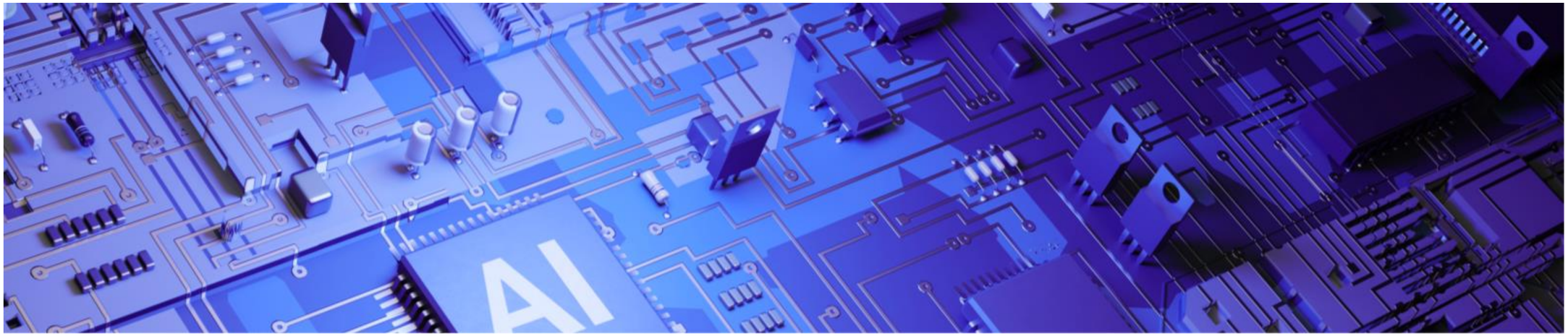


Technical Trends of Time-series Data Imputation & MICE

김홍범

202202144

01

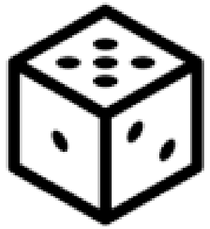
Missing data

01 결측치 데이터 종류

02 결측치 데이터 보간 방법

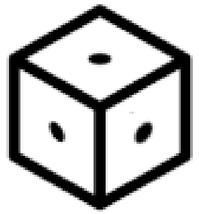
Missing Data

결측치 데이터 종류



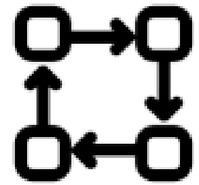
Missing Completely at Random

(MCAR)



Missing at Random

(MAR)



