SDTM Mapping based on Natural Language Process and Machine Learning

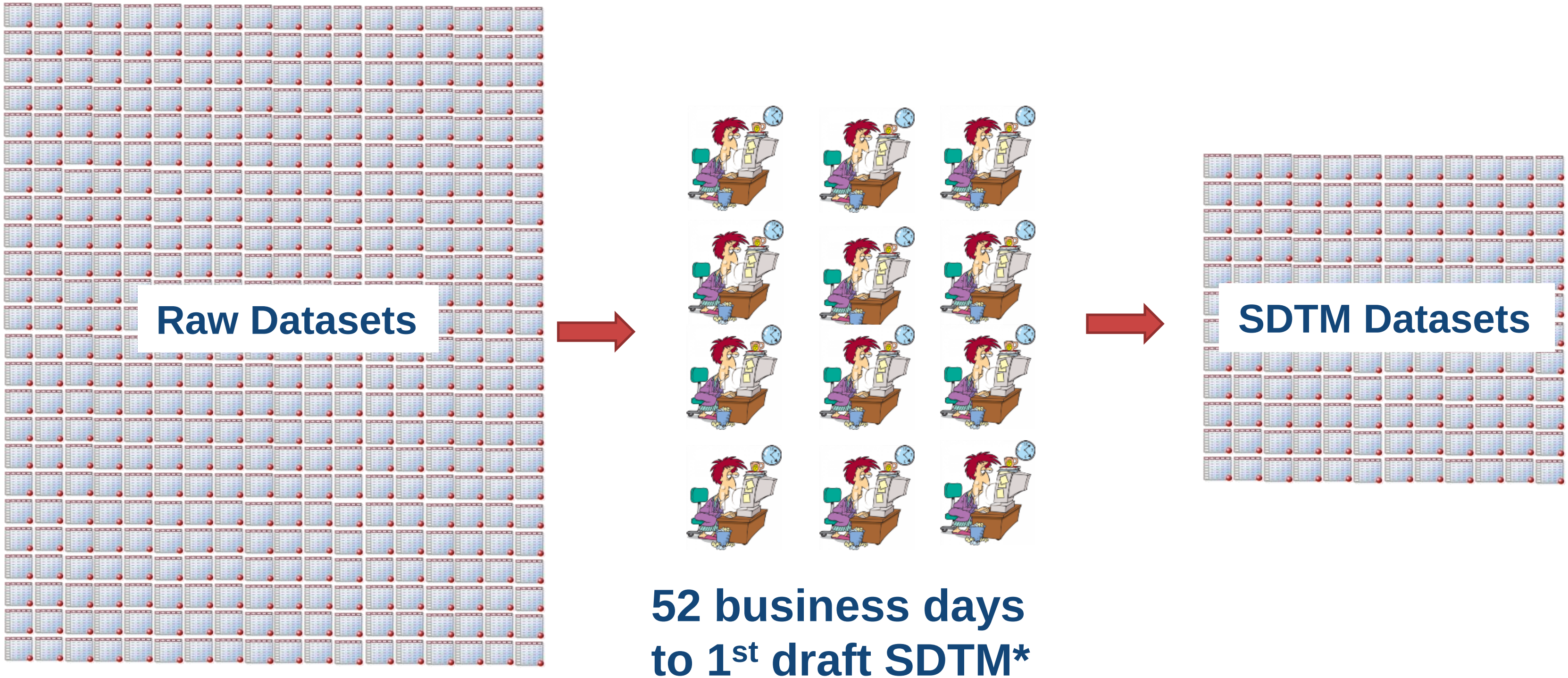
Sam Tomioka,

Director of CDP - Data Science, SUNOVION Pharmaceuticals

10.10.2018

A Problem We Want to Solve

A PROBLEM WE WANT TO SOLVE



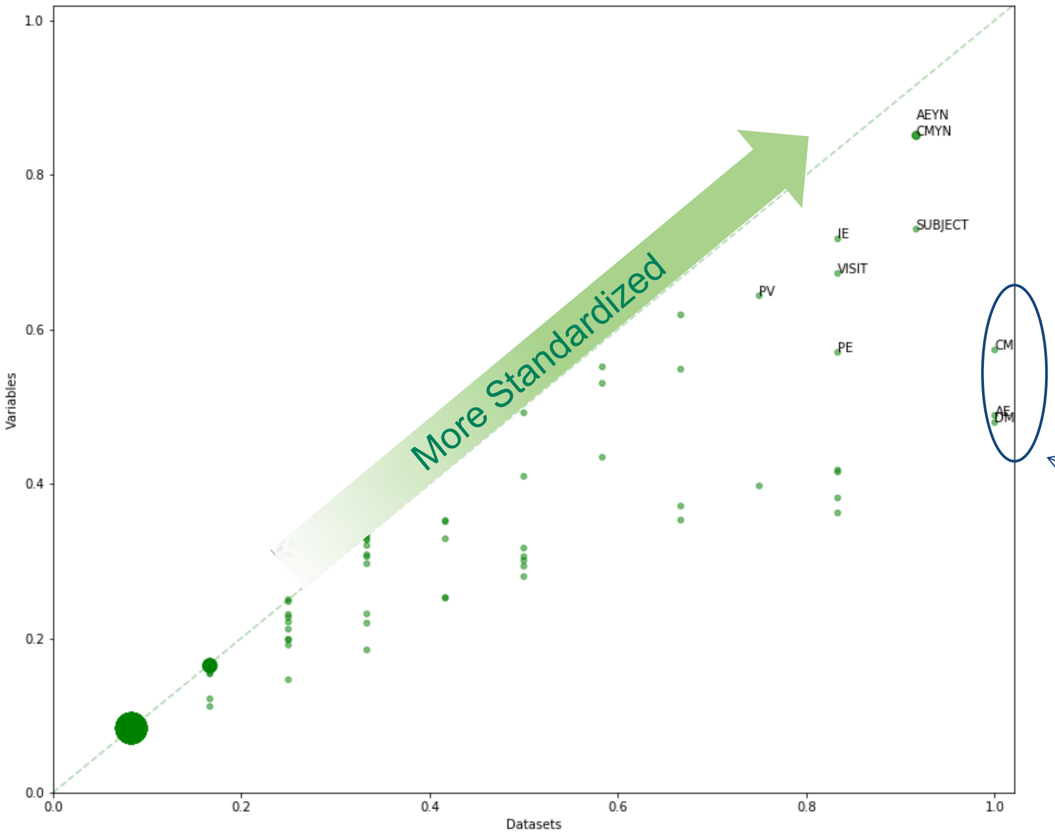
Raw Datasets

SDTM Datasets

52 business days to 1st draft SDTM*

* 2017 data

Reuse of datasets and variables
across 12 Sunovion studies delivered by one CRO in 2017



