新人向け データ分析研修

2021/10 Sekikawa

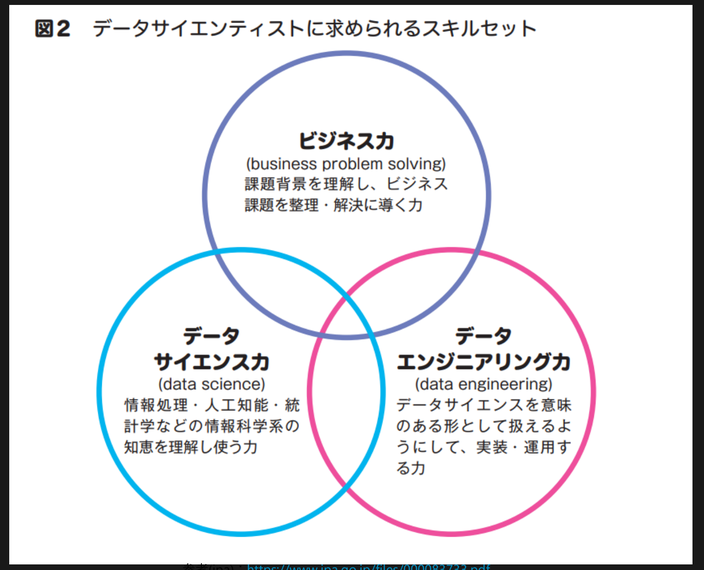
この研修のゴール・目的

- **データ分析のイメージを感じ取ってもらう**
  - データ分析を興味を持ってもらうきっかけ作り
- **簡単なデータ加工・集計・モデリングができる**
  - Pythonのプログラムを触って実体験する
- **楽しんでもらう！**
  - 業務で必要不可欠な内容ではないので気楽にやりましょう！

データ分析の基礎知識

データ分析関連の職種

- 単にデータ分析といっても, 多くの職種がある
  - 色々な人が協力してデータ分析を行う



参考(ipa)　https://www.ipa.go.jp/files/000083733.pdf

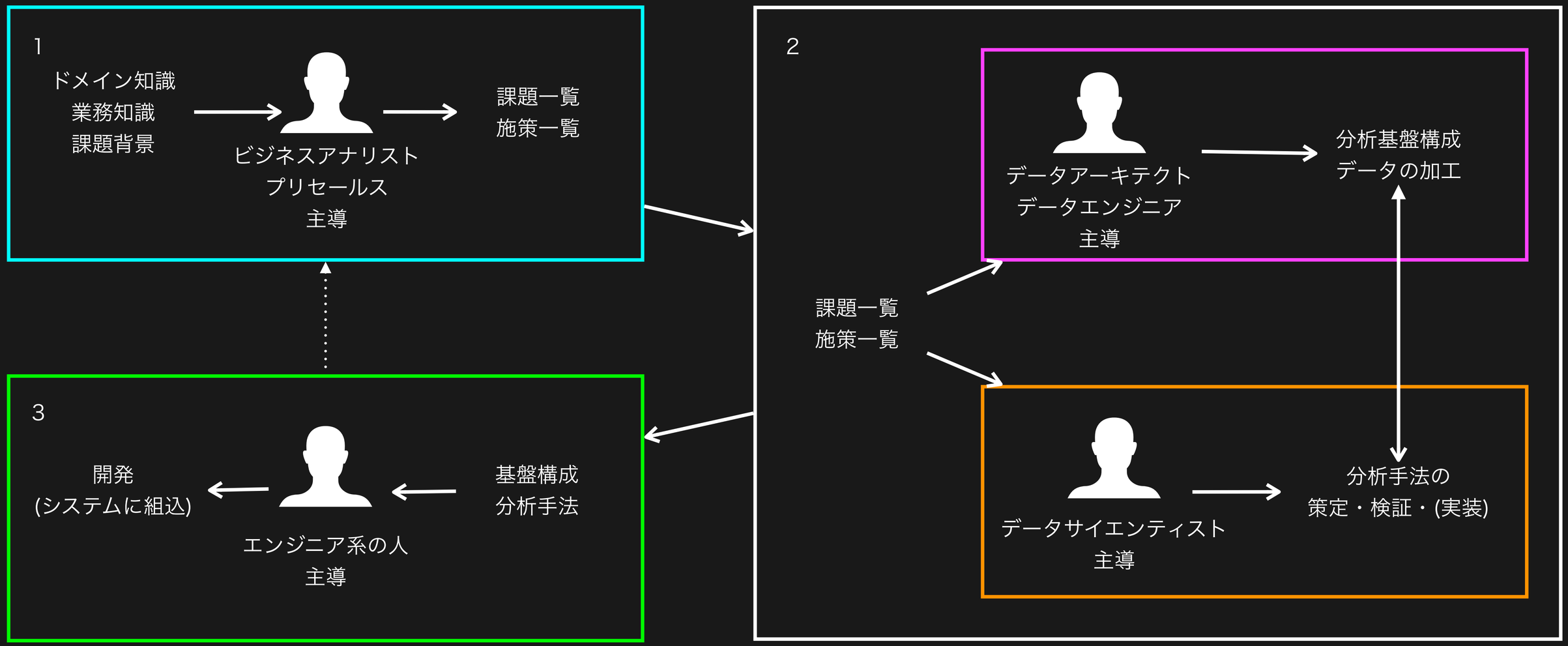
職種名	BZ	DS	DE	内容
ビジネスアナリスト	★★★	★★☆	★☆☆	統計的な分析を用いてビジネスの施策を導く
データサイエンティスト	★★☆	★★★	★☆☆	機械学習・統計を用いたモデリング・検証・レポート
データエンジニア	★☆☆	★☆☆	★★★	データの集計・抽出の仕組みを作る
機械学習エンジニア	★☆☆	★★☆	★★☆	機械学習モデルをシステムに組み込む
プリセールス	★★★	★☆☆	★☆☆	抽象度の高い顧客要望をデータ分析課題に落とし込む

