Paper 3825-2019

Make SAS Viya Realize Strategic Data Science
~Data Governance, AI platform and AI SAS programmer~

Ryo Kiguchi, Katsunari Hirano, Yoshitake Kitanishi;

SHIONOGI & CO., LTD.

ABSTRACT

Data scientist is always necessary to construct various models and to acquire the latest analysis method for various kinds of data in order to make use of maximizing “Data Science”. Fashionable AI is also the same. The Artificial Intelligence (AI) that we define is the system with the series of processes of “Recognition”, “Learning” and “Action”, which assists people's activities. There are various types of data used in AI, and so that, the models or methods used recognition, learning, and action are different depending on the data format. However, in talking about data science, data governance is very important regardless of the data format. We have made the strategy about data governance using Python and SAS via SAS Viya, and have been maximizing data science based on effective matching such as machine learning and deep learning (CNN, RNN etc.). As one example, we introduce "AI SAS programmer" system developed by our company which is semi-automatically creates SAS programs to analyze clinical data. This system is constructed from machine learning and deep learning etc. by selecting a programming language in Data Driven via SAS Viya, and this system was led to 33% reduction of standard work time in analysis work. Now, based on this data driven's data governance strategy, we are making the strategy to utilize data science for product innovation of new drug development, and we introduce a part of strategy on the day.

INTRODUCTION

In order to maximize data science, it is necessary to construct various models and to acquire analysis methods for various types of data. The same is true for Artificial Intelligence (hereinafter, called AI) which is currently in fashion. Before talking about AI, I would like to define our AI. Because, AI has begun to spread to the world in various scenes, but the definitions of AI are variously defined by researchers. The AI that we define is the system with the series of processes of “Recognition”, “Learning” and “Action”, which assists people's activities. Now, there were three booms of AI so far, and we summarize the main analytical approaches in each boom in table 1.

	Main analytical approaches
1st AI boom (1950s ~ 60s)	heuristic search, inference
2nd AI boom (1980s)	knowledge engineering
3rd AI boom (2010s)	Machine Learning, Deep Learning, Reinforcement Learning

Table1 Main analytical approaches in each boom of AI

As shown in Table 1, Machine Learning, Deep Learning and Reinforcement Learning serve as an engine of the current AI boom's. Among these technologies, Convolutional Neural Network (hereinafter called CNN), which is good at analyzing image data, or Recurrent Neural Networks (hereinafter, called RNN), which is good at analyzing time series data and text data, are one of the innovation technologies that brought about great impact on the world. The technologies of CNN and RNN are based on Neural Network (hereinafter, called NN). The history of NN is old, and NN was originally modeled by W. McCulloch and W. Pitts

in 1943 [1]. For example, CNN is good for image data analysis, and even in the healthcare industry, CNN is applied various scenes such as image diagnosis.

