¡¡¡¡¡¡¡ HEAD ======= ¿¿¿¿¿¿¿ 8b06f277a20cc3d6e2ac8260de58bd4f3f6fca06
¡¡¡¡¡¡¡ HEAD ======= ¿¿¿¿¿¿¿ 8b06f277a20cc3d6e2ac8260de58bd4f3f6fca06

# seis-ml-api 中間レポート

情報科学類 江畑 拓哉 (201611350)、畑中 智之 (201611402)、栗本真太郎 (201511366)

## Contents



## 1 情報特別演習概要

本演習は江畑、栗本、畑中の 3 名により実施する。それぞれ演習を進め、最終的に大規模データベースを用いた機械学習 API（seis-ml-api）を作成することが目標である。

### 1.1 演習範囲に関して

以下に示す範囲を演習し、その成果を組み合わせることで seis-ml-api の作成を目指す。

Table 1: 演習範囲に関して

| 氏名 | 分野 | 内容 |
|---|---|---|
| 江畑 | データベースと機械学習の統合 | 大規模データを利用した機械学習の作成 |
| 栗本 | 機械学習 | 機械学習モデルの調整 |
| 畑中 | データベース | 大規模データベースの作成 |

### 1.2 seis-ml-api 概要

seis-ml-api は、機械学習を提供する Web API である。大規模データベースを採用する。
seis-ml-api のフローをに示す。

ユーザは API に自分の持っているデータを登録する。API は、API に登録されているデータを用いて機械学習処理を行い、その結果をユーザに返す。

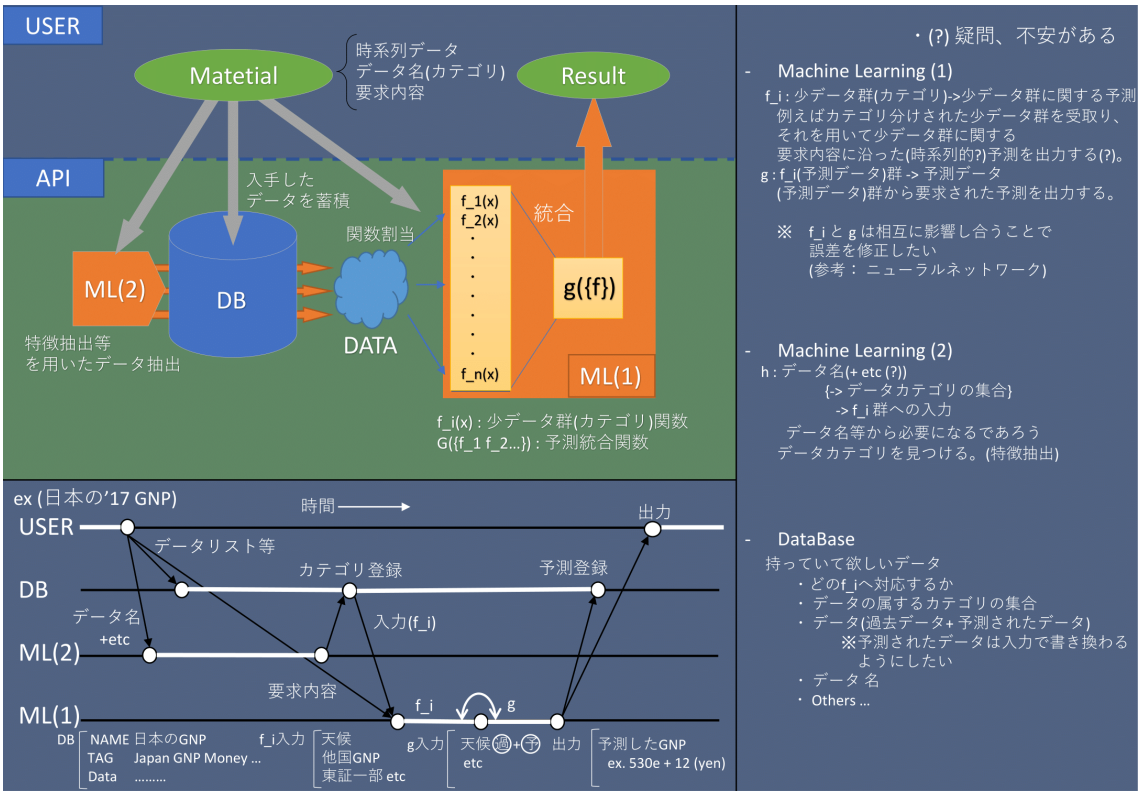


Figure 1: seis-ml-api フロー

### 1.3 実験に用いるデータ

以下のデータをデータベースに登録し、実験段階においても用いたいと考えている。
データの相関を求める関係上、これらの入手元から日本国内の経済についてのデータを集めたいと考えている。

- Quandl
  ほとんどすべてのデータはここで手に入る。ただし、長期間のデータは乏しいようである。
  主には、ここから得た株価データを用いて分析を行う予定である。