Each dot represents raw dataset.Annotation added for top 10 most consistent datasets

Note: Scatterplot x-axis: proportion of studies using a particular dataset to 12 studies

Note: Scatterplot y-axis: the mean proportion of the studies which used a particular variable within a dataset to 12 studies

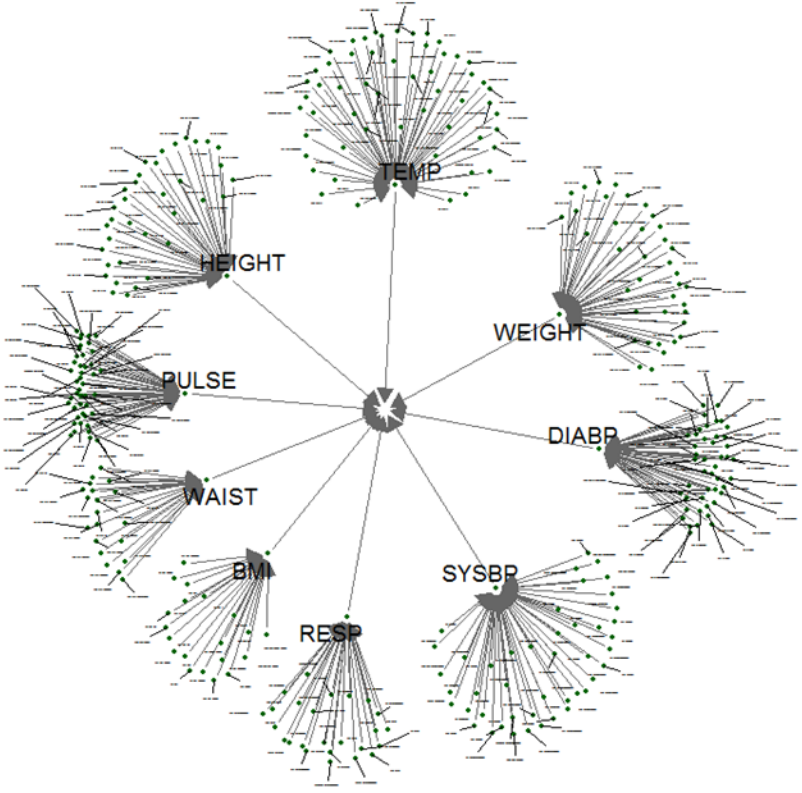
Data Source: Raw dataset metadata from 12 clinical studies managed by one CRO in 2017

PROGRAMMERS ARE FRUSTRATED WITH NON-STANDARDIZED RAW DATA

Only 3 datasets used
in all 12 studies

ENDLESS MAPPING…

SDTM.VS.VSORRES mapping for
12 Sunovion studies delivered by one CRO



Data Source: ~/My Documents/r/programming_challenges_201711xx/inventory20171114_t.xlsx

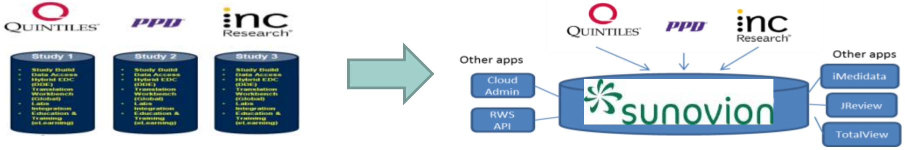
```
20 sources for BMI

"VS.BMI"
"VS.BMI_RAW"
"VS.BMI_Z"
"VS.BMI_Z_RAW"
"VS.BMIS"
"VS.BMIS_RAW"
"VS1.BMI"
"VS1.BMI_RAW"
"VS1.BMI_Z"
"VS1.BMI_Z_RAW"
"VS1.VS1BMI"
"VS2.BMI"
"VS2.BMI_RAW"
"VS2.D_BMI"
"VS2.D_BMI_RAW"
"VS2.VS2BMI"
"VSMSTR.BMI"
"VSMSTR.BMI_RAW"
"VSMSTR.D_BMI"
"VSMSTR.D_BMI_RAW"
```

Embrace Metadata Standards

Target up to 20% reduction in time
Target up to 40% data consistency

RaveX Implementation Project

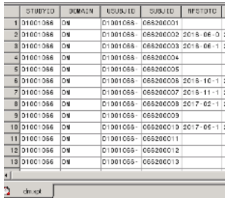


Go live FY 2018 2Q

SDTM Automation Project

SDTM automation macros
Domain templates

Started FY 2017 1Q



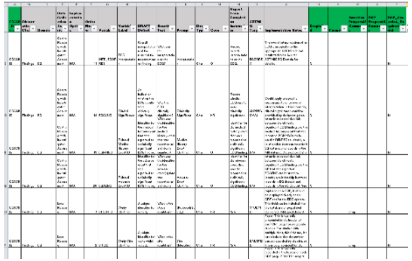
	STUDYID	DOMAIN	USUBJID	SUBJID	RFSTDTC
1	D1001066	DM	D1001066-	066200001	
2	D1001066	DM	D1001066-	066200002	2016-06-0
3	D1001066	DM	D1001066-	066200003	2016-06-1
4	D1001066	DM	D1001066-	066200004	
5	D1001066	DM	D1001066-	066200005	
6	D1001066	DM	D1001066-	066200006	2016-10-1
7	D1001066	DM	D1001066-	066200007	2016-11-1
8	D1001066	DM	D1001066-	066200008	2017-02-1
9	D1001066	DM	D1001066-	066200009	
10	D1001066	DM	D1001066-	066200010	2017-05-1
11	D1001066	DM	D1001066-	066200011	
12	D1001066	DM	D1001066-	066200012	
13	D1001066	DM	D1001066-	066200013	

SDTM

Standard CRF Working Group

- Sunovion Standard CRF Final v1.0_Central_EG ...
- Sunovion Standard CRF Final v1.0_Central_LB ...
- Sunovion Standard CRF Final v1.0_DA ...
- Sunovion Standard CRF FINAL v1.0_EX ...
- Sunovion Standard CRF FINAL v1.0_EX-DA ...
- Sunovion Standard CRF FINAL v1.0_PV ...
- Sunovion Standard CRF FINAL v1.0_RAND ...
- Sunovion Standard CRF FINAL v1.0_SU ...
- Sunovion Standard CRF FINAL v1.0_TC ...
- Sunovion Standard CRF FINAL v2 0 DS1 ...

