Data Mining Introduction

“(　　)은 절하고, (　　)은 악수한다”

“(　　)은 절하고, (　　)은 악수한다”?

귀납적 추론

- 과거사례/데이터로부터 명제나 패턴 추출
- 인간 생존을 위한 가장 중요한 지적 활동

귀납적 추론

- 내일 아침에도 해가 뜬다.
- 눈이 오면 길이 미끄럽다.
- 비가 오면 젖는다.

귀납적 추론

- “미국인들은 부유하다.”

귀납적 추론

- “부유한” 미국인?

귀납적 추론

- "일본인들은 날씬하다."

귀납적 추론

- “날씬한” 일본인?

귀납적 추론

- 귀납적 추론의 정확도는
  - 과거 사례의 수가 충분한가?
  - 과러 사례의 질이 좋은가?

데이터마이닝

- 정의:
  - 대규모 데이터에 대한 귀납적 추론

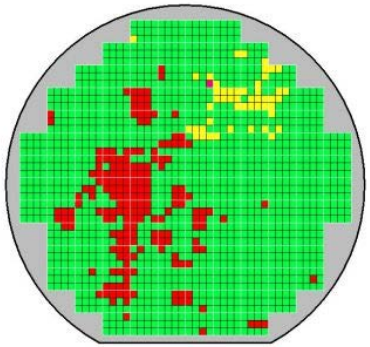
- 대규모 데이터의 탐색 분석을 통하여 의미 있는 명제나 패턴을 찾는 과정

다른 이름

- 애널리틱스
- Advanced 애널리틱스
- Predictive 애널리틱스
- 기계학습 (Machine Learning)
- 패턴인식 (Pattern Recognition)

애널리틱스 분류

- 3 가지 데이터



숫자 애널리틱스

- 자연 현상 IoT 데이터
  - 지진, 기상, 토질
- 자동차, 장비, 공정에서 발생하는 IoT
- 인간의 행동 데이터
  - 이동, 구매, 시선, 뇌파

텍스트 애널리틱스

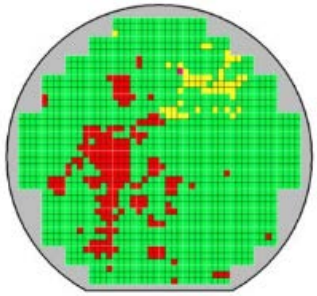
- 제품/서비스에 대한 AS 내역은?
- 공정 엔지니어의 검사/이상상황 코멘트에 대한 체계적인 이해
- 소셜 미디어에 표출된 소비자 반응
- 소비자가 원하는 제품/서비스는?

이미지 애널리틱스

- **Object recognition**
  - 테러리스트 인식

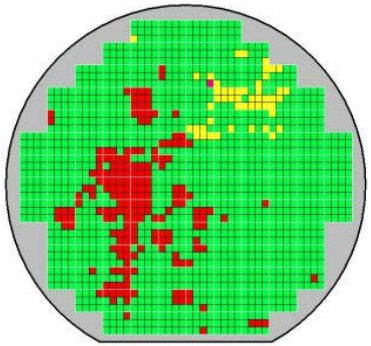
- **Image 자동 태깅 및 검색**

- **Image classification and clustering**
  - 불량 분류



데이터마이닝


enterprise
infrastructure
technology
operations
information
scorecards
objectives
capitalize
analyze
text mining
manage
metrics
applications
connection
technical
solution
stakeholder



애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

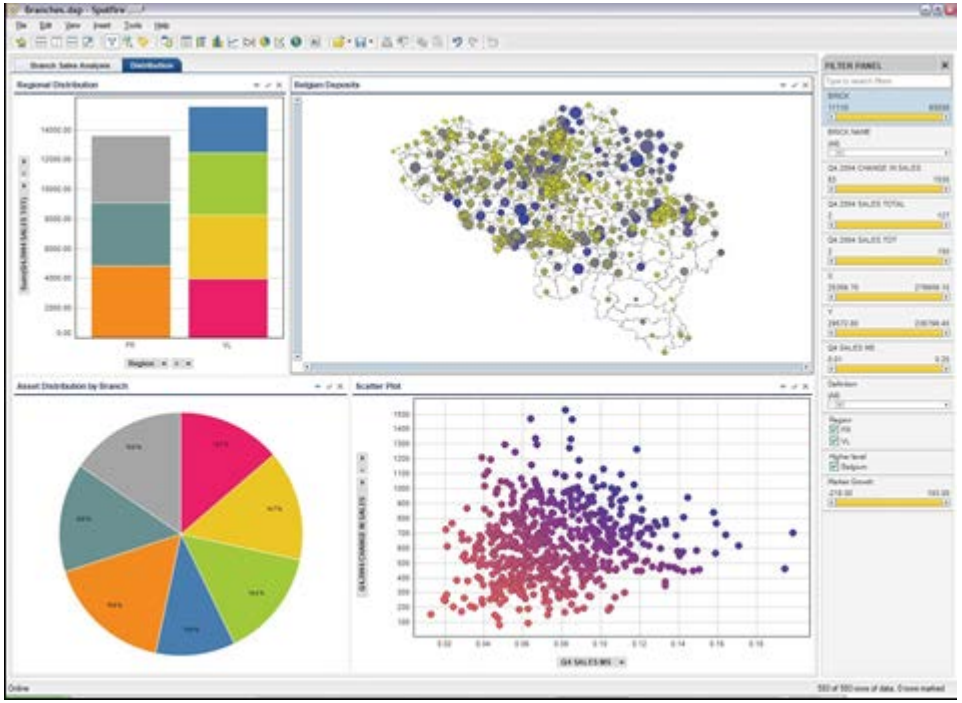
애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

자동차 Dashboard

electronics.howstuffworks.com

BI “dashboard”



bitools.org

BI “dashboard”

- 마케팅 “이메일 캠페인 효과 측정”
- 의사결정 변수
  - 채널, 메시지, 가격, 날짜, 시간, 고객, 상품
- 효과
  - 반응률 및 수익률

