1. 모집단과 표본

1.1 통계학이란 무엇인가?

"인류사회가 **제조업 기반의 경제**에서
지식과 데이터 위주의 사회로 **이동**할 것"

엘빈 토플러(Alvin Toffler, 1928~2016) 미래학자

"인공지능의 시대에 **통계학**이
매우 중요한 학문이 될 것이며,
미래의 종교는 **데이터 종교**"

-사피엔스(2015)-

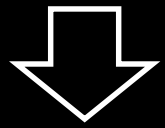
유발 하라리(Yuval Noah Harari, 1976~) 역사학자

統計學

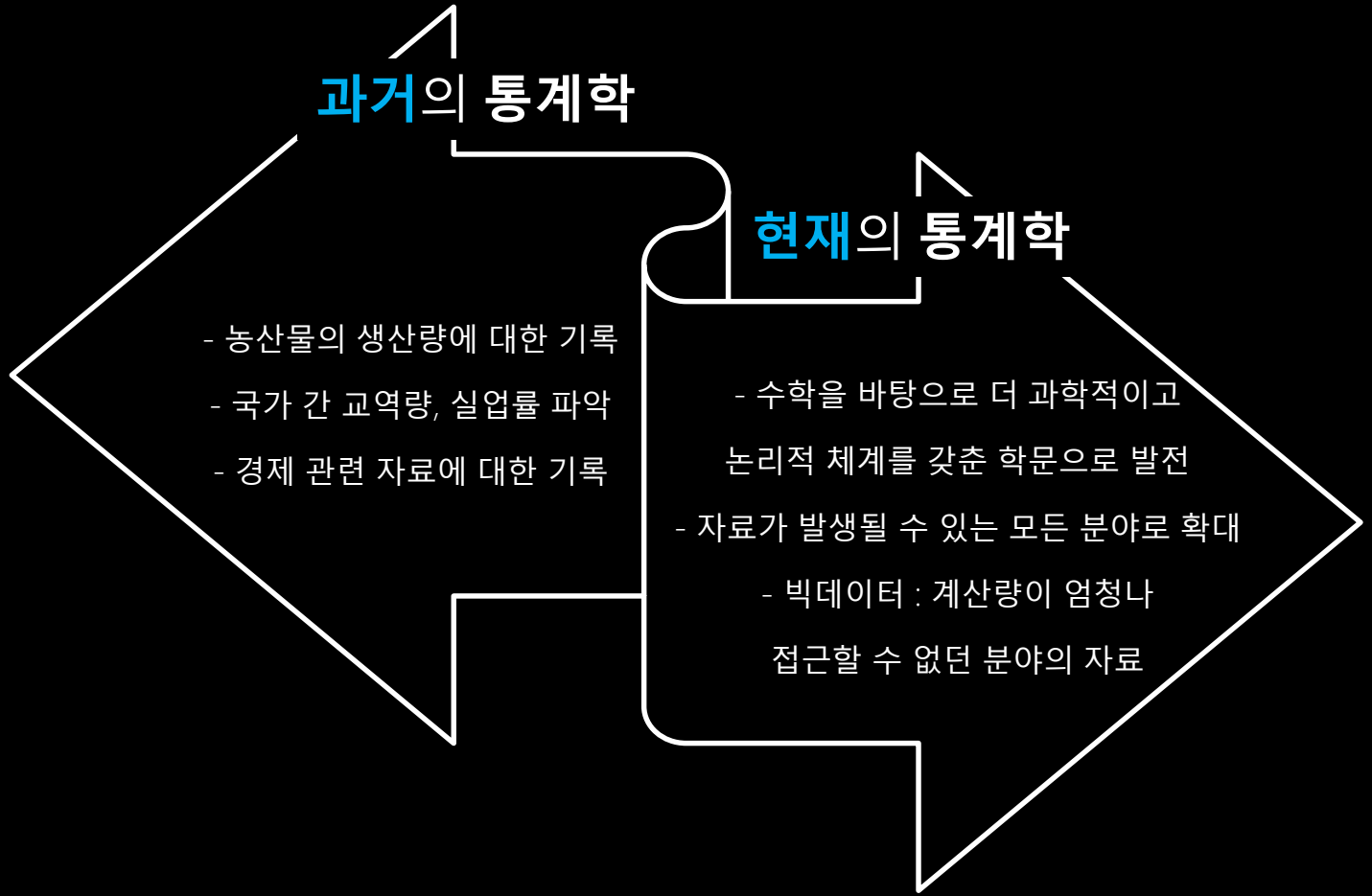
Statistics

“Status” + “ics”

