



SDTM Mapping and Data Conversion Automation using ML/NLP

www.Sunovion.com

www.Symbiance.com

THE PROBLEM

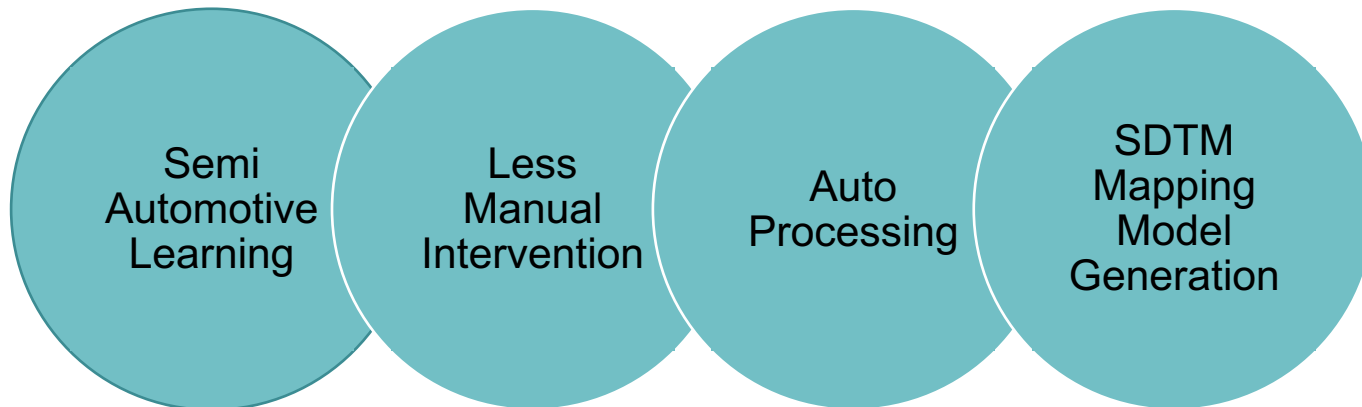
- Mapping of non standard raw data to SDTM is manual and time consuming.
- Mapping is repetitive process.
- Endless mapping and multiple mapping decisions for same scenario.
- Require highly skilled resources to execute.
- No automotive way to handle errors and multiple iterations.
- Identifying data types and code lists are highly time consuming.

