Introduction

Tadashi Mori

2019.9.26

Ver.2.0

これ説明できますか？

- AI、機械学習、ディープラーニングの違い
- データサイエンティストとは
- ビジネスとデータの関係とは（目的意識）

Agenda

1. Python Cafeの目的とゴール
2. AI、機械学習、ディープラーニングの違い
3. データサイエンティストとは？
4. ビジネスにデータがどう活かされているのか？
5. 分析プロセス・アルゴリズム
6. 教師あり学習　回帰
7. 教師あり学習　分類
8. 過学習・評価指標
9. 最終課題・最終プレゼン

1.1. Python Cafeの目的とゴール

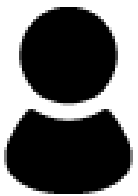
目的：関係性について

「アルムナイ・リレーションシップ」という立場で三菱自動車と関係性を保ち、会社の成長への寄与と新しい価値観を共有していく。

参考：https://news.mynavi.jp/article/20181208-736941/

目的：Win x Winな関係

僕



・インプットしたことをアウトプットすることで知識の定着化と理解を深める
・人に教えるワザを磨く

みなさん

・データ分析、機械学習の世界がイメージできるようになる
・データサイエンティストのスターターになれる

ゴールイメージ（メンバー）

❶データサイエンスに関する言葉を知る。理解する。

❷Python(juypter notebook)の使い方を知る。
自分で調べてコードの書き方やエラー解決ができるようになる。

❸単純な業務用データを活用して、アウトプットできるようになる。
アウトプットの根拠が説明できるようになる・・・？

ゴールイメージ（マネジメント）

❶活動結果を数値で評価する。

❷社内活動（部内教育）ができる仕組みのネタとする。

❸単純な業務用データを活用したアウトプットのプレゼンに対して良し／悪しを判断できる・・・かも・・・

数値評価

■データサイエンススキルチェックで成長ぶりを振り返ろう
できれば、下記URLを実施して、最初と最後の自分の成果を振り返ってみてください。

https://check.datascientist.or.jp/skillcheck-full/

スケジュール

day1	イントロ 予測 統計学によるCS分析 アンケート調査結果 統計学の触りを知る	day5	分類 サポートベクターマシン 手書き文字データ スクリプト実行、読み解く
day2	予測 単回帰・重回帰 Bostonデータ jupyter notebookに慣れる	day6	予測 オープンデータ活用 決定木・ランダムフォレスト データクレンジング（欠損値処理）を学ぶ
day3	分類 ロジスティック回帰 Irisデータ jupyter notebookに慣れる	day7	最終プレゼンの事前学習 これまでの振り返り
day4	過学習について ホールドアウト法、交差検証 検証方法について 評価指標	day8	最終課題のプレゼン 目的の達成度確認

