

# 빅데이터 분석이란

© TemplatesWise.com

심탁길

terryshim@naver.com

# 목차

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



# 빅데이터 개요



# 빅 데이터란?

# Big Data

기존의 방식으로

저장/관리 분석하기 어려울 정도의 큰 규모의 자료



**Variety:** 데이터의 다양성



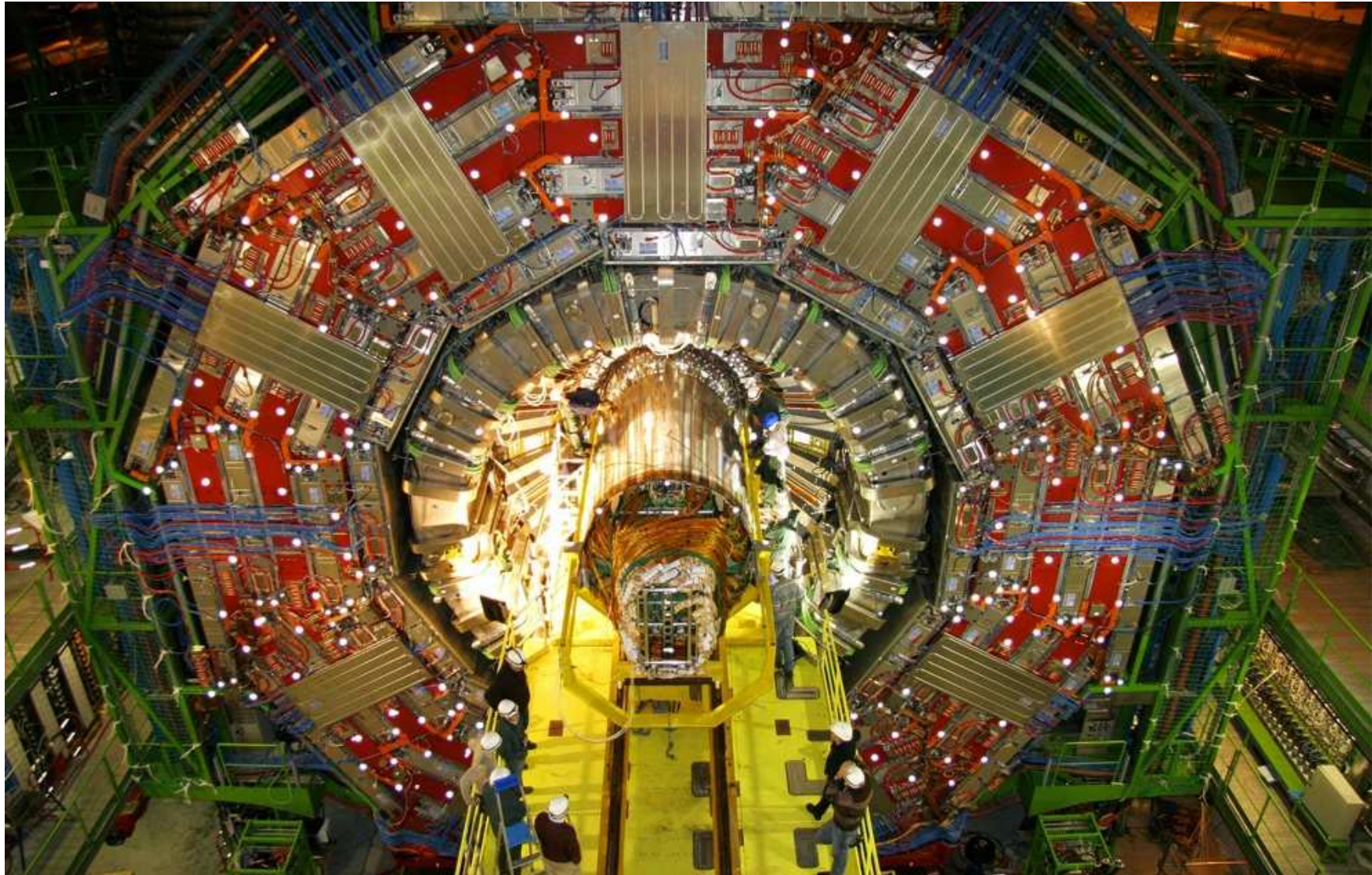
**Volume:** 데이터의 크기



**Velocity:** 데이터의 처리 속도



# 빅데이터 발생지 (1)



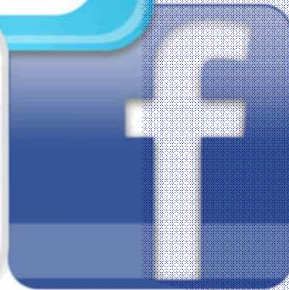


# 빅데이터 발생지 (2)

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

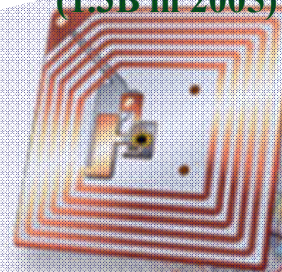


**? TBs** of  
data every day



**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 million  
s of  
GPS  
enable**

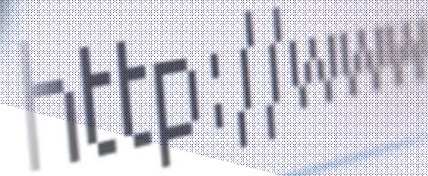


**2+ billion**  
d devices  
sold  
annually

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



**n** people  
on the  
Web by  
end 2011



# 데이터 증가



Source: The Information Explosion, 2009

# 빅 데이터는 옛날(?)부터 있었다

## *How Big is Big Data?*



### <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



**hadoop**

APACHE  
**HBASE**





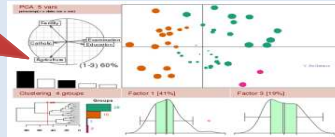
하지만...



# 오픈소스 빅데이터 기술들



**R  
프로젝트**



**A P A C H E  
HBASE**

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

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

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



[Fig]

We are creating infrastructure to support ad-hoc analysis of very large data sets. Parallel processing is the name of the game. Our system runs on a cluster computing architecture, on top of which sit several layers of abstraction that ultimately bring the power of parallel computing into the hands of ordinary users. The layers in between automatically translate user queries into efficient parallel evaluation plans, and orchestrate their execution on the raw cluster hardware.



The highest abstraction layer in Pig is a query language interface, whereby users express data analysis tasks as queries, in the style of SQL or Relational Algebra. Queries articulate data analysis tasks in terms of ad-hoc transformations, e.g., apply a function to every record in a set, or group records according to some criterion and apply a function to each group. Set-oriented transformations are inherently amenable to parallel evaluation, because the processing logic for each record (or group of records) is self-contained, and the order in which outputs are produced is immaterial. The layers between the query interface and the raw cluster hardware are responsible for planning and executing efficient parallel evaluation strategies for queries. In designing these intermediate layers, we focus on re-use of derived data, joint evaluation of multiple (bulk) queries, and intelligent data placement and replication strategies.





