WITH STANDARDS – UNLOCK THE POWER OF DATA

2022
US
INTERCHANGE
26-27 OCTOBER | AUSTIN

Practical Steps for Implementing ML for SDTM Mapping

Presented by Sharon Rossouw, Director, Biostatistics, Bioforum

Meet the Speaker

Sharon Rossouw

Title: Director, Biostatistics

Organization: Bioforum

There is a farm in Africa…. I grew up on a farm in the Zimbabwean bushveld. I completed my schooling and university education in South Africa culminating with a Masters in Biostatistics.

A biostatistician over 25 years of experience providing biostatistical and medical writing services to the pharmaceutical industry and academic institutions.

I am passionate about the training and development of biostatisticians and statistical programmers and have a special interest in process development and implementation.

Disclaimer and Disclosures

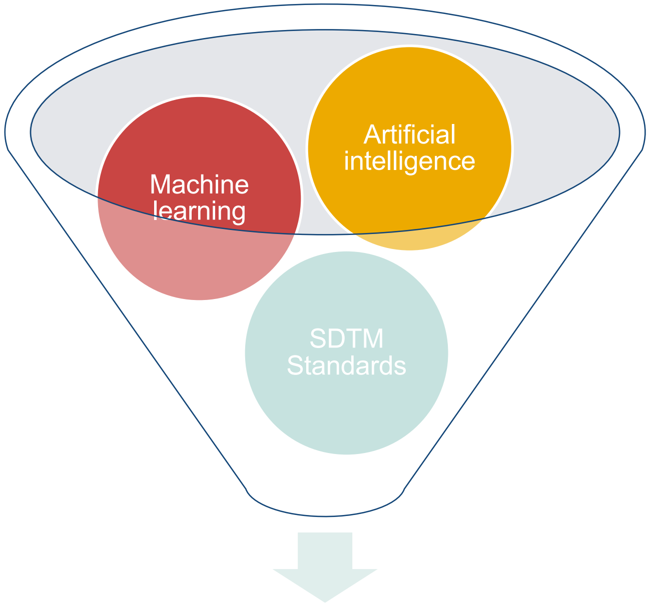
- *The views and opinions expressed in this presentation are those of the author and do not necessarily reflect the official policy or position of CDISC.*

- *The author has no real or apparent conflicts of interest to report.*

Agenda

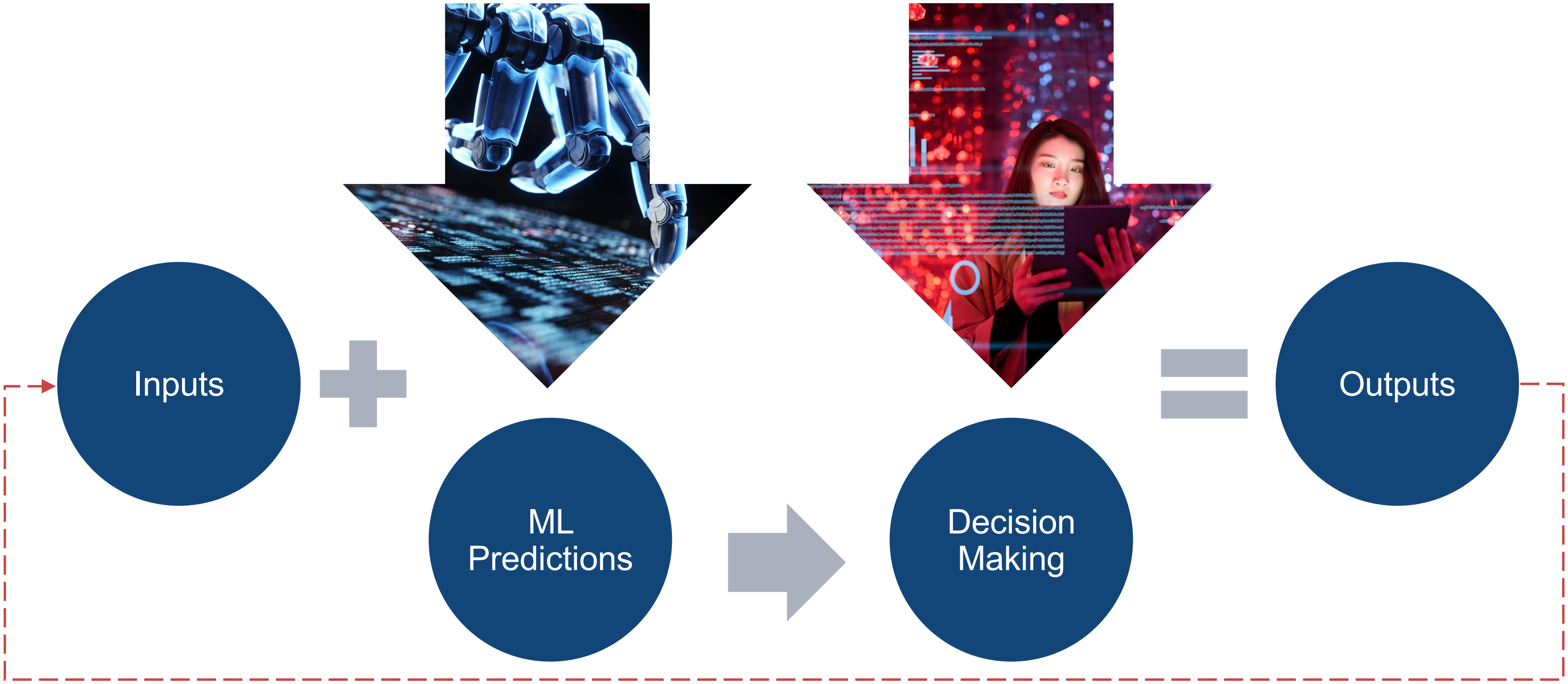
1. Background
2. Build steps: Develop the models
3. Taking a step back
4. Refine steps: Applying the models
5. The view from our destination

What are the steps from concept to reality?



