Big Data Analytics

Sungzoon Cho
Seoul National University

- **Volume** Lots of Data
- **Velocity** Stream Data
- **Variety** Text, Image

- Where?
  - **Internet of Things (IoT)**
  - **Bring your own Data (BYOD)**

Big Data

amazon.com

Walmart
Save money. Live better.

Big Data

Big Data

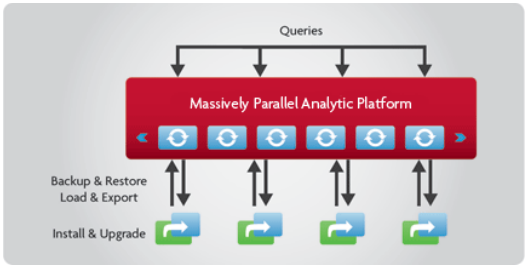
Big Data

Big Data

Big Data

Why Big Data

- Distributed Database
- Parallel Processing



Why Big Data

Decision Making

- Based on Insight & Foresight
- Evidence based

Is that new?

- "stat"istics = state
  - "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,"

State is Data

Science is Data

- Electronic microscope => Nano Science

$$ is Data

- Mayer Rothschild in 19th century
  - Amschel (Frankfurt), Salomon (Vienna), Nathan (London), Calmann (Naples), Jakob (Parks)

Management is Data

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

  – Peter Drucker

Big Data = Statistics?

- Yes and No
- Statistics
  - Hypothesis => Data => Test
- Big Data
  - Data => Hypothesis/Knowledge => Test
  - Knowledge Discovery

Asking Right questions

- What to do with Big Data?
- How to deal with Big Data?

Questions of "what"

- "How to use it for decision making?"
  - Change in Society Trend?
  - What products do consumers want now?
  - Which customer will churn?
  - Who is going to purchase?
  - How to predict defects from process data?
  - Who should we hire?
- Return?

Questions of “how”

- How to deal with Big Data?
  - How to manage incoming stream of Data?
  - How to make realtime decisions?
  - Which SW packages to use?
  - Who is going to analyze data?

- Infrastructure Invest?

Which questions to ask first?

Return

Invest

Three kinds of data

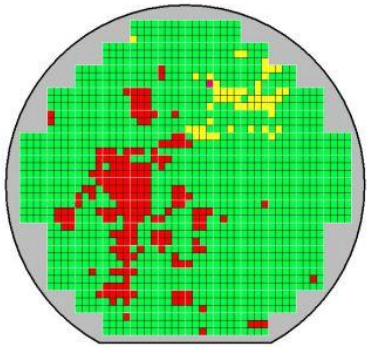
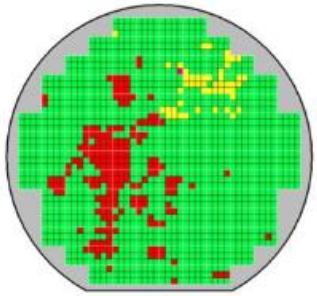


Image mining

- **Object recognition**

- **Image automatic tagging**

- **Image classification and clustering**



Text mining

- AS or field claim data?
- Test comment or anomaly condition comment in engineering process
- Voters' voice from Social Media
- Understanding products/functions consumers want, from Social Media

Text Mining

- Text is “Unstructured” Data
- **Natural Language Processing (NLP)**

NLP pre-processing

- Filtering
- Tokenization
- Stop-word removal
- Stemming
- Pruning
- Vectorization

1. Tokenization

Before

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.

After

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

Before

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

After

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

Before(After)

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)** vector space model
  - TF (Term Frequency):
  - IDF (Inverse Document Frequency):