2. AI、機械学習、ディープラーニングの違い

AI、機械学習、ディープラーニングの違い

https://news.mynavi.jp/itsearch/article/devsoft/3120

・AIは、処理をプログラムで定義した技術として広い意味で解釈される

・機械学習は、処理（振る舞い）をコンピューターが学習すること

・DLは、ニューラルネットワーク技術を用いて、処理（振る舞い）をコンピューターが学習すること

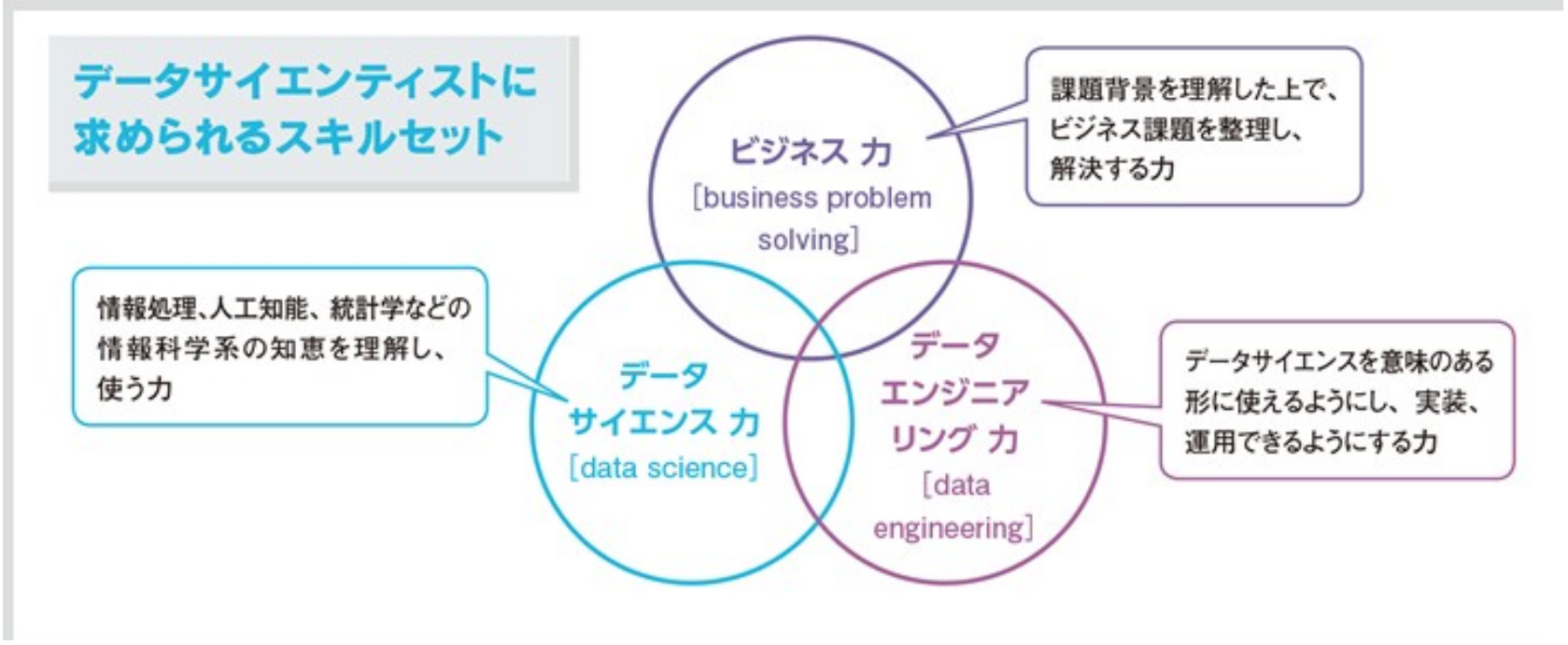
機械学習（DL）は３つに区別される

https://www.itmedia.co.jp/enterprise/articles/1901/07/news015.html

3. データサイエンティストとは？

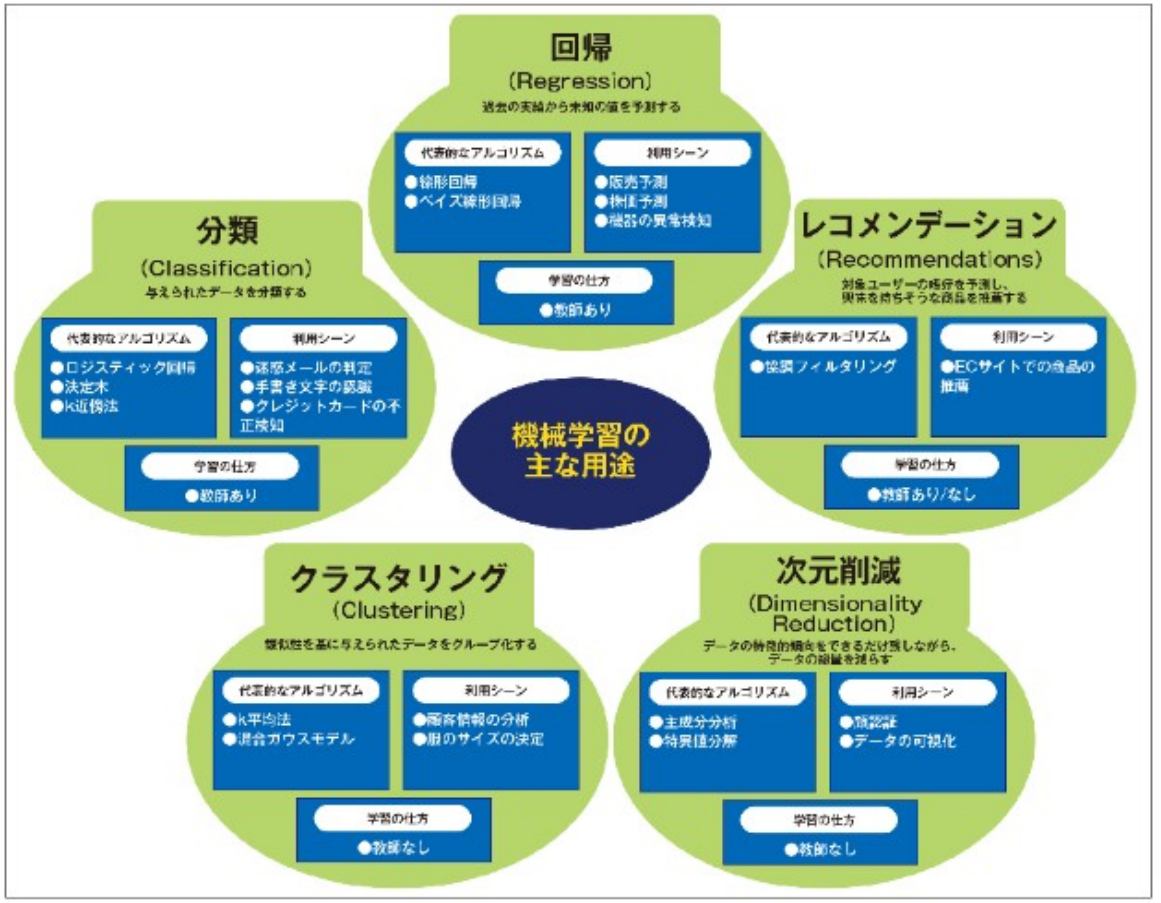
データサイエンティストとは

データサイエンティストとは、**大量のデータ（ビッグデータ）からビジネスに活用できる情報を引き出す**専門技術者のこと。
データサイエンティストは、**ビッグデータの分析及び分析結果をもとに、問題の解決や状況改善のための施策立案**を行う。



