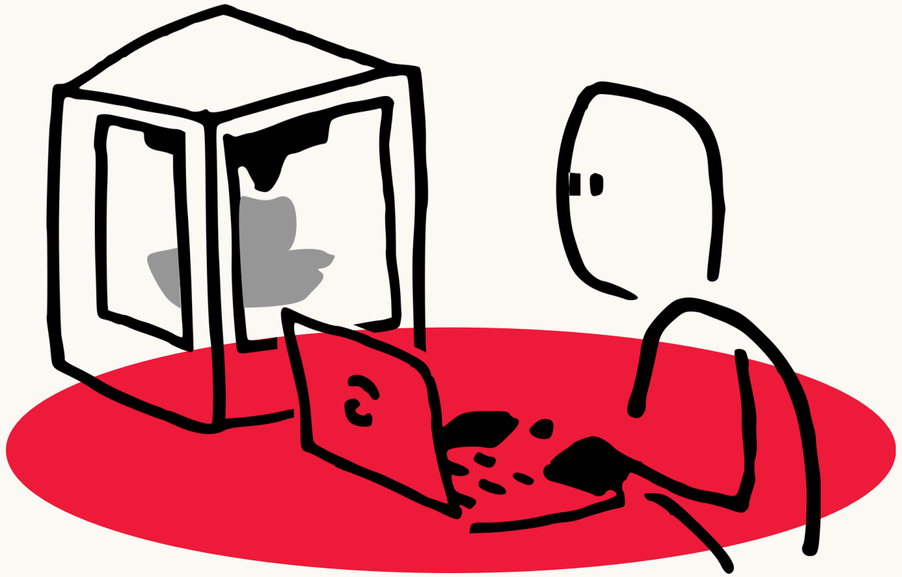
APPLIED DATA SCIENCE PROJECT

Patient Preference Studies
Classification System

Francesco Giuseppe Gillio & Cesar Augusto Seminario Yrigoyen



UNIVERSITÀ DI TORINO

Politecnico di Torino

APPLIED DATA SCIENCE PROJECT

Checkpoint

UNIVERSITÀ DI TORINO

Politecnico di Torino

Part 1

Table of Contents

The
Data

Patient Preference Studies

Patient Preference Studies explore
therapy attributes important to **patient communities**,
their significance, and acceptable trade-offs

Available on

PubMed

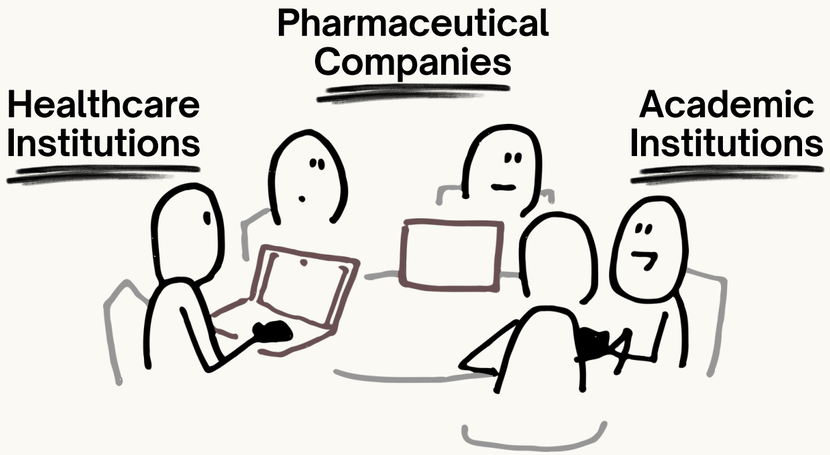
UNIVERSITÀ
DI TORINO

Politecnico
di Torino

The Stakeholders

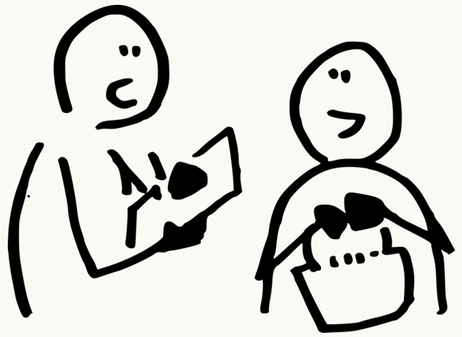
Medical Researchers

CORE



Healthcare Ecosystem

DIRECT



Patient Communities

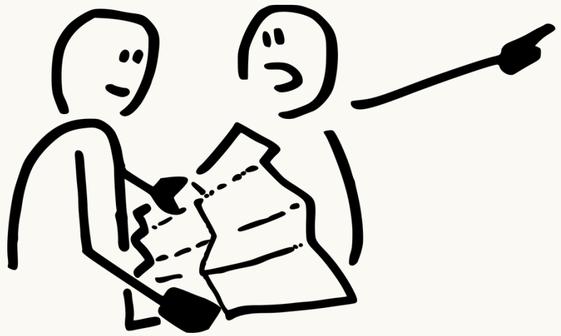
INDIRECT

The Challenges

Large Volume

of Scientific Literature

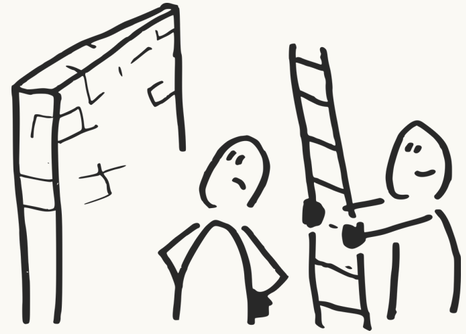
search strings in citation databases (**PubMed**) return a large amount of content, often irrelevant



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

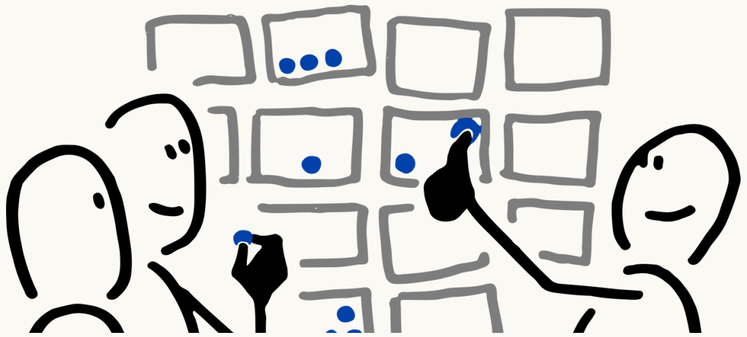
manual supervision struggles to cope with the publication scale of scientific literature

The Project Values

Improve Relevance

by Classification System

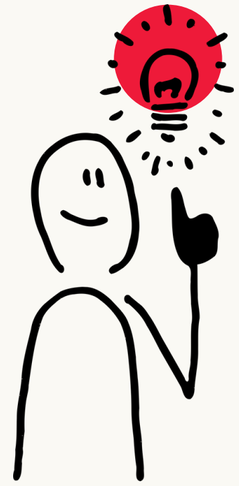
to bypass irrelevant content and return high-value literature



