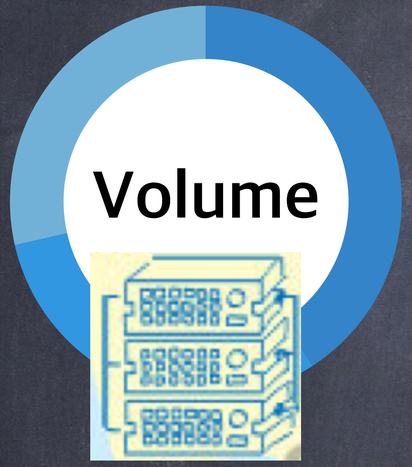
Big Data Near You

2016. 06. 16

Prof. Sehyug Kwon

Dept. of Statistics

4V's of Big Data

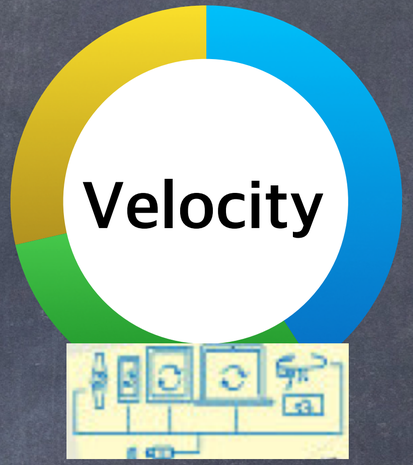


대용량

- tera(1,000^4) - peta -exa- zetta(1000^7) bytes in 2020
- 통화내역, 카드사용내역
- 회사, 정부: DB-DW-Data mart

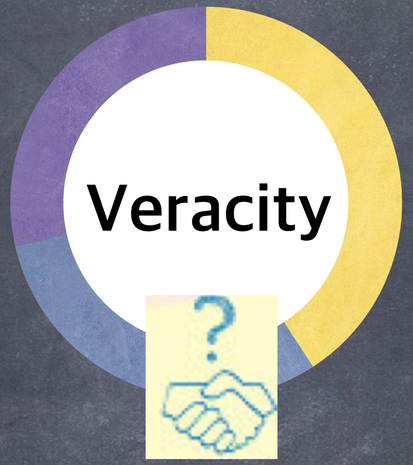
다양한 유형

- 텍스트 마이닝
- 이미지, 멀티미디어
- (비)정형화 포멧



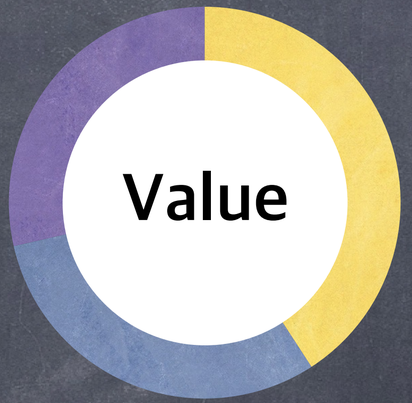
실시간 정보

- Streaming data
- 실시간 분석결과



(불)확실성

- 데이터 이력
- 불확실성=비용
- 1/3 CEO 의사결정 불신



가치

- 딥러닝, 학습효과
- 비즈니즈 정보

Three Experts in Big Data

- Hadoop (분산파일처리)
- MapReduce (분산프로그래밍모델)
  - Java / Python / Ruby
  - NoSQL, DB
  - Apache Spark

개발자

관리자

- 하둡, 리눅스 관리
  - Cluster Management
  - Cluster Performance
  - Virtualization

- 데이터 과학 - 모델링
- 기계학습, 마이닝기법
- 빅데이터 벤더 : R/SAS
- 데이터 Visualization

데이터 분석가

Leaders in Big Data

Deep Learning Tools

New this year was a category of Deep Learnir

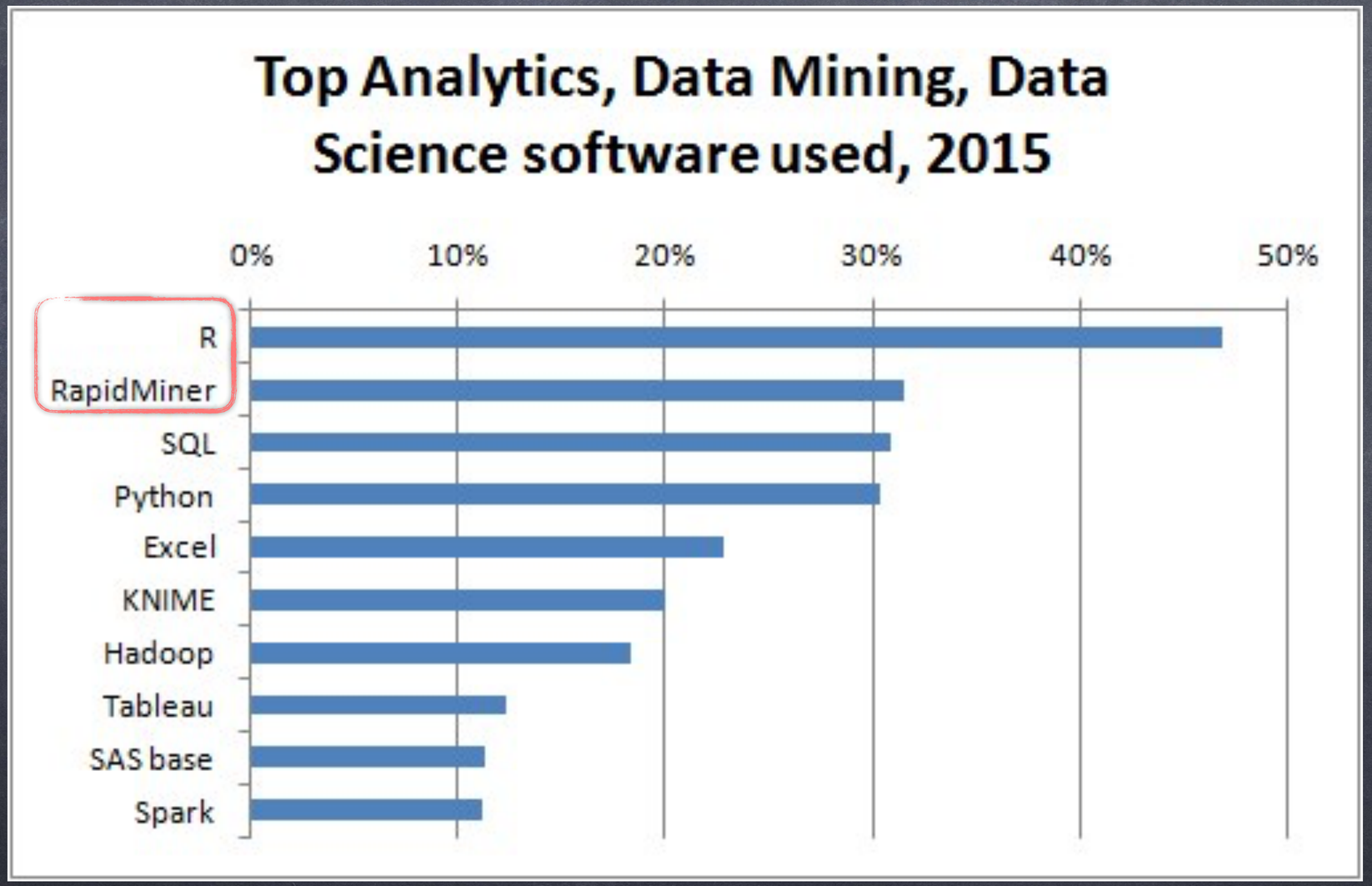
- Pylearn2 (55 users)
- Theano (50)
- Caffe (29)
- Cuda-convnet (17)
- Deeplearning4j (12)
- Torch (27)

