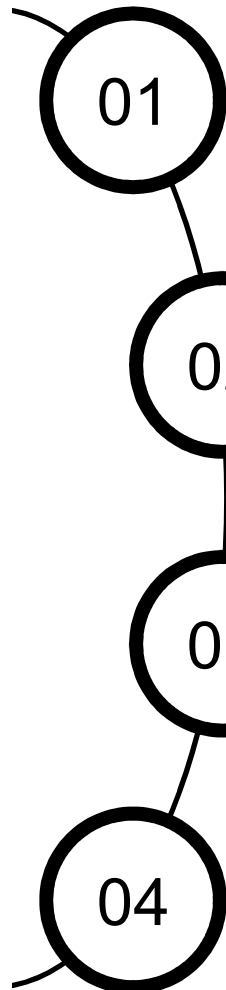


# 자기소개

한국원자력연구원  
미래전략본부  
지능형컴퓨팅연구실  
이유한 선임연구원

# Table of Contents

---

- 
- 01 Background of Youhan Lee
  - 02 Youhan Lee as Researcher
  - 03 Youhan Lee as Kaggle
  - 04 Dream of Youhan Lee in KAERI

# 01 Background of YH

# Profile

---

## # 학력

학부 : 부산대학교 화공생명공학부 졸업 (2008.03 ~ 2014.02)

석사: KAIST 생명화학공학과 졸업 (2014.03 ~ 2016.02)

박사: KAIST 생명화학공학과 졸업(예정) (2016.03 ~ 2020.02)

실험실: Molecular Simulation Laboratory (Prof. Jihan Kim)

세부 전공: Molecular simulation, computational chemistry,  
Machine learning, deep learning

## # 병역

육군 기계화학교 K-1 전차 조종수 조교 (만기 전역)



이유한

# Profile

---

## # 테크니컬 스킬

### Programming

: C, MATLAB, Python

### Computational chemistry

: Quantum Espresso, VASP, COTA

### Data science

: Tensorflow, Keras, Pytorch, Sklearn, various visualization libraries

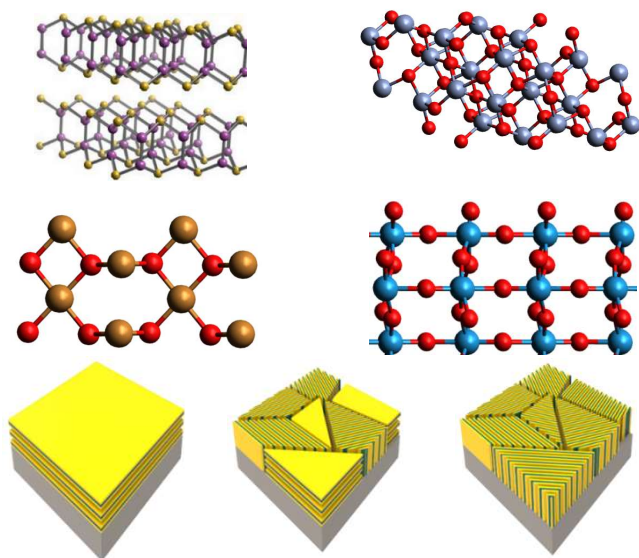


이유한

## 02 YH as researcher

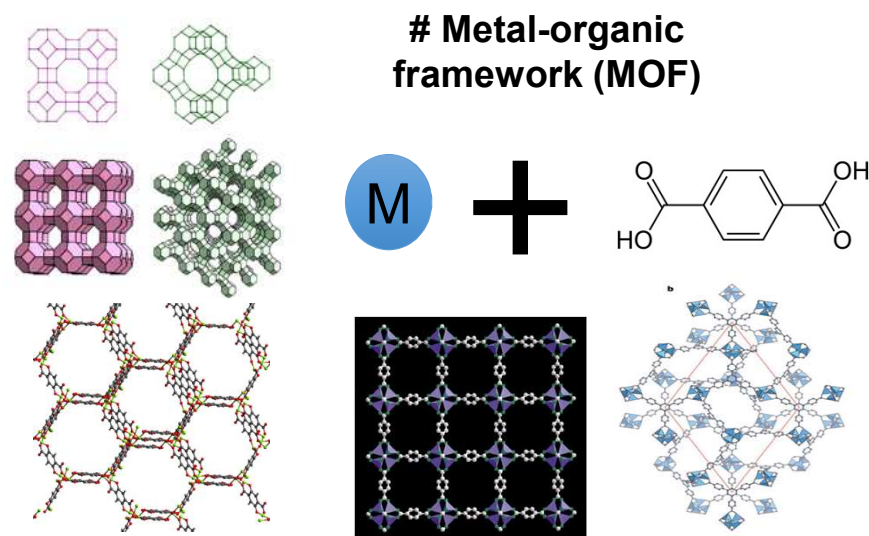
# Two main subjects of my degree

## # Gas sensing



- I've simulated the **sensing behavior of various chemi-resistors** for various applications.
  - Harmful gas sensing
  - Hydrogen monitoring

## # Porous material

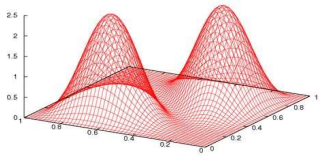


- I've simulated the **adsorption properties of MOFs** for various applications.
  - CCS (carbon capture storage)
  - Hydrogen storage

# Two main skills of my degree – Molecular simulation

# From master to 2<sup>nd</sup> year of Ph.D

# Quantum mechanics    # Classical mechanics

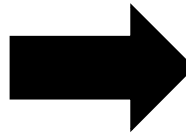


$$\hat{H}\psi = E\psi$$

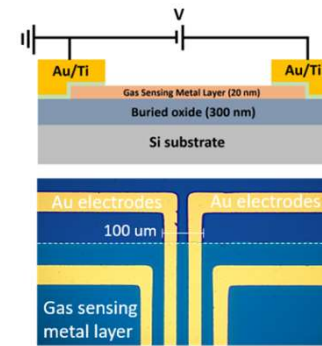
Electron 1

Electron 2

$$E_{ij} = \frac{q_i q_j}{4\pi\epsilon_o r_{ij}} + 4\epsilon \left[ \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right]$$

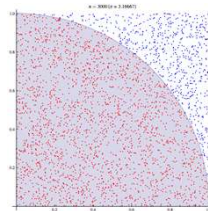
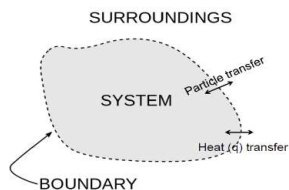


# Gas sensing

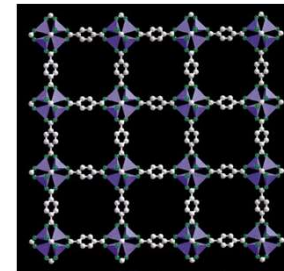


- Binding energy
- Density of states

# Grand canonical Monte Carlo simulation



# Porous material



- Working capacity
- Heat of adsorption
- Binding energy
- Henry coefficient
- Adsorption isotherm



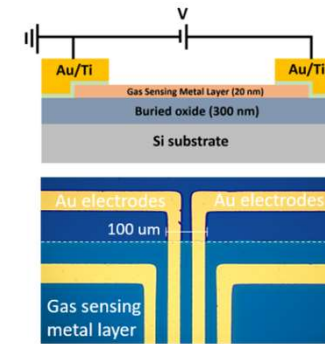
# Two main skills of my degree – Data science