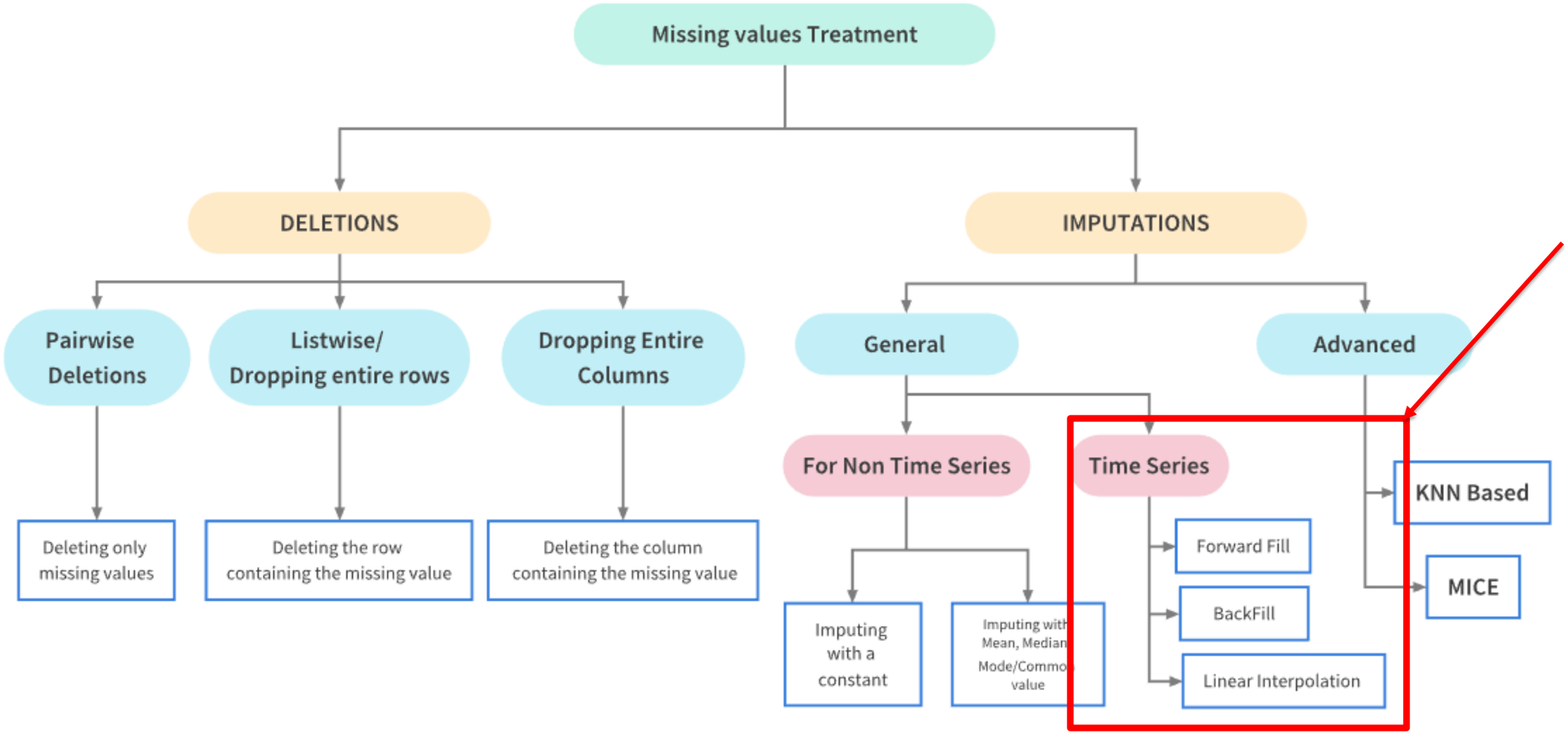
Missing Not at Random

(MNAR)

- 완전 무작위 결측(MCAR) – 전체에 걸쳐 무작위하게 누락된 경우, 변수의 종류, 변수의 값과 상관없이 비슷한 분포로 누락된 데이터를 의미한다.
- 무작위 결측(MAR) – 어떤 특정 변수에 대하여 데이터가 누락되는 경우를 의미, 결측값의 경우가 자료 내의 다른 변수와 관련이 있다.
  - Ex) 30대 남성의 용돈 설문시 결측값이 자주 발생, 30대 남성과 용돈 설문 결측에는 관련이 있다 -> 단 얻고자 하는 결과(소득 수준)과 용돈 설문과는 상관관계가 없는 경우를 의미함
- 비무작위결측(MNAR) – 누락되는 부분들이 무작위로 누락되는 것이 아닌 누락된 변수의 값이 누락된 이유가 있는 경우, 대부분의 결측치에 해당
  - Ex) 의료 데이터의 경우 환자 개인 사정 혹은 어떤 질병이냐에 따라서 누락되는 경우가 존재함
- MCAR, MAR의 경우 결측값을 제거하고 진행하는 것이 좋음, 하지만 MNAR같은 경우 결측치 데이터 제거시 모델이 편향적 학습될 위험이 있음, 따라서 적절한 결측치 보간 및 처리 방법이 매우 중요