★はスキルが求められる頻度

BZ：ビジネス力
DS：データサイエンス力
DE：データエンジニアリング力

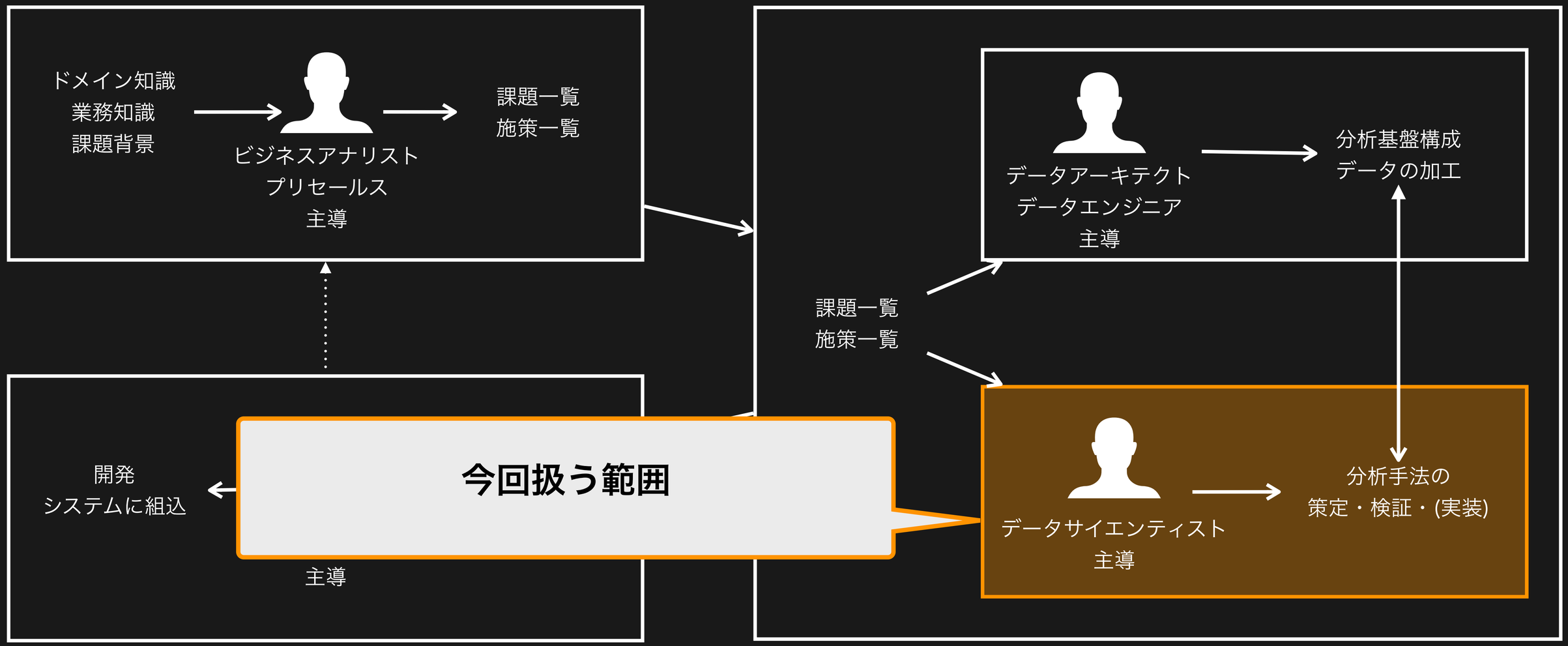
データ分析の流れ

注意：あくまで自分のイメージ, 実際にはここまで綺麗に分業されていないはず



データ分析の流れ

注意：あくまで自分のイメージ, 実際にはここまで綺麗に分業されていないはず



データサイエンティストの担当内容

- **どんな分析技術を使えば良いか判断して, 分析手法の実装と検証を行うことがメイン**
  - 施策を考えたり, 分析基盤を整えることにも参加する(はず)

- **施策の例) 住宅販売価格をデータに基づいて, 適切な値で販売する**
  - どのようなデータが必要で, どのような加工が必要か
  - どのような分析手法で実現するのが最適か, どのような改善策が考えらえるか