라틴어로서state, 국가를 의미 학문이라는 뜻의 접미어



국가의 경영 또는 통치에 필요한 학문


과거의 통계학
- 농산물의 생산량에 대한 기록
- 국가 간 교역량, 실업률 파악
- 경제 관련 자료에 대한 기록
현재의 통계학
- 수학을 바탕으로 더 과학적이고
논리적 체계를 갖춘 학문으로 발전
- 자료가 발생될 수 있는 모든 분야로 확대
- 빅데이터 : 계산량이 엄청나
접근할 수 없던 분야의 자료

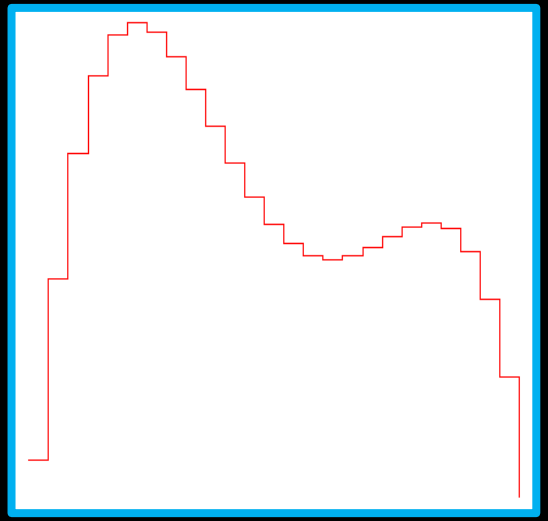
인공지능(AI : Artificial Intelligence)

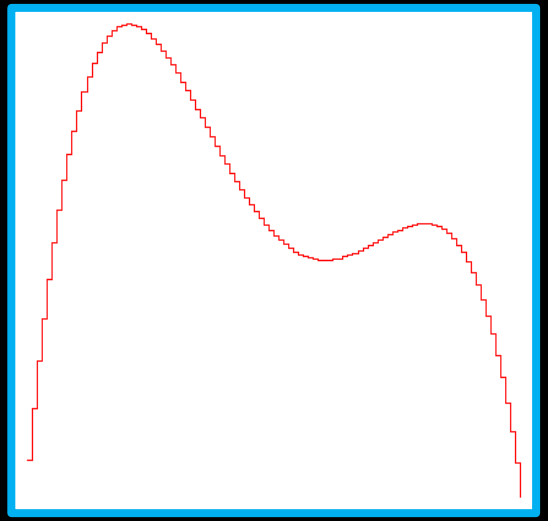
Basic Axiom in AI

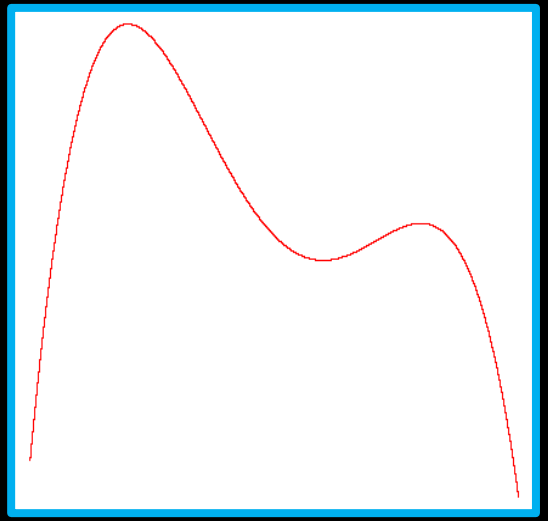
인공지능의 입력은 반드시 데이터의 형태를 가진다

인공지능(AI : Artificial Intelligence)

인공지능(AI : Artificial Intelligence)







인공지능(AI : Artificial Intelligence)

