Lecture 7:

Clinical Data and Electronic Health Records

- Digital health transformation
- EHR adoption rates
- Data-driven medicine

Introduction to Biomedical Datascience

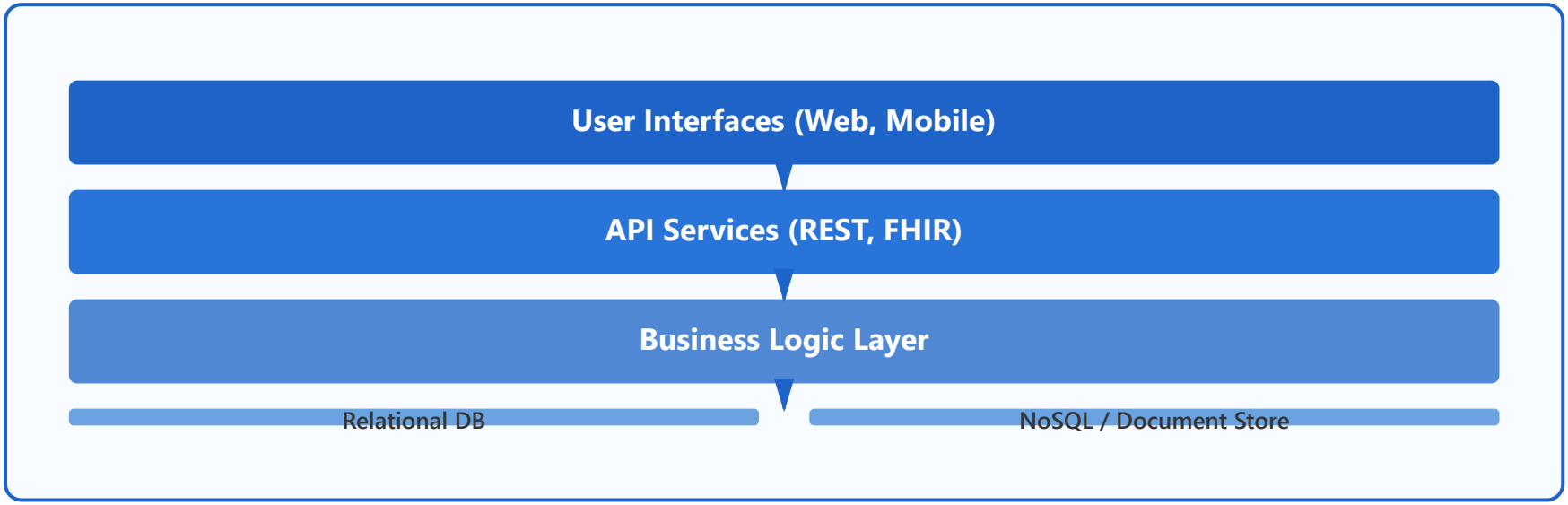
Lecture Contents

Part 1/3:

EHR Systems

- System components
- Data models
- Interoperability
- Security requirements

EHR Architecture



Database Design

- Relational databases (PostgreSQL, MySQL)
- NoSQL for unstructured data
- Data normalization strategies
- Indexing for performance

Application Layers

- Presentation layer (UI/UX)
- Business logic layer
- Data access layer
- Microservices architecture

💻 User Interfaces

- Web-based portals
- Mobile applications
- Clinical workflow integration
- Responsive design patterns

🔌 API Services

- RESTful APIs
- FHIR endpoints
- Authentication & authorization
- Rate limiting & monitoring

☁ Cloud Deployment

Modern EHRs leverage cloud infrastructure (AWS, Azure, GCP) for scalability, disaster recovery, and compliance with healthcare regulations (HIPAA, GDPR)

Data Types in EHR

Demographics

- Patient name, DOB, gender
- Address, contact information
- Insurance details
- Emergency contacts

Diagnoses/Procedures

- ICD-10 coded diagnoses
- CPT procedure codes
- Problem lists
- Surgical history

