CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

BIG DATA AS A SERVICE

Asst. Prof. Natawut Nupairoj, Ph.D.
Dept. of Computing Engineering
Faculty of Engineering
Chulalongkorn University

natawut.n@chula.ac.th
@natawutn
http://natawutn.wordpress.com
http://www.slideshare.net/natawutnupairoj

“Data is a new class of economic asset, like currency and gold” - World Economic Forum

WELCOME TO DATA-DRIVEN ECONOMY

In July 2014, the European Commission outlined a new strategy on Big Data, supporting and accelerating the transition towards a data-driven economy in Europe

In Feb 2015, The White House appointed the first US chief data scientist

As of today, US Government's open data publishes more than 190,000 datasets to the public

(our data.go.th has 506 datasets as of this morning)

The Smart City Of The Future Will Bring Big Data To A New Level

18 JUN

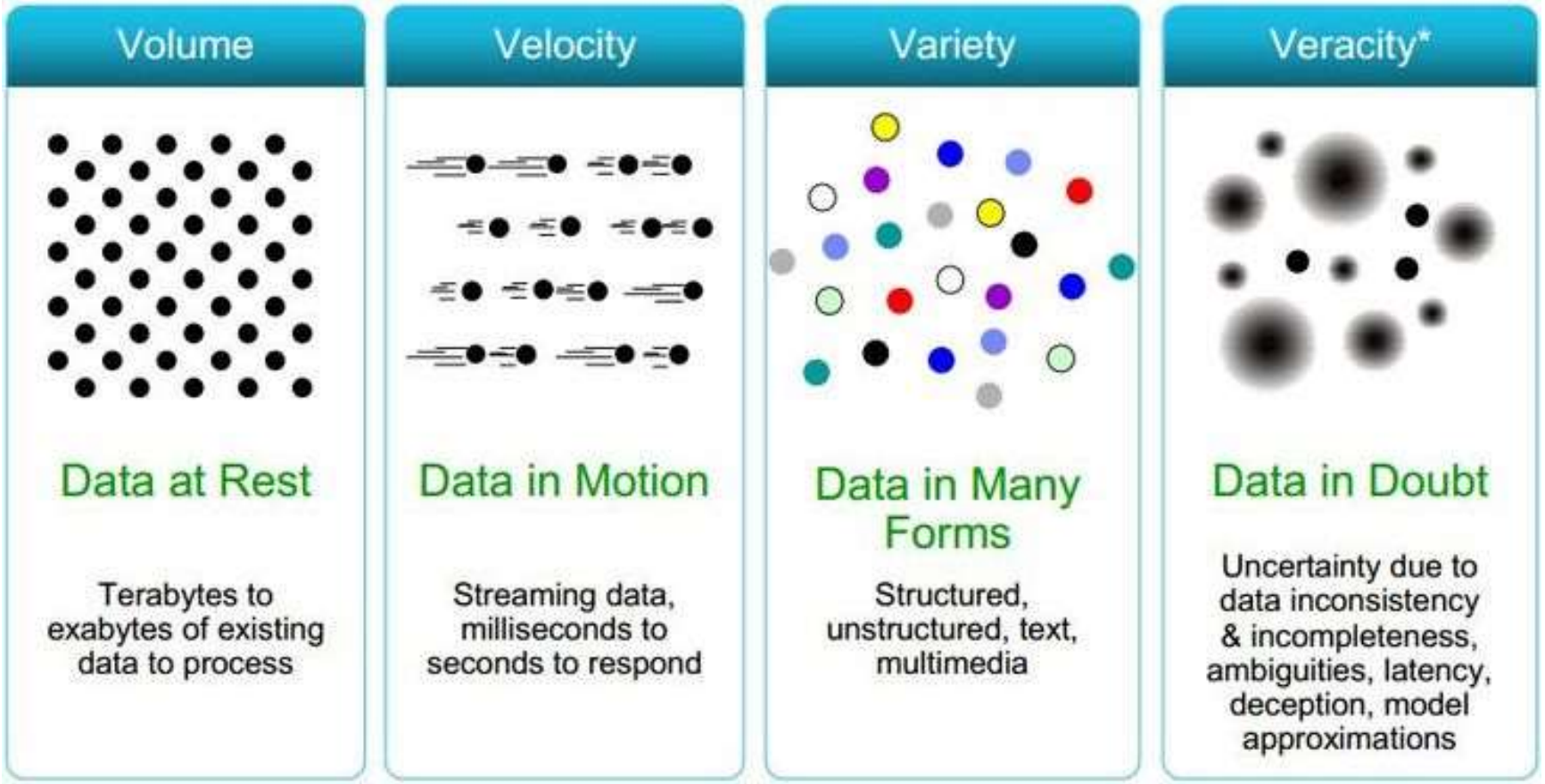
4488 views

403 shares

floq.to/LKz15

The smart city of Songdo, the world's first 'City in a Box' will be ready in 2015. It encompasses 1.500 acres of reclaimed land in South Korea and it will be a revolution in city design. Located just 40 miles from Seoul and 7 miles from Incheon International Airport. Songdo will have commercial office spaces, retail shops, residences, hotels as well as civic and cultural facilities spread out over 100 million square foot. A consortium of partners consisting of Cisco, 3M, Posco E&C and United Technology are currently developing the city of Songdo.

DATA CHARACTERISTICS



Source: IBM

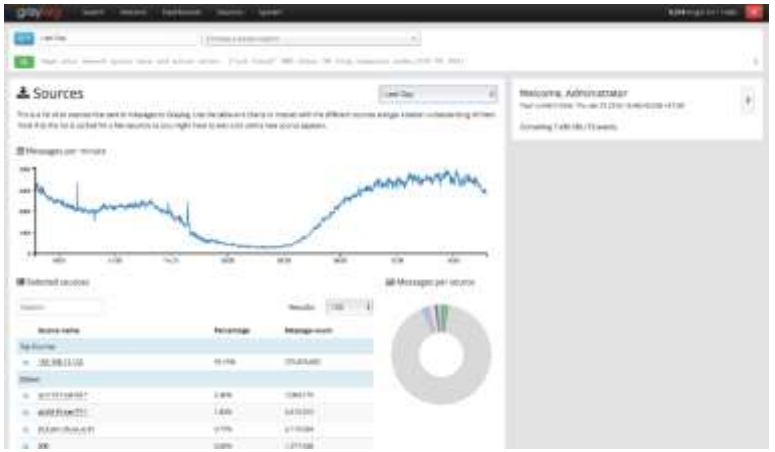
พระราชบัญญัติ
ว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์

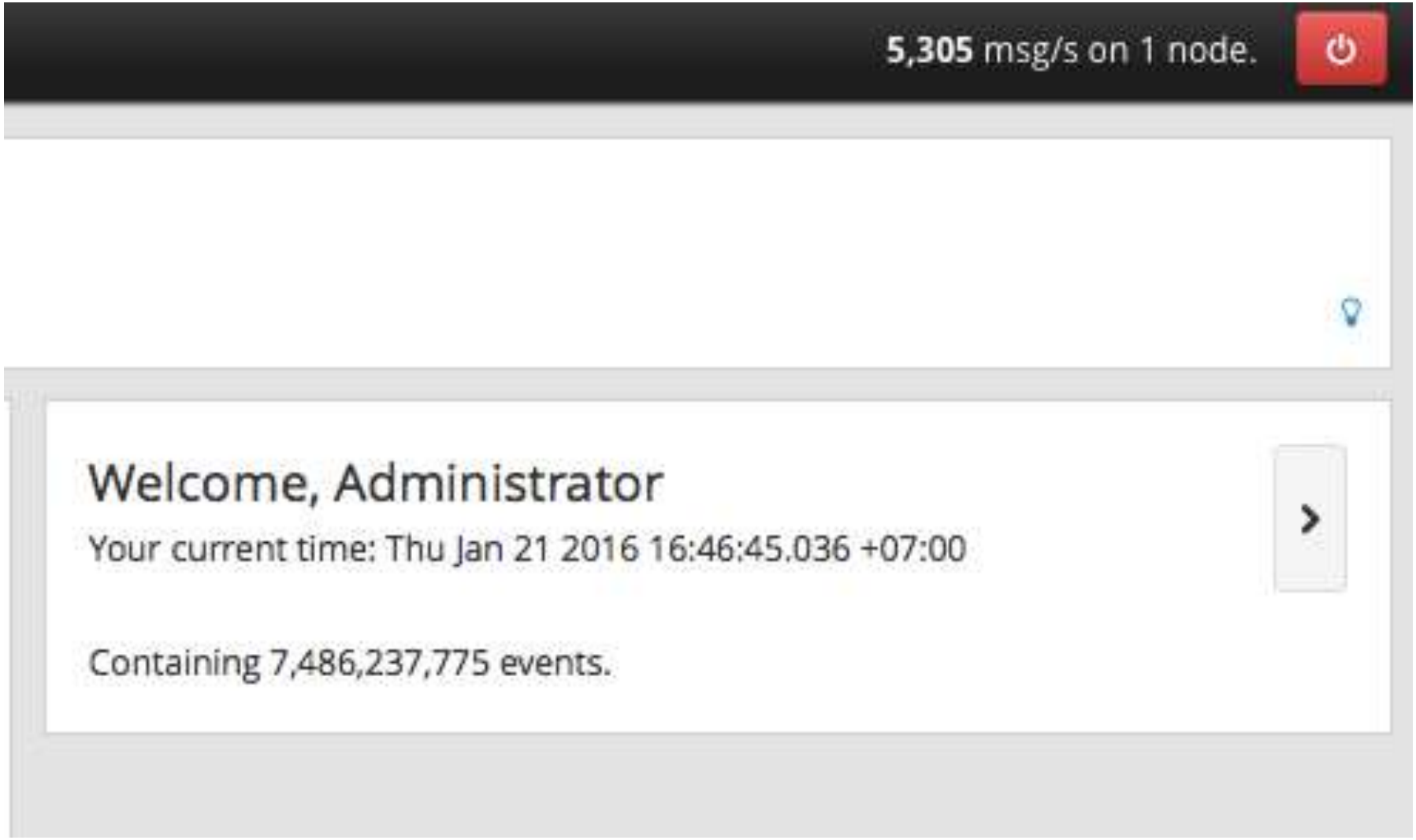
พ.ศ.๒๕๕๐

IT LOG AT CHULALONGKORN UNIVERSITY

Users 40,000+
Servers = 500+
Wifi + NAT

Manual processes

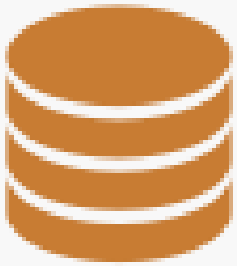


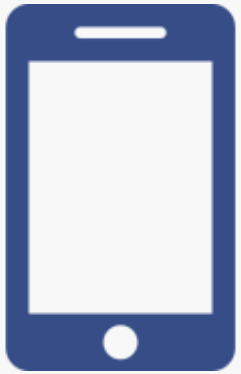


Storage Requirements 90 days = 39,000,000,000 events (6.5TB)

Internal

External





Structured

Unstructured

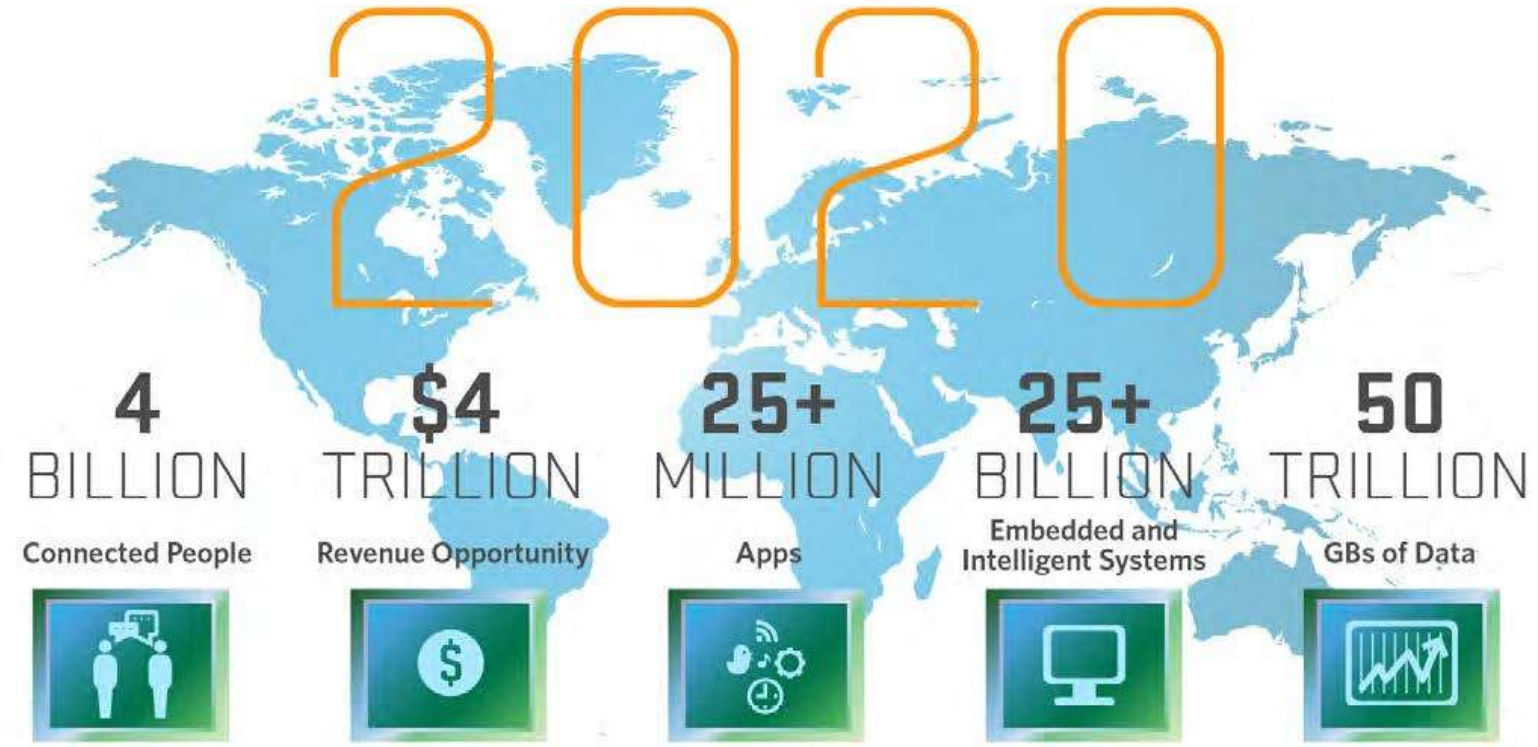
BIG DATA'S DRIVERS
MOBILE & DEVICES - COMPUTING EVERYWHERE

Thailand's rate is 147% (smartphone = 49%)

Wearable devices' shipment will be doubled in 4 years (from 72m in 2015 to 155m in 2019)