Top Hadoop/Big Data tools were

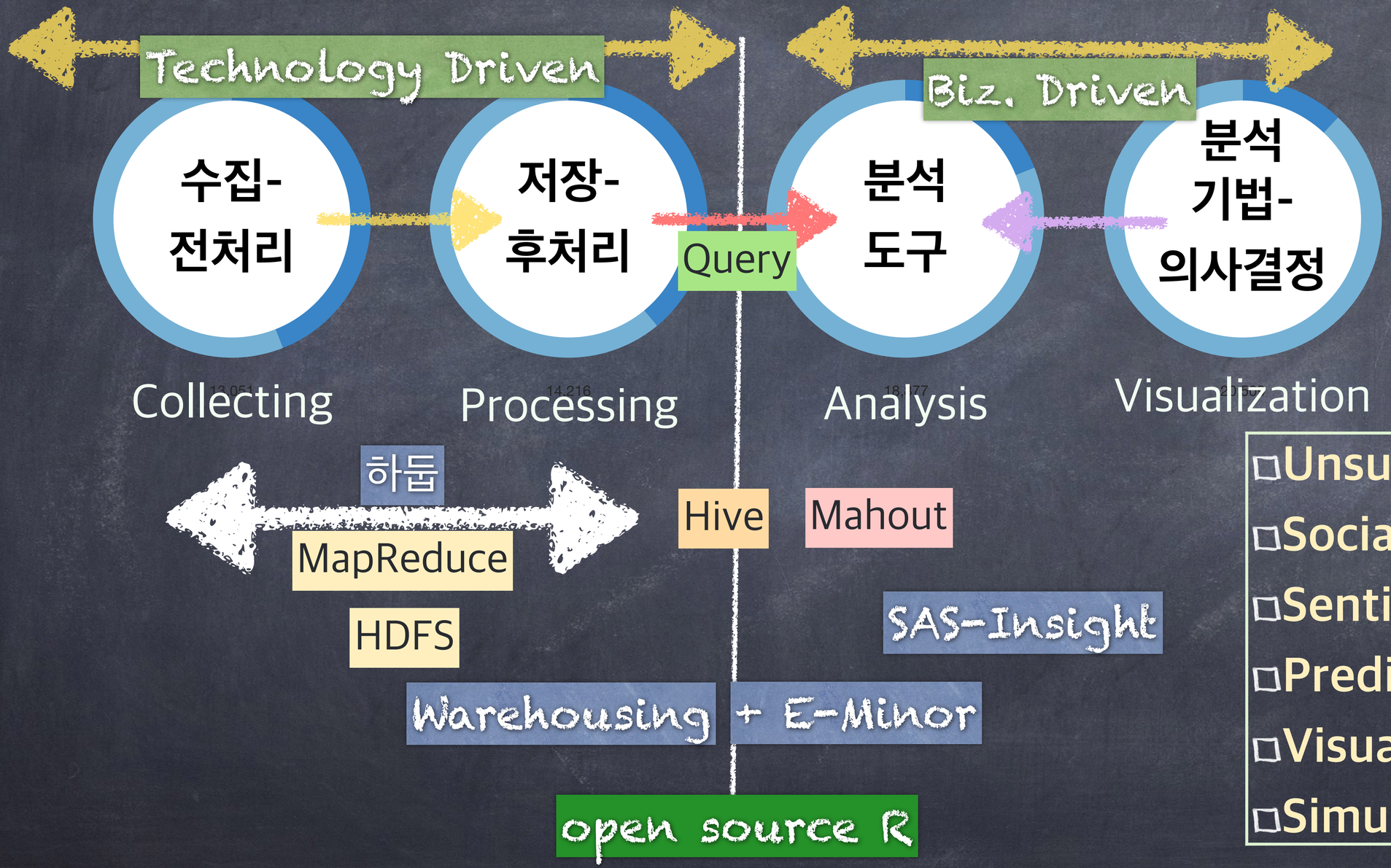
- Hadoop, 18.4% share (507 votes)
- Spark, 11.3% (311)
- Hive, 10.2% (282)
- SQL on Hadoop tools, 7.2% (198)
- Pig, 5.4% (150)
- HBase, 4.6% (127)
- Other Hadoop/HDFS-based tools, 4.5% (125)
- MLlib, 3.3% (91)
- Mahout, 2.8% (76)
- Datameer, 0.8% (23)

http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html

Leaders in Big Data (cont.)

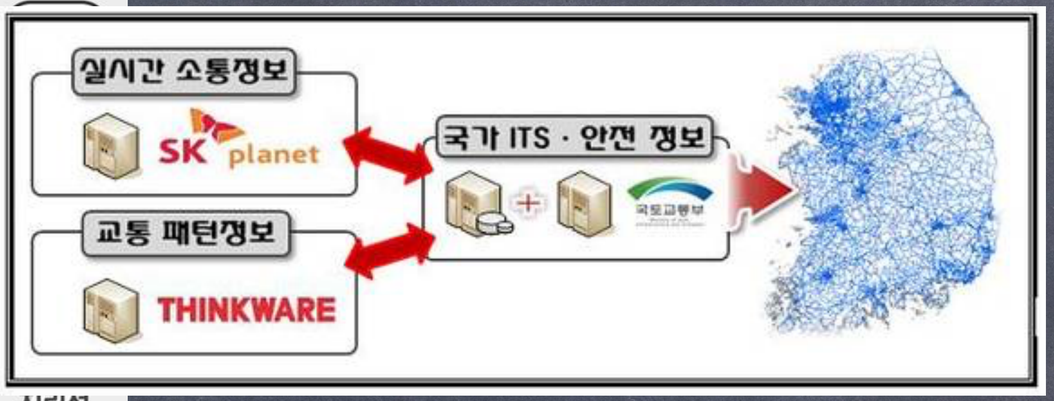


Big Data Flow



- Unsupervised Learning
- Social Media analytics
- Sentiment analysis
- Predictive modeling
- Visualization
- Simulation

Case in Big Data



yelp

Kakao + = KakaoTaxi

[그림] 허츠 빅데이터 적용 프로세스

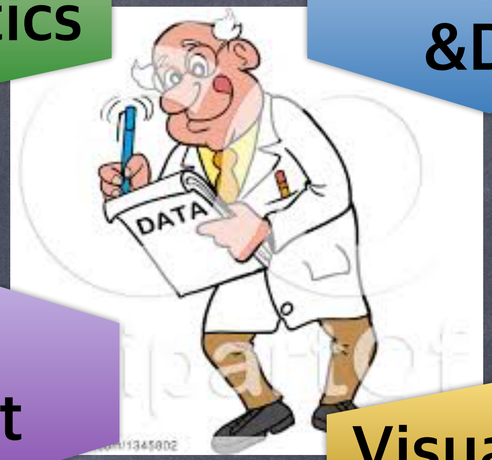
수집	분석	적용
전세계 8,300개 지점의 웹 서베이, 이메일, 콜센터, 문자메시지, 온라인 VOC 등 방대한 고객 의견 데이터 수집	데이터로부터 '지점-서비스-고객불만'의 연관성 파악 → VOC 고객불만 키워드 자동 추출	전사적 서비스 및 지점별 불만 요소 추출 및 개선에 적용 → 운영 수익률 증가 → 고객 만족도 향상

Does this Data Scientist exist in the world?

