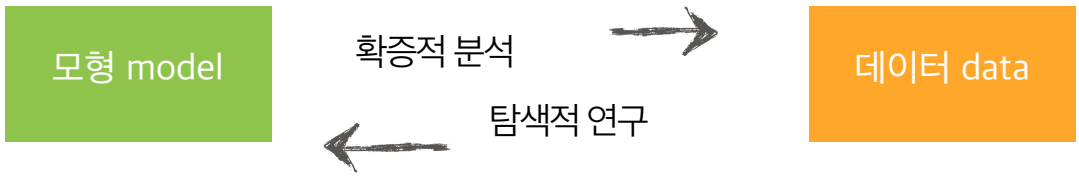
1) 데이터 철학

과학은 이론적 통찰 (예: 상대성 이론), 새로운 현상의 관찰 (Kepler 행성 궤도 관련 법칙)이나 경험을 (Student T-분포) 통한 새롭고 혁신적인 이론이 만들어지는 경우는 극히 드물고 대부분 관찰, 실험, 분석 등의 반복을 통해 이론이 정립된다. 벼 품종 개량, 새 의약품 개발, 화학 공정 개선 등이 실험 계획에 의한 연구 결과가 이에 해당된다.

통계 전문가는 제시된 이론을 통계적 가설이나 통계 모형으로 설정하고 관련 데이터를 수집하여 가설(모형)의 유의성을 검정하거나(confirmatory data analysis) 수집된 데이터를 탐색하여 가능한 모형이나 이론을 제시하는 역할을(exploratory data analysis) 담당하고 있다. 이처럼 (탐색적) 데이터 분석이 타 분야의 새로운 이론 발견에 기여할 수 있으려면 1)그 분야에 대한 지식 2)모형과 데이터 3)그리고 모형과 데이터의 사이클 개념을 올바로 이해해야 한다.

모형과 데이터 사이클



$$X_i \sim f(x;\theta) \qquad x_i \sim iidf(x;\theta)$$

$$f(x_1, x_2, \ldots, x_n) \sim [f(x;\theta)]^n$$

과학에서 이론이 제안되고 데이터 분석이 이루어지는 경우보다는 데이터로부터 새로운 이론이나 모형을 도출하는 경우가 많고 탐색적 자료 분석에 의해 제안된 이론이나 모형은 다시 confirmatory 방법에 의해 유의성이 (significance) 검증되므로 모형과 데이터는 순환 사이클을 갖는다.

통계적 모형은 과학적 진실이기 보다는 분석 대상이 되는 사실(현황)의 대표적 모형이다. 예를 들어, 회귀모형에서는 ($y = a + b \times x + e$) 선형함수(모형)이 설명하지 못하는 오차항(e)이존재하고이오차항은 평균 0, 분산 σ^2인 정규분포를 따른다고 가정한다.

데이터 정의

통계학에서 분석 대상인 데이터는 행과 열로 이루어져 있으며, 행은 개체 subject 관측치, 열은 관심 특성, 변수 variable로 이루어진 숫자 행렬이다.

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

행의 첨자는 개체를 나타내고, 열의 첨자는 변수를 나타낸다.

(정의) 추론, 토론, 계산에 사용되는 실제의 정보 (측정값 혹은 통계) (웹스터), 정보를 가진 숫자의 모임

2) 데이터 수집

+ Veracity(데이터 정확성) + Value(데이터 가치) = 5V of Big Data

데이터는 객관적인가? NO <- 수집되는 데이터는 목적이 있다. even 빅데이터 - 분석자의 의도없이 매초 단위로 엄청난 자동 저장되는 데이터도 일단 분석 대상이 되는 순간 그 데이터는 목적을 가지게 되므로 객관성을 상실한다.

데이터는 관심을 갖는 모집단 개체로부터 분석 대상 특성을 관측, 측정 등을 통하여 얻어지는 숫자 (고전적 데이터), 문자(텍스트 마이닝), 음성, 이미지(빅데이터) 향식이다.

3) 데이터 종류

(1)고전적 정의

통계학에서는 변수, 사회과학방법론에서는 척도라 한다.

✴ Metric (측정형 변수, measurable) : 실험 개체의 측 정 가능한 특성을 측정한 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측 정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다. 예) 교통량

• 구간척도 interval : 0 의미 없고 배율도 의미 없음 (예) 온도

• 비율척도 ratio : 0이라는 숫자의 의미가 있고 배율의 의미 존재 (예) 모든 측정형 소득

✴ Non-metric (분류형 변수, classified, 범주형 categorical) : 개체를 분류하기 위해 측정된 변수를 의미 하며 성별, 결혼여부 등이 그 예이다.

• 명목척도 (nominal): 분류만 성별, 결혼여부, 소득 (단위: 만원)

• 순서척도 (ordinal): 순서를 가진 분류 성적(A, B, ..) 소득수준(상, 중, 하), 5점 척도

	구간	비율	순서	명목
빈도표	X	X	X	X
순서 있음		X	X	X
최빈값	X	X	X	X
평균	X	X		
중위수	X	X	X	
+, - 가능	X	X		
곱셈, 나누셈		X		
0의 개념, 배율		X		

(2) 시간적 정의

✴ 횡단 변수 cross section : 일정 시점의 조사 데이터

✴ 종단변수 time series : 시간적 순서를 갖는 데이터 - X_i 대신 X_t로 표현, 경제 지표(환율, 수출량) 나 기업의 연차별 자료(연도별 매 출액), 연도별/월별 청년 실업율

(3) 인과관계 casual relationship

인과 관계 모형은 통계분석에 의해 검증되는 것이지 발견하는 것은 아니다. 모형 설정은 이론, 경험적 타당성에 근거하여 이루어진다.

