Data Science for Business

ON, Byung-Won(溫炳原)[*pronounced as "ohn, byongwon"*], PhD/Prof
Department of Software Convergence Engineering
Kunsan National University, South Korea
bwon@kunsan.ac.kr

July 12, 2018

DILAB

dilab.kunsan.ac.kr

Home
Members
Projects
Papers
Demo
Contact

Gallery
Internal

Introduction to Myself

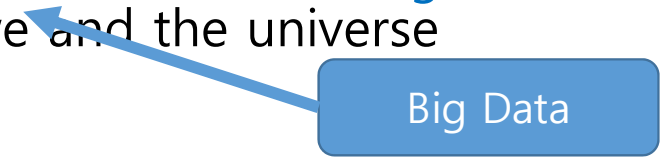
Byung-Won On is an associate professor in Department of Software Convergence Engineering, Kunsan National University, Gunsan-si, Jeollabuk-do, Korea. He has been also leading Data Intelligence Lab at the same university. In 2007, he earned his PhD degree in Department of Computer Science and Engineering at the Pennsylvania State University at University Park, PA, USA. He also received his MS degree in Department of Computer Science and Engineering at Korea University, Seoul, Korea in 2000. His recent research interests are around Text Data Mining, Machine Learning, and Artificial Intelligence, mainly working on *Big Data Summarization*, *Creative Computing*, *Fake News Detection*, and *Distributed Deep Learning Models*. He is an editor of ETRI journal and Frontiers in Big Data journal. Nowadays, he has served as a committee member of ISO/IEC JTC 1/SC 32 - Data Management and Interchange, Korean Association of Data Science, and SIG on Human Language Technology in Korean Institute of Information Scientists and Engineers. He is also a committee member of Informatization Committee in Jeollabuk-do Provincial Government.

Work Experience

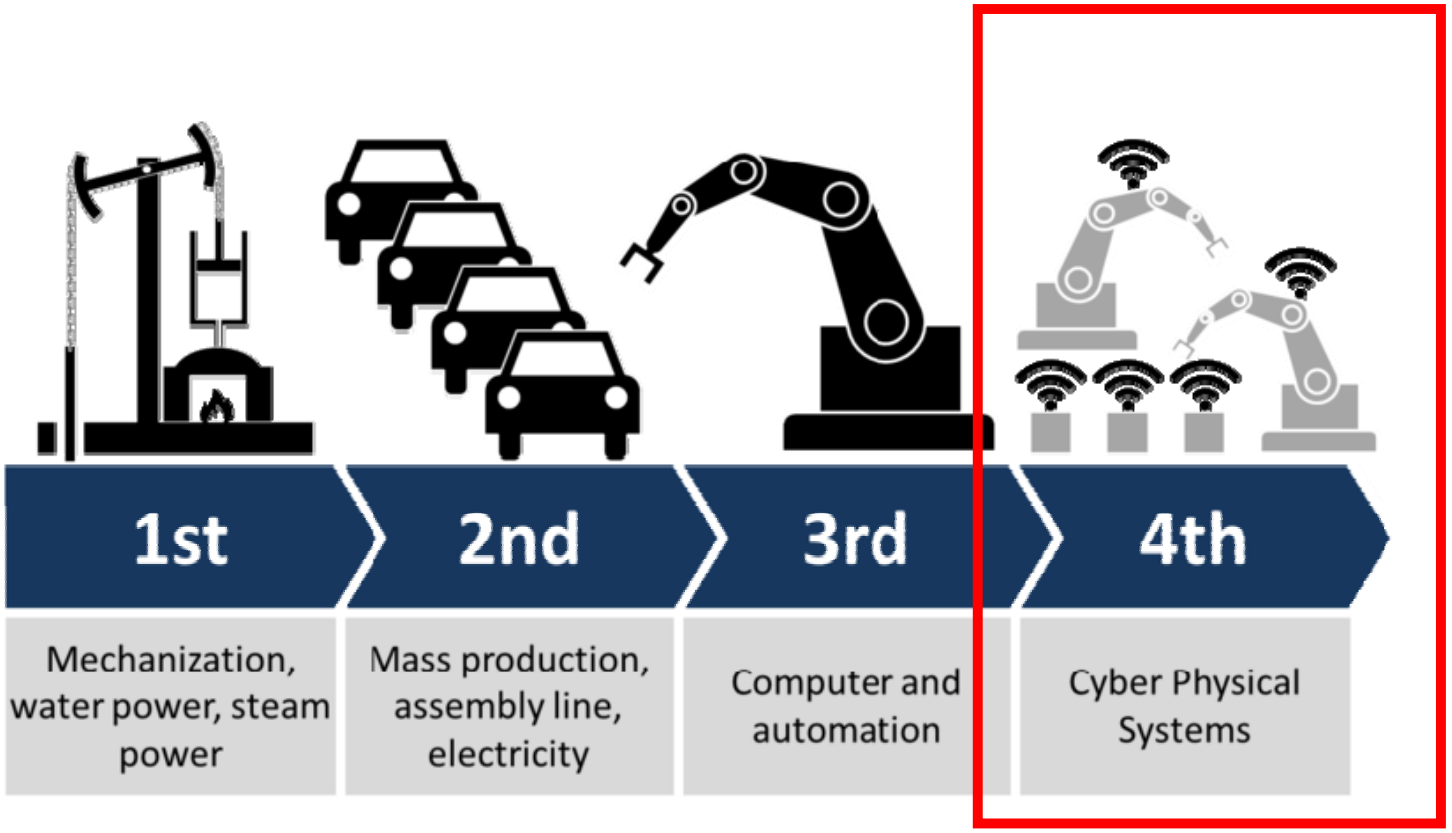
- Associate Professor, Department of Software Convergence Engineering, Kunsan National University (2018 ~ present)

My Long-Term Research Goal

- If a machine can write a book report ☺
  - Let me show Carl Sagan's "Cosmos" to it
  - It reads the book
  - It writes a one-page report
- I want to know why the universe exists (I have been curious since I was a high school student)
  - Because machines never sleep, let them study all the world's knowledge
  - Hope that it would show us any insight as to why we and the universe exist
- This is not going to happen in my life, but I want to make a small contribution anyway ~~ !!



History of Industrial Revolution (1/2)



History of Industrial Revolution (2/2)

- The First Industrial Revolution (1769)
  - James Watt invented the steam engine
- The Second Industrial Revolution (early 1900)
  - Mass production of Ford Motors
- The Third Industrial Revolution (1980 ~ 2010)
  - Information society through computers and Internet
- The Fourth Industrial Revolution (since 2010)
  - ??

The Fourth Industry Revolution

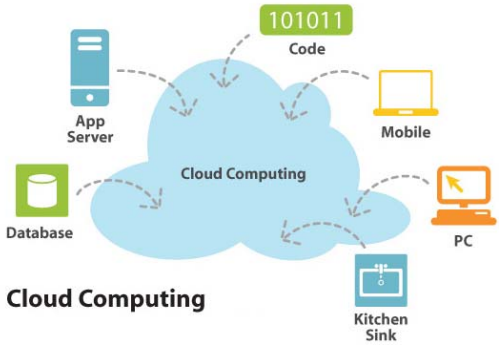
- Davos Forum in 2016
  - Klaus Schubert, President of the World Economic Forum
- Now is significantly different from the past …
  - Accelerating innovation through the Internet of Things, Big Data, and Artificial Intelligence
  - Economic center shifts from hardware industry to software industry
    - Cars: Engine (mechanical engineering) -> Batteries (electronic engineering)
    - Now, if you want to work for a car company, choose software major
  - Most companies turns into service companies
    - Manufacturer like GE -> Transformed into a software company
    - IT companies like Google -> Hardware manufacturer
  - Integration of physical space with virtual space
  - Convergence (interdisciplinary)
    - One study (law, chemistry, nursing, entrepreneur, …) + software
    - Improve technological innovation and productivity in the study

Internet of Things (IoT)

- Attaching a microcomputer to an object or tool used by humans
- Connect computerized objects to the Internet
- Exchange of information among objects connected to the Internet
- Better and intelligent services

Smart Devices

SNS



- Data Creation
  - Whenever, wherever
- Data Types
  - Location information, behavior information, transactions, text messages, images, text, links, ...
- IDC (according to Market Research)
  - Worldwide digital data volume
    - 1 Zeta Bytes (approximately 1 trillion GB)
  - Doubled every two years
    - In 2020, approximately 40 zeta bytes
    - Equivalent to about 57 times the amount of sand in the world's beaches)

Big Data Technologies

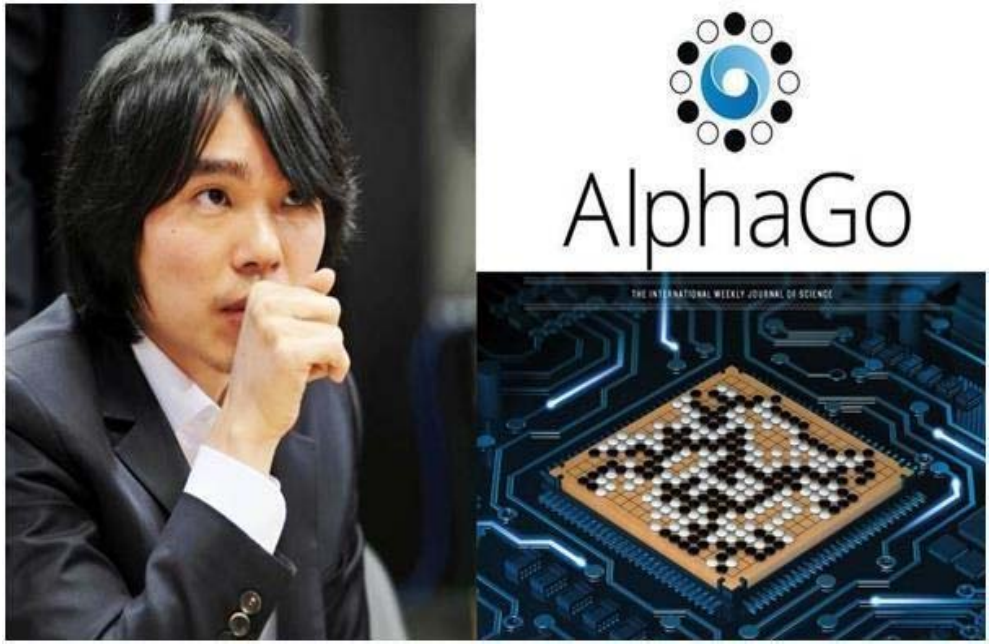
Big Data

Analysis

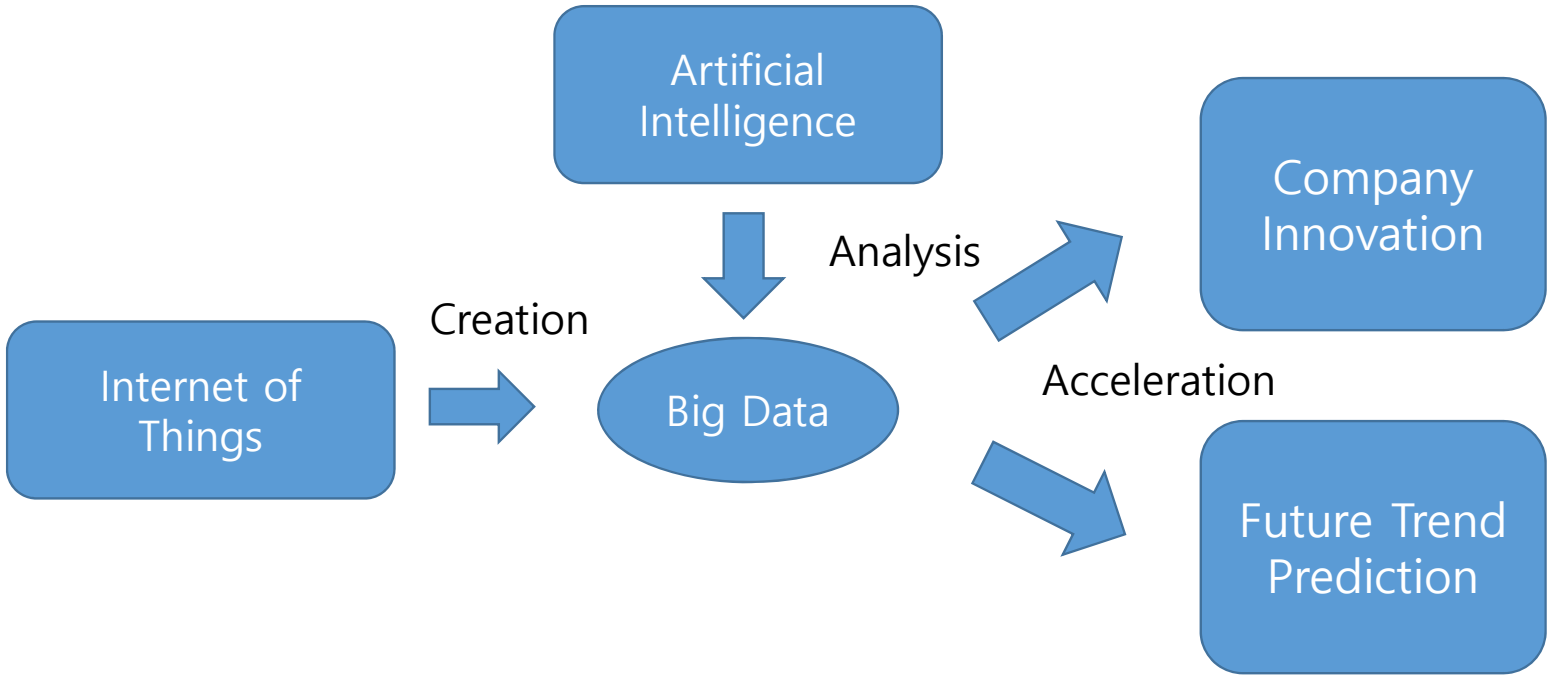
Through Big Data analysis, we want to become 21st century fortunetellers !!

