

Machine Learning & Data Mining

- Introduction -

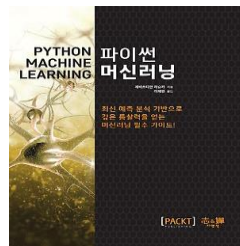
Kyung-Ah Sohn

Ajou University

수업 소개

Logistics

- Instructor
 - Kyung-Ah Sohn (손경아) , kasohn@ajou.ac.kr
 - 산학협력원 507호 (Tel: 031-219-2434)
- Office hours
 - By appointment
- Textbook
 - No official textbook (lecture slides will be posted)
 - Python Machine Learning, by Sebastian Raschka (원서 혹은 번역본, <https://github.com/rasbt/python-machine-learning-book>)
 - 밑바닥부터 시작하는 딥러닝(Deep learning from scratch), 사이토 고키
- TA
 - 문정현



Logistics

- Final grade will be based on
 - Midterm/final exam (30%/30%)
 - Homework (20%)
 - Term project (15%)
 - Class participation (5%)
- Late homework policy
 - Homework is worth full credit at the beginning of class on the due date, or as specified on e-class
 - You will be allowed 3 total late day without penalty for the entire semester
 - Once those days are used,
 - It is worth zero credit after that

Logistics

- Homework
 - Mixture of problem solving and code writing
 - python: highly recommended
- Zero score for copied homework

Logistics

- Academic misbehavior during exam → F grade (no exception)
- No negotiation about the final grade
 - Examples of excuses that are not allowed: [scholarship/graduation](#)
- If you are sick or have emergency situation, tell (email, call) me in advance and bring *official documents*
- Use of [electronic devices](#) must be restricted to class-related activities

당당하게 거절해야 할 14가지 부정 청탁 대상

- 1 인·허가, 면허, 승인, 검사 인증 등 처리
- 2 과태료, 범칙금, 이행강제금 등 각종 행정처분 또는 형벌부과의 감경·면제
- 3 채용, 승진, 보직 등 공직자 인사
- 4 각종 위원회의 위원, 시험·선발 위원 등 공공기관 의사결정에 관여하는 직위자의 선정·탈락
- 5 각종 수상, 포상 등의 선정·탈락
- 6 입찰, 경매, 개발, 시험 특허, 군사, 과세 등에 관한 직무상 비밀
- 7 계약 당사자 선정·탈락
- 8 보조금, 출연금, 교부금, 기금 등의 배정 지원 또는 투자
- 9 공공기관의 재화와 용역의 거래
- 10 각급 학교의 입학·성적 업무
- 11 징병검사·입영기일 연기, 부대 배속 및 보직 부여 등 병역 업무
- 12 장기요양등급 판정 등 공공기관이 실시하는 각종 평가·판정
- 13 행정지도·단속·감사·조사 대상에서 특정 개인·법인 선정·배제
- 14 사건의 수사·재판·심판·결정·조정·중재·화해 등의 처리

Term project (15%)

- Team of 4 members
 - Proposal (1 page)
 - Proposal presentation
 - Final presentation
 - Final report (max 5 pages)
- Some project ideas will be distributed later
- Project score will be based on **peer-review**
 - Team score: Originality/Difficulty/Completeness/ Presentation/Final report
 - Proportional to team contribution

Peer-review form

팀명 (조번호)	이름 (본인포함)	1.아이디어	2. 구현	3. 보고서/ 발표자료	4. 팀워크	계
	계	100	100	100	100	400

* 1~4 각 기여도 항목에 대해 팀원들의 점수 합이 100이 되어야 합니다.

(예: 조원 3인의 경우, 1.아이디어: 30/30/40, 2. 구현: ...)

* 추가하고 싶은 의견이 있는 경우, 아래에 작성하기 바랍니다.

기타의견:

Past project titles (2016, 2017)

- 영화, 드라마 흥행 예측
- 스포츠 예측
- Ah! 이게 뜨네/소속사 차려서 성공하려면?/음악 순위 예측
- 랭킹아이돌: SNS 활용
- 대박자리를 찾아서: Café profit prediction
- 건강보조식품 추천 서비스
- 싸다고 별로인게 아니야: 성분 유사도 기반 화장품 추천
- 니들이 뭔데 나를 판단해: 인물 평판 분석
- 서울의 미세먼지 예측
- 건강보험 심평원 자료를 활용한 심혈관계 질환예측

Past project titles (2018-1)

- DoAjou(도아주): 아주대 학생들을 위한 질의응답 챗봇
- 손그림 기반 캐릭터 검색 웹서비스
- FaceFaker
- VGG-19 모델과 histogram을 기반으로 한 색상, 스타일, 패턴에 따른 패션이미지의 유사성 분석
- 냉장고 재료 인식을 통한 레시피 추천
- 자동 해시태그 추천
- 스타일에 따른 한국어 시 작성 모델
- 텐서플로우 Fast-Rcnn모델을 활용한 재활용 물품의 구성요소 파악

INTRODUCTION

Machine Learning (기계학습)

- 기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구 분야 (Arthur Samuel, 1959)
- Machine Learning (ML) gives computer systems the ability to "learn" with data (컴퓨터 시스템이 데이터를 학습할 수 있는 능력 부여하기)

Machine Learning

- Definition by Tom M. Mitchell (1997)
 - A computer program is said to learn from experience E if
 - it improves its performance P
 - at some task T
 - with experience E



Well-defined learning task: $\langle P, T, E \rangle$

ML applications – everyday life

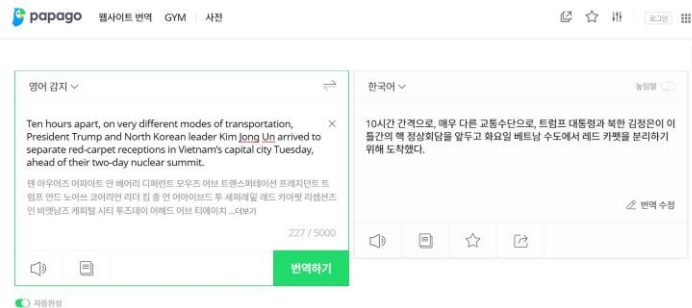
Virtual Personal Assistant



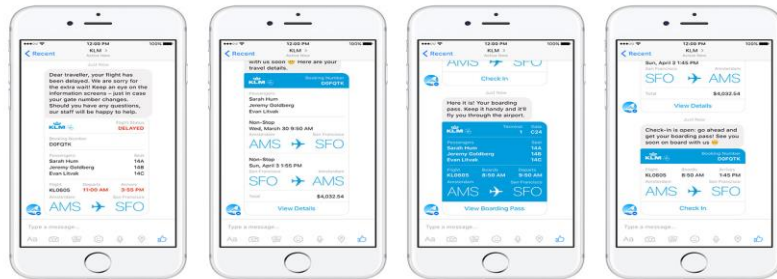
Social Media Services

- *People You May Know:*
- *Face Recognition*
- *Similar Items / Places*

Machine Translation



Chatbot

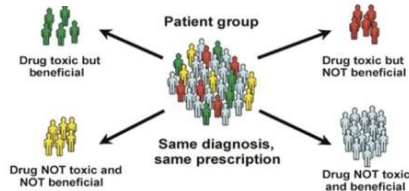
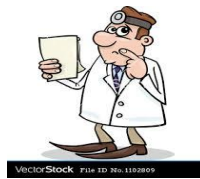
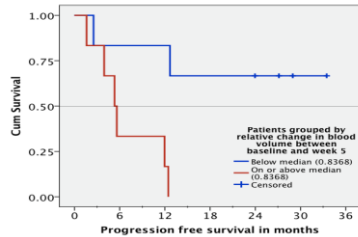


