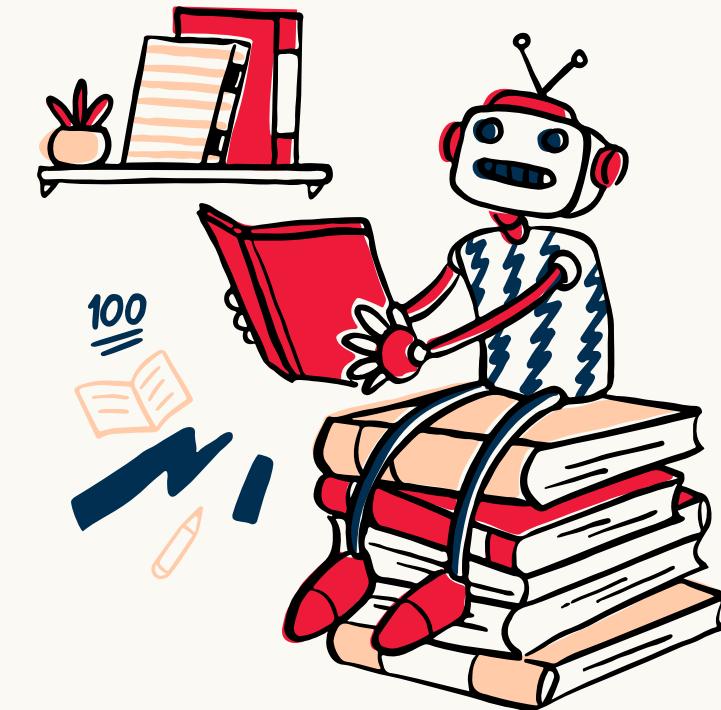


APPLIED DATA SCIENCE PROJECT

# Patient Preference Studies Classification System



UNIVERSITÀ  
DI TORINO

Cesar Augusto Seminario Yrigoyen  
Francesco Giuseppe Gillio



Politecnico  
di Torino

## APPLIED DATA SCIENCE PROJECT



# Checkpoint



UNIVERSITÀ  
DI TORINO



Politecnico  
di Torino

## Part 1

# Table of Contents

- ▶▶▶ Project **Background**
- ▶▶▶ Project **Value Proposition**
- ▶▶▶ Project **General Objectives**
- ▶▶▶ Project **Design**
- ▶▶▶ Project **Work Breakdown Structure**

# The Data

---

## Patient Preference Studies

**Patient Preference Studies** evaluate what treatment attributes patients value, their importance, and the trade-offs patients choose to inform healthcare decisions



UNIVERSITÀ  
DI TORINO



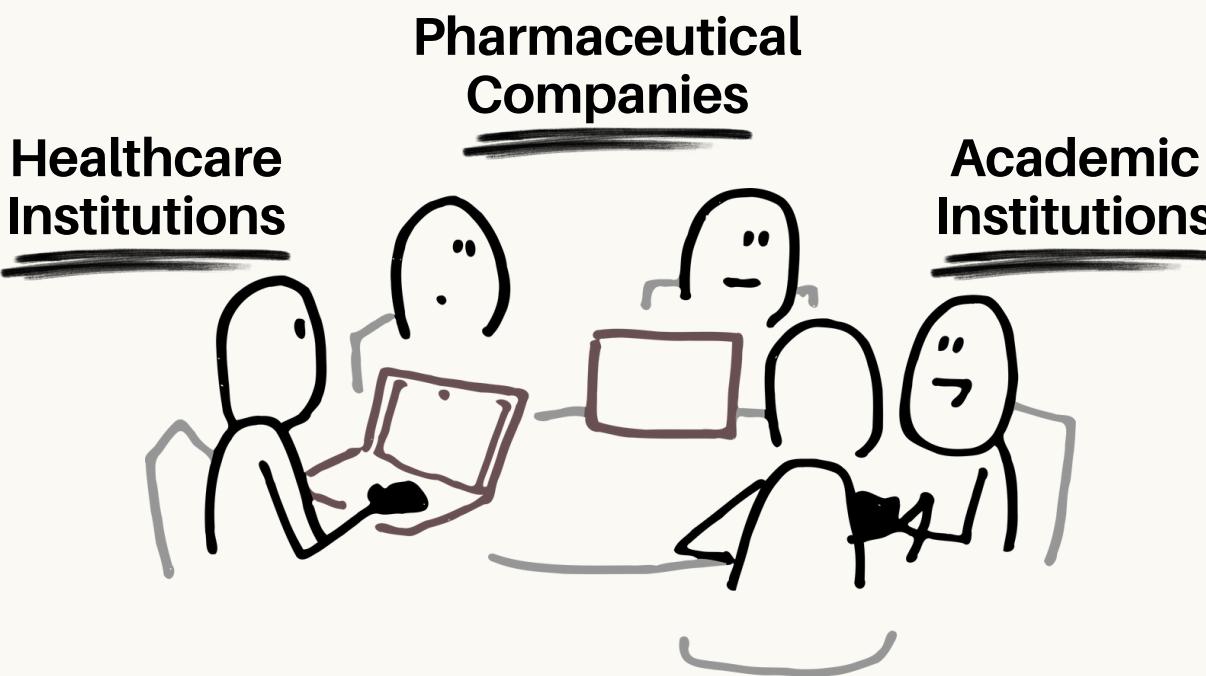
Politecnico  
di Torino

# The Stakeholders



**Medical  
Researchers**

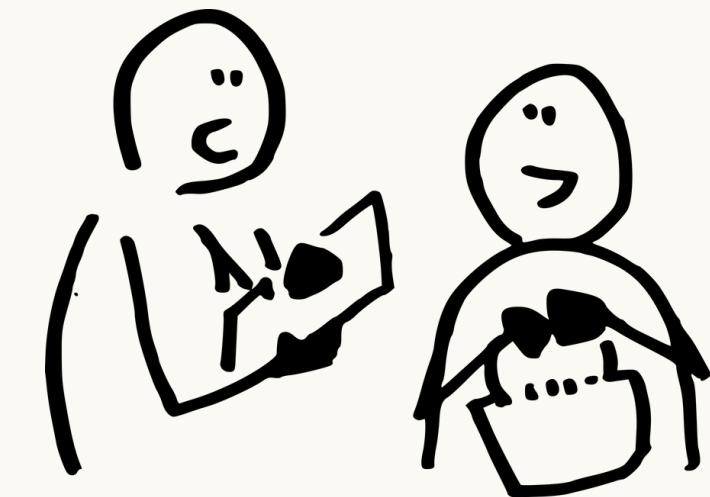
CORE



**Healthcare  
Ecosystem**

DIRECT

valuable insights to



**Patient  
Communities**

INDIRECT

valuable services to

# The Challenges

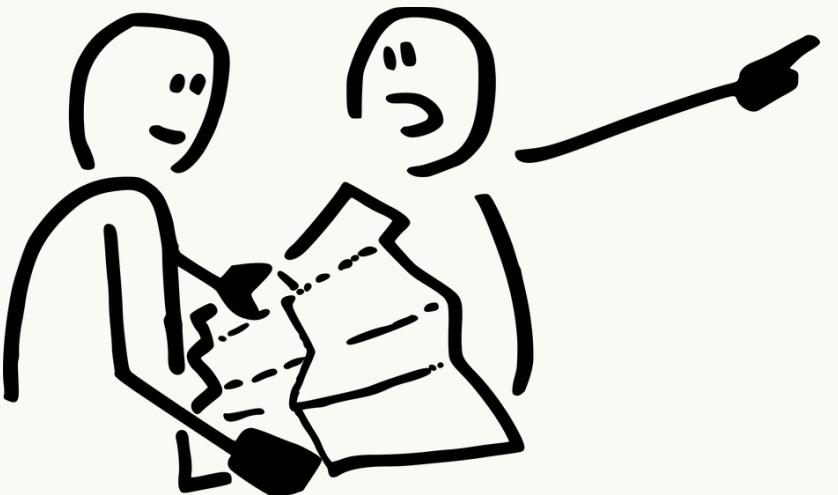
---



## Large Volume

### of Scientific Literature

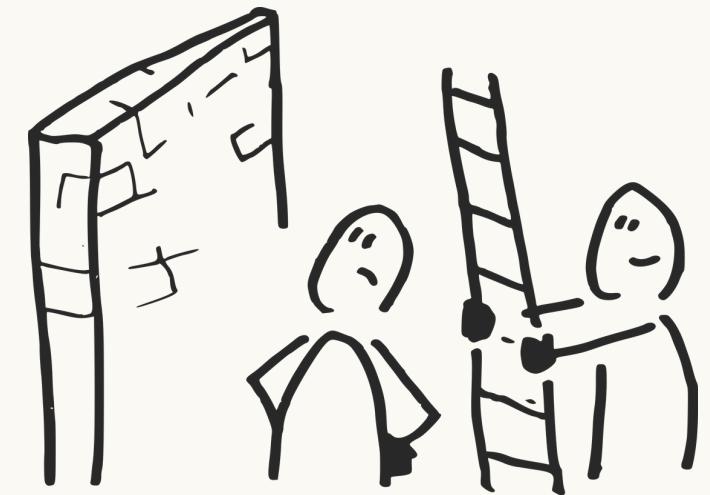
search strings in citation databases (**PubMed**) return a large amount of content, often irrelevant



## Broad Scope

### of Scientific Areas

PPS cover a wide range of clinical areas and accurate searches require manual supervision

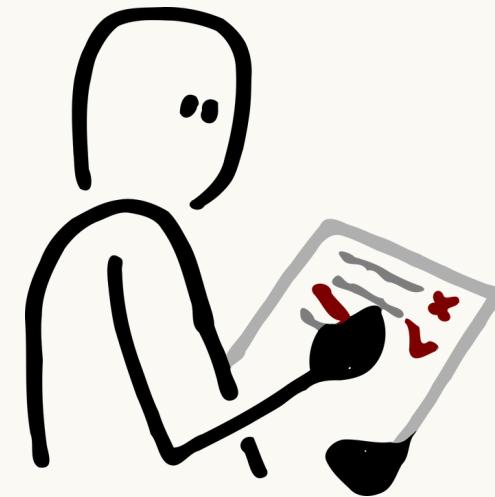


## Adaptation to Scale

### of Scientific Databases

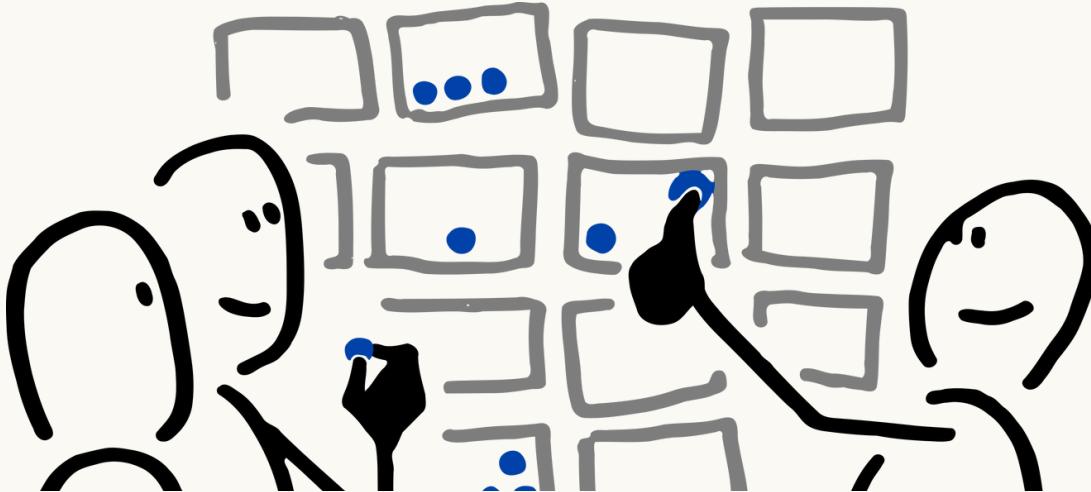
manual supervision struggles to cope with the publication scale of scientific literature

