

いちばんやさしい
Python機械学習の教本

から見る
機械学習を学習する課題

Takanori Suzuki

#stapy 48 / 2019 Aug 8

今日話すこと

- 書籍の紹介
- この内容になった経緯
- 機械学習を学習する課題

最初にお願い

- ・撮影、ツイートぜひどうぞ
- ・スライドは公開します
- ・ぜひ懇親会でフィードバックください

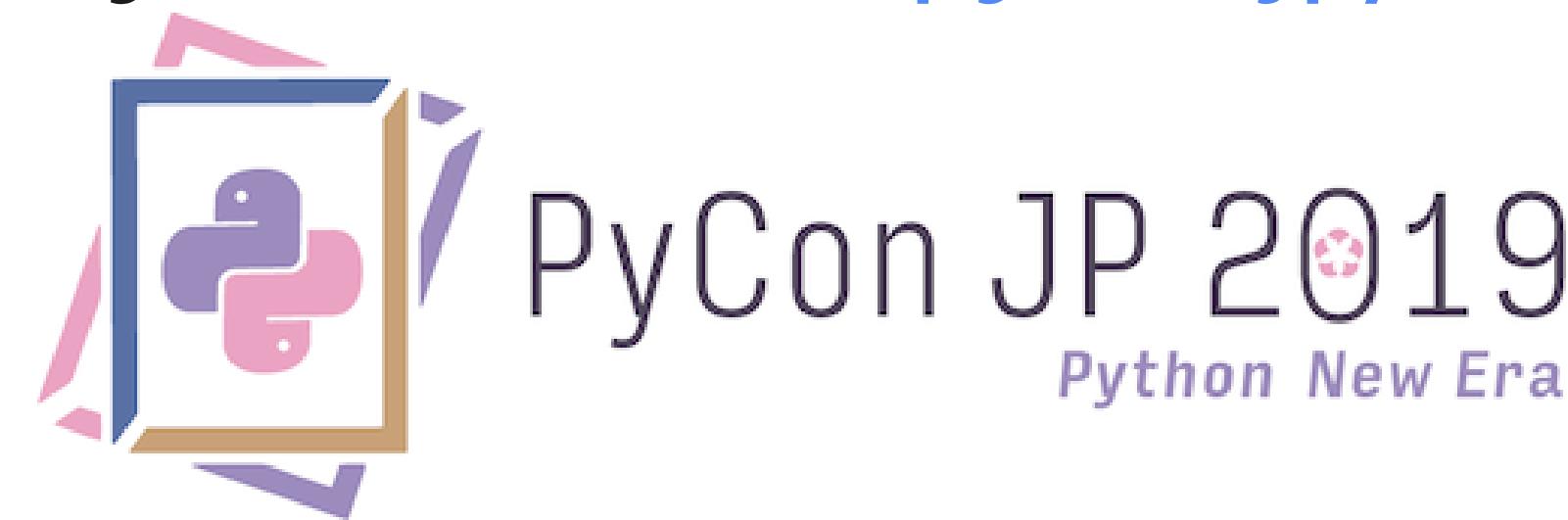
Who am I?(お前誰よ)

- 鈴木たかのり / Takanori Suzuki
- Twitter: [@takanory](#)
- 一般社団法人PyCon JP 副代表理事: #pyconjp
- 株式会社ビープラウド 役員
- Python mini Hack-a-thon 主催: #pyhack
- Pythonボルダリング部 部長: #kabepy



PyCon JP 🐍

- PyCon JP 2019: pycon.jp/2019



BeProud

- connpass: connpass.com
- PyQ: pyq.jp
- 書籍: Pythonプロフェッショナルプログラミング他



2019年海外PyCon発表に挑戦!!



- 2月: PyCon APAC
- 5月: US PyCon
- 6月: PyCon Thailand
- 7月: EuroPython
- 8月: PyCon Malaysia
- 9月: PyCon JP
- 9月: PyCon Taiwan
- 10月: PyCon Singapore
- 詳しくは gihyo.jp/news/report にレポート掲載中

書籍の紹介

いちばんやさしいPython機械学習 の教本

- 発売日: 2019年6月21日
- ページ数: 304
- 2,600円+税
- 著者: 鈴木 たかのり、降旗 洋行、平井 孝幸、株式会社ビープラウド



20件の結果 "いちばんやさしいPython"

並べ替え: アマゾンおすすめ商品 ▾

Amazonプライム

 ✓prime 通常配送料無料 (条件あり)Amazon.co.jpが発送する¥2000以上
の注文は通常配送料無料 (日本国内のみ)

配達日

 明日お届け

カテゴリー

本
コンピュータ・IT
Kindleストア
工学
コンピュータ・IT

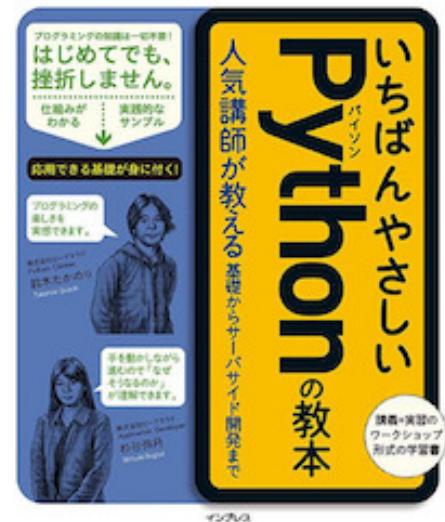
著者

 鈴木たかのり
 大澤文孝
 辻真吾

フォーマット

 単行本(ソフトカバー)
 単行本
 Kindle版

海外配送

 配送対象いちばんやさしいPythonの教本 人気講師が教える基礎からサーバサイド開発まで ('いちばんやさしい教本'シリーズ)
鈴木たかのり, 杉谷弥月他

★★★★★☆ ~ 12

単行本(ソフトカバー)

¥2,376

Amazon ポイント: 24pt (1%)

✓prime 明日中8/7までにお届け
通常配送料無料こちらからもご購入いただけます
¥1,676 (20点の中古品と新品)

その他のフォーマット: Kindle版

いちばんやさしい Python 入門教室
大澤文孝

★★★★★☆ ~ 18

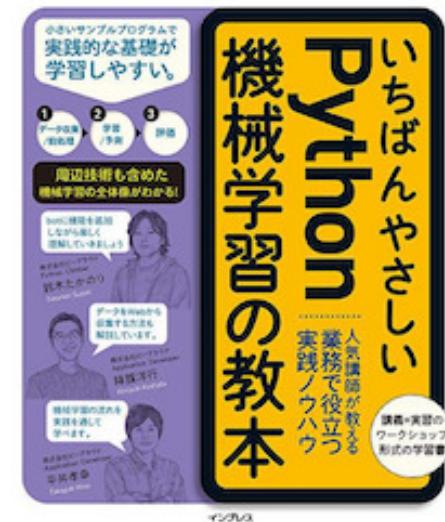
単行本

¥2,462

Amazon ポイント: 25pt (1%)

✓prime 明日中8/7までにお届け
通常配送料無料

その他のフォーマット: Kindle版

いちばんやさしい Python 機械学習の教本 人気講師が教える業務で役立つ実践ノウハウ
鈴木たかのり, 降旗洋行他

単行本(ソフトカバー)

¥2,808

Amazon ポイント: 28pt (1%)

✓prime 明日中8/7までにお届け
通常配送料無料こちらからもご購入いただけます
¥2,307 (6点の中古品と新品)

その他のフォーマット: Kindle版

いちばんやさしい Python の本
Python スタートブック [増補改訂版]
辻真吾

★★★★★☆ ~ 17

大型本

¥2,700

Amazon ポイント: 27pt (1%)

✓prime 明日中8/7までにお届け
通常配送料無料こちらからもご購入いただけます
¥2,056 (13点の中古品と新品)

その他のフォーマット: Kindle版

いちばんやさしい Python の本
Python スタートブック
辻真吾

★★★★★☆ ~ 57

大型本

こちらからもご購入いただけます
¥736 (26点の中古品と新品)

その他のフォーマット: Kindle版



ベストセラー

プログラミング
超初心者が初心者
になるためのPython入門

1

@takanory | #stapy

シリーズの特徴

「いちばんやさしいPython機械学習の教本」の読み方

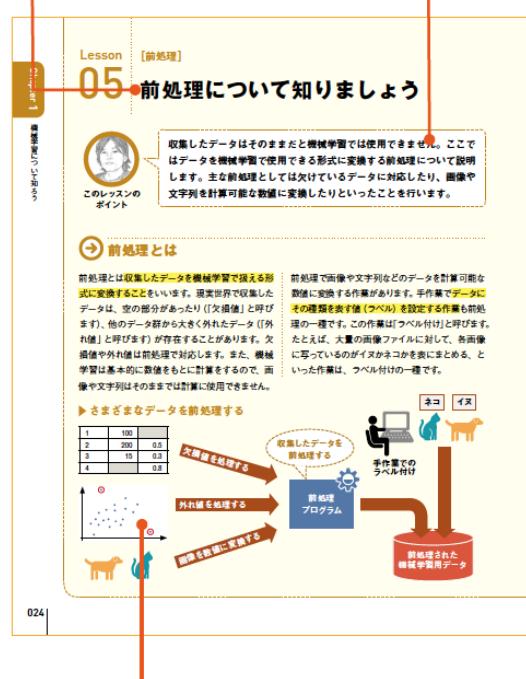
本書の読み方

「何のためにやるのか」がわかる！

「いちばんやさしいPython機械学習の教本」は、はじめての人でも迷わないように、わかりやすい説明と大きな画面でPythonを使ったプログラムの書き方を解説しています。

