

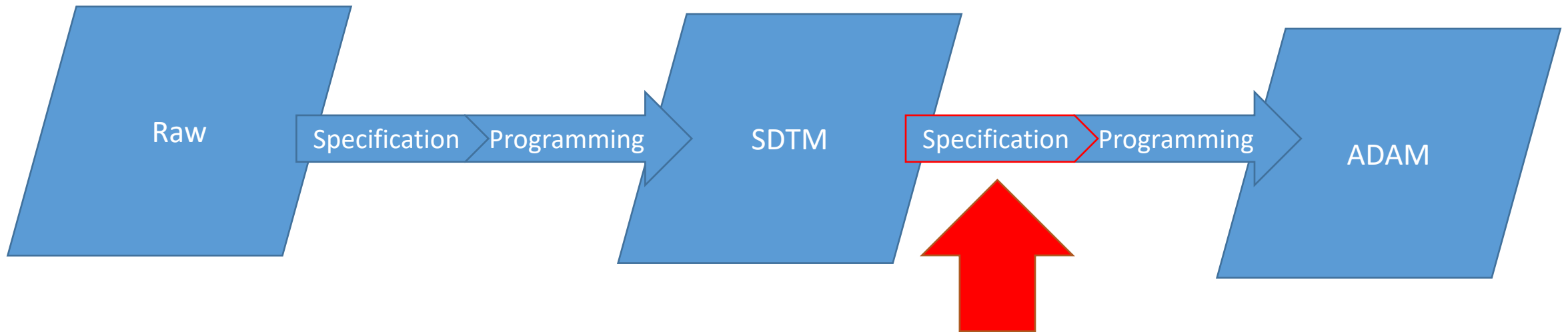
Machine Learning For CDISC Specification Definitions

ML09

Preston Burns

Brandon Welch

CDISC Clinical Data Pipeline: Our Focus



Spec'ing is a pain!

Dataset ▾	Name612 ▾	Label612 ▾	Submitdb ▾	FlagVal ▾	InAllDatasets ▾	SortOrd ▾	FullWhere ▾	Definition ▾
ADAE	AEACNN	Action Taken with Study Treatment (N)	Yes	☐	☐	119.2		Assign a numeric code for AEACN using the following: 1='DOSE NOT CHANGED'
ADAE	AEACNOTH	Other Action Taken	Yes	☐	☐	120		SDTM.AE.AEACNOTH
ADAE	AEACNDS	Caused Study Withdrawal	Yes	☐	☐	120.2		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="AEACNDS".
ADAE	INSPDTC	Interruption of Study Trt: Stopped	Yes	☐	☐	120.3		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="INSPDTM".
ADAE	INRSTDTC	Interruption of Study Trt: Restarted	Yes	☐	☐	120.4		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="INRSTDTC".
ADAE	AEREL	Causality	Yes	☐	☐	120.8		SDTM.AE.AEREL
ADAE	AREL	Analysis Causality	No	☐	☐	121		Set to "RELATED" if ADAE.AEREL is missing. Otherwise, set to AE.AEREL.

Typing out for every study is a pain.
Can we automate?

Goal: Recommend Repetitive Definitions

Programmer specification time

Programmer specification pain



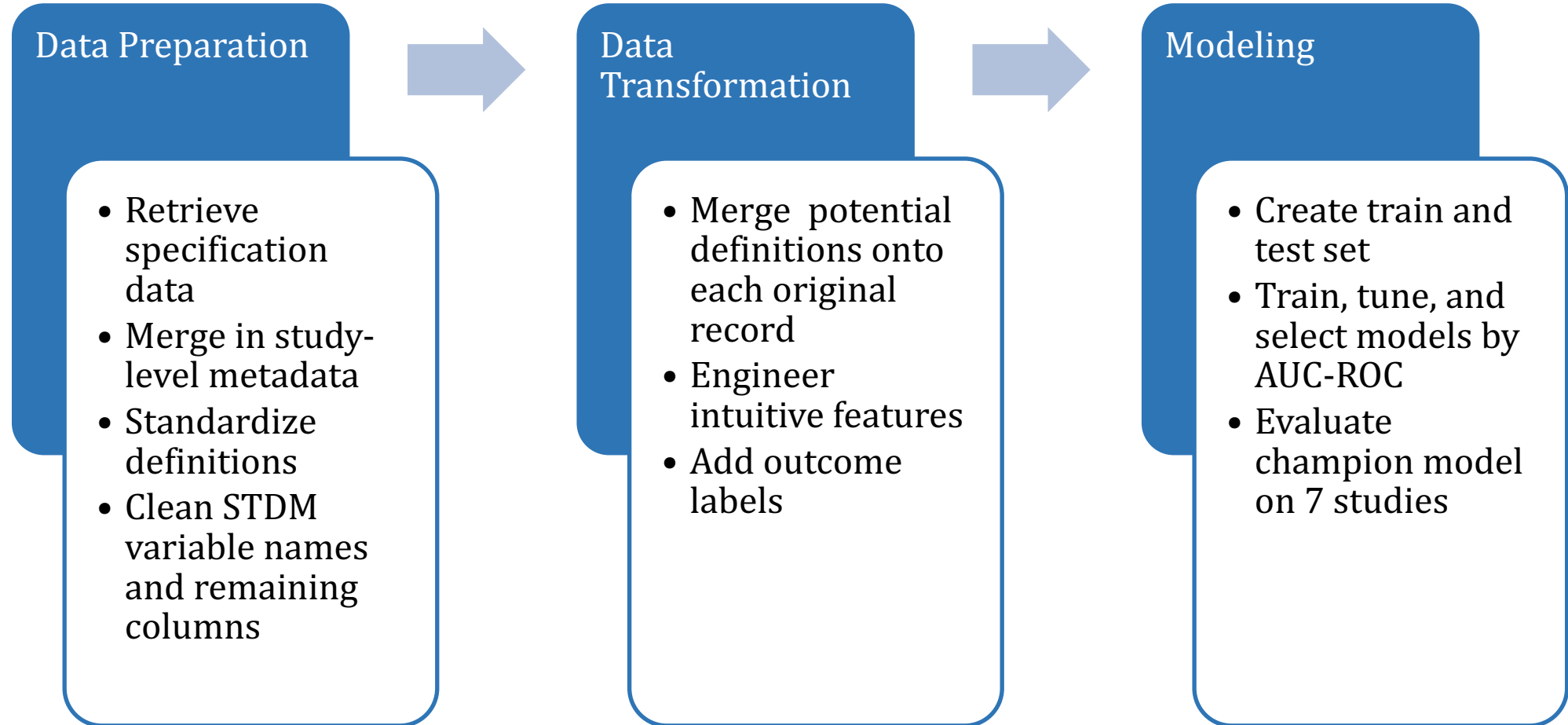
Previous Paper

- First attempt at automating definition recommendations
- Leveraged historical specification database
- Introduced new approaches to finding similar definitions
 - Cosine Similarity
- Presented at PHUSE EU Connect 2018
- Only using one input, why can't we use more?

