Introduction to Big Data

Chapter 1 & 2 (Week 1)

Course overview & introduction

Asst. Prof. Minseok Seo
mins@korea.ac.kr

Course Overview

Introduction to Big Data

01

Contents

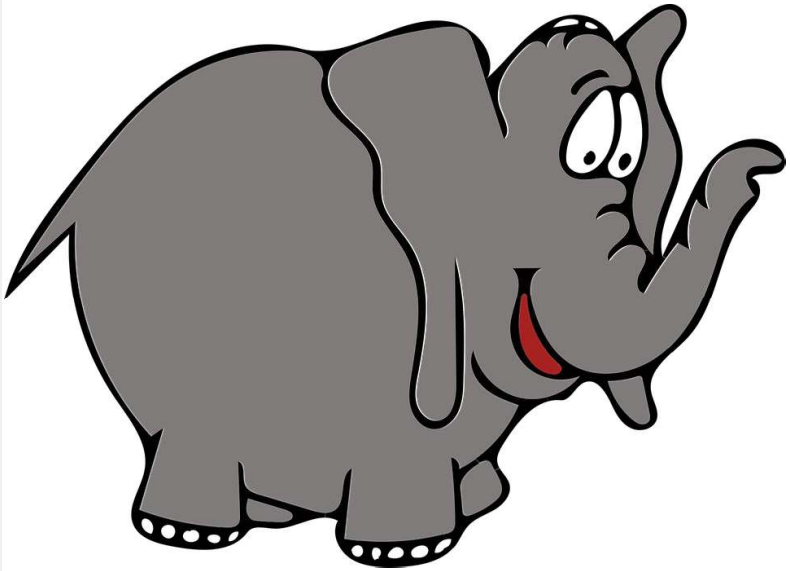
1. **Course Overview**
   - Brief introduction of professor & course
   - Object & Aim of the course
   - Assignments & Quiz
   - Evaluation

2. **Introduction to Big Data**
   - Definition of Big Data
   - Key techniques in Data Science
   - Core technology of Informatics

Course Overview

Definition of Big Data (Cont.)

Big Data



vs.

Which is bigger, elephant or rat?

Course Overview

Definition of Big Data (Cont.)