# The Project Values



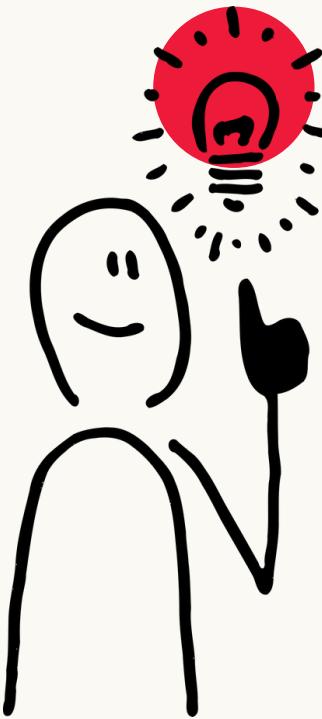
## Improve Relevance

**by Classification System**  
to bypass irrelevant content and return  
high-value literature



## Improve Retrieval

**by Categorization System**  
to categorize search results and improve  
area-specific retrieval



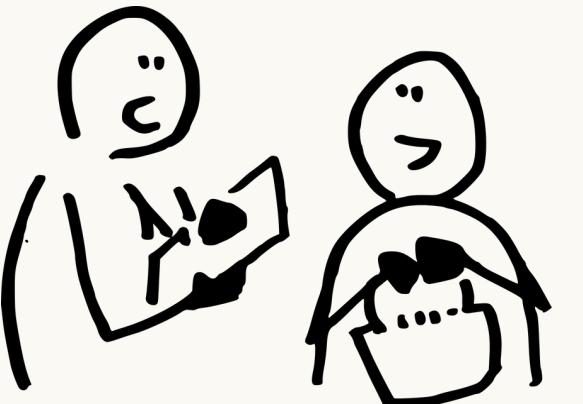
## Improve Efficiency

**by System Automation**  
to reduce manual effort and improve  
access to up-to-date research

# The **Values**



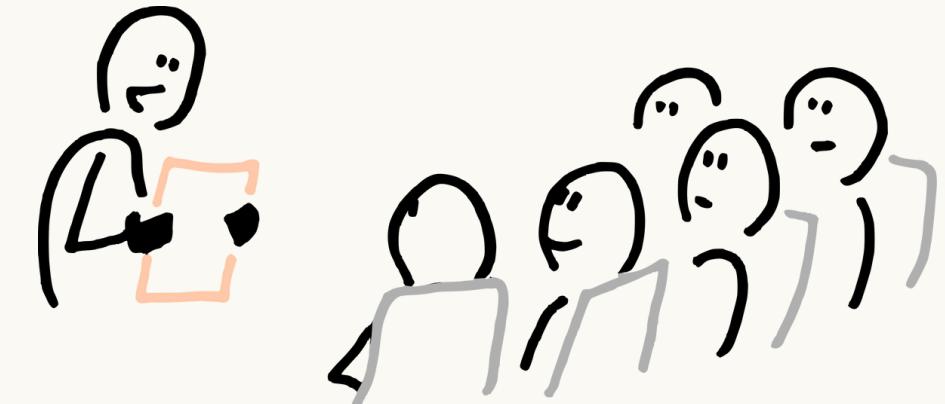
## Sustainable Development Goals 2030



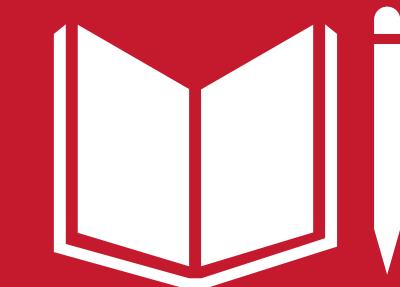
**3** GOOD HEALTH  
AND WELL-BEING



Support the advancement of  
**patient care systems and processes**



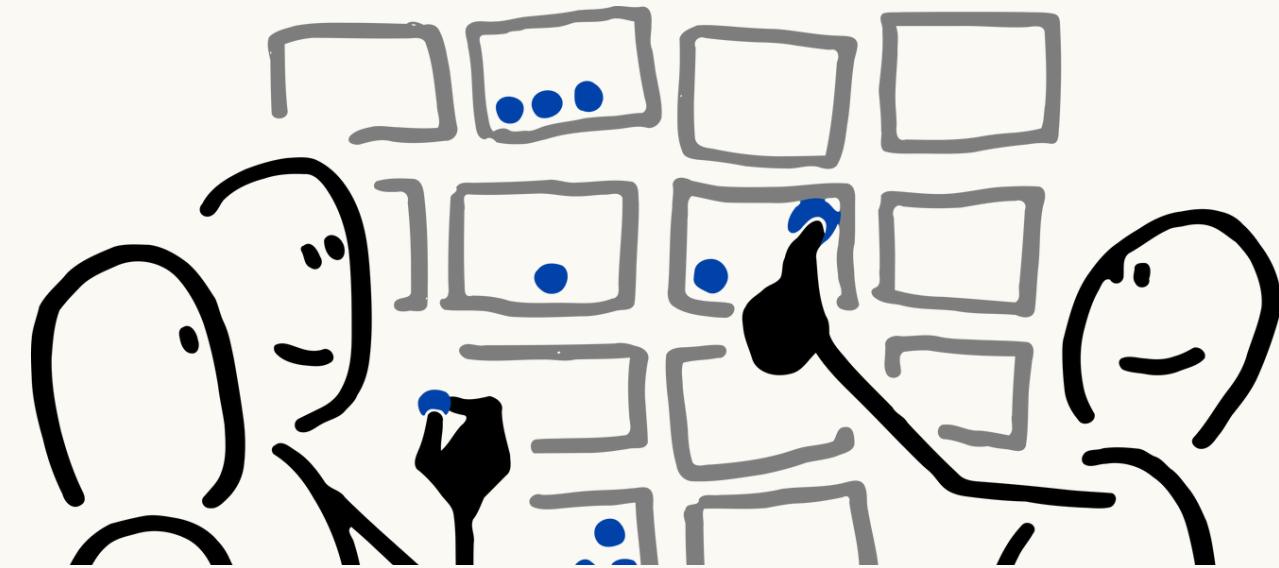
**4** QUALITY  
EDUCATION



Support the advancement of  
**information retrieval in medical research**

# The **Objective**

---



## Classifier Model for Medical Research Papers

- ▶▶▶ Papers classification by relevance to **Patient Preference Studies (PPS)**
  
- ▶▶▶ Papers classification by relevance to **Clinical Areas**

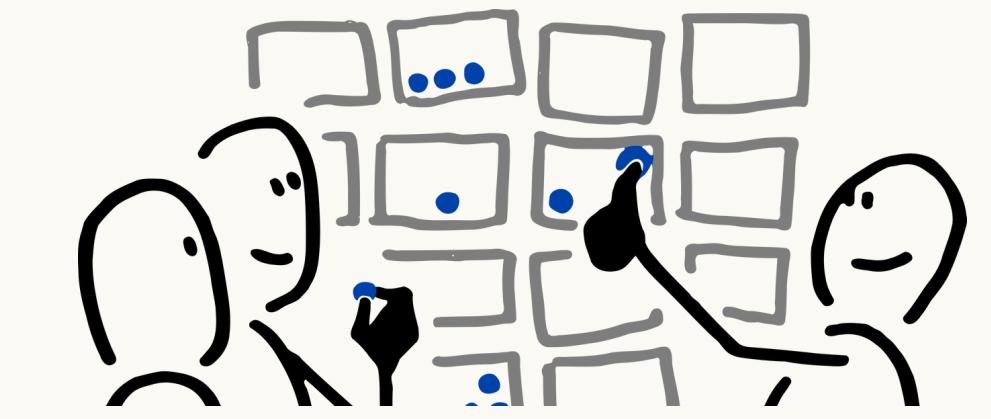
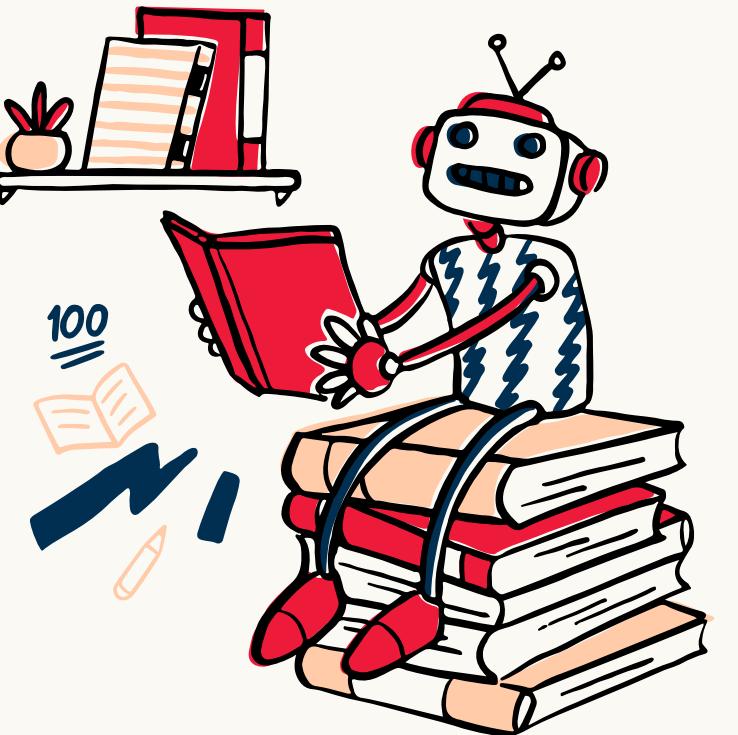
# The **Project Objective**