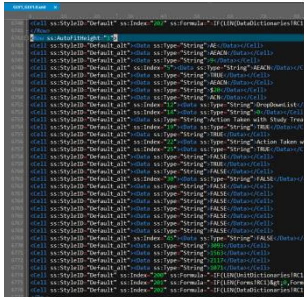
paper



metadata

CDASH

Global Library Volume Implementation Project



```
<Cell ss:StyleID="Default" ss:Index="202" ss:Formula="=IF(LEN(DataDictionaries!RC1
</Row>
<Row ss:AutoFitHeight="1">
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">AE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">AEACN</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">9</Data></Cell>
<Cell ss:StyleID="Default_alt" ss:Index="5"><Data ss:Type="String">AEACN</Data></C
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">TRUE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">AEACN</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">$20</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">ACN</Data></Cell>
<Cell ss:StyleID="Default_alt" ss:Index="12"><Data ss:Type="String">DropDownList</
<Cell ss:StyleID="Default_alt" ss:Index="14"><Data ss:Type="String">0</Data></Cell
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">Action Taken with Study Trea
<Cell ss:StyleID="Default_alt" ss:Index="19"><Data ss:Type="String">TRUE</Data></C
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">TRUE</Data></Cell>
<Cell ss:StyleID="Default_alt" ss:Index="22"><Data ss:Type="String">Action Taken w
<Cell ss:StyleID="Default_alt" ss:Index="25"><Data ss:Type="String">TRUE</Data></C
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">TRUE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt" ss:Index="30"><Data ss:Type="String">FALSE</Data></
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">FALSE</Data></Cell>
<Cell ss:StyleID="Default_alt" ss:Index="45"><Data ss:Type="String">FALSE</Data></
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">3093</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">1563</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">2117</Data></Cell>
<Cell ss:StyleID="Default_alt"><Data ss:Type="String">1071</Data></Cell>
<Cell ss:StyleID="Default" ss:Index="200" ss:Formula="=IF(LEN(UnitDictionaries!RC1
<Cell ss:StyleID="Default" ss:Index="201" ss:Formula="=IF(LEN(Forms!RC1)&gt;0,Form
<Cell ss:StyleID="Default" ss:Index="202" ss:Formula="=IF(LEN(DataDictionaries!RC1
```

metadata in ODM XML

SDTM

CT

Started FY2018 1Q

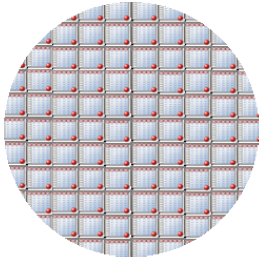
WHAT NEXT?

- Can I use **natural language** model or **machine learning** algorithms to map raw data variables to SDTM variables?
- Can I build a word model to represent the mapping specifications for ML training?

