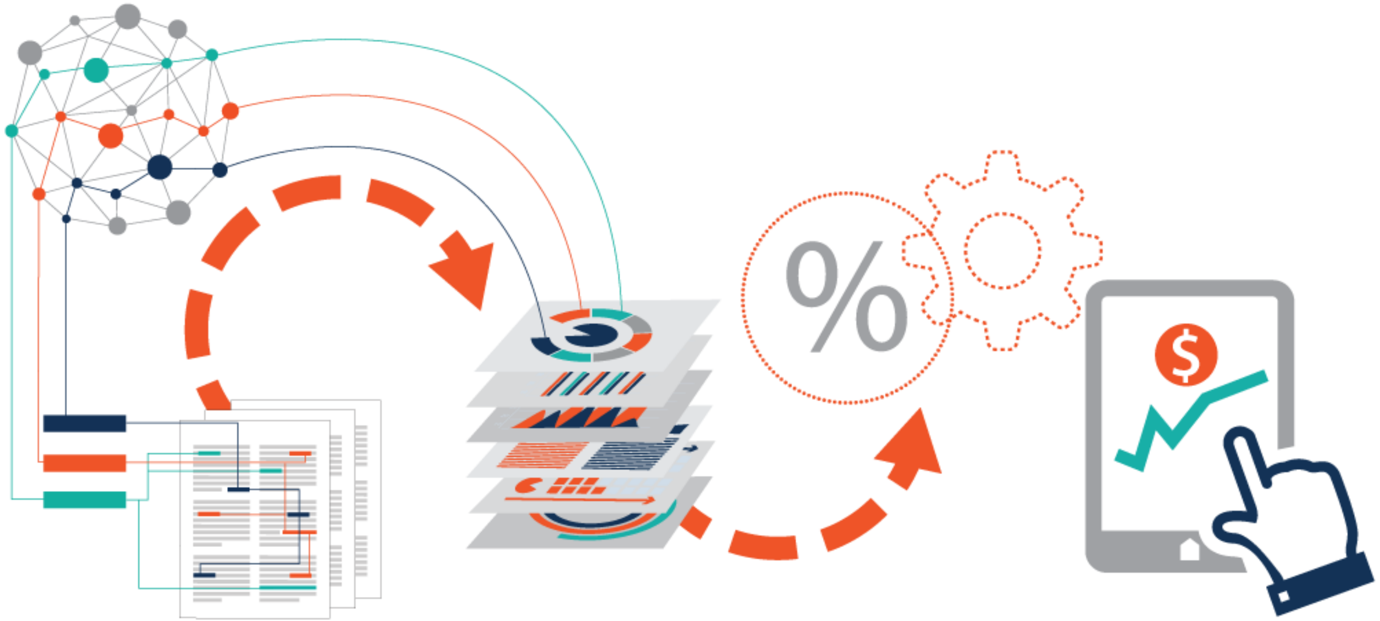
Big Data Analytics

Assoc. Prof. Dr. Tiranee Achalakul

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

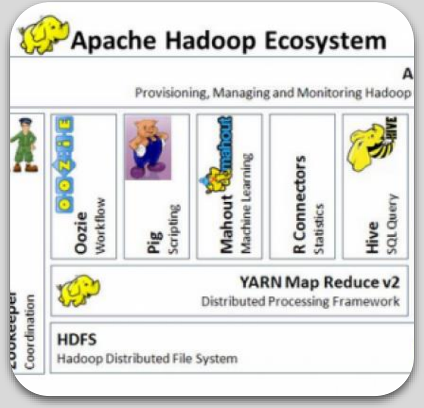


Course objectives

- Learners see practical examples of big data in action
- Learners see the overview of current big data technology
- Learners understand big data technology

Schedule

Module 1: Introduction to Big Data

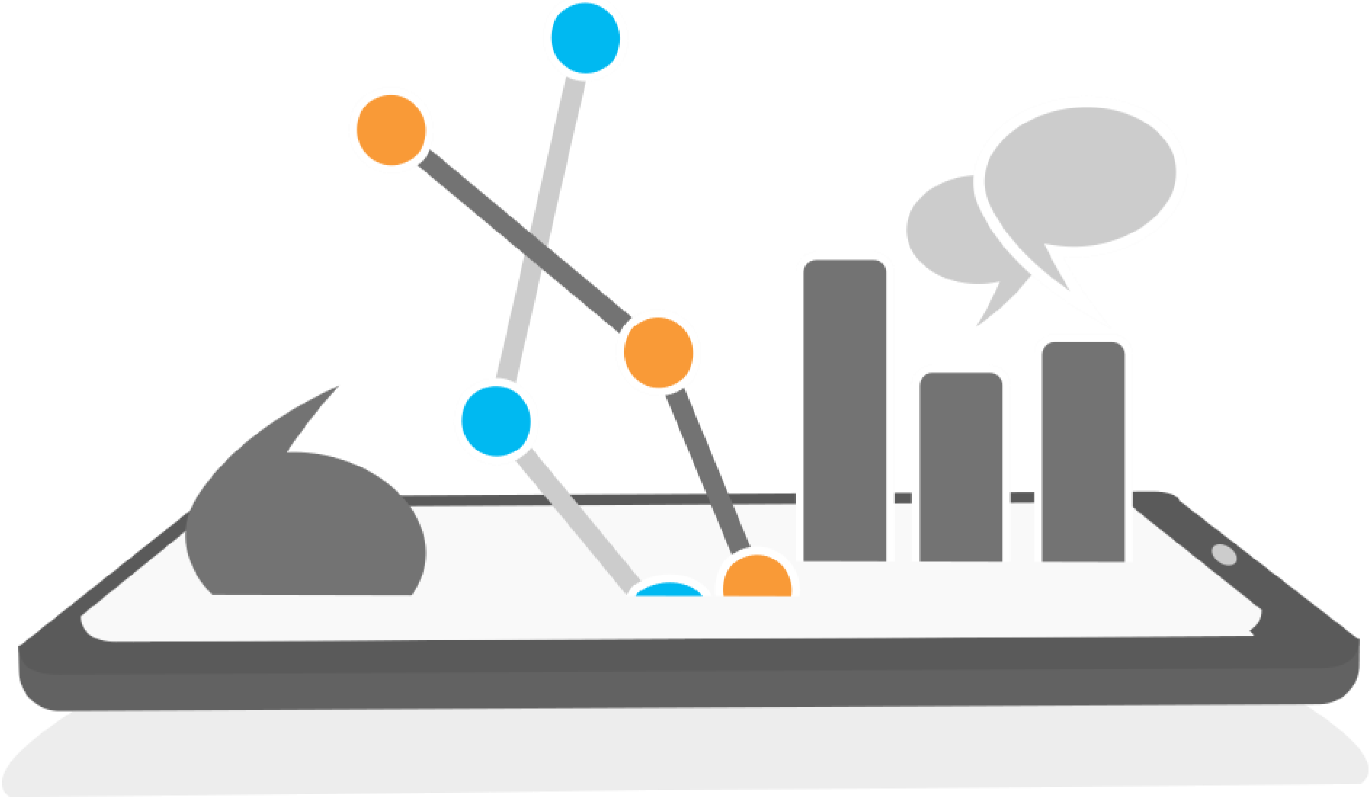


Module 2: Introduction to Big Data Technology

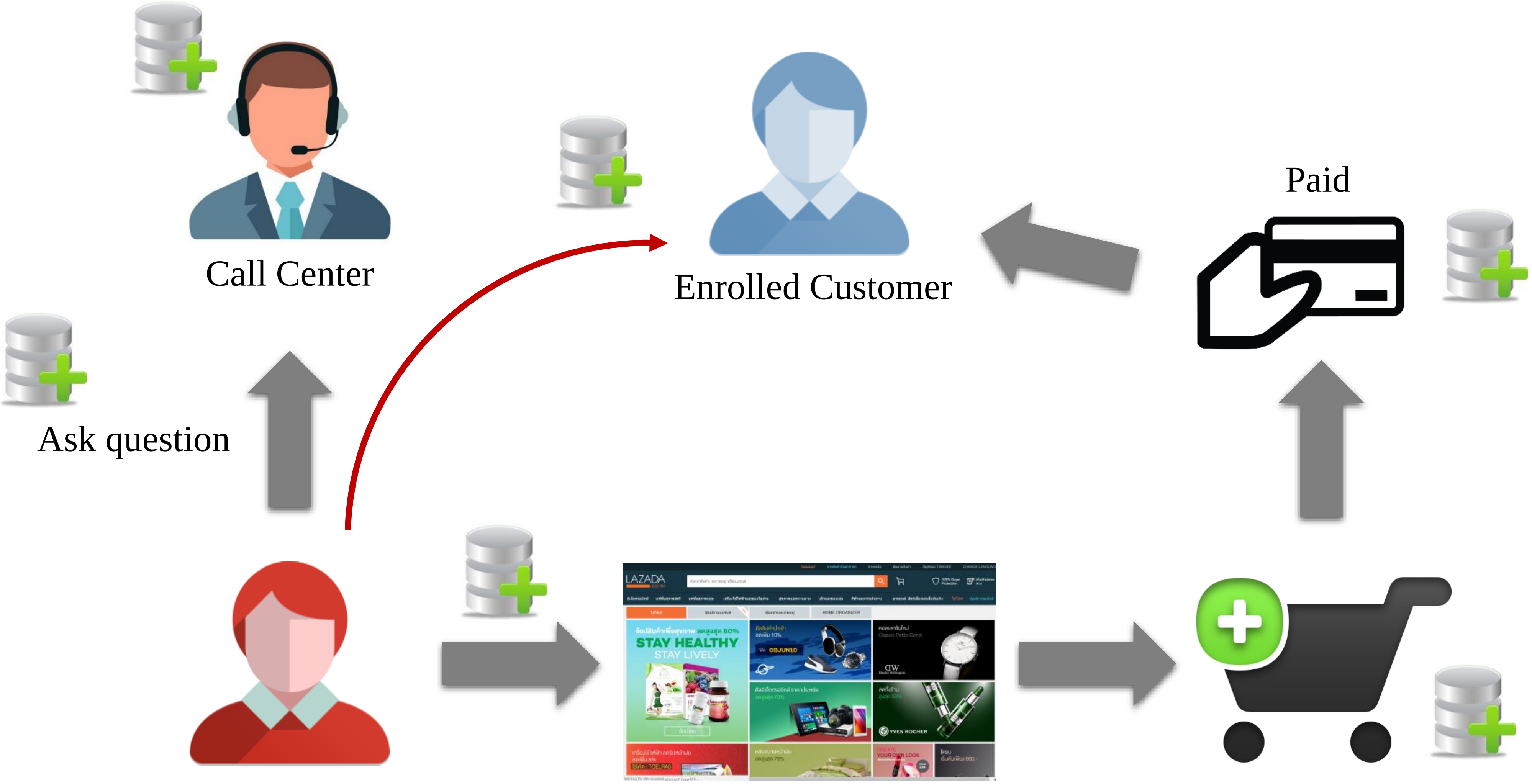
Data Mining
Machine Learning
Text Mining an NLP
Apache Hadoop

BIG DATA
EXPERIENCE

Introduction to Big Data



KMUTT g-able



Call Center

Enrolled Customer

Paid

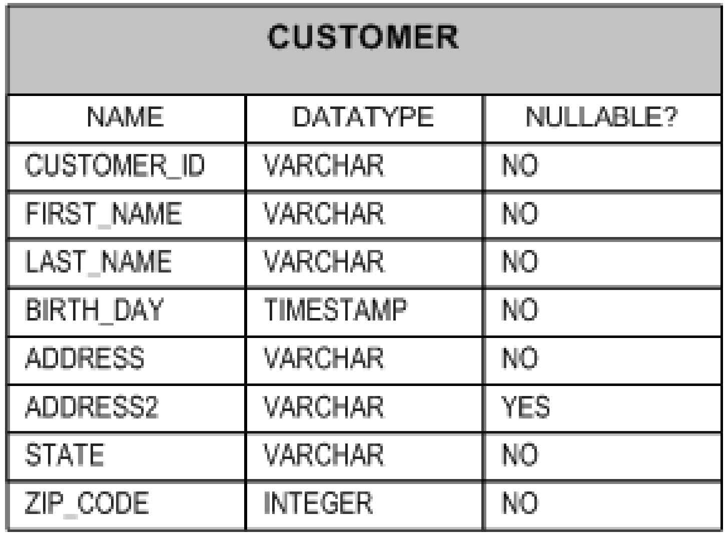
Ask question

Prospective
Customer

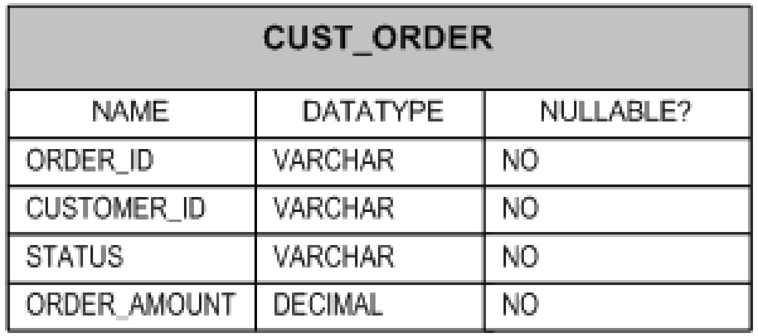
Online Store

Cart

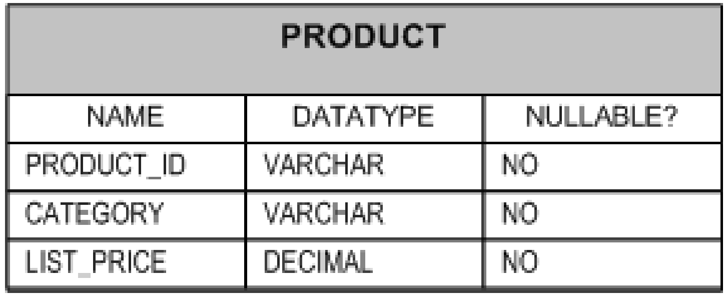
Traditional database



CUSTOMER		
NAME	DATATYPE	NULLABLE?
CUSTOMER_ID	VARCHAR	NO
FIRST_NAME	VARCHAR	NO
LAST_NAME	VARCHAR	NO
BIRTH_DAY	TIMESTAMP	NO
ADDRESS	VARCHAR	NO
ADDRESS2	VARCHAR	YES
STATE	VARCHAR	NO
ZIP_CODE	INTEGER	NO

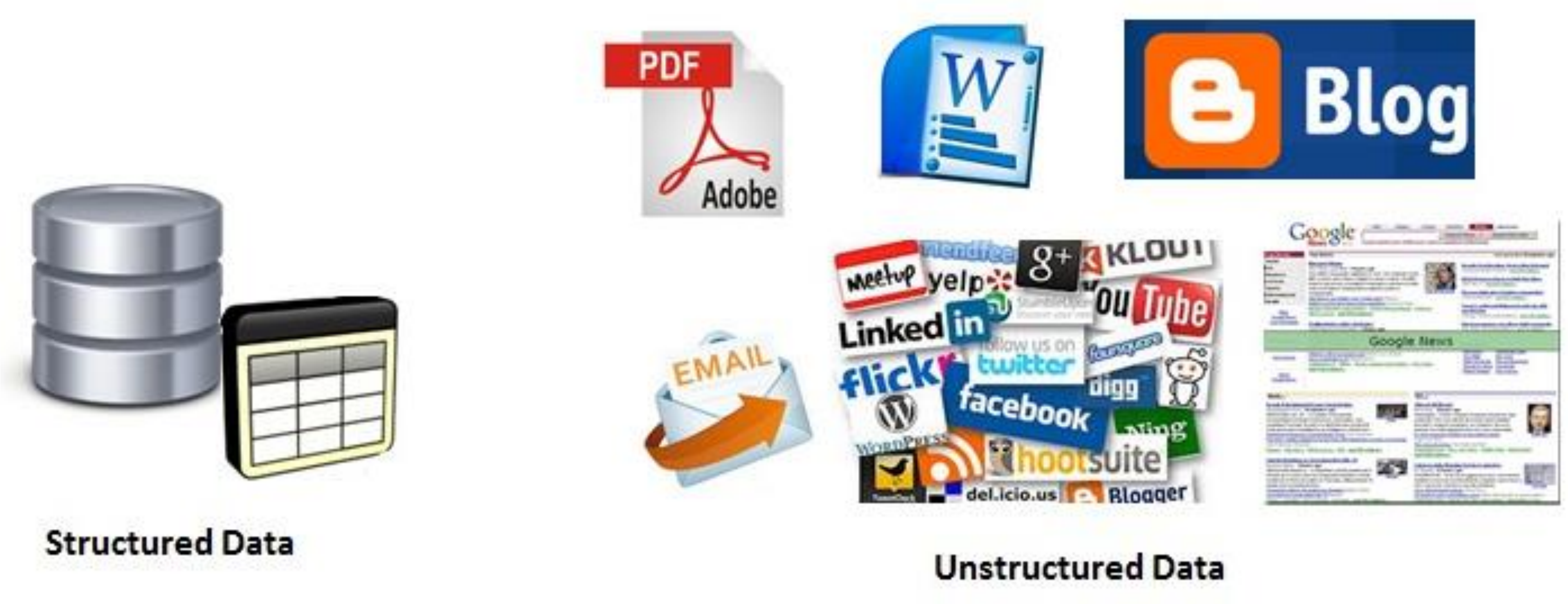


CUST_ORDER		
NAME	DATATYPE	NULLABLE?
ORDER_ID	VARCHAR	NO
CUSTOMER_ID	VARCHAR	NO
STATUS	VARCHAR	NO
ORDER_AMOUNT	DECIMAL	NO



PRODUCT		
NAME	DATATYPE	NULLABLE?
PRODUCT_ID	VARCHAR	NO
CATEGORY	VARCHAR	NO
LIST_PRICE	DECIMAL	NO

More than just tables

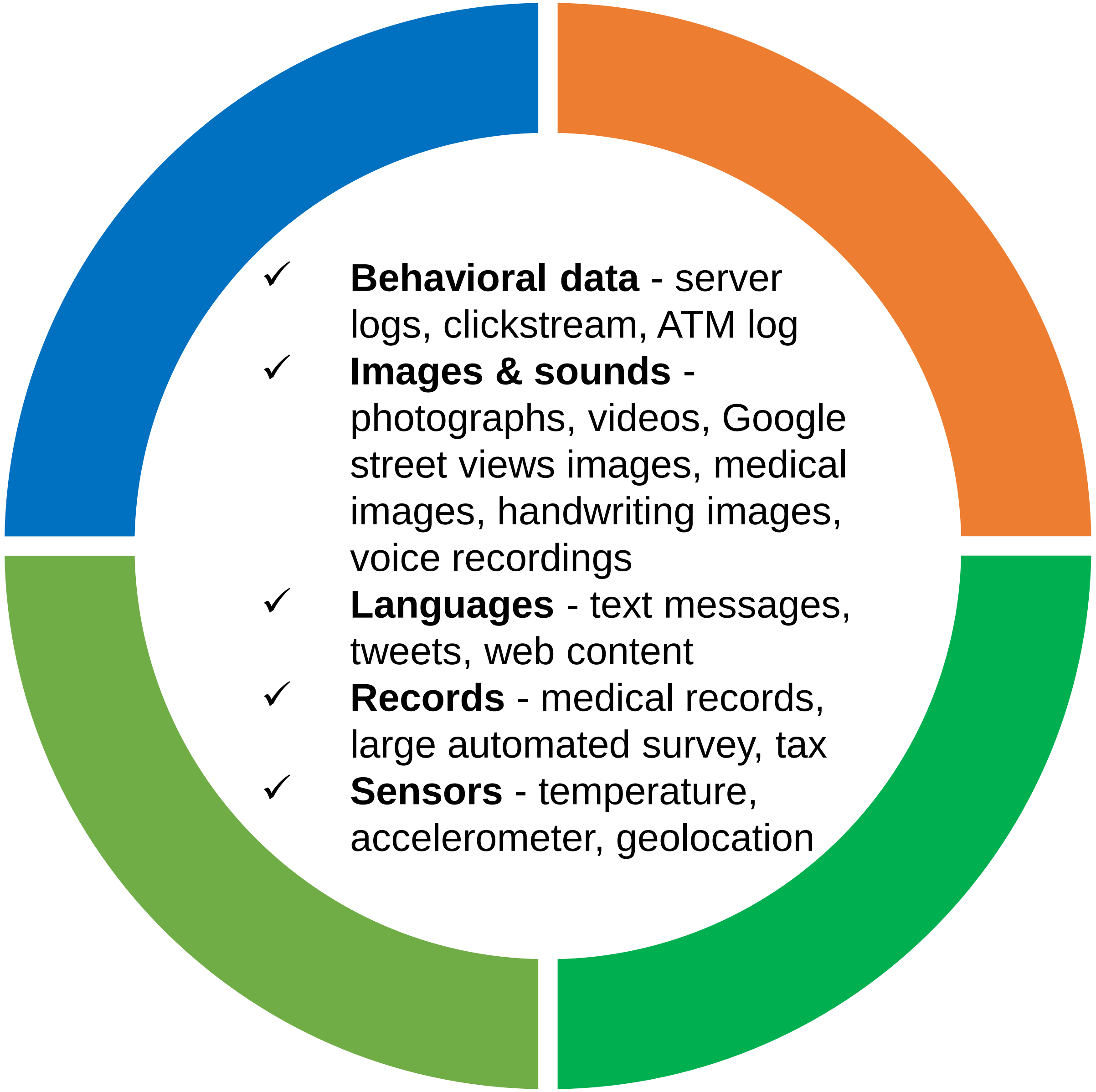


VOLUME

Online & Offline
Automatically generated
Manually created

VELOCITY

Speed of Generation
Rate of Analysis



- ✓ **Behavioral data** - server logs, clickstream, ATM log
- ✓ **Images & sounds** - photographs, videos, Google street views images, medical images, handwriting images, voice recordings
- ✓ **Languages** - text messages, tweets, web content
- ✓ **Records** - medical records, large automated survey, tax
- ✓ **Sensors** - temperature, accelerometer, geolocation

VARIETY

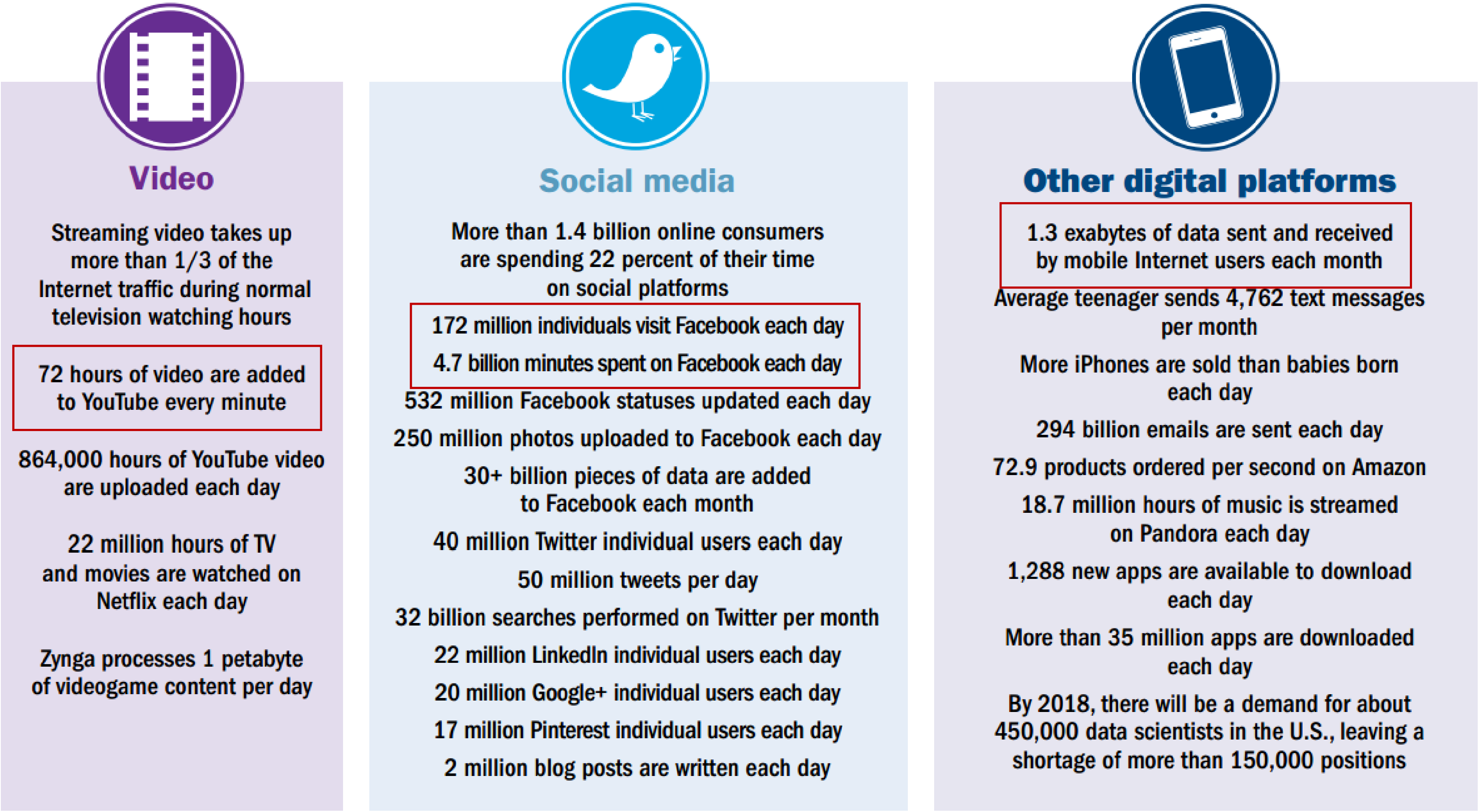
Unstructured
Structured

VERACITY

Untrusted
Uncleansed
Unclear

Big !

