

# Introduction to Data Mining: Big Data and Data Mining

---

**Ph.D. 이 기 천**

**Industrial Engineering, Hanyang University**

**2015**



# Method & Grading

- 강의중심 (강의 70%, 과제/실습 30%)
- 교재  
주교재: Data Mining for Business Intelligence (ISBN 978-0-470-52682-8)  
& additional handouts
- Software – XLMiner (online help: [www.resample.com/xlminer/help/Index.htm](http://www.resample.com/xlminer/help/Index.htm))  
R
- Grading

Attendance/Participation	4+1=5%
Homework	25% reading assignments, book questions, small-scale projects, etc.
Group Projects/Quiz	15%
Midterm	25% to be announced two weeks before the midterm week
Final exam	30% to be announced two weeks before the final week

# Schedule

- 1 Introduction, R
- 2 Data exploration, visualization
- 3 Dimensionality Reduction
- 4 Evaluation of Classification
- 5 Regression
- 6 Classification by K-nn, Minimum Distance
- 7 Naïve Bayes, Classification by Support Vector Machines
- 8 Midterm Exam
- 9 Regression Tree
- 10 Logistic Regression
- 11 Neural Network
- 12 Discriminant Analysis
- 13 Association
- 14 Clustering Analysis
- 15 Text Mining, Further Topics
- 16 Final Exam

# Data Mining?

- 데이터마이닝이란 “대량의 데이터 집합으로부터 유용한 정보를 추출하는 것”이다(**Hand *et al.*, 2001**)
- **Data Mining is a collection of systemic and reasonable ways to extract meaningful information from extensive data sets.**

*Data Mining in the era of Big Data?*

# Cases of Big Data Applications

## 삼성-신한카드, 빅데이터 시대 연다

삼성카드 빅데이터 전문가 영입, 신한카드 '코드나인' 출시



'코드 나인'은 신한카드가 보유한 2200만 고객의 소비 패턴과 특성을 분석해 내놓은 상품개발 체계로, 기존 성별, 연령, 소득 등으로 고객을 구분하는 것에 반해 소비패턴을 기준으로 고객을 분류했다.

9개의 코드는 예를 들면 '프렌드 대디'(여행 등을 같이 하는 친구 같은 아빠), '그레이 젠틀맨'(필수 소비만 하는 시니어), '프리마돈나'(문화, 여가를 즐기는 싱글 직장인) 등으로 고객을 분류한다.

위성호 신한카드 사장은 "고객의 빅데이터를 분석해보니 세대, 계층과 무관하게 유사한 소비경향을 가진 집단이 있었다"며 "이를 바탕으로 남녀 각각 9개의 코드로 분류해 고객 맞춤형 상품과 서비스를 제공할 것이다"라고 밝혔다. [미디어펜=장영일 기자]

# Cases of Big Data Applications



## 빅데이터 분석으로 탄생한 '대용량 요구르트'

기사입력 2014-08-13 11:40 기사수정 2014-08-13 11:40

빅데이터 분석을 통해 기존 요구르트보다 용량을 크게 늘린 대용량 요구르트가 출시됐다.

편의점 CU는 20~30대 성인 여성을 겨냥해 60~150mL인 기존 요구르트보다 용량을 크게 늘린 270mL 'CU 빅 요구르트(1250원)'를 14일 출시한다.

# Cases of Big Data Applications

## "빅데이터로 대박 예감"...중기청, 상권분석 서비스 오픈

안지영 기자 ▾



빅데이터 분석 기술을 활용해 상권 정보와 창업 아이템별 매출 추이를 한눈에 알려주는 서비스가 문을 연다.

중소기업청과 소상공인진흥공단은 비씨카드와 공동으로 소상공인의 성공적인 창업을 지원하기 위해 '점포 이력·평가서비스'를 시작한다고 7일 밝혔다.

# Cases of Big Data Applications

## 자동차 결함 빅데이터로 사전 예측한다

교통안전공단 '제작결함 빅데이터' 시스템 구축  
잠재적 위험요소 출시전 보완 ... 리콜부담 해소

이형근 기자 bass007@dt.co.kr | 입력: 2014-07-28 19:00  
[2014년 07월 29일자 10면 기사]

는 자동차 결함과 관련된 정보를 취합해 분석하는 '자동차 제작결함전산망 빅데이터 분석' 시스템을 구축한다. 자동차 결함 빅데이터 시스템은 국내 자동차 관련 웹사이트, 포털 등에 있는 정보 중 자동차 결함과 관련된 정형, 비정형 데이터를 수집해 결함 요소를 예측, 분석하는 것이 목적이다. 이 시스템을 기존 결함신고센터와 연계해 소비자들이 결함에 관련된 정보를 쉽게 파악할 수 있게 할 계획이며, 응용프로그래밍인터페

이와 관련 BMW는 제품 생산과 수리, 유지 보수 업무 전반을 개선하기 위해 IBM 빅데이터 분석 기술을 도입하고 있다. BMW는 엔진 주물의 주입 온도과 압력, 원재료의 배합 비중, 외부 온도와 습도 등 다양한 데이터를 기반으로 마지막 냉각 공정 전에 미리 최종 불량 가능성을 실시간으로 예측해 불량 가능성이 높은 반제품은 재활용 공정으로 이동한다. BMW는 빅데이터 분석 기술을 이용해 실린더 헤더의 불량률을 16주 동안 50% 줄인 것으로 알려졌다.

미국 자동차회사 포드는 차량에 설치된 센서로 운전자의 주행습관 데이터를 수집·분석해 고객의 숨은 요구(needs)를 찾아내 신제품에 반영한다.



# Cases of Big Data Applications

## [미국] 빅데이터 분석가, 대졸 초봉 1억2500만원

윤소정 기자 | sojeong10@agnews.com

### ●美 빅데이터 분석가, 대졸 초봉 1억2500만원

- 현재 직업을 고르는 사람들은 빅데이터 분석가를 고려해 볼 만 함, 빅데이터는 다양한 분야에서 쓰이고 있음, 결과값을 도출할 필요가 있는 대부분의 영역에서 빅데이터가 쓰이고 있음
- 수요는 많은데 공급이 부족한 상태라 급여조건이 대단히 좋으며, 관련 학과를 졸업하면 바로 12만5000달러(약 1억2500만원)정도를 벌 수 있음
- 1~2년정도 경력 쌓고 자신만의 노하우가 있으면 2배 이상 수입도 가능

국제

미국, 캐나다

입력 2014-08-10 20:09:00, 수정 2014-08-11 01:22:10

## 美 ‘빅데이터 과학자’ 몸값 천정부지

분석 수요 폭증하는데 공급 적어  
초임연봉 1억원대 ... 유치전 치열

미국에서 데이터 과학자는 그 희귀성으로 인해 ‘유니콘’(뿔이 하나인 소)으로 불린다. 데이터 과학 관련 분야를 전공해 박사학위를 받으면 초임 연봉이 10만달러(약 1억360만원) 이상이고, 2년 이내에 연봉이 20만~30만달러로 된다고 월스트리트저널(WSJ)이 9일(현지시간) 보도했다. 전문직 네트워크인 ‘링크드인’에 자신을 데이터 과학자라고 소개하면 하루에 100명 이상으로부터 스카우트 제의를 받는다고 WSJ가 전했다.

# Current Status of Big Data

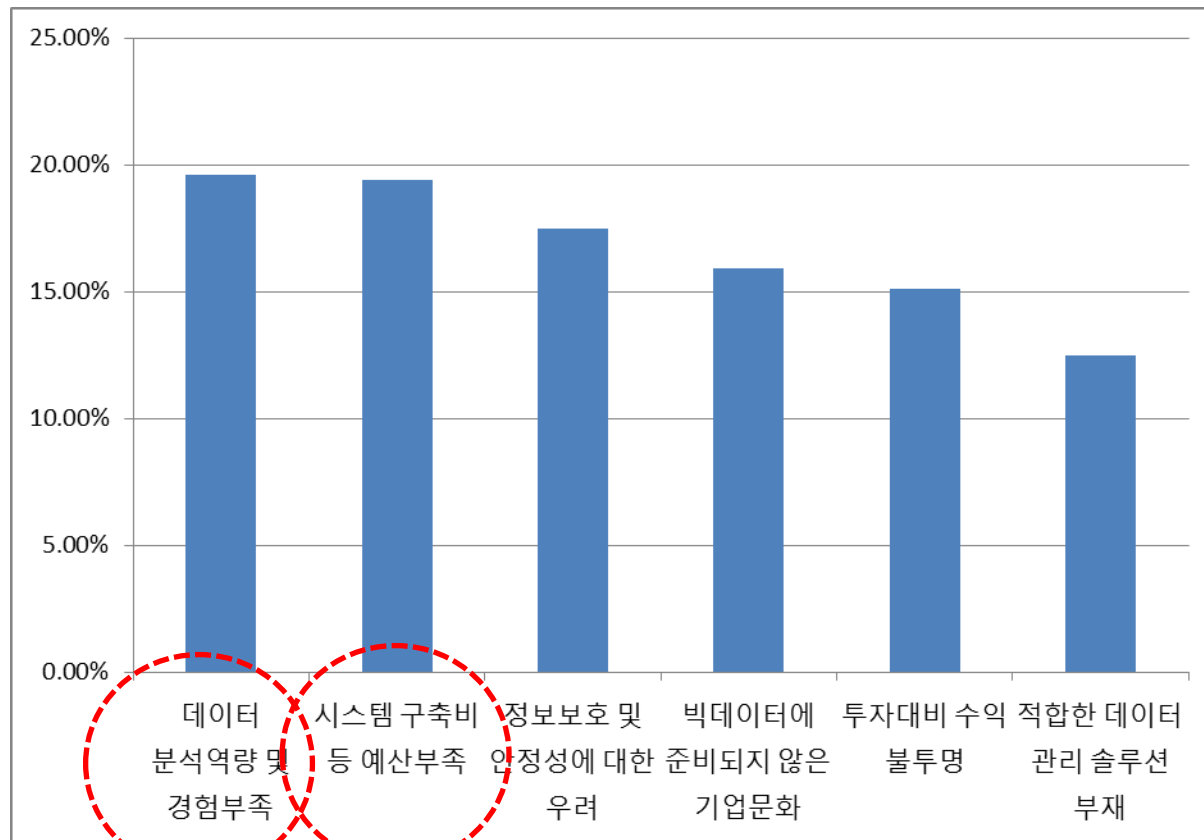


IT강국인데 '빅데이터'는 약소국...기업 82% "활  
용안해"

