
How to augment your time-series data?



2022. 12. 16

Data Mining & Quality Analytics Lab.

황 순 혁

발표자 소개



❖ 황순혁 (Sunhyeok Hwang)

- 고려대학교 일반대학원 산업경영공학과 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석·박사 통합과정 2학기차 (2022. 03 ~)

❖ Research Interest

- Augmentation Strategy
- Representation Learning for Time-series data
- Self-supervised Learning

❖ Contact

- shhwang1@korea.ac.kr

Contents

❖ Introduction

- Background
 - Time-series data
 - Data Augmentation

❖ Data Augmentation for Time-series Data

- Basic Data Augmentation Methods
- Advanced Data Augmentation Methods

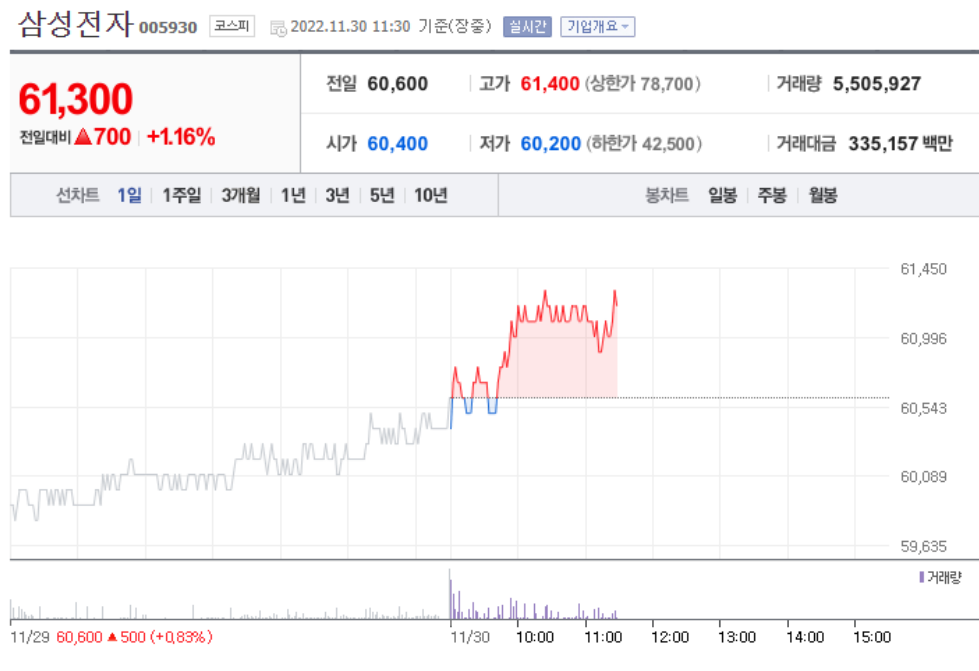
❖ Conclusion

Introduction

Introduction

Background

Time-series data란 무엇일까?



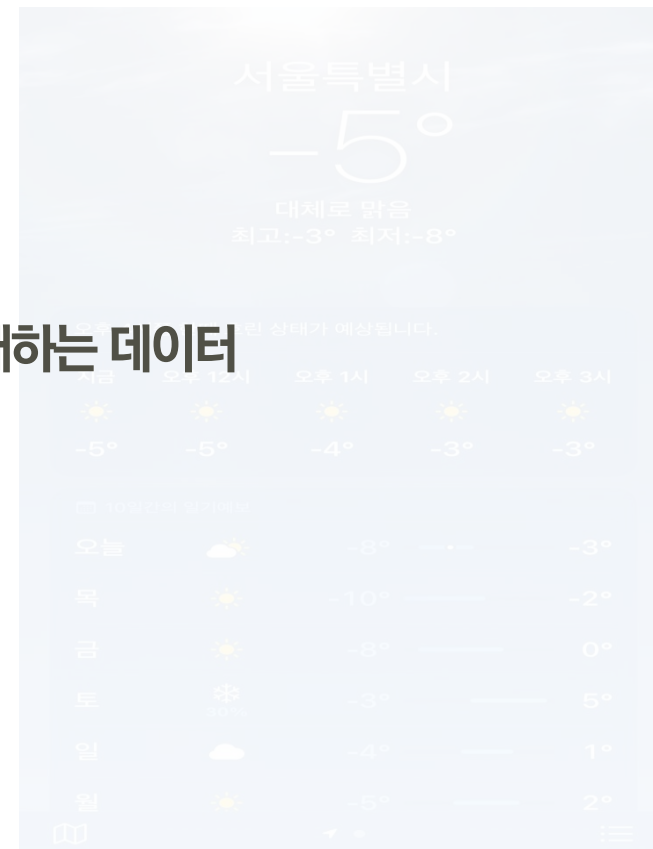
Introduction

Background

Time-series data란 무엇일까?