Medications

- Current medications
- Prescription history
- Allergies & adverse reactions
- Dosage and frequency

Laboratory Results

- Blood tests, imaging
- Pathology reports
- Vital signs
- LOINC coded values

Clinical Notes

- Progress notes
- Consultation reports

- Discharge summaries
- Nursing documentation

Structured vs Unstructured Data

Structured Data

- Predefined fields & formats
- Easily queryable
- Standardized codes (ICD, LOINC)
- Direct database storage
- Machine-readable

Unstructured Data

- Free text clinical notes
- Medical images (X-ray, MRI)
- Scanned documents
- Voice recordings
- Requires NLP for extraction

🔀 Hybrid Documents

Many clinical documents combine structured fields (dates, vital signs) with unstructured narratives (clinical impressions)

HL7 and FHIR Standards

HL7 v2 Messages

- Pipe-delimited format
- ADT, ORM, ORU message types
- Widely adopted legacy standard
- Complex parsing required

FHIR Resources

- JSON/XML formats
- Patient, Observation, Medication
- Modern web-based standard
- Easy to implement

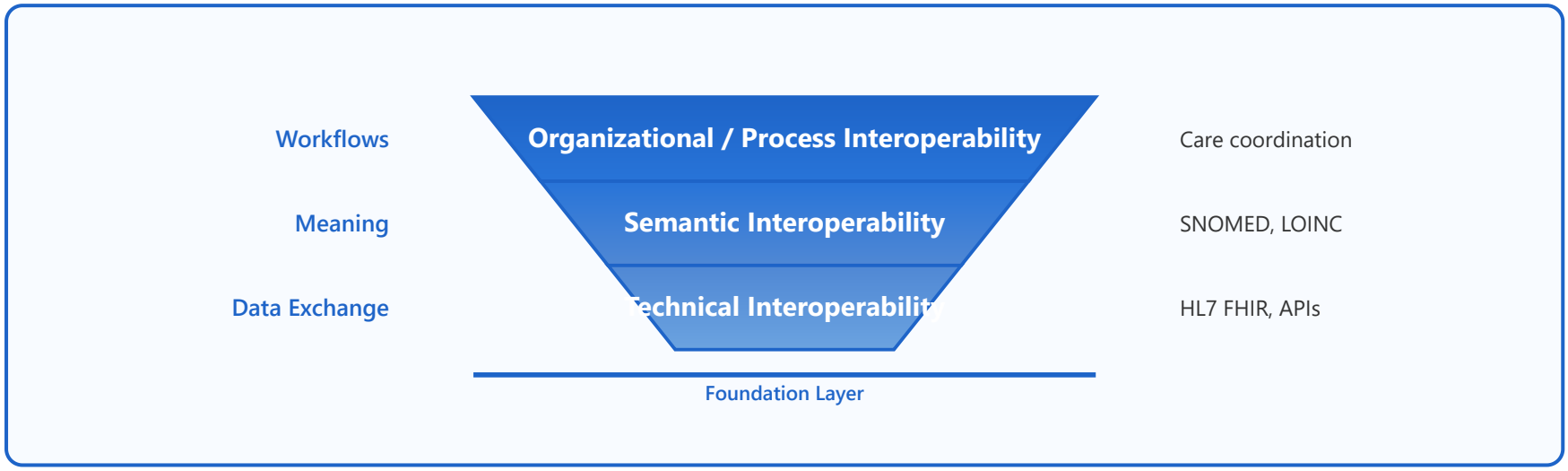
RESTful APIs

- HTTP GET, POST, PUT, DELETE
- Resource-based URLs
- OAuth 2.0 authentication
- SMART on FHIR apps

Implementation Guides

- US Core profiles
- Argonaut specifications
- Country-specific extensions
- Validation tools

Interoperability



Technical Standards

- HL7 FHIR
- Direct messaging
- APIs and web services
- Transport protocols (HTTPS, SFTP)

Semantic Standards

- Common terminologies (SNOMED, LOINC)
- Value set harmonization
- Concept mapping
- Unified Code Management