# From 3<sup>rd</sup> year of Ph.D to 4<sup>th</sup> year of Ph.D

# Data science

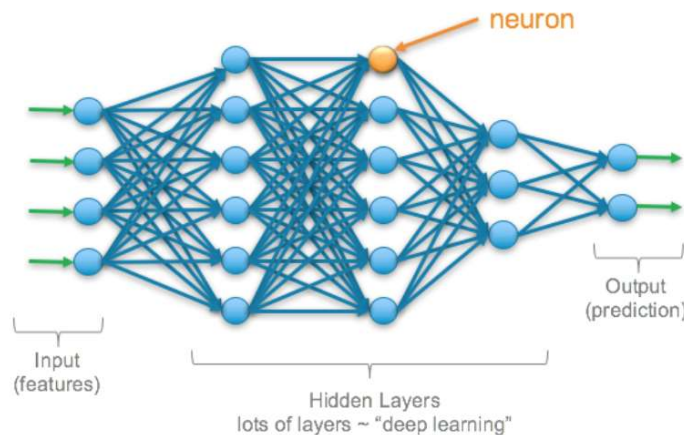


# Gas sensing

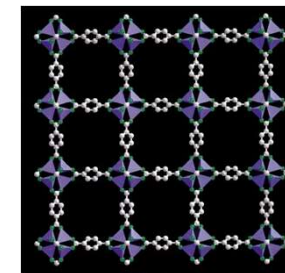


- Anomaly detection (Sensitivity)
- Classification (Sensitivity)

# Deep learning



# Porous material



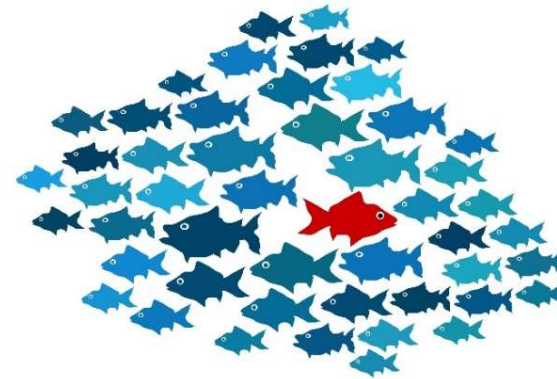
- Henry coefficient
- Defect

# 02.1 Research 1

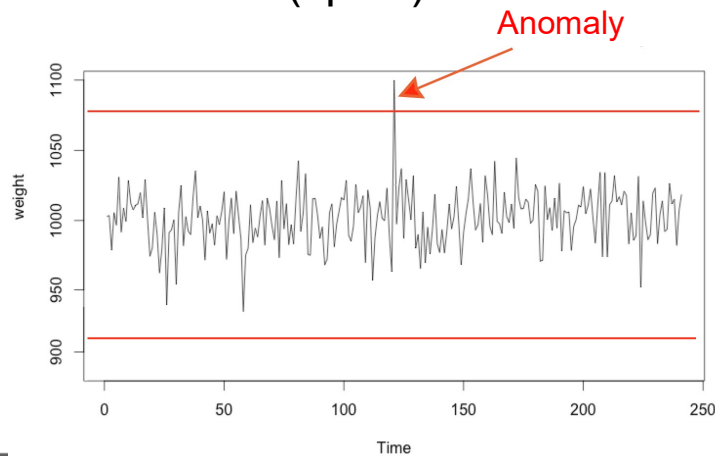
# Methods – Anomaly detection

## # Anomaly

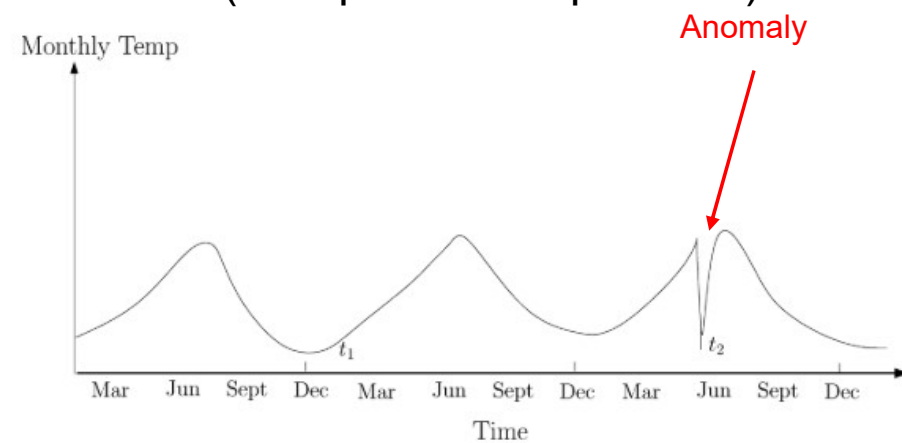
- All patterns(point, behavior) which are unexpected and different from the normal patterns.
- Type:
  - (1) Point anomaly.
  - (2) Contextual anomaly.
  - (3) Collective anomaly (point + contextual).



## # Point anomaly (spike)



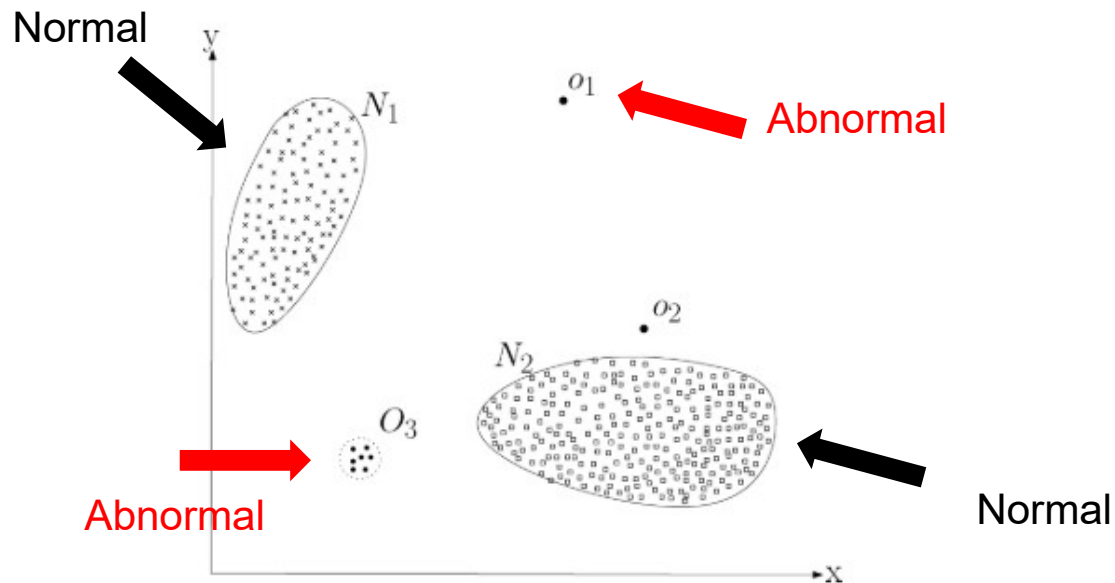
## # Contextual anomaly (unexpected temperature)



Chandola et al, ACM Computing Surveys, 41, 15 (2009).

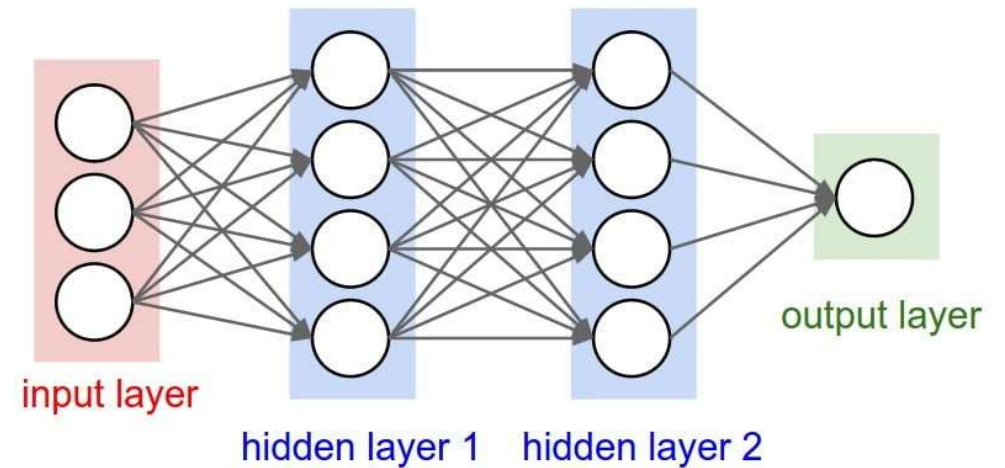
# Methods – Deep learning based anomaly detection

## # Anomaly detection



Chandola et al, ACM Computing Surveys, 41, 15 (2009).

