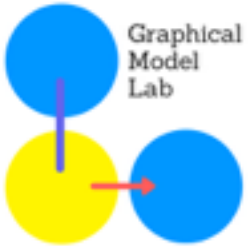
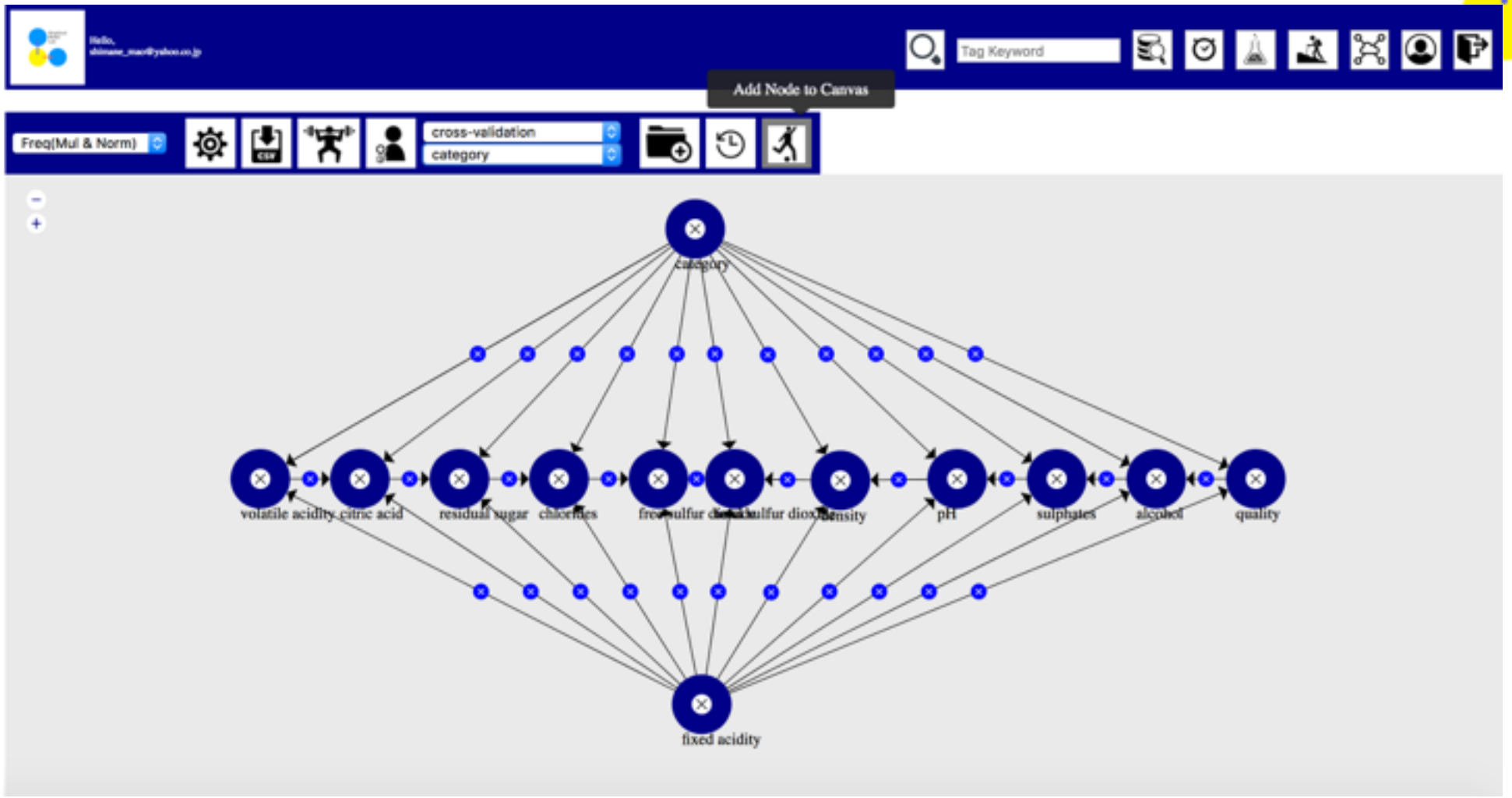


Graphical Model Lab

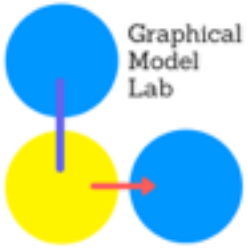
伊藤 真央

開発中の画面



* 後ほどデモします

コミュニティ活動 状況



- Blog
  https://graphicalmodeling.com/
  英語です。日本だけじゃなくて、海外でもアピールしたいということで、この辺は英語主体で行く予定です

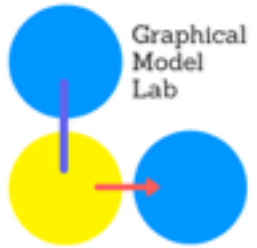
- Twitter
  https://twitter.com/gml55892067
  まさかのフォロワー０人ですので、興味のある方は、フォローしてください！

- ミートアップ
  第一回：今回です

- 国際会議
  来年の3月に東京で開催される国際会議で、発表する予定になっています。

  今後もミートアップを継続して行う予定です。その他の団体のイベントや会議とも、どんどん絡んで、ツールを成熟して、使えるものにしたいと思っています

グラフィカルモデルとは？

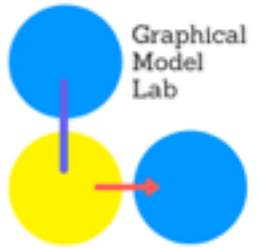


- 一般的には、ベイジアンネットワークや、マルコフ確率場と呼ばれるようなモデルのより一般的な名称

- グラフを使ってデータの関連性を計算し、以下のような計算ができる

  (1) 分類問題
  (2) 回帰
  (3) 順序列解析・時系列解析 (単語に対する品詞列、株価予測、など)
  …
  応用次第で、いろんな使い方ができる

どんなOSS？



- グラフベースの機械学習のSaaS
  => ディープラーニングも含む

- インタラクティブな UI/UX
  現在は、React で Simple Page Application
  ViewとLogicを一体化して開発できるというポイントで採用中
  => よりオブジェクトオリエンティッドなUIの開発が可能なので、簡単に複雑なUIが作れる