WHY AUTOMATION

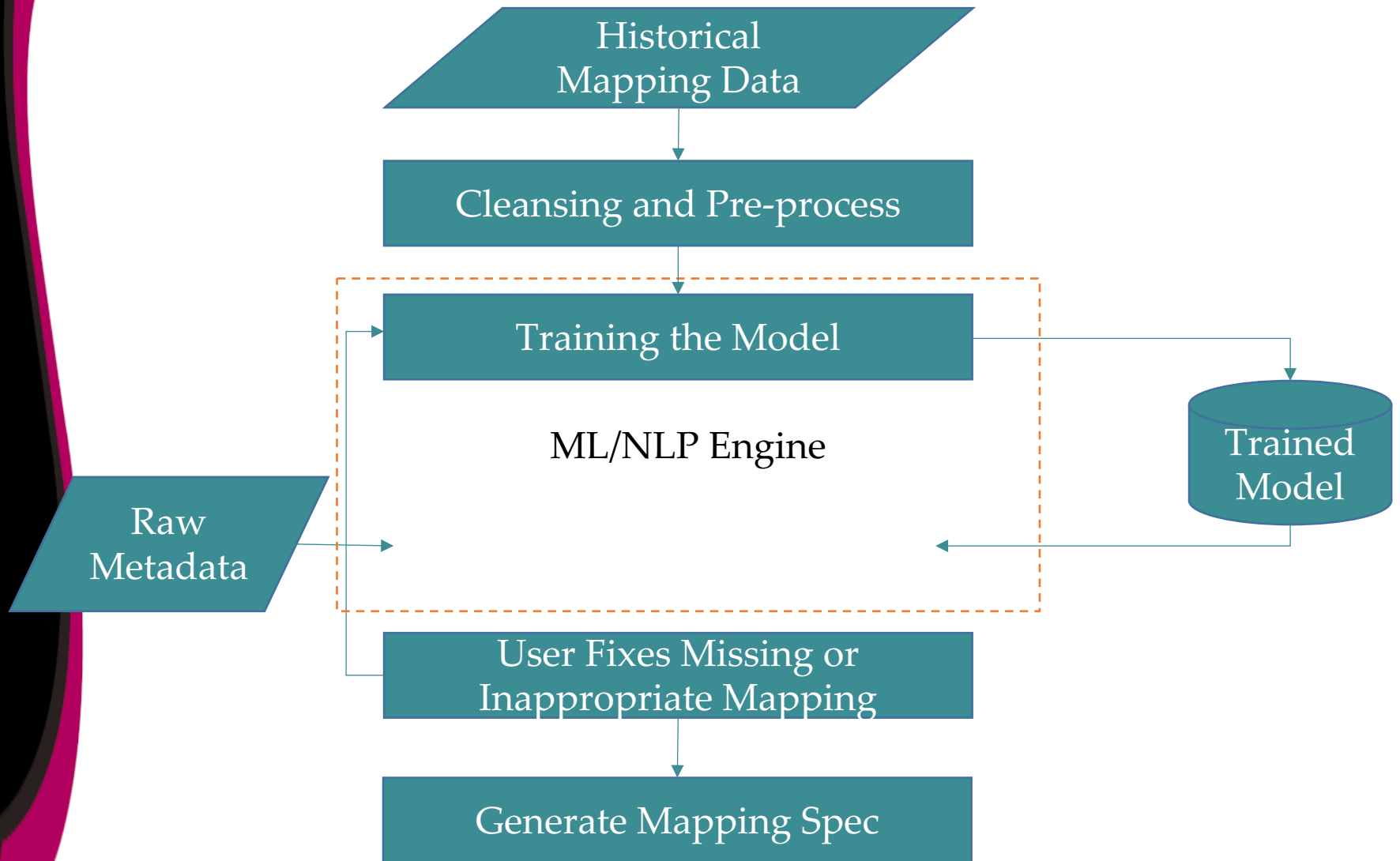
- 60%-80% of manual effort can be saved with more study data using the ML/NLP tool.
- Effective use and reuse of Study and Project Metadata.
- Mapping and reuse of Metadata between CDISC Metadata/Sponsor defined standards (including Metadata, IG Version, Control Terminologies, etc.).
- The ML engine is constantly trained for greater accuracy and prediction. Effective reuse of mapping decisions.
- The ML/NLP tool creates the SDTM Mapping Specs and standalone SAS® programs for each domain.
- Greatest advantage for choosing ML/NLP - automate all possible derivation logic, which is key time savings for programmers.

