빅데이터 분석이란

심탁길

terryshim@naver.com

목차

1. 빅데이터 개요
2. 빅데이터 활용 사례
3. 빅데이터 분석 데모 - 헬스케어

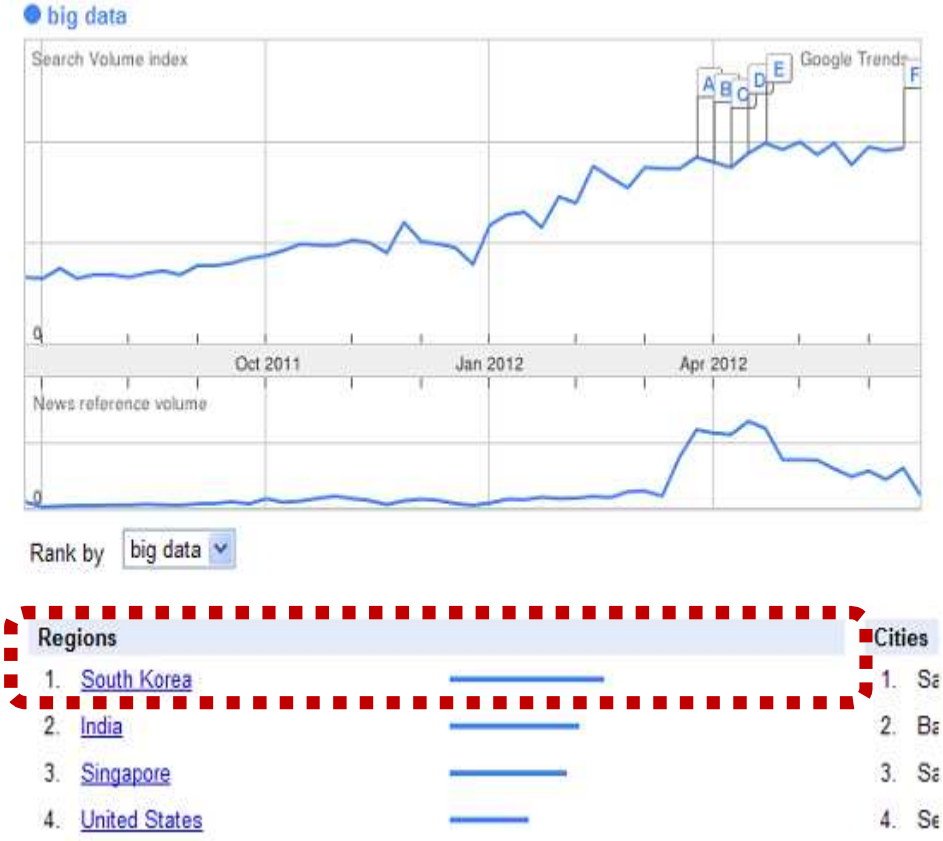
빅데이터 개요

빅데이터란?

