지도학습

a. 분류(Classification)
b. 회귀분석(Regression)

비지도 학습

a. K-Means Clustering
b. Hierarchical Clustering

Anomaly detection

a. LOF(Local Outlier Factor)
b. Isolation Forest

Lecture 1: Introduction

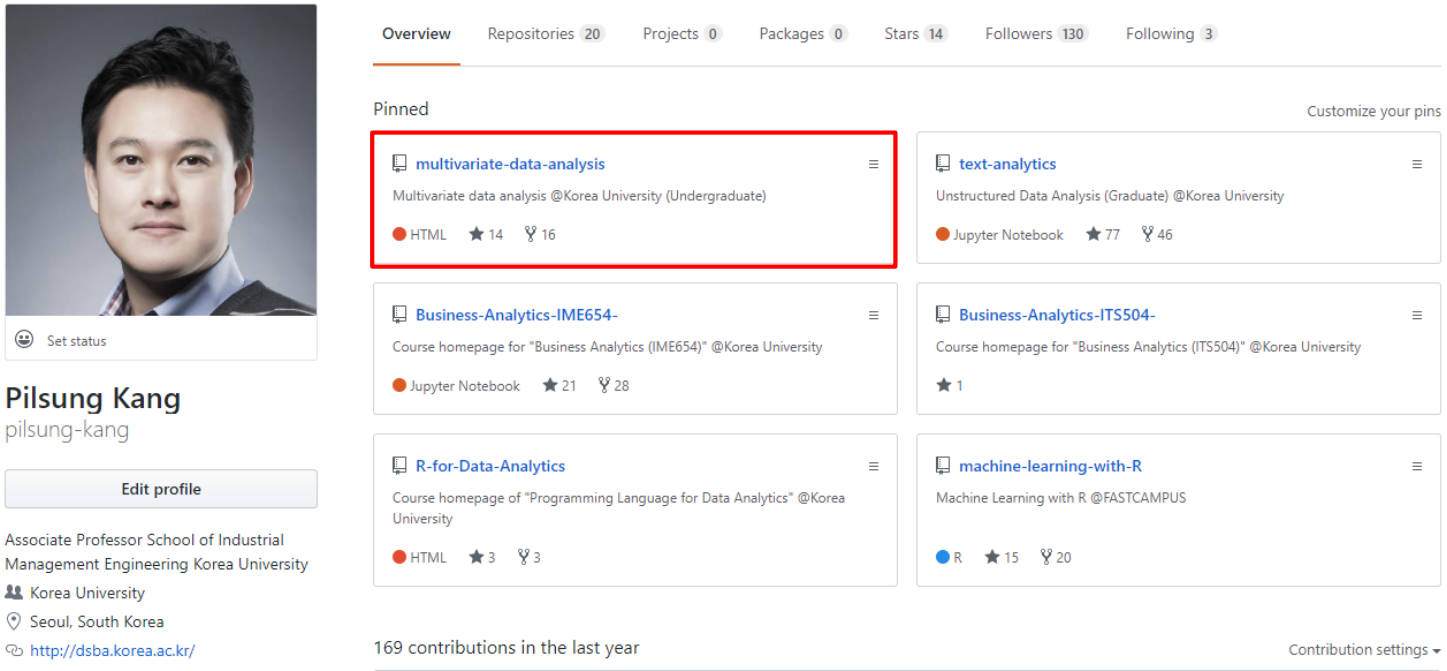
Pilsung Kang

School of Industrial Management Engineering

Korea University

Course Information

- Lecturer
  - ✓ Pilsung Kang, 801A, Innovation Hall
  - ✓ pilsung_kang@korea.ac.kr
- Course homepage
  - ✓ http://github.com/pilsung-kang/multivariate-data-analysis



AGENDA

Amazon: Anticipated Shipping

Amazon은 고객이 구매한 과거이력들을 바탕으로 짧은 미래, 즉 단시간 내에 특정 고객들이 어떠한 상품을 살 것인지 미리 예측을 진행

고객들의 구매이로구나 데이터 분석을 통해서 특정 고객이 어떤 제품을 구매할 것인지를 예측을 하고 예측된 상품들을 가까운 물류창고에다가 옮겨놓으면 고객이 실제로 구매를 하는 순간(클릭)을 할때 가까운 물류센터에서 출발을 하대되면 훨씬 더 짧은 배송시간을 가질 수 있게 된다는 것을 의미

물류센터 (핵심)

1) 데이터 분석을 통해 예측을 진행
2) 예측에 해당하는 action을 수행
3) 고객 시간에서는 더 빠른 시간에 편리하게 물건을 알아볼 수 있다.

landing.ai: What is defective product and where?

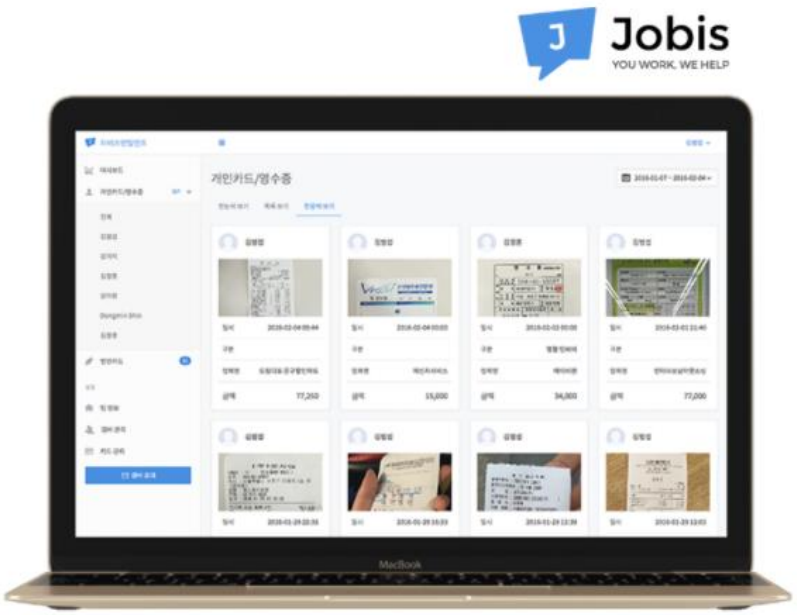
제조업 관점

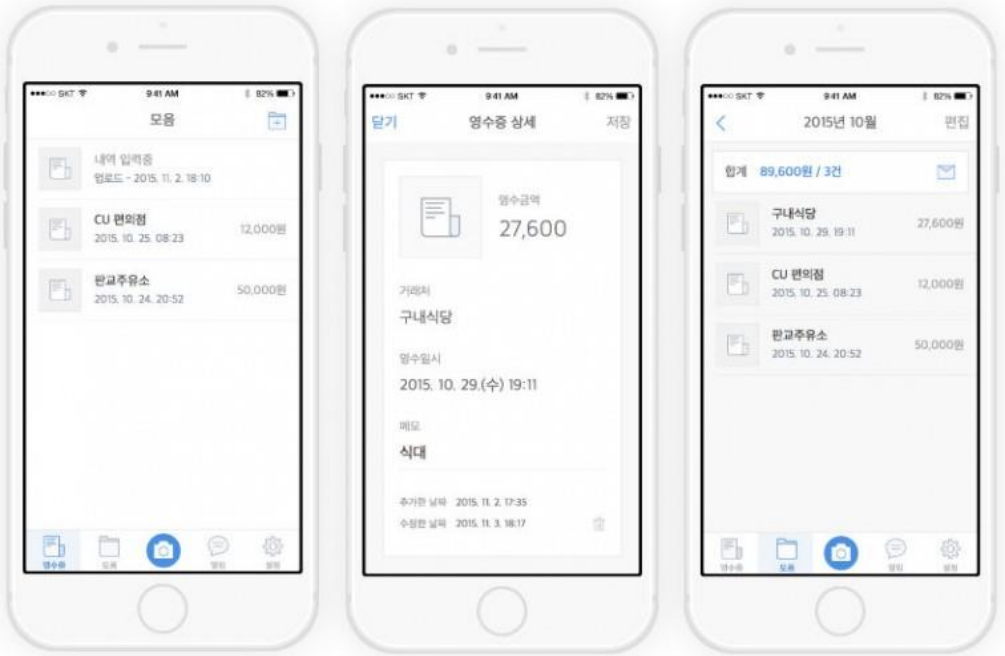
1) Adaptive manufacturing
2) Automated Quality Control
3) Predictive Maintenance

충분히 많은 data set을 통해서 어떤 제품, 즉 사람 검사원이 양품/불량품을 확인했다면 데이터 분석 인공지능 알고리즘에 제공을 하여 과거의 입력데이터(사진)과 출력 데이터 (실제 제품이 불량한지 아닌지에 대한 판정 결과물의 주며 학습을 진행한다.

* 어떠한 영역 때문에 불량으로 판정되었는지 까지도 판별

Is Data a Tool or Purpose of Business?

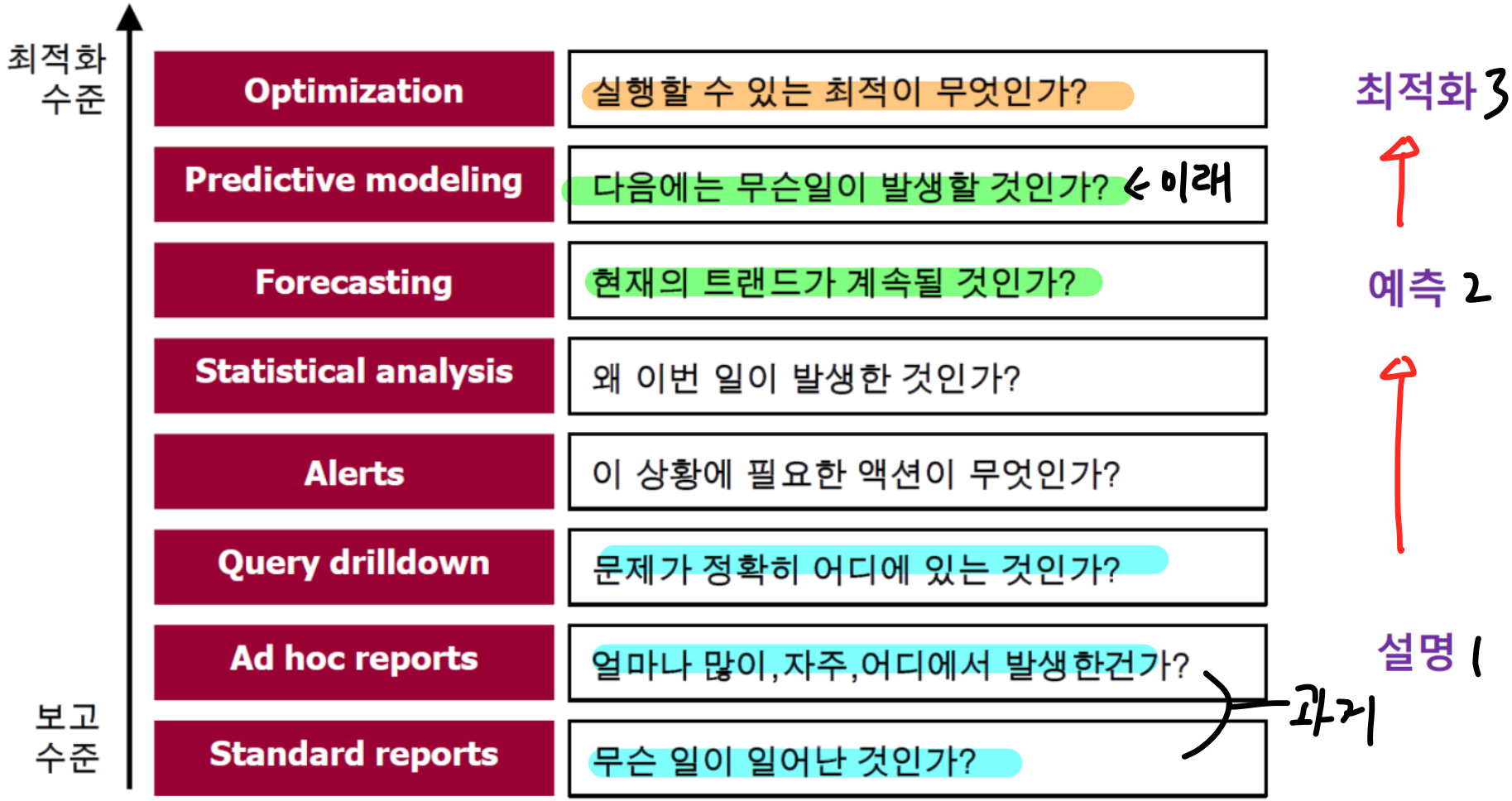




Data-driven Decision Making
(데이터 기반 의사결정)

- What we want to know

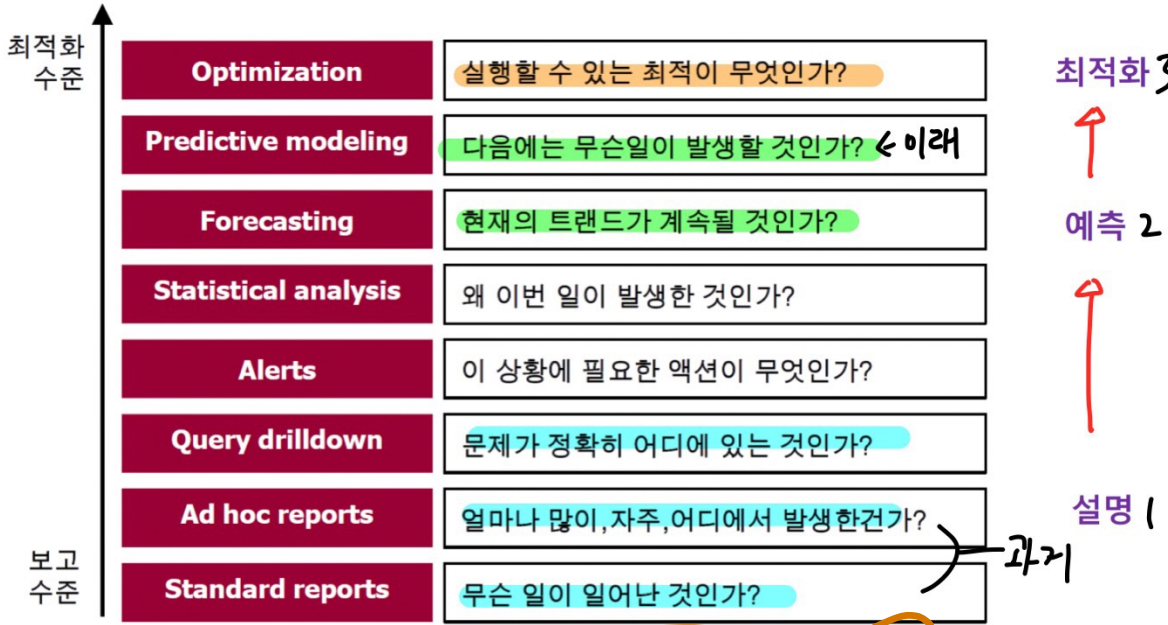
어떤 회사 및 개인 기관, 공공기관 중 어떤 특정한 시스템에 의해 의사결정을 하는데 있어서 데이터를 기반으로 객관적인 수치를 통해서 최적의 의사결정을 하자!

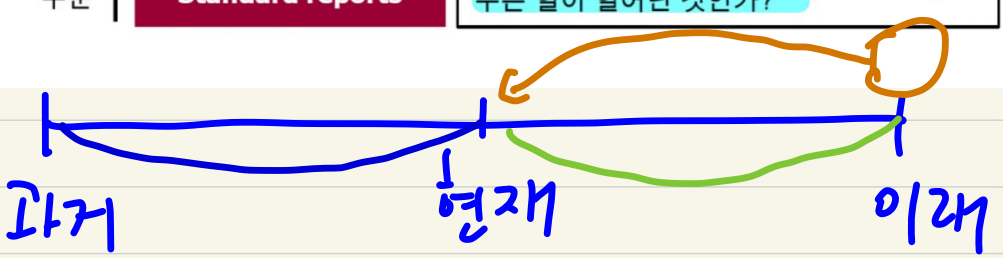


Data-driven Decision Making
(데이터 기반 의사결정)

- What we want to know

어떤 회사 및 개인 기관, 공공기관 중 어떤 특정한 시스템에 의해 의사결정을 하는데 있어서 데이터를 기반으로 객관적인 수치를 통해서 최적의 의사결정을 하자!





설명 단계

설명의 관점에서는 과거에서부터 현재까지 무슨 일이 일어났는지 보고, 그러한 일들이 얼마나 자주 어디서 발생 하였는가? 문제는 도대체 어디있는가?

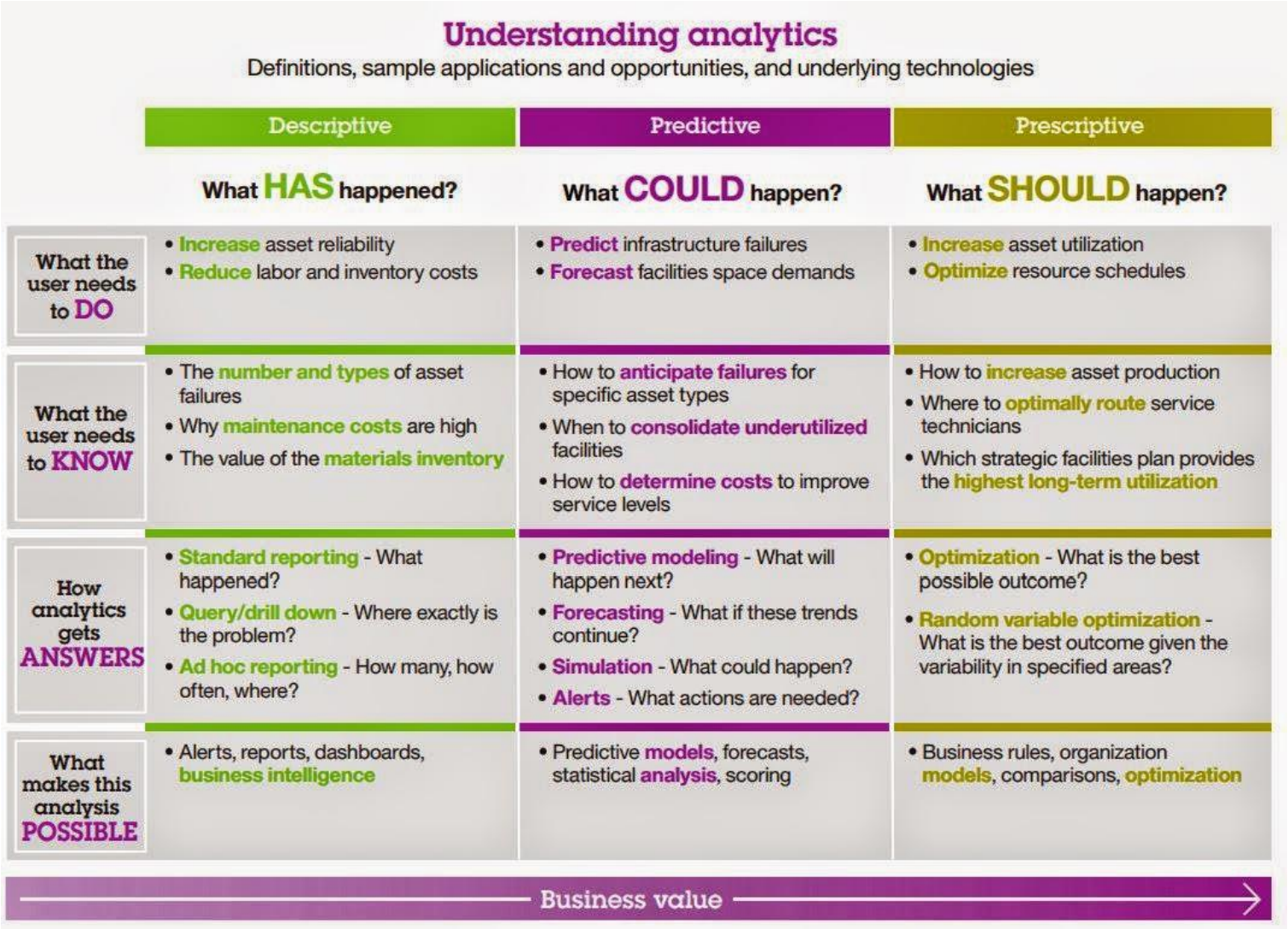
예측 관점에서는 과거에서부터 현재까지 있었던 상황을 바탕으로 미래에 무슨 일이 일어날 것인지 선제적으로 알아보는 것

미래단계에서는 미래에 발생할 일을 바탕으로 현재의 우리가 어떤 액션을 취해야 하는지 바로 실행할 수 있는 최적의 답을 찾는 것이 최적화 단계이다.

