수학,확률, 통계 그리고 AI와 ML (Wanna be Data Scientist?)

Kim Byung Chum
KAIST
bckim@kaist.ac.kr

미래 어떤 전문가가 되고 싶습니까?

전산전문가

CEO

- 일단 전산 전문가가 된후, 다음에 회사를 세워 CEO가 되는 것

어떤 전산 전문가?

막연함(AI, ML ???)

Google이나 Amazon 취업(어떤 분야에서 일할 수 있을 까?)

아니면, 삼성, LG, 네이버, 다음에 취업

그들은 나의 어떤 면을 보고 뽑아 줄까?

- 학교 성적, 프로그래밍 경력, 코딩실력

나의 Background는?

- 전산학과 또는 전산 관련 학과
- 수학과
- 통계학과
- 경영 또는 경제학과(숙제 정도 패키지 사용)
- 기타(코딩 해본 적도 없다)

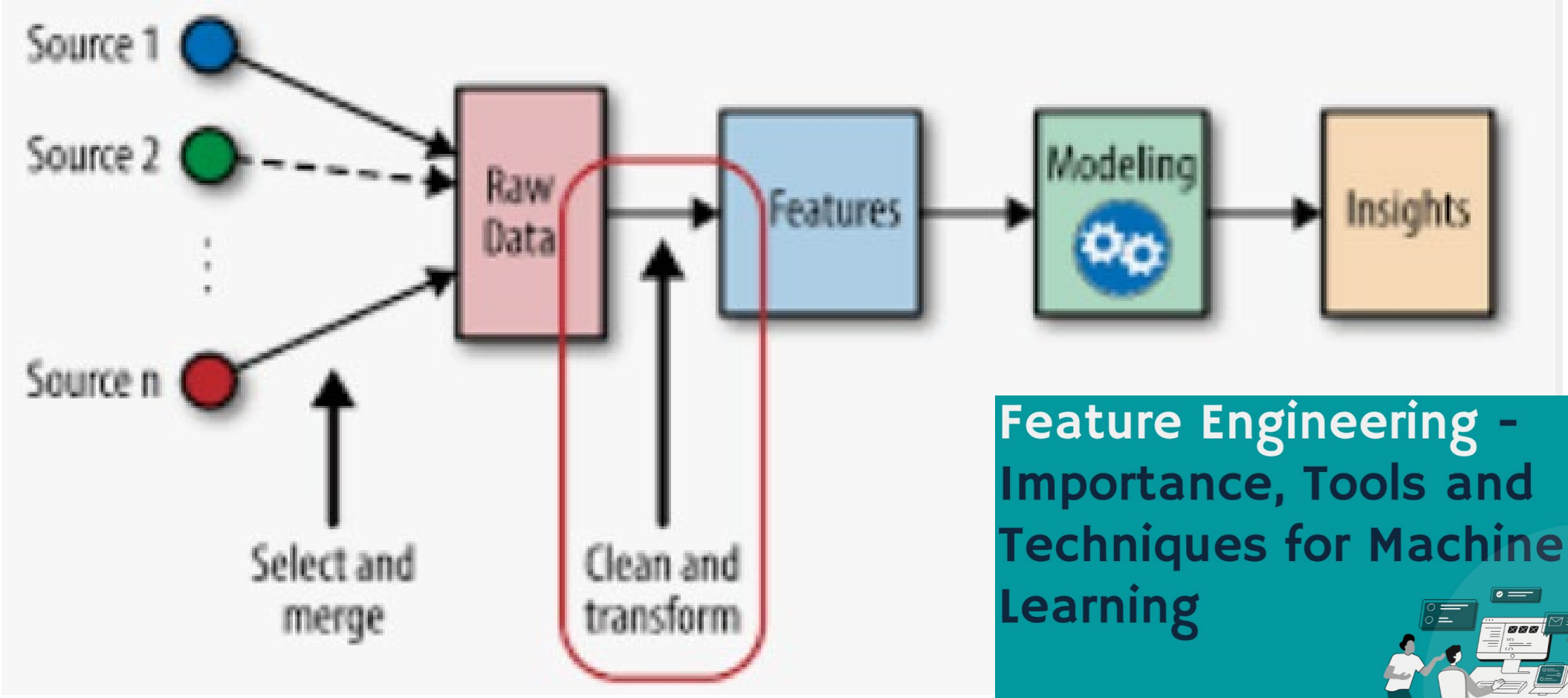
코딩 배우면 전문가가 될 가능성이 높다?

Study recommendation

- A good foundation in algebra(especially linear algebra), calculus, probability, and statistics
- Python or R as a programming language, and their corresponding libraries for Data Science.
- Knowledge of SQL to make queries about databases
- Obtaining data from different sources (API queries, web scrapping, …)
- Cleaning and preprocessing of data (and the famous *feature engineering*)
- Machine Learning (algorithms, modeling, evaluation, optimization, etc).
- Deep Learning, Reinforcement Learning, Natural Language Processing, Computer Vision, …

특징 공학 (Feature engineering)

도메인 지식을 이용해 원시 데이터를 가공하여 특징을 추출하는 것 .

