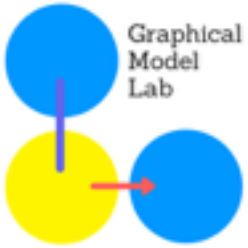


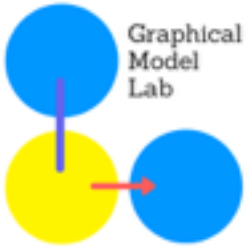
Graphical Model Lab: Towards the Development of Graphical Modelling Open Source SaaS

Mao Ito

Mao Ito

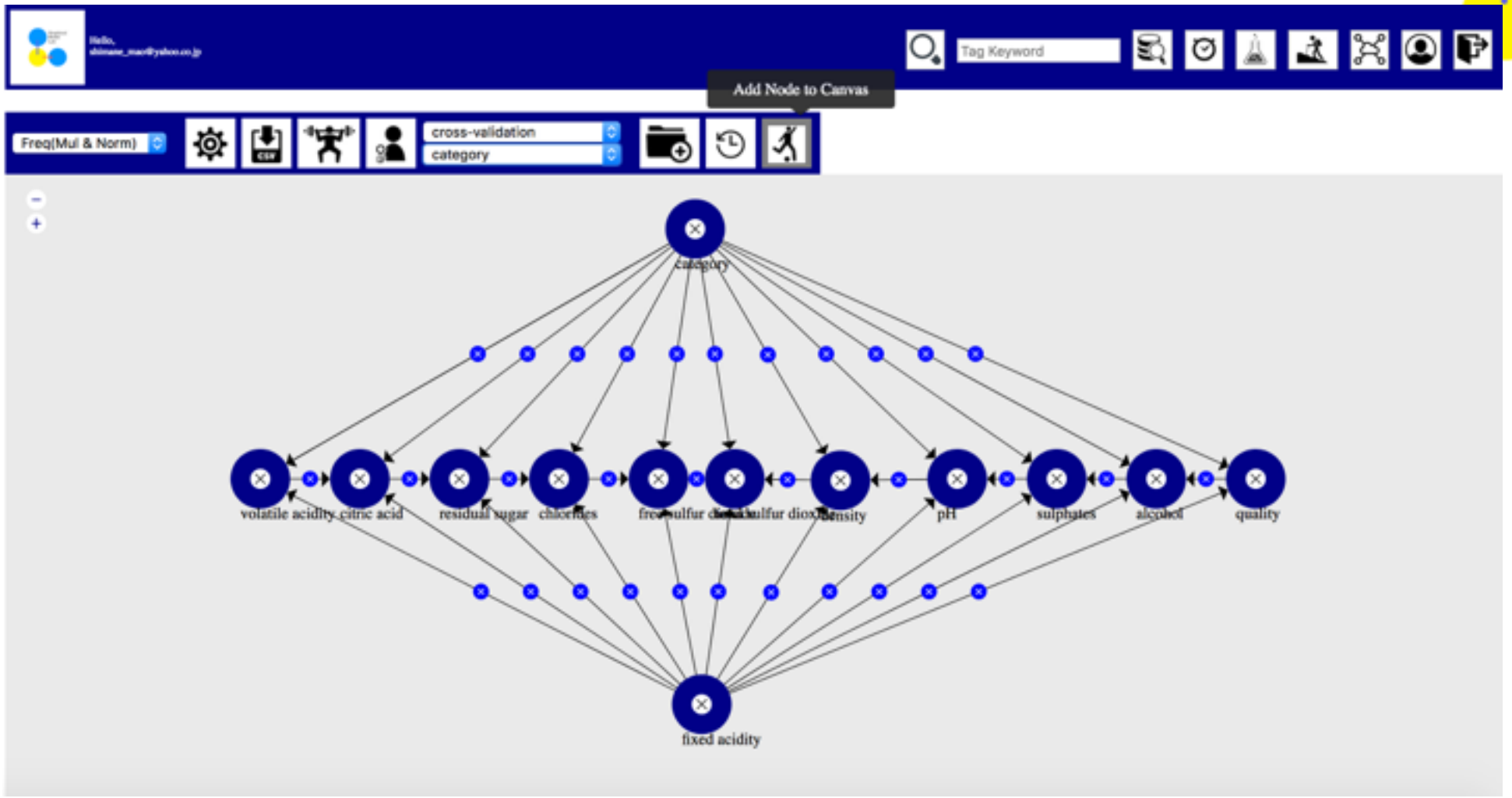
- Senior Data Engineer at Teradata Japan
- M.S from UW-Madison, U.S.A
  B.S from Shimane University, Japan

Today’s Agenda



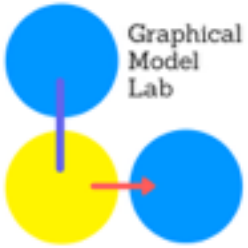
- What is Graphical Model Lab ?
- Architecture of Graphical Model Lab
- Models for calculating Graph
- Demo

Screen Shot of Current UI



* Does Demo later

Current Activity of The Community



- Tech Blog
  https://graphicalmodeling.com/

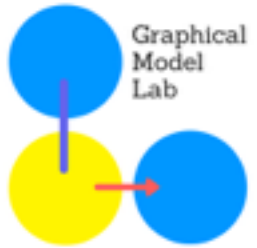
- Twitter
  https://twitter.com/gml55892067

- github
  https://github.com/GraphicalModelLab/GML_SaaS

- Meetups
  https://graphicalmodellab.connpass.com/

  The first meetup：2018 - 9/26 - Japan

What is Graphical Model ?

