

Data Mapping using Machine Learning

Daniel Choi, Eric Bolender, Stijn Rogiers
Daniel.Choi@sas.com Eric.Boledner@sas.com Stijn.Rogiers@sas.com

Introduction

Mapping raw data to standards is one of the most challenging process in the healthcare industry. Reusing or reapplying the information collected during mapping processes from previously mapped studies and building upon that knowledge inference is the most important part of the mapping process. Most companies struggle with building knowledge inferences and reapplying them through the data process efficiently. In addition, sometimes, there are multiple tools/programs required to go through full-cycle of data mapping. Therefore, it is difficult for the standard user to know all the different versions and tools, and use them correctly throughout the data mapping process.

The SAS Data Mapping Tool is a web-based tool that provides a user-friendly interface for everything from mapping raw data to generating SDTM standards (including domain templates). Simple User Interface (UI) and click-away concept design provides access to all the required information on a single screen. Auto-mapping and smart-mapping features in the SAS Data Mapping Tool, which are based on knowledge inference derived from machine learning algorithms, reduce time and effort for the user. This leads to improvements in quality, efficiency and consistency.

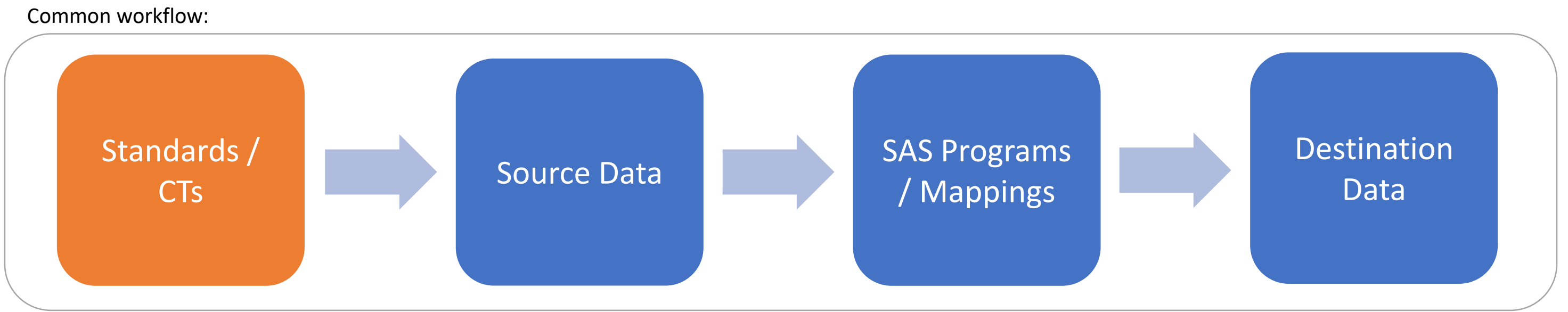
Problem Definition

Source Data is mapped to Destination Data based on Standards for Clinical Studies

Standards are guidelines rather than instructions, leaving room for individual interpretation

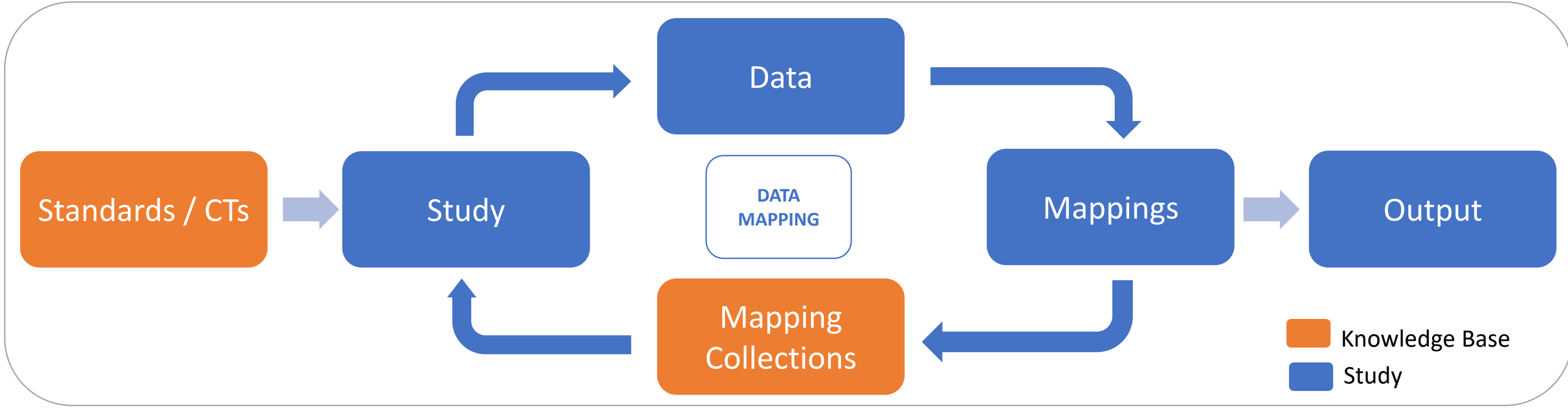
Mapping is currently done in individual SAS programs which:

- 1. Lacks collection of central metadata
- 2. Leads to inconsistencies in way mapping is done across different studies
- 3. Only way to re-use previous information is copy-and-paste of programs



Solution

To solve this problem, we have created an application that follows the flow shown below:



Components	Description
Standards/CTs	Define Standards and Controlled Terminologies. (ex SDTM, ADaM or company specific)
Study	Define metadata such as Standards/CTs to be used.
Data	Define Data Sources.
Mappings	Provides ability to map source to destination data through these steps: data transform, table/variable mapping, and table/variable transform. This is where auto/smart mapping (Machine Learning) will be applied.
Mapping Collections	Ability to publish mapping metadata into different libraries.
Output	Generate SAS Datasets/Programs

Automap is a process where the application will suggestion mappings based upon what has been published to the mapping collections.