Process Interoperability

- Clinical workflows

Health Information Exchange (HIE)

- Regional/national networks

- Care coordination protocols
- Consent management
- Data governance policies

- Query-based vs push
- Patient matching algorithms
- Blockchain potential for trust

Data Warehousing for EHR

ETL Processes

- Extract from operational systems
- Transform & clean data
- Load into warehouse
- Incremental updates

Data Marts

- Disease-specific repositories
- Quality improvement data
- Research cohorts
- Departmental analytics

Star Schema

- Fact tables (encounters, labs)

Real-time vs Batch

- Batch: overnight processing

- Dimension tables (patient, time)
- Optimized for queries
- Aggregate calculations

- Real-time: streaming analytics
- Near real-time: micro-batching
- Trade-offs in complexity

Part 2/3: Clinical Coding

Clinical Coding

- Terminology systems
- Ontology relationships
- Mapping challenges
- Use cases

ICD-10 Coding

Code Structure

- 3-7 alphanumeric characters
- Chapter (A-Z)
- Body system/condition
- Laterality & severity

Diagnosis Coding

- E11.9 - Type 2 diabetes
- I10 - Essential hypertension
- J45.909 - Asthma, unspecified
- Combination codes available

Procedure Coding (ICD-10-PCS)

- 7-character codes
- Used for inpatient procedures
- Section-Body System-Root Operation
- Detailed anatomical specificity

Coding Guidelines

- Principal vs secondary diagnoses
- Specificity requirements
- Excludes1 vs Excludes2
- Use additional code notes

CPT Codes

Procedure Classification

- 5-digit numeric codes
- Category I: Common procedures
- Category II: Performance measures
- Category III: Emerging technology

E&M Codes

- Evaluation and Management
- 99201-99499 range
- Office visits, consultations
- Level based on complexity

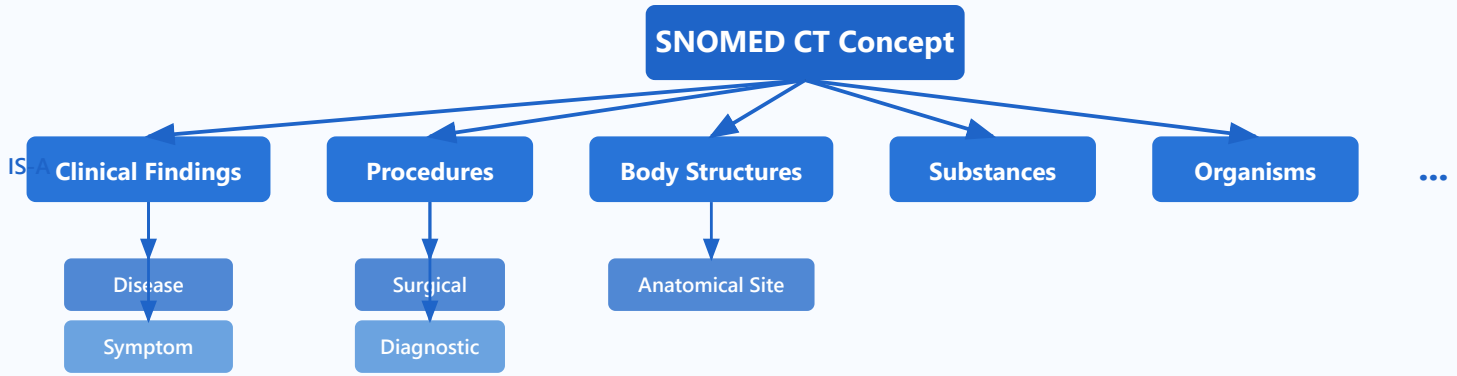
Modifier Usage

- -25: Significant separate E&M
- -59: Distinct procedural service
- -76: Repeat procedure
- Bilateral, multiple procedures

RVU Values & Billing

- Relative Value Units
- Work, practice expense, malpractice
- Medicare reimbursement formula
- Commercial payer variations