ML applications – everyday life

Spam filtering



<https://www.flickr.com/photos/notoriousxl/3030271346/>



Text recognition



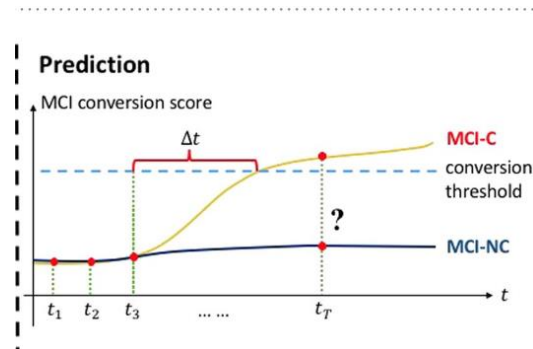
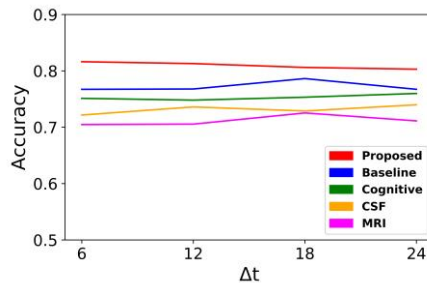
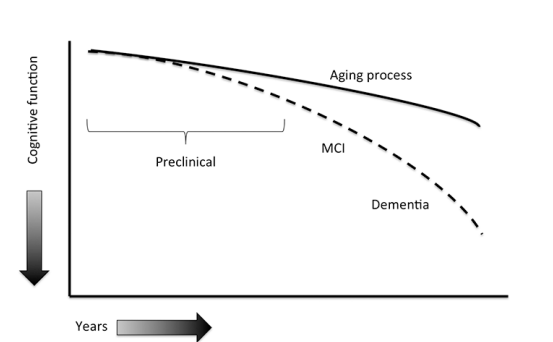
maicm.ittdmc.net

Biology and Medicine

Learning to predict Alzheimer progression

[Lee et al. 2019]

- Task: 경도인지장애 (MCI) 상태에서 x 개월 후 치매로의 변환 가능성 예측
- Data: 나이, 성별, MRI, CSF, APOE status, cognitive score, etc.
- Performance
 - 예측 정확도
(Training accuracy, Test accuracy)



Learning to recognize images

- Task
- Data
- Performance

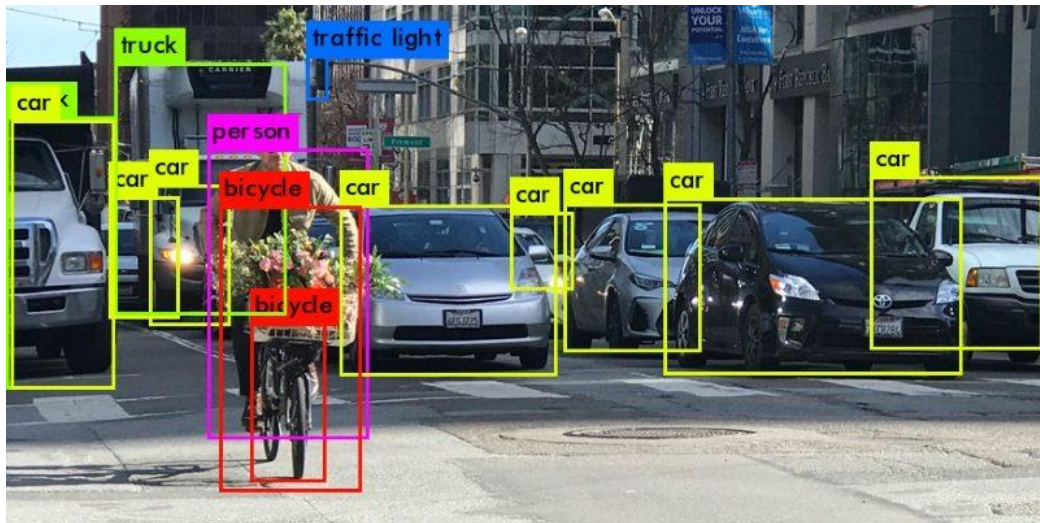
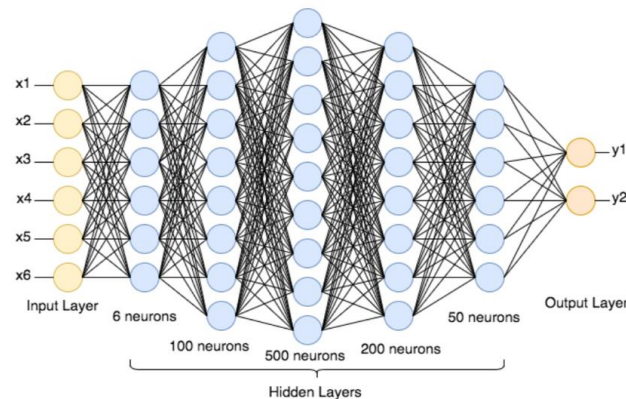


Image source: <https://azati.ai/image-detection-recognition-and-classification-with-machine-learning/>

Deep Learning

- Part of machine learning methods
- Mostly based on **deep neural network** architecture
- Great success in many area
 - **Computer vision, speech recognition**, natural language processing, machine translation, social network filtering, bioinformatics, etc.
- Sometimes comparable to or superior to human experts



What current DL can do

- GAN: generative adversarial network
(생성 모델)



Ian Goodfellow

@goodfellow_ian

Follow



4.5 years of GAN progress on face generation.

arxiv.org/abs/1406.2661

arxiv.org/abs/1511.06434

arxiv.org/abs/1606.07536

arxiv.org/abs/1710.10196

arxiv.org/abs/1812.04948



4:40 PM - 14 Jan 2019

Style transfer



Source photograph



A Picasso, used as a style image

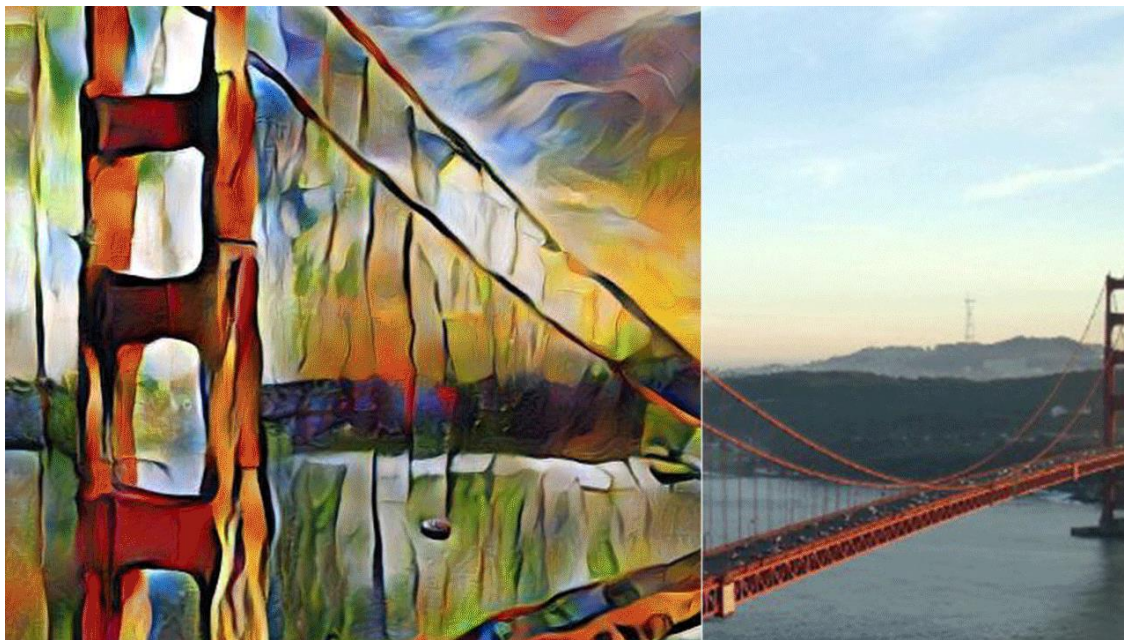


Stylization with Gatys' method



Stylization with WCT, a much faster method

IMAGE CREDIT: Yijun Li et al.



