

# 빅데이터 분석

2012. 2. 28



1. Big Data?
2. Big Data & Biz
3. Big Data Analysis
4. Saltlux



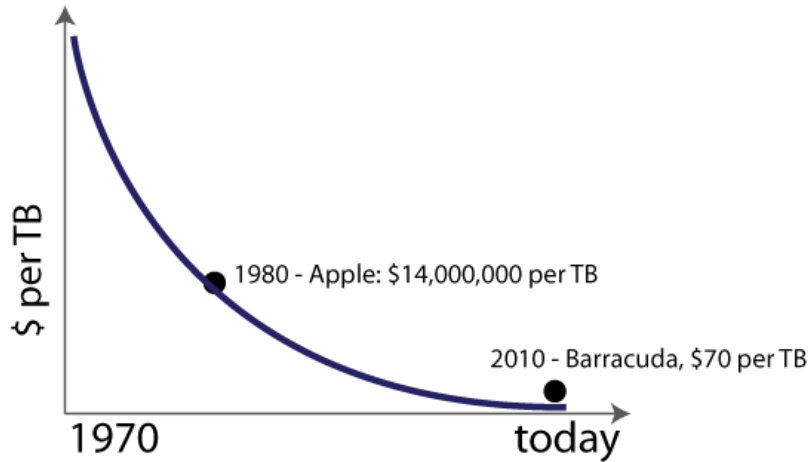
# 1. Big Data ?

↓ Big Data ↓

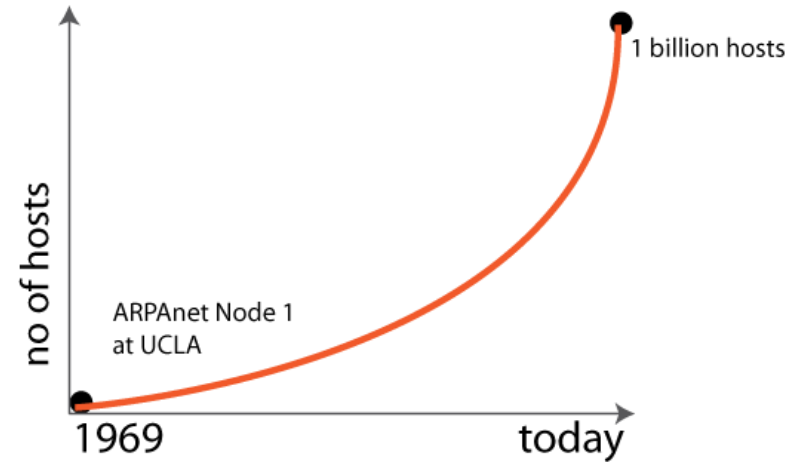


# Why Big Data?

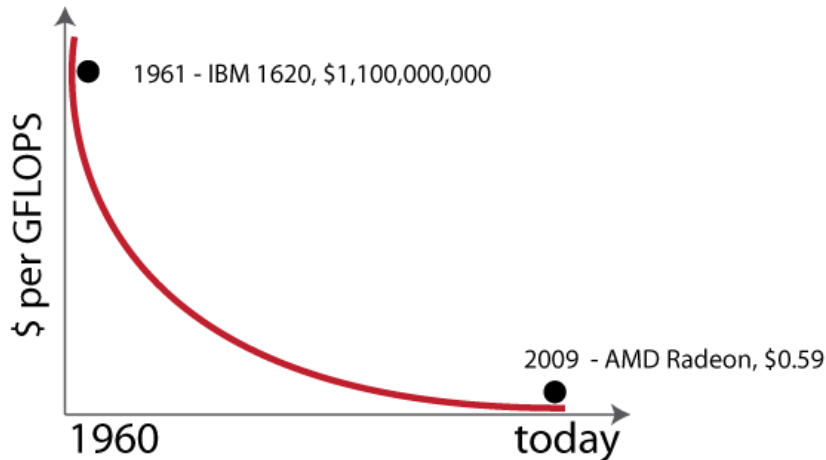
## storage cost



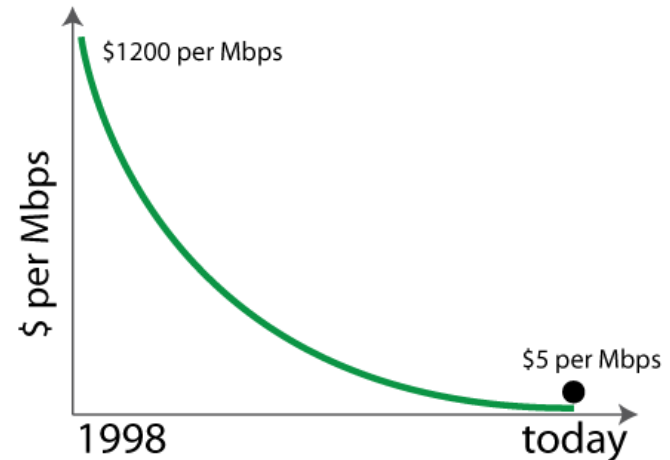
## network access



## CPU cost

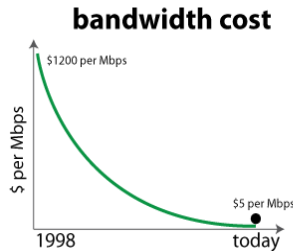
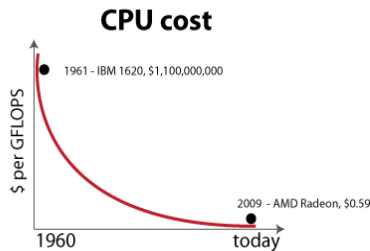
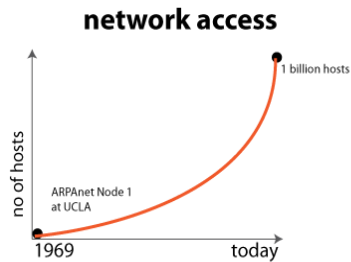
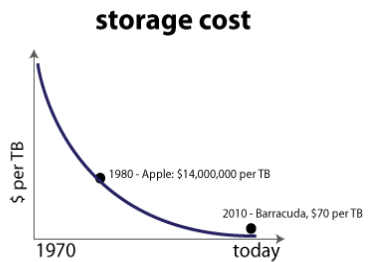


## bandwidth cost



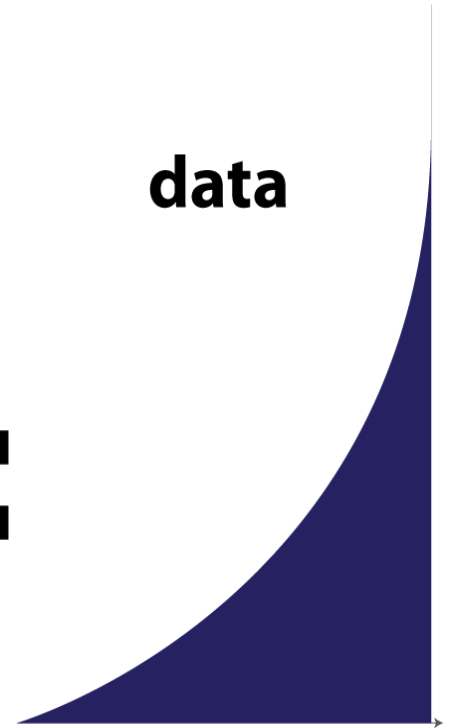
데이터 저장, 관리 비용이 낮아지면서 데이터는 폭발적으로 증가 추세

f (



) =

data



Source: Mike Driscoll, CTO Metamarkets: The Three Sexy Skills of Data Scientists (& Data Driven Startups)

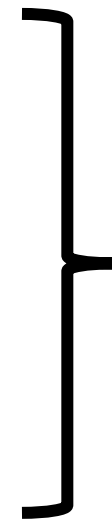


It's

too Huge ,

Fast and

Heterogeneous



Big Data

Issues

(3V)

to understand and utilize them.

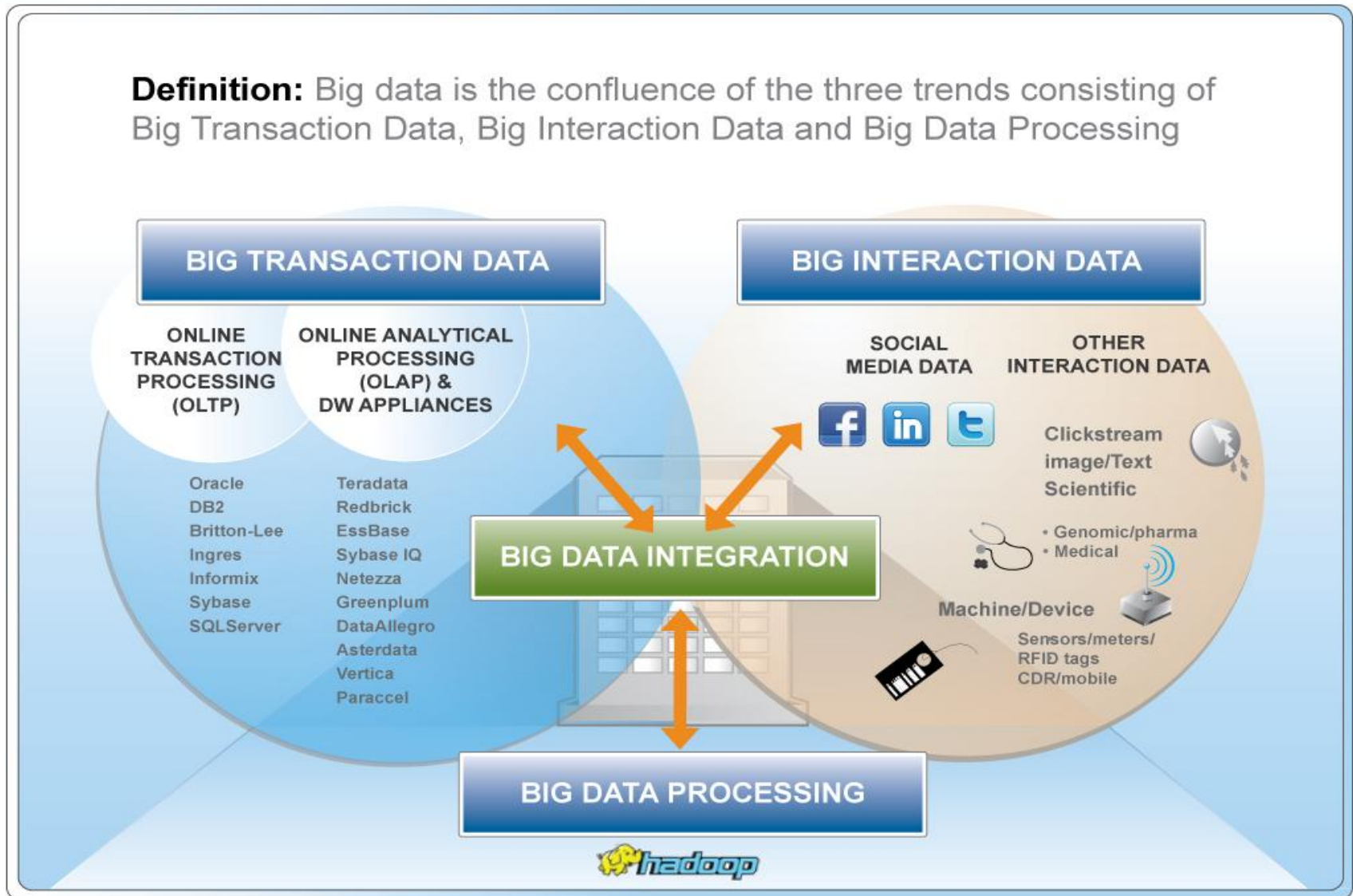
정형  
데이터

My overall experience was positive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Please complete the section below if your contact with us involved permitting/licensing/registration assistance.</b>					
The regulations were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The application instructions were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The terms and conditions of the permit, license, or registration were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
<b>Please indicate the name(s) of any staff person you would like to commend:</b>					
<input type="text"/>					
<b>Comments:</b>					
<input type="text"/>					
<b>If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:</b>					
<input type="text"/>					

비정형  
데이터

# Why Big Data?

**Definition:** Big data is the confluence of the three trends consisting of Big Transaction Data, Big Interaction Data and Big Data Processing



Source: EMC



## 2. Big Data 와 Biz

5. Big Data 와 Biz



### STAGE 1



“The money is in the **hardware**,  
not the software”

### STAGE 2

*Microsoft*

“Actually, the **money** is in the software”

### STAGE 3



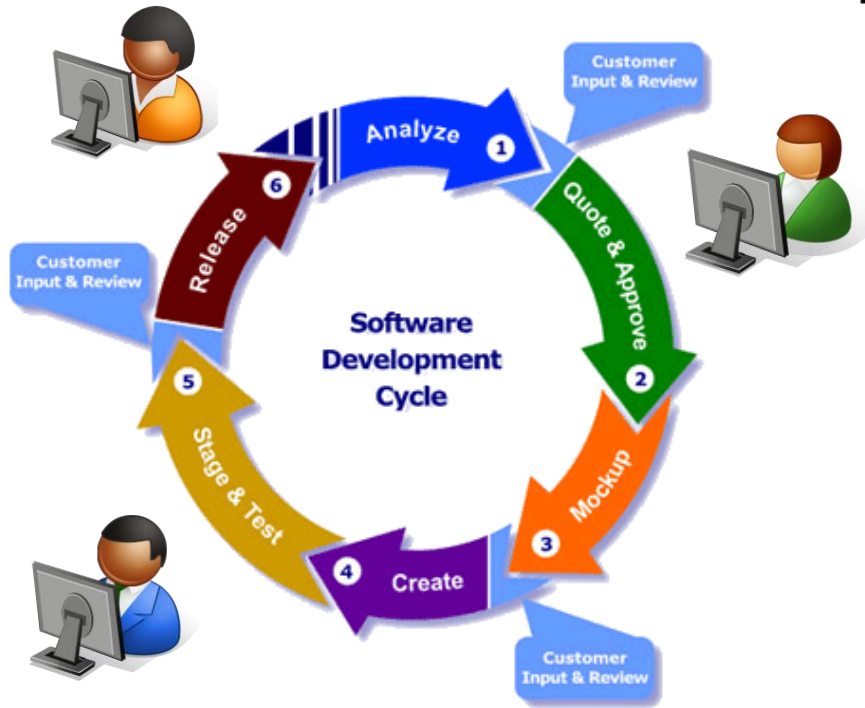
“The money is not in the **software**,  
but it is differentiating”

### STAGE 4



“Software is not even differentiating,  
the value is the **data**”