발행일 2014.07.14

500 companies surveyed

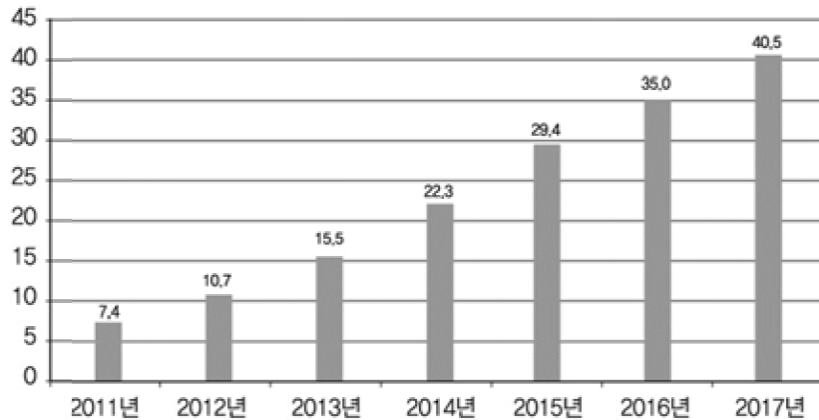
## Reason of not applying big data



# Market size of Big Data

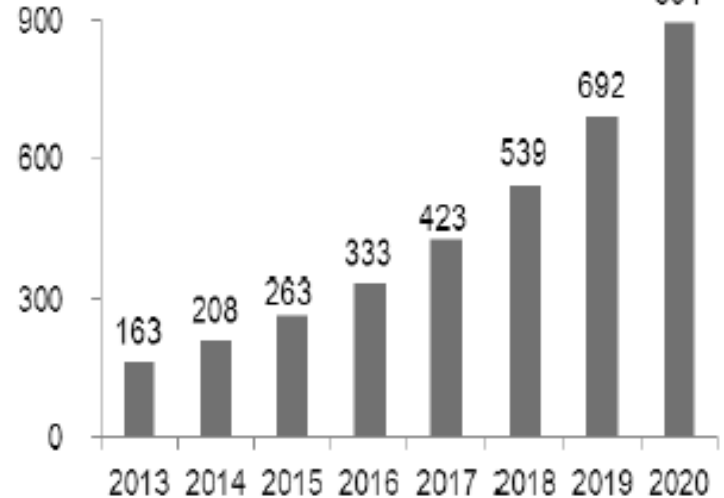
- International Big Data Market
- Domestic Big Data Market

단위: 십억 달러



Source: IDC와 Wikibon의 예상치 평균  
(2013년)

(백만\$)



Source: KISTI (2013년)

Expected to grow annually about 25%

# What is Big Data?

- Data in high volume, of high variety, at high velocity enough not to handle by previous approaches



Volume, Variety, Velocity are called 3Vs.

# 기존 데이터와 빅데이터 비교

- In addition to 3V, Veracity + Visualization + **Value**를 will make up 6V

구분	기존 데이터	빅데이터
Data volume	< terabytes	> = terabytes
Data type	Structured data	Including unstructured data
Analysis process	Structured ways of analysis	Highly variable covering many aspects; finding a broad correlation structure first

# Unstructured Data

- Data types that are hard to handle by classical data-management
  - Including semi-structured data such as XML, HTML and unstructured data such as general text data, images, video clips, and so forth

- Log Data

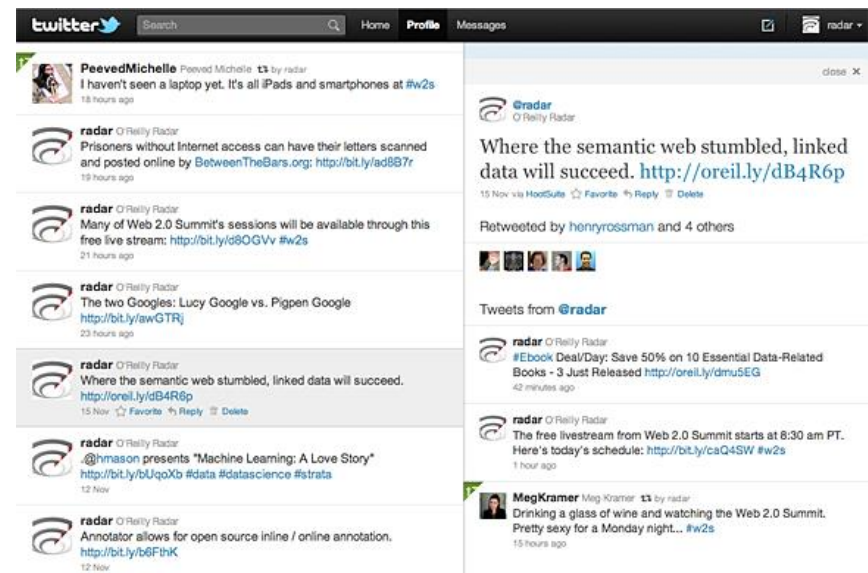
Client                      Date                      Request                      Status                      Size  
10.0.0.1 - - [11/Dec/2008:16:01:22 +0100] "GET /catalog/categories.html HTTP/1.1" 200 958  
"http://myshop.org/index.html"                      User-agent  
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.0.4) Gecko/2008102920 Firefox/3.0.4"

10.0.0.1 - - [11/Dec/2008:16:01:23 +0100] "GET /catalog/books.html HTTP/1.1" 200 1030 "http://myshop.org/catalog/categories.html"  
"Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.0.4) Gecko/2008102920 Firefox/3.0.4"

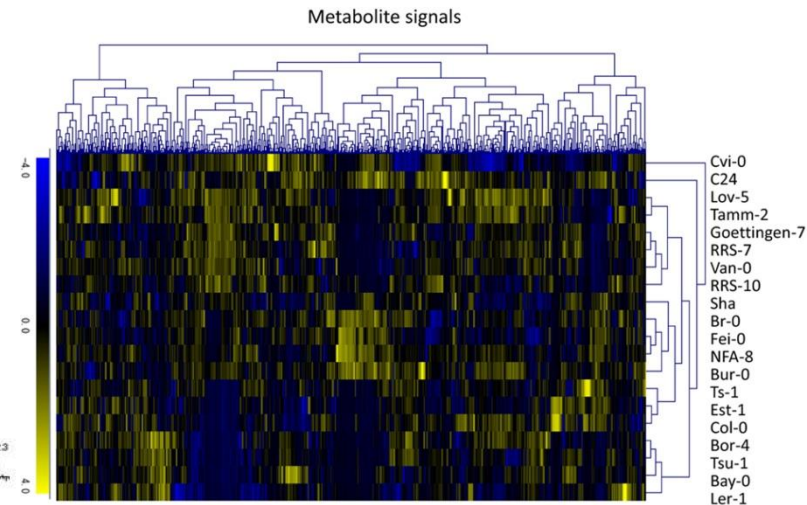
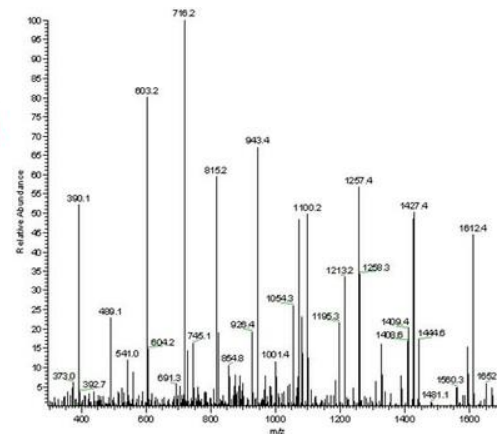
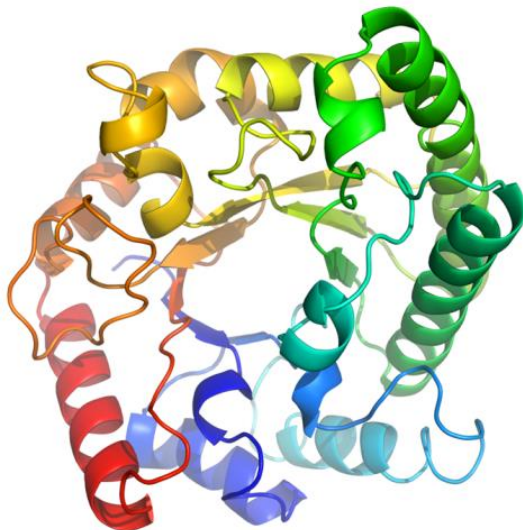
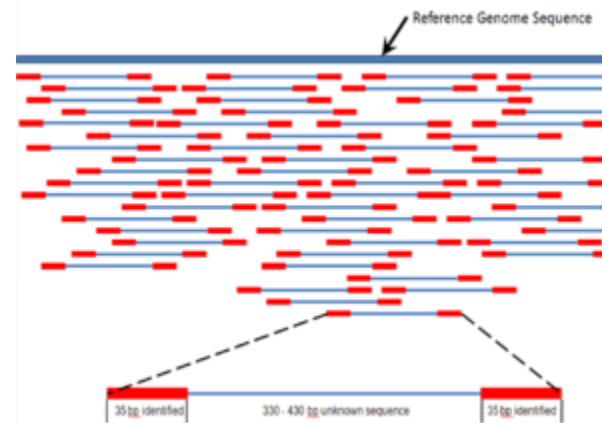
- Image Data