Big Data

기존의 방식으로

저장/관리 분석하기 어려울 정도의 큰 규모의 자료

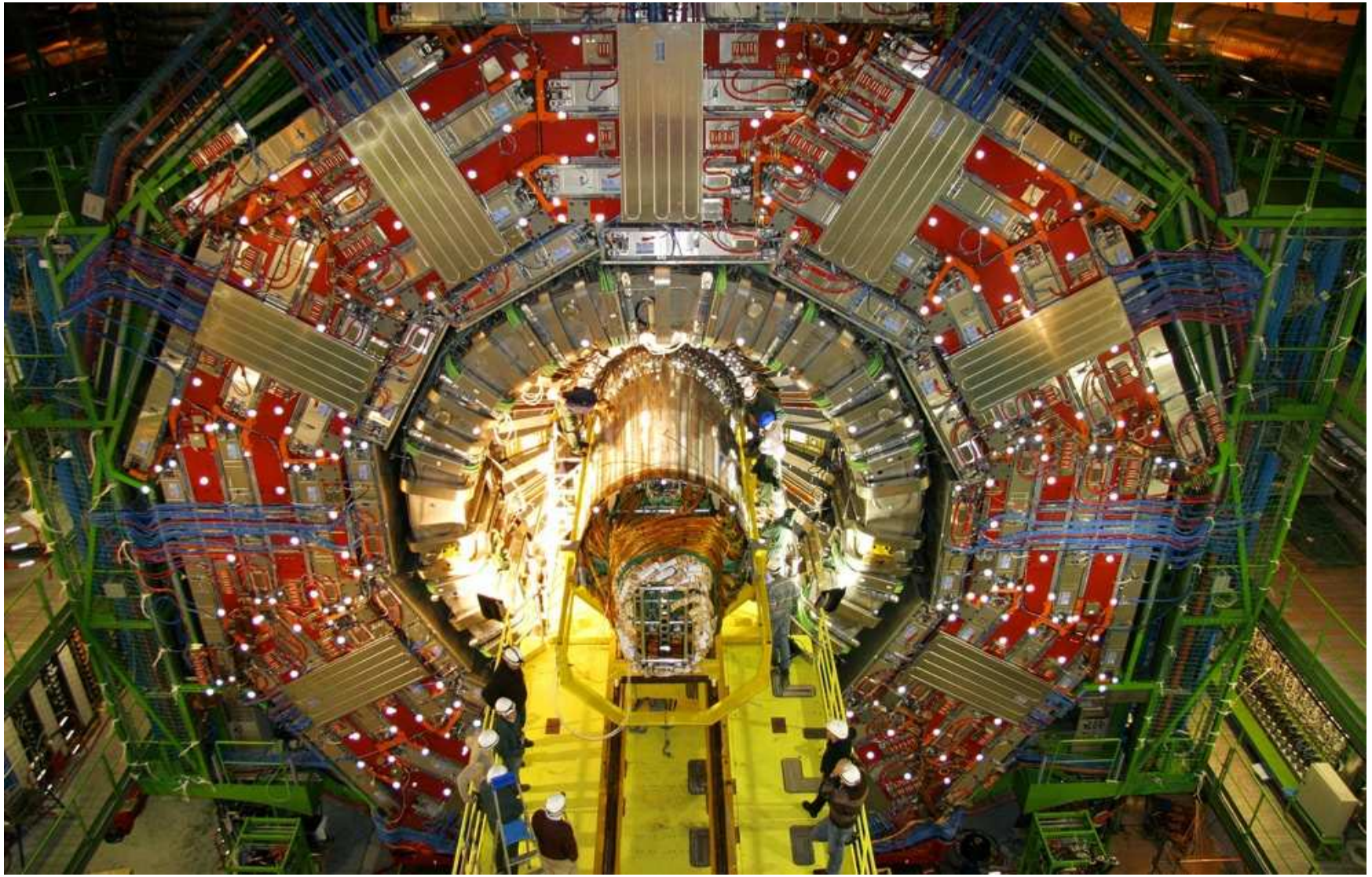


Variety: 데이터의 다양성

Volume: 데이터의 크기

Velocity: 데이터의 처리 속도

빅데이터 발생지 (1)



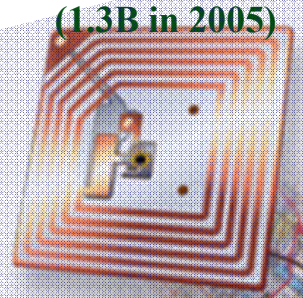
빅데이터 발생지 (2)

? TBs **of data every day**

12+ TBs **of tweet data every day**

25+ TBs **of log data every day**

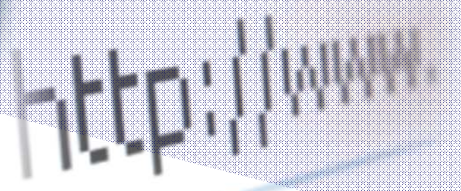
30 billion **RFID tags today (1.3B in 2005)**



4.6 billion **camera phones world wide**

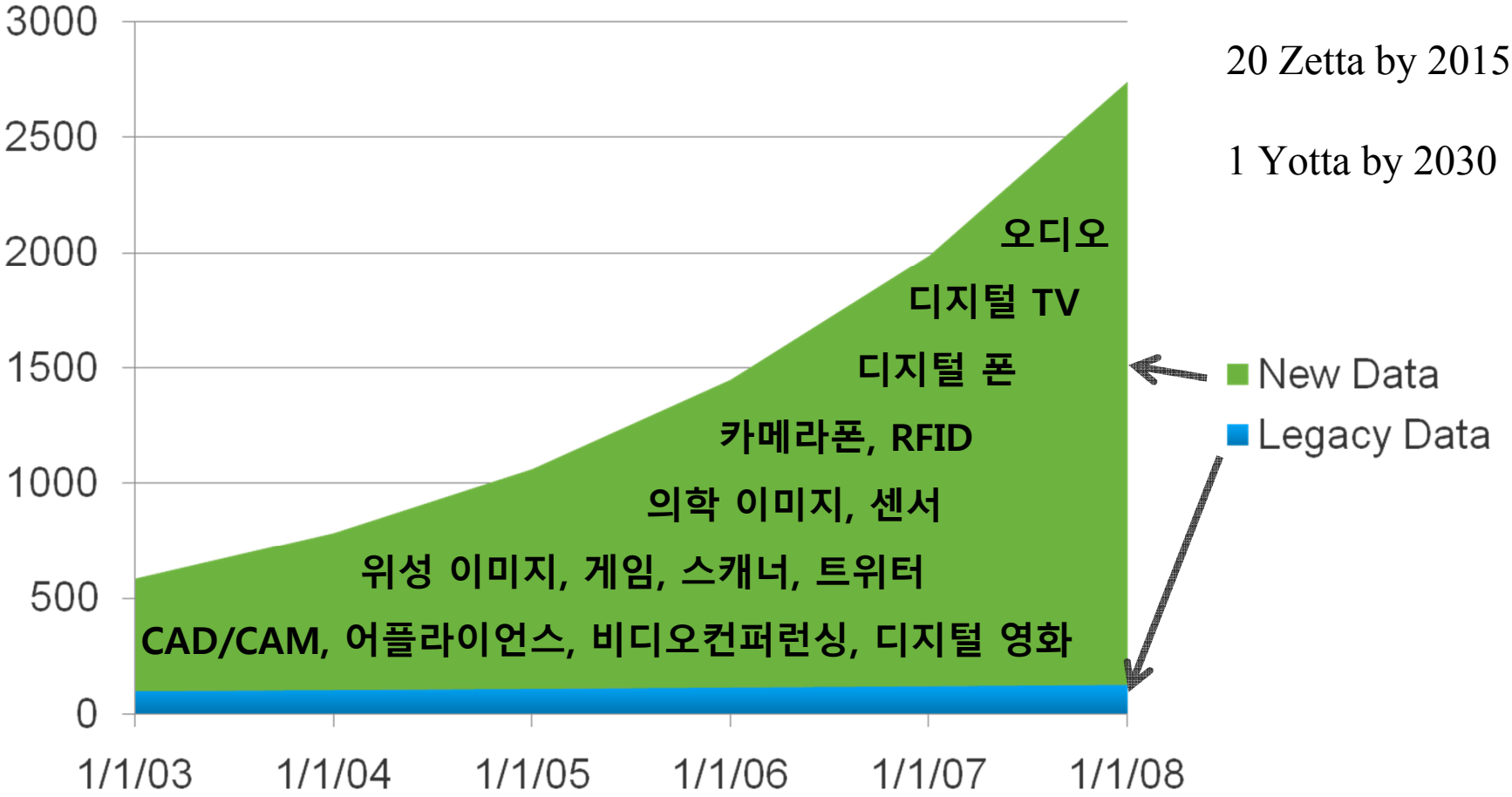
100s of millions of GPS enabled devices sold annually