# The Age of Software



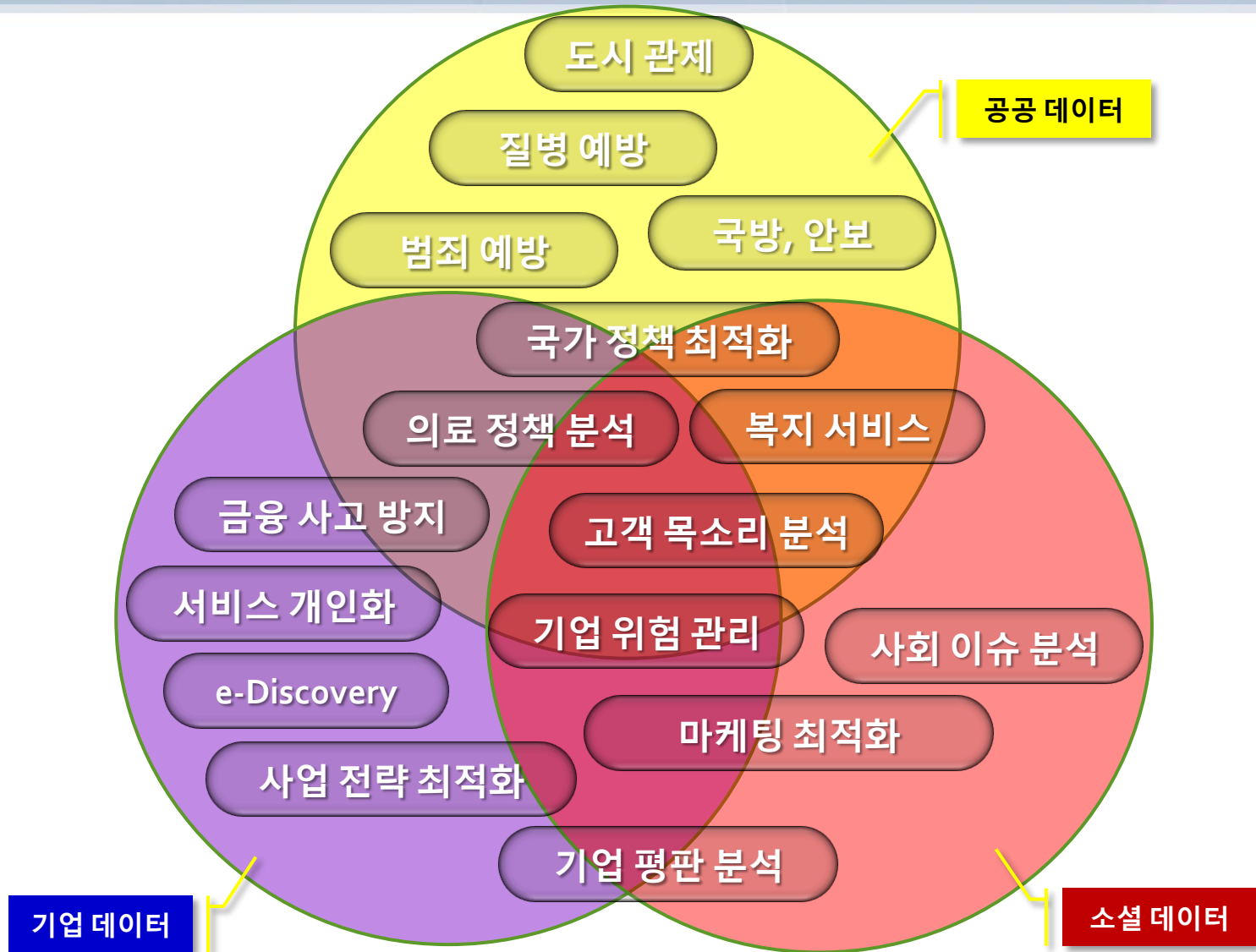
Software Expert

# The Age of **Data**

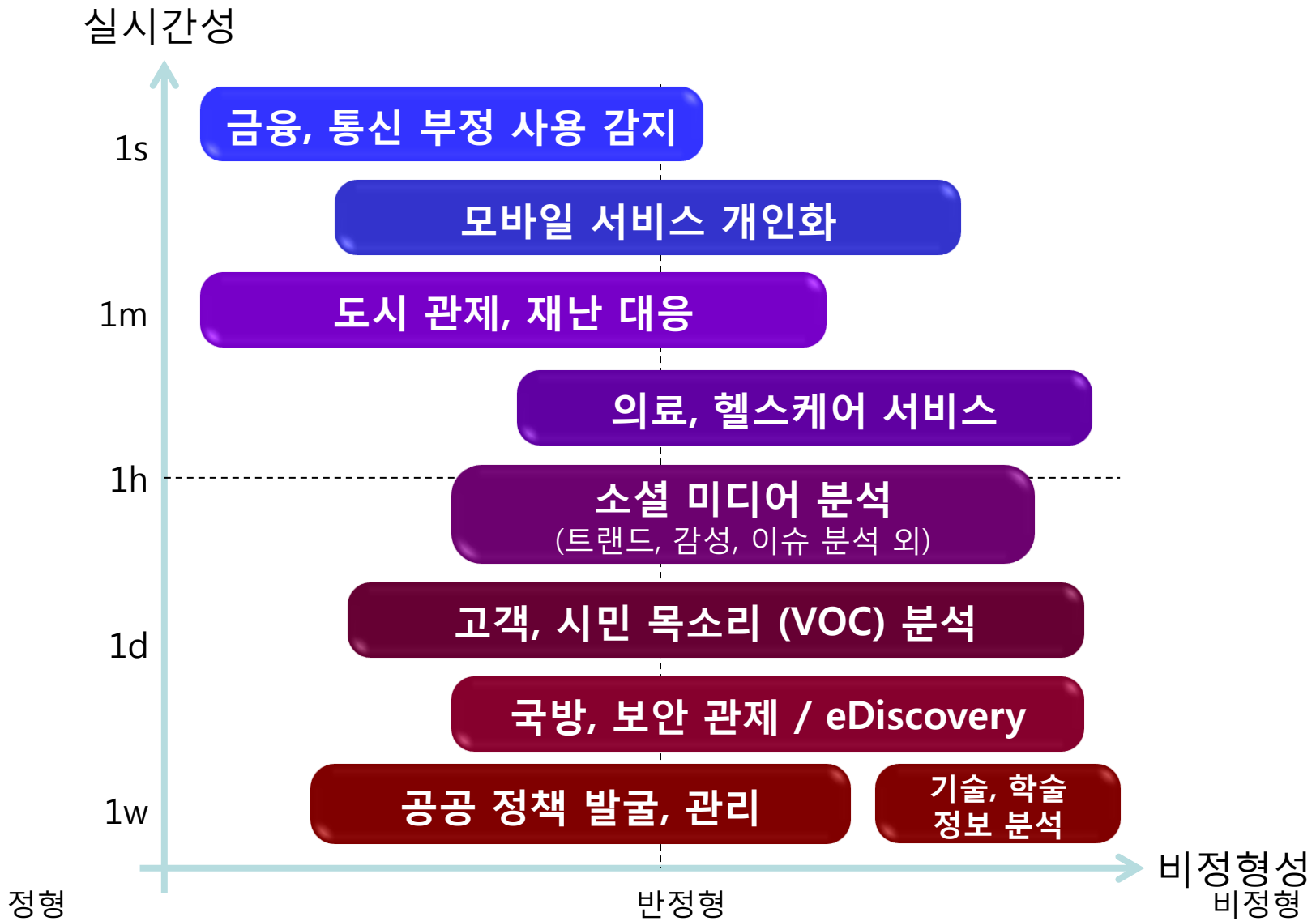


Getty

Data Scientist



# 빅데이터의 응용 영역



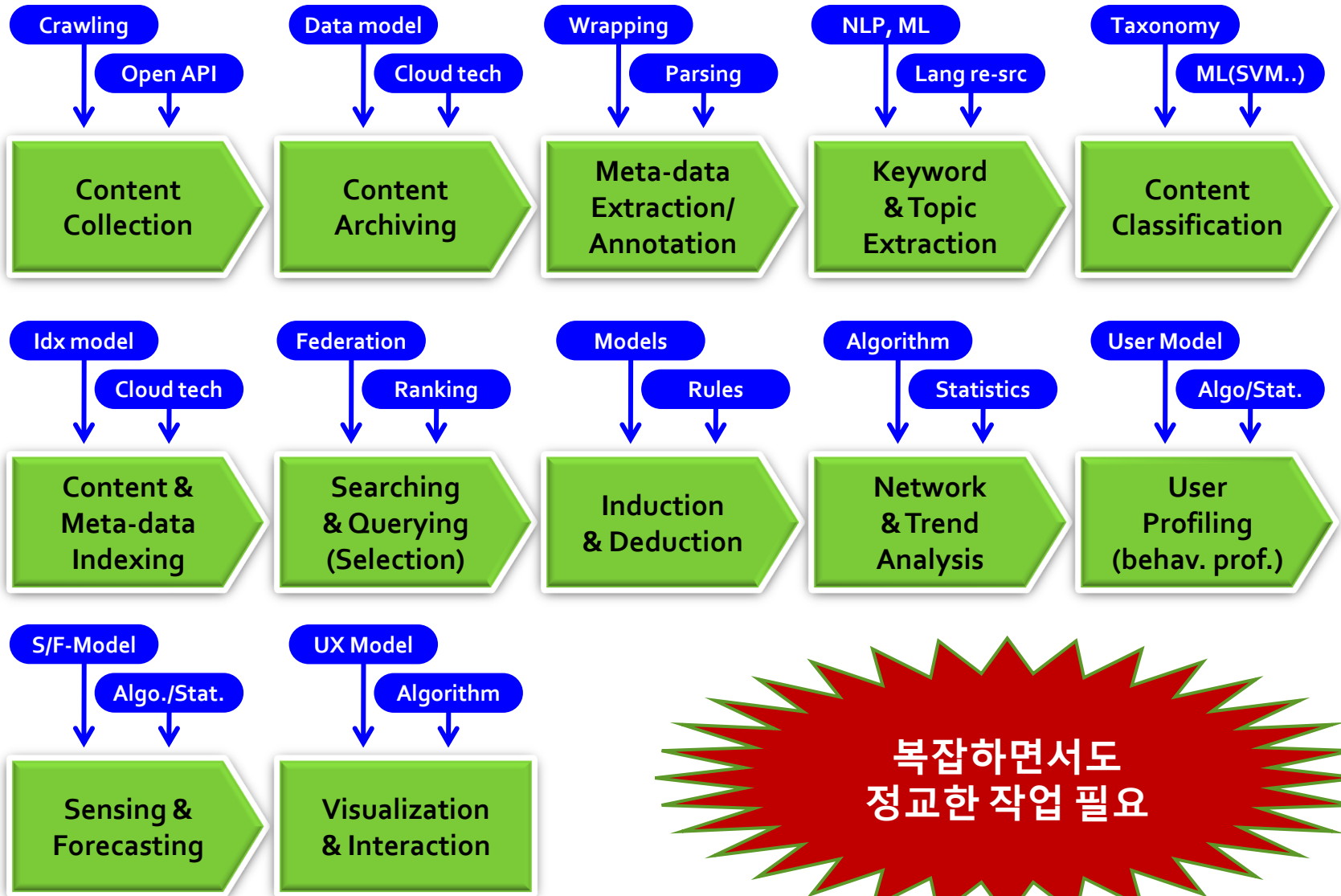


# 3. Big Data 분석

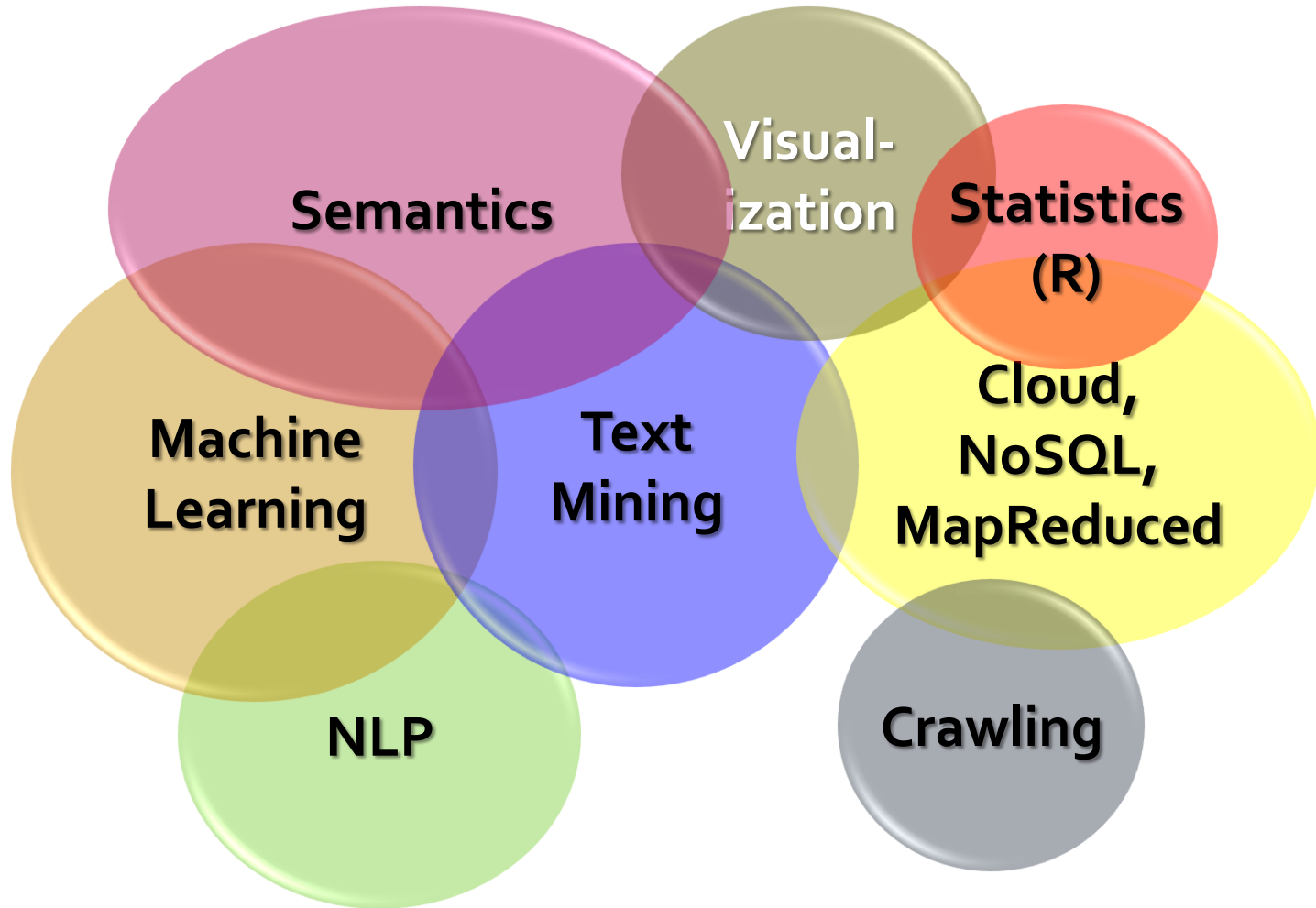
3' 빅데이터 분석



# Big Data 분석 절차



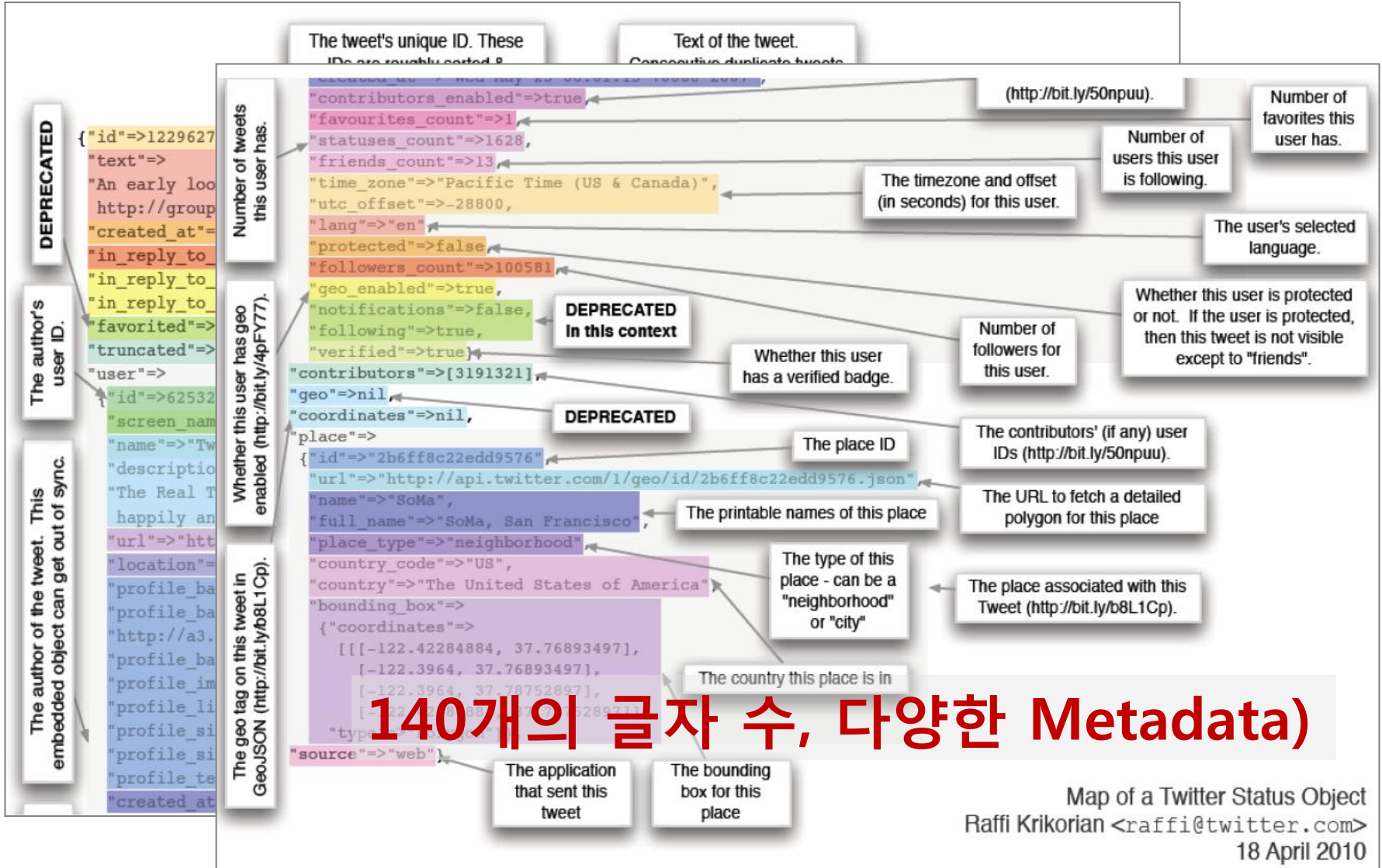
**복잡하면서도  
정교한 작업 필요**



## 필요기술: Crawling(Content Collection)

Method	News	Blog	Café	Twitter	FB	FSQ
Crawling	○	○	△	○	△	△
Feeding (RSS)	○	○	△	X	X	X
Push (Streaming)	△	X	X	△	X	X
Open API	△	X	X	△	△	△
Agent Install	X	X	X	△	△	X
Col. Interval	1hrs	6hrs	6hrs	1mins	20mins	1days
Min. Life-time	5yrs	3yrs	2yrs	1yrs	1yrs	6mons

# 필요기술: Crawling(Twitter Data)



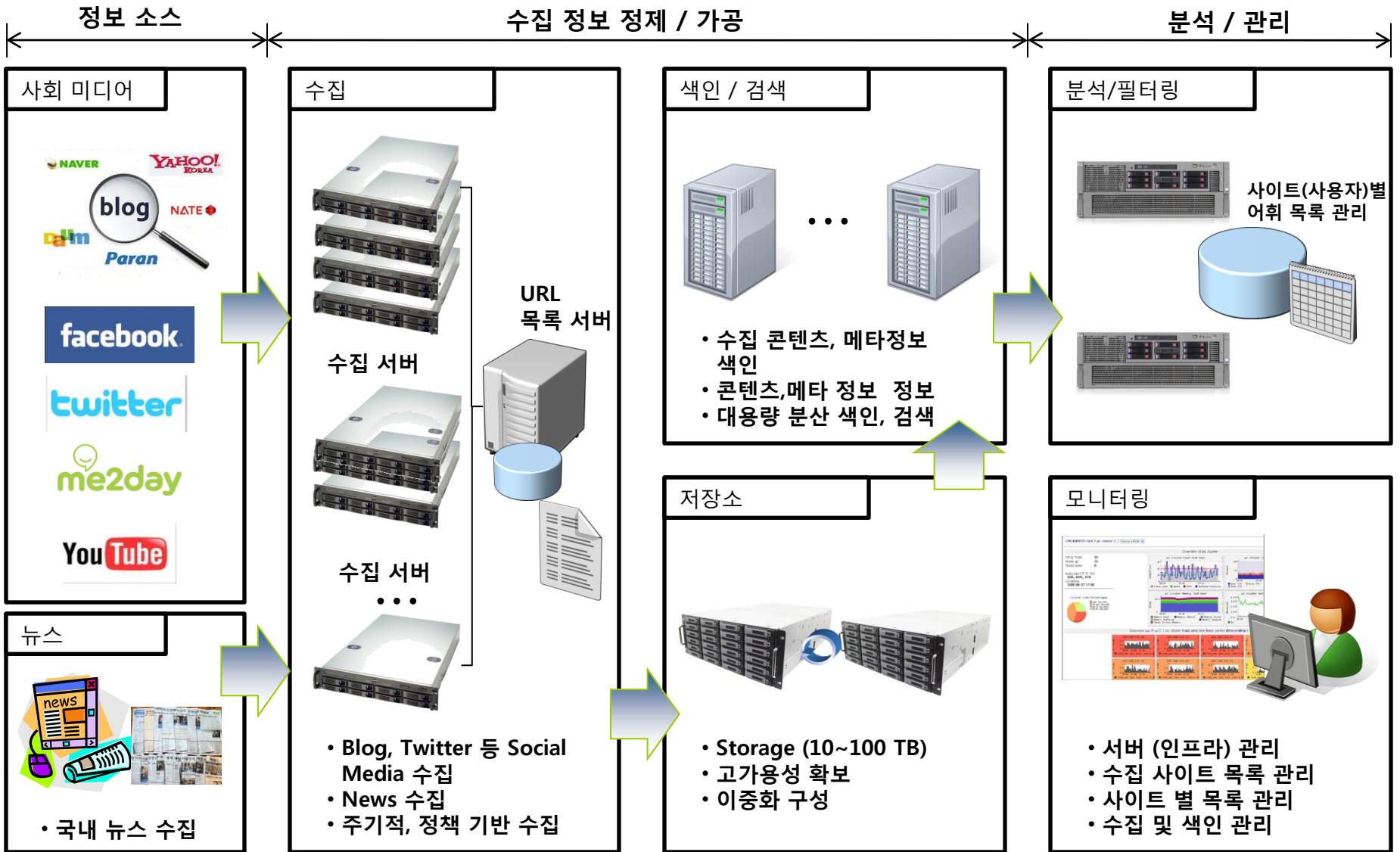


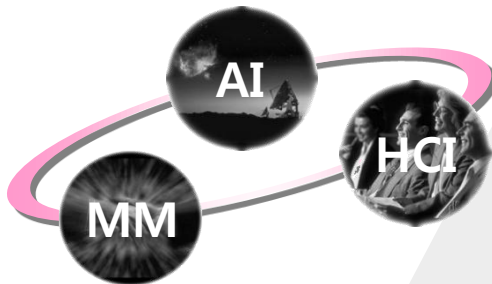
## Explicit metadata

user name, e-mail, pictures, videos, links,  
demography, group, membership, location