## # Deep learning



- Artificial neural network can find the distribution of normal state automatically.

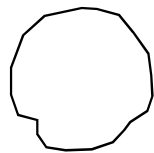
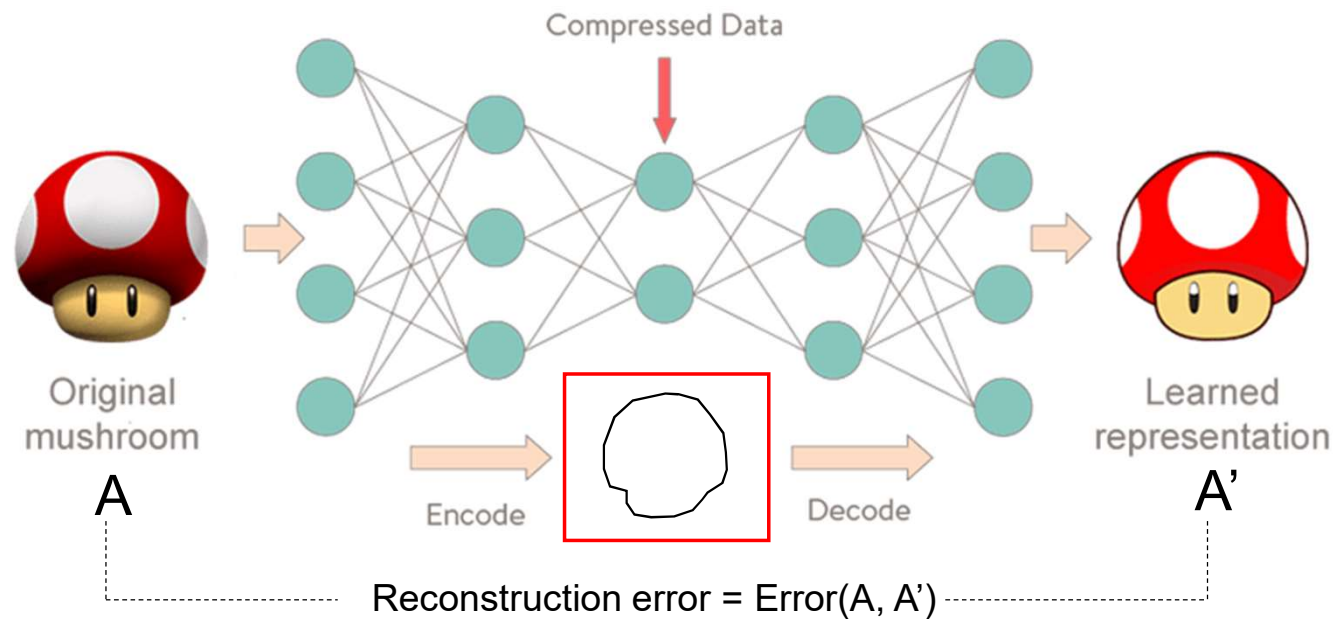
# Methods – Representation learning using auto-encoder

## # Auto-encoder (AE)

: Neural network architecture which learn the features during reconstruction of input.

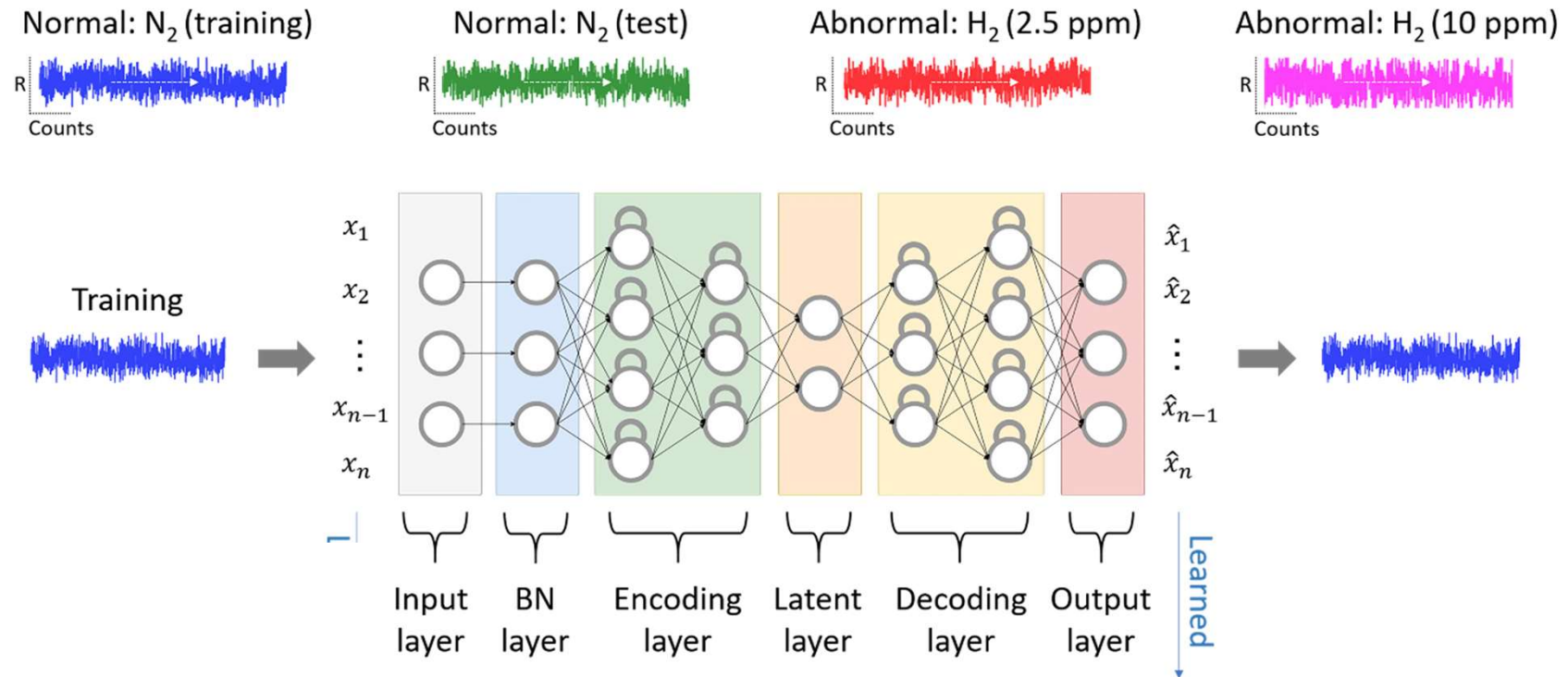
Rumelhart et al, *Parallel Distributed Processing*. Vol1: Foundations. MIT Press, Cambridge, MA. (1986)

Vincent et al, *J.Mach. Learn. Res.* **11**, 3371-3408 (2010)



