

Data Harmonization Using Machine Learning

Paper ML09

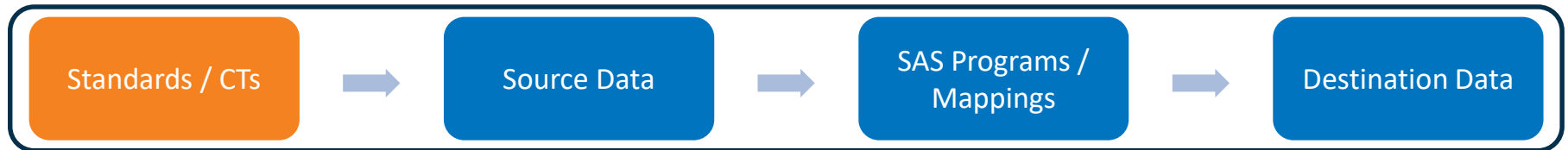
Presented by Lex Jansen, SAS Institute Inc.



PhUSE US Connect,
February 27, 2019
Baltimore, MD

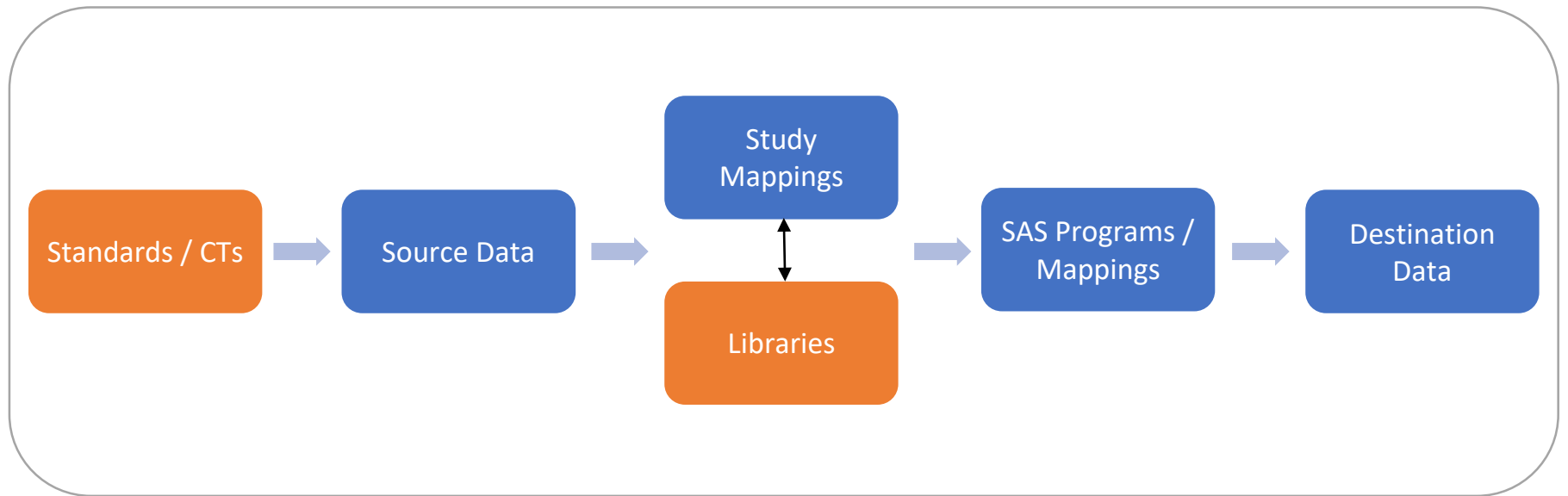
Data Mapping

Problem Definition



- Source Data is mapped to Destination Data based on Standards for Clinical Studies
 - Time-consuming process
- Standards are guidelines rather than instructions, leaving room for individual interpretation
 - Inconsistent Mappings – Not what a “Standard” should be
- Mapping is currently done in individual SAS programs which:
 1. Lacks collection of central metadata
 2. Only way to re-use previous information is copy-and-paste of programs