Data from 2014



Video

Streaming video takes up more than 1/3 of the Internet traffic during normal television watching hours

72 hours of video are added to YouTube every minute

864,000 hours of YouTube video are uploaded each day

22 million hours of TV and movies are watched on Netflix each day

Zynga processes 1 petabyte of videogame content per day

Social media

More than 1.4 billion online consumers are spending 22 percent of their time on social platforms

172 million individuals visit Facebook each day

4.7 billion minutes spent on Facebook each day

532 million Facebook statuses updated each day

250 million photos uploaded to Facebook each day

30+ billion pieces of data are added to Facebook each month

40 million Twitter individual users each day

50 million tweets per day

32 billion searches performed on Twitter per month

22 million LinkedIn individual users each day

20 million Google+ individual users each day

17 million Pinterest individual users each day

2 million blog posts are written each day

Other digital platforms

1.3 exabytes of data sent and received by mobile Internet users each month

Average teenager sends 4,762 text messages per month

More iPhones are sold than babies born each day

294 billion emails are sent each day

72.9 products ordered per second on Amazon

18.7 million hours of music is streamed on Pandora each day

1,288 new apps are available to download each day

More than 35 million apps are downloaded each day

By 2018, there will be a demand for about 450,000 data scientists in the U.S., leaving a shortage of more than 150,000 positions

Ref: Mushroom Networks, Deep Blue Analytics, MBAOnline, IBM, Gartner

Big Data Adoption Goal

Improves operational efficiency and drive productivity

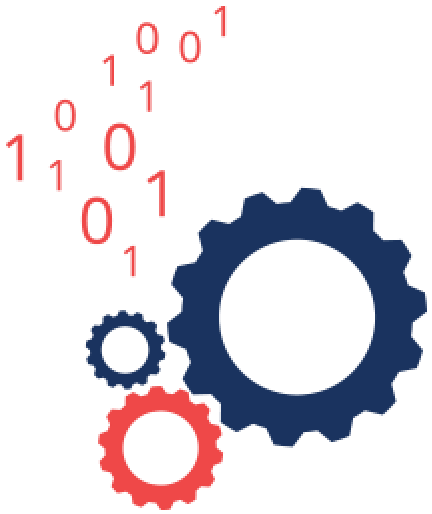
Improves profit through cost reduction

Business sustainability through customer satisfaction

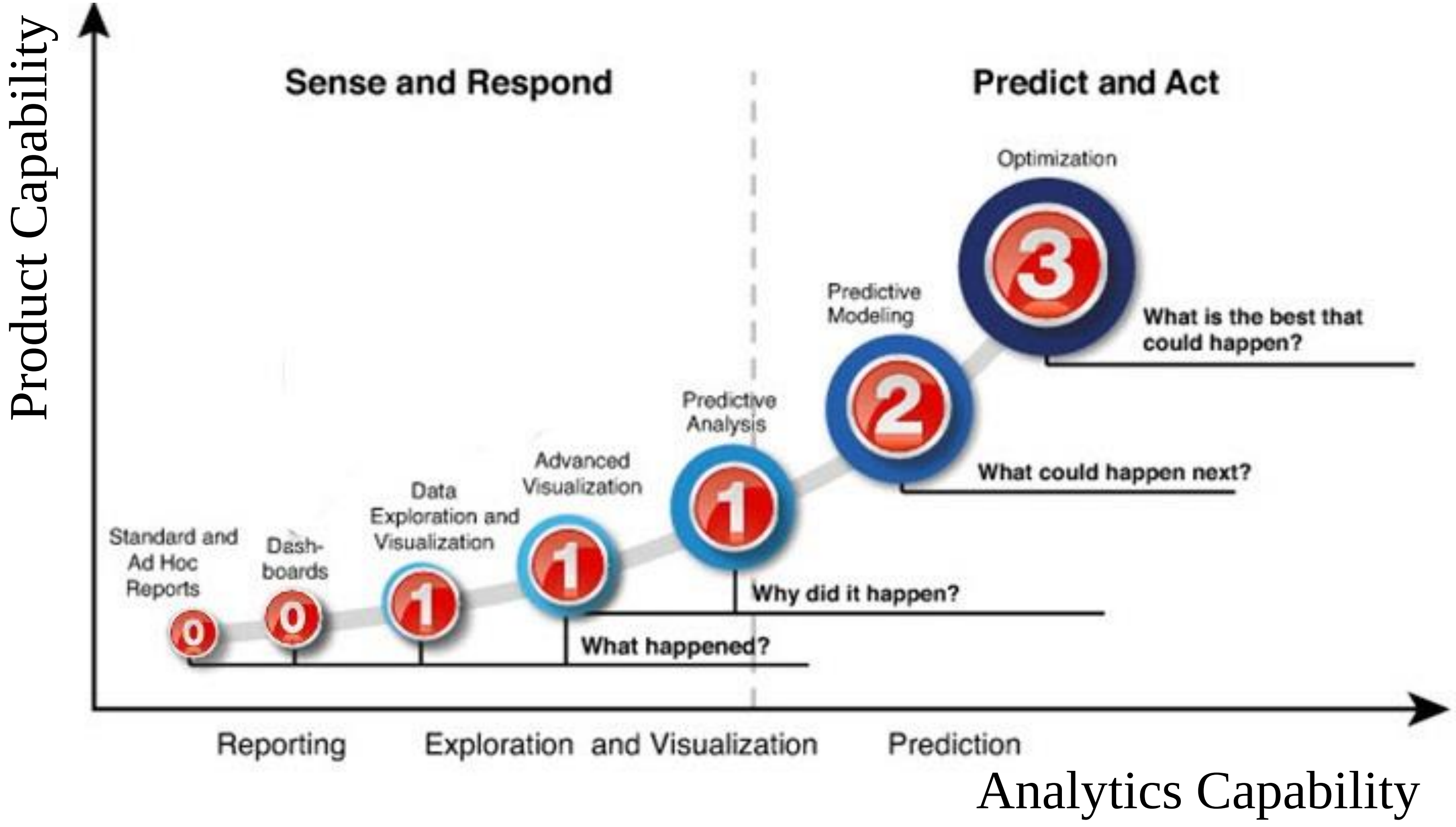
Creates new revenue sources

Big Data Analytics

A set of fundamental concepts/principles that underlie techniques for extracting useful knowledge from large datasets containing a variety of data types. To uncover hidden patterns, unknown correlations, market trends, customer preferences, and other useful business information

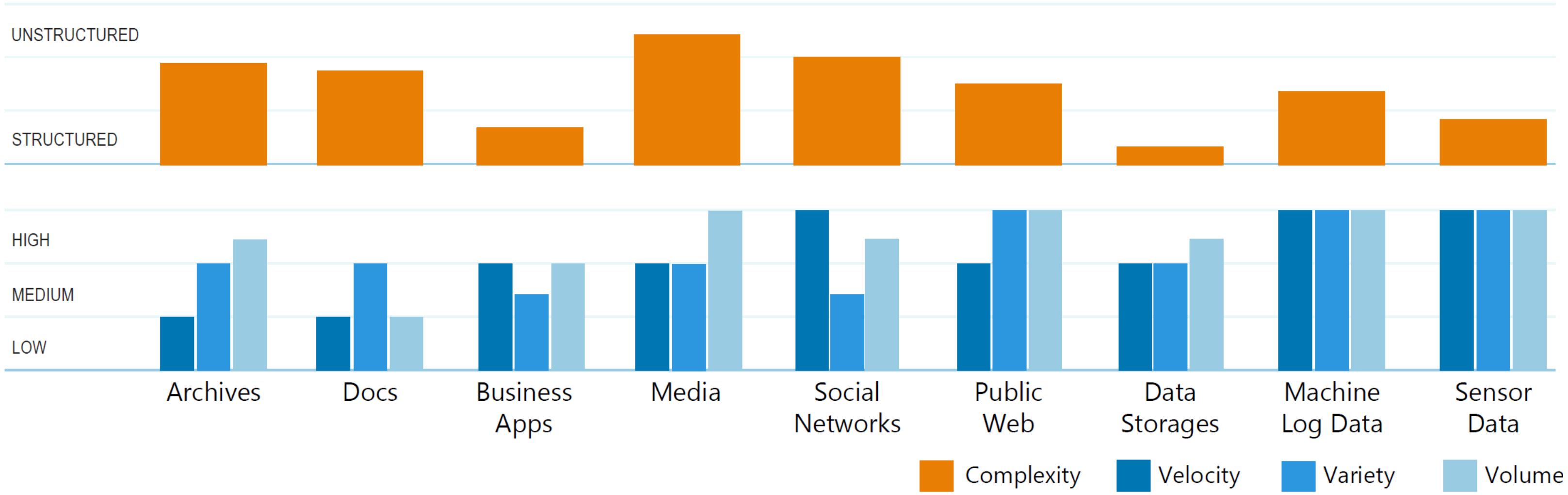


Big Data Maturity



http://fisherdaniel.com/

Big data challenge



Archives
Scanned documents, statements, medical records, e-mails etc..

Docs
XLS, PDF, CSV, HTML, JSON etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Media
Images, video, audio etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

Public Web
Wikipedia, news, weather, public finance etc

Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

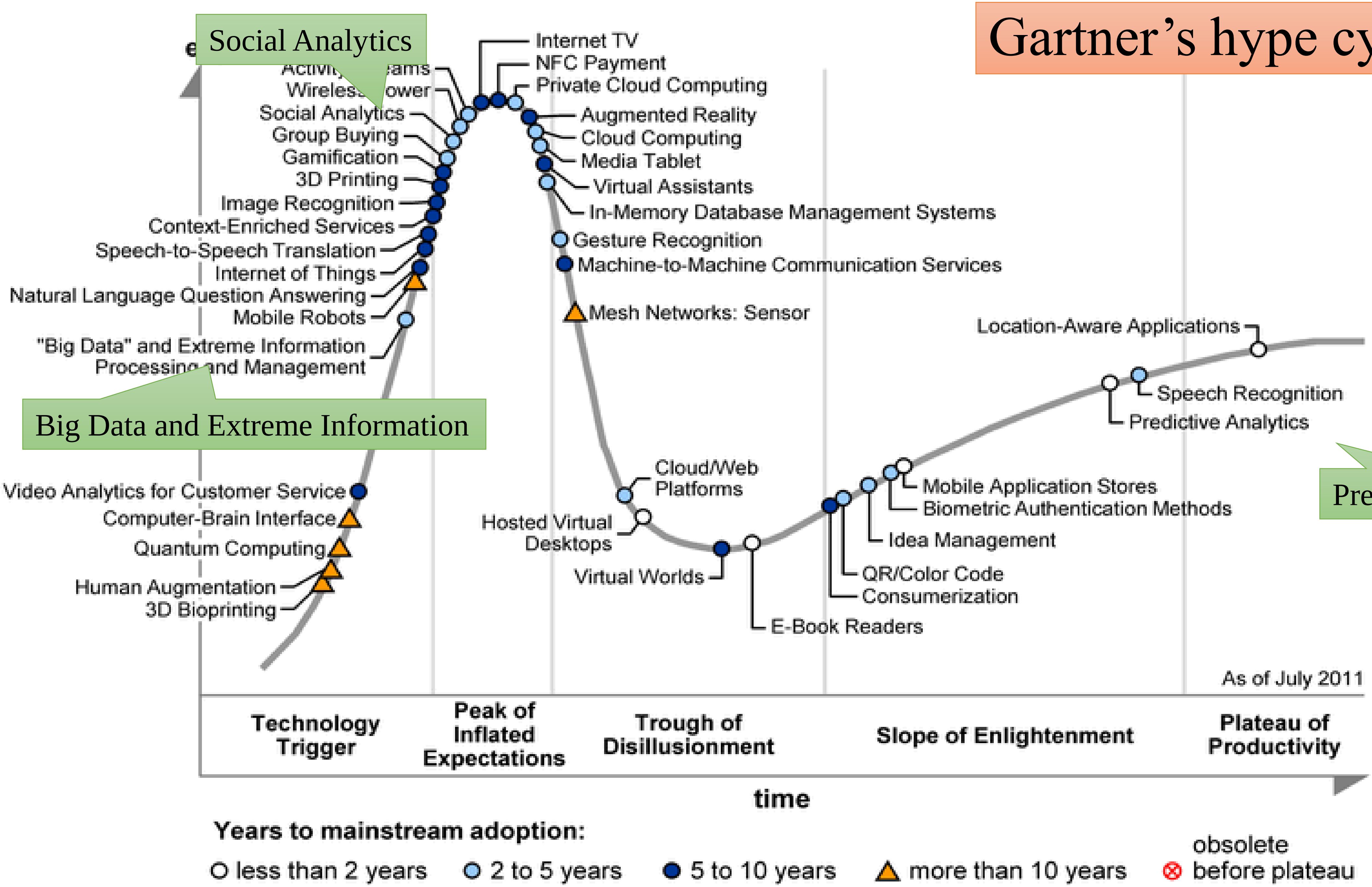
Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

Big data analytics

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

Gartner’s hype cycle 2011

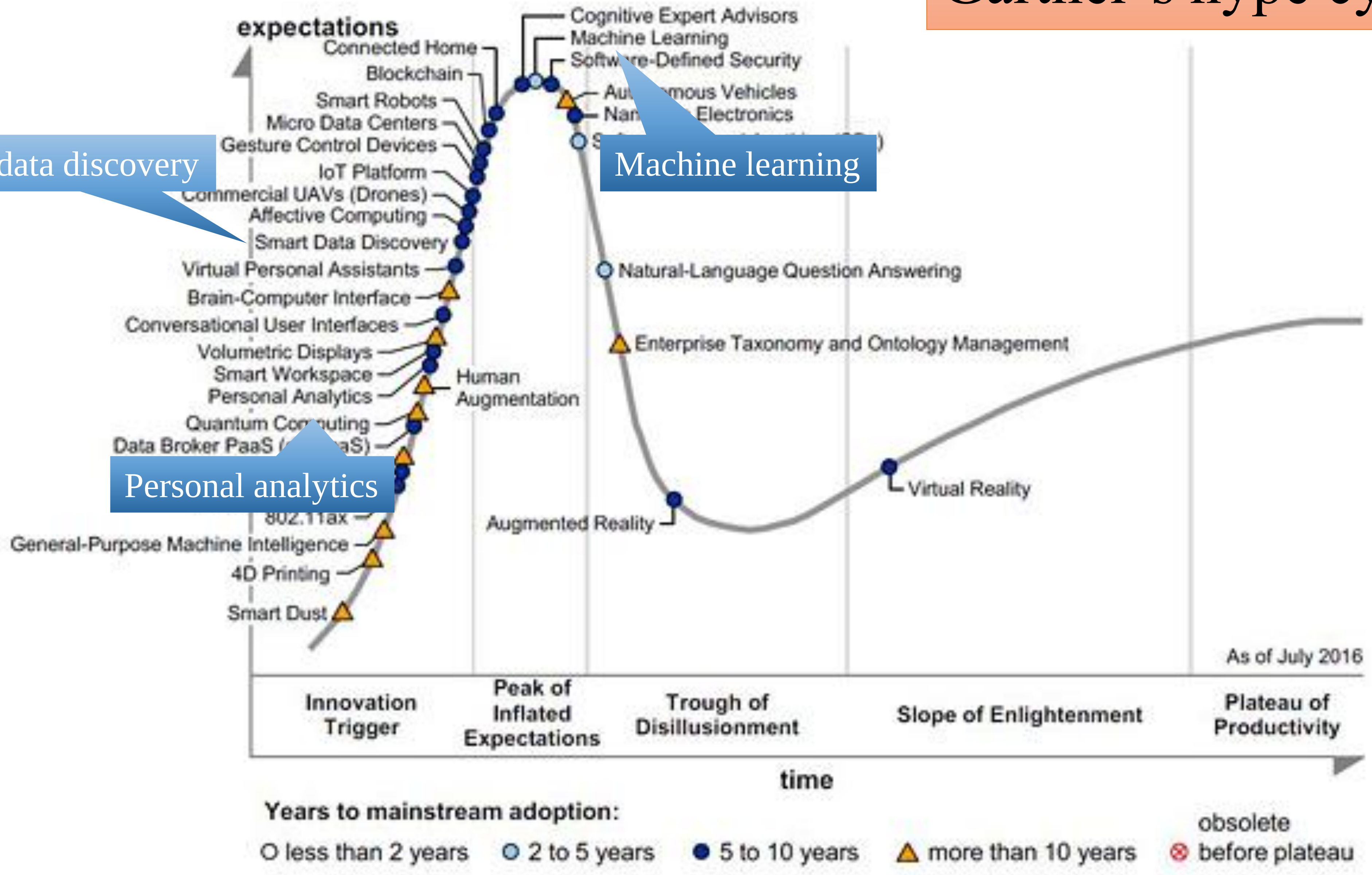


Social Analytics

Big Data and Extreme Information

Predictive Analytics

Gartner's hype cycle 2016



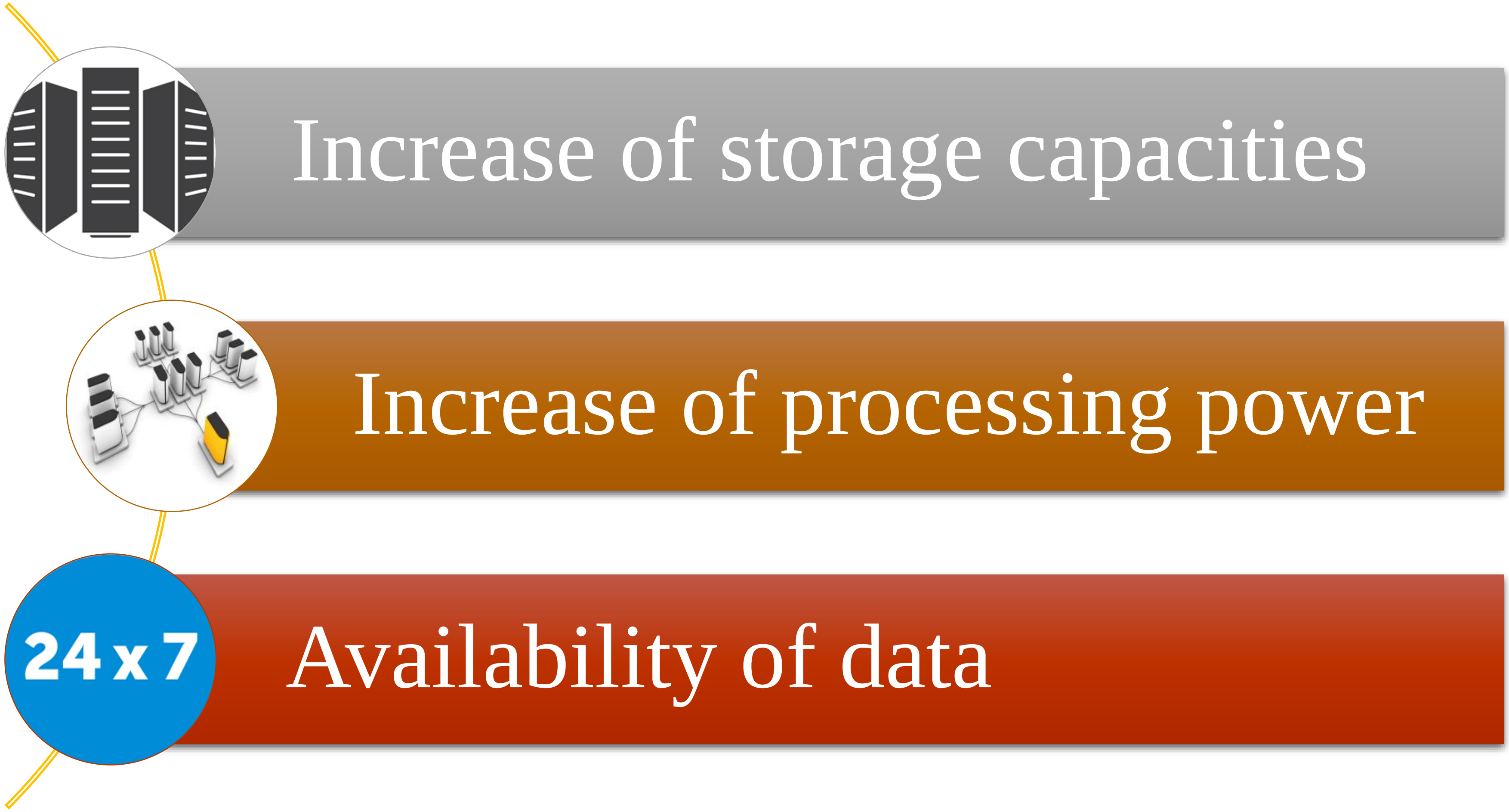
Smart data discovery

Machine learning

Personal analytics

Source: Gartner (July 2016)

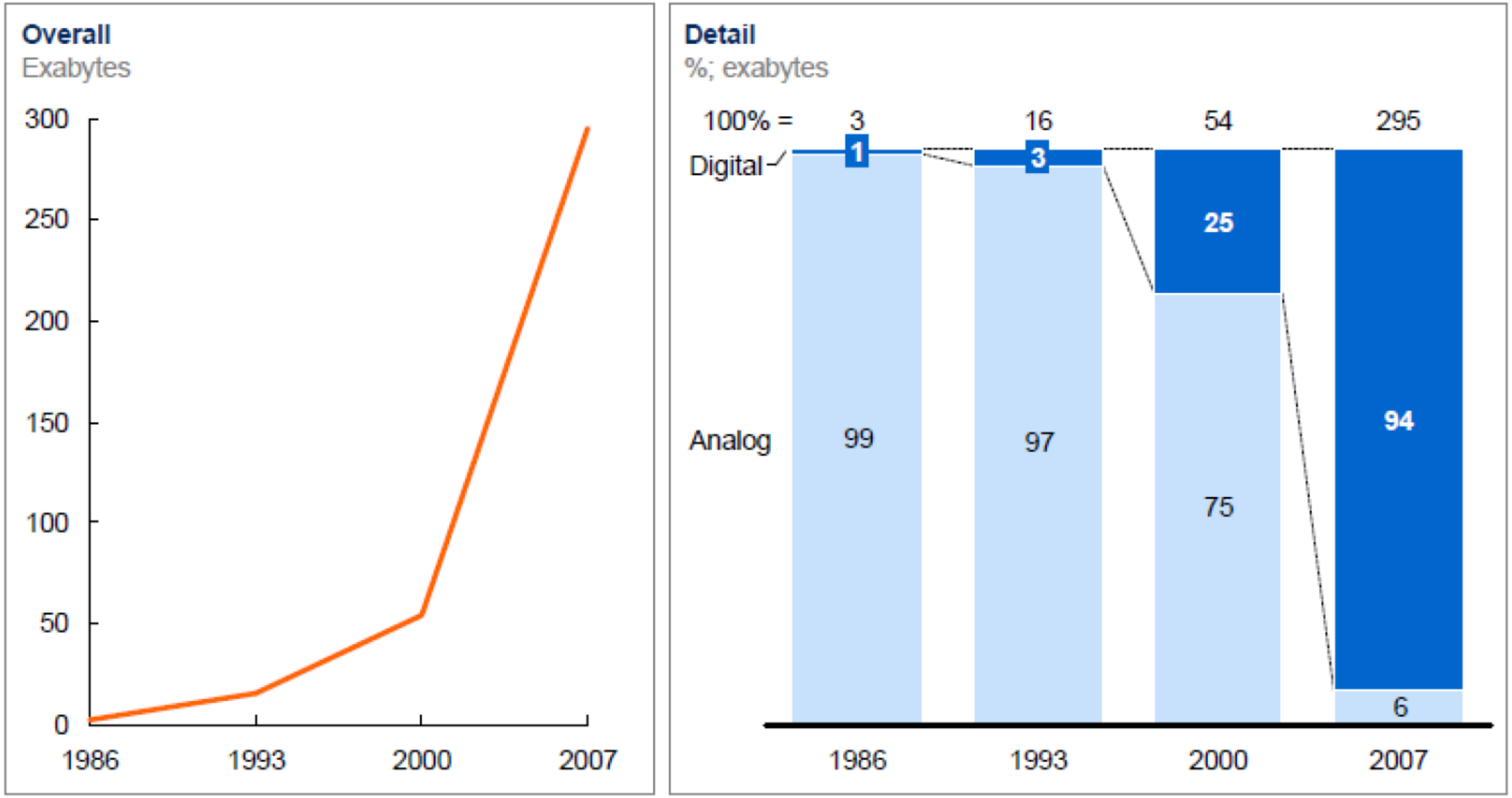
Why big data?



