データサイエンス特論
Data Science

Introduction

1. データサイエンスとは
2. Pythonとパッケージ
3. Jupyter NotebookとColaboratory
4. 授業の進め方とレポート書き方

本資料は、1～12回目（担当：橋本）の説明です。
授業の進め方はmanaba「掲示板」（授業計画）をご覧ください。
13～15回目（担当：大久保）の授業の進め方は別途指示があります。

(C) 創造技術コース　橋本洋志／大久保友幸
{hashimoto, ohkubo-tomoyuki}@aiit.ac.jp
http://hhlab.org/

自己紹介

研究分野：
- 知能機械学，ロボット工学，システム制御工学
- 知能情報処理，画像・信号処理
- 認知科学，行動心理学
- 福祉・医療工学と機器の設計開発
- ヒューマンインタフェース，学習支援システム
- 中小企業支援工学，加工・物流のIoT導入

PBL情報	http://hhlab.org/ ⇒ “PBLチーム活動”
科学研究費	https://nrid.nii.ac.jp/ja/nrid/1000060208460/
researchmap	https://researchmap.jp/captain_hashimoto/

我が国から公的研究予算を受ける
研究者は全員登録

データサイエンスとは

データサイエンスとは

科学（science）とは

- 科学（science）とは，ある領域を対象にして**科学的方法**により知識体系を築き上げる研究活動を言う。

科学的方法とは

- 問題の発見，仮説の設定，それを測る手段，実験による観察・データなどに対する客観的な分析，考察，結論を導くこと。

データサイエンスとは

- データサイエンスは「データを科学的に扱う」学問分野である。
- すなわち，科学的方法（様々なデータの収集，可視化，分析と解析，マイニング，評価，考察など）により，仮説発見・仮説検証を通して，データの産み出されたメカニズム（原因，因果性，モデルなども含む）を明らかにして，その知識体系を築くことである。さらに，意思決定や行動に役立たせることも行う。

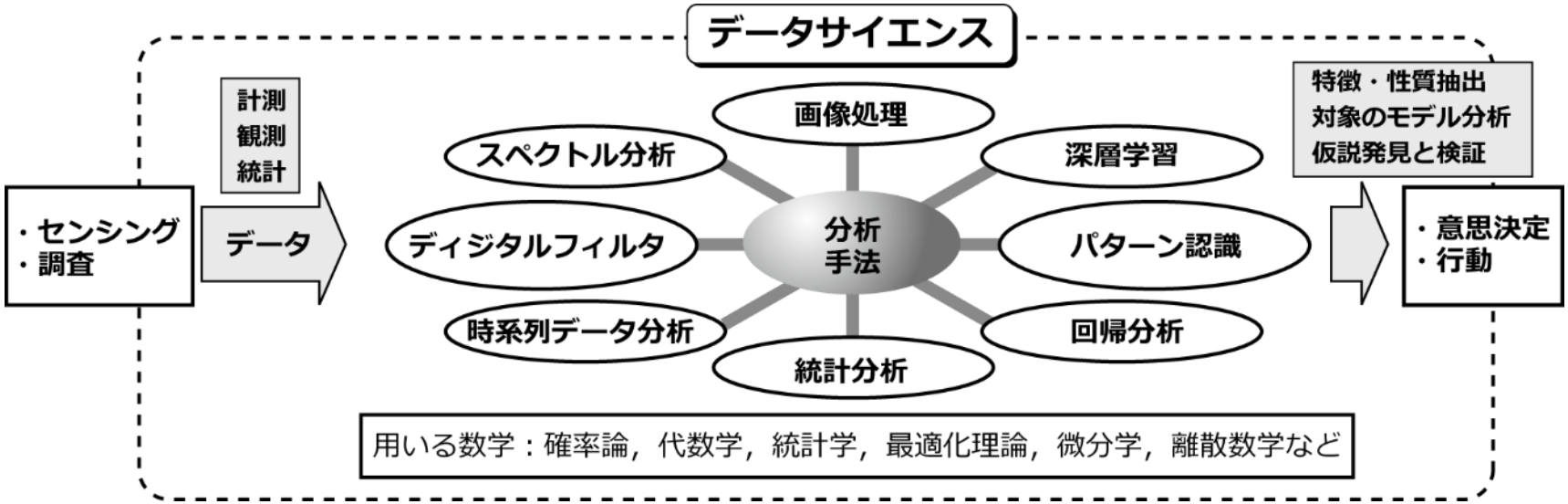
データサイエンティストの必要性

- 超スマート社会の実現に必要，キーワード：インダストリー4.0，Society 5.0,データサイエンティストの育成が必須
- http://iot-jp.com/iotsummary/iotbusiness/科学技術白書「超スマート社会」/.html
- http://monoist.atmarkit.co.jp/mn/articles/1610/12/news034_2.html

データを見る眼を養う

- **夕張市，病院数を減らしたら市民が元気になった**
  - 相関関係は，統計から言えたとすれば，事実である
  - 因果関係（メカニズムともいう）はあるのか？
  - 誤解を誘導する表現
- **ワインの味のグレード（grade）は成分から数値で表現できる**
  - ソムリエは必要なくなる？　（最近のAI論調に近い言い方）
  - センシング技術は？
- **TVの視聴率15%，内閣支持率50%，緊急電話調査50%**
  - この数字をどう理解する？　どう表現する？
  - サンプリングの方法を吟味すべき（レコーダは？　特定支持層の隠れ，等）
- **電力需要や雨を予測**
  - なぜ，予測が必要？
  - 何時間先が重要？

データサイエンスの領域と役割



備考

- パターン認識は工学を起源とする。機械学習は計算機科学の分野から生じた。両者の交わる分野は多く，このため，最近，機械学習の中にパターン認識を取り入れた表現が多い。実際，狭義のパターン認識は機械学習の一つとして捉えることができる（本「パターン認識と機械学習」より）。
- 機械学習とは，現在において，「データの中で、見えているものから、見えていないものを予測する（ことを、コンピュータにやらせるための）」技術と定義されている。
- システム同定は，時系列データを離散時間の動的モデルに当てはめ，どのモデルの性質を考察するものであり，AICという統計学で重宝している指標を産み出したり，雑音データの下でのシステム制御や金融分野での有力なツールを提供している
- ディジタル信号は，ここでは画像データも含んでいる。ディジタル信号処理はセンサ／IoTの出力部の扱いで必須であり，かつ，他の分野と多くの点で関わっている。
- 本書では，扱う内容がわかりやすいよう，目次の書き方を，これまでの分類になるべく沿った。
- 数学の厳密な展開や証明は他書に譲り，必要最小限な記述に留めている。

データサイエンティストの心構え

データの取得，分析，考察に科学的方法を取り入れる

- 客観性，再現性，定量的表現，仮説発見・設定，その検証，など

巨人の肩の上に立つ

- 意味，Wikipedia　https://ja.wikipedia.org/wiki/巨人の肩の上
- 我々は，巨人の肩の上に立って，独創性，創造性を発揮できる

独創性

- 武部啓先生の言葉:「科研費申請の正攻法とコツ」,羊土社,1998
- 本研究（仕事，開発）はきわめて独創的であると言える
- 世界あるいは日本国内で、第1人者あるいはそれに並ぶ高い評価を受けている研究であること。招聘、広く読まれている、等[中略]
- 自分が樹立した系統（株）、あるいは自分が見つけた疾患（患者）などを有し、それを多くの人が使ったり、引用したりされている場合
- 自分が開発した研究技術、解析手法などを駆使する研究
- 多くの研究者が注目している分野で、自分の研究の位置づけと意義を明確に示すことができ、しかも具体的な研究計画が示されている場合
- 全く新しい技術、方法を強い説得力で提案する研究。

これだけの文章で終わるような独りよがりの主張は、大抵、退けられる

創造性とは？

多数の定義や言い回しがあり、一つに限定できない。この多様性は、JSPSはぼやかしており、申請者に期待している

Japan Society for the Promotion of Science

日本大百科全書(ニッポニカ)

- https://kotobank.jp/word/%E5%89%B5%E9%80%A0%E6%80%A7-1556439
- G.Wallas： 創造の過程⇒(1)準備、(2)孵化(ふか)、(3)啓示、(4)検証の4段階に分けた、これより・・・(詳細は上記サイト)
- 創造性は突然真空から出現するものではなく、やはり長年月を要する基礎的学習という努力に加えて、当面の問題へ没入する集中のうえに築かれる。それは単なる思い付きではなく、まして無知や白紙状態と両立するものではない。

日本創造学会

- http://www.japancreativity.jp/definition.html
- 多数の言い回しがあり、
- 伊東俊太郎(麗澤大学)：「創造とは、問題を解決する、素材の新しい組み合わせ、新しい理論への変換を可能にする新たな視点の発見である」

この言葉は、「創造性」をイメージするのに、良い例になると思います。
既出の武部啓先生の言葉と重なっています。
以上のことから、独自性、創造性のイメージは重なり合っていると考えられ、しゃかりきになって、この二つを分ける必要はないことがわかります。

Pythonとパッケージ

1. Pythonを使う理由
2. ANACONDAとColaboratory
3. パッケージの概要

Pythonの開発歴史，名前の由来は，Wikipediaなどに詳述されているので，それらを参照されたい。

Pythonを使う理由

人気が高い

- IEEE(Institute of Electrical and Electronics Engineers，世界最大規模の学会)Spectrum 学術雑誌より，2017年学術分野で1位であった。
  - https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages
- 米国大学Computer Science コースの教育用言語として人気が高い
  - https://www.computersciencedegreehub.com/best/computer-science-online/
  - https://www.onlinecoursesreview.org/computer_science/
  - http://www.pgbovine.net/
- 民間指標(Tiobe)でも人気が高い
  - https://www.tiobe.com/tiobe-index/
- Pythonを使っている製品・商品が豊富
  - https://en.wikipedia.org/wiki/List_of_Python_software

