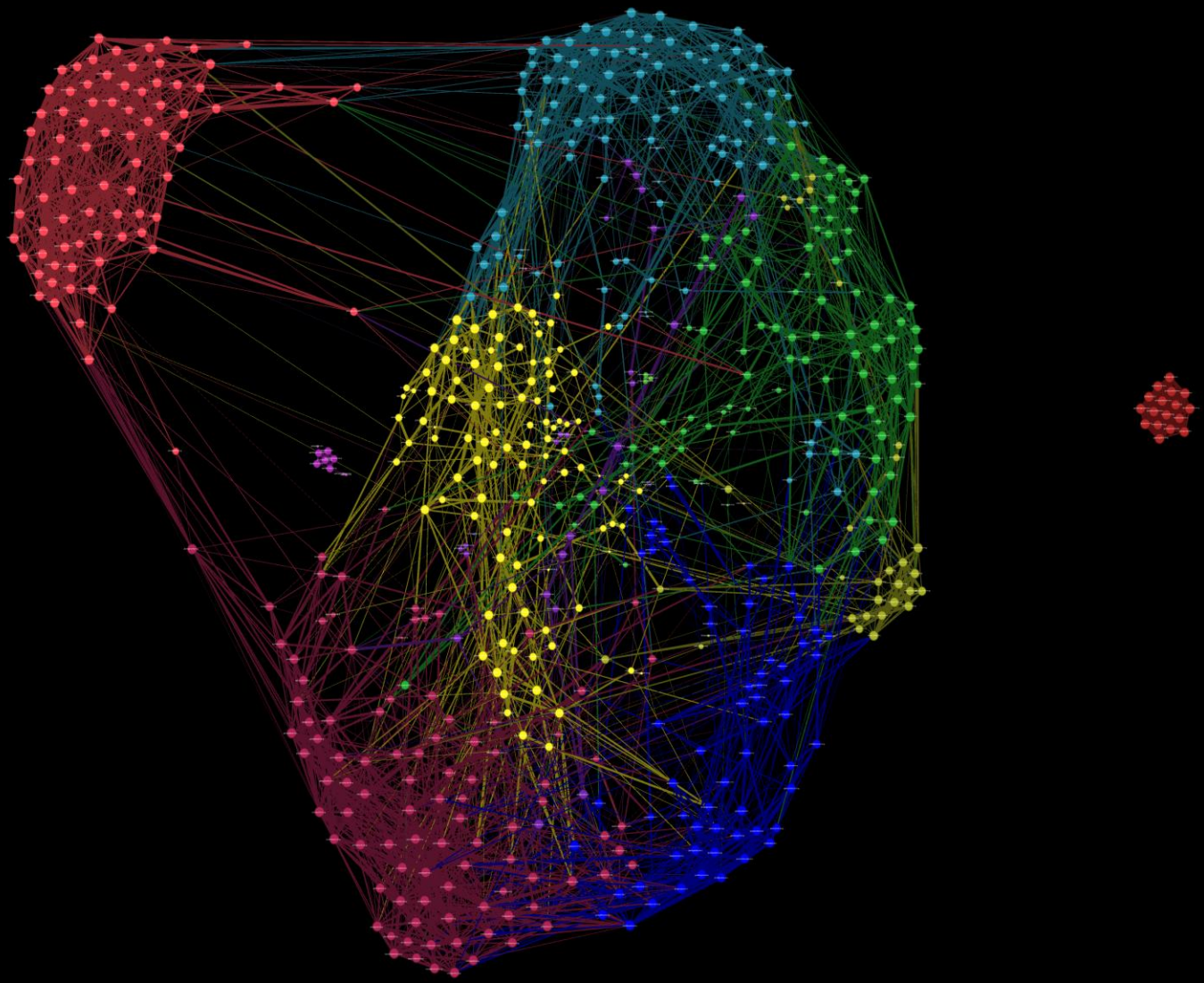
Growth in Baseball Data



Moneyball—a Breakthrough in 2003

- **Moneyball was born *before* the data explosion**
- **Built on Box Score and Play-by-Play data**
  - MLB games from 1963-2002
  - 80,000 games
  - < 1.5 G of data
- **Based on "outcome" data**

Clustering Pitchers by Attributes



Most Similar/Opposite Pairs

Profile	Vs. LHB	Vs. RHB
Similar	K Drabek—R Delgado B Chen—R Wolf J Arrieta—J Hammel F Dubront—M Harrison B Arroyo—F Garcia	C Friedrich—D Smyly J Niese—W Rodriguez L Lynn—M Scherzer P Maholm—J Russell D Price—F Dubront
Opposite	J Santana—M Harrison J Garcia—M Moore J Tomlin—M Scherzer B Morrow—K Lohse M Buehrle—M Moore	N Eovaldi—S Marcum J Collmenter—E Volquez J Cueto—J Tomlin R Porcello—P Humber M Garza—C Young

Discussion on Big Data and Technology

Current Big Data approaches focus on SEARCH

What is the RIGHT answer to a question?

But the high value Analytics is in DISCOVERY

What is the RIGHT question to ask?

