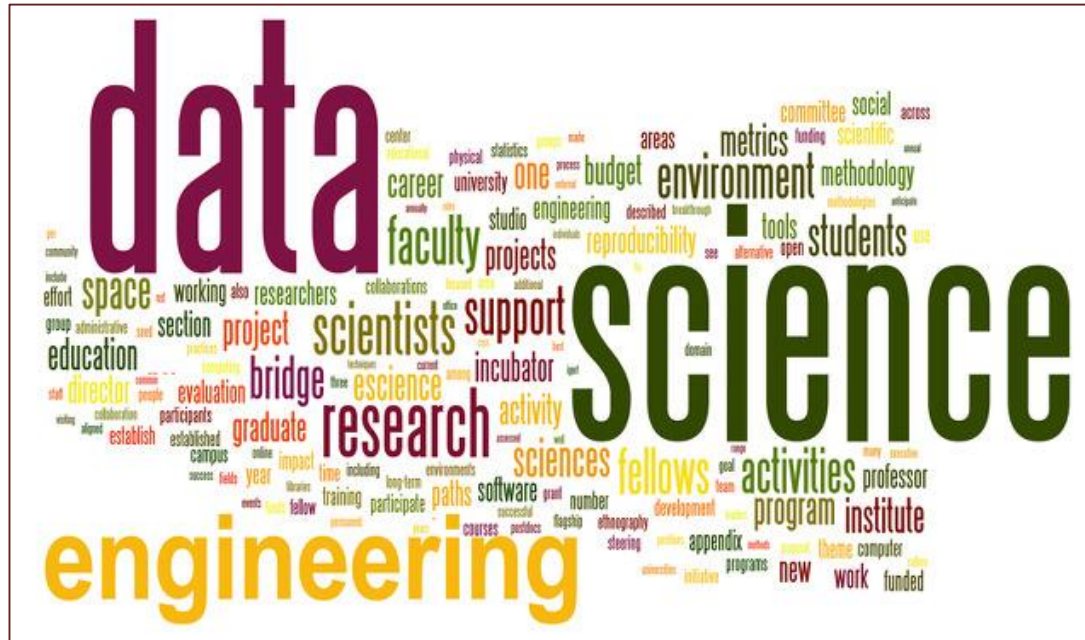


# 데이터사이언스



호서대학교 빅데이터경영공학부 연구필

---

# 데이터 사이언스

---

**1.1 데이터분석의 이해**

**1.2 빅데이터 분석**

**1.3 통계학과 기계학습**

**1.4 데이터 분석 툴**



# 1.1 데이터 분석의 이해

## ■ 데이터와 수량화의 역사

- 메소포타미아의 5,000년 된 점토판
  - 기록내용: 보리 29,086 자루, 37개월, ???
- : 유발 하라리의 <Sapiens: A brief history of humankind>

## ■ 수량화의 열기

- 유럽의 전 도시는 과도한 재정 지출을 감내하면서도 최고급 시계(시간을 나타내는 측정)의 설치 경쟁에 뛰어 들었다. 시간의 수치화가 도시의 클래스를 보여주는 기준이었다 (유럽 여행에서 주로 보는 시계탑 들)
- 출생, 사망 등의 수치 기록이 이후에 물리 등의 발전에 영향을 줌.
- : <수량화 혁명: 유럽의 패권을 가져온 세계관의 탄생> 엘프리드 W. 크로스비(지은이)



# 1.1 데이터 분석의 이해

## ■ 숫자 표기

- 로마숫자 : I, II, III, IV, V, VI, VII, VIII, IX, X, ...
- 아라비아숫자 : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ...
- ✓ 16세기 이탈리아 메디치가(무역상 집)는 로마식 숫자 표기가 유치하다는 것을 인정하고 인도의 아라바아 숫자로 부기를 하기 시작하였다.
- ✓ 전 유럽에 확산하여 실질적 숫자로 아라비아 숫자를 사용하는 인류 최대의 문화적 혁명을 가져왔다.
- ✓ 부실한 숫자 체계에서는 과학혁명 발전이 어렵다.

## ■ 수량화의 열기

- 세금 제도에 대한 여론조사, 세종 12년 (1430년)
- ✓ 표본 크기 172, 648명 (전국 8도)
- ✓ 결과: 개혁안 찬성 57%, 반대 43%
- 측우기 사례: 측정의 안정성 (세계 최초), 그릇으로 재면 어디서 재든 안정한 결과를 준다. 땅을 뚫으면 땅 종류에 따라 변이가 크다.



# 1.1 데이터 분석의 이해

## ■ 데이터 과학의 역사

- 국가(state)의 산술(arithmetic): 인구, 수명, 농업생산, 군사, ...->통계(statistics)
- 확률론이 적용되면서 통계학(statistics)
- 컴퓨터(+ IT)의 발달로 데이터 과학(data science) -> Big Data 가 다시 시작 됨
- ✓ 데이터의 시작은 통계 용어에서 볼 수 있듯이 빅데이터가 시작이다:
- ✓ Big data -> Small data가 활발해 짐 : 조사(survey)(보통 1,000표본 정도, 비교실험(100샘플 정도) -> 다시 최근에 Big data 가 다시 revival 된 것임.

## ■ 데이터 분석의 유형

- 분석(分析)의 의미: 나누고 쪼갬다. 이를 통해 데이터 안에 담겨져 있는 내용과 의미를 알아가는 것
- EDA(exploratory data analysis: 탐색적 데이터 분석) : 데이터의 특징과 구조에 대한 탐구, 先데이터, 後분석 → 빅데이터 분석은 EDA에 가깝다.
- CDA(confirmatory data analysis:확증적 데이터 분석): 가설, 모형의 타당성, 일반성, 재현성 평가, 모형 적합도, 가설검정, 계획->데이터 확보->분석



# 1.1 데이터 분석의 이해

## ■ EDA vs CDA

**보기 1.** 감기에 걸리는 사람들과 걸리지 않는 사람들 간에 어떠한 차이가 있는가를 수 십 가지 측면에서 살펴보았다. 그 결과, 비타민 C를 복용하는 사람들이 감기에 잘 걸리지 않음을 알게 되었다. 그러면, 비타민 C를 복용하면 감기에 덜 걸린다고 말할 수 있는가? [EDA]

- EDA로 이에 대한 답을 하긴 어렵다. 비교실험을 설계하여 새로 자료를 수집해 가설을 확인해 볼 필요가 있다 [CDA]

**보기 2.** 대형마켓에서 고객들의 구매 내역 자료를 분석한 결과, 일부 고객들은 다른 고객들에 비해 유기농 식재료 비중이 크게 나타났다. 그들이 어떤 생각을 하는 사람들인가? 이에 대해 몇 개의 추측이 생성되었다. [EDA]

- 추측이 맞는가? 이를 확인하기 위하여 전체 고객의 일부를 선택하여 몇 가지 인구사회학적 속성과 연소득, 그리고 소비와 삶에 대한 태도를 조사하여 구매내역과 연결해 확인해 볼 필요가 있다 [CDA]



# 1.1 데이터 분석의 이해

---

## ■ 데이터와 모형의 진화

- 복잡성(현실)의 이해 : 현실 세계는 많은 요인이 얽혀져 있다. 그렇다고 절대적으로 불가지(不可知) 한 것은 아님
- 데이터 : 우리가 포착한 복잡계의 한 단면이다.
- 모형(Model) : 모형은 우리가 현실을 이해하는 ‘틀’이다. 진리는 아니다.

“ All models are wrong but some are useful.” - George Box-

- 데이터와 모형의 진화(evolution)

데이터 ➔ 모형 ➔ New 데이터 ➔ New 모형 ➔ 계속적 진화



## 1.2 빅데이터 분석

### ■ 빅데이터(Big Data)

데이터의 양이 거대하고 형태도 다양해서 기존 방법으로 수집, 저장, 분석이 어려운 데이터  
=> 3V+1C(?)

