

빅데이터의 이해와 활용

# 오리엔테이션

방송대 통계·데이터과학과  
이금희 교수

- 이금희 교수(한국방송통신대 통계·데이터과학과)
- 관심분야 : 시계열분석, 비모수평활법, 국가통계
- 빅데이터를 이용한 지표 작성에 관심
  - IT인증기기 분석을 통한 기술동향 연구 (2007, 정보통신부)
  - 네이버 검색 경기지수 작성(2015, 한국은행)
  - 스캐너 데이터를 이용한 소비자물가지수 작성 (2017, 통계청)
  - 뉴스 데이터를 이용한 경제불확실성지수 작성 (2020, KDI)

# 1. 강의 개요

## ● 빅데이터의 이해와 활용

내용	강사	내용	강사
1. 빅데이터의 개요	이금희 (방송대)	7. 빅데이터 의사결정	함유근 (건국대)
2. 빅데이터의 수집과 활용		8. 빅데이터 기업 경영	
3. 텍스트 빅데이터	이준환 (서울대)	9. 빅데이터 기술	원중호 (서울대)
4. 빅데이터 시각화			
5. 추천시스템	김용대 (서울대)	10. 개인정보와 프라이버시 보호	이금희 (방송대)
6. 기계학습			

## 2. 어떻게 학습할까?

- 빅데이터 (Big Data)는 빠르게 발전 : 새로운 개념, 새로운 법률
  - 데이터과학 : 빅데이터와 관련된 학문
  - 변하는 세상에 열린 마음을 가지고 학습
- 이 교과목은 다른 통계·데이터과학과 교과목과 달리 통계학 및 수학의 사전 지식이 크지 않음
  - 코딩 실습은 하지 않음



## ■ 데이터과학, 어떻게 배울까?

≡

kaggle

+

Create

🏠

Home

🏆

Competitions

📁

Datasets

⌂

Code

💬

Discussions

🎓

Courses

✓

More

🔍

Search

Sign In

Register

🏛️

GettingStarted Prediction Competition

### Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

k

Kaggle · 14,166 teams · Ongoing

Overview

Data

Code

Discussion

Leaderboard

Rules

New Notebook

...

🔍

Search notebooks

≡

Filters

All


Your Work

Shared With You

Bookmarks

Hotness

▼



**Titanic\_Logistic-Regression\_ for beginners**


Updated 7h ago

1 comment · Titanic - Machine Learning from Disaster

▲

3

...

한국방송통신대학교  
Korea National Open University

# 01 강 빅데이터의 이해와 활용

## 빅데이터의 개요 1







# 학습목차

- 1 빅데이터 시대
- 2 빅데이터의 확산 배경
- 3 빅데이터의 정의



빅데이터의  
이해와 활용

# 1 빅데이터 시대





# 1. 빅데이터 시대

## ● 빅데이터(Big Data)

- 크고 다양한 형태, 빠르게 생산·유통·소비되는 데이터
- 빅데이터로부터 통찰
- 빅데이터는 21세기 원유

### < 빅데이터의 예 >

- SNS 데이터, 검색 데이터, 뉴스 데이터
- 사진, 동영상, 위치 데이터
- 유전체 데이터

# 1. 빅데이터 시대

- 지식(DIKW) 피라미드



# 1. 빅데이터 시대

- **데이터** : 데이터, 정보, 지식을 모두 포함

## <지능정보화기본법>의 데이터

- 부호, 문자, 음성, 음향 및 영상 등으로 표현된 모든 종류의 자료 또는 지식

# 1. 빅데이터 시대

## • 데이터의 구분

형태

정형, 비정형

보안

비밀, 민감, 일반

개인정보

식별, 비식별

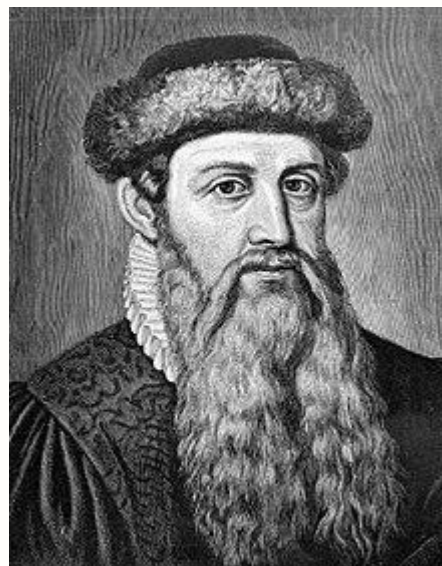
개인  
비개인

개인, 비개인(비식별, 기계)

# 1. 빅데이터 시대

## ● 데이터의 역사

- 1450년 : 구텐베르크 인쇄술의 대중화
- 1835년 : 케틀레의 통계
- 1991년 : 팀 버너스리의 인터넷
- 2007년 : 스티브 잡스의 스마트폰
- 2017년 : 테슬라의 모빌리티 데이터