**Latent, represented features**

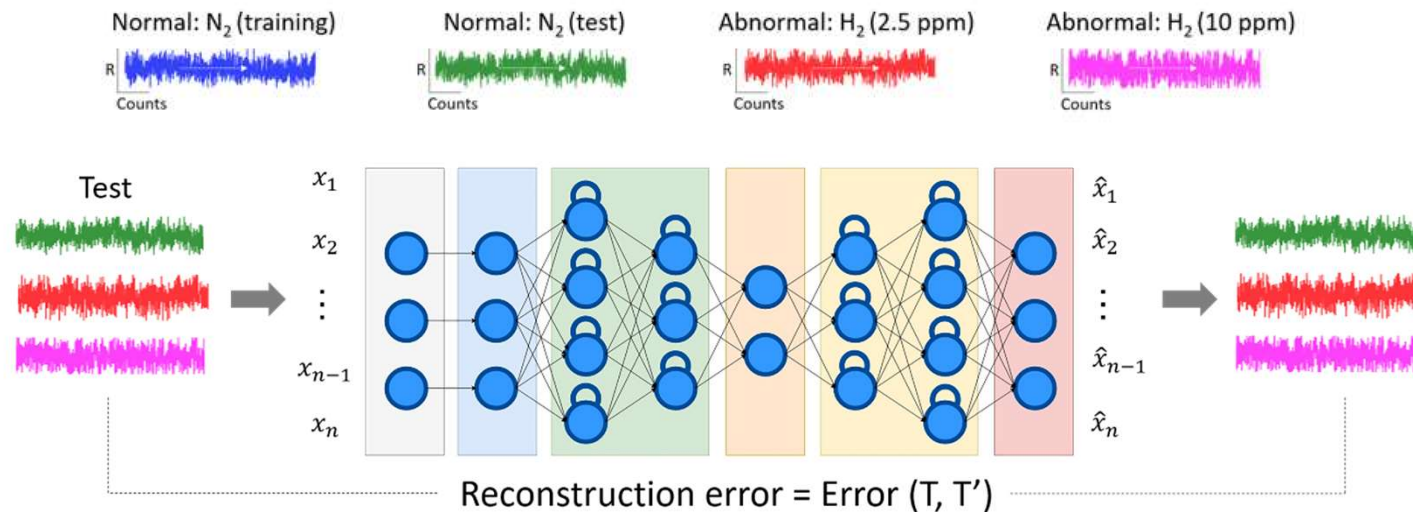
# Methods – Overall process: semi-supervised learning



- Learning the normal exclusively.
- Model are optimized to represent the normal statistics.
- 1,000 reconstruction errors of training( $N_2$ ) were calculated.

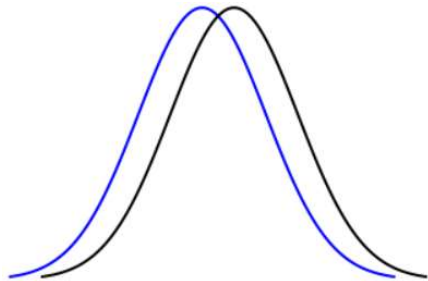


# Methods – Overall process: semi-supervised learning

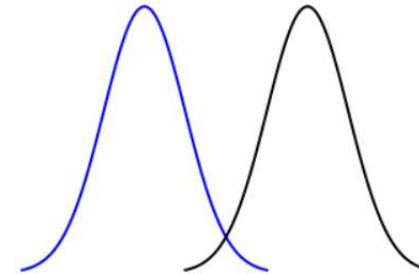


- 3,000 reconstruction errors of test(N<sub>2</sub>, H<sub>2</sub>) were calculated.

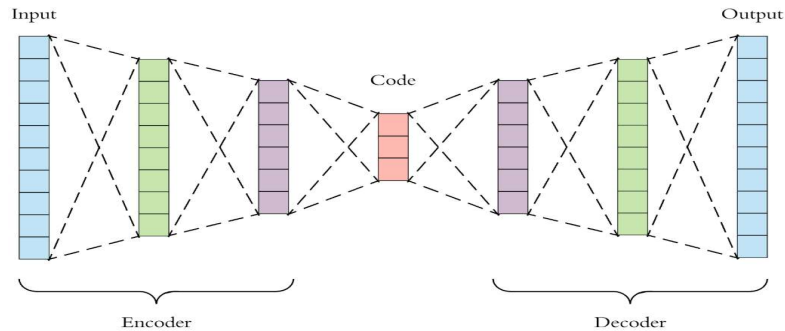
**# Poor classification**



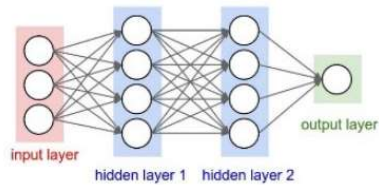
**# Good classification**



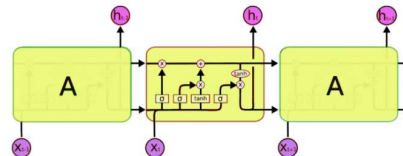
# Results – Finding an optimal architecture



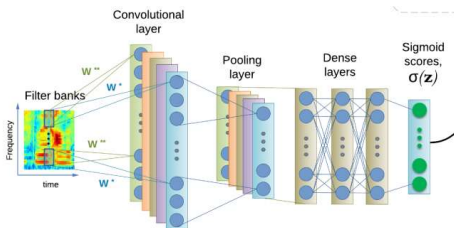
## # Fully-connected layer



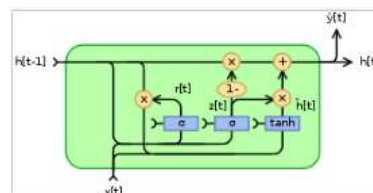
## # LSTM layer



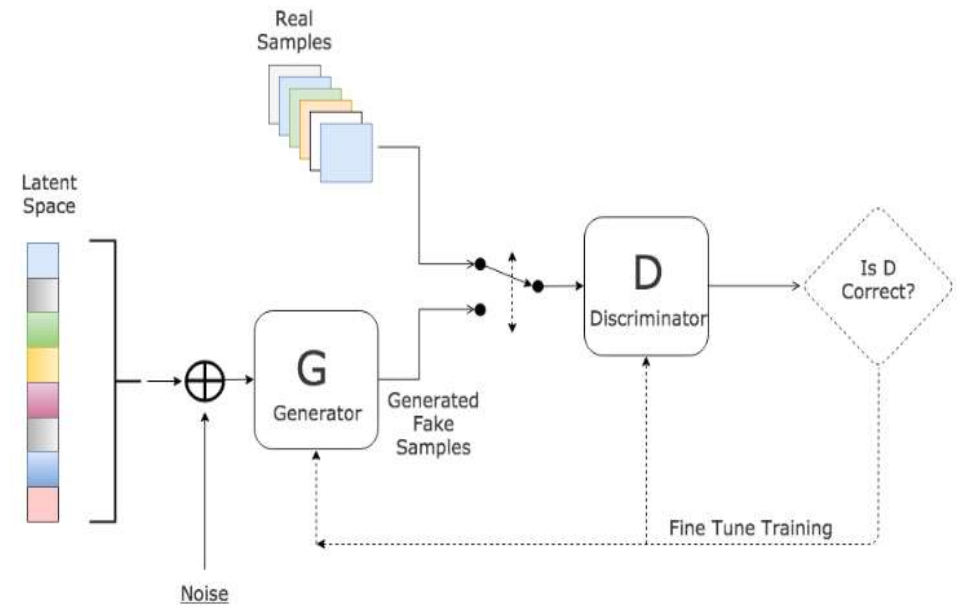
## # 1-D convolutional layer



## # GRU layer



## # Generative adversarial networks



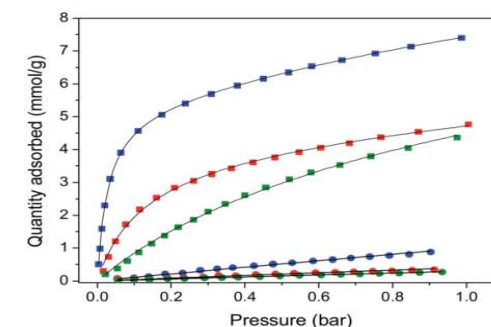


## 02.2 Research 2

# Introduction – Adsorption properties of MOFs

Property	Definition
Adsorption isotherm	Gas uptake as a function of pressure at const. T
Heat of adsorption	Heat given off by gas adsorption
Binding energy	Lowest potential energy between host material and guest molecule
Henry adsorption coefficient	Slope of the adsorption isotherm at low external P

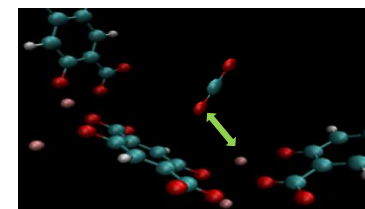
**Adsorption isotherm**



**Heat of adsorption**

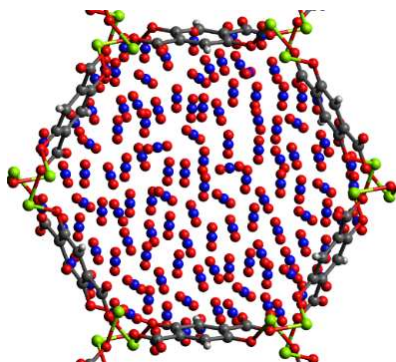
$$Q_{st} = \frac{\partial K_H}{\partial \beta} \text{ where } \beta = k_B T$$