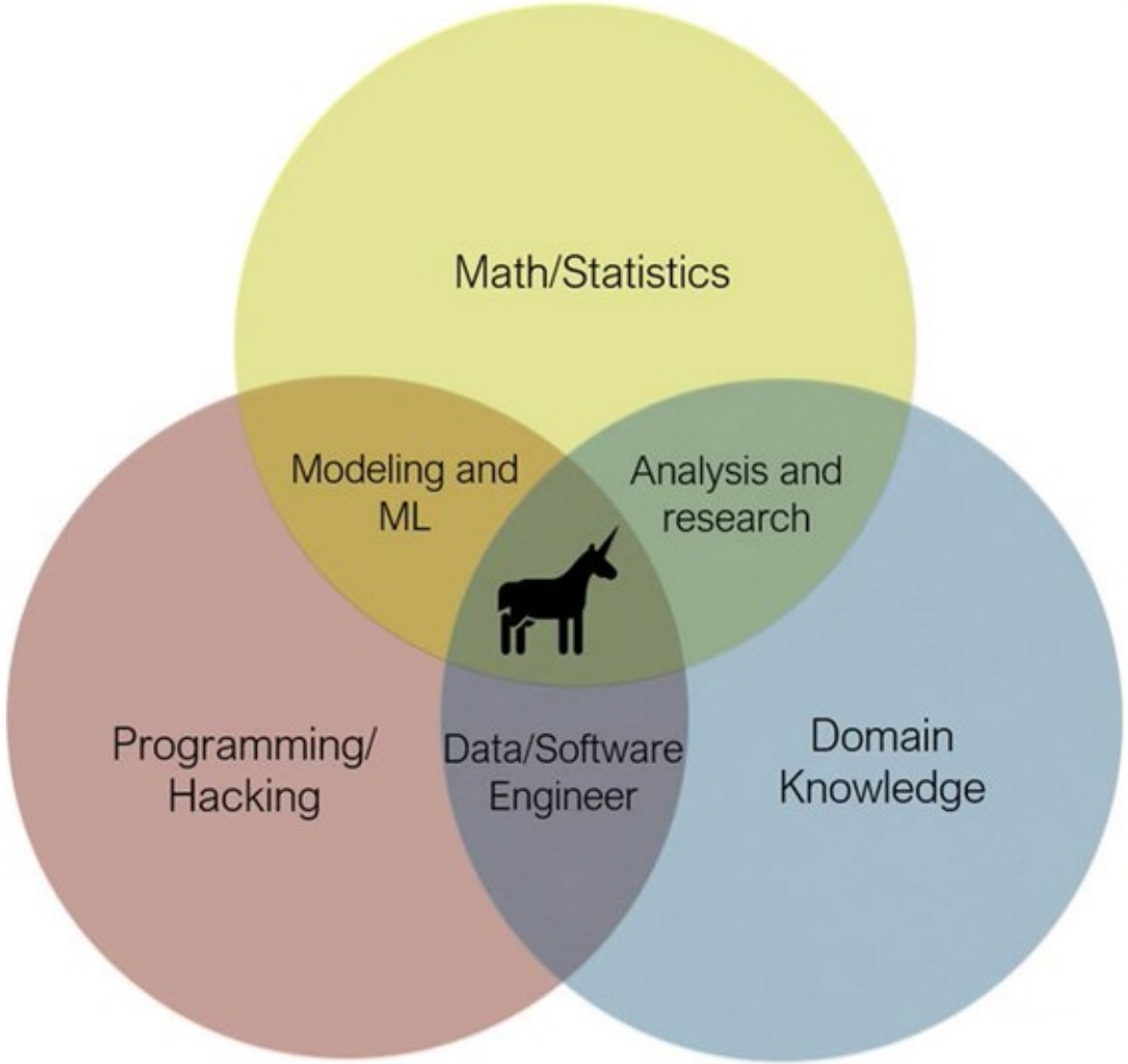


Big Trash VS Big Data

정돈 안된 내 사진들은 쓰레기 VS 방송국 또는 신문사의 사진들은 귀한 자료

쓰래기(trash, garbage,....)

필요한 지식은?


Math/Statistics
Modeling and ML
Analysis and research
Programming/ Hacking
Data/Software Engineer
Domain Knowledge

제일 기본인 학문은?

- 전문가가 되기 위해서는 지식의 바탕에는 어떤 학문?

수학, Mathematics
그리고 확률과 통계
이를 바탕으로
ML, AI

수(數):Numbers

- 숫자에 익숙하다고 생각하십니까?
- 숫자를 가지고 무엇을 할 수 있나요?
- 왜 숫자가 필요한가요?
- 숫자가 모든 세상을 설명할 수 있나요?
- 그럼 한계가 있을 까요?

- 경제학, 인공지능, 철학등을 비롯한 많은 학문들이 수학과 많은 관련이 있다고 한다

- 정말일까? 그럼 왜?
- But Math is Boring!?!?

수학은 지루하고 따분한 학문?

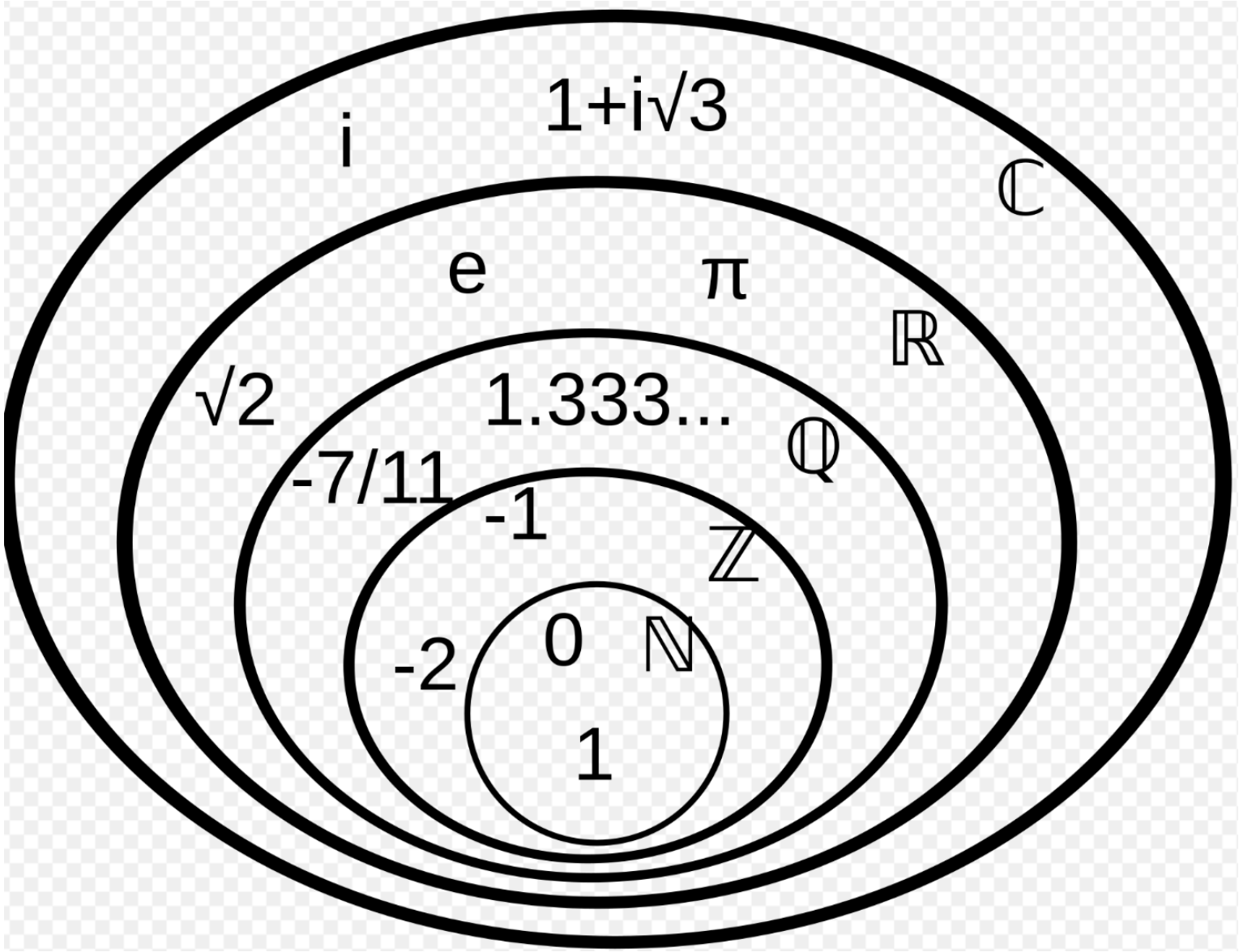
- 수학은 Boring할까? 왜?

