

R 전문가로 가는 길

-- 빅 데이터 활용 바로 보기 --



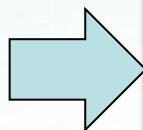
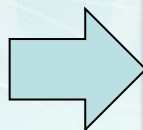
Heewon Jeon (NexR Corp.)

- Author/Maintainer of KoNLP package.
- Admin of Korea CRAN server

Interactive Data Analysis

레거시 데이터 분석

- 컴퓨팅 리소스가 굉장히 비쌌다.
- 많은 입력 값
- 많은 출력 값
- 부담없이 여러 번 수행하기 힘들
- 모든 결과를 쓰는 건 아님

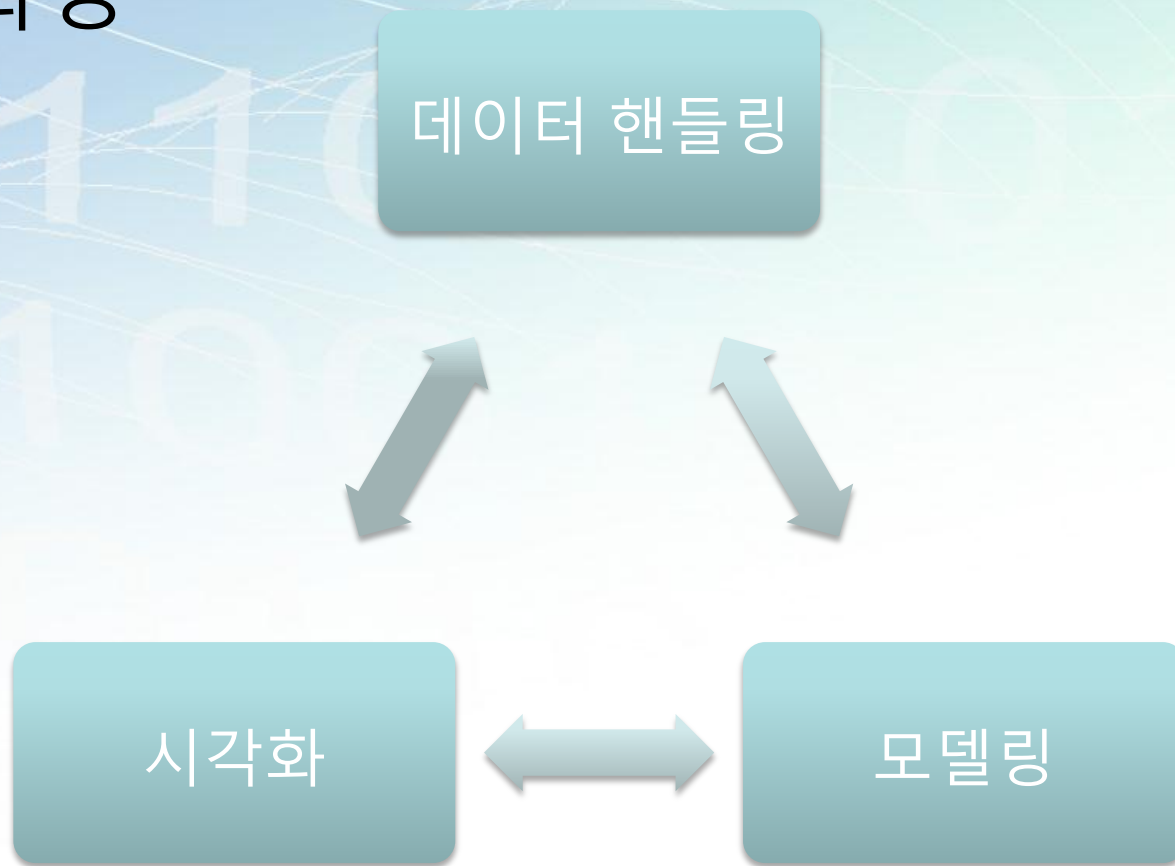


현재 데이터 분석

- 컴퓨팅 리소스가 굉장히 싸졌다.
- 어떤 분석을 수행하든 부담이 없어짐
- 데이터 입력, 변환, 무응답 대체, 데이터 핸들링, 시각화, 모델링 등 분석 등 재반의 작업을 반복 수행하면서 알고자 하는 의문을 하나 둘씩 풀어가는 분석이 수행 가능해짐

역동적인 분석에 적합한 언어 R

일반화된 데이터 분석 과정



R is an environment for...

- 데이터 핸들링
 - 데이터 소스에 접근하고
 - 자르고, 붙이고, 변형하고...
 - 모델링/시뮬레이션
 - 통계 모델
 - 통계 시뮬레이션
 - 데이터 시각화
 - 일반적인 통계 시각화
 - 진보되고, 다양한 시각화를 위한 패키지
-
- ```
graph LR; A[데이터 핸들링] --> B[모델링/시뮬레이션]; B --> C[데이터 시각화];
```
- The diagram illustrates the workflow of R. It starts with '데이터 핸들링' (Data Handling), which leads to '모델링/시뮬레이션' (Modeling/Simulation), and finally to '데이터 시각화' (Data Visualization). Each step is represented by a light blue box with its name written inside. Arrows indicate the sequential flow from left to right. The background of the slide features a green and blue gradient with binary code (0s and 1s) and a network of white lines.

# Why R?

- R은 공짜다.
- R은 문서화가 잘 되어있다.
- R은 대부분의 플랫폼에서 잘 돌아간다.
- R은 오픈소스이다.
- R은 다양한 통계 패키지를 포함하고 있다.
- R은 시각화에 강하다.
- R은 직관적인 데이터 핸들링을 제공한다.
- R은 복잡한 일을 처리하기 적합하다.
- R은 재현성을 충분히 발현할 환경을 갖추고 있다.
- R은 교과서에 나온 통계적인 용어를 그대로 사용한다.
- R은 학생들로 하여금 프로그래밍을 하도록 유도한다.
- R이 배우는데 많은 시간이 걸리지만, 일단 학습후에는 사용자로 하여금 다양한 분석을 할 수 있는 자유로움을 준다.
- R은 ~~빅 데이터용 분석 환경~~이다.