**Binding energy**



# Introduction – Calculation of henry adsorption coefficient

Random insertion of guest molecules in unit cell



Potential energy calculation  
more than 100,000  
 $U_i$



Henry adsorption coefficient ( $K_H$ ) calculation

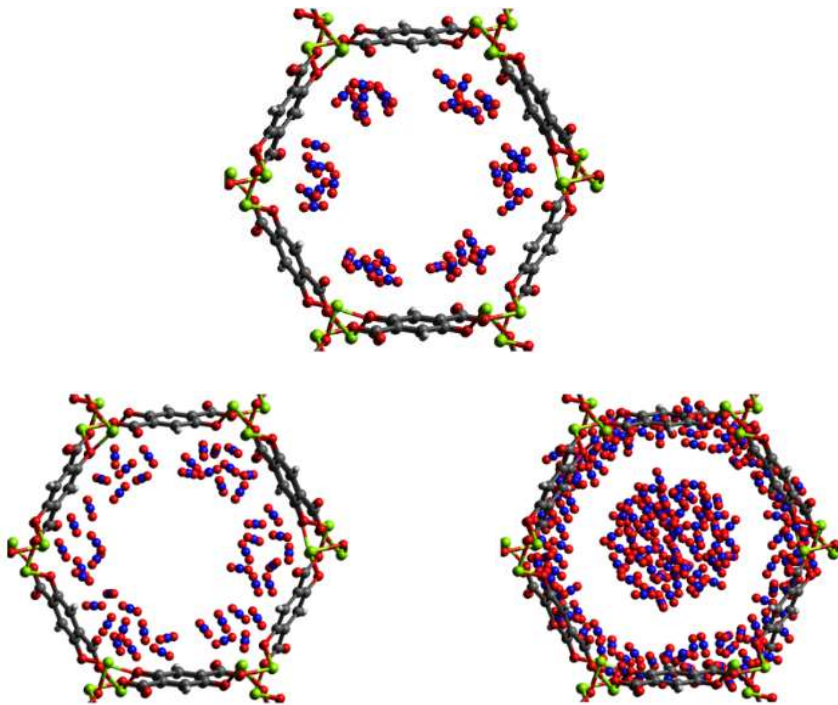
$$K_H = \frac{1}{N} \sum_{i=1}^N \exp\left(-\frac{U_i}{k_B T}\right)$$

CO<sub>2</sub>-MOFs DFT interaction energy:  $U_i = U_{(MOF+CO_2,i)} - (U_{MOF} + U_{CO_2,i})$

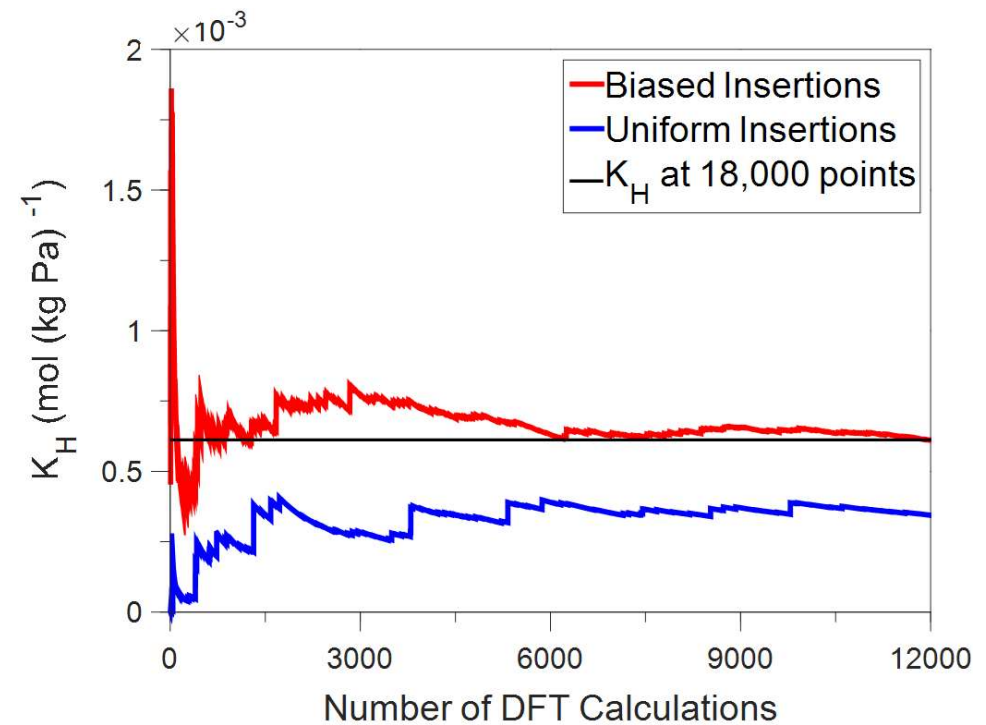
Number of CO <sub>2</sub> molecules ( $N$ )	Time to compute $U_{MOF}$ (minutes)	Time to compute $U_{CO_2}$ (minutes)	Time to compute $U_{MOF+CO_2}$ (minutes)	Total time
1	1 to 2	0.1 to 0.2	1 to 2	<b>5-10 minutes</b>
100,000	1 to 2	10,000 to 20,000	100,000 to 200,000	<b>3 to 5 months</b>
1,000,000	1 to 2	100,000 to 200,000	1,000,000 to 2,000,000	<b>2 to 4 years</b>

# Results – Efficient sampling to reduce computational costs

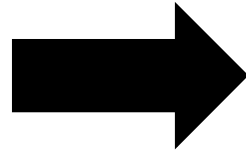
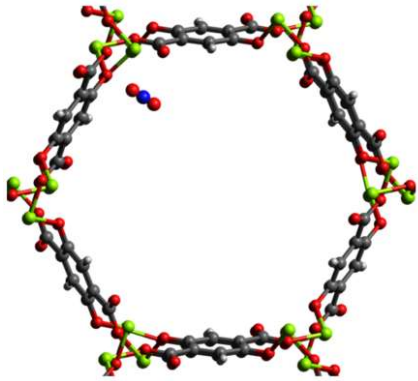
## # Efficient sampling



## # Reduced computational costs

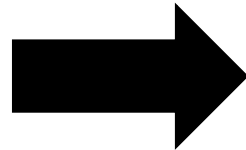


# Introduction – Efficient calculation of $K_H$ with DL



$$E = F(r)$$

Approximation of Energy potential  
of gas adsorption

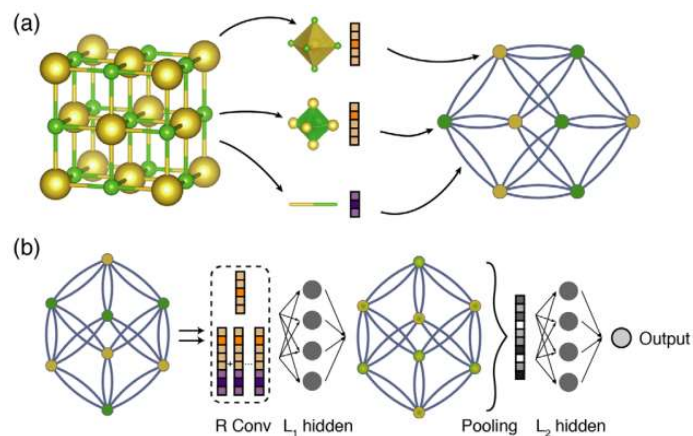


Reduction of computation cost for adsorption  
property calculation

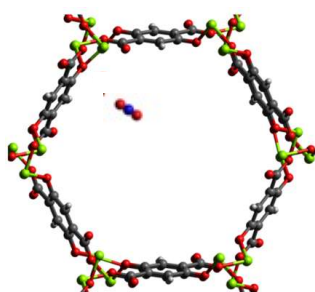
## # Hypothesis/motives

- If we have a model to calculate the adsorption energy of gases, we can reduce the number of calculation.
- This leads to fast adsorption property calculation.