---



**Supervised  
Binary  
Classifier Model**

to classify  
search string outputs  
by relevance to **PPS**



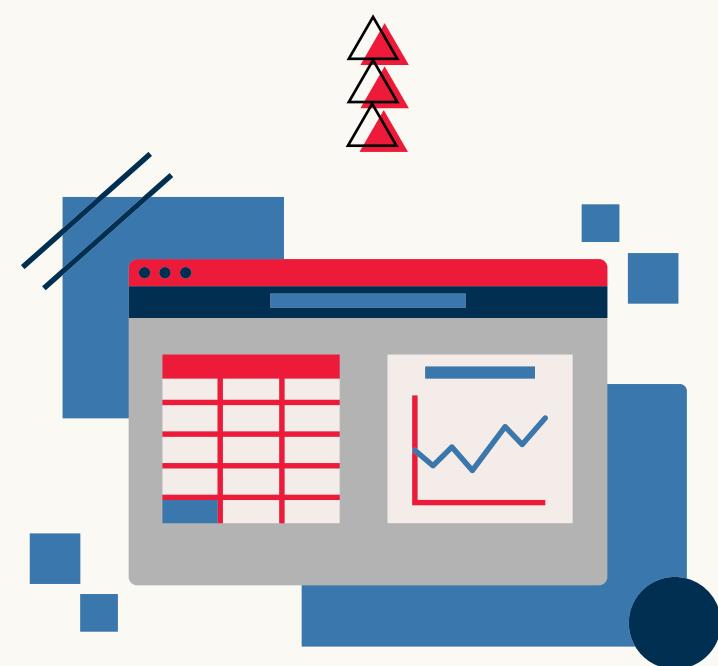
**Self-Supervised  
Multi-Label  
Classifier Model**

to categorize  
PPS-relevant content  
into **Clinical Areas**

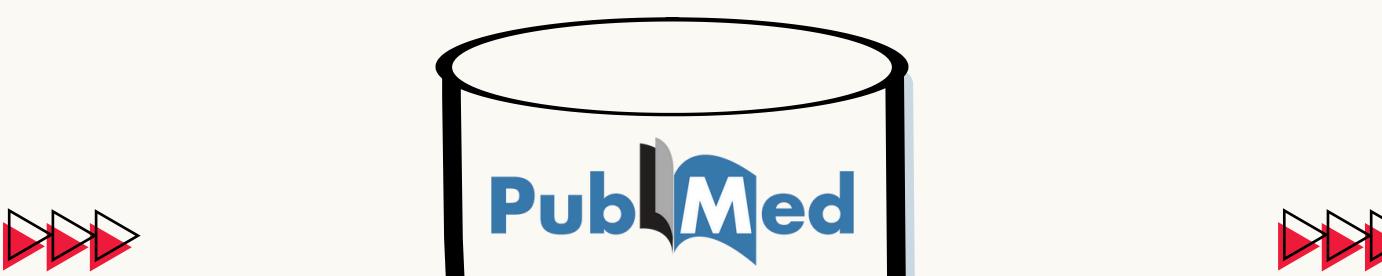
# The Project Design



Medical Researcher



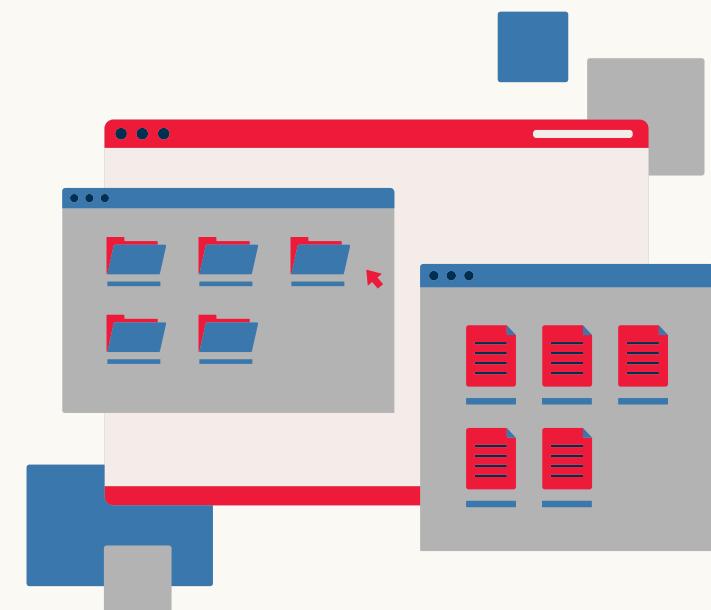
High-quality PPS-relevant literature



Large Search String on  
**Patient Preference Studies**

**Self-Supervised  
Multi-Label  
Classifier Model**

to categorize PPS-relevant  
content into Clinical Areas



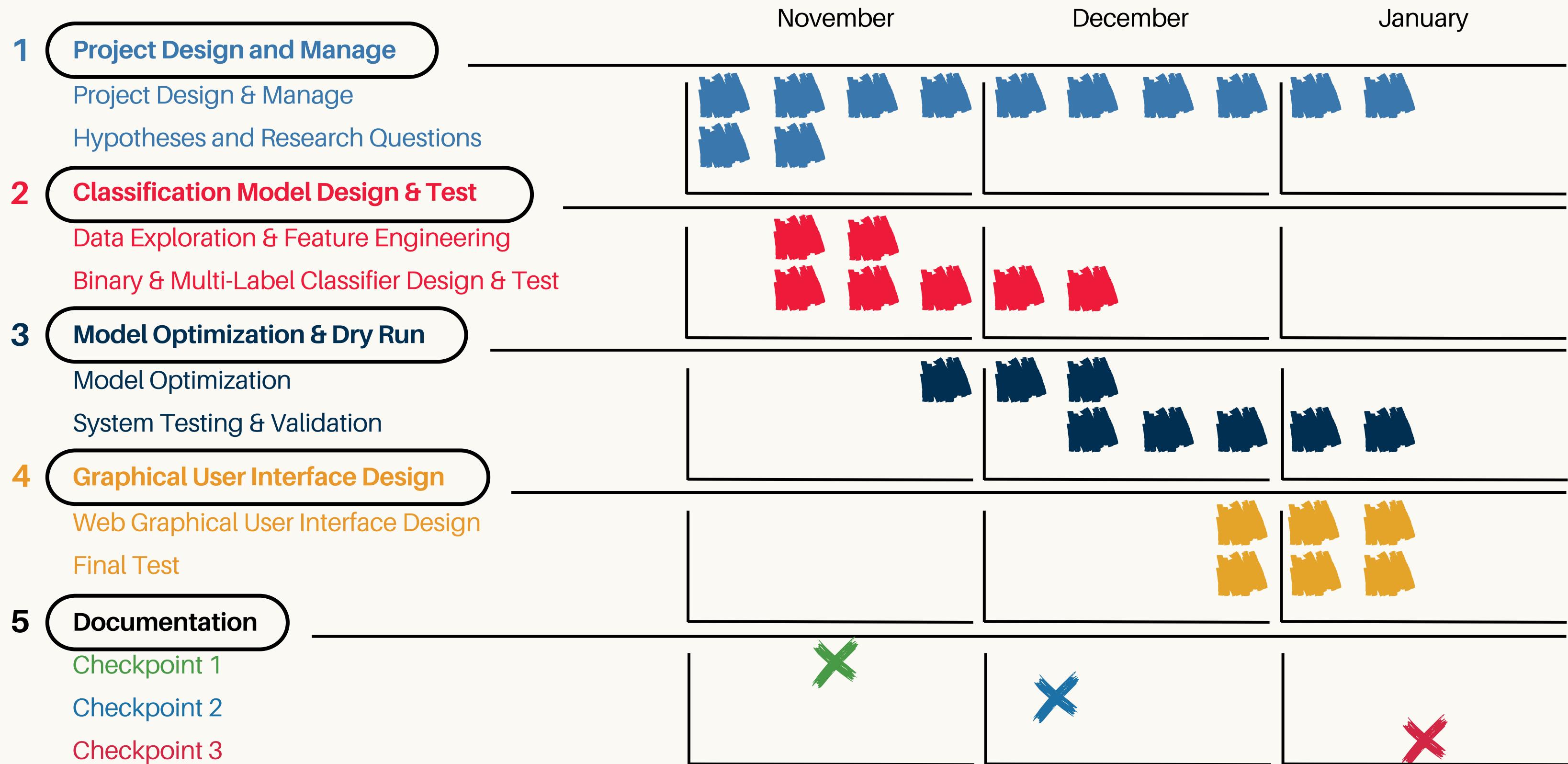
PPS-relevant and irrelevant  
content across clinical areas



**Supervised  
Binary  
Classifier Model**

to classify search string outputs  
by relevance to PPS

# Project Work Breakdown Structure



# Table of Contents

## Part 2

Project  
**Picture**

Project  
**Objectives**

Binary  
**Text Classification Problem**

Multi-Label  
**Text Classification Problem**

Project  
**Pipeline Design**

# The Picture



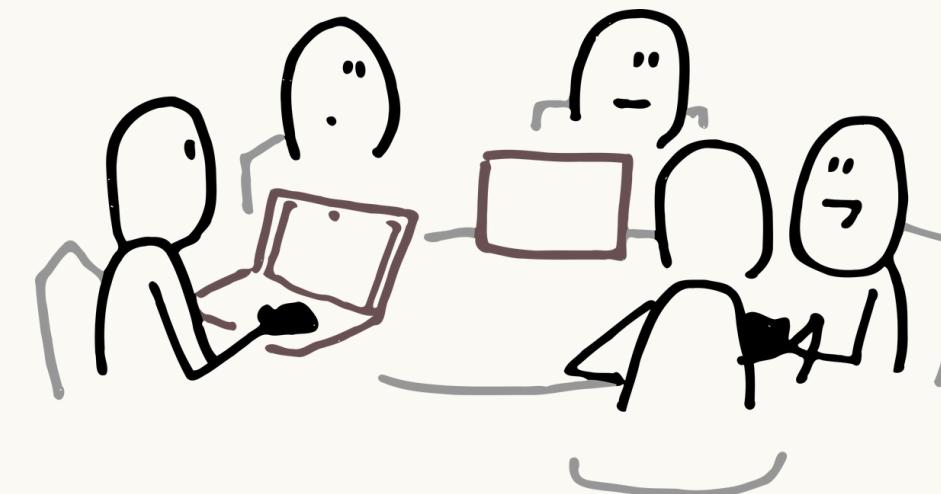
**"The urgent demand for tools that support efficient access, integration, and analysis of health data to derive actionable insights from patient-reported outcomes and real-world evidence"**



