Big Data

조성준
서울대학교 산업공학과

- **정의**
- **Volume** 대량 Lots of Data
- **Velocity** 순식간에 발생 Stream Data
- **Variety** 문서,이미지 비정형 데이터

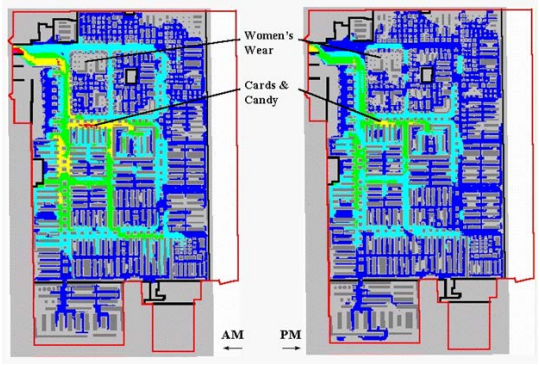
- 어디서?
  - 기계 Internet of Things (IoT)
  - 사람 Bring your own Data (BYOD)

- **어떤 형태?**
  - **숫자** numbers
  - **문자** text
  - **이미지, 동영상** image

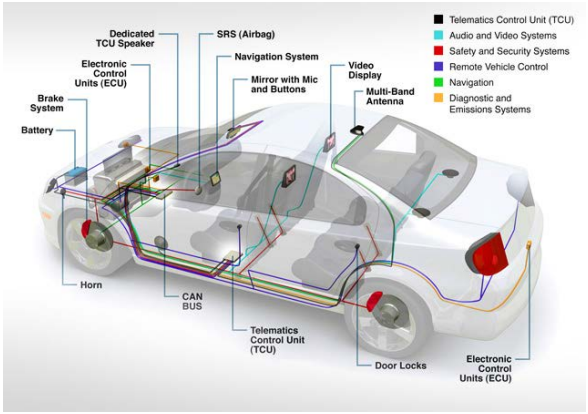
Numbers

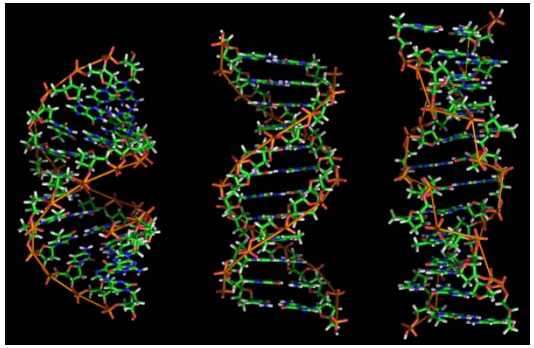
Walmart
Save money. Live better.

amazon.com



Numbers


Telematics Control Unit (TCU)
Audio and Video Systems
Safety and Security Systems
Remote Vehicle Control
Navigation
Diagnostic and Emissions Systems
Dedicated TCU Speaker
SRS (Airbag)
Navigation System
Video Display
Electronic Control Units (ECU)
Mirror with Mic and Buttons
Multi-Band Antenna
Brake System
Battery
Horn
CAN BUS
Telematics Control Unit (TCU)
Door Locks
Electronic Control Units (ECU)



Numbers

Text

Text

twitter

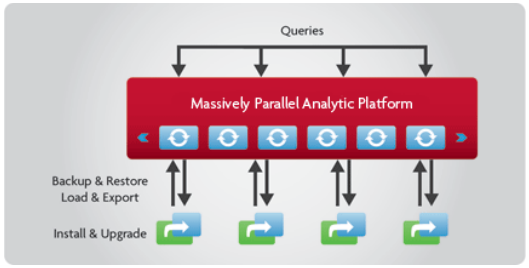
KAKAO TALK

NAVER

Images

Why "빅데이터" talk?

- 드디어, 빅데이터 저장, 분석 가능!
  - 데이터의 분산화
  - 계산의 병렬화



Why “빅데이터” talk?

빅데이터, 근데 정말 왜 하는가?

빅데이터, 근데 정말 왜 하는가?

- 데이터로부터 **Insight** 도출
- 데이터로부터 **Foresight** 도출

빅데이터 특징 FIFA

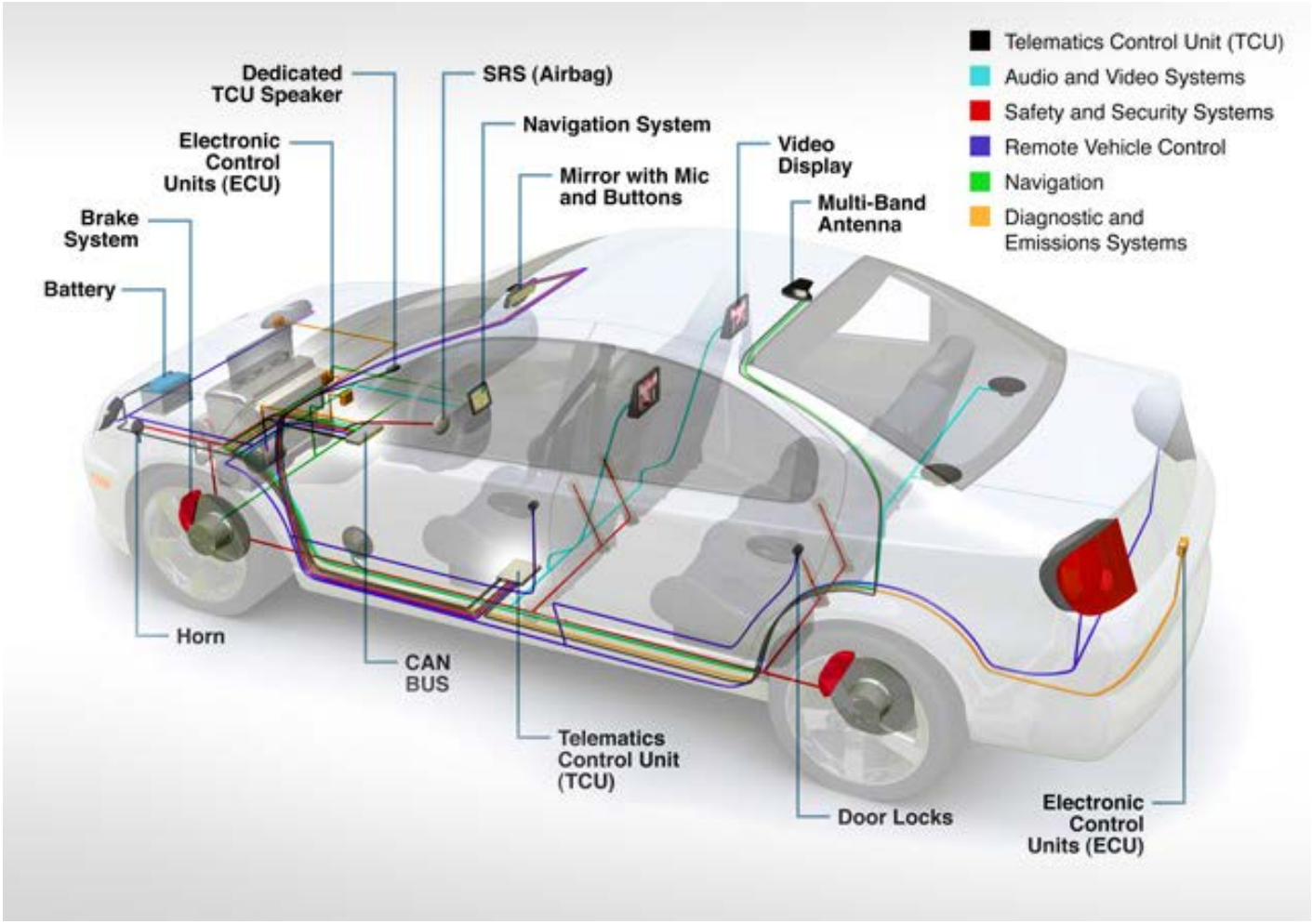
- Fair 편견 없는
- Individual 개별적인
- Far reaching 큰 변화를 가져오는
- Always 상시적인

빅데이터 실현 IOC

- Individual 개별화
- Objective 객관적
- Continuous 상시적

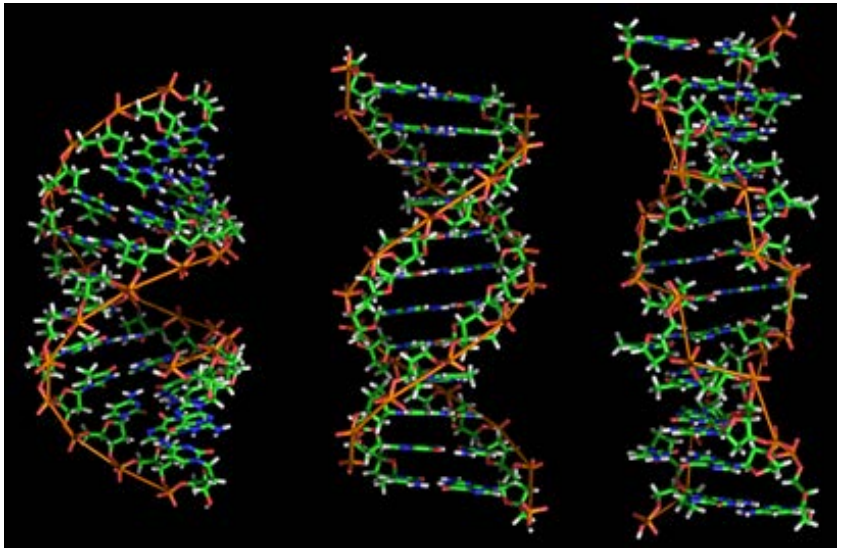
개별화 처리

예방 vs 예측


Telematics Control Unit (TCU)
Audio and Video Systems
Safety and Security Systems
Remote Vehicle Control
Navigation
Diagnostic and Emissions Systems
Dedicated TCU Speaker
SRS (Airbag)
Navigation System
Video Display
Electronic Control Units (ECU)
Mirror with Mic and Buttons
Multi-Band Antenna
Brake System
Battery
Horn
CAN BUS
Telematics Control Unit (TCU)
Door Locks
Electronic Control Units (ECU)

