THE AGE OF BIG DATA

Asst. Prof. Natawut Nupairoj, Ph.D.

Mobile Application and System Services Research Group

Department of Computing Engineering

Chulalongkorn University

natawut.n@chula.ac.th

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

“By 2020, total digital information will be 40ZB or 5.2TB for each human” - IDC

“Just 3% of all data is currently tagged and ready for manipulation, and only one sixth of this - 0.5% - is used for analysis” - IDC

B | KB | MB | GB | TB | PB | EB | ZB

BIG DATA MOVEMENTS

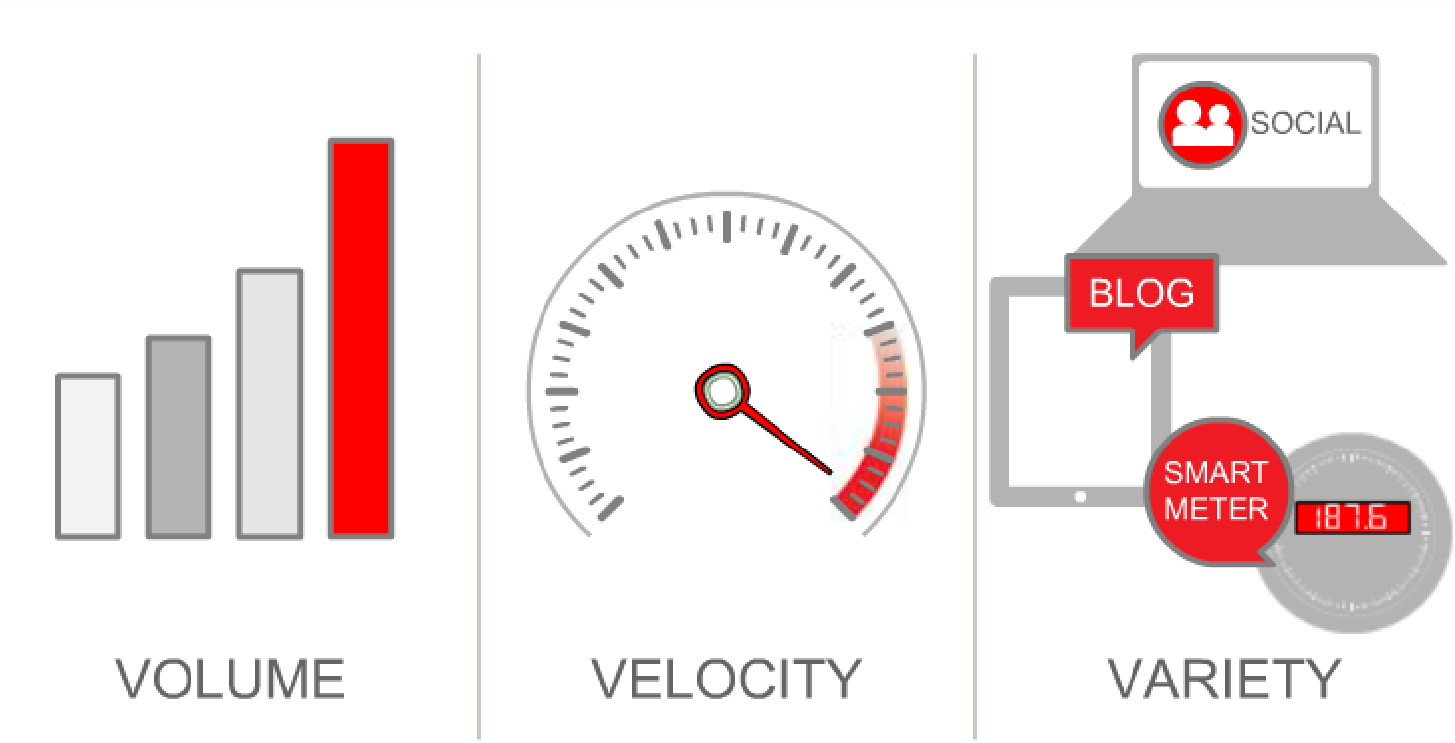
Big Data Analytic worldwide market will reach $50 billion by 2018

Big data technology and services will grow worldwide at annual growth rate of 40% – about seven times that of the ICT market overall

There are at least 23 US University offering Master’s Programs in Data Science

By 2018, there will be 140,000–190,000 data scientist job postings that go unfulfilled

“3VS” MODEL OF BIG DATA BY GARTNER (2012)