Data-driven Decision Making

운영(설명) 예측 최적화

- Descriptive vs. Predictive vs. Prescriptive Analytics



Understanding analytics

Definitions, sample applications and opportunities, and underlying technologies

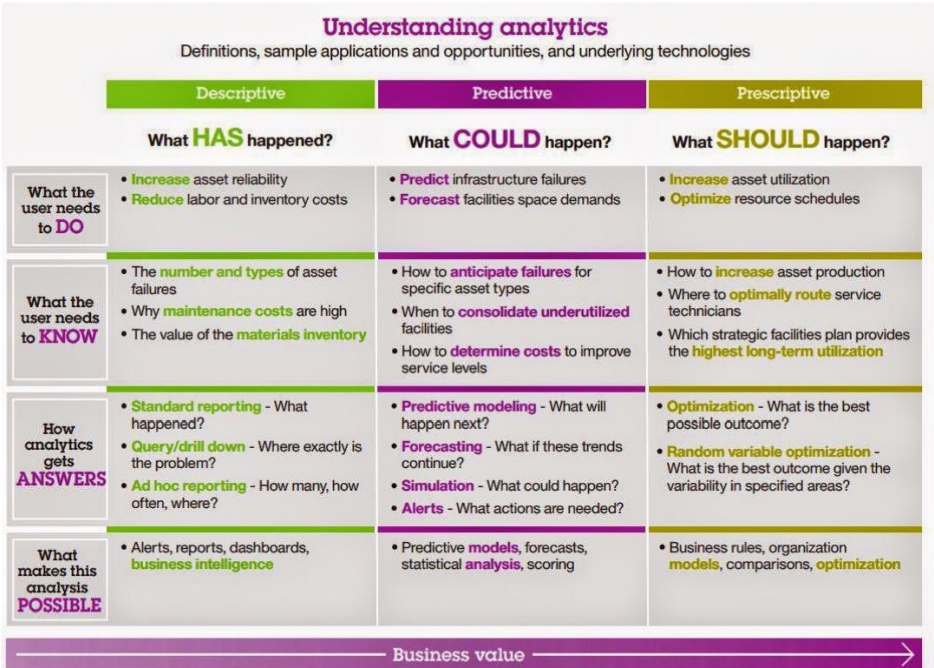
	Descriptive	Predictive	Prescriptive
	What **HAS** happened?	What **COULD** happen?	What **SHOULD** happen?
What the user needs to **DO**	• **Increase** asset reliability • **Reduce** labor and inventory costs	• **Predict** infrastructure failures • **Forecast** facilities space demands	• **Increase** asset utilization • **Optimize** resource schedules
What the user needs to **KNOW**	• The **number and types** of asset failures • Why **maintenance costs** are high • The value of the **materials inventory**	• How to **anticipate failures** for specific asset types • When to **consolidate underutilized** facilities • How to **determine costs** to improve service levels	• How to **increase** asset production • Where to **optimally route** service technicians • Which strategic facilities plan provides the **highest long-term utilization**
How analytics gets **ANSWERS**	• **Standard reporting** - What happened? • **Query/drill down** - Where exactly is the problem? • **Ad hoc reporting** - How many, how often, where?	• **Predictive modeling** - What will happen next? • **Forecasting** - What if these trends continue? • **Simulation** - What could happen? • **Alerts** - What actions are needed?	• **Optimization** - What is the best possible outcome? • **Random variable optimization** - What is the best outcome given the variability in specified areas?
What makes this analysis **POSSIBLE**	• Alerts, reports, dashboards, **business intelligence**	• Predictive **models**, forecasts, statistical **analysis**, scoring	• Business rules, organization **models**, comparisons, **optimization**

Business value

Data-driven Decision Making

운영(설명) 예측 최적화

- Descriptive vs. Predictive vs. Prescriptive Analytics



Understanding analytics

Definitions, sample applications and opportunities, and underlying technologies

	Descriptive	Predictive	Prescriptive
	What HAS happened?	What COULD happen?	What SHOULD happen?
What the user needs to DO	• Increase asset reliability • Reduce labor and inventory costs	• Predict infrastructure failures • Forecast facilities space demands	• Increase asset utilization • Optimize resource schedules
What the user needs to KNOW	• The number and types of asset failures • Why maintenance costs are high • The value of the materials inventory	• How to anticipate failures for specific asset types • When to consolidate underutilized facilities • How to determine costs to improve service levels	• How to increase asset production • Where to optimally route service technicians • Which strategic facilities plan provides the highest long-term utilization
How analytics gets ANSWERS	• Standard reporting - What happened? • Query/drill down - Where exactly is the problem? • Ad hoc reporting - How many, how often, where?	• Predictive modeling - What will happen next? • Forecasting - What if these trends continue? • Simulation - What could happen? • Alerts - What actions are needed?	• Optimization - What is the best possible outcome? • Random variable optimization - What is the best outcome given the variability in specified areas?
What makes this analysis POSSIBLE	• Alerts, reports, dashboards, business intelligence	• Predictive models, forecasts, statistical analysis, scoring	• Business rules, organization models, comparisons, optimization

Business value

① 무슨 일이 일어났는가? (descriptive) Knowledge base 구축

→ 과거부터 현재까지 무슨 일이 일어났는가?

(과거에서부터 현재까지의 상황을 요약해주고 인사이트 추출

② 무슨 일이 일어날 것인가? (predictive)

→ 이래의 예측에 대해서 보는것

③ 무슨일이 일어나도록 만들어야 하는가? (prescriptive)

* ② 와 ③ 의 차이는 Predictive Analysis 에서는 우리가 우연과를 조절하지는 않는다.

지금 상황이 지속되면 어떤일이 발생할것인가 예측

Prescriptive 에서는 우리 조직에 가장 유리한 상황이 발생하도록

하려면 어떤 일이 일어나도록 만들게 하는가? (변화시킬수 있는 요소들을 어떻게 변화시키는가?)

Machine Learning

- Definition
  - ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)

(경험)

Supervised Learning

- Predict a single "target" or "outcome" variable
- Finds relations between X and Y.
- Train (learn) data where target value is known
- Score data where target value is not known

Unsupervised Learning

- Explores intrinsic characteristics.
- Estimates underlying distribution
- Segment data into meaningful groups or detect patterns
- There is no target (outcome) variable to predict or classify

Machine Learning

- Definition
  - ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)

(경험)

Supervised Learning

- Predict a single "target" or "outcome" variable
- Finds relations between X and Y.
- Train (learn) data where target value is known
- Score data where target value is not known

Unsupervised Learning

- Explores intrinsic characteristics.
- Estimates underlying distribution
- Segment data into meaningful groups or detect patterns
- There is no target (outcome) variable to predict or classify

→ 제조업에서는 불량탐지같은경우는 새로들어온 제품이 정상인지 아닌지 예측 (task)

→ 이 task를 얼마나 잘수행하는지 측정할수 있는 성능 (Performance P)

※ 어떠한 task에 대해서 P라는 측정지표를 가지고 측정을 했을때 경험이 제공될수록 개선이 된다.

(E)=데이터를 제공을 하면 이 데이터로부터 task를 수행하는데 있어 평가 성능이 점점 향상될 수 있는것이 ML.

→ X (원인) ∨ (예측결과)

Machine Learning

- Definition
  - ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)

(경험)

Supervised Learning

- Predict a single "target" or "outcome" variable
- Finds relations between X and Y.
- Train (learn) data where target value is known
- Score data where target value is not known

Unsupervised Learning

- Explores intrinsic characteristics.
- Estimates underlying distribution
- Segment data into meaningful groups or detect patterns
- There is no target (outcome) variable to predict or classify

9/58

지도학습은 X와 Y가 핵심이다. (X)원인이고 (Y)예측하고자 하는 타겟이다.

입력과 출력의 기준으로 제품의 불량품 판별에서는

제품의 이미지(입력)
제품의 불량 인지 아닌지(출력)

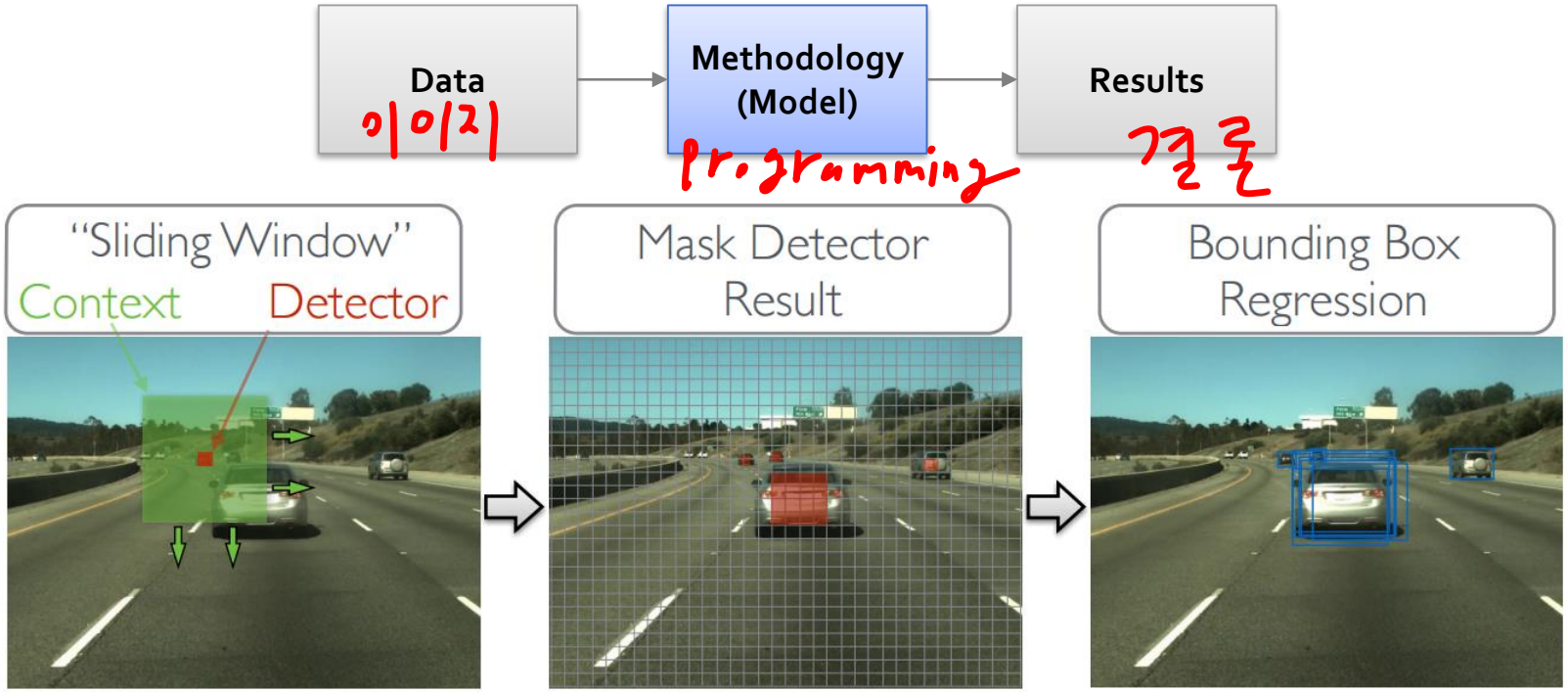
즉 X와 Y의 관계식을 찾아주는 것이 지도학습이다.

비지도학습은 Y가 없다.

즉 X라는 데이터로 부터 가질 수 있는 분포 라던지 특징 및 패턴을 찾는다.

Machine Learning

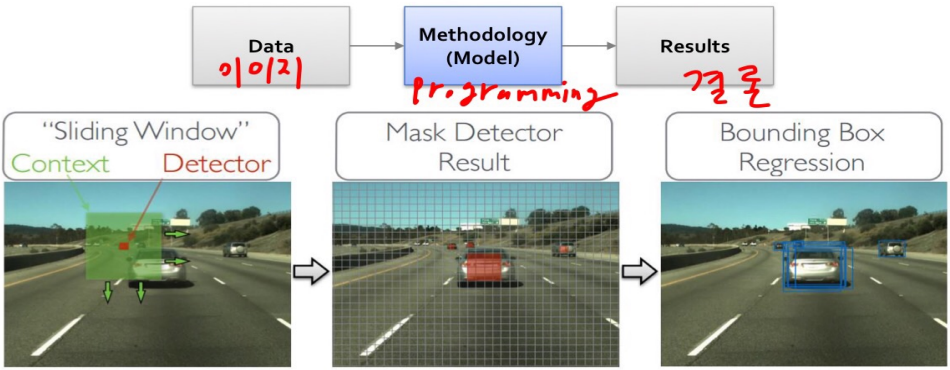
- Definition
  - ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)



Machine learning models in Self-driving cars

Machine Learning

- Definition
  - ✓ A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at task in T, as measured by P, improves with experience E," – Mitchell (1997)



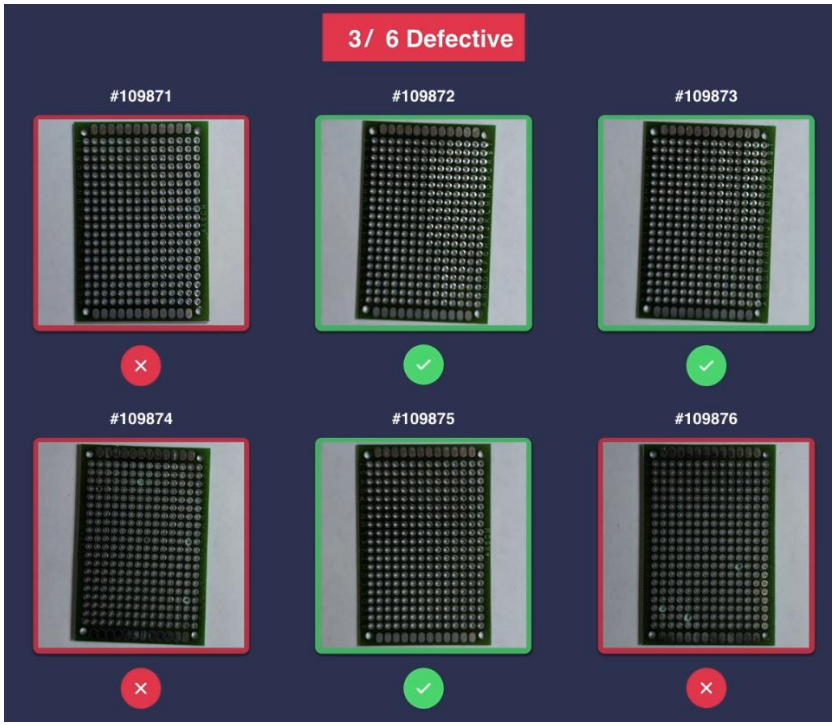
Machine learning models in Self-driving cars

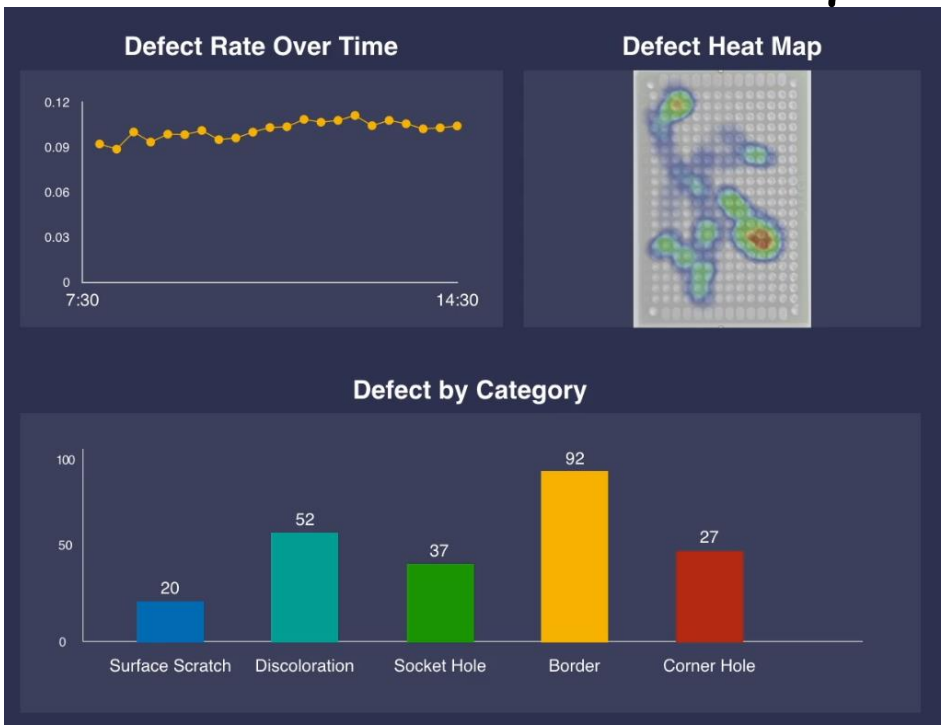
→ 데이터는 각각 촬영된 이미지다. (번호판 인식)
→ 내앞에 차가 얼마나 떨어진지

Machine Learning

- Machine learning in manufacturing industry
  - ✓ Based on product images and class labels (good/bad),
  - ✓ Models are trained to correctly predict which product are defective and where suspicious areas are

의심되는 영역





→기본산업의 구성요소 (빅데이터)

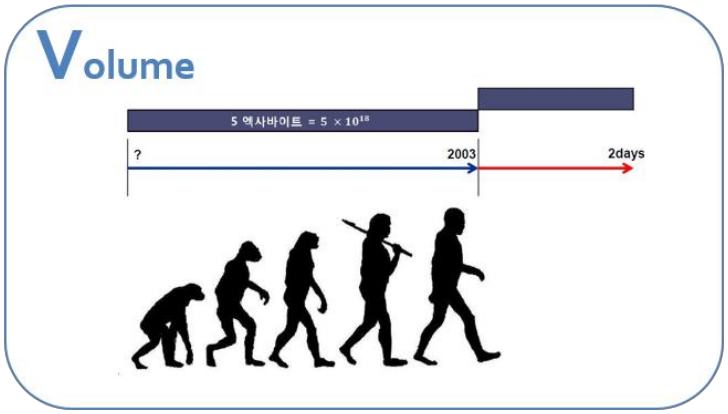
Big Data

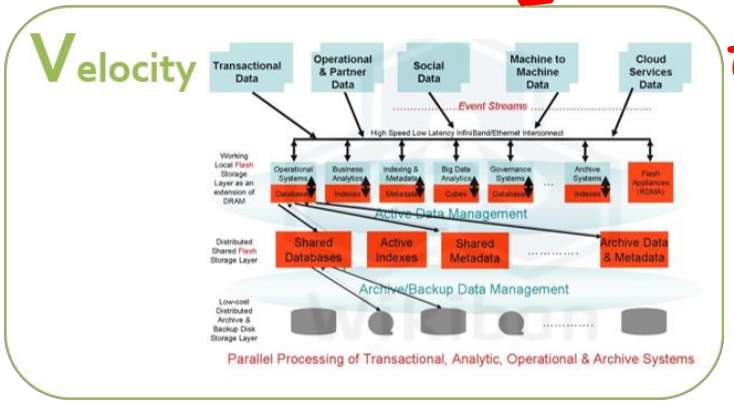
velocity

→데이터가 생성되는 요소들 뿐만이 아니라 처리되는 속도가 매우 빠른 시간이 요구된다 (실시간처리)

- 4Vs in Big Data
  - ✓ Volume, Velocity, Variety, Value

분석대상이 되는 데이터의 원천이 다양함





효과표시

→데이터의 존재만으로 복잡한 분석방법론이 없어도 문제가 해결될수 있는 상황이 존재

Big Data

- Big Data itself is valuable without any complicated analytics methods

Type	Institution	Forecast
데이터의 중요성 Economic Feasibility in Industries	Economist (2010)	☐ " Data are becoming the new raw material of business: an economic input almost on a par with capital and labour."
	Gartner (2011)	☐ "Intelligence about Information is the Oil of the 21st Century." Future competitive advantage depends on data. ☐ Winning organizations understand the stage of the data economy and overcome information silos through effective information sharing.
	McKinsey (2011)	☐ Big Data is the next frontier for innovation, competitiveness and productivity ☐ Big Data will create values worth more than $600 billion in 5 areas including medicine and public administration
National Competitiveness	US PCAST	☐ Advisors emphasize that US government organizations should focus on the strategy for transformation of data into knowledge, and of knowledge into action.
	Singapore	☐ Singapore looks to evaluate threatening risks and detect environmental changes based on data

*현재 상황을 얼마나 잘 파악하는가?

Big Data

- Case Study: Navigation system

늙대폰 navigation

순정 네비게이션

VS

→ 데이터 자체로서의 의미를 가지는것보다는 것은 어떤 의미냐면 교통상황을 보는데, 핸드폰 가입자들의 정보를 기반으로 실시간으로 보여주기에 정확하다.

*데이터 양의 문제

*현재 상황을 얼마나 잘 파악하는가?

Big Data

- Case Study: Navigation system

휴대폰 navigation

순정 네비게이션

VS

→ 데이터 자체로서의 의미를 가지는것보다는 것은 어떤 의미냐면 교통상황을 보는데, 핸드폰 가입자들의 정보를 기반으로 실시간으로 보여주게에 정확하다.