- Increase of storage capacities
- Increase of processing power
- Availability of data

Enabler: data storage

Global data storage has grown significantly to digital after 2000

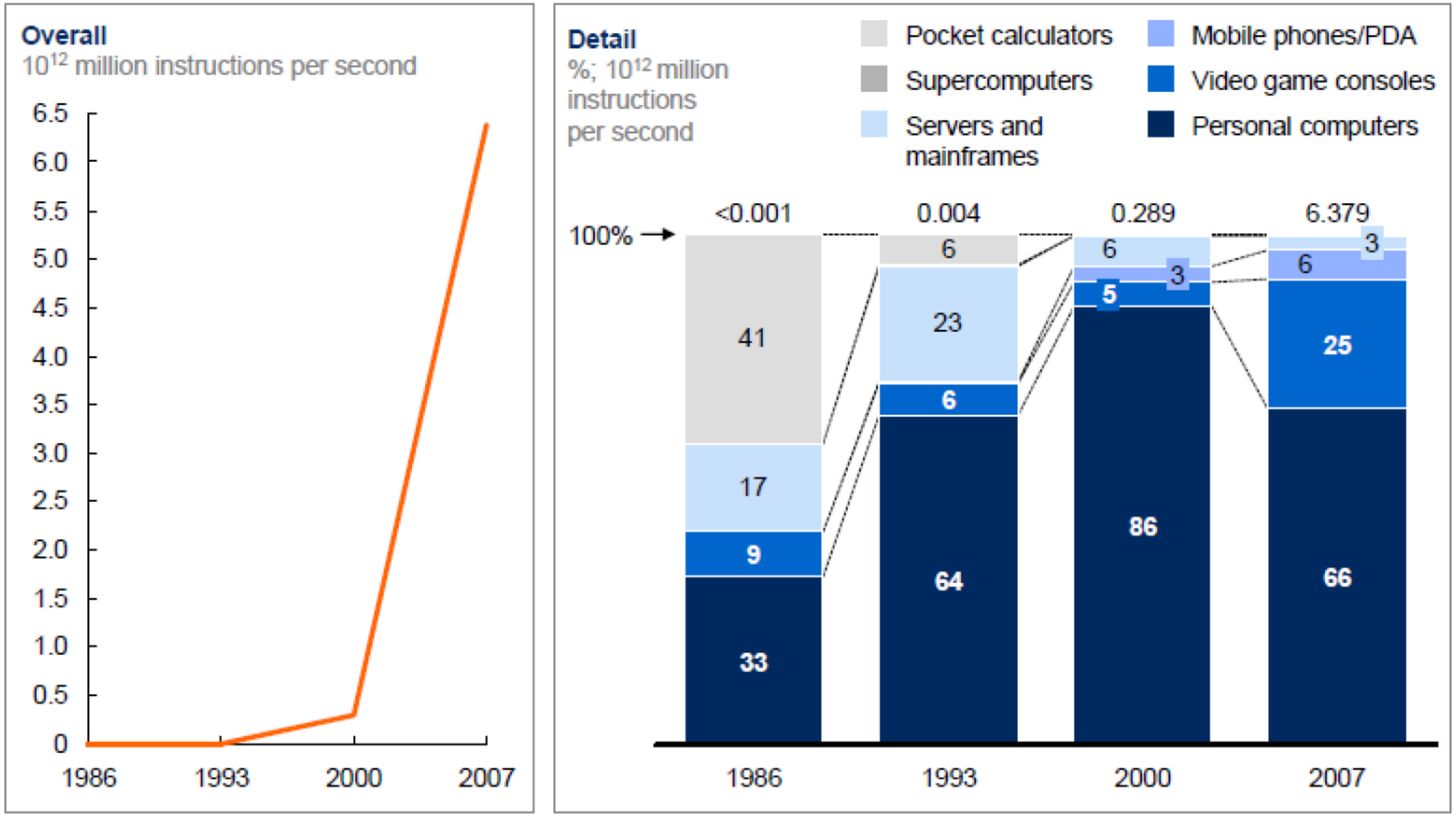


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Computation capacity

Computation capacity has also risen sharply

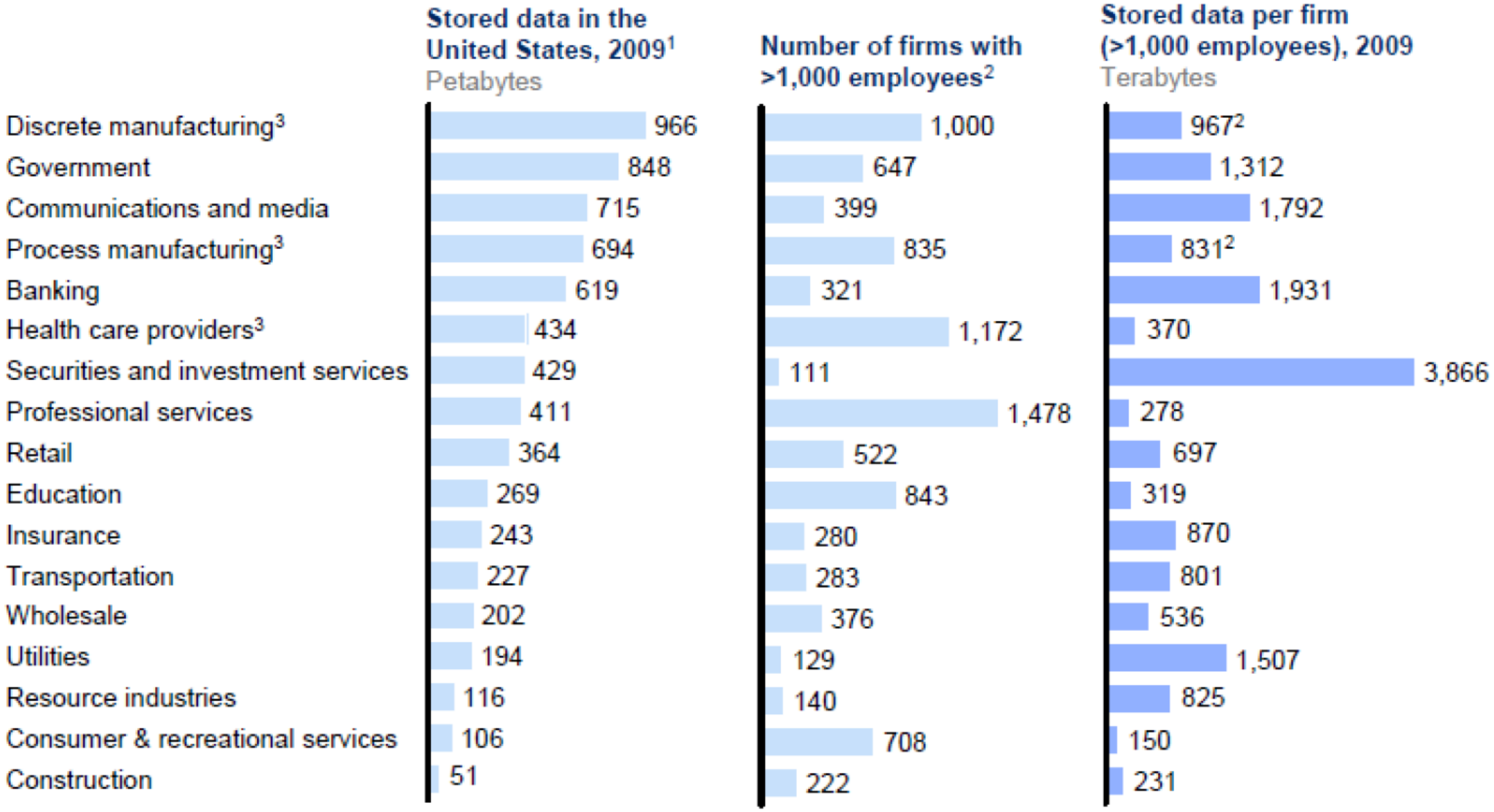


NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

Enabler: Data availability

Companies in all sectors have at least 100 TB of stored data in US



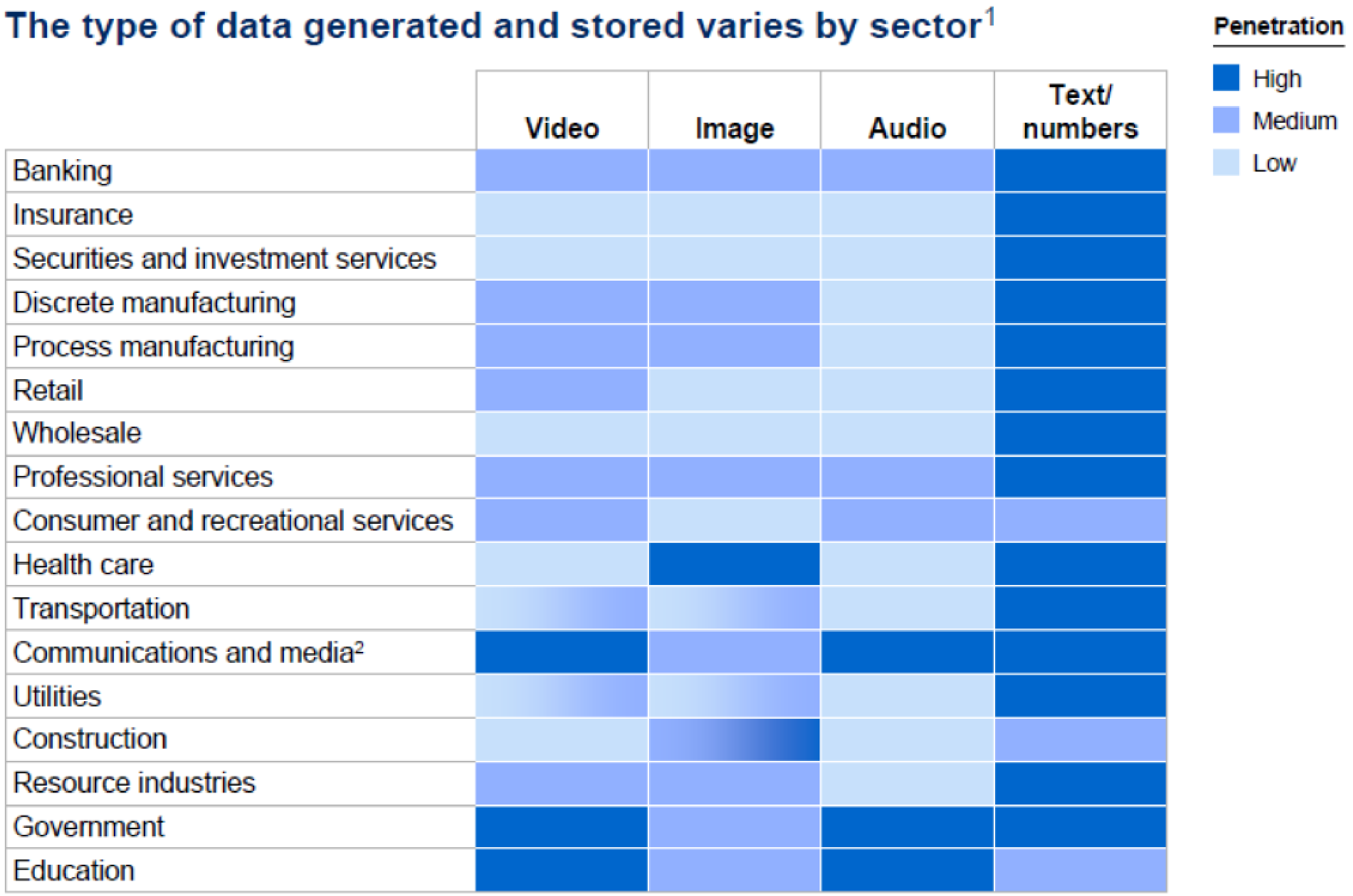
[1] Storage data by sector derived from IDC.

[2] Firm data split into sectors, when needed, using employment

[3] The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

Type of available data



The type of data generated and stored varies by sector[1]

Penetration

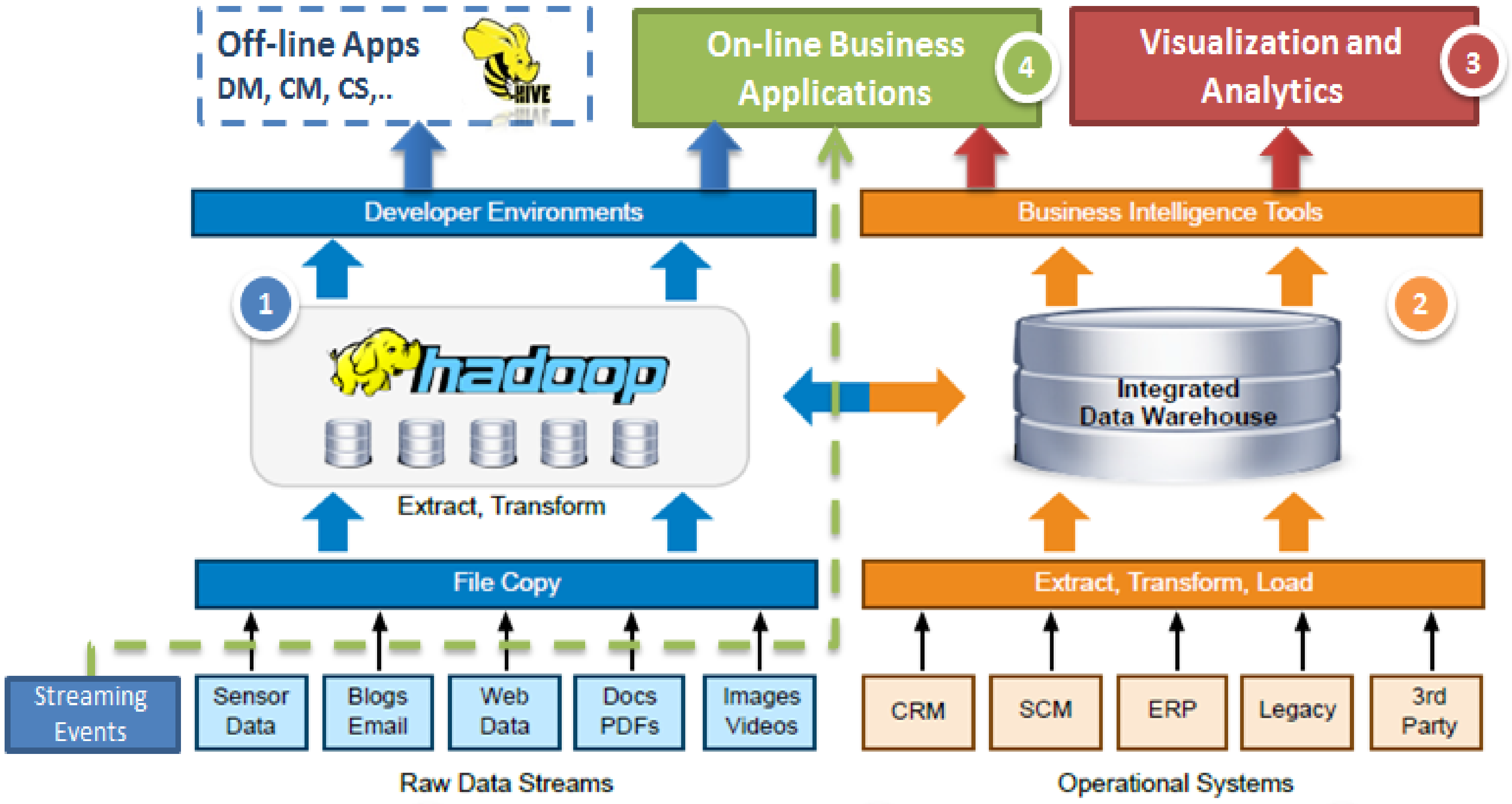
- ■ High
- ▨ Medium
- □ Low

	Video	Image	Audio	Text/ numbers
Banking	▨	▨	▨	■
Insurance	□	□	□	■
Securities and investment services	□	□	□	■
Discrete manufacturing	▨	▨	□	■
Process manufacturing	▨	▨	□	■
Retail	▨	□	□	■
Wholesale	□	□	□	■
Professional services	▨	▨	▨	■
Consumer and recreational services	▨	□	▨	▨
Health care	□	■	□	■
Transportation	□ ▨	□ ▨	□	■
Communications and media[2]	■	▨	■	■
Utilities	□ ▨	□ ▨	□	■
Construction	□	▨ ■	□	▨
Resource industries	▨	▨	□	■
Government	■	▨	■	■
Education	■	▨	■	▨

[1] We compiled this heat map using units of data (in files or minutes of video) rather than bytes.

[2] Video and audio are high in some subsectors.

SOURCE: McKinsey Global Institute analysis



1 Data Lake - Raw Data Storage & processing

- Handles **structured and unstructured** data
- Hadoop-based
- Map-reduce algorithms

2 Data warehouse

- Handles only structured data
- **MPP** based (massively parallel processing)
- **Column based**
- **Cloud (Google/Redshift)**

3 Visualization and Analytics

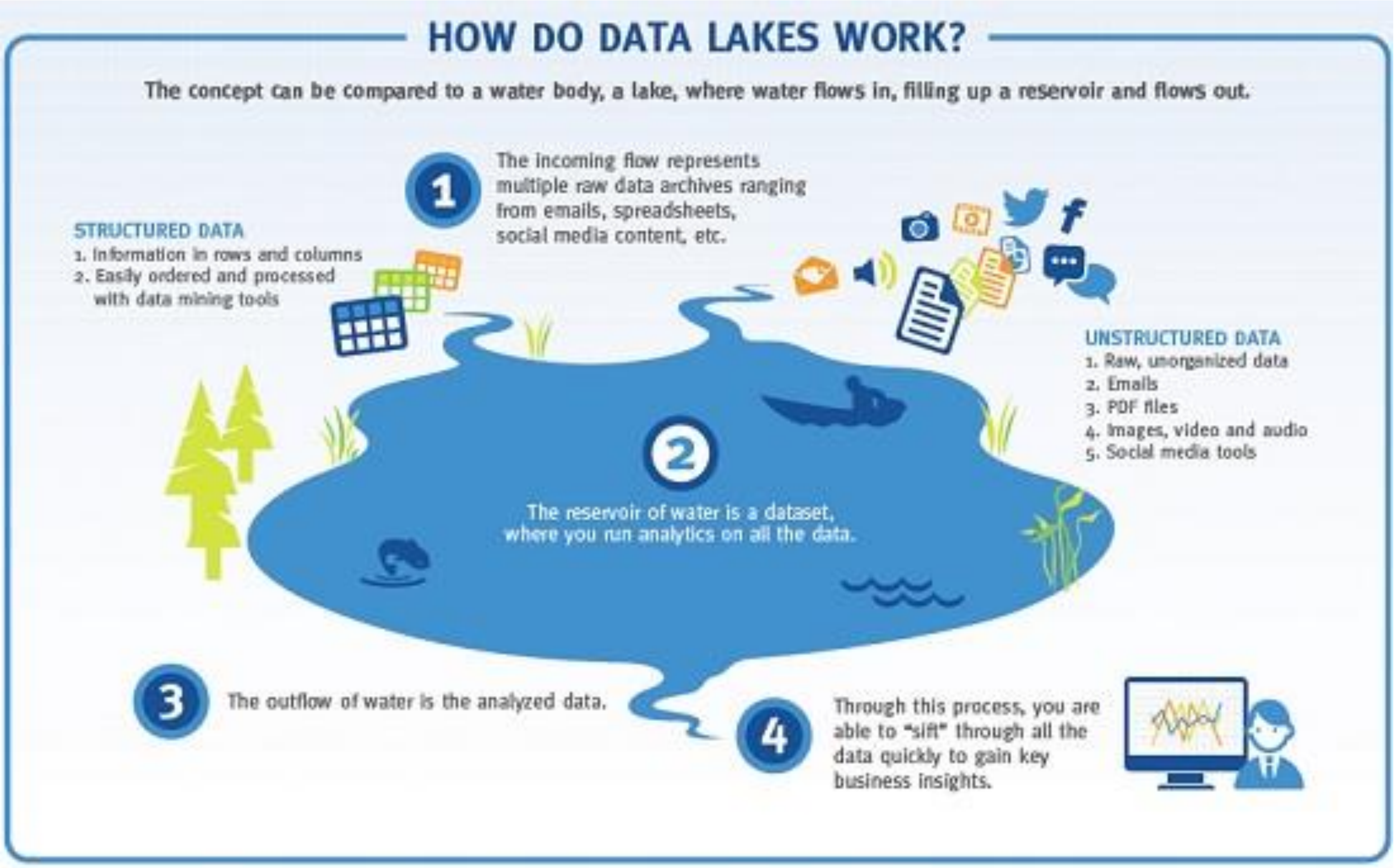
- Handles structured data
- Supports visualization and reporting in “**exploratory**” mode

4 On-line Business Apps

- RT Applications
- Recommendations engines
- Machine learning
- No-SQL solutions (Cassandra, Riak, MongoDB, Hbase,...)

BIG DATA
EXPERIENCE

Data Warehouse	Vs.	Data Lake
Structured, processed	Data	structured / semi-structured / unstructured, raw
Schema-on-write	Processing	Schema-on-read
Expensive for large data volumes	Storage	Designed for low-cost storage
Less agile, fixed configuration	Agility	Highly agile, configure and reconfigure as needed
Mature	Security	Maturing
Business Professionals	Users	Data Scientist et. Al.



HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

1 The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools

UNSTRUCTURED DATA

1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools

2 The reservoir of water is a dataset, where you run analytics on all the data.

3 The outflow of water is the analyzed data.

4 Through this process, you are able to "sift" through all the data quickly to gain key business insights.

https://www.linkedin.com/pulse/building-data-lake-using-open-source-technologies-aneel

Big data in public transportation

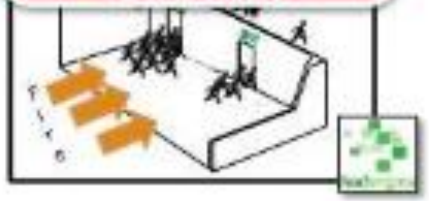
AFC Tap-in & tap-out

Event ticketing

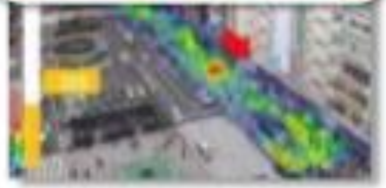
Crowd Counting

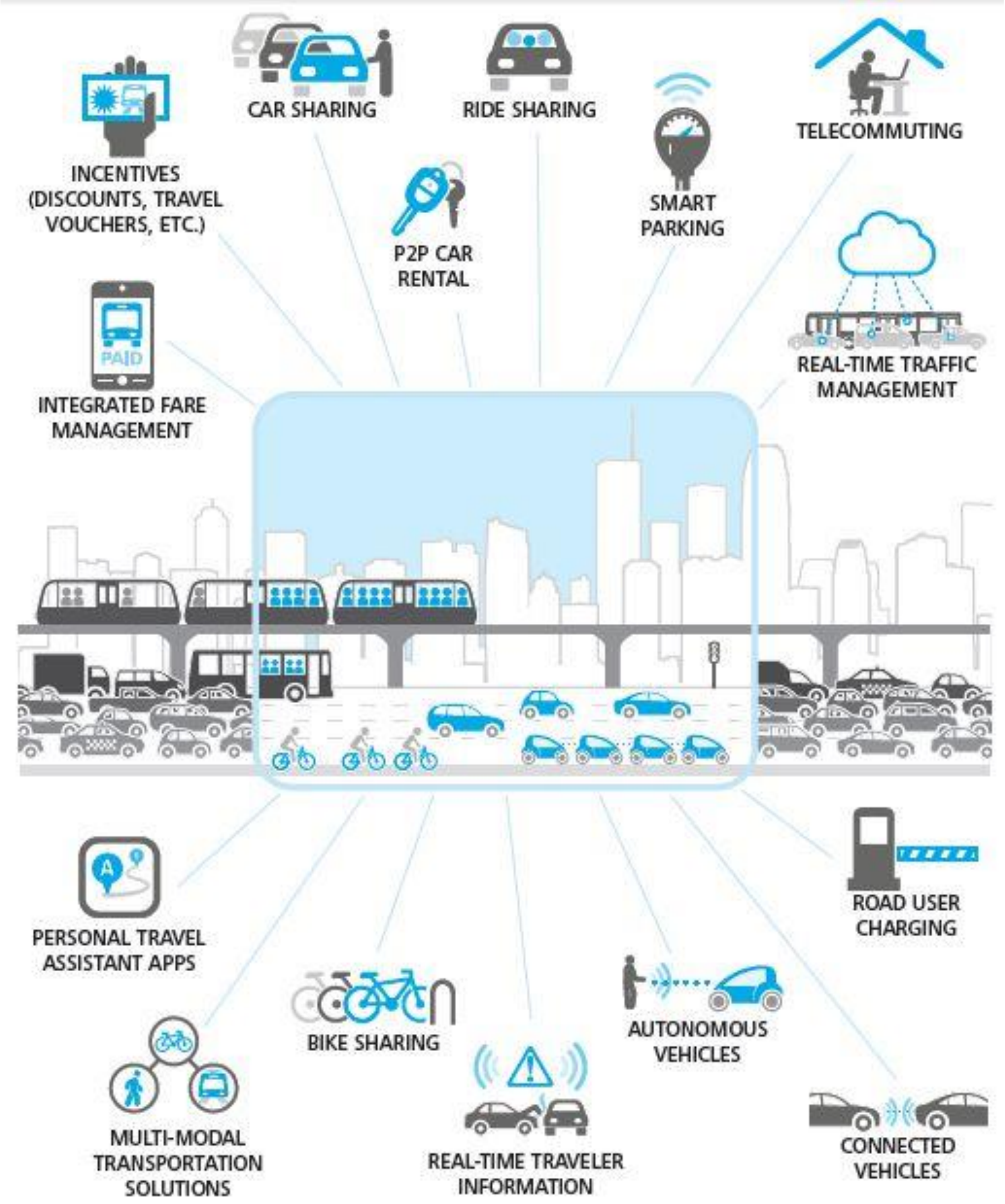
Transport App

Sensor-based Crowd Detection



Video-based Crowd Behavior Analysis



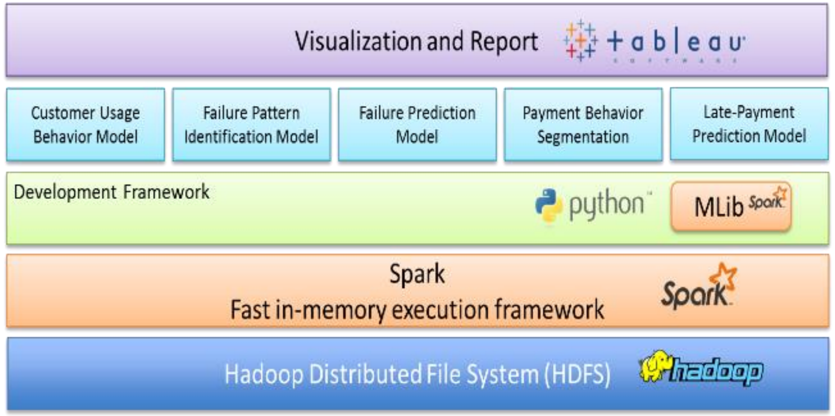

INCENTIVES (DISCOUNTS, TRAVEL VOUCHERS, ETC.)
CAR SHARING
RIDE SHARING
TELECOMMUTING
P2P CAR RENTAL
SMART PARKING
INTEGRATED FARE MANAGEMENT
PAID
REAL-TIME TRAFFIC MANAGEMENT
PERSONAL TRAVEL ASSISTANT APPS
ROAD USER CHARGING
BIKE SHARING
AUTONOMOUS VEHICLES
MULTI-MODAL TRANSPORTATION SOLUTIONS
REAL-TIME TRAVELER INFORMATION
CONNECTED VEHICLES

BIG DATA
EXPERIENCE

การไฟฟ้านครหลวง
Metropolitan Electricity Authority

VOICE OF CUSTOMERS

To understand customers and improve quality of services



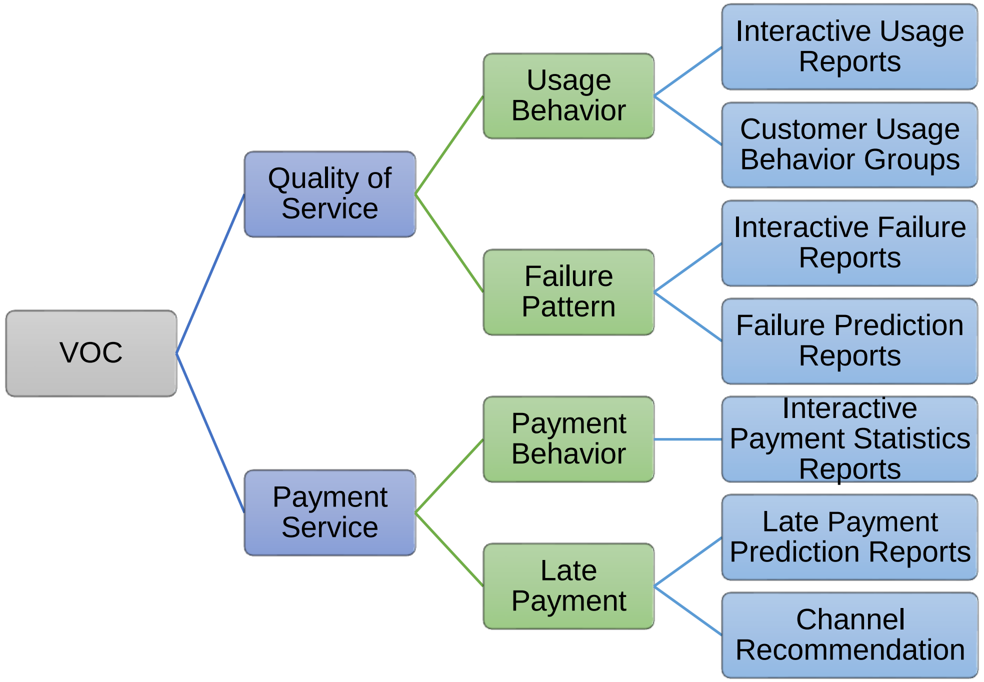
Visualization and Report tableau

Customer Usage Behavior Model | Failure Pattern Identification Model | Failure Prediction Model | Payment Behavior Segmentation | Late-Payment Prediction Model

Development Framework python MLib Spark

Spark
Fast in-memory execution framework Spark

Hadoop Distributed File System (HDFS) hadoop



- VOC
  - Quality of Service
    - Usage Behavior
      - Interactive Usage Reports
      - Customer Usage Behavior Groups
    - Failure Pattern
      - Interactive Failure Reports
      - Failure Prediction Reports
  - Payment Service
    - Payment Behavior
      - Interactive Payment Statistics Reports
    - Late Payment
      - Late Payment Prediction Reports
      - Channel Recommendation

KMUTT g-able

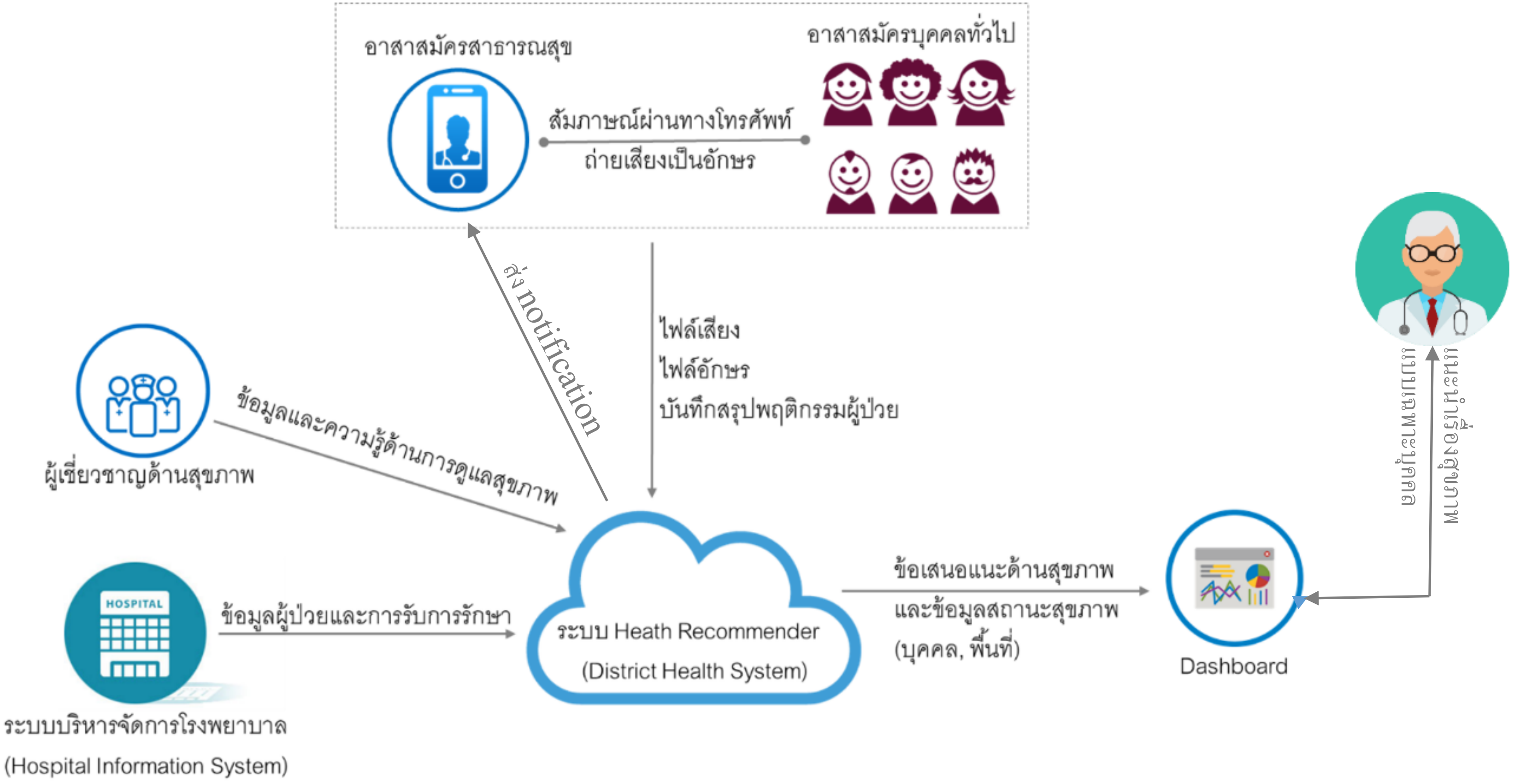
Health Predictive Modeling

- Databases of information about the state of the health of the general public can be built.
  - Genetic factors (Patient records)
  - Life style (social media, etc.)
  - Wearable sensor data
  - medical and insurance records
- Person’s data can be compared and analyzed alongside thousands of others
  - Highlight specific threats and issues through patterns that emerge during the comparison
  - Enables sophisticated predictive modelling to take place

Ref: http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/

HEALTH RECOMMENDER SYSTEM

In Collaboration with Sodsri-Saritwong Foundation



DIALYSIS SOLUTION DELIVERY

In Collaboration with NHSO

- Goal: Optimize dialysis solution order and delivery
- Methods
  - Treatment data exploration (cleaning, filtering, selection) – 300 M Records, 61 M people
  - Predictive model creation (predict time period dialysis state for each patients in the system)
  - Dialysis solution amount estimation
  - Delivery schedule optimization

GENETIC ANALYSIS AUTOMATION

Create software that can automate the genetic analysis process
Convolutional Neural Network (Deep Learning) is applied
A new startup is underway

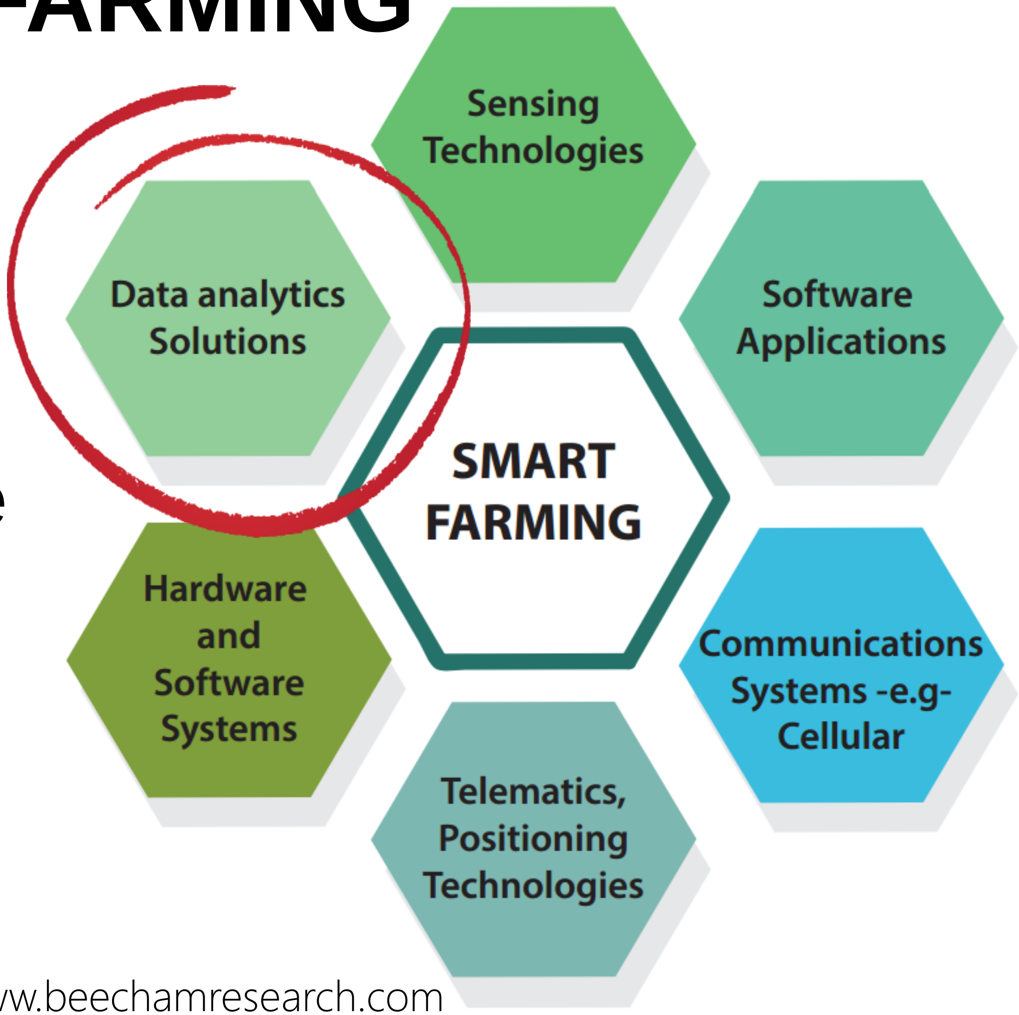
ศูนย์พันธุศาสตร์การแพทย์

ให้บริการตรวจวิเคราะห์ทางพันธุกรรม เพื่อค้นหาความผิดปกติในระดับโครโมโซม ไปจนถึงความผิดปกติในระดับดีเอ็นเอ
เพื่อให้แพทย์สามารถใช้ผลการตรวจวิเคราะห์ดังกล่าวในการดูแลผู้ป่วยและครอบครัวได้อย่างมีประสิทธิภาพ
ดำเนินงานโดยแพทย์เฉพาะทางด้านพันธุศาสตร์ และทีมนักวิทยาศาสตร์ผู้เชี่ยวชาญ

SMART FARMING

To optimize the yield / unit of farming land

We need to farm smarter utilizing the latest and the greatest technologies

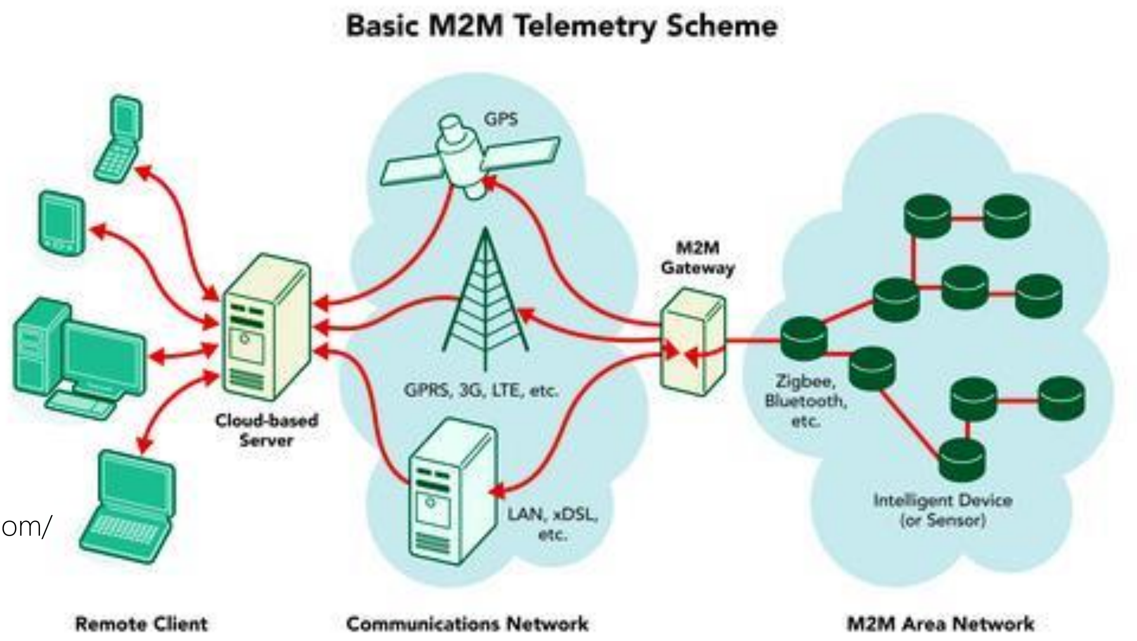

Sensing Technologies
Data analytics Solutions
Software Applications
SMART FARMING
Hardware and Software Systems
Communications Systems -e.g- Cellular
Telematics, Positioning Technologies

www.beechamresearch.com

ICT-based decision support systems

TECHNOLOGY INTEGRATION

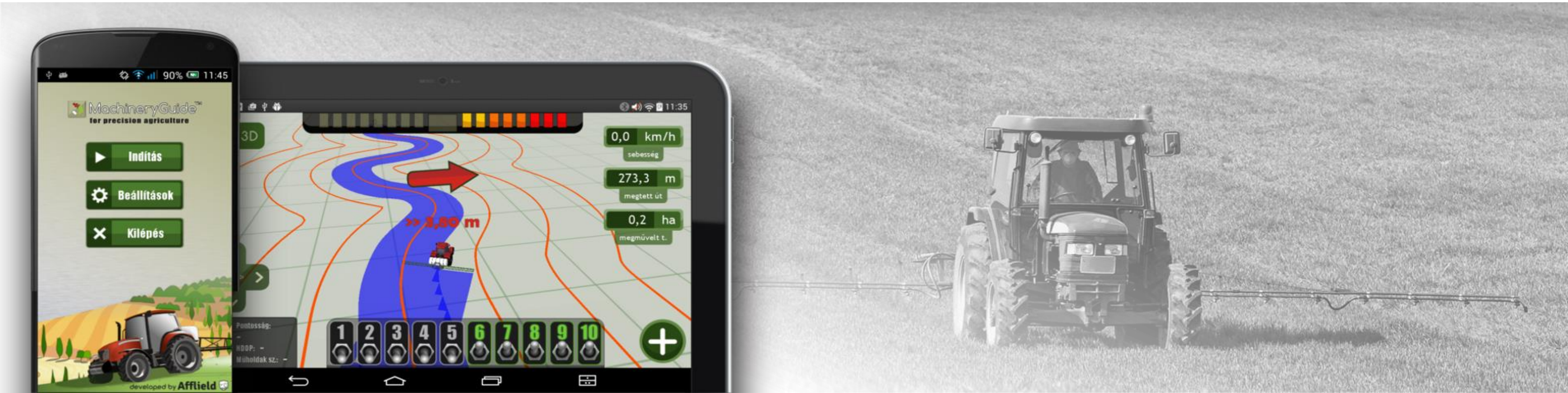
Machine-to-machine (M2M) telemetry plays an essential part in the Internet of Things revolution that is rapidly reshaping farming



http://www.designnews.com/

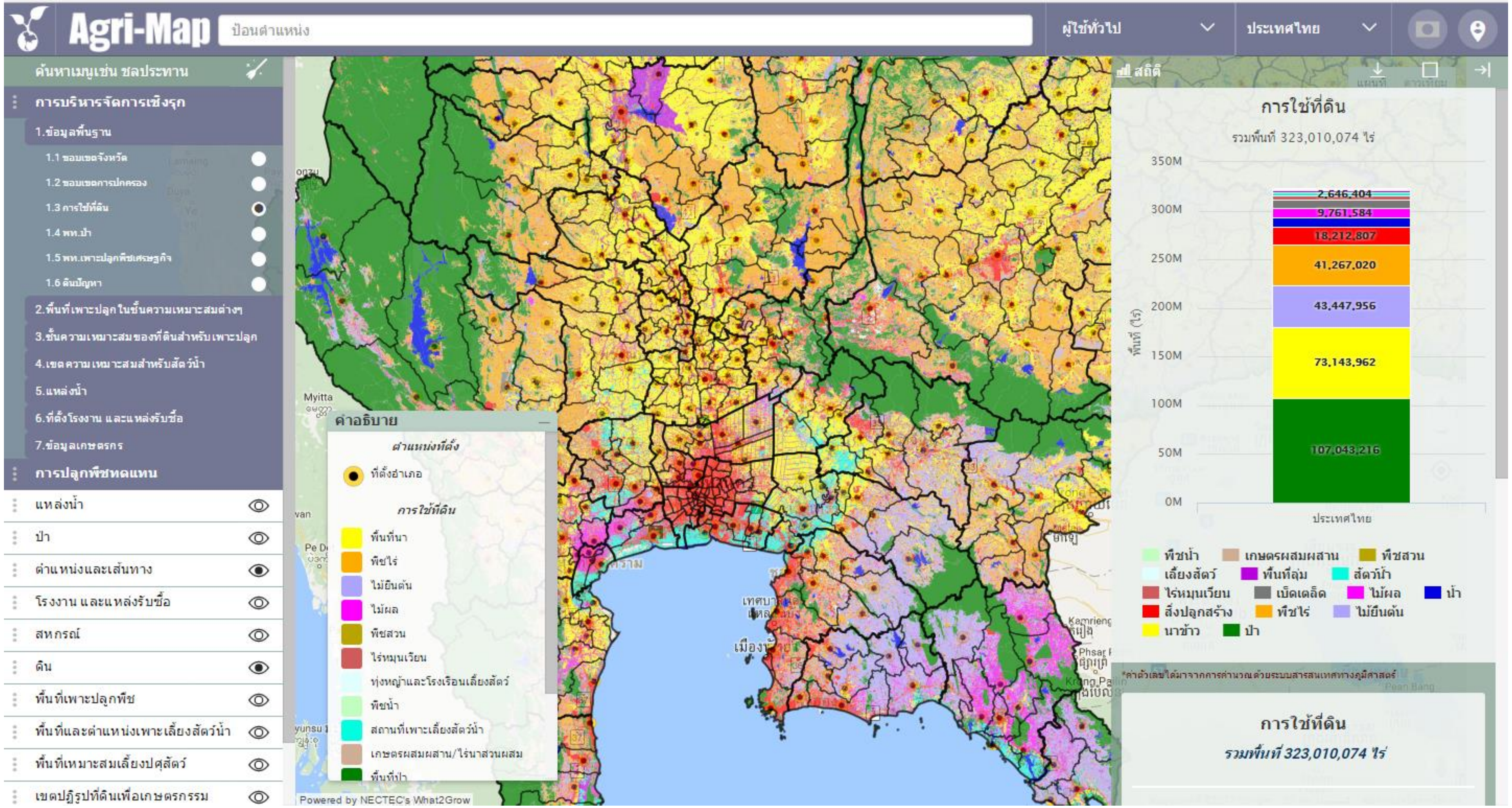
Precision Farming Data

- Optimize farming decisions in order to maximize yields.
- Farmers can make proactive decisions based on future conditions
  - when to plant, fertilize and harvest crops
- Adopt wireless, cloud-connected systems, and place sensors throughout the fields
  - Provide real-time monitoring: measure temperature and humidity of the soil and air
  - Take pictures of fields using satellite imagery and robotic drones. The images over time show crop maturity.
  - Predictive weather modeling show pinpoint conditions 24-48 hours in advance
  - Automate everyday agriculture operations
  - Provide data analysis for smart decision making (day-to-day, season-to-season)