- Volume(대규모)
- Velocity(실시간 생성)
- Variety(숫자, 문자, 영상 등)
- Complexity(복잡성)

➤ Value(가치 창출 ?)

- ⇒ 한 시점에서 굉장히 크다가 다른 시점에서는 별거 아닐 수도 있으므로 좋은 정의가 아니다. 즉, 한시점에서는 좋을 지 모르지만 넓게는 좋지 않다.
- ⇒ “범람하는(넘쳐나는) 데이터와 정보” : 이제까지 다루지 못한(안해본 것, 통찰력이 없어서) 데이터의 양, 융합의 어려움 등을 망라하여 정의되어야 한다.
- ⇒ ‘빅’ 데이터=데이터: ‘빅’ 데이터로부터의 정보와 지식=통찰력을 얻는 것. 새로운 것에 도전하는 정성과 실력(자질)이 필요함.

### ■ 스몰데이터(Small Data)

조사/실험 데이터 등 과학적이고 체계적임 → 재현성이 확인될 필요가 있을 때.



## 1.2 빅데이터 분석

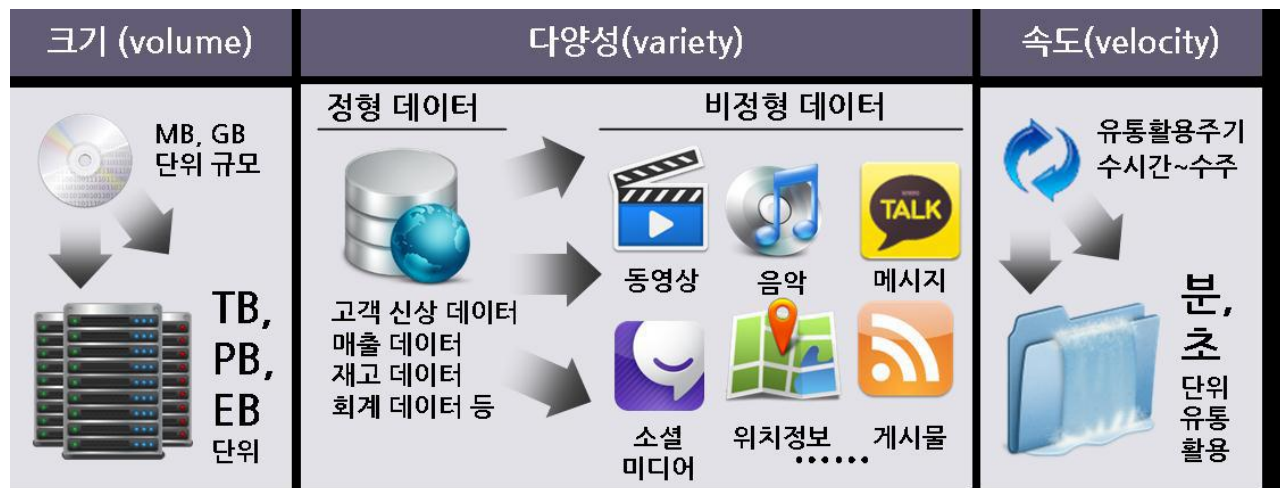
### ■ 빅데이터의 정의

- 삼성경제연구소

기존의 관리 및 분석 체계로는 감당할 수 없을 정도의 거대한 데이터의 집합으로 대규모 데이터와 관계된 기술 및 도구를 모두 포함하는 개념

- 국가전략위원회

대용량 데이터를 활용, 분석하여 가치 있는 정보를 추출하고 생성된 지식을 바탕으로 능동적으로 대응하거나 변화를 예측하기 위한 정보화 기술





## 1.2 빅데이터 분석

### ■ 예전의 데이터

이름	생년월일	신장
유리	89.12.05	167
효연	89.09.22	160
서현	91.06.28	168
수영	90.02.10	170
써니	89.05.15	158
태연	89.03.09	162
...	...	...

## 1.2 빅데이터 분석

### ■ 최근의 데이터



## 1.2 빅데이터 분석

### ■ 최근의 데이터



## 1.2 빅데이터 분석

### ■ 최근의 데이터





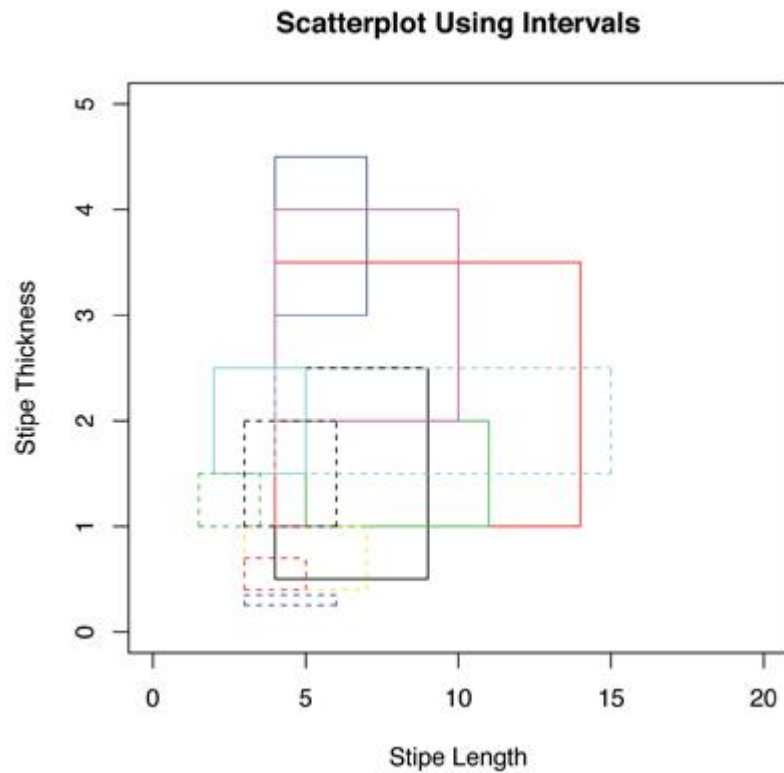


- 수치 데이터 5%
- 非수치 데이터 95% ➔ 새로운 분야의 데이터 분석 전문가가 필요  
텍스트, 이미지, Audio, Video, ...



## 1.2 빅데이터 분석

### Interval data의 회귀분석



출처: [https://chance.amstat.org/2013/09/big\\_picture\\_26-3/](https://chance.amstat.org/2013/09/big_picture_26-3/)

## 1.2 빅데이터 분석

### ■ 이미지 세그멘테이션



출처: [https://wiki.tum.de/display/lfdv/Image+ Semantic+ Segmentation](https://wiki.tum.de/display/lfdv/Image+Semantic+Segmentation)





## 1.2 빅데이터 분석

■ Small Data vs. Big Data : 실험(조사) 자료 vs. 관찰자료

	<u>Small Data(실험)</u>	<u>Big Data(관찰)</u>
목적	연구	업무활용
가치	과학	상업
수집	통제된 현재 자료	관찰된 과거 자료
크기	작다	크다
정도	정제되어 있다	정제되어 있지 않다
상태	정적	동적



## 1.2 빅데이터 분석

---

### ■ 사례 : 2차 세계대전의 생존 전투기 탄흔 기록

- 생존전투기의 탄흔 기록을 보면 전투기에서 탄흔이 없는 부분(그곳에 맞은)에 해당되는 전투기들은 돌아오지 않았다(이런 데이터는 없다)
- Big data는 관찰된 데이터라 편향된 데이터가 올 수 있다.
- 관찰된 부분의 데이터만 보면 매우 위험할 수 있다.
- 관측 데이터는 특정 목적에 맞게 기획되어 수집된 자료가 아니라 표본이 모집단을 대표하지 않는다
- Small data(조사 혹은 실험 데이터)는 데이터를 잘 디자인하여 균형 있게 추출하여 이런 문제점이 없다
- 즉, 이 사례는 빅데이터의 기계학습 예측의 위험을 주의해야 됨을 말해 준다