Data Mapping Solution

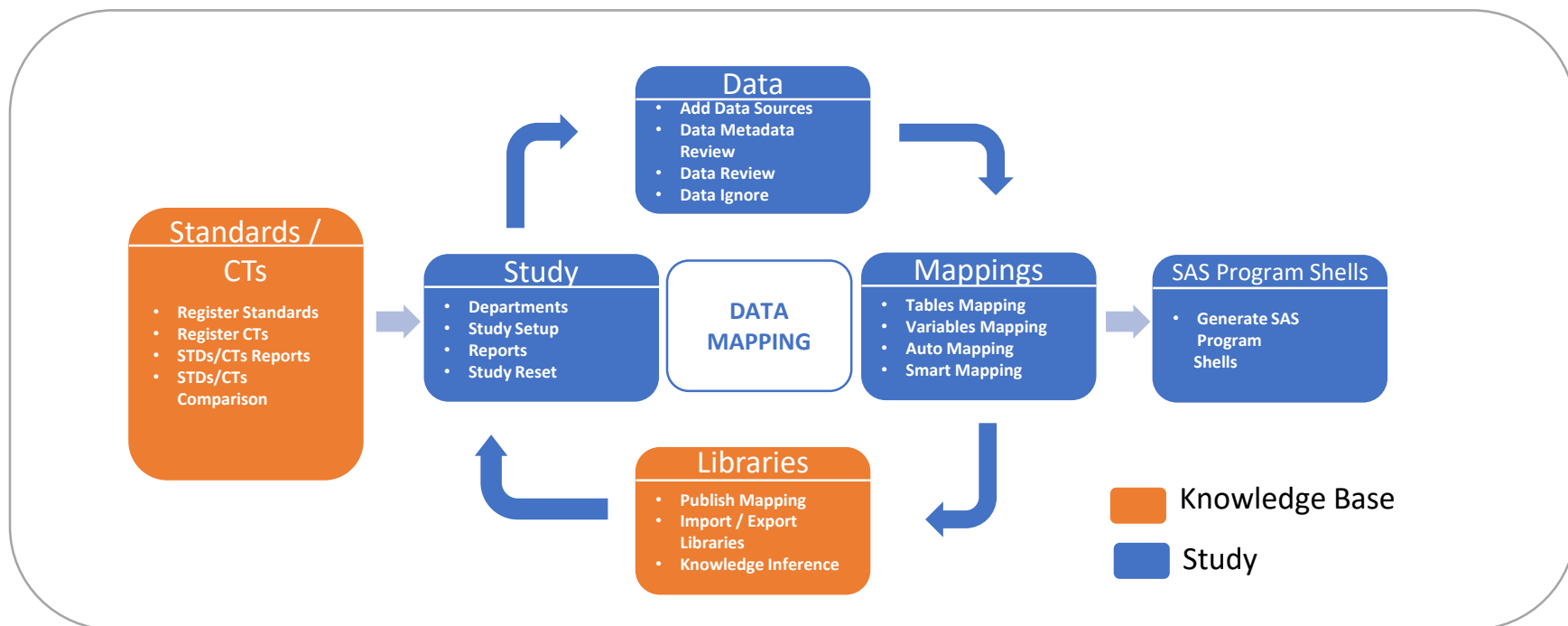


Data Mapping

Solution

- User interface to mapping process
- **Libraries** - Collection of mapping rules in central database
- Collect data and standards, and allow user to define rules for mapping
- Ability to re-use mapping rules in future studies
- **Auto Mapping** - One-to-One mapping Rule between Source Data and its variable, to Destination Data and its variable
 - e.g. adverse.patid → AE.USUBJID
 - Next study, adverse.patid is automatically mapped
- **Smart Mapping** - Similar Variables mapped in past provide guidelines to map new variables
 - e.g. adverse.ptid → AE.USUBJID
 - Suggested mapping based on previous mapping, adverse.Patid → AE.USUBJID
- Generate the SAS programs based on defined mapping

