

Recommender System for Study Build: Using the Protocol to Select CRFs from Standard Libraries

March 2023

Komal Govil & Karan Pande



Protocol



Case Report
Forms

Recommender Systems

Collaborative Filtering Model



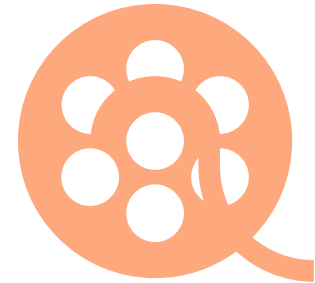
Based on User Profiles which in this case is each study protocol

Content Filtering Model



Based on Item Profiles which in this case is each CRF

A Hybrid Model



Combines both filtering models together

Challenges with using a Machine Learning Decision Tree

Unstructured Data

- Protocol PDFs are great for maintaining formatting
 - But very large
 - Difficult to parse
- English language
 - Contains stop words
 - Use of synonyms
 - Scientific jargon
 - Context may be required to decode meaning

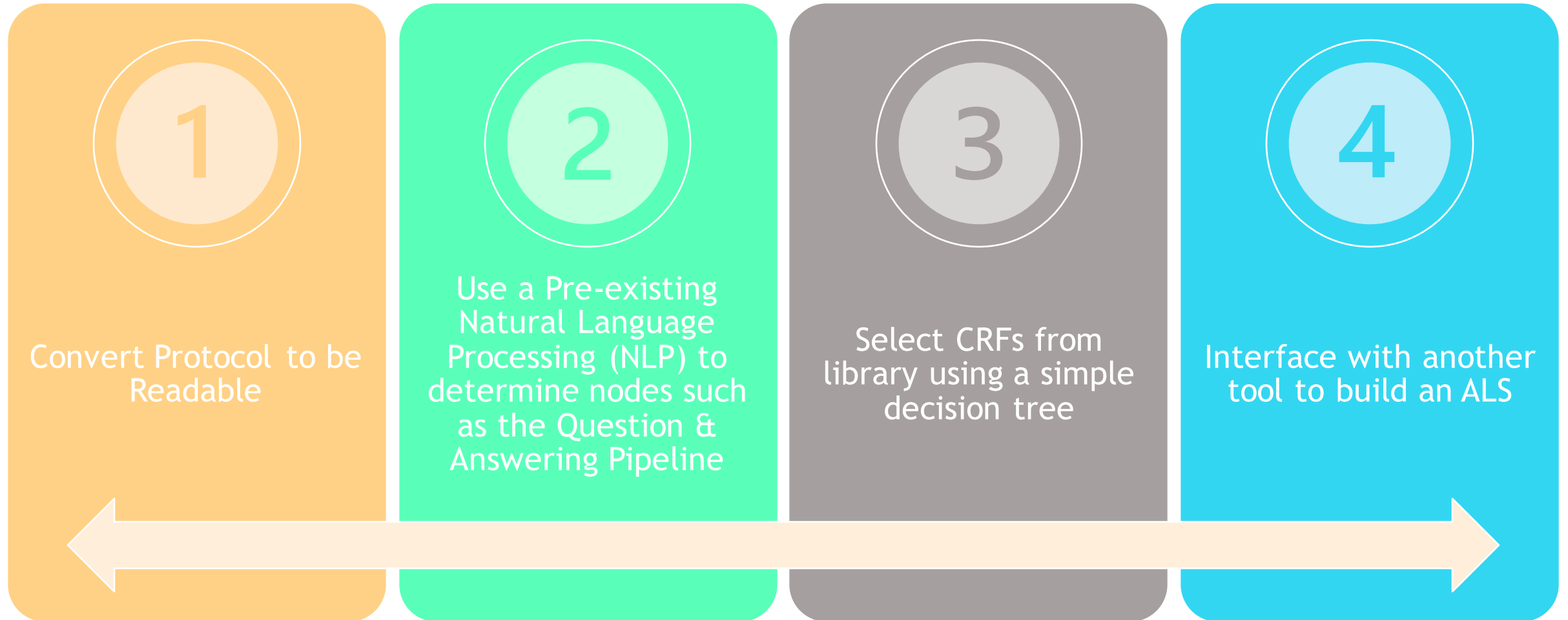
Black Box

- Feature Selection
 - Difficult to determine what to include in the model and how to weight factors for good predictability
- Machine Learning Decision Tree
 - decision nodes cannot be controlled, changed or even viewed
 - Non-explicit

