SDTM Mapping Based on TF-IDF and Neural Network Probabilistic Models

Sam Tomioka

NJ CDISC Users Group Meeting

March 12, 2019

CDISC Strength Through Collaboration

Brief Introduction

- Sam Tomioka
- Director, Clinical Data Programming – Data Science
- Sunovion Pharmaceuticals

- Current ML Projects
  - SDTM Mapping
  - Protocol Optimization
  - Digital Endpoints (Seizure)

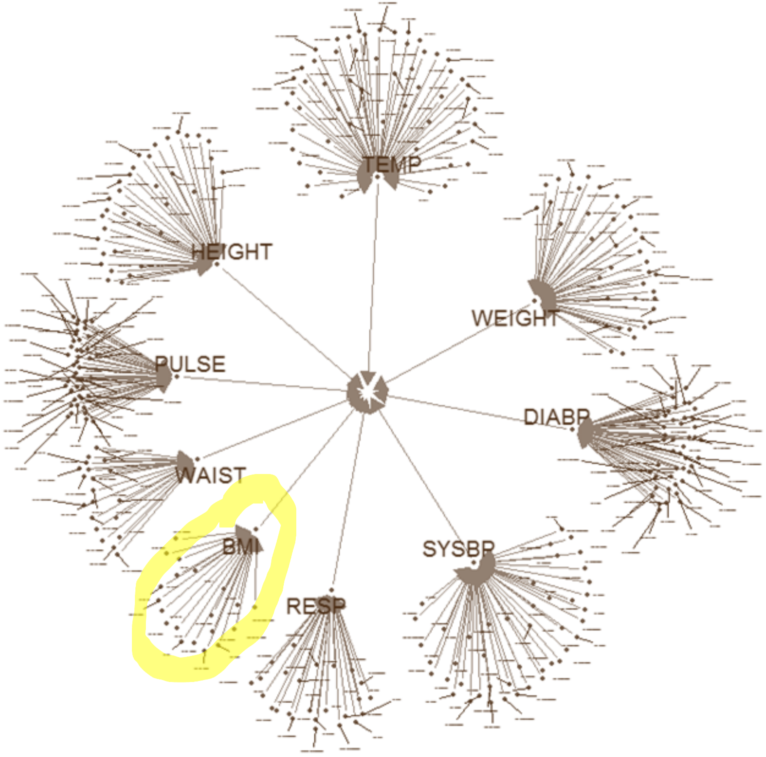
- Past ML Projects
  - Adverse Events
  - Dose Titration
  - Digital Endpoints (Stroke, Depression)
  - SDTM Mapping

AGENDA

- Problem to Solve
- SDTM Mapping with “Machine Learning“
- Tools Used
- Thought

ENDLESS MAPPING…

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



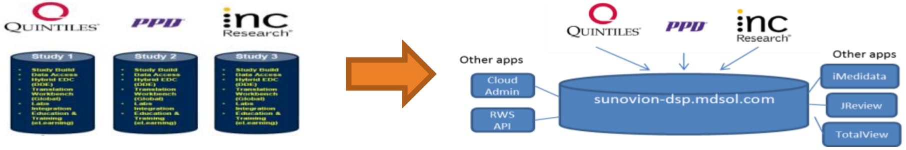
Data Source: ~/My Documents/r/programming_challenges_201711xx/inventory20171114_t.xlsx

```
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"
```

Embrace Metadata Standards

RaveX Implementation Project



Go live FY 2018 2Q

MDR Based SDTM Automation Project

Started FY 2018 4Q

SDTM Automation Project

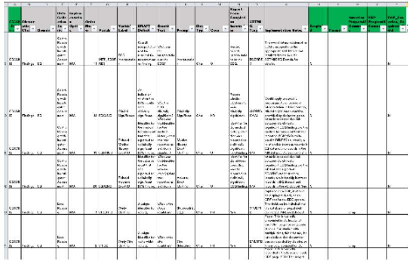
SDTM automation macros
Domain templates
Go Live FY 2018 1Q

SDTM

Standard CRF Working Group

- Sunovion Standard CRF Final v1.0_Central_EG ...
- Sunovion Standard CRF Final v1.0_Central_LB ...
- Sunovion Standard CRF Final v1.0_DA ...
- Sunovion Standard CRF FINAL v1.0_EX ...
- Sunovion Standard CRF FINAL v1.0_EX-DA ...
- Sunovion Standard CRF FINAL v1.0_PV ...
- Sunovion Standard CRF FINAL v1.0_RAND ...
- Sunovion Standard CRF FINAL v1.0_SU ...
- Sunovion Standard CRF FINAL v1.0_TC ...
- Sunovion Standard CRF FINAL v2 0 DS1 ...