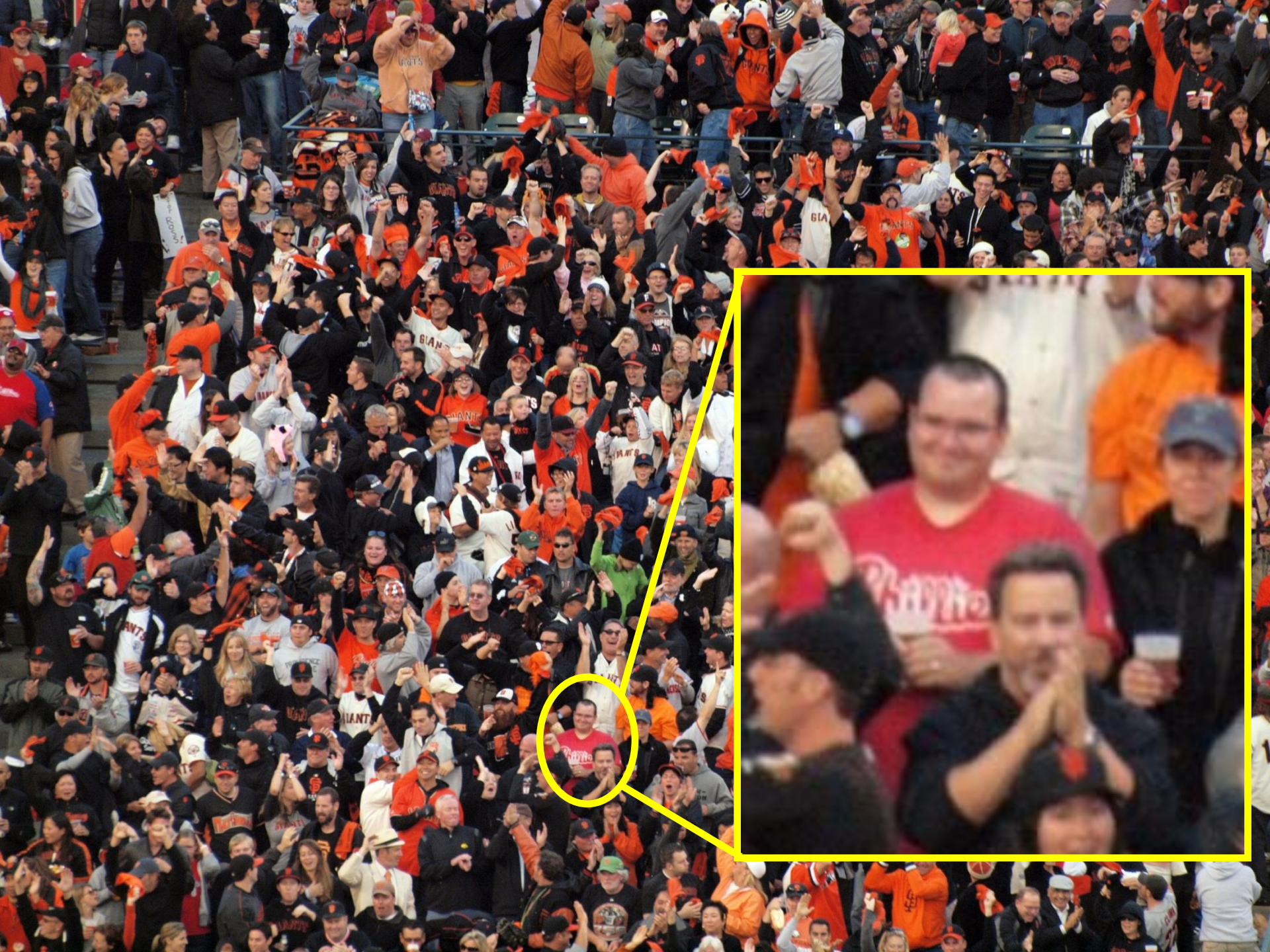
Search Problems…

- How many fans are there?
- What demographics?
- What apparel? By color/style/size?
- What concessions?

…

Maybe even…

- Fan density issues for safety, ticketing?
- Based on lighting and face direction, where in the park was this shot taken?



What is a Search?

Scalable and Efficient analysis of various of metrics on lots of Giants fans you expect to find.

