Change the world with a data revolution

CGM을 활용한 시계열 데이터 분석

from classical to ML-based approach

Global No.1 Health Data Tech Company

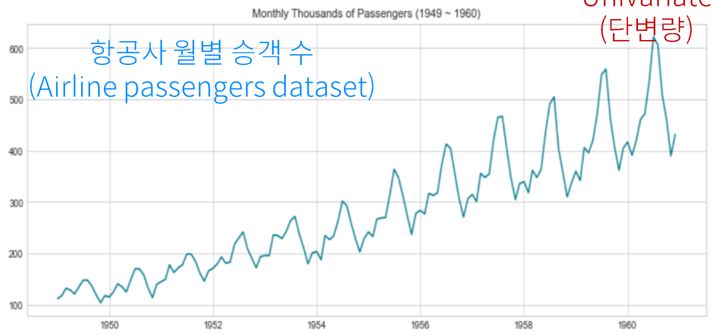
SEP 2022

Background

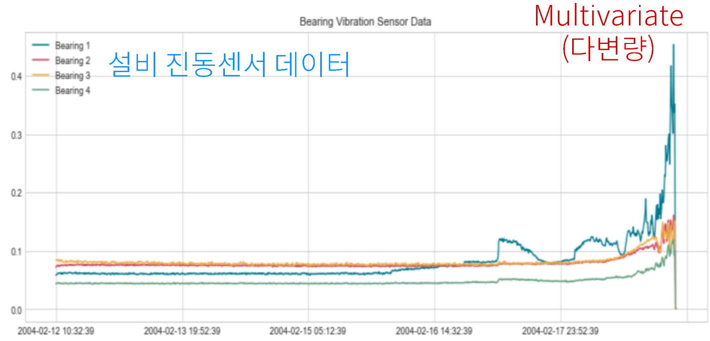
- **시계열(Time series)** : 일정한 간격으로 배치된 데이터의 수열(sequence). 각 샘플은 시간적인 순서를 가짐.
- **사례** : 항공사 월별 승객 수, 연간 제품 판매량, KOSPI 기업 주가, 설비 진동 센서 값, 그리고 CGM 데이터

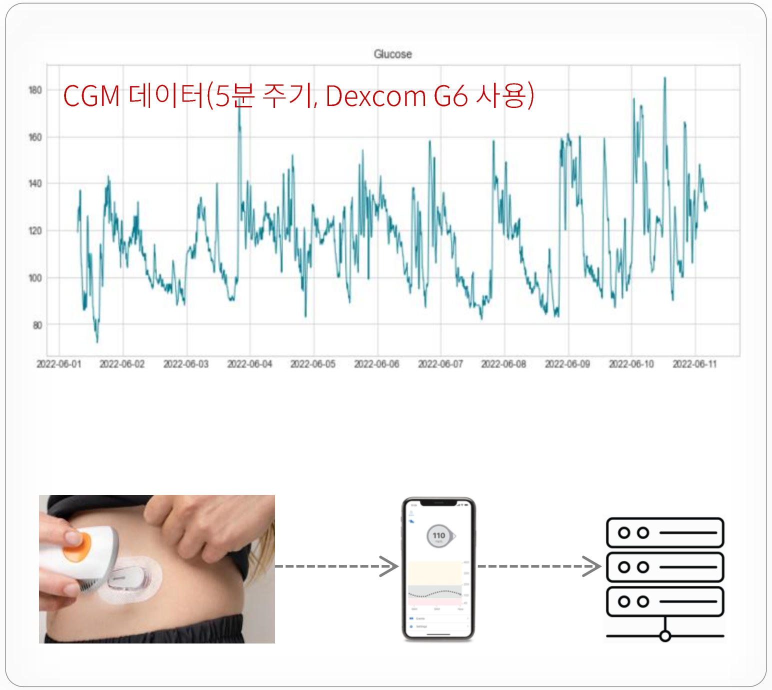
timestamp	glucose
2022-06-01 06:55:00	119.0
2022-06-01 07:00:00	122.0
2022-06-01 07:05:00	125.0
2022-06-01 07:10:00	128.0
2022-06-01 07:15:00	129.0
...	...

Univariate
(단변량)



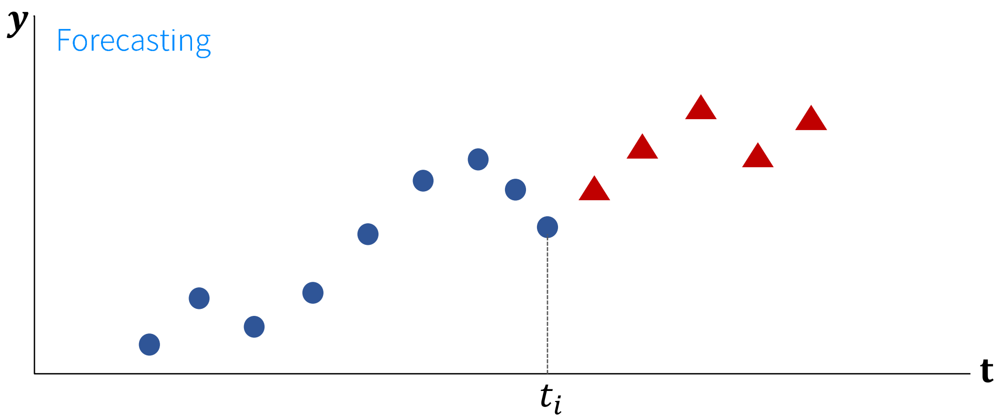
Multivariate
(다변량)

