SDTM Mapping Based on Natural Language and Machine Learning Models

CJUG 2019

Sam Tomioka,
Director of Clinical Data Programming
Data Science
SUNOVION Pharmaceuticals
02.07.2019

sunovion
Innovation today, healthier tomorrows

Brief Introduction

- Sam Tomioka (When I am outside of Japan)
- Director, Clinical Data Programming – Data Science
- Clarity Pharma -> Dainippon -> Dainippon Sumitomo -> Sunovion

AGENDA

Problem to Solve

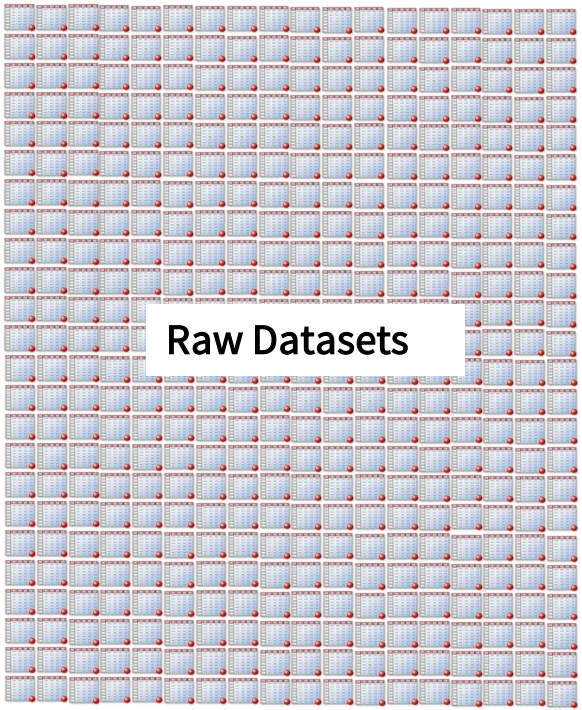
SDTM Mapping with “Machine Learning“

Tools Used

Thought

sunovion

A PROBLEM WE WANT TO SOLVE

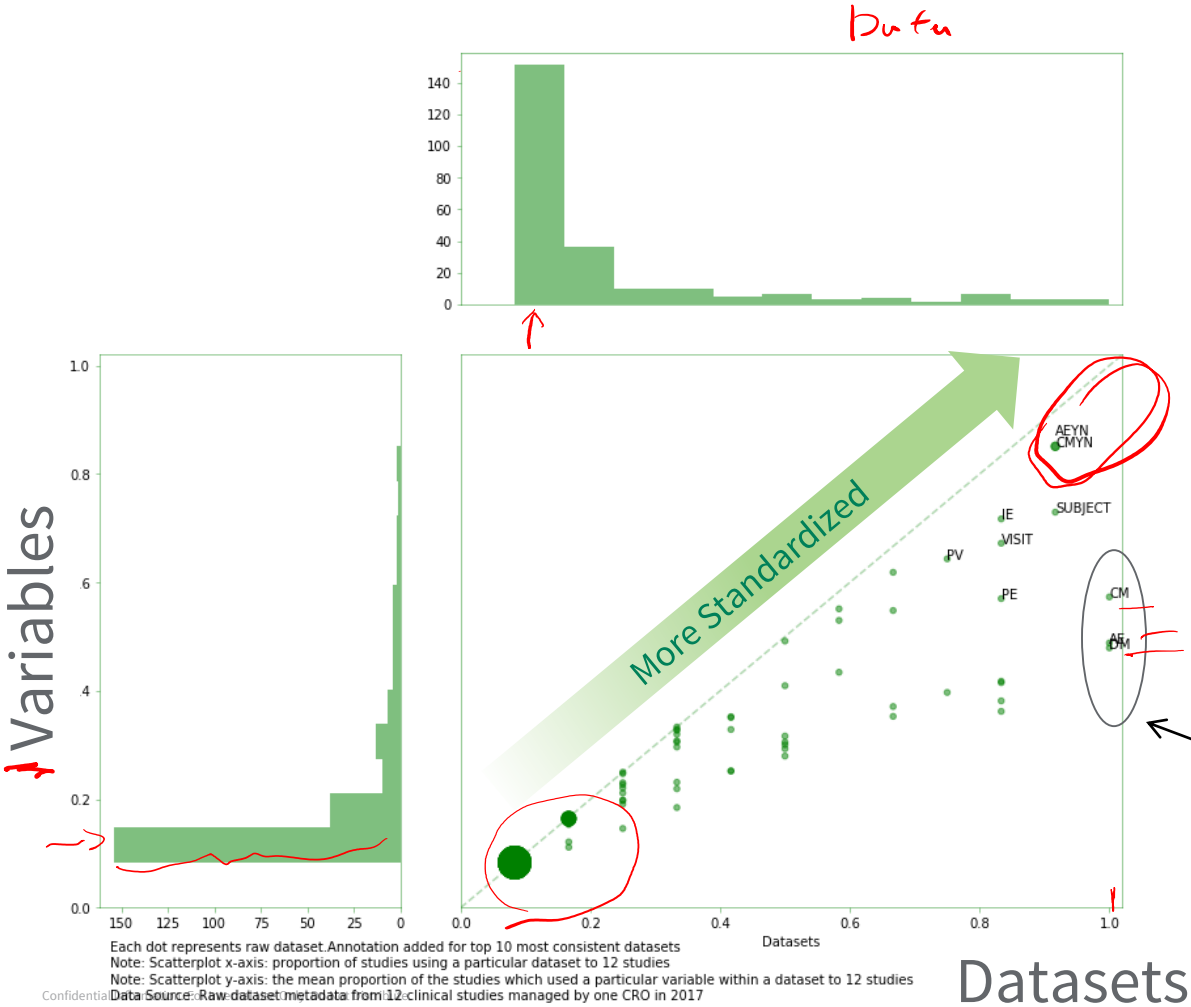


52 business days to 1st draft SDTM*

* 2017 data

sunovion

Reuse of datasets and variables across 12 studies



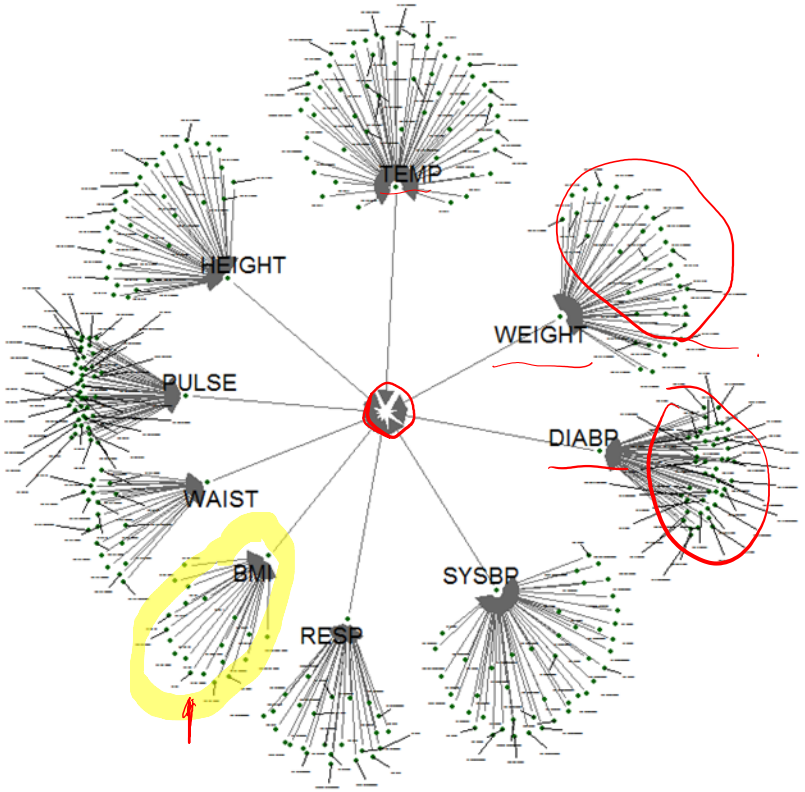
PROGRAMMERS ARE FRUSTRATED WITH NON-STANDARDIZED RAW DATA

Only 3 datasets used in all 12 studies

sunovion

ENDLESS MAPPING...

SDTM.VS.VSORRES mapping for
12 Sunovion studies delivered by one CRO