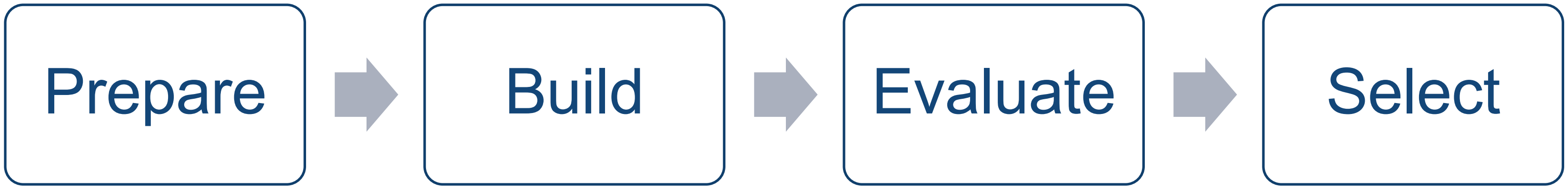
Implementation

Where to use ML in the SDTM mapping process?



Build steps: Develop the models

Build steps: Develop the models



Prepare → Build → Evaluate → Select

1st step: Identify the pieces

- Domain
- Variable
- Controlled terminology
- …

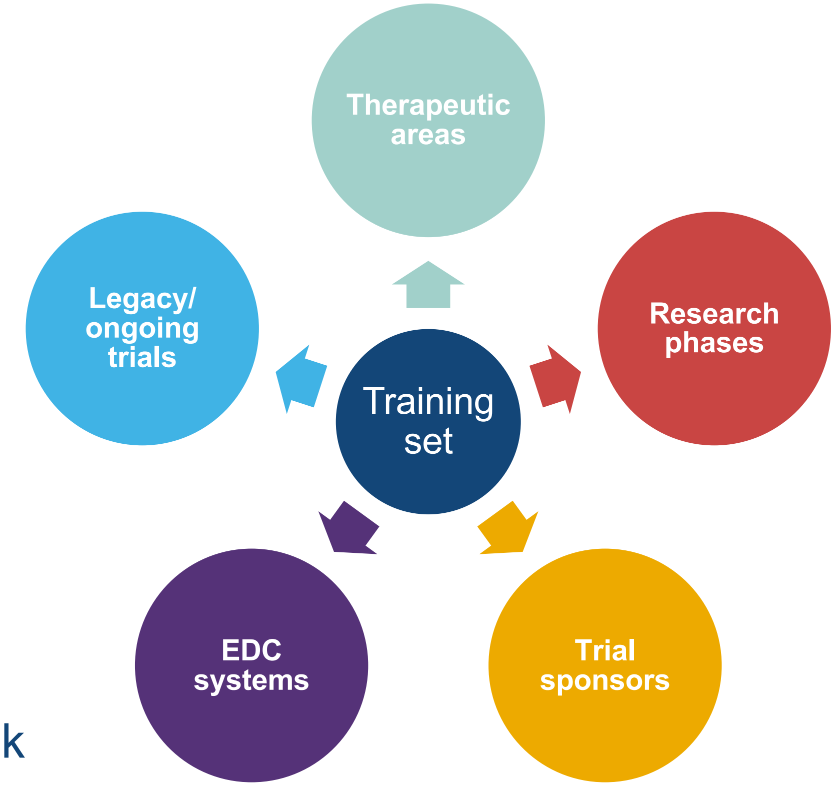
STEP 1: Prepare

- Training set ➔ Pre-mapped trials

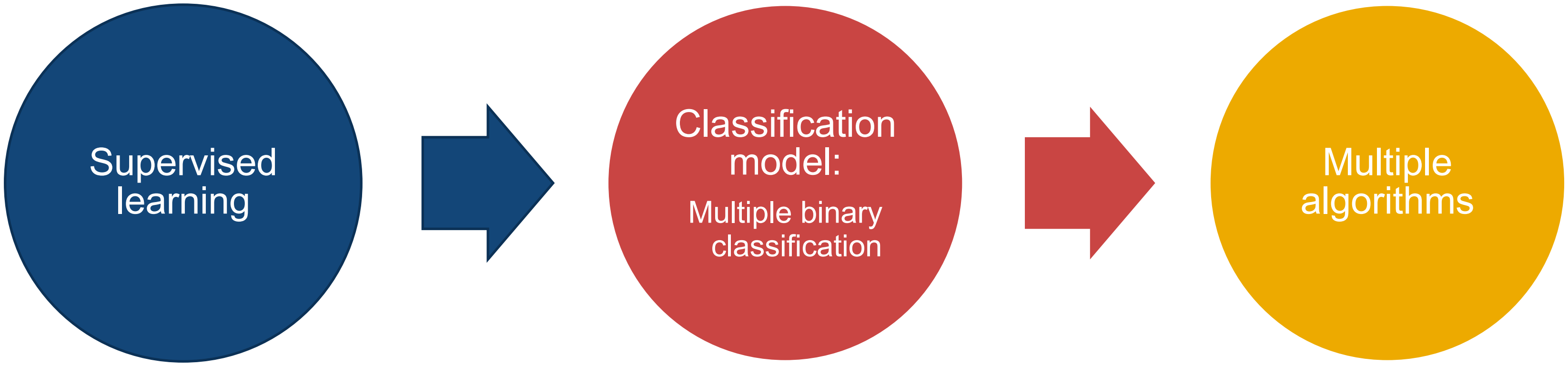
- Raw variables labelled:
  - Domain(s)
  - Variable, etc.

