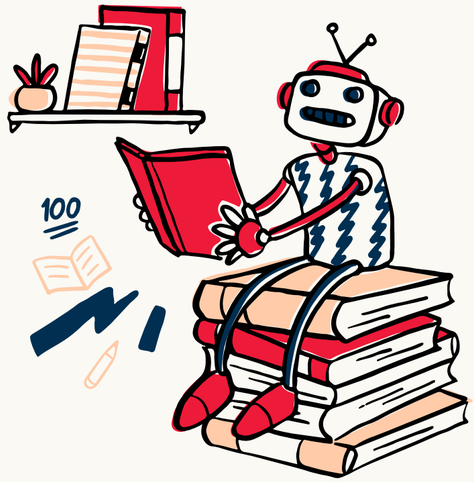
APPLIED DATA SCIENCE PROJECT

Patient Preference Studies
Classification System



Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

UNIVERSITÀ DI TORINO

Politecnico di Torino

Part 1
Table of Contents

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

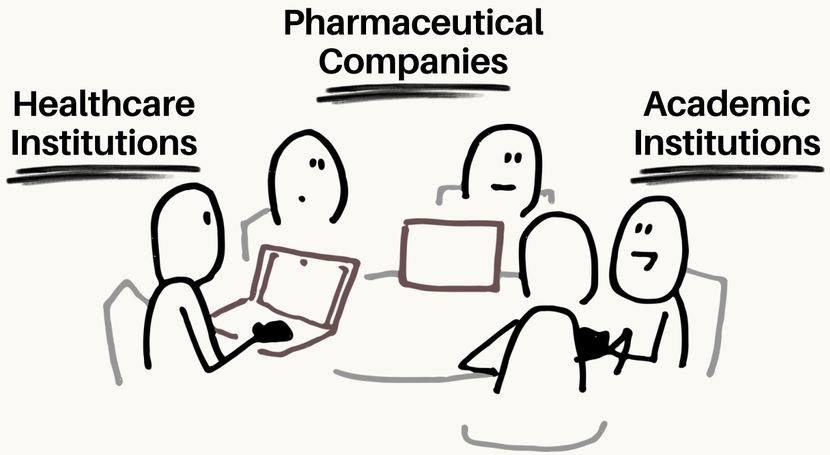
The **Stakeholders**

Project
Background

Medical Researchers

CORE

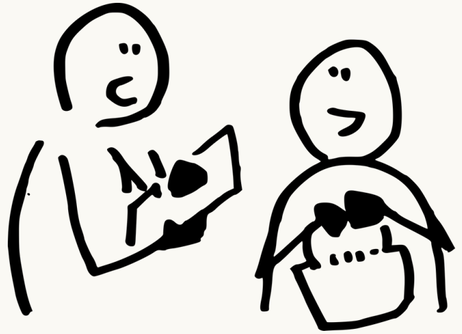
actionable insights to



Healthcare Ecosystem

DIRECT

valuable services to



Patient Communities

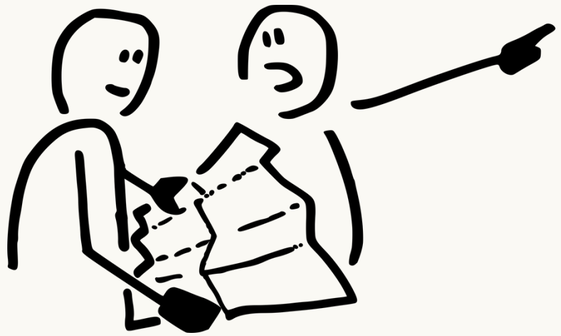
INDIRECT

The **Challenges**

Large Volume

of Scientific Literature

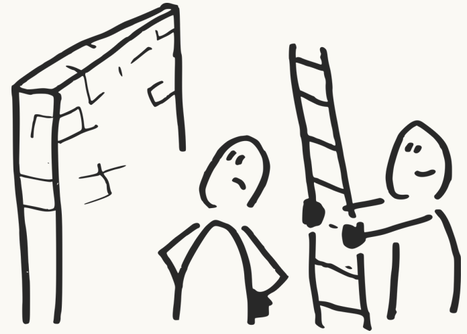
search strings in citation databases (**PubMed**) return a large amount of content, often irrelevant



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

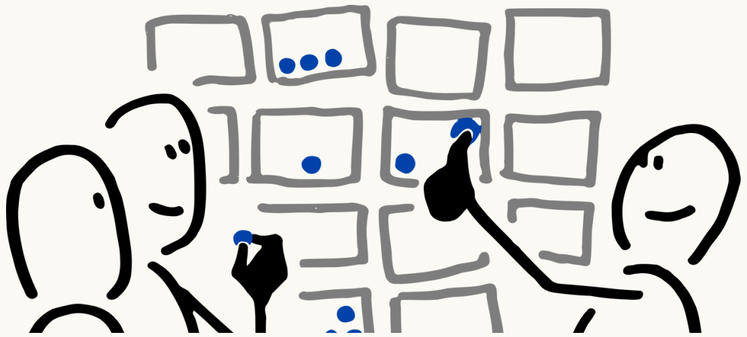
The Values

Project
Value Proposition

Improve Relevance

by Classification System

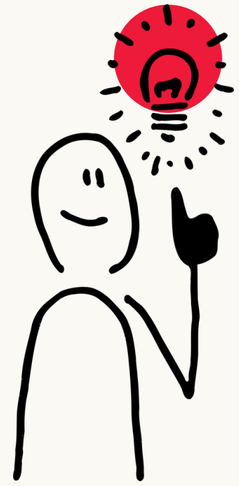
to bypass irrelevant content and return high-value literature



Improve Retrieval

by Categorization System

to categorize search results and improve area-specific retrieval



Improve Efficiency

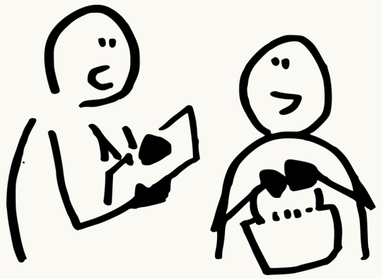
by System Automation

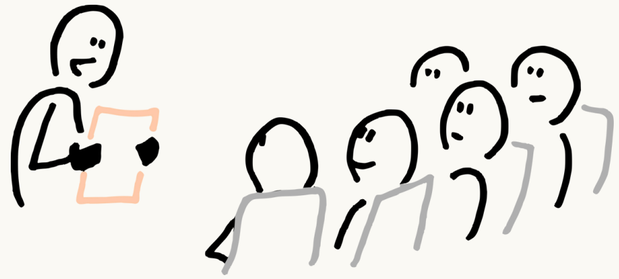
to reduce manual effort and improve access to up-to-date research