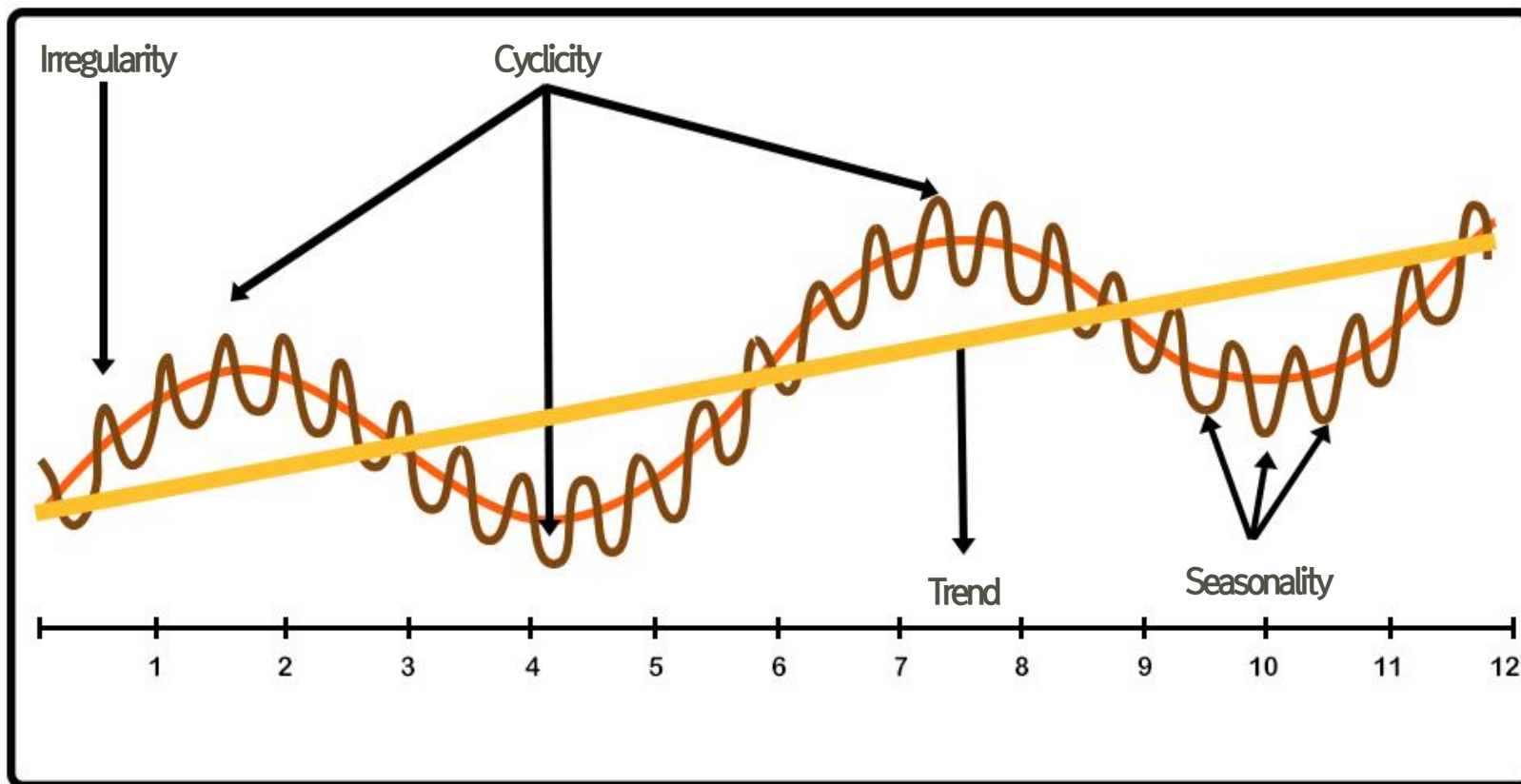
입력 데이터에 **시계열성**이 존재하는 데이터



Introduction

Background

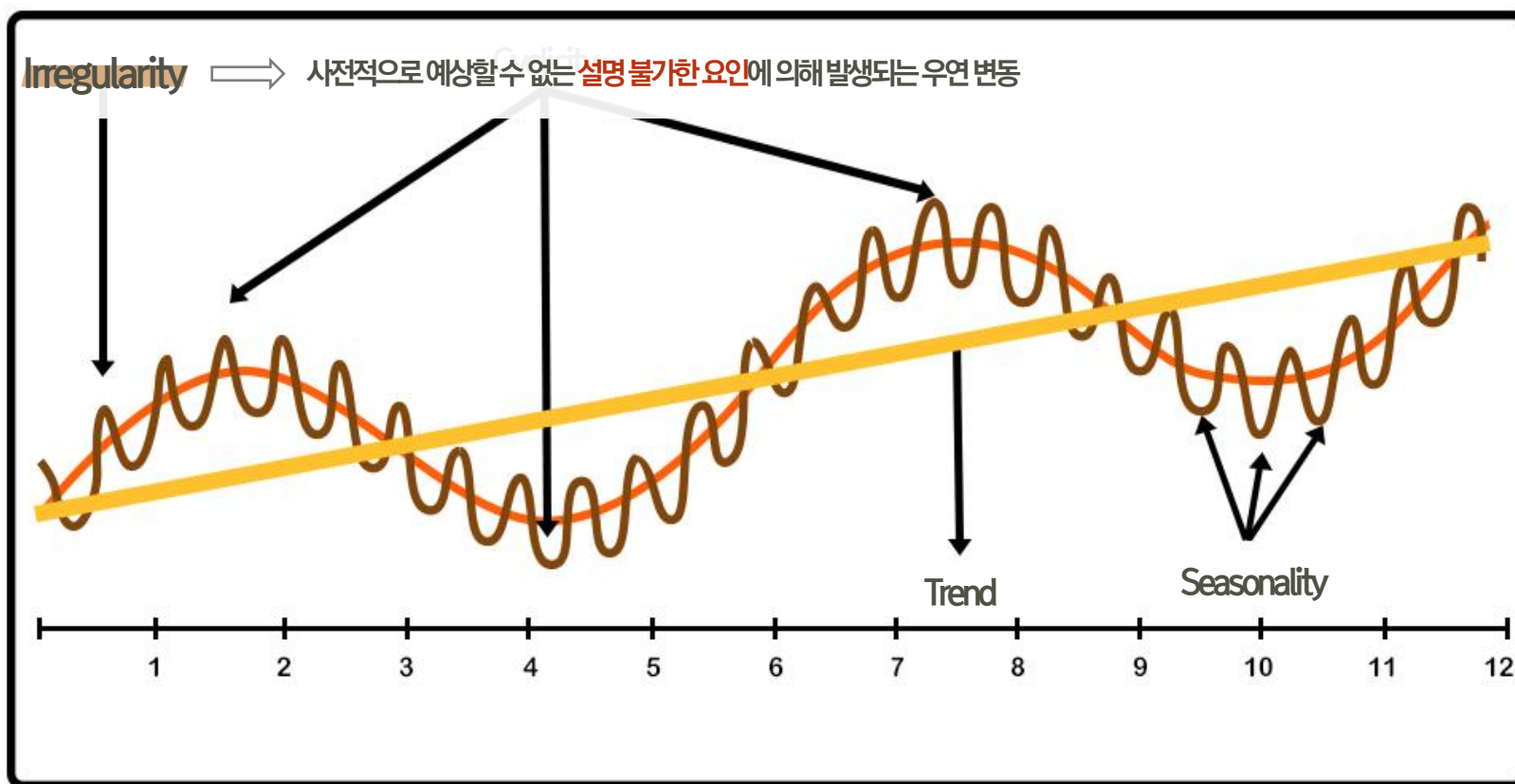
Time-series data 내 존재하는 특성



Introduction

Background

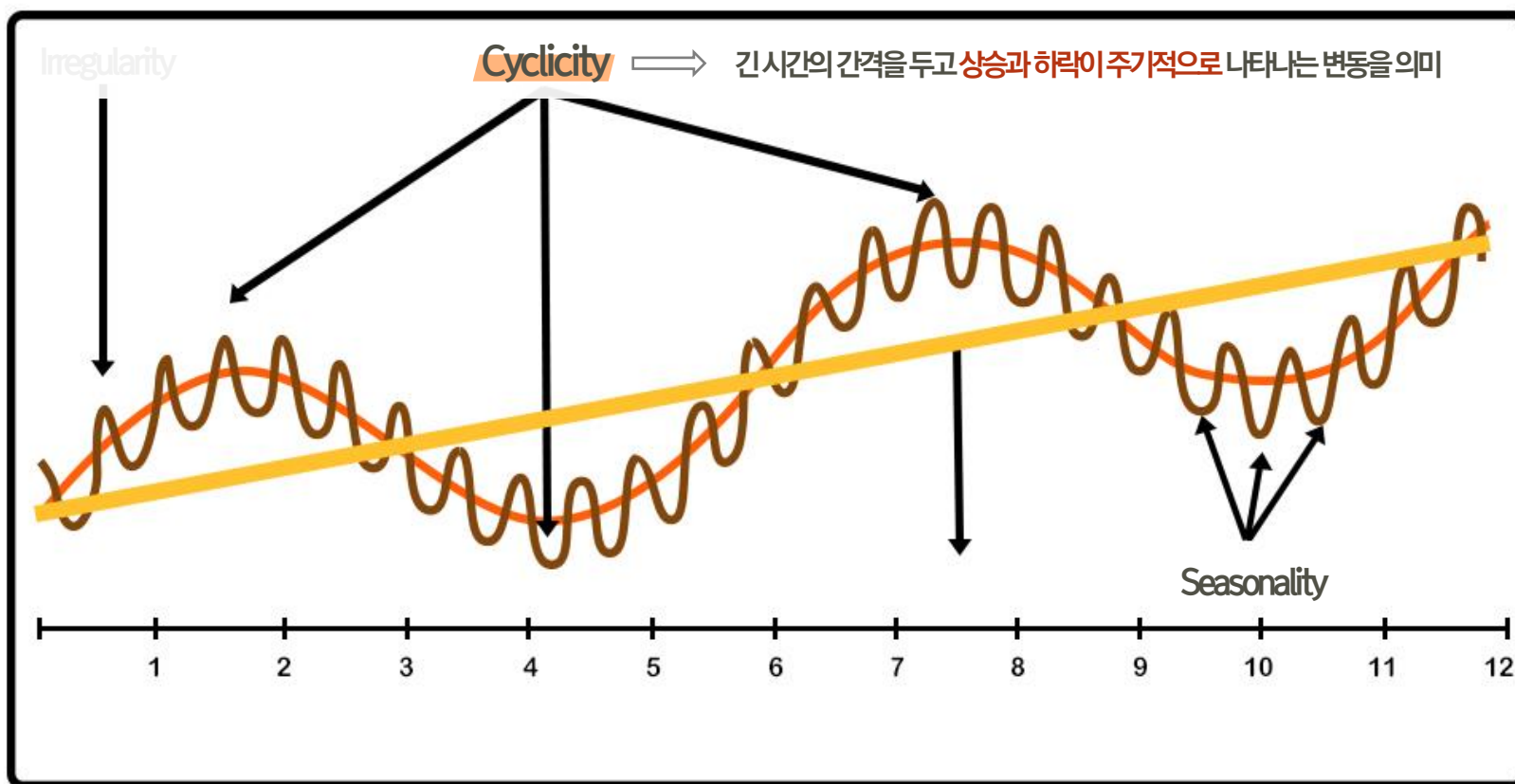
Time-series data 내 존재하는 특성



Introduction

Background

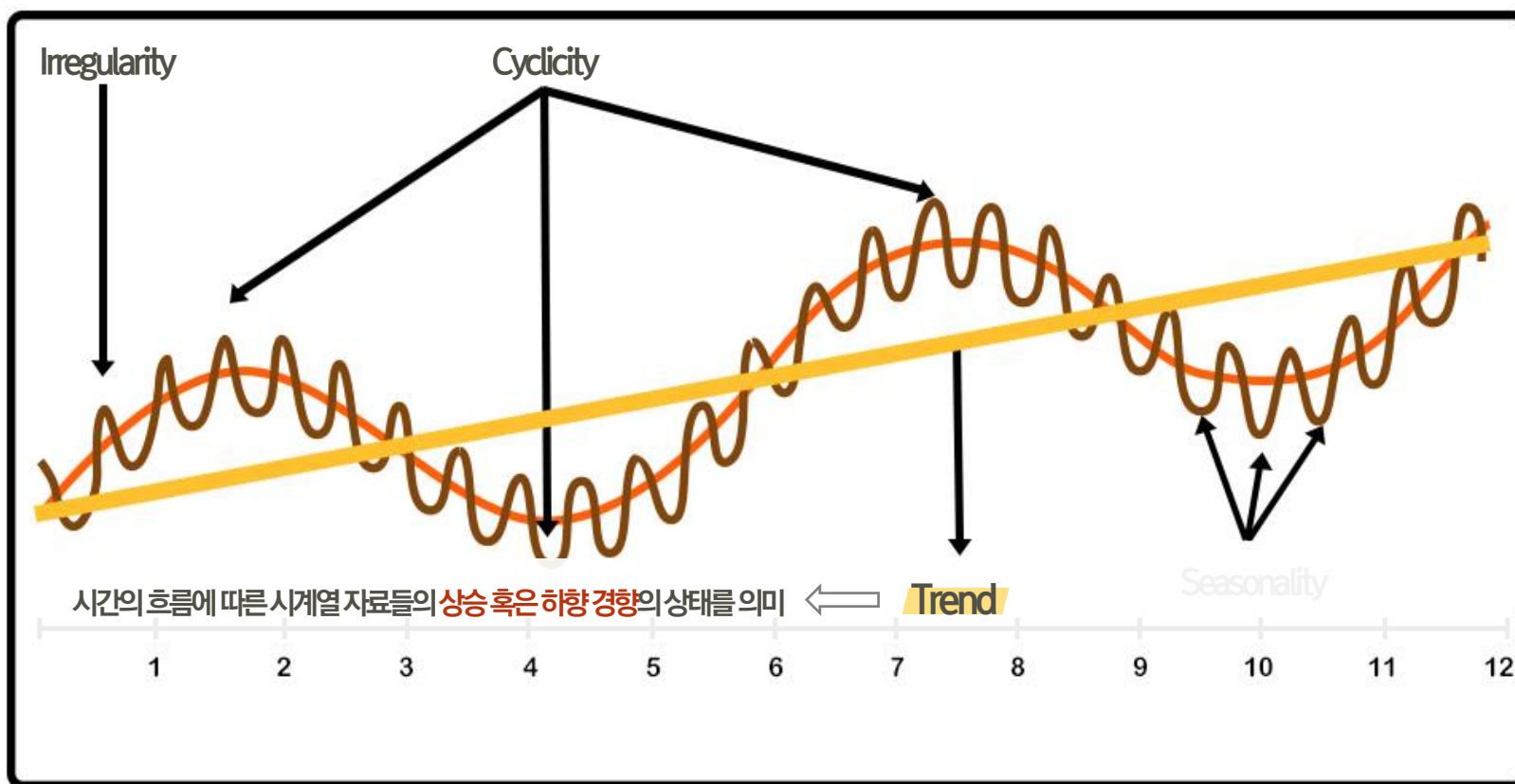
Time-series data 내 존재하는 특성



Introduction

Background

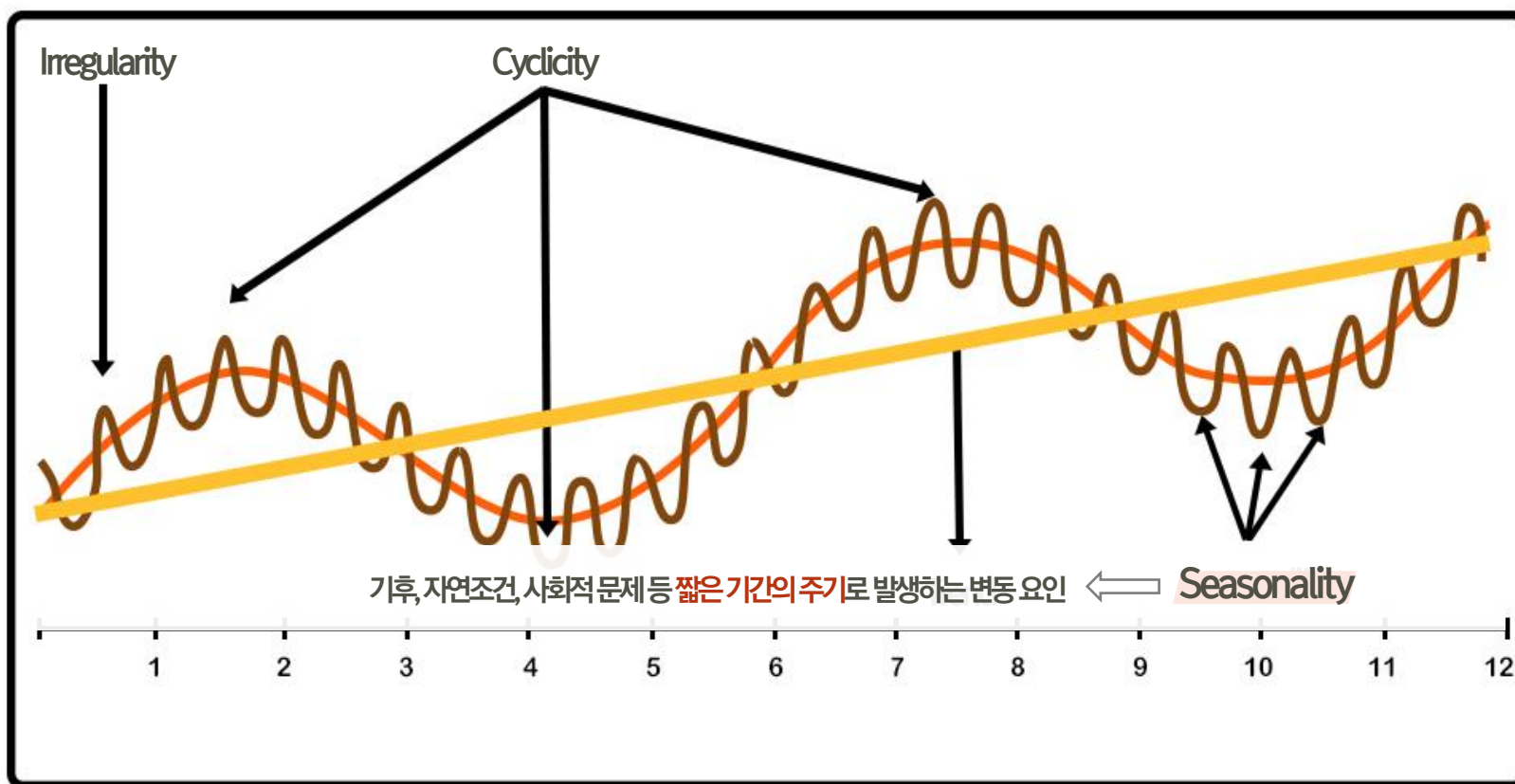
Time-series data 내 존재하는 특성



Introduction

Background

Time-series data 내 존재하는 특성

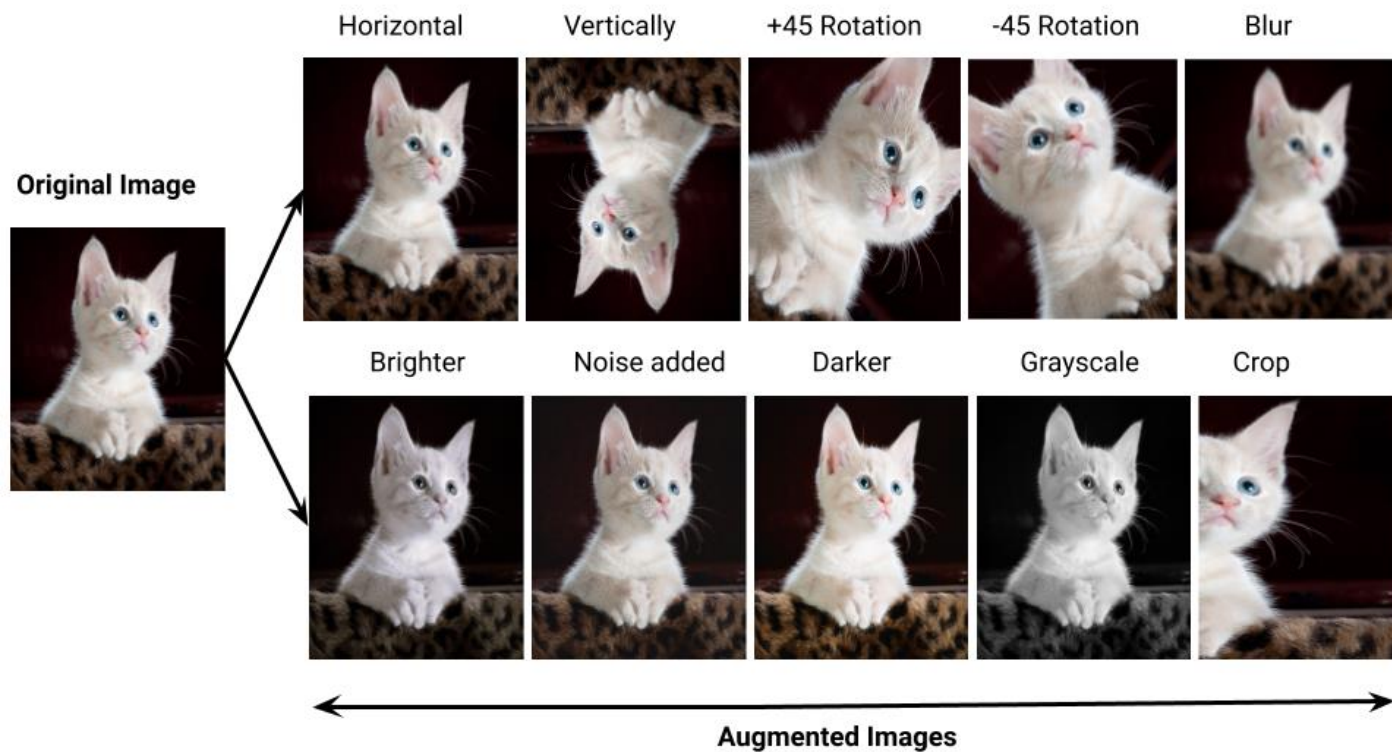


Introduction

Background

Data Augmentation

왜 멀쩡한 이미지를 이렇게 변형할까?