タイトル
レッスンの目的をわかりやすくまとめています。

レッスンのポイント
このレッスンを読むとどうなるのか、何に役立つかを解説しています。



解説
Webサイトを作る際の大変な考え方を、画面や図解をまじえて丁寧に解説しています。

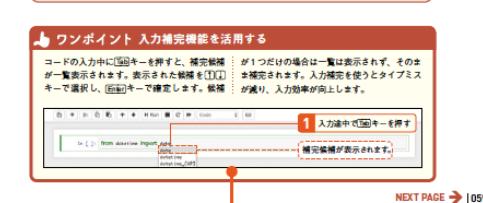
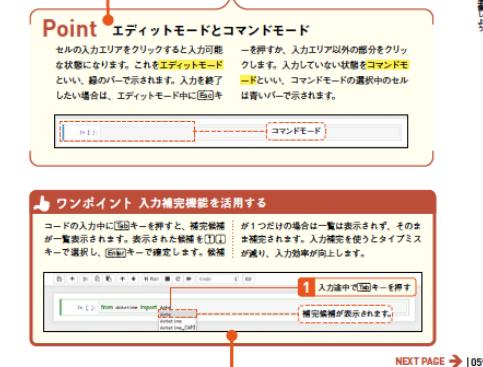
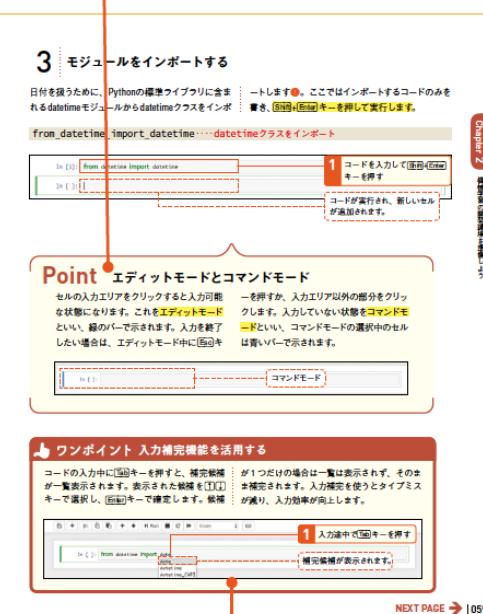
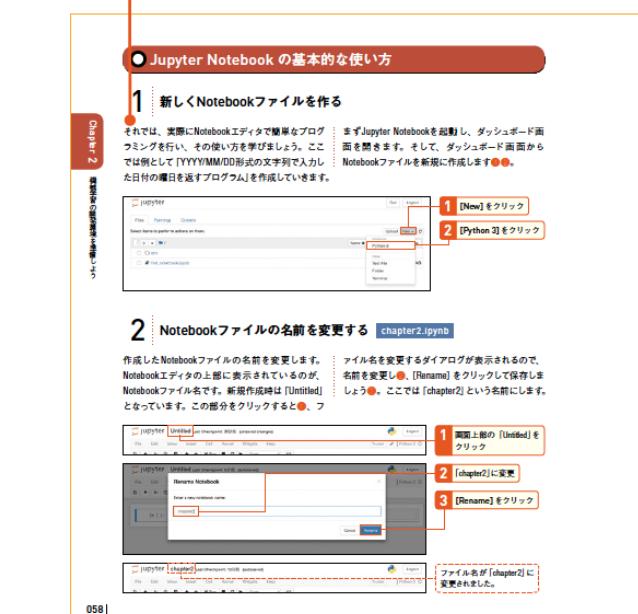
薄く色の付いたページでは、プログラムを書く際に必要な考え方を解説しています。実際のプログラミングに入る前に、意味をしっかり理解してから取り組めます。

「どうやってやるのか」がわかる！

プログラミングの実践パートでは、1つ1つのステップを丁寧に解説しています。途中で迷いそうなところは、Pointで補足説明があるのでつまづきません。

手順
番号順に入力をしています。入力時のポイントは赤い線で示しています。また、一部のみ入力するときは赤字で示します。

Point
その入力作業を行う際の注意点や補足説明です。



ワンポイント
レッスンに関連する知識や知っておくと役立つ知識を、コラムで解説しています。

本書の読み方

005

004

対象読者

- 「いちばんやさしいPythonの教本」を読んでいる
- Pythonの基本文法は知っている
- 機械学習学びたい人

主な内容

- 1章: 機械学習について知ろう
- 2章: 開発環境の準備
- 3章: スクレイピング(データ収集)
- 4章: 文章自動生成(自然言語処理)
- 5章: 手書き文字認識(機械学習)
- 6章: 表データの前処理
- 7章: 気温予測(回帰分析)
- 8章: 次のステップ

本書の特徴

- 1章は完全な読み物
- 3章/4章/5章/6章+7章は好きなところから
- 各章の最後はpybotでコマンド化

Lesson 11

[本書の読み進め方]

本書での学習の仕方を理解しましょう



このレッスンの
ポイント

次のChapterからいよいよ機械学習に関する技術を学びます。しかし、機械学習に関する技術は非常に多様であるため本書ではその一部しか解説、実践していません。本書で学習を進める上でのポイントや、本書では扱わない内容について解説します。

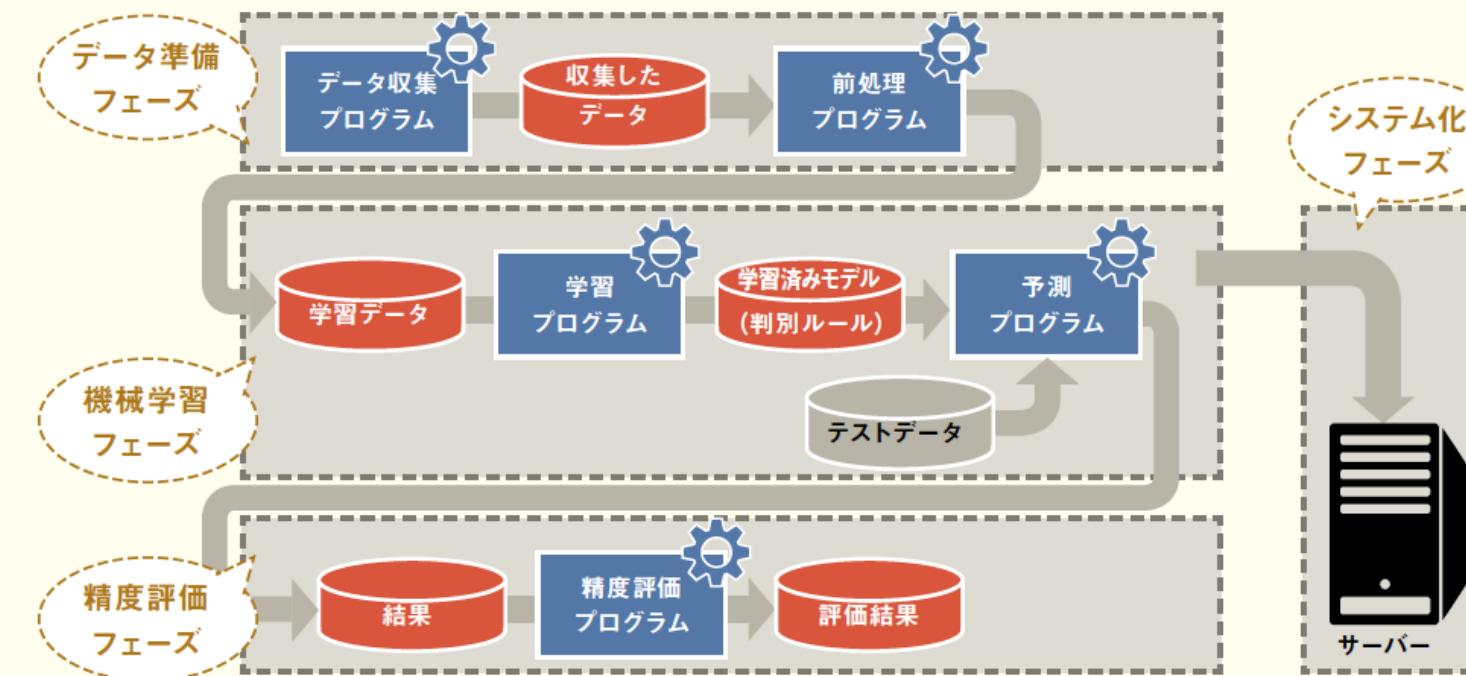
→ 機械学習プロジェクトの全体の流れ

機械学習を組み込んだシステムを作成する機械学習プロジェクトの全体的な流れを改めて見てみましょう。ここでは便宜上4つのフェーズに分けて説明します。

最初のフェーズは機械学習に使用するデータを集め前処理を行う、データ準備フェーズです。データが準備できたら、プロジェクトの中心となる機械学

習フェーズで学習済みモデルを作成します。次に結果から精度を計算して評価を行うのが精度評価フェーズです。PoCではこれら3つのフェーズを何度も繰り返すことになります。最終的にシステム化フェーズで機械学習プログラムをシステムに組み込んで運用します。システム運用後も追加データでの再学習や精度評価を繰り返し行う必要があります。

► 機械学習プロジェクトの全体像

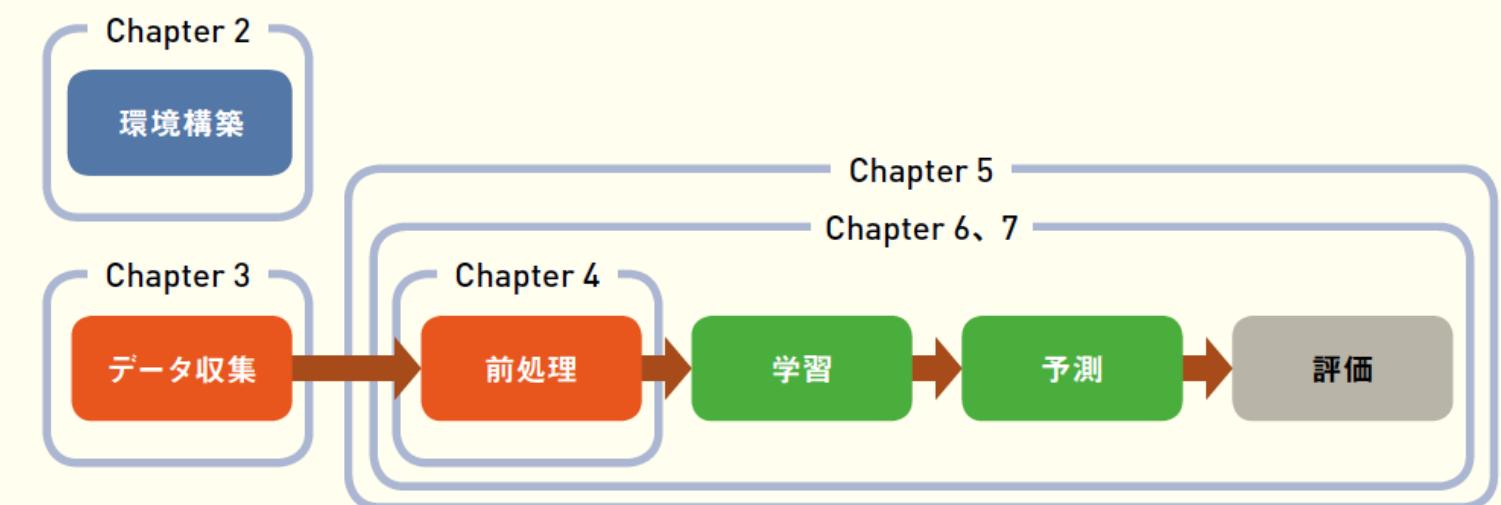


→ 本書の読み進め方

ここまで機械学習の全体像や各要素について説明をしてきました。しかし、本書の中ではデータ収集→前処理→学習→予測→評価というような全体の流れを使用した演習は行いません。このような手順を踏むと、最終的な成果物ができるまでの手順が長すぎるため、本書の最後まで進めてやっと1つの機能ができる形式となってしまいます。本書ではそ

の代わりに、機械学習プロジェクトで必要となる要素を使用した小さなプログラムを、各Chapterで作成していきます。Chapter 2で機械学習プログラミングのための環境を構築したあとは、興味のあるChapterから実践してみてください。なお、Chapter 6、7は連続しているため、続けて実践することをおすすめします。

