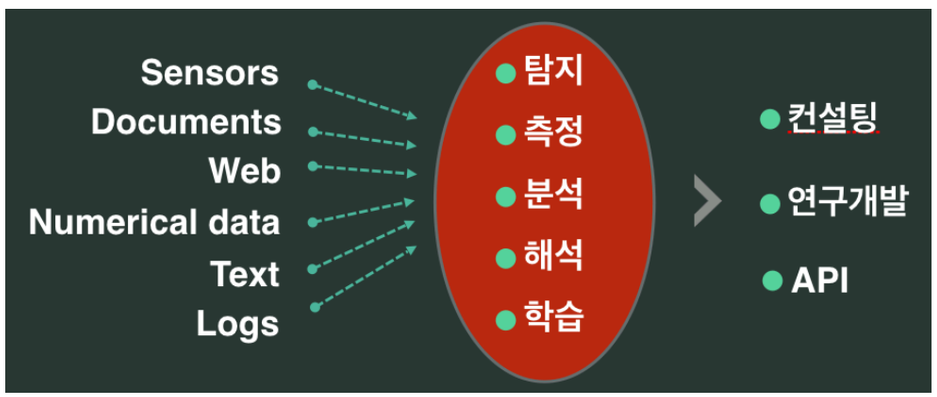
Pattern detection & interpretation

from Web log files to machine vibrations

데이터의 홍수.
늘 새로 생성되고 변화하는 데이터.
잘 활용하려면…

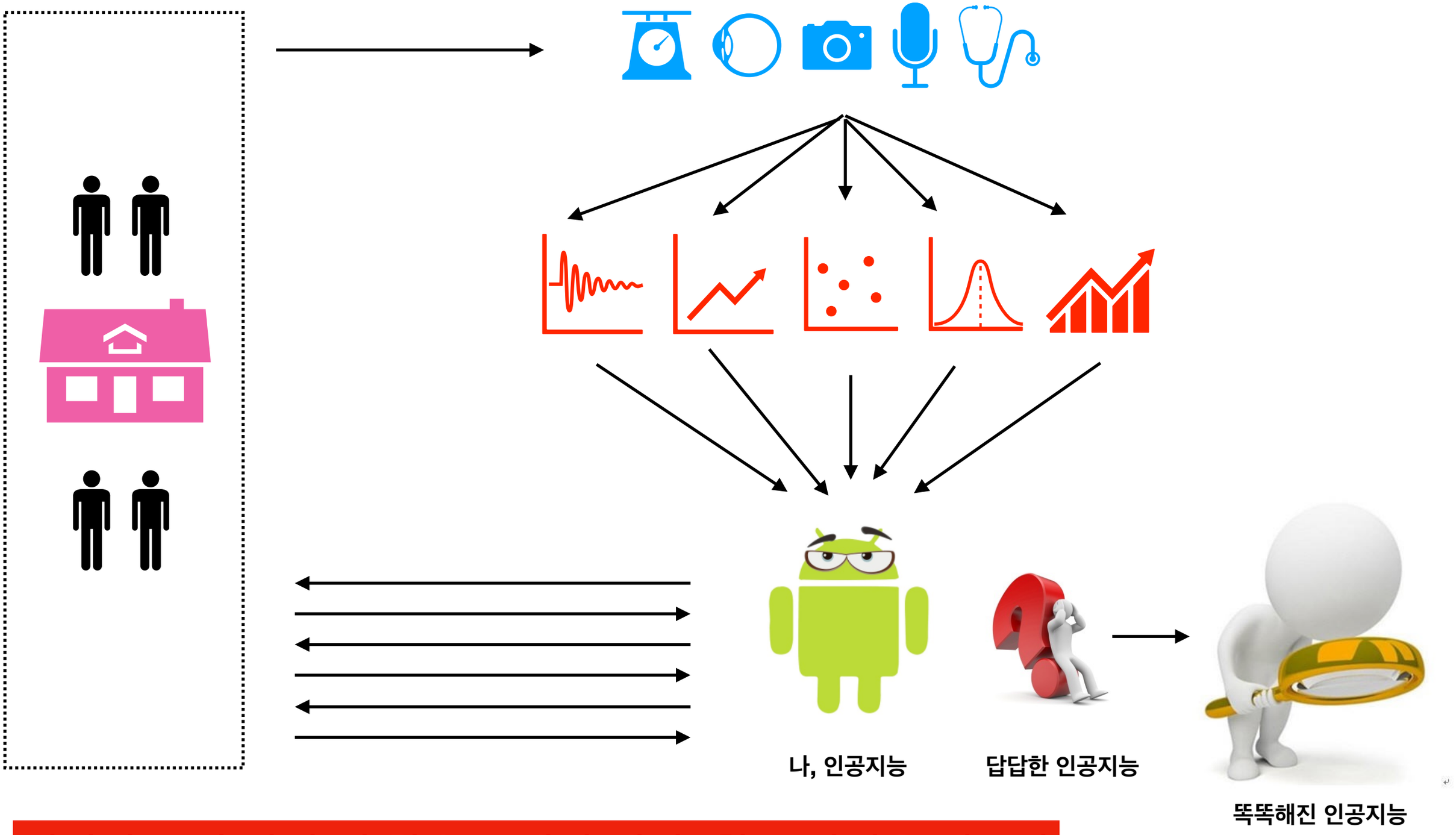
Who am I?

- **최진혁 (Jinhyuk Choi) Ph.D. @ KAIST 전산학과**
  - **Human-Computer Interaction / Machine Learning / Data Mining**
  - https://sites.google.com/site/choi31u/
- **ETRI 선임연구원, KAIST 연구교수**
  - 스마트 홈 미들웨어를 위한 데이터 마이닝 기술
  - Web & SNS Mining (usage analysis, text analysis…)
- **Inforience @Daejeon** (https://inforience.net/)
  - Information + Experience / Information + Science



근데 요즘은…

Goooooooooal



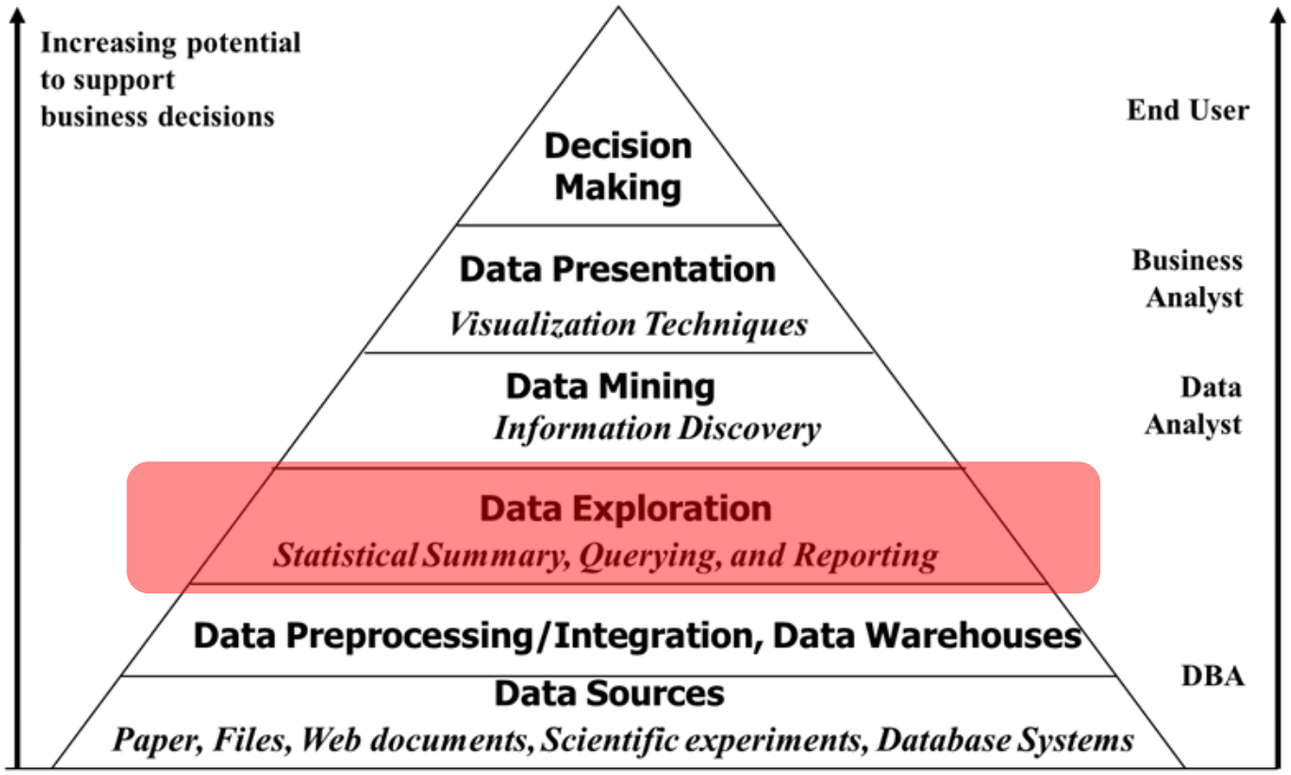
데이터 분석 과정 및 학습 모델의 생성+활용 과정을 대중화

This seminar ?

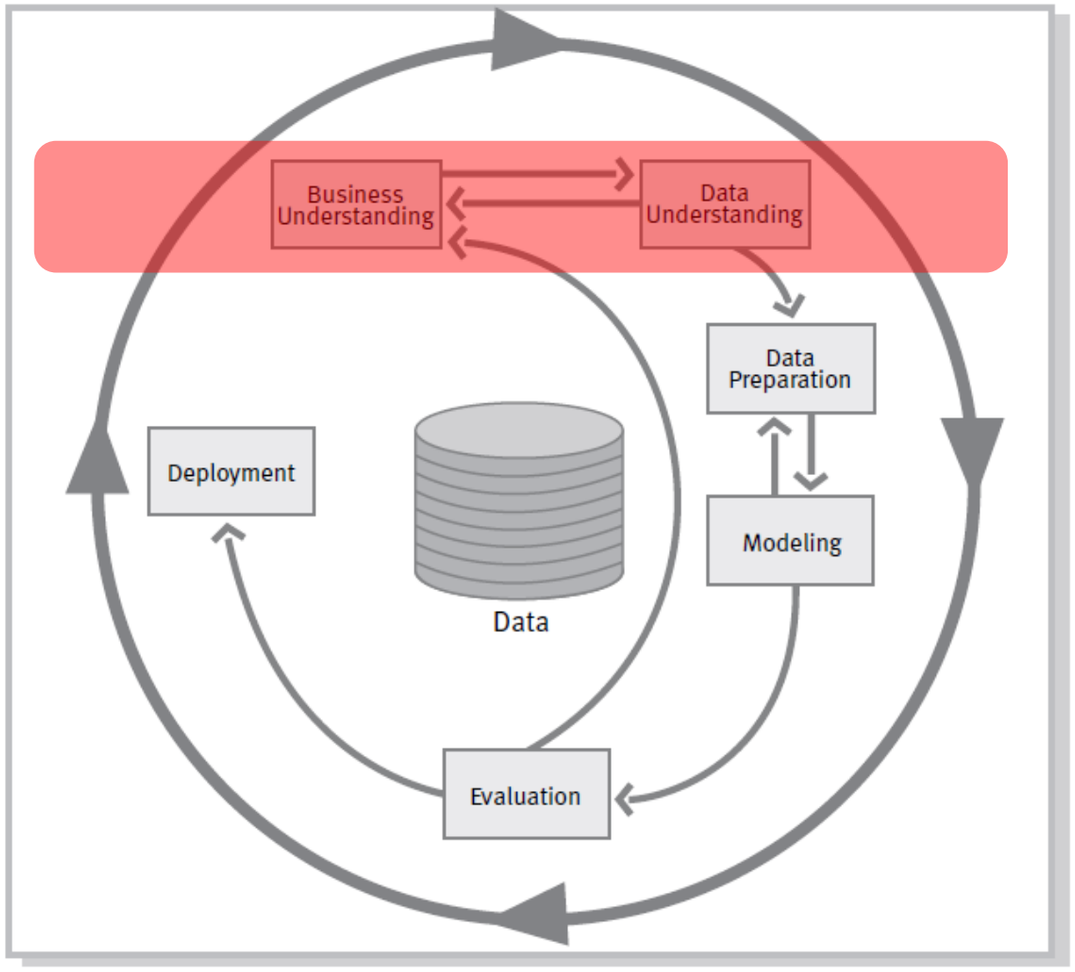
- 데이터 분석 과정 소개? (데이터 분석 프로젝트의 특성?)
- 특정 알고리즘이나 시스템의 개념 소개?
- Inforience 보유 기술 소개?
- Interaction btw. Human and Data?
- Anomaly detection 소개?