Improve Retrieval

by Categorization System

to categorize search results and improve area-specific retrieval



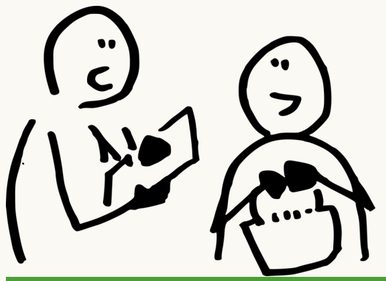
Improve Efficiency

by System Automation

to reduce manual effort and improve access to up-to-date research

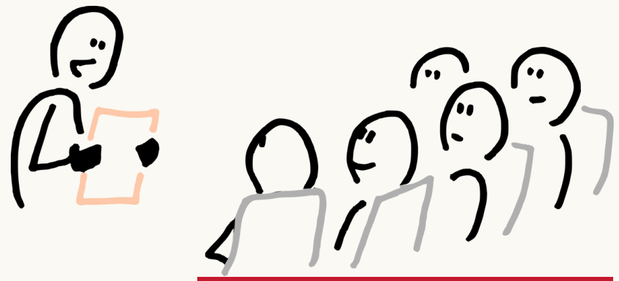
The Project Values

Sustainable Development Goals Alignment



3 GOOD HEALTH AND WELL-BEING

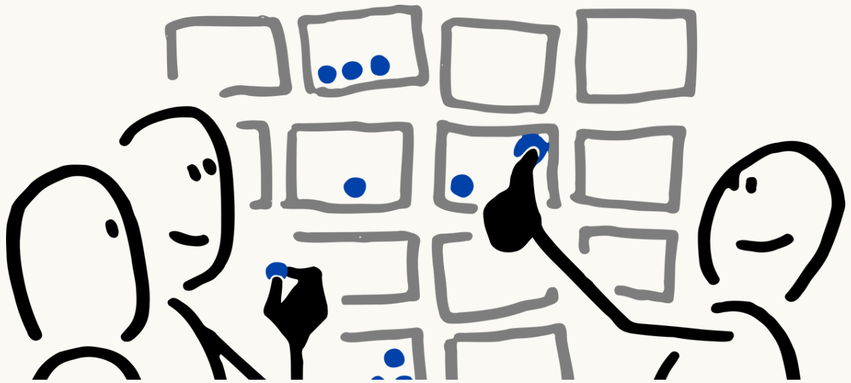
Support the advancement of **patient care** through accurate and high-quality patient-relevant data



4 QUALITY EDUCATION

Improve information retrieval in **medical and health research** through accurate, high-quality and relevant literature

The Project Objective



Classifier Model for Scientific Papers

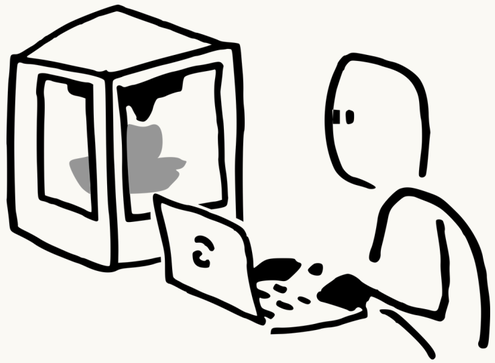
to detect and categorize **Patient Preference Studies** into Clinical Areas by

- **Titles**
- **Abstracts**

The
Project Objective

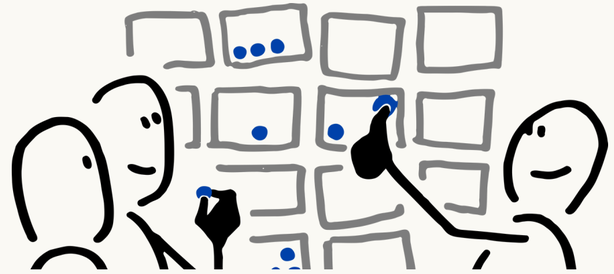
Supervised
Binary
Classifier Model

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



Titles

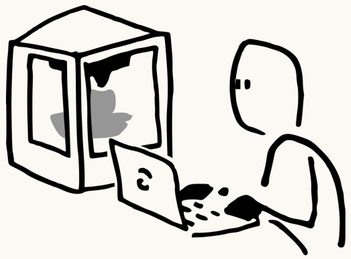
Abstracts



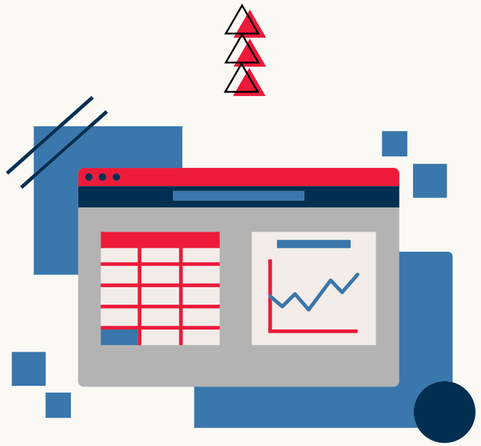
Self-Supervised
Multi-Label
Classifier Model