*데이터 양의 문제

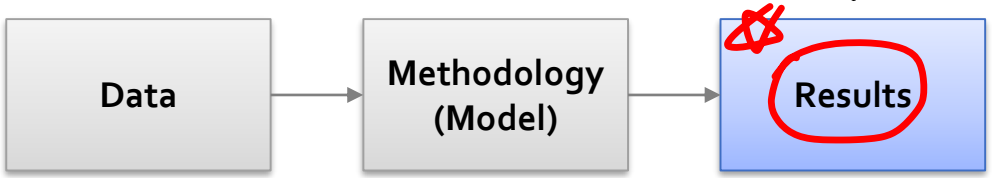
14/58

* 교통상황을 측정하는데 있어서 휴대폰 네비게이션은 휴대폰을 사용하고 있는 사람들 그럼 특정 통신사를 고를수 없지만, 통신사별로 수백만에서 이천만에 가까운 사람들의 이동경로를 수집하는것과 네비게이션 시스템을 구축하기 위해서 센서가 장착된 수십만대 자동차를 이용해서 교통흐름을 측정하는 것을 비교하면 휴대폰 네비게이션이 훨씬 더 정확하게 측정이된다.

* 사실 최단거리 알고리즘은 크게차이가 나지는 않지만 현재상황을 얼마나 잘 파악하고 있느냐에 따라서 휴대폰 네비가 유리하다. (정확↑ 속도↑, 최단경로 알고리즘의 차이문제 X 데이터 양의 문제

Data Mining

- Definitions



결과물을 어떻게 활용하고 추정할 것, 뽑아낼 것이냐에 대해 방점이 찍혀있는 것이다.

- ✓ Extracting useful information from large datasets. (Hand et al., 2001)

대량의 데이터로부터 유용한 정보를 뽑아낸다.

- ✓ The process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, 2000)

대량 데이터를 탐색

↘ 탐색 후에 pattern을 찾는다.

- ✓ The process of discovering meaningful new correlations, patterns and trends by sifting through large amount data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. (Gartner Group, 2004)

↳ 데이터를 통해서 그 안에 숨겨진 지식을 발굴하자!

Data Mining

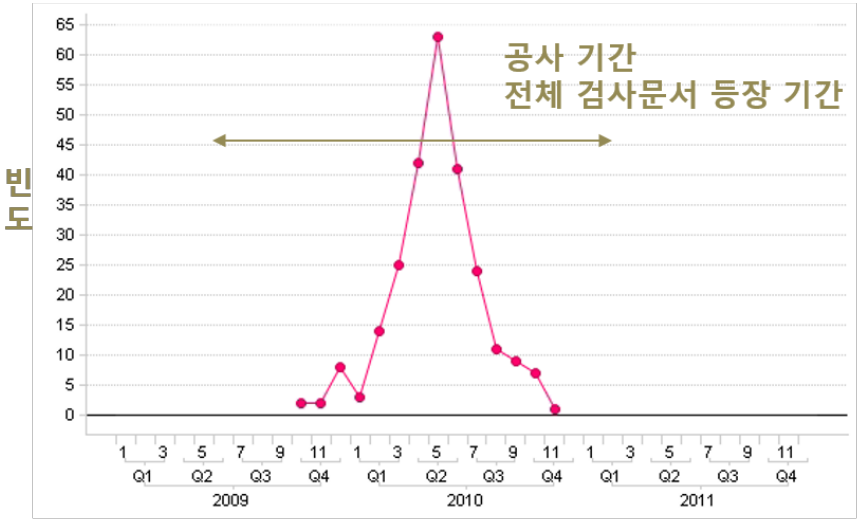
결과물을 어떻게 활용하고 측정할것, 뽑아낼것이냐에 대해 방점이 찍혀있는것이다.

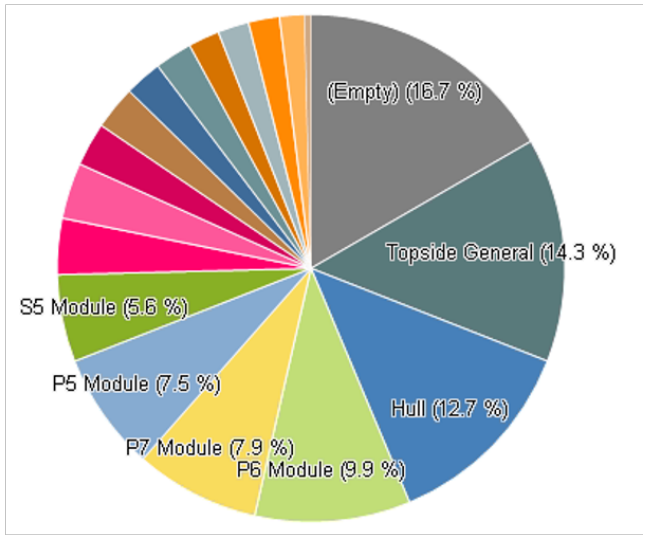
- Definitions

"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"





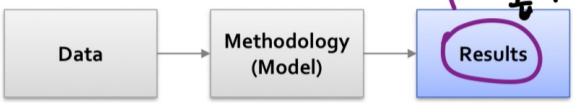
발생하는 장소

→ 문제 원인 및 조치결과

Data Mining

결과물을 어떻게 활용하고 측정할것, 뽑아낼것이냐에 대해 방점이 찍혀있는것이다.

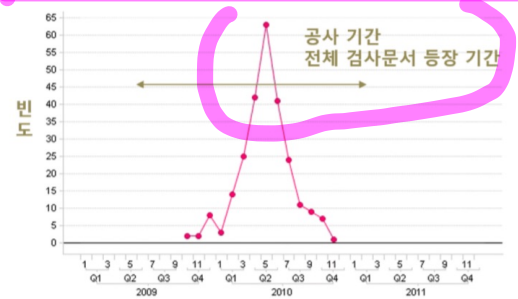
- Definitions

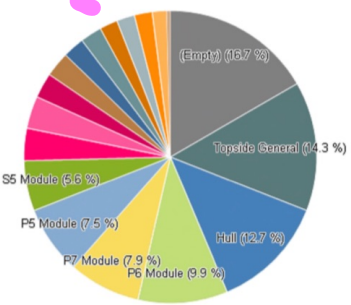


"파이프(pipe)가 흔들리니(shake), 지주(support)를 추가(add)하라"
언제, 어디서?
"공사 중반, Topside General, Hull, P5,6,7 Module 등에서 주로 발생한다"

Data 적용 mining

다양 검사문서





→ 문제 원인 및 조치결과

16/58

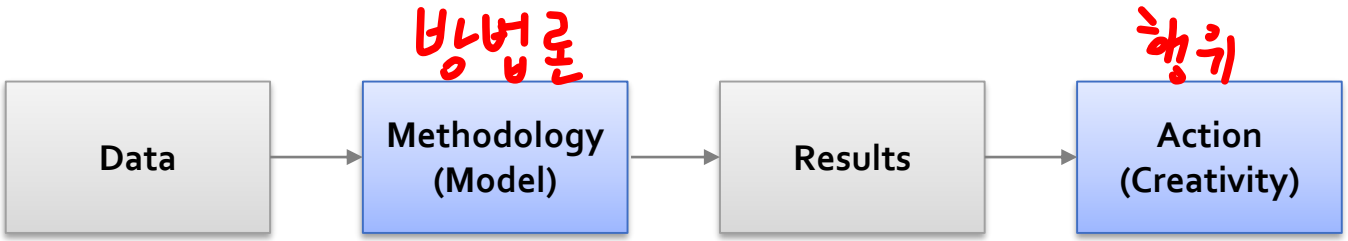
→ 예를 들어서 특정한 어떤설비를 생산하는데 있어서 강비나 콘크리트를 건축하는데 있어서는 공사기간이 오래걸리게 때문에 매일매일 사람들이 검사한들이 과거에 선전작업 내역에 대해 기록을 하고 혹시문제가 있으면 원인은 무엇이었는지 그리고 그것을 어떻게조치했는지에 대한검사문서를 만든다.

데이터 마이닝 기법적용후 전체 공사기간中 PIPE가흔들림 (지속적)

↳ +솔루션추가, 어떠한 위치에서 발생한지 알고있다면 나중에 유사한 구조물이나 건축물들을 만들 때 결과물들을 활용을 해서 사후조치가 취해지지 않도록 공정의 시작관점에서 이러 계획을 세울수 있음을 의미

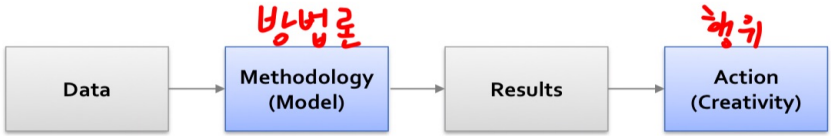
Artificial Intelligence

- Definition
  - ✓ Computers and computer software that are capable of intelligent behavior
  - ✓ Intelligent agent perceives its environment and takes actions that maximize its chance of success



Artificial Intelligence

- Definition 방법론+행위 지능화된 역량을 지님
  - ✓ Computers and computer software that are capable of intelligent behavior
  - ✓ Intelligent agent perceives its environment and takes actions that maximize its chance of success



17/58

→ 약한 인공지능은 지금까지의 경험을 토대로 가장 최적의 의사결정을 낸다. (창의적 솔루션을 도출할 가능성은 상대적으로 낮다)

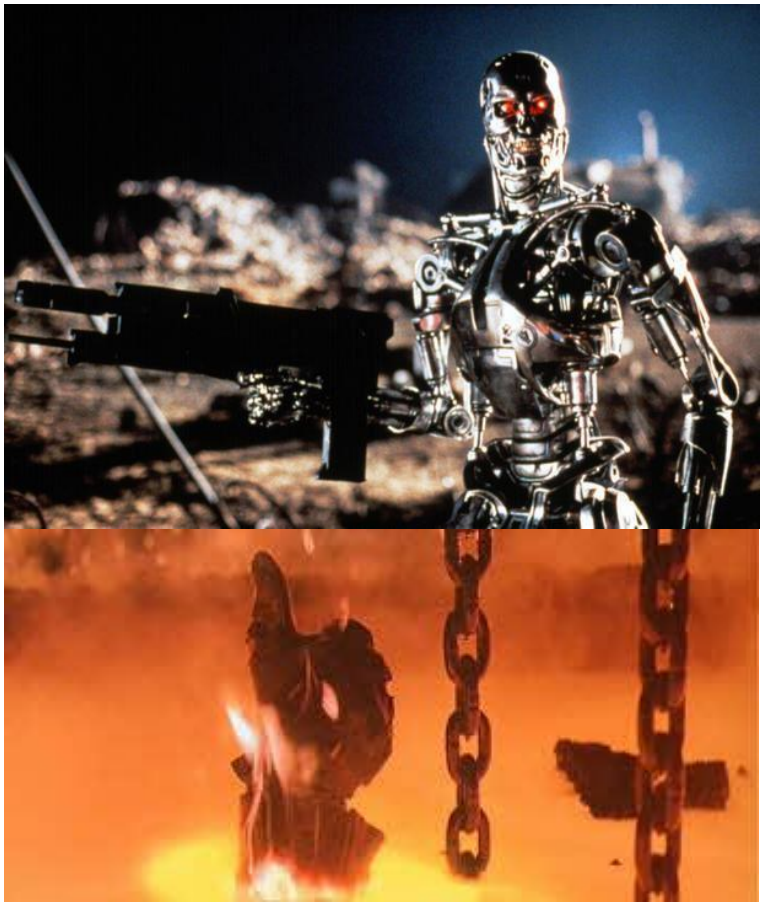
→ 강한 인공지능은 한번도 보지못한 상황에서 어떤선택을 해야할것인가 대해서도 창의성을 발휘해서 사람처럼 생각하고 행동할수 있는 인공지능을 의미한다.

※ 행위에 대해서 사람처럼 정말 창의성이 있는가? 이게 이제 약한 or 강한 인공지능을 구분하는 여부

Artificial Intelligence

- AI should be…

터미네이터



사만다(인공지능)

Artificial Intelligence

- AI in nowadays…

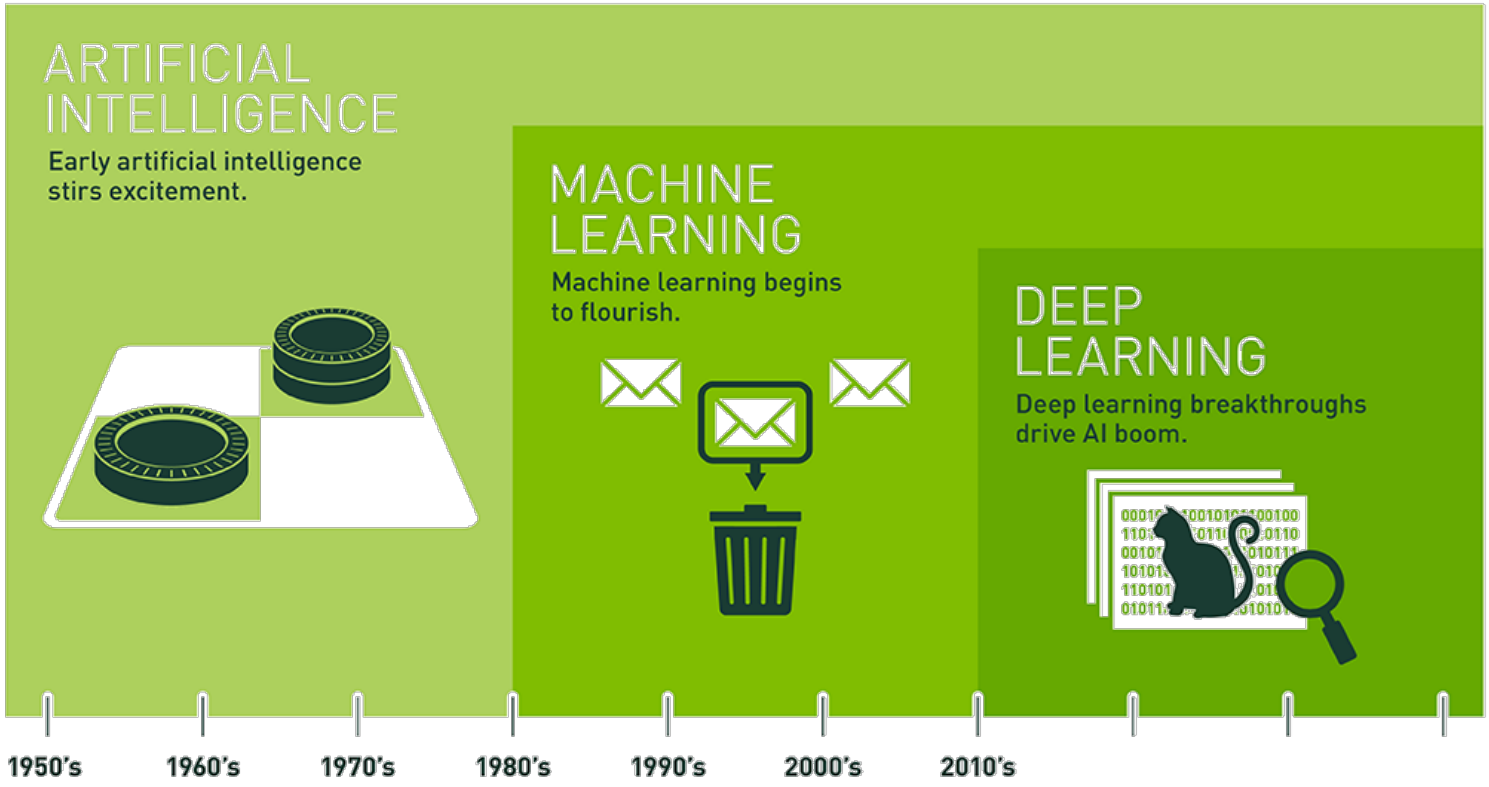
Artificial Intelligence

- AI in nowadays…

Artificial Intelligence

가장 상위 개념이 AI

- AI vs. Machine Learning vs. Deep Learning



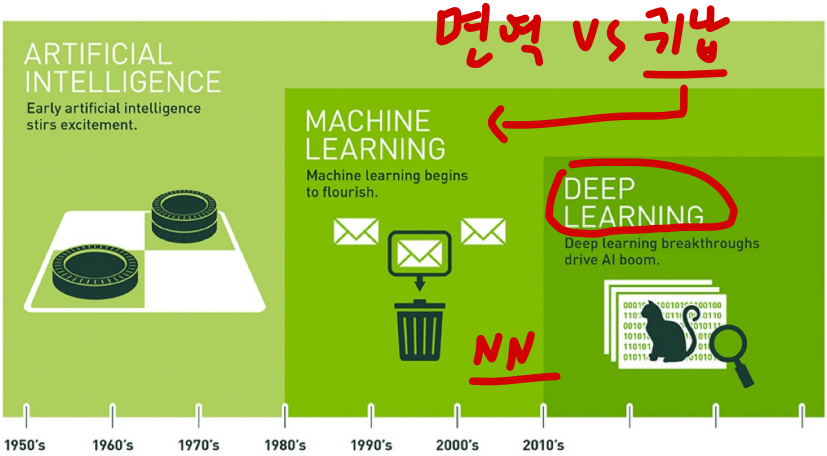
Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Artificial Intelligence

가장 상위 개념이 AI

- AI vs. Machine Learning vs. Deep Learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

→ 인공지능은 사람처럼 생각하는 기계를 만들자라고 생각

인공지능을 만드는 2가지 지능

① 면적적 지능 ② 귀납적 지능

①은 3단논법처럼 A면 B고, B면 C고, C면 D고 논리적인 추론

Symbolic 연산을 진행시 추론가능

②는 경험을 통해 습득

ML은 ② 귀납적 지능을 구현하고자 한다.

AGENDA

데이터 사이언스 분야에서의 응용

대량의 데이터를 시각화 및 요약

Data Science Applications

1

Visualization for intuitive understanding

Word Cloud

대량의 데이터를 시각화 및 요약

Data Science Applications

취잉연설문

1

Visualization for intuitive understanding

WordCloud

WordCloud는 어떠한 문서에서 특정한 단어가 등장을 많이쓰면 크기를 크게 표현해주고 특정한 단어가 적게 나올수록 크기를 작게 표현해주는 시각화 도구

→ 대량의 데이터를 시각화 혹에 요약

* 짧은 시간 내에 2명의 전임 대통령이 어떤 방향의 국정에 초점을 맞추고 있는지 보기좋다.

현대자동차 (News 기반)

Data Science Applications

데이터 분석기법을 통해서 빠르게 시각화및요약

1

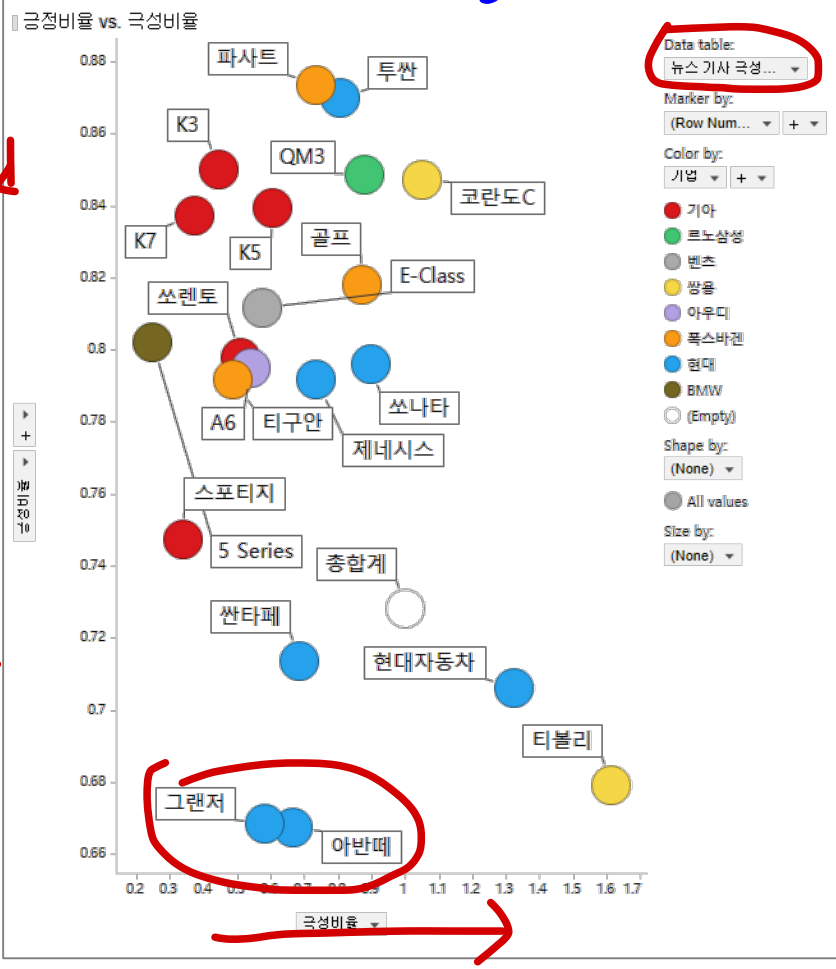
Visualization for intuitive understanding