출처 : 위키피디아



# 1. 빅데이터 시대

## ● 데이터의 역사

- 1450년 : 구텐베르크 인쇄술의 대중화
- 1835년 : 케틀레의 통계
- 1991년 : 팀 버너스리의 인터넷
- 2007년 : 스티브 잡스의 스마트폰
- 2017년 : 테슬라의 모빌리티 데이터



출처 : 위키피디아

# 1. 빅데이터 시대

## ● 데이터의 역사

- 1450년 : 구텐베르크 인쇄술의 대중화
- 1835년 : 케틀레의 통계
- 1991년 : 팀 버너스리의 인터넷
- 2007년 : 스티브 잡스의 스마트폰
- 2017년 : 테슬라의 모빌리티 데이터

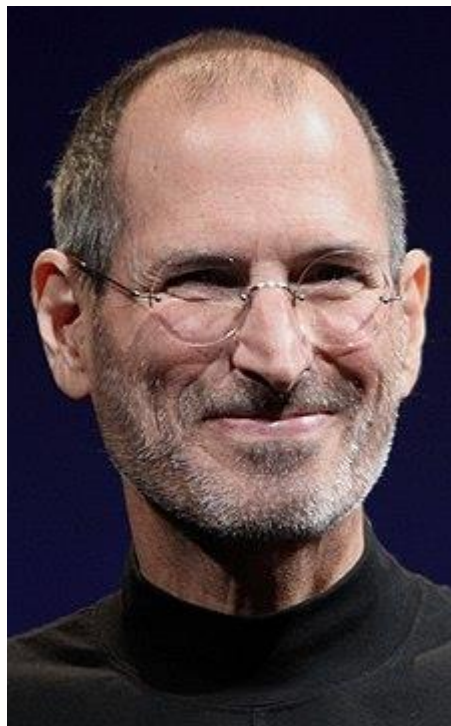


출처 : 위키피디아

# 1. 빅데이터 시대

## ○ 데이터의 역사

- 1450년 : 구텐베르크 인쇄술의 대중화
- 1835년 : 케틀레의 통계
- 1991년 : 팀 버너스리의 인터넷
- 2007년 : 스티브 잡스의 스마트폰
- 2017년 : 테슬라의 모빌리티 데이터



출처 : 위키피디아

# 1. 빅데이터 시대

## ○ 데이터의 역사

- 1450년 : 구텐베르크 인쇄술의 대중화
- 1835년 : 케틀레의 통계
- 1991년 : 팀 버너스리의 인터넷
- 2007년 : 스티브 잡스의 스마트폰
- 2017년 : 테슬라의 모빌리티 데이터

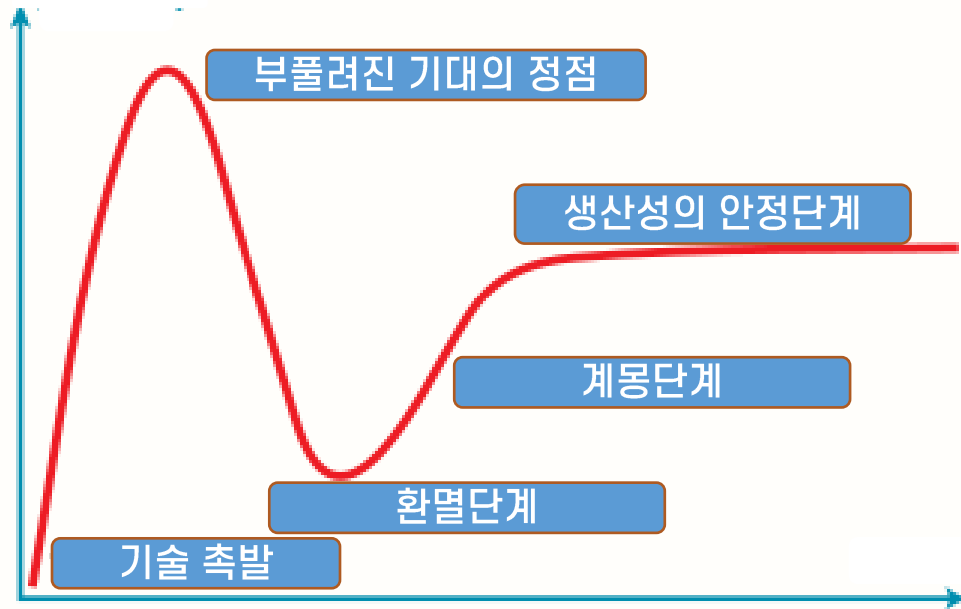


출처 : 위키피디아



# 1. 빅데이터 시대

- 가트너(Gartner)의 **하이프 사이클(Hype Cycle)**
  - 기술 촉발 → 부풀려진 기대의 정점 → 환멸단계  
→ 계몽단계 → 생산성 안정 단계