NEW POWER: CNN

Therefore, we considered using CNN, which is NEW POWER, to acquire AI which assists people's work, and to improve business. CNN can be implemented by SAS Viya. We described below the CNN program by SAS Studio interface and by Python interface. Depending on the data's characteristics, you should select which program is the best one, but we thought Python could be written the program more concise than SAS. So Using Python, we thought about the business improvement which will be described later.

```
/* Program1 CNN by SAS Studio interface */
proc cas ;
        loadactionset 'image';

        image.loadImages / casout={name='image_files', replace=1}
                path="/var/viya_data/image"
                recurse= TRUE decode=TRUE labellevels=-2;
        image.processImages / casout={name='resized_image', replace=1}
        imageTable={name='image_files'}
                imageFunctions={{functionOptions={functionType="RESIZE", height=224, width=224}}};
        shuffle / casout={name='resized_shuffled', replace=1}
        table={name='resized_image'};
        deepLearn.buildModel /
                modelTable={name="SAS_TEST", replace=TRUE} type="CNN";
        deepLearn.addLayer /
                layer={type="INPUT" nchannels=3 width=224 height=224}
                modelTable={name="SAS_TEST"} name="data";
        deepLearn.addLayer /
                layer={type="CONVO" nFilters=8 width=7 height=7}
                modelTable={name="SAS_TEST"} name="conv1" srcLayers={"data"};
        deepLearn.addLayer /
                layer={type="POOL" width=2 height=2 }
                modelTable={name="SAS_TEST"} name="pool1" srcLayers={"conv1"};
        deepLearn.addLayer /
                layer={type="CONVO" nFilters=8 width=7 height=7}
                modelTable={name="SAS_TEST"} name="conv2" srcLayers={"pool1"};
        deepLearn.addLayer /
                layer={type="POOL" width=2 height=2 }
                modelTable={name="SAS_TEST"} name="pool2" srcLayers={"conv2"};
        deepLearn.addLayer /
                layer={type="OUTPUT" act='softmax' n=2}
                modelTable={name="SAS_TEST"} name="outlayer" srcLayers={"pool2"};
        deepLearn.modelInfo / modelTable={name="SAS_TEST"};
        deepLearn.dlTrain /
                inputs={{name="_image_"}}
                modelTable={name="SAS_TEST"}
                modelWeights={name="SAS_TEST_W" replace=TRUE}
                table={name="resized_shuffled"}
                optimizer={ maxepochs=1 miniBatchSize=128};
quit;

/* Program2 CNN by Python interface */
%matplotlib inline

import swat

sess = swat.CAS("sasviya-address", 1234, "user-id", "password")

from dlpy.images import ImageTable

img_path='/var/viya_data/image'

my_images = ImageTable.load_files(sess, path=img_path)
my_images.resize(width=224)

from dlpy.splitting import two_way_split
```

```
tr_img, te_img = two_way_split(my_images, test_rate=20, seed=123)
tr_img.as_patches(width=200, height=200, step_size=24, output_width=224, output_height=224)

from dlpy import Model, Sequential
from dlpy.layers import *
from dlpy.applications import *

model1 = Sequential(sess, model_table='CNN sample program')
model1.add(InputLayer(3, 224, 224, offsets=tr_img.channel_means))
model1.add(Conv2d(8, 7))
model1.add(Pooling(2))
model1.add(Conv2d(8, 7))
model1.add(Pooling(2))
model1.add(Dense(16))
model1.add(OutputLayer(act='softmax', n=2))
```

There is a drawback in utilizing CNN. It is the point which it is difficult for us to interpret the analysis result. In other words, it is unknown what decision criteria the analysis results were obtained from, and the decision process is black boxed. Recently, there are various approaches to elucidate the black box, and SAS Viya made it possible to interpret the results by visualizing CNN 's decision criteria information with heat map. This approach made it easier to figure out where the machine is looking at where in the graphics. In the heat map, it is possible to check the places affecting judgement based on density of blue, green, and red. This analysis can be implemented using the heat_map_analysis () method of the DLPy package.

```
/* Program3 CNN by Python interface */
model1.heat_map_analysis(data=test_imge, mask_width=56, mask_height=56, step_size=9)
```

For example, based on the model for judging ladybugs and cats constructed from ladybug image (1597 pictures) and cat images (792 pictures), confirm in Figure1 that how the machine judged the ladybug when analyzed the ladybug image. The left image is the original image, the right image is the heat map, and the center image is the superimposed image. When confirm this result, you can see that the machine is judging whether it is a ladybug or a cat, paying attention to information around the ladybug. As you can see, it is relatively easy to implement CNN with the Python interface of SAS Viya, and the CNN's program is concise, short and readable. In addition to that, it is attractive to be able to interpret CNN.