VOLUME

One-fifth of organizations store more than 1 petabyte of data

100TBs of data are uploaded to facebook each day

Walmart collects more than 2.5PBs from 1 million customers each hour

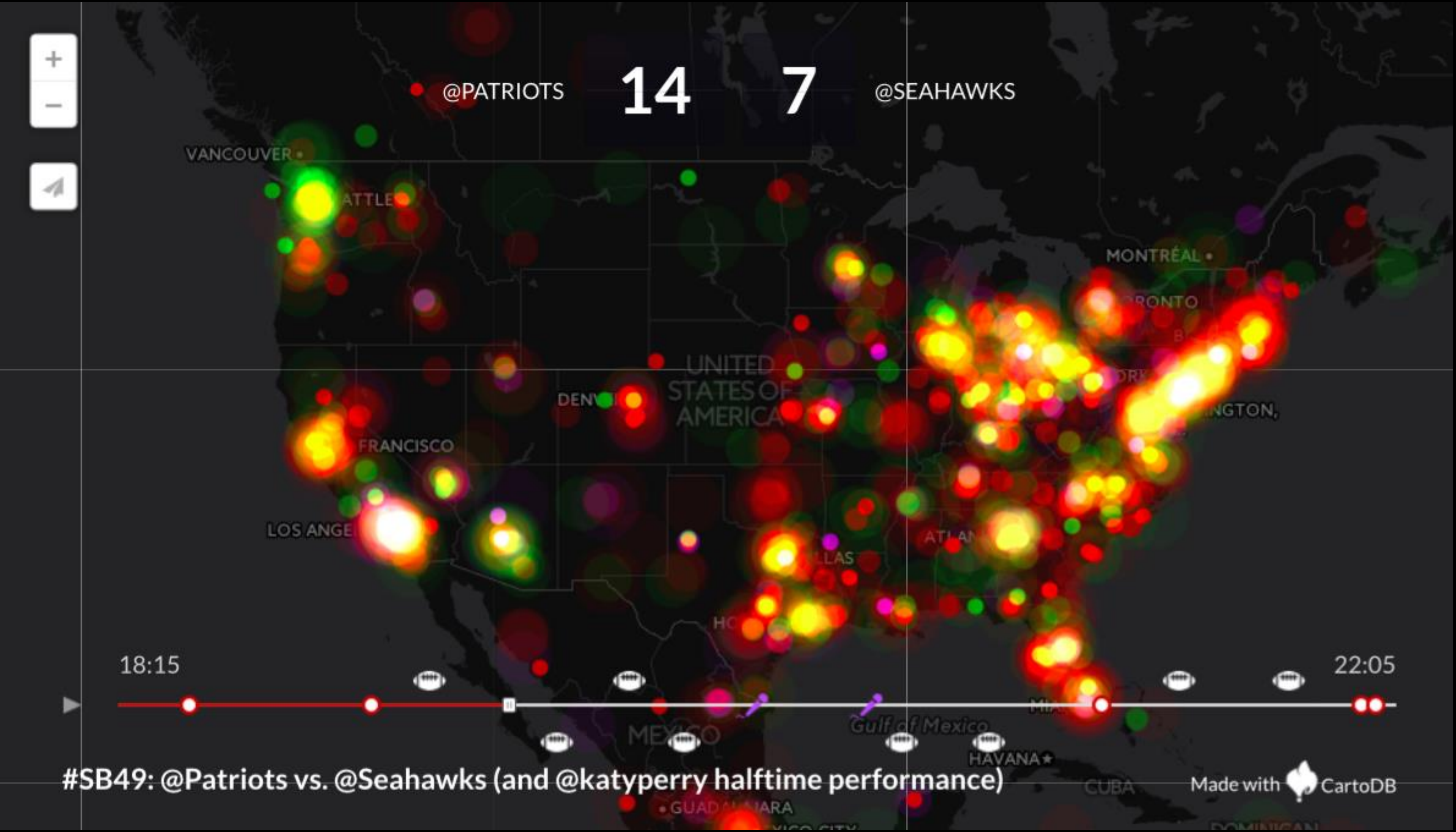
VELOCITY

Real-time data processing aka. Fast Data becomes common

Large Hadron Collider has about 150 million sensors delivering data 40 million times per second

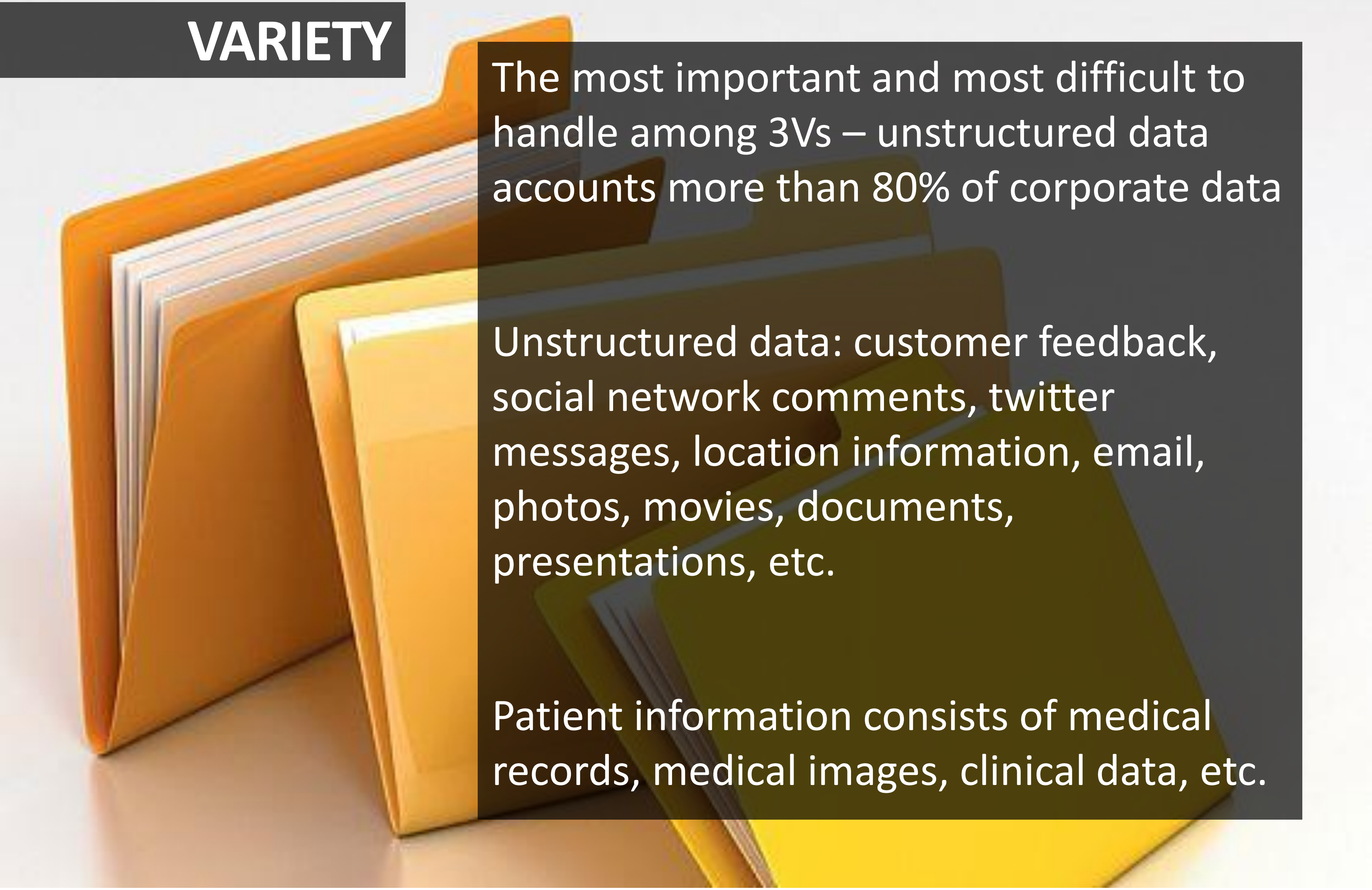
Google handles 2m search queries per minute

Superbowl 2015 final stats: 28.4M tweets with a peak of 395k tweets-per-minute



#SB49: @Patriots vs. @Seahawks (and @katyperry halftime performance)