**Medical  
Researchers**



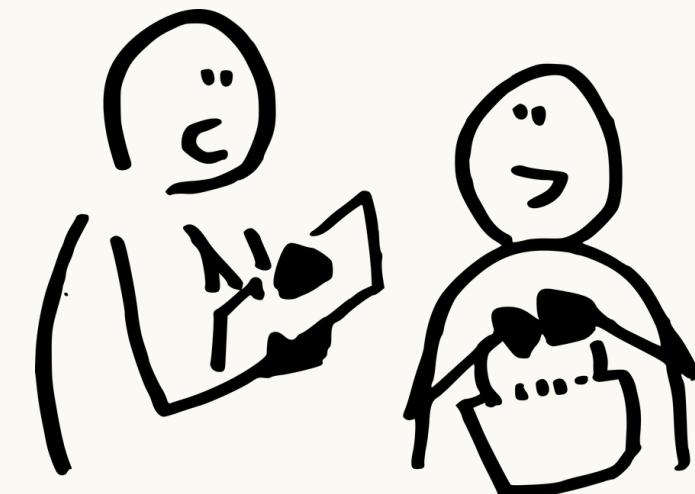
actionable insights to



**Healthcare  
Ecosystem**



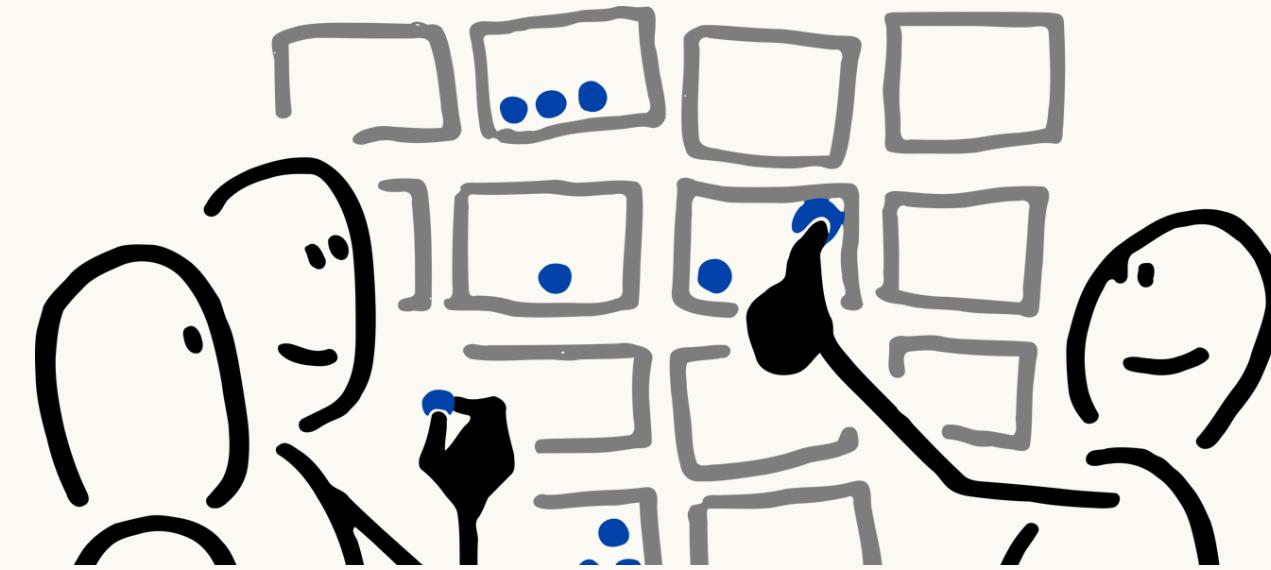
valuable services to



**Patient  
Communities**

# The **Objective**

---



## Classifier Model for Medical Research Papers

- ▶▶▶ Papers classification by relevance to **Patient Preference Studies (PPS)**
  
- ▶▶▶ Papers classification by relevance to **Clinical Areas**

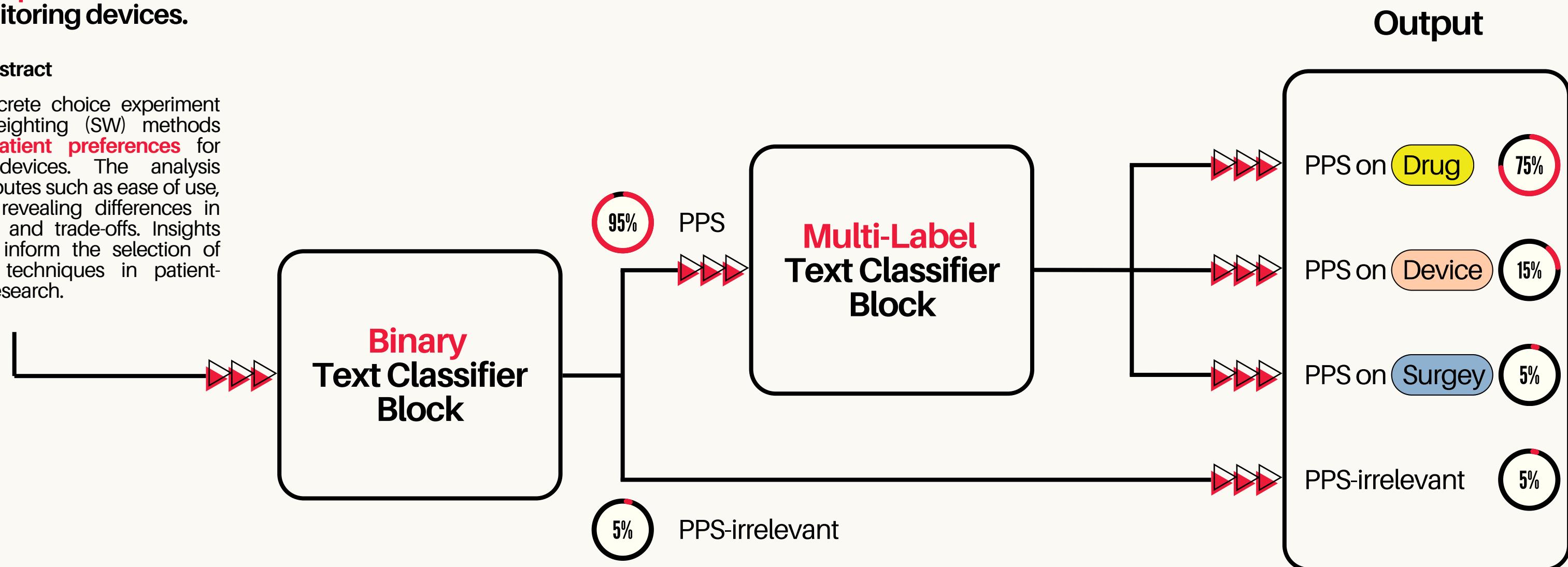
# The Task

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

## Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic patient preferences for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