WHAT2GROW

By NECTEC



*ค่าตัวเลขได้มาจากการคำนวณด้วยระบบสารสนเทศทางภูมิศาสตร์

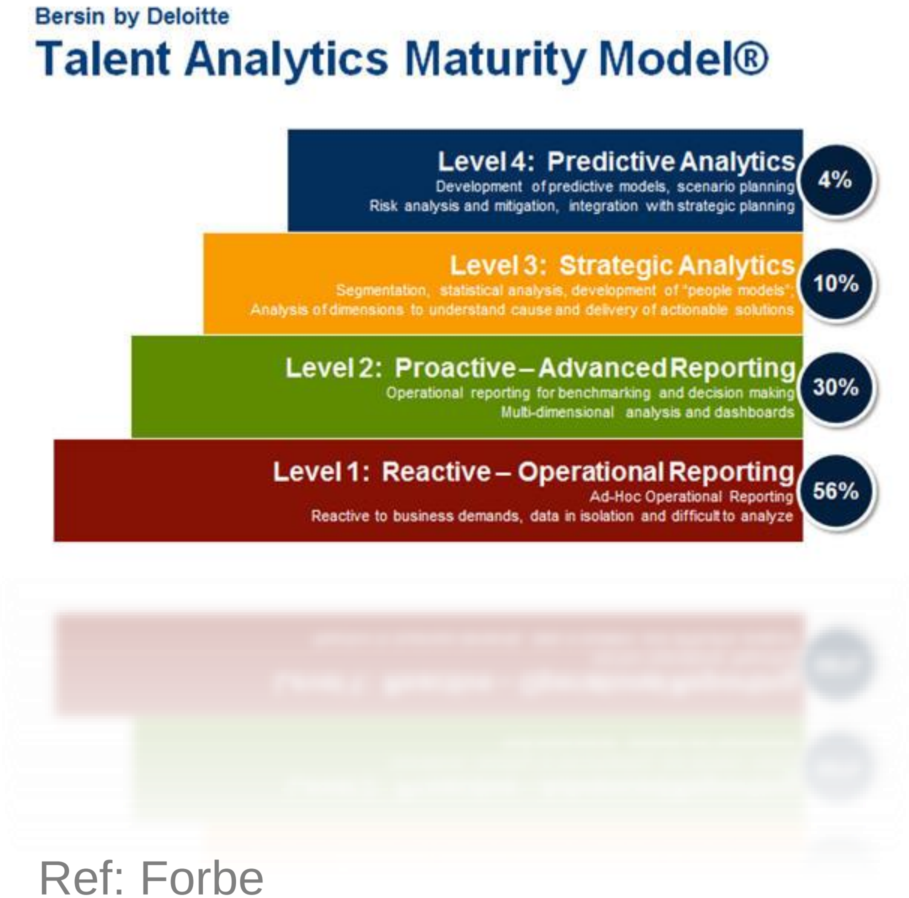
การใช้ที่ดิน

รวมพื้นที่ 323,010,074 ไร่

Powered by NECTEC's What2Grow

Big Data For HR

- Talent acquisition, retention, placement, promotion, compensation, or workforce and succession planning.
- Analyzing the skills and attributes of high performers in the present; build a template for future quality hiring.
- Non-traditional data gathering sources
  - Social media channels where prospective candidates usually leave their digital *'thought prints'.*
- Statistical analysis of productivity and turnover
  - Old indicators (such as GPA and education) were far less critical to performance and retention.



Bersin by Deloitte

Talent Analytics Maturity Model®

Level 4: Predictive Analytics
Development of predictive models, scenario planning
Risk analysis and mitigation, integration with strategic planning
4%

Level 3: Strategic Analytics
Segmentation, statistical analysis, development of "people models";
Analysis of dimensions to understand cause and delivery of actionable solutions
10%

Level 2: Proactive – Advanced Reporting
Operational reporting for benchmarking and decision making
Multi-dimensional analysis and dashboards
30%

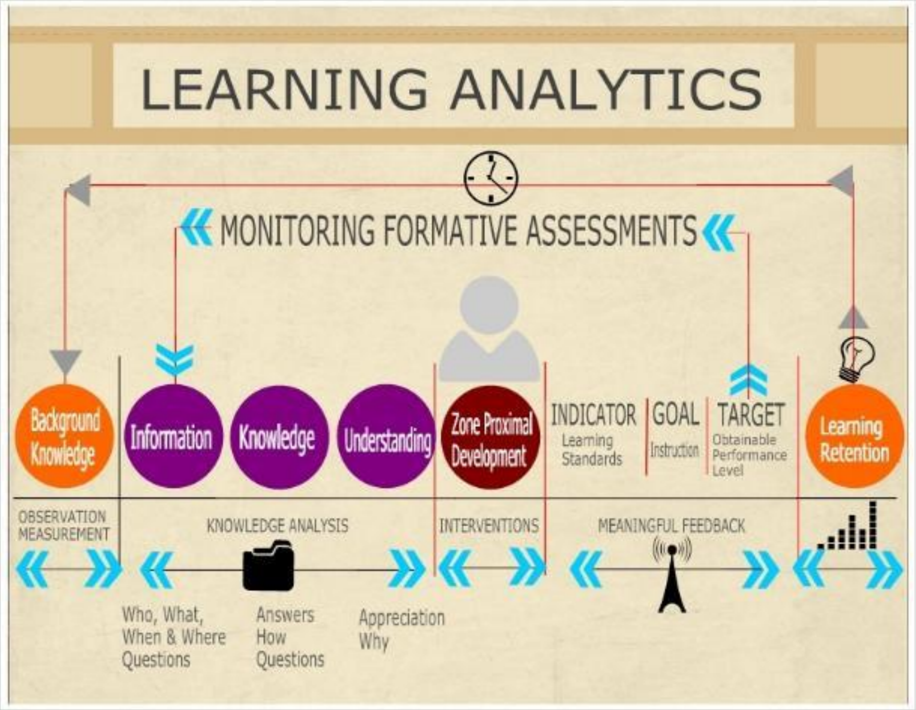
Level 1: Reactive – Operational Reporting
Ad-Hoc Operational Reporting
Reactive to business demands, data in isolation and difficult to analyze
56%

Ref: Forbe

Big Data and Learning

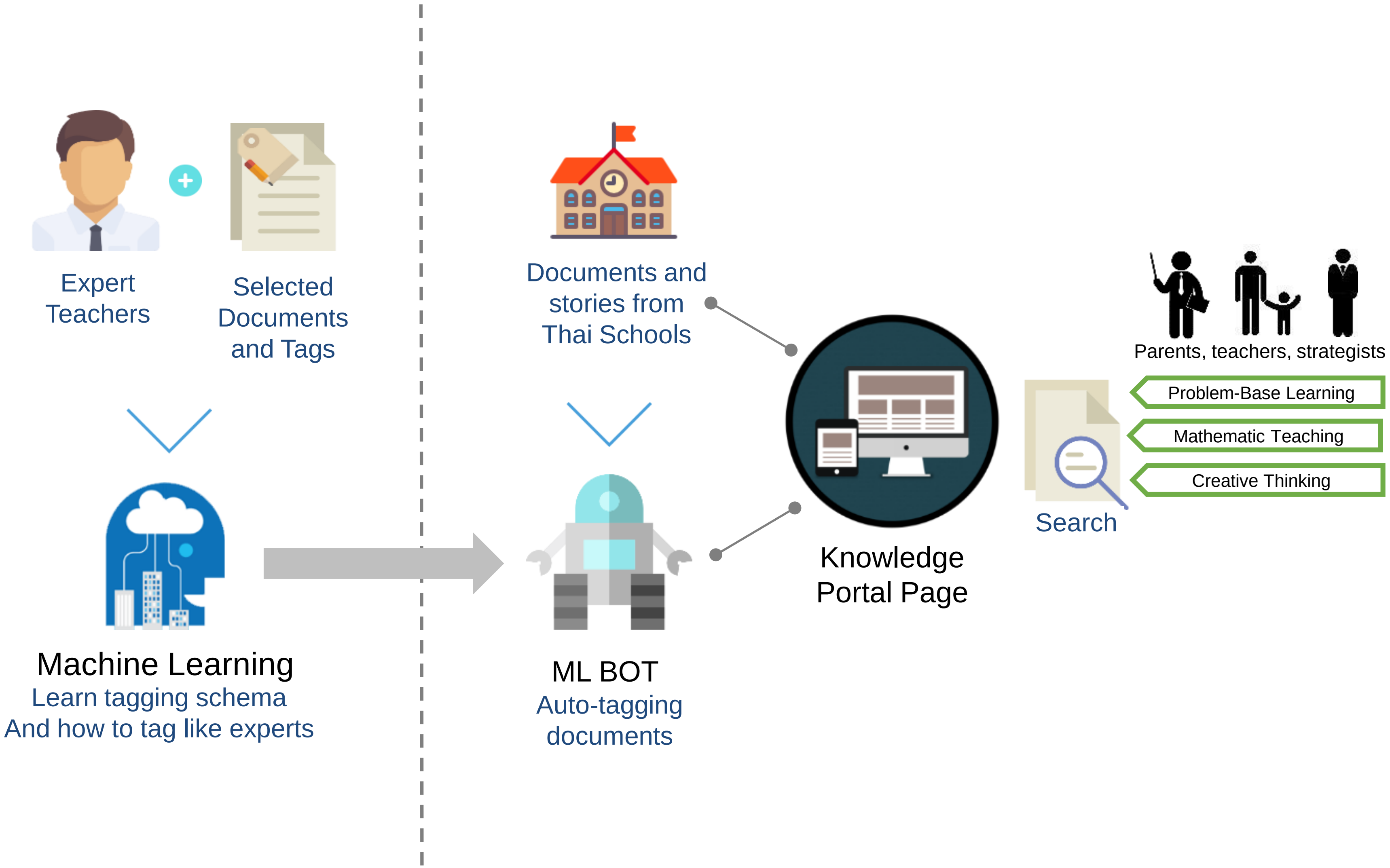
The measurement, collection, analysis and reporting of data about learners and their contexts.

- Focuses on applying techniques at larger scales in instructional systems.
  - ✓ Track what students know or does not know
  - ✓ Monitor student behaviors through level of engagement
  - ✓ Track individual student performance in each class through opinions and scores
  - ✓ Track course outcomes and student achievements
- Questions that can be answered:
  - ✓ When are students ready to move on to the next topic
  - ✓ When is a student at risk to not completing a course
  - ✓ What grade is a student likely to receive without intervention
  - ✓ Should a student be referred to a counselor for help



BUILD A PEDAGOGY PORTAL

In Collaboration with The Knowledge Management Institute



Expert Teachers

Selected Documents and Tags

Machine Learning

Learn tagging schema
And how to tag like experts

Documents and stories from Thai Schools

ML BOT

Auto-tagging documents

Knowledge Portal Page

Search

Parents, teachers, strategists

Problem-Base Learning

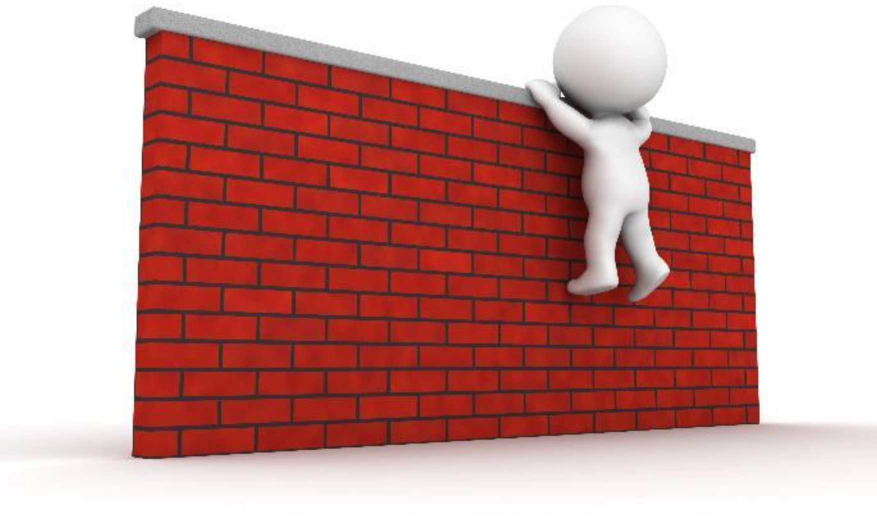
Mathematic Teaching

Creative Thinking

BIG DATA
EXPERIENCE

THE OBSTACLES

- The absence of data
- Lack of data gathering tools
- Existing data quality (consistency, accuracy, completeness, conformity)
- Lack of concept understanding
- Data sharing within and across organizations
- Competing instead of collaborating among internal teams
- Maintainability after initiatives



Initiations requires that

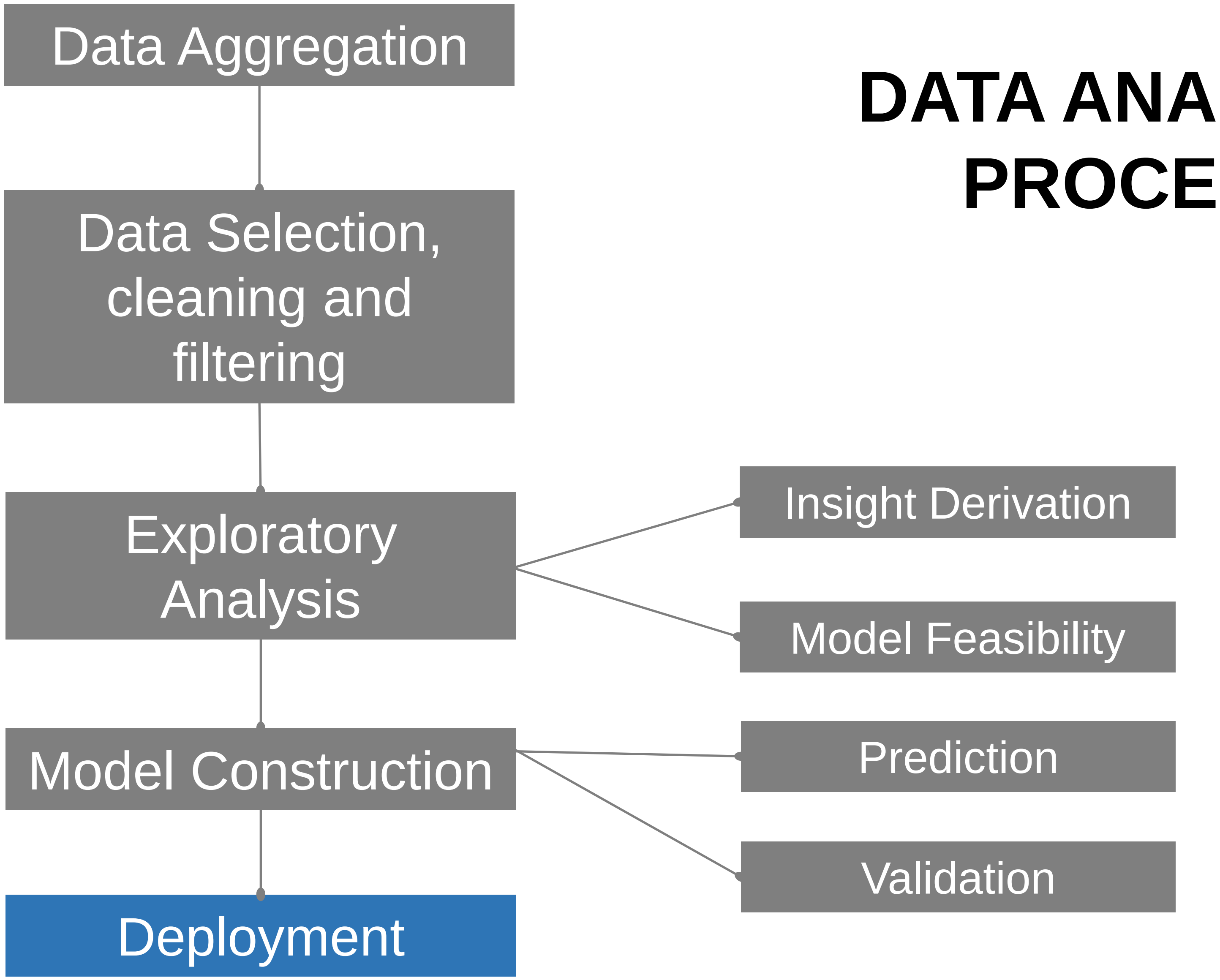
- Managers understand the principles well enough to envision data science opportunities.
- A diverse team of data scientists and business analysts be formed and work closely together.
- The business problem be well specified.
- Data teams be educated and trained on the science of data.
- Community be built for show and share of experiences.
- Management commits to prototyping efforts and initial investments.

KMUTT g-able

Introduction to Big Data Technology

13:00 – 16:00

DATA ANALYTIC PROCESS


Data Aggregation

Data Selection, cleaning and filtering

Exploratory Analysis

Model Construction

Deployment

Insight Derivation

Model Feasibility

Prediction

Validation

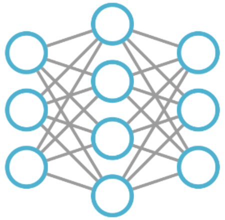
Analytic Methods

Data Mining

The Computational process of discovering patterns in large data sets involving methods at the intersection of statistics, machine learning, and database systems.

Text Analytics

The process of deriving high-quality information from **text**. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.



Machine Learning / Deep Learning

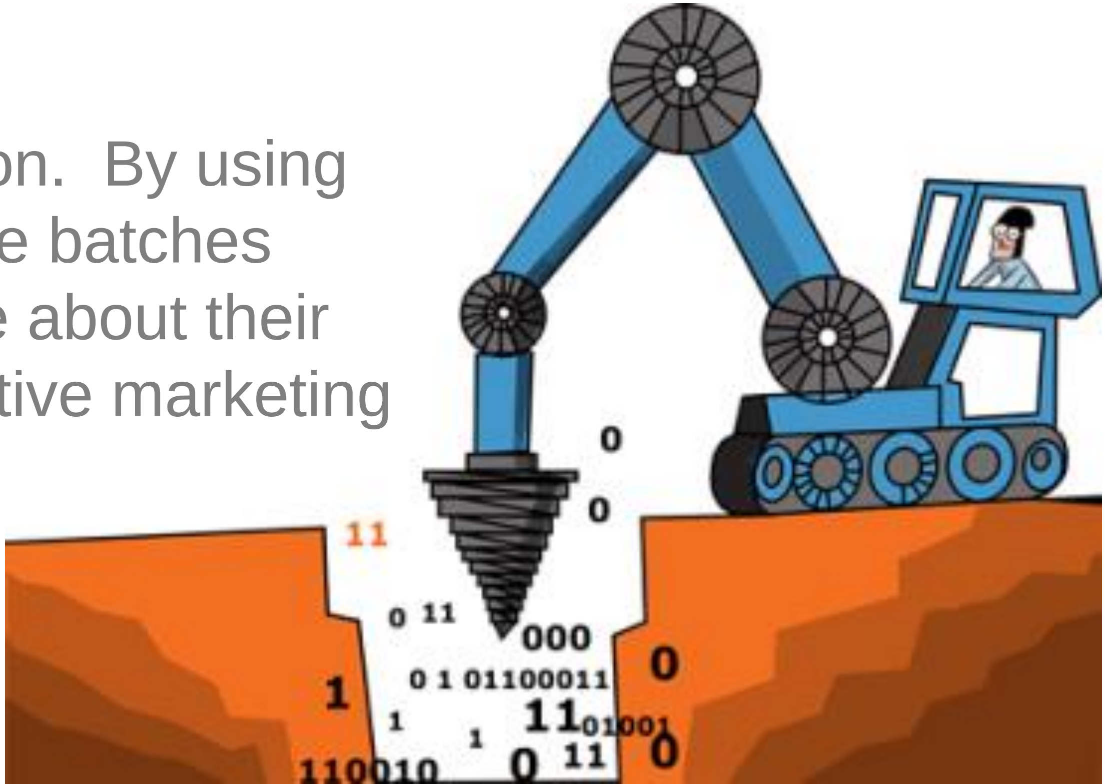
The science of getting computers to learn from data without having to be explicitly programmed by humans. Machine model can teach themselves to grow and change when exposed to new data.

Big Data Technology

Technology designed to manage and process extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

DATA MINING

Turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies.



Common Tasks

1. Classification

   predict, for each individual in a population, which of a set of classes this individual belongs to.

   - Among the customers of Telco, which are likely to respond to a given offer ? (Classes: will respond, will not respond)

2. Regression

   produce a model that, given an individual, estimates the value of the particular variable specific to that individual.

   - How much will a given customer use the service? (variable: service usage)

3. Similarity matching

   identify similar individuals based on data know about them.

   Similarity underlie solutions to other tasks.

   - Finding people who are similar to you in terms of products they have purchased.

Common Tasks

4. Clustering

   group individuals in a population by their similarity (not driven by any specific purpose).

   - Do our customers form natural groups or segments?

