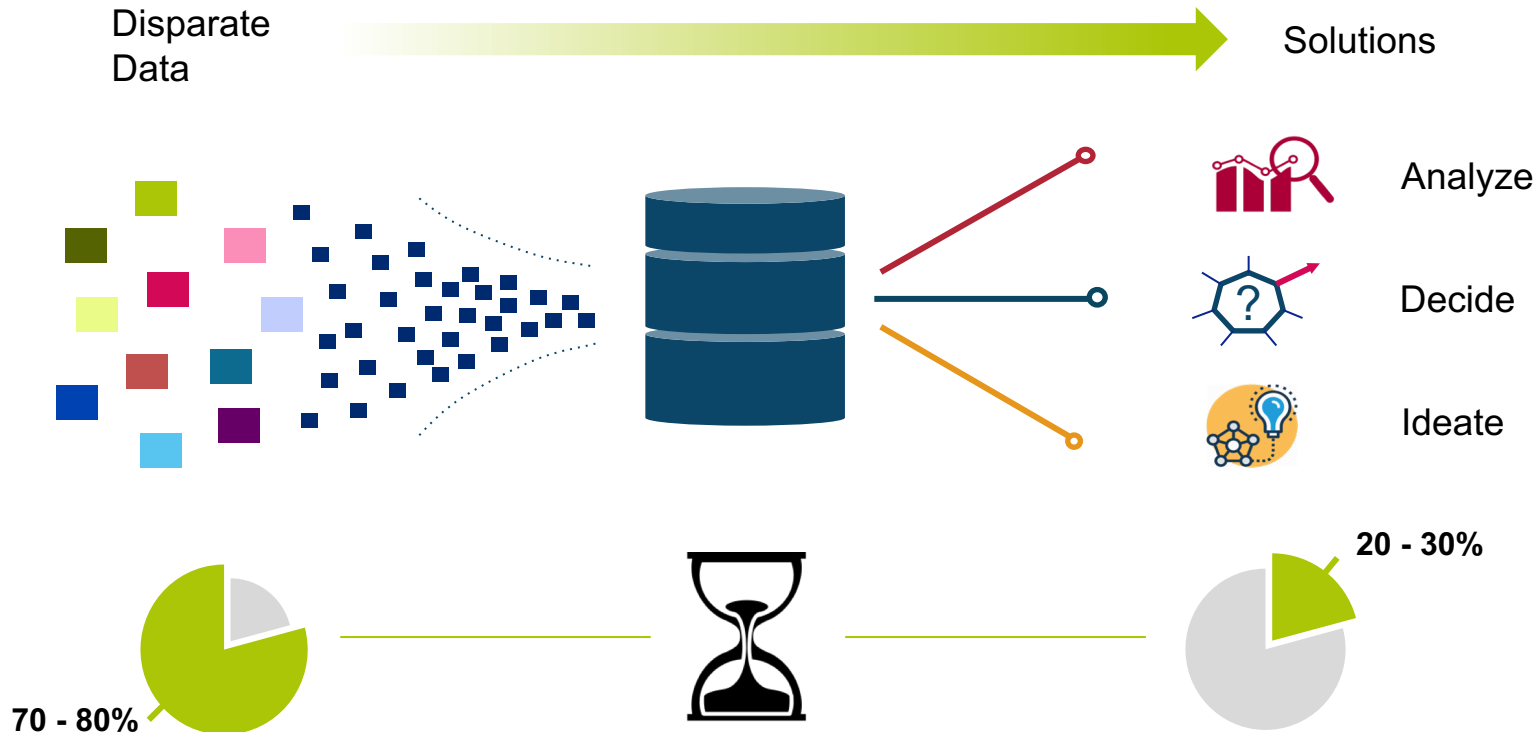


How Machine Learning Can Automate Data Standardization in Clinical Trials

*Fanyi Zhang, Medidata Solutions, Inc.
November, 2018 at Phuse, Frankfurt*

The **Big Data** Revolution in Healthcare and Clinical Research?