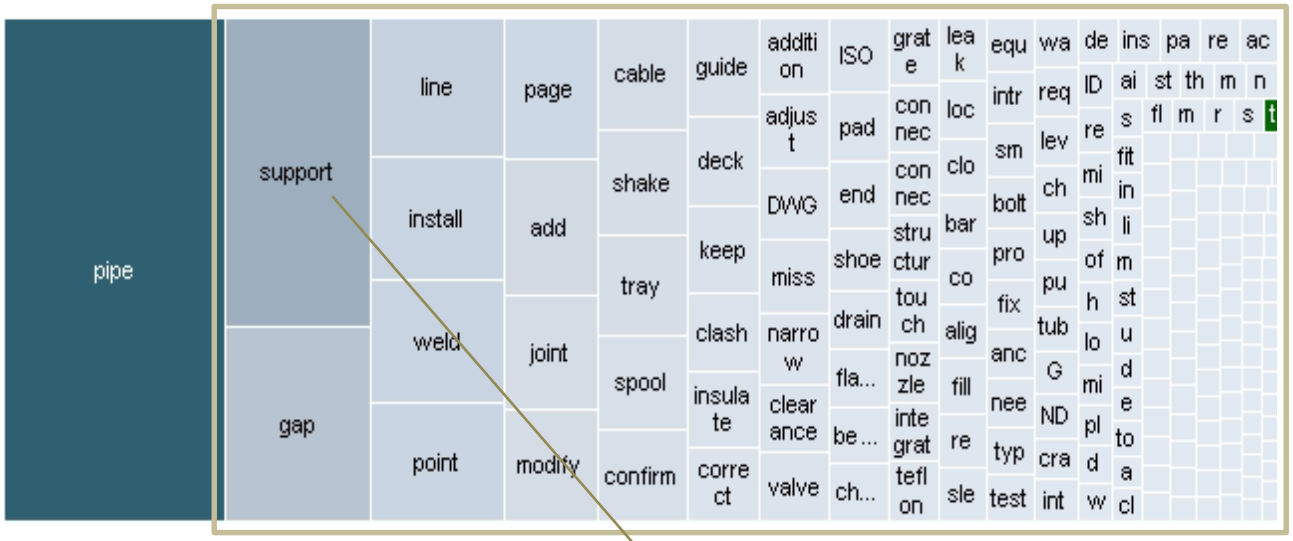
Stephen Yu, Computer World

BI “dashboard”

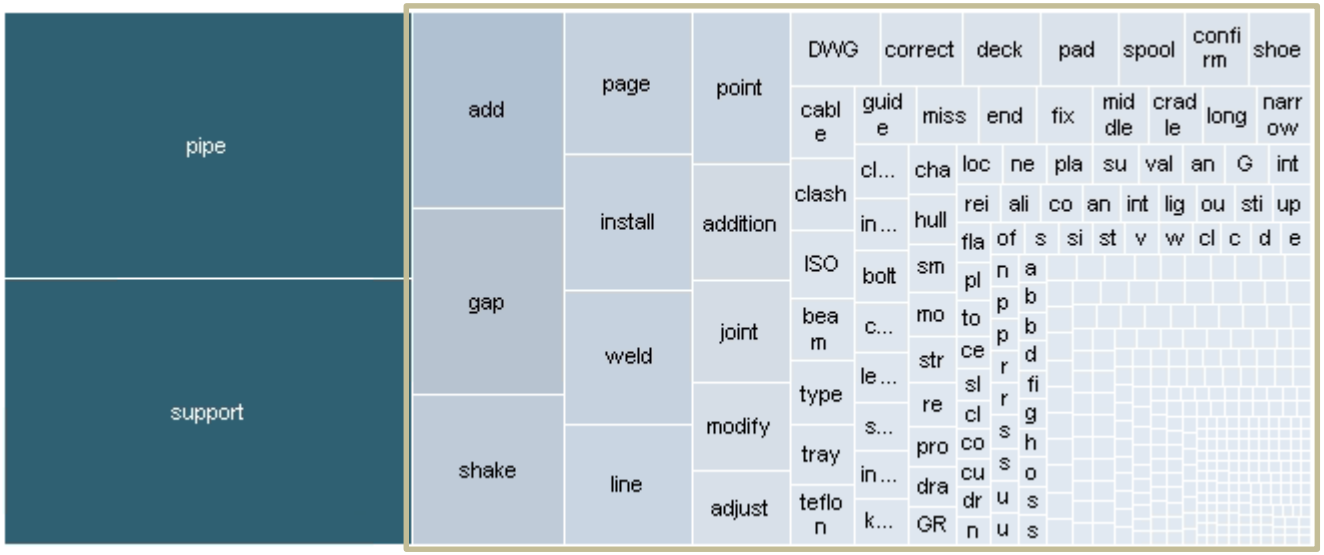
- 제조공정 “공정 파라미터 효과 측정”
- 의사결정 변수
  - 온도, 속도, 압력, 농도, 장력
- 효과
  - 불량률 및 품질수준

문제들 간에 어떤 관계가 있는가?