Pythonを使う理由

特徴

- 下記は全て無料かつ許諾無し(と言っても無制限の自由はありません)に使用でき，そのライセンス内容はPSFL(Python Software Foundation License)次を参照されたい。
  https://docs.python.org/3/license.html
- スクリプト言語(scripting language):プログラムの記述が比較的簡易に行え，かつ，インタープリタ言語でもあるので，コンパイル不要で，逐次実行が可能。
- グルー言語(glue language):グルーとは接着剤を意味し，他言語(C/C++，Fortran(現在はオブジェクト言語))と結合(接着のニュアンス)させることができる。
- 豊富なライブラリ:パッケージとも呼ばれ，科学技術計算としての，数値計算，数式処理，統計，パターン認識，信号処理，システム制御，機械学習の他に，サーバ開発におけるサーバシステム開発，クローリング，クレイピングなどのライブラリを提供している。
- 開発環境の提供:Webベース開発の**Jupyter Notebook**，統合開発環境Spyderなどが提供されている
- 実行速度を上げることができる
  - スクリプト言語として実行すると遅い。
  - Numpyを上手に使うと，C/C++よりも早くなる場合がある
  - GPU向けにコンパイルできる。例えば，CUDA(NVIDIA開発)向けにnumba，CuPyなどのライブラリが提供されている。

参考図書

Pythonデータエンジニアリング入門 高速化とデバイスデータアクセスの基本と応用、橋本、他、オーム社(2020)
https://www.ohmsha.co.jp/book/9784274225345/

Python ドキュメント

Python 公式HP

- https://www.python.org/

ドキュメント

- https://docs.python.org/3/ ， 日本語:https://docs.python.jp/3/index.html

Python言語リファレンス

- https://docs.python.org/3/reference/
- 言語の構文とその構造について解説
- Pythonの思想を学びたい人向け

Python標準ライブラリ

- https://docs.python.org/3/library/
- 標準ライブラリ，組込み関数の使い方の説明
- 例えば，上記サイトから 2. Built-in Functions -> open() を選ぶと
  - open(file, mode='r', buffering=-1, encoding=None, errors=None, newline=None, closefd=True, opener=None)
- これから，open()文の引数パラメータの説明がある。英語に自信の無い人は，上記の日本語ドキュメントサイトの「ライブラリーリファレンス」から入ると，同じ組込み関数の使い方の日本語版を見ることができる。

ANACONDA

概要

- Pythonとそのライブラリ(パッケージ)，および，開発環境用のソフトウェアは多数あり，一つ一つインストールするのは面倒
- ANACONDAは，それら(100以上)を一括してダウンロードして，インストールを行う。
- 公式HP　https://www.anaconda.com/
- 注意：Python3.6以上を用いる。Python2.7系は用いない。

インストール方法

- 各自が所有するPC(Win10)にインストールしたい場合には，下記を参照。
  - https://sites.google.com/site/datasciencehiro/install
- ただし、インストールに関する質問は受け付けませんので、各自の責任で行ってください。

ANACONDA

CONDA

- ANACONDAをインストールすると使えるコマンドの名称
- コマンドラインからの入力となるので，次のウィンドウを開くこと。
  - Windows→コマンドプロンプト，または，PowerShell
  - Mac/Linux→ターミナル
- 簡単な使い方のまとめは次のサイトのPDFにまとめてある。
  - https://sites.google.com/site/datasciencehiro/install
- 授業で用いるパッケージのインストールもここに書いてある。

Python プログラムの拡張子

注意：下記はいずれもANACONDAパッケージに含まれているので，インストールは不要。

“.ipynb”

- Jupyter NotebookまたはColaboratory環境内で，Webベース上で記述されたファイル
- 公式HP　http://jupyter.org/
- 変換：”.py”，“.html”などのファイルに保存することができる。
- “.html”にはスクリプト，図，式，ドキュメントを一体的に含めることができる。

“.py”

- Python スクリプトで記述されたファイル
- 実行方法，2種ある
  - コマンドライン（Windowsではコマンドプロンプト，Macではターミナル）で行う

    > python filename.py

    - ただし，filename.py にパスが適合していることに注意されたい
  - 開発統合環境
    - その環境ウィンドウ内の実行ボタン（Run）をクリックする

パッケージとは

パッケージはライブラリ

- ほぼ同義で使われている。各サイトで表現が統一されていない。

本講義の使い方

- 関数（function）：def文で定義された個々の関数をいう。オブジェクト言語風にメソッドと称することもある。
- モジュール（module）：関数を幾つか（1つも可）まとめて1つのファイルに記述したもの。例えば，import scipy.signalの宣言ではscipy.signalがモジュールとなる。
- パッケージ（package）：モジュールを複数まとめたものをいう。Pythonドキュメントではパッケージをimportするという表現がある。この場合，各モジュールに機能分担させて，これら取りまとめたものをパッケージと言っている。ライブラリと同義で使う場合もある。また，chainerはパッケージの代わりにフレームワークと言っている。
- ライブラリ（library）：Pythonドキュメントではライブラリという用語が出現する。しかし，scikit-learnなどではライブラリは使わずパッケージと称している。本書ではこれらを同義として扱う。

講義で用いるパッケージの概要

- **NumPy** http://www.numpy.org/
  - ベクトル演算を利用した高速演算に特徴，他のパッケージからも参照される。
  - 注意：標準偏差numpy.stdで，不偏標準偏差を求めるにはパラメータの指定が必要。
- **SciPy** https://www.scipy.org/
  - サイパイと発する。
  - 科学技術計算に関する多様なツールを提供するパッケージである。例えば、補間、積分、最適化、画像処理、統計、特殊関数等がある。
  - Scipy Lecture Notes: https://scipy-lectures.org/
    日本語: http://www.turbare.net/transl/scipy-lecture-notes/index.html
- **pandas** https://pandas.pydata.org/
  - Pythonプログラムの統計分野でよく用いられるデータフレーム(DataFrame)，この意味はデータにラベルやインデックスを付けて変数に蓄える形式（フレーム）を意味する。
  - 簡単なデータベースのような操作（抽出，ソートなど）やグラフを描くことができる
  - 本ドキュメント，後半に，簡単な説明在り
- **statsmodels** https://www.statsmodels.org/
  - 統計分析用のパッケージ
- **scikit-learn** http://scikit-learn.org
  - 機械学習用のパッケージ

Jupyter NotebookとColaboratory

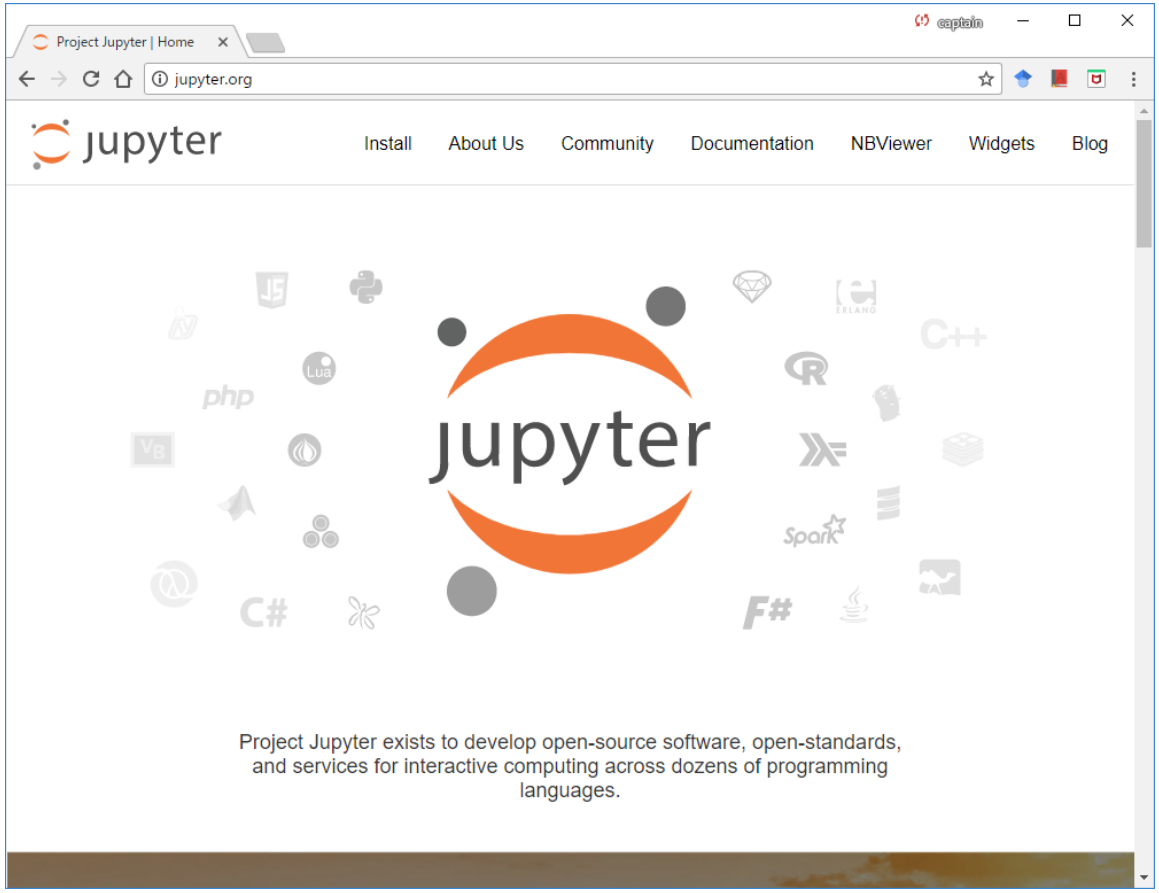
1. Jupyter
2. Jupyter Notebook
3. Colaboratory