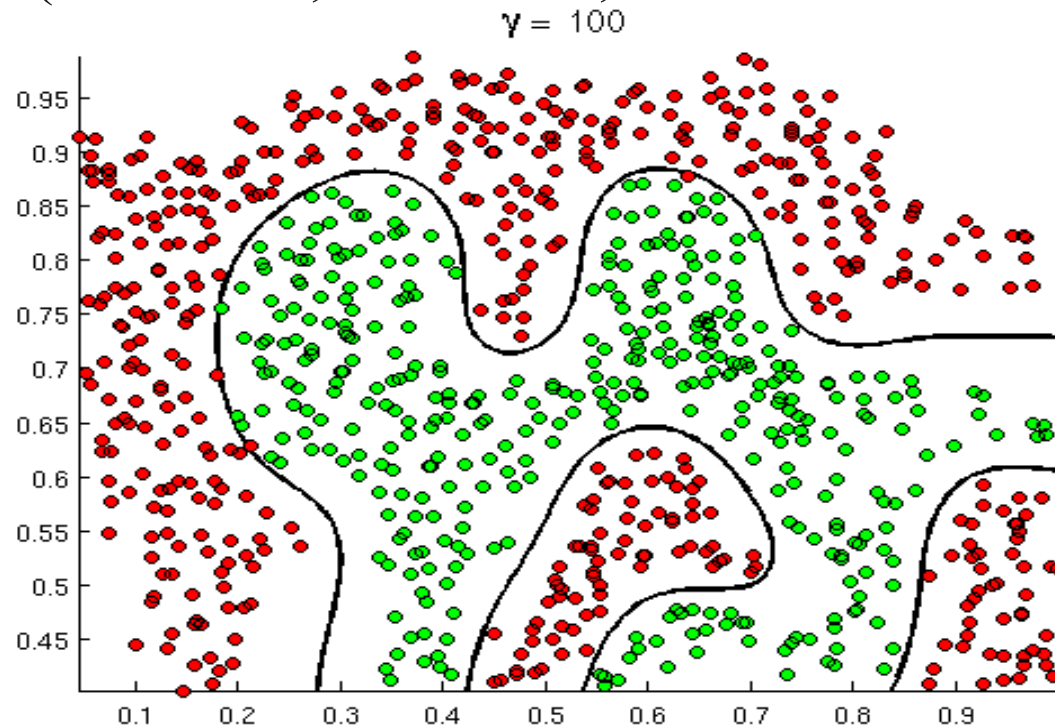
# 빅데이터 분석의 핵심

## **예측 분석(Predictive Analytics)**



# 예측 분석 – 분류(Classification)

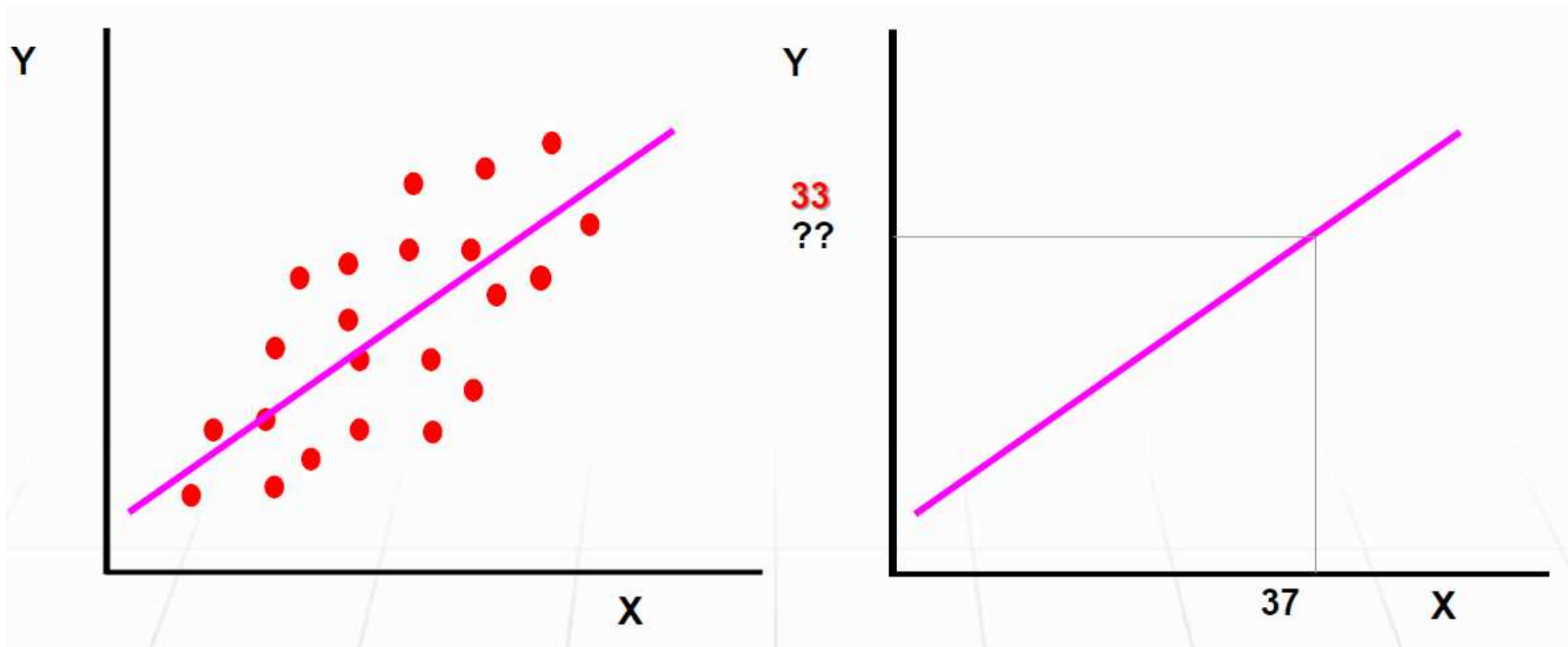
- 사전에 소속 그룹(Class)을 알고 있는 관측치들을 이용하여, 미래에 소속 그룹(Class)이 알려지지 않은 관측치가 어떤 그룹에 분류될 것인가를 예측하는 분석 방법 (품종분류, 품질예측, 고객 이탈방지 예측)