Optimization is key…

Yet, requires you to know what you are going to search

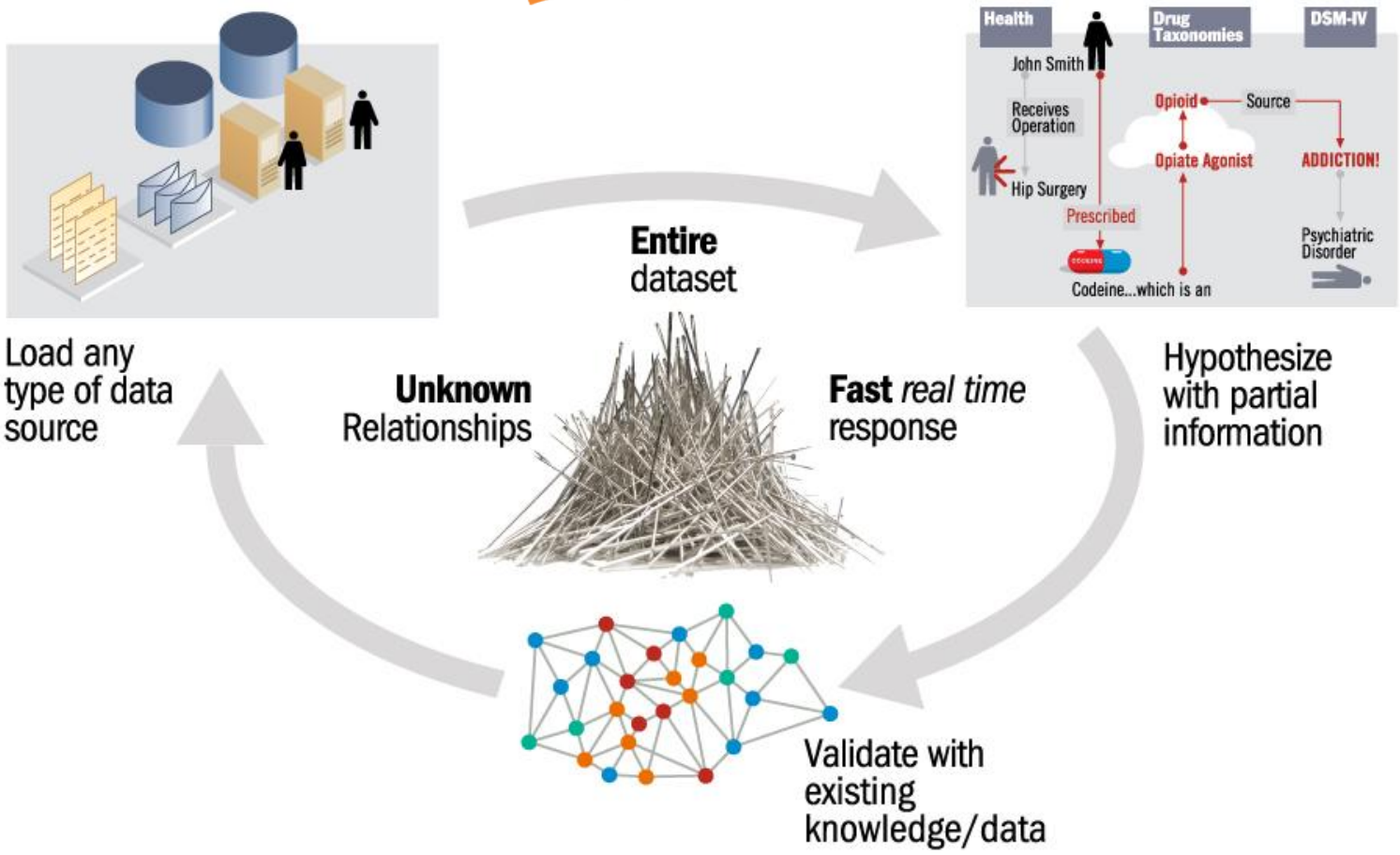
What is Discovery?

Finding the implication of the one Phillies fan you didn't expect to find

"I do not know what I am going to search"

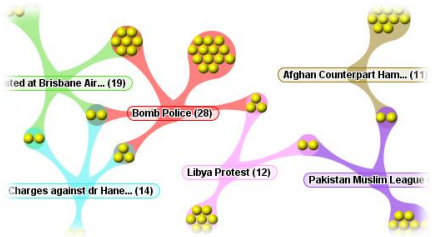
No reference Locality – No start point

Discovery through fast hypothesis validation



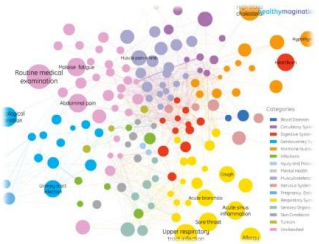
> “ In the amount of time it takes to validate one hypothesis, we can now validate 1000 hypotheses– massively improving our success rate and systematizing serendipity. ” YarcData *Government Customer*

Discovery/Graph Analytics is everywhere…



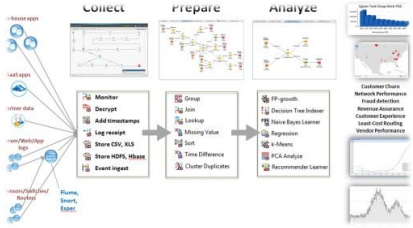
Government/Security
- Patterns of Activity Analytics
- CyberThreat Discovery
- Tax Fraud Discovery
- Crime Prediction

…



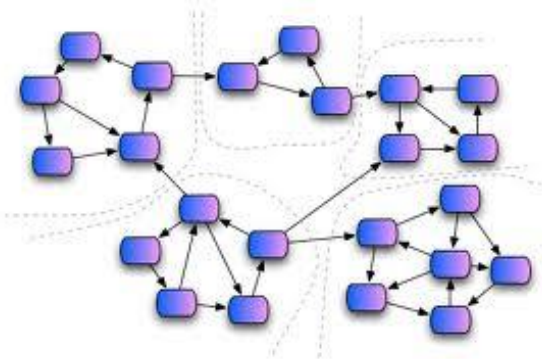
Healthcare
- Personalized Treatment
- Fraud Detection
- Efficacy of Care
- Adverse Event Clustering
- Disease Prediction

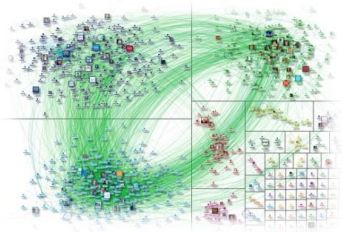
…



Energy/Resources
- Location Discovery
- Field Production Analysis
- Contingency Analysis
- Climate Modeling

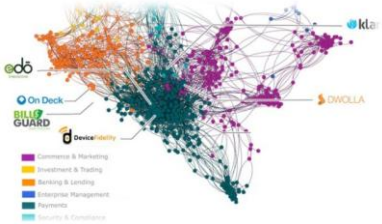
…





Telecom/Media
- Influencer Discovery
- Churn Analytics
- Behavior Analytics

…

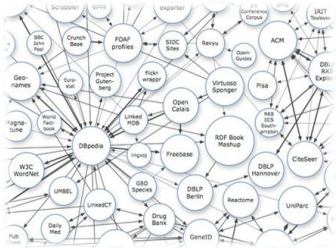


Financial Services
- Market Sensing
- News/Trading Analytics
- Counterparty/Risk
- Insider Threat
- AML/Compliance