예방 vs 예측

맞춤형 의료



Mass ➔ Target ➔ 1-1

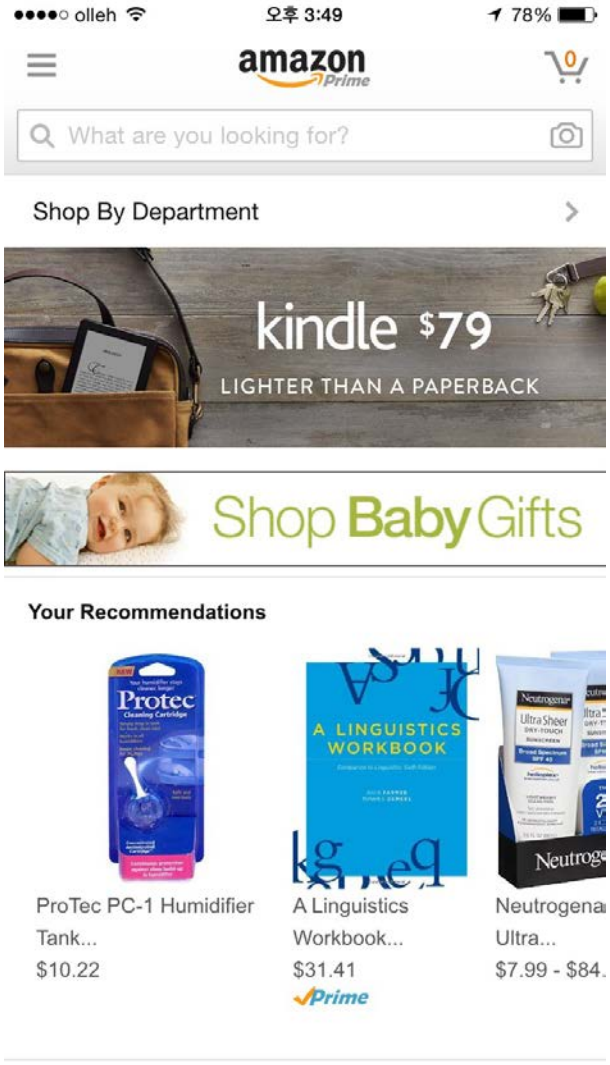
- Mass
- 모든 이들에게 같은 제품, 같은 메시지

Mass ➔ Target ➔ 1-1

- Targeting
- 시간대, 위치에 따라 차별화

Mass ➔ Target ➔ 1-1

- 1-1
- 개별화



Mass ➔ Target ➔ 1-1

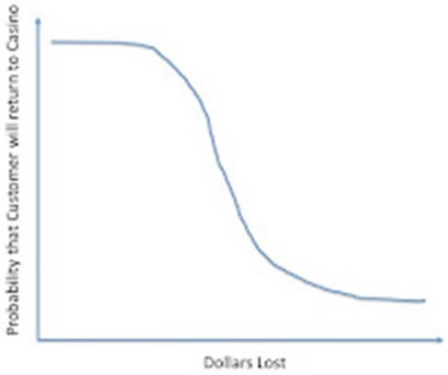
- 1-1
- 추천

Mass ➔ Target ➔ 1-1

- 1-1
- 추천


TOTAL
REWARDS®
GOLD
ROBIN KLINE
12345678901
EXP 02/2050

Customer Intelligence: The Pain Point



객관적 “증거기반” 의사결정

비 증거 기반 의사결정

SOUTHWEST AIRLINES

비 증거 기반 의사결정

- United benchmarking Southwest, 1994
  - 유니폼 및 기내식 없애고
  - 보잉 737 만 운행
  - 지상 체류 시간 최소화
  - 운항 횟수 최대화

비 증거 기반 의사결정

SOUTHWEST AIRLINES

비 증거 기반 의사결정

- 실패 원인
  - 실제 성공의 요인 이유를 입증하지 않고…
  - 눈에 보이는 것만 copy
  - 중요하지 않거나 나쁜 관행 copy

비 증거 기반 의사결정

- Southwest 의 성공 요인

믿음 #1 “stock option”

- “스톡옵션이 없다면 회사에 주주로 임하는 직원은 없고, 피고용인만 남게 될 것이다”

  ~Cyprus 반도체 CEO J. Rogers

믿음 #1 “stock option”

- 반증
  - 주주의 이익에 도움이 안 됨, National Bureau of Economic Research, 2002
  - 주식 소유와 재무 실적과 무관, Dalton et al, 2003
  - 비중 클수록 실적 조작 가능성 높아, NYT, 2001

믿음 #2 "선점 효과"

- "시장에 처음 진입한 기업에게 경쟁적 우위가 있다"
- 증거 불충분 및 의견 분분
- 반증
  - Amazon?
  - Internet Explorer? (Mosaic)
  - IBM compatible PC?

증거 기반 진료

- 방혈 bloodletting
  - 환자의 피를 빼내는 치료법
  - 히포크라테스 시절에도 존재
  - 이후 19세기까지

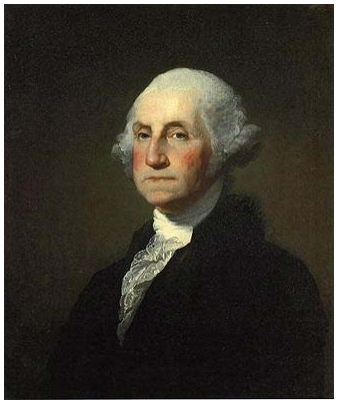
증거 기반 진료

- Pierre Louis 1836
  - Clinical Trial 시조

  - 폐렴 환자들 대상 방혈 vs 비방혈 test,
  - 방혈 환자가 더 많이 사망

증거 기반 진료

- 미국 대통령 사망 1797,
  - 급성 인후염, 3인의 주치의가 혈액의 ½ 방혈
  - 당시 미국에서 가장 유행하던 치료법



선수 평가

- 축구나 농구에서 득점을 지원하는 선수의 contribution?

- Assist
  - 득점에 성공하면 1 어시스트
  - 득점에 실패하면 0 어시스트

선수 평가

- https://youtu.be/OVY15CHD6FQ

Sports Analytics

- 경기장 양쪽 3대씩 총 6 대의 카메라로
- 모든 선수와 공의 X,Y,Z 위치 데이터를
- 1초 당 25회 기록

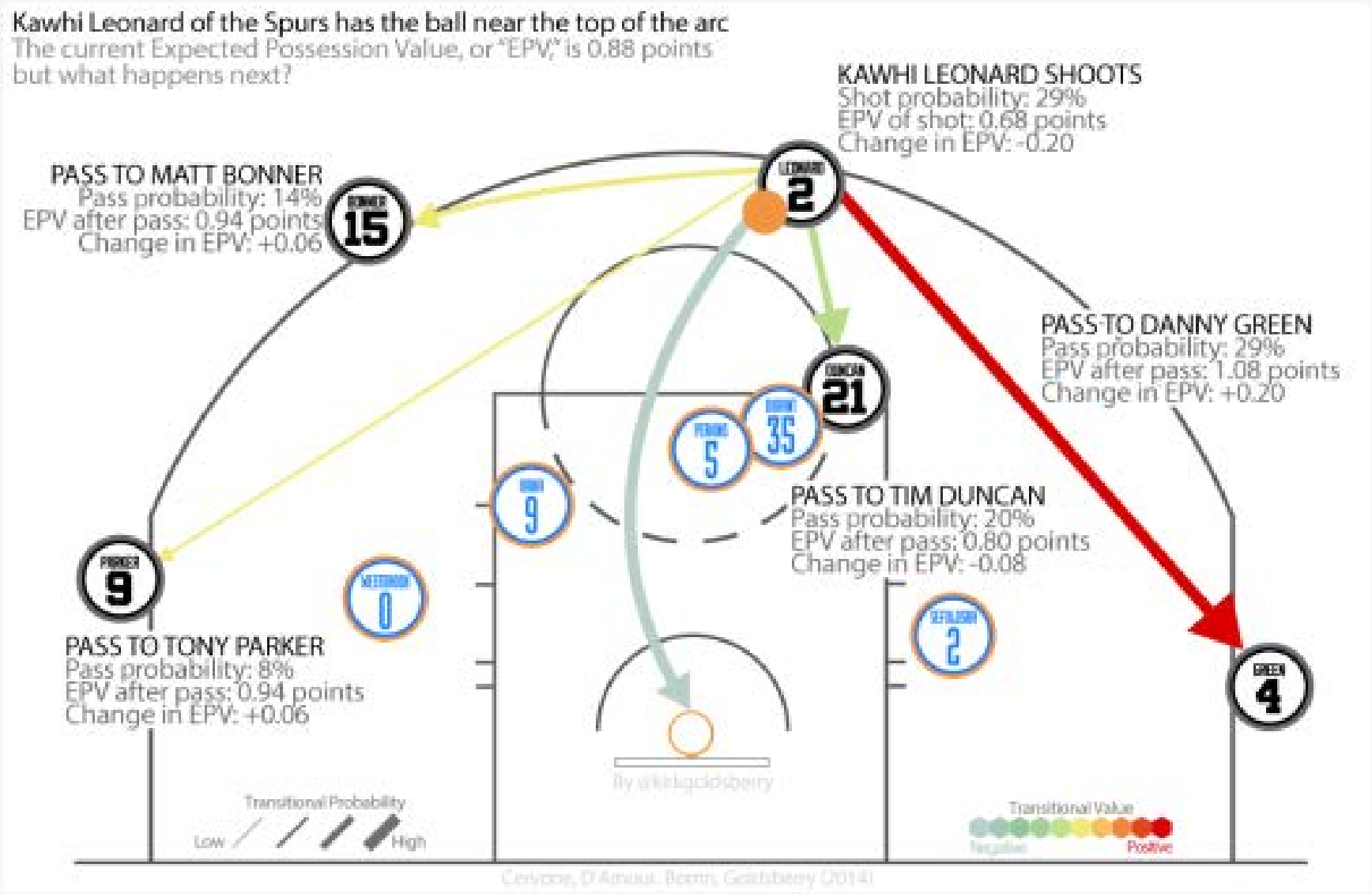
From http://www.stats.com

Sports Analytics

- Expected Possession Value (EPV)
- 공을 소유한 팀의 순간 예상득점치
  - 해당 순간에 공과 선수들의 위치가 주어졌을 때, 공격하는 팀이 득점할 수 있는 점수기대치
  - 예) 상대 팀 골대 아래 혼자 공을 가지고 있는 경우?

- 마르코프 모형

POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data, by Dan Cervone, et al 2014



POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data, by Dan Cervone, et al 2014

