

クレジット:

UTokyo Online Education データマイニング入門 2018 森 純一郎

ライセンス:

利用者は、本講義資料を、教育的な目的に限ってページ単位で利用することができます。特に記載のない限り、本講義資料はページ単位でクリエイティブ・コモンズ 表示-非営利-改変禁止 ライセンスの下に提供されています。

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

本講義資料内には、東京大学が第三者より許諾を得て利用している画像等や、各種ライセンスによって提供されている画像等が含まれています。個々の画像等を本講義資料から切り離して利用することはできません。個々の画像等の利用については、それぞれの権利者の定めるところに従ってください。



# データマイニング入門

2018年度A semester 水曜5限

数理・データサイエンス教育プログラム

# 本日の授業の流れ

- ガイダンス
  - 本授業について
    - 授業計画、学習目標、成績評価方法など
- Pythonプログラミング環境について
- データマイニングの導入

# 授業について

- 講義題目：データマイニング入門
- 時間割コード：0590103
- 曜限：水曜・5限
- 開講区分：Aセメスター
- 単位数：2.0
- 学年：B3,4,5,6,M,D
- 教室：理学部1号館287講義室

# 講師

- 自己紹介
  - 森 純一郎
    - 数理・情報教育研究センター 准教授
    - 大学院情報理工学系研究科 電子情報学専攻 兼任
    - 専門分野：人工知能、大規模データ解析、ネットワーク分析
- コンタクト・オフィスアワー
  - ITC-LMSの講義ページの掲示板で質問を受け付けます

# 授業の概要

- データ分析・データマイニングの基礎について学ぶとともに演習を通して実際にデータを分析するプロセスを学ぶ
- 学部後期課程におけるデータ分析関連の講義・研究の基礎となる知識を習得する
  - データサイエンス、人工知能、機械学習、自然言語処理などの関連講義
  - データ分析を用いた実験・演習・プロジェクトや卒業研究
    - 理学部・工学部に限らず、ほぼすべての学部においてデータ分析を用いる機会がある

# 後期課程におけるデータ分析関連講義

- 工：機械学習の数理、統計的機械学習、知能機械情報学、言語・音声情報処理、確率数理工学、システムデータ解析、応用データ解析 など
- 理：統計的機械学習、知能システム論、生物データマイニング論、生命情報表現論、地球物理データ解析 など
- 農：バイオインフォマティクス など
- 薬：生物統計学 など
- 医：医学データの統計解析、バイオインフォマティクス など
- 経：経済データ分析、計量経済学、ベイズ計量経済学、経営科学 など
- 文：社会心理学、心理学統計、社会学のためのデータ分析法 など
- 教育：教育政策研究方法論、教育社会学調査実習 など
- 教養：言語データ分析、計算社会科学、意思決定・知的システム論 など

# 数理・データサイエンス教育プログラム

- 学部横断型プログラム

<http://www.mi.u-tokyo.ac.jp/mds-oudan/index.html>

- 理系・文系にまたがる体系化された数理・データサイエンスに関する講義科目
- 理系・文系を問わず将来の研究あるいは実務の面において必要になる数理・データサイエンス分野に関する基礎的知識と技術を身に付ける
- 後期課程学生対象、大学院学生も履修可能

- 科目群

- 確率論、確率過程論、工学のための現代数学入門、最適化手法、時系列解析、統計データ解析、確率統計学基礎、**Pythonプログラミング入門**、**データマイニング入門**、コンピュータシステム概論、社会科学のための統計分析



# 学内のデータ分析・サイエンス公開講座

- 東京大学データサイエンティスト養成講座
  - <http://dss.i.u-tokyo.ac.jp/about.html>
  - 大学院生対象
  
- 東京大学グローバル消費インテリジェンス寄付講座
  - <http://gci.t.u-tokyo.ac.jp/>
  - データサイエンス講座
  - 本学学生対象