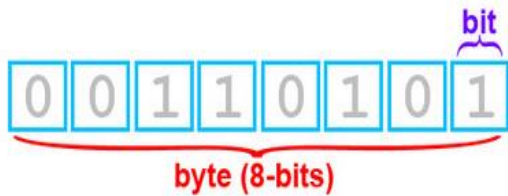
- Blog/SNS Data



# Unstructured Data: in bio fields







1 Megabyte =  
1E6 Bytes

1 gigabytes =  
1E9 Bytes

1 terabytes =  
1E12 Bytes

1 Petabyte =  
1E15 Bytes or  
250,000 DVDs



Facebook is known to hold 100  
petabytes of data



## ■ Friendship map of Facebook



source: Sebastien  
Pierre

Twitter known to generate  
7TB data a day

Facebook known to  
generate 10TB data a day

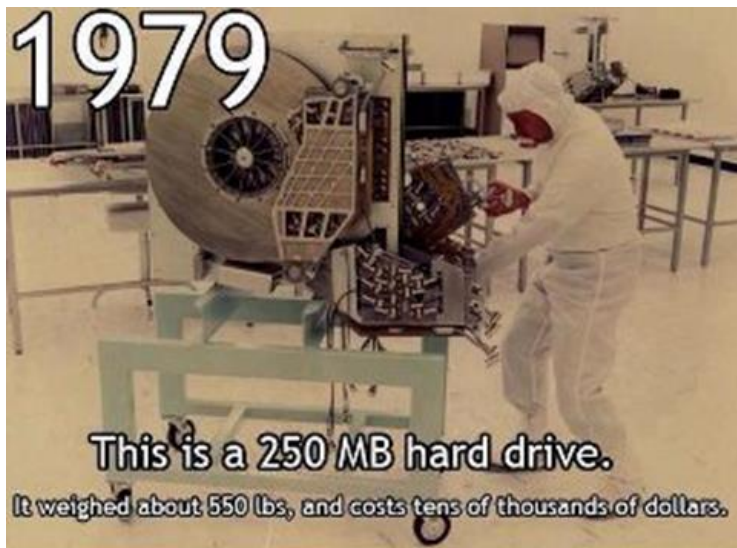
## ■ Satellite image at night by NASA



Social network services  
getting globalized and  
tightly connected to  
our daily life

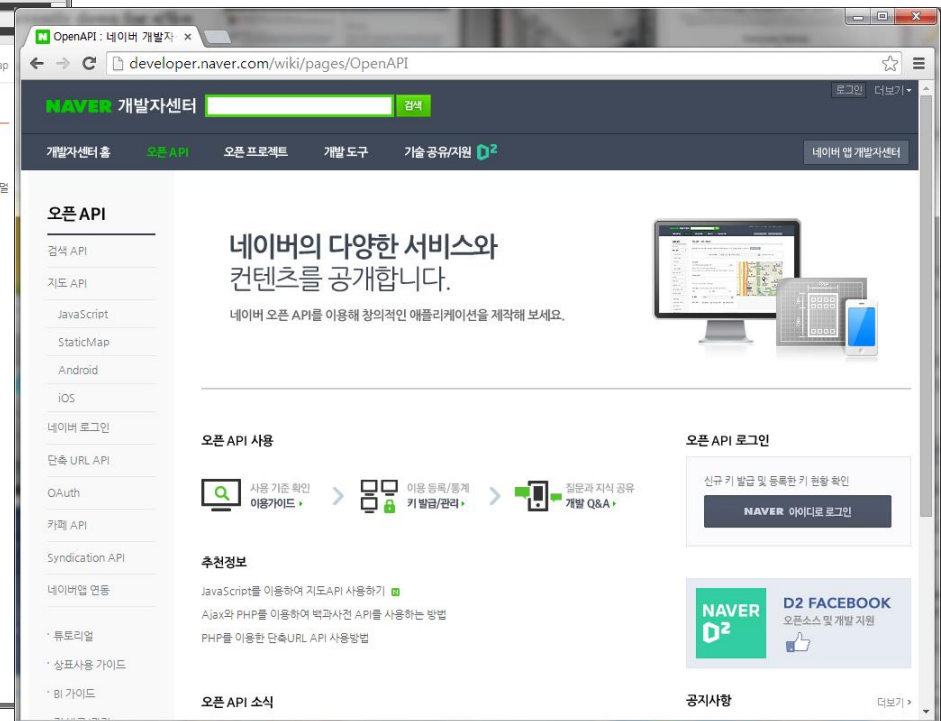
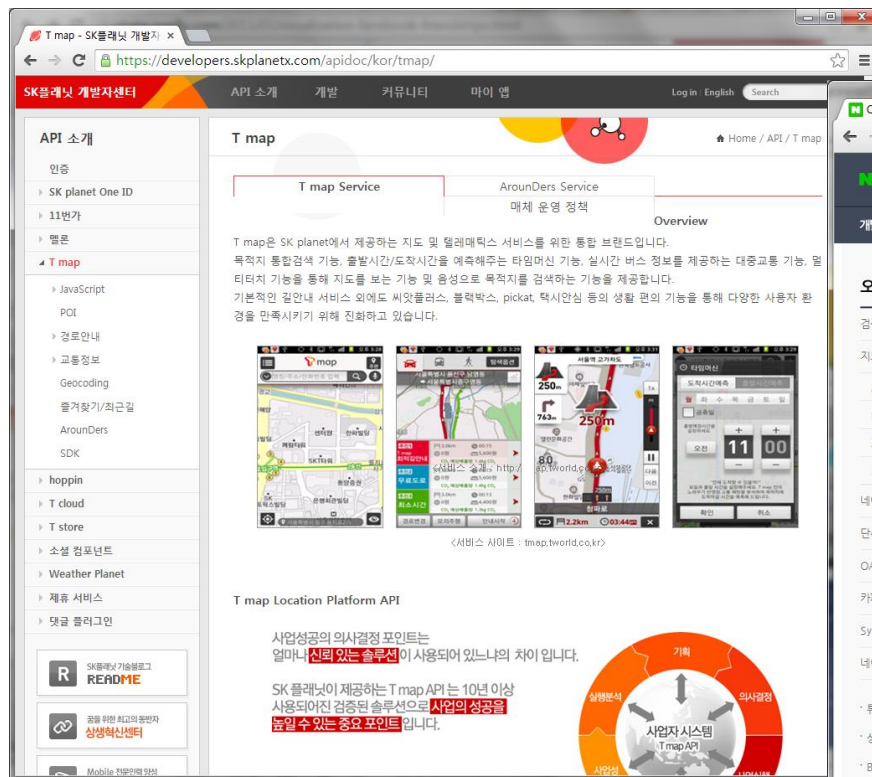
# Background of Big Data Era

- Spread of smart devices, advancement of data storage and retrieval technology
  - Many sensors enable to collect data in private/group/network scopes
  - Substantial drop of storage price and communication costs
  - Increase of computational capability