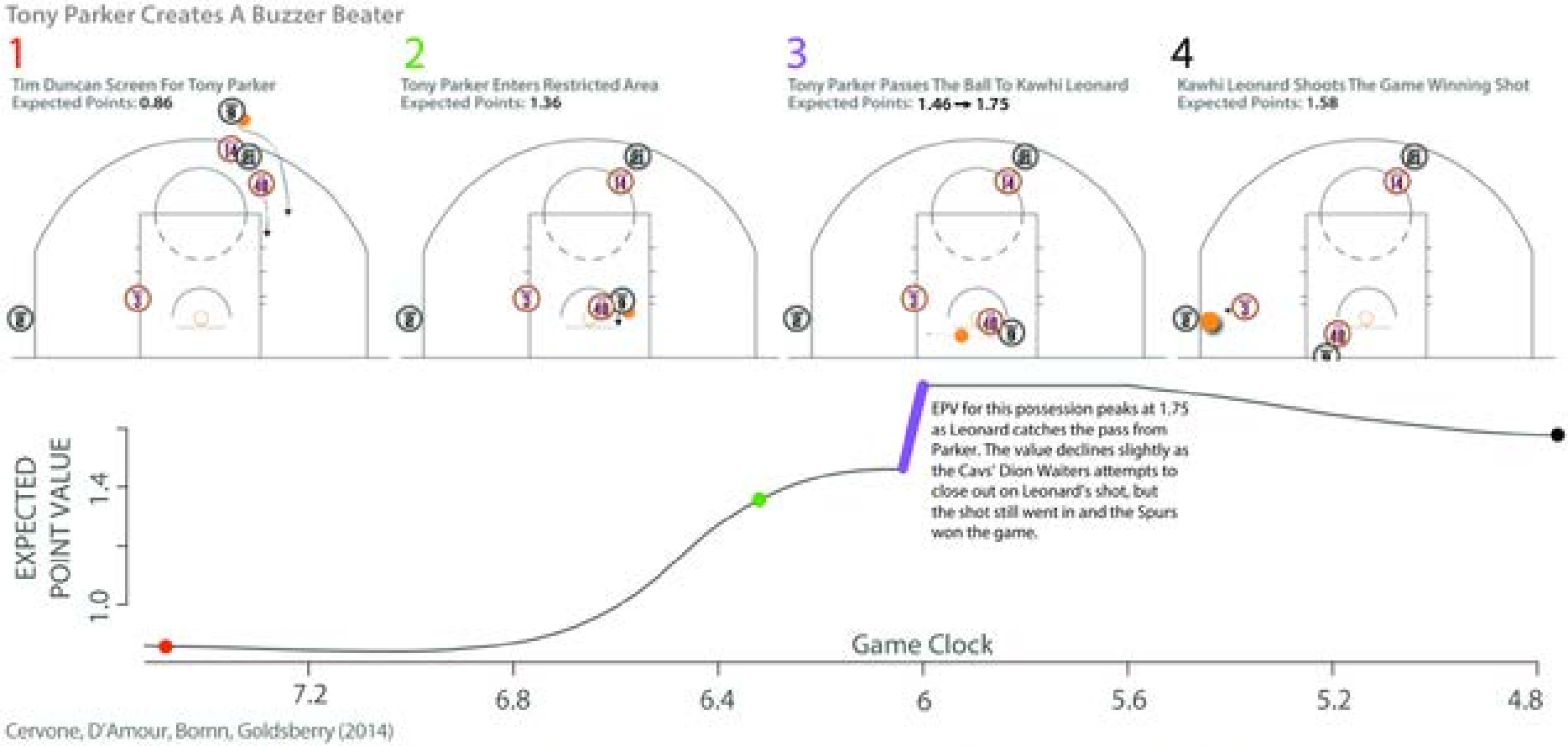


Figure 2. EPV throughout the Spurs' final possession, with annotations of major events.

POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data, by Dan Cervone, et al 2014

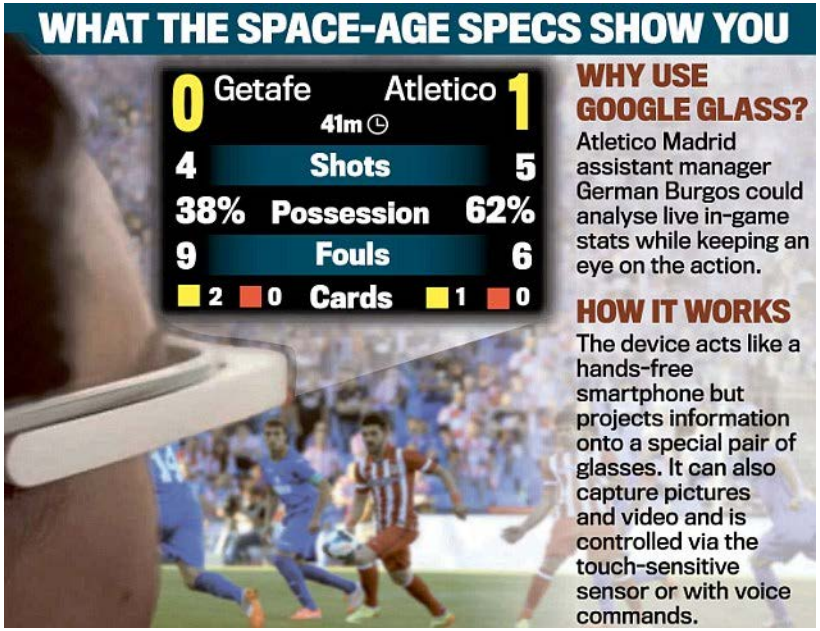
Football

From http://www.stats.com

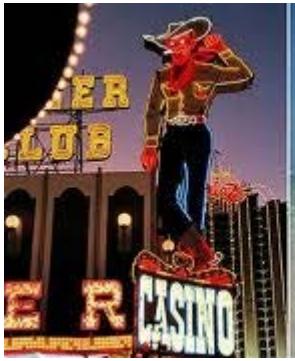
Football

Football


0 GET ATM 0
0 GET ATM 1
0 GET ATM 2


WHAT THE SPACE-AGE SPECS SHOW YOU
0 Getafe Atletico 1
41m
4 Shots 5
38% Possession 62%
9 Fouls 6
2 0 Cards 1 0
WHY USE GOOGLE GLASS?
Atletico Madrid assistant manager German Burgos could analyse live in-game stats while keeping an eye on the action.
HOW IT WORKS
The device acts like a hands-free smartphone but projects information onto a special pair of glasses. It can also capture pictures and video and is controlled via the touch-sensitive sensor or with voice commands.

카지노 Myths



카지노 Myths 파괴

- 데이터
  - 고개 별 매출액, 좌석 이용률, 수익성, 직원 이직률
- Gary Loveman 1998 Harrah's COO
  - 3금 정책: 절도/성희롱/무실험정책

HARVARD BUSINESS SCHOOL

Harrah's CASINOS

CAESARS ENTERTAINMENT

카지노 Myths 파괴

카지노 Myths 파괴

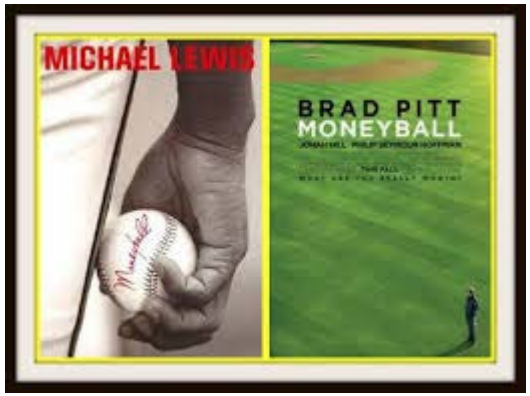
카지노 Myths 파괴

MLB Myths

- 비싼 선수가 잘 한다
- 인건비를 많이 쓰면 승리한다
- 번트/도루/히트앤드런, 작전으로 승리한다

MLB Myths 파괴

- 가장 데이터가 많이 생성되는 스포츠



MLB Myths 파괴

- 비싼 선수가 잘 한다

- 과거의 화려한 경력이 미래를 보장하지 못한다
- 나이 많고 부상에 시달린다
- 데이터로 "지금 막 뜨는 선수" 발굴

MLB Myths 파괴

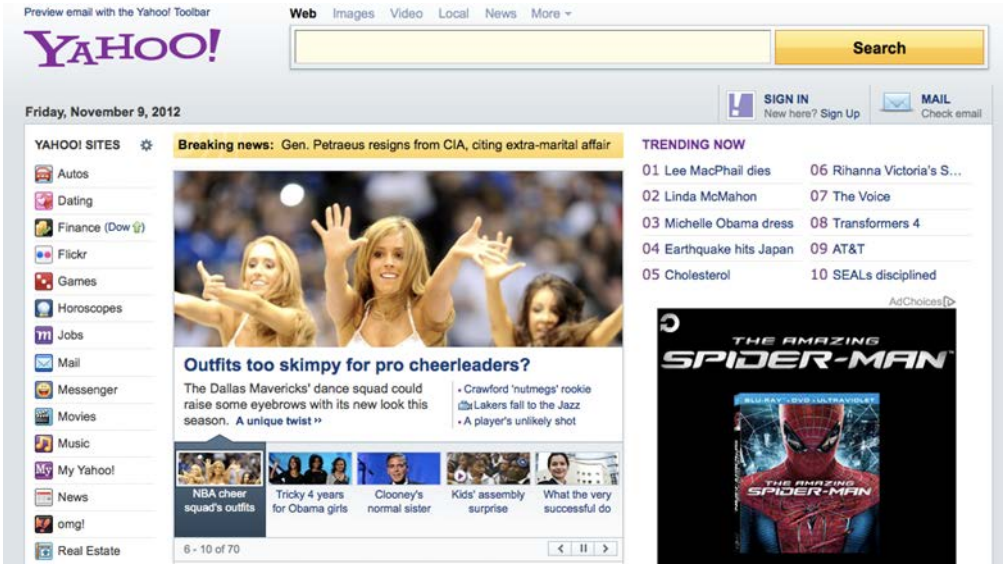
- 인건비를 많이 쓰면 승리한다

- 양키즈 103승에 $130M, 2002년
- A's 103승에 $40M, 2002년

MLB Myths 파괴

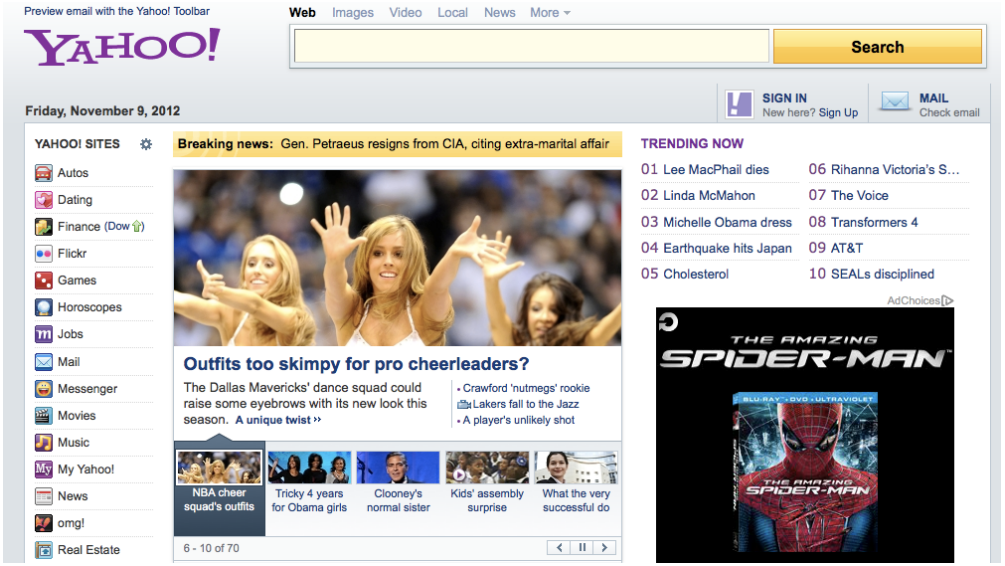
- 번트/도루/히트앤드런, 작전으로 승리한다
- 번트, 도루 최소, 2002 A's (AL West #1)

믿음 대신... 증거 활용



- 매달 평균 5억명 방문, 30개 언어
- 2013.7 미국 방문자 2억명 > Google
- 1년 사이 21% 증가

Plan / Test / See cycle

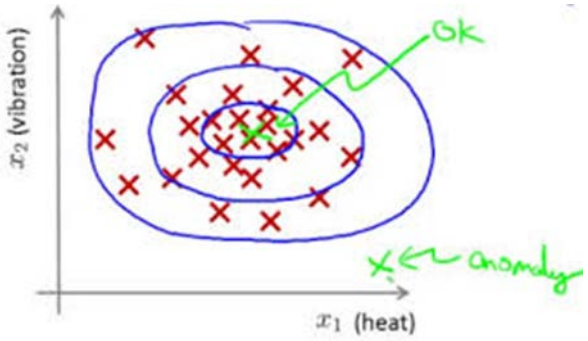


- 기사, 광고, 버튼의 위치/컬러 실험
- 10~20만 treatment vs 나머지 control
- 실시간 효과 검증 => 탁상 논쟁 무의미

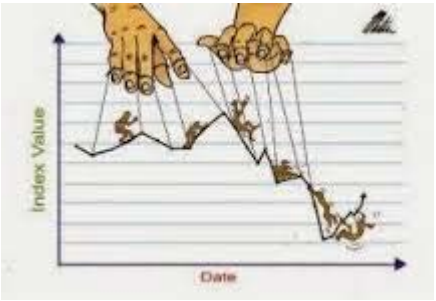
상시적 관찰


Überkopplung
A


1:47
65°


Ok
anomaly
x_2 (vibration)
x_1 (heat)


GOOG
Price Gap 1
Price Gap 2
Price Gap 3


Index Value
Date


MARKET
MANIPULATION
FREE
TRANSACTIONS
PAINTINGS
ORDER
SELLING
FRAUD
SUBSTANTIALLY
RARE
SHARE
MARKET
PLACES
QUICK
INCREASE
PERIOD
GIVE
GROUP
ARTIFICIALLY
CURRENCY
STOCK
TRADER
COMMODITY
SPECIFIC
ATTRACT
CREATE
MAINTAIN
OCCUR
IMPRESSION
VOLUMINOUS

빅데이터 누가하지? WHO?