- **What is Data?**

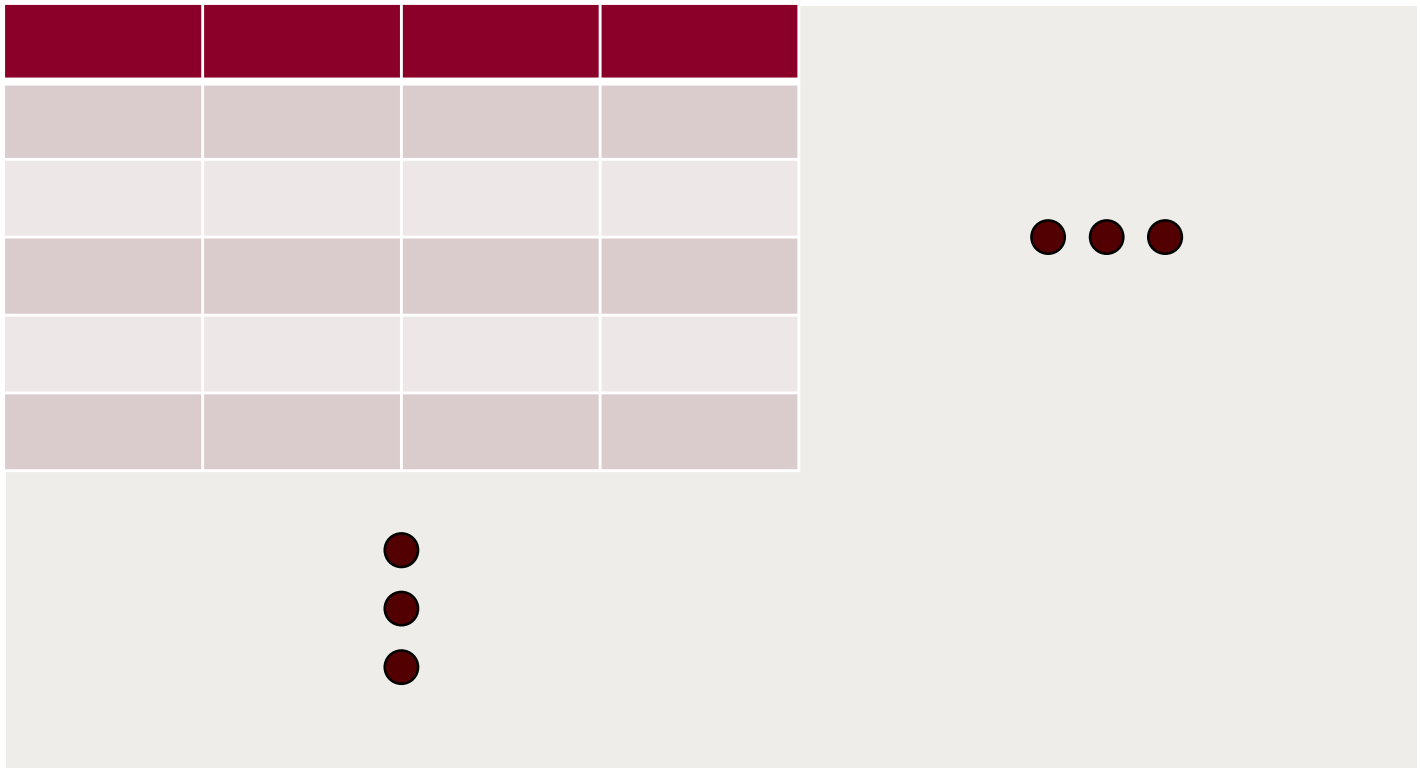
Attributes (Dimension; Features; Variables)

Objects (Samples, Individuals)

ID	Height	Weight	Age
Student 1	189 cm	81 kg	24
Student 2	210 cm	90 kg	26
Student 3	191 cm	92 kg	27
…	…	…	…
Student N	162 cm	71 kg	21

Course Overview

Definition of Big Data (Cont.)