VARIETY

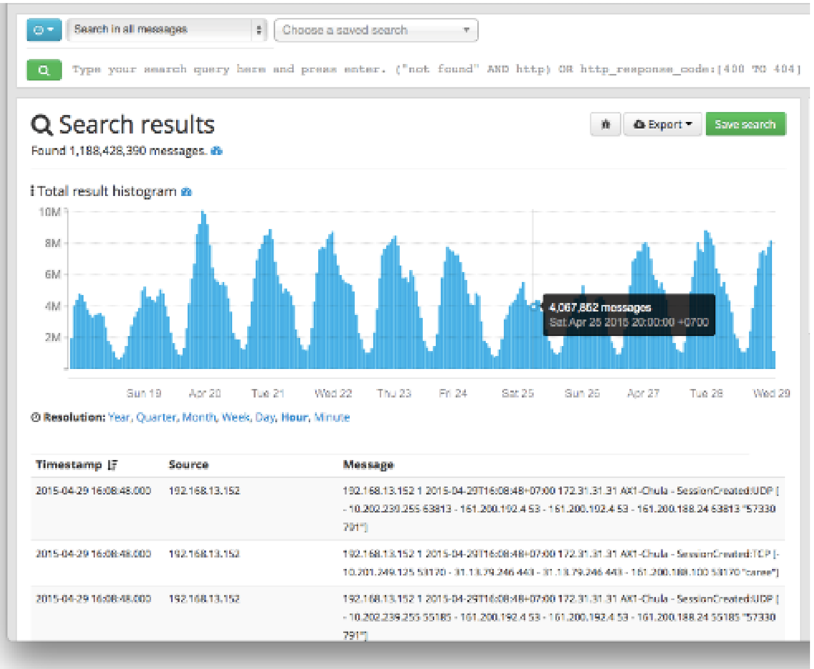


The most important and most difficult to handle among 3Vs – unstructured data accounts more than 80% of corporate data

Unstructured data: customer feedback, social network comments, twitter messages, location information, email, photos, movies, documents, presentations, etc.

Patient information consists of medical records, medical images, clinical data, etc.

SIMPLE EXAMPLE



Type your search query here and press enter. ("not found" AND http) OR http_response_code:[400 TO 404]

Search results

Found 1,188,428,390 messages.

Timestamp	Source	Message
2015-04-29 16:08:48.000	192.168.13.152	192.168.13.152 1 2015-04-29T16:08:48+07:00 172.31.31.31 AX1-Chula - SessionCreated:UDP [- 10.202.230.255 63813 - 161.200.192.4 53 - 161.200.192.4 53 - 161.200.188.24 63813 "57330 791"]
2015-04-29 16:08:48.000	192.168.13.152	192.168.13.152 1 2015-04-29T16:08:48+07:00 172.31.31.31 AX1-Chula - SessionCreated:TCP [- 10.201.249.125 53170 - 31.13.79.246 443 - 31.13.79.246 443 - 161.200.188.100 53170 "caree"]
2015-04-29 16:08:48.000	192.168.13.152	192.168.13.152 1 2015-04-29T16:08:48+07:00 172.31.31.31 AX1-Chula - SessionCreated:UDP [- 10.202.239.255 55185 - 161.200.192.4 53 - 161.200.192.4 53 - 161.200.188.24 55185 "57330 791"]

Chula IT collects various logs from 40 servers, specifically for computer crime act 2007

Velocity = 2,000 events/sec

Volume = approx. 30+ GBs each day (or 3TBs for 3 months)

Variety = Different data sources, different data schema

BIG DATA’S DRIVERS
MOBILE PHONES AND PERSONAL DEVICES

World mobile penetration rate is more than 95%

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

Facebook has 798M daily active mobile users

Future: wearable computers

BIG DATA'S DRIVERS
INTERNET OF THINGS AND SENSOR NETWORKS

RFID, GPS, sensors, surveillance cameras, smart meters, appliances, medical devices, toys, etc.

212b sensors will be available with 30b IOT are expected to be connected to networks by 2020

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

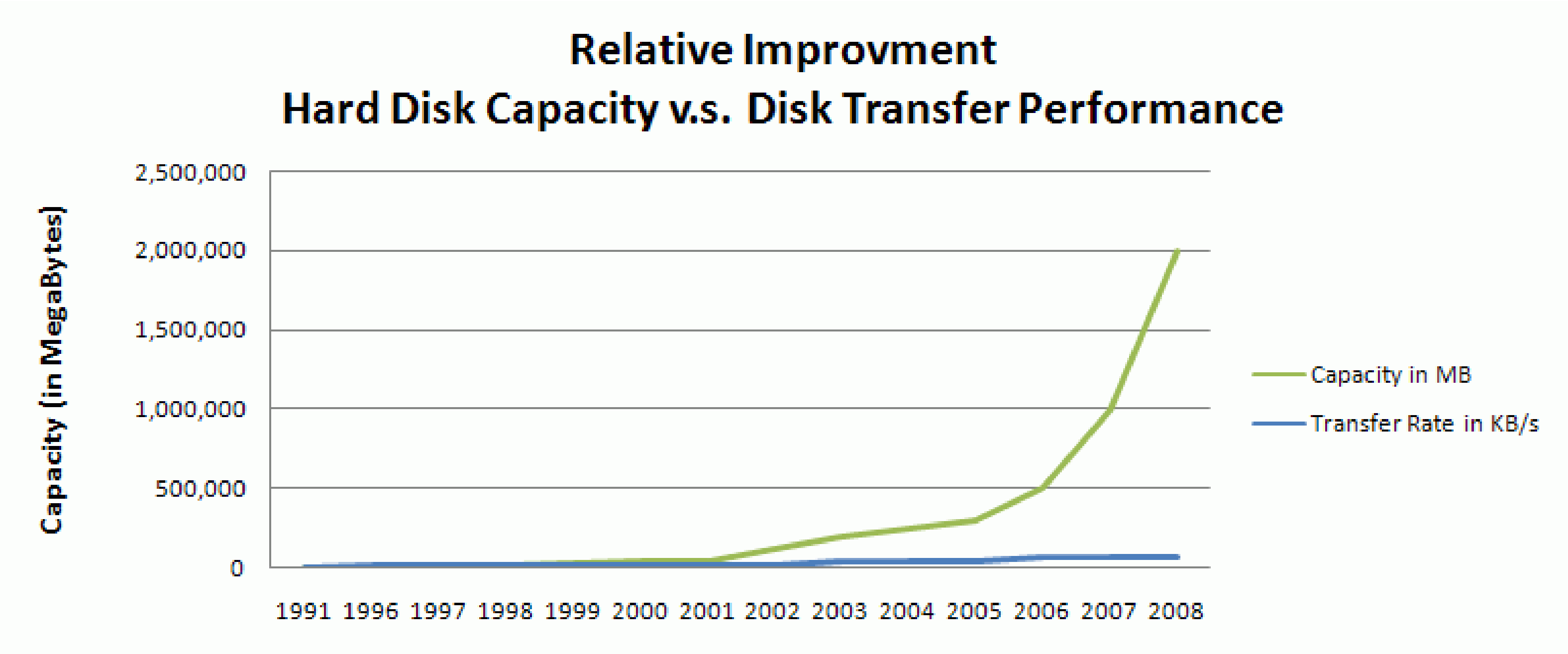
BIG DATA'S DRIVERS
CLOUD COMPUTING

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

HOW CAN WE RESPOND TO THESE CHALLENGES?