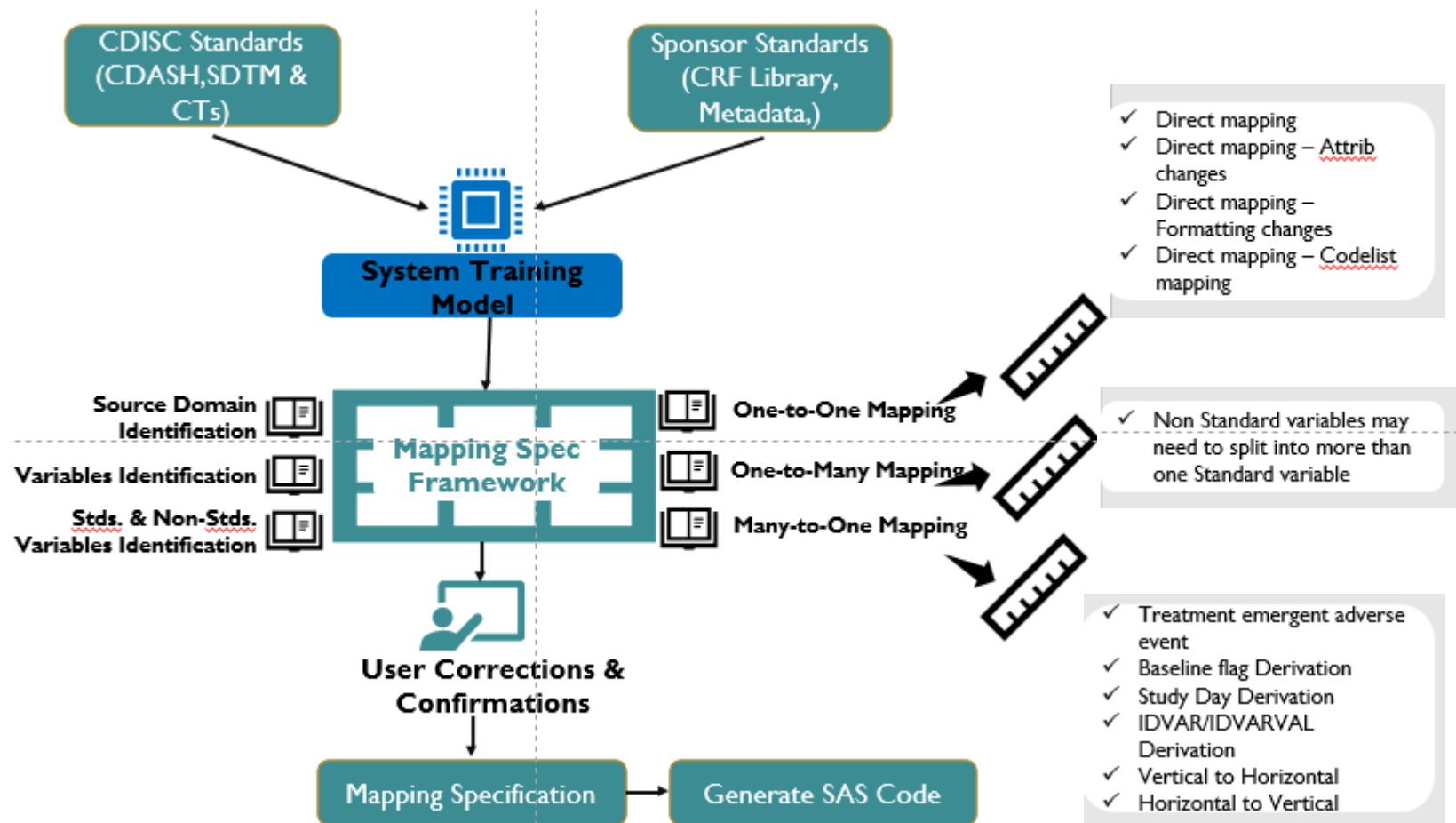
NEED OF MACHINE LEARNING

- Self-learning improves the accuracy over the time.
- Works well on unknown data.
- System acts as an expert by itself from the earlier learnings to fix issues.
- Percentage of fallouts are very minimum to handle manually.