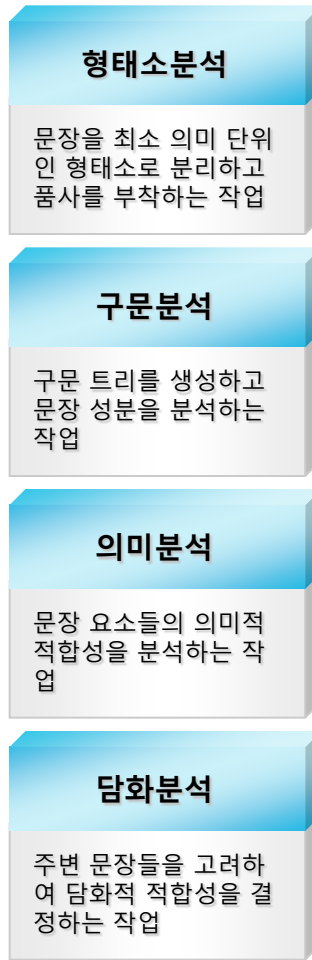
## Implicit metadata

retweets, replies, follows, comments, likes,  
page views, interests





**자연어 처리**  
 인간의 언어를 컴퓨터가 이해할 수 있도록 하기 위한 방법론을 연구하는 것



노트북을 사고 싶습니다.

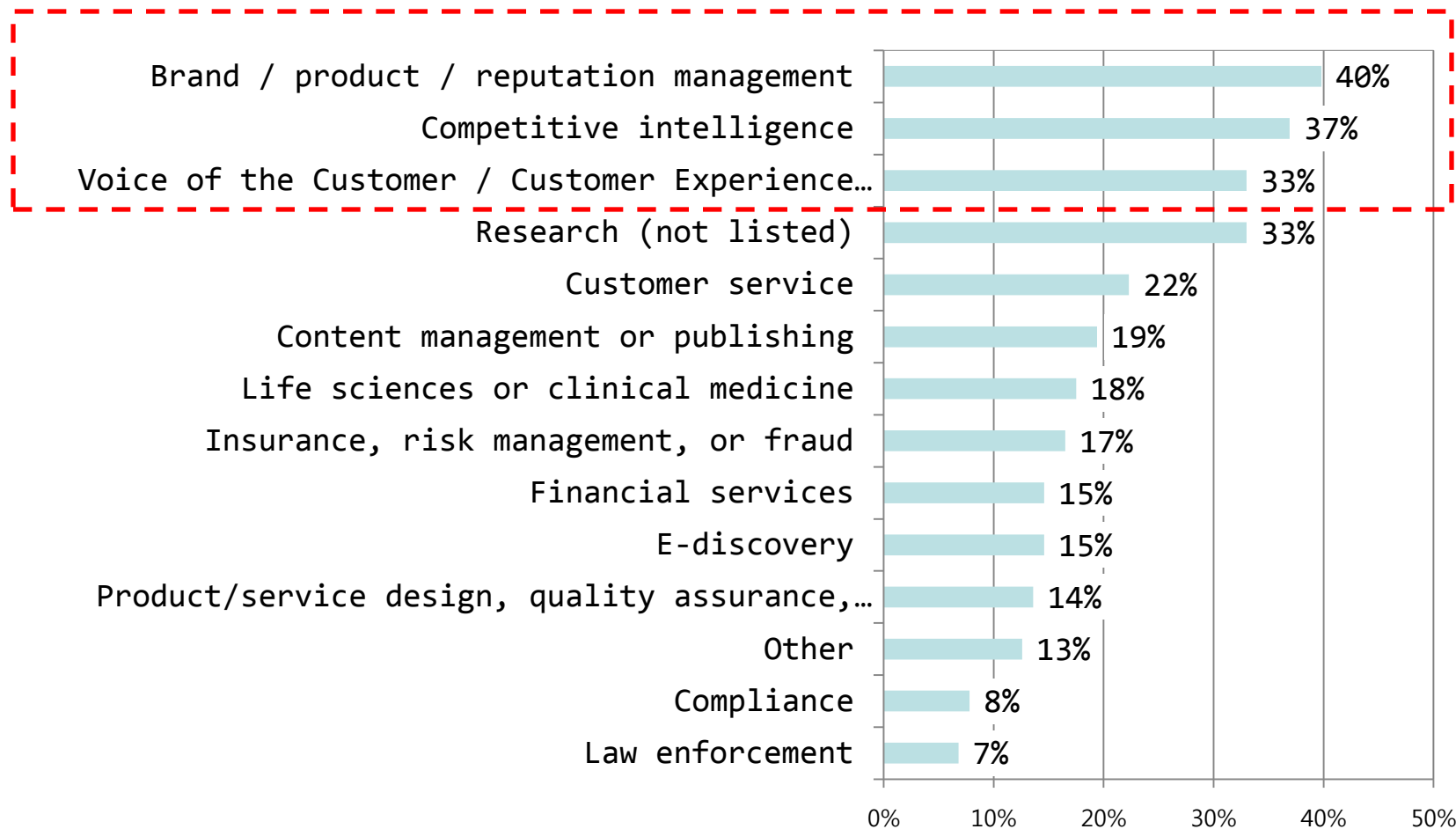
노트북/명사 | 을/조사 | 사/동사 | ...

노트북을 | 사고 | 싶습니다.

buy (modal; want, object; notebook)

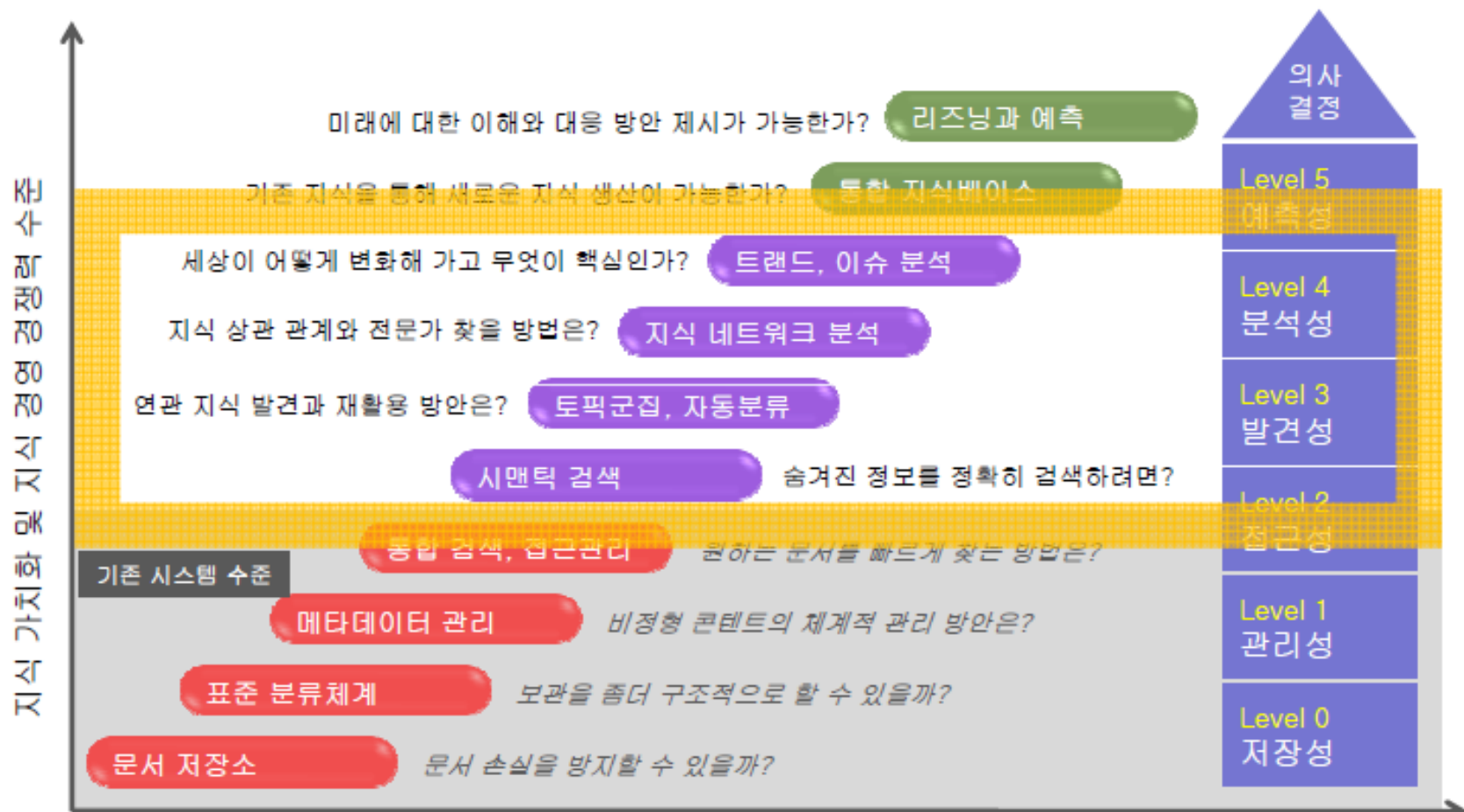
Request →  
Ask\_brand (어떤 브랜드를 찾으세요?)

## Text 분석이 핵심인 중요한 분야는?



Source: Text Analytics Summit (2011)

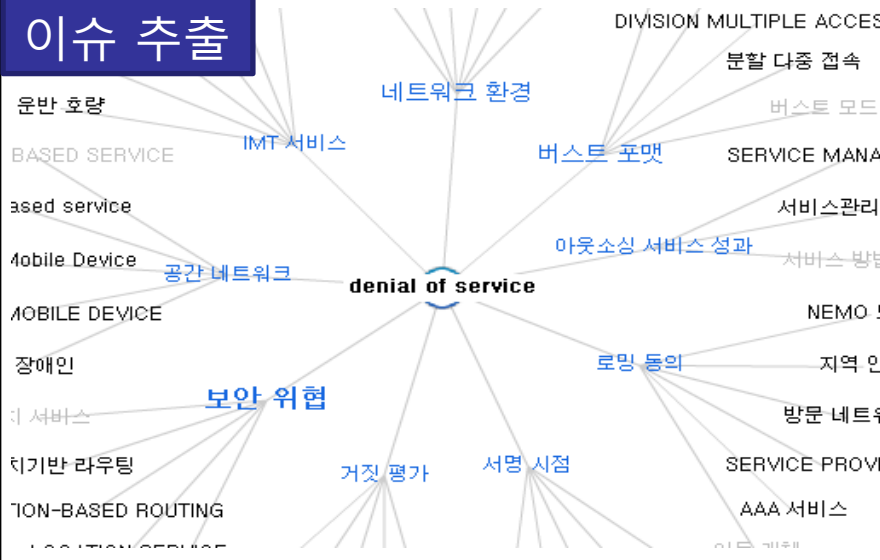
## 비정형 콘텐츠 거버넌스의 발전 단계



비정형 콘텐츠 거버넌스 및 지능화 수준



## 이슈 추출



## 자동 군집

결과 더보기 보기 (상위 200건)

기술 (111)...

- 정보, 의미, xml, ... (48)
- 검색, 사용자, 공유 (10)
- 웹서비스, 처리 (13)
- 개념, 사람, 기계, 방... (10)
- 언어, 마크, 구성 (7)
- 지능, 차세대 (9)
- 그림, 형태 (6)
- 표준화, 중심 (1)
- 프레임워크, 모델, 비즈니스 (2)
- semantic web, 통합 (2)
- 메타데이터 (2)
- OTHERS (1)

온톨로지 (29)...

정의, 관계, 사이, 실현 (5)

시스템, 요소, 관리 (11)

환경, 분석, 유비쿼터스 (13)

시맨틱 기술, 미인, 전용 (3)

## 자동 분류

- 제2편 기본규정
- 제3편 조직규정
- 제4편 업무규정
- 지침서
- 사업지식
- 업무지침
- 참고 자료실
- 공지사항
- 매스컴 및 외신 보도
- BELT사업 커뮤니티
- KOTRA경영뉴스
- 주요업무메모
- 해외무역관알림
- Great Work Place
- 노조소식
- 사우회홍치
- 버락시장
- 부서공람

부산무역관 주간업무보고 (12. 4 - 8) (22.7%)

... ○ 담당자: 김군호 부장 ○ 추진업무: 상담결과 취합 □ 동 29 ○ 파견지: 방콕, **하노이** ○ 파견규모: 10개사 내외 ○ 다. 해외전시회 □ 라스베가스 자동차부품 박람회 ○ 개...

광주, 전남 무역관 주간업무보고 (06.12.04~12.08) (22.7%)

... 체: 51업체, 68개 지사 2. 수출마케팅지원 ◇ 시장개척단 포 ~ 12월 4일 - 파견지역: 시드니, 멜버른, **하노이** - 참가업체수: 10개사 전시회 □ 모스크바 정보통신박람회 홍보 - 일정: 2007.5.14 ~ 5.18

## 자동 요약

0303\_semantic\_v2.pdf

- **주요 키워드**  
방법론, DL, 기반, DL, 기반, 엔진, Description Framework
- **간략보기**  
이 문서는 Berners Lee, CYC **방법론, DL** 기반, DL 기반 엔진, Description Framework에 대한 문서입니다.
- **중요문**  
**미래**에 대해 언급하면서, **컴퓨터**가 **디지털화된 정보**를 이해하고 **논리적으로 추론**할 수 있도록 해 주는 **Semantic Web**를 핵심으로 하는 **Web 3.0**의 구현을 위한 **기술**에 대해 소개하고 있다. 그리고 **EU 차원**에서 추진되고 있는 2010 전략에서 **Semantic Gov 프로젝트**와 **EASTWEB 프로젝트**가 추진되고 있다. **메타데이터**들은 XML(eXtensible Markup Language) 구문에 기반한 **RDF 트리플**(triple) **형식**으로 표현되며, **RDFS(RDF Schema)**와 **온톨로지**는 **RDF** 트리플 생성 시에 필요한 **클래스(class)**와 **속성(property)**을 정의하고 **계층** 관계를 설정한다. **SemanticGov 프로젝트의 목적**은 **행정(PA: Public Administration) 시스템** 기반으로 서 시맨틱 웹 서비스를 활용하는 것이다.

EDMS>CTO>SD사업부>AST그룹 | 정윤일 | 20090414 | 554KB | 조회수: 88 | 스크랩건수: 0

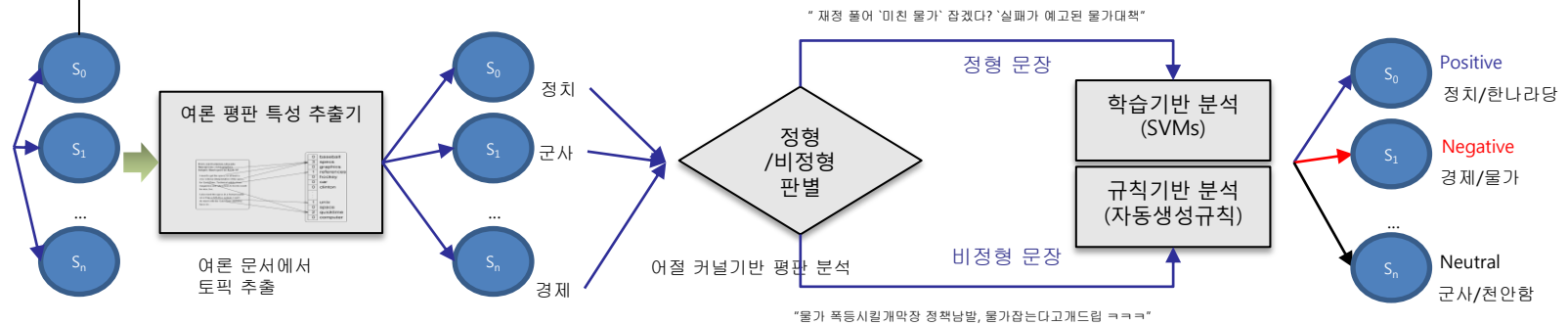
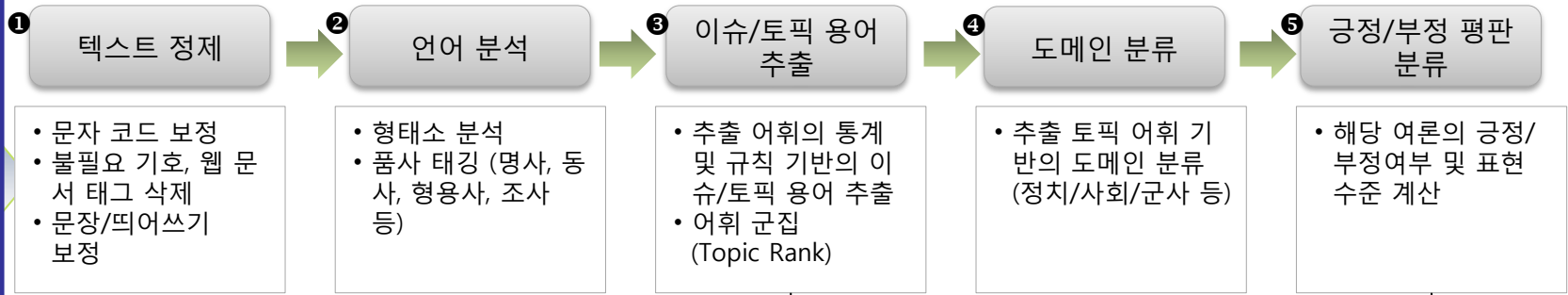
스크랩 | 닫기

# 여론 데이터 평판 분석

최신의 평판 분석 시스템은 문장에서 토픽(주제)을 추출하고, 음절 커널에 기반하여 분석하는 이단계 평판 분석 모델을 사용합니다. 정형/비정형 문장에 따라 학습기반의 SVMs 또는 규칙기반 모델로 다르게 적용하는 방식을 사용합니다.