Source: http://wiki.r1soft.com/pages/viewpage.action?pageId=3016608

Disk data transfer improves very slowly

Increase throughput with **scale-out** – processing data on multiple disks in parallel

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

Spark

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.

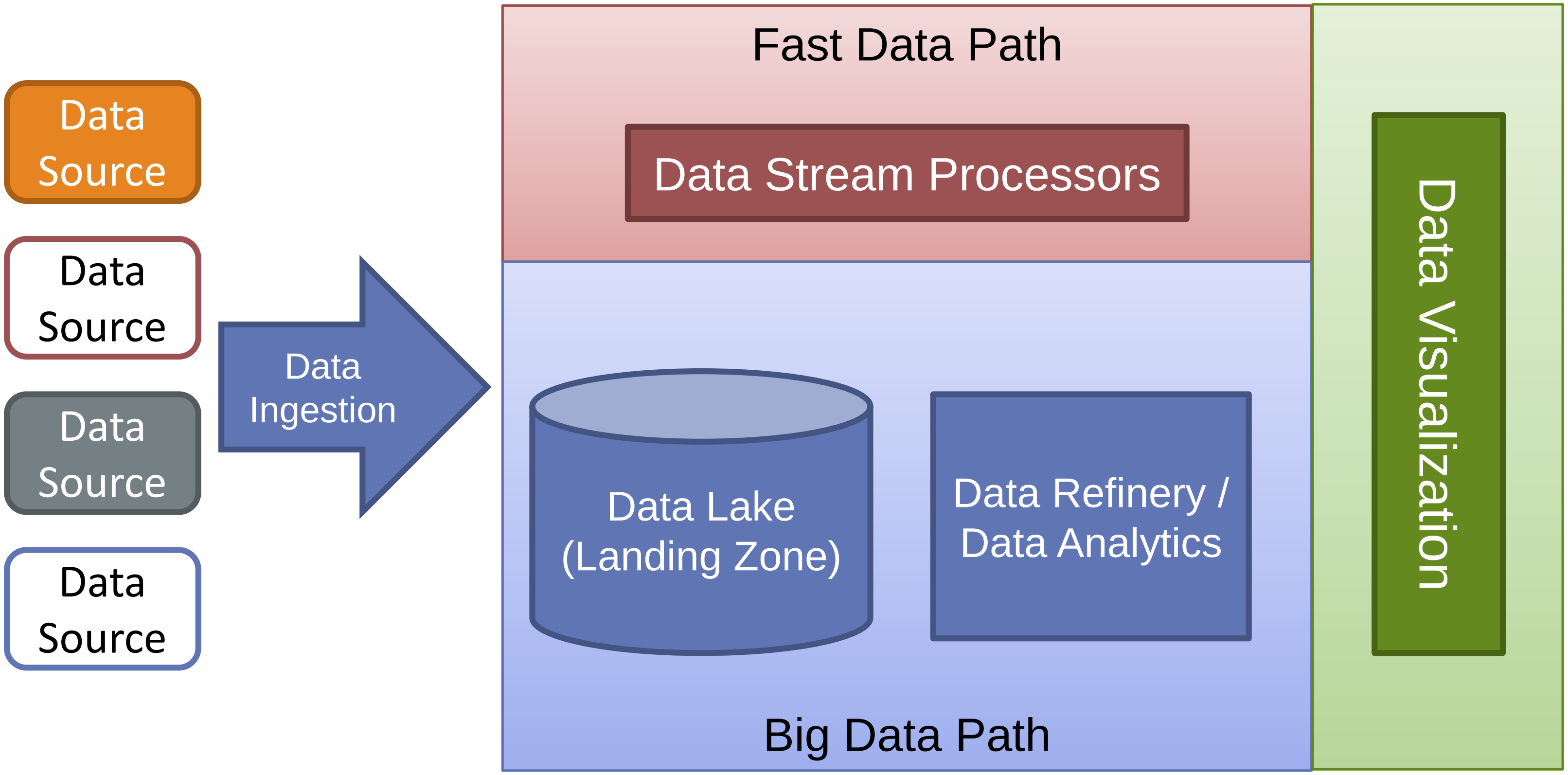
PREDICTIVE ANALYTICS

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.

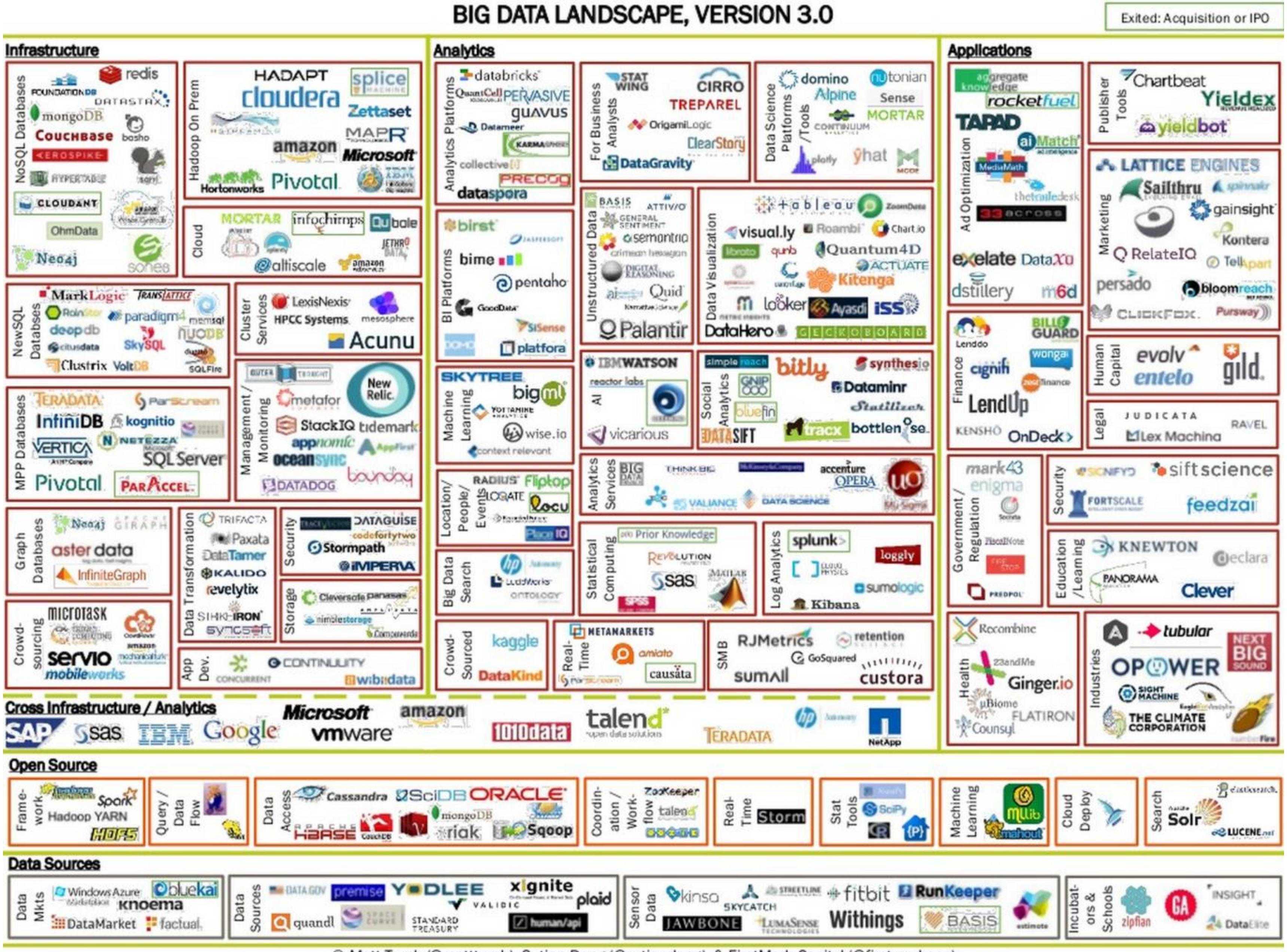
Utilize other techniques e.g. text analytics, image processing, location analytics, etc.