  - 実装した結果として, 望むような結果が得られるか

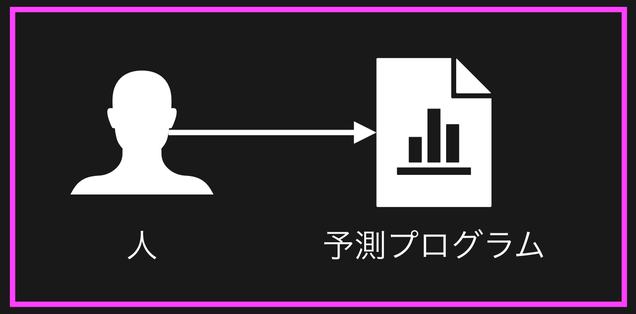
2日間の研修のスケジュール(予定)

- 1日目
  - 機械学習の基礎
  - Pythonの基礎
  - データ分析ライブラリの基礎
- 2日目
  - データ分析コンテストのデータを使って実践してみる

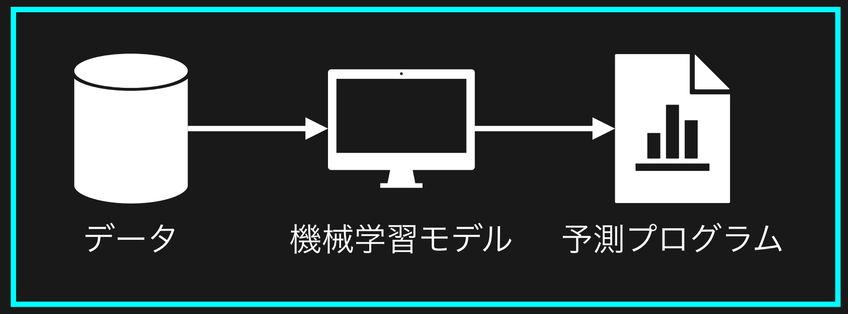
機械学習の基礎知識

機械学習とは

- 目的を達成するための知識や行動を，データから機械に獲得させる技術
  - パターンを探し出すアルゴリズムの総称
  - 自らルールを記述せずとも，プログラムがルール(パターン)を見つけてくれる
  - データを元に何かを予測するタスクなどで使用される