Cold-Start Problem

- New Design Elements
 - The model may not be able to predict forms if the protocol has a new design, indication or compound
- Continuous Learning
 - Newly added forms and fields to standard libraries may not get added to studies if older similar forms/fields exist already

Solution



Data Engineering -Prepare Data for Model Ingestion



Read PDF file using
PYPDF2 python
package

Stemming and other
data engineering
techniques avoided



Hugging Face

HuggingFace NLP based model with the help of BERT and T5 encoder-decoder model

It is pre-trained on generic raw data. No data type biases.

Simplified output using text dictionary leads to plain language

— Example| Oncology, Carcinoma, Tumor Malignancy = Cancer

Combination of exact match and F-1 score are used to predict answers in the question and answering pipeline

Examples of the Question-Answering Pipeline



Exact Match

⚡ Question Answering demo

using [deepset/roberta-base-squad2](#)

📖 Question Answering Example 1 ▾

Compute

Context

A Phase II, Randomized, Open-label Trial of drug1 in Combination with drug2 versus standard of care as First-line Therapy for Melanoma

CheckMate ABC: CHECKpoint pathway and MAb clinical Trial Evaluation ABC

John Doe, MD
Clinical Trial Physician- Medical Monitor/Clinical Scientist
Bristol-Myers Squibb Company
Route 206 & Province Line Road
Lawrenceville, NJ 08543

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.050 s

Randomized 0.266

[JSON Output](#) Maximize

F-1 score

⚡ Question Answering demo

using [deepset/roberta-base-squad2](#)

📖 Question Answering Examples ▾

Compute

Context

A Phase II, Randomized, Open-label Trial of drug1 in Combination with drug2 versus standard of care as First-line Therapy for Melanoma

CheckMate ABC: CHECKpoint pathway and MAb clinical Trial Evaluation ABC

John Doe, MD
Clinical Trial Physician- Medical Monitor/Clinical Scientist
Bristol-Myers Squibb Company
Route 206 & Province Line Road
Lawrenceville, NJ 08543

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.066 s

John Doe, MD 0.301

[JSON Output](#) Maximize

Text dictionary

⚡ Question Answering demo

using [deepset/roberta-base-squad2](#)

📖 Question Answering Example 1 ▾

Compute

Context

A Phase II, Randomized, Open-label Trial of drug1 in Combination with drug2 versus standard of care as First-line Therapy for Melanoma

CheckMate ABC: CHECKpoint pathway and MAb clinical Trial Evaluation ABC

John Doe, MD
Clinical Trial Physician- Medical Monitor/Clinical Scientist
Bristol-Myers Squibb Company
Route 206 & Province Line Road
Lawrenceville, NJ 08543

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.049 s

Melanoma 0.022

[JSON Output](#) Maximize

Benefits of using the Question-Answering Pipeline

⚡ Question Answering demo
using [deepset/roberta-base-squad2](#)

📄 Question Answering Example 1 ▾

Who is the sponsor? Compute

Context

A Phase II, Randomized, Open-label Trial of drug1 in Combination with drug2 versus standard of care as First-line Therapy for Melanoma

CheckMate ABC: CHECKpoint pathway and nivolumAb clinical Trial Evaluation ABC

John Doe, MD
Clinical Trial Physician- Medical Monitor/Clinical Scientist
Bristol-Myers Squibb Company
Route 206 & Province Line Road
Lawrenceville, NJ 08543

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.048 s

Bristol-Myers Squibb Company 0.011

</> JSON Output 🖥 Maximize

Black Box typical of ML models

- Answers to questions predicted by the model can be printed for validation by a user as needed