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

PUTTING THEM ALL TOGETHER



BIG DATA LANDSCAPE, VERSION 3.0



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

TYPICAL USE CASES

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

Product Recommendation

Social Listening

Fraud Detection and Risk Management

Micro Customer Segmentation

Demand Sensing for Supply Chain

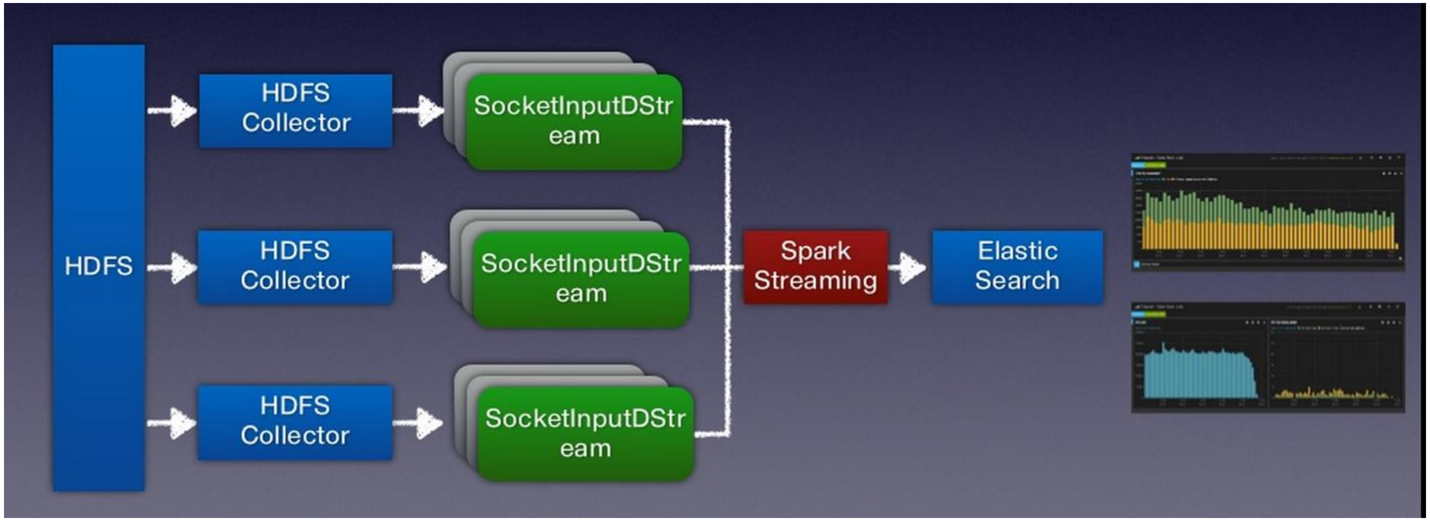
Bio Informatics

CASE STUDY:
SK TELECOM'S USAGE PATTERN ANALYSIS

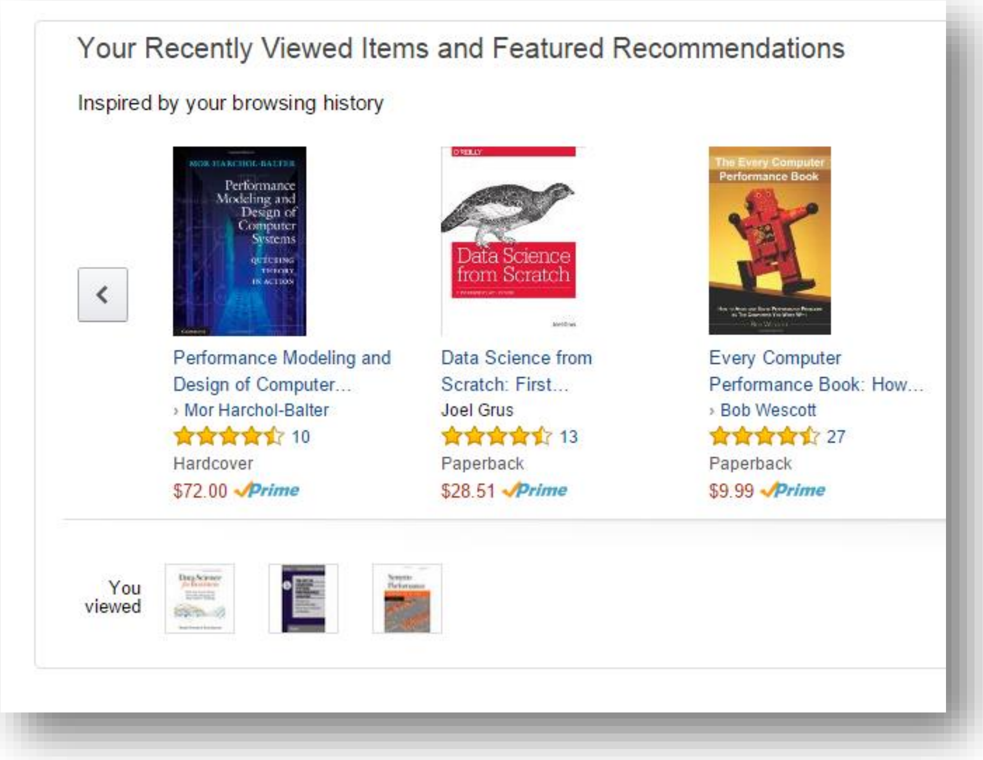
Process usage data from 28 millions subscribers: 40TB/day – 15PB total

