# Clinical Trial Recommendation System using NLP

Semantic Matching of Patient Profiles to Trial Eligibility Criteria

# Problem Statement

- Patient recruitment is a major bottleneck in clinical research
- Eligibility criteria are written in unstructured natural language
- Manual trial matching is time-consuming and error-prone
- Need for automated, explainable trial recommendation systems

# Project Objectives

- Build an NLP-based system to match patients to relevant trials
- Apply eligibility constraints like age and sex
- Provide explainable and ranked trial recommendations
- Enable decision-support for clinicians and coordinators

# Dataset Overview

- Publicly available clinical trials data from Kaggle
- Unstructured inclusion and exclusion criteria
- Synthetic patient profiles for privacy compliance
- ~3,500 trials covering COVID-19 studies

# Sample Synthetic Patient Profiles

- Patient 1: Age 45, Male, Condition: Hypertension
- Patient 2: Age 30, Female, Condition: Diabetes
- Patient 3: Age 60, Male, Condition: COVID-19 positive
- Used for testing model recommendations

# Eligibility Criteria Examples

- Inclusion: Age 18-65, PCR positive for COVID-19
- Exclusion: Pregnant, history of cardiovascular disease
- Unstructured text requires NLP for semantic understanding
- Variation in terminology across trials

# Data Wrangling & Preprocessing

- Cleaned and normalized eligibility text
- Separated inclusion and exclusion conditions
- Standardized age and sex constraints
- Generated synthetic patient profiles programmatically

# Preprocessing Workflow

- Raw eligibility text -> Tokenization -> Stopword removal -> Embedding generation
- Inclusion/Exclusion parsing -> Eligibility filters applied
- Prepared data for baseline and semantic models
- Ensured reproducibility and modular pipelines

# Modeling Approach

- Baseline: TF-IDF keyword-based retrieval
- Semantic Model: Sentence-BERT embeddings
- Similarity-based trial ranking
- Eligibility rule filtering applied post-ranking

# TF-IDF Baseline Model

- Convert eligibility text to TF-IDF vectors
- Compute cosine similarity with patient profile
- Rank trials based on similarity score
- Apply age/sex filters for eligibility

# Sentence-BERT Semantic Model

- Convert trial eligibility and patient profiles to embeddings
- Compute semantic similarity
- Rank trials based on embedding similarity
- Apply eligibility constraints and explainability annotations

# Evaluation Strategy

- Used Precision@K to evaluate retrieval quality
- Compared TF-IDF baseline vs semantic model
- Weak supervision due to lack of labeled data
- Visualized performance differences with plots

# Precision@K Comparison

- TF-IDF Precision@5: 0.8
- Sentence-BERT Precision@5: 1.0
- Semantic model consistently outperformed baseline
- Improved trial relevance and ranking stability

# Top Recommendations Example

- Patient: Age 45, Male, COVID-19 positive
- Top Trial 1: NCT001, COVID-19 Vaccine Study, Similarity Score: 0.95
- Top Trial 2: NCT002, Antiviral Therapy Trial, Similarity Score: 0.92
- Top Trial 3: NCT003, Hospitalization Reduction Study, Similarity Score: 0.91

# Explainable Recommendations

- Highlighted matched conditions and keywords
- Displayed eligibility alignment (age, sex)
- Rationale for trial suggestions provided
- Enhanced trust for clinicians and coordinators

# Key Results & Insights

- Semantic embeddings improved trial retrieval over TF-IDF
- Eligibility rules ensured compliance with patient constraints
- Precision@K metrics confirmed retrieval accuracy
- Visualizations aided intuitive understanding and communication

# Conclusion & Future Work

- Demonstrated semantic NLP can enhance patient-trial matching
- Future work: integrate richer eligibility parsing, medical ontologies
- Potential integration with EHR systems for real-world deployment
- Dashboard and interactive visualization for clinicians