- 수학은

  numbers, operations, laws

- 세상을 수학으로 표현 할 수 있을 까?

수의 분류



1 Natural numbers
2 Integers
3 Rational numbers
4 Real numbers
5 Complex numbers

수=현상, 상태

숫자(numbers)로 문장을?

- 숫자로 문장을 만들 수 있나?

  '명사+동사'가 마치 문장이 되듯이(잠+자다)

- 연산자(+, -, x, / 그외 ∀,∃ 등)를 이용하여 문장을 만드는 데, 결과는 true, false(Boolean).

예: $\forall y \exists x (x^2 = y)$

Numbers + Operators

Basic Operations

- The basic operations used in mathematics are

addition(+), **subtraction(-)**, **multiplication(*)**, and **division(/)**

- In addition to these operations, there are also inequalities (relational operators):

equals (=),

greater than (>),

less than (<),

greater than or equal to (≥),

L **ess than or equal to** (≤),

not equal (≠).

Law(법칙)

- Commutative Law of Addition: $a + b = b + a$
- Associative Law of Addition: $(a + b) + c = a + (b + c)$
- Commutative law of multiplication: $x \times y = y \times x$
- Associative law of multiplication: $a \times b \times c = (a \times b) \times c = a \times (b \times c)$
- Distributive law of multiplication:

$$(a + b) \times c = a \times c + b \times c$$

$$c \times (a + b) = c \times a + c \times b$$

예: $5 \times (6 - 2) = 5 \times 6 - 5 \times 2 = 30 - 10 = 20$

Field(체; 體, 体)

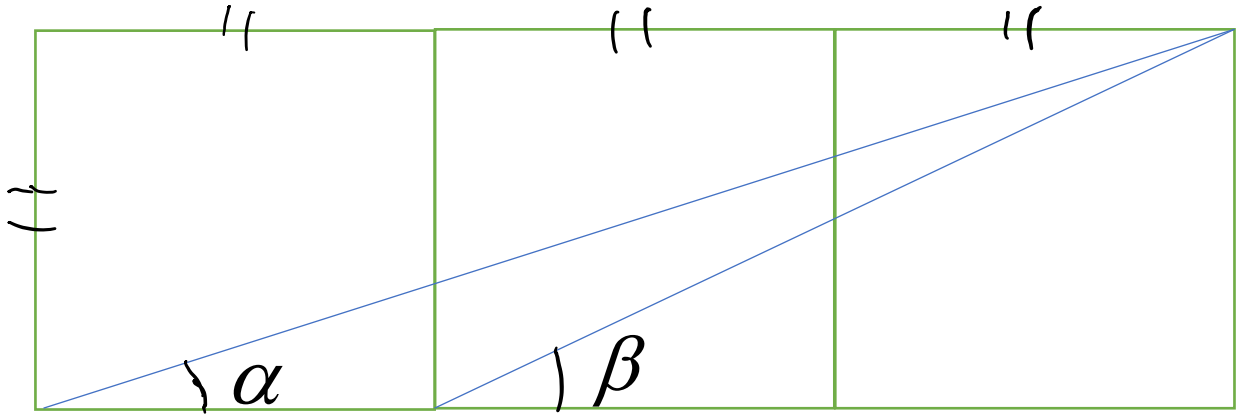
A field is a set F together with two binary operations of F called *addition* and *multiplication*. A binary operation on F is a mapping $F \times F \to F$. These operations are required to satisfy the following properties, referred to as *field axioms*(in these axioms, a, b, and c are arbitrary elements of the field F):