## Two-Stage Text Classification



# The Binary Text Classifier

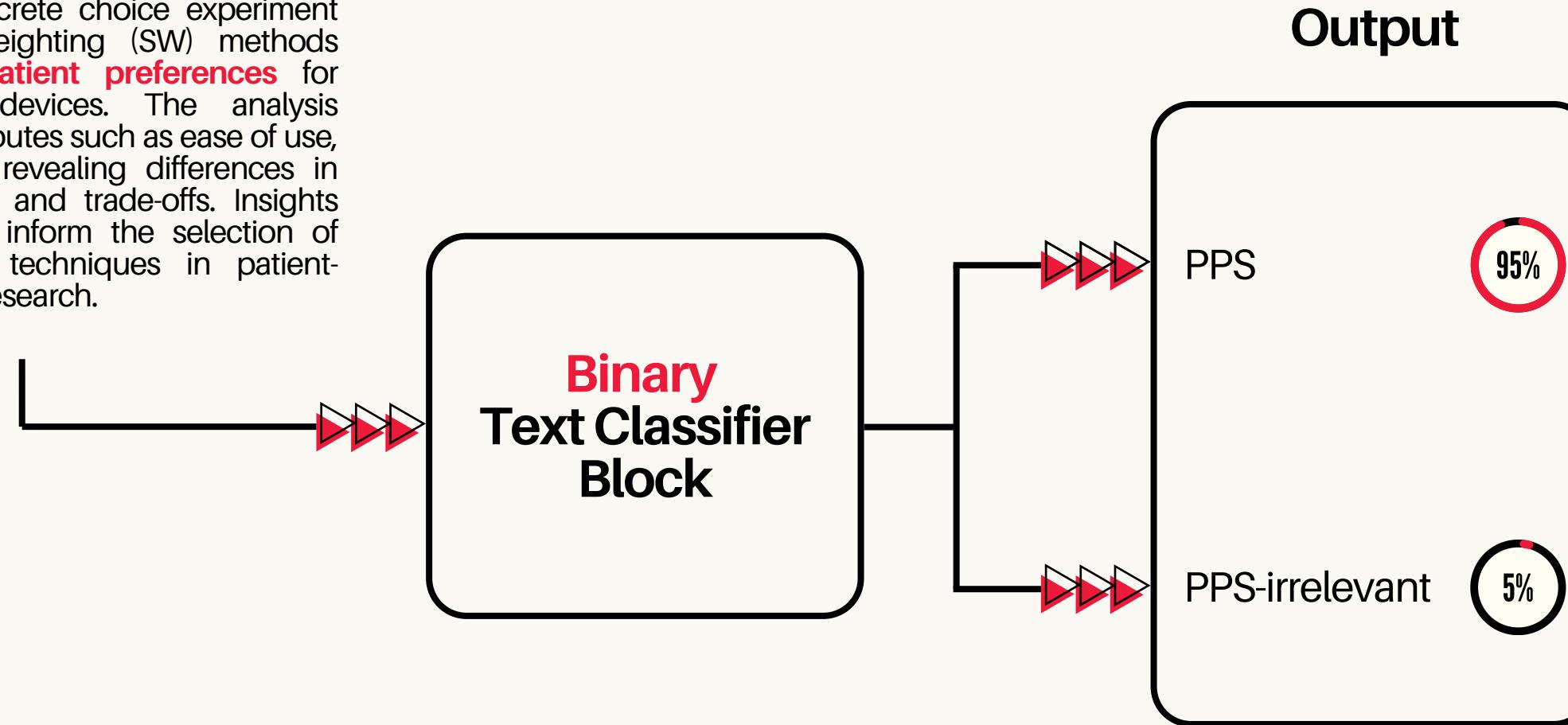


## Classification by relevance to Patient Preference Studies

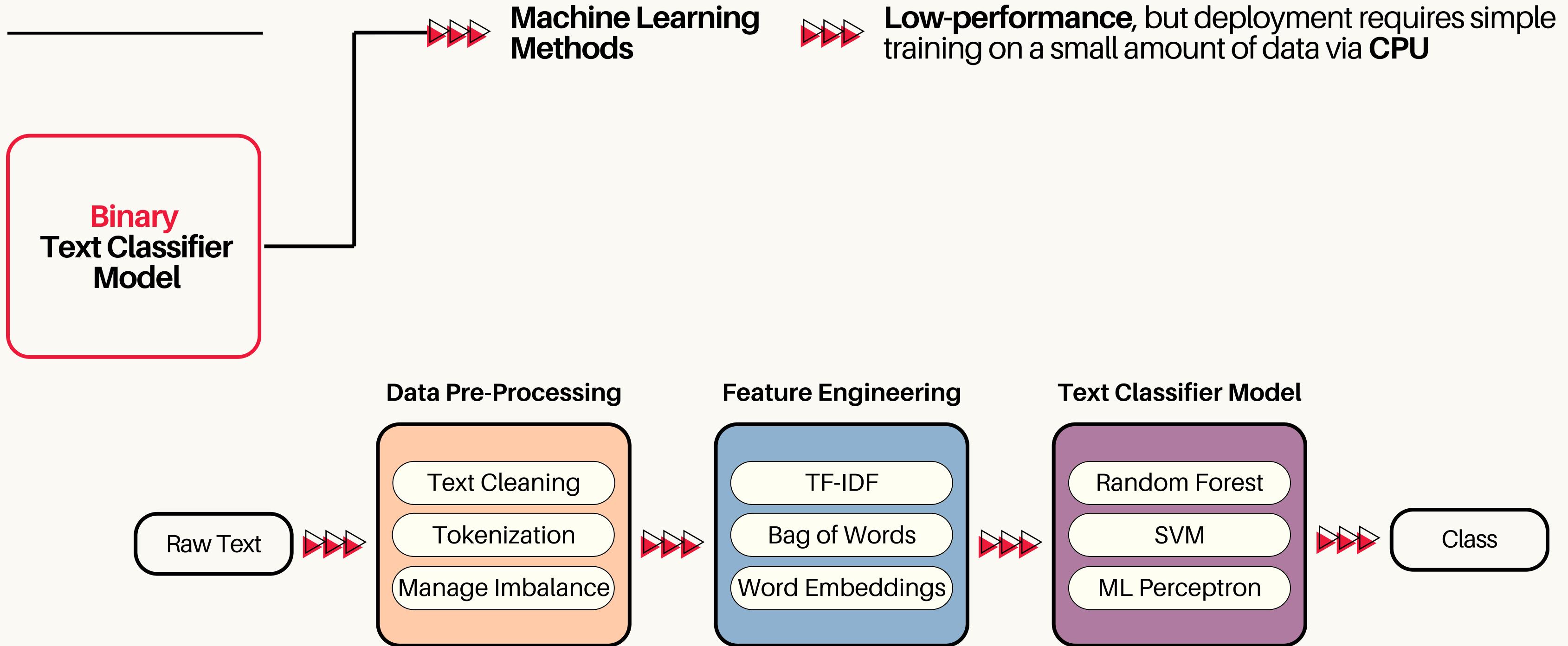
The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

### Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic patient preferences for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.

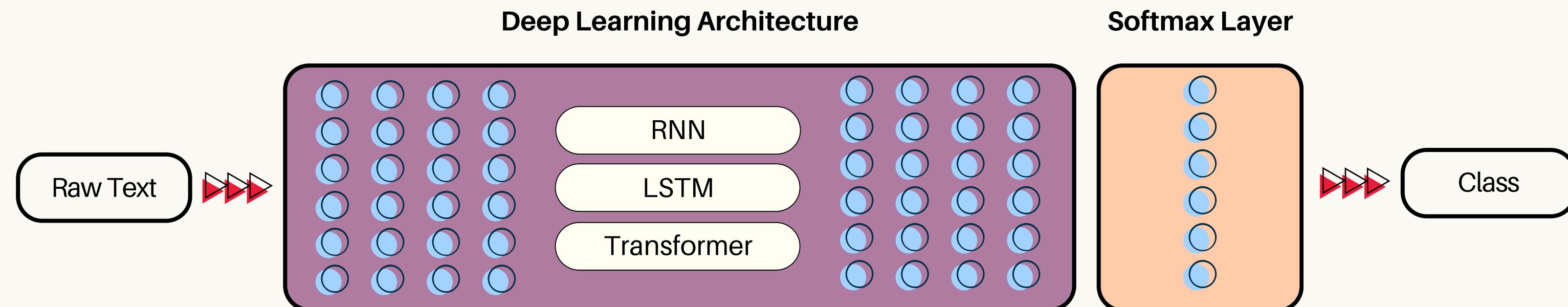
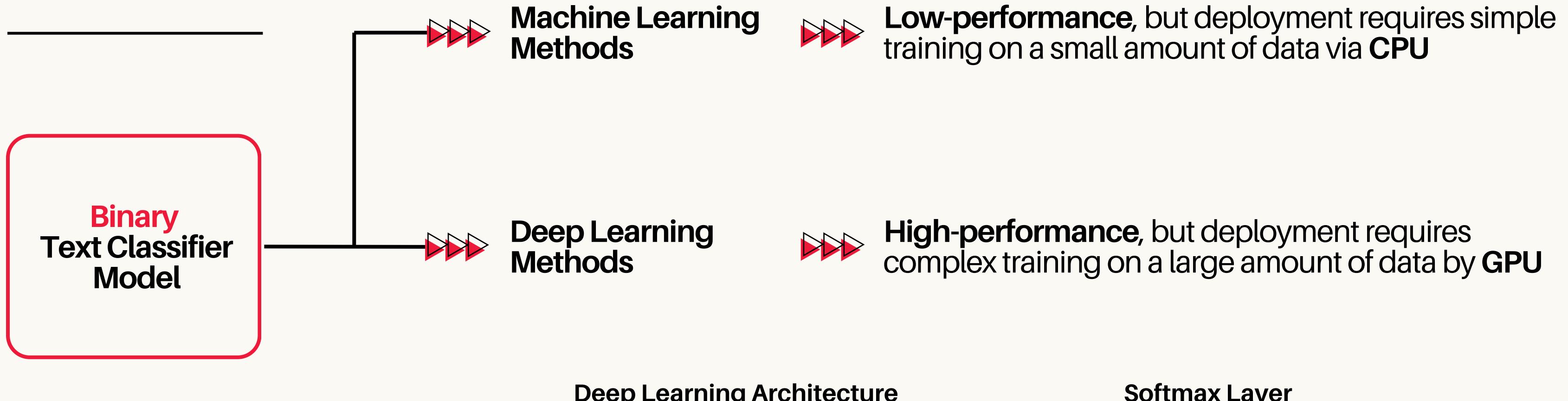