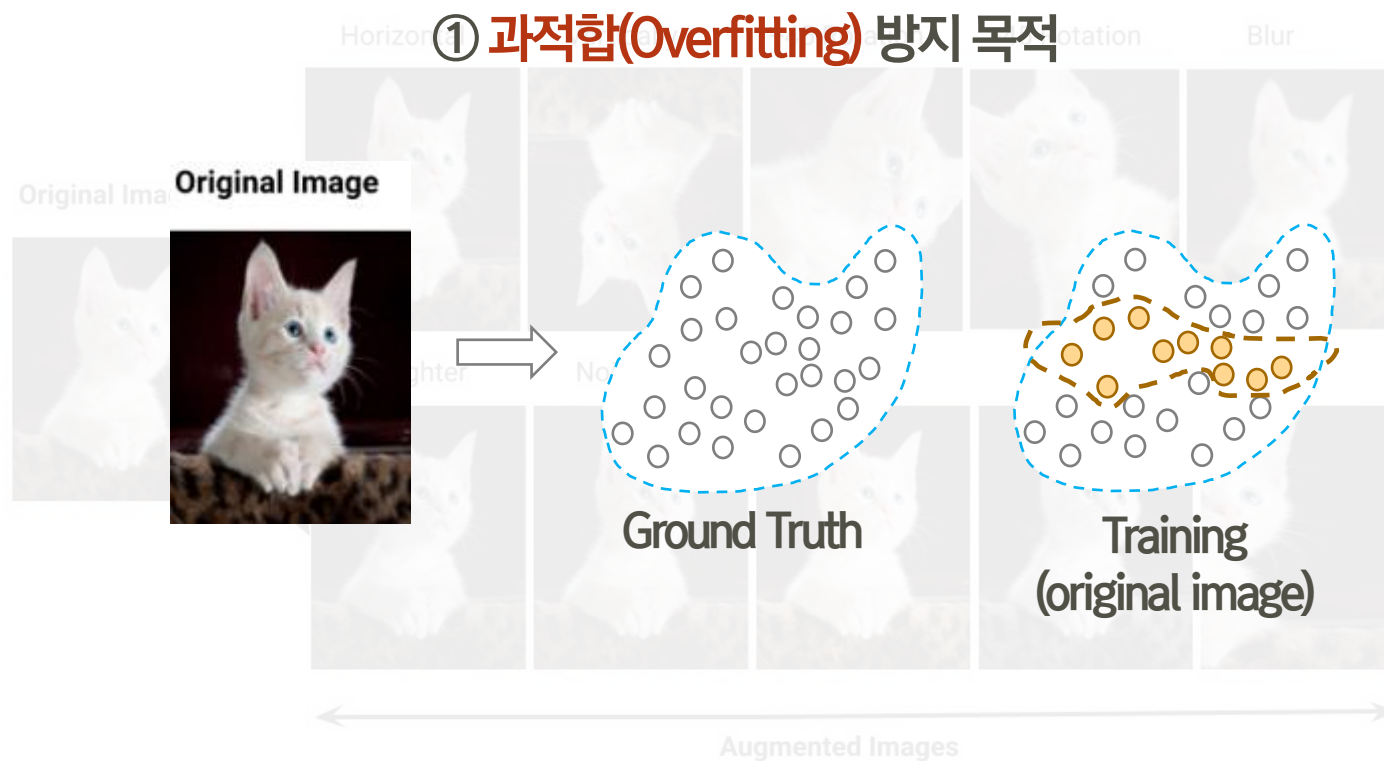


Introduction

Background

Data Augmentation

왜 멀쩡한 이미지를 이렇게 변형할까?



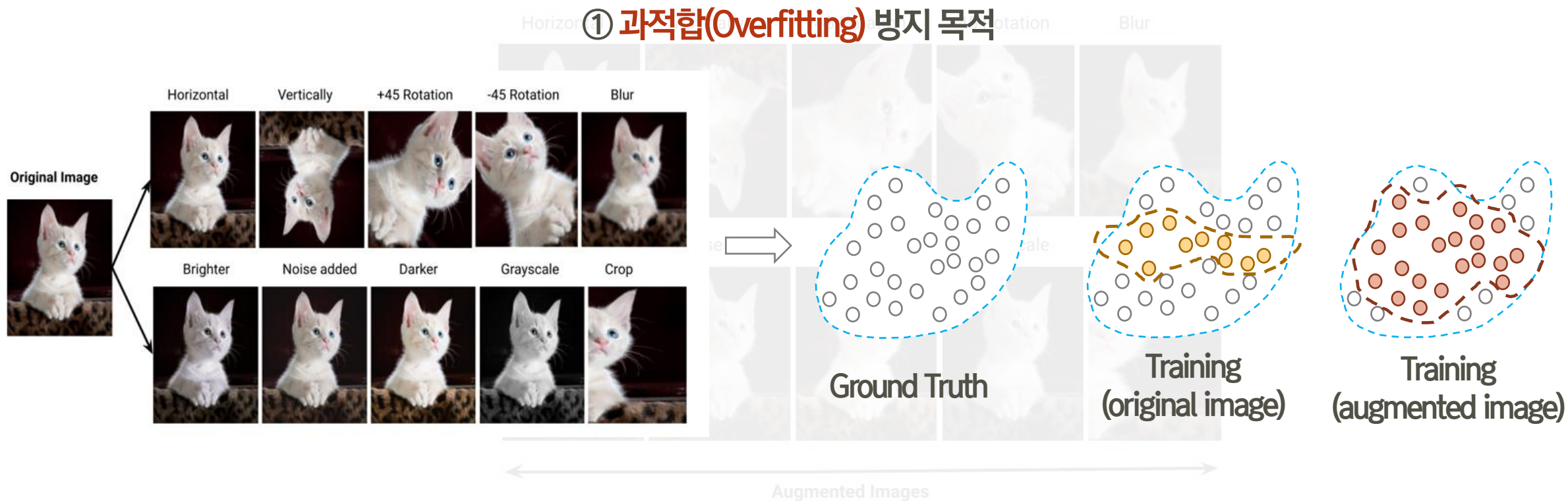
Introduction

Background

Data Augmentation

왜 멀쩡한 이미지를 이렇게 변형할까?

① 과적합(Overfitting) 방지 목적



Introduction

Background

Data Augmentation

왜 멀쩡한 이미지를 이렇게 변형할까?

① 과적합(Overfitting) 방지 목적



② 데이터 특징(Feature)에 대한 강건한 학습 가능

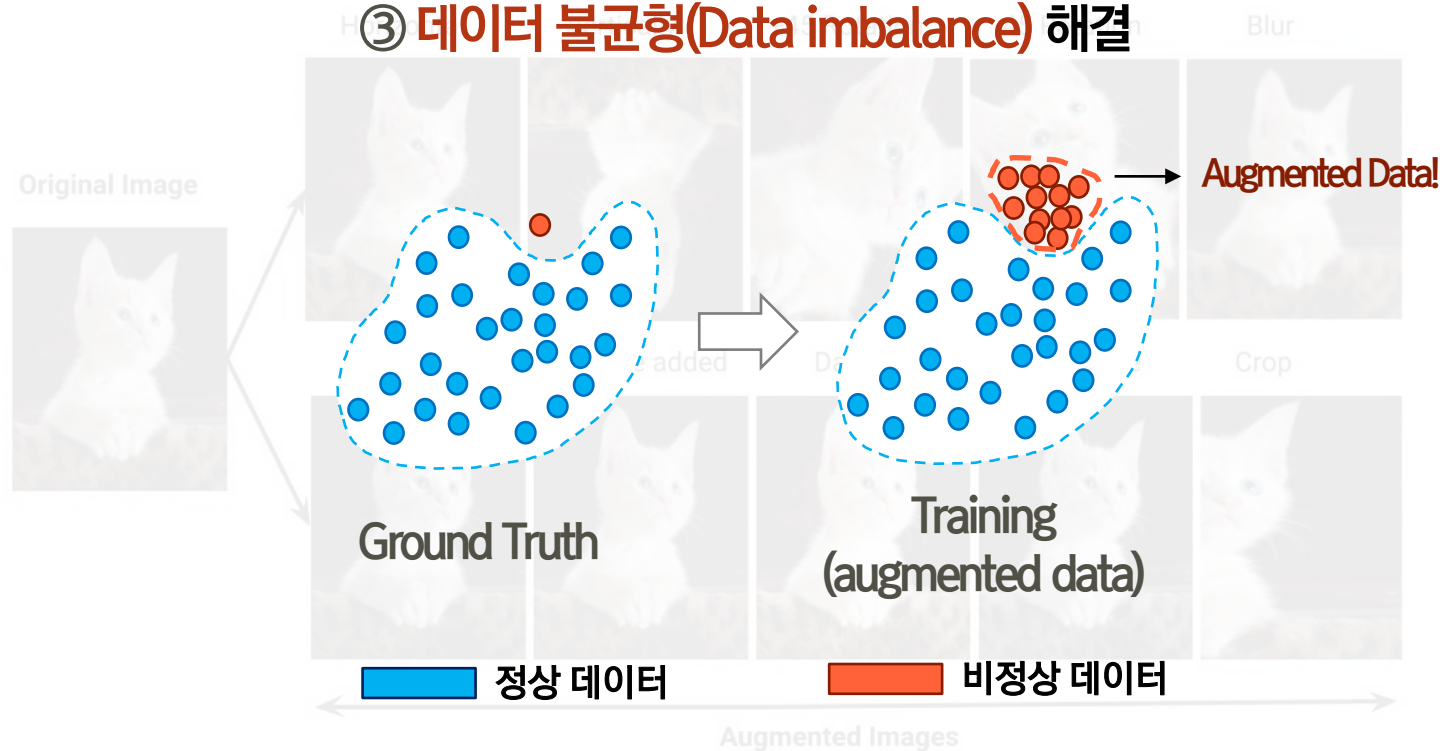
Introduction

Background

Data Augmentation

왜 멀쩡한 이미지를 이렇게 변형할까?

③ 데이터 불균형(Data imbalance) 해결



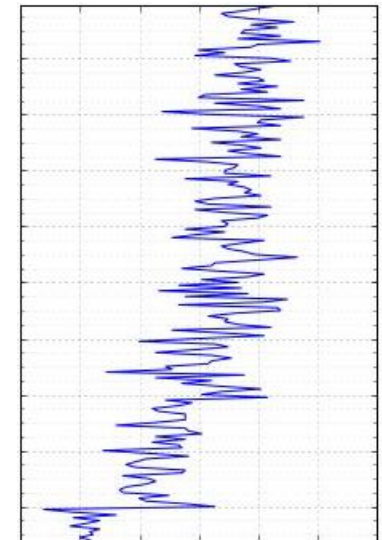
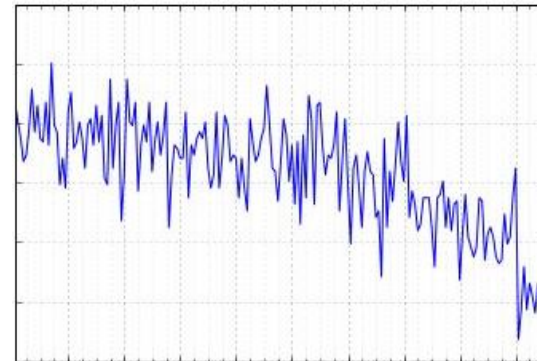
✓ 단, **Class가 유지**되는 선에서 Augmentation을 적용하는 것이 중요!

Introduction

Background

❖ Difficulty in applying augmentation to Time-series data

- 시간 종속적(Temporal dependency)인 시계열의 본질적 특성을 활용하지 못함
- 하고자하는 Task에 따라 데이터 증강의 적합성 여부가 달라지는 문제



90° Rotation Augmentation

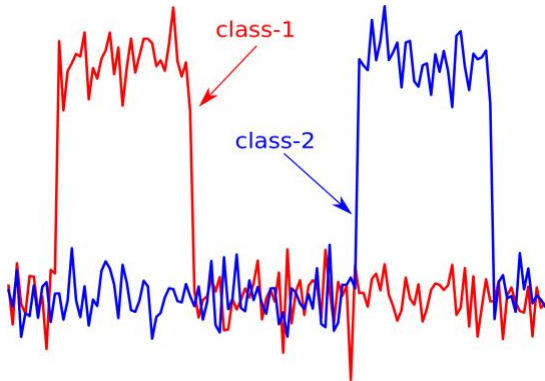
- ✓ 시간에 따른 분포와 특성이 변하는 문제

Introduction

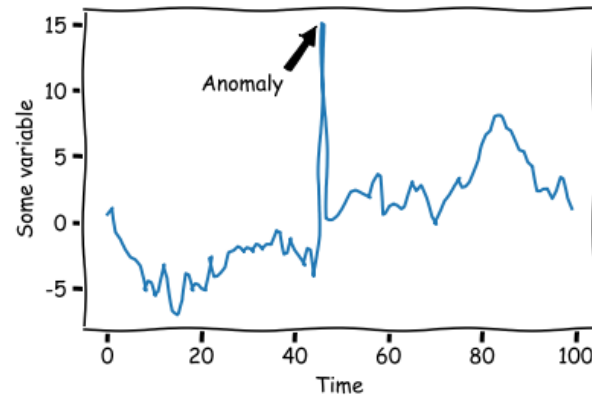
Background

❖ Difficulty in applying augmentation to Time-series data

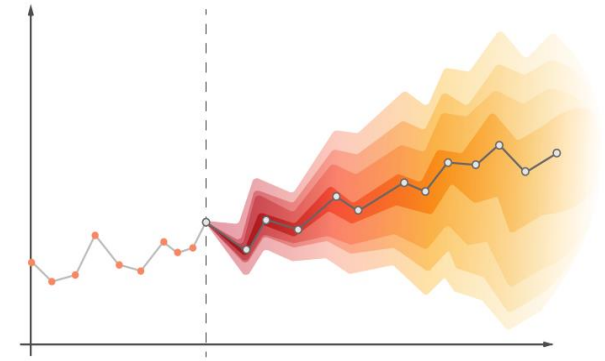
- 시간 종속적(Temporal dependency)인 시계열의 본질적 특성을 활용하지 못함
- 하고자하는 Task에 따라 데이터 증강의 적합성 여부가 달라지는 문제