Data Mapping Process Flow



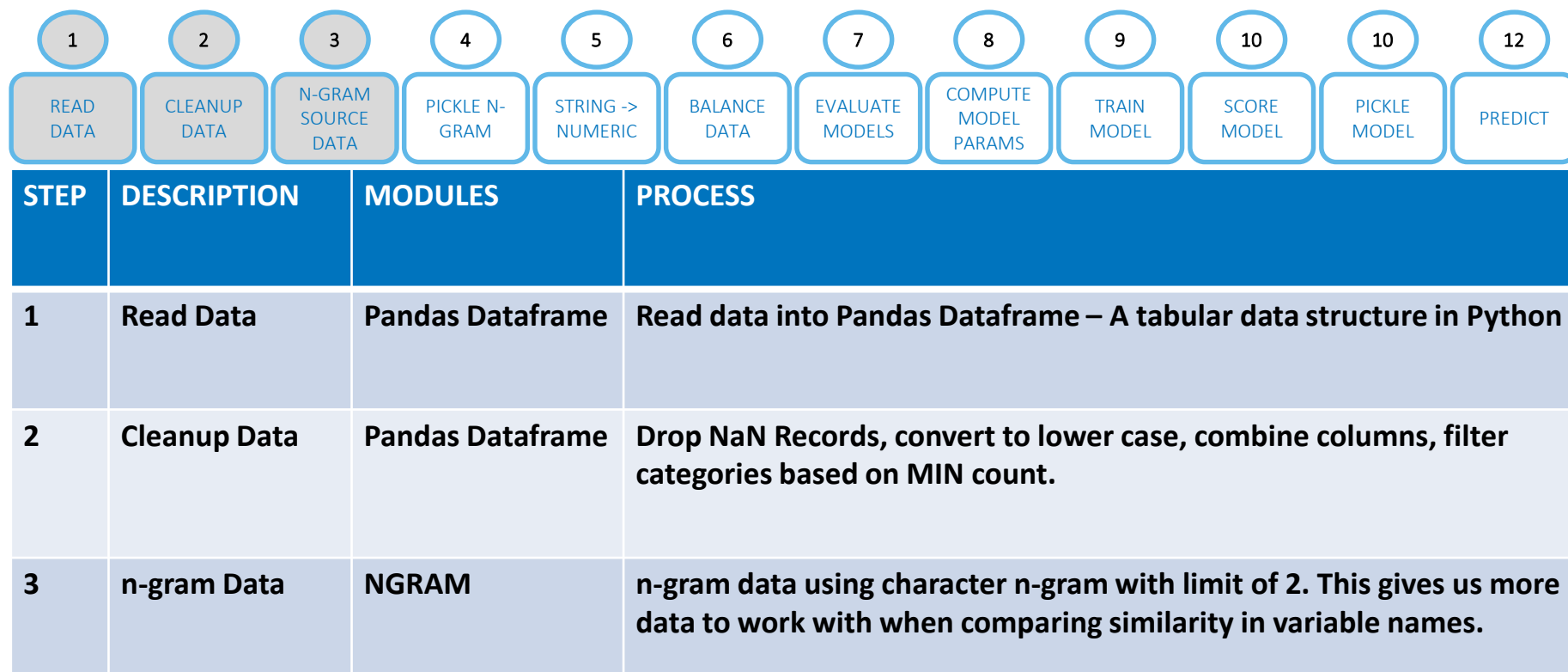
Data Mapping

Mapping Types

MAPPING TYPE	DESCRIPTION	SOURCE			DESTINATION		
TABLES	1-1 MATCH (Top Similarity Value)	DATASET			DOMAIN		
VARIABLE (AUTOMAP)	1-1 MATCH (Top Similarity Value)	DATASET	VARIABLE		DOMAIN	VARIABLE	
VARIABLE (SMART MAP)	1-3 MATCH (>0.25 Similarity)	DATASET	VARIABLE		DOMAIN	VARIABLE	
CONTROLLED TERMS	1-1 MATCH (Top Similarity Value)	DATASET	VARIABLE	VALUE	CONTROLLED TERM	VALUE	

