

I. 금융 빅데이터와 데이터 분석 (Big data in finance & financial time series data analysis)

① 빅데이터 기본 개념

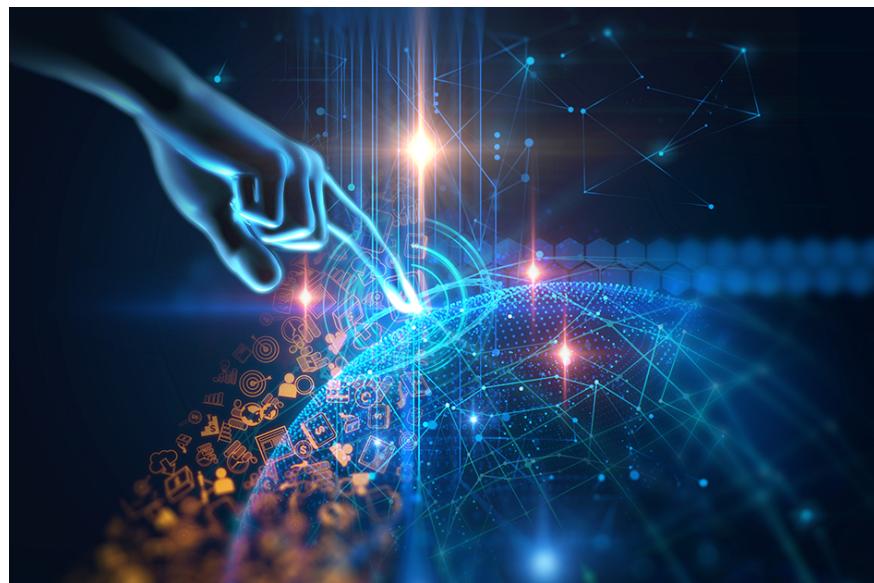
금융에서의 빅 데이터(Big Data)

- 빅 데이터란 기존 데이터베이스 관리도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술을 의미한다.
- **금융 빅 데이터**는 금융 분야의 데이터로 주로 거래 시장 데이터와 공공정보, 뉴스정보, 공시정보, 그리고 소셜 데이터 등 금융에 관련된 데이터를 포함한다. 거래 시장 데이터는 정형화된 데이터이며, 공공정보 등은 비정형화된 데이터이다.



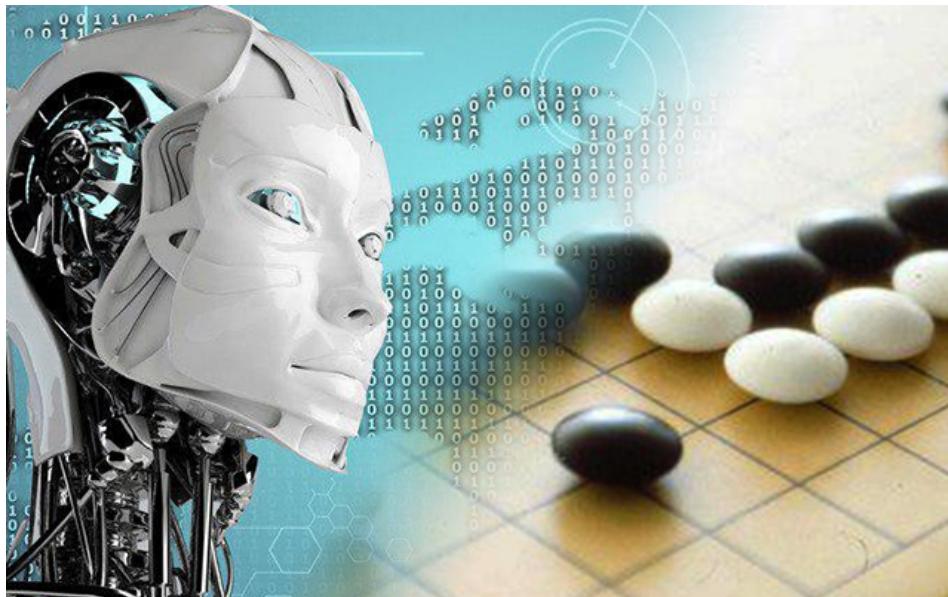
인공지능(AI; Artificial Intelligence)

- 인공지능은 기계로부터 만들어진 지능을 말한다. 컴퓨터 공학에서 이상적인 지능을 갖춘 존재, 혹은 시스템에 의해 만들어진 지능, 즉 인공적인 지능을 뜻한다.
- **인공지능의 궁극적인 목표** : 인간과 같은 지능의 개발
- 1940년대 후반과 1950년대 초반에 이르러서 수학, 철학, 공학, 경제 등 다양한 영역의 과학자들에게서 인공적인 두뇌의 가능성이 논의되었고, 1956년에 이르러서, 인공지능이 학문 분야로 들어섰다.



