GTC 2018

S8371 - How We Can Analyze Profile from Real-Time Conversation by Unsupervised Learning

03/28/2017

dAIgnosis,Inc.

dAIgnosis,INC.

GTC 2018

COMPANY PROFILE

Circumstances

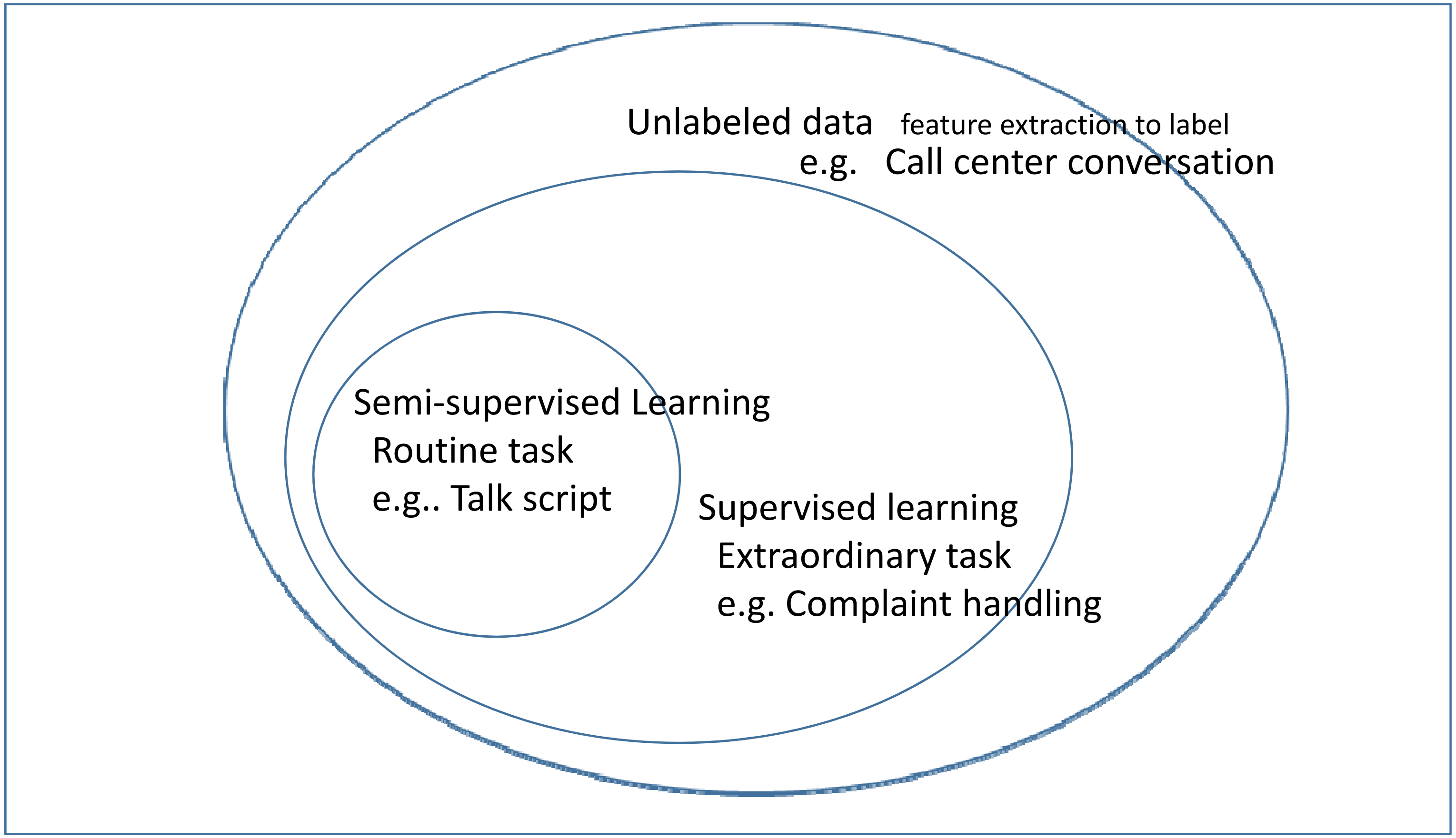
- Design / development engineers who dedicated to Computing services gathered.
- Started research on AI technology based on medical system development in a national project
- Established the company May 2017 with the theme of deep using GPU.
  VP of Google head office joined as a director.
- Advance technology development to build the original models while stu multiple cloud platforms.
- Started research using NVIDIA DGX-1 *7 +1 units (Volta in April 2018) from affiliates.
- Planned to start real-time analysis of text combined with image,etc. from the beginning of 2018.

OWNED TECHNOLOGY

Highly Unique Technology

- Development of Booster Pack for building TensorFlow based on DGX-1
- Medical diagnosis support by combined processing of text analysis and image recognition
- Model optimization of business flow from business system program and model to speed up business processing with GPU

