빅데이터 분석

2012. 2. 28

내용

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

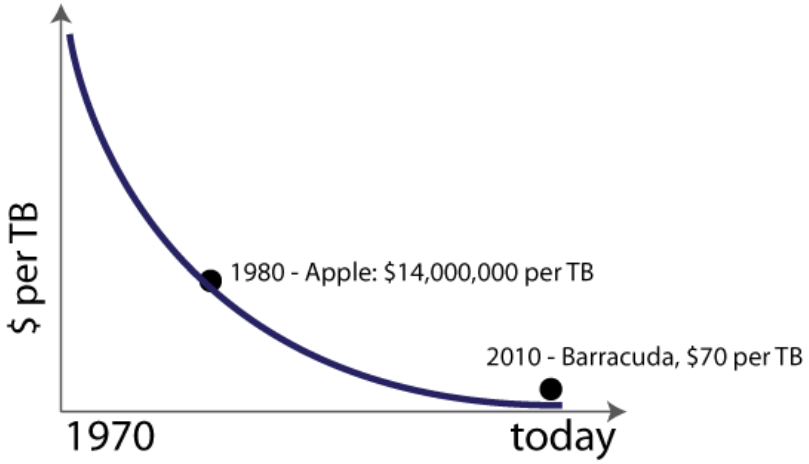
Saltlux

1. Big Data ?

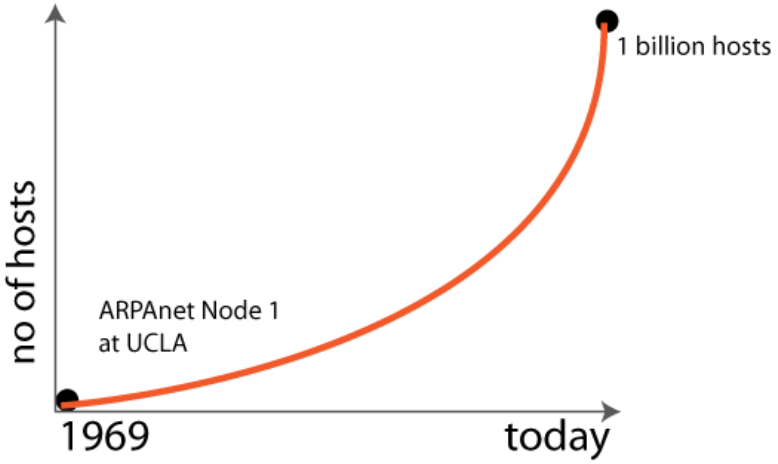
Saltlux

Why Big Data?

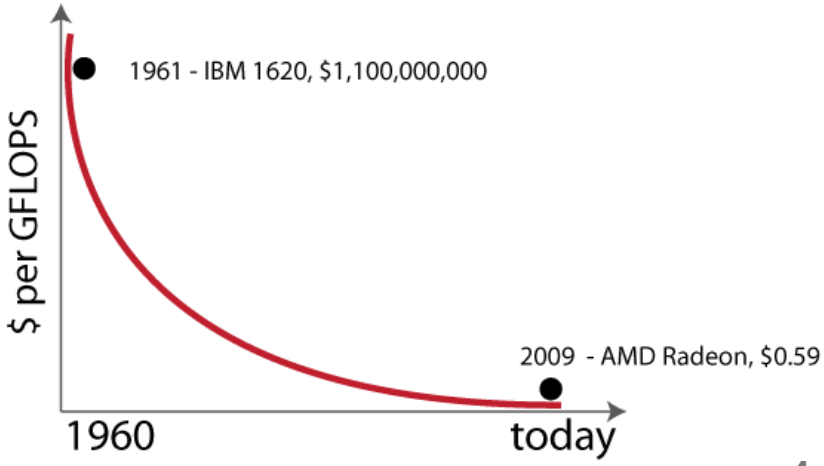
storage cost



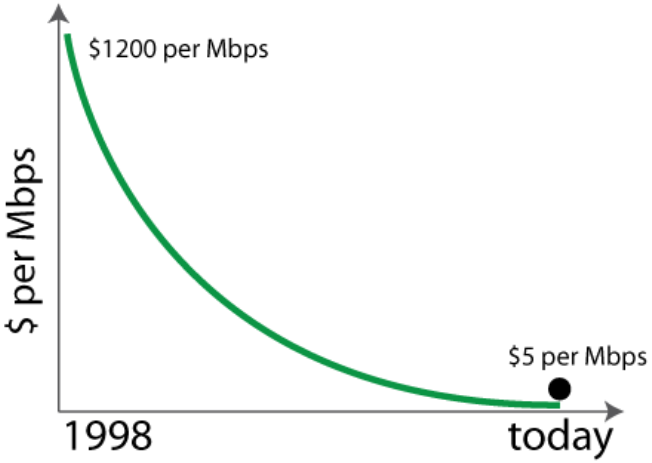
network access



CPU cost



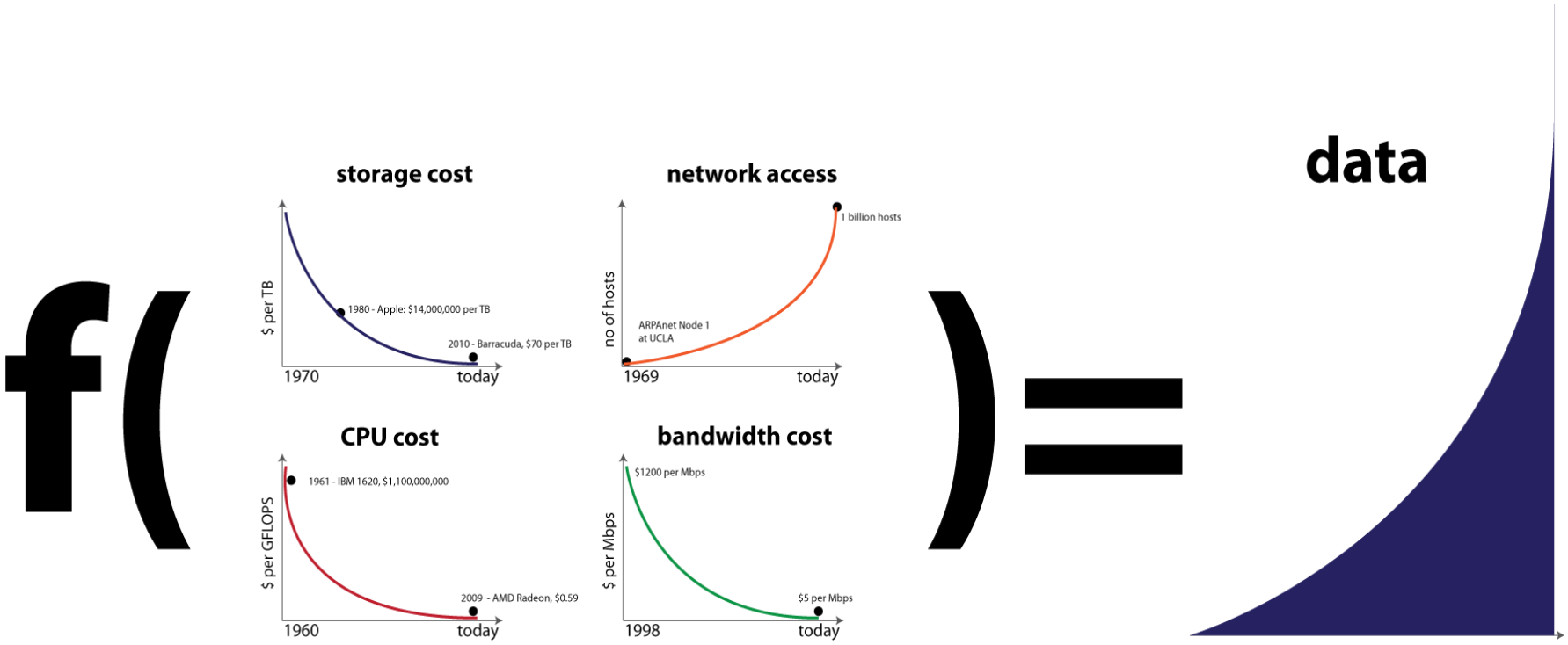
bandwidth cost



Saltlux

Why Big Data?

데이터 저장, 관리 비용이 낮아지면서 데이터는 폭팔적으로 증가 추세



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

Saltlux

It's

too Huge ,

Fast and

Heterogeneous

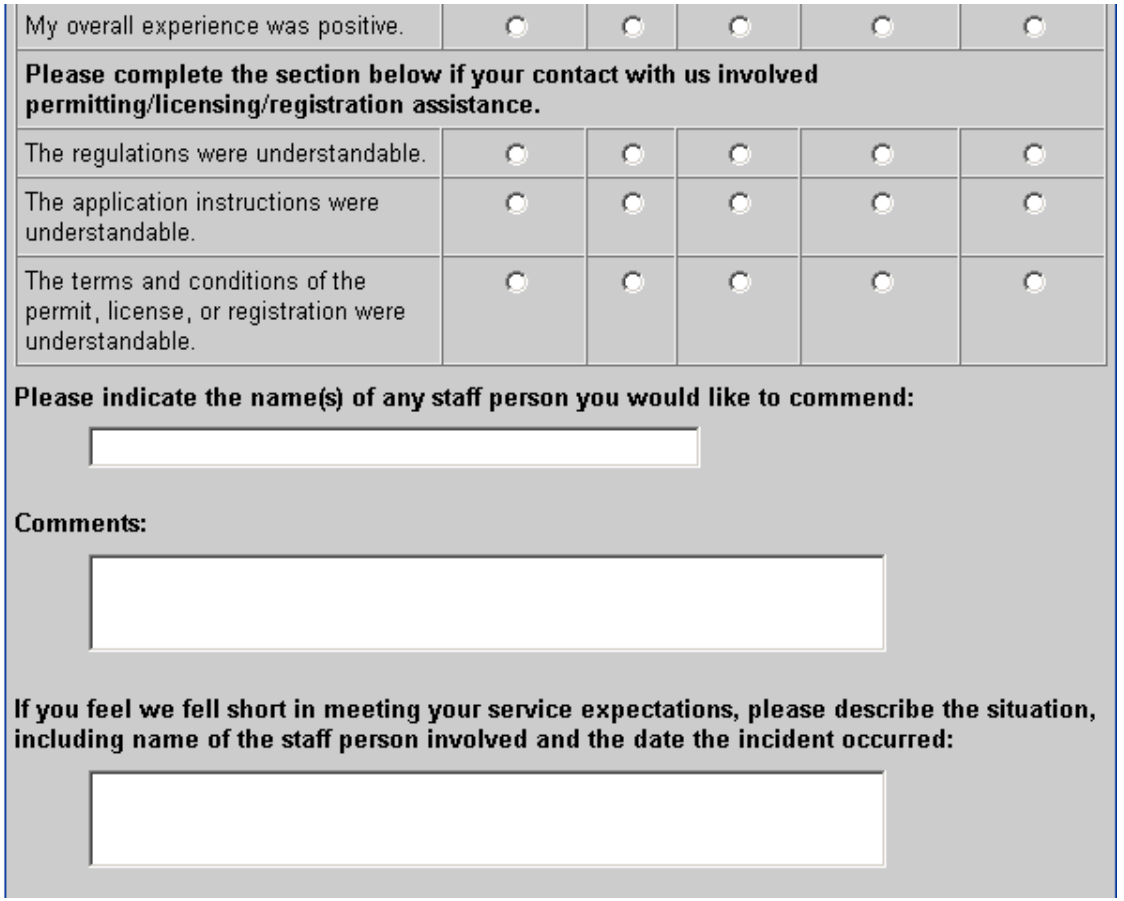
} Big Data Issues (3V)

to understand and utilize them.

정형 vs. 비정형

정형 데이터

비정형 데이터



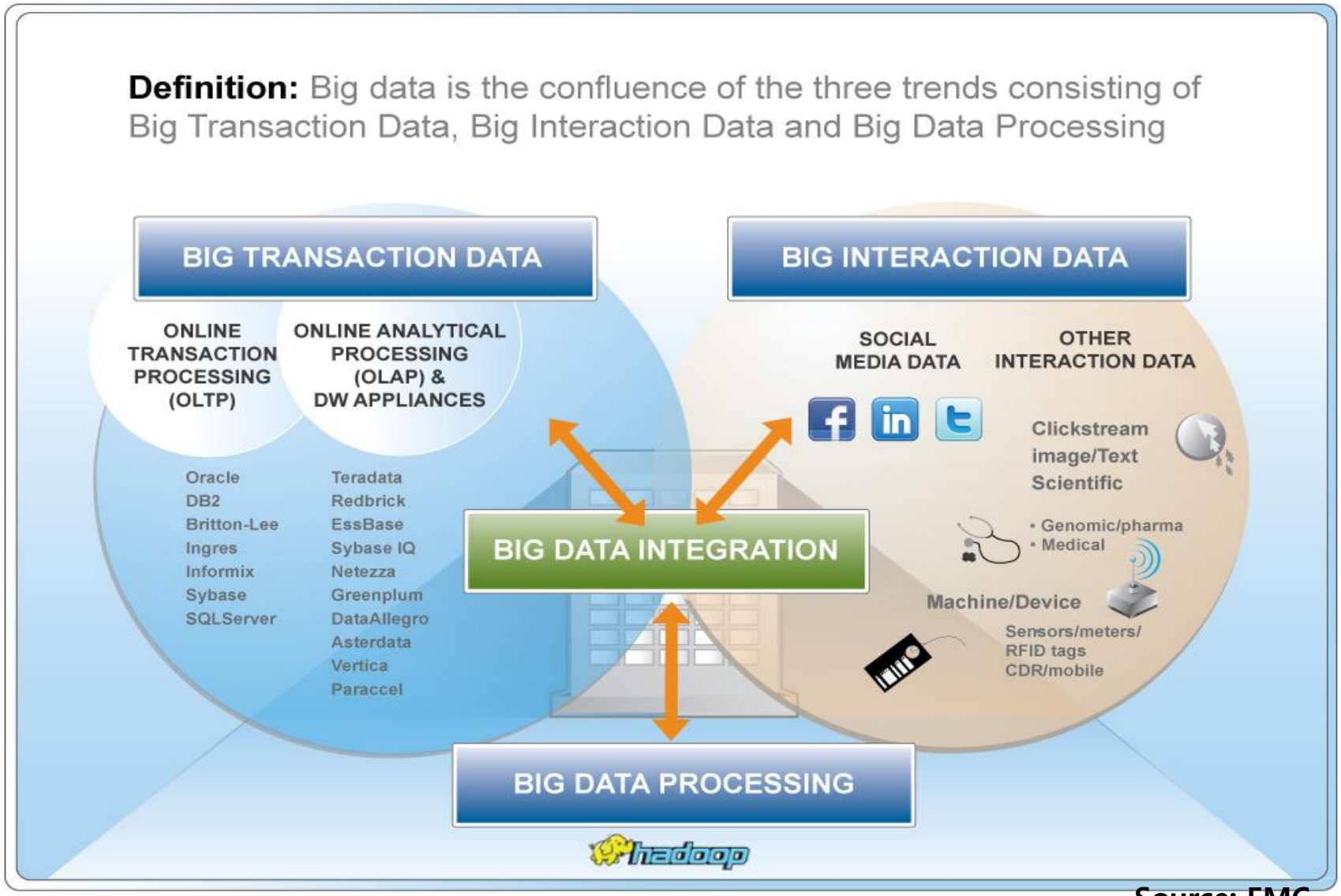
My overall experience was positive.	○	○	○	○	○
Please complete the section below if your contact with us involved permitting/licensing/registration assistance.					
The regulations were understandable.	○	○	○	○	○
The application instructions were understandable.	○	○	○	○	○
The terms and conditions of the permit, license, or registration were understandable.	○	○	○	○	○

Please indicate the name(s) of any staff person you would like to commend:

Comments:

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:

Why Big Data?



Source: EMC

Saltlux

2. Big Data 와 Biz

Saltlux

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"

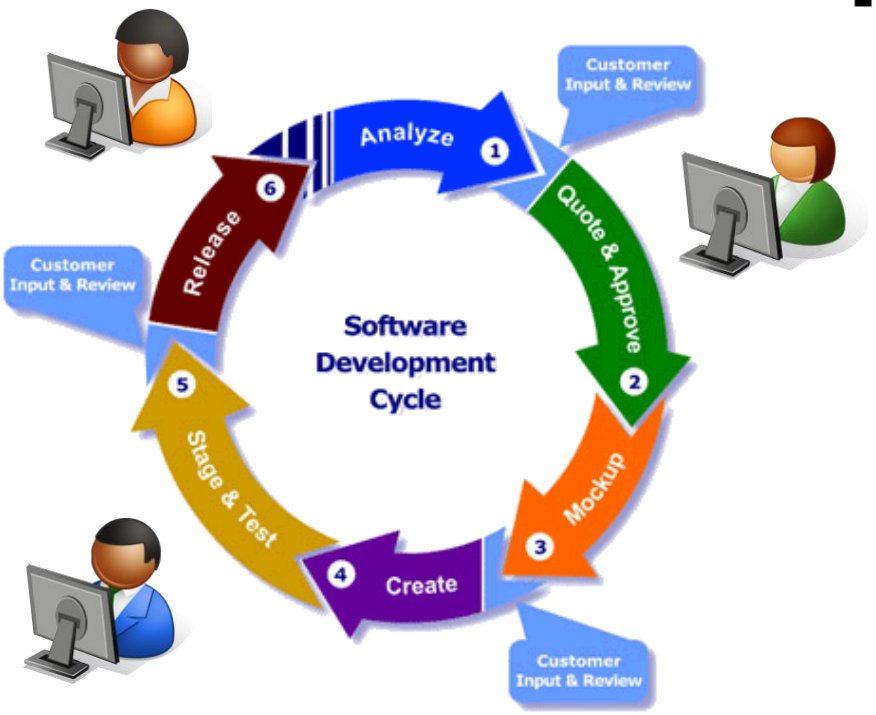
STAGE 4

"Software is not even differentiating, the value is the data"

Saltlux

Age of Data?