- 통계모델링
- 기계학습
- 최적화
- 지도학습 : 의사결정나무, 판별분석
- 자율학습 : 군집분석. 차원축소

Statistics

Program &DB



- 프로그래밍언어 (Python)
- 통계소프트웨어(R)
- DB SQL/NoSQL
- MapReduce
- Hadoop-Hive/ Pig

Field Expert

- 경험과 충분한 지식
- 비즈니스 열정
- 데이터 관심
- 문제 해결능력
- 전략, 창의, 협력

Visualization

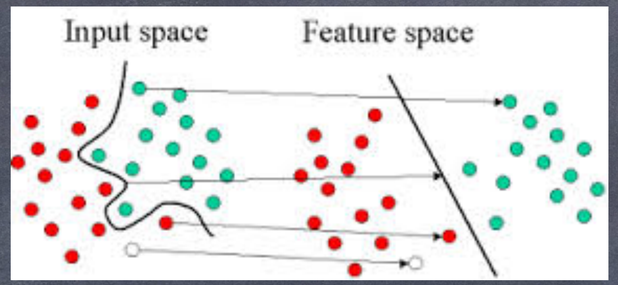
- 비즈니스 보고서
- 스토리텔링
- 데이터 기반 인사이트 -> 의사결정
- 비쥬얼레포팅
- 시각화 : Flare, Taleau
- 그래픽툴 : R (ggplot, lattice)

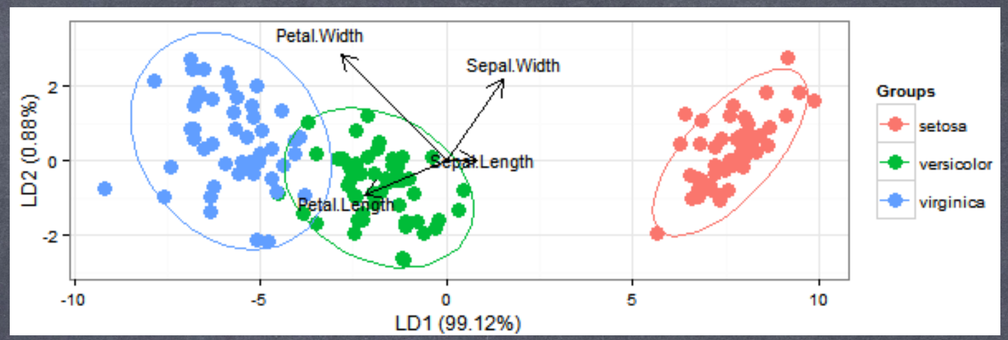
Supervised Learning 기계학습

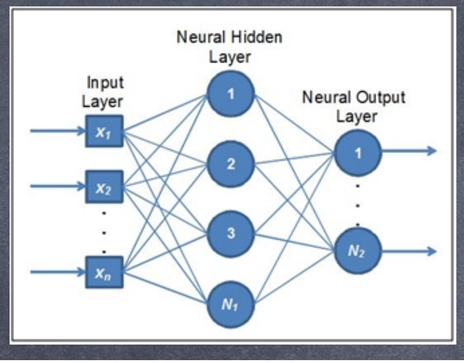
- 기계학습 기법에서는 최적 모형의 탐색 및 선정은 통계분석가에 의해 진행 - 예측 및 판별이 주 목적
- 훈련 Trained(목표변수(target)=함수(예측변수 predictors) 관계를 도출)-검증 Validation(모형의 타당성을 검증)-평가 Test(서로 다른 통계적 방법들의 평가
- 인공지능 (컴퓨터 학습할 수 있도록 하는 알고리즘과 기술) 반복 학습 알고리즘 이용한 데이터 인사이트 탐색 - (예) 스팸 메일 검증여부
- Deep Learning: (Unsupervised Learning)
  - 기계 학습의 한 영역으로, 특히 음성/텍스트/이미지 인식 분야에서 획기적인 발전을 거듭하며 급성장하고 있음,
  - 기본 원리 : 다수의 히든 레이어를 갖춘 신경망을 통해 컴퓨터가 태스크를 학습하고 정보를 체계화하여 스스로 패턴을 찾아낼 수 있게 하는 것입니다.

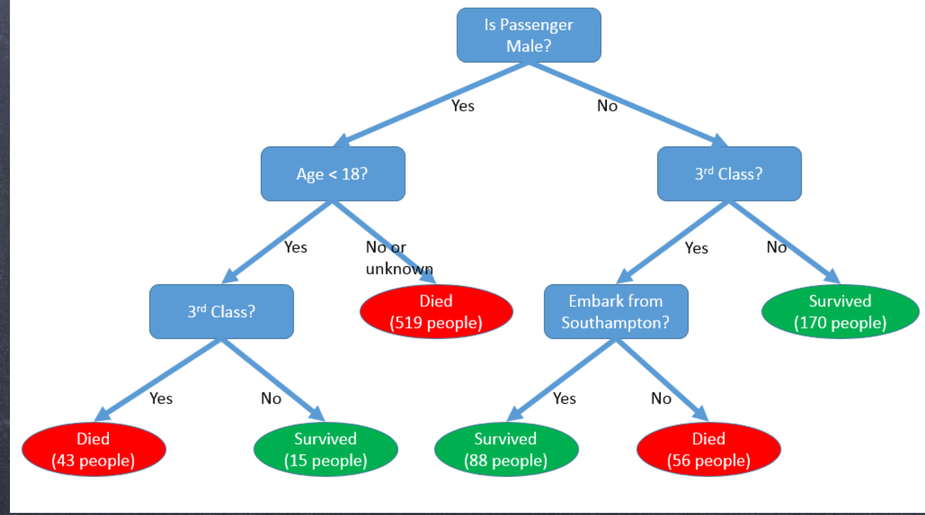
Supervised Learning (cont.)

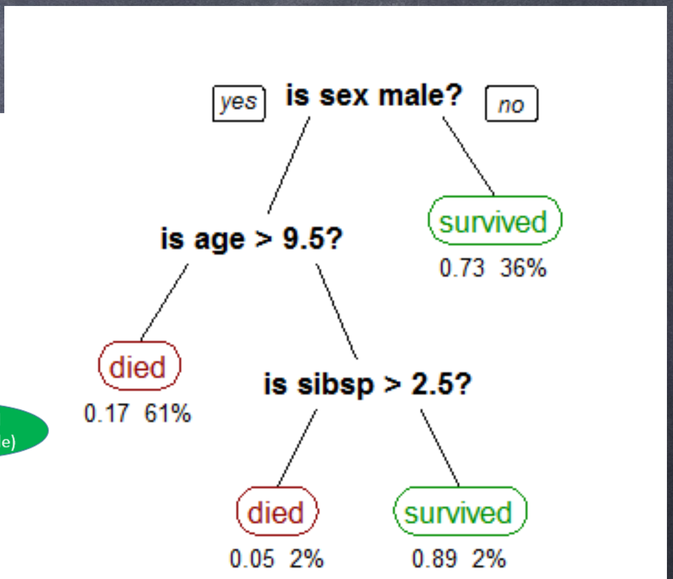
- 판별분석, 예측모형 : 신용평가 모형
  - 예측변수(input) 활용 판별규칙
- SVM
  - 패턴인식, 지도학습
  - 분류, 회귀분석에 사용
- Artificial Neural Network
  인공신경망
  - 인공뉴런(노드)이 학습을 통한 문제해결
  - 교사학습 vs. 비교사 학습(기계학습)
- 의사결정나무
  - 타이타닉 생존자 분류