Missing Data

결측치 데이터 보간 방법



해당 부분에서
최근 연구가
이루어지고 있음

02

Time series imputation model

01 통계적 기법

02 행렬 기반 기법

03 RNN 기반 기법

04 GAN 기반 기법

Time-series imputation model

통계적 기법, 행렬 기반 기법

- **ImputeTS(단순대입법)**
  표준 오차가 과소 추정되는 단점이 존재 -> 간단하게 평균 대치법으로 생각하면 편함(Kalman smoothing을 이용)

- **Hot-Deck Imputation**
  상관성이 존재하거나 유사성이 존재하는 변수 집합으로부터 랜덤하게 데이터를 뽑아 결측값을 대체하는 방법
  가장 많이 나온 수로 대치 혹은 변하지 않았을 거라 가정하고 대체

- **MICE(다중 대치법) - 2011**
  각 복원 모델에 따라 대치를 진행하는 실용적인 방법, 간단하게 시뮬레이션된 여러 결측치 데이터 셋을 만든 후에 해당
  데이터셋의 대치 평균을 이용하여 결측값을 대체하는 방법

행렬 기반 기법

- **행렬분해(Matrix Factorization) - 2009**
  행렬을 분해 및 재구성함으로써 데이터 간 상관관계를 도출하여 결측치를 대치하는 방법
  1. 데이터의 행렬을 2개의 저차원 행렬로 분해한다.
  2. 원래 행렬로 재구성하는 시도를 거치면서 누락된 값을 대치하는 방식으로 진행

- **TRMF(Temporal Regularization Matrix Factorization) – 2016**
  결측값이 있는 고차원 시계열 데이터에 매우 적합하고 확장 가능한 행렬분해를 사용
  시간적 종속성 뿐만 아니라 데이터 기반 종속성 특징도 학습하여 우수한 결과를 도출

- **PSMF(Probabilstic Sequential Matrix Factorization) – 2019**
  고차원 시계열로 구성된 시변량 및 비정상성 데이터 세트를 분해하는 방법
  Markovian 종속성이 있어 그 구조를 이용하여 시간 종속성에 대한 속성을 저차원 특징 공간으로 인코딩이 가능
  또한, 칼만 필터 기법을 통하여 효율적으로 모델을 구성함으로써 미분 가능한 비선형 부분적 공간 모델을 보정하고 추정하는 결측치 처리 방법