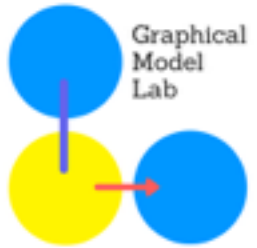


- In General, Graphical Model means something like Bayesian Network, Markov Random Field
  e.g. Naive Bayes

- Using Graph as a Representation, we can solve the following kinds of tasks:

  (1) Classification
  (2) Regression
  (3) Sequence・TimeSeries (e.g. finding parts of speech、predicting stock price, etc)
  …

What kind of OSS ?



- OSS SaaS for Machine Learning which can be represented as “Graph”
  => Include Deeplearning

- Interactive UI/UX
  Currently, React is used for building Simple Page Application
  View & Logic can be developed together
  => We can develop UI in a object oriented design

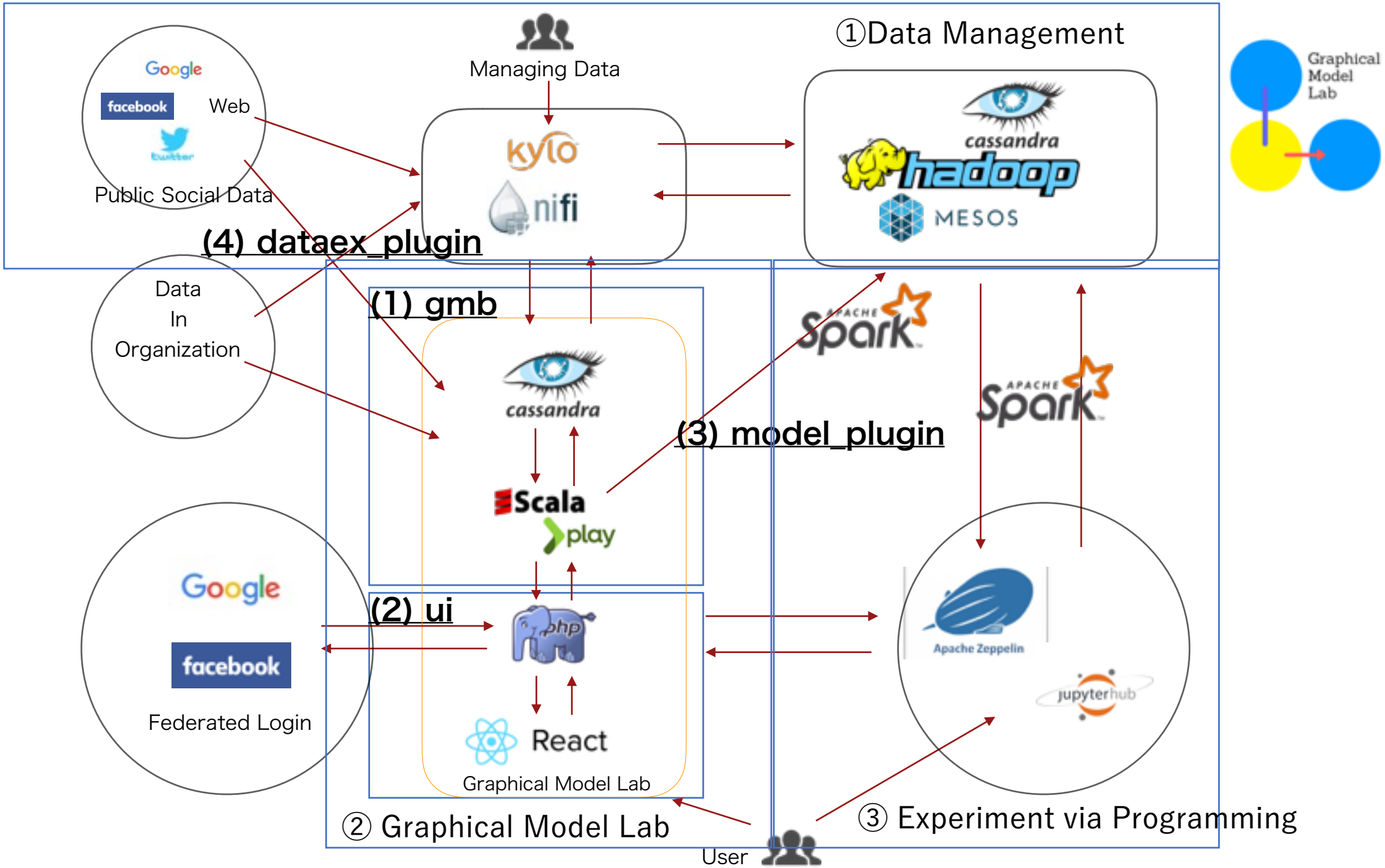
- Utilizing Big Data
  (1) Utilizing Social Data in Web
  => Develop new Web Query Language