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

The
Project Design

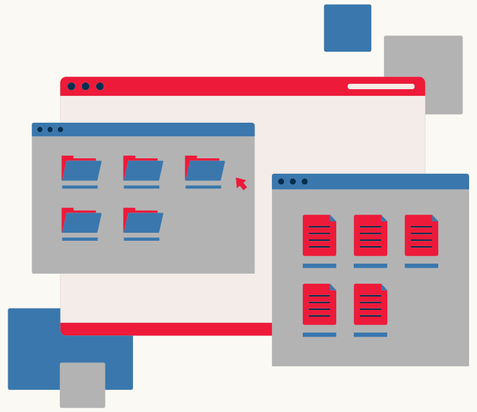


Medical Researcher



High-quality PPS-relevant literature

Large Search String on
Patient Preference Studies



PPS-relevant and irrelevant content across clinical areas

Supervised
Binary
Classifier Model

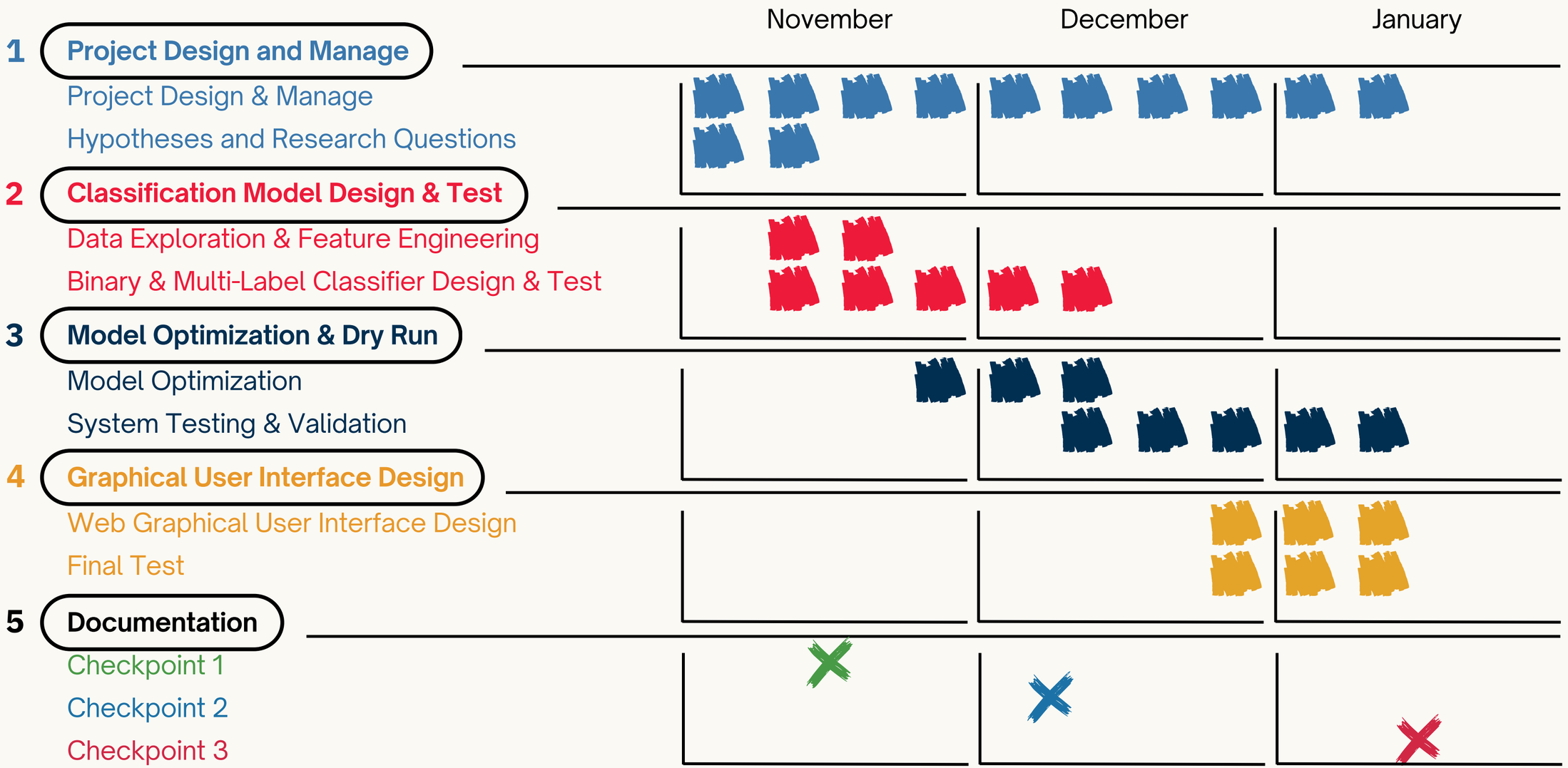
to classify search string outputs by relevance to PPS

Self-Supervised
Multi-Label
Classifier Model

to categorize PPS-relevant content into Clinical Areas

Project

Work Breakdown Structure



Part 2

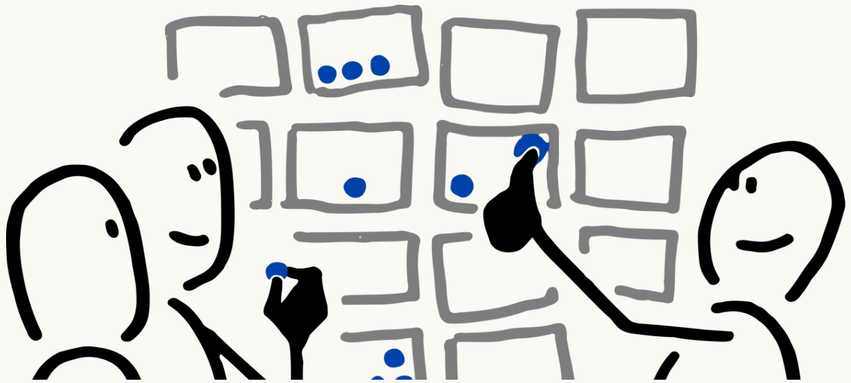
Table of Contents

The
Picture

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

- EU Commission

The
Project Objective



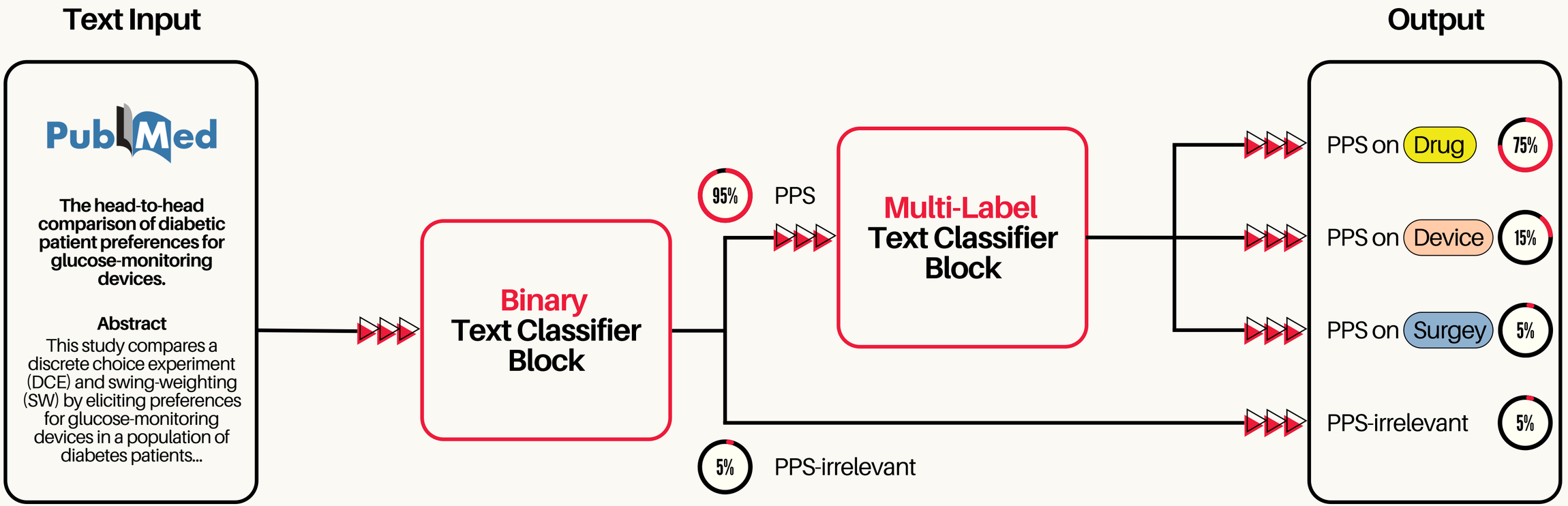
Classifier Model for Scientific Papers

- Detect **Patient Preference Studies** from PubMed outputs
- Categorize PPS-relevant papers into **Clinical Areas**