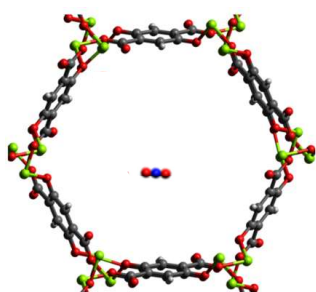
# Methods – Graph input generation (adsorption state)



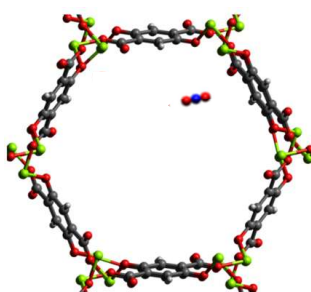
- Crystals are also represented as Graph.
- Adsorption in crystal is also represented as Graph.



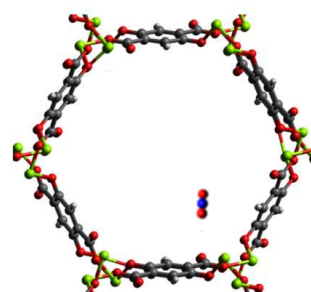
**Graph1**  
**U1**



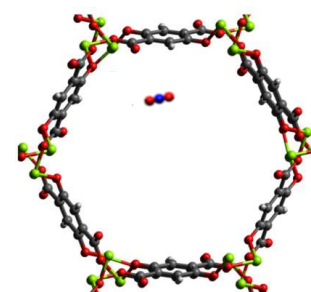
**Graph2**  
**U2**



**Graph3**  
**U3**



**Graph4**  
**U4**



**Graph5**  
**U5**

# Summary

---

01

Analyzed various properties of MOFs using computational chemistry and deep learning.

02

Analyzed gas sensing performance of chemi-resistors using computational chemistry and deep learning.

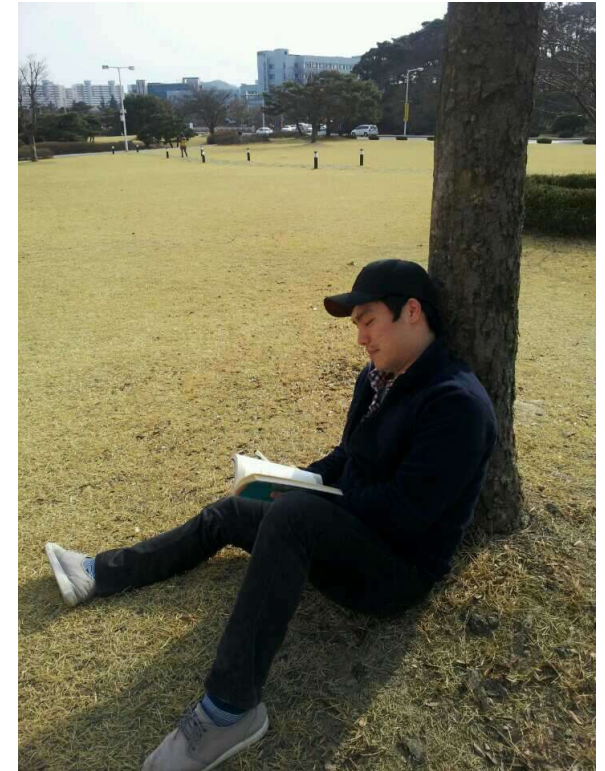
# 03 YH as Kaggle



# 취미!! 특기!!

---

# 취미!! 특기!!



취미!! 특기!!

---

kaggle

취미!! 특기!!

---

kaggle

머신러닝 대회

## 캐글 as a company

---

- 2010년 설립된 빅데이터 솔루션 대회 플랫폼 회사
- 2017 년 3월에 구글에 인수

kaggle

## 캐글 as a community

---

- 현재 200만명의 회원 보유
- Data science, ML, DL 을 주제로 모인 community

kaggle

# Competition - Data Race for 데이터 과학자!

---

기업, 정부기관, 단체, 연구소, 개인

**Dataset  
With Prize**

kaggle

**Dataset & Prize  
개발 환경(kernel)  
커뮤니티(follow, discussion)**

전 세계 데이터 사이언티스트

# 현재 프로필

## 다수의 머신 러닝 대회 경험 (약 30번 이상)

ex) 정형 데이터, 이미지 대회, 자연어 처리 대회.

## 유수의 국제 머신 러닝 대회 수상 이력

- 분자 특성 예측 대회 **3등 0.11%** (2749팀).
- 안구 당뇨병성 망막증 분류 대회 **3등 0.10%** (2943팀).
- 신용카드 fraud detection 대회 **11등, 0.17%** (6381팀).

## 머신 러닝 대회 플랫폼 상위 랭커 (마스터)

- 머신 러닝 랭킹 전세계 **0.10% (109/122,351)**.
- 국내 **3등**.

**YouHan Lee**  
Ph.D student at KAIST  
Daejeon, South Korea  
Joined 2 years ago · last seen in the past day  
in

**Competitions Master**

Home Competitions (30) Datasets Kernels (33) Discussion (447) ...

Competitions Master	Datasets Contributor	Kernels Master
<b>Current Rank</b> <b>109</b> of 125,456	<b>Unranked</b>	<b>Current Rank</b> <b>97</b> of 108,188
<b>Highest Rank</b> <b>95</b>		<b>Highest Rank</b> <b>36</b>
3 6 7	0 0 0	4 6 15
<b>APTOS 2019 Bli...</b> 4 months ago Top 1% 3rd of 2931	No dataset results	<b>My EDA - I want...</b> a year ago 196 votes
<b>Predicting Mole...</b> 4 months ago Top 1% 3rd of 2749		<b>Simple quant fe...</b> a year ago 136 votes
<b>IEEE-CIS Fraud ...</b> 3 months ago Top 1% 11th of 6381		<b>YH EDA - I want...</b> 10 months ago 102 votes



# 해왔던 대회들

---

Porto: 고객이 내년에 자동차 **보험금 청구**를 할 것인가?

Home Credit: 고객이 앞으로 **대출 상환**을 할 것인가?

