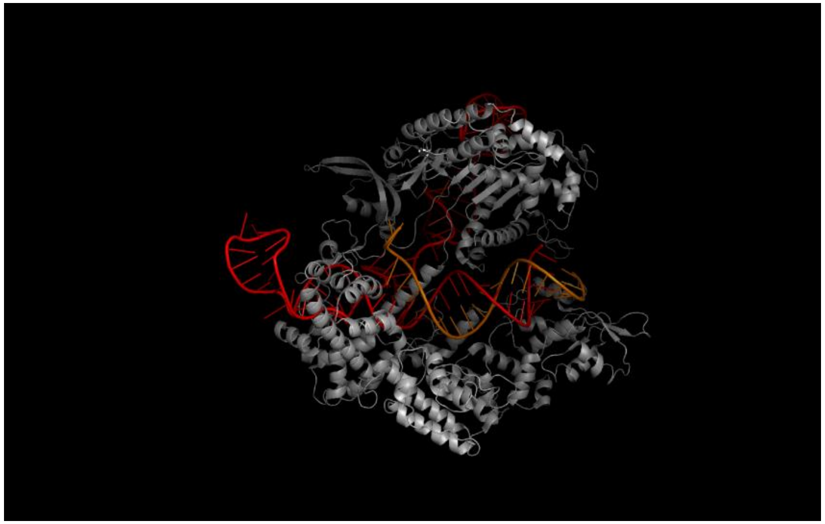
HPC meets Big Data

Dr. Putchong Uthayopas
Department of Computer Engineering,
Faculty of Engineering, Kasetsart University
Email: putchong@ku.th

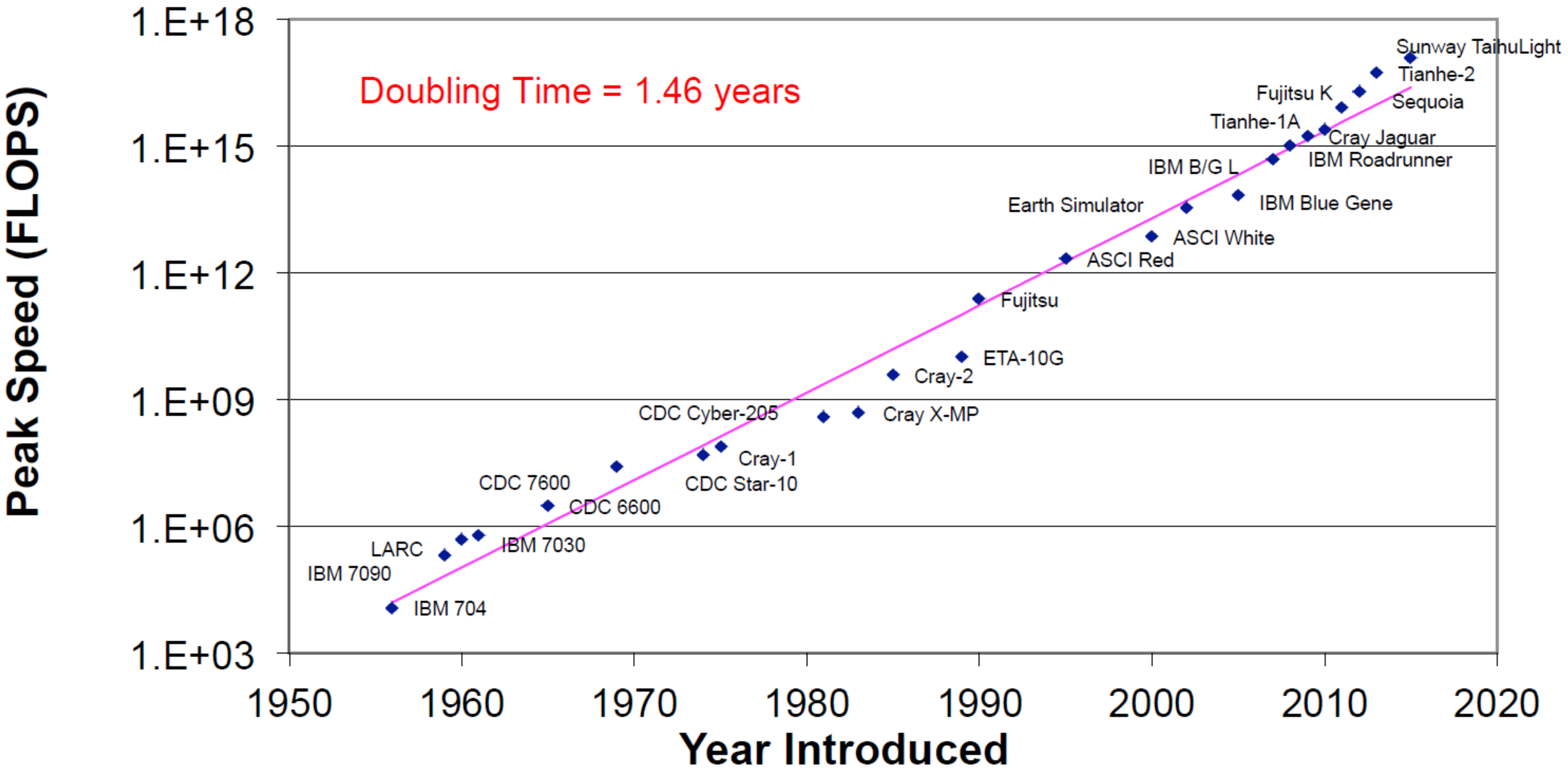
Brief History of HPC Platform



- HPC on single processor
- HPC using Vector machine
- HPC with SMP, SIMD
- MPP and Cluster Computin
- GPU computing
- Heteronomous computing

Mostly build for compute-intensive application!