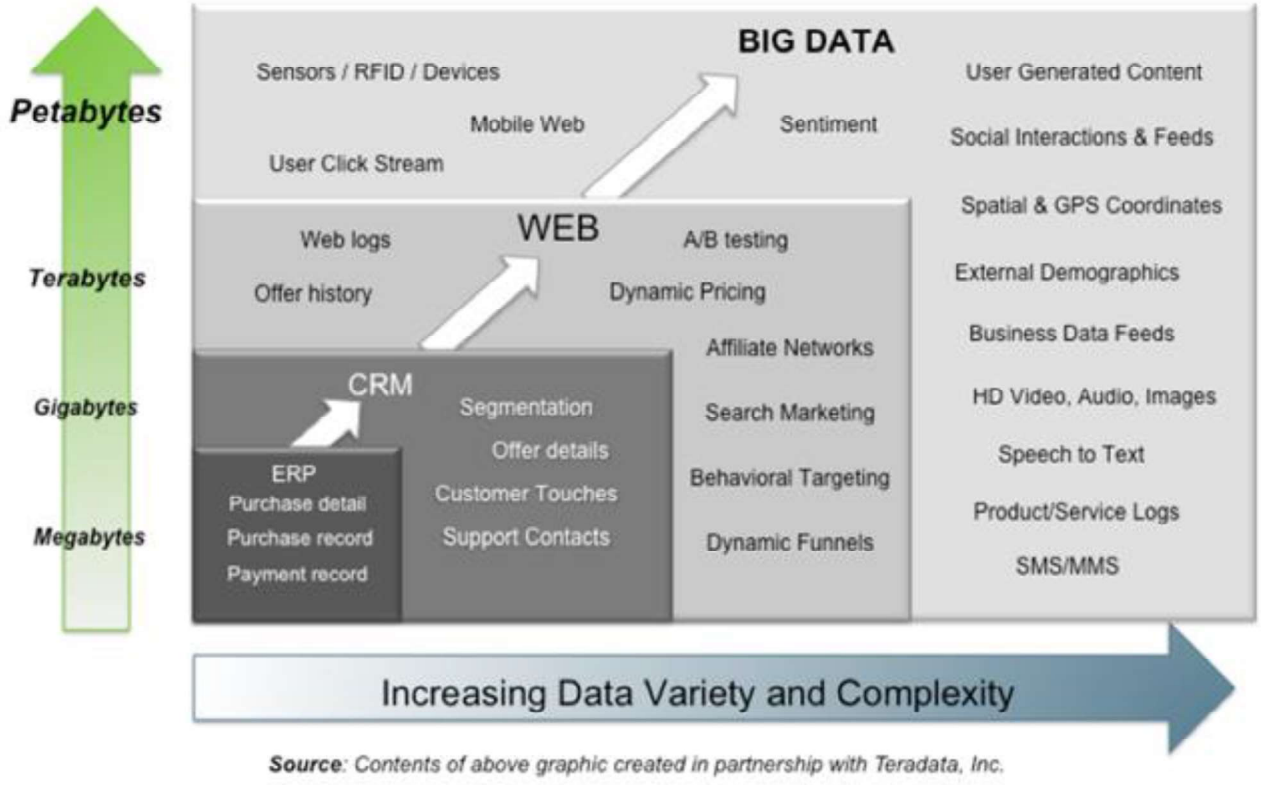


- In a narrow sense, Big Data means only sample size.
- In a broad sense, Big Data represents both sample size and dimensionality.

Course Overview

Definition of Big Data (Cont.)

- **3V's (Volume, Velocity, and Variety)**



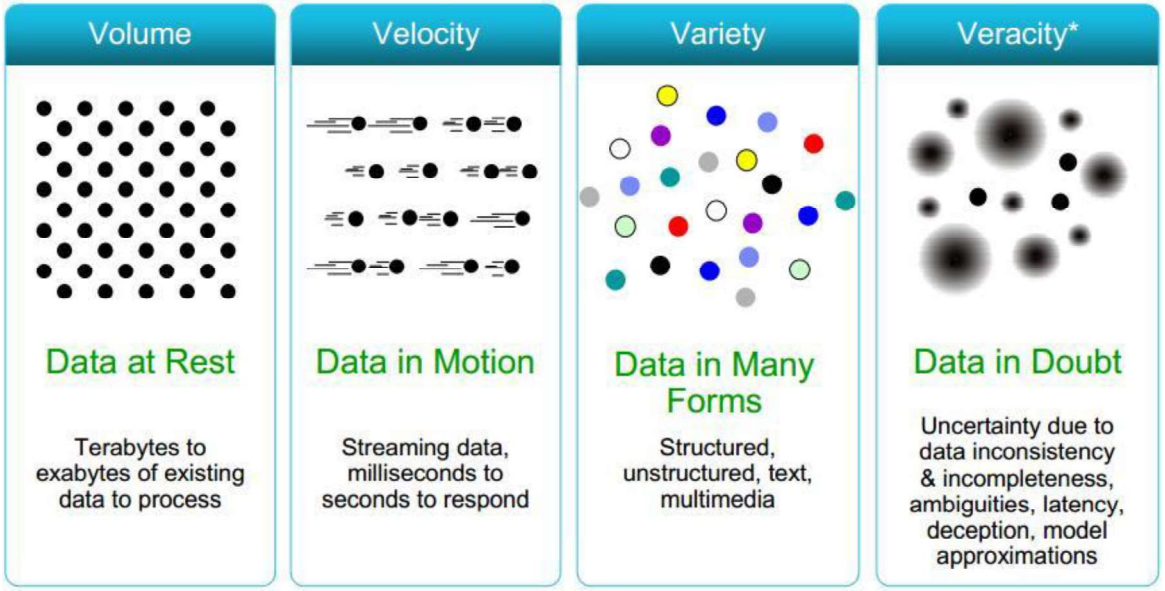
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Course Overview

Definition of Big Data (Cont.)