人がプログラム(ルール)を記述する

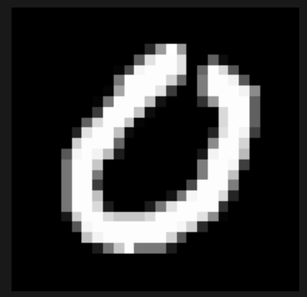


機械学習がデータを元にルールを出力する

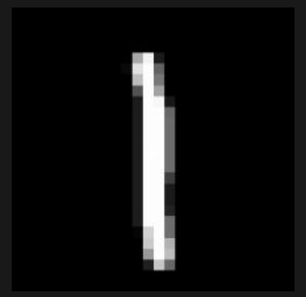
画像分類の例

- 0と1が書かれた画像をプログラムで分類することを考える
  - どんなプログラムを書いて実現する？

0のグループ

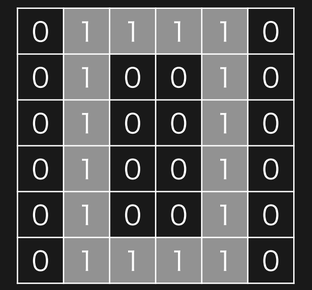



1のグループ

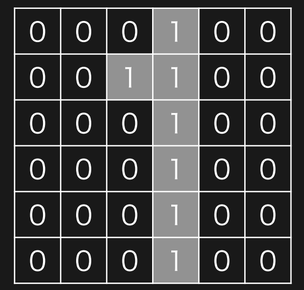



自分でルールを記述する場合

```
> data_0 = [[0,1,1,1,1,0], [0,1,0,0,1,0], […], […], […], […]]
> data_1 = [[0,0,0,1,0,0], [0,0,1,1,0,0], […], […], […], […]]
```

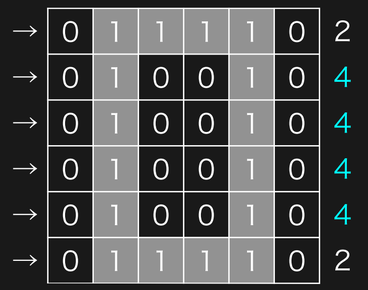


0の画像

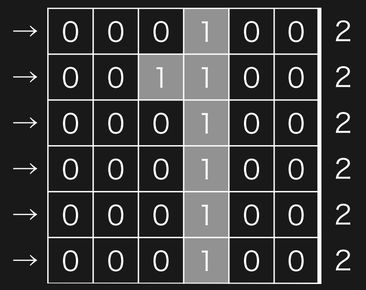


1の画像

1. 「行数」回ループする
    1. 行方向(横方向)に対して差分を計算する
    2. [0→1], [1→0]になった回数が4回以上なら, [count += 1]
2. [count > 行数/3] の場合, 0を出力する