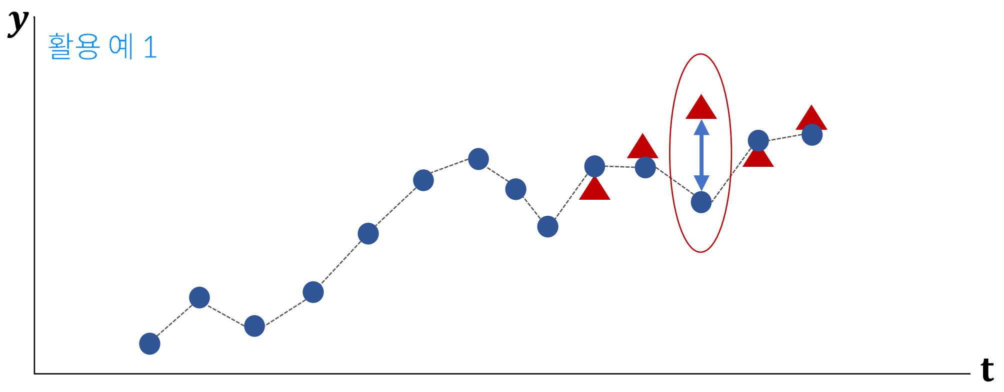


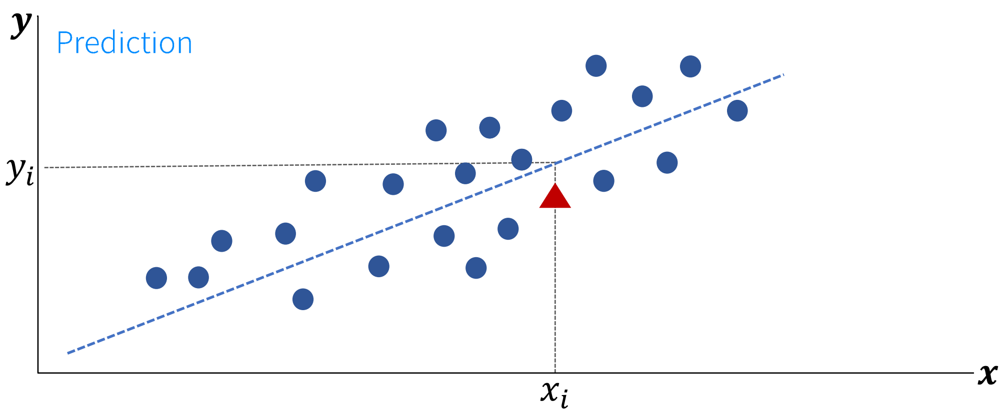


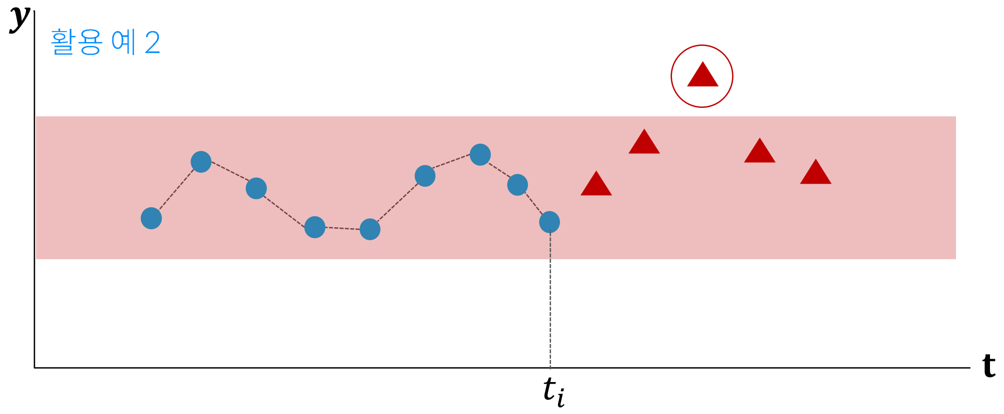
Background

- 활 용
  - forecasting(예측) (≠prediction)
  - detection(탐지), diagnosis(진단)