paper

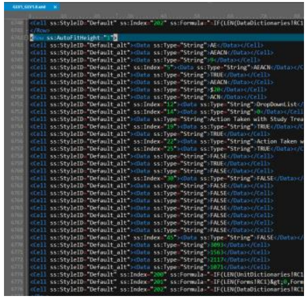


metadata

CDASH

cdisc

Global Library Volume Implementation Project



SDTM

CT

cdisc

metadata in ODM XML

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“

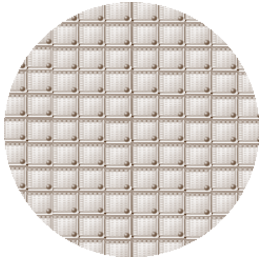
Tools Used

Thought

ML based SDTM mapping for
<u>fast</u>, <u>accurate</u>, <u>consistent</u> SDTM generation

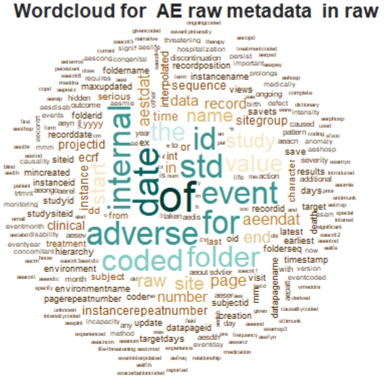
Input

Raw data



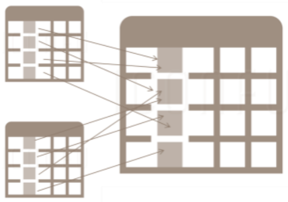
CDASHIG, SDTM IG

NLP & ML



Output

Mapping Specification



Steps

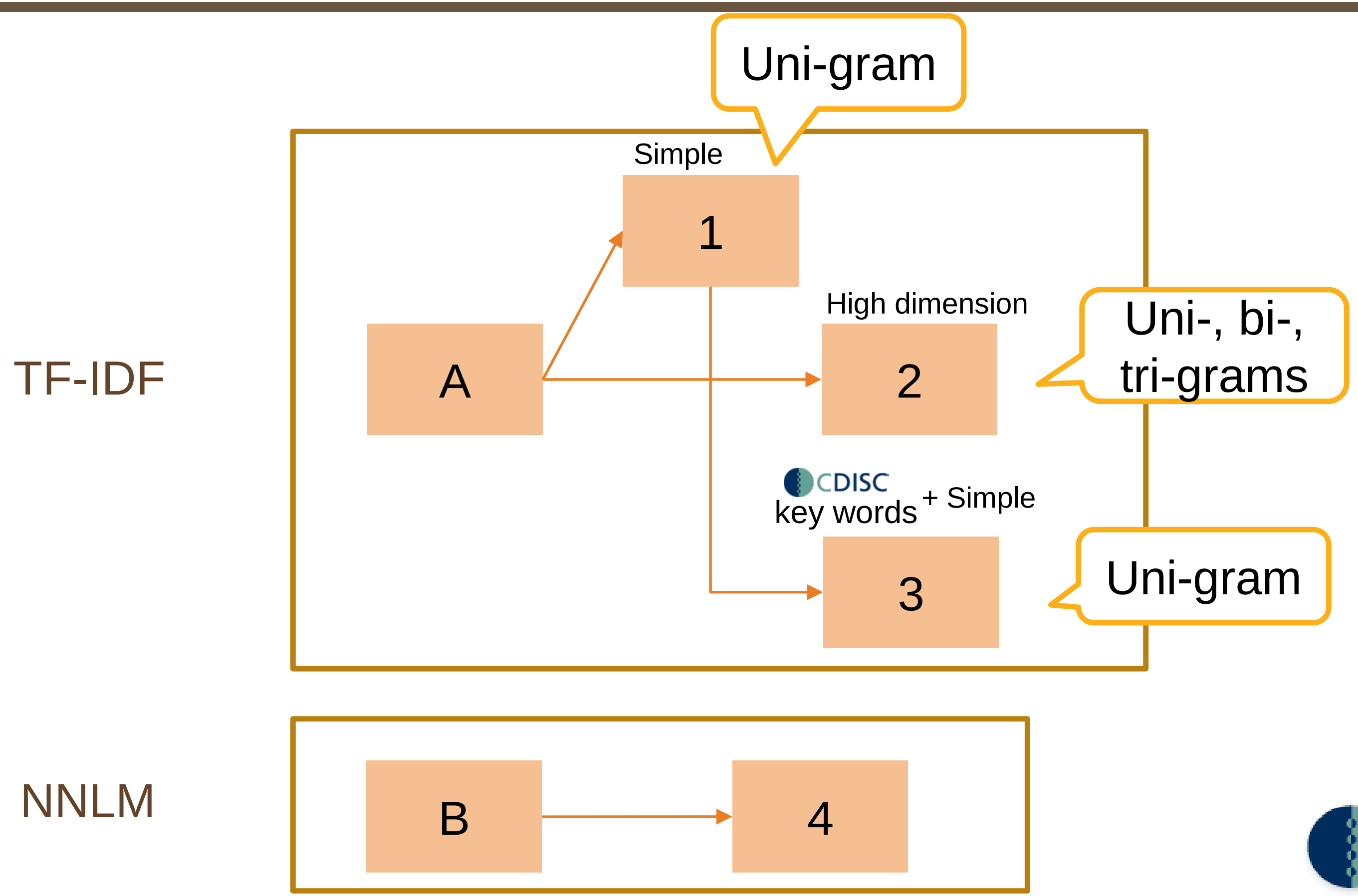
- Obtain raw data and CDISC Standards
- Create mapping specifications
- Create and optimize language models
- Train and optimize supervised ML algorithms
- Validate the models
- Test the model