4. ビジネスにデータがどう活かされているのか？

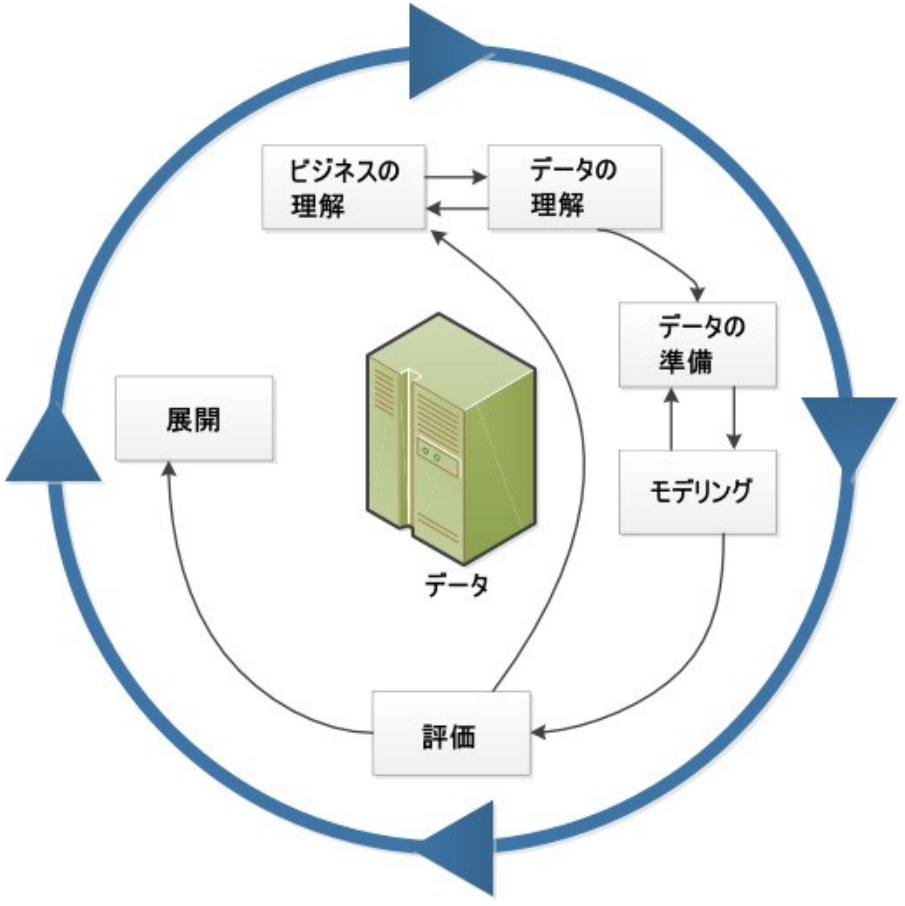
ビジネスにデータがどう活かされているか



https://xtrend.nikkei.com/atcl/contents/18/00076/00002/

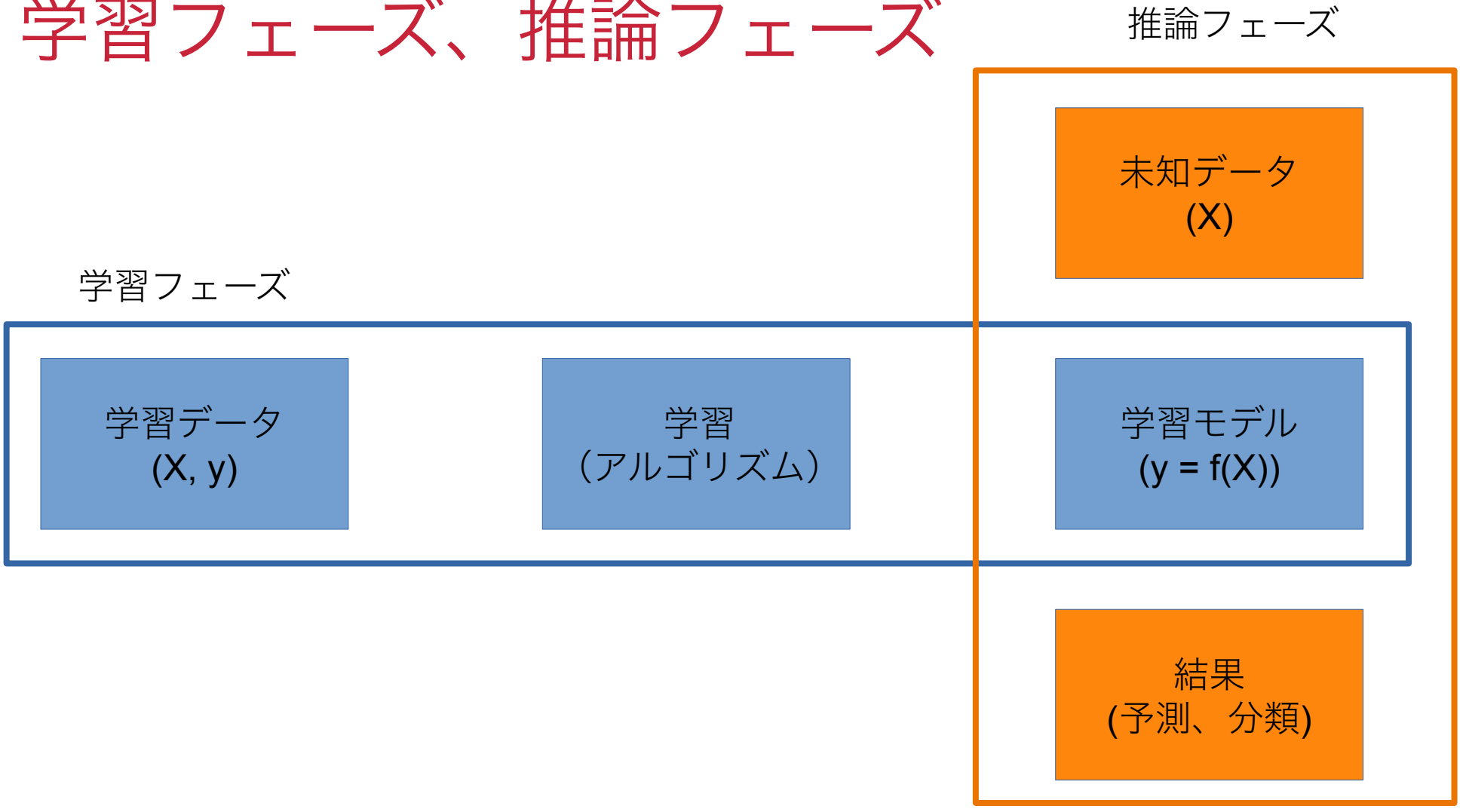