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)



CASE STUDY:
AMAZON'S RECOMMENDATION ENGINE



Your Recently Viewed Items and Featured Recommendations

Inspired by your browsing history

Performance Modeling and Design of Computer...
› Mor Harchol-Balter
★★★★☆ 10
Hardcover
$72.00 ✓Prime

Data Science from Scratch: First...
Joel Grus
★★★★☆ 13
Paperback
$28.51 ✓Prime

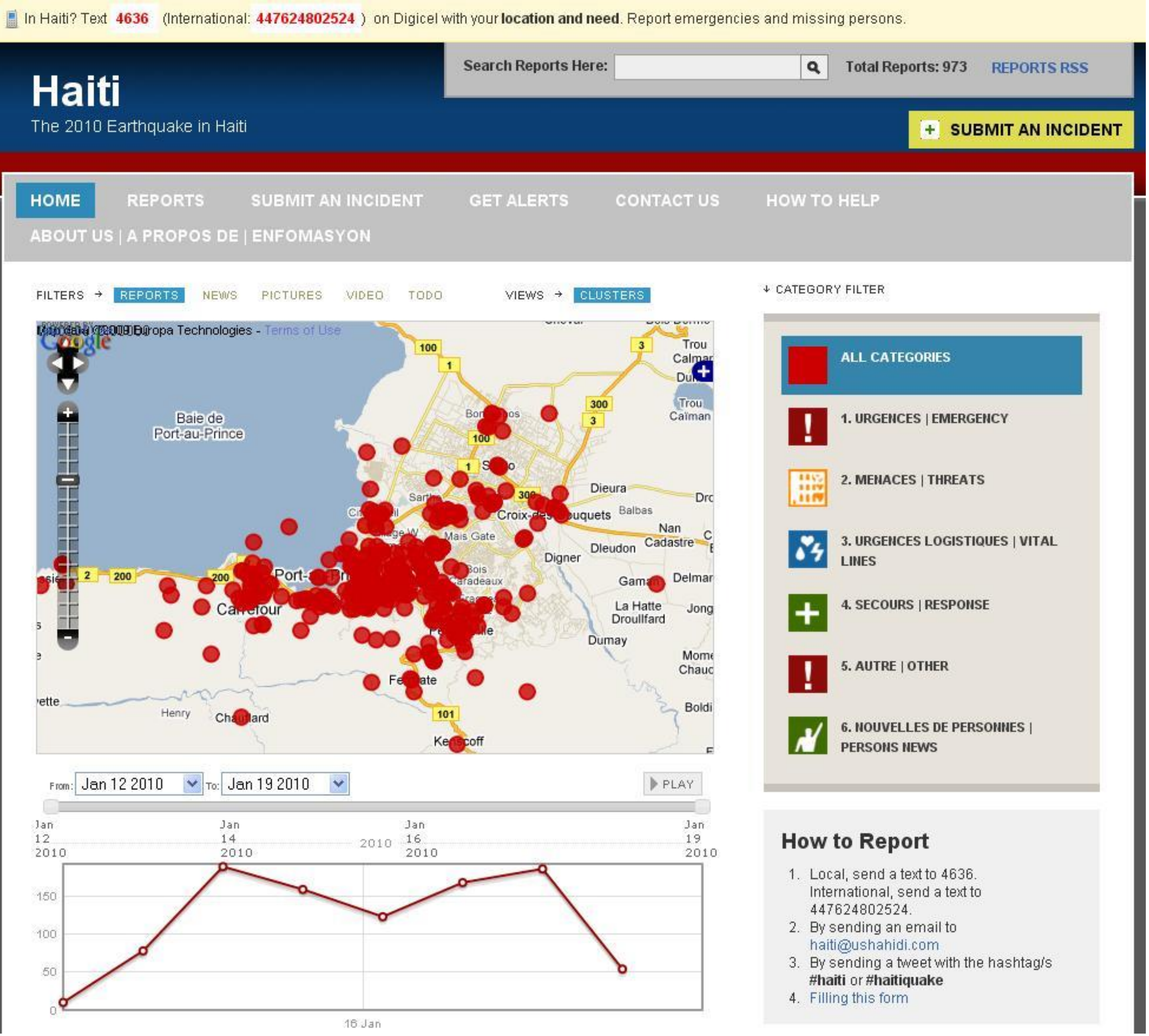
Every Computer Performance Book: How...
› Bob Wescott
★★★★☆ 27
Paperback
$9.99 ✓Prime

You viewed

Amazon mines data from 152 million customers to suggest products to customers

Use Hadoop + DynamoDB to perform collaborative filtering, click-stream analysis, historical purchase data analytics

Other similar offerings: Facebook, LinkedIn, Netflix



In Haiti? Text 4636 (International: 447624802524) on Digicel with your **location and need**. Report emergencies and missing persons.

Haiti

The 2010 Earthquake in Haiti

Search Reports Here: Total Reports: 973 REPORTS RSS

SUBMIT AN INCIDENT

HOME REPORTS SUBMIT AN INCIDENT GET ALERTS CONTACT US HOW TO HELP

ABOUT US | A PROPOS DE | ENFOMASYON

FILTERS → REPORTS NEWS PICTURES VIDEO TODO VIEWS → CLUSTERS

From: Jan 12 2010 To: Jan 19 2010 PLAY

CATEGORY FILTER

- ALL CATEGORIES
- 1. URGENCES | EMERGENCY
- 2. MENACES | THREATS
- 3. URGENCES LOGISTIQUES | VITAL LINES
- 4. SECOURS | RESPONSE
- 5. AUTRE | OTHER
- 6. NOUVELLES DE PERSONNES | PERSONS NEWS

How to Report

1. Local, send a text to 4636. International, send a text to 447624802524.
2. By sending an email to haiti@ushahidi.com
3. By sending a tweet with the hashtag/s **#haiti** or **#haitiquake**
4. Filling this form