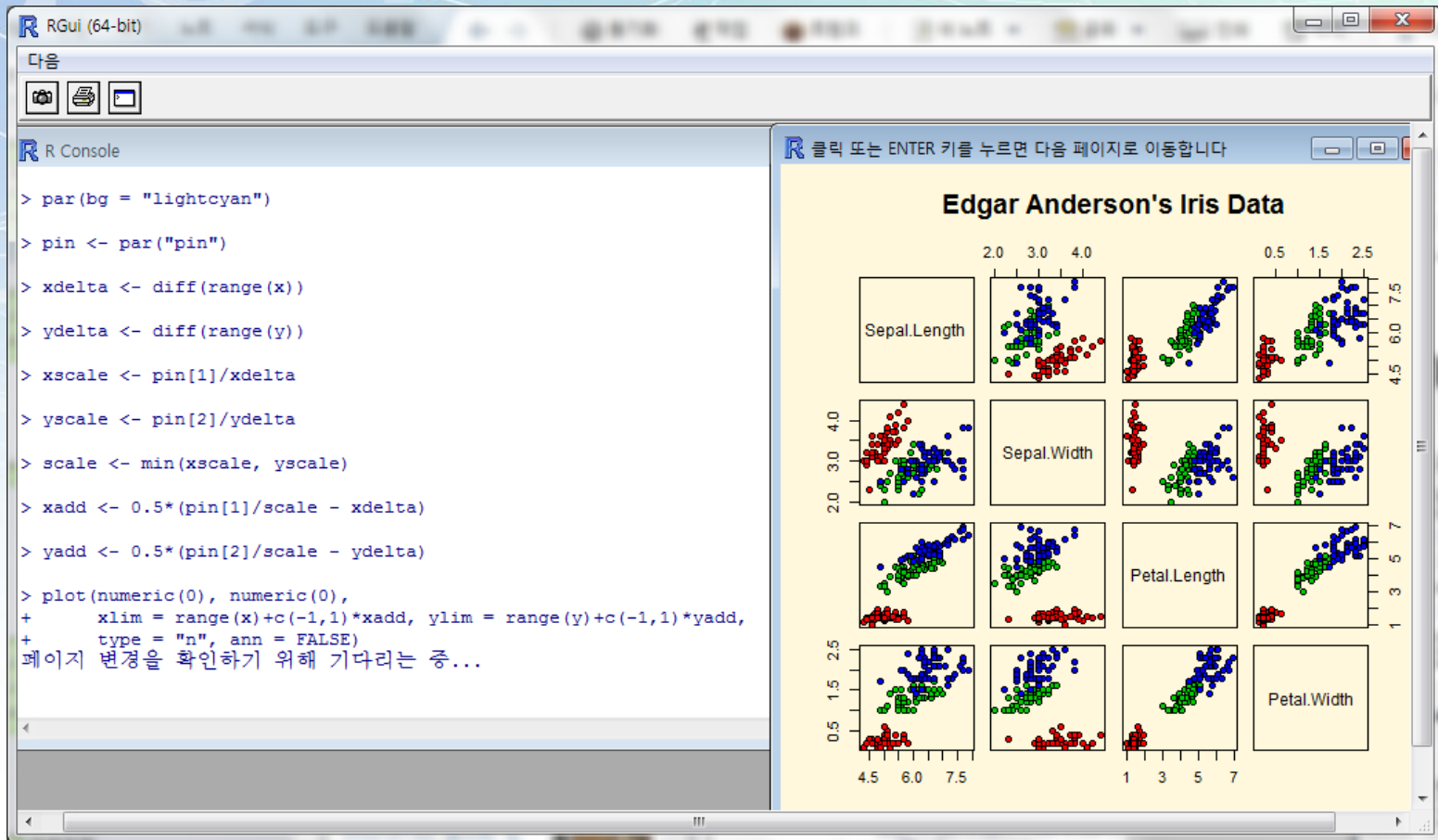
# Really R?

- 개발자도 배우기 쉽지 않은 언어
  - 함수형, 절차형 언어의 혼합
  - 통계용어 난무
- 통계학자도 배우기 쉽지 않은 언어
  - 프로그래밍의 어려움
- 자유로운 분석 추구  $\neq$  쉽다.

# But R!

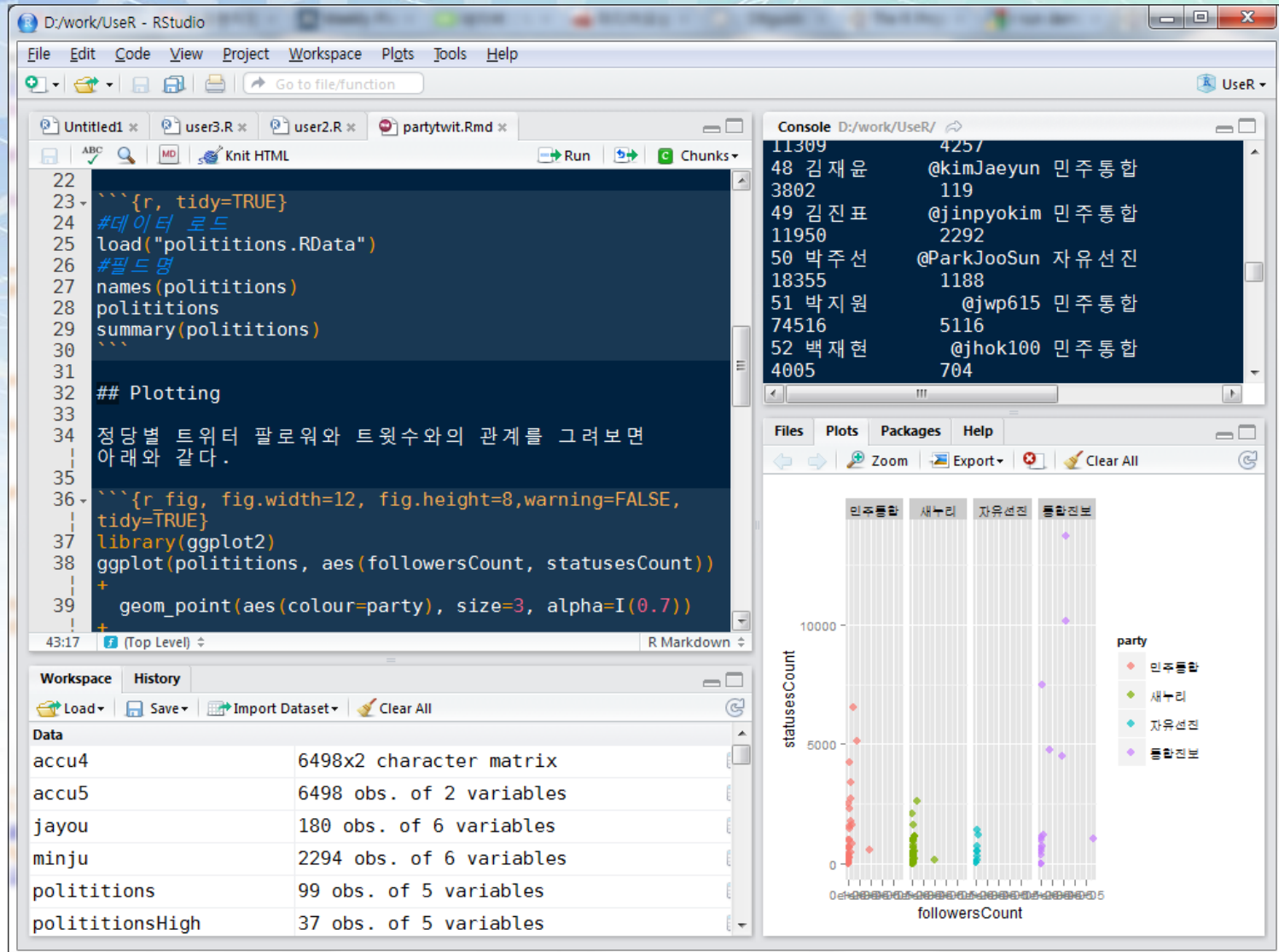
- 해외 리서치 영역에서는 R이 기본이 되었음
  - Reproducible Research
  - Literate Programming
- 국내 대학에서 R을 가르치기 시작
- 대형 벤더에서 R을 인터페이싱 함
  - Oracle, Teradata, SAS, SPSS ...