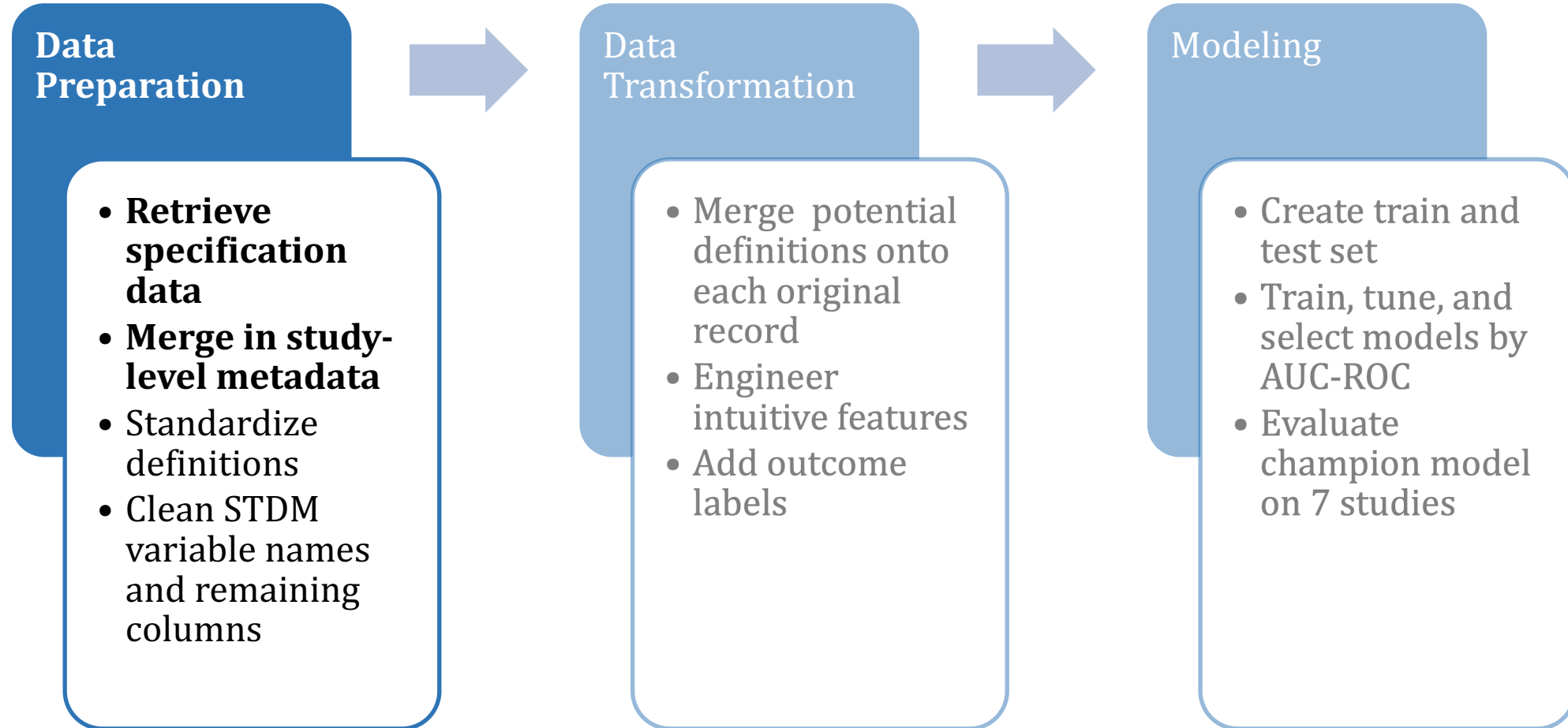
Method: Supervised Machine Learning

- Need to rank potential definitions by likelihood that a programmer would select them...
- We can train a machine learning model to do this by providing it with examples of definitions that were chosen and those that were not.

Workflow



Workflow (cont.)