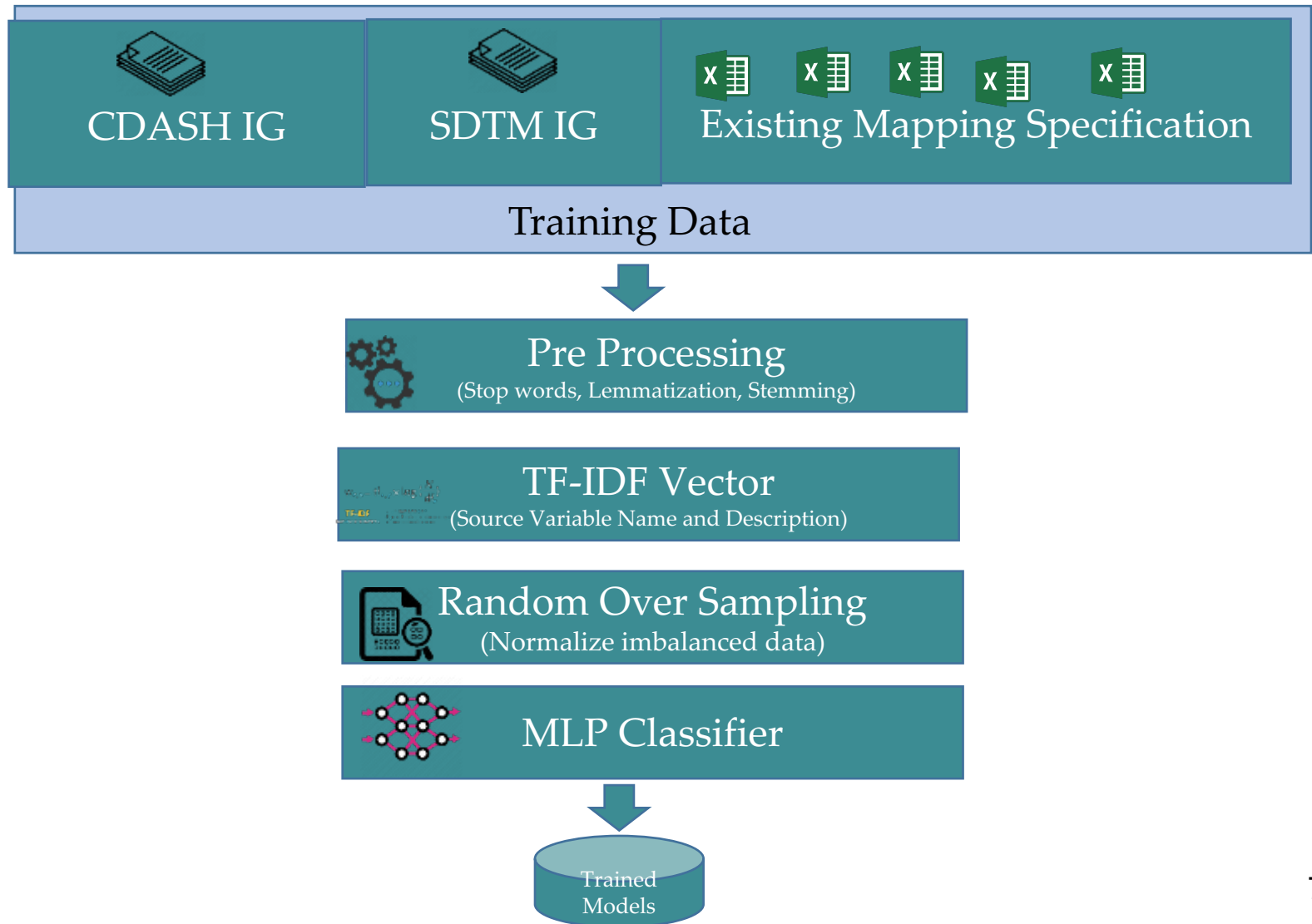


HOW TO APPLY ML/NLP ?





PROCESS TO GENERATE CLASSIFICATION MODEL



Continued

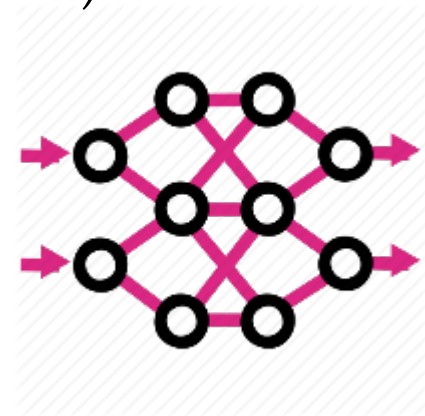
- Term Frequency – Inverse Document Frequency - TF-IDF
 - *Statistical representation of importance's of a word in a document*
 - *Used to weight the term in variable information in a vector*
 - *Vector will feed in to MLP Classifier*

