UNI
FR

UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Big Data Integration

Philippe Cudré-Mauroux

eXascale Infolab, University of Fribourg
Switzerland

eXascale Infolab

CSNDSP, Budapest
July 20, 2018

Instant Quiz

- n-Vs of Big Data?
- Yarn?
- Spark
- Knowledge Graph?

eXascale Infolab (XI)

- New lab @ U. of Fribourg–Switzerland
- Big Data/AI infrastructures for social / semantic / scientific data

https://exascale.info/

erc European Research Council FNSNF HORIZON 2020 HASLER STIFTUNG SAP VERISIGN Google armasuisse amazon

Exascale Data Deluge

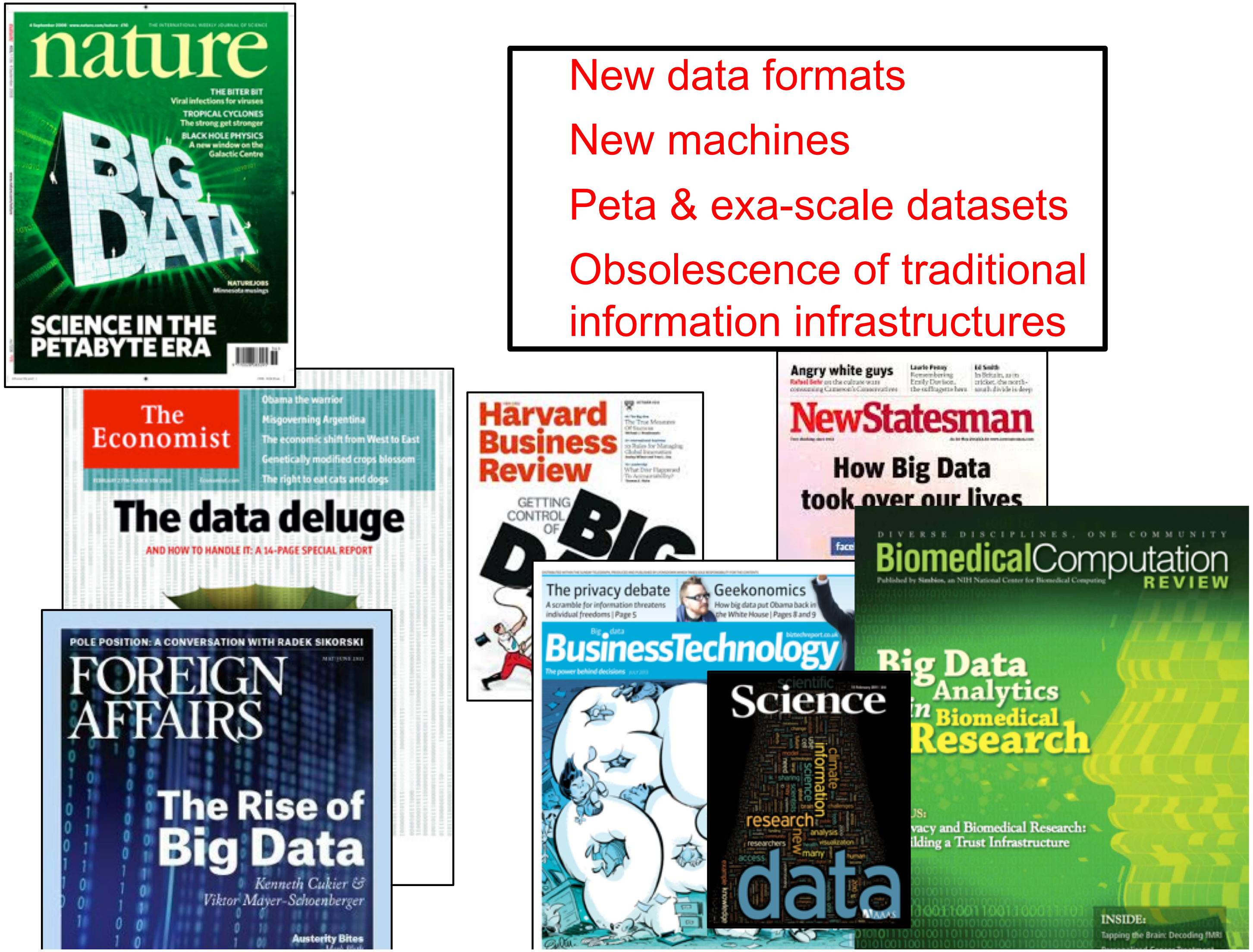
- Web companies
  - Google
  - Ebay
  - Yahoo
- Science
  - Biology
  - Astronomy
  - Remote Sensing
- Financial services, retail companies governments, etc.

New data formats

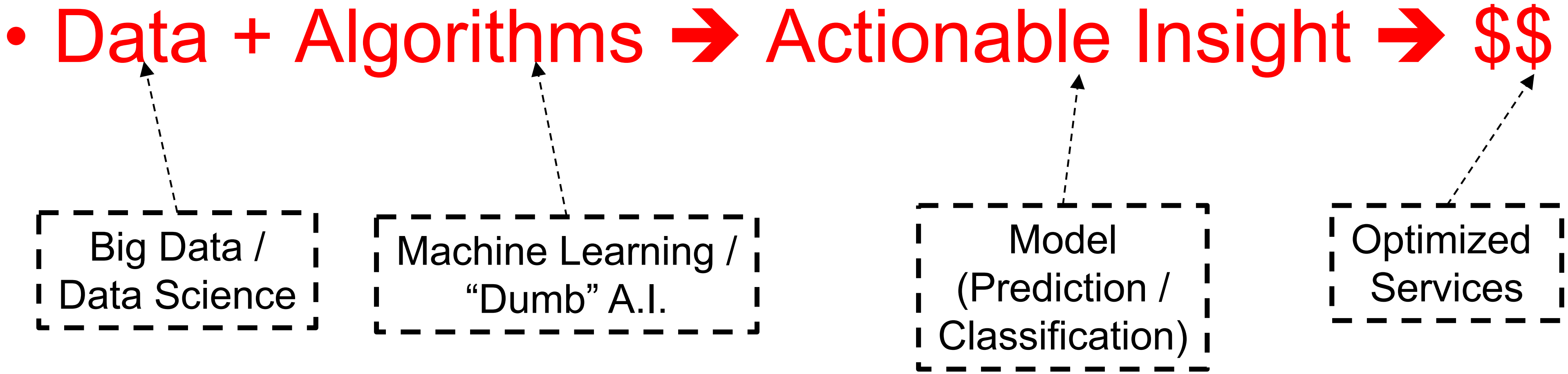
New machines

Peta & exa-scale datasets

Obsolescence of traditional information infrastructures


nature
BIG DATA
SCIENCE IN THE PETABYTE ERA
The Economist
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT
Harvard Business Review
GETTING CONTROL OF BIG DATA
NewStatesman
How Big Data took over our lives
BiomedicalComputation REVIEW
Big Data Analytics in Biomedical Research
FOREIGN AFFAIRS
The Rise of Big Data
Kenneth Cukier & Viktor Mayer-Schoenberger
BusinessTechnology
The privacy debate
Geekonomics
Science
data

Data is the new Oil



- Data + Algorithms ➔ Actionable Insight ➔ $$

Big Data / Data Science

Machine Learning / “Dumb” A.I.