  (2) Utilising private data existing in the organization

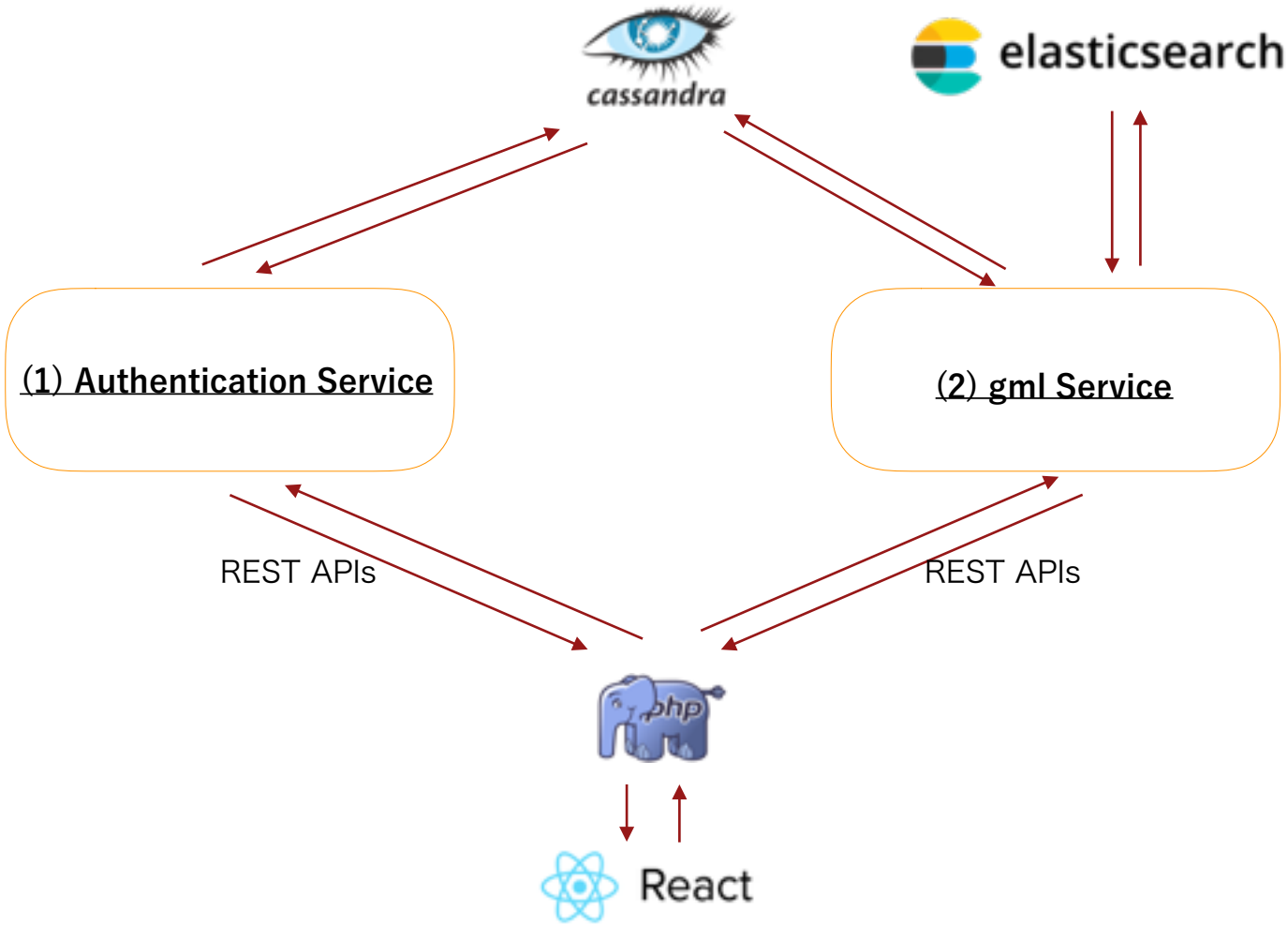
  (3) SNS
  => can share other people’s result
  => Transfer learning

  (4) Sharing/Saving domain knowledge

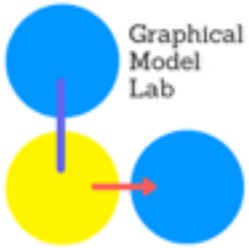
- system to improve analytics
  e.g. how can you improve your model after you got 95 % accuracy in the experiments ?


