Lecture 02 - Contents

An overview of the main sections in this lecture.

This outline is for guidance. Navigate the slides with the left/right arrow keys.

Lecture 2:

Medical Data Preprocessing and Curation

데이터 품질이 모델 성능의 80%를 결정합니다

Ho-min Park

homin.park@ghent.ac.kr

powersimmani@gmail.com

Lecture Overview

Part 1: Clinical Text Processing Pipeline

Part 2: Medical Ontologies and Coding Systems

Part 3: Multimodal Data Integration

Part 1/3:

Clinical Text Processing Pipeline

1. De-identification Techniques
2. PHI Detection and Removal
3. Clinical Text Normalization
4. Abbreviation Expansion
5. Negation Detection
6. Temporal Expression Extraction
7. Section Segmentation

De-identification Techniques

Safe Harbor Method

18가지 식별자를 제거하는 HIPAA 표준 방법

- 이름, 주소, 날짜 등 명시된 항목 제거
- 구현이 상대적으로 간단
- 규정 준수 용이

Expert Determination

전문가가 재식별 위험을 매우 낮은 수준으로 판단

- 통계적 방법 활용
- 더 많은 데이터 활용 가능
- 전문가 검증 필요

Rule-based Pattern Matching

정규표현식을 사용한 자동 탐지

- 날짜 패턴: \d{2}/\d{2}/\d{4}
- 전화번호: \d{3}-\d{3}-\d{4}
- 빠른 처리 속도

ML-based Detection

머신러닝 기반 NER 모델 활용

- BiLSTM-CRF, BERT 모델
- 문맥 기반 탐지 가능
- F1 score 95%+ 달성

정확도 메트릭 비교

Precision
정밀도
탐지된 항목 중 실제 PHI 비율

Recall
재현율
실제 PHI 중 탐지된 비율

F1 Score
조화 평균
Precision과 Recall의 조화 평균

FP Rate
오탐률
비PHI가 PHI로 잘못 탐지

PHI Detection and Removal

1 이름	**2** 주소	**3** 날짜	**4** 전화번호	**5** 팩스번호	**6** 이메일
7 사회보장번호	**8** 의료기록번호	**9** 건강보험번호	**10** 계좌번호	**11** 면허번호	**12** 차량번호
13 기기 ID	**14** 웹 URL	**15** IP 주소	**16** 생체인식정보	**17** 사진	**18** 기타 고유식별자

하이브리드 접근법: 규칙 기반 + 머신러닝

규칙 기반 (Rule-based)

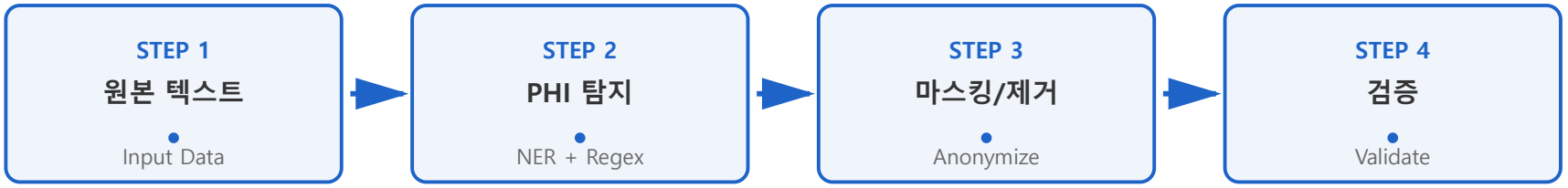
정규표현식으로 명확한 패턴 탐지

- 날짜: MM/DD/YYYY
- 전화: (123) 456-7890
- 높은 정밀도, 낮은 재현율

머신러닝 (ML-based)