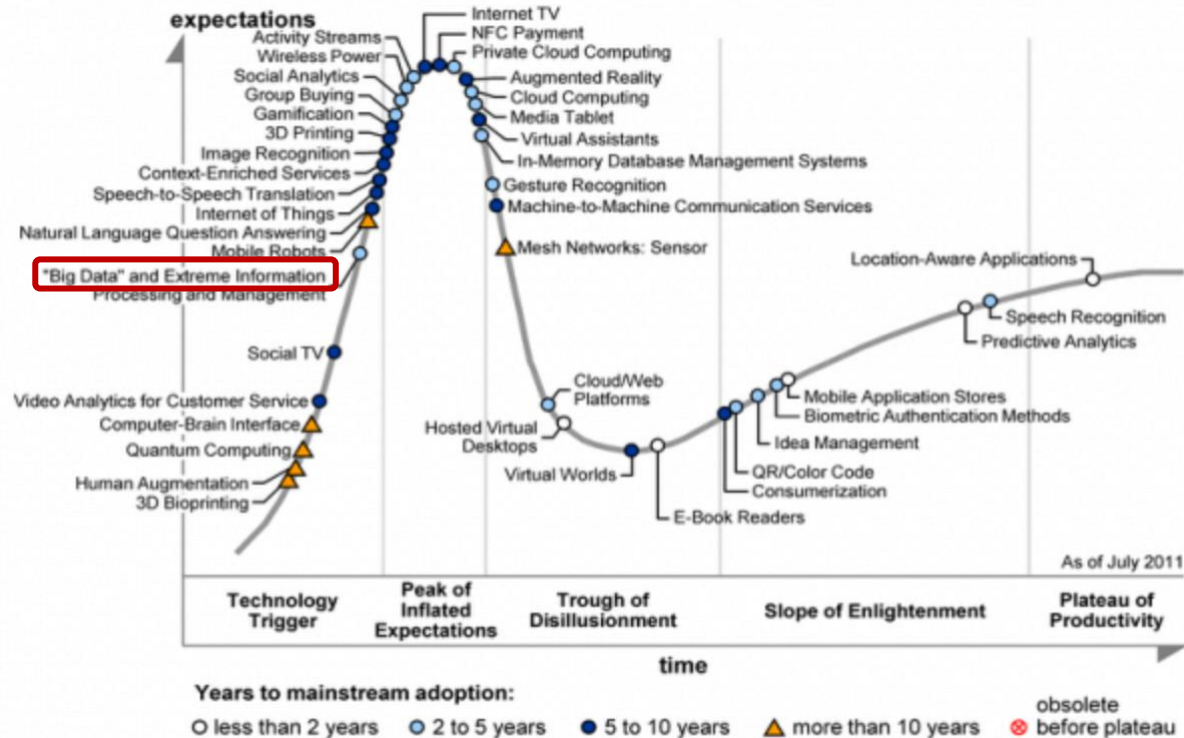




# 1. 빅데이터 시대

## 가트너(Gartner)의 하이프 사이클

2011년

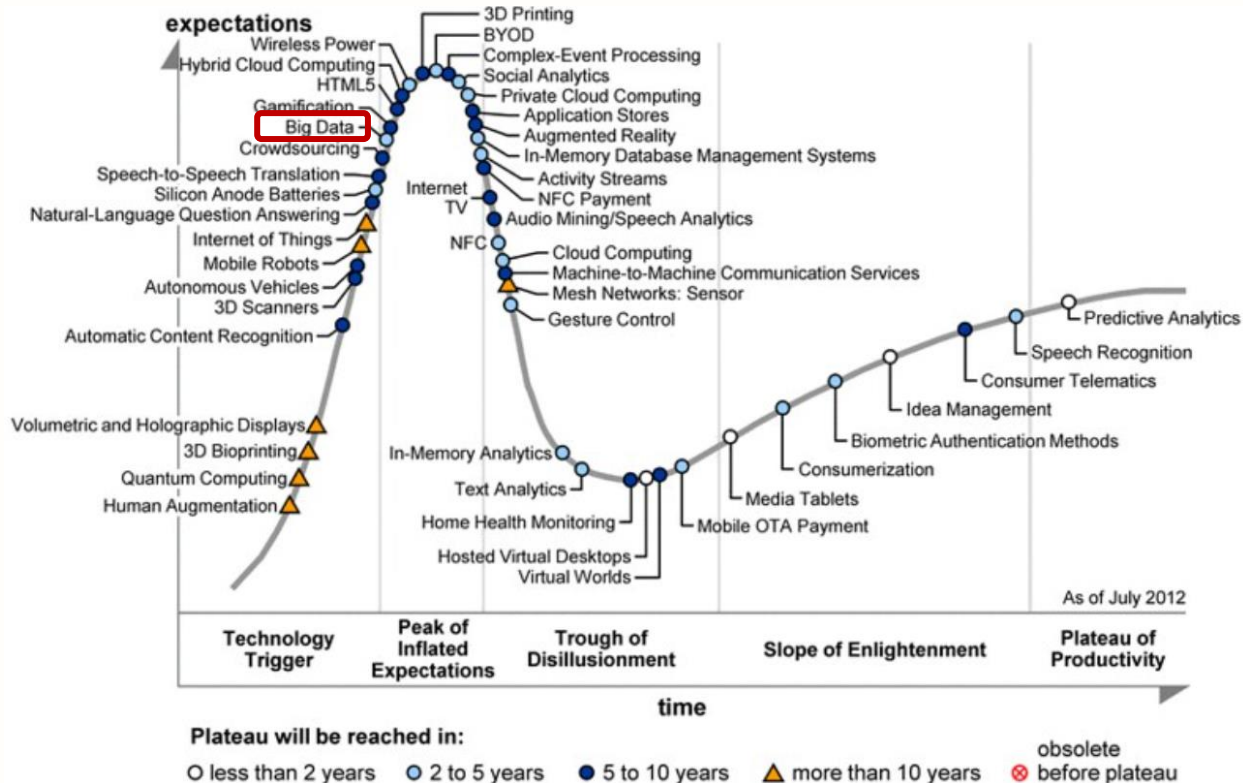


출처 : Gartner

# 1. 빅데이터 시대

## 가트너(Gartner)의 하이프 사이클

2012년

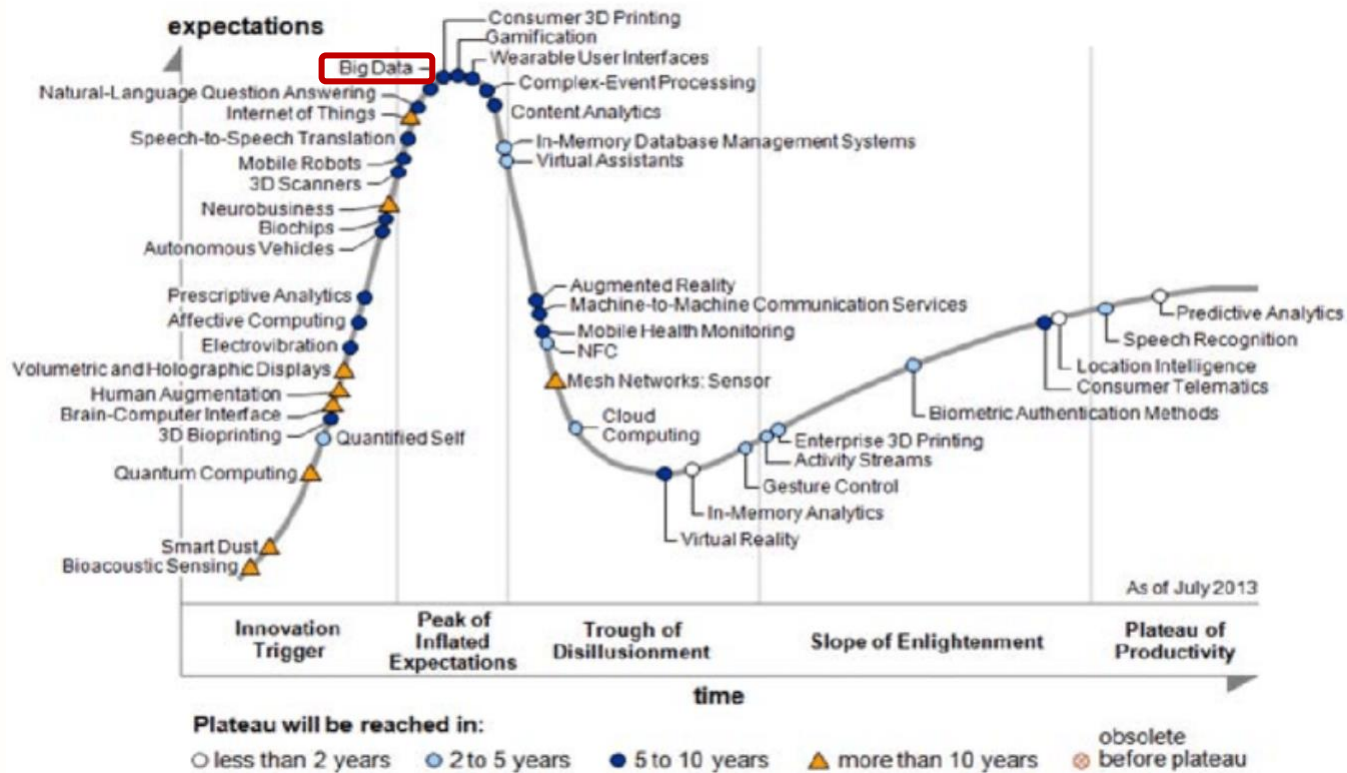


출처 : Gartner

# 1. 빅데이터 시대

## 가트너(Gartner)의 하이프 사이클

2013년



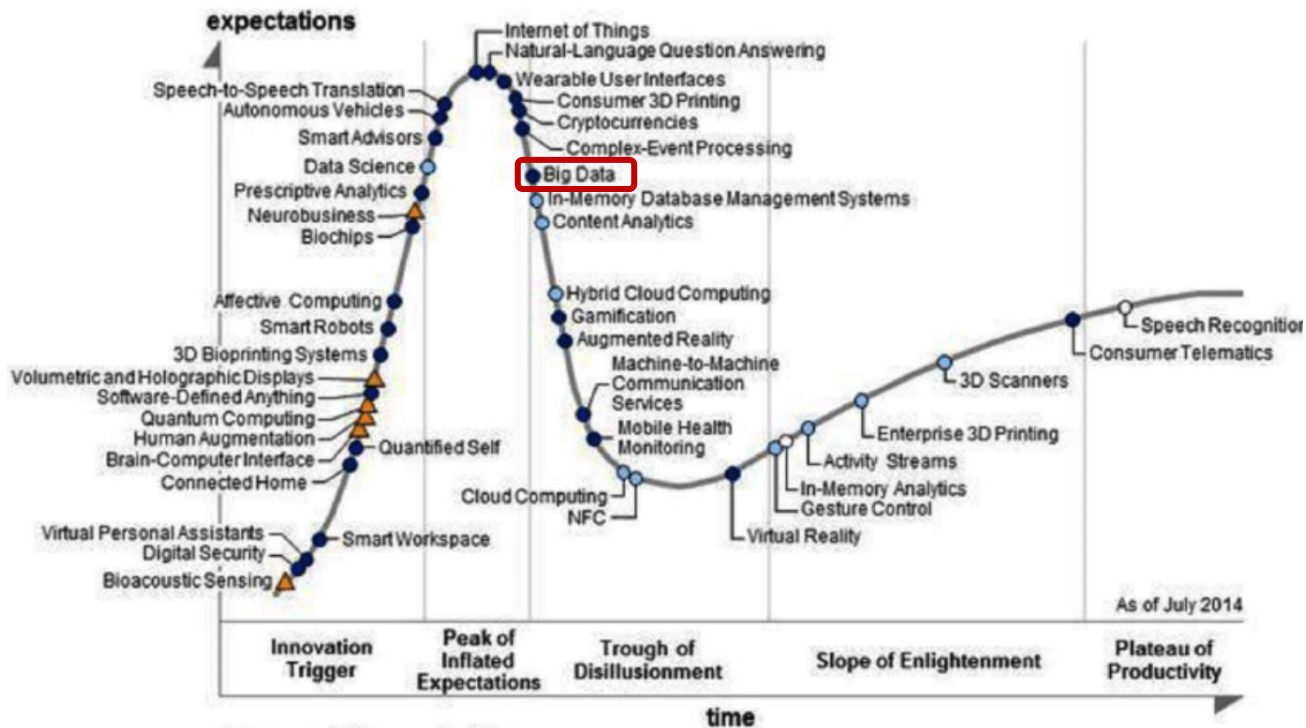
출처 : Gartner



# 1. 빅데이터 시대

## 가트너(Gartner)의 하이프 사이클

2014년