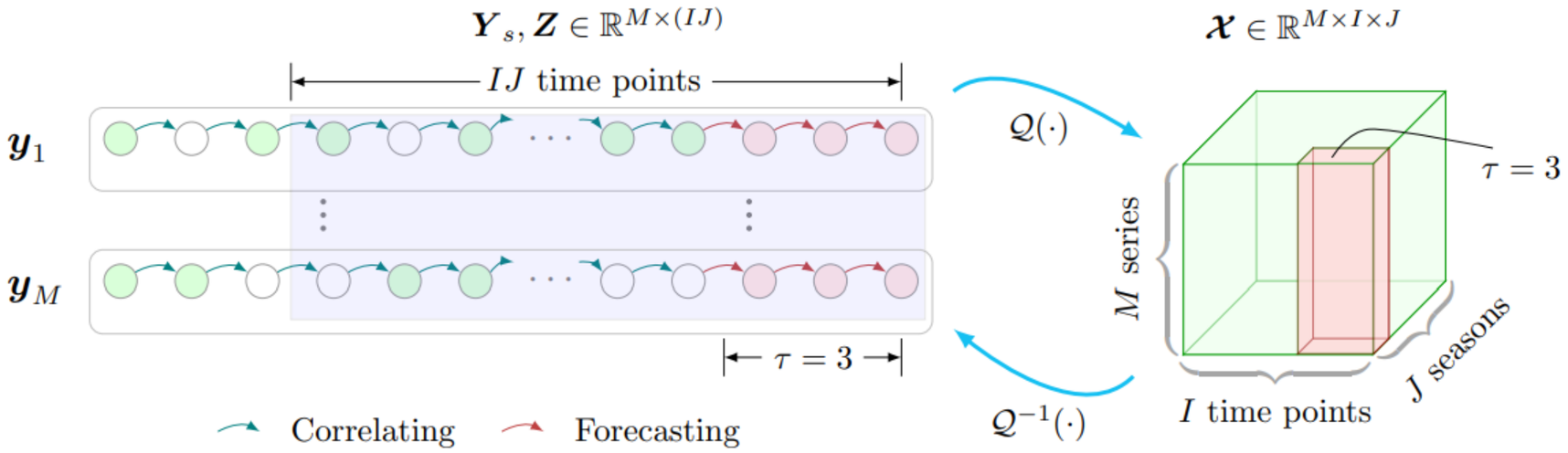
Time-series imputation model

회귀분석 기법

- 과거의 데이터를 통하여 모델을 학습하여 예측하는 방법
- 빠르고 간단하게 처리가 가능하나, 전체적인 시계열 특성을 반영하지 못함
- 대표적인 모델 : 자기회귀모형(AR), ARIMA등이 있음

LATC(Low-Rank-AutoRegressive Tensor Completion) – 2020

- ✓ AR 모델을 발전시킨 모델로서 다변량 시계열 데이터를 3차원의 텐서로 변환시킴
- ✓ 시간, 계절성, 다변량 변수 3가지의 기준을 고려하여 진행함

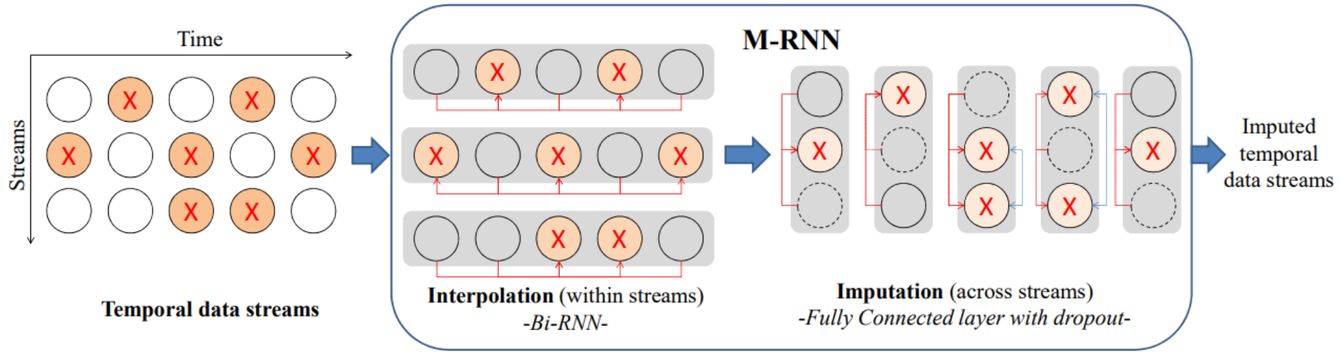


LATC Illustration

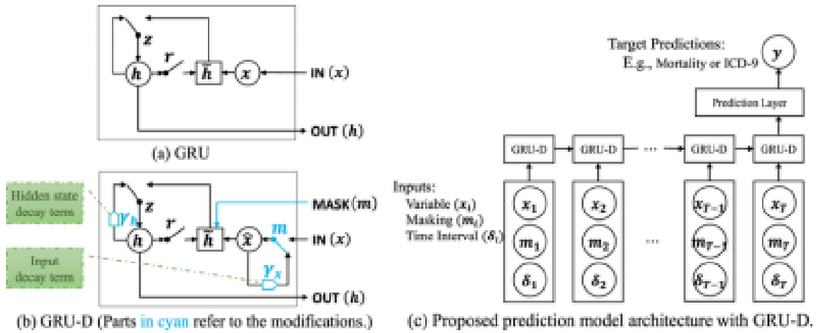
Time-series imputation model

RNN 기반 기법

- **M-RNN(Multi – Directional Recurrent Neural Network)**
  데이터 스트림 내에서 보간과 데이터 스트림에 대치하는 두 방법을 접목하는 방식으로 동작
  특히, 데이터 스트림의 관계가 중요한 의료 데이터에 관하여 향상된 기능을 보여줌