…

Life Sciences
- Drug Discovery
- Drug Repurposing
- Clinical Trial Mining

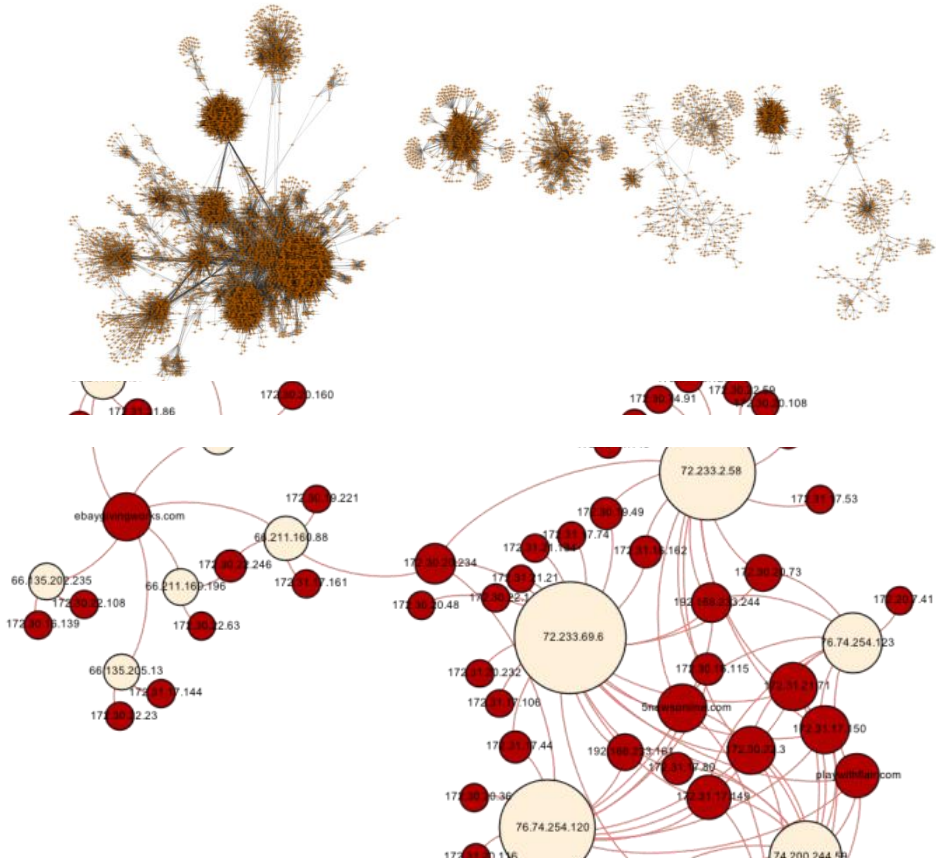
…



. . .

Discovering New Cyber Threats

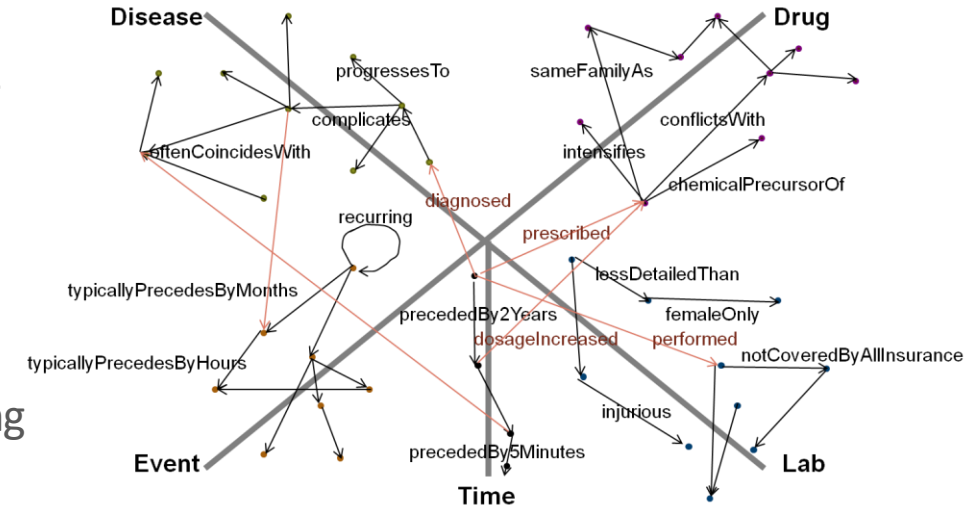
- **Goal:** Proactively identify unknown cyber threats by examining all relationships
- **Data sets:** IP, MAC, BGP, Firewall, DNS, Netflow, Whois, NVD, CIDR…
- **Technical Challenges:** Volume and Velocity of data; Temporal dependencies; Real-time response
- **Users:** Cyber Analysts
- **Usage model:** Iterative analysis of all patterns across all traffic to explore deviations in frequency of occurrence, derivative patterns of known threats and linking patterns through relationships in offline data
- **Augmenting:** Existing data appliances