Model (Prediction / Classification)

Optimized Services

Big data can generate significant financial value across sectors

US health care

- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration

- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data

- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail

- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

Manufacturing

- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

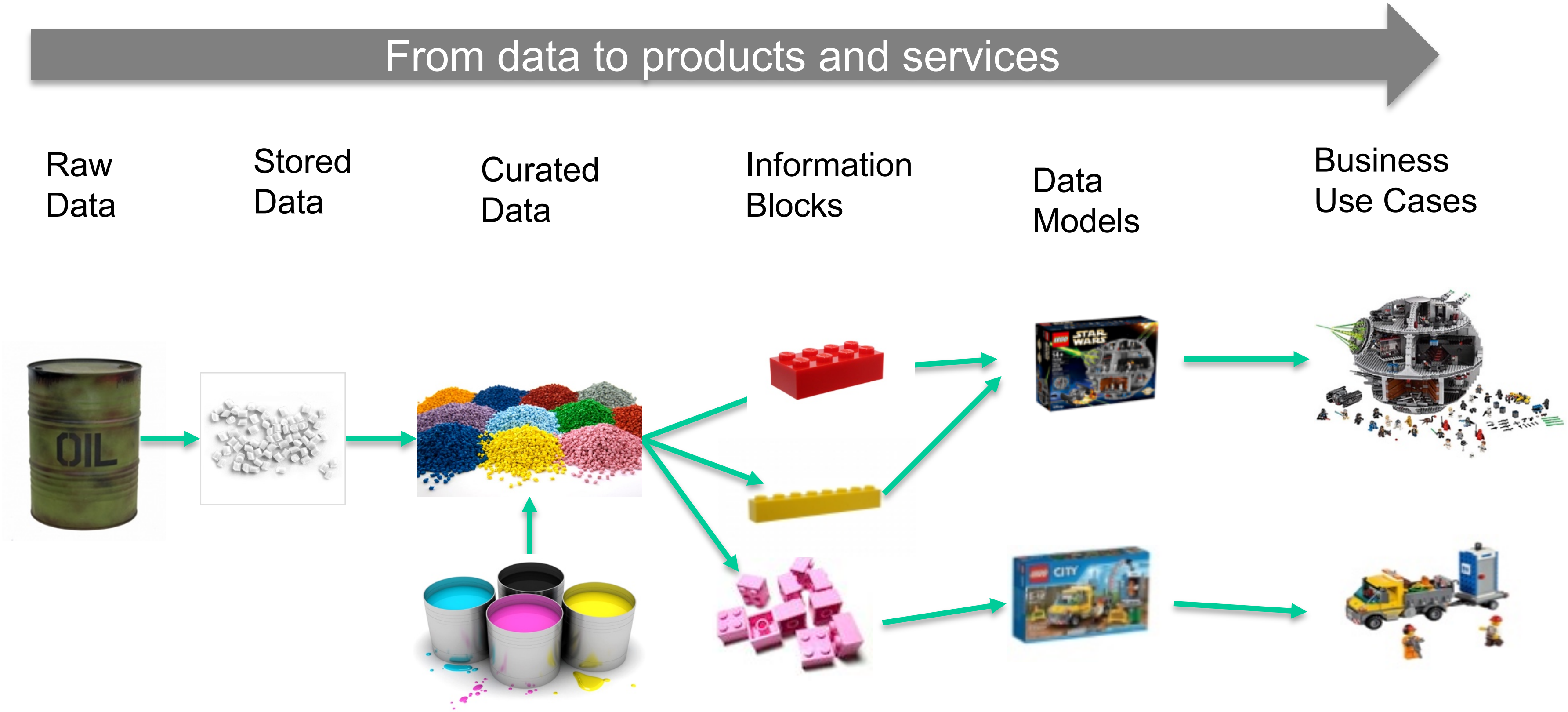
SOURCE: McKinsey Global Institute analysis

The *n*-Vs of Big Data

- Volume
  - amount of data (scale *out* not *up*)
- Velocity
  - speed of data in and out
- Variety
  - range of data types and sources