0の画像

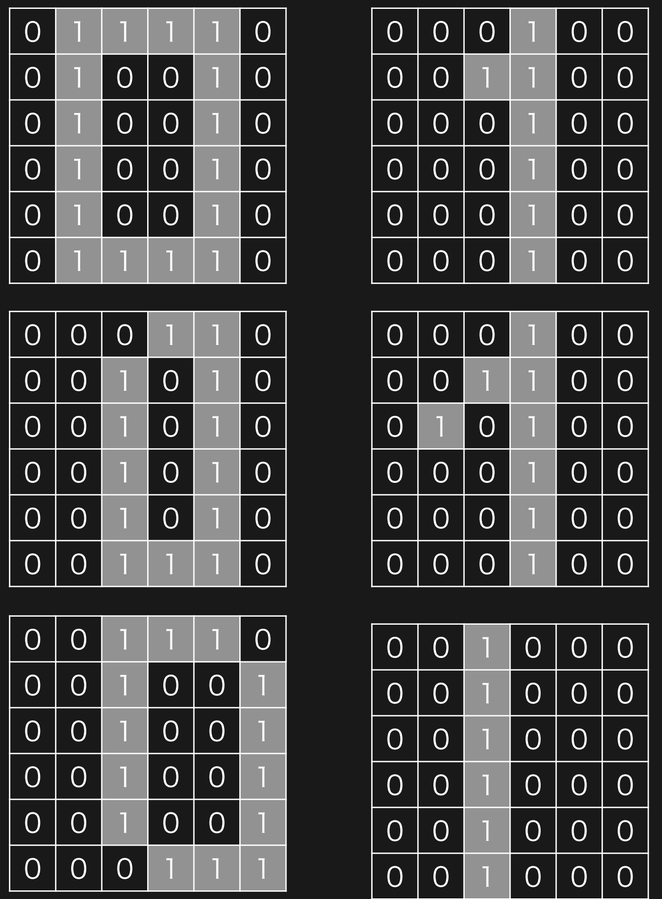


1の画像

機械学習を使用する場合

- data_0 = [[0,1,1,1,1,0], [0,1,0,0,1,0], […], […], […], […]]
- data_1 = [[0,0,0,1,0,0], [0,0,1,1,0,0], […], […], […], […]]

- dataを集めて, そのデータをモデルに渡して学習を実行する
  - data = [data_0_1, …, data_0_100, data_1_1, …, data_1_100]
    - # 0,1のデータを100個ずつ集める
  - label = [0, 0, …, 0, 1, 1, …, 1]
    - # dataに対して出力して欲しい値(答え)を作る
  - model.fit(data, label)
    - # データを読み込ませて機械学習モデルで学習



0の画像　　1の画像

機械学習の強み

1. 人が言葉でうまく説明できないパターンも自分でルールを見つけられる
   - 犬と猫の画像を分類分けするタスクも実行可能
2. データから勝手にパターンを覚えてくれる
   - 自分でルールを考えなくて良い
3. パターンが変わってもデータを変えてモデルを再作成すれば最新化出来る
   - データから学ぶため, データを変えれば最新化される

機械学習の弱み

1. 入力に対する出力を得る過程がブラックボックス化している
    - 何を元に出力を決定しているか分かりにくい → 倫理面の問題とか
2. 100%を作りにくい
    - 明示的にルールを記述しないので, ある条件下の時に必ず1と出力させるのは難しい
    - ルールが分かっているなら, If文で記述するべき
3. データがないと無力
    - データから学ぶため, データがないと無力