76 million **smart meters in 2009… 200M by 2014**



2+ billion **people on the Web by end 2011**

데이터 증가



20 Zetta by 2015

1 Yotta by 2030

Source: The Information Explosion, 2009

빅데이터는 옛날(?)부터 있었다

How Big is Big Data?

Google

<2007년>
- 데이터 센터 당 약 40,000 ea 서버
- 총 서버 대수 1,000,000 ea
- 일 단위 평균 400PB 데이터 처리
- 한 작업 당 180GB 입력 데이터

<빅 데이터 기술 논문 공개>
- Google Filesystem(2004년)➔ 분산 파일시스템
- MapReduce(2005년)➔ Hadoop 프로젝트
- BigTable(2006년)➔ NoSQL 프로젝트
- Chubby, Sawzall 등➔ Hive, ZooKeeper 프로젝트

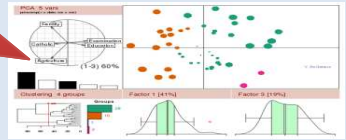
오픈소스 빅데이터 프로젝트의 동기

APACHE HBASE

하지만...

오픈소스 빅데이터 기술들

R
프로젝트



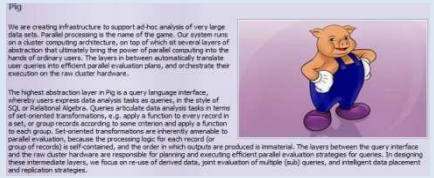
APACHE HBASE

사용자 분류, 군집 by behaviors
(구매 성향, 방문 페이지, 검색어 ...)

$$L_C^m(\mathbf{X}^m, \mathbf{Z}^m(\Phi^{m-1}), \Phi) = L_C^{m-1}(\mathbf{X}^{m-1}, \mathbf{Z}^{m-1}, \Phi) + \underbrace{\sum_q z_{mq}\left(\log \alpha_q + \sum_l \sum_{j \neq m} z_{jl} \log(\pi_{ql}^{x_{mj}}(1-\pi_{ql})^{1-x_{mj}})\right)}_{L_C(\mathbf{X}_m, \mathbf{Z}^m, \Phi)}$$

mahout

select * where group by cohort
(weeks, month, quarters, years, event period and geography, behaviors)