[Gartner 2012] *"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization"*

Data vs. Traditional Assets



From data to products and services

Raw Data

Stored Data

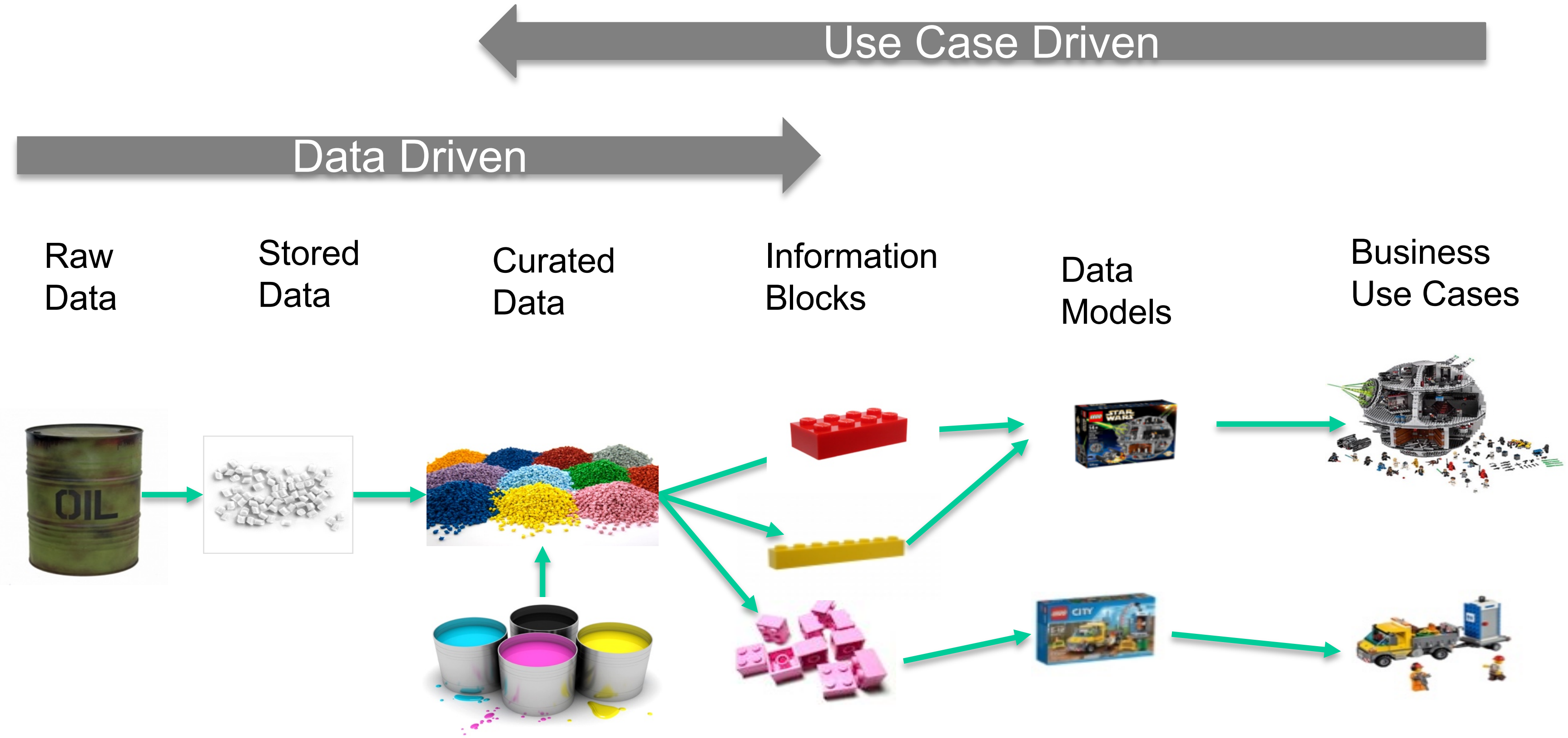
Curated Data

Information Blocks

Data Models

Business Use Cases

Use Case or Data Driven



Use Case Driven

Data Driven

Raw Data

Stored Data

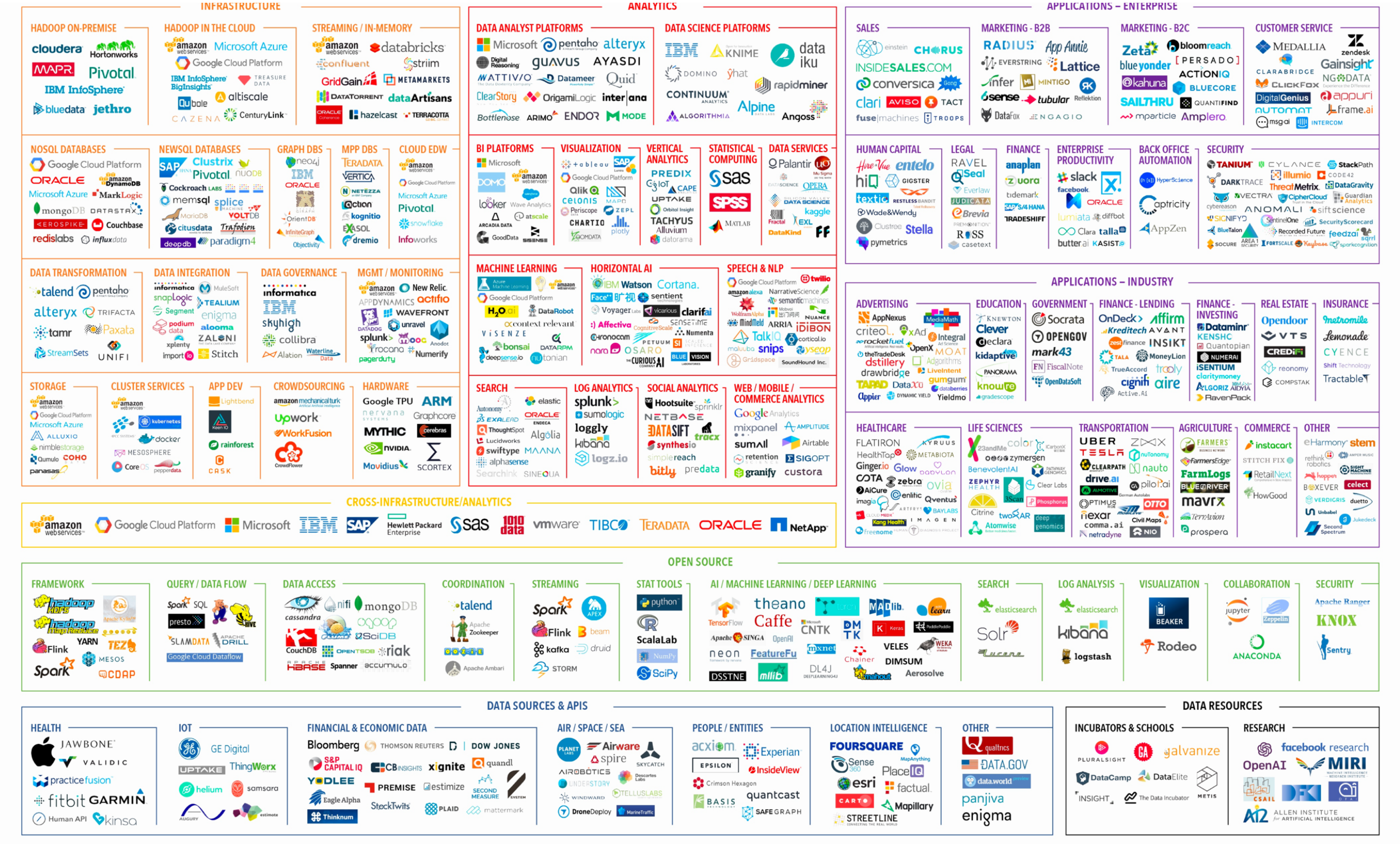
Curated Data

Information Blocks

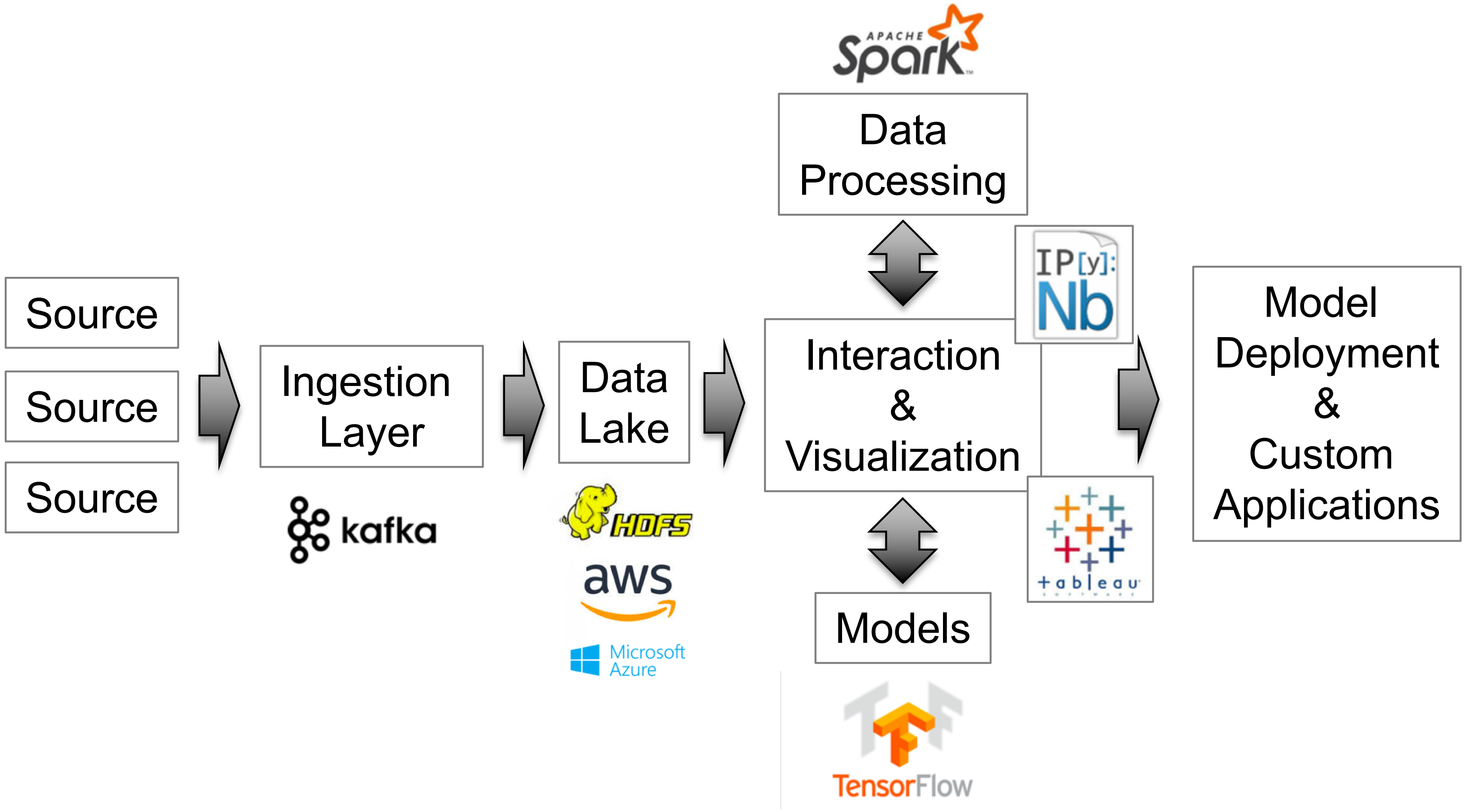
Data Models

Business Use Cases

Big Data Landscape



Typical Big Data Architecture (circa 2018)