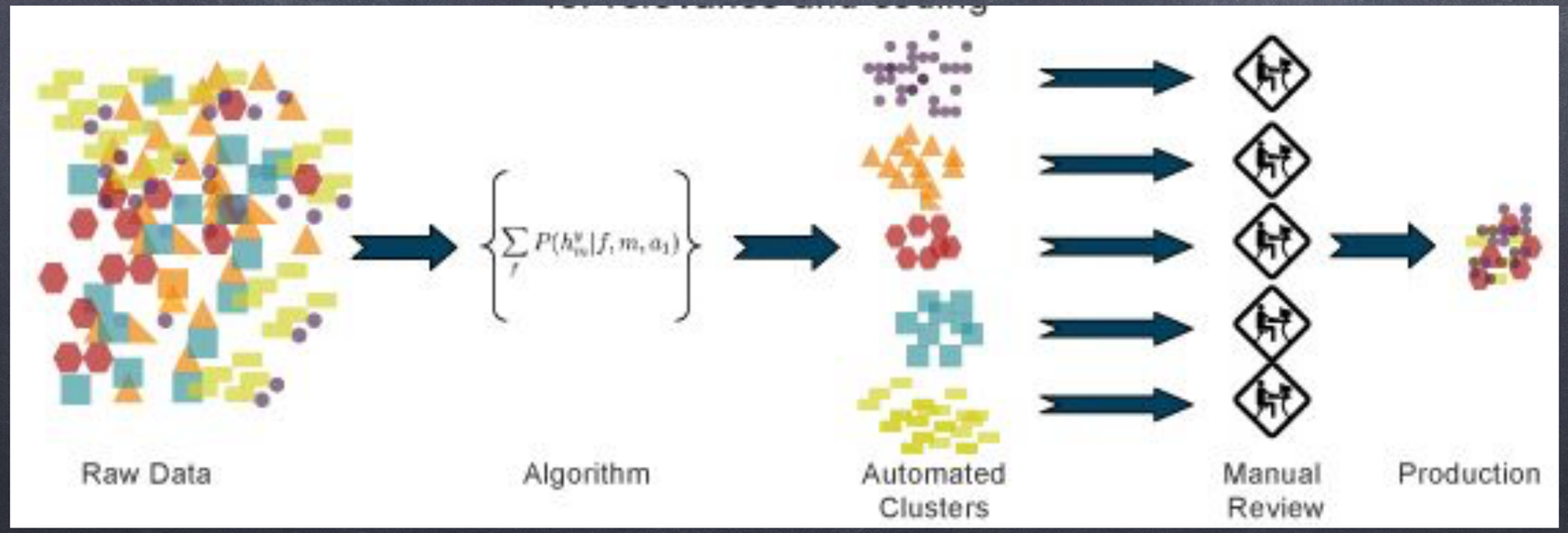






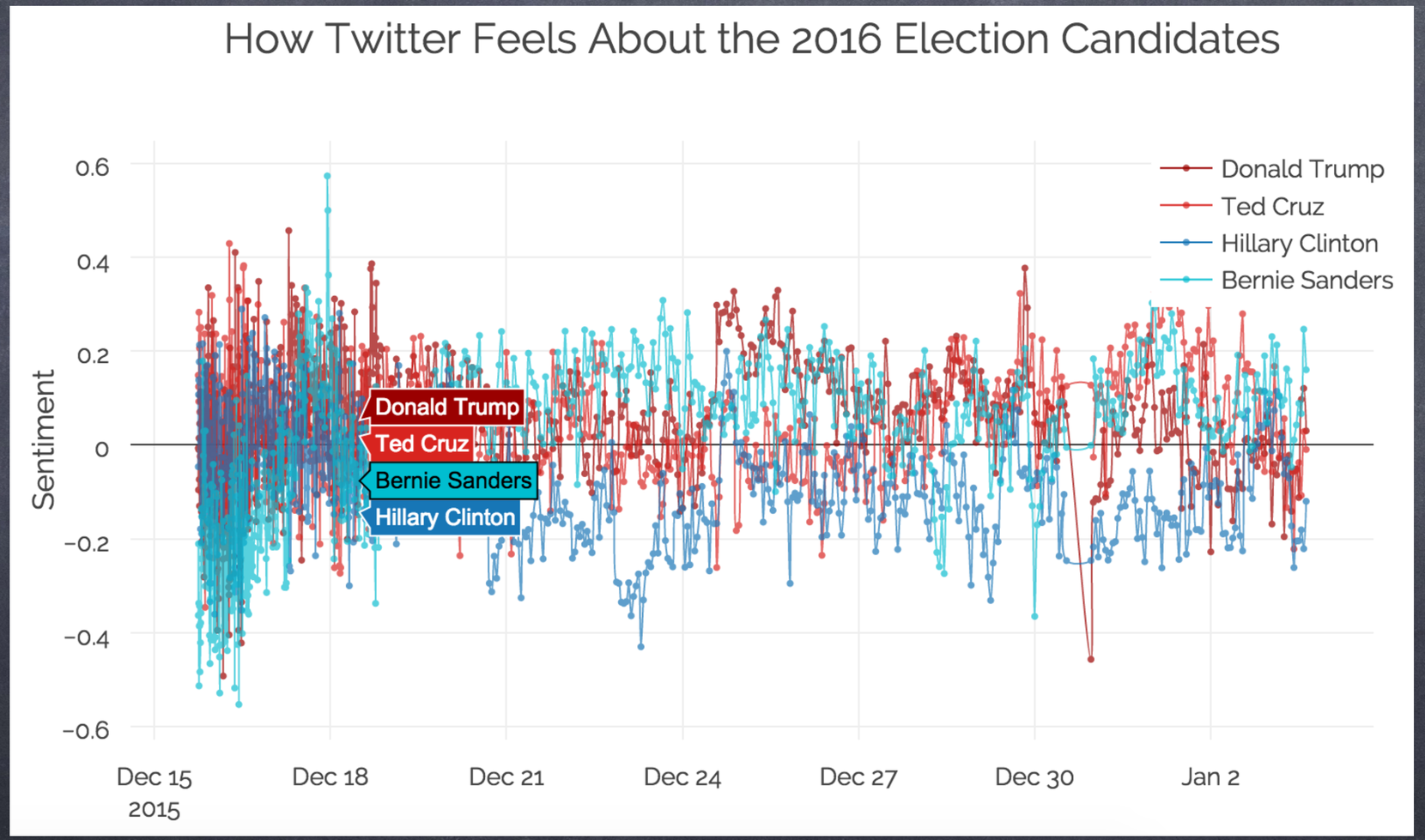
Un-Supervised Learning 자율학습

- 군집분석 : 집단 분류가 없는 개체들을 개체 내의 내재된 관계를 설명하는 함수(유사성의 함수)를 활요하여 분류
  - 계층적 군집 - (Linkage 연결-덴드로그램)
  - 비계층적 군집 - K-means
- 신경망 이론
  - SOM(self organizing map), ART(adaptive resonance theory) 알고리즘