## 1.2 빅데이터 분석

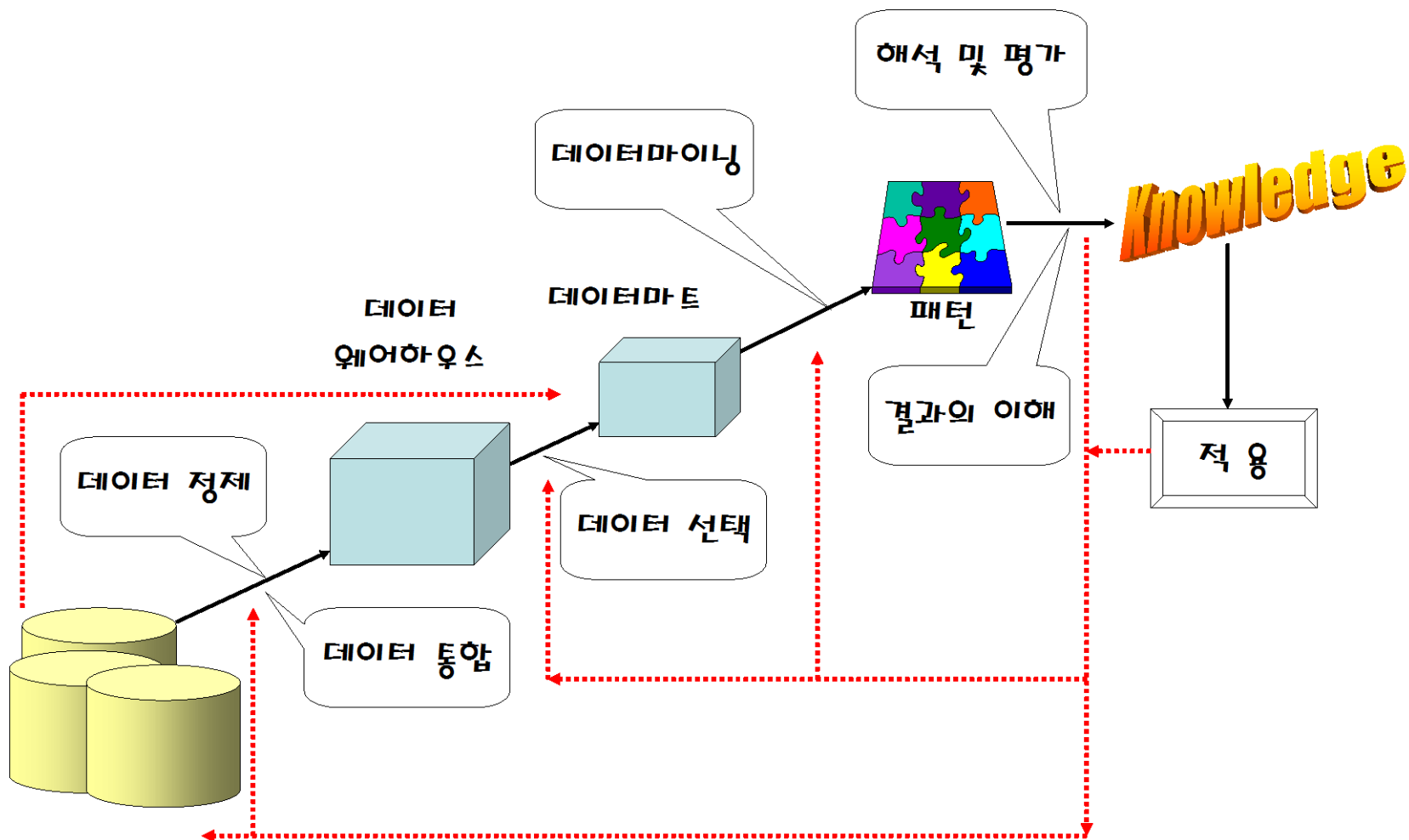
---

### ■ 사례 : 구글의 독감 트렌드 (Google Flu Trend)

- Ignaz Semmelweis (산부인과 의사), 1844 오스트리아 빈
- ✓ 2008년, 독감과 관련 검색어의 빈도를 집계 독감을 예측하는 서비스
- ✓ 미국 질병관리본부(CDC)보다 1주 이상 빠르게 독감의 유행을 예측 (Nature지에서 소개)
- ✓ 2009년, 신종 인플루엔자(H1N1)의 세계적 유행을 놓침
- ✓ 2013년, 실제 발생률의 2배에 달하는 예측치 → 서비스 중단
- ✓ 2008년 사이트를 조회하면서 조회 수가 늘고 독감이 과대 예측됨
- ✓ 예측(prediction)과 설명(cause-and effect)의 문제를 인식할 필요가 있다.