The Values

Project
Value Proposition

United Nations
Sustainable Development Goals
2030





3 GOOD HEALTH AND WELL-BEING

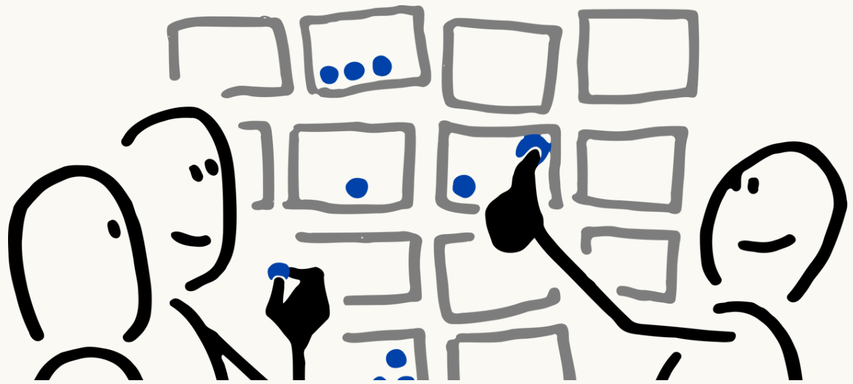
Support the advancement of
patient care systems and processes

Support the advancement of
information retrieval in medical research

4 QUALITY EDUCATION

The **Objective**

Project
General Objectives



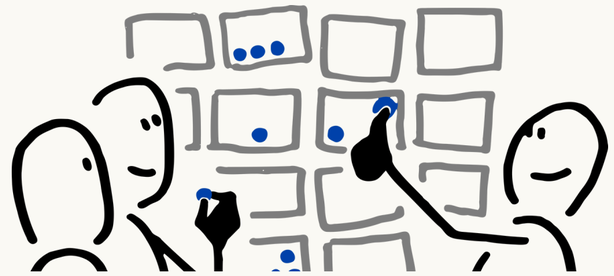
Classifier Model for Medical Research Papers

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

The **Objective**

Project
General Objectives

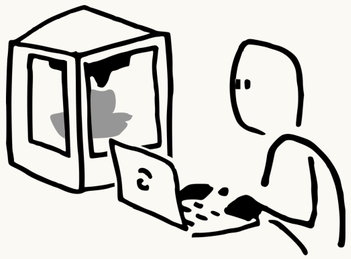
Supervised
Binary
Classifier Model



Self-Supervised
Multi-Label
Classifier Model

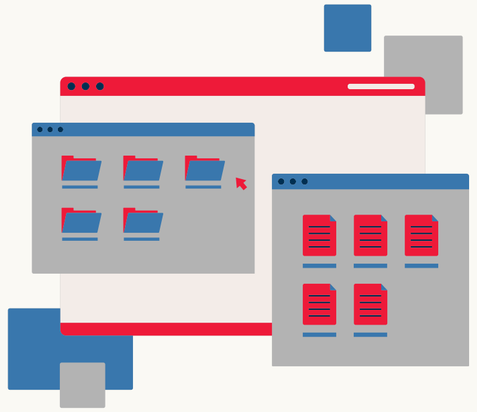
The Design

Project
Design



Medical Researcher

Large Search String on
Patient Preference Studies



PPS-relevant and irrelevant
content across clinical areas

High-quality PPS-relevant literature

Self-Supervised
Multi-Label
Classifier Model

to categorize PPS-relevant
content into Clinical Areas

Supervised
Binary
Classifier Model

to classify search string outputs
by relevance to PPS

The **Work Breakdown Structure**

Project
Work Breakdown Structure

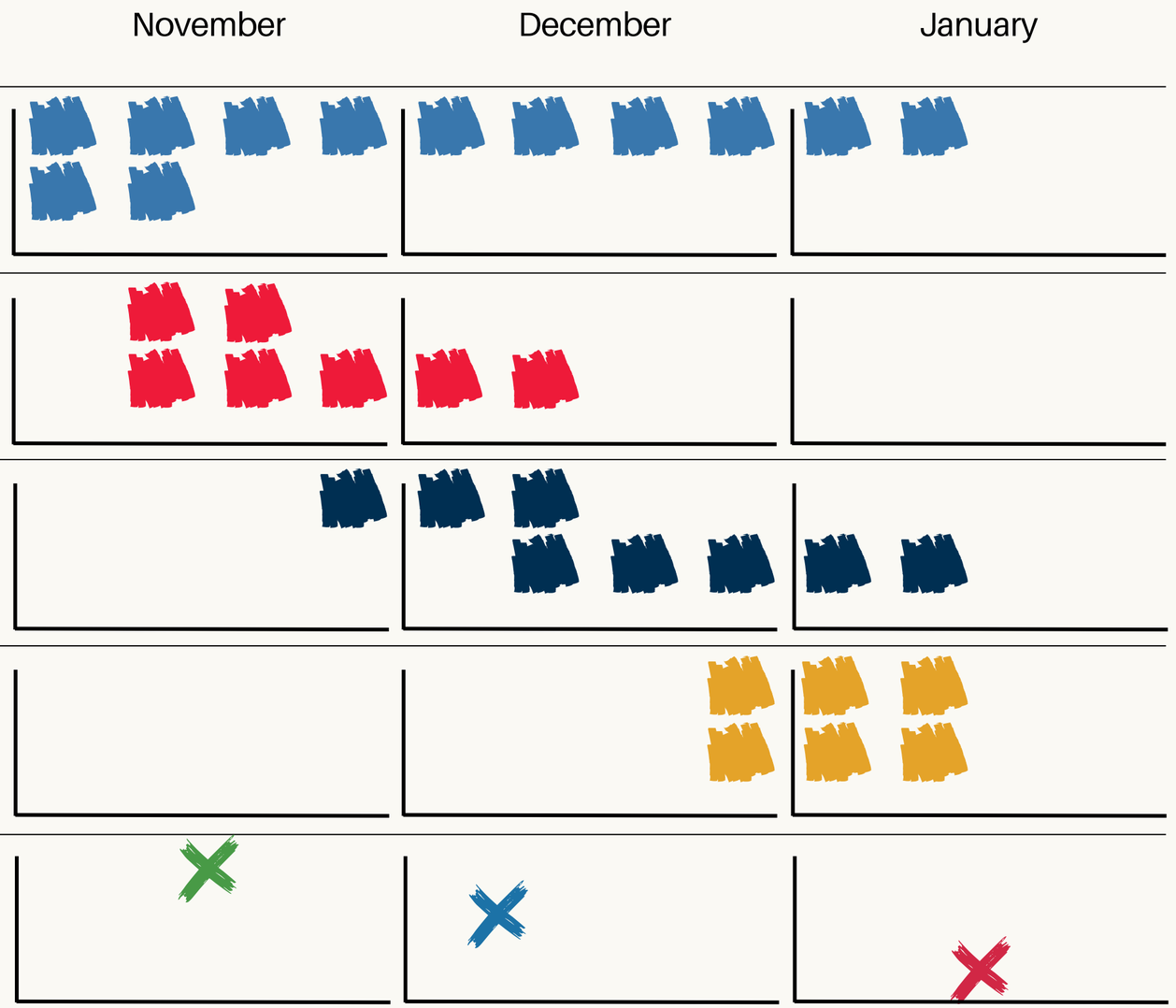
1 **Project Design and Manage**
- Project Design & Manage
- Hypotheses and Research Questions