► 各Chapterの範囲



► 各Chapterの概要

| Chapter | キーワード | 概要 |
|-----------|--------------|------------------------------------|
| Chapter 2 | 環境構築 | 機械学習プログラミングのための環境を構築する |
| Chapter 3 | データ収集 | Webスクレイピングを使用してWebページからデータを収集する |
| Chapter 4 | 前処理 | 日本語テキストの形態素解析を行い、文章を自動生成する |
| Chapter 5 | 前処理、学習、予測、評価 | 数字が書かれた画像を教師あり学習（分類）し、手書き文字を認識する |
| Chapter 6 | 前処理 | オープンデータを機械学習が行える形式にデータ変換する |
| Chapter 7 | 学習、予測、評価 | Chapter 6で作成したデータを元に、教師あり学習（回帰）を行う |



Chapter8
本書を学んだあとの
次のステップ

機械学習の学習
Webサイト、書
籍、コミュニティ
Lesson.63,64

AI-bot開発

サンプルプログラム
pybotの
起動と仕組み
Lesson.17,18

スクレイピングで取
得したデータの検索
コマンド作成
Lesson.25

Chapter2 機械学習の 開発環境準備

Python
インストール
Lesson.12

仮想環境準備
(venv)
Lesson.13

Jupyter Notebook
インストールと使い方
Lesson.14,15

機械学習とは何か、
注目される理由
Lesson.1, 2

機械学習関連技術
前処理、データ収集
前処理、精度評価
Lesson.3,4,5,9

機械学習の手法
アルゴリズム
Lesson.6,7

データ収集 Chapter3 スクレイピング

スクレイピングの
基礎知識
ライブラリ
Lesson.19,20

スクレイピングの
実行、データ保存
実行時注意点
Lesson.21,22,23,24,26

PoC
Lesson.8

機械学習システムを
運用する仕組み
Lesson.10

本書での学習の仕方
Lesson.11

自然言語処理 Chapter4 日本語文章生成

分かち書き
Lesson.27

形態素解析
(Janome)
Lesson.28

自然言語処理の
アルゴリズム
(Bag of Word,TD-IDF)
Lesson.29

マルコフ連鎖
Lesson.30

日本語データ用意
(青空文庫,SNS)
Lesson.31

マルコフ連鎖用
辞書データ作成
Lesson.32

マルコフ連鎖での
文章自動生成
プログラム
Lesson.33

文章自動生成
前処理
Lesson.34

マルコフ連鎖用
辞書データ生成
Lesson.35

マルコフ連鎖による
文書自動生成
コマンド作成
Lesson.36

機械学習 Chapter5 手書き文字認識

手書き文字認識
の基本
(教師あり学習+分類)
Lesson.37

機械学習のライブラリ
インストール
(scikit-learn,
Pillow,NumPy,Matplotlib)
Lesson.38

UCI手書き数字データ
セットによる機械学習
Lesson.39,49,41

自分で手書きした
文字の予測(分類)
Lesson.42,43,44,45

分類モデルの
精度評価と比較選択
Lesson.46,47

学習済みモデルの作成
Lesson.48

手書き文字
分類コマンド作成
Lesson.49

Chapter6 表形式データ の前処理

表形式データの処理
(pandas)
Lesson.50~55

データの可視化
(気温データのグラフ
表示)
Lesson.56

気温データ
検索コマンド作成
Lesson.57

Chapter7 データ予測 回帰分析

回帰分析の基本

線形単回帰分析
(緯度から
気温を予測)

線形重回帰分析
(緯度と高度から
気温を予測)

回帰分析モデルの評価

気温予測コマンド作成

<https://shacho.beproud.jp/entry/ichiyasa-pythonml>

この内容になった経緯

2018年1月10日に打診あり

January 10th, 2018



Iwohtsu 18:20

@takanory @shimizukawa

あけましておめでとうございます。本年もどうぞよろしくお願ひ申し上げます。

新年早々ですが、実はまたインプレスさんの書籍でご協力いただきたい案件がありまして、近々打ち合わせをお願いできないでしょうか.....?

企画しているのは次の2つです。

- ・ビジネスで使えるPythonスクリプト集（目標5月刊。前後の構成などがないので多人数でワッとやりやすい書籍ではないかと思います）
- ・いややさPythonの続編（2018年中）



takanory 🍻 18:20

お

とりあえず打ち合わせ了解です

社内的にGO!ができる

- 2018年1月12日: 編集者と打合せ
- 2018年1月18日: 社内でGoができる
- 2018年1月24日: 平井がjoin
- 2018年2月1日: 第2回目打合せ
- →まずはネタ出しを進める



5 2 ~ 100% 1 Arial 26 B I U A 11 12 13 14 15 16 17 18

編集

31

アイデア出し

- 機械学習は、次のステップにしてはハードル高すぎな気がする(nakagami)
- scikit lern だったら、あんまり理論わからなくても、なんかそれっぽく判断してくれる感があるの？(nakagami)
- pandas で集計なら、Excel の代わりっぽいことができるってイメージがあるかな(nakamiga)
- 理論には踏み込まない？(xiao)
 - 最低限の説明はする、しないと面白くないのでする、ただできるだけしたくない(takanory)
 - PyPro3 的な？(shiro)
 - そんな感じ(takanory)
- とはいえるやめやつても面白くないよね(takanory)
- いや続編Pythonで会話するbotプログラムを作ったので、それを拡張して対応できるコマンドを作りたい(takanory)
- Web API になってるサービスは使いたくない (Watson 的なやつ) (nakagami)
 - 画面ぱちぱちするだけになる(takanory)
 - API系はデベロッパー登録が大変なので、使うとしても1種類だけかなと(takanory)
 - bot を人間っぽくするために MS の AI の API を使うとか(xiao)
 - API 叱くだけだと、天気予報と同じ(takanory)
- 次のステップ
- スクレイピング -> データ分析いい(shiro)
 - 不動産のお買い得物件を探すというのがあった(shiro)
 - 機械学習を使って東京23区のお買い得賃貸物件を探してみた
 - <http://www.analyze-world.com/entry/2017/11/09/061023>
- オープンデータでいいものがあるか
- グラフはやってる感がある(nakagami)
 - プログラミングで何かやってる感
 - jupyter notebook は変化が激しくて、本に向いてない(takanory)
 - IPythonでもいいかも(takanory)
 - deprecation warning で心配になる(takanory)
- Wookieepedia: スターウォーズの辞書(takanory)
 - <http://ja.starwars.wikia.com/wiki/%E3%83%A1%E3%82%A4%E3%83%B3%E3%83%9A%E3%83%BC%E3%82%B8>
- kaggle で面白データないかな(takanory)
- xiaoの興味は?(takanory)
 - sushi, 銀英伝(xiao)
 - あと興味あるのはPinterest(xiao)

企画案を編集に伝える

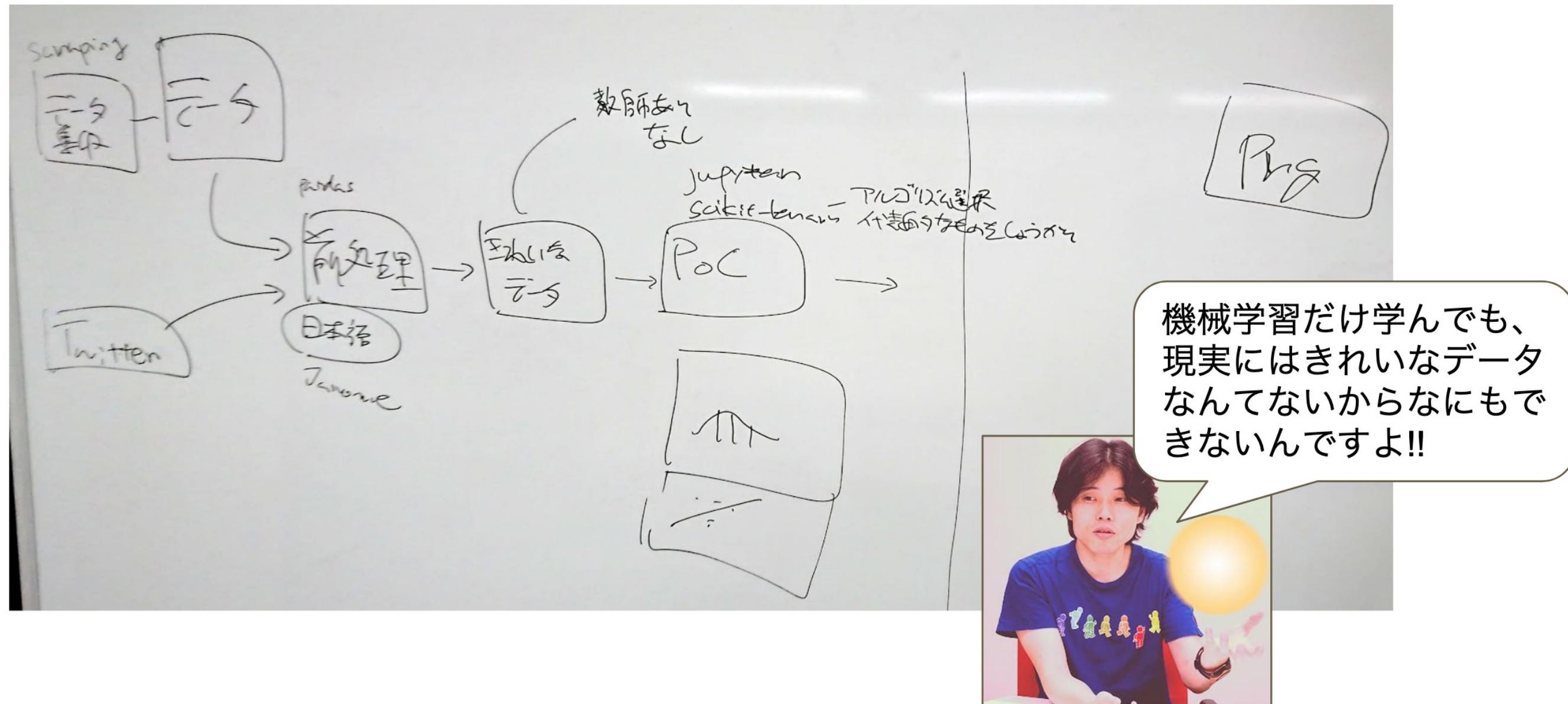
- 2018年7月6日
- 環境構築、pybot準備
- スクレイピングでデータ収集
- マルコフ連鎖で文章生成
- 手書き文字認識
- 前処理と予測
- 回帰分析
- 次のステップ