ML based SDTM mapping tool for fast, accurate, consistent SDTM generation

Input

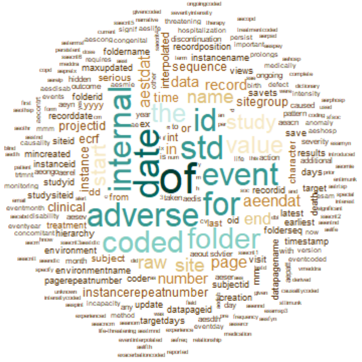
Raw data



CDASHIG, SDTM IG

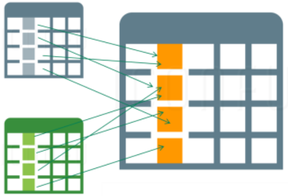
NLP & ML

Wordcloud for AE raw metadata in raw

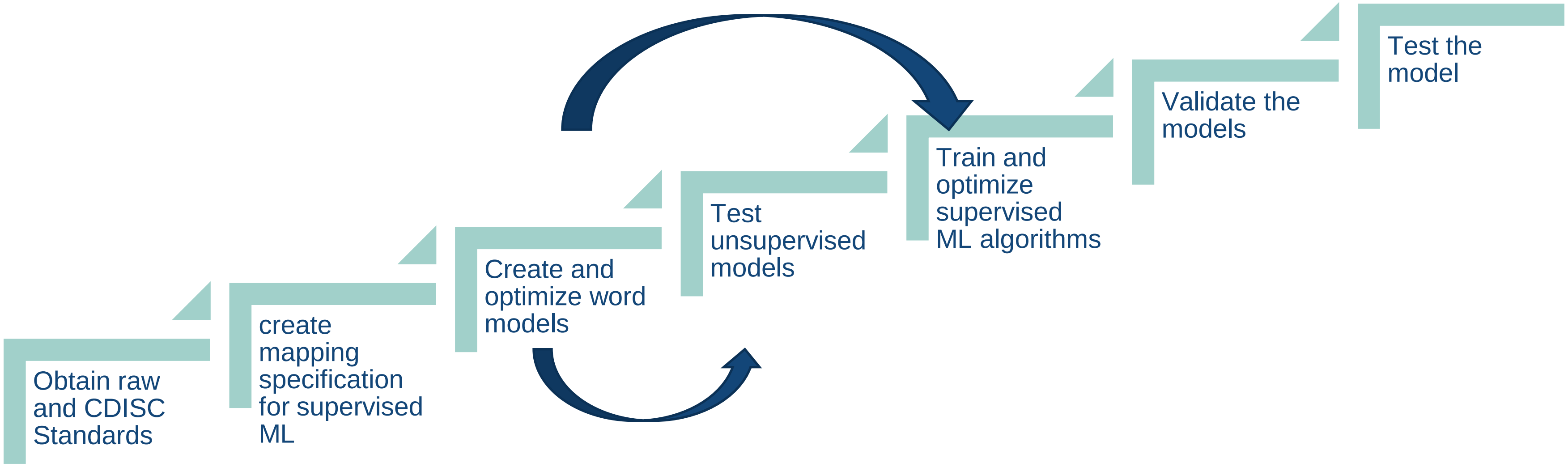


Output

Mapping Specification



Steps



Obtain raw and CDISC Standards

create mapping specification for supervised ML

Create and optimize word models

Test unsupervised models

Train and optimize supervised ML algorithms

Validate the models

Test the model

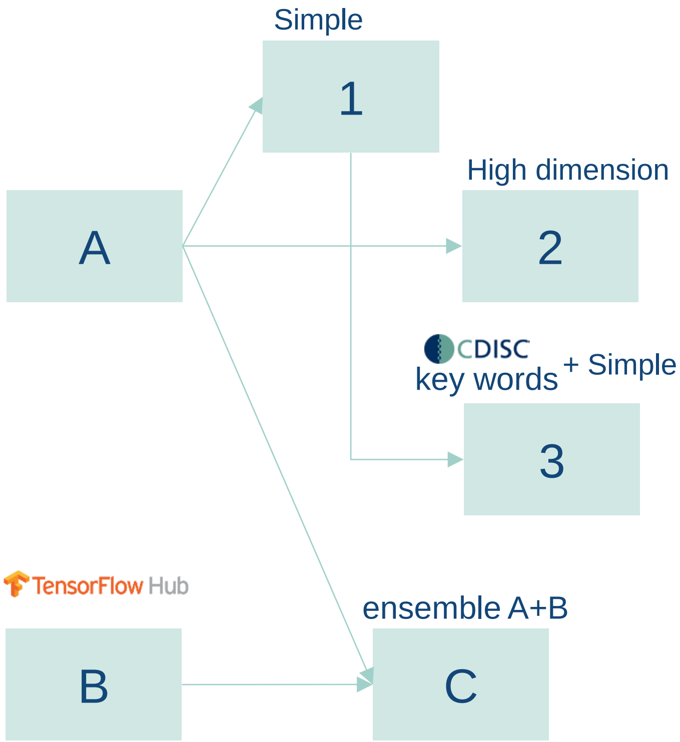
Use as input

Create mapping specifications (with Human Intelligence)

Raw Variable	SDTM Variable
PT	AEDECOD
SOC	AEBODSYS
PTNAME	AEDECOD
SOCNAME	AEBODSYS
…	…

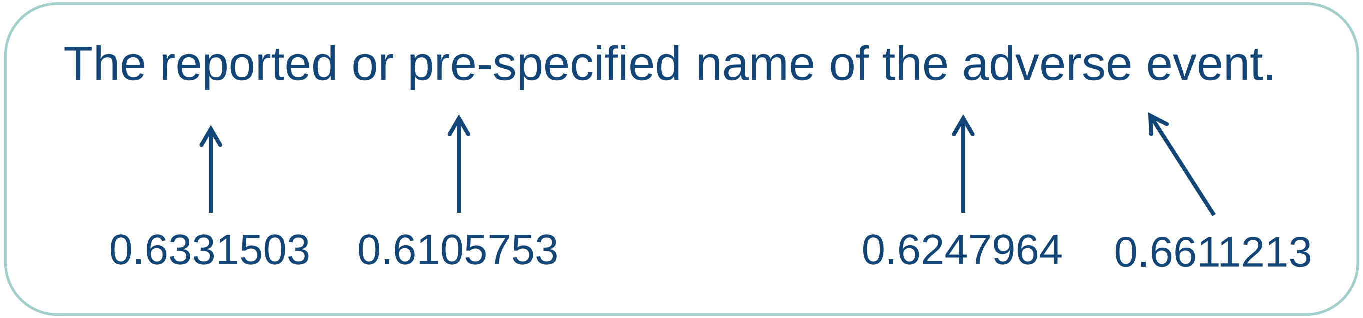
*illustration purpose only

Natural Language Models



Natural Language Model A

AETERM definition in CDASH IG

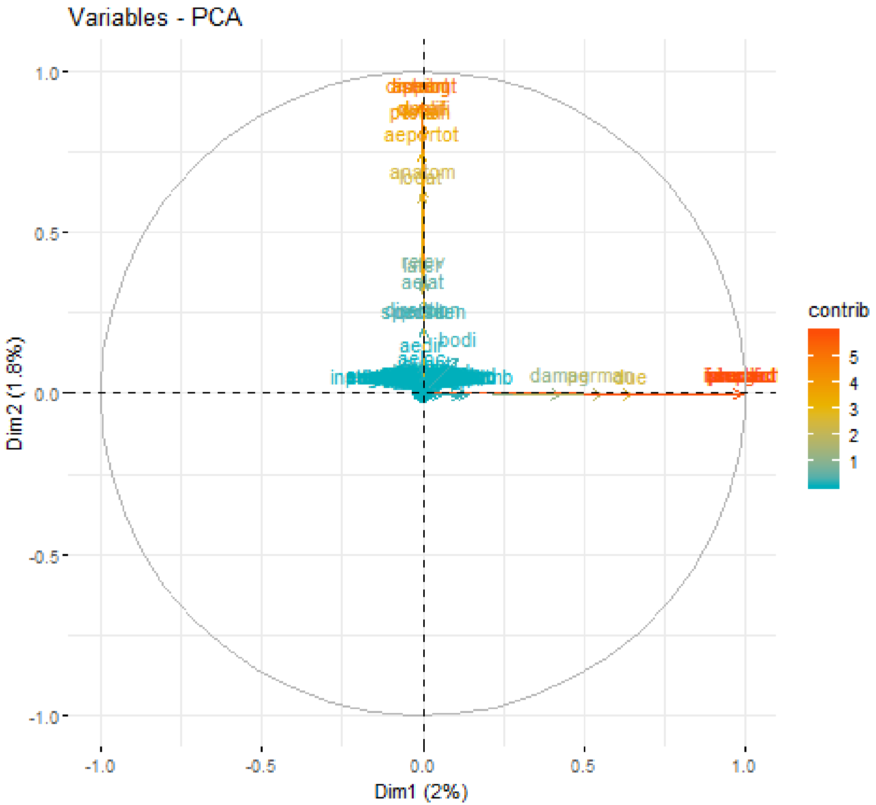


The reported or pre-specified name of the adverse event.

0.6331503 0.6105753 0.6247964 0.6611213

Natural Language Model A

This model yield **0.64058** accuracy with logistic regression from 10 fold cross validations in training