5. 分析プロセス・アルゴリズム

CRISP-DM

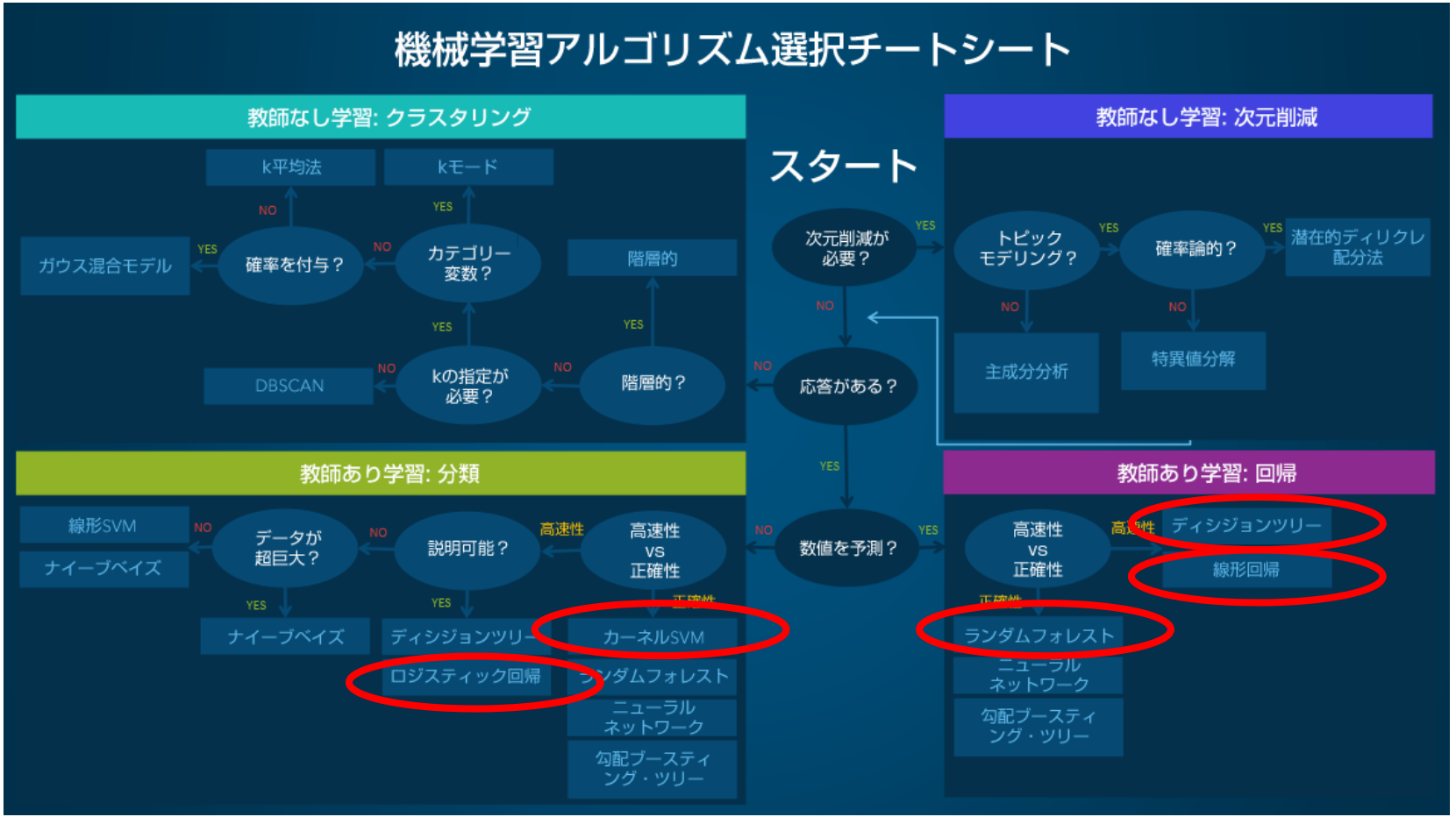


https://www.ibm.com/support/knowledgecenter/ja/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html