- ビッグデータの活用
  (1) ソーシャルデータの活用
  => 新しいWeb Query Languageの構想・開発

  (2) プライベートなデータの活用
  => いわゆる組織内のプライベートなデータ

  (3) SNS的な要素 (他の人の結果の活用など)
  => いわゆる転移学習的な要素を取り入れたい

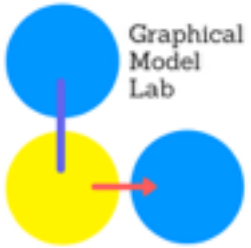
  (4) ドメイン知識の集積・共有

- 改善の仕組み化
  例えば、95%の精度が出た後に、この後、どう改善していくか？

OSSとして公開予定ですが、
まだ完成していないので、
公開していません。

現在は、私のポケットマネーで、PrivateのGitで開発しています。
5人まで追加できますので、開発参加に興味のある方は、
私に話して頂ければ、追加します。

本当にビッグデータが必要か？

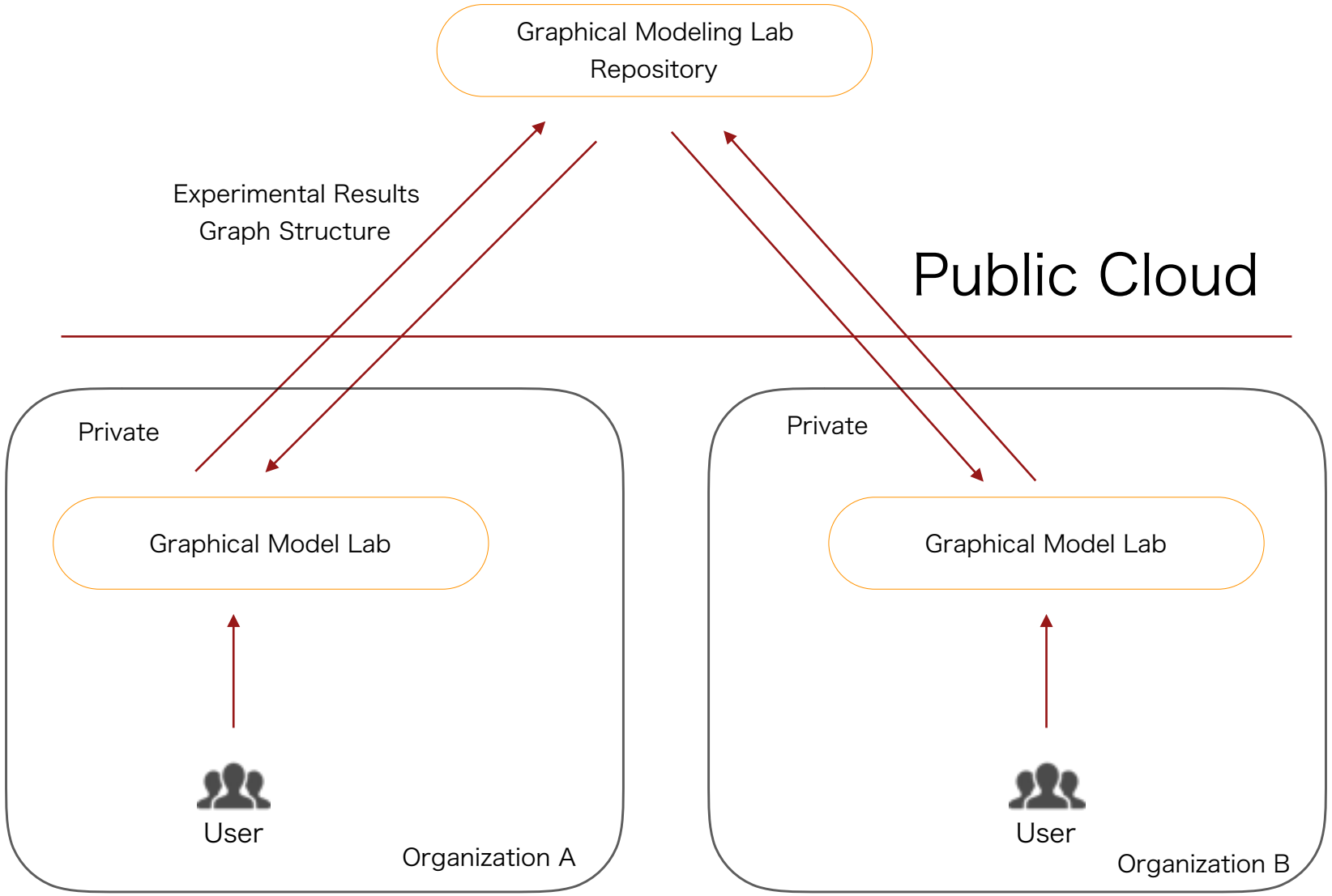


- 以下のような理由からビッグデータは必要：

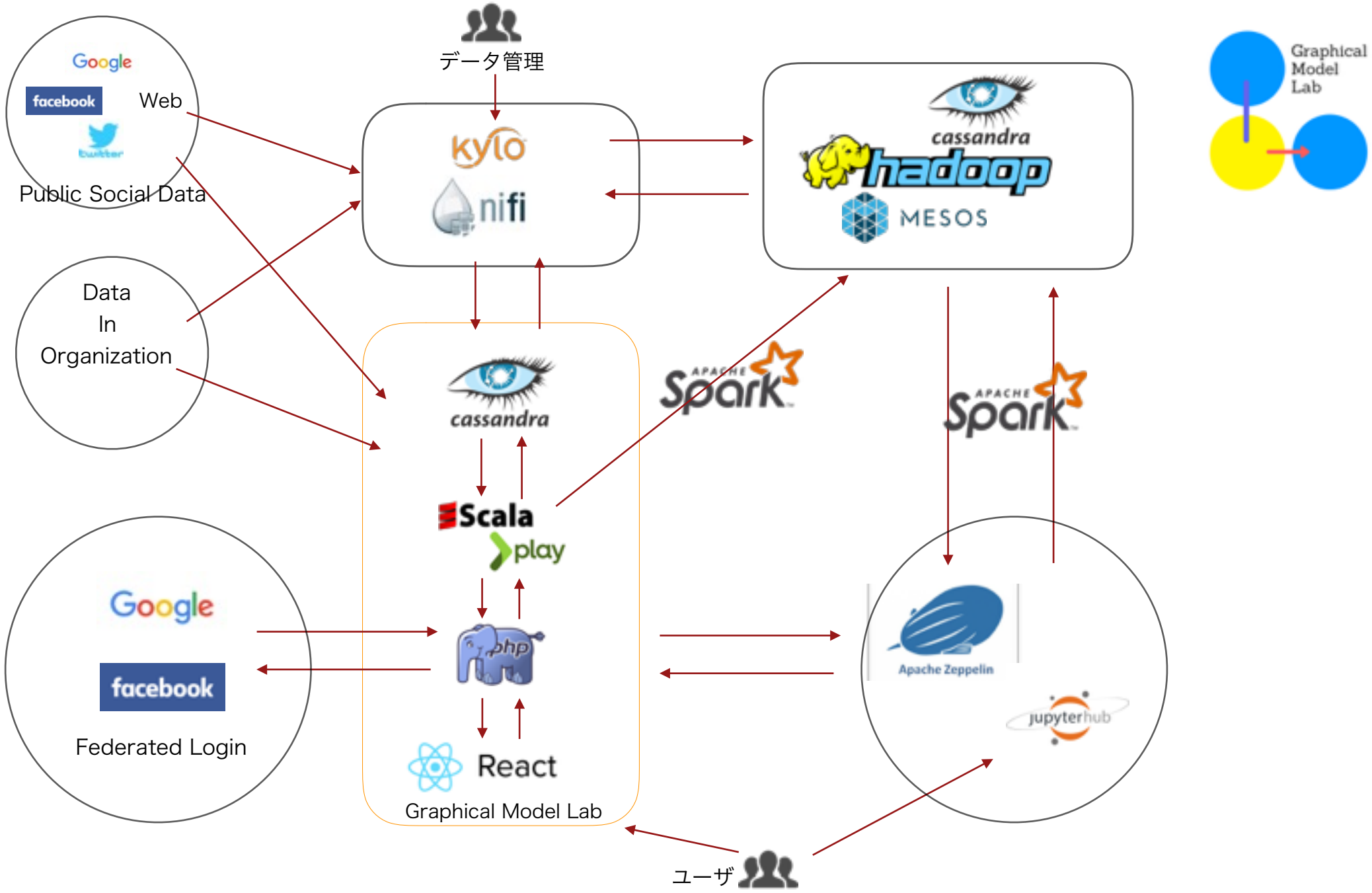
  (1) 常に新しいデータは存在する

  (2) Small データは間違った結論に繋がる
  （ドメインの知識が無い場合）

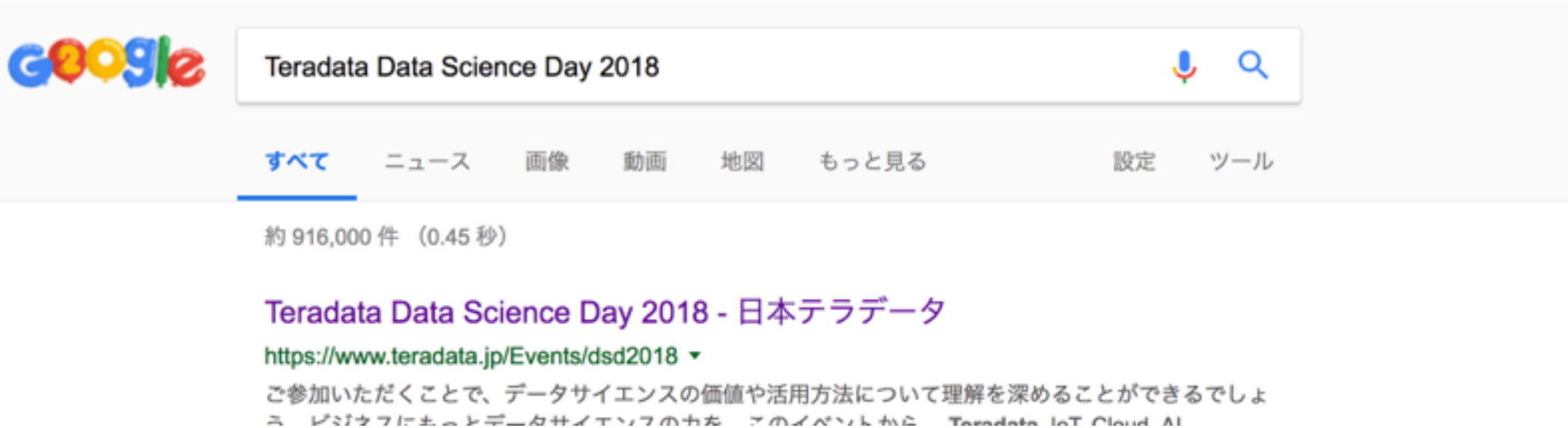
アーキテクチャ