여론 문서 집합 (블로그, 카페, 게시판)

## 여론 평판 분석 시스템



TrueStory 시즌1 정치인 Beta

트루스토리의 분석 지수는 정치인의 지도도 및 당선율과는 관계가 없습니다.

2012년 02월 01일 (현재부터 1년전까지의 분석입니다.)

세누리당 박근혜 +추가

종합  뉴스  블로그  트위터 분석하기 1 분석대상은 여러개 선택이 가능합니다.

뉴스vs소셜 비교분석



박근혜 상세정보

- 정당 세누리당
- 선거구 대구 달성군
- 소속위원회 기획재정부원회
- 당선연수 4선 (15대, 16대, 17대, 18대)

분석대상 링크 관심지수

분석대상	링크	관심지수
종합	3	6.46%
뉴스	5	6.22%
블로그	1	4.8%
트위터	5	6.08%

○ "박근혜 개혁"에 대한 국민 관심의 변화는?



○ "박근혜 개혁"에 대한 관심 키워드의 변화는?

상세보기

09월 10월 11월 12월 01월 02월

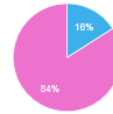
- |            |           |         |             |         |              |
|------------|-----------|---------|-------------|---------|--------------|
| 1 안철수      | 1 나경원     | 1 전 대표  | 1 BBK 의류 폐기 | 1 오영환위  | 1 사회적 약자 반대편 |
| 2 이회철      | 2 선거 판세   | 2 신우철   | 2 한나라당      | 2 한나라당  | 2 세누리        |
| 3 은지현      | 3 전 대표    | 3 권 거침말 | 3 북주        | 3 비대위원장 | 3 강금석        |
| 4 정동준이     | 4 나경원 구상기 | 4 안철수   | 4 김학산       | 4 김수경학회 | 4 안철수        |
| 5 한나라당 정동준 | 5 도가니2    | 5 사실무근  | 5 알광        | 5 문봉투 쇼 | 5 인터넷 기사     |

○ "박근혜 개혁"에 대한 연관관 비교분석의 의견은?

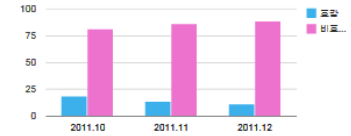
(분석대상: 트위터)

관심지수 ★★★★★ 관련 의견 표 관련 표 관련 의견의 척도입니다.

<연말 호관 지수>



<연말 호관 트렌드>



<세부 주제별 호관 지수 - 세부 주제별 고려 시 대상자 호관 지수>

	2011.10	2011.11	2011.12
나경원	도과 28% 비도과 71%	도과 8% 비도과 91%	도과 16% 비도과 86%
안철수	도과 24% 비도과 76%	도과 19% 비도과 81%	도과 17% 비도과 83%
무상급식	도과 16% 비도과 86%	도과 7% 비도과 93%	도과 8% 비도과 92%
박정순	도과 16% 비도과 86%	도과 42% 비도과 68%	도과 17% 비도과 83%
서유석장	도과 20% 비도과 80%	도과 12% 비도과 88%	도과 7% 비도과 93%
한철수	도과 8% 비도과 91%	도과 19% 비도과 81%	도과 17% 비도과 83%
한나라당	도과 19% 비도과 81%	도과 7% 비도과 93%	도과 8% 비도과 92%
FTA	도과 16% 비도과 86%	도과 42% 비도과 68%	도과 17% 비도과 83%
신당	도과 16% 비도과 86%	도과 42% 비도과 68%	도과 17% 비도과 83%
박정희	도과 20% 비도과 80%	도과 12% 비도과 88%	도과 7% 비도과 93%
최신	도과 8% 비도과 91%	도과 19% 비도과 81%	도과 17% 비도과 83%
비대위	도과 19% 비도과 81%	도과 7% 비도과 93%	도과 8% 비도과 92%
재정당	도과 16% 비도과 86%	도과 42% 비도과 68%	도과 17% 비도과 83%
이명박	도과 16% 비도과 86%	도과 42% 비도과 68%	도과 17% 비도과 83%
BBK	도과 20% 비도과 80%	도과 12% 비도과 88%	도과 7% 비도과 93%

○ 호관메시지

- 이렇게 세심하지 박근혜는 유하나 이런 분을 비대위원으로 모셔야
- @GH\_PARK [누 한나라당] 한철수 조류현 이재우가 비대위
- 송리의 정@mikban0108: @anbjkxt: 새해원 모은
- 현재 한나라당과 민주당의 지도부 무게 비교를 권해 드립니다. 한
- @GH\_PARK [가소롭다] 정이계가 일처리를 한나라당을 박근혜
- @GH\_PARK [최신] 박근혜 위원장은 <국민이 납득할만
- RT @sukhe212: RT @ak40ey: "박근혜"
- @GH\_PARK [박근혜 세심은 성실] 박근혜 비대위가 강력한
- 한나라당 박근혜 비대위원장은 물론 "자다가도 새벽에 깨다. 생각
- 같은인 비대위원장이 안철수 등장 대충이라도 모두 분명, 박근혜

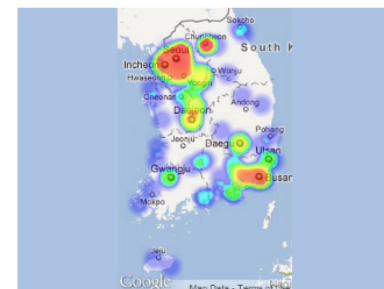
○ 내오관메시지

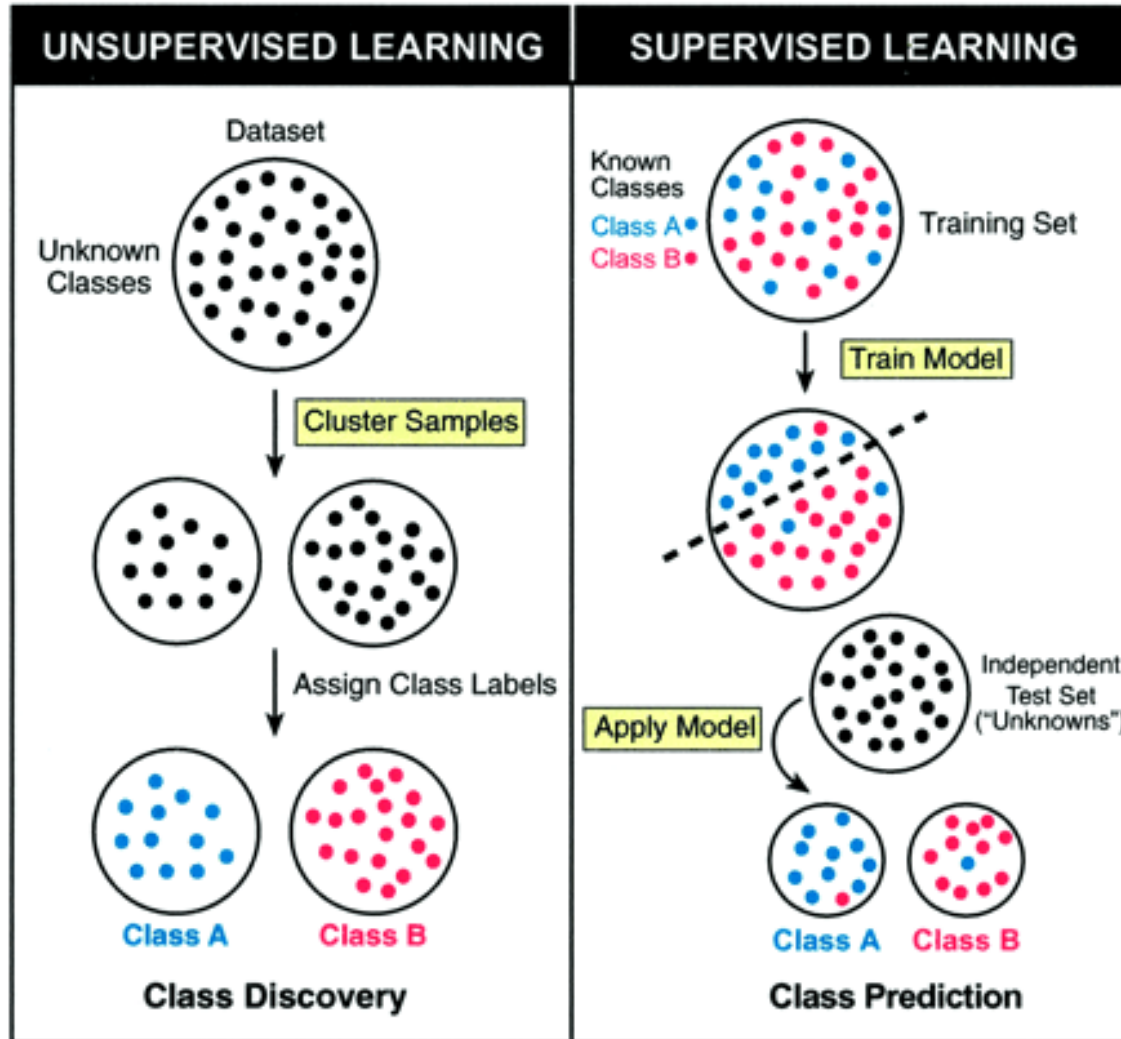
- RT @chojungdong444: 한나라당이 지금 박근혜 앞세
- 박근혜한나라당비대위원장! 지금같이 있는일이목적이지 세심이라면?
- @dogeul 박근혜라는 이름이 대중주자로 음모는거 지않가 글
- 음으로 관심주자로 하고 누구한테 승낙할줄 하나?? RT @n
- RT @twobentline: RT @mettayocon: 한나라당
- RT @mettayocon: 박근혜 "문봉투, 구리한 정치로통론
- RT @du0290: "그건 얘기할게 없다. 여기까지 와서 난무
- 문봉투로 살아 왔으면서 무슨 구만 RT @newface21 박
- 순석회 둘다 환경 불타진, 풍사역 은 "2007년 경선당시 박근
- RT @mangchubun: 박희태문봉투사건 터드린 이유는 박

○ "박근혜 개혁"에 대한 연관관 키워드



○ "박근혜 개혁"에 대한 지역별 관심 분포





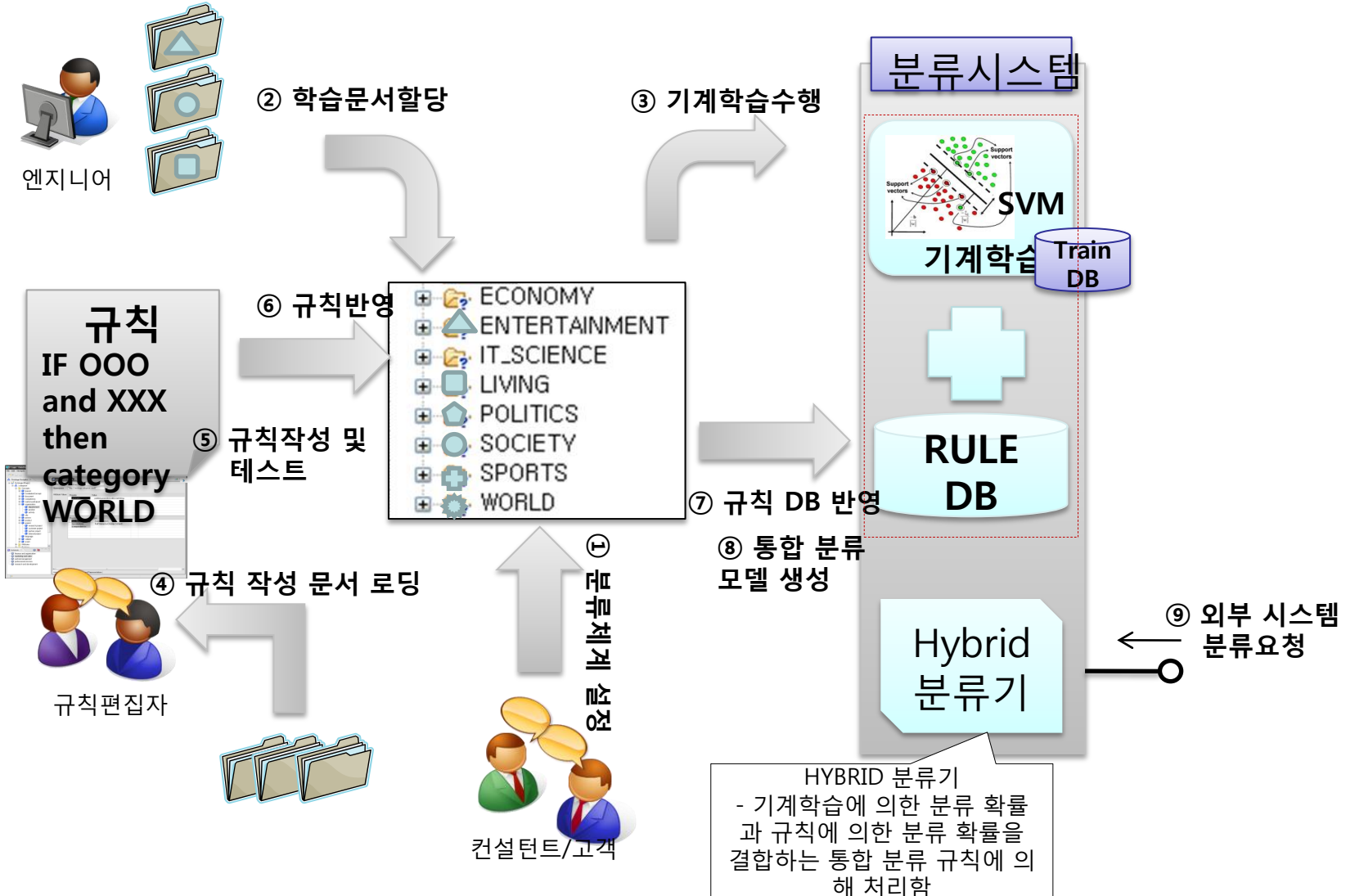
사용자의 개입  
불필요

군집(Clustering)

사용자의 개입  
필요

분류(Classification)

# Saltlux – 하이브리드 분류기 적용 프로세스





# Saltlux – 하이브리드 자동분류

**1** 분류체계 설계

**2** 기계학습

ID	Created date	Title	Author
319	1-12월-2009	조영교외국어학원에 오시기를 환영하... N한스한국어지하오스 - http...	saltlux
318	1-12월-2009	외국어 교육 N Intensi	회화 관련...
317	1-12월-2009	'98년 1월 영어 일어 강	외국어...
316	1-12월-2009	광송지도 Words N Bc	...
315	1-12월-2009	아바토 소개 N 아바토	류&가격 ...
314	1-12월-2009	아바토 강의 N 12월 28일 N 1. 아바토 영어관 N 2. 독해(예제) N Why is t...	saltlux
313	1-12월-2009	제 79회 - 1999년 6월 14일 N I'll put in a good word for you. N 사회...	saltlux
312	1-12월-2009	TOEIC의 개요 N TOEIC(Test of English for International Communi...	saltlux
309	1-12월-2009	어학센터는 학생들의 어학교육을 위하여 다양한 어	
308	1-12월-2009	어학코스 원서작성법 N 1. 필요 서류 N 입학원서 사	
307	1-12월-2009	어학코스의 교육과정 N 다양한 어학코스의 교육과정	
306	1-12월-2009	초급영어회화 메뉴 N EBS 라디오 초급 영어 회화	
307	1-12월-2009	Frequently Asked Questions N 생활편 질문과 답	

Total positive documents in training set: 49, total page: 1, current page: 1

**3** 학습문서

**4** 문서확인

\* TOEIC의 개요 TOEIC(Test of English for International Communication)은 우리 잘 알려진 TOEFL 시험 을 출  
 Testin...가 사어...이 국제적 공용어로서의 영어 숙 달정도를 측정하기 위해 개발한 시험이다. 1979년  
 우리... 도입되어 1,379명이 처음 응시한 것을 시작으로 1995년 현재에는 30만명  
 형은... 시되는데, 정기시험은 매월 시행되며 그 결과는 ETS에서 직접 채점하고  
 에 행... 로 기 업체에서 신입사원 선발, 인사고과 반영, 해외파견 연수자 선발 등  
 에 따라... 40인 이상을 경우 수시로 실시되고 있다. \* TOEIC 점수 현황 1993년 상반기(1-6월)에 총  
 22,201명(22,201명)이 TOEIC 시험을 응시하였고 그 중 15,717명(70.8%)은 정기시험을 응시하여 그 점수 평균