출처 : Gartner

# 1. 빅데이터 시대

## 가트너(Gartner)의 하이프 사이클

2015년

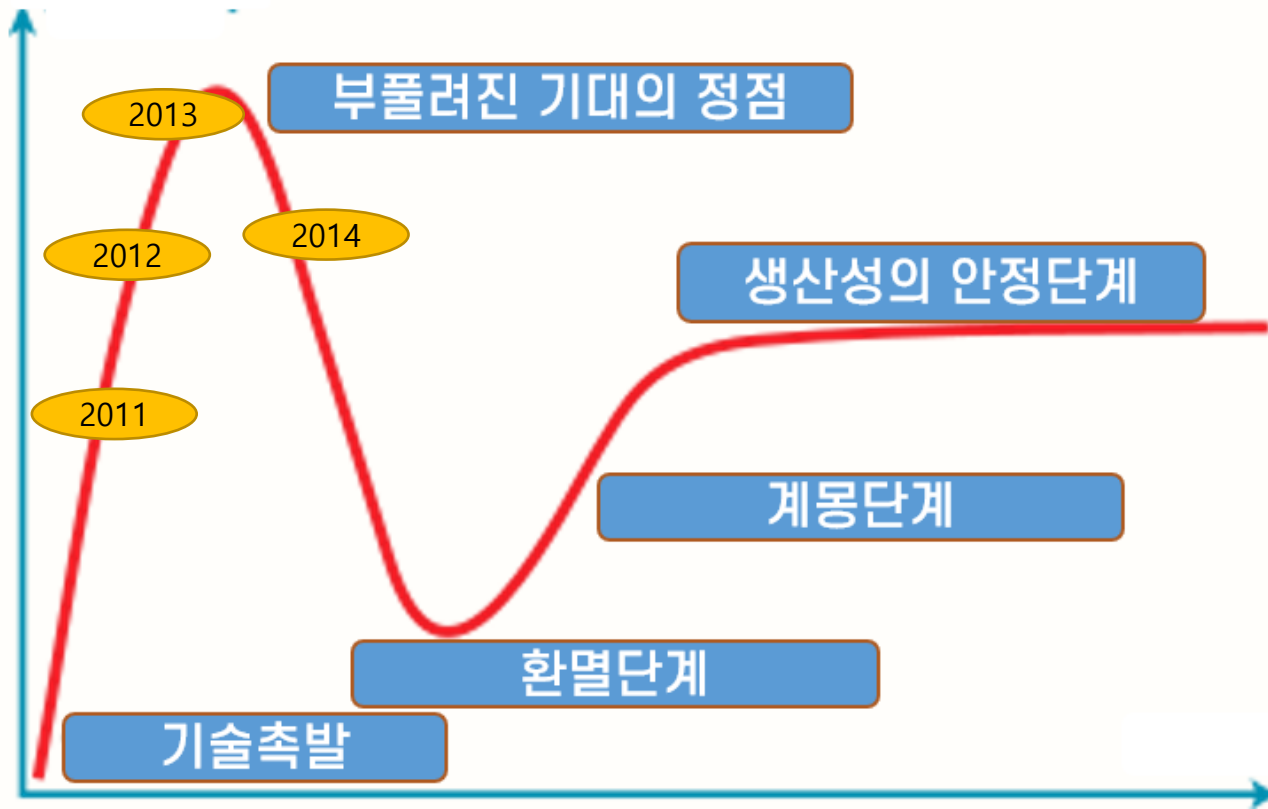


출처 : Gartner



# 1. 빅데이터 시대

- 가트너(Gartner)의 **하이프 사이클** : Big Data



출처 : 위키피디아

# 1. 빅데이터 시대

## • 데이터의 변화

### 2000년대

- PC 기반 데이터
- 검색 데이터
- 블로그, 뉴스 등 텍스트 데이터

### 2010년대

- 2000년대 데이터
- 모바일 데이터
- 사진, 동영상
- 센서 데이터

### 2020년대

- 2010년대까지 데이터
- 모빌리티 데이터
- 메타버스 데이터

빅데이터의  