전부 다 조금씩… + 깨알 회사 홍보

데이터 활용 과정



Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011



CRISP 데이터 마이닝 프로세스 모델

Exploration vs. Confirmation

- Data is always incomplete
- Data is objective, but its collection and interpretation are subjective
- Any one set of data supports an infinite number of narratives
- The accuracy and relevance of data decay with time

Exploration vs. Confirmation in reality…

당신의 프로젝트는 어디에?

https://chiefmartec.com/2013/09/14-rules-data-driven-marketing/

Can you interpret this?

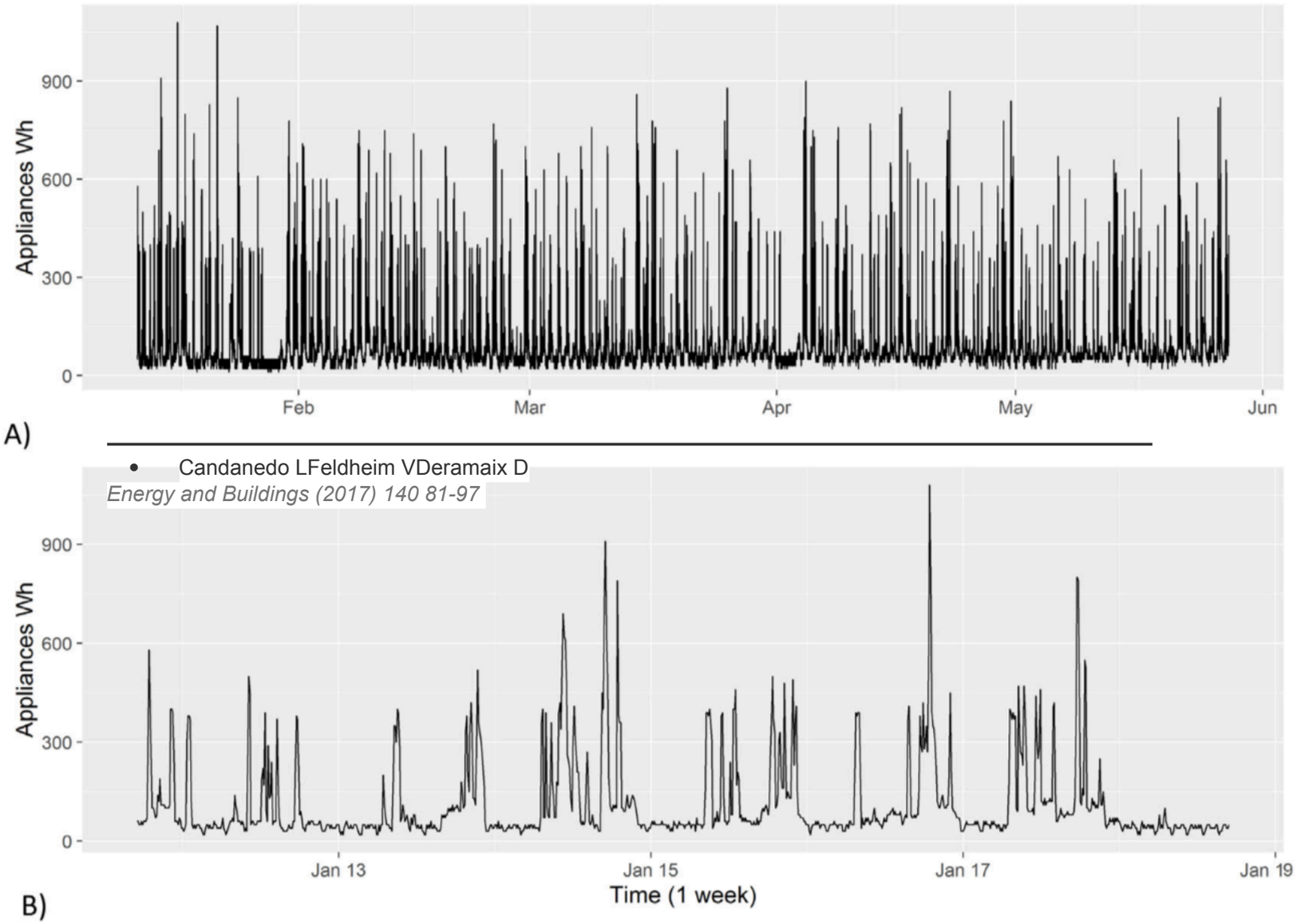


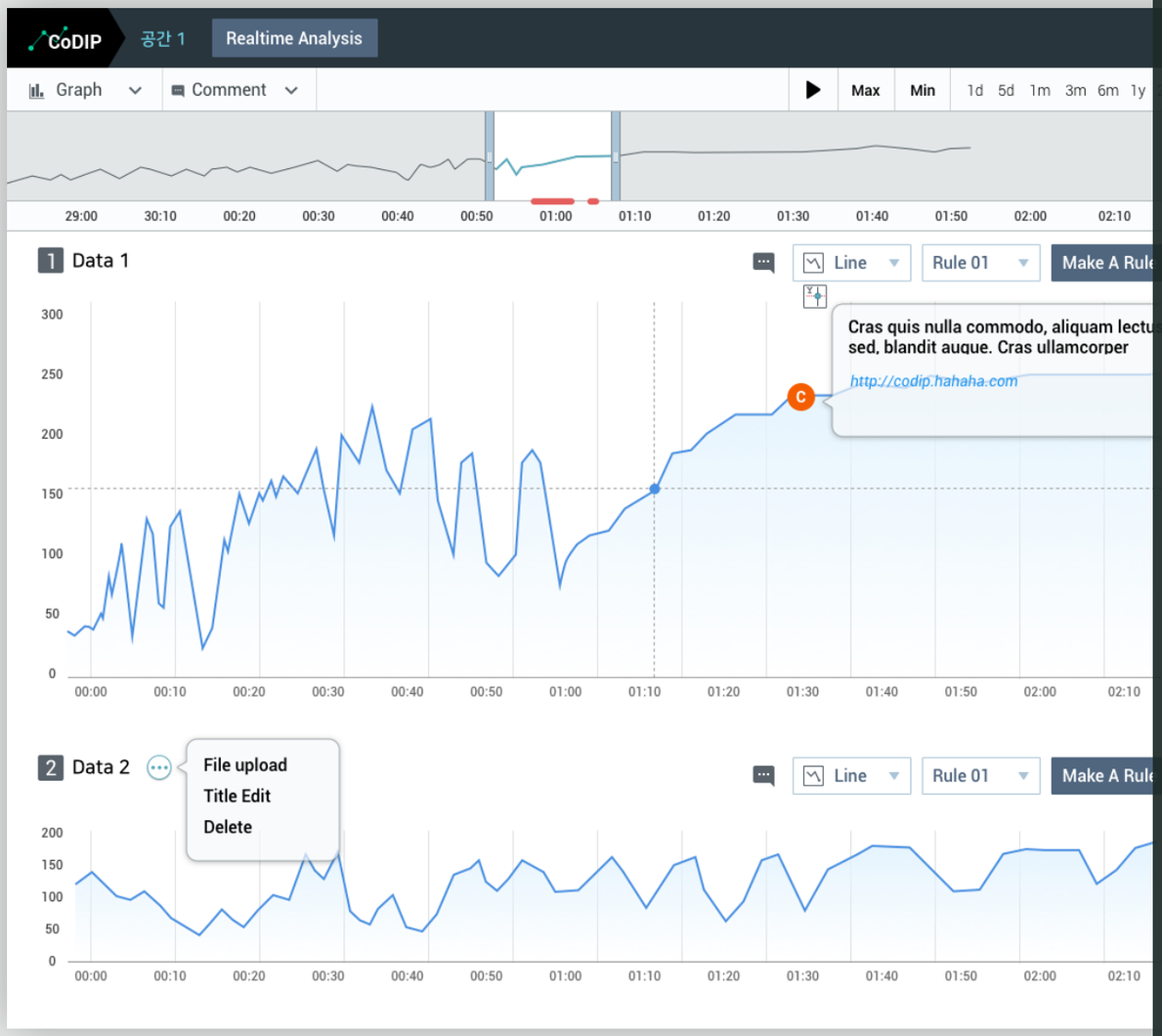
Fig. 7. (A) Appliances energy consumption measurement for the whole period, (B) A closer look at the first week of data.

Candanedo L, Feldheim V, Deramaix D, Data driven prediction models of energy use of appliances in a low-energy house, Energy and Buildings (2017), 140, 81-97

데이터에 대해 얼마나 알고있느냐가 학습 모델의 능력을 좌우합니다.

데이터에 대한 당신의 작은 배경 지식이 인공지능의 능력을 좌우합니다.