機械学習の種類

機械学習

教師あり学習

分類問題　　回帰問題

説明変数 + 目的変数の組み合わせを使って
学習するモデル

教師なし学習

クラスタリング　　次元削減

自動生成

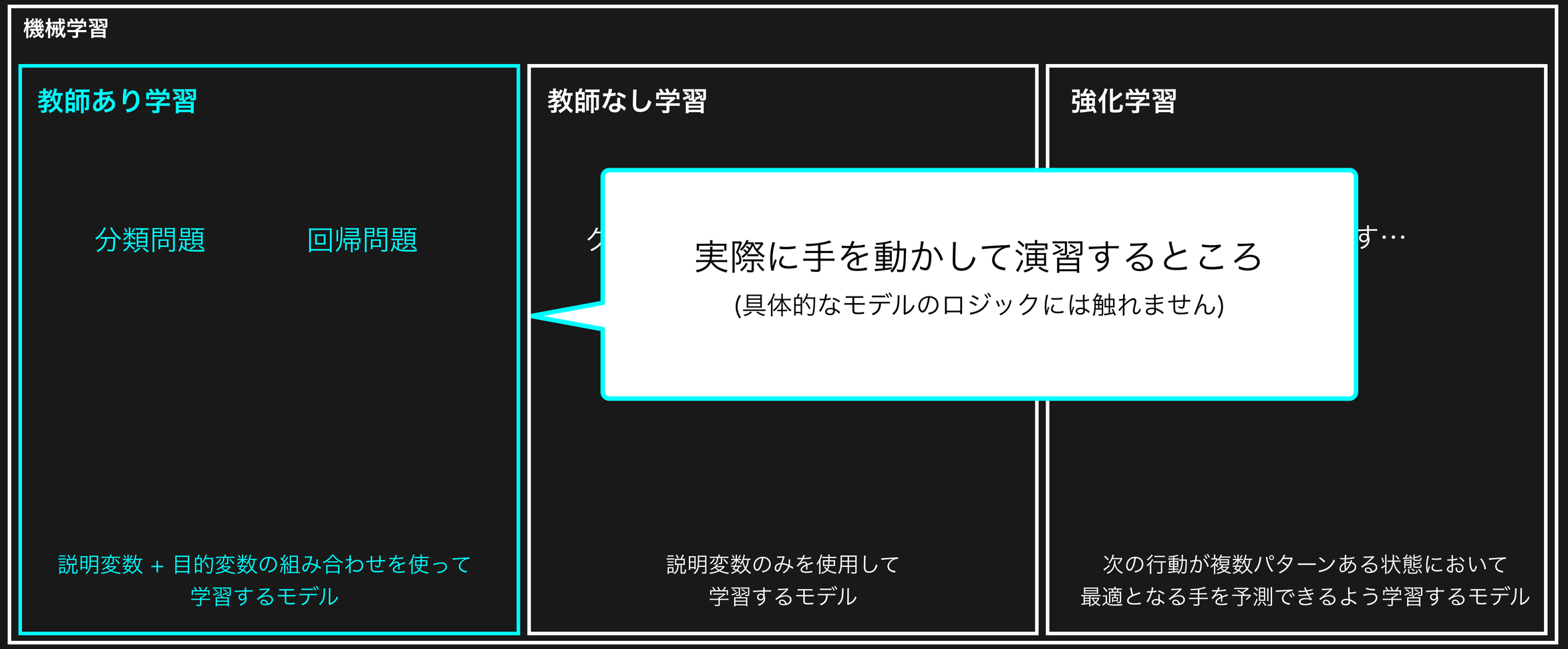
説明変数のみを使用して
学習するモデル

強化学習

詳しくないです…

次の行動が複数パターンある状態において
最適となる手を予測できるよう学習するモデル

機械学習の種類


機械学習
教師あり学習
分類問題
回帰問題
説明変数 + 目的変数の組み合わせを使って
学習するモデル
教師なし学習
説明変数のみを使用して
学習するモデル
強化学習
次の行動が複数パターンある状態において
最適となる手を予測できるよう学習するモデル
実際に手を動かして演習するところ
(具体的なモデルのロジックには触れません)

教師あり学習

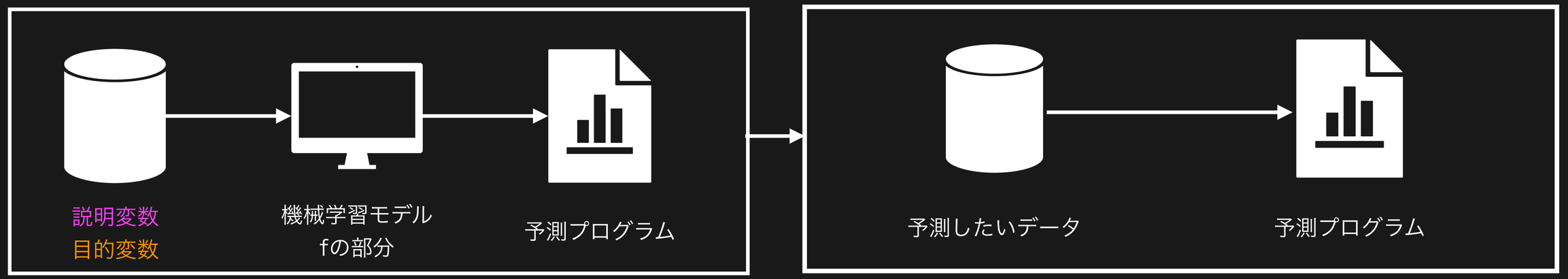
- 説明変数と目的変数の組み合わせを使用して学習するモデル
  - 説明変数(入力)から，目的変数(答え)を予測する
  - 画像の例も教師あり学習　「ピクセル値が説明変数」　「0,1が目的変数」
  - y = f(x)のイメージ
    - 目的変数 = f(説明変数)