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	

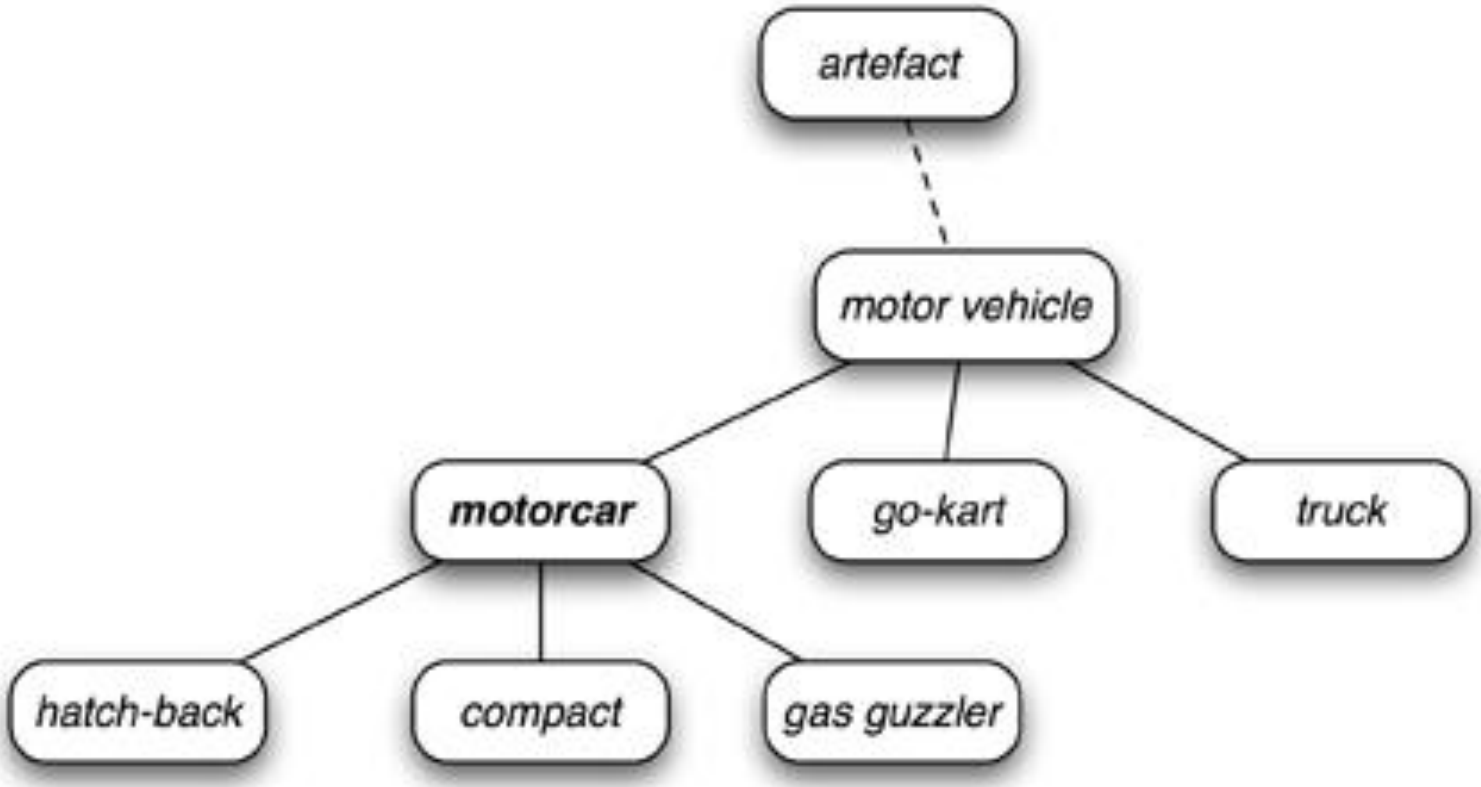
4. Vectorization

- Dictionary of terms, predefined in advance
  - For each domain: Sports, Electronics, Politics
  - Time consuming

- Unstructured Text into Structured vectors

  => Data Mining applied

WORDNET "artefact"