이해와 활용

## 2 빅데이터의 확산 배경



## 2. 빅데이터의 확산 배경

### ● 빅데이터의 확산 배경 개요

스마트  
기기

모바일 스마트 기기와 센서의 확산

인프라

빅데이터를 저장, 처리할 수 있는 하드웨어  
인프라와 통신의 빠른 발전

소프트  
웨어

소프트웨어의 빠른 발전

## 2. 빅데이터의 확산 배경

### ● 센서 포함한 **스마트 기기**의 확산

- 스마트폰의 센서 : 위치, 온도, 습도, 조도, 지자기, 자이로 센서
  - 센서 데이터 스마트폰으로 연결 → 빅데이터 → 의사결정
- 기계들 IoT 통해 연결, 데이터 축적 → 최적의 결정
  - 스마트공장, 자율주행차(카메라, 레이더, 라이다)



## 2. 빅데이터의 확산 배경

- 빅데이터 기반 하드웨어 및 네트워크의 고도화
  - CPU, 메모리, 하드디스크 성능의 지수적 향상, 가격은 하락
    - GPU와 같은 새로운 연산 프로세서 등장
  - 클라우드 컴퓨팅 : IT 자원 가상화 형태의 인터넷 서비스
    - 병렬 연결된 서버에서 분산처리
    - AWS, Azure, GCP, NCP
    - 클라우드 컴퓨팅 기반 서비스

## 2. 빅데이터의 확산 배경

- 빅데이터 기반 **하드웨어 및 네트워크**의 고도화

- 유무선 네트워크 고도화 → 초고속, 초저지연성
  - 가상현실, 자율주행, 스마트 시티, 메타버스 등이 가능
  - 유선 인터넷 속도 : 1Mbps(1998년) → 10Gbps(2018년)
  - 이동통신 속도 : 9.6Kbps(90년대 중반) → 20Gbps(2020년)