# R has simple GUI





# RStudio is better



# R Package System

[Bayesian](#)  
[ChemPhys](#)  
[ClinicalTrials](#)  
[Cluster](#)  
[DifferentialEquations](#)  
[Distributions](#)  
[Econometrics](#)  
[Environmetrics](#)  
[ExperimentalDesign](#)  
[Finance](#)  
[Genetics](#)  
[Graphics](#)  
[HighPerformanceComputing](#)  
[MachineLearning](#)  
[MedicalImaging](#)  
[Multivariate](#)  
[NaturalLanguageProcessing](#)  
[OfficialStatistics](#)  
[Optimization](#)  
[Pharmacokinetics](#)  
[Phylogenetics](#)  
[Psychometrics](#)  
[ReproducibleResearch](#)  
[Robust](#)  
[SocialSciences](#)  
[Spatial](#)  
[Survival](#)  
[TimeSeries](#)  
[gR](#)

## Total 3,921 Packages

- 오픈소스 라이선스의 파워
- Fortran, C++, C, Java 등 대부분의 언어와 연동 가능한 R의 유연성
- 리서치 영역에서 활발한 사용

# If you want to do twitter analysis.

- Data Source
  - **twitterR**
- Data Preprocessing
  - **KoNLP**
- Visualization
  - **wordcloud**

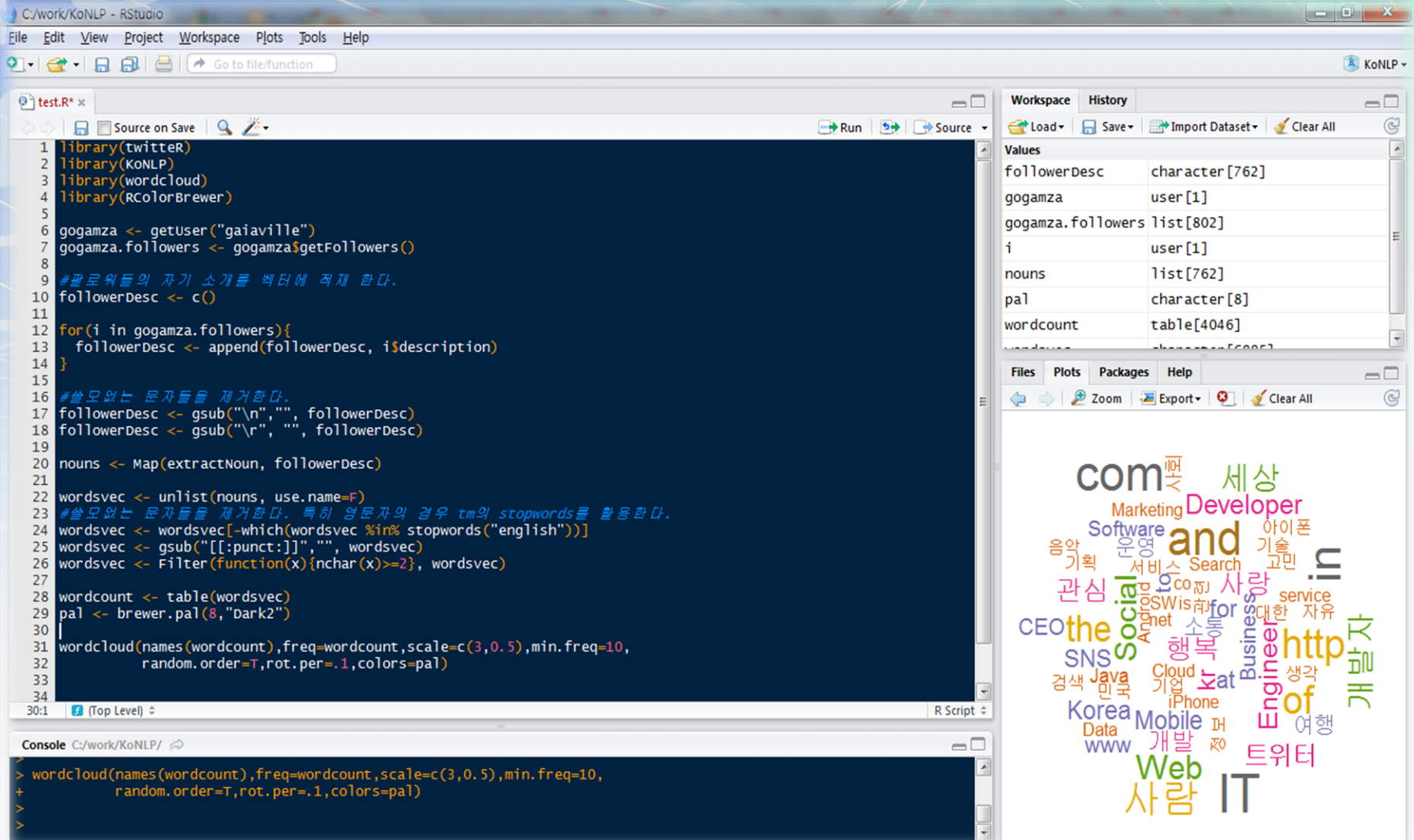
분석 방법 구상

적은 시간으로 구현  
(약 30라인)