- **GRU-D**
  기존 GRU 모델에서 결측 여부를 보여주는 마스킹 정보와 결측된 시간 간격을 고려하여 작동함
  이를 감쇠율이라 정의하고, 다른 실제 값들이 시간 간격 정보에 따라 결측치에 대하여 얼마나 미치는지에 대한 모델링을 통해 결정된다.

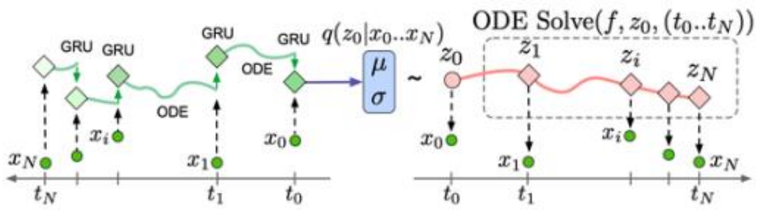


a) - GRU의 구조
b) - GRU-D의 구조
c) - GUR-D의 모델 아키텍처

Time-series imputation model

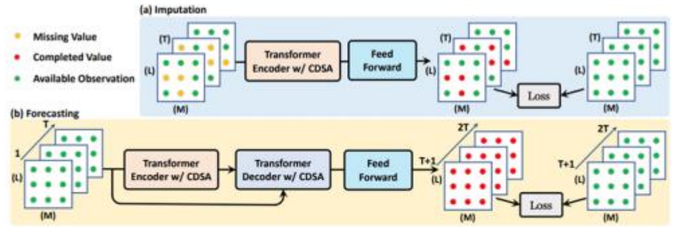
RNN 기반 기법

- **BRITS(Bidirectional Recurrent Imputation for Time Series) – 2018**
  특정 데이터의 분포 가정 없이 결측값을 대치하기 위하여 동적 시스템을 양방향 RNN으로 조정함
  여러 개의 상관된 결측값 처리가 가능, M-RNN은 결측값을 상수 취급하지만, 이러한 결측값 사이의 상관관계는 고려하지 않았음
  즉, 결측치간의 상관관계를 추가로 고려한 모델임
- **SSIM(Sequence-to-Sequence Imputation Model) – 2020**
  무선 센서 네트워크 상황에서 누락된 데이터를 복구하기 위해 새롭게 제안된 모델
  슬라이딩 윈도우 알고리즘을 사용하여 데이터양에 비해 많은 훈련 샘플을 생성한다는 특징을 가짐
- **ODE-RNN**
  간격이 균일하지 않은 시계열에 대해 단순 RNN 모델이 적용하기 어려운 점을 보완함
  미분 방정식(ODE)에 의해 RNN의 은닉 계수의 관계를 도출하여 학습함
- **Latent ODE**
  ODE-RNN과 달리 VAE(Variational Au-toEncoder)를 기반으로 하여 ODE-RNN을 인코더 구조로 ODE를 디코더 구조를 이용하여 처리하는 방식을 이용
  불규칙적으로 샘플링된 데이터도 Latent-ODE는 적용 가능하며 RNN보다 뛰어난 성능을 보여줌
- **CDSA(Cross-Dimensional Self-Attention)**
  다변량의 지리적인 위치가 지정된 시계열 데이터 처리의 경우 적합한 모델 – 참고
  위치 정보가 결합된 시계열의 데이터의 경우 더 효과적으로 처리함



출처 Reprinted with permission from Author[19]

그림 2 Latent ODE 모델



출처 Reprinted with permission from Author[20]