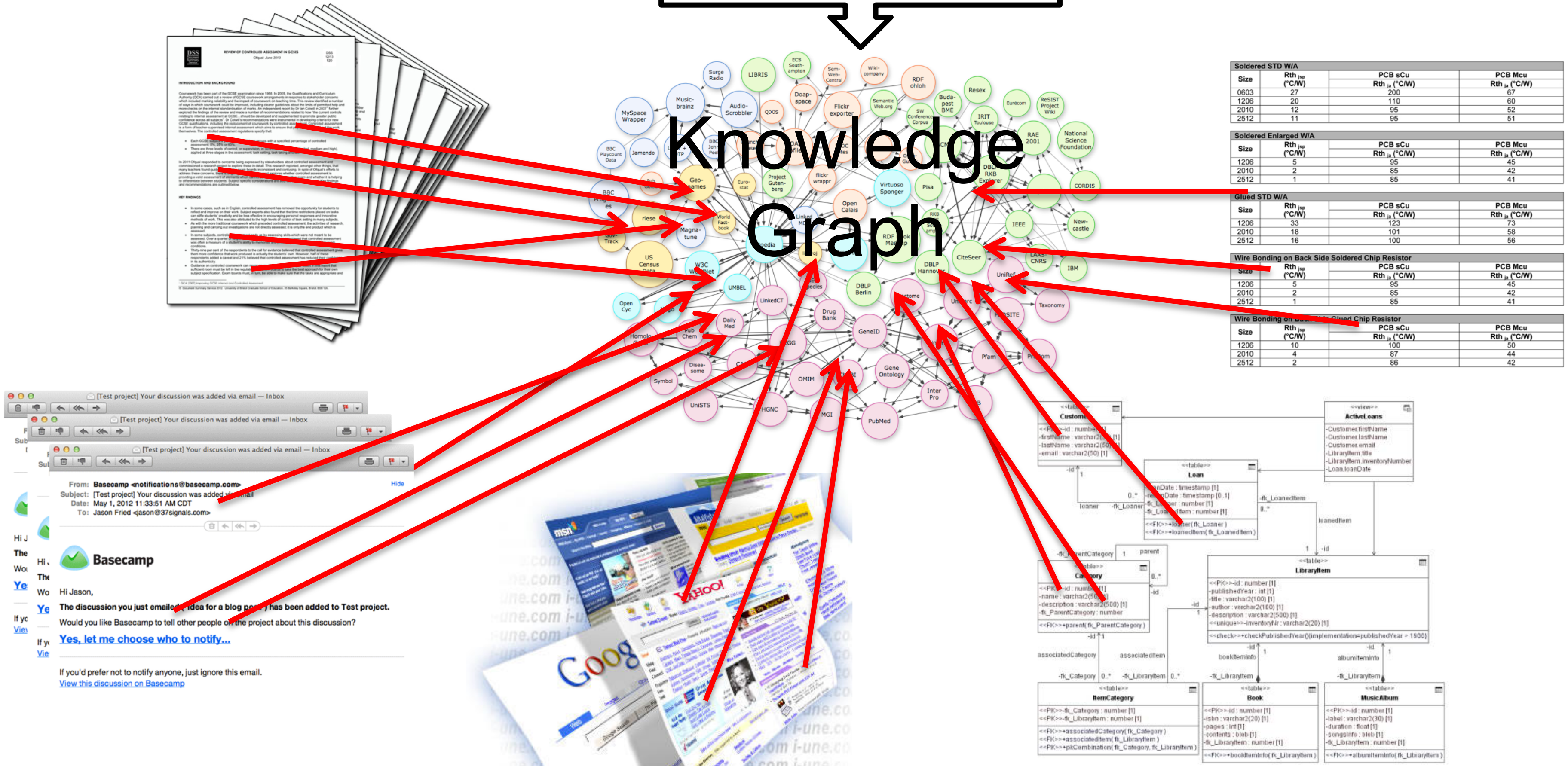
The *n*-Vs of Big Data

- Volume
  - amount of data
- Velocity
  - speed of data in and out
- Variety (*fusing* n data sources as an input to a model)
  - range of data types and sources

Entity-Centric Data Fusion

Captures both direct and indirect relationships

Higher-level apps



From: Basecamp <notifications@basecamp.com>
Subject: [Test project] Your discussion was added via email
Date: May 1, 2012 11:33:51 AM CDT
To: Jason Fried <jason@37signals.com>

Hi Jason,

The discussion you just emailed (Idea for a blog post) has been added to Test project.