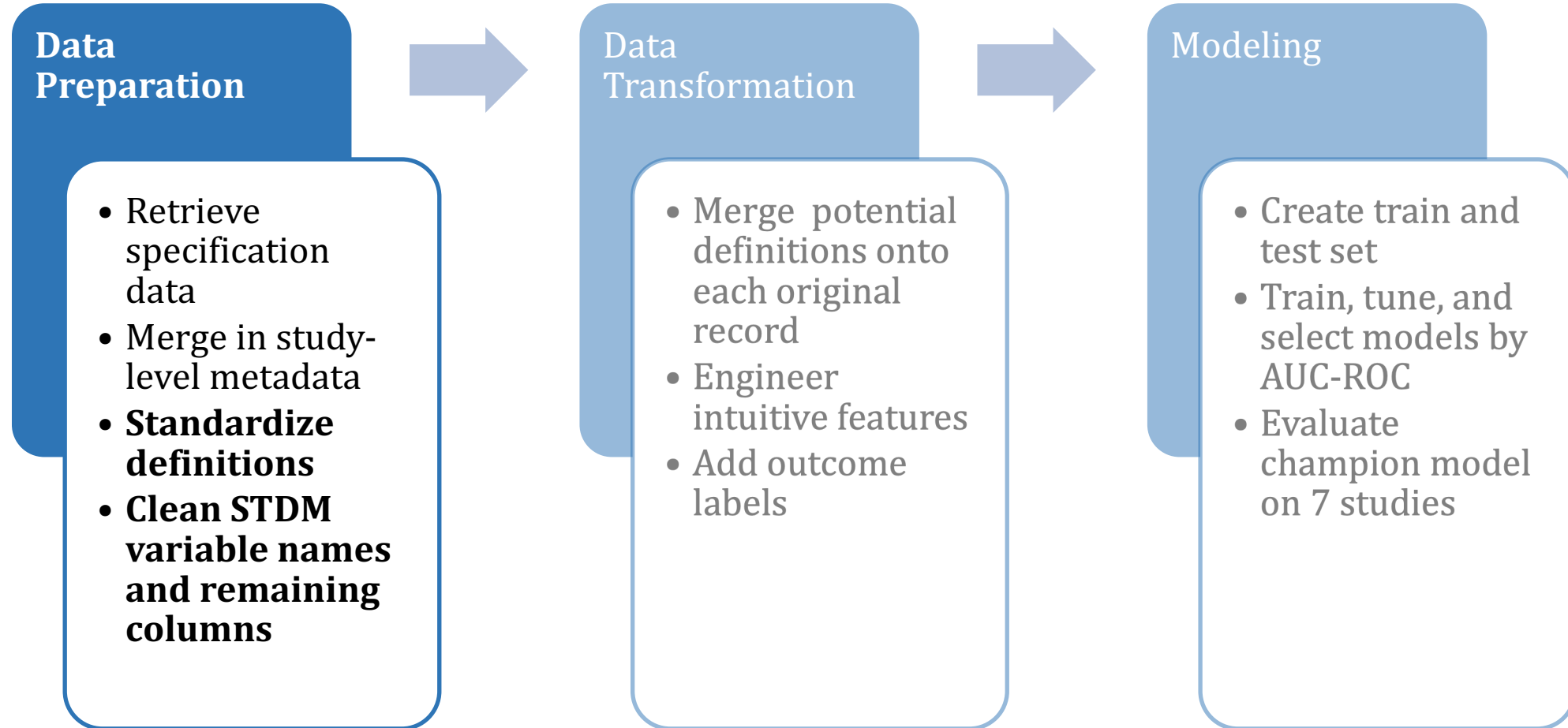
Historical Data

- All variable specifications from the last 12 years for 36 ADaM datasets
- 8,720 Variable names
- 199 Studies
- 46,425 definitions
- 21,768 definitions occurred only once
 - 47% of definitions are completely unique

Why Continue?

- There's a wide variability in the uniqueness of variables and studies
- We don't need to predict every variable
- They are not as unique as they seem. Example:
 "Y" if ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.TRTEDT + 3 missing
 otherwise
 ="Y" if TRTSDT <= ADAE.ASTDT <= ADSL.TRTEDT + 3 =missing
 otherwise

Workflow (cont.)

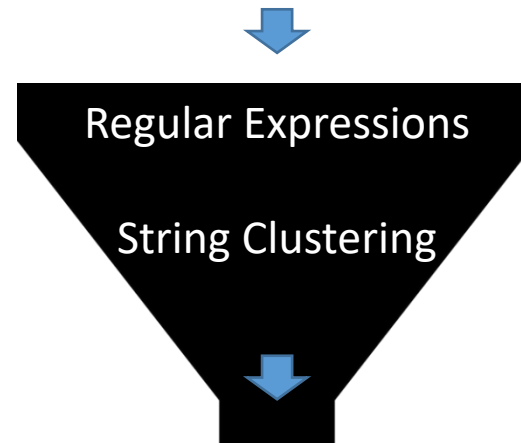


Standardizing Definitions

TRTEMFL Definitions:

"Y" if ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.TRTEDT + 3 missing
otherwise

= "Y" if TRTSDT <= ADAE.ASTDT <= ADSL.TRTEDT + 3 = missing otherwise



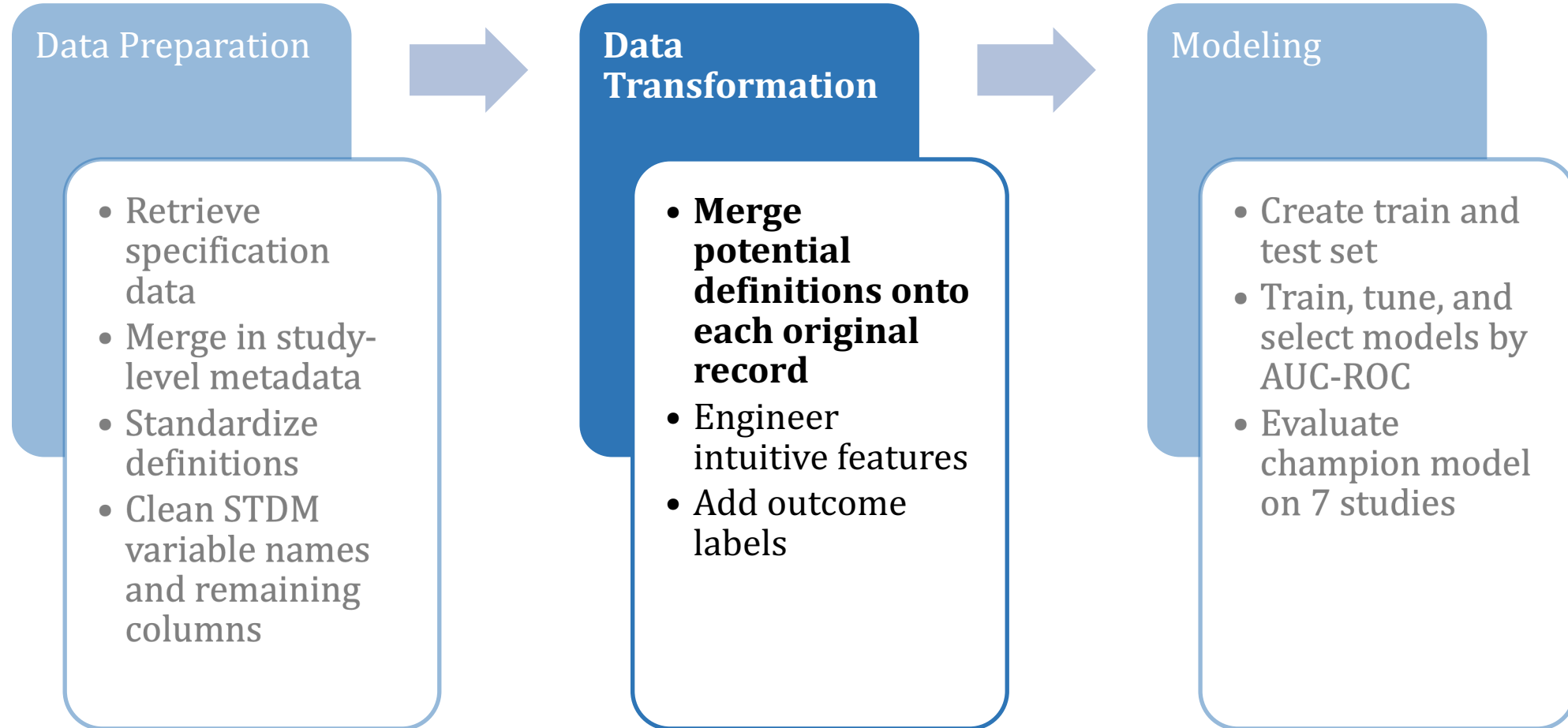
"Y" if ADSL.TRTSDT <= ADAE.ASTDT <= ADSL.TRTEDT + 3 missing otherwise

47% Unique



36% Unique

Workflow (cont.)