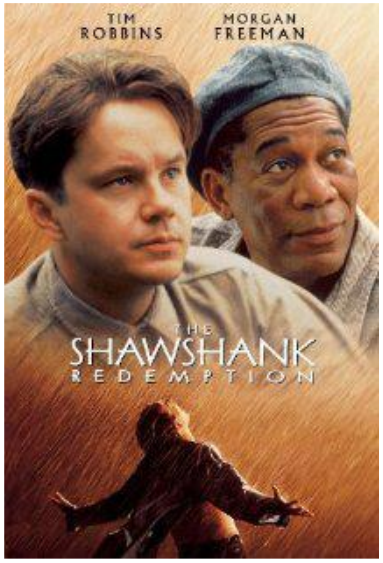
Big Data + Text

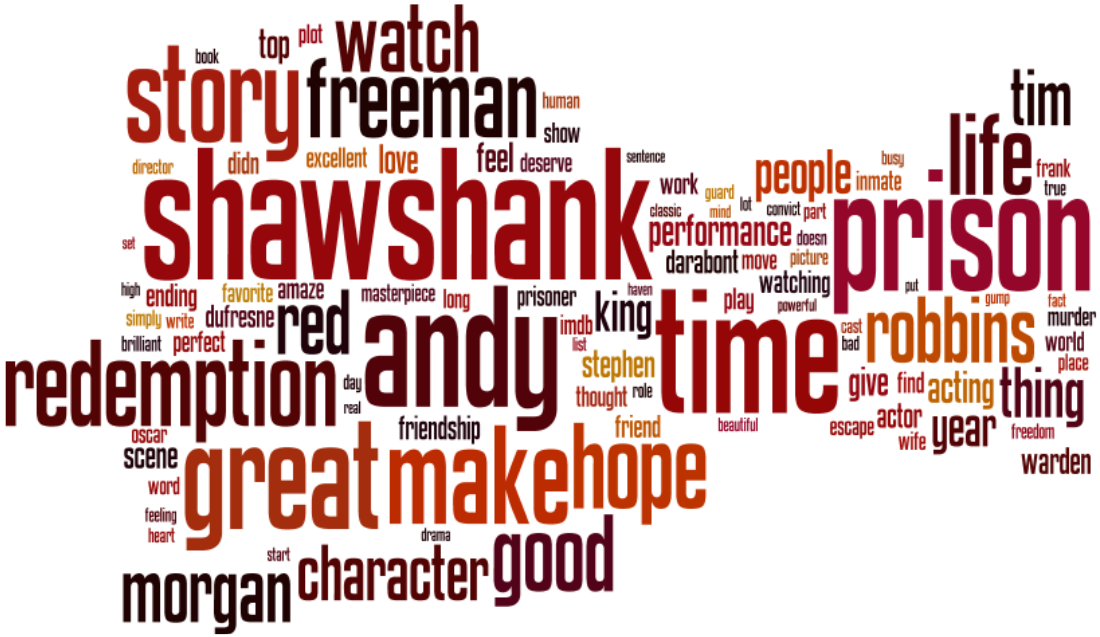
- Information Retrieval (정보검색)

Big Data + Text

- Topic Extraction
  - Understanding of doc topics
  - Clustering docs based on topics covered
- Classification into classes
  - Automatic News feed
  - Spam filter
- Sentiment Analysis
  - Polarity of doc

Topic Extraction





Text Classification

- Spam Filter

Text Classification

- 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

Text Classification

- 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, resulting in insufficient sealing in short term after the machine is used, tra

Sentiment Analysis

- *" 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 and **attribute**

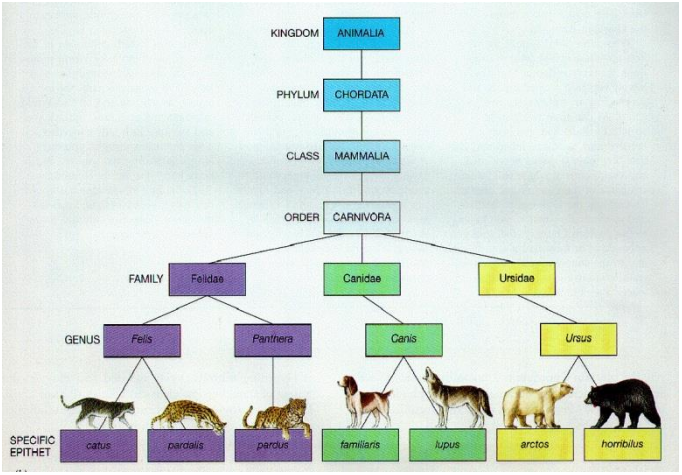
- *The **TV** is wonderful. Great **size**, great **picture**, easy **interface**. It makes a cute little song when you **boot** it up and when you shut it off. I just want to point out that the **43"** does not in fact play videos from the USB. This is really annoying because that was one of the major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

Positive and **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.*

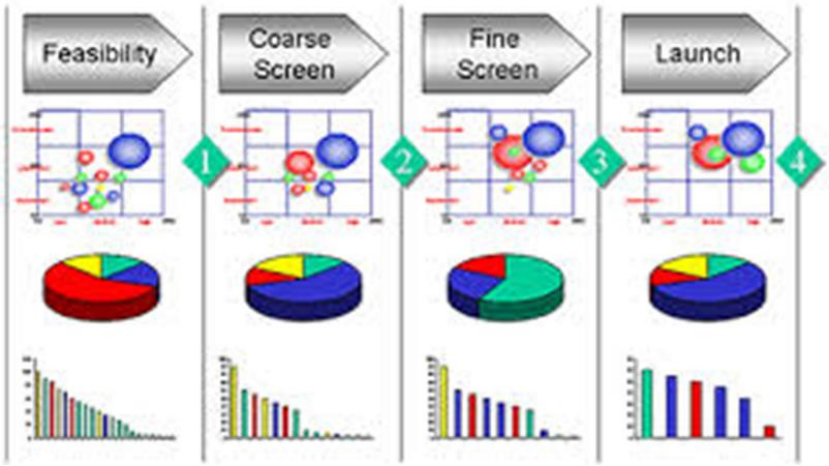
Sentiment Analysis

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



Difficulty of Sentiment Analysis

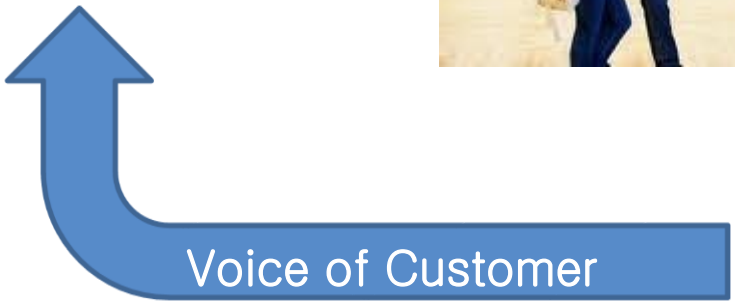
- NLP issues
  - POS tagging, disambiguating terms and lexicons, spelling error correction
- Context dependent terms
  - "The size seems small", of a USB vs TV
- Long text
  - Blog postings harder than tweets or reviews