Smartmap is a process where machine learning process will be used to find mapping suggestions to display to users what is recommended to map based upon their current study.

This application allows users to train the model and efficiently map their raw datasets to a standard that they have defined which ultimately improves quality, efficiency, and consistent output results.

Machine Learning Design Flow

STEP	DESCRIPTION	MODULES	PROCESS
1	Read Data	Pandas	Read data into Dataframe
2	Cleanup Data	Pandas Dataframe	Drop NaN Records, convert to lower case, combine columns, Filter categories based on MIN count
3	NGRAM Data	NGRAM	NGRAM data using Character ngram with limit of 2
4	Pickle NGRAM	Pickle	Ngram pair of (Source, Destination) value and pickle it
5	String to Numeric Value conversion	TfidfVectorizer, MinMaxScaler	Generate Dictionary with TF-IDF values and re-scale it using MinMaxScaler
6	Balance Data	SMOTETomek	Balance data to handle under-sampling and over-sampling
7	Evaluate Models	sklearn	LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, GaussianNB, MultinomialNB, OneVsRestClassifier (LinearSVC), OneVsRestClassifier (SGDClassifier)
8	Compute Model Params	sklearn	GridSearchCV
9	Train Model	sklearn	Fit model
10	Score Model	sklearn.metrics	Classification_report – Precision, Recall, F1 Score
11	Pickle Model	Pickle	Save the trained model in pickle file
12	Predict	Sklearn	Predict based on trained model

Variables Mapping Sample and Scores

After word_ngram created for all dest_var

Source var	Dest var	model_ngram	word_ngram	Classification Report for OVR_SVC After Model has been scored				
adverse.subject_id	ae.usubjid	ad dv ve er rs se e.	(ae.usubjid, adverse.subject_id)	precision	recall	f1-score	support	
	ae.usubjid	.s su ub bj ji id						
aedata.patid	ae.usubjid	ae ed da at ta a. .p	(ae.usubjid, aedata.patid)	ae.aerel	1.00	1.00	1.00	3
	ae.usubjid	pa at ti id		ae.aeser	1.00	1.00	1.00	2
adverse.startdt	ae.aestdtdc	ad dv ve er rs se e.	(ae.aestdtdc, adverse.startdt)	ae.aestdtdc	1.00	1.00	1.00	3
	ae.aestdtdc	.s st ta ar rt td dt		ae.aetert	1.00	1.00	1.00	2
				ae.usubjid	1.00	1.00	1.00	2
				cm.cmdosu	1.00	1.00	1.00	2
				cm.cmroute	1.00	0.60	0.75	5
				cm.usubjid	1.00	1.00	1.00	4
				dm.age	1.00	1.00	1.00	3
				dm.arm	1.00	1.00	1.00	2
				dm.dmdtc	1.00	1.00	1.00	2
				dm.height	0.00	0.00	0.00	2
				dm.investigator	1.00	1.00	1.00	2
				dm.race	1.00	1.00	1.00	2
				dm.sex	1.00	1.00	1.00	1
				dm.siteid	0.67	1.00	0.80	2
				dm.subjid	0.67	1.00	0.80	2
				dm.usubjid	1.00	0.33	0.50	3
				dm.weight	0.00	0.00	0.00	3
				ex.dose	1.00	1.00	1.00	5
				ex.freq	1.00	1.00	1.00	2
				ex.route	0.00	0.00	0.00	0
				ex.usubjid	1.00	1.00	1.00	1
				ex.visitnum	1.00	1.00	1.00	2
				lb.usubjid	1.00	1.00	1.00	3
				avg / total	0.89	0.85	0.86	60

Results From Machine Learning

If similarity > 0.9, show 1 match, else if similarity < 0.9, show Top 3 match

Test Data:

```
# Define TEST Data and Expected Outcomes
X_test = np.array(['adverse.ptid', 'hema.subjid', 'conmed.patid', 'dm.birthdate', 'exposure.dosage'])
y_test = np.array(['AE.USUBJID', 'LB.USUBJID', 'CM.USUBJID', 'DM.AGE', 'EX.DOSE'])
```

Test Results:

```
Predictions: ['ae.usubjid' 'lb.usubjid' 'cm.usubjid' 'dm.age' 'ex.dose'] Expected: ['AE.USUBJID', 'LB.USUBJID', 'CM.USUBJID', 'DM.AGE', 'EX.DOSE']
```

	Search_Term	Model_Matched_Term	Model_Similarity	Ngram_Matched_Term	Ngram_Similarity
0	adverse.ptid	ae.usubjid	0.412617	ae.usubjid	1.000000
1	hema.subjid	lb.usubjid	0.474403	lb.usubjid	1.000000
2	conmed.patid	cm.usubjid	0.699372	cm.usubjid	1.000000
3	dm.birthdate	dm.age	0.682097	dm.age	0.578947
4	dm.birthdate	dm.usubjid	0.022968	dm.age	0.444444
5	dm.birthdate	dm.subjid	0.021280	dm.brthdtdc	0.444444
6	exposure.dosage	ex.dose	0.684286	ex.dose	0.684211
7	exposure.dosage	dm.usubjid	0.022682	ex.dose	0.684211
8	exposure.dosage	dm.subjid	0.020918	ex.route	0.434783

Use combination of Model + NGRAM similarity to predict output values.