Conceptual diagram



GTC 2018

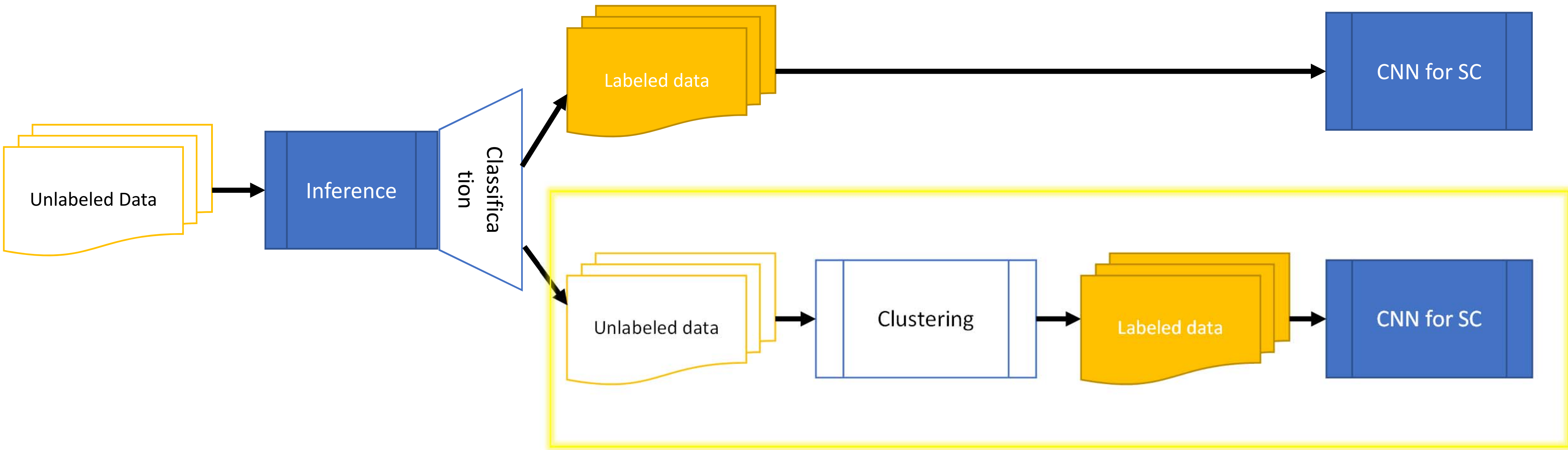
Responding to issues of speech recognition through phoneme-text conversion system

- Adaptation to business systems of machine learning
- Machine learning in Japanese(End to End)
- Business fitting for clustering
- Efficient data collection
- Improvement of fault tolerance on DGX-1

GTC 2018

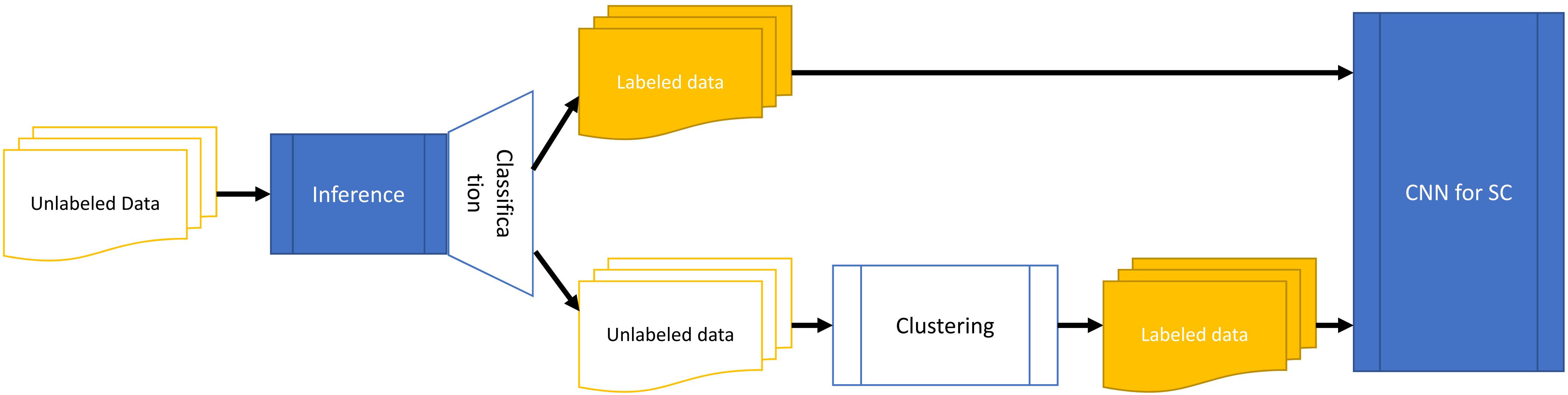
Data Flow for CNN for SC(1st trail)

SC sentence classification



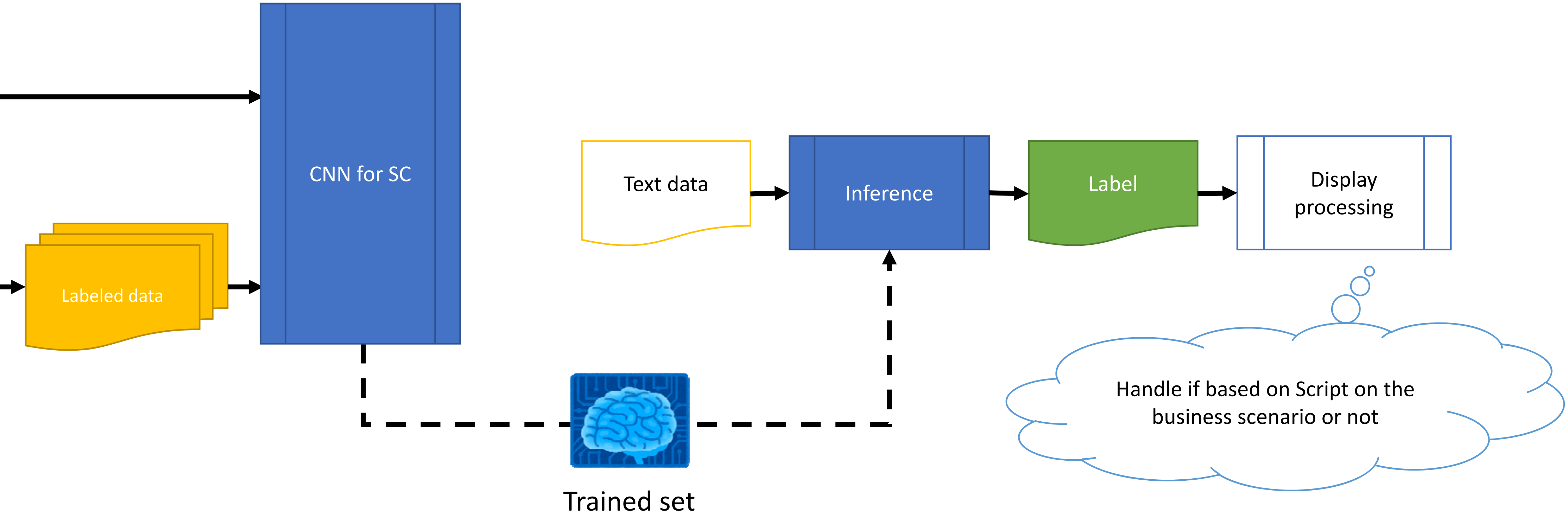
dAIgnosis,INC.