Growth of Supercomputing



Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks

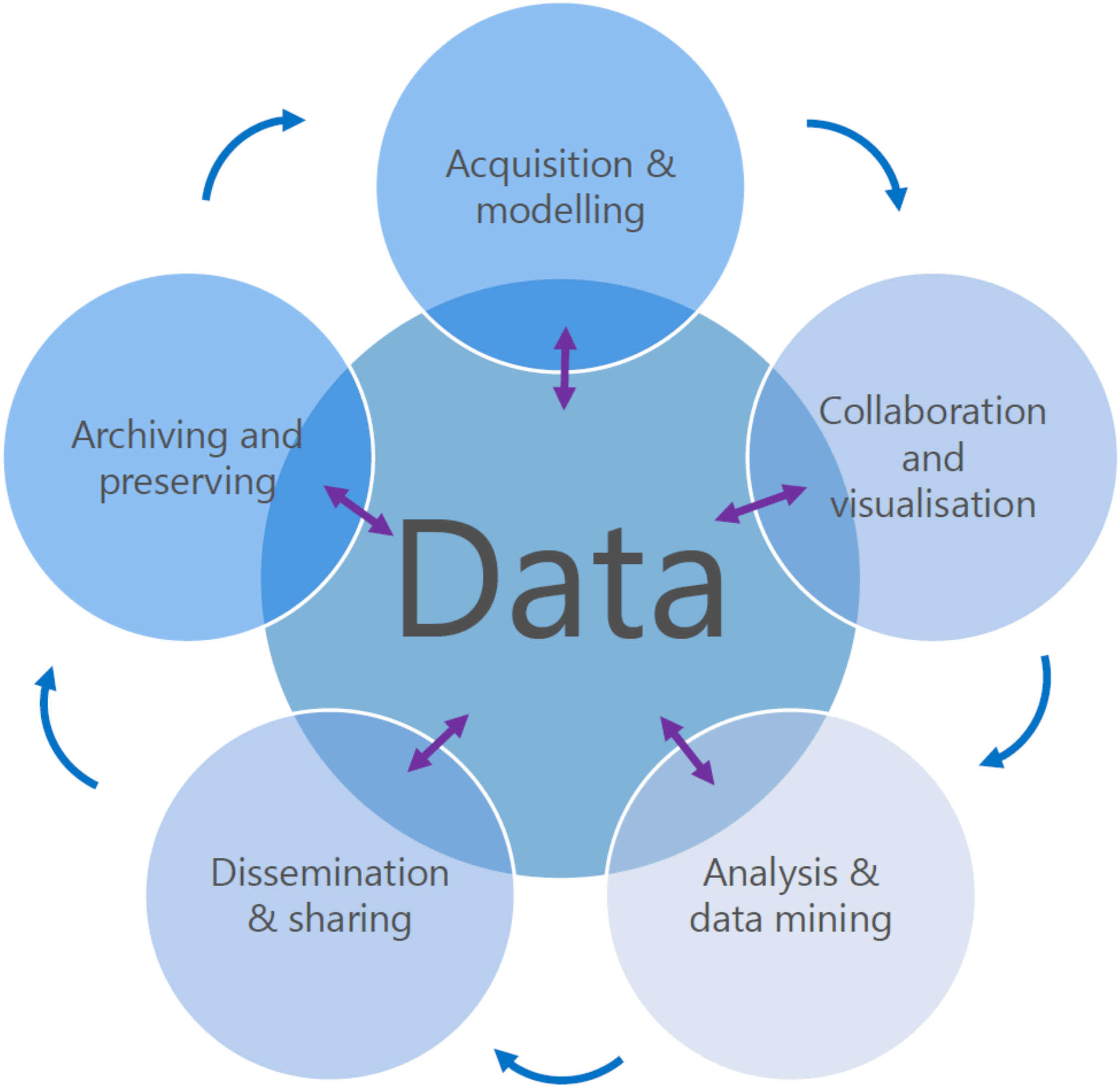
➢ eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

(With thanks to Jim Gray)