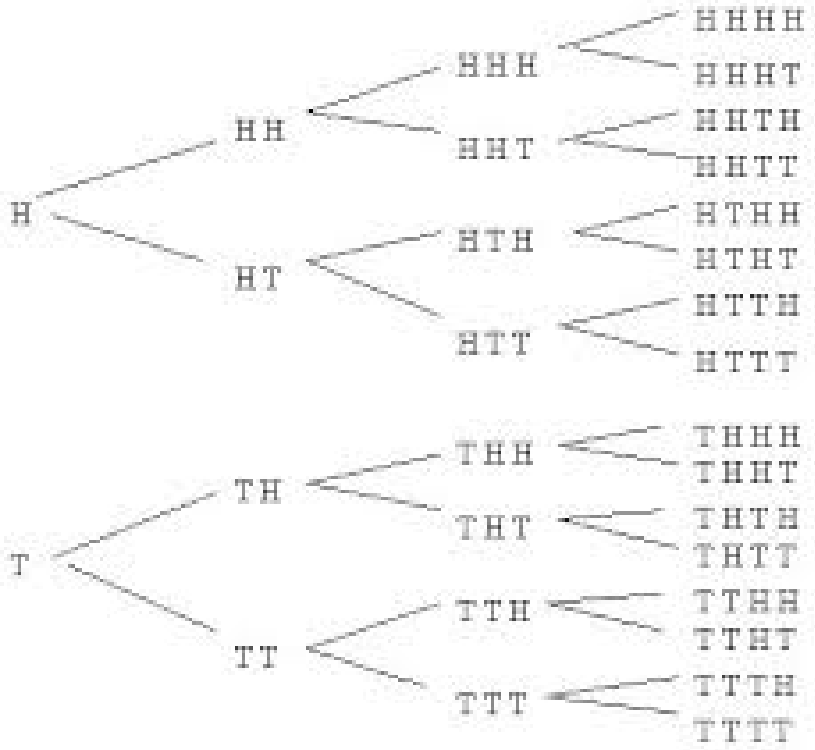
빅데이터 핵심 역량

리더

- Who? CEO, 본부장, 담당 임원,
- To Do
  - 의사결정자 **지지**
  - 분석가 및 인프라 **투자**
- 조건: Big Data 분석의 value 이해

리더

- 데이터 기반 의사결정 안 하는 리더
  - Intuition 으로 성공



리더

NOKIA

리더

- 1st thing to do
  - 자체 확보 / 외부 접근 가능 데이터 파악
- 2nd thing to do
  - Business value / Big Data 활용 분야 선정
- 3rd thing to do
  - 분석 전략 수립

리더

의사결정자

- Who? 마케터, 상품기획자, 공정 엔지니어
- To Do
  - 빅데이터 분석 인사이트를 의사결정에 적용
  - 데이터만을 고려한 의사결정 (x)
  - 데이터도 고려한 의사결정 (o)

의사결정자

- 조건
  - 빅데이터 분석에 대한 이해 필수
    - 어떤 종류의 분석이 있고,
    - 어떤 결과를 주고,
    - 어떤 한계가 있는지
- 방법
  - 교육 1~3주 fulltime

분석가

- Who?
  - 데이터사이언티스트, 데이터마이너, 분석가
- To Do
  - 데이터에서 비즈니스 인사이트/포사이트 도출
- 조건
  - 분석 지식: 컴퓨터, 통계학, 산업공학
  - 적용 지식: 해당 비즈니스
  - 스킬: 커뮤니케이션, fast learning

Big Data Analysis

"Insight"

Business Decision Maker

Visualization

BI tools

Stopfire, Tableu, ...

Visualization module

Apache Server & Tomcat

Toad, Beeswax ...

SQL

Data Miner

Text Mining

KDD

Clustering

Classification

Association rule analysis

NLP

Tokenizing

PoS tagging

Keyword extraction

Big data analyst

Storage, Low-level API

Map/Reduce job

Hive

Hadoop cluster

NameNode

M/R job submit

SQL query

ETL

RDBMS

Sqoop

HDFS, Map/Reduce operation

DataNode

DataNode

DataNode

...

...

...

DataNode

DataNode

DataNode

Data collection

Crawler

Crawler

Crawler

Crawler

twitter

blog

"Fact"

System engineer

Big Data People

Business Decision Maker

Analytical Mind; business understanding and interpersonal communication

Data Miner

Statistical analysis; multiple regression, logistics regression, nonlinear least regression

Descriptive analytics; clustering, association analysis

Predictive analytics; classification, prediction, time series, novelty detection

Optimization; linear/nonlinear programming, integer programming, combinatorial optimization

Natural language processing; word segmentation, PoS tagging, keyword extraction

Image processing; image segmentation, filtering & neighborhood processing

Big data analyst

Analytic for big data; Hadoop, map-reduce, hive

Data visualization

System engineer

Data management & Information processing; database design …

Data warehousing & workflow management; ETL, OLAP, workflow, business intelligence,

system performance evaluation…

분석가

- 분석가 양성 및 임원 재교육
  - 2012. 2 New York Times 예측
  - 분석가 14~19만 명, 매니저 150만 명

The New York Times

NEWS ANALYSIS "The Age of Big Data" By STEVE LOHR
Published: February 11, 2012 82 Comments

분석가

- 방법
  - 내부 인력 "특화" **교육** (최소 4주, SK, SDS)
  - 외부 인력 **채용** (GE) 별도 조직 구성 여부
  - 외부 인력 **활용** (대다수 기업)
    - 내부 counterpart 중요

Big Data

- 가스 터빈
- 하루에 500GB의 센서 데이터 발생,
- 전 세계 12,000대 작동 중

Big Data

- 제트 엔진
- 엔진 데이터 + 비행 정보
- 최근 1년 발생 데이터 > 지난 96년 발생 누적 데이터

GE Global Research

- **실리콘밸리에 “Global SW Center” 건립**
  - 2011년, 총 10억 달러를 투자
  - 소프트 파워를 강화하고 데이터 분석을 통한 GE의 사업 역량 강화가 목적

GE Global Research

- **전 세계 유수의 데이터 과학자(Data Scientist)영입**
  - 구글, 애플 등에서 기계 학습, 통계학 등의 분야 총 400여 명의 인재 확충
  - 부사장 윌리엄 러(前, CISCO systems): “지원자의 5% 이내 최고의 실력 인재만을 선발”

GE Global Research

"The Software Sciences and Analytics team is bringing together our knowledge of data, analytics, and computing to grow GE's businesses in exciting new ways. We develop advanced computing and decision-making tools to analyze, interpret and utilize data, creating software systems, solutions and architectures that will change the way our customers create, deliver and manage their businesses."

GE Global Research

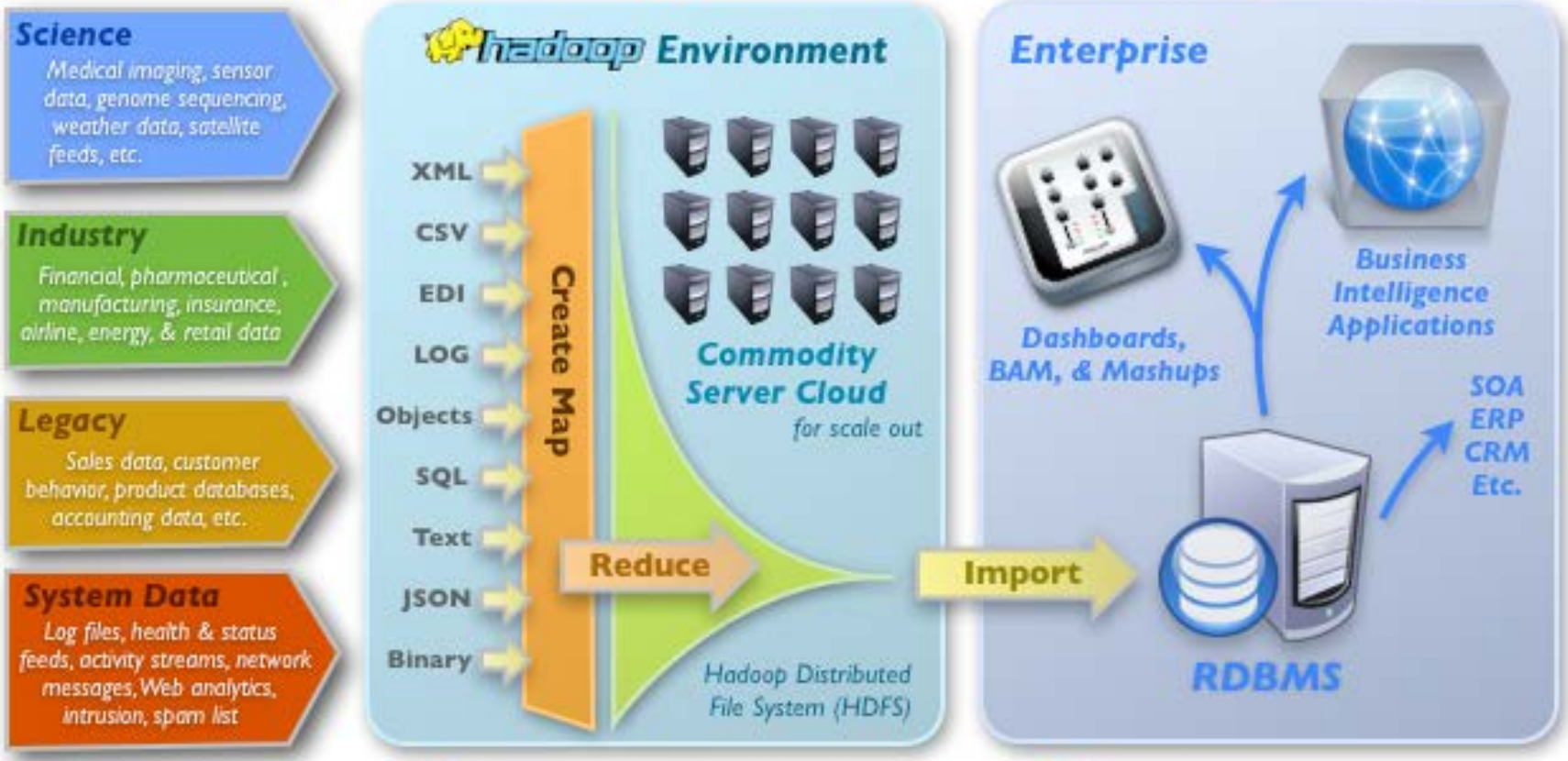
- 빅데이터/분석가 들을 별도 조직에 모은 이유

- **고급 인재 부족**
  - BI 개발자, 프로그래머는 많지만 데이터 속의 지식을 발견할 수 있는 인재는 태부족
  - 높은 수준의 분석은 이러한 소수의 고급 인재들에 의해 수행
- **퇴사 방지**
  - 현업 배치되는 경우, 이들의 승진, 비젼이 어두움
  - 데이터 분석에 집중하여 그 곳에서 승진할 수 있도록
- **기술의 공유**
  - 다양한 현업 문제를 해결하는 공통된 분석 기술

빅데이터 인프라?