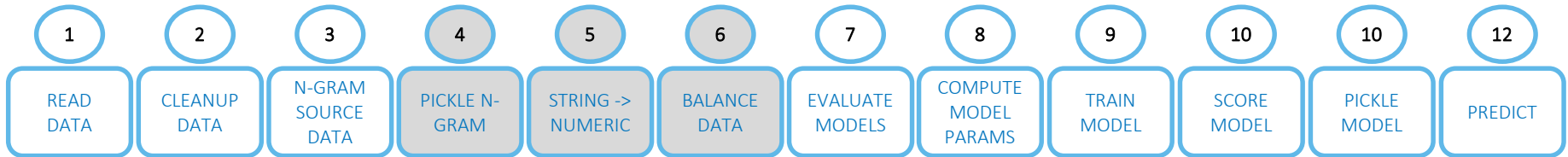
Data Mapping

Design Flow



Data Mapping

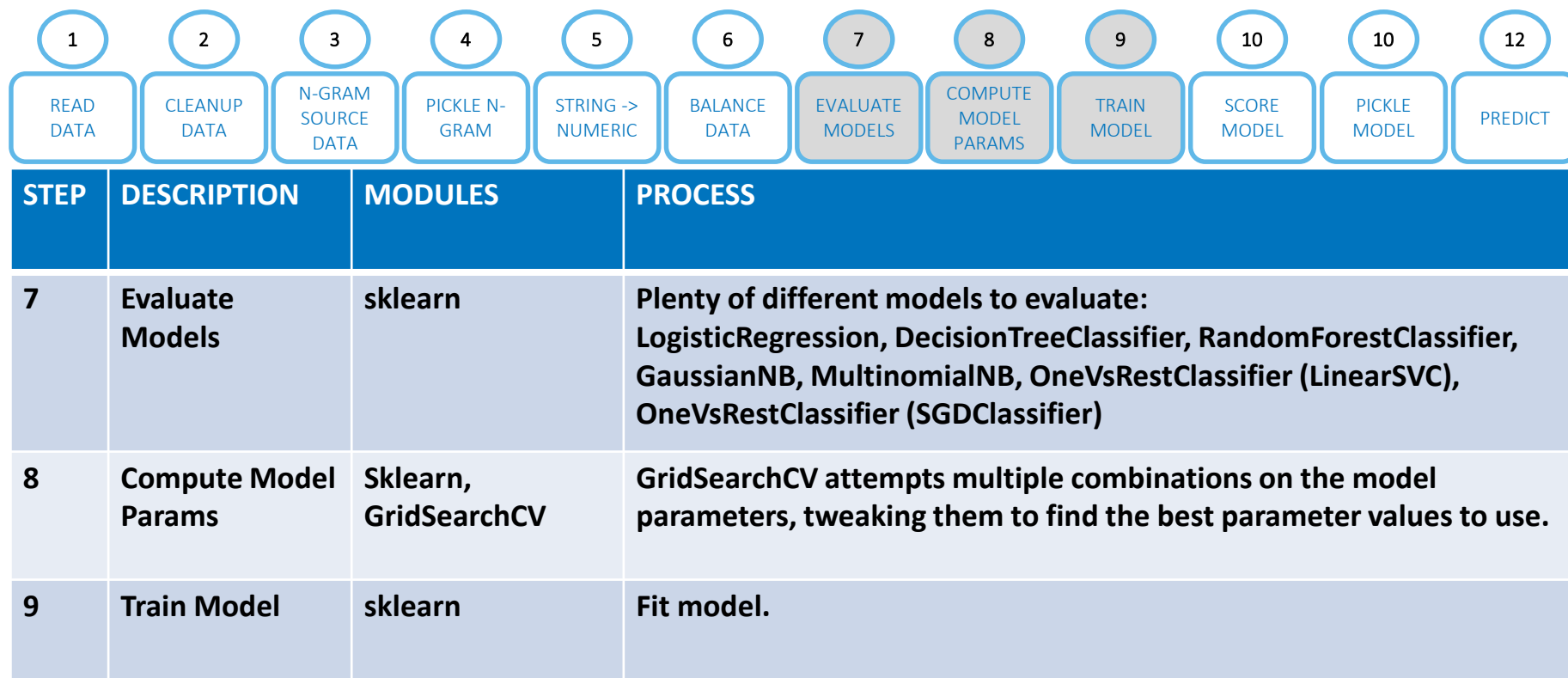
Design Flow



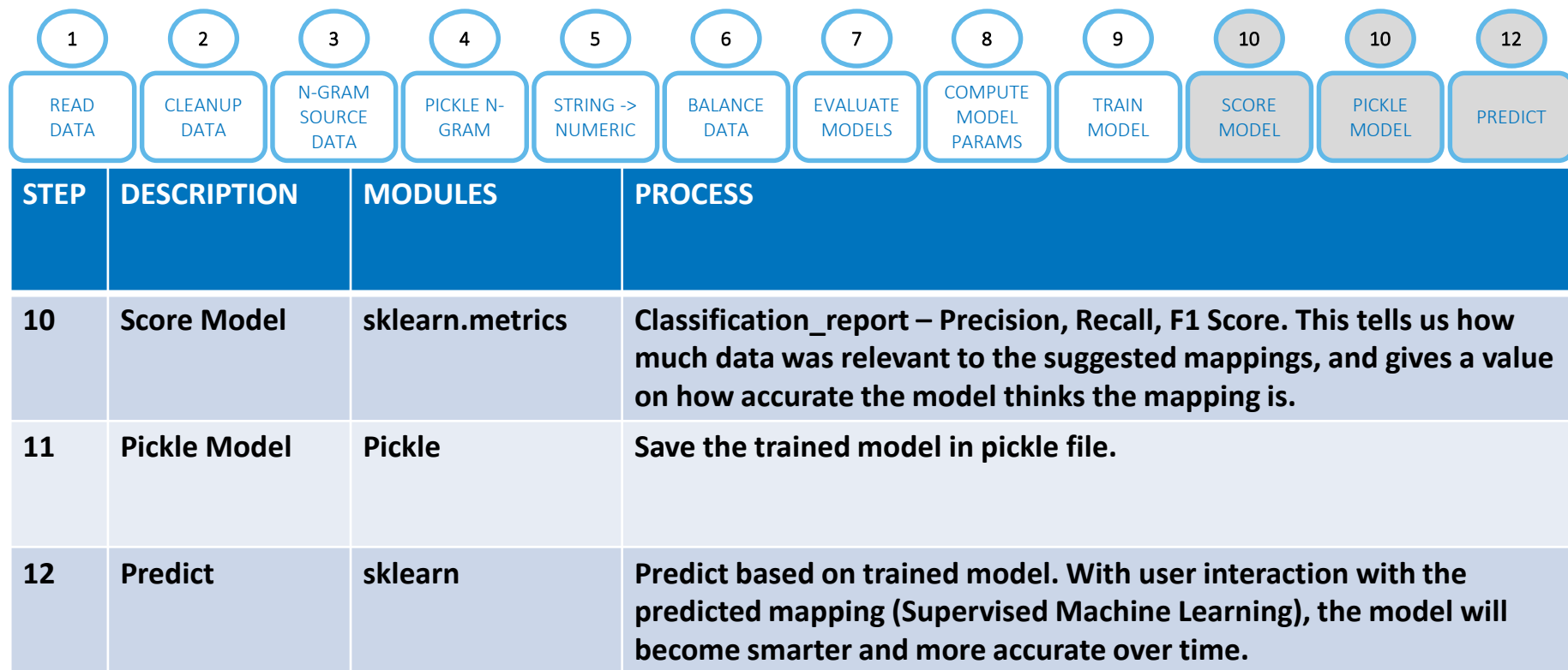
STEP	DESCRIPTION	MODULES	PROCESS
4	Pickle n-gram	Pickle	n-gram pair of (Source, Destination) value and pickle it. The pickle file is the python data object after serialization, and is an efficient way to store and use any python object data.
5	String to Numeric Value conversion	TfidfVectorizer, MinMaxScaler	Generate Dictionary with TF-IDF values and re-scale it using MinMaxScaler. For machine learning to work, the values must be numeric between 0 and 1.
6	Balance Data	SMOTETomek	Balance data to handle under-sampling and over-sampling. If the data we have is skewing our results, we want to normalize closer to 0.5.

Data Mapping

Design Flow



Data Mapping Design Flow



Data Mapping

Tables - Example