Sentiment Analysis

- Opinion Mining 오피니언 마이닝



Twitter mining with R

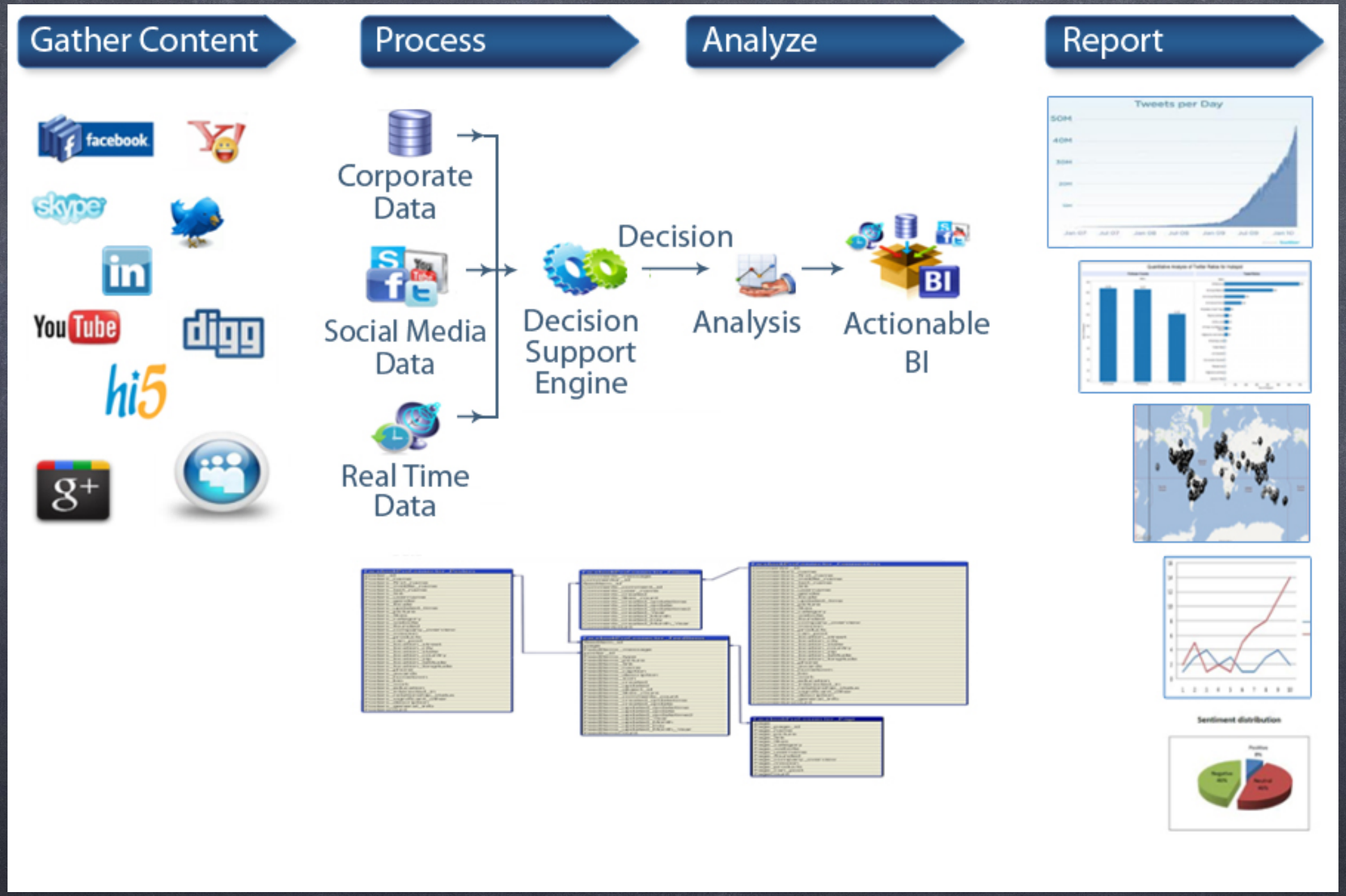
☑R-설치

http://r-project.org

```
1  #R=Twitter mining with R
2  Sys.getlocale() #set default
3  library(twitteR)     #install.packages("twitteR")
4  library(RCurl)       #install.packages("RCurl")
5  library(ROAuth)      #install.packages("ROAuth")
6  library(base64enc)   #install.packages("base64enc")
```

```
[[95]]
[1] "3RMarket0: 빅데이터의 분류\nhttps://t.co/Apjm3ybiwy\n★-2014213605★"

[[96]]
[1] "eejongho: RT @nalm: 누리미디어, 빅데이터 추천기술 적용 학술정보서비스 오픈  https:/

[[97]]
[1] "TheCatcherNews: 중국 정부가 의료 빅데이터 산업을 키우기 위해 나섰습니다 https://t.c
://t.co/X6bcdF88CE"

[[98]]
[1] "3RMarket0: [1편] '자원' : 활용할 수 있는 빅데이터 발견하기\nhttps://t.co/Un7g3F
```

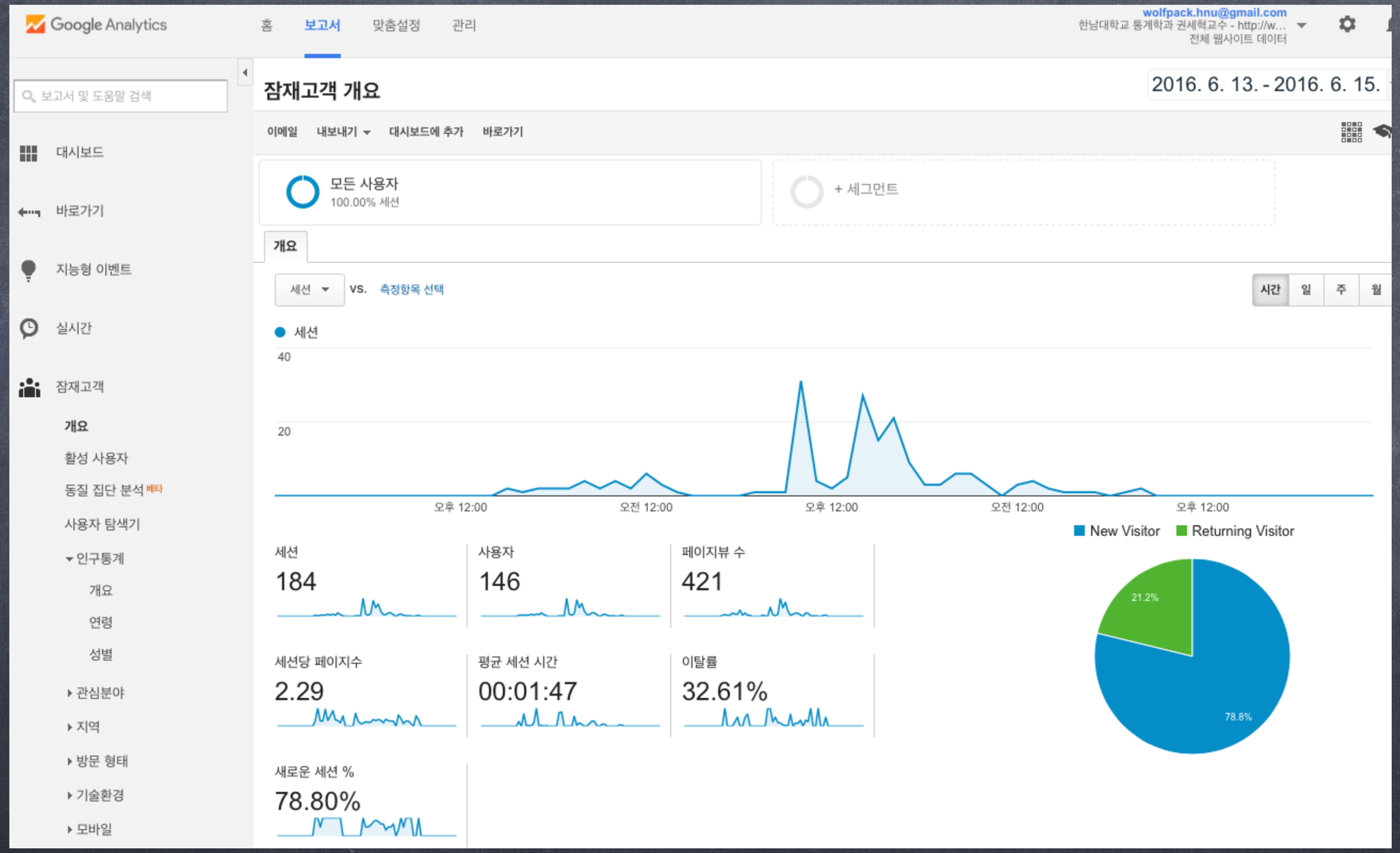
Social Media Analytic



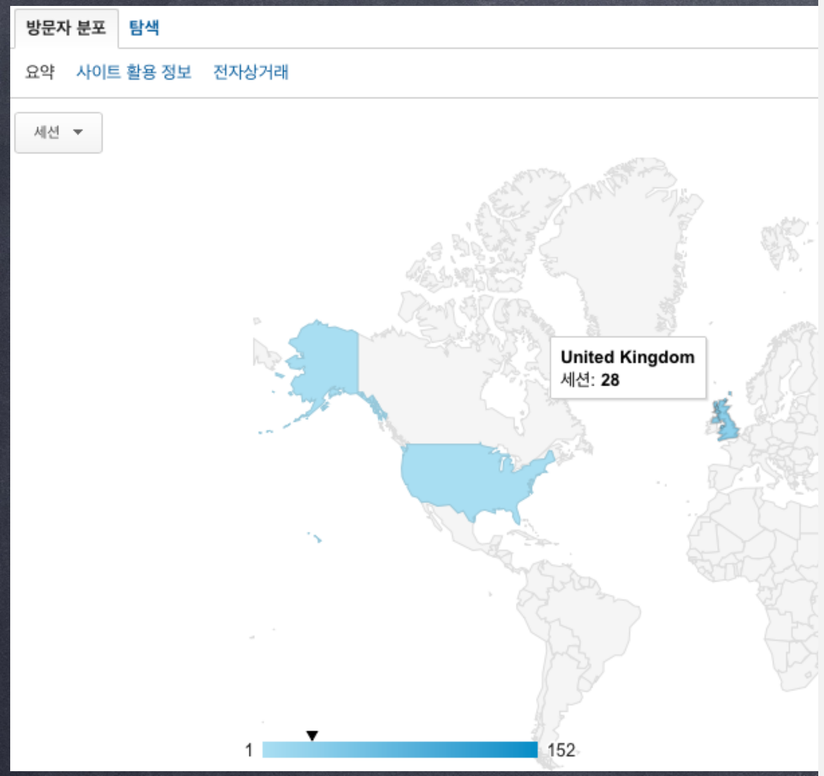
Social Media Analytic with Google Analytics

http://analytics.google.com

My webpages : Google analytics



My webpages : Google analytics (cont.)