Data-intensive Research



Acquisition & modelling

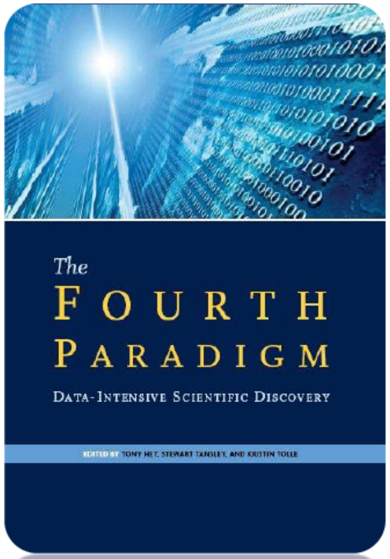
Collaboration and visualisation

Analysis & data mining

Dissemination & sharing

Archiving and preserving

Data

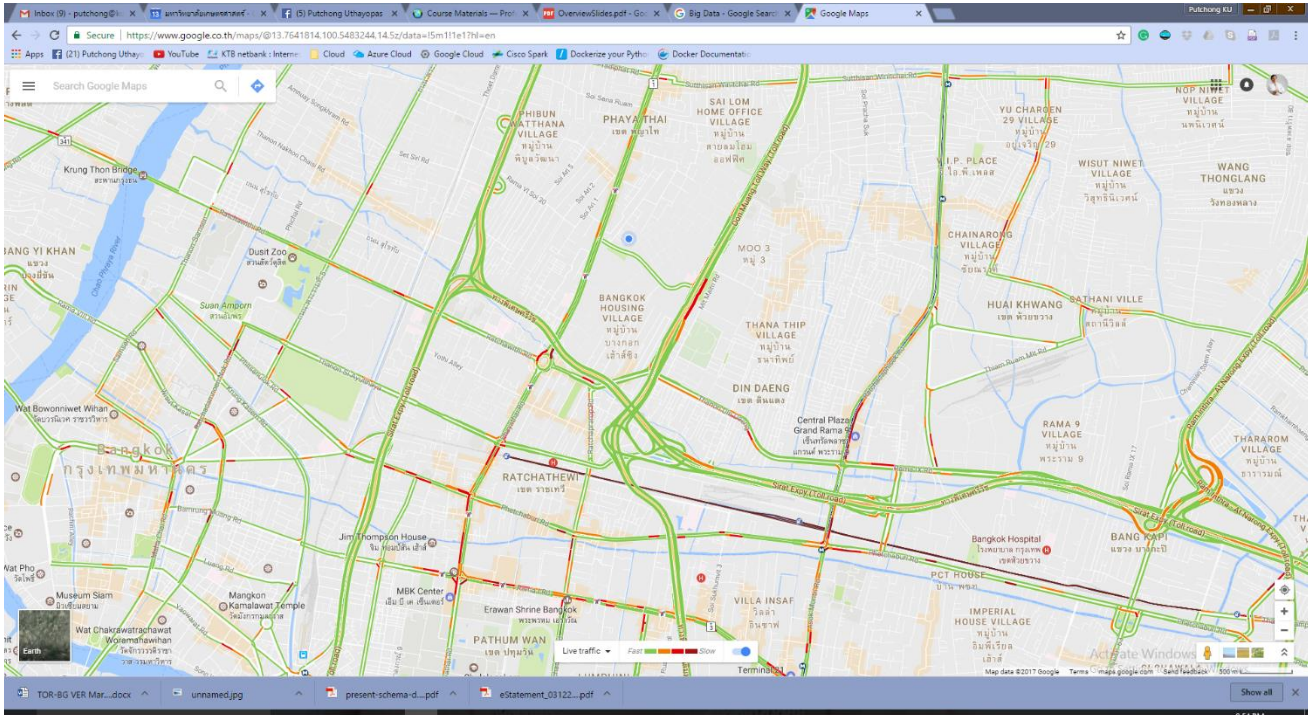

The
FOURTH PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

fourthparadigm.org

Science@Microsoft
THE FOURTH PARADIGM IN PRACTICE

We are using Big Data All the time

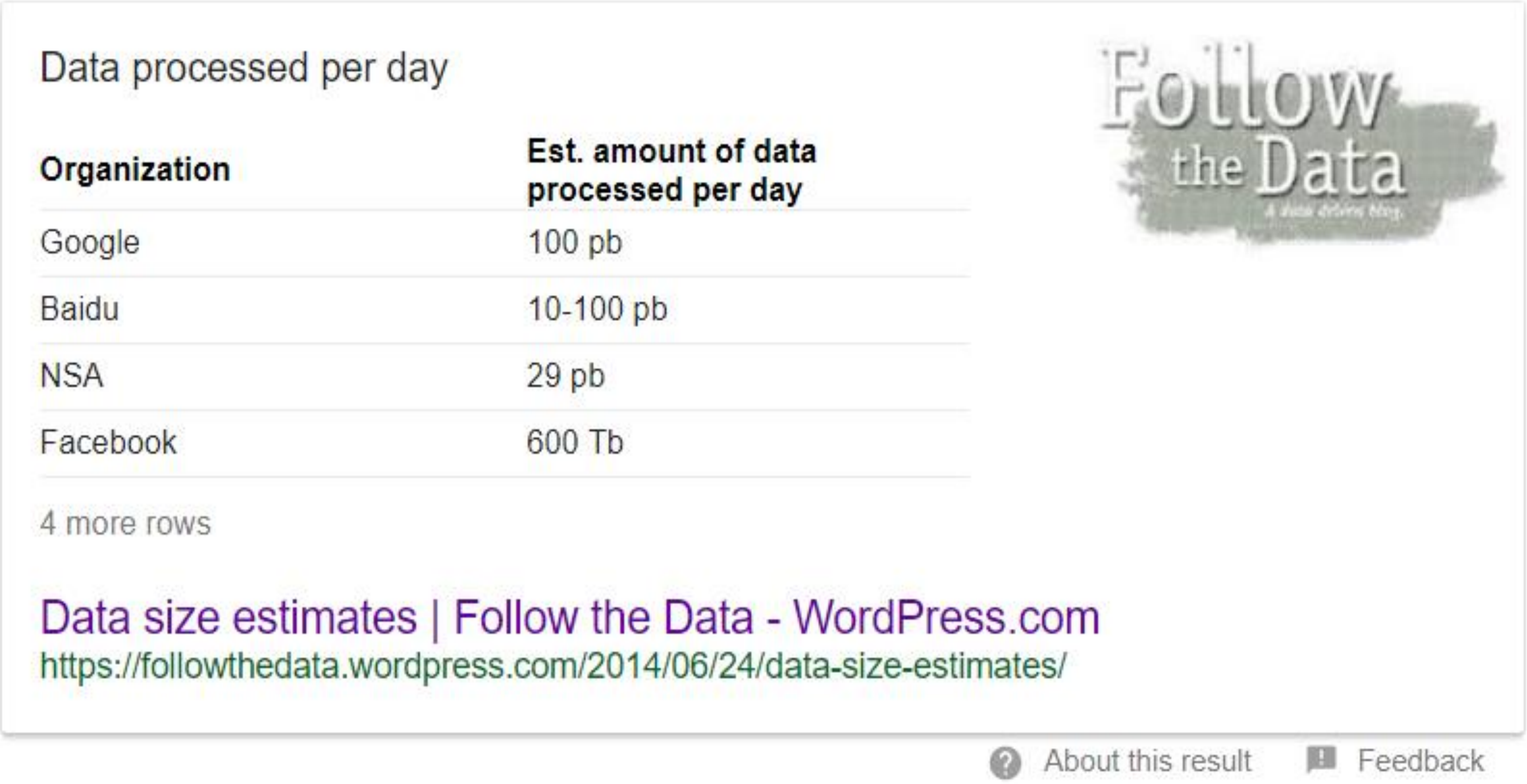
- How can google map know about the traffic condition?



Facebook Usage Statistics (March 2016)

- **1.09 billion daily active users**
- **989 million mobile daily active users**
- 1.65 billion monthly active users
- 1.51 billion mobile monthly active users