평가 or 리포팅

실제 분석가의 상상력의 한계만 있을 뿐이며,  
어떤 분석이든지 필요한 것 대부분은  
패키지에서 커버하고 있음

# R Packaging System 사용 예 - 1





# R Packaging System 사용 예 - 2

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for loading libraries (KoNLP, arules, igraph, combinat), reading a text file, extracting nouns, filtering rules, and creating a network graph.
- Workspace:** Lists data objects: edgelist (18086x2 character matrix), f2 (71 obs. of 2 variables), rulemat (27x2 character matrix), rulequality (43 obs. of 3 variables), wordtab (1547x1547 integer matrix), and KtoS (character[71]).
- Plots:** Displays a network graph with nodes labeled in Korean (e.g., 다문화, 연구, 사회, 시대, 정책, 중심, 음악, 교육, 가족, 다문화.연구, 가정.다문화, 가정, 교사, 정보과학) and edges representing relationships.
- Console:** Shows the execution of the R code, including the final plot command.

```
1 library(KoNLP)
2 library(arules)
3 library(igraph)
4 library(combinat)
5
6 f <- file("다문화_utf8.txt", encoding="UTF-8")
7 fl <- readLines(f)
8 close(f)
9 tran <- Map(extractNoun, fl)
10 tran <- unique(tran)
11 tran <- sapply(tran, unique)
12 tran <- sapply(tran, function(x) {Filter(function(y) {
13 nchar(y) <= 4 && nchar(y) > 1 && is.hangul(y)}
14 ,x)})
15 tran <- Filter(function(x){length(x) >= 2}, tran)
16 names(tran) <- paste("Tr", 1:length(tran), sep="")
17 wordtran <- as(tran, "transactions")
18
19 ares <- apriori(wordtran, parameter=list(supp=0.05, conf=0.05))
20 rules <- labels(ares, ruleSep="/", setStart="", setEnd="")
21 rules <- sapply(rules, strsplit, "/", USE.NAMES=F)
22 rules <- Filter(function(x){!any(x == "")}, rules)
23 rulemat <- do.call("rbind", rules)
24
25 ruleg <- graph.edgelist(rulemat, directed=F)
26 plot.igraph(ruleg, vertex.label=V(ruleg)$name, vertex.label.cex=0.5,
27 vertex.size=20, layout=layout.fruchterman.reingold.grid)
28
29
```

```
> rules <- labels(ares, ruleSep="/", setStart="", setEnd="")
> rules <- sapply(rules, strsplit, "/", USE.NAMES=F)
> rules <- Filter(function(x){!any(x == "")}, rules)
> rulemat <- do.call("rbind", rules)
> ruleg <- graph.edgelist(rulemat, directed=F)
> plot.igraph(ruleg, vertex.label=V(ruleg)$name, vertex.label.cex=0.5, vertex.size=20,
> layout=layout.fruchterman.reingold.grid)
>
```



# R data structures for DBA - 1

Vectors

numeric

```
x <- c(0, 2:4)
```

character

```
y <- c("cat", "dog", "cat", "cat")
```

logical

```
z <- c(TRUE, TRUE, T, F)
```

# R data structures for DBA - 2

Object

list

```
a <- list(x,y,z)
```

matrix

```
b <- matrix(rep(x,3), ncol=3)
```

data.frame

```
c <- data.frame(x,y,z)
```

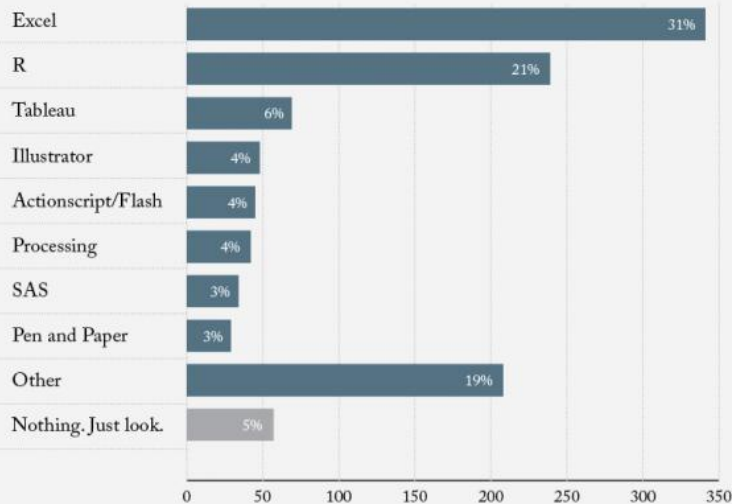
# R data structures for DBA - 3

```
1 library(sqldf)
2
3 iris3 <- sqldf("select Sepal_Length, Sepal_Width, Species from iris")
4
5
6 head(iris3)
7 sqldf("select * from iris3 limit 6")
8 # Sepal_Length Sepal_Width Species
9 # 1 5.1 3.5 setosa
10 # 2 4.9 3.0 setosa
11 # 3 4.7 3.2 setosa
12 # 4 4.6 3.1 setosa
13 # 5 5.0 3.6 setosa
14 # 6 5.4 3.9 setosa
15 |
16
17 subset(iris3, Sepal_Length > 5)
18 sqldf("select * from iris3 where Sepal_Length > 5")
19
20
21 with(iris3,
22 aggregate(list(Sepal_Length, Sepal_Width), by=list(Species), mean)
23)
24 sqldf("select Species, avg(Sepal_Length), avg(Sepal_Width)
25 from iris3 group by Species")
26
27
28 head(iris3[order(iris3$Sepal_Length, decreasing = TRUE),], 3)
29 sqldf("select * from iris3 order by Sepal_Length desc limit 3")
30
```

# Popularity of R

## What do you use to analyze and/or visualize data?

1,112 responses

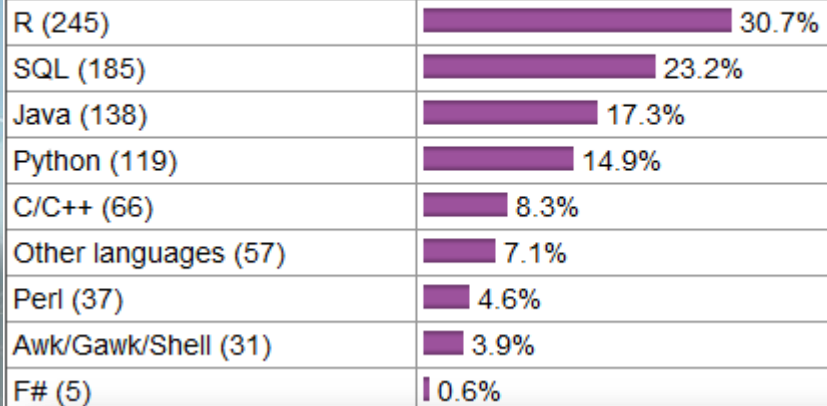


FlowingData Poll, September 2010

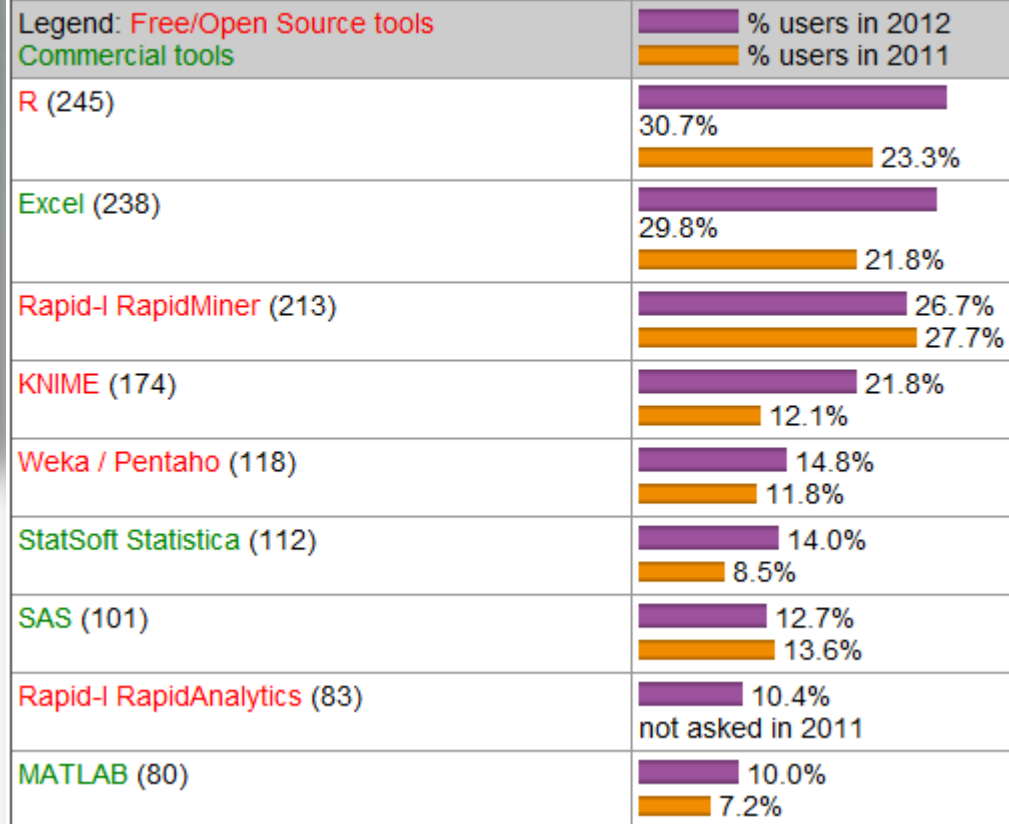
<http://flowingdata.com/2010/09/28/poll-results-what-do-you-use-to-analyze-and-or-visualize-data/>

