PhUSE US Connect 2019
Paper ML11

Smart Transformation of Clinical & Nonclinical Data for Insights

Isaac Mativo, Raja Ramesh, Phaneendra Bonda
PointCross Life Sciences Inc.

ABSTRACT

Disparate data with inconsistent data models, terminologies, and unstructured descriptions from studies need to be ingested into a searchable data store. Smart transformation replaces fixed adaptors and mappers as an important part of curation to make study data searchable across studies to gain insights.

Smart transformation uses machine learning to transform clinical, nonclinical and biomarker data from data lakes to a target model with automation. Supervised, expertly curated datasets train multiple deep neural network models that transform disparate source data. Recommendation engines using ontologies and vocabularies referenced in the target data model definition harmonize the transformed data. The smart transformers continually improve, learn and adaptively evolve as data managers intervene, assert or correct errors in transformation or users make decisions on metadata, content and terminology recommendations. This artificial intelligence augmented automation promotes data normalization and harmonization for search analytics as well as for regulatory packaging of eData.

INTRODUCTION

Most BioPharma companies accumulate collected data on nonclinical and clinical studies, and the molecular biomarker data from their bio-samples and hold them in their native format (SAS, Excel, flat files, etc.). These are the "data lakes" from which precisely the data that serves a business purpose should be read, and transformed for that business purpose. Curating this data for scientific uses such as cross-study cohort identification or analysis is well known to be time and labor consuming.

There are many opportunities for automation that reduces the time and effort while improving quality. These include identifying the needed data, semantically mapping and transforming using deep neural networks based recommendation engines, to self-organize the data using supervised machine learning. This improves the speed and quality of such curation through "Smart Transformation".

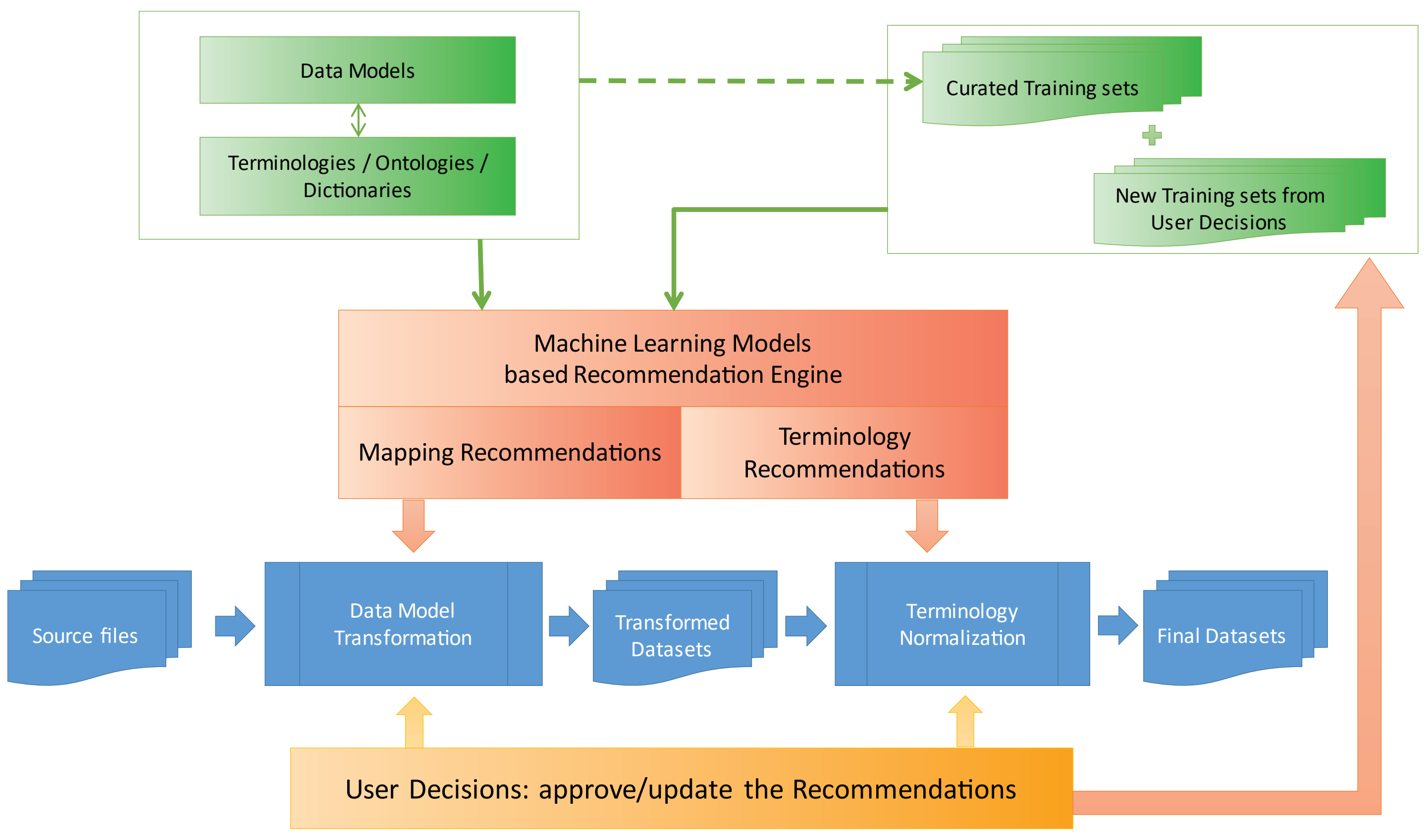
METHODOLOGY

Multiple Deep Neural Network based models are trained on Clinical, Non-clinical and Biomarker data, and these are executed in hierarchical sweeps with some heuristics. This process may result in more than one recommendation with different confidence levels, and these are further summarized to provide final recommendations.