cdisc

Natural Language Model B

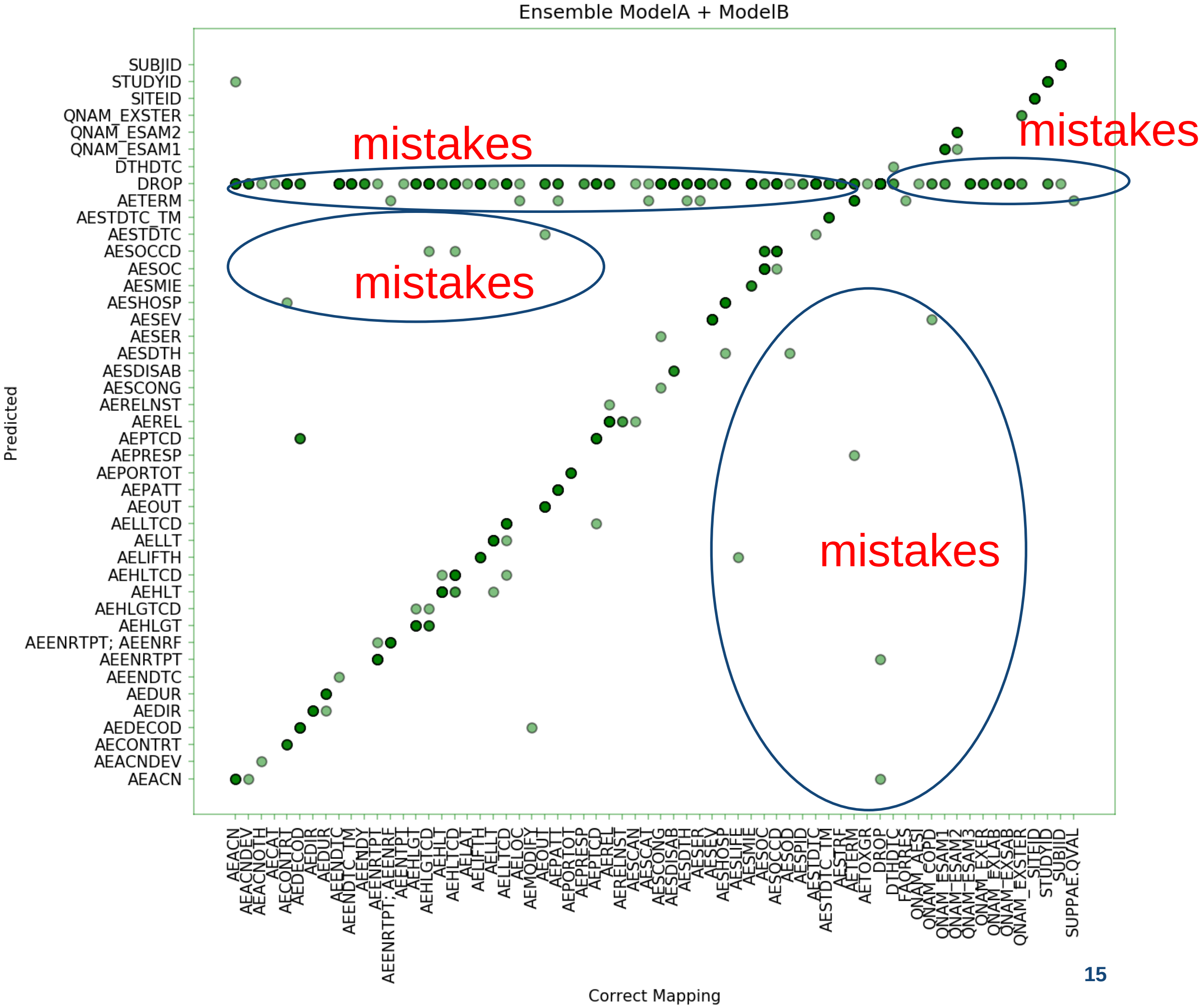
Language model based on feed forward neural network language models with 3 hidden layers trained on English Google News

TensorFlow Hub

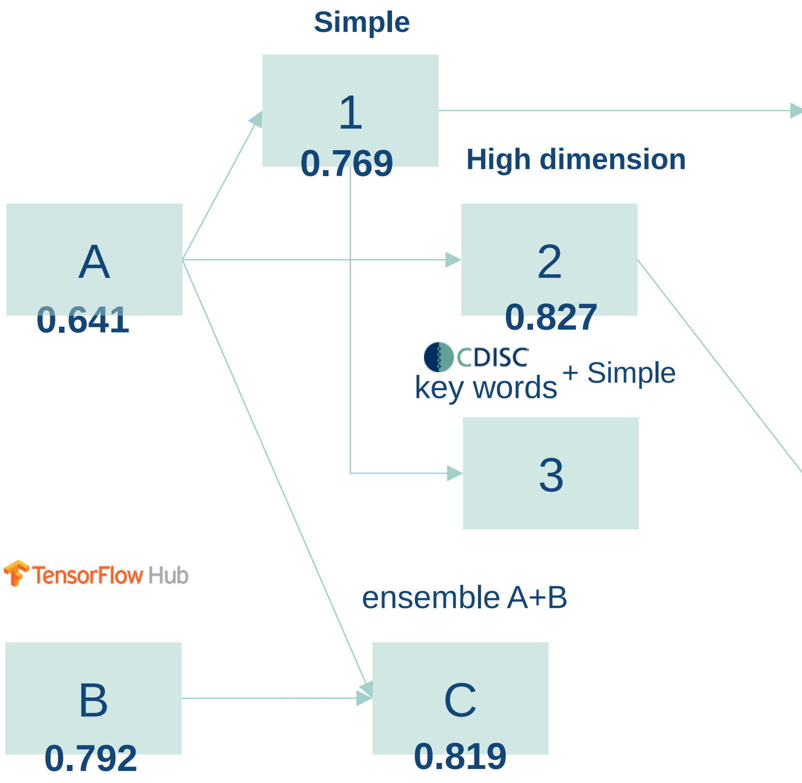
This model yield **0.79249** accuracy with a logistic regression from 10 fold cross validations in training

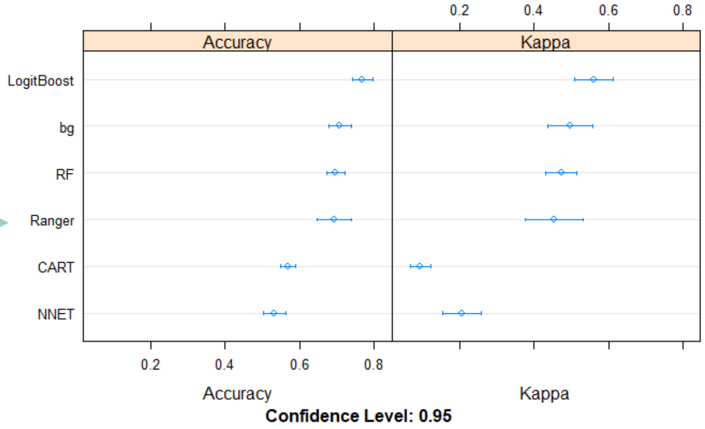
Mapping Accuracy on Training Data

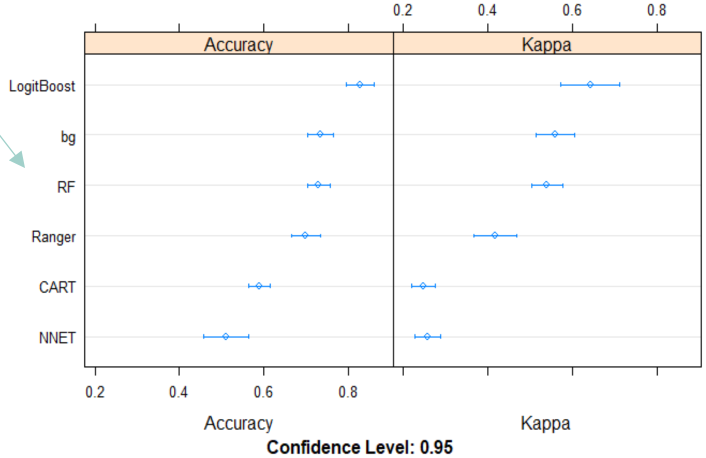
Model A + Model B: **0.81910** with logistic regression