Data Flow for CNN for SC(2nd trail)



GTC 2018

Following Data Flow for CNN for SC



dAIgnosis,INC.

Demonstration data

- We learned the conversation that is answering the question out of 6000 data of the telephone correspondence conversation.

  Using conversation data on the telephone reception of the hotel

- In order to show the change in the amount of data to be learned, inference is made in a two-pattern model with a learning amount of 1,700 cases / 800 cases.

Demonstration (Learning Phase)

Labeling learning data with unsupervised learning by clustering.

Is the room available on dd/mm?	The next room is noisy
What time is check-in?	Is breakfast served?
Can I make a reservation on dd/mm?	What time can I check in?
Do you have breakfast?	Room xx is noisy though

Demonstration Overview (Learning Phase) 1

1. Labeling learning data by unsupervised learning (k-means method etc.) and clustering.

Is the room available on dd/mm?
What time is check-in?

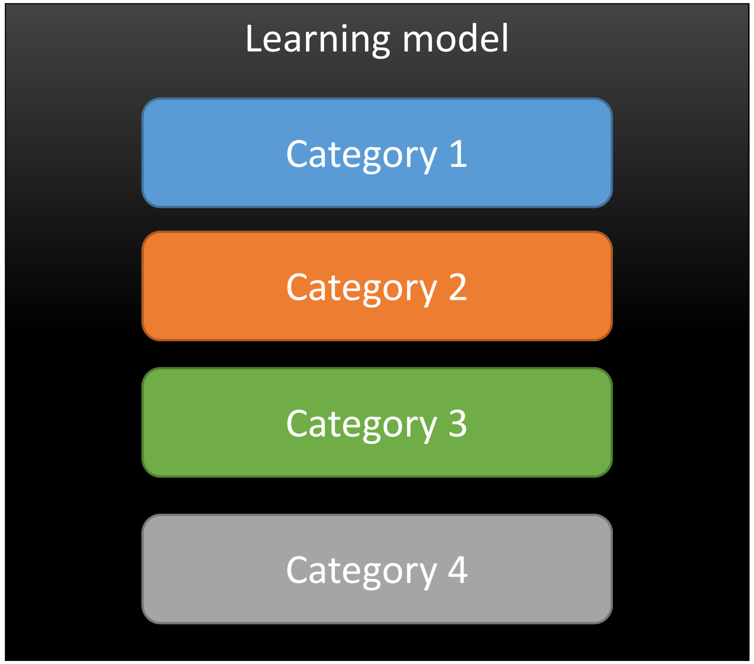
The next room is noisy
Is breakfast served?

Can I make a reservation on dd/mm?
Do you have breakfast?

What time can I check in?
Room xx is noisy though

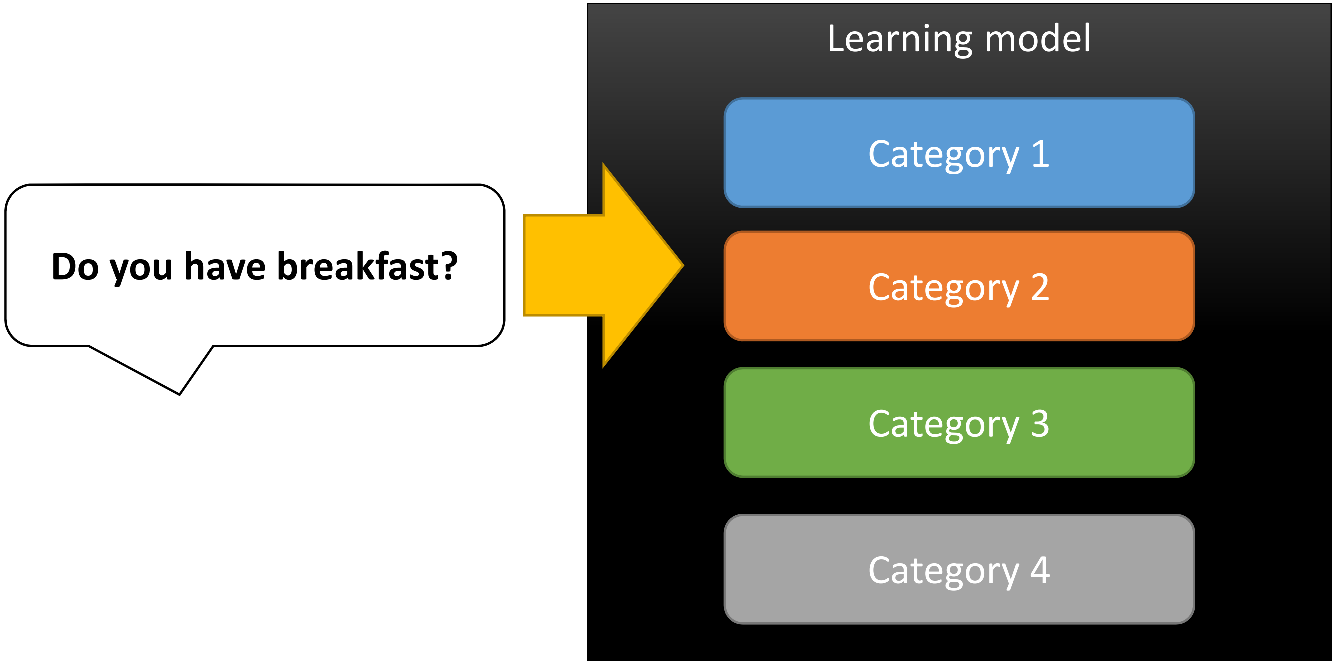
Demonstration Overview (Learning Phase) 2