教師あり学習の例

- **例) 住宅価格の予測**

説明変数　　目的変数

- やりたいこと：立地や広さなどのデータから住宅価格を予測したい
  - 住宅価格の予測値 = f(立地などのデータ)



教師あり学習の例

目的変数を予測するために使用する値

予測したい値

例) 住宅価格の予測

説明変数

目的変数

	ID	種類	地域	市区町村コード	都道府県名	市区町村名	地区名	最寄駅：名称	最寄駅：距離（分）	間取り	面積（㎡）	土地の形状	間口	延床面積（㎡）	建築年	建物の構造	用途	今後の利用目的	前面道路：方位	前面道路：種類	前面道路：幅員（m）	都市計画	建ぺい率（%）	容積率（%）	取引時点	改装	取引の事情等	取引価格（総額）_log
0	1060685	中古マンション等	NaN	1108	北海道	札幌市厚別区	大谷地東	大谷地	8	3LDK	80	NaN	NaN	NaN	平成7年	SRC	住宅	NaN	NaN	NaN	NaN	準工業地域	60.0	200.0	2009年第4四半期	未改装	NaN	7.079181
1	1005580	中古マンション等	NaN	1101	北海道	札幌市中央区	南9条西	中島公園	5	1DK	30	NaN	NaN	NaN	昭和57年	SRC	NaN	住宅	NaN	NaN	NaN	近隣商業地域	80.0	300.0	2018年第3四半期	未改装	NaN	6.755875

教師あり学習の種類

- 大きく分けると**分類**と**回帰**に分けられる
  - 分類：クラス分けを予測する
    - 例) 0,1の画像の予測 (0クラス, 1クラス)
  - 回帰：連続値を予測する
    - 例) 住宅価格の予測 (7.56とか6.98とか)

教師あり学習の代表的なモデル

分類

Random forest (classifier)

ロジスティック回帰

SVM(SVC)

MLP

回帰

Random forest (regressor)

単回帰分析

SVM(SVR)

MLP

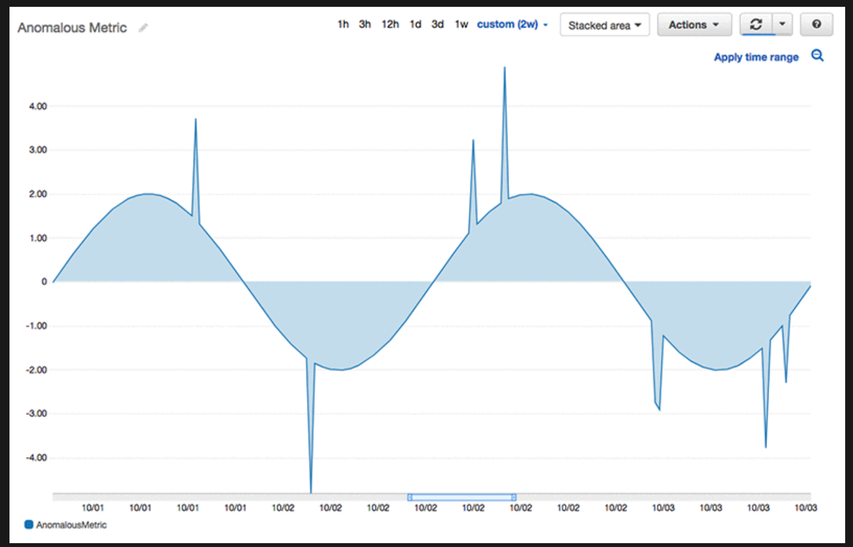
教師なし学習

- 目的関数を必要とせずに学習できるモデル
  - 説明変数のみで学習する
  - 出力の様式は様々(答えがなくともできそうなタスク)
    - グループ分けしてくれたり, 学習時のデータとそっくりなデータを生成してくれたり
  - 統計の範囲と被っているように感じる

教師なし学習の例

- 例)異常値の除去
- やりたいこと：入力値の中から異常値を見つけたい
  - 説明変数のみから学習して, 異常値か推定する
    - (目的変数を必要としない)

このデータだけでも異常値
は見つけられそうじゃない？