Artificial Intelligence (AI)

- The art of computers thinking and behaving like humans
  - Self-driving cars, drones, robots



Relationship among IoT, Big Data, and AI



The Fourth Industry Revolution

- Davos Forum in 2016
  - Klaus Schubert, President of the World Economic Forum
- Now is significantly different from the past …
  - Accelerating innovation through the Internet of Things, Big Data, and Artificial Intelligence
  - Economic center shifts from hardware industry to software industry
    - Cars: Engine (mechanical engineering) -> Batteries (electronic engineering)
    - Now, if you want to work for a car company, choose software major
  - Most companies turns into service companies
    - Manufacturer like GE -> Transformed into a software company
    - IT companies like Google -> Hardware manufacturer
  - Integration of physical space with virtual space
  - Convergence (interdisciplinary)
    - One study (law, chemistry, nursing, entrepreneur, …) + software
    - Improve technological innovation and productivity in the study

A competition among automakers at CES ... The acceleration of automobiles ' home appliances

Why did the world's top automakers participate in the CES, the world's largest consumer electronics exhibition?

The answer is simple. In fact, cars are becoming home appliances that move by themselves. In addition, as next-generation cars such as self-driving cars and electric cars became home appliances, cooperation with other industries such as electric and electronic industries is inevitable. The fact that a car has become capable of TVs and smartphones also plays a role.

The Fourth Industry Revolution

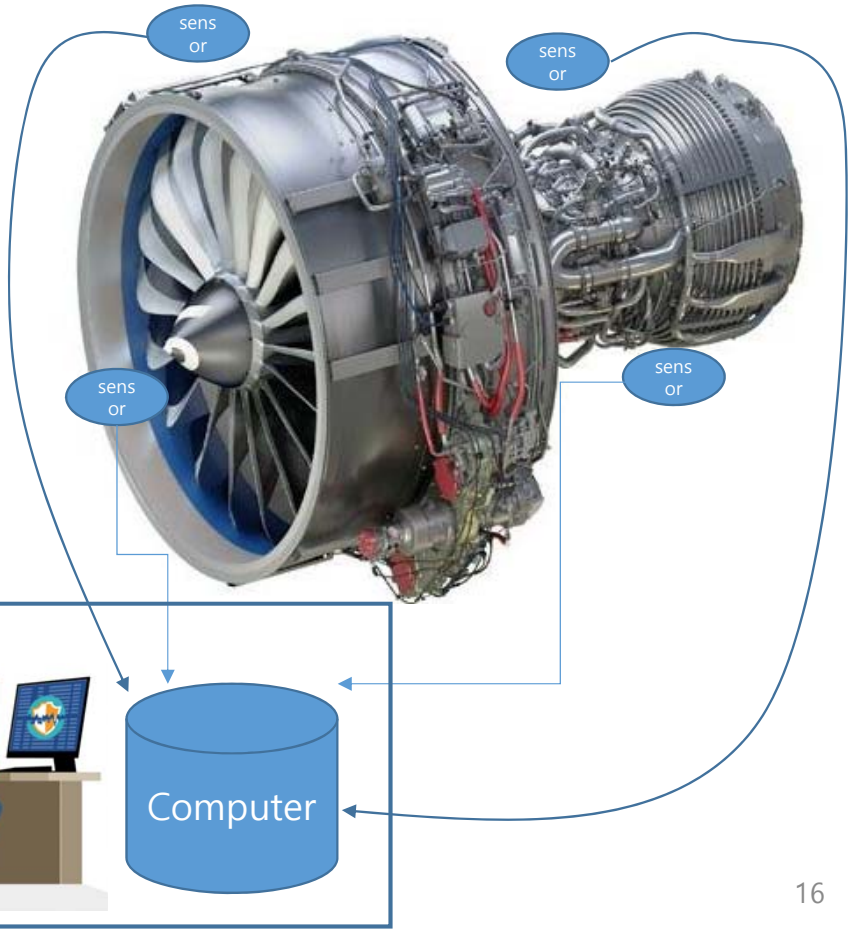
- Davos Forum in 2016
  - Klaus Schubert, President of the World Economic Forum
- Now is significantly different from the past …
  - Accelerating innovation through the Internet of Things, Big Data, and Artificial Intelligence
  - Economic center shifts from hardware industry to software industry
    - Cars: Engine (mechanical engineering) -> Batteries (electronic engineering)
    - Now, if you want to work for a car company, choose software major
  - Most companies turns into service companies
    - Manufacturer like GE -> Transformed into a software company
    - IT companies like Google -> Hardware manufacturer
  - Integration of physical space with virtual space
  - Convergence (interdisciplinary)
    - One study (law, chemistry, nursing, entrepreneur, …) + software
    - Improve technological innovation and productivity in the study


Data analytics and management services are 75 % of revenue
GE become a software company by 2020

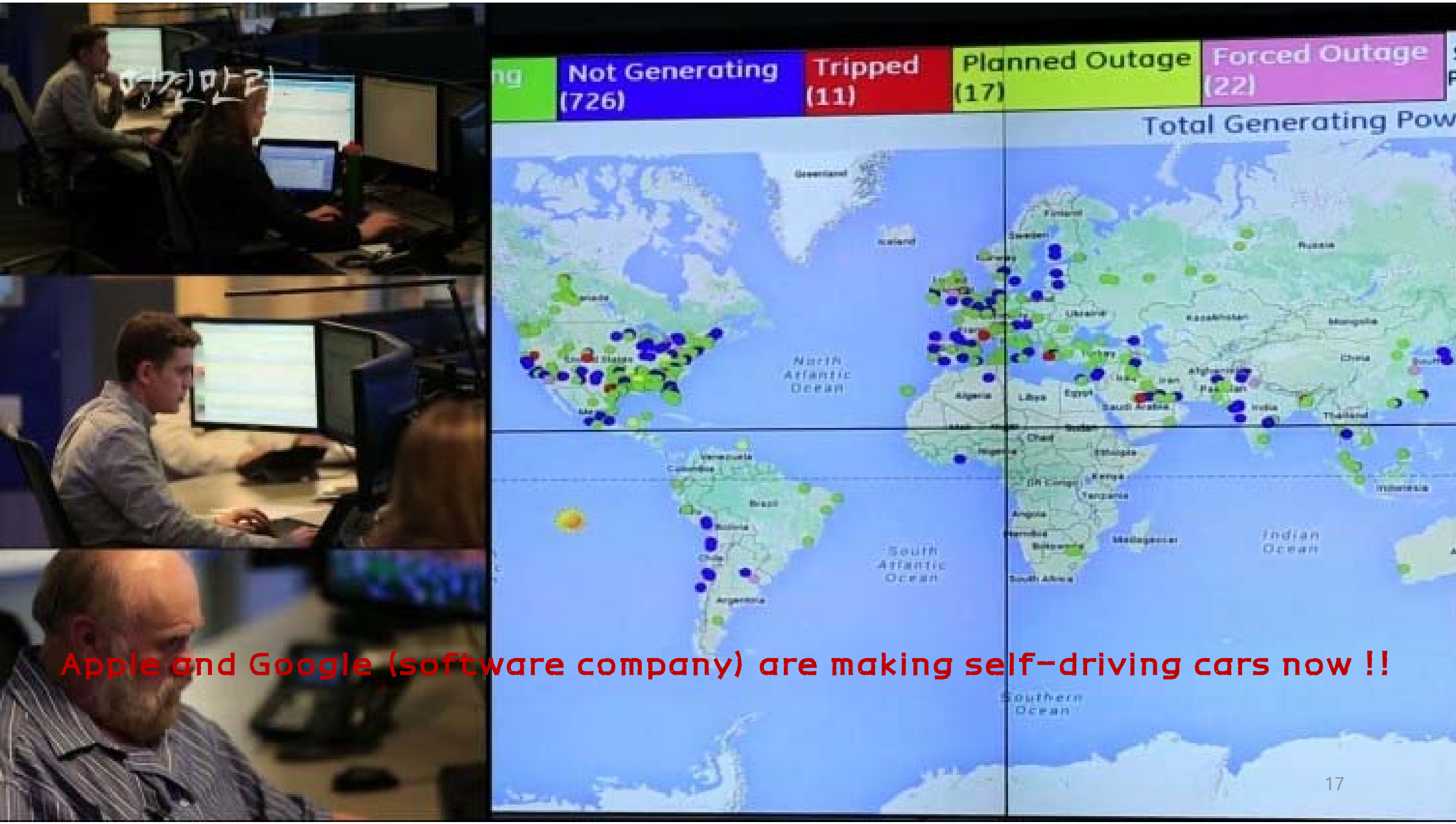
Past

Real-time transfer of anomaly in the aircraft engine

Current



GE business model


Not Generating
(726)
Tripped
(11)
Planned Outage
(17)
Forced Outage
(22)
Total Generating Pow
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Southern Ocean
Apple and Google (software company) are making self-driving cars now !!

The Fourth Industry Revolution

- Davos Forum in 2016
  - Klaus Schubert, President of the World Economic Forum
- Now is significantly different from the past …
  - Accelerating innovation through the Internet of Things, Big Data, and Artificial Intelligence
  - Economic center shifts from hardware industry to software industry
    - Cars: Engine (mechanical engineering) -> Batteries (electronic engineering)
    - Now, if you want to work for a car company, choose software major
  - Most companies turns into service companies
    - Manufacturer like GE -> Transformed into a software company
    - IT companies like Google -> Hardware manufacturer
  - Integration of physical space with virtual space
  - Convergence (interdisciplinary)
    - One study (law, chemistry, nursing, entrepreneur, …) + software
    - Improve technological innovation and productivity in the study

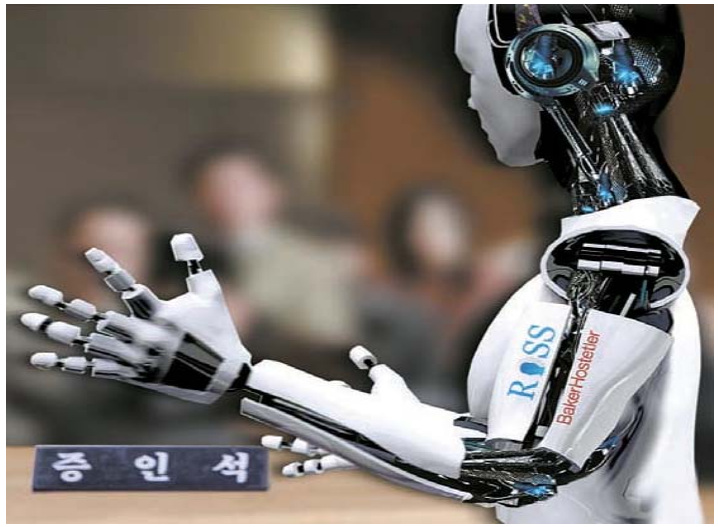
Integration of Physical and Cyber Spaces

The Fourth Industry Revolution

- Davos Forum in 2016
  - Klaus Schubert, President of the World Economic Forum
- Now is significantly different from the past …
  - Accelerating innovation through the Internet of Things, Big Data, and Artificial Intelligence
  - Economic center shifts from hardware industry to software industry
    - Cars: Engine (mechanical engineering) -> Batteries (electronic engineering)
    - Now, if you want to work for a car company, choose software major
  - Most companies turns into service companies
    - Manufacturer like GE -> Transformed into a software company
    - IT companies like Google -> Hardware manufacturer
  - Integration of physical space with virtual space
  - Convergence (interdisciplinary)
    - One study (law, chemistry, nursing, entrepreneur, …) + software
    - Improve technological innovation and productivity in the study

Convergence Technologies

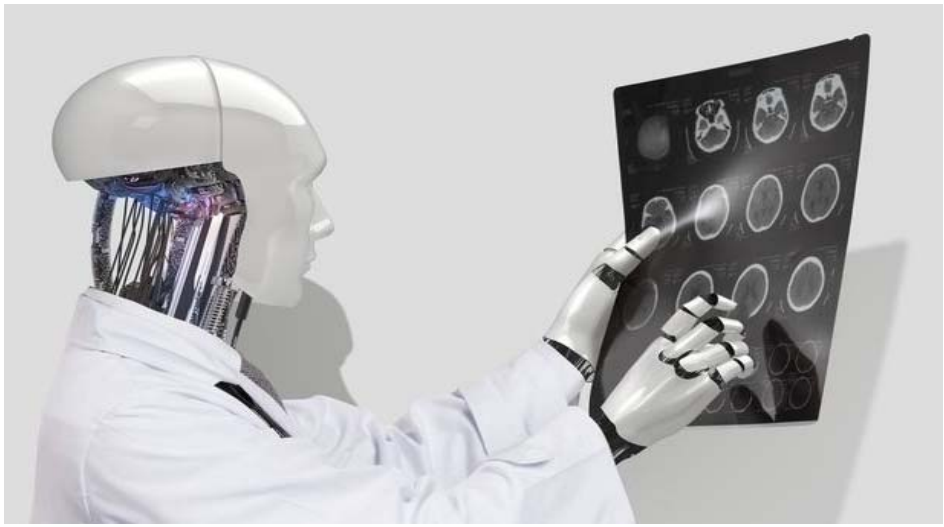
Artificial Intelligence Attorney Ross gets job at New York Law Firm