by **Titles** and **Abstracts**

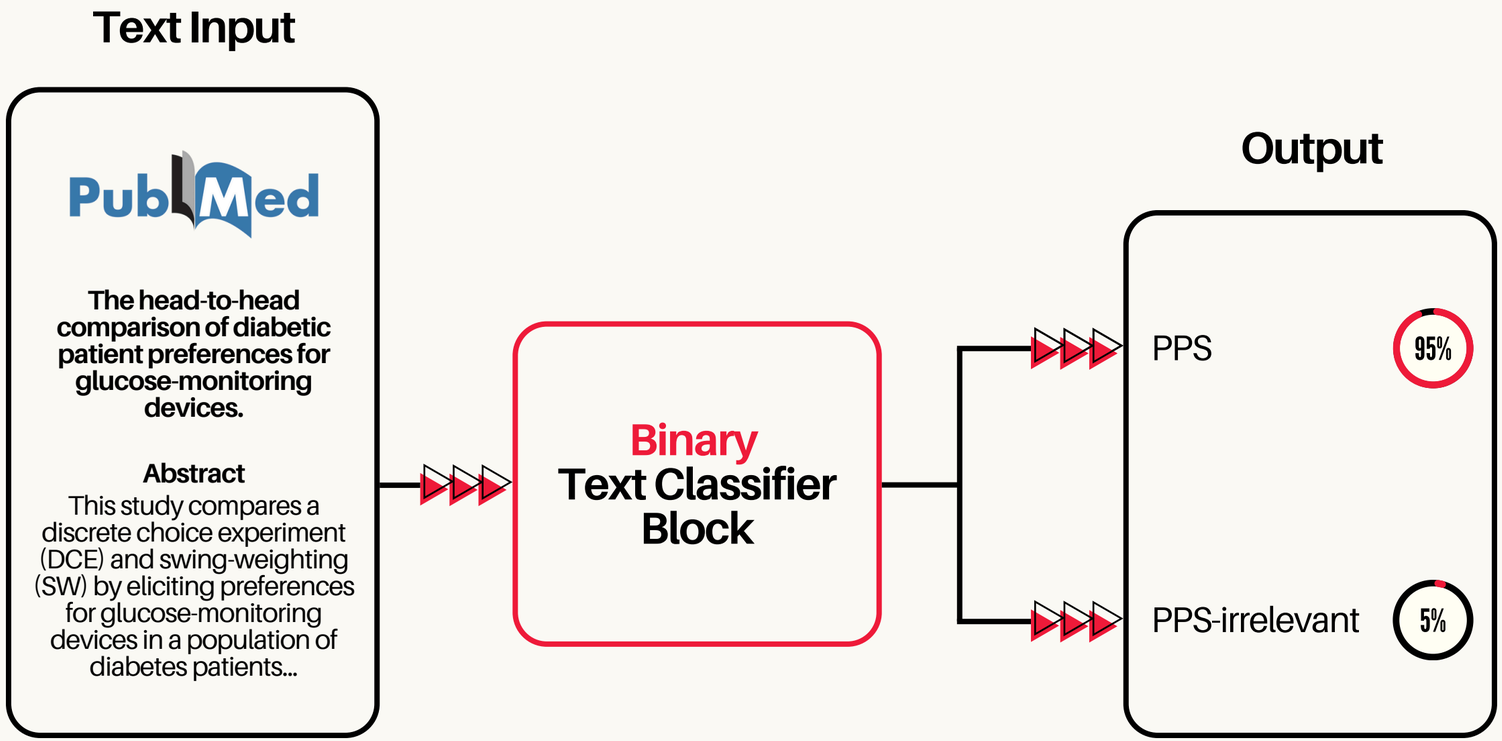
The Task

Two-Stage Text Classification



The **Binary** Text Classifier

Detect **Patient Preference Studies** from PubMed outputs by **Titles** and **Abstracts**

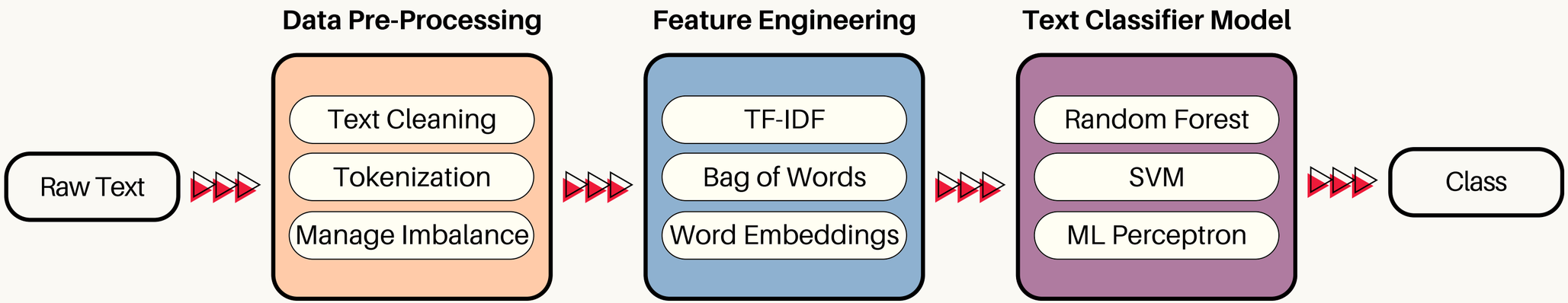


The **Binary** Classifier Methods

Binary
Text Classifier Model

Machine Learning Methods

Low-performance, but deployment requires simple training on a small amount of data via **CPU**