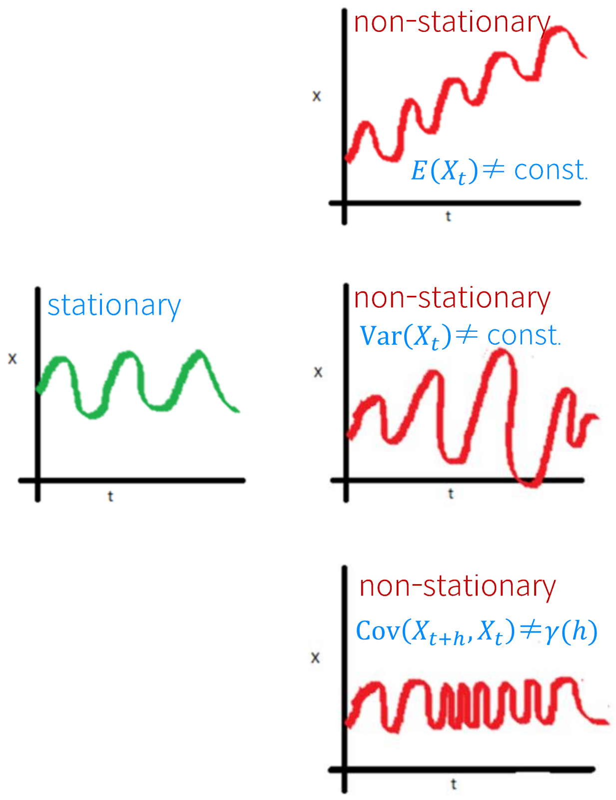


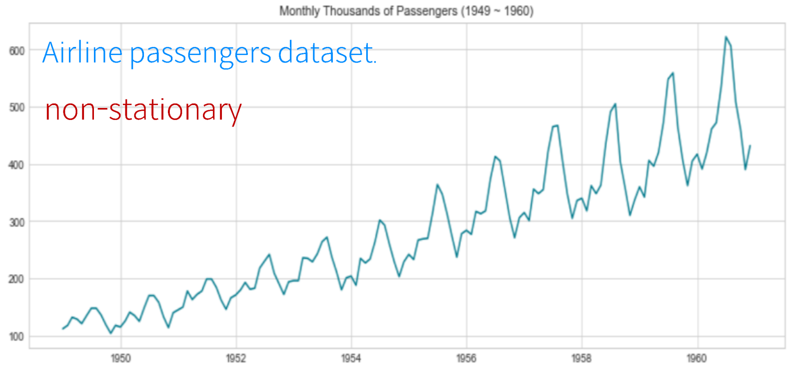


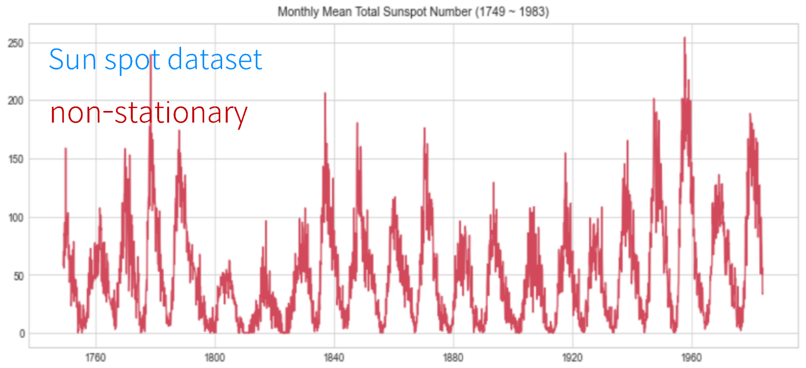


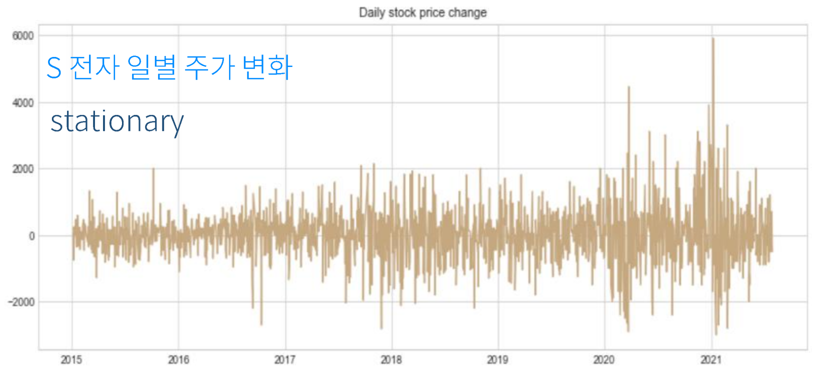
Background

- 정상성(Stationary) vs 비정상성(Non-stationary)
  - stationary : 시계열 데이터가 관측된 시간에 대해서 독립일 때(ex: white noise)
  - Non-stationary : Stationary 판별기준을 만족하지 않을 때(ex: trend, seasonality를 포함하는 시계열)
  - 판별기준 : 평균($E(X_t) = \mu$), 분산($Var(X_t) = \sigma$), 공분산($Cov(X_{t+h}, X_t) = \gamma(h)$)









Modeling Procedure

- 예측모델 개발 절차

데이터 전처리 (Preprocessing)

- 결측 치 처리
- Smoothing
- Resampling

변환 (Transformation)

- Differencing
- Log transformation
- Normalization

학습 준비 (Preparation)

- Train Dataset
- Test Dataset

모델 학습 (Train)

- Classical method
  - ARIMA, SARIMA, SIRMAX
- ML-Based method
  - Single-step
  - Multi-step

평 가 (Evaluation)