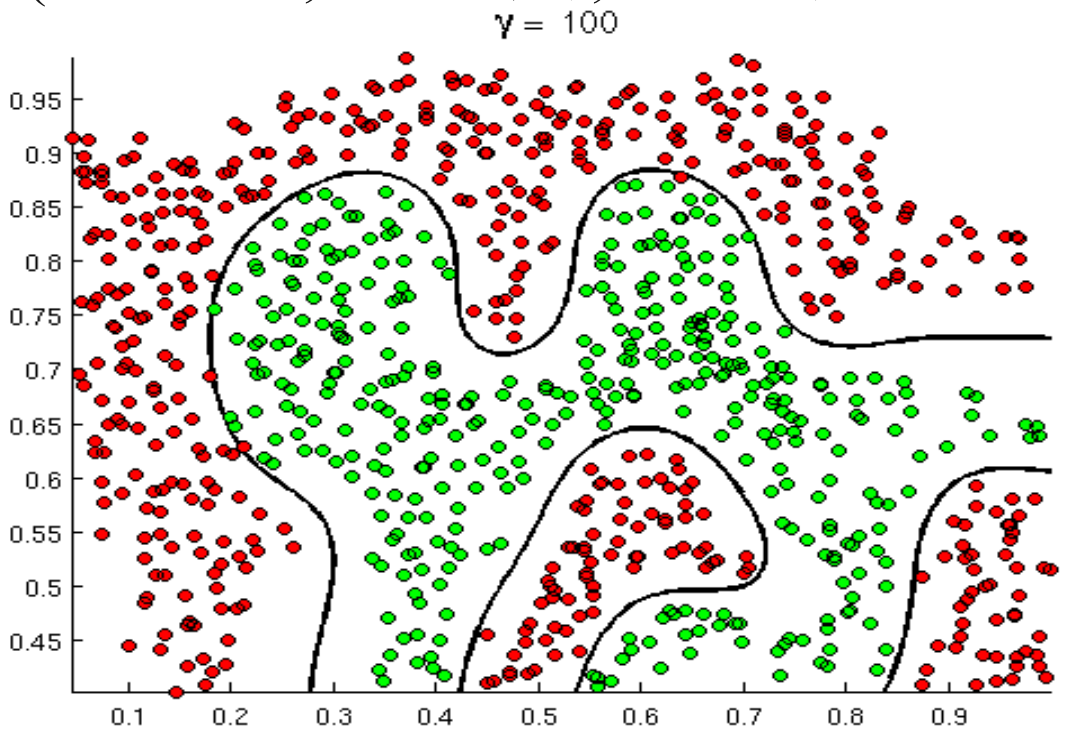
hadoop

빅데이터 분석의 핵심

예측 분석(Predictive Analytics)

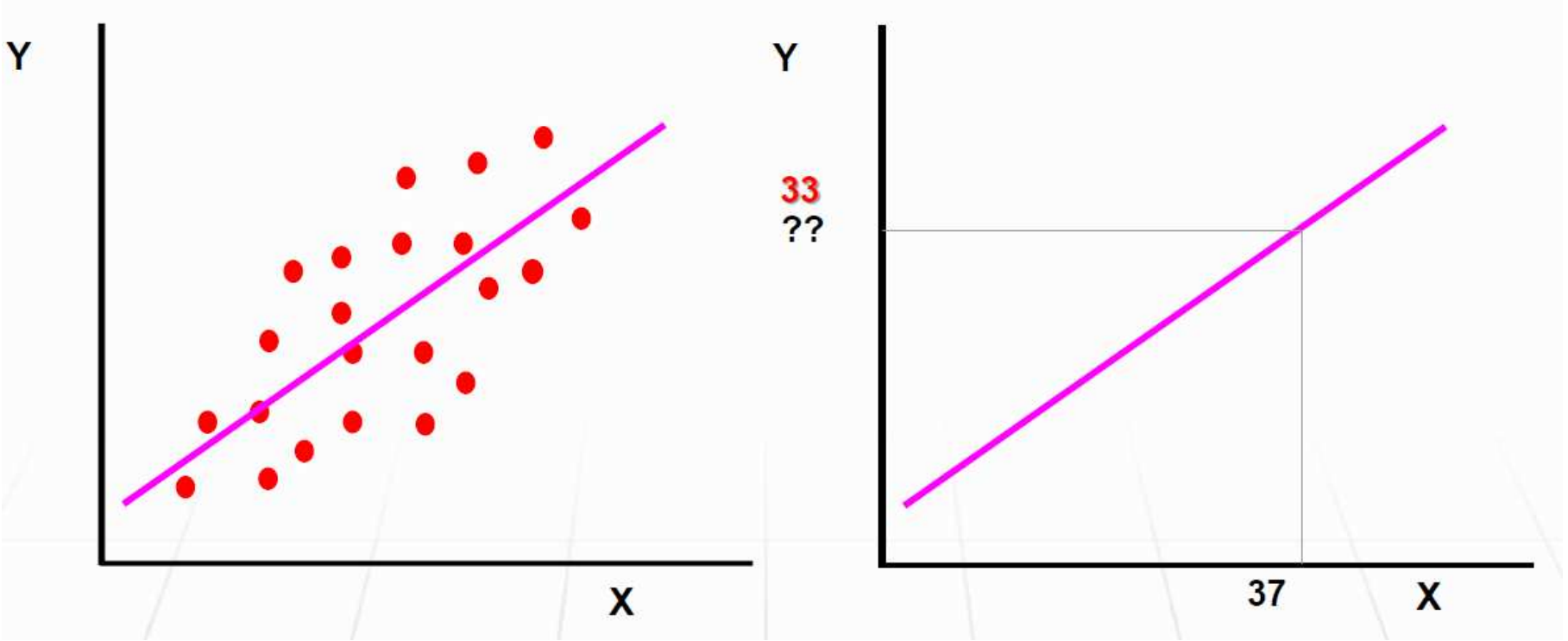
예측 분석 – 분류(Classification)

- 사전에 소속 그룹(Class)을 알고 있는 관측치들을 이용하여, 미래에 소속 그룹(Class)이 알려지지 않은 관측치가어떤 그룹에 분류될 것인가를 예측하는 분석 방법 (품종분류, 품질예측, 고객 이탈방지 예측)