- X - 독립변수, 요인(처리효과), 예측변수, 설명변수, 내생변수 : 원인이 되는 변수
- Y - 종속변수, 반응변수, 목표변수, 외생변수 : 영향을 받는 변수

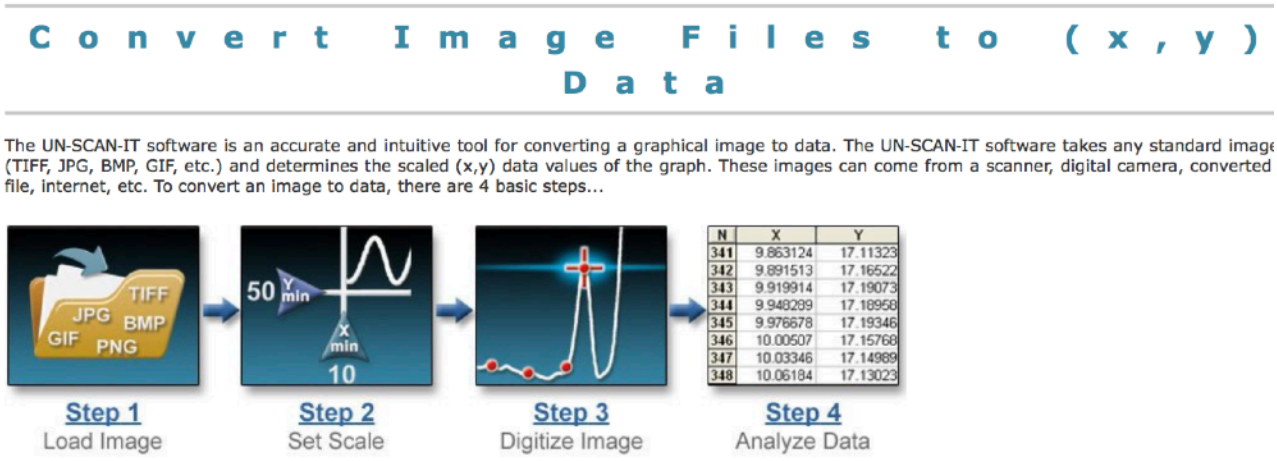
(4) 포멧

- 숫자 : 명목형 데이터는 문자(범주)이지만 분석 시 class 변수로 설정하여 숫자처럼 사용한다.
- 문자 : 텍스트 마이닝, 자연어 처리, word cloud
- 음성 : 오디오 파일 포멧을 데이터로 변환/역변환 가능 (convert audio format into data format)

```
> s7<-readWave("mysong.wav", from = 1, to = 5, units = "seconds")
> s7
```

```
Wave Object
        Number of Samples:      32000
        Duration (seconds):     4
        Samplingrate (Hertz):   8000
        Channels (Mono/Stereo): Mono
        PCM (integer format):   TRUE
        Bit (8/16/24/32/64):    16
```

- 이미지 : 이미지 파일 (x, y) 좌표나 6자리 숫자 단위로 변환 가능



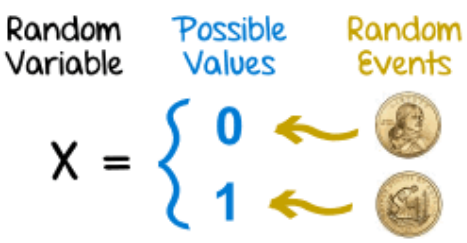
Assuming all your .img files are in one folder and the ENVI file format suffices as the binary output this R code works.

```
library(raster)

directory <- "/path/to/IMG/files"
setwd(directory)

## create vector containing all image filenames
Images <- dir(directory, pattern="\\.img$") #edit the pattern (case sensitive!) if yo

## create the subdirectory
dir.create(file.path(directory, 'binary'))

for (i in 1:length(Images)){
  outRaster = raster(Images[i])
  writeRaster(outRaster, filename=paste(directory, '/binary/', Images[i], sep=""), fo
  }
```

✳ 동영상 : CCTV 범죄 프로파일링, 3차원 CT/MRI

4) 데이터와 확률변수 random variable

통계분석의 대상이 되는 데이터는 변수들로 구성되어 있는데, 변수의 관측결과와 숫자를 일대일 매칭한 함수를 확률변수라 한다. 변수의 관측치가 숫자인 경우는 데이터 값 자체가 확률변수이다.



확률변수가 가질 수 있는 값과 대응하는 확률을 일대일 매칭한 함수를 확률분포함수(pdf probability density function)라 한다. PDF 는 데이터(변수)가 가진 모든 정보를 표현하고 있다. 변수의 관측값이 가질 수 있는 가장 큰 값, 작은 값, 중앙 위치, 그리고 어느 구간이 관측 가능성이 가장 높은지 등, 모든 정보를 확률밀도함수가 가지고 있다.