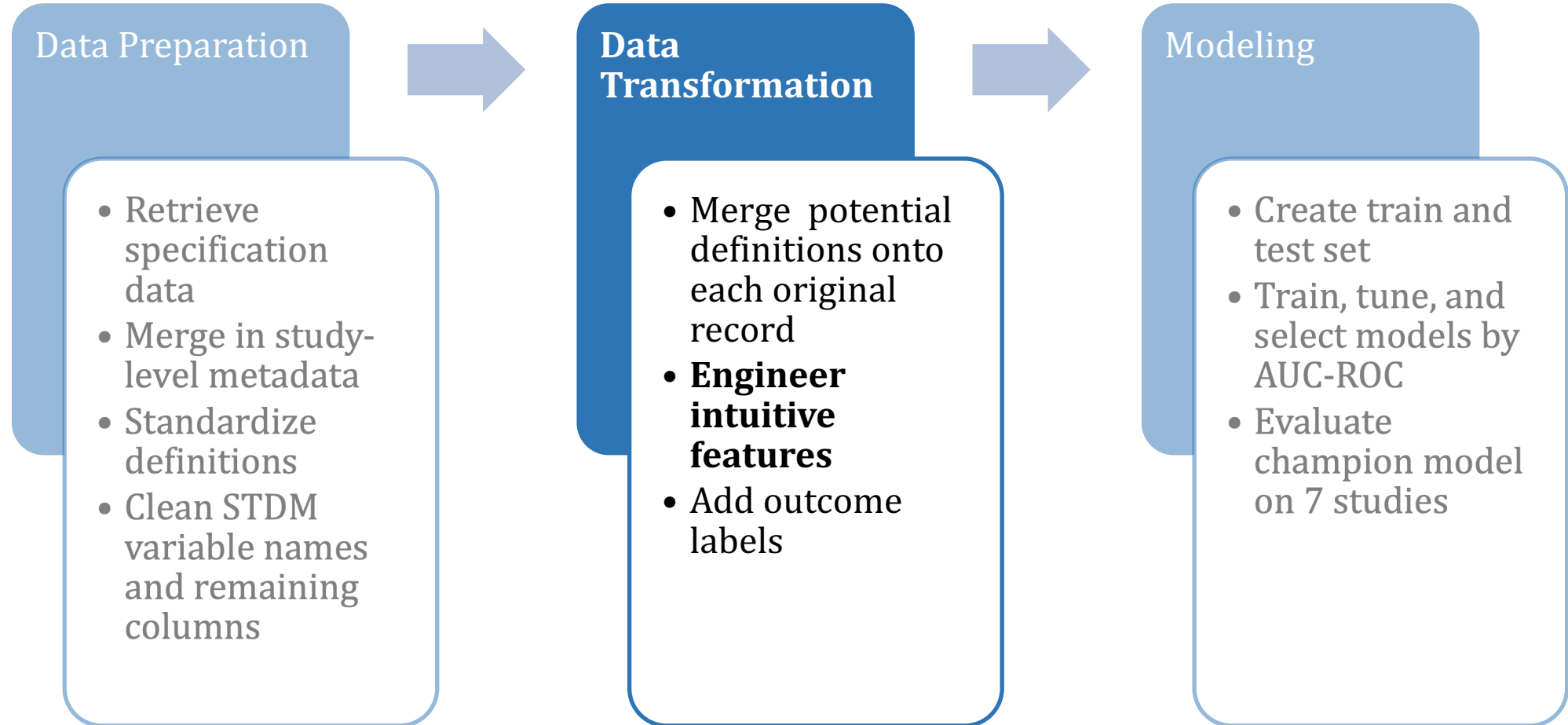
Merging Potential Definitions

Study	Variable Name	Historical Definition
A	AGEU	DM.AGEU
B	AGEU	= 'years'
C	AGEU	DM.AGEU



Study	Variable Name	Historical Definition	Potential Definition	Potential Definition Study
A	AGEU	DM.AGEU	DM.AGEU	A
A	AGEU	DM.AGEU	= 'years'	B
A	AGEU	DM.AGEU	DM.AGEU	C
B	AGEU	= 'years'	DM.AGEU	A
B	AGEU	= 'years'	= 'years'	B
B	AGEU	= 'years'	DM.AGEU	C
⋮				

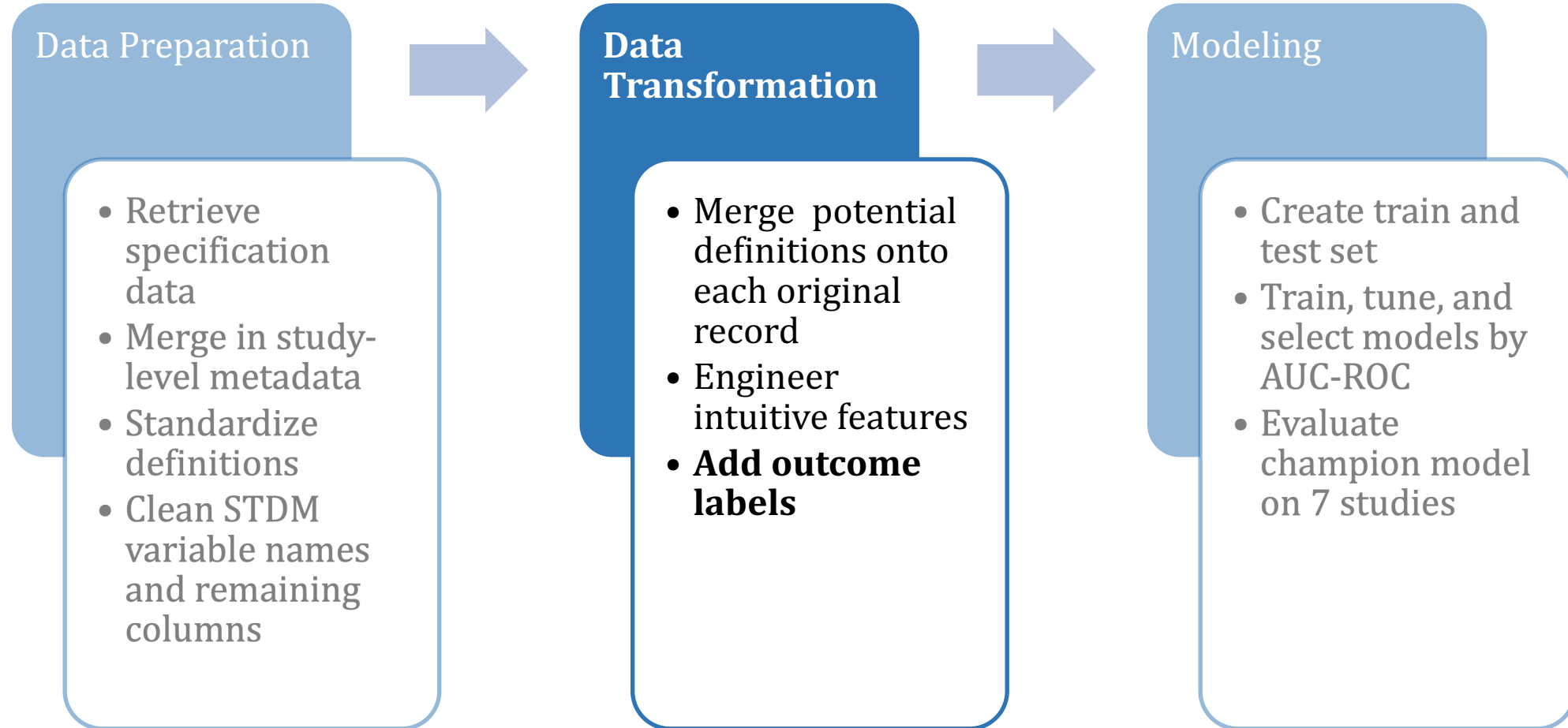
Workflow (cont.)