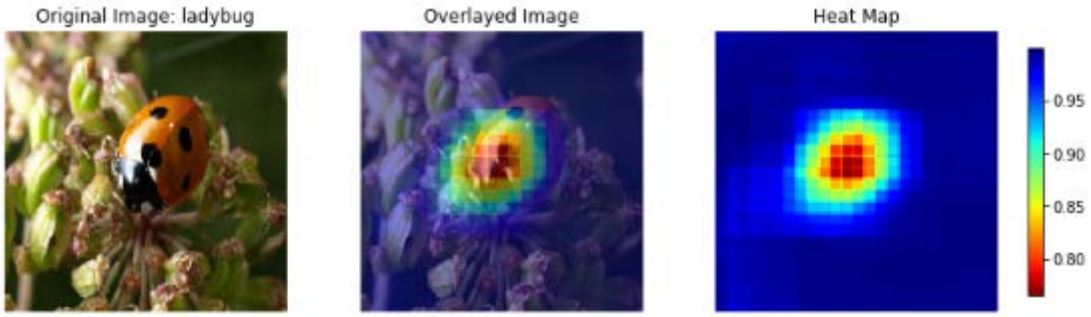

Figure1 CNN's analysis result of ladybug (Heat map)

Then, we introduce specific examples of Business Process Reengineering. What do you imagine when you hear image analysis? Many people imagine the analysis using photographs. Actually, in the healthcare industry, there are many analysis examples in MRI images, eye fundus photographs and oncology pathology images. However, all data can be made into images. For example, if you take a screenshot of text and convert it to .jpg

(or .jpeg etc.) file, it becomes image data. What we focused on this time is the "TLF shells" that we use in clinical trials. In clinical trials, we make a protocol and also make a Statistical Analysis Plan (SAP) which describes the detailed analysis method for clinical trial data. After that, we make TLF shells that summarizes the specific output results' image (Tables, Figures, Listings) as shown Figure2. In this TLF shells, usually, several hundred images of demographic, efficacy and safety analysis are described. Based on the TLF shells, we make the analysis programs using the SAS software, and we obtain the analysis results by running the programs. The CNN analysis case introduced here is the case where this TLF shells was converted into image data and classified into 5 classes according to the characteristics of image by CNN. The model summary is shown in Figure3, and the analysis result of TLF shells by heat map is shown in Figure4. This model is very simple, but accuracy rate is about 65%. Based on this result, we could discover new findings by considering where the machine watched and classified. This technology is configured as part of "AI SAS Programmer" developed by our company. AI SAS Programmer is the semi-automatic generation system for analysis SAS program based on SAP and TLF shells etc. This system was led to 33% reduction of standard work time in analysis work, and we are utilizing CNN as a help for large process innovation.

Table 14.1.1.1 Patient Disposition
All Randomized Patients

		S-****** N=** n (%)	Placebo N=** n (%)
Completed		** (**.*)	** (**.*)
Withdrawn		** (**.*)	** (**.*)
Reason for withdrawal	ex. Ineligibility	** (**.*)	** (**.*)
	ex. Adverse event	** (**.*)	** (**.*)
	ex. Lack of efficacy	** (**.*)	** (**.*)
	ex. Recovery	** (**.*)	** (**.*)
	ex. Withdrawal by subject	** (**.*)	** (**.*)
	ex .Lost to follow-up	** (**.*)	** (**.*)
	ex. Other	** (**.*)	** (**.*)

