R 전문가로 가는 길

-- 빅 데이터 활용 바로 보기 --

Heewon Jeon (NexR Corp.)

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

fppt.com

Interactive Data Analysis

레거시 데이터 분석

- 컴퓨팅 리소스가 굉장히 비쌌다.
- 많은 입력 값
- 많은 출력 값
- 부담없이 여러 번 수행하기 힘듦
- 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

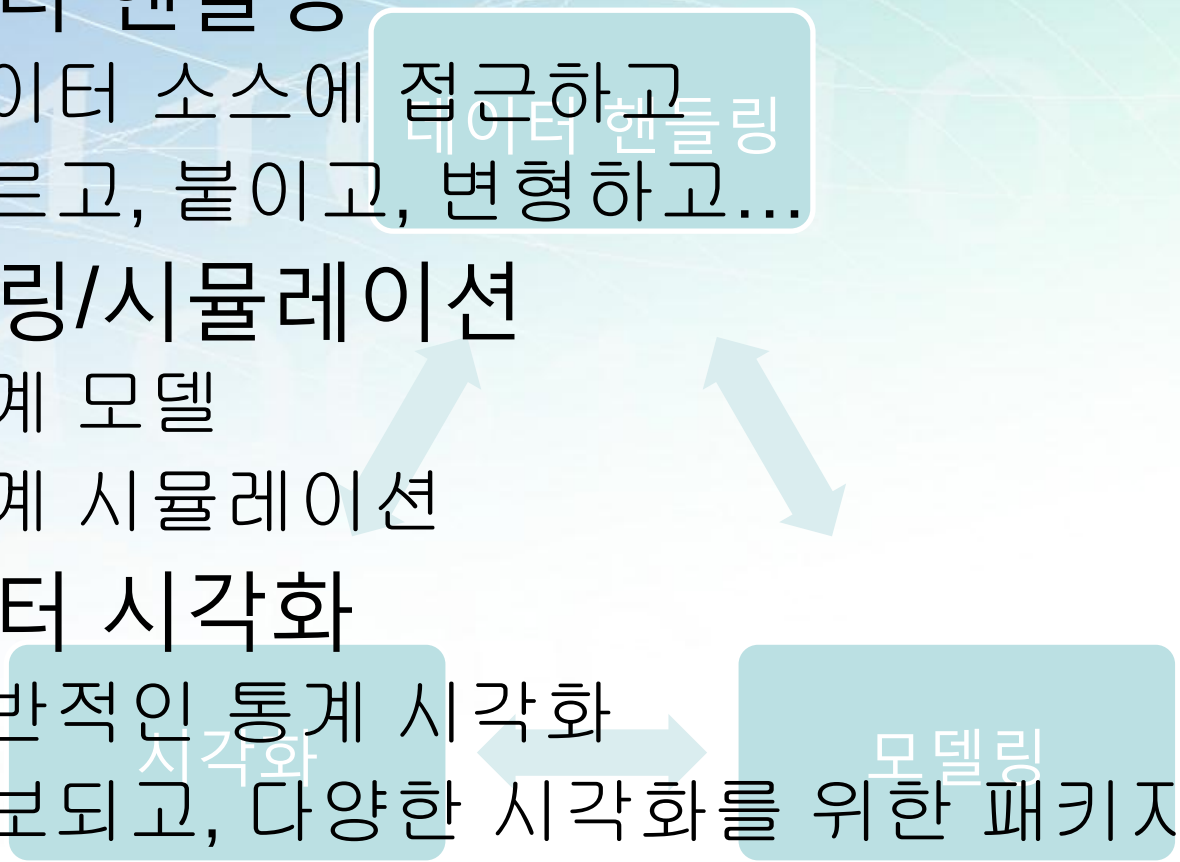
fppt.com

일반화된 데이터 분석 과정

fppt.com

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



fppt.com

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 환경이다.

http://robjhyndman.com/researchtips/rvsexcel/

fppt.com

Really R?

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

fppt.com

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 ...

fppt.com

R has simple GUI

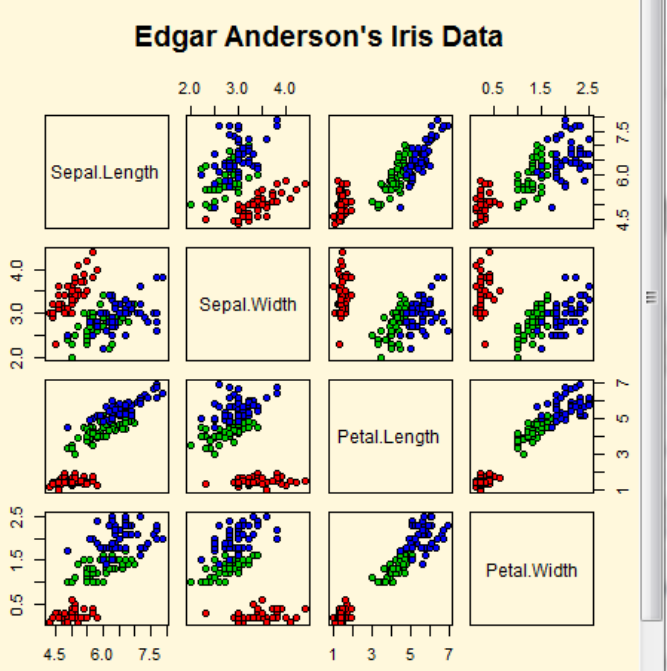
RGui (64-bit)