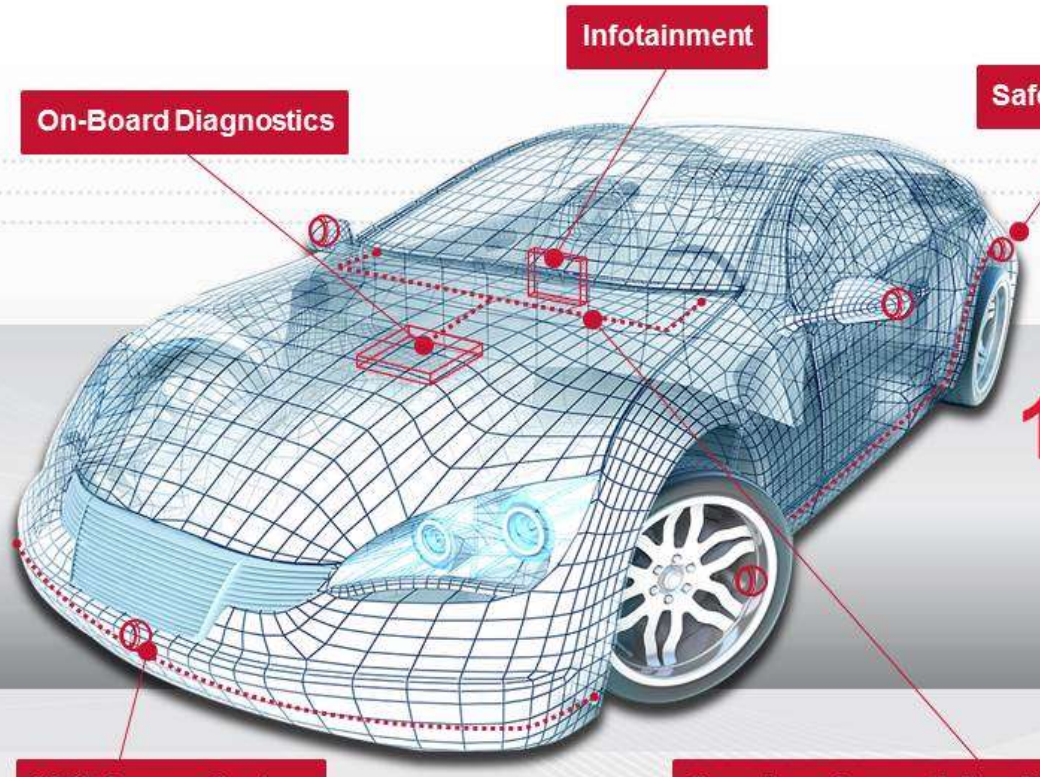
20% will be healthcare related devices

THE INTERNET OF THINGS



Source: Mario Morales, IDC

THE CONNECTED CAR

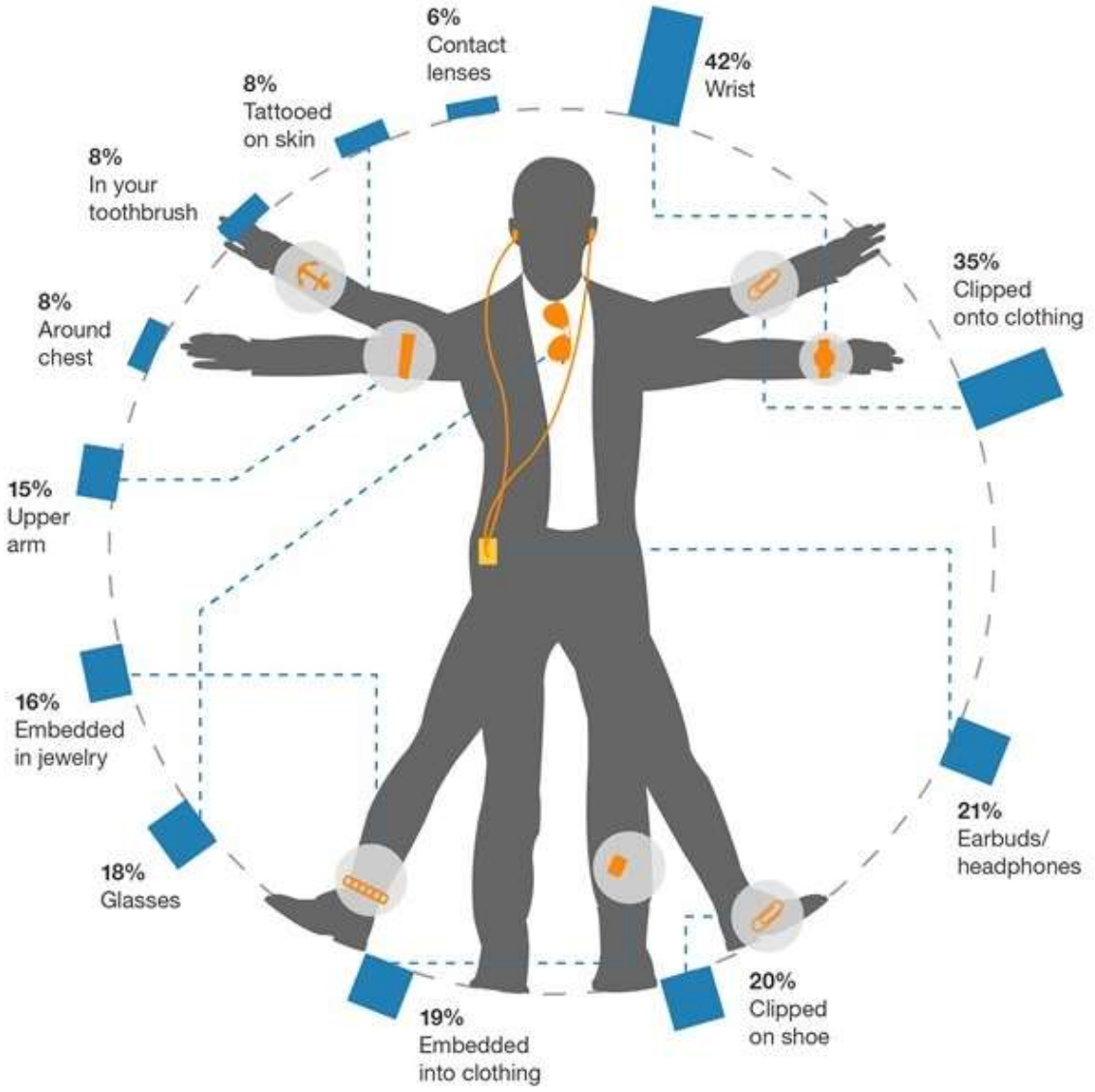


100% Of cars will be connected by 2025[1]

75% Of cars on the road will be autonomous by 2035[2]

Source: [1]GSMA 2013, [2]Navigant Research 2013

"How would you be interested in wearing/using a sensor device, assuming it was from a brand you trust, offering a service that interests you?"



Base: 4,556 US online adults (18+)
(multiple responses accepted)

Source: Forrester's North American Consumer Technographics® Consumer Technology Survey, 2014

INTRODUCING FDA-APPROVED INGESTIBLE SENSORS IN PILLS

http://www.forbes.com/sites/singularity/2012/08/09/no-more-skipping-your-medicine-fda-approves-first-digital-pill/

BIG DATA'S DRIVERS
USER GENERATED CONTENTS AND CROWDSOURCING

Blogging, reviewing commenting, forum, digital video, podcasting, mobile phone photography, social networking, crowdsourcing, etc.

Highly influential to consumer behavior and also enable the study of consumer behavior

Generate lots of both structured and unstructured data

BIG DATA'S DRIVERS
CLOUD COMPUTING

Deliver computing services over a network

Evolution of technology, but revolution of economy

One of Big Data accelerators: significant big data sources and enabling platform for big data processing

USE CASES BY SUBJECT AREAS

- Infrastructure and Information Management
- Social Listening / Customer Understanding
- Health Improvement
- Logistics and Planning
- Operation / Product Improvement

INFRASTRUCTURE AND INFORMATION MANAGEMENT

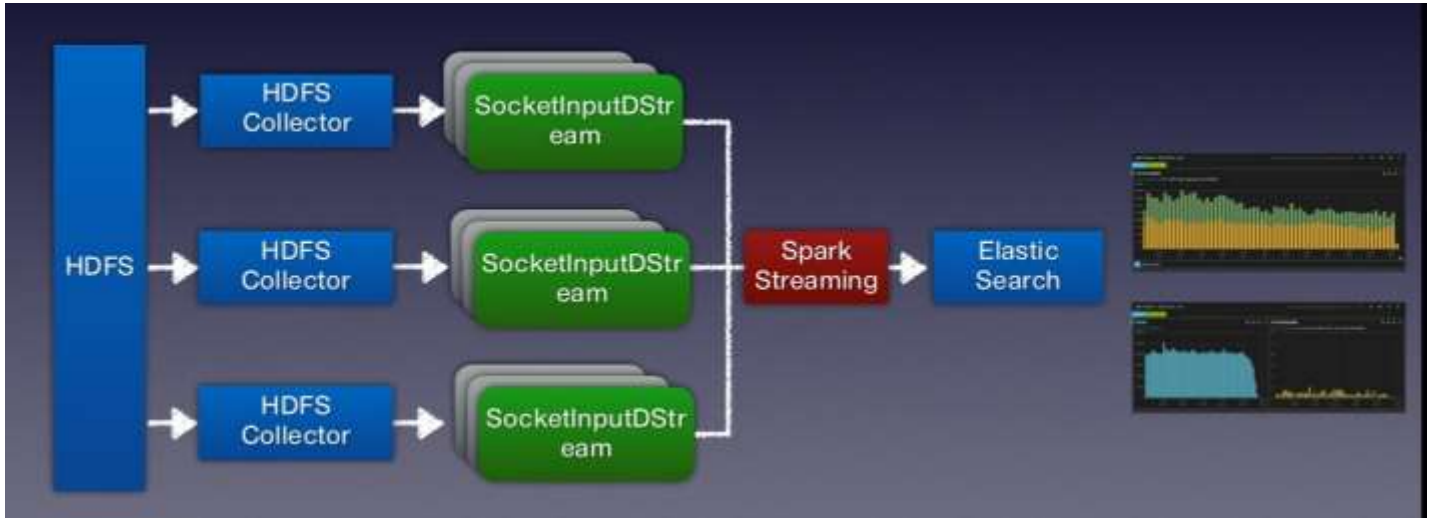
- Bigger and Faster Data Warehouse
- Information Archival and Management

CASE STUDY:
SK TELECOM'S USAGE PATTERN ANALYSIS

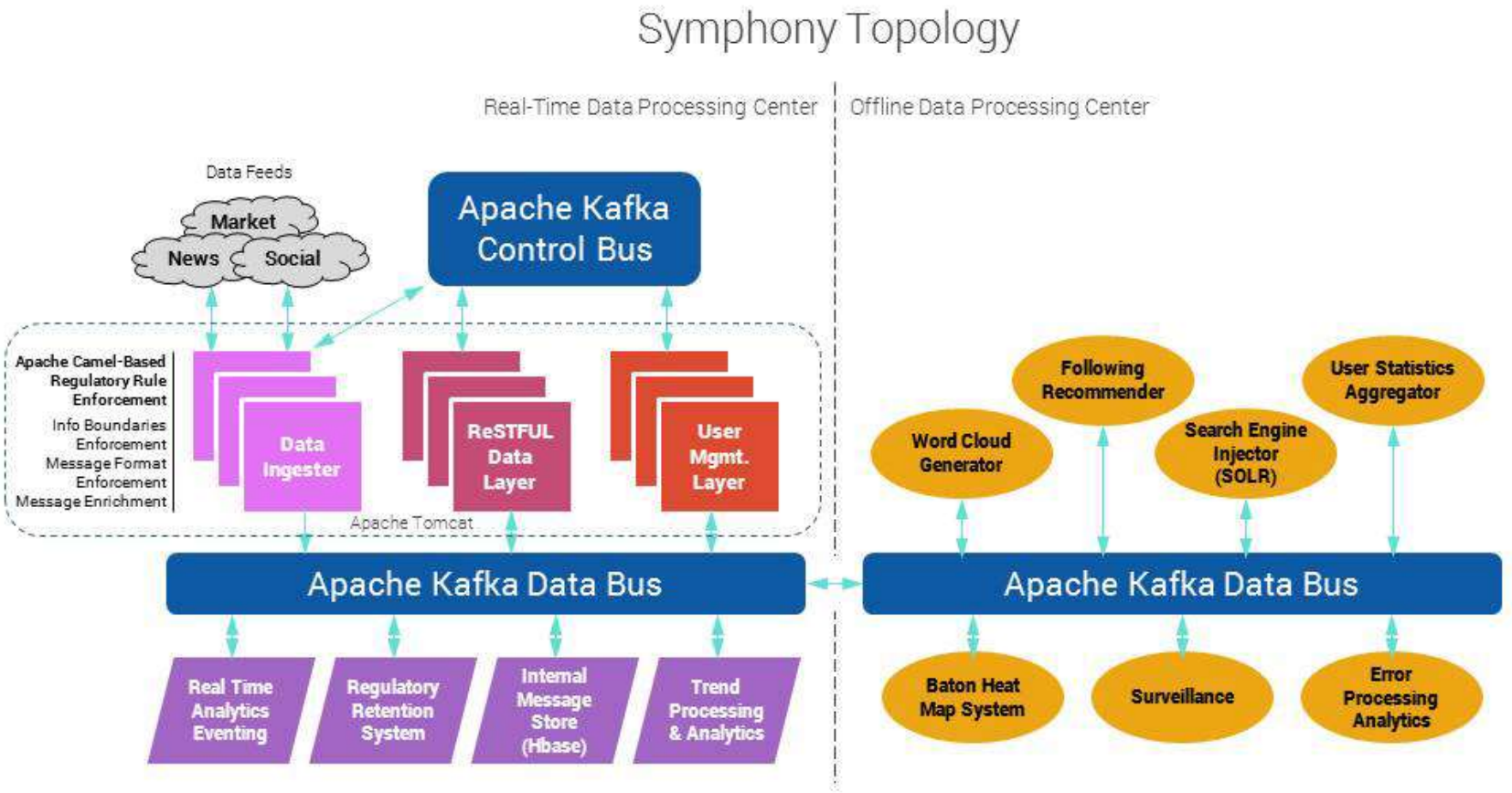
Process usage data from 28 millions subscribers: 40TB/day – 15PB total

Must process data with 530MB/sec or 1 million records/sec

Use Hadoop, Spark, and ElasticSearch to provide mobile usage pattern analytics with low latency ad-hoc query (< 2 secs)



GOLDMAN SACHS – EFFECTIVE MESSAGING PLATFORM



http://www.goldmansachs.com/what-we-do/engineering/see-our-work/inside-symphony.html

SOCIAL LISTENING / CUSTOMER UNDERSTANDING

- Sentimental Analysis / Social Network Trends
- Customer 720
- Customer Segmentation
- Customer Retention
- Targeted Marketing / Personalization Offering
- Click-Stream Analysis
- In-store Tracking