$$w_{x,y} = \text{tf}_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

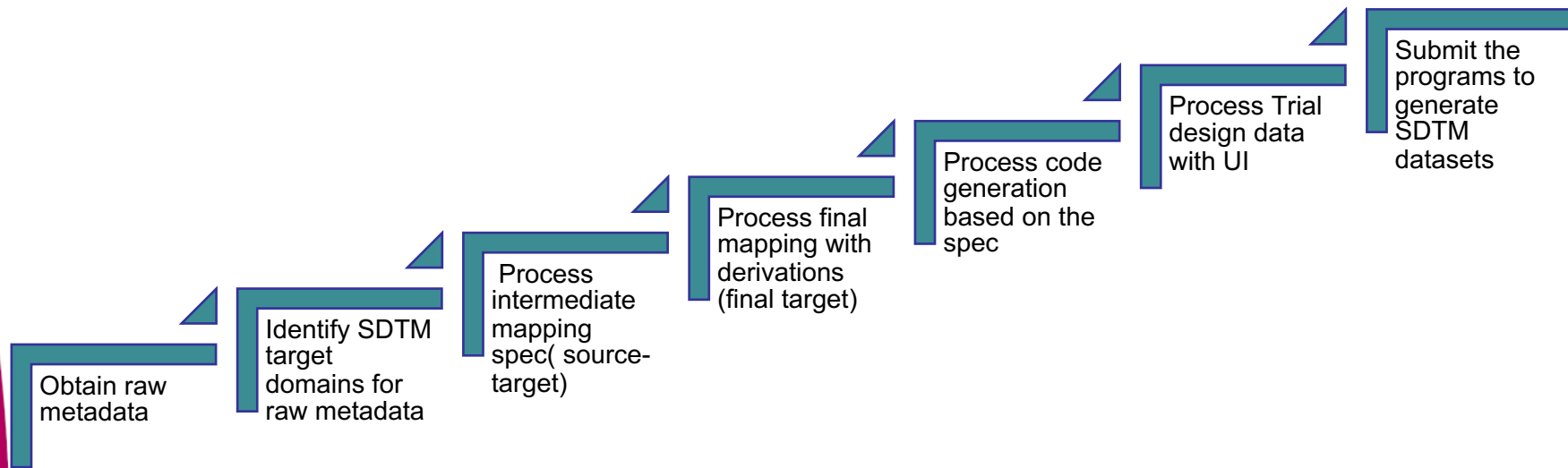
TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
 df_x = number of documents containing x
 N = total number of documents

- Multi-layer Perceptron classifier (MLP Classifier)
 - *Simple and lightweight neural network*
 - *Works better with minimal training data*
 - *Less resource intensive*



STEPS



MINNAL - DOMAIN MAPPING

- System trained based on CDASH and CDISC standards. System can also be retrained based on the Sponsor standards.
- System will automatically predict the primary domains and secondary domains based on “Topic” variables collection.
- System will also provide the functionality to edit the domain prediction. Once it is changed, System will retrain itself and it will be available for further studies within the same project.
- Retraining can be controlled by governance team/authorized role.

Upload

Domain Mapping

Intermediate Mapping

Trial Datasets

Final Mapping

Download Specification

Raw Dataset Name	Primary SDTM Domain	Secondary SDTM Domains	Action
demo	Choose Domain DM	Choose Domains EX, SC	Edit
comlog	Choose Domain CO	Choose Domains SC	Edit
aecdisc	Choose Domain SC	Choose Domains AE	Edit
aed	Choose Domain SC	Choose Domains QS	Edit
ecgdet	Choose Domain SC	Choose Domains QS	Edit
fnt301	Choose Domain QS	Choose Domains	Edit
formats	Choose Domain QS	Choose Domains	Edit
final	Choose Domain SC	Choose Domains	Edit
epmfx	Choose Domain MH	Choose Domains QS, SC	Edit
label	Choose Domain SC	Choose Domains	Edit

MINNAL - INTERMEDIATE MAPPING

- Based on the keyword, System will also predict the target variable, and will also provide the possible matches.
- User can also manually map the target variable or create the custom domain. System will get retrained based on the user changes. The mapping will be available for subsequent studies.
- System will allow the user to provide a mapping and derivation comments that are used in the specification and eventually the program generator.