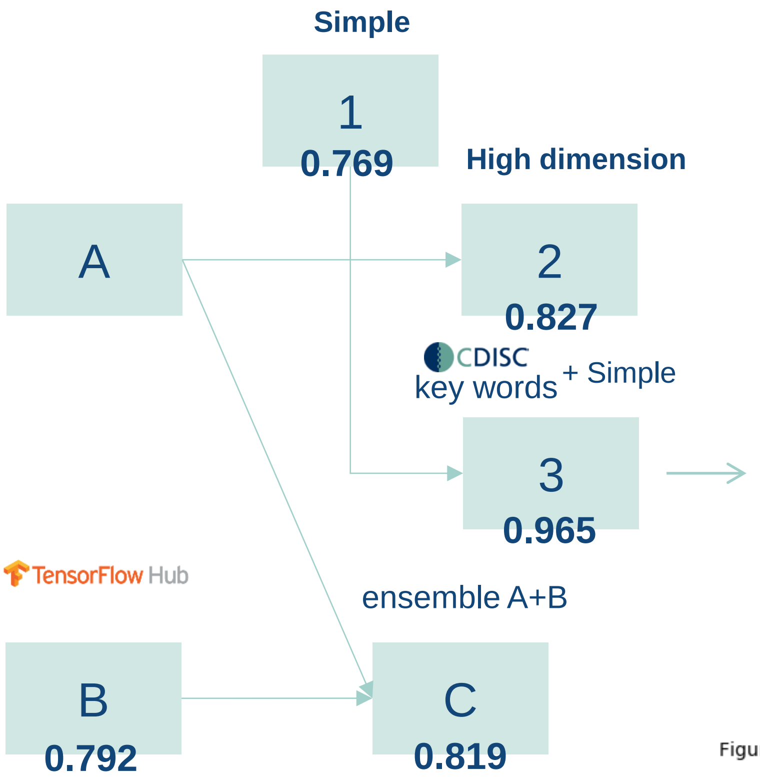
Natural Language Models







Natural Language Models



Simple

1
0.769

High dimension

A

2
0.827

CDISC key words + Simple

3
0.965

TensorFlow Hub

ensemble A+B

B
0.792

C
0.819

Machine Learning Models

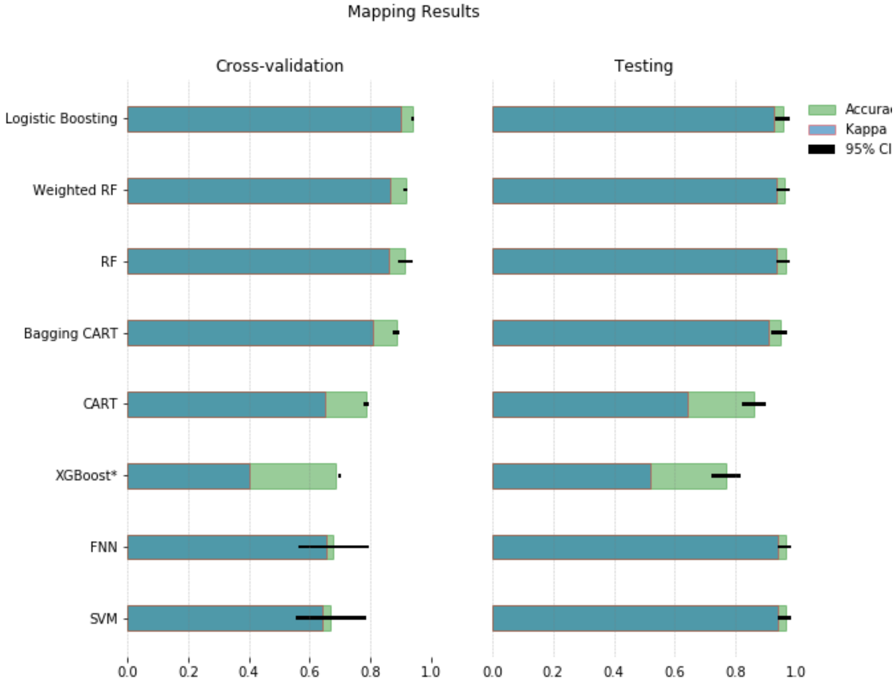
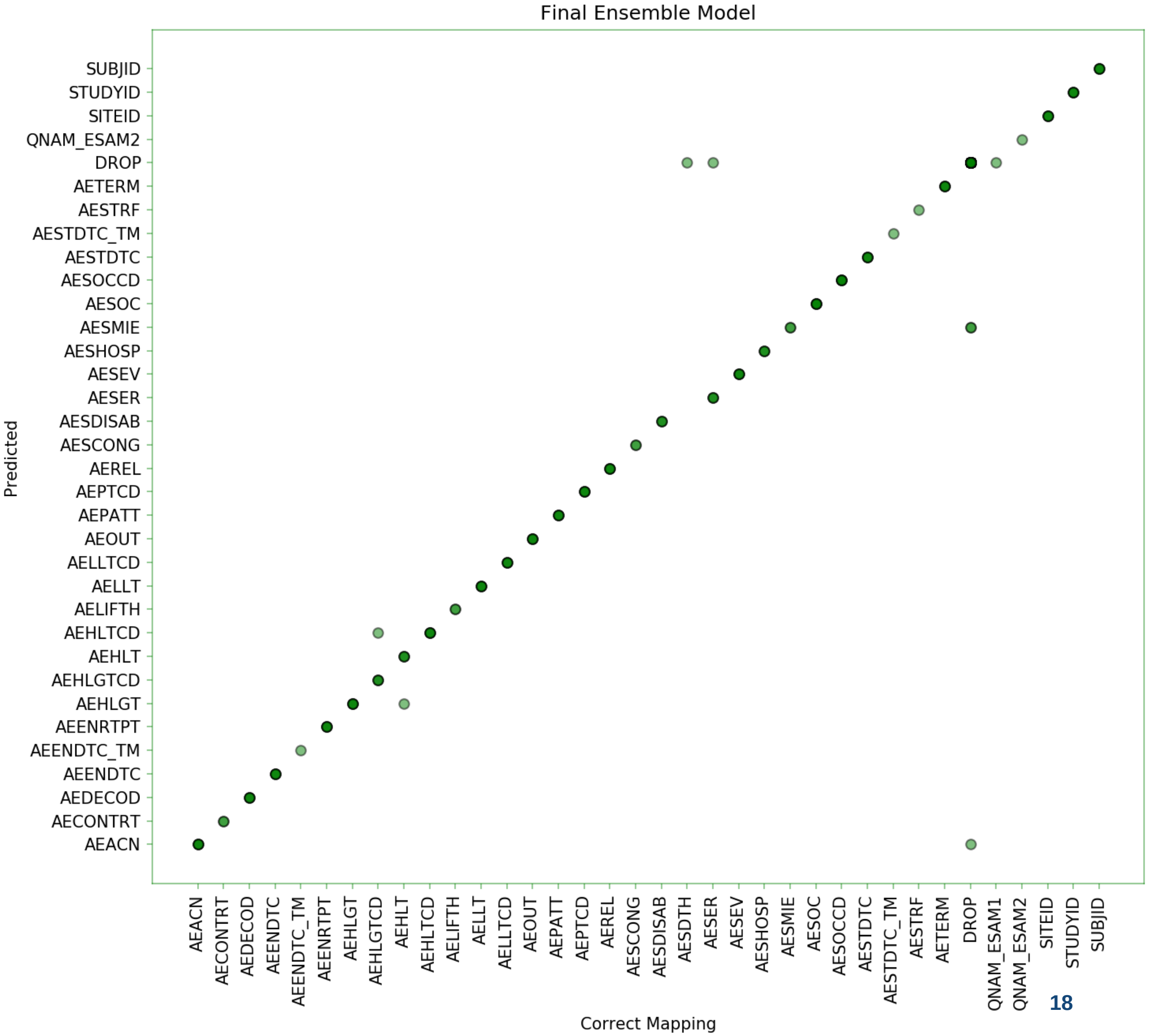