学習フェーズ、推論フェーズ



アルゴリズム



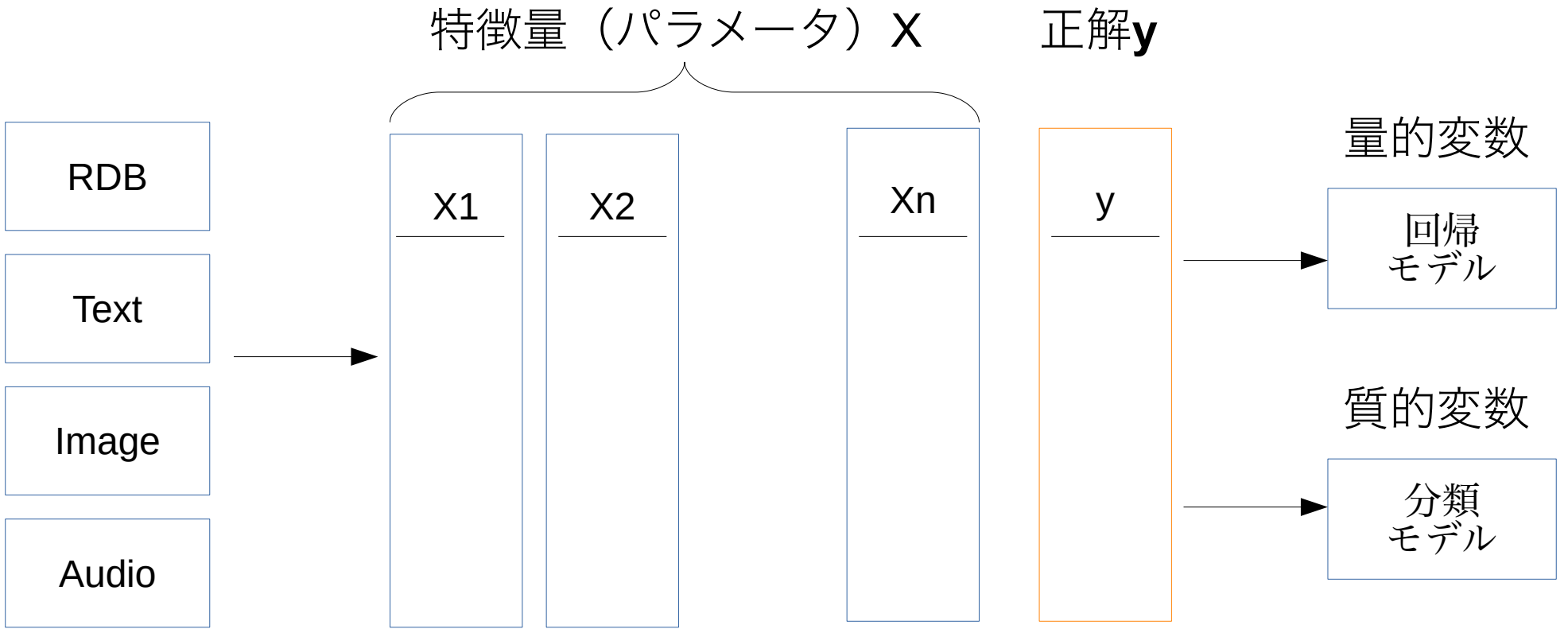
引用：SAS Japan
http://www.sascom.jp/blog/2017/11/21/machine-learning-algorithm-use/