다음

R Console

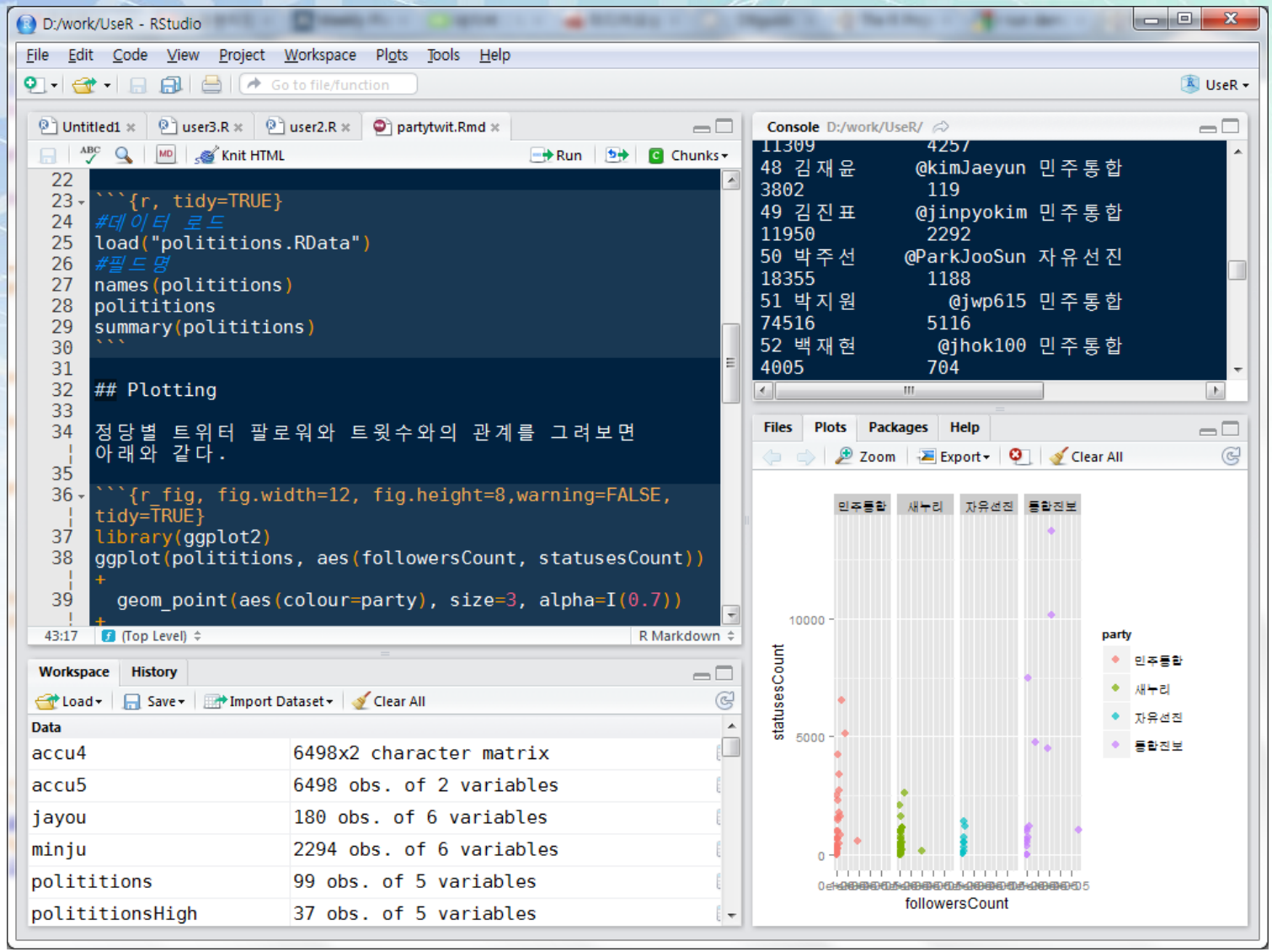
```
> par(bg = "lightcyan")

> pin <- par("pin")

> xdelta <- diff(range(x))

> ydelta <- diff(range(y))

> xscale <- pin[1]/xdelta

> yscale <- pin[2]/ydelta

> scale <- min(xscale, yscale)

> xadd <- 0.5*(pin[1]/scale - xdelta)

> yadd <- 0.5*(pin[2]/scale - ydelta)

> plot(numeric(0), numeric(0),
+      xlim = range(x)+c(-1,1)*xadd, ylim = range(y)+c(-1,1)*yadd,
+      type = "n", ann = FALSE)
페이지 변경을 확인하기 위해 기다리는 중...
```

클릭 또는 ENTER 키를 누르면 다음 페이지로 이동합니다



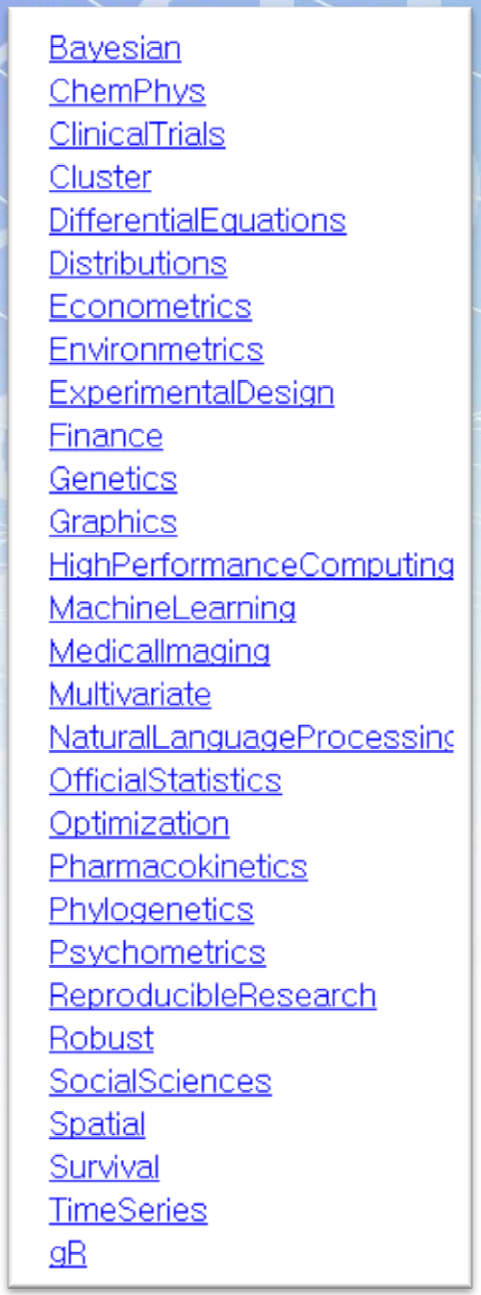
fppt.com

RStudio is better



fppt.com

R Package System



Total 3,921 Packages

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

fppt.com

If you want to do twitter analysis.