Time-series Classification



Time-series Anomaly detection

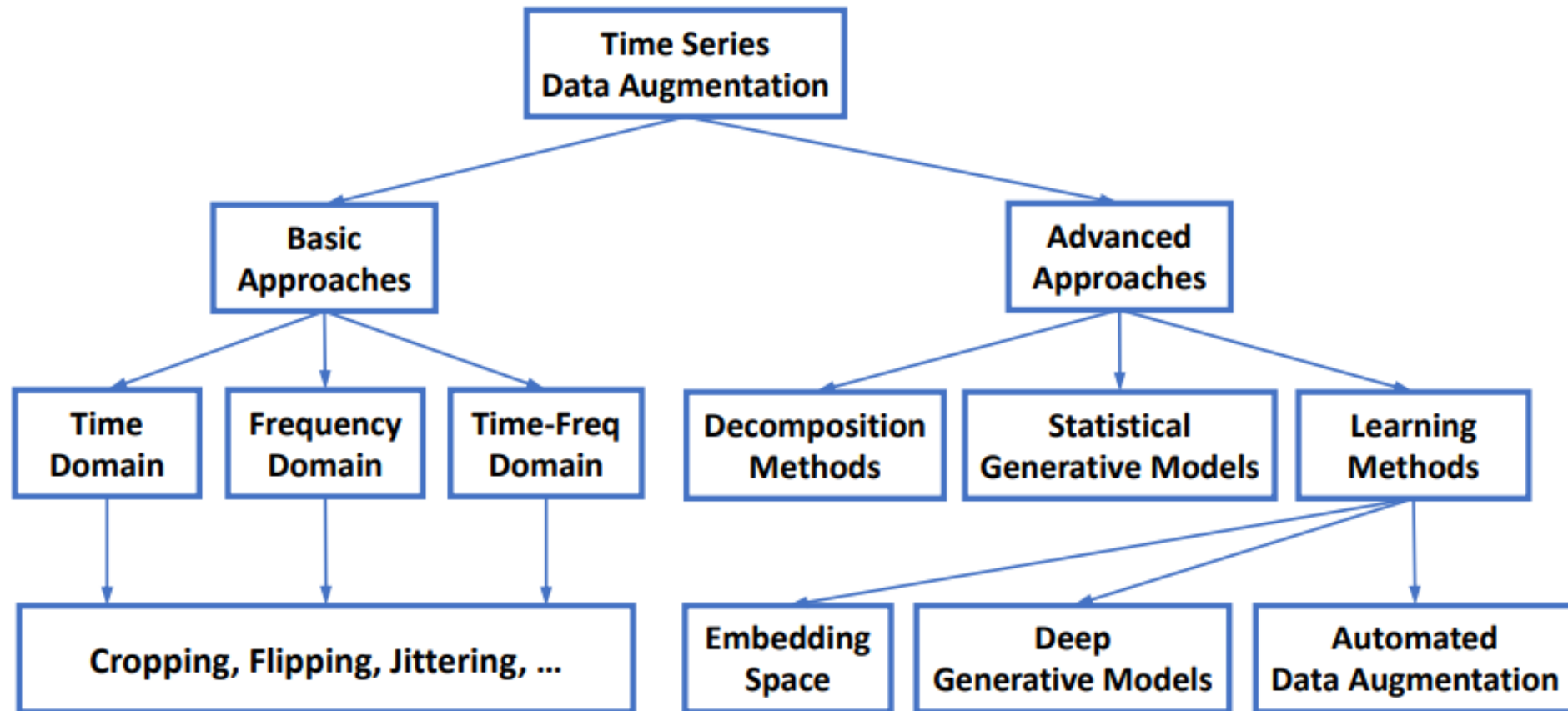


Time-series Forecasting

Data Augmentation for Time-series Data

Data Augmentation for Time-series Data

Taxonomy



Taxonomy of Time-series Data Augmentation

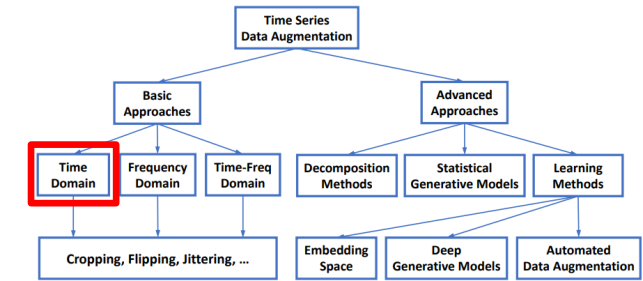
Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

- 매우 직관적인 augmentation 방법으로 많이 사용됨

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

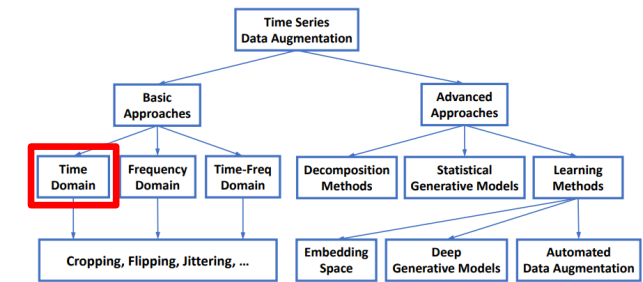


Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

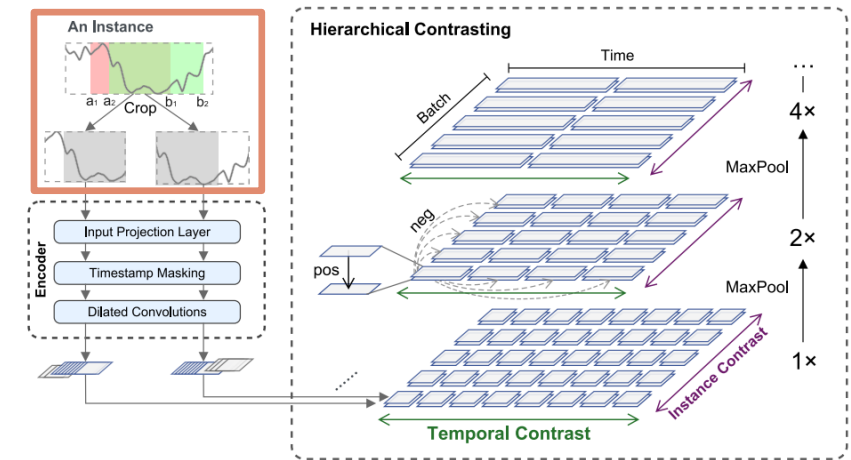
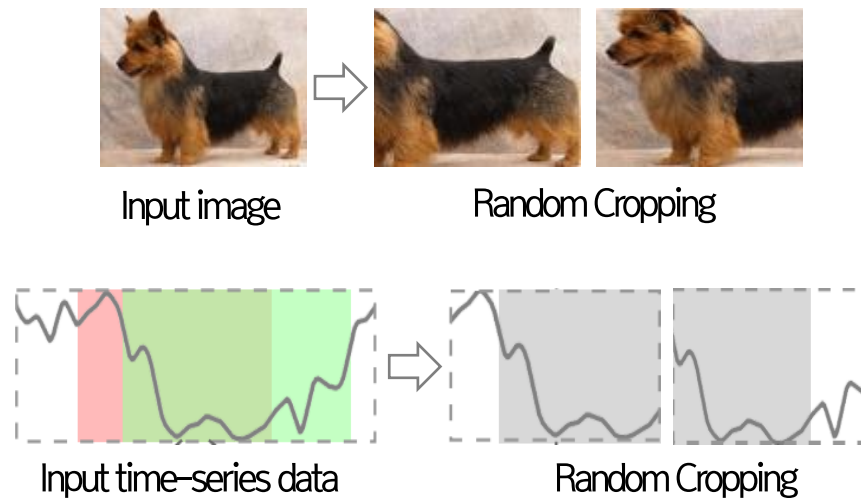
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Window cropping (Le guennec et al., 2016)

원본 데이터에서 Random하게 연속적인 일부분을 잘라내는 증강 기법
이미지 데이터를 랜덤하게 잘라내어 증강하는 방법과 유사

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion



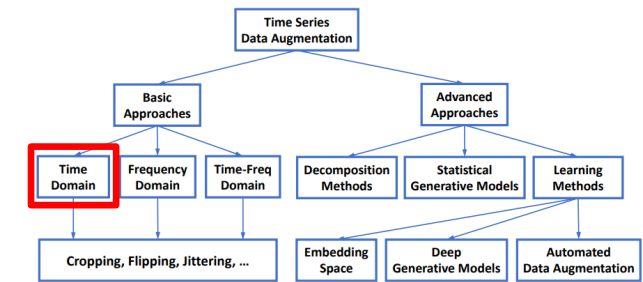
TS2Vec Algorithm

Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

- 매우 직관적인 augmentation 방법으로 많이 사용됨

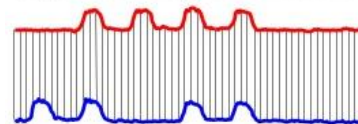


Window warping

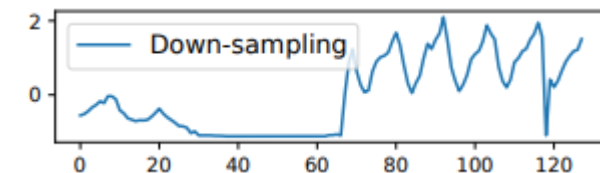
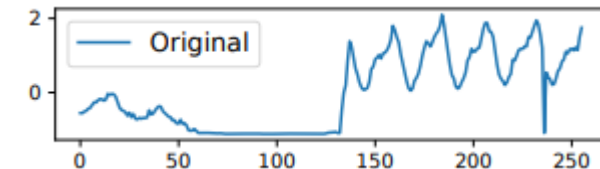
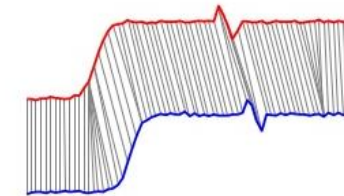
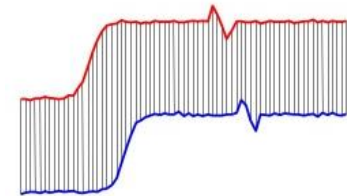
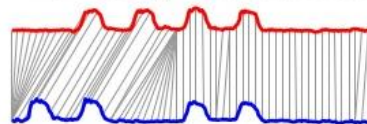
Random한 range의 시간 동안의 값들을 압축하거나 확장하는 증강 기법
Dynamic Time Warping(DTW)와 유사한 방법

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

Euclidean Distance
Sequences are aligned "one to one"



DTW
Nonlinear alignments are possible

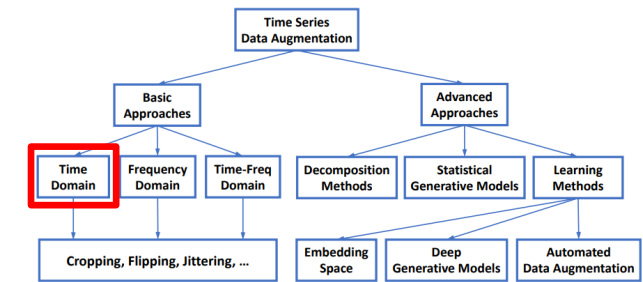


Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

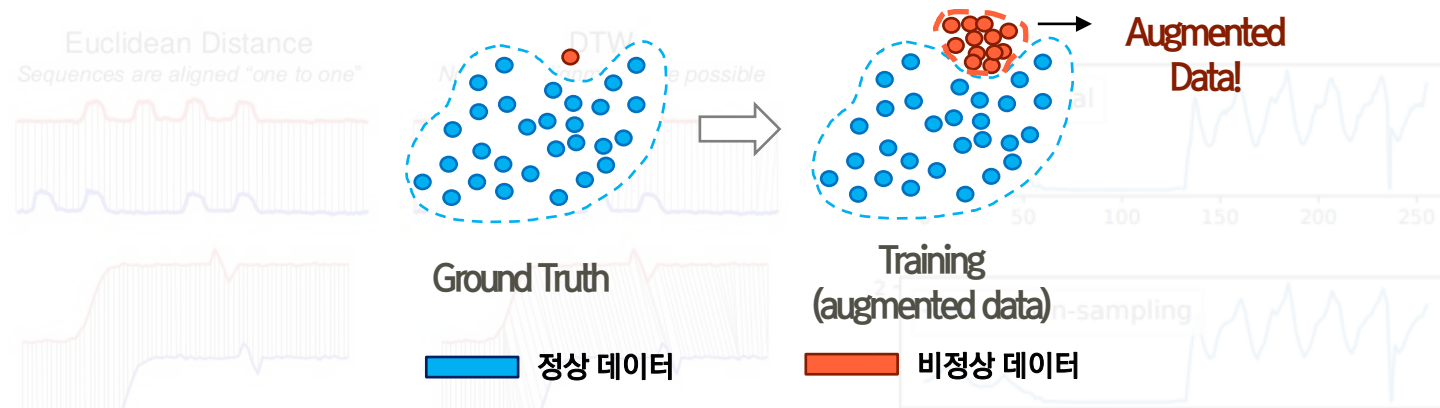
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Window warping

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

Random한 range의 시간 동안의 값들을 압축하거나 확장하는 증강 기법
Dynamic Time Warping (DTW)와 유사한 방법
③ 데이터 불균형(Data imbalance) 해결 - Upsampling



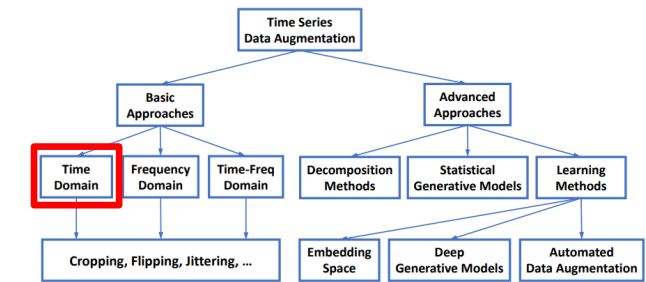
✓ 단, **Class가 유지**되는 선에서 Augmentation을 적용하는 것이 중요!

Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

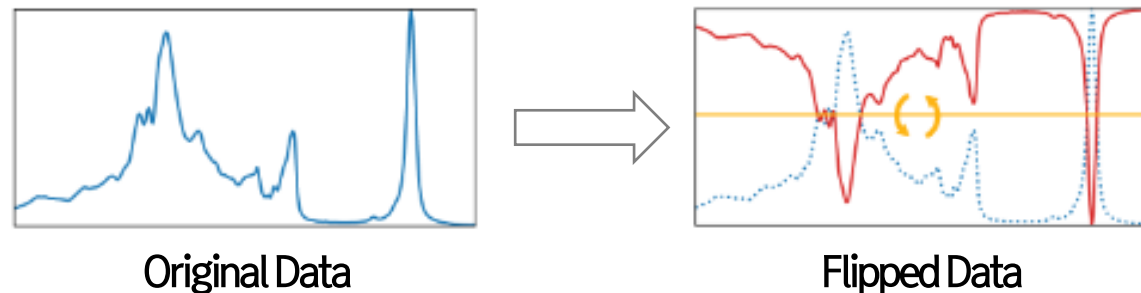
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Flipping

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

특정 지점을 기준으로 기존 데이터를 뒤집는 증강 기법 ($x'_t = -x_t$)
이상치 탐지 및 분류 문제에서 위 아래 방향의 대칭이 있다고 가정할 때 레이블은 여전히 동일

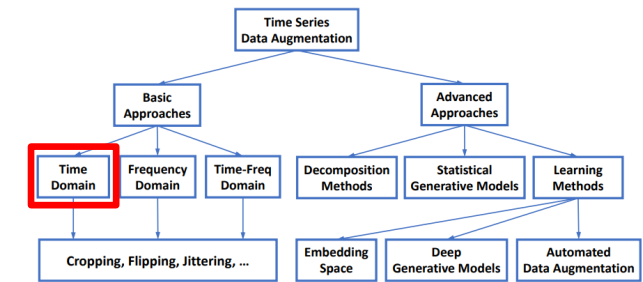


Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

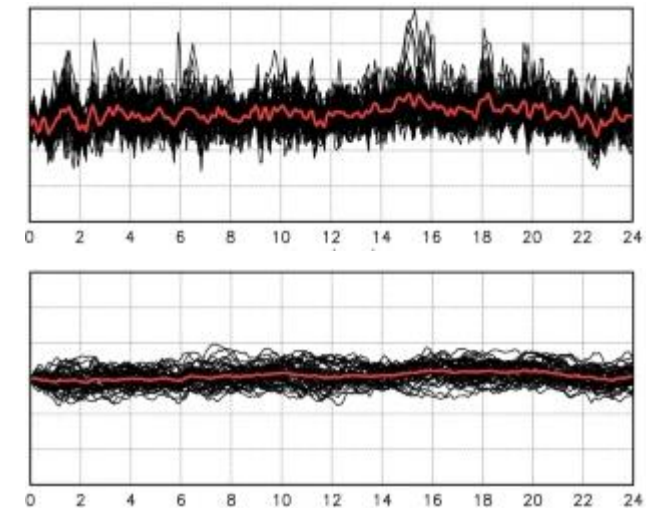
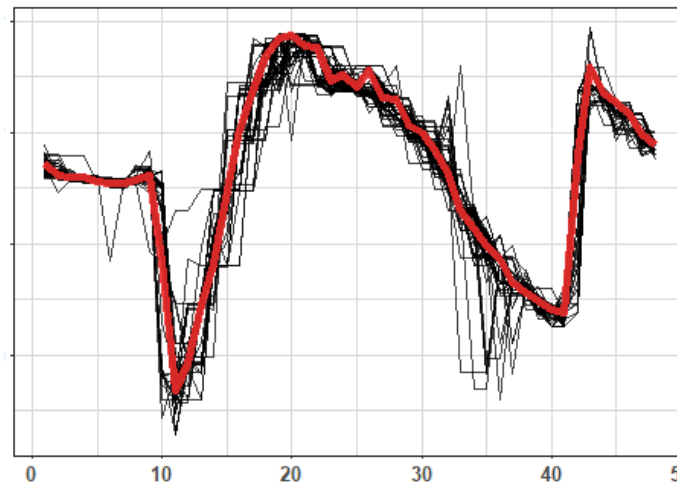
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Perturbation (& Ensemble)

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

- ① 증강 기법이 적용된 다양한 시계열 데이터들을 DTW 과정을 통해 원본 데이터의 Time length로 통일
- ② Time length가 통일된 데이터들을 Ensemble한 후 평균값을 이용하여 데이터 증강

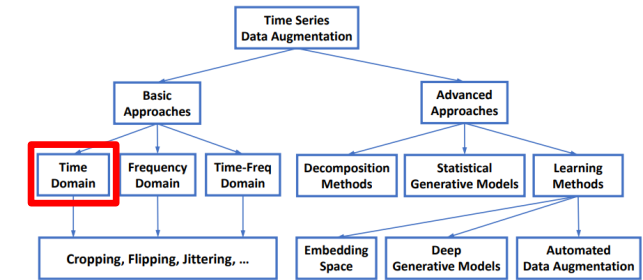


Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

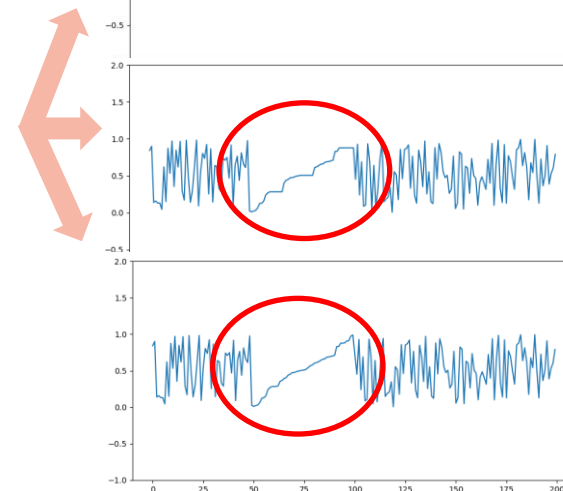
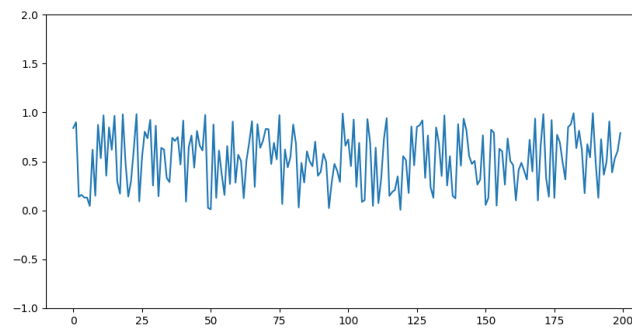
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Noise injection

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

기존 Class를 변경시키지 않는 선에서 약간의 noise/outlier를 추가하는 증강기법
대표적으로 Spike, Step-like trend, Slope-like trend 세가지 유형 존재



① Spike

② Step-like trend

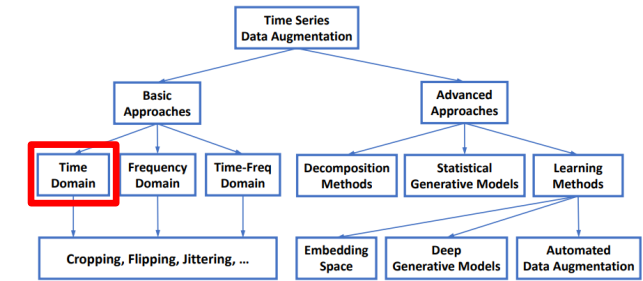
③ Slope-like trend

Data Augmentation for Time-series Data

Basic Approaches – Time domain

❖ Basic Approaches – Time domain

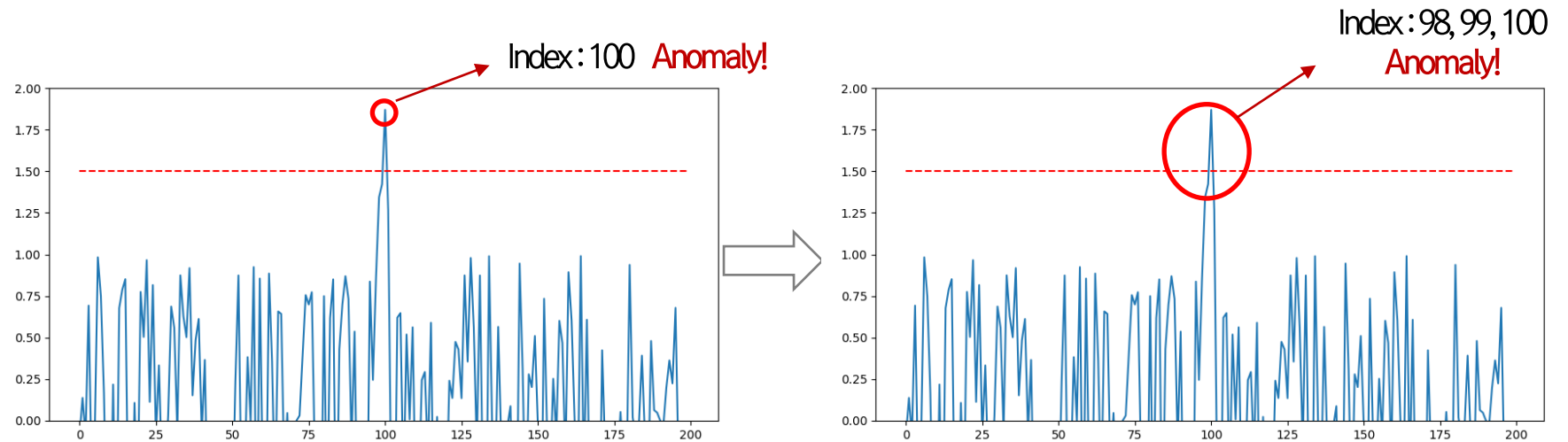
- 매우 직관적인 augmentation 방법으로 많이 사용됨



Label Expansion

1. Window cropping
2. Window warping
3. Flipping
4. Perturbation
5. Noise injection
6. Label Expansion

일반적으로 시계열 이상치 탐지 문제에서 이상치가 발견되기 전에 경향들이 존재
이상치가 발견된 한 지점뿐만 아니라 전후 경향이 발생하는 시점 또한 불량으로 판단하는 label augmentation 기법

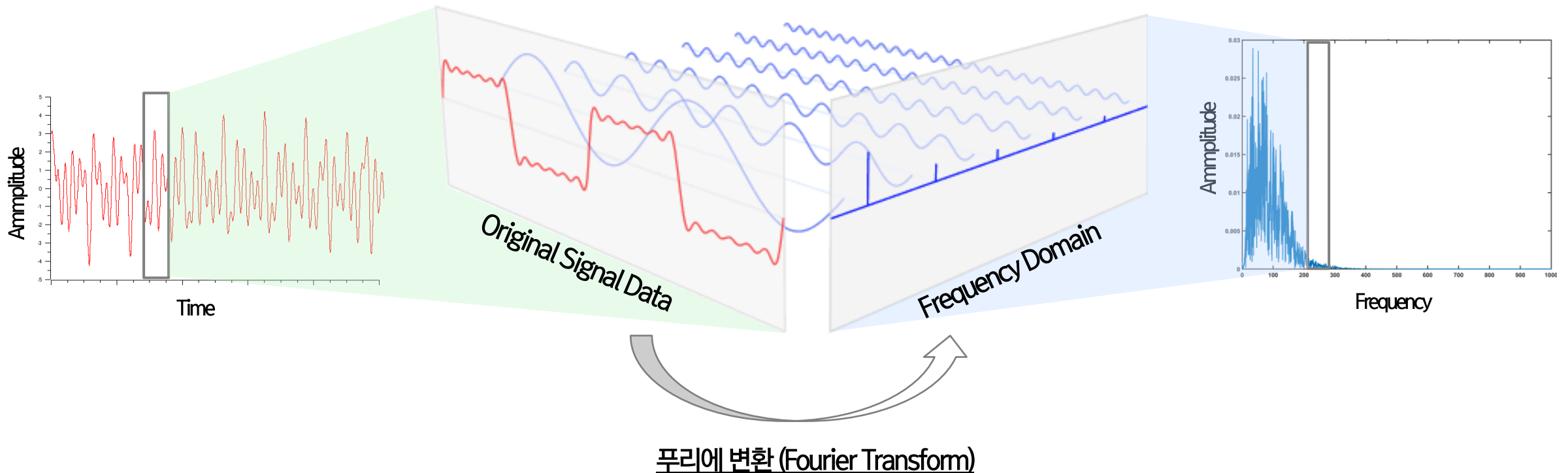
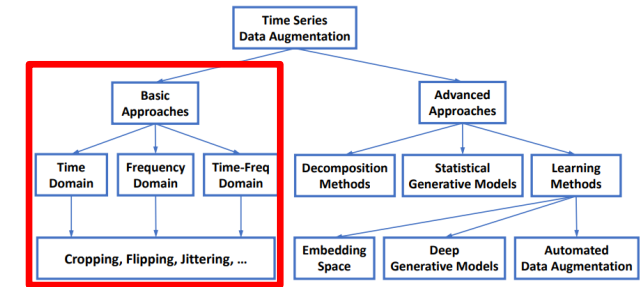