# 예측 분석 - 회귀(Regression)

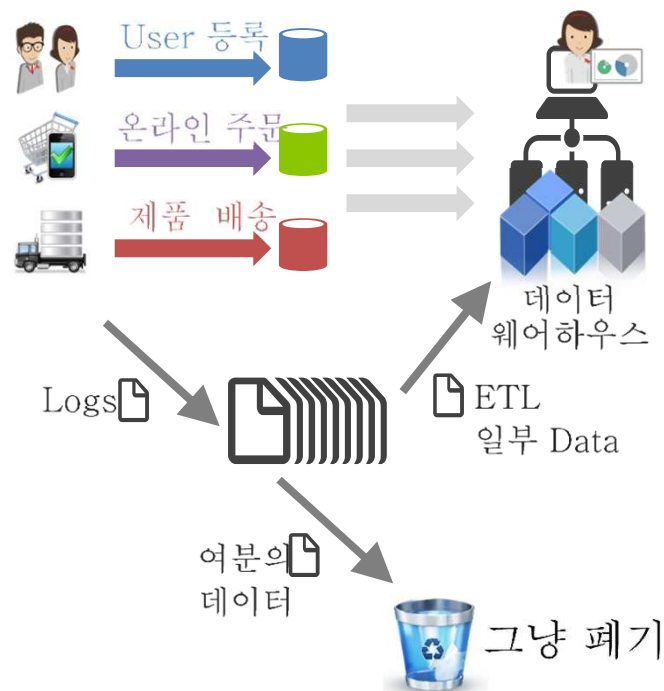
- 기존 데이터의 종속변수와 독립변수를 이용하여 모델을 만들고, 미래의 관측치의 독립변수 값이 주어졌을 경우 종속변수의 값을 예측