Create mapping specifications (with Human Intelligence)

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

*illustration purpose only

Natural Language Models



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

- Words that are too frequent are not important

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

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

Logistic regression, accuracy=0.64058

AETERM definition in CDASH IG

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

0.6331503 0.6105753 0.6247964 0.6611213

→ LR → $\hat{y}$

Document term matrix of 17 studies data and IG

Natural Language Model B

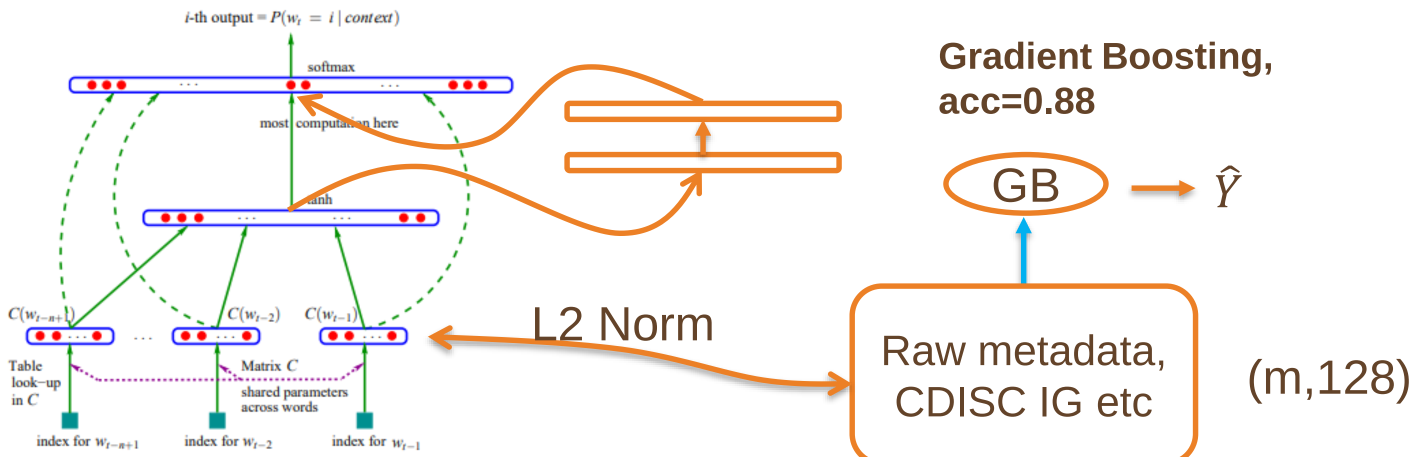


Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i-th word feature vector.