"pipe"와 함께 등장하는 단어

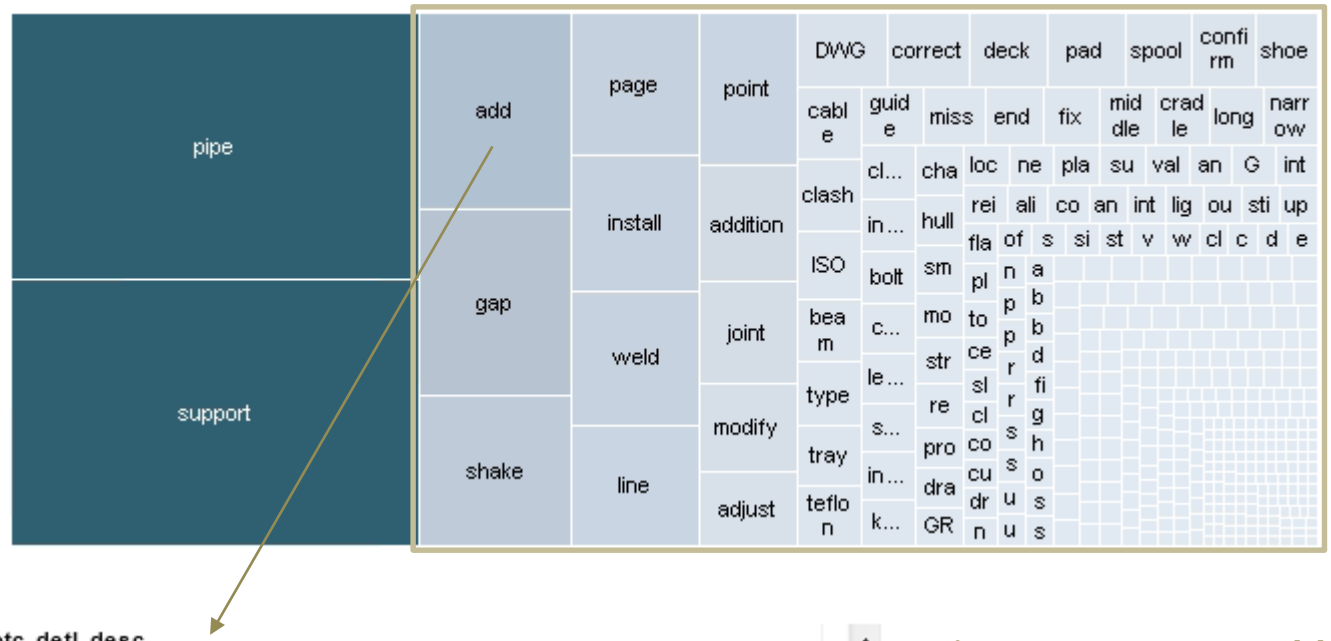


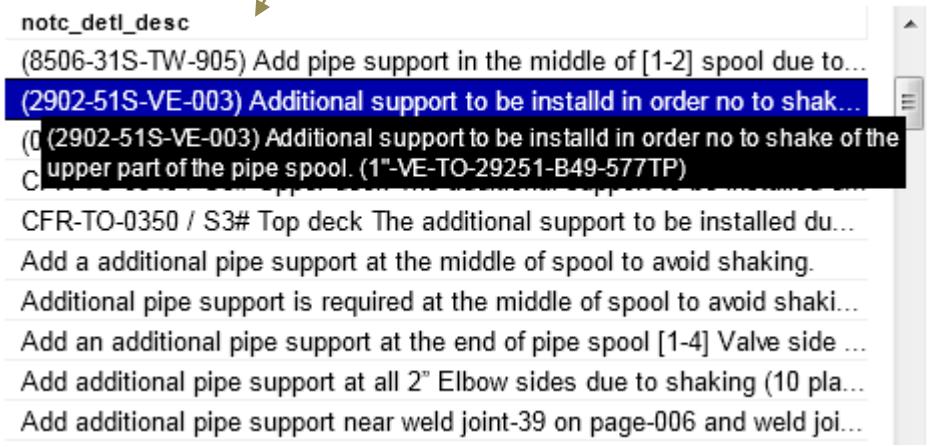
"pipe", "support"와 함께 등장하는 단어



문제들 간에 어떤 관계가 있는가?

"pipe", "support"와 함께 등장하는 단어





"pipe", "support", "add"가 함께 등장하는 검사문서

"shake"도 함께 등장함

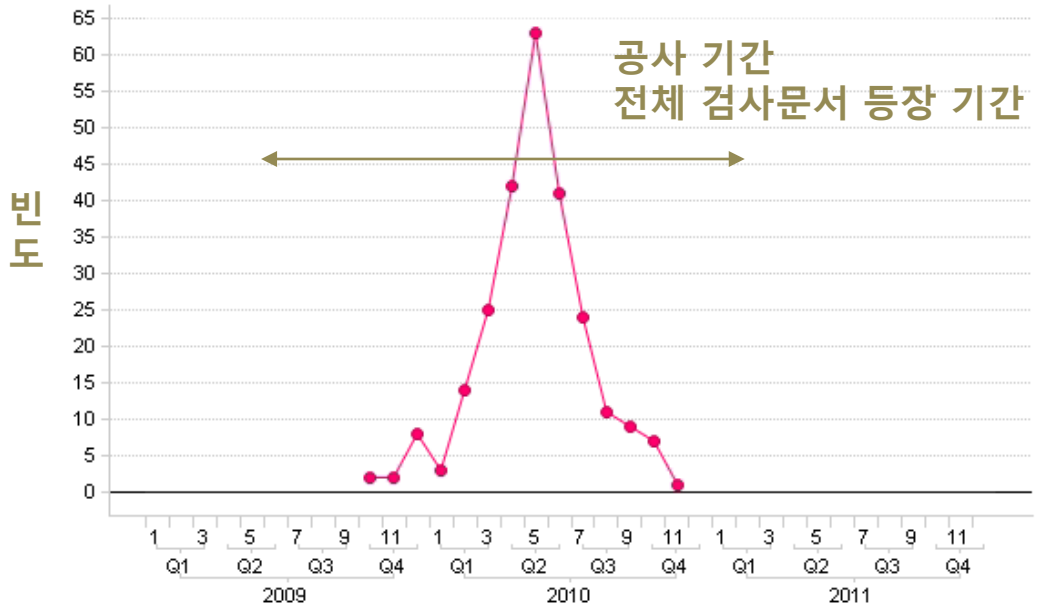
→"파이프(pipe)가 흔들리니(shake),
지주(support)를 추가(add)하라"

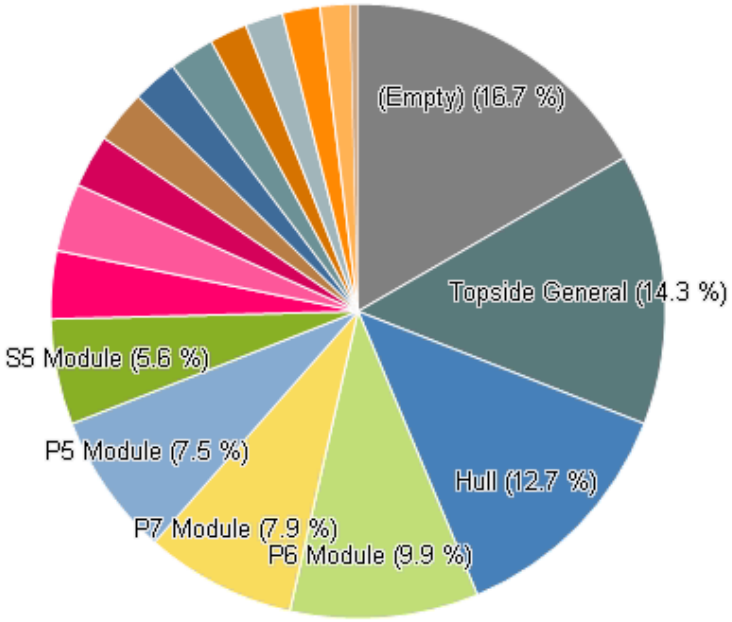
문제들 간에 어떤 관계가 있는가?

"파이프(pipe)가 흔들리니(shake), 지주(support)를 추가(add)하라"

언제, 어디서?