- Data Source
  - **twitteR**
- 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

fppt.com

R Packaging System 사용 예 - 1

C:/work/KoNLP - RStudio

File Edit View Project Workspace Plots Tools Help

test.R*

```
library(twitteR)
library(KoNLP)
library(wordcloud)
library(RColorBrewer)

gogamza <- getUser("gaiaville")
gogamza.followers <- gogamza$getFollowers()

#팔로워들의 자기 소개를 벡터에 적재 한다.
followerDesc <- c()

for(i in gogamza.followers){
  followerDesc <- append(followerDesc, i$description)
}

#쓸모없는 문자들을 제거한다.
followerDesc <- gsub("\n","", followerDesc)
followerDesc <- gsub("\r", "", followerDesc)

nouns <- Map(extractNoun, followerDesc)

wordsvec <- unlist(nouns, use.name=F)
#쓸모없는 문자들을 제거한다. 특히 영문자의 경우 tm의 stopwords를 활용한다.
wordsvec <- wordsvec[-which(wordsvec %in% stopwords("english"))]
wordsvec <- gsub("[[:punct:]]","", wordsvec)
wordsvec <- Filter(function(x){nchar(x)>=2}, wordsvec)

wordcount <- table(wordsvec)
pal <- brewer.pal(8,"Dark2")

wordcloud(names(wordcount),freq=wordcount,scale=c(3,0.5),min.freq=10,
          random.order=T,rot.per=.1,colors=pal)
```

Workspace History

Values	
followerDesc	character[762]
gogamza	user[1]
gogamza.followers	list[802]
i	user[1]
nouns	list[762]
pal	character[8]
wordcount	table[4046]

Files Plots Packages Help

Console C:/work/KoNLP/

```
> wordcloud(names(wordcount),freq=wordcount,scale=c(3,0.5),min.freq=10,
+           random.order=T,rot.per=.1,colors=pal)
>
>
```

fppt.com

R Packaging System 사용 예 - 2

File Edit View Project Workspace Plots Tools Help

gogamza | Sign Out

multi-culture

analysis.R* Untitled1* Untitled2*

Source on Save Run Source

```
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
```

24:1 (Top Level) R Script

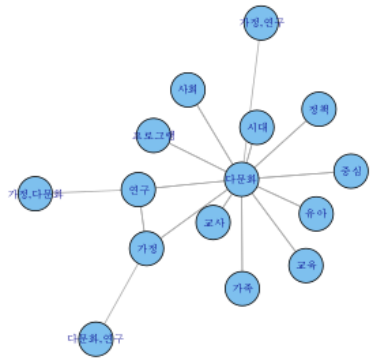
Workspace History

Load Save Import Dataset Clear All

Data	
edgelist	18086x2 character matrix
f2	71 obs. of 2 variables
rulemat	27x2 character matrix
rulequality	43 obs. of 3 variables
wordtab	1547x1547 integer matrix
Values	
KtoS	character[71]