2. A learning model is created by performing supervised learning with categories clustered by 1 as labels.



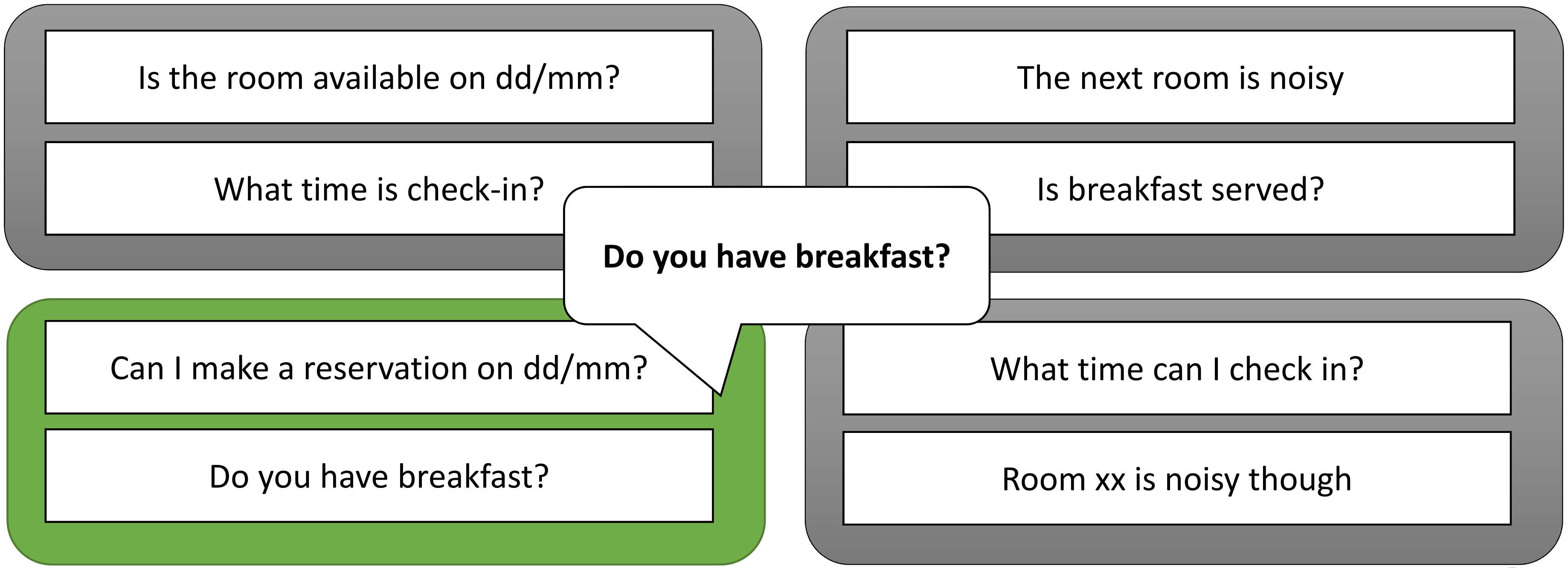
Overview of demo (inference phase)

Using a learning model, infer which category a message entered will be.



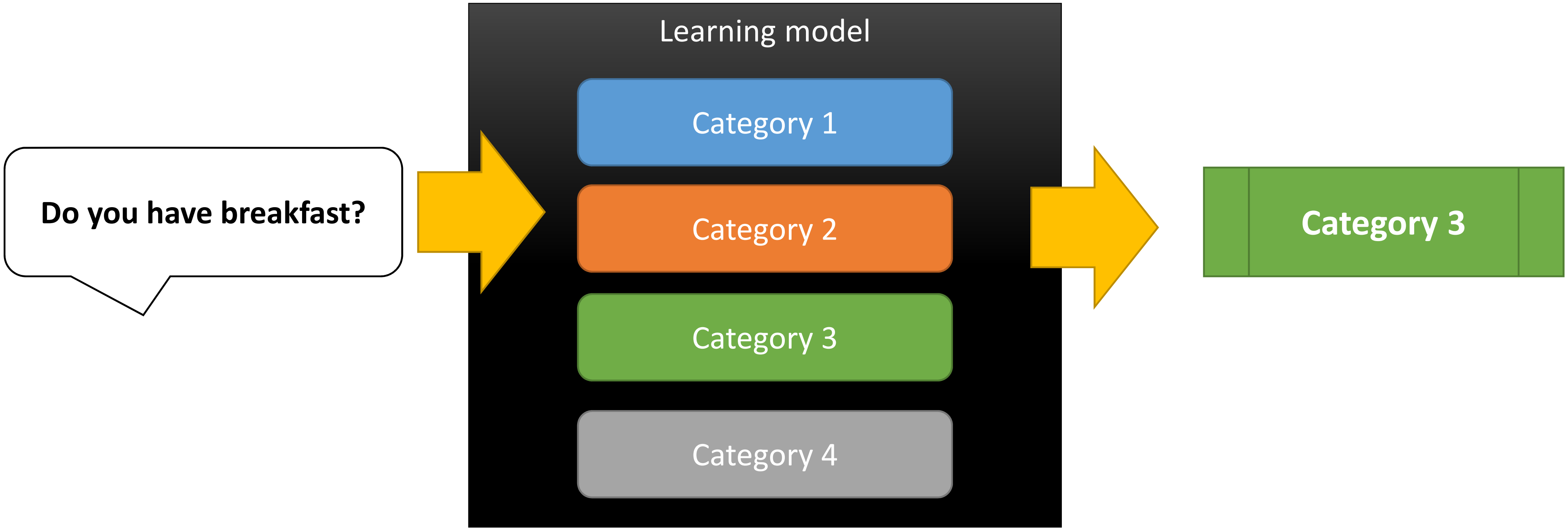
Overview of demo (inference phase) 1

1.Using a learning model, infer which category a message entered will be.