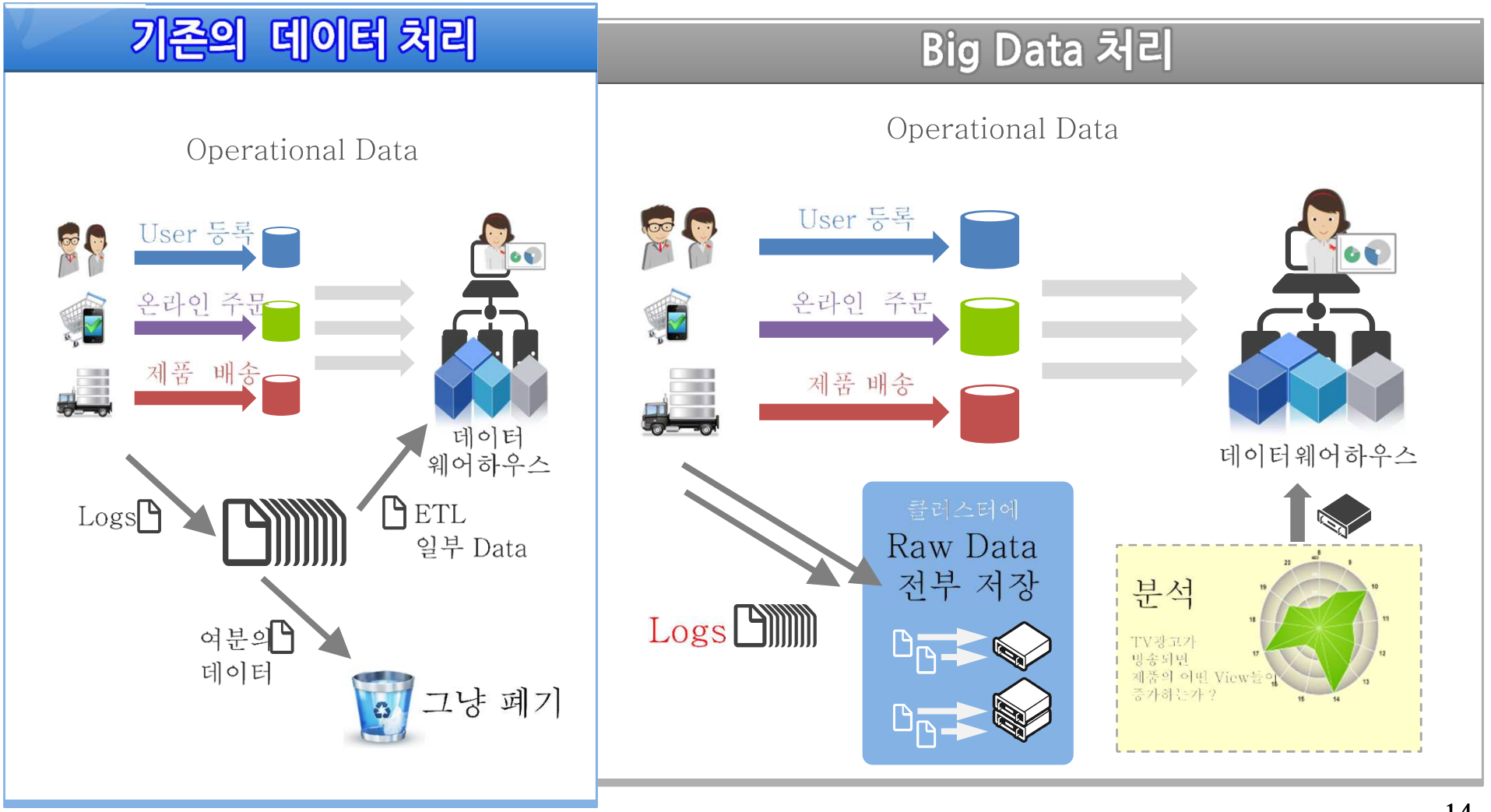


예측 분석 - 회귀(Regression)