<https://research.adobe.com/news/image-stylization-history-and-future-part-3/>

What current DL can do

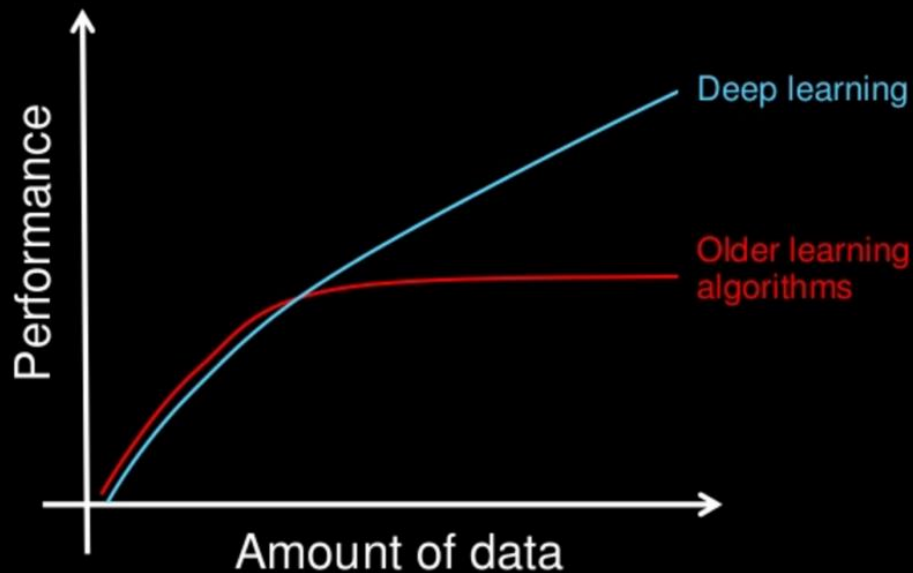
- Vehicle tracking and speed estimation for the NVIDIA AI City Challenge Workshop at CVPR 2018

https://www.youtube.com/watch?v=_i4numqiv7Y

- Deep and Machine Learning Uses that made 2019 a new AI Age.

<https://towardsdatascience.com/14-deep-learning-uses-that-blasted-me-away-2019-206a5271d98>

Why deep learning

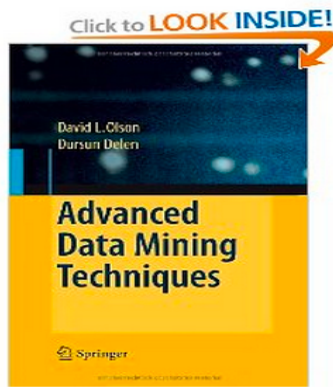


How do data science techniques scale with amount of data?

Slide by [Andrew Ng](#), all rights reserved

데이터 마이닝 (data mining)

- 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아 내는 것
- **KDD** (데이터베이스 속의 지식 발견, knowledge-discovery in databases)
- 응용: 신용평가모형, 사기탐지시스템, 장바구니 분석, 최적 포트폴리오 구축 등



Click to open expanded view

Advanced Data Mining Techniques [Paperback]

[David L. Olson](#) (Author), [Dursun Delen](#) (Author)

[Be the first to review this item](#)

List Price: ~~\$139.00~~

Price: **\$102.95** & **FREE Shipping**. [Details](#)

You Save: **\$36.05 (26%)**

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it tomorrow, Nov. 19? Order within **8 mins** and choose **One-Day Shipping** at checkout. [Details](#)

25 new from **\$44.95** **18 used** from **\$55.28**

amazonstudent Miss your FREE Two-Day Shipping? [Save 50% on a Prime membership.](#)

Formats

Amazon Price **New from** **Used from**

Customers Who Viewed This Item Also Viewed



Data Mining: Practical Machine Learning ...

> [Ian H. Witten](#)

★★★★☆ (33)

Paperback

\$37.80



Data Mining Techniques: For Marketing, Sales, ...

> [Gordon S. Linoff](#)

★★★★☆ (9)

Paperback

\$33.80



Data Mining: Concepts and Techniques, Third ...

> [Jiawei Han](#)

★★★★☆ (19)

Hardcover

\$60.52



Data Mining with R: Learning with Case ...

> [Luis Torgo](#)

★★★★☆ (11)

Hardcover

\$66.55

Your Recently Viewed Items and Featured Recommendations

You viewed



Buy New

\$102.95

Quantity:

☐ Yes, I want **FREE Two-Day Shipping** with **Amazon Prime**

Add to Cart

or

Sign in to turn on 1-Click ordering

Add to Wish List

Sell Us Your Item

For a **\$1.80** Gift Card

Trade in

[Learn more](#)

Kindle Edition

Read instantly on your iPad, PC, Mac, Android tablet or Kindle Fire

Buy Price: **\$94.99**

Rent From: **\$38.64**

[Get Kindle Edition Here](#)

www.amazon.com

Mining the transaction data

- Association analysis (연관분석)



Wal-Mart data mining example:

On Friday afternoon, young American males who buy **diapers** also have a predisposition to buy **beer**.

Data mining vs. Machine learning

- Implementing data mining techniques typically involves two components: **database** for data management, and **machine learning** for data analysis.
- To drive a business both techniques have to work hand to hand, one technique will define the problem and other will give you the solution in the much accurate way.

DM

Introduced in 1930'

Traditional DB with structure data

Used to find patterns (or rules) from data

ML

Introduced in 1950'

Existing data as well as algorithms

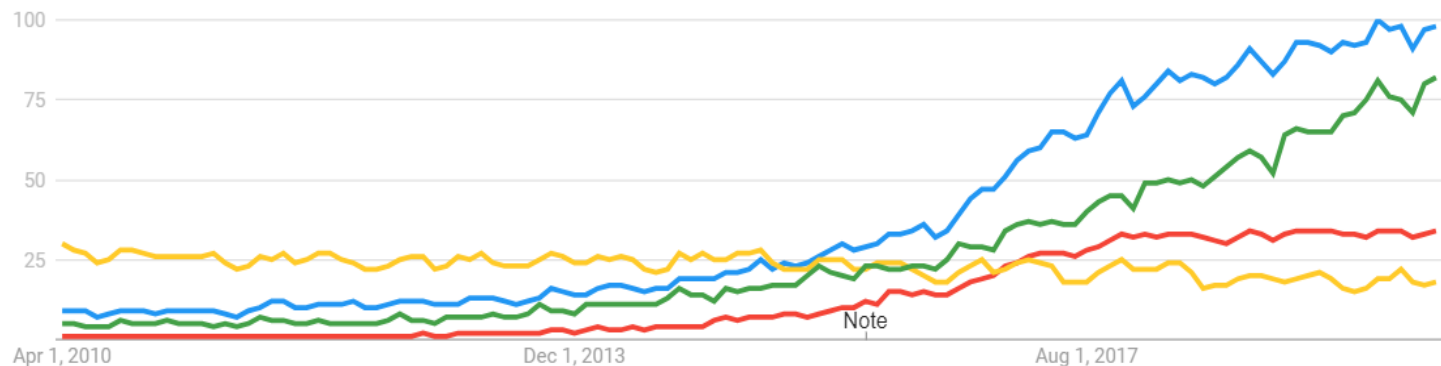
ML learns and understand the rules.