Overview of demo (inference phase) 2

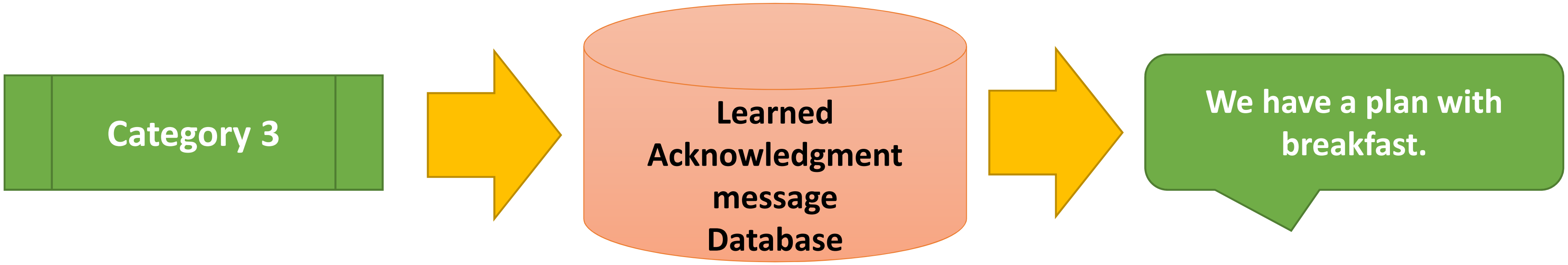
2. Use a learning model to infer which category the message entered will be.



GTC 2018

Overview of demo (inference phase) 3

3. Display messages tied to categories inferred by the learning model



dAIgnosis,INC.

Adaptation of business systems of machine learning

- When building business systems in Japan, object oriented languages such as java, C # etc. are preferred. Because object - oriented languages are preferred, inevitably there are many engineers in object - oriented languages such as java, C # in Japan.

- On the other hand, in the field of machine learning, python is overwhelmingly popular. There are also python engineers in Japan, but it is difficult to acquire as numbers enough as human resources. In consideration of current situation, we made the learning part of machine python and the inference part Java.

- By setting the learning part to python, it is possible to investigate / validate the new model as soon as possible. By setting the reasoning part to java, it becomes possible to build business applications with a familiar language, so that engineers can concentrate on the learning part more.

Differences of phoneme between in English and Japanese

English（20 vowels＋24 consonant＝44 phoneme）:
/iː/, /ɪ/, /e/, /æ/, /ʌ/, /ɑː/, /ɒ/, /ɔː/, /ʊ/, /uː/, /ɜː/, /ə/, /eɪ/, /aɪ/, /ɔɪ/, /əʊ/, /aʊ, ɑʊ/, /ɪə/, /eə/, /ʊə/; /p/, /b/, /t/, /d/, /k/, /g/, /ʧ/, /ʤ/, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/, /h/, /m/, /n/, /ŋ/, /l/, /r/, /w/, /j/

Japanese（5 vowels＋16 consonants＋3 peculiars＝24phoneme）:
/a/, /i/, /u/, /e/, /o/; /j/, /w/; /k/, /s/, /c/, /t/, /n/, /h/, /m/, /r/, /g/, /ŋ/, /z/, /d/, /b/, /p/; /N/, /T/, /R/

Reference: http://user.keio.ac.jp/~rhotta/hellog/2012-02-12-1.html

Machine learning in Japanese

GTC 2018

- Language features

Unlike languages with spaces between words like Japanese, Japanese has a structure in which Hiragana “あめりか”, Katakana “アメリカ” ,and Chiese character:Kanji“亜米利加” are arranged equally to the same characters at a time.

From the viewpoint of diversity of linguistic expression, there are places depending on the granularity of the language, but in the case of Japanese, the notation also occurs. (ex. apple, apples, Apple) and the subject and the object are omitted, and the predicate comes to the end of the sentence.

- Due to the above characteristics, we devised the way of machine learning Japanese as compared with English and others.

As in the English-speaking style of Japanese notation method, “Machine learning” is carried out after "spacing" which puts a space between word and word. It is possible to carry out machine learning more efficiently by applying "separating".

- It will be touched on from the future perspective.

Business fitting for clustering

- Due to the characteristics of clustering, select data similar. As you know, clusters of selected data do not necessarily become divisions according to business.

- In order to solve this problem, semi supervised learning is used. By supervised learning to be created at the beginning, by improving classification according to work in advance, we can improve learning model suitable for work.

- Also, for data that is not subject to learning by semi-supervised learning, there is a high possibility that it is data deviating from fixed form in the first place, so automatic clustering is performed using clustering.

- Using semi supervised learning and clustering, we use it as a flow to make effective use, not to discard data.

Data handling at character level

- In Japanese, documents are generally not languages expressed in a form divided for each word by "division". In the present situation, we divide into words using morphological analysis (Kaomoji). In the preprocessing, as we do `` ingenuity '', we can not deny the possibility that the precision of `` ingenuity '' affects the learning model of this process.
- As a future prospect, we will examine the method of advancing machine learning without "separating". (Non separation model)
- Machine learning considering the character level and the following are available, but it is premised that words are recognized with a space delimiter. In addition, since the number of representations of characters is limited (ASCII only), lots of ingenuity is required.
- Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts
- Character-level Convolutional Networks for Text Classification

Efficient data collection

- In machine learning, it is a problem to prepare a large amount of data collection, especially training data (evaluation data) in which input data and correct label are paired.
- However, in order to give direction to the learning result, the correct label is indispensable.
- In business systems, we try to generate data in which input data and correct labels are paired based on user's operation.
- Ex. Presentation of answer alternates to inquiry contents
- Generate additional training data from the inferred inquiry contents and the answer selected from the presented answer alternates