CoDIP



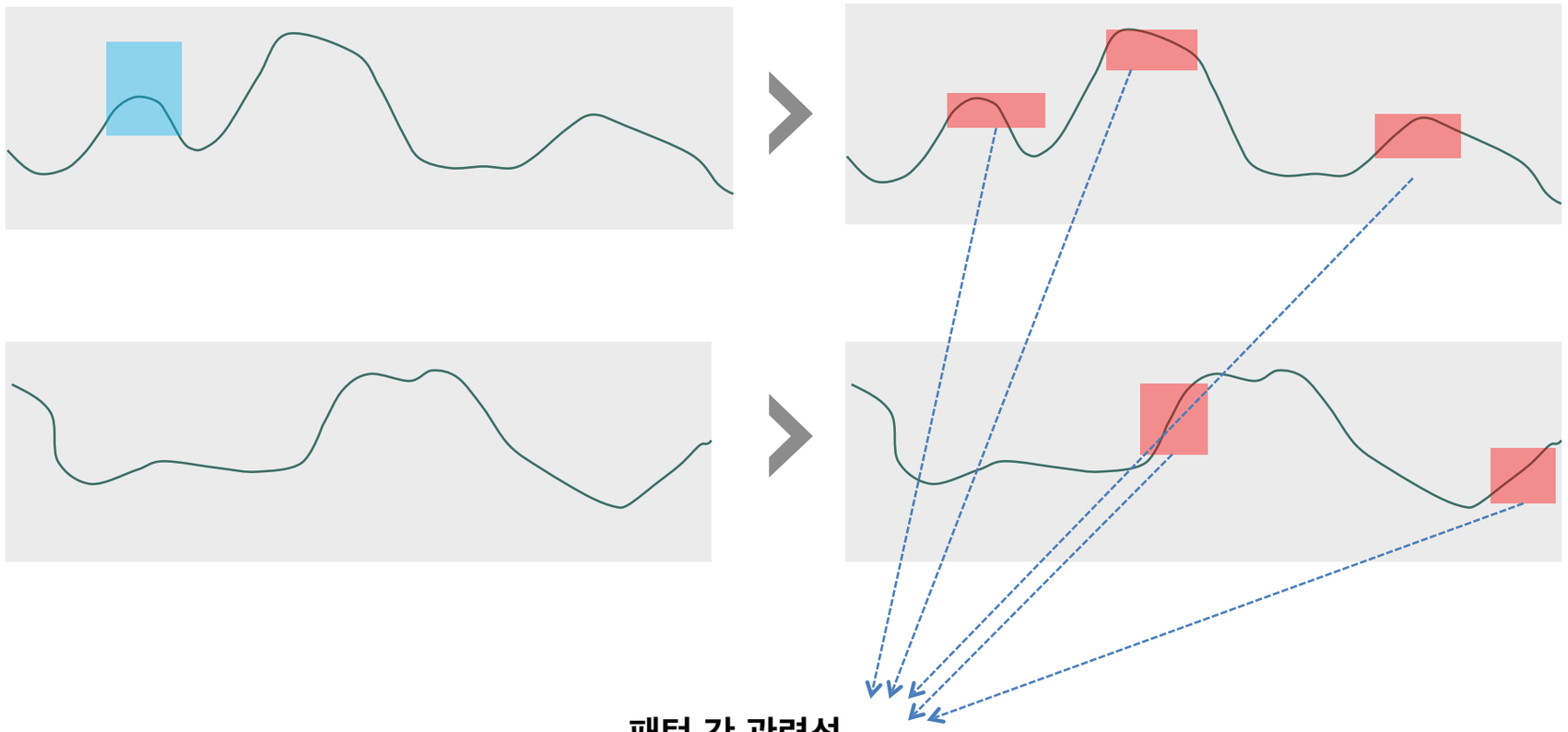
- Open API를 통한 데이터 연결 (업로드)
- 데이터에 포함된 패턴 탐색 (수동, 자동)
- 유의미한 패턴에 대한 해석 기록
- 패턴과 해석 저장
- 공유, 협업
- 탐색 및 해석 결과를 실시간 패턴 탐지 및 사용자와의 인터랙션에 활용

2 ways to teach CoDIP

Query

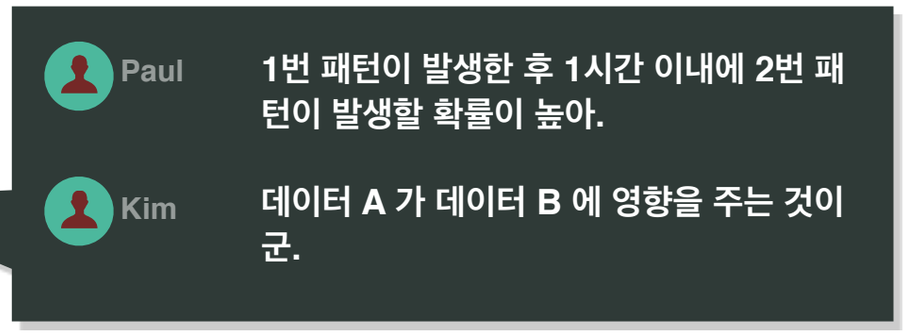
탐색할 패턴 설정 및 입력

패턴 탐지 결과



패턴 간 관련성

5. 해석 입력과 공유



6.실시간으로 입력되는 데이터로부터 패턴 탐지

핵심 기능 요소

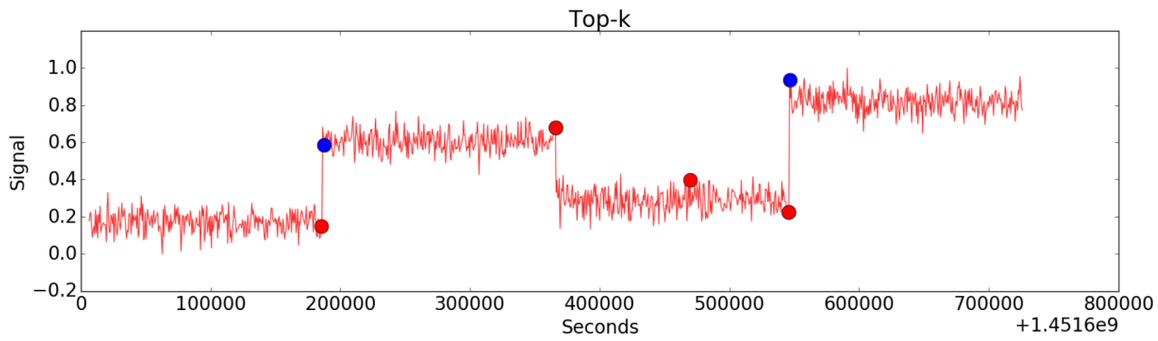
항목	개요	알고리즘	시스템
VA	Visual analysis	완료	완료
PA	Automatic pattern detection	완료	–
QA	Query analysis	완료	완료
MA	Machine learning	완료	–
Communicator	User 와의 대화 인터페이스	–	–