Feature Engineering

Feature	Python Code
Popularity	<code>data.groupby(['reco_varname','reco_definition'])['reco_definition'].transform('count')</code>
Same Sponsor	<code>np.where(data['seed_sponsor'] == data['reco_sponsor'], 1, 0)</code>
Same Therapeutic Area	<code>np.where(data['seed_thera_area'] == data['reco_therea_area'], 1, 0)</code>
Same Indication	<code>np.where(data['seed_indication'] == data['reco_indication'], 1, 0)</code>
Contains Variable Name in Definition	<code>[x[0].lower() in x[1] for x in zip(data['seed_varname'], data['reco_definition'])]</code>
Sponsor	Transformed into dummy coded columns by <code>pd.get_dummies(data)</code>

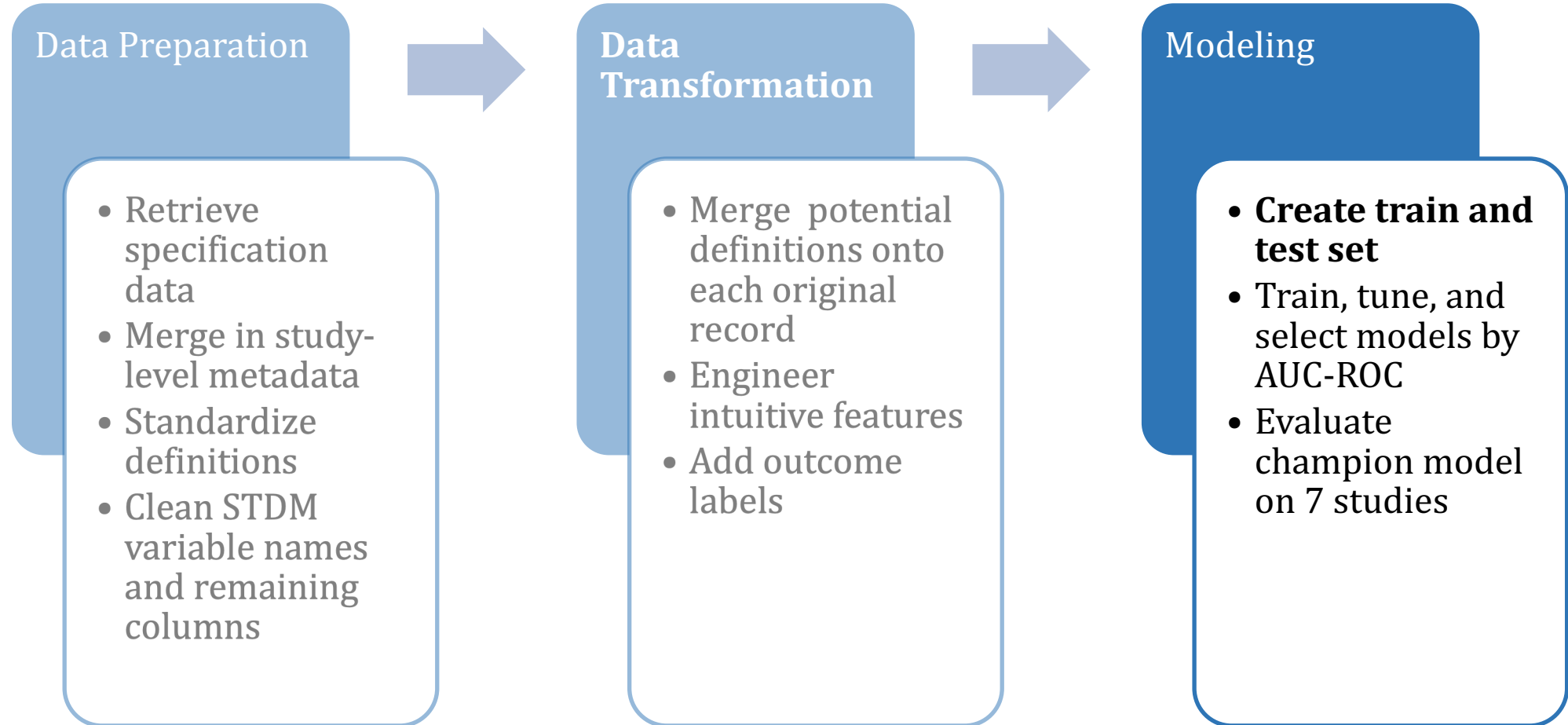
Workflow (cont.)