5. Co-occurrence grouping

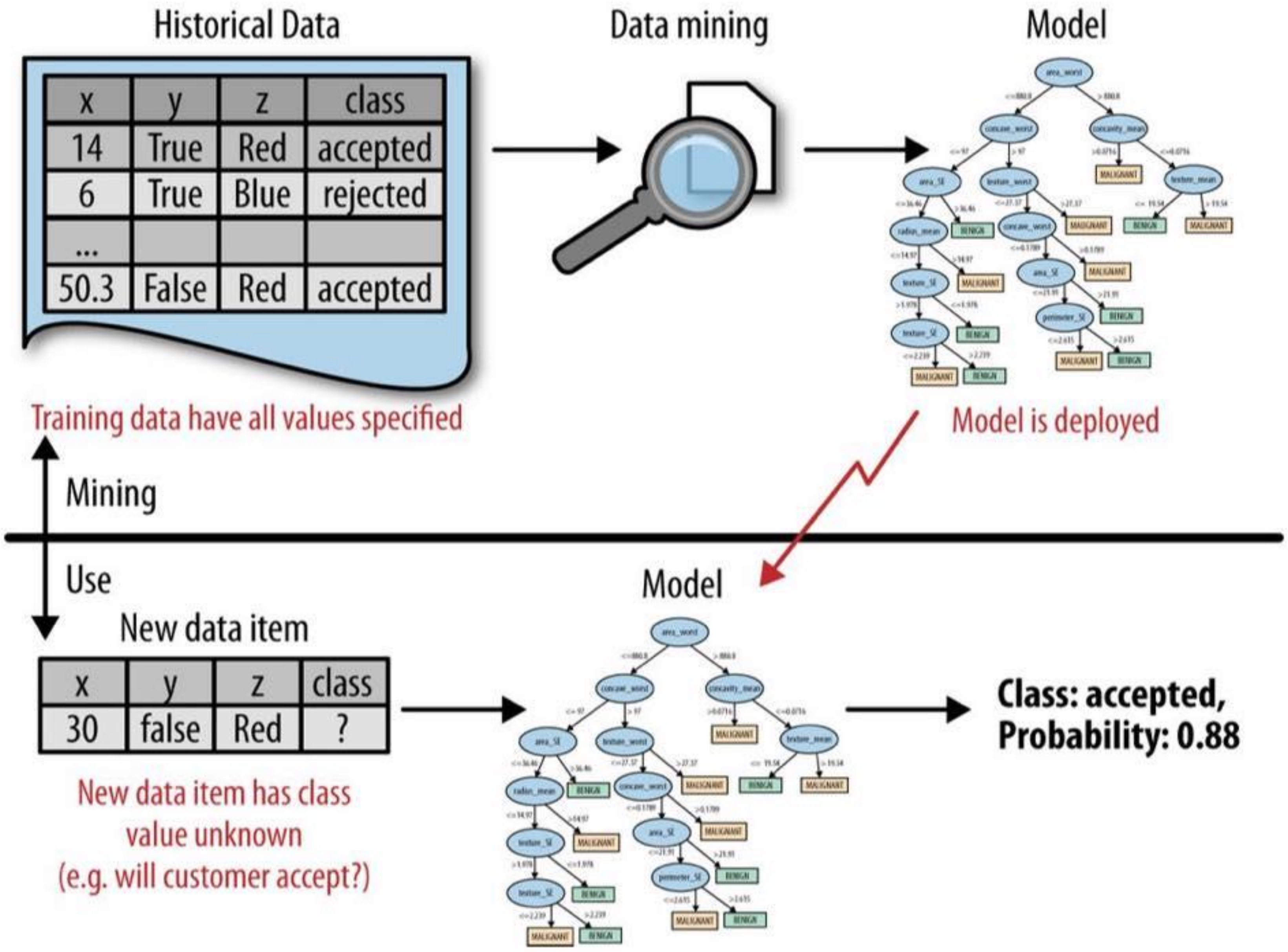
   find associations between entities based on transactions involving them.

   - What items are commonly purchased together?

6. Profiling

   characterize the typical behavior of an individual, group, or population.

   - What is the typical cell phone usage of this customer segment ?
   - Used to establish behavior norms for anomaly detection (fraud detection)



Historical Data

x	y	z	class
14	True	Red	accepted
6	True	Blue	rejected
...			
50.3	False	Red	accepted

Training data have all values specified

Data mining

Model

Model is deployed

Mining

Use

New data item

x	y	z	class
30	false	Red	?

New data item has class
value unknown
(e.g. will customer accept?)

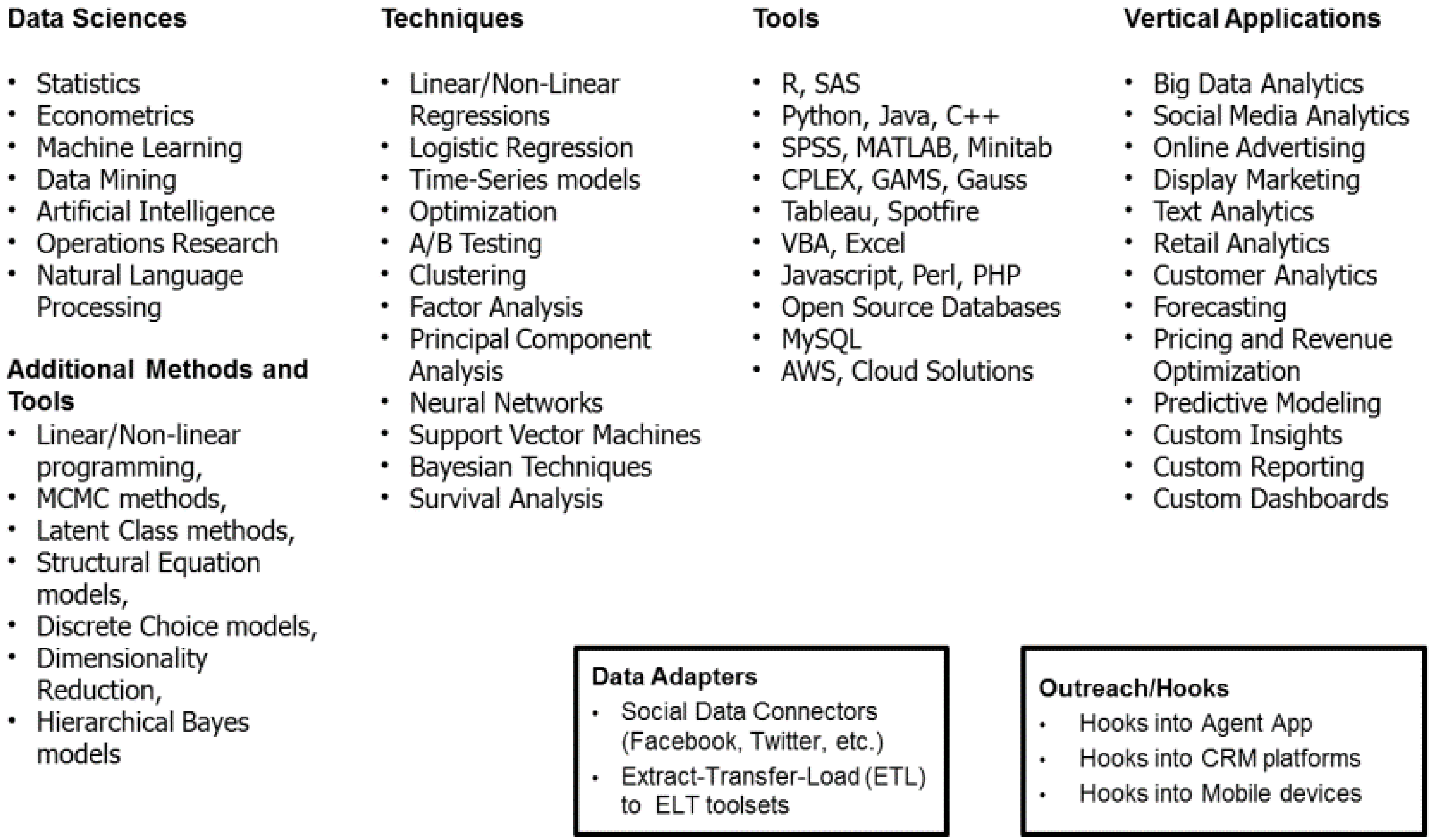
Model

Class: accepted,
Probability: 0.88

Answer Business Questions

- Who are the most profitable customers?
  - A straightforward database query, if “profitable” can be defined clearly.
- Is there really a difference between the profitable customers and the average customer?
  - Statistical Hypothesis testing
- But who really are these customers? Can I characterize them?
  - Automated pattern finding
- Will some new customer be profitable ? How much revenue can I expect?
  - Predictive model of profitability

Knowledge



Data Sciences

- Statistics
- Econometrics
- Machine Learning
- Data Mining
- Artificial Intelligence
- Operations Research
- Natural Language Processing

Additional Methods and Tools

- Linear/Non-linear programming,
- MCMC methods,
- Latent Class methods,
- Structural Equation models,
- Discrete Choice models,
- Dimensionality Reduction,
- Hierarchical Bayes models

Techniques

- Linear/Non-Linear Regressions
- Logistic Regression
- Time-Series models
- Optimization
- A/B Testing
- Clustering
- Factor Analysis
- Principal Component Analysis
- Neural Networks
- Support Vector Machines
- Bayesian Techniques
- Survival Analysis

Tools

- R, SAS
- Python, Java, C++
- SPSS, MATLAB, Minitab
- CPLEX, GAMS, Gauss
- Tableau, Spotfire
- VBA, Excel
- Javascript, Perl, PHP
- Open Source Databases
- MySQL
- AWS, Cloud Solutions

Vertical Applications

- Big Data Analytics
- Social Media Analytics
- Online Advertising
- Display Marketing
- Text Analytics
- Retail Analytics
- Customer Analytics
- Forecasting
- Pricing and Revenue Optimization
- Predictive Modeling
- Custom Insights
- Custom Reporting
- Custom Dashboards

Data Adapters

- Social Data Connectors (Facebook, Twitter, etc.)
- Extract-Transfer-Load (ETL) to ELT toolsets

Outreach/Hooks

- Hooks into Agent App
- Hooks into CRM platforms
- Hooks into Mobile devices

Ref: https://practicalanalytics.wordpress.com/2015/05/25/big-data-analytics-use-cases/

MACHINE LEARNING

Learn from data and make predictions about data by using statistics to develop self learning algorithm

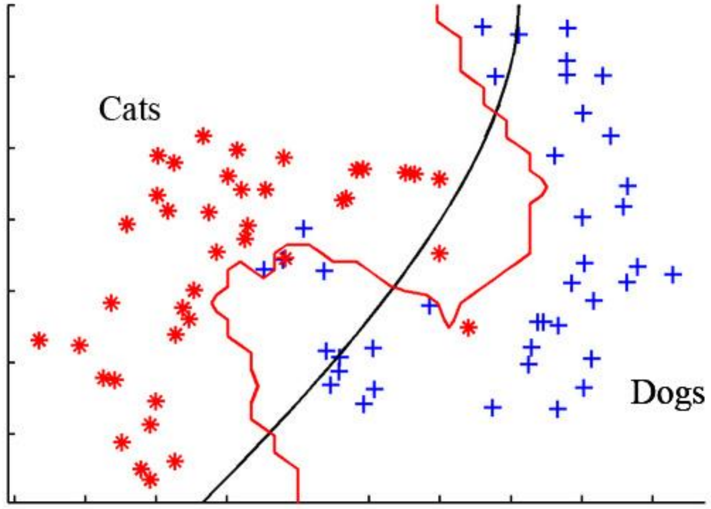
MACHINE LEARNING

Machine Learning

“The science of getting computers to learn from data without having to be explicitly programmed by humans.”

Machine learning is surrounding you

- Google search
- Auto Facebook photo tagging
- Email Spamming
- Games
- Chat bot
- Recommender



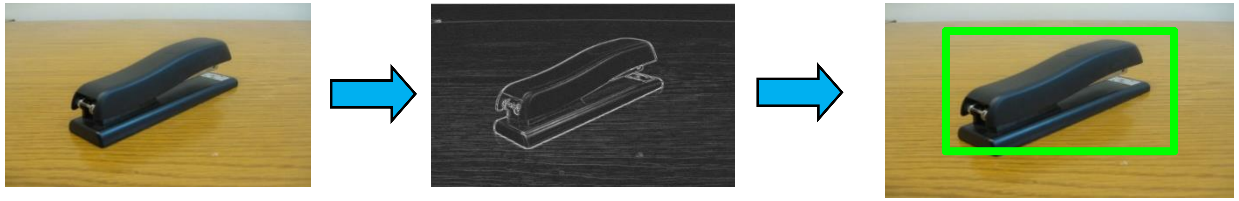
ML Basic Understanding

- It’s all about taking in the ‘input’, pushing out the ‘output’ prediction
  - Example: given the robot’s sensor and camera input, the algorithm pushes out the appropriate movement command.
  - Example: given the search engine terms as input, the algorithm output predictions of what the person is looking for.
- It’s all about letting computer learns what ‘input’ is associated to what ‘output’.

Machine Learns to Do House Chores

How is machine perception done?

Images/video

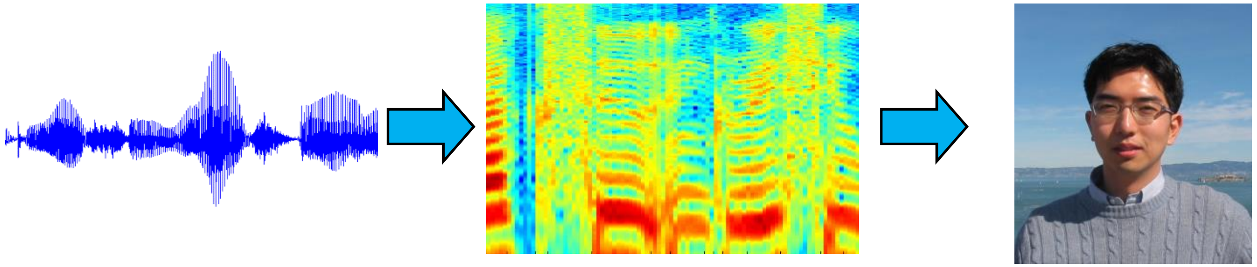


Image

Vision features

Detection

Audio

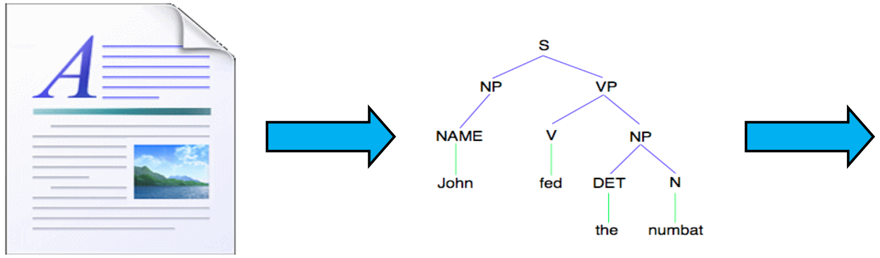


Audio

Audio features

Speaker ID

Text



Text classification,
Machine translation,
Information retrieval, ...

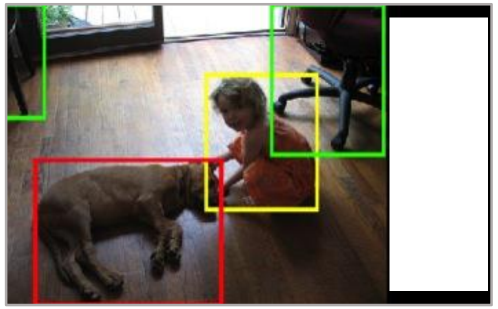
Text

Text features

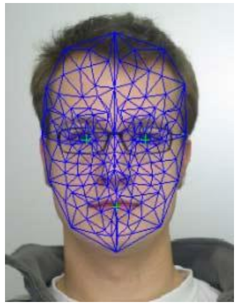
Slide courtesy of Andrew Ng, Stanford University

Early Use Cases

Image Classification, Object Detection, Localization

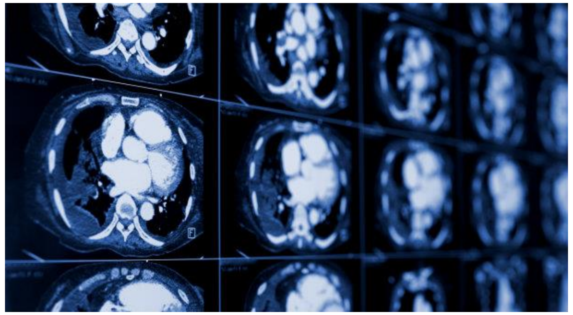


Face Recognition

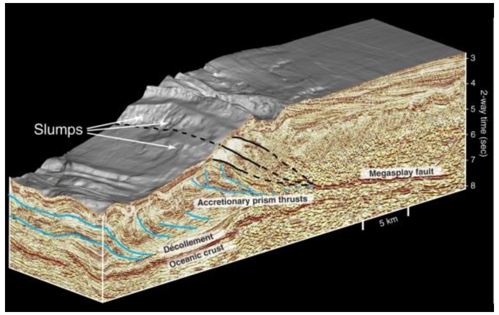


Speech & Natural Language Processing

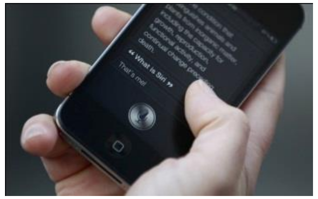
Medical Imaging & Interpretation



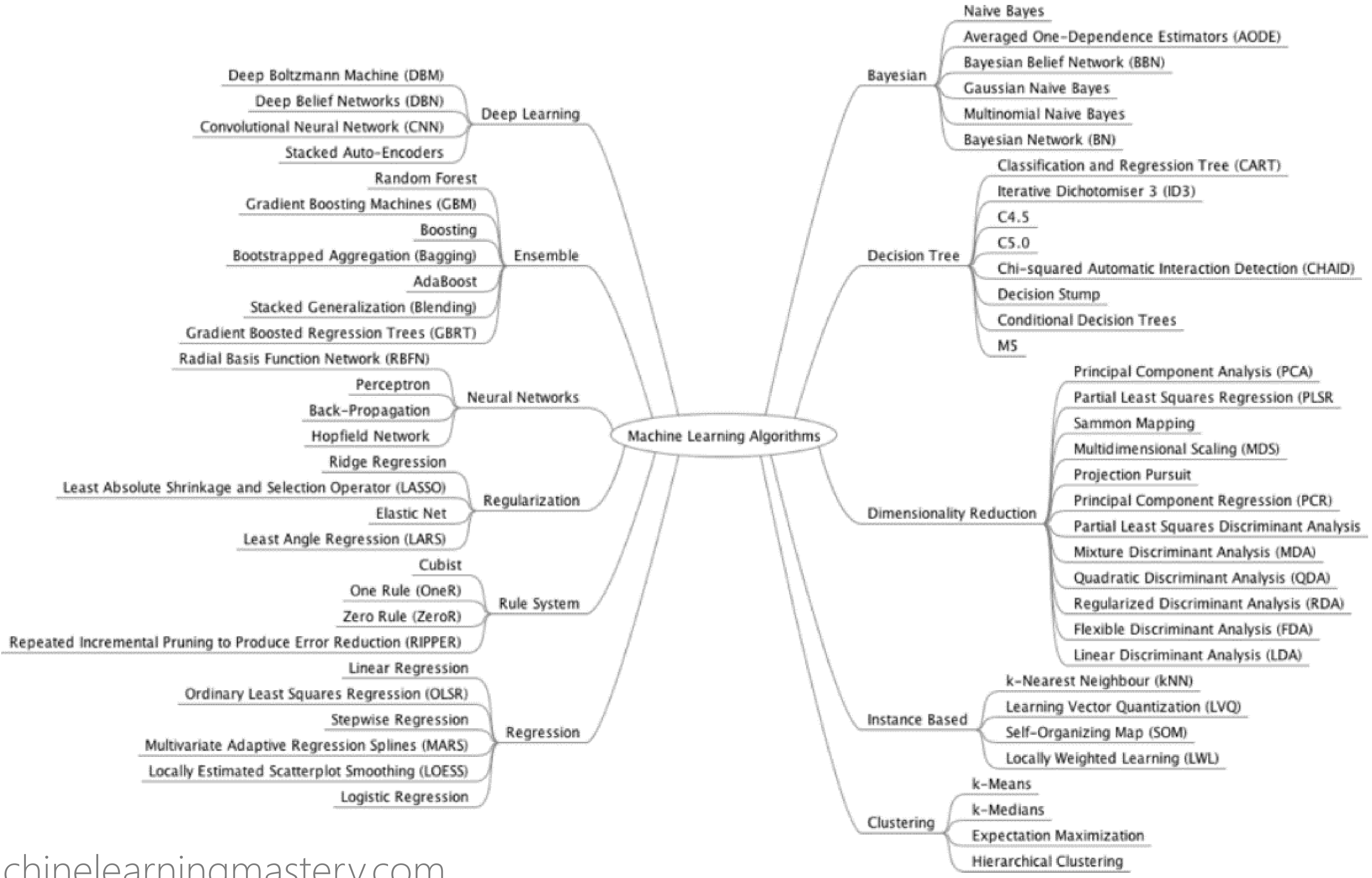
Seismic Imaging & Interpretation



Recommendation



There are so many models around



- Machine Learning Algorithms
  - Deep Learning
    - Deep Boltzmann Machine (DBM)
    - Deep Belief Networks (DBN)
    - Convolutional Neural Network (CNN)
    - Stacked Auto-Encoders
  - Ensemble
    - Random Forest
    - Gradient Boosting Machines (GBM)
    - Boosting
    - Bootstrapped Aggregation (Bagging)
    - AdaBoost
    - Stacked Generalization (Blending)
    - Gradient Boosted Regression Trees (GBRT)
  - Neural Networks
    - Radial Basis Function Network (RBFN)
    - Perceptron
    - Back-Propagation
    - Hopfield Network
  - Regularization
    - Ridge Regression
    - Least Absolute Shrinkage and Selection Operator (LASSO)
    - Elastic Net
    - Least Angle Regression (LARS)
  - Rule System
    - Cubist
    - One Rule (OneR)
    - Zero Rule (ZeroR)
    - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
  - Regression
    - Linear Regression
    - Ordinary Least Squares Regression (OLSR)
    - Stepwise Regression
    - Multivariate Adaptive Regression Splines (MARS)
    - Locally Estimated Scatterplot Smoothing (LOESS)
    - Logistic Regression
  - Bayesian
    - Naive Bayes
    - Averaged One-Dependence Estimators (AODE)
    - Bayesian Belief Network (BBN)
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Bayesian Network (BN)
  - Decision Tree
    - Classification and Regression Tree (CART)
    - Iterative Dichotomiser 3 (ID3)
    - C4.5
    - C5.0
    - Chi-squared Automatic Interaction Detection (CHAID)
    - Decision Stump
    - Conditional Decision Trees
    - M5
  - Dimensionality Reduction
    - Principal Component Analysis (PCA)
    - Partial Least Squares Regression (PLSR
    - Sammon Mapping
    - Multidimensional Scaling (MDS)
    - Projection Pursuit
    - Principal Component Regression (PCR)
    - Partial Least Squares Discriminant Analysis
    - Mixture Discriminant Analysis (MDA)
    - Quadratic Discriminant Analysis (QDA)
    - Regularized Discriminant Analysis (RDA)
    - Flexible Discriminant Analysis (FDA)
    - Linear Discriminant Analysis (LDA)
  - Instance Based
    - k-Nearest Neighbour (kNN)
    - Learning Vector Quantization (LVQ)
    - Self-Organizing Map (SOM)
    - Locally Weighted Learning (LWL)
  - Clustering
    - k-Means
    - k-Medians
    - Expectation Maximization
    - Hierarchical Clustering

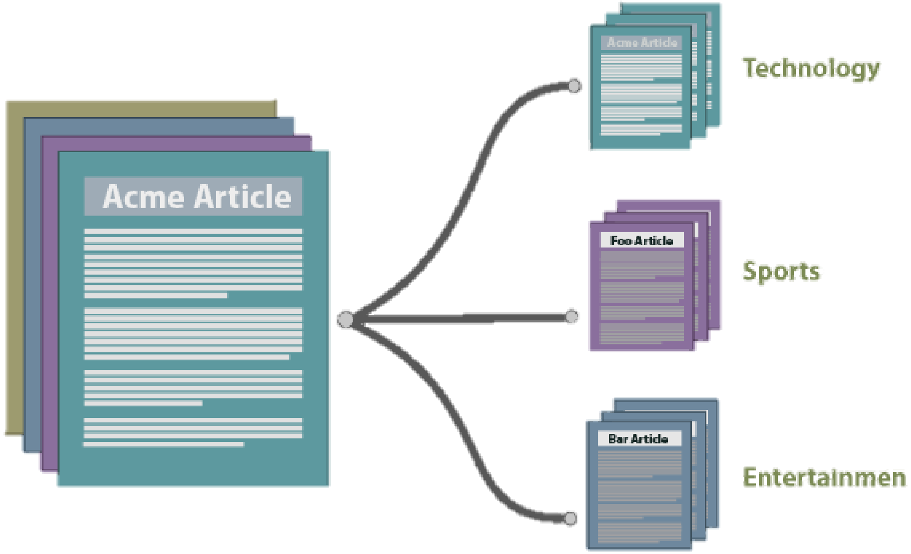
Ref: machinelearningmastery.com

TEXT MINING AND NLP

Deriving high-quality information from text by devising of patterns and trends through means such as statistical pattern learning.