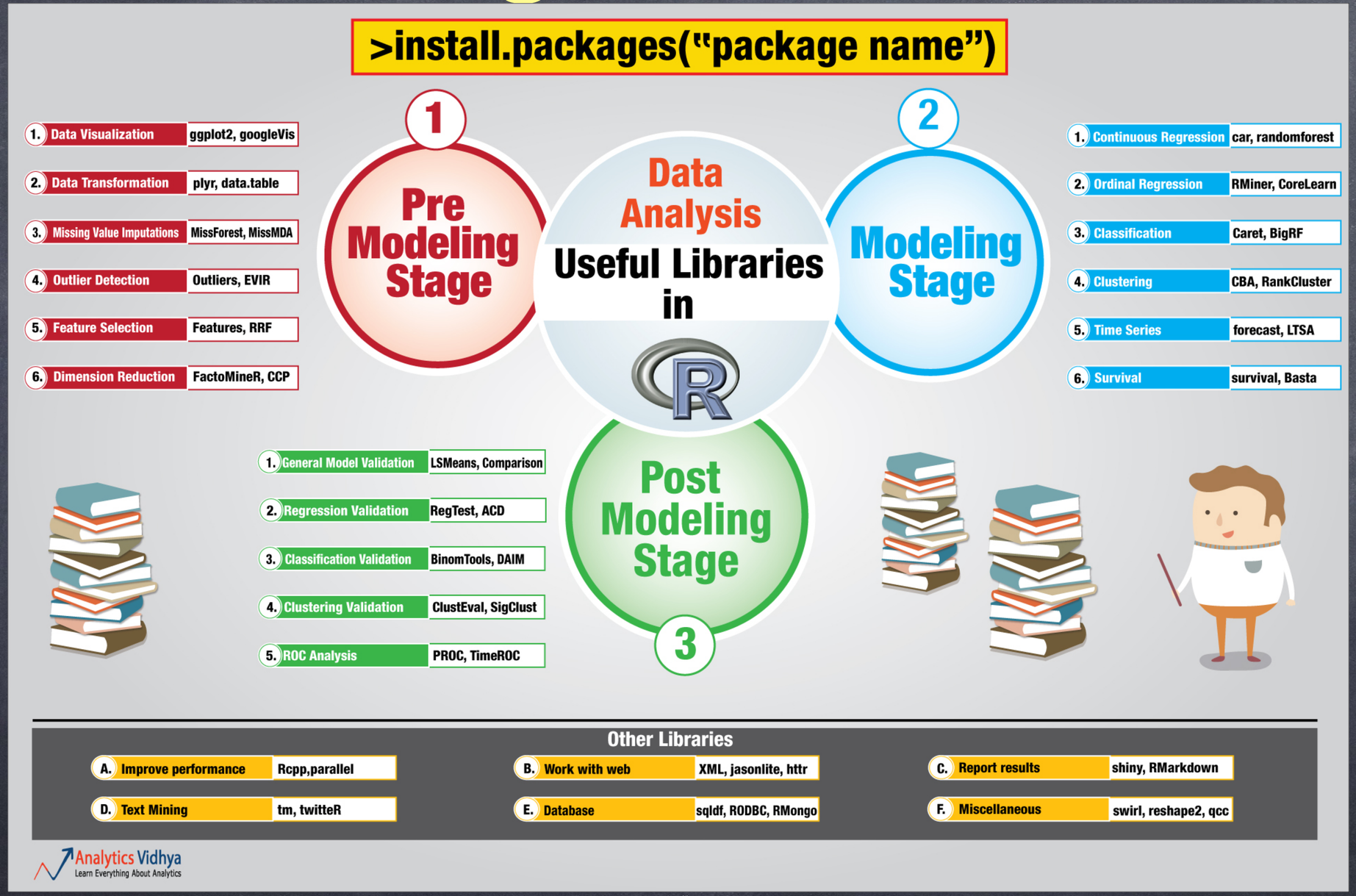
- 지역
  - 언어
  - 위치
- 방문 형태
  - 신규 방문 vs. 재방문
  - 게재빈도 및 방문빈도
  - 참여도
- 기술환경
- 모바일
  - 개요
  - 기기
- 맞춤
- 벤치마킹

사용자 흐름

획득

	휴대기기 정보	획득		
		세션 ↓	새로운 세션 %	신규 방문자
		57 전체 대비 비율 (%): 30.81% (185)	63.16% 평균 조회: 78.38% (-19.42%)	36 전체 대비 비율 (%): 24.83% (145)
☐	1. Apple iPhone	**28** (49.12%)	50.00%	14 (38.89%)
☐	2. Samsung SM-N916K Galaxy Note 4	**3** (5.26%)	33.33%	1 (2.78%)
☐	3. Microsoft Xbox One	**2** (3.51%)	100.00%	2 (5.56%)
☐	4. Samsung SHV-E300K Galaxy S IV	**2** (3.51%)	50.00%	1 (2.78%)
☐	5. Samsung SHV-E330S Galaxy S IV	**2** (3.51%)	100.00%	2 (5.56%)
☐	6. Samsung SM-G900K Galaxy S5	**2** (3.51%)	50.00%	1 (2.78%)
☐	7. Samsung SM-G920L Galaxy S6	**2** (3.51%)	50.00%	1 (2.78%)
☐	8. Samsung SM-N910S Galaxy Note 4	**2** (3.51%)	50.00%	1 (2.78%)
☐	9. Samsung SM-N920S Galaxy Note 5	**2** (3.51%)	100.00%	2 (5.56%)
☐	10. (not set)	**1** (1.75%)	100.00%	1 (2.78%)