①Data Management
Managing Data
Google
facebook
Web
Public Social Data
kylo
nifi
cassandra
hadoop
MESOS
Graphical Model Lab
(4) dataex_plugin
Data In Organization
(1) gmb
cassandra
Scala
play
(3) model_plugin
APACHE Spark
APACHE Spark
(2) ui
php
React
Graphical Model Lab
Google
facebook
Federated Login
Apache Zeppelin
jupyterhub
② Graphical Model Lab
③ Experiment via Programming
User

② Graphical Model Lab



List of Available Models



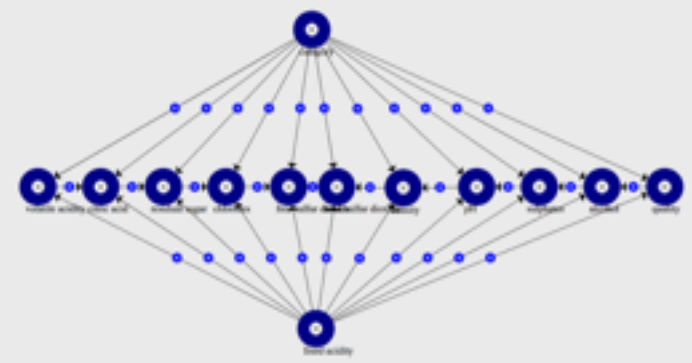
In big picture, we categorise models in the following 2 big categories with each 3 subcategories:

- Generative Model
  (1) Directed Model
    * one parametric model with Multivariate Gaussian
    * one non parametric model with Kernel Density Estimation
  (2) Undirected Model
  (3) Mixture of Directed and Undirected Model

- Discriminative Model
  (1) Directed Model
    * one model (only theory part has been created)
  (2) Undirected Model
  (3) Mixture of Directed and Undirected Model

Plugin For Calculation model

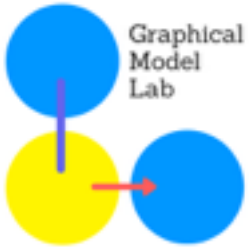
- It is hard to create one universal model which is effective for all use cases.
  => Research Level
- By creating plugin functionality, I want people to develop algorithms.



How to calculate ?

ans. Depends on the model

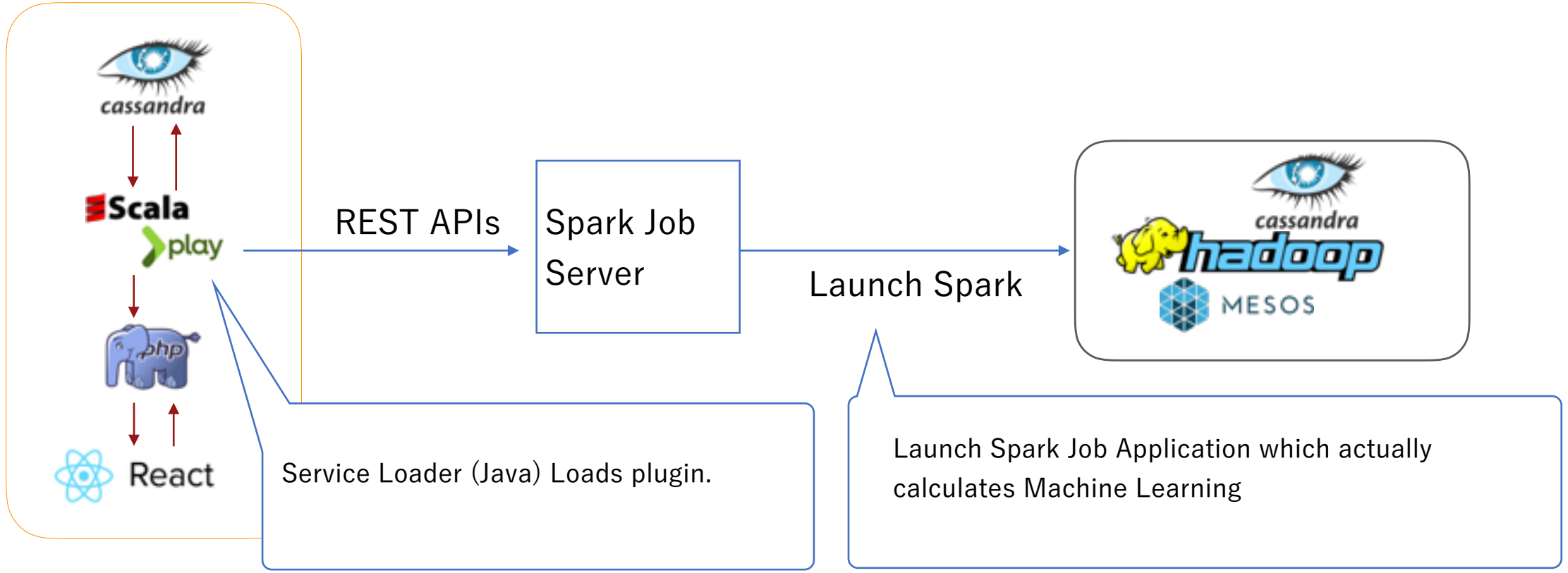
Two Generative Models



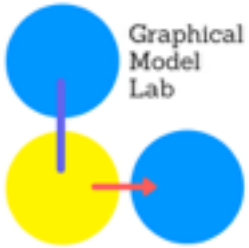
(1) One parametric model for Directed Graph