- 아파치 오픈 소프트웨어 프로젝트
- 구성요소
  - MapReduce Framework
  - Hadoop Distributed File System (HDFS)
  - Hive

Using Hadoop in the Enterprise



From http://www.ebizq.net/blogs/enterprise

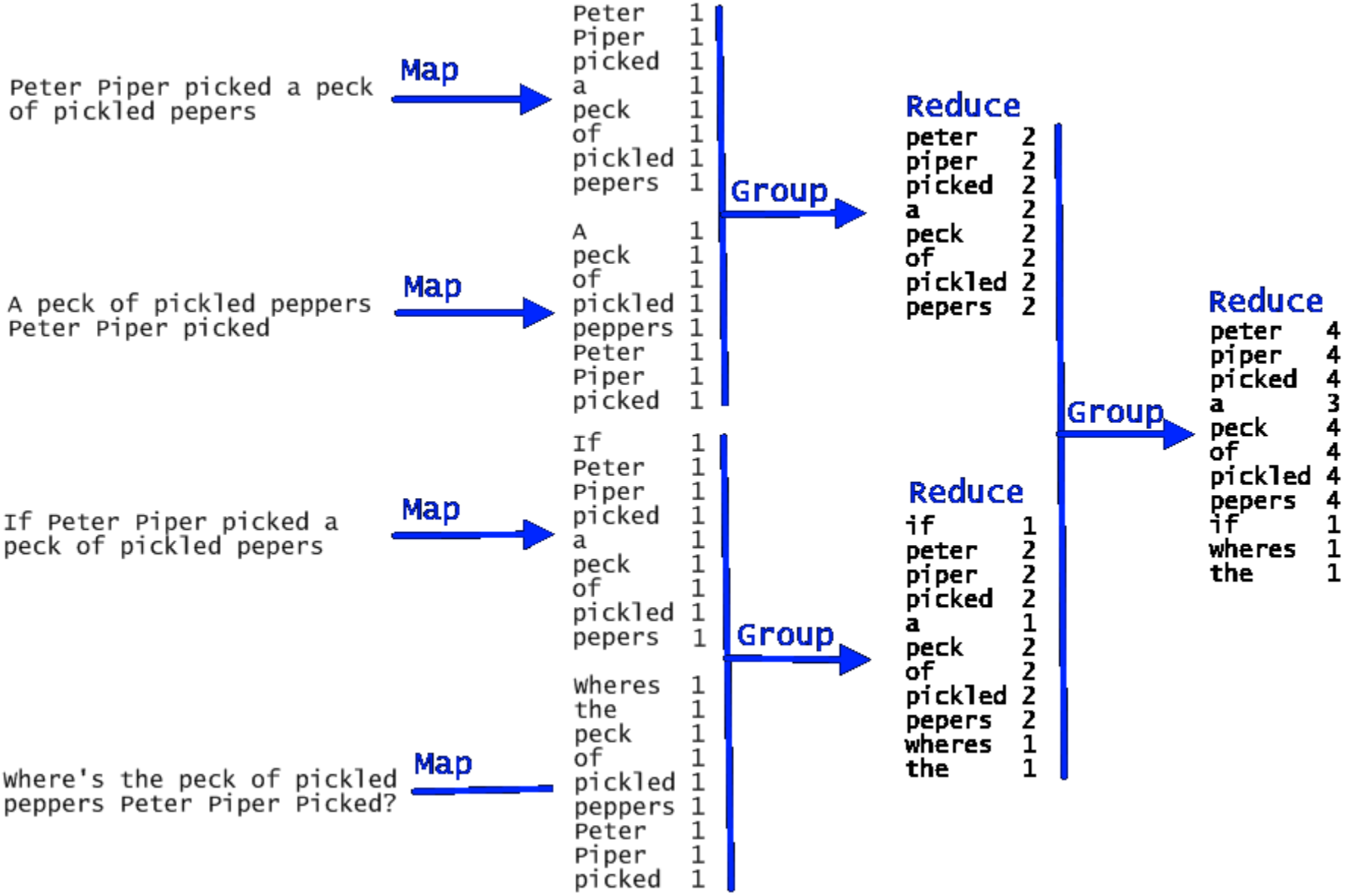
MapReduce

- 수십 수백 TB의 데이터를
- 수 천 대의 일반 PC (클러스터) 상에서
- 결함포용적으로 병렬로 처리하는
- 응용 소프트웨어를
- 쉽게 작성할 수 있게 해주는 소트프웨어 프레임워크

MapReduce *job*

- Input 데이터 셋을 map 태스크가 병렬적으로 처리할 수 있는 독립적인 청크로 분할
- 또한 Map 의 output을 순서대로 걸러서 reduce 태스크로 보냄.

- MapReduce 프레임워크는 다양한 태스크의 스케쥴링, 모니터링, 그리고 실패한 태스크의 재실행등을 관리


Peter Piper picked a peck of pickled pepers
Map
Peter 1
Piper 1
picked 1
a 1
peck 1
of 1
pickled 1
pepers 1
A peck of pickled peppers Peter Piper picked
Map
A 1
peck 1
of 1
pickled 1
peppers 1
Peter 1
Piper 1
picked 1
If Peter Piper picked a peck of pickled pepers
Map
If 1
Peter 1
Piper 1
picked 1
a 1
peck 1
of 1
pickled 1
pepers 1
Where's the peck of pickled peppers Peter Piper Picked?
Map
Wheres 1
the 1
peck 1
of 1
pickled 1
peppers 1
Peter 1
Piper 1
picked 1
Group
Group
Reduce
peter 2
piper 2
picked 2
a 2
peck 2
of 2
pickled 2
pepers 2
Reduce
if 1
peter 2
piper 2
picked 2
a 1
peck 2
of 2
pickled 2
pepers 2
wheres 1
the 1
Group
Reduce
peter 4
piper 4
picked 4
a 3
peck 4
of 4
pickled 4
pepers 4
if 1
wheres 1
the 1

- 계산 노드와 데이터저장 노드 동일
- 맵리듀스 프레임워크와 HDFS 는 동일한 노드에서 돌아감
- 이런 구조 덕분에 모든 클러스터에서 높은 수준의 통합 bandwidth 가 가능함

빅데이터 걸림돌: 프라이버시 & 법 체계

“개인정보” 기본 태도	적절한 활용	생래적 정보 보유자의 인권 보호 배타적 지위의 부여

“공공데이터 제공 및 이용활성화에 관한 법률” (2013.10.31 시행)
“공공기관의 정보공개에 관한 법률”

2013. 안정행정부

개인 정보

- 비공개대상
- "당해 정보에 포함되어 있는 이름 주민등록번호 등 개인에 관한 사항으로서
- "공개될 경우 개인의 사생활의 비밀 또는 자유를 침해할 우려가 있다고 인정되는 정보"

- 판례: 식별 가능성 없어도...

보호 체계 [테크앤로 2013]

	개인정보 보호법	정보통신망법
동의 없는 개인정보 수집	5,000만원 이하 과태료	5년 이하 징역
동의 원칙 예외	다양한 예외 인정	한정적
업무 위탁	공개 또는 고지로 충분 위반 시 과태료	동의 필요 위반시 과징금

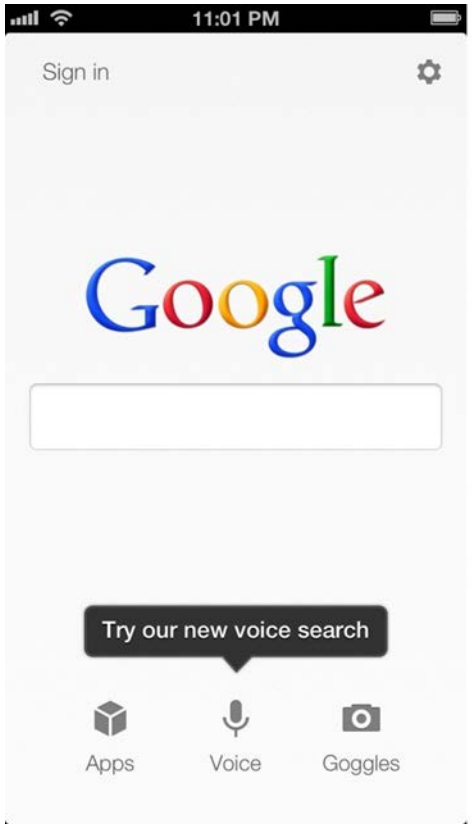
보호 체계 [테크앤로 2013]

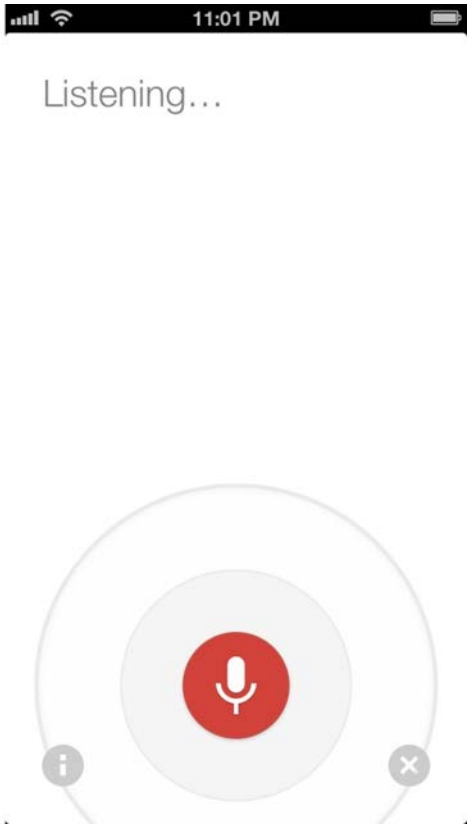
	개인정보 보호법	정보통신망법
동의 없는 개인정보 수집	5,000만원 이하 과태료	5년 이하 징역
동의 원칙 예외	다양한 예외 인정	한정적
업무 위탁	공개 또는 고지로 충분 위반 시 과태료	동의 필요 위반시 과징금