NEW PRODUCT DEVELOPMENT

What do consumers think?

Voice of Customer



“Smart” Home appliaces

- Technology centric

해양구조물 검사문서 “펀치, punch”

- 검사 시행 이후 조건부 합격의 부산물로 미흡한 부분에 대한 수정 사항
- 100여명의 검사원, 1인당 1일 300여 건의 검사문서 입력 (1년에 1천만 건)
- 한 개의 구조물에서 약 2년 간 약 28,800여 건이 입력

Four Frameworks of Analytics

- Prediction/Classification
- Association mining
- Clustering
- Anomaly Detection

Prediction /Classification

- Fit y as a function of x's
- 즉 y = f (x)
- **Regression, Neural Network, Decision Tree**

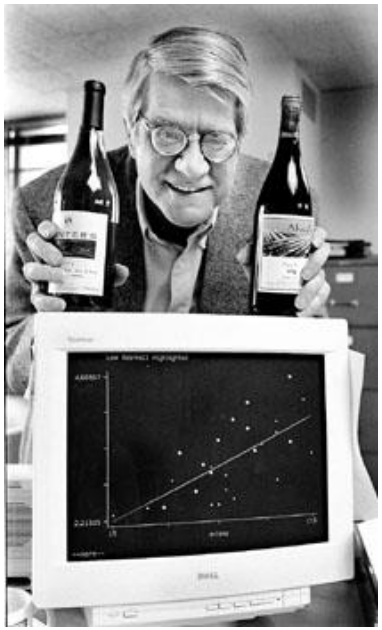

CHÂTEAU MARGAUX
2006
GRAND VIN
PREMIER GRAND CRU CLASSÉ
CHÂTEAU MARGAUX
2009
GRAND VIN
PREMIER GRAND CRU CLASSÉ
WINERY DIRECT
Ch Margaux Margaux 2006
$499.99 750ml Single Bottle
$449.99* 750ml MIX 6
95pts
Wine Spectator
Margaux, Bordeaux, France- "This has a wonderful nose of perfume, sandalwood, ripe plum and berries. Incredible. Full-bodied, with seamless tannins and a long finish. Gorgeous."
Concentrated, Cassis, Violet, Full-
WINERY DIRECT
Margaux Margaux
2009
98 pts
$1499.97
750ml
Wine Enthusiast
Truffle, Full-

Bordeaux

6개월~18개월

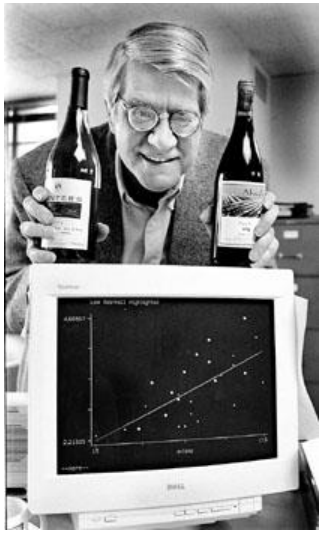
6개월

Quality of Wine?