2 **Classification Model Design & Test**
- Data Exploration & Feature Engineering
- Binary & Multi-Label Classifier Design & Test

3 **Model Optimization & Dry Run**
- Model Optimization
- System Testing & Validation

4 **Graphical User Interface Design**
- Web Graphical User Interface Design
- Final Test

5 **Documentation**
- Checkpoint 1
- Checkpoint 2
- Checkpoint 3



Part 2

Table of Contents

The **Picture**

Project
Picture

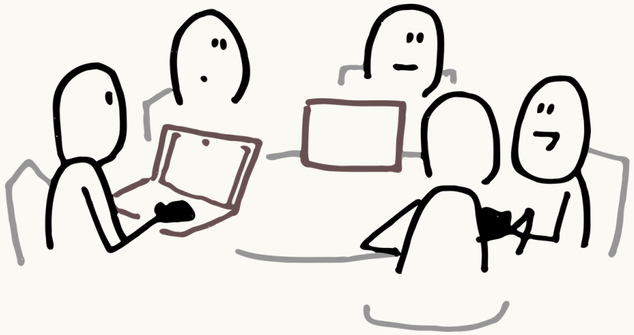
European Commission

"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"

- European Commission

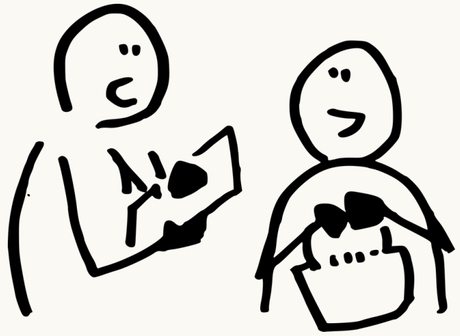
Medical Researchers

actionable insights to



Healthcare Ecosystem

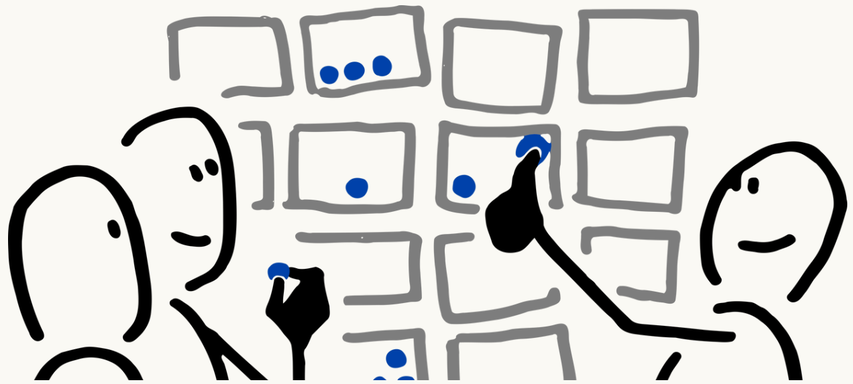
valuable services to



Patient Communities

The **Objective**

Project
General Objectives



Classifier Model for Medical Research Papers

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

The **Task**

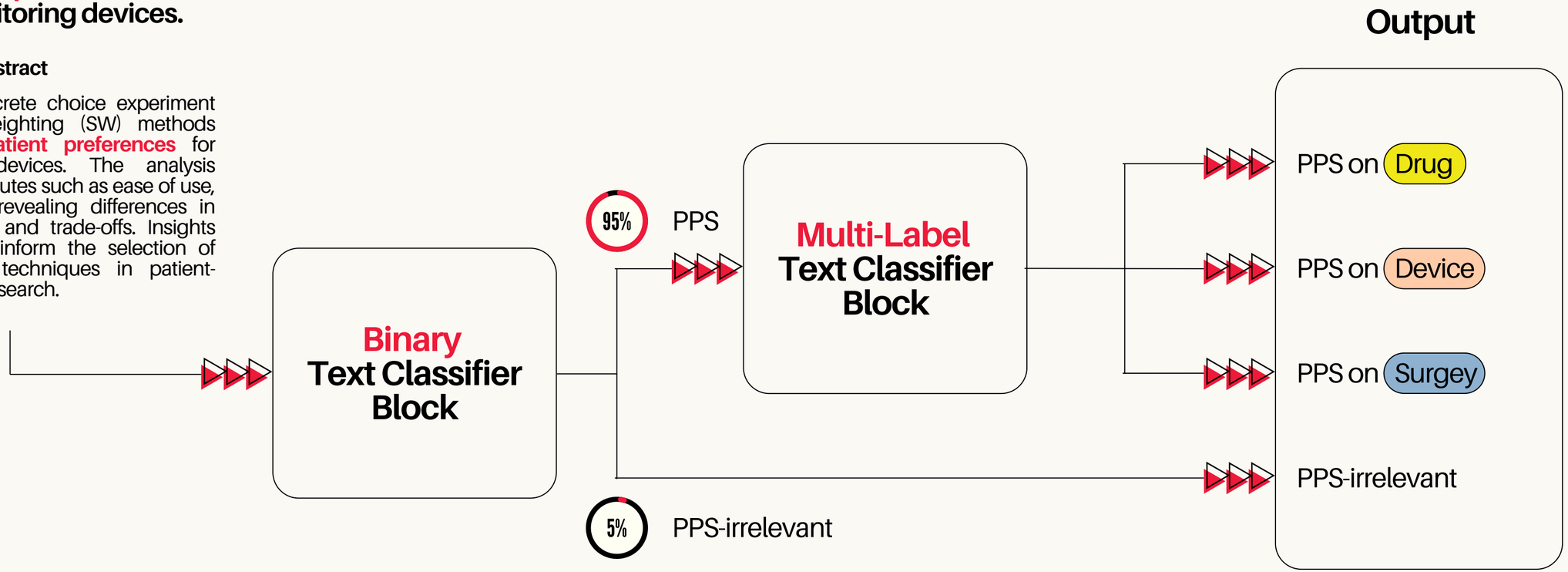
Project
Objectives

Two-Stage
Text Classification

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** Text Classifier

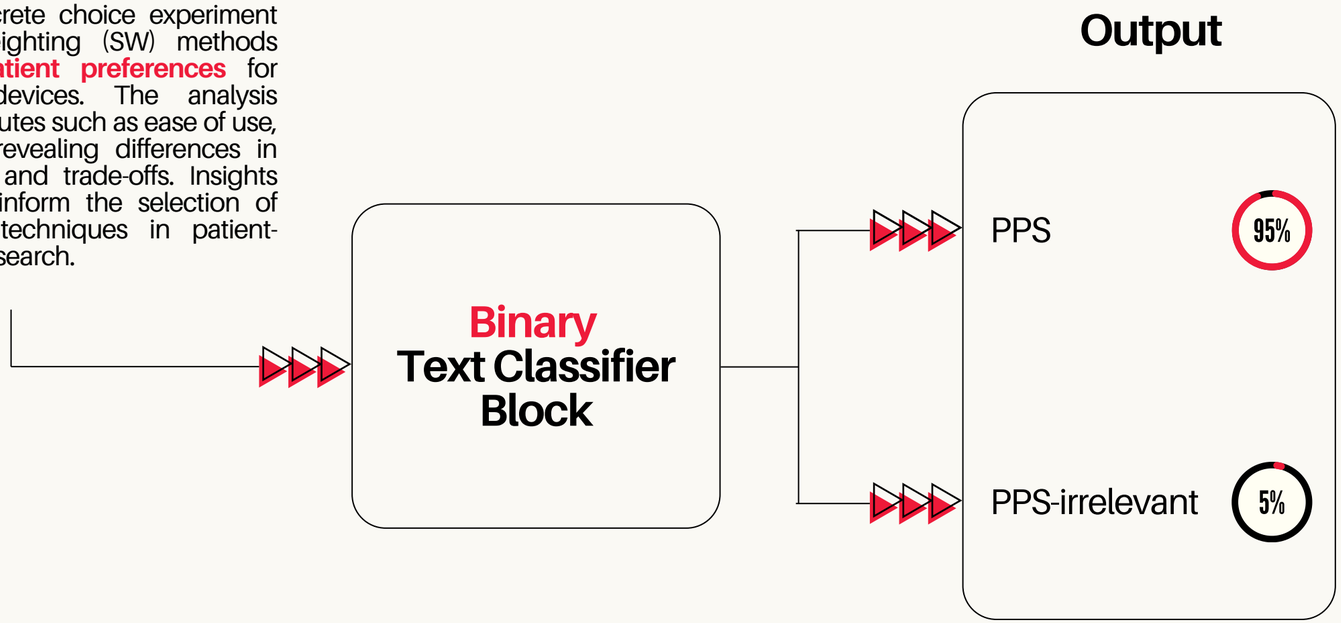
Binary
Text Classification Problem

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.