Graphical Modeling Lab
Repository
Experimental Results
Graph Structure
Public Cloud
Private
Graphical Model Lab
User
Organization A
Private
Graphical Model Lab
User
Organization B

アーキテクチャ


Google
facebook
Web
Public Social Data
データ管理
kylo
nifi
cassandra
hadoop
MESOS
Graphical
Model
Lab
Data
In
Organization
cassandra
APACHE
Spark
APACHE
Spark
Scala
play
Google
facebook
Federated Login
php
Apache Zeppelin
jupyterhub
React
Graphical Model Lab
ユーザ

Kylo - Data Science Day 2018

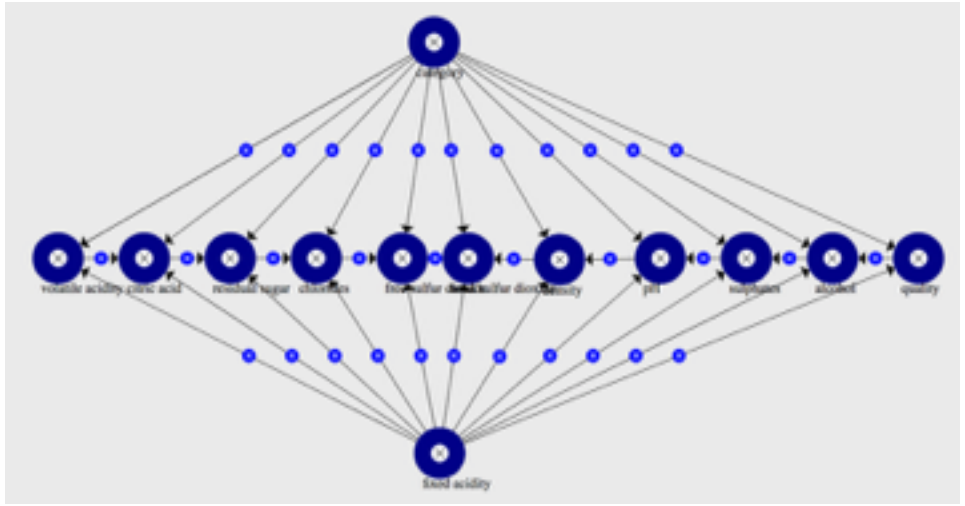


Kyloは、私の雇用主のオープンソースになってますので、このミートアップシリーズでの話題としては、控えさせてもらいます。このミートアップは私の個人プレーの趣味でやってますので。
今度、会社のイベントで、話しますので、興味のある方は、お申し込みくださいー