- 기존 데이터의 종속변수와 독립변수를 이용하여 모델을 만들고,미래의 관측치의 독립변수 값이 주어졌을 경우 종속변수의 값을 예측

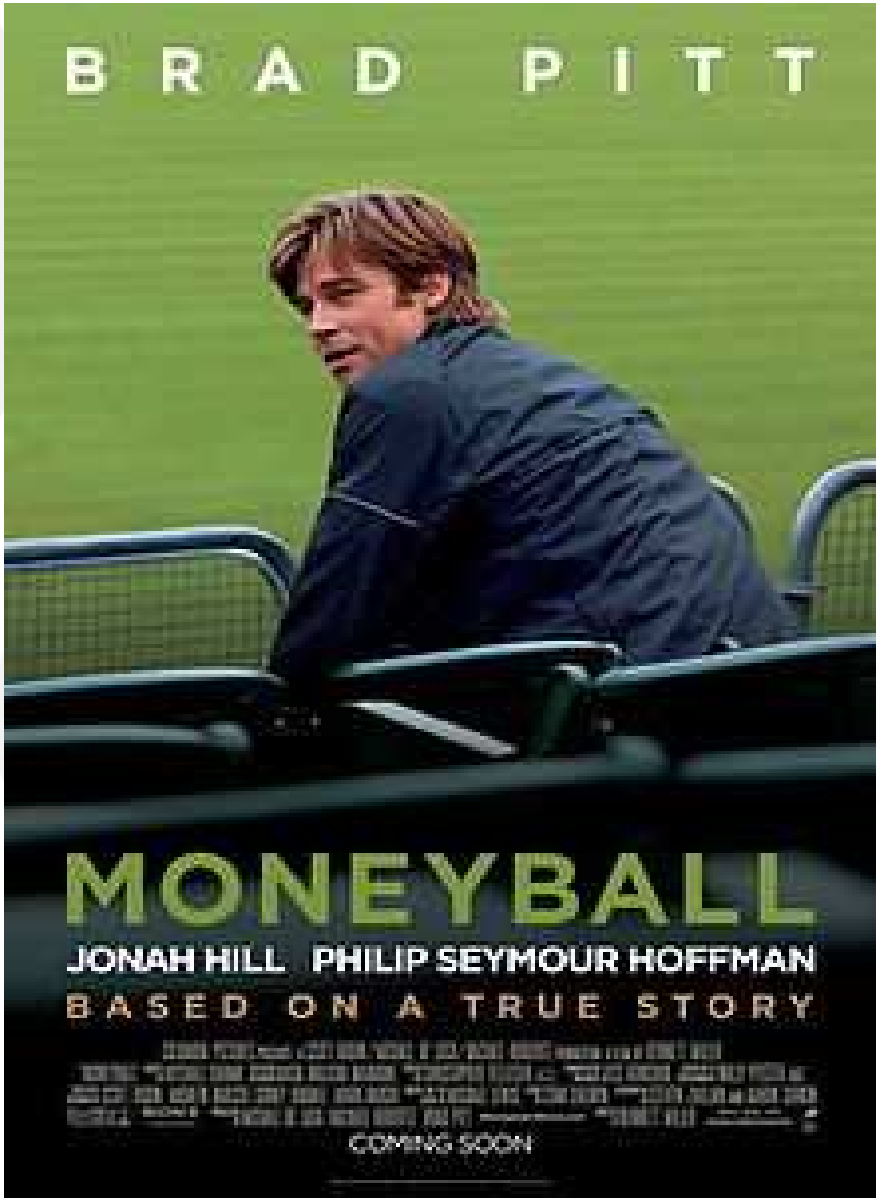


무엇이 달라졌나?


기존의 데이터 처리
Operational Data
User 등록
온라인 주문
제품 배송
데이터
웨어하우스
Logs
ETL
일부 Data
여분의
데이터
그냥 폐기
Big Data 처리
Operational Data
User 등록
온라인 주문
제품 배송
데이터웨어하우스
Logs
클러스터에
Raw Data
전부 저장
분석
TV광고가
방송되면
제품의 어떤 View들이
증가하는가?

빅데이터 활용 사례

비즈니스 활용 사례 – 메이저리그 분석

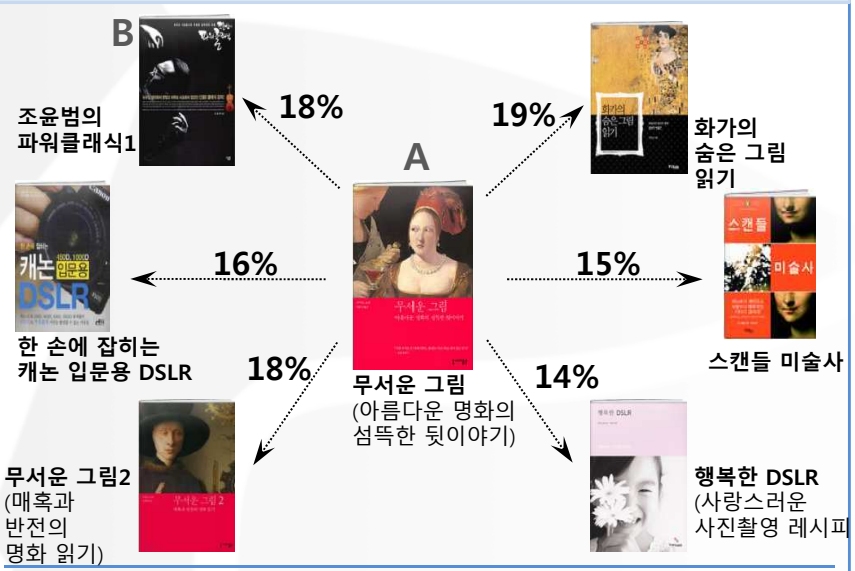


비즈니스 활용 사례 – 2012년 미국 대선

비즈니스 활용 사례 – 개인화/추천

구매성향이 비슷한 고객들을 그룹화하여 구매할 가능성이 높은 상품에 대해 점수를 매겨서 추천함

상품 Network 구성(도서 예시)



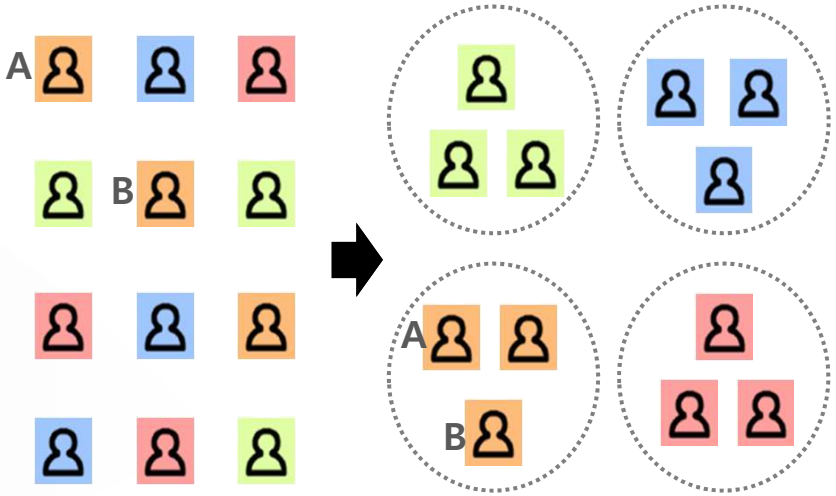
상품 간 관계 분석

- 특정 기간 동안 판매된 도서의 상관관계를 나타낸 상품 Network 구성

비고

- 책이 팔릴 때마다 상관관계가 변동되어 적절한 주기로 상품 Network를 업데이트
- 가중치 적용(구매시기/판매수량/구매자수)