# 무엇이 달라졌나?

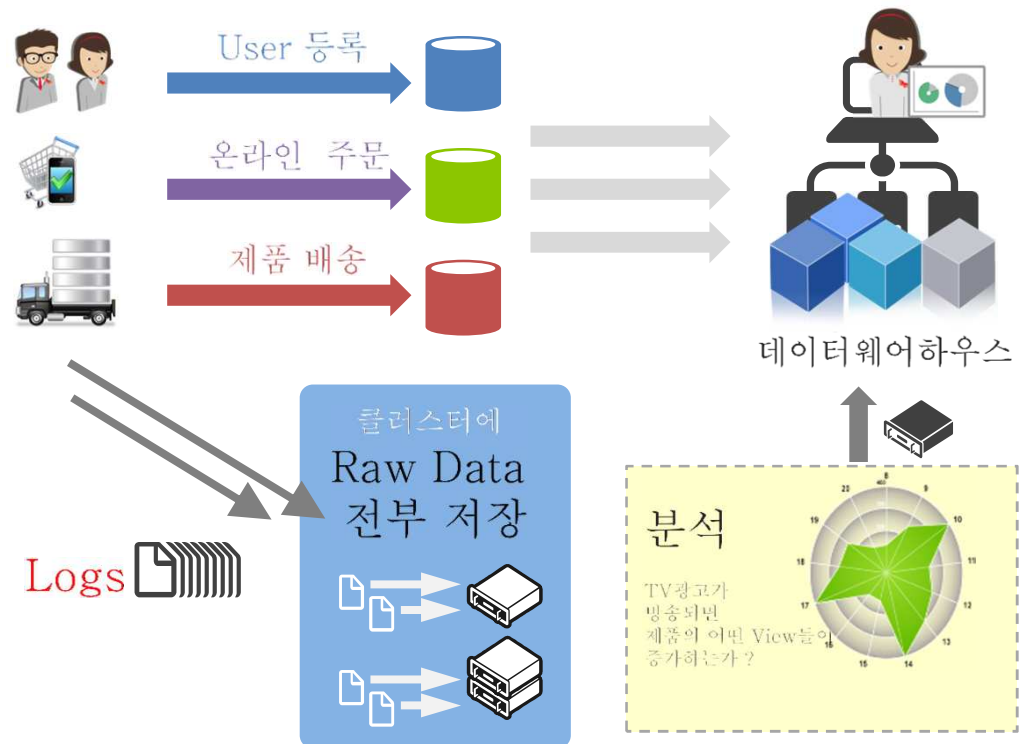
## 기존의 데이터 처리

Operational Data