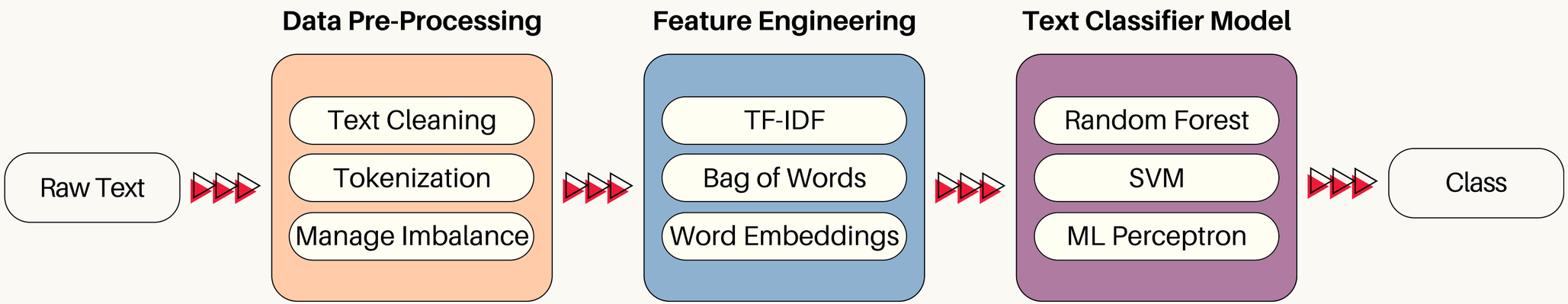
The **Binary** Classifier Methods

Binary
Text Classification Problem

Binary
Text Classifier Model

Machine Learning Methods

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



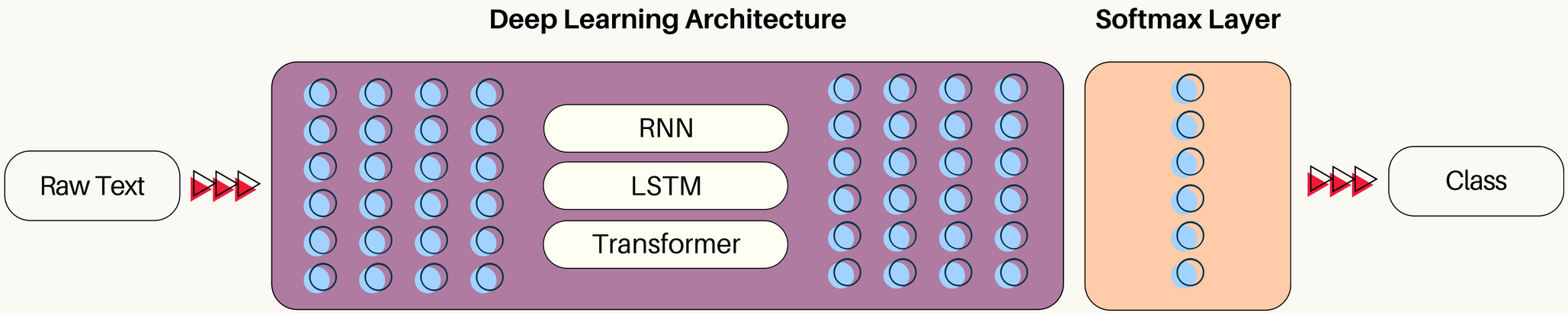
The **Binary** Classifier Methods

Binary
Text Classification Problem

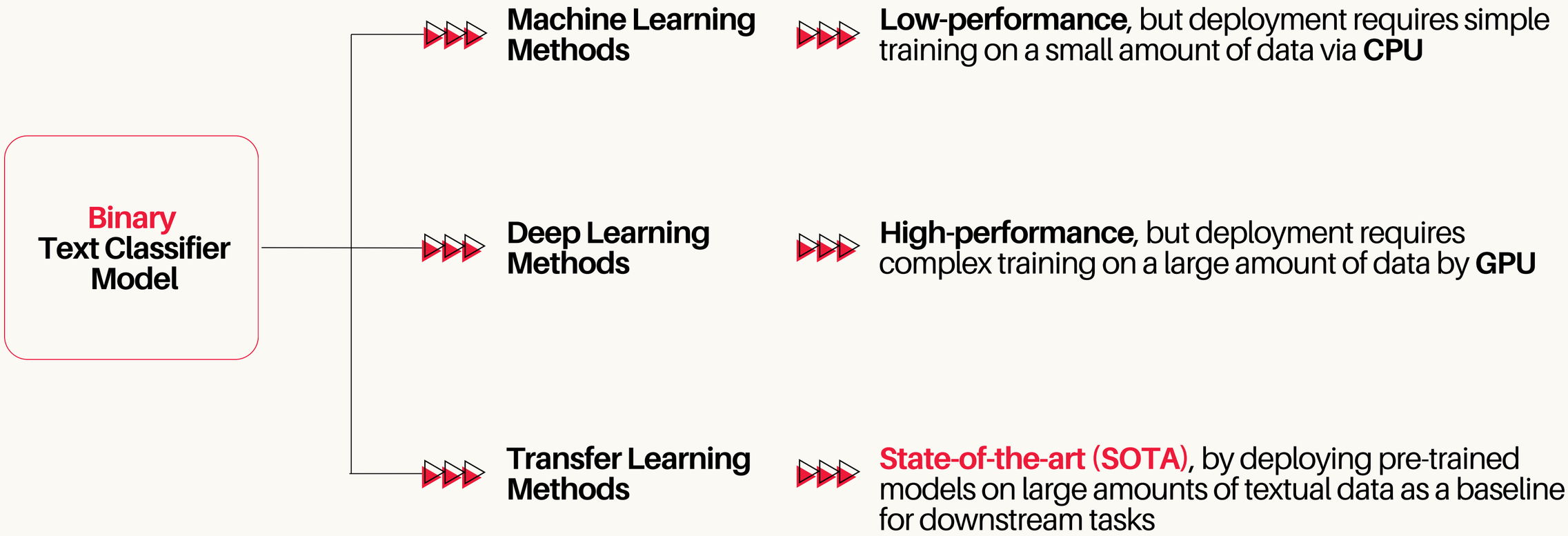
Binary
Text Classifier Model

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**

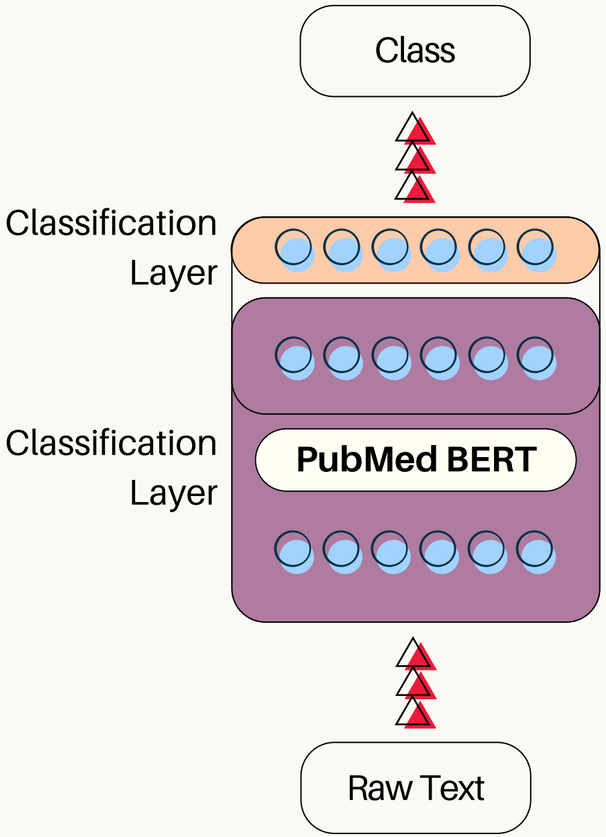


The **Binary** Classifier Methods

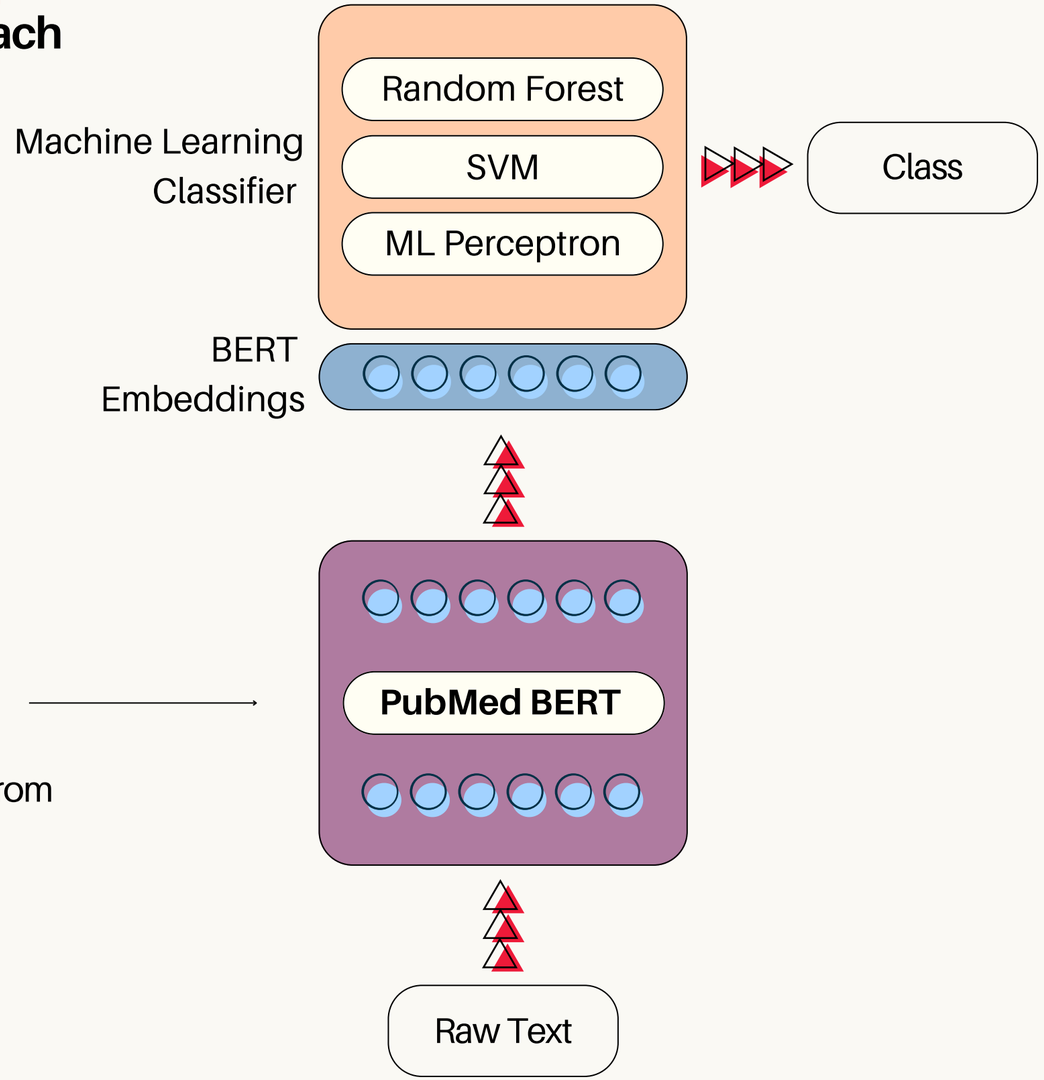


The Choices

The Standard Approach



The Hybrid Approach



Pre-Trained BERT Model