EMERGING TECHNOLOGIES

자연어 처리

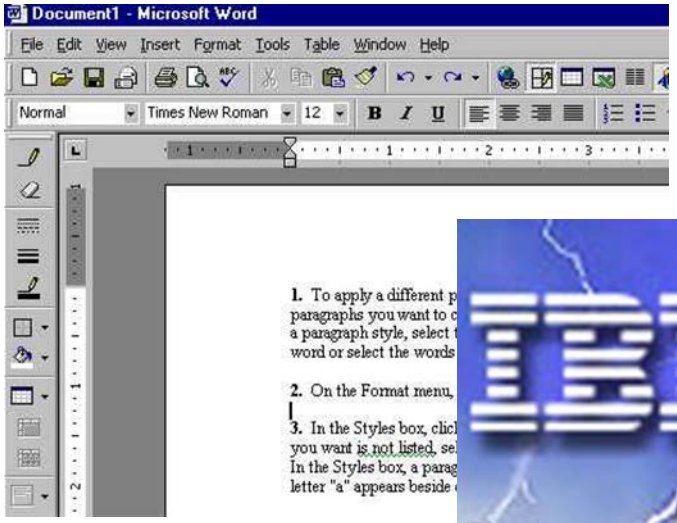
음성 인식


11:01 PM
Sign in
Google
Try our new voice search
Apps
Voice
Goggles

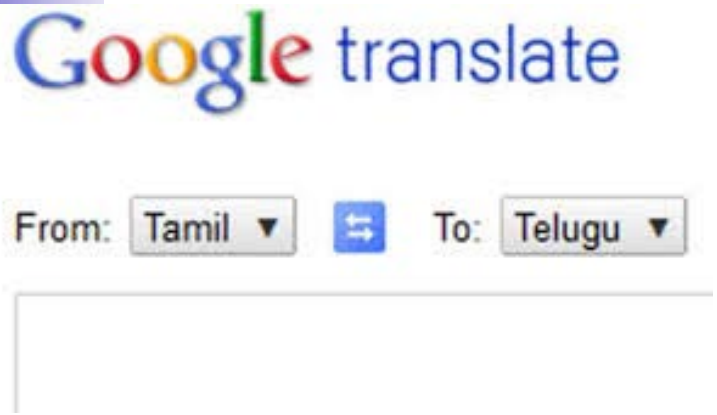

11:01 PM
Listening...

NUANCE®

자동 번역

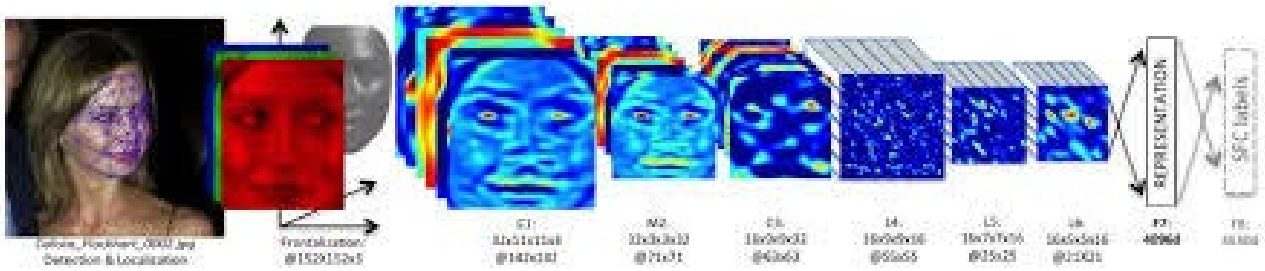

Document1 - Microsoft Word
File Edit View Insert Format Tools Table Window Help
Normal
Times New Roman
12


IBM


Google translate
From: Tamil
To: Telugu

Deepface

- Deep neural network trained on
- 4M photos of faces from around 4,000 users
- Resulting in 97.25% accuracy



CONCLUSIONS

Summary

- 1가지 목적
  - 분석 기반 객관적 의사결정
- 2가지 프레임워크
  - Descriptive Analytics: insight
  - Predictive Analytics: foresight
- 3가지 데이터
  - Numbers, Texts, Images

성공 전략

The Economist

- 빅데이터: 작은 gain, trial-and-error 반복을 통한 누적

- Plan Do See cycle

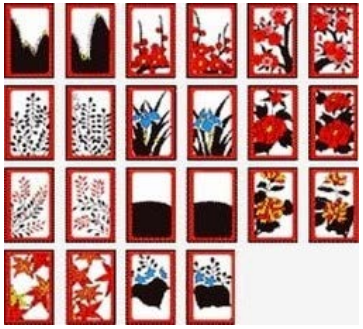
성공 전략

- 6만대 트럭, 1마일 단축, $50m

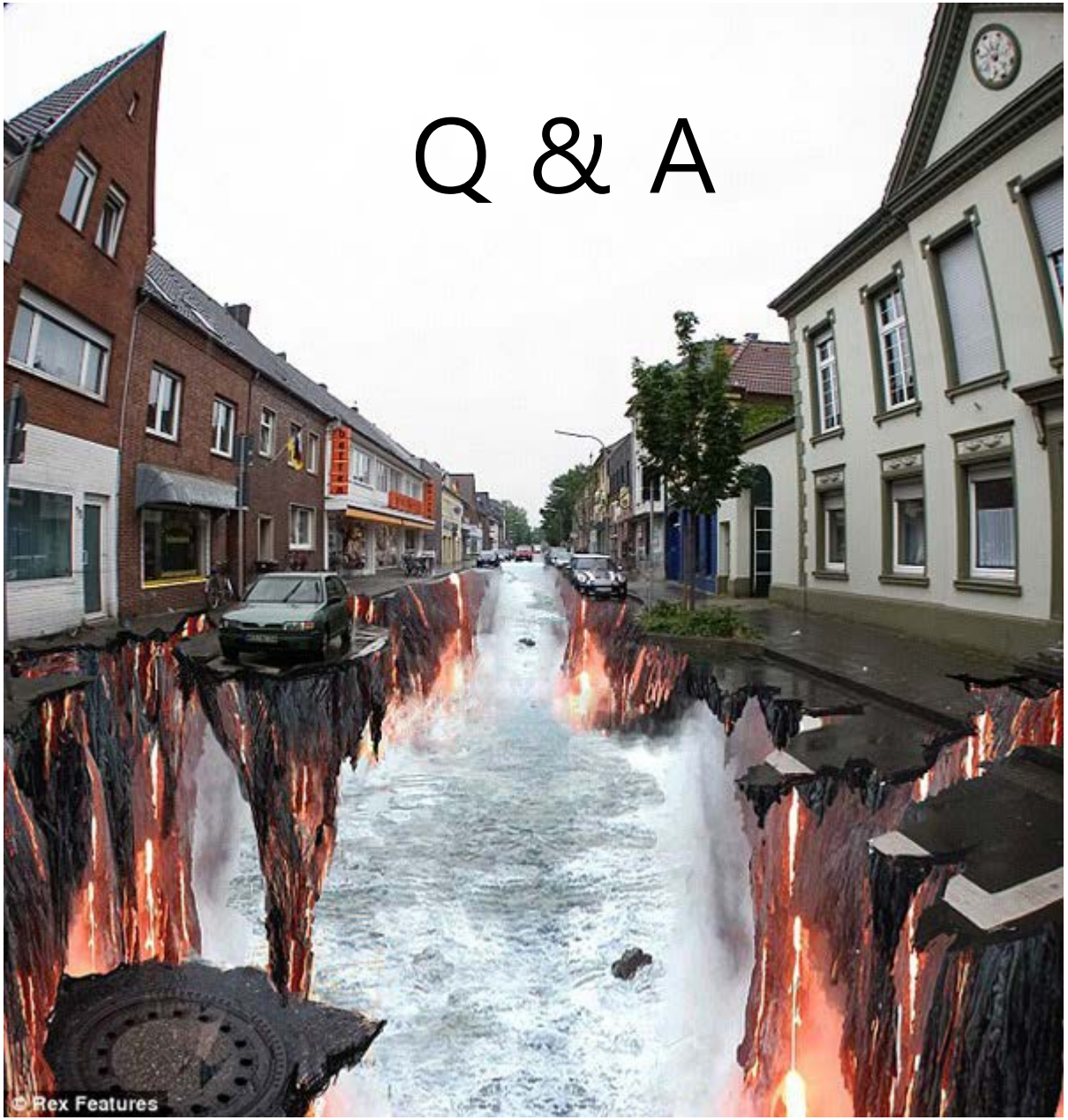
성공 전략

- 인사 생산성, 팀 별 0.5~1% gain 이 전체 22% gain
  - 7명 오버 +, 다른 시차 멤버 -, 서로 알면+

성공 전략



Q & A



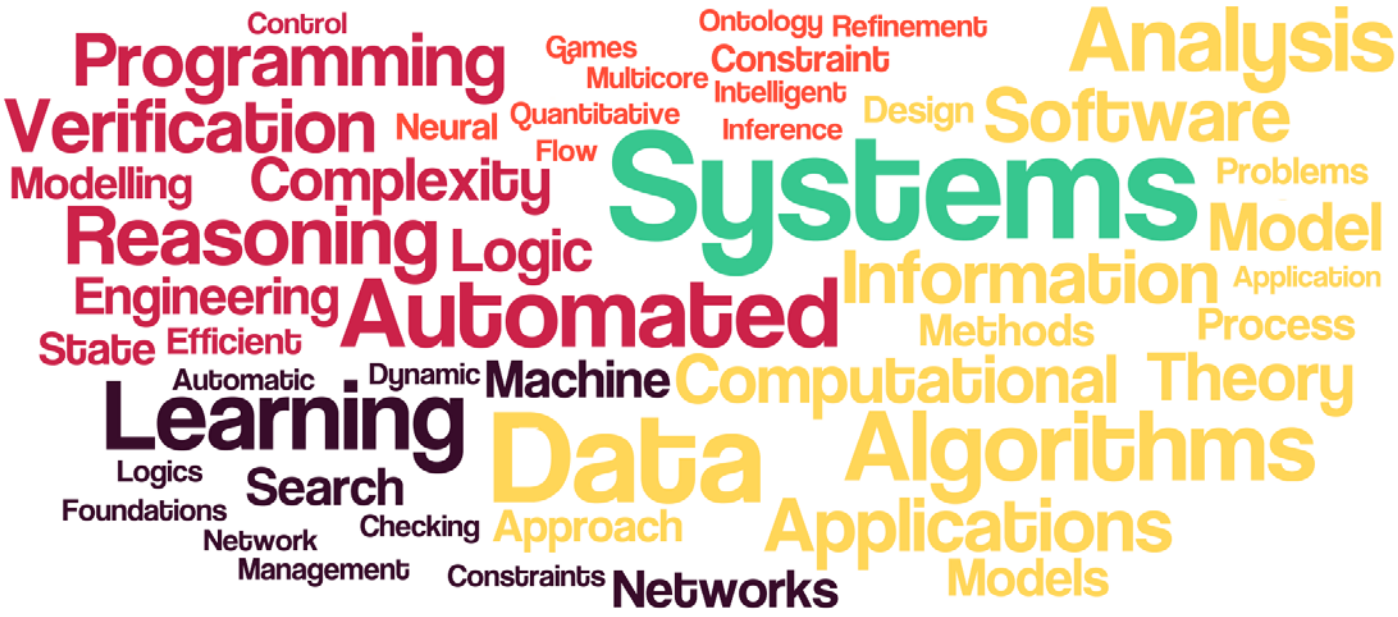
FAQ 1 IS THIS NEW?

What's new?