Refining category  Alpha 0.8 Threshold 0.0

```

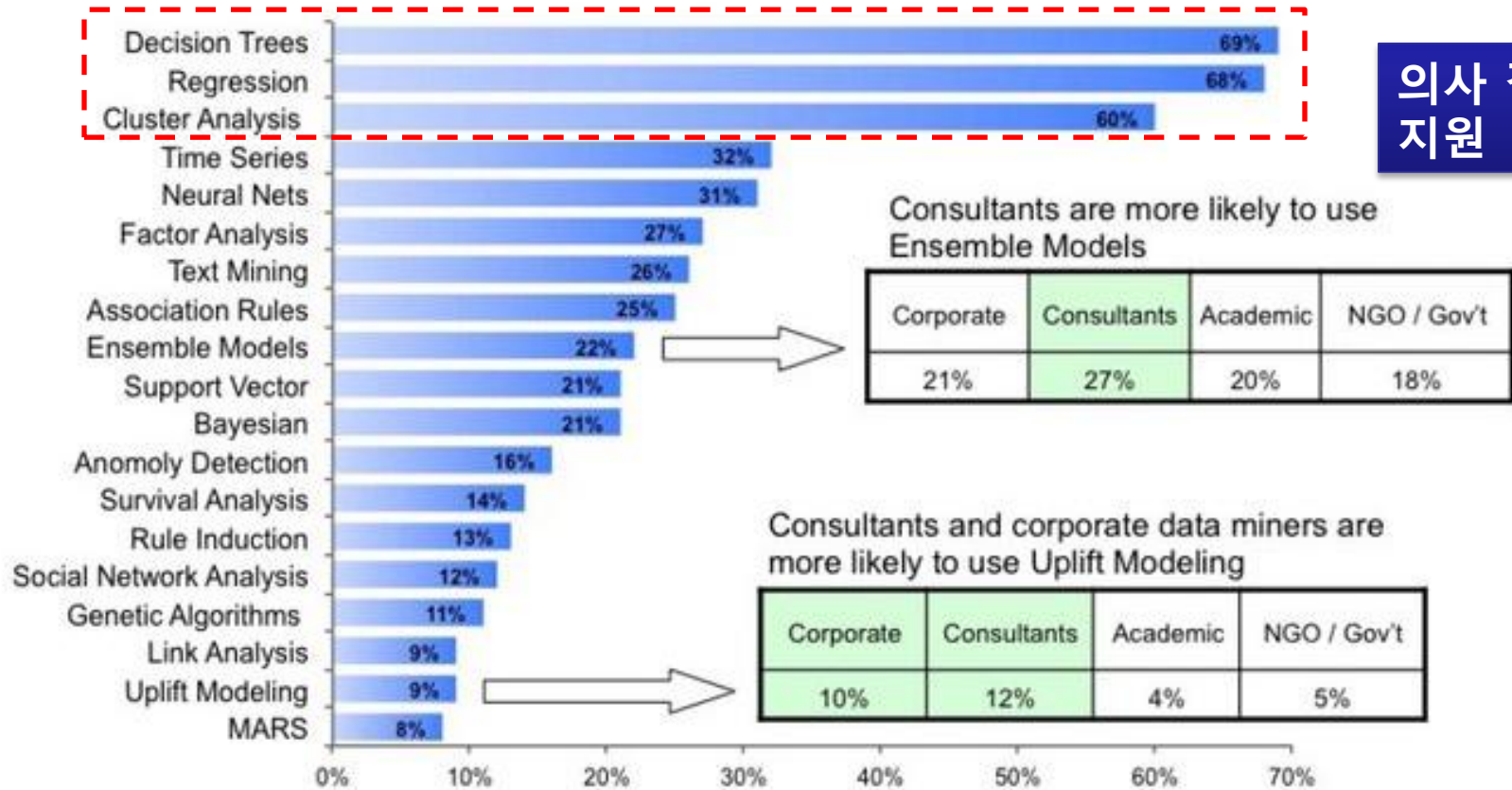
1 (
2   DOCUMENT <CONTAIN> culture
3   <OR>
4   SVM_SCORE > 5
5   <OR>
6   AUTHOR <EQUALTO> Keynes
7   <OR>
8   AUTHOR <EQUALTO> hello
9 )
10

```

Education hello world

규칙편집

- Decision trees, regression, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been very consistent over time.
- However, a wide variety of algorithms are being used.



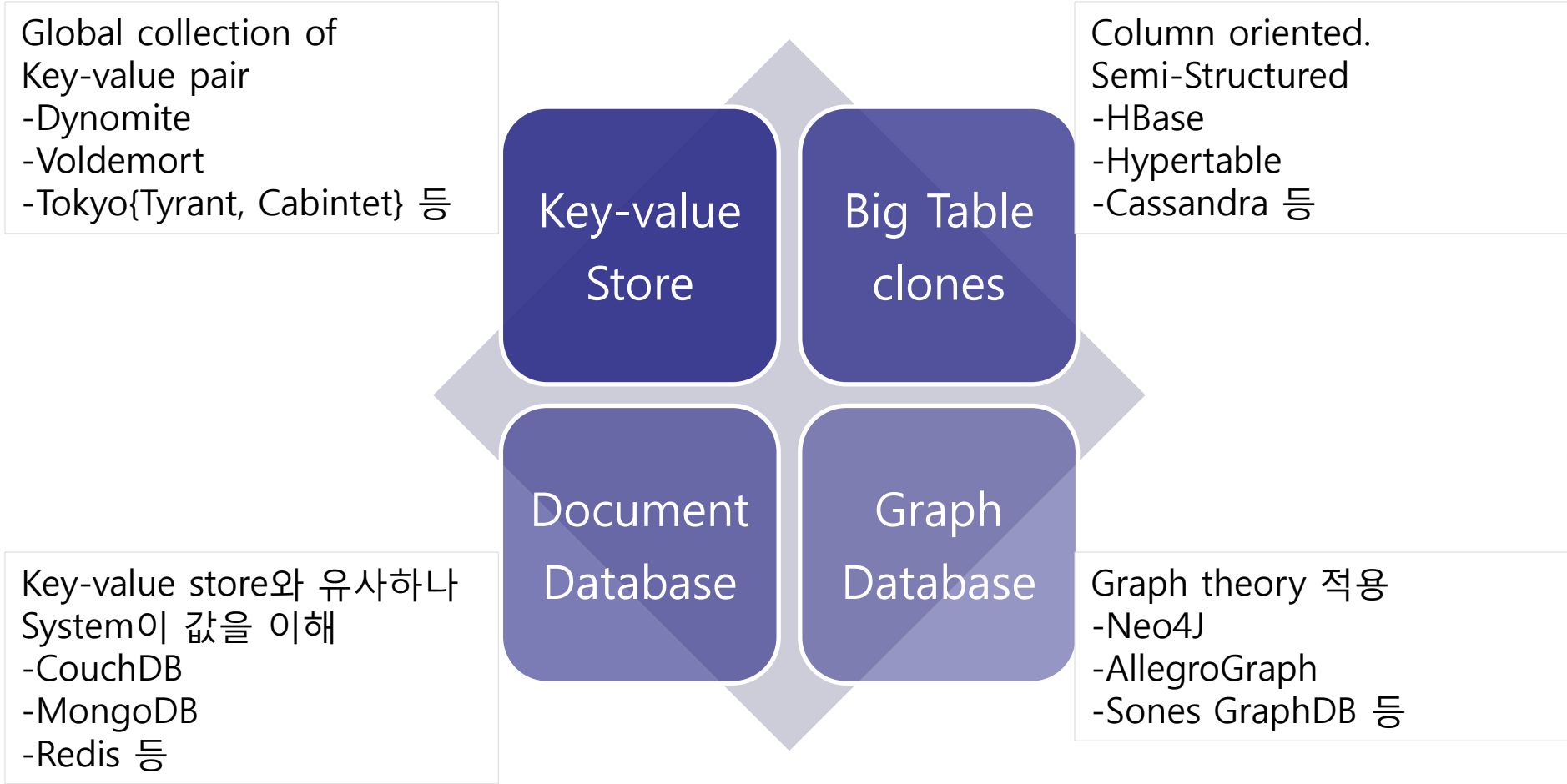
의사 결정  
지원

Question: What algorithms/analytic methods do you TYPICALLY use? (Select all that apply)

Vendors were excluded from this analysis.

Source : Lexer Analytics (2011)

# 필요기술: NoSQL (Scalability)

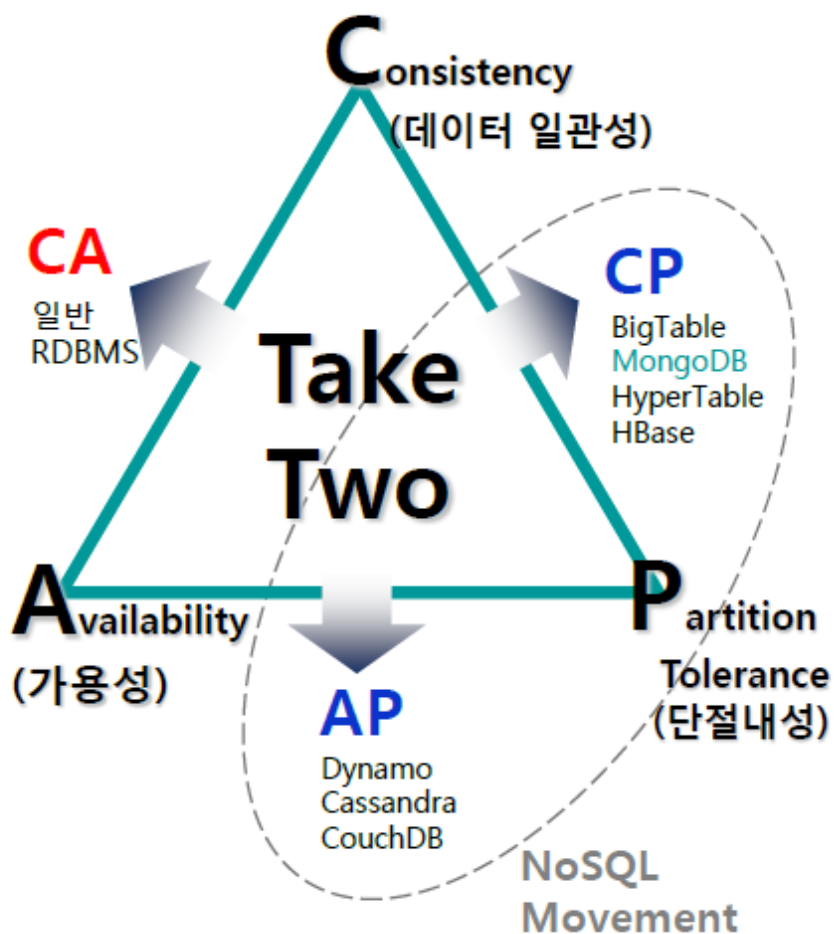


Source : Tobias Ivarsson (2010)

## 필요기술: NoSQL(Indexing / Archiving)

대용량 분산 데이터 저장소는 데이터 일관성(Consistency), 가용성(Availability), 단절내성(Partition Tolerance)을 모두 만족시키는 것이 불가능하므로 두 가지만 전략적으로 선택해야 한다는 이론

### CAP Theorem과 데이터 관리전략

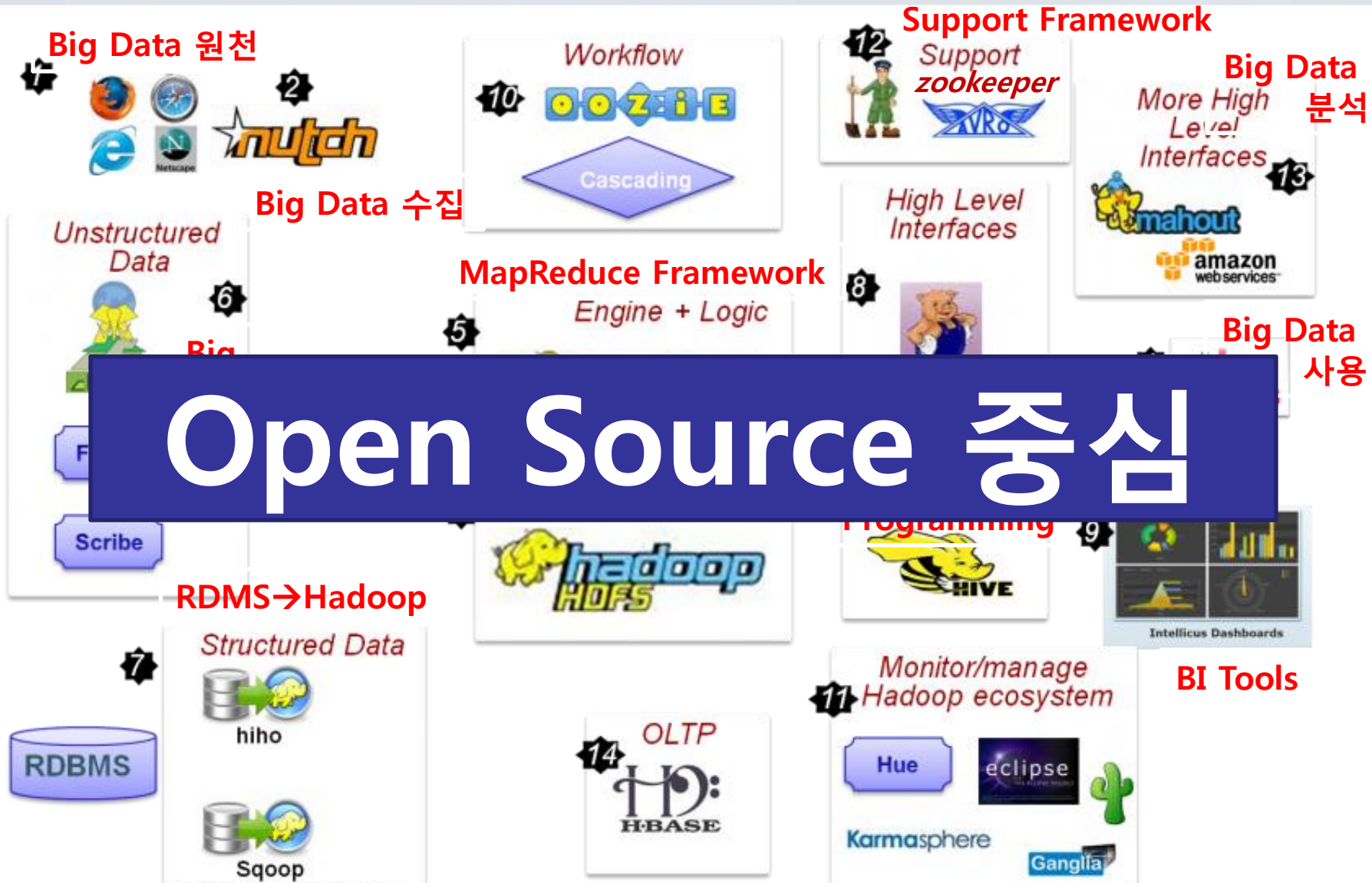


### CAP Theorem의 시사점

- 초고용량 데이터의 효율적인 처리 (비용, 성능 등)를 위해서는 수평적 확장이 가능한 네트워크 기반 병렬컴퓨팅 시스템이 필요함
- 분산/병렬 시스템의 기본인 Partition Tolerance를 중심으로 한 Database Model을 선택해야 함.
- 또, 비즈니스 데이터 처리 유형에 따라 CP 또는 AP 기반의 적합한 Data Model과 store 기술선택이 중요
- 분산파일시스템기반의 분산DB(NoSQL DB) 아키텍처 구성필요



# 필요기술: Big Data (Scalability) → Hadoop



Source : <http://indoos.wordpress.com/2010/08/16/hadoop-ecosystem-world-map/>

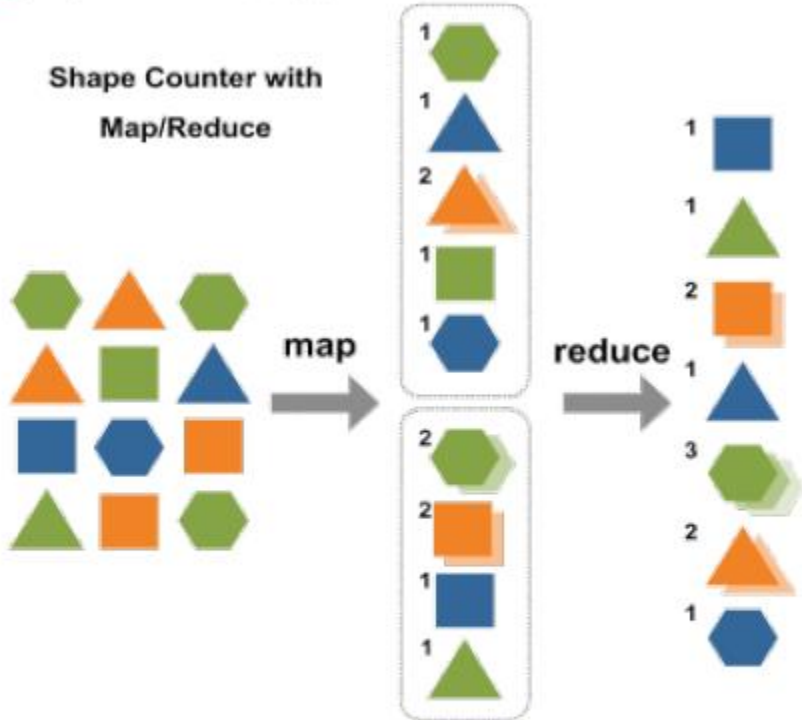
# 필요기술: Big Data 저장관리

주요 기술	설명
<b>빅 테이블 (Big Table)</b>	<ul style="list-style-type: none"> <li>• 구글 파일 시스템(Google File System) 상에 구축된 상용 분산 데이터 베이스 시스템</li> <li>• HBase에 영향을 미침</li> </ul>
<b>카산드라 (Cassandra)</b>	<ul style="list-style-type: none"> <li>• 분산 시스템에서 방대한 분량의 데이터를 처리할 수 있도록 디자인된 오픈소스(무료) 데이터베이스 관리 시스템</li> <li>• 원래 페이스북에서 개발했으며 지금은 아파치 소프트웨어 재단의 한 프로젝트로 관리되고 있음</li> </ul>
<b>분산 시스템 (Distributed System)</b>	<ul style="list-style-type: none"> <li>• 동시에 일을 처리하기 위해 네트워크로 연결된 컴퓨터들의 집합으로 단일 또는 다수의 컴퓨터의 리소스를 부분적으로 활용함으로써 시스템의 가성비, 안정성 그리고 확장성을 향상시킬 수 있음</li> </ul>
<b>구글 파일 시스템 (Google File System)</b>	<ul style="list-style-type: none"> <li>• 구글에서 개발한 분산 파일 시스템</li> <li>• Hadoop과 관련되어 있음</li> </ul>
<b>Hadoop</b>	<ul style="list-style-type: none"> <li>• 분산 시스템 상에서 대용량 데이터 처리 분석을 지원하는 오픈소스 프레임워크</li> <li>• 구글이 개발한 맵리듀스 (MapReduce)를 오픈소스로 구현한 결과물</li> <li>• 원래 야후!에서 최초 개발되었으며 지금은 아파치 소프트웨어 재단의 한 프로젝트로 관리됨</li> </ul>
<b>HBase</b>	<ul style="list-style-type: none"> <li>• 구글(Google)의 빅테이블(Big Table)을 참고로 개발된 오픈소스 분산 비관계형 데이터베이스</li> <li>• 원래 Powerset에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 Hadoop의 일환인 프로젝트로 관리되고 있음</li> </ul>
<b>MapReduce</b>	<ul style="list-style-type: none"> <li>• 분산 시스템 상에서 대용량 데이터 세트를 처리하기 위해 구글(Google)이 소개한 소프트웨어 프레임워크로 Hadoop에 구현되어 있음</li> </ul>
<b>비관계형 데이터베이스/ Key Value Store</b>	<ul style="list-style-type: none"> <li>• 비관계형 데이터베이스는 데이터를 테이블(행, 컬럼)에 저장하지 않는 데이터베이스이며 관계형 데이터베이스와 대조되는 개념임</li> <li>• Key Value Stores를 사용하면 스키마 없는 엔티티(noSQL)를 관리할 수 있음</li> </ul>