Files Plots Packages Help

Zoom Export Clear All

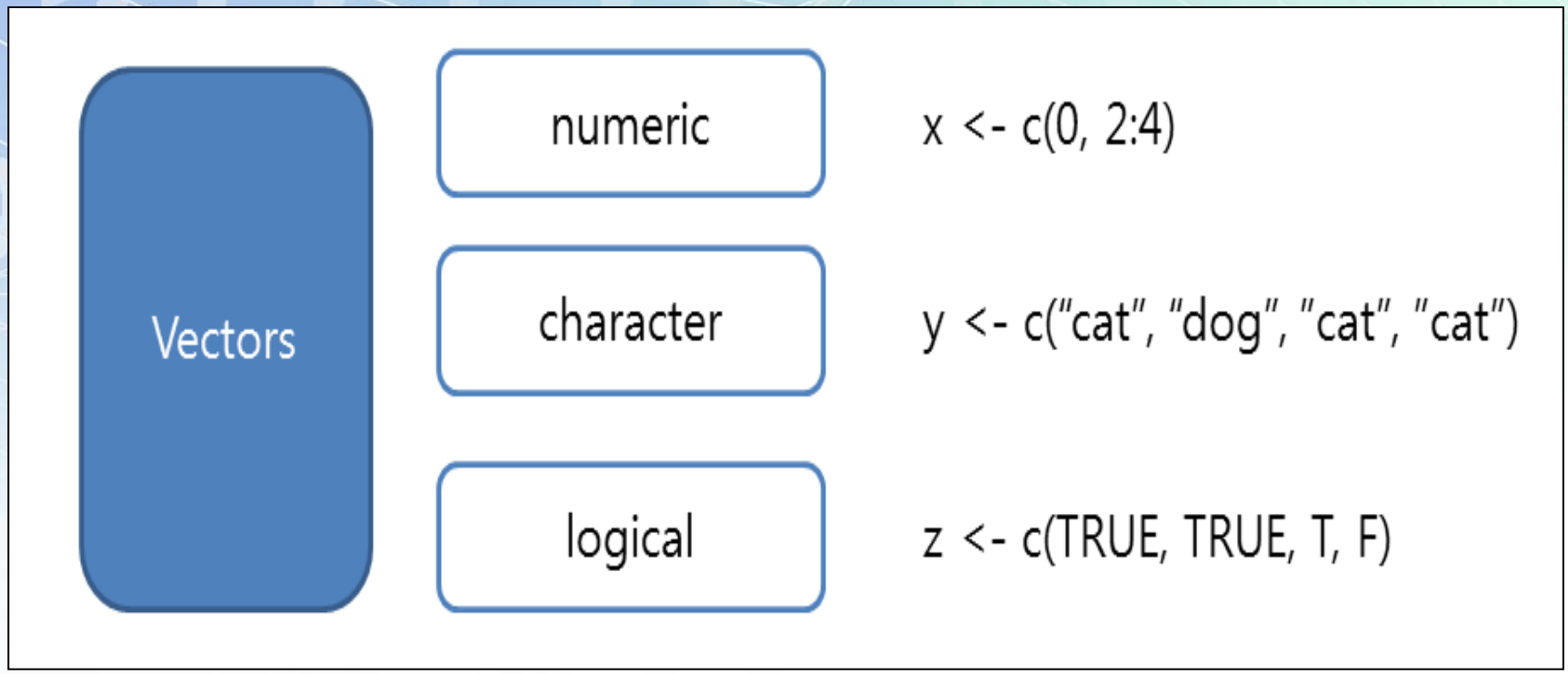


Console ~/work/multi-culture/

```
> 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)
>
```