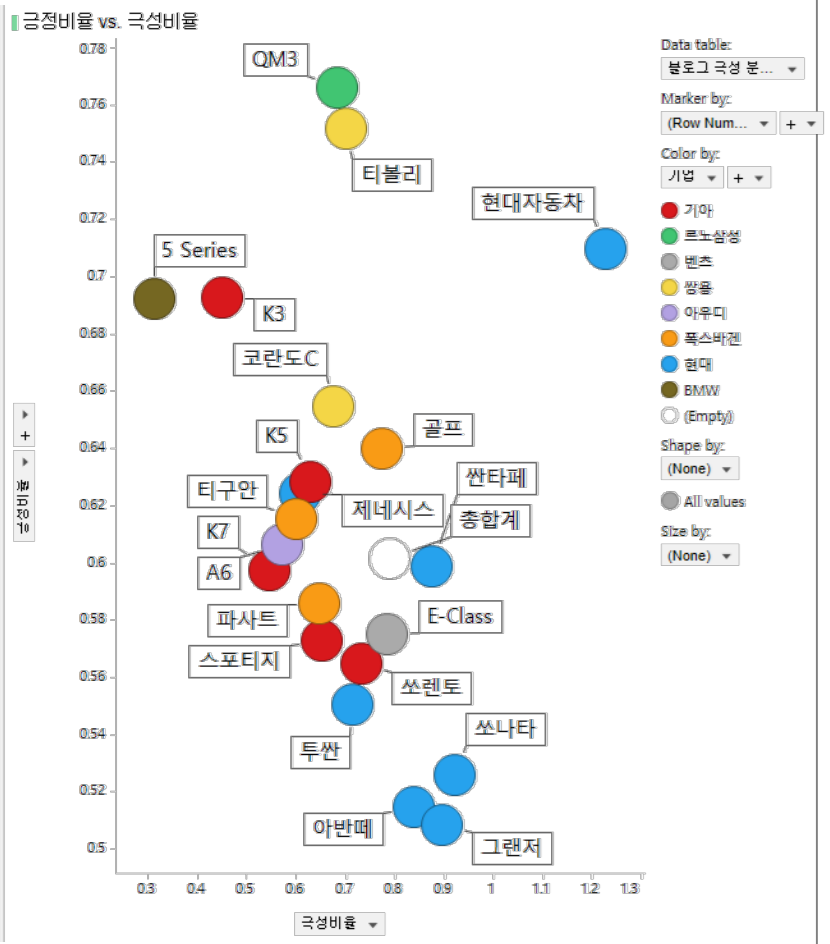
긍정

부정

Good

Bad





Data Science Applications

중앙은행

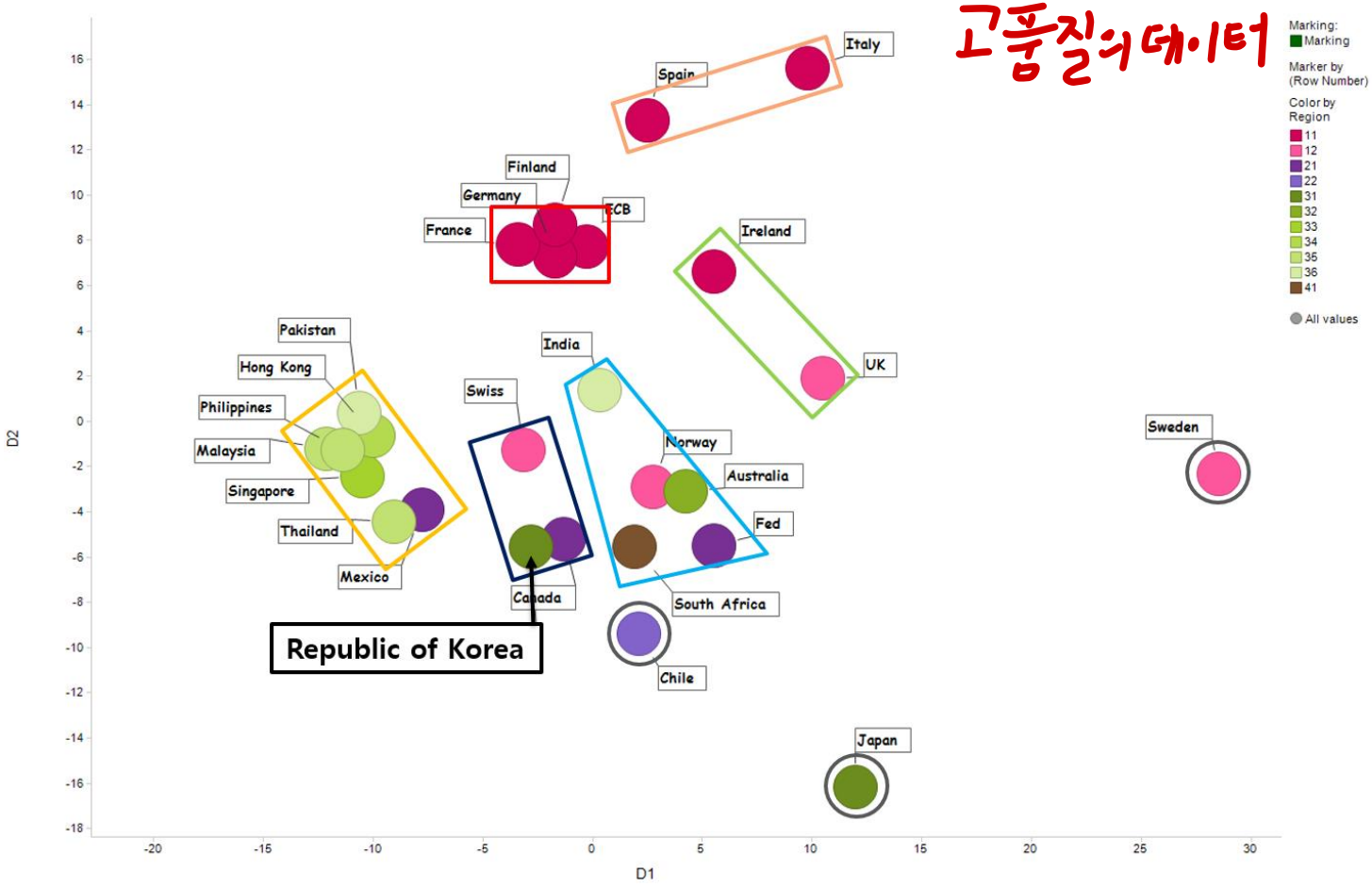
전세계 중앙은행 총재 연설문 분석

1

Visualization for intuitive understanding

각국의 통화정책을 고찰

고품질의 데이터



Data Science Applications

중앙은행

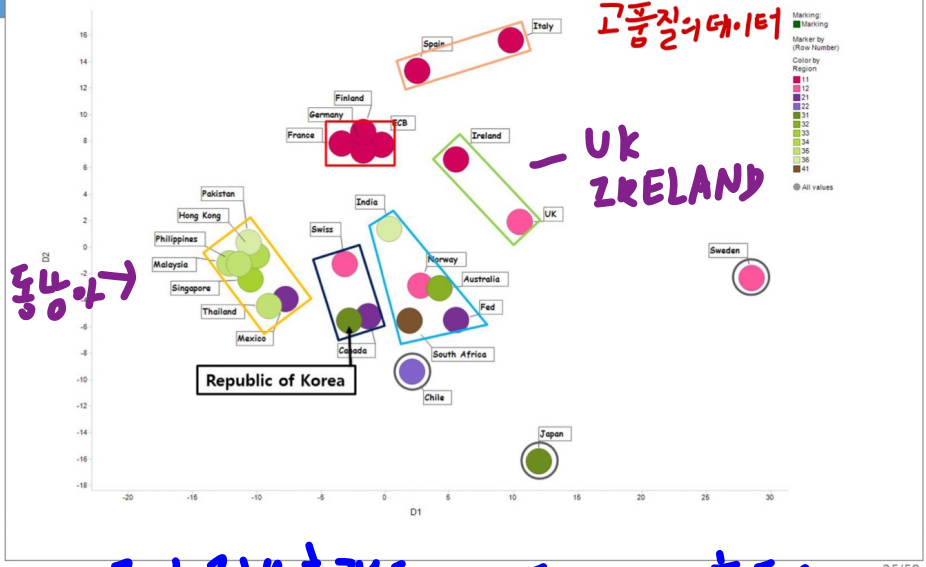
전세계 중앙은행 총재 연설문 분석

1

Visualization for intuitive understanding

각국의 통화정책을 관장

고품질의 데이터



UK IRELAND

동남아

각국의 중앙은행 총재들이 어떤 단어, 어휘들을 시간의 흐름에 따라서 많이 사용했는지를 분석을 해보면 각국의 통화정책 방향성에 유사성을 살펴볼 수 있다.

10년치 기준분석을 했을때 일본과 스위덴은 독자적인 통화 정책을 사용했다.

연도별로 각각 중앙은행들이 특정시점에 어떠한 단어들을 중점적으로사용
Data Science Applications
했는지를 살펴보면 해당하는 국가가 해당시점에 무엇을 목표로 삼는지 파악

1

Visualization for intuitive understanding

Year	Common Concern	Federal Reserve System	European Central Bank	Deutsche Bundesbank	Bank of England	Bank of Japan
2004	Sustainability Credibility	Expansion Imports	Parliaments Cooperation	Retirement Ages Working Hours	Household-Spending	QE Deflation
2005	China Inflation	Deficits Competitive	Financial Integration	Global Imbalance	Households Future Inflation	QE Recession
2006	Competitive Global Imbalance	Incentives Risk Taking	Administered Price Indirect Taxes	Inflation	China / India Commodities	Domestic and-External Demand
2007	Subprime-Mortgage	Subprime-Mortgage	Price Stability Turmoil	Banking Supervision Disclosure	Credit	Subprime-Mortgage
2008	Financial Turmoil Commodity Prices	Financial Turmoil Funding Markets	Financial Turmoil Liquidity	Financial Turmoil Subprime	Commodity Prices Housing Market	Securitized Product
2009	Financial Crisis Lehman Brothers	Financial Crisis ABS	Non-standard-Measure	Financial Crisis Rescue	Asset Purchase Recovery	Credit Bubble Financial Crisis
2010	Recovery Reform	Recovery Recession	ESRB/ FSB Deficits	Microprudential Macroprudential	Recovery/ QE VAT/ TAX	Deflation
2011	Sovereign Debt Basel III	Dodd-Franc Act Recovery	Sovereign Debt EFSF	Debt Crisis Basel III	Commodity Prices Basel III	Asset Purchase ETFs / REITs
2012	Europe Deleveraging	Recovery (has been) Labor Market	OMT/ ESM/ SSM Fragmentation	Banking Union Taxpayer	Investment-Banking	European Debt Deleveraging
2013	Real Economy Price Stability	(At least as long as) Unemployment	SRM/ SSM/ OMT	SRM / SSM Banking Union	Prudential-Regulation	Price Stability QE

Data Science Applications

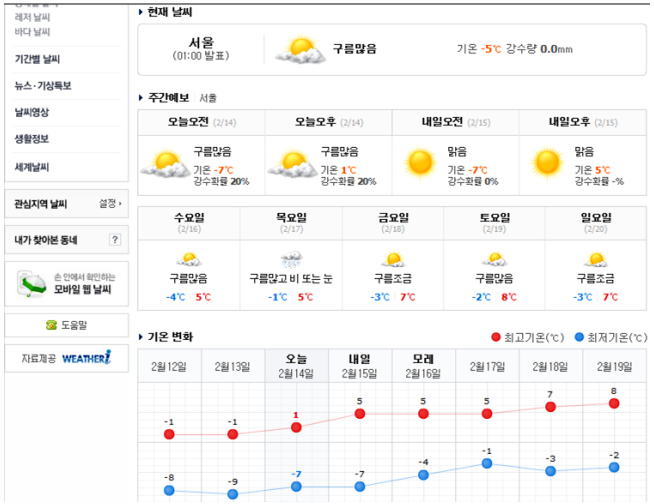
이래 예측/탐색, 진단

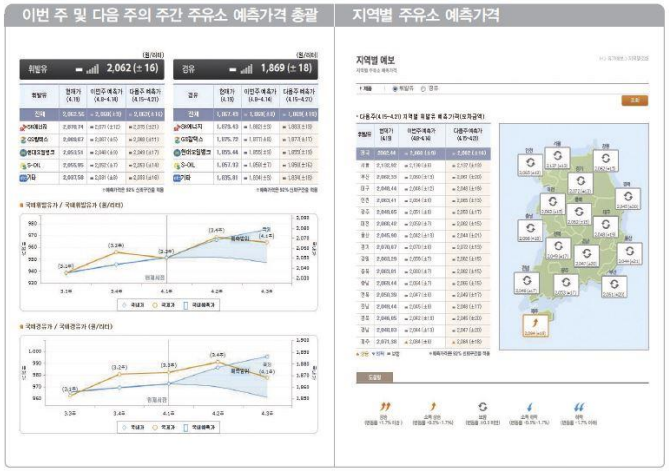
전부 과거실적에 관련된 데이터 이용해 이래예측

2

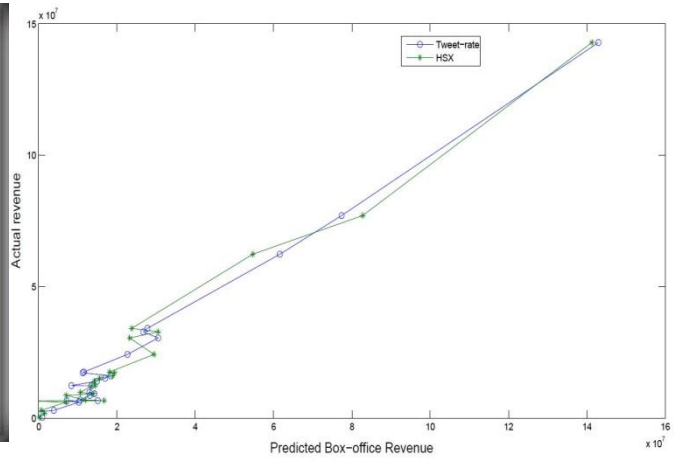
Predict, Diagnosis, and Detection

주식/경제지표/날씨





자료 : Opinet 유가정보서비스



Asur and Huberman (2010) Predicting the Future with Social Media, WI-IAT10: 492-499

삼성전자

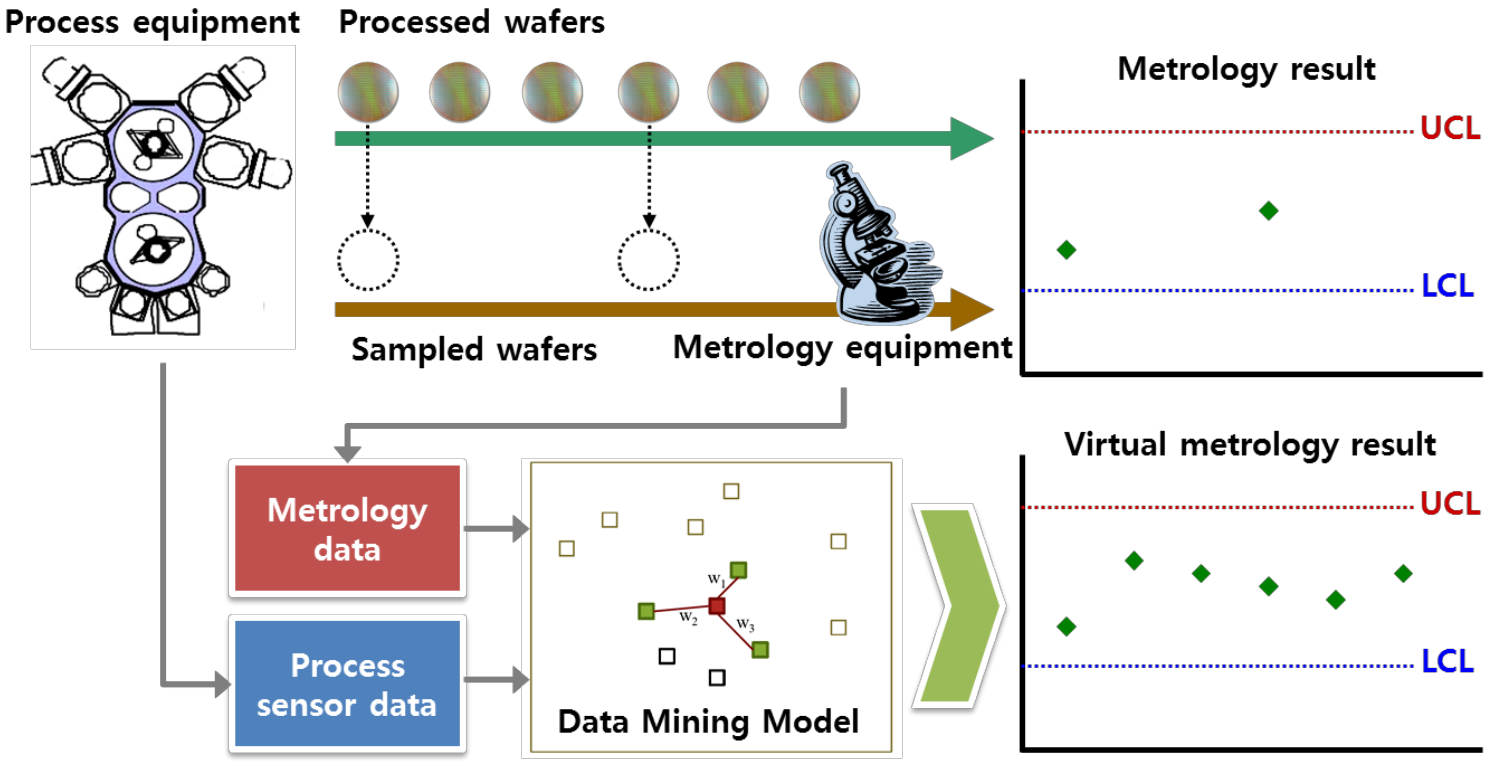
Data Science Applications

가상계측

2

Predict, Diagnosis, and Detection

Virtual Metrology in Semiconductor Manufacturing



Kang et. al. (2009) A virtual metrology system for semiconductor manufacturing, ***Expert Systems with Applications*** **36**(10): 12554-12561.

삼성전자.

Data Science Applications

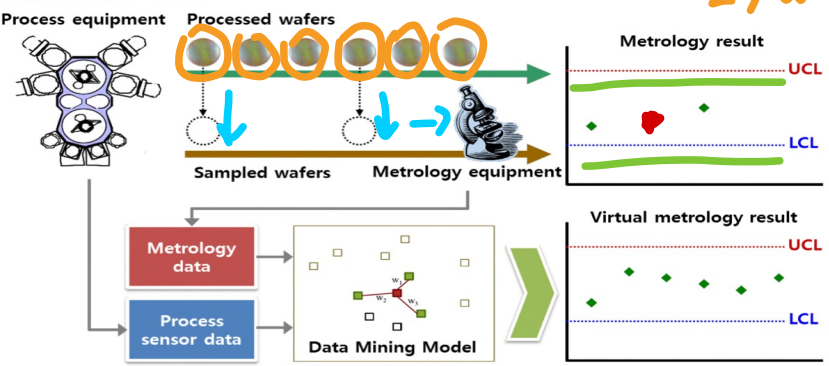
가상계측

2

LOT = 25 wafer 묶음

Predict, Diagnosis, and Detection

Virtual Metrology in Semiconductor Manufacturing



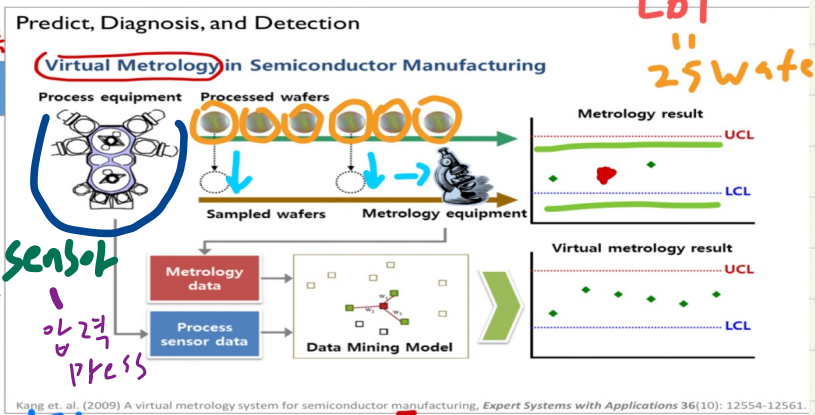
Kang et. al. (2009) A virtual metrology system for semiconductor manufacturing, *Expert Systems with Applications* 36(10): 12554-12561.

→ 산업공학에서 품질관리, 통계적 sampling에 기반한 품질관리를 배운다.

과거의 전통적 품질관리는 장비로부터 어떠한 이반도체 공정, wafer들이 하나씩 이렇게 가공이 되고난 뒤에는 해당하는 wafer들이 특정한 품질지표의 상한/하한을 만족시켰는지 안시켰는지 평가하기 위해서 sampling

※ sampling된 wafer들을 계측장비에 투입해서 예측수행 그랬을때 해당하는 wafer가 정상영역에 위치하게 되면 품질지표가 해당 wafer를 sampling한 나머지집단에 대해서는 전부 통과, 그렇지 않으면 문제가 발생하면 해당 wafer가 추출이 된 LOT, → 해당하는 LOT에 포함된 wafer들을 재가공하거나 다시 검사하거나 rollback을 시킨다.

삼성전자



LOT = 25 wafer 묶음

예상 계측

2

Sensor

temp 온도

압력 press

→ 산업공학에서 품질관리, 통계적 sampling에 기반한 품질관리를 배운다.