20 sources for BMI

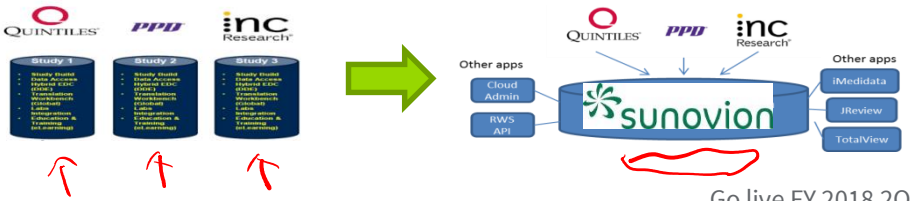
"VS.BMI"
"VS.BMI_RAW"
"VS.BMI_Z"
"VS.BMI_Z_RAW"
"VS.BMIS"
"VS.BMIS_RAW"
"VS1.BMI"
"VS1.BMI_RAW"
"VS1.BMI_Z"
"VS1.BMI_Z_RAW"
"VS1.VS1BMI"
"VS2.BMI"
"VS2.BMI_RAW"
"VS2.D_BMI"
"VS2.D_BMI_RAW"
"VS2.VS2BMI"
"VSMSTR.BMI"
"VSMSTR.BMI_RAW"
"VSMSTR.D_BMI"
"VSMSTR.D_BMI_RAW"

Data Source: ~/My Documents/r/programming_challenges_201711xx/inventory20171114_t.xlsx

sunovion

Embrace Metadata Standards

RaveX Implementation Project



Go live FY 2018 2Q

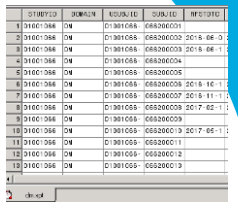
MDR Based SDTM Automation Project

Started FY 2018 4Q

SDTM Automation Project

SDTM automation macros
Domain templates

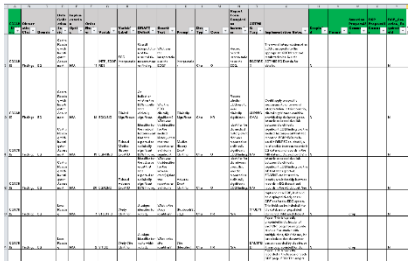
Go Live FY 2018 1Q



SDTM

Standard CRF Working Group

paper

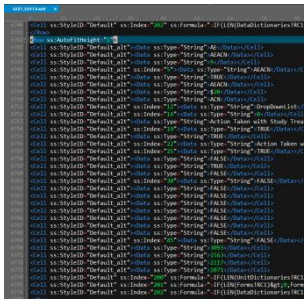


metadata

CDASH

cdisc

Global Library Volume Implementation Project



metadata in ODM XML

SDTM

CT

cdisc

Started FY2018 1Q

WHAT NEXT?