Standardizing clinical data is mandatory by FDA

“In 2014, FDA mandated that data collection in clinical trials adhere to Study Data Tabulation Model (SDTM) developed by Clinical Data Interchange Standards Consortium (CDISC)”

eCRF Example*: Adverse Event

Event Experienced [lgAE_UseCase1_YN]

1. Were any adverse events experienced? [Any AEs?]

2. What is the sex of the subject? [Sex]

2.1* AE Group ID [AE Group ID]

2.2* What is the adverse event term? [Adverse Event]

2.3* What is the ethnicity of the subject? [Ethnicity]

AE Number Start Date

3.

eCRF Example*: Demographics 1

Birth Year Req (2012-2014) Birth Month NReq

eCRF Example*: Demographics 2

1.* Birth Date [Birth Date]	[ETHNIC_C17459_1] [ETHNIC_C17459_2] [ETHNIC_C17459_3] [ETHNIC_C17459_4] [ETHNIC_C17459_5] [ETHNIC_C17459_6] [ETHNIC_C17459_7] [ETHNIC_C17459_8] [ETHNIC_C17459_9] [ETHNIC_C17459_10] [ETHNIC_C17459_11] [ETHNIC_C17459_12] [ETHNIC_C17459_13] [ETHNIC_C17459_14] [ETHNIC_C17459_15] [ETHNIC_C17459_16] [ETHNIC_C17459_17] [ETHNIC_C17459_18] [ETHNIC_C17459_19] [ETHNIC_C17459_20] [ETHNIC_C17459_21] [ETHNIC_C17459_22] [ETHNIC_C17459_23] [ETHNIC_C17459_24] [ETHNIC_C17459_25] [ETHNIC_C17459_26] [ETHNIC_C17459_27] [ETHNIC_C17459_28] [ETHNIC_C17459_29] [ETHNIC_C17459_30] [ETHNIC_C17459_31] [ETHNIC_C17459_32] [ETHNIC_C17459_33] [ETHNIC_C17459_34] [ETHNIC_C17459_35] [ETHNIC_C17459_36] [ETHNIC_C17459_37] [ETHNIC_C17459_38] [ETHNIC_C17459_39] [ETHNIC_C17459_40] [ETHNIC_C17459_41] [ETHNIC_C17459_42] [ETHNIC_C17459_43] [ETHNIC_C17459_44] [ETHNIC_C17459_45] [ETHNIC_C17459_46] [ETHNIC_C17459_47] [ETHNIC_C17459_48] [ETHNIC_C17459_49] [ETHNIC_C17459_50] [ETHNIC_C17459_51] [ETHNIC_C17459_52] [ETHNIC_C17459_53] [ETHNIC_C17459_54] [ETHNIC_C17459_55] [ETHNIC_C17459_56] [ETHNIC_C17459_57] [ETHNIC_C17459_58] [ETHNIC_C17459_59] [ETHNIC_C17459_60] [ETHNIC_C17459_61] [ETHNIC_C17459_62] [ETHNIC_C17459_63] [ETHNIC_C17459_64] [ETHNIC_C17459_65] [ETHNIC_C17459_66] [ETHNIC_C17459_67] [ETHNIC_C17459_68] [ETHNIC_C17459_69] [ETHNIC_C17459_70] [ETHNIC_C17459_71] [ETHNIC_C17459_72] [ETHNIC_C17459_73] [ETHNIC_C17459_74] [ETHNIC_C17459_75] [ETHNIC_C17459_76] [ETHNIC_C17459_77] [ETHNIC_C17459_78] [ETHNIC_C17459_79] [ETHNIC_C17459_80] [ETHNIC_C17459_81] [ETHNIC_C17459_82] [ETHNIC_C17459_83] [ETHNIC_C17459_84] [ETHNIC_C17459_85] [ETHNIC_C17459_86] [ETHNIC_C17459_87] [ETHNIC_C17459_88] [ETHNIC_C17459_89] [ETHNIC_C17459_90] [ETHNIC_C17459_91] [ETHNIC_C17459_92] [ETHNIC_C17459_93] [ETHNIC_C17459_94] [ETHNIC_C17459_95] [ETHNIC_C17459_96] [ETHNIC_C17459_97] [ETHNIC_C17459_98] [ETHNIC_C17459_99] [ETHNIC_C17459_100]	[BIRTHDAY] [BIRTHYR] Birth Year Req (2012-2014) Birth Month NReq	[BIRTHMO]
2.* Sex [Sex]	[SEX] [A:F] [A:M]	[SEX] [A:F] [A:M]	
3.* Ethnicity [Ethnicity]	[ETHNIC] [A:HISPANIC OR LATINO] [A:NOT HISPANIC OR LATINO] [A:NOT REPORTED] [A:UNKNOWN]	[ETHNIC] [A:HISPANIC OR LATINO] [A:NOT HISPANIC OR LATINO] [A:NOT REPORTED] [A:UNKNOWN]	Hispanic or Latino Not Hispanic or Latino Not reported Unknown
4.* Race [Race]	[RACE] [A:AMERICAN INDIAN OR ALASKA NATIVE] [A:ASIAN] [A:BLACK OR AFRICAN AMERICAN] [A:NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER] [A:WHITE]	[RACE] [A:AMERICAN INDIAN OR ALASKA NATIVE] [A:ASIAN] [A:BLACK OR AFRICAN AMERICAN] [A:NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER] [A:WHITE]	American Indian or Alaska Native Asian Black or African American Native Hawaiian or Other Pacific Islander White