計算モデル

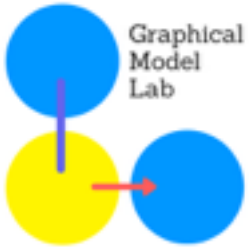
Graphical Model Lab

- 全てのケースで、有効なモデルは無い。
  => 研究レベル
- プラグイン機能を設けて、いろんな人に、計算モデルを作ってもらう



どう使う？
どうやって計算する？

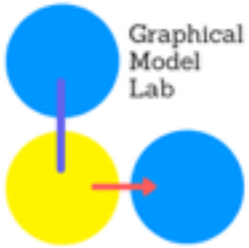
計算モデル - プラグイン機能



- リフレクションでプラグイン化

  (1) Scala Reflection API
  (2) Reflection API (Google)
  (3) ServiceLoader (Java)

  => (3)をとりあえず、現在は採用中

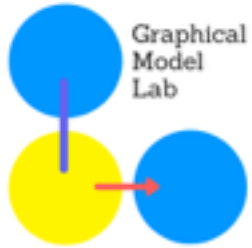


今回 紹介するモデルは、

(1) パラメトリックなアプローチ

(2) ノンパラメトリックなアプローチ

Inductive Biasとは？



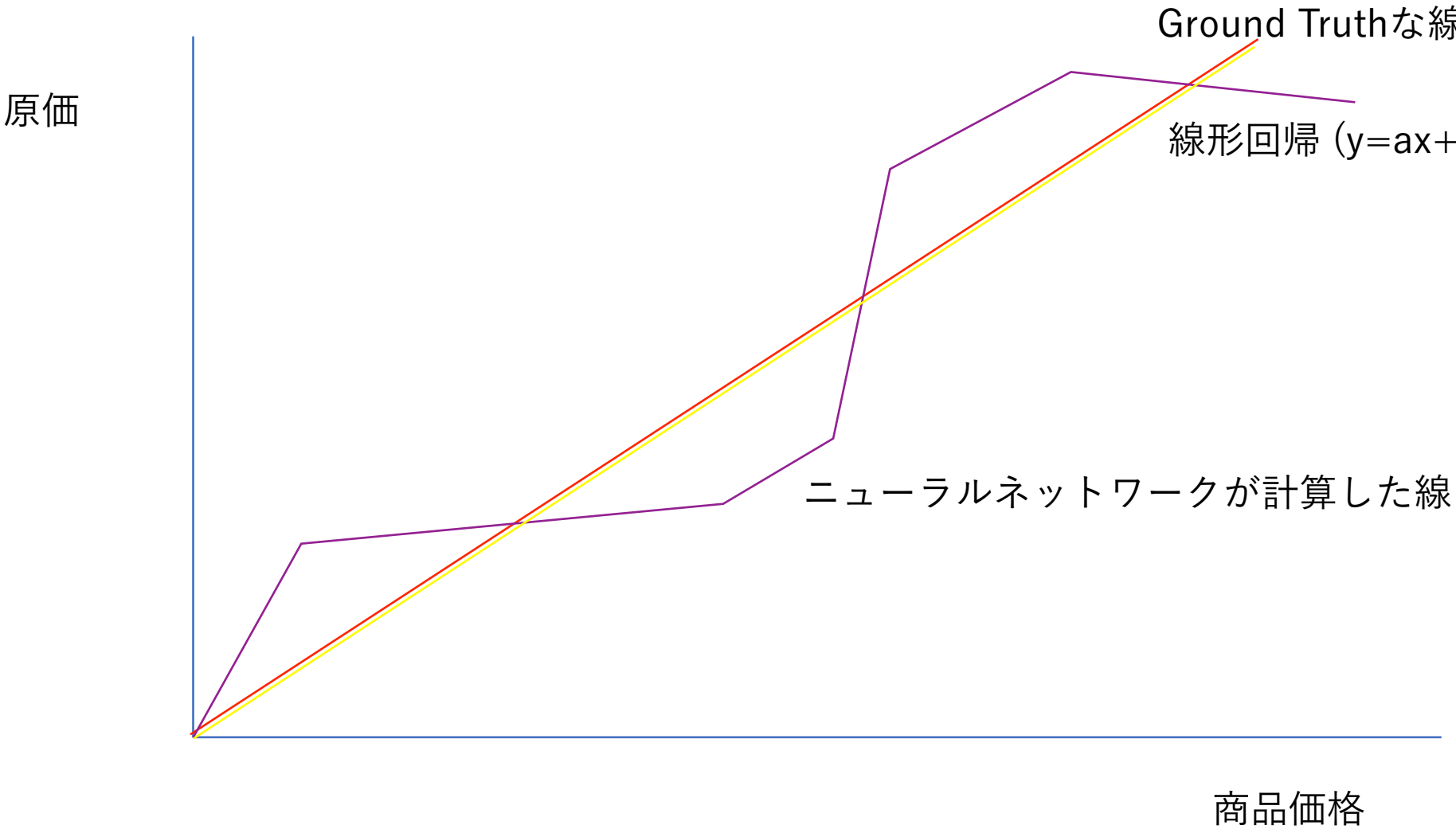
- 例えば、コンビニで売っている商品の商品価格から原価の値を、予測したいと云う問題設定を考える。つまり、

  原価 = F（商品価格）が成り立つような関数 Fを探しに行く。

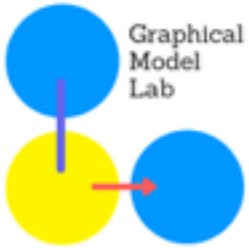
  そこで、以下の２つのモデルを比べてみる

  (1) ニューラルネットワーク
  (2) 線形回帰 (y = ax + b)

ニューラルネットワーク vs 線形回帰



線形回帰の方が予測が良さそう



- なぜ線形回帰の方が良いのか？

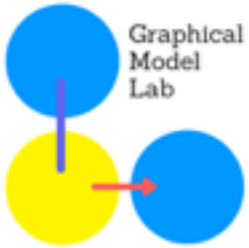
  この問題の場合、原価から商品価格を決める際、”人間”が、単純に計算した式である可能性が高い。例えば、原価の120%の値を商品価格と仕様など

  つまり、原価は商品価格の”単純な掛け算”である可能性が高い

  => 線形回帰はまさにこの”単純な掛け算”の線しか予測できない為、逆に良い
  　　ニューラルネットワークのようなジグザクな線は計算できない
  => ニューラルネットワークは、”単純な掛け算”以上の線を予測してしまう為、逆に悪くなる