Basic Axiom in AI

인공지능의 **입력**은 반드시 **데이터**의 형태를 가진다

"AlphaGo(기보)"

인공지능(AI : Artificial Intelligence)

"AlphaGo"

- 19x19 = 361 => 361! = 10^{500}
- (17, 165, 33, 211, …) :
  모든 기보는 숫자로 표현됨
- cf. Google <= Googol = 10^{100}

인공지능(AI : Artificial Intelligence)

Basic Axiom in AI

인공지능의 **입력**은 반드시 **데이터**의 형태를 가진다

"AlphaGo(기보)"

"Siri(음성)"

인공지능(AI : Artificial Intelligence)

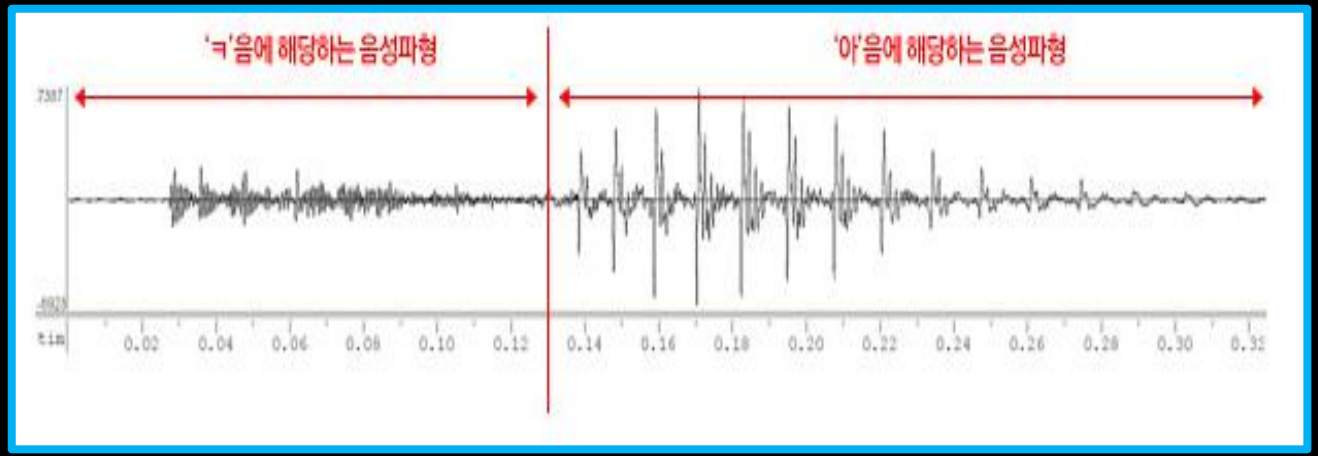
"Siri"

Speech Interpretation and Recognition Interface

"Bixby"

– 음성인식을 이용한 통역 및 번역

인공지능(AI : Artificial Intelligence)



인공지능(AI : Artificial Intelligence)