* 데이터 사이언스

- 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합 분야
- 데이터마이닝 및 빅데이터와 연관
- 응용: 추천, 광고 서비스 등 다양한 분야

Google Trend: ML / DL / DM / DS

Worldwide



machine learning

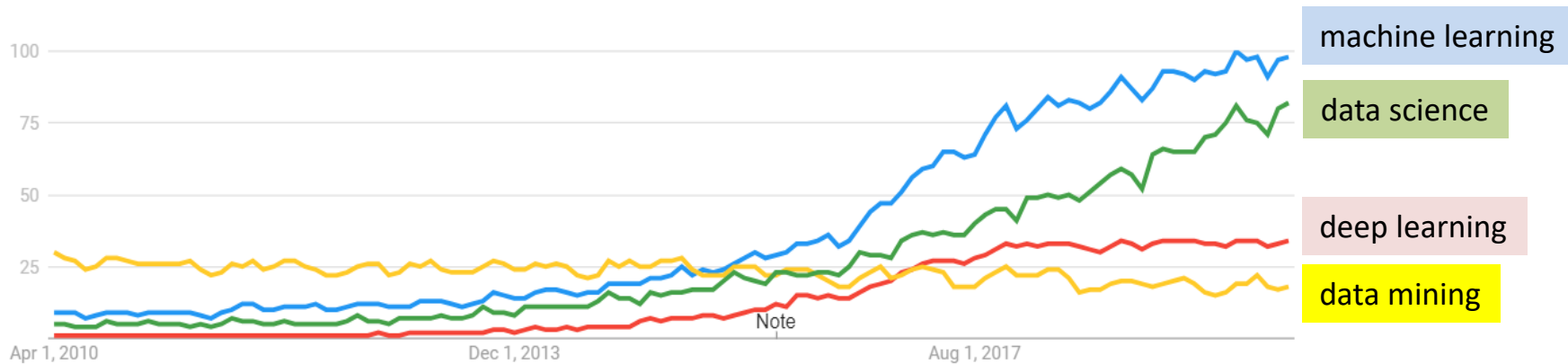
deep learning

data mining

data science

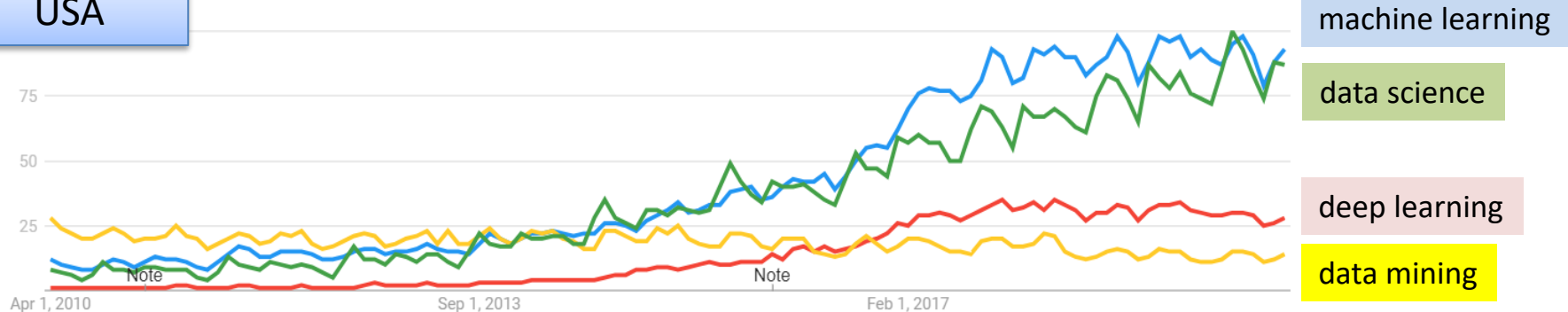
Google Trend: ML / DL / DM / DS

Worldwide

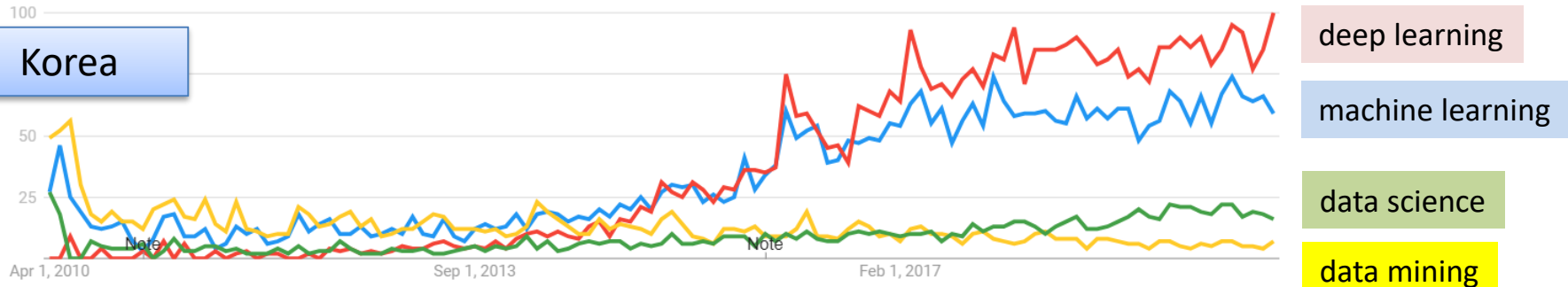


Google Trend: ML / DL / DM / DS

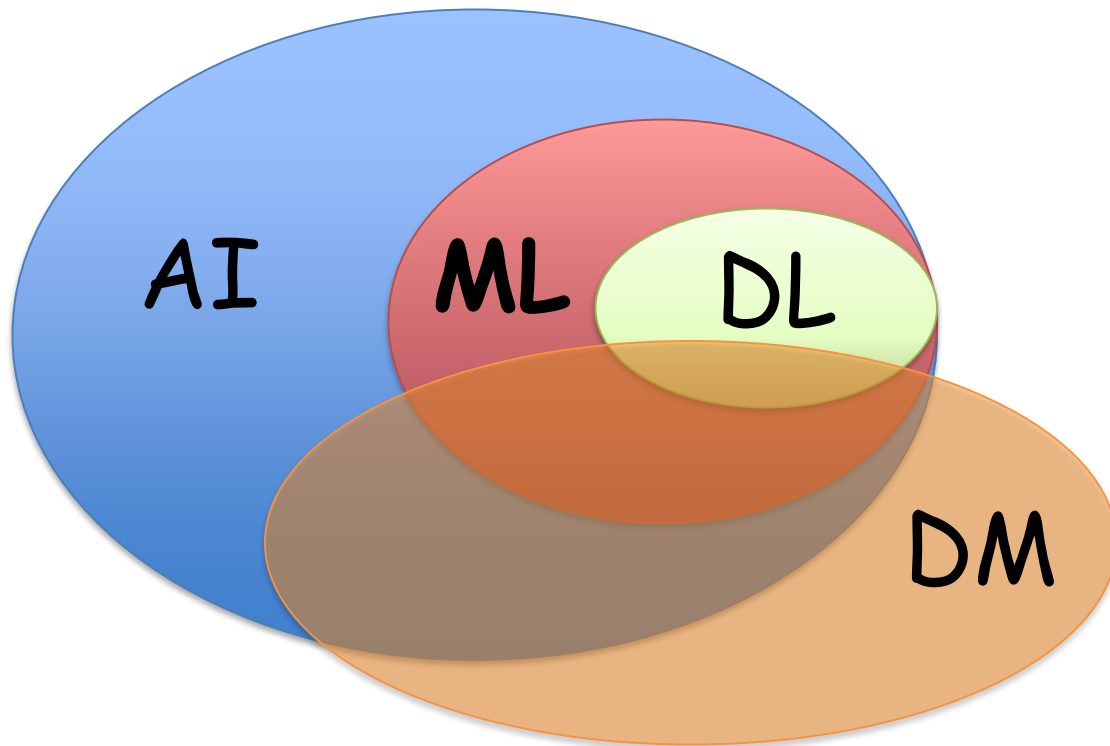
USA



Korea



AI / ML / DL / DM



금융/산업 빅데이터



의료 빅데이터

국민 10만명 유전체 정보 빅데이터 구축

기사입력 2013-12-04 03:00:00 | 기사수정 2013-12-04 11:19:25

Like 2 Tweet 기사보내기

2021년까지... 맞춤형의료에 활용

2021년까지 국민 10만 명의 유전체 정보가 빅데이터로 구축된다. 이 정보들은 개인 맞춤형 의료, U-헬스케어 등 융복합형 신산업을 창출하는 데 활용된다. 정부는 3일 경제관계장관회의를 열어 이 같은 내용의 '4대 국민 생활 분야(안전 건강 편리 문화) 융합 신산업시장 활성화 전략'을 확정 발표했다. 4대 전략은 4월 전문가 100여 명으로 구성된 산업융합포럼이 만든 40개 신사업 모델 중 최종 선정된 항목들이다.

4대 전략의 핵심인 유전 정보 빅데이터가 구축되면 개인 맞춤형 의료 서비스가 실현된다. 예를 들어 유전체 분석 서비스를 받으면 10년 뒤 신체 어느 부위에 암이 생길 가능성이 높다는 결과를 얻을 수 있다. 암 발병을 막기 위한 식이 요법, 운동 처방, 예방 약물치료를 받을 수 있다.

정부는 빅데이터 구축을 위해 공공기관이 보유한 유전체 자원, 국가 연구 결과물 등을 의무적으로 소관 부처로 제출하게 할 방침이다. 복지부 관계자는 "빅데이터 구축으로 유전체 분석해독 컨설팅 사업, 맞춤 의료, 맞춤 제약, U-헬스 등 신비즈니스 모델이 창출될 것"이라고 기대했다.