Thailand Facebook place ranking

Add Place

Facebook place ranking | Swarm ranking

Rank		Place	Total Checkin ▼	Yesterday Checkin ▼	Like ▼	Talking About This ▼
		ท่าอากาศยานสุวรรณภูมิ | Suvarnabhumi Airport BKK	3,294,222	6,780	148,009	44,005
2.		CentralWorld | เซ็นทรัลเวิลด์	2,431,384	2,431,384	102,562	38,864
3.		Future Park Rangsit	947,015	32	19,977	762
4.		Central Plaza Pinklao | เซ็นทรัลปิ่นเกล้า	788,508	673	28,463	3,609
5.		MBK Center	747,205	1,333	66,567	5,970
6.		เยาวราช	655,299	2,673	15,850	10,438

Top brand ranking in Thailand

67,868 Facebook place

Last update at 2016/03/20 20:00

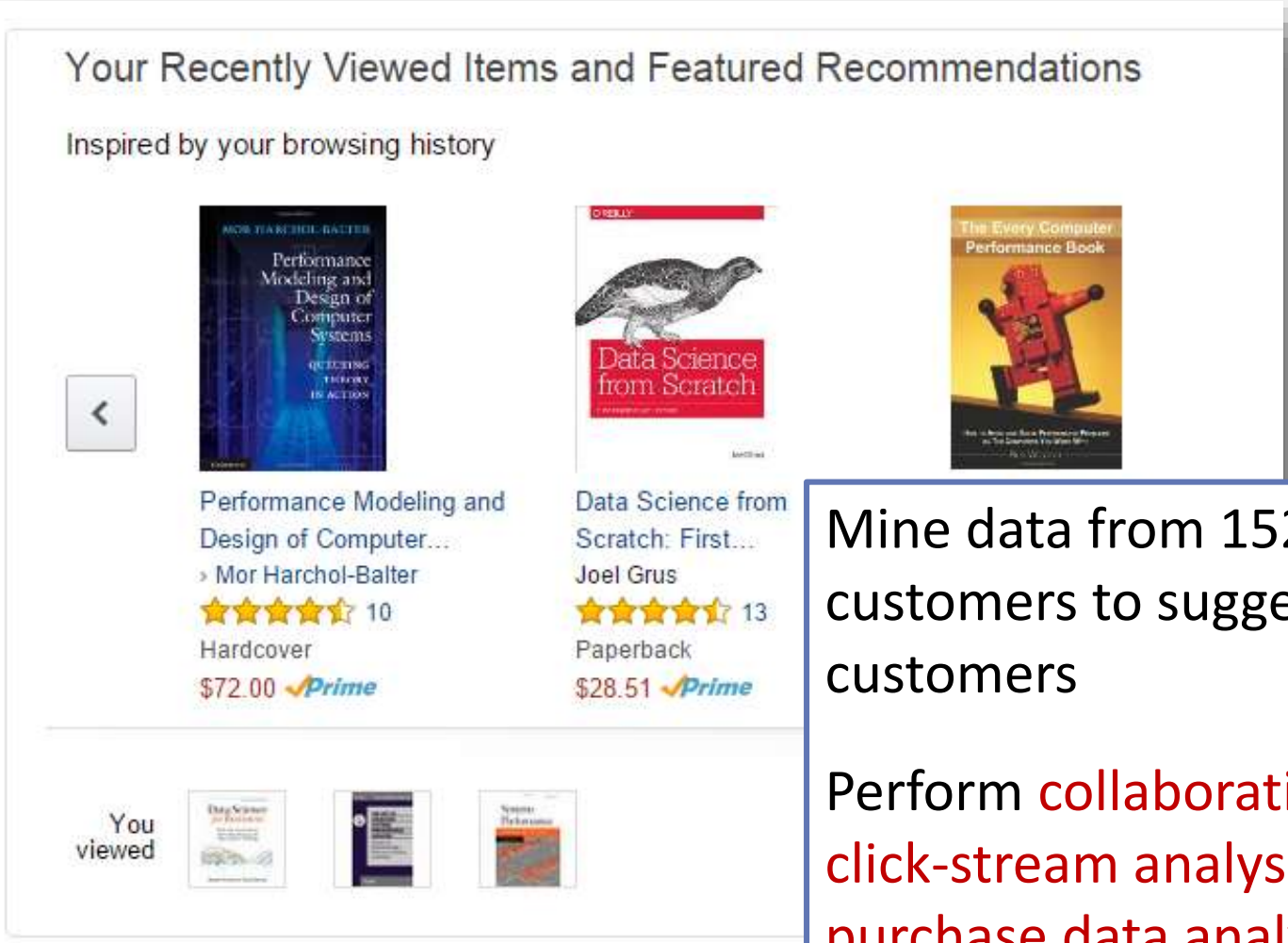
CASE STUDY: JETBLUE SENTIMENT ANALYSIS

JetBlue gets 45,000 customer feedbacks per months

Read as many as possible – 300 feedbacks per day per analyst

Utilize text-mining to analyze customer sentiment + combine with aircraft and seat numbers to fix direct problems

CASE STUDY:
AMAZON'S RECOMMENDATION ENGINE



Mine data from 152 million customers to suggest products to customers

Perform collaborative filtering, click-stream analysis, historical purchase data analytics

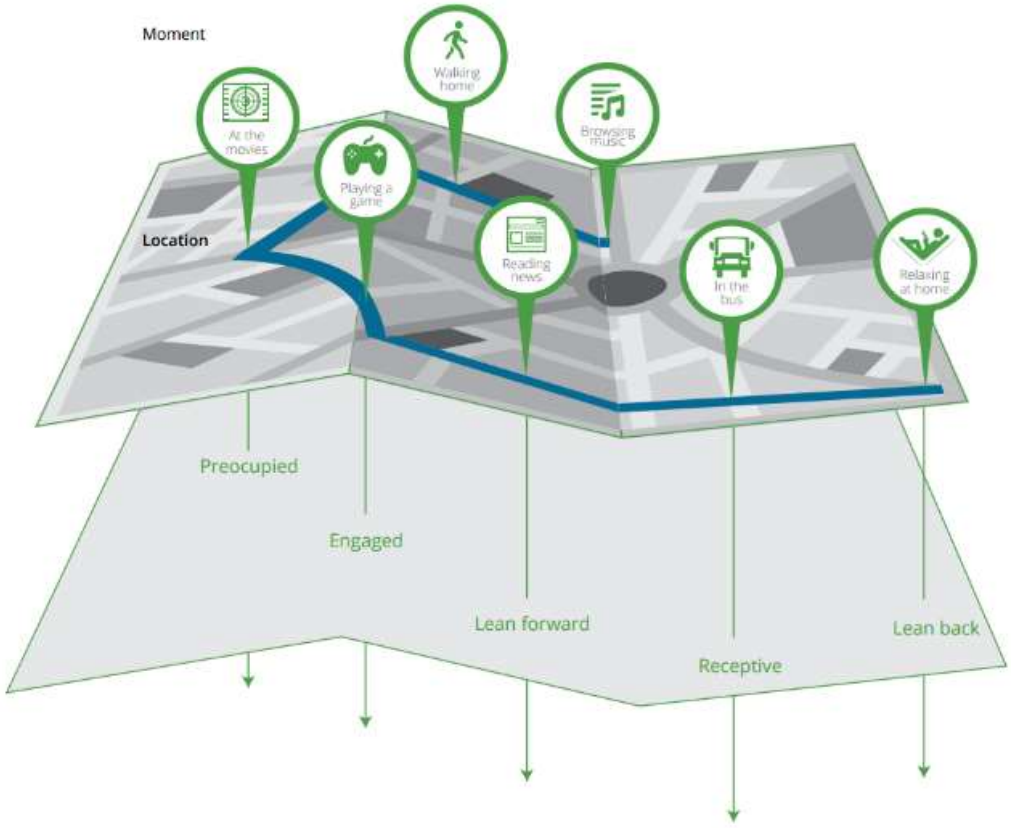
CASE STUDY:
UBER'S DYNAMIC PRICING FARES

Uber's entire business model is based on the very Big Data principle of crowd sourcing

"dynamic pricing" fares are calculated automatically, using GPS, street data, demand forecast, and predictive algorithms

Due to traffic conditions in New York on New Year's Eve 2011, the fare of journey of one mile rose from $27 to $135

CASE STUDY:
INMOBI'S TARGETED MARKETING



User behaviour changes dramatically across work, home, commute, and other location contexts

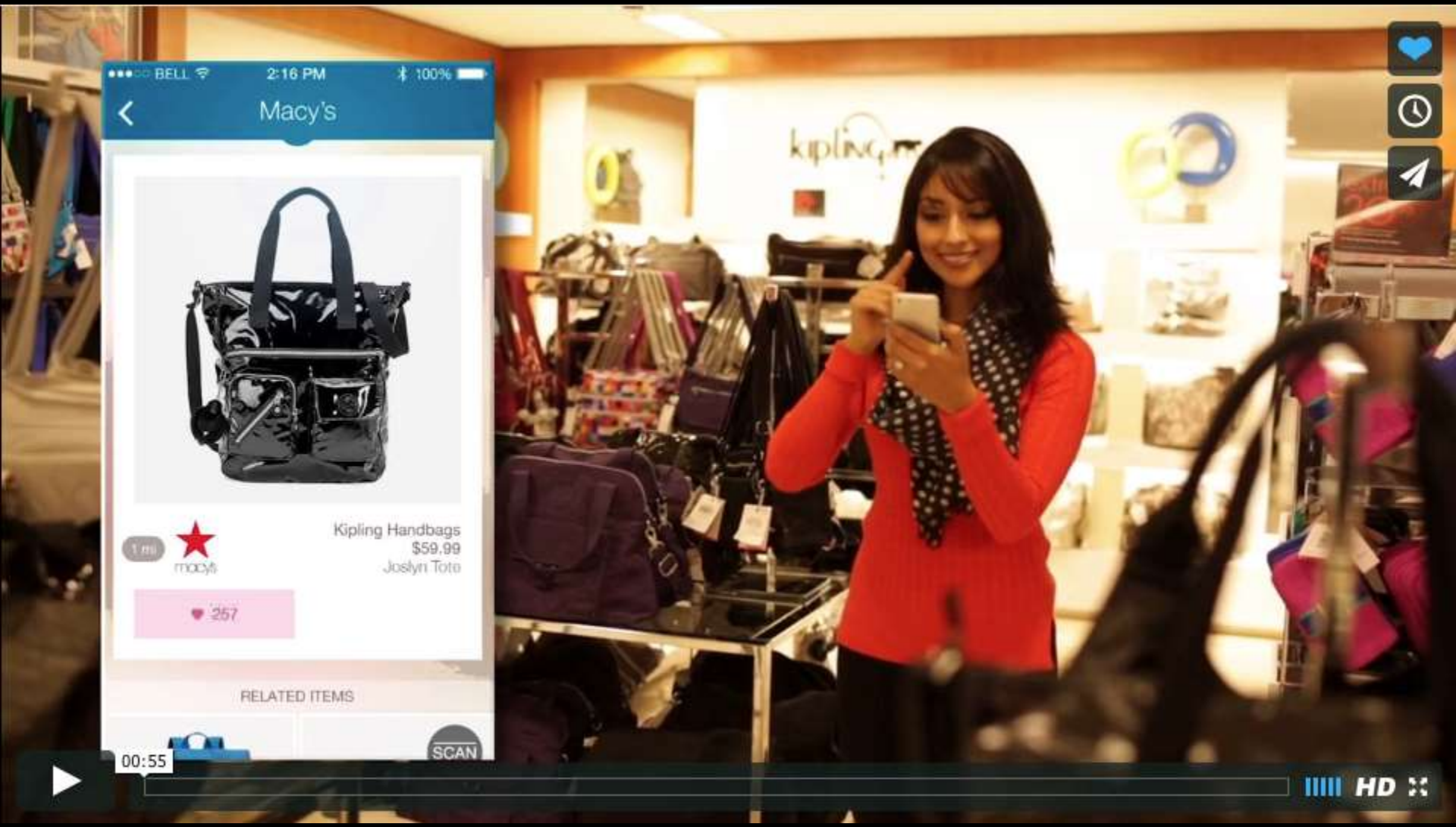
Geo context targeting: create customer micro segmentation from customer's location activities, time of day, and app being used

CASE STUDY: MARCY'S

Mid-range to upscale department store chain

Goal is to offer more localized, personalized and smarter customer experience across all channels

Deploy 4,000 sensors inside 768 stores to identify customers' in-store locations


BELL
2:16 PM
100%
Macy's
Kipling Handbags
$59.99
Joslyn Toto
257
RELATED ITEMS
SCAN
00:55
HD

HEALTH IMPROVEMENT

- eHR / Care Coordination Record / Patient 360
- Text Analytics for Medical Classification
- Machine Learning for Diagnosis and Screening
- Genome Analytics / Precision Medicine
- Risk Prediction for Patient Care / Urgent Care Management
- After-discharge monitoring
- Population Health Management / Preventive Healthcare

Prof. Michael Snyder
Stanford University School of Medicine

- Genome indicates high risk for Type-2 diabetes
- Perform extensive blood tests every two months
- Into the 14-month study, analyses showed he developed diabetes
- The illness was treated successfully while in its early stages

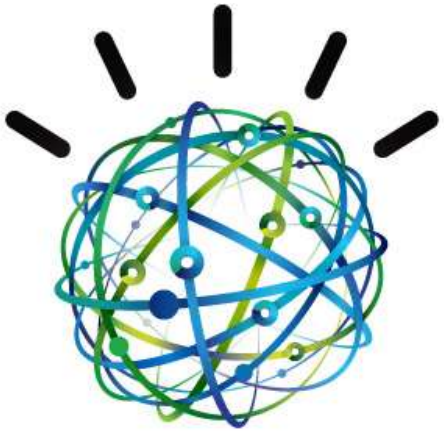
Dexcom
THE GLUCOSE SENSOR COMPANY

Owlet

Rest Assured.