Basic Axiom in AI

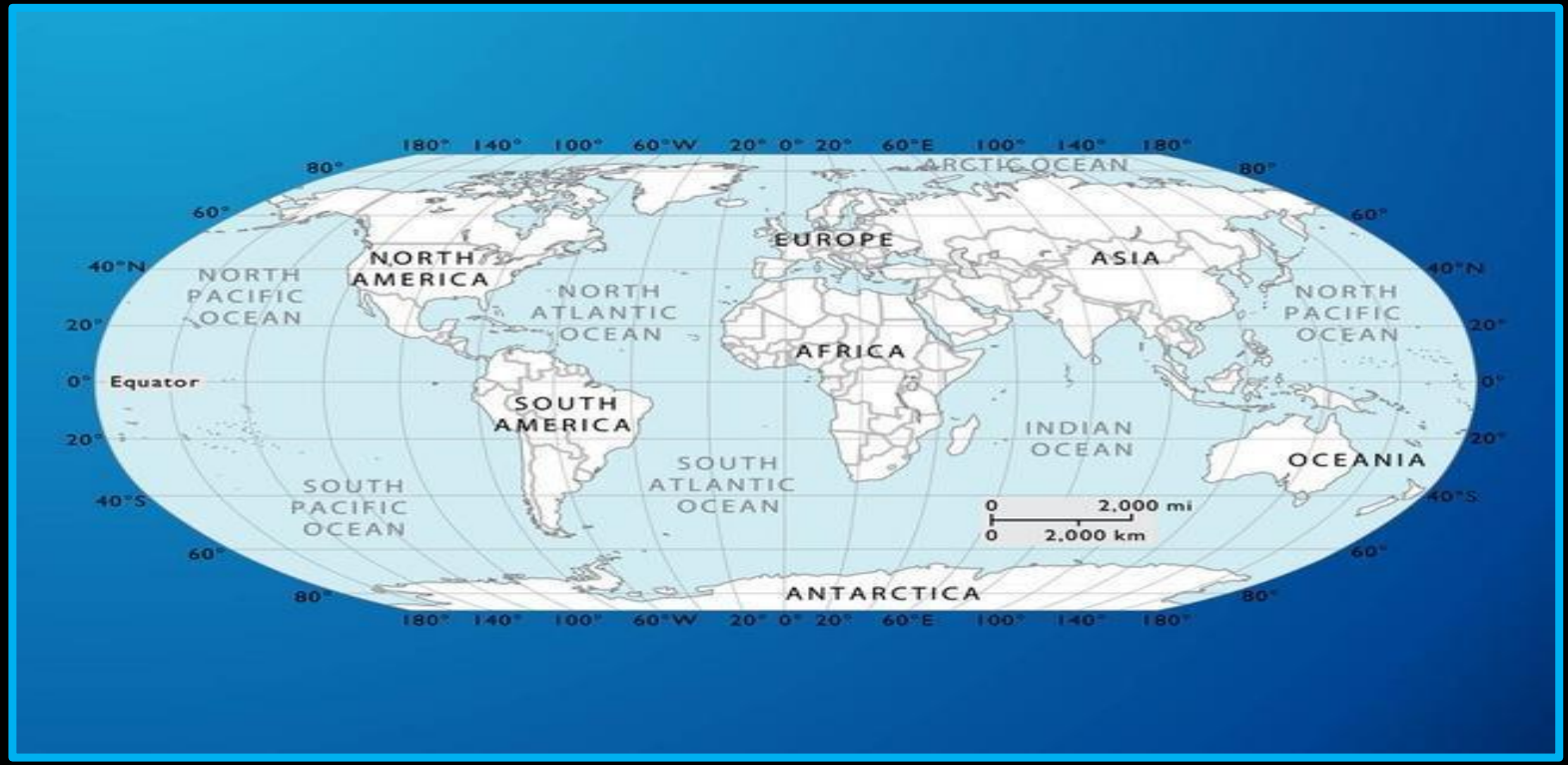
인공지능의 **입력**은 반드시 **데이터**의 형태를 가진다

"AlphaGo(기보)"

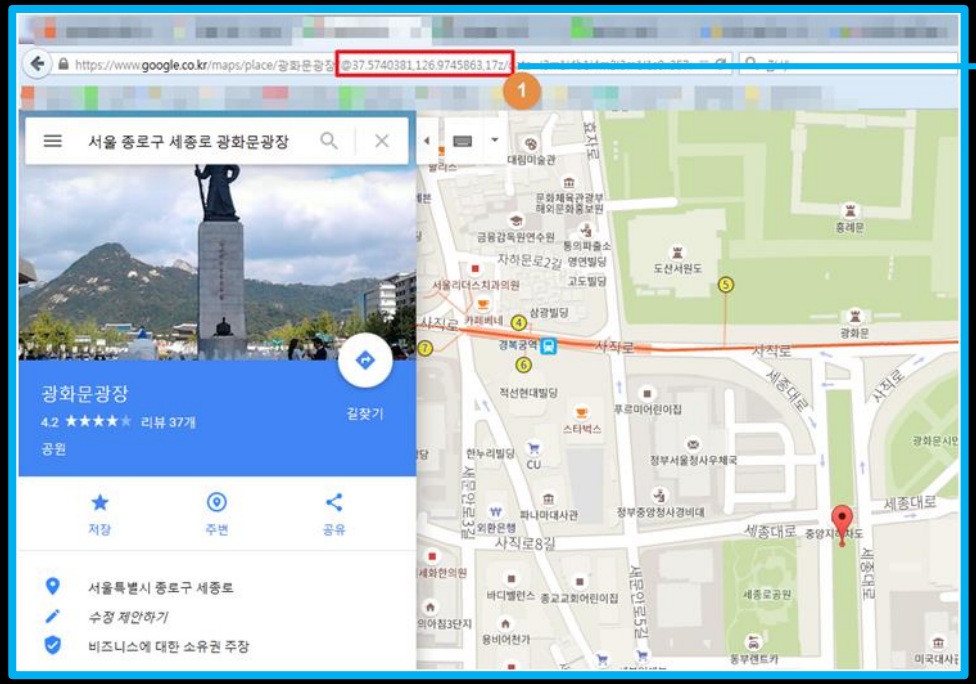
"Siri(음성)"