Big Data & R

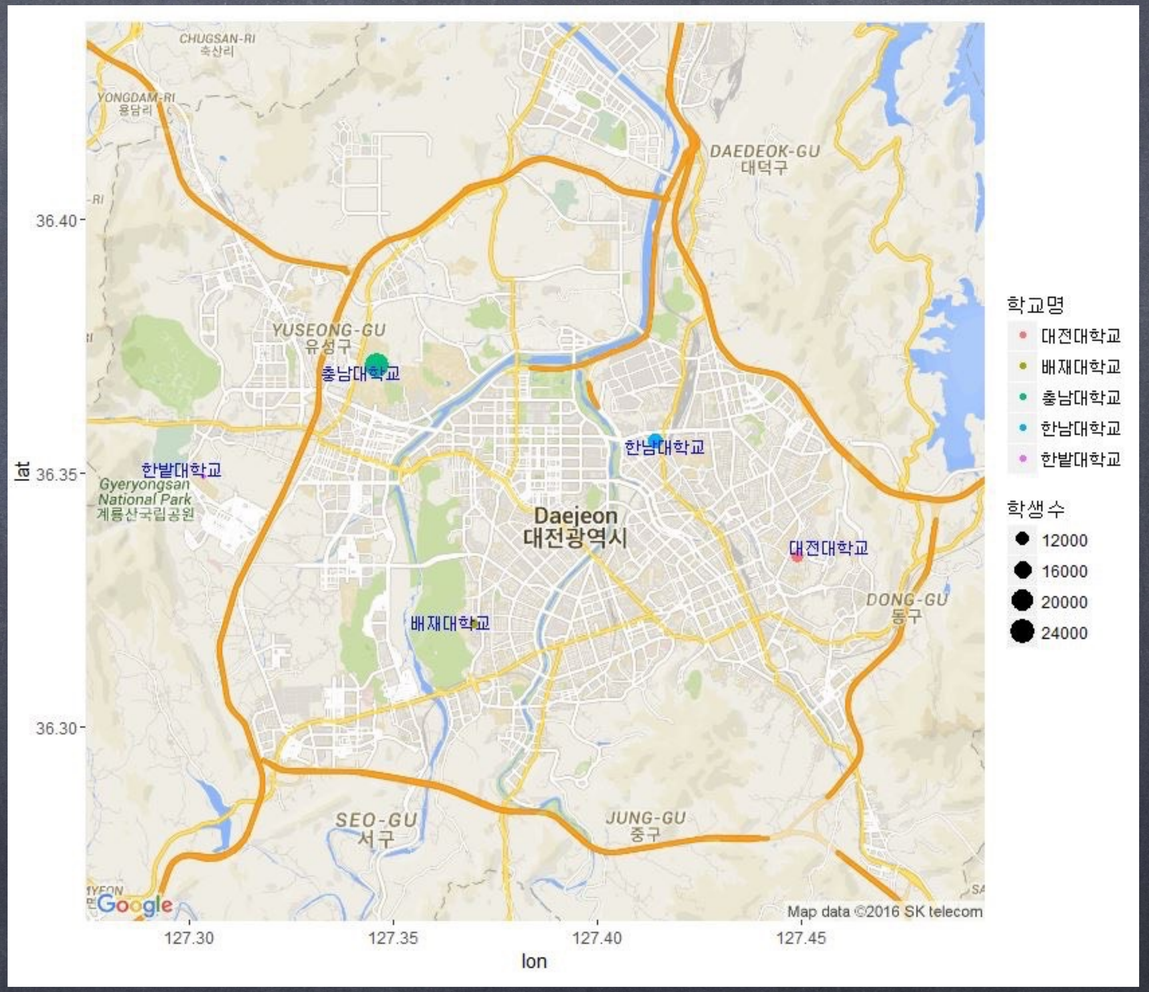

>install.packages("package name")
1
Pre Modeling Stage
1. Data Visualization ggplot2, googleVis
2. Data Transformation plyr, data.table
3. Missing Value Imputations MissForest, MissMDA
4. Outlier Detection Outliers, EVIR
5. Feature Selection Features, RRF
6. Dimension Reduction FactoMineR, CCP
Data Analysis
Useful Libraries in
R
2
Modeling Stage
1. Continuous Regression car, randomforest
2. Ordinal Regression RMiner, CoreLearn
3. Classification Caret, BigRF
4. Clustering CBA, RankCluster
5. Time Series forecast, LTSA
6. Survival survival, Basta
3
Post Modeling Stage
1. General Model Validation LSMeans, Comparison
2. Regression Validation RegTest, ACD
3. Classification Validation BinomTools, DAIM
4. Clustering Validation ClustEval, SigClust
5. ROC Analysis PROC, TimeROC
Other Libraries
A. Improve performance Rcpp,parallel
B. Work with web XML, jasonlite, httr
C. Report results shiny, RMarkdown
D. Text Mining tm, twitteR
E. Database sqldf, RODBC, RMongo
F. Miscellaneous swirl, reshape2, qcc
Analytics Vidhya
Learn Everything About Analytics

Big data with R ggplot2()

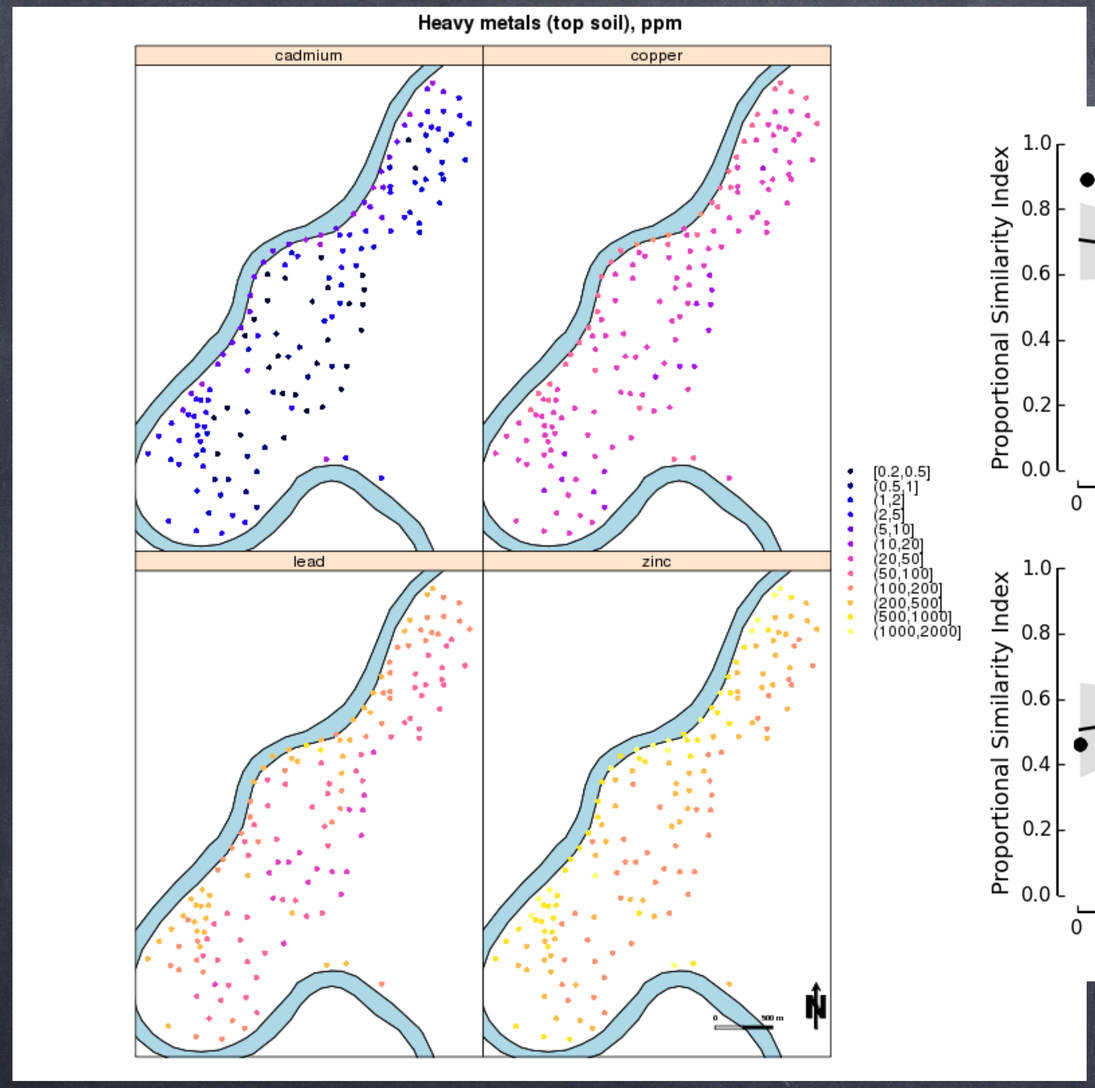
```
install.packages(“ggmap”); install.packages("ggplot2")
library(ggplot2); library(ggmap)

#  maptype=c("terrain", "terrain-background", "satellite",
#  "roadmap", "hybrid", "toner", "watercolor", "terrain-labels", "terrain-lines",
#  "toner-2010", "toner-2011", "toner-background", "toner-hybrid",
#  "toner-labels", "toner-lines", "toner-lite")

uni_seoul <- read.csv("서울 11개 대학.csv", header=T)
seoul <- get_map("seoul", zoom=11, maptype = "roadmap")

seoul_map <- ggmap(seoul)
seoul_map <- seoul_map + geom_jitter( data=uni_seoul, aes(x=경도, y=위도, size = 학생수,color=학교명)) +
scale_size(name="학생수")
seoul_map + geom_text(data=uni_seoul, aes(x = 경도, y = 위도, label=학교명),size=3,col="blue")
#=============================================================
uni_daejeon <- read.csv("대전 5개 대학.csv", header=T)
daejeon <- get_map("daejeon", zoom=12, maptype = "roadmap")

daejeon_map <- ggmap(daejeon)
daejeon_map <- daejeon_map + geom_jitter( data=uni_daejeon, aes(x=경도, y=위도, size = 학생수,color=학교명)) +
scale_size(name="학생수")
```