## 2. 빅데이터의 확산 배경

- 빅데이터 기반 **소프트웨어**의 발전
  - 빅데이터 저장 및 처리 기술 : 하둡(Hadoop), 스파크(Spark) 등
  - 딥러닝 기술 : Tensorflow, PyTorch, Keras 등
  - 데이터 기반 기계학습, 딥러닝 기반 각종 서비스
    - 챗봇, 번역, 음성인식, 추천시스템 등



## 2. 빅데이터의 확산 배경

- 엔드류 응(Andrew Ng) 교수

- 빅데이터 기반 인공지능은 20세기 산업발전의 전기 역할



출처 : 위키피디아

빅데이터의  
이해와 활용

### 3 빅데이터의 정의





### 3. 빅데이터의 정의

- **빅데이터** : 새로이 생긴 데이터를 저장, 분석, 처리, 활용하는 하드웨어, 소프트웨어 기술이 바뀌면서 생긴 용어
- 가트너(Gartner)의 레이니(Doung Laney)는 **빅데이터의 속성**을 3V로 정의
  - 3V : 규모(Volume), 다양성(Variety), 속도(Velocity)
  - 5V : 3V + 정확성(Veracity), 가치(Value)

### 3. 빅데이터의 정의

#### ● 규모(Volume)

- 빅데이터는 크다는 ‘빅(Big)’ : 빅데이터로부터 가치를 얻기 위한 속성
- 데이터가 커지면 정확성이 높고 세분화된 분석 가능
- 데이터 규모가 커야만 의미 있는 머신러닝, 딥러닝 모델을 작성

영문단위	값(이전단위 기준)	값(bytes)
Kilobytes(KB)	1,000 B(bytes)	$10^3$ B
Megabytes(MB)	1,000 KB	$10^6$ B = $1,000^2$ B
Gigabytes(GB)	1,000 MB	$10^9$ B = $1,000^3$ B
Terabytes(TB)	1,000 GB	$10^{12}$ B = $1,000^4$ B
Petabytes(PB)	1,000 TB	$10^{15}$ B = $1,000^5$ B
Exabytes(EB)	1,000 PB	$10^{18}$ B = $1,000^6$ B
Zettabytes(ZB)	1,000 EB	$10^{21}$ B = $1,000^7$ B
Yottabytes(YB)	1,000 ZB	$10^{24}$ B = $1,000^8$ B

### 3. 빅데이터의 정의

#### ● 다양성(Variety)

- 정형/비정형/반정형 데이터로 구분
  - 정형 데이터 : 엑셀데이터 등
  - 비정형 데이터 : 사진, 동영상, 음성, 텍스트 등
  - 반정형 데이터 : HTML, XML, JSON
- 기업 : 정형 데이터(관계형 DB) 중심 → 비(반)정형 활용
  - 전세계 데이터 중 비정형/반정형 데이터 비중이 80% 이상

### 3. 빅데이터의 정의

- 속도(Velocity)

- 유무선 네트워크 환경 고도화
  - 데이터의 생성-유통-소비 주기가 빨라짐
  - 빅데이터 시대의 뉴스
    - : 신문(방송)사 취재보다 트위터, 유튜브가 빠름



### 3. 빅데이터의 정의

- **정확성(Veracity)과 가치(Value) → 5V**
  - 빅데이터는 규모가 크고 다양한 형태로 수집되지만 정확성은 낮음
    - 정확성(Veracity)이 크다면 분석 결과를 더 신뢰
  - 빅데이터로부터 가치(Value)를 얻어야 의미 있는 의사결정이 가능

### 3. 빅데이터의 정의

#### ● 빅데이터의 정의

- 협의의 정의 : 3V 또는 5V로 정의
  - 큰 규모, 다양한 형태, 생성-유통-소비가 빨라서 기존 방식으로 관리·분석이 어려운 데이터
- 광의의 정의
  - 빅데이터로부터 의미 있는 가치를 도출할 수 있는 빅데이터 관련 기술, 인력, 조직과 인프라 포함