UNDER ARMOUR

IBM

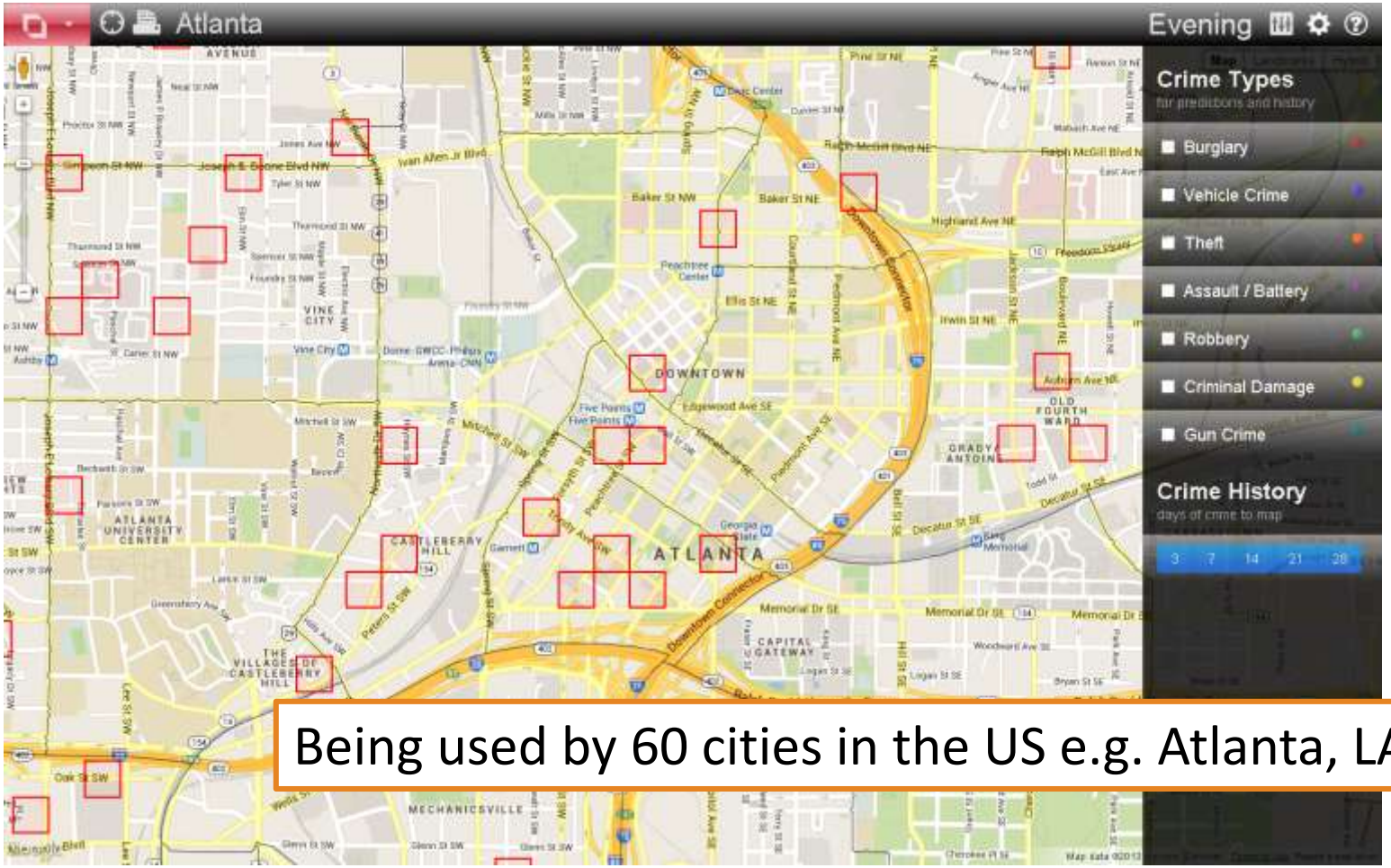


Behavioral trend tracking – customize fitness program setup
Food intake tracking - visual recognize food intake
Environment factor tracking – modify fitness program recommendation

LOGISTICS AND PLANNING

- Route Optimization
- Location Planning
- Crowdsourcing
- Remote-Sensing-Aided Marketing Research
- Urban Planning

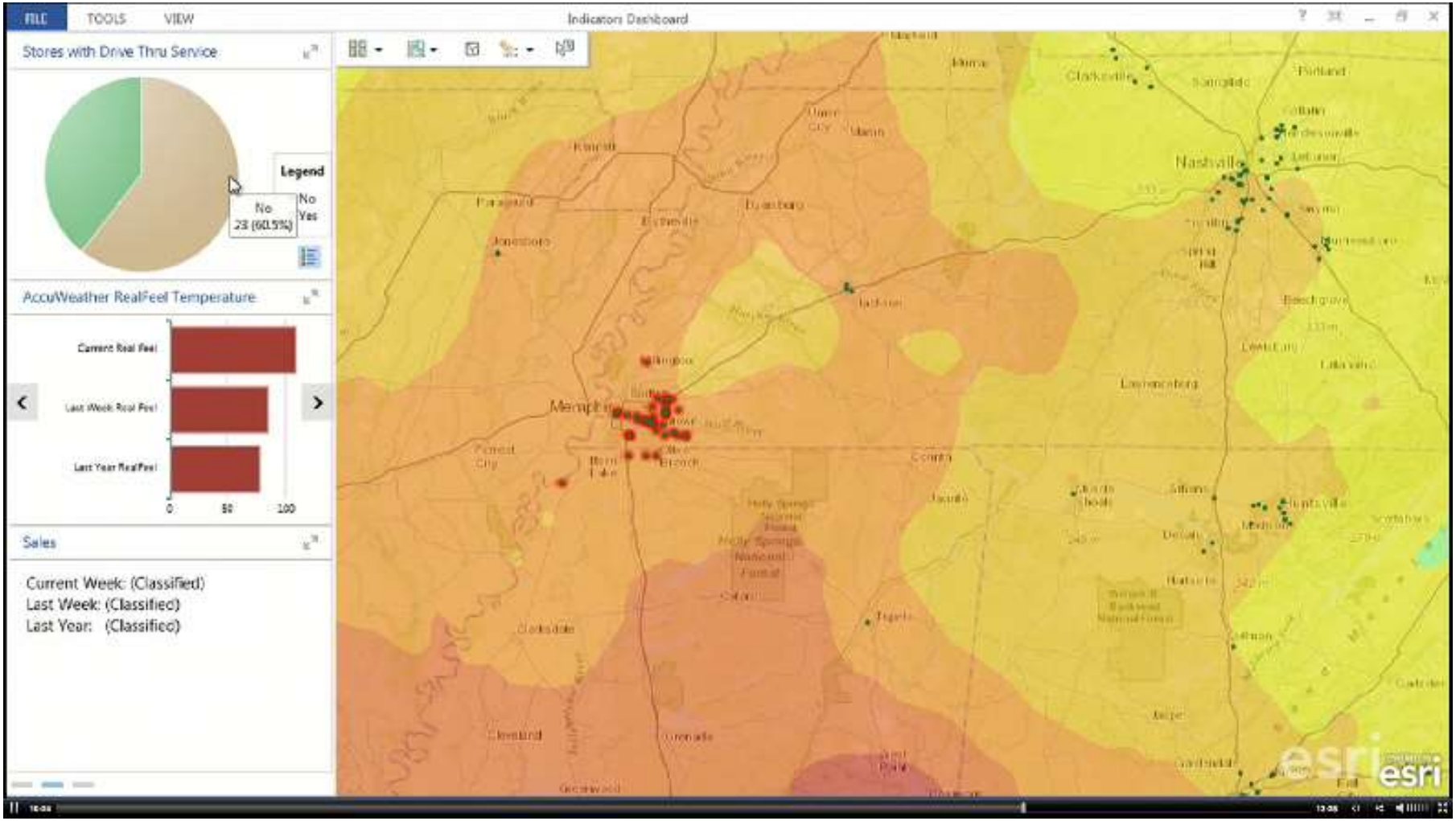
CASE STUDY: PREDICTIVE POLICING



Being used by 60 cities in the US e.g. Atlanta, LA, etc.

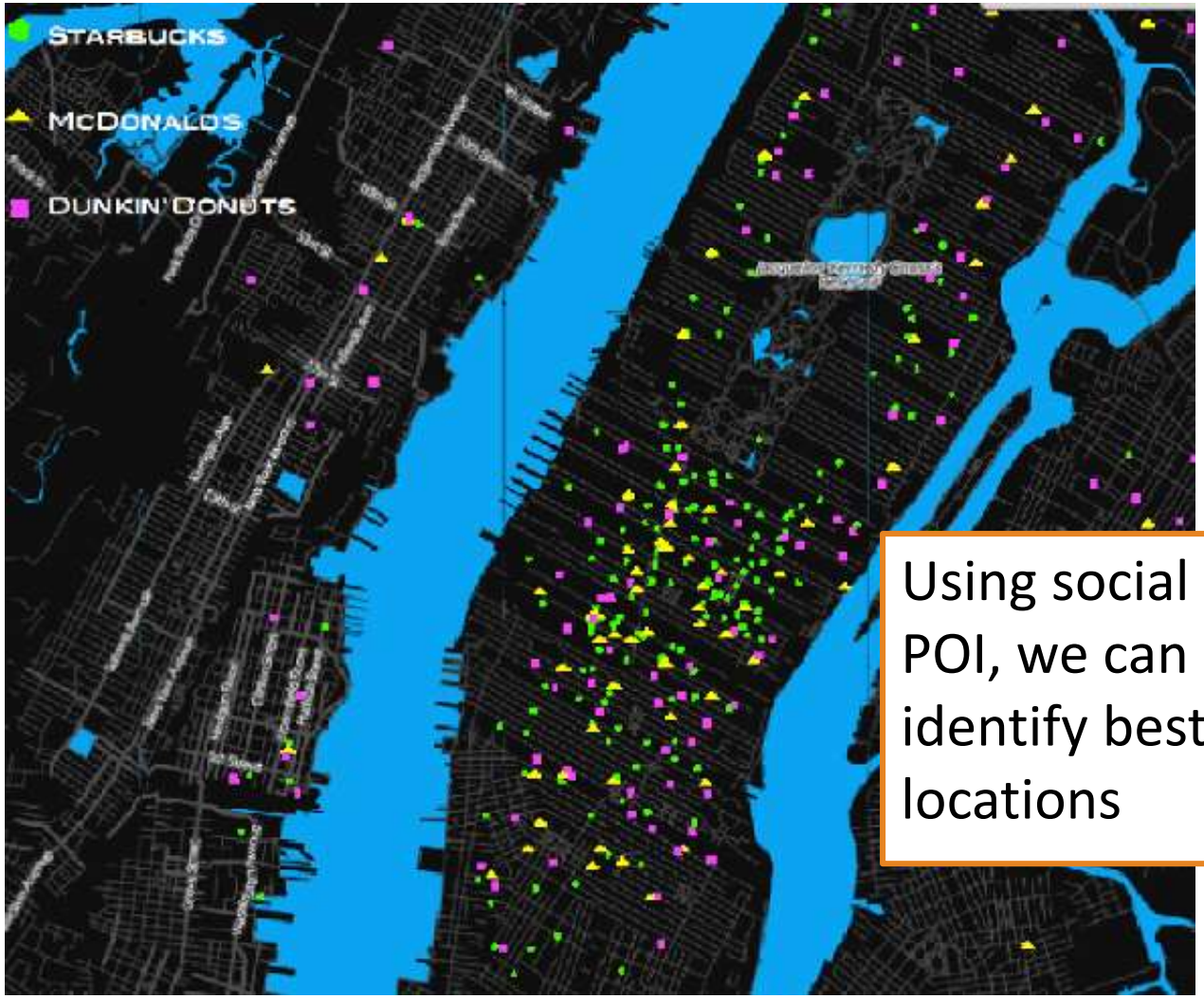
Source: http://www.forbes.com/sites/ellenhuet/2015/02/11/predpol-predictive-policing

CASE STUDY: STARBUCKS OPERATION PLANNING



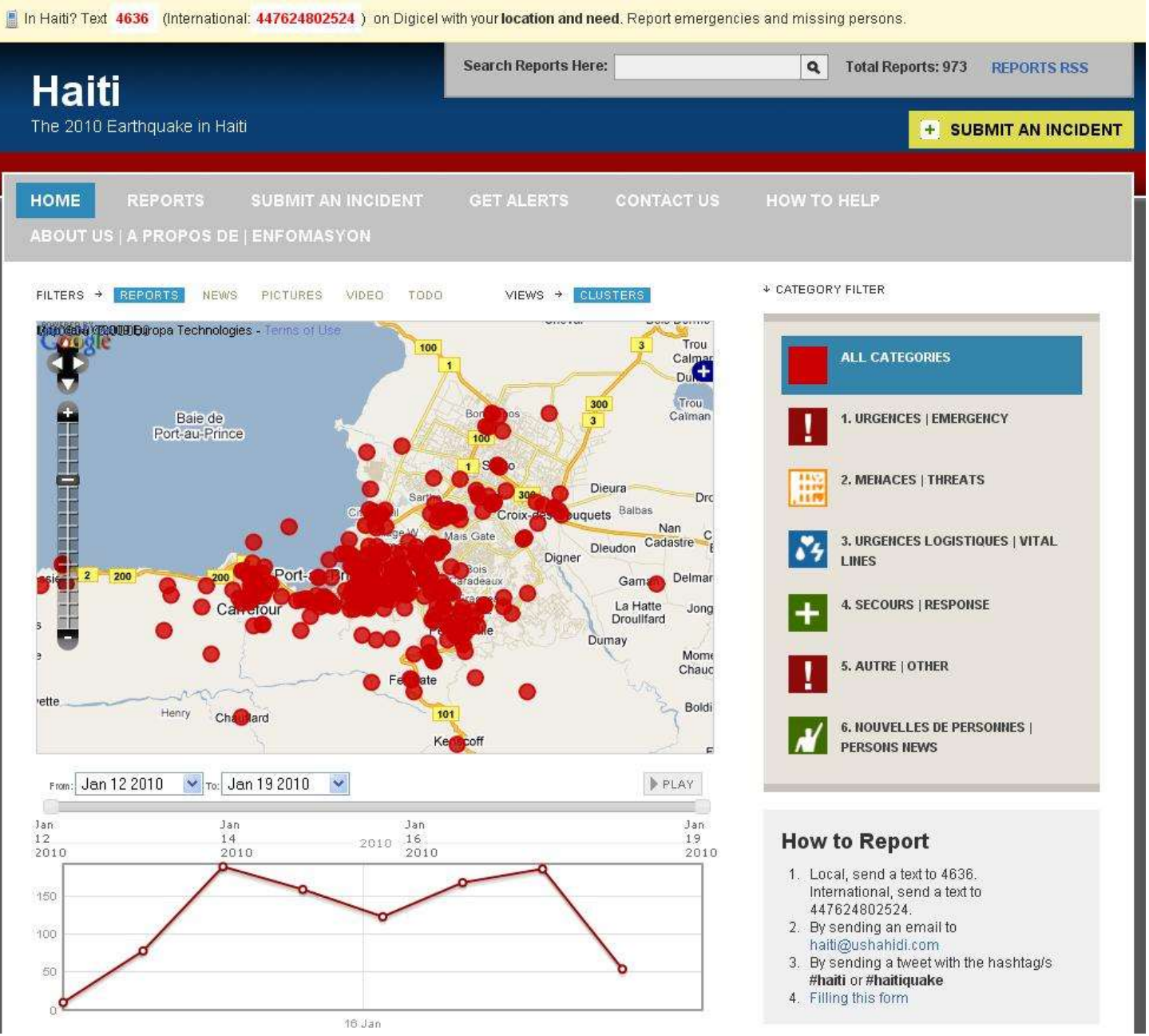
http://www.fastcompany.com/3034792/how-fast-food-chains-pick-their-next-location

CASE STUDY: FASTFOOD STORE PLANNING



Using social network and POI, we can effectively identify best store locations

http://www.fastcompany.com/3008621/tracking/github-reveals-a-formula-for-your-hacker-persona



USHAHIDI

2007
Kenya

2010
Haiti
Chile
Washington DC
Russia

2011
Christchurch
Middle East
India
Japan
Australia
US
Macedonia

2012
Balkans

2014
Kenya

CASE STUDY: NIELSEN - GEO ANALYTICS AND MARKETING RESEARCH

Stratified sampling divides members of the population into homogeneous subgroups to improve effectiveness