These neural network models will learn and adapt to the data that is being brought into the system and transformed, i.e. the model learns from users' decisions on metadata, content and terminology mappings. Every time a user modifies the system recommendation or makes a new decision, the same is remembered for use in future recommendations with the given context of file/column/contents.

RECOMMENDATION ENGINE

Smart Transformation Process Flow


Data Models
Terminologies / Ontologies / Dictionaries
Curated Training sets
New Training sets from User Decisions
Machine Learning Models based Recommendation Engine
Mapping Recommendations
Terminology Recommendations
Source files
Data Model Transformation
Transformed Datasets
Terminology Normalization
Final Datasets
User Decisions: approve/update the Recommendations

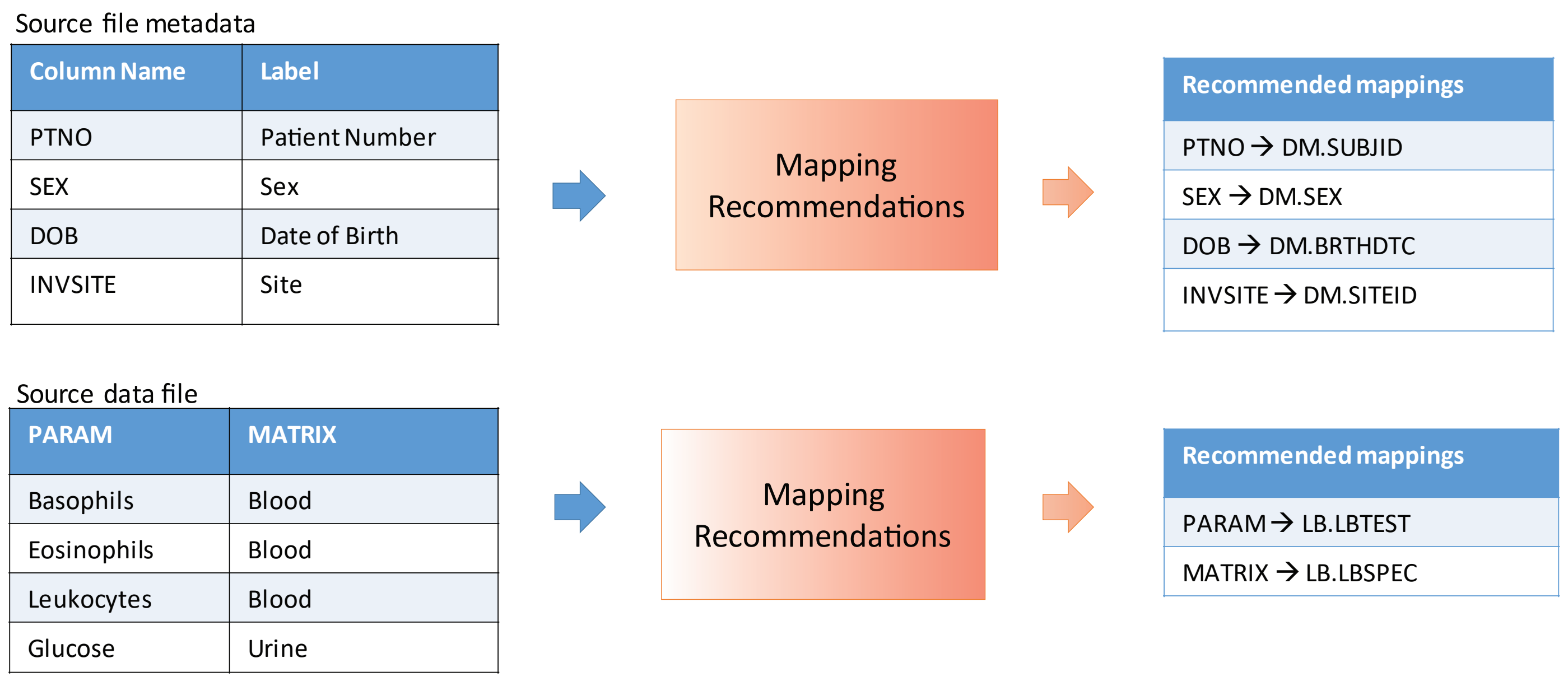
Data Models: Metadata Definition containing data domains, variables/columns names, labels, datatypes and references to terminologies/codelists.

Terminologies/Ontologies/Dictionaries: Prepared from standard CTs, public and subscription based databases / dictionaries (CDISC, MedDRA etc.) covering Preferred Terms, Synonyms, IDs

Curated Training sets: Prepared from the corpus of source files and corresponding standardized data based on each neural network's input and prediction needs.