注意：
Colaboratory (Colabと略す)はJupyter Notebook（Notebookと略す）の環境に多くの部分で似ているため、特に断らない限り、Notebookという記述はColaboratoryを意味すると読み進められたい。

Jupyter

Jupyter プロジェクト

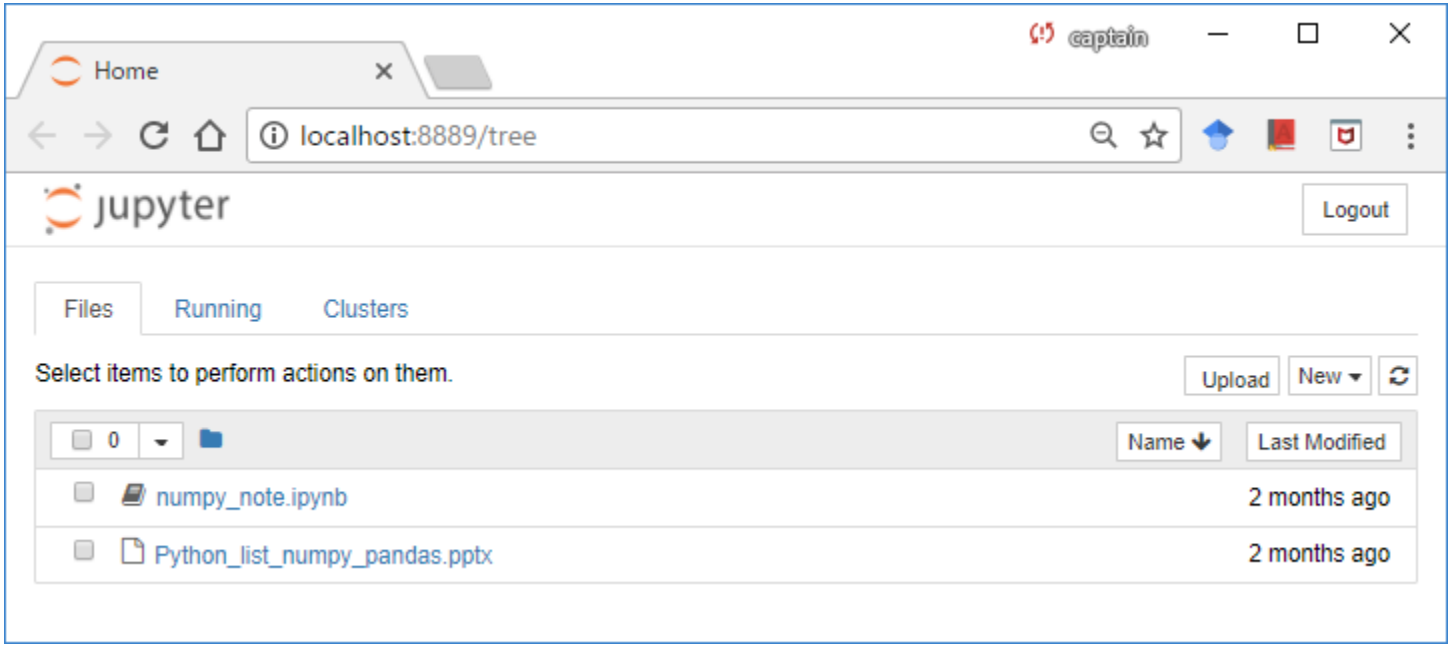
- http://jupyter.org/，Julia + Python + R を指向しているが多言語もサポート
- Ipythonプロジェクトからスピンオフしたプロジェクト
- ブラウザ用のインタラクティブインタフェースを複数提供



Jupyter Notebook

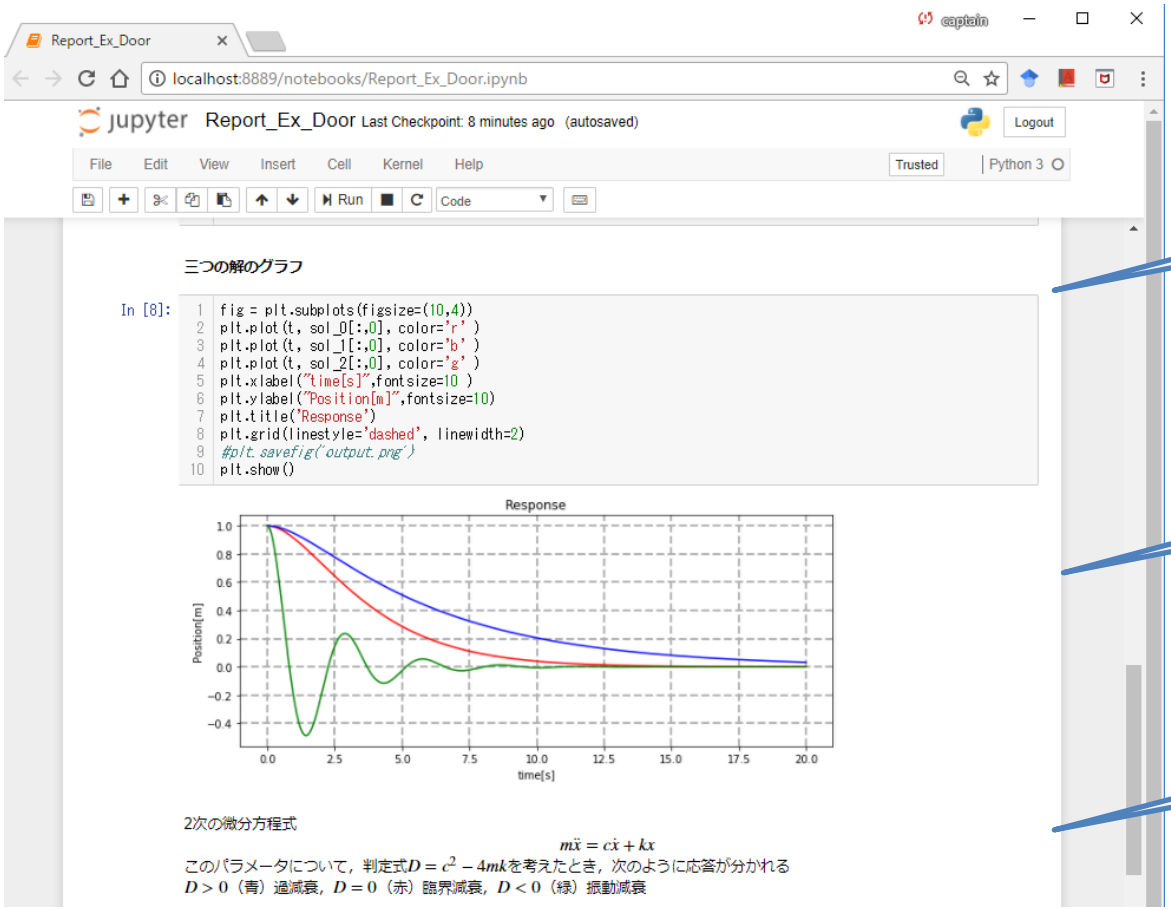
特徴

- https://jupyter-notebook.readthedocs.io/en/stable/notebook.html
- インタラクティブな実行が可能
- ブラウザベース（Chrome, IE, Firefoxなど）での編集，実行，表示のインタフェースを提供
- ANACONDAの中に含まれている。
- 特別なエディタやソフトウェアを必要としない。
- 本講義はこれを用いて，Pythonプログラムを編集，実行する。



スクリプト，グラフ，テキスト，式

- Notebookは、次の3種をブラウザ内に表現できる



スクリプト

グラフ

ドキュメントと式

Colaboratoryとは（概説）

概要：

- クラウドで実行され，Jupyter NotebookをGoogleドライブに保存する無料のJupyter Notebook環境である。

仕様（一部）：

- Intel(R) Xeon(R) CPU @ 2.20GHz （2 core）
- GPU NVIDIA Tesla K80
- TPU（Tensor Processing Unit）
- アイドル状態が90分続くと停止，連続使用は最大12時間

ドキュメント

- “Colaboratory へようこそ” https://colab.research.google.com/notebooks/welcome.ipynb

本授業での使用

- Colaboratory（クラウド上、推奨）、Jupyter Notebook（PC上）どちらを使用しても構わない
- Jupyter Notebookは大学PCで利用できる。個人PCにインストールは各自の責任で行うこと（インストールに関する質問は受け付けない）。
- インストールに自信の無い人はColaboratoryの使用を推奨する。
- Colaboratoryの使い方：次を参照されたい https://sites.google.com/site/datasciencehiro/python-kai-fa-huan-jing/colaboratoryno-shii-fang

授業の進め方とレポート書き方

テキストコンテンツ、動画、レポート

配布と動画の見方

- **manaba「掲示板」**（manaba→データサイエンス特論→「掲示板」）の第1回目に、テキストコンテンツ配布と動画の見方を説明する。
- テキストコンテンツは数回に分けて配布する。これを**各自所有するUSBメモリ**（USB 2.0以上，1GB以上）に保存して利用すること。
- 動画について、毎週定められた日時に動画を見ることのできるURLを示す。