Can I use **natural language** model and **machine learning** algorithms to map raw data variables to SDTM variables?

AGENDA

Problem to Solve

SDTM Mapping with “Machine Learning“

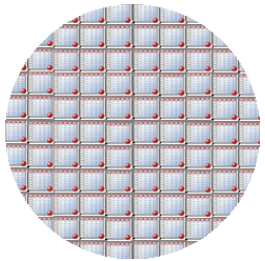
Tools Used

Thought

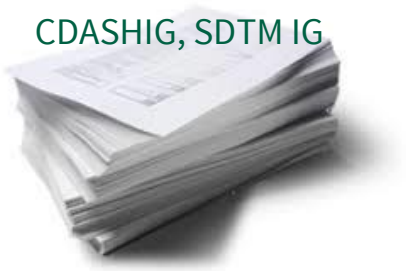
ML based SDTM mapping for fast, accurate, consistent SDTM generation

Input

Raw data



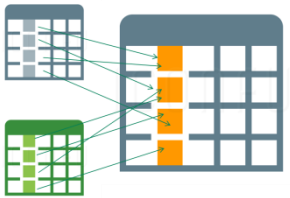
CDASHIG, SDTM IG



NLP & ML

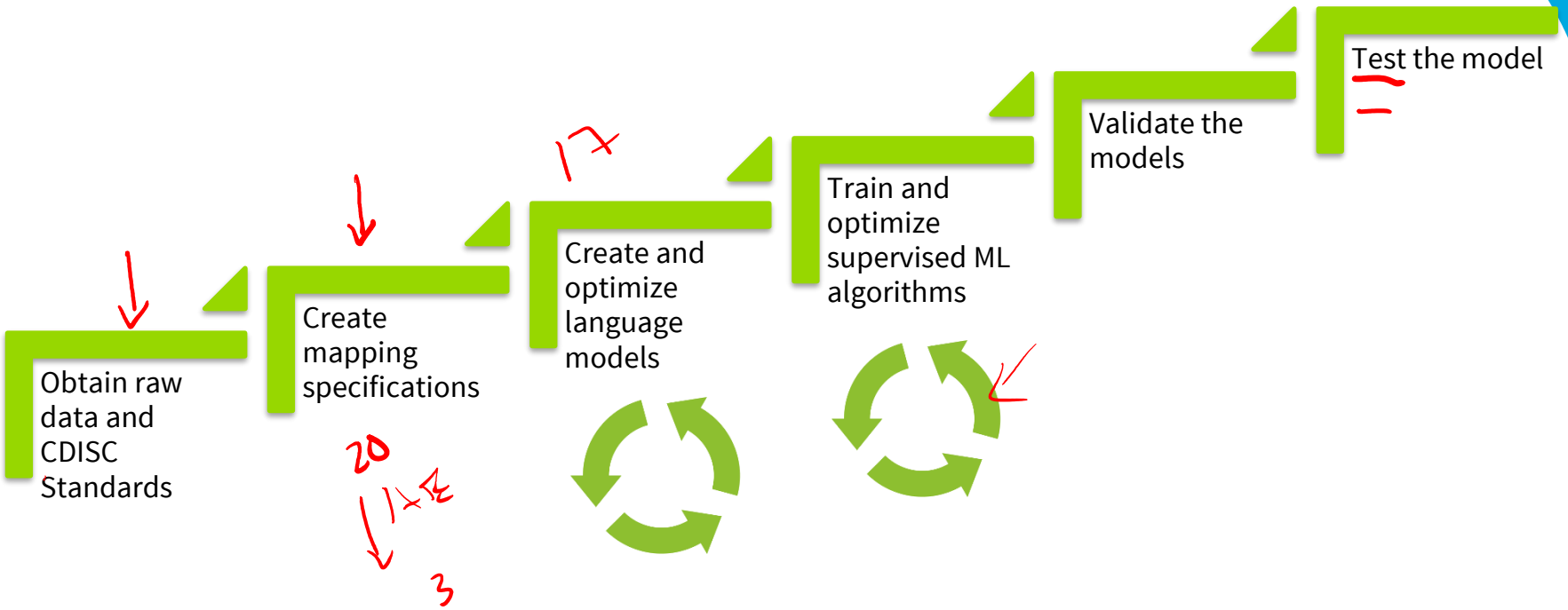
Output

Mapping Specification



sunovion

Steps



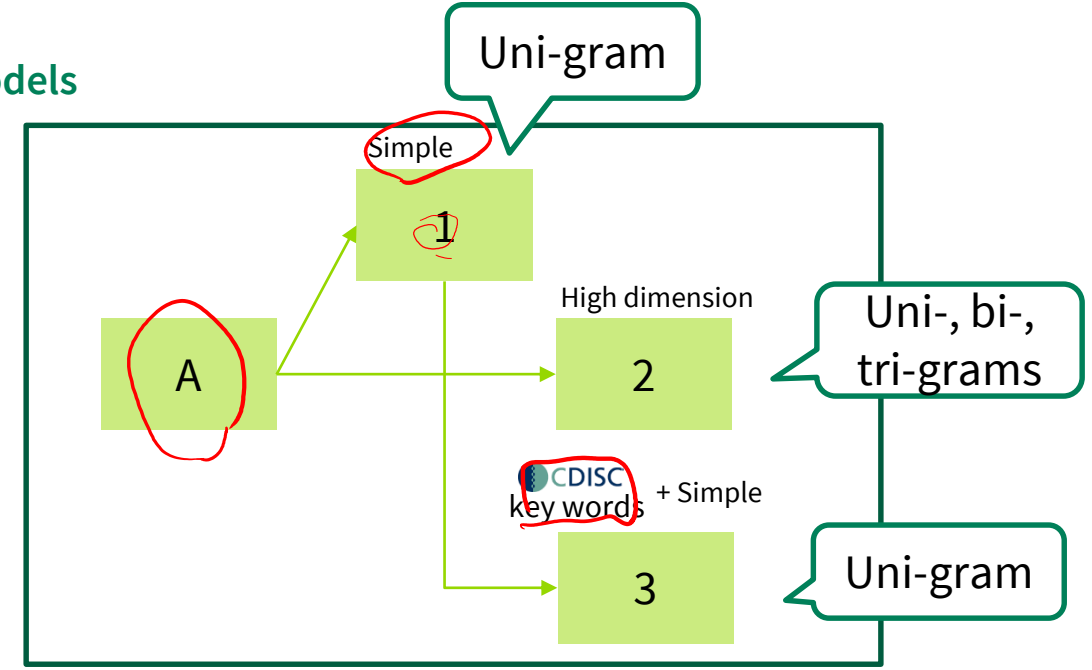
Create mapping specifications (with Human Intelligence)

Raw Variable	SDTM Variable
PT	AEDECOD
SOC	AEBODSYS
PTNAME	AEDECOD
SOCNAME	AEBODSYS
…	…

*illustration purpose only

Natural Language Models

TF-IDF



Neural Nets probabilistic language model

sunovion

Natural Language Model A

TF-IDF algorithm: Weighing terms