Figure2 an example of TLF shells

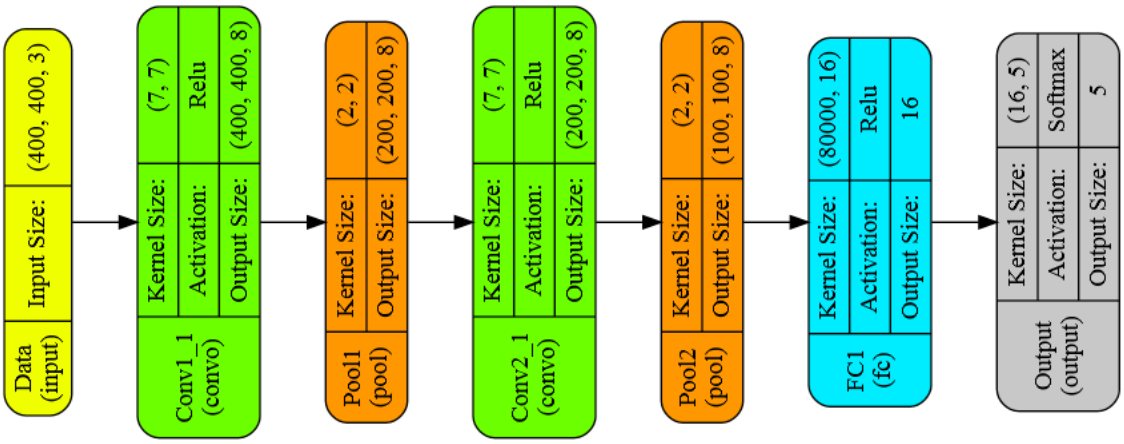

Figure3 analysis model summary of TLF shells

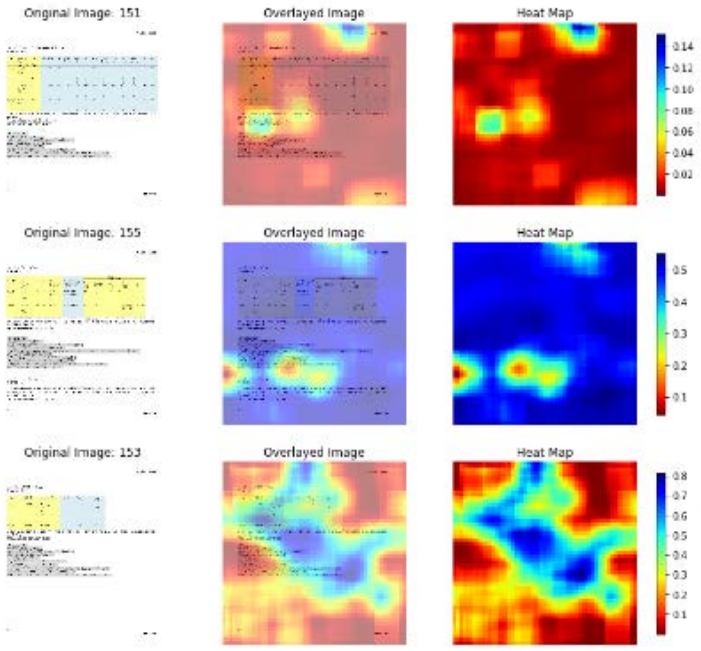

Figure4 analysis result of TLF shells (Heat map)

IMPORTANCE OF DATA GOVERNANCE IN USING AI

There are many technologies used in "AI SAS Programmer" system developed by our company, introduced in the previous chapter. In order to construct this system, in addition to Data Science Group technology until now, we have acquired several technologies from outside companies, and completed the system by combining optimal knowledge(technology) among them. Table 2 shows a part of technologies our Data Science group currently possesses.

Knowledge of Data Science Group	Database / Database technology	FAERS (FDA Adverse Event Reporting System)	JADER (Japanese Adverse Drug Event Report database)	Clinical trial data
		Real World Data	SQL/NoSQL	Hadoop
	Analysis software/ Programming language	SAS	SAS Viya	Python
		Lua	R	Fortran
		QGIS	Spotfire	
	Analysis method / Analysis technique	Machine Learning	Deep Learning	Neural Network (NN)
		Recurrent Neural Networks (RNN)	Convolutional Neural Network (CNN)	Clustering methods
		Random Forest	Regression methods	Text Mining
		Data Mining	Support Vector Machine (SVM)	Multivariate Analysis
		Causal Inference	Propensity Score Matching	Optimization
		Association Analysis	Time Series Analysis	Bayesian theory

		Spatial Statistics	Inference Analysis	Normalization Technique
		Anonymization	Large-Scale Simulation	Data Visualization
Knowledge from outside companies	Database	Dictionary Database for Text Mining		
	Analysis method	Reinforcement Learning	Code Analysis	