The Age of Software



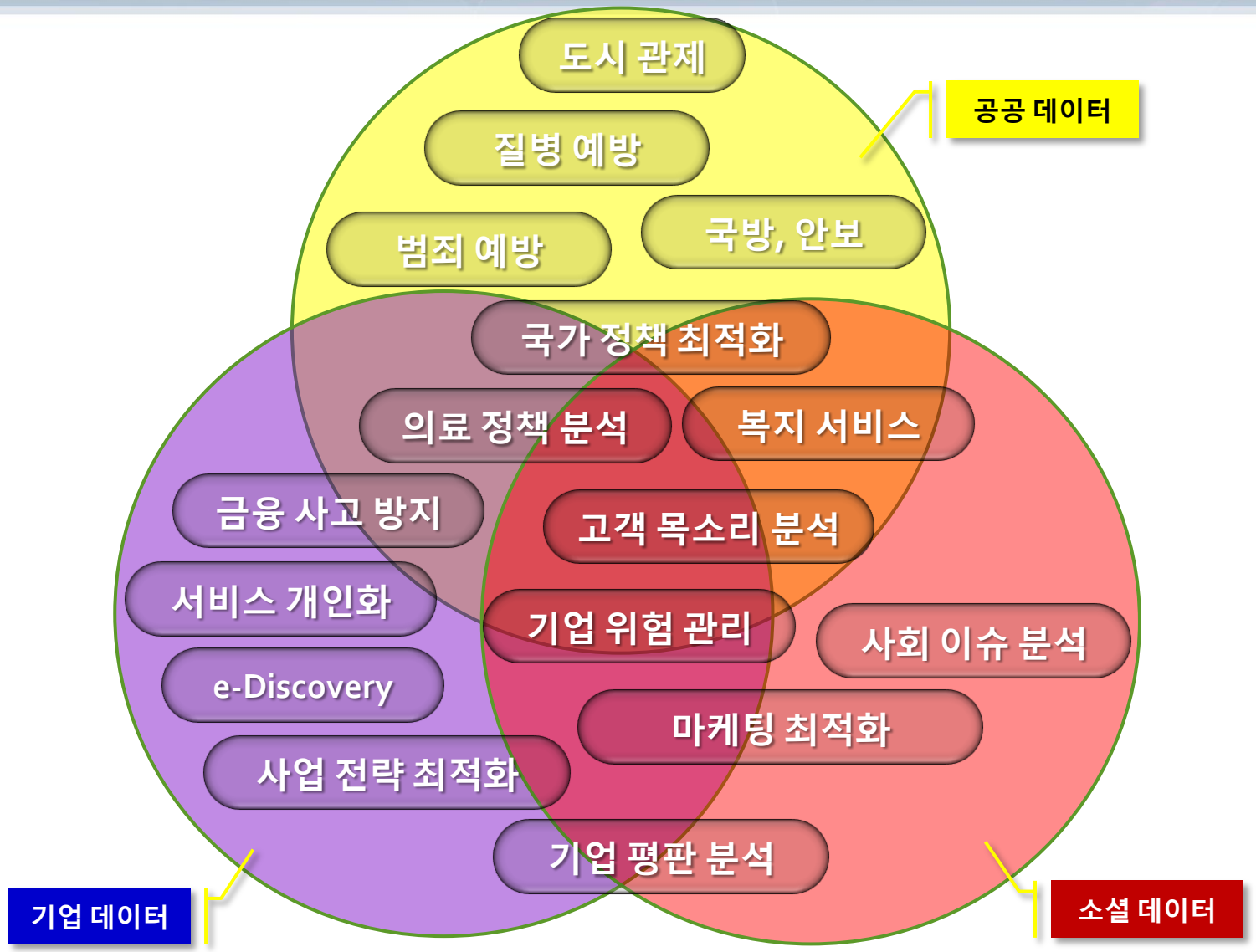
Software Expert

The Age of Data

Data Scientist

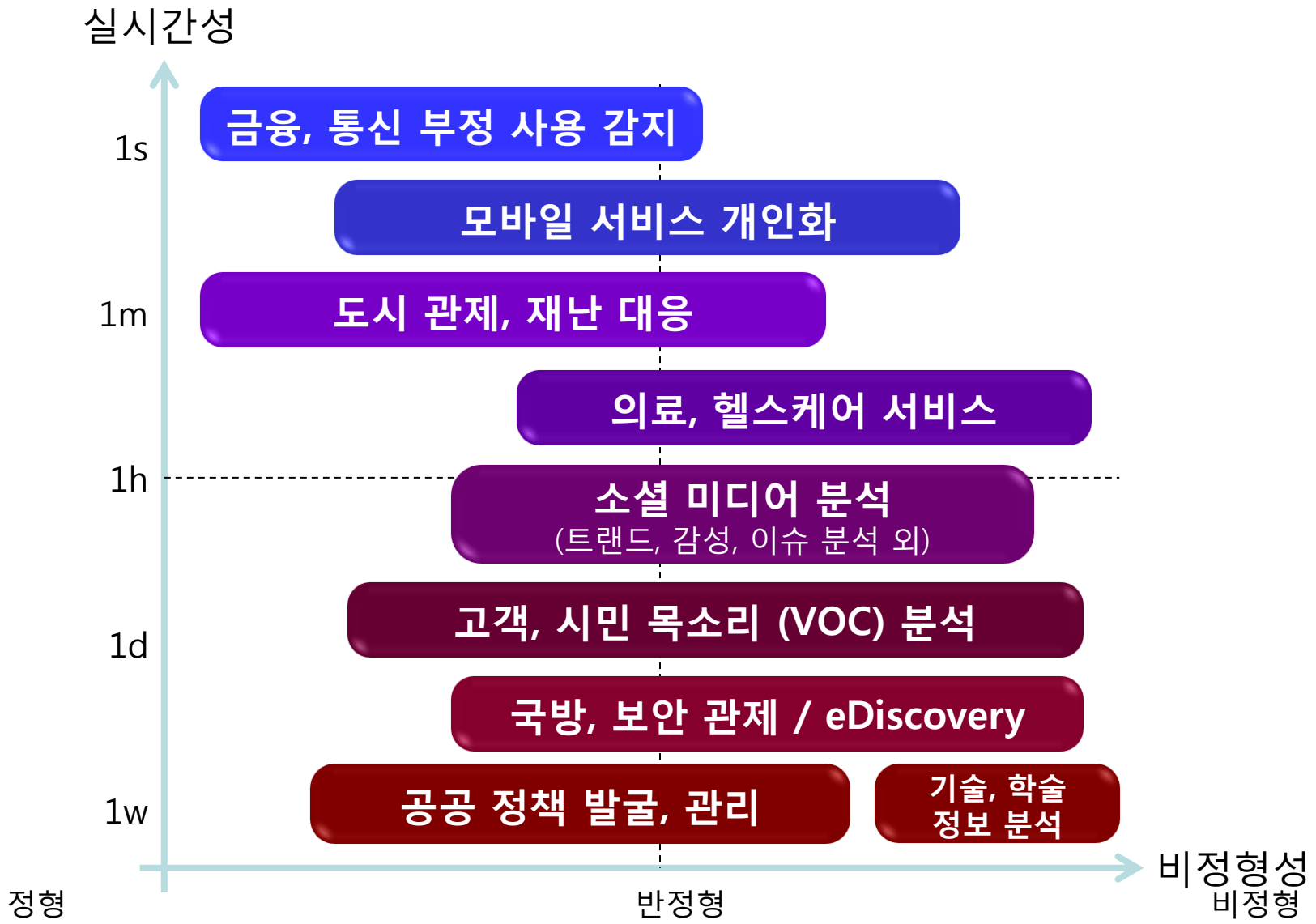
Saltlux

빅데이터 응용 Biz 분야



Saltlux

빅데이터의 응용 영역

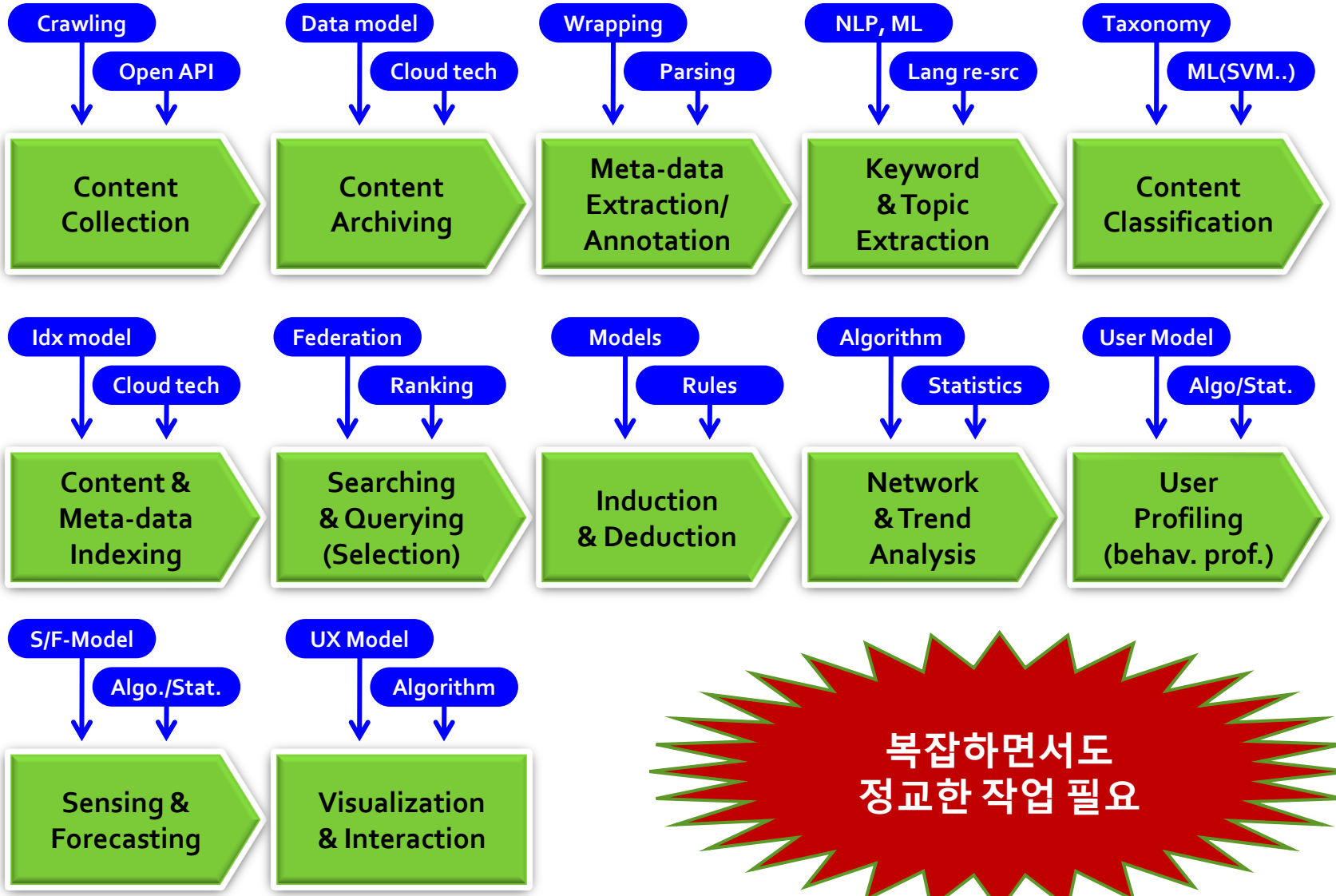


Saltlux

3. Big Data 분석

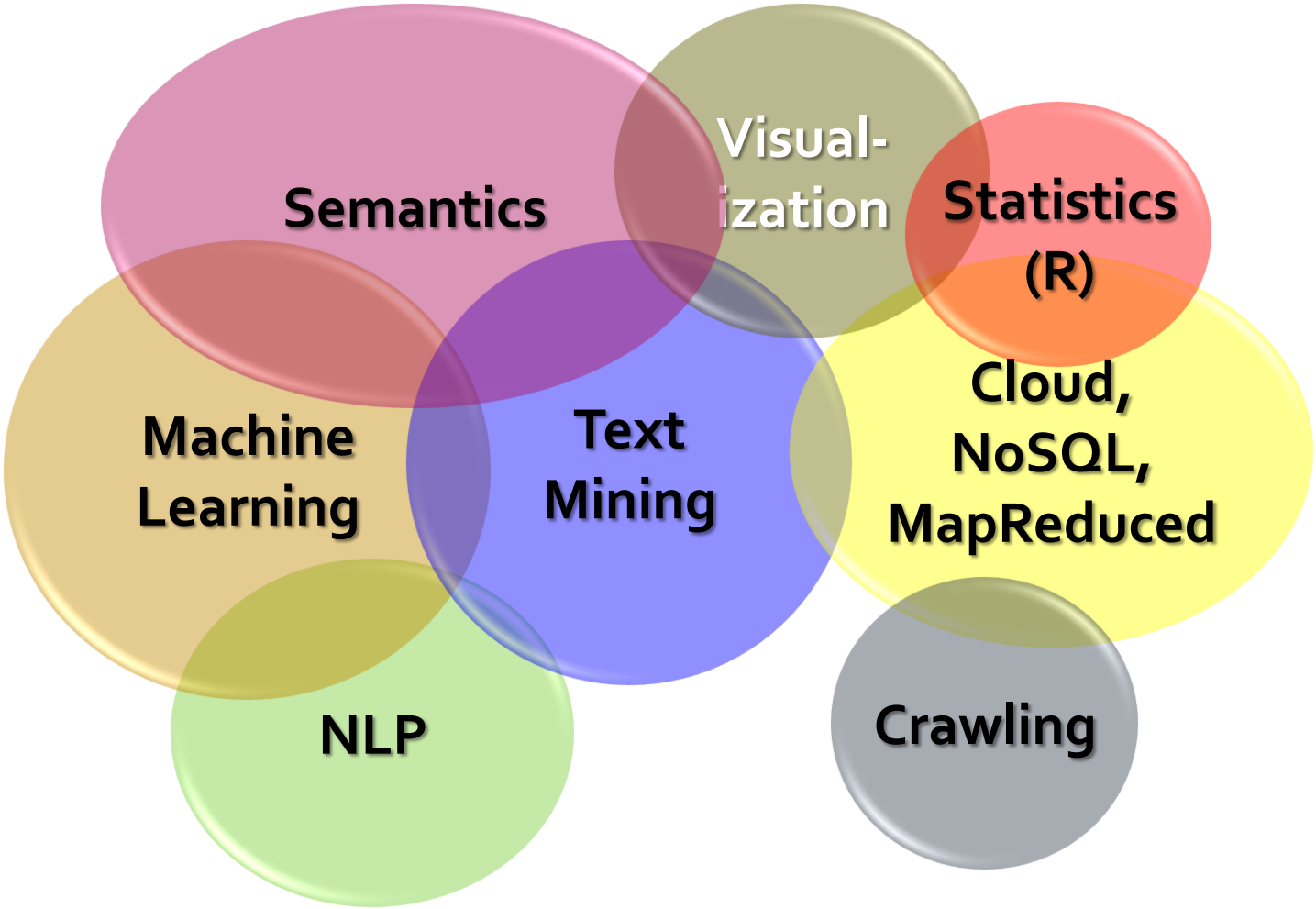
Saltlux

Big Data 분석 절차



Saltlux

Big Data 분석을 위해 필요한 기술

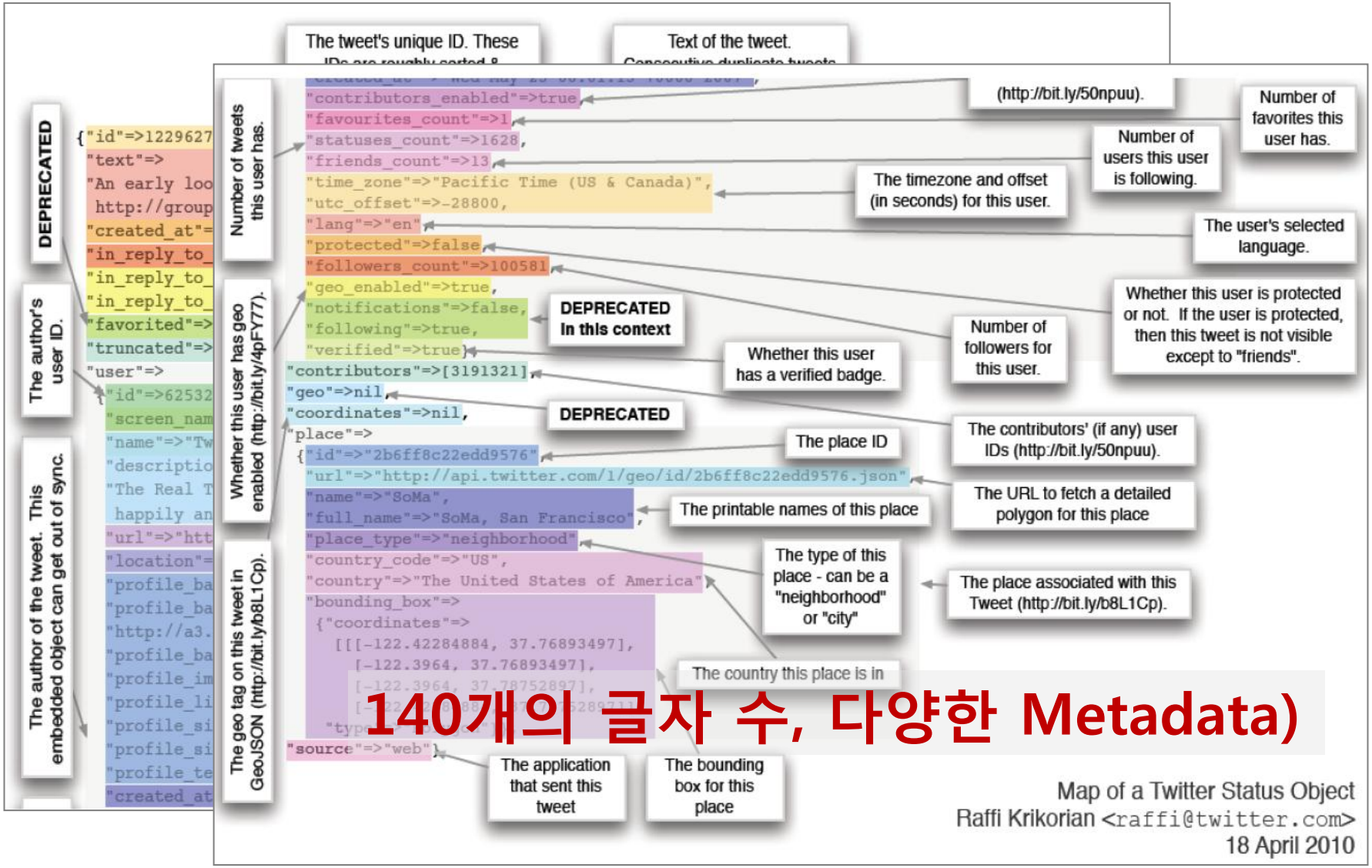


Saltlux

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)