Would you like Basecamp to tell other people on the project about this discussion?

Yes, let me choose who to notify...

If you'd prefer not to notify anyone, just ignore this email.
View this discussion on Basecamp

Soldered STD W/A

Size	Rth $_{jsp}$ (°C/W)	PCB sCu	PCB Mcu
		Rth $_{ja}$ (°C/W)	Rth $_{ja}$ (°C/W)
0603	27	200	67
1206	20	110	60
2010	12	95	52
2512	11	95	51

Soldered Enlarged W/A

Size	Rth $_{jsp}$ (°C/W)	PCB sCu	PCB Mcu
		Rth $_{ja}$ (°C/W)	Rth $_{ja}$ (°C/W)
1206	5	95	45
2010	2	85	42
2512	1	85	41

Glued STD W/A

Size	Rth $_{jsp}$ (°C/W)	PCB sCu	PCB Mcu
		Rth $_{ja}$ (°C/W)	Rth $_{ja}$ (°C/W)
1206	33	123	73
2010	18	101	58
2512	16	100	56

Wire Bonding on Back Side Soldered Chip Resistor

Size	Rth $_{jsp}$ (°C/W)	PCB sCu	PCB Mcu
		Rth $_{ja}$ (°C/W)	Rth $_{ja}$ (°C/W)
1206	5	95	45
2010	2	85	42
2512	1	85	41

Wire Bonding on Back Side Glued Chip Resistor

Size	Rth $_{jsp}$ (°C/W)	PCB sCu	PCB Mcu
		Rth $_{ja}$ (°C/W)	Rth $_{ja}$ (°C/W)
1206	10	100	50
2010	4	87	44
2512	2	86	42

Three Big Data Fusion Applications

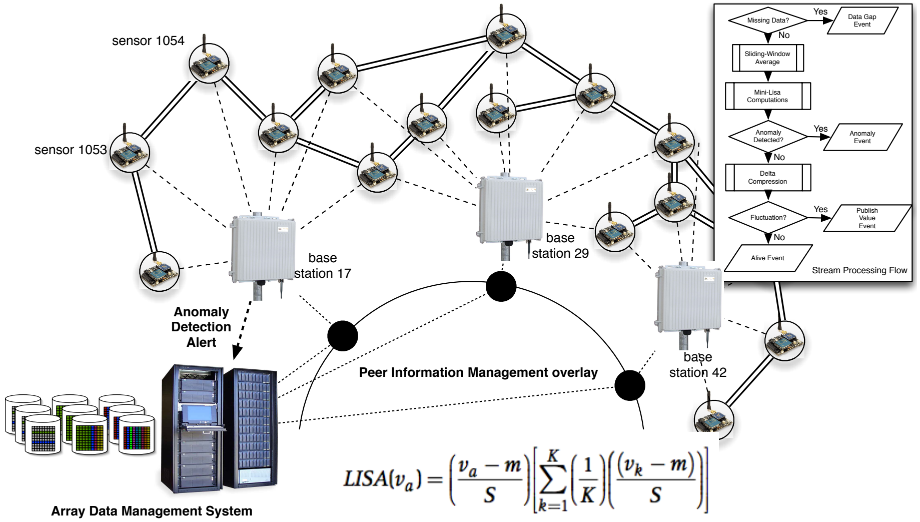
1. Anomaly Detection for Smart Cities
2. Crime Prediction using Data Fusion
3. Infrastructure Monitoring

1. Anomaly Detection for Smart Cities

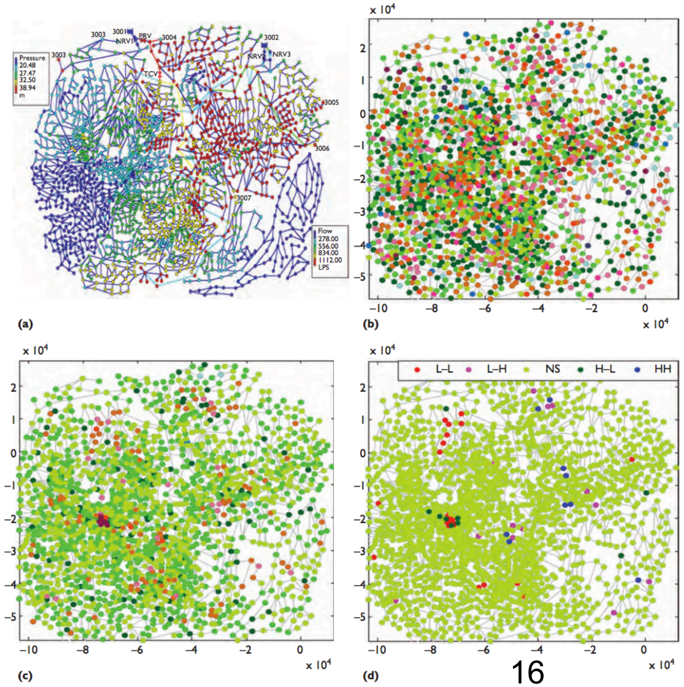
- Detecting leaks / pipe bursts / contamination in real-time for water distribution networks

Sensors installed in the water pipes!