알파고 (AlphaGo)

- 알파고(AlphaGo)는 구글(Google)의 딥마인드(DeepMind Technologies Limited)가 개발한 인공지능(AI, Artificial Intelligence) 바둑 프로그램이다.
- 영국의 스타트업 기업이었던 딥마인드가 2014년 구글에 인수되면서 개발이 본격적으로 진행되었다. 2015~2017년 프로토타입 버전인 알파고 판, 알파고 리, 알파고 마스터가 공개되었고, 2017년 10월에 최종 버전인 알파고 제로를 발표하였다. 2018년 12월에는 바둑을 포함한 보드게임에 적용할 수 있는 범용 인공지능 알파 제로(Alpha Zero)를 발표하였다.



알파고는 어떻게 인간을 이겼을까?

- 알파고는 포석단계에서부터 직관과 감각을 계량화하여 각 경우의 수의 가치와 우선순위를 계산
- 불과 몇 년 전까지만 해도 바둑은 체스와 달리 경우의 수가 너무 많아 인간이 기초 데이터를 입력하는데 한계가 있으므로 컴퓨터가 인간을 이길 수 없을 것이라고 믿었다.
- 체스는 10의 120승, 장기는 10의 220승, 바둑은 10의 360승의 경우의 수를 가진다.
- 하지만 사람이 데이터를 입력하지 않아도 컴퓨터가 스스로 학습해서 깨우치는 머신러닝/딥러닝 기술이 개발되면서 드디어 기계가 바둑에서 인간을 능가하게 됩니다.
- 인간의 뇌는 무수한 정보 가운데 자신의 판단에 필요한 내용을 순간적으로 선택하는 능력과 정보를 무의식적으로 저장하는 능력은 기계 대비 우월합니다. 인간의 뇌는 출생 당시 미성숙한 상태로 시작하여 이후 성장과정에서 자극을 받고 정보를 취사선택하고 조합하는 "자기조직 원리"에 의해 완성된다.
- 이때 판단과 선택 그리고 조합은 매우 빨리 작동한다. 이것은 수백만 년 이상 의식적 또는 무의식적으로 인간 DNA에 축적된 유전자 정보 덕분이다.

알파고는 어떻게 인간을 이겼을까?

- 21세기에 접어들어 뇌의 특정 부분과 신경세포들이 어떻게 정보를 처리하는지 밝혀지면서 뇌 작동 방법을 모방한 컴퓨터 알고리즘이 만들어지기 시작하였다.
- 알파고는 **딥러닝 신경망과 몬테카를로 트리 검색을 결합하여 스스로 학습**해 작동 원리를 깨우치며 이를 통해 각각의 바둑 수에 대한 확률 계산
- 거기에 더해 알파고는 전문가 지도학습과 자체 경기를 통한 강화학습으로 훈련해 왔고,
- 프로 바둑기사들의 기보 16만개를 입력해 그들의 수법을 모방하였으며 스스로 학습해 원리를 깨우치는 머신러닝 기법이 더해진 것이다.
- 알파고가 프로 바둑기사의 기보 학습에 더해 자체 경기를 통해 강화학습을 한 이유는 인공지능이 기존 학습 데이터에 너무 고정되어 새로운 상황에 직면하면 잘못된 결정을 내리는 오류를 범했기 때문이다.
- 또한 프로 바둑기사의 기보에 의한 지도학습 결과가 최적의 선택이라고 판단하기 어렵기 때문이다.
- 그래서 이러한 문제를 보완하기 위해 자체 경기를 통해 얻은 3천만 개의 기보를 통해 스스로 바둑의 원리를 깨우치게 되었다.