＊注意：これはバイアスを説明する為だけなので、ニューラルネットワークで頑張れば、精度は出せるはず

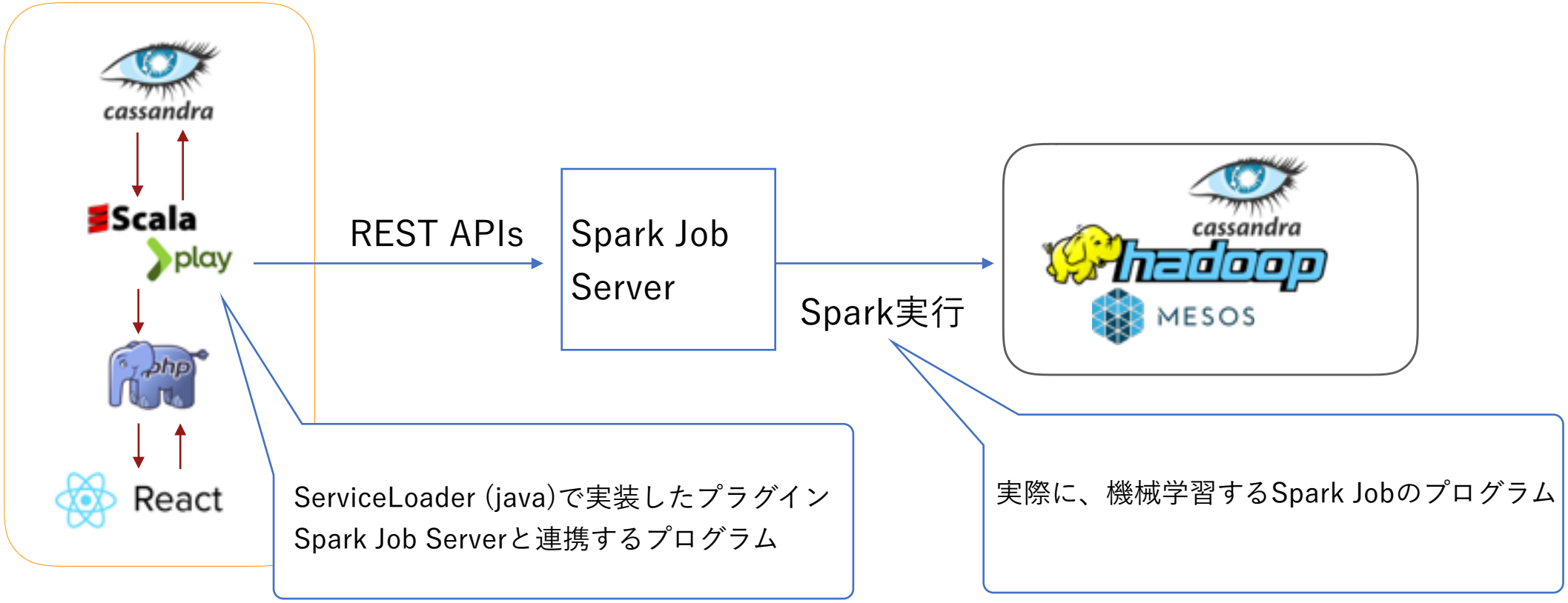
今回の計算モデル



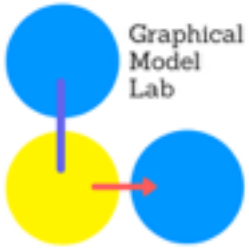
- 簡単なサンプルモデルをご紹介します。
  以下の条件で、使用できる:

  1. 特徴量は、カテゴリカルな値または数値 (IntまたはDouble)
  2. カテゴリカルな変数だけで依存してる項目はUniform Distributionを仮定

今回の計算モデルのプラグイン



ワインの分類問題



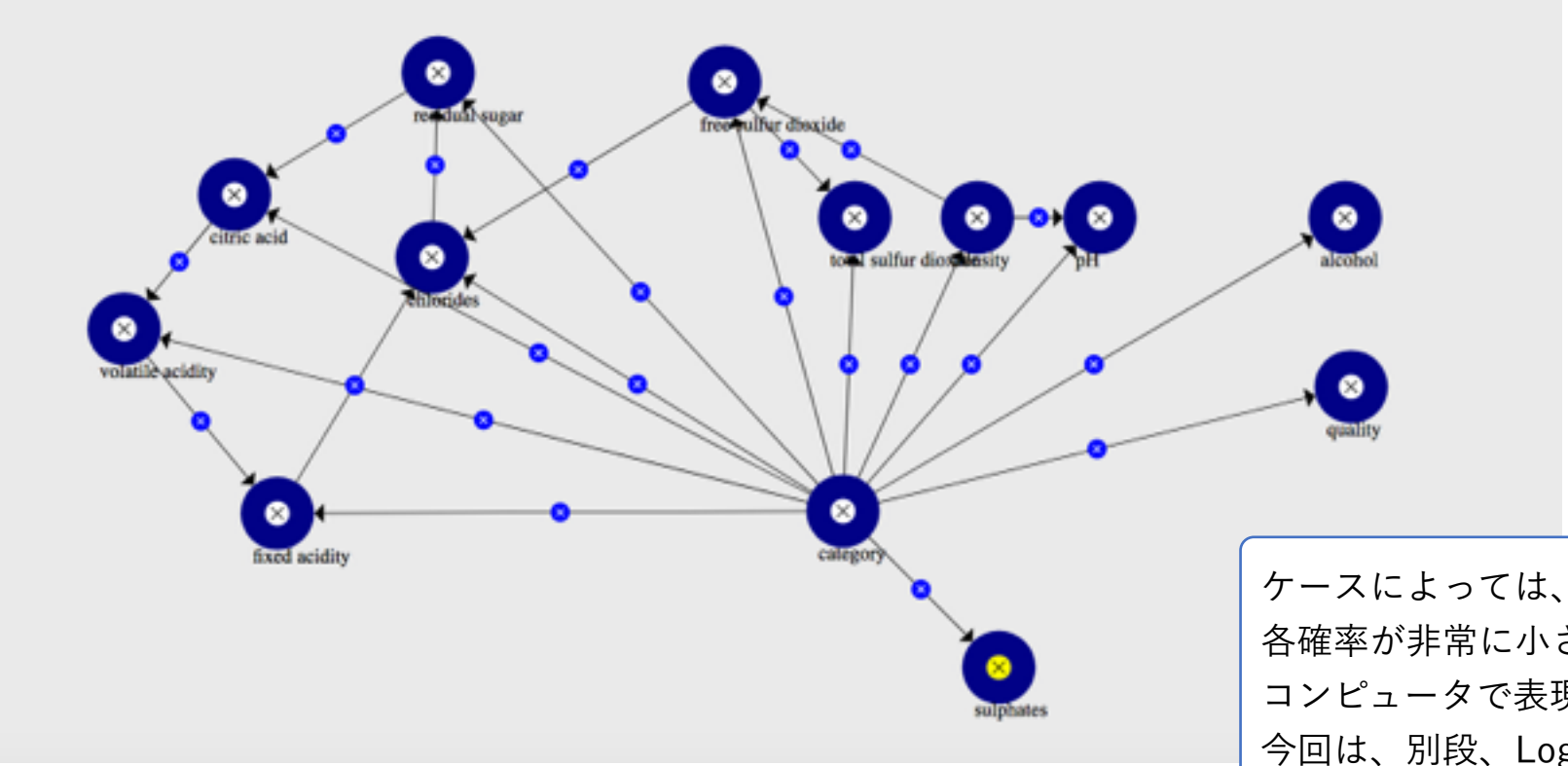
- https://archive.ics.uci.edu/ml/machine-learning-data wine-quality/