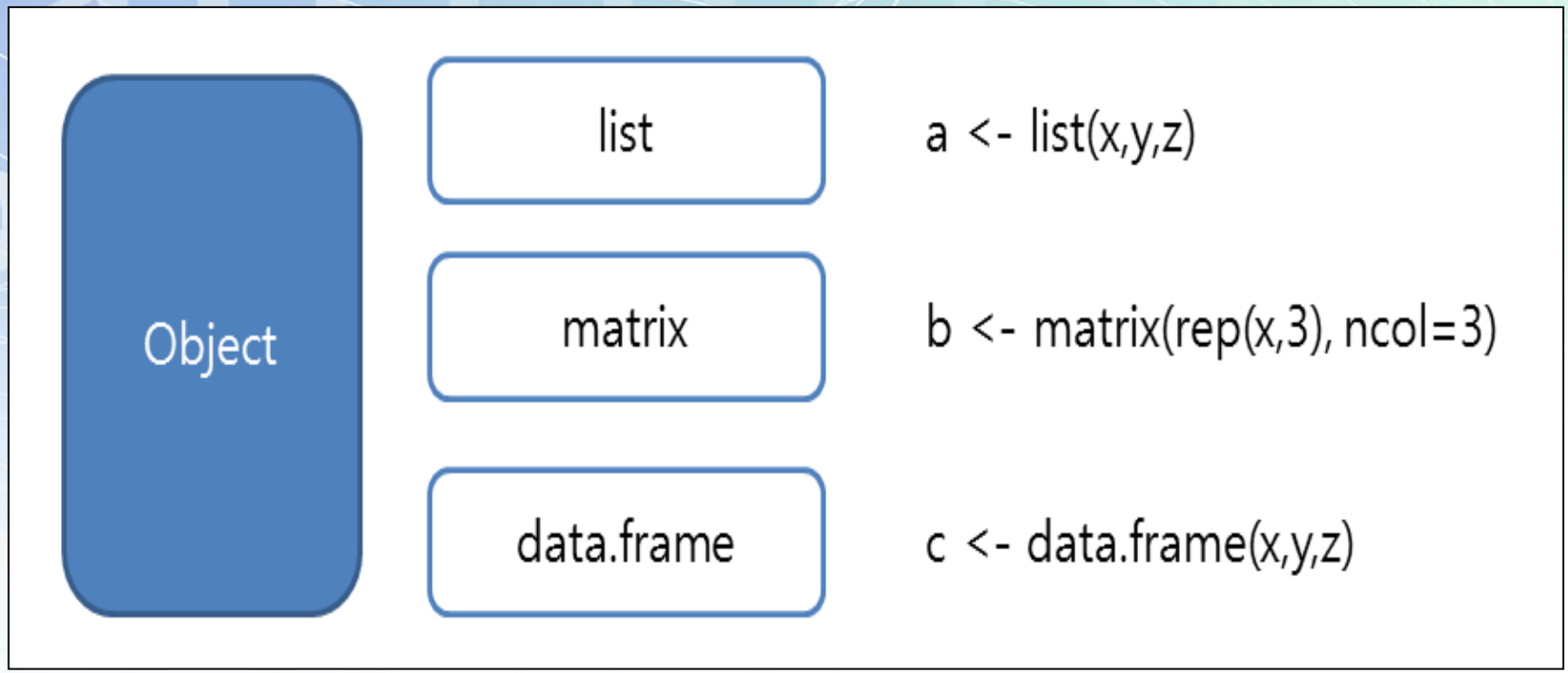
fppt.com

R data structures for DBA - 1



fppt.com

R data structures for DBA - 2



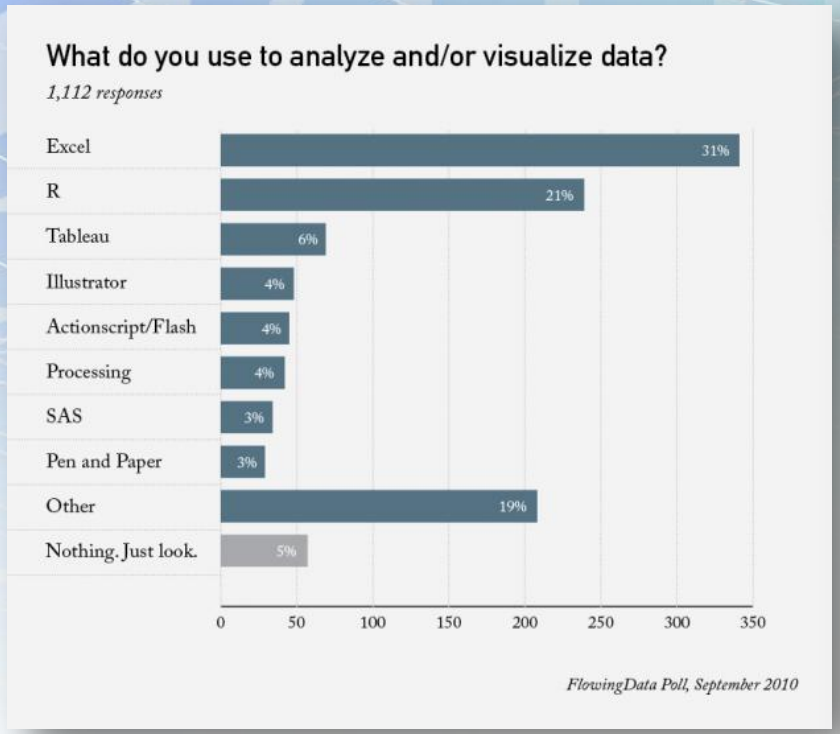
fppt.com

R data structures for DBA - 3

```r
library(sqldf)

iris3 <- sqldf("select Sepal_Length, Sepal_Width, Species from iris")


head(iris3)
sqldf("select * from iris3 limit 6")
# Sepal_Length Sepal_Width Species
# 1          5.1         3.5  setosa
# 2          4.9         3.0  setosa
# 3          4.7         3.2  setosa
# 4          4.6         3.1  setosa
# 5          5.0         3.6  setosa
# 6          5.4         3.9  setosa


subset(iris3, Sepal_Length > 5)
sqldf("select * from iris3 where Sepal_Length > 5")


with(iris3,
     aggregate(list(Sepal_Length,Sepal_Width), by=list(Species), mean)
     )
sqldf("select Species, avg(Sepal_Length), avg(Sepal_Width)
      from iris3 group by Species")


head(iris3[order(iris3$Sepal_Length, decreasing = TRUE), ], 3)
sqldf("select * from iris3 order by Sepal_Length desc limit 3")

```

fppt.com

Popularity of R



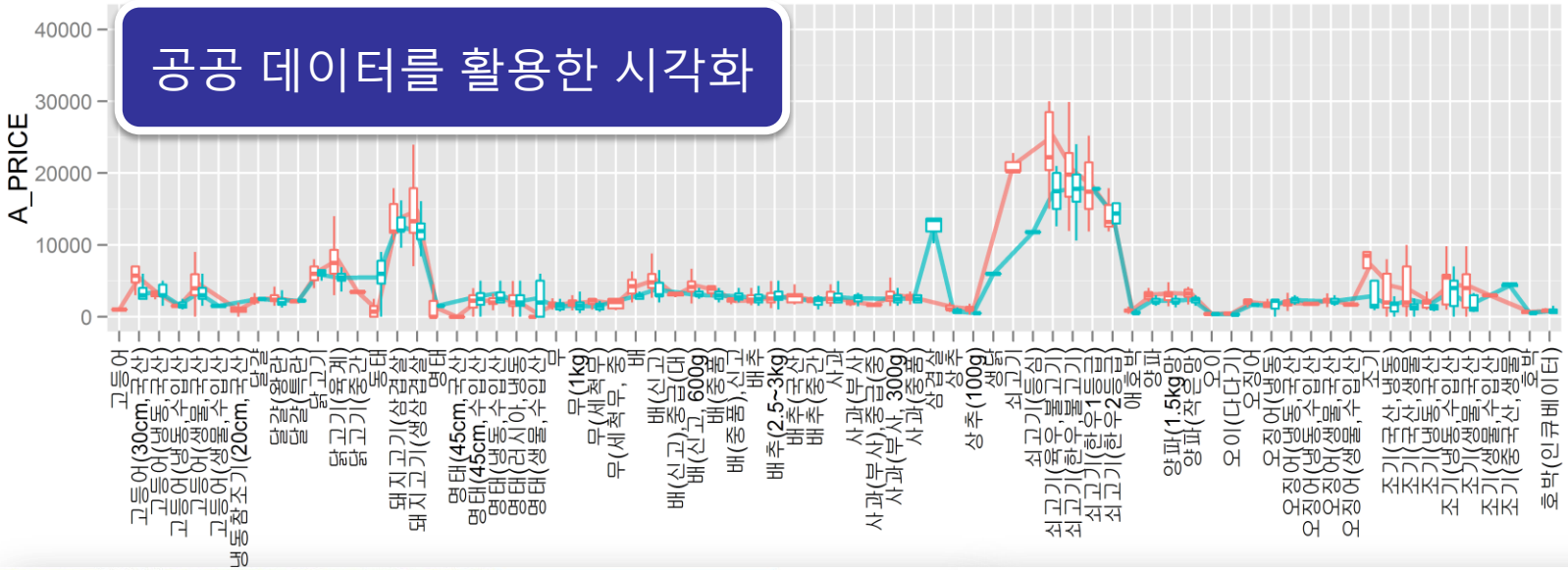
http://flowingdata.com/2010/09/28/poll-results-what-do-you-use-to-analyze-andor-visualize-data/