NER 모델로 문맥 기반 탐지

- BiLSTM-CRF, BERT
- spaCy, MedCAT
- 높은 재현율, 문맥 이해



Clinical Text Normalization

대소문자 통일

~~CHF~~ → chf
~~Diabetes~~ → diabetes

약어 확장

~~BP~~ → blood pressure
~~Dx~~ → diagnosis

철자 교정

~~diabetis~~ → diabetes
~~hypertention~~ → hypertension

날짜/시간 표준화

~~3/15/23~~ → 2023-03-15
~~5pm~~ → 17:00

단위 변환

~~98.6°F~~ → 37°C
~~150 lbs~~ → 68 kg

특수문자 처리

~~HTN-1~~ → HTN 1
~~pt./patient~~ → patient

Abbreviation Expansion

50,000+
의료 약어 사전 항목

85%
문맥 기반 확장 정확도

약어 유형

- 일반 약어: BP → blood pressure
- 약물 약어: ASA → aspirin
- 진단 약어: MI → myocardial infarction
- 검사 약어: CBC → complete blood count

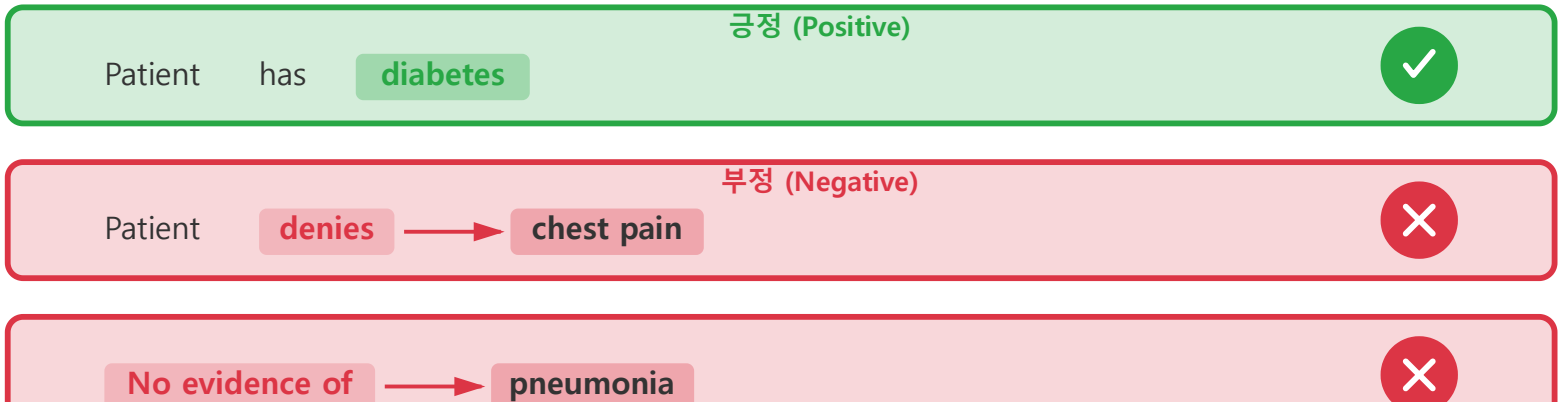
모호성 해결

- MS → multiple sclerosis vs. mitral stenosis
- RA → rheumatoid arthritis vs. right atrium
- 문맥 분석 필수
- UMLS 활용

Negation Detection

NegEx 알고리즘

부정 표현과 그 영향 범위를 결정하는 규칙 기반 알고리즘



부정 트리거

no, not, denies, without, absent, negative, rule out, free of

가능성 트리거

possible, probable, likely, suspected, questionable, consider

Temporal Expression Extraction

날짜 (Date)

- 2023-03-15
- March 15, 2023
- 03/15/2023

기간 (Duration)

- 3 weeks
- for 2 months
- since 2020

빈도 (Frequency)