- 데이터 학문
  - 컴퓨터 과학
  - 통계학

Computer Science

- 데이터의 표현, 저장, 및 검색
- 빅데이터 인프라 제공



Source: http://www.macs.hw.ac.uk/texturelab/EPSRC/clusters.html

Statistics

- "stat"istics = 국가
  - "science dealing with data about the condition of a state or community" [Barnhart], from German *Statistik*, by German political scientist Gottfried Aschenwall, 1770
  - from Modern Latin *statisticum (collegium)* "(lecture course on) state affairs,"

국가는 데이터

- 국가의 존립에 핵심 역할

과학도 데이터

- 새로운 측정 기기 => breakthrough
- 전자현미경 => 의료 및 나노~ 반도체 산업

$$도 데이터

- 19세기 유럽의 Rothchild 가문
  - Amschel (프랑크푸르트), Nathan(런던), Calmann(나폴리), Jakob(파리), Salomon(비엔나) 파견
  - 데이터 수집, 시차 이용

경영도 데이터

- Peter Drucker
  - *you can't manage what you can't measure.*

Statistics vs Big Data

- Statistics
  - 데이터 수 30개
  - 개개 데이터의 정합성 critical
  - 샘플링
- Big Data
  - 데이터 수 Big
  - 개개 데이터의 정합성 less critical
  - 전수, 모집단 대상

Statistics vs Big Data

- Statistics
  - 가설 => 데이터 수집 => 검증
  - Q) 가설은 누가 어떻게 만드나요?
  - A) 그 분야 전문가, 즉 지식이 있는 사람

- Big Data
  - 데이터 => 가설 발견 => 검증
  - 가설이 데이터로부터 "등장"
  - 즉, 지식 Knowledge 발견 Discovery

Statistics vs Big Data

- Statistics
  - 전체에 대한 이해
  - "대표값": 평균, 중간값
  - "차량의 90%가 2년 후에 브레이크 문제 발생"
- Big Data
  - 개체에 대한 이해
  - "이 차량은 30일 후에 브레이크 문제 발생"

FAQ 2 TEXT MINING

텍스트마이닝

- 문서는 “비정형” 데이터
- 문서 요약, 정서 분석
- **Natural Language Processing (자연어 처리 기법)**

Don Swanson

- 문헌정보학자
- 의료 논문 검색 Medline
  - 편두통 ~ 간질, 혈전발생" 관련 논문
  - 간질, 혈전발생 ~ 마그네슘 결핍" 관련 논문
- "마그네슘 결핍 => 편두통?"
  - 향후 실험으로 검증 확인
- 데이터 => 가설 => 검증 (cf 통계적 검정)

텍스트마이닝 전처리

- Filtering
  - 특수 문자, 문장부호 (예: “, !, (,), etc.) 제거
- Tokenization
  - 문장을 term들의 sequence로 분리하는 과정
- Stop-word removal
  - 관사와 같이 변별력이 없는 term (예: and, this, it) 제거

텍스트마이닝 전처리

- Stemming
  - Term 들을 어간(stem) 형태로 변환
- Pruning
  - 매우 낮은 빈도로 나타나는 term 들을 제거
- Vectorization
  - 하나의 문서를 하나의 숫자 벡터로 변환

1. Tokenization

처리 전

circuits within the main distribution panel that are doubled up (referred to as "double taps") should be separated. Each circuit should be served by a separate fuse or breaker. All junction boxes should be fitted with cover plates in order to protect the wire connections.

처리 후

circuits + within + the + main + distribution + panel + that + are + doubled + up + referred + to + as + double + taps + should + be + separated + Each + circuit + should + be + served + by + a + separate + fuse + or + breaker + All + junction + boxes + should + be + fitted + with + cover + plates + in + order + to + protect + the + wire + connections

Tokenizing with pre-defined punctuators:

사전에 정의한 분리기호를 기준으로 텍스트를 분리함

현재 사용 중인 분리기호는 다음과 같음: [)] [(] [[] []] [{] [}] [:] [;] [,] ["] [”] [〃] [‘] [&] [=] [+] [>] [<] [@] 등.

2. Stop-word removal

처리 전

circuits + within + the + main + distribution + panel + that + are + doubled + up + referred + to + as + double + taps + should + be + separated + Each + circuit + should + be + served + by + a + separate + fuse + or + breaker + All + junction + boxes + should + be + fitted + with + cover + plates + in + order + to + protect + the + wire + connections

처리 후

circuits + **(within)** + **(the)** + main + distribution + panel + **(that)** + **(are)** + doubled + **(up)** + referred + **(to)** + **(as)** + double + taps + **(should)** + **(be)** + separated + **(Each)** + circuit + **(should)** + **(be)** + served + **(by)** + **(a)** + separate + fuse + **(or)** + breaker + **(All)** + junction + boxes + **(should)** + **(be)** + fitted + **(with)** + cover + plates + **(in)** + order + **(to)** + protect + **(the)** + wire + connections

Stop-word removal step:

Stop-word는 변별력이 없는 단어들을 말하며, Stop-word list에 포함된 단어들을 분석에서 제외함

3. Stemming

처리 전(후)

circuits**(circuit)** + main**(main)** + distribution**(distribut)** + panel**(panel)** + doubled**(doubl)** + referred**(refer)** + double**(doubl)** + taps**(tap)** + separated**(separ)** + circuit**(circuit)** + served**(serv)** + separate**(separ)** + fuse**(fuse)** + breaker**(breaker)** + junction**(junction)** + boxes**(box)** + fitted**(fit)** + cover**(cover)** + plates**(plate)** + order**(order)** + protect**(protect)** + wire**(wire)** + connections**(connect)**

Porter stemmer:

영문에 대한 형태소 분석 알고리즘인 porter algorithm을 사용하여, 영단어의 어근 형태로 변환함

4. Vectorization

Term	circuit	main	distribut	panel	doubl	…
Frequency	1	1	1	1	2	

- **TF (x IDF)**의 벡터 공간 모형(vector space model)
  - TF (Term Frequency): 문서 내 용어(term)의 출현 빈도
  - IDF (Inverse Document Frequency): 용어를 포함하는 문서 수의 역수

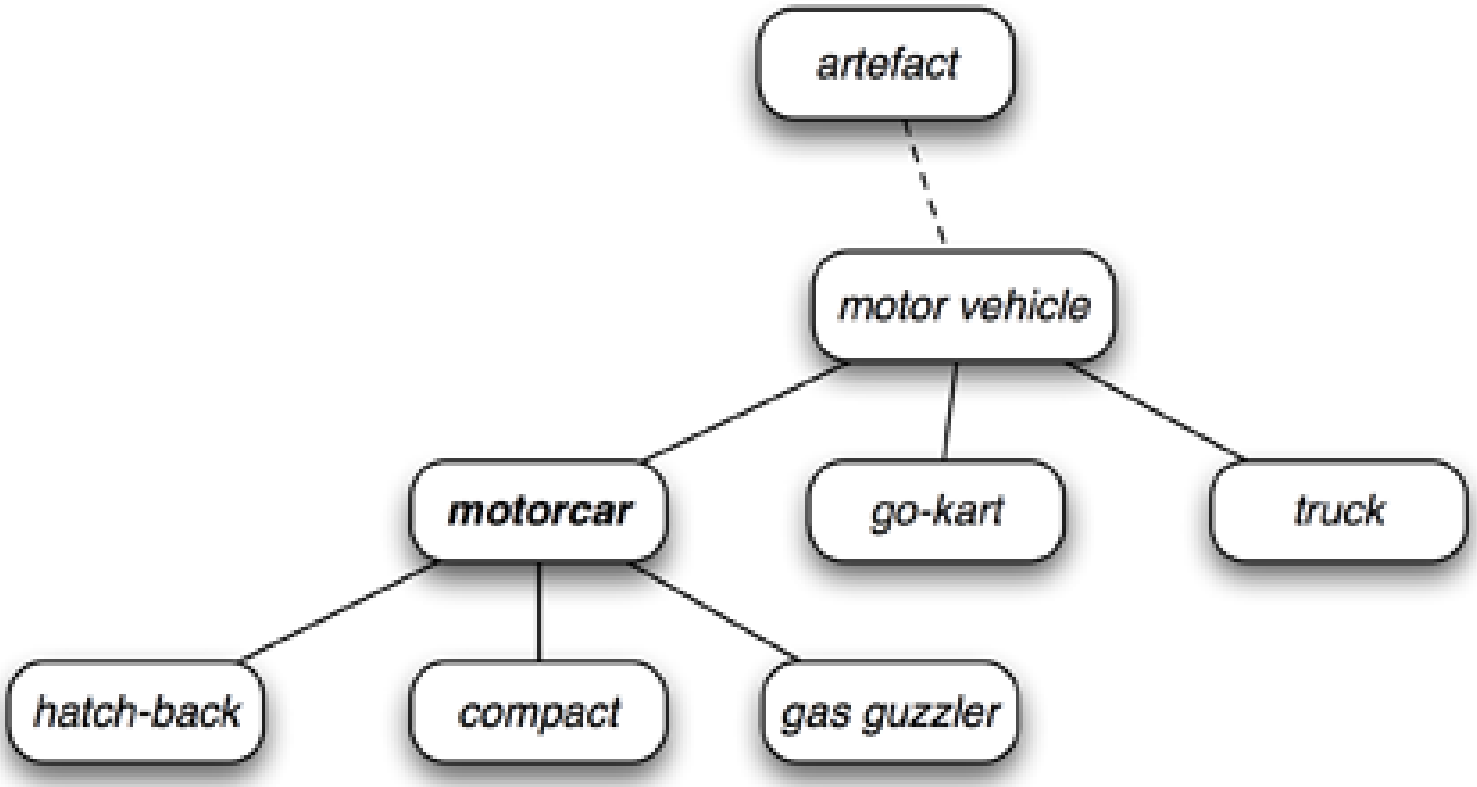
Term	circuit	main	distribut	panel	doubl	…
Doc #1	1	1	1	1	2	
Doc #2	2	4	3	2	1	
Doc #3	0	0	0	0	0	
Doc #4	2	1	2	1	1	

4. Vectorization

- 분석에 사용할 모든 용어들의 사전(dictionary)을 미리 정의
  - 문서의 domain 별로: 스포츠 vs 전자제품 vs 정치
  - Time consuming

- 비정형적인 텍스트 문서의 정형화 => 데이터마이닝 적용

WORDNET "artefact"