Indonesia is a large country which can be expensive for sampling

Use crowdsourcing + satellite imagery + K-Mean to better measure urbanization and lead to optimal allocation of interviewers to respondents

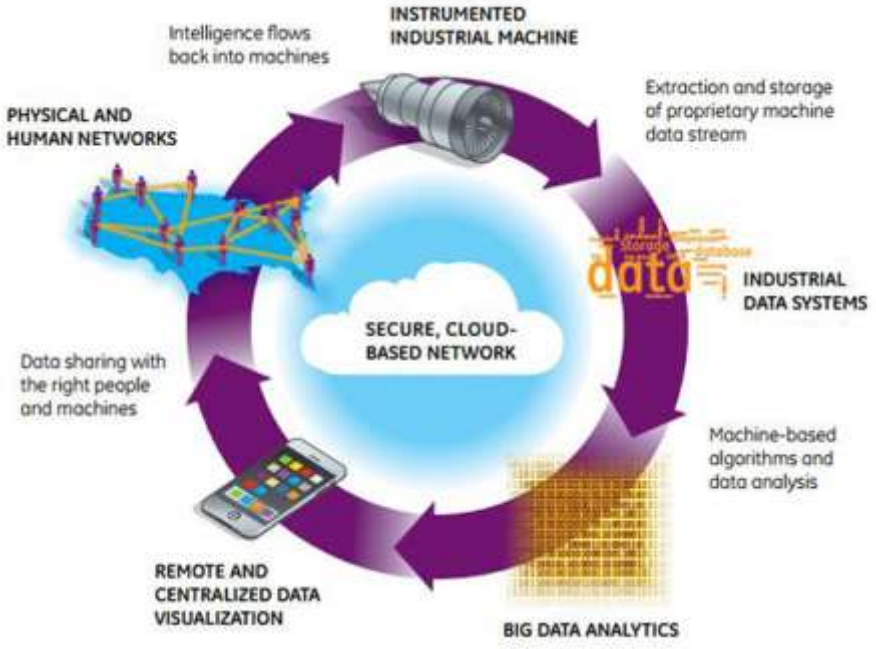

5.9 million
transactions per day
commuters board and alight trains and buses on a daily basis


MATSim
senozon

OPERATION / PRODUCT IMPROVEMENT

- New Products / New Services
- Risk Management / Fraud Detection
- Predictive Maintenance

CASE STUDY:
GE'S SMART MACHINES



GE has launched Industrial Internet initiative

Jet engine has 20 sensors generating 5,000 data samples per second

Data can be used for fuel efficiency and service improvements

"In the future it's going to be digital. By the time the plane lands, we'll know exactly what the plane needs."

CASE STUDY:
JP MORGAN CHASE

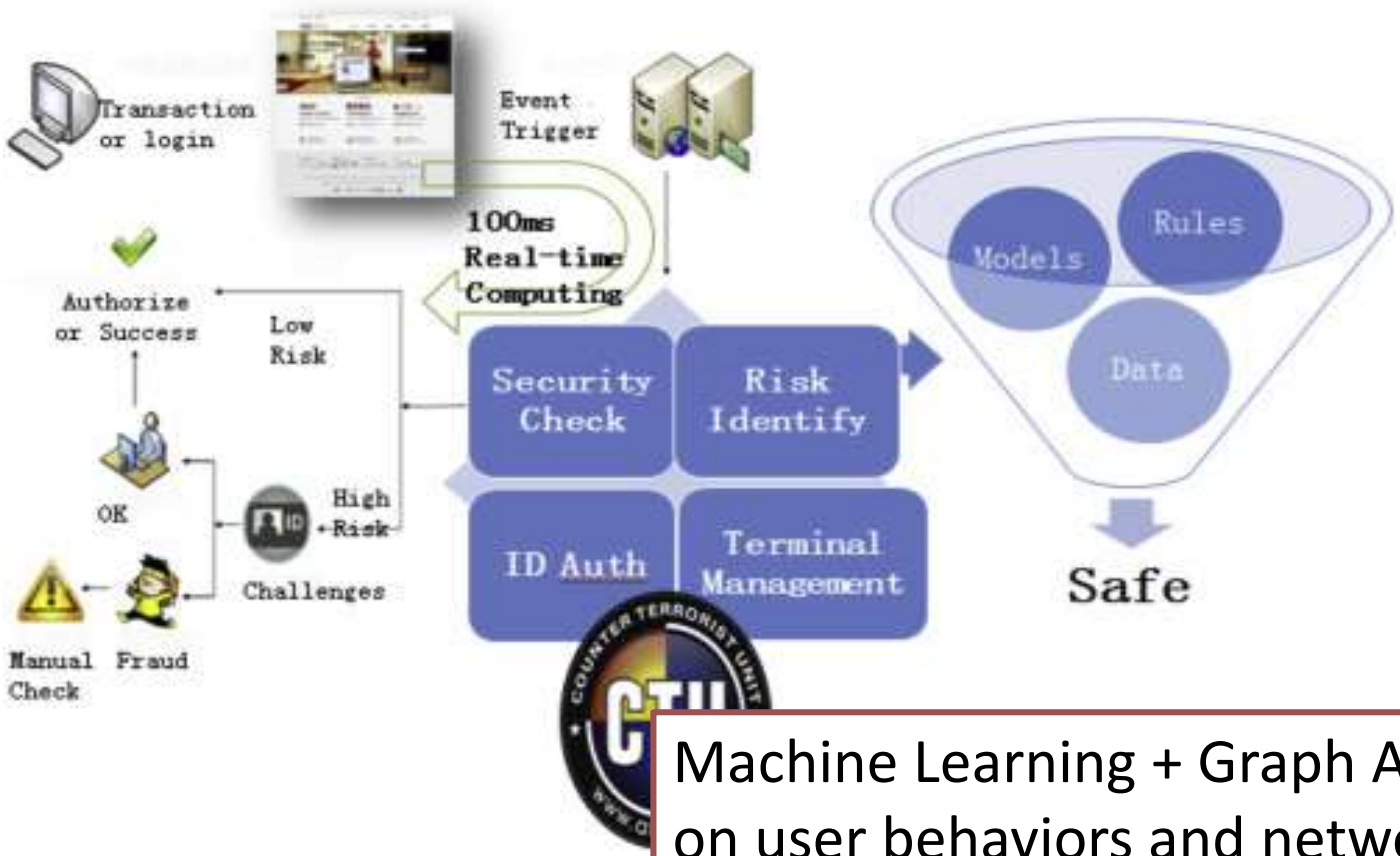
JP Morgan Chase & Co use Big Data to aggregate all available information about a single customer

Data included monthly balances, credit card transactions, credit bureau data, demographic data

This allowed bank to offer lower interest rates by reducing credit card fraud

Aggregating data of 30 million customers, they provide US economic outlooks with “Weathering Volatility: Big Data on the Financial Ups and Downs of U.S. Individuals”

CASE STUDY: ALIBABA FRAUD DETECTION



Machine Learning + Graph Analytics on user behaviors and network

Source: http://www.sciencedirect.com/science/article/pii/S2405918815000021

How EDUCATIONAL DATA MINING & LEARNING ANALYTICS can help:

Educational data mining focuses on developing new tools and algorithms for discovering data patterns

EDUCATIONAL DATA MINING CAN ANSWER QUESTIONS LIKE:

- What sequence of topics is most effective for a specific student?
- Which student actions are associated with better learning and higher grades?
- Which actions indicate satisfaction and engagement?
- What features of an online learning environment lead to better learning?

Learning analytics focuses on applying tools and techniques at larger scales in instructional systems

LEARNING ANALYTICS CAN ANSWER QUESTIONS LIKE:

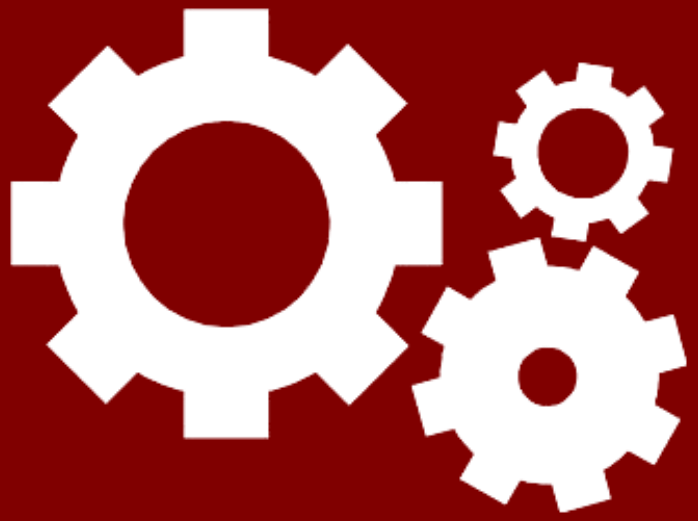
- When are students ready to move on to the next topic?
- When is a student at risk for not completing a course?
- What grade is a student likely to receive without intervention?
- Should a student be referred to a counselor for help?

Source: collegestats.org

CASE STUDY: THYSSENKRUPP ELEVATOR

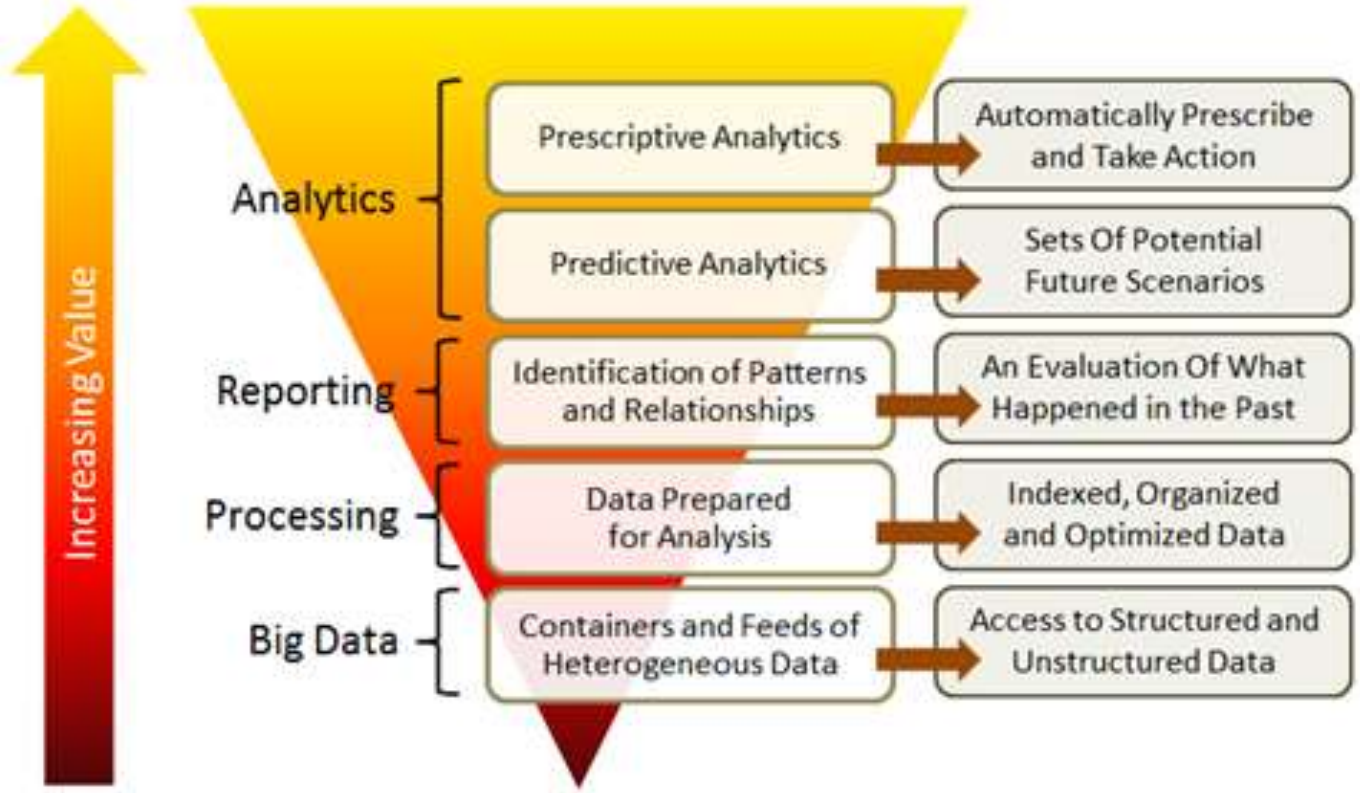
- Continuously monitor equipment condition from motor temp to shaft alignment, cab speed and door functioning using thousands of sensors
- Use predictive analytics to schedule planned downtime
- Reduced downtime
- Improved cost forecasting, resource planning and maintenance scheduling

Data Engineering (Big Data)



Data Science (Data Analytics)