- twice daily
- every 6 hours
- once a week

HeidelTime

규칙 기반 시간 표현 추출

SUTime

Stanford 시간 표현 인식기

Section Segmentation

Chief Complaint
주호소 - 환자가 내원한 주된 이유

HPI
현병력 - 현재 질병의 발병 과정

Past Medical History
과거력 - 이전 질병 및 수술 기록

Physical Exam
신체검사 - 활력징후 및 진찰 소견

Assessment
평가 - 진단 및 임상 판단

Plan
계획 - 치료 방침 및 추적 계획

경계 탐지 방법

- 헤더 키워드 매칭 (HISTORY:, ASSESSMENT:)
- 머신러닝 기반 세그멘테이션
- F1 Score: 92-96%

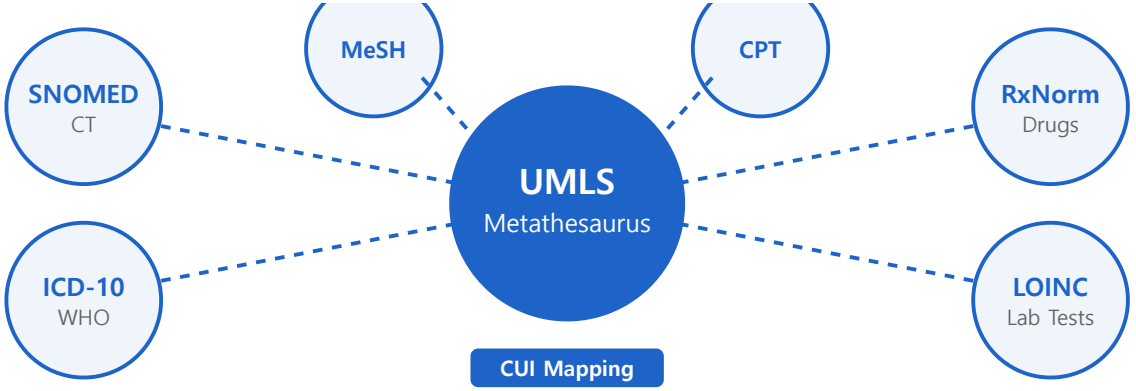
Part 2/3:

Medical Ontologies and Coding Systems

1. UMLS Metathesaurus
2. SNOMED CT Hierarchy
3. ICD-10/11 Coding
4. RxNorm Drug Normalization
5. LOINC Lab Values
6. Entity Linking Techniques

UMLS Metathesaurus

200+	380만	1400만
소스 어휘	개념 (CUI)	고유 이름



Metathesaurus

다양한 의료 용어 체계를 통합한 개념 데이터베이스

- CUI (Concept Unique Identifier)
- 동의어 및 번역
- 소스 간 매핑

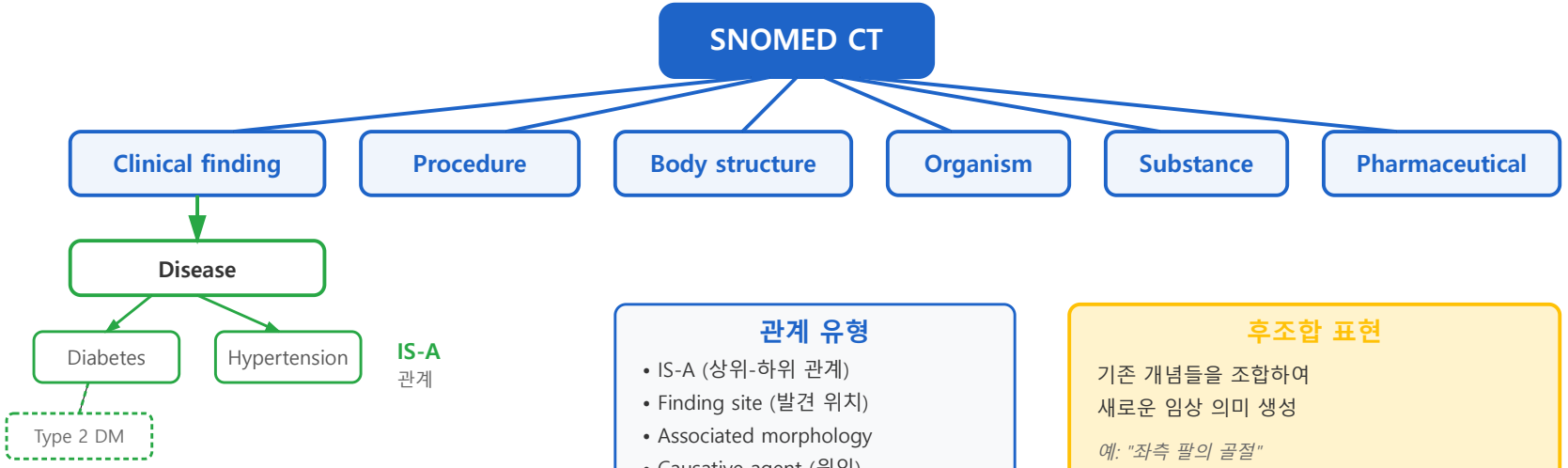
Semantic Network

135개 의미 유형과 54개 관계로 구성

- is_a 관계
- associated_with
- treats, causes 등

SNOMED CT Hierarchy

350,000+ 개념 | 19개 최상위 계층 | 150만+ 관계



관계 유형

- IS-A (상위-하위 관계)
- Finding site (발견 위치)
- Associated morphology
- Causative agent (원인)