"Navigation(위도와 경도)"

인공지능(AI : Artificial Intelligence)



인공지능(AI : Artificial Intelligence)



위도 : 37.574

경도 : 126.974

인공지능(AI : Artificial Intelligence)

Basic Axiom in AI

인공지능의 **입력**은 반드시 **데이터**의 형태를 가진다

“AlphaGo(기보)”

“Siri(음성)”

“Navigation(위도와 경도)”

“Watson for Oncology”

인공지능(AI : Artificial Intelligence)

Watson

“Watson for Oncology”

종양학과 관련된 전문 지식과
의학 학술지 300권, 의학서 200권 등
1500만 쪽 분량의 의료 정보 탑재

- Watson for Genomics
- Watson for Solution

인공지능(AI : Artificial Intelligence)

Basic Axiom in AI

인공지능의 **입력**은 반드시 **데이터**의 형태를 가진다

"AlphaGo(기보)" "Siri(음성)"

"Navigation(위도와 경도)" "Watson for Oncology"

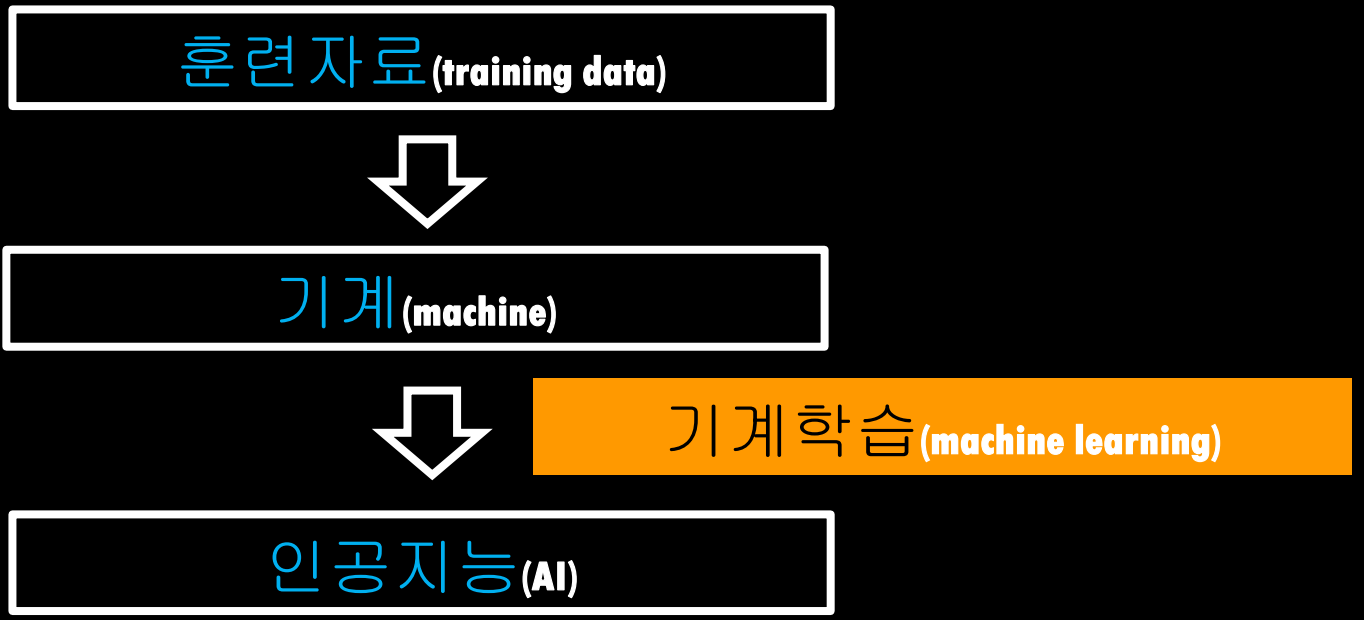
Two Core Technologies in AI

- Voice Recognition(음성인식) : Alexa, Siri, Assistant, Bixby, …
- Pattern Recognition(형상인식) : Drone, Autonomous Car, 3D, …

인공지능(AI : Artificial Intelligence)

"자율주행 자동차"

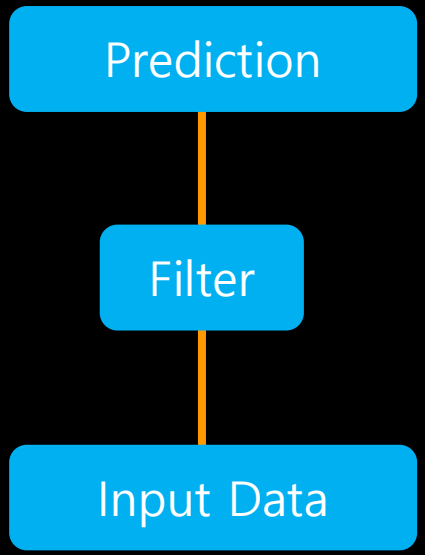
- 음성인식과 형상인식
- 현재 개발중인 인공지능의 최첨단 기술의 집약체
- 5~10년 내 실용화 가능