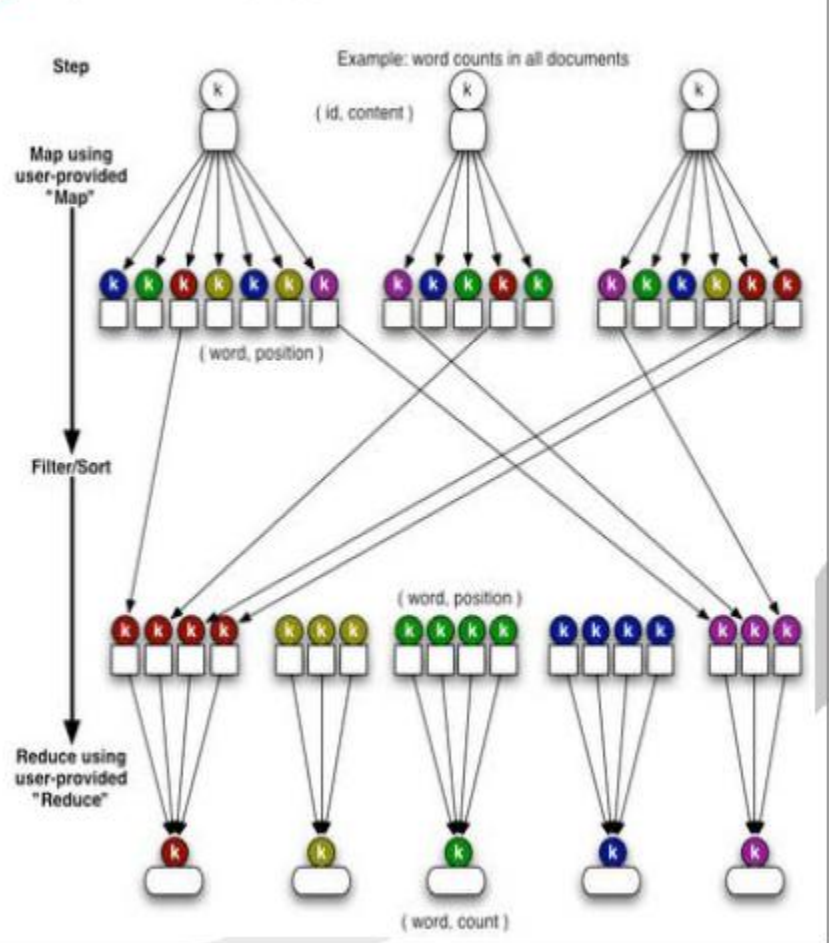
# 필요기술: Map Reduce

## Map / Reduce의 개념

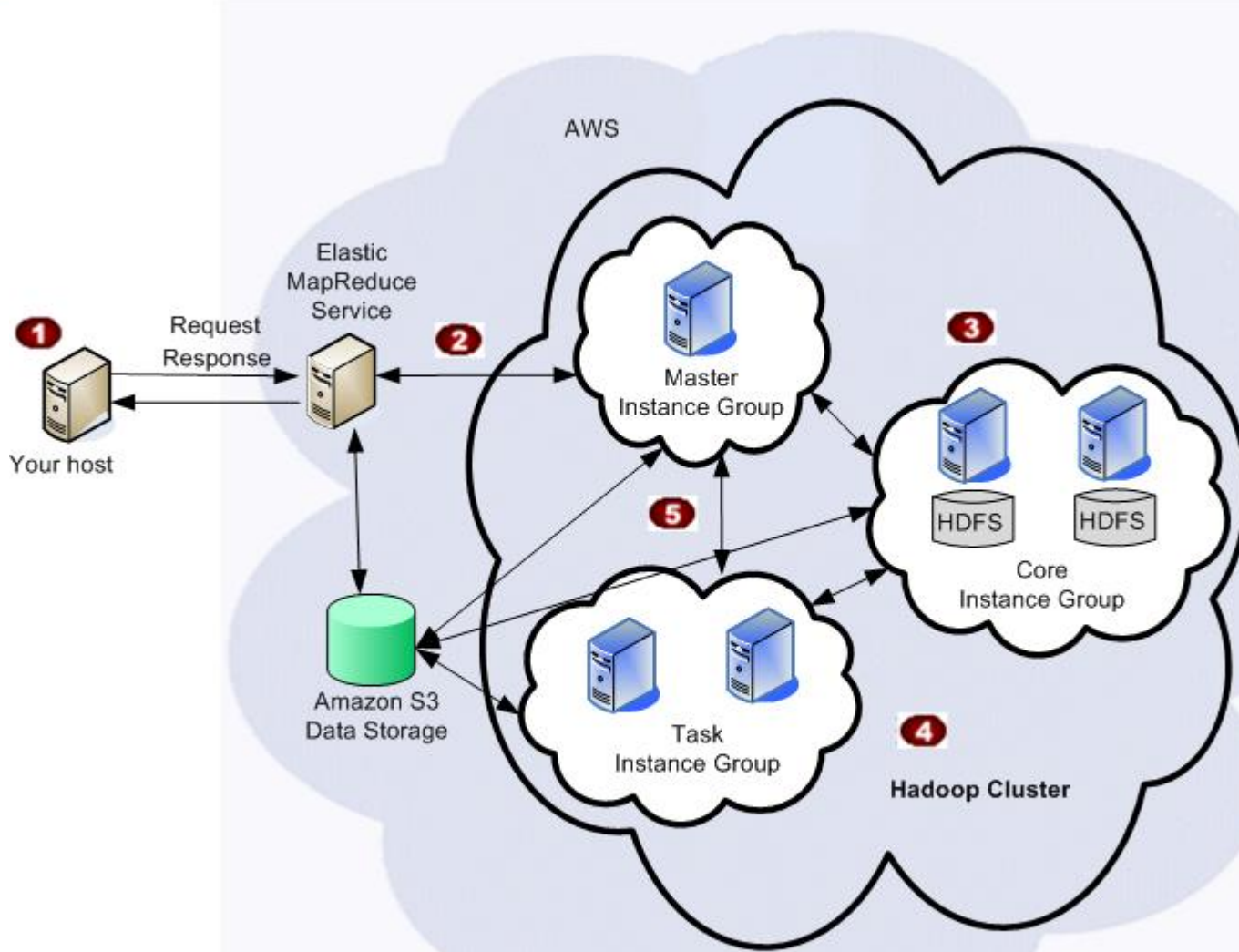


Map : 흩어져 있는 데이터를 Key, Value 의 형태로 연관성 있는 데이터 분류로 묶는 작업.  
 Reduce : Map 화 한 작업 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 단계.

## Map / Reduce의 예시



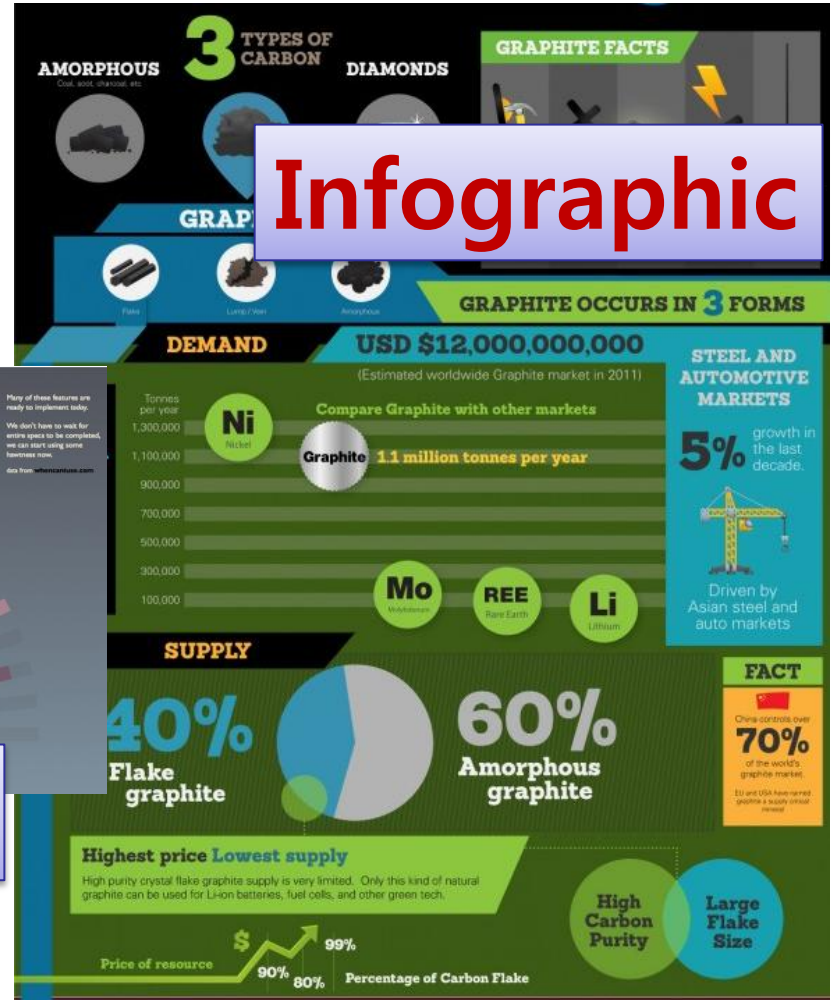
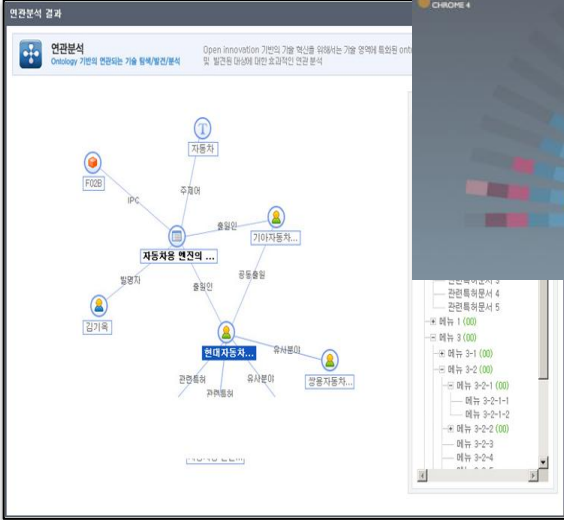
# Big Data 처리 인프라 서비스 (Amazon Elastic Map Reduce)



Source: <http://docs.amazonwebservices.com/ElasticMapReduce>



## HTML5





## R is a tool for...

### Data Manipulation

- connecting to data sources
- slicing & dicing data

### Modeling & Computation

- statistical modeling
- numerical simulation

### Data Visualization

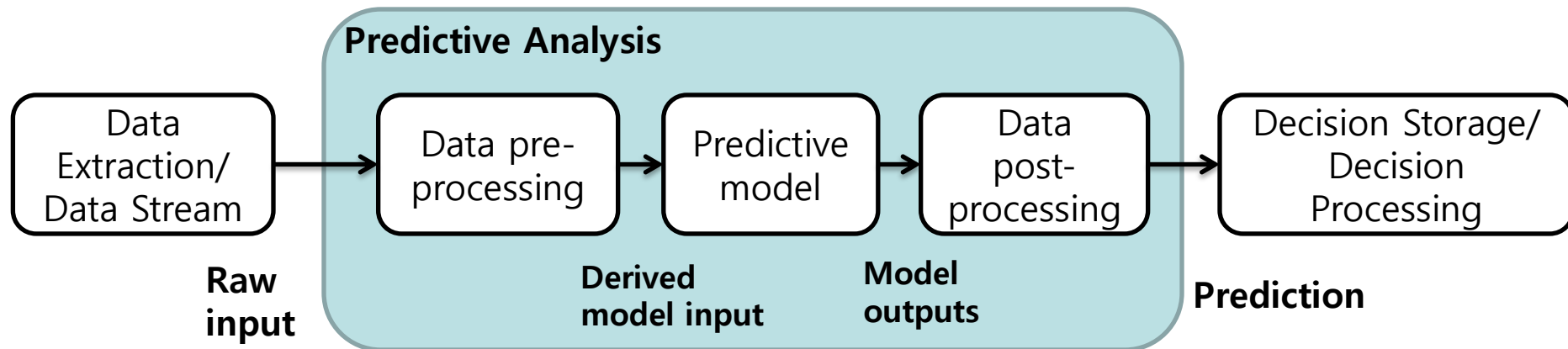
- visualizing fit of models
- composing statistical graphics

The screenshot displays the RStudio interface with several windows open:

- Source Editor:** Contains R code for data manipulation and visualization:

```
rgl.sp.ylen <- ylen[Z] - ylen[1] + 1
rgl.sp.colorlut <- terrain.colors(ylen)
rgl.sp.col <- colorlut[y - ylen[1] + 1]
rgl.sp <- rgl.clear()
rgl.sp <- rgl.surface(x, z, y, color = col)
rgl.viewpoint(az = 0, el = 0)
rgl.render()
border <- function(data, repts = 200, k = c(0, 1),
  odd = TRUE, col = 1L, border = FALSE, col1 = 1L) {
  dim <- dim(data)
  dx <- dim[2]
  dy <- dim[1]
  plot(0, 0, axes = F, main = "", xlim = x, ylim = y,
    ylab = " ")
  if (orientation == "page") {
    dx2 <- (dx - min(dx)) / (max(dx) - min(dx)) * (x[2] - x[1])
    dy2 <- (dy - min(dy)) / (max(dy) - min(dy)) * (y[2] - y[1])
    aspect <- rep(y[1:], length(dx))
    if (fill == F) {
      confshade(dx2, sebelow, dy2, col = col)
    }
    if (border == TRUE) points(dx2, dy2, type = "l", col = col)
  } else {
    dx2 <- (dx - min(dx)) / (max(dx) - min(dx)) * (x[2] - x[1])
    dy2 <- (dy - min(dy)) / (max(dy) - min(dy)) * (y[2] - y[1])
  }
}
```
- R Data Editor:** Shows a table with columns 'height' and 'weight'. The 'height' column contains values from 115 to 164, and the 'weight' column contains values from 120 to 164.
- Workspace Browser:** Lists objects in the workspace, including 'data' (data.frame), 'g' (factor), 'i' (numeric), 'n' (numeric), 'opar' (list), 'pic.sides' (numeric), 'pic' (numeric), 'scale' (numeric), 'sar' (numeric), 'Y.women' (data.frame), 'height' (numeric), 'weight' (numeric), and 'x' (numeric).
- Package Manager:** Shows installed and available packages, including 'graphics', 'grid', 'lattice', and 'methods'.
- 3D Plot:** A 3D surface plot showing a topographic map with a color gradient from green to yellow.

**Predictive analytics** encompasses a variety of statistical techniques from modeling, [data mining](#) and [game theory](#) that analyze current and historical facts to make predictions about future events. (Wikipedia)



Source : IBM

**BBC**  
**NEWS TECHNOLOGY**

Home | UK | Africa | Asia-Pac | Europe | Latin America | Mid-East | South Asia | US & Canada | Business

9 September 2011 Last updated at 14:57 GMT

## Supercomputer predicts revolution

**Feeding a supercomputer with news stories could help predict major world events, according to US research.**

A study, based on millions of articles, charted deteriorating national sentiment ahead of the recent revolutions in Libya and Egypt.

While the analysis was carried out retrospectively, scientists say the same processes could be used to anticipate upcoming conflict.

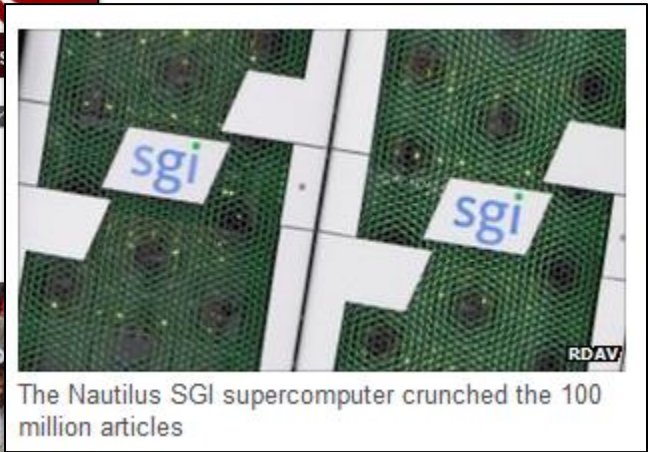
The system also picked up early clues about Osama Bin Laden's location.

Kalev Leetaru, from the University of Illinois' Institute for Computing in the Humanities, Arts and Social Science, **presented his findings** in the journal First Monday.