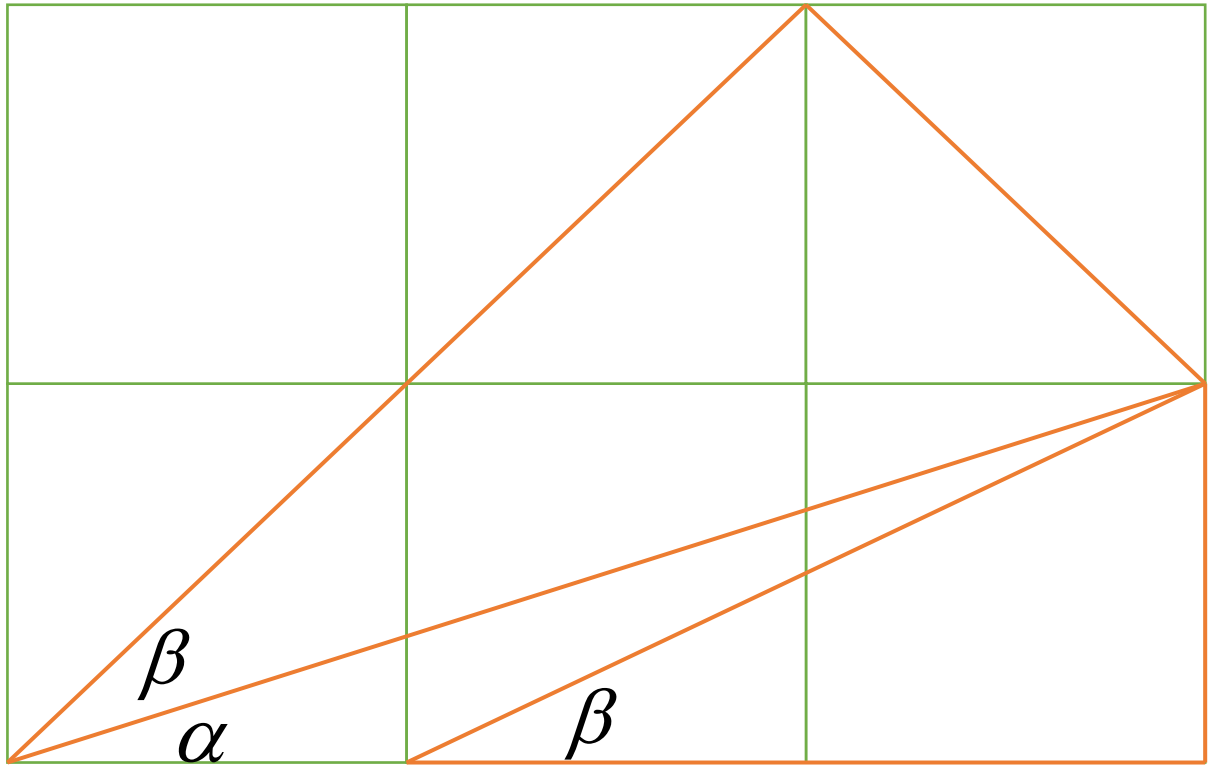
- Associativity of addition and multiplication: $a + (b + c) = (a + b) + c$, and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- Commutativity of addition and multiplication: $a + b = b + a$, and $a \cdot b = b \cdot a$.
- Additive and multiplicative identity: there exist two different elements 0 and 1 in F such that
  $a + 0 = a$ and $a \cdot 1 = a$.
- Additive inverses: for every a in F, there exists an element in F, denoted $-a$, called the *additive inverse* of a, such that $a + (-a) = 0$.
- Multiplicative inverses: for every $a \neq 0$ in F, there exists an element in F, denoted by a^{-1} or $1/a$, called the *multiplicative inverse* of a, such that $a \cdot a^{-1} = 1$.
- Distributivity of multiplication over addition: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

$$\alpha + \beta = ?$$

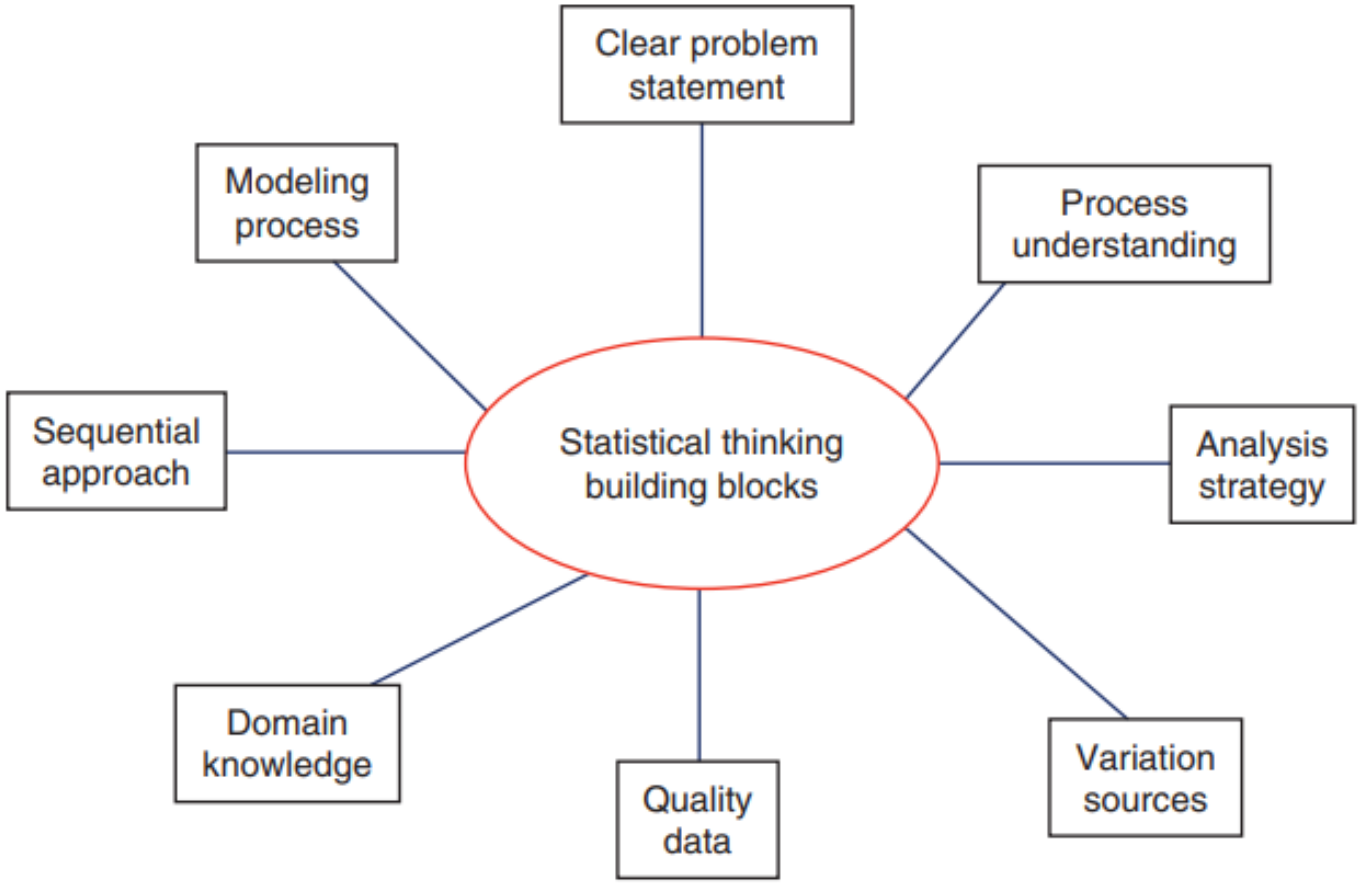
삼각함수를 이용하지
않고 푸는 방법은?