Key: [*] = Item is required

*Screenshots from Medidata Rave

MEDS: One of the Industry's Largest Give to Get Clinical and Operational Data

>14K

studies

99% in study count*

>4M

trial subjects

73% in study count*

57K

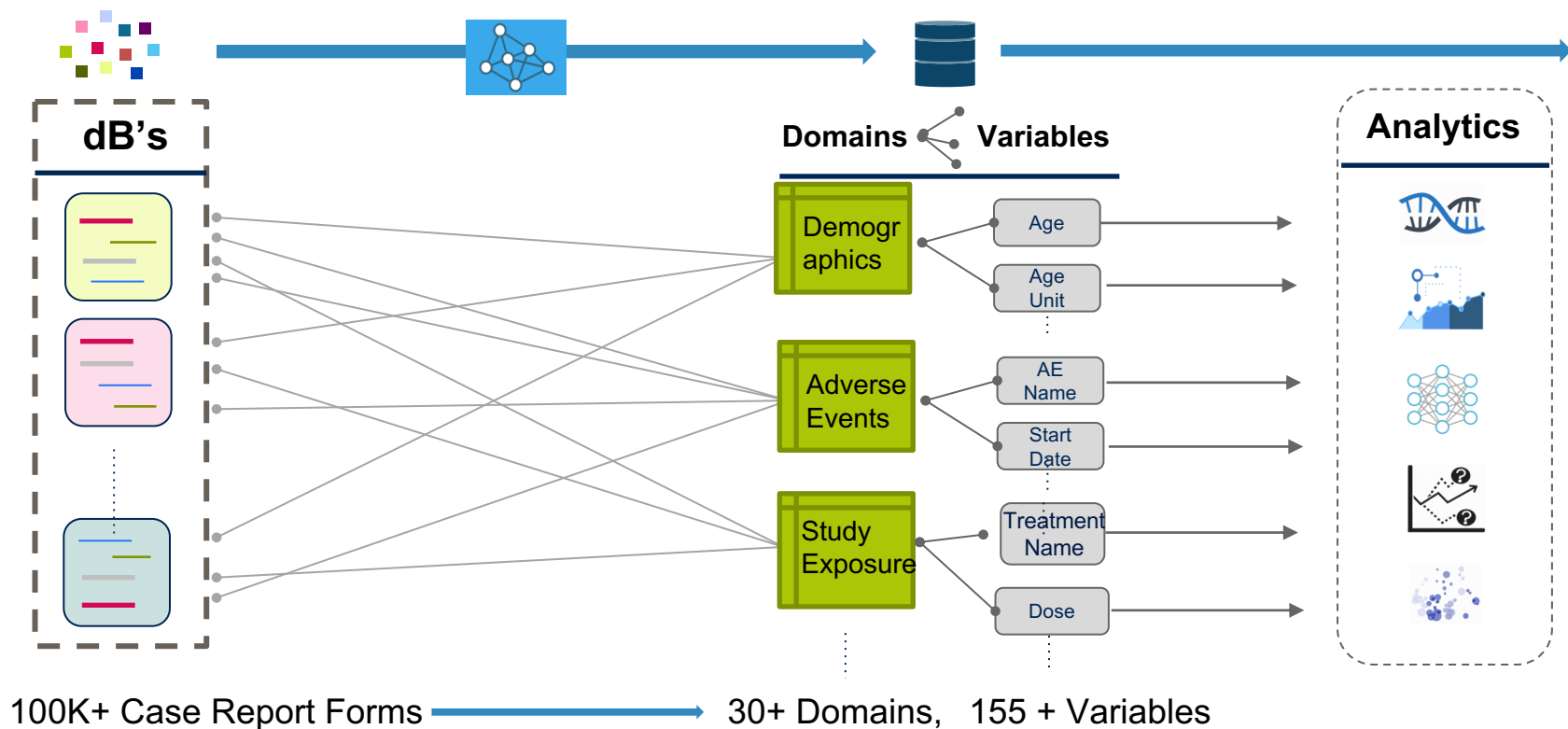
sponsor/site relationships

90% in number of study sites*

***Growth over 3 year period**

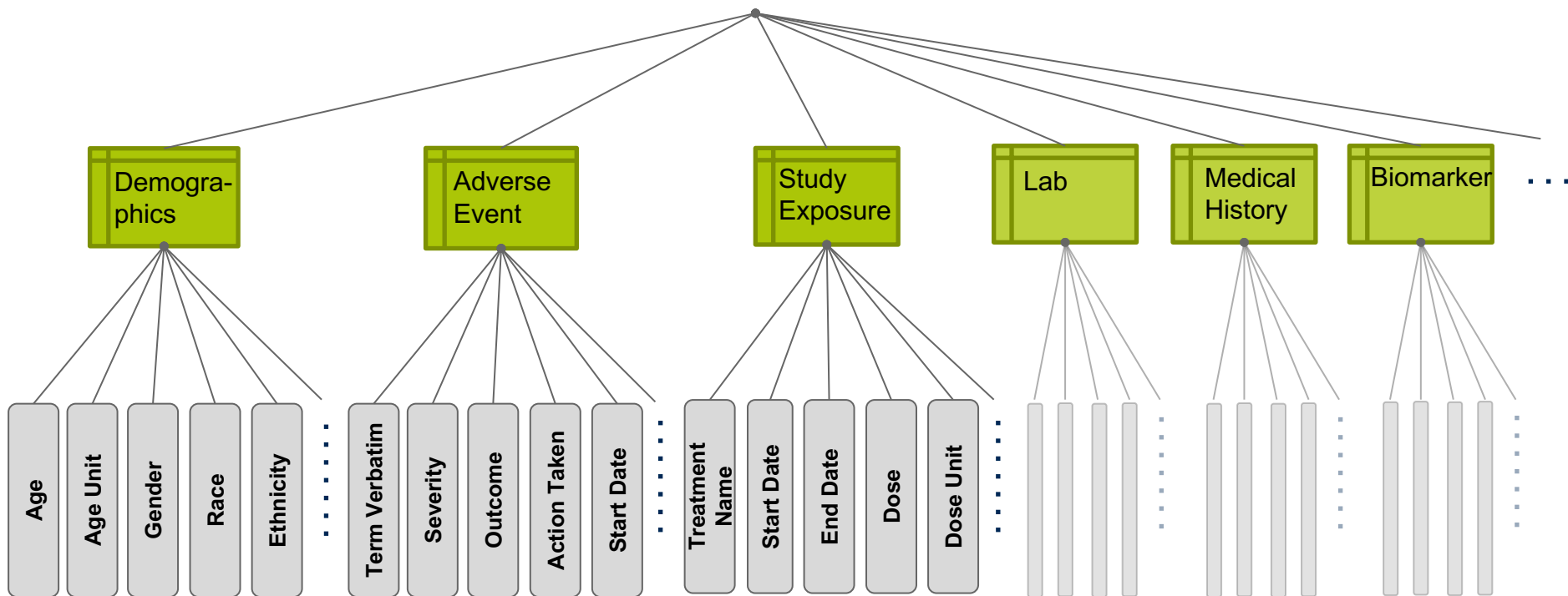
Standardize Data Across Multiple Clinical Trials