Temp, Sunlight, Rain, …

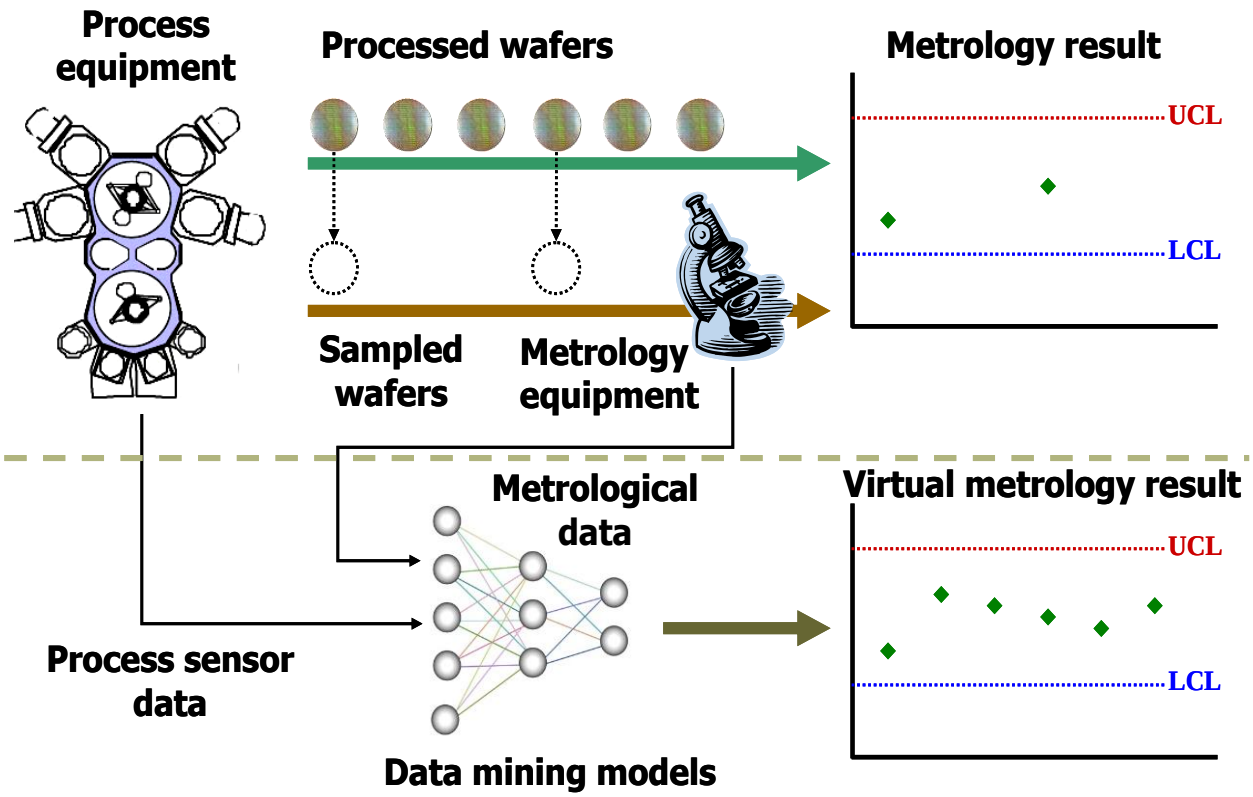
Quality of Wine



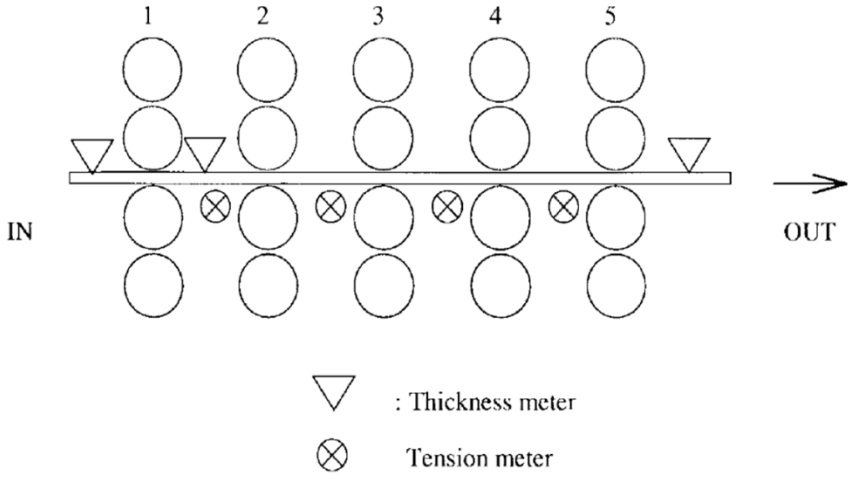
Quality = 12.145
+ 0.00117 * Winter Rainfall
+ 0.06140 * Average Temp
- 0.00386 * Autumn Rainfall

1986 vintage???

Virtual metrology



Optimal Roll Force



Ship Building Time

- Only 4,887 Platinum users among 300,000 premium members
- Who to target among 295,123 for Platinum card?
- Those who use cards in a similar fashion as Platinum users!!

- ***Five Star Hotel*** *> $100* & ***Airline purchase***
  - 787 (Platinum 93.1%)
- ***Golf*** > $480 & ***Sushi Restaurant*** > $100 & no ***Airline purchase***
  - 151 (Platinum 92.7%)
- ***Golf*** > $70 & ***Sushi Restaurant*** > $240 & ***Five Star Hotel*** <$100 & ***Airline purchase***
  - 90 (Platinum 93.3%)

Platinum 지수*	고객 수
90~100	9,116
80~90	6,890
70~80	19,916
60~70	8,908
50~60	11,798
40~50	12,513
30~40	24,974
20~30	35,968
10~20	165,040

THE WALL STREET JOURNAL.

xerox

evolv

Who is likely to quit?