과거의 전통적 품질관리는 장비로부터 어떠한 이반도체 공정, wafer들이 하나씩 이렇게 가공이 되고난 후에는 해당하는 wafer들이 특정한 품질지표의 상한/하한을 만족시켰는지 안시켰는지 평가하기 위해서 sampling

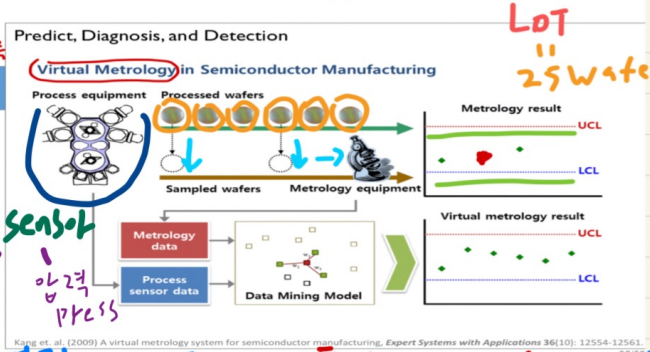
※ sampling된 wafer들을 계측장비에 투입해서 예측수행 그랬을때 해당하는 wafer가 정상범위에 위치하게 되면 품질지표가 해당 wafer를 sampling한 나머지집단에 대해서는 전부 통과, 그렇지 않으면 문제가 발생하면 해당 wafer가 추출이 된 LOT, → 해당하는 LOT에 포함된 wafer들을 재가공하거나 다시 검사하거나 rollback을 시킨다.

※ 현장에 있던 엔지니어분들이 생각한 것은 기본적으로 wafer가 잘 가공이 되고있는지 그렇지 아닌지에 대한정보는 그림.

※ 실질적인 반도체 공장 FAB에서는 이장비가 최소해도 수십억단위 장비투입. 장비자체에 센서가 부착

※ 센서 데이터(온도, 압력)을 전부 수집하면 분명히 이번 wafer가 가공이 잘되던 시점에서 반도체 설비가 작동이 되는대로의 설계 recipe 대로 실질적으로 동작을 하지 않았을것이다. 그러면 결국 process의 또는 장비의 설비 sensor data를 입력 X로 놓고 계측했던 정보인 metrology data(과거 계측했던 품질지표)를 Y로 놓고 $Y=f(x)$ 2개사이의 관련성을 찾아내는 예측회귀식

삼성전자



Kang et. al. (2009) A virtual metrology system for semiconductor manufacturing, *Expert Systems with Applications* 36(10): 12554-12561.

LOT = 25 wafer 묶음

센서 temp 온도 압력 press

→ 산업공학에서 품질관리, 통계적 sampling에 기반한 품질관리를 배운다.

과거의 전통적 품질관리는 장비로부터 어떠한 이반도체 공정, wafer들이 하나씩 이렇게 가공이 되고난 후에는 해당하는 wafer들이 특정한 품질지표의 상한/하한을 만족시켰는지 안시켰는지 평가하기 위해서 sampling

sampling된 wafer들을 계측장비에 투입해서 예측수행 그랬을때 해당하는 wafer가 정상범위에 위치하게 되면 품질지표가 해당 wafer를 sampling한 나머지집단에 대해서는 전부 통과, 그렇지 않으면 문제가 발생하면 해당 wafer가추출이 된 LOT. → 해당하는 LOT에 포함된 wafer들을 재가공하거나 다시 검사하거나 rollback을 시킨다.

현장에 있던 엔지니어들이 생각한 것은 기본적으로 wafer가 잘 가공이 되고있는지 그렇지 아닌지에 대한 정보는 그림.

실질적인 반도체 공장 FAB에서는 이장비가 저렴해도 수십억단위 장비투입. 장비자체에 센서가 부착

센서 데이터(온도,압력)을 전부 수집하면 불명의 이런 wafer가 가공이 잘되던 시점에서 반도체 설비가 작동이 되는대로의 설계 recipe 대로 실질적으로 동작을 하지 않았을것이다. 그러면 결국 process의 또는 장비의 설비 sensor data를 입력X로 놓고 계측했던 전후인 metrology data(과거 계측했던 품질지표)를 Y로 놓고 $Y=f(x)$ 2개사이의 관련성을 찾아내는 예측모형수립

실질적으로는 wafer에 대해서 전수품질검사를 하지 않아도 예측모델이 어느정도 정확하다면 모든 wafer에 대한 품질지표 값을 측정할수 있지 않는가에 대한 가정

Data Science Applications

추천시스템 (의사결정)

3

Support decision making in everyday life (recommendation system)

Roll over image to zoom in

Apple iPad Pro (11-inch, Wi-Fi, 64GB) - Space Gray (Latest Model)

by Apple

★★★★☆ 129 customer reviews | 141 answered questions

List Price: ~~$799.00~~
Price: $699.99
You Save: $99.01 (12%)

In Stock.

This item does not ship to **Seoul, Korea; Republic of (South Korea)**. Please check other sellers who may ship internationally. Learn more
Ships from and sold by Amazon.com.

Style: **Wi-Fi**

Wi-Fi | Wi-Fi + Cellular

Color: **Space Gray**

Size: **64GB**

1TB | 64GB | 256GB | 512GB

- 11-Inch edge-to-edge Liquid Retina display with Promotion, true Tone, and wide Color
- A12X Bionic chip with Neural Engine
- Face ID for secure authentication and Apple Pay
- 12MP back camera, 7MP True Depth front camera
- Four speaker Audio with wider Stereo sound
- 802. 11AC Wi-Fi and gigabit-class LTE cellular data
- Up to 10 hours of battery life

Show more

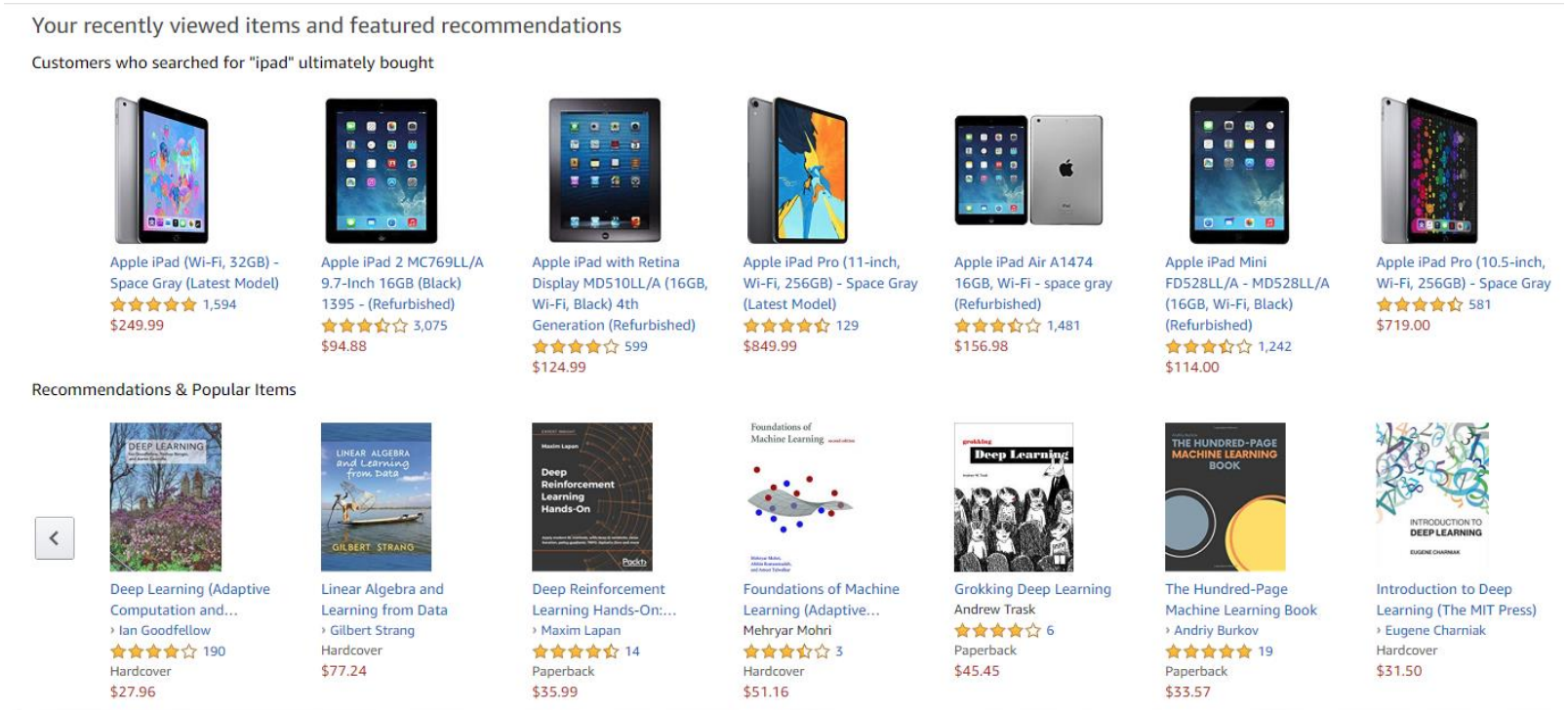
Jump to: Compare devices | Technical details

This product has a serial number that uniquely identifies the item. Should your order go missing before it arrives, Amazon may register the serial number with loss and theft databases to prevent fraudulent use or resale of the item.

Data Science Applications

3

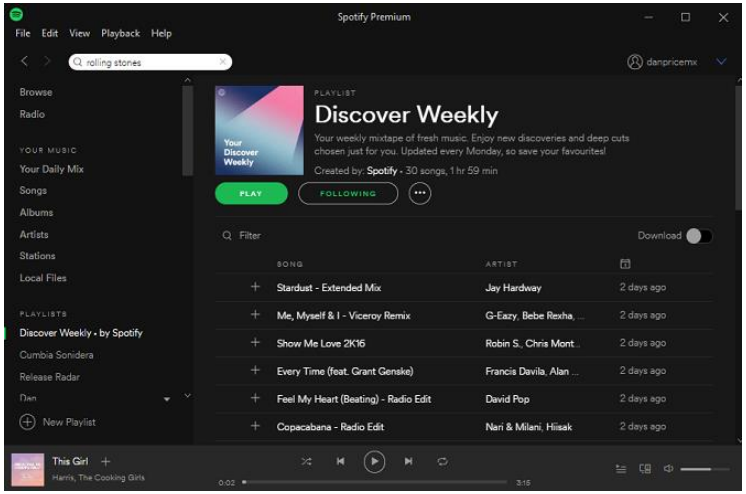
Support decision making in everyday life (recommendation system)

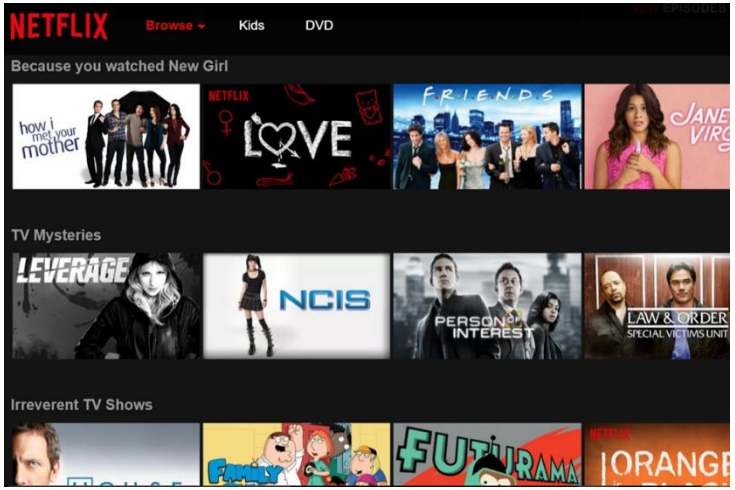


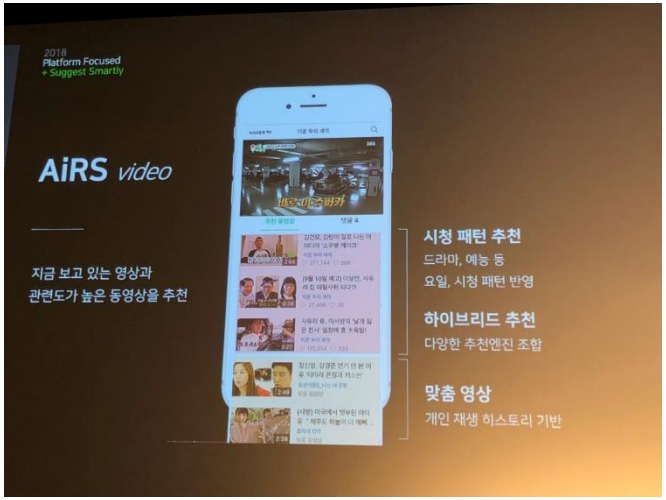
Data Science Applications

3

Support decision making in everyday life (recommendation system)







AGENDA

데이터 사이언스 분석 절차 또는 분석 방법론에 있어서는
다변량 데이터 분석의 방법을 탐색

(5개 Categories)

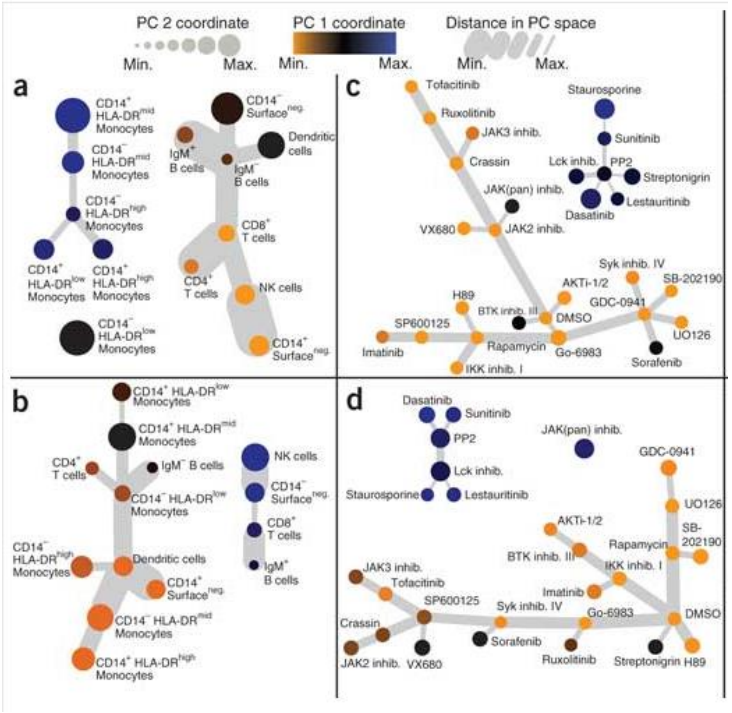
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

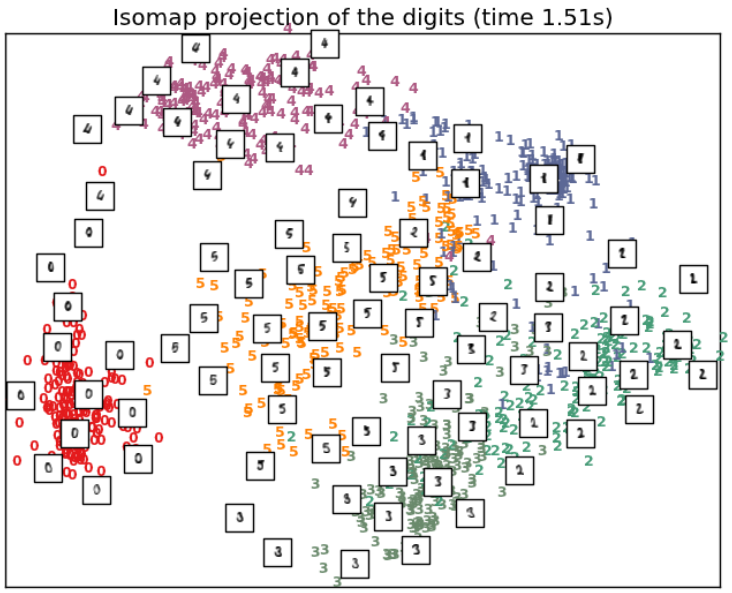
The phenomenon being studied is represented as simply as possible without sacrificing valuable information.

It is hoped that this will make interpretation easier.

Principal Component Analysis



Variable Reduction



http://www.nature.com/nbt/journal/v30/n9/fig_tab/nbt.2317_F6.html

(5개 Categories)

Multivariate Data Analysis for Data Science

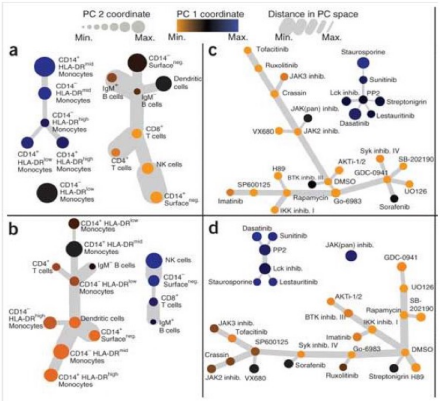
Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

데이터 축소

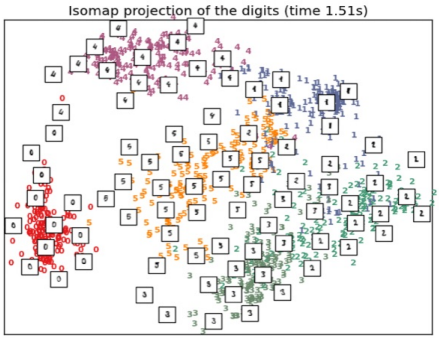
The phenomenon being studied is represented as simply as possible without sacrificing valuable information.

It is hoped that this will make interpretation easier. 해석이 훨씬 쉬워진다.

Principal Component Analysis



Variable Reduction



http://www.nature.com/nbt/journal/v30/n9/fig_tab/nbt.2317_F6.html

→ 주어진 데이터들을 이용해서 어떻게 하면 데이터가 가지고 있는 본질적인 특징을 최대한 보존하면서 훨씬 더 차원이 작은 데이터셋으로 변환할 수 있을까? * 중요한 정보는 잃지 않고 최대한 단순하게 표현하자.

↳ 변수 축소기법 PCA

Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

The phenomenon being studied is represented as simply as possible without sacrificing valuable information.

It is hoped that this will make interpretation easier.

- Applications
  - ✓ Using data on several variables related to cancer patient responses to radio-therapy, a simple measure of patient response to radiotherapy was constructed
  - ✓ Track records from many nations were used to develop an index of performance for both male and female athletes
  - ✓ Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions
  - ✓ Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants

Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

The phenomenon being studied is represented as simply as possible without sacrificing valuable information.
It is hoped that this will make interpretation easier.

- Applications
  - ✓ Using data on several variables related to cancer patient responses to radio-therapy, a simple measure of patient response to radiotherapy was constructed
  - ✓ Track records from many nations were used to develop an index of performance for both male and female athletes
  - ✓ Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions
  - ✓ Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants

→ Application 같은 경우에는 데이터를 축소하고 단순화하는것이 어디에 사용될수 있는가?

→ 암 환자 데이터에 대해 치료제 및 생존율을 판단하는데 있어서 중요한 변수가 무엇인지 찾아내는것

→ 올림픽 참가한 여러 국가들에 대해서 여러가지 경제적인 지표를 사용해서 해당하는 국가별로 index를 생성하는데도 만들기 별수있다.

유사한것끼리 grouping

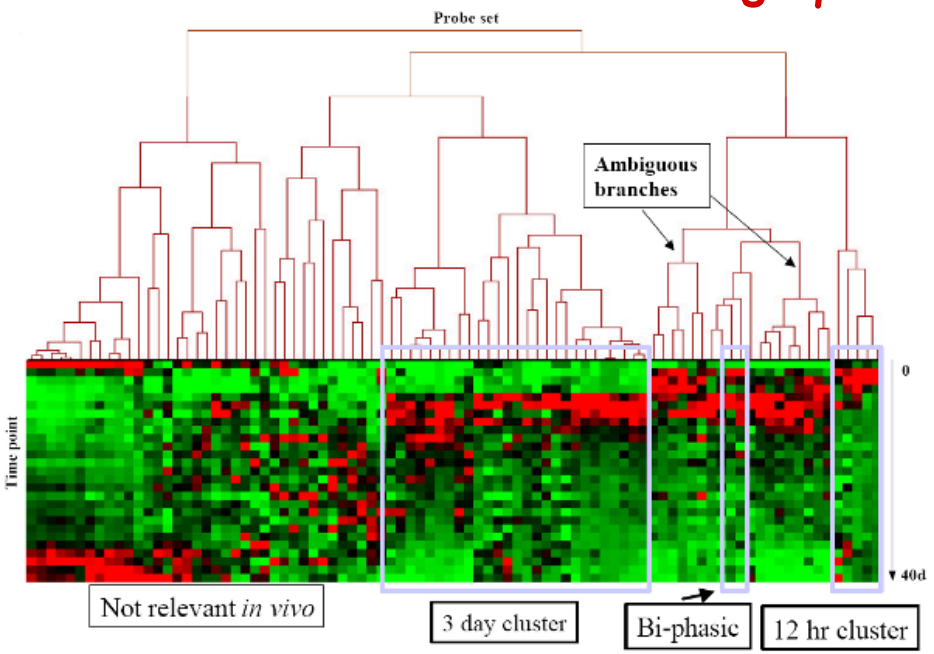
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	**Sorting and Grouping**	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

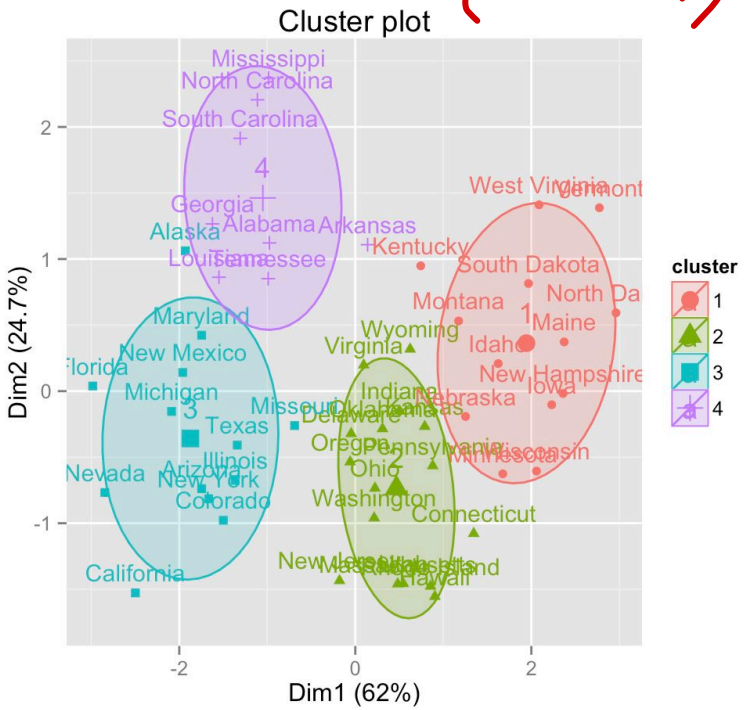
Groups of "similar" objects or variables are created, based upon measured characteristics.