- **5V's (Volume, Velocity, Variety, Veracity, and Value)**

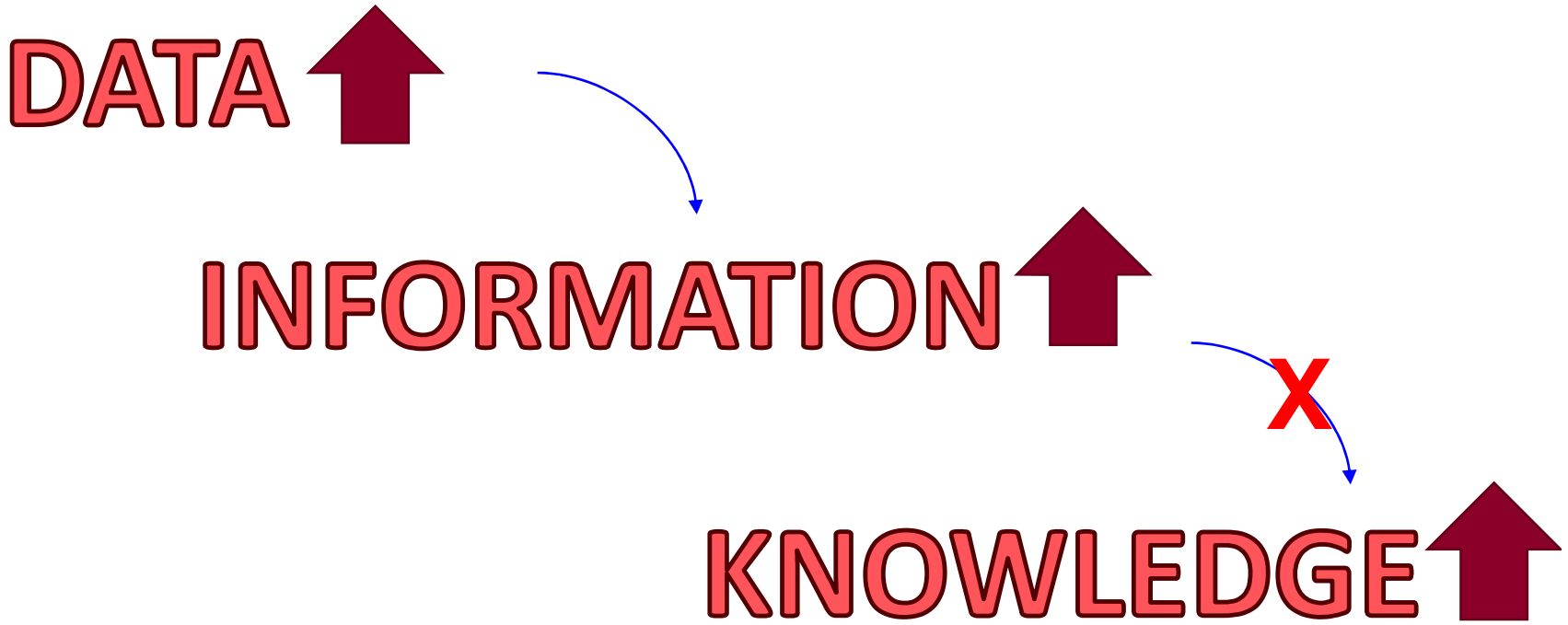
- **Volume: Data size**
- **Velocity: Data production speed**
- **Variety: Data oriented from various things**
- **Veracity: Data accuracy (Trustworthy)**
- **Value: Data value**