視聴の確認

- 毎回、視聴後速やかにmanaba「小テスト」（短い提出期限が設定されています）を提出すること。出席代わりにもなる。

レポート

- **manaba「レポート」**（manaba→データサイエンス特論→「レポート」）に課題と提出法が示される。

上記の説明は、1～12回目（担当：橋本）の説明です。

13～15回目について、担当の大久保先生の指示に従ってください。

- この指示は、manaba「掲示板」に示されます。

授業の進め方

学習

- テキストコンテンツ内容を事前に理解し，動画を見て、Pythonを実行し、その結果を考察する。
- 復習：主に，レポート作成と提出

Python・コンピュータの使い方

- 遠隔授業ではColaboを使用（個人PCでAmacondaインストールしての使用は各自の責任で行ってください）。対面授業では、さらに大学PCのAnacondaを用いてもよい。Pythonの説明は各自で独学で学ぶこと。

評価

- シラバスに書いてあるとおり
- レポートは全部提出が必要条件
- 最終試験を受けることも必要条件
- プログラムの実行の注意：大学PCで実行する場合は，個人の作業領域は，デスクトップ上，または，USBメモリ上で行うこと。

授業の進め方

参考書

- シラバスに記載したもの
- 統計の教養は重要。授業中では詳しく説明しない。自学習が必要。上記のシラバスにも良書を記載
- それ以外に

無料で読める統計の良書

- オンラインで無料で読める統計書22冊　http://id.fnshr.info/2013/08/11/online-stat-books/
- オンラインで無料で読める統計書プラス32冊　http://id.fnshr.info/2016/08/15/online-stat-books-2/

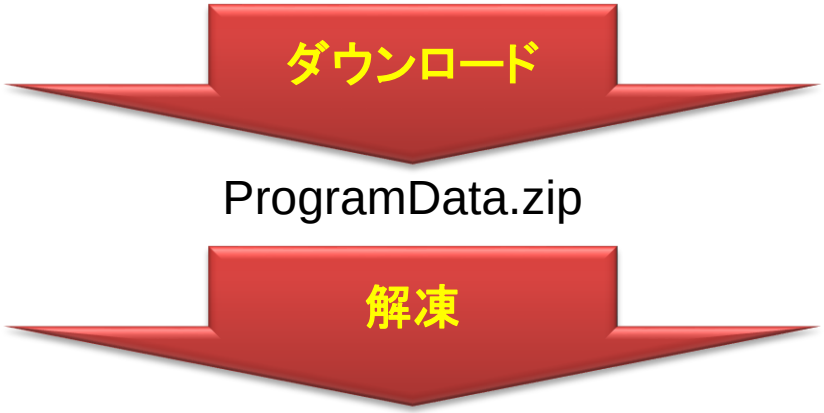
レポートの書き方と提出

1. 実行手順
2. スクリプト表記の約束
3. レポート作成と提出の方法

授業で用いるスクリプトの入手

サイト名：　データサイエンス

https://sites.google.com/site/datasciencehiro/download



注意：

- プログラム実行時に，ネットワーク接続を行っておくこと！　データをネットワーク経由で取得する例が多いため。
- オフライン（ネットワークに接続していない）で作業を行いたい場合には，予め，ネットワーク接続でデータを取得，これを個人の記憶領域に保存して用いること。

実行手順

Windowsの場合での，Jupyter Notebookの起動，プログラム実行概要を説明する。

1. 作業フォルダを開く（スクリプトかNotebookのあるフォルダ）
2. このフォルダのアドレスバーに，コマンドプロンプト（cmd）またはPoweShell(powershell)を入力して，どちらかを起動する。
3. ブラウザが起動して，その中で，スクリプトやテキスト，さらにLaTeX流の数式の記述が行える。
4. Run Allで，数値やグラフの結果はブラウザ上に表示される。
5. 終了は，Fileメニューから Close and Halt またはブラウザの×印をクリック，コマンドプロンプトウィンドウはCTRL＋Cまたは×印をクリック

この詳細は次のPDFを参照されたい

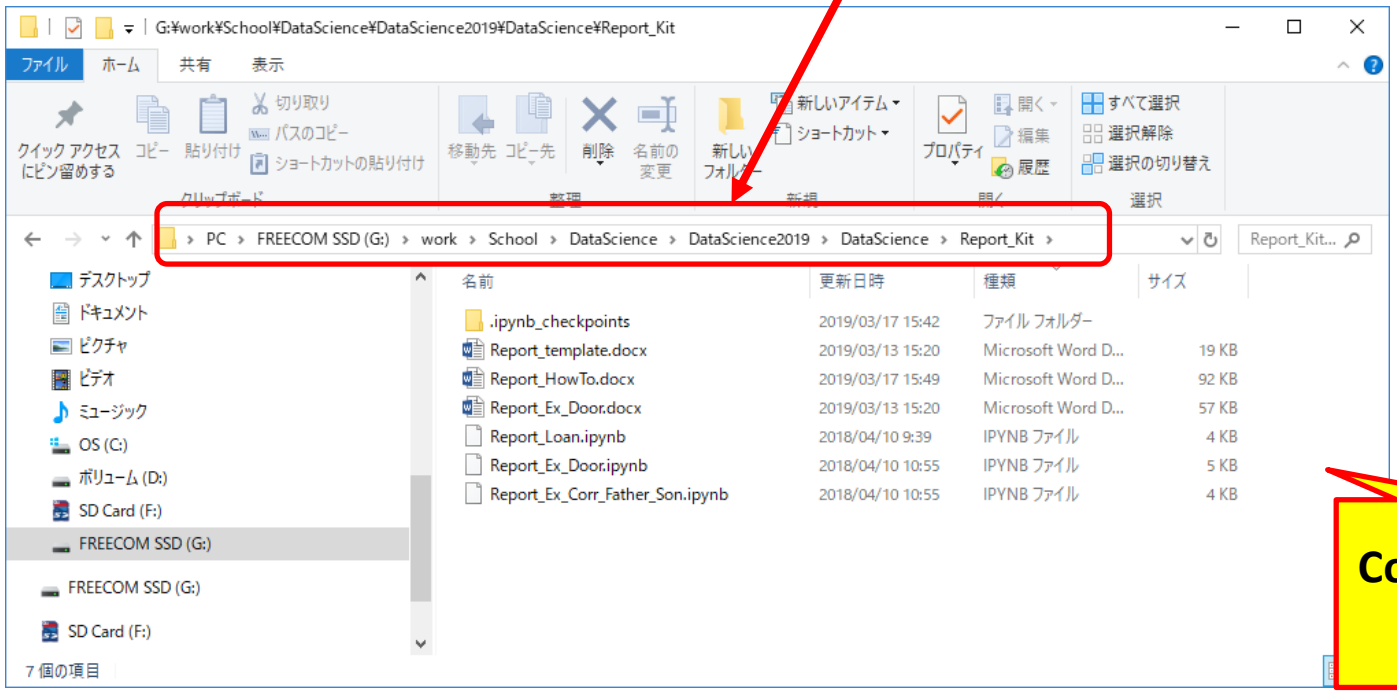
URL: https://sites.google.com/site/datasciencehiro/python-kai-fa-huan-jing/jupyter
にある“HowTo_Jupyter_Notebook.pdf” を掲載。

こちらを利用

注意：Notebookの代わりに、Colaboratoryを用いる場合の説明：
https://sites.google.com/site/datasciencehiro/python-kai-fa-huan-jing/colaboratoryno-shii-fang

Notebookの起動

1. 実行したい".ipynb"のあるフォルダを開く
2. アドレスバーのパスを消去する。
3. cmd または powershell を入力
4. 開いたウィンドウから > jupyter notebook を入力
5. Junpyter Notebookが起動する。



Colaboの利用とは違います

この後の使い方は，Jupyter Notebookの使い方　を見ること。

スクリプト表記の約束

プログラム中の日本語表記

- 1行目に次を記述すれば，スクリプト中に日本語を書くことができる
  - # -*- coding: utf-8 -*-
- ここに，utf-8はエンコーディング名である。他に，Shift_JIS，cp932，EUC-JP などがある。
- 注意：日本語に限らず，英語や他言語でも，どのエンコーディングを用いたかが重要。Python はディフォルトがutf (Unicode Transformation Format)

パッケージの省略語

本文中で次の省略語を用いる。

- import numpy as np
- import scipy as sp
- import pandas as pd
- import matplotlib.pyplot as plt
- import statsmodels.api as sm
- import statsmodels.formula.api as smf

レポート作成と提出の方法

フォルダ DataScience¥Valuable_Kit

- このフォルダの中に，レポート書き方，レポート用Notebookなどがある。
- この中を用いて，以下，説明する

フォルダDataScienceは、配布するテキストコンテンツのフォルダ名、フォルダValuable_Kitはこの中にあります。

レポートの書き方，例題，テンプレート

- Report_HowTo.docx (.pdf)　　レポートの作成手順（初めに読む）
- Report_Ex_Spring_Damper.docx (.pdf)　　レポートはこのように書く，という例
- Report_template.docx (.pdf)　　レポートのひな型，これを用いること（類似も可）

例題用のNotebook

- Report_*.ipynb
- 拡張子”.ipynb”は　IPython Notebookの略称
- 注意，他の章のNotebookは，
  - https://sites.google.com/site/datasciencehiro/download　の中の“ProgramData.zip”をダウンロードすること。