그림 3 CDSA 프레임워크

Time-series imputation model

GAN 기반 기법

- ✓ GAN의 기본적인 원리인 입력 데이터의 확률적 분포를 알아내고, 학습하여 데이터를 생성하는 원리를 이용함
- ✓ 생성자(G) : 실제 데이터와 비슷한 데이터를 만들어 낼 수 있도록 학습을 진행
- ✓ 구분자(D) : 실제 데이터와 생성자가 생성한 가짜 데이터를 잘 구분할 수 있도록 설계됨
- ✓ 이와 같이 생성자와 구분자가 서로 대립하면서 성능을 개선하는 원리가 GAN임

- **GAIN(Generative Adversarial Imputation Networks)**
  생성자(G)는 실제 데이터의 일부 구성요소를 관찰하고 실제로 관찰된 데이터에 따라서 결측된 데이터를 대치한다.
  구분자(D)는 대치된 데이터와 실제 데이터가 맞는지 판별한다.

  이 때, 구분자에 벡터 형식으로 몇가지 원본 샘플 데이터의 누락에 대한 부분 희트를 제공함
  이를 통하여 생성자는 실제 데이터 분포에 따라서 생성하는 법을 학습함

- **GRUI-GAN**
  GRU-D구조를 약간 변형하여 GAN의 구조에 결합한 기술
  GAN과 GRU의 구조적 결합을 통해 정확성을 높인 모델이다.
  단점 : 모델 학습시간이 길고, 임의의 노이즈가 입력으로 들어가 정확도가 안정적이지 못함

- **E2GAN**
  GRUI-GAN 개발진들이 단점을 보완한 모델임
  임의의 노이즈가 아닌 오토 인코더 구조를 추가하는 방향이 주된 기여점

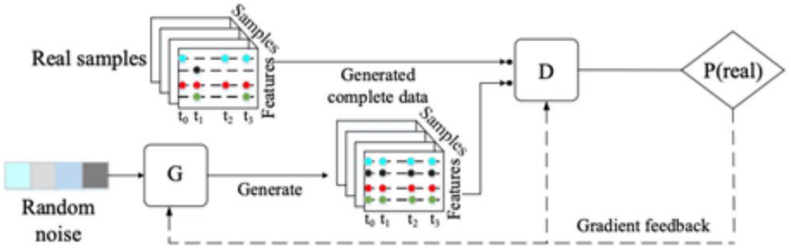
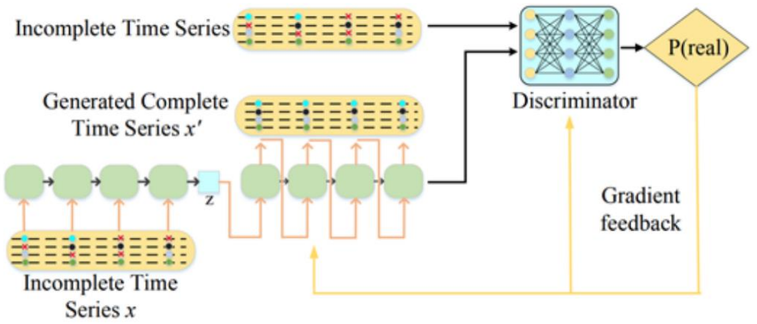
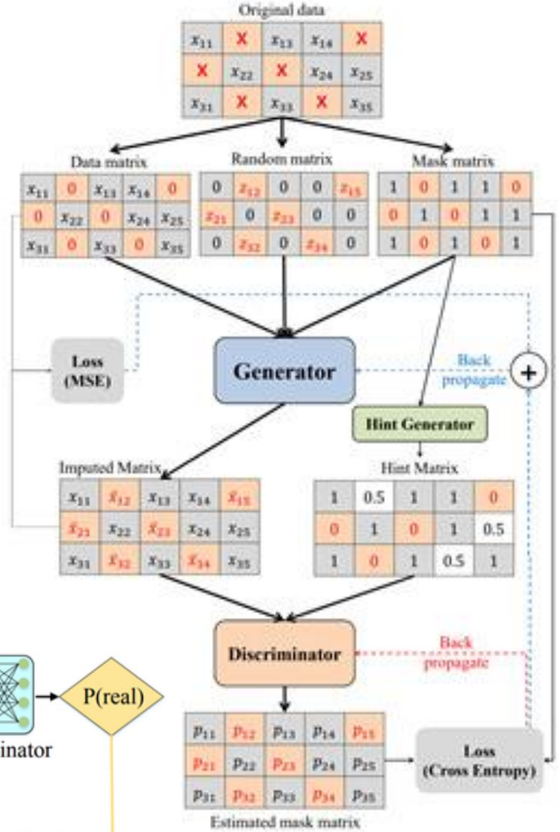