SNOMED CT



🧩 Concept Model

- Concepts, descriptions, relationships
- Unique concept IDs (SCTID)
- Fully specified names
- Synonyms and translations

📊 Hierarchies

- Clinical findings
- Procedures
- Body structures
- Substances
- IS-A relationships

🔗 Relationships

🎯 Post-coordination

- Finding site
- Associated morphology
- Causative agent
- Procedure site
- Compositional grammar

- Combine multiple concepts
- Express complex clinical meanings
- Example: 'Fracture of left femur'
- International adoption by 40+ countries

LOINC for Lab Tests

Test Identification

- Laboratory observations
- Clinical measures
- Survey instruments
- Unique numeric codes

Six-part Structure

- Component (analyte)
- Property (e.g., mass, volume)
- Timing (point in time)
- System (specimen type)
- Scale (quantitative, ordinal)
- Method (optional)

Panel Organization

- Grouping related tests
- Complete blood count (CBC)
- Basic metabolic panel (BMP)
- Lipid panel

Units Harmonization

- UCUM units of measure
- Conversion factors
- Reference ranges
- Mapping local lab codes

RxNorm for Medications

Drug Concepts

- Normalized drug names
- Ingredient + strength + dose form
- Links to other terminologies
- Unique RxNorm CUI

Hierarchy Levels

- Ingredient (e.g., Metformin)
- Clinical drug (Metformin 500 MG)
- Branded drug (Glucophage 500 MG)
- Drug pack (combination products)

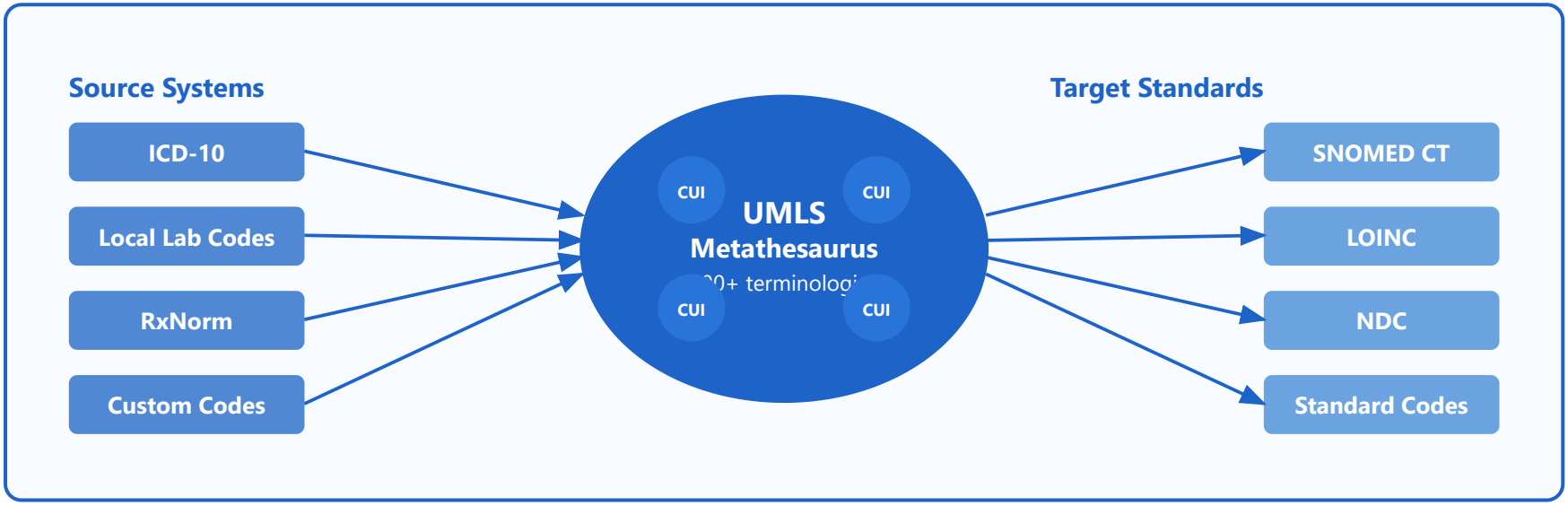
Dose Forms

- Oral tablet
- Injectable solution
- Transdermal patch
- Inhalation powder
- Topical cream