Data Augmentation for Time-series Data

Basic Approaches

❖ Domain in Time-series data

- 시간 영역(Time-Domain): 시계열 데이터의 가로 축에 해당하고, 시간의 정보를 담고 있는 영역
- 주파수 영역(Frequency-Domain): 시계열 데이터의 주파수 정보를 담고 있는 영역으로 푸리에 변환을 통해 시간 영역을 주파수 영역으로 변환

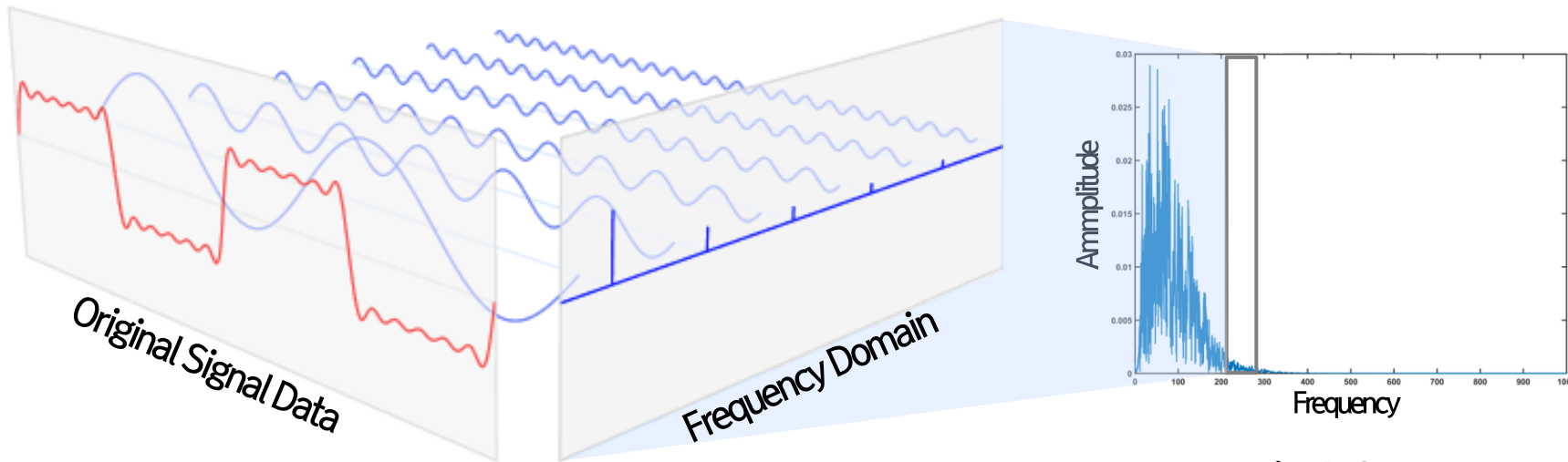


Data Augmentation for Time-series Data

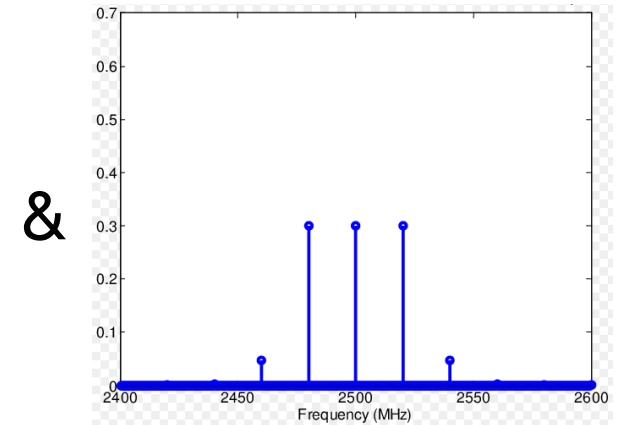
Basic Approaches

❖ Domain in Time-series data

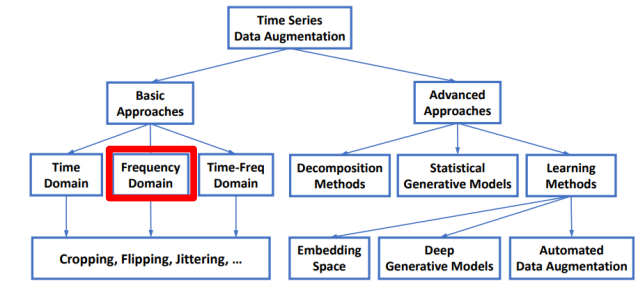
- 주파수 영역(Frequency domain)을 구성하는 두가지 스펙트럼
 - 1. Amplitude Spectrum : 주파수 영역에 대한 신호 값의 세기(진폭)을 나타내는 스펙트럼
 - 2. Phase Spectrum : 주파수 영역이 어느 위치에서 시작하는지 대한 시간 축 위치 정보를 담고있는 스펙트럼



Amplitude Spectrum



Phase Spectrum

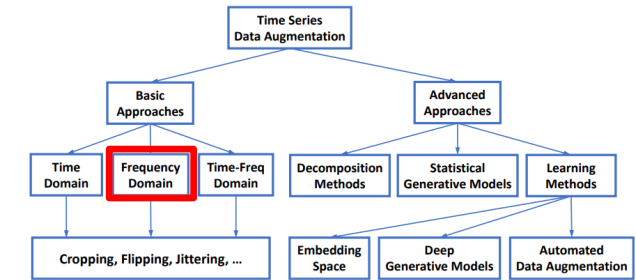


Data Augmentation for Time-series Data

Basic Approaches

❖ Domain in Time-series data

- 주파수 영역(Frequency domain)을 구성하는 두가지 스펙트럼
 - 1. **Amplitude Spectrum** : 주파수 영역에 대한 신호 값의 세기(진폭)을 나타내는 스펙트럼
 - 2. **Phase Spectrum** : 주파수 영역이 어느 위치에서 시작하는지 대한 시간 축 위치 정보를 담고있는 스펙트럼



종료 Dive into Audio Transformer

2022.04.01
Data Mining & Quality Analytics Lab.
발표자: 고은지
data2100@korea.ac.kr

Dive into audio transformer

발표자:  고은지

📅 2022년 4월 1일
🕒 오후 1시 ~
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

종료 Introduction to Analysis for Sound data

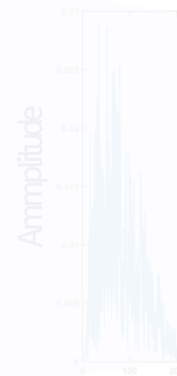
2020. 01. 08
Data Mining & Quality Analytics Lab.
발표자: 정 기원
dis517@korea.ac.kr

Introduction to Analysis for Sound data

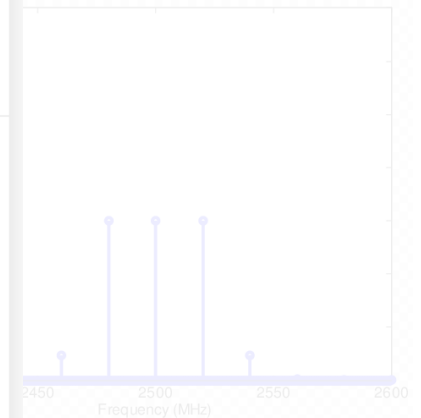
발표자:  정기원

📅 2021년 1월 8일
🕒 오후 1시 ~
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →



Amplitude Spectrum



Phase Spectrum



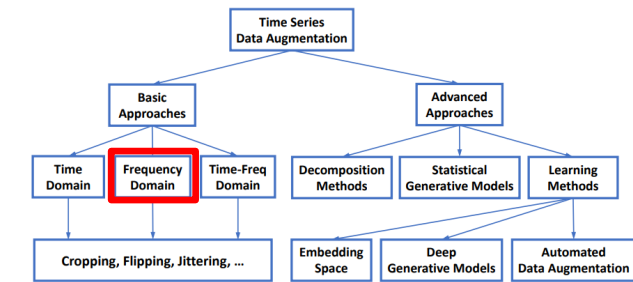
Original Signal Data

Data Augmentation for Time-series Data

Basic Approaches – Frequency domain

❖ Basic Approaches – Frequency domain

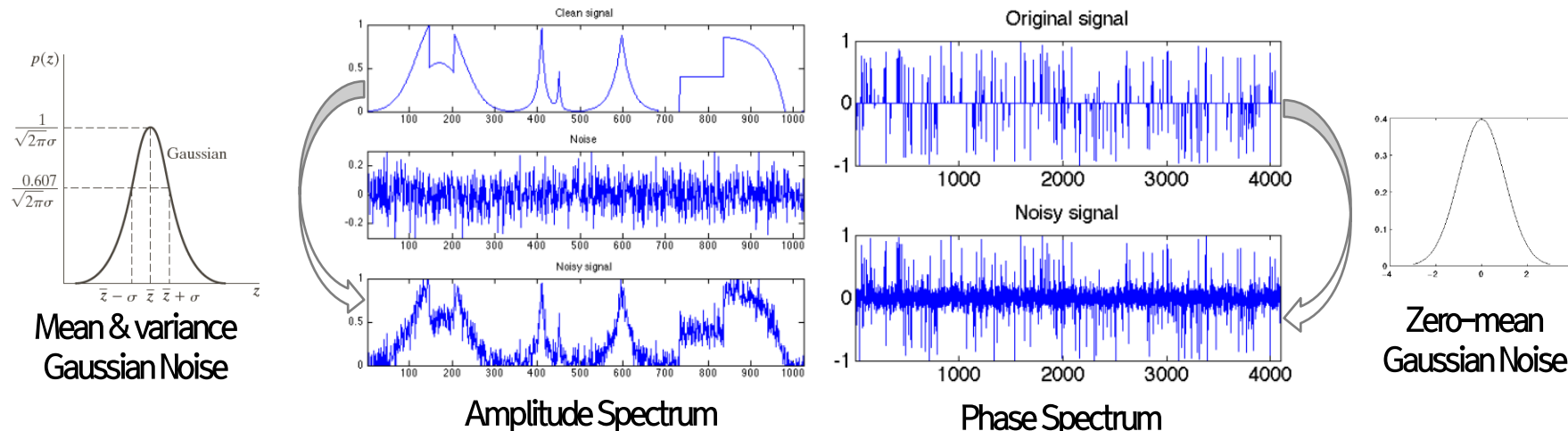
- Time-domain보다 현재까지의 기법 수는 적지만, 최근 연구가 활발히 진행됨
 - Amplitude & Phase Perturbations (APP)
 - Amplitude adjusted Fourier transform (AAFT)



Amplitude & Phase Perturbations (APP)

Amplitude spectrum → 일부 Amplitude 값에 대해 전체 데이터의 평균&분산 값을 갖는 Gaussian Noise를 부여

Phase spectrum → 일부 Phase 값에 대해 zero-mean Gaussian noise를 부여

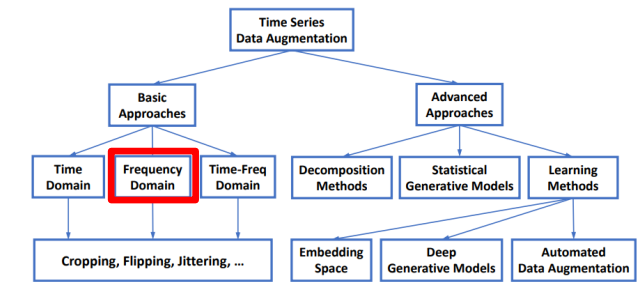


Data Augmentation for Time-series Data

Basic Approaches – Frequency domain

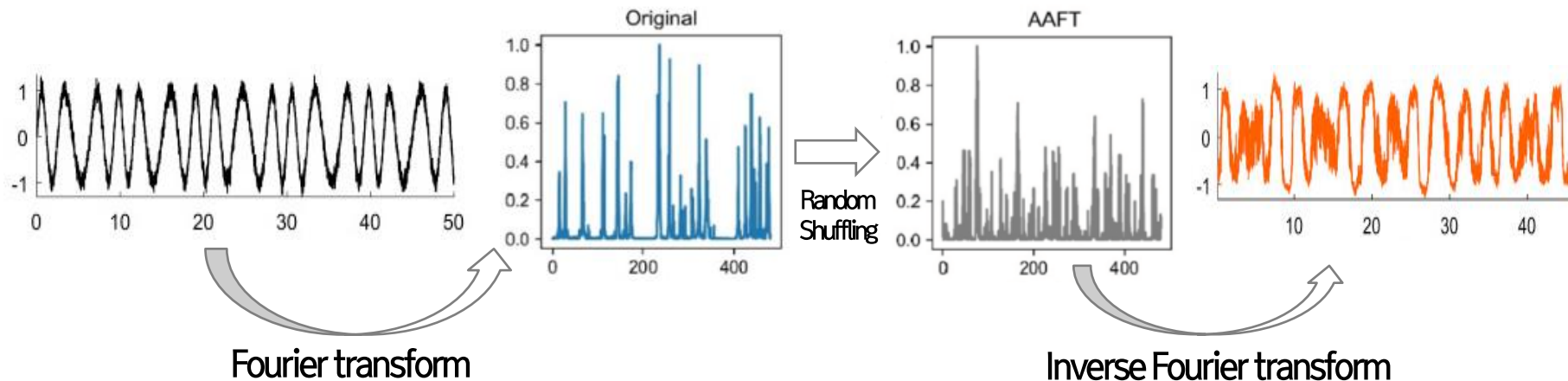
❖ Basic Approaches – Frequency domain

- Time-domain보다 현재까지의 기법 수는 적지만, 최근 연구가 활발히 진행됨
 - Amplitude & Phase Perturbations (APP)
 - Amplitude adjusted Fourier transform (AAFT)