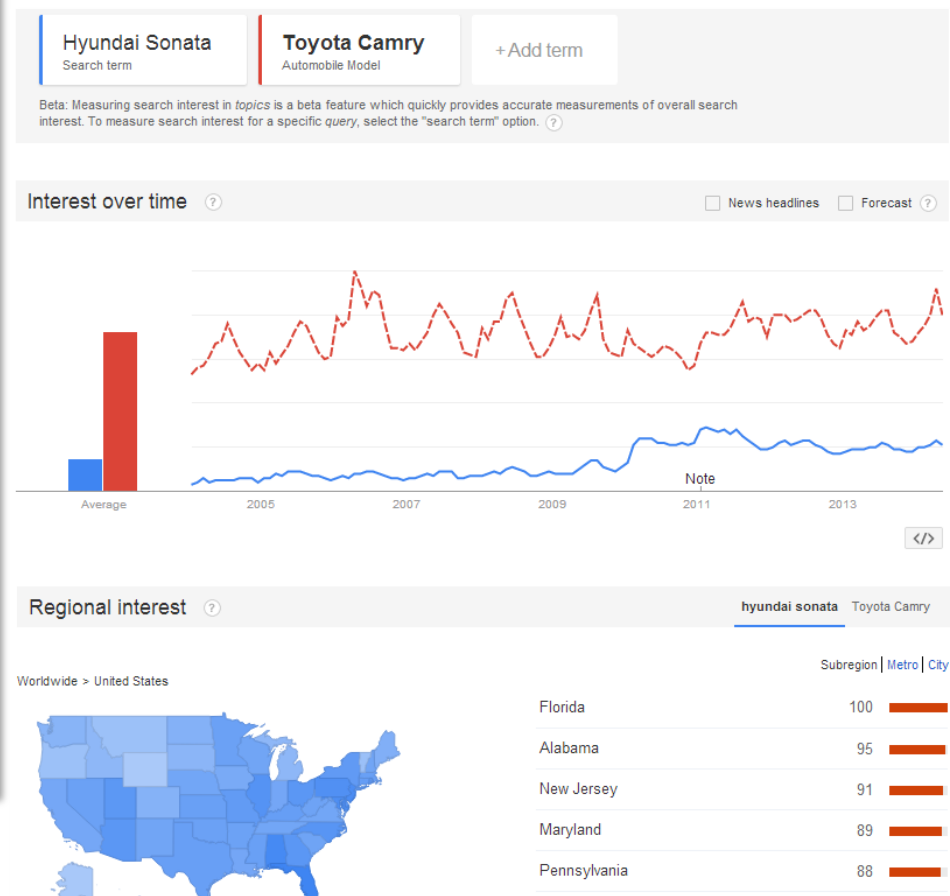
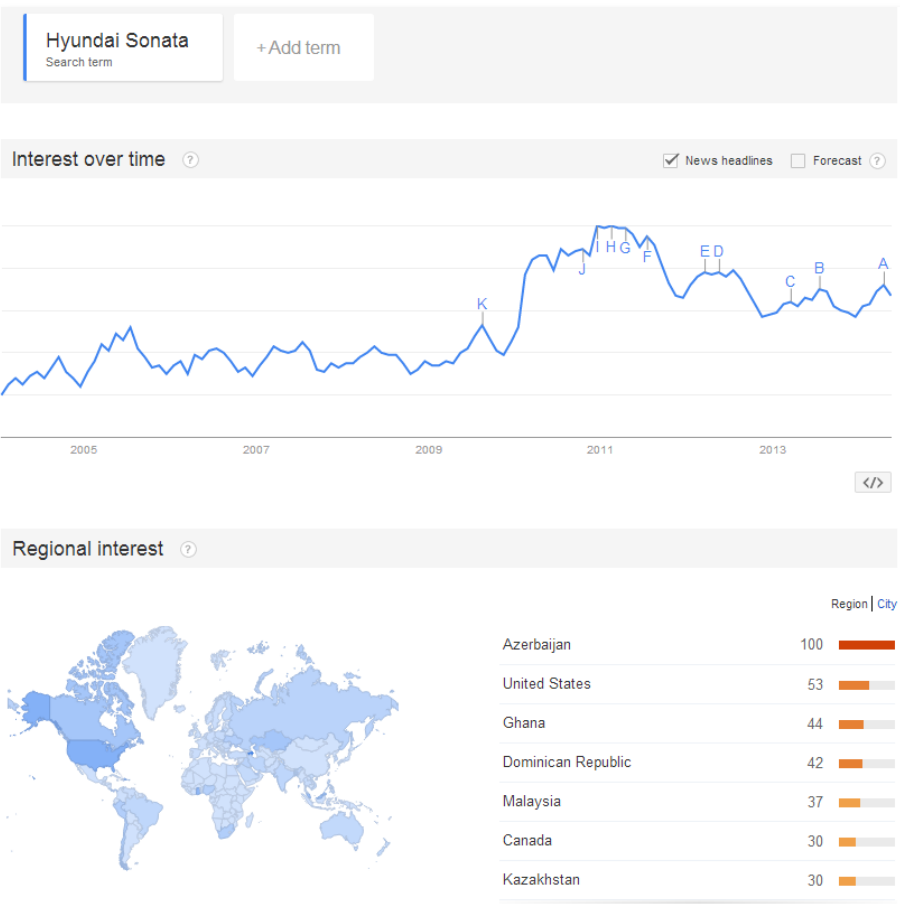
# Background of Big Data Era

- Many information-oriented services are getting cheaper, more available, smarter
  - SK Planet's T map, Naver's Open API, Google's translation, Google's Trend services, and so forth



# Background of Big Data Era

- Google trend service: <http://www.google.com/trends>



# Service providers in Big Data

- Providers in Infrastructure, Software, Service

*Main players*

Infra:  
Storage,  
server,  
network



*Big data service providers*

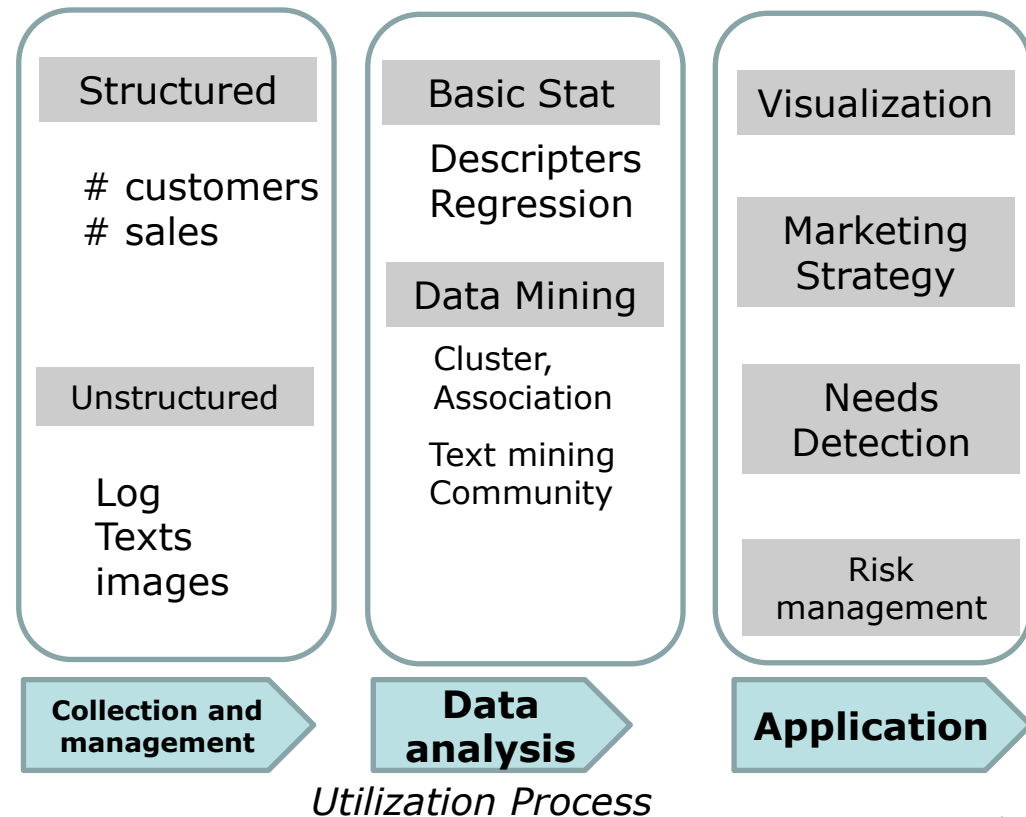
Software



Service



# Utilization Process of Big Data



Big data  
service  
providers

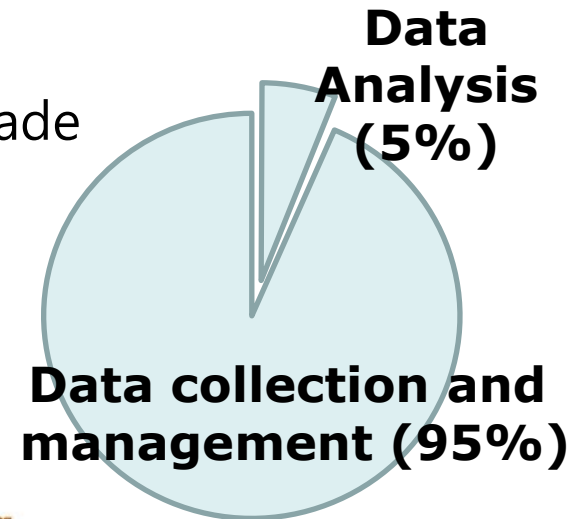




# Infra Techniques in Big Data

## ■ Importance

At the early stage, most efforts are made in establishing infra structure



Source: Gianmarco

# Hadoop



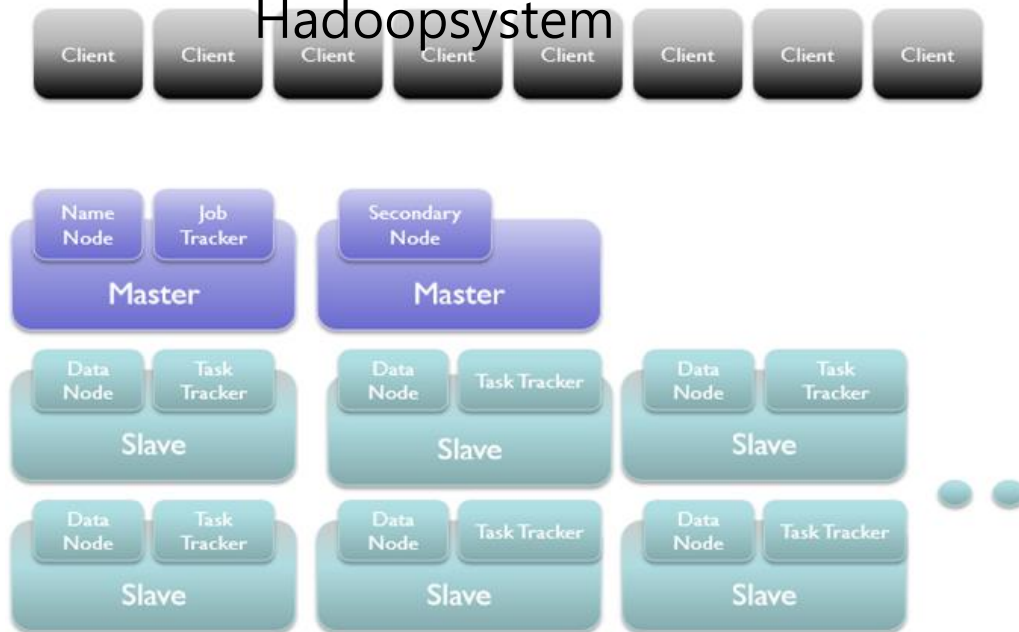
- Open source based HDFS (Hadoop Distributed File System)
- Using many data servers, it constructs a virtual HDFS
- Providing a MapReduce framework to handle data of a big size
- Users
  - Amazon
  - Facebook
  - New York Times
  - Google
  - IBM
  - Etc.