"공사 중반, Topside General, Hull, P5,6,7 Module 등에서 주로 발생한다"





발생하는 장소

Year	Common Concern	Federal Reserve System	European Central Bank	Bank of England	Deutsche Bundesbank	Bank of Japan
2004	Sustainability Credibility	Corporate Gover-nance Scandals	Parliaments Growth and Job	Low Inflation	Government-Deficit	QE Deflation
2005	China Inflation	Oil/Natural Gas Basel II	Domestic Inflat-ionary Pressure	Repo Rate Crystallising	Debt Levels	Private-Consumption
2006	Competitive Global Imbalance	Risk Management Creditworthiness	Monetary-Expansion	Exchange Rates China / India	Future Inflation Two Pillar Strategy	Domestic and-External Demand
2007	Subprime-Mortgage	Foreclosures	Risk to Price Stability	Growth of Money and Credit	Local Currency Bond Market	Price Stability
2008	Financial Turmoil Commodity Prices	Primary Dealers Foreclosures	Price Stability Supply of Liquidity	Failing Banks Spare Capacity	Resilience of-Financial System	Securitized-Product
2009	Financial Crisis Lehman Brothers	TALF SCAP	Non-standard Macroprudential	Asset Purchase	Expansionary-Monetary Policy	Outright Purchase Credit Bubble
2010	Recovery Reform	Unemployed SCAP	Macroprudential Excessive Deficit	VAT Depreciation of £	Microprudential Reform of Basel II	Overcoming-Deflation
2011	Sovereign Debt Basel III	Job Growth Dodd-Franc Act	Economic Governance	Real Incomes PRA	No Bail Shadow Banking	After Earthquake Monetary Easing
2012	Europe Deleveraging	Maturity Extension Forward Guidance	OMT / SSM Fragmentation	FLS	Liability Rescue Package	European Debt
2013	Real Economy Price Stability	(At least as long as) Unemployment	Fragmentation SSM / SRM	FLS PRA	Liability SSM	QQE

애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

연관분석
Association Mining

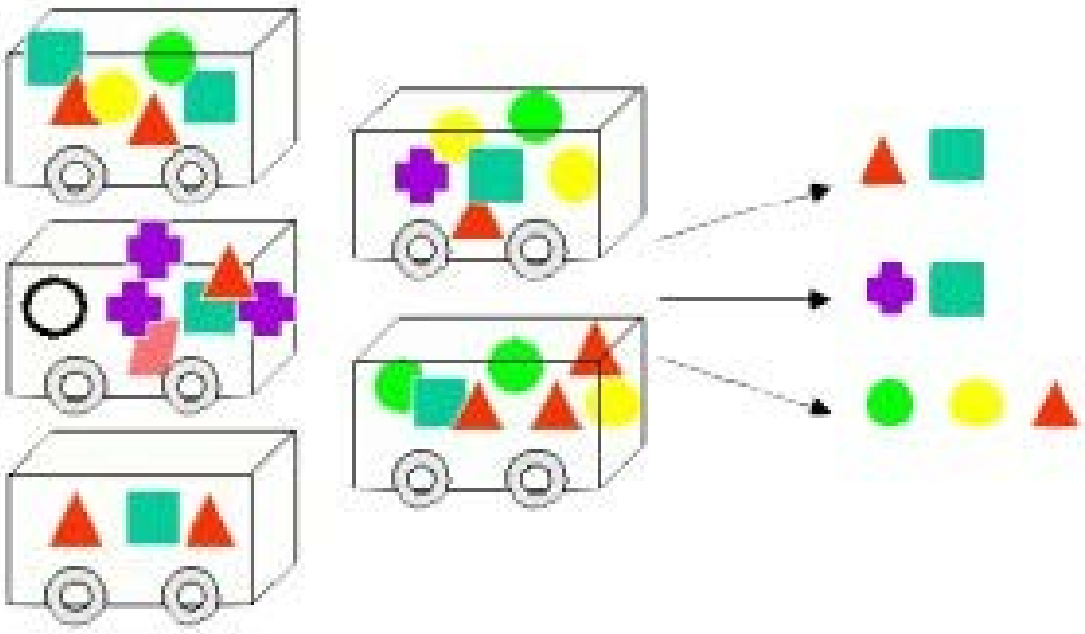
- 슈퍼에서 함께 팔리는 물품은 ?
- 영화 A를 본 사람은 영화 B를 본다 ?
- 74번 공정을 거치는 블록은 35번 공정을 거친다.

연관분석
Association Mining

- 동시에 발생하는 사건/조건 파악 또는
- 동시에 구매되는 물품 파악

- 방법론
  - A Priori algorithm

연관분석 결과



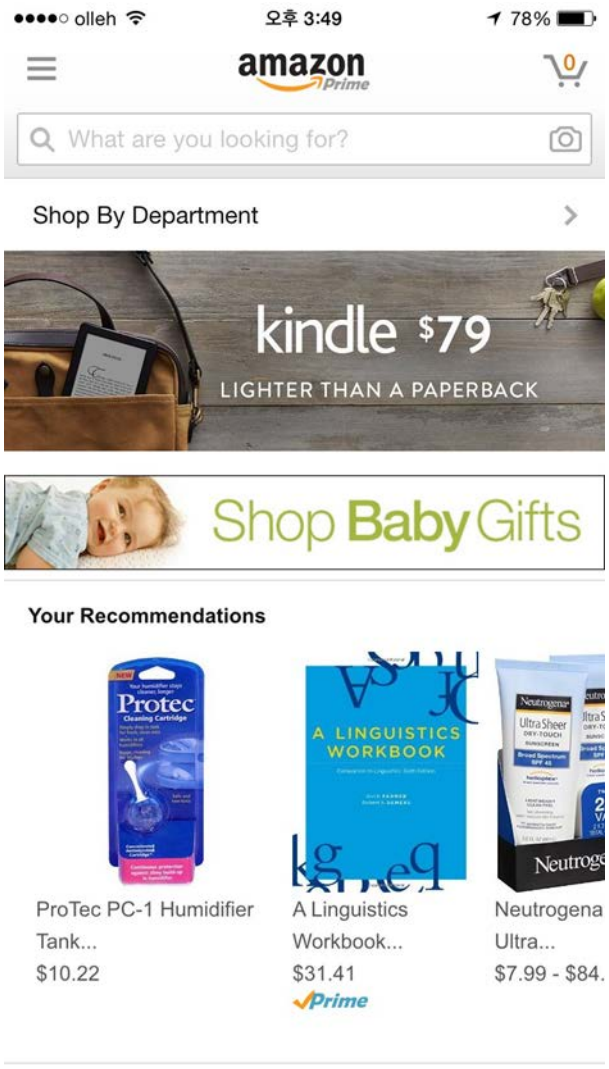
http://gerardnico.com/wiki/data_mining/association