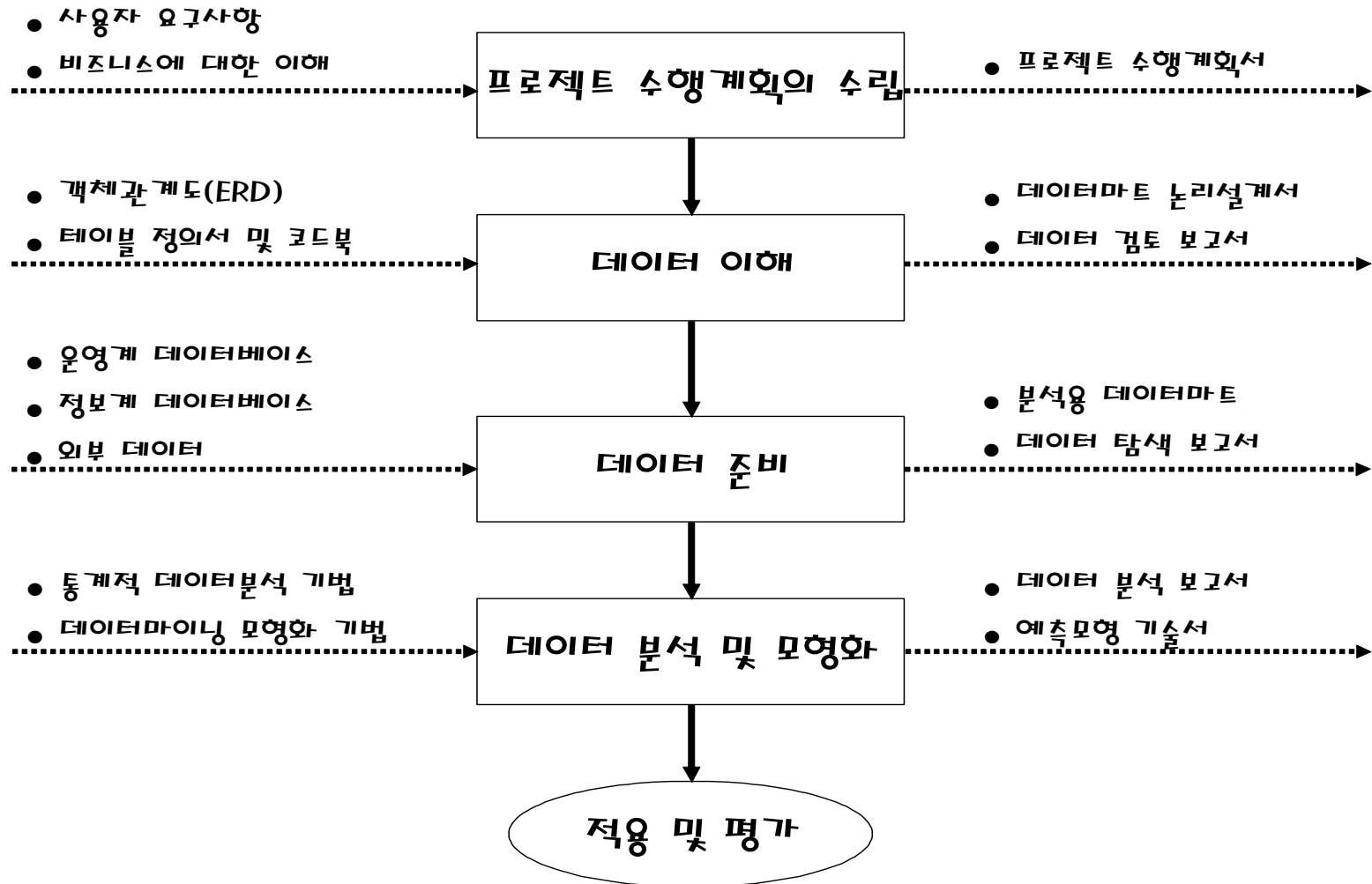
## 1.2 빅데이터 분석

### ■ DB로부터의 지식발견(KDD) 과정



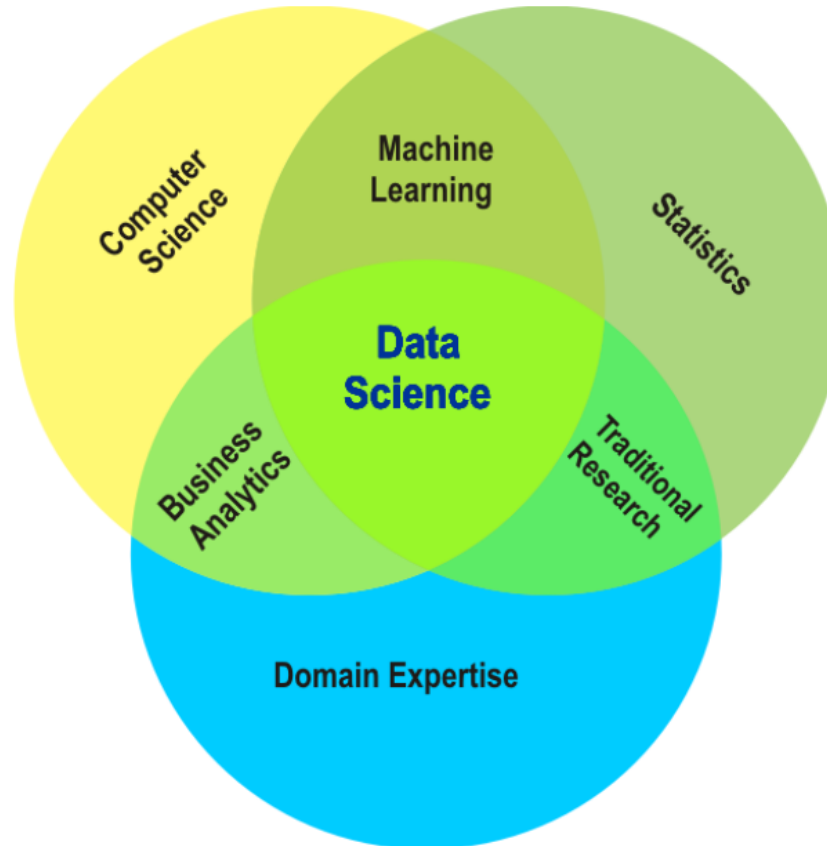
## 1.2 빅데이터 분석

### ■ 빅데이터 분석 프로세스



## 1.3 통계학과 기계학습

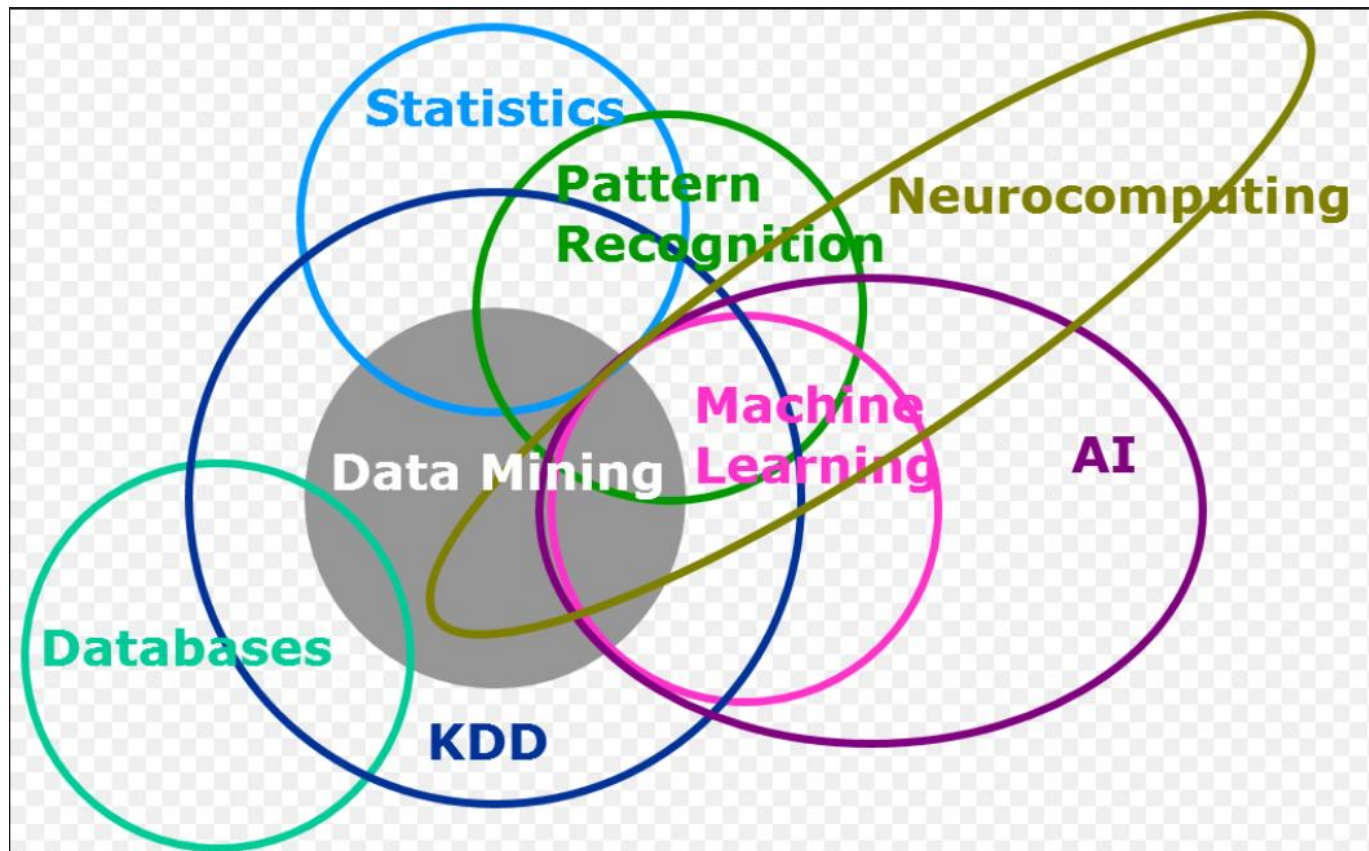
### ■ 데이터 사이언스



<https://marsiantech.com/data-science-analytics-training-course-pune.php>

## 1.3 통계학과 기계학습

### ■ 통계학과 기계학습



IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics (Hoyt et al., 2017)

## 1.3 통계학과 기계학습

### ■ 통계학과 기계학습

	<i>MACHINE LEARNERS</i>	<i>STATISTICIANS</i>
<i>Network/Graphs vs. Models</i>	<i>Network/Graphs to train and test data</i>	<i>Models to create predictive power</i>
<i>Weights vs. Parameters</i>	<i>Weights used to maximize accuracy scoring and hand tuning</i>	<i>Parameters used to interpret real-world phenomena - stress on magnitude</i>
<i>Confidence Interval</i>	<i>There is no notion of uncertainty</i>	<i>Capturing the variability and uncertainty of parameters</i>
<i>Assumptions</i>	<i>No prior assumption (we learn from the data)</i>	<i>Explicit a-priori assumptions</i>
<i>Distribution</i>	<i>Unknown a priori</i>	<i>A-priori well-defined distribution</i>
<i>Fit</i>	<i>Best fit to learning models (generalization)</i>	<i>Fit to the distribution</i>