정부는 △u-안심생활 서비스 활성화(안전) △스마트 홈에너지 관리 서비스 확산(편리) △문화예술 체험형 콘텐츠 비즈니스 창출(문화) 등도 유망한 융합 신산업 분야로 선정해 지원해 왔다.

인공지능 의사 왓슨



- 출생: IBM 수퍼컴퓨터
- 학습: 의학저널·전문문헌 290종, 의학과교서 200종, 전문자료 1200만 쪽
- 경력: 2012년 미국 메모리얼슬로언케터링 암센터에서 '레지던트' 시작
- 분석력: 내년 중 암의 85% 분석 예상



결핵 환자 영상에 대한 인간과 컴퓨터의 판독 결과

결핵 전문의

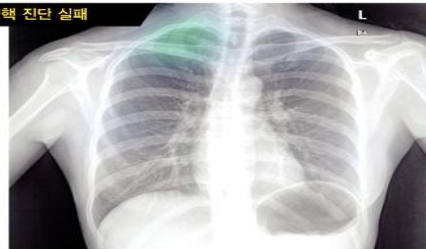
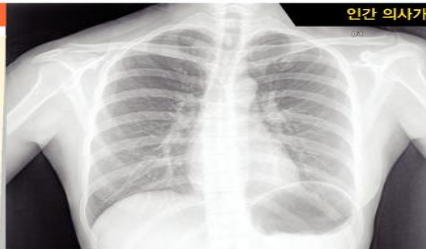


병변이 없는 것으로
파악되어 정상 판정

서법석 루닛 의료담당 이사
"저도 발견하지 못했어요. 컴퓨터 판독
결과를 보고 리부를 해봤더니, 폐에
가려져서 잘 안 보였던 것 같아요.
신체구조상 쉽게 놓칠 수 있는 병변
이라고 판단됩니다."



인간 의사가 결핵 진단 실패



루닛 인공지능(DIB)

결핵 가능성(abnormality score)

33.66% 결핵 확인



김양중 <한겨레> 의료전문기자
"이렇게 낮게 평가된 수치를 보고
결핵인지 여부를 판단할 수 있느냐의
문제는 앞으로 풀어야 할 중요한
과제입니다."



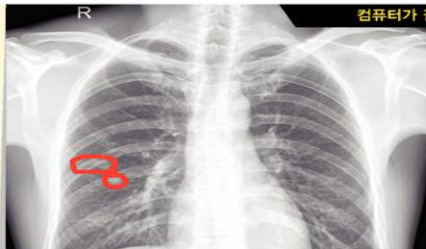
결핵 확인해
병변 위치를 표시



서 이사
"왼쪽에 병변이 확실해 있는 것 같아요.
결정 모양의 것들이 여러 개 있어요."
김 기자
"이 정도 이상이면 못 찾기 어려울 것
같은데요."



컴퓨터가 결핵 진단 실패



결핵 가능성(abnormality score)

3.1% 정상 판정



백승욱 루닛 대표
"이 결과는 사실상 '정상' 소견을
냈다고 봐야 해요. 컴퓨터에게
어려운 게 걸렸네요. 좀 더 학습을
시키면 나아지겠죠."



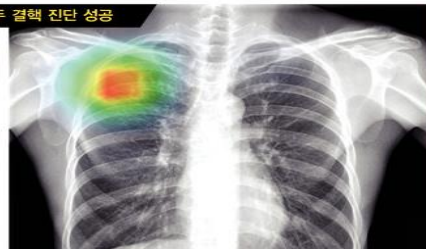
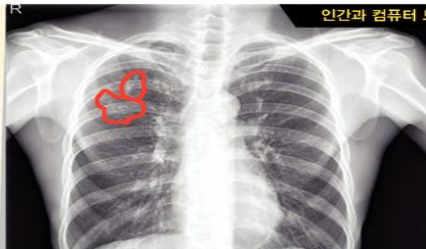
결핵 확인해
병변 위치를 표시



서 이사
"왼쪽에 병변이 나타나 있죠.
음영이 증가된 결절이 보이고 가스가
찬 공동도 보여 전반적으로 결핵이라
시사하는 소견입니다."



인간과 컴퓨터 모두 결핵 진단 성공



결핵 가능성(abnormality score)

96.45% 결핵 확인



백 대표
"학습에 사용된 데이터의 결핵과
유사한 패턴을 컴퓨터가 발견한
것입니다."