Law + Software

Analysis of a million volumes of legal documents

Artificial Intelligence Doctor Watson gets Job at Gacheon University - Gil Hospital



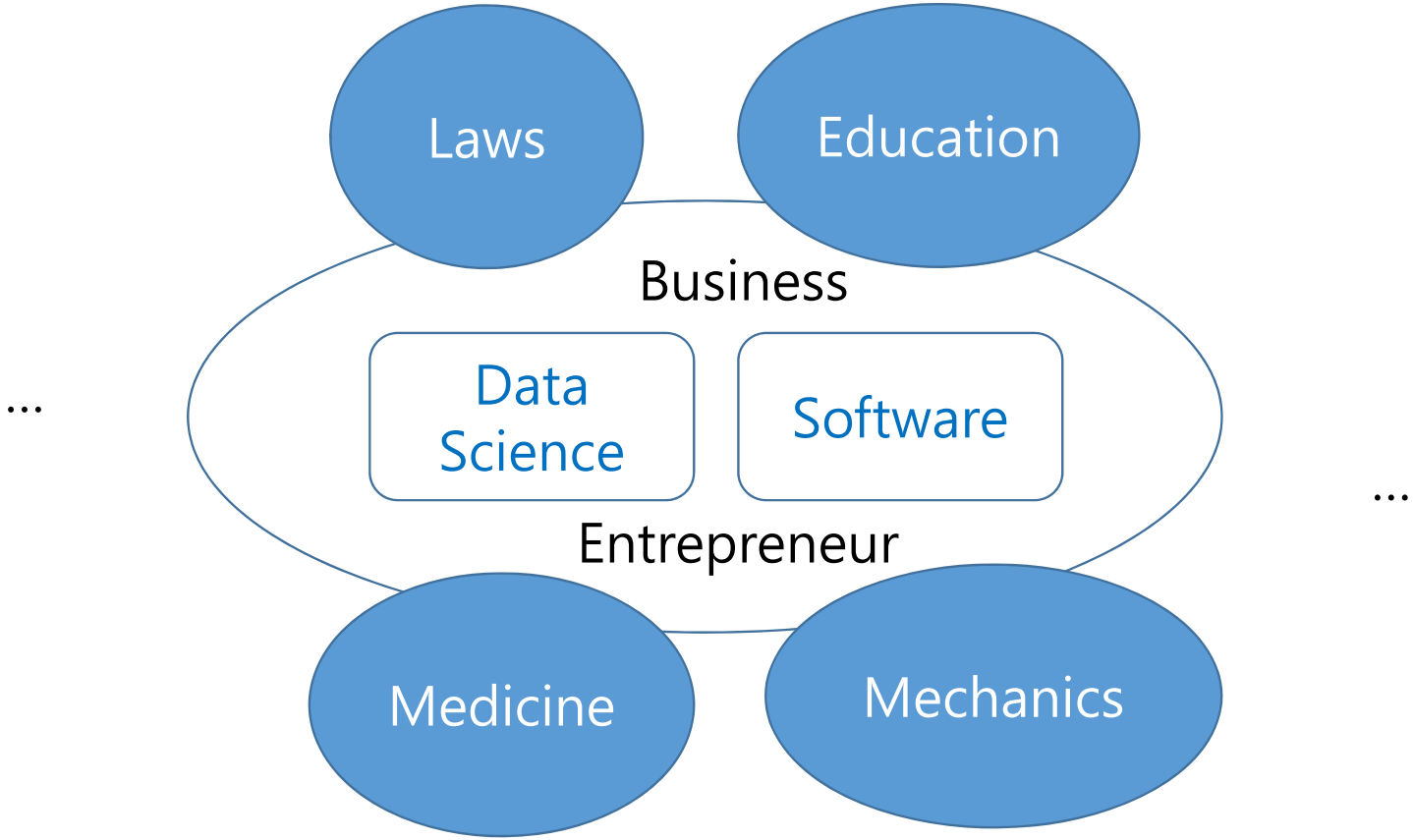
Medicine + Software

Gain knowledge of 15 million pages of cancer

Why should we pay attention to "data"?

- Core of the Fourth Industrial Revolution: Convergence
  - Each study (law, chemistry, nursing, ...) + software
- Increase technological innovation and productivity in the study
- After all, the most important thing in the fourth industrial revolution is "data analysis"
- Data Analysis Methods
  - Data Mining : Analyze data in the database (find patterns or rules of data)
  - Machine Learning : Mathematics-based data analysis
  - AI : Analysis of data by techniques mimicking the human brain
- Skills for Data Analysis
  - Data processing : C/C++/Java/Python, data structures, computer algorithms
  - Large data processing : Hadoop, MapReduce, Spark
  - Unstructured data processing : Natural language processing, image processing
  - Data analysis methods

In This Lecture, You Can Learn Basic Concept about Data Science, especially Big Data !!

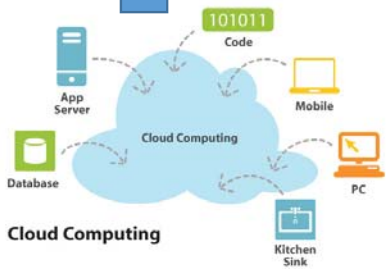


Digital Data

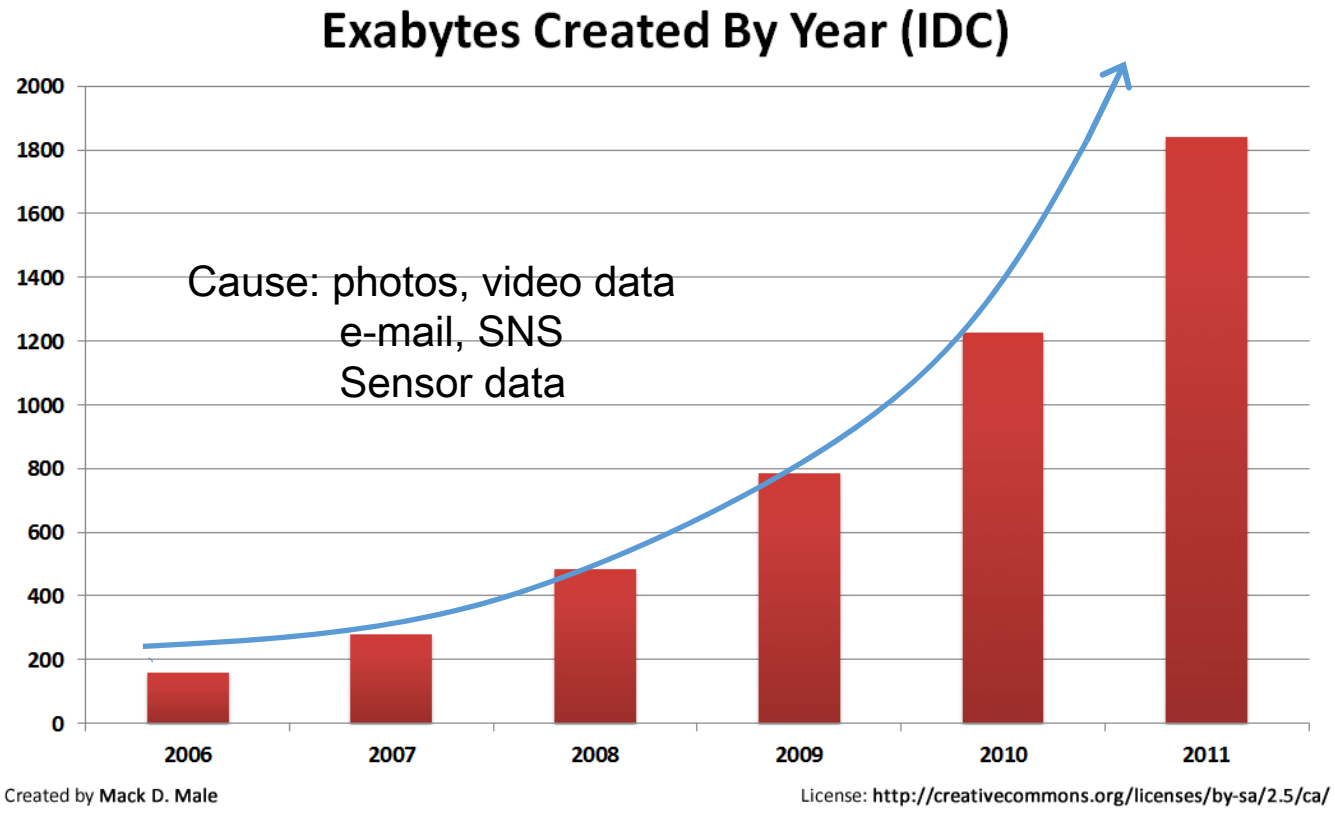
1 byte : One character (number) representation

- mega = 10^6 = 1,000,000
- giga = 10^9 = 1,000,000,000
- terra = 10^{12} = 1,000,000,000,000
- peta = 10^{15} = 1,000,000,000,000,000
- exa = 10^{18} = 1,000,000,000,000,000,000
- zetta = 10^{21} = 1,000,000,000,000,000,000,000
- iotta = 10^{24} = 1,000,000,000,000,000,000,000,000

Emergence of new technologies and services



Information Explosion



McKinsey Big Data Report

- Released in May 2011
- Big Data: The next frontier for innovation, competition, and productivity
- Big Data is not a problem but a new challenge

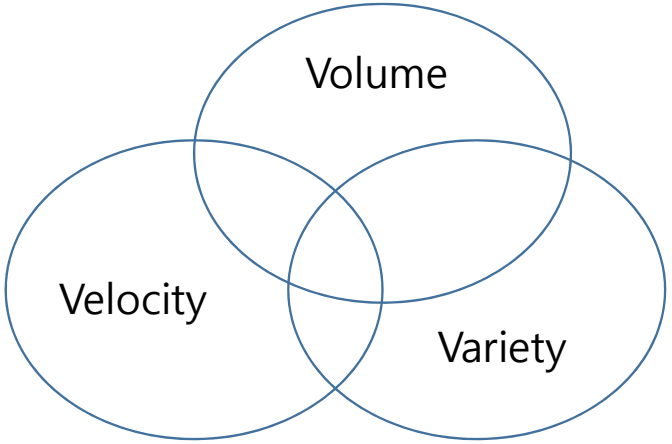
What is Big Data?

McKinsey&Company

- A huge amount of data that cannot be stored and processed in the current database technologies

Gartner

- 3V: Volume, Variety, Velocity



IBM

- 3V + Veracity = 4V

ORACLE

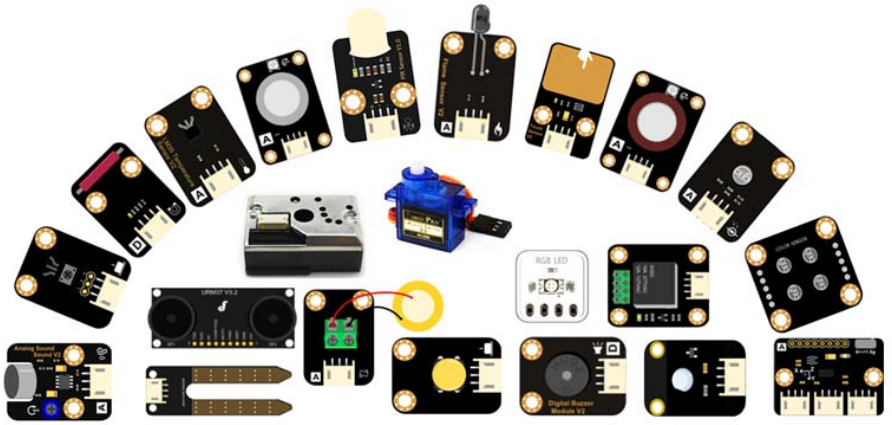
- 4V + Value = 5V

Big Data Property : Volume

- Data size
- Large amount of data that can not be processed with current technology
- Decision
  - More than 10 TB (Tera Bytes) = 10,000,000,000,000 bytes
    - cf) 1 byte: one character (numeric) representation

Big Data Property : Velocity

- Data creation speed
  - Data generated in seconds
- Data collection through sensors
  - Fine dust, vibration, temperature and humidity, gyroscope, ...
- Accelerating by Internet of Things



Big Data Property : Variety

Type	Description	Example
Structured Data	Rows, columns	-Database -Excel Spreadsheet
Unstructured Data	Keys, values	-Text -Voice -Image -Moving picture -Social network

Structured Data

- Data consisting of rows and columns
- Machine-understandable format
- Example

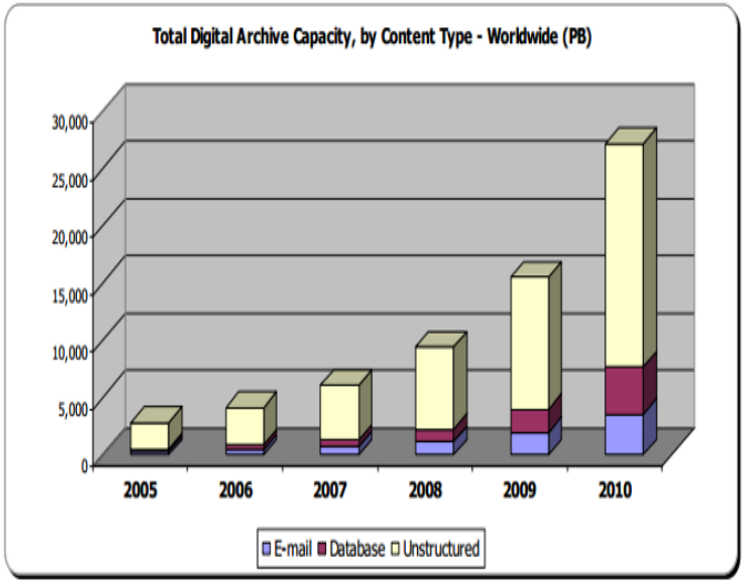
Student ID	Name	Phone	Email
1711021	Minji Kim	010-6134-3568	kimmj@kunsan.ac.kr
1711041	Seri Na	010-9865-7622	naseri@kunsan.ac.kr
1711055	Boeun Suh	010-1373-2489	suhbe@kunsan.ac.kr
1711083	Eunji Lee	010-2832-7890	leeji@kunsan.ac.kr

Unstructured Data

- Voices, Images, Moving pictures, Texts

DATA INTELLIGENCE LAB

Welcome to the home of DILAB (**Data** Mining & Artificial **Intelligence Lab**oratory) in Department of Software Convergence Engineering, Kunsan National University, Gunsan, Jeollabuk-do, Korea. Now, it is obvious that data is the new oil of the 21st century. We also believe that we will be able to achieve the innovation of technology based on Data-driven Artificial Intelligence, of which the goal is the study of extracting insights hidden in *raw* data and proposing state-of-the-art data-driven applications. Today, we are still enjoying resolving challenging problems related to data science and engineering areas for paradigm shift and our better life in the near future. Our primary research fields lie in *text data mining*, *natural language processing*, *big data*, and *artificial intelligence* in recent time.