- Raw variable feature extraction:
  - Raw data file characteristics
  - Raw variable metadata
  - Variable values
  - Trial documents

- Training set was tailored to the task



STEP 2: Build the models



Supervised learning

Classification model:

Multiple binary classification

Multiple algorithms

STEP 3: Evaluate the models

- Test model performance
  - Cross validation methodology
- Results
  - Vector of probabilities for each raw variable
  - Probability = Likelihood of mapping to target
- Simple decision rule
  - Select target with highest probability
- Compare
  - Selected target ⇔ Pre-mapped target

STEP 4: Select the models

Domain & Variable mapping

Random Forest

Controlled terminology mapping

Natural Language Processing

- For simplicity, the domain model results are presented
  - Variable model results are briefly mentioned
  - Controlled terminology results are a “forthcoming attraction”

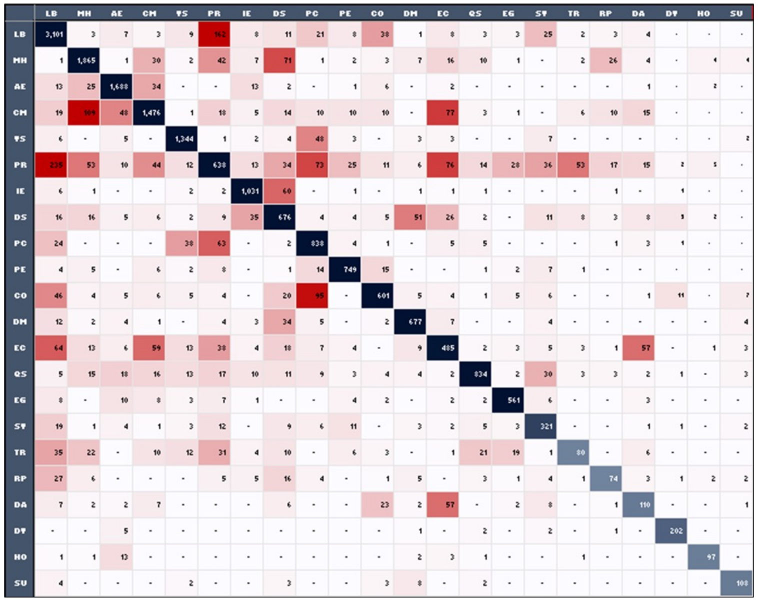
Where did the first few steps lead?

Model accuracy

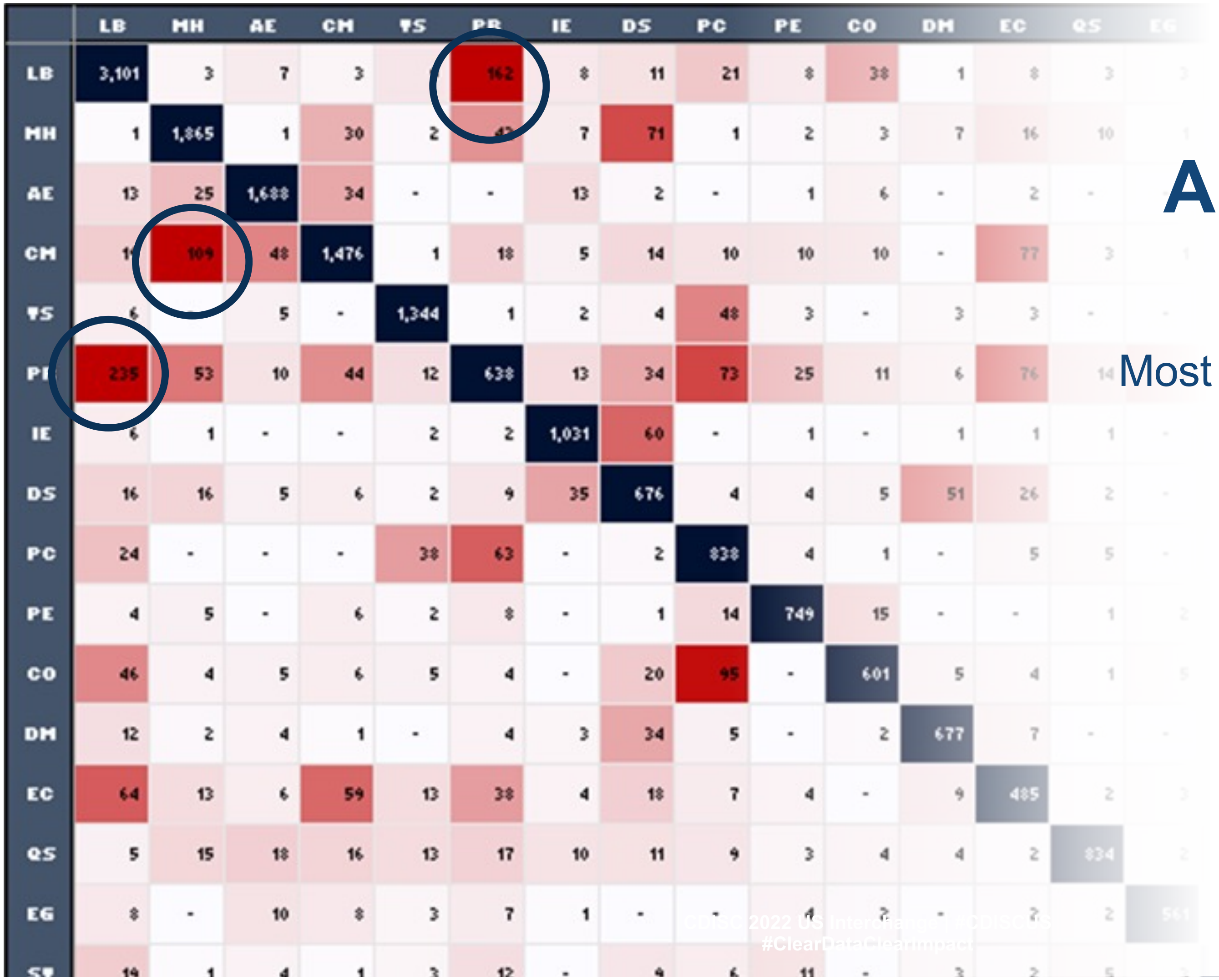
- Confusion Matrix
  ➔ Indication of model quality
- Diagonal frequencies = correct predictions
- Model is mostly correct
- Accuracy:
  - Domain-level ⇨ 71.5% (41 trials)
  - Variable-level ⇨ 83.8% (61 trials)