(2) One non parametric model for Directed Graph

How this two plugin works



Wine Data Set



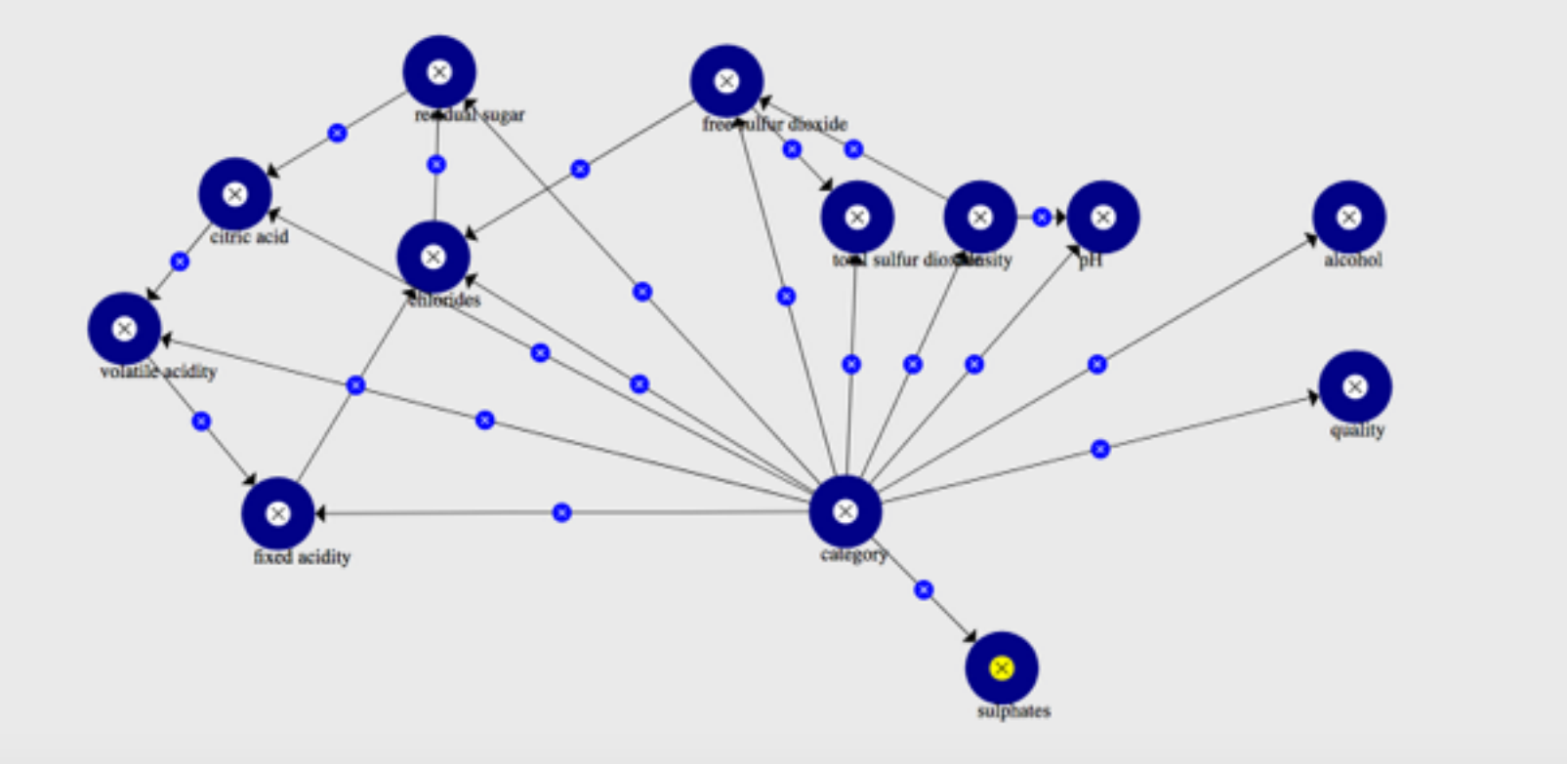
- https://archive.ics.uci.edu/ml/machine-learning-data wine-quality/