## Big Data 처리

Operational Data

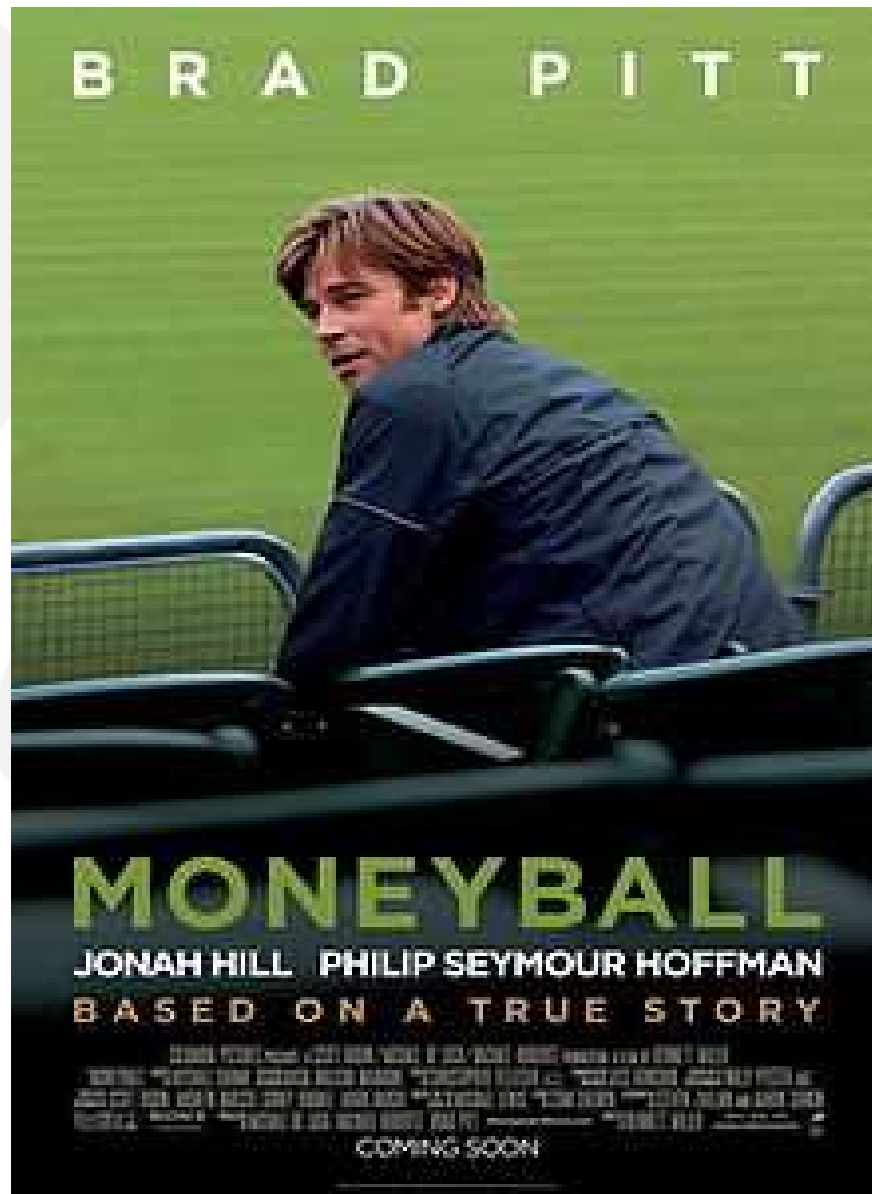




# 빅 데이터 활용 사례