Columns = Predicted domain

Rows = Pre-mapped domain



	LB	MH	AE	CM	VS	PR	IE	DS	PC	PE	CO	DM	EC	QS	EG	SV	TR	RP	DA	DV	HO	SU
LB	3,101	3	7	3	9	162	8	11	21	8	38	1	8	3	3	25	2	3	4	-	-	-
MH	1	1,865	1	30	2	42	7	71	1	2	3	7	16	10	1	-	2	26	4	-	4	4
AE	13	25	1,688	34	-	-	13	2	-	1	6	-	2	-	-	-	-	-	1	-	2	-
CM	19	109	48	1,476	1	18	5	14	10	10	10	-	77	3	1	-	6	10	15	-	-	-
VS	6	-	5	-	1,344	1	2	4	48	3	-	3	3	-	-	7	-	-	-	-	-	2
PR	235	53	10	44	12	638	13	34	73	25	11	6	76	14	28	36	53	17	15	2	5	-
IE	6	1	-	-	2	2	1,031	60	-	1	-	1	1	1	-	-	-	1	-	1	-	-
DS	16	16	5	6	2	9	35	676	4	4	5	51	26	2	-	11	8	3	8	1	2	-
PC	24	-	-	-	38	63	-	2	838	4	1	-	5	5	-	-	-	1	3	1	-	-
PE	4	5	-	6	2	8	-	1	14	749	15	-	-	1	2	7	1	-	-	-	-	-
CO	46	4	5	6	5	4	-	20	95	-	601	5	4	1	5	6	-	-	1	11	-	7
DM	12	2	4	1	-	4	3	34	5	-	2	677	7	-	-	4	-	-	-	-	-	4
EC	64	13	6	59	13	38	4	18	7	4	-	9	485	2	3	5	3	1	57	-	1	3
QS	5	15	18	16	13	17	10	11	9	3	4	4	2	834	2	30	3	3	2	1	-	3
EG	8	-	10	8	3	7	1	-	-	4	2	-	2	2	561	6	-	-	3	-	-	-
SV	19	1	4	1	3	12	-	9	6	11	-	3	2	5	3	321	-	-	1	1	-	2
TR	35	22	-	10	12	31	4	10	-	6	3	-	1	21	19	1	80	-	6	-	-	-
RP	27	6	-	-	-	5	5	16	4	-	1	5	-	3	1	4	1	74	3	1	2	2
DA	7	2	2	7	-	-	-	6	-	-	23	2	57	-	2	8	-	1	110	-	-	1
DV	-	-	5	-	-	-	-	-	-	-	-	1	-	2	-	2	-	1	-	202	-	-
HO	1	1	13	-	-	-	-	-	-	-	-	2	3	1	-	-	1	-	-	-	97	-
SU	4	-	-	-	2	-	-	3	-	-	3	8	-	2	-	-	-	-	-	-	-	108



	LB	MH	AE	CM	VS	PR	IE	DS	PC	PE	CO	DM	EC	QS	EG
LB	3,101	3	7	3	[illegible]	162	8	11	21	8	38	1	8	3	3
MH	1	1,865	1	30	2	42	7	71	1	2	3	7	16	10	1
AE	13	25	1,688	34	-	-	13	2	-	1	6	-	2	-	
CM	[illegible]	109	48	1,476	1	18	5	14	10	10	10	-	77	3	1
VS	6	[illegible]	5	-	1,344	1	2	4	48	3	-	3	3	-	-
PR	235	53	10	44	12	638	13	34	73	25	11	6	76	14	
IE	6	1	-	-	2	2	1,031	60	-	1	-	1	1	1	-
DS	16	16	5	6	2	9	35	676	4	4	5	51	26	2	-
PC	24	-	-	-	38	63	-	2	838	4	1	-	5	5	-
PE	4	5	-	6	2	8	-	1	14	749	15	-	-	1	2
CO	46	4	5	6	5	4	-	20	95	-	601	5	4	1	5
DM	12	2	4	1	-	4	3	34	5	-	2	677	7	-	-
EC	64	13	6	59	13	38	4	18	7	4	-	9	485	2	3
QS	5	15	18	16	13	17	10	11	9	3	4	4	2	834	2
EG	8	-	10	8	3	7	1	-	-	4	2	-	2	2	561
ST	19	1	4	1	3	12	-	9	6	11	-	3	2	5	2

A closer look

Most frequently confused:

MH instead of CM

PR instead of LB

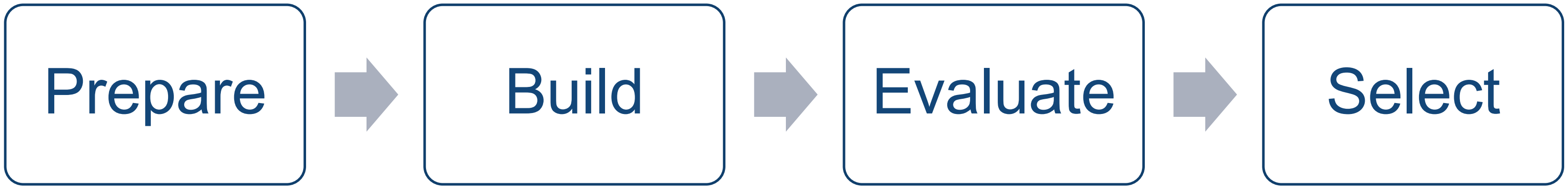
LB instead of PR

…

Taking a step back

Reviewing the journey

So what was the next step?