Value*

Course Overview

Relationship between Big-data & Data Science

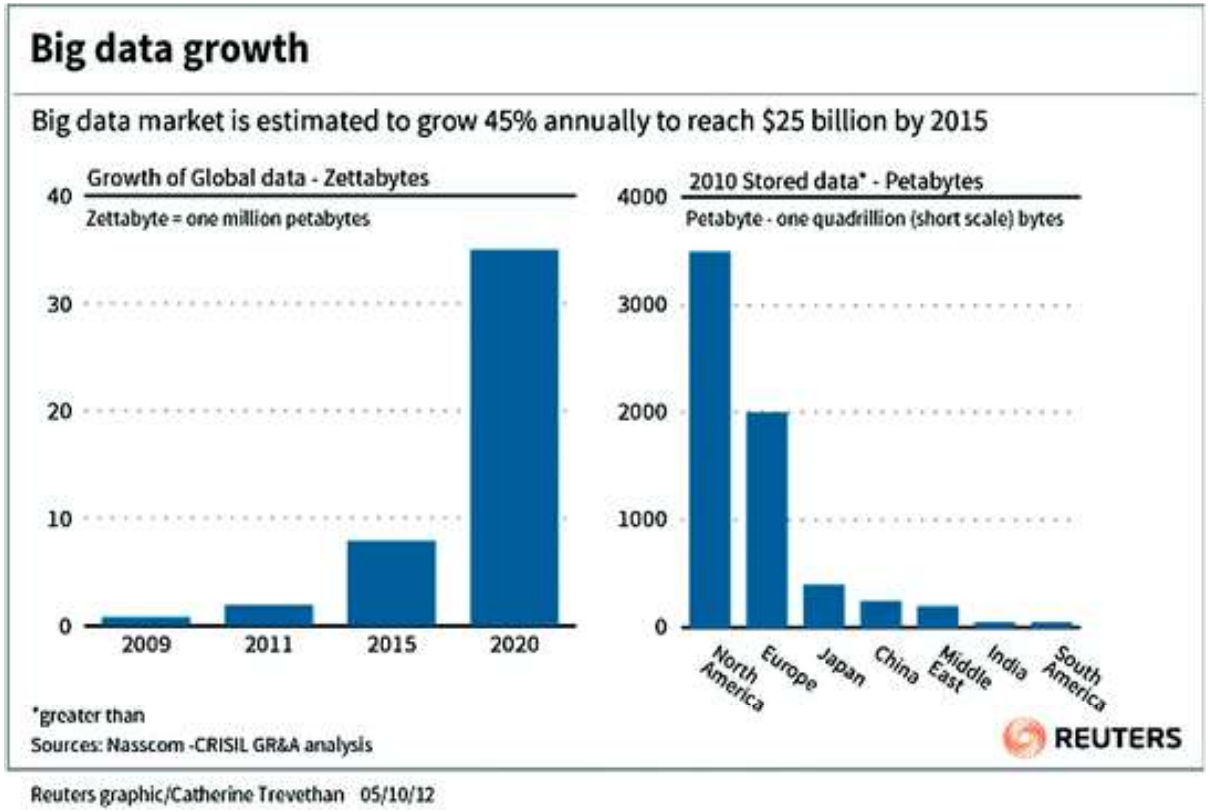


- **The amount of data and information is not directly correlated with knowledge generation.**
- **But the demand for data scientists will be growing.**

Course Overview

Job market of Big data

Furht B., Villanustre F. (2016) Introduction to Big Data. In: Big Data Technologies and Applications. Springer, Cham