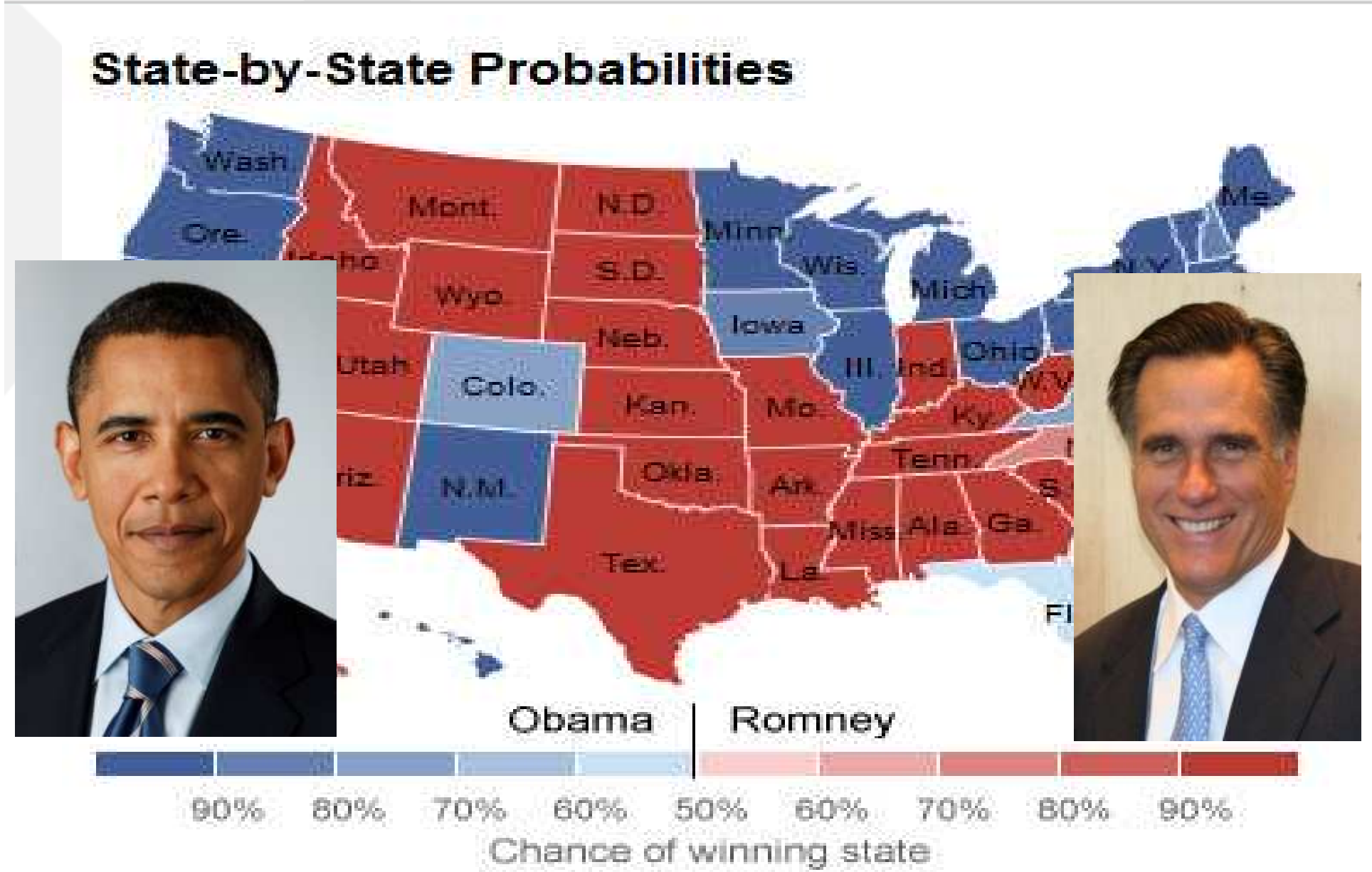
## 비즈니스 활용 사례 – 메이저리그 분석