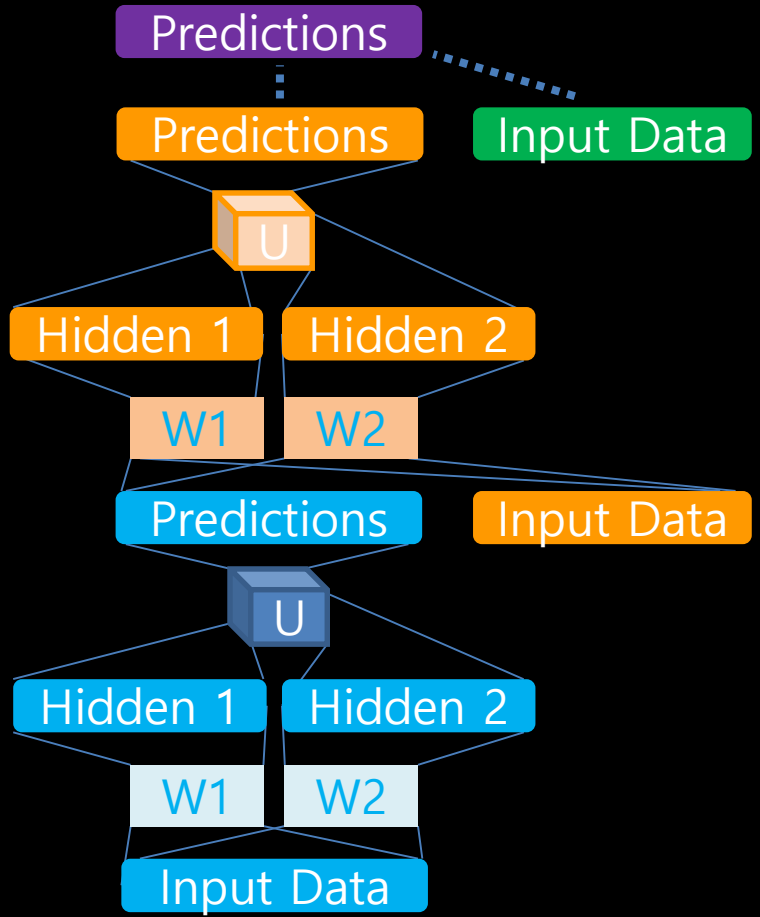


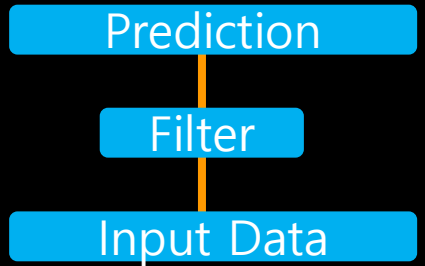
기계학습: 수학, 통계학, 전산학, 인문학, 사회과학, …

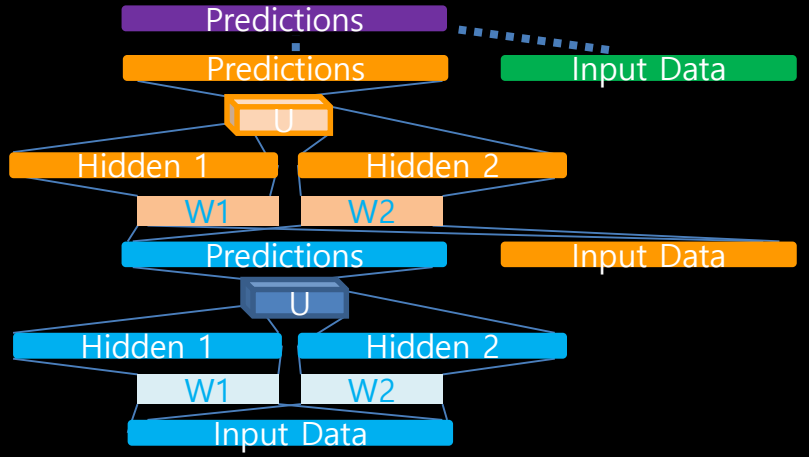
기계학습의 주요 방법: 심화학습(deep learning)

cf. 단순학습(shallow learning)


Prediction
Filter
Input Data


Predictions
Predictions
Input Data
U
Hidden 1
Hidden 2
W1
W2
Predictions
Input Data
U
Hidden 1
Hidden 2
W1
W2
Input Data