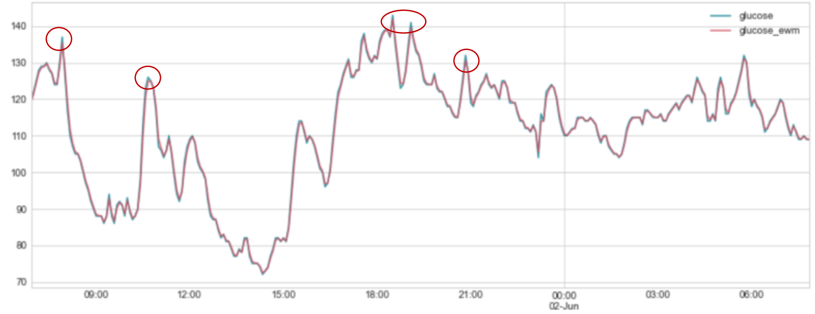
- MAE, RMSE, MAPE

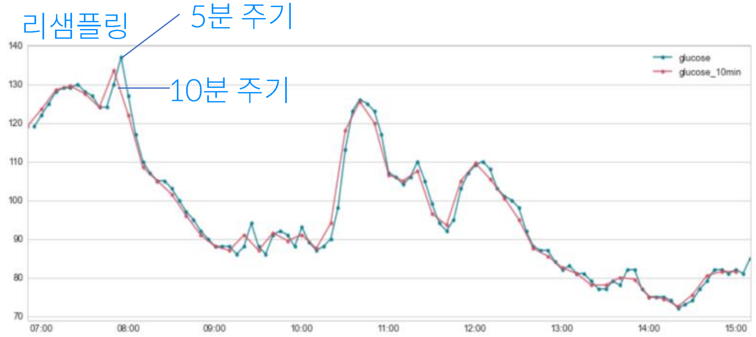
Preprocessing/ Transformation

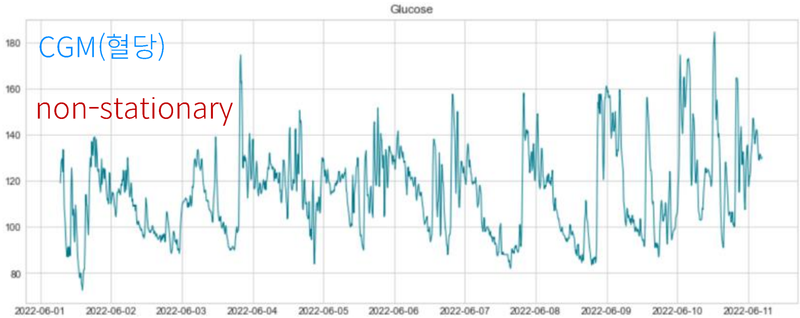
- **결측 치(Missing values) 처리** : Forward/ Backward fill, Linear/ Cubic interpolation
- **스무딩(Smoothing)** : 노이즈 제거
- **리샘플링(Resampling)** : 측정 주기 조정
- **변환(Transformation)** : 차분(Differencing), 로그(Log) 변환, 정규화(Normalization)

timestamp	glucose 원본	glucose_ewm 스무딩	glucose_10min 리샘플링
2022-06-01 06:55:00	119.0	119.000000	NaN
2022-06-01 07:00:00	122.0	121.500000	123.5
2022-06-01 07:05:00	125.0	124.322581	NaN
2022-06-01 07:10:00	128.0	127.269231	128.5
2022-06-01 07:15:00	129.0	128.654289	NaN
...	...	...	...
2022-06-11 04:30:00	129.0	129.246638	129.5
2022-06-11 04:35:00	130.0	129.849328	NaN
2022-06-11 04:40:00	130.0	129.969866	130.0
2022-06-11 04:45:00	130.0	129.993973	NaN
2022-06-11 04:50:00	130.0	129.998795	130.0

스무딩(α=0.9) $S_t = \alpha X_t + (1-\alpha) S_{t-1}$







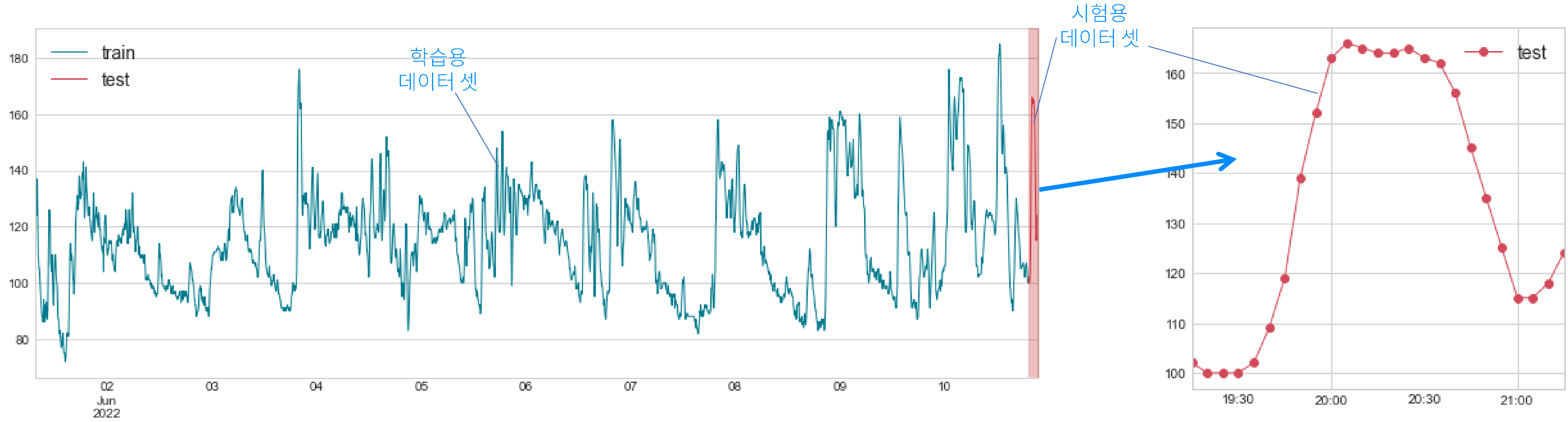
차분
(differencing)

Preparation

- 데이터 셋 준비 : 총 2,760 샘플(2022-06-01 07:20 ~ 06-10 21:15)

- 학습용 2,735 샘플 (2022-06-01 07:20 ~ 06-10 19:10)

- 시험용 25 샘플 (2022- 06-10 19:15 ~ 06-10 21:15)



	timestamp	glucose(t)
0	2022-06-01 07:20:00	129.0
1	2022-06-01 07:25:00	130.0
2	2022-06-01 07:30:00	128.0
3	2022-06-01 07:35:00	127.0
4	2022-06-01 07:40:00	124.0

...

...

	timestamp	glucose(t)
2755	2022-06-10 20:55:00	125.0
2756	2022-06-10 21:00:00	115.0
2757	2022-06-10 21:05:00	115.0
2758	2022-06-10 21:10:00	118.0
2759	2022-06-10 21:15:00	124.0

Train : classical method

ARIMA(p, d=1, q)

$$\Delta\hat{y}_t = c + a_1\Delta y_{t-1} + a_2\Delta y_{t-2} + \cdots + a_p\Delta y_{t-p} + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

- ARIMA (Auto-Regressive Integrated Moving Average)
  - step 1 : 최적의 p, d, q 조합 찾기
  - step 2 : 해당 p, d, q로 모델 학습

학습을 통해 찾아야 할 파라미터 들

$a_1, a_2, \ldots, a_p$

$\theta_1, \theta_2, \cdots, \theta_q$

Step 1

p, d, q 탐색 범위 설정

(0, 0, 0), (0, 0, 1), (0, 0, 2), (0, 0, 3), (0, 0, 4), (0, 1, 0), (0, 1, 1), (0, 1, 2), (0, 1, 3), ···