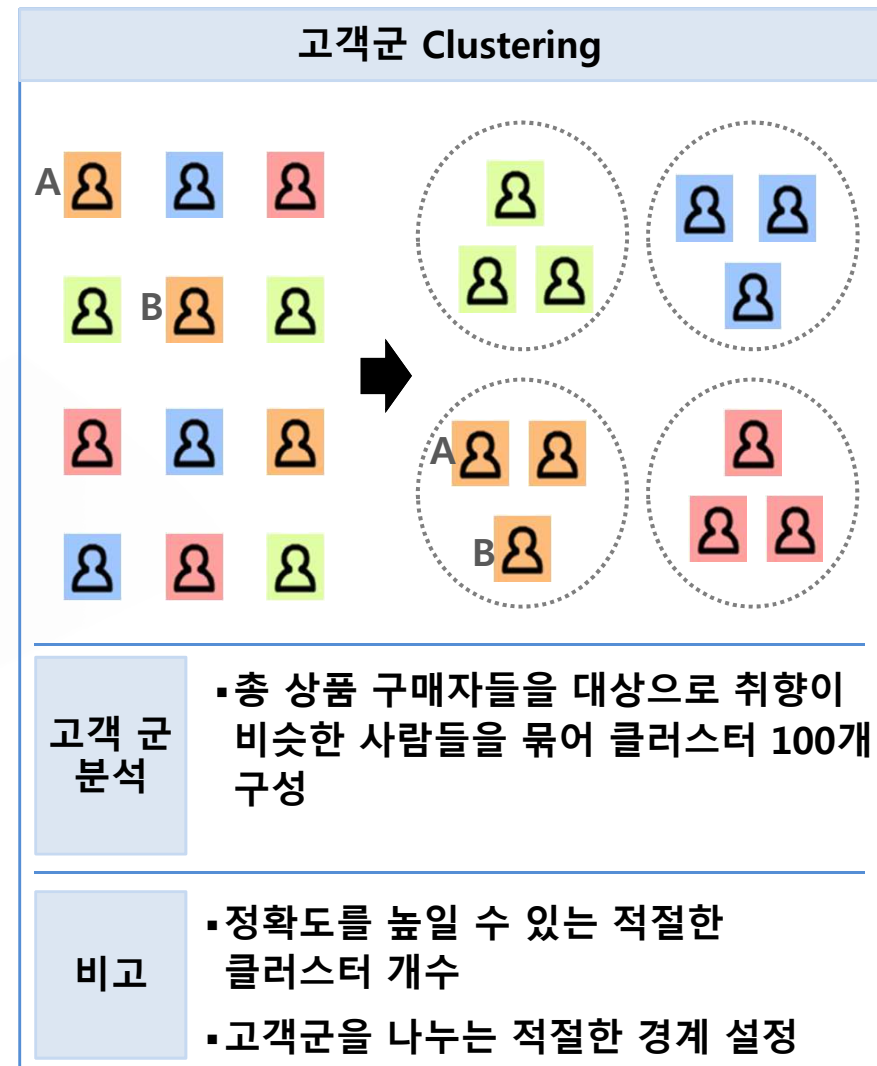
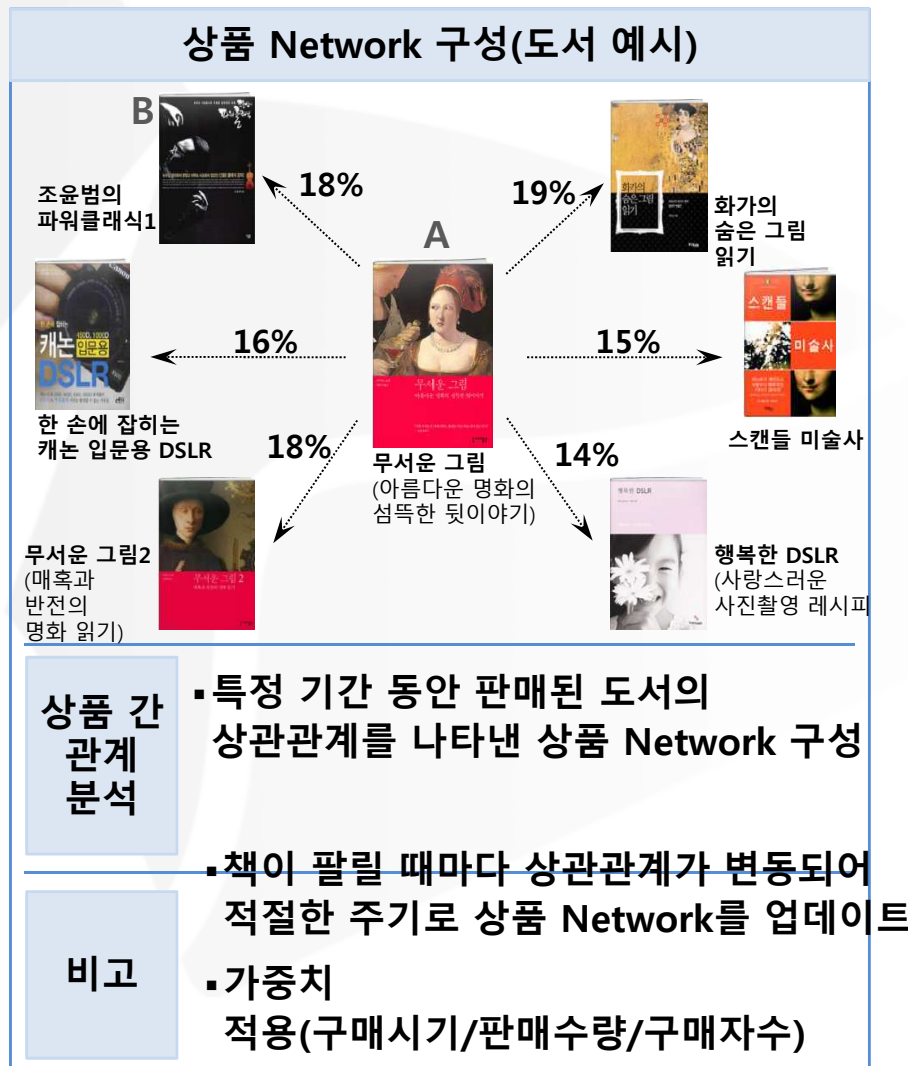
## 비즈니스 활용 사례 – 2012년 미국 대선