140개의 글자 수, 다양한 Metadata)

Saltlux

Explicit metadata

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

Implicit metadata

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

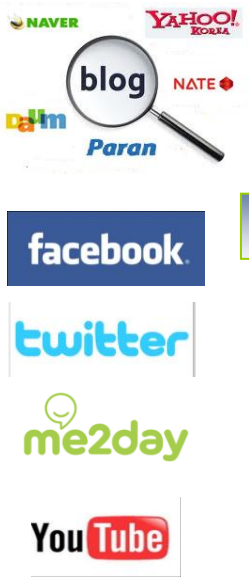
Saltlux – O2 수집/분석 infra

정보 소스

수집 정보 정제 / 가공

분석 / 관리

사회 미디어



뉴스

• 국내 뉴스 수집

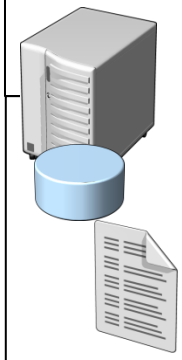
수집

수집 서버

수집 서버

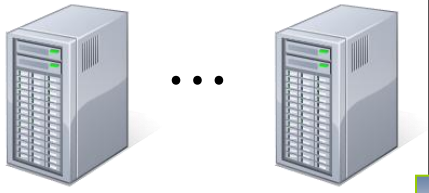
• • •

URL
목록 서버



• Blog, Twitter 등 Social Media 수집
• News 수집
• 주기적, 정책 기반 수집

색인 / 검색

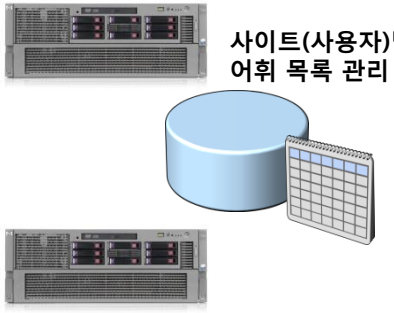


• 수집 콘텐츠, 메타정보 색인
• 콘텐츠,메타 정보 정보
• 대용량 분산 색인, 검색

저장소

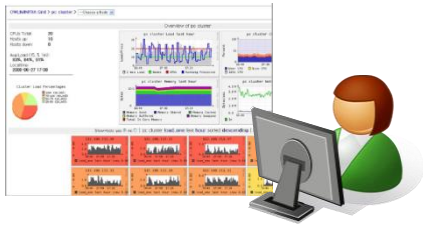
• Storage (10~100 TB)
• 고가용성 확보
• 이중화 구성

분석/필터링



사이트(사용자)별 어휘 목록 관리

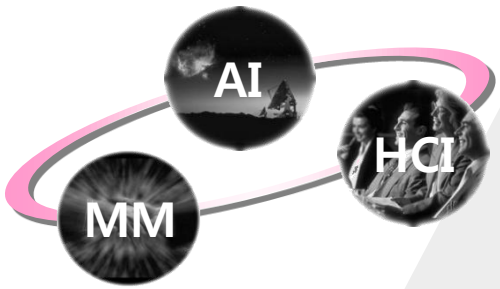
모니터링



• 서버 (인프라) 관리
• 수집 사이트 목록 관리
• 사이트 별 목록 관리
• 수집 및 색인 관리

Saltlux

필요기술: NLP



자연어 처리

인간의 언어를 컴퓨터가 이해할 수 있도록 하기 위한 방법론을 연구하는 것

노트북을 사고 싶습니다.

형태소분석

문장을 최소 의미 단위인 형태소로 분리하고 품사를 부착하는 작업

노트북/명사 | 을/조사 | 사/동사 | ...

구문분석

구문 트리를 생성하고 문장 성분을 분석하는 작업

노트북을 → 사고 → 싶습니다.

의미분석

문장 요소들의 의미적 적합성을 분석하는 작업

buy (modal; want, object; notebook)

담화분석

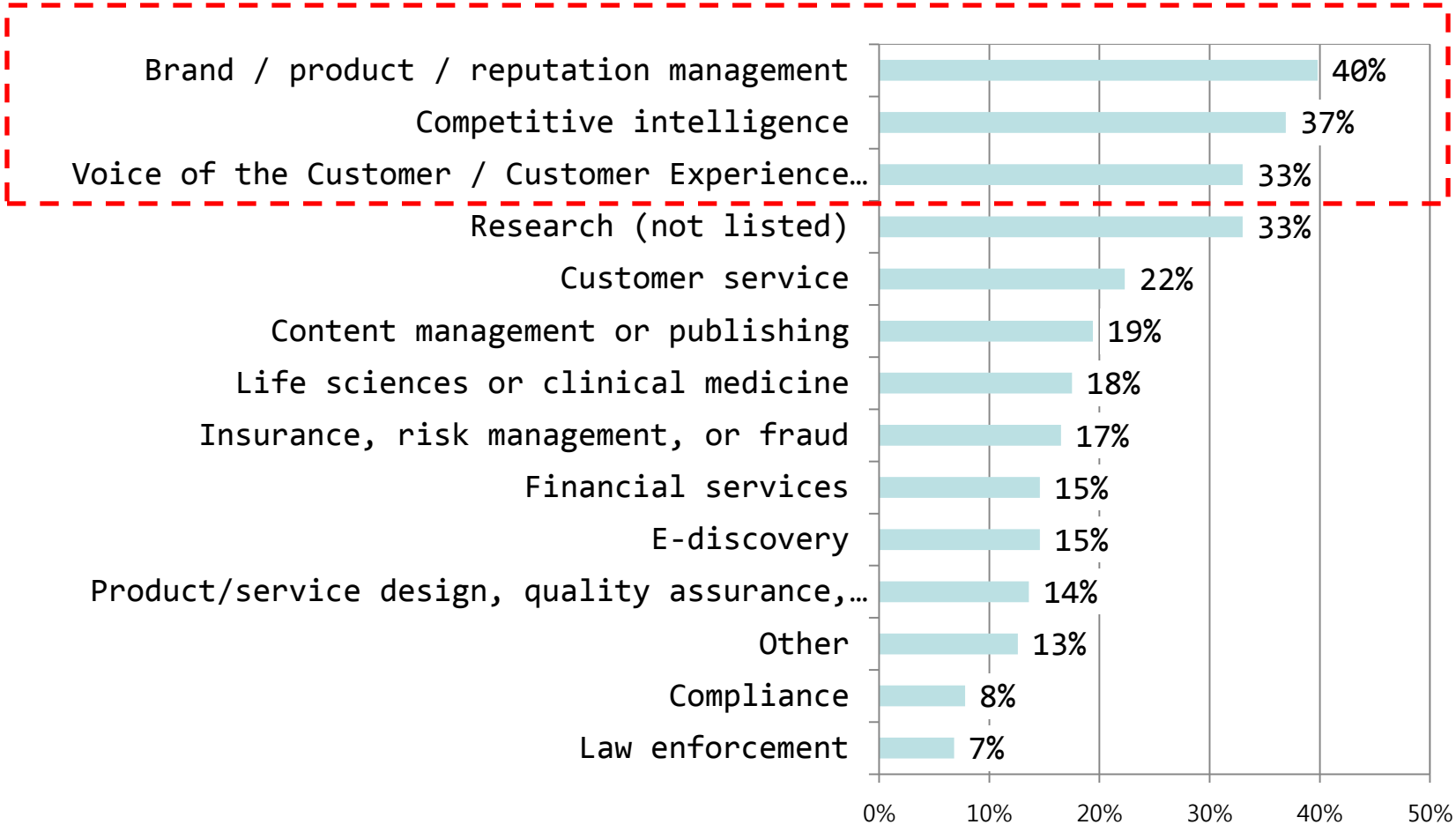
주변 문장들을 고려하여 담화적 적합성을 결정하는 작업

Request →
Ask_brand (어떤 브랜드를 찾으세요?)

Saltlux

필요기술: NLP

Text 분석이 핵심인 중요한 분야는?

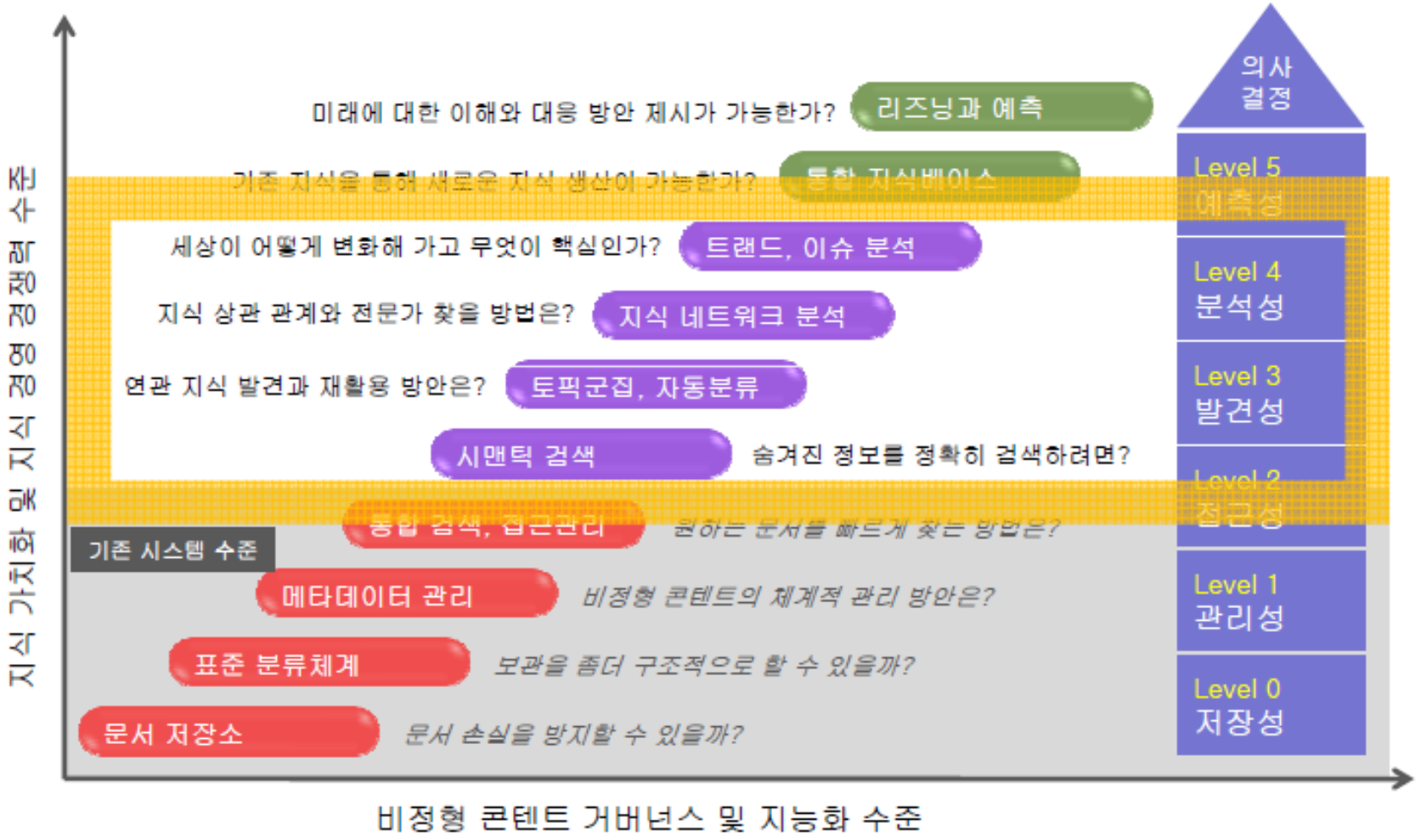


Source: Text Analytics Summit (2011)

Saltlux

필요기술: Text Mining(비정형 데이터 분석)

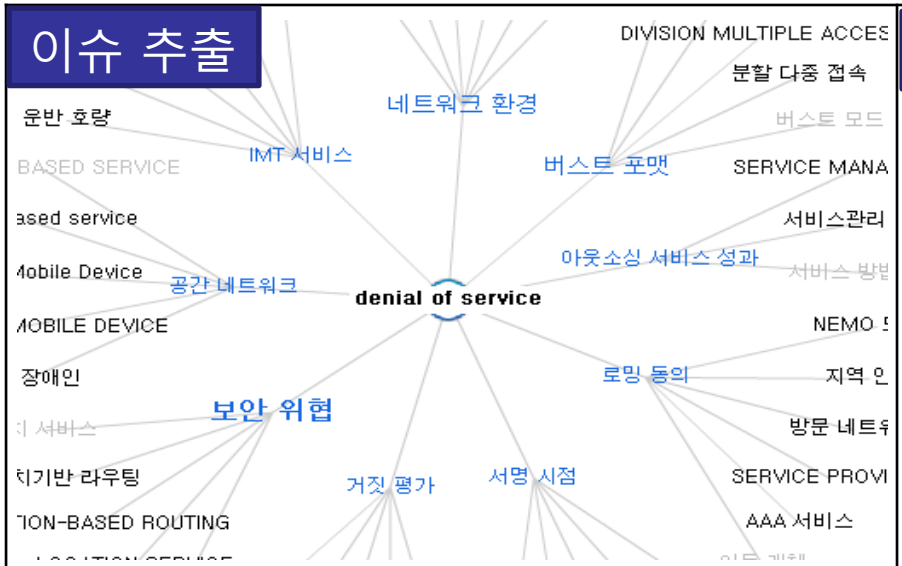
비정형 콘텐트 거버넌스의 발전 단계



Saltlux

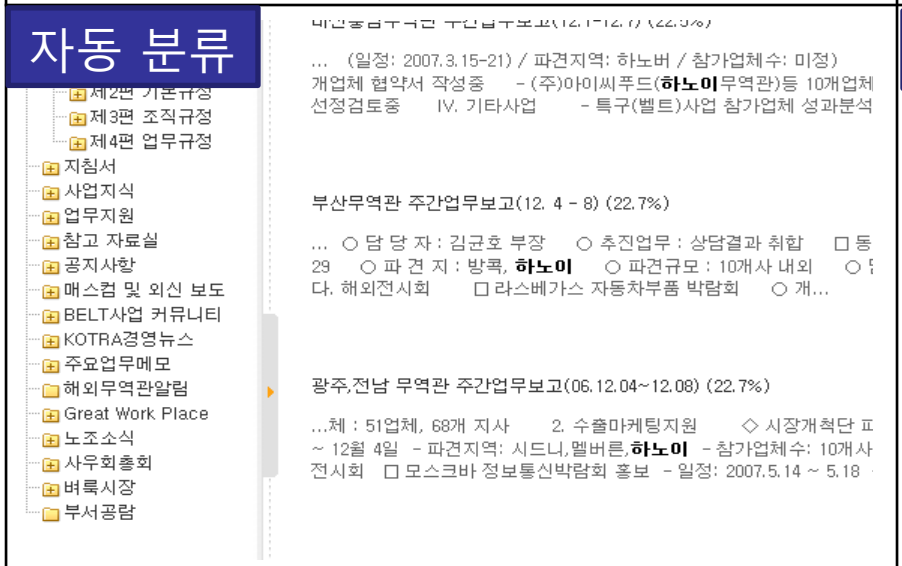
필요기술: Text Mining(비정형 콘텐츠 분석)