<http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html>

## Your own code you used for analytics/data mining in the past 12 months in:

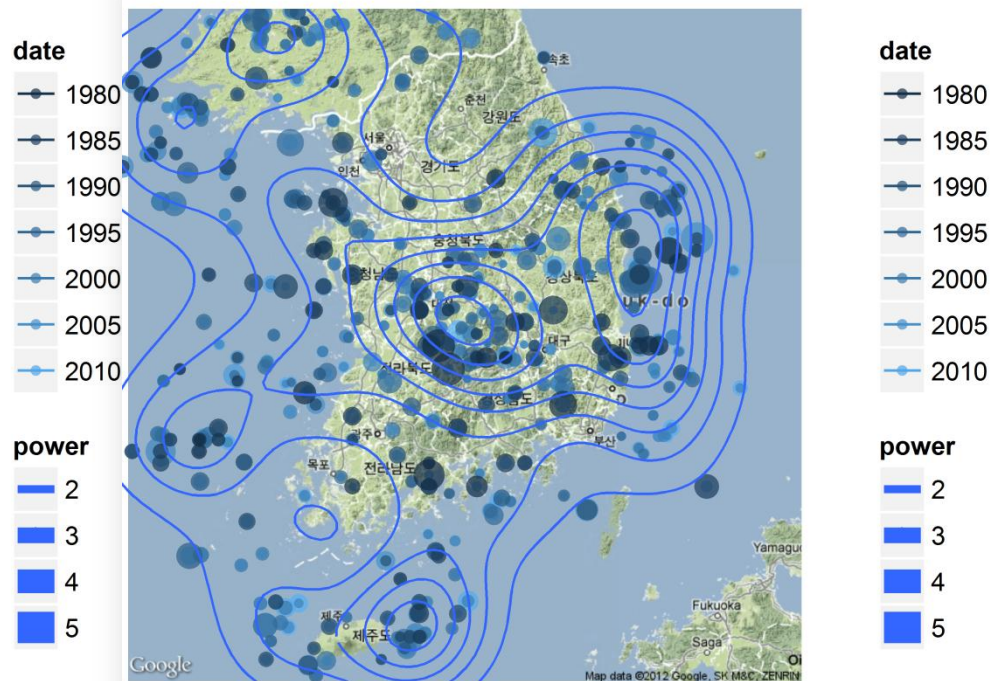
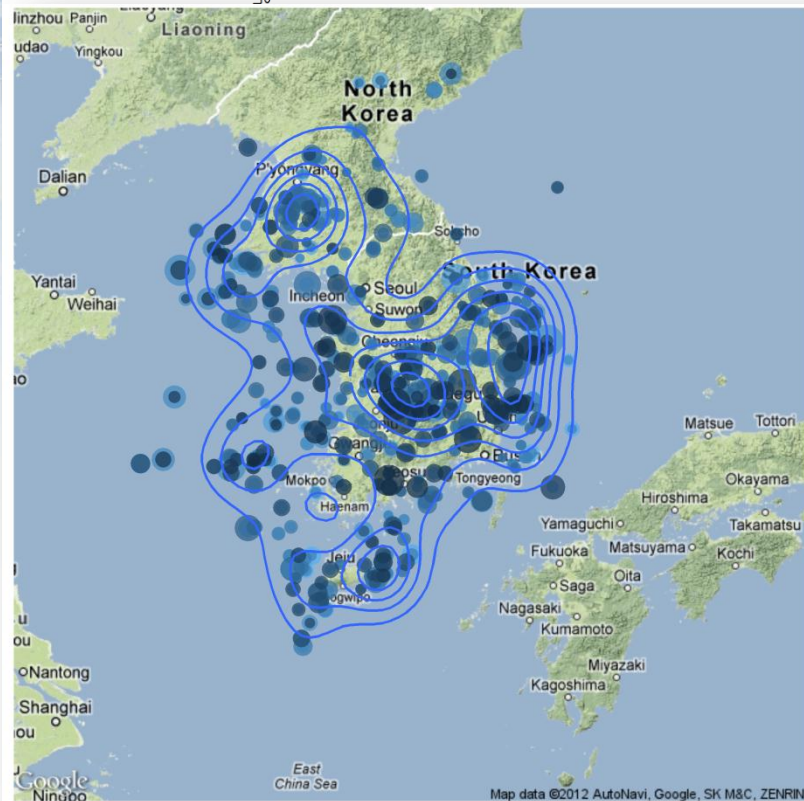
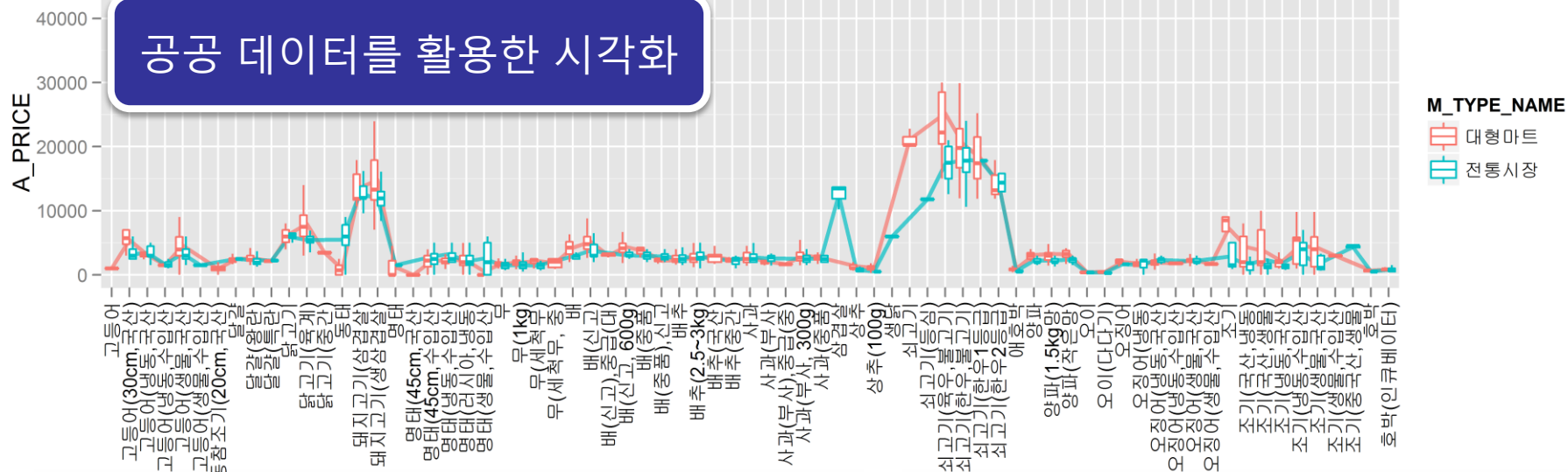