- Words occur nearby frequently are important than words that only appear once or twice

  Frequency (TF)

- Words that are too frequent are not important

  Inverse Document Frequency (DF)

- Weight

$$\mathrm{tf}_{t,d} = \begin{cases} 1+\log_{10}\mathrm{count}(t,d) & \text{if } \mathrm{count}(t,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{idf}_t = \log_{10}\left(\frac{N}{\mathrm{df}_t}\right)$$

$$w_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t$$

AETERM definition in CDASH IG

The reported or pre-specified name of the adverse event.

0.6331503 0.6105753 0.6247964 0.6611213

Logistic regression, accuracy=0.64058

Document term matrix of 17 studies data and IG

sunovion

Natural Language Model B

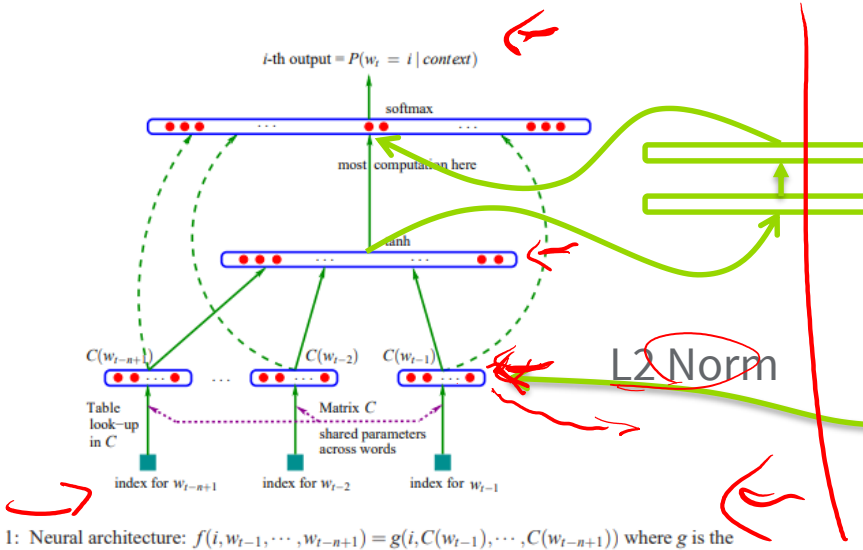


Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i-th word feature vector.

L2 Norm

ML → $\hat{Y}$

Raw metadata, CDISC IG etc

(m,128)