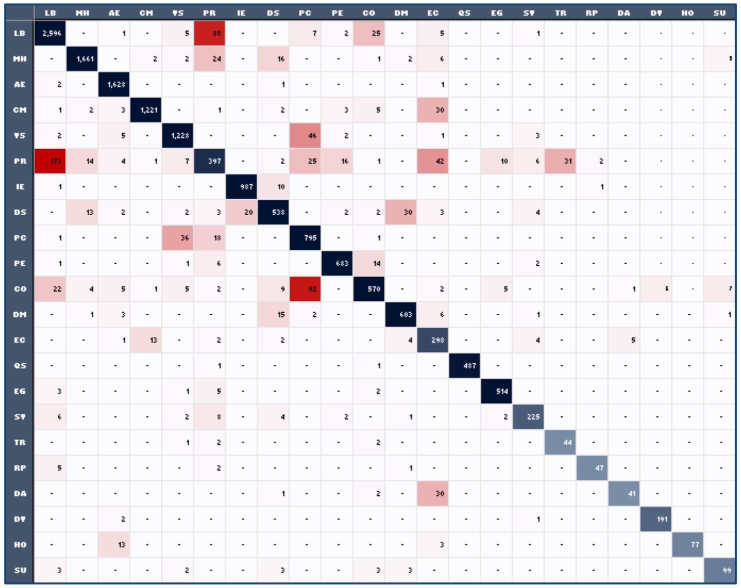
Refine…

Refine steps: Domain models

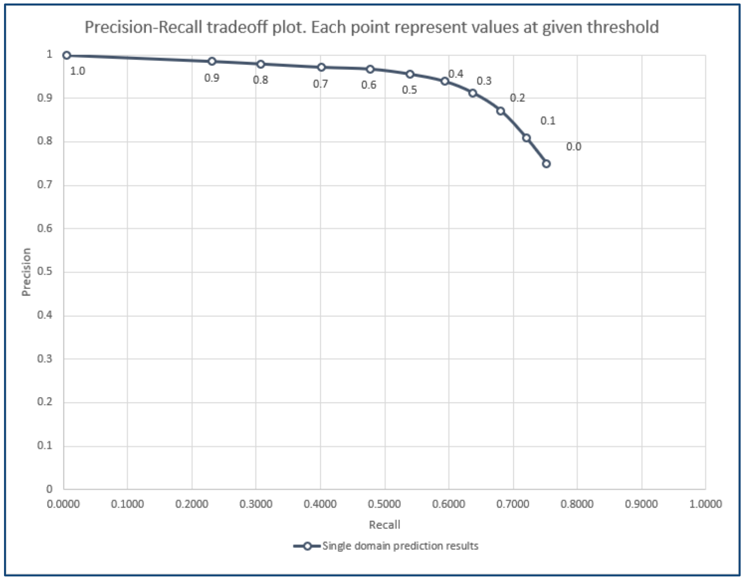
STEP 1 – Remove predictions with low confidence levels

Remove predictions if confidence ≤ 0.3 ➔ 69.7% of variables retained

- Domain precision: 91.4%, recall of 63.7% (41 trials)



	LB	MH	AE	CM	VS	PR	IE	DS	PC	PE	CO	DM	EC	QS	EG	SV	TR	RP	DA	DV	HO	SU
LB	2,596	-	1	-	5	88	-	-	7	2	25	-	5	-	-	1	-	-	-	-	-	-
MH	-	1,661	-	2	2	24	-	16	-	-	1	2	6	-	-	-	-	-	-	-	-	5
AE	2	-	1,628	-	-	-	-	1	-	-	-	-	1	-	-	-	-	-	-	-	-	-
CM	1	2	3	1,221	-	1	-	2	-	3	5	-	30	-	-	-	-	-	-	-	-	-
VS	2	-	5	-	1,228	-	-	-	46	2	-	-	1	-	-	3	-	-	-	-	-	-
PR	173	14	4	1	7	397	-	2	25	16	1	-	42	-	10	6	31	2	-	-	-	-
IE	1	-	-	-	-	-	987	10	-	-	-	-	-	-	-	-	-	1	-	-	-	-
DS	-	13	2	-	2	3	20	538	-	2	2	30	3	-	-	4	-	-	-	-	-	-
PC	1	-	-	-	36	18	-	-	795	-	1	-	-	-	-	-	-	-	-	-	-	-
PE	1	-	-	-	1	6	-	-	-	683	14	-	-	-	-	2	-	-	-	-	-	-
CO	22	4	5	1	5	2	-	9	92	-	570	-	2	-	5	-	-	-	1	6	-	7
DM	-	1	3	-	-	-	-	15	2	-	-	603	6	-	-	1	-	-	-	-	-	1
EC	-	-	1	13	-	2	-	2	-	-	-	4	298	-	-	4	-	-	5	-	-	-
QS	-	-	-	-	-	1	-	-	-	-	1	-	-	487	-	-	-	-	-	-	-	-
EG	3	-	-	-	1	5	-	-	-	-	2	-	-	-	514	-	-	-	-	-	-	-
SV	6	-	-	-	2	8	-	4	-	2	-	1	-	-	2	225	-	-	-	-	-	-
TR	-	-	-	-	1	2	-	-	-	-	2	-	-	-	-	-	44	-	-	-	-	-
RP	5	-	-	-	-	2	-	-	-	-	-	1	-	-	-	-	-	47	-	-	-	-
DA	-	-	-	-	-	-	-	1	-	-	2	-	30	-	-	-	-	-	41	-	-	-
DV	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	191	-	-
HO	-	-	13	-	-	-	-	-	-	-	-	-	3	-	-	-	-	-	-	-	77	-
SU	3	-	-	-	2	-	-	3	-	-	3	3	-	-	-	-	-	-	-	-	-	99