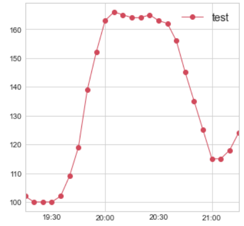
with p, d, q ARIMA Model

평가 $AIC = 2k - 2\ln(\hat{L})$

	(p,d, q)	AIC
0	(4, 1, 3)	12750.441986
1	(4, 1, 1)	12759.075456
2	(2, 1, 4)	12761.322766
3	(1, 1, 4)	12770.993652
4	(2, 1, 1)	12771.278911

Step 2

학습 데이터



시험 데이터

with 4, 1, 3 ARIMA Model

평가 MAE, RMSE, MAPE, ···

ARIMA(p=4, d=1, q=3)

$$\Delta\hat{y}_t = 0.0028 + 1.409\Delta y_{t-1} + 0.401\Delta y_{t-2} - 1.3589\Delta y_{t-3} + 0.5397\Delta y_{t-4} - 0.7506\varepsilon_{t-1} - 0.9779\varepsilon_{t-2} + 0.7286\varepsilon_{t-3}$$

Train : ML-based method

- Single-step forecasting
  - 사용모델 : Xgboost
  - 주요 파라미터 : n_estimators 1000, learning_rate 0.05

	timestamp	glucose(t)
0	2022-06-01 07:20:00	129.0
1	2022-06-01 07:25:00	130.0
2	2022-06-01 07:30:00	128.0
3	2022-06-01 07:35:00	127.0
4	2022-06-01 07:40:00	124.0

…

<원본 데이터>

지도학습 형태로 변환

입력/ 피처(X) : 현재, 과거 값

타겟(y)

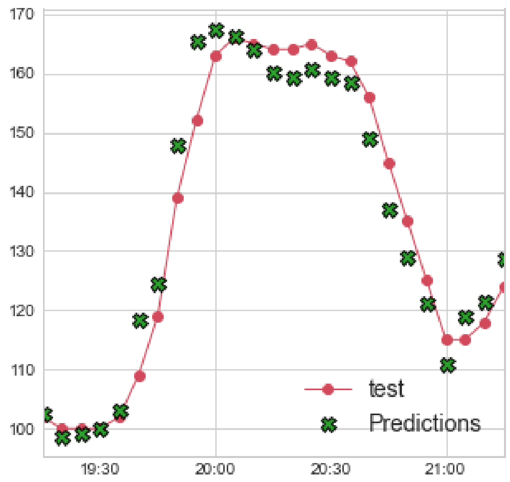
timestamp	glucose(t-4)	glucose(t-3)	glucose(t-2)	glucose(t-1)	glucose(t)	glucose(t+1)
2022-06-01 07:20:00	122.0	125.0	128.0	129.0	129.0	130.0
2022-06-01 07:25:00	125.0	128.0	129.0	129.0	130.0	128.0
2022-06-01 07:30:00	128.0	129.0	129.0	130.0	128.0	127.0
2022-06-01 07:35:00	129.0	129.0	130.0	128.0	127.0	124.0
2022-06-01 07:40:00	129.0	130.0	128.0	127.0	124.0	124.0
...	...	...	...	...	...	...