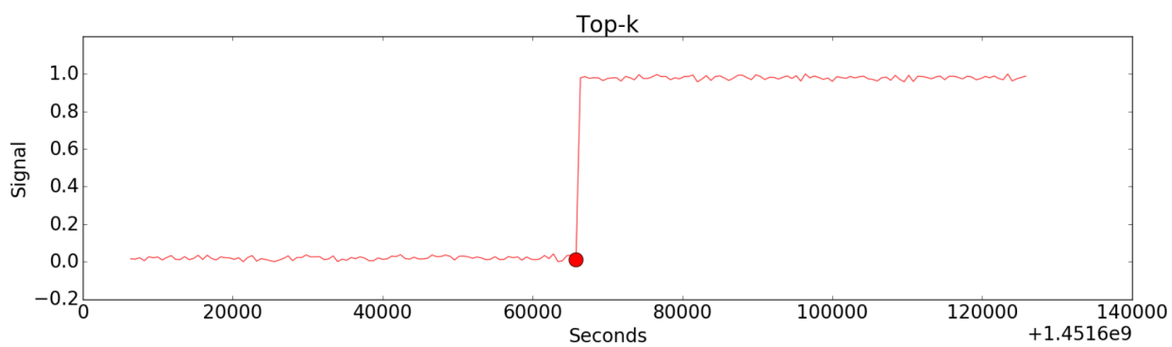
핵심 기능 요소

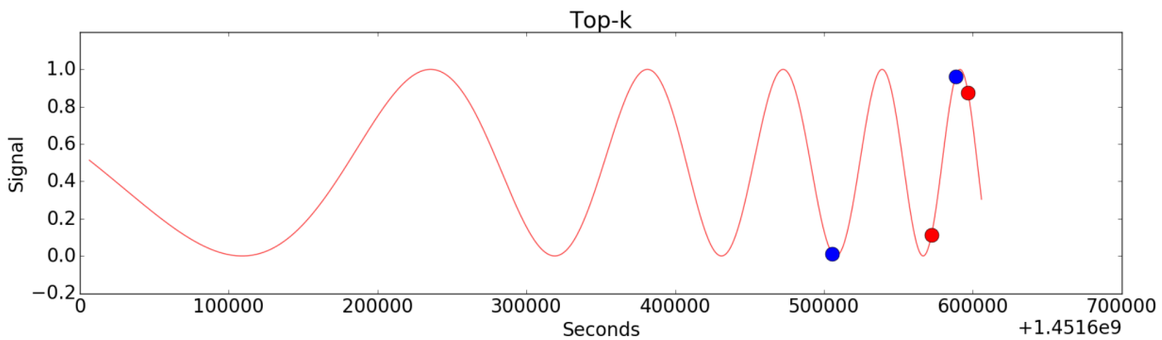
VA

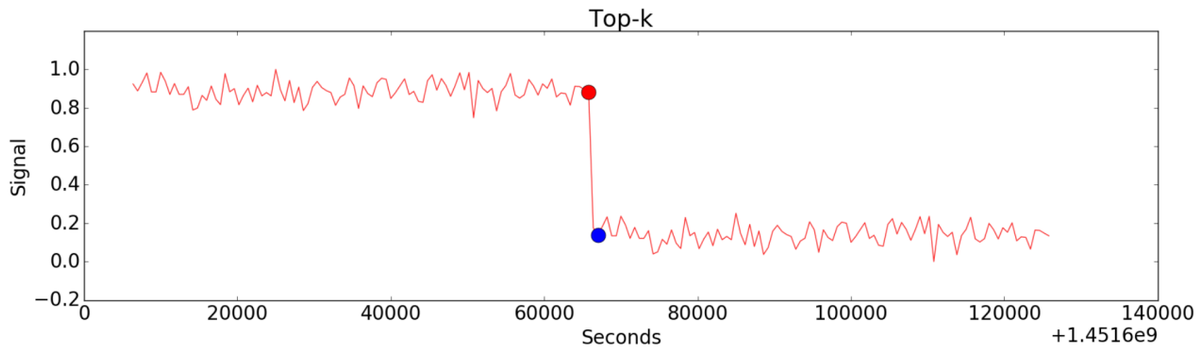
핵심 기능 요소

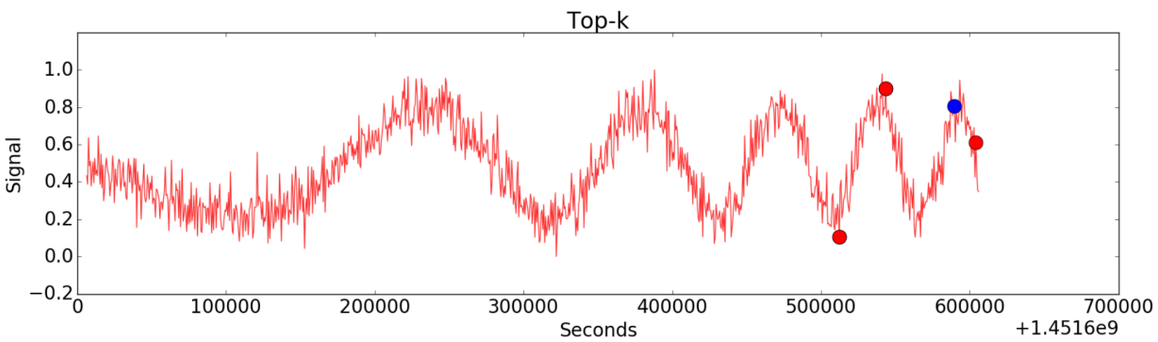
PA







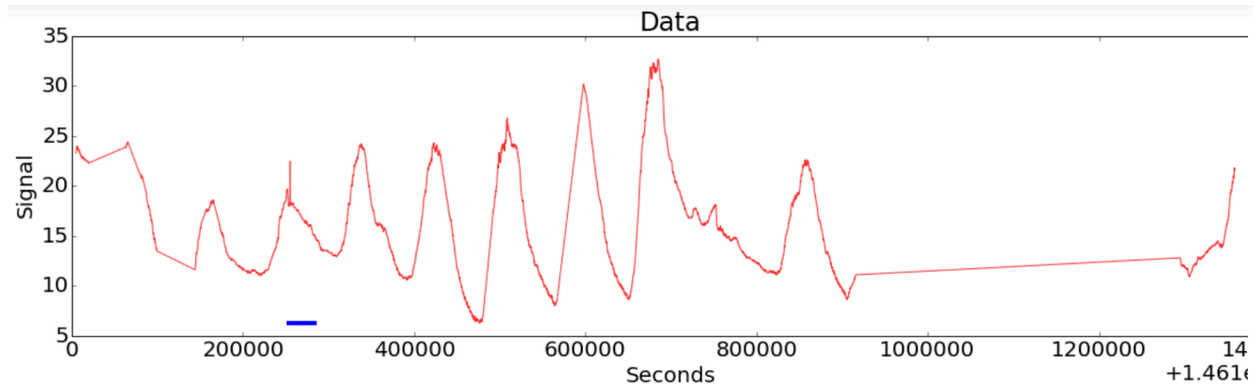


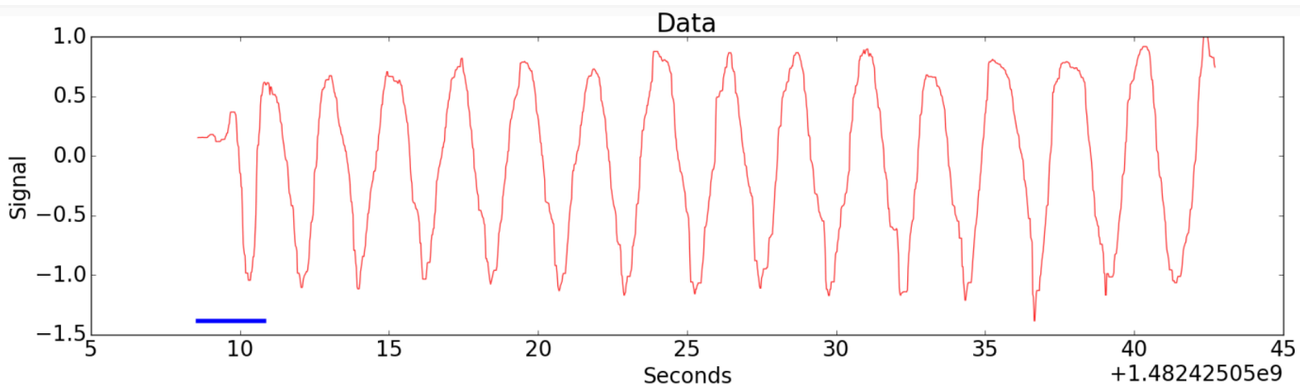


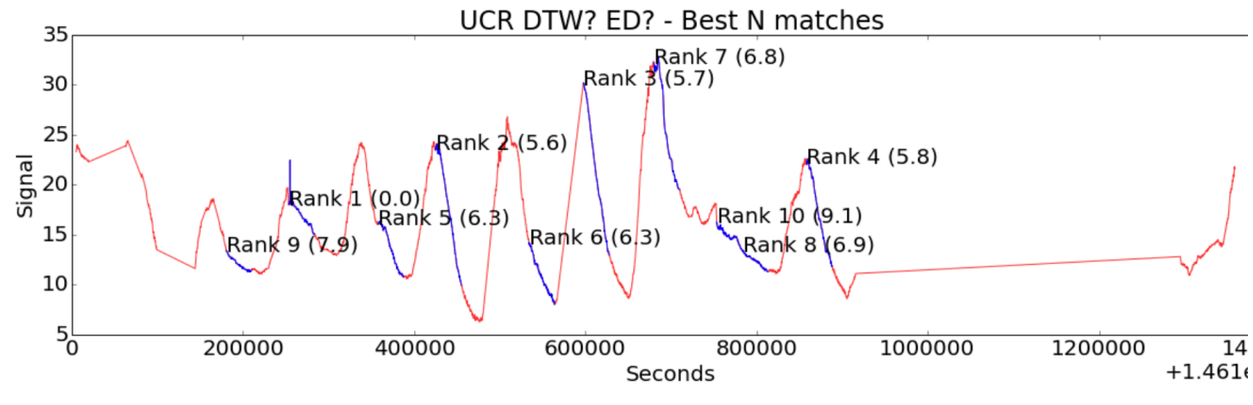
다양한 종류의 anomaly를 놓치지 않는 빠른 알고리즘

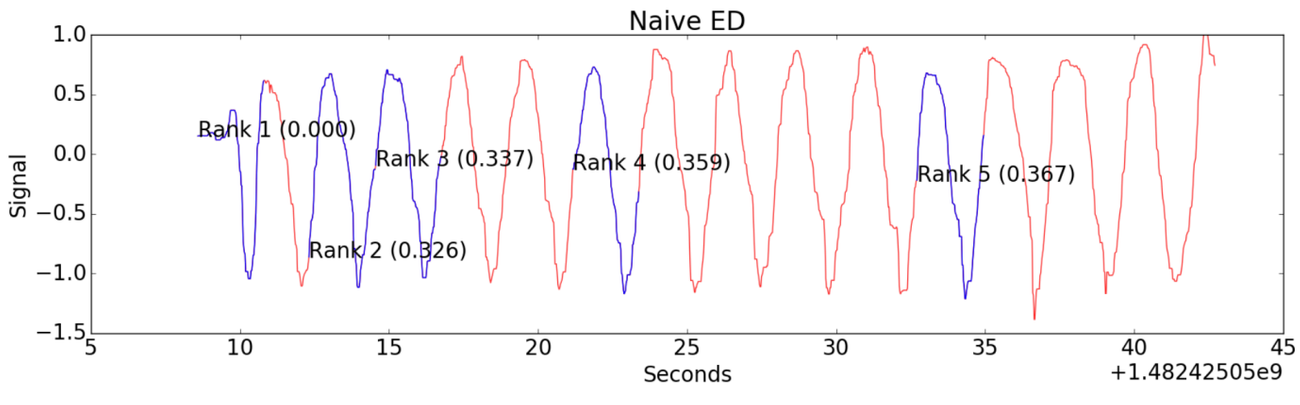
핵심 기능 요소

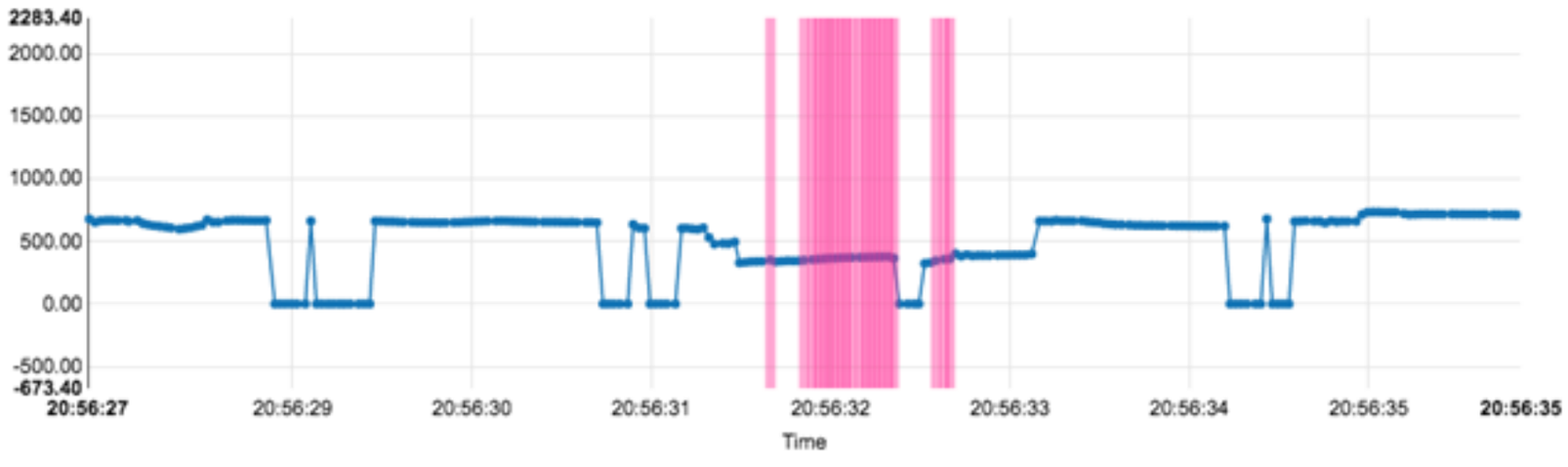
QA











핵심 기능 요소

MA - Hidden Markov Model

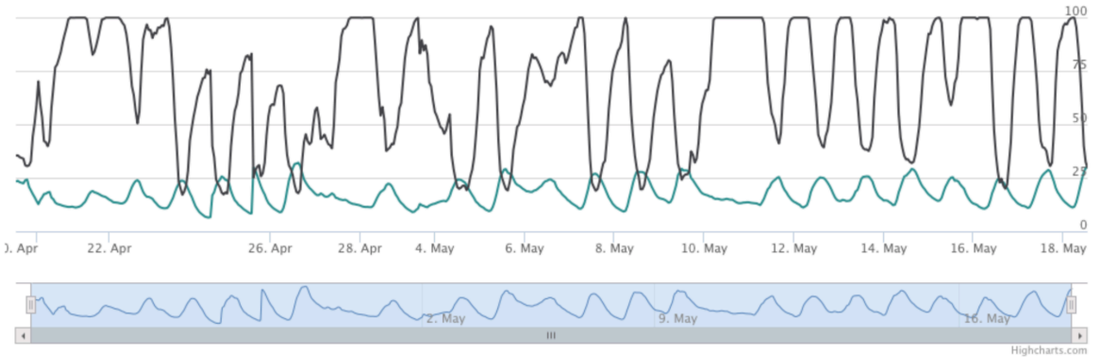


그림 #3. 약 1개월 간 수집한 육묘장 내 온,습도 데이터의 전체 패턴 (1분당 1샘플 수집)

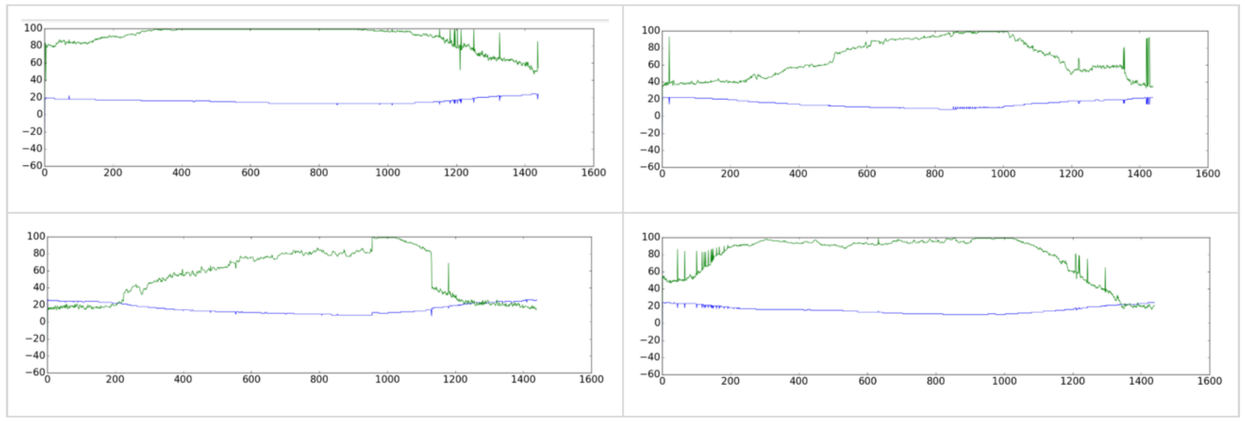
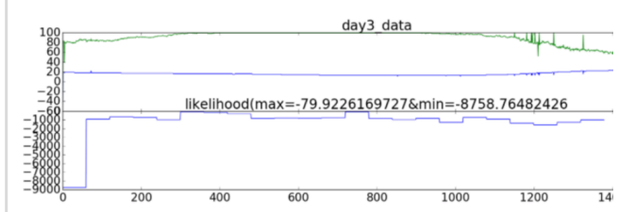
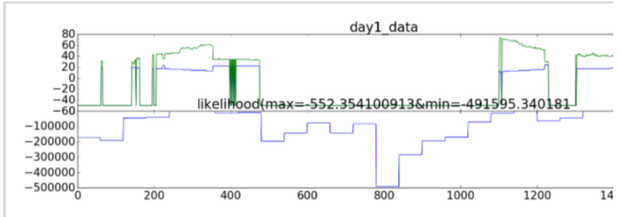