Precision-Recall tradeoff plot. Each point represent values at given threshold

STEP 1 – Impact

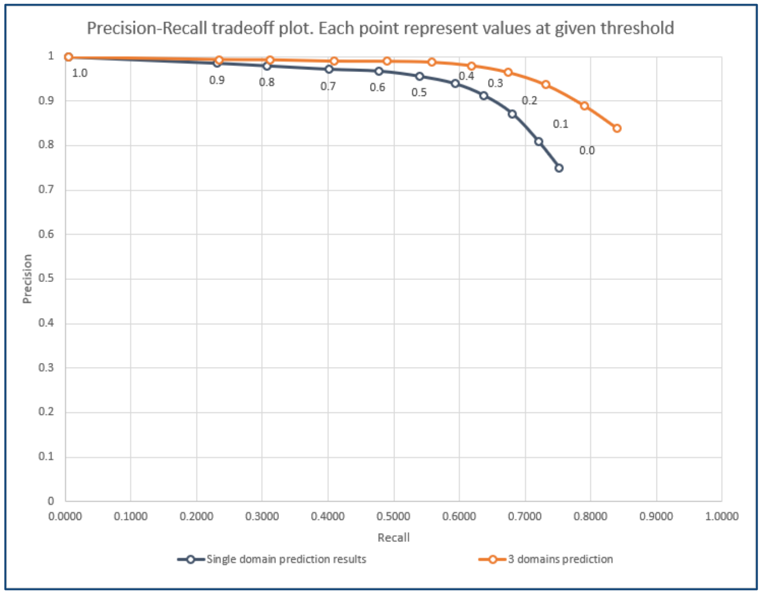
- Now our users were interested…
  - Trust had increased
  - Some trials: nearly error-free

- Investigate mistaken predictions
  - “Correct” recommendation was 2nd/3rd on list
  - How can we adapt the implementation of the models?

STEP 2 - Provide the top 3 most likely predictions

- Change the decision rule:
  - “Correct” target in the top 3 most likely recommendations

- Using the 0.3 threshold:
  - Precision ➔ 96.6% (STEP 1: 91.4%)
  - Recall ➔ 67.3% (STEP 1: 63.7%)



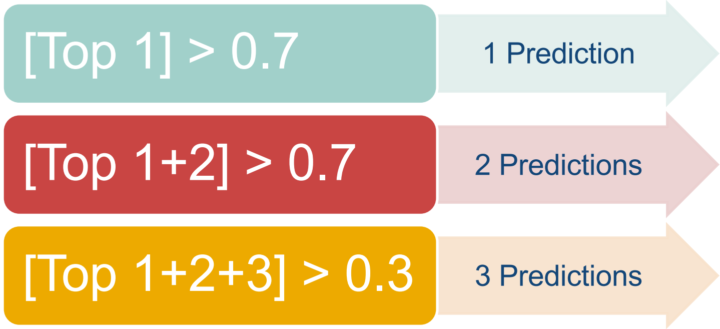
STEP 2 – Impact

- Users loved this!
- ↑ accuracy = ↑ interest
- Next step:
  - Users had to continuously evaluate 3 options, even when the “perfect” fit was obvious
  - Is there somewhere in between?

STEP 3 - Dynamic predictions based on cumulative confidence thresholds

- Alternative method investigated:
    - Dynamic cumulative approach

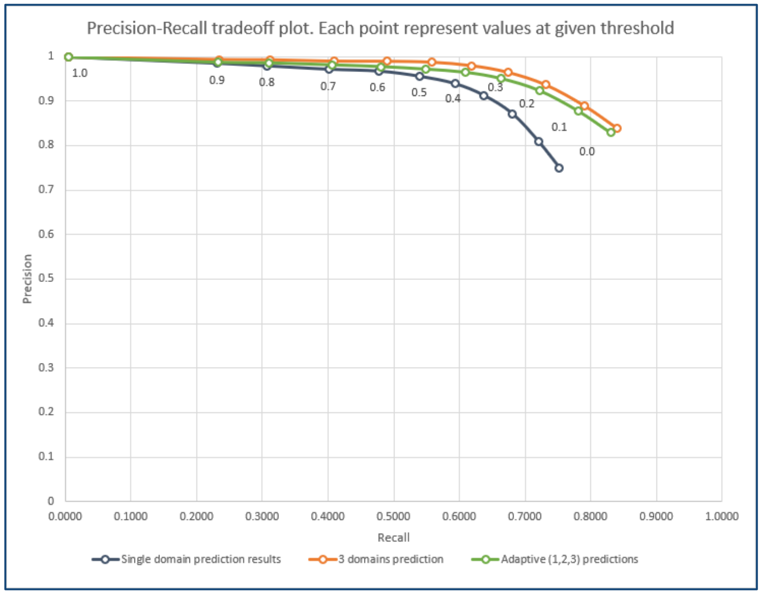
- Tradeoff:
    - Provide a single target
    - Need for high confidence