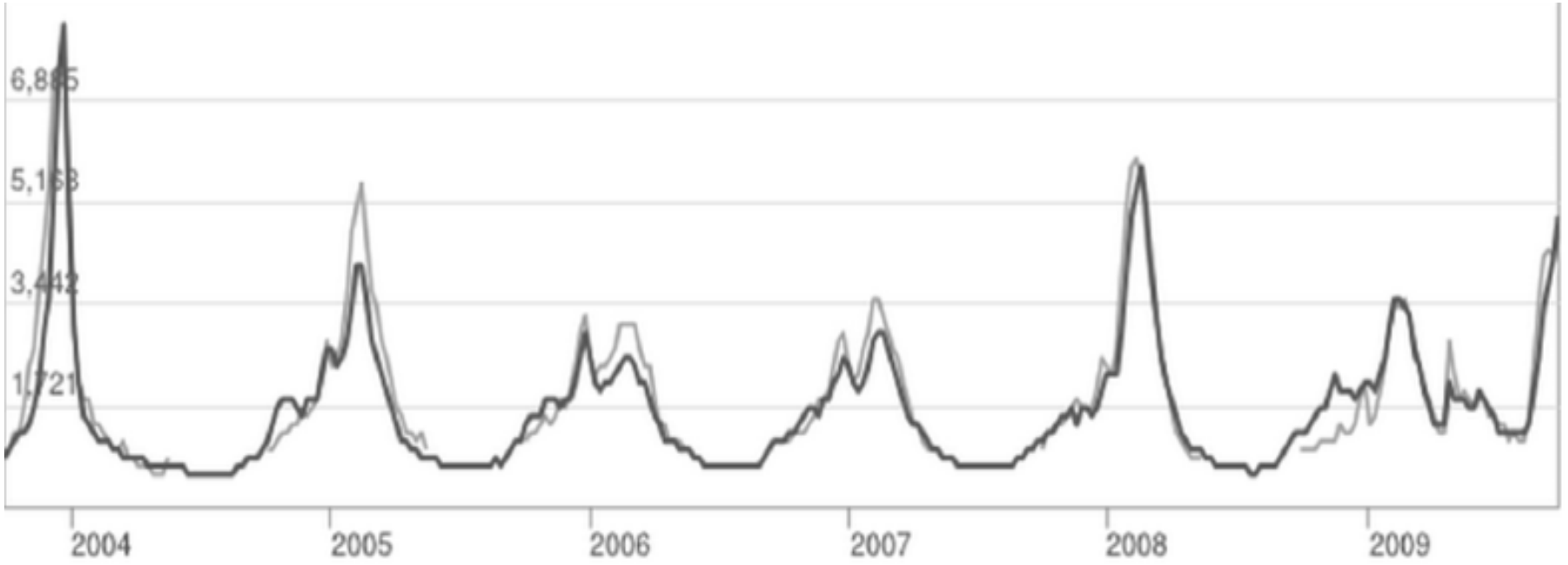
More than 90% of all currently generated data is unstructured data

Examples of Big Data

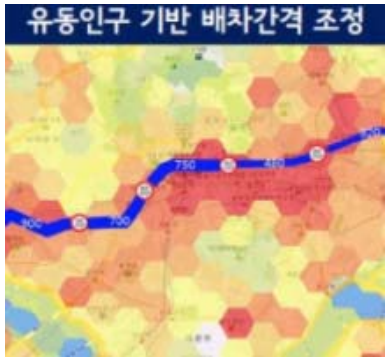
- Facebook
  - Generate 33 billion monthly content
- YouTube
  - One hour of video registration per second
- Manufacturing (Semiconductor)
  - Samsung Electronics: Generate 600TB log data per year
- SK Telecom
  - Store usage history of 10 million customers
- Walmart
  - 1 million transactions per hour
- …

Google Trend

- Gaining influenza information by searching "flu" words included in Google emails



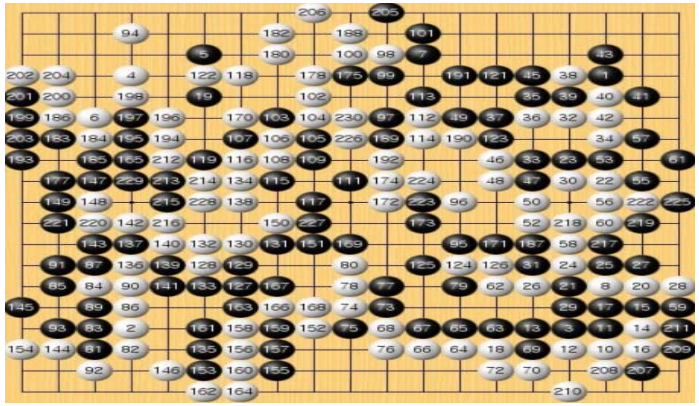
Seoul Midnight Bus Route Optimization



Target's Big Data Strategy

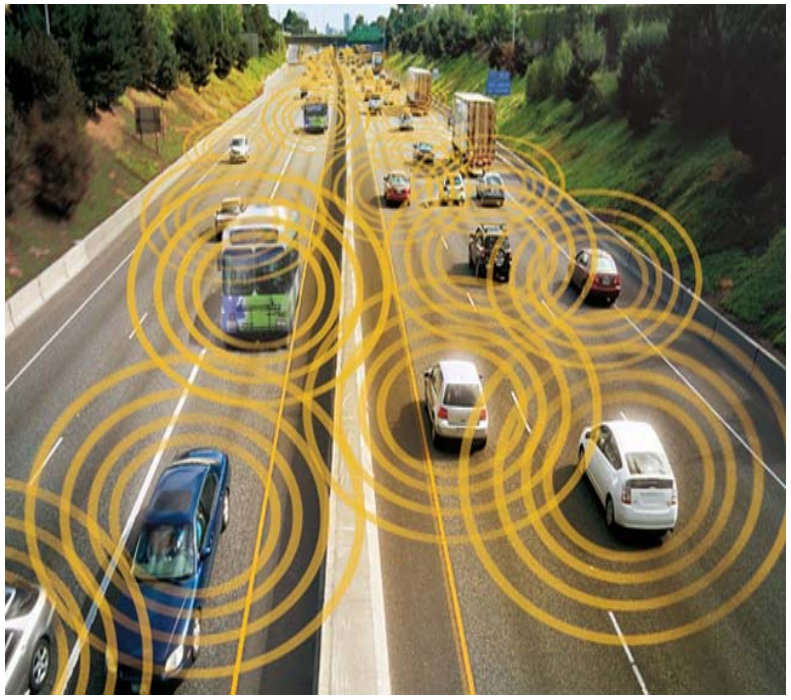
- Insight from Target's Big Data analysis
  - Early pregnancy
    - Buy calcium and magnesium supplements
  - Middle of pregnancy
    - Odorless Lotion
  - Birthday
    - Buy fragrance soap, detergent, cotton balls
- Target's new business model
  - Predict customers' pregnancy in advance
  - Give promotion (discount coupons) to pregnant women

Lee Sedol vs AlphaGo



- The game of go
  - Number of cases in which a stone can be placed
  - 250^{150} wins
    - cf) All atoms in the universe (about 10^{80} wins)
- Learning the past 30 million mark of the first-class professional Go article
- Twin alpaca programs and over 100,000 titles
- Alpha Go is one million times and takes four weeks to learn
  - cf) People take 1,000 years
- Alpha has recently improved its skills through 4 million times

Connected Car



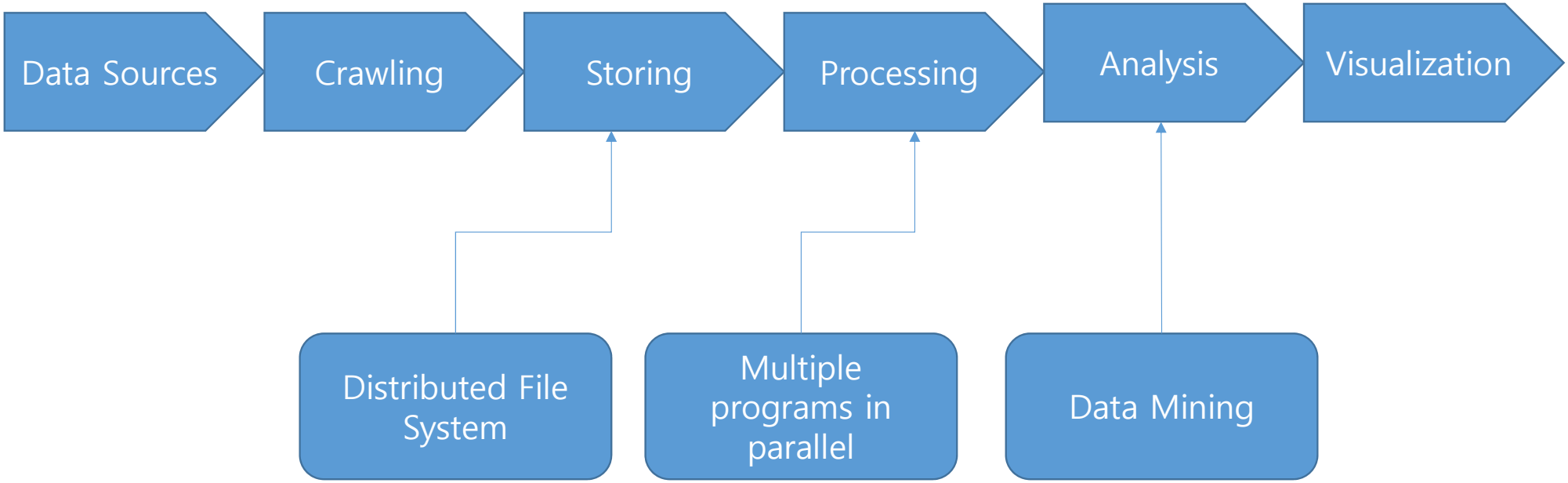
- Connected car
  - Car version of IoT
  - Car is not a machine, but a "household appliance"
- Information collected in seconds
  - Vehicle operation information (engine, brake, noise ...)
  - Biometric information of the driver (heartbeat, stress, ...)
  - Video information through 3D camera
  - Location information such as GPS
  - Information from Internet
  - Communication information with other vehicles
  - ...

Why is big data important?

- Utilization of unused data
  - We can see what we did not see before
- Enable fact-based decisions
  - Use analysis results
- New approach, mind-set requirement
  - Re-examine the use of IT technology
- Start from analysis of small data

Big Data Processing Phases

- To collect, store, process, analyze, and visualize Big Data



Basic Idea of Big Data

✓ Big Data Solution

→

Scale up

- ✓ High Cost
- ✓ Low Scalability
- ✓ Single Computer

Scale out

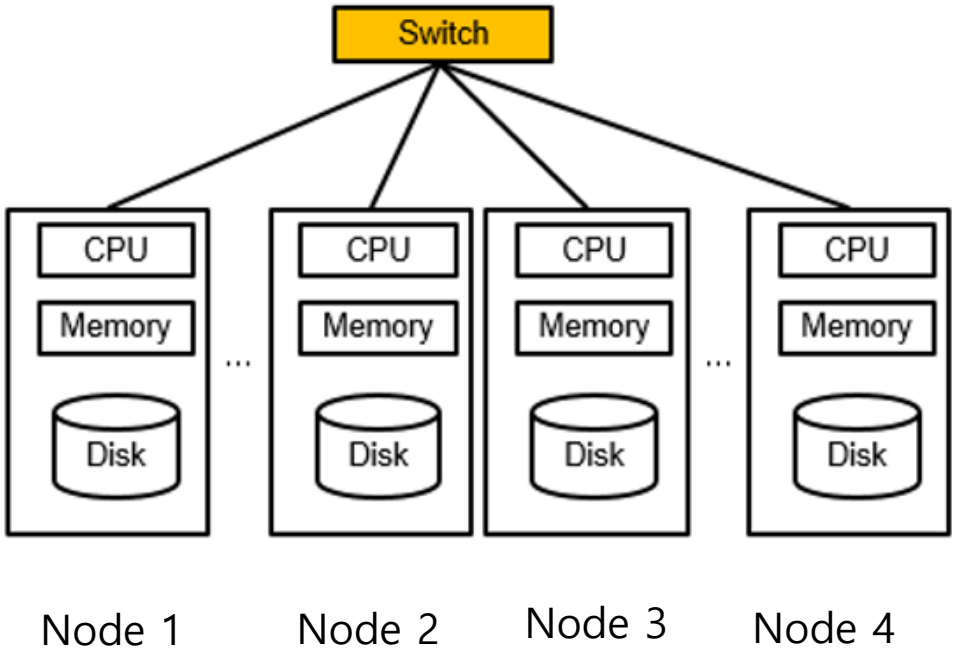
- ✓ Low Cost
- ✓ High Scalability
- ✓ Parallel Computer