# Hadoop

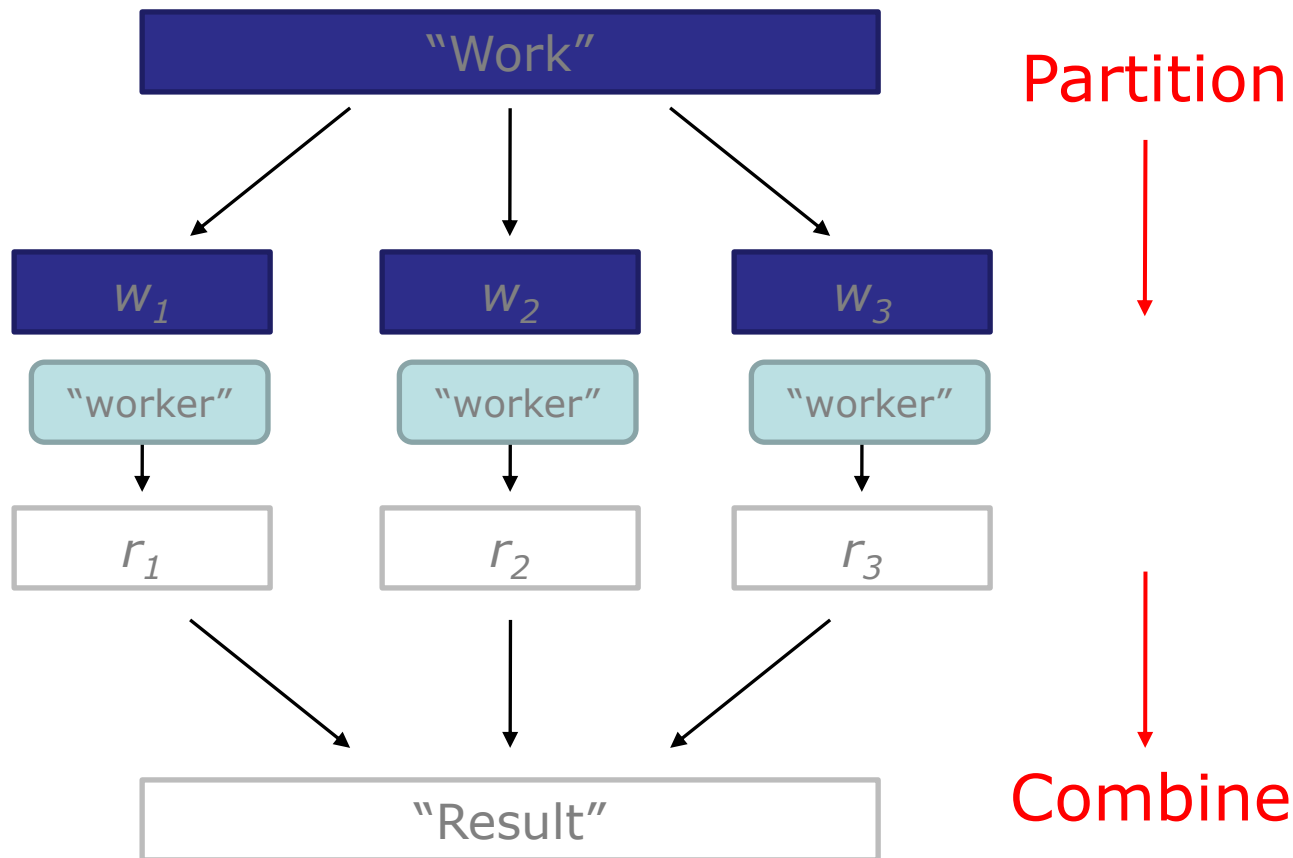
- Philosophy: Let us collect the results of many sub-machines
- It is an implementation of the MapReduce framework
- Led by Yahoo in 2006 and released
  - For example, more than 40,000 machines consists a