Alternatively, rules for classifying objects into well-defined groups may be required.

Clustering: Hierarchical Clustering 계층적



Clustering: K-Means Clustering (K-means)



유사한것끼리 grouping

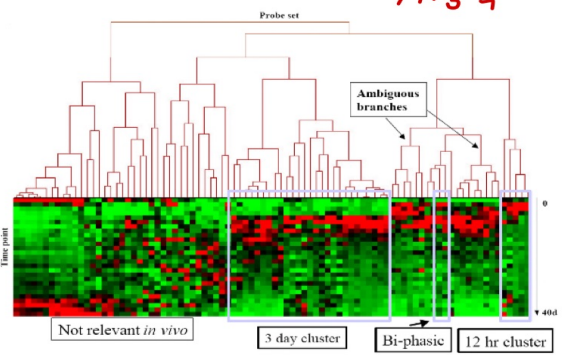
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	**Sorting and Grouping**	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

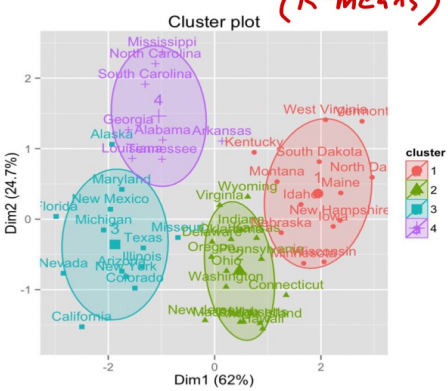
Groups of "similar" objects or variables are created, based upon measured characteristics.

Alternatively, rules for classifying objects into well-defined groups may be required.

Clustering: Hierarchical Clustering 계층적



Clustering: K-Means Clustering (K-means)



→ 데이터에 대해 유사한 변수들을 묶어낸다

→ 비슷한 개체들을 하나의 군집으로 살펴보고 정의를 하고 각각의 서로 다른 군집들은 어떠한 차이가 있는지 살펴본다.

Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	**Sorting and Grouping**	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

Groups of "similar" objects or variables are created, based upon measured characteristics.

Alternatively, rules for classifying objects into well-defined groups may be required.

- Applications
  - ✓ Data in several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing computer utilization
  - ✓ Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from nonalcoholics
  - ✓ Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease

→ 사용자들이 어떠한 패턴을 보이는지 (유사한 패턴)

→ 의심가는 패턴 (정상패턴)과 비교

Multivariate Data Analysis for Data Science

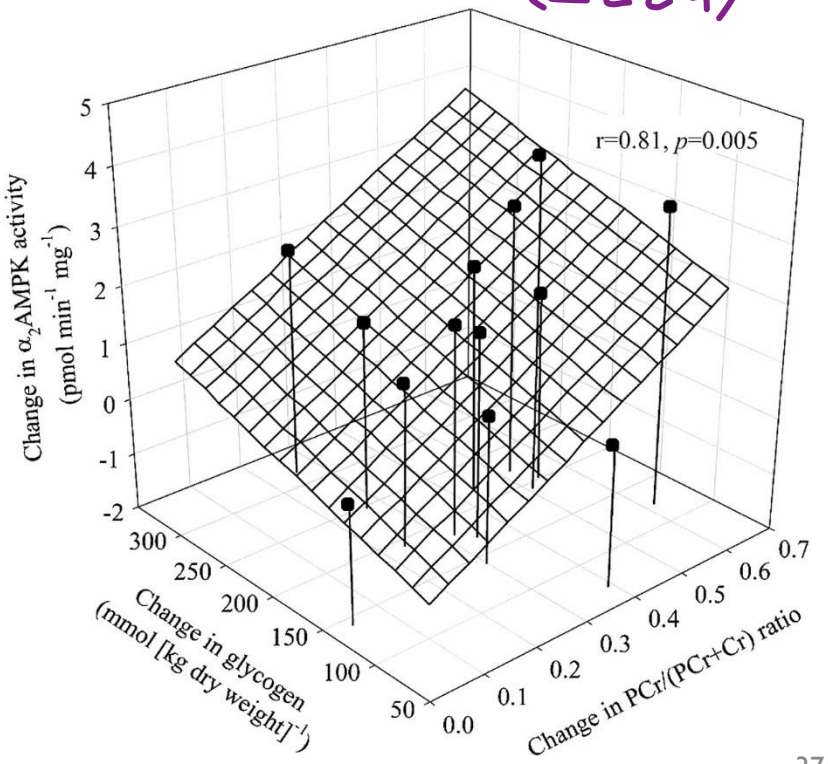
Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

The nature of the relationships among variables is of interest

Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?

Association Rule Mining

Factor Analysis



Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	**Investigation of the dependence among variables**	Prediction	Hypothesis construction and testing

The nature of the relationships among variables is of interest

Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?

- Applications
  - ✓ Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.
  - ✓ Measurements of variables related to innovation, on the on hand, and variables related to the business environment and business organization, on the other hand, were used to discover why some firms are product innovators and some firms are not.
  - ✓ The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to asses the relation between risk-taking behavior and performance

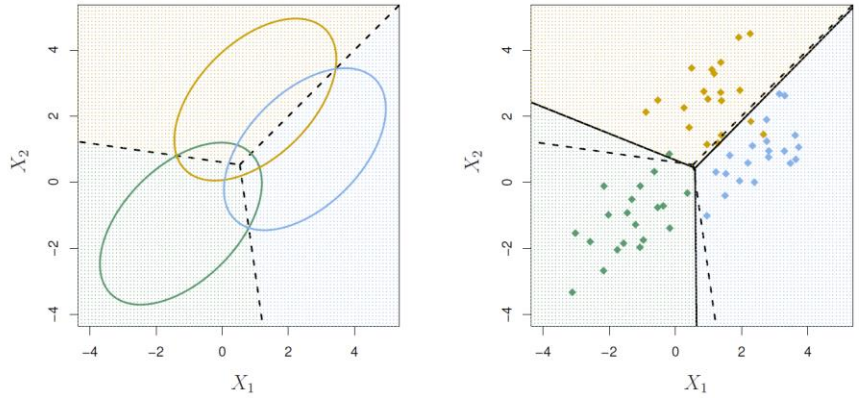
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	**Prediction**	Hypothesis construction and testing

예측

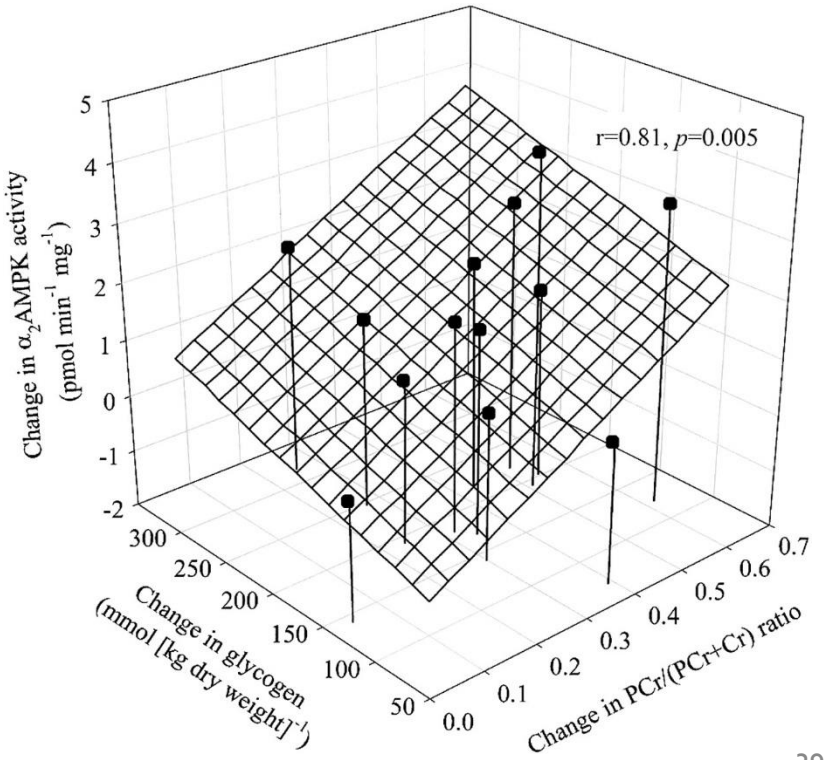
Relationships between variables must be determined for the purpose of predicting the value of one or more variables on the basis of observations on the other variables.

Discrimination and Classification → 분류



Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.
The dashed lines are known as the *Bayes decision boundaries*.
Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Multivariate Linear Regression



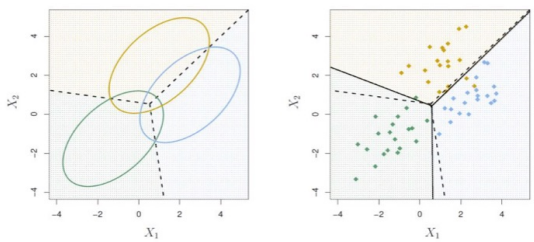
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

예측

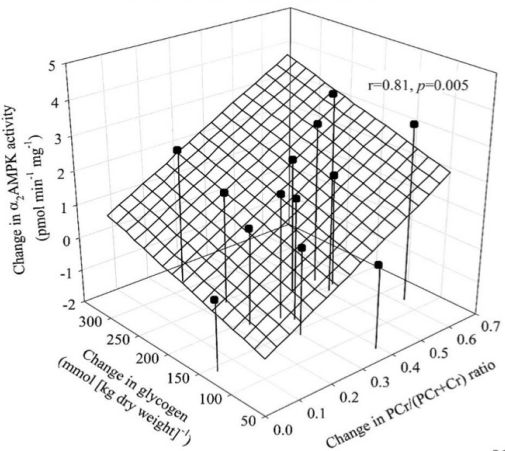
Relationships between variables must be determined for the purpose of predicting the value of one or more variables on the basis of observations on the other variables.

Discrimination and Classification → 분류



Here $\pi_1 = \pi_2 = \pi_3 = 1/3$.
The dashed lines are known as the *Bayes decision boundaries*.
Were they known, they would yield the fewest misclassification errors, among all possible classifiers.

Multivariate Linear Regression



39/58

① 예측

→ 어떠한 목적을 가지고 하나의변수들에 대해서 하나의 변수를 다른 변수들의 관측치로부터 예측을 한다

→ 어떠한 객체가 특정한 범주에 속하는지 아닌지를 예측하게되면 분류문제 (새로운 이미지분류)

② 회귀 → (실수값을 예측하는것)

→ 새로운 고객을 유치했을때 이고객이 우리회사에 평생동안 생애가치로서 얼마정도 수익이 될것인지 확인

Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	**Prediction**	Hypothesis construction and testing

Relationships between variables must be determined for the purpose of predicting the value of one or more variables on the basis of observations on the other variables.

- Applications
  - ✓ The associations between test scores, and several high school performance variables, and several college performance variables were used to develop predictors of success in college
  - ✓ Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments
  - ✓ Measurements on several accounting and financial variables were used to develop a method for identifying potentially insolvent property-liability insurers

예시 읽어보기

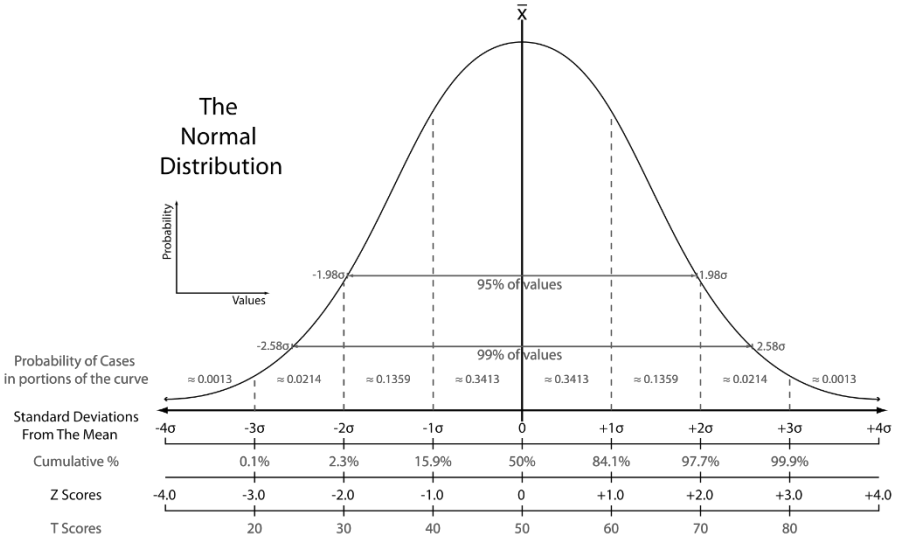
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

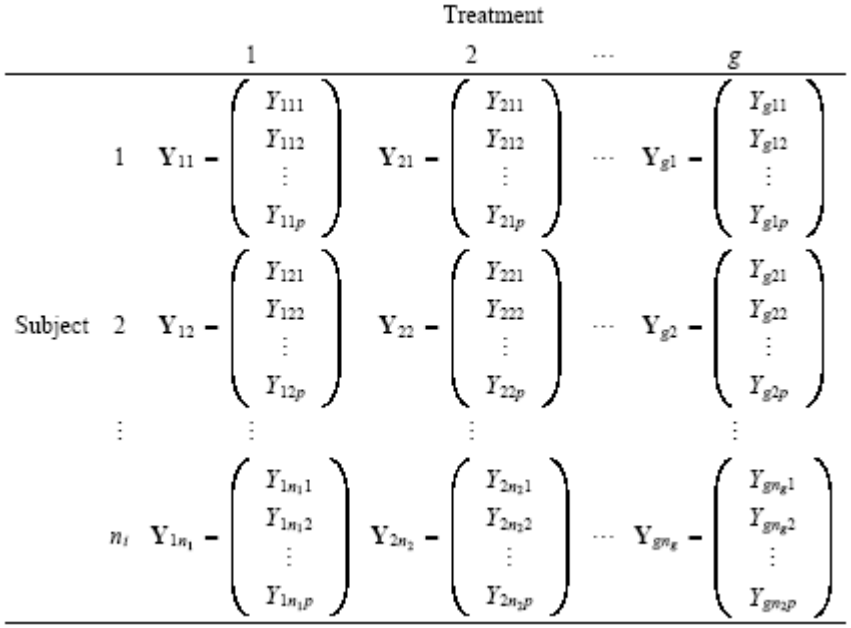
Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested.

This may be don to validate assumptions or to reinforce prior convictions.

Inferences about a Mean Vector



Comparisons of Several Multivariate Means



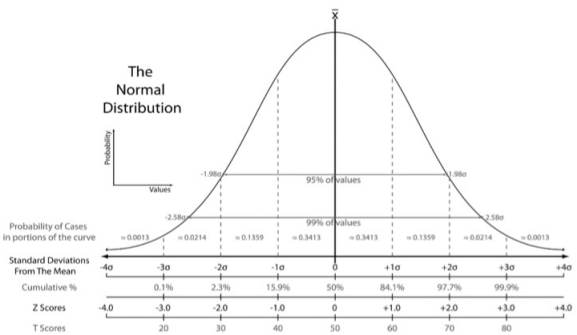
Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested.

This may be don to validate assumptions or to reinforce prior convictions.

Inferences about a Mean Vector



Comparisons of Several Multivariate Means

$$\begin{array}{cc|ccccc} & & \text{Treatment} & & & & \\ & & 1 & & 2 & \cdots & g \\ \hline 1 & & \mathbf{Y}_{11} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{11p} \end{pmatrix} & & \mathbf{Y}_{21} = \begin{pmatrix} Y_{211} \\ Y_{212} \\ \vdots \\ Y_{21p} \end{pmatrix} & \cdots & \mathbf{Y}_{g1} = \begin{pmatrix} Y_{g11} \\ Y_{g12} \\ \vdots \\ Y_{g1p} \end{pmatrix} \\ \text{Subject} & 2 & \mathbf{Y}_{12} = \begin{pmatrix} Y_{121} \\ Y_{122} \\ \vdots \\ Y_{12p} \end{pmatrix} & & \mathbf{Y}_{22} = \begin{pmatrix} Y_{221} \\ Y_{222} \\ \vdots \\ Y_{22p} \end{pmatrix} & \cdots & \mathbf{Y}_{g2} = \begin{pmatrix} Y_{g21} \\ Y_{g22} \\ \vdots \\ Y_{g2p} \end{pmatrix} \\ \vdots & & \vdots & & \vdots & & \vdots \\ n_i & & \mathbf{Y}_{1n_1} = \begin{pmatrix} Y_{1n_1 1} \\ Y_{1n_1 2} \\ \vdots \\ Y_{1n_1 p} \end{pmatrix} & & \mathbf{Y}_{2n_2} = \begin{pmatrix} Y_{2n_2 1} \\ Y_{2n_2 2} \\ \vdots \\ Y_{2n_2 p} \end{pmatrix} & \cdots & \mathbf{Y}_{gn_g} = \begin{pmatrix} Y_{gn_g 1} \\ Y_{gn_g 2} \\ \vdots \\ Y_{gn_g p} \end{pmatrix} \end{array}$$

※가설 → 1조/2조 오류

→ 평균벡터에 대한 가설검정/여러평균벡터에 비교/분산에 대한 가설검정

Multivariate Data Analysis for Data Science

Data Reduction/ Structural Simplification	Sorting and Grouping	Investigation of the dependence among variables	Prediction	Hypothesis construction and testing

Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested.

This may be don to validate assumptions or to reinforce prior convictions.