✓ Scale out Solution

- Distributed computing
- Failover
- Easy parallel programming

Computer Hardware for Big Data

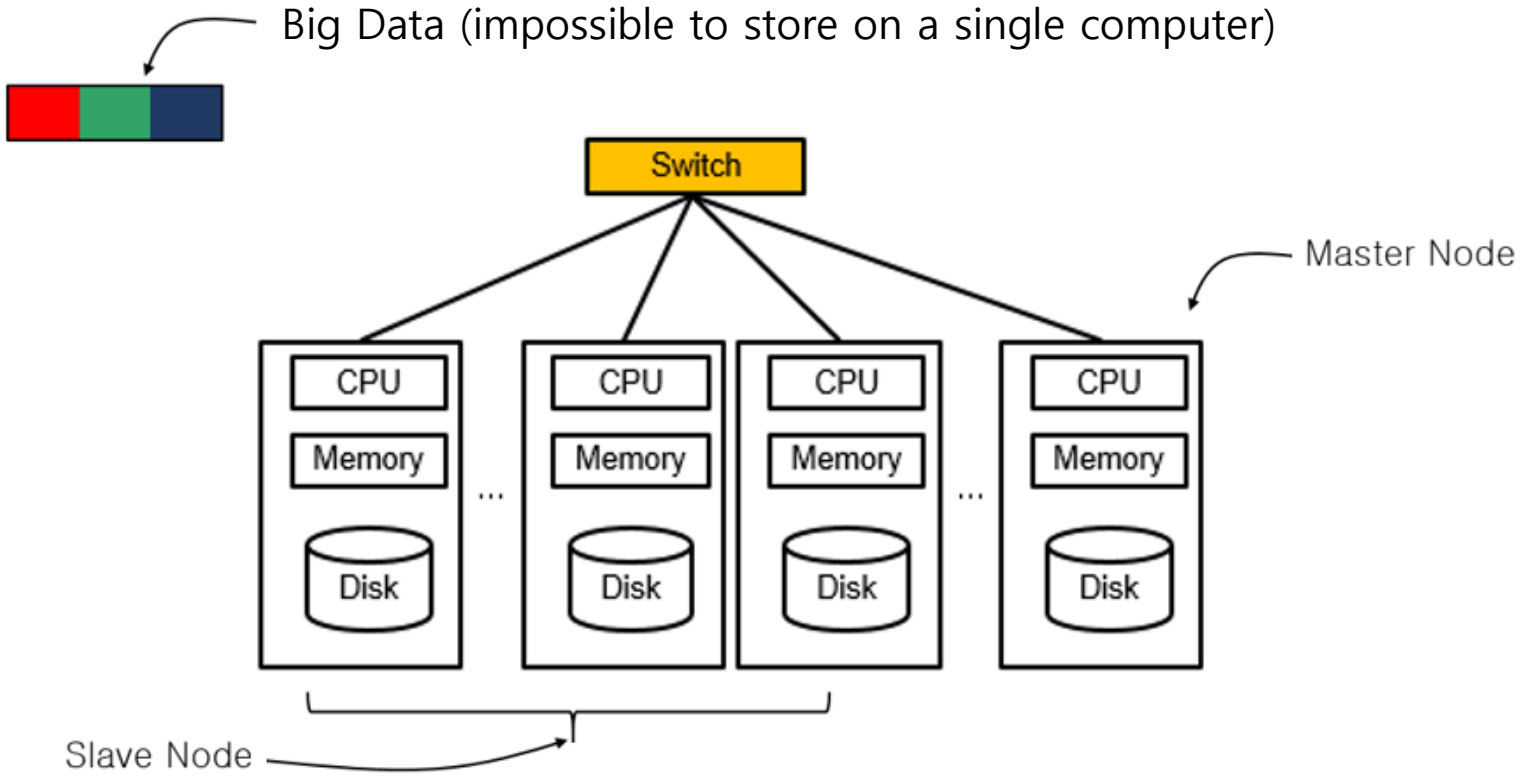
- Cluster system or distributed processing system



Computer Software for Big Data

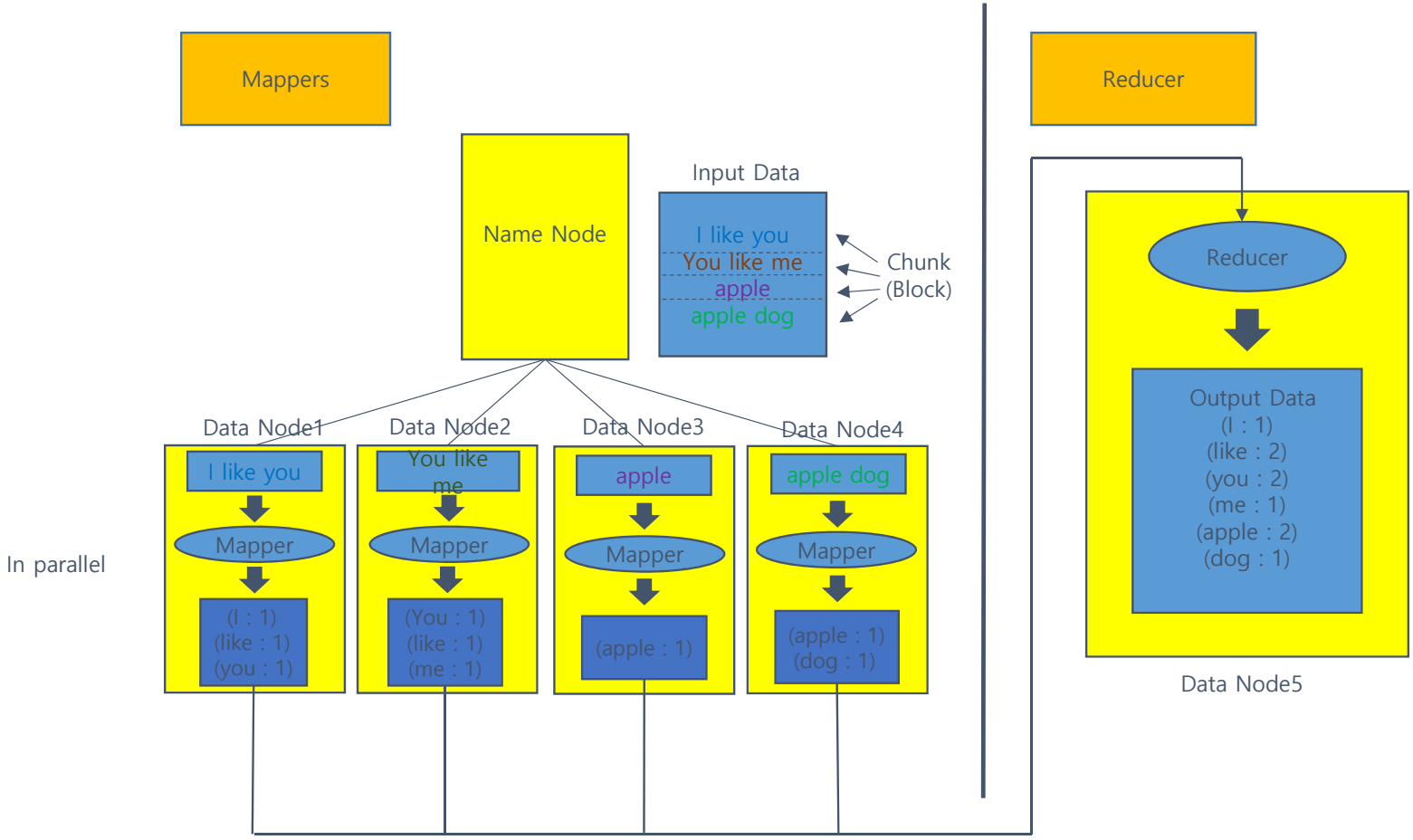
- Hadoop: High Availability Distributed Object-Oriented Platform
- Founder of Apache Lucene: Doug Cutting
- Open source distributed processing technology project
- Used in Yahoo, Facebook, etc.
- Major components (Hadoop Eco-system)
  - The Hadoop Distributed File System (HDFS)
  - MapReduce programming
  - HBase database
  - Pig / Hive
  - …

Distributed File System



Distributed computing : A system that organizes low-cost, general-purpose computers into high-performance networks and acts like a single computer

MapReduce: Running Processes in Parallel



Big Data Analysis

- Data Mining
  - Find meaningful knowledge, such as patterns or rules of formal data stored in the database
  - Use as resources for management activities
- Machine Learning
  - Computers automatically analyze data to find patterns or predict future
  - Statistics and mathematics base
- Artificial Intelligence
  - Analyze data, imitate human brain action, find patterns or rules, predict future
  - Autonomous unmanned vehicles, drones, robots

Data Mining

		EXPLANATORY VARIABLES													
UNIQUE_ID	USER DEFINED	CUSTOMER			VOICE CALL USAGE (MONTHLY AGGREGATES)						DATA USAGE (MONTHLY AGGREGATES MB)				CHURN
LINE_NUMBER	REFERENCE_DATE	AGE_YEARS	GENDER	TENURE_MTHS	CALL_CNT_M0	CALL_CNT_M1	CALL_CNT_M2	CALL_DUR_M0	CALL_DUR_M1	CALL_DUR_M2	DATA_M0	DATA_M1	DATA_M2	...	TARGET
6132435172	2016-06-30	24	Female	12	3	6	11	120	557	538	337	1146	578	...	0.2856
6132461613	2016-06-30	56	Male	7	12	34	4	248	389	640	2585	2845	2469	...	-0.1945
6132464181	2016-06-30	18	Male	9	26	20	35	319	279	170	228	3700	5618	...	0.0024
6132465666	2016-06-30	22	Female	10	6	7	6				597	256	149	...	1.4896
6132470392	2016-06-30	51	Female	18							3990	259	3890	...	-0.7267
6132470615	2016-06-30	29		5										...	0.6678
6132471047	2016-06-30													...	1.0256
6132472127	2016-06-30													...	-0.0049
6132499775	2016-06-30													...	
6132500423	2016-06-30													...	
6132510447	2016-06-30													...	
...	...													...	
...	...													...	

Model Apply Analytical Data Set

User-Defined Reference Date = 2016-06-30

To apply the model every month, increase reference date + 1 month to update the explanatory variables in the correct time frame.

SAP PREDICTIVE ANALYTICS

Apply Classification Model

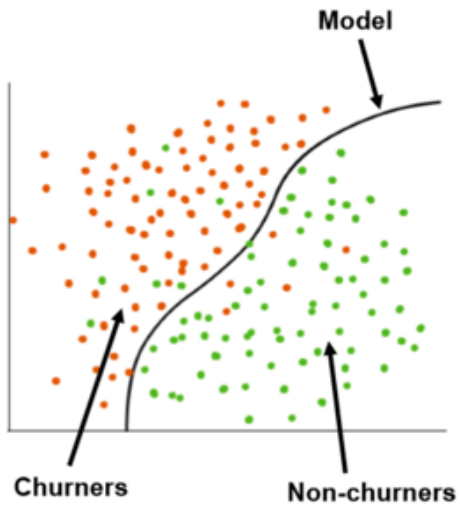
$Y = a + b_1 * x_1 + b_2 * x_2 + \ldots b_n * x_n$

Apply the model to calculate a score based on all of the explanatory variables for each unique ID.

- A high score indicates that the unique ID has a high potential to churn.
- A low score indicates that the unique ID has a low potential to churn.

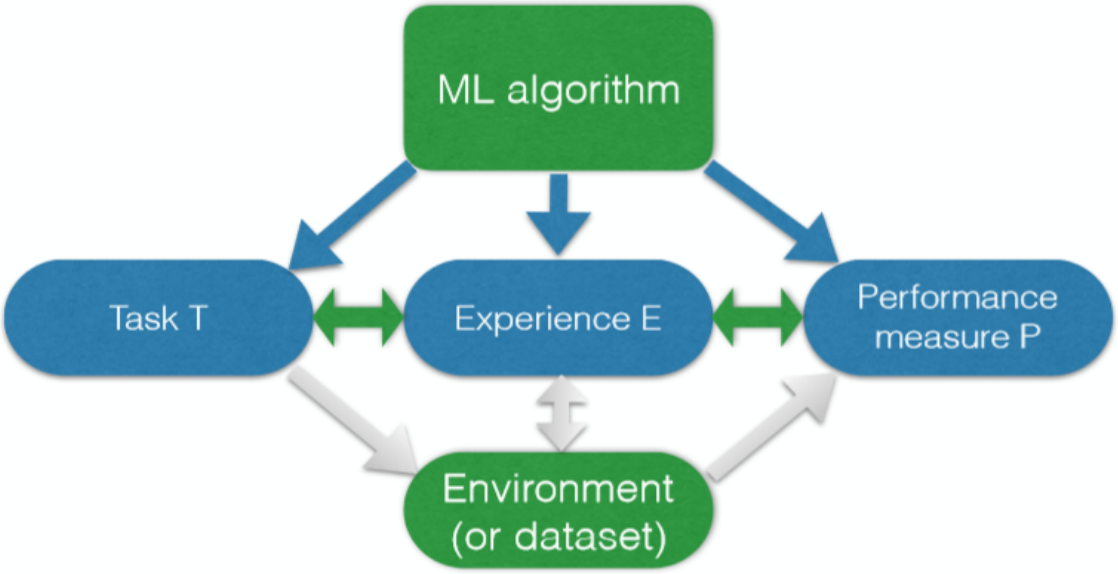
Please notice that the score is simply the output from the model equation. It can have negative values, and can have values greater than 1.