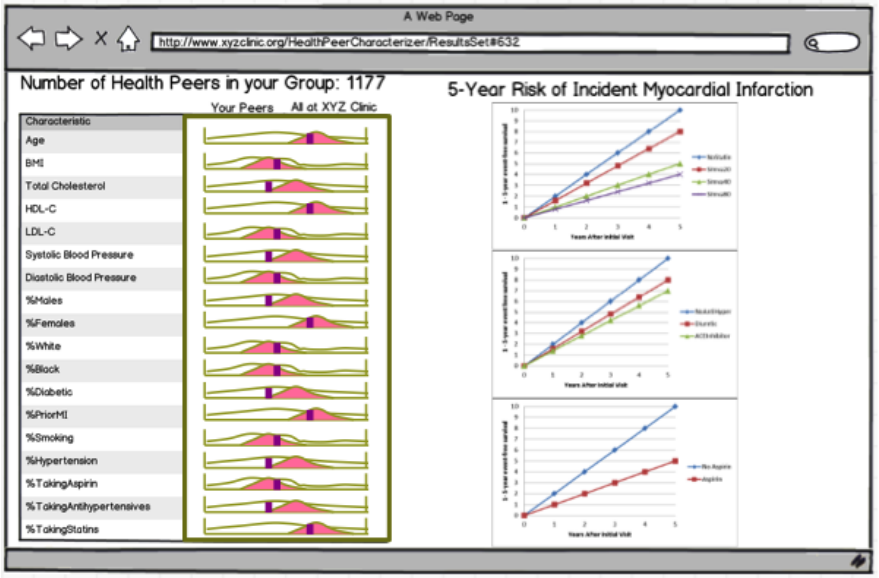


DTG	SIP	DIP	PROT	SPORT	DPORT	PKTS	BYTES	AVG BYTES per PKT	blacklist
2012-07-10T07:52:29	172.31.21.234	212.117.170.54	6	51200	80	6	558	93.00	Tor exit node
2012-07-10T08:37:59	172.31.21.234	212.117.170.54	6	51845	80	6	558	93.00	Tor exit node
2012-07-10T08:38:25	172.31.21.234	212.117.170.54	6	51846	80	8	944	118.00	Tor exit node
2012-07-10T10:59:25	172.31.21.29	212.117.170.54	6	52834	80	7	598	85.43	Tor exit node
2012-07-10T12:04:43	172.31.21.234	212.117.170.54	6	53341	80	6	558	93.00	Tor exit node
2012-07-10T14:04:19	172.31.21.29	212.117.170.54	6	54373	80	6	558	93.00	Tor exit node
2012-07-10T15:59:32	172.31.21.29	212.117.170.54	6	54830	80	6	558	93.00	Tor exit node
2012-07-10T16:45:57	172.31.20.53	65.164.25.47	6	47155	80	33	2347	71.12	SSH Brute Force
2012-07-10T16:45:58	172.31.20.53	65.164.25.47	6	47166	80	8	932	116.50	SSH Brute Force
2012-07-10T16:45:58	172.31.20.53	65.164.25.47	6	47165	80	8	933	116.63	SSH Brute Force

Discovering New Treatments

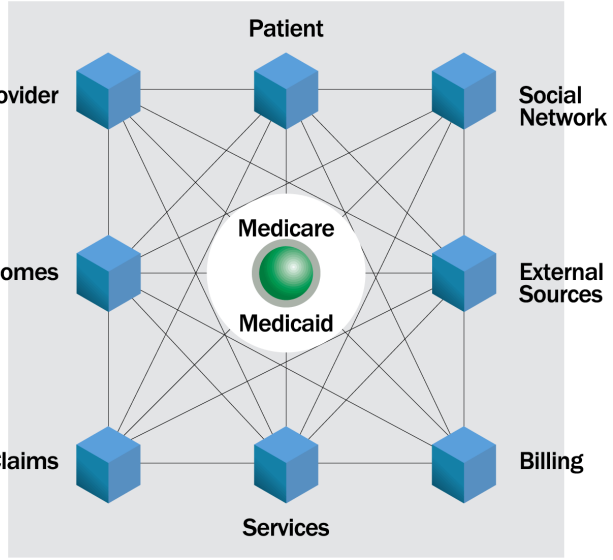
- **Goal:** Proactively identify optimal treatments for patients based on treatment results for "similar" patients
- **Data sets: Longitudinal** patient data, Family history, Genetics, Reference data
- **Technical Challenges: Ad**-hoc constantly changing definition of "similarity" across thousands of constantly changing attributes
- **Users: Doctors**
- **Usage model: Compare** results of candidate treatment options for "similar" patients based on ad-hoc physician specified patterns
- **Augmenting: Existing** data warehouse