# 授業計画（全13回予定）

- 9/26 ガイダンス
- 10/3 データ分析のためのプログラミング基礎
- 10/10 データ分析のためのプログラミング基礎
- 10/17 データの前処理・加工、データベースの基礎
- 10/24 テキストデータ分析の基礎
- 10/31 休講予定
- 11/7 ネットワーク・グラフデータ分析の基礎
- 11/14 教師なし学習の基礎
- 11/21 教師なし学習の基礎
- 11/28 教師あり学習の基礎
- 12/5 教師あり学習の基礎
- 12/12 データ分析の実践
- 12/19 ミニプロジェクト
- 1/9 発展的な内容と全体まとめ

# 学習目標

- データ分析の基本的なプロセスを理解する
  - 前処理、管理、分析、評価
- データ分析のためのプログラミング基礎を理解する
  - Pythonを用いる
- データ分析のための数理的基礎を理解する
  - 特に、確率・統計、線形代数、解析学
- 代表的なデータの数理モデルと基礎的な処理を理解する
  - テキストデータ、グラフデータ、時系列データ
- データマイニング・機械学習の基礎的な手法を理解する
- データ分析プロジェクトの基礎的な設計と実施ができる

# 学習内容（予定）

- データ分析のためのPythonプログラミング
  - Pythonの基本的な文法とNumpy, Scipy, Pandas, Matplotlibなどの科学技術計算のための主要なモジュール
- データ分析のための数理的基礎
  - 記述統計、分布・検定、ベクトル・行列、固有値分解、最適化基礎など
- データの前処理・加工とデータベース
  - サンプルングと信頼区間、欠損値・ノイズ・外れ値の処理、関係データベース など
- テキストデータ分析の基礎
  - tfidf、Bag of Words、ベクトル空間モデル、形態素解析、類似度、潜在意味解析 など
- ネットワーク・グラフデータ分析の基礎
  - 接続行列, 最短距離, クラスタリング係数, 中心性, コミュニティ抽出, ネットワークの数理モデル など
- 機械学習の基礎（教師あり学習）
  - 線形回帰、ロジスティック回帰、正則化、モデル選択、ニューラルネットワーク など
- 機械学習の基礎（教師なし学習）
  - k-means、階層化クラスタリング、EMアルゴリズム、主成分分析など

# 授業の方法

- 授業前半
  - スライドを用いた講義
- 授業後半
  - 前半の講義内容を元に、jupyter notebook上で実装と課題の説明
  - 後半部分は各人のノートPCで実際にコードを実行しながら説明を聞いてもらえる  
とよいですが、コードの実行ができなくても理解できるように説明します
  - プログラミング課題に次週の授業までに取り組んでもらい提出してもらいます
    - 授業の説明を聞いていれば実装できるような内容に設定します

# 履修上の注意

- ITC-LMSでコース参加登録をしてください
- プログラミング課題に取り組める環境（後述）の準備をしてください
- 高校数学の知識を前提とします

# Python

- フリー・オープンソースのインタプリタ・スクリプト言語
- 科学技術計算に利活用
  - データ分析、機械学習、数値計算、最適化、画像・信号処理、可視化 など
- 設計思想「シンプルで読みやすいコード」
  - The Zen of Python
- Pythonエコシステム

著作権の都合により  
ここに挿入されていた画像を削除しました

Jake Vanderplas - Keynote - PyCon 2017  
<https://www.youtube.com/watch?v=ZyjCqQEUa8o>

# Anaconda と Jupyter Notebook

- Anaconda
  - <https://anaconda.org/>
  - Pythonと主要なライブラリ（モジュール）を一括でインストール可能にしたディストリビューション
    - 以下のJupyter Notebookも含む
  - ECCSにはすでにインストールされている
    - Python2系とPython3系があり、講義ではPython3系を使用
- Jupyter Notebook
  - <http://jupyter.org/>
  - Pythonコードをブラウザ上で対話的に実行可能な環境
  - ノートブックドキュメントを通して、ブラウザ上でコードの記述と実行結果の保存と共有が可能
  - ECCSではAnaconda Navigatorから起動できる