It is not a probability.



Machine Learning: core idea

"A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in **T**, as measured by **P**, improves with experience **E**." (Mitchell, 1997)



Next, we need to specify what we mean by **Tasks T**, **Performance measure P** and **Experience E**

Performance measure P

Typically, performance measure **P** is specific to the task **T**

One possible choice for **classification tasks**

$$Error\ rate = \frac{N_{incorrectly\,classified}}{N_{total}} \Leftrightarrow Accuracy = \frac{N_{correctly\ classified}}{N_{total}} = 1 - Error\ rate$$

Error rate = expected 0-1 loss

The main problem with the the 0-1 loss is that it is not differentiable. A smooth version is available for probabilistic model: the **log-probability** given by the model to training examples

Possible choices for **regression tasks**:

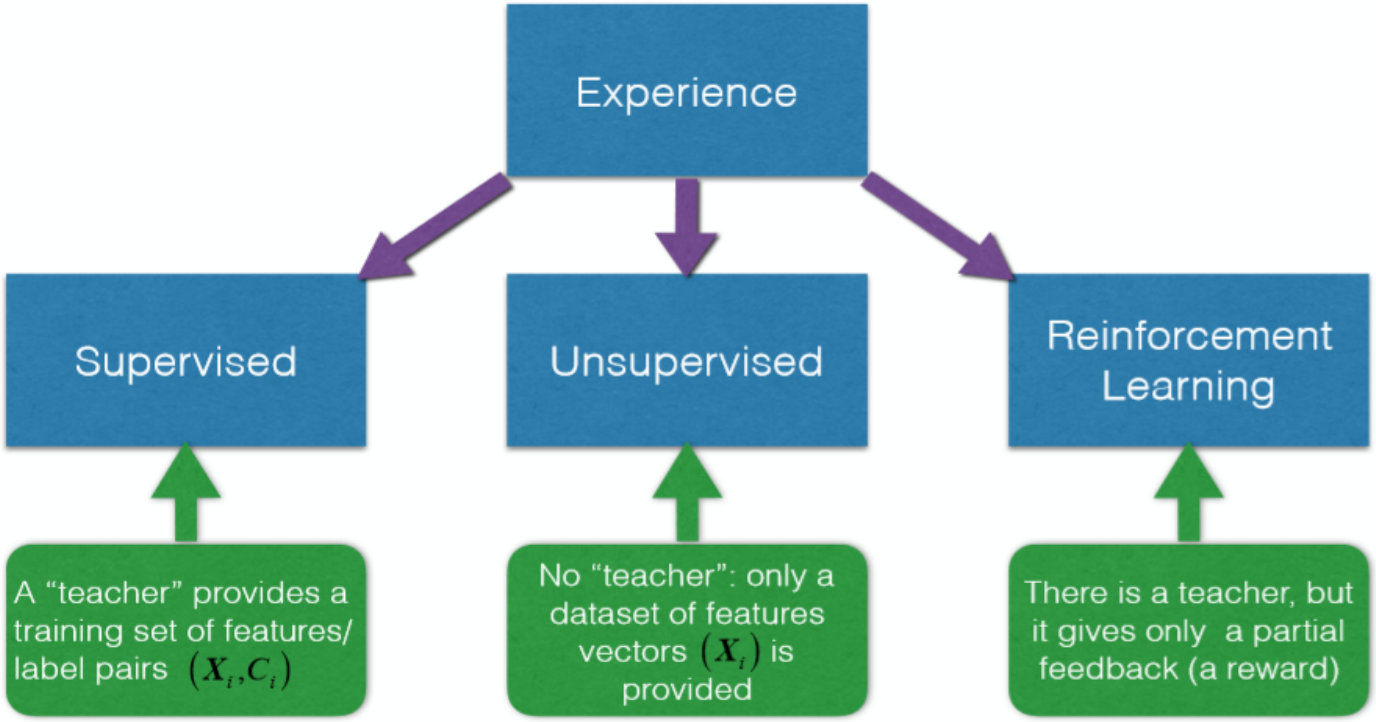
Mean square loss: $$L = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)^2$$

L1-loss: $$L = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$$

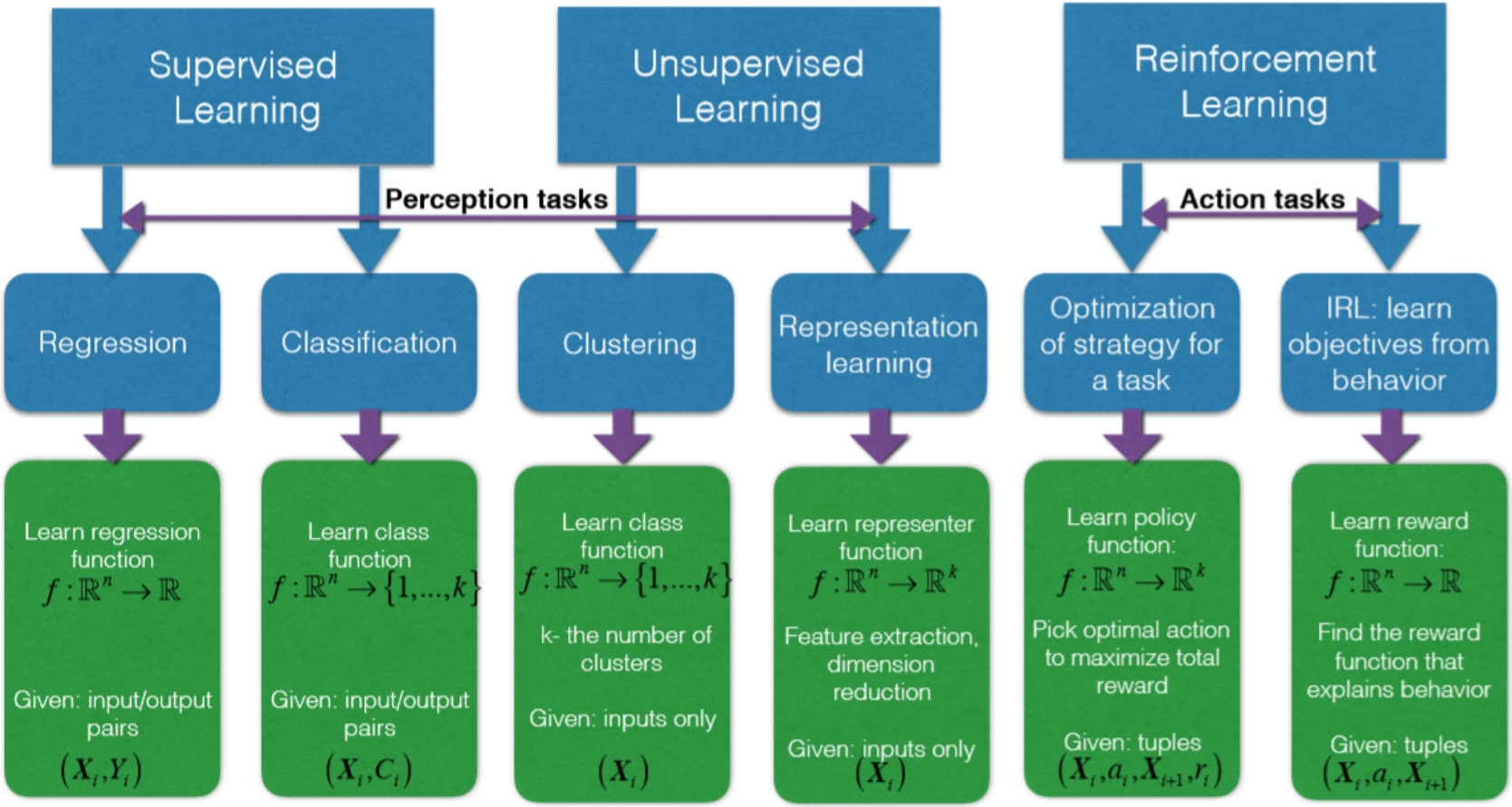
Learning from Experience E

The performance measure **P** improves with Experience **E** as a result of learning