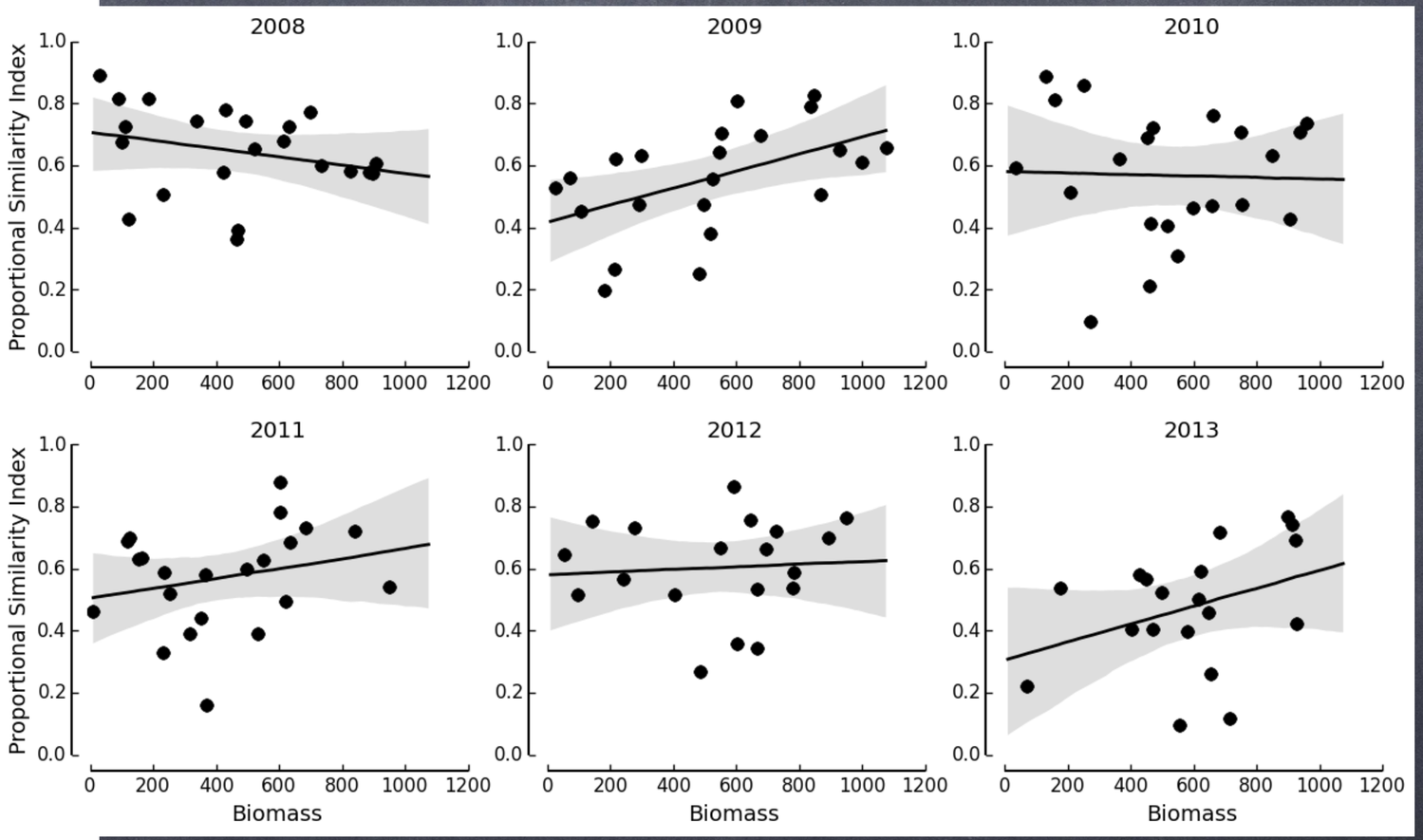
ggplot2 결과


학생수
10000
20000
30000
학교명
건국대학교
고려대학교
동국대학교
서강대학교
서울대학교
서울시립대학교
성균관대학교
숭실대학교
연세대학교
한국외국어대학교
한양대학교
lat
lon
37.7
37.6
37.5
37.4
126.8
126.9
127.0
127.1
Goyang
고양시
Bukhansan National Park
북한산국립공원
DOBONG-GU
도봉구
NOWON-GU
노원구
GANGBUK-GU
강북구
Guri
구리시
JONGNO-GU
종로구
SEODAEMUN-GU
서대문구
MAPO-GU
마포구
YONGSAN-GU
용산구
GWANGJIN-GU
광진구
GANGDONG-GU
강동구
SONGPA-GU
송파구
SEOCHO-GU
서초구
GWANAK-GU
관악구
DONGJAK-GU
동작구
GURO-GU
구로구
YANGCHEON-GU
양천구
OJEONG-GU
오정구
SOSA-GU
소사구
JUNGWON-GU
중원구
SUJEONG-GU
수정구
Bucheon
부천시
Gwangmyeong
광명시
Gwacheon
과천시
Anyang
안양시
Seongnam
성남시
서울특별시
김포국제공항
Seonyudo
선유도
노들섬
채빛섬
한국외국어대학교
고려대학교
성균관대학교
서울시립대학교
연세대학교
서강대학교
한양대학교
건국대학교
숭실대학교
서울대학교
Google
Map data ©2016 SK telecom


학교명
대전대학교
배재대학교
충남대학교
한남대학교
한밭대학교
학생수
12000
16000
20000
24000
lat
lon
36.40
36.35
36.30
127.30
127.35
127.40
127.45
CHUGSAN-RI
축산리
YONGDAM-RI
용담리
DAEDEOK-GU
대덕구
YUSEONG-GU
유성구
충남대학교
한남대학교
한밭대학교
Gyeryongsan
National Park
계룡산국립공원
Daejeon
대전광역시
대전대학교
배재대학교
DONG-GU
동구
SEO-GU
서구
JUNG-GU
중구
Google
Map data ©2016 SK telecom

Big data with R lattice()





Still going…

R-bloggers
받는 사람: wolfpack_HNU
답장 받는 사람: R-bloggers
[R-bloggers] R, Yelp and the Search for Good Indian Food – An Open Course (and 6 more aRticles)

[R-bloggers] R, Yelp and the Search for Good Indian Food – An Open Course (and 6 more aRticles)

- **R, Yelp and the Search for Good Indian Food – An Open Course**
- **Using Microsoft R Server on a single machine for experiments with 600 million taxi rides.**
- **githubinstall: New R Package for Easy to Install R Packages on GitHub**
- **My knitr LaTeX template: manuscript and supplement interleaved in one source file**
- **R Hero saves Backup City with archivist and GitHub**
- **Le Monde puzzle [#965]**
- **R holds top ranking in KDnuggets software poll**

R, Yelp and the Search for Good Indian Food – An Open Course

Posted: 14 Jun 2016 03:09 PM PDT

(This article was first published on **DataCamp Blog**, and kindly contributed to **R-bloggers**)

New Free Course by Springboard and DataCamp

Are all Yelp restaurant reviews created equal? Should we place greater trust in reviews made by people who know the cuisine well? How about reviews of eth

Play Course!

This **free interactive tutorial** will walk you through importing data, data manipulation, and data visualization. More importantly, it will teach you how to "cut thr