# The Binary Classifier Methods



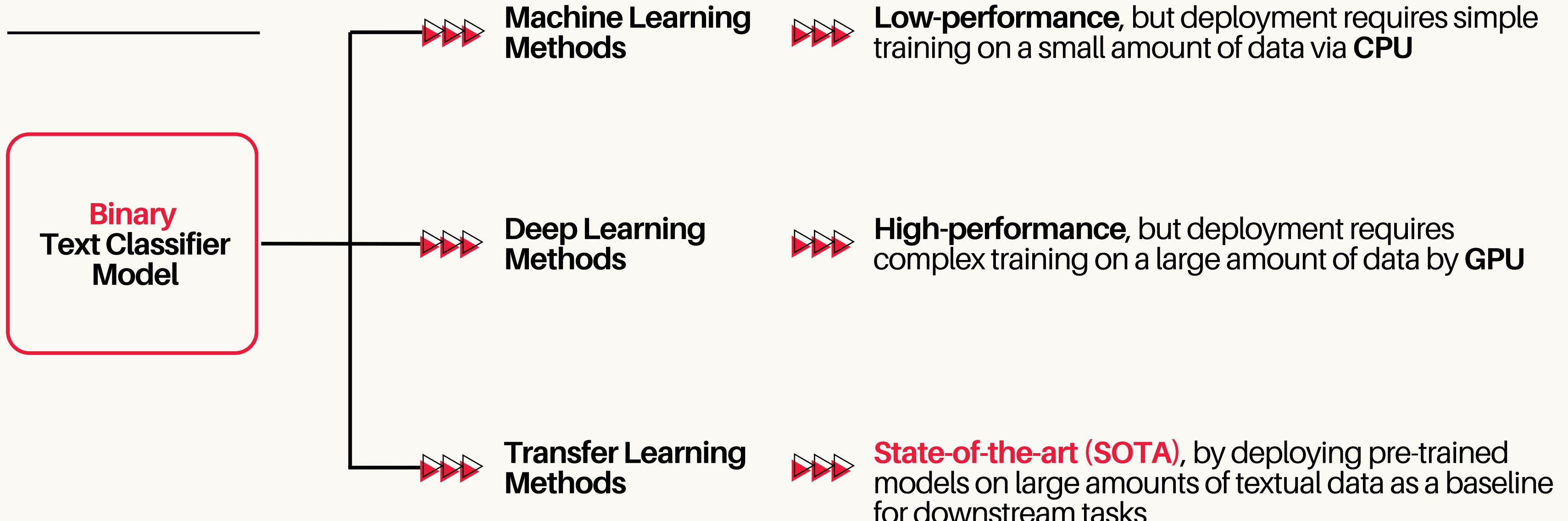
## Binary Text Classification Problem

# The Binary Classifier Methods

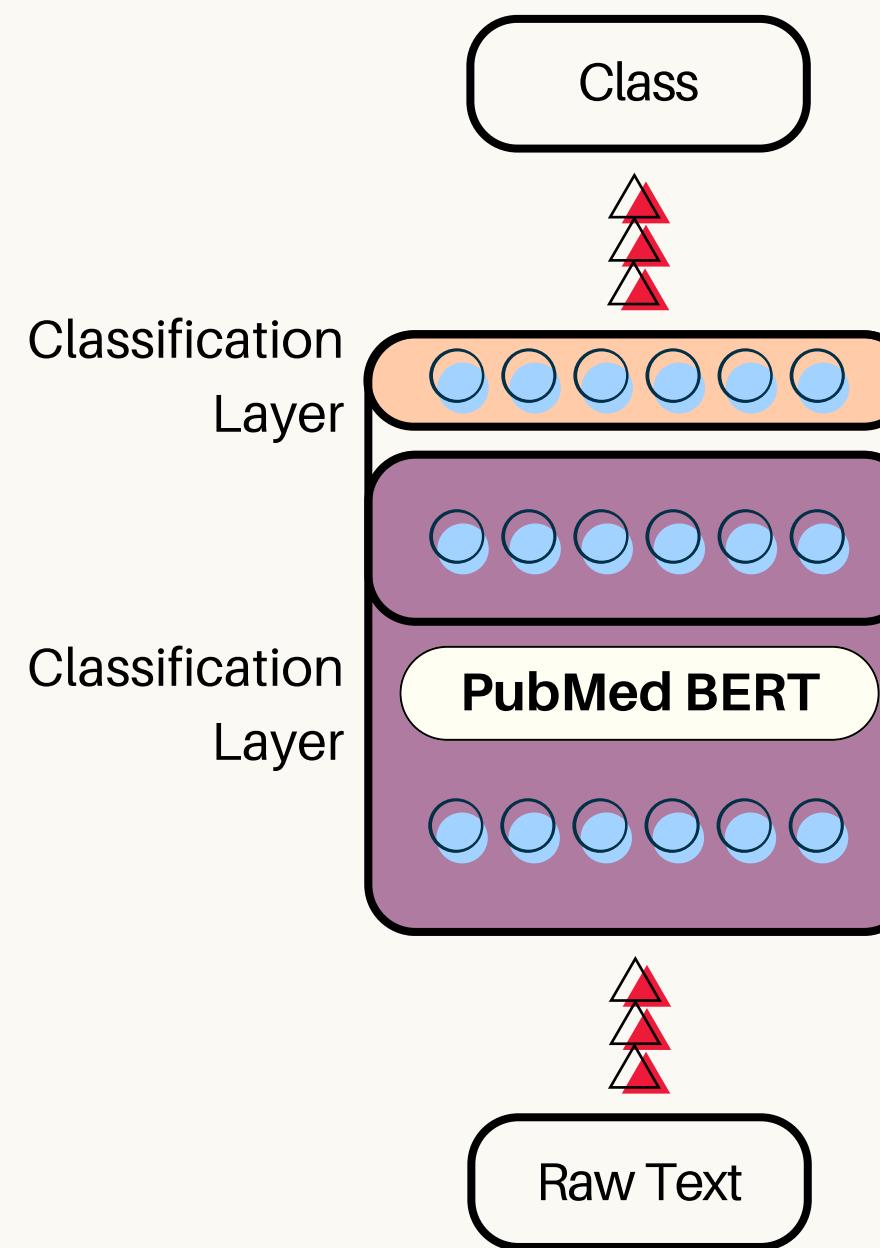


## Binary Text Classification Problem

# The Binary Classifier Methods

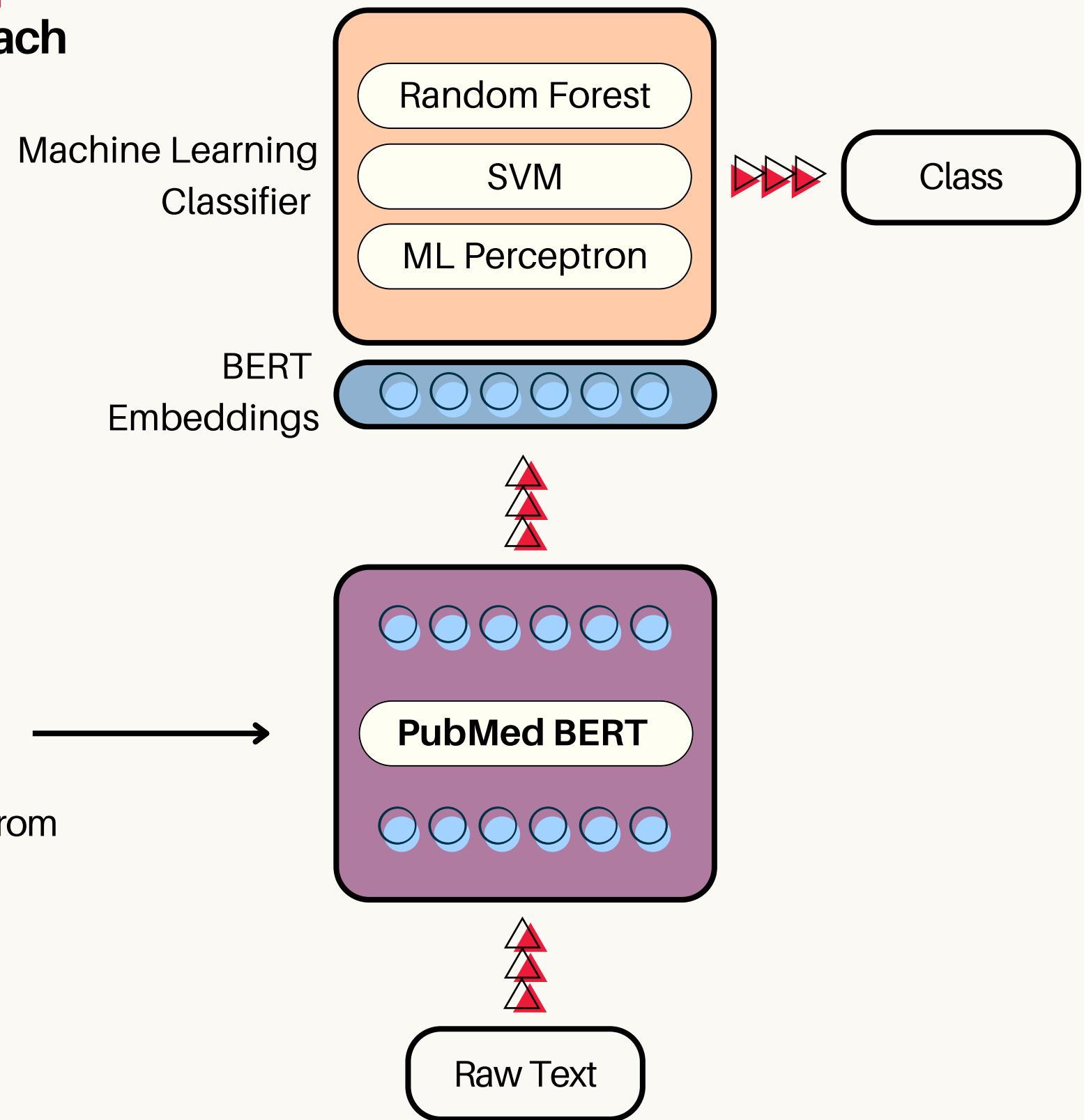


# The Choices



## The Standard Approach

## The Hybrid Approach



Pre-Trained  
BERT Model

from scratch on abstract from  
PubMed

# The Multi-Label Text Classifier

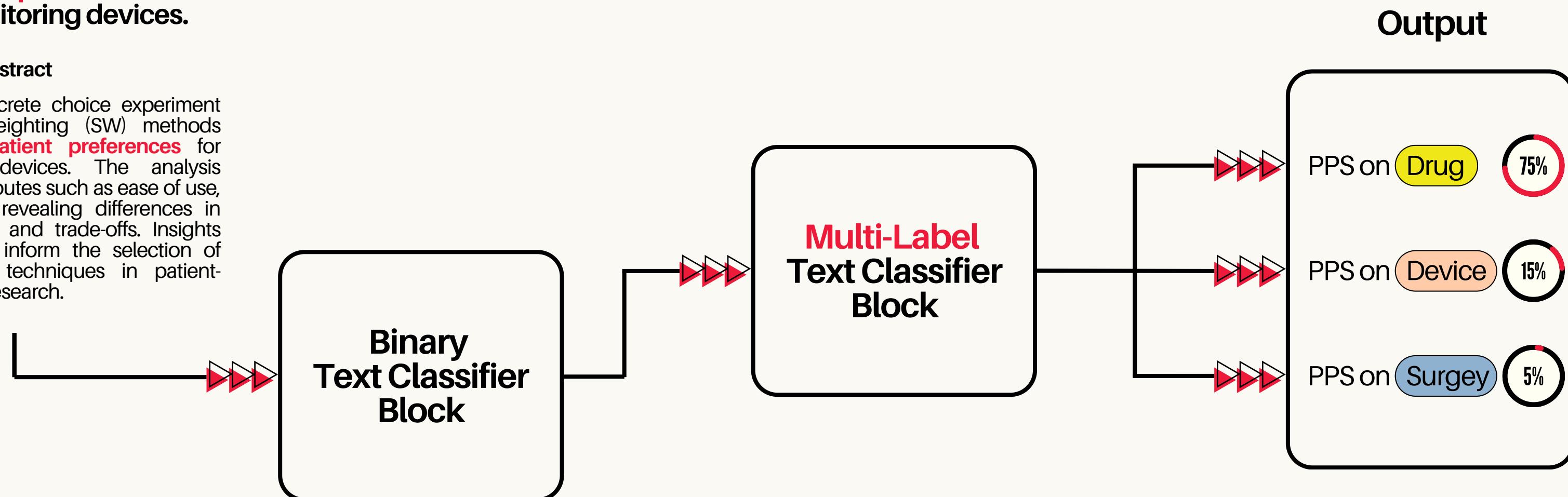


## Classification by relevance to Clinical Areas

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

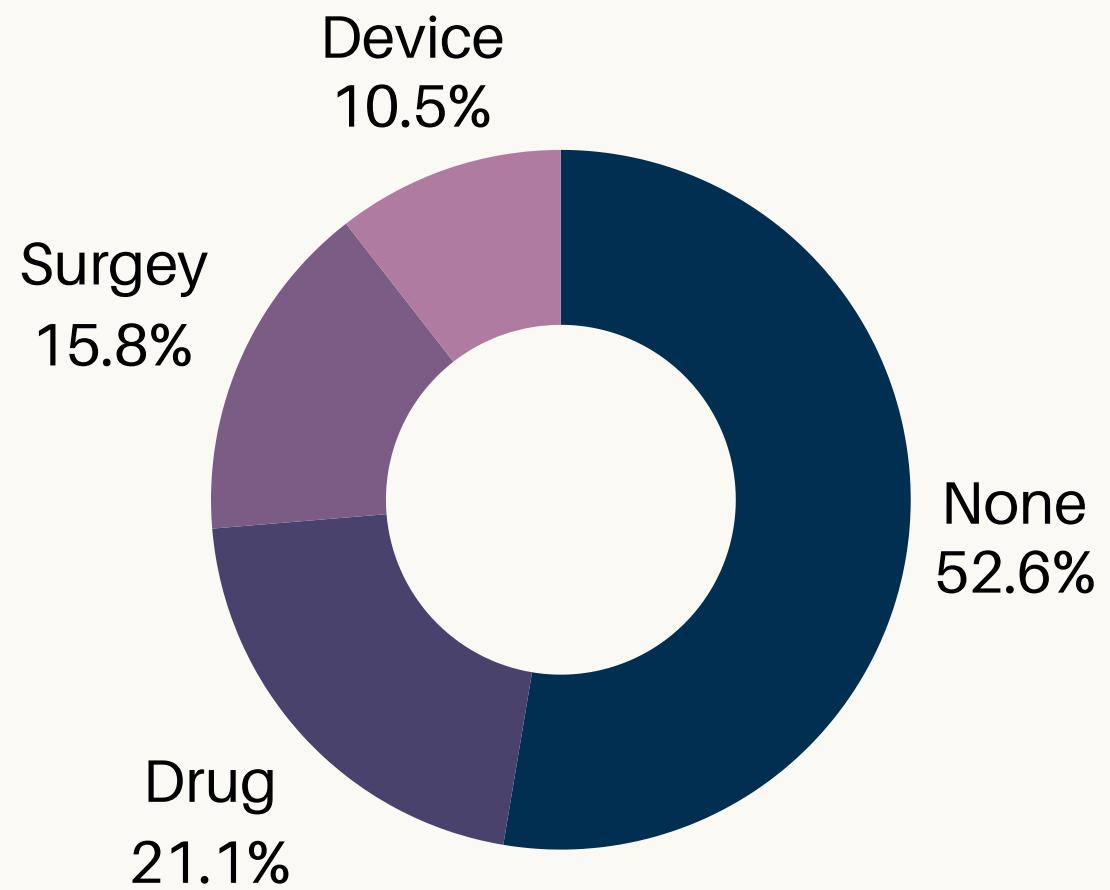
### Abstract

A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic patient preferences for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs. Insights from this evaluation inform the selection of preference-elicitation techniques in patient-centered healthcare research.