$$\tan(\alpha+\beta) = \frac{\tan\alpha + \tan\beta}{1 - \tan\alpha \cdot \tan\beta}$$




β
α
β

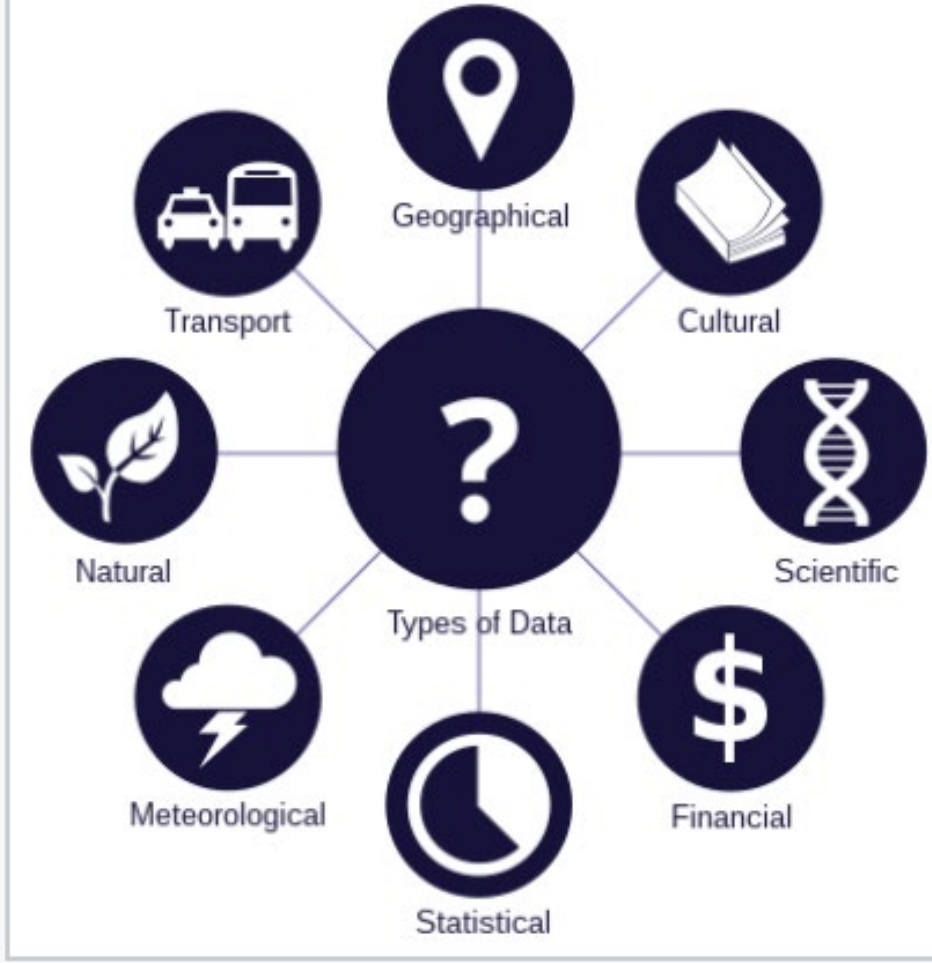
통계적 관점에서의 구성요소



우리가 갖고 있는 데이터는?

숫자, 문자, 그림(image)들

Set

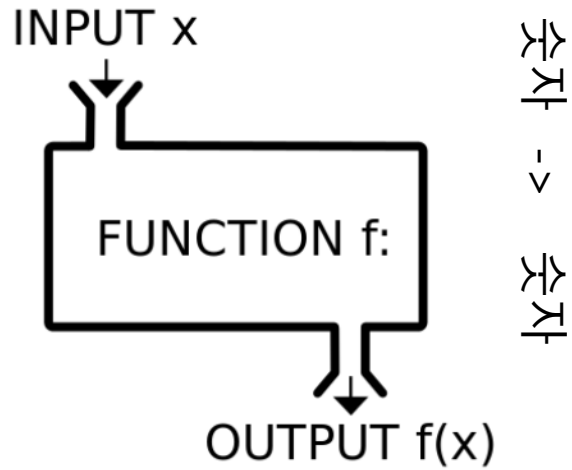


Some of the different types of data.

함수:Function

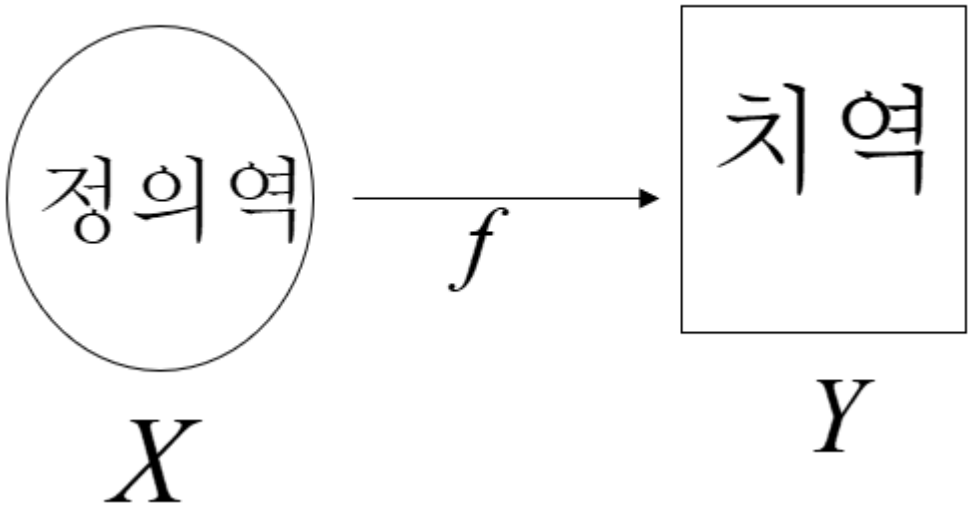
Function
$x \mapsto f(x)$

- A function described metaphorically as a "machine" or "black box" that for each input yields a corresponding output



문자 → 숫자로 가는 함수는 ?