6. 教師あり学習　回帰

回帰は、数値データを扱い数値を予測する
数値データによるビジネスニーズに適応される

教師あり学習（Supervised learning）

特徴量（パラメータ）Xと結果yの関係性を学習するのが「教師あり学習」

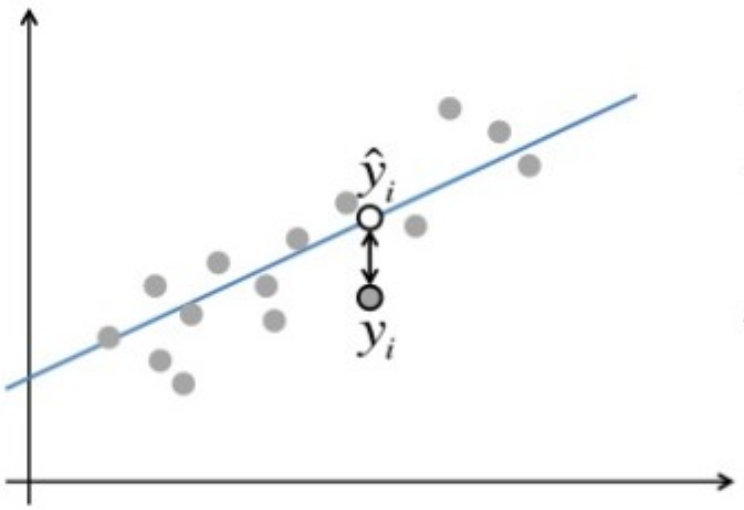


線形回帰・重回帰

・線形回帰は、最小二乗法で解く。説明変数から目的変数を求める手法。
・重回帰は説明変数（パラメータ）が複数あるもの

最小二乗法による推定

実測値と予測値の二乗誤差を最小化する

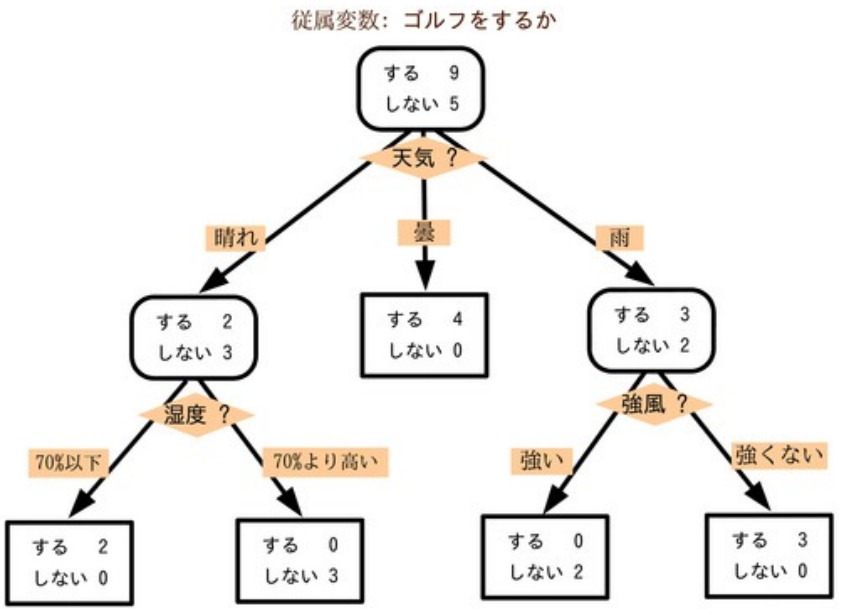


二乗誤差Ewを
以下のように定義:

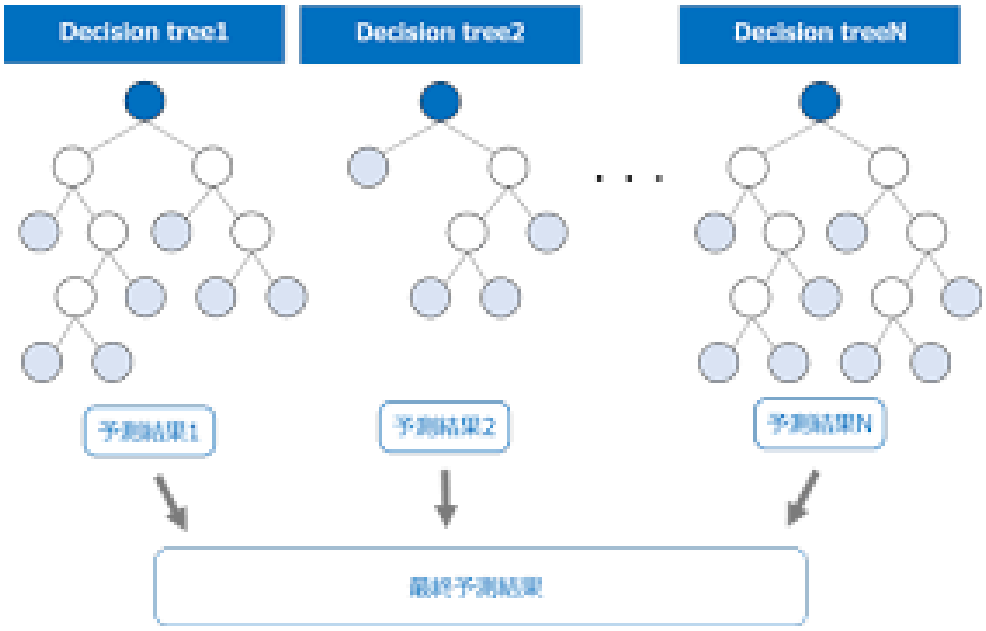
$$E_D(w) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