Table2 a part of technologies our Data Science group currently possesses

The reason for constructing the system by combining multiple knowledge is the feature of the data used in AI. From the viewpoint of data structure, data used in AI is roughly divided into structured data and unstructured data. And from the viewpoint of data types, there are various data such as values, text, sound, image, and the like. Incidentally, "AI SAS Programmer" system uses value, text and image. On the other hand, even in one field of machine learning, there are various analysis methods (Ex. Generalized linear regression, Decision tree, Convolutional Neural Network, Naive Bayes, AR model and Clustering etc.), and it is necessary to select an appropriate analysis method according to the data format. Therefore, as shown introduction, it is self-evident that it is necessary to select an analysis method depending on the data format even in a series of processes of recognition, learning and action in AI.

However, data governance is important regardless of the data format used for AI. Because we think that data governance realizes the smooth process of recognition, learning and action, and constructs the optimum AI Platform. The outline of our data governance strategy is shown in Figure5. Ordinarily, there are multiple in-house systems in a company, and various databases are utilized. So many data pour into from inside and outside company. The state which the data is simply accumulated and stored in a database is called “Data Lake” in this paper. However, it is inefficient to simply store the data in Data Lake. The level we aim is to store the internal and external data with “high quality”. We mean here that “high quality” is the state which databases are linked to each other, and data extraction, processing and analysis can be performed smoothly from databases. This state is called “Data Reservoir”. This Data Reservoir should be quality controlled as far as possible on the upstream. Because management cost is less in the upstream than in downstream. We think that ideal data governance is to efficiently manage data and to make using of data smoothly according to purpose.

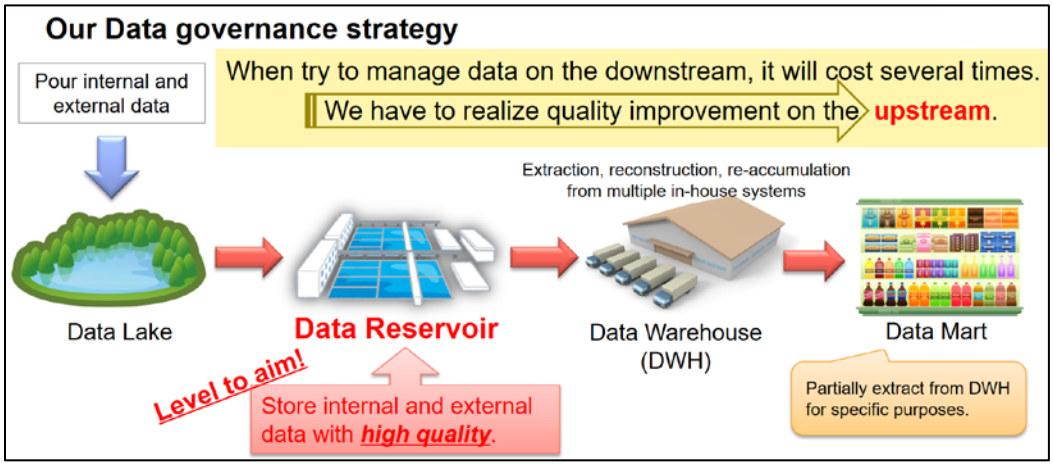

Figure5 outline of our data governance strategy

In order to realize this ideal data governance, we utilized SAS Viya. The advantage of SAS Viya is that we can centralize data management with single platform called CAS and strengthen governance. This makes it possible to combine multiple Data Reservoirs into one. Besides that, by being able to select programming language with data driven, it became possible to put the data in Data Reservoir without depending on data format. So, we can also include New Power's CNN in this platform. Figure 6 shows AI platform constructed using SAS Viya as foundation tool of data governance. Building a strategically data driven AI platform using SAS Viya led to the AI technology acquisition. At the same time, as a result, it became clear that the key to our strategic Data Science activities is "data driven".