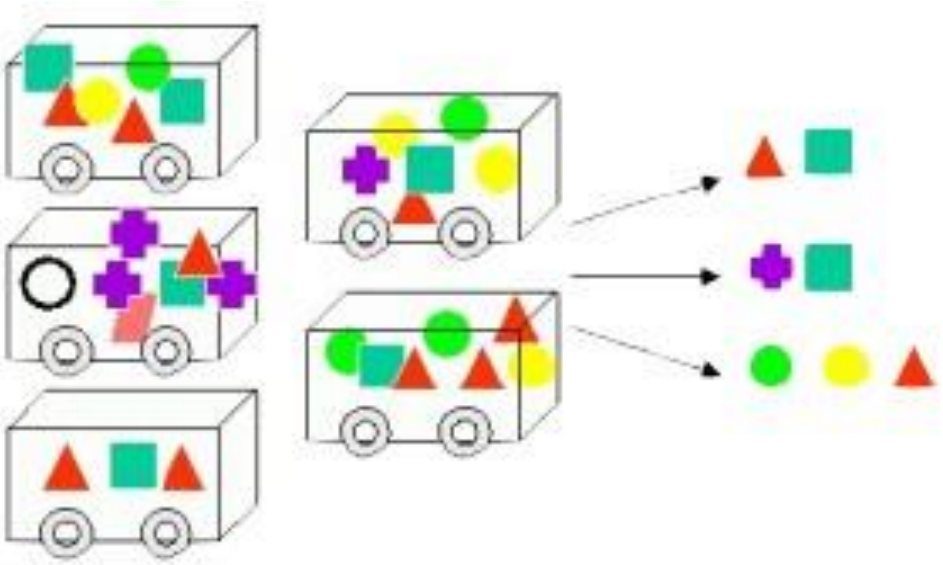
- Live far away
- No reliable transportation
- No Social Network, or Too many (>5)
- Inquisitive (cf. Creative)
- Too Sympathetic

Association Mining

- Identify Event/condition that occur together
- Identify items bought together

- A Priori algorithm

Association Mining



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

Result

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

amazon.com

PANDORA
internet radio

Spotify

Assembling ships

- 200 blocks on Shipyard
- 6,000 movements

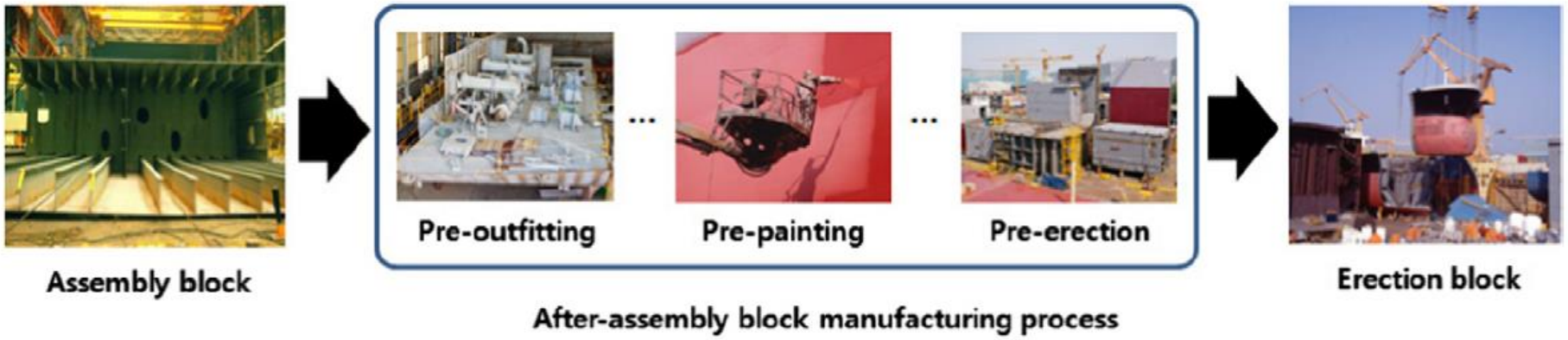


Fig. 1. After-assembly block manufacturing process.