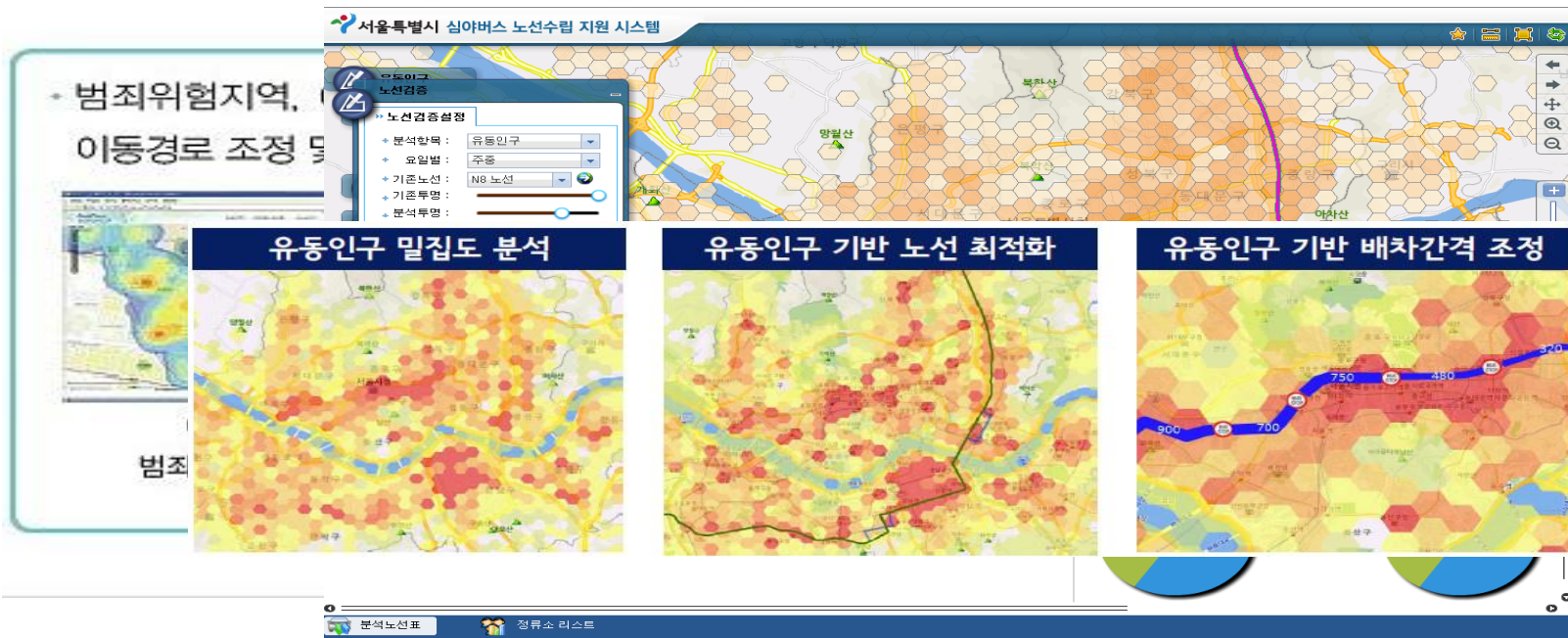
사용자 빅데이터

스마트폰/태블릿 PC/SNS/Game/...

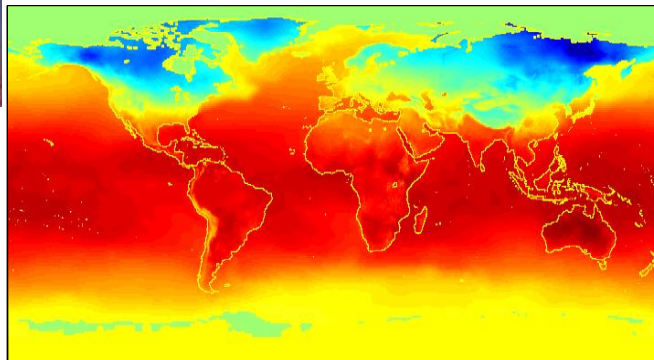


공공 빅데이터

〈 빅데이터 분석 · 활용 사례 : 경찰청 범죄예측 지리적 프로파일링 시스템 〉

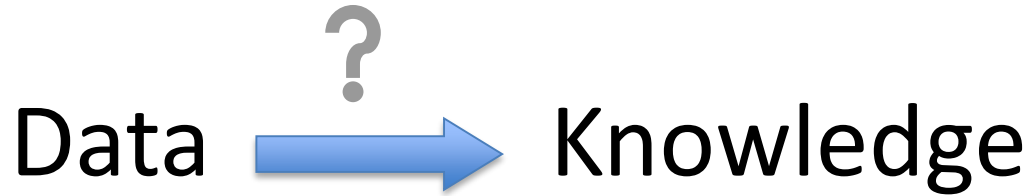


자연과학 빅데이터



The promise of Big Data

- Data contains information of great business value



- ✓ Big data + High performance computing +

Good algorithms



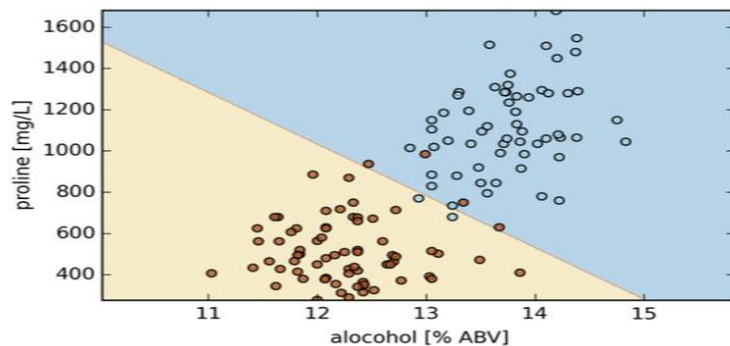
- ✓ Domain knowledge

학습 주제

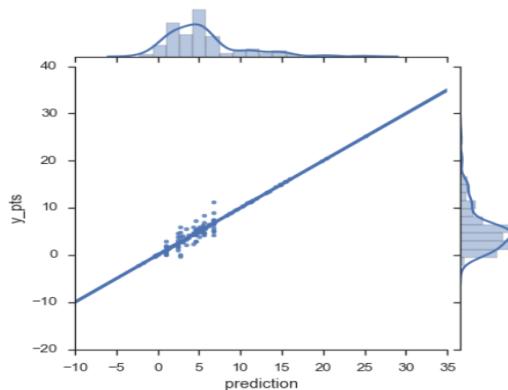
Machine Learning Approaches

- **Supervised learning**

- Labeled data
- Direct feedback
- “**Predict** outcome/future”



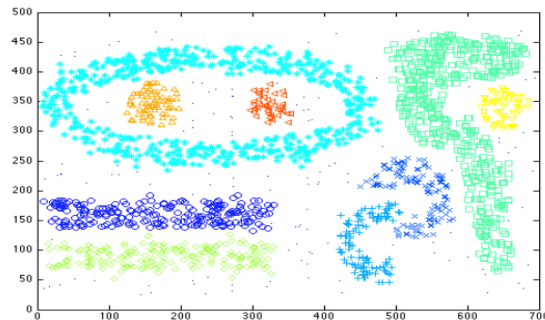
Classification



Regression

- **Unsupervised learning**

- No labels
- No feedback
- “**Find** hidden structure”



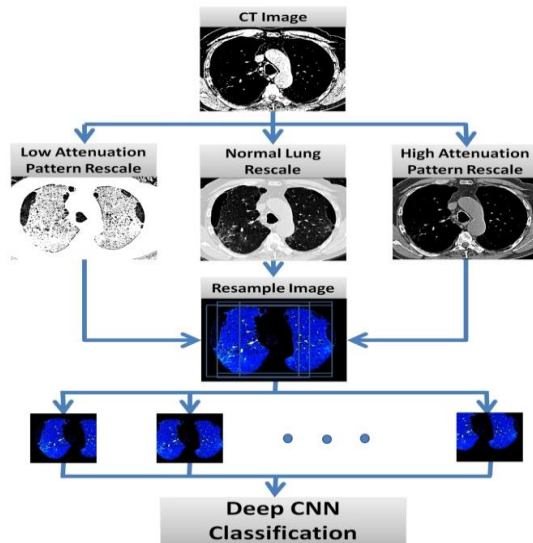
Clustering

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Supervised learning: classification

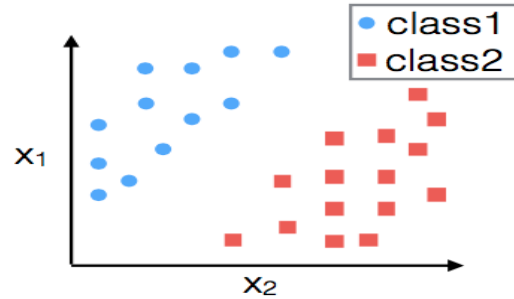
- Given an object, which class of objects does it belong to?
- Given object x , predict its class label y

- ✓ **Medical image classification**
- ✓ **Computer vision:** is this a cat?