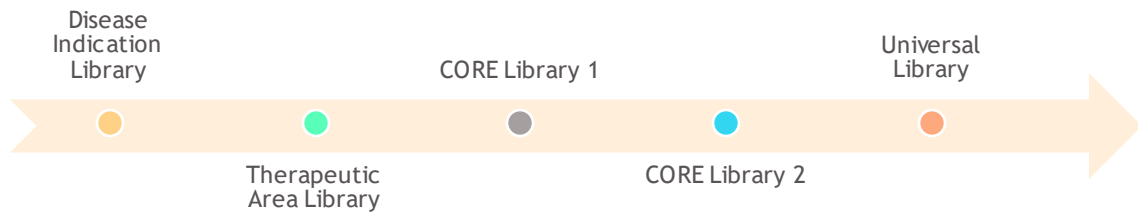
Cold-Start Problem

- Easily modify questions as the protocol development landscape changes

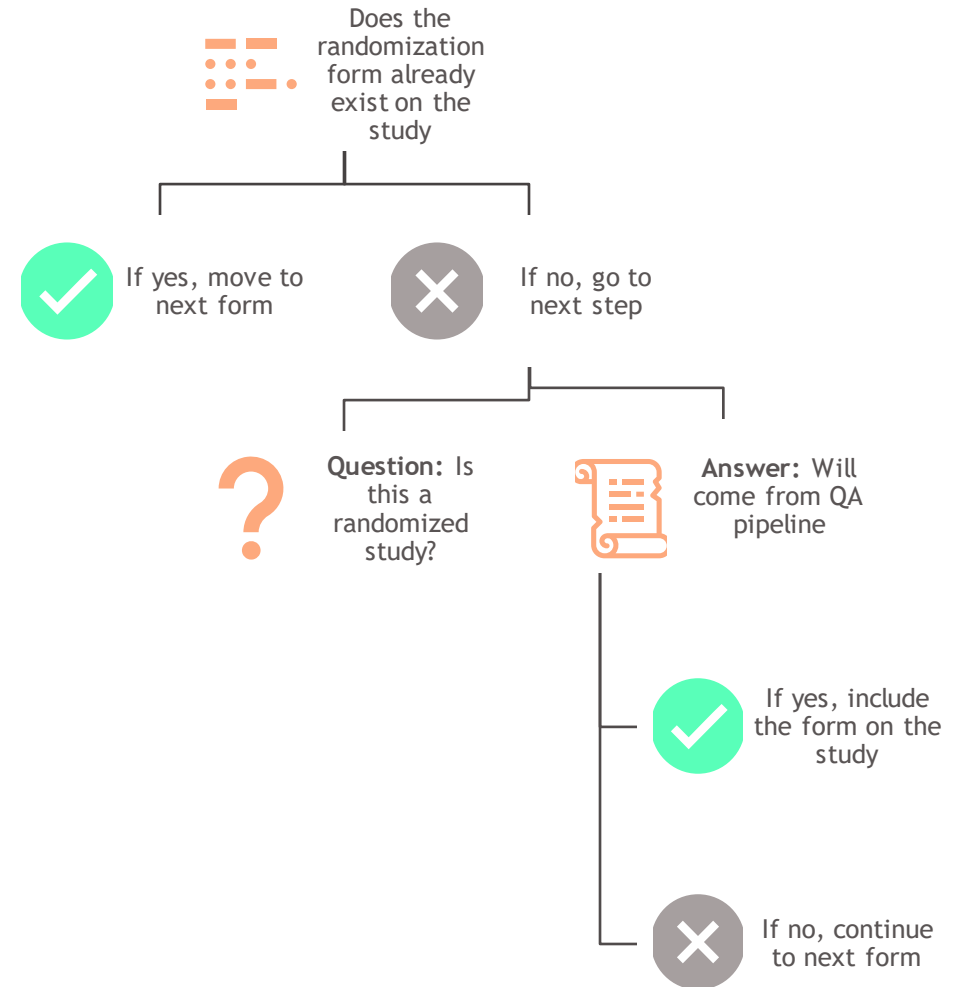
Simple Decision Tree for CRF Selection

Selection of Library

Ability to follow library hierarchy and choose the correct version of the form if multiple versions of the form exist



Maintain control and the ability to interact with the predictions to add libraries for integrations to other systems that may not be defined in the protocol



Be Scrappy!