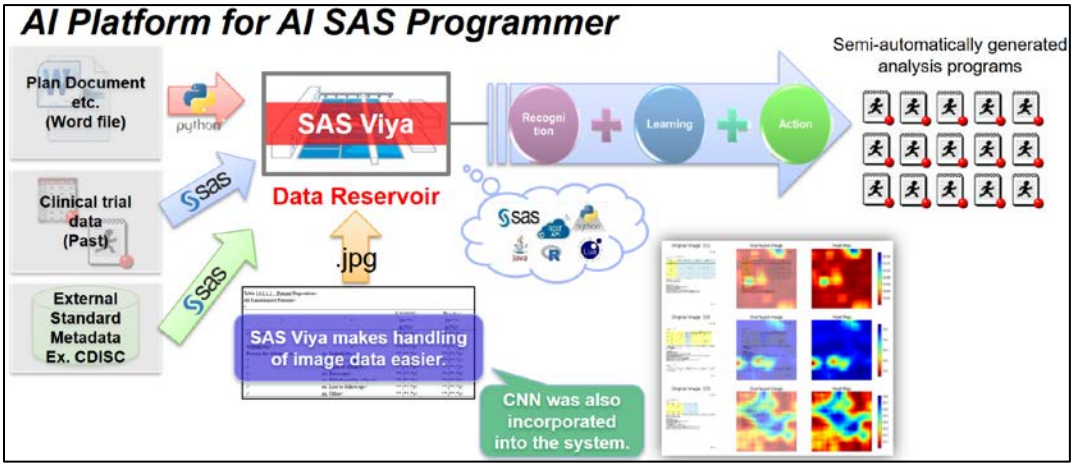

Figure6 AI platform constructed using SAS Viya as foundation tool of data governance

CONCLUSION

The key word of strategic data science is "data driven". In data science, it is important to effectively combine knowledge. In data driven data science activities, it is essential to properly utilize from basic analysis method to latest analysis method based on data. In this paper, as an example of data driven data science activity, we introduced a case of data driven business improvement using SAS Viya, that is, case of process innovation (AI SAS programmer). By selecting analysis methods and programming languages according to the type of data, we could construct an efficient and effective system. Nonetheless, the strategy of product innovation in drug development remains unchanged.

However, there are two major innovations in pharmaceutical companies - process innovation and product innovation, but the latter product innovation is more important. And yet, the strategy of product innovation in drug development remains unchanged. By maintaining internal and external data in Data Reservoir and analyzing the data, we formulate new hypotheses. And by proving the hypothesis, we will maximize product value and achieve product innovation. we show one idea of product innovation. In clinical trials until now, image data, voice data and stream data like biosensor data are difficult to manage and had not been used so much. But it is possible to analyze the data by selecting programming language in data-driven based on data governance independent of data format. Patients' facial expressions and voices, life logs, medical images etc. are valuable data with objectivity for us. Because it makes it possible to perform objective evaluation based on the objective data, especially for diseases which could only be evaluated subjectively so far. The benefits of objective evaluation are immeasurable. For example, due to objectivity, homogenization of medical care will be promoted by reducing the variance of evaluation among evaluators such as doctors. And it will also be easy to imagine that

objectivity is useful for Health Technology Assessment (HTA). Furthermore, objectivity will also help patients themselves to understand their illnesses correctly.

The case introduced in this paper is only the first step in strategic data science. But, it sits on the doorstep that strategic data science causes product innovation.

REFERENCES

[1] McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 7:115 - 133.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at the following addresses:

Ryo Kiguchi

Shionogi & Co., Ltd.

ryo.kiguchi@shionogi.co.jp

Katsunari Hirano

Shionogi & Co., Ltd.

katsunari.hirano@shionogi.co.jp

Yoshitake Kitanishi

Shionogi & Co., Ltd.

yoshitake.kitanishi@shionogi.co.jp

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.