from scratch on abstract from PubMed

The **Multi-Label** Text Classifier

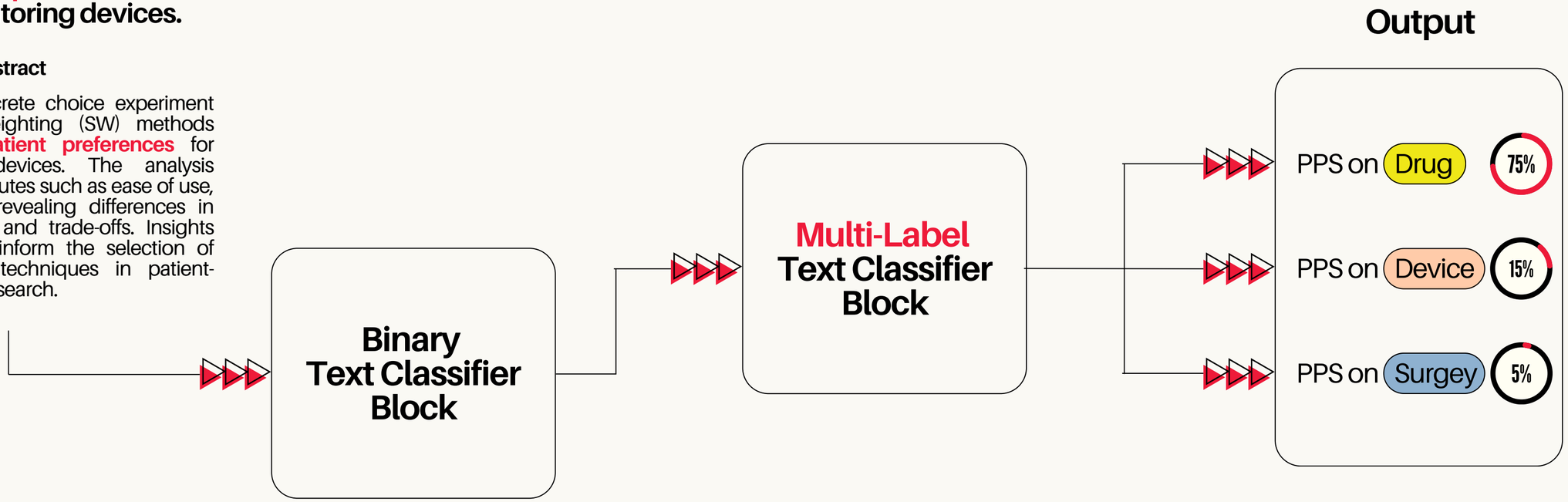
Multi-Label
Text Classification Problem

Classification by relevance to Clinical Areas

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

Abstract

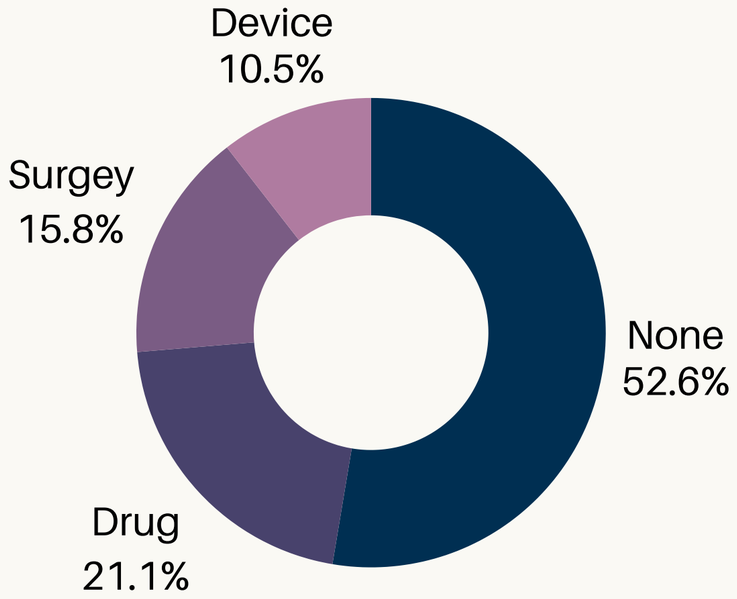
A comparison of discrete choice experiment (DCE) and swing-weighting (SW) methods assesses diabetic **patient preferences** for glucose-monitoring devices. The analysis highlights critical attributes such as ease of use, accuracy, and cost, revealing differences in attribute prioritization and trade-offs.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