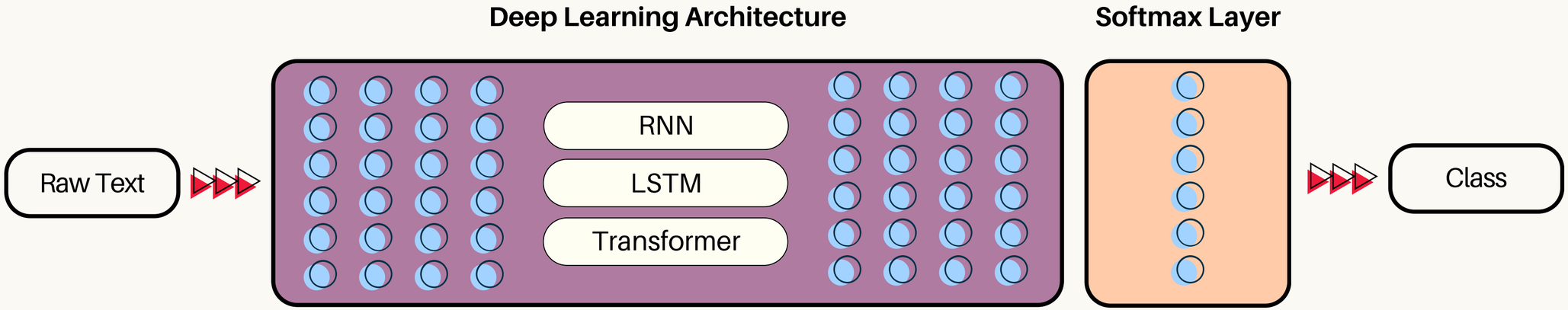


The **Binary** Classifier Methods

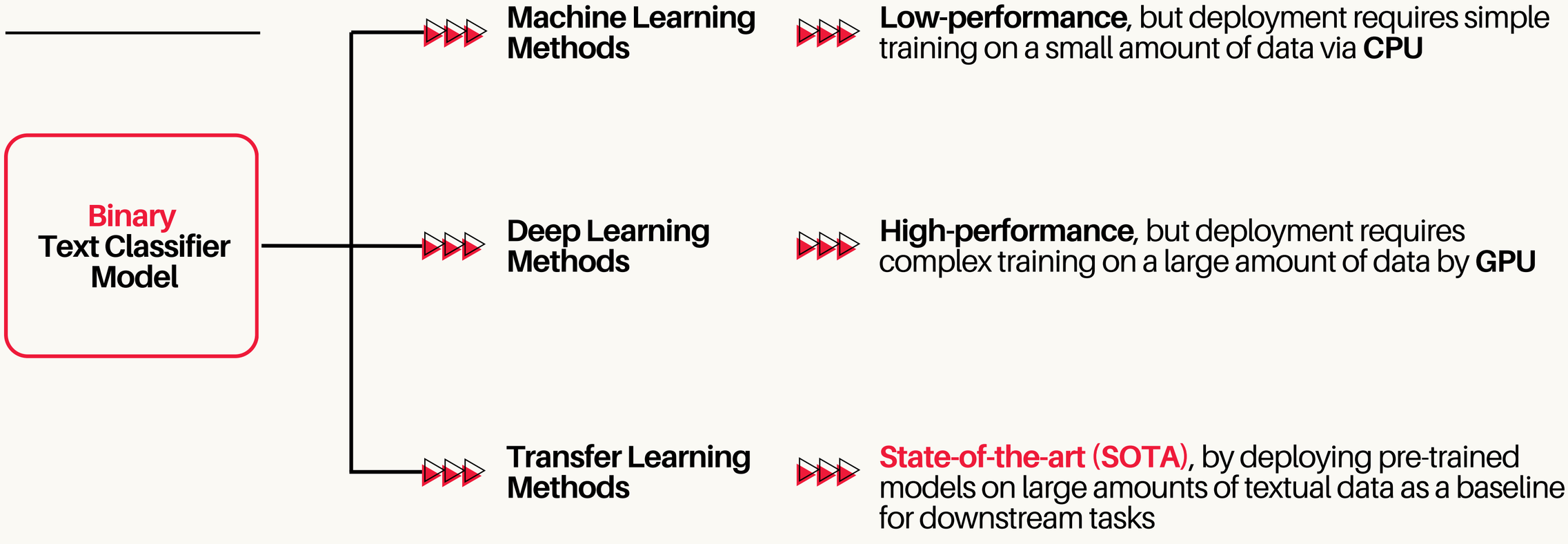
Binary
Text Classifier Model

Machine Learning Methods → **Low-performance**, but deployment requires simple training on a small amount of data via **CPU**

Deep Learning Methods → **High-performance**, but deployment requires complex training on a large amount of data by **GPU**