- google finance
  日本のデータを csv 形式で入手することは困難だが、海外のデータは容易に手に入る。

- 総務省統計データ
  めぼしいデータは少ないが、ゼロではないため活用していきたい。

## 2 それぞれの進捗について

- 江畑
  以下に引用されている文書を学習し、その結果より機械学習アルゴリズムの策定を行った。
  それに加え Java 、Python 、Clojure での Hbase の利用方法について学習した。

- 栗本
  「機械学習」の履修及び関連書籍の学習を行った。また、主専攻実験「ヒューマンセンシング」の自主実験において、サポートベクターマシンを用いた簡易画像分類器を作成した。

- 畑中
  HBase 、Hadoop を用いて疑似分散環境を構築した。また、Java から HBase にアクセスする方法を、実際にプログラムを動作させて確認した。

## 3 今後の演習について

- 江畑

策定した機械学習のモデルを実際のコードに実現する作業と HBase に入力されたデータを送る API を作成する。

- 栗本

機械学習モデルの理解を深め、Python によるデータ分析に慣れ、より良いモデル調整が可能なように学習していきたいと考えている。

- 畑中

完全分散環境を構築して、HBase の性能テストを行いたいと考えている。

## 4 参考文献

以下にそれぞれで用いた参考文献を示す。なお、これらの文献は今後より深く読み進めていく予定である。

- SARIMA モデルについて [12] [18] [14] [17] [1]

- RandomForest について [8] [20] [7] [11] [4] [10] [6] [3] [15] [2]

- 決定木について [19]

- データベースについて [13] [9] [5]

## References


[1] Carolia Garcia-Martos Andres M. Alonso. *Time Series Analysis*. URL: http://www.etsii.upm.es/ingor/estadistica/Carol/TSAtema4petten.pdf.

[2] Teppei Baba. 機械学習ハッカソン：ランダムフォレスト. SlideShare. URL: https://www.slideshare.net/teppeibaba5/ss-37143977.

[3] Leo Breiman. "Random Forests". In: *Mach. Learn.* 45.1 (Oct. 2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. URL: http://dx.doi.org/10.1023/A:1010933404324.

[4] Hemant Ishwaren Fei Tang. "Random Forest Missing Data Algolithms". In: (Jan. 2017). URL: https://arxiv.org/pdf/1701.05305.pdf.

[5] *HBase Tutorial*. tutorials point. URL: https://www.tutorialspoint.com/hbase/index.htm.

[6] *Imputation with Random Forests*. Cross Validated. URL: https://stats.stackexchange.com/questions/49270/imputation-with-random-forests.

[7] *Imputing missing values before building an estimator*. scikit learn. URL: http://scikit-learn.org/stable/auto_examples/missing_values.html.

[8] Satoshi Kato. *Imputation of Missing Values using Random Forest*. SlideShare. URL: https://www.slideshare.net/kato_kohaku/imputation-of-missing-values-using-random-forest?ref=http://kato-kohaku-0.hatenablog.com/entry/2016/05/01/155908.

[9] Christopher Miles. *All Your HBase Are Belong to Clojure*. Jan. 2012. URL: https://twitch.nervestaple.com/2012/01/12/clojure-hbase/.

[10] "Random Forests". In: (). URL: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

[11] *rfImpute*. RDocumentation. URL: https://www.rdocumentation.org/packages/randomForest/versions/4.6-12/topics/rfImpute.

[12] George Athanasopoulos Rob J Hyndman. *Forecasting: principles and practice*. Oct. 2017. URL: https://www.otexts.org/fpp/8.

[13] David Santiago. *clojure-hbase*. June 2014. URL: https://github.com/davidsantiago/clojure-hbase.

[14] The Pennsylvania State University. *Seasonal ARIMA models*. 2017. URL: https://onlinecourses.science.psu.edu/stat510/node/50.

[15] *What is the proper way to use rfImpute? (Imputation by Random Forest in R)*. Cross Validated. URL: https://stats.stackexchange.com/questions/226803/what-is-the-proper-way-to-use-rfimpute-imputation-by-random-forest-in-r?rq=1.

[16] 重穂 中村. "論文に於ける「だ」と「である」の選択条件に関する試行的考察". In: (Dec. 2009). URL: https://eprints.lib.hokudai.ac.jp/dspace/handle/2115/45684.

[17] 時系列解析. URL: https://upo-net.ouj.ac.jp/tokei/contents/sub_contents/c01_06_00.xml.

[18] 時系列解析__理論編. Logics of Blue. June 2017.

[19] 決定木. Matlab. URL: https://jp.mathworks.com/help/stats/classification-trees-and-regression-trees.html#bsw6a62.

[20] 石岡恒憲. *Random Forest* を用いた欠測データの補完とその応用. 大学入試センター研究開発部, Nov. 2010. URL: http://www.rd.dnc.ac.jp/~tunenori/doc/jjasRf2010slide.pdf.