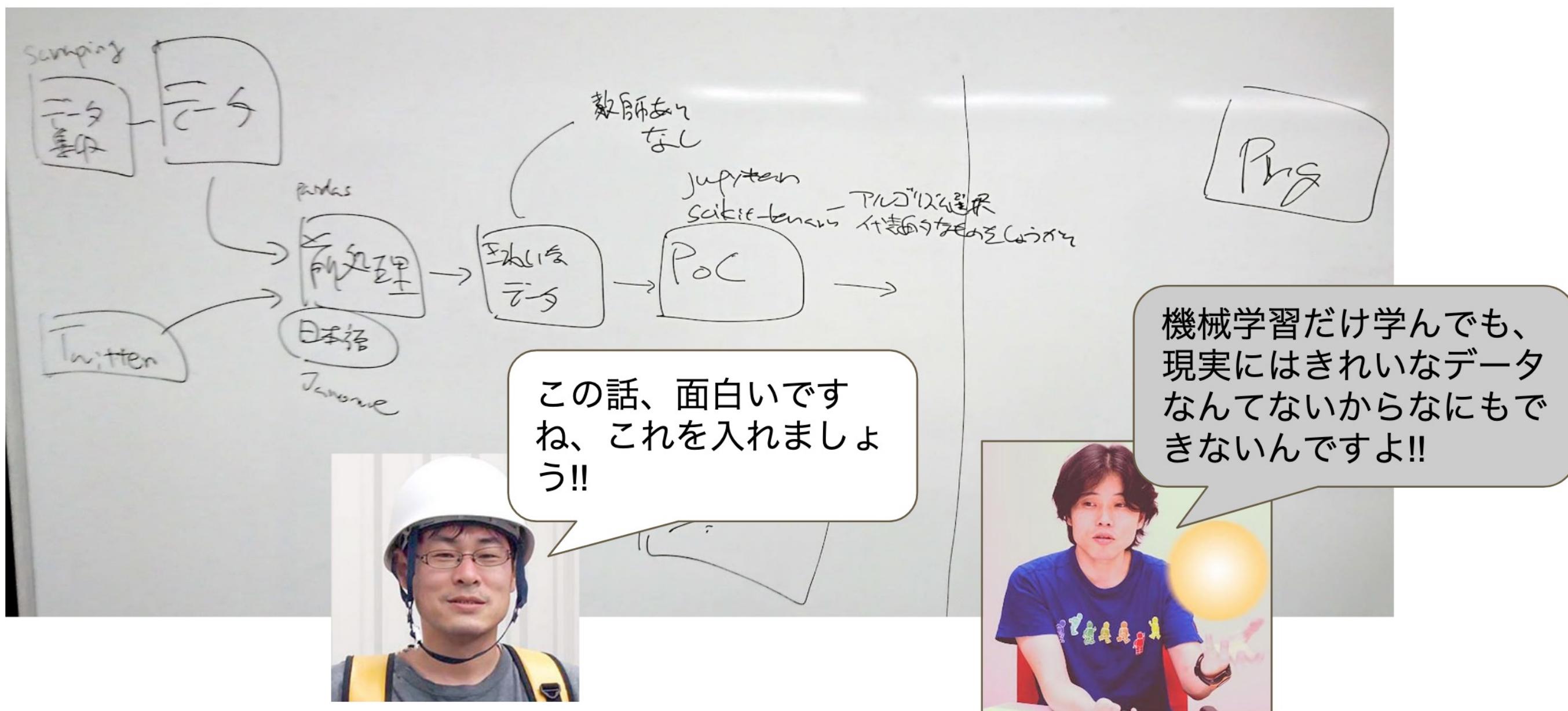
なんかバラバラでつながら
ないんですよねー

リブロワークス大津さん

ホワイトボードで各要素の関連を説明

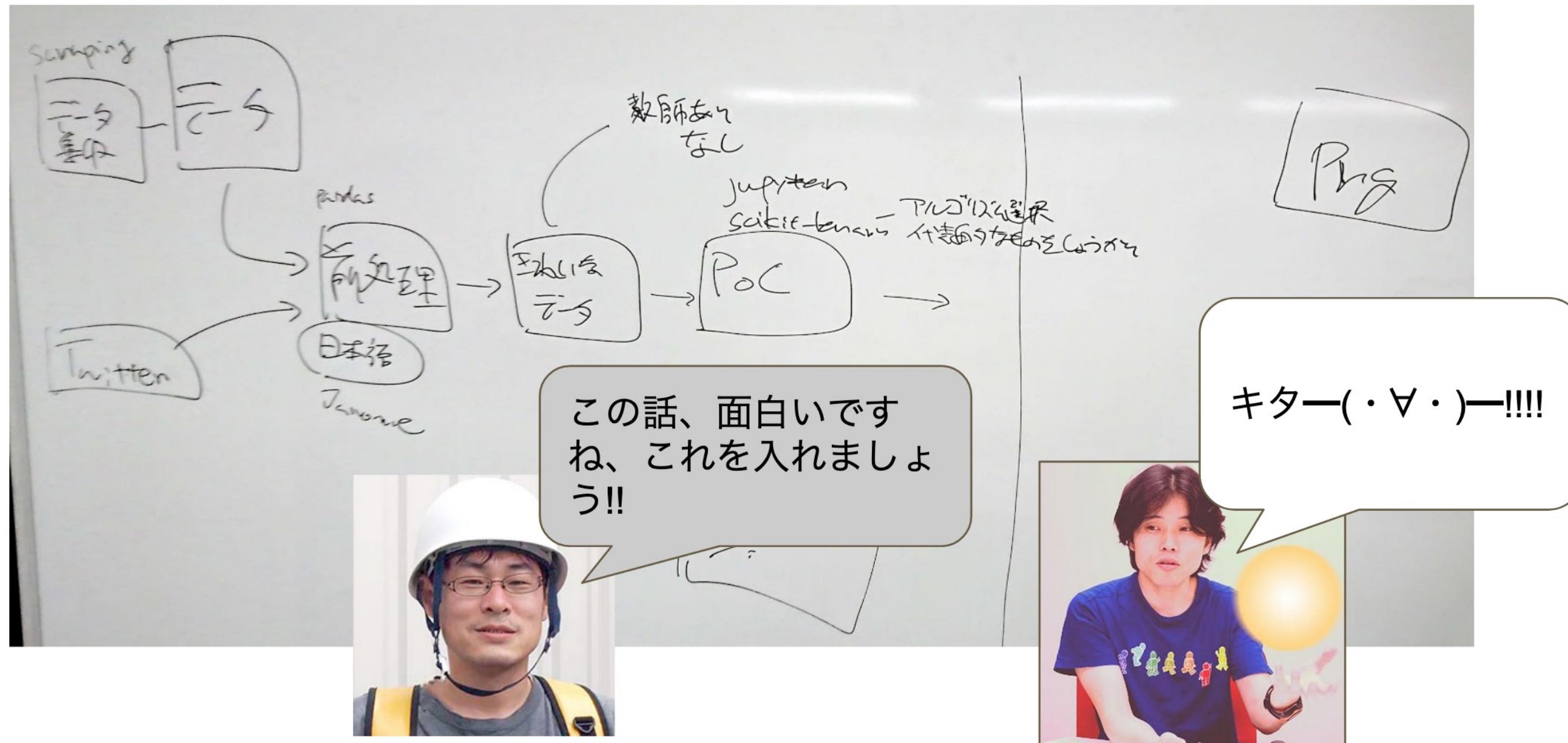


ホワイトボードで各要素の関連を説明



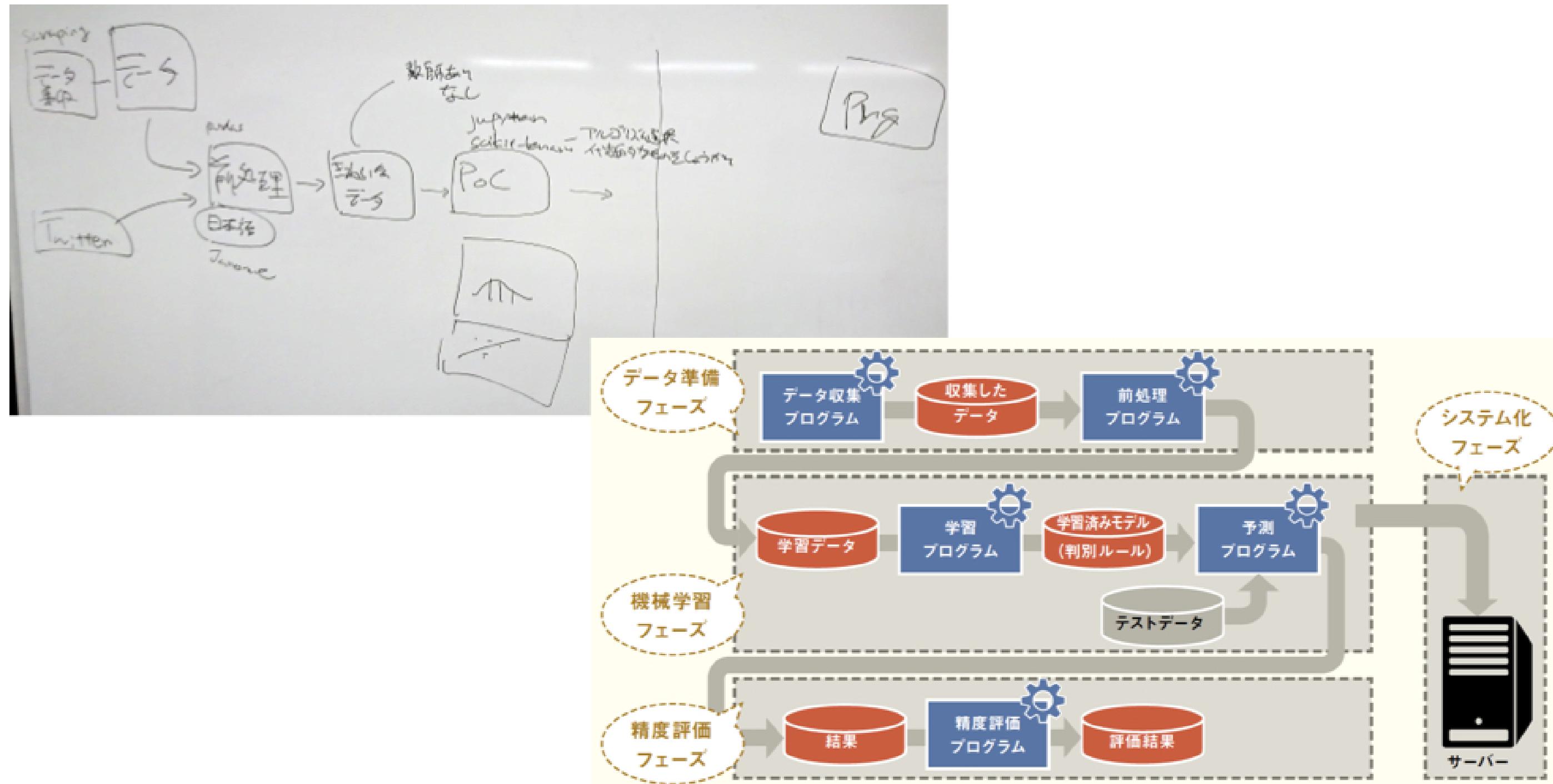
リブロワークス大津さん

ホワイトボードで各要素の関連を説明



リブロワークス大津さん

そして現在の構成に



長い戦いがはじまった!!!



- 2018年7月31日: 降旗がjoin
- 2018年8月: 執筆開始
 - 8ヶ月に渡る戦い
- 2019年4月5日: 脱稿
 - 当初は1月中旬で3月発売予定...