이슈 추출

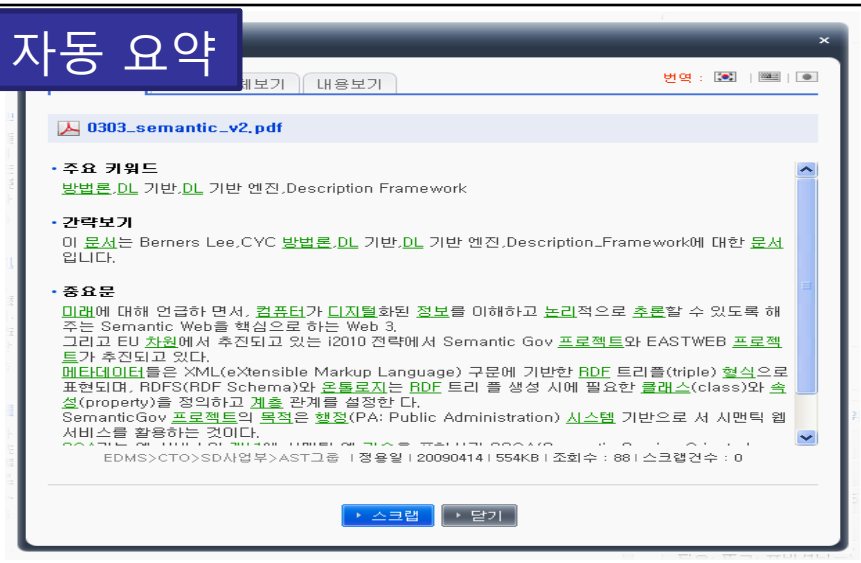


자동 군집

자동 분류



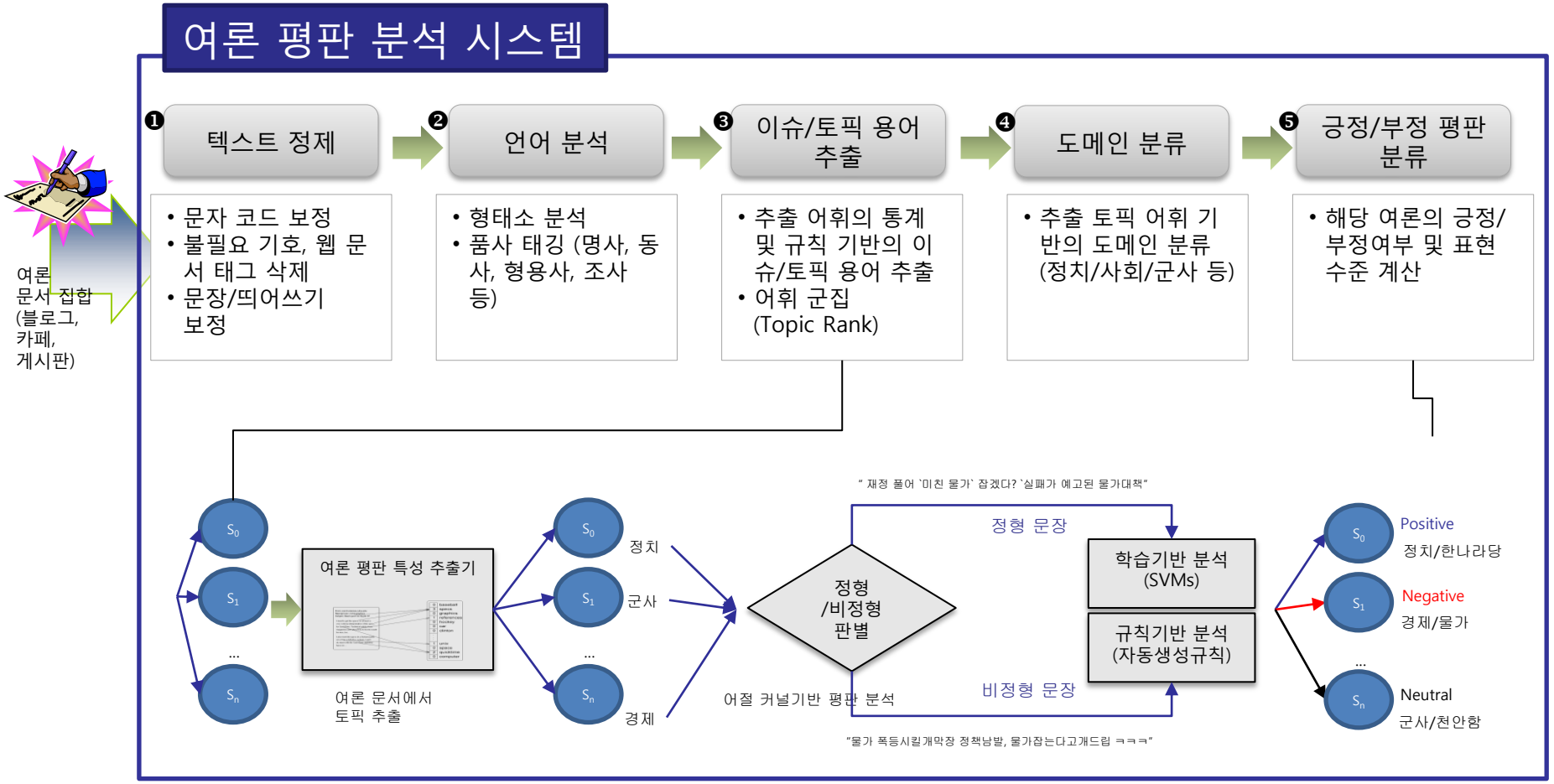
자동 요약



Saltlux

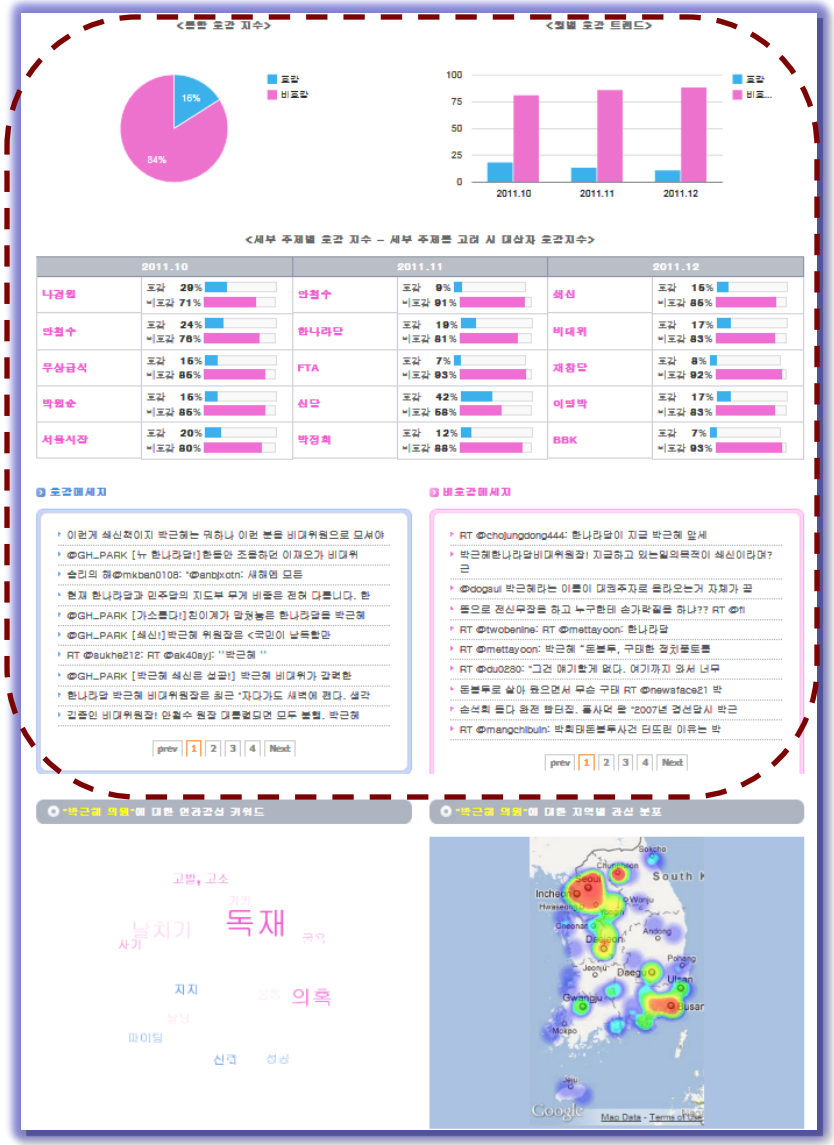
여론 데이터 평판 분석

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.



Saltlux

Sentiment Analysis

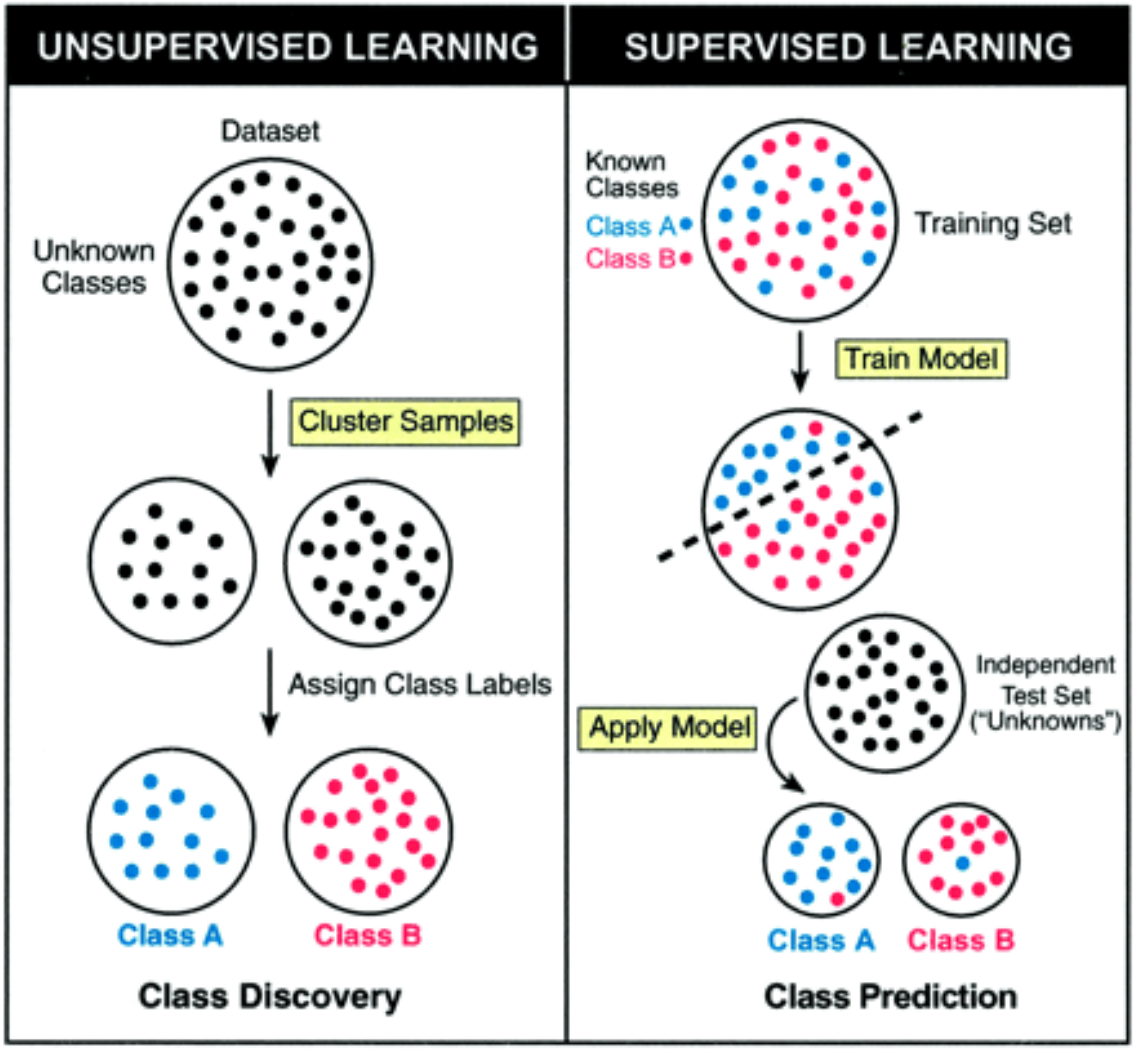


Saltlux

필요기술: Machine Learning

사용자의 개입 불필요

군집(Clustering)

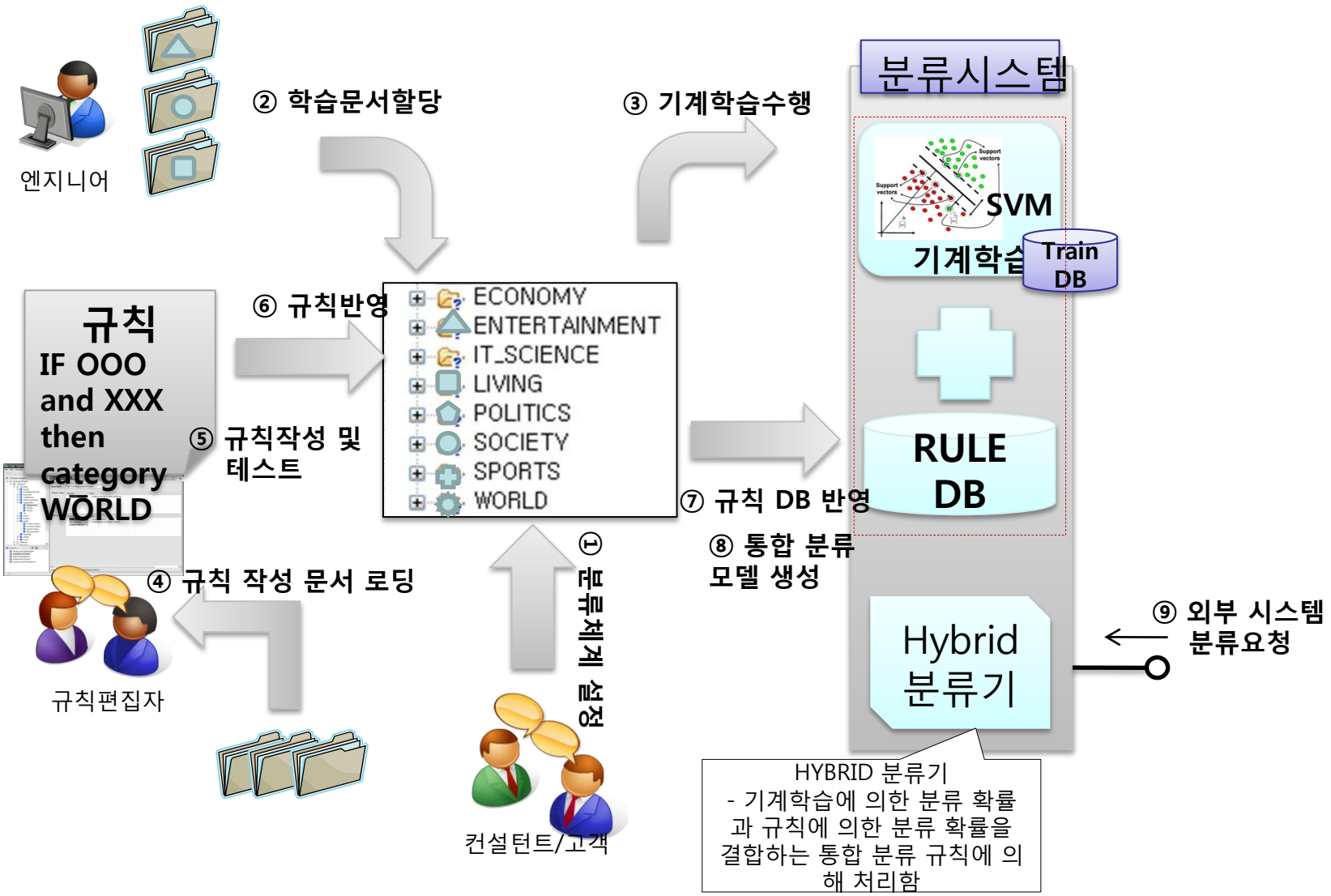


사용자의 개입 필요

분류(Classification)