From Rule-Based Engines to Auto-SDTM



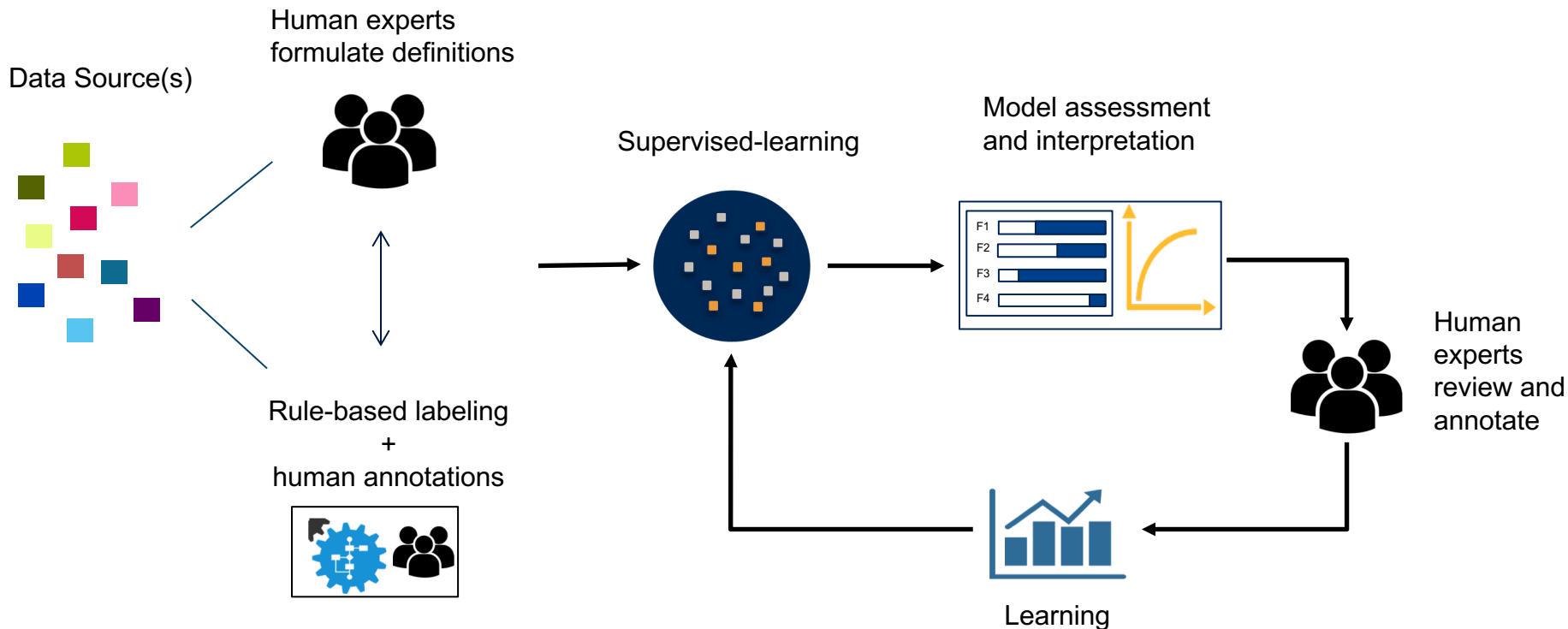
MEDS-SDTM Standard

30+ Domains, 155+ Variables

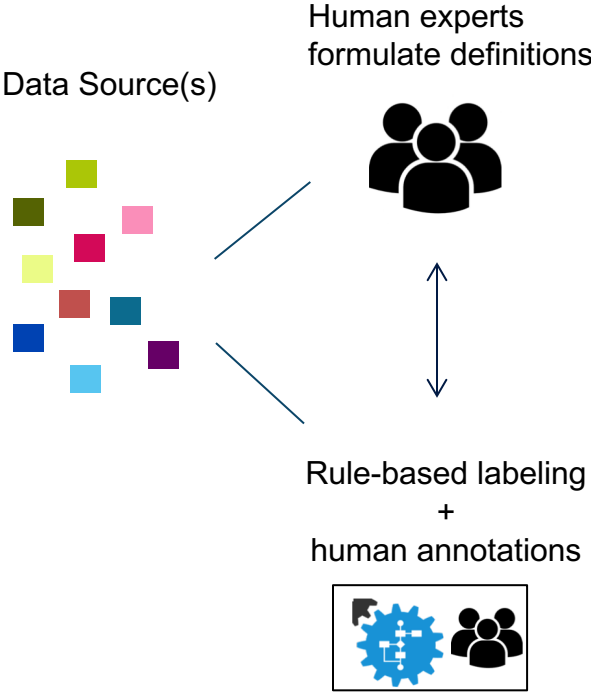


Human (Expert) in the Loop Machine Learning

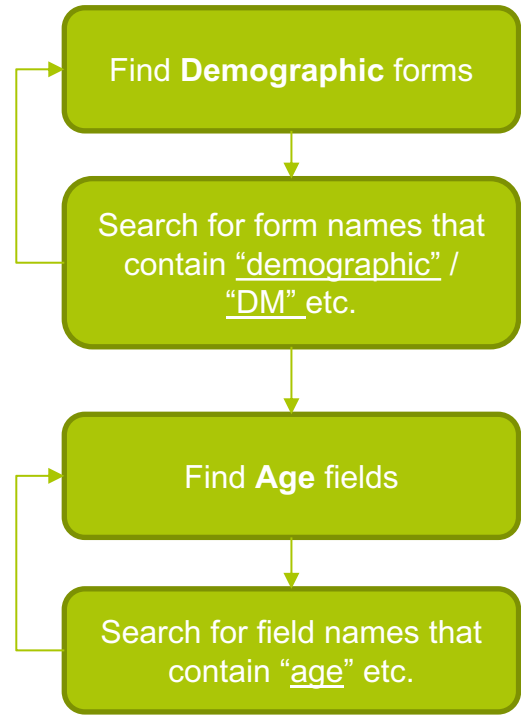
Faster in speed, Better in quality



Create Initial Labels



Example from DM domain



Filter Data and Specify RegEx Terms Here

ControlType_FP Predicted Form Y Filter Data Filter

Field Rule DataPageName Contains RegEx Rule Definition

Expression Submit

Download Enter Filename here Submit Rule Combination Apply Rule

Column visibility

PredictedForm	PreText_FP	SASLabel_DM	DataPageName	FolderName_DP	DictionaryName_DD
AE	Date Ended or R...	Stop Date Year...	adverse events	adverse events ...	
AE	Date Started	Start Date Info...	adverse events	adverse events ...	
AE	Severity	Severity,Sever...	adverse events	adverse events ...	CTCAE_SEVERITY...