そして出版へ

- 2019年4月22日: 著者初校、校閲
- 2019年5月11日: 再校作成(編集)
- 2019年5月20日: 著者再校
- 2019年5月23日: 念校・仕上げ
- 2019年5月27日: 印刷所入稿
- 2019年05月30日: 校了
- 2019年06月21日: 発売!!



機械学習を学習する課題

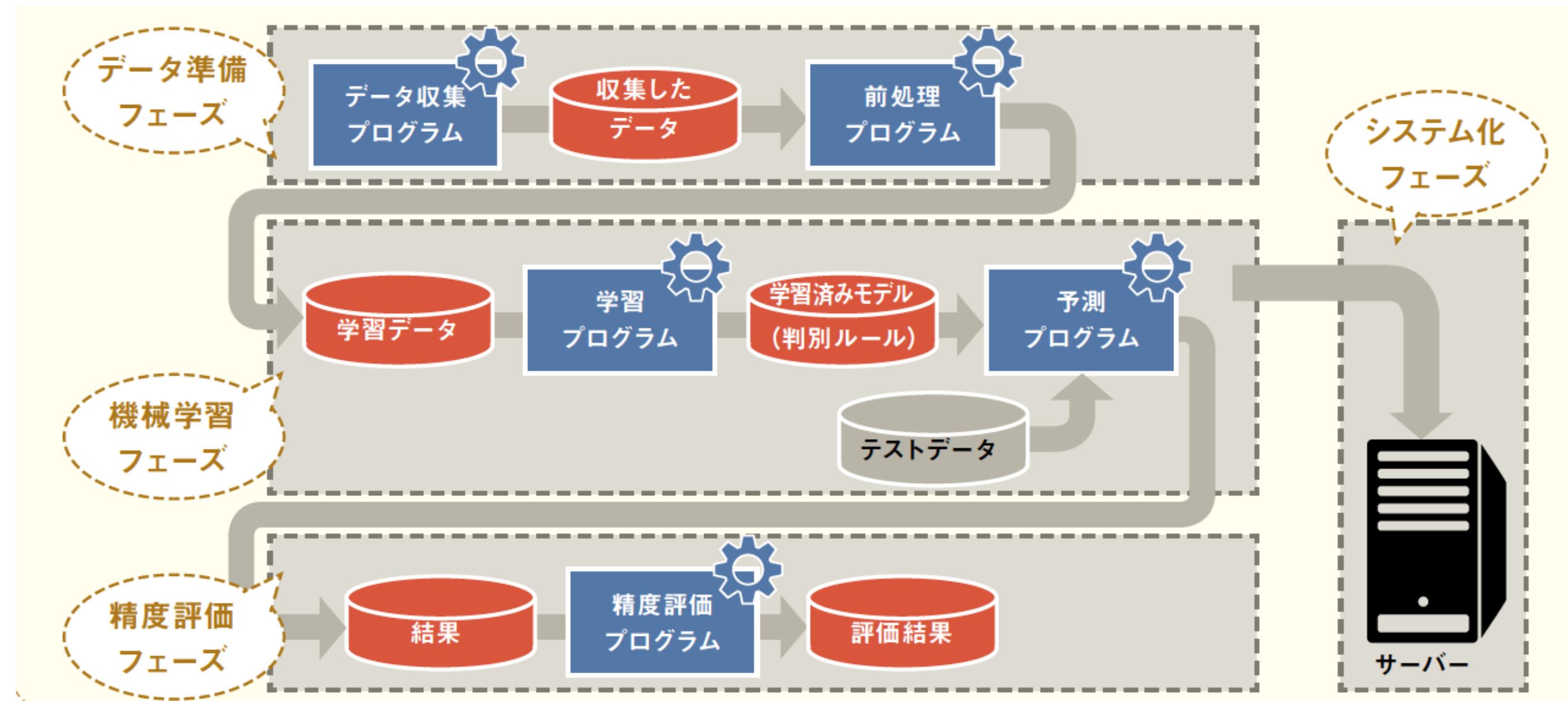
(注)個人的な意見であり、特定の書籍、人物などを批判するものではありません。

課題: 機械学習だけじゃなにもで きない

課題: 機械学習だけじゃなにもで きない

- 課題の設定
- データ収集
- 前処理

そこで: 全体の流れを解説

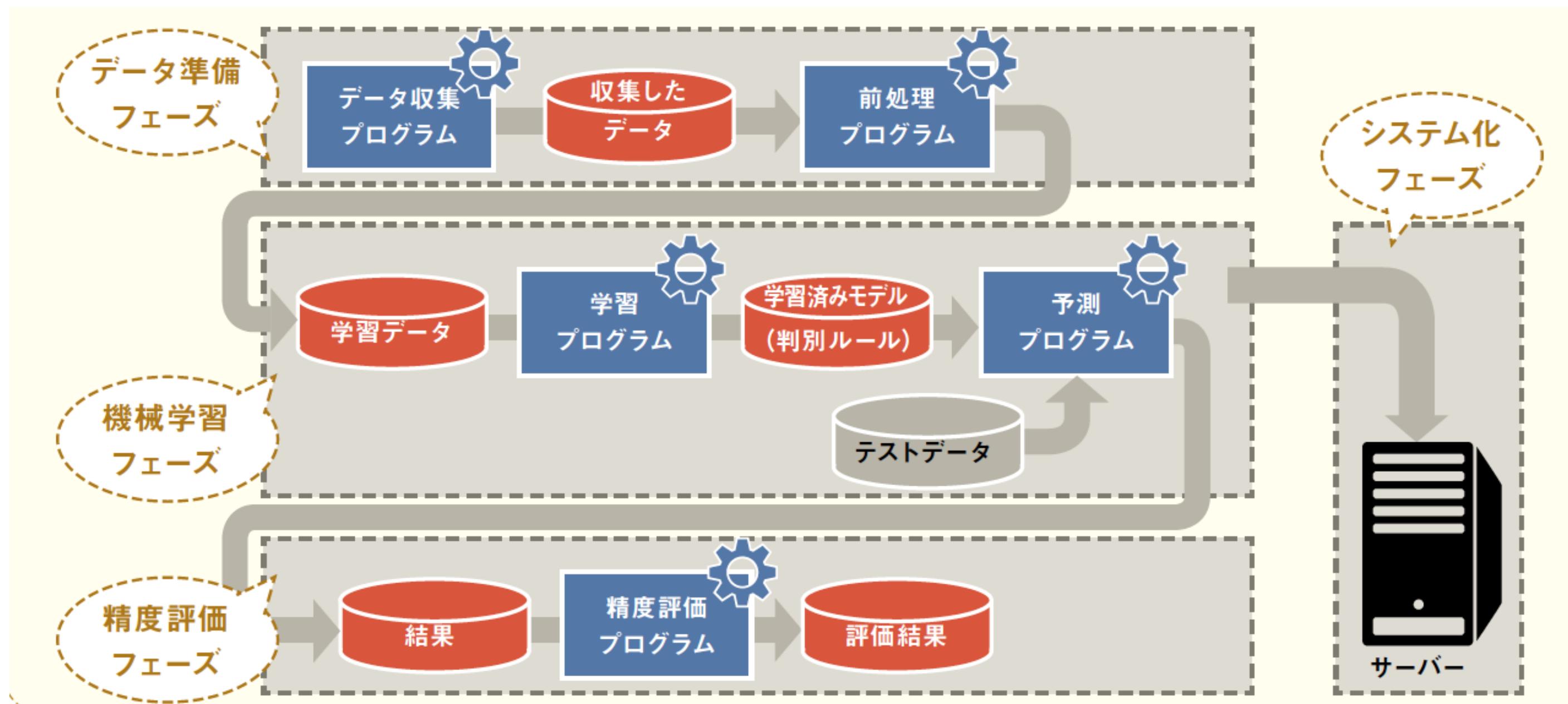


そこで: 周辺技術も体験

- スクレイピングでデータ収集(Ch.3)
- 日本語の形態素解析(Ch.4)
- データの前処理(Ch.6)
- 精度の評価(Ch.5, 7)

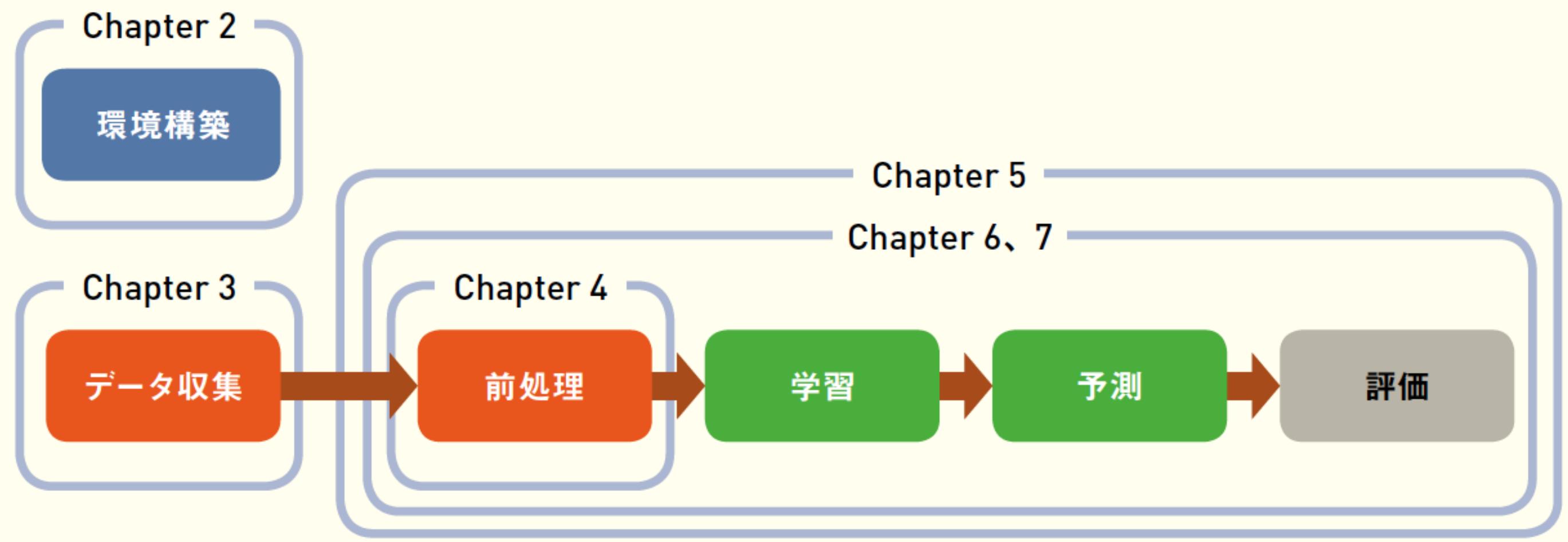
課題: 最後までがとても長い

課題: 最後までがとても長い



そこで: 断片的に体験できる

▶各Chapterの範囲



課題: 達成感がない

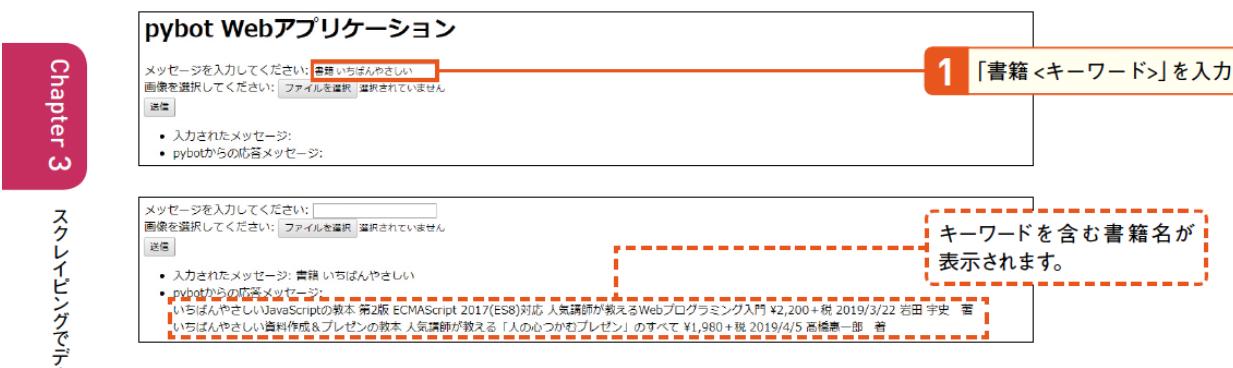
課題: 達成感がない

- ・コードを実行してグラフ表示→で?