## What Analytics, Data mining, Big Data software you used in the past 12 months for a real project (not just evaluation) [798 voters]





# 공공 데이터를 활용한 시각화





# 빅 데이터 분석에서의 R의 문제점/해결책

## 메모리 한계 이슈

모든 데이터를 메모리에 로딩 후 처리하는 작업 방식

ff, bigmemory, RevoScaleR

10GB 이상 데이터는 처리 가능하나 너무 느리다는 단점

불필요한 데이터 저장으로 인한 메모리 부족 현상

gc(), rm()

32비트에서 표현 가능한 숫자만이 사용,  $2^{31}-1$

R 2.15부터  $2^{51}$  이상의 벡터 길이 사용 가능

No int64

int64 package from Google

메모리 단편화

64bit 머신 사용

더 많은 메모리

## Single Core 이슈

멀티코어 CPU에서 1코어만 사용한다.






R 2.14 부터 parallel 패키지 기본 탑재

**TB급 빅 데이터  
는 여전히 처리  
하기 힘들**

# 독보적인 Hadoop기반 Big Data 분석 플랫폼

## Big Data

Big data tools use grew 5-fold, from about 3% to about 15% of respondents.

| Big Data software you used in the past 12 months  |                                                                                          |
|---------------------------------------------------|------------------------------------------------------------------------------------------|
| Apache Hadoop/Hbase/Pig/Hive (67)                 |  8.4% |
| Amazon Web Services (AWS) (36)                    |  4.5% |
| NoSQL databases (33)                              |  4.1% |
| Other Big Data Data/Cloud analytics software (21) |  2.6% |
| Other Hadoop-based tools (10)                     |  1.3% |

- 세계적인 데이터 분석 커뮤니티인 Kdnugget의 설문조사
- 작년에 비해 5배 이상 빅 데이터 응답자가 늘어났다.
- 작년에 이어 Hadoop 기반의 오픈소스 플랫폼이 1위

# RHipe

```
1 bigmeanmap <- expression({
2 y <- do.call("rbind", lapply(map.values, function(r){
3 as.numeric(strsplit(r, ",")[[1]])
4 })
5 })
6 summed<- colSums(y, na.rm=T)
7 nr <- nrow(y)
8 nc <- ncol(y)
9 #accumulate # of NAs
10 for(i in 1:nc){
11 nanum <- length(which(is.na(y[,i])))
12 if(nanum == nr) next
13 rhcollect(i, list(val=summed[i], len=(nr-nanum)))
14 }
15 })
16
17 bigmeanreduce<-expression(
18 pre={
19 total <- 0
20 cnt <- 0
21 },
22 reduce={
23 total <- total + sum(sapply(reduce.values, function(x) sum(x$val)))
24 cnt <- cnt + sum(sapply(reduce.values, function(x) sum(x$len)))
25 },
26 post={rhcollect(reduce.key, total/cnt)}
27)
28
29 z <- rhmr(map=bigmeanmap, reduce=bigmeanreduce,
30 ifolder="/rhipe/airline/hl_airline.csv",
31 ofolder="/rhipe/airline/out5",
32 inout=c("text", "sequence")
33)
34
35 jobid <- rhex(z, async=TRUE)
```

- RHIFE(R and Hadoop Integrated Processing Environment)는 Purdue Univ.의 통계학 박사과정 학생이었던 Saptarshi Guha에 의해 개발된 R 라이브러리
- R을 Hadoop 환경에서 MapReduce 개념의 분산처리가 가능하게 해 줌
- Amazon의 EC2에서 사용 가능함 (<http://www.stat.purdue.edu/~sguha/rhipe/doc/html/ec2.html> )
- 최근에 RHadoop이라는 Revolution Analytics에서 나온 오픈소스 패키지 출시



Facebook에서의 R+RHIFE에 대한 Guha's lecture  
<http://www.lecturemaker.com/2011/02/rhipe/>

# RHive – Hive



- <http://hive.apache.org>
- **A data warehouse system for Hadoop**
- **Open Source (Apache License)**
- **ANSI SQL Support**
- **Facebook's Main Data Warehousing System**

# RHive



- ◆ Language : R or ANSI-SQL
- ◆ R-Hive Bridge
- ◆ R Export
- ◆ R 기반 분산 처리 Framework

- ◆ 가장 널리 사용하는 Analytic Tool
- ◆ CRAN : 4,000+ Rich R library Set
- ◆ 용이한 Library/Procedure 제작
- ◆ 다양한 Visualization, IDE 도구

- ◆ Hadoop 기반 분산 병렬 처리
- ◆ ANSI SQL : Low Learning Cost
- ◆ 용이한 기능 확장 : UDF, UAF

```
> install.package("RHive")
```



# RHive - Demo

```
1 library(RHive)
2 rhive.init()
3 rhive.connect()
4
5 rhive.hdfs.ls("/")
6
7 rhive.query("SHOW TABLES")|
8 rhive.desc.table("weights")
9 rhive.query("select * from weights limit 10")
10
11 map <- function(k, vs){
12 if(is.null(vs)) {
13 put("NA", 0)
14 }
15 for(n in 1:length(vs)){
16 put(as.character(n),vs[n])
17 }
18 }
19
20 reduce <- function(k,v){
21 put(k,mean(as.double(v)))
22 }
23
24
25 rhive.mrapply("weights", map, reduce,
26 as.character(rhive.desc.table("weights")[,1]),
27 c("rowname", "one"), by="rowname",
28 c("rowname", "one"), c("rowname", "count"))
29
```

- HDFS interface
  - Hive query interface
  - Map/Reduce
- Programming with R

# RHive - RHive Analytics

RHive 위에 구현된 대용량 분산 데이터 마이닝 시스템

## Clustering

K-means

## Prediction

Multi-variate linear regression

Classification tree

## Sampling

random, stratified, cluster, quota, sampling

## Modeling

model parameter tuning

feature selection

# Data Scientist's way to solve real world problem

Raw포맷은 다양하며,  
이들을 효과적으로  
처리할 수 있어야 한다.