本講義で用いる用語
Notebook　　“.ipynb”ファイルをさす
スクリプト　　“.py”ファイルをさす

Colaboratoryでは”.py”を実行しません

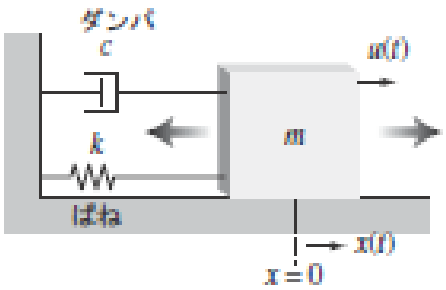
レポート作成と提出の方法

提出するレポートのファイル名

1. 提供するワードの**ひな型**の “Report_template.docx”ファイル
   - この他に、指定されたファイル（スクリプト等）の提出を求めることがあるが、特に何も指示が無ければ、このファイルのみを提出すること。
   - ファイル名の管理は各自に任せますが、例えば、次のようにすれば管理はしやすくなります。
     - Report_氏名_n回レポート.docx

次ページ以降では、このレポートの解説を動画で示します。
この動画と次ページ以降の説明の両方を理解してください。

演習：ばね・ダンパ・質量系



m [kg] ：質量（mass）
c [Ns/m] ：減衰係数（damping coefficien
k [kg/s^2] ：ばね定数（spring coefficient

図1　質量・ダンパ・ばね系

図に示すダンパ（damper）とは，ピストン状の中に，空気や粘性物を入れて物体の速度を減じるために使われるもの．身近なところでは，ドアダンパや車のショックアブソーバなどがある．式で表現すると速度 v（$=dx/dt$）に比例して減衰する力が働くと考えて，この力は次式で表される．

$$f = cv \tag{1}$$

ここに，c は粘性減衰定数（viscous damping coefficient）または簡単に減衰定数（damping coefficient）と言う．

初めに，外部の力 $u(t)=0$ の場合を考える．このとき，図1に示す質量・ダンパ・ばね系の運動方程式は

$$m\ddot{x}+c\dot{x}+kx=0 \tag{2}$$

で表され，これは2階斉次方程式である．

次で、このシミュレーション例を見る

既に説明した「実行手順」を参照

Report_Ex_Spring_Damper

この囲みは，
拡張子".ipynb"，
フォルダValuable_Kit内にある

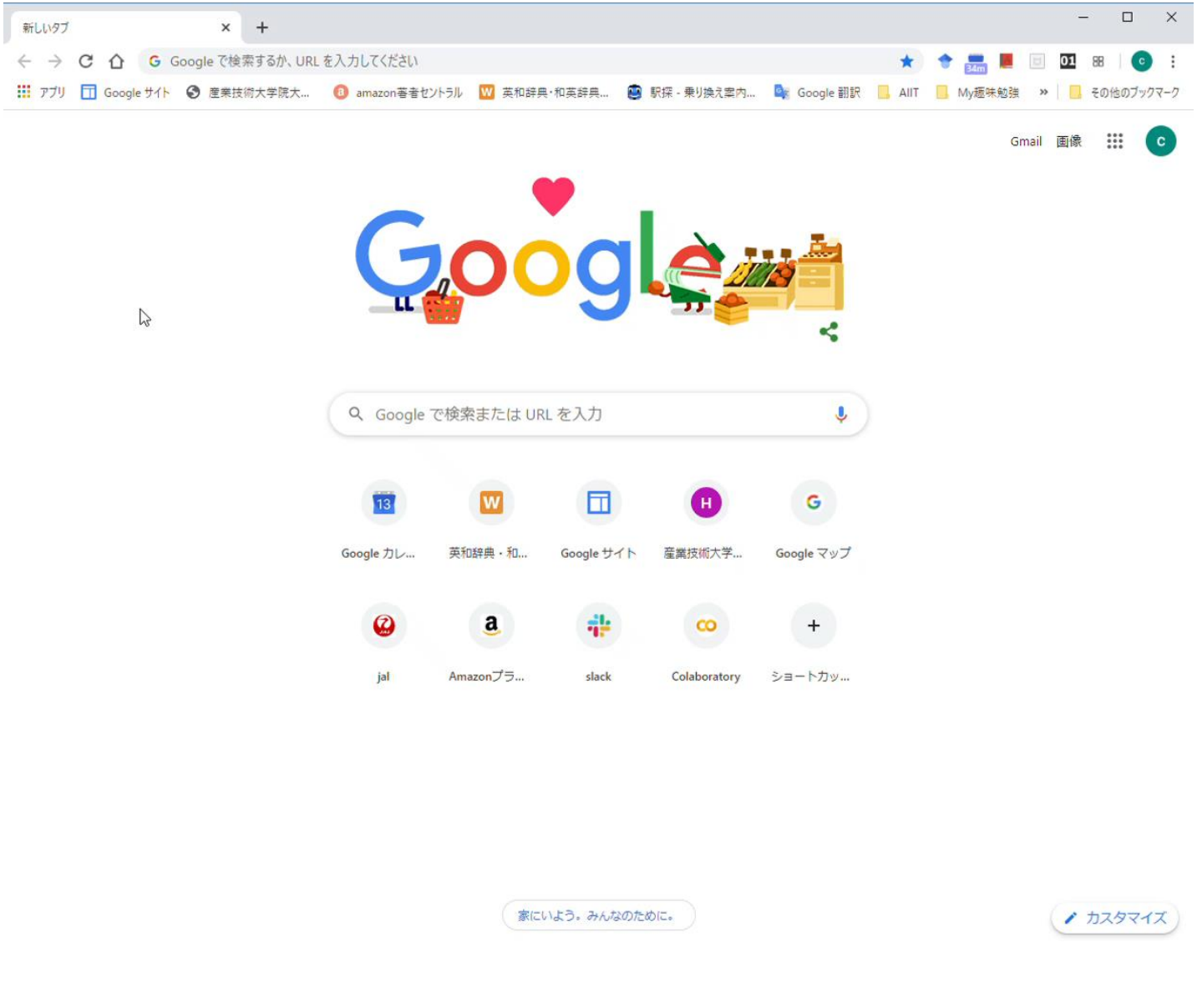
レポート例：
"Report_Ex_Spring_Damper.docx"，
"ditto.pdf"

シミュレーション結果のレポートへの貼付け

ノートブック内のスクリプトを実行し、結果として得られたグラフ（または表）だけを切り取り、それをレポート用紙（ワードファイル）に張り付ける手順を学ぶ。

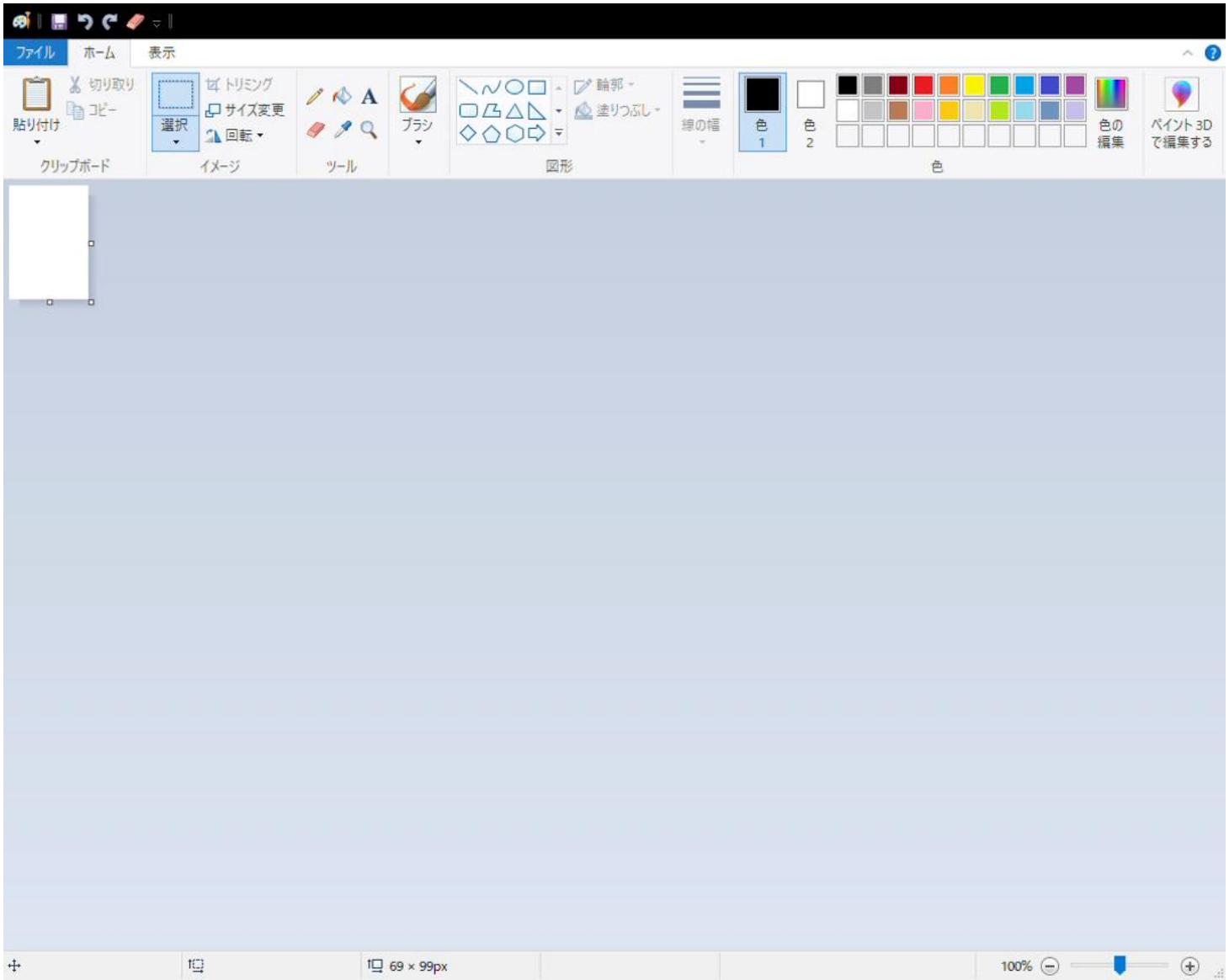
1. フォルダ　DataScience¥Valueable_Kit　の中の“Report_Ex_Spring_Damper.ipynb”をColaboにアップロード
2. 実行（[ラインタイム]⇒[すべてのセルを実行]）
3. 画面キャプチャ（Windows10：Altキー＋Prt SCキー）で画面の画像をPC内のバッファにコピー
4. それを何かのペイントソフト（ここでは、Windowsの「ペイント」を用いる）に貼付ける（ペースト）。
5. そこで、欲しい部分だけを切り取り、コピーを選ぶ。
6. ワードファイル状で、貼り付け（ペースト）を行う。