The screenshot displays the Minnal.AI web application interface for the 'DEMOSTUDY' project. The 'Intermediate Mapping' tab is active, showing a table with columns: Source Domain, Source Variable, Source Label, Datatype, SDTM Domain, and SDTM Variable. The table lists mappings for various adverse event (AE) related variables. A 'Possible Variations' tooltip is visible over the 'USUBJID' mapping.

Source Domain	Source Variable	Source Label	Datatype	SDTM Domain	SDTM Variable
AE	studyid	Internal Id for the study	Numeric	AE	STUDYID
AE	Subject	Subject name or identifier	Text	AE	USUBJID
AE	AETERM	Adverse Event	Text	AE	AETERM
AE	AETERM_LL1	Adverse Event_LL1	Text	AE	AELL1
AE	AETERM_LL1_CODE	Adverse Event_LL1_CODE	Text	AE	AELL1CD
AE	AETERM_PT	Adverse Event_PT	Text	AE	AED000
AE	AETERM_PT_CODE	Adverse Event_PT_CODE	Text	AE	AEP1CD
AE	AETERM_HLT	Adverse Event_HLT	Text	AE	APHLT

MINNAL - FINAL MAPPING SPEC

- System will provide the Key variables and sorting suggestion based on CDISC. User can modify based on the data collection, that will be used in generated program.
- System will have a standard derivation from global metadata and user provided derivations and automatically identify transpose finding type variables.

The screenshot displays the Minnal.AI interface for the 'Final Mapping' stage. The top navigation bar includes 'Upload', 'Domain Mapping', 'Intermediate Mapping', 'Final Mapping', and 'Trial Datasets'. The 'Final Mapping' tab is active, showing a table of CDISC domains and their corresponding variables.