https://www.slideshare.net/siritori/6-35685092

決定木・ランダムフォレスト



https://ja.wikipedia.org/wiki/%E6%B1%BA%E5%AE%9A%E6%9C%A8

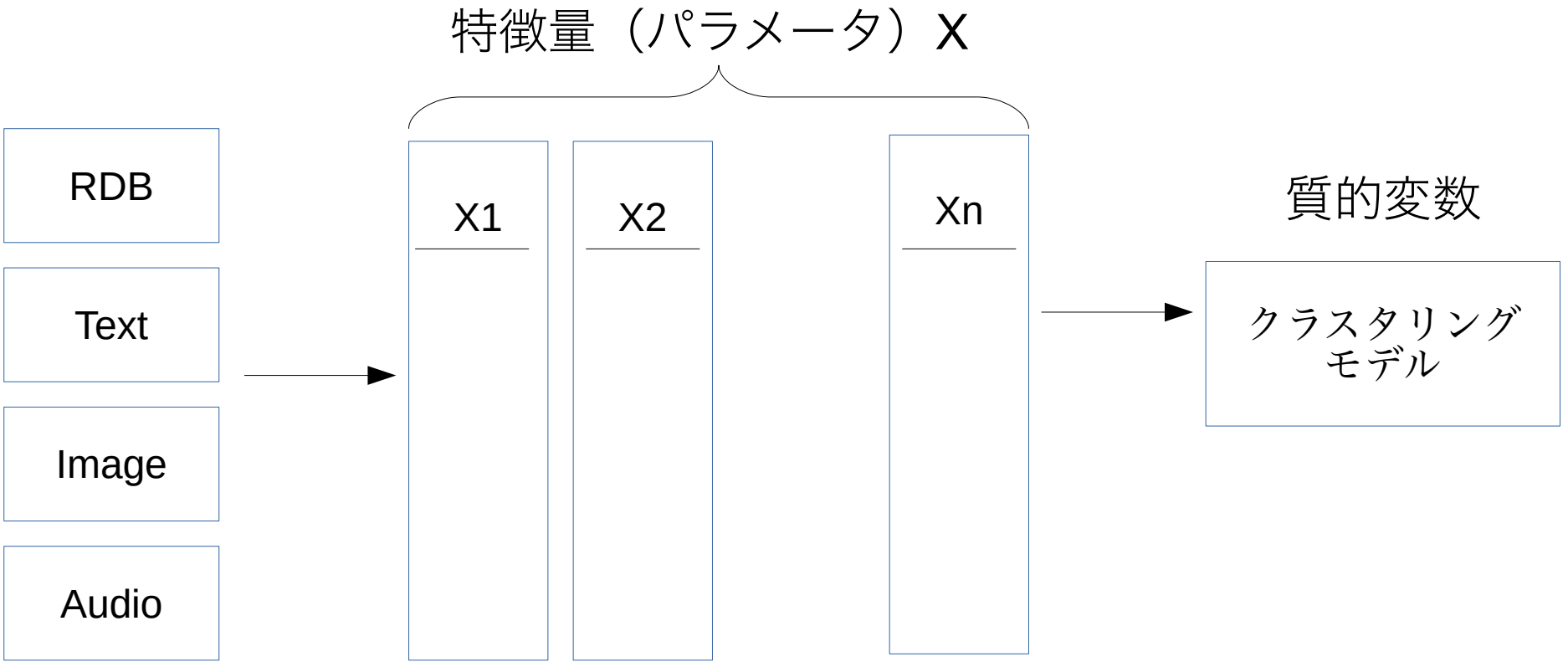


http://www.stats-guild.com/analytics/12543

7. 教師あり学習　分類

教師なし学習（Unsupervised learning）

データをパターン化する「教師なし学習」



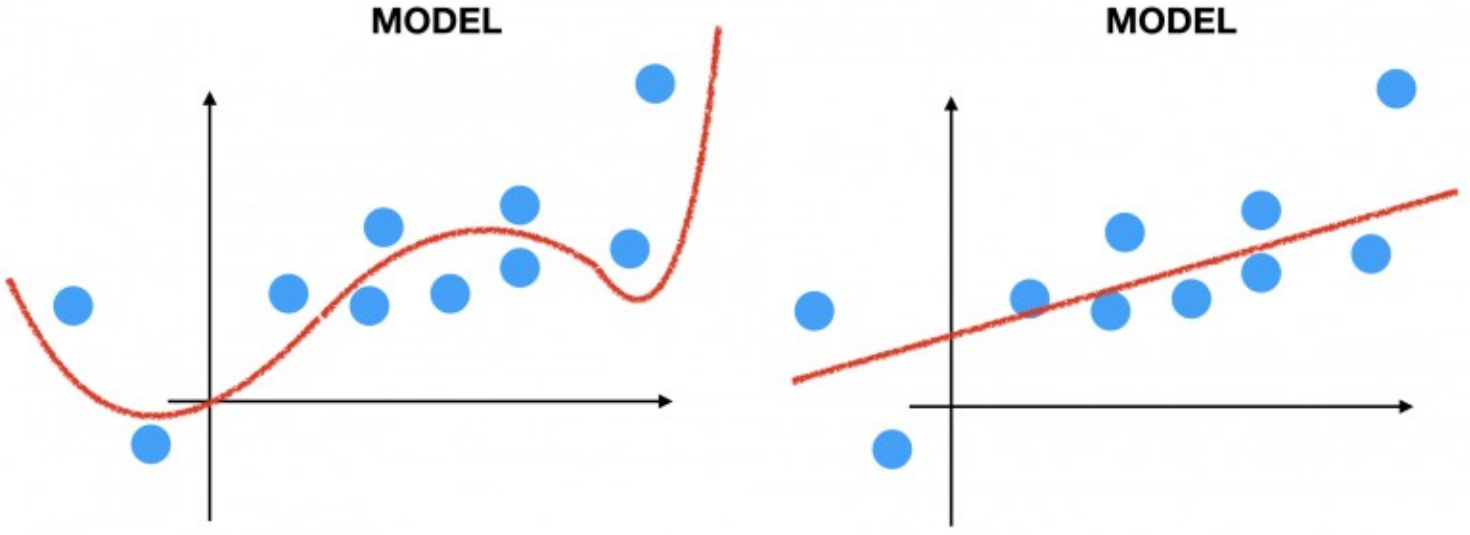
ロジスティック回帰

サポートベクターマシーン（SVM）

8. 過学習・評価指標

過学習とは

過学習（Over fitting）とは、学習データに適合しすぎたモデルのこと。
未知のデータに対して汎化能力が低下している。
機械学習は過学習との戦いとも言われる！



https://research.miidas.jp/2019/01/ã ã ‡pythonã€ æ£å‰‡åŒ–ã £ã ¦ä½•ï¼Ÿpart-1/

ホールドアウト法

交差検証（クロスバリデーション）

評価指標

9. 最終課題・最終プレゼン

最終課題・最終プレゼン