シミュレーション結果のレポートへの貼付け



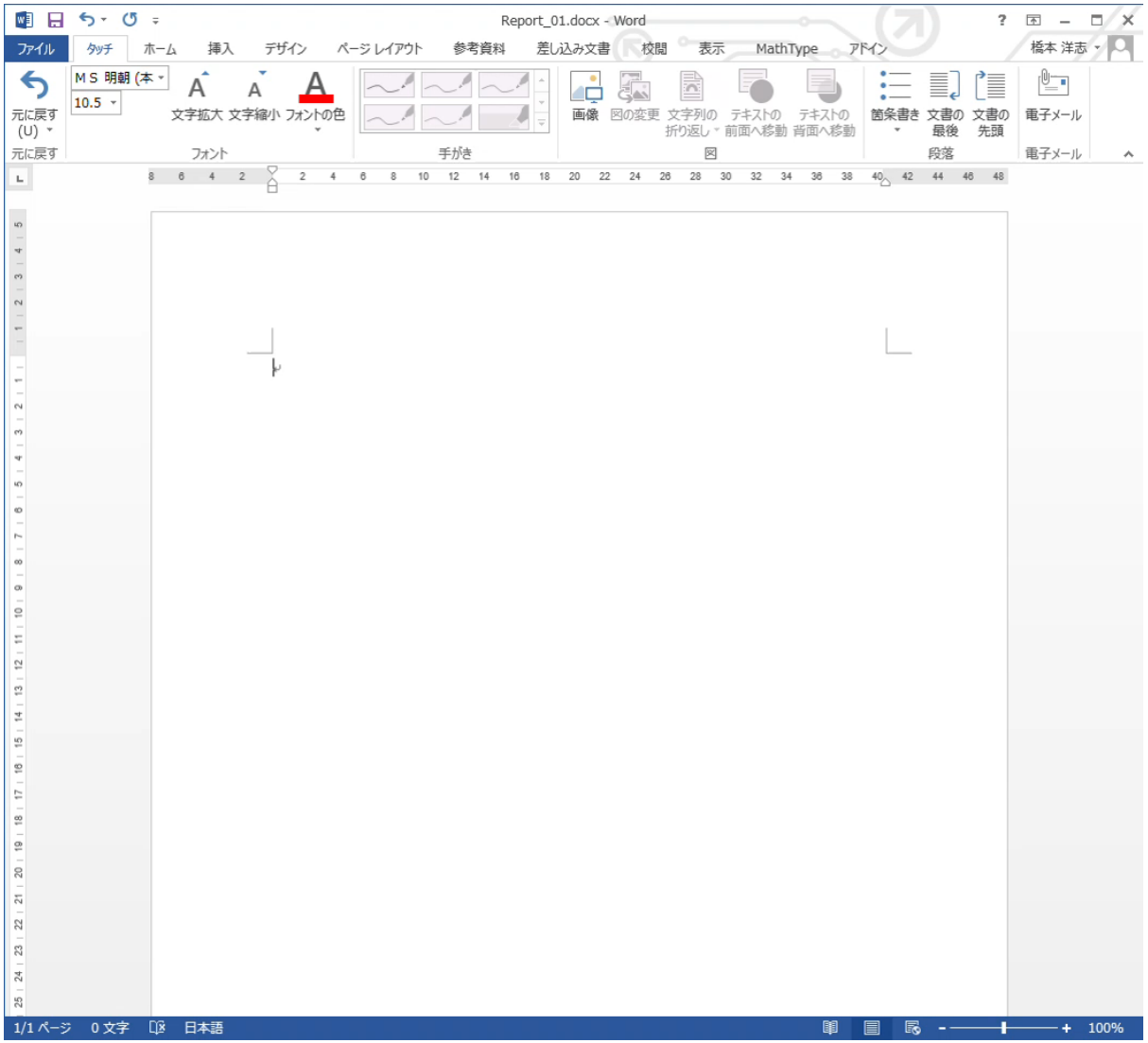
シミュレーション結果のレポートへの貼付け

3. 適当なペイントソフトで、グラフだけを切り抜く。



シミュレーション結果のレポートへの貼付け

4. これをワードに貼り付ける



演習：父親と息子の身長の相関

問題の内容

- 父親と息子の身長[cm]に相関があるのか
- 先行研究：F.ゴルトンの線形回帰、http://www.stat.go.jp/dss/⇒“ビジネスに役立つ統計講座”⇒“未来がわかる方程式”

演習の内容

- 問題の内容の理解は、上記のサイト、および、提供するNotebookを参照されたい。
- 線形回帰、相関、t検定量などは、本講義の後に現れ、これらを見て、考察することを本講義は要求するので、予め、予習しておくこと。
- さらに、表やグラフをレポート用紙に張り付けられるようにしておくこと。

フォルダValuable_Kit内の
Report_Ex_Corr_Father_Son

付録
各自で自習すべき内容
授業を理解するための必須項目

1. 用語の使い方
2. 数学，簡単表記
3. pandas 概要

言葉は変わる

言葉は分野によりその意味が異なる。例えば，

interfaceは，化学分野での異物質の界面を意味するが，情報分野では人とコンピュータ，またはコンピュータとコンピュータの接続部のソフトウェアやハードウェアを意味する。

controlについて，system controlは制御工学では思い通りに動かすこと（激しく動かすこともある），医療分野でのcontrolled study は「ある（対照群を置いた）比較試験」，航空分野でのcontrol towerは航空管制棟を意味する。

このように，同じ単語でも分野が異なれば，単語が有するイメージが異なる。

データサイエンスの分野は異なる文化を持つ分野を網羅するので，ある用語のイメージが異なることがある。ここでは，各章における分野に沿った用語の使い方を行う。

次ページ以降，しばらく，言葉の意味が異なることの説明を述べる。

説明変数，目的変数，入力，出力

説明変数と目的変数について幾つかの表現

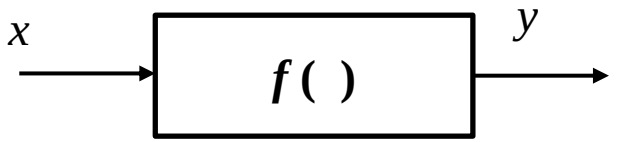
- 【説明変数 x】
  - 説明変数　explanatory variable
  - 予測変数　predictor variable
  - 独立変数　independent variable
  - 外生変数　exogenous variable -> python statsmodels で exog　（経済系が多い）
- 【目的変数 y】
  - 目的変数、応答変数、反応変数　response variable
  - 結果変数　outcome variable
  - 従属変数　dependent variable
  - 基準変数　criterion variable
  - 内生変数　endogenous variable -> python statsmodels で endog（経済系が多い）
- ○○変数(variable)ではなく○○変量(variate)とする本も多い。

endog, exog, what's that? in statsmodels
http://www.statsmodels.org/devel/endog_exog.html

説明変数，目的変数，入力，出力

入力と出力

- 工学の多くの分野では，統一的な表現を用いるため，
  右のように，ブロック線図の表現を用いて
  システムの入力と出力で表現することが多い。
- 入力と出力には因果性があり，ある関数 f ( )で，その
  対応が数学的に対応付けられているとする。
- f ()が線形で，有理多項式で表現されるとき，
  伝達関数として表現されることも多い。
- これにより，システムを数学的に解析できる。
- 実は，上記の目的変数と説明変数もこの表現で説明できるが，
  本講義では，用途により，説明変数／目的変数，入力／出力　を使い分ける。



予測，推定

先のブロック線図表現で

- f()を既知として，xからy を求めることを考える。
- 統計，パターン認識分野は，xに基づきyを求めることを予測という
- 理工学分野の多くでは，ダイナミカルシステムを扱うため，yを求める行為においても，予測と推定を使い分けている。時間 t を意識する。

ダイナミカルシステムとは

- x, yは時々刻々と変化する
- xとyに因果性がある
- xがゼロになっても，yはしばらく値を示す
  - 鐘を撞くと，音と振動がしばらく続く。鐘がシステム，振動と音が出力y，撞くことが入力xを与える。
  - 加熱後に熱湯を冷ます。熱湯がシステム，温度が出力y，加熱が入力x

ダイナミカルシステムの予測

- ある時刻t0 において、先の時刻 t1 (= t0 + d (>0))における状態を推定することを予測(prediction)という。時間因子 t が根底にある。
- この点で、統計や機械学習で用いる予測とはニュアンスが異なる。すなわち、時間因子 t が無いことが多い。