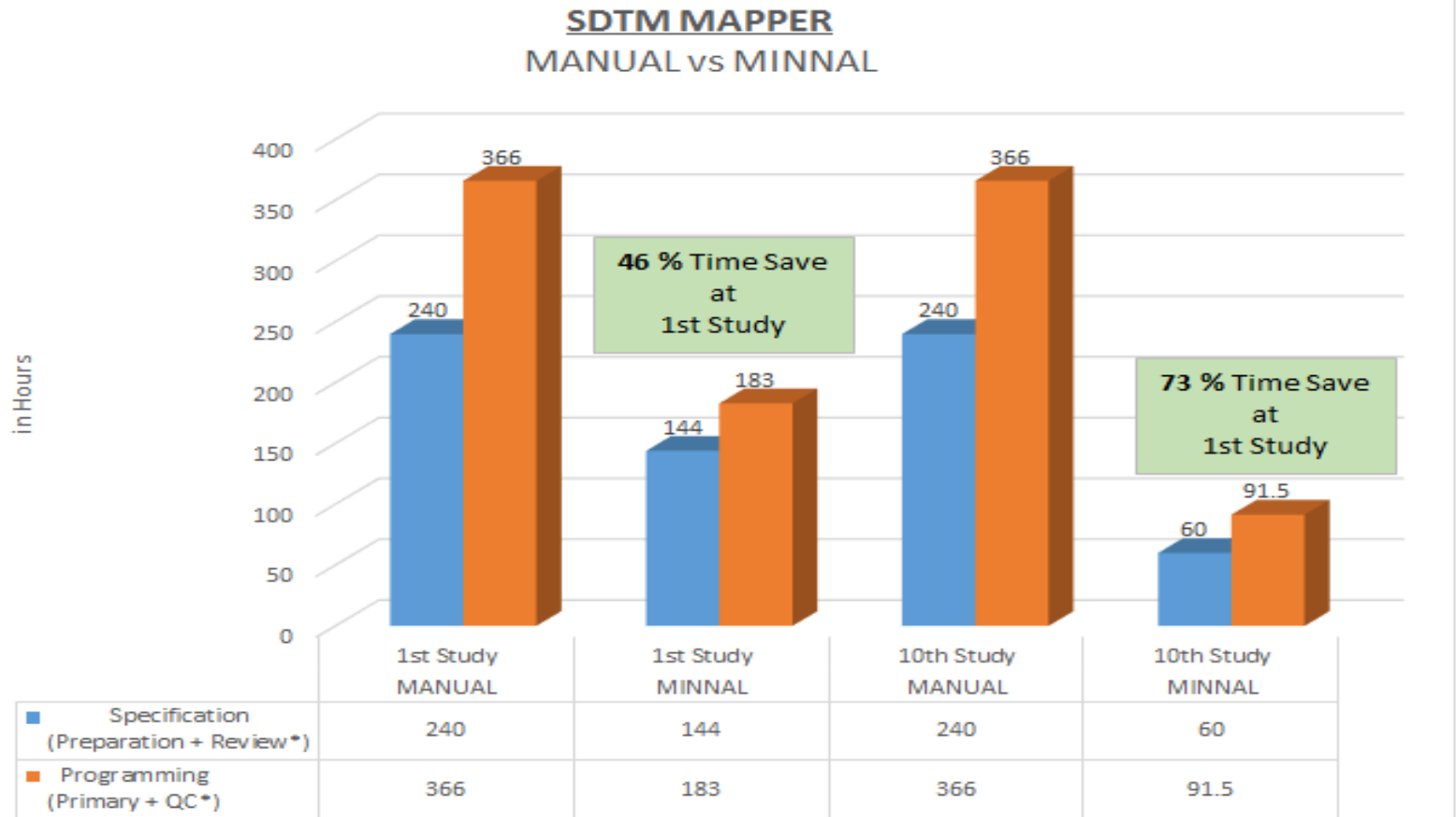
CDISC Variable	Controlled Vocab	EDTM Label	Case
STUDYID		Study Identifier	Required
DOMAIN	DOMAIN	Domain Abbreviation	Required
USUBJID		Unique Subject Identifier	Required
PROBID		Prod Identifier	Permissible
SPONSID		Sponsor Device Identifier	Permissible
ASGID		Assignment Number	Required
GRPID		Group ID	Permissible
REFID		Reference ID	Permissible
REFID		Reference/Ordered Identifier	Permissible
LINKID		Link ID	Permissible

The right-hand panel shows the 'SDTM Domain' details for 'STUDYID'. It includes a 'Mapping' section with the text 'source domain source variable' and a 'Derivation' section with the text 'get(studyid, best)'.

MINNAL - ADDITIONAL FEATURES

- System will generate separate spec for SUPP-- domains.
- System will perform default consistency check, like --DTC and --DY. It can be configurable by the user requirement.
- System will produce a SAS® program per domain, whereas the primary domain program will have the functionality to create its SUPP-- domain.
- System will take SDTM domain's variable order and its attributes to meet the submission guideline.

MANUAL VS MINNAL



* By Manual

Input as based on High criticality study which containing 50 sources datasets

Derived on basis of 21 standard safety domain, 4 Custom domains with SE, SV, RELREC and five Trial domains

CONCLUSION

- ML/NLP helps to solve the problem with the historical data.
- It can predict new mapping scenarios.
- It can learn from the user action when System cannot determine a scenario.
- The accuracy will improve over time based on the volume of data.
- System was trained using 8 historical studies and tried on 3 new studies.
- Concluded that this approach could save significant time and resources.