연관분석 결과

Association Rules	Support Count of Antecedent	Support Count of Rule	Support of Rule	Confidence of Rule
Diaper → Beer	4	3	$3/5 = 0.6$	$3/4 = 0.75$
{Milk, Diaper} → Beer	3	2	$2/5 = 0.4$	$2/3 = 0.67$
Bread → Milk	4	3	$3/5 = 0.6$	$3/4 = 0.75$
{Bread, Milk} → Diaper	3	2	$3/5 = 0.6$	$2/3 = 0.67$
{Bread, Milk} → Coke	3	1	$3/5 = 0.6$	$1/3 = 0.33$

http://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter002/section001/blue/page002.html

연관분석을 통한 추천


olleh
오후 3:49
78%
amazon
Prime
What are you looking for?
Shop By Department
kindle $79
LIGHTER THAN A PAPERBACK
Shop Baby Gifts
Your Recommendations
Protec
A LINGUISTICS WORKBOOK
Neutrogena
Ultra Sheer
ProTec PC-1 Humidifier Tank...
$10.22
A Linguistics Workbook...
$31.41
Prime
Neutrogena Ultra...
$7.99 - $84.


Digits
Tech News & Analysis From the WSJ
COMPANIES
MOBILE
PRIVACY
SOCIAL MEDIA
HOT TOPICS: APPLE TWITTER BITCOIN NSA SURVEILLANCE MICROSOFT CEO SEARCH
AMAZON
Amazon Wants to Ship Your Package Before You Buy It

애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

군집화
Clustering

- 고객들을 유사한 사람들끼리 군집화, 타겟 마케팅
- 유사한 주식 시장 상황끼리 군집화
- 사용 패턴이 유사한 에어컨 실내기 들을 군집화

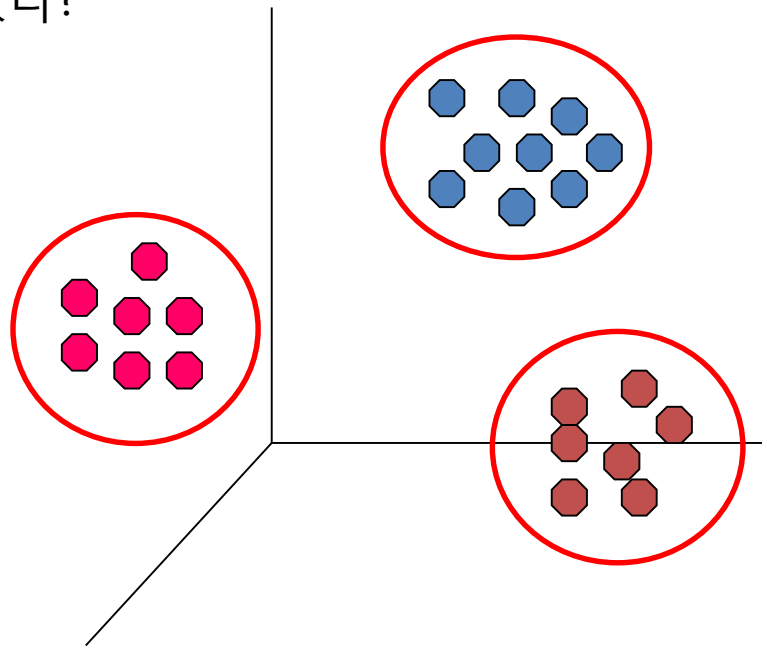
군집화
Clustering

- 유사 데이터들의 군집화
- Unsupervised, Exploratory Knowledge Discovery

- 방법론
- K-means, Agglomerative, Competitive Learning

군집화

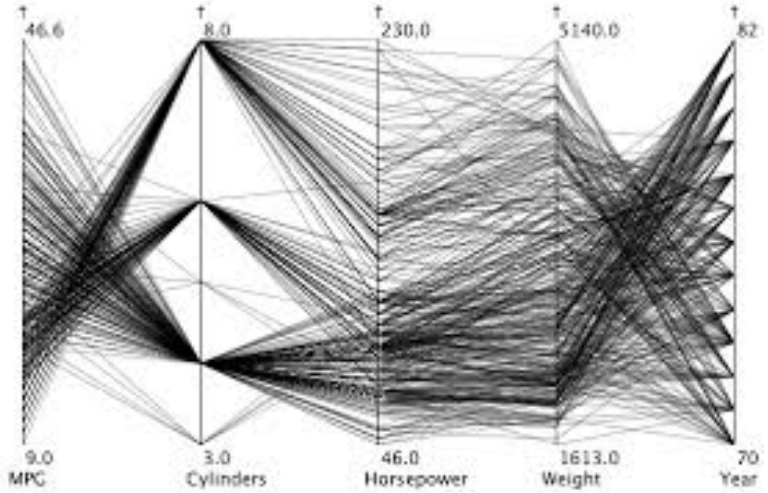
Recency 구입한지 얼마나 지났나?

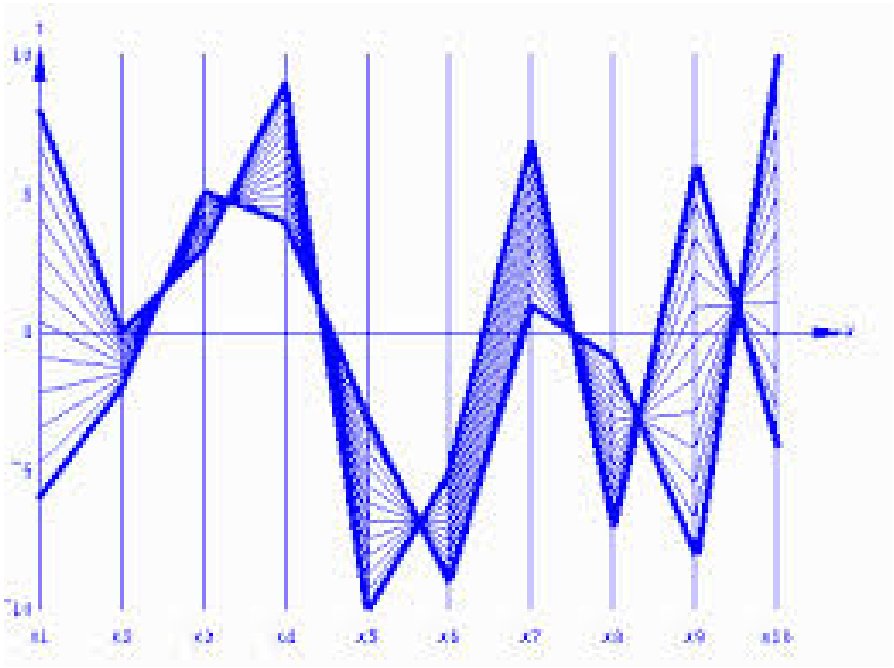


Frequency 최근 1년간 몇 번 구매?