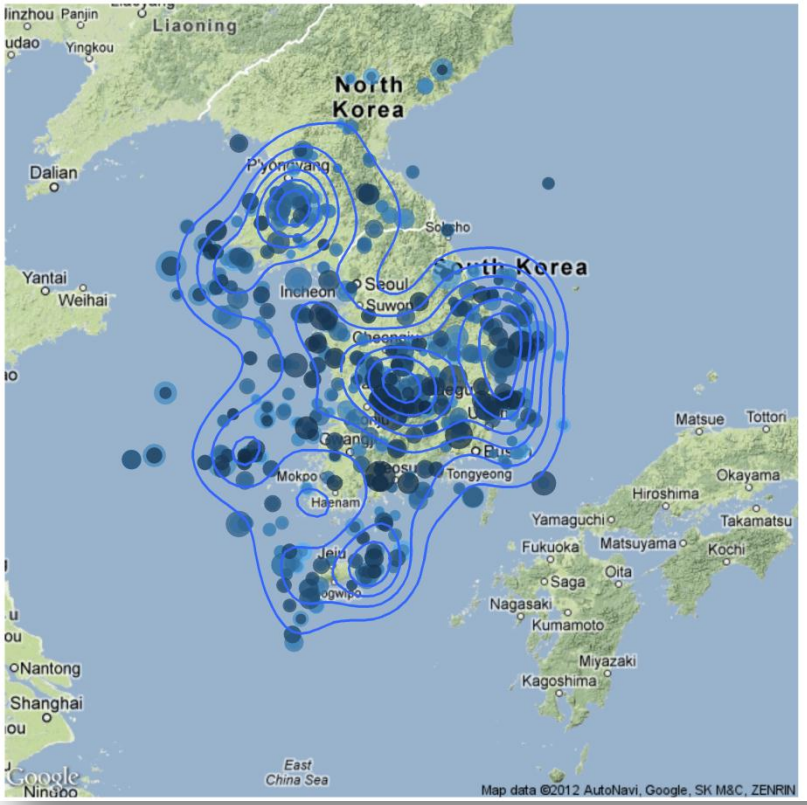
http://www.kdnuggets.com/polls/2012/analytics-data-mining-big-data-software.html

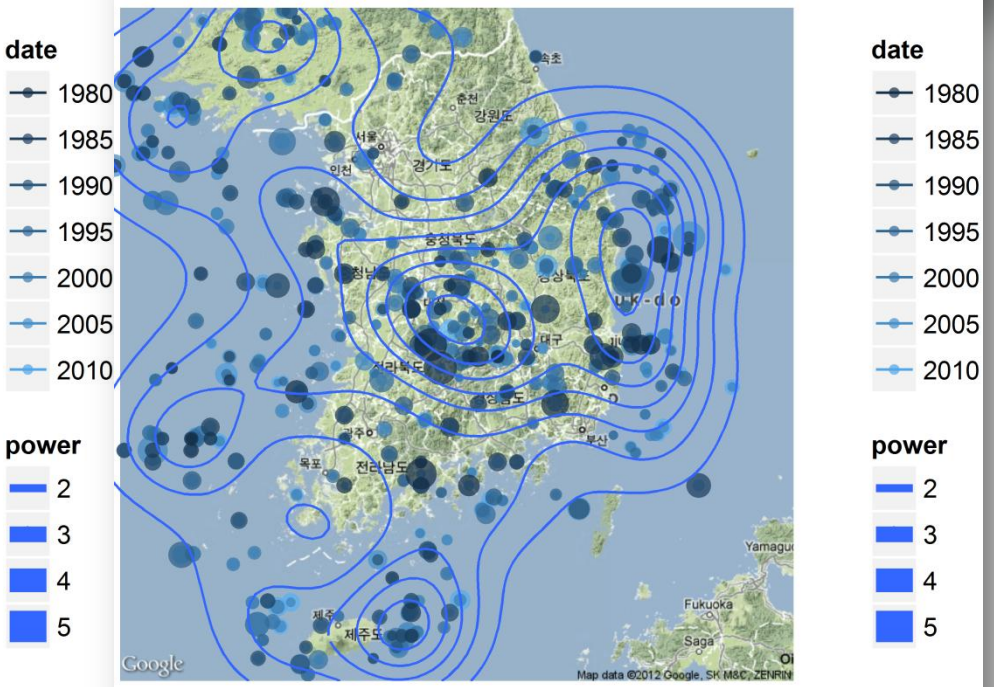
Your own code you used for analytics/data mining in the past 12 months in:	
R (245)	30.7%
SQL (185)	23.2%
Java (138)	17.3%
Python (119)	14.9%
C/C++ (66)	8.3%
Other languages (57)	7.1%
Perl (37)	4.6%
Awk/Gawk/Shell (31)	3.9%
F# (5)	0.6%

What Analytics, Data mining, Big Data software you used in the past 12 months for a real project (not just evaluation) [798 voters]	
Legend: Free/Open Source tools Commercial tools	% users in 2012 % users in 2011
R (245)	30.7% 23.3%
Excel (238)	29.8% 21.8%
Rapid-I RapidMiner (213)	26.7% 27.7%
KNIME (174)	21.8% 12.1%
Weka / Pentaho (118)	14.8% 11.8%
StatSoft Statistica (112)	14.0% 8.5%
SAS (101)	12.7% 13.6%
Rapid-I RapidAnalytics (83)	10.4% not asked in 2011
MATLAB (80)	10.0% 7.2%

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 데이터는 여전히 처리하기 힘듦

fppt.com

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

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

fppt.com

RHipe

```
bigmeanmap <- expression({
  y <- do.call("rbind", lapply(map.values, function(r){
        as.numeric(strsplit(r,",")[[1]])
      }
  ))
  summed<- colSums(y, na.rm=T)
  nr <- nrow(y)
  nc <- ncol(y)
  #accumulate # of NAs
  for(i in 1:nc){
    nanum <- length(which(is.na(y[,i])))
    if(nanum == nr) next
    rhcollect(i, list(val=summed[i], len=(nr-nanum)))
  }
})

bigmeanreduce<-expression(
 pre={
   total <- 0
   cnt <- 0
   },
 reduce={
   total <- total + sum(sapply(reduce.values, function(x) sum(x$val)))
   cnt <- cnt + sum(sapply(reduce.values, function(x) sum(x$len)))
   },
 post={rhcollect(reduce.key,total/cnt)}
)

z <- rhmr(map=bigmeanmap, reduce=bigmeanreduce,
   ifolder="/rhipe/airline/hl_airline.csv",
   ofolder="/rhipe/airline/out5",
   inout=c("text", "sequence")
)

jobid <- rhex(z, async=TRUE)
```