Brand vs Generic

- Generic ingredient mapping
- Brand name relationships
- NDC code linkage
- Drug interactions database

Ontology Mapping



Crosswalk Creation

- ICD-10 to SNOMED CT
- LOINC to local lab codes
- RxNorm to NDC
- Manual and automated approaches

Automated Mapping

- String similarity algorithms
- Lexical matching
- Machine learning classifiers
- Natural language processing

✓ Validation Methods

UMLS Metathesaurus

- Expert review
- Dual coding
- Inter-rater reliability
- Continuous quality improvement

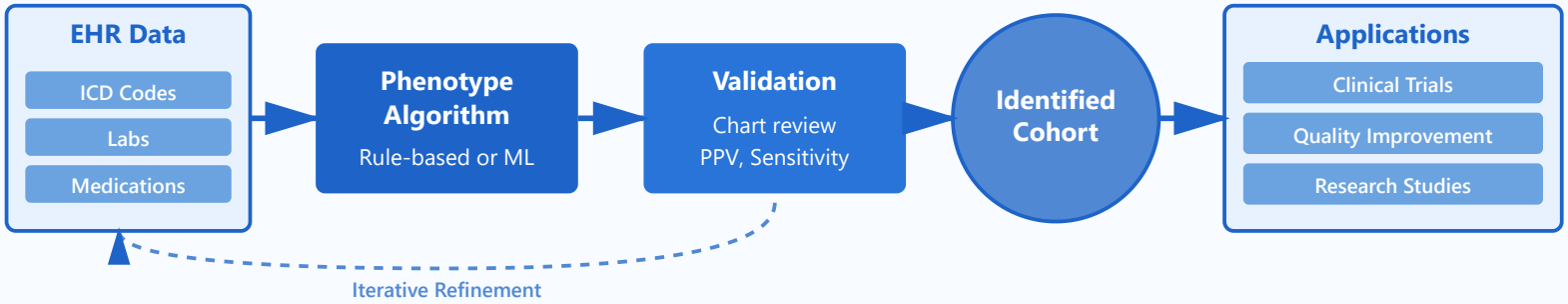
- Unified Medical Language System
- Integrates 200+ terminologies
- Concept Unique Identifiers (CUI)
- Relationship mappings across systems

Part 3/3: Data Analytics

Clinical Coding

- Phenotyping algorithms
- Predictive models
- Quality measures
- Population health

Clinical Phenotyping



📋 Computable Phenotypes

- Standardized disease definitions
- ICD codes + labs + meds
- Temporal logic criteria
- Inclusion/exclusion rules

📐 Rule-Based Methods

- Boolean logic (AND, OR, NOT)
- Diagnosis code combinations
- Lab value thresholds
- Medication orders

Machine Learning Approaches

- Supervised classification

✓ Validation Strategies

- Chart review (gold standard)

- Feature engineering from EHR
- Random forests, deep learning
- Semi-supervised learning

- PPV, NPV, sensitivity, specificity
- Cross-institutional validation
- Phenotype libraries (PheKB, eMERGE)

Cohort Identification

Inclusion Criteria

- Age range, gender
- Diagnosis codes
- Procedure history
- Medication exposures

Exclusion Criteria

- Comorbidities
- Prior treatments
- Missing data patterns
- Follow-up requirements