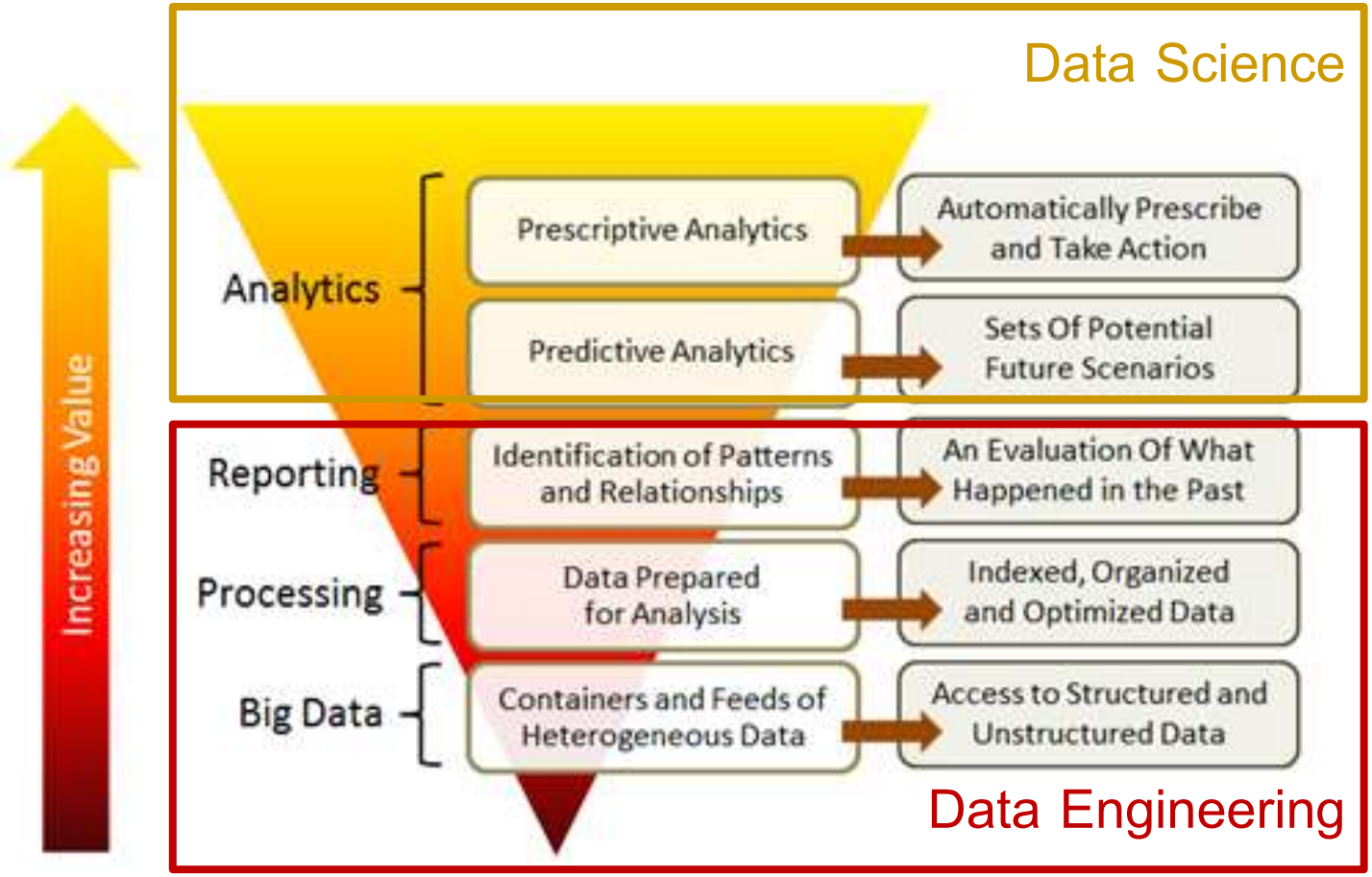
DATA VALUE CHAIN



Source: http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

DATA VALUE CHAIN

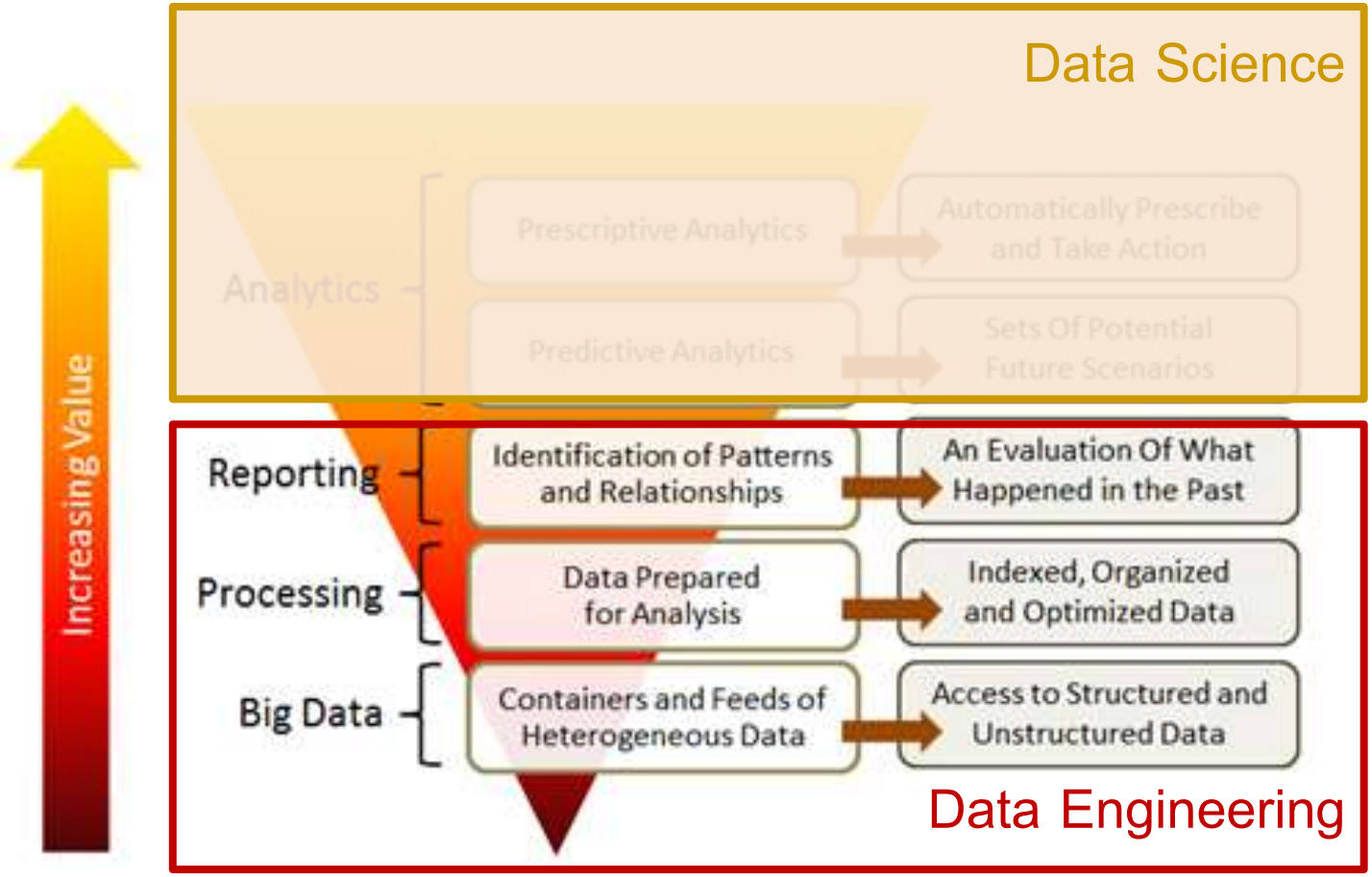
มองโจทย์เป็นตัวตั้ง



มองข้อมูลเป็นตัวตั้ง

Source: http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

DATA VALUE CHAIN กับ IT LOG



Source: http://steinvox.com/blog/big-data-and-analytics-the-analytics-value-chain/

การวิเคราะห์ข้อมูลติดตามรถขนส่ง

ข้อมูลการทำงานของเครื่องยนต์ (ความเร็ว วงเลี้ยว ฯลฯ)

ข้อมูลตำแหน่ง GPS ของรถ

ข้อมูล VDO Streaming จากกล้องที่ติดด้านหน้า/หลังของรถ

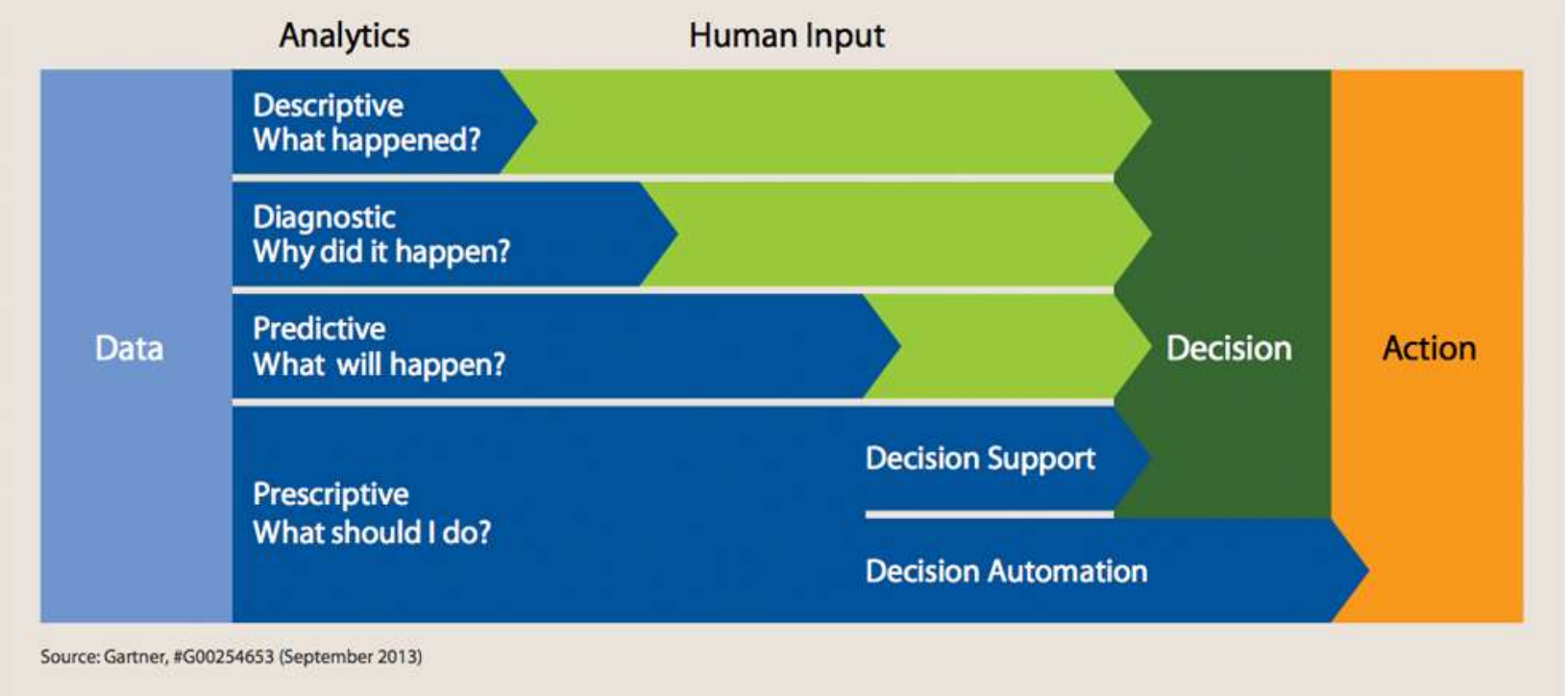
คำถาม:

คนขับรถ มีพฤติกรรมการขับที่ปลอดภัยหรือไม่?

มีปัจจัยสภาพอากาศมาเกี่ยวข้อง?

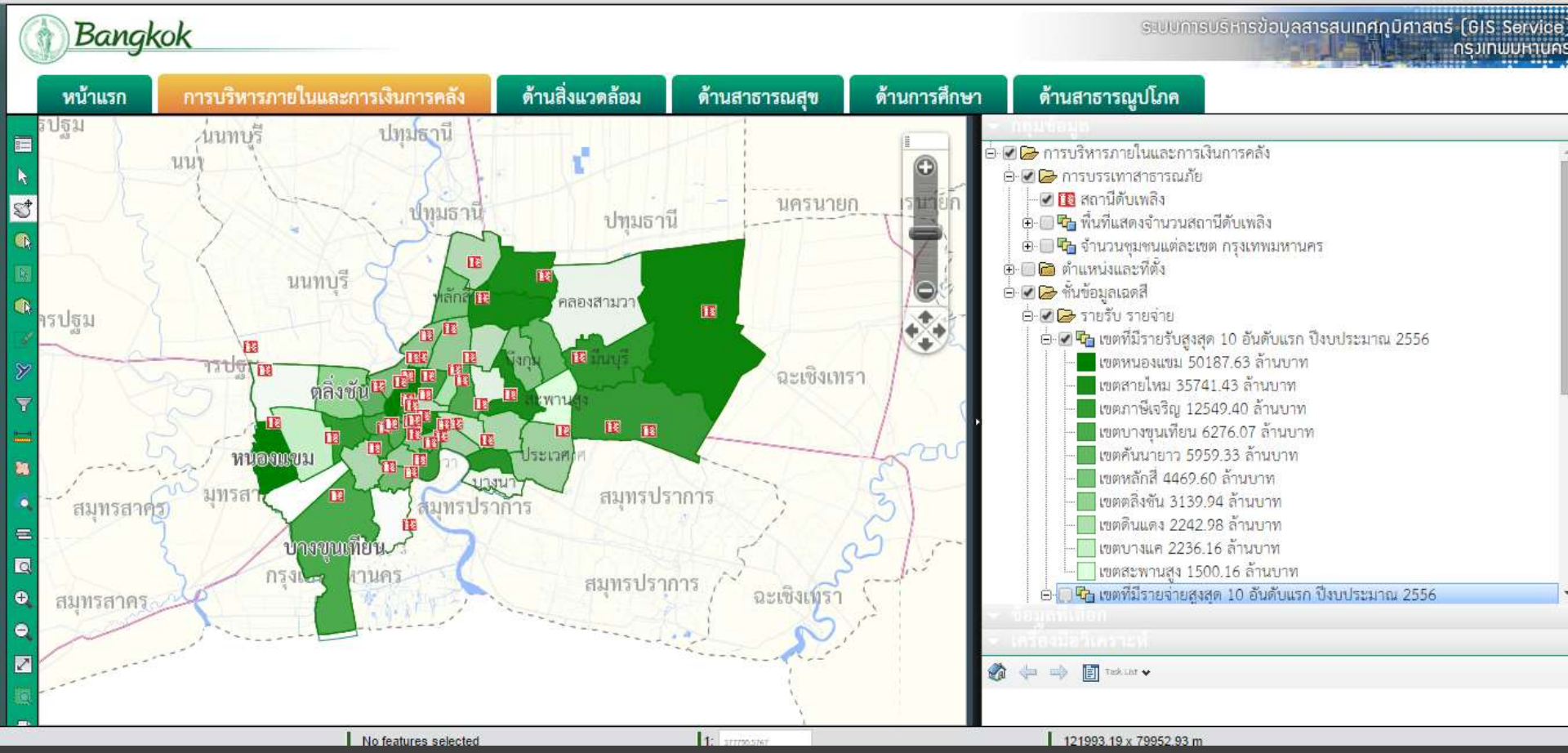
ถ้าต้องรองรับรถจำนวนหลายพันคันจะต้องทำอย่างไร?