X

y

Machine learning Model

시험 결과

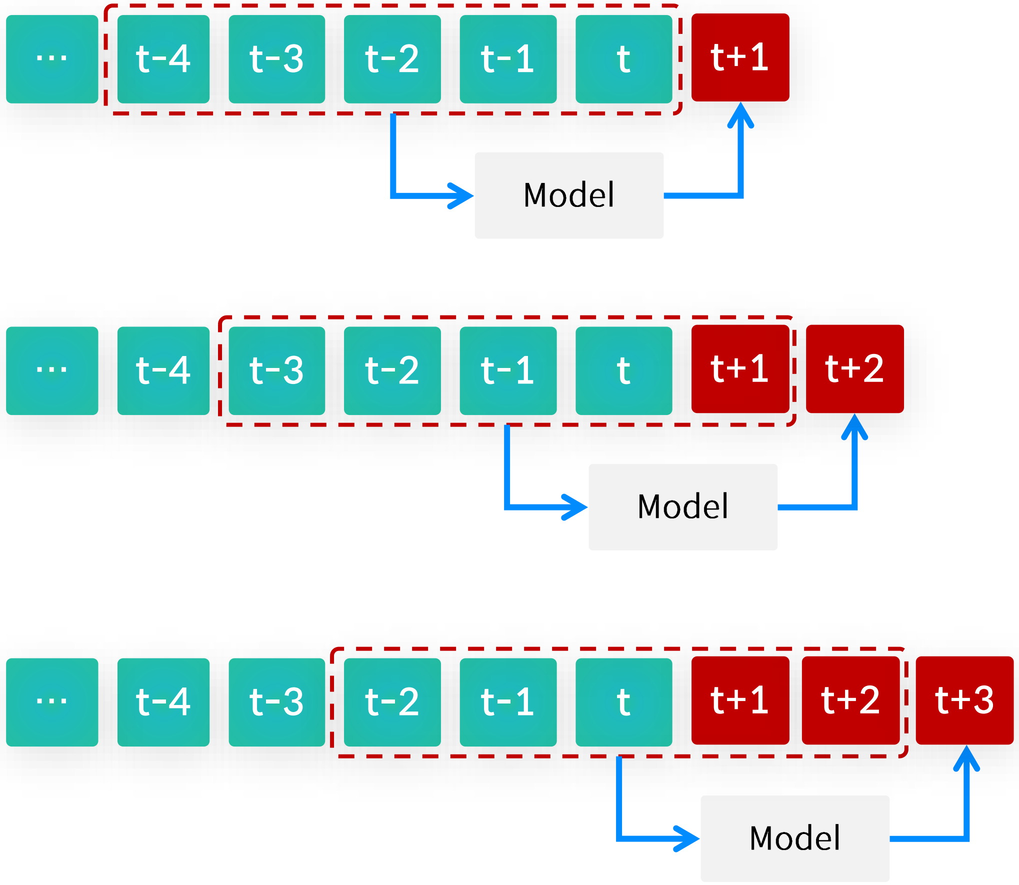


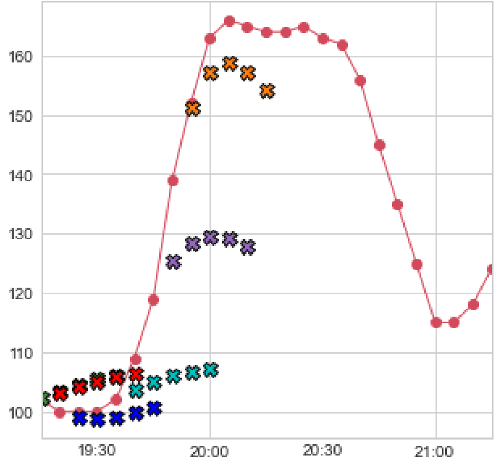
<시험결과>

Train : ML-based method

- Multi-step forecasting

  - recursive method : 예측 결과를 다음 스텝의 입력으로 사용



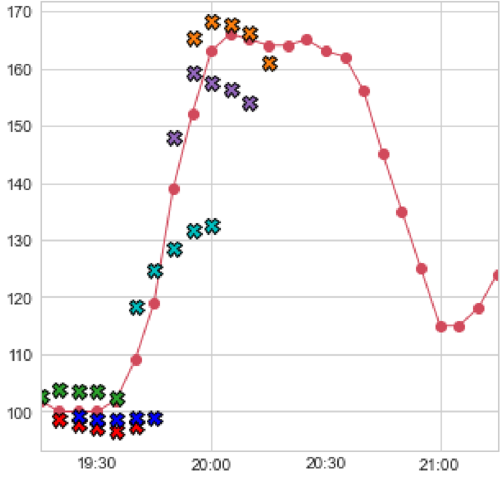


<(S)ARIMA>

샘플 수 10,000개 미만,

월/분기/연간 계절성이 있을 때 잘 동작

<시험결과 : (S)ARIMA>



<Machine learning>

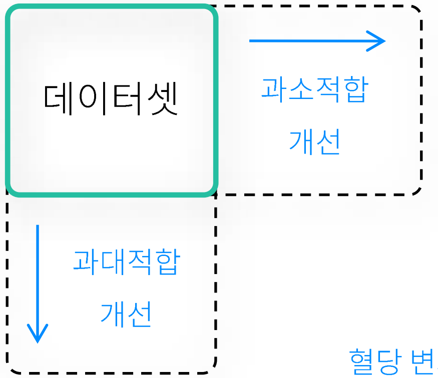
- ARIMA 잔차가 white noise가 아닐 때
- 알지 못하는 계절성이 존재할 때
- 패턴이 비선형 특성을 보이는 경우
- 계절성 패턴이 두개 이상일 때
  (예: 온도변화)

<시험결과 : Xgboost>

Train : ML-based method

- Multi-step forecasting

- recursive method : 피처 추가



혈당 변화량 피처 추가
$(y_t - y_{t-1})$

입력/ 피처(X) : 현재, 과거 값

타겟(y)