- ワインの成分情報やラベリング情報を併せ持つ

- 目標は、ワインの成分情報から”赤”か”白”を推定すること

  ＊このデータセットは、本来ワインの質の推定を、プロの人間ではなく、機械学習で置き換えようというミッションで作られてるものです

グラフと確率

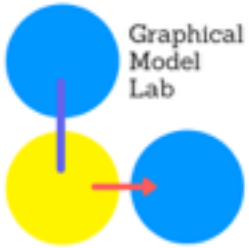


ケースによっては、Logで式展開したりします
各確率が非常に小さいと、掛け算していくと、
コンピュータで表現できなくなる。
今回は、別段、Logは取ってない

P(category, sulphates, quality, pH, …) <= 共起確率

= P(category)P(sulphates | category) P(quality | category) P(pH | category, dioxide)…

どうやってラベルを推定する？



- 数学的に書くと…

推定ラベル = argmax P(category = x | あるWine成分情報)

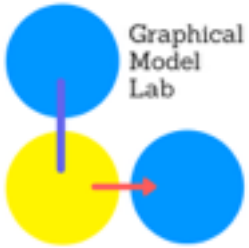
P(category = red　| あるWine成分情報)

P(category = white | あるWine成分情報)

上記の２つの式を計算して、高い方のcategoryのラベルを推定ラベルとする。

つまり、ある成分情報が与えられた時に、そのワインが、“red”か”white”とそれぞれ”仮定”して、確率を計算し、高い方が、そのワインの色と結論づける。

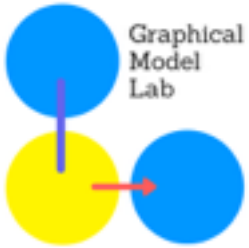
どうやってラベルを推定する？



- 以下で式展開：

$$P(\text{category} = \text{red} \mid \text{あるWine成分情報}) = \frac{P(\text{category} = \text{red}, \text{あるWine成分情報})}{P(\text{あるWine成分情報})}$$

どうやってラベルを推定する？



- 結論から言うと、下記の式が計算できれば、色を推定できる。

$$P(\text{category} = x \text{ , あるWine成分情報})$$

今回のタスクでは、分母の計算 P(あるWine成分情報）は必要ない。なぜなら、”red”と”white”で大きい方の値を色として推定するから。全てのラベルで分母は同じなので、省略できる。

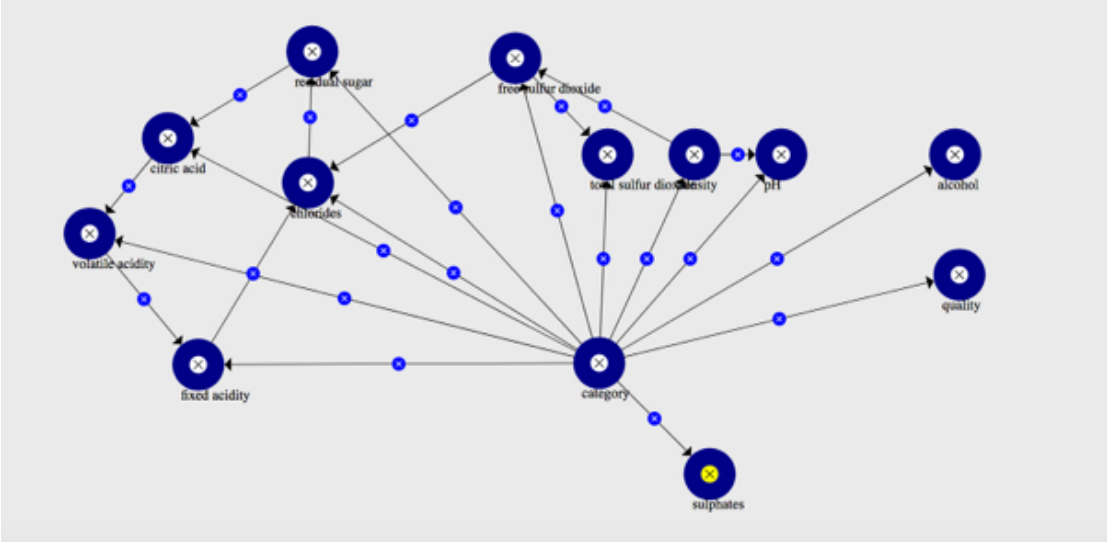
=> グラフの共起確率が求まれば、Wineの色を推定できる

どうやってラベルを推定する？

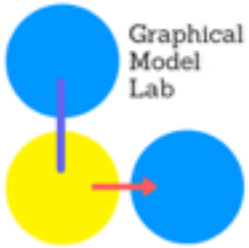
- P(category = red , あるWine成分情報)

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

グラフの構造によって、この式展開の結果が変わる



どうやってラベルを推定する？



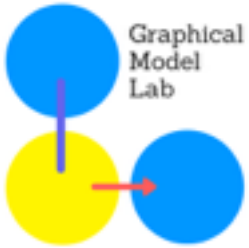
- P(pH=0.11 | category=red, dioxide=192.2)

  = P(pH=0.11 | dioxide=192.22)

  * データの中から category=redのラインだけを学習時に考慮する

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

どうやってラベルを推定する？

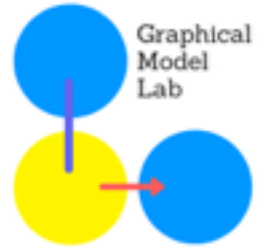


- あとは、分母と分子の計算式をどう計算するか？
  P(pH=0.11, dioxide=192.22)
  P(dioxide=192.22)