# データマイニングとは

- KDD (Knowledge Discovery in Databases)
  - The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data [Fayyad 96]
    - 汎用的で、新しく、有用で、理解できるパターンの発見
      - データに内在する規則や特徴的なパターン
  - データマイニング
    - KDDプロセスの中のパターン発見の段階
- KDDのプロセス
  - データの同定・収集・選択
  - 前処理・変換
  - マイニング (パターン発見)
  - 結果の解釈・評価・活用

最終的には意思決定・問題解決

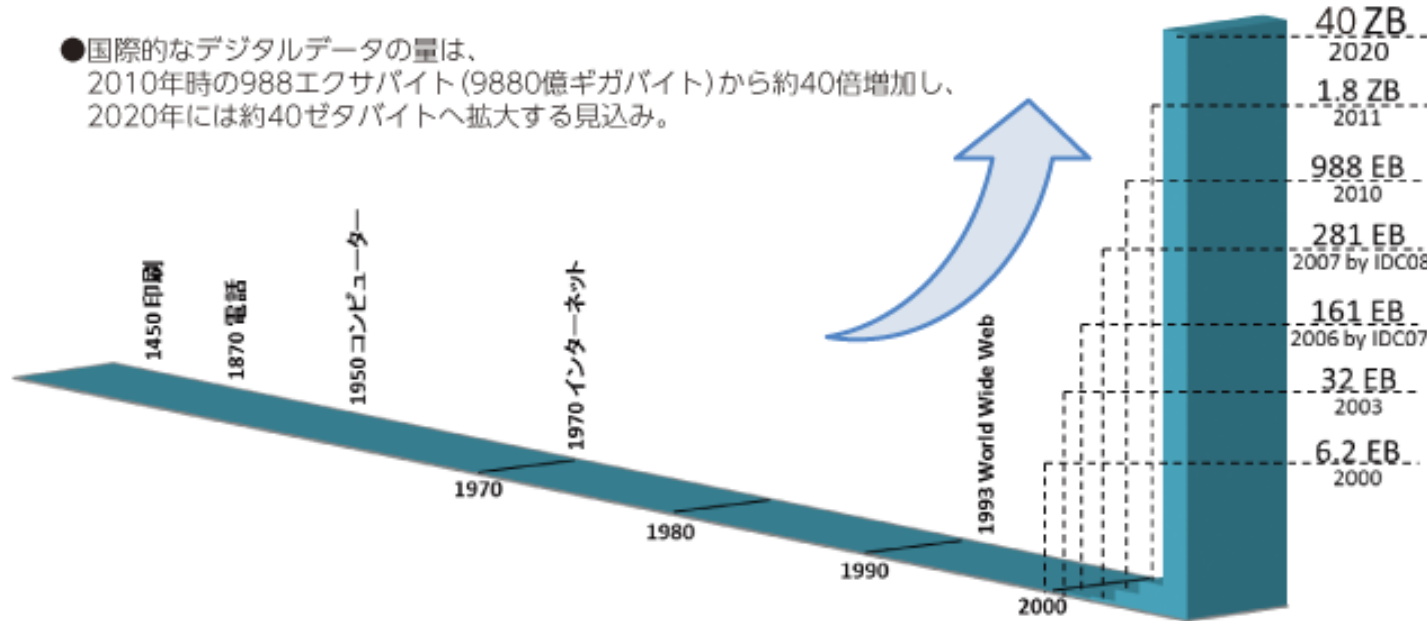
# 古典的なデータマイニング

- データウェアハウス [Inmon 96]
  - 継続的に収集されるデータを分析しやすい形式で運用するための枠組み
    - 例：POS（販売時点情報）データ
      - 属性（フィールド）からなるレコード
        - [ID、店舗、商品名、値段、個数、日時、…]
- データマイニング in 90's
  - データからの規則やパターンの発見
    - 頻出パターン
      - 相関ルール (Association Rule)
  - 購買データの分析からスーパーマーケットでビールとおむつが一緒に買われていることがわかった [wall street journal 92]