Amplitude adjusted Fourier transform (AAFT)

- ① 기존 Signal data에 Fourier transform 적용
- ② Phase spectrum 내 무작위 Phase shuffle 적용
- ③ Inverse Fourier transform 적용 후 rank-ordering을 거쳐 augmented data 생성

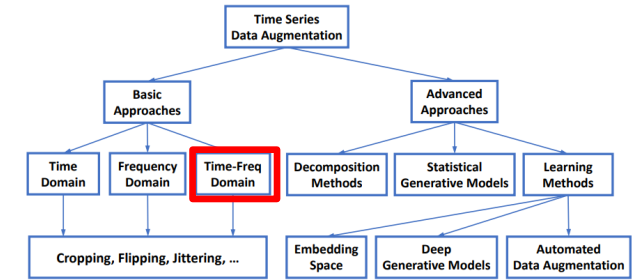


Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

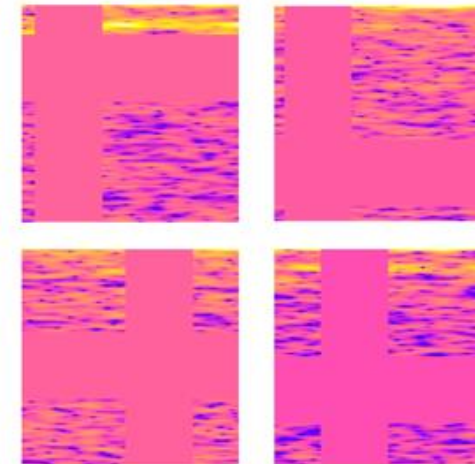
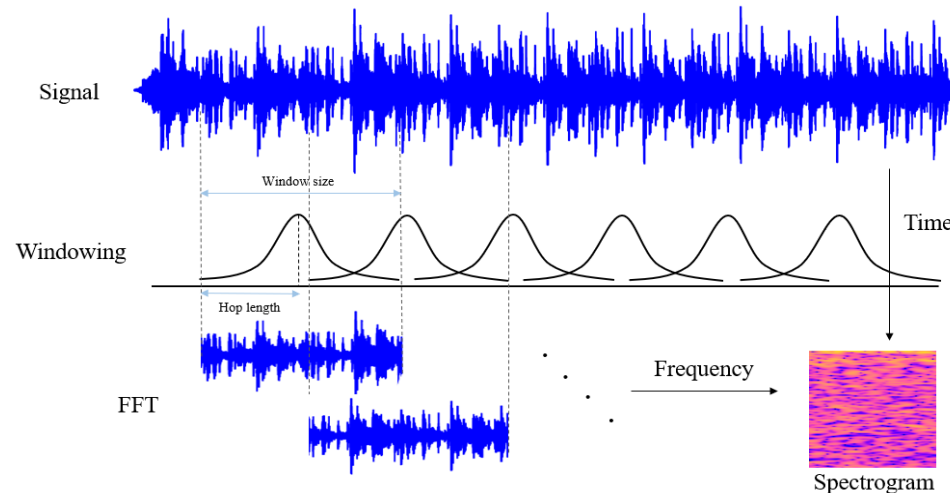
❖ Basic Approaches – Time & Frequency domain

- 단시간 푸리에 변환 (Short Term Fourier Transform; STFT)



Short Term Fourier Transform(STFT)

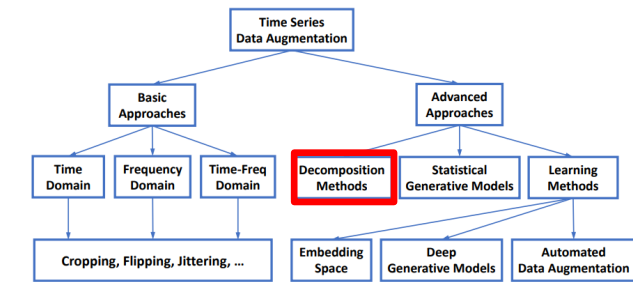
시간의 흐름에 따른 amplitude의 정보를 담고 있는 기존의 signal 데이터를 변형하는 기법
사전에 정의한 시간 단위로 windowing하여 FFT 수행
시간의 흐름에 따른 주파수 영역의 특성 변화를 파악할 수 있는 장점



SpecAugment Masking

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

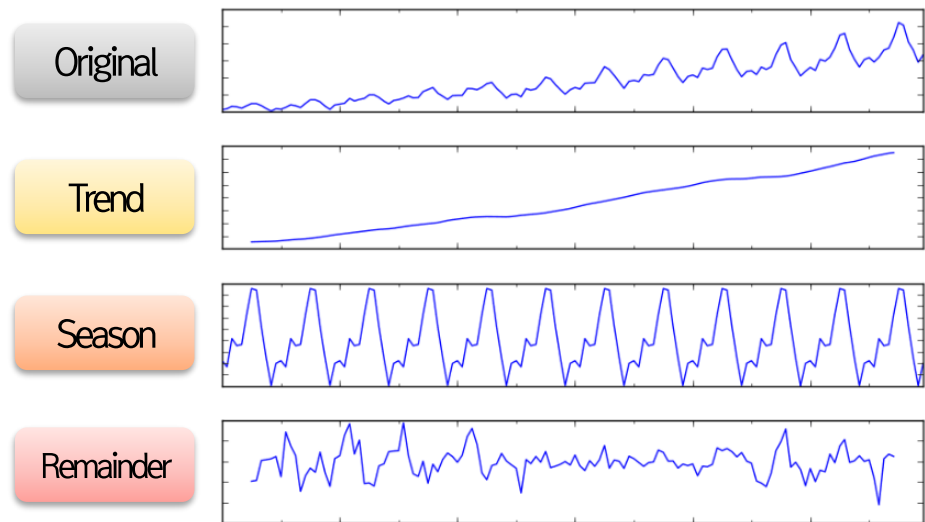


❖ Advanced Approaches – Decomposition Methods

- Seasonal and Trend decomposition using Loess (STL)

Seasonal and Trend decomposition using Loess(STL)

시계열 데이터를 Trend, Season, Remainder로 분해하는 STL 기법을 활용한 증강 기법



$$x_t = \tau_t + s_t + r_t$$

1) Deterministic Method

Trend(τ_t), Season(s_t), Remainder(r_t)의 가중치를 조절하여 데이터를 생성하는 증강 기법

2) Stochastic Method

잔차에 기반하여 구성된 확률적인 모델을 바탕으로 데이터를 생성하는 방법

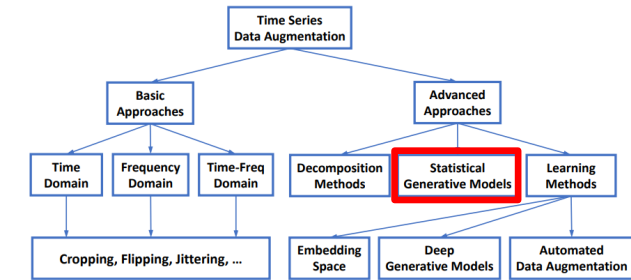
3) Bootstrapped-based Method

STL을 거치면서 생긴 잔차에 Bootstrap을 적용하여 데이터를 증강시키는 방법

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Statistical Generative Model



Statistical Generative Model

시계열 데이터의 **conditional distribution**을 반영한 데이터 증강 기법
 t 시점에 증강 기법을 통해 생성된 데이터는 이전 point의 영향을 받는다고 가정

1) Parsimonious statistical model

Mixture of Gaussian trees(MoG)

Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

2) LGT-based statistical model

Local and Global Trend(LGT)

Non-linear한 모델 및 Student T-distribution을 활용

LGT statistical model을 통해 계산된
parameters와 forecast paths의 샘플을
사용하여 데이터를 증강하는 방법

3) MAR-based statistical model

Mixture of Gaussian trees(MoG) 사용

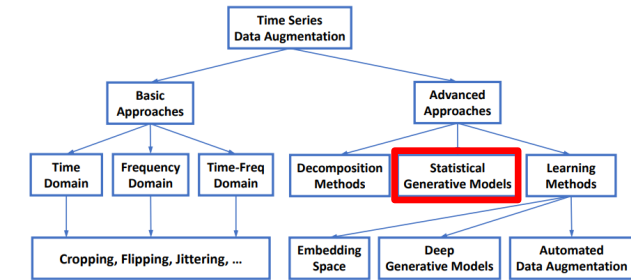
Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Statistical Generative Model



Statistical Generative Model

시계열 데이터의 **conditional distribution**을 반영한 데이터 증강 기법
 t 시점에 증강 기법을 통해 생성된 데이터는 이전 point의 영향을 받는다고 가정

1) Parsimonious statistical model

Mixture of Gaussian trees(MoG)

Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

2) LGT-based statistical model

Local and Global Trend(LGT)

Non-linear한 모델 및 Student T-distribution을 활용

LGT statistical model을 통해 계산된
parameters와 forecast paths의 샘플을
사용하여 데이터를 증강하는 방법

3) MAR-based statistical model

Mixture of Gaussian trees(MoG) 사용

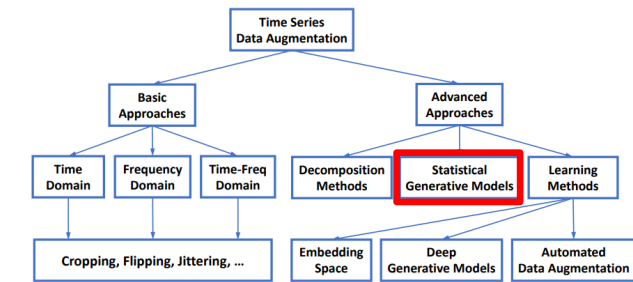
Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Statistical Generative Model



Statistical Generative Model

시계열 데이터의 **conditional distribution**을 반영한 데이터 증강 기법
 t 시점에 증강 기법을 통해 생성된 데이터는 이전 point의 영향을 받는다고 가정

1) Parsimonious statistical model

Mixture of Gaussian trees(MoG)

Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

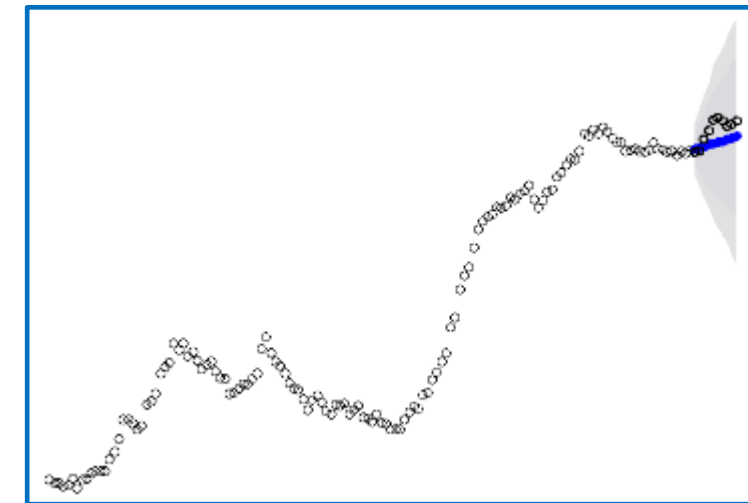
기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

2) LGT-based statistical model

Local and Global Trend(LGT)

Non-linear한 모델 및 Student T-distribution을 활용

LGT statistical model을 통해 계산된
parameters와 forecast paths의 샘플을
사용하여 데이터를 증강하는 방법

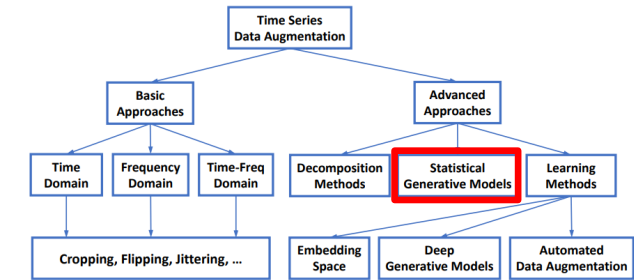


[LGT-based model example]

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Statistical Generative Model



Statistical Generative Model

시계열 데이터의 **conditional distribution**을 반영한 데이터 증강 기법
 t 시점에 증강 기법을 통해 생성된 데이터는 이전 point의 영향을 받는다고 가정

1) Parsimonious statistical model

Mixture of Gaussian trees(MoG)

Multi-modality 환경을 가정하고,
Minor class에 대한 oversampling 적용

기존 oversampling 방법과는 다르게 인접 지점간의
시계열 상관관계를 고려하는 oversampling 방법

2) LGT-based statistical model

Local and Global Trend(LGT)

Non-linear한 모델 및 Student T-distribution을 활용

LGT statistical model을 통해 계산된
parameters와 forecast paths의 샘플을
사용하여 데이터를 증강하는 방법

3) MAR-based statistical model

Mixture of Auto-regressive model(MAR)