후조합 표현

기존 개념들을 조합하여
새로운 임상 의미 생성

예: "좌측 팔의 골절"
= Fracture + Finding site: Left arm

ICD-10/11 Coding

ICD-10

- 70,000+ 코드
- 21개 챕터
- A00-Z99 범위
- 7자리 세부 코드

ICD-11

- 55,000 코드 (간소화)
- 26개 챕터
- 온라인 검색 최적화
- 확장성 개선

자동 코딩 알고리즘

- NLP 기반 임상 노트 분석
- 85%+ 정확도
- 규칙 기반 + ML 하이브리드
- BERT 기반 코딩 모델

RxNorm Drug Normalization

RxNorm 개념 모델

- **Ingredient**: 활성 성분 (aspirin)
- **Clinical Drug**: 성분 + 용량 (aspirin 81 MG)
- **Branded Drug**: 제품명 (Bayer Aspirin 81 MG)
- **RxCUI**: 고유 식별자

NDC 매핑

National Drug Code와 연결
제조사, 패키지 정보 포함

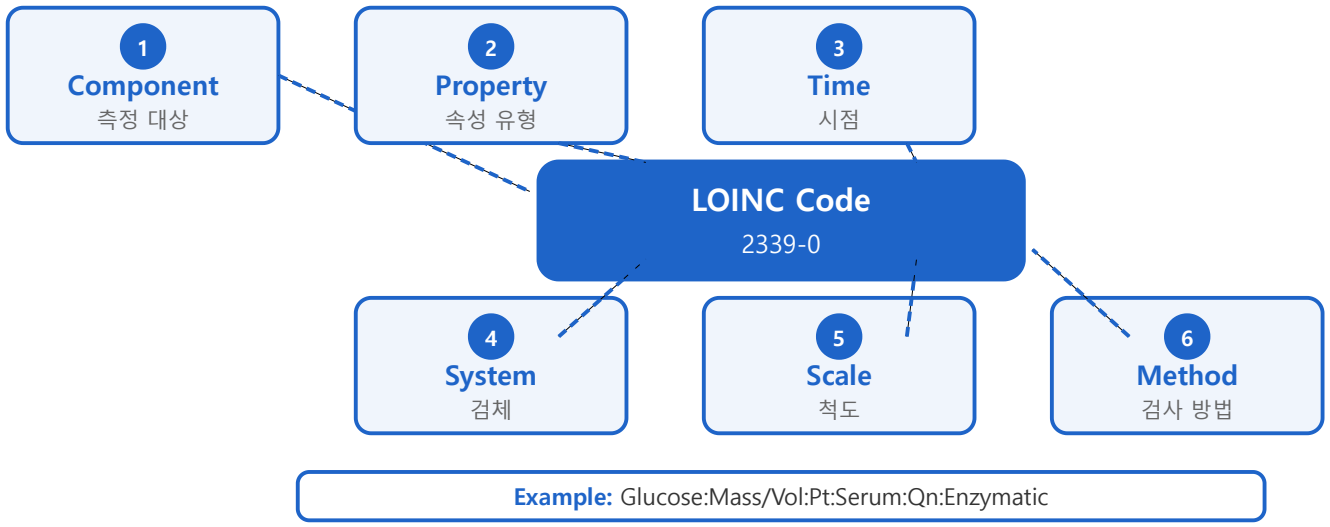
상호작용 체크

약물 간 상호작용 데이터
금기사항 확인

LOINC Lab Values

96,000+ 검사 코드 | **6개 축** 구조

LOINC 6-Part Structure



Entity Linking Techniques

문자열 매칭

- Exact matching
- Fuzzy matching
- Levenshtein distance
- Soundex, Metaphone

의미 유사도

- Word embeddings
- BERT embeddings
- Cosine similarity
- Semantic distance

문맥 기반 링킹 (앙상블)