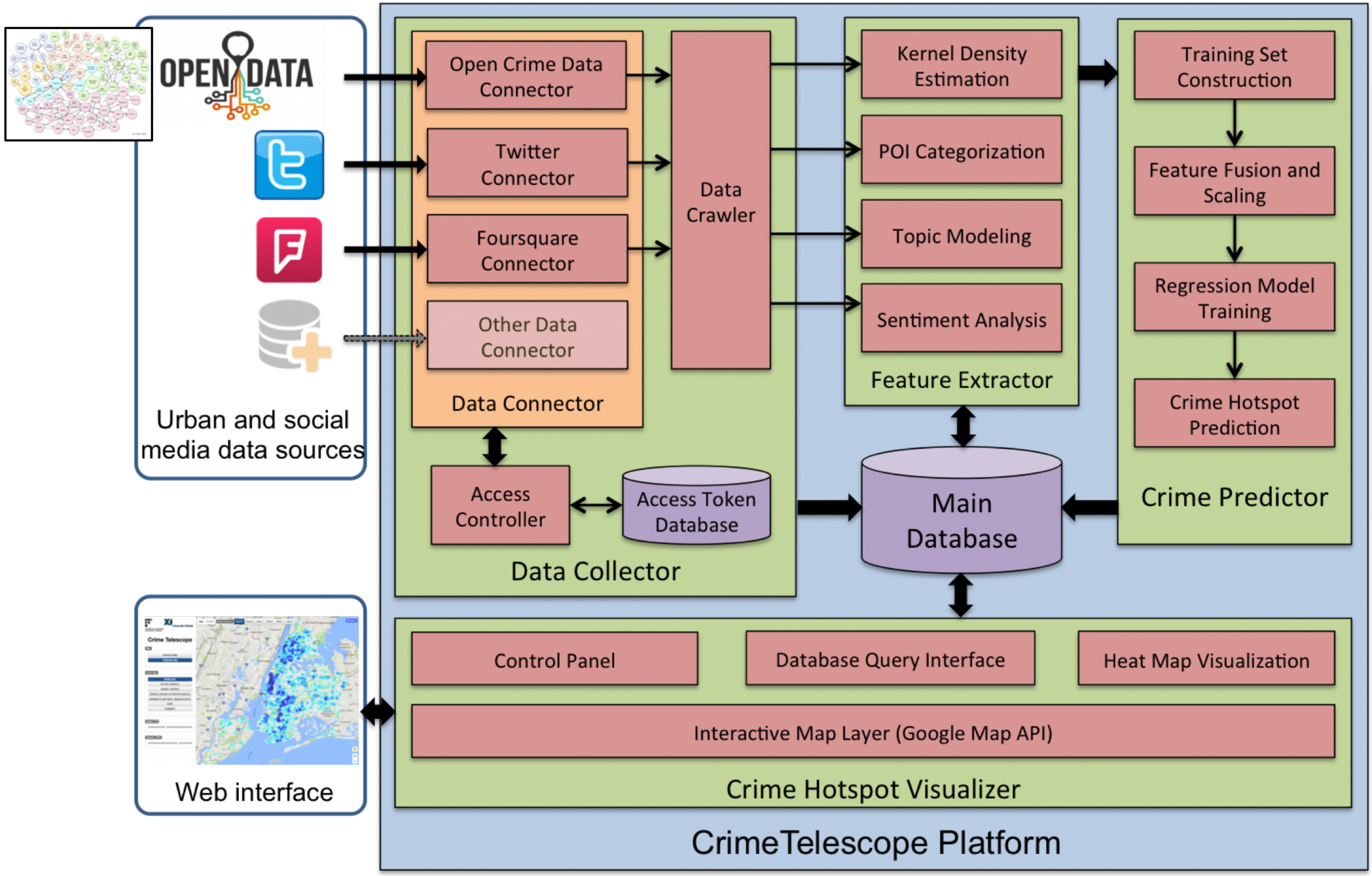
- Spatial + temporal statistical processing (mini-Lisas)
- Stream processing (Storm) + Array processing (SciDB)



$$LISA(v_a) = \left(\frac{v_a - m}{S}\right)\left[\sum_{k=1}^{K}\left(\frac{1}{K}\right)\left(\frac{(v_k - m)}{S}\right)\right]$$



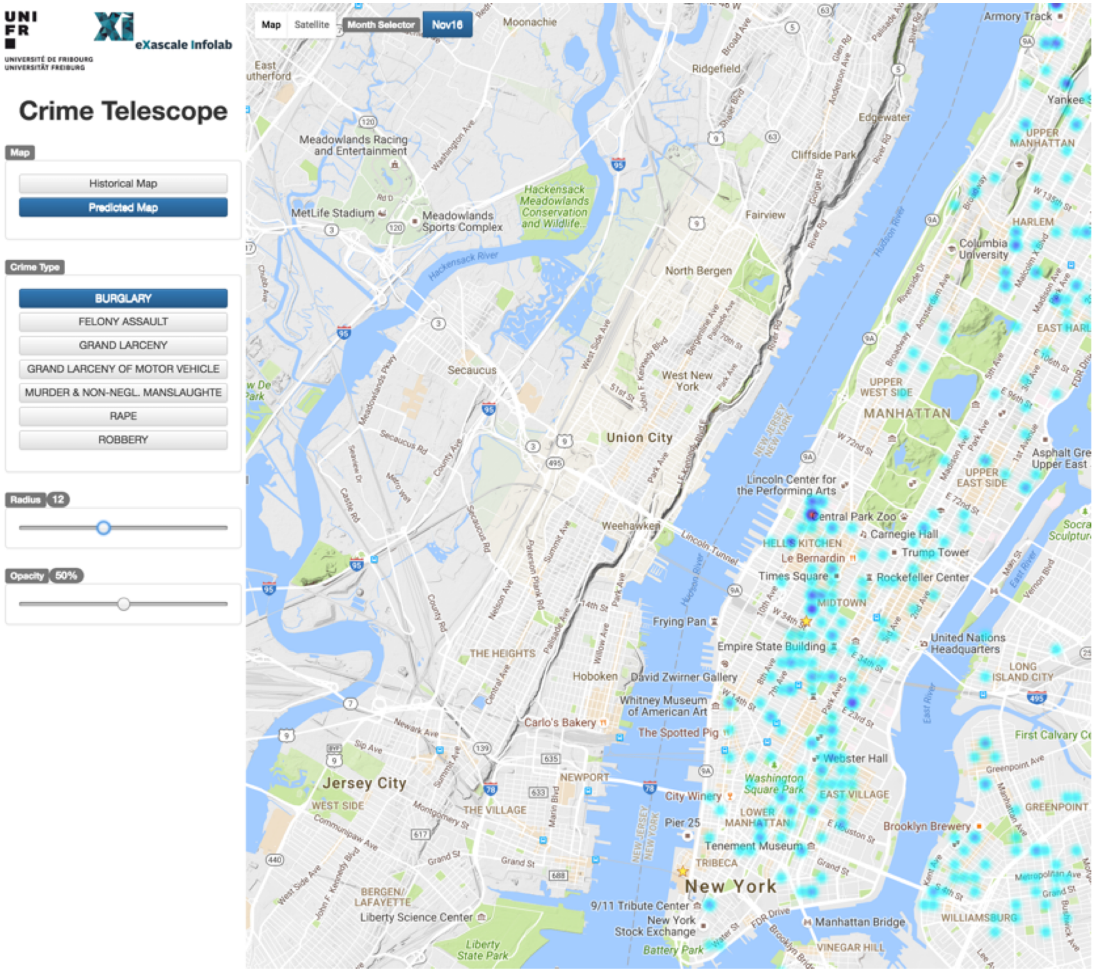
2. Crime Prediction using Big Data Fusion



- How to predict crime hotspots more accurately?
- Fusion of historical, urban & social data

CrimeTelescope: Crime Hotspot Prediction based on Urban and Social Media Data Fusion. D.Yang, T. Heaney, A. Tonon, L. Wang, P. Cudre-Mauroux. WWWJ 2017.