Hive는 분석 인원이  
최적으로 운영할 수  
있는 정도의 컴퓨팅  
리소스를 가져야 한다.

최대한 많은 양의  
메모리를 확보한다.

preprocessing  
for input

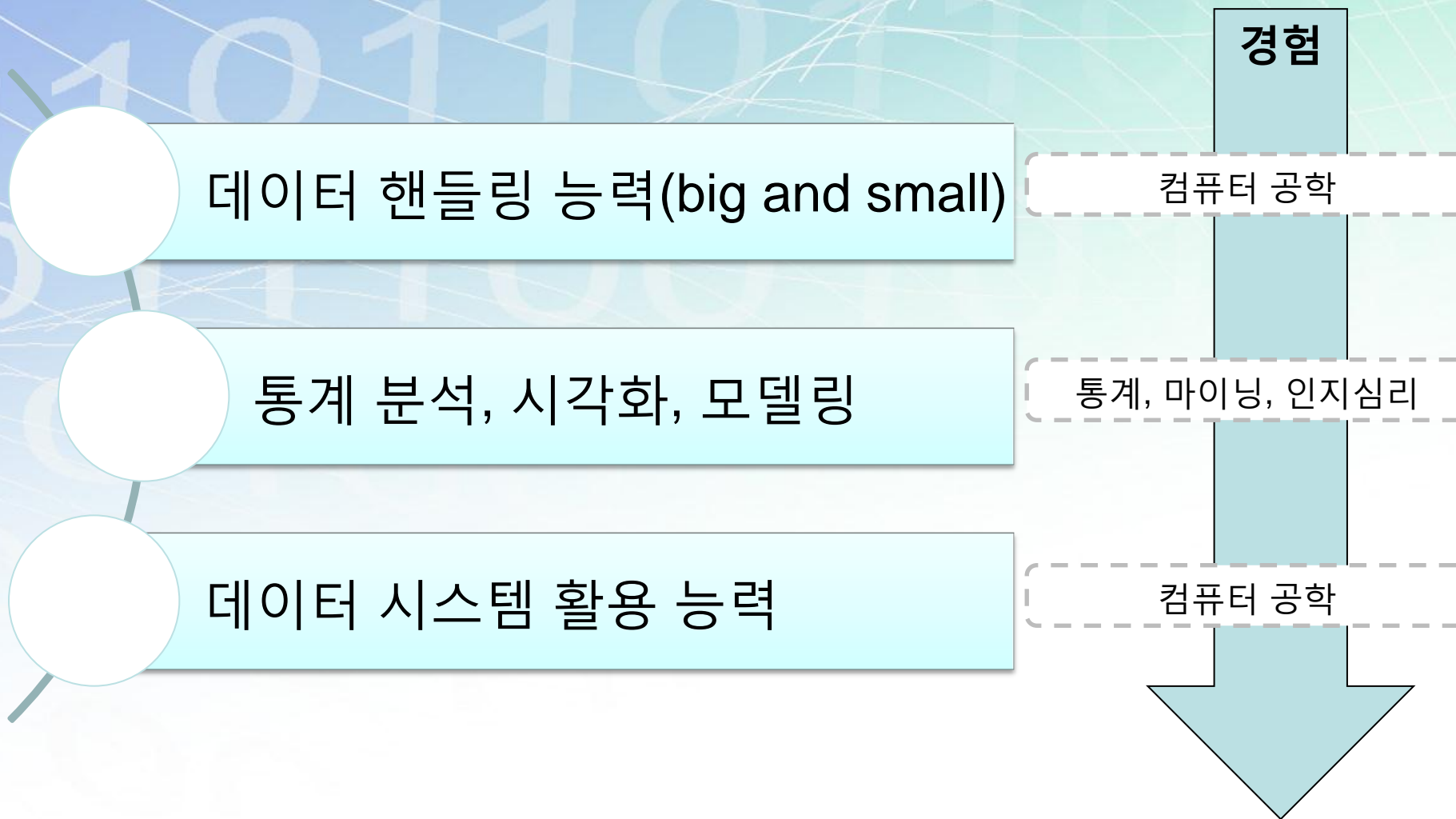
ETL on Hive

R analysis


Hive에 입력에 맞는  
포맷으로 데이터 처리  
(using Python or Perl)

Aggregate, Filtering 작업  
혹은 샘플링

# 데이터과학자로서 요구되는 기술



# 데이터과학자로서 요구되는 자질



창의력 : 분석 스토리를 만드는 능력

적극성 : 데이터 그리고 자신에 대한 믿음을 바탕으로...

커뮤니케이션 또는 프리젠테이션 능력



# R 데이터 분석가가 되기 위해서는?

## R언어 이해

- 학습
- 경험 혹은 연습

## 통계/마이닝 능력

- 통계학
- 데이터 마이닝, 기계학습

## 시각화

- 가르쳐 주는 곳 없음
- 책기반으로 독학/실습 혹은 인터넷 참고

## 경험

- 오픈 데이터를 이용한 분석 실습/해석/호기심 필수
- 데이터 마이닝 대회를 통한 노하우 습득

마지막 한 꼭지!