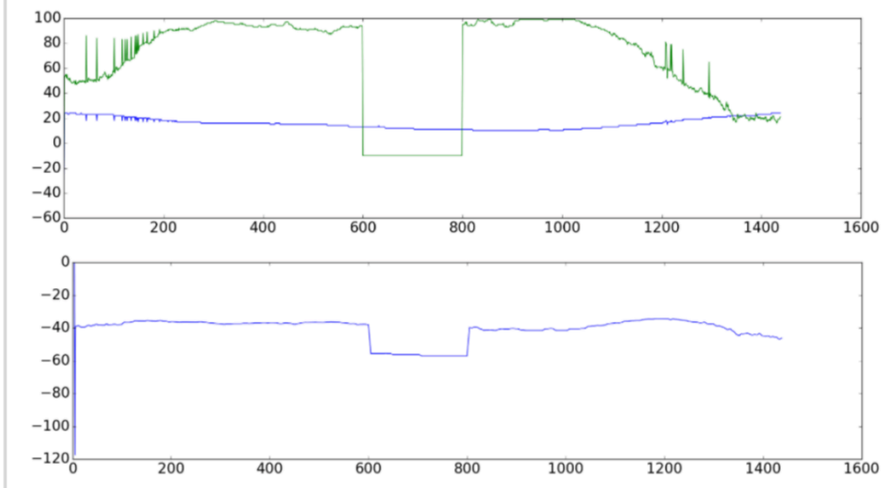
그림 #4. 비교적 정상적인 하루 동안의 온습도 변화 (1일 당 1440 샘플)



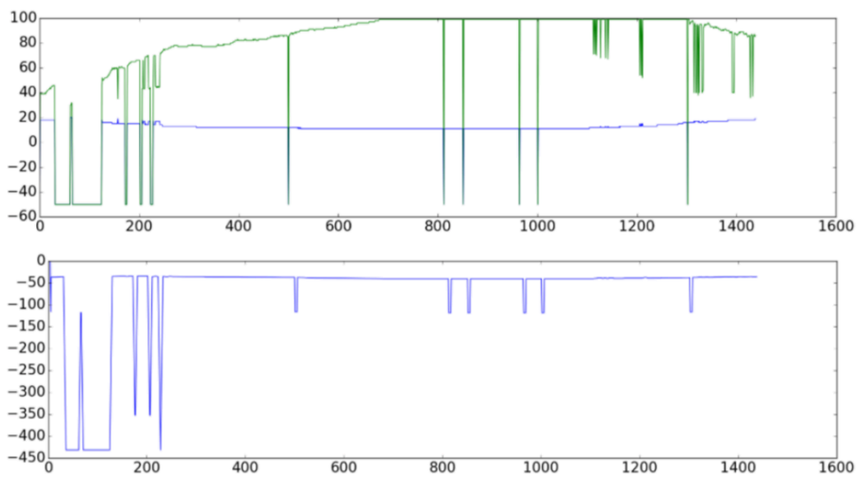
정상적인 데이터의 경우 likelihood 값이 전체적으로
을 유지



비정상적인 패턴을 포함한 경우 likelihood 값이 현저히
아지는 구간이 발견됨



하루에 1440 번 수집되는 데이터 중에서 600~800 번째 샘플 데이터에 인위적으로 노이즈를 삽입하였을 때 해당 타이밍의 데이터가 입력되자마자 노이즈의 존재 여부를 탐지함.



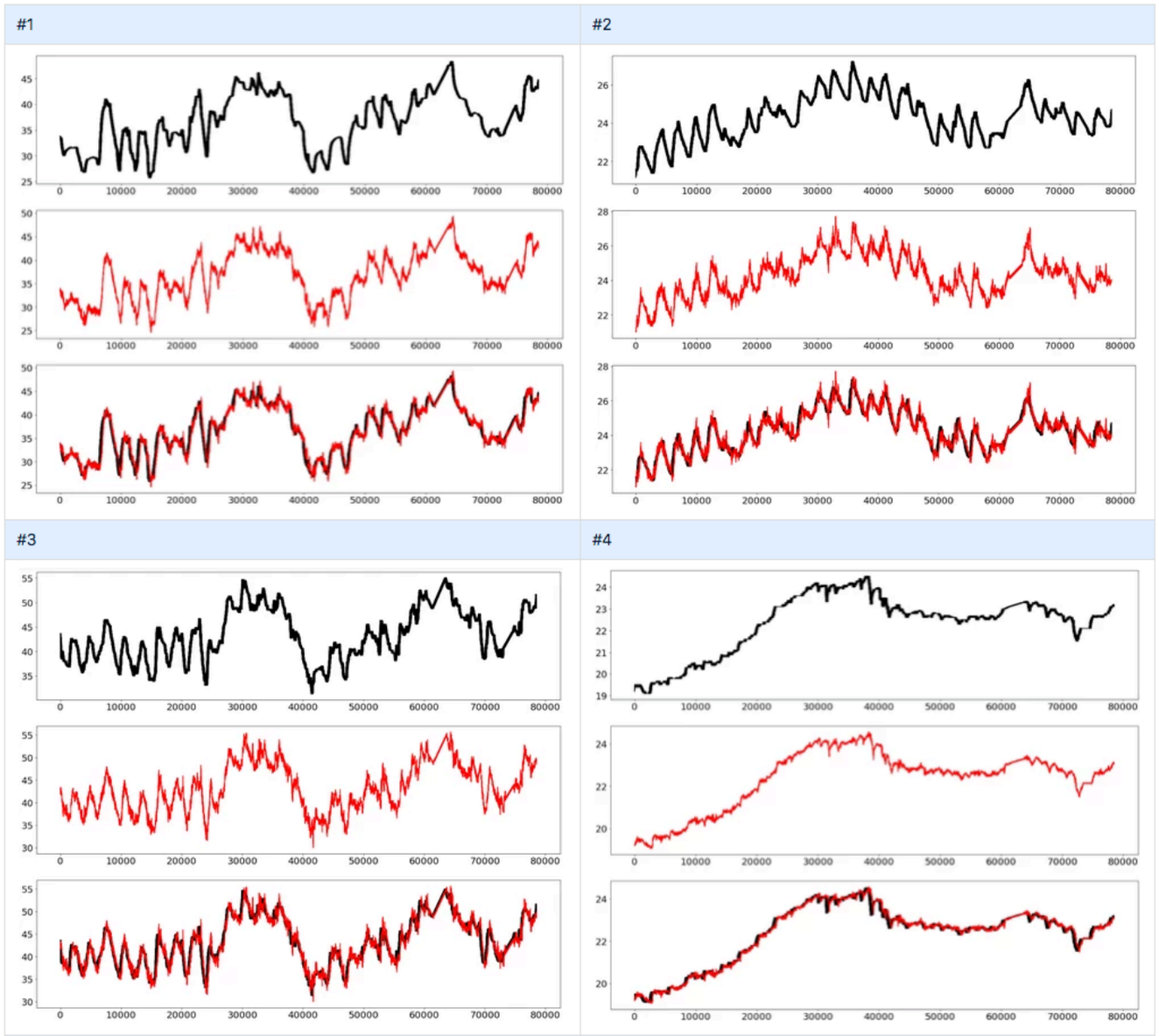
이상 패턴이 포함된 구간을 모두 탐지해 냄.

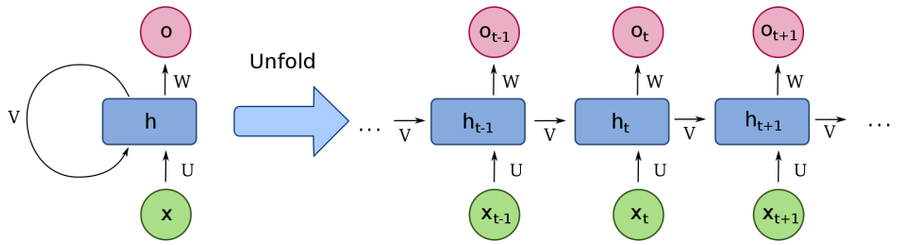
Global Likelihood

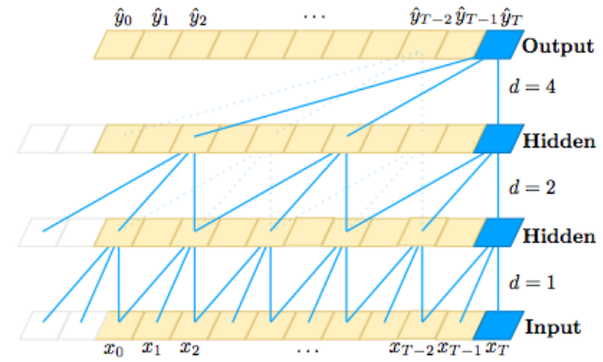
Local Likelihood

핵심 기능 요소

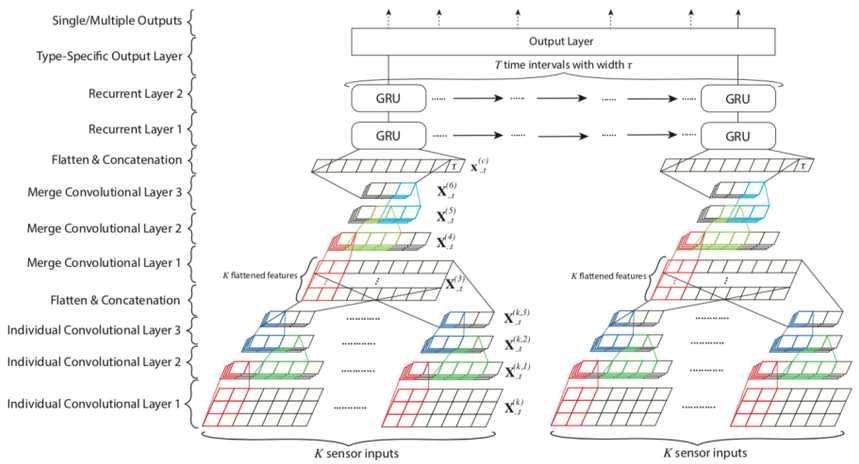
MA - Deep Learning







An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling
https://arxiv.org/abs/1803.01271