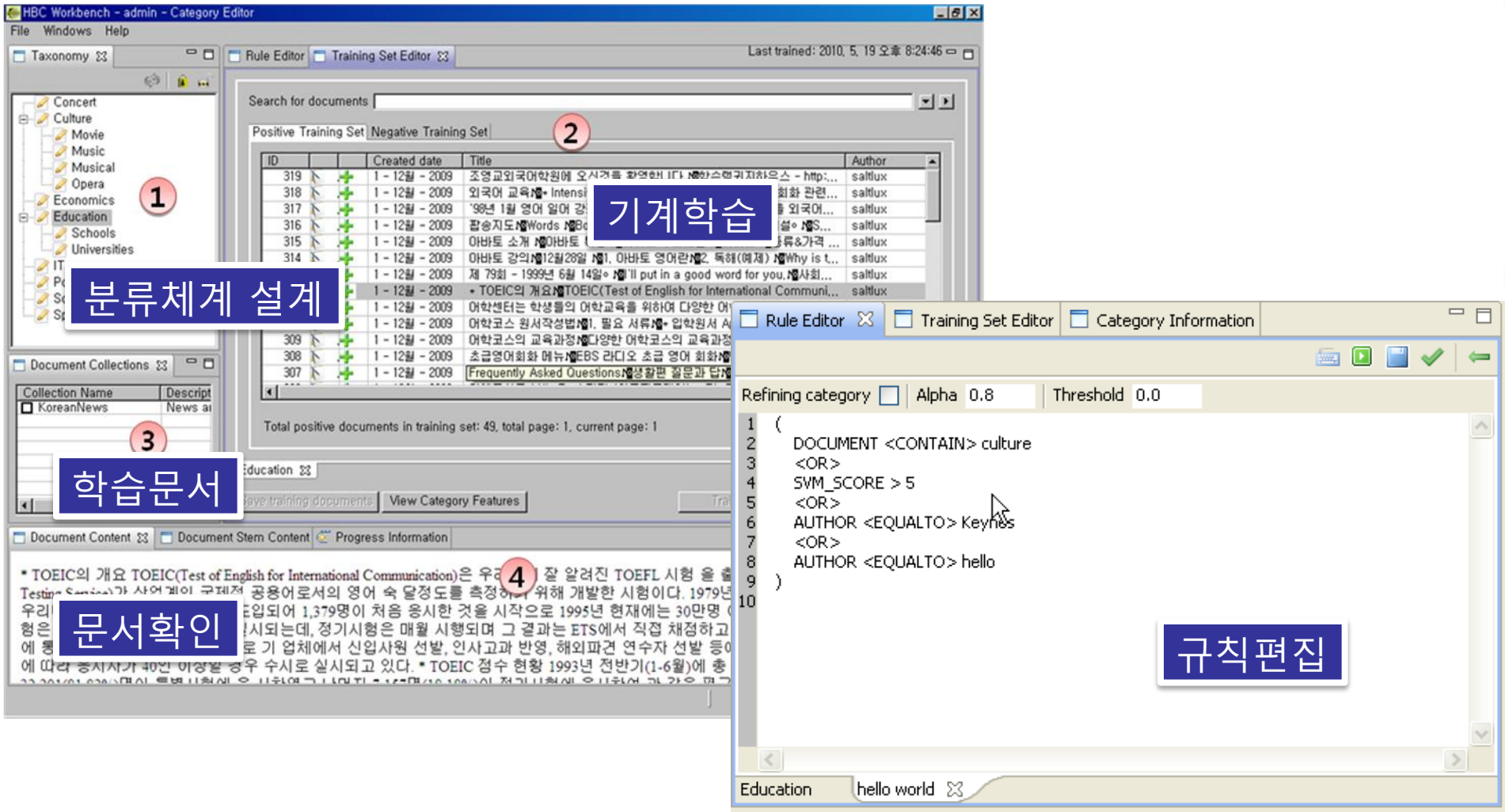
Saltlux

Saltlux – 하이브리드 분류기 적용 프로세스



Saltlux

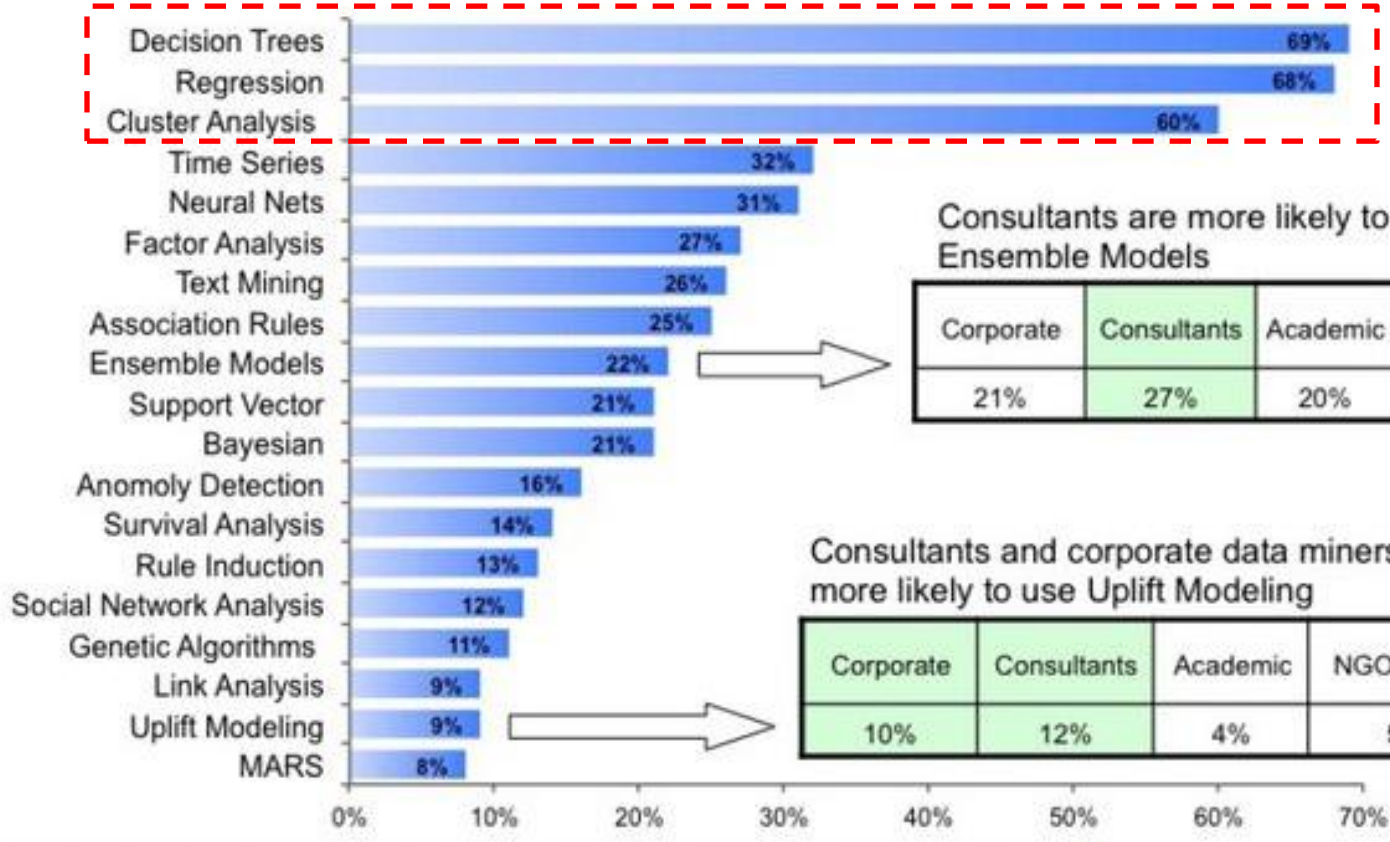
Saltlux – 하이브리드 자동분류



Saltlux

필요기술: Mining

- 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

Consultants are more likely to use Ensemble Models

Corporate	Consultants	Academic	NGO / Gov't
21%	27%	20%	18%

Consultants and corporate data miners are more likely to use Uplift Modeling

Corporate	Consultants	Academic	NGO / Gov't
10%	12%	4%	5%

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

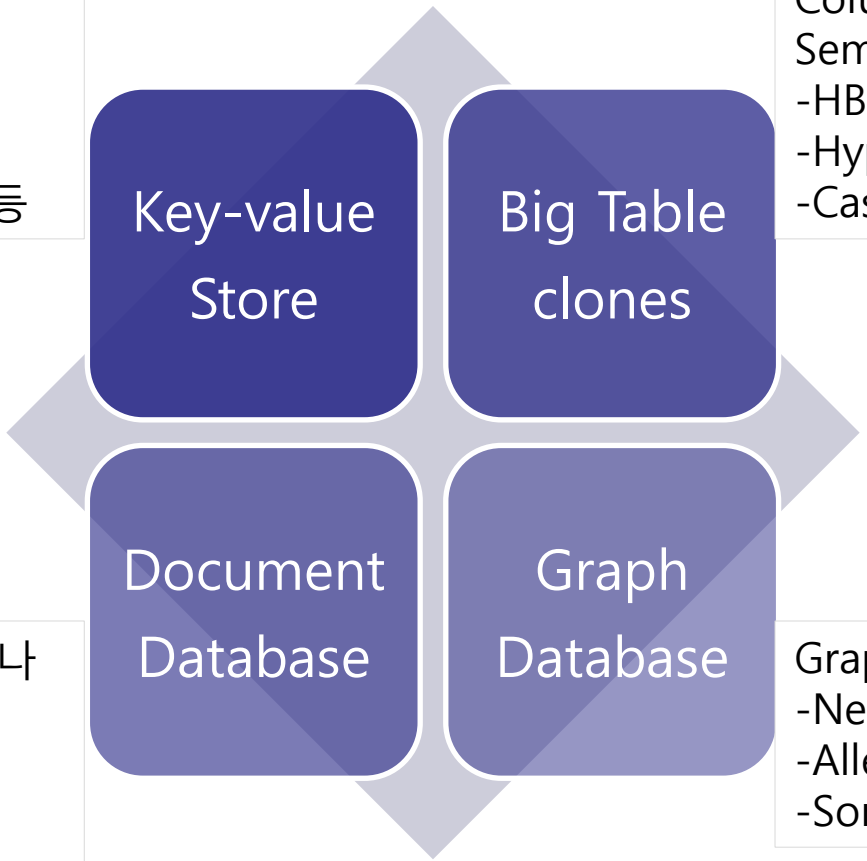
Vendors were excluded from this analysis.

Source : Lexer Analytics (2011)

Saltlux

필요기술: NoSQL (Scalability)

Global collection of
Key-value pair
-Dynomite
-Voldemort
-Tokyo{Tyrant, Cabintet} 등



Column oriented.
Semi-Structured
-HBase
-Hypertable
-Cassandra 등

Key-value store와 유사하나
System이 값을 이해
-CouchDB
-MongoDB
-Redis 등

Graph theory 적용
-Neo4J
-AllegroGraph
-Sones GraphDB 등

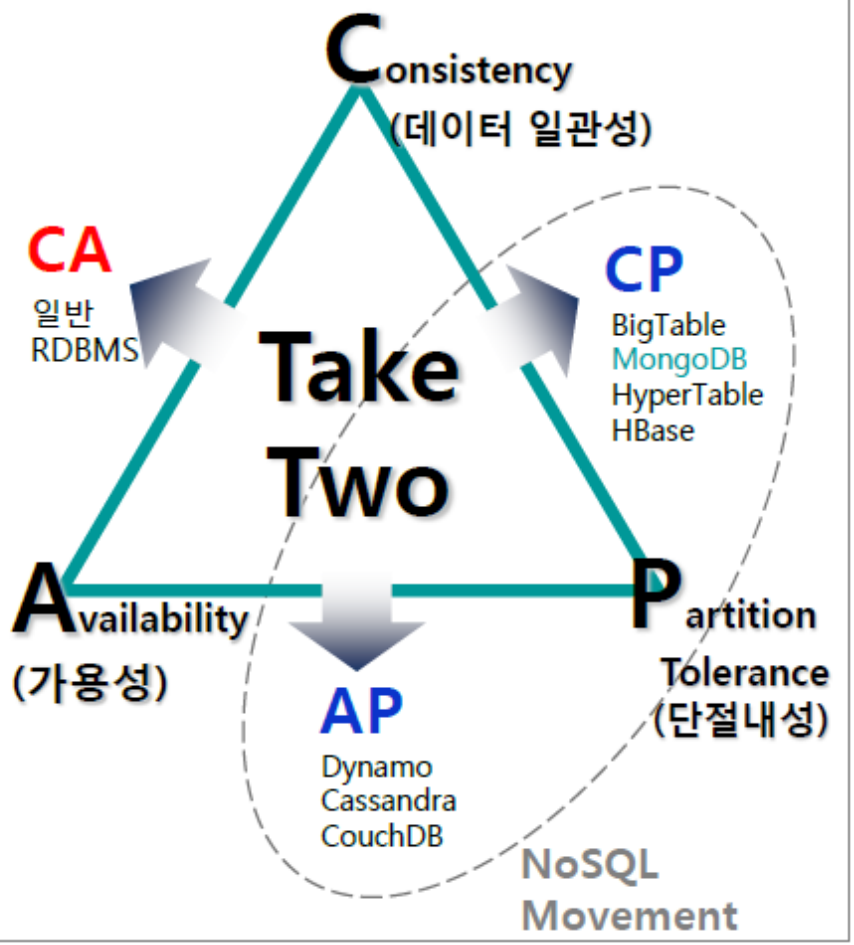
Source : Tobias Ivarsson (2010)

Saltlux

필요기술: NoSQL(Indexing / Archiving)

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

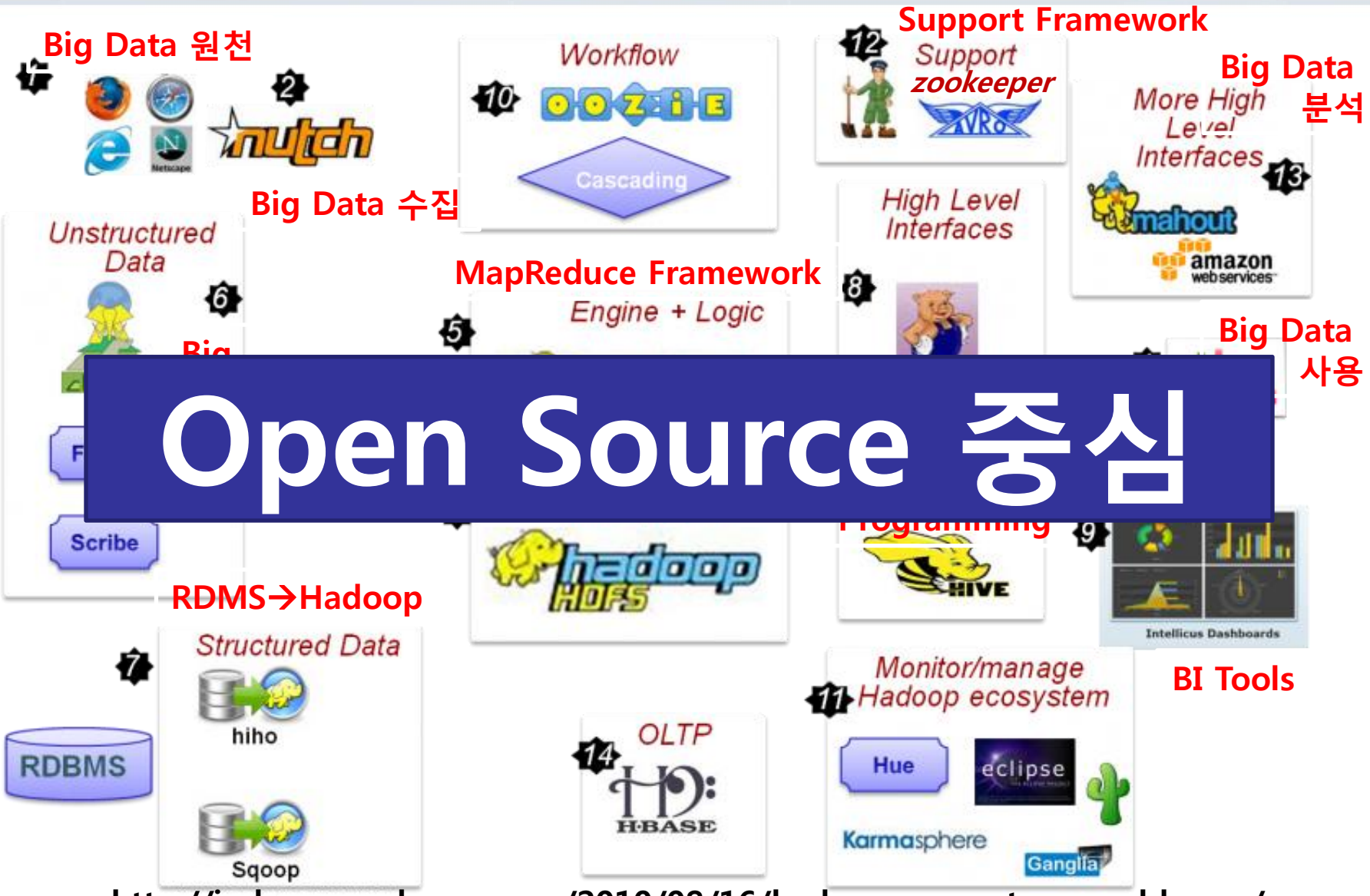
CAP Theorem과 데이터 관리전략



CAP Theorem의 시사점

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

필요기술: Big Data (Scalability) ➔ Hadoop



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

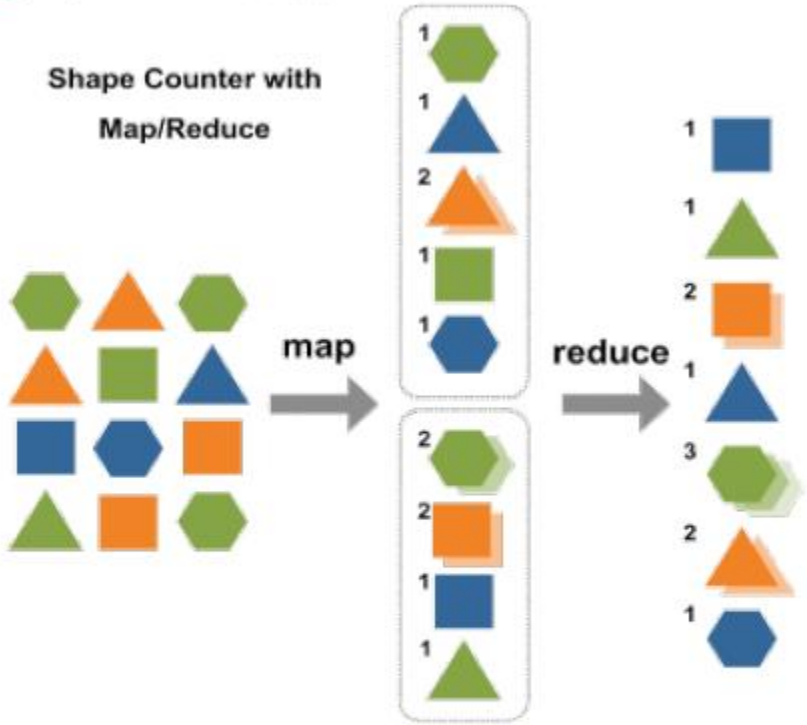
Saltlux

필요기술: Big Data 저장관리

주요 기술	설명
빅 테이블 (Big Table)	• 구글 파일 시스템(Google File System) 상에 구축된 상용 분산 데이터 베이스 시스템 • HBase에 영향을 미침
카산드라 (Cassandra)	• 분산 시스템에서 방대한 분량의 데이터를 처리할 수 있도록 디자인된 오픈소스(무료) 데이터베이스 관리 시스템 • 원래 페이스북에서 개발했으며 지금은 아파치 소프트웨어 재단의 한 프로젝트로 관리되고 있음
분산 시스템 (Distributed System)	• 동시에 일을 처리하기 위해 네트워크로 연결된 컴퓨터들의 집합으로 단일 또는 다수의 컴퓨터의 리소스를 부분적으로 활용함으로써 시스템의 가성비, 안정성 그리고 확장성을 향상시킬 수 있음
구글 파일 시스템 (Google File System)	• 구글에서 개발한 분산 파일 시스템 • Hadoop과 관련되어 있음
Hadoop	• 분산 시스템 상에서 대용량 데이터 처리 분석을 지원하는 오픈소스 프레임워크 • 구글이 개발한 맵리듀스 (MapReduce)를 오픈소스로 구현한 결과물 • 원래 야후!에서 최초 개발되었으며 지금은 아파치 소프트웨어 재단의 한 프로젝트로 관리됨
HBase	• 구글(Google)의 빅테이블(Big Table)을 참고로 개발된 오픈소스 분산 비관계형 데이터베이스 • 원래 Powerset에서 개발했으며, 현재는 아파치 소프트웨어 재단에서 Hadoop의 일환인 프로젝트로 관리되고 있음
MapReduce	• 분산 시스템 상에서 대용량 데이터 세트를 처리하기 위해 구글(Google)이 소개한 소프트웨어 프레임워크로 Hadoop에 구현되어 있음
비관계형 데이터베이스/ Key Value Store	• 비관계형 데이터베이스는 데이터를 테이블(행, 컬럼)에 저장하지 않는 데이터베이스이며 관계형 데이터베이스와 대조되는 개념임 • Key Value Stores를 사용하면 스키마 없는 엔티티(noSQL)를 관리할 수 있음