- Includes wine data, e.g. chemical features of wine, etc

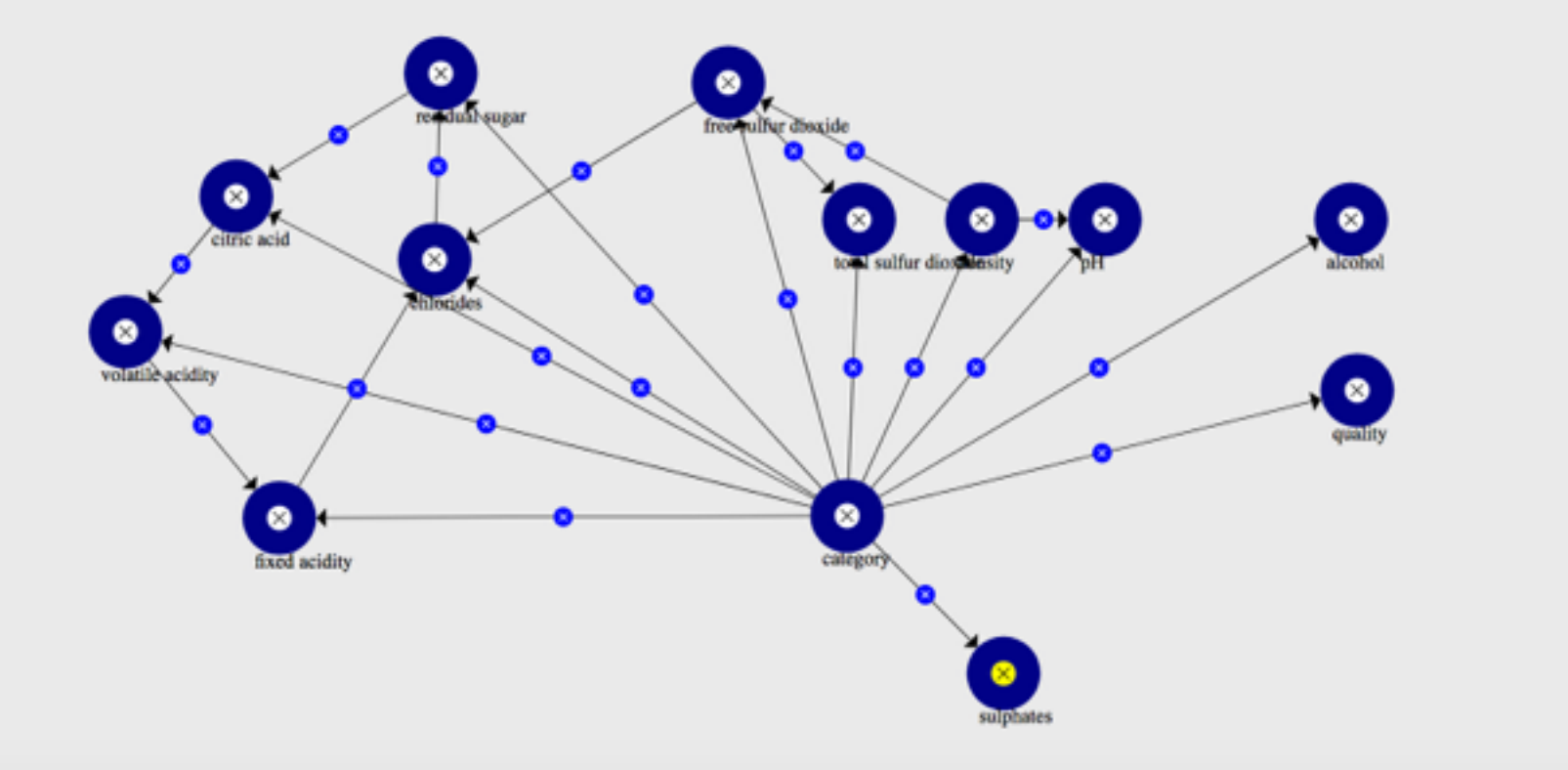
The goal of task

Graphical Model Lab

- Predict if a liquid is red or white wine, give chemical features

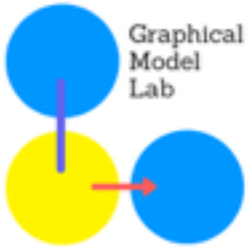


Graph And Probability



P(category, sulphates, quality, pH, …) : Joint Probability

How do we infer red or white ?



- The mathematical formula for inference is ..

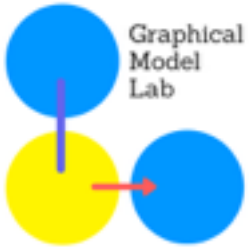
Inferred Label = argmax P(category = x | given wine chemical features)

P(category = red | give wine chemical features)

P(category = white | give wine chemical features)

Calculating the above 2 formula, we pick the higher one. Then, that label is the predicted label.

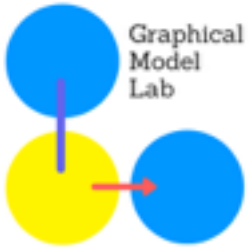
How do we infer red or white ?



•

$$P(\text{category} = \text{red} \mid \text{given chemical features}) = \frac{P(\text{category} = \text{red, given chemical features})}{P(\text{given chemical features})}$$

How do we infer red or white ?