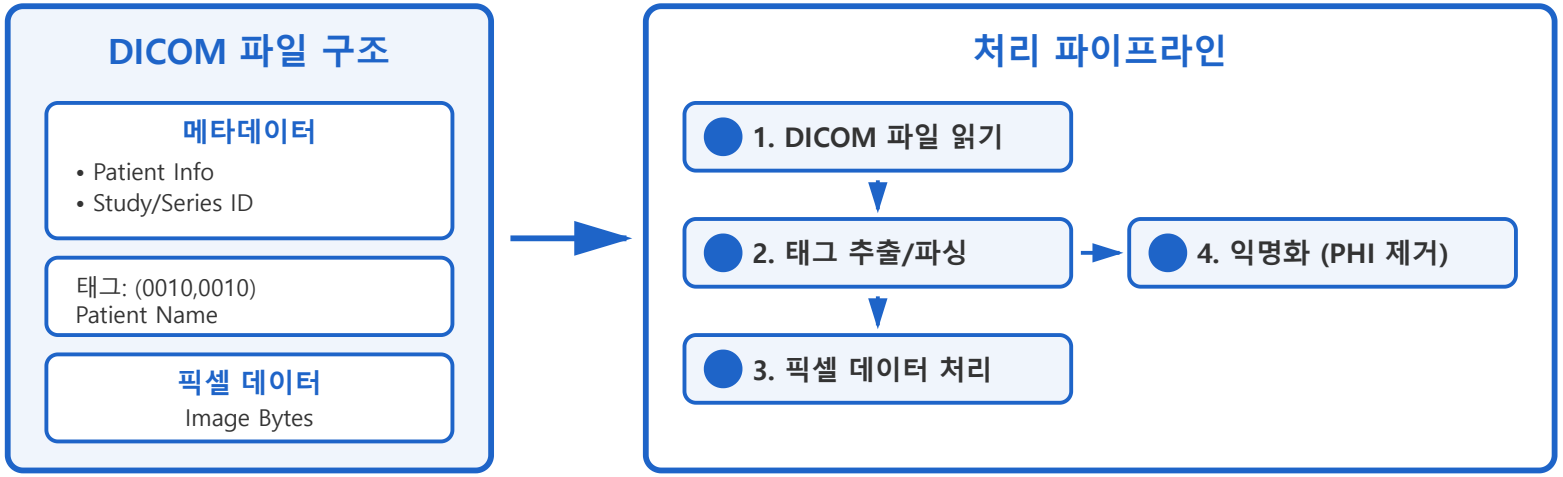
규칙 기반 + 문자열 매칭 + 의미 유사도를 결합
주변 단어와 문맥 고려
정확도 90%+ 달성 가능

Part 3/3:

Multimodal Data Integration

1. DICOM Image Handling
2. HL7 FHIR Integration
3. Waveform Signal Processing
4. Lab Value Normalization
5. Data Quality Assessment
6. Bias Detection & Mitigation