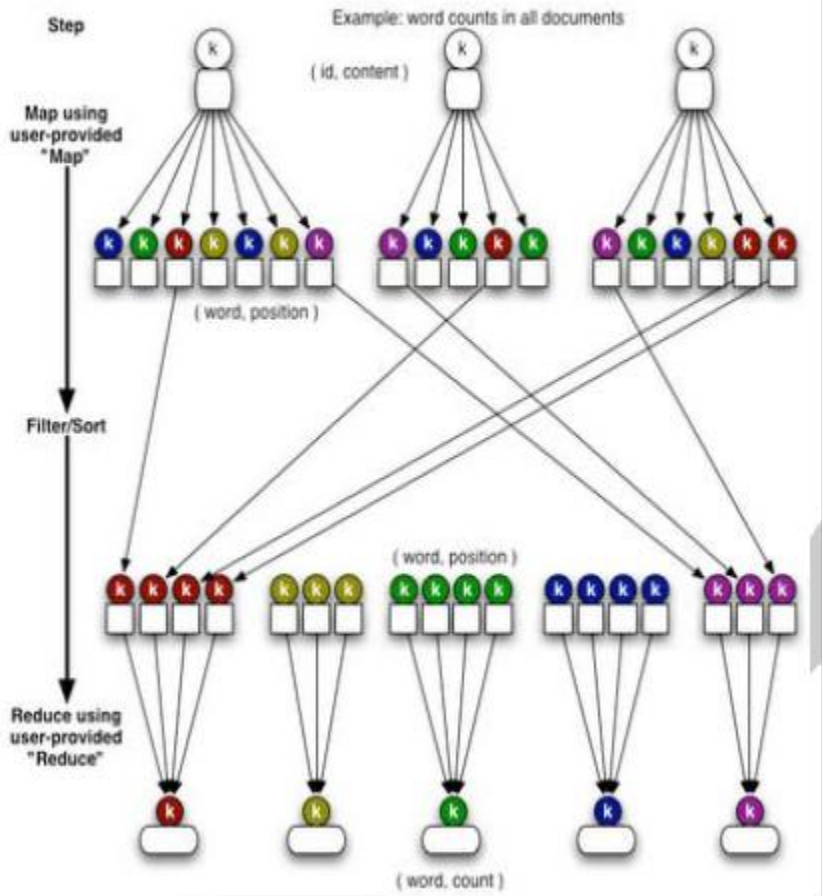
필요기술: Map Reduce

Map / Reduce의 개념



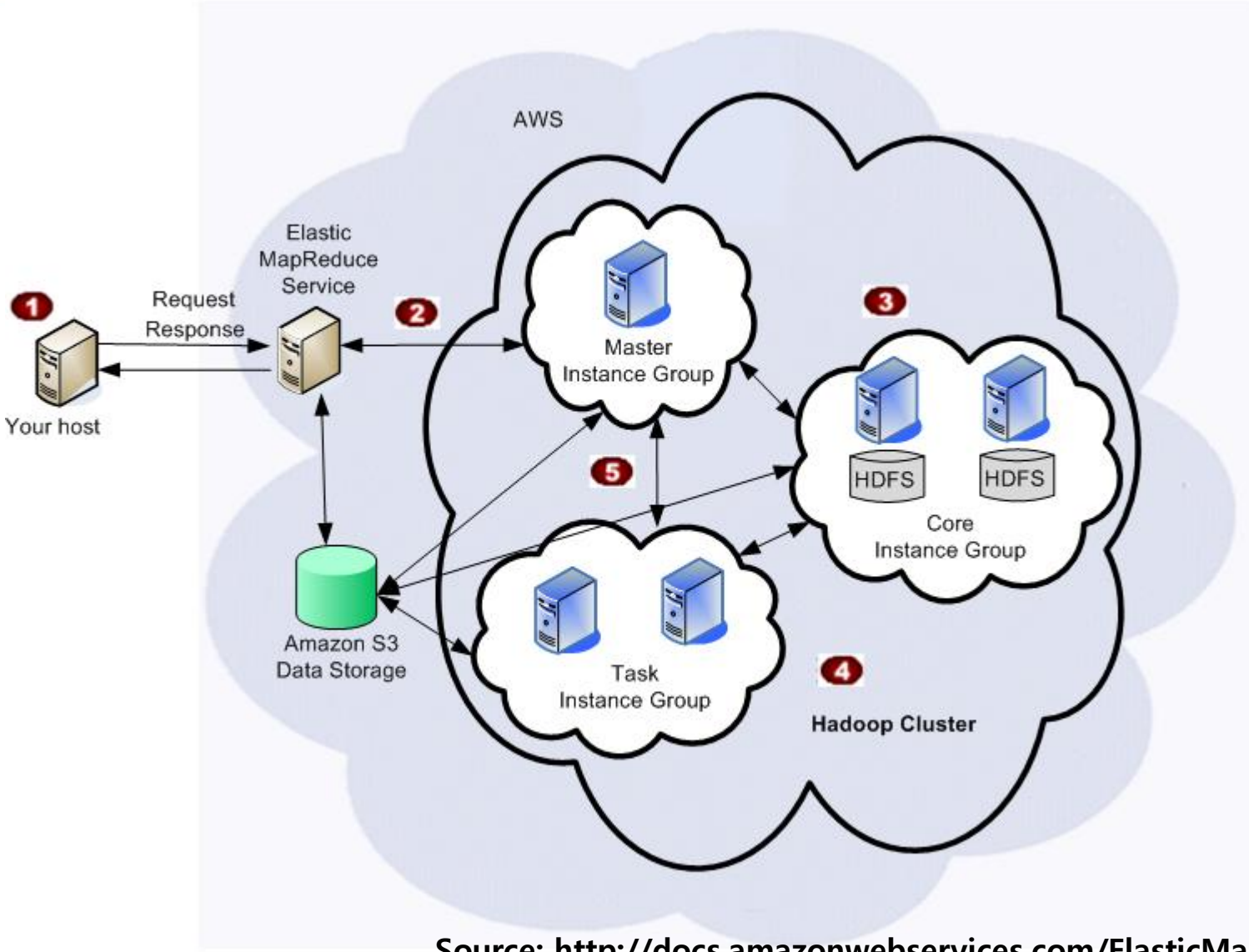
Map : 흩어져 있는 데이터를 Key, Value 의 형태로 연관성 있는 데이터 분류로 묶는 작업.
Reduce : Map 화 한 작업 중 중복 데이터를 제거하고 원하는 데이터를 추출하는 단계.

Map / Reduce의 예시



Saltlux

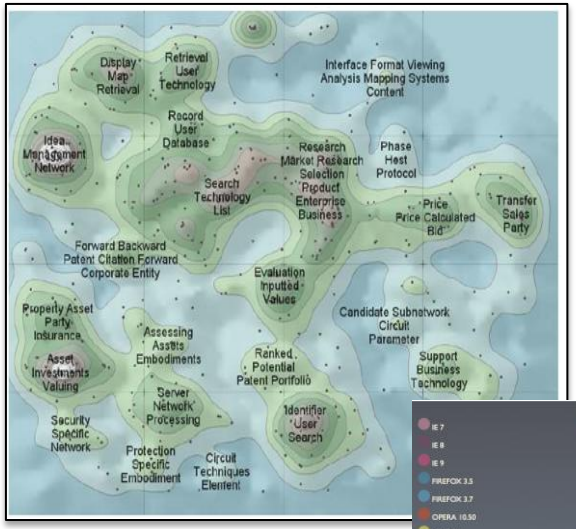
Big Data 처리 인프라 서비스 (Amazon Elastic Map Reduce)



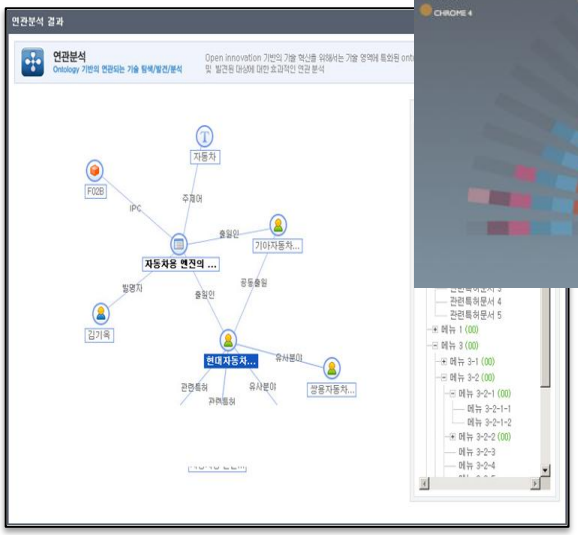
Source: http://docs.amazonwebservices.com/ElasticMapReduce

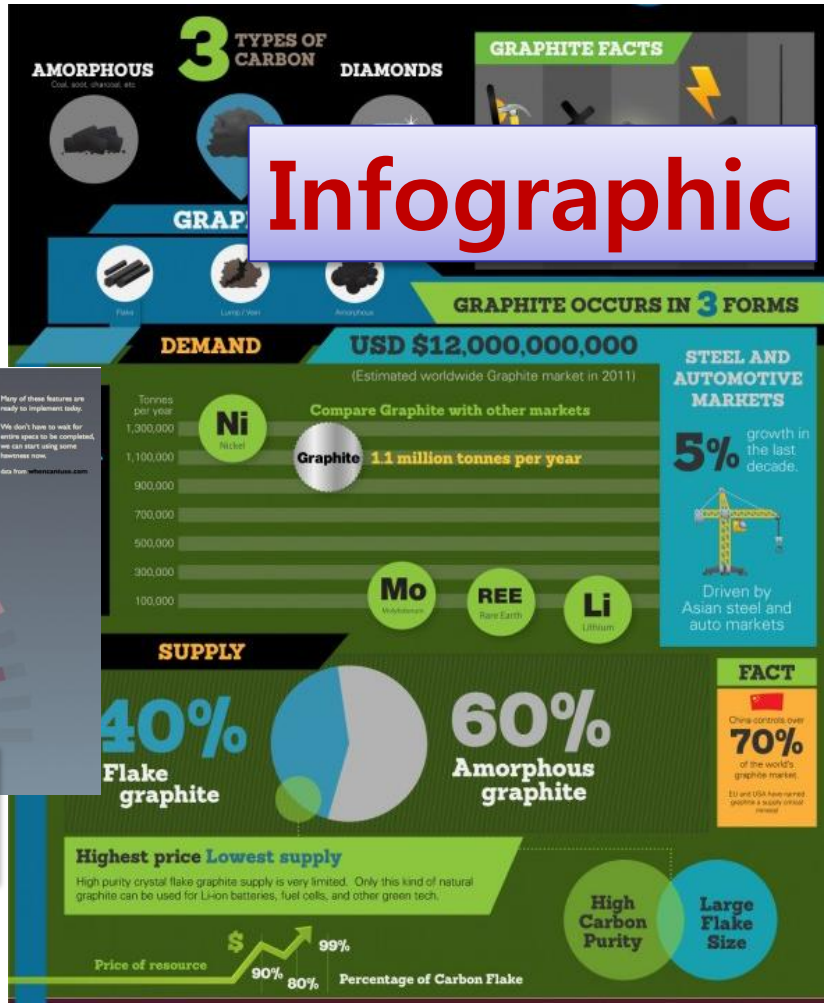
Saltlux

필요기술: Visualization



HTML5





Infographic

Saltlux

R is a tool for…

Data Manipulation

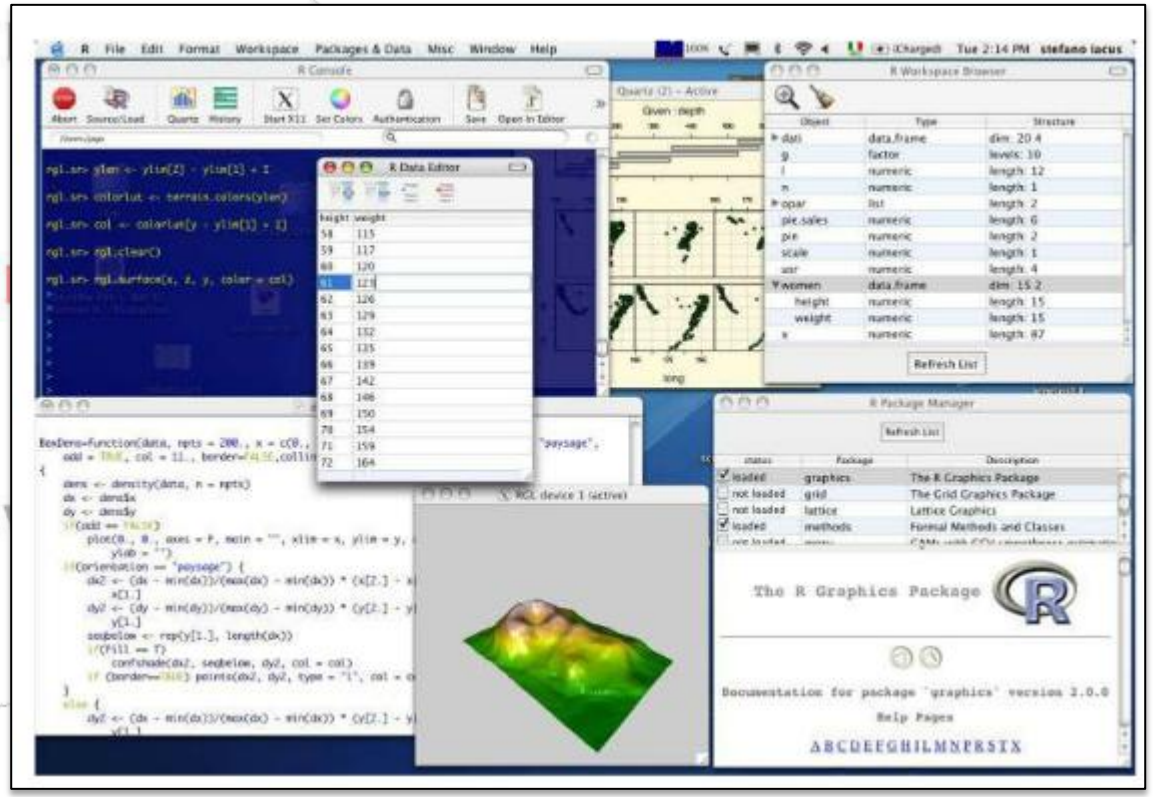
- connecting to data sources
- slicing & dicing data

Modeling & Computation

- statistical modeling
- numerical simulation

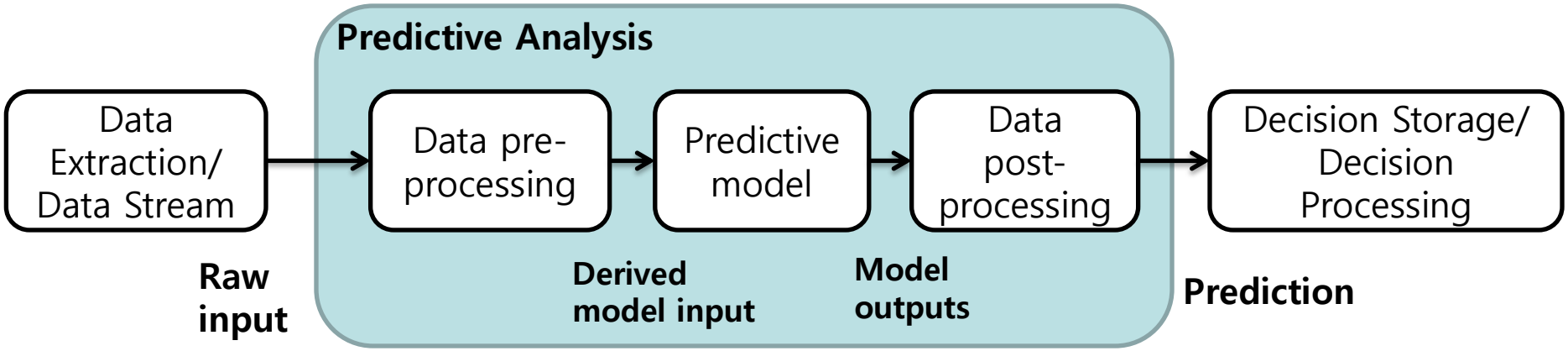
Data Visualization

- visualizing fit of models
- composing statistical graphics



필요기술: 통계(Predictive Analysis)

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

Saltlux

필요기술: 통계(Predictive Analysis)

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.

Sentiment mining showed a sharp change in ton around Egypt ahead of President Mubarak's oust

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

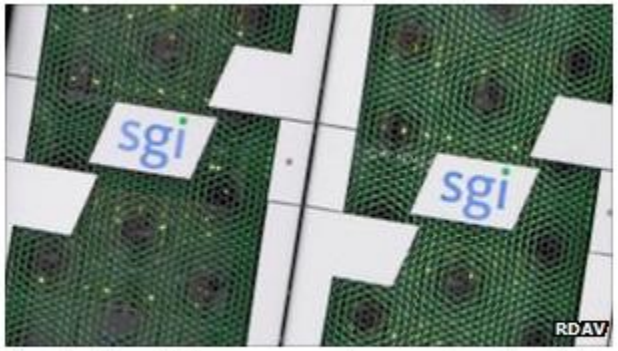
Kalev Leetaru, from the University of Illinois' Institute for Computing in the Humanities, Arts and Social Science, **presented his findings** in the journal First Monday.