[Top 1] > 0.7 → 1 Prediction

[Top 1+2] > 0.7 → 2 Predictions

[Top 1+2+3] > 0.3 → 3 Predictions



STEP 3 – Impact

- Maintained high precision
- Single target for about 60% of predictions
  - Other variables: Users could select the prediction from the list
- User approved

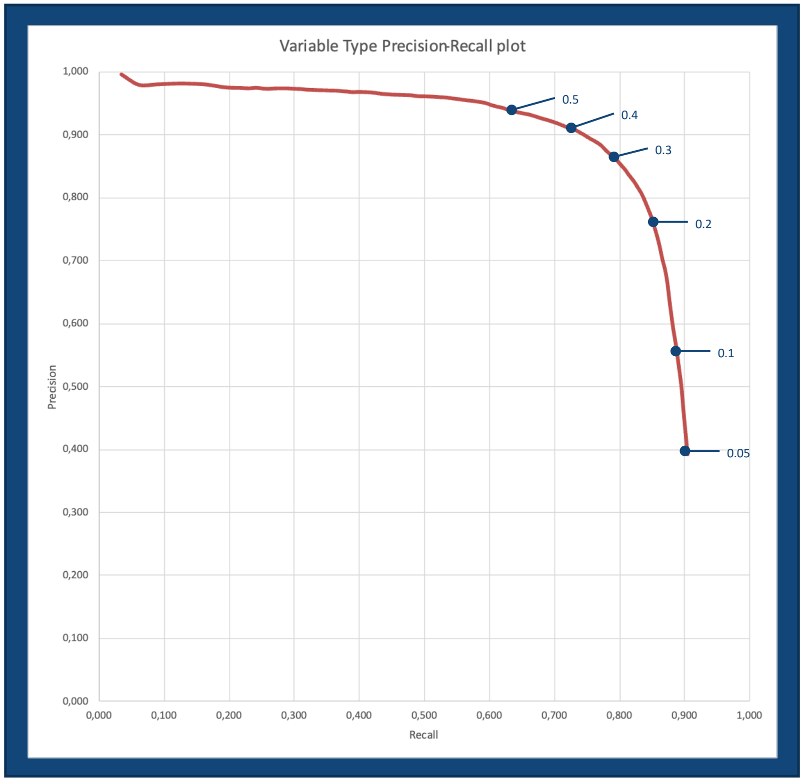
Refine steps: Variable models

STEP 1 – Provide predictions above a static threshold

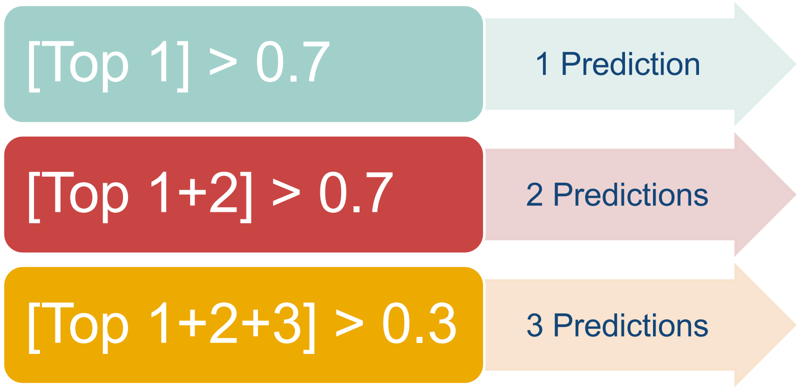
- Present if likelihood > threshold
  - Up to a maximum of 3 predictions

- Variable-level at 0.3 threshold (61 trials):
  - Precision: 86.9%
  - Recall of 78.6%

- It works to only suggest what you are sure about!



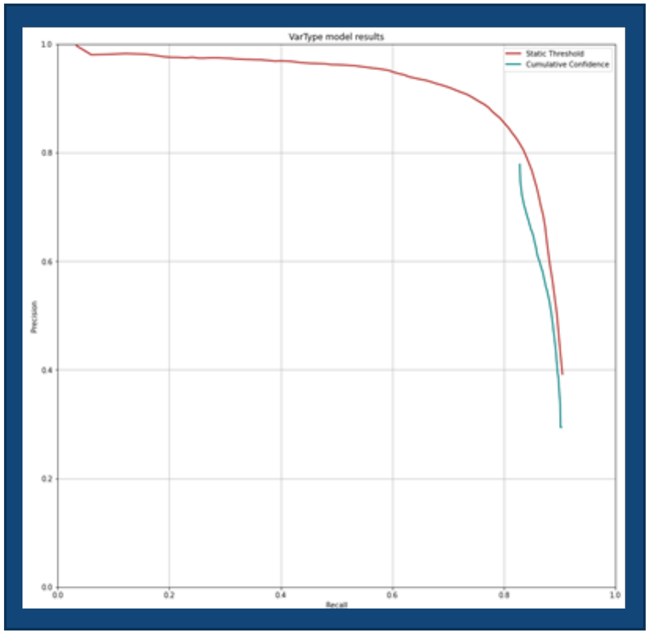
STEP 2 - Dynamic predictions based on cumulative confidence thresholds