Source: http://newsroom.fb.com/company-info/

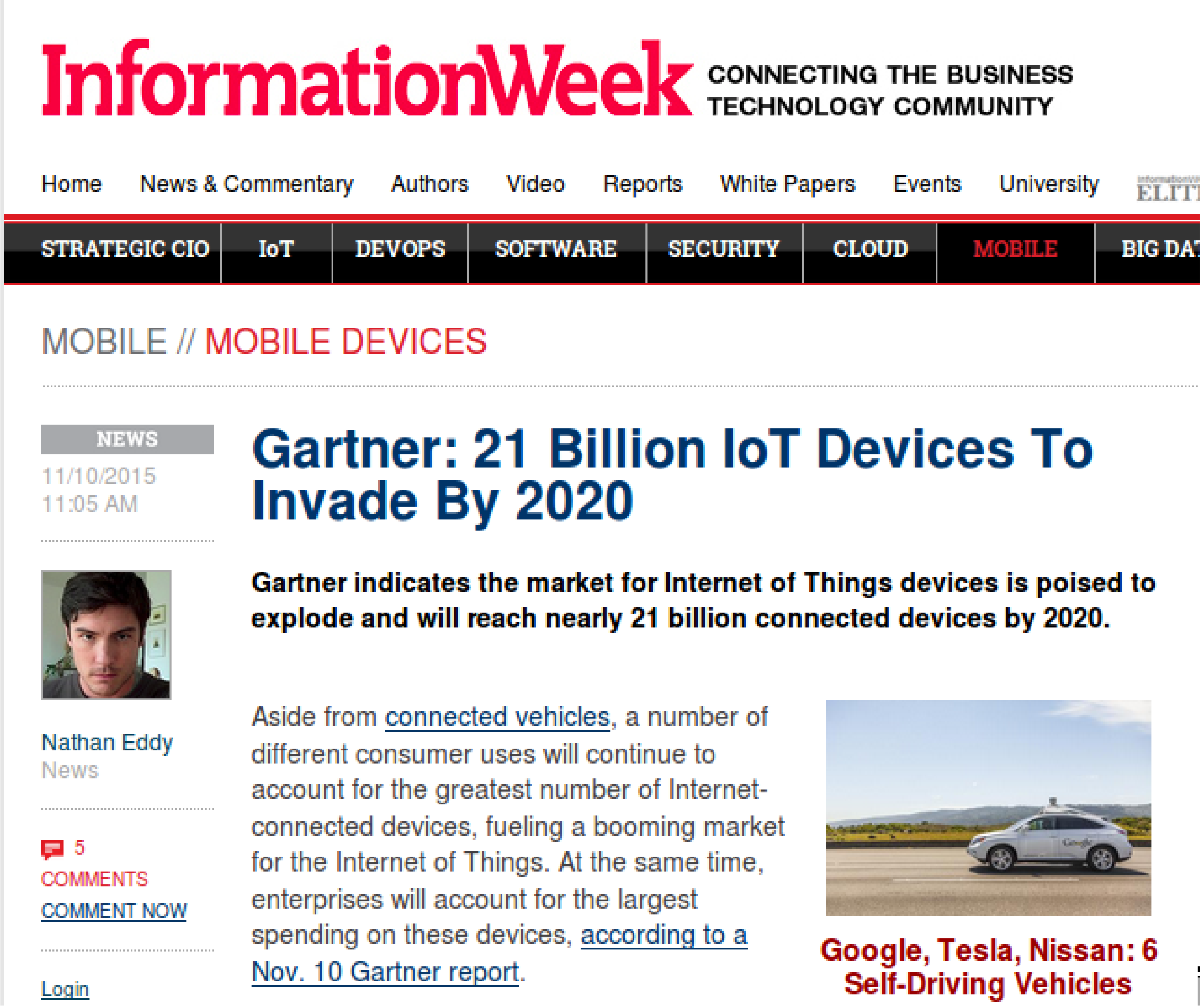


Data processed per day

Organization	Est. amount of data processed per day
Google	100 pb
Baidu	10-100 pb
NSA	29 pb
Facebook	600 Tb

4 more rows

Data size estimates | Follow the Data - WordPress.com
https://followthedata.wordpress.com/2014/06/24/data-size-estimates/



InformationWeek CONNECTING THE BUSINESS TECHNOLOGY COMMUNITY

Home News & Commentary Authors Video Reports White Papers Events University ELITE

STRATEGIC CIO IoT DEVOPS SOFTWARE SECURITY CLOUD MOBILE BIG DAT

MOBILE // MOBILE DEVICES

NEWS

11/10/2015
11:05 AM

Nathan Eddy
News

5
COMMENTS
COMMENT NOW

Login

Gartner: 21 Billion IoT Devices To Invade By 2020

Gartner indicates the market for Internet of Things devices is poised to explode and will reach nearly 21 billion connected devices by 2020.

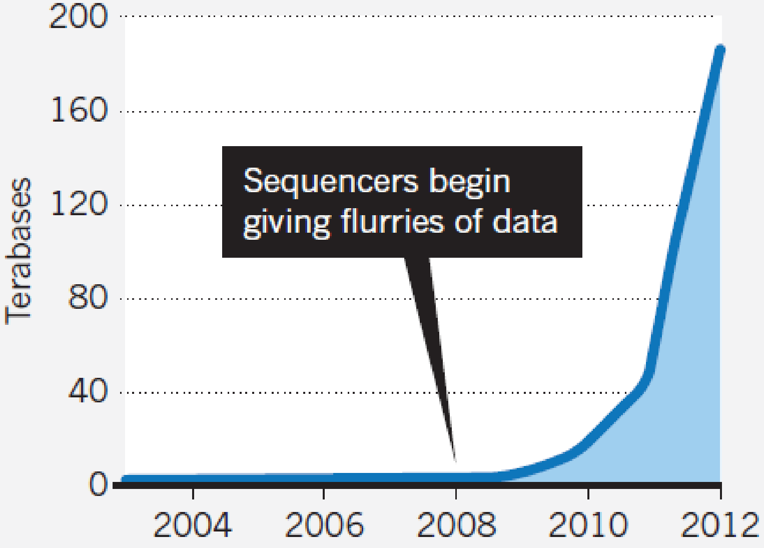
Aside from connected vehicles, a number of different consumer uses will continue to account for the greatest number of Internet-connected devices, fueling a booming market for the Internet of Things. At the same time, enterprises will account for the largest spending on these devices, according to a Nov. 10 Gartner report.

Google, Tesla, Nissan: 6 Self-Driving Vehicles

SOURCE: EMBL–EBI

DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.

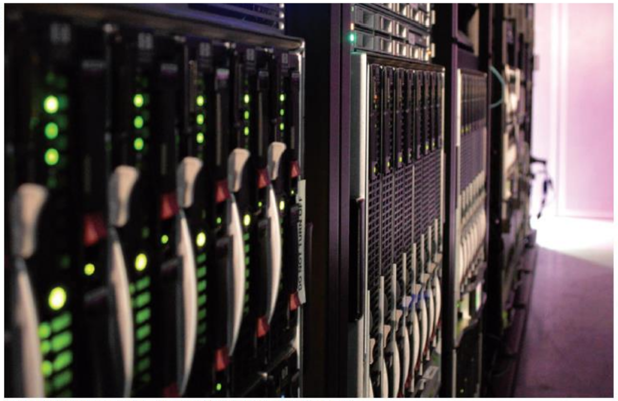


TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

EMBL-EBI



Extremely powerful computers are needed to help biologists to handle big-data traffic jams.

BY VIVIEN MARX

Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[1].

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why coastal algae bloom, what microbes dwell where in human body cavities and how the genetic make-up of different cancers influences how cancer patients fare[2]. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year[3] (see 'Data explosion').

This data pile is just one-tenth the size of the data store at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Every year, particle-collision events in CERN's Large Hadron Collider generate around 15 petabytes of data — the equivalent of about 4 million high-definition feature-length films. But the EBI and institutes like it face similar data-wrangling challenges to those at CERN, says Ewan Birney, associate director of the EBI. He and his colleagues now regularly meet with organizations such as CERN and the European Space Agency (ESA) in Paris to swap lessons about data storage, analysis and sharing.