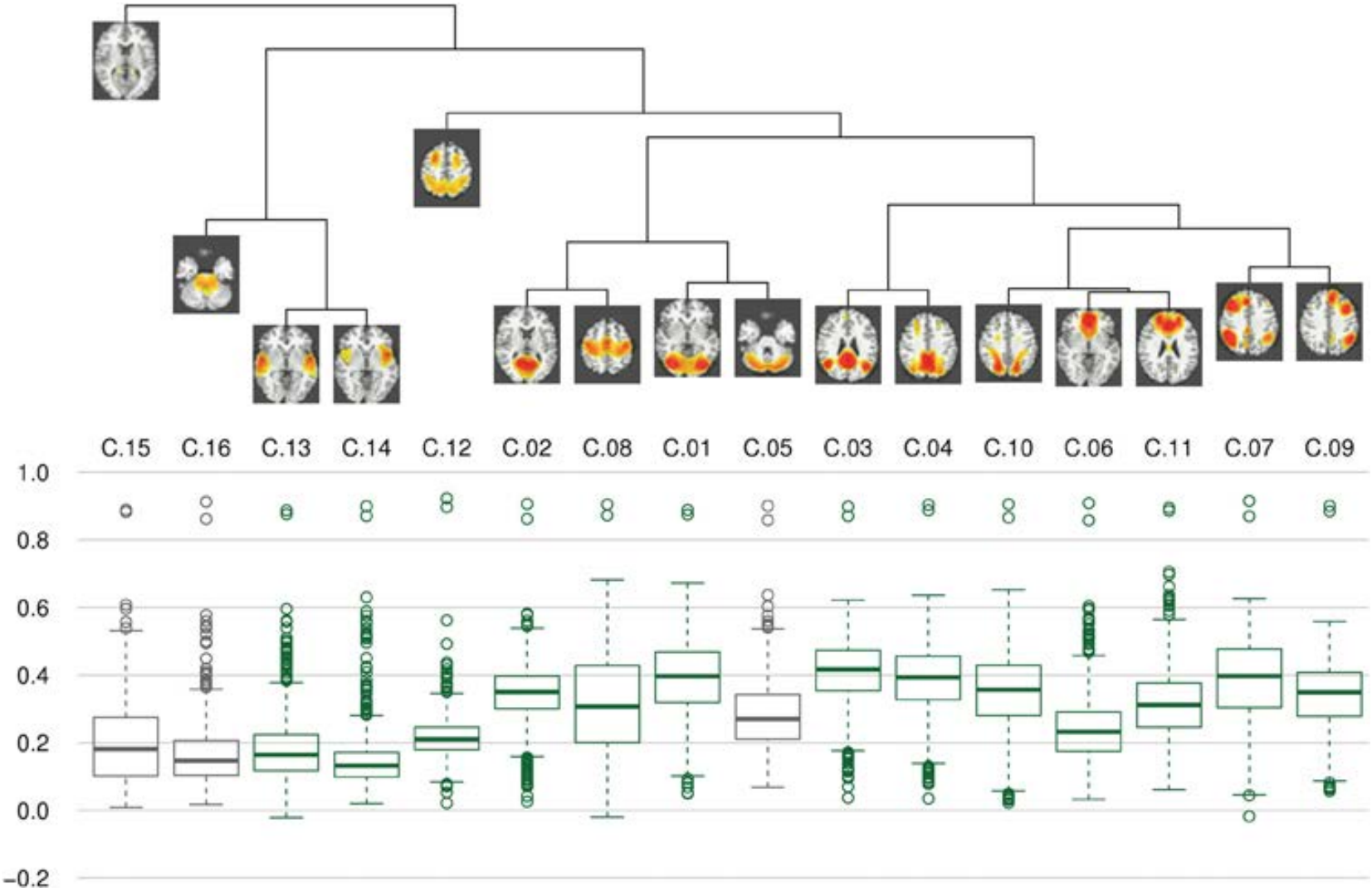
Monetary 최근 1년간 총 구매액?

다차원?


46.6
8.0
230.0
5140.0
82
9.0
MPG
3.0
Cylinders
46.0
Horsepower
1613.0
Weight
70
Year



군집화 결과



http://journal.frontiersin.org/Journal/10.3389/fnhum.2012.00301/full

군집 특징 파악

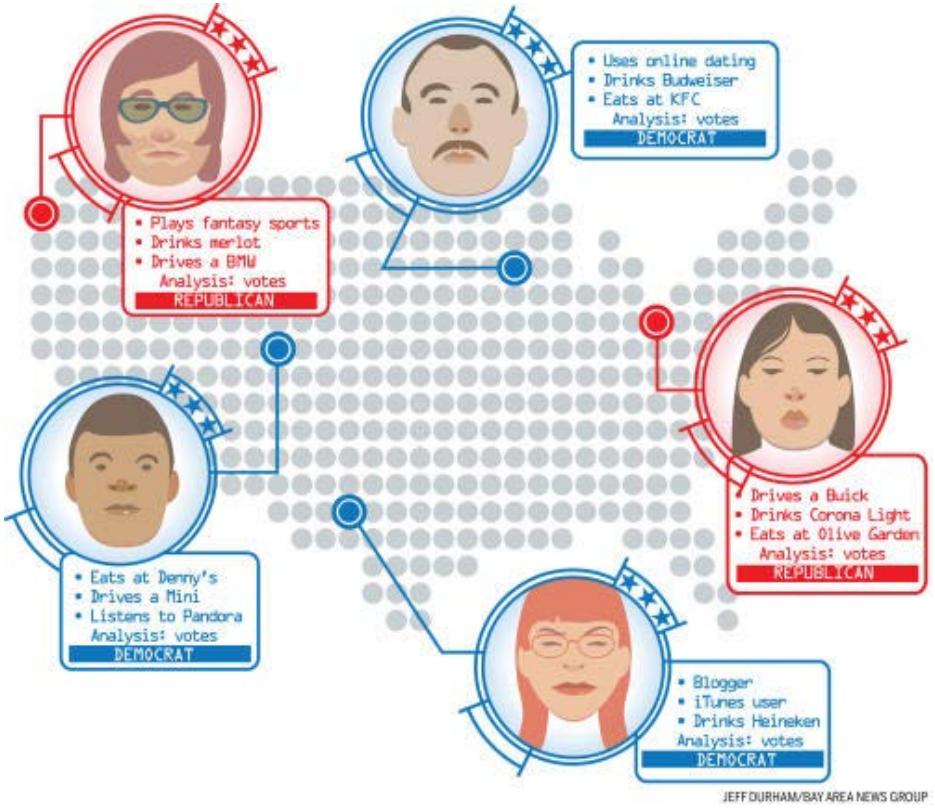
- 마케팅
- "주로 30~40대 여성이며 학부형일 가능성이 많으며 그들의 주 거주지는 청담 지역이며 생활수준이 매우 높으며, 피트니스와 맛집 탐방의 라이프 스타일"
- 식으로 우량 고객을 묘사