[Top 1] > 0.7 → 1 Prediction

[Top 1+2] > 0.7 → 2 Predictions

[Top 1+2+3] > 0.3 → 3 Predictions



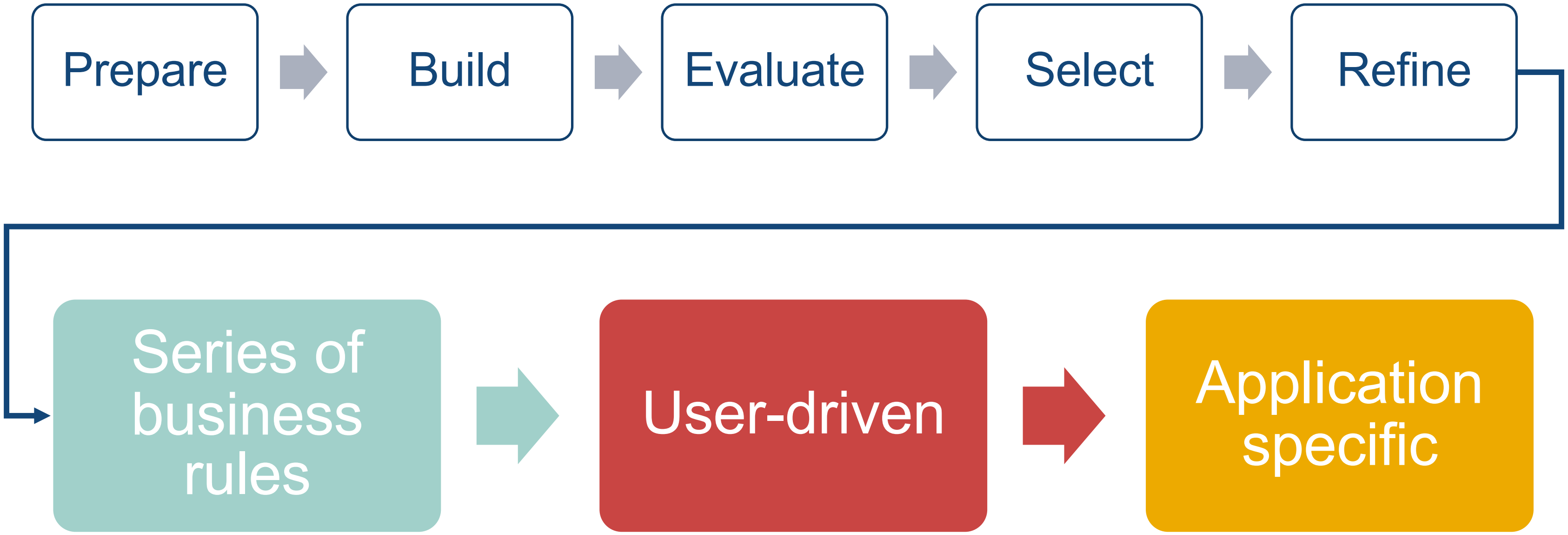
Be practical when implementing ML models!

The view from our refined destination

Where did the journey take us?

To summarize…

Practical steps for implementing ML for SDTM Mapping



Prepare

Build

Evaluate

Select

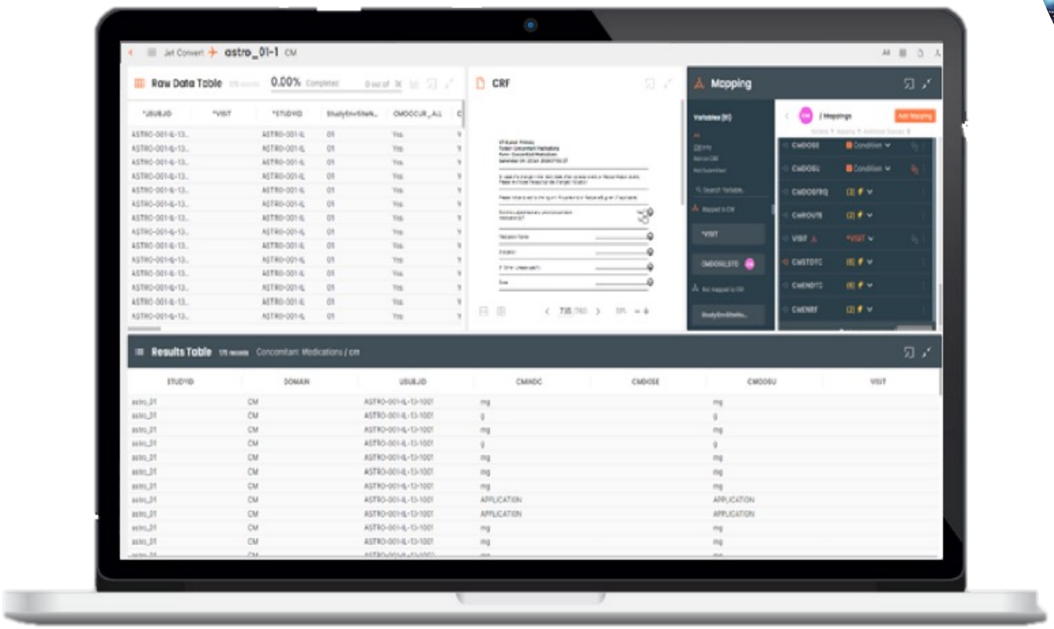
Refine

Series of business rules

User-driven

Application specific

Using ML in the SDTM mapping process is a reality



*USUBJID	*VISIT	*STUDYID	StudyEnvSiteN...	CMOCCUR_ALL	C
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y
ASTRO-001-IL-13...		ASTRO-001-IL	01	Yes	Y

STUDYID	DOMAIN	USUBJID	CMINDC	CMDOSE	CMDOSU	VISIT
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	g		g	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	g		g	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	APPLICATION		APPLICATION	
astro_01	CM	ASTRO-001-IL-13-1001	APPLICATION		APPLICATION	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	
astro_01	CM	ASTRO-001-IL-13-1001	mg		mg	

cdisc

Thank You!

I would like to extend my thanks to the following individuals for assisting in the preparation of this presentation:

- Sergei Merson
- Shahar Cohen
- Lena Hazanov
- Mor Meyerovich
- Eyal Wultz
- Bremer Louw