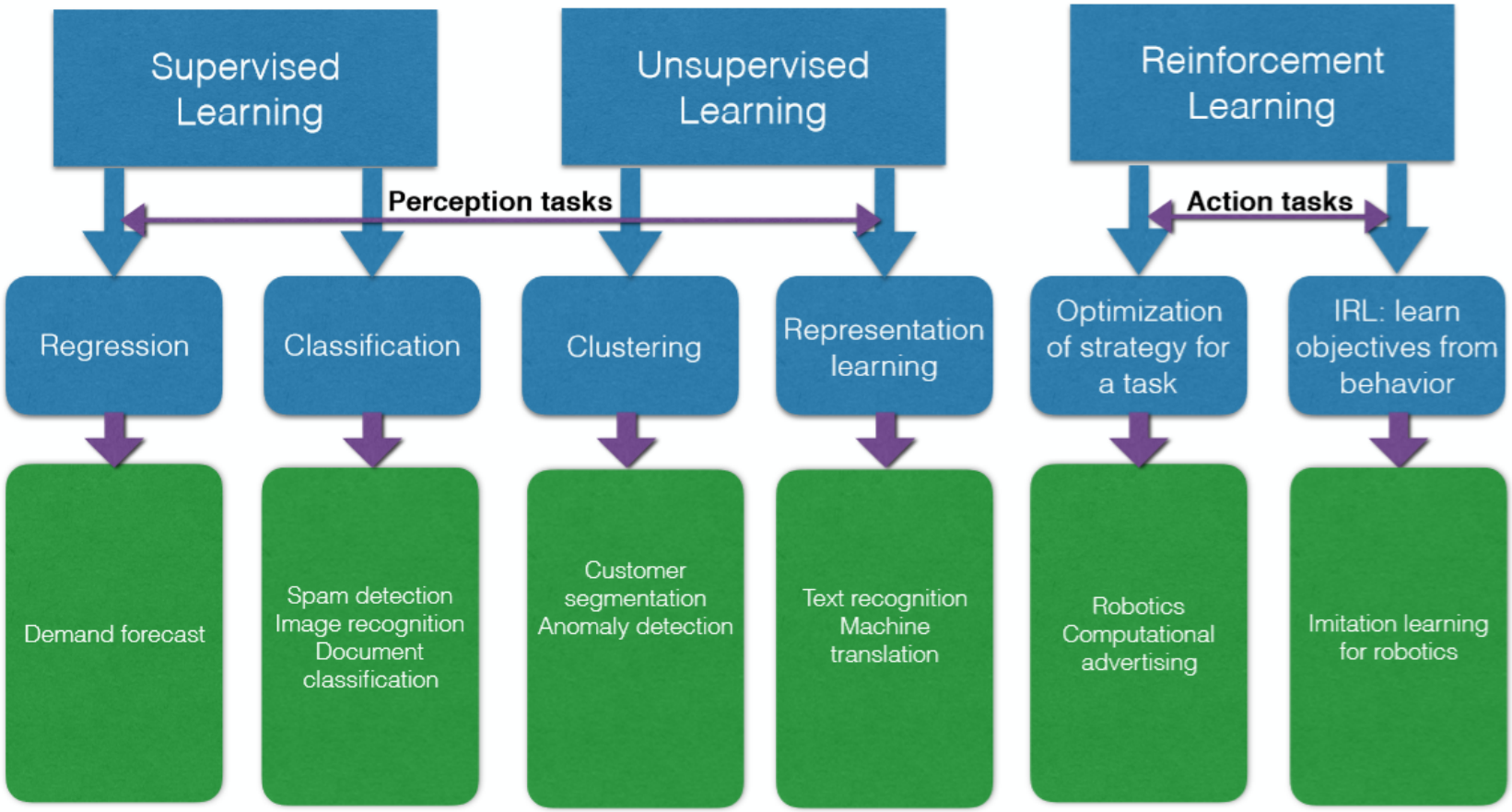
Types of learning from experience **E**



Machine Learning landscape

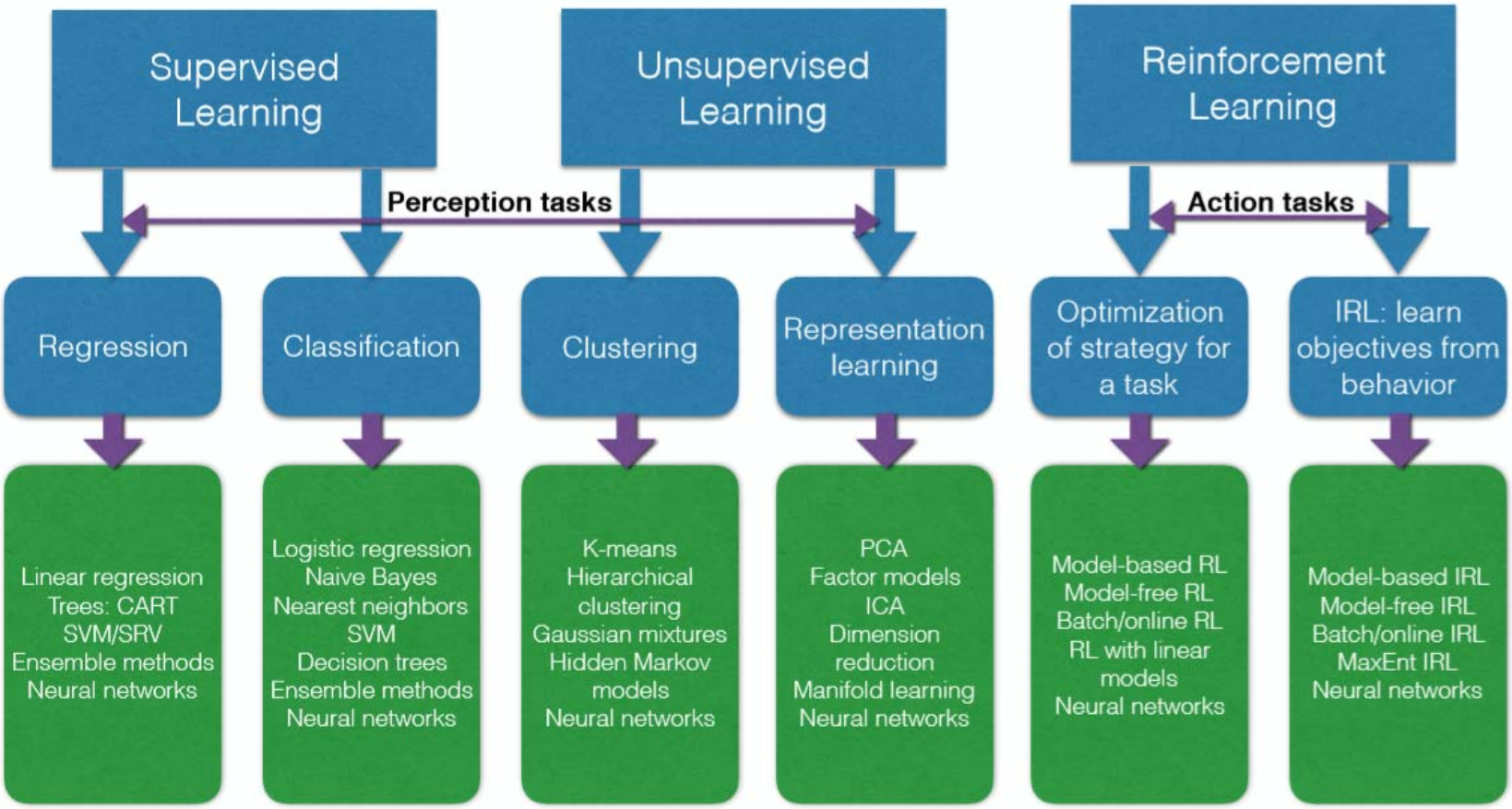


Machine Learning: examples



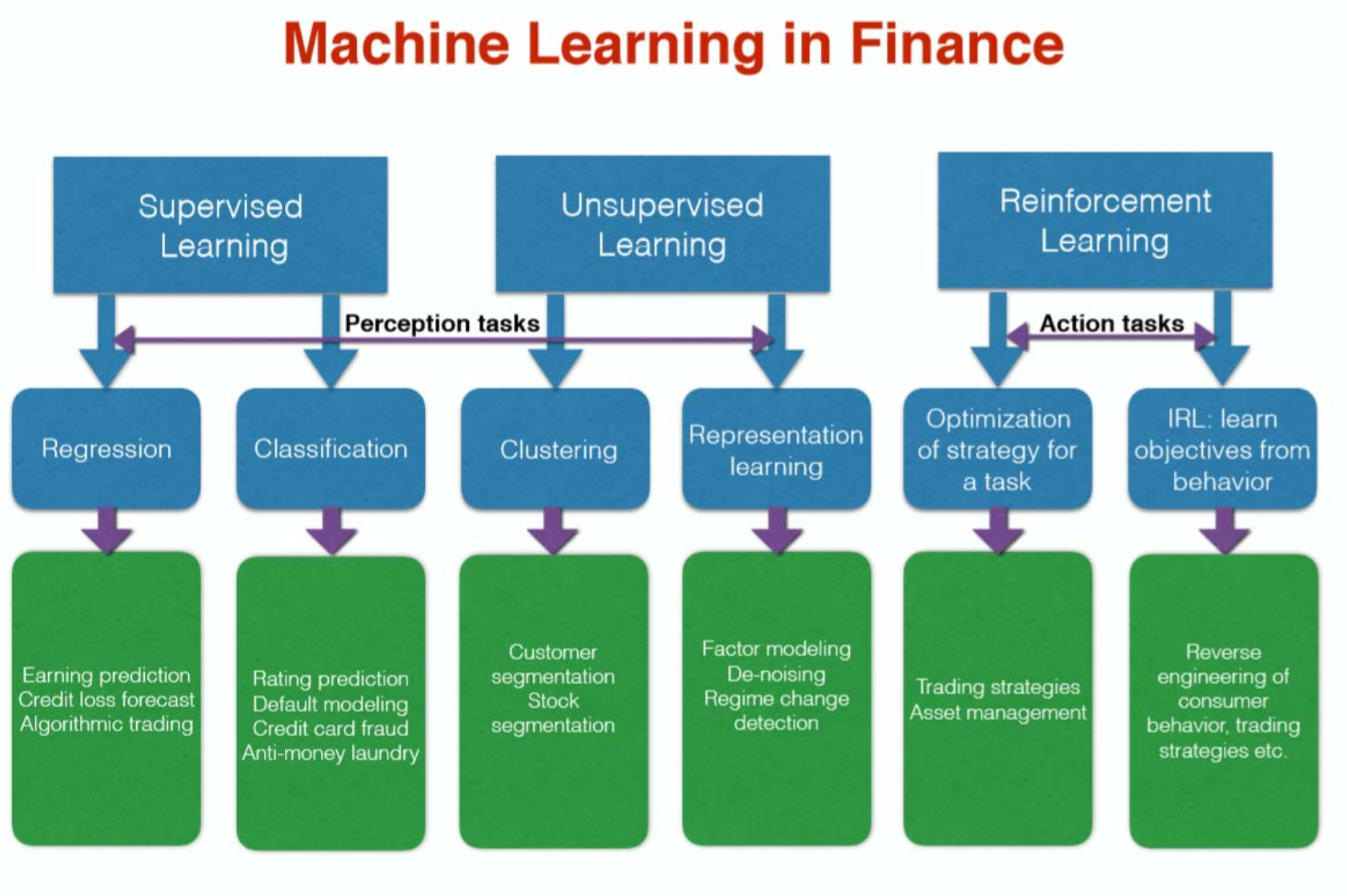
- These are general industrial applications
- Will be referred to as “ML in Tech” for short

Machine Learning: methods



- **Neural networks** is the most universal (and scalable) approach
- **Deep Learning** revolution (2007-present)!

Machine Learning in Finance



Which business question do you need to answer?

Classification

Who will (buy | fraud | churn …) next (week | month | year…)?

Regression

What will the (revenue | # churners) be next (week | month…)?

Segmentation or Clustering

What are the groups of customers with similar (behavior | profile …)?

Forecasting (Time Series Analysis)

What will the (revenue | # churners…) be over next year on a monthly basis?

Link Analysis

Analyze interactions to identify (communities | influencers…)

Association or Recommendation Engines

Provides recommendations on web sites or to retailers – basket analysis

Predictive Modeling Methodology – Overview

Use predictive analytics to solve a variety of business challenges

- Churn Reduction
- Customer Acquisition
- Lead Scoring
- Product Recommendation
- Campaign Optimization
- Customer Segmentation
- Next Best Offer/Action

- Predictive Maintenance
- Load Forecasting
- Inventory/Demand Optimization
- Product Recommendation
- Price Optimization
- Manufacturing Process Optimization
- Quality Management
- Yield Management

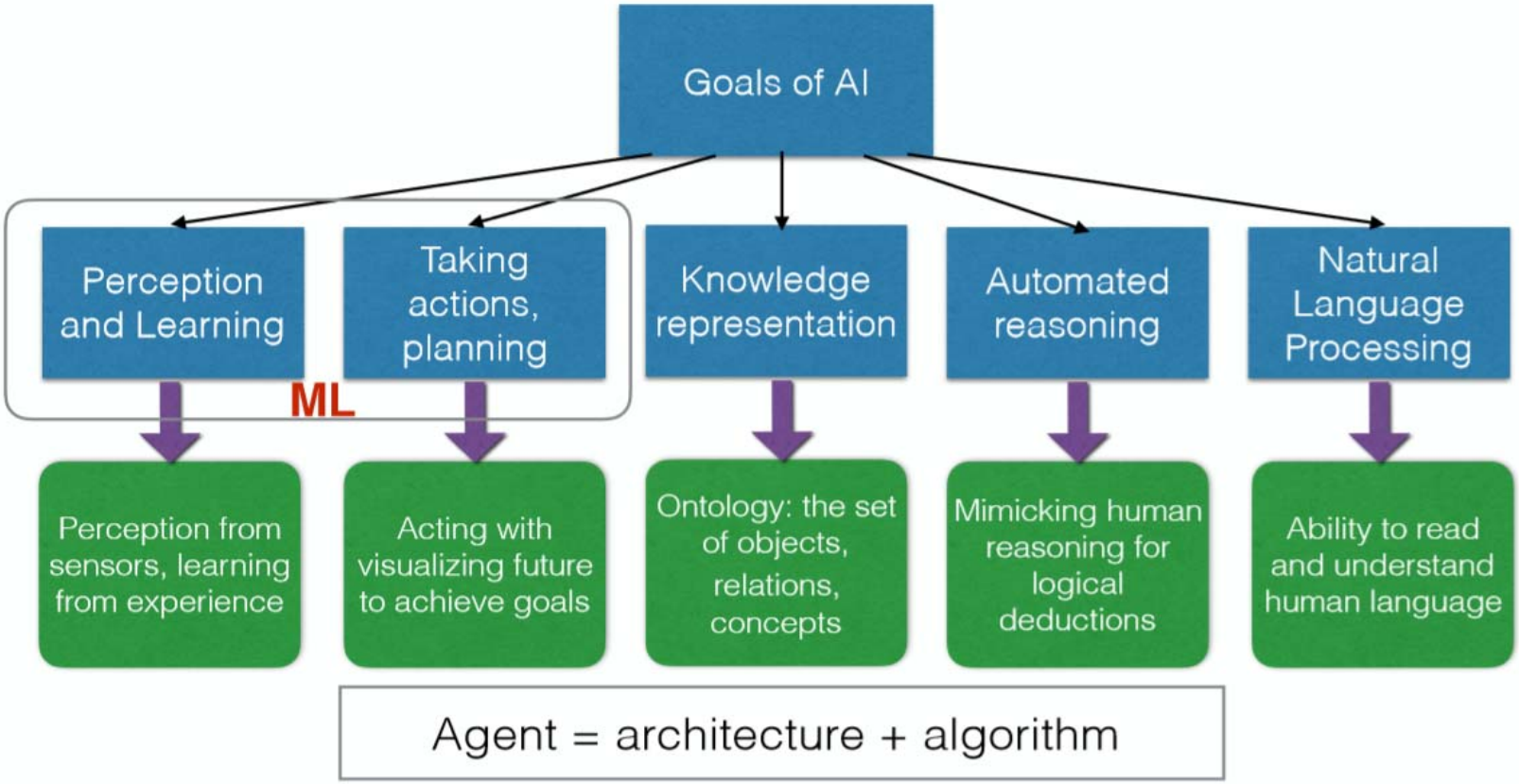
- Fraud and Abuse Detection
- Claims Analysis
- Collection and Delinquency
- Credit Scoring
- Operational Risk Modeling
- Crime Threat
- Revenue and Loss Analysis

- Cash Flow and Forecasting
- Budgeting Simulation
- Profitability and Margin Analysis
- Financial Risk Modeling
- Employee Retention Modeling
- Succession Planning

- Life Sciences
- Healthcare
- Media
- Higher Education
- Public Sector / Social Sciences
- Construction and Mining
- Travel and Hospitality
- Big Data and IoT

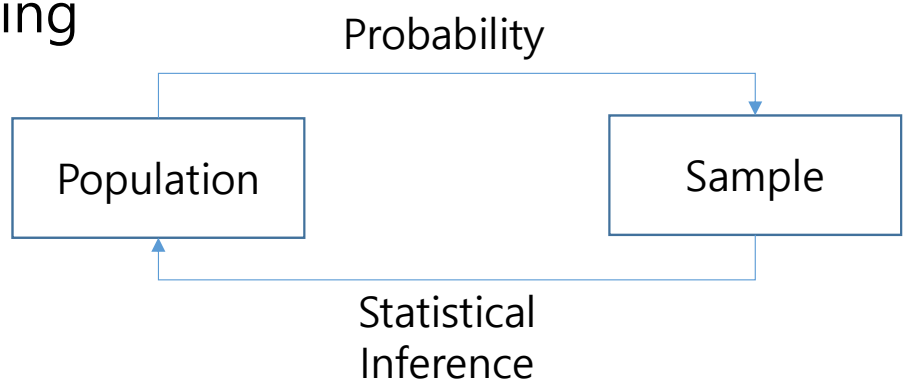
What is AI?