そこで: botでアプリケーション化

5 pybot Webアプリケーションから書籍検索コマンドを実行する

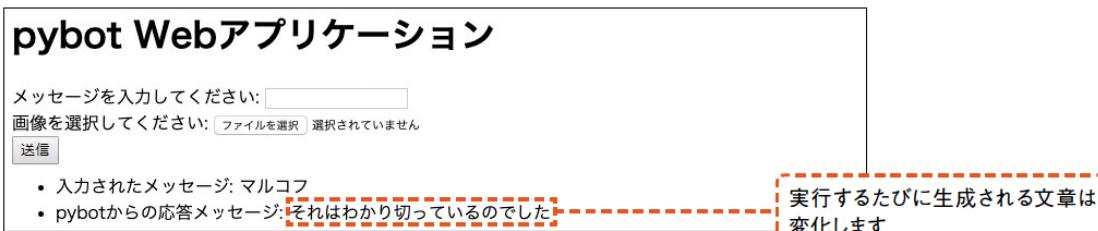
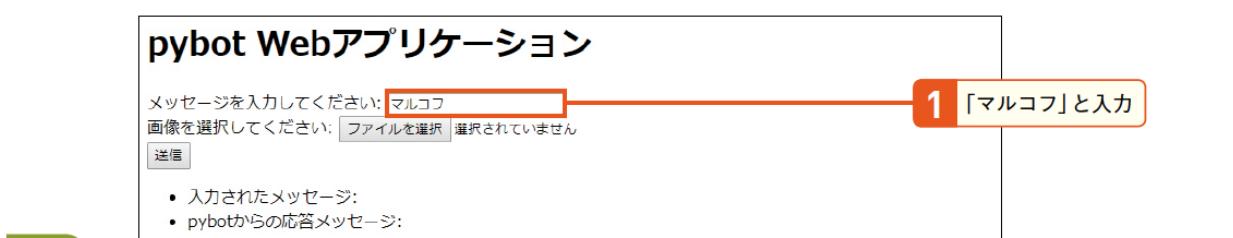
コマンドプロンプト上で「python pybotweb.py」を実行してpybotサーバーを起動します（Lesson 18参照）。書籍検索コマンドが実行され、検索結果が表示されます。ブラウザで「<http://localhost:8080/hello>」にアクセスして、pybot Webアプリケーションの画面を表示し、



5 pybot Webアプリケーションからマルコフコマンドを実行する

コマンドプロンプト上で「python pybotweb.py」を実行してpybotサーバーを起動します。ブラウザで「<http://localhost:8080/hello>」にアクセスして、pybot

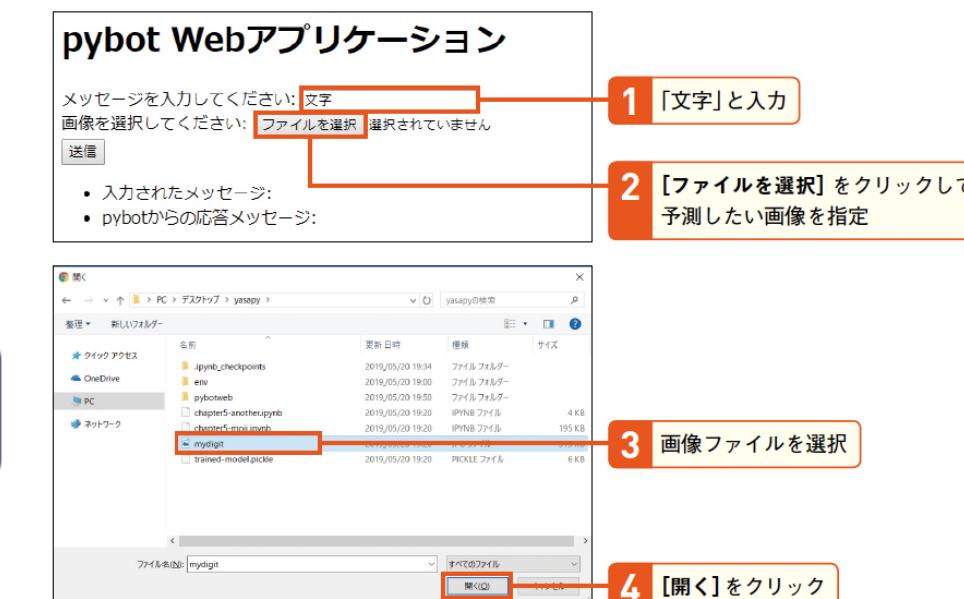
Webアプリケーションの画面を開き、「マルコフ」と入力して送信しましょう。マルコフコマンドが実行され、自動生成された文章が表示されます。



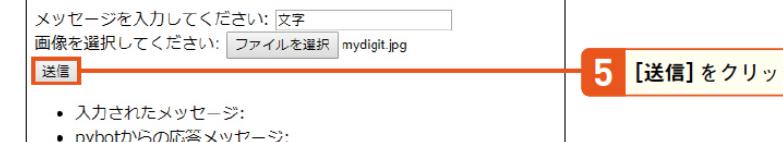
4 文字コマンドを使ってみる

コマンドプロンプト上で「python pybotweb.py」を実行してpybotサーバーを起動します。Webブラウザ上で「<http://localhost:8080/hello>」を開いて、pybot Webアプリケーションの初期画面を開きます。

テキストボックスへ「文字」と入力して①、ファイルに認識させたい手書き文字のファイルを選択します②③④。入力できたらフォームを送信すると⑤ pybotから応答が返ってきます。



pybot Webアプリケーション



pybot Webアプリケーション



そこで: botでアプリケーション化

Chapter 6

表形式のデータを前処理しよう

5 気温データコマンドを使ってみる

コマンドプロンプト上で「python pybotweb.py」を実行してpybotを起動しましょう。Webブラウザ上で「<http://localhost:8080/hello>」を開いて、pybot Webアプリケーションの初期画面を開きます。テキスト

pybot Webアプリケーション

メッセージを入力してください: 1 「気温データ 東京」と入力
画像を選択してください: ファイルを選択 指定されていません
送信 2 [送信] をクリック

- 入力されたメッセージ: 気温データ 東京
- pybotからの応答メッセージ:

pybot Webアプリケーション

メッセージを入力してください: 1 「気温データ 35.7」と入力して送信
画像を選択してください: ファイルを選択 指定されていません
送信

- 入力されたメッセージ: 気温データ 35.7
- pybotからの応答メッセージ: 平均気温は23.4度です

東京の平均気温が表示されます。

4 緯度を入力して予測された気温をみる

緯度がわかったので、いよいよpybotに気温を予測させることができます。神保町の緯度は約35.7でした。「気温データ 35.7」と入力して予測された気温を

見てみましょう①。「タブン24.4度クライ」という応答があれば、気温データコマンドの改造は無事完了です！

pybot Webアプリケーション

メッセージを入力してください: 1 「気温データ 35.7」と入力して送信
画像を選択してください: ファイルを選択 指定されていません
送信

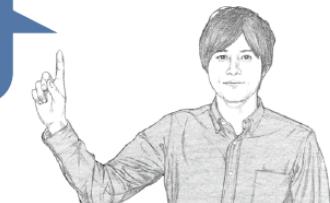
- 入力されたメッセージ: 気温データ 35.7
- pybotからの応答メッセージ:

pybot Webアプリケーション

メッセージを入力してください: 1 「気温データ 35.7」と入力して送信
画像を選択してください: ファイルを選択 指定されていません
送信

- 入力されたメッセージ: 気温データ 35.7
- pybotからの応答メッセージ: タブン24.4度クライ

予測された気温が表示されます。



ここでは既存のデータの中から検索する機能を追加しました。次のChapter 7では既存のデータから未知のデータを予測する機能を追加します。



インターネットでは、住所を緯度経度を変換する(ジオコーディング)APIが提供されています。Chapter 3で紹介したライブラリと組み合わせて、未知の住所の緯度を自動で調べる機能を追加してみると面白いでしょう。

課題: データが面白くない

課題: データが面白くない

- アヤメとかボストンの住宅価格とか興味ある?