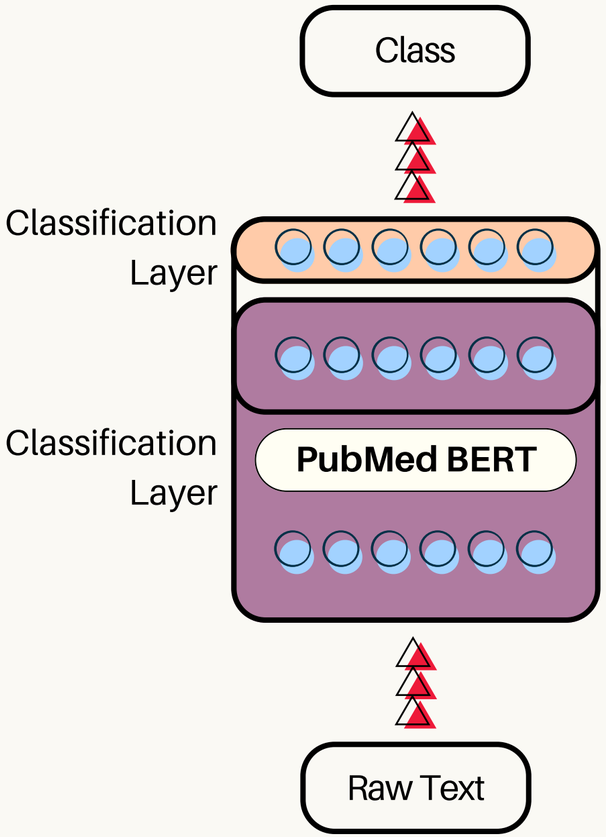


The **Binary** Classifier Methods

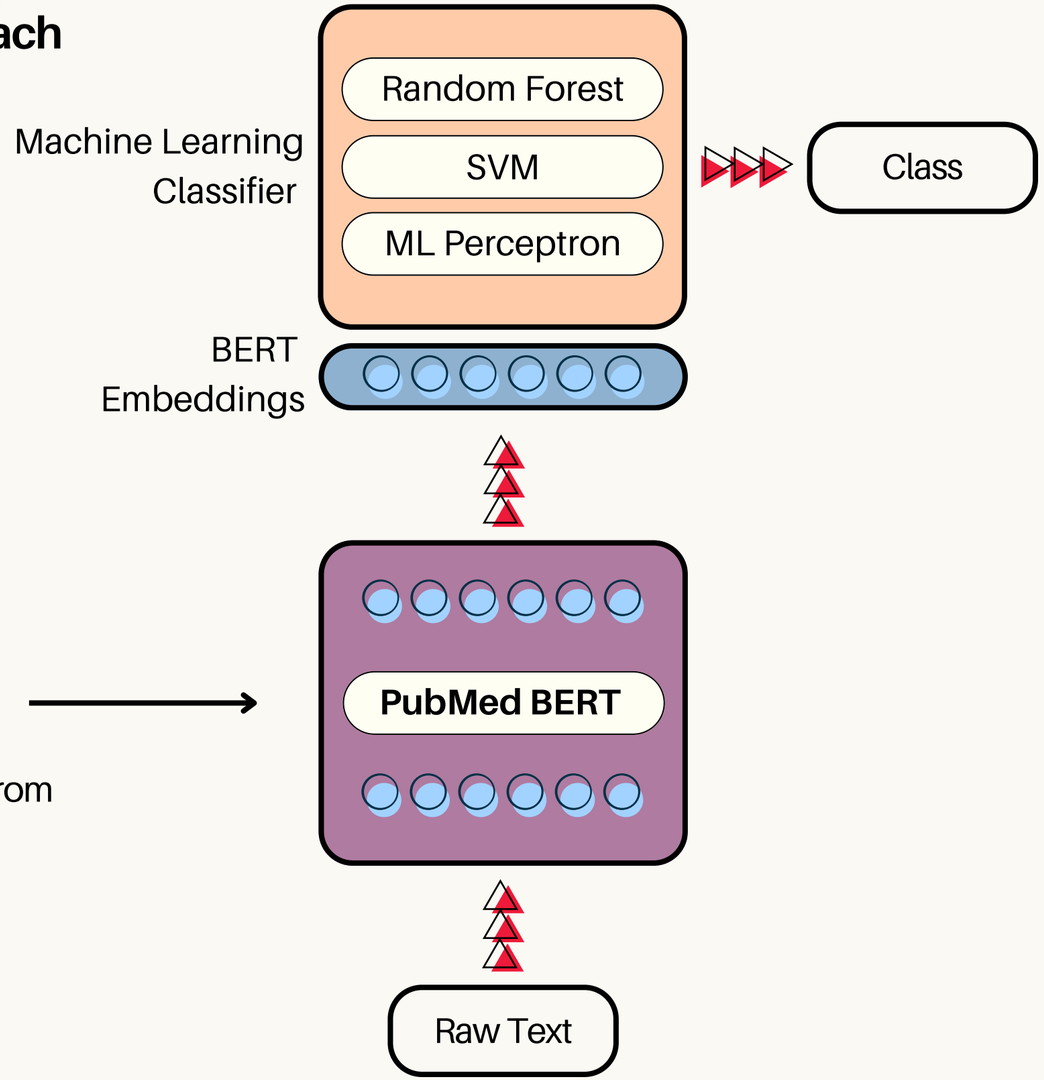


The Choices

The Standard Approach



The Hybrid Approach

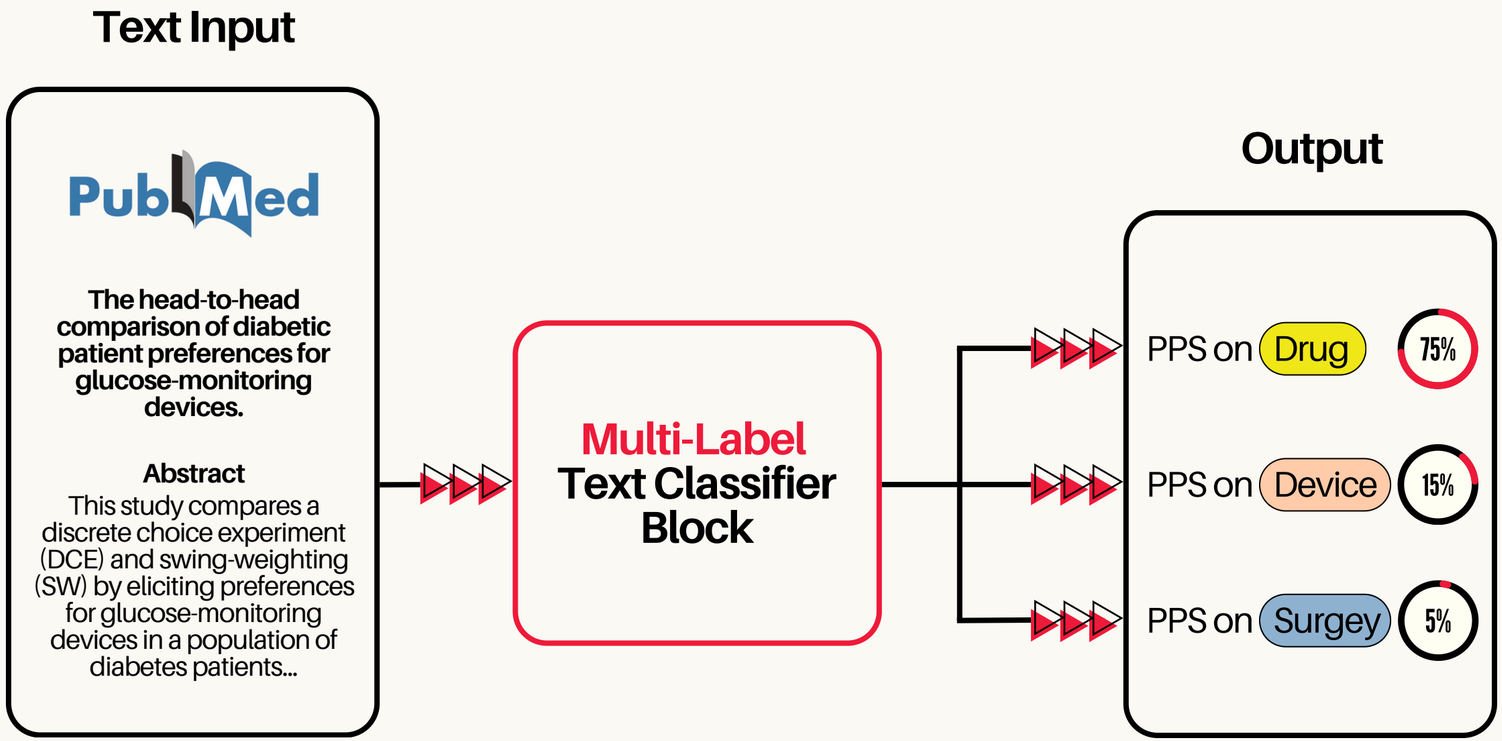


Pre-Trained BERT Model

from scratch on abstract from PubMed

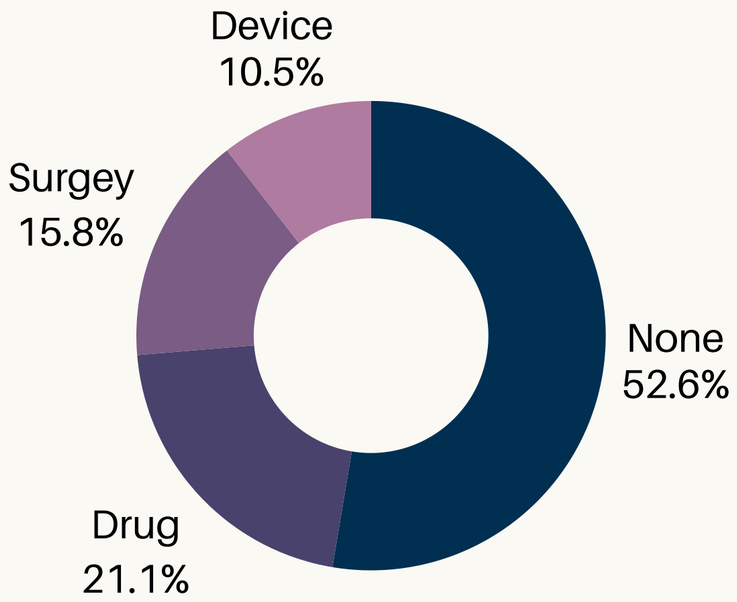
The **Multi-Label** Text Classifier

Categorize PPS-relevant papers into **Clinical Areas** by **Titles** and **Abstracts**