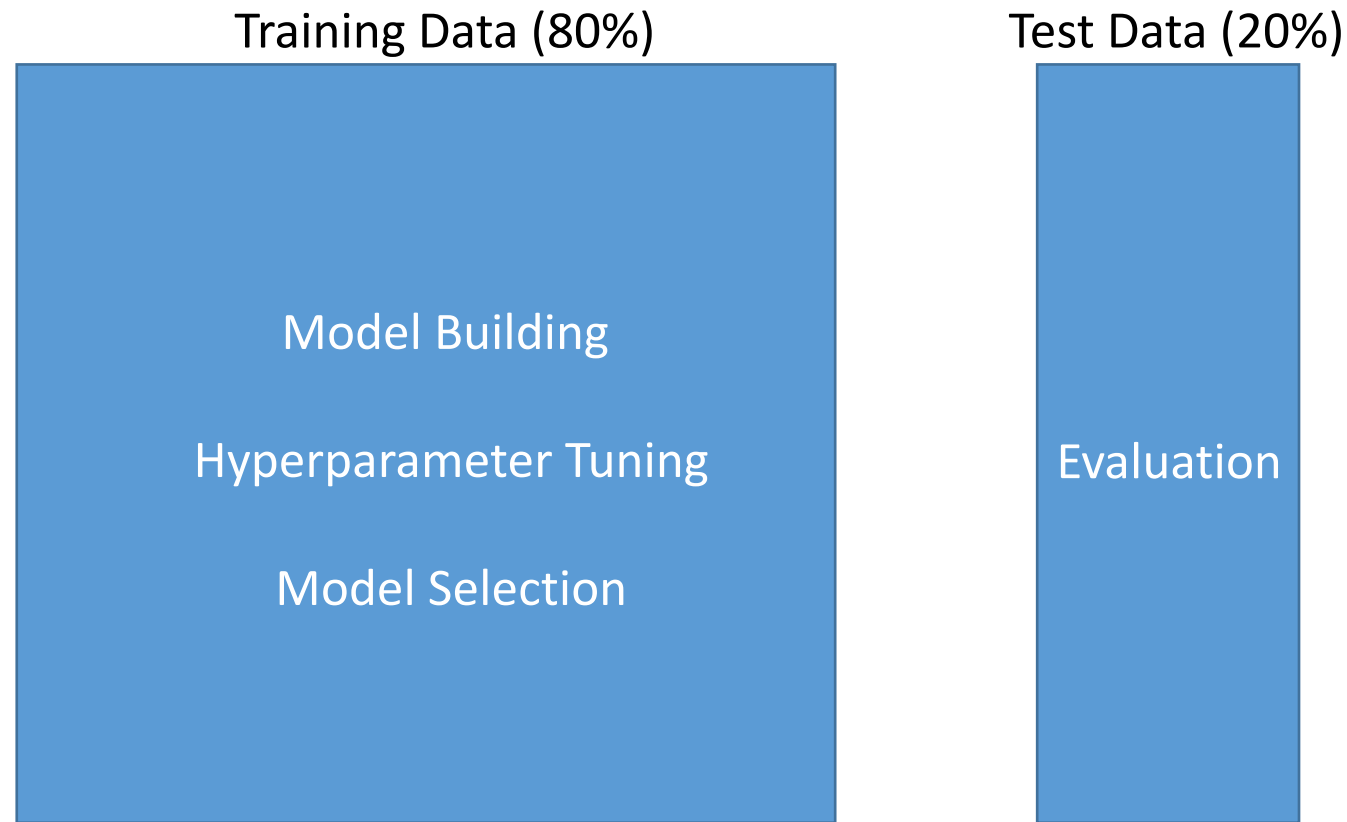
Add Outcome Variable

Study	Variable Name	Historical Definition	Potential Definition	Potential Definition Study	Correct Definition
A	AGEU	DM.AGEU	DM.AGEU	A	4
A	AGEU	DM.AGEU	= 'years'	B	0
A	AGEU	DM.AGEU	DM.AGEU	C	1
B	AGEU	= 'years'	DM.AGEU	A	0
B	AGEU	= 'years'	= 'years'	B	1
B	AGEU	= 'years'	DM.AGEU	C	0
⋮					

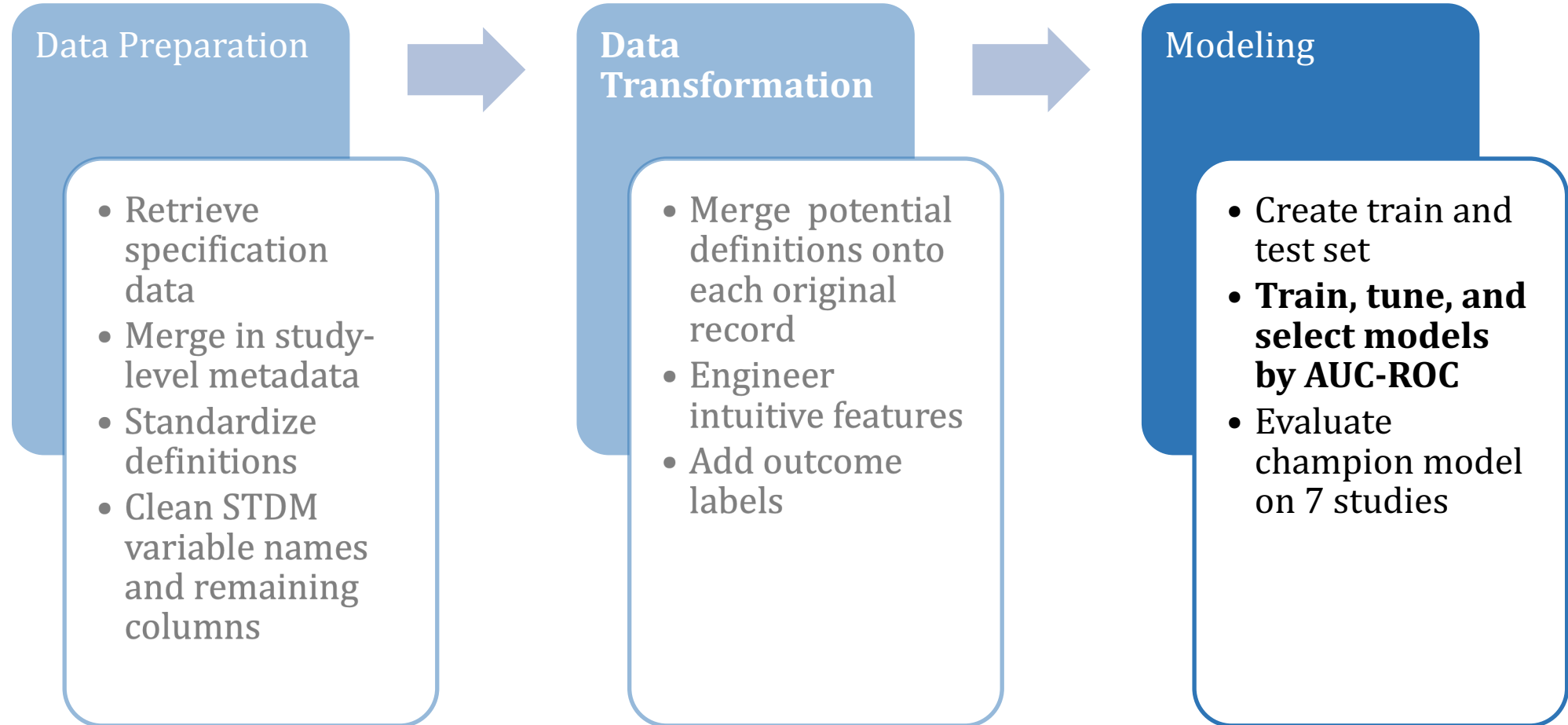
Workflow (cont.)