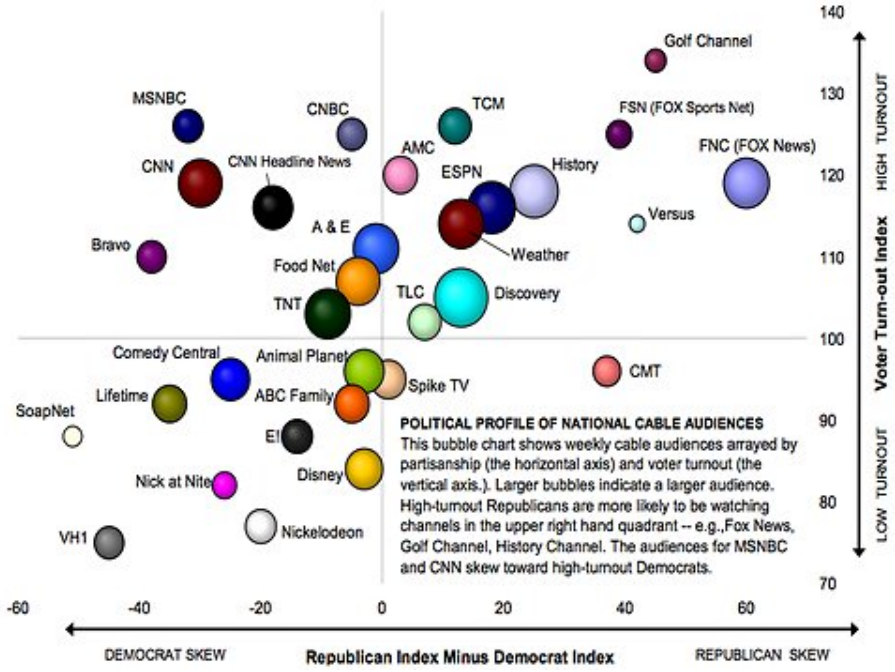
Stephen Yu, Computer World

Other “customers”?

- Voters
- Employees
- Central Bankers
- People

유권자





www.mercurynews.com

campaignstops.blogs.nytimes.com

애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

Predictive analytics

- 마케팅
  - 특정 고객의 미래의 특정 행동 가능성을 통계적 확률 또는 점수 (score) 로 추정
- 제조
  - 특정 lot 가 경험한 현재 제조 환경으로부터 미래의 품질 수준을 확률 또는 점수 (score) 로 추정

Stephen Yu, Computer World

관심 event 또는 target

- 마케팅의 경우
  - 어떤 물품이나 서비스의 사용자와 비사용자,
  - 캠페인에 반응한 사람과 무관심한 사람,
  - 크레딧을 줄 만한 사람과 그렇지 않은 사람,
  - 고정 고객과 단발성 고객,
  - 높은 수익률을 주는 고객과 그렇지 않은 고객,
  - 자주 찾는 고객과 가끔씩만 들르는 고객

Stephen Yu, Computer World

관심 event 또는 target

- 제조/품질의 경우
  - Lot 단위의 정상과 불량
  - 선박 건설 소요 기간
  - 최적 sinter belt 스피드
  - 최적 Roll Force
  - 필드 클레임 불량

Stephen Yu, Computer World

관심 event 또는 target

- 1 단계: 타겟에 대한 수학적 정의
- "돈을 많이 쓰는 고객"의 수학적 의미
  - 얼마 이상? 5백만 원?
  - 어느 정도의 기간을 고려하는가?
  - 어느 채널로?
  - 한 번에? 아니면 1년 동안? 외상은?

Stephen Yu, Computer World

- 우량 고객 30만 가운데 4,887명 플래티넘 카드 사용자
- 나머지 295,123 명 비 사용자 가운데 누구를 타켓팅 할 것인가?

관심 event 또는 target

- 2단계: 타깃과 타깃이 아닌 두 집단의 차이를 변별하는 변수를 찾는 작업
- 빅데이터는 수천 개의 변수 고려
- 이 모든 과정이 수학적으로 이루어짐
  - 직관적으로 오랫동안 사용해온 변수들도 유용하지 않다면 버려지고
  - 많은 변수들은 합쳐지기도 하고, 숫자의 경우 그룹으로 나누어지던가 공식을 통한 변환을 겪기도 함

Stephen Yu, Computer World

- ***특급호텔*** 11만원 이상 & ***항공사*** 이용
  - 787명 (Platinum 93.1%)
- ***골프장*** 48만원 이상 & ***일식*** 10만원 이상 & ***항공사*** 이용 안 함
  - 151명 (Platinum 92.7%)
- ***골프장*** 7만원 이상 & ***일식*** 24만원 미만 & ***특급호텔*** 11만원 미만 & ***항공사*** 이용
  - 90명 (Platinum 93.3%)

관심 event 또는 target

- **“알고 있는 확실한 정보”로 “지금은 알 수 없는 대상”을 예측**

Stephen Yu, Computer World

애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

예측/분류
Prediction /Classification

- 특정 고객이 마케팅 켐페인에 반응할 확률
- 휴대폰 고객이 향후 6개월 내에 이탈할 여부
- 다음 주 주가 상승 여부 = f(최근 주가 추이, 경제환경)
- 숙성 중인 와인의 품질 예측
- 제조 중인 반도체 웨이퍼의 수율 예측
- 어느 지원자를 선발해야 들어와서 일 잘할까?