そこで：イメージが湧くデータ

○書籍ページから書籍名と値段を取得する

ここでは、前のLessonで取得した『いちばんやさしいPythonの教本』の書籍ページのHTMLをスクレイピングして、「書籍名」と「値段」を取得します。

▶取得する書籍名と値段



1 HTMLを取得する chapter3-scraping.ipynb

まずは前回と同様に、『いちばんやさしいPythonの教本』の書籍ページのHTMLを取得します。そして、そのHTMLをスクレイピングするために、BeautifulSoup

オブジェクトを作成します。次のコードをJupyter Notebookに入力し、Shift+Enterキーで実行してください①。

```
import _requests
from bs4 import BeautifulSoup

res = _requests.get('https://book.impress.co.jp/books/1116101151')
html_doc = res.text
soup = BeautifulSoup(html_doc, 'html.parser')
```

```
In [12]: import requests
from bs4 import BeautifulSoup

res = requests.get('https://book.impress.co.jp/books/1116101151')
html_doc = res.text
soup = BeautifulSoup(html_doc, 'html.parser')
```

..... BeautifulSoupオブジェクトを作成

1 コードを入力して実行

Chapter 3 スクレイピングでデータを収集しよう

○青空文庫からテキストをダウンロードしよう

1 サイトにアクセスして、作家や作品名で探す

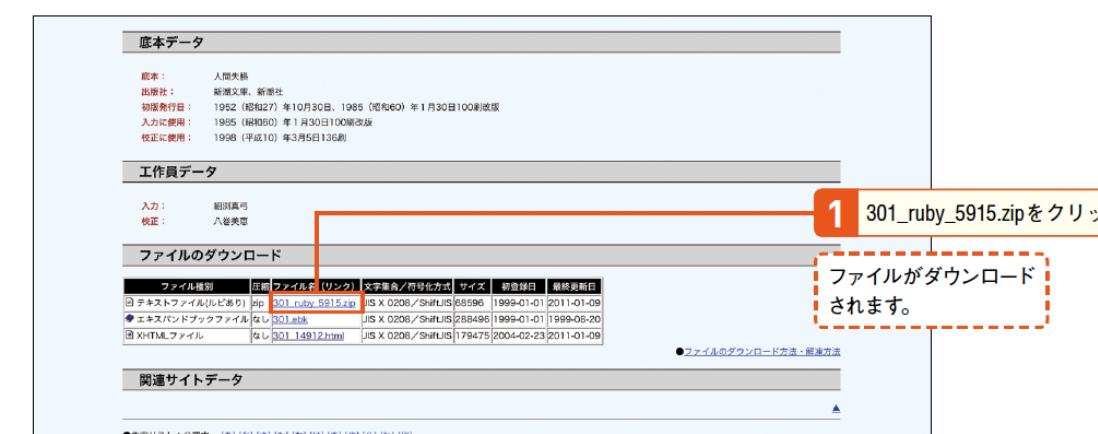
ブラウザを開いて青空文庫 (<https://www.aozora.gr.jp/>) のURLにアクセスします。作家別、作品別にタグして対象となる小説を探します①。ここでは「太宰治」の『人間失格』を選択します。

頭文字で分類されているので、任意のリンクをクリック



2 ファイルをダウンロードする

任意の小説のページを開くと、画面下部に「ファイルのダウンロード」という項目があります。この中の「テキストファイル（ルビあり）」の横にあるファイル名のリンクをクリックしてダウンロードします（作品ごとにファイル名は異なります）①。



Chapter 4 日本語の文章を生成しよう

課題: 用語多過ぎ

課題: 用語多過ぎ

欠損値、外れ値、形態素解析、ラベル、教師あり学習、教師なし学習、分類、回帰、クラスタリング、アンサンブル学習、モデル、線形回帰、ロジスティック回帰、サポートベクターマシン、決定木、ランダムフォレスト、混同行列、真陽性、偽陽性、偽陰性、真陰性、正解率、適合率、再現率、F値、、、

そこで：用語集を用意

④ 機械学習フェーズ

機械学習についてはLesson 6、Lesson 7で解説しています。用語については検索などでも利用することができます。

▶ 機械学習に関する用語

| 用語 | 説明 | 参照先 |
|-----------------------------------|--|---------------------|
| 教師あり学習 (Supervised Learning) | 正解となるデータをもとに機械学習を行う手法。データの分類や数値の予測などに使用する | Chapter 5、Chapter 7 |
| 教師なし学習 (Unsupervised Learning) | 正解が用意されていないデータに対して行う手法。データのクラスターリングなどに使用する | — |
| 強化学習 (Reinforcement Learning) | ある環境の中での行動に対して報酬を与えて学習させる手法。ゲームや自動運転などにおいて、振る舞いを最適化するために使用する | — |
| 分類 (Classification) | 教師あり学習でデータがどのグループに属するか（ラベル）を予測すること | Chapter 5 |
| 回帰 (Regression) | 教師あり学習でデータに対して数値を予測すること | Chapter 7 |
| クラスタリング (Clustering) | 教師なし学習で、似ているデータをグループ化すること。分類とは異なり正解が存在しない | — |
| アルゴリズム (Algorithm) | 機械学習ではそれぞれの機械学習を行うための手順のことを指す。主要なアルゴリズムはscikit-learnで用意されている | Chapter 5、Chapter 7 |
| アンサンブル学習 (Ensemble Learning) | 複数のモデルの結果を組み合わせて多数決などで決定する手法 | — |
| ラベル (Label) | 分類で、データの正解を表す値 | — |
| モデル (Model) | 機械学習アルゴリズムが作成した、予測を行うためのパラメータの集まり。予測プログラムで使用する | — |

▶ 教師あり学習の主な手法

| 手法 | 説明 | 参照先 |
|---|---|-----------|
| 線形回帰 (Linear Regression) | 回帰に使用するアルゴリズムの1つ | Chapter 7 |
| ロジスティック回帰 (Logistic Regression) | アルゴリズムの名前には回帰が付いているが、主に分類に使用するアルゴリズム | Chapter 5 |
| サポートベクターマシン (Support Vector Machine : SVM) | 分類、回帰に使用できるアルゴリズム | — |
| 決定木 (Decision Tree) | データを分割するルールを定義して分類を行うアルゴリズム | — |
| ランダムフォレスト (Random Forest) | 複数の決定木の予測結果から、多数決で予測を行うアルゴリズム。アンサンブル学習の1つ | Chapter 5 |

④ 精度評価フェーズ

PoC（概念実証）についてはLesson 8で解説しています。精度については混同行列の図を見なします。また、精度評価については、Lesson 9で解説します。

▶ 精度に関する用語

| 用語 | 説明 | 参照先 |
|---------------------------|---|------------------------------|
| 学習データ (Data) | 学習済みモデルを作成するため、機械学習アルゴリズムの入力に使用するデータの集まり。あらかじめ用意したデータを学習データとテストデータに分割する。教師データともいう | Lesson 9 |
| テストデータ (Test Data) | モデルの精度評価を行うために使用するデータ | Lesson 9、Lesson 46、Lesson 61 |
| 混同行列 (Confusion Matrix) | 分類の精度を計算するために予測と正解の組み合わせを集計した表 | Lesson 9 |
| 陽性 (Positive) | 分類で目的としているデータの持つ性質 | Lesson 9 |
| 陰性 (Negative) | 分類で目的としていないデータの持つ性質 | Lesson 9 |
| 真陽性 (True Positive : TP) | 陽性と予測して (Positive)、予測が当たった (True) データの性質 | Lesson 9 |
| 偽陽性 (False Positive : FP) | 陽性と予測して (Positive)、予測が外れた (False) データの性質 | Lesson 9 |
| 偽陰性 (False Negative : FN) | 陰性と予測して (Negative)、予測が外れた (False) データの性質 | Lesson 9 |
| 真陰性 (True Negative : TN) | 陰性と予測して (Negative)、予測が当たった (True) データの性質 | Lesson 9 |
| 正解率 (Accuracy) | 全体のうち予測が当たった割合。 $(TP + TN) / (TP + FP + FN + TN)$ | Lesson 9 |
| 適合率 (Precision) | 陽性と予測したうち、実際に陽性だった割合。 $TP / (TP + FP)$ | Lesson 9 |
| 再現率 (Recall) | 陽性のデータのうち、陽性と予測した割合。 $TP / (TP + FN)$ | Lesson 9 |
| F値 (F-Value) | 適合率と再現率のバランスをとった値。適合率と再現率の調和平均で求める | Lesson 9 |

▶ 混同行列



課題: やってみても成果がない

課題: やってみても成果がない

- 「AIでなんかやって」と無茶振り

そこで: データがないとダメだよ

Lesson [データ収集]

04

データ収集について知りましょう



このレッスンの
ポイント

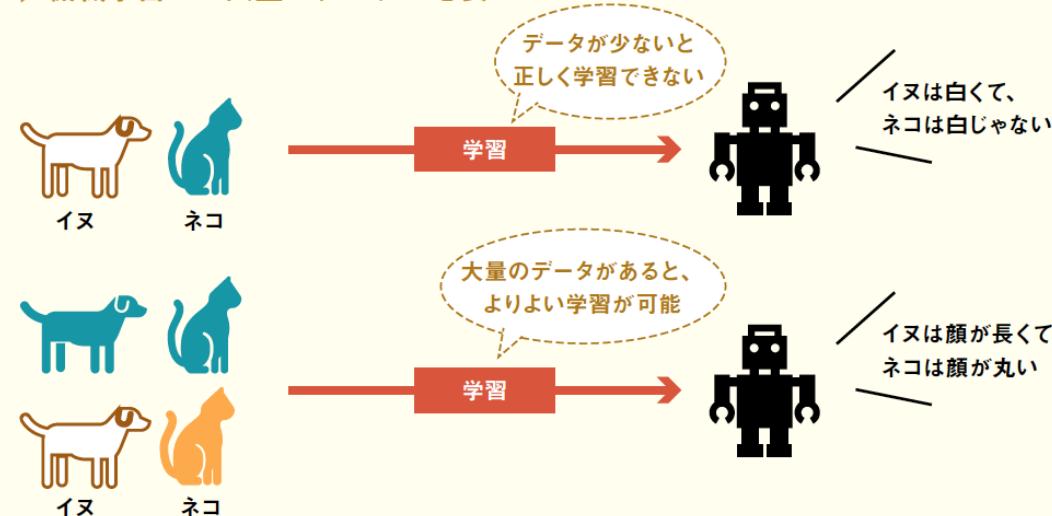
機械学習を行うためには学習の対象となるデータを用意する必要があります。ここではなぜ大量のデータが必要であるのかについてと、大量データを集めためのさまざまな手法について解説します。

→ 機械学習にはなぜ大量のデータが必要なのか

機械学習では大量のデータを集めないと、効果的な学習が行えず、よい学習済みモデル(判別ルール)を作成できません。たとえばイヌとネコの画像を分類するモデルの作成を目標とした場合、イヌとネコの画像が1枚ずつしか用意できなかったら、そのたった2枚を基準に誤った形でイヌとネコを学

習してしまいます。適切な学習済みモデルを作成するためには、大量の学習データを用意する必要があります。その作業をデータ収集と呼びます。なお、必要なデータ量は目的やデータの傾向によって異なります。