확률 함수



정의역:: set(원소가 문자 또는 숫자)

치역: 실수—real numbers

Probability(확률) 공리(Axiom)

- Let (Ω, F, P) be a measure space with $P(E)$ being the probability of some event E. and $P(\Omega) = 1$

  1. the probability of an event is a non-negative real number;

$$P(E) \in \mathbb{R}, P(E) \geq 0 \qquad \forall E \in F$$ where F is the event space.

  2. $P(\Omega) = 1.$

  3. Any countable sequence of mutually exclusive events, $E_1, E_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i).$$

집합 연산자

- 두 집합 A와 B가 mutually exclusive이면, $A \cap B = \varnothing$

$$P(A \cup B) = P(A) + P(B)$$

- 조건부 확률

$$p(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

$$P(A \mid B) = P(A) \quad or \quad P(B \mid A) = P(B)$$

$$\Leftrightarrow A \text{ and } B \text{ } are \text{ statistically independent}$$

The Total Probability Rule and Bayes' Theorem

- Bayes' Theorem
  - Given a set of prior probabilities for an event and some new information, the rule for updating the probability of the event is called **Bayes' theorem**.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)}$$

or

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}$$

Example : Bayes' Theorem

- Assume that 99% of the individuals taking a polygraph test tell the truth. These tests are considered to be 95% reliable (i.e., a 95% chance of actually detecting a lie). Let there also be a 0.5% chance that the test erroneously detects a lie even when the individual is telling the truth.
- An individual has just taken a polygraph test and the test has detected a lie. What is the probability that the individual was actually telling the truth?
- Let D denote the outcome that the polygraph detects a lie and T represent the outcome that an individual is telling the truth.

- Given the following probabilities,

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(T) = 0.99$	$P(D \mid T) = 0.005$	$P(D \cap T) = 0.00495$	$P(T \mid D) = 0.34256$
$P(T^c) = 0.01$	$P(D \mid T^c) = 0.95$	$P(D \cap T^c) = 0.00950$	$P(T^c \mid D) = 0.65744$
$P(T) + P(T^c) = 1$		$P(D) = 0.01445$	$P(T \mid D) + P(T \mid D^c) = 1$

We find $P(T \mid D) = \dfrac{P(D \cap T)}{P(D \cap T) + P(D \cap T^C)}$

$$P(T \mid D) = \frac{(0.005)(0.99)}{(0.005)(0.99) + (0.95)(0.01)} = \frac{0.00495}{0.01445} = 0.34256$$

Example

- 국내에 코로나 환자는 10명중 1명꼴이다. 코로나를 확인 하는 가정용 키트는 코로나 환자를 대상으로 검사하면 90% 양성반응을 나타내고, 건강한 사람을 대상으로 하면 99%가 음성반응을 나타낸다. 어느 날 몸이 이상하여 집에서 테스트를 해 보니 양성 반응이 나타났다. 그럼 우리가 알아야 하는 것은, 테스트 키트에서 양성 반응이 나타났을 때 코로나에 걸릴 확률은?

풀이

P(A)=1/10;event A는 코로나 환자

P(+|A)=0.9, P(-|A)=0.1

P(+|A^c)=0.01, P(-|A^c)=0.99

풀어야 할 식은 **P(A|+)**

$$P(A|+) = \frac{P(A \cap +)}{P(+)} = \frac{P(A)P(+|A)}{P(A)P(+|A) + P(A^c)P(+|A^c)}$$

$$= \frac{(0.1)(0.9)}{(0.1)(0.9)+(0.9)(0.01)} = \frac{0.09}{0.09+0.009} = \frac{0.09}{0.099} = 0.909$$

=90.9%

- **예를 들어, 국민 100명 중 3명이 감염된 상황을 가정했을 때** 민감도 90%·특이도 99%인 자가검사키트를 현장에서 사용하면, **자가검사키트로 양성이 나타난 사람 중 진짜 감염자가 나타나는 비율(양성예측도)는 73.6% 정도 나타납니다.** (계산식은 너무 어려워서 차마 못 가져 왔습니다ㅜㅜ)[출처] 민감도 90%로 허가받은 자가검사키트, 검사 현장에서 양성예측도가 76%인 이유는?|작성자 식약지킴이

P(A)=3/100;event A는 코로나 환자

P(+|A)=0.9(민감도), P(-|A)=0.1

P(-|A^c)=0.99(특이도), P(+|A^c)=0.01,

풀어야 할 식은 **P(A|+)-------양성예측도**

$$P(A|+)=\frac{P(A\cap+)}{P(+)}=\frac{P(A)P(+|A)}{P(A)P(+|A)+P(A^c)P(+|A^c)}$$

$$=\frac{(0.03)(0.9)}{(0.03)(0.9)+(0.97)(0.01)}=\frac{0.027}{0.027+0.0097}=0.73569$$

Random variable(확률변수)

$X(\omega) = x,\ \ X$ is function,

ω : element in Domain Set

x: real number

Then X is called the random variable

예: X(앞면)=1, X(뒷면)=0

확률분포(Probability Distribution)

확률변수 X는 여러 개 $(X = x)$의 값을 가진다

- X is discrete(이산) rv, when x is 1-1 to integer
  주사위 눈금, 성별, 골프스코어 등

- X is continuous(연속) rv, when x is 1-1 to real numbers
  체중, 시간…..