AE	Serious?	Serious? Coded ...	adverse events	adverse events ...	NO_YES_2
AE	Did event occur...	Response No/Yes...	adverse events	adverse events ...	NO_YES_2
AE	Relationship: ...	Related to Chem...	adverse events	adverse events ...	NO_YES_2
AE	Relationship: ...	Related to Beva...	adverse events	adverse events ...	NO_YES_2
AE	Relationship: ...	Related to Inve...	adverse events	adverse events ...	NO_YES_2
AE	Did event start...	Event Started P...	adverse events	adverse events ...	NO_YES_2
AE	Intermittent?	Intermittent? C...	adverse events	adverse events ...	NO_YES_6

Showing 1 to 10 of 48,964 entries

2

Supervised-learning



Bag of Words



Feature transformation



Model selection



Form Classifier in 3 fold Cross Validation

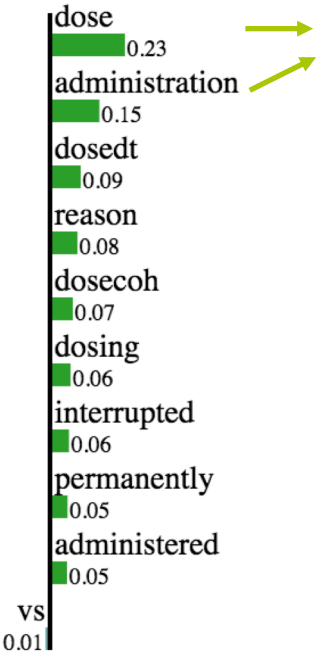
Method*	CV Accuracy	Build time (minutes)
Logistic Regression w/ regularization	0.95	7.5
Linear SVM	0.95	6.6
SGD Classifier	0.95	7.7
Random Forest	0.96	7.8
XGBoost Classifier	0.96	5.1