予測，推定

推定

- ある真のモノがあるが，もやがかかったり，暗かったりして，そのモノを厳密に知ることができないとき，何らかの手がかりを用いて，そのモノを何であろうかと知ること
- この定義から
- 統計分野では，点推定，区間推定という
- システム制御論，計測論，ディジタル信号処理，画像処理・認識論からも推定という用語をよく用いる
- パターン認識，機械学習の分野では，未知のものを探り当てることを予測と言っている。
- 同じ行為であっても，用語の用い方が異なる。

予測区間

- 予測区間（よそくくかん）とは統計学用語で、母集団を仮定した上で、将来観察されるであろう標本値（現在は測定できない）に対して「どの範囲にあると予測されるか」を示すものである。
- これに対し、信頼区間とは、母集団の母数（標本から測定できない）に対して「どの範囲にあると推定できるか」を示すものである。混同しないように注意。
- https://ja.m.wikipedia.org/wiki/%E4%BA%88%E6%B8%AC%E5%8C%BA%E9%96%93

予測

- 予 測（よそく） 将来のことを前もって推測すること。 将来のことを前もって推測した内容。
- https://ja.m.wiktionary.org/wiki/Special:Search?search=%E4%BA%88%E6%B8%AC&fulltext=1&searchToken=2wv1kj8gq4sm8dejsrfnj1eyy
- prediction , a statement of what you think will happen in the future (by https://dictionary.cambridge.org/)

予測，推定

統計の分野では

- https://en.wikipedia.org/wiki/Prediction
- In statistics, **prediction** is a part of statistical inference. One particular approach to such inference is known as **predictive inference**, but the **prediction** can be undertaken within any of the several approaches to statistical inference. Indeed, one possible description of statistics is that it provides a means of transferring knowledge about a sample of a population to the whole population, and to other related populations, which is not necessarily the same as **prediction** over time. When information is transferred across time, often to specific points in time, the process is known as **forecasting**. **Forecasting** usually requires time series methods, while **prediction** is often performed on cross-sectional data.
- google 翻訳
- 統計では、予測(prediction)は統計的推論の一部です。そのような推論に対する1つの特定のアプローチは、予測推論(predictive inference)として知られているが、予測は、統計的推論に対するいくつかのアプローチのうちのいずれかの中で行われ得る。実際、統計の1つの可能な記述は、母集団のサンプルについての知識を母集団全体および他の関連母集団に伝達する手段を提供することであり、これは時間の経過とともに必ずしも予測と同じではない。情報が一定の時間内に、しばしば特定の時点に転送される場合、プロセスは予測と呼ばれます。予測には通常時系列の手法が必要ですが、予測は断面データに対して行われることがよくあります。

分野別表現

- 推定（点推定，区間推定），予測（予測区間）を用いる
- パターン認識（機械学習）では，情報伝達に基づく推論を行うことから予測を用いる。これは，時系列データを扱う分野では推定の意味となる
- 時系列データを扱う分野（制御工学，信号処理，経済分野）では，時間軸を念頭において，現在時刻以前を推定，将来時刻での推定を予測（prediction，forecast）と言う。

予測，推定

分野別表現

- 推定（点推定，区間推定），予測（予測区間）を用いる
- ここで，
  - 平均値など母集団パラメータを求めることはestimation
  - Xに基づきYを求めることは prediction
  - ref: Bret Larget; Estimation and Prediction, Dept. of Botany and of Statistics Univ. of Wisconsin, 2007, http://www.stat.wisc.edu/courses/st572-larget/Spring2007/handouts03-1.pdf

predictの語源

- ラテン語 praedict , Early 17th century: from Latin praedict- ‘made known beforehand, declared’, from the verb praedicere, from prae- ‘beforehand’ + dicere ‘say’.
  - ref Oxford Dictionaries Online – English Dictionary and Thesaurus https://en.oxforddictionaries.com/
- 1620s (implied in *predicted*), ″foretell, prophesy,″ a back formation from *prediction* or else from Latin *praedicatus*, past participle of *praedicere* ″foretell, advise, give notice,″ from *prae* ″before″ (see pre-) + *dicere* ″to say″ (from PIE root *deik- ″to show,″ also ″pronounce solemnly″). Related: *Predicted*; *predicting*.
  - Online Etymology Dictionary https://www.etymonline.com/

分類，識別，判別

本講義では，クラス分類(classification)を用いる

- classificationの訳が分類であるから，クラス分類は屋上屋を重ねている感がある。しかし，単に分類と言うと，grouping，categorization などと混乱するかもしれず，何を分類しているかが認識できないことがある。ここでは，クラスを分類することがここでの目的であることを明確にするために，敢えて「クラス分類」という用語を用いる。
- パターン認識では，従来，判別(識別)分析(discriminant analysis)という用語がある。判別 のもともとの意味は，違いを認識または知覚するなどがある。したがって，分類したものを判別するという言い方ができる。この分類と判別の目的を考えると，クラス分類と判別分析はほぼ同義と考えても良い。
- 判別と識別(discrimination) （わかりやすいパターン認識，オーム社）
  http://nosyan.hatenablog.com/entry/20100313/1268437040
- 識別 goo辞書 https://dictionary.goo.ne.jp/thsrs/3319/meaning/m0u/
- 統計学の分類 https://ja.wikipedia.org/wiki/分類_(統計学)
- クラス分類とクラスタリングの違い http://blogs.itmedia.co.jp/takafumi/2015/09/post_3.html
- 注意：racial [sexual] discrimination （人種(男女)差別)という使われ方があるように， discriminationに差別と言う意味があるので，一般の場合の使用には注意が必要である。

こんな用例では

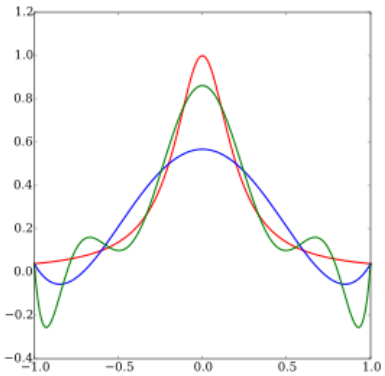
- 群衆の中から，いるかどうかわからない犯人の顔を(識別，判別，分類，同定)する。この例では，まず，分類は用いないであろう。
- 本講義では，このような事例は無いが，読者がさらに発展した学びでは，活動フィールドに適する用語を用いられたい。

トレーニングデータ，テストデータ

- **統計分析，パターン認識，機械学習の分野で多種の用語**
- **トレーニングデータ(training data)**
  - 同じ意味として，訓練データ，学習データ，教師データ
- **テストデータ（test data）**
  - 同じ意味として，評価データ，検証データなど
- **カタカナ用語に統一して使う**

- **備考**
  - trainingには，訓練、養成、練習、鍛練、調教、訓練課程がある。ニュアンスとしては，あるスキルを発揮するのに訓練・調教するということである。
  - もともと，train（列車）の原義が，「引っ張る(pull，draw)」であり，これから，所望の形に成長させる，という意味に繋がった。
  - 参照： https://en.oxforddictionaries.com/

Overfitting

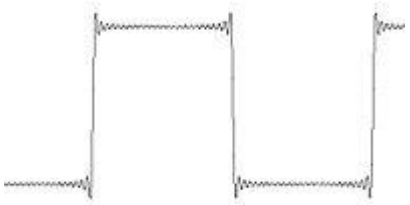


過剰適合

- 過学習は言い過ぎの感がある。
- オーバーフィッティングとして用いる

類似の解析

- ルンゲ現象(Runge‘s phenomenon) 数値解析で高次の多項式で多項式補間する際に発生する問題である。
  - 多項式の次数を大きくしていくと補間誤差が無限大に漸近する
  - https://en.wikipedia.org/wiki/Runge%27s_phenomenon
    - https://en.wikipedia.org/wiki/File:Runge_phenomenon.svg
  - https://ja.wikipedia.org/wiki/%E3%83%AB%E3%83%B3%E3%82%B2%E7%8F%BE%E8%B1%A1
  - 対策の1案:Chebyshev nodes https://en.wikipedia.org/wiki/Chebyshev_nodes
- ギブス現象(Gibbs phenomenon)
  - https://ja.wikipedia.org/wiki/%E3%82%AE%E3%83%96%E3%82%BA%E7%8F%BE%E8%B1%A1
  - https://en.wikipedia.org/wiki/Gibbs_phenomenon



分析と解析

- **どちらも英語でanalysis，日本語と英語は1対1には対応しない。**
- **分析**
  - 複雑な事柄を一つ一つの要素や成分に分け、その構成などを明らかにすること。（デジタル大辞泉，小学館）
  - 分析化学，リスク分析，ポートフォリオ分析，財務分析，定性分析，というようにデータに基づく行為を指すことが多い。
  - これらの用例に示すように，分析は，どちらかというと，対象を数学モデルよりは図表やグラフ，言葉などで表現し，それらの互いの関係を明らかにして対象の特性や特徴を調べる際に用いることが多い。
- **解析**
  - 事物の構成要素を細かく理論的に調べることによって、その本質を明らかにすること。（デジタル大辞泉，小学館）
  - 解析学／フーリエ解析（数学），ノード解析（電気工学），暗号解析（情報学），流体解析（機械工学），形態素解析（言語学），というように数式，または定理に基づく行為を指すことが多い。
  - これらの用例に示すように，解析は，数学的に扱う場合に用いることが多く，「解析的に解く」「解析解」「解析学」などと言う。対象を数学モデル（数式，関数）で表すことが多く，解析の結果は定量的（数量的）なものになることが多い。