Emotional grasp from documents

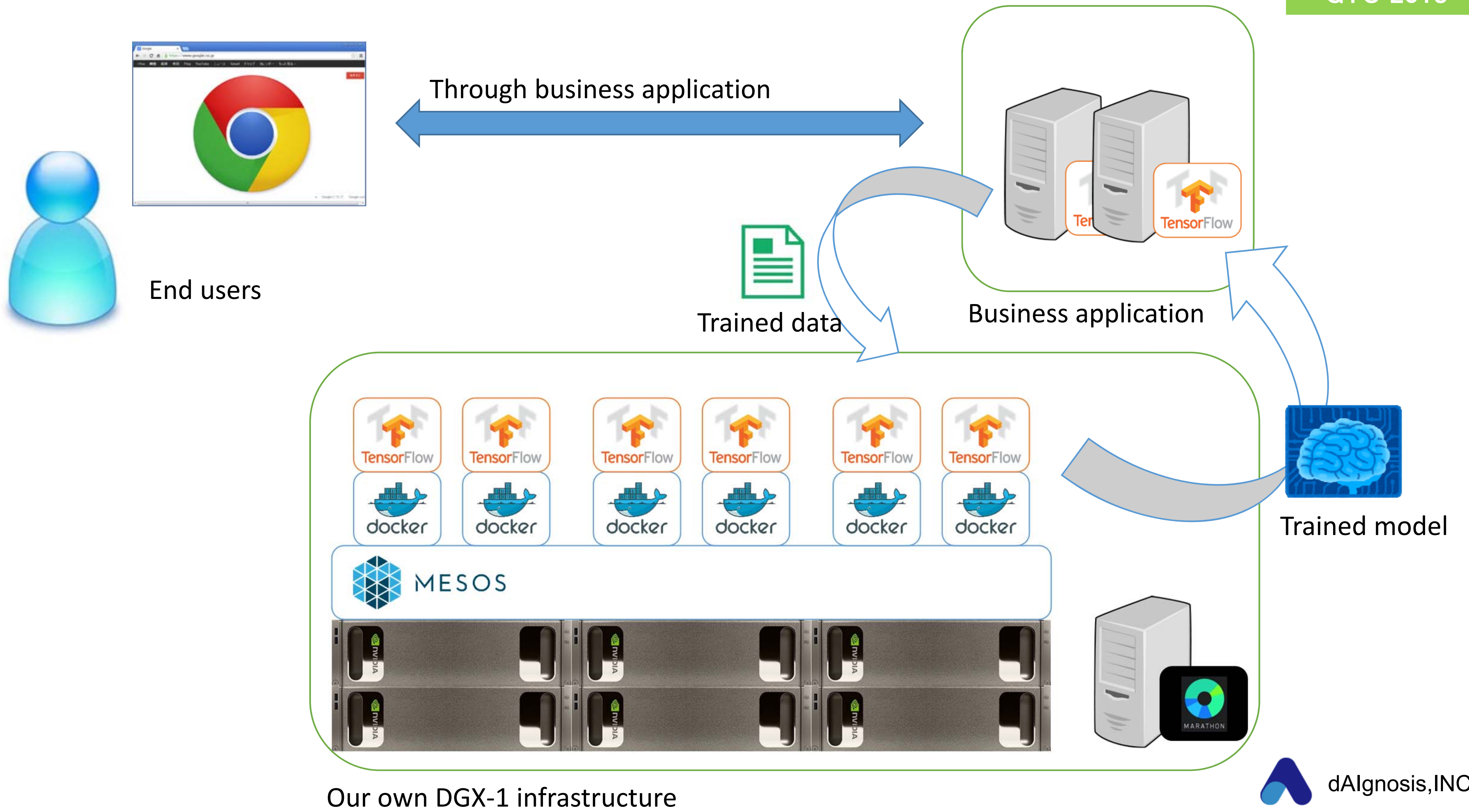
- In the next version of the third party voice to text solution, it will be possible to link traditional keyword hook type emotional information.

- A certain range of emotional information can be extracted by not only voice to text information but also emotional information. Through machine learning based on the text information and the obtained emotional information, we will demonstrate more emotional grasp from the entire text information as well as from keywords.

Improvement of fault tolerance on DGX-1

- The DGX - 1 is a high - performance enclosure. Since it is a physical one, to utilize and operate it for a production environment, it is necessary to improve fault tolerance by ourselves.

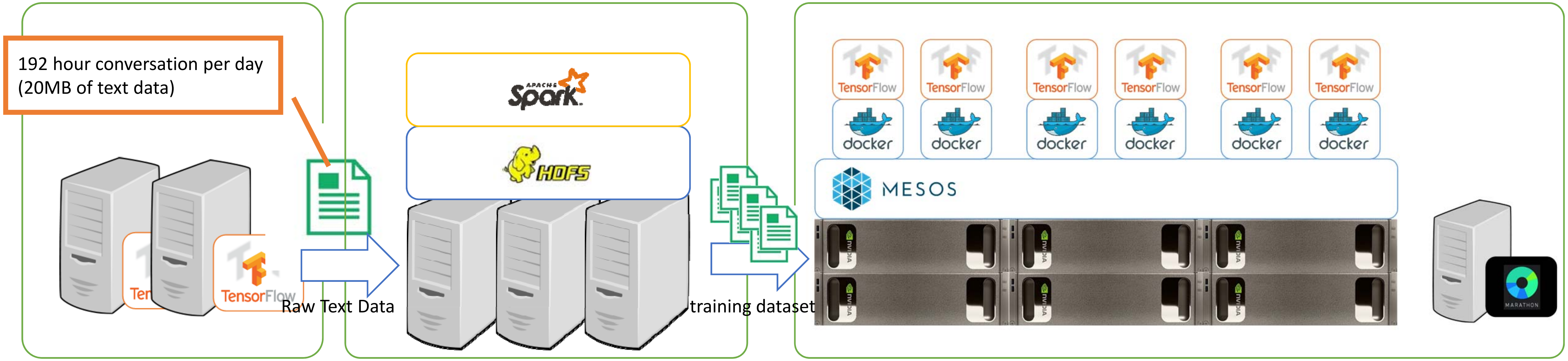
- We cametto bundle DGX - 1 and applied Mesos to treat it as a high - performance resource pool. By using Docker container via Mesos, we could realize abstraction of difference between development environment and production one. And for fault tolerance, now we can minimize downtime of Docker container by adopting framework around marathon.