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 Models

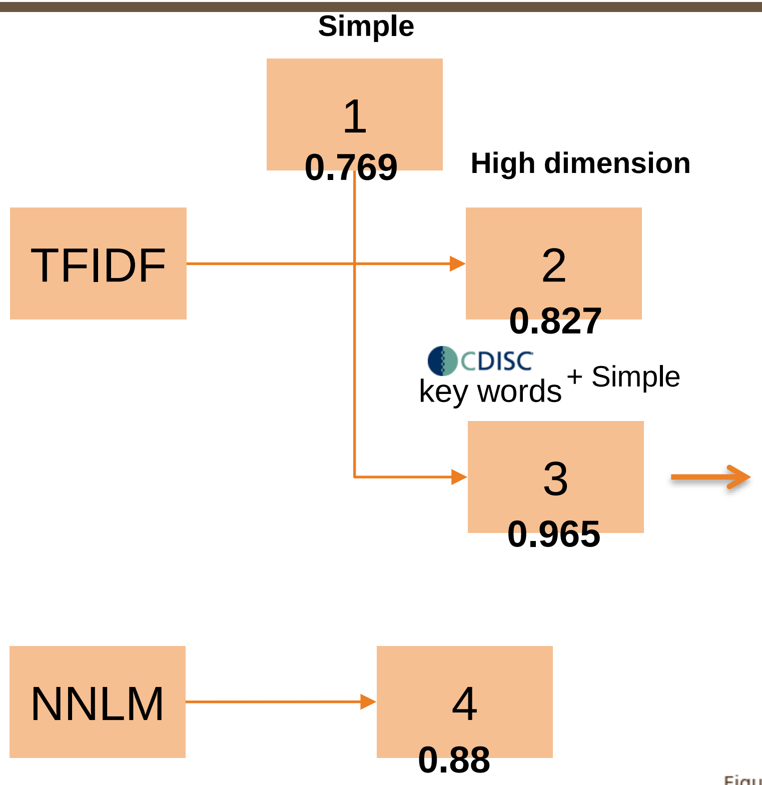
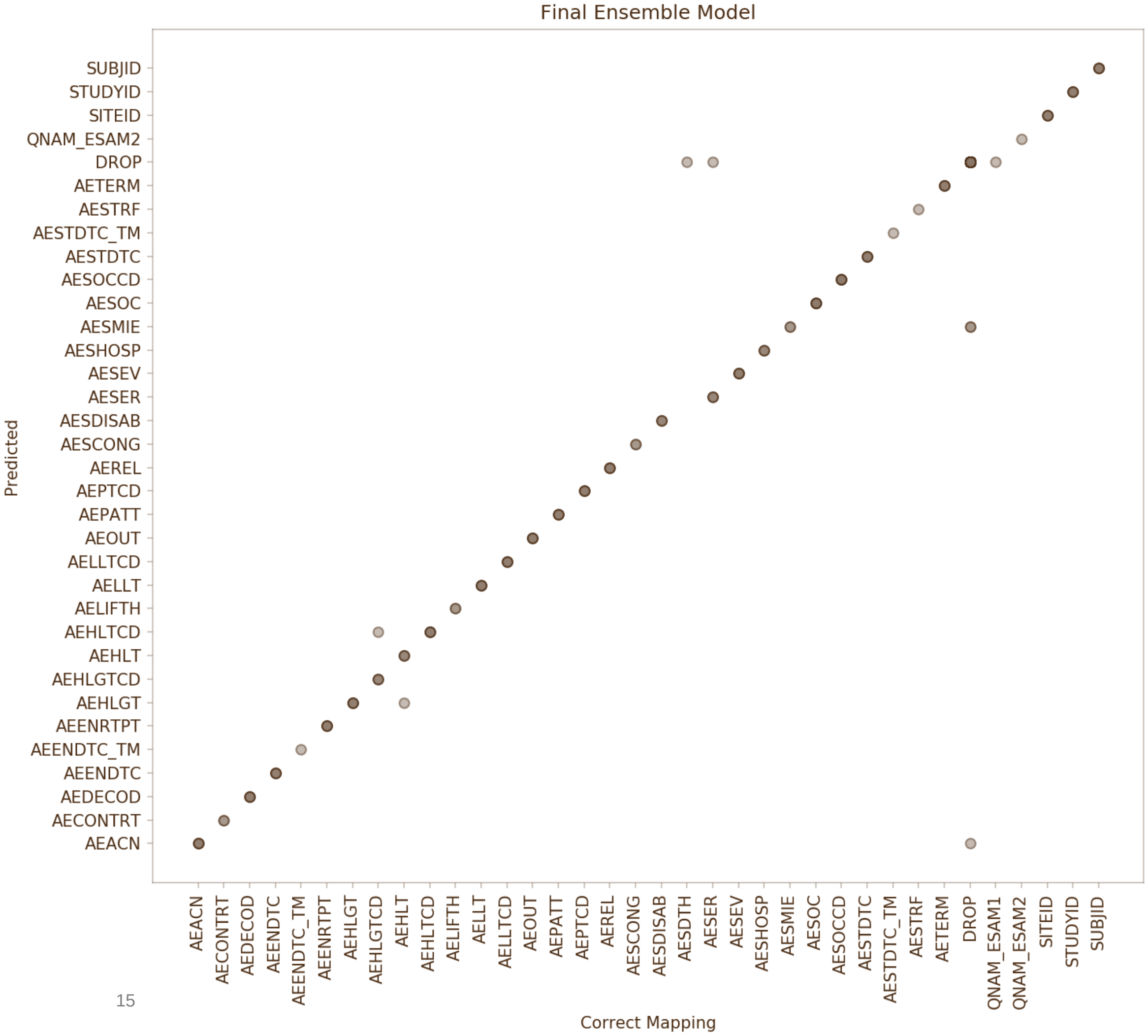