고객군 Clustering



고객 군 분석

- 총 상품 구매자들을 대상으로 취향이 비슷한 사람들을 묶어 클러스터 100개 구성

비고

- 정확도를 높일 수 있는 적절한 클러스터 개수
- 고객군을 나누는 적절한 경계 설정

추천을 위한 준비 : 사용자 선호도 테이블 생성

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
u1	1		0	1		1		0	1	1		1
u2	1	1		0	1			1	0		0	
u3	1		1		1		1	1	0	0		0
u4	1		1		1	1	1		0	0	0	1
u5	0			0		1		0	1		1	
u6	1	1		0	0			1	1			0
u7	0				0		1	0	1	1	1	
C1	1	1	1	0	1	1	1	1	0	0	0	0,5
C2	0,3	1		0	0	1	1	0,3	1	1	1	0

추천을 위한 준비 : 사용자-사용자(또는 상품-상품)간의 유사도 메트릭스 생성

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
u1	1		0	1		1		0	1	1		1
u2	1	1		0	1			1	0		0	
u3	1		1		1		1	1	0	0		0
u4	1		1		1	1	1		0	0	0	1
u5	0			0		1		0	1		1	
u6	1	1		0	0			1	1			0
u7	0				0		1	0	1	1	1	
C1	1	1	1	0	1	1	1	1	0	0	0	0,5
C2	0,3	1		0	0	1	1	0,3	1	1	1	0

$$simil(x,y) = \frac{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})(r_{y,i} - \bar{r_y})}{\sqrt{\sum_{i \in I_{xy}} (r_{x,i} - \bar{r_x})^2 \sum_{i \in I_{xy}} (r_{y,i} - \bar{r_y})^2}}$$

	u1	u2	u3	u4	u5	u6	u7	c1	c2
u1	1,00	-0,58	-0,71	-0,32	0,41	-0,41	0,58	-0,54	0,18
u2		1	1	1	-0,67	0,25	-1	1	-0,29
u3			1	0,73	-1	0,167	-0,71	0,942	-0,29
u4				1	-0,58	-0,58	-0,71	0,943	-0,59
u5					1	0,333	1	-0,33	0,962
u6						1	0,333	0,281	0,795
u7							1	-0,75	0,966
c1								1	-0,12
c2									1

실시간 또는 배치 추천

사용자 선호도 테이블

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
u1	1		0	1		1		0	1	1		1
u2	1	1		0	1			1	0		0	
u3	1		1		1		1	1	0	0		0
u4	1		1		1	1	1		0	0	0	1
u5	0			0		1		0	1		1	
u6	1	1		0	0			1	1			0
u7	0				0		1	0	1	1	1	
C1	1	1	1	0	1	1	1	1	0	0	0	0,5
C2	0,3	1		0	0	1	1	0,3	1	1	1	0

사용자 유사도 테이블

	u1	u2	u3	u4	u5	u6	u7	c1	c2
u1	1,00	-0,58	-0,71	-0,32	0,41	-0,41	0,58	-0,54	0,18
u2		1	1	1	-0,67	0,25	-1	1	-0,29
u3			1	0,73	-1	0,167	-0,71	0,942	-0,29
u4				1	-0,58	-0,58	-0,71	0,943	-0,59
u5					1	0,333	1	-0,33	0,962
u6						1	0,333	0,281	0,795
u7							1	-0,75	0,966
c1								1	-0,12
c2									1

상품 추천 테이블

한 명(U1)에게 상품 추천을 위해 필요한 정보들

(1) U1과 유사한 성향을 가진 사용자(S)들 목록 추출
➔ S1, S2, S3 3명이라고 가정

(2) U1이 구매하지 않은 상품들 추출
➔ 미구매 상품을 p2, p5, p7 라고 가정

(3) 추천 점수 계산
➔ p2의 추천 점수 : (S1의 유사도값 X p2 + S2의 유사도 값 X p2 + S3의 유사도값 X p3) / (S1의 유사도값 + S2의 유사도값 + S3의 유사도값)