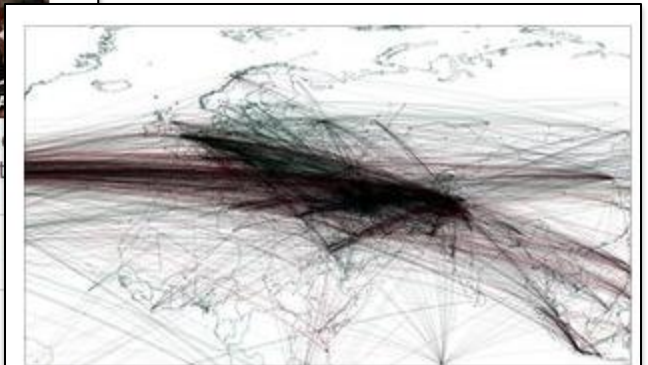
**Mood and location**



Sentiment mining showed a sharp change in tone around Egypt ahead of President Mubarak's ouster



The Nautilus SGI supercomputer crunched the 100 million articles



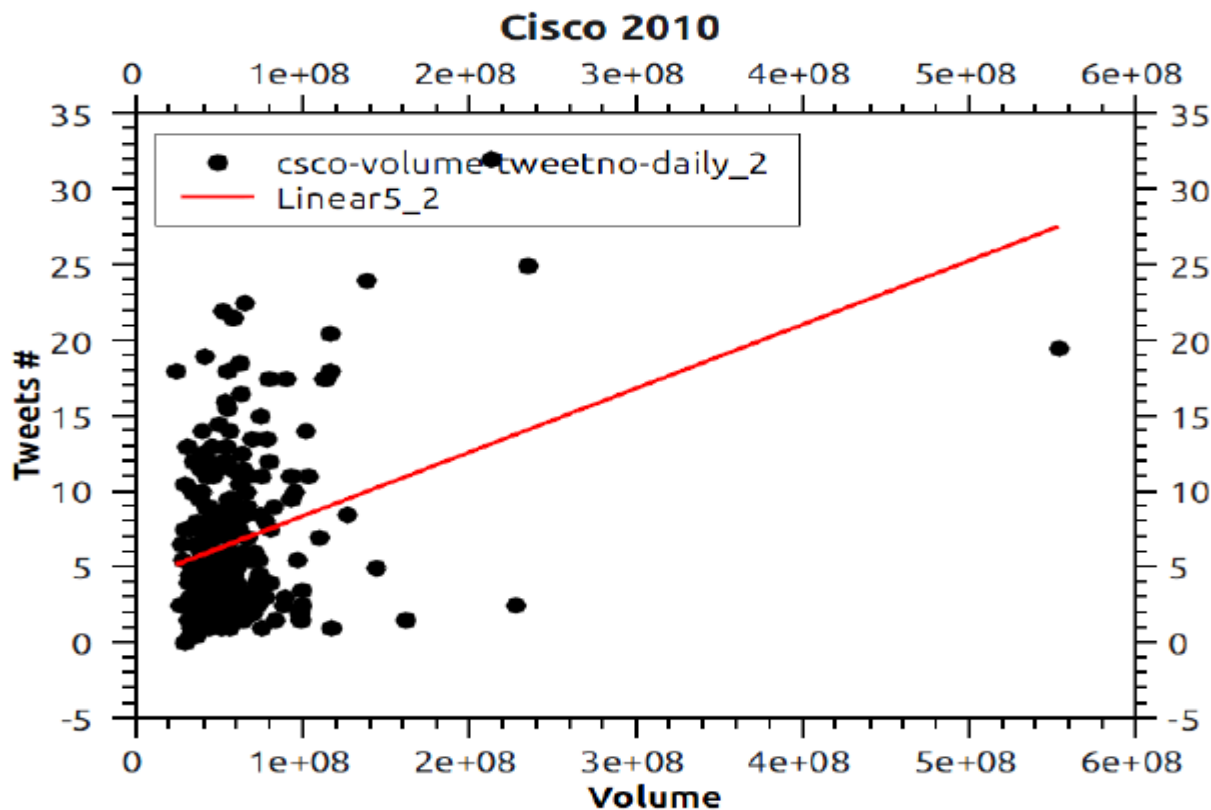
Media reports mentioning Osama Bin Laden may have helped narrow down his location

Source : BBC

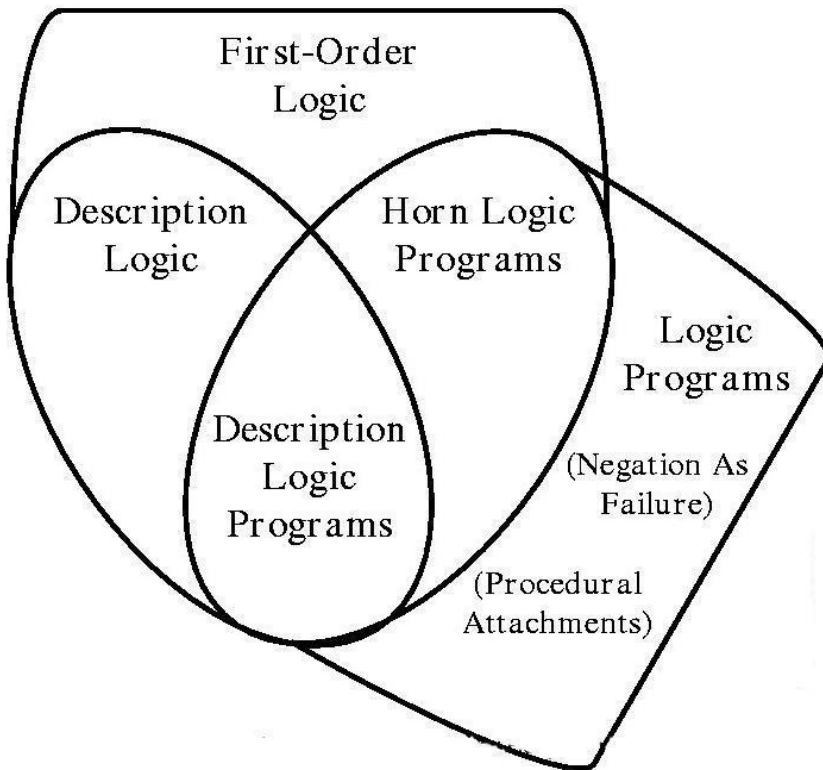




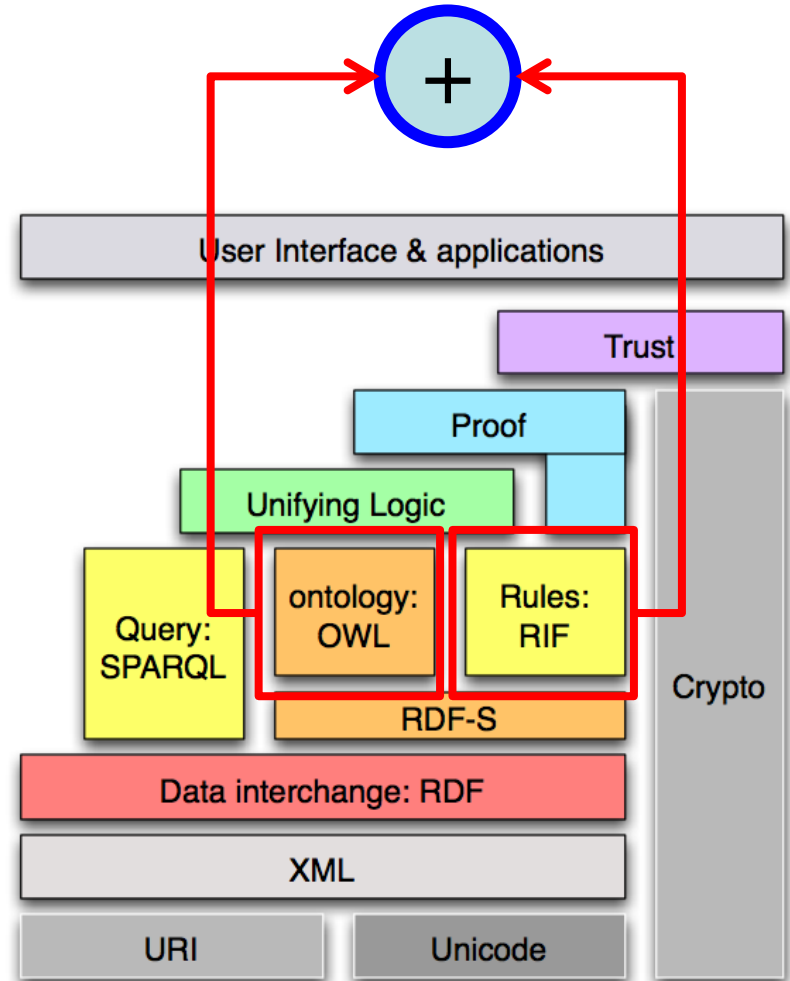
The relationship between Cisco daily volume of stock trades and the volume of tweets – fairly good correlation was found



Source : SEMTECH 2011

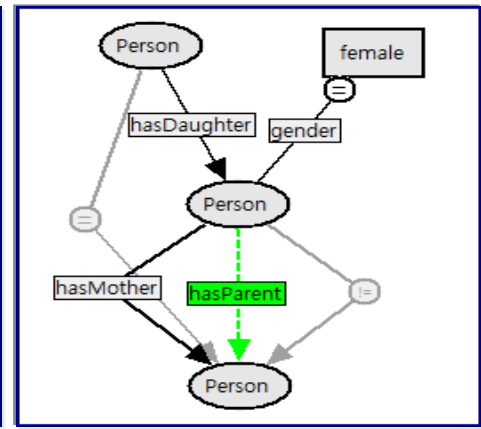
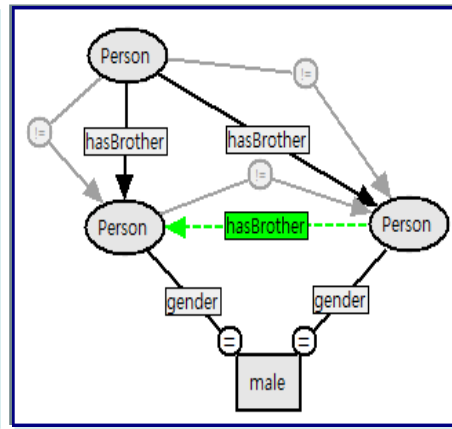
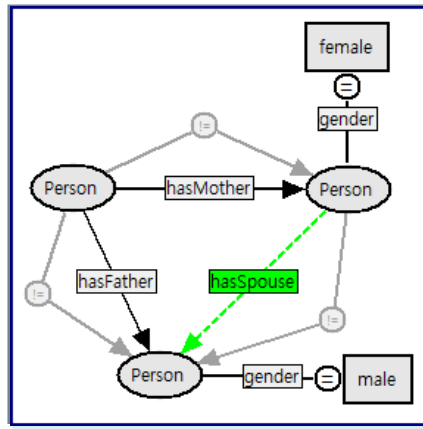
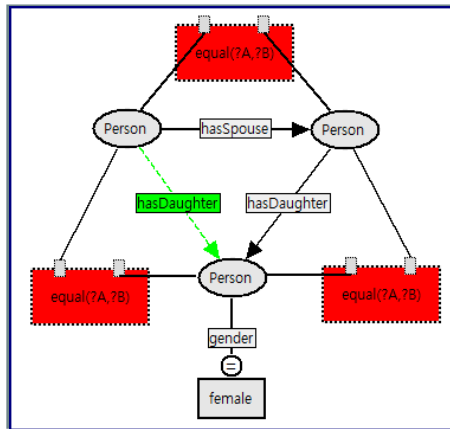
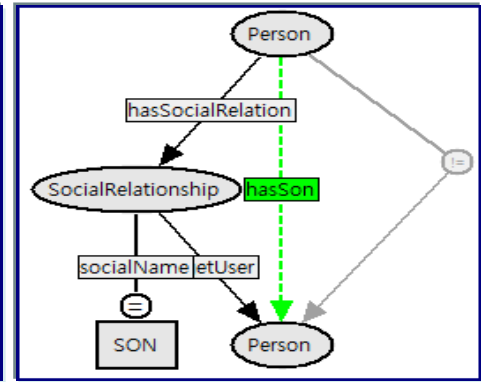
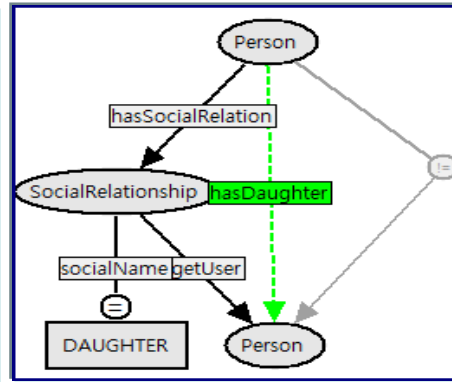
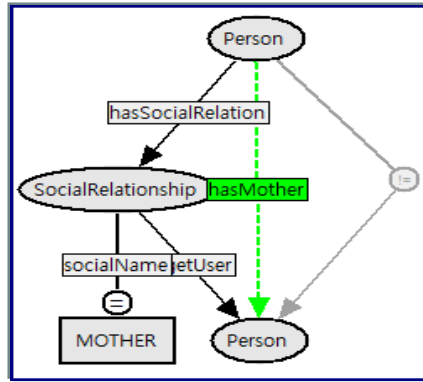
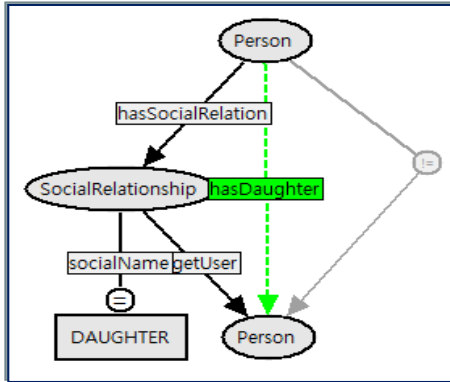


The relationships among different formalisms (Benjamin Grosf)

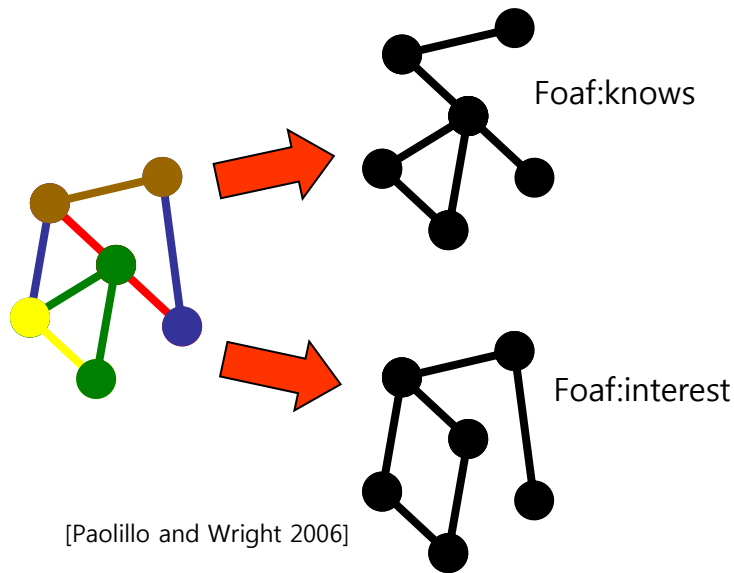


Semantic Web Architecture

# 필요기술: Semantic (Hybrid Reasoning: DL + Rules)

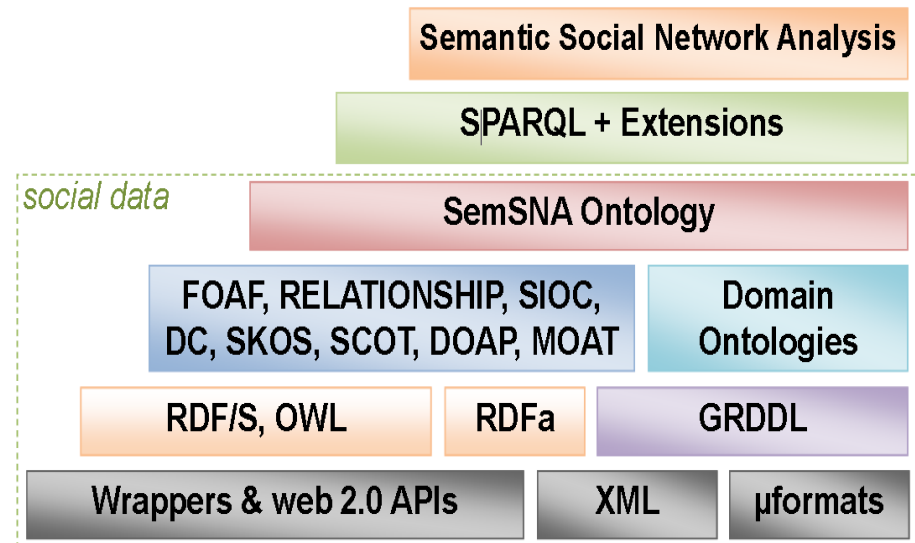


- **Social Networks** : networks based on the relation between people
- **Semantic Social Network** : RDF representations of social network and data



Rich graph representations reduced to simple untyped graphs in order to apply SNA

### Abstraction stack for semantic SNA



[Semantic Social Network Analysis, [http://journal.webscience.org/141/2/websci09\\_submission\\_43.pdf](http://journal.webscience.org/141/2/websci09_submission_43.pdf)]

시맨틱 네트워크의 구성과 분석

$$Density = \sum_k v_k / g(g-1)$$

$v_k$  : weighting of relation  
 $n$  : number of relations  
 $g$  : total number of entity

가족, 동료,  
부서, 동아리

Density,  
n-Clan, n-Clique

$$C'_{jk}(i) = \frac{\sum_{jkt} g_{jk}(t) / g_{jk}}{[(g-1)(g-2)/2]}$$

$g_{jk}$  : j와k 사이에 존재하는 최단 경로 수  
 $g_{jk}(i)$  : j와 k 경로 중 i를 경유하는 수)