The **Multi-Label** Classifier Problem

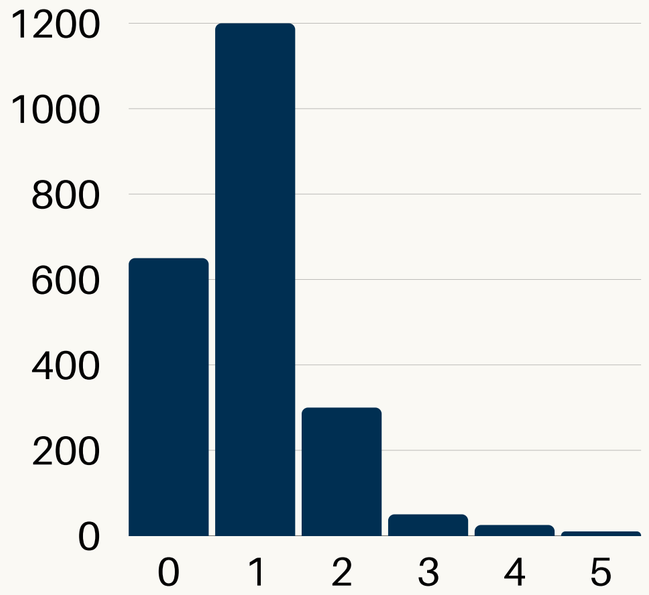
Multi-Label
Text Classification Problem

Class Distribution



Data Imbalance

Multi-Label Distribution



The **Hypotheses** and Research Questions

Multi-Label
Text Classification Problem

"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

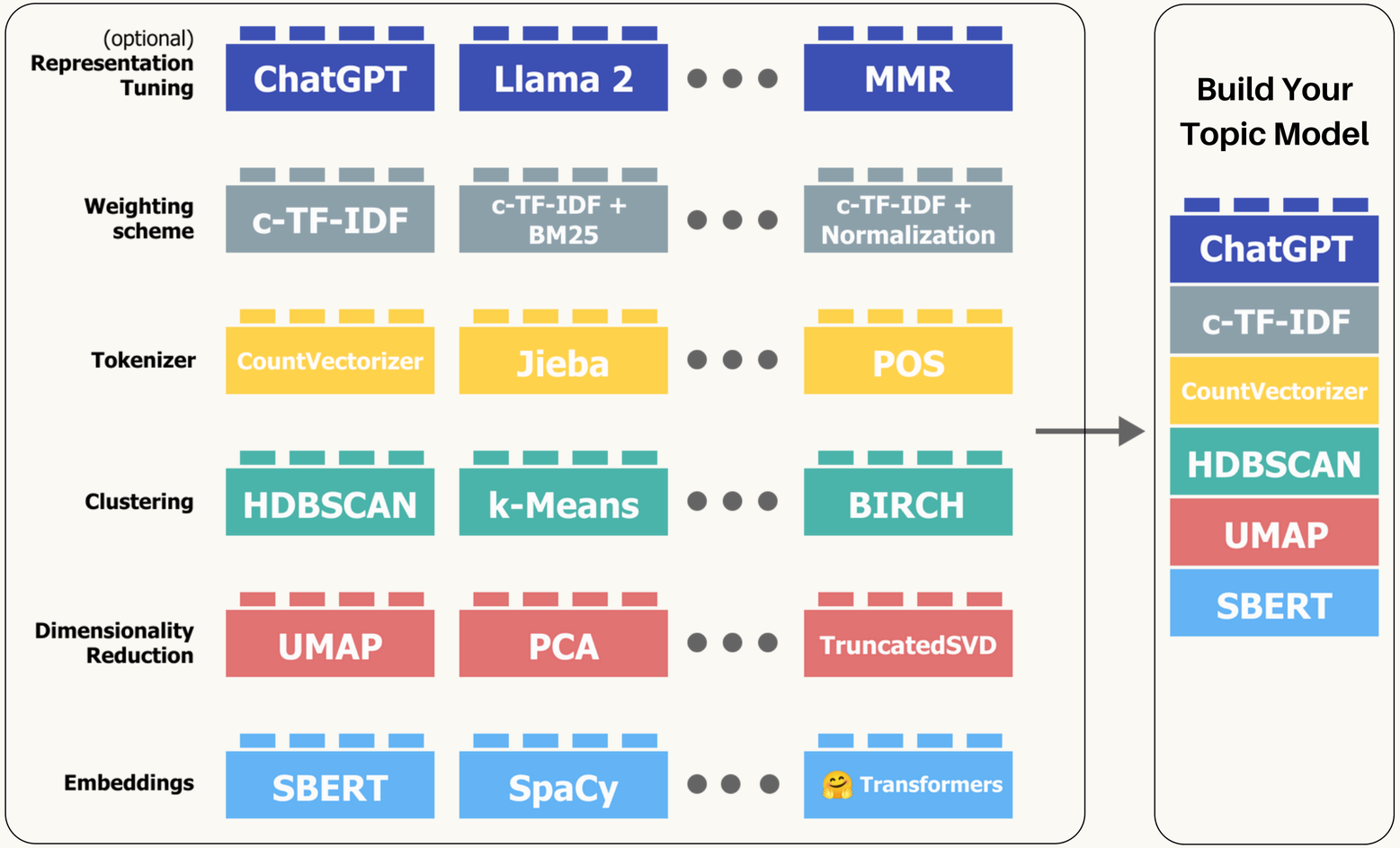
Problem Adaptation

Algorithm Adaptation