Mood and location

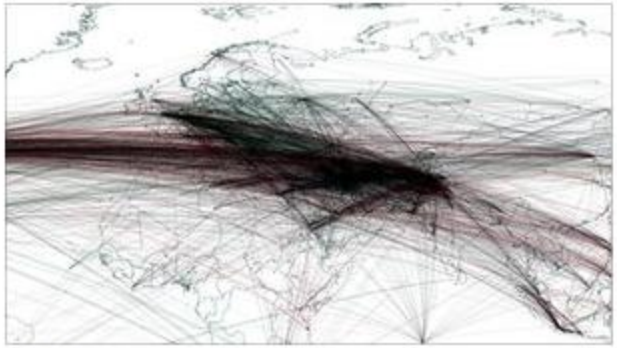
Related Stories

Scientists aim to simulate Earth

New weather supercomputer test



The Nautilus SGI supercomputer crunched the 100 million articles

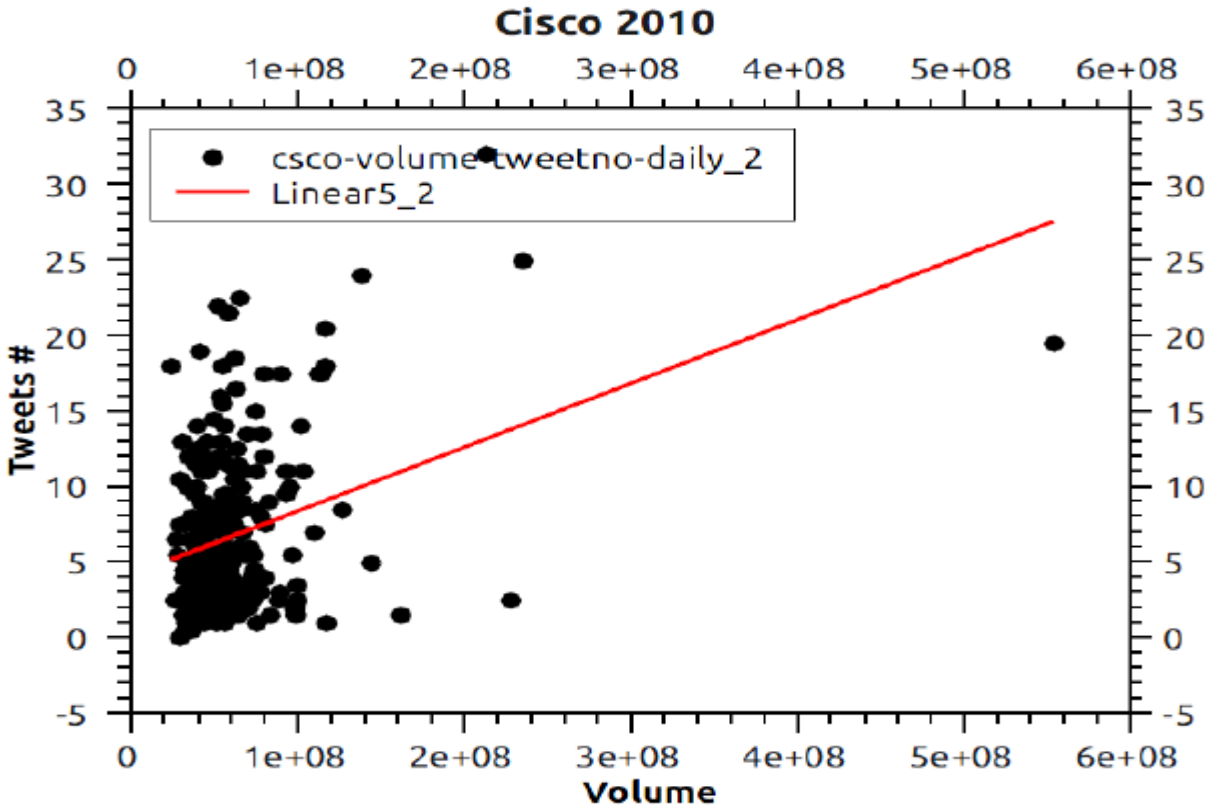


Media reports mentioning Osama Bin Laden may have helped narrow down his location

Source : BBC

Saltlux

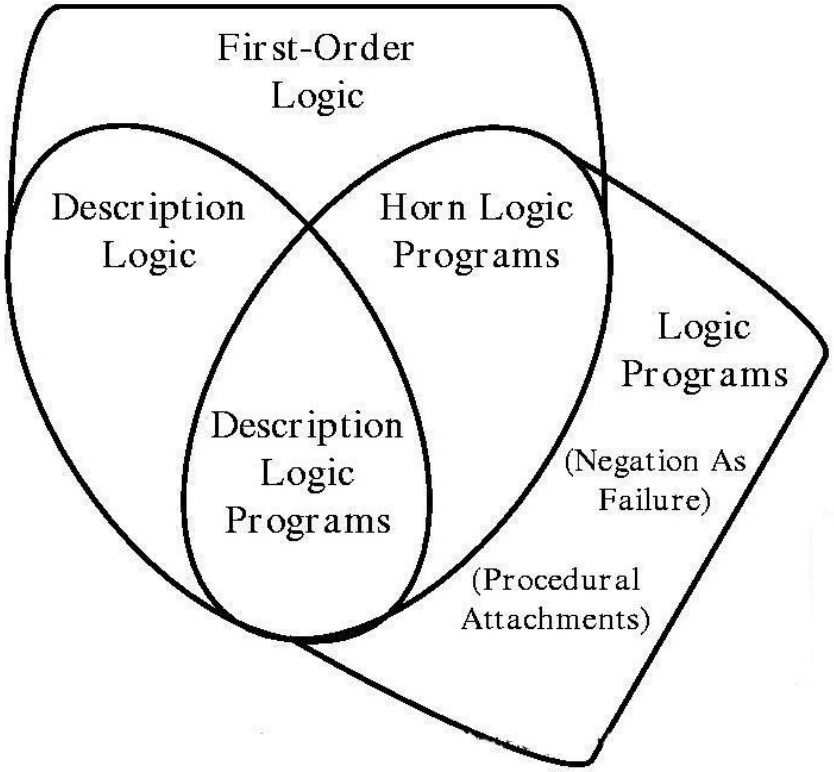
The relationship between Cisco daily volume of stock trades and the volume of tweets – fairly good correlation was found



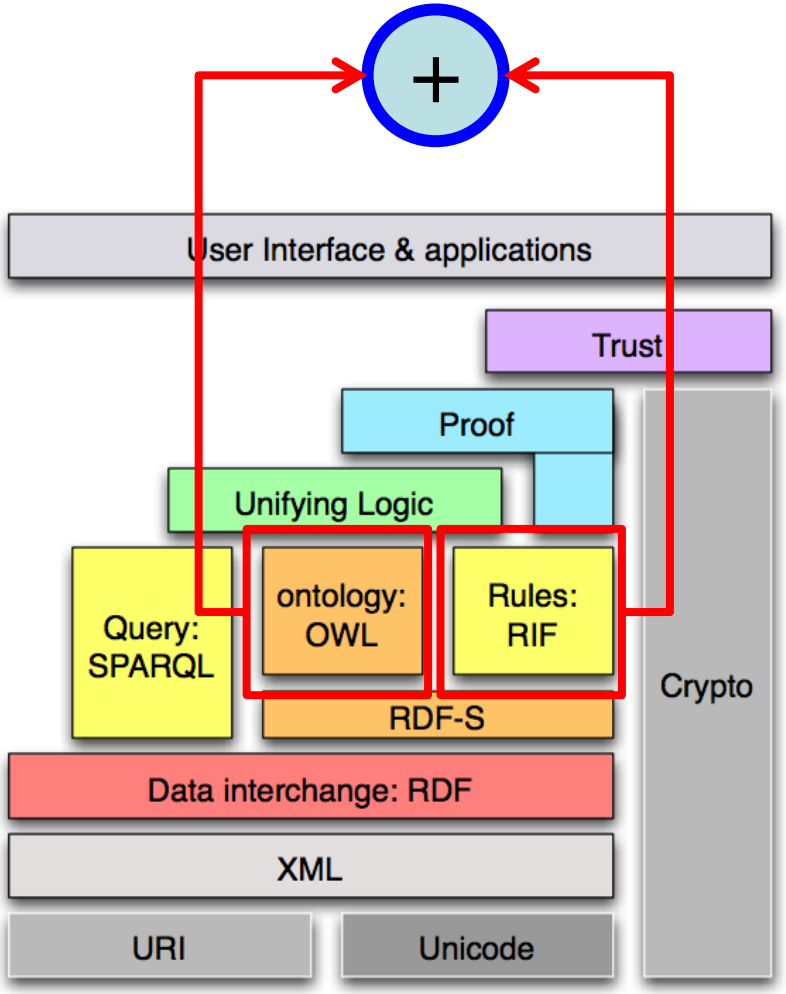
Source : SEMTECH 2011

Saltlux

필요기술: Semantic



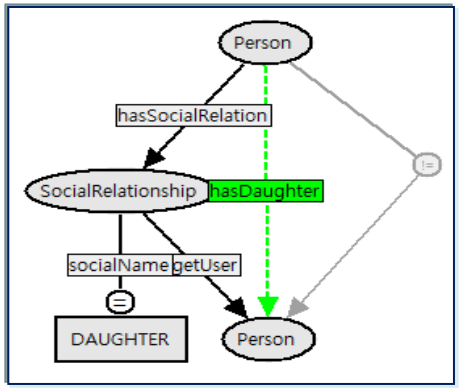
The relationships among different formalisms (Benjamin Grosof)

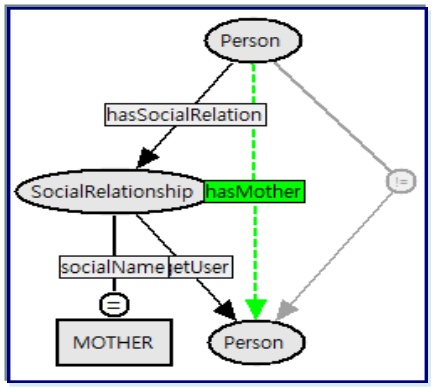


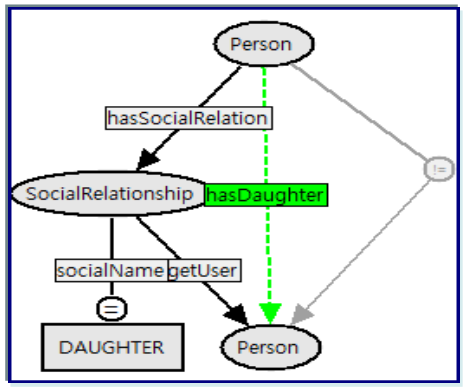
Semantic Web Architecture

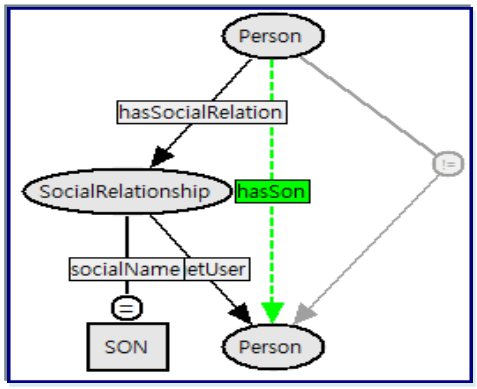
Saltlux

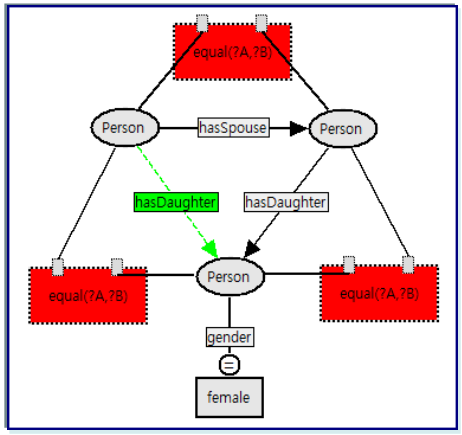
필요기술: Semantic (Hybrid Reasoning: DL + Rules)

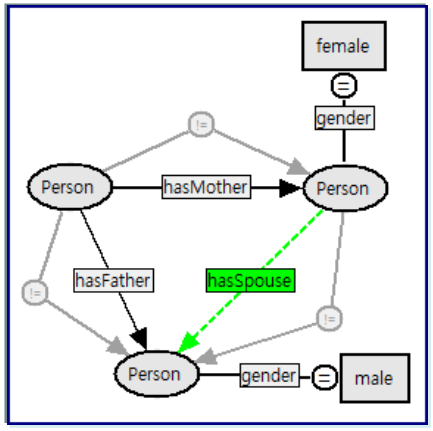


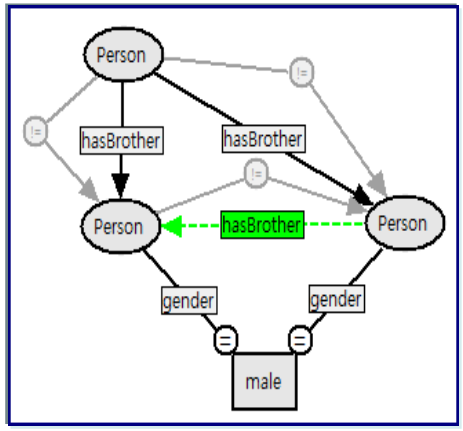


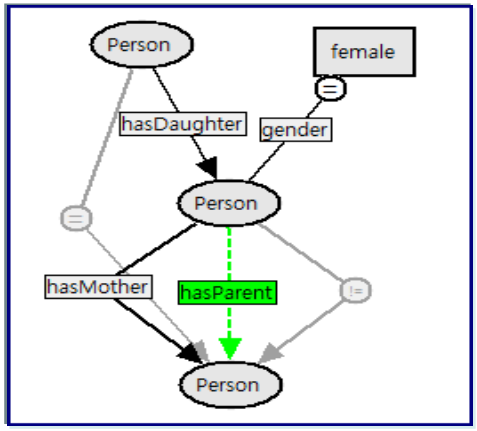






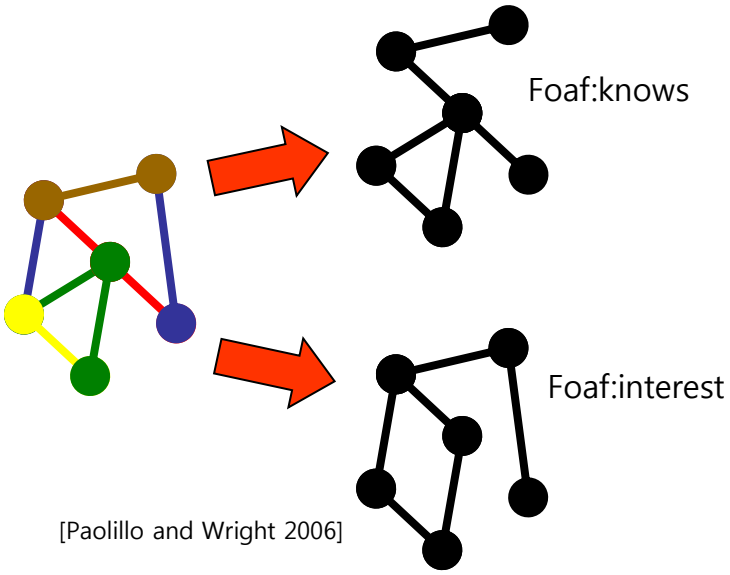






Saltlux

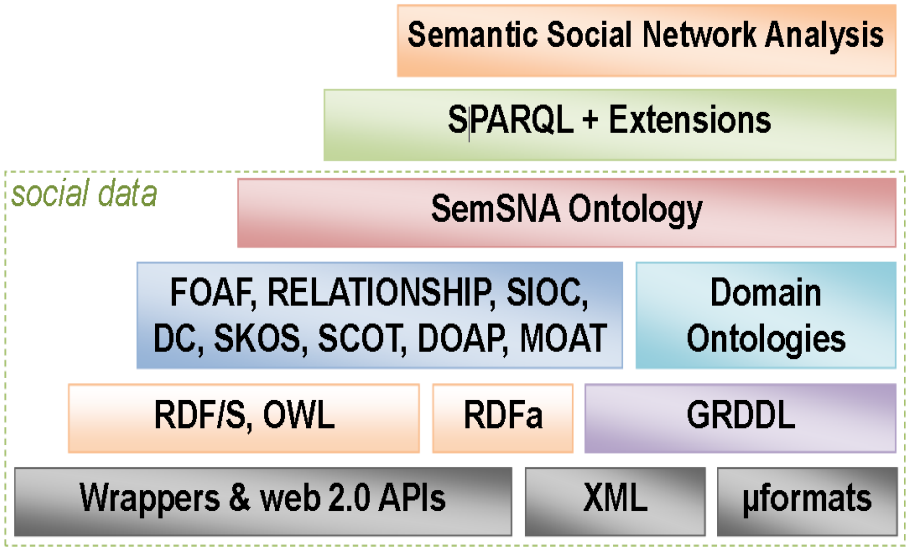
- **Social Networks** : networks based on the relation between people
- **Semantic Social Network** : RDF representations of social network and data



[Paolillo and Wright 2006]

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]

시맨틱 네트워크의 구성과 분석

$$Density = \sum_{k=1}^{n} v_k / g(g-1)$$

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

가족, 동료,
부서, 동아리

Density,
n-Clan, n-Clique

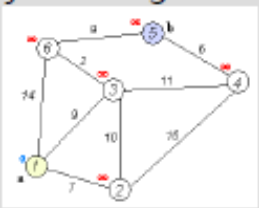
$$C'_B(i) = \frac{\sum_{j<k} g_{jk}(i) / g_{jk}}{[(g-1)(g-2)/2]}$$