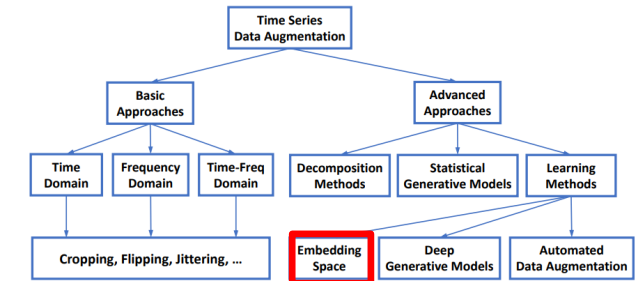
MAR 모델을 바탕으로 시계열 데이터 집합에 대한
simulation 및 statistical한 정보 파악

해당 정보를 바탕으로 생성된 데이터 중
noise가 과부여된 augmented data를 복원 가능

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Learning Methods

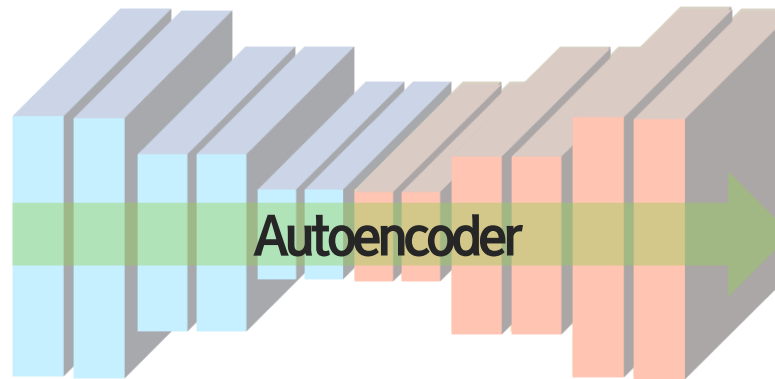


Learning Methods – Embedding Space

Raw input에 대한 augmentation보다 **embedded input에 대한 augmentation**이 효과적임을 가정
Embedded input에 대한 interpolation & extrapolation 적용

V1	V2	V3	V4	V5
0.291	0.680	0.512	0.128	1.401
0.282	Interpolation →	0.528	0.128	1.339
Interpolation →	0.294	0.551	Interpolation →	1.363
0.488	0.339	0.595	0.128	Interpolation →
0.391	0.412	0.587	0.128	1.288

[Original Time-series Data]



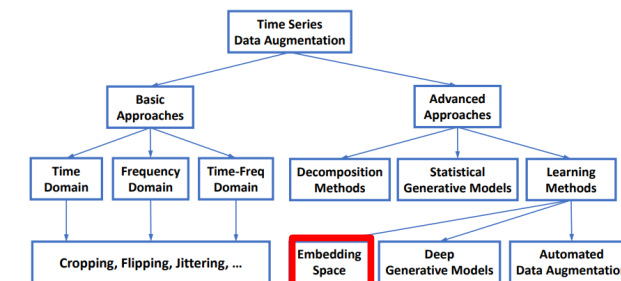
V1	V2	V3	V4	V5
0.217	0.618	0.501	0.169	1.281
0.255	0.625	0.474	0.163	1.179
0.918	0.247	0.681	0.164	1.366
0.410	0.330	0.621	0.170	1.208
0.331	0.406	0.603	0.174	1.217

[Reconstructed Time-series Data]

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Learning Methods

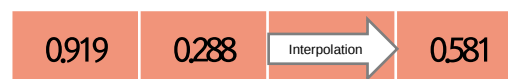
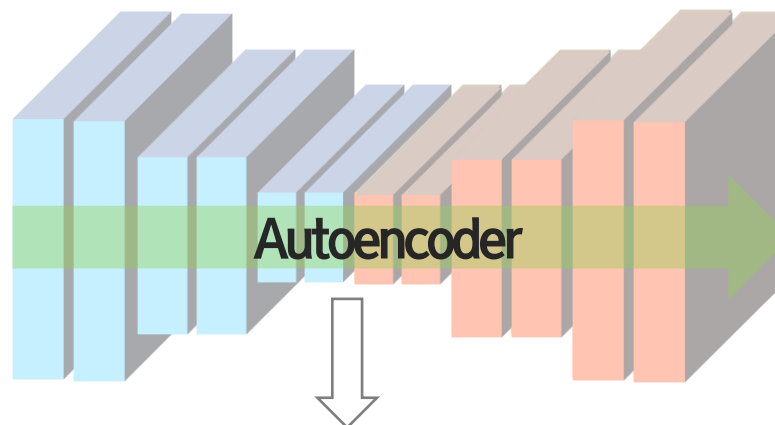


Learning Methods – Embedding Space

Raw input에 대한 augmentation보다 embedded input에 대한 augmentation이 효과적임을 주장
Embedded input에 대한 interpolation & extrapolation 적용

V1	V2	V3	V4	V5
0.291	0.680	0.512	0.128	1.401
0.282	0.661	0.528	0.128	1.339
1.271	0.294	0.551	0.129	1.363
0.488	0.339	0.595	0.128	1.291
0.391	0.412	0.587	0.128	1.288

[Original Time-series Data]



[Embedding Space]

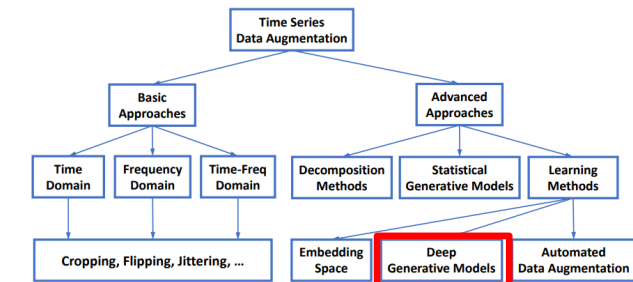
V1	V2	V3	V4	V5
0.279	0.653	0.531	0.140	1.337
0.277	0.647	0.517	0.142	1.319
1.190	0.266	0.583	0.142	1.320
0.471	0.330	0.546	0.141	1.278
0.380	0.406	0.593	0.142	1.260

[Reconstructed Time-series Data]

Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Deep Generative Models



Deep Generative Models

이미지 생성에 주로 사용되는 **Deep Generative Models(DGM)**을 시계열 데이터 생성에 적용한 사례
Recurrent GAN(RGAN), TimeGAN 두 알고리즘이 대표적

Recurrent GAN (RGAN)

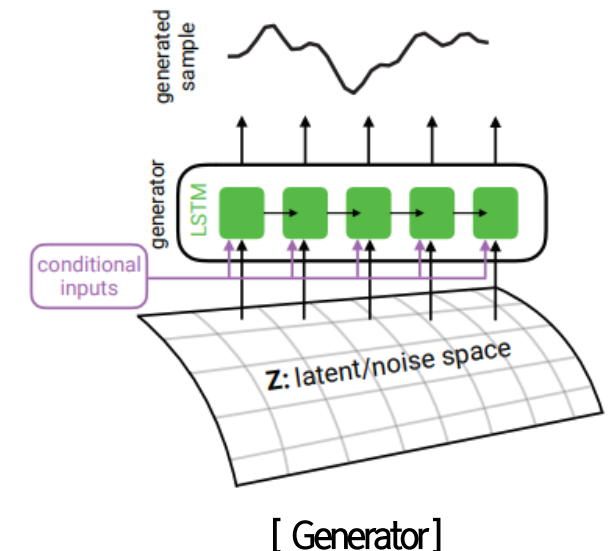
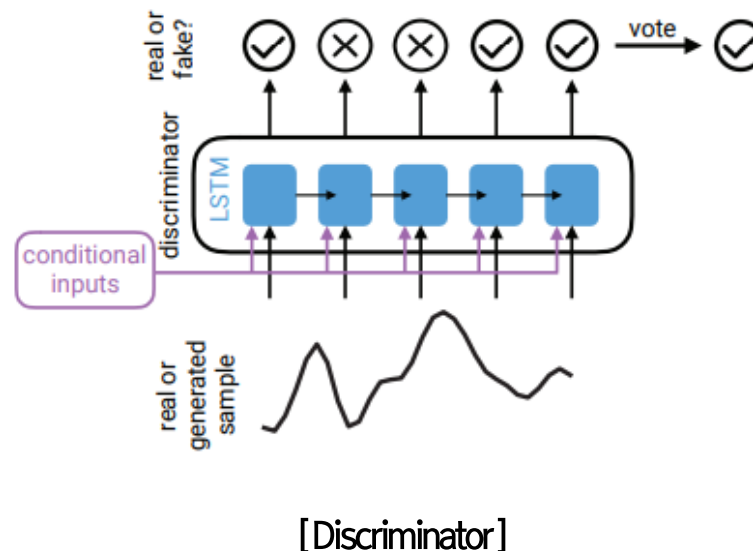
Synthetic medical data를 generation하기 위해
[Esteban et al., 2017]에서 고안된 모델

Discriminator

LSTM 모델을 통해 순차적으로 time-step마다
해당 데이터의 진위 여부를 판단

Generator

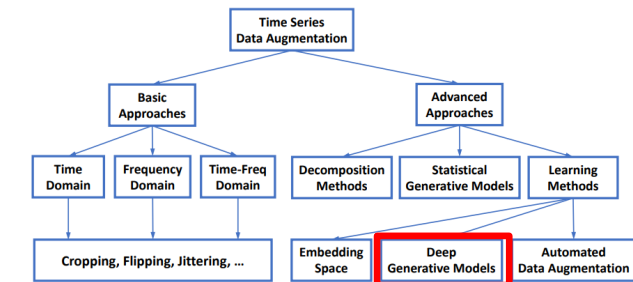
Latent/noise space로부터 실제 데이터의
Distribution을 고려하여 새로운 데이터를 만드는 생성기



Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Deep Generative Models



Deep Generative Models

이미지 생성에 주로 사용되는 Deep Generative Models(DGM)을 시계열 데이터 생성에 적용한 사례
Recurrent GAN(RGAN), TimeGAN 두 알고리즘이 대표적

TimeGAN

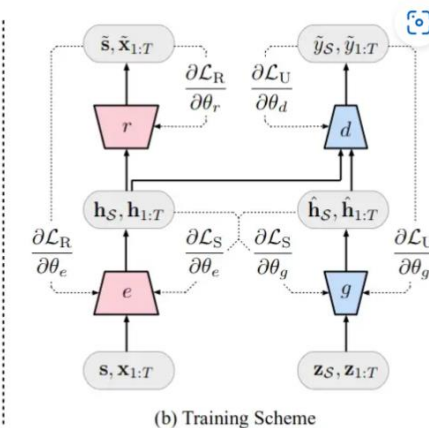
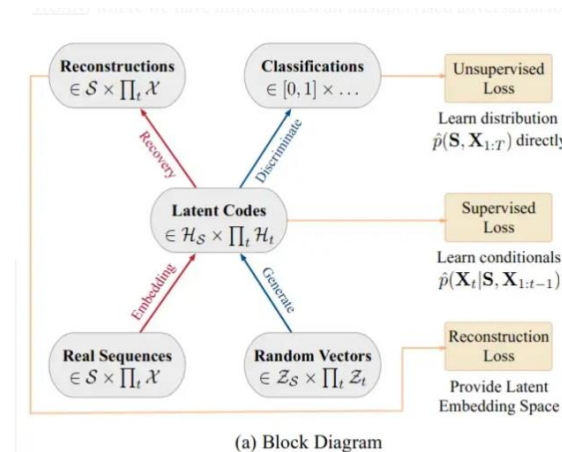
기존의 Discriminator/Generator 형태 유지

Supervised-loss

기존 데이터를 활용하여 time-conditional distribution을 고려하는 loss function 추가

Embedding network

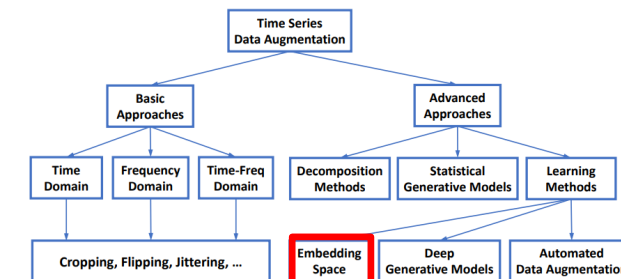
적대적 학습 공간 차원을 줄이는 것을 목표로 하는 네트워크가 추가적으로 도입



Data Augmentation for Time-series Data

Basic Approaches – Time & Frequency domain

❖ Advanced Approaches – Deep Generative Models



Deep Generative Models

종료

준비 중입니다

ACGAN : Auxiliary Classifier GANs

발표자: 신옥수

📅 2022년 2월 18일
🕒 오후 1시 ~
📍 고려대학교 신공학관 218호
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

종료

Fusion Probabilistic Models (DPM)

Score-based Generative Models and Diffu

발표자: 조한샘

📅 2022년 2월 11일
🕒 오후 1시 ~
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

(a) Block Diagram

(b) Training Scheme

Conclusion

Conclusions

How to augment your time-series data?

❖ Time-series data만을 위한 데이터 증강 기법의 필요성

- 높은 성능 향상을 가져온 이미지 데이터에 적합한 증강 기법은 다음의 이유로 시계열 데이터에 적용하기 어려움
 - 시계열 데이터 내에 존재하는 시간 종속적(Time dependency)인 특성을 해결하지 못함
 - 하고자하는 Task에 따라 증강 기법 적합성의 여부가 달라지는 문제

❖ Basic Approaches for Time-series Data Augmentation

- Time domain - Cropping, Warping, Flipping, Perturbation(w/ Ensemble), Noise injection, Label expansion
- Frequency domain - Amplitude & Phase Perturbations (APP), Amplitude adjusted Fourier Transform (AAFT)
- Time & Frequency domain - Short-Term Fourier Transform (STFT)

❖ Advanced Approaches for Time-series Data Augmentation

- Decomposition Methods, Statistical Generative Models
- Learning Methods - Embedding Space, Deep Generative Models (RGAN, TimeGAN)

감사합니다.