Text Classification and Clustering

- Classification
  - ✓ To assign a document to one or more classes or categories.
- Clustering:
  - ✓ The application of cluster analysis to textual documents

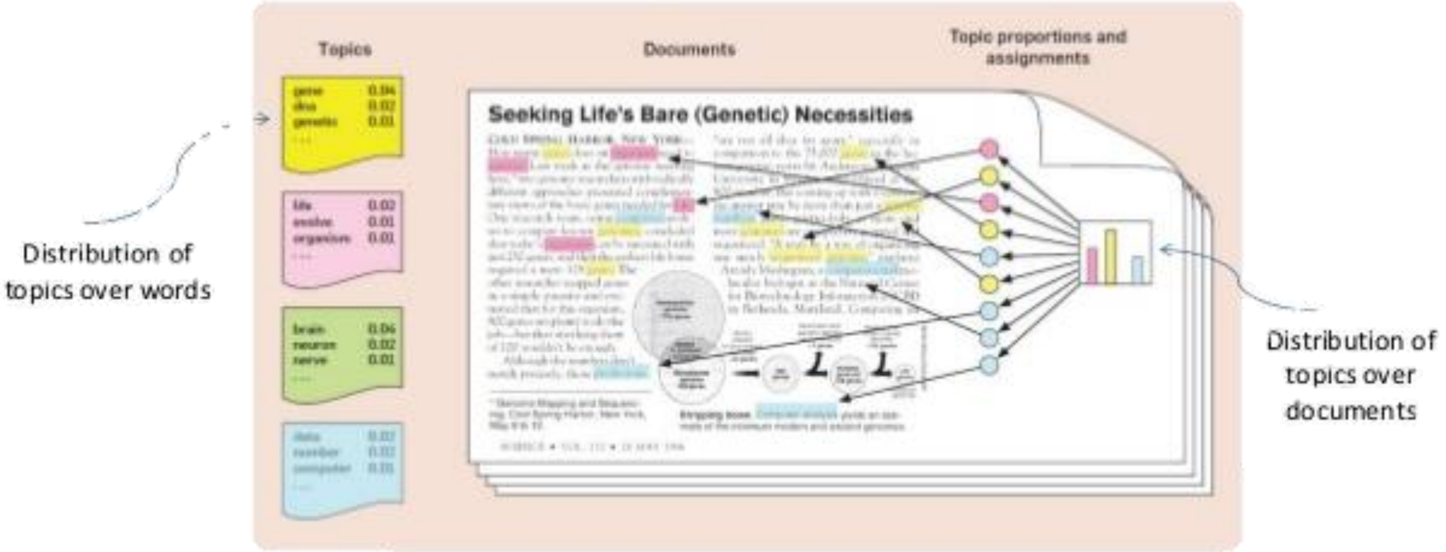


Sentiment Analysis

- To determine the attitude of a writer with respect to some topic or the overall contextual polarity of a document.
- Widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service

Topic Discovery

- Characterizes document according to topics
  - ✓Discover topics mentioned about “ประชามติ” on the social network
  - ✓Discover topics mentioned about “พร้อมเพย์” on the social network

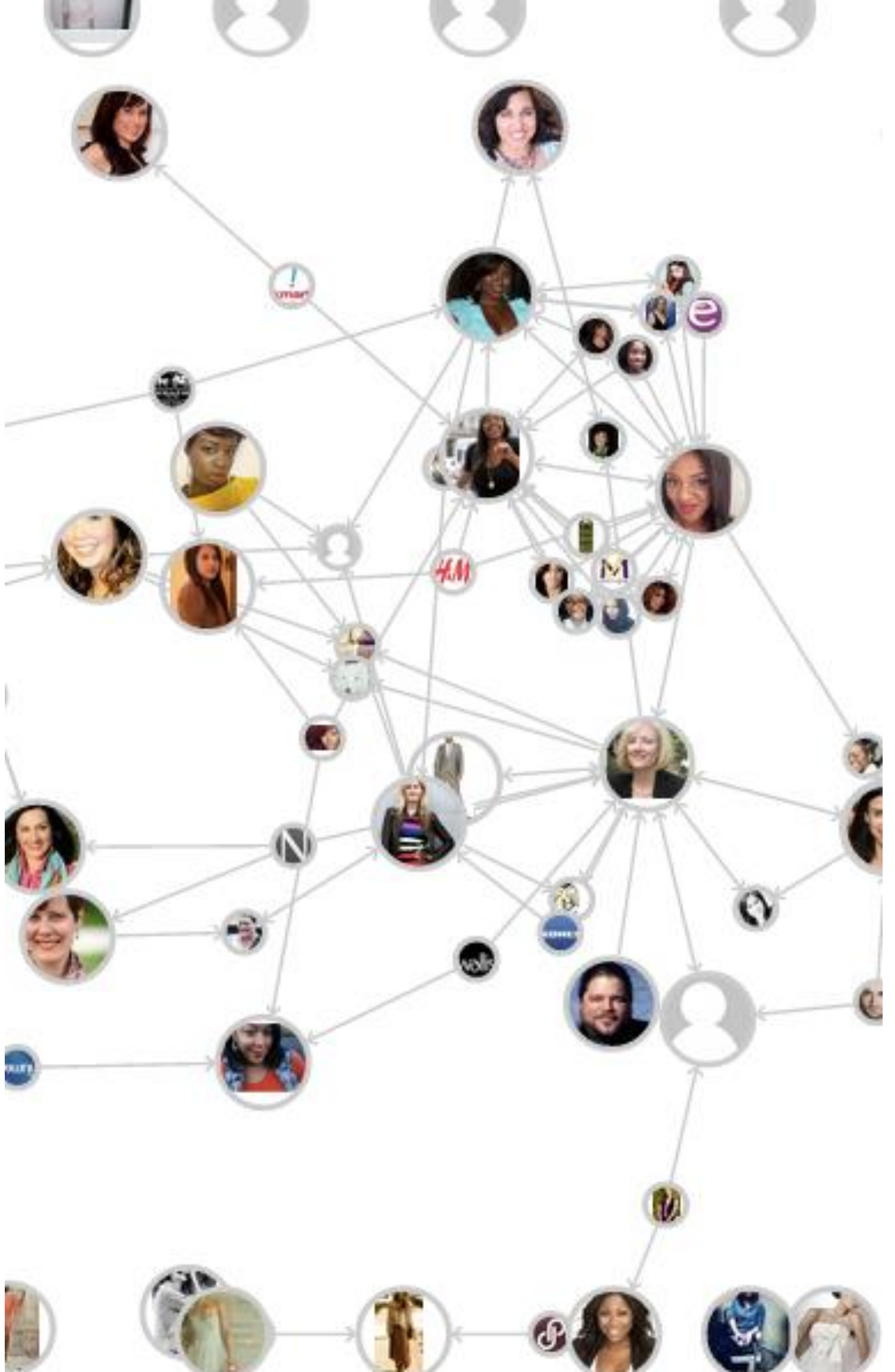

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
life 0.02
evolve 0.01
organism 0.01
brain 0.04
neuron 0.02
nerve 0.01
data 0.02
number 0.02
computer 0.01
Seeking Life's Bare (Genetic) Necessities
Distribution of topics over words
Distribution of topics over documents

[Image from Blei, D. *Probabilistic Topic Models*, Communication of the ACM, 2012]

Influencer Analysis

- An influencer is an individual who has above-average impact on a specific niche process.

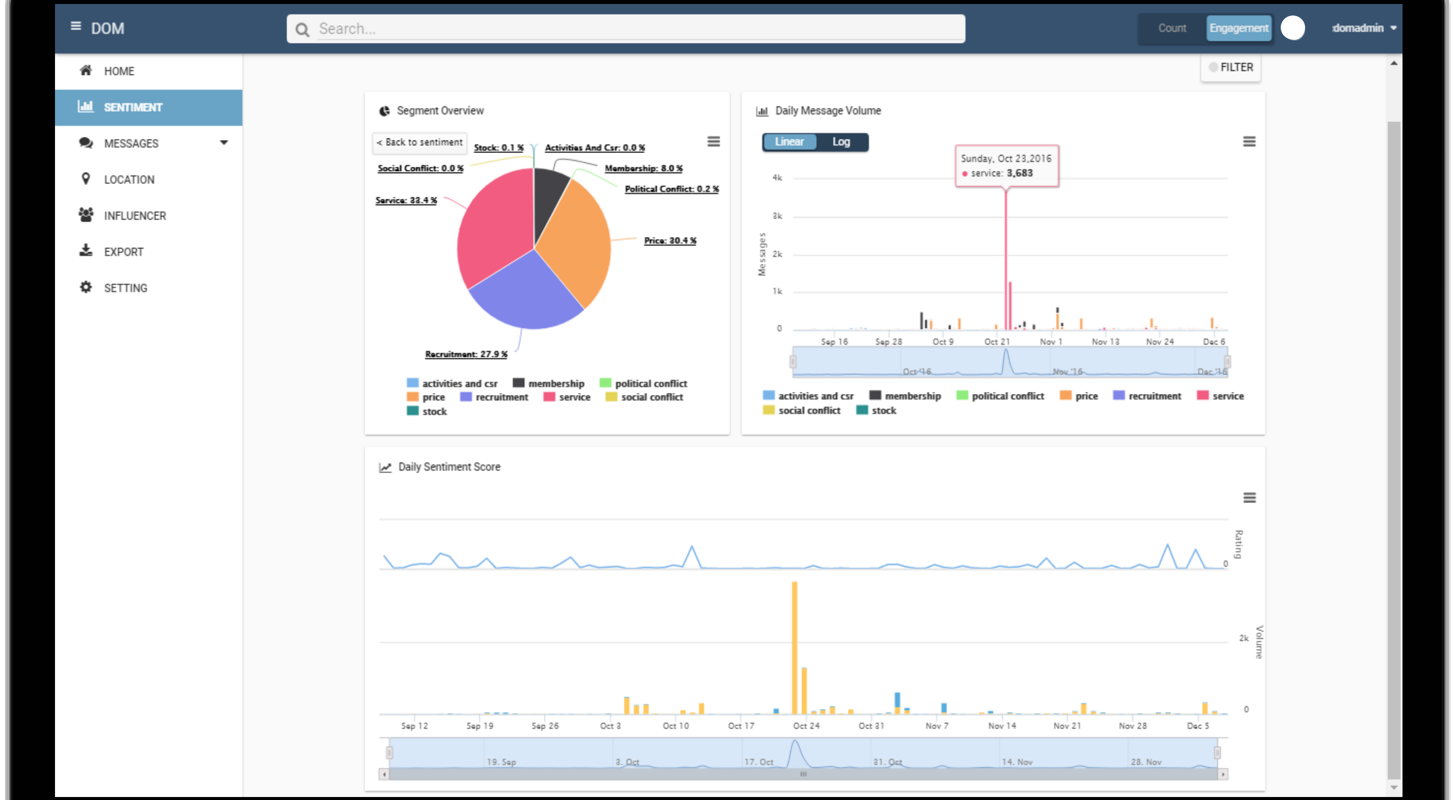
- On the social network, a influencer can referred to the most shaping a discussion about a brand or topic.



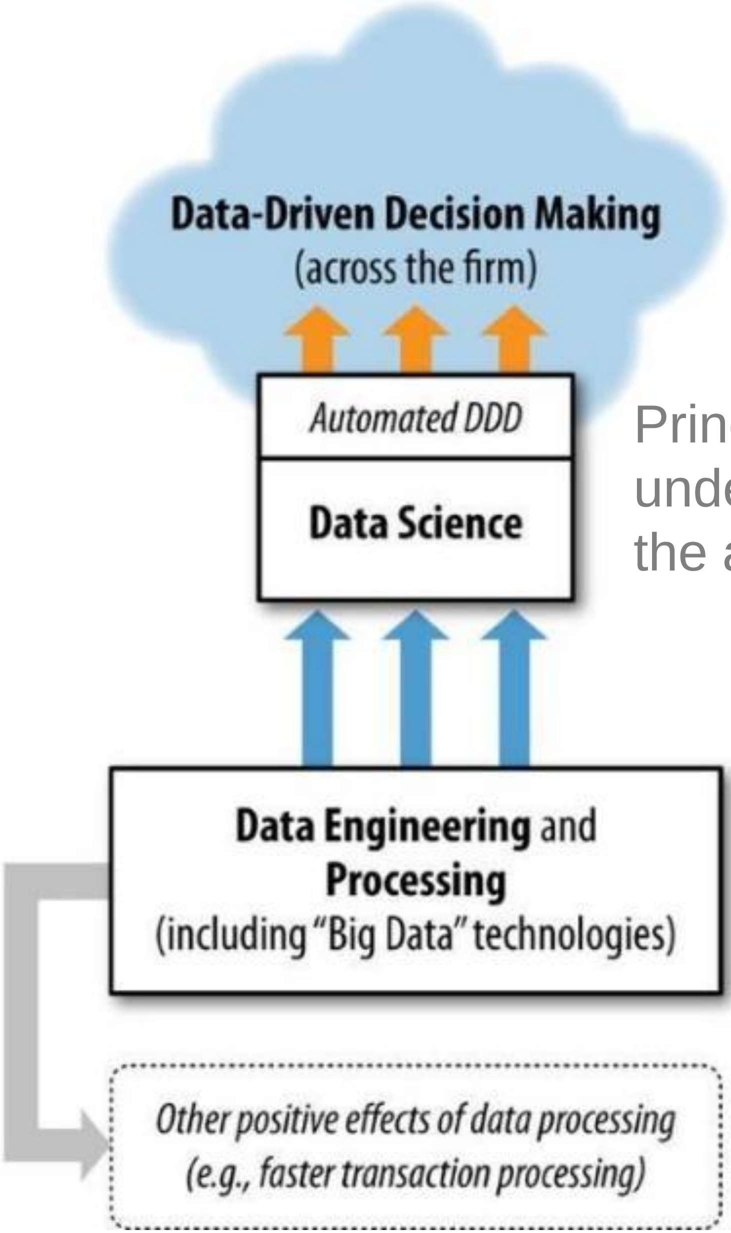
Social Analytics

- Keep tracking your brand & competitors
- Real-time monitoring your feedback
- Real-time detect anomaly issues
- Knows your feedback sentiment
- Knows where your target audiences are
- Find out who influences your brand

BIG DATA
EXPERIENCE



KMUTT g·able


Data-Driven Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
(including "Big Data" technologies)
Other positive effects of data processing
(e.g., faster transaction processing)

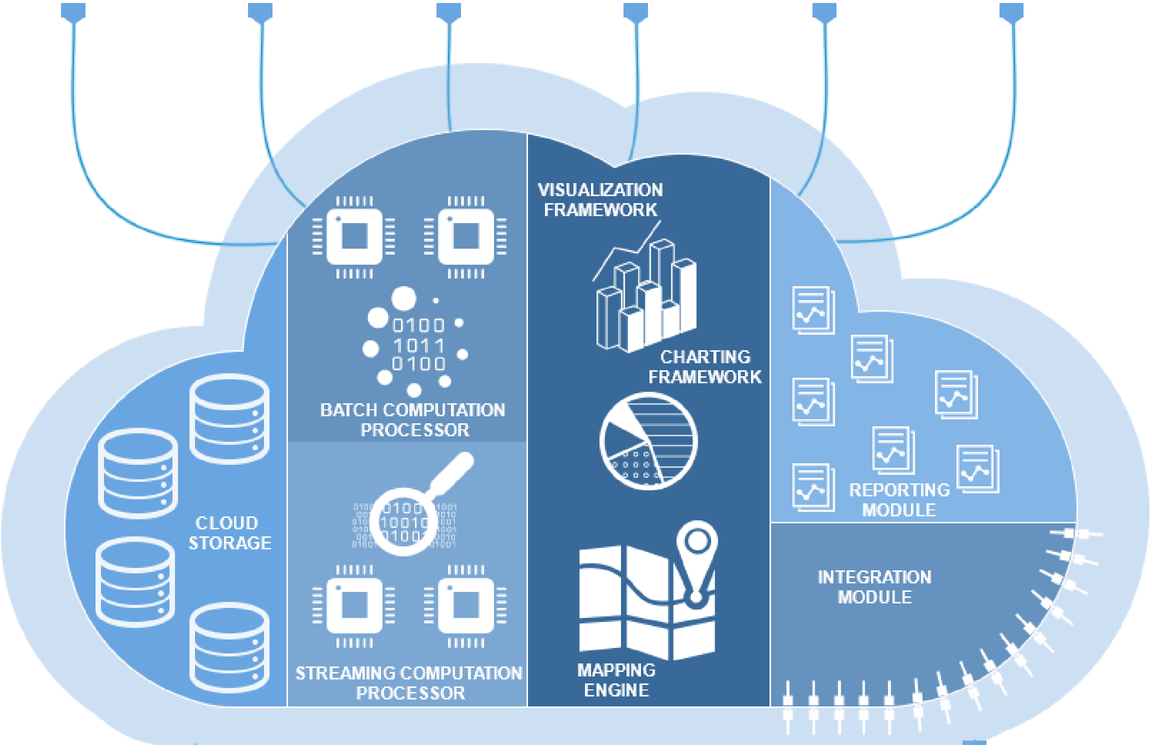
DDD = practice of basing decision on the analysis of data, rather than intuition

Principles and techniques for understanding phenomena via the analysis of data.

The data analysis is not testing a simple hypothesis, but the data are explored with the hope that something useful will be discovered.

Accessing and processing of massive-scale data flexibly and efficiently with Big Data technologies

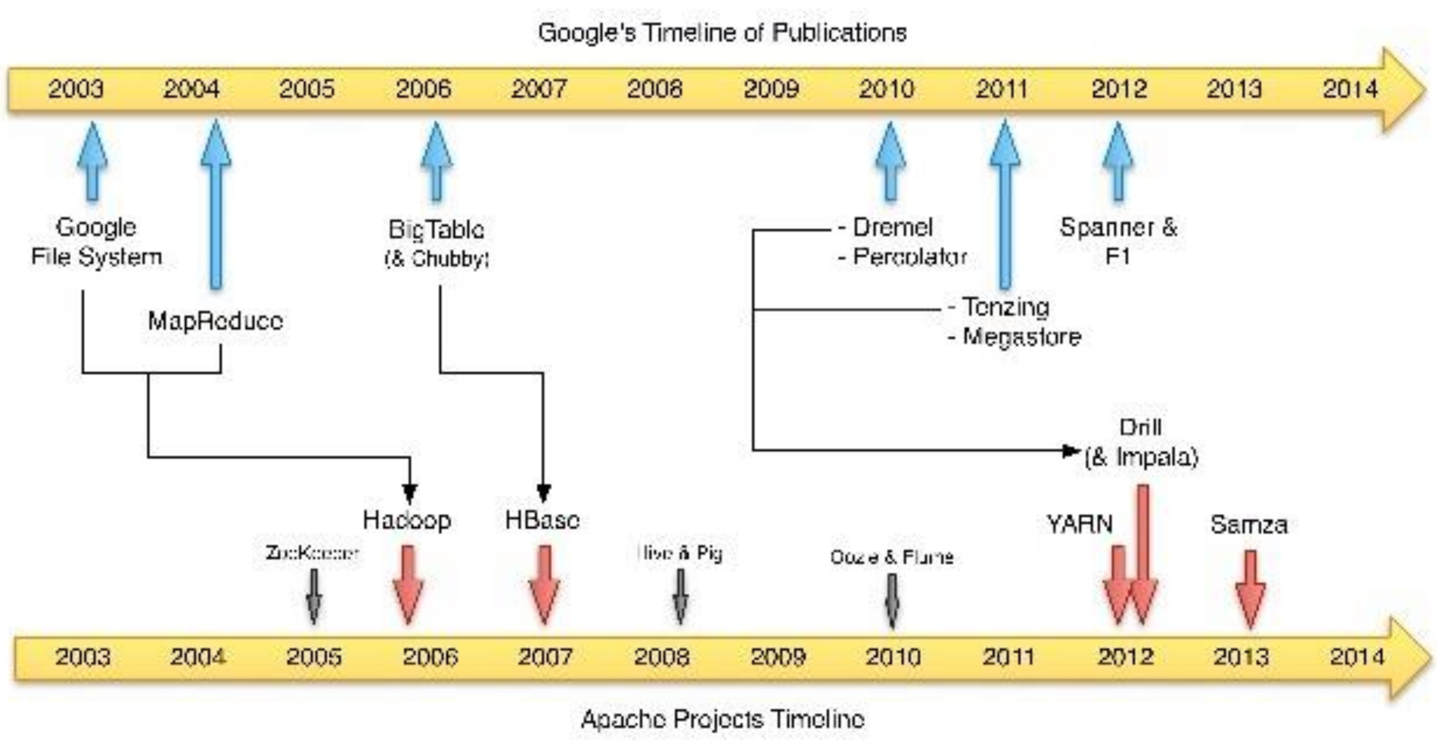
Hadoop Big Data Platform



Introduction to Hadoop

Hadoop is apache open source framework which provides **reliable, scalable, distributed storage and processing** of large data sets across clusters of computers using simple programming models

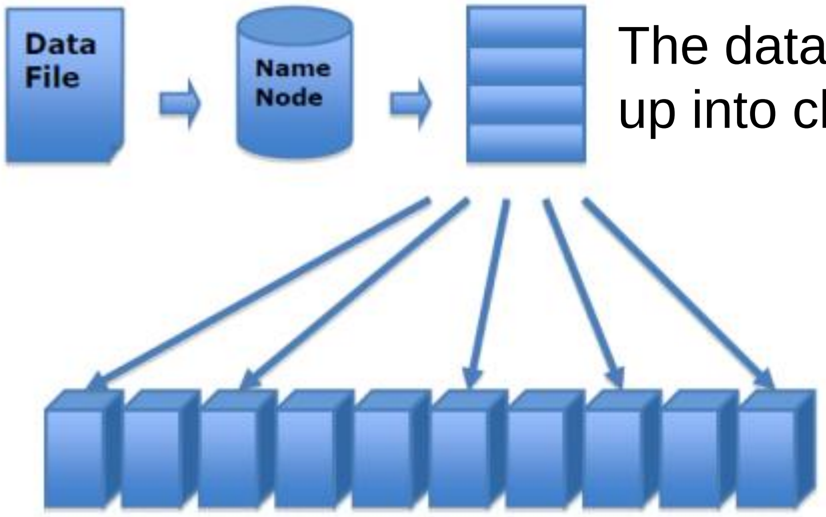
Hadoop Timeline



Hadoop is based to a large degree on ideas crafted by Google. Possibly to develop a competitive market Google published many technical papers describing the technologies driving the world’s largest search engine provider – and data acquirer of modern times

Hadoop Core Concept (1)

- Big data (Social network, scientific, Clickstream, etc.) is here and we are struggling to store, access, and analyze it.
- To reduce reading/writing time from/to data storage, multiple disks may be used in parallel.