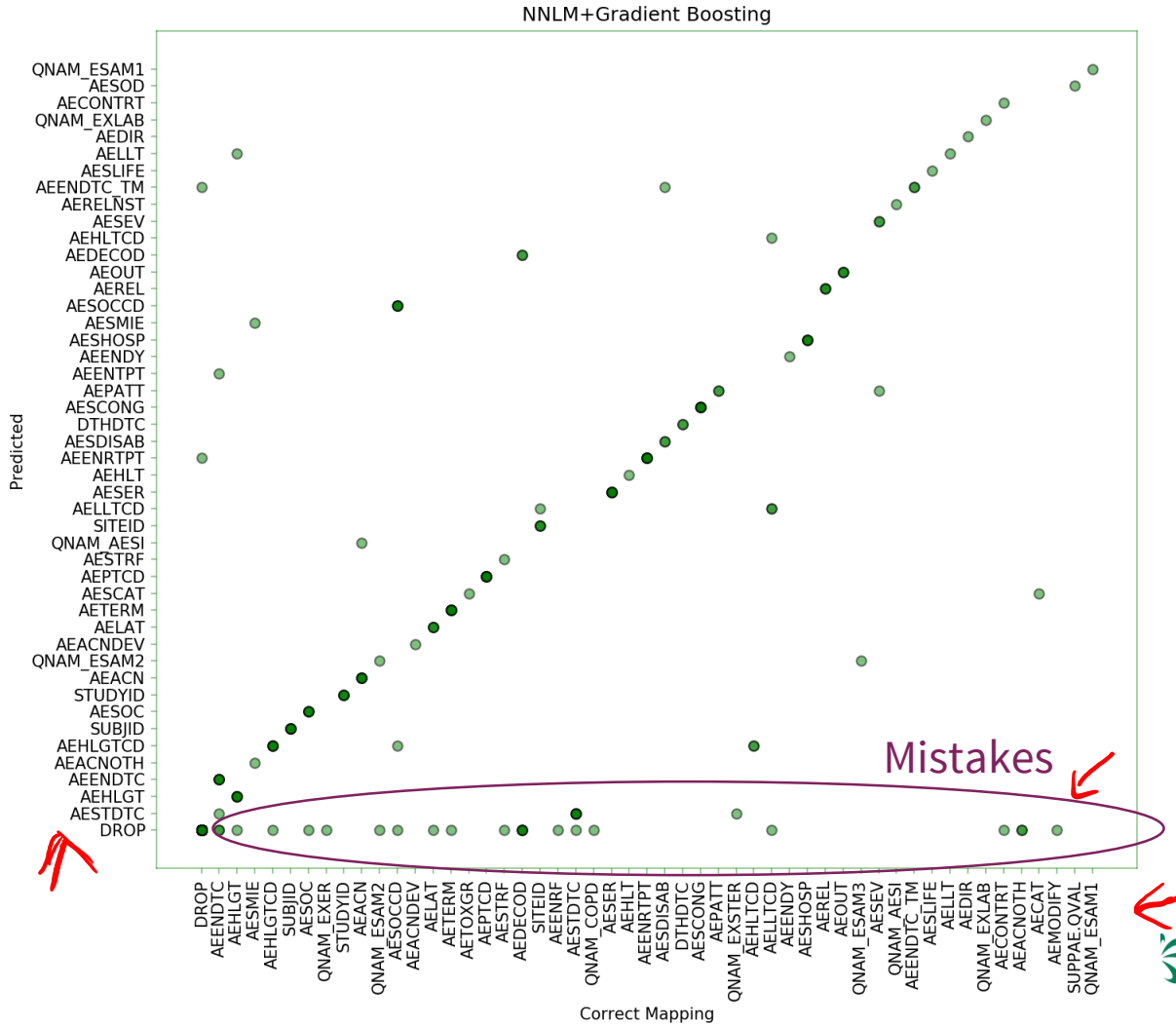
200,000,000,000 English Google News corpus

Figure from Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3:1137-1155, 2003.