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

Topic Modelling

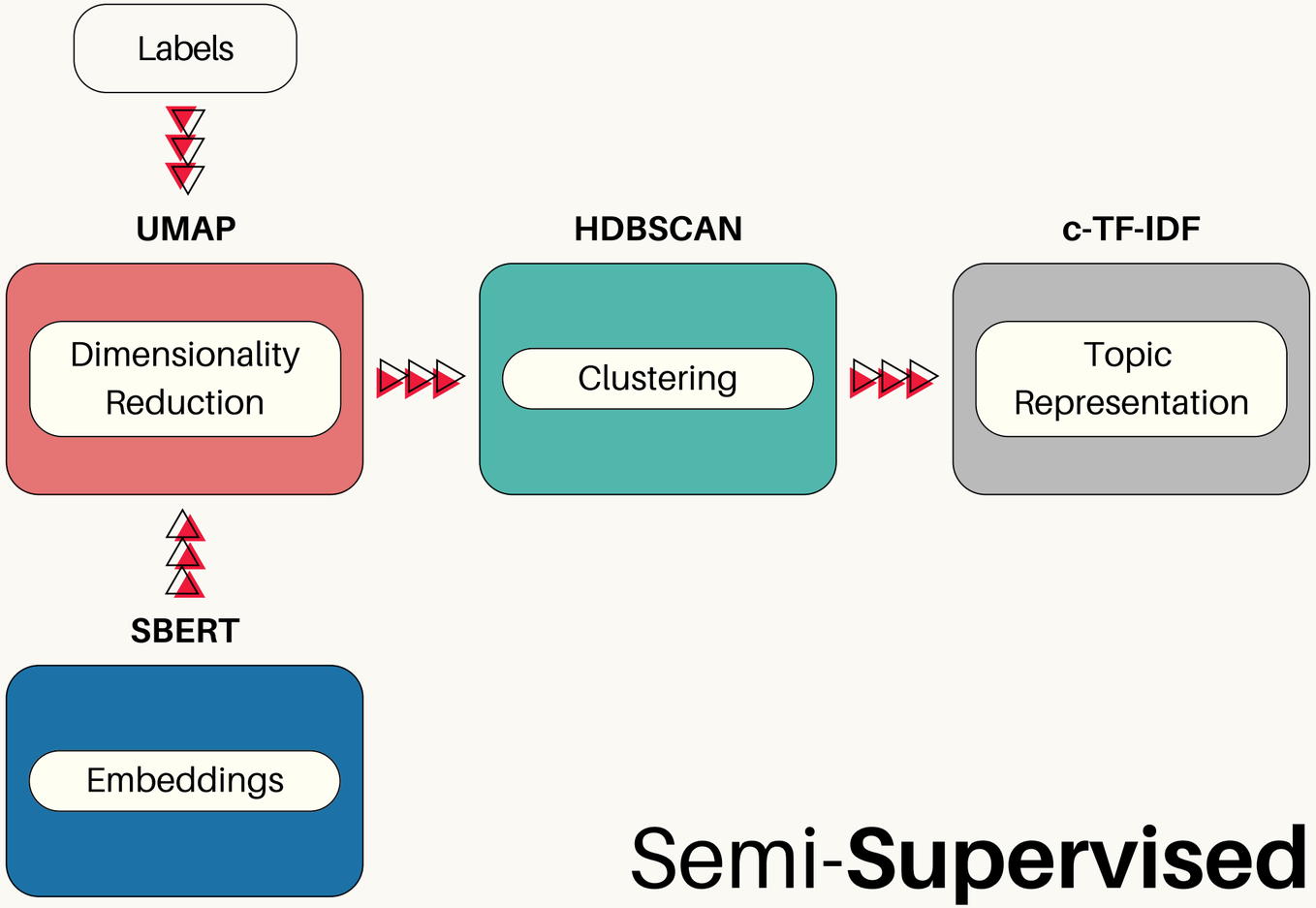
- LDA
- NMF
- Top2Vec
- **Bertopic**



BERTopic Project Benefits

Multi-Label
Text Classification Problem

- **Topic Exploration**
- **Categories Not Fixed**
- **Improve & Speed Up Manual Labeling**
  - Binary Classifier
  - Multi-Label Classifier
- **Classification Approaches:**
  - Unsupervised
  - Semi-Supervised
  - Supervised: Classification