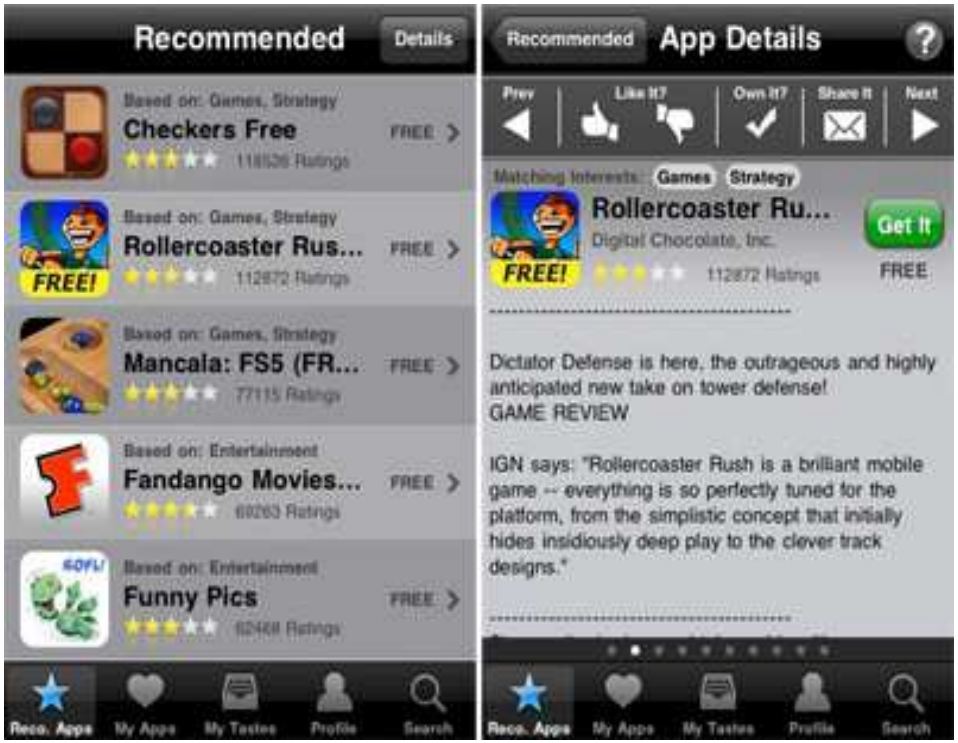
➔ p5의 추천 점수 : …

➔ p7의 추천 점수 : …

(4) p2, p5, p7 중 가장 추천 점수가 높은 상품을 최종적으로 사용자에게 추천함

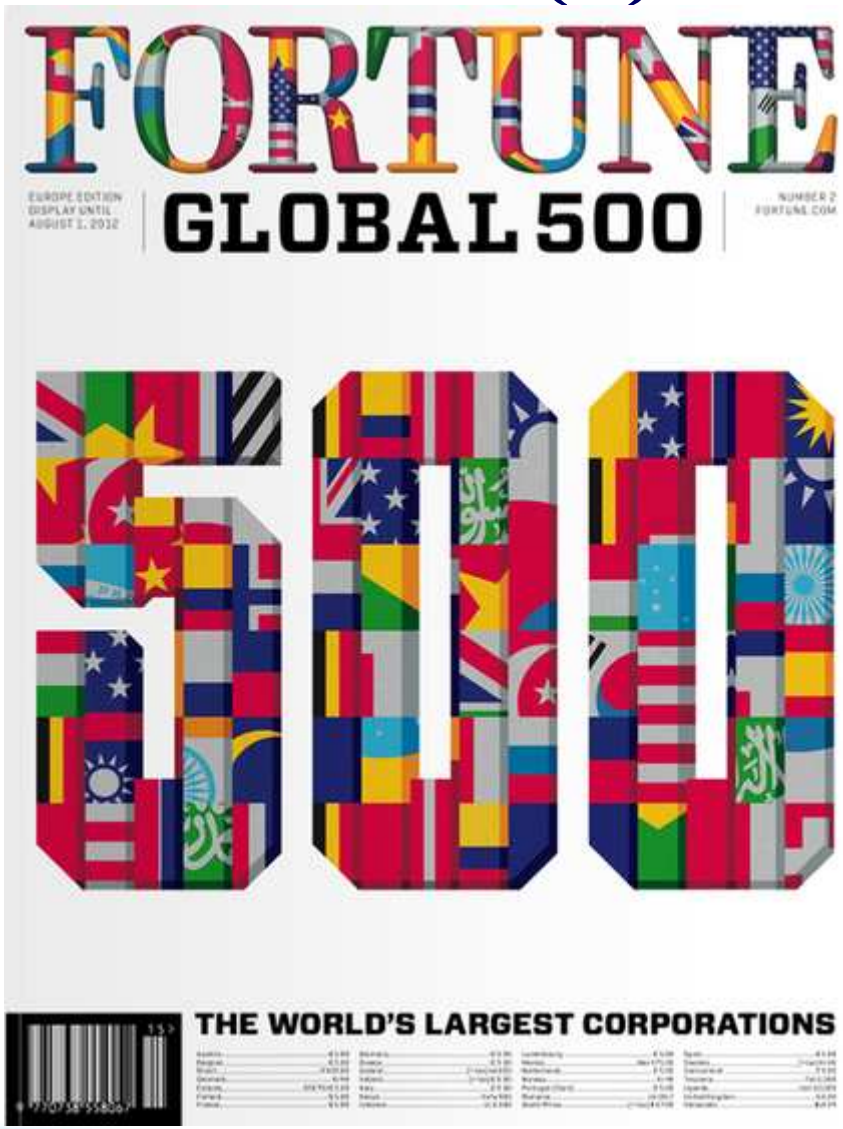
비즈니스 활용 사례 – 개인화/추천

분석 결과 평가 및 반영



<Netflix 인용> **<Intomobile.com 인용>**

이제는 매우(?) 많음



빅데이터 분석 데모 - 헬스케어

Q&A