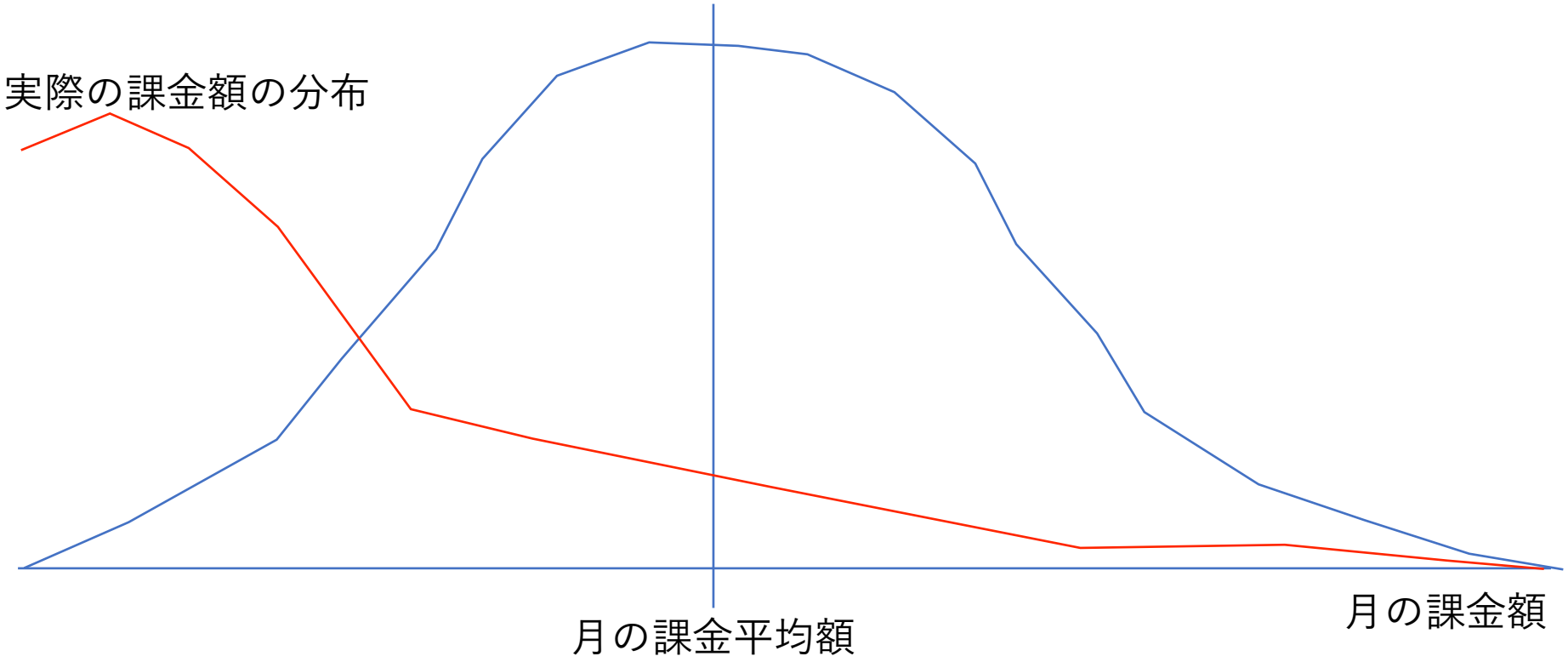
  以下の２つのアプローチを紹介。

  (1) 多変量正規分布　（パラメトリック）
  (2) カーネル密度推定（ノンパラメトリック）

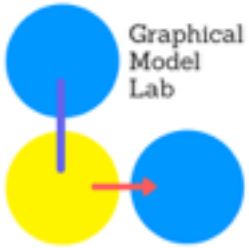
正規分布



ソシャゲにおけるか金額を正規分布で見てみる



(1) 多変量正規分布



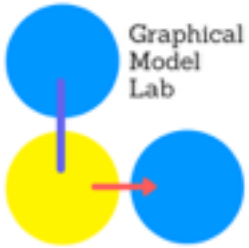
- あとは、分母と分子の計算式をどう計算するか？

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

PDF：多変量正規分布の確率密度関数 で近似という意味合いで今回してます
　=> つまり、厳密には確率の値は計算しない

Mean VectorとCovariance Matrixがあれば、ようやく、最終的な値を計算できる。
どうやって計算する？
=> この計算が、今回のモデルで言えば、“学習”と呼ばれるフェーズ

(1) 多変量正規分布

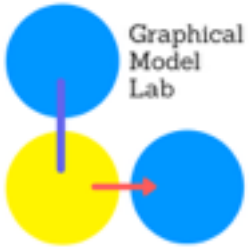


- Mean Vector と Covariance Matrix
  今回は、

  Mean Vector = サンプルの平均ベクトル
  Covariance Matrix = サンプルからの共分散行列

  => つまり、最尤推定と同じ
  => 本当に？Categoryでデータセット絞ってるけど、大丈夫？
  => 偏微分しますし、大丈夫かと思いますが、数学的な証明はしていません

(2) カーネル密度推定



- P(pH=0.11, dioxide=192.22)

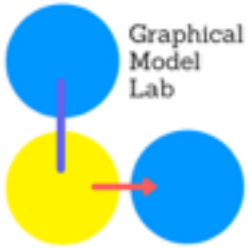
  = f(pH=0.11, dioxide=192.22)
  = 1/nh ΣK(ph=0.11, dioxide=192.22)
  = 1/n Σ∏ 1/hj K(xj = xxx)　　* xj は, phやdioxideの事

* 今回は多変量の変数に対して、Multiplicative Kernelを適用し式展開しています。

* Bandwidthは変数ごとに設定できる式展開にしてますが、今回のデモでは一律で同じBandwidthを適用

* 多変量正規分布と何が違うのか？
＝＞”学習”すべきパラメーターが存在しない。パラメーターを学習せずとも、計算できる。
　パラメーターを学習しない代償として、計算量が増大。 計算を最適化する研究もあるかもしれませんが、調べてません

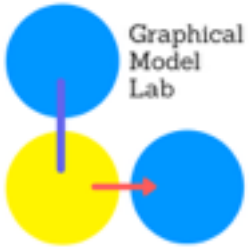
カテゴリカル変数はどうするのか？



- P(category=red)はどう計算する？

  今回は、カテゴリカル変数だけ依存する確率は、すべて”1”を返すように組んでいます。

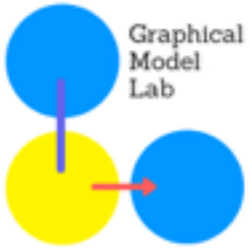
  => つまり、一様分布と同じ。すべてのパターンは同じ確率を持つ。
  => “1”を返すと、一様分布じゃないって話がありますが、今回は、分類という的に絞っと実装してますので、”1”を返すことは、一様分布にニュアンス的に近いという意味で言ってます。



Sparkで実装

実際に、ソースを見せながら説明

新しいWeb Query Languageの開発



- どのようにソーシャルデータを”簡単”に”抜き出し”, そして”結合”するか？
  例えば、

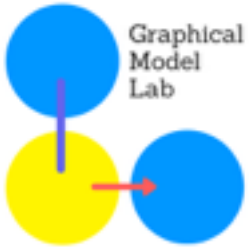
  SELECT userid_tweet
  FROM twitter
  WHERE “mycompany” in tweet