Semi-**Supervised**

BERTopic Project Benefits

Multi-Label
Text Classification Problem

Topic Exploration

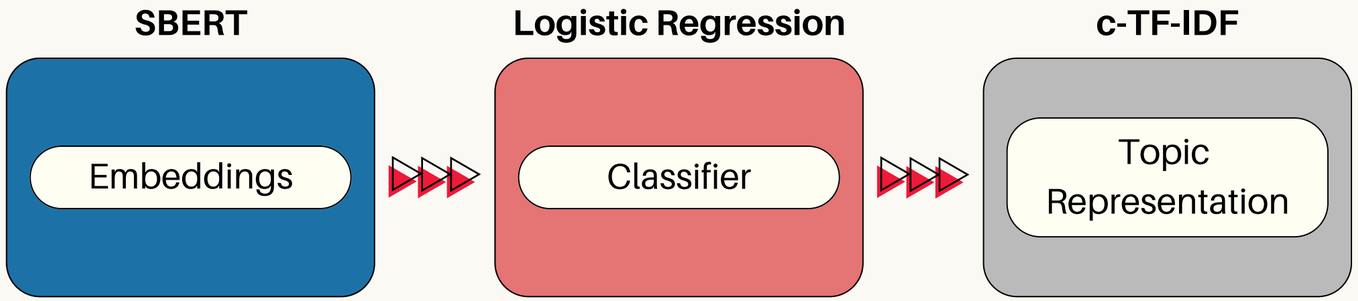
Categories Not Fixed

Improve & Speed Up Manual Labeling

- Binary Classifier
- Multi-Label Classifier

Classification Approaches:

- Unsupervised
- Semi-Supervised
- Supervised: Classification



Supervised

The **Evaluation** Metrics

Multi-Label
Text Classification Problem

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

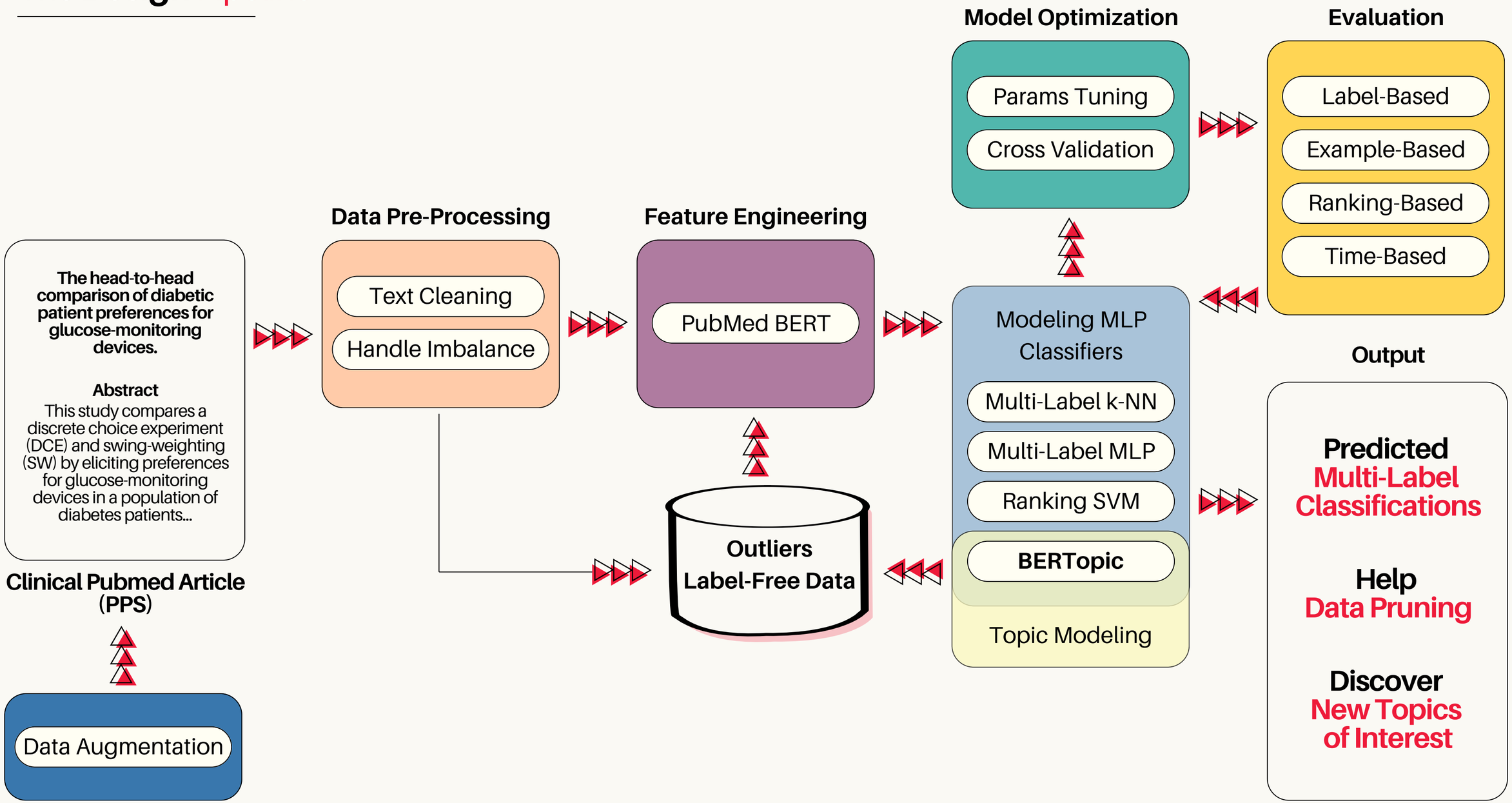
Topic Modelling

- Diversity
- Coherence

System Time

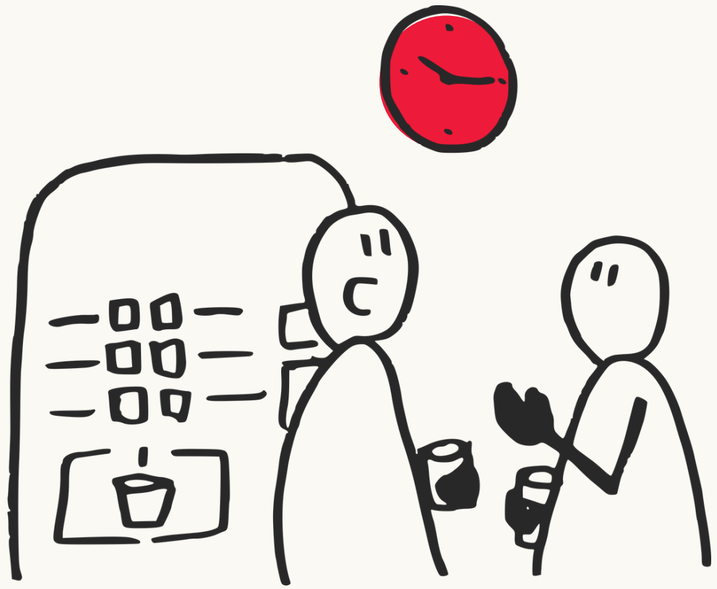
- Train Time
- Inference Time
- Latency
- Throughput
- Scalability

The **Design** Pipeline



APPLIED DATA SCIENCE PROJECT

Thank You



Cesar Augusto Seminario Yrigoyen
Francesco Giuseppe Gillio

UNIVERSITÀ DI TORINO

Politecnico di Torino