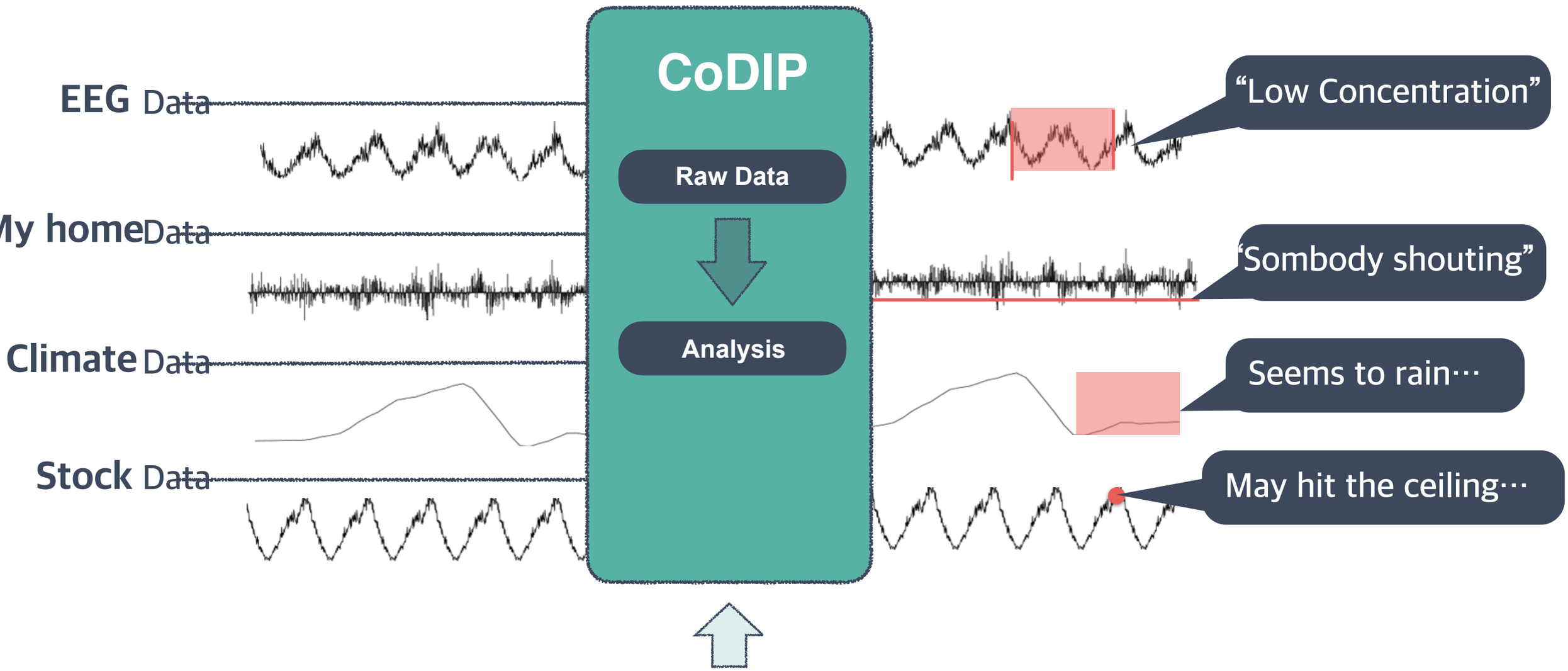
DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing
https://arxiv.org/abs/1611.01942

응용?

- 시스템 동작 패턴에 대한 지능적 탐지 및 응용
  - 기계 장치
  - 사람
  - 정보 시스템
  - 온라인 서비스
- 환경과 상황의 변화를 탐지하고 이해
- 데이터 흐름 패턴 분석 및 관련된 최신 정보의 자동 검색

CoDIP

Dynamic patterns, Analysis, Interpretation, & Collaboration…



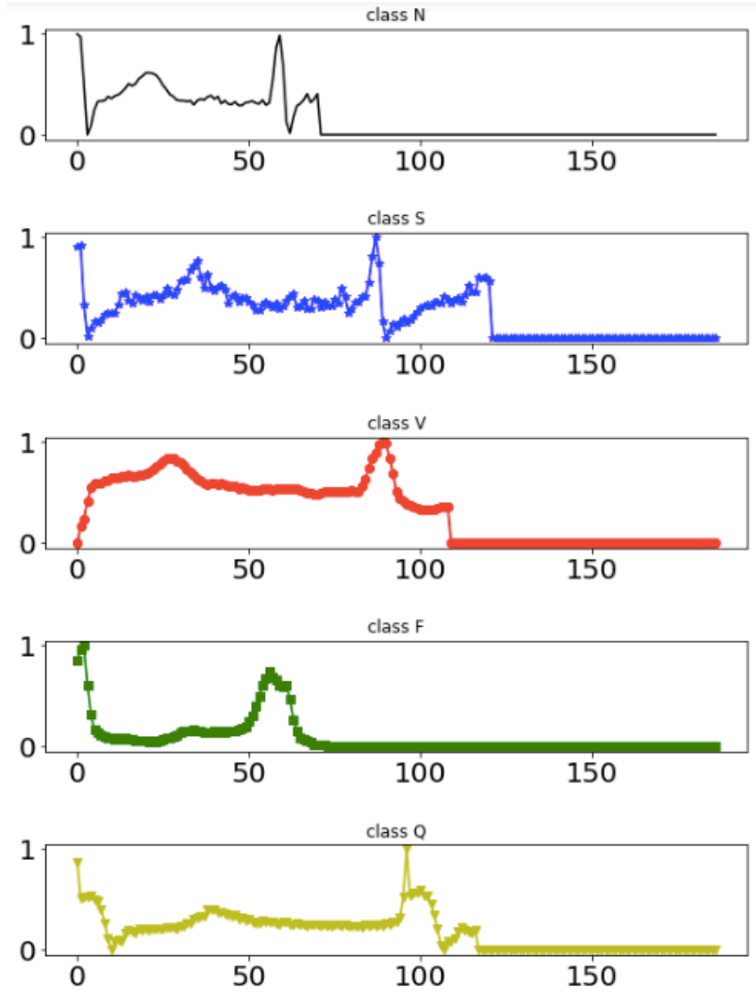
Science
Analysis
Algorithm

Knowledge
Experience

Applying Deep Learning to Anomaly Detection

ECG classification

Category	Annotations
N	• Normal • Left/Right bundle branch block • Atrial escape • Nodal escape
S	• Atrial premature • Aberrant atrial premature • Nodal premature • Supra-ventricular premature
V	• Premature ventricular contraction • Ventricular escape
F	• Fusion of ventricular and normal
Q	• Paced • Fusion of paced and normal • Unclassifiable




```
confusion matrix:
[[737  38   9  12   4]
 [ 45 732  15   3   5]
 [ 14   4 759  18   4]
 [  9   6  16 769   0]
 [  6   2   4   0 788]]
```

stacked LSTM : 94.6%

```
confusion matrix:
[[800   0   0   0   0]
 [131 665   4   0   0]
 [ 21   2 768   8   0]
 [ 32   0  15 753   0]
 [  9   0   0   0 791]]
```

1D-CNN : 94.4%

Classification vs. Anomaly detection

Anomaly in Machine Vibration

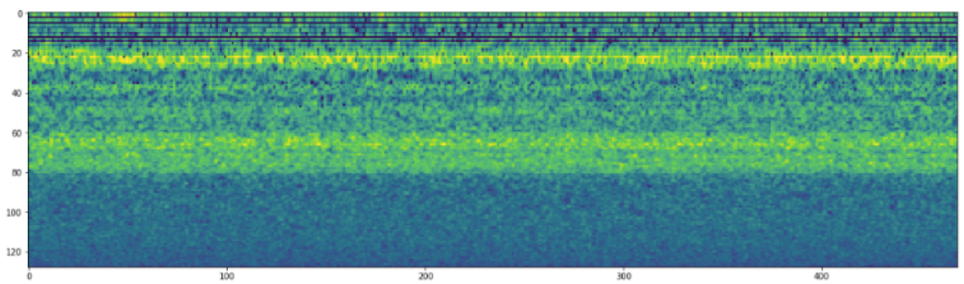


그림 1. Normal 상태의 기어 진동(소음)의 mel-spectrogram

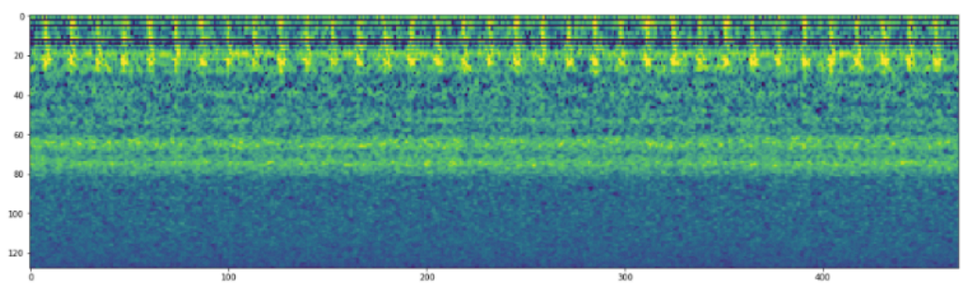
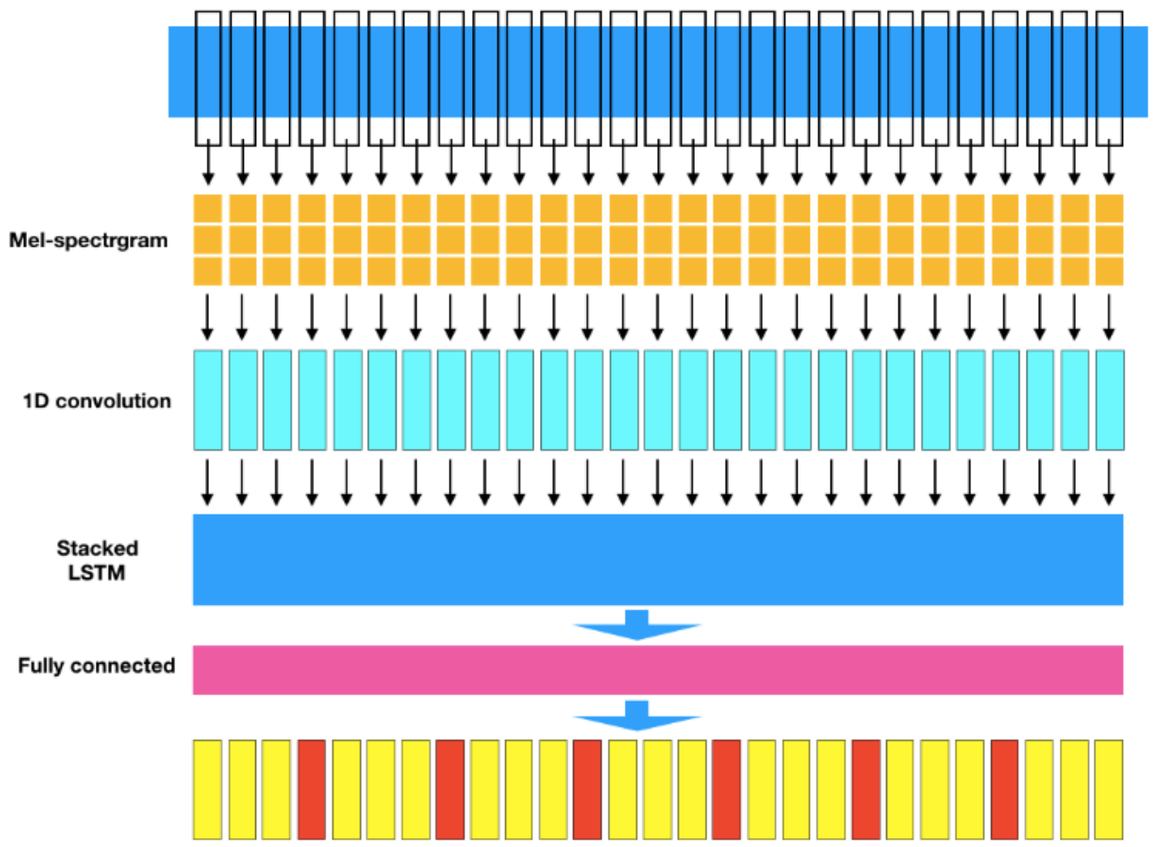
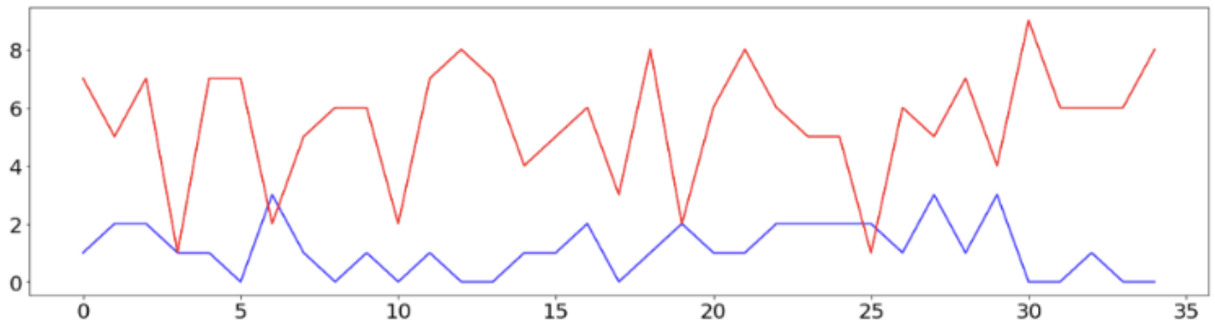