  INJECT userid_tweet

  https://www.w3.org/TandS/QL/QL98/pp/wql.html
  20年前あたりのIBMがやってたやつを、改良する？

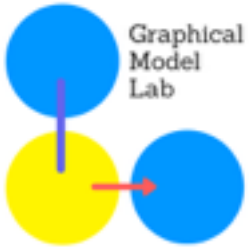
  SQLは古い？

未解決問題



- 違う確率分布に従う複数の確率変数の共起確率の定義
- ディープラーニングとのコラボ？
- どう転移学習をサービスとして展開するか
- 新しいWebQuery Languageの構想・開発
- 改善の仕組み化

いろんな確率分布のMix

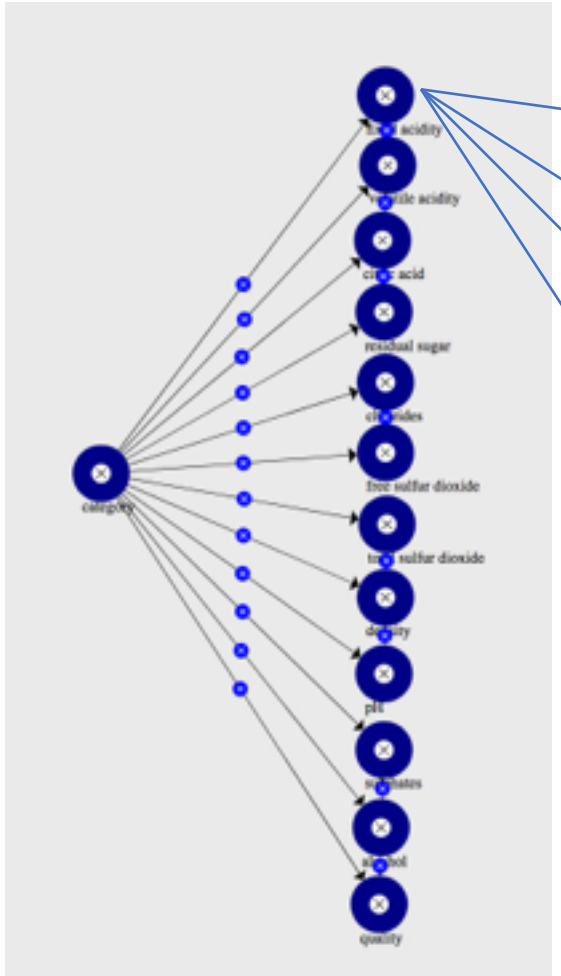


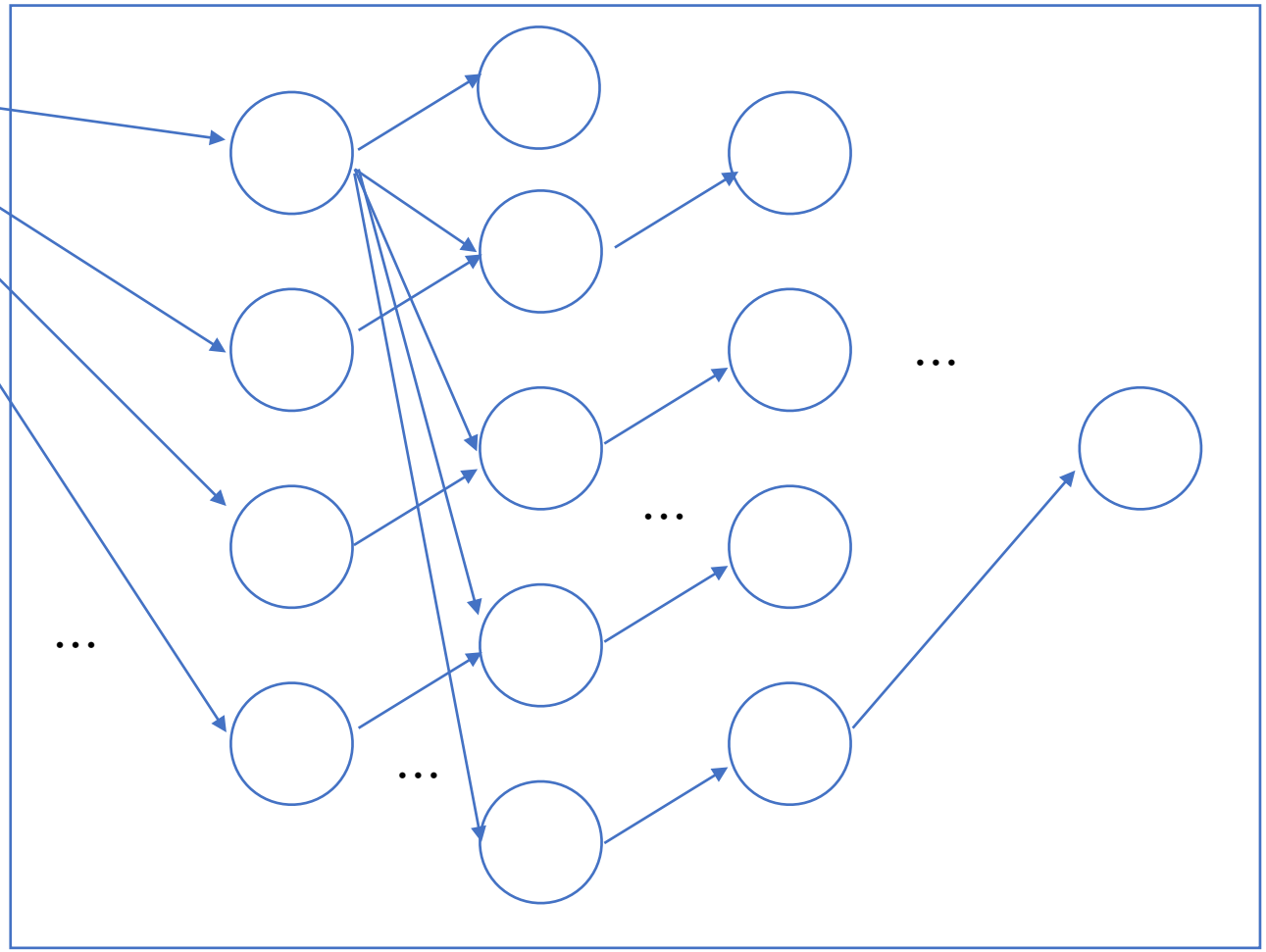
- 複数の違う確率分布の計算式の定義は存在しない
  （私が知る限り。あったら教えてもらえると有り難いです）

$$P(X, Y) = k\,P(X) + f\,P(Y)$$

線形結合して、EMなどで、kとfを求める？

ディープラーニング？





Graphical Model as Pre-Processing

Deep Learning : Not Developing Yet