USHAHIDI

2007
Kenya

2010
Haiti
Chile
Washington DC
Russia

2011
Christchurch
Middle East
India
Japan
Australia
US
Macedonia

2012
Balkans

2014
Kenya

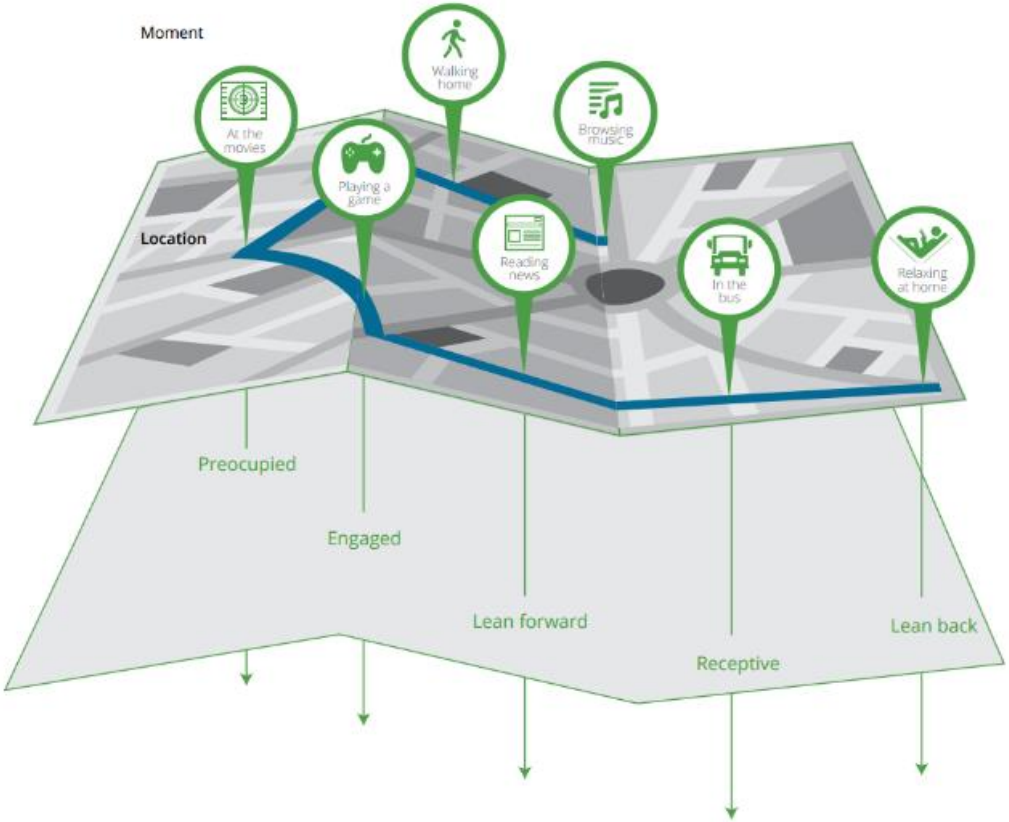
CASE STUDY: JP MORGAN

In 2013, 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

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

Analyze 250b monthly activities of 750m customers for 65m POI in 50+ countries -- 500TB/day

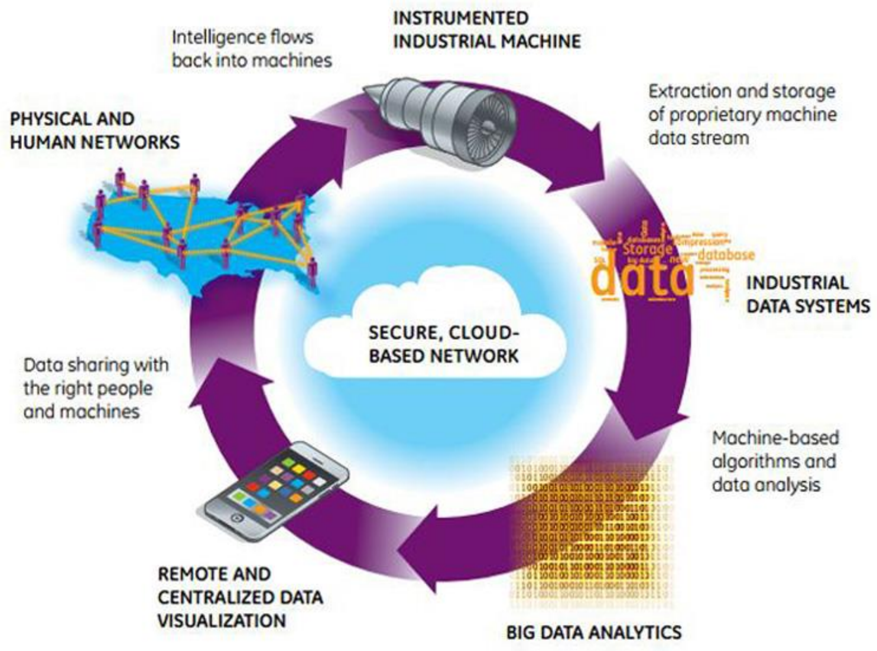
CASE STUDY:
UBER’S DYNAMIC PRICING FARES

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:
GE’S SMART MACHINES


INSTRUMENTED INDUSTRIAL MACHINE
Extraction and storage of proprietary machine data stream
INDUSTRIAL DATA SYSTEMS
Storage
database
data
Machine-based algorithms and data analysis
BIG DATA ANALYTICS
REMOTE AND CENTRALIZED DATA VISUALIZATION
Data sharing with the right people and machines
PHYSICAL AND HUMAN NETWORKS
Intelligence flows back into machines
SECURE, CLOUD-BASED NETWORK