TYPES OF DATA ANALYTICS



DATA ANALYTICS SIMPLIFIED

Descriptive	• “A.Natawut drinks about 1 cup of coffee a day”
Diagnostic	• “Number of cups that A.Natawut drinks depend on number of meetings he has each day”
Predictive	• “Tomorrow, A.Natawut has 2 meetings, it is very likely that A.Natawut will drink 2 cups tomorrow”
Prescriptive	• “Inform secretary to prepare 1 cup in the morning and one in the afternoon for A.Natawut”



Descriptive = รายงานมูลค่าที่จัดเก็บได้

Diagnostic = วิเคราะห์เหตุผลว่ามาจากแหล่งใด

Predictive = ทำนายอนาคตว่าจะได้เท่าไหร่ (ที่แม่นยำขึ้น)

Prescriptive = แนะนำว่าจะต้องเตรียมการอย่างไร

แนวทางการใช้งาน BIG DATA กับงานราชการ

Bigger / Faster / More Up-to-Date Data Warehouse

Social Listening / Crowdsourcing

Workforce Planning / Economics Planning

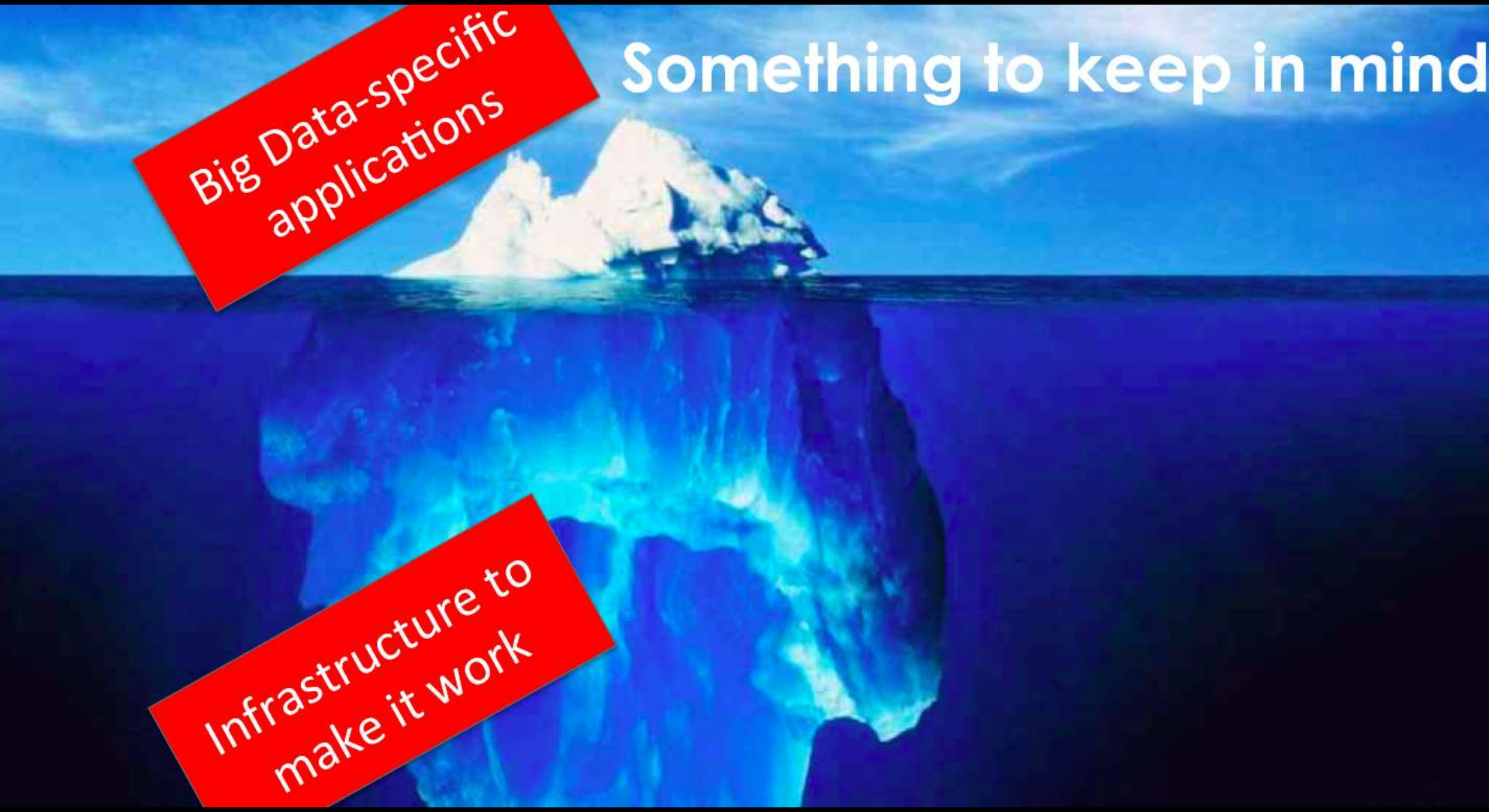
Smart Education

Precision Agricultural / Resource Management

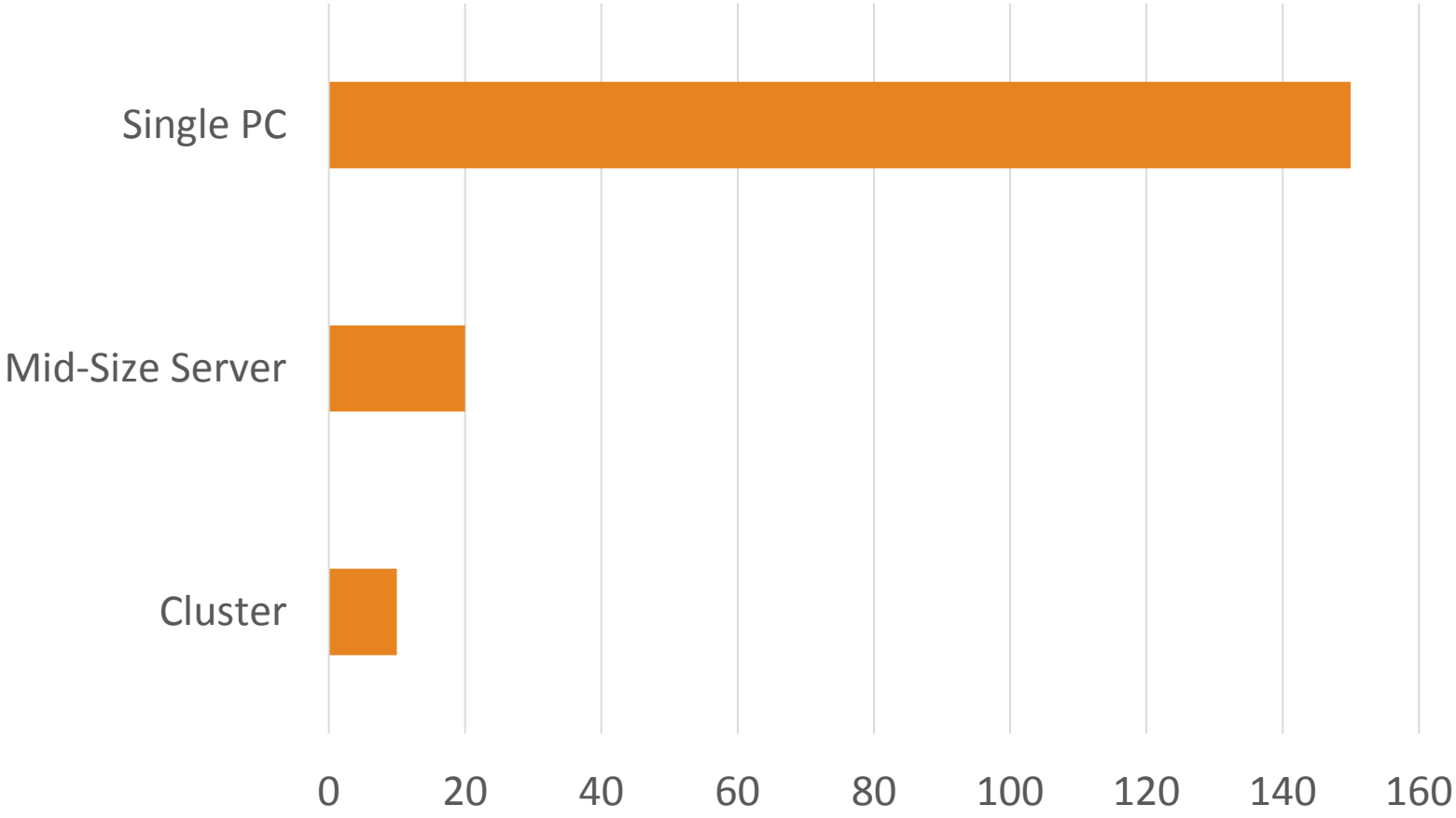
Preventive Healthcare

Fraud Detection (e.g. Tax, Social Security, etc.)

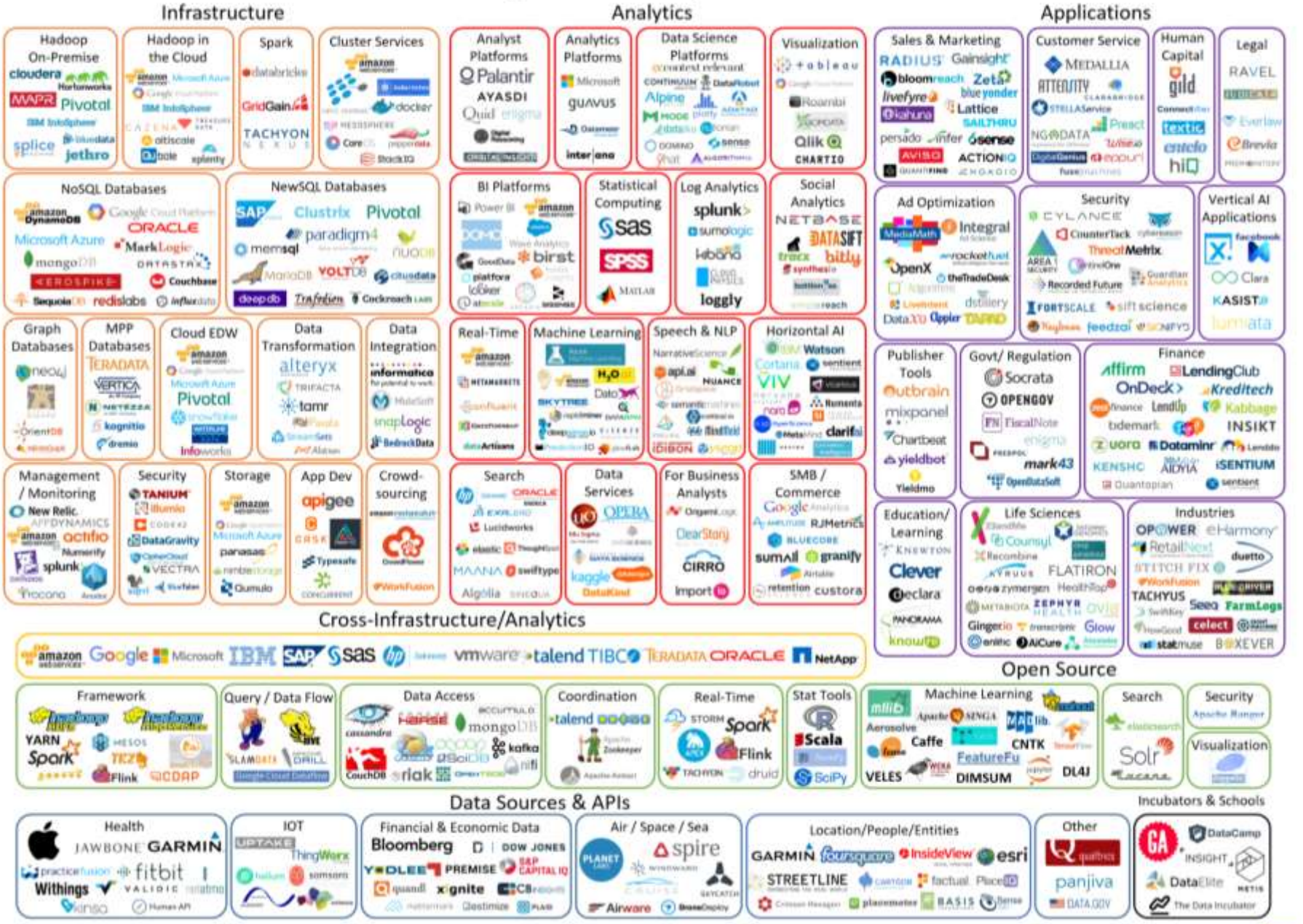
Video Analytics / Satellite Image Analytics


Something to keep in mind
Big Data-specific applications
Infrastructure to make it work