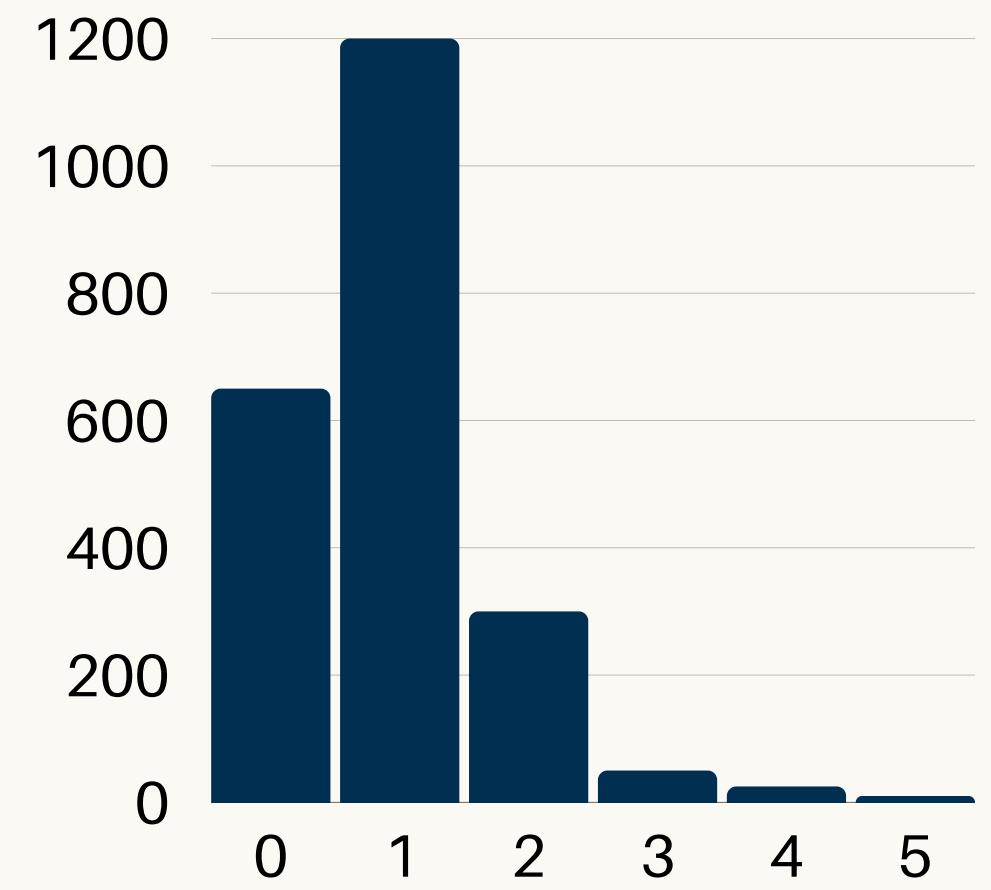
# The Multi-Label Classifier Problem

Class Distribution



## Data Imbalance

Multi-Label Distribution



# Research **Questions**

---

“What are the most effective techniques for **building a multi-label classification model** to categorize scientific articles while addressing data imbalance?”

“How **can topic modeling** be used to determine if the identified **areas are sufficient, discover new areas** of interest and can the same model be leveraged for **classification?**”

# The Multi-Labels Classifier Models

Multi-Label  
Text Classification Problem



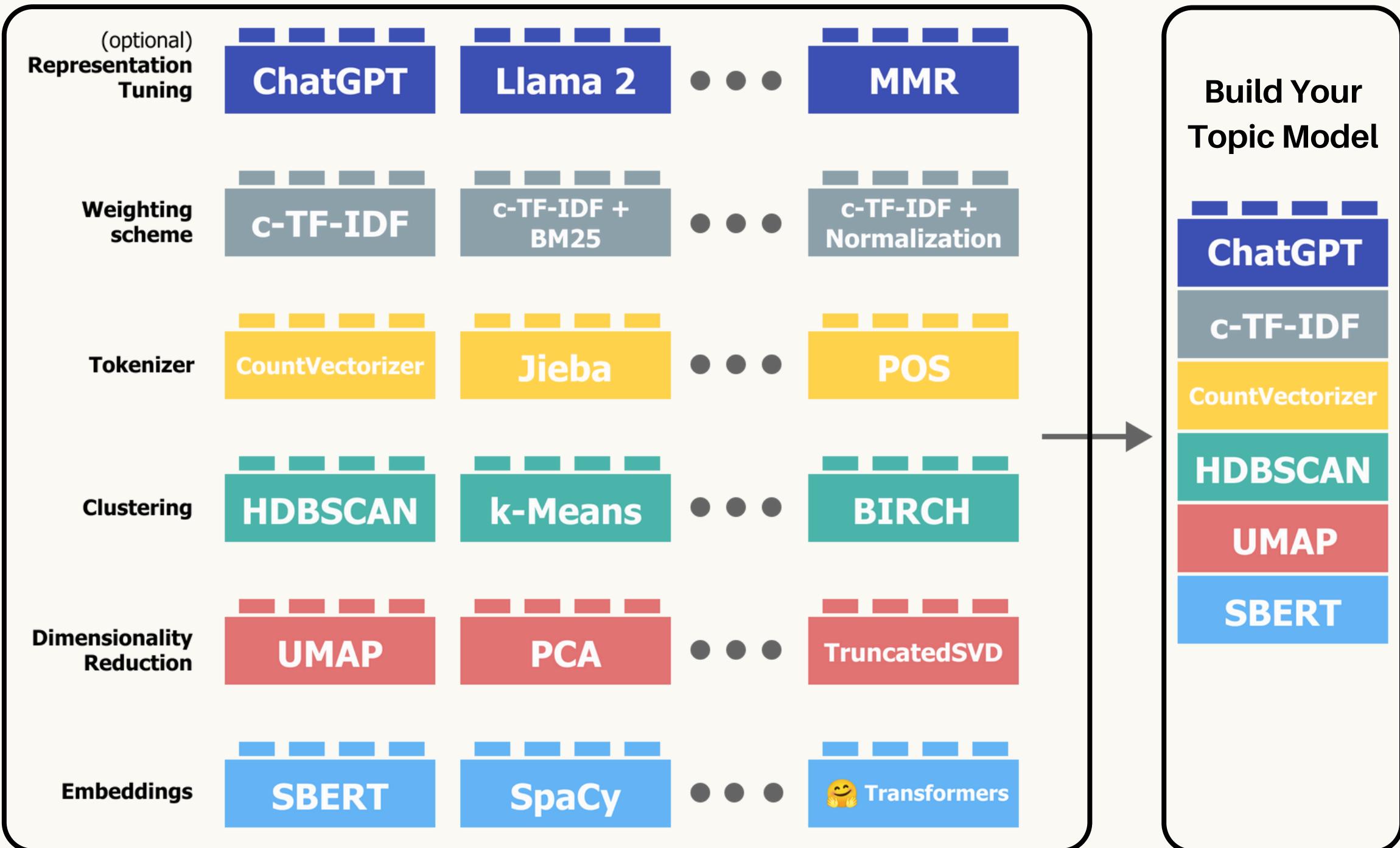
Problem Adaptation

Algorithm Adaptation

- Multi-Label k-NN
- Multi-Class MLP
- Ranking SVM

Topic Modelling

- LDA
- NMF
- Top2Vec
- Bertopic



# BERTopic Project Benefits

---

▶▶▶ Topic Exploration

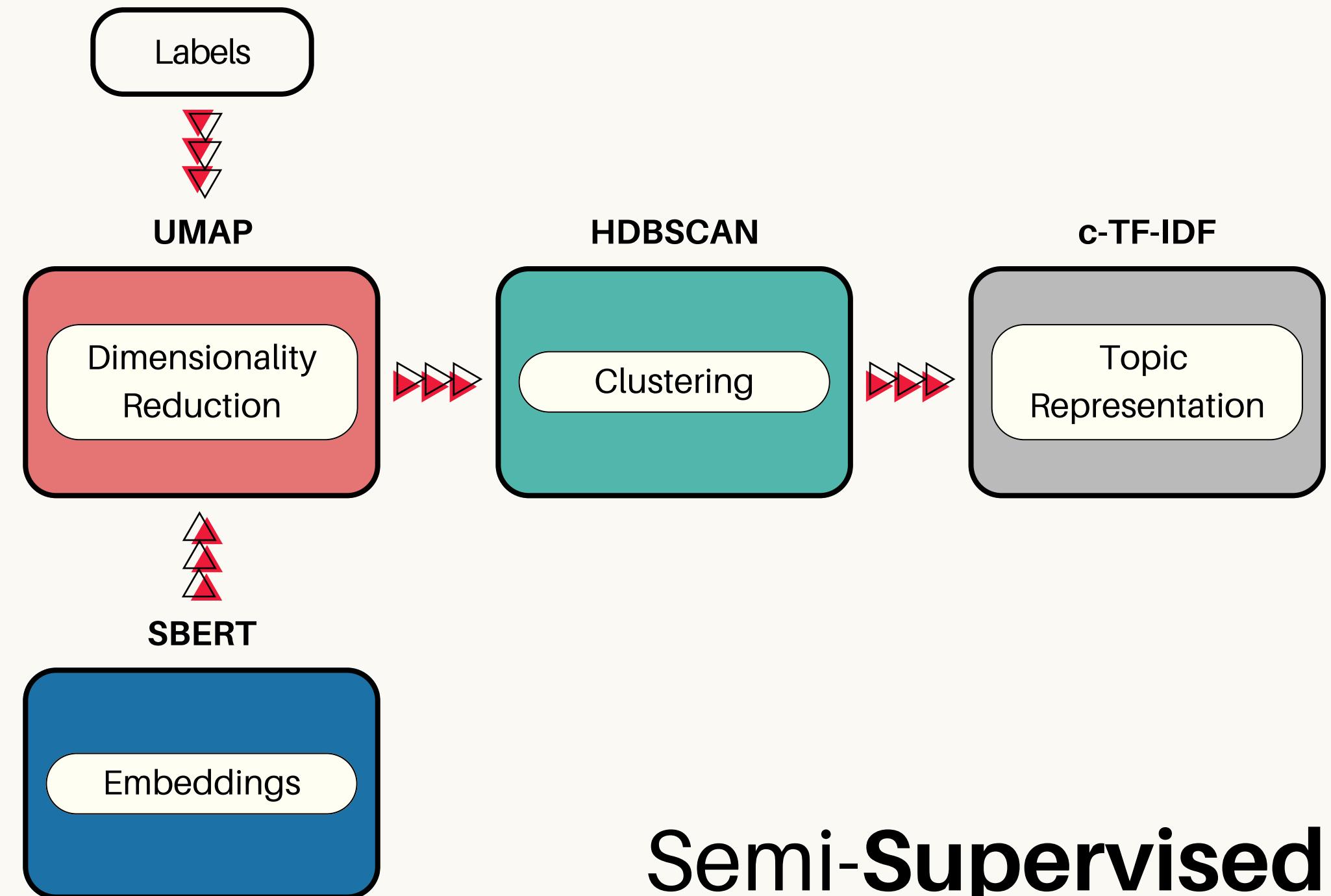
▶▶▶ Categories Not Fixed

▶▶▶ Improve & Speed Up  
Manual Labeling

- Binary Classifier
- Multi-Label Classifier

▶▶▶ Classification Approaches:

- Unsupervised
- Semi-Supervised
- Supervised: Classification



# BERTopic Project Benefits

---

▶▶▶ Topic Exploration

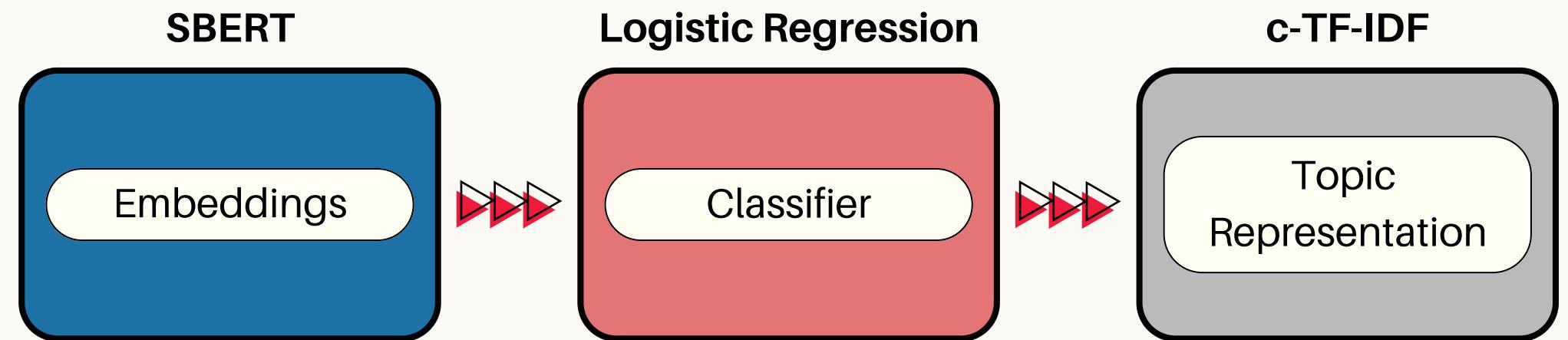
▶▶▶ Categories Not Fixed

▶▶▶ Improve & Speed Up  
Manual Labeling

- Binary Classifier
- Multi-Label Classifier

▶▶▶ Classification Approaches:

- Unsupervised
- Semi-Supervised
- Supervised: Classification

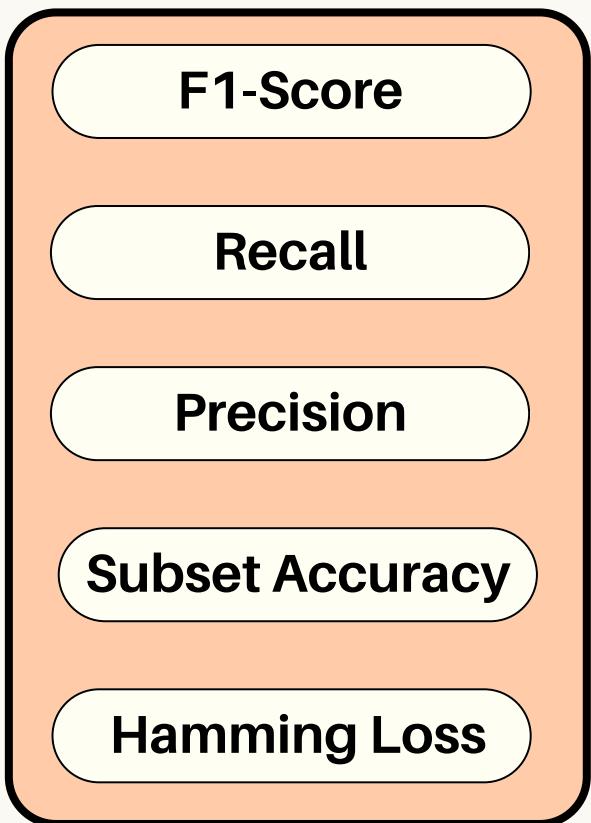


Supervised

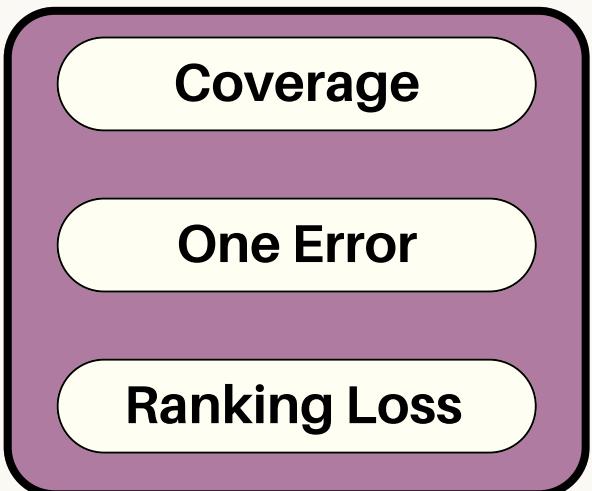
# The Evaluation Metrics

## Classification Models

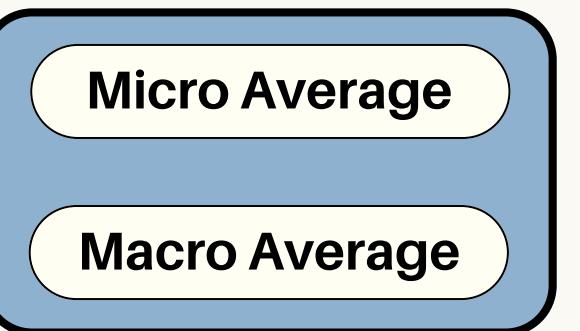
Example-Based



Ranking-Based



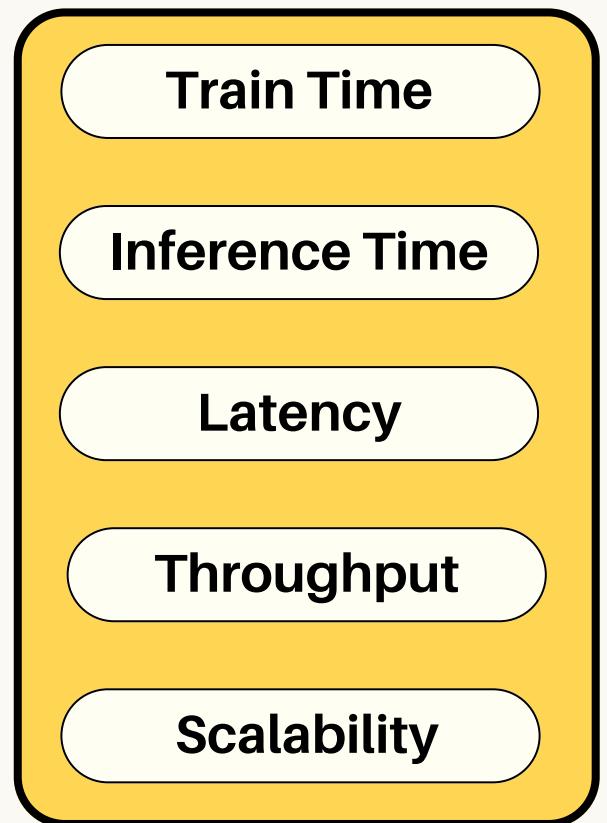
Label-Based



## Topic Modelling

- Diversity
- Coherence

## System Time



# The Design Pipeline

The head-to-head comparison of diabetic patient preferences for glucose-monitoring devices.

**Abstract**  
This study compares a discrete choice experiment (DCE) and swing-weighting (SW) by eliciting preferences for glucose-monitoring devices in a population of diabetes patients...

Clinical Pubmed Article (PPS)

Data Augmentation

## Data Pre-Processing

Text Cleaning  
Handle Imbalance

## Feature Engineering

PubMed BERT

Outliers  
Label-Free Data

## Model Optimization

Params Tuning  
Cross Validation

## Evaluation

Label-Based  
Example-Based  
Ranking-Based  
Time-Based

## Output

**Predicted Multi-Label Classifications**

**Help Data Pruning**

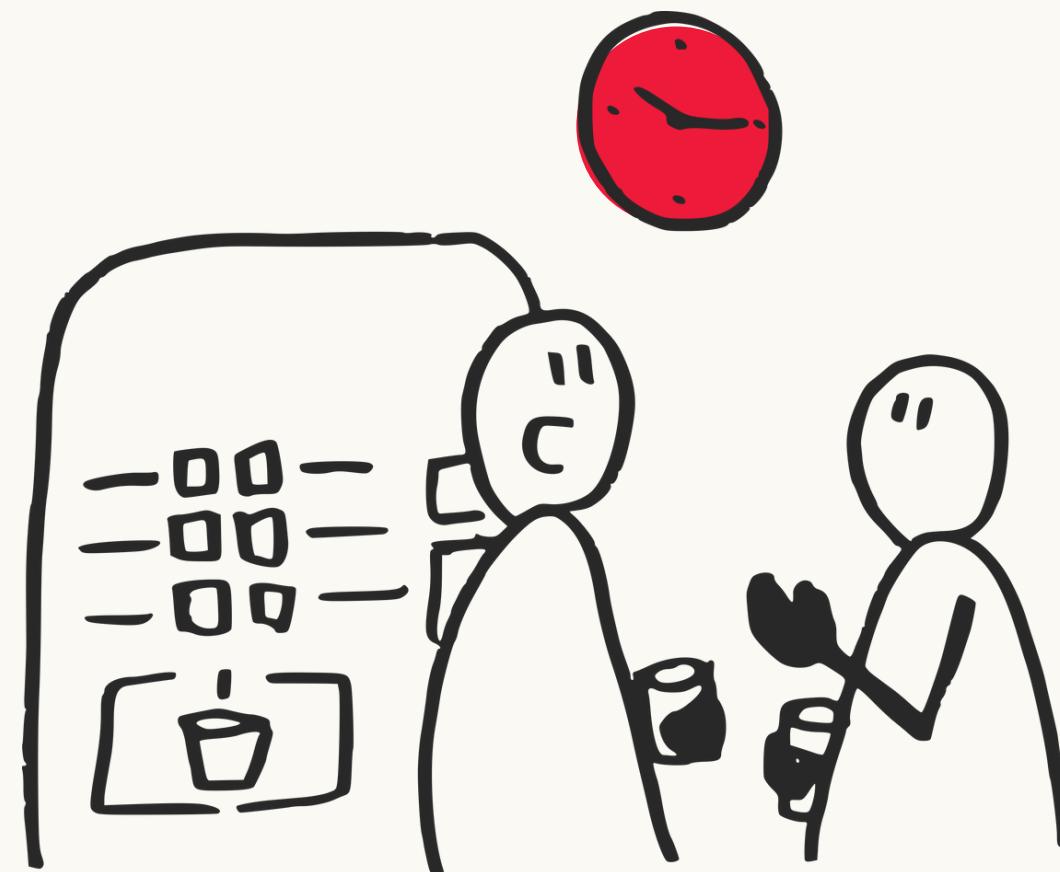
**Discover New Topics of Interest**

## Modeling MLP Classifiers

Multi-Label k-NN  
Multi-Label MLP  
Ranking SVM  
**BERTopic**  
Topic Modeling

## APPLIED DATA SCIENCE PROJECT

# Thank You



Cesar Augusto Seminario Yrigoyen  
Francesco Giuseppe Gillio



UNIVERSITÀ  
DI TORINO



Politecnico  
di Torino