**빅 데이터가 정말 도움이 되는가?**

# 어떤 그래프이길 원하시나요?

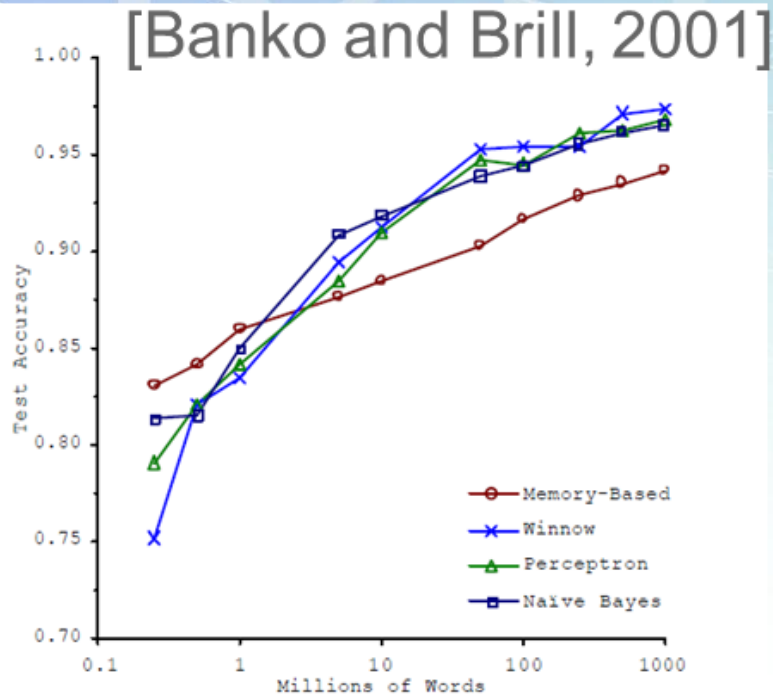
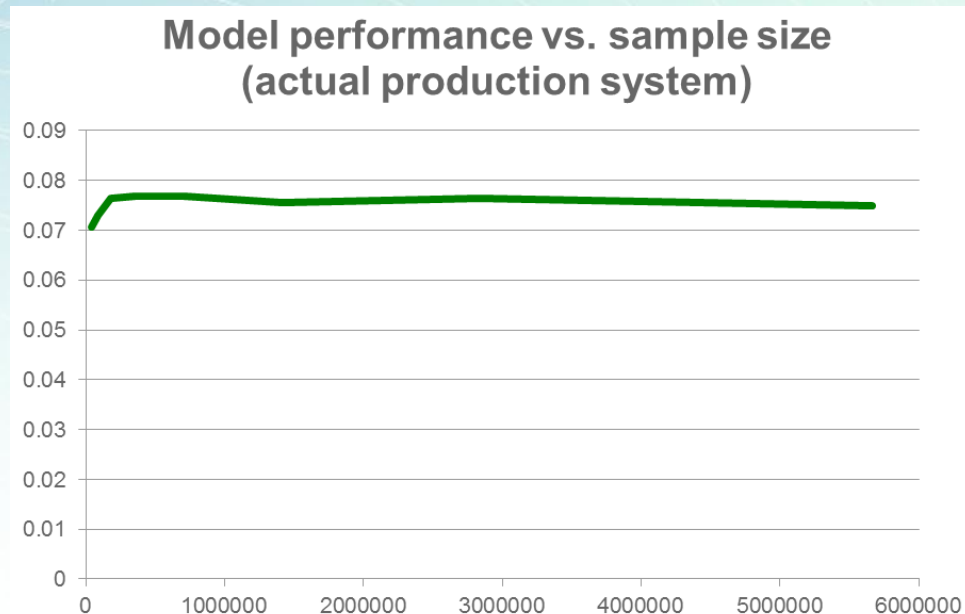


Figure 1. Learning Curves for Confusion Set Disambiguation



# 빅 데이터가 항상 도움이 되는 건 아니다!

빅 데이터 붐을 초래한 ....

“We don’t have better algorithms. We just have more data.” –Peter Norvig--

왜 그런가?

많은 예측변수는 많은 데이터를 필요로 한다. 변수가 적다면? 혹은 쓸모없는 변수를 넣는다면?

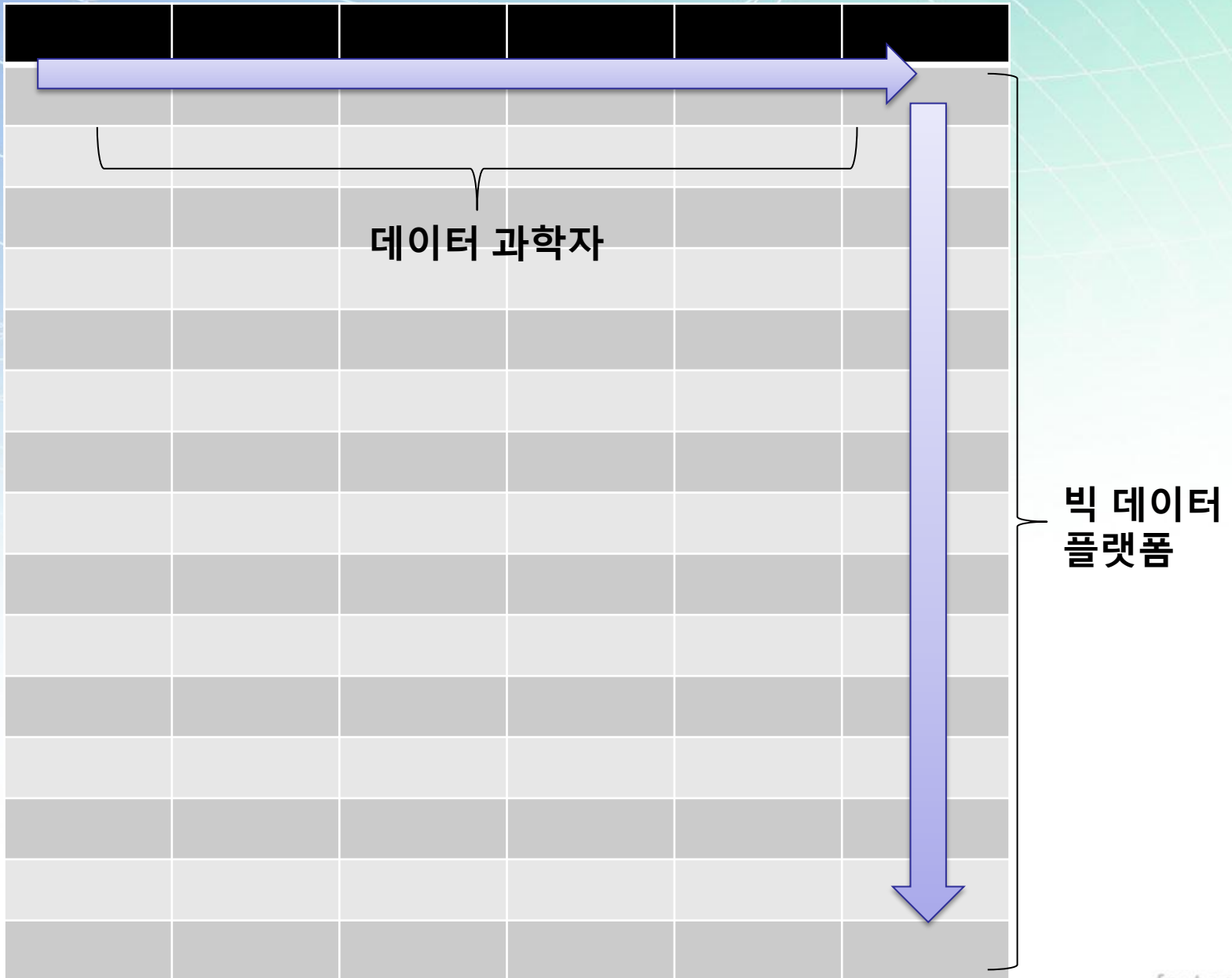
결론

큰 데이터에 적합한 접근방법을 사용하지 않는다면 그 데이터는 쓰레기밖에 되지 못한다.

**빅 데이터를 확인하고 접근 방법을 결정하는 데이터과학자 혹은 분석가의 역할이 무엇보다 중요하다.**



# 플랫폼이 대체할 수 없는 데이터과학자



# 빅 데이터 분석에 있어 데이터과학자의 요구사항

**빠르게** 눈으로 직접 확인해야 될 것들이 많아졌다.

**빠르게** 다양한 포맷의 데이터를 병합하고 쪼개보고 꼬아봐야 된다.

**빠르게** 최신의 알고리즘부터 오래된 알고리즘까지 적용 가능한지 시도해야 된다.



# Q & A



**madjakarta@gmail.com**

**<http://freesearch.pe.kr>**