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

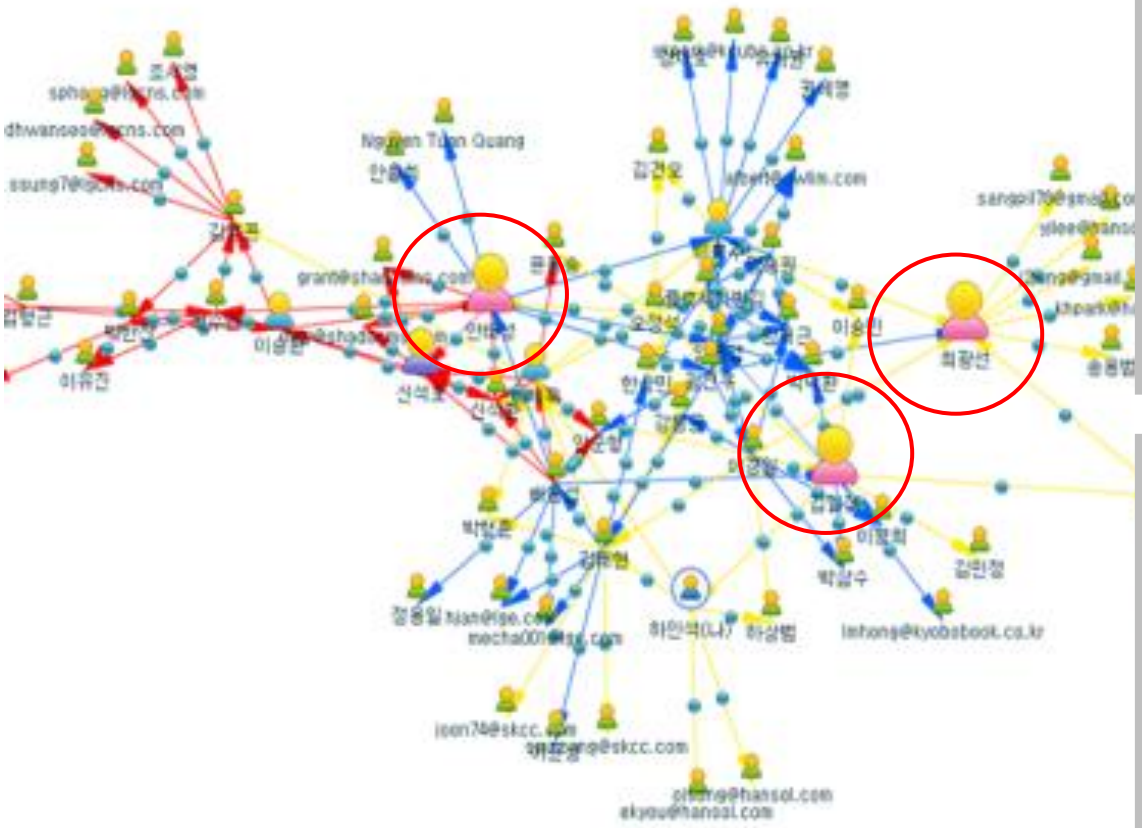


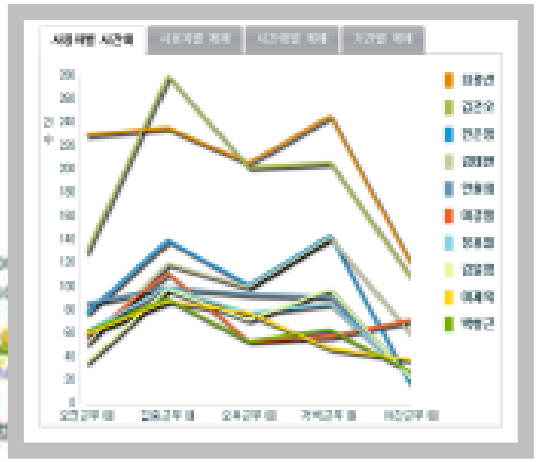
$O(|E| + |V| \log |V|)$

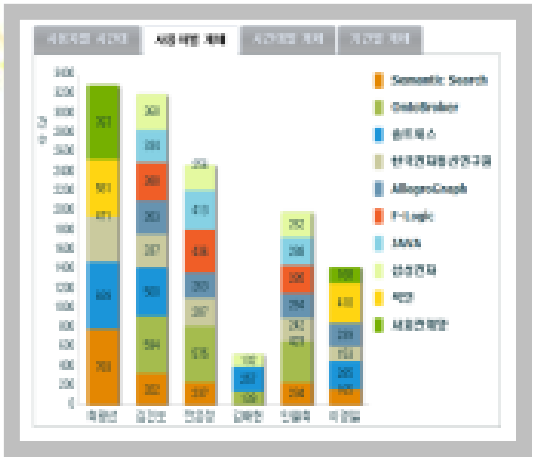
빨리 사람
찾기

Shortest Path
Algorithms

필요기술: Semantic



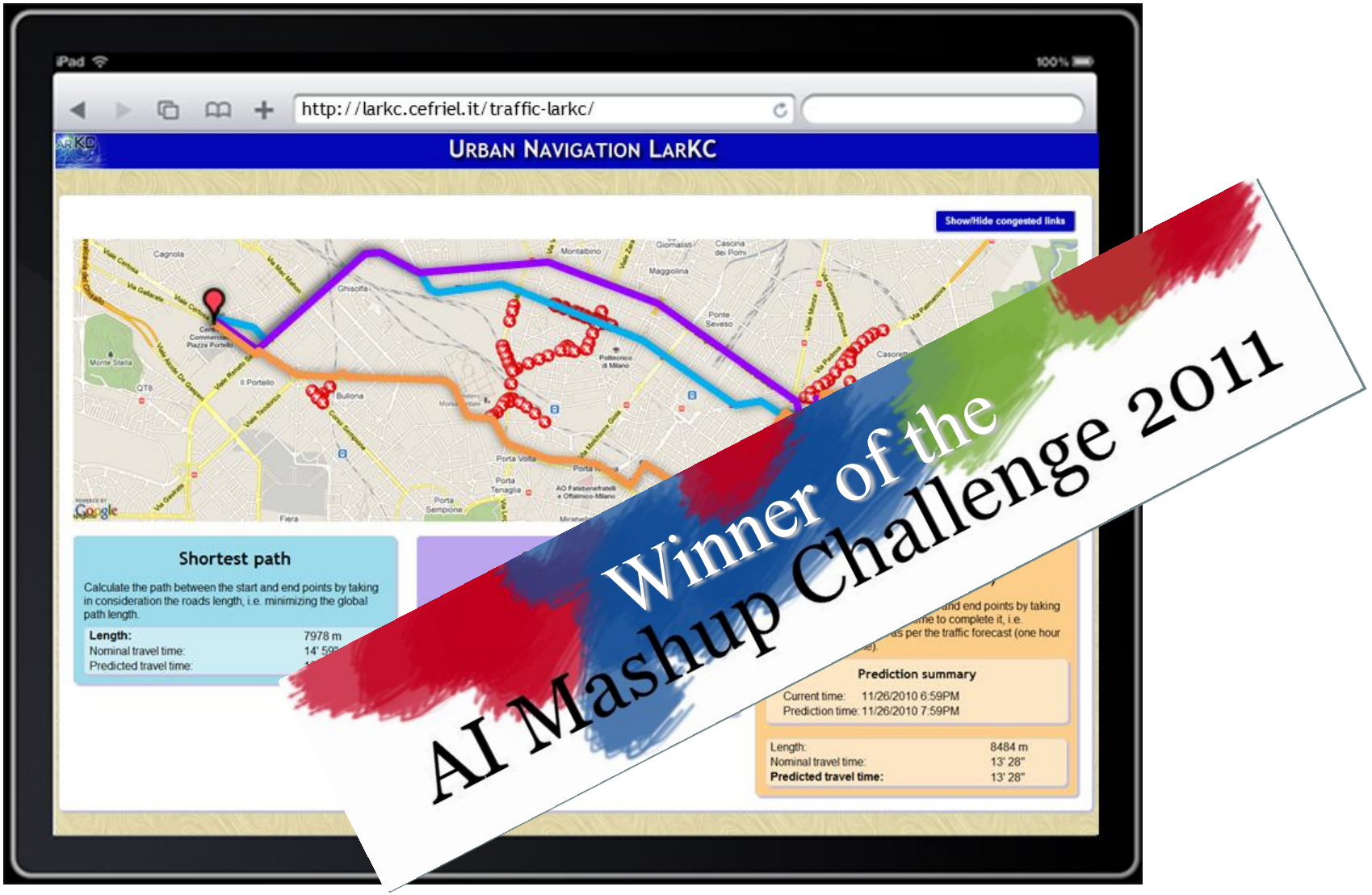




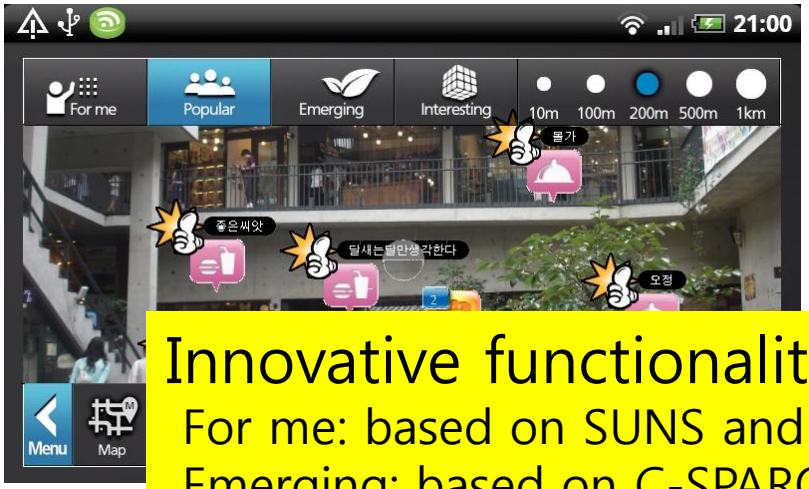
Saltlux

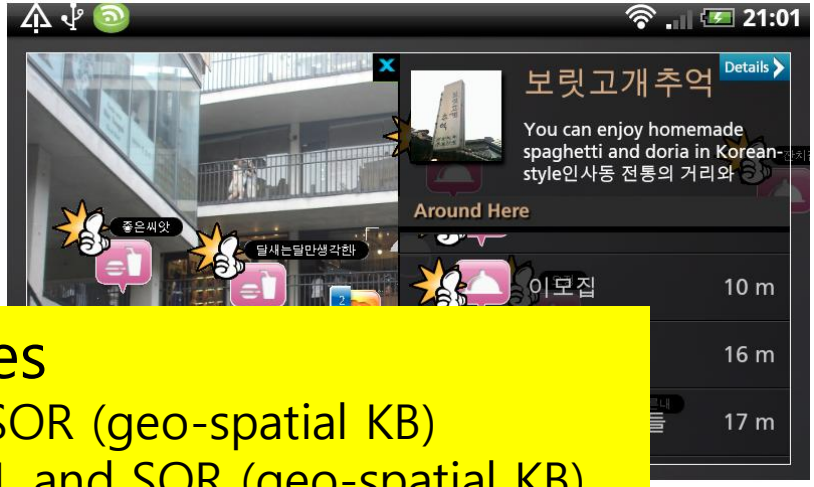
LarKC Urban Computing

LarKC Urban Computing



LarKC Urban Computing



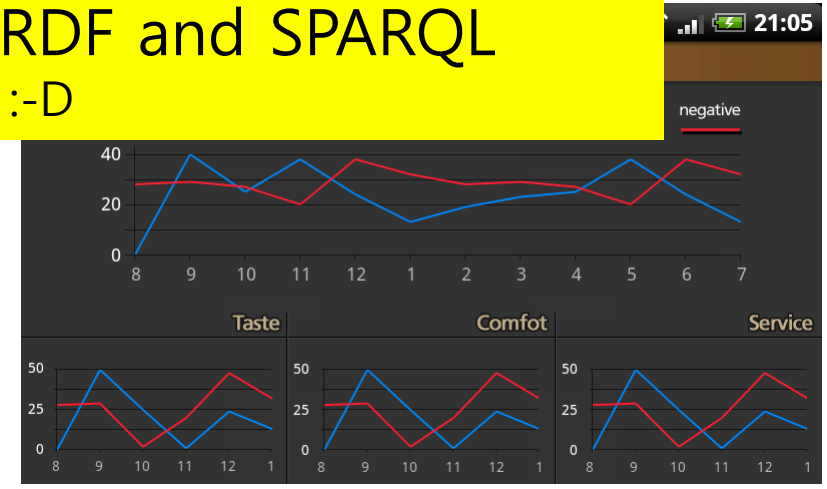


Innovative functionalities

- For me: based on SUNS and SOR (geo-spatial KB)
- Emerging: based on C-SPARQL and SOR (geo-spatial KB)

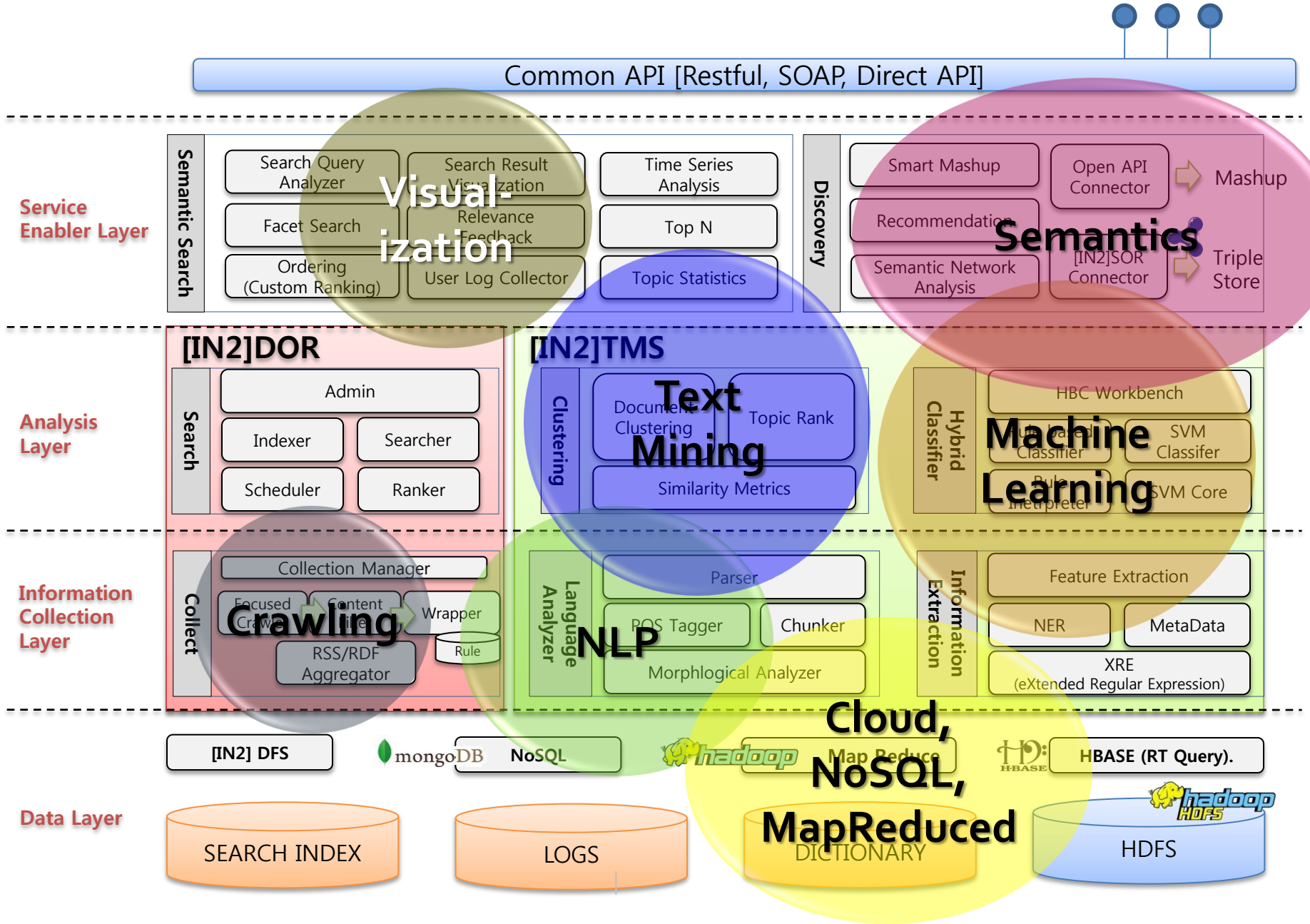
Fully implemented on RDF and SPARQL

- First Commercial Mobile APP :-D



Saltlux [IN2] Discovery Architecture 2.x

13 FEB. 2012



4. Saltlux 기업 소개

Saltlux

“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

기업 신뢰성

TTA, ISO, 지식경제부, 소프트웨어진흥원, 과학기술부, 행정안전부 등으로 부터 다양한 인증과 수상실적을 가진 신뢰할만한 기업입니다.

아시아 IT 200대 기업 선정

RED HERRING THE BUSINESS OF TECHNOLOGY

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)

GOOD Software

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

디지털이노베이션 대상	INNO-BIZ 선정	ISO9001:2000	신 소프트웨어 상품대상	2010년 IT 히트 상품	과기부 신기술 마크	행망 S/W 등록
Digital Innovation Awards 2007	INNOBIZ		KSWA 신소프트웨어 상품대상 KOREA SOFTWARE AWARDS	HIT	KT Excellent Korean Technology	
[IN2]	기술혁신형 중소기업 선정	품질관리	[IN2]DOR &TMS/Discovery	[IN2]Discovery, 품질 우수부문	자동번역 솔루션	[IN2]DOR/TMS

Saltlux

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**

자연어 처리

기계 학습

인공 지능

시맨틱 모호성 배제

텍스트 마이닝

시맨틱 어노테이션

추론

온톨로지 기반 시맨틱 기술

Search 2.0 (시맨틱 검색)

시맨틱 BPM

웹 3.0 (시맨틱 웹)

유비쿼터스/모바일(상황인지)

Saltlux

주요고객

㈜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

Saltlux

Thank you!

135-848 서울특별시 강남구 대치동 **967** 덕일빌딩 **5, 6, 7** 층
Tel : 02-3402-0081 Home : www.saltlux.com
Fax: 02-3402-0082 E-mail : saltluxinc@saltlux.com