- Word_ngram created for all counts of dest_var
- Minimum count of 3 is required for Model class categories

dsname	dsname_to
adverse	AE
ae	AE
AE	AE
conmeds	CM
adverse_event	AE
cm1	CM
Concomittant_meds	CM
ecg	EG
Electrocardiogram	EG
EG	EG
Chemsitry	LB
Hematology	LB
Preg	LB



source_var	dest_var	Model_ngram	Word_ngram
adverse	AE	ad dv ve er rs se	(ae,adverse)
ae	AE	ae	(ae,ae)
AE	AE	ae	(ae,ae)
conmeds	CM	co on nm me ed ds	(cm,conmeds)
adverse_event	AE	ad dv ve er rs se e__e ev ve en nt	(ae,adverse_event)
cm1	CM	cm m1	(cm,cm1)
Concomittant_meds	CM	co on nc co om mi it tt ta an nt t__m me ed ds	(cm,concomittant_meds)
ecg	EG	ec cg	(eg,ecg)
Electrocardiogram	EG	el le ec ct tr ro oc ca ar rd di io og gr ra am	(eg,electrocardiogram)
EG	EG	eg	(eg,eg)
Chemsitry	LB	ch he em mi is st tr ry	(lb,chemistry)
Hematology	LB	he em ma at to ol lo og gy	(lb,hematology)
Preg	LB	pr re eg	(lb,preg)