Hadoopsystem

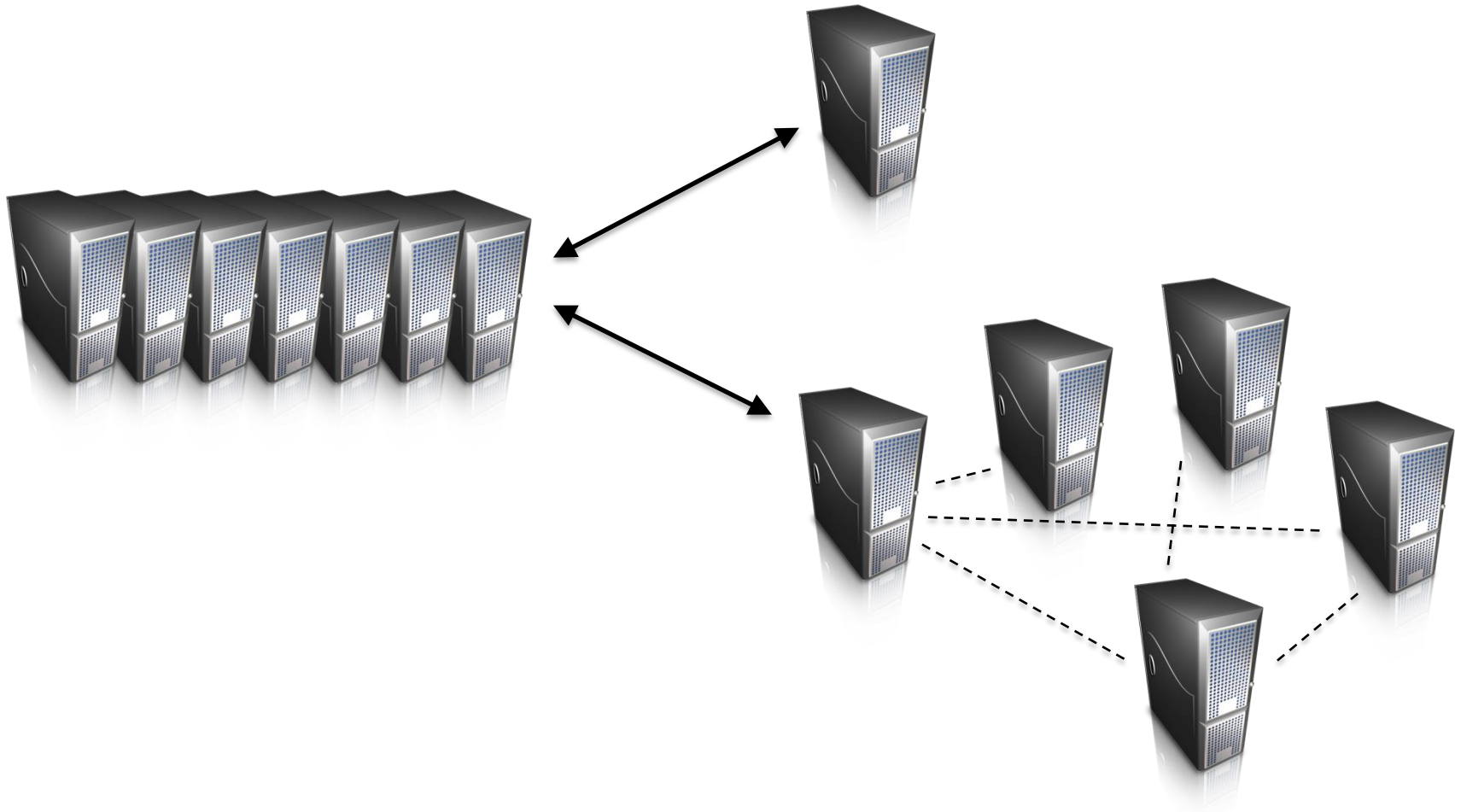


# Hadoop

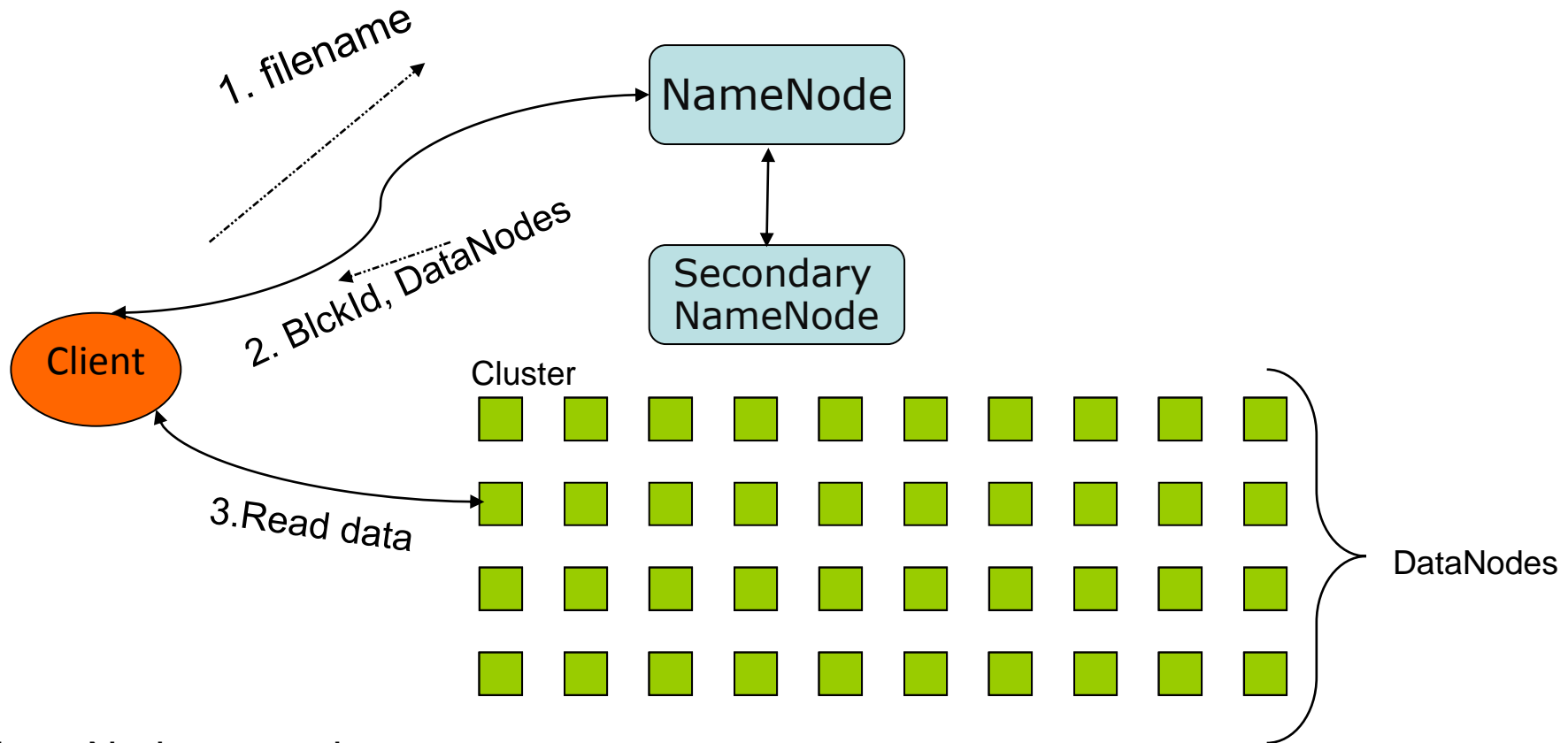
## Divide and Conquer



# How data are managed?



# HDFS (Hadoop Distributed File System)



NameNode : metadata management

DataNode : file data management, connection of block-id and disk position

Secondary NameNode: for fault tolerance, periodically backing up

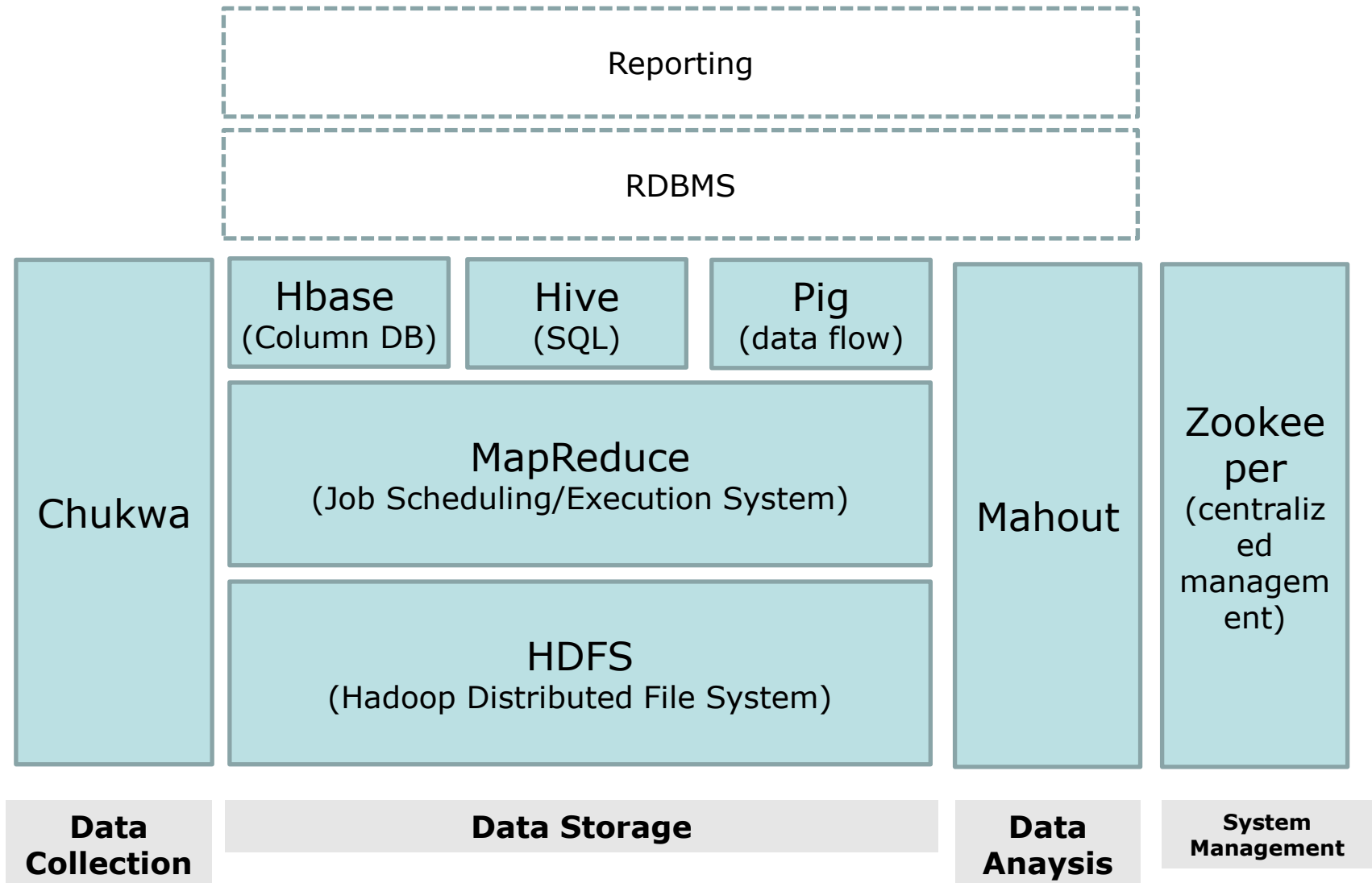
# Hadoop in praxe: Search Assist™



- Search Assist™ shows an application of Hadoop
- 3 year-long log data are analyzed by 20 hadoop steps

	Before Hadoop	After Hadoop
Processing time	26 days	20 minutes
Language	C++	Python
Developing time	2-3 weeks	2-3 days

# Hadoop Ecosystem



# Big Data Tools: public

구분	명칭	역할	사이트
collection	Apache Sqoop	Collection of RDBMS data	<a href="http://sqoop.apache.org/">http://sqoop.apache.org/</a>
Collection	Apache Flume	Collection of event data	<a href="http://flume.apache.org/">http://flume.apache.org/</a>
Collection	Scribe	Collection of log data	<a href="http://en.wikipedia.org/wiki/Scribe_(log_server)">http://en.wikipedia.org/wiki/Scribe_(log_server)</a> <a href="https://github.com/facebookarchive/scribe">https://github.com/facebookarchive/scribe</a>
Storage	Elastic Search	Search engine	<a href="http://camel.apache.org/elasticsearch">http://camel.apache.org/elasticsearch</a>
Analysis	Mahout	Data Analysis	<a href="https://mahout.apache.org/">https://mahout.apache.org/</a>
Analysis/Visualization	R	Data Analysis	<a href="http://www.r-project.org/">http://www.r-project.org/</a>
Infra	Cascading	Developing Hadoop-based systems	<a href="http://www.cascading.org/">http://www.cascading.org/</a>
Infra	Apache Whirr	Developing cloud services	<a href="http://whirr.apache.org/">http://whirr.apache.org/</a>
Management	Apache Oozie	Hadoop management	<a href="http://oozie.apache.org/">http://oozie.apache.org/</a>

# Big Data Tools: platforms



제품명	업체명	수집	저장/처리	분석/표현	관리
BAAS	그루터	Flume Scribe Chukwa	Hadoop Hbase Cassandra Pig Hive ElasticSearch		Oozie Cascading Zookeeper
NDAP	KT NexR	Flume Sqoop	Hadoop Hbase Hive	R	Oozie Zookeeper
CDH	Cloudera	Flume Sqoop	Hadoop Hbase Hive Pig	Mahout	Oozie Whirr Zookeeper
HDP	Hortonworks	Flume Sqoop	Hadoop Hbase Hive Pig		Oozie





# Big Data Tools: players



# Methods of Big Data Analysis

- Basic statistical analysis
  - Descriptive statistics, visualization, regression, and so forth
- Association analysis
  - Frequent-item mining, rule mining
- Text mining
  - Natural language processing, word clustering
- Opinion mining
  - Sentiment analysis
- Network analysis
  - Finding clusters, centroids, and influential nodes

# Purposes of Big Data Analysis

- Classification and prediction
- Association and recommendation
- Clustering
- Abnormality detection

# Classification and prediction

- Is this person going to buy this product?
- How the response rate of a marketing plan will go?
- How good the wine quality of this year will be?
- Does a stock price go up in September?
- How the usage of electricity in this winter move?

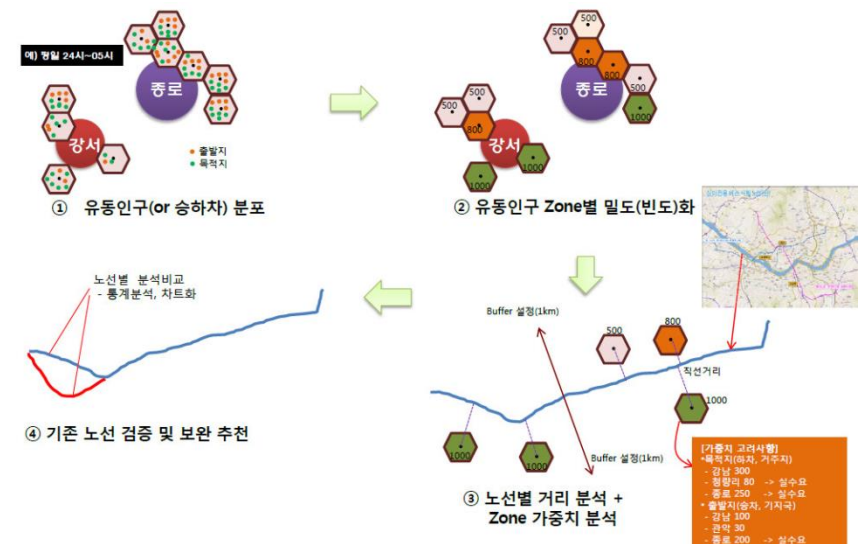
# Classification and prediction

- Google's flue trend: <http://www.google.org/flutrends/>



Can we use this in the marketing plan?