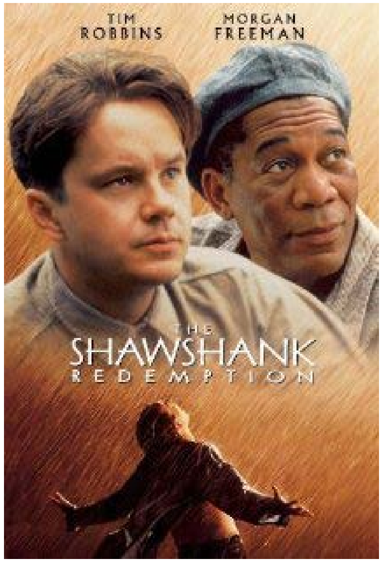
Big Data + Text

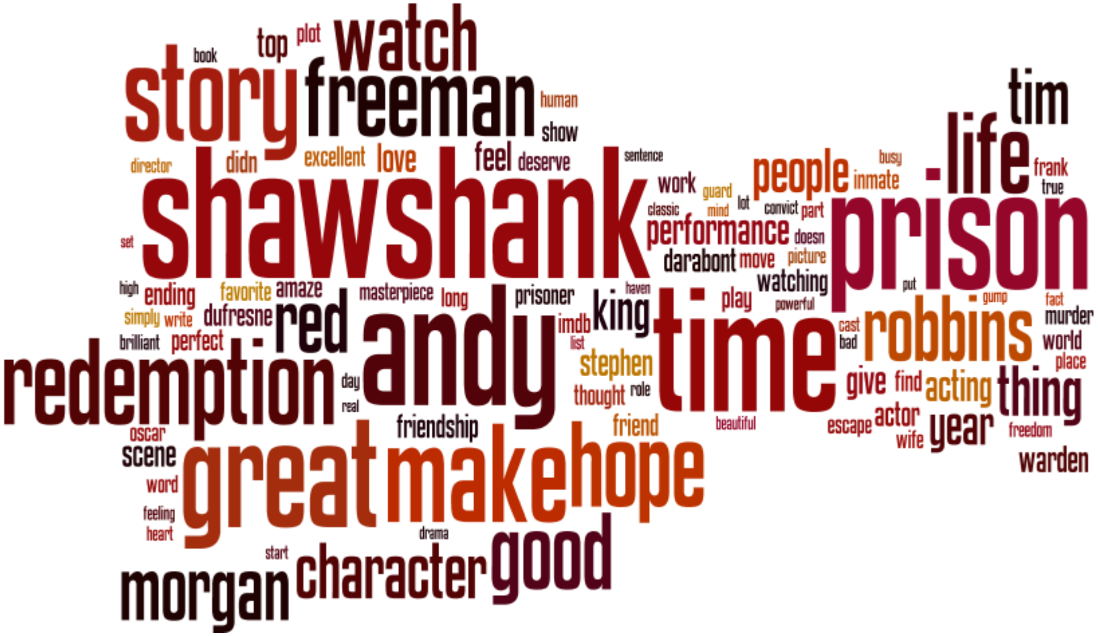
- Information Retrieval (정보검색)

빅데이터 + 텍스트

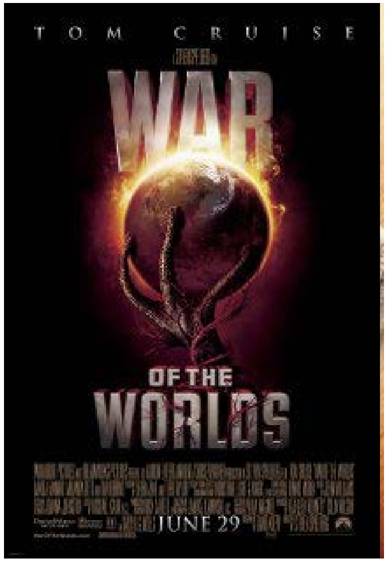
- Topic 추출
  - Understanding of doc topics
  - Clustering docs based on topics covered
- 텍스트 분류
  - Automatic News feed
  - Spam filter
- 정서 분석
  - Polarity of doc

Topic 추출

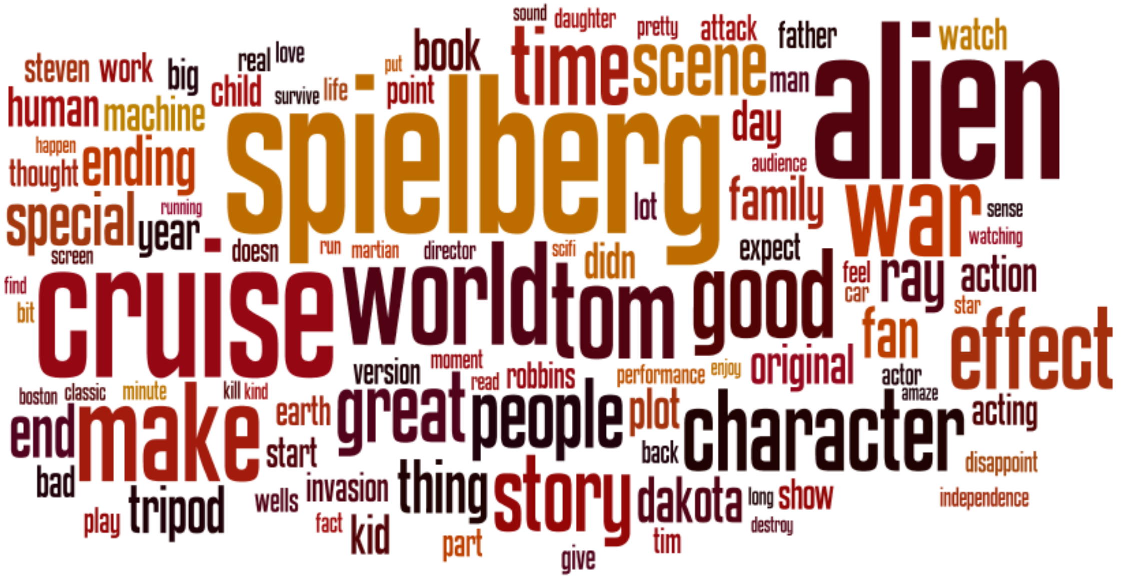

TIM ROBBINS
MORGAN FREEMAN
THE SHAWSHANK REDEMPTION


top
plot
watch
book
story
freeman
human
show
tim
life
director
didn
excellent
love
feel
deserve
sentence
busy
frank
true
shawshank
work
guard
people
inmate
classic
mind
lot
convict
part
prison
set
performance
doesn
darabont
move
picture
high
ending
favorite
amaze
masterpiece
long
prisoner
haven
watching
put
gump
fact
simply
write
dufresne
red
king
play
powerful
murder
brilliant
perfect
andy
imdb
list
time
cast
robbins
world
bad
place
redemption
day
stephen
give
find
acting
thing
thought
role
real
friendship
friend
beautiful
escape
actor
year
freedom
oscar
wife
scene
great
make
hope
warden
word
feeling
heart
drama
start
character
good
morgan

Topic 추출



TOM CRUISE
WAR OF THE WORLDS
JUNE 29



sound daughter pretty attack father watch
steven work big real love put book time scene man alien
human machine child survive life point spielberg day
happen thought ending audience
special running year lot family war sense watching
screen doesn run martian director scifi didn expect
find bit cruise world tom good feel car ray action star
fan effect
boston classic minute kill kind version moment read robbins performance enjoy original actor amaze
end make earth great people plot character acting
start back disappoint
bad tripod wells invasion thing story dakota long show independence
play fact kid part tim destroy give

Text 분류

- Spam Filter

Text 분류

- Field Claim classification

THE FORKLIFT WILL DRIVE IN REVERSE AND FORWARD BUT WHEN NEUTRAL APPLIED THE TRUCK STILL DRIVES IN FORWARD
FORKLIFT HAD A TRAVEL PROBLEM AT SECOND SPEED□
CLAIM IS FOR FREIGHT FOR PARTS RETURNED, PARTS WERE SENT FOC BY DOOSAN
NO HIGH SPEED
FORKLIFT HAD A TRAVEL PROBLEM AT SECOND SPEED□
TRANSMISSION PROBLEMS
NO HIGH SPEED
TRANSMISSION PROBLEMS
TRANSMISSION WILL NOT PULL IN HIGH GEAR.
UNIT WOULD NOT MOVE
no second gear
No second gear, see mail
TRANSMISSION HAD METAL IN IT.PER DAN SUMMERS WE INSTALLED A EXCHANGE TRANSMISSION.□
NO HIGH SPEED IN TRANSMISSION
After drain the oil, the techcian see that has pieces of metal with the oil□
AS PER HELPDESK CALL AUPS02814. E-MAILS BETWEEN ANDY CRIPPS AND DAVIDCHUNG ON SEPTEMBER 13 2012. SUBJECT: TB2 13-2010817 CD40S-5 FDB02-1520-
As per e-mail corrospondece between YC Kim and Rod MB on the 14th June2012. Subject:AUPS02471 CG55C-5 FGB05-1410-00012 - Noisy Tranmission□
As per e-mail corrospondence between David Chung and Rod M. on the 07/02/2013. Subject: AUPS03071 CG40S-5 FGB02-1410-00016 - Transmission Failure
FAULTY TRANSMISSION
UNIT CAN'T MOVE CAUSE THE INNER PART TRANSMISSION WAS BROKEN□
TRAVELED TO LOCATION; FOUND WET AROUND BELL HOUSING PLUG COMING DOWN ALSO AROUND SIDE OF STARTER; PULLED PLUG OUT OF BELL HOUSING; FL
CRACKED TRANSMISSION HOUSING
Transmission used at assembly has mismatched bolt thread patterns between the halves,□esulting in insufficient sealing in short term after the machine is used, tra

Text 분류

- Field Claim classification
  - To determine the cause of the problem
  - To determine who is to blame

Supplier	Buyer
•Wants to sell commodity	•Wants to buy commodity
•Does not want to divulge excessive information	•Wants lots of data to assess risks

정서 분석

- 제품 리뷰
- "The TV is wonderful. Great size, great picture, easy interface. It makes a cute little song when you boot it up and when you shut it off. I just want to point out that the 43" does not in fact play videos from the USB. This is really annoying because that was one of the major perks I wanted from a new TV."

From "Text Mining and Analysis", SAS institute

Object 와 **attribute**

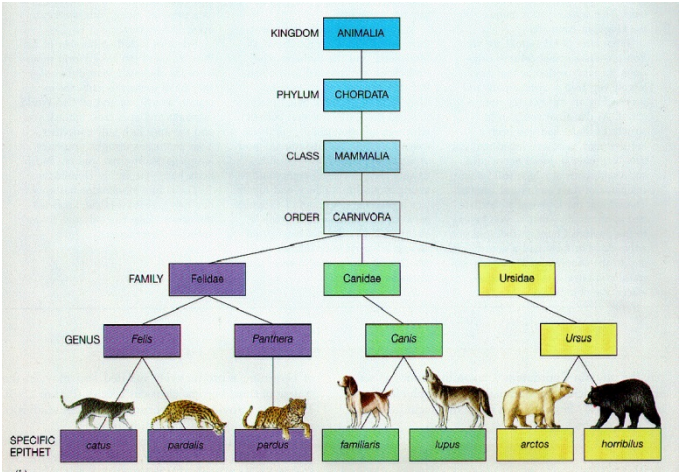
- The **TV** is wonderful. Great **size**, great **picture**, easy **interface**. It makes a cute little song when you **boot** it up and when you shut it off. I just want to point out that the **43”** does not in fact play videos from the USB. This is really annoying because that was one of the major perks I wanted from a new **TV.**

Positive 와 **negative**

- The TV is **wonderful**. **Great** size, **great** picture, **easy** interface. It makes a **cute** little song when you boot it up and when you shut it off. I just want to point out that the 43" **does not** in fact play videos from the USB. This is really **annoying** because that was one of the major perks I wanted from a new TV.

정서 분석

- Document 수준: 5 Pos and 2 Neg
- Sentence 수준: P, P, P, N, N
- Object / feature 수준
  - Use of Taxonomy



정서분석의 어려운 점

- NLP 수행
  - POS tagging, disambiguating terms and lexicons, spelling error correction
- 컨텍스트에 따라 다른 의미
  - “The size seems small”, USB vs TV
- 장문
  - Blog postings harder than tweets or reviews