→ 機械学習には大量のデータが必要



Chapter 1

Chapter 1

機械学習について知ろう

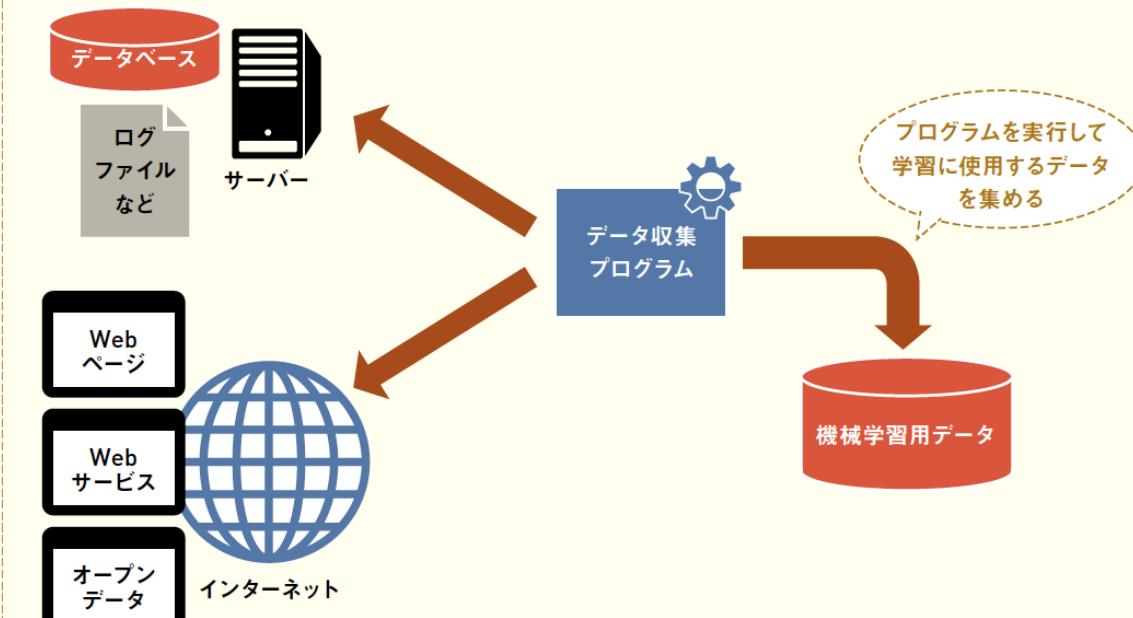
機械学習について知ろう

→ データ収集とは

データ収集とは名前の通り、機械学習に使用するデータを集めることです。機械学習を行うためには大量の学習用のデータが必要となります。あらかじめデータが存在しない場合、データを集めめる必要

があります。データを集めるには各種サーバー、インターネット、公開されているデータセット、データベースといったさまざまなデータソースを利用することが考えられます。

▶ さまざまなデータソースからデータを集め



→ サーバーからデータを取得する

自身が運用しているサーバーでサービスを提供している場合、そのサーバーが利用するデータベースや、ログから各種情報を集めて機械学習に利用できます。たとえば、ECサイトを運営している場合は、データベースから顧客情報と販売情報を取得して、どういう属性の人が何を購入しているか、というよう

な学習用データが作成できます。また、サーバーのアクセスログをもとに侵入を検知したり、CPUなどの各種パフォーマンスに関するログからハードウェアの異常検知を行うといったことが考えられます。Webサーバーへのアクセスログはユーザーの行動分析などに利用することが考えられます。

そこで: 成果が出ない場合あるよ

Chapter 1 機械学習について知ろう

Lesson 08 [PoC: Proof of Concept] PoCについて理解しましょう

このレッスンのポイント

機械学習では収集したデータから目的とする結果が得られるかどうかは試してみないとわかりません。実際に成果が得られるかどうかを検証する作業をPoCと呼びます。ここでは、なぜPoCが必要なのか、具体的に何を行うのか、について理解しましょう。

→ 機械学習におけるPoCとは

PoC（ポックまたはピーオーサー）は Proof of Conceptの略で日本語では概念実証と訳されます。PoC自体は機械学習の専門用語ではなく、新しい概念やアイデアが実現可能かを確認するために、部分的に成果物を作成することをいいます。いきなり実際のプロジェクトを進めて失敗すると多くの費用が発生するため、PoCによって実現可能性をかかります。試供品を顧客が試用した結果を確認する

ことや、新薬を投与して安全性を確認することなども概念実証と呼ばれます。機械学習プロジェクトにおけるPoCとは何を行うのでしょうか？機械学習ではデータを元に目的となる成果、たとえばイヌとネコの画像をどの程度正しく分類できるかは学習させてみないとわかりません。成果が得られそうかどうかを実際にシステム化する前に検証することが、機械学習でのPoCの目的です。

→ PoCによって事前に検証する

これは「イヌ」です → うまく分類できそう
これは「ネコ」です → 似ているので、うまく分類できなそう
これは「ネコ」です → 似ているので、うまく分類できなそう
これは「トラ」です → 似ているので、うまく分類できなそう

→ PoCでは何を行うのか

PoCでは、実際に「学習用のデータを用意し、前処理を行い、学習する」という一連の処理を行い、結果の精度を評価します。精度がPoCでの目標値に達していない場合はデータを見直したり、アルゴリズムを変更したりといった改善を行います。機械学習プロジェクトを進めるべき（またはやめるべき）かを判断するための検証を行います。

→ PoCは繰り返し行う

→ PoCを行う場合に注意すること

本書では紹介のみでPoCの実作業については説明しませんが、機械学習プロジェクトを行う場合にはPoCは重要な作業となります。ここではPoCに関する注意点をまとめておきます。PoCは試行回数を増やすほどよい成果が得られる可能性が高いです。しかし、無限に時間を使うことはできませんから、あらかじめ期間を決めて、期間内に成果が得られるかを判断すべきです。期待する成果が得られない場合は、ビジネス要件を見直す必要があります。ときにはプロジェクト全体をあきらめる場合もあります。

機械学習プロジェクトでは、データから想定する成果が得られるかを、PoCで事前に検証しましょう。

Chapter 1 機械学習について知ろう

そこで：費用対効果が大事だよ

Lesson
09

[精度評価]

機械学習の精度について理解しましょう



このレッスンの
ポイント

Lesson 8で成果が得られそうかどうかを検証するPoCについて説明しました。この成否を決める指標となるのが「精度」です。精度の求め方についてはあのChapterでも何度か説明しますが、ここでは共通する基本的な考え方を説明します。

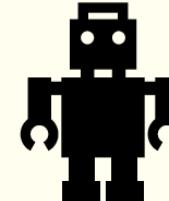
→ 精度がプロジェクトの成否を決める

機械学習で作成した予測プログラムは100%正解を導き出すことはまずありません。どのくらい正確に予測できているかを表す精度という指標を使用して、どのアルゴリズムがよいか、といった比較を行います（精度は0から1の間の数値で表します）。

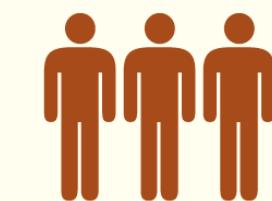
PoCでは、精度の値をもとにプロジェクトがビジネス的に意味があるか（費用対効果があるか）を判断します。たとえば、人が行っていた作業を機械学習プログラムに置き換える場合、人件費の減少が見込めるので、人が行うときより結果が多少悪くても問題ない、ということも考えられます。

▶ 精度がわかれれば費用対効果を割り出せる

機械学習で
精度90%



VS
3人がかりで
精度100%



精度を指標にして、「精度が多少落ちても機械学習を導入するメリットがある」のか「精度が落ちるから見送るべき」なのかを判断します。



036

Chapter 1

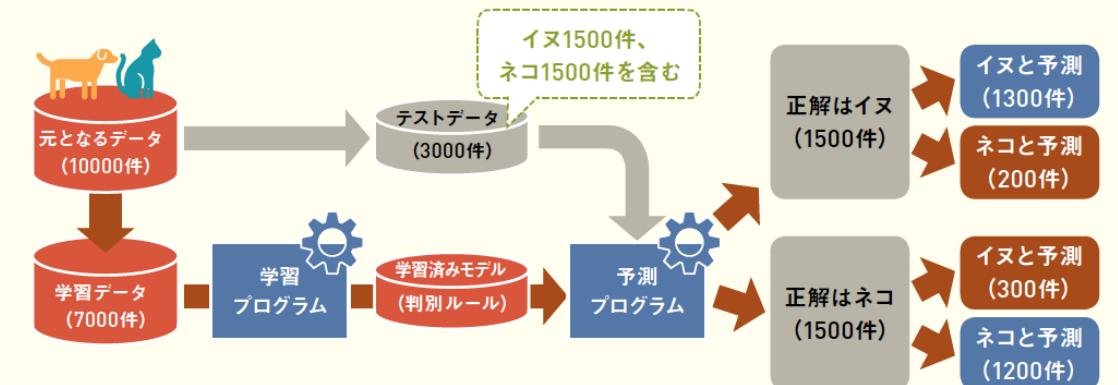
機械学習について知ろう

→ 精度を求めるための準備

教師あり学習で精度を求めるためには、学習して予測を行い、その結果を使用します。一般的に機械学習では、用意したデータを学習データ（教師データともいいます）とテストデータに分割して使用します。学習データで機械学習を行い、そこから作

成した予測プログラムを利用して、テストデータに対して結果を予測し、正解と不正解の数を数えます。次の図では右端の青い四角が正しく予測された結果、赤色が間違って予測した結果です。

▶ データを分割して、正解／不正解を数える



→ 予測結果を混同行列にまとめる

次の図のように、正解と予測の組み合わせがそれぞれ何件あったかをまとめた表を「混同行列」といいます。この例では「イヌとネコの画像からイヌを見つける」というシナリオを想定して、イヌを陽性（Positive）、ネコを陰性（Negative）として扱います。また、正解と予測が合っているもの（青色）はTrue、

異なるものの（赤色）はFalseといいます。この図を使用して、機械学習でよく使われる精度指標をいくつか紹介します。精度の計算式では4つの枠を表す略語（TP、FP、FN、TN）を使用するので、合わせて覚えましょう。

▶ 混同行列

| 予測はイヌ（陽性） | 予測はネコ（陰性） |
|--|--|
| 正解はイヌ（陽性: Positive） | 正解はネコ（陰性: Negative） |
| イヌの画像をイヌと予測(1300件) TP (True Positive) | イヌの画像をネコと予測(200件) FN (False Negative) |
| ネコの画像をイヌと予測(300件) FP (False Positive) | ネコの画像をネコと予測(1200件) TN (True Negative) |

037

Chapter 1

機械学習について知ろう

課題まとめ

- なにもできない→周辺技術を知って体験
- 最後までが長い→断片的に体験
- 達成感がない→botのコマンド化
- データ面白くない→興味が湧くデータ
- 用語多過ぎ→用語集とLessonの対応
- 成果でない→データ量、成果でないことがある、費用対効果

まとめ

- 書籍の紹介
- この内容になった経緯
- 機械学習を学習する課題

Thank You!

- Twitter: [@takanory](https://twitter.com/takanory)
- Slides: github.com/takanory/slides

おまけページ

執筆時の工夫

アウトラインでレビュー

- 執筆初期にアウトライン状態で著者レビュー
 - 見出し、何を書くつもりか箇条書き
- ストーリー的によさそうかをチェック
- 説明の抜け漏れを防ぐ
- 全体的な流れが統一しやすかった

レビュー説明会

- BeProud社内から17名のレビューが参加
 - 機械学習チョットテ“キル/テ”キナイ勢
- レビューの観点、心の持ち方を説明
 - (日本語の)バグを憎んで人を憎まず
 - 思ったらとりあえず書く
 - 指摘が採用されなくても気にしない
- レビューシートでバランス調整

Adobe Document Cloud

- 前半はPDFをDropboxでレビュー
- コメント增多ると重たい問題
- Adobe Document Cloud超便利!!!
 - 依頼者は有料アカウント必要
 - Acrobat Readerでレビューできる
 - 激列に軽い!
 - オフラインでレビューできる!!

Thank You!

- Twitter: [@takanory](https://twitter.com/takanory)
- Slides: github.com/takanory/slides