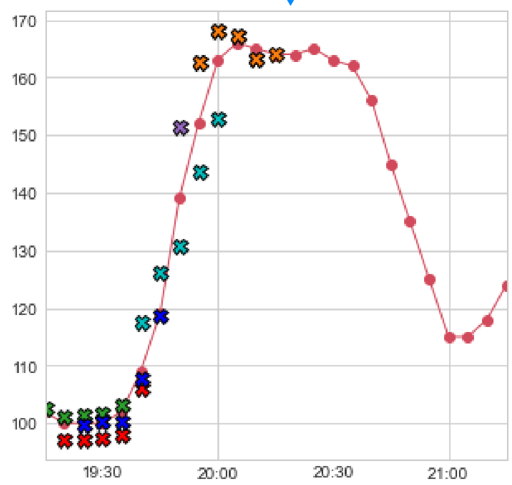
X

y

Machine learning
Model

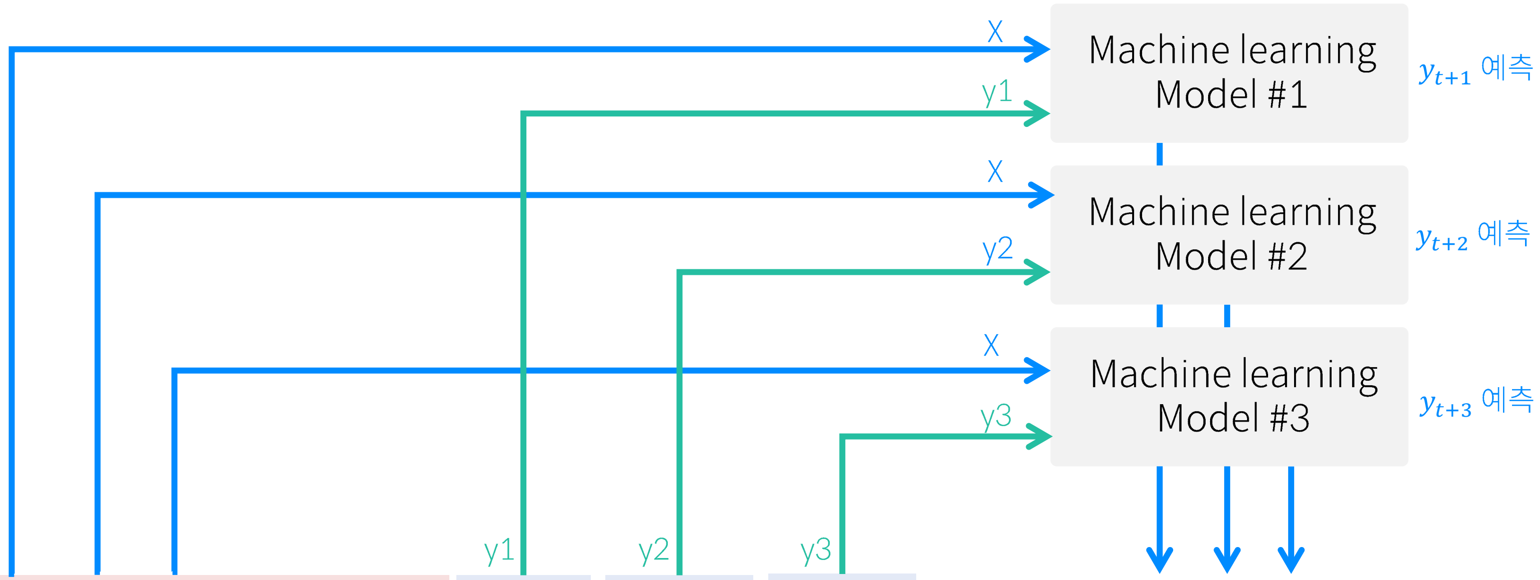
시험 결과

timestamp	diff(t-4)	diff(t-3)	diff(t-2)	diff(t-1)	diff(t)	glucose(t-4)	glucose(t-3)	glucose(t-2)	glucose(t-1)	glucose(t)	glucose(t+1)
2022-06-01 07:20:00	3.0	3.0	3.0	1.0	0.0	122.0	125.0	128.0	129.0	129.0	130.0
2022-06-01 07:25:00	3.0	3.0	1.0	0.0	1.0	125.0	128.0	129.0	129.0	130.0	128.0
2022-06-01 07:30:00	3.0	1.0	0.0	1.0	-2.0	128.0	129.0	129.0	130.0	128.0	127.0
2022-06-01 07:35:00	1.0	0.0	1.0	-2.0	-1.0	129.0	129.0	130.0	128.0	127.0	124.0
2022-06-01 07:40:00	0.0	1.0	-2.0	-1.0	-3.0	129.0	130.0	128.0	127.0	124.0	124.0
...	...	...	...	...	...	...	...	...	...	...	...

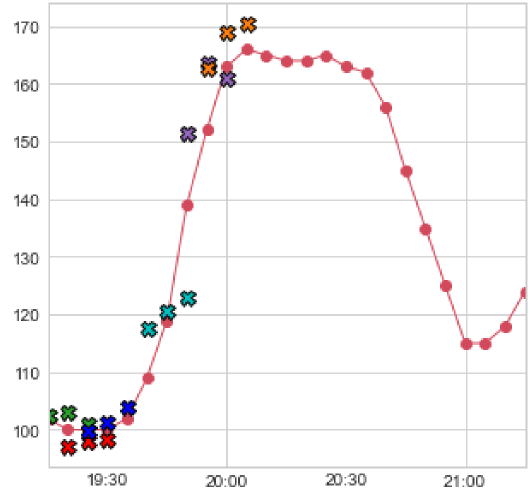


Train : ML-based method

- Multi-step forecasting
  - **direct method**(multi-output repressor)



timestamp	diff(t-4)	diff(t-3)	diff(t-2)	diff(t-1)	diff(t)	glucose(t-4)	glucose(t-3)	glucose(t-2)	glucose(t-1)	glucose(t)	glucose(t+1)	glucose(t+2)	glucose(t+3)
2022-06-01 07:20:00	3.0	3.0	3.0	1.0	0.0	122.0	125.0	128.0	129.0	129.0	130.0	128.0	127.0
2022-06-01 07:25:00	3.0	3.0	1.0	0.0	1.0	125.0	128.0	129.0	129.0	130.0	128.0	127.0	124.0
2022-06-01 07:30:00	3.0	1.0	0.0	1.0	-2.0	128.0	129.0	129.0	130.0	128.0	127.0	124.0	124.0
2022-06-01 07:35:00	1.0	0.0	1.0	-2.0	-1.0	129.0	129.0	130.0	128.0	127.0	124.0	124.0	130.0
2022-06-01 07:40:00	0.0	1.0	-2.0	-1.0	-3.0	129.0	130.0	128.0	127.0	124.0	124.0	130.0	137.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