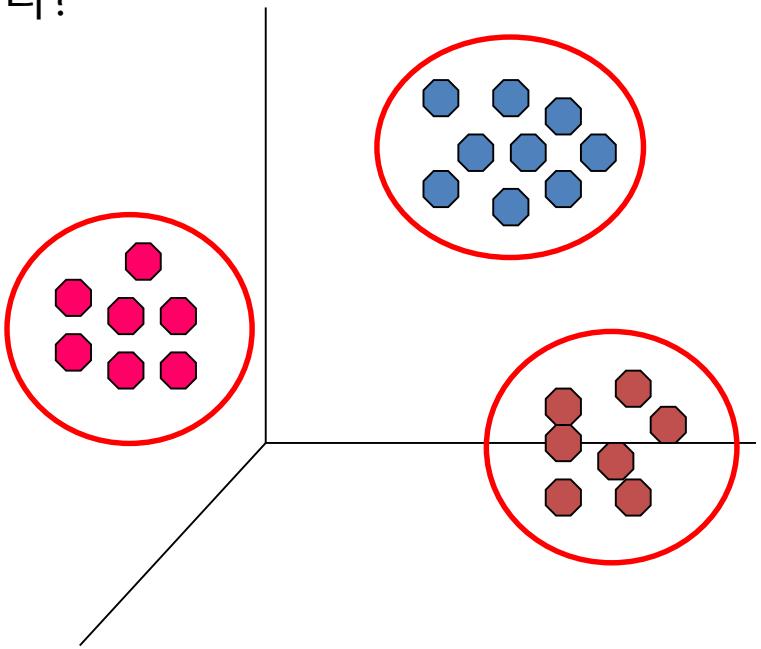
Clustering

- Grouping similar objects into same clusters
- Unsupervised, Exploratory Knowledge Discovery

- K-means, Agglomerative, Competitive Learning

Clustering

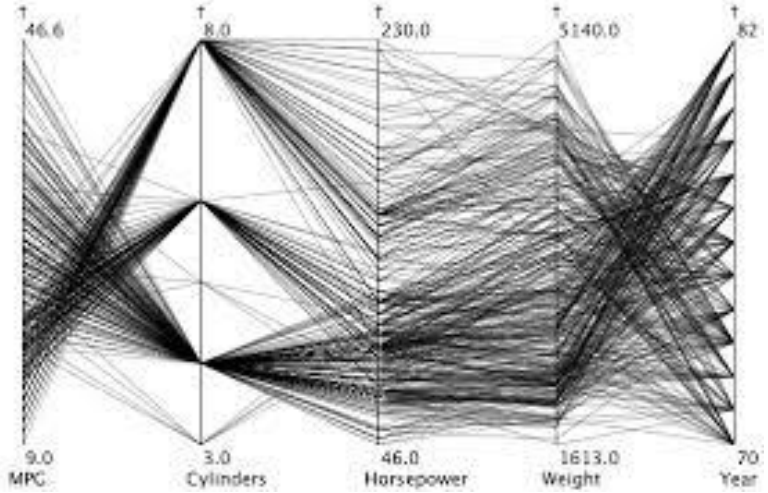
Recency 구입한지 얼마나 지났나?

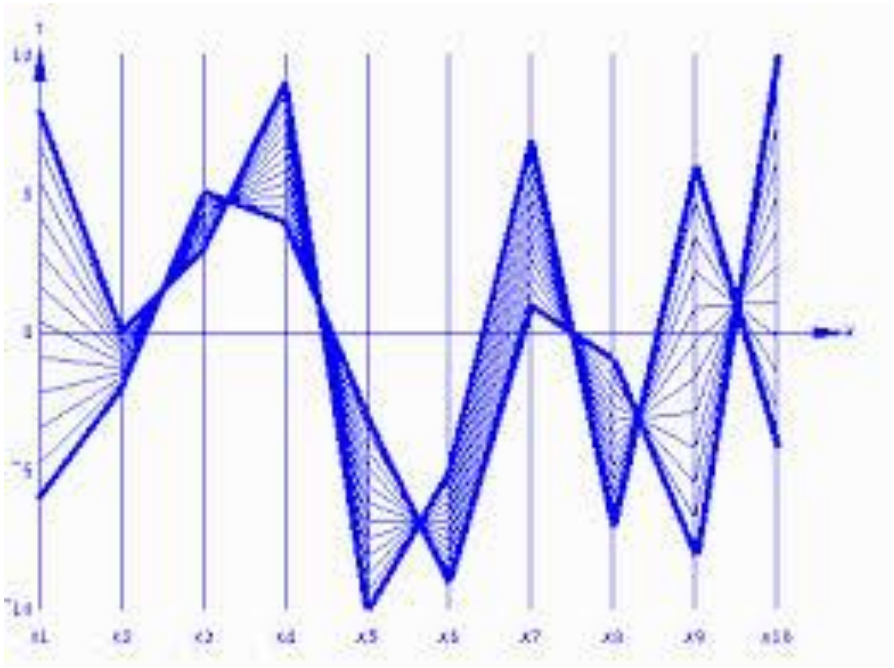


Frequency 최근 1년간 몇 번 구매?

Monetary 최근 1년간 총 구매액?