### State-by-State Probabilities



## 비즈니스 활용 사례 – 개인화/추천

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



# 비즈니스 활용 사례 - 개인화/추천

## 추천을 위한 준비 : 사용자 선호도 테이블 생성

```
2010-02-10 00:01:07 W3SVC1446 WEB100 216.167.204.29 GET /tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ - 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://www.google.com/search?hl=en&client=safari&rs=eq&itunes%3A+your-current-security-settings-do-not-allow-this-program-to-be-downloaded&asf=0&oeq= blog.caneja.com 200 0 0 8530 621 982
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/contact-form-7/stylesheets.css ver=2.0.7 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 811 479 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-includes/js/wp-ajax-response.js ver=2.9.1 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 1537 440 124
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/js/wp-ajax-edit-comments.js ver=2.3 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 5941 478 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/lightbox-plus/css/elegant/colorbox.css - 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 1365 474 93
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/css/themes/circular/edit-comments.css - 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 1414 495 109
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/css/colorbox/colorbox.css - 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 4495 483 109
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/lightbox-plus/js/jquery.colorbox-min.js ver=1.3.1 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 4495 470 296
2010-02-10 00:01:08 W3SVC1446 WEB100 216.167.204.29 GET /wp-content/plugins/wp-ajax-edit-comments/js/jquery.colorbox-min.js ver=2.9.1 80 - 12.178.189.252 HTTP/1.1 Mozilla/5.0+(Macintosh;+u;+Intel+Mac+OS+X+10_6_2;+en-us)+AppleWebKit/531.9+(KHTML,+likeGecko)+version/4.0.3+Safari/531.9 - http://blog.caneja.com/tips-tricks/fix-your-current-security-settings-do-not-allow-this-file-to-be-downloaded-error-in-ie/ blog.caneja.com 200 0 0 4495 470 296
```

	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

## 비즈니스 활용 사례 – 개인화/추천

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



# 비즈니스 활용 사례 – 개인화/추천

## 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 : ...

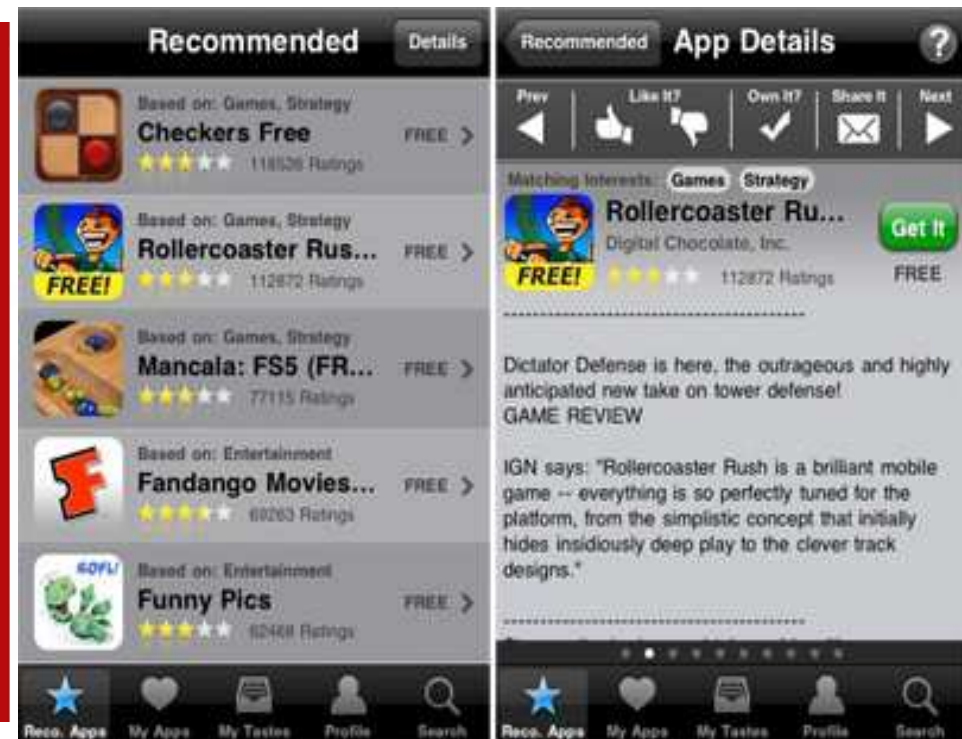
(4) p2, p5, p7 중 가장 추천 점수가 높은 상품을 최종적으로 사용자에게 추천함

# 비즈니스 활용 사례 – 개인화/추천

## 분석 결과 평가 및 반영



<Netflix 인용>



<Intomobile.com 인용>

# 이제는 매우(?) 많음





# 빅데이터 분석 데모 - 헬스케어







Q&A