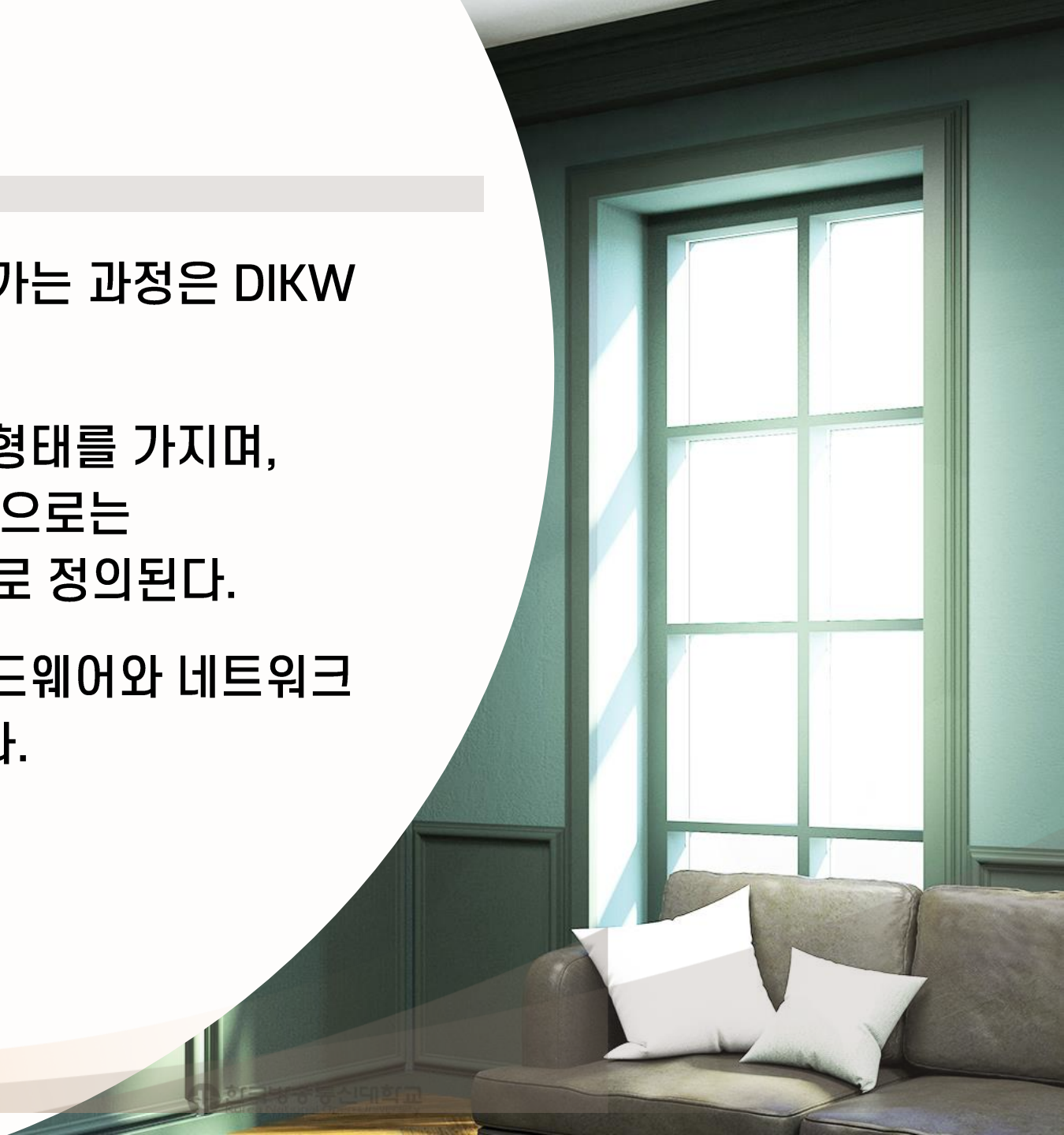


## 교재 읽기 : 3p

빅데이터에는 나름의 숨겨진 패턴과 이야기가 있고, 이로부터 우리는 새로운 세상을 열 수 있는 통찰(insight)과 지혜(wisdom)를 얻을 수 있다. 빅데이터로부터 통찰을 얻으려면 우리는 데이터를 수집·저장하는데 한정하지 않고, 데이터 시각화, 데이터분석, 머신러닝과 딥러닝 등을 통해 데이터에 숨겨진 패턴을 찾아야 한다. 기업과 정부는 빅데이터에 존재하는 패턴을 찾아서 혁신성장, 효율화가 가능한 새로운 서비스를 만들어서 제공하고, 개인은 맞춤형 서비스를 통해 편리한 생활을 누린다. 빅데이터는 새로운 생산요소가 되어서 빅데이터를 21세기 원유라고 부른다.

## 정리하기

- 데이터로부터 정보, 지식과 지혜를 만들어 가는 과정은 DIKW 피라미드로 표현된다.
- 빅데이터란 데이터의 규모가 크고, 다양한 형태를 가지며, 생성-유통-소비가 매우 빨라서 기존의 방식으로는 관리·분석이 어려운 데이터로 3V 또는 5V로 정의된다.
- 빅데이터의 확산은 스마트 기기의 확산, 하드웨어와 네트워크 고도화, 관련 소프트웨어의 발전에 기인한다.





02

강

다음시간안내

## 빅데이터의 개요 2

수고하셨습니다!