- In conclusion, if we can calculate the following formula, we can infer red or white:

$$P(\text{category} = x \text{ , given chemical features})$$

This task is a classification task so that we don't need to calculate denominator part

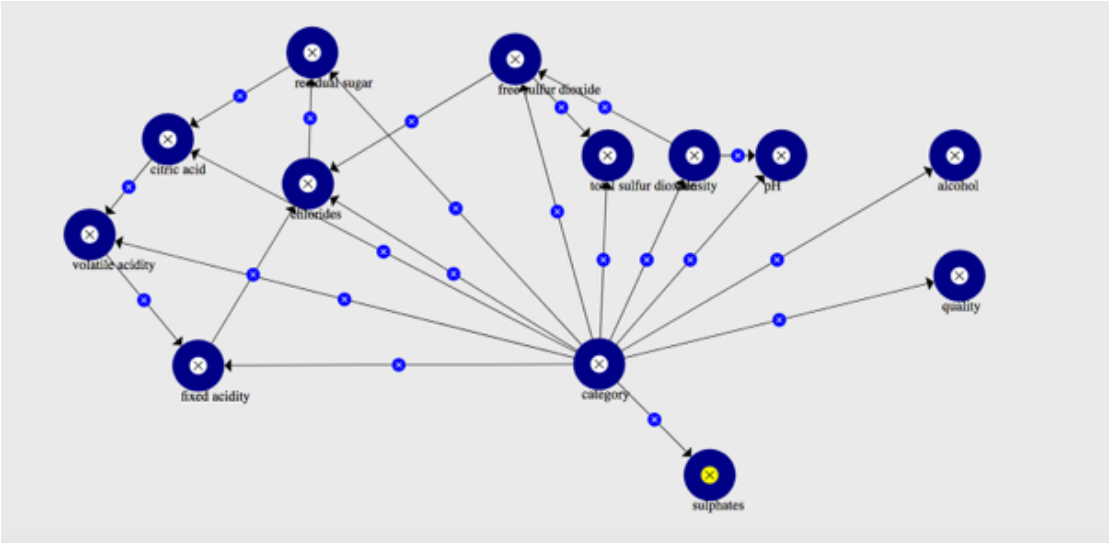
=> i.e. if joint probability can be calculated, then we can classify data

How do we infer red or white ?

- P(category = red , given chemical features)

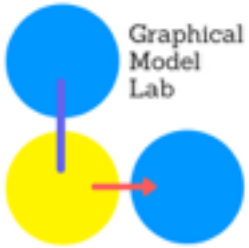
  = P(category=red) P(pH=0.11 | category=red, dioxide=192.2)…

Depending on graph structure, this formula changes.



Depending on the case, we would take “log”. But, in this example plugins, we do not take “log”

How do we infer red or white ?



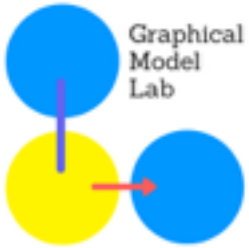
- P(pH=0.11 | category=red, dioxide=192.2)

  = P(pH=0.11 | dioxide=192.22)

    * we only consider a subset of dataset where category = red.

  = P(pH=0.11, dioxide=192.22) / P(dioxide=192.22)

How do we infer red or white ?