# ビッグデータの時代に

ビッグデータ：Volume, Variety, Velocity [Zikopoulos 11]

●国際的なデジタルデータの量は、  
2010年時の988エクサバイト(9880億ギガバイト)から約40倍増加し、  
2020年には約40ゼタバイトへ拡大する見込み。



(出典) 総務省「ICTコトづくり検討会議」報告書

ビッグデータの応用

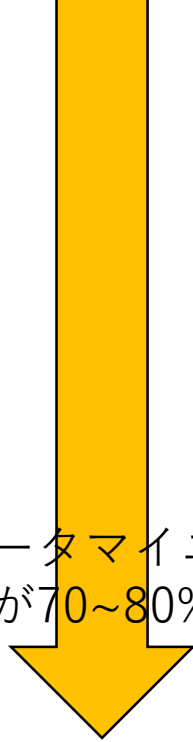
- 創薬
- 気象 (ゲリラ豪雨)
- 地震・津波
- ゲノム
- パンデミック
- 宇宙 (望遠鏡、衛星)
- 生物発生
- 農業
- 科学技術文献
- ...

# ビックデータとデータマイニング

- より大規模なデータからのマイニングに
  - 店舗での購買 (POS) データから…
  - インターネットの普及とビックデータ in 2000's
    - オンラインショッピング、検索、ソーシャルメディア
      - 購買履歴、趣味・嗜好、共有
  - デバイス・センサーの発展・普及とビックデータ in 2010's
    - スマートフォン、ホームスピーカー
      - 位置、生体情報、文脈
- 新たな課題
  - スケーラビリティ、高次元、非構造化、マルチモーダル など
- パターン発見から知識抽出、価値創造へ
  - データの見えざる手 [矢野 14]
    - 仮説検証から仮説生成へ
    - データを活用して稼ごうとすることで、社会の幸福や共感が高まる

# データマイニングのプロセス

- データの同定・収集・選択
  - 対象領域の理解
  - データ集合からマイニングに必要なデータセットを選択
- 前処理
  - データセットからノイズや異常値をクレンジング
    - 連続値の離散化・カテゴリー化、欠損の補完、外れ値・重複除去
    - 複数データの統合、正規化、表記・単位の統一、矛盾解消 など
- 変換
  - 前処理されたデータをアルゴリズム入力用に変換
    - 属性選択、属性抽出、属性構築・統合、事例選択
      - サンプルング、相関、次元削減、外部知識利用 など
    - 構造データと非構造データ
- マイニング（パターン発見）
  - 統計分析、機械学習、データマイニング手法
- 結果の解釈・評価・活用
  - 抽出したパターンを解釈・評価して知識を得る
    - 可視化 など



実際のデータマイニングでは  
変換までが70~80%とされている

# データマイニングのプロセス（産業）

CRIPS-DM (cross industry standard process for data mining) [Shearer 00]

- ビジネスの理解
  - 解決したい課題・問題
- データの理解
  - データセットの選択と仮説
- データの準備
  - 前処理・変換
- モデリング
  - 統計分析、機械学習、データマイニング手法
    - モデル選択とパラメータ推定
- 評価
  - モデルの検証
- 適用
  - モデルの組み込み

# データサイエンスとは

- データサイエンティストのプロフィール [Schutt 14]
  - 計算機科学、数学、統計学、機械学習、専門知識、コミュニケーション、プレゼン、可視化
- Data Science for Undergraduates [USADS 16]
  - 数学基礎
  - 統計基礎
  - 計算（コンピュータ科学・情報）基礎
  - データ管理とキュレーション
  - データ記述と可視化
  - データモデリングと評価
  - ワークフローと再現性
  - コミュニケーションとチームワーク
  - 領域知識の考慮
  - 倫理的な問題への対応