Cumulative distribution function (누적 분포 함수)

CDF of X is $P(X \le x)$

X가 이산인 경우,

$$P(X \le x) = \sum_{\text{all } x \le t} P(X = t) = \sum_{\text{all } x \le t} f_X(t)$$

we call $P(X = t)$ is probabity mass fuction

X가 연속인 경우

$$P(X \le x) = \int_{-\infty}^{x} f_X(t)dt$$

$f_X(t)$ is called probabity density fuction

Probability function

- Probability mass function(확률질량함수: 이산인 경우에만)

예: 주사위 경우,P(X=1)=1/6=$f_X(1)$=1/6

P(X=1.7)=0=$f_X(1.7)$=0

Probability density function(확률밀도함수: 연속인 경우에만 사용)

예:정규분포의 pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

P(X=0)=?

확률은?

Probability function(pmf 또는pdf), cumulative distribution function중 어느 것인가?

우리가 다루는 것은 cdf: $P(X \leq x)$: 넓이의 개념

그러면 $P(X = x) = P(X \leq x) - P(X < x)$

Example:

Expectation: $E(X)$: 기대값

$$E(X) = \sum_{all\ x} x \cdot P_X(X = x)$$

$$= \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \mu$$

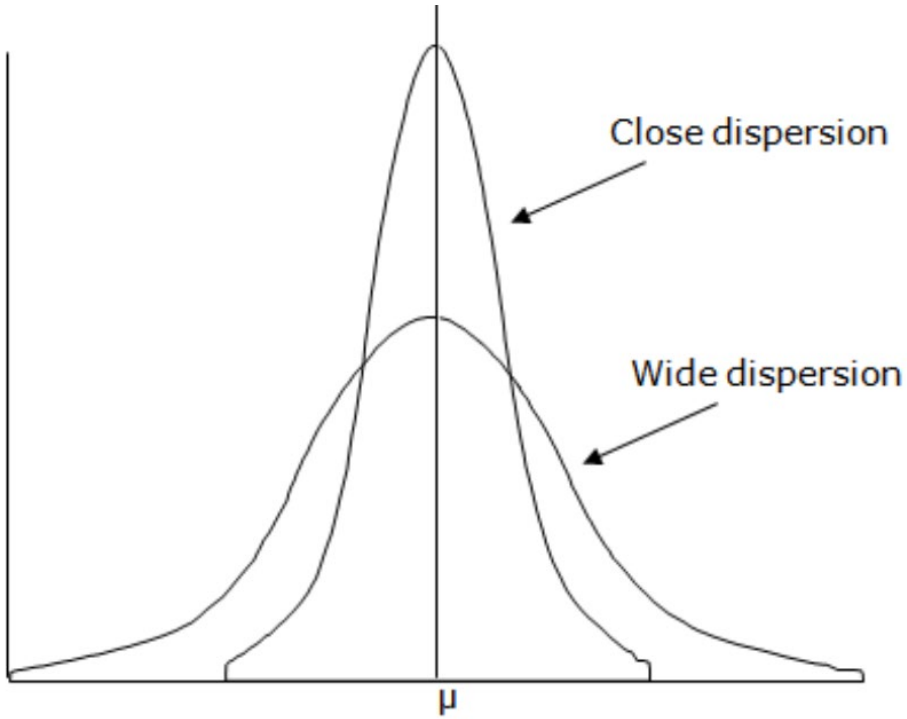
$$E(aX + b) = aE(X) + b$$

(a and b are constants)

예: 주사위 경우

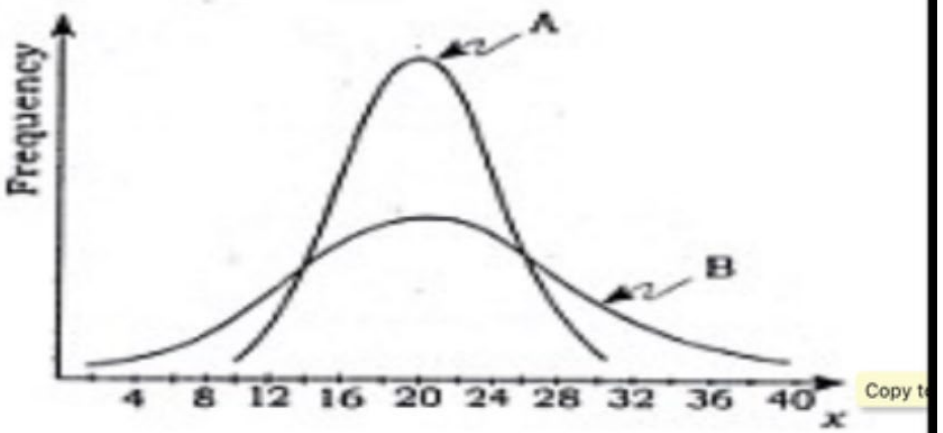
$$E(X) = \sum_{x=1}^{6} x \cdot P(X = x) = \sum_{x=1}^{6} x \cdot \frac{1}{6} = \frac{21}{6}$$

자료의 퍼짐의 정도는 어떻게?



Same mean and differnet dispersion

Data A: 10 11 14 20 20 20 22 24 28 31
Data B: 2 9 13 14 20 20 24 26 32 40



어떻게 숫자로 표현?

The above data sets (A& B) have similar mean but have different dispersion.
Data set B is more dispersed or spread from its mean.
Data set A is more clustered about the mean (majority of data cluster around the mean).

Variance : $V(X)$: 분산

$$|x_i - \mu| \Rightarrow \sum_i |x_i - \mu| \Rightarrow \frac{\sum_i |x_i - \mu|}{N}$$

$$\frac{\sum_i (x_i - \mu)^2}{N} (= \sigma^2)$$

퍼짐의 정도를 평균개념으로

$$V(X) = E[(X - \mu)^2] \left(= \sigma^2\right)$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sigma$$

인간과 컴퓨터, 누가 우위에 있을까?

- The human brain, according to a 2010 article in *Scientific American*, the memory capacity of the human brain was reported to have the equivalent of 2.5 *petabytes* of memory capacity. As a number, a "*petabyte*" means 1024 terabytes or a million gigabytes, so the average adult human brain has the ability to store the equivalent of *2.5 million gigabytes* digital memory.