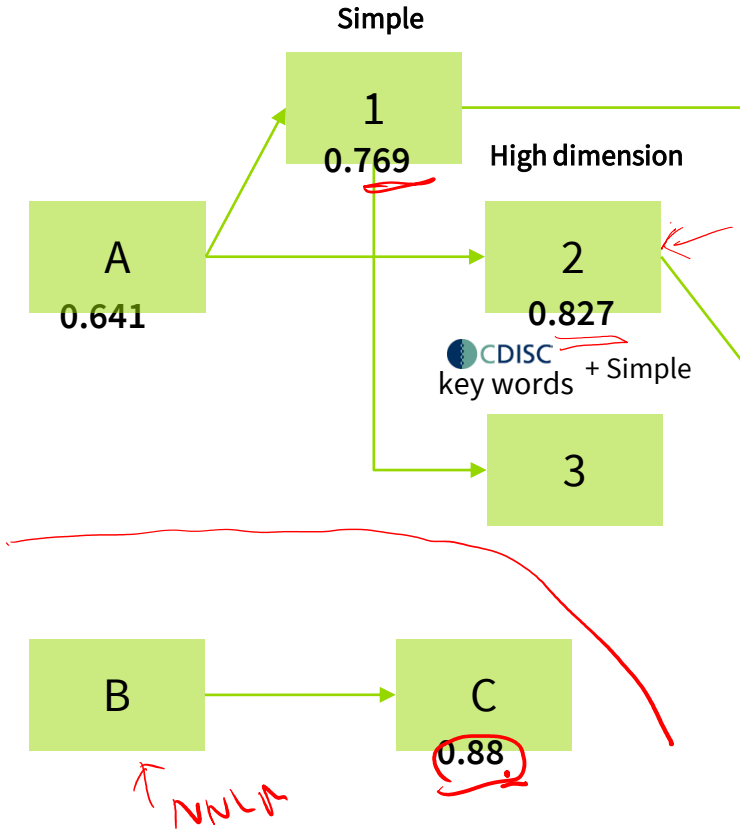
Natural Language Model B

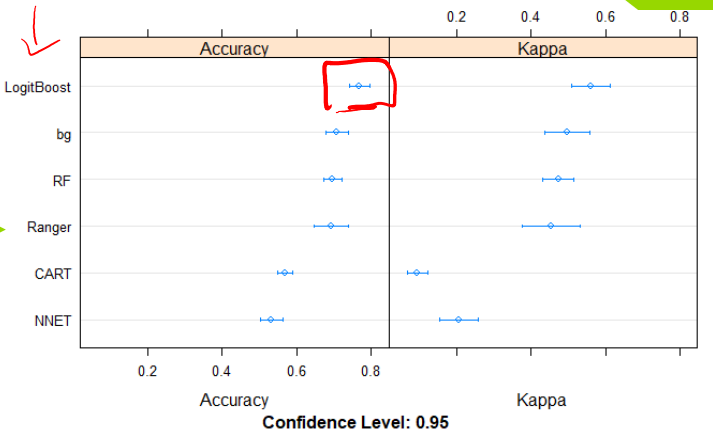
```
(1, 128)
The reported or pre-specified name of the adverse event. =  [[ 0.4712333   0.24399556  0.04435228 -0.10534722 -0.12959094  0.01965462
  0.00061709  0.11828943  0.06076331  0.11396483  0.21145034  0.09158859
 -0.16773823 -0.06314589 -0.00201162  0.17754264 -0.06478326  0.04580432
 -0.08311215  0.51585186  0.1620522   0.08714737  0.0600881   0.04144981
  0.08530827  0.01127403 -0.09444343  0.06110086  0.06098798  0.08598933
  0.16139089 -0.10500295  0.03692641 -0.06367766 -0.02176712 -0.0580758
  0.04609384 -0.01650539  0.04840343  0.07840265 -0.04859402 -0.12693526
  0.04102192 -0.08049966  0.00896453 -0.04968296 -0.04114896  0.08899469
  0.04490945 -0.0051559   0.03986294  0.08894438  0.17566486  0.0614834
 -0.07853805 -0.01386436  0.1349262  -0.04074099  0.08051076  0.05174435
  0.11551834 -0.06626088  0.01692365  0.0140045   0.04898851  0.0913236
  0.00335585 -0.01931612  0.08324809  0.01950045  0.02538943 -0.00366753
  0.1219741   0.09413652 -0.0340781   0.07892839 -0.11236215 -0.17992431
  0.04012968  0.0132846   0.09931402 -0.09699554  0.05006879  0.00603813
 -0.11555596 -0.10244549 -0.24568063 -0.14376119  0.01996353  0.15219717
 -0.00676261 -0.01049919  0.09959473 -0.21953513 -0.06734764 -0.01071521
 -0.14200029 -0.10907198 -0.06779519  0.04904051 -0.15821819  0.14623989
  0.01035561 -0.01564043  0.00844536  0.06118653 -0.12851241 -0.09161683
 -0.16958985 -0.10710518 -0.01085521 -0.1233172  -0.08650532 -0.05034445
  0.04382536 -0.12378055  0.17017077 -0.04361667 -0.00717157  0.16402934
  0.16050048 -0.19754027  0.0992441   0.00550133 -0.0244873   0.01215202
  0.05600908 -0.0494738 ]]
```

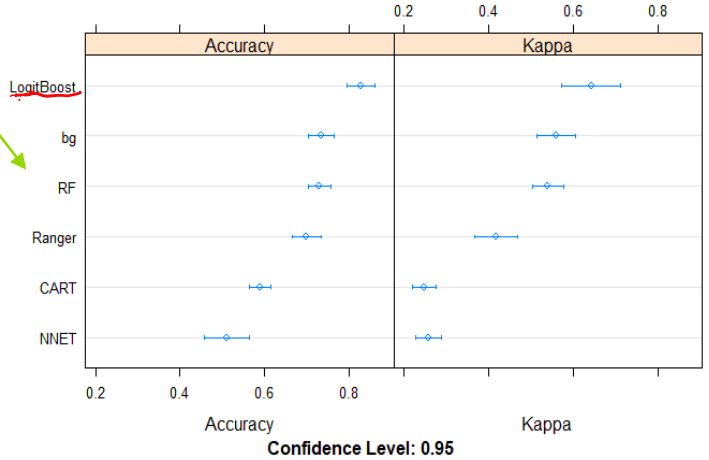
This model yield **0.88** accuracy with a gradient boosting from 10 fold cross validations