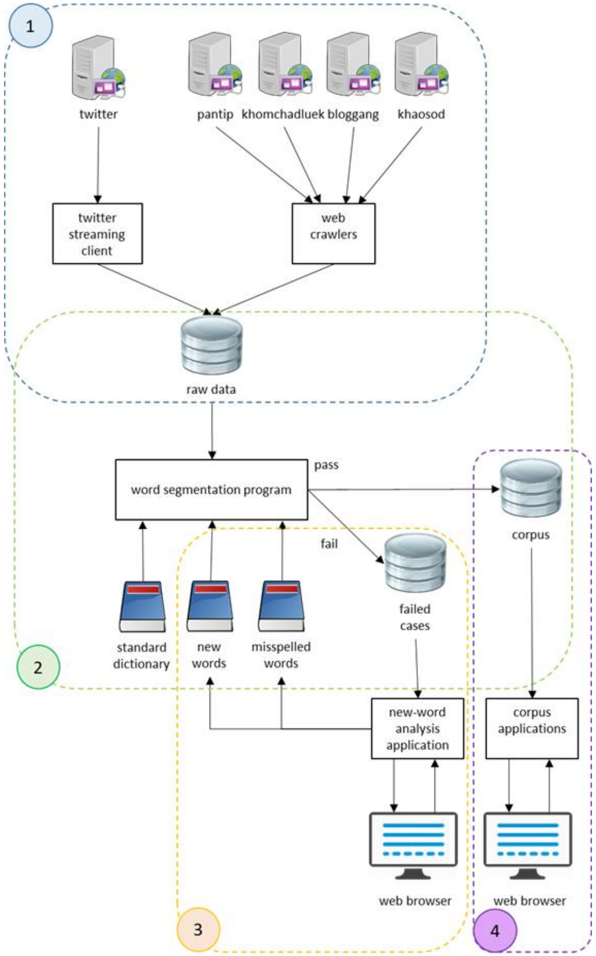
GE has launched Industrial Internet initiative

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:
SOCIAL NETWORK-BASED THAI MONITOR CORPUS



Cooperation with Department of Linguistics

Construct Thai Monitor Corpus by extracting data from social networks e.g. Twitter, forums, blogs, websites, etc.

Provide corpus repository with temporal analytic capabilities that can keep expanding

Target: 500M words

OTHER USE CASES

Automatic – Data Processing for IoT

Baidu – Business insights data analytics

Cisco – Entity resolution for Whois clean-up

eBay – Log Transaction Analytics

ESRI – Geospatial and temporal analysis

Goldman Sachs – Data curation processing

NBC Universal – Multimedia data distribution prediction

Novartis – Genomic data analytic

Salesforce – Recommendation engine

Shopify – E-commerce transaction data analytics

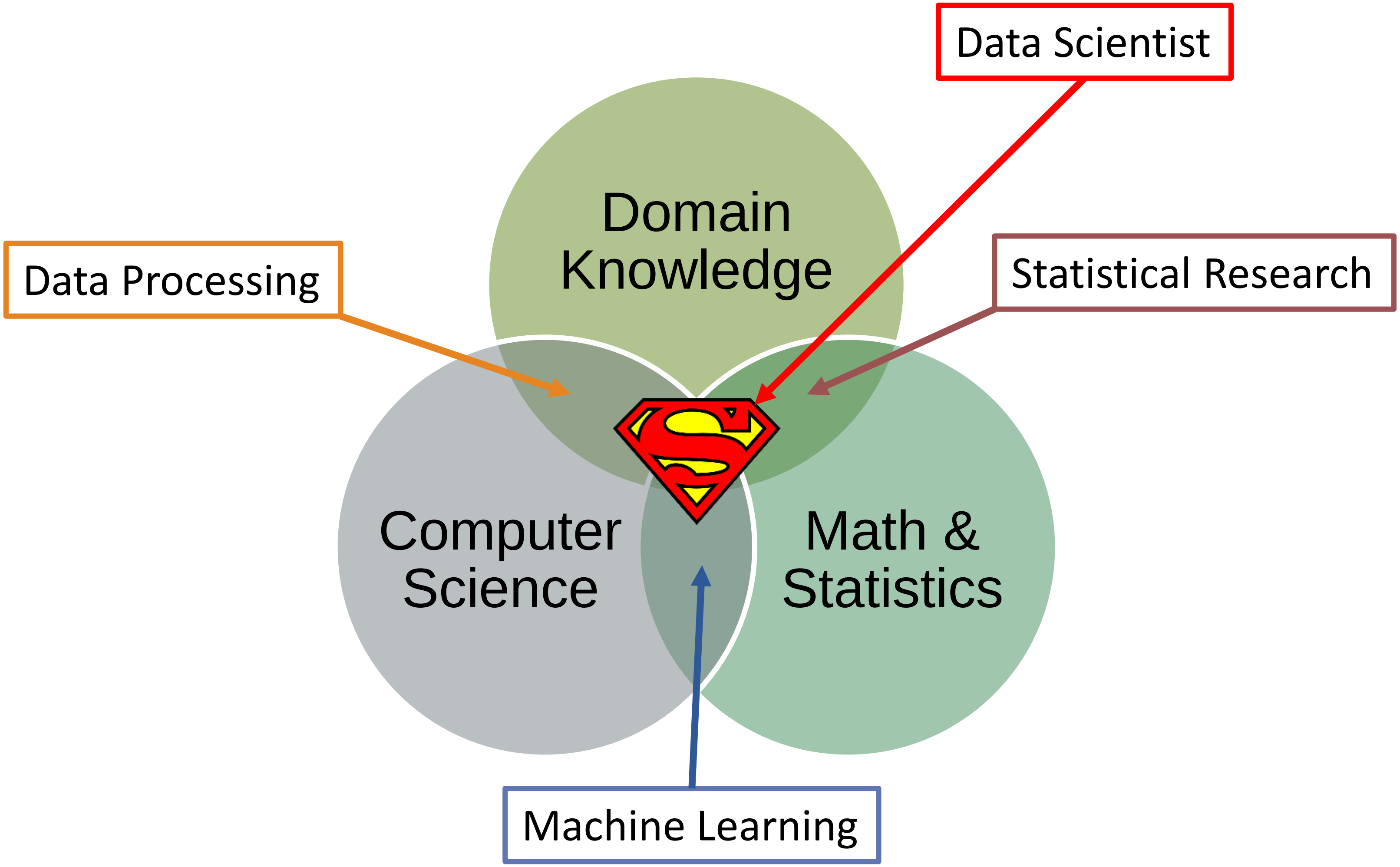
Tresata – Anti-Money Laundering real-time graph analytic

Yahoo – Real-time personalization engine

“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 SCIENTIST IS A TEAM SPORT



OUR CAPABILITIES

Big Data Architecture (6+ faculties)

- Map-Reduce and Hadoop
- In-Memory Processing
- NoSQL
- Data Ingestion
- Social Network Information Retrieval
- Internet of Things
- Embedded System
- Wireless Sensor Network
- Cloud Computing
- Mobile Computing

OUR CAPABILITIES

Big Data Analytics (7+ faculties)

- Machine Learning
- Clustering / K-Mean
- Data Mining
- Neuron Networks
- Social Network Analytics
- Text Analytics
- Image Processing
- Time-Series Analytics
- Location Analytics
- Data Visualization

COME ALONG 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 150,000 datasets to the public

FINAL THOUGHTS

We are definitely in the age of Big Data
(for at least 2 years)

Big Data is more than just “Big” data

Next Battlefield … Data-Driven Economy

Are you ready?