컴퓨터

아직도 메모리가 부족, 속도가 느리다 =>Algorithm이 필요

컴퓨터(AI)가 인간을 지배할까?

Computer > Human or Human > Computer?

바둑: AlphaGo > 이세돌 (2016.3.9 ~3.15, 4:1)

Jeopardy Vs Watson(자연어 처리가 가능한 AI 컴퓨터)
(2011.2.16): Watson은 기계일 뿐, 진짜 '생각'은 이것을 만든 인간에게서 나왔다

Chess: Computer > Human

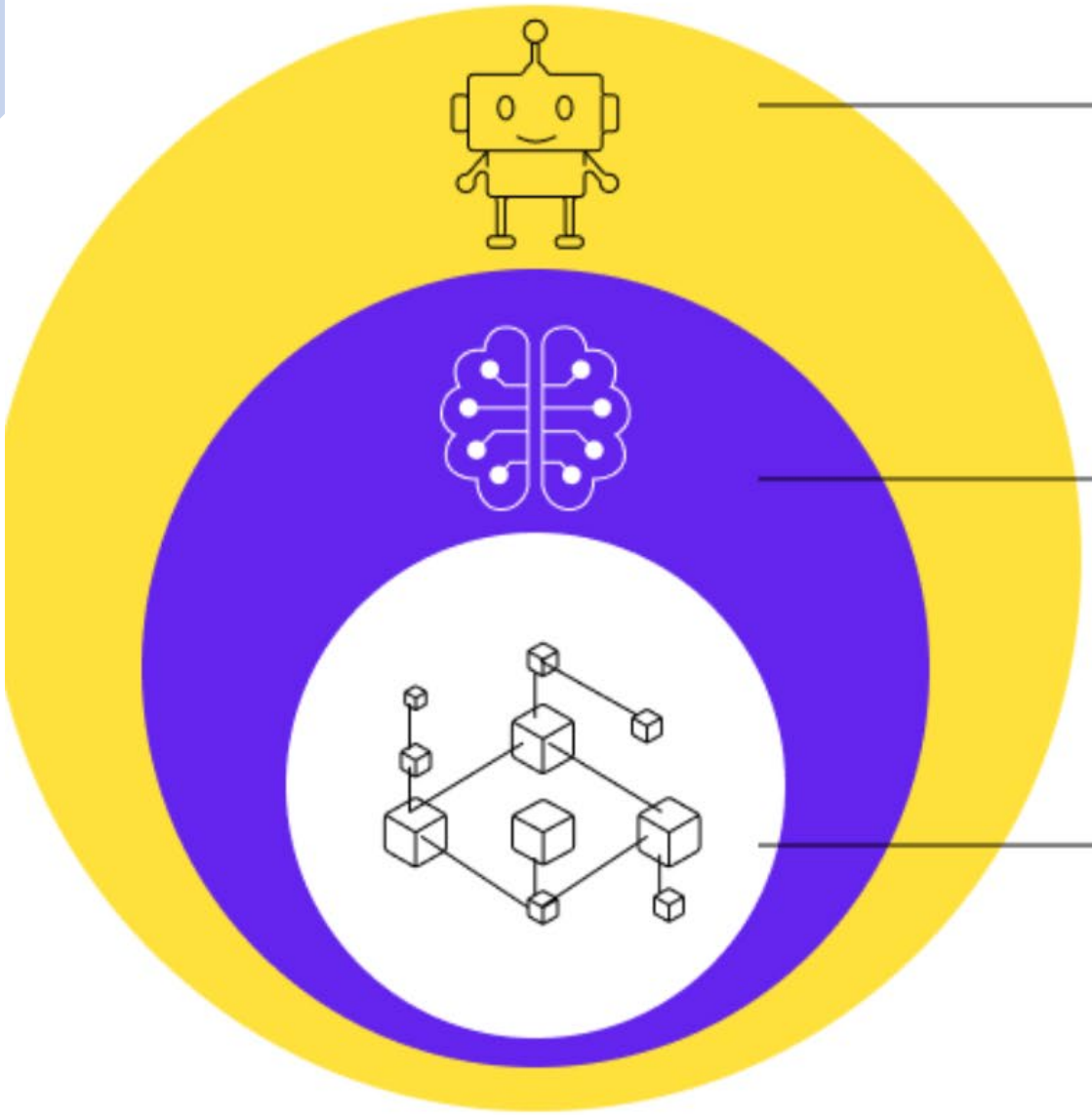
Q: Can Humans beat computers at chess?

Modern Data Scientist는

21세기의 가장 섹시한 직업인 데이터를 분석하는 과학자는 수학, 확률, 통계, Computer Science, 커뮤니케이션 및 비즈니스의 교차점에 이르기까지 다양한 분야의 기술이 혼합된 지식을 가지고 있어야 한다.
데이터를 분석하는 훌륭한 과학자를 찾는 것이 쉽지 않다. 제대로 지식을 갖춘 데이터 과학자가 누구인지 제대로 알고 있는 사람을 찾는 것도 마찬가지로 어렵다

이와 같은 전문가들을 찾기도 힘들고, 그 전문가를 평가하는 사람도 거의 없다.

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard.



Artificial Intelligence

Computers that can imitate human intellect and behavior.

Machine Learning

Statistical algorithms that enable AI implementation through data.

Deep Learning

Subset of machine learning which follows neural networking.

Courses for Learning Data Science

- **Step 1: Learn to code**

Learn a programming language before going deep into the math and theory behind data science models.

- **Step 2: Statistics**

Statistics is at the core of every data science workflow — it is required when building a predictive model, analyzing trends in large amounts of data, or selecting useful features to feed into your model.

- **Step 3: Foundational Math Skills**

Calculus and linear algebra are two other branches of math that are used in the field of machine learning.

- **Step 4: Machine Learning & AI**

Finally, you can use the knowledge gained in the courses above to take Introduction to Machine Learning & AI courses. This program will walk you through the implementation of predictive models in Python. Apart from just working with structured datasets, you will also learn to process image and sequential data.

Thanks!

- Math=Logic