Multi-dimensional …



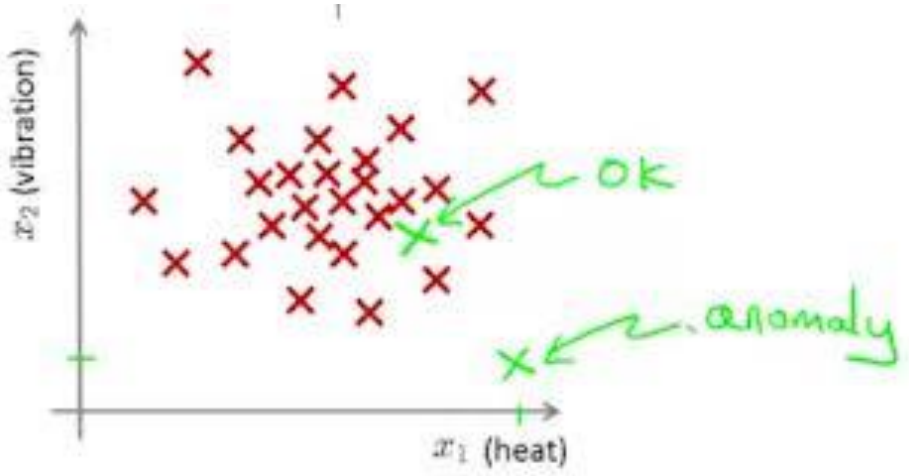


Anomaly Detection

- Learn “Normal” situation,
- Alarm if it does not belong to “Normal”
- Normal Data abundant, yet abnormal Data very few

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

Normal vs Anomaly



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

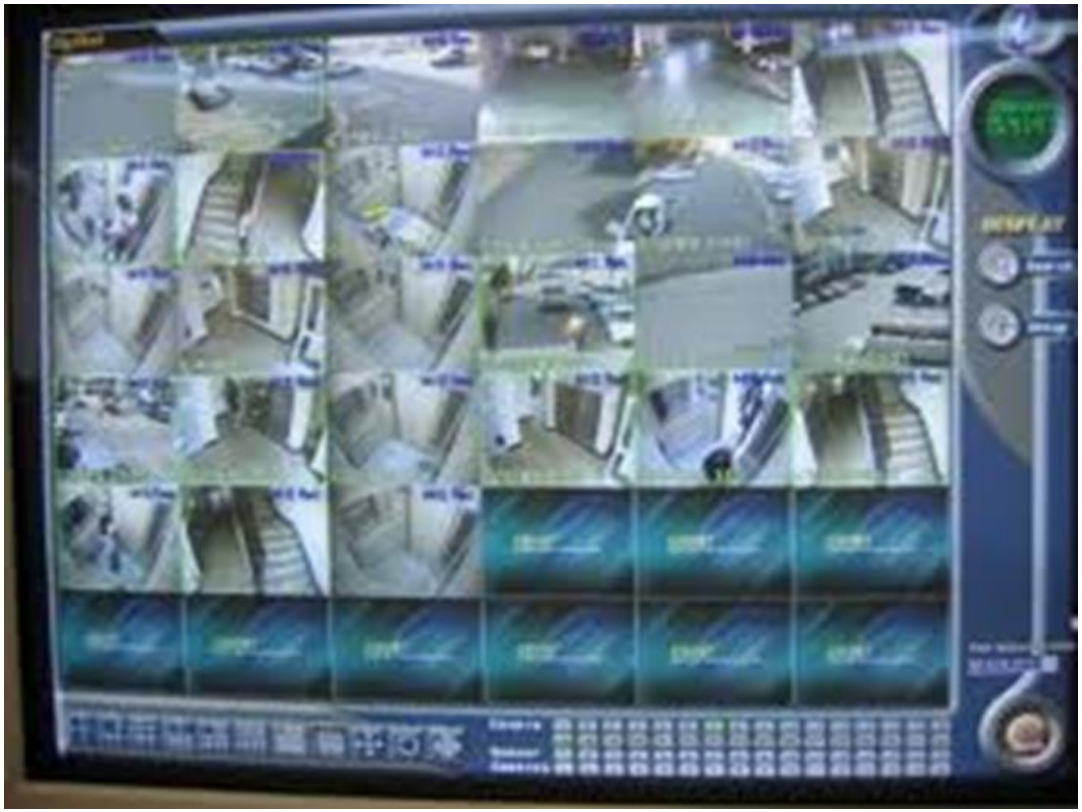
The Economist

2010년 4월 10일자

- When your carpet calls your doctor
- The coming convergence of wireless communications, social networking and medicine will transform health care

CCTV

- Real time interception



Fraud detection

Conclusions

- Big Data?
  - It's okay!
- Big Data?
  - What do we do with it?
  - What values does it provide?
  - Where?
- Great opportunities