- Applications
  - ✓ Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends
  - ✓ Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores
  - ✓ Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories

AGENDA

데이터 사이언스 수행절차

*학술적 측면에서는 재미있고 의미있는 문제면 충분/비즈니스 관점에서는 효율향상

Data Science Procedure (5단계)

① 질문을 던지는 단계 ② 데이터수집 ③ 데이터 탐색 ④ 모델링 ⑤ 내용적용/시각화

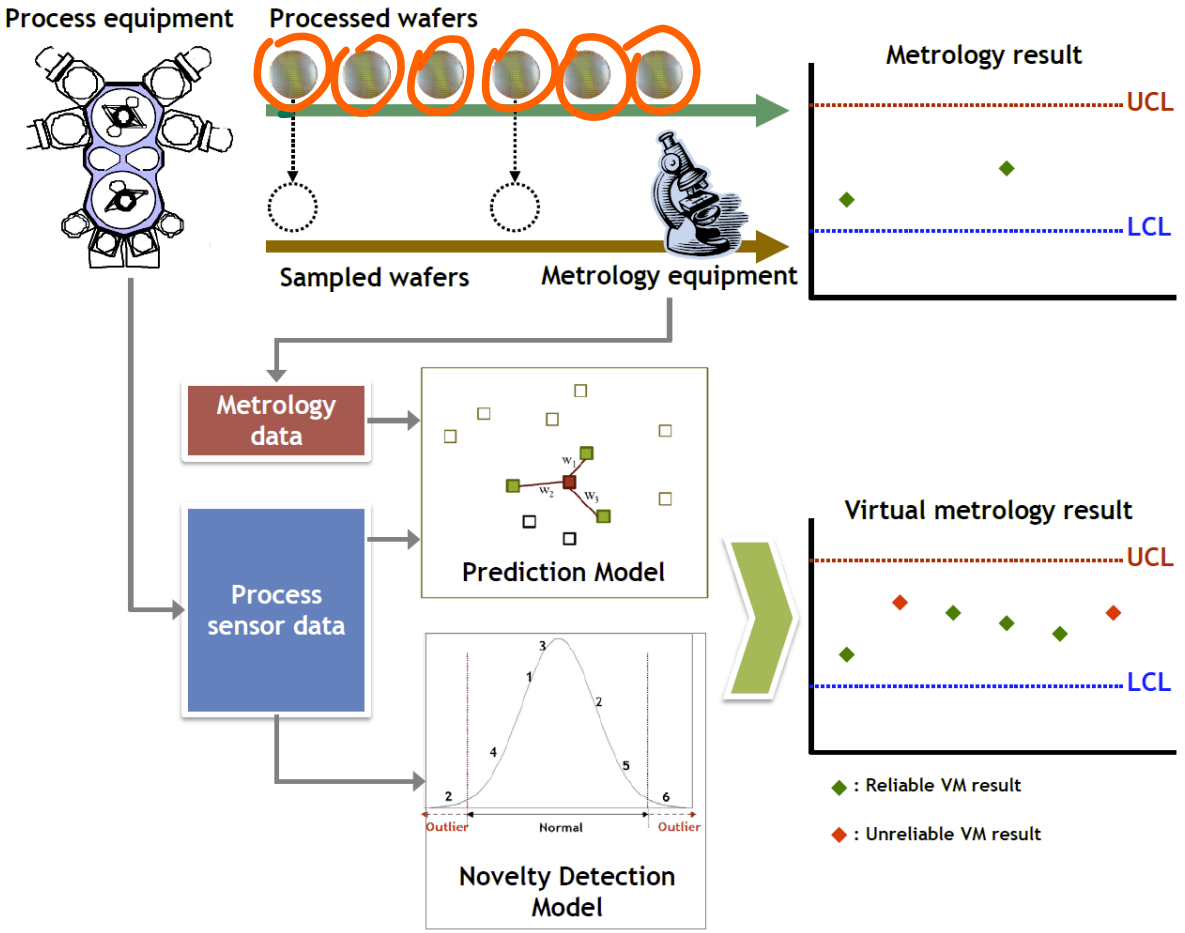
Ask an interesting question	Get the data	Explore the data	Model the data	Communicate and visualize the results
▸ 풀고자 하는 문제가 무엇인가?	▸ 데이터는 어떻게 샘플링할 것인가?	▸ 데이터를 그려보며 데이터의 속성과 구조를 알아보기	▸ 모델 수립	▸ 결과 요약 및 시사점 분석
▸ 만약 관련해 모든 데이터를 보유하고 있다면 무엇을 할 것인가?	▸ 어떤 데이터와 정보가 우리 목표와 관련이 있는가?	▸ 데이터에서 이상한 점은 없는가?	▸ 모델 적합화	▸ 결과가 타당한가?
▸ 무엇을 예측하고 추정하기를 원하는가?	▸ 프라이버시나 개인정보 이슈는 없는가?	▸ 데이터에 어떠한 패턴이 존재하는가?	▸ 모델 검증	▸ 스토리를 말할 수 있는가? (전략 수립)

ML/데이터분석 알고리즘에서는 학습데이터를 통해서 충분히 학습한 경우에 신뢰

Ask an Interesting Question

흥미로운 질문을 해라. → 우리가 무엇인가 질문을 던졌을때 그질문이 해결이 되게 되면 시스템/운영조직이 조율이 될수있는가?

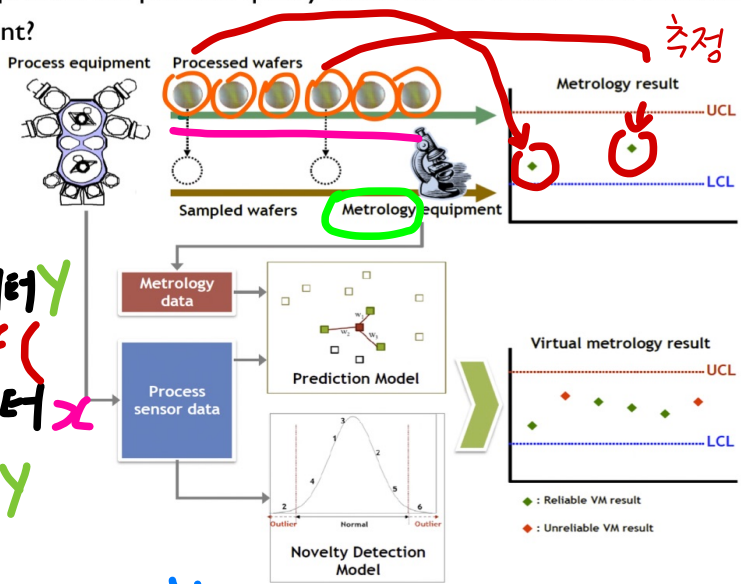
- Step 1: Ask an interesting question
  - ✓ Can we predict the product quality based on the sensor data collected from equipment?



Ask an Interesting Question

★ 흥미로운 질문을 해라. → 우리가 무엇인가 질문을 던졌을때 그질문이 해결이 되게 되면 시스템/운영조직에 도움이 될수있는가?

- Step 1:Ask an interesting question
  - ✓ Can we predict the product quality based on the sensor data collected from equipment?



실제측데이터 Y
f (
센서데이터 X
f(x) = y

45/58

→ 반도체 설비에서 해당하는 반도체 들이 이렇게 쭉 가공이 되는데 가공이 되는 반도체들의 공정이 너무 정밀하다보니 정밀한 공정에서 wafer들이 제대로 가공이 되었는지 아닌지를 판단하기 위한 장치로써 계측이라는것을 수행

→ metrology를 수행하게 되면 각각의 개별적인 wafer들이 정상범위내에 들어왔는지를 측정을 한다.

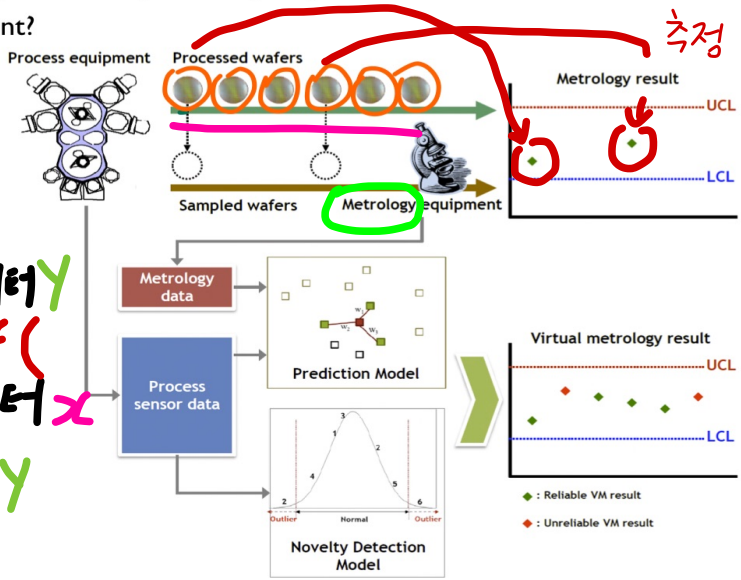
★문제는 정상으로 계측이 되지않은 wafer들은 과연 어떻게 해당하는 품질지표를 알수있을것인가?

★위 질문에 대답을 할수 있다면 실질적인 계측을 수행하지 않고 개별적인 wafer의 품질정보를 알수있다. 생산성 향상에 도움이 됨

Ask an Interesting Question

★ 흥미로운 질문을 해라. → 우리가 무엇인가 질문을 던졌을때 그질문이 해결이 되게되면 시스템/운영조직이 도움이 될수있는가?

- Step 1:Ask an interesting question
  - ✓ Can we predict the product quality based on the sensor data collected from equipment?



실제측데이터 y
f(
센서데이터 x
f(x) = y

→ 계측된 센서데이터에 대해서 과거 우리가 배웠던 센서데이터 기술로 봤을때 일반적인 형태 데이터 인지 아닌지 그렇지 않아서 계측의 신뢰도가 낮아질것인지 이런 고백하는 포돌

Ask an Interesting Question

text분석

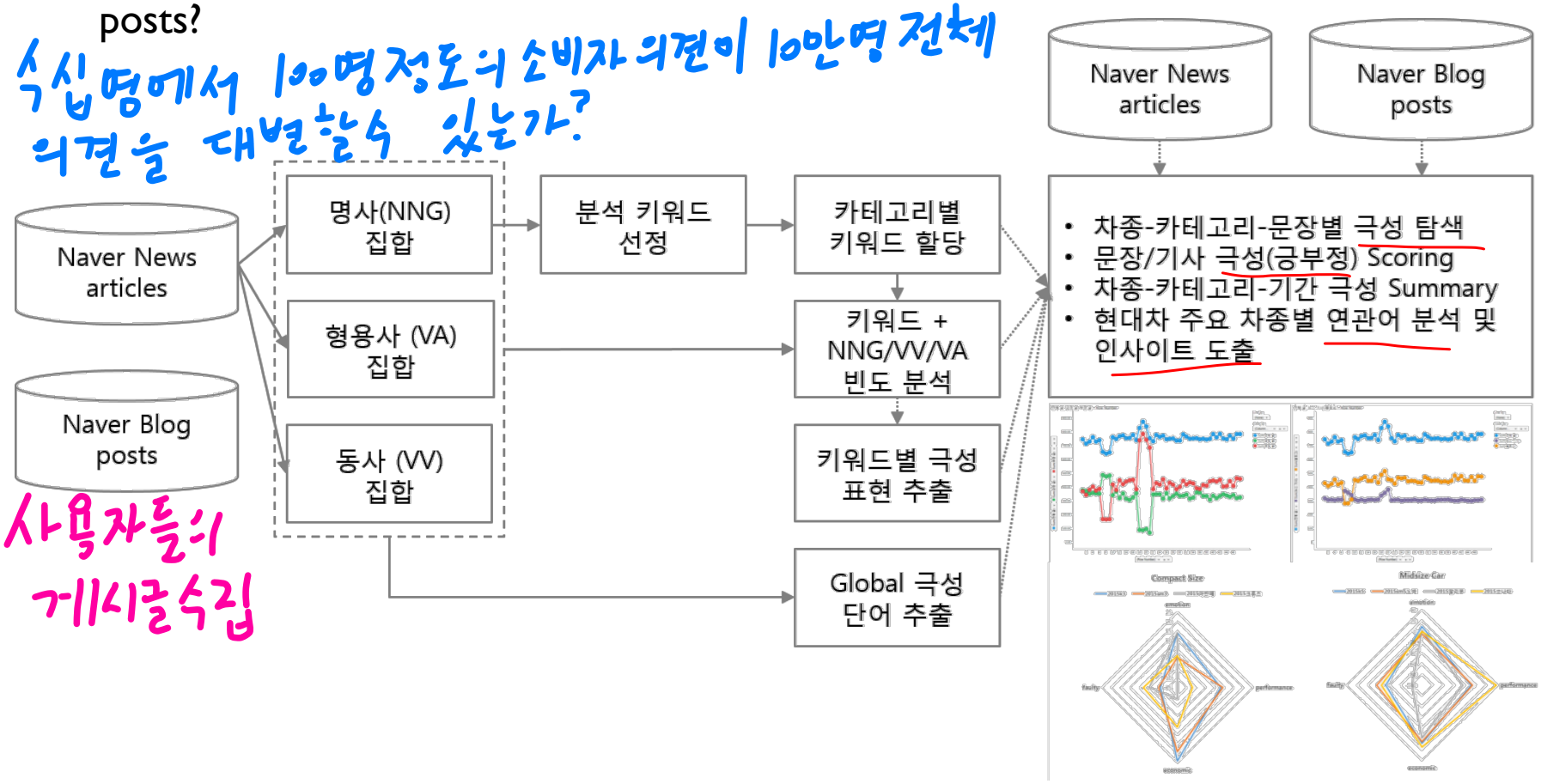
자동차 각각의 차종들에 대해서 만족도 분석

① 고객이 주로 언급하는 내용

② 긍/부정

- Step 1: Ask an interesting question
  - ✓ Is it possible to understand customers' preference based on news articles and blog posts?

수십명에서 100명정도의 소비자 의견이 10만명 전체 의견을 대변할수 있는가?

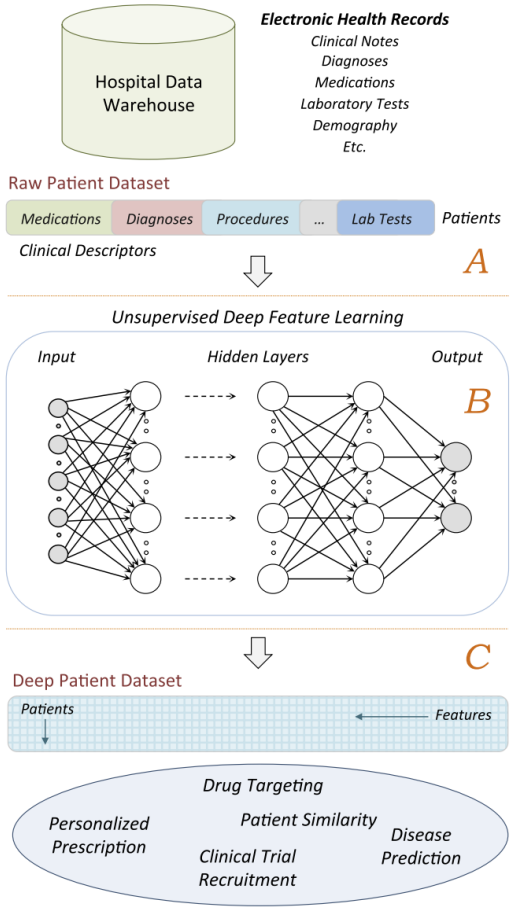


Ask an Interesting Question

★논문에서 해결하고자 던졌던 문제점

- Step 1: Ask an interesting question
  - ✓ Can we predict various diseases based on the electric health records (EHR)?

(전자의료기록)



	Area under the ROC curve		
Disease	**RawFeat**	**PCA**	**DeepPatient**
Diabetes mellitus with complications	0.794	0.861	**0.907**
Cancer of rectum and anus	0.863	0.821	**0.887**
Cancer of liver and intrahepatic bile duct	0.830	0.867	**0.886**
Regional enteritis and ulcerative colitis	0.814	0.843	**0.870**
Congestive heart failure (non-hypertensive)	0.808	0.808	**0.865**
Attention-deficit and disruptive behavior disorders	0.730	0.797	**0.863**
Cancer of prostate	0.692	0.820	**0.859**
Schizophrenia	0.791	0.788	**0.853**
Multiple myeloma	0.783	0.739	**0.849**
Acute myocardial infarction	0.771	0.775	**0.847**
Personality disorders	0.787	0.788	**0.846**
Inflammatory conditions of male genital organs	0.659	0.825	**0.841**
Endometriosis	0.697	0.765	**0.839**
Inflammatory diseases of female pelvic organs	0.714	0.799	**0.830**
Cancer of ovary	0.646	0.788	**0.824**
Sickle cell anemia	0.567	0.689	**0.822**
Nephritis, nephrosis and renal sclerosis	0.763	0.775	**0.821**
Cancer of bladder	0.711	0.744	**0.818**
Chronic kidney disease	0.764	0.758	**0.814**
Cancer of testis	0.508	0.771	**0.811**
Menopausal disorders	0.681	0.772	**0.808**
Delirium, dementia and amnestic (and other) cognitive disorders	0.728	0.720	**0.803**
Peritonitis and intestinal abscess	0.689	0.747	**0.801**
Cardiac arrest and ventricular fibrillation	0.711	0.747	**0.799**
Developmental disorders	0.705	0.737	**0.798**

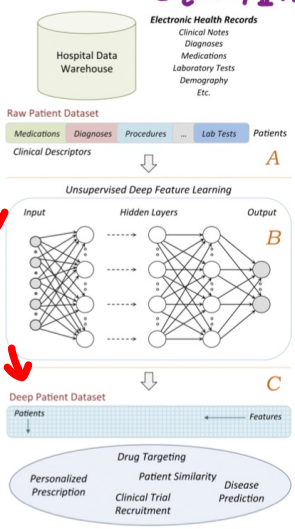
Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6.

Ask an Interesting Question

※논문에서 해결하고자 던졌던 문제점

- Step 1:Ask an interesting question
  - ✓ Can we predict various diseases based on the electric health records (EHR)?

(전자의료기록)



Disease	Area under the ROC curve RawFeat	PCA	DeepPatient
Diabetes mellitus with complications	0.794	0.861	**0.907**
Cancer of rectum and anus	0.863	0.821	**0.887**
Cancer of liver and intrahepatic bile duct	0.830	0.867	**0.886**
Regional enteritis and ulcerative colitis	0.814	0.843	**0.870**
Congestive heart failure (non-hypertensive)	0.808	0.808	**0.865**
Attention-deficit and disruptive behavior disorders	0.730	0.797	**0.863**
Cancer of prostate	0.692	0.820	**0.859**
Schizophrenia	0.791	0.788	**0.853**
Multiple myeloma	0.783	0.739	**0.849**
Acute myocardial infarction	0.771	0.775	**0.847**
Personality disorders	0.787	0.788	**0.846**
Inflammatory conditions of male genital organs	0.659	0.825	**0.841**
Endometriosis	0.697	0.765	**0.839**
Inflammatory diseases of female pelvic organs	0.714	0.799	**0.830**
Cancer of ovary	0.646	0.788	**0.824**
Sickle cell anemia	0.567	0.689	**0.822**
Nephritis, nephrosis and renal sclerosis	0.763	0.775	**0.821**
Cancer of bladder	0.711	0.744	**0.818**
Chronic kidney disease	0.764	0.758	**0.814**
Cancer of testis	0.508	0.771	**0.811**
Menopausal disorders	0.681	0.772	**0.808**
Delirium, dementia and amnestic (and other) cognitive disorders	0.728	0.720	**0.803**
Peritonitis and intestinal abscess	0.689	0.747	**0.801**
Cardiac arrest and ventricular fibrillation	0.711	0.747	**0.799**
Developmental disorders	0.705	0.737	**0.798**

-확률

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6.

→보통 병원에가면 전부 의료보험의 다기본적으로 국민에게 제공
언제 어떤 병원에서 어떤처방/치료를 받았는지 기록

→이러한 기록들을 모두 한꺼번에 통합을 해서 예측모델에 투입을 하면 해당하는 일반적인 환자가 주요질병들에 대해서 얼마만큼의 확률로 질병진단을 받을것인지 예측

기존방법론들에 비해서 DL기법을 쓸때 상대적으로 5~10% 이상높은 정확도를 보여줌

Q)지금까지의 진료기록을 가지고 한국사람이 특정한질병의 진단을 받을지 아 특정한 질병에 걸릴것인지 예측

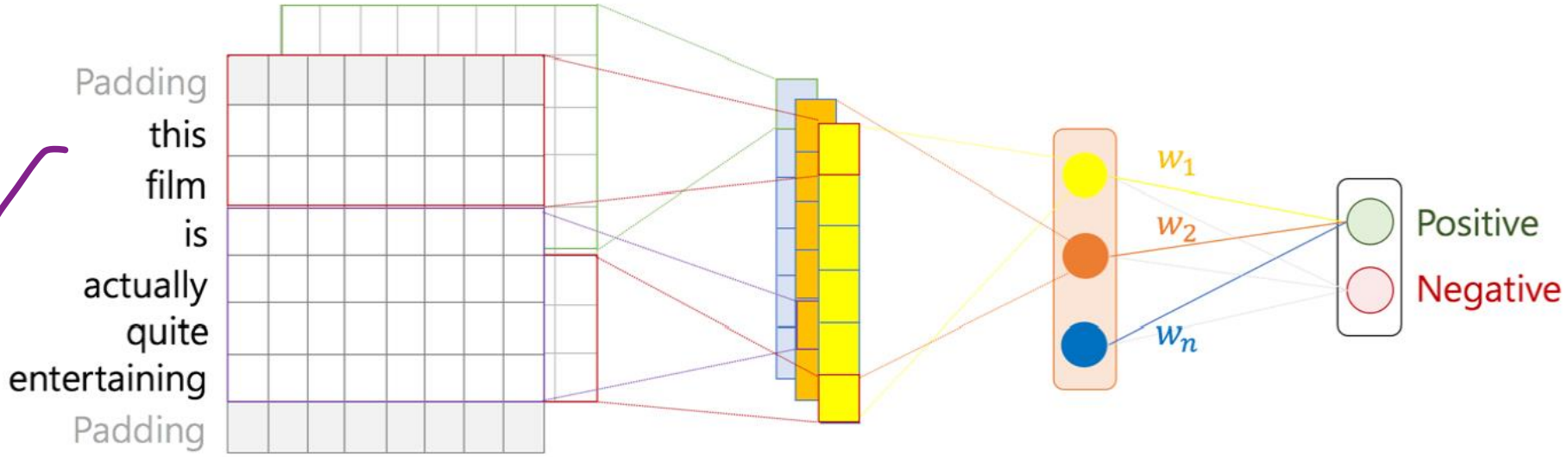
A) 예측을 할수있다면 데이터를 가지고 시도 ← interesting question

Ask an Interesting Question

✱ 어떤 문장이 들어왔을 때 해당하는 음절들이 이 문장이 과연 긍/부정인지 뉘앙스를 살펴본다.