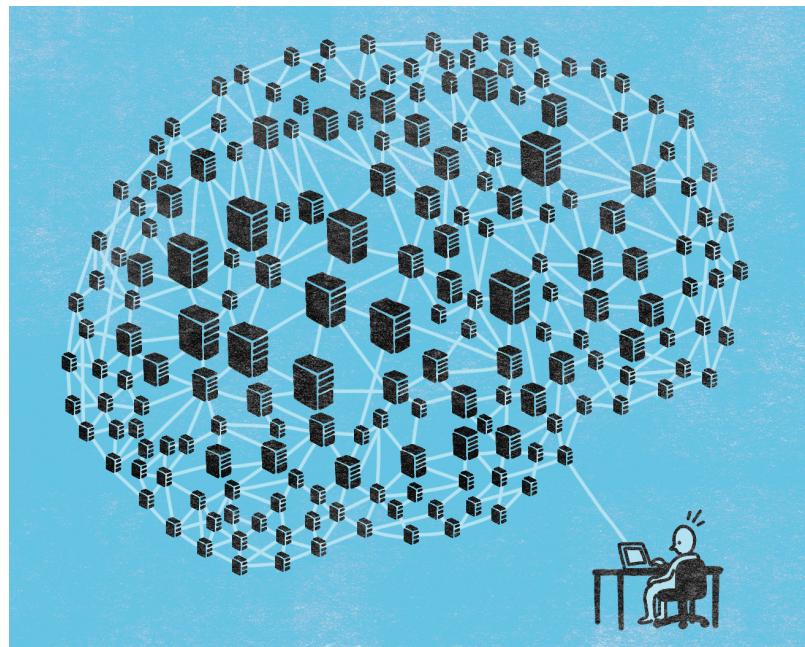
알파고는 어떻게 인간을 이겼을까?

- 바둑에서는 경우의 수가 워낙 방대해서 추가적인 알고리즘을 통하여 경우의 수를 대폭 축소시켜 계산 시간을 줄여야 했으며 그 과정에 데이터의 양도 엄청나기 때문에 이를 빨리 저장하고 처리할 수 있는 기술이 동반되어야 한다.
- 인공지능을 위한 엄청난 양의 데이터를 저장하고 처리할 수 있는 기술은 **클라우드 컴퓨팅과 빅데이터**를 통해 가능해 졌고, 빨리 처리할 수 있는 컴퓨터를 위해서는 **고성능의 CPU와 GPU**가 도입되었다.
- 알파고는 인간의 신경망 구조를 모방한 딥러닝 알고리즘과 여기에 엄청난 양의 정보를 빠르게 처리하는 빅데이터 기술이 결합되면서 가능해진 것이다.
- 딥러닝은 목표 내용을 주입하는 것이 아니라 무수한 데이터를 걸러내는 과정에서 그 내용을 컴퓨터가 알아서 찾아내도록 설계하는 것이다.
- 따라서 방대한 경우의 수를 탐색하여 가장 주도적인 관련성을 찾아내는 작업이 딥러닝의 핵심이라 할 수 있다.
- 토론토 대학의 제프리 힌튼 교수는 슈퍼컴퓨터를 기반으로 딥러닝 개념을 증명하는 알고리즘을 **병렬화**하는데 성공하였으며 **병렬 연산에 최적화된 GPU의 등장**은 신경망의 연산 속도를 획기적으로 높여 딥러닝 기반 인공지능의 등장을 가속화하였다.

머신러닝(Machine Learning)과 딥러닝(Machine Learning)