- There is an existing excel-macro based home-grown tool that can load universal libraries and subset libraries from SharePoint
- User must select libraries and forms from each library and the excel macro generates the ALS
- Use of existing infrastructure and processes leads to faster implementation
- By having the model interface with the current tool, the process of going from the protocol to the ALS will be complete



1

Convert Protocol to be Readable

2

Use a Pre-existing Natural Language Processing (NLP) to determine nodes such as the Question & Answering Pipeline

3

Select CRFs from library using a simple decision tree

4

Interface with another tool to build an ALS

Next Steps to Implementation

Build the decision tree

This will be done alongside the pre-existing standard library. The question/condition will be added alongside each existing form and library

Fine-tune

Best practices for phrasing of the questions for the model need to be established

Deploy on a subset of studies

Test implementation time, ease of use and predictive capability on a handful of studies

Continuous Improvement

Continue to integrate with existing tools and infrastructure to improve the program



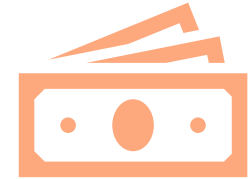
Scalability

Greater adherence to standards
Reusable edit checks, reports,
programming



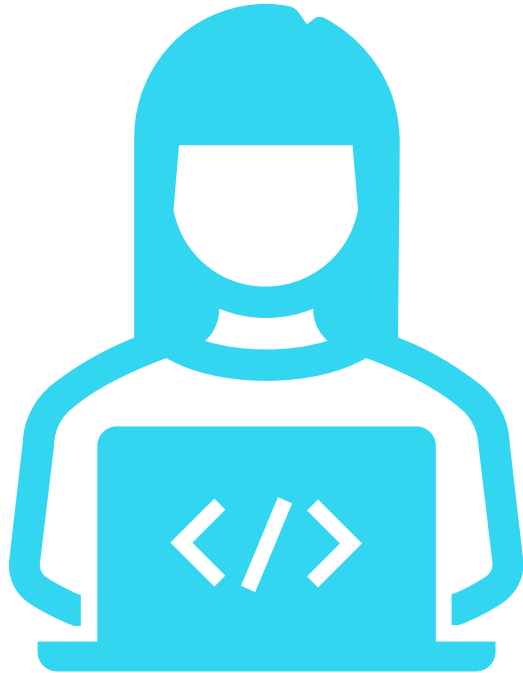
Save Time

Instead of 12 weeks for a standard
build maybe 6 weeks
Greater study team interaction
within the actual system



Save Money

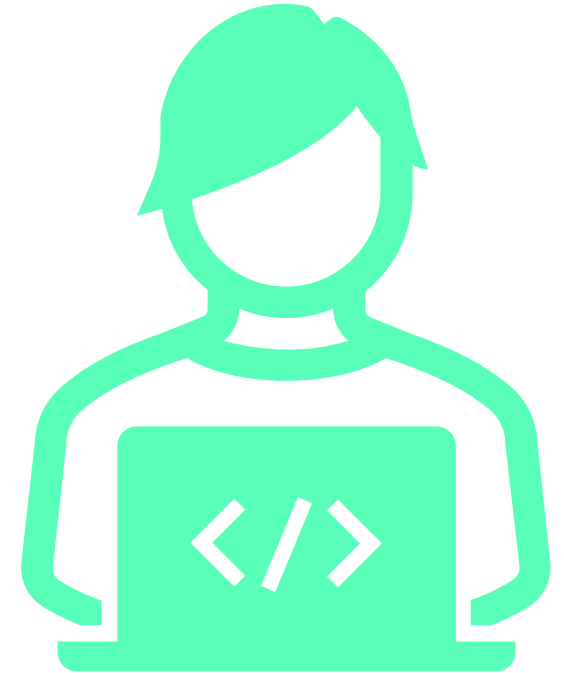
Time is money
Less resource intensive in the
future after initial implementation



We use artificial intelligence everywhere in our daily lives.

Be curious and responsible.
Find ways to apply these advances to the domains that you already have expertise in.

Who know what the drug development space can become within the next few years.



Thank you