Through business application
End users
Trained data
Business application
TensorFlow
docker
MESOS
MARATHON
Trained model
Our own DGX-1 infrastructure

For Deep Leaning, Execution Time In The Batch Processing Flow

GTC 2018

25 minutes	5 minutes	150 minutes by 8 GPUs (Training on 50 Epochs)
Raw Data Export 〜 File Transfer	Generating the training dataset	Deep learning



Business Application | Distributed processing | Deep Learning processing

Note: Execution time of prediction on the machine powered by only CPU.
- 220 ms (the training of SCDV using 3,000 dimensions, then)
- 430 ms (the training of SCDV using 6,000 dimensions, then)

dAIgnosis,INC.

Unsupervised clustering (X-means method)

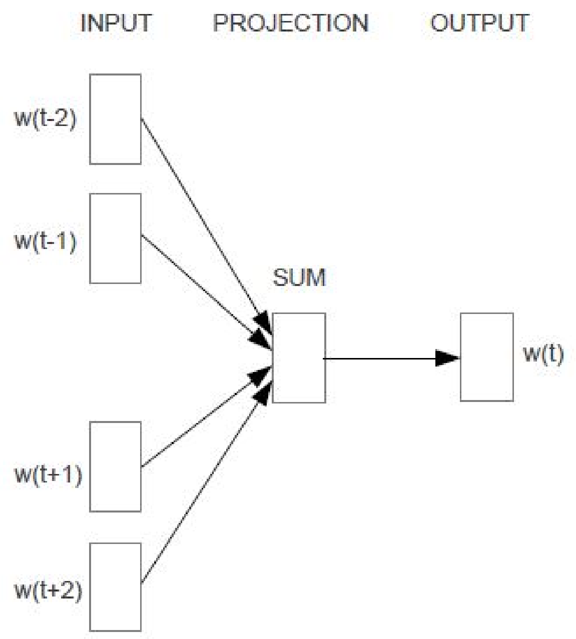
- In the case of the K-means method, it is necessary to give the number of clusters as an initial value, but in the case of the X-means method, the number of clusters is automatically estimated.

- Clustering problem
- In division of the cluster, since the division method becomes uncertain, it does not necessarily divide it suited for business when applying business.
- In order to solve this problem we introduce semi-teacher learning and respond.

Word2Vec

- It is a basic technique in the field of machine learning of NLP as an efficient learning method of word vector

- Learn the relation of surrounding words to a certain word

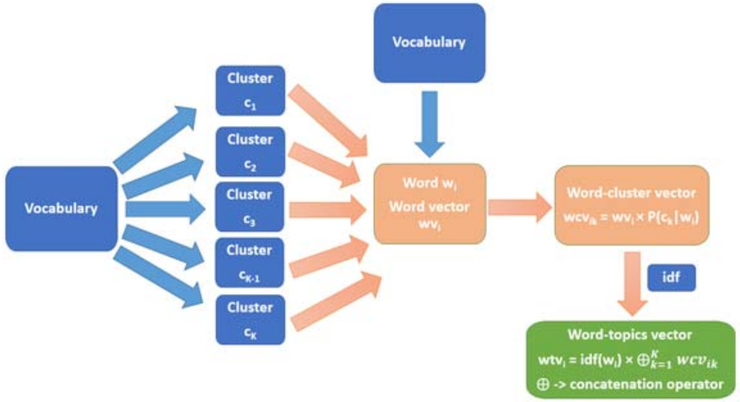


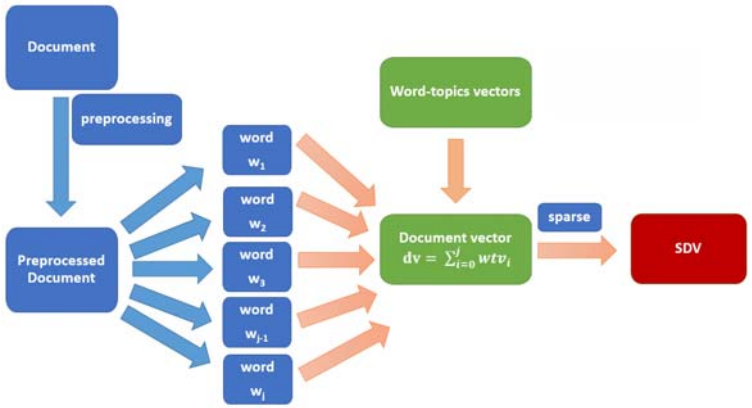
CBOW

Reference: https://arxiv.org/pdf/1301.3781.pdf

SCDV

- A method of improving the vector representation of a document by considering both the clustering result and the probability distribution in the vector space of Word2Vec



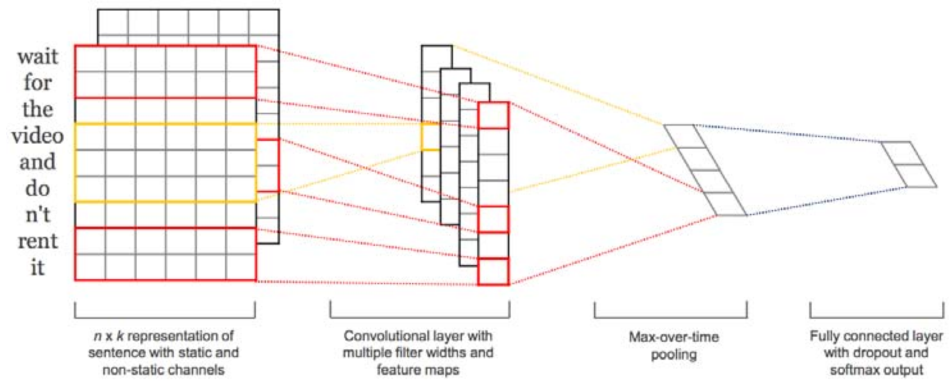


Reference: https://arxiv.org/pdf/1612.06778.pdf

CNN for Sentence Classification

(Convolutional Neural Networks for Sentence Classification)