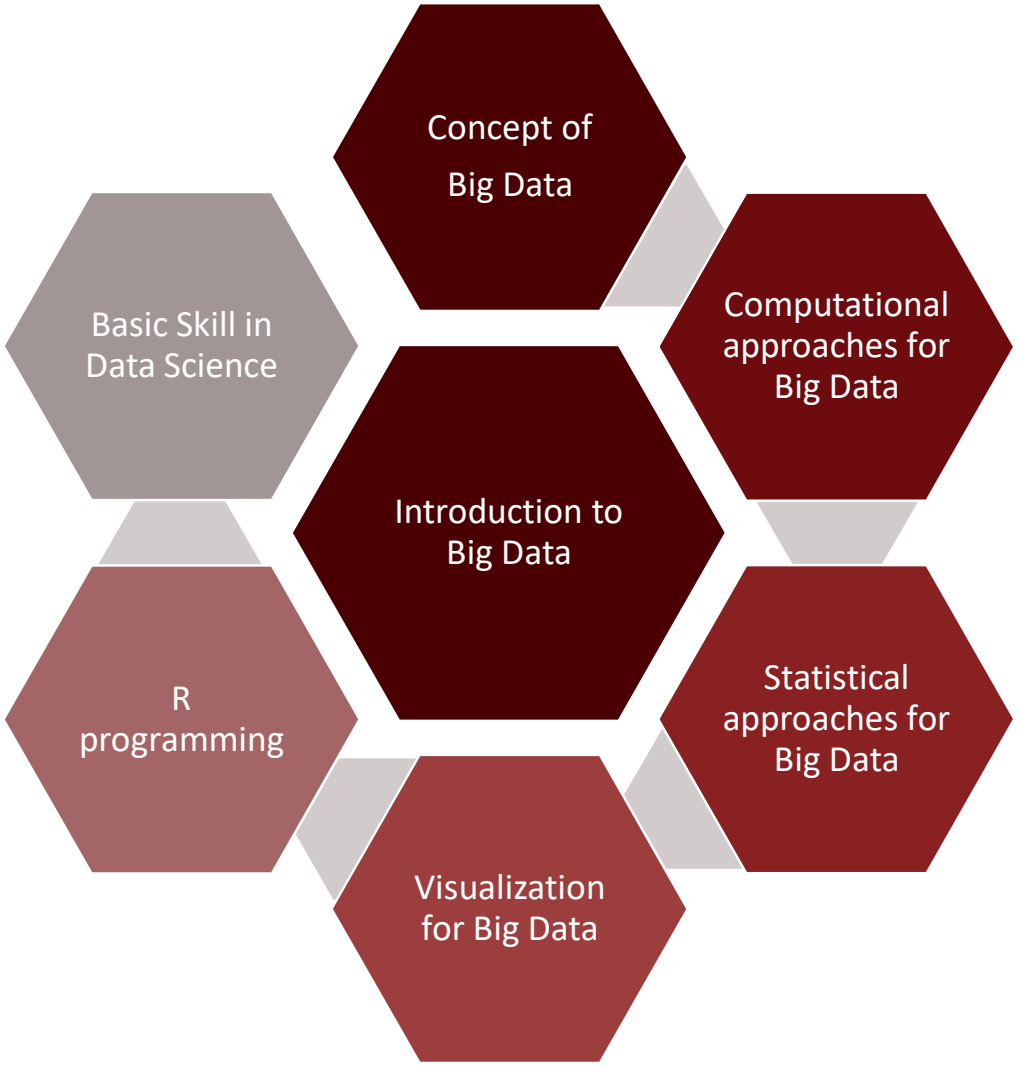


- It is the time to prepare for an academic course to cultivate data analysts commensurate with demand.

Course Overview

Object & Aim of the course

- **Students who have taken this course expect to be able to learn:**



Contents

1. Course Overview

- Brief introduction of professor & course
- Object & Aim of the course
- Assignments & Quiz
- Evaluation