<https://blog.galvanize.com/why-a-mathematician-statistician-machine-learner-solve-the-same-problem-differently-2/>





## 1.3 통계학과 기계학습

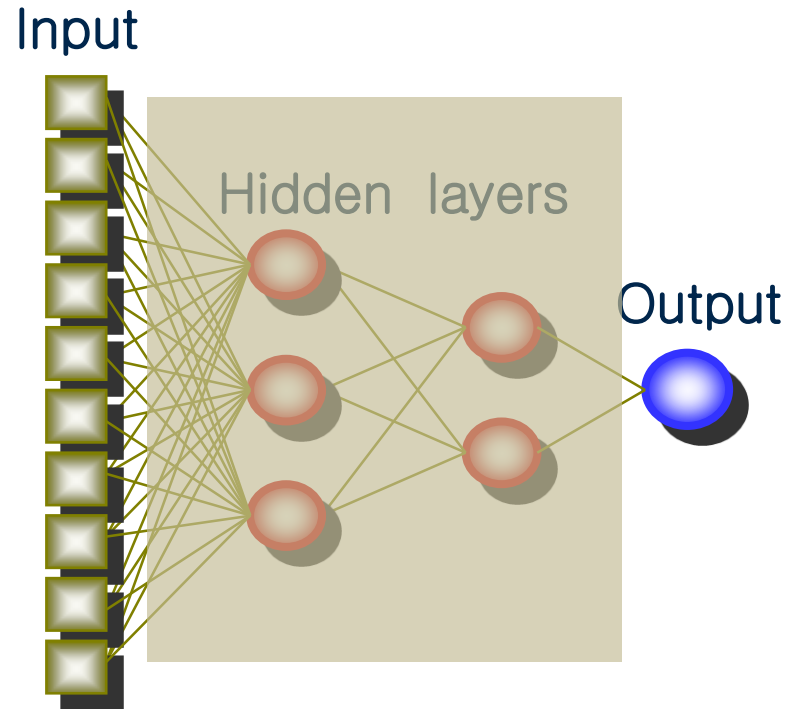
### ■ 통계학과 기계학습

Statistics	Machine Learning
Model	Network, Graphs
Parameter	Weight
Fitting	Learning
Test set performance	Generalization
Regression Classification	Supervised learning
Density estimation Clustering	Unsupervised learning
통계학은 해석가능성에, 기계학습은 예측정확도에 더 많은 관심을 둠	

## 1.3 통계학과 기계학습

### ■ 예측 vs 설명

- 기계학습: 주로 예측 목적, 블랙박스 모형



- 설명 없는 예측이 과학(science)인가?

Shmueli (2010). "To explain or to predict", *Statistical Science* 25(3), 289-310.



## 1.3 통계학과 기계학습

---

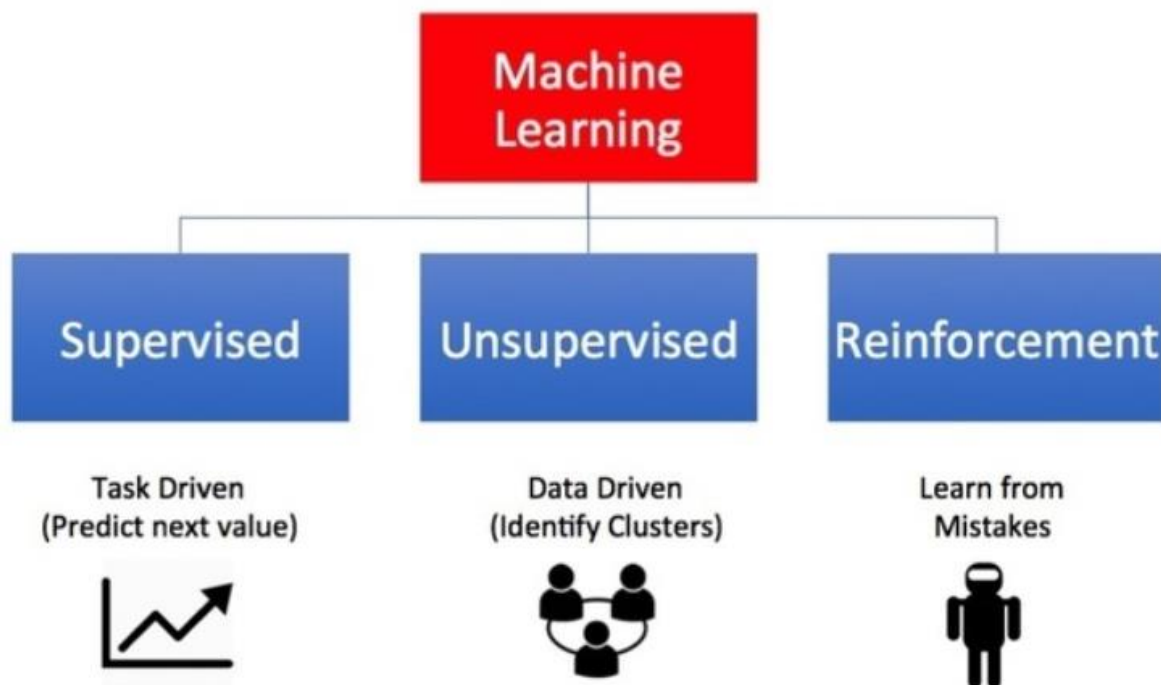
### ■ 데이터분석과 기계학습을 위한 시각

- 빅데이터 → Exploratory Data Analysis, Machine Learning
  - Confirmatory Studies with Small Data
  - 지식화
  - 빅데이터