Costa rican: 고객의 **소득 수준**을 ML 로 구분하라

Elo: 거래 내역 데이터를 가지고, **고객 충성도**를 예측하라

New York taxi: Taxi **탑승 시간**을 예측하라

직방: **아파트 거래가격** 예측하라

INFOCARE: **아파트 경매가격** 예측해라

# 해왔던 대회들

---

Tensorflow: 30개 단어를 구분하는 AI 만들어라

Quora: 성실한, 불성실한 질문을 구분해내라

Doodle: 340개의 클래스 별 낙서를 ANN 으로 구분하라

Protein: 28개의 클래스 별 Protein 을 ANN 으로 구분하라

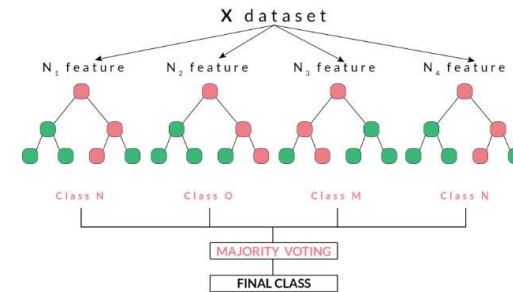
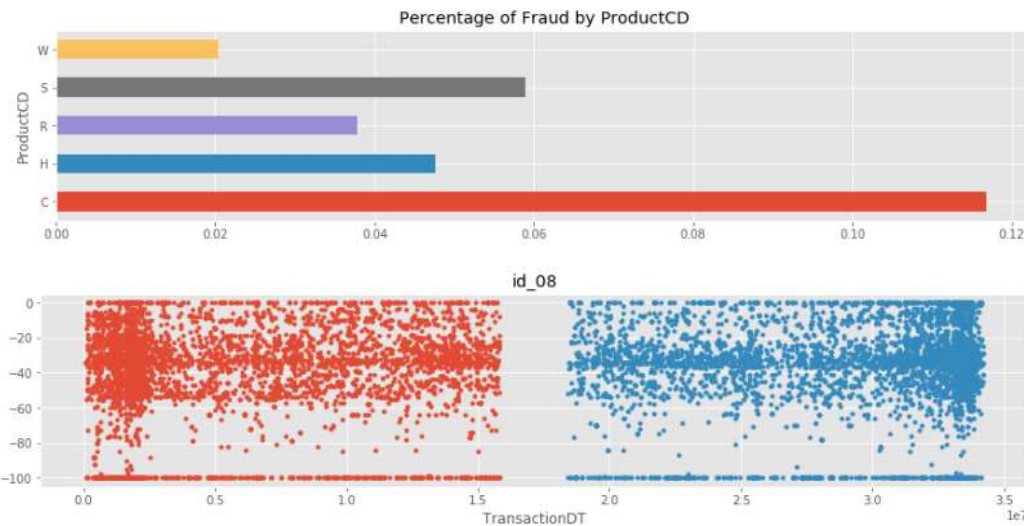
Airbus: 바다 위 배를 찍은 위성 사진에서 배의 위치를 찾아내라

Statoil: 바다 위 빙산과, 배를 구분하라

# My capacity for AI – Tabular data

정형

- Exploratory data analysis.
  - Visualization.
- Feature engineering, preprocessing techniques.
  - Category encoding, feature engineering.
- Conventional ML algorithms.
  - Lightgbm, xgboost, catboost.



dmlc  
**XGBoost**

LightGBM

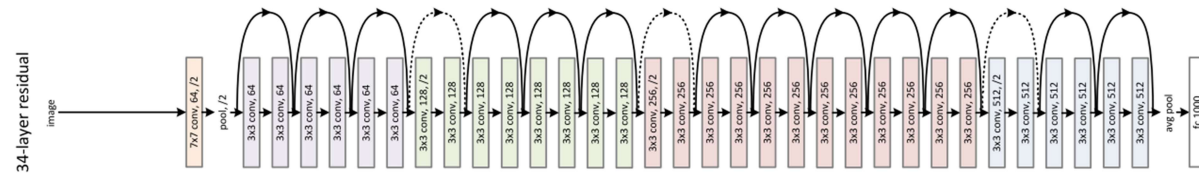
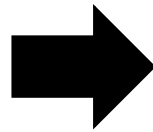
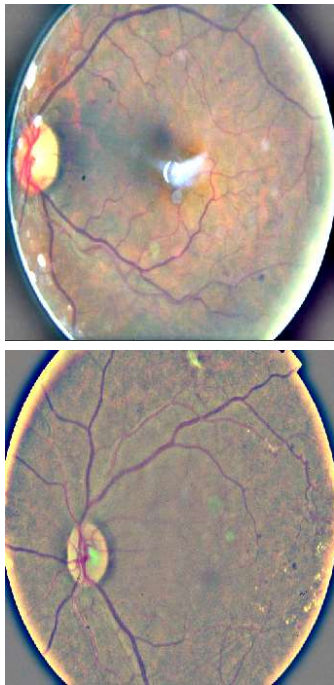
 CatBoost

신용카드 Fraud detection 대회: 11등 0.17% (6347팀 참여).

# My capacity for AI – Computer vision

이미지

- Image classification.
  - Category classification, regression.
- Object detection.
  - Semantic segmentation, instant segmentation.



Pre-trained  
model  
(ImageNet)



Fine-tuned model  
(eyeball)

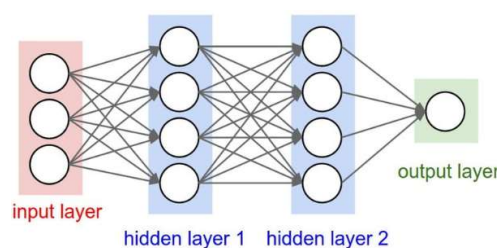
당뇨성 망막 병증 중증도 분류 대회: **3등 0.10%** (2943팀 참여).

# My capacity for AI – Graph data

그래프  
(graph data)

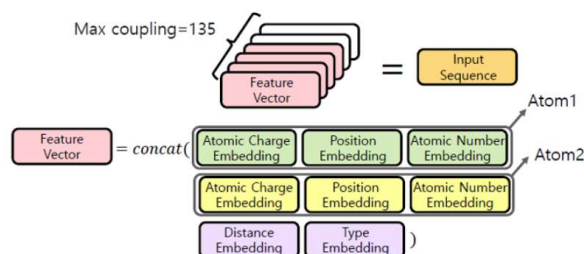
- Graph neural networks.
- Embedding with transformer.

Graph neural network



Molecule transformer

Chemical  
property



# My capacity for AI – Natural Language Processing

자연어 처리

- Various preprocessing.
- Various embedding with LSTM.
- BERT based models.

2013

2017

2018

Word **encoding**

One hot encoding

Bag of words

TF-IDF

Word **embedding**

Word2Vec  
(2013)

Fasttext  
(2017)

GloVe  
(2014)

**Sentence embedding**

ELMo  
(2018)

GPT  
(2018)

BERT  
(2018)

Xlnet  
(2019)

RoBERTa  
(2019)

독성 문장 분류 대회: **26등 0.80%** (3165팀 참여).

## Profile

---

### # 활동

캐글 코리아 페이스북 페이지 운영자 (현재 7,298명)

# 캐글 코리아

## Kaggle Korea

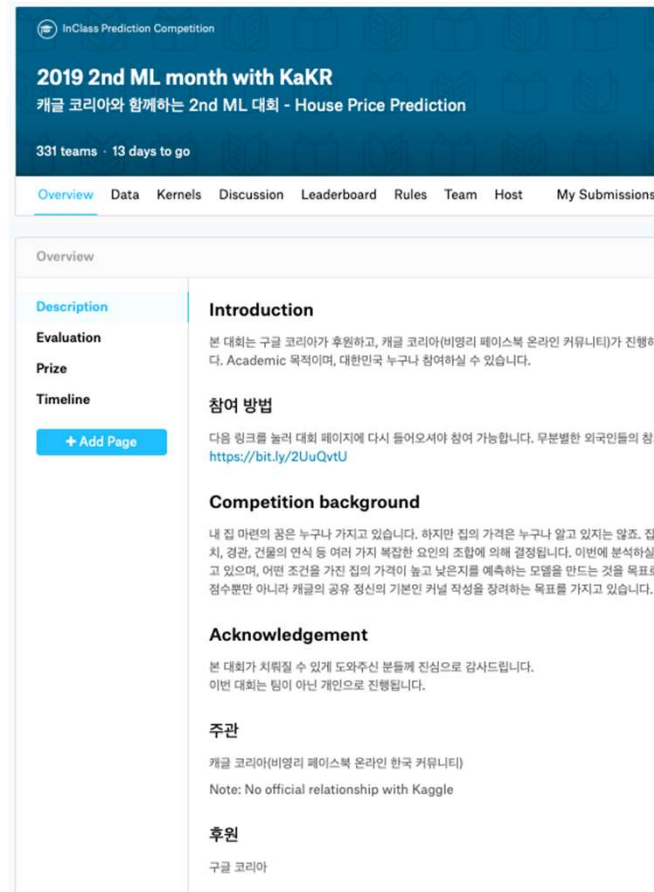
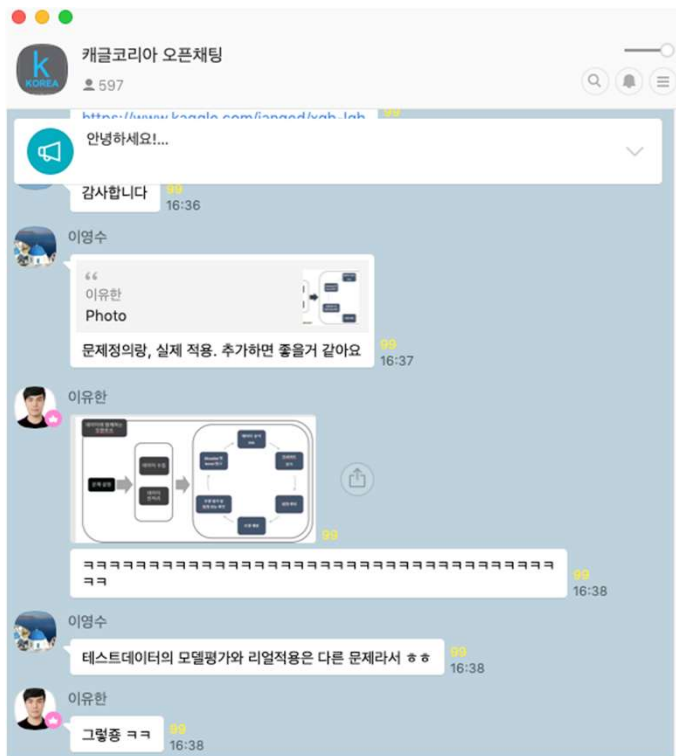
Non-Profit Facebook Group Community