TIME (IN MINUTES) TO READ 1TB OF DATA

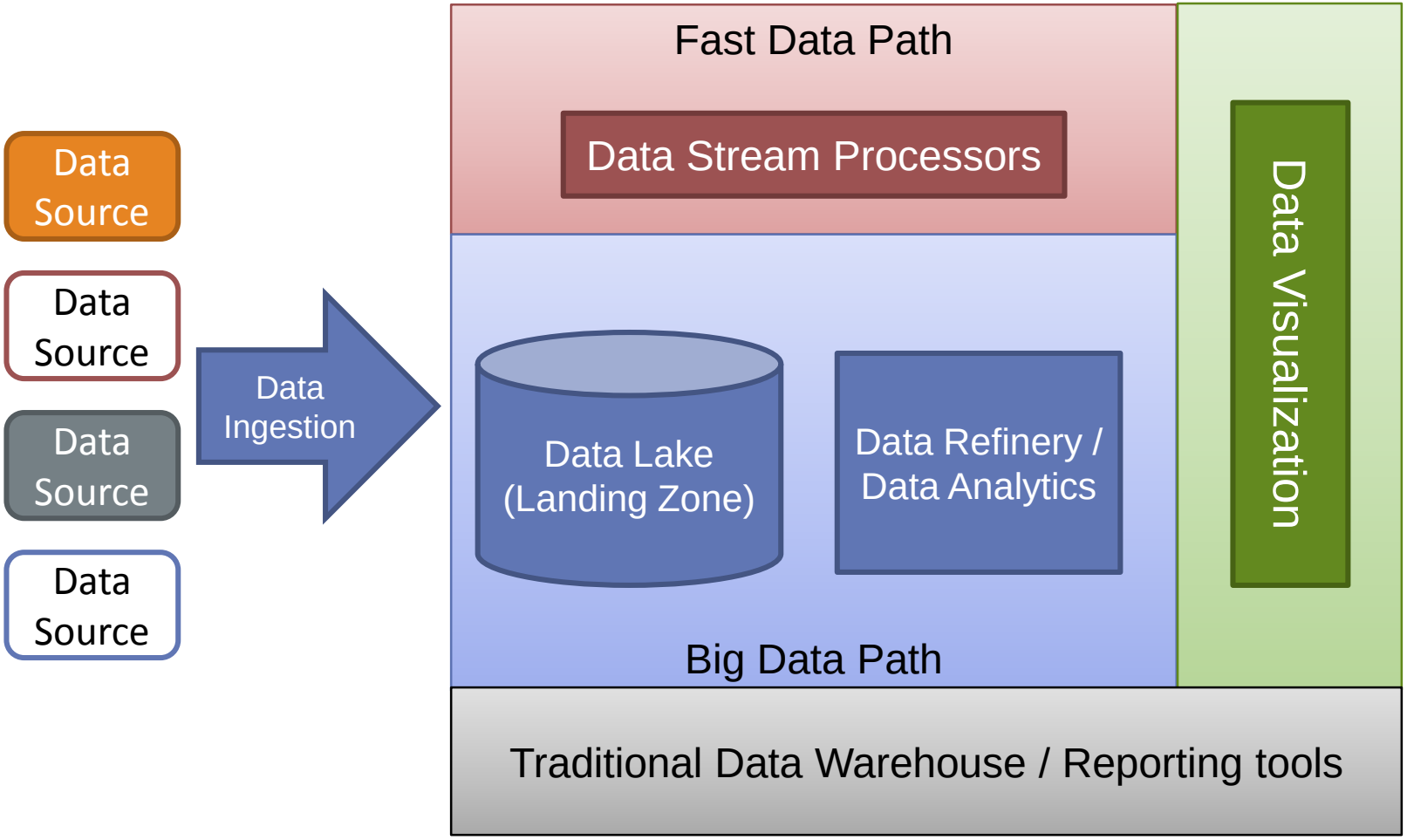


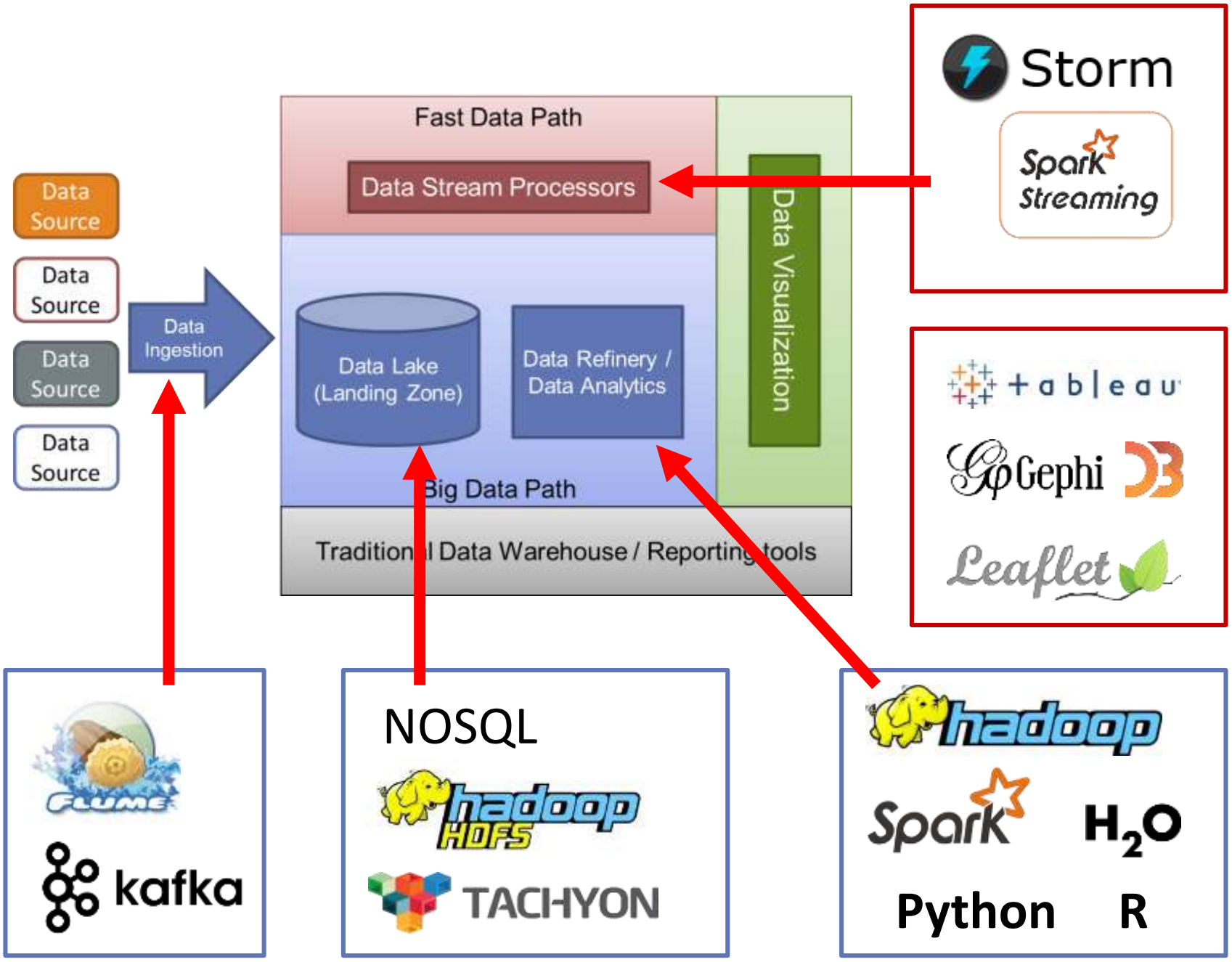
Big Data Landscape 2016



FIRSTMARK

TYPICAL BIG DATA ARCHITECTURE




Storm
Spark Streaming
Fast Data Path
Data Stream Processors
Data Visualization
Data Source
Data Source
Data Source
Data Source
Data Ingestion
Data Lake (Landing Zone)
Data Refinery / Data Analytics
Big Data Path
Traditional Data Warehouse / Reporting tools
tableau
Gephi
D3
Leaflet
FLUME
kafka
NOSQL
hadoop HDFS
TACHYON
hadoop
Spark
H2O
Python
R

Opensource software framework inspired by Google Search Engine Architecture

Provide easy-to-program scale-out foundation for data-intensive applications on large clusters of commodity hardware

Hadoop File System (HDFS) has been widely used

Users: Yahoo!, Facebook, Amazon, eBay, American Airline, Apple, Google, HP, IBM, Microsoft, Netflix, New York Times, etc.

Products: IBM InfoSphere BigInsights, Google App Engine, Oracle Big Data Appliance, Microsoft HDInsight

In-Memory Data Processing from UC Berkeley

Extend MapReduce model to support batch executions, interactive queries, and stream processing

Support various languages (Java, Python, Scala, R) with built-in analytic libraries (machine learning, graph processing)

Strong and growing community

High performance, based on sorting benchmarks, Spark is 10x – 100x faster than Hadoop

NOSQL – NOT ONLY SQL

Special DBMS for large data that does not require relational model e.g. unstructured data

Various types: Document Store, Graph, Key-Value store, etc.

Products: Parquet, Cassandra, HBASE, ElasticSearch, Accumulo, DynamoDB, Redis, Riak, CouchDB, MangoDB, Neo4j, etc.

299 systems in ranking, March 2016

Rank			DBMS	Database Model	Score		
Mar 2016	Feb 2016	Mar 2015			Mar 2016	Feb 2016	Mar 2015
1.	1.	1.	Oracle	Relational DBMS	1472.01	-4.13	+2.93
2.	2.	2.	MySQL	Relational DBMS	1347.71	+26.59	+86.62
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1136.49	-13.73	-28.31
4.	4.	4.	MongoDB	Document store	305.33	-0.27	+30.32
5.	5.	5.	PostgreSQL	Relational DBMS	299.62	+10.97	+35.19
6.	6.	6.	DB2	Relational DBMS	187.94	-6.55	-10.91
7.	7.	7.	Microsoft Access	Relational DBMS	135.03	+1.95	-6.66
8.	8.	8.	Cassandra	Wide column store	130.33	-1.43	+23.02
9.	↑ 10.	↑ 10.	Redis	Key-value store	106.22	+4.14	+9.17
10.	↓ 9.	↓ 9.	SQLite	Relational DBMS	105.77	-1.01	+4.06
11.	↑ 12.	↑ 15.	Elasticsearch	Search engine	80.17	+2.33	+21.24
12.	↓ 11.	↓ 11.	SAP Adaptive Server	Relational DBMS	76.64	-3.39	-8.72
13.	13.	13.	Teradata	Relational DBMS	74.07	+0.69	+1.29
14.	14.	↓ 12.	Solr	Search engine	69.37	-2.91	-12.52
15.	↑ 16.	↓ 14.	HBase	Wide column store	52.41	+0.39	-8.32
16.	↓ 15.	↑ 17.	Hive	Relational DBMS	50.51	-2.26	+11.18
17.	17.	↓ 16.	FileMaker	Relational DBMS	47.93	+0.90	-4.41
18.	18.	↑ 19.	Splunk	Search engine	43.73	+0.90	+8.01
19.	19.	↑ 21.	SAP HANA	Relational DBMS	39.99	+1.91	+7.82
20.	↑ 21.	↑ 23.	Neo4j	Graph DBMS	32.36	+0.07	+4.73

Source: http://db-engines.com/en/ranking

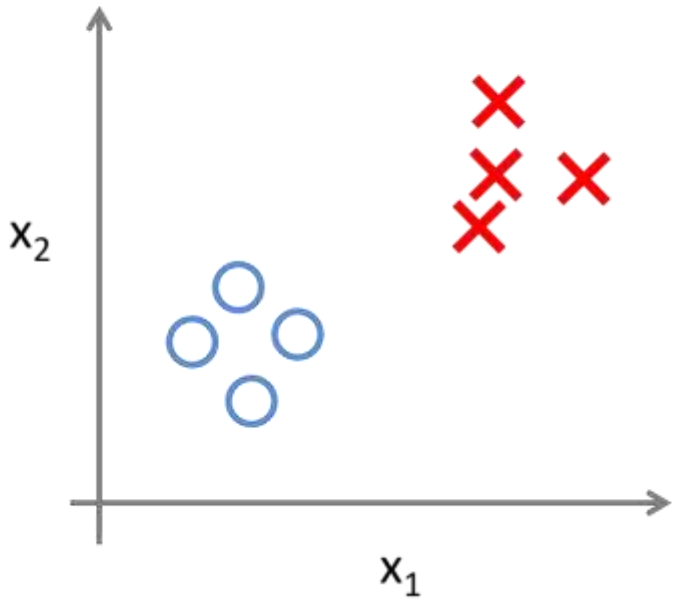
PREDICTIVE ANALYTICS

Analyze current and historical data to automatically find patterns based on several techniques e.g. statistics, modeling, machine learning, data mining, time series analysis, deep learning, etc.

Utilize other techniques e.g. text analytics, image processing, location analytics, etc.

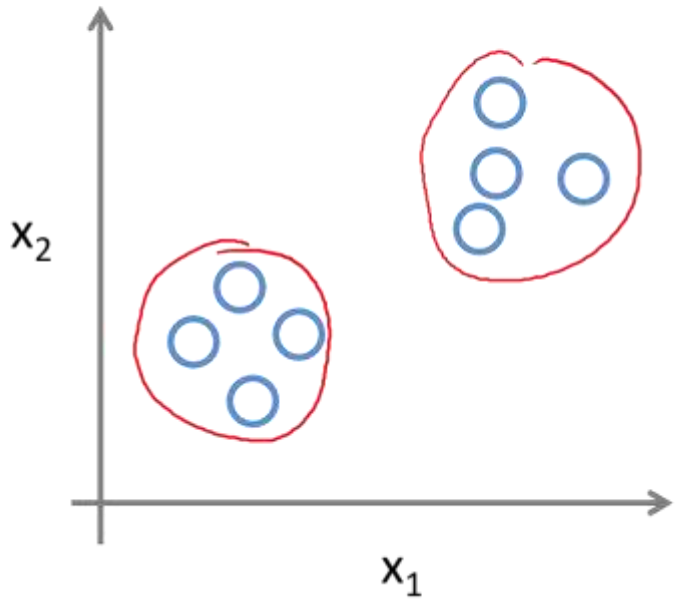
Applications: Micro Customer Segmentation, Sentiment Analysis, Customer retention, Fraud detection, etc.

Supervised Learning



Database marketing
Fraud detection
Pattern detection
Churn customer detection
Web classification

Unsupervised Learning



Customer Segmentation
Collaborative Filtering

OTHER ANALYTICS

Spatial Analytics

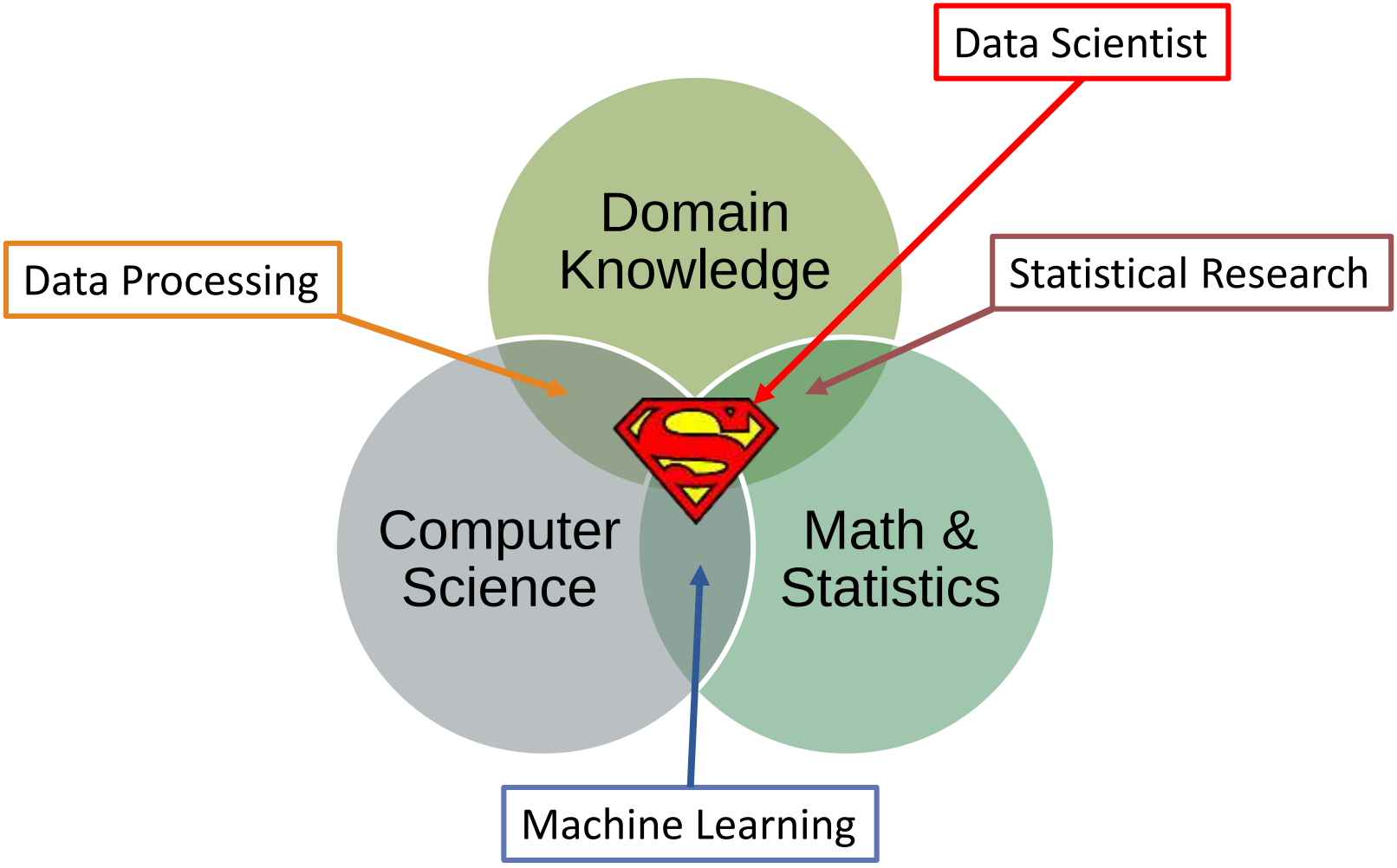
Mobility Analytics

Social Network Analytics

"Big data is about having the technology and people with the appropriate analysis skills to allow firms to make sense of huge volumes of data in an affordable manner."

Source: Forrester Research, 2012

“Data Science is a Team Sport” – DJ Patil



Data Driven Organization

CS PROGRAM

Architecture Track

- Map/Reduced
- In-Memory Processing
- Cloud Computing
- Mobile and Networks

Analytics Track

- Machine Learning
- Data Mining
- Big Data Analytics
- Social Network Analysis