Artificial Intelligence (AI) studies "intelligent agents" that perceive their environment and perform different actions to solve tasks that involve mimicking cognitive functions of humans (Russell, Norvig, "Artificial Intelligence: A Modern Approach", 2009)

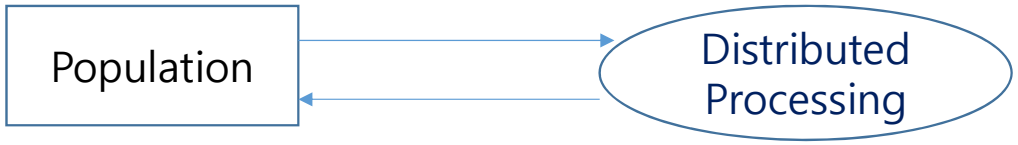


Differences between ML and Statistical Modeling

- Statistical Modeling



- ML for Big Data
  - Advances in computer and software technology



Differences between ML and Statistical modeling

Statistical Modeling	Machine Learning
Parametric models that try to "**explain**" the world. The focus is on modeling causality	**Non-parametric** models that try to "**mimic**" the world rather than "explain" it. Often uses correlations as proxies to causality
Deduce relations for observed quantities by parameter estimation for a pre-specified model of the world	**Induce** relations between observable quantities, main goal is predictive power
Small data (1-100 attributes,100-1000 examples)	Large data (10-100K attributes, 1K-100M examples)
Scalability is typically not the major concern	Scalability is often critical in applications
Based on a probabilistic approach	Some ML methods are not probabilistic (SVM, neural networks, clustering, etc.)

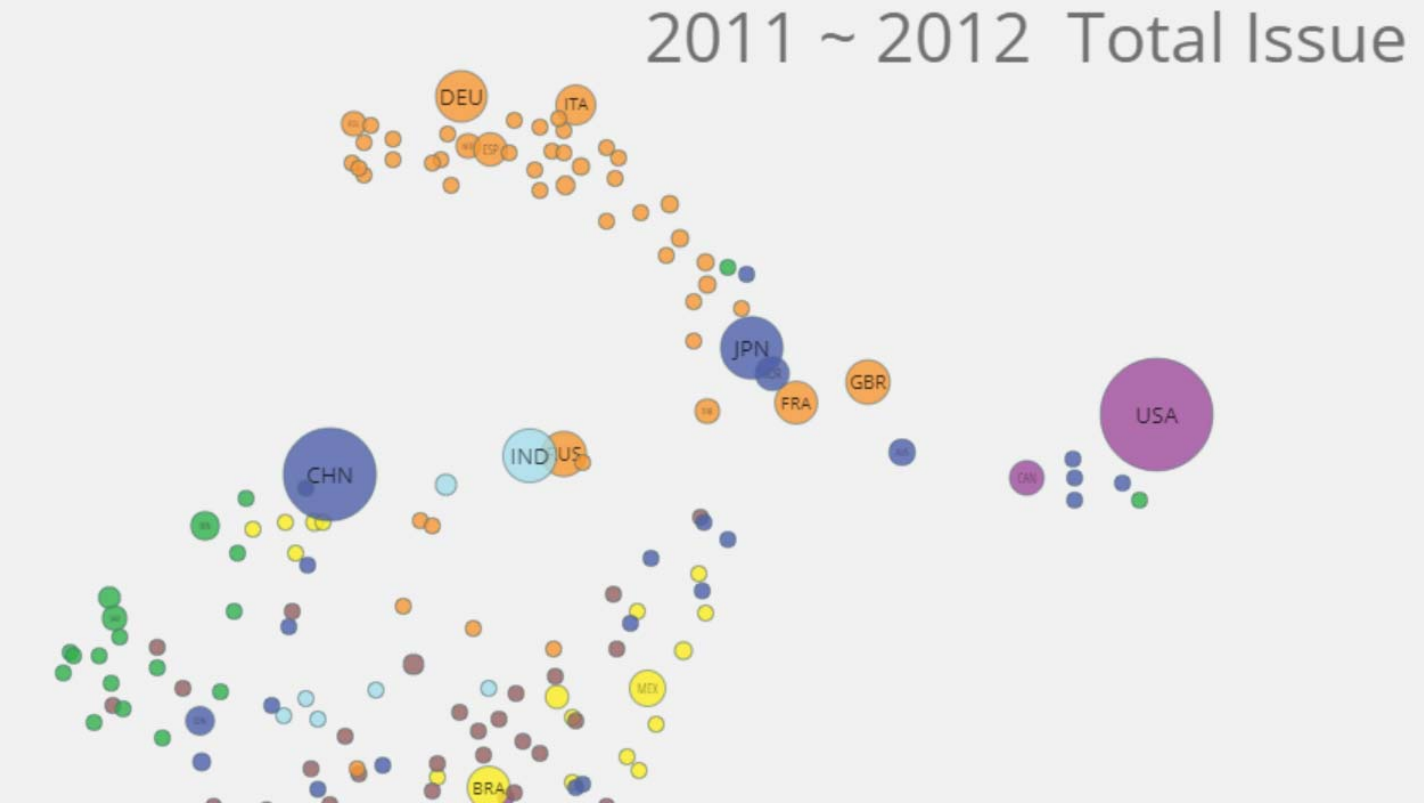
Programming for Data Analysis

- R
  - Open source (free)
  - Statistical, modeling, and data mining programs
  - Specialized language for visualizing and displaying results of graphs
- Python
  - Open source (free)
  - Data collection, machine learning, artificial intelligence programming
  - A programming language that anyone can easily learn
  - Secondary students, Political Science, Law, Food and Nutrition, etc.
- Data Mining/Machine Learning/Artificial Intelligence tools
  - Weka, Rapid Miner, Microsoft AzureML, TensorFlow, Torch, Keras

Big Data Visualization

- Effectively deliver data analysis results to management
- Difficult and complex information
  - Information expression technology expressed in simple charts or 3D images for easy understanding at a glance
- Example: 2009 Google Fusion Tables
  - An online service that expresses vast amounts of data

Example of Data Visualization

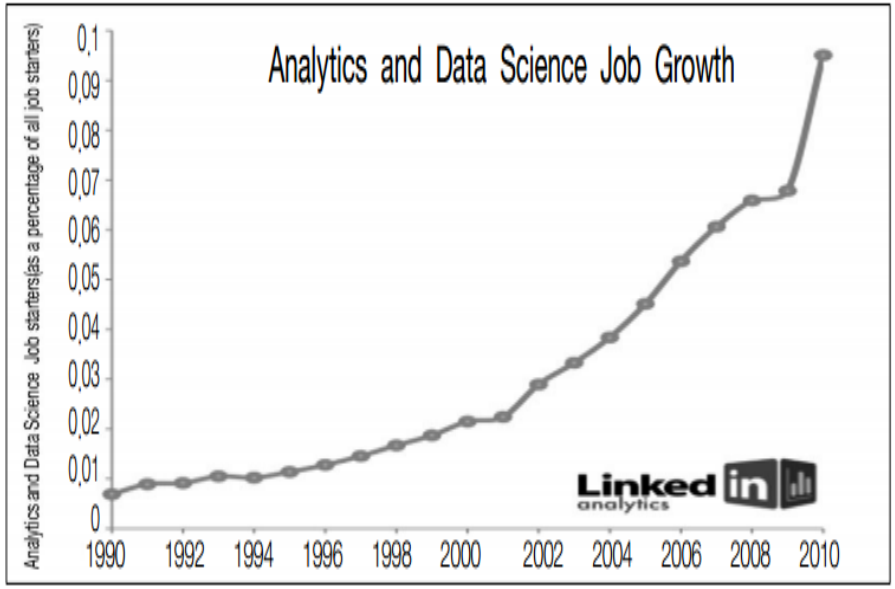


Data Analysis Occupation

- Over 4,900 jobs (LinkedIn)
  - Statistical analysts, data miners, business analysts, data analysts, mathematical economists, medical statisticians, insurance analysts, financial analysts, marketing researchers, ...
- Enterprise
  - Chief Data Officer (CDO)
  - Data Scientist
  - Data analyst (past)

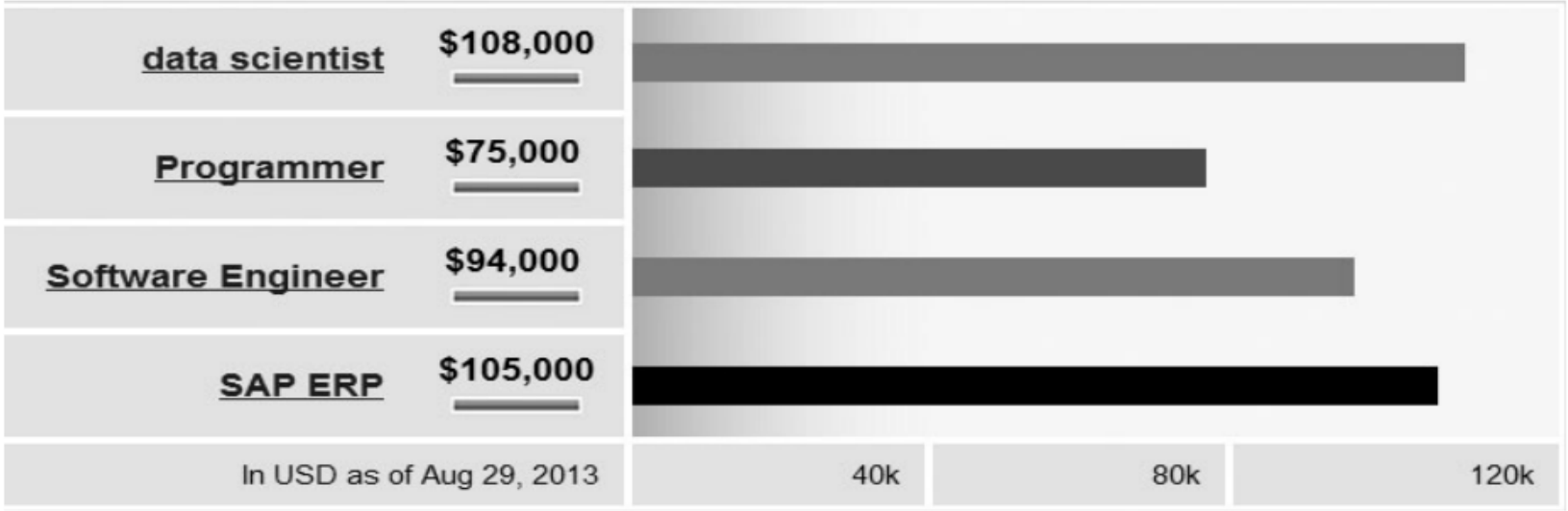
Data Scientist

- Harvard Business Review
  - The most attractive occupations of the 21st century
- McKinsey:
  - Nearly 200,000 data analysts in the US by 2018
  - 1.5 million data-base managers needed in the US



Data Scientist Salary and Employment Rate

Items / Universities	Employment rate(%)			Avg. Base Salary($)	
	2008	2009	2010	2009	2010
Master of Science in Analytics at North Carolina State Univ.	100.0	100.0	97.0	73,000	83,500
Master of Info. Sys. Mgt. at Carnegie Mellon	88.0	77.0	78.0	N/A	89,400
Master of OR and Info. Eng. At Cornell	88.0	73.0	85.0	79,200	N/A
Master of Finance at MIT	N/A	N/A	89.5	N/A	79,600



IBM - What is a Data Scie

Byung-Won

www-01.ibm.com/software/data/infosphere/data-scientist/

IBM

About data scientists

Rising alongside the relatively new technology of big data is the new job title data scientist. While not tied exclusively to big data projects, the data scientist role does complement them because of the increased breadth and depth of data being examined, as compared to traditional roles.

So what does a data scientist do?

A data scientist represents an evolution from the business or data analyst role. The formal training is similar, with a solid foundation typically in computer science and applications, modeling, statistics, analytics and math. What sets the data scientist apart is strong business acumen, coupled with the ability to communicate findings to both business and IT leaders in a way that can influence how an organization approaches a business challenge. Good data scientists will not just address business problems, they will pick the right problems that have the most value to the organization.

The data scientist role has been described as "part analyst, part artist." Anjul Bhambhri, vice president of big data products at IBM, says, "A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization."

Whereas a traditional data analyst may look only at data from a single source – a CRM system, for example – a data scientist will most likely explore and examine data from multiple disparate sources. The data scientist will sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem. A data scientist does not simply collect and report on data, but also looks at it from many angles, determines what it means, then recommends ways to apply the data.

Data scientists are inquisitive: exploring, asking questions, doing "what if" analysis, questioning existing assumptions and processes. Armed with data and analytical results, a top-tier data scientist will then communicate informed conclusions and recommendations across an organization's leadership structure.

Contact IBM

Considering a purchase?

Email IBM

Request a quote

Or call us at: 1-877-426-3774
Priority code: Info Mgmt

Forbes Article

Learn what a data scientist is from IBM's Anjul Bhambhri

Read the article

Blog posts by James Kobielus

- Data Scientists: Myths and mathemagical superpowers
- Data Scientist: Closing the Talent Gap
- Data Scientist: Master the Basics, Avoid The Most Common Mistakes
- Data Scientist: Exploration in the Age of the Unstructured
- Data Scientist: Bringing True Science into the Business Process

More data scientist resources

Data Scientist

- Expert to handle Big Data
- An expert who can analyze and visualize the large amount of data that the company has so that the executives in the enterprise can make appropriate decisions about the business in the future.
- Combines thought and expertise in computer, mathematics, statistics, management, industrial engineering, and visual design.
- Computer programming, algorithms, databases, distributed processing, basic statistics, machine learning, mathematical economics, time series, signal processing

Skills for Data Scientist

- At least 5 to 8 years of field experience
- Data Quality Expert
- Computer Programming
- Experience with various platforms
- Legacy data, SaaS, PaaS, IaaS
- Analysis software
- Communication ability
- Mindfulness, sincerity, curiosity, honesty

Summary

- Big data
  - Volume, Velocity, Variety
- Big data technologies
  - Hadoop, MapReduce, data analytics
- Data scientists
  - Data analytics handling big data in industry
  - Necessary academic skills: Computer Science, Statistics, Economics, and Substantive Expertise