- Deep Learning: 인간의 신경망(neural network)을 본뜬 알고리즘으로 기계 스스로 학습

(예) Shallow Learning: 아침 -> morning

Deep Learning: 아침(밝다, 흐리다, …) -> morning

아침(맛있다, 과일, …) -> breakfast

인공지능: 입력된 자료를 분석하여

최적의 분류(classification)를 할 수 있는 기계

통계의 오용에 대한 경고

"세상에는 세가지 거짓말이 있다.
거짓말, 새빨간 거짓말, 그리고 통계"

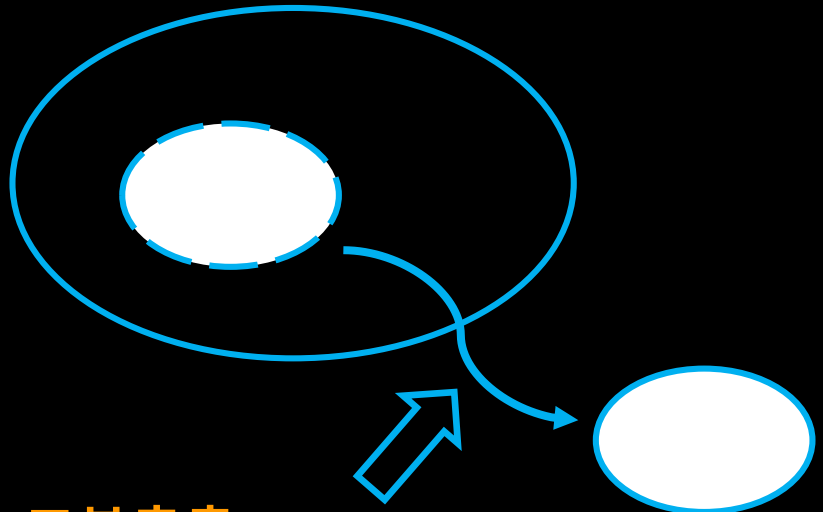
There are three kinds of lies:
Lies, damned lies, and statistics.