예측/분류

- 종속변수(y)를 독립 변수 (x) 들의 함수 (f) 로 적합,
- 즉 y = f (x)
- **방법론: 회귀분석**, 신경회로망, 사례기반 추론, **의사결정나무**


WAL★MART

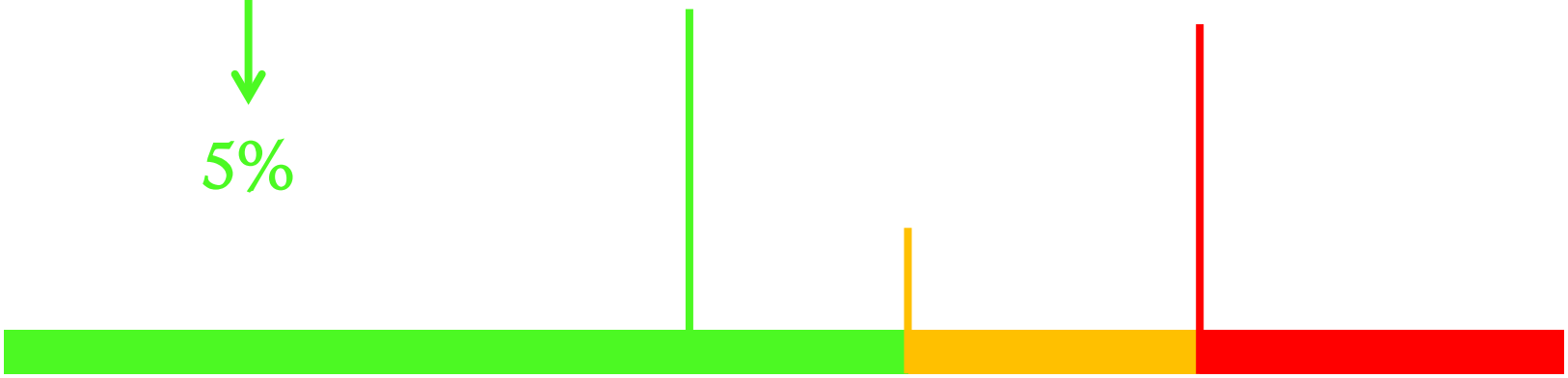

Kellogg's
pop-tarts
Blueberry
7 Vitamins and Minerals

cablecom®

20%

7개월 차
고객관리!

5%



ROYAL SHAKESPEARE COMPANY

관객수 70% 증가

애널리틱스

- 인사이트 (Descriptive)
  - Business Intelligence
  - 연관분석
  - 클러스터링
- 포사이트 (Predictive)
  - 예측/분류
  - 이상탐지

이상 탐지
Anomaly Detection

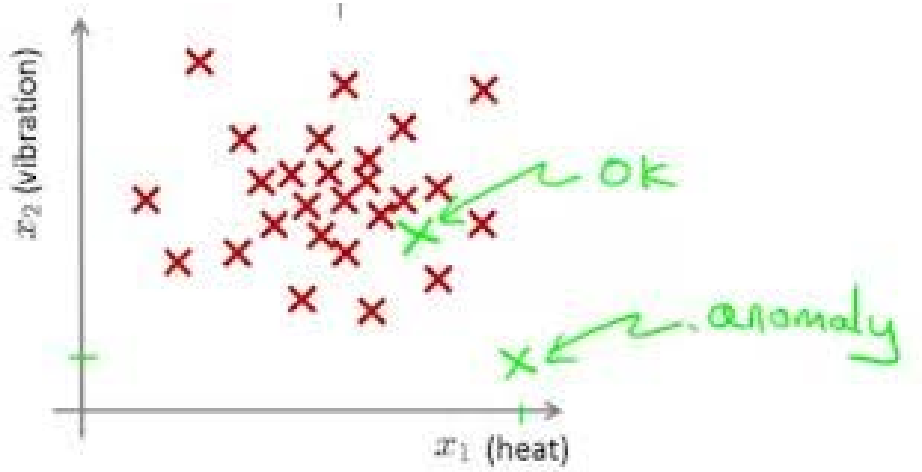
- 보험, 세금, 신용카드 사용 사기 탐지
- 반도체 공정의 이상 상황 예지
- 헬스 모니터링에서 이상 상황 예지

이상 탐지
Anomaly Detection

- 정상적인 상황을 학습하여, "어디까지가 정상인지" 이해한 후,
- 비정상적인 상황이 발생하였을 때에, 반응함.
- 정상 상태 데이터는 풍부하나, 비정상 상태 데이터가 아주 부족하거나 없음

- **방법론**
- **Unsupervised Learning, Normal boundary Learning, Distance based Model**

이상 탐지 결과



http://dnene.bitbucket.org/docs/mlclass-notes/lecture16.html

프로세스

Diamond 마이닝

Diamond 프로세싱

Diamond 디자인

Data

- 마이닝
  - 데이터의 산에서 패턴, 명제, 수식 발췌
- 프로세싱
  - 비즈니스 문제와의 연관성 파악 및 검증
- 디자인
  - 비즈니스 문제에 대한 솔루션으로 변환

Data

- 여기에서는 역순으로

- 디자인: 비즈니스 문제 정립, 적용 및 평가
- 프로세스: 인사이트, 포사이트 도출
- 마이닝: 전처리, 모델링

5 단계 문제 해결 절차

1. 비즈니스 문제 정립
2. 분석 계획 도출
3. 데이터 확보
4. Insight / Foresight 도출
5. 비즈니스에 적용 및 평가

어떻게 시작하지?

옳은 질문하기

- 데이터를 가지고 무엇을 할 것인가?
- 데이터 생성,저장,처리 어떻게 할 것인가?

필수 질문: 유형 1

- "어떤 의사결정에 사용하지?"
  - 우리 사회 트렌드 변화는?
  - 소비자가 원하는 제품/서비스는?
  - 어느 고객이 이탈할 것인가?
  - 어떤 고객이 이 제품을 구매할까?
  - 공정 데이터로부터 불량 탐지를 어떻게?
  - 어느 지원자를 선발해야 하는가?
- Return 은 얼마나?

필수 질문: 유형 2

- 데이터 분석을 어떻게 하지?
  - 쏟아져 나오는 Data의 저장은 어디에?
  - 실시간 의사결정을 어떻게?
  - 데이터 통합은 어떻게?
  - 누가 분석하지?
  - 분석가들에게 어떤 SW package를?
  - 외부 데이터는 어떻게 확보하지?
- 인프라 Invest 를 얼마나 하지?

질문의 순서

Return 먼저

Invest 나중에

질문의 순서

Why 먼저

How 나중에

R first, I later

- Proof of Concept
- Why?