## 1.3 통계학과 기계학습

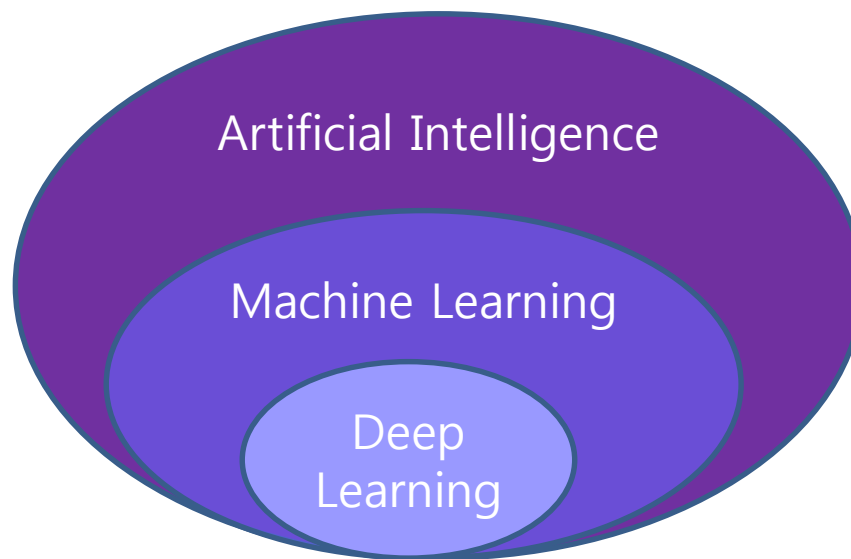
### ■ 기계학습 분류

#### Types of Machine Learning



## 1.3 통계학과 기계학습

### ■ Machine Learning, Deep Learning, AI

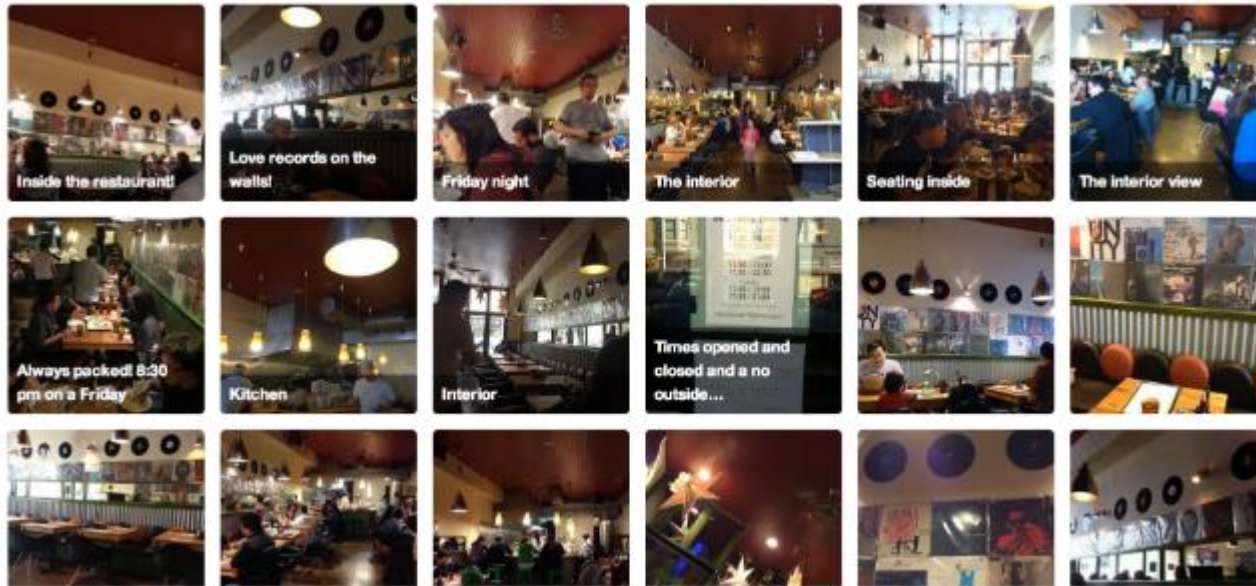


## 1.3 통계학과 기계학습

### ■ Companies using Machine Learning in cool ways

(<https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications>)

#### 1. Yelp – Image Curation at Scale



## 1.3 통계학과 기계학습

### ■ Companies using Machine Learning in cool ways

#### 2. Google – Neural Networks and ‘Machines That Dream’





## 1.3 통계학과 기계학습

### ■ Companies using Machine Learning in cool ways

#### 3. Edgcase – Improving Ecommerce Conversion Rates

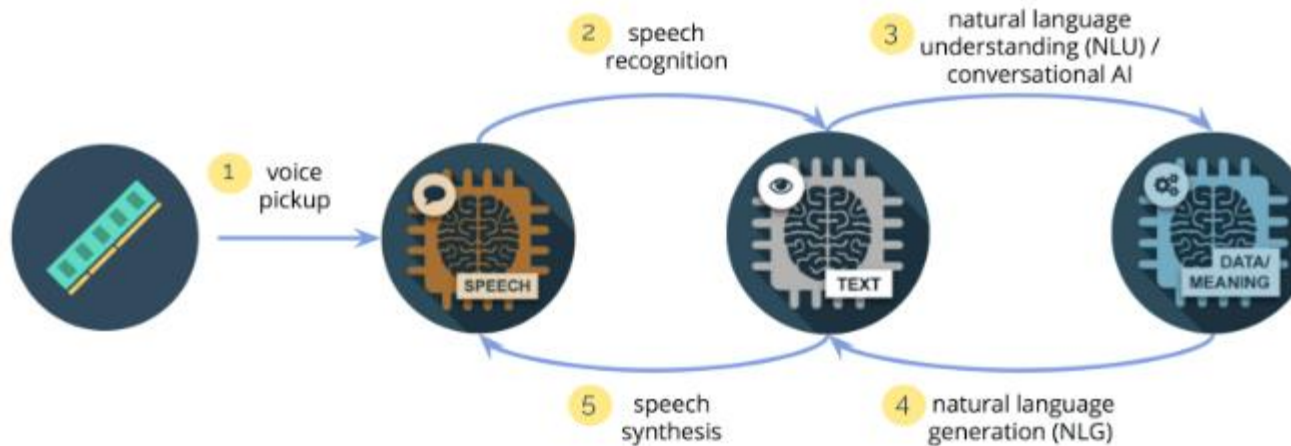




## 1.3 통계학과 기계학습

### ■ Companies using Machine Learning in cool ways

#### 4. Baidu – The Future of Voice Search



## 1.3 통계학과 기계학습

### ■ Companies using Machine Learning in cool ways

#### 5. Pinterest – Improved Content Discovery





## 1.4 데이터 분석 툴

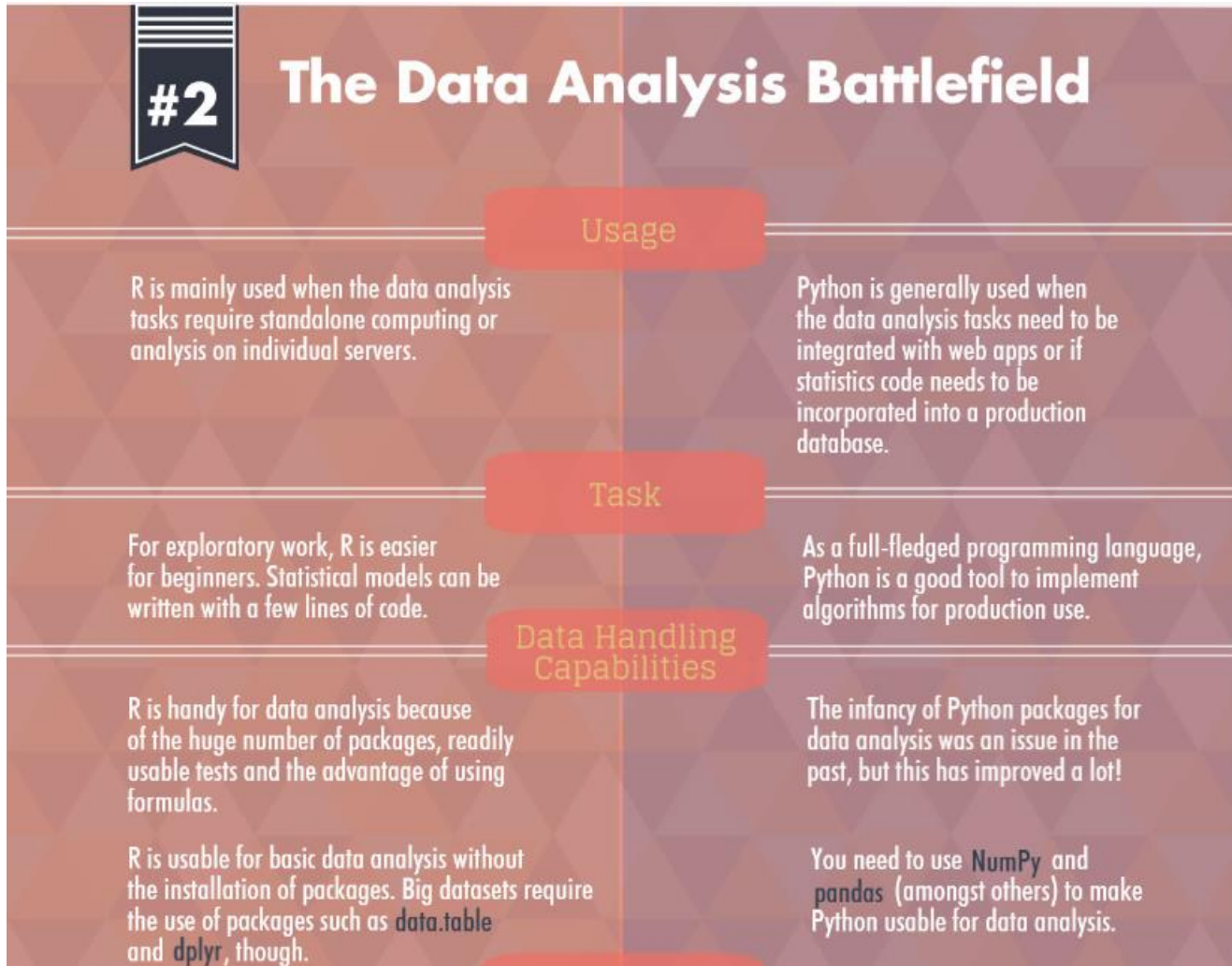
### ■ 데이터 분석 도구

프로그램	주사용자	특징
SAS	통계 전공자	통계분석의 막강한 기능 가격이 비쌈, 코딩 기반, 배우기 어려움
SPSS	사회과학 전공자	메뉴 방식의 분석, 배우기 쉬움, 유료
R	통계 전공자 등	무료 소프트웨어, 다양한 통계 분석 기능 , 그래픽 기능, 패키지를 통한 확장
Python	프로그래머 등	무료 소프트웨어, 빅데이터분석도구로 각광
Spotfire	데이터 분석가 등	데이터 시각화에 강함, 유료

## 1.4 데이터 분석 툴

### ■ R vs Python

(<https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>)



## 1.4 데이터 분석 툴

### ■ R vs Python

(<https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis>)

**#4 And The Winner is...**

It's a tie!  
It's up to you, the data scientist,  
to pick the language that best fits your needs.  
The following questions can guide you in your decision.