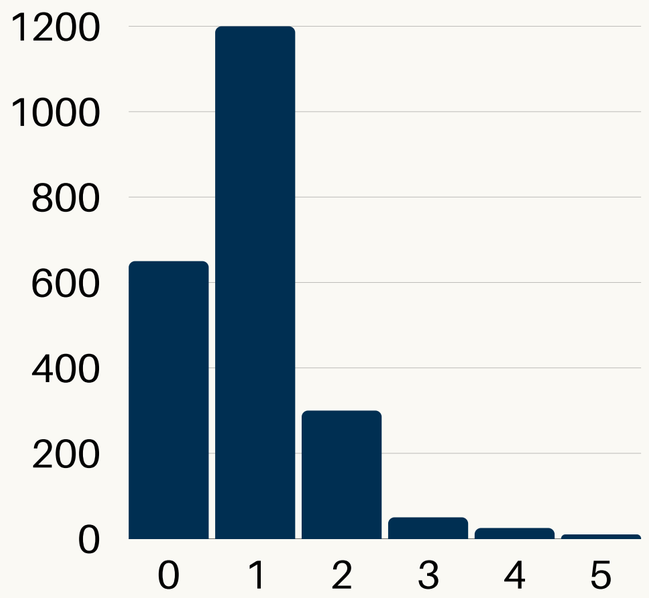
The **Multi-Label** Classifier Problem

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

BERTopic

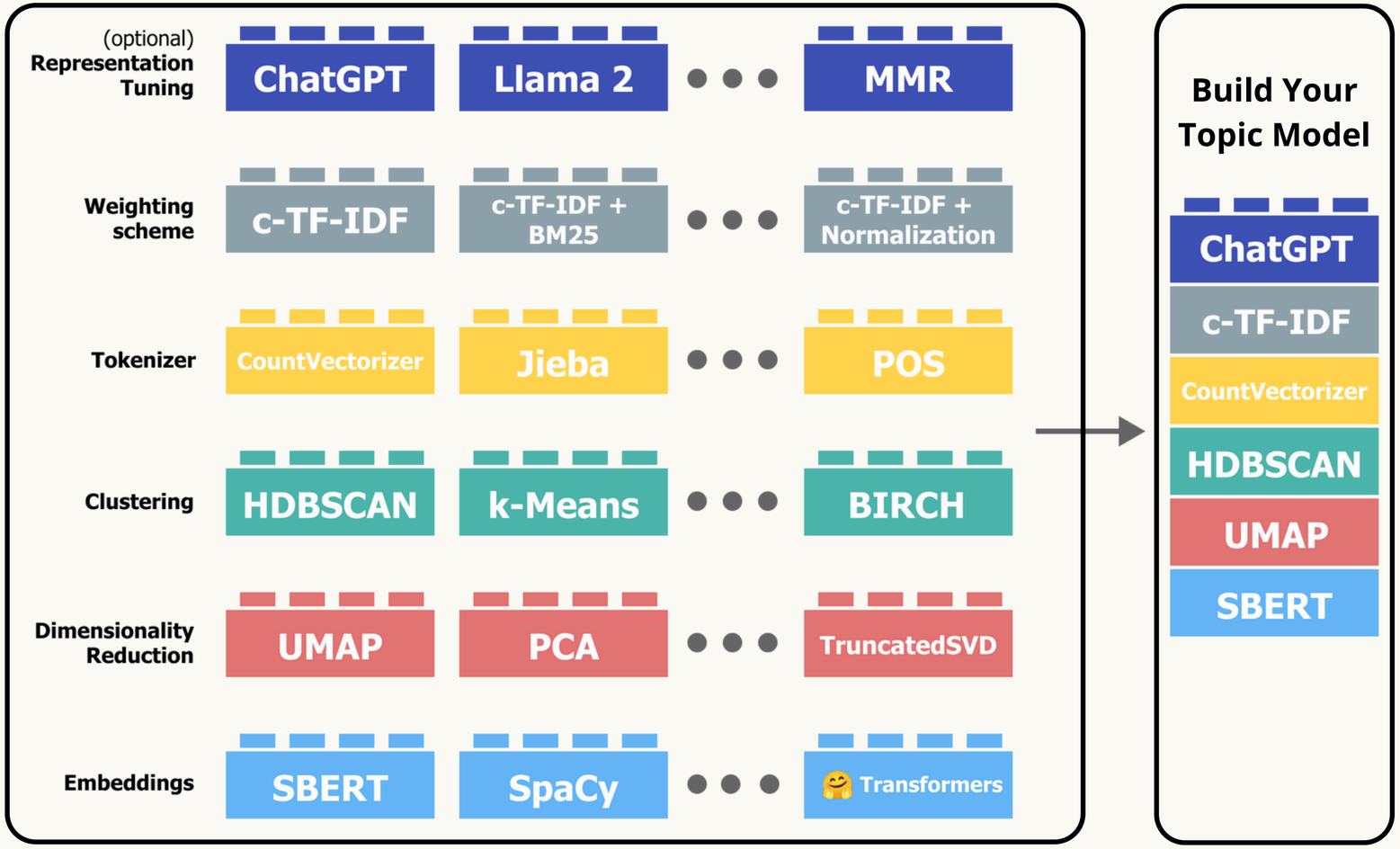
Problem Adaptation

Algorithm Adaptation

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

Topic Modelling

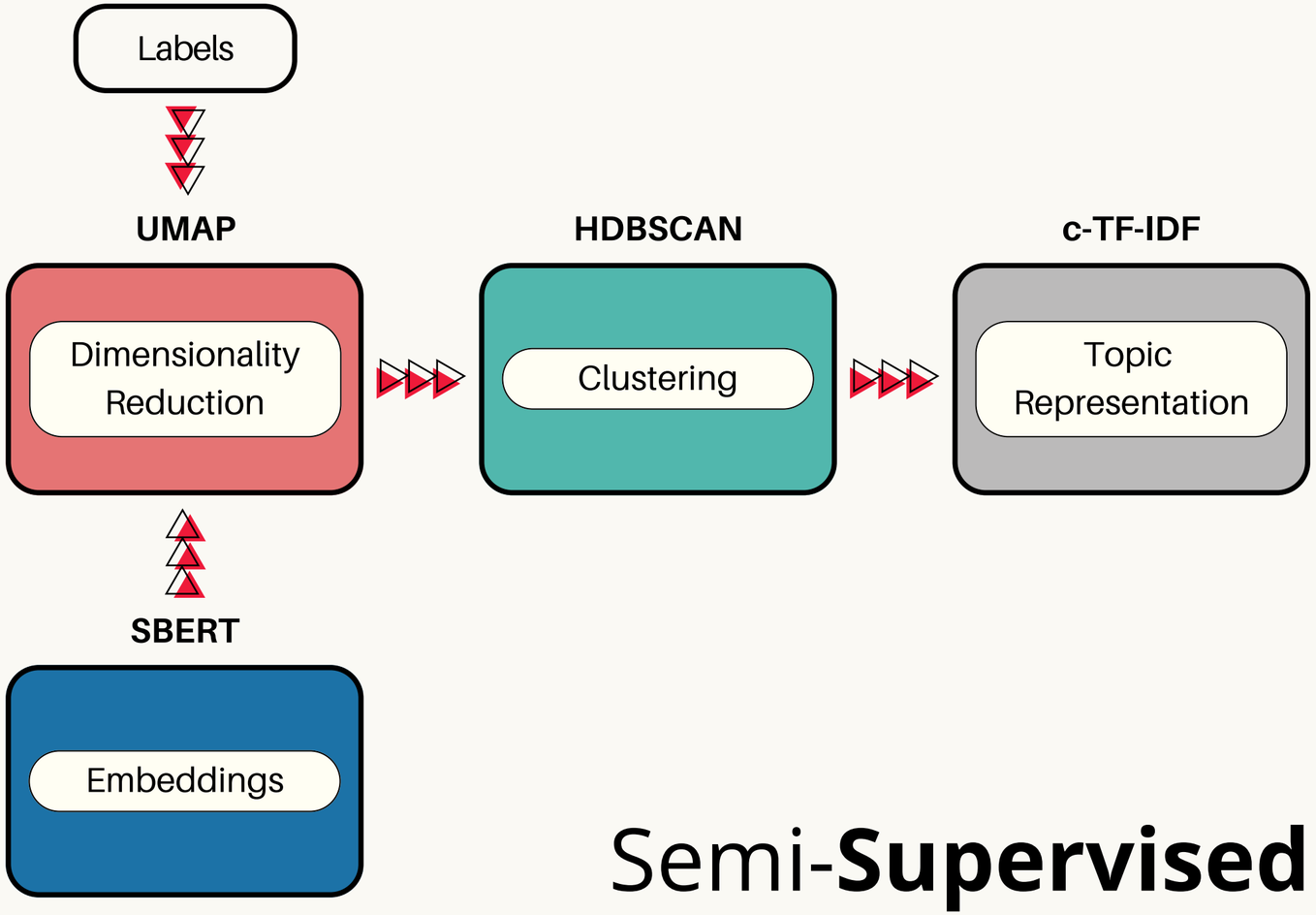
- LDA
- NMF
- Top2Vec
- **Bertopic**



Multi-Label
Text Classification Problem

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



Semi-**Supervised**

Multi-Label
Text Classification Problem

BERTopic
Project Benefits

Topic Exploration

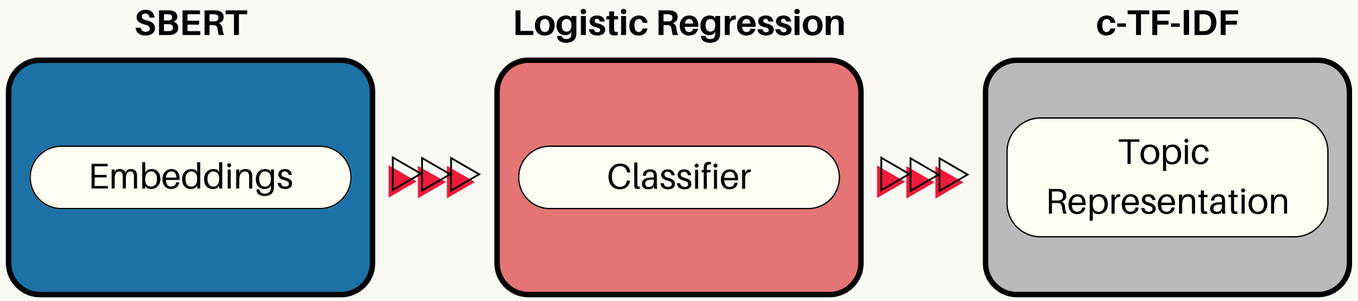