2. Introduction to Big Data

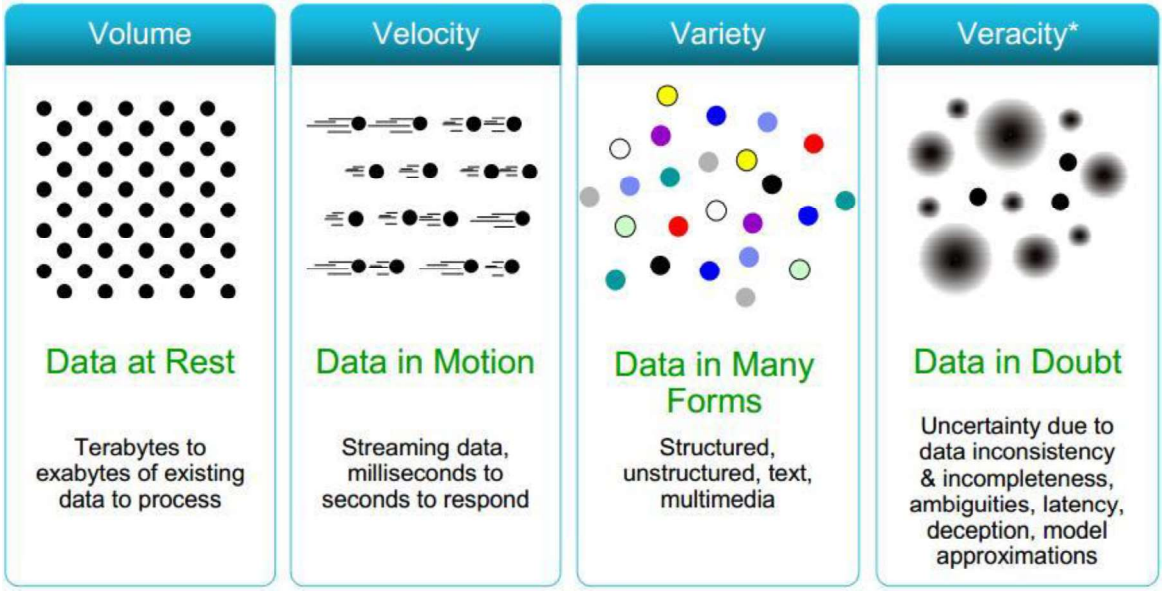
- Concept of Big Data
- Key techniques in Data Science for Big data

Characteristics of Big Data

Remind concept of Big Data

- **5V's (Volume, Velocity, Variety, Veracity, and Value)**

- **Volume: Data size**
- **Velocity: Data production speed**
- **Variety: Data oriented from various things**
- **Veracity: Data accuracy (Trustworthy)**
- **Value: Data value**



Value*

Petabyte era

1 PB = 1000000000000000B = 1015bytes = 1000terabytes

1000 PB = 1 exabyte (EB)

- AT&T transferred about 197 PB of data thorough its network each data (2018)
- processed about 24 petabytes daily (2009)

In fact, we can say that we have already entered the exabyte era.

Characteristics of Big Data

How do you recognize if it's big data or not?

Computer Scientist

- My computer is low on memory for handling this data!!

  That is Big Data

- No!!!! This data is over 2TB. Where do I store it?????

  That is Big Data

- In short, if you're having trouble with data processing on your computer (**멘붕에 빠지면**), it will be due to the Big Data.

Characteristics of Big Data

How do you recognize if it's big data or not?