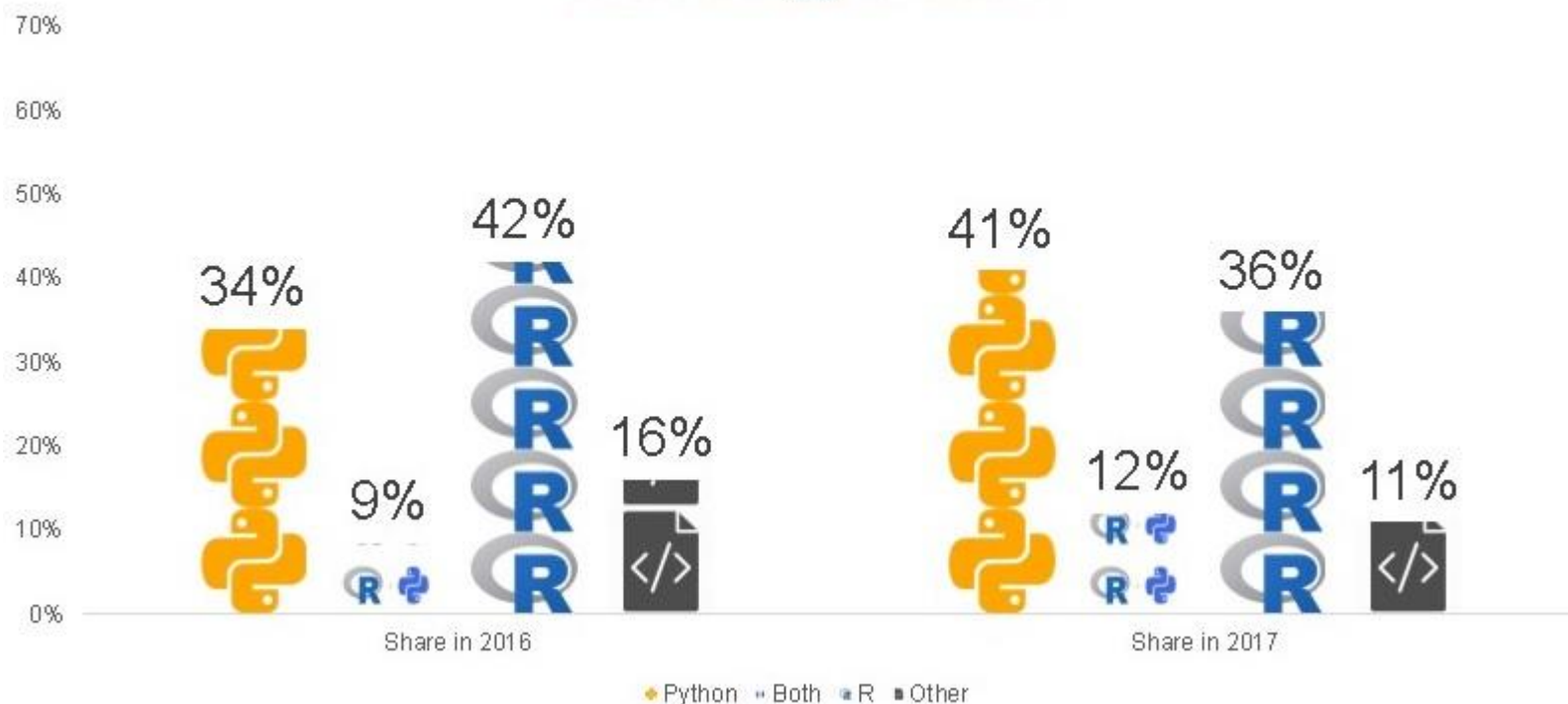
- 1 What problems do you want to solve?
- 2 What are the net costs for learning a language?\*
- 3 What are the commonly used tool(s) in your field?
- 4 What are the other available tools in your field and how do these relate to the commonly used tool(s)?

\* it will cost time to learn a new system that is better aligned for the problem you want to solve, but staying with the system you know may not be made for that kind of problem.

## 1.4 데이터 분석 툴

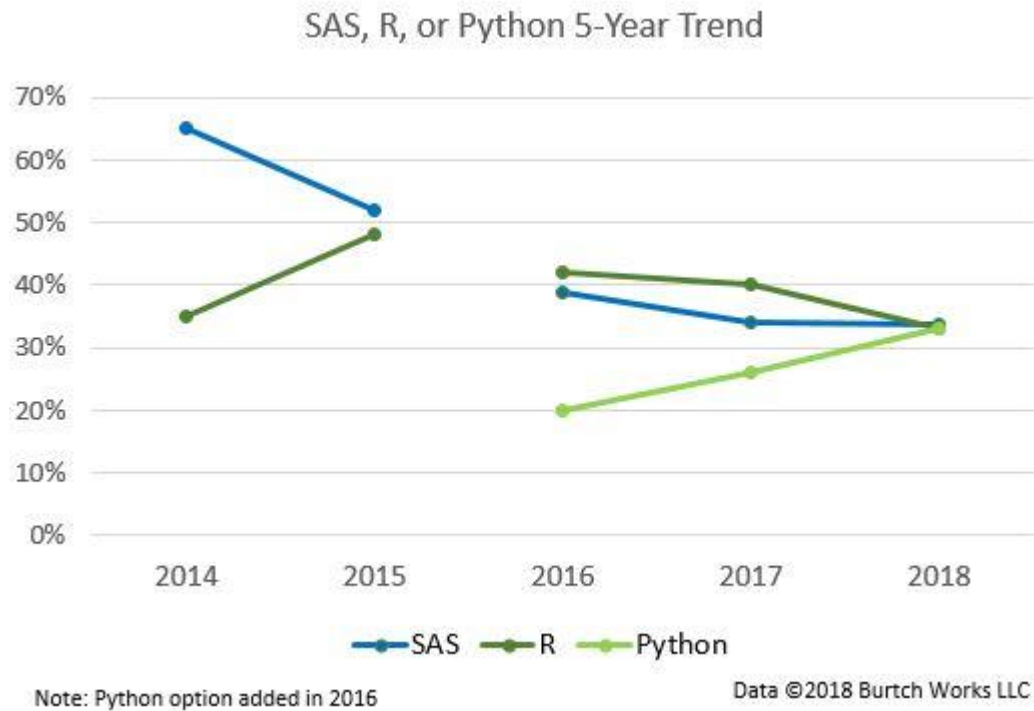
### ■ R vs Python

Python, R, Both or Other Platforms for Data Science  
Source: [KDnuggets Poll 2017](#)