The data file is broken up into chunks

The chunks are replicated 3 times
And scattered amongst the disks

Ref: http://reportbank.com/

Hadoop Core Concept (2)

- Applications are written in high-level code
  - Developers do not worry about network programming, temporal dependencies etc.
- Nodes talk to each other as little as possible
  - Developer should not write code which communicates between nodes
  - “Share Nothing” architecture
- Data is spread among machines in advance
  - Computation happens where the data us stored, whenever possible
  - Data is replicated multiple times on the system for increased availability and reliability

Hadoop vs. Traditional RDBMS

RDBMS

- Refined
- Has a lot of features
- Accelerates very fast
- Pricey
- Expensive to maintain

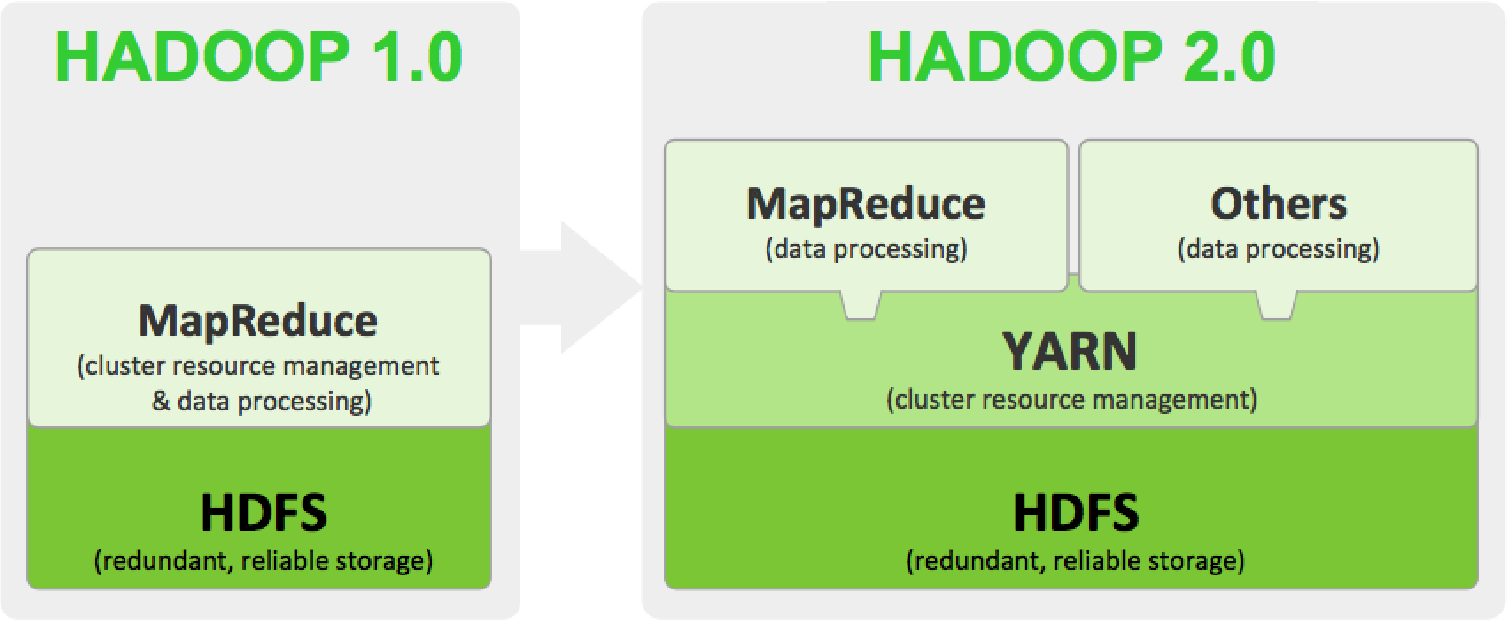
Hadoop

- Rough
- Missing a lot of “luxury”
- Slow to accelerate
- Carries almost anything
- Moves a lot of stuff very efficiently

http://www.hadoopsphere.com/2012/08/analogies-romans-train-dabbawalla-oil.html

Core Components of Hadoop

- Shared storage – HDFS (Hadoop Distributed File System)
- Data processing – MapReduce
- Resource management –YARN* (Yet Another Resource Negotiator)


HADOOP 1.0
MapReduce
(cluster resource management & data processing)
HDFS
(redundant, reliable storage)
HADOOP 2.0
MapReduce
(data processing)
Others
(data processing)
YARN
(cluster resource management)
HDFS
(redundant, reliable storage)

Hadoop : HDFS

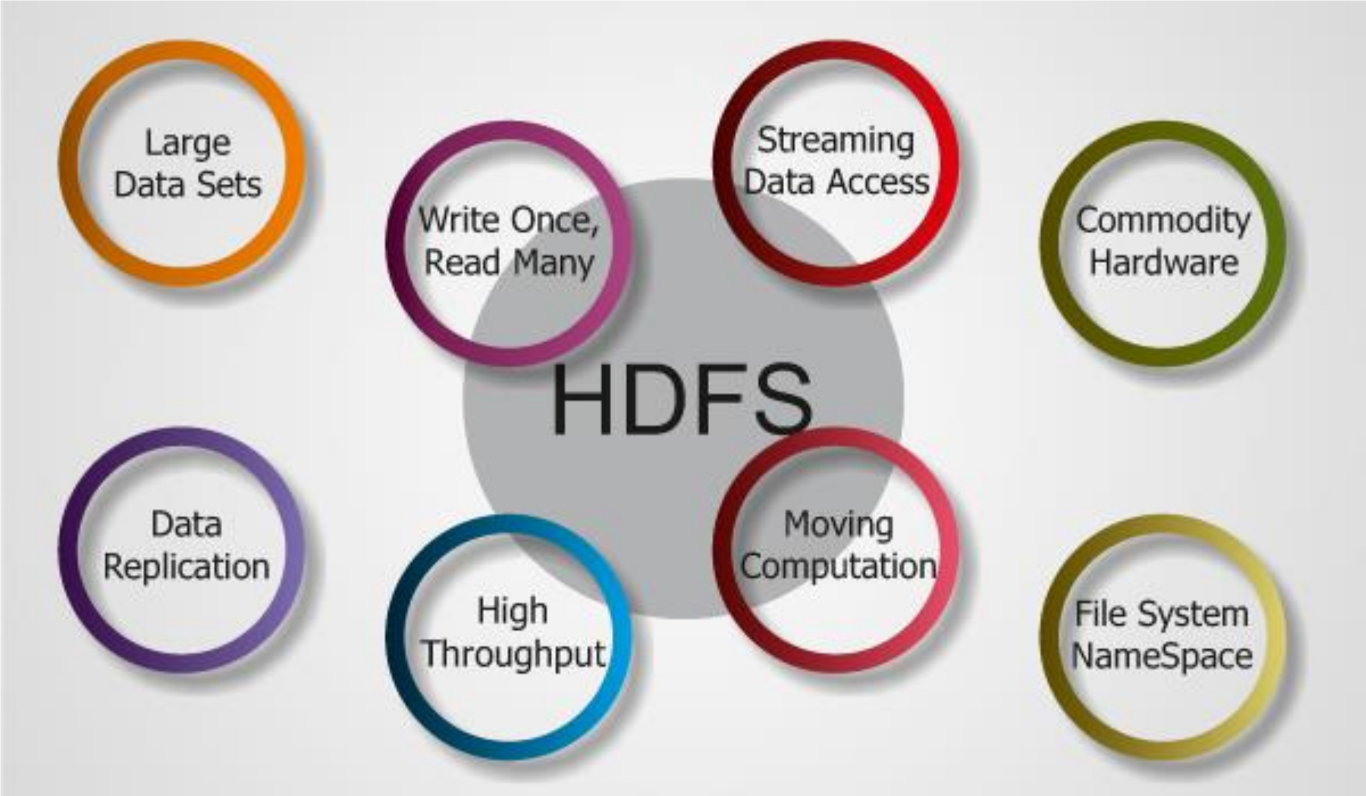
- HDFS, the Hadoop Distributed File System, is responsible for storing data on the cluster.

- Data files are split into blocks and distributed across multiple nodes in the cluster.

- Each block is replicated multiple times, with the default set to three times. Replicas are stored on different nodes, which ensures both reliability and availability

Hadoop HDFS

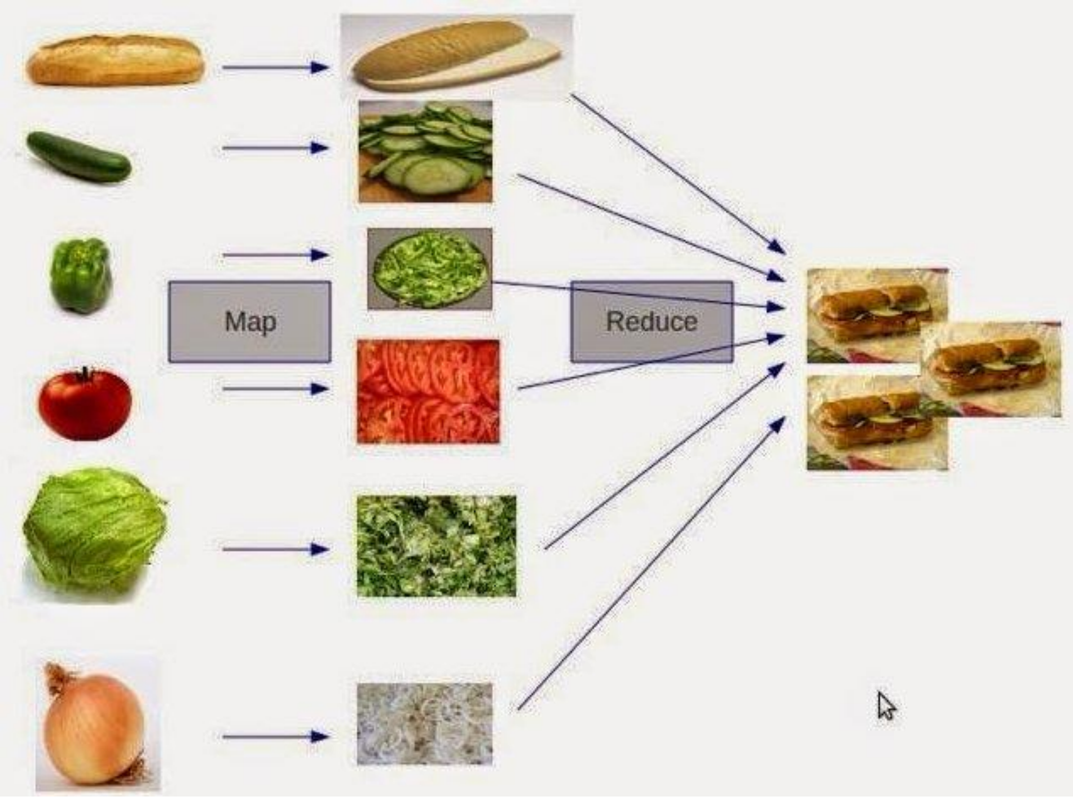
- HDFS is a file system written in Java. It is based on Google’ GFS.
- HDFS sits on top of a native file system, e.g. ext3, ext4, xfs etc.
- It provides redundant storage for massive amounts of data, using cheap, unreliable computers.



https://developer.yahoo.com/hadoop/tutorial/module2.html

Hadoop MapReduce

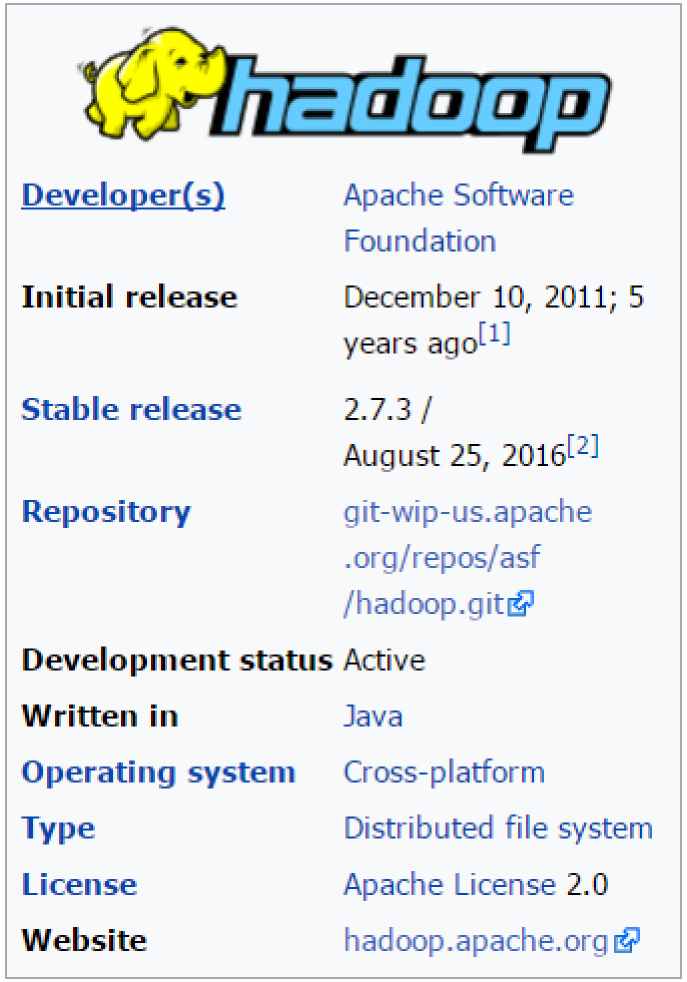
- MapReduce is a programming model which enables batch processing for large volumes of data on a cluster of computers.
- The processing is split into two phases, allowing the computation to run in parallel across multiple nodes.



https://developer.yahoo.com/hadoop/tutorial/module4.html

Hadoop : MapReduce

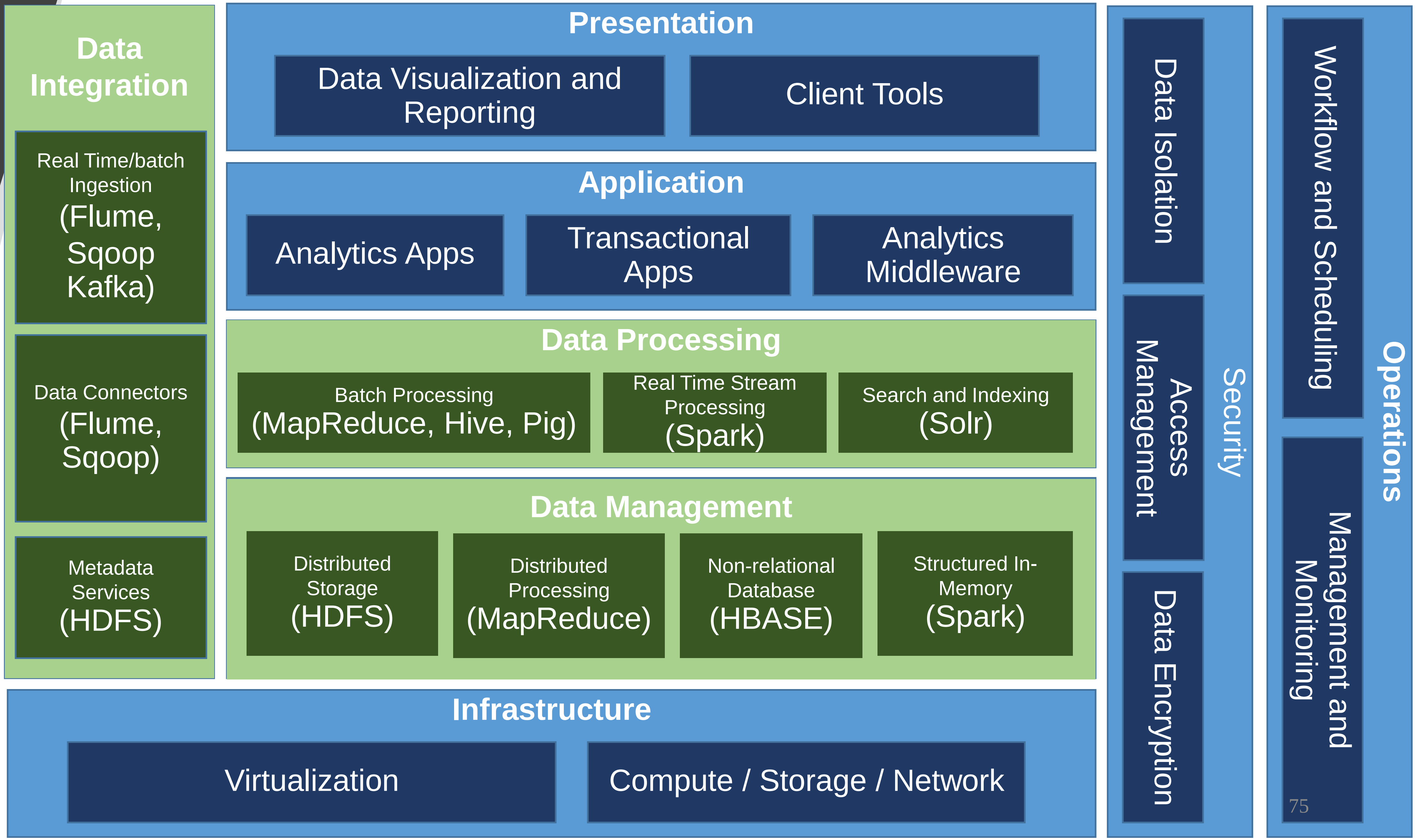
- MapReduce is a system (one of many) used to process data in the Hadoop cluster.
- It consists of two phases: Map and then Reduce.
- Each Map task operates on a discrete portion of the overall dataset, typically one HDFS data block.
- After all Maps are complete, the MapReduce system distributes the intermediate data to fewer nodes which perform the Reduce Phase.



Developer(s)	Apache Software Foundation
Initial release	December 10, 2011; 5 years ago[1]
Stable release	2.7.3 / August 25, 2016[2]
Repository	git-wip-us.apache.org/repos/asf/hadoop.git
Development status	Active
Written in	Java
Operating system	Cross-platform
Type	Distributed file system
License	Apache License 2.0
Website	hadoop.apache.org

Apache Hadoop is an open-source software framework for distributed storage and distributed processing of very large data

Cloudera was the first commercial software vendor to release a Hadoop Distribution with enterprise features security and governance



Data Integration

Real Time/batch Ingestion (Flume, Sqoop Kafka)

Data Connectors (Flume, Sqoop)

Metadata Services (HDFS)

Presentation

Data Visualization and Reporting

Client Tools

Application

Analytics Apps

Transactional Apps

Analytics Middleware

Data Processing

Batch Processing (MapReduce, Hive, Pig)

Real Time Stream Processing (Spark)

Search and Indexing (Solr)

Data Management

Distributed Storage (HDFS)

Distributed Processing (MapReduce)

Non-relational Database (HBASE)

Structured In-Memory (Spark)

Infrastructure

Virtualization

Compute / Storage / Network

Security

Data Isolation

Access Management

Data Encryption

Operations

Workflow and Scheduling

Management and Monitoring

Hadoop and Ecosystem Overview

- Data Storage: HDFS
- Processing Framework: MapReduce, APACHE Spark™
- Workload Management: hadoop YARN
- Coordinator and Workflow Scheduling: ZooKeeper, Oozie

  OOZiE Apache ZooKeeper™

- Data Integration: Flume, Sqoop, Kafka

  FLUME sqoop kafka

Hadoop and Ecosystem Overview

- SQL Engine and Analytics: Hive, Impala
- Search and Indexing:
- NoSQL Storage: Hbase
- User Interface: HUE

The right components for the right solution

Apache Solr
kafka
Spark
Lightning-Fast Cluster Computing
APACHE HBASE

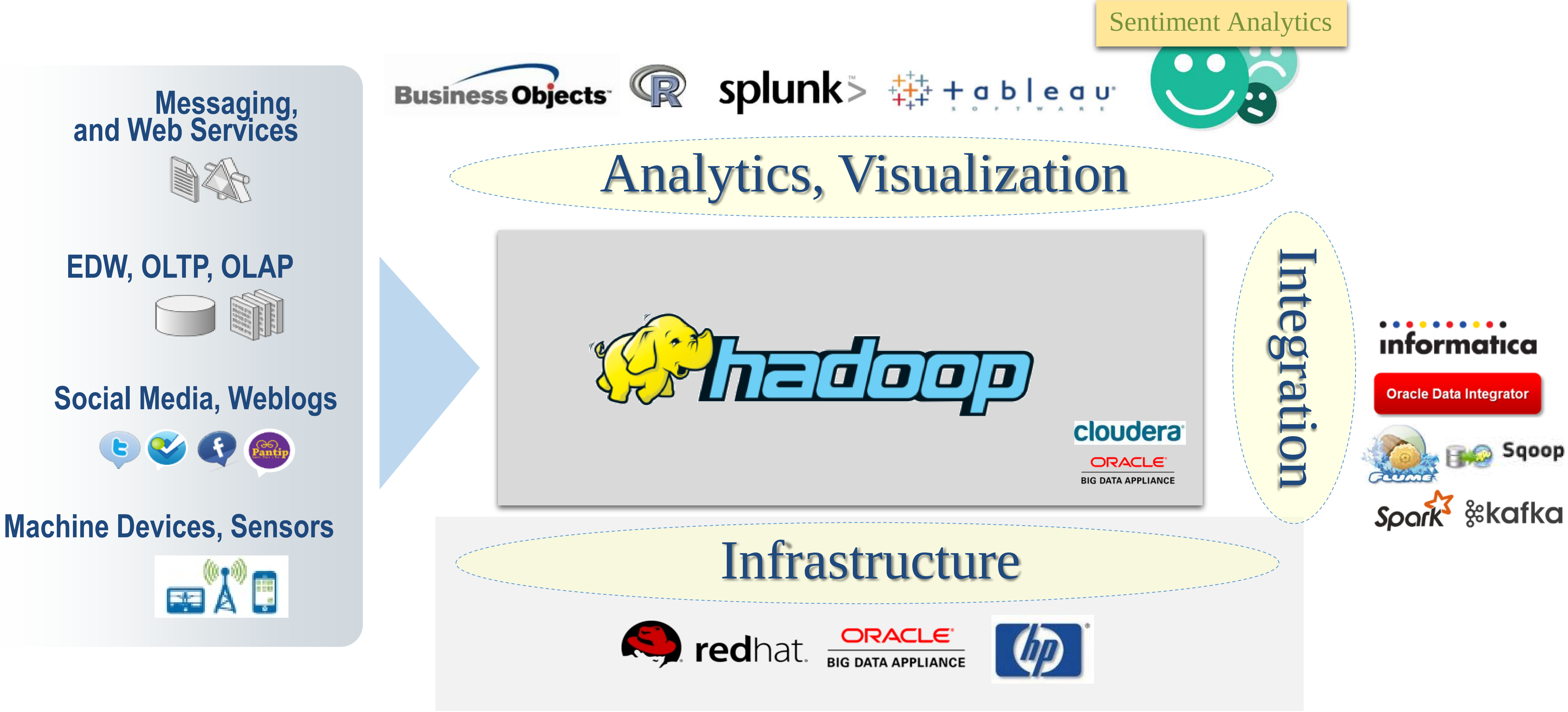
Full Text Search and Indexing

Interactive Analytic SQL Engine

Batch/Real-time Processing

NoSQL Storage

BIG DATA Ecosystem
for Data Lake Solutions



Messaging, and Web Services

EDW, OLTP, OLAP

Social Media, Weblogs

Machine Devices, Sensors

Sentiment Analytics

Analytics, Visualization

Integration

Infrastructure

Big Data In The cloud

- “Picking between Spark or Hadoop isn’t the key to big data success. Picking the right infrastructure is”, www.infoworld.com.
- The key is running both real-time and batch processing on elastic infrastructure. Thus, cloud has a big role in big data analytics.
- Hundreds of terabytes or petabytes of data are hard to move across the network, Hadoop clusters should be on premise and on various clouds.
- Analytics should be performed wherever the bulk of the data has landed.
- When the newer data sets (social network data, machine and sensor data) originate outside the enterprise, the public cloud becomes a natural place to do the processing.
- Cloud service providers can offer Hadoop clusters that scale automatically with the demand of the customer for a cost.

“Information is the oil of the 21st century, and analytics is the combustion engine”

Peter Sondergaard, Senior Vice President, Gartner