Statistician

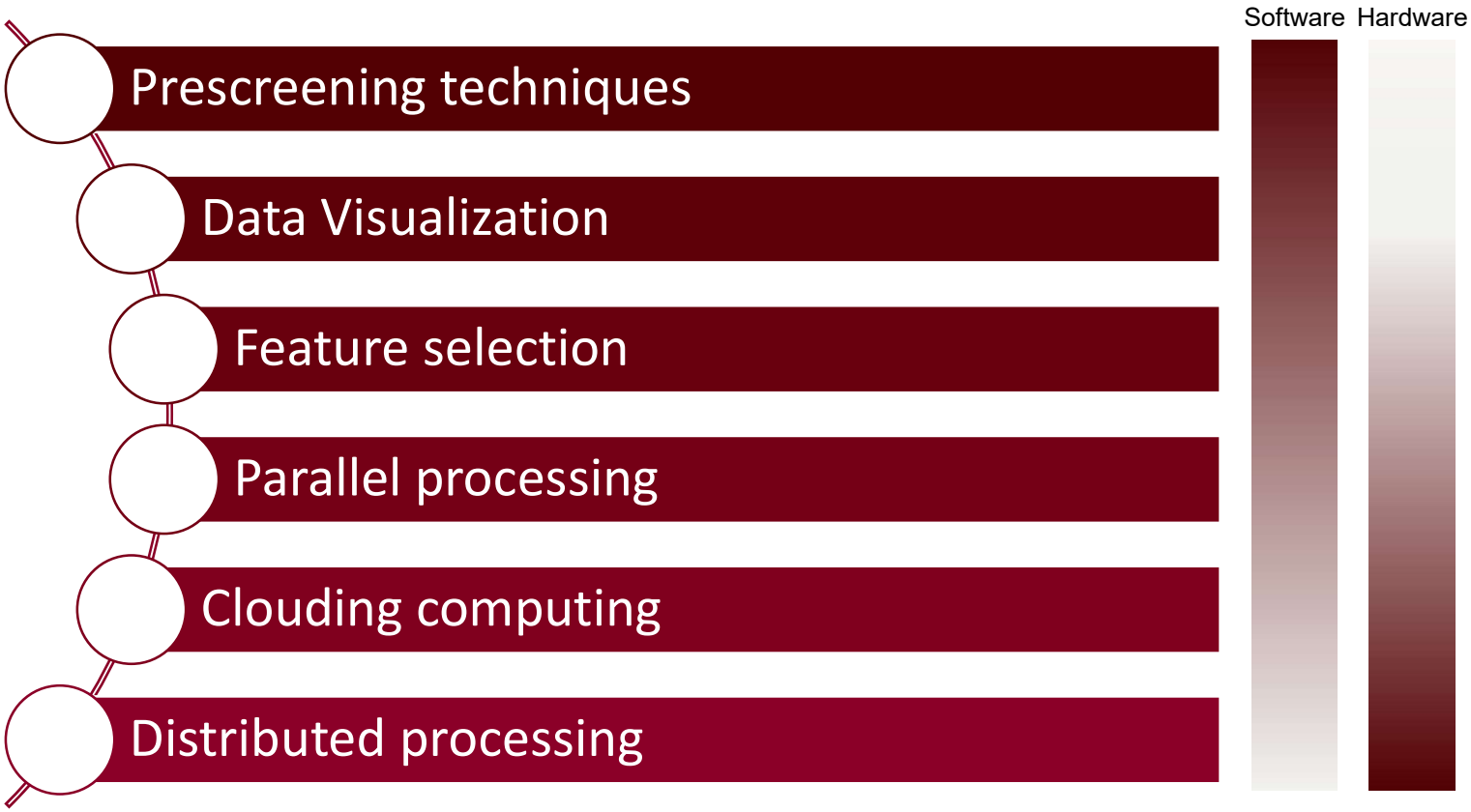
- When does this calculation end? I was only waiting for 10 years ...
- Dimensionality is too high!!!! I can't build statistical model using this data!!!

That is Big Data

- In short, if you're having trouble with data analysis on your computer (**멘붕에 빠지면**), it will be due to the Big Data.

Core technologies of Big Data era

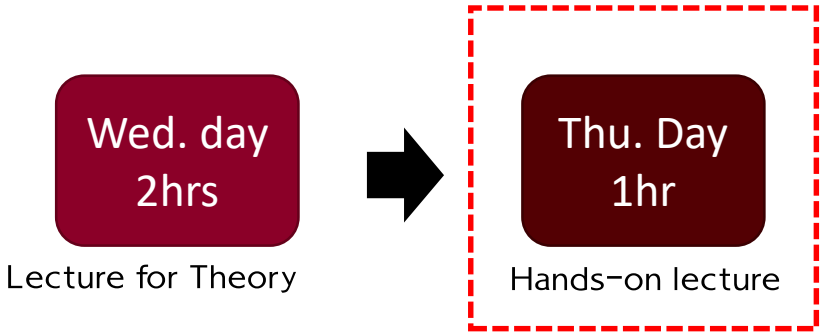
IT technologies to resolve issue derived from the Big data



- Difficulties arise in both hardware and software.
- But students can approach software difficulties.

Computational language for Big Data

R and Python



- There are two representative computer language for Big data analysis, R and Python.
- R programming language (free and relatively easy) for hands-on lecture.
- Let's connect R homepage

https://cran.r-project.org/