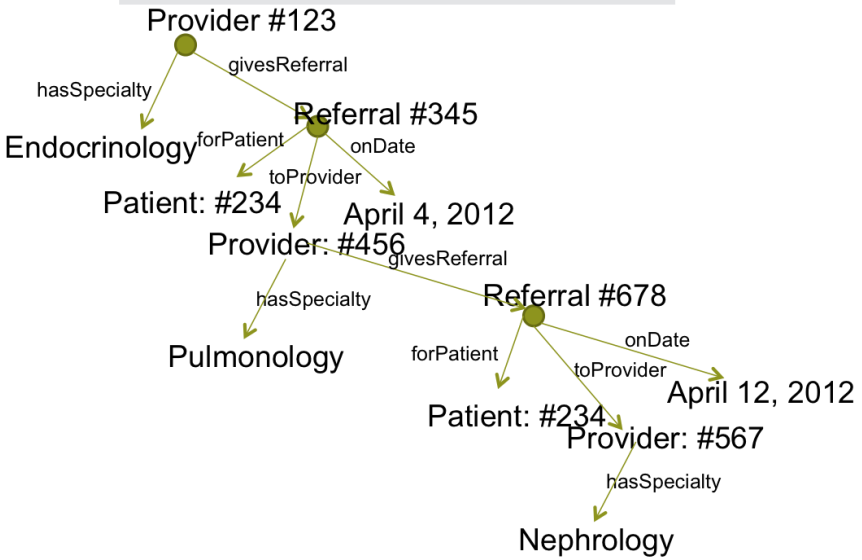




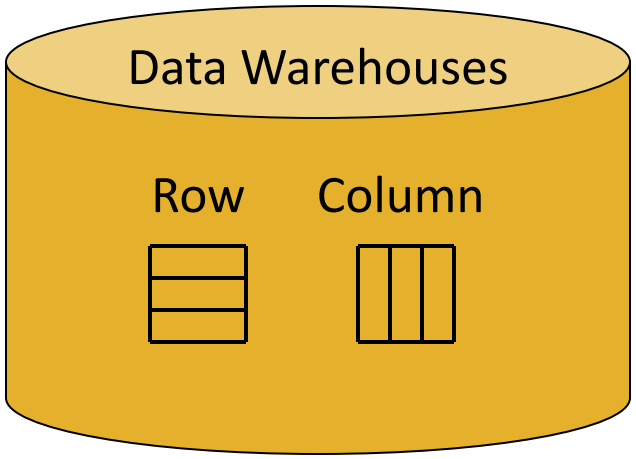
Discovery New Fraud Patterns

- **Goal:** Proactively identify new patterns of healthcare fraud (perpetrator/provider/patient nexus) by examining all healthcare relationships
- **Data sets:** Provider, Beneficiary, Policy, Claims, Billing, Services, Outcomes, Social Network, Organization…
- **Technical Challenges:** Volume and Velocity of data; Entity Resolution; Complex Inter-relationships; Temporal dependencies
- **Users:** Fraud Inspectors/Analysts
- **Usage model:** Analyze outcome and cost for various disease trajectories and identify outliers for treatment optimization and fraud investigation; Separate fraud patterns from benign treatment or legitimate errors
- **Augmenting:** Existing data warehouse, Predictive Analytics