- Late-night Bus routes by KT's cell phone usages



# Association and recommendation

amazon.com<sup>®</sup>



## Customers Who Bought This Item Also Bought



★ Shine Whitening ★  
Professional Teeth  
Whitening Kit ★ (2) 5cc  
Syringes and Mouth ...  
★★★★☆ (124)  
\$35.00



Oral-B Pro-Health Clinical  
Pro-Flex Medium  
Toothbrush 2 Count  
★★★★☆ (127)  
\$6.47



Oral-B Glide Pro-Health  
Comfort Plus Mint Flavor  
Floss Twin Pack 80 M  
★★★★☆ (46)  
\$5.77



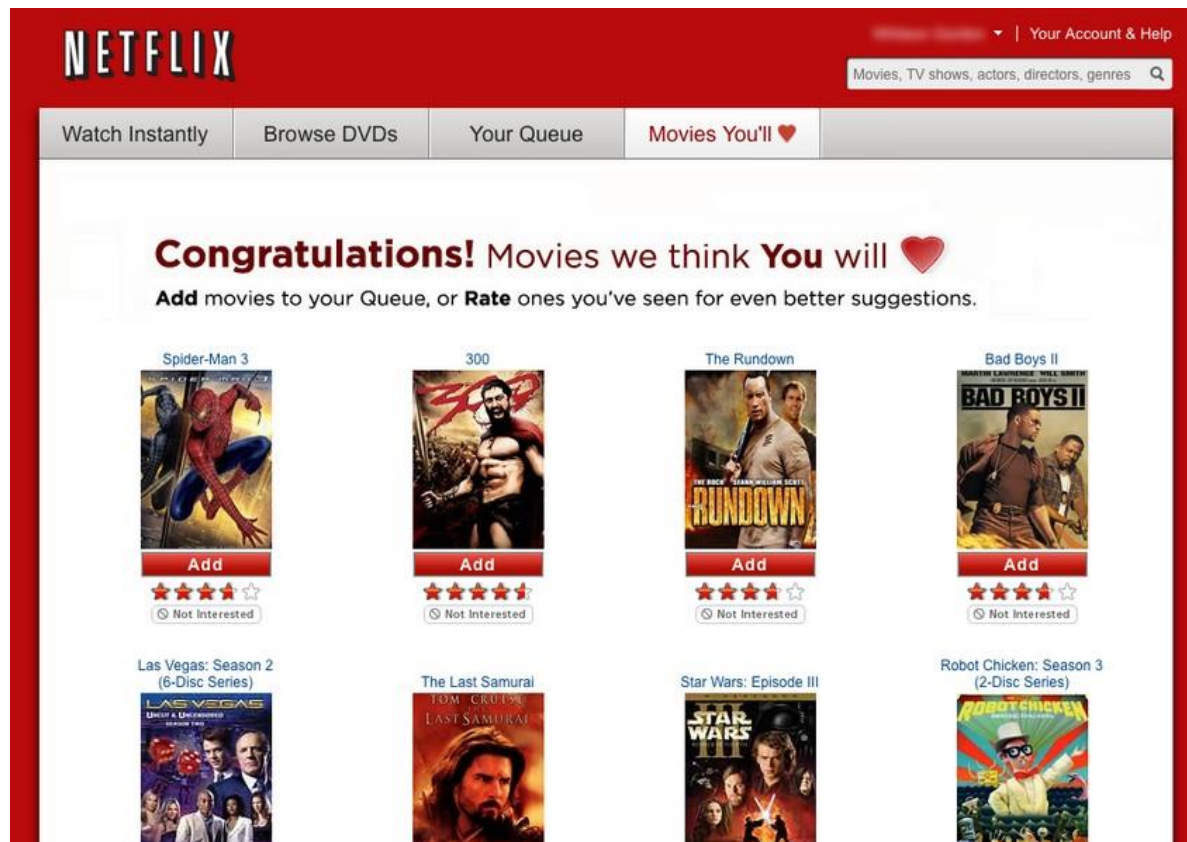
Old Spice High Endurance  
Fresh Scent Men's  
Deodorant Twin Pack 4.5  
Oz  
★★★★☆ (94)  
\$3.65



Opalescence Whiten  
Toothpaste COOL MI  
with flouride wt. 4.7oz  
(GUARANTEED ...  
★★★★☆ (133)  
\$8.00

# Association and recommendation

# NETFLIX



# Association and recommendation

You may also like



Related hotels...



Hotel 41

★★★★★ 1,170 Reviews

London, England

Show Prices



Search for people, places and things

Suggested Groups

Friends' Groups

Local Groups



패션프리마켓

Open Group

패션과 관련된 모든것들을 자유롭게 거래  
28,728 members

+ Join Group

수학(Mathematics) 그룹

Closed Group

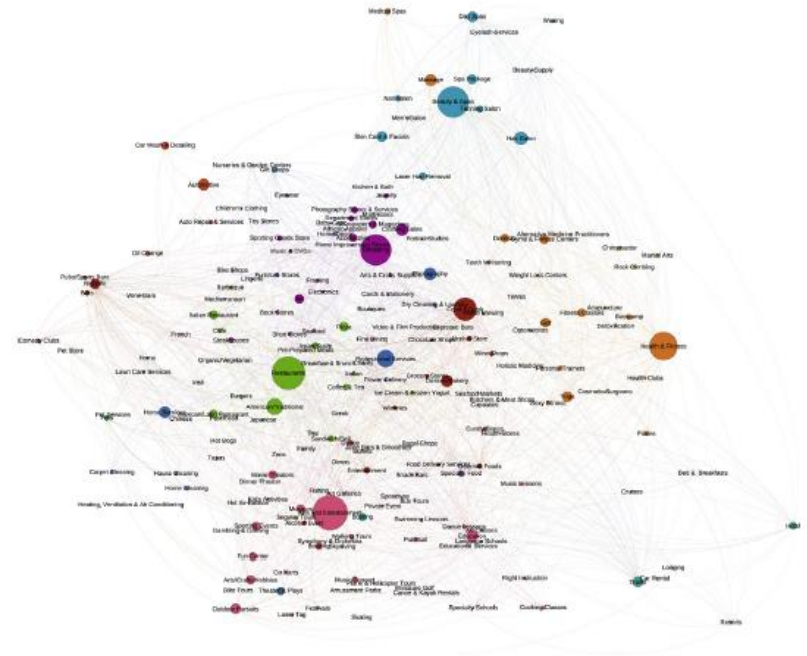
수학에 관한 모든 질문을 토론하는 Q&A 그룹  
토론규칙 http://goo.gl/K0mDBr 을 유념  
24,183 members