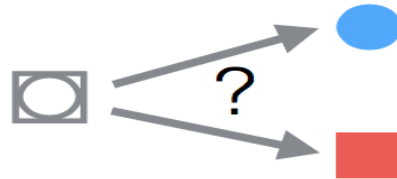


Classification

1) Learn from training data



2) Map unseen (new) data



Common classifiers

- SVM
- K-NN
- Naïve Bayes
- Logistic regression
- Decision tree
- Random Forest (ensemble methods)
- Neural network / Deep learning

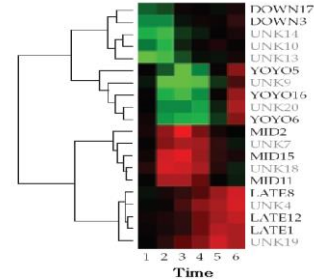
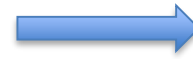
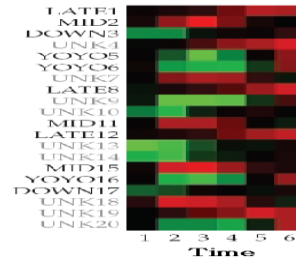
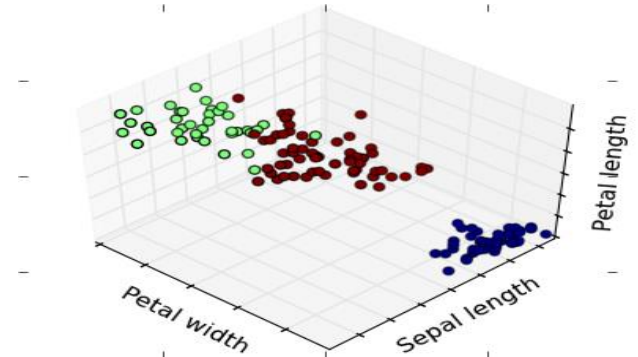
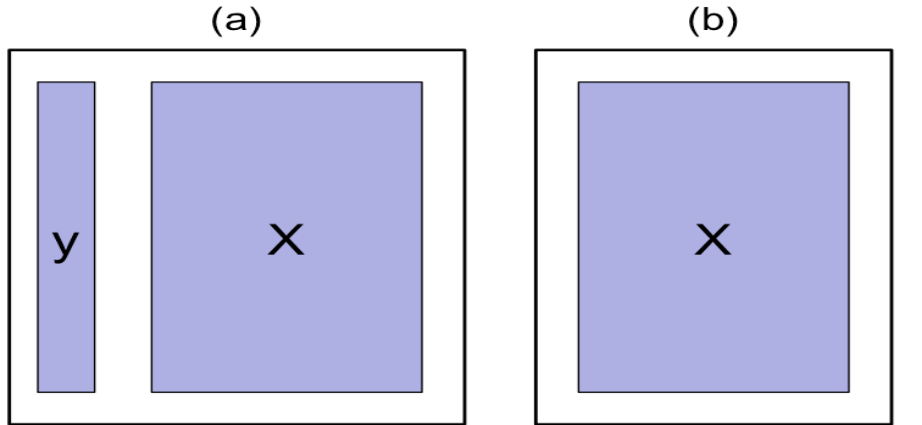
Regression

- Supervised learning: predict one variable Y given a set of other variables X
 - Classification: Y is categorical
 - Regression: Y is numeric
- Assume a linear or non-linear model
- Applications
 - Sales forecasting
 - Stock market prediction

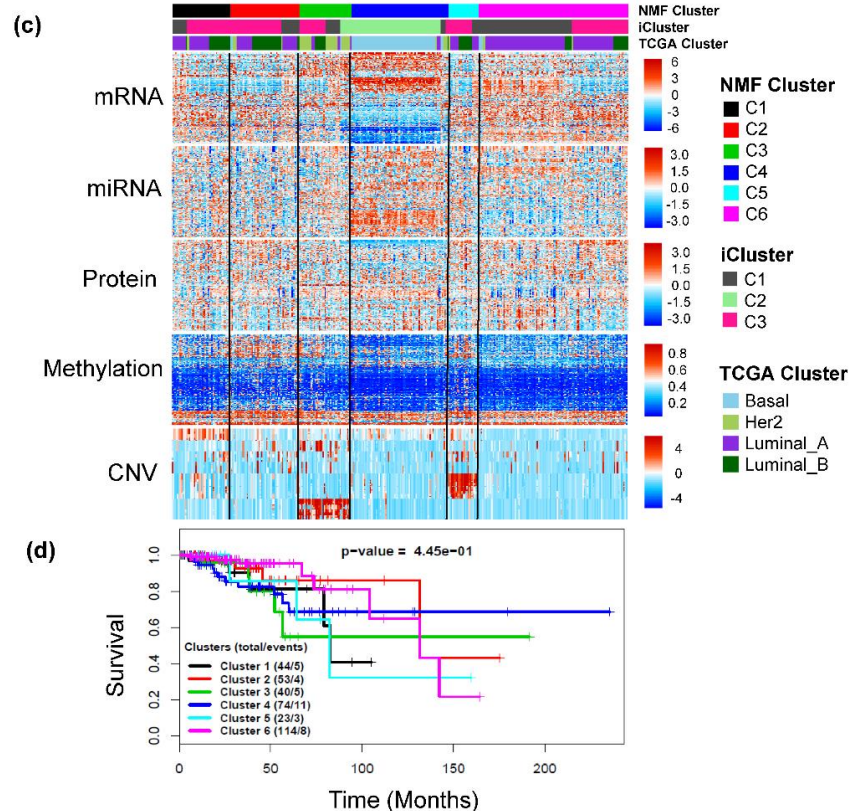


Clustering

- Finding groups of items that are similar
- Clustering is **unsupervised**



Cancer subtype clustering



Chalise P, Fridley BL (2017) Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLOS ONE 12(5): e0176278.
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176278>

Choice of Distance Measure

- Euclidean distance
- L_p norm
- Mahalanobis distance
- 1 - Correlation
- 1 - Cosine similarity
- ...

Data and Application-dependent

Dimension reduction

- High-dimensional data \rightarrow low-dimensional data
 - For feature extraction / selection
 - For visualization
- Principal Component Analysis (PCA), tSNE, Autoencoder, etc.