Data Mapping Tables – Model Scores

Model	Definition	Score	Precision	Recall	F1-Score	Support
NB	CalibratedClassifierCV(GaussianNB())	0.772727272727	0.81	0.77	0.77	22
MNB	CalibratedClassifierCV(MultinomialNB(alpha=0.5, fit_prior=True, class_prior=None))	0.727272727273	0.75	0.73	0.72	22
LR	CalibratedClassifierCV(LogisticRegression(C=1))	0.727272727273	0.72	0.73	0.70	22
OVR_SVC	CalibratedClassifierCV(OneVsRestClassifier(LinearSVC(random_state=RANDOM_STATE)))	0.883333333333	0.89	0.88	0.88	22

Classification Report for OVR_SVC

	precision	recall	f1-score	support
ae	1.00	1.00	1.00	3
cm	1.00	0.67	0.80	3
dm	1.00	1.00	1.00	1
ds	0.00	0.00	0.00	1
eg	1.00	1.00	1.00	1
ex	1.00	1.00	1.00	1
fa	1.00	1.00	1.00	1
ho	0.00	0.00	0.00	1
ie	0.25	1.00	0.40	1
lb	1.00	1.00	1.00	3
mh	1.00	1.00	1.00	1
pe	1.00	1.00	1.00	4
vs	1.00	1.00	1.00	1
avg / total	0.88	0.86	0.85	22

Data Mapping

Tables – Model Test Results

- Test Data

```
# Define TEST Data and Expected Outcomes
```

```
X_test = np.array(['adverse2', 'conmed', 'chemistry', 'follow', 'electrocardiac', 'inclusion'])
```

```
y_test = np.array(['AE', 'CM', 'LB', 'FA', 'EG', 'IE'])
```

- Test Result

```
Predictions: ['ae' 'cm' 'lb' 'fa' 'eg' 'ie'] Expected: ['AE', 'CM', 'LB', 'FA', 'EG', 'IE']
```

```
Model_Matched_Term  Model_Similarity NGram_Matched_Term  NGram_Similarity
```

```
Search_Term
```

adverse2	ae	0.717276	ae	0.583333
chemistry	lb	0.515414	lb	1.000000
conmed	cm	0.703553	cm	1.000000
electrocardiac	eg	0.699650	eg	0.521739
follow	fa	0.683542	fa	1.000000
inclusion	ie	0.537428	ie	1.000000

- Use combination of Model + NGRAM Similarity to predict output values

Data Mapping

Variables – Data Sample

- Word_ngram created for all counts of dest_var
- Minimum count of 3 is required for Model class categories

dsname	varname	dsname_to	varname_to
adverse	SUBJID	AE	USUBJID
aedata	PATID	AE	USUBJID
adverse	startdt	AE	AESTDTC
aedata	stdt	AE	AESTDTC
adverse	relation	AE	AEREL
ae_info	relinfo	AE	AEREL
hema	ptid	LB	USUBJID
chem	ptid	LB	USUBJID
conmed	patid	CM	USUBJID
cmmeds	patient	CM	USUBJID



source_var	dest_var	model_ngram	word_ngram
adverse.subjid	ae.usubjid	ad dv ve er rs se e. .s su ub bj ji id	(ae.usubjid,adverse.subjid)
aedata.patid	ae.usubjid	ae ed da at ta a. .p pa at ti id	(ae.usubjid,aedata.patid)
adverse.startdt	ae.aestdct	ad dv ve er rs se e. .s st ta ar rt td dt	(ae.aestdct,adverse.startdt)
aedata.stdtd	ae.aestdct	ae ed da at ta a. .s st td dt	(ae.aestdct,aedata.stdtd)
adverse.relation	ae.aerel	ad dv ve er rs se e. .r re el la at ti io on	(ae.aerel,adverse.relation)
ae_info.relinfo	ae.aerel	ae e__i in nf fo o. .r re el li in nf fo	(ae.aerel,ae_info.relinfo)
hema.ptid	lb.usubjid	he em ma a. .p pt ti id	(lb.usubjid,hema.ptid)
chem.ptid	lb.usubjid	ch he em m. .p pt ti id	(lb.usubjid,chem.ptid)
conmed.patid	cm.usubjid	co on nm me ed d. .p pa at ti id	(cm.usubjid,conmed.patid)
cmmeds.patient	cm.usubjid	cm mm me ed ds s. .p pa at ti ie en nt	(cm.usubjid,cmmeds.patient)