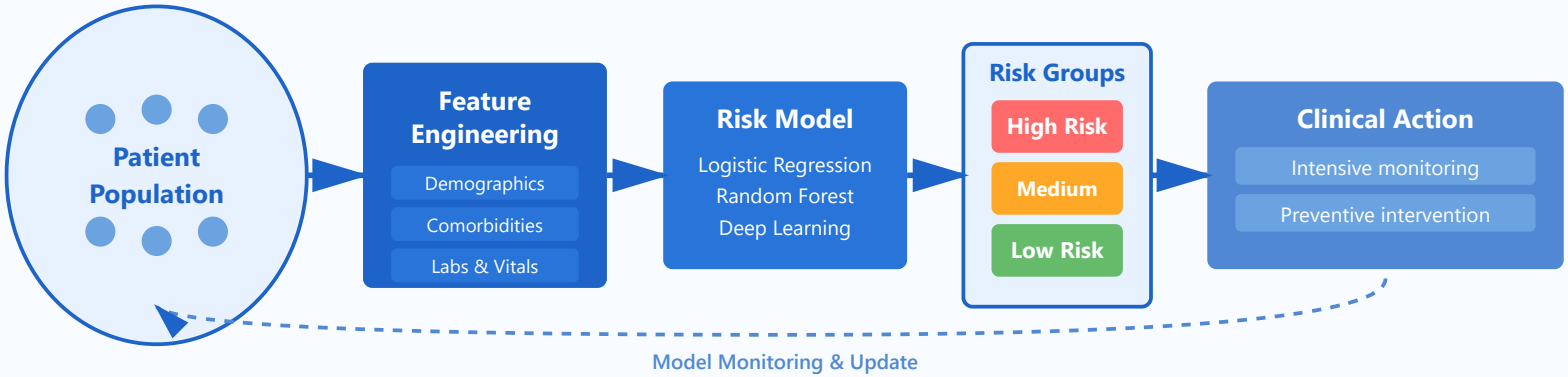
Temporal Constraints

- Index date definition
- Washout periods
- Follow-up windows
- Event ordering

Query Optimization

- Database indexing
- Efficient SQL queries
- Cohort definition tools (ATLAS)
- Sample size estimation

Risk Stratification



📊 Clinical Risk Scores

- CHADS-VASc (stroke risk)
- MELD (liver disease)
- GRACE (cardiac events)
- Point-based scoring systems

Model Development

- Logistic regression
- Cox proportional hazards
- Gradient boosting machines
- Neural networks

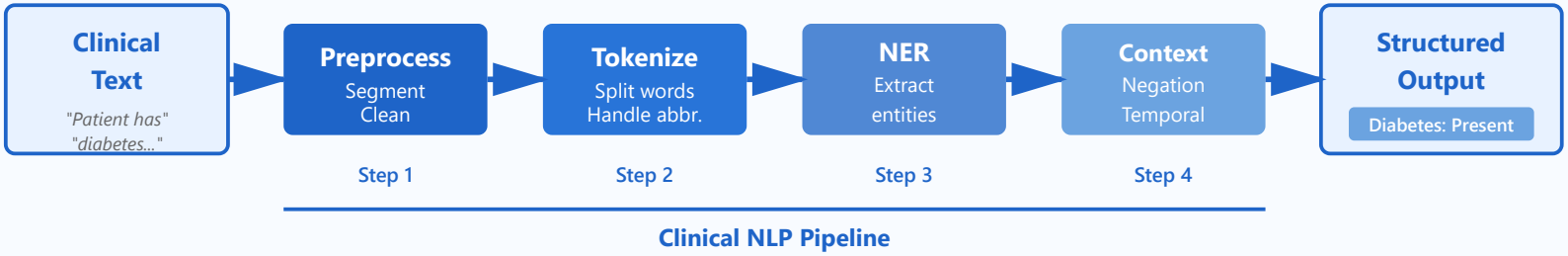
Feature Engineering

📈 Calibration & Implementation

- Aggregating encounter data
- Temporal patterns
- Medication burden scores
- Comorbidity indices (Charlson, Elixhauser)

- Calibration plots
- Decision curve analysis
- Integration into EHR alerts
- Continuous model monitoring

Clinical NLP Basics



Text Preprocessing

- Sentence segmentation
- Lowercasing, punctuation removal
- Handling abbreviations
- PHI removal