Recommendation Engine consists of multiple Neural Network models created using TensorFlow™ (an open-source software library for dataflow programming developed by Google) and Keras™ library (an open source neural network library developed using Python at MIT). These models are executed in a hierarchy with higher level models trained to classify data domains and variable classes [1]. Classification results from each model are collected and summarized with a confidence score. Next set of neural network models are chosen from the hierarchy based on the shortlisted data domains and variable classes to further identify and recommend mappings to the target domain and variable.

Mapping Recommendation (Target Domain & Variables)



Source file metadata

Column Name	Label
PTNO	Patient Number
SEX	Sex
DOB	Date of Birth
INVSITE	Site

Mapping Recommendations

Recommended mappings
PTNO → DM.SUBJID
SEX → DM.SEX
DOB → DM.BRTHDTC
INVSITE → DM.SITEID

Source data file

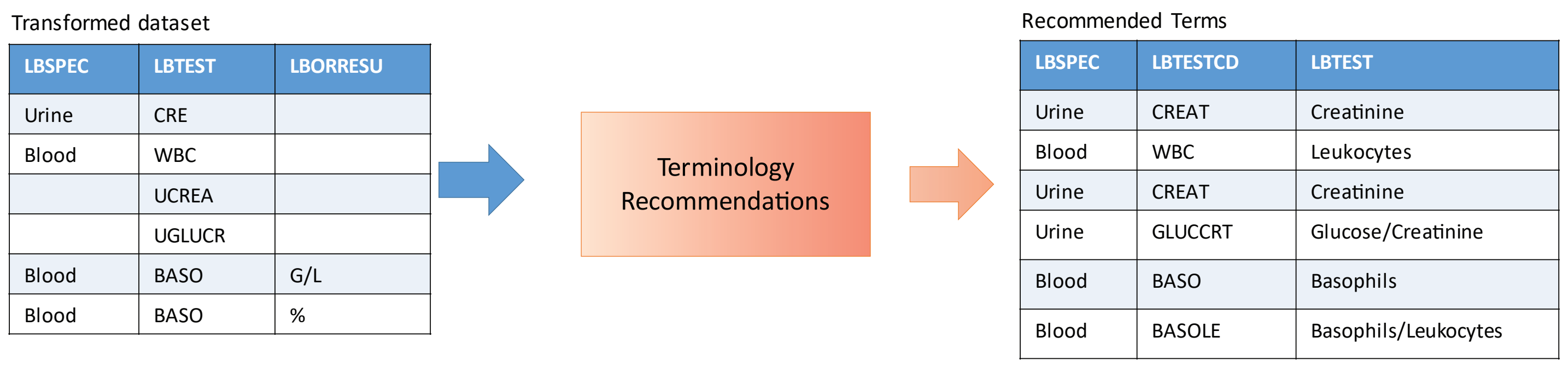
PARAM	MATRIX
Basophils	Blood
Eosinophils	Blood
Leukocytes	Blood
Glucose	Urine

Mapping Recommendations

Recommended mappings
PARAM → LB.LBTEST
MATRIX → LB.LBSPEC

Users review these mapping recommendations and can revise where required. A data transformation module accepts these mappings and transforms data to target data structure. Terminology recommendation algorithms process these transformed datasets and propose Preferred Terms from respective dictionaries or code lists based on the model definition. A user can review and approve terminology recommendations to create final standardized datasets.

Terminology Recommendation (Target Terms)



Transformed dataset

LBSPEC	LBTEST	LBORRESU
Urine	CRE	
Blood	WBC	
	UCREA	
	UGLUCR	
Blood	BASO	G/L
Blood	BASO	%

Terminology Recommendations

Recommended Terms

LBSPEC	LBTESTCD	LBTEST
Urine	CREAT	Creatinine
Blood	WBC	Leukocytes
Urine	CREAT	Creatinine
Urine	GLUCCRT	Glucose/Creatinine
Blood	BASO	Basophils
Blood	BASOLE	Basophils/Leukocytes

Every mapping transformation and terminology recommendation change made by the user is tracked. These user decisions are used to create new training datasets automatically from source file metadata and data.

Relevant neural network models are retrained considering the new training dataset, calibrated for accuracy using train and test split of data using Grid Search [2]. The retrained neural network model is versioned and saved for further use when the prediction accuracy is greater than the current version.

CONCLUSION

Variability in study datasets makes it difficult to perform analytics across different datasets. In this paper, we have described how neural networks based machine learning techniques can be used to build a recommendation engine to transform datasets from their original state to a target, consumable structure. The critical evaluation of the neural network model's performance and machine learning techniques greatly improves the data transformations process with cost and time savings.

REFERENCES

[1] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep Learning based Recommender System: A Survey and New Perspectives," arXiv [cs.IR], 24-Jul-2017.

[2] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," J. Mach. Learn. Res., vol. 13, no. Feb, pp. 281–305, 2012.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Isaac Mativo, PointCross Life Sciences Inc., isaac@pointcross.com