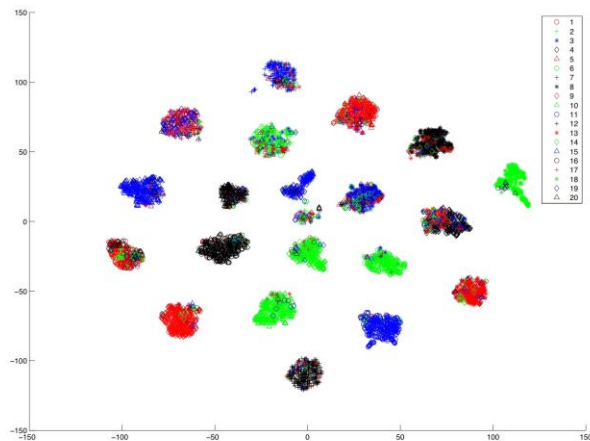


Image source: <https://lvdmaaten.github.io/tsne/>

Reinforcement learning (강화학습)

- For sequential decision making
 - game, robot path planning, self-driving car
- Not supervised learning
: 매 상태에 대해 취해야 할 ground-truth (정답)를 알지 못한다
- Not unsupervised
: 환경으로부터 reward(포상) 형태를 통해 피드백을 받음
- 가장 큰 차이점
: Sequential & correlated data

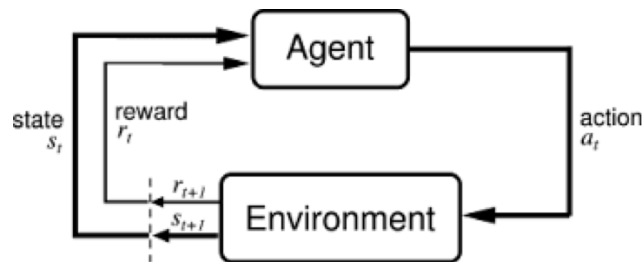


Image credit: [Sutton & Barto](#)

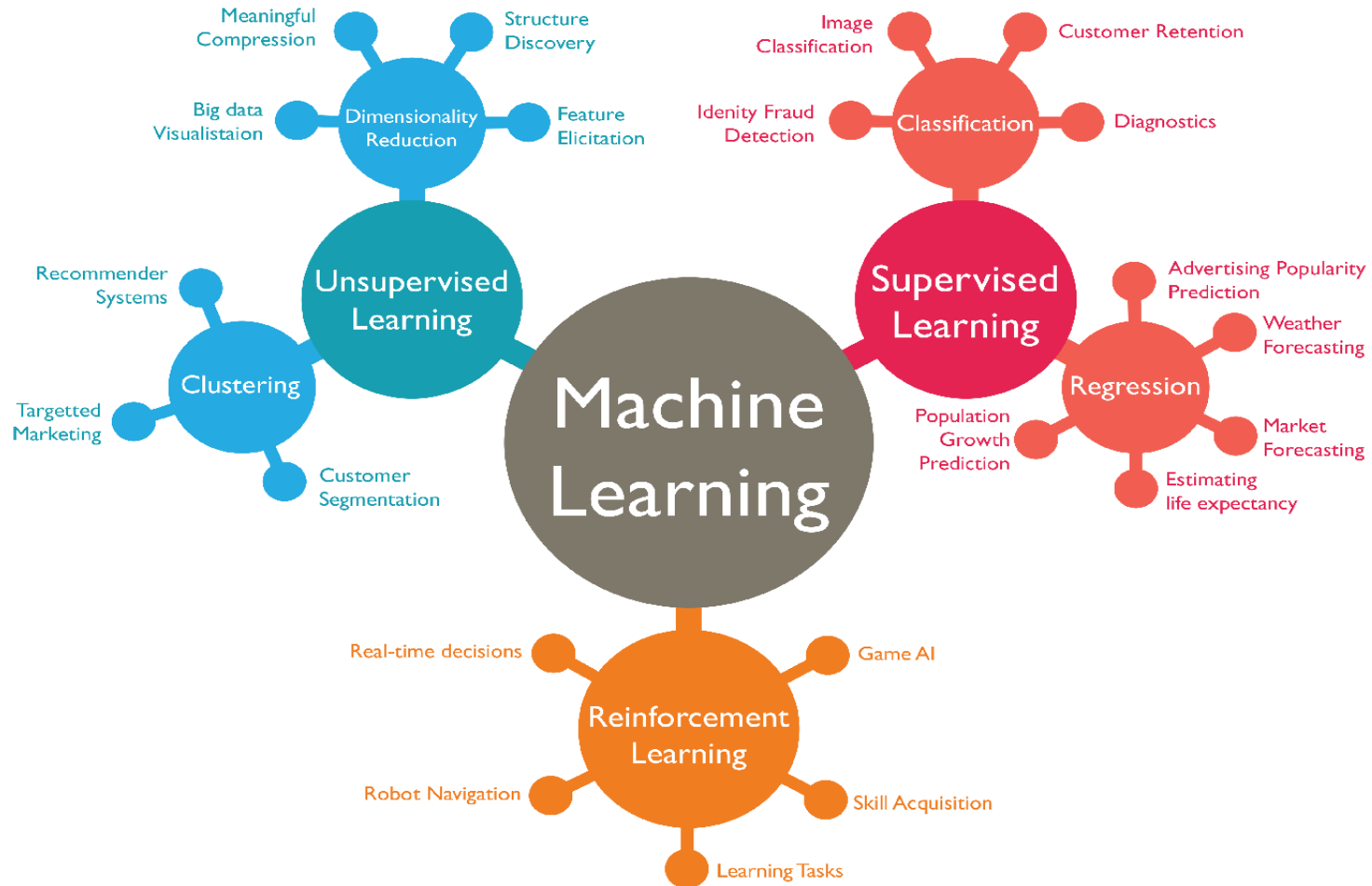
Text analysis

- Text pre-processing
- Text data representation
 - TF-IDF (term-frequency inverse-document frequency)
 - Word embedding (word2vec)
- Content analysis
 - Latent Dirichlet Allocation (LDA)
- Various RNN based models

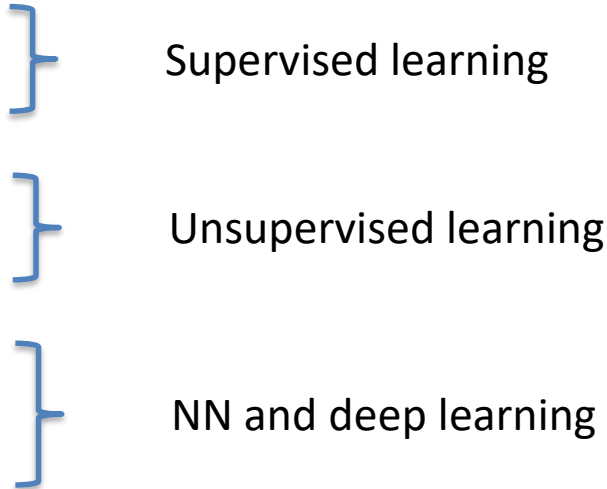
Image analysis

- Image representation, understanding and analysis
- Mostly with deep learning (CNN) models – Object detection, recognition, segmentation, super-resolution, style transfer, etc.
- Will not discuss about raw image processing or traditional computer vision techniques

학습 정리



Topics to cover

- Introduction
 - Classification
 - Regression
 - Evaluation and model selection
 - Clustering
 - Dimension reduction / visualization
 - Recommender system
 - Neural network
 - CNN and image analysis
 - RNN and text analysis
 - Generative model
 - Reinforcement learning
 - Advanced techniques recent applications
- 
- The diagram uses blue curly braces to group the topics into three categories:
- Supervised learning**: Includes Classification, Regression, and Evaluation and model selection.
 - Unsupervised learning**: Includes Clustering, Dimension reduction / visualization, and Recommender system.
 - NN and deep learning**: Includes Neural network, CNN and image analysis, RNN and text analysis, and Generative model.