+ Join Group

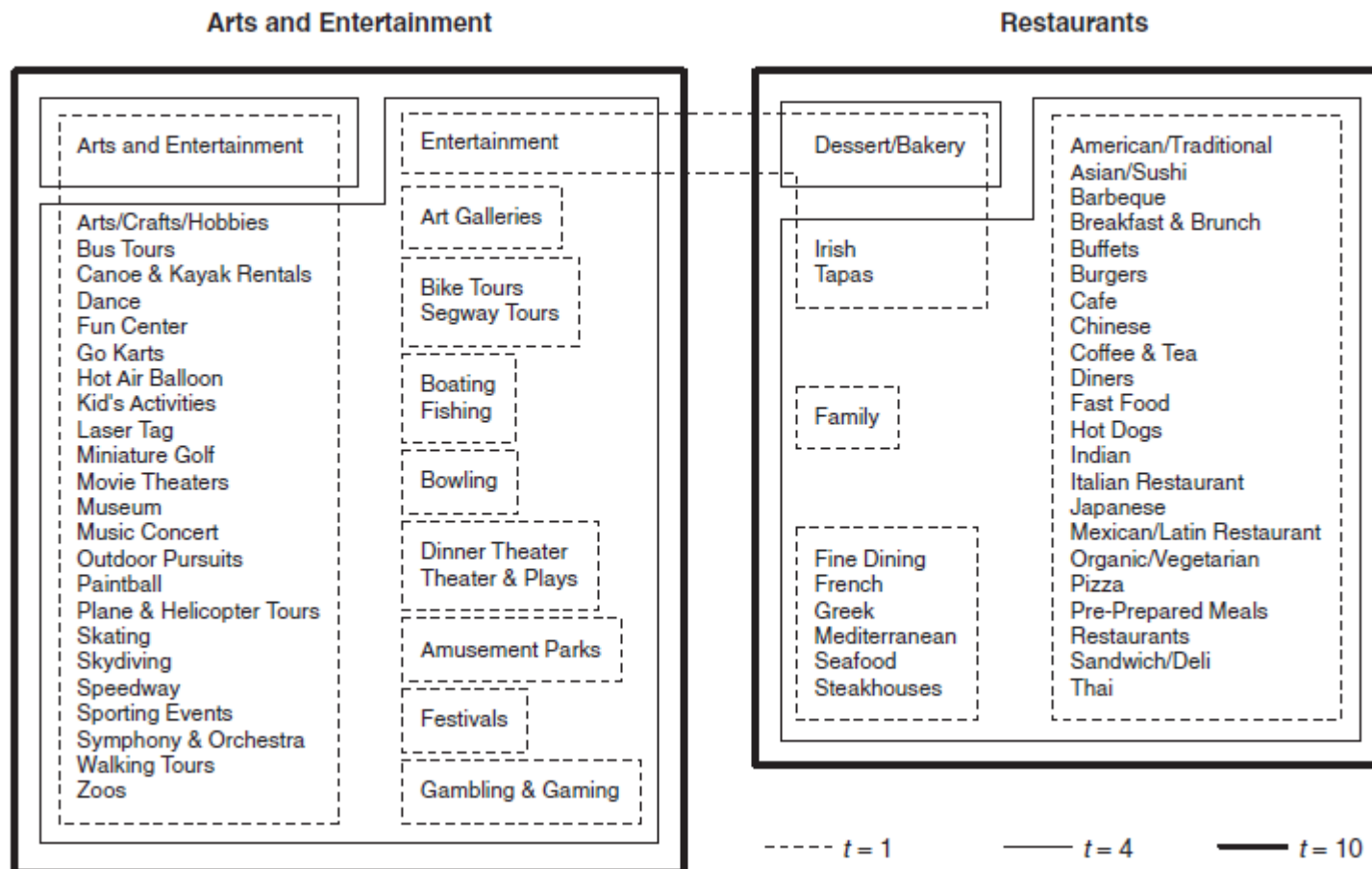


# Clustering

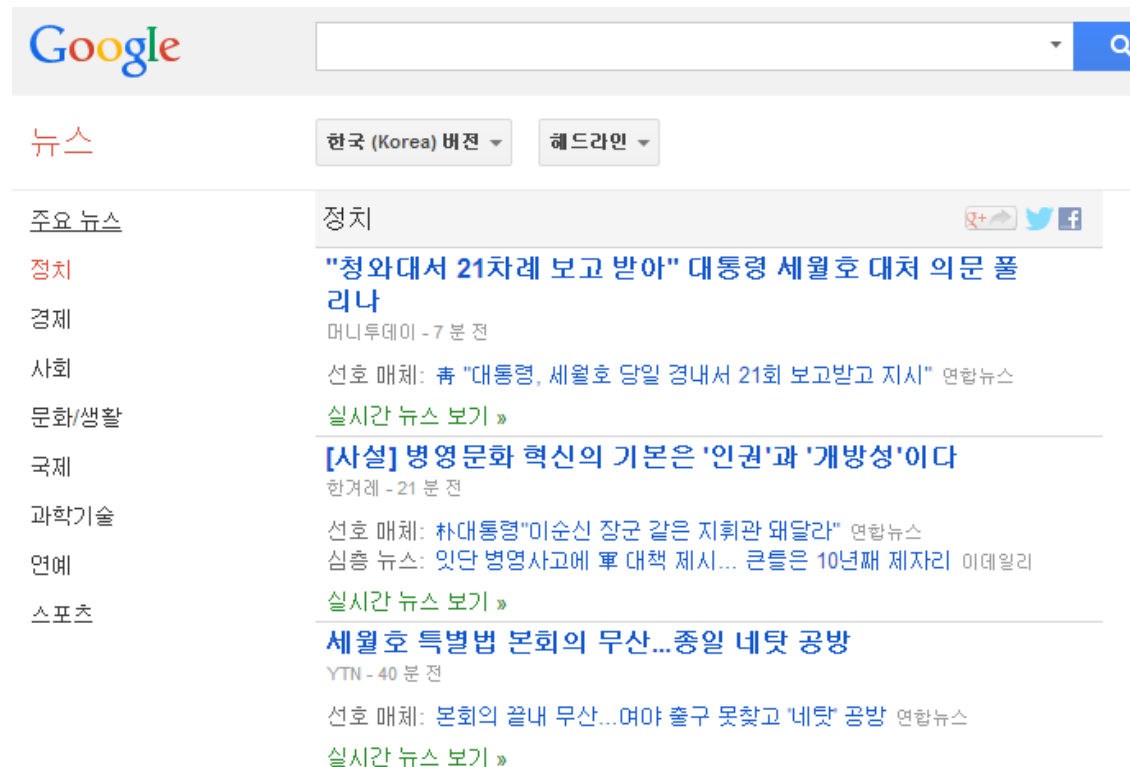
The image shows two screenshots of the Groupon website. The top screenshot is the Groupon Korea homepage (www.groupon.kr/app/index) featuring a green header with navigation links like '로그인' (Login), '회원가입' (Sign Up), and '마이페이지' (My Page). It includes a search bar and a main banner for SK-II products. The bottom screenshot is the Groupon.com homepage (https://www.groupon.com/browse/chicago) with a white header, a search bar, and a 'All Deals' section. The 'All Deals' section lists categories like 'Beauty & Spas', 'Health & Fitness', and 'Events & Activities', and features a prominent deal for 'FTD.com - Half Off Flowers and Gifts' with a bouquet of flowers.



# Clustering



# Clustering



The image shows the Google News homepage in Korean. At the top is the Google logo and a search bar. Below it, there are tabs for '뉴스' (News) and '한국 (Korea) 버전' (Korea version). The main content area is divided into sections for different news categories: 정치 (Politics), 경제 (Economy), 사회 (Society), 문화/생활 (Culture/Life), 국제 (International), 과학기술 (Science/Technology), 연예 (Entertainment), and 스포츠 (Sports). The '정치' section is highlighted, showing a headline about President Yoon Suk-yeol's 21st report to the National Assembly. Other headlines include news about the 'Ssangnok' (Ssangnok) and 'Ganghwa' (Ganghwa) and the 'Ganghwa' (Ganghwa) and 'Ganghwa' (Ganghwa).

Google

뉴스

한국 (Korea) 버전

헤드라인

주요 뉴스

정치

경제

사회

문화/생활

국제

과학기술

연예

스포츠

정치

"청와대서 21차례 보고 받아" 대통령 세월호 대처 의문 풀리나

머니투데이 - 7 분 전

선호 매체: 靑 "대통령, 세월호 당일 경내서 21회 보고받고 지시" 연합뉴스

실시간 뉴스 보기 »

[사설] 병영문화 혁신의 기본은 '인권'과 '개방성'이다

한겨레 - 21 분 전

선호 매체: 朴대통령 "이순신 장군 같은 지휘관 왜달라" 연합뉴스

심층 뉴스: 잇단 병영사고에 靑 대책 제시... 큰 틀은 10년째 제자리 이데일리

실시간 뉴스 보기 »

세월호 특별법 본회의 무산...종일 네트 공방

YTN - 40 분 전

선호 매체: 본회의 끝내 무산...여야 출구 못찾고 '네트' 공방 연합뉴스

실시간 뉴스 보기 »

# Abnormality detection

## 한국경제

2014-07-07 A19면

### 위성 영상 빅데이터 분석해 적조 예방



한국과학기술정보연구원(KISTI)과 한국해양과학기술원(KIOST)은 최근 위성 영상의 빅데이터 분석 기법을 활용해 한반도 연안 적조를 찾아내는 시스템을 개발했다.

# Abnormality detection

**MK 뉴스**

2014.08.11 15:17:15

## 해킹 막는 `빅데이터 보안` 뜬다

홈피 이용자 급증 포착해 다운차단 등 솔루션 개발

사진 파일 형태로 보관된 휴대폰 가입신청서, 신제품 설계도면, 직원 이력서 등을 빼돌리거나, 유출한 CCTV 영상데이터를 기반으로 경쟁사가 매장 내 고객 구매패턴을 빅데이터로 분석하는 식이다. 날로 지능화되는 해킹 위협에 대비해 보안업계도 빅데이터를 응용한 암호 기술을 대거 도입하고 있다.

### 빅데이터 보안 솔루션 사례

업체	내용
보메트릭	글로벌 기업 최초 국내 빅데이터 암호모듈인증 획득
SK C&C	자회사 인포섹과 빅데이터 기반 통합보안로그분석 플랫폼 구축
시만텍	빅데이터로 소프트웨어 사용 패턴 분석
IBM	국내 업체 SGA와 보안 기술 개발



# Other Application Areas

Smart Healthcare



Multichannel sales



Finances



Log analysis



Public safety



Control of traffic



Communication



Search



Manufacturing



Trading



Fraud detection



Customer mgt



# Applications of Big Data

- Applied in marketing by customer analysis
- Predict in customer behavior by its pattern
- Combine inside data with outside data
- Manage risk factors by real-time analysis

# Other cases of Big Data

## ■ [www.decide.com](http://www.decide.com)

- Predict the prices of electronics
- Merged by Ebay in Sep. 2013

## ■ [www.meteo-logic.com](http://www.meteo-logic.com)

- Forecasting of energy, weather



## ■ Renaissance Technologies: [www.rentec.com](http://www.rentec.com)

- Investment management by statistical and mathematical methods

**MK 뉴스**

2014.08.07

### 미다스의 손으로 부상한 노정석

2008년 스타트업 구글에 매각... 또 창업해 1년반만에 400억 대박  
美광고사 텃밭에 파이브락스 매각

### Customer targeting by game data analysis

파이브락스는 모바일 게임 사용자 그룹을 세분해 주요 그룹을 파악하고 이들의 행동 패턴을 분석해 고객사의 타겟 마케팅을 돕는다. 이 회사는 자체 개발한 '코호트 분석'을 사용하는데 사용자들 앱 구매 여부, 스마트폰 종류까지 나눠 분석한다.



# Implications and what to do

- Infra of Big Data is a starting place; however it should not be the goal itself.
  - Consider ROI(return of investment)
- Need to come up with utilization of inside data and outside data
- Setting a goal is important
- Basic statistical analysis and data mining is important