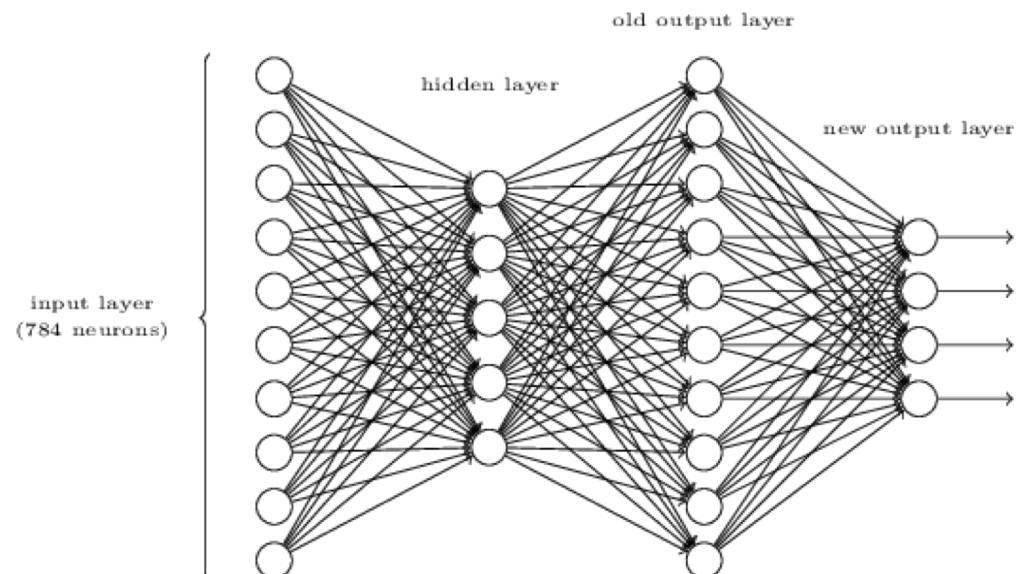
- 머신러닝

- "명시적인 프로그래밍 없이 컴퓨터가 스스로 학습할 수 있는 방법에 대한 학문"
- Arthur Samuel (1959)