## 1.4 데이터 분석 툴

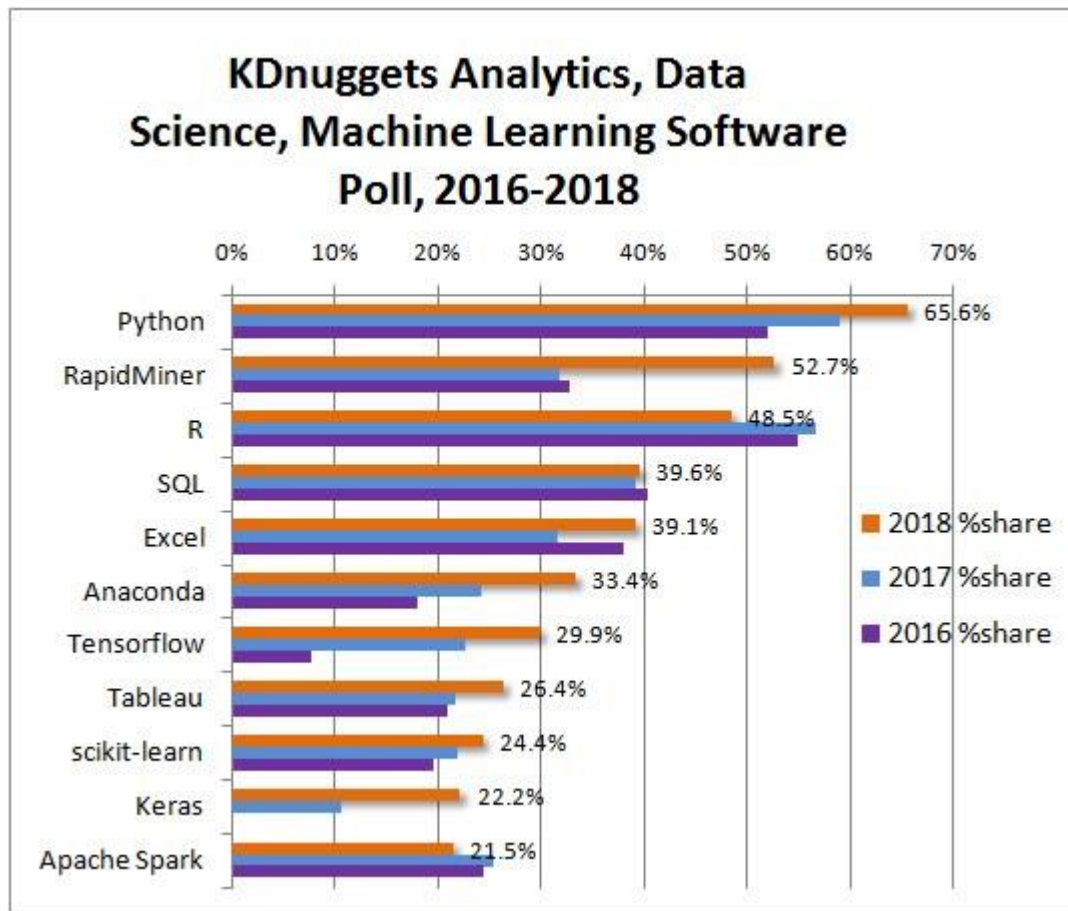
### ■ R vs Python





## 1.4 데이터 분석 툴

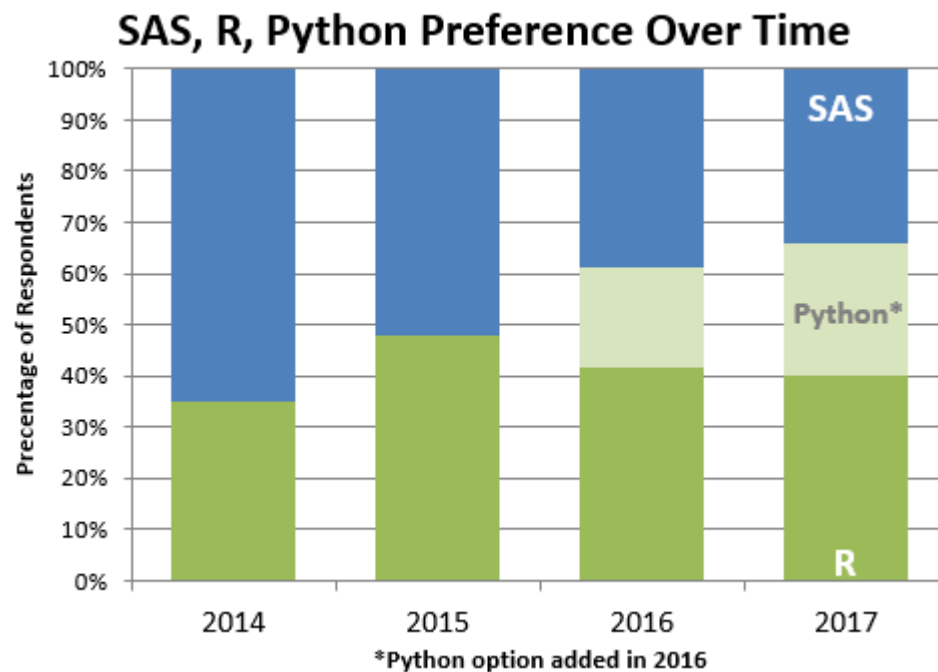
### ■ R vs Python





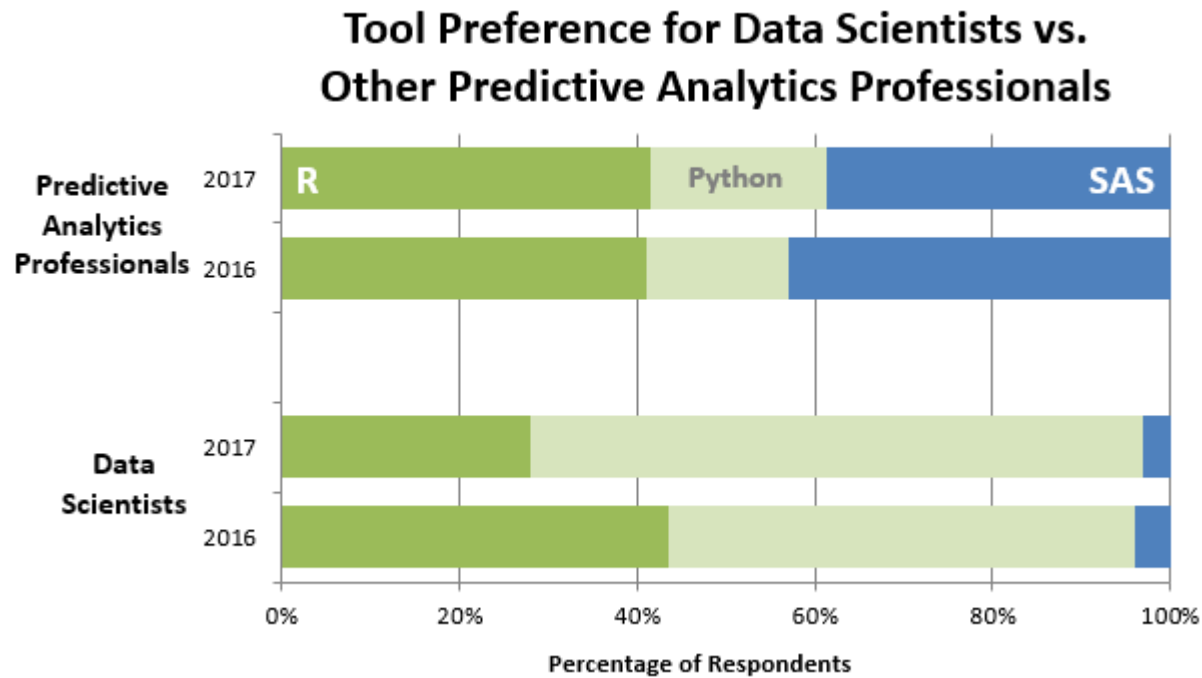
## 1.4 데이터 분석 툴

### ■ R vs Python




## 1.4 데이터 분석 툴

### ■ R vs Python



- traditional predictive analytics professionals working with structured data
- data scientists working with unstructured or streaming data.



---

The key word in "Data  
Science" is **not Data**, it is  
**Science**

– Jeff Leek, Simply Statistics(2013)