Data Mapping

Variables - Scores

Model	Definition	Score	Precision	Recall	F1-Score	Support
NB	CalibratedClassifierCV(GaussianNB())	0.8	0.81	0.80	0.78	60
MNB	CalibratedClassifierCV(MultinomialNB(alpha=0.5, fit_prior=True, class_prior=None))	0.6666666666666667	0.75	0.67	0.67	60
LR	CalibratedClassifierCV(LogisticRegression(C=1))	0.83	0.85	0.79	0.83	60
OVR_SVC	CalibratedClassifierCV(OneVsRestClassifier(LinearSVC(random_state= RANDOM_STATE)))	0.85	0.89	0.85	0.86	60

Classification Report for OVR_SVC

	precision	recall	f1-score	support
ae.aerel	1.00	1.00	1.00	3
ae.aeser	1.00	1.00	1.00	2
ae.aestdte	1.00	1.00	1.00	3
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.dmdte	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

Data Mapping

Results - Mapping

- If Similarity >0.9, show 1 match , 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 Result

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

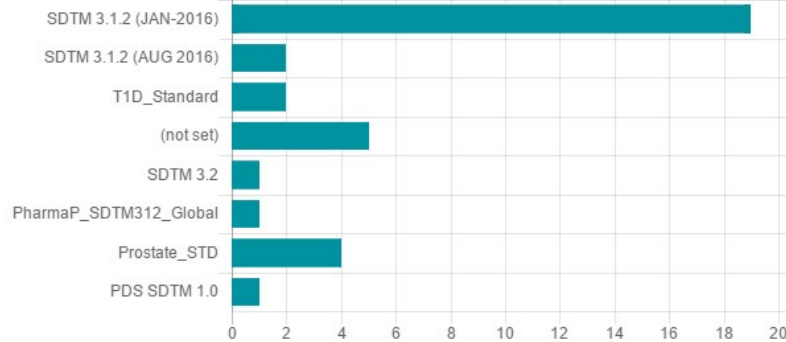
	Search_Term	Model_Matched_Term	Model_Similarity	NGram_Matched_Term	NGram_Similarity
0	adverse.ptid	ae.usubjid	0.412617	ae.usubjid	1.000000
1	hema.subjid	lb.usubjid	0.474403	lb.usubjid	1.000000
2	conmed.patid	cm.usubjid	0.699372	cm.usubjid	1.000000
3	dm.birthdate	dm.age	0.682097	dm.age	0.578947
4	dm.birthdate	dm.usubjid	0.022968	dm.age	0.444444
5	dm.birthdate	dm.subjid	0.021280	dm.brthdte	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

Data Mapping

Reporting - Standards

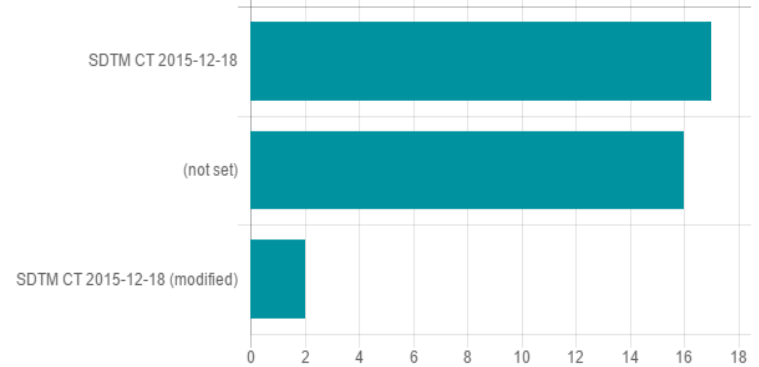
Standards by Studies Used



1

Standards By Studies used

Controlled Terminologies by Studies Used



2

CT By Studies used

Standards by Status

Active (82.4%) Inactive (11.8%) In Review (5.9%)