정보소통의  
중심, 매개자

Degree, Closeness  
Betweenness  
Centrality

Dijkstra's algorithm

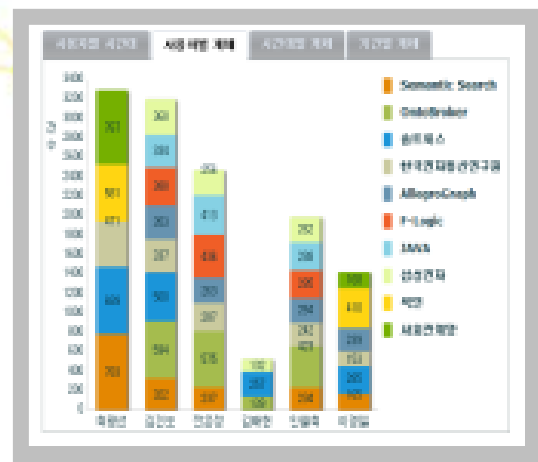
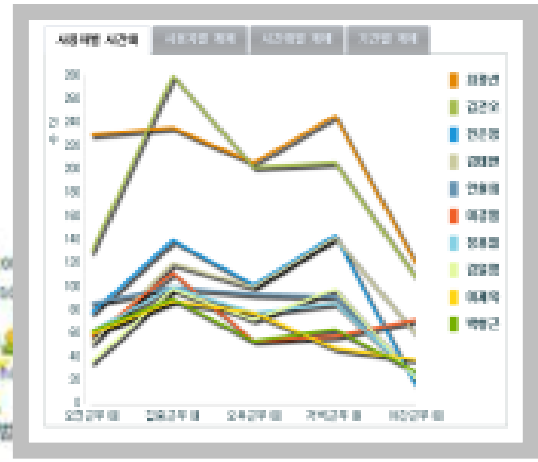
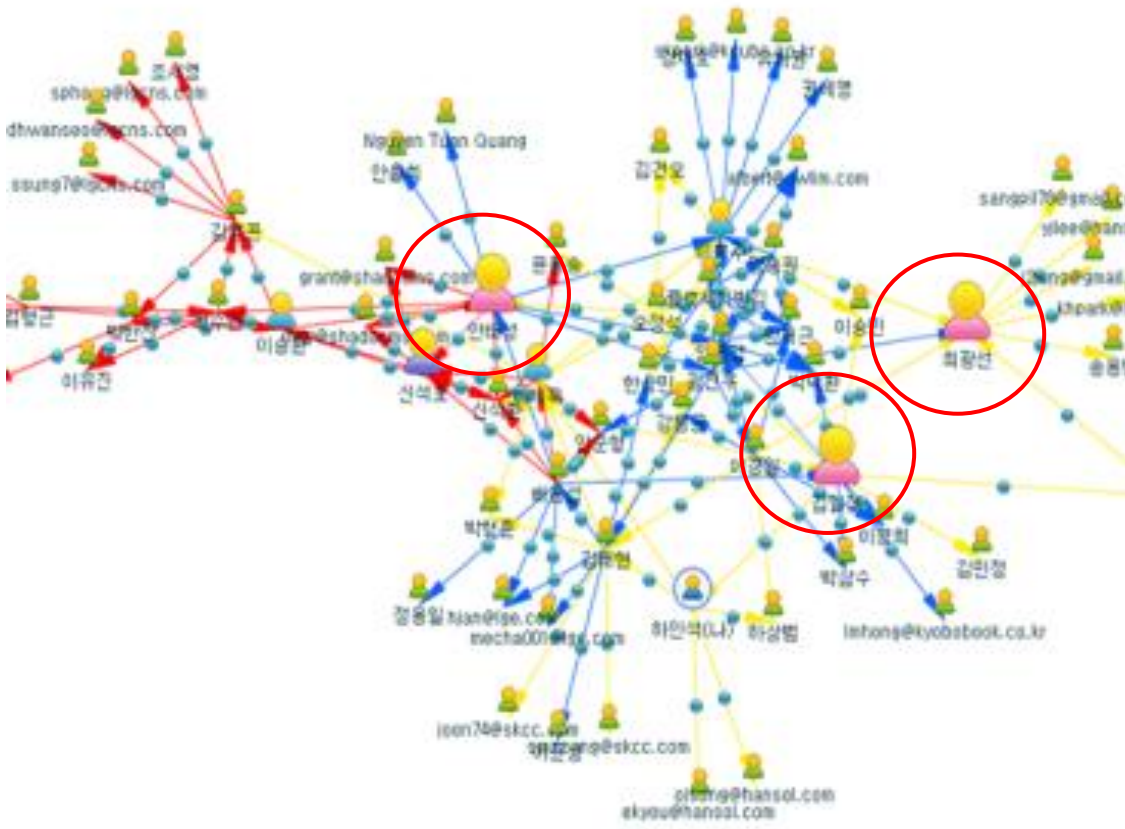


$O(|E| + |V| \log |V|)$

빨리 사람  
찾기

Shortest Path  
Algorithms

# 필요기술: Semantic





# LarKC Urban Computing

Road Sign ID				<a href="http://www.saltlux.com/rsm#rsp_3871">http://www.saltlux.com/rsm#rsp_3871</a>
Road Sign Name				UR-92[남부순환로]-DN-4
Left Long	Semi Left Long	Forward Long	Semi Right Long	
.	.	삼성역	.	
Left Short	Semi Left Short	Forward Short	Semi Right Short	
영동세브란스	.	대치역	구룡터널	



# LarKC Urban Computing

URBAN NAVIGATION LARKC

Shortest path

Calculate the path between the start and end points by taking in consideration the roads length, i.e. minimizing the global path length.

Length:	7978 m
Nominal travel time:	14' 59"
Predicted travel time:	14' 59"

Prediction summary

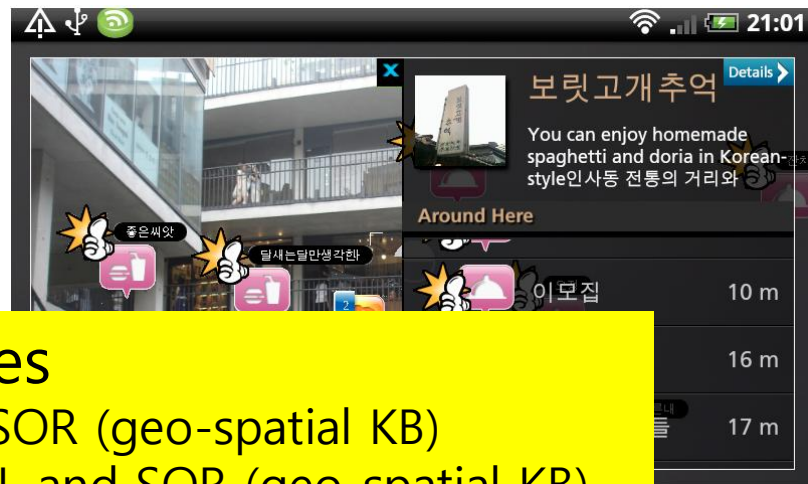
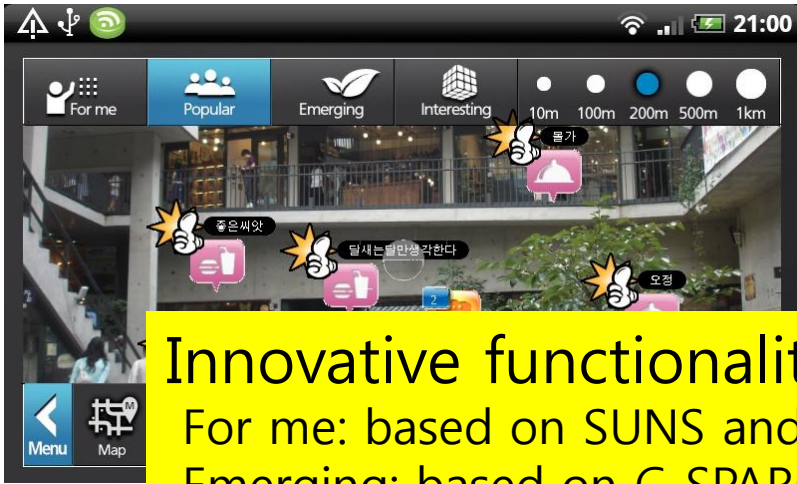
Current time: 11/26/2010 6:59PM  
Prediction time: 11/26/2010 7:59PM

Length:	8484 m
Nominal travel time:	13' 28"
Predicted travel time:	13' 28"

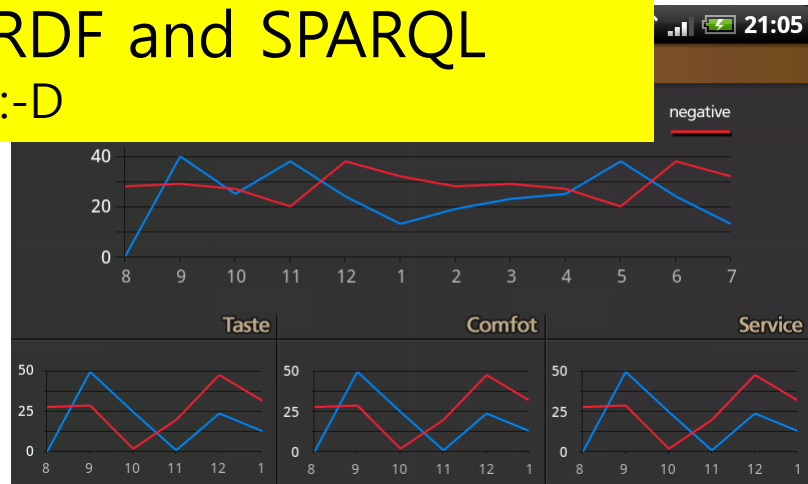
Winner of the AI Mashup Challenge 2011



# LarKC Urban Computing



**Innovative functionalities**  
 For me: based on SUNS and SOR (geo-spatial KB)  
 Emerging: based on C-SPARQL and SOR (geo-spatial KB)  
**Fully implemented on RDF and SPARQL**  
**First Commercial Mobile APP :-D**



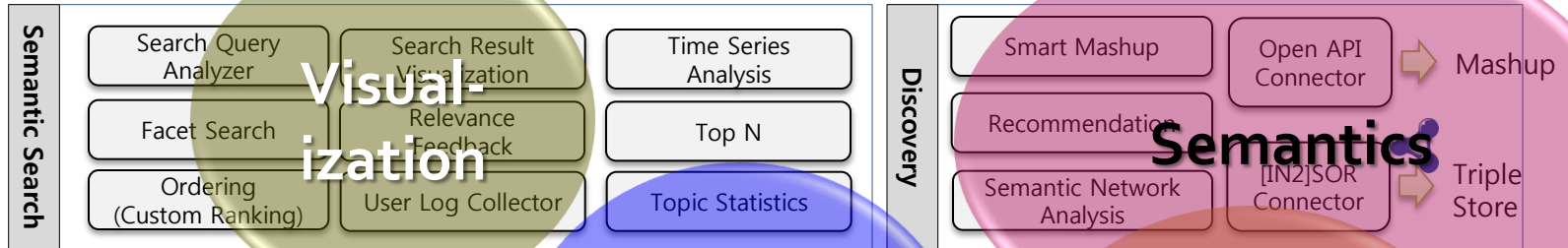
# Saltlux [IN2] Discovery Architecture 2.x

13 FEB. 2012

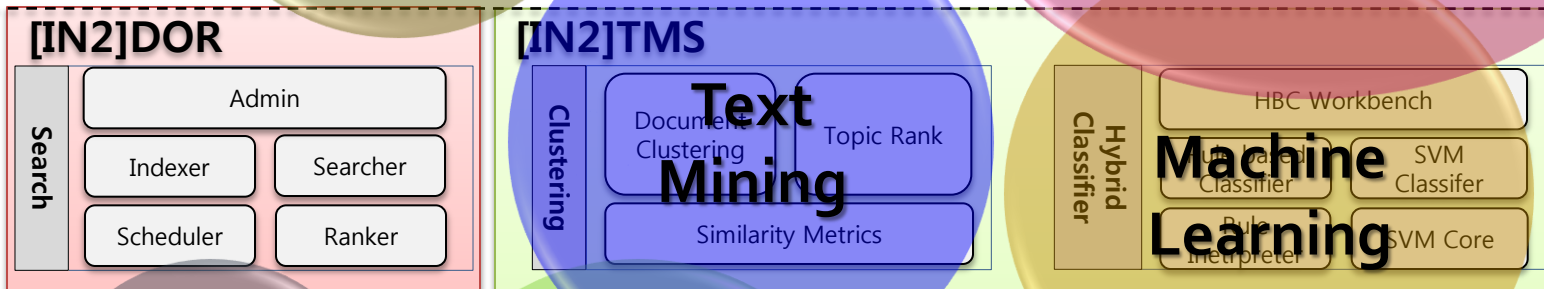


Common API [Restful, SOAP, Direct API]

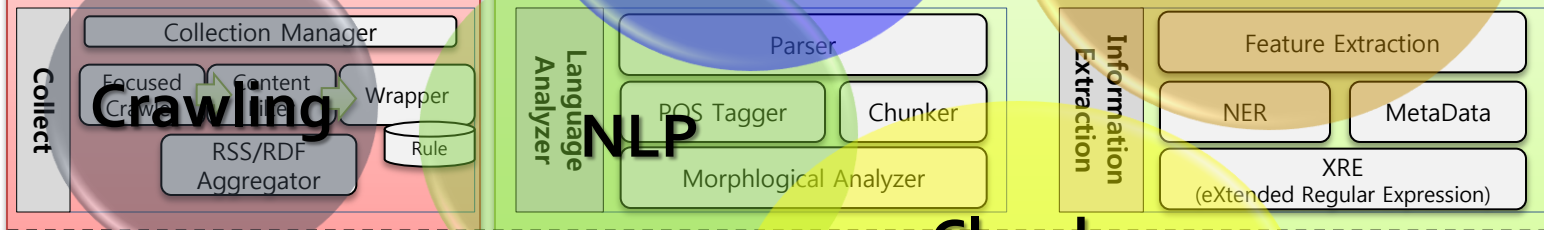
Service Enabler Layer



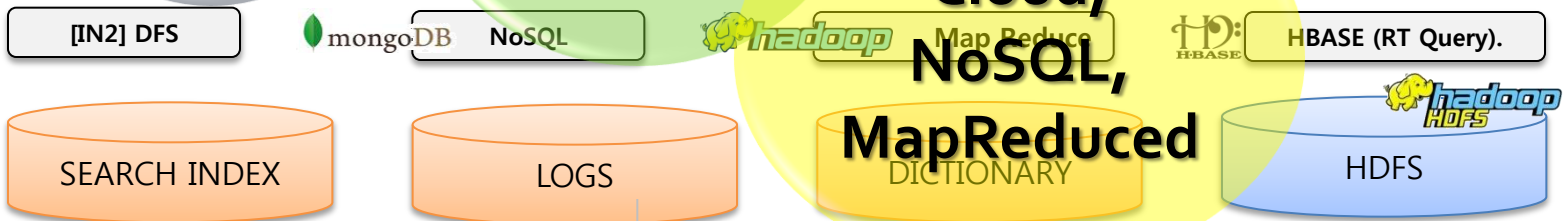
Analysis Layer



Information Collection Layer



Data Layer



# 4. Saltlux 기업 소개

4. 2014년 1월 1일



“세상 사람들이 자유롭게 지식 소통 하도록 돕는 일”,  
이것이 솔트룩스의 사명(mission)입니다.

- 회 사 명    **주식회사 솔트룩스 (Saltlux Inc.) / 1979. 6. 1. 설립**
- 대 표 이 사   **이 경 일**
- 본 사 주 소 지   **서울시 강남구 대치동 967번지 덕일빌딩 (02-3402-0081)**
- 해외법인/지사   **솔트룩스 Japan, 베트남 개발센터(VDC)**
- 홈 페 이 지    **www.saltlux.com**
- 기 술 연 구 소   **HLT Laboratory (한국 최초 EU FP6/FP7 프로젝트 수행)**
- 주 요 제 품    **[IN2] : Search & Discovery Platform**  
                  **STORM: Semantic Business Platform**  
                  **OWLIM: Semantic Web Search Service**



TTA, ISO, 지식경제부, 소프트웨어진흥원, 과학기술부, 행정안전부 등으로 부터 다양한 인증과 수상실적을 가진 **신뢰할만한 기업**입니다.

## 아시아 IT 200대 기업 선정



세계적인 미디어 회사 Red Herring으로부터 RH 200 Asia Awards를 수상하여 기술력을 국제적으로 인정 받았습니다. (검색, 정보 마이닝 분야 국내 최초)

## 2010 대통령상 수상



2010' 대한민국 소프트웨어대상 상품상 부문 대통령상을 수상하며 2010년 최고의 SW로 선정되었습니다.

## GS 인증 (Good Software)



SW 품질 신뢰성과 상호 호환성 등의 까다로운 심사를 통과 하였습니다. (정보 마이닝 / 시맨틱 검색 엔진 국내 유일)

## 디지털이노베이션 대상



[IN2]

## INNO-BIZ 선정



기술혁신형 중소기업 선정

## ISO9001:2000



품질관리

## 신 소프트웨어 상품대상



[IN2]DOR & TMS/Discovery

## 2010년 IT 히트 상품



[IN2]Discovery, 품질 우수부문

## 과기부 신기술 마크



자동번역 솔루션

## 행망 S/W 등록



[IN2]DOR/TMS

# ○ 핵심 기술 : 시맨틱 기술과 정보 마이닝

컴퓨터가 정보를 스스로 처리하여 상황에 맞게 사람과 컴퓨터가 상호 협력  
할 수 있도록 돕는 차세대 웹, 지식 처리 기술



## ○ 주요고객

(주)솔트룩스는 다양한 산업 분야에서 ECM, KMS, EDM, GW 등 다양한 문서 및 지식 콘텐츠 시스템들과 연동한 많은 레퍼런스를 보유하고 있습니다.

산업분야	고객사이트
건설 및 제조	삼성전자, 현대기아자동차, 삼성중공업, 한화그룹, 포스코, 현대제철, 현대상선, 농심, 현대엔지니어링, 현대시멘트, LG전자, 한진중공업, 두산인프라코어, 동국제강, GS파워, SFA엔지니어링, 동서, 동서물산, LS전선, LS산전, 유한킴벌리, 동양기전, 한국타이어, 한일시멘트, 현대하이스코, 코아로직, 금호건설, GS건설, LIG건설, 성원건설, 포스코건설, 태영건설, 경남기업, 도화종합기술공사, 하이트진로, 광동제약 등
공공	행정안전부, 통일부, 국방부, 노동부, 외교부, 환경부, 국가기록원, 대통령기록관, 특허청, 조달청, KOTRA, 한국도로공사, 한국전기안전공사, 한국가스안전공사, 한국전력, 인천공항공사, 한국철도공사, 대한송유관공사, 한국항공우주연구원 등
정보통신	SK Telecom, KT(QOOK TV), Yahoo Korea, LG CNS, POSCO ICT, 쌍용정보통신, 롯데정보통신, 한화S&C, 한국스마트카드 등
금융	국민은행, 기업은행, 대구은행, 푸르덴셜생명보험, 교보생명, BC카드, 현대캐피탈, 산은캐피탈, IBK연금보험, 코람코자산신탁 등
기타	KBS, SBS, 아리랑TV, 경인방송, 한솔교육, 딜로이트, 태평양법무법인, 골프존



# Thank you!

135-848 서울특별시 강남구 대치동 967 덕일빌딩 5, 6, 7 층

Tel : 02-3402-0081 Home : [www.saltlux.com](http://www.saltlux.com)

Fax: 02-3402-0082 E-mail : [saltluxinc@saltlux.com](mailto:saltluxinc@saltlux.com)