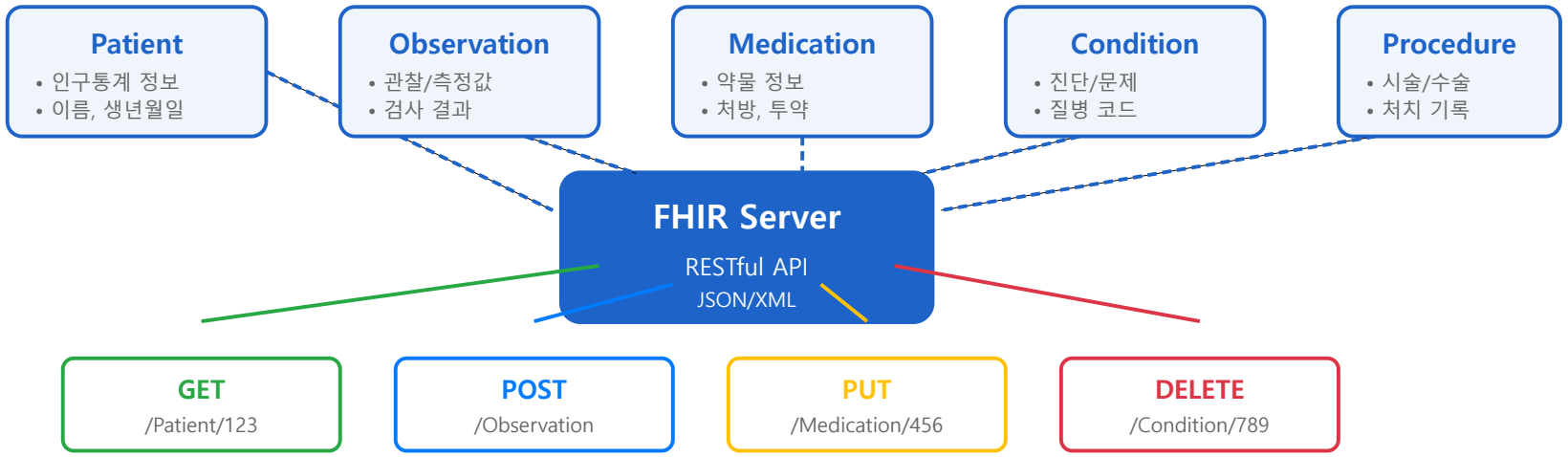
DICOM Image Handling



Python 라이브러리: pydicom

```
import pydicom
ds = pydicom.dcmread('image.dcm')
pixel_array = ds.pixel_array
patient_name = ds.PatientName
```

HL7 FHIR Integration



RESTful API 특징

- **GET** /Patient/123 - 환자 정보 조회
- **POST** /Observation - 관찰 데이터 생성
- **JSON 형식**으로 데이터 교환
- 표준 리소스 구조로 상호운용성 확보

Waveform Signal Processing

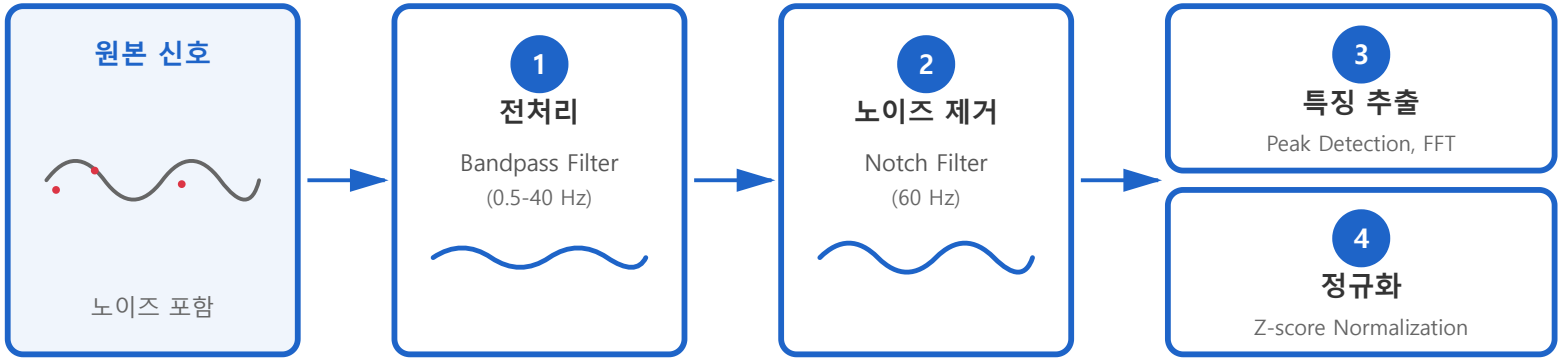
ECG (심전도)

- 샘플링: 250-500 Hz
- P, QRS, T 파형 탐지
- 부정맥 분류
- R-R 간격 계산

EEG (뇌파)

- 샘플링: 256-512 Hz
- 주파수 대역 분석
- 아티팩트 제거
- 발작 탐지

신호 처리 파이프라인



Lab Value Normalization

단위 통일

- **Glucose**: mg/dL ↔ mmol/L
- **Hemoglobin**: g/dL ↔ g/L
- **Creatinine**: mg/dL ↔ μmol/L
- SI units vs US conventional units