Figure 3: The structure of the proposed model.

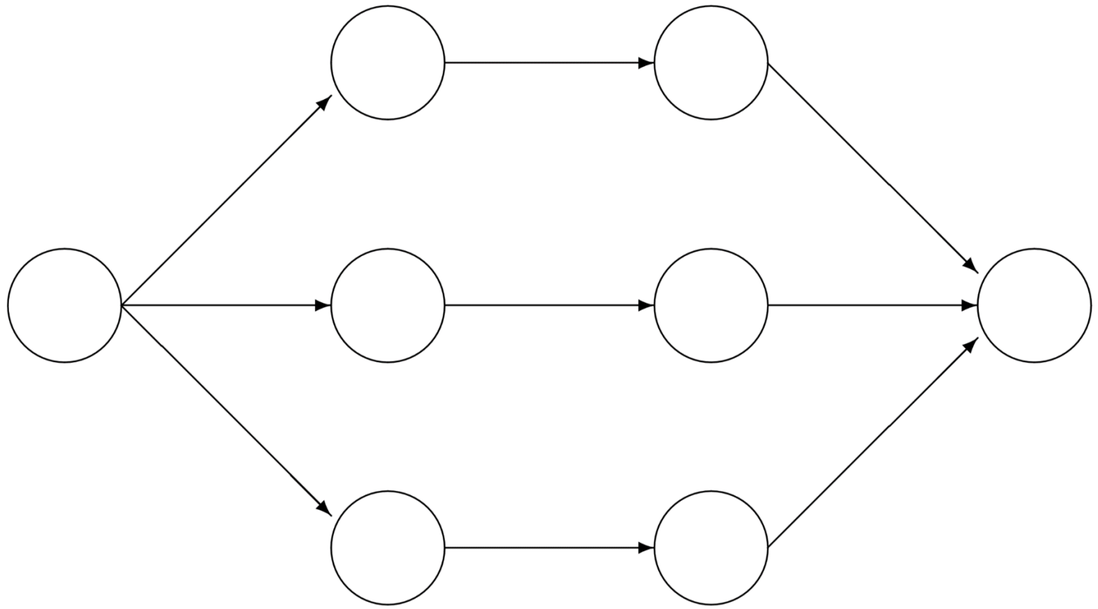




Time-series imputation model

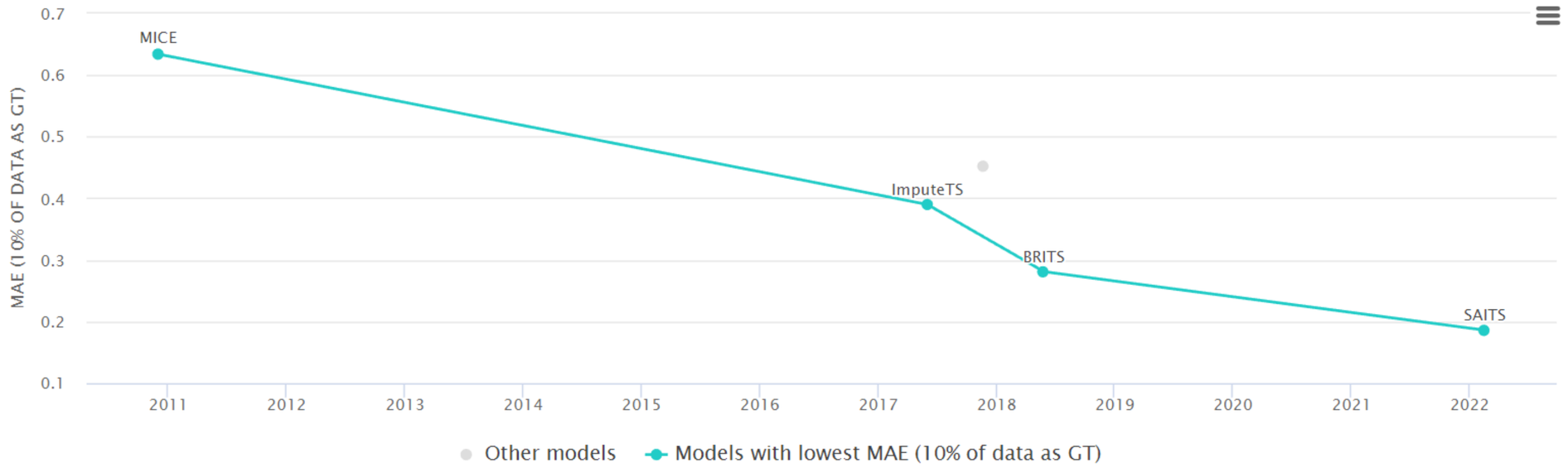
MICE

데이터셋 생성 : 특정 알고리즘에 따라 결측 값을 대체 값(평균)값으로 바꾼 m개의 데이터 셋 생성
분석과 추정 : m개의 완전한 데이터셋을 각각 원하는 분석 기법에 대해 분석하고, 그 결과에 대해 추정치와 표준오차 계산
결합 : 각 데이터셋의 결과를 Rubin’s rule에 의거하여 결합
이후에 해당 추정치의 평균을 결과값으로 나타냄



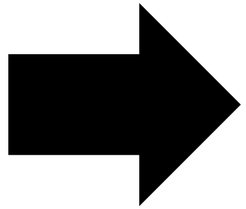
Time-series imputation model

PhysioNet



Rank	RNN-VAE	Latent-ODE + Poisson	Latent-ODE	MICE	M-RNN	ImputeTS	BRITS	SAITS
MAE	5.930(mse)	2.789(mse)	2.118(mse)	0.634	0.451	0.390	0.281	0.186

방문횟수	Counts	환자 수	누적 방문(방문횟수 이상)
5	100	20	625
6	162	27	605
7	182	26	578
8	216	27	552
9	243	27	525
10	390	39	498
11	352	32	459
12	480	40	427
13	689	53	387
14	728	52	334
15	585	39	282
16	656	41	243
17	748	44	202
18	720	40	158
19	912	48	118
20	820	41	70
21	441	21	29
22	176	8	8



- 방문 횟수 마다 M-RNN모델 수행
- 방문횟수 : 20이하

Time-series imputation model

PPMI

Rank	Mrnn(I)	Mrnn(F)	Mrnn(B)	MICE	Imputation(mean)
10%	0.3827(233%)	0.1148	0.1368(19%)	0.1237(7%)	0.1384(20%)
20%	0.3813(234%)	0.1140	0.1375(20%)	0.1282(12%	0.1387(21%)
30%	0.3806(231%)	0.1148	0.1376(19%)	0.1439(25%)	0.1402(22%)

- MRNN(I) : 환자의 Visit 마다 mrnn 수행
- MRNN(F) : 환자의 Visit 5부터 mrnn수행
- MRNN(B) : 환자의 Visit 20부터 역으로 mrnn 수행

Thank You