Figure 6. Results of SDTM mapping (Accuracy with 95% CI and Kappa)

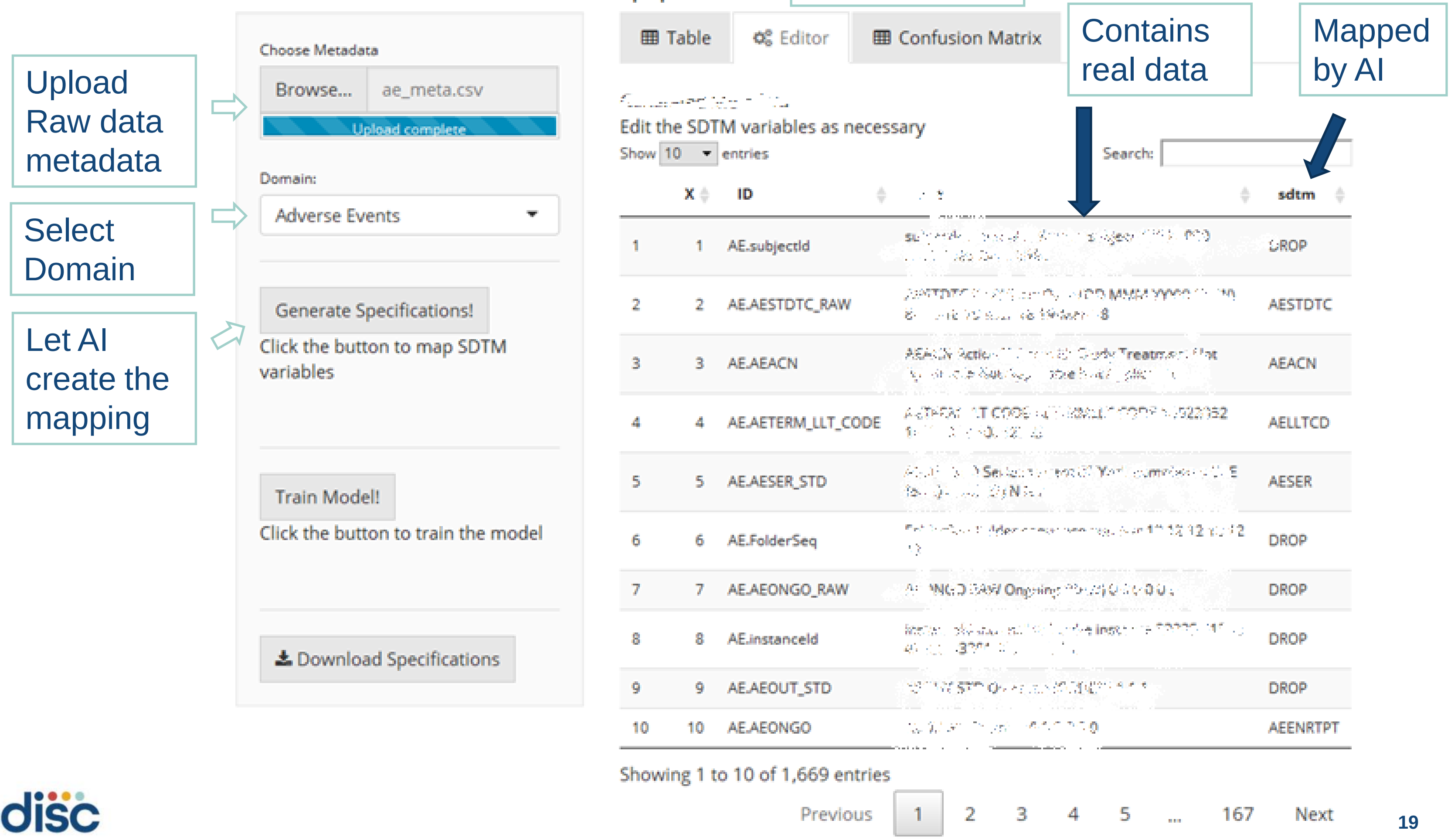
Mapping Accuracy on New Data using Final Ensemble Model

0.97484



Machine SDTM Mapper

Prototype GUI



Upload Raw data metadata

Select Domain

Let AI create the mapping

Contains real data

Mapped by AI

Choose Metadata

Browse... ae_meta.csv

Upload complete

Domain:

Adverse Events

Click the button to map SDTM variables

Click the button to train the model

[redacted]

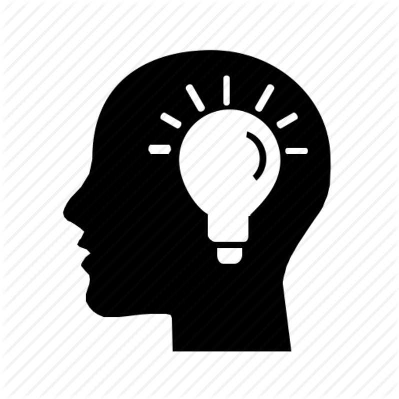
Edit the SDTM variables as necessary

Show 10 entries

	X	ID	[redacted]	sdtm
1	1	AE.subjectId	[redacted]	DROP
2	2	AE.AESTDTC_RAW	[redacted]	AESTDTC
3	3	AE.AEACN	[redacted]	AEACN
4	4	AE.AETERM_LLT_CODE	[redacted]	AELLTCD
5	5	AE.AESER_STD	[redacted]	AESER
6	6	AE.FolderSeq	[redacted]	DROP
7	7	AE.AEONGO_RAW	[redacted]	DROP
8	8	AE.instanceId	[redacted]	DROP
9	9	AE.AEOUT_STD	[redacted]	DROP
10	10	AE.AEONGO	[redacted]	AEENRTPT

Showing 1 to 10 of 1,669 entries

Previous 1 2 3 4 5 ... 167 Next



- This proof of concept demonstrated that **machine learning** along with a decent **natural language model** can produce a pretty accurate SDTM mapping specification document.
- Light weight simplistic approach works.
- Optimizing the model using a laptop without NVIDIA GPU is very challenging.
- As in any ML models, as you feed more mapping specs, the model will learn them and become more robust.

THANK YOU

www.linkedin.com/in/STomioka

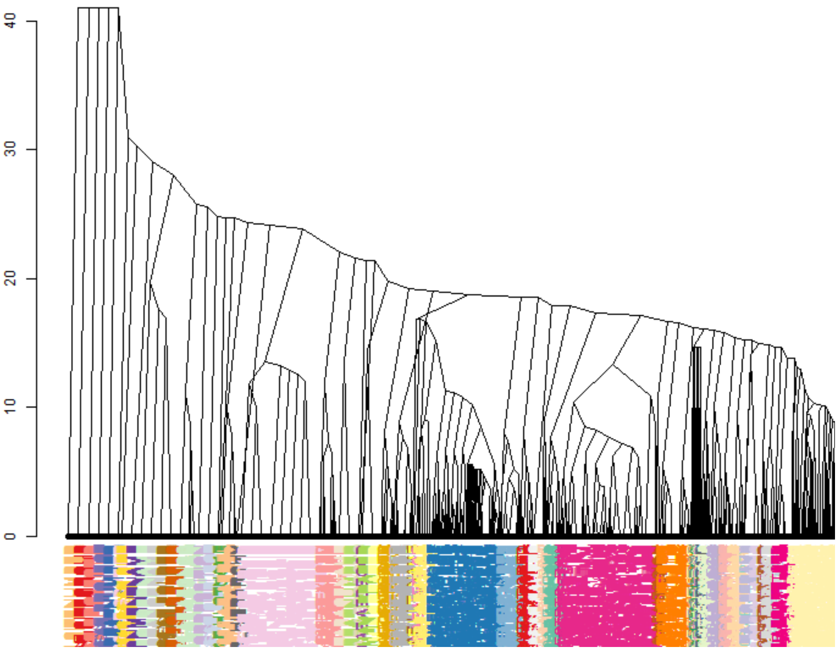
Sam.Tomioka @ Sunovion.com

cdisc

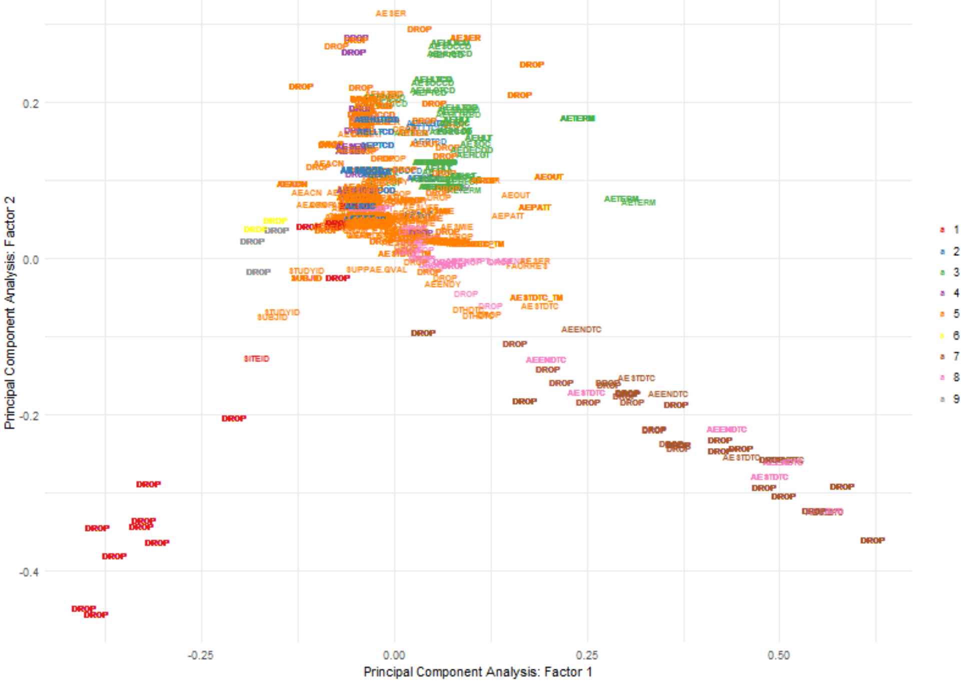
Back up

Unsupervised Machine Leaning Approach

Raw Variables - Euclidean Distance with Ward's linkage



K-Means Cluster: 9 clusters of SDTM variables on PCA Features


Principal Component Analysis: Factor 2
Principal Component Analysis: Factor 1
AESER
DROP
AETERM
AEOUT
AEPATT
AEACN
AEENDTC
AESTDTC
AESTDTC_TM
AEENDY
STUDYID
SUBJID
SUPPAE.QVAL
SITEID
DTHDTC
AESME
FAORRES