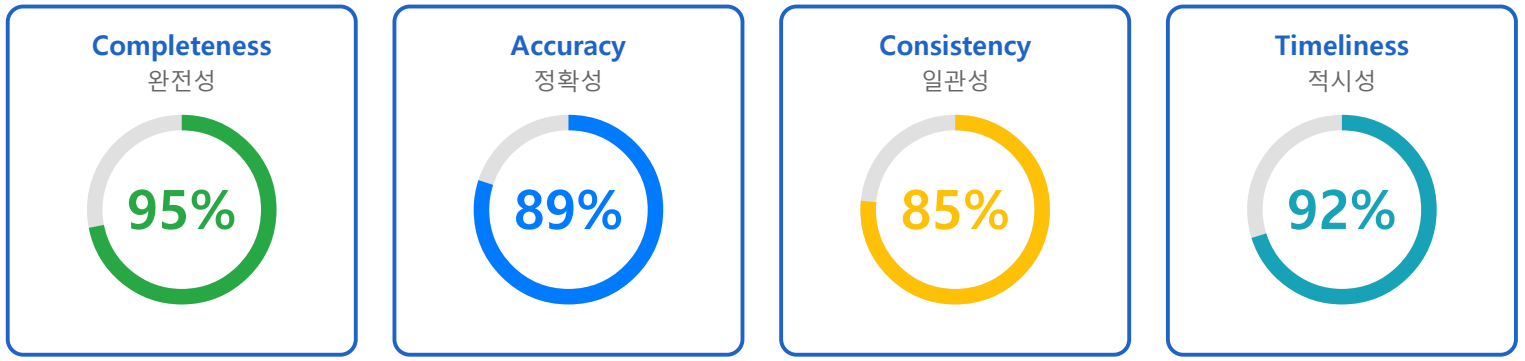
정상 범위 표준화

- 연령별 기준값
- 성별 기준값
- 임신 기준값
- Z-score 계산

이상치 탐지 및 시계열 정렬

- IQR (Interquartile Range) 방법
- 3-sigma rule
- 시간 동기화 및 결측값 처리

Data Quality Assessment



품질 점수 계산

- 필드별 완전성 측정
- 데이터 타입 검증
- 범위 검사
- 종합 품질 점수 (0-100)

개선 전략

- 자동 검증 규칙
- 이상치 플래그
- 데이터 프로파일링
- 품질 대시보드

품질 트렌드 모니터링

시계열 품질 메트릭 추적, 주기적 리포트 생성, 알림 시스템 구축

Bias Detection & Mitigation

편향 유형

- **인구통계 편향**: 특정 인종/성별 과소표집
- **선택 편향**: 비무작위 샘플링
- **측정 편향**: 측정 도구 차이
- **라벨 편향**: 주석자 편견

공정성 메트릭

- **Demographic Parity**
- **Equalized Odds**
- **Disparate Impact**
- **Individual Fairness**

완화 기법

- 재샘플링 (over/under sampling)
- 가중치 조정
- 공정성 제약 조건 추가
- 편향 완화 후처리

Missing Data Strategies

결측 패턴

- **MCAR**: 완전 무작위 결측
- **MAR**: 무작위 결측
- **MNAR**: 비무작위 결측

결측 히트맵으로 패턴 시각화

대체 방법

- **평균/중간값** 대체
- **KNN** 대체
- **MICE**: Multiple Imputation
- **Deep learning** 기반