Tokenization

- Word-level tokens
- Subword tokenization (BPE)
- Clinical-specific tokenizers
- Handling medical jargon

Named Entity Recognition

- Diseases, symptoms

Negation & Section Detection

- NegEx, ConText algorithms

- Medications, dosages
- Anatomical sites
- Procedures

- Identifying negated findings
- Section headers (HPI, ROS, A&P)
- Temporal expressions

Named Entity Recognition (NER)

Medical Entities

- Diseases & conditions
- Drugs & treatments
- Signs & symptoms
- Lab tests & values
- Anatomical structures

Rule-Based Systems

- Dictionary lookups (UMLS)
- Regular expressions
- Pattern matching
- Fast but limited coverage

Machine Learning Models

- CRF, SVM models
- LSTM, BiLSTM
- Contextual embeddings
- Better generalization

Deep Learning & Hybrid

- BioBERT, ClinicalBERT
- Transfer learning
- Combining rules + ML
- State-of-the-art performance

Temporal Reasoning

Time Expression Extraction

- Absolute dates (Jan 1, 2024)
- Relative dates (3 days ago)
- Durations (for 2 weeks)
- Frequencies (twice daily)

Event Ordering

- BEFORE, AFTER, OVERLAP
- Medication start/stop
- Surgery dates
- Symptom onset timing

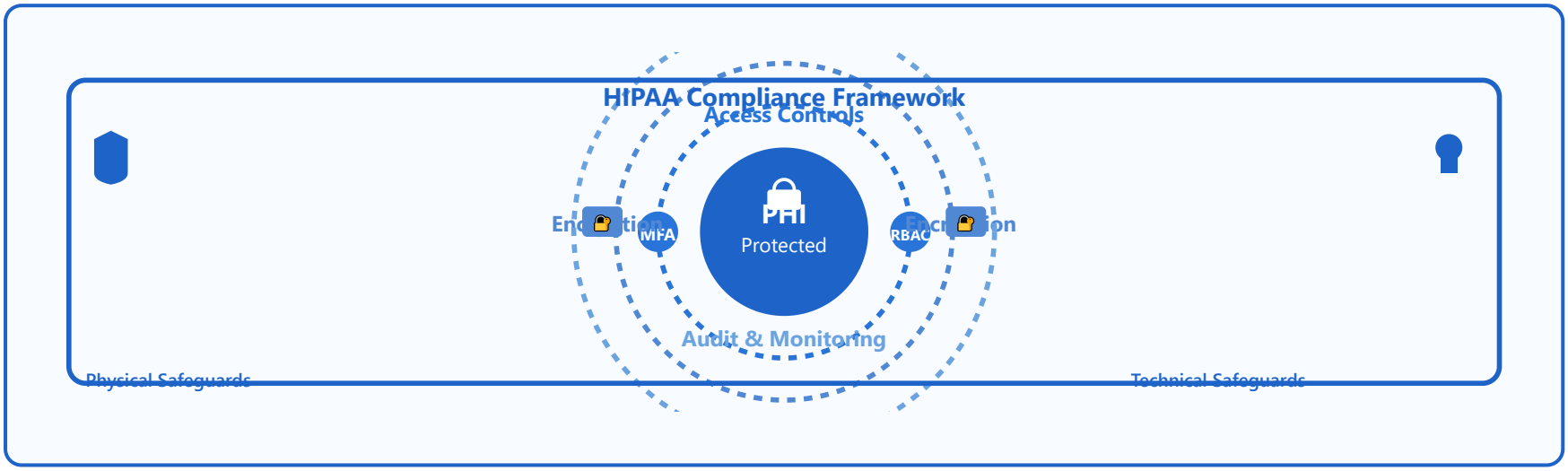
Timeline Construction

- Patient journey visualization
- Merging multi-source data
- Handling conflicts
- Uncertainty representation

Clinical Applications

- Disease progression tracking
- Treatment efficacy evaluation
- Adverse event detection
- Longitudinal phenotyping