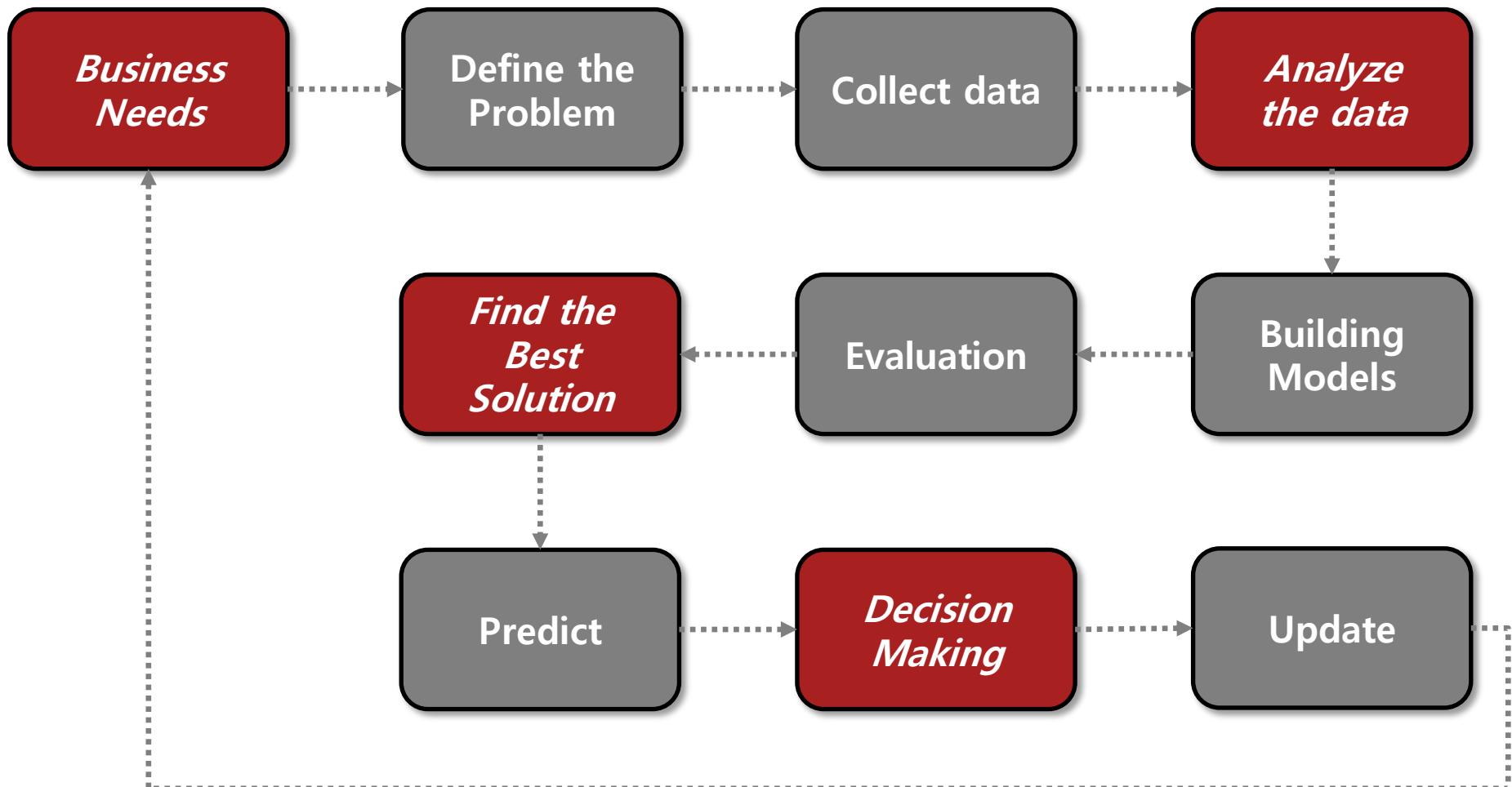
- 딥러닝

- 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화를 시도하는 기계학습(machine learning) 알고리즘의 집합



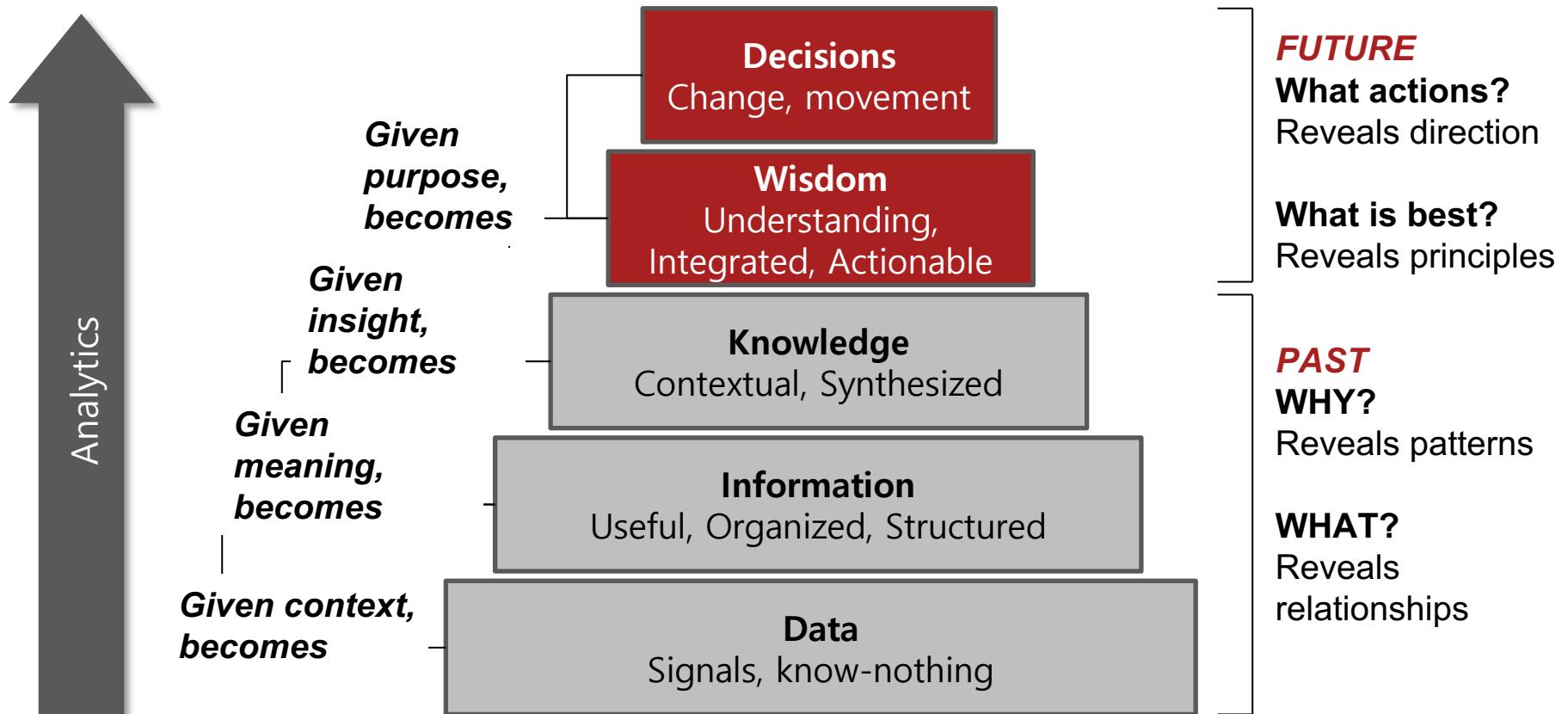
비즈니스 애널리틱스(BA; Business Analytics)란?