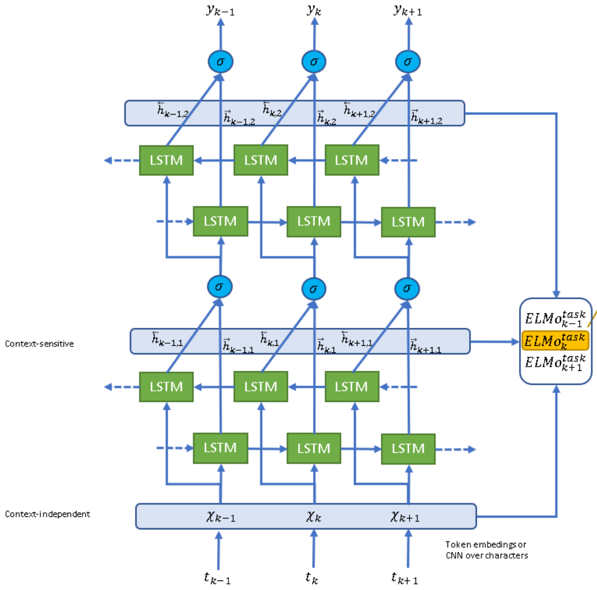
Figure 6. Results of SDTM mapping (Accuracy with 95% CI and Kappa)



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

0.97

More robust approach

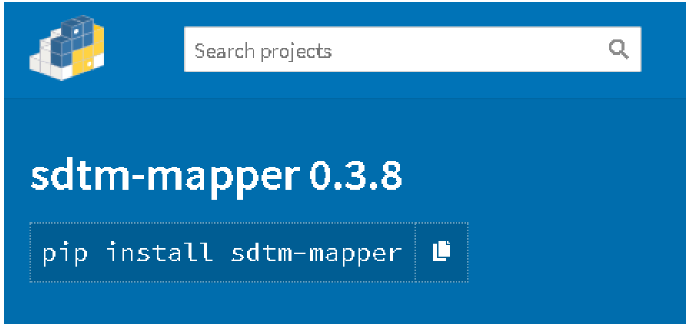


$$E(R_k;\Theta^{task}) = \gamma^{task}(s_0^{task}x_k + s_1^{task}h_{k,1}^{LM}) + s_2^{task}h_{k,2}^{LM})$$
$$= \gamma^{task}\sum_{j=0}^{L} s_j^{task} h_{k,j}^{LM}$$

$R_k = \{x_k, \overleftarrow{h}_{k,j}, \vec{h}_{k,j} \mid j = 1, \dots, L\}$

s_j^{task} *softmax-normalized weights*

γ^{task} *scalar parameter*



AGENDA

Problem to Solve

SDTM Mapping with “Machine Learning“

Tools Used

Thought

Tools used for POC

Programming	IDE	ML Framework	Purpose
sas	SAS	NA	Metadata extraction from sas7bdat
R	RStudio jupyter	caret	NLP ML Visualizations
python™	jupyter	scikit learn	Transfer learning ML Visualizations

AGENDA

Problem to Solve

SDTM Mapping with “Machine Learning“

Tools Used

Thought

Summary

- This proof of concept demonstrated that **machine learning** along with a **natural language model** can produce a pretty accurate SDTM mapping specification document.
- As in any ML models, as you feed more mapping specs, the model will learn them and become more robust.

Thank You

www.linkedin.com/in/STomioka

Sam.Tomioka@Sunovion.com