영향 분석

대체 전후 모델 성능 비교, 민감도 분석, 결측률에 따른 영향 평가

Data Augmentation Techniques

텍스트 증강

- **역번역**: EN→KO→EN
- **동의어 치환**: WordNet
- **패러프레이징**: T5, GPT
- **임의 삽입/삭제**

합성 데이터

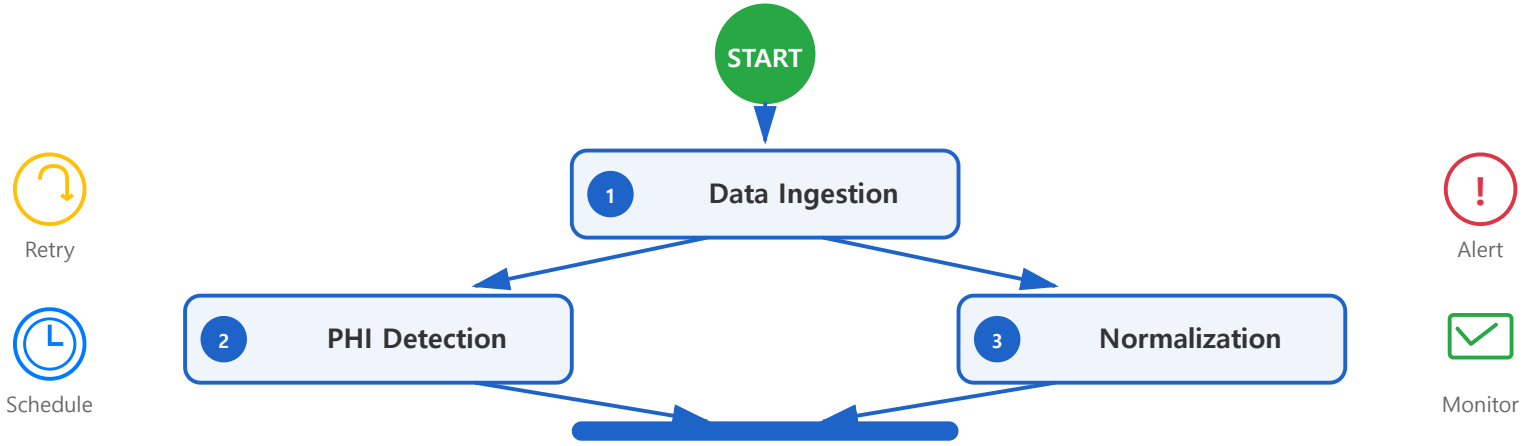
- **GPT-4** 기반 생성
- **템플릿** 기반 생성
- **SMOTE**: 소수 클래스
- **GAN**: 이미지 생성

증강 효과

데이터 다양성 증가, 과적합 방지, 소수 클래스 성능 향상, F1 score +5-15%

Pipeline Orchestration

DAG (Directed Acyclic Graph)



Apache Airflow

- DAG (Directed Acyclic Graph)
- 태스크 의존성 관리
- 스케줄링 및 모니터링
- 재시도 및 알림

Kubeflow

- Kubernetes 기반
- ML 파이프라인 전용
- 컴포넌트 재사용
- 실험 추적

주요 기능

- 의존성 자동 해결
- 에러 처리 및 재시도

- 로그 중앙 집중화
- 성능 모니터링 대시보드

Hands-on: Preprocessing with MIMIC-III

Python 코드 예제

```
import pandas as pd
from presidio_analyzer import AnalyzerEngine
from presidio_anonymizer import AnonymizerEngine

# PHI 제거
analyzer = AnalyzerEngine()
anonymizer = AnonymizerEngine()

text = "Patient John Doe, MRN 123456"
results = analyzer.analyze(text, language='en')
anonymized = anonymizer.anonymize(text, results)

# 약어 확장
abbrev_dict = {'BP': 'blood pressure', 'HR': 'heart rate'}
text = text.replace('BP', abbrev_dict['BP'])

# LOINC 매핑
loinc_code = '2339-0' # Glucose [Mass/volume] in Blood
```

- ✓ PHI 탐지 및 제거
- ✓ 텍스트 정규화
- ✓ 약어 확장
- ✓ 부정 탐지

✓ 개념 링킹

✓ 품질 검증

Best Practices Checklist

전처리 체크리스트

- ✓ PHI 완전 제거 확인
- ✓ 약어 일관성 검증
- ✓ 날짜/단위 표준화
- ✓ 부정 표현 처리
- ✓ 온톨로지 매핑
- ✓ 결측값 처리
- ✓ 이상치 탐지
- ✓ 편향 평가
- ✓ 품질 메트릭 계산
- ✓ 문서화 완료

품질 보증 프로세스

1. 샘플 검증: 무작위 100건 수동 검토
2. 자동 테스트: 단위 테스트 및 통합 테스트
3. 성능 벤치마크: 처리 속도 및 정확도
4. 문서화: 처리 단계 및 의사결정 기록

Thank you

다음 강의 예고: Lecture 3 - Advanced LLM Training

Ho-min Park

homin.park@ghent.ac.kr
powersimmani@gmail.com