Raw Text with Highlighted Words

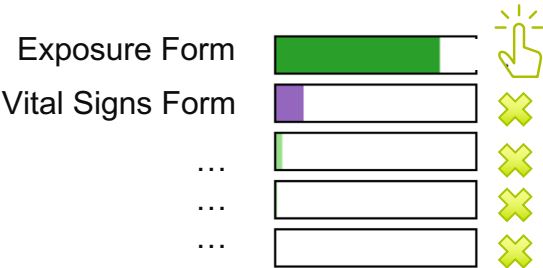
doseadminiv docetaxel administration iv
dosedlrs;dose;doserdsp;dosecoh;dosedel;dose
dlsp;doses;dosstrtm;doseint;dosedisc;dosered
;dosedssp;dosedt;dosedsrs;dosmodyn;dostpt
m;z_bsaxdosediff;doscalt;dosereas;doserdrs;
z_bsaxdose actual dose given (mg);bsa *
cohort dose per cycle (site calculated);cohort
dose per cycle (nominal dose);bsa * cohort
dose per cycle (site vs database
calculated);stop time (hh:mm 00:00-
23:59);dose delayed?;were there any dose
modifications?;if dose reduced, reason;if
infusion interrupted, reason;specify;was the
infusion interrupted?;if dose delayed,
reason;bsa * cohort dose per cycle (database
calculated);dosing permanently
discontinued?;if dosing permanently
discontinued, reason;dose reduced?;date
administered;start time (hh:mm 00:00-23:59)

Keywords Ranking

Not Exposure Form Exposure Form



Predicted Probabilities

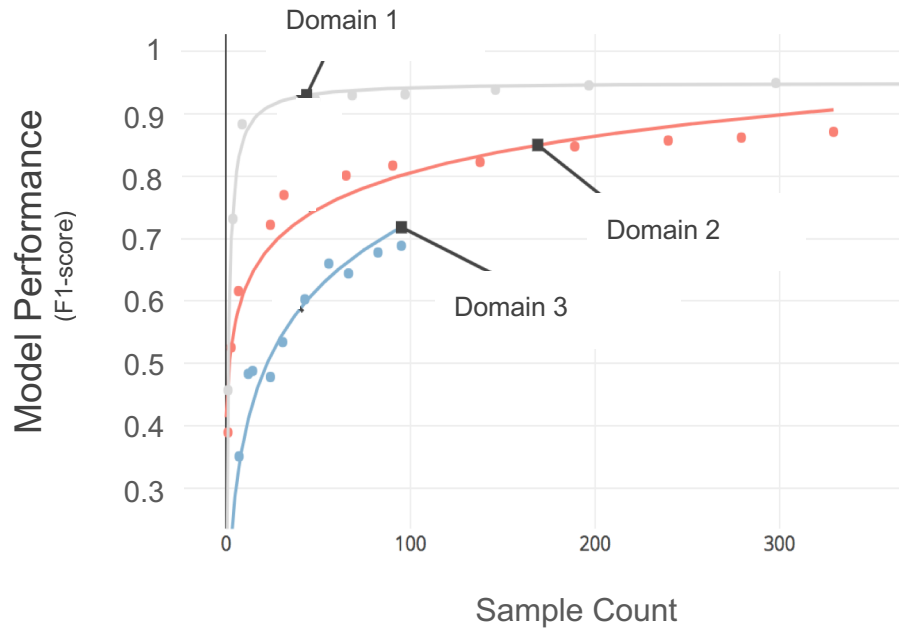


3

Model assessment and interpretation



Do we need more labeled data ?