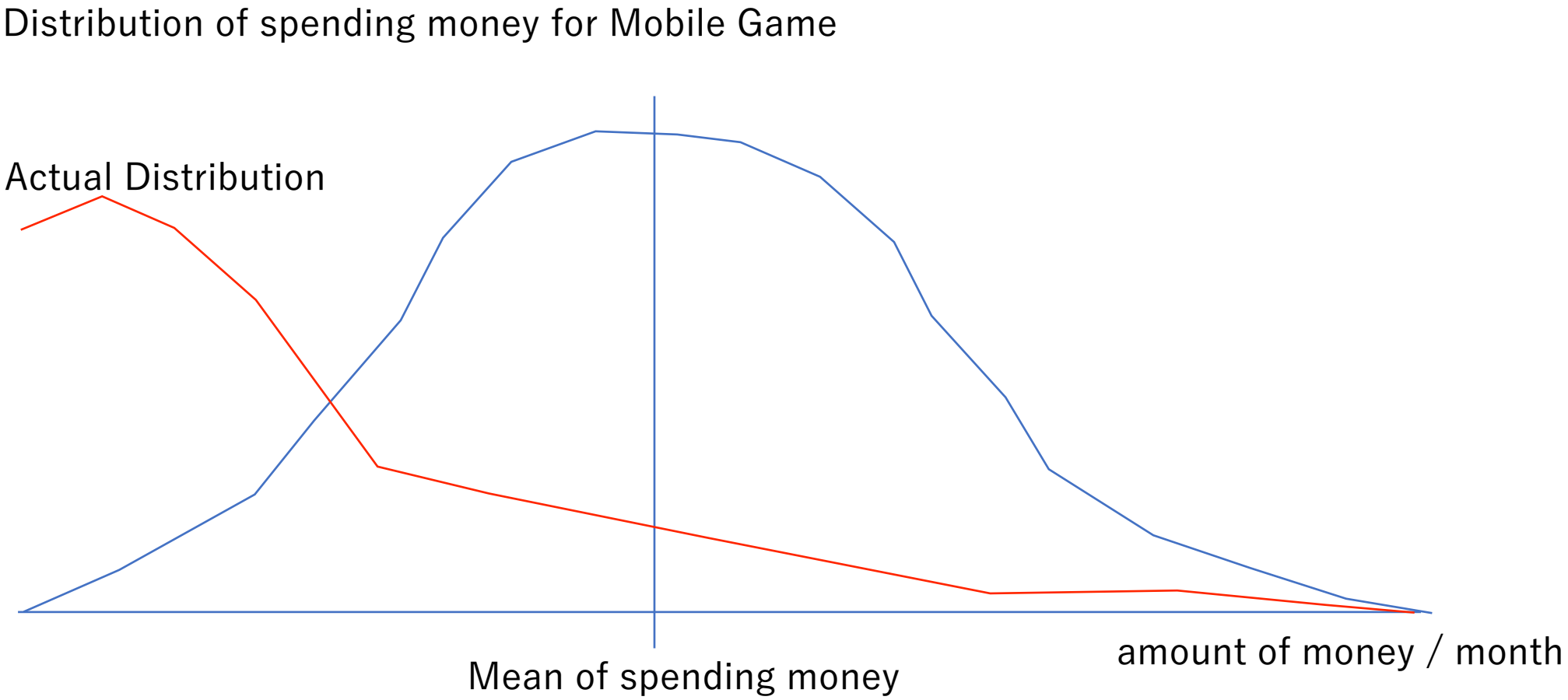
- Then, we just need to calculate the following terms:

  P(pH=0.11, dioxide=192.22)
  P(dioxide=192.22)

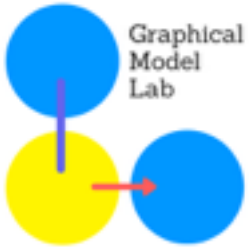
  For calculating these quantity, there are two approaches:

  (1) Multivariate Gaussian Distribution (Parametric)
  (2) Kernel Density Estimation (Non Parametric)

Gaussian Distribution



(1) Multivariate Gaussian



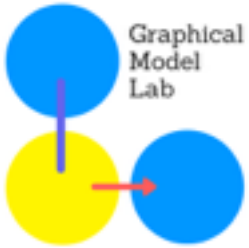
- How can we calculate via Multivariate Gaussian ?

  P(pH=0.11, dioxide=192.22)
  = PDF (Mean Vector, Covariance Matrix)

  PDF : Probability density function of multivariate gaussian
  => i.e. does not calculate the exact probability

  Now, if we have Mean Vector and Covariance Matrix, we can calculate this term
  => Finding Mean Vector and Covariance Matrix is “training” phase in this model.

(1) Multivariate Gaussian

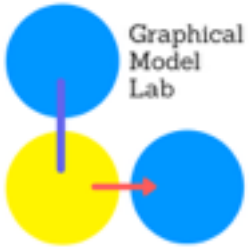


- Mean Vector と Covariance Matrix

  Mean Vector = Mean vector of training data
  Covariance Matrix = covariance matrix of training data

  => i.e. same as maximum likelihood estimate

(2) Kernel Density Estimation



- P(pH=0.11, dioxide=192.22)

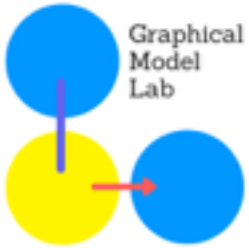
  = f(pH=0.11, dioxide=192.22)
  = 1/nh Σ K(ph=0.11, dioxide=192.22)
  = 1/n Σ Π 1/hj K(xj = xxx) : Multiplicative Kernel

  * xj means chemical features like “ph”, “dioxide”, etc.

-

  * What is the difference from Multivariate Guassian ?
  ＝＞We don’t have any parameters we need to learn.
  In stead of non parametric, the amount of calculation increases.
  There might be some research which optimize this calculation. But, I have not investigated.

How about Categorical Term ?

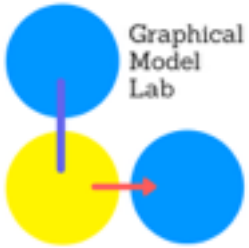


- How do we calculate P(category=red) ?

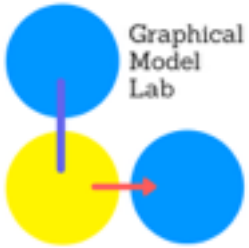
  In this example model, we just return “1” for the probability terms.

  => i.e. same as uniform distribution. All patterns have the same probability
  => Uniform distribution actually does not return “1”. But, this is classification task. So, the behaviour is the same.

  Currently, I am using multinomial Distribution.



Demo Time



Thanks !!!!!!