- どういう目的で参加したか
- 学んだこと（言葉の説明、モデリングの説明）
- データサイエンティストスキルチェックの結果
  見せなくてもOK！
- 感想
- プレゼンのフォーマットは自由
  juypter notebookを使うのが好ましい！

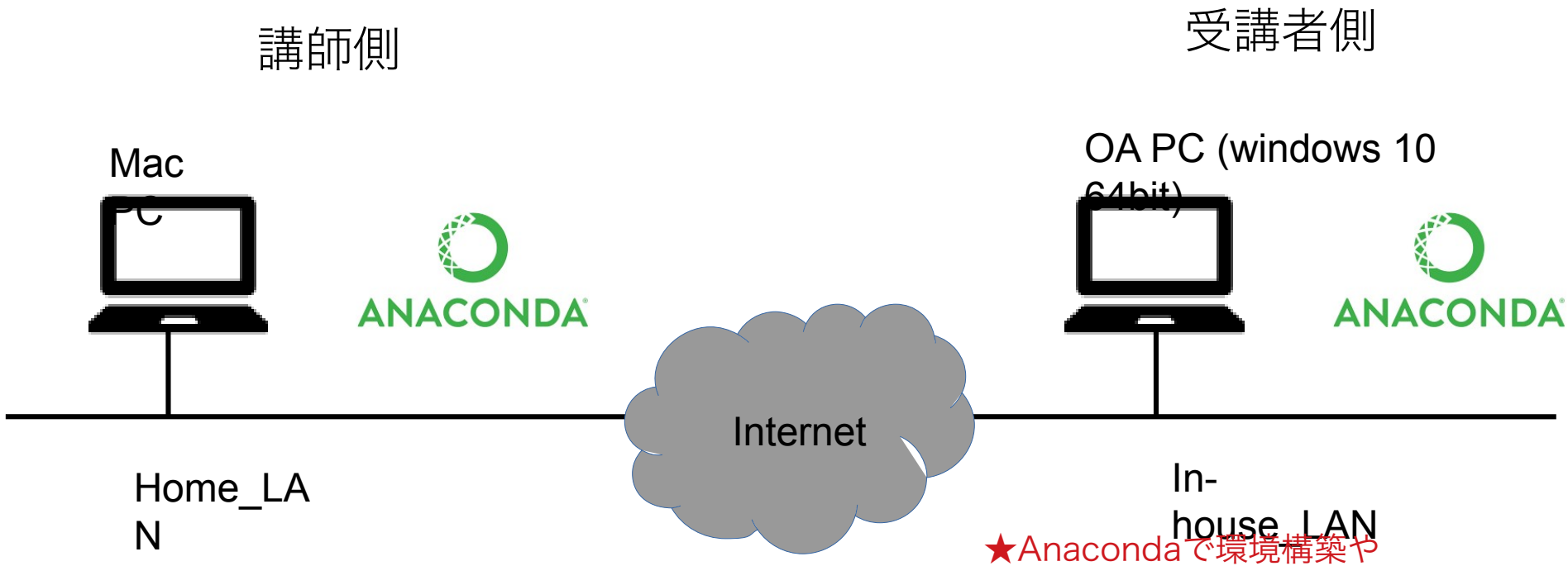
END

コニュミケーション方法

Tools	Advantages	DIsadvantages
Skype	・画面共有できる ・扱いやすい ・グループ接続可能	・for Businessと 一般とで制限がある ・for BusinessのWeb版で繋がるかも？
Googleハングアウト https:// https://hangouts.google.com/	・画面共有できる ・ブラウザで動作 ・グループ接続可能 https://cloud-work.jp/productivity/google-hangout/	・IEはプラグインをインストールしないといけない（管理者権限が必要かも・・・）

ファイル共有はBoxを活用する
https://app.box.com/folder/87062787914

プログラミング環境



★Anacondaで環境構築やパッケージインストールなどするので一定期間の管理者権限付与が必要と思われる