그림 2. Anomaly 상태의 기어 진동(소음)의 mel-spectrogram



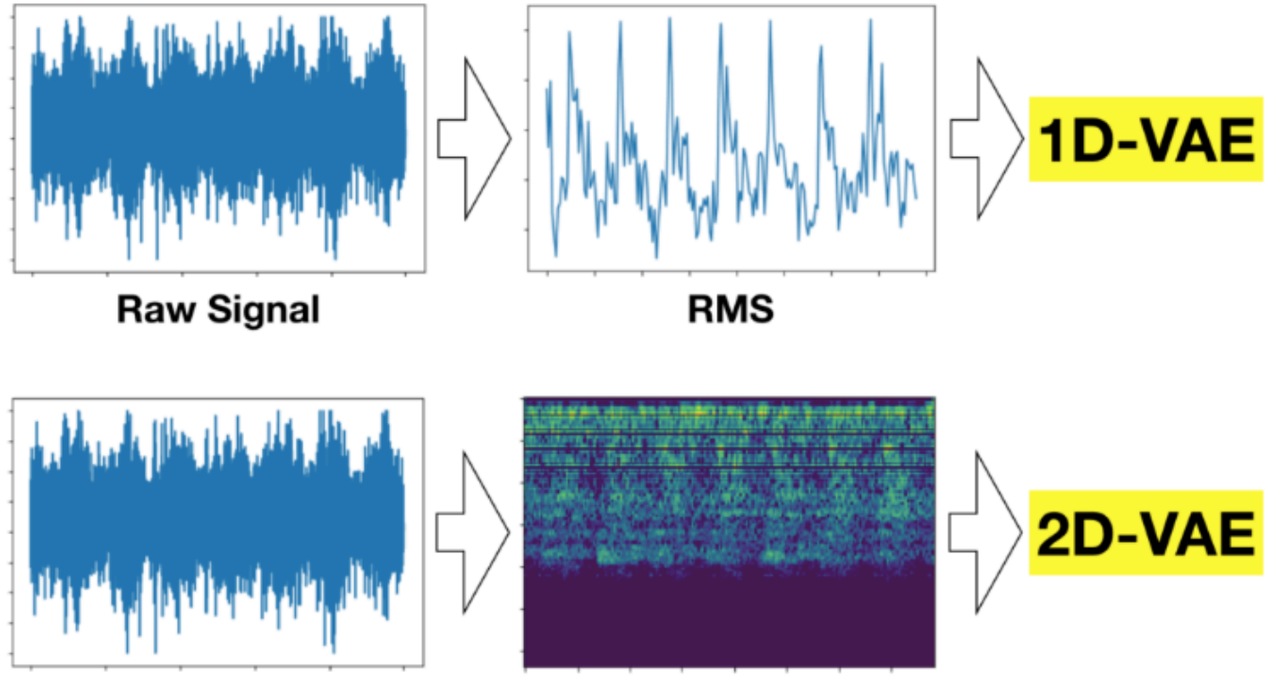
CRNN



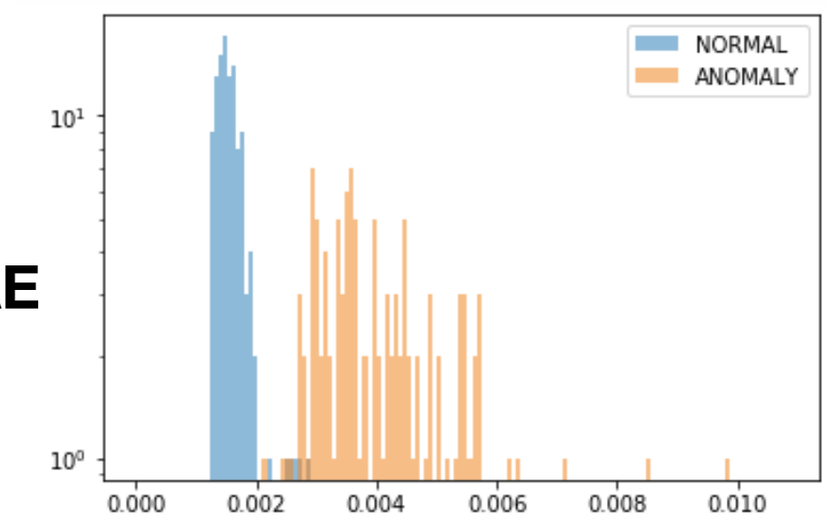
```
confusion matrix:
[[35  0]
[ 6 29]]

roc auc score:
0.9142857142857144
```

Anomaly in Machine Vibration

Nishchal K. Verma, R. K. Sevakula, S. Dixit and A. Salour, Intelligent Condition Based Monitoring using Acoustic Signals for Air Compressors, IEEE Transactions on Reliability, vol. 65, no. 1, pp. 291-309, 2016.

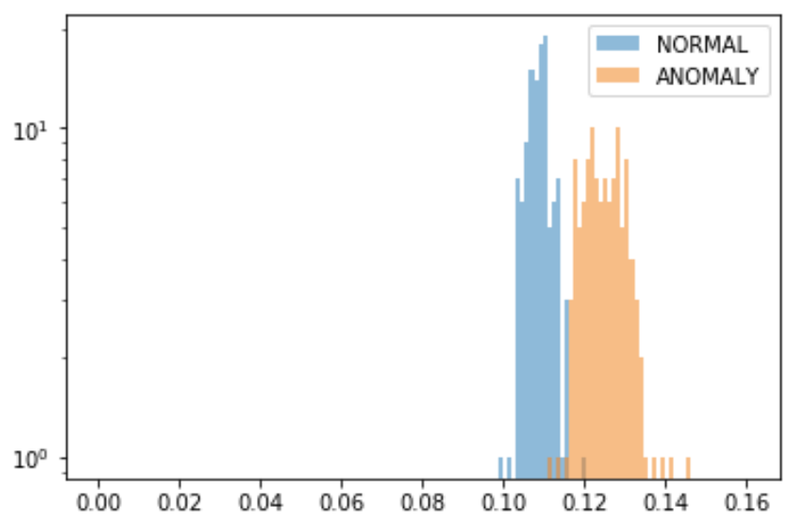


1D-VAE




```
+++++Total Results++++++
Accuracy: 0.9690265486725663
F1 score: 0.9688888888888889
Recall: 0.9646017699115044
Precision: 0.9732142857142857
```

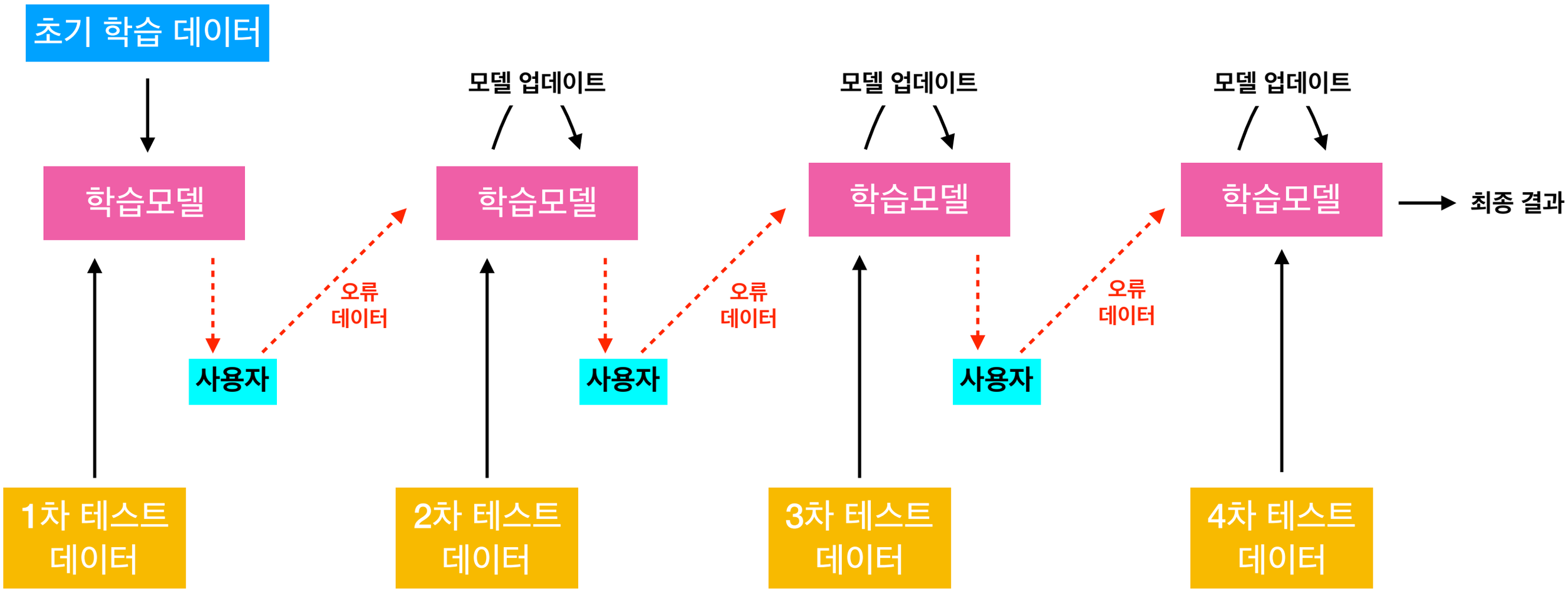
2D-VAE




```
+++++Total Results++++++
Accuracy: 0.9778761061946902
F1 score: 0.9777777777777777
Recall: 0.9734513274336283
Precision: 0.9821428571428571
```