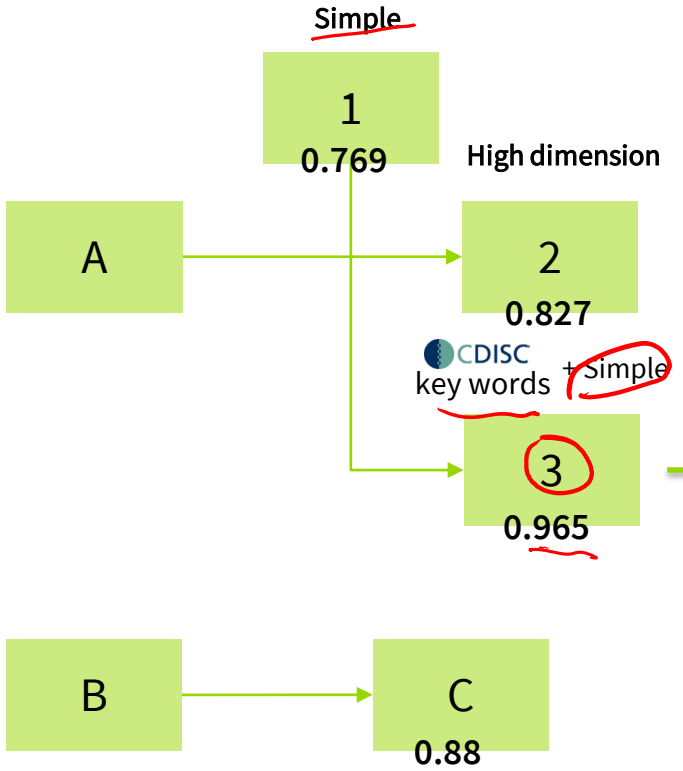
Natural Language Models







Natural Language Models



Machine Learning Models

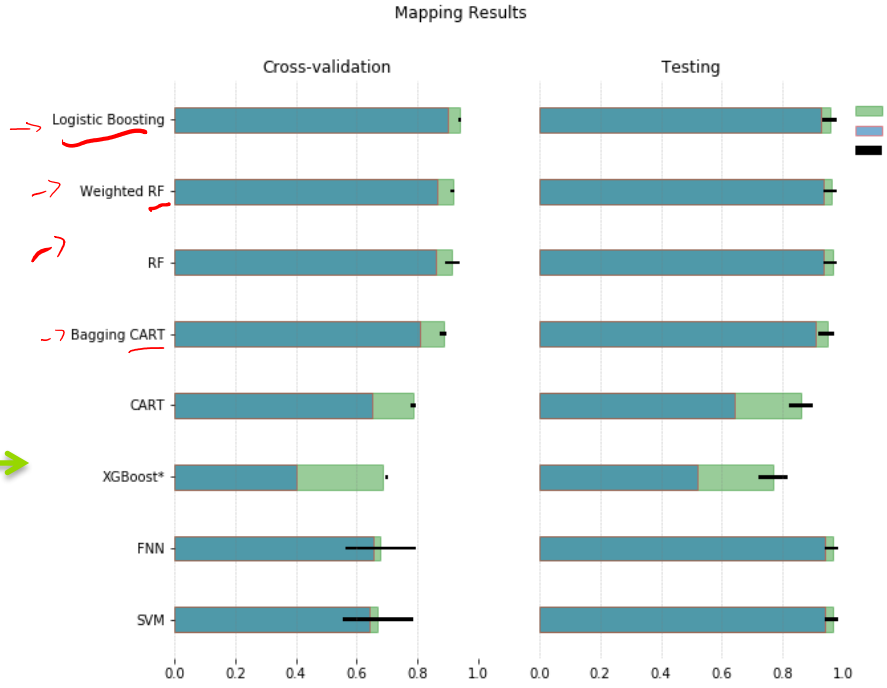
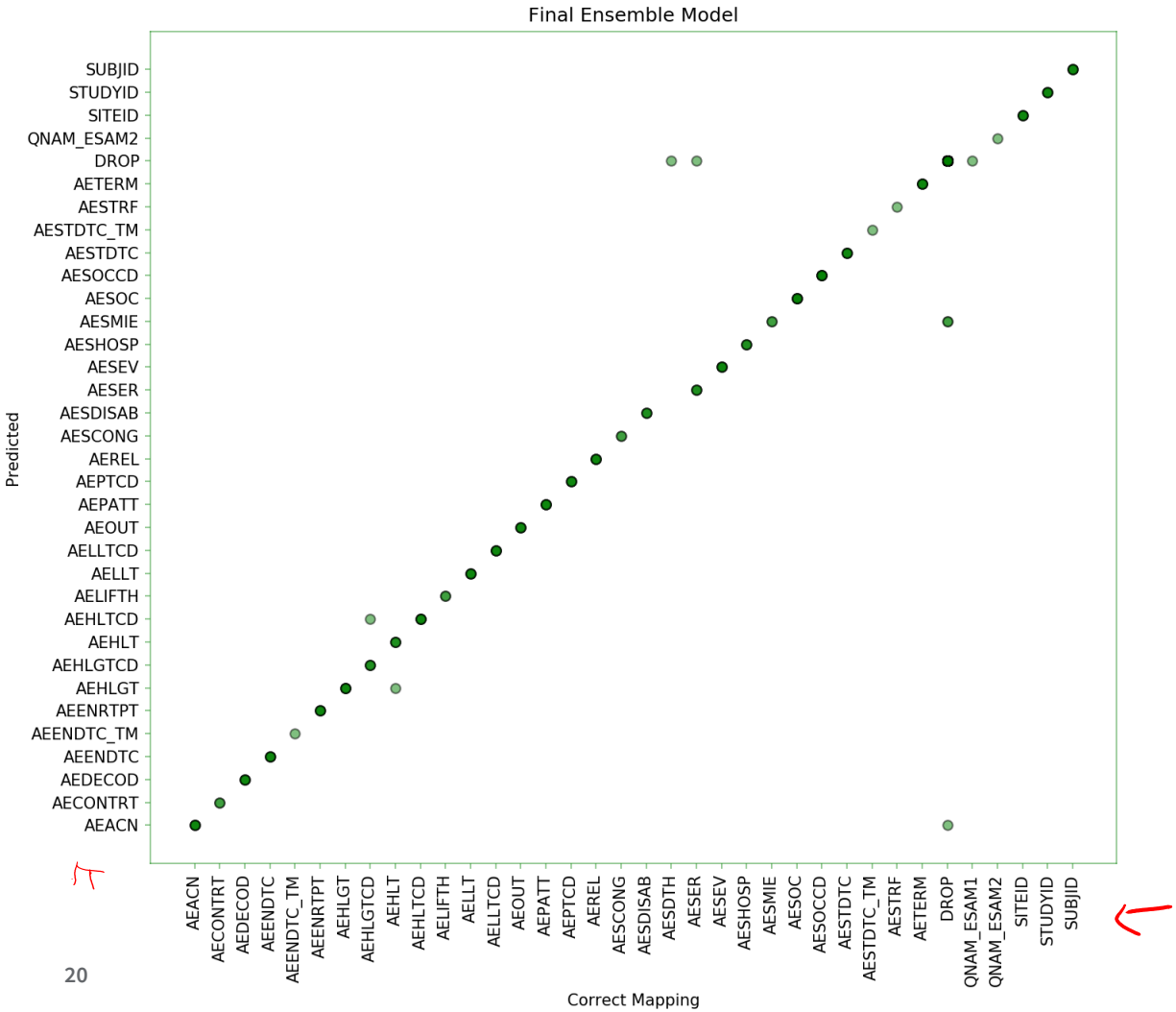