- RHIPE(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+RHIPE에 대한 Guha's lecture
http://www.lecturemaker.com/2011/02/rhipe/

fppt.com

RHive - Hive

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

fppt.com

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 Leaning Cost
- 용이한 기능 확장 : UDF, UAF

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

fppt.com

RHive - Demo

```
library(RHive)
rhive.init()
rhive.connect()

rhive.hdfs.ls("/")

rhive.query("SHOW TABLES")
rhive.desc.table("weights")
rhive.query("select * from weights limit 10")

map <- function(k, vs){
  if(is.null(vs)) {
    put("NA", 0)
  }
  for(n in 1:length(vs)){
    put(as.character(n),vs[n])
  }
}

reduce <- function(k,v){
  put(k,mean(as.double(v)))
}


rhive.mrapply("weights", map, reduce,
              as.character(rhive.desc.table("weights")[,1]),
              c("rowname","one"), by="rowname",
              c("rowname","one"), c("rowname","count"))
```

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

fppt.com

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

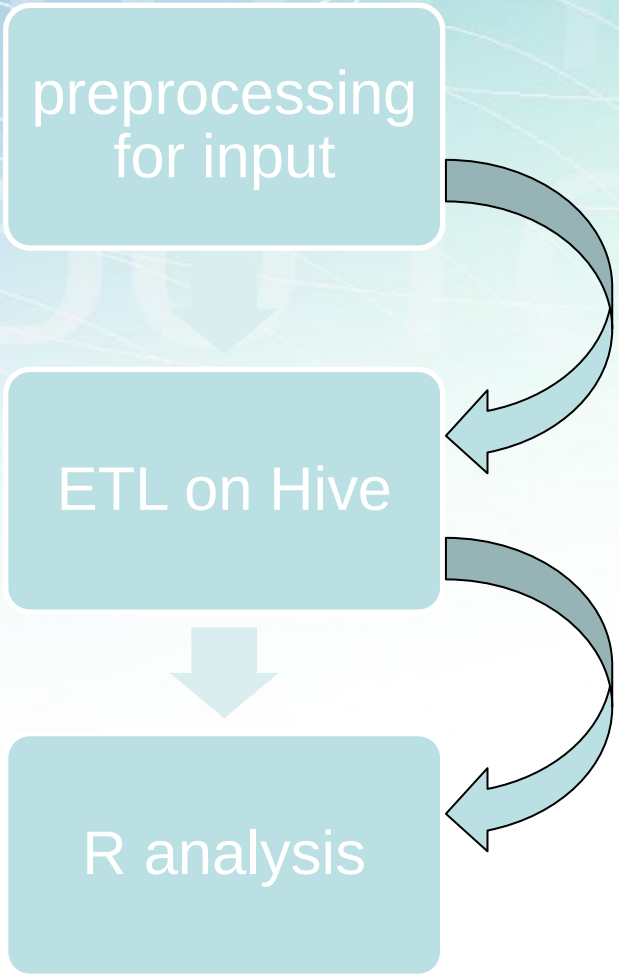
fppt.com

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.



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

Aggregate, Filtering 작업 혹은 샘플링

fppt.com

데이터과학자로서 요구되는 기술

- 데이터 핸들링 능력(big and small)
- 통계 분석, 시각화, 모델링
- 데이터 시스템 활용 능력

fppt.com

데이터과학자로서 요구되는 자질

- 창의력 : 분석 스토리를 만드는 능력
- 적극성 : 데이터 그리고 자신에 대한 믿음을 바탕으로...
- 커뮤니케이션 또는 프리젠테이션 능력

fppt.com

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

fppt.com

마지막 한 꼭지!

빅 데이터가 정말 도움이 되는가?

fppt.com

어떤 그래프이길 원하시나요?

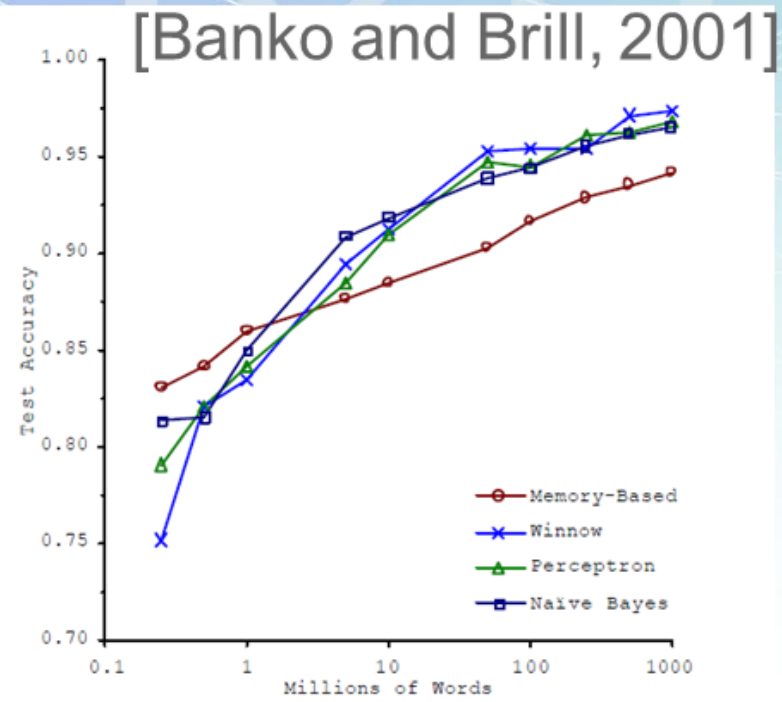
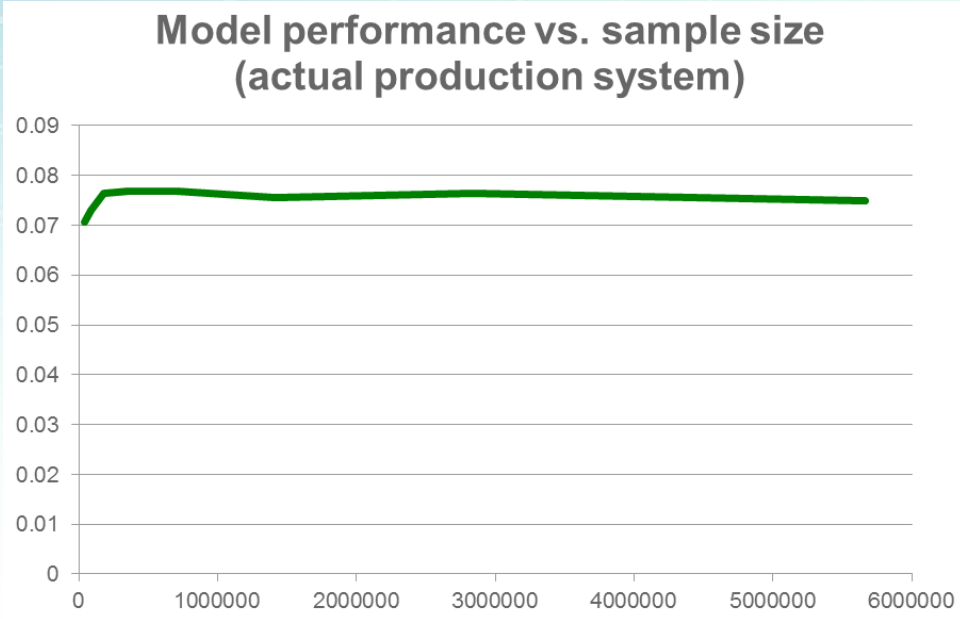