Anomaly in Machine Vibration

Interactive



Anomaly in Machine Vibration

Interactive

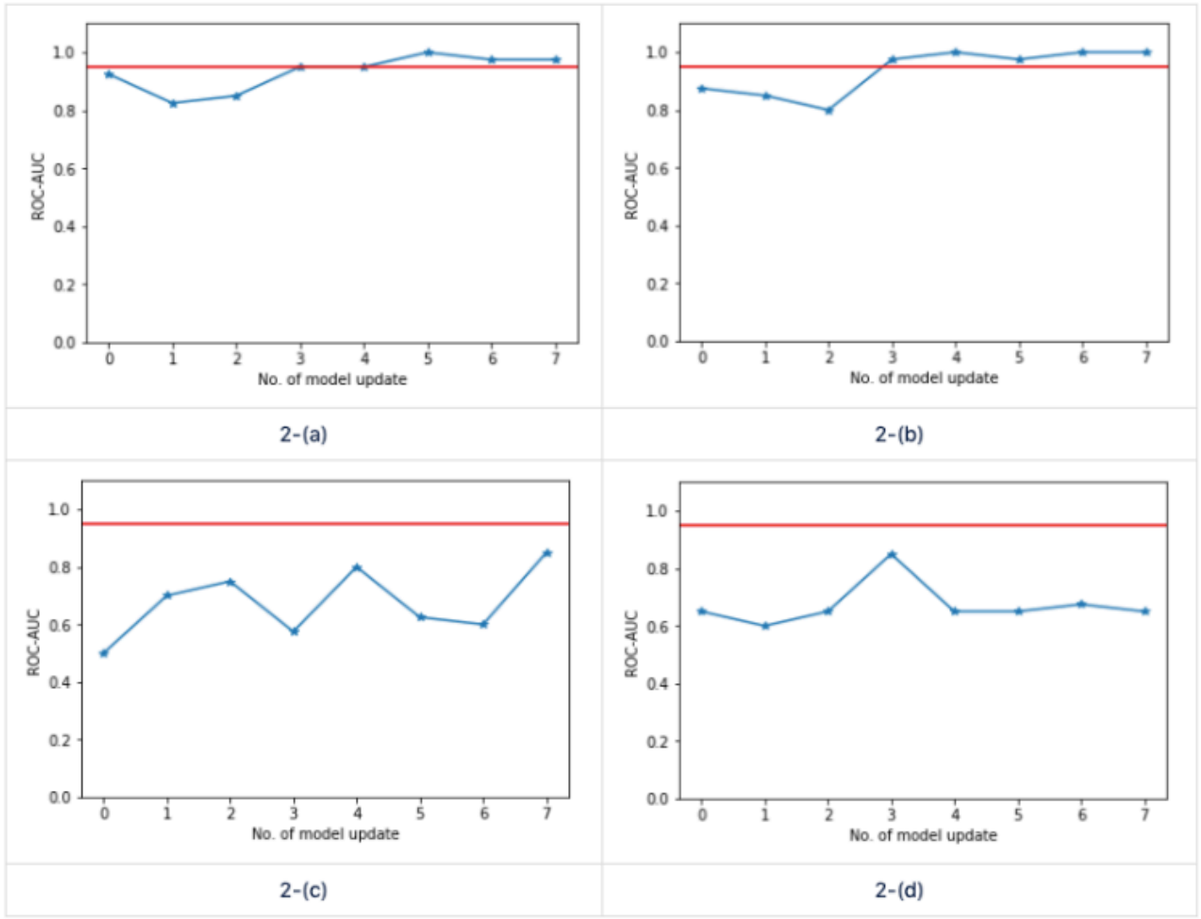


그림 2. 가벼운 업데이트를 적용했을 때의 성공 케이스와 실패 케이스

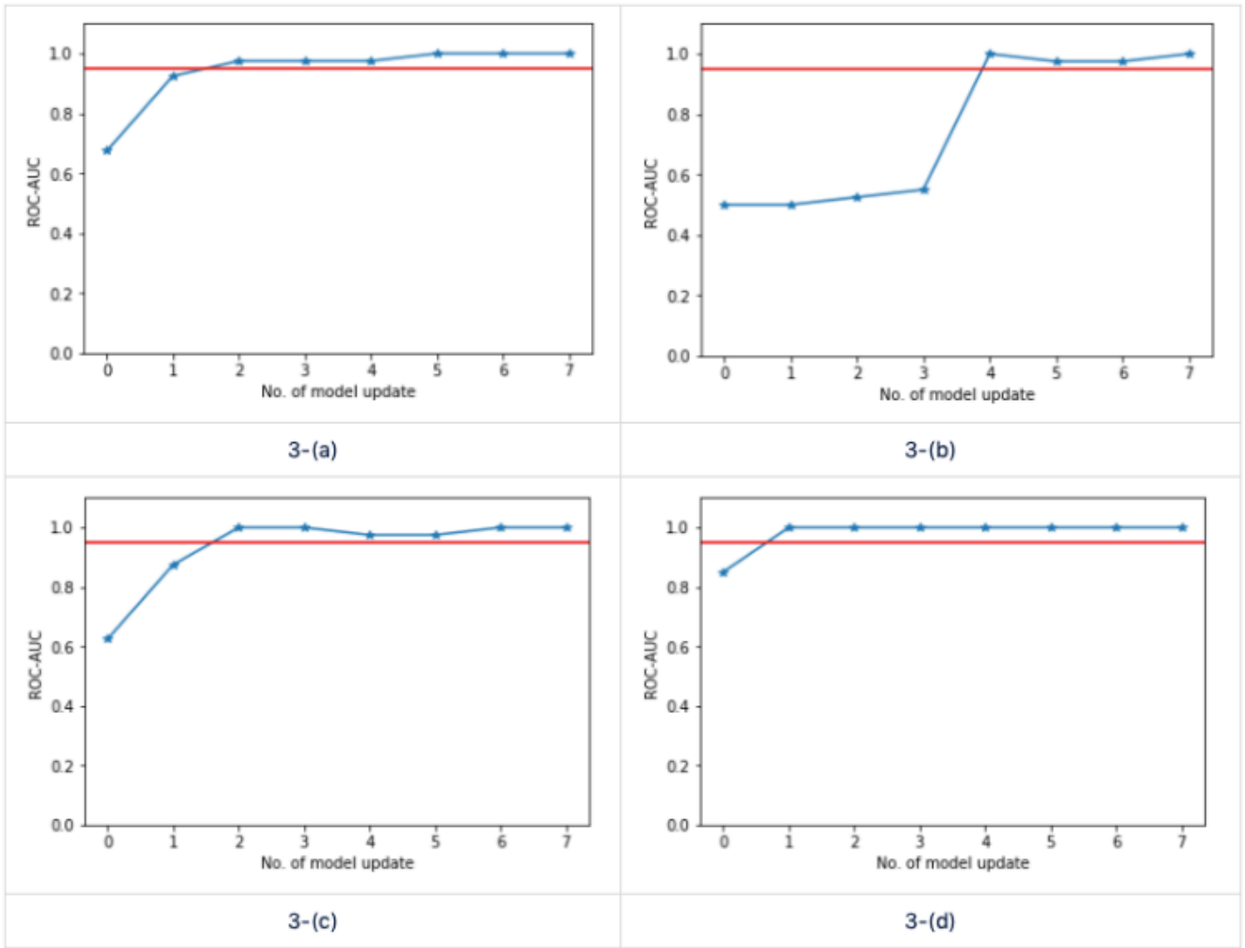
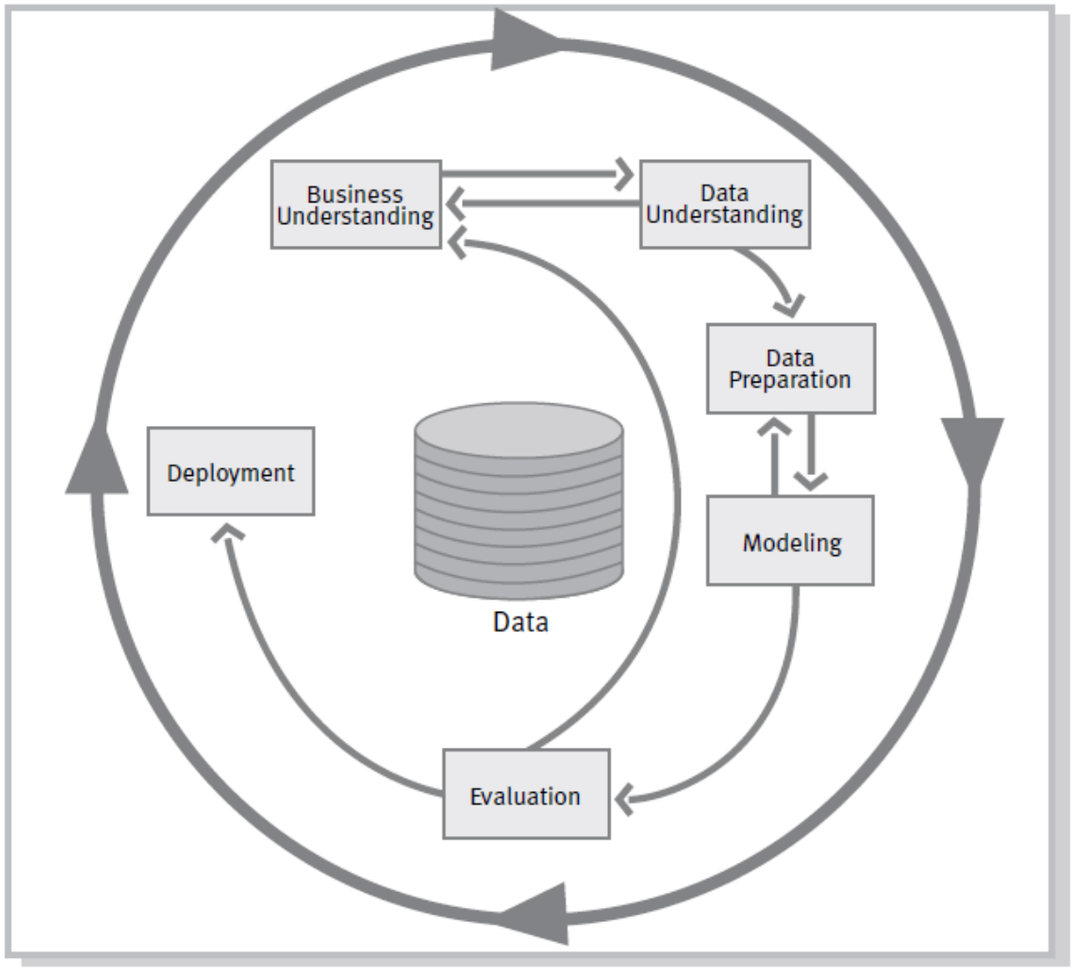


그림 3. 약간 깊은 업데이트를 적용했을 때의 성공 케이스

What about your data & your application?

What about your data & your application?

데이터 분석 프로젝트 vs. 소프트웨어 개발 프로젝트



- 도메인 지식의 중요성
  - 체계화 - 공유
- 데이터 초기 분석의 중요성
  - 데이터 품질 / 용량 / 활용도

데이터 + 도메인 지식 제공자의 역할이 매우 중요!!!

What about your data & your application?

데이터 준비 (수집) → 1차 분석 결과 → 2차 분석 결과 → 실제 응용에 반영

The most important step!!!

Product를 만드는 것이 전부가 아니다.

What about your data & your application?

어떤 사람(비전문가)이 AI 기술을 자신의 업무나 비즈니스에 제대로 활용할 수 있는가?

Data

데이터 제공자 / 도메인 지식 보유자

데이터 분석자 / 학습 모델 개발자

Pattern Detection & Interpretation

Application Development

What about your data & your application?