Privacy and HIPAA



PHI Definition

- 18 identifiers under HIPAA
- Names, addresses, dates
- Medical record numbers
- Biometric identifiers

Minimum Necessary Rule

- Access only what's needed
- Role-based permissions
- Need-to-know principle
- Limit data sharing

Access Controls

- User authentication (MFA)

Breach Notification

- Report within 60 days

- Authorization levels
- Audit logs
- Encryption at rest & in transit

- Notify affected individuals
- Inform HHS if >500 patients
- Penalties for non-compliance

De-identification

Safe Harbor Method

- Remove 18 HIPAA identifiers
- Dates shifted or generalized
- Ages >89 aggregated
- Geographic areas >20,000

Expert Determination

- Statistical/scientific analysis
- Very small re-identification risk
- Certified expert assessment
- More data retained

Automated Tools

- NER for PHI detection
- Date shifting algorithms
- Scrubber software (Philter, BoB)
- Validation required

Re-identification Risk & Synthetic Data

- K-anonymity, l-diversity
- Differential privacy
- Synthetic data generation (GANs)
- Trade-off: utility vs privacy

Real-World Evidence (RWE)

Randomized Controlled Trials

- Gold standard for efficacy
- Strict inclusion criteria
- Controlled environment
- Expensive & time-consuming
- Limited generalizability

Real-World Evidence

- Effectiveness in practice
- Diverse patient populations
- Natural clinical settings
- Lower cost, faster
- Confounding & bias challenges

Regulatory Acceptance

FDA increasingly accepts RWE for drug approvals, label expansions, and post-market surveillance. Key: rigorous study design and bias mitigation.

Clinical Trials Data

EDC Systems

- Electronic Data Capture
- eCRFs (electronic case report forms)
- Real-time data validation
- Audit trails

CDISC Standards

- SDTM (Study Data Tabulation)
- ADaM (Analysis Data Model)
- CDASH (data collection)
- Regulatory submissions

Data Monitoring

- Data Safety Monitoring Boards
- Interim analyses
- Stopping rules
- Adverse event tracking

EHR Integration

- Recruitment from EHR cohorts
- Automated eligibility screening
- Pulling baseline data
- FHIR for trial data exchange

Hands-on: OMOP CDM

Data Model Overview

- Common Data Model by OHDSI
- Standardized tables & vocabularies
- Observation period, Visit, Condition
- Drug exposure, Measurement, Procedure

ETL Implementation

- Map source codes to standard concepts
- Transform dates to OMOP format
- Preserve provenance
- White Rabbit, Rabbit-in-a-Hat tools

OHDSI Tools

- ATLAS: Cohort definition & analytics
- ACHILLES: Data quality reports
- PatientLevelPrediction: ML models
- CohortMethod: Causal inference

Example Analysis

- Define diabetes cohort (Type 2)
- Characterize baseline demographics
- Predict HbA1c control
- Compare metformin vs sulfonylureas

Hands-on: Clinical NLP with Python

spaCy Medical (scispaCy)

- pip install scispacy
- en_core_sci_sm model
- NER for biomedical entities
- Abbreviation detection

MedCAT Tutorial

- Medical Concept Annotation Tool
- Unsupervised learning from UMLS
- Link entities to CUIs
- Negation & context detection

BioBERT Usage

- Pretrained on PubMed + PMC
- Fine-tune on clinical notes
- NER, relation extraction
- Hugging Face transformers

Pipeline Creation

- Load clinical notes
- Preprocess & tokenize
- Extract entities
- Evaluate with precision, recall, F1

Thank You!
Future Directions in Clinical Informatics

AI in Healthcare

- Diagnostic assistance
- Drug discovery
- Personalized treatment
- Ambient clinical documentation

Precision Medicine

- Genomics integration
- Multi-omics data
- Pharmacogenomics
- Targeted therapies

Policy & Ethics

- Algorithmic bias
- Health equity
- Data governance
- International collaboration

Career Opportunities

- Clinical data scientist
- Health informatics researcher
- Bioinformatics engineer
- Healthcare AI developer