벤자민 디즈레일리(Benjamin Disraeli, 1804~1881)

1. 모집단과 표본

1.2 모집단과 표본

모집단(母集団：population)



표본추출(標本推出：sampling)

표본(標本：sample)

예) 대선 후보자 지지도 여론조사

"본 조사는 500명을 대상으로
실시하였으며 95% 신뢰구간에서
표준오차는 3.1%입니다 "

지지도를 **정확**하게 계산하려면
약 4천만 명의 유권자를 대상으로
전부 여론조사를 실시

엄청난 경비와 시간을 요구

지역, 성별, 나이 등 여러 변수를 고려하여
유권자의 일부를 임의로 선택

예를 들어, 1,000명이라는
일부의 유권자를 대상으로
전체 유권자의 표심을 추정하고자 하는 것

1. 모집단과 표본

1.3 R 들어가기

1.3.1 R 설치하기

(1) go to the site called CRAN (Comprehensive R Archive Networks)

http://www.r-project.org/

(2) execute "Download R“

(3) choose Korea http://cran.nexr.com/

(4) click "Download for R Windows“

(5) click "base“

(6) click “Download R 3.3.1 for Windows”

1.3.2 R 사용법

A. 주의할 점

(1) case sensitive

(2) commands are separated by ; or newline

(3) comments can be put anywhere starting with #

(4) subsequent commands are made by +

B. 내장 기능 (Inbuilt facilities)

(1) help, example, demo

help(solve) 또는 **?solve**

solve 라는 명령어 사용법에 대한 설명#

example(solve)

solve 라는 명령어에 대한 예제#

demo(persp)

persp 라는 명령어에 대한 예시#

(2) Data

data()

내장되어 있는 자료파일을 불러올 수 있음#

- women (height, weight, n=15)
- stackloss (Air.Flow, Water.Temp, Acid.Conc., stack.loss, n=21)
- faithful (eruptions, waiting, n=272)
- sleep (extra, group, n=20)

(3) Libraries

- Some useful libraries in R
  - lattice : lattice graphics
  - MASS : Modern Applied Statistics using S-Plus
  - mgcv : generalized additive models
  - nlme : mixed effects models
  - nnet : neural networks and multinomial log-linear models
  - spatial : spatial statistics
  - survival : survival analysis
- To see contents of "survival" library, for example, type

```
library(help=survival)
```

(4) Packages

You can install packages using "install packages".

You have to open "library" to use packages.

e.g.) packages -> install packages -> choice of country ->download "lars" ->
library(lars) -> ?lars

(5) data editing

To use a "bacteria" dataset in the "MASS" library, type

library(MASS); attach(bacteria); bacteria

C. Simple Manipulations : Numbers and Vectors

(1) Vectors and assignment

▶ R code

```
x <- c(10.4, 5.6, 3.1, 6.4, 21.7)

assign("x", c(10.4, 5.6, 3.1, 6.4, 21.7))

c(10.4, 5.6, 3.1, 6.4, 21.7) -> x

1/x

y <- c(x, 0, x); y
```

(2) Vector arithmetic

▶ R code

```
v <- 2*x + y + 1 ; v

15/7 : real
15%/%7 : integer part
15%%7 : remainder part
sum((x-mean(x))^2)/(length(x)-1)
var(x)
sqrt(-17) : NaN (not a number)
```

(3) Generating regular sequences

▶ R code

```
1:30 ; c(1:30); t <-c(1:30)

s3 <- seq(-5, 5, by=.2); s3

s4 <- seq(length=51, from=-5, by=.2) ; s4

s5 <- rep(x, times=5); s5

s6 <- rep(x, each=5); s6
```

** R에 내장되어 있는 자료 중 강좌에서 이용할 자료

- faithful

자료설명 : 미국 Yellowstone 국립공원 내에 있는 여러 간헐천 중에서 Old Faithful Geyser 에서 수집된 자료로서 2개의 변수와 272개의 관측치로 구성

변수 : eruptions (분출시간 (단위:분))

waiting (다음 분출될 때까지의 시간 (단위:분))

- Stackloss

자료설명 : 어떤 화학공정에서 여러 환경변화에 따른 암모니아의 산화비율을 측정한 자료로 4개의 변수와 21개의 관측치로 구성

변수 : Air.Flow (공기 주입량)

Water.Temp (물의 온도)

Acid.Conc. (질소농도)

stack.loss (암모니아 산화비율)