Figure 1. Learning Curves for Confusion Set Disambiguation



fppt.com

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?

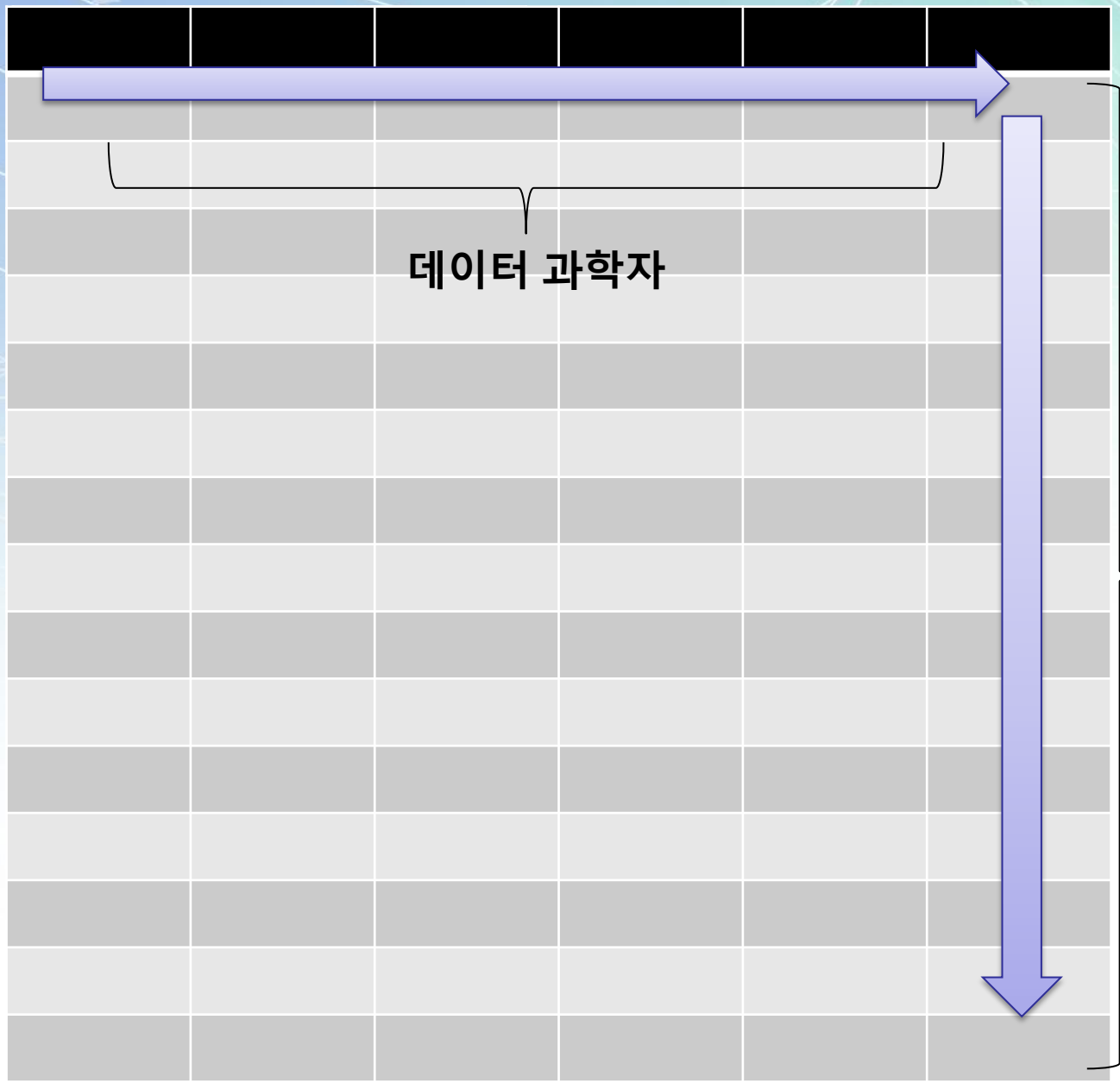
결론

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

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

fppt.com

플랫폼이 대체할 수 없는 데이터과학자



fppt.com

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.

fppt.com

Q & A

madjakarta@gmail.com

http://freesearch.pe.kr

fppt.com