4

Human experts review and annotate



Both, Classifier and RBL* agree

Classifier = RBL* = VS

Prediction probabilities

NOT VS

VS

VS	0.98
OTH	0.02
AE	0.00
BI	0.00
Other	0.00

vital	0.24
signs	0.24
pulse	0.12
dbp	0.08
sbp	0.07

Text in the raw data ...

vitals_1 vital signs
temp_c;visittm;visitdt;route;sbp;pulse;tempraw;respir;

*RBL: Rule-based labels

**VS: vital sign forms

4

Human experts review and annotate



Classifier is correct and RBL* is incorrect

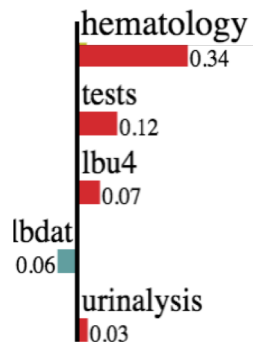
Classifier = LB**, RBL* = OTH**

Prediction probabilities

LB	0.69
OTH	0.35
AE	0.00
CM	0.00
Other	0.00

NOT LB

LB



Text in the raw data ...

hematology tests urinalysis dxx lbdatt;lbpoint;lbvalue

*RBL: Rule-based labels

**LB: study exposure forms

**OTH: other / unlabeled forms

4

Human experts review and annotate



Classifier is incorrect and **RBL*** is correct

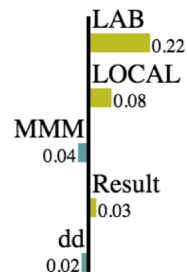
Classifier = LB**, RBL* = BI**

Prediction probabilities

NOT LB

LB

LB	0.39
OTH	0.37
LB-NR	0.13
BI	0.07
Other	0.05



Text in the raw data ...

gemu idh gene mutation analysis neg;geswap;sample;gemutdc;gemund;swapdte;p
Done;Bone Marrow;Sample type;;IDH1 Gene Result Negative;;CENTRAL **LAB**
(dd-MMM-yyyy);Date of sample: (dd-MMM-yyyy;LOCAL TESTING - IDH1 GI
(dd/MMM/yyyy)

*RBL: Rule-based labels

**LB: lab forms

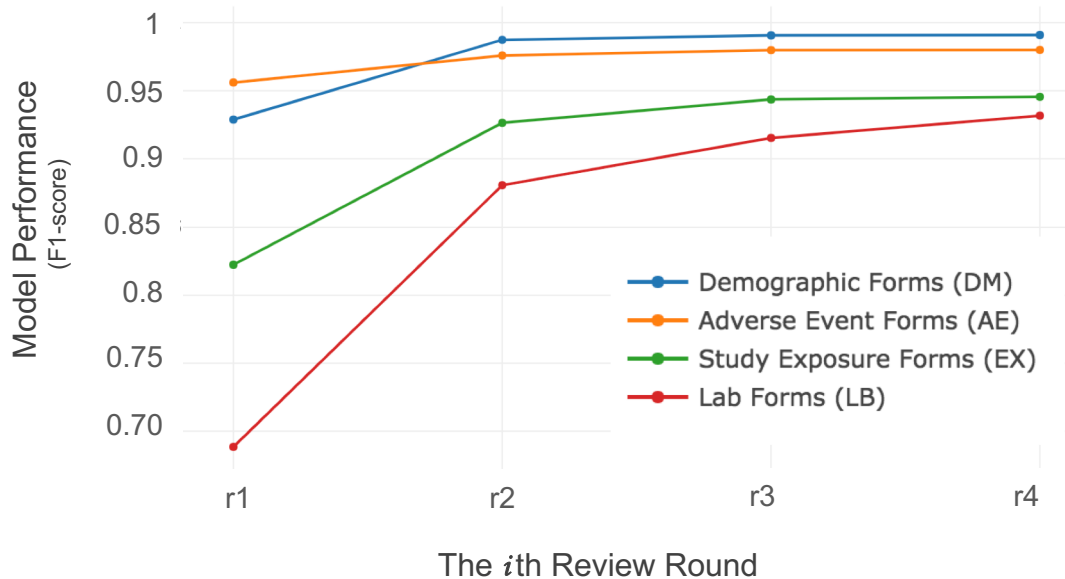
**BI: biomarker forms

5

Learning



**Model improved continuously
over 4 rounds of iterations**



Thank you.