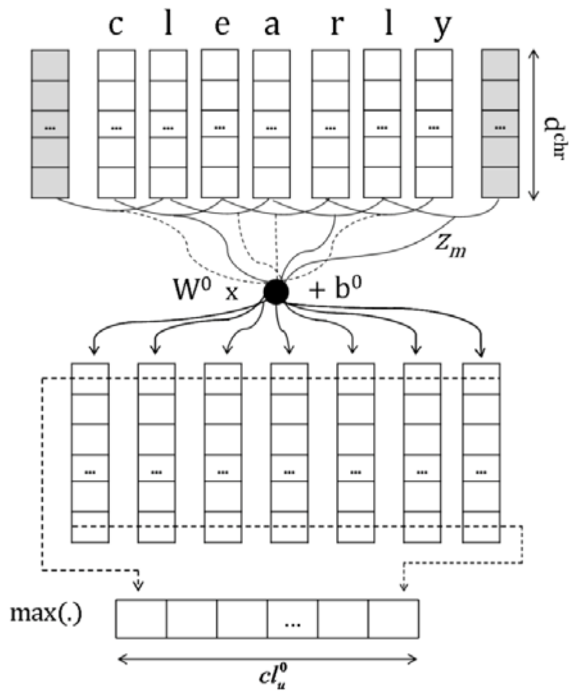
- A method of expressing and classifying a document by expressing the document as a word vector string and using CNN for the vector sequence



Reference: https://arxiv.org/pdf/1408.5882.pdf

Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

- In order to deal with the problem that feature extraction is difficult for short documents because of limited context information, not only vector expression of word level which is usually used but also vector expression of character level is constructed to obtain vector representation of document Technique to improve performance by doing

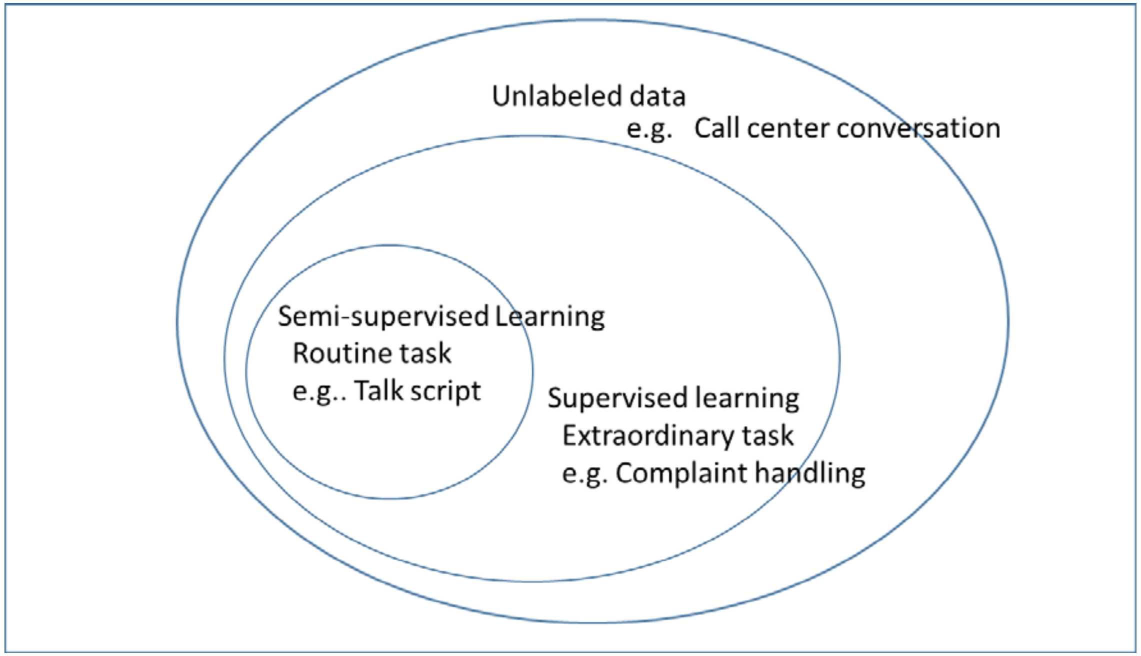


Reference: http://www.aclweb.org/anthology/C14-1008

Semi supervised learning (Self-training method)

- Semi-supervised learning is a learning method in which inference is made on a learned model learned by supervised learning, unsupervised data is learned by generating a hypothetical label (teacher) on the condition of accuracy etc.

- By introducing semi-supervised learning, it becomes possible to make enforcement to the event that clustering division method becomes indefinite.



Technical References.

Word2Vec	Efficient Estimation of Word Representations in Vector Space	Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean https://arxiv.org/pdf/1301.3781.pdf
	word2vec Parameter Learning Explained	Xin Rong https://arxiv.org/pdf/1411.2738.pdf
SCDV	Sparse Composite Document Vectors using soft clustering over distributional representations	Dheeraj Mekala, Vivek Gupta, Bhargavi Paranjape, Harish Karnick https://arxiv.org/pdf/1612.06778.pdf https://dheeraj7596.github.io/SDV/
CNN for Sentence Classification	Convolutional Neural Networks for Sentence Classification	Yoon Kim https://arxiv.org/pdf/1408.5882.pdf
Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts	Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts	Cicero dos Santos, Maira Gatti http://www.aclweb.org/anthology/C14-1008

FUTURE PLAN

Automated call center

- Agreement with a large call center for 1,400,000 membership of home delivery service to develop Semi-Automated call center.

- Based on widely used Phoneme recognition system, Semi-Automated call center can be a good show case to shorten the education period usually longer than half a year and complement veteran dialogue skills to newcomers by deep learning.

- Having collected more conversation data will enable us of Automated call center sooner by DGX-1.

GTC 2018

http://www.daignosis.com
omatsu@daignosis.com

Thank you.

dAIgnosis,INC.