All labs need to manipulate data to yield research answers. As prices drop for high-throughput instruments such as automated ▶

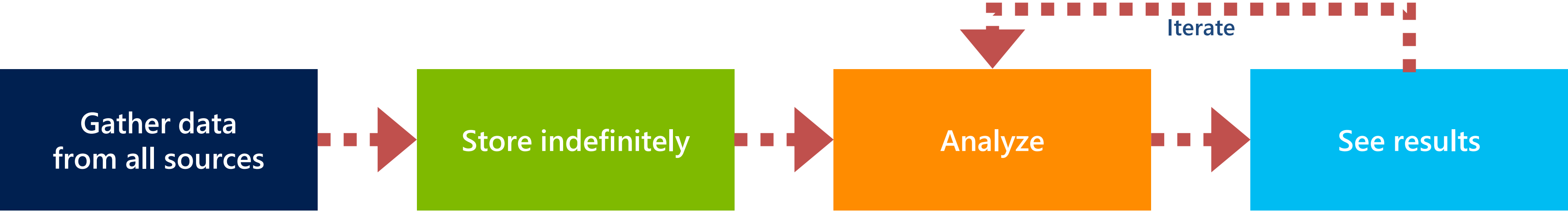
Google Open Image Dataset

- **The Open Images Dataset**
- **YouTube-8M Dataset**
- **Google Books Ngrams**
- **Google Trends Datastore**

https://www.infoworld.com/article/3131515/artificial-intelligence/4-google-data-sets-to-kickstart-machine-learning.html

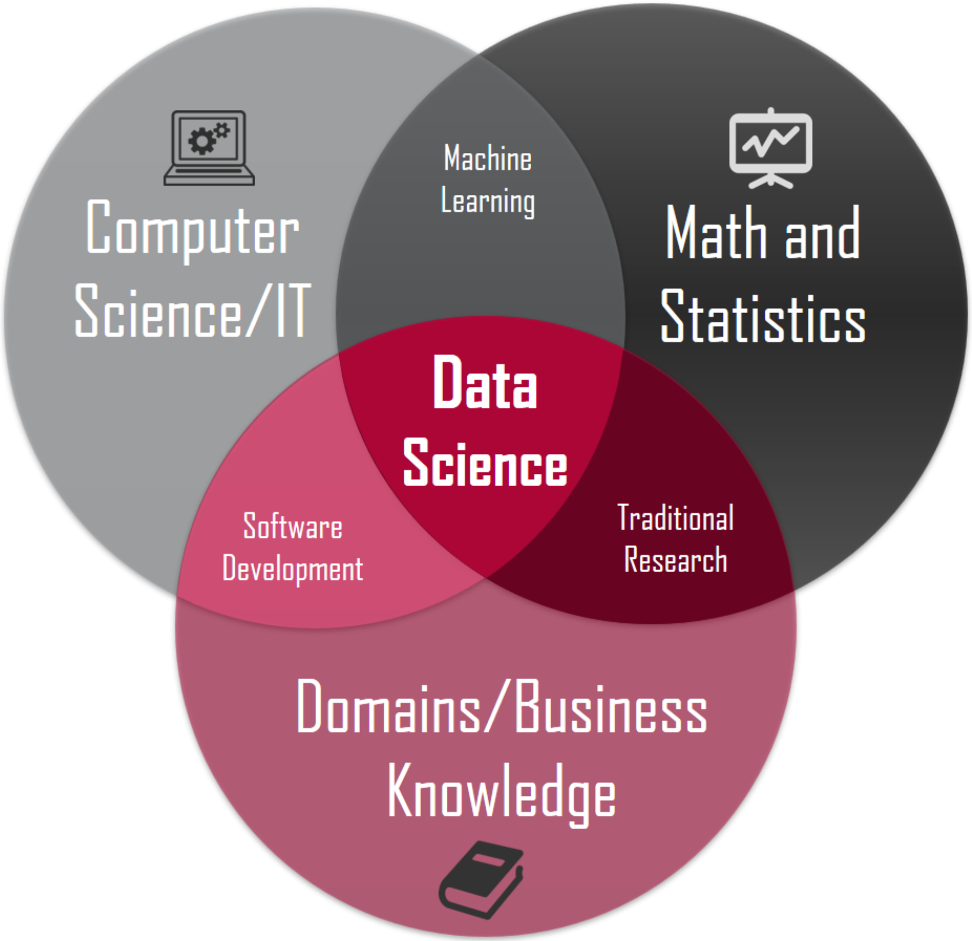
New big data thinking: All data has value

- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read).*
- Apps and users interpret the data as they see fit



Microsoft

What is Data Science?



- **Data Science** is the extraction of knowledge from large volumes of data that are structured or unstructured.