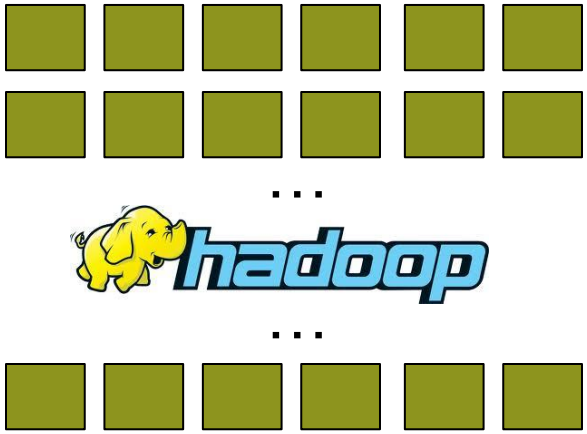


Big Data Analytics covers a lot of technologies…

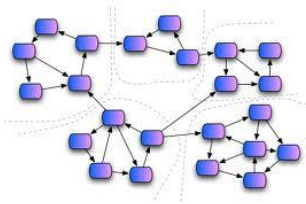


XML Databases

Document Stores

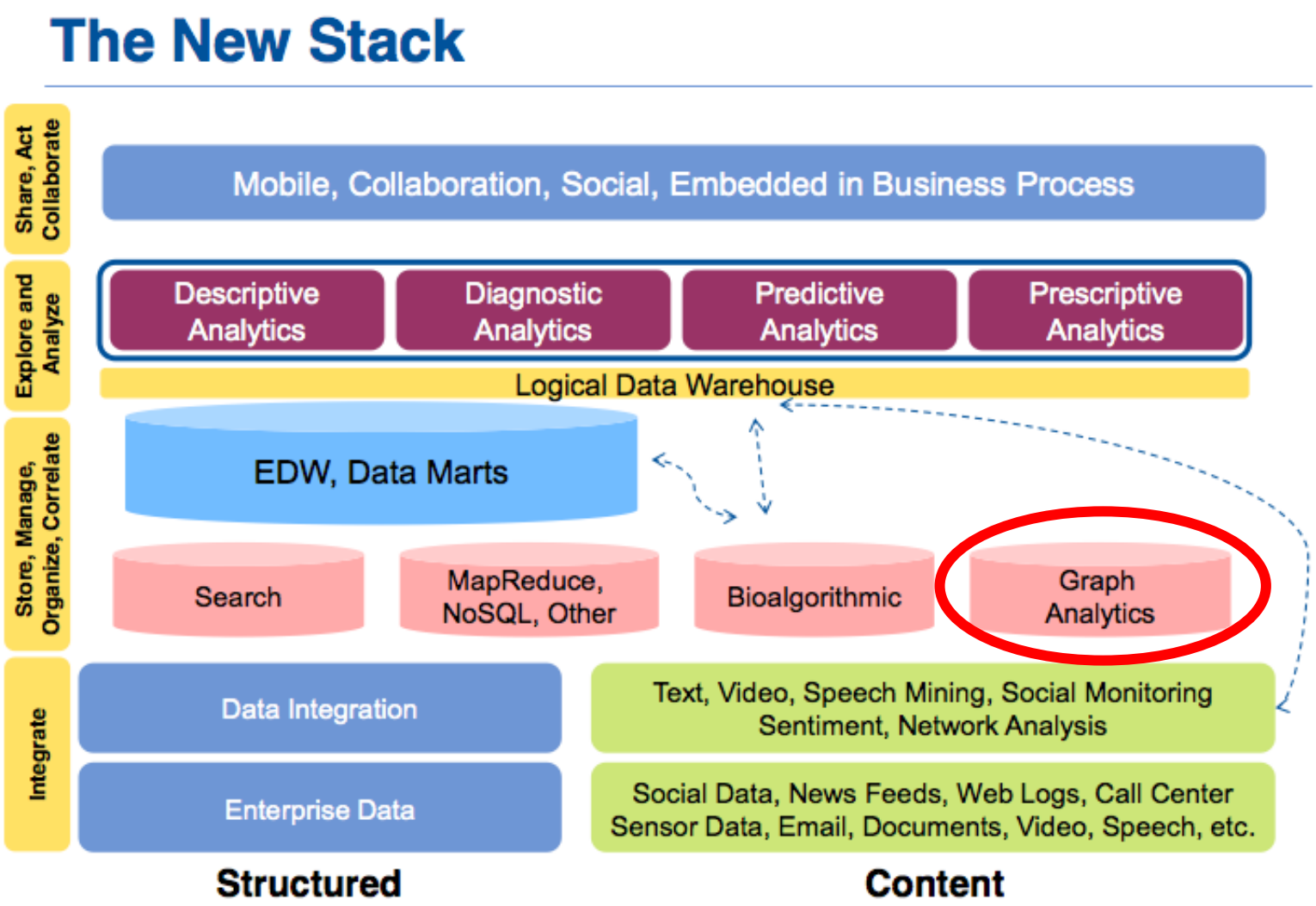


Hybrid Datastores



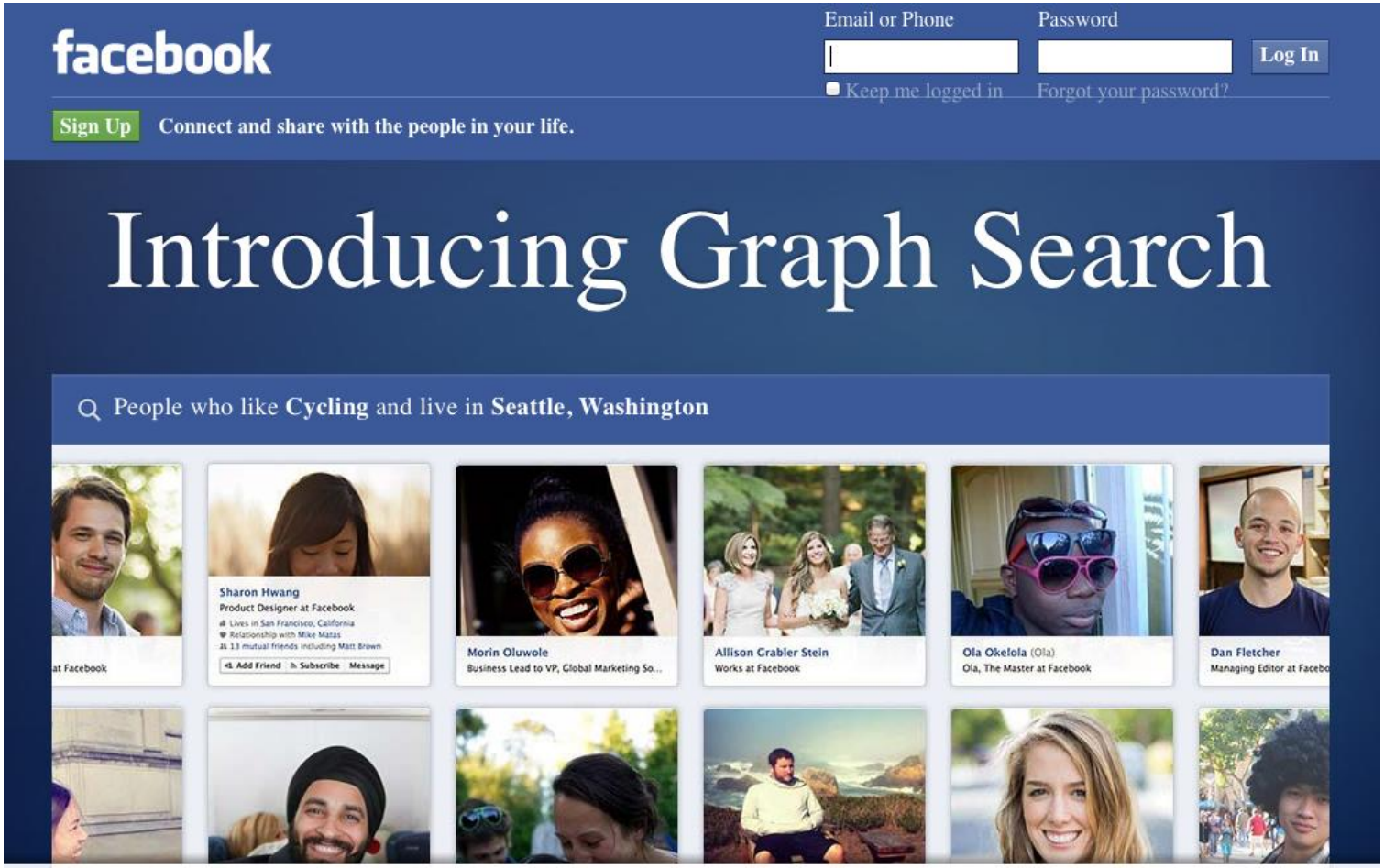
Graph Analytics

Graph Analytics is reaching an Inflection Point…



Source: Gartner BI Summit, Mar 2013

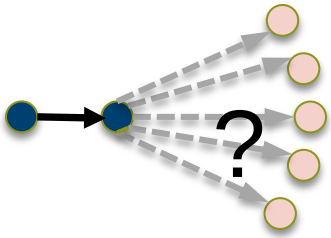
Graph Analytics is reaching an Inflection Point…