# データサイエンスのプロセス

以下のサイクルを回す

- 問題設定
- データ収集
- データ処理（前処理含む）
- 探索（Exploration） ・ 可視化
- 分析 ・ 予測 ・ 推論
  - 探索と分析を合わせてExploratory Data Analysisとも呼ばれる
- 洞察（Insight） ・ 決定



# データサイエンティスト スキルレベル

## 見習いレベルの例

- 統計数理の基礎知識を有している（代表値、分散、標準偏差、正規分布、条件付き確率、母集団、相関、ベイズの定理など）
- データ分析の基礎知識を有している
  - 予測（回帰係数、標準誤差・・・）、検定（帰無仮説、対立仮説・・・）、グルーピング（教師あり学習、教師なし学習・・・） など
- 適切な指示のもとに、データ加工を実施できる
  - 基本統計量や分布の確認、および前処理（外れ値・異常値・欠損値の除去・変換や標準化など）
- データ可視化の基礎知識を有している（ヒストグラム、散布図、積み上げ棒グラフなど）

# データサイエンティスト スキルレベル

## 独り立ちレベルの例

- 基礎的なデータ加工については自律的に実施できる
  - 外れ値・異常値・欠損値の対応
  - 適切な学習データとテストデータの作成
- 基礎的な分析活動については自律的に実施できる
  - 多重共線性を考慮した重回帰分析
  - パラメトリックな 2 群の検定の活用 (t 検定)
  - 適切な初期値設定を行った非階層クラスター分析
  - 主成分分析や因子分析
  - 機械学習における過学習の理解
  - 形態素解析などを用いた基本的文書構造解析 など

# データソース

- ウェブ
  - 検索、スクレイピング、API、ソーシャルメディア …
- オープンデータ
  - 各国のオープンデータ
    - <http://opendata-portal.metro.tokyo.jp/www/index.html>
    - <http://www.data.go.jp/> <https://www.data.gov/> <https://data.gov.uk/> <http://data.un.org/>
    - <https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/>
    - [https://qiita.com/tmp\\_llc/items/7296c5d6bb8769b18d24](https://qiita.com/tmp_llc/items/7296c5d6bb8769b18d24)
- データ分析コンペ
  - <https://www.kaggle.com/datasets>
- 研究用データセット
  - <https://archive.ics.uci.edu/ml/index.php>
  - [https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_for\\_machine\\_learning\\_research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research)
- データセット検索
  - <https://toolbox.google.com/datasetsearch>

# 次回までに

- Pythonプログラミング環境（jupyter notebookが動作する環境）を準備してください
- 配布教材のpython\_for\_data\_analysis1.ipynbを予習しておいてください
- 内容が難しいと感じたらPython入門の本やPythonプログラミング入門の教材で自学を勧めます

- Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From Data Mining to Knowledge Discovery in Databases', *AI Magazine* , 37-54.
- W. H. Inmon : Building the Data Warehouse, John Wiley & Sons, Inc,
- Wilke, J. R. Retailing: Supercomputers manage holiday stock. Wall Street Journal. 1992-12-23
- Paul Zikopoulos, Chris Eaton. 2011. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (1st ed.). McGraw-Hill Osborne Media.
- データの見えざる手: ウェアラブルセンサが明かす人間・組織・社会の法則 単行本 – 2014/7/17
- Shearer, C. (2000), 'The CRISP-DM Model: The New Blueprint for Data Mining', *Journal of Data Warehousing* **5** (4).
- Rachel Schutt and Cathy O'Neil. 2013. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc.
- National Academies of Sciences, Engineering, and Medicine. 2018. Data Science for Undergraduates: Opportunities and Options. Washington, DC: The National Academies Press.