Source:Wikipedia

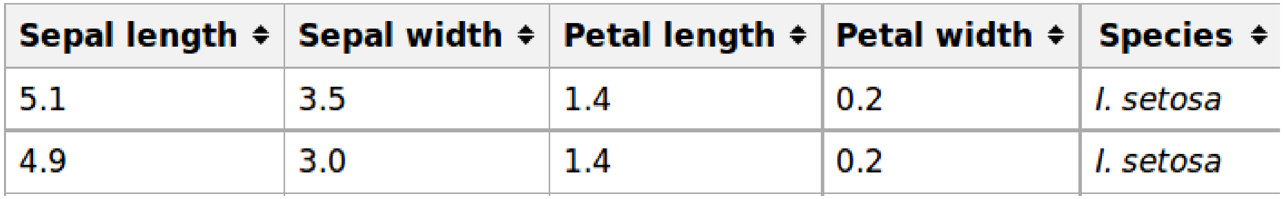
K-Mean on iris data

Iris setosa

Iris versicolor

Iris virginica

Ref: https://en.wikipedia.org/wiki/Iris_flower_data_set

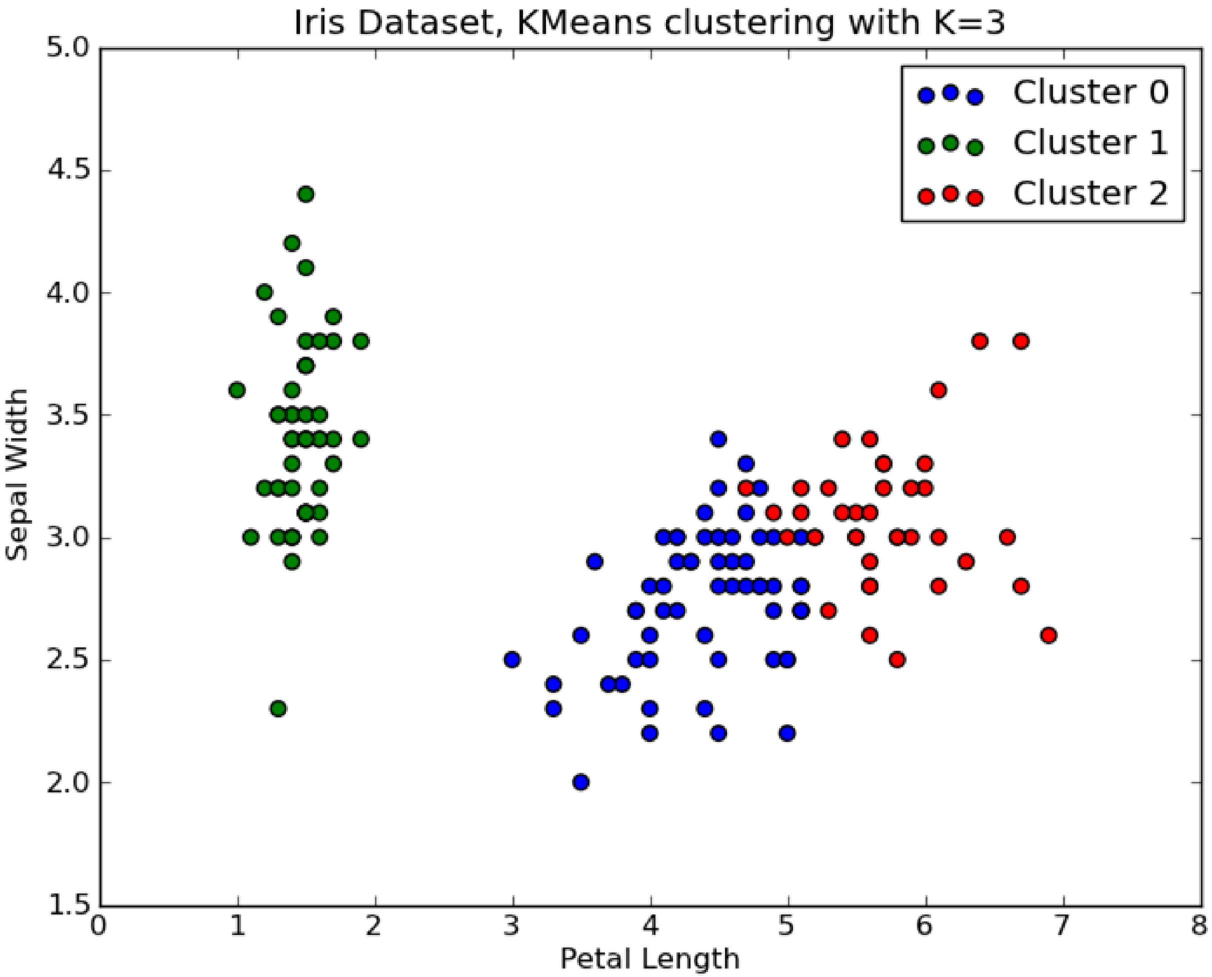


Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*

Iris data: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

SparkML kmean app:
https://github.com/apache/spark/tree/branch-1.5/examples/src/main/python/mllib

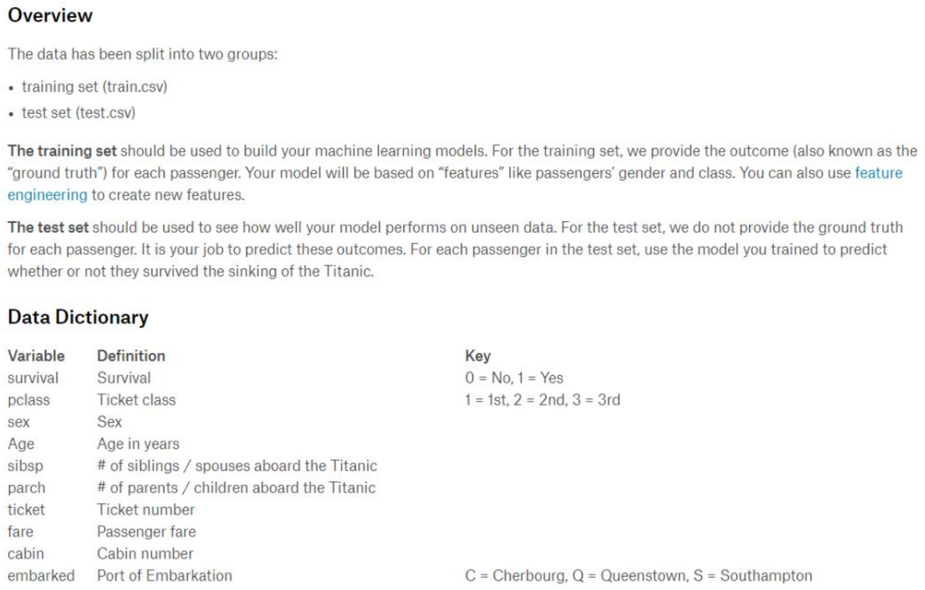
Ex. Iris K-mean



Source: http://stackoverflow.com/questions/6645895/calculating-the-percentage-of-variance-measure-for-k-means

Kaggle

- Titanic competition
  - What is the factor involve in surviving Titanic
- Data set of the passengers has been provided



Overview

The data has been split into two groups:

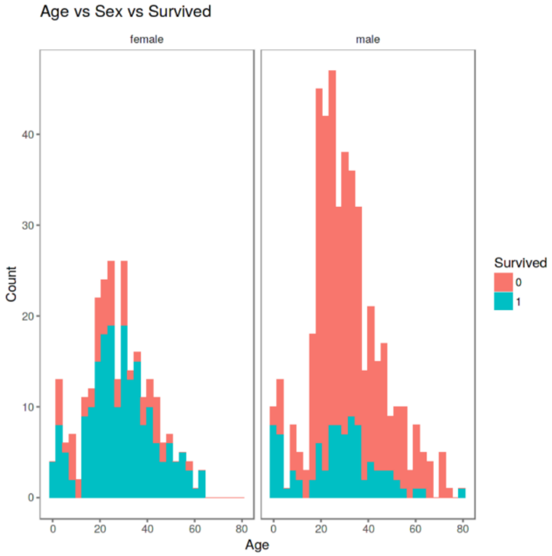
- training set (train.csv)
- test set (test.csv)

The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the "ground truth") for each passenger. Your model will be based on "features" like passengers' gender and class. You can also use feature engineering to create new features.