- Step 1: Ask an interesting question

  ✓ Can we find the emotional expressions automatically from review texts?

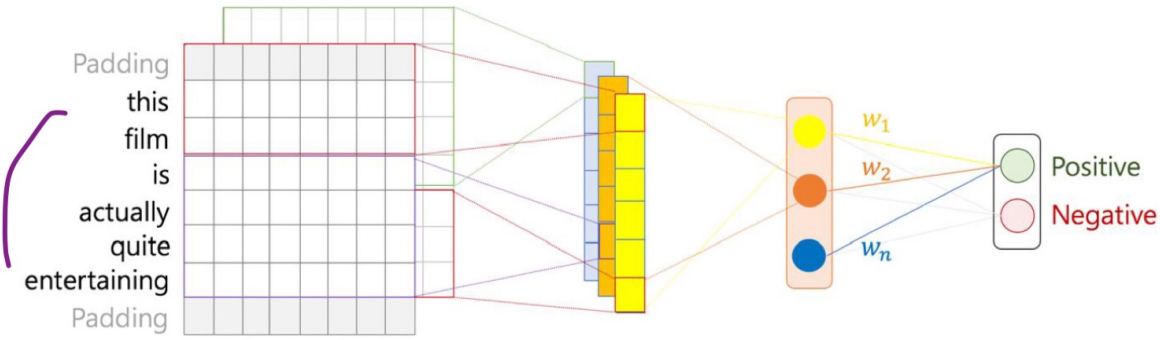


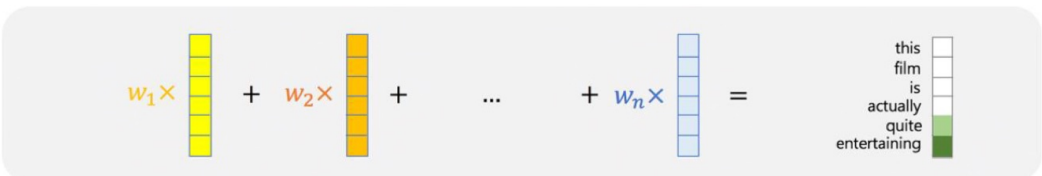
Ask an Interesting Question

✖ 어떤 문장이 들어왔을 때 해당하는 문장들이 이문장이 과연 긍정인지 부정인지 뉘앙스를 살펴본다.

- Step 1:Ask an interesting question

  ✓ Can we find the emotional expressions automatically from review texts?





→ 어떠한 단어를 때문에 긍정이라고 표현했는지 역으로 추적할수 있는 mechanism

Ask an Interesting Question

- Step 1: Ask an interesting question
  - ✓ Can we find the emotional expressions automatically from review texts?

Method	Sentence
Raw text	One of the funniest most romantic and most musical movies ever; definitely worth renting/buying especially if you have a taste for older style of cinematography. The animals and the songs alone will make you smile while watching the movie. A definite must for Madonna fans. :o) *(10 / 10 points)*
Rand	One of the **the funniest most romantic** and musical movies ever definitely worth renting buying especially if you have a taste for older style cinematography The animals songs alone will make smile while watching movie A definite must Madonna **fans** **Positive**
Static	One of the **funniest most romantic and** musical movies ever definitely worth renting buying especially if you have a taste for older style cinematography The animals songs alone will make smile while watching movie **A** definite must Madonna fans **Positive**
NStatic	One **of the funniest most** romantic and musical movies ever definitely worth renting buying especially if you have a taste for older style cinematography The animals songs alone will make smile while watching movie A definite must Madonna **fans** **Positive**
2ch	One **of the funniest most** romantic and musical movies ever definitely worth renting buying especially if you have a taste for older style cinematography The animals songs alone will make smile while watching movie A definite must Madonna **fans** **Positive**

4가지의 경우는 어떻게 입력변수를 구성했는지에 대한차이
주어진 문장은 긍정(positive)로써 예측모형에 의해판단

Ask an Interesting Question

- Step 1: Ask an interesting question
  - ✓ Can we find the emotional expressions automatically from review texts?

Method	Sentence
Raw text	This is one of the most boring films I've ever seen. The three main cast members just didn't seem to click well. Giovanni Ribisi's character was quite annoying. For some reason he seems to like repeating what he says. If he was the Rain Man it would've been fine but he's not. *(3 / 10 points)*
Rand	This is one of the most boring films I ve ever seen The three main cast members just didn t seem to click w ell Giovanni Ribisi s character was quite annoing For some reason he seems like repeating what says If Rain Man it **would been fine but he s** not **Negative**
Static	This is one of the **most** boring **films** I ve ever seen The three main cast members just didn t seem to click w ell Giovanni Ribisi s character **was quite annoying For** some reason he seems like repeating what says If Rai n Man it would been fine but he s not **Negative**
NStatic	This is one of **the most boring films** I ve ever seen The three main cast members just didn t seem to click w ell Giovanni Ribisi s character was quite **annoying For** some reason he seems like repeating what says If Rai n Man it wouldbeen fine but not **Negative**
2ch	This is one of the **most boring films** I ve ever seen The three main cast members just didn t seem to click w ell Giovanni Ribisi s character was **quite annoying For** some reason he seems like repeating what says If Rai n Man it would beenfine but not **Negative**
4ch	This is one of the **most boring films** I ve ever seen The three main cast members just didn t seem to click w ell Giovanni Ribisi s character was **quite annoying For** some reason he seems like repeating what says If Rai n Man it would been fine but not **Negative**

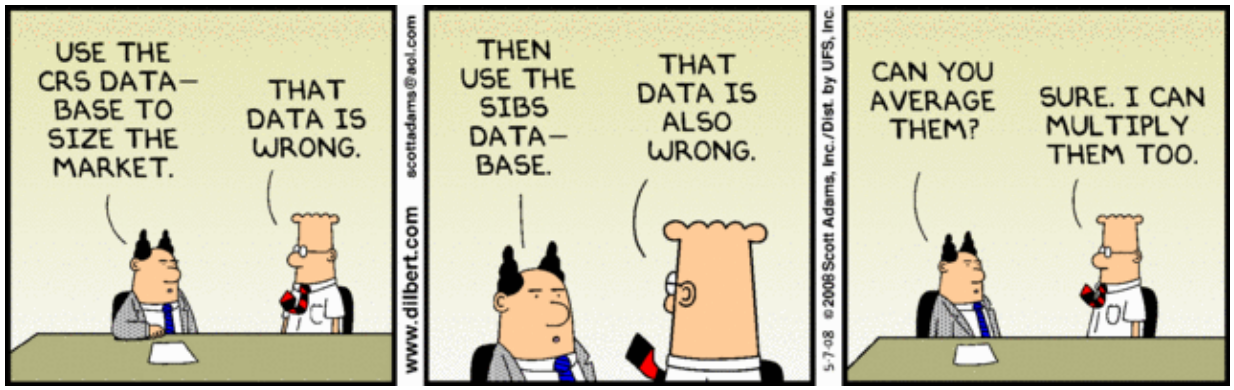
부정적인 평가는 어떤 부분을 보는가? (영화가 재미없다)

Get the Data (데이터 획득)

→ 뭔가 문제를 풀고자했을때 흥미로운질문 task를 정의

- Step 2: Get the Data ✔ 전체 시스템을 아울러서 여기에서 어떤문제를 해결해야 될것이다.
  - ✓ Garbage in, garbage out

데이터 정제



GIGO

  - ✓ The larger, the better

주어진문제를 푸는것은 누구나 잘할수있지만 어떤문제를 풀어야 될것인가를 정하는것은 굉장히 중요한 연습이 필요하다.

크면 클수록 좋다

"We don't have better algorithms than anyone else. We just have more data."

(가용 가능한 컴퓨팅 자원에서)

※ 같은 알고리즘이면 데이터가 많을수록 더우수한 퍼포먼스 낼수있다.

데이터작업
Data annotation
↳ start up business

Get the Data

- Step 2: Get the Data 어떤조직에서 문제를 풀기 위해 label이 없는데이터에 label을 달아야 하는작업 필요
  - ✓ Use domain knowledge from experts (especially when making answer sets)

도메인전문가 중요

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

IMPORTANCE Deep learning is a family of computational methods that allow an algorithm to program itself by learning from a large set of examples that demonstrate the desired behavior, removing the need to specify rules explicitly. Application of these methods to medical imaging requires further assessment and validation.

OBJECTIVE To apply deep learning to create an algorithm for automated detection of diabetic retinopathy and diabetic macular edema in retinal fundus photographs.

DESIGN AND SETTING A specific type of neural network optimized for image classification called a deep convolutional neural network was trained using a retrospective development data set of 128 175 retinal images, which were graded 3 to 7 times for diabetic retinopathy, diabetic macular edema, and image gradability by a panel of 54 US licensed ophthalmologists and ophthalmology senior residents between May and December 2015. The resultant algorithm was validated in January and February 2016 using 2 separate data sets, both graded by at least 7 US board-certified ophthalmologists with high intragrader consistency.

Editorial

Supplemental content

→ 안저영상에서 부터 당뇨성망막 병증이 있는지 없는지 파악하는 모델을 만든다.

↳ 13만장 image

→ 그만큼 좋은 데이터를 고품질 전문가로부터 확보한다.

논문의 결론은 오픈소스를 사용하면 안과전문의보다 새로운 데이터에 대한 판별력이 비슷하다.

Get the Data

Table. Baseline Characteristics[a]

Characteristics	Development Data Set	EyePACS-1 Validation Data Set	Messidor-2 Validation Data Set
No. of images	128 175	9963	1748
No. of ophthalmologists	54	8	7
No. of grades per image	3-7	8	7
Grades per ophthalmologist, median (interquartile range)	2021 (304-8366)	8906 (8744-9360)	1745 (1742-1748)
Patient demographics			

[a] Summary of image characteristics and available demographic information in the development and clinical validation data sets (EyePACS-1 and Messidor-2). Abnormal images were oversampled for the development set for algorithm training. The clinical validation sets were not enriched for abnormal images.

[b] Unique patient codes (deidentified) were available for 89.3% of the development set (n = 114 398 images).

[c] Individual-level data including age and sex were available for 66.1% of the development set (n = 84 734 images).

[d] Image quality was assessed for a subset of the development set.

[e] Referable diabetic retinopathy, defined as the presence of moderate and worse diabetic retinopathy and/or referable diabetic macular edema according to the International Clinical Diabetic Retinopathy Scale,[14] was calculated for each ophthalmologist before combining them using a majority decision. The 5-point grades represent the grade that received the highest number of votes for diabetic retinopathy alone. Hence, the sum of moderate, severe, and proliferative diabetic retinopathy for the 5-point grade differs slightly from the count of referable diabetic retinopathy images.

(데이터를 탐색하는 단계)

Explore the Data

※ 모델링하기 전에 데이터를 충분히 탐색하고 이해하는 단계

- Step 3: Explore the data before modeling 단변량 분석/통계량 분석과 같은 descriptive analytics 해당하는 방법 사용

(데이터를 탐색하는 단계)

Explore the Data

모델링하기 전에 데이터를 충분히 탐색하고 이해하는 단계

- Step 3: Explore the data before modeling

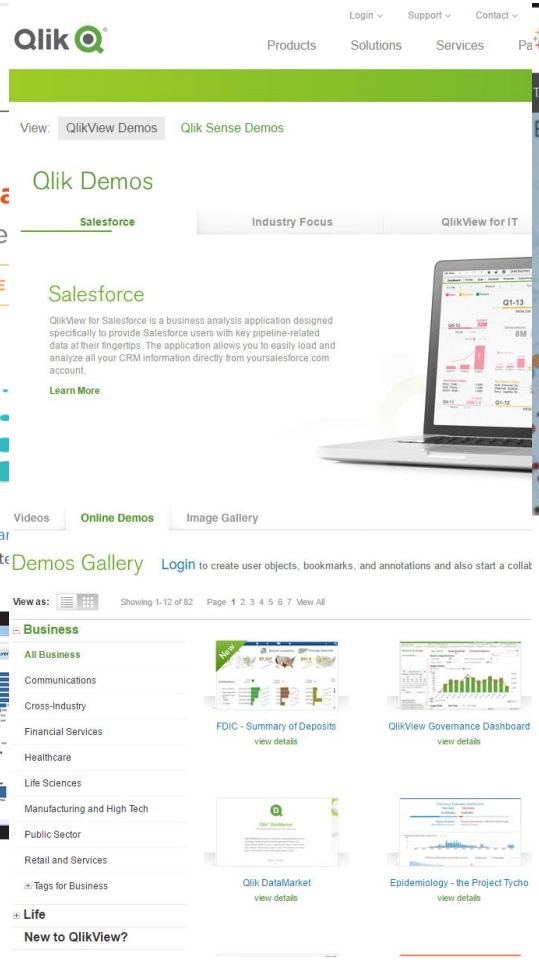
단변량 분석/통계량 분석과 같은 descriptive analytics 해당하는 방법 사용

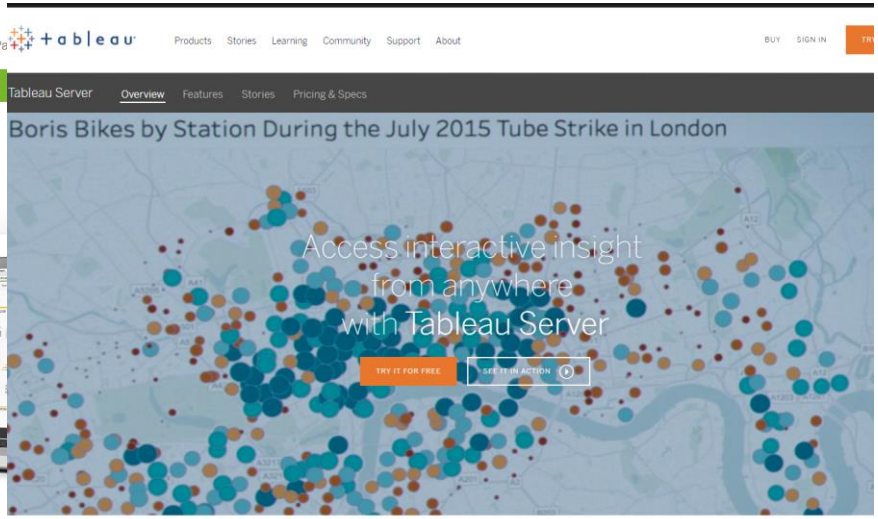
→ 데이터가 어떤 특성을 가지고 있는지를 모델러의 머리속에 숙지를 시킨 다음에 적절한 알고리즘을 적용해야지 만족할만한 결과를 얻을수 있다.

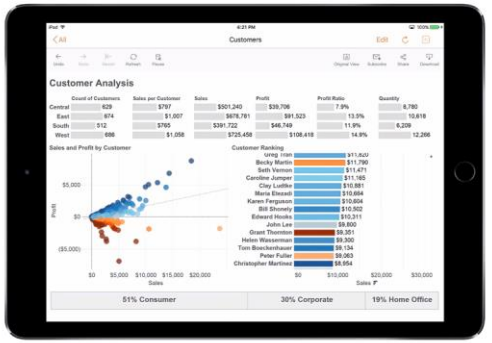
(다양한 예시들)

Explore the Data

- Step 3: Explore the data before modeling
  - ✓ Data visualization software can be helpful







Model the Data

- Step 4: Model the data
  - ✓ Select an appropriate algorithm based on the purpose of task, data characteristics, requirement for interpretability.

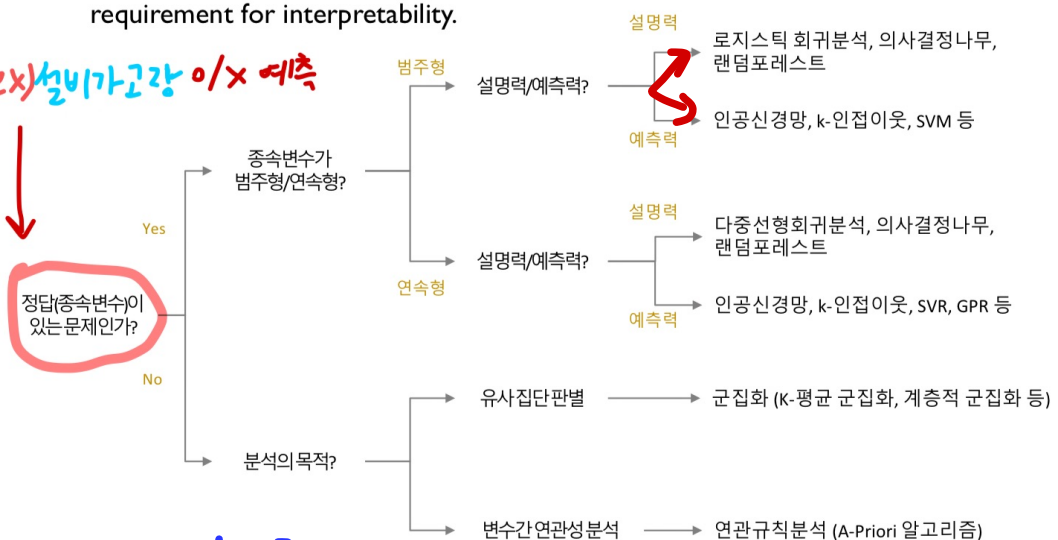
모델링

Model the Data

→ 내 똑같은 문제를 해결하기 위해서 여러가지 알고리즘을 배워야 하는가?

- Step 4: Model the data — 상황에 따라서 적합한 알고리즘이 다르다.
  - ✓ Select an appropriate algorithm based on the purpose of task, data characteristics, requirement for interpretability.

ex) 설비가고장 o/x 예측



주어진 데이터를 통해서 데이터가 생성된 분포를 찾아라.

원가 이상한 pattern을 찾아라. (정답이 없는 문제)

정답이 있을때도 해당하는 종속변수가 범주형인지 연속형인지에 따라서 사용한 알고리즘이 달라진다.

→ 연속형은 내일당장 S&P지수가 몇포인트 될것이냐? (숫자를 정확히 예측해야 하는 부분)

→ 과연 알고리즘이 설명력을 제공해야 하는가 우수한 예측력만 있으면 되는가? (AI에서는 예측력에 방점) → 질확률 최소화

→ 마지막 단계는 결과물을 가지고 실제 운영진과 communication

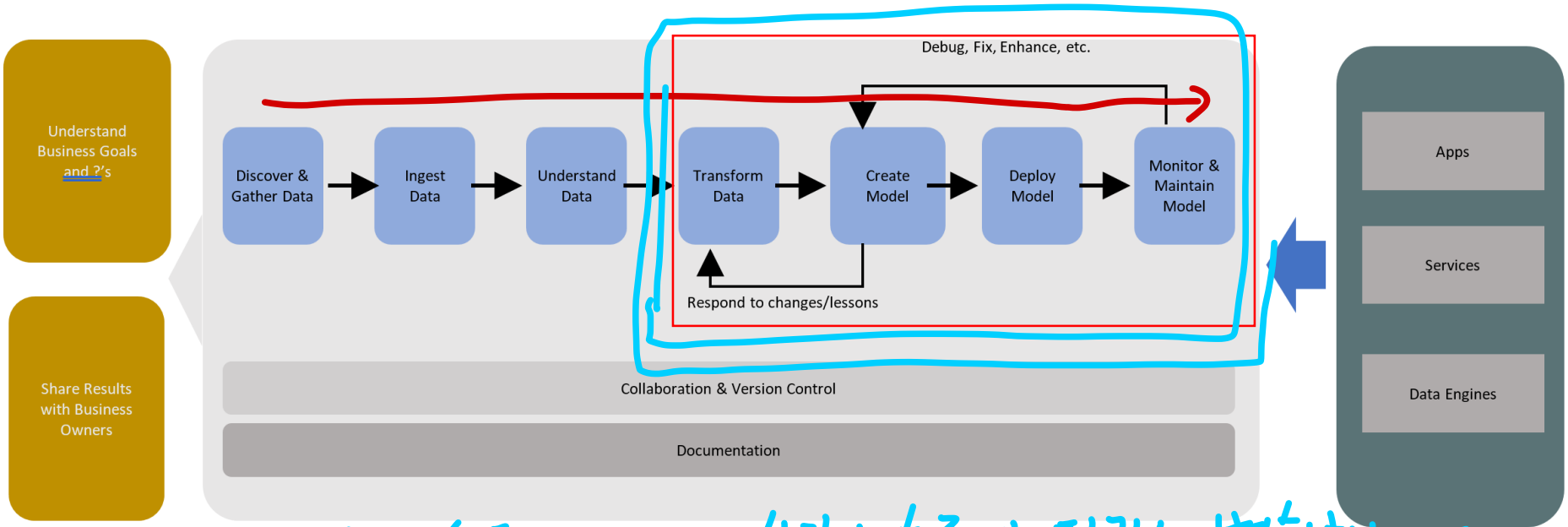
Communicate and Visualize the Results

시각화 (비전공자 게게)

- Step 5: Communicate and Visualize the Results
  - ✓ System implementation, A/B test, model updates, etc.

↳ implementation을 통해 계산하는 단계

Loop



Feedback이 계속 돌아가면서 시간의 흐름에 따라서 변화하는것들을 반영하고 update해서 지속적으로 꾸준히 좋은 모델을 만들수있도록 유지하기.


ANY
questions?