- BA는 복잡한 경영환경에서 직면한 문제를 해결하기 위해 “현재 우리의 문제가 무엇인지”를 정의하고, 이에 맞는 데이터 수집 및 최적의 Solution을 찾아 의사결정을 내리는 일련의 과정이다.



비즈니스 애널리틱스(BA; Business Analytics)

- 경영환경에서 무수히 많은 데이터로부터 얻은 정보를 기반해 불확실한 미래에 대한 예측 및 의사결정



BA의 넓은 의미

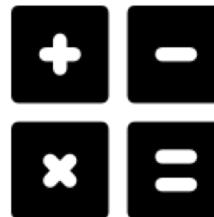
Sourcing & Data Preparation



“ 어떤 자료를
쓸 것인가? ”

- ✓ Data 발생 주체
- ✓ DB 확인
- ✓ Data 수집
- ✓ Data 전처리

Modeling & Analysis



“ 당면 문제를 어떻게
풀어낼 것인가? ”

- ✓ 분석목표 선정
- ✓ 가용 Data 확인
- ✓ 분석 알고리즘 선정
- ✓ 최적 모델 도출

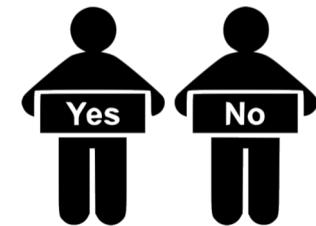
Create report & Dashboard



“ 분석결과를 어떻게
표현할 것인가? ”

- ✓ 결과 Visualization
- ✓ Report 작성
ex) PPT, Markdown
- ✓ Dashboard 구성

Supporting Decision Making



“ 어떤 의사결정을
지지할 것인가? ”

- ✓ Best Decision
도출
- ✓ 예상결과 제시

BA의 좁은 의미

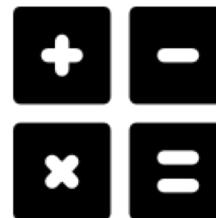
Sourcing & Data Preparation



“ 어떤 자료를
쓸 것인가? ”

- ✓ Data 발생 주체
- ✓ DB 확인
- ✓ Data 수집
- ✓ Data 전처리

Modeling & Analysis



“ 당면 문제를 어떻게
풀어낼 것인가? ”

- ✓ 분석목표 선정
- ✓ 가용 Data 확인
- ✓ 분석 알고리즘 선정
- ✓ 최적 모델 도출

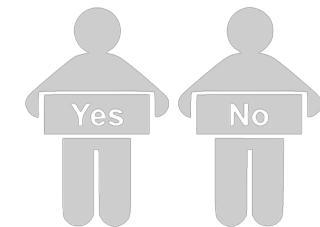
Create report & Dashboard



“ 분석결과를 어떻게
표현할 것인가? ”

- ✓ 결과 Visualization
- ✓ Report 작성
ex) PPT, Markdown
- ✓ Dashboard 구성

Supporting Decision Making



“ 어떤 의사결정을
지지할 것인가? ”

- ✓ Best Decision
도출
- ✓ 예상결과 제시

애널리틱스(Aalytics)를 위한 기초통계

1. 자료의 분포 특성(Properties) : 중심경향 정도와 산포 정도
2. 모수(Parameter)와 통계량(Statistic)
3. 기술 통계량(Descriptive Analytics)

중심경향 : 평균(Mean), 중앙값(Median), 최빈값(Mode)

평균 (Mean)	<ul style="list-style-type: none">일반적으로 평균이라고 하면 산술평균을 의미함. 평균은 지나치게 크거나 작은 값에 영향을 받을 수 있음 $\text{모평균} : \mu = \frac{x_1 + \dots + x_N}{N} = \frac{1}{N} \sum_{i=1}^N x_i, \text{ 표본평균} : \bar{x} = \frac{x_1 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$
중앙값 (Median)	<ul style="list-style-type: none">주어진 값들을 크기의 순서대로 정렬했을 때 가장 중앙에 위치하는 값을 의미지나치게 크거나 작은 값에 영향을 받지 않아 이상치를 판단할 수 있음
최빈값 (Mode)	<ul style="list-style-type: none">가장 빈도가 높은 자료
1분위수 (1 st quantile)	<ul style="list-style-type: none">측정값이 가장 낮은 순에서 높은 순으로 4등분 했을 때, 아래에서부터 25%에 해당하는 값
3분위수 (3 rd quantile)	<ul style="list-style-type: none">측정값이 가장 낮은 순에서 높은 순으로 4등분 했을 때, 아래에서부터 75%에 해당하는 값