The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. It is your job to predict these outcomes. For each passenger in the test set, use the model you trained to predict whether or not they survived the sinking of the Titanic.

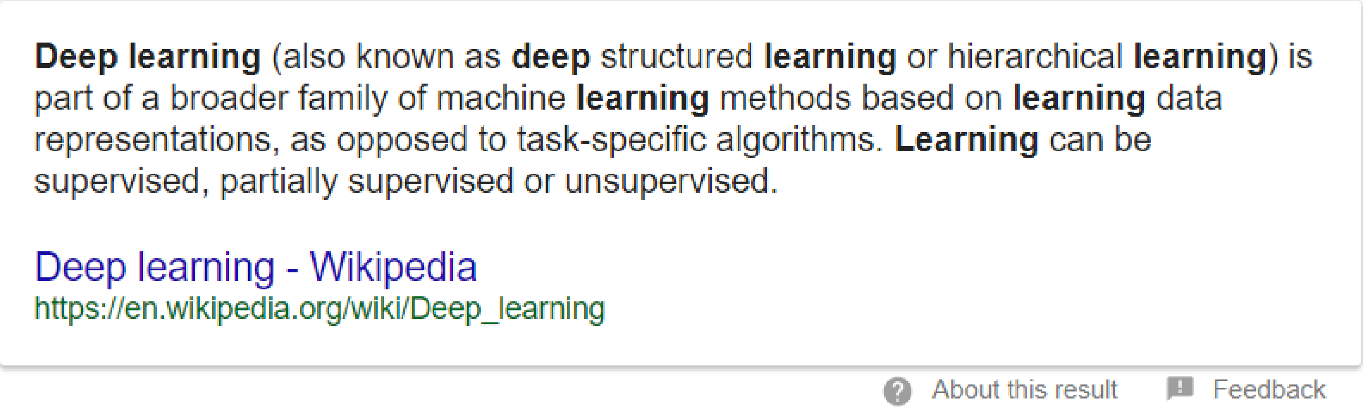
Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton



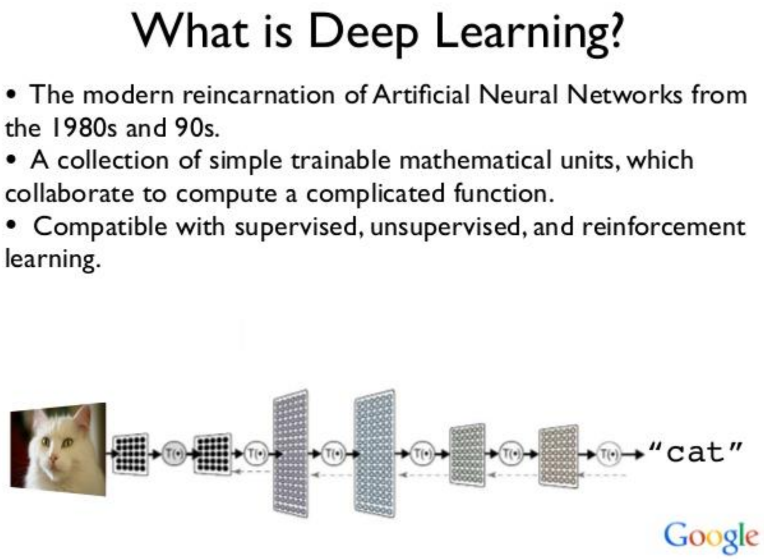
https://www.kaggle.com/c/titanic

Deep Learning



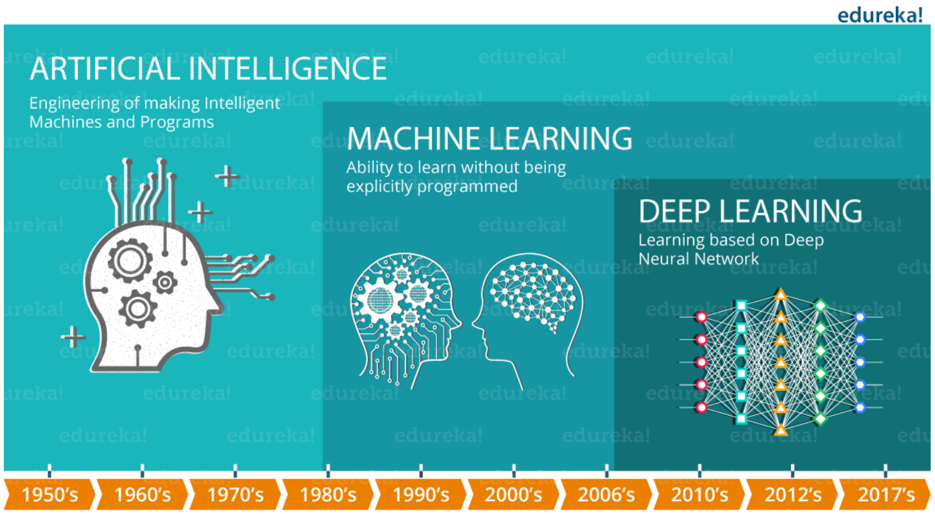
Deep learning (also known as **deep** structured **learning** or hierarchical **learning**) is part of a broader family of machine **learning** methods based on **learning** data representations, as opposed to task-specific algorithms. **Learning** can be supervised, partially supervised or unsupervised.

Deep learning - Wikipedia
https://en.wikipedia.org/wiki/Deep_learning



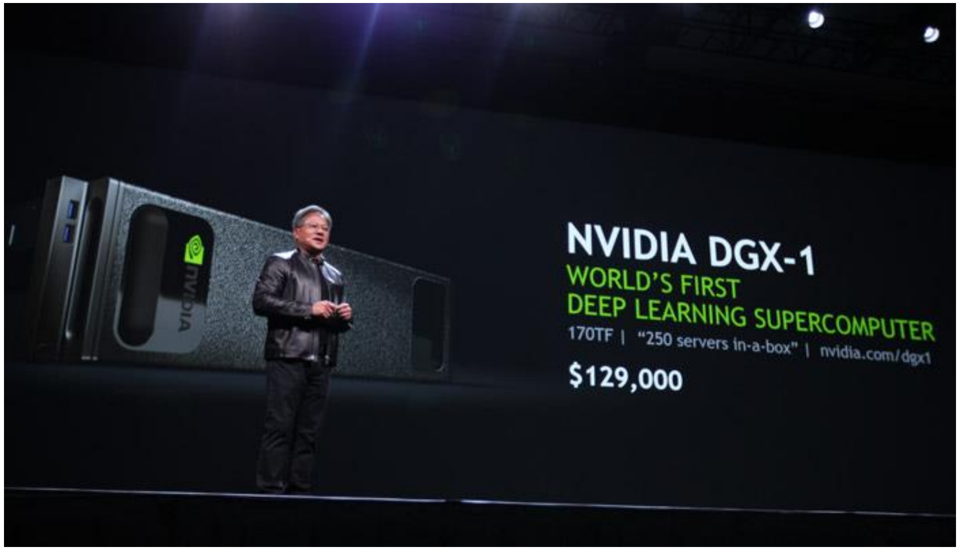
What is Deep Learning?