Train Test Split



Workflow (cont.)

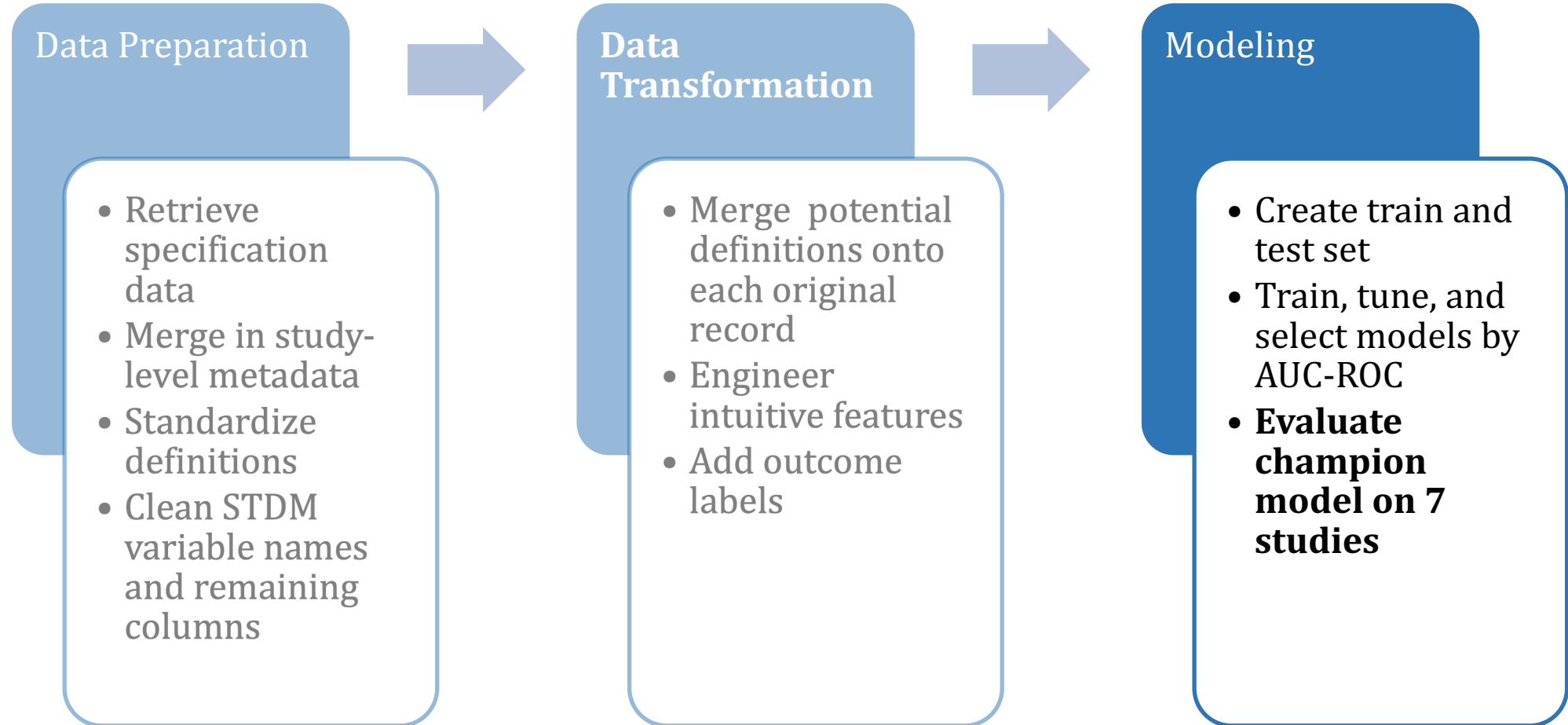


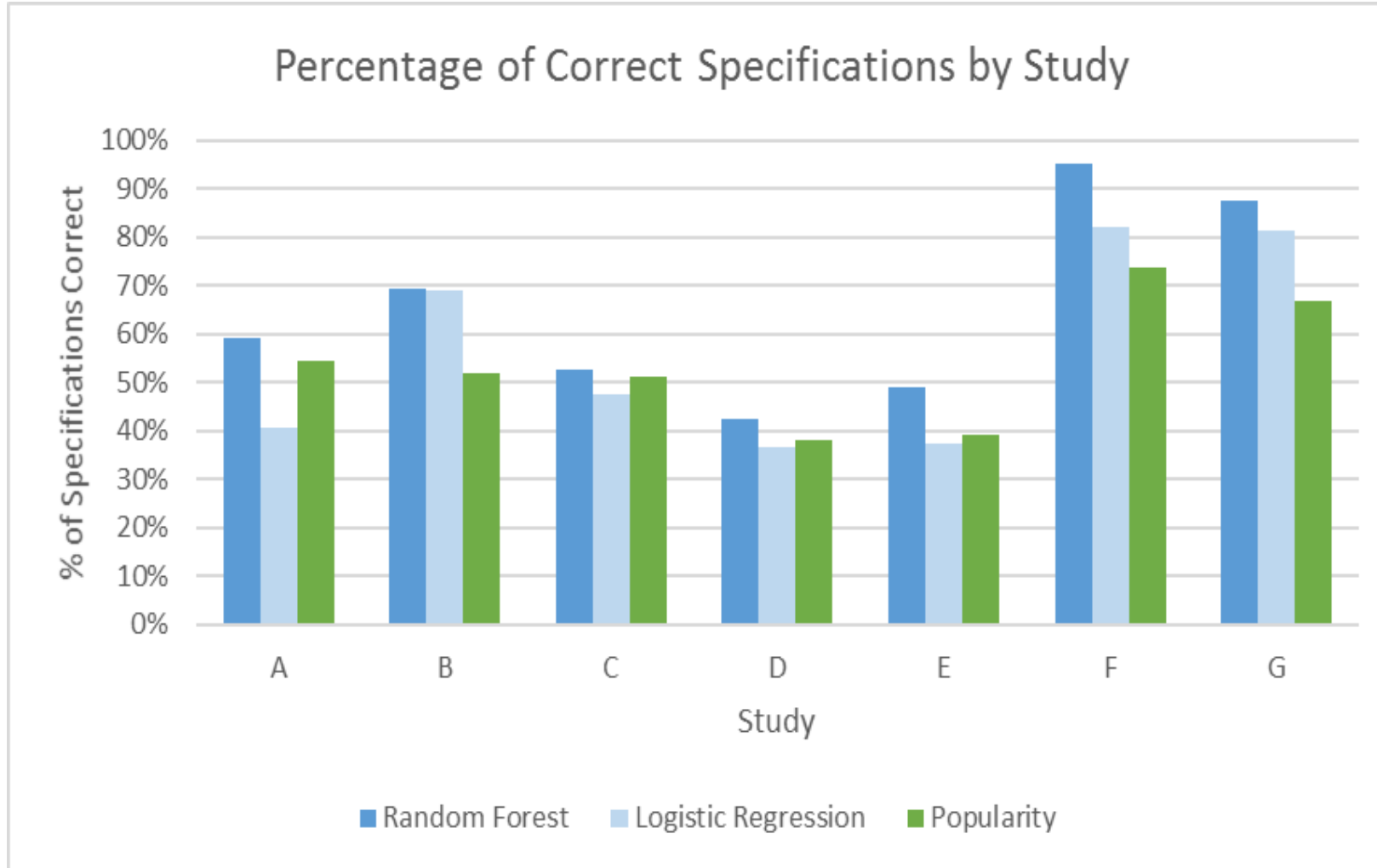
Model Performance

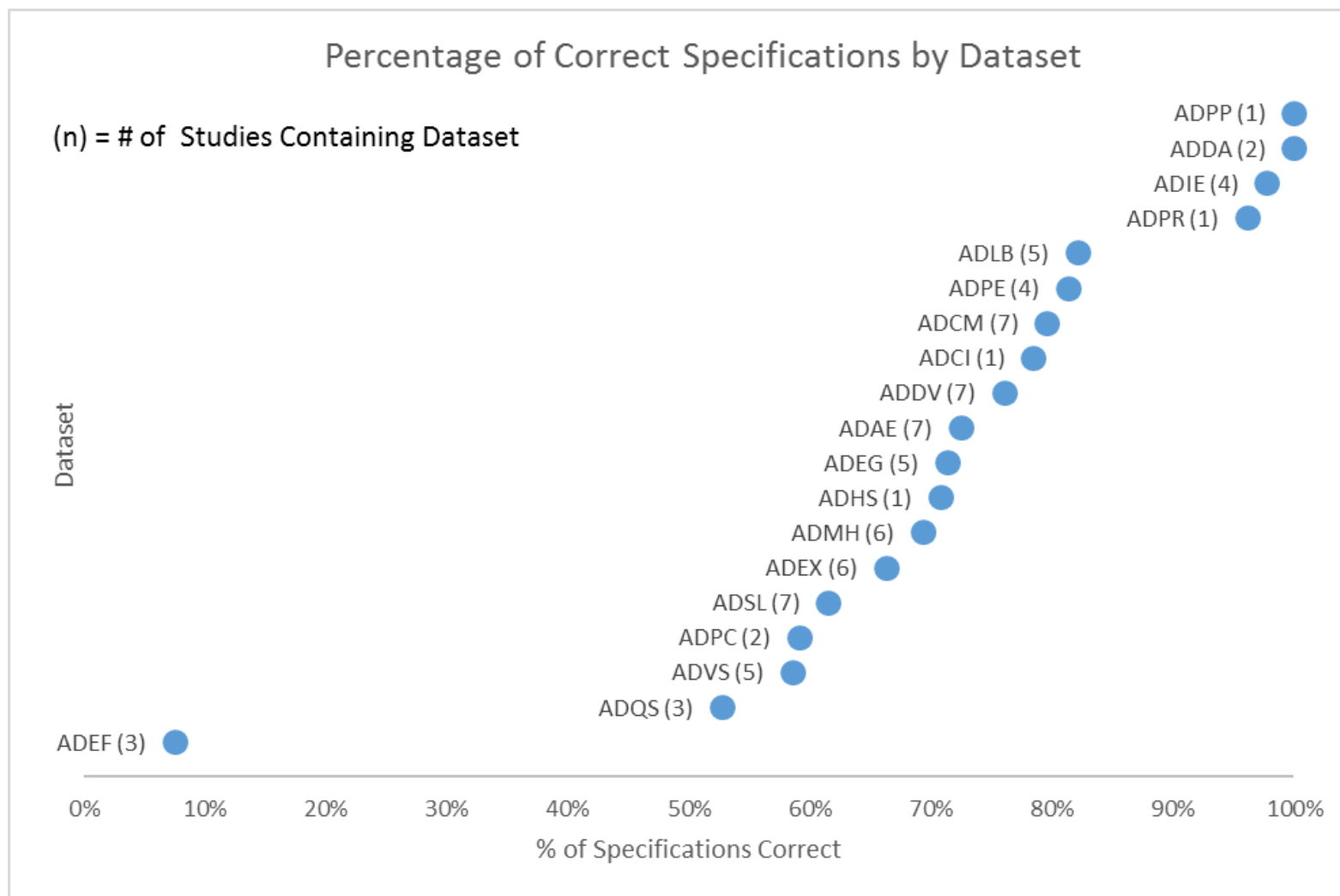
Evaluated on test dataset

	Popularity	Logistic Regression (Baseline)	Random Forest (TPOT)
AUC - ROC	.865	.893	.978

Workflow (cont.)







Implementation

Dataset ▾	Name612 ▾	Label612 ▾	Submitdb ▾	FlagVal ▾	InAllDatasets ▾	SortOrd ▾	FullWhere ▾	Definition ▾
ADAE	AEACNN	Action Taken with Study Treatment (N)	Yes	☐	☐	119.2		Assign a numeric code for AEACN using the following: 1='DOSE NOT CHANGED'
ADAE	AEACNOTH	Other Action Taken	Yes	☐	☐	120		SDTM.AE.AEACNOTH
ADAE	AEACNDS	Caused Study Withdrawal	Yes	☐	☐	120.2		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="AEACNDS".
ADAE	INSPDTC	Interruption of Study Trt: Stopped	Yes	☐	☐	120.3		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="INSPDTM".
ADAE	INRSTDTC	Interruption of Study Trt: Restarted	Yes	☐	☐	120.4		SDTM.SUPPAE.QVAL where SDTM.SUPPAE.QNAM="INRSTDTC".
ADAE	AEREL	Causality	Yes	☐	☐	120.8		SDTM.AE.AEREL
ADAE	AREL	Analysis Causality	No	☐	☐	121		

1 st Choice
2 nd Choice
3 rd Choice

Conclusion

- Correctly recommended more than 40% of definitions on every evaluation study
- Almost perfectly (95%) predicted one study from a previous sponsor and indication
- Performed well on repetitive definitions
- Flexible, multi-step process so plenty of room for improvement
- Limited by use of historical definitions



Thank You!

Preston Burns
Data Scientist
Preston_Burns@rhoworld.com

Brandon Welch
Principal Statistical Programmer
Brandon_Welch@rhoworld.com