Crime Prediction using Data Fusion



Crime Telescope

Map: Predicted Map

Crime Type: BURGLARY

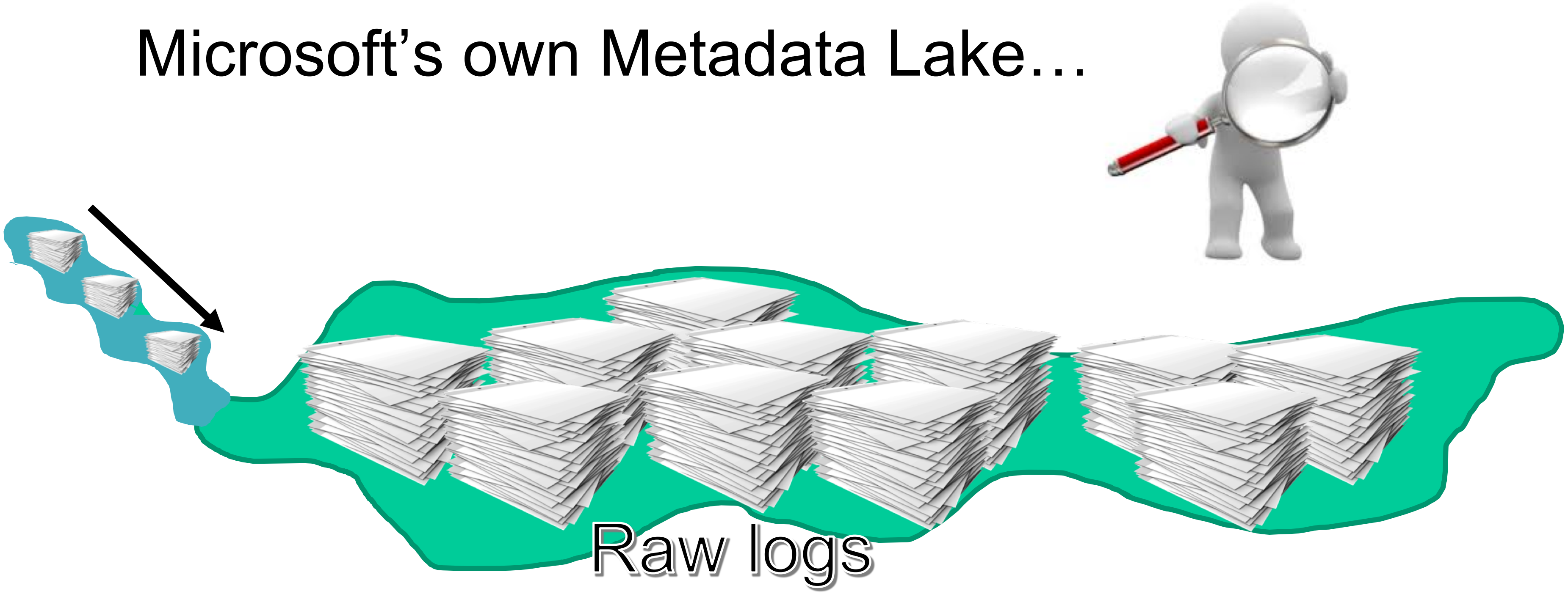
Radius 12

Opacity 50%

Month Selector: Nov16

3. Infrastructure Monitoring

Microsoft’s own Metadata Lake…

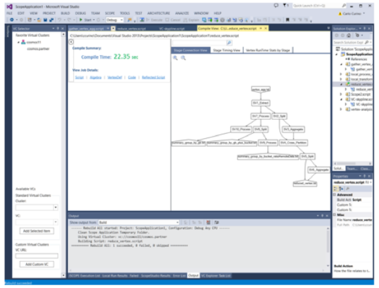


Example 1: Job pipeline analysis (state-of-the-art)

- *User: "I need help with my ML experiment processing Clicklogs"*
- *Ops / Engineer:*

Dig through many tables (from cooked logs)

+



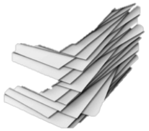
```
------ Rebuild All started: Project: ScopeApplication1, Configuration: Debug Any CPU ------
   Clean Scope Application Temporary Folder.
   Using Virtual Cluster: vc://cosmos11/cosmos.partner
   Building Script: reduce_vertex.script
========== Rebuild All: 1 succeeded, 0 failed, 0 skipped ==========
```

Write ad-hoc analytics

+

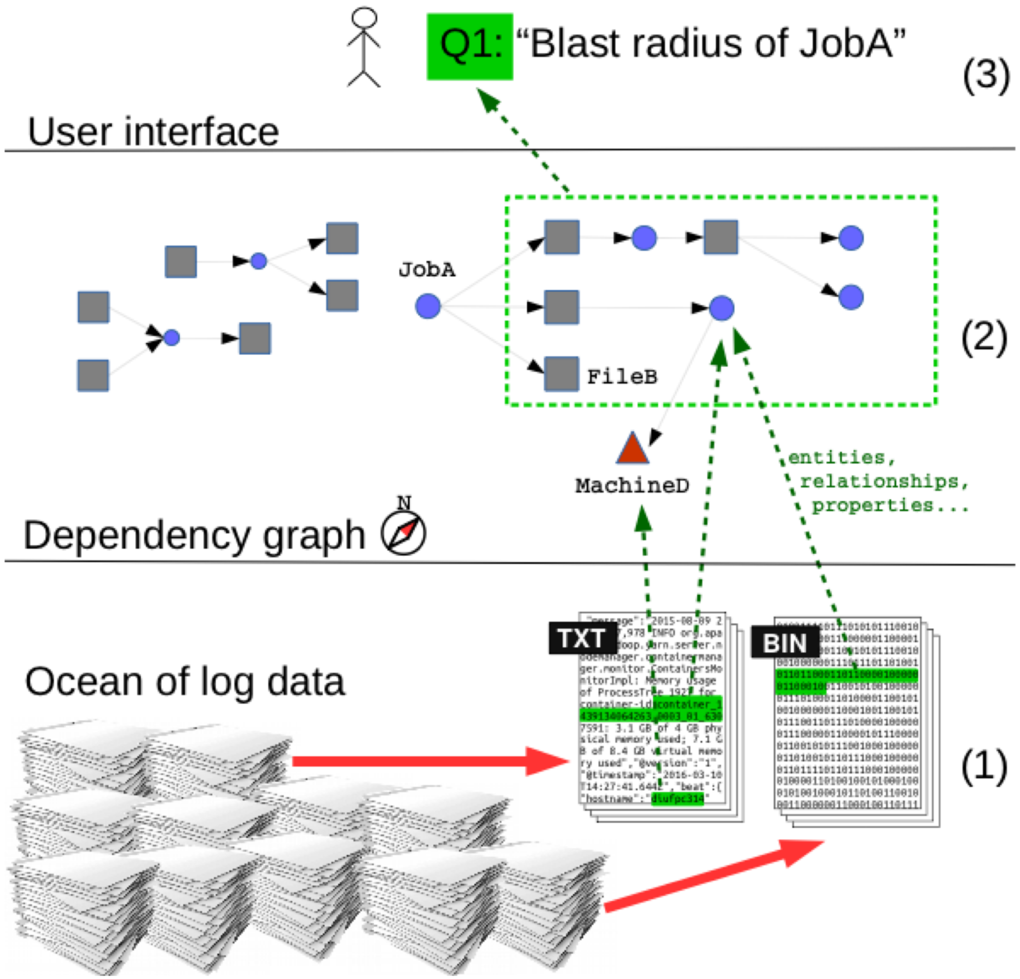
Wait 2 hours (drink coffee)