Figure 6. Results of SDTM mapping (Accuracy with 95% CI and Kappa)



Mapping Accuracy on data from 3 new studies using Final Ensemble Model

0.97484

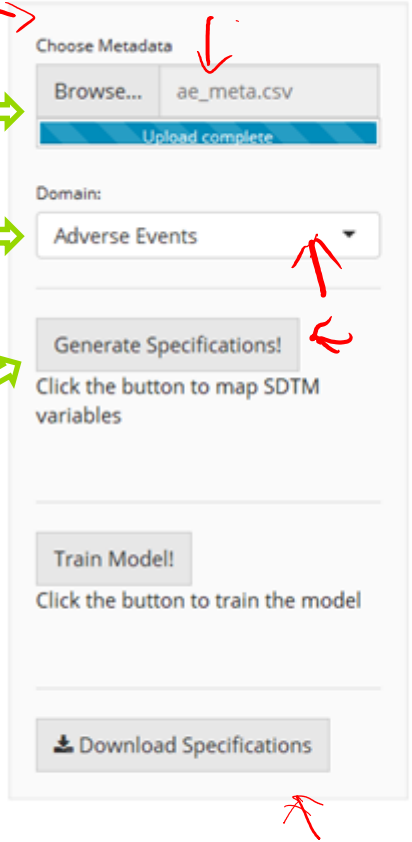
Machine SDTM Mapper

Prototype GUI

Upload Raw data metadata

Select Domain

Let AI create the mapping



Contains real data

Mapped by AI

Table | Editor | Confusion Matrix

Edit the SDTM variables as necessary

Show 10 entries

Search:

	X	ID		sdtm
1	1	AE.subjectId		DROP
2	2	AE.AESTDTC_RAW		AESTDTC
3	3	AE.AEACN		AEACN
4	4	AE.AETERM_LLT_CODE		AELLTCD
5	5	AE.AESER_STD		AESER
6	6	AE.FolderSeq		DROP
7	7	AE.AEONGO_RAW		DROP
8	8	AE.instanceId		DROP
9	9	AE.AEOUT_STD		DROP
10	10	AE.AEONGO		AEENRTPT

Showing 1 to 10 of 1,669 entries

Previous 1 2 3 4 5 ... 167 Next

AGENDA

Problem to Solve

SDTM Mapping with “Machine Learning“

Tools Used

Thought

sunovion

Tools used for development

Programming	IDE	ML Framework	Purpose
	NA	NA	Metadata extraction from sas7bdat
		caret	NLP ML Visualizations
			Transfer learning ML Visualizations

AGENDA

Problem to Solve

SDTM Mapping with “Machine Learning“

Tools Used

Thought

sunovion

Summary

- This proof of concept demonstrated that **machine learning** along with a **natural language model** can produce a pretty accurate SDTM mapping specification document.
- Light weight classic approach works.
- Transfer training has potential for reducing the amount of time to prepare training data

sunovion

Current/Future Works

- Optimization of NNLM based algorithm
- Transfer learning from other state of the art word embeddings
- End to end machine learning to step wise approach
- Development of language model for CDISC documentations

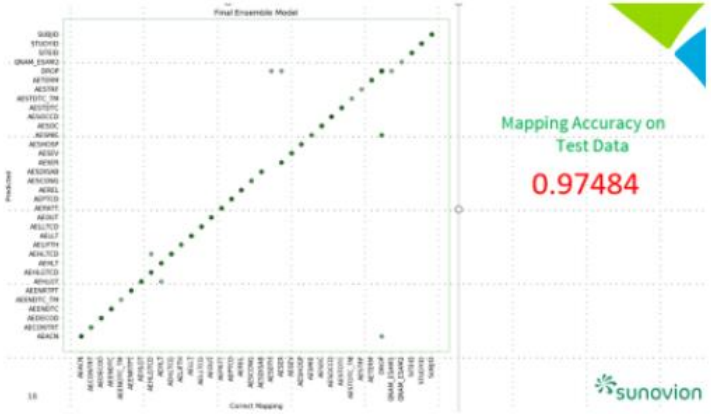
sunovion

Machine Learning Project Sub-team

Computational Science Working Groups

SDTM Mapping

- Project initiated based on the positive outcome of the POC presented at CDISC US Interchange 2018
  - weighting scheme for word-variable matrices
  - neural network language model
- Purpose
  - To explore and develop the SDTM mapping algorithms using NLP, ML, and NN.
  - To publish whitepaper on SDTM mapping



phuse.eu/cs-working-groups.aspx

Thank you

www.linkedin.com/in/STomioka

Sam.Tomioka@Sunovion.com

sunovion

Innovation today, healthier tomorrows

Back up

Unsupervised Machine Leaning Approach