Categories Not Fixed

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

- F1-Score
- Recall
- Precision
- Subset Accuracy
- Hamming Loss

Ranking-Based

- Coverage
- One Error
- Ranking Loss

Label-Based

- Micro Average
- Macro Average

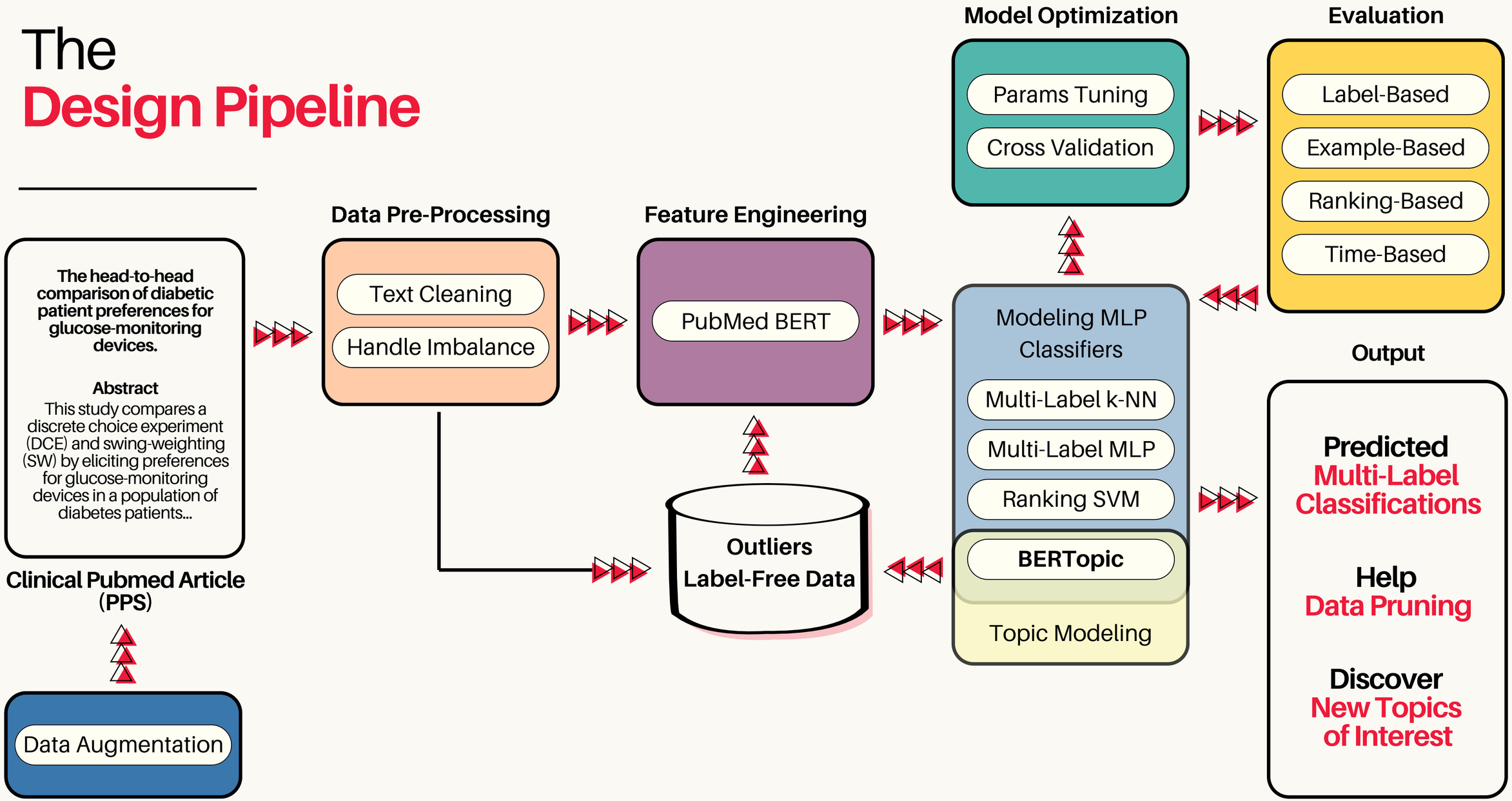
Topic Modelling

- Diversity
- Coherence

System Time

- Train Time
- Inference Time
- Latency
- Throughput
- Scalability

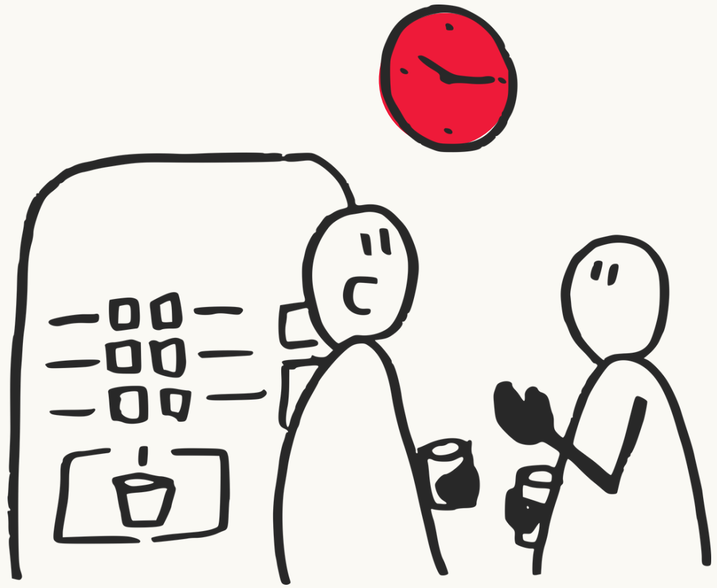
The Design Pipeline



APPLIED DATA SCIENCE PROJECT

Thank You

Francesco Giuseppe Gillio & Cesar Augusto Seminario Yrigoyen



UNIVERSITÀ DI TORINO

Politecnico di Torino