=



Manual inspect
Job exec plan
(XML blobs)

Our Solution: Guider

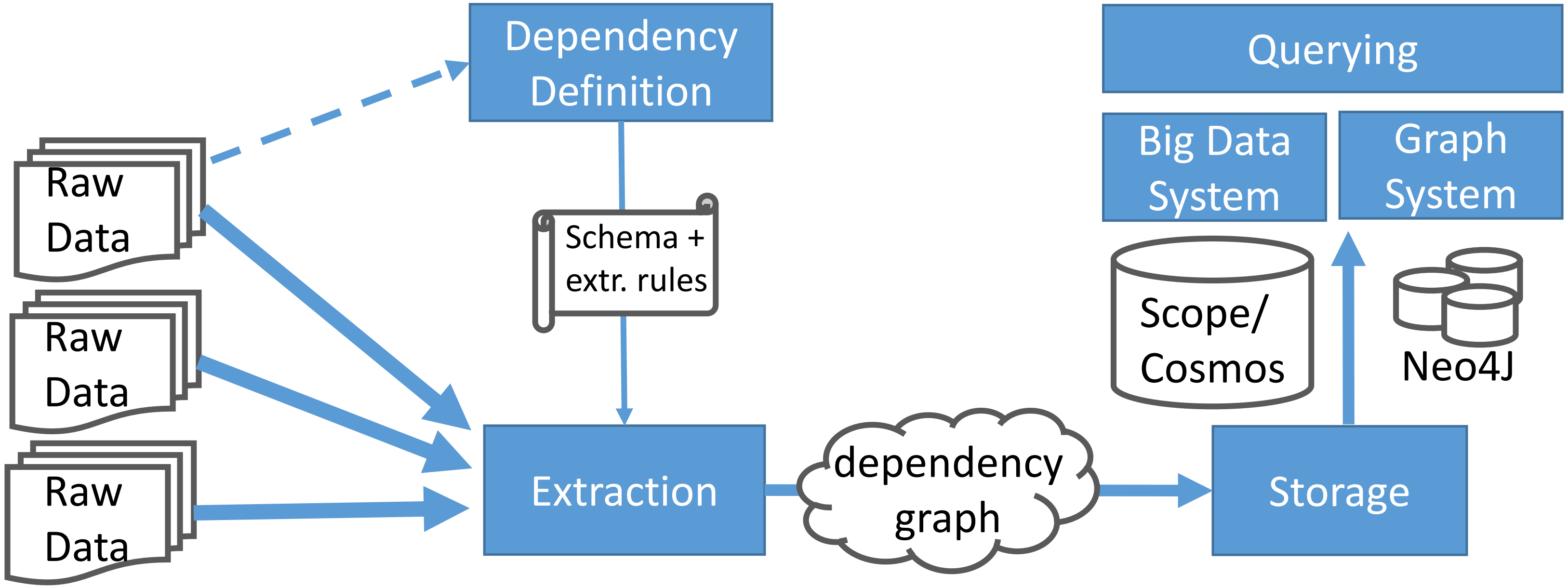


(3) User-level queries return bytes of aggregated data.

(2) Entity graph that represents a lightweight "skeleton" of the logs used for navigation

(1) Petabytes of daily logs

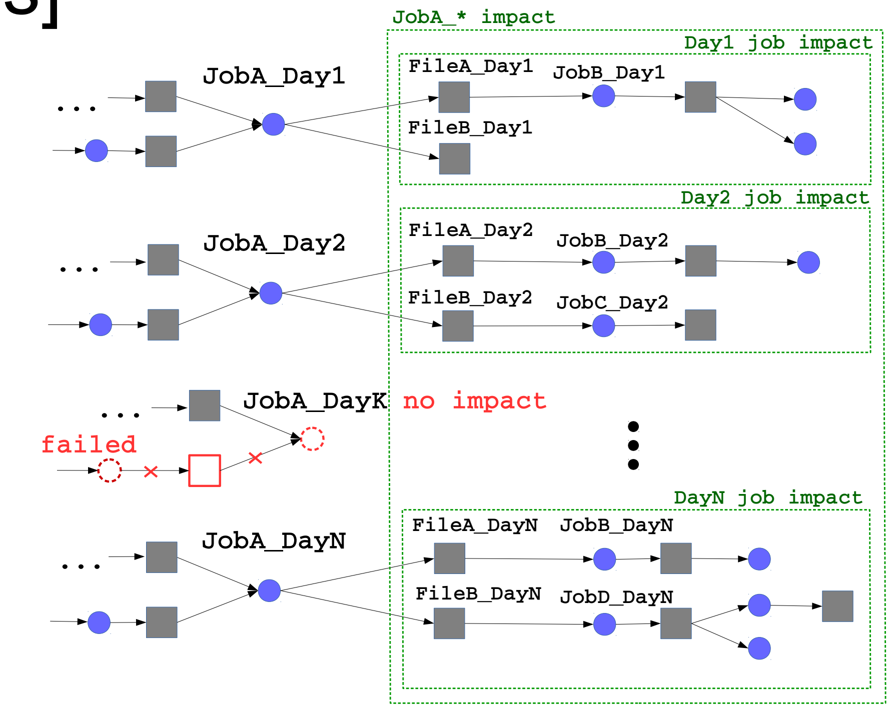
Guider Architecture



Dependency-Driven Analytics: a Compass for Uncharted Data Oceans.
Ruslan Mavlyutovm, Carlo Curinom, Boris Asipovm, and Phil Cudre-Mauroux. CIDR 2017

Guider Use-Cases

1. Auditing and Compliance [in production]
2. Job Scheduling [Morpheus]
3. Global Job Ranking
4. Datacenter migration



Thanks for your Attention!

https://exascale.info