- 2018년 6월 12일 Kaggle 본사로부터 캐글 코리아 이름 사용 허가
- 2018년 6월 15일 캐글 코리아 개설

# Profile

## # 활동

캐글 코리아 페이스북 페이지 운영자 (현재 7,298명)





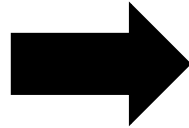
캐글 (머신러닝 대회) 만한  
공부 방법이 없어요 😊

04

# Dream of YH

# Transferability of AI

머신 러닝의 장점



데이터 도메인이 달라도  
같은 방법론을 적용할 수 있다.

시계열 데이터  
From gas sensor

시계열 데이터  
From KAERI

이미지 데이터  
From 안구

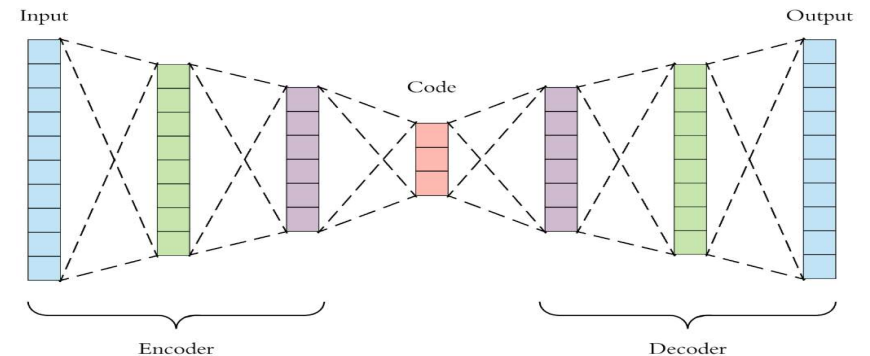
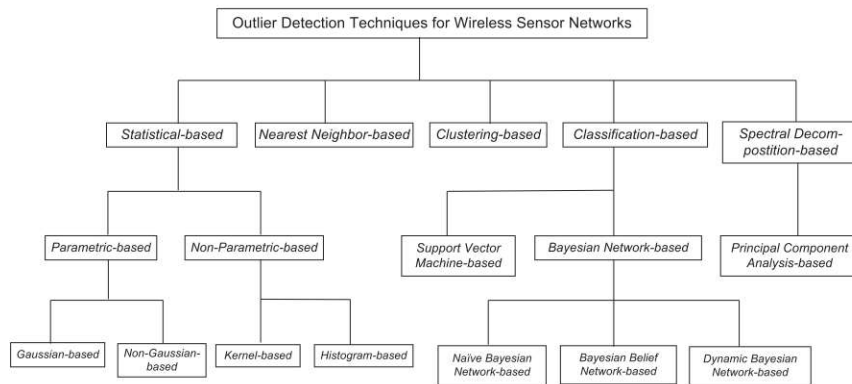
이미지 데이터  
From KAERI

- Various time series analysis.
- Anomaly detection using deep learning.
- Classification, regression.

- Image classification.
  - Category classification, regression.
- Object detection.
  - Semantic segmentation, instant segmentation.

# 지원 부서 요구 역량 – 직무수행 내용

원자로 운전 중 발생하는 센서 데이터의 시계열 데이터 분석 및  
Anomaly detection 기법 연구



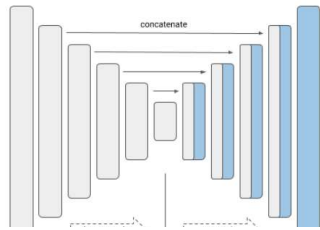
Zhang et al, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 12, SECOND QUARTER 2010

- 안전을 위해 모든 인공지능 기술을 동원하여 안전 극대화

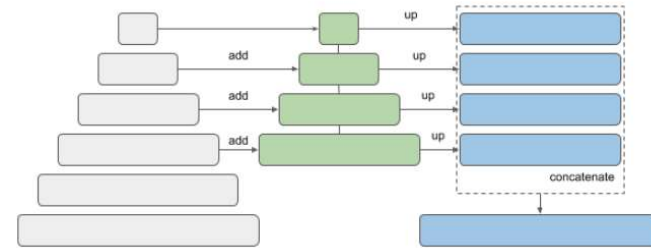
## 지원 부서 요구 역량 – 직무수행 내용

인공지능 기반 객체 인식 수행 및  
비파괴/의료영상을 위한 인공지능 기반 객체 인식.

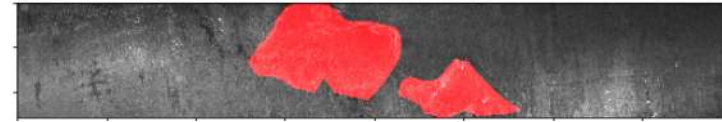
Unet



FPN  
(Feature Pyramid Networks)



예시) Surface defect on steel sheet.



- 최신 딥러닝 Computer vision 기술 활용해 자동화 실현

# KAERI 에서 해보고 싶은 꿈들

---

## # 세계 최고 수준의 안전한 에너지에 대한 비전

- 인공지능 기술 기반한 원전 내 센서 성능(속도, 민감도, 선택도) 극대화
- 인공지능 기반 위험요소 판단 기술 자동화

## # 안전재료 기술 개발

- 인공지능 기술 기반 재료 디자인

## # 원내 AI 역량 증진

- 스터디 운영

## # KAERI 소속으로 대회 1등 수상하기

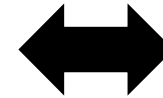
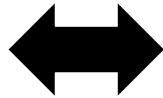
- 캐글 그랜드마스터 (국내 3 or 4 호!)

# Data scientist in KAERI

# 함께 공부해서 함께 일합시다

## # 도메인 전문가

- 데이터 명세
  - 도메인 지식이 핵심!
- 데이터 전처리
  - 그냥 주면 못해요
  - 알아야 처리를...
- AI 기본 지식
  - 다 되는게 아닙니다 ☺



## # AI 전문가

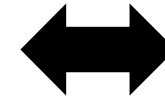
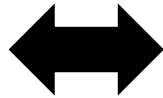
- 자유로운 아이디어 제시
  - 해봐야 알아요.
- 새로운 기술 트렌드 공유
  - 같이 공부해요.
- 유기적인 협업
  - 효율적인 툴 사용.



# 함께 공부해서 함께 일합시다

## # 도메인 전문가

- 데이터 명세
  - 도메인 지식이 핵심!
- 데이터 전처리
  - 그냥 주면 못해요
  - 알아야 처리를...
- AI 기본 지식
  - 다 되는게 아닙니다 ☺



## # AI 전문가

- 자유로운 아이디어 제시
  - 해봐야 알아요.
- 새로운 기술 트렌드 공유
  - 같이 공부해요.
- 유기적인 협업
  - 효율적인 툴 사용.

터놓고 공유하며 소통합시다.

잘 부탁드립니다.  
많이 배우겠습니다!