3

Standards by Status

Controlled Terminologies by Status

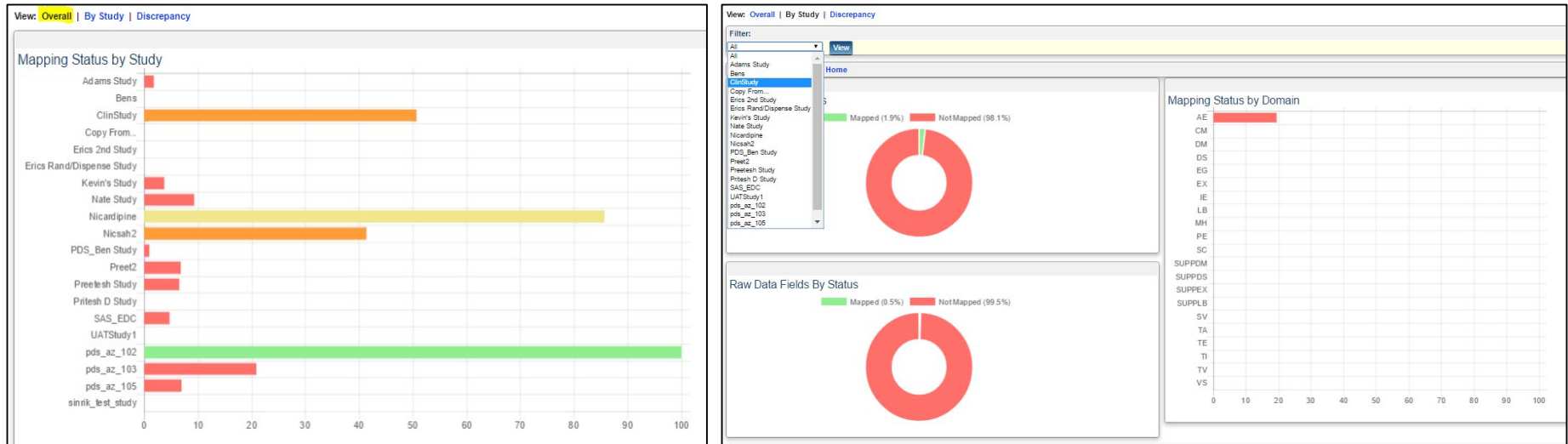
In Review (33.3%) Active (66.7%)



4

CTs By Status

Data Mapping Reporting - Global



View: **Overall** | By Study | Discrepancy

Study Mapping Difference to Libraries

Study: **Adams Study** Library:

Study: **Bens** Library:

Study: **ClinStudy** Library: **DeptA**

Domain / Variable	Library Mapping	Library Pseudo	Study Mapping	Study Pseudo
AE.AEDECOD		(none present)	ADVERSE.AEDECOD	
AE.AEENDTC		(none present)	ADVERSE.AEENDTC	
AE.AEOUT		(none present)	ADVERSE.AEOUT	
AE.AEREL		(none present)	ADVERSE.AEACTTTRT	
AE.AERELNST		(none present)	ADVERSE.AEACTTTRT	
AE.AESEQ		(none present)		%seq(AESEQ)
AE.AESER		(none present)	ADVERSE.AESEV	
AE.AESTDTC		(none present)	ADVERSE.AESEV	
AE.AETERM		(none present)	ADVERSE.AEACTTTRT	
AE.DOMAIN		(none present)		Domain="AE"
AE.STUDYID	ADVERSE.BODYSYS		ADVERSE.STUDYIDADVERSE.BODYSYS	
AE.USUBJID	ADVERSE.SUBJID		ADVERSE.SUBJID	
CM.STUDYID	CHEM.STUDYID		CHEM.STUDYIDCONMED.STUDYID	
CM.USUBJID	CHEM.SUBJID		CHEM.SUBJIDCONMED.SUBJID	



Questions?

sas.com

Thank you for your time



sas.com

Lex Jansen

Principal Solution Consultant
Health and Life Sciences

100 SAS Campus Dr
Cary, NC 27513 USA

T +1 919 531 9860

lex.jansen@sas.com

sas.com