- The modern reincarnation of Artificial Neural Networks from the 1980s and 90s.
- A collection of simple trainable mathematical units, which collaborate to compute a complicated function.
- Compatible with supervised, unsupervised, and reinforcement learning.



edureka!
ARTIFICIAL INTELLIGENCE
Engineering of making Intelligent Machines and Programs
MACHINE LEARNING
Ability to learn without being explicitly programmed
DEEP LEARNING
Learning based on Deep Neural Network
1950's
1960's
1970's
1980's
1990's
2000's
2006's
2010's
2012's
2017's

New Technology to Handle Big Data and Machine learning


NVIDIA DGX-1
WORLD'S FIRST
DEEP LEARNING SUPERCOMPUTER
170TF | "250 servers in-a-box" | nvidia.com/dgx1
$129,000
NVIDIA

- Hadoop/Spark Ecosystem
- GPU system
- GPU Cluster
- AI supercomputer using dense GPUs

A DECADE OF SCIENTIFIC COMPUTING WITH GPUS

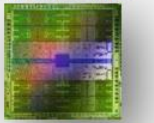
Stream Processing @ Stanford

CUDA Launched

NVIDIA CUDA

World's First GPU Top500 System

Fermi: World's First HPC GPU



Oak Ridge Deploys World's Fastest Supercomputer w/ GPUs

Discovered How H1N1 Mutates to Resist Drugs

AlexNet beats expert code by huge margin using GPUs

IMAGENET

World's First Atomic Model of HIV Capsid

Stanford Builds AI Machine using GPUs

World's First 3-D Mapping of Human Genome

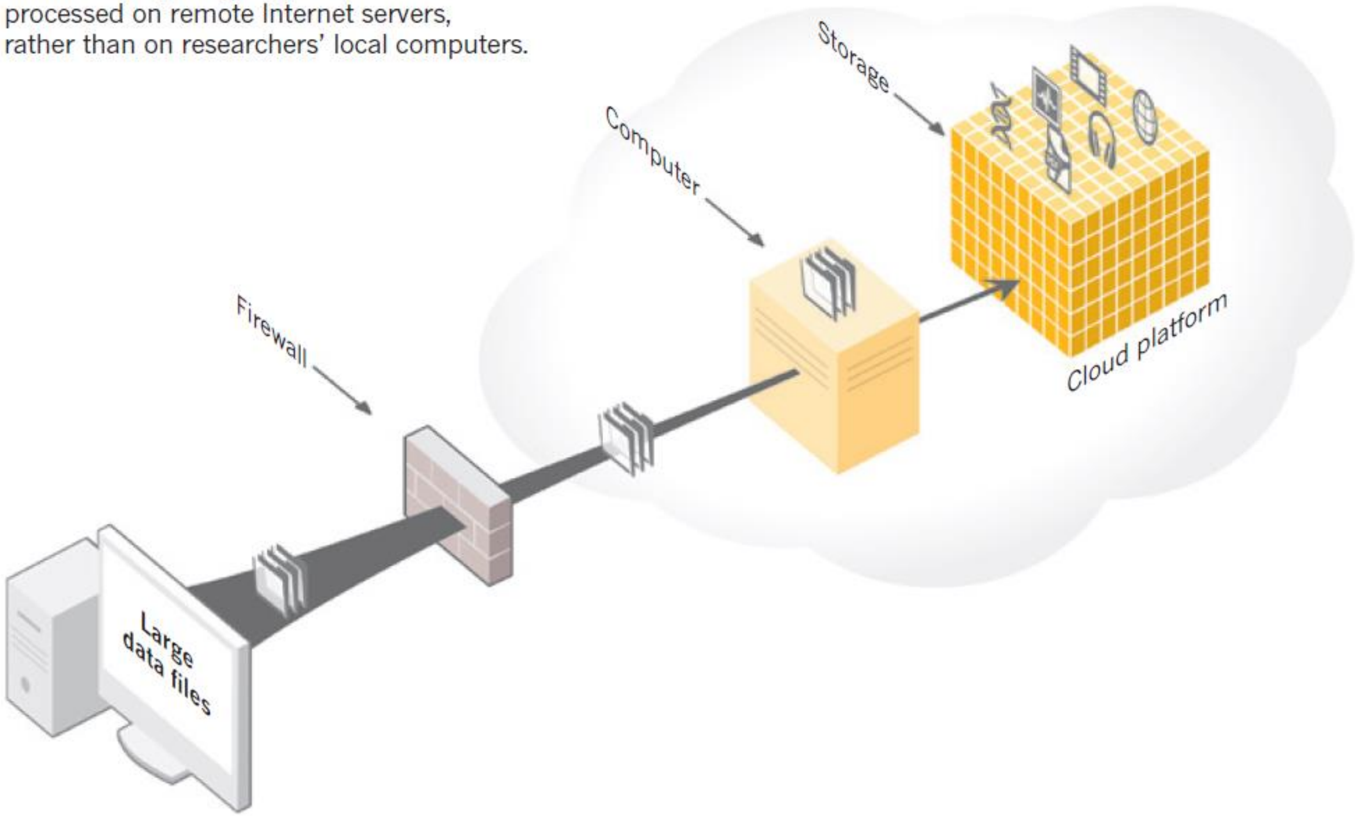
Google Outperform Humans in ImageNet

GPU-Trained AI Machine Beats World Champion in Go

2006 2008 2010 2012 2014 2016

HEAD IN THE CLOUDS

In cloud computing, large data sets are processed on remote Internet servers, rather than on researchers' local computers.



SOURCE: ASPERA

Summary

- Scientific Research is rapidly changing to Data Intensive Research
  - Driving by Big data analytics and Machine learning
- Innovative Platform is needed that put data storage, and very high computing power in one place
  - Hadoop move computing to data not traditional data move to computing
- Everything going to the CLOUD