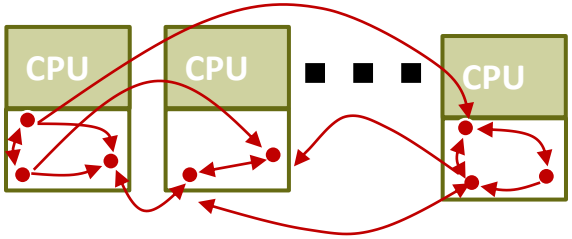
The YarcData Approach

Business Challenge:

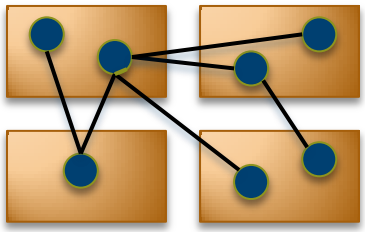
uRiKA



Large Shared Memory Architecture
Up to 512 TB



XMT2 Massively Multi-Threaded Processors
128 Threads

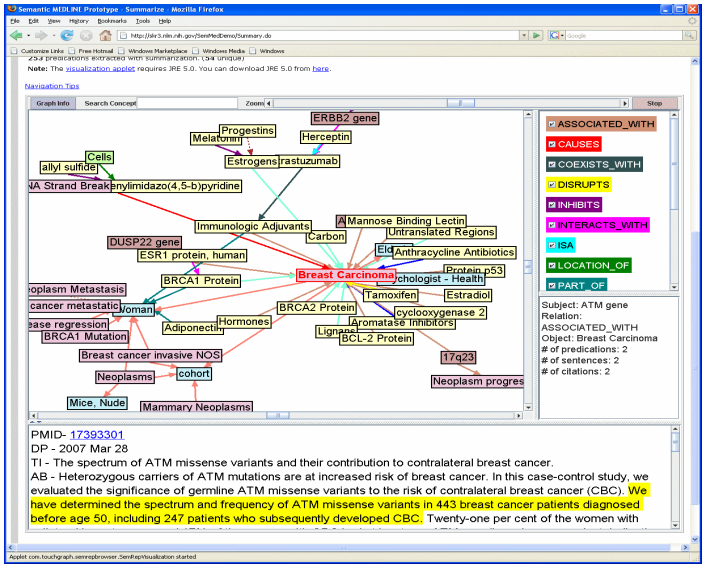


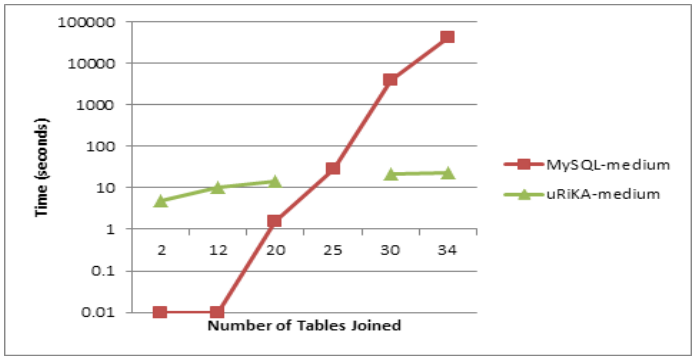
Scalable IO
Up to 350TB per Hour

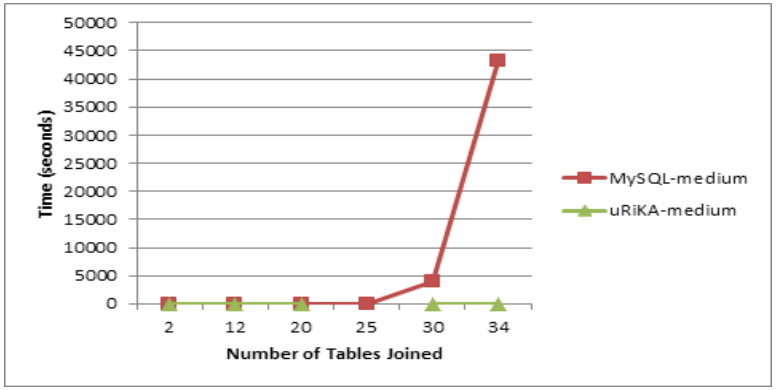
Real-time, Interactive Analytics
on Large Graph Problems

Semantic Medline at NIH-NLM

- **Current :** Web based research tool.
- **Transition:** Current systems re-engineered to leverage Urika (less than 5 days)
- **Purpose:** Build a platform users to perform increasingly complex analysis
- **Immediate Requirement :** Replicate current capability
- **Future:** Allow for increasingly complex analysis. Ability to capture and share analytics in addition to sharing data. Tailor Urika to less complex queries.







Transition:
Semantic Medline Demonstration

-Q&A Post Demo

Tim@yarcdata.com