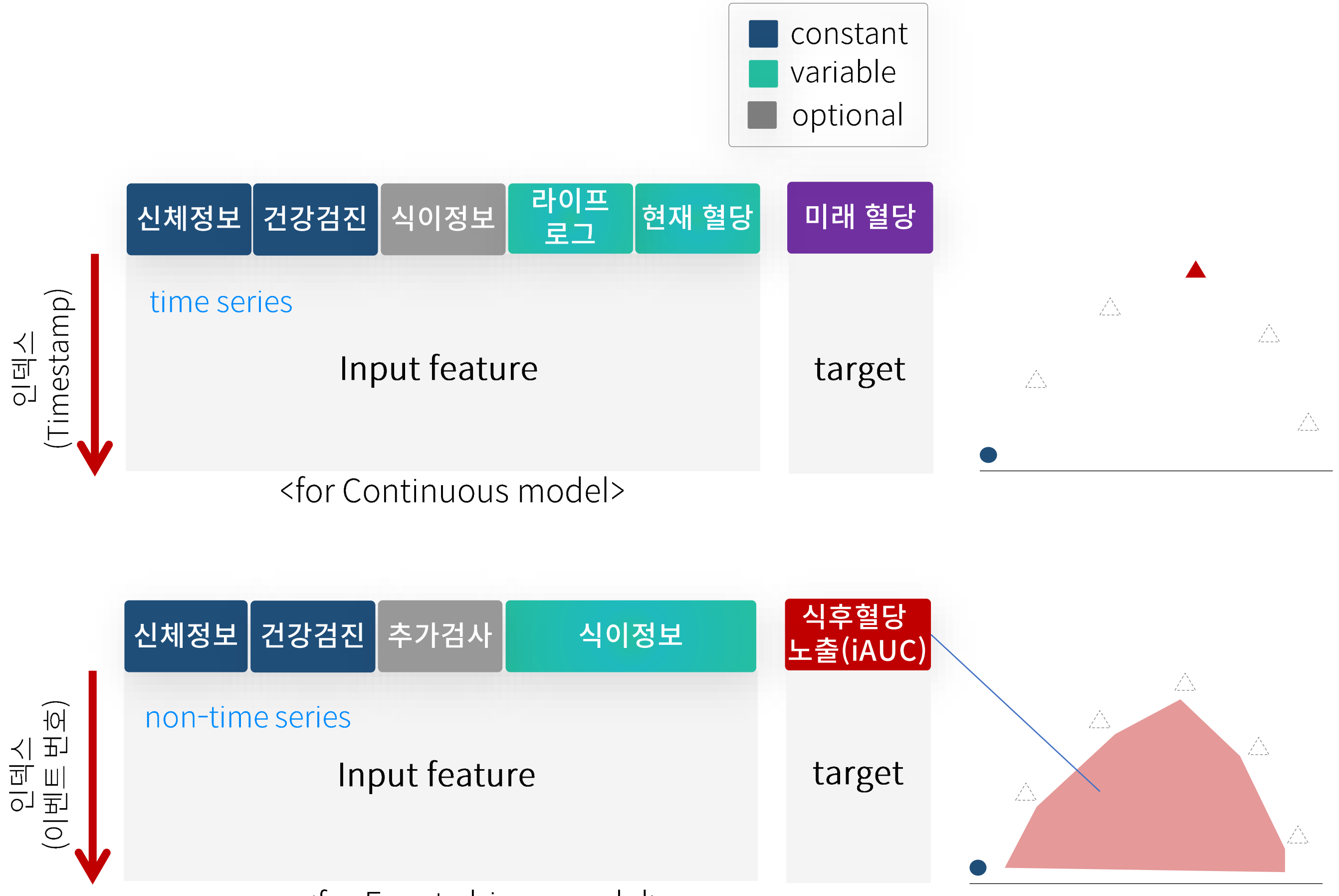


Conclusion

- **향후계획**

- 임상 데이터 셋 구축(multivariate)

<데이터 수집 환경>

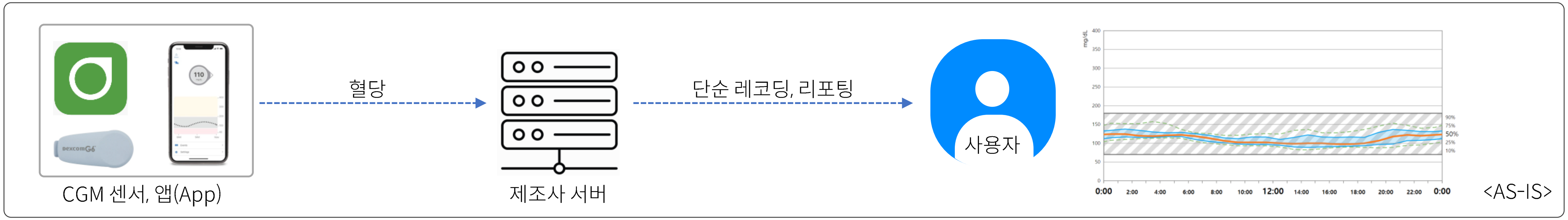


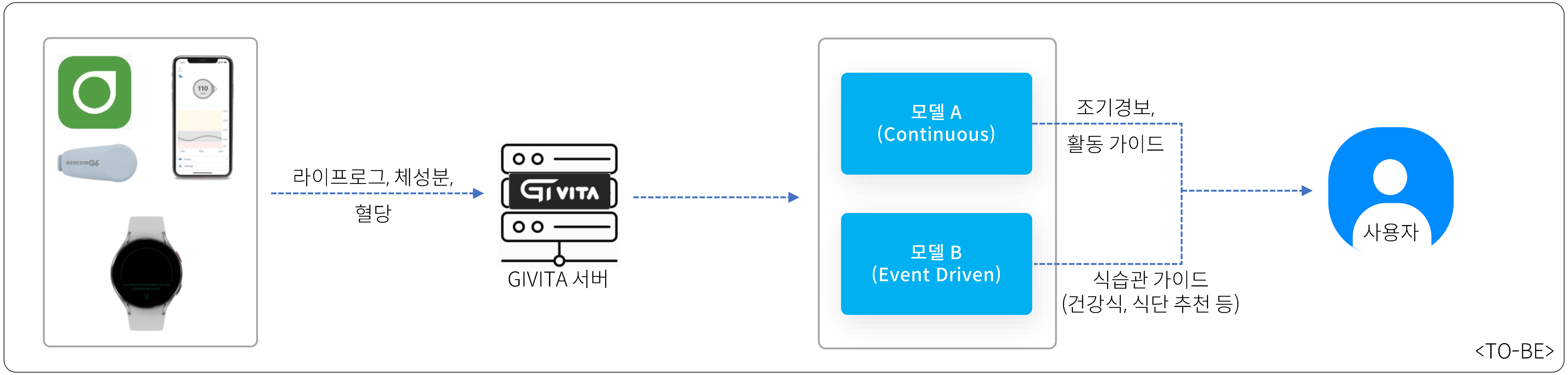
Conclusion

- **향후계획**

- 모델 개발, 개념 검증

- BM 구체화





Change the world with a data revolution

THANK YOU

Global No.1 Health Data Tech Company
SEP 2022