分析と解析

本書では

- 統計，回帰，パターン認識の分野はデータに基づくため用語「分析」を用いる。
- 章のタイトルは苦慮した。主成分分析などは，従来，「多変量解析」の分野で発展してきた。しかし，筆者の判断で，主成分分析，因子分析，対応分析を選定し，これを一つの章にまとめざるをえなかった。そのため，苦慮した上で，「多次元データ分析」という章にまとめた。このまとめ方にご容赦願いたい。

変数と変量

次の解説がある

- 「統計集団をなす個体が“担っている”数量を抽象化して変量（ variate ）と呼ぶことが多い．数学の変数（ variable ）の概念に対応するが、個体に応じて変化し，物理的，経済的な意味をもつ量であるとの意識が強い．データは変量がとる値（ value ）である．しかし，変量とデータは変数と変数値のように混同されがちであり，**うるさく区別しないほうが便利である**．変量と変数も混同されがちで，本辞典内でも区別しない場合が多い．変量と変数も混同されがちで，本辞典内でも区別しない場合が多い．」,竹内啓 編集委員代表，統計学辞典，東洋経済新報社，1.2.1 データと変量より抜粋
- 本授業でも，うるさく区別しない。

- なお，統計分野では「多変量解析」という用語はあるが，「多変数解析」はない。これは，数学に「多変数解析関数論」という用語があるためであろう。

相関，共分散

- Wikipedia　によると
  https://en.wikipedia.org/wiki/Cross-correlation
  - In probability and statistics, the term cross-correlations is used for referring to the correlations between the entries of two random vectors X and Y, while the autocorrelations of a random vector X are considered to be the correlations between the entries of X itself, those forming the correlation matrix (matrix of correlations) of X. This is analogous to the distinction between autocovariance of a random vector and cross-covariance of two random vectors. One more distinction to point out is that in probability and statistics the definition of correlation always includes a standardising factor in such a way that correlations have values between −1 and +1.

引用：http://lambdalisue.hatenablog.com/entry/201

上記は英語版 Wikipedia の Cross-correlation から引用したものですが、調べものをする際は基本的に読み飛ばす癖があるので、この部分を見つけるまでにかなりの時間を要してしまいました。

引用によりますと、自己相関（autocorrelation）や相互相関（cross-correlation）と言った場合に統計学（statistics）および信号処理（signal processing）にて定義に若干の違いが有るようです。統計学においては「相関（correlation）」と言った場合は常に正規化されており、値は -1 ~ 1 を取るのに対し、信号処理では「相関」と言った場合は処理対象の値をそのまま返し、正規化されたものは特別に「相関係数（correlation coefficient）」と呼ばれるようです。

要約すると正規化の有無により以下のような呼称の違いが有るようです。

正規化	統計学 (statistics)	信号処理 (signal processing)
×	相互共分散 (cross-covariance)	相互相関 (cross-correlation)
○	相互相関 (cross-correlation)	相互相関係数 (cross-correlation coefficient)
×	自己共分散 (autocovariance)	自己相関 (autocorrelation)
○	自己相関 (autocorrelation)	自己相関係数 (autocorrelation coefficient)

また統計学の定義では「共分散（covariance）」および「相互共分散（cross-covariance）」に数式的な違いが見いだせませんでしたが、これは概念的な違いを明確にするための言葉のようでした（英語版 Wikipedia Cross-covariance より引用）。

> In the case of two random vectors X=(X_1, X_2, ... , X_n) and Y=(Y_1, Y_2, ... , Y_n), the cross-covariance would be a square n by n matrix C{XY} with entries C{XY}(j,k) = cov(X_j, Y_k).\, Thus the term cross-covariance is used in order to distinguish this concept from the "covariance" of a random vector X, which is understood to be the matrix of covariances between the scalar components of X itself.

実際の解析時に使用する numpy では信号処理的な定義が用いられているため（関数の返り値が正規化されていない）今回は**信号処理**の定義および呼称を用いることとします。

数学　簡単表記

- **ベクトル　　：小文字・太文字・イタリック体　$\boldsymbol{a}$**
- **行列　　：大文字・太文字・イタリック体　$\boldsymbol{A}$**
- **距離**
  - ノルム（norm）の意味で用いている。良く知られたユークリッド距離はこの一部である。
  - https://ja.wikipedia.org/wiki/ノルム
- **三角関数**
  - サイン関数（正弦），コサイン関数（余弦），タンジェント関数（正接）とカタカナ表記する

初心者向けの参考文献：
- 高校数学の美しい物語：https://mathtrain.jp/
- 統計学の時間：https://bellcurve.jp/statistics/course/

pandas: 基本関数

- Pandasのデータオブジェクト
- データI/O
- データを眺める
- 各種統計量
- データの選択
- データの操作
- データの結合
- グルーピング
- 階層インデックス
- 時系列データの基本処理
- タイムゾーン

上の引用：https://qiita.com/tanemaki/items/2ed05e258ef4c9e6caac

データの抽出
http://akiyoko.hatenablog.jp/entry/2017/04/03/081630
http://sinhrks.hatenablog.com/entry/2014/11/15/230705

グルーピングの例：
https://openbook4.me/projects/183/sections/1670

行，列の挿入
https://pythondatascience.plavox.info/pandas/pandas%E3%81%AE%E3%83%87%E3%83%BC%E3%82%BF%E3%83%95%E3%83%AC%E3%83%BC%E3%83%A0%E3%81%AB%E8%A1%8C%E3%82%84%E5%88%97%E3%82%92%E8%BF%BD%E5%8A%A0%E3%81%99%E3%82%8B
by Hiroshi

```
df = pd.read_csv('World_GDP_Popu_CBR.csv')
tmp_rate = df['GDP']/df['Population'] #一人当たりのGDP
df['GDPpop'] = tmp_rate
```

pandas の loc、iloc、ix の違い – python
http://ailaby.com/lox_iloc_ix/

pandas: DataFrame と Series

DataFrame

- 2次元のテーブルを表す（右上図）
- index: 1月，・・・，12月を指す
- columns: 平均気温，・・・，降水量を指す

Series

- 1次元配列（右下図）
- 右上図の列（column），または，行（row）に相当する。
- ラベルを付与できる。

2つのtypeを見る

df = pd.DataFrame(右上図)とおく
print(type(df.describe())
<class 'pandas.core.frame.DataFrame'>

print(type(df.mean())
<class 'pandas.core.series.Series'>

typeがSeriesだから，例えば
mean = df.mean[‘日照時間’]
とすれば，平均の数値だけを抽出できる。

	平均気温	日照時間	最低気温	最高気温	降水量
1月	5.2	184.5	0.9	9.6	52.3
2月	5.7	165.8	1.7	10.4	56.1
3月	8.7	163.1	4.4	13.6	117.5
...	...	...	...	...	...
10月	17.5	131.0	14.2	21.5	197.8
11月	12.1	147.9	8.3	16.3	92.5
12月	7.6	178.0	3.5	11.9	51.0

	平均気温	日照時間	最低気温	最高気温	降水量
1月	5.2	184.5	0.9	9.6	52.3
2月	5.7	165.8	1.7	10.4	56.1
3月	8.7	163.1	4.4	13.6	117.5
...	...	...	...	...	...
10月	17.5	131.0	14.2	21.5	197.8
11月	12.1	147.9	8.3	16.3	92.5
12月	7.6	178.0	3.5	11.9	51.0

参照： http://www.yunabe.jp/docs/pandas_basics.html

pandas: Tips for DataFrame Operation

- **2次元DataFrameから，ある1要素を抽出**
  - rxy = corr_coef.loc[['father'],['son']].values # rxy = [[0.63427]]　配列で抽出
  - rxy = corr_coef['father']['son'] #　rxy = 0.63427　数値で抽出
- **ある行を削除したい，indexカラムが無い下のデータ列に対して**
  - 注意：indexカラムが無いので，例えば，df.drop('Brazil')はできない。
  - Country列のBrazilを含む行を削除したい，排他的操作を用いて

df2 = df[df['Country'] != 'Brazil']

	A	B	C	D	E	F	G
1	Country	Currency	Year	GDP	Population	CBR	
2	Afghanistan	US$	2015	2.03E+10	33736	28.7	
3	Americas	US$	2015	2.50E+13	319926	4.3	
4	Bolivia	US$	2015	3.3E+10	450	16.8	
5	Brazil	US$	2015	1.77E+12	205962	9.1	
6	Colombia	US$	2015	2.92E+11	48229	10.4	
7	Ethiopia	US$	2015	5.99E+10	99873	26.1	
8	Finland	US$	2015	2.32E+11	5482	1.3	
9	France	US$	2015	2.42E+12	64457	3.4	
10	Germany	US$	2015	3.36E+12	81708	-2.4	
11	Greece	US$	2015	1.95E+11	11218	-1.2	

pandas:統計量

- pandas.DataFrame.describe
  - std（標準偏差）は不偏標準偏差を求める（サンプル数−1で除算）
  - https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html
  - ココを見ると， DataFrame.count, DataFrame.max, DataFrame.min, DataFrame.mean, DataFrame.std, DataFrame.select_dtypes　とある。例えば，stdは次に説明がある。
  - https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.std.html#pandas.DataFrame.std
- pandas.DataFrame.corr
  - 出力はDataFrame
  - https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html

リスト、NumPy、Pandas間の変換

- http://blog.pepese.com/entry/2016/09/04/144109
- ndarrayのリスト変換
- https://www.python-izm.com/data_analysis/numerical/numpy/ndarray_tolist/
- numpy tips
- https://qiita.com/wwacky/items/4f475d44753819f98083

よく、復習を行ってください。
また、課題が出ているかは、毎回、manabaを確認してください。