산포 정도 : 분산(variance) 및 표준편차(Standard deviation)

분산 (Variance)

- 자료가 평균(Mean) 혹은 기대값(Expectation)으로부터 얼마나 떨어진 곳에 분포하고 있는지를 가늠하는 지표로 자료가 퍼져 있는 정도를 나타냄
- 일반적으로 예측모형의 경우, 예측값의 분산을 최소로 하는 모형 즉, 예측값의 변동이 적은 모형을 효율적인(Efficient) 모형이라고 평가함. 하지만, 단순히 분산이 작다고 예측값의 성능이 좋은 것은 아님

$$\text{모분산} : \sigma^2 = \frac{(x_1 - \mu)^2 + \dots + (x_N - \mu)^2}{N} = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2,$$

$$\text{표본분산} : s^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

표준편차 (Standard Deviation)

- 표준편차 역시 자료의 산포도를 나타내며 분산의 양의 제곱근으로 정의됨

$$\text{모표준편차} : \sigma = \sqrt{\frac{(x_1 - \mu)^2 + \dots + (x_N - \mu)^2}{N}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2},$$

$$\text{표본표준편차(표준오차)} : s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

One-slide summary

모수(Parameter); 모집단의 특성/속성

모평균
(Population Mean) 모분산
(Population Variance) 모표준편차
(Population Standard Deviation)

$$\mu$$

$$\sigma^2$$

모표준편차

$$\sigma$$

모집단 모형의 계수

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

모집단 계수(모수)는 일반적으로
그리스 문자로 표현!

모집단 (Population)

Ex) 한양대 전체 학생의
토익성적
전국 학생의 평균 수면량
공부량과 학점의 관계
⋮

표본추출
(Sampling)

표본추출
(Sampling)

통계량(Statistic); 표본집단의 특성/속성

표본평균
(Sample Mean)

$$\bar{X}$$

표본분산
(Sample variance)

$$S^2$$

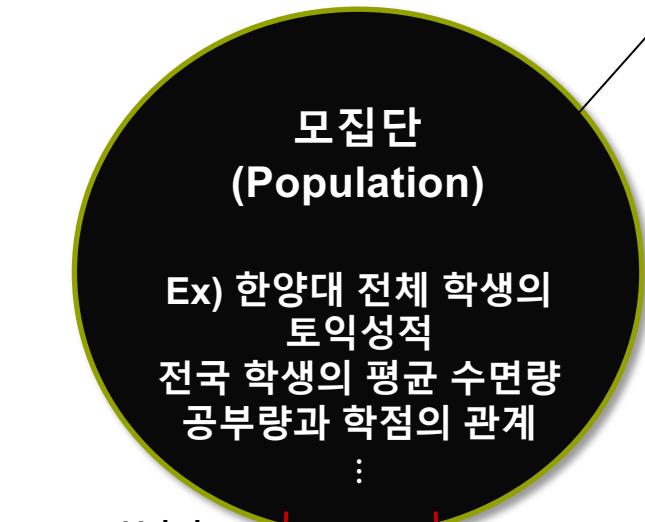
표본표준편차
(Sample standard deviation)

$$S$$

표본집단 모형의 계수

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + e$$

표본의 계수는 일반적으로
그리스 문자에 햇(hat; $\hat{}$) 표시를 해서 표현!



표본집단1
(Sample)

...

표본집단2
(Sample)

모평균
(Population Mean) 모분산
(Population Variance) 모표준편차
(Population Standard Deviation)

$$\mu$$

$$\sigma^2$$

모표준편차

$$\sigma$$

모집단 모형의 계수

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

모집단 계수(모수)는 일반적으로
그리스 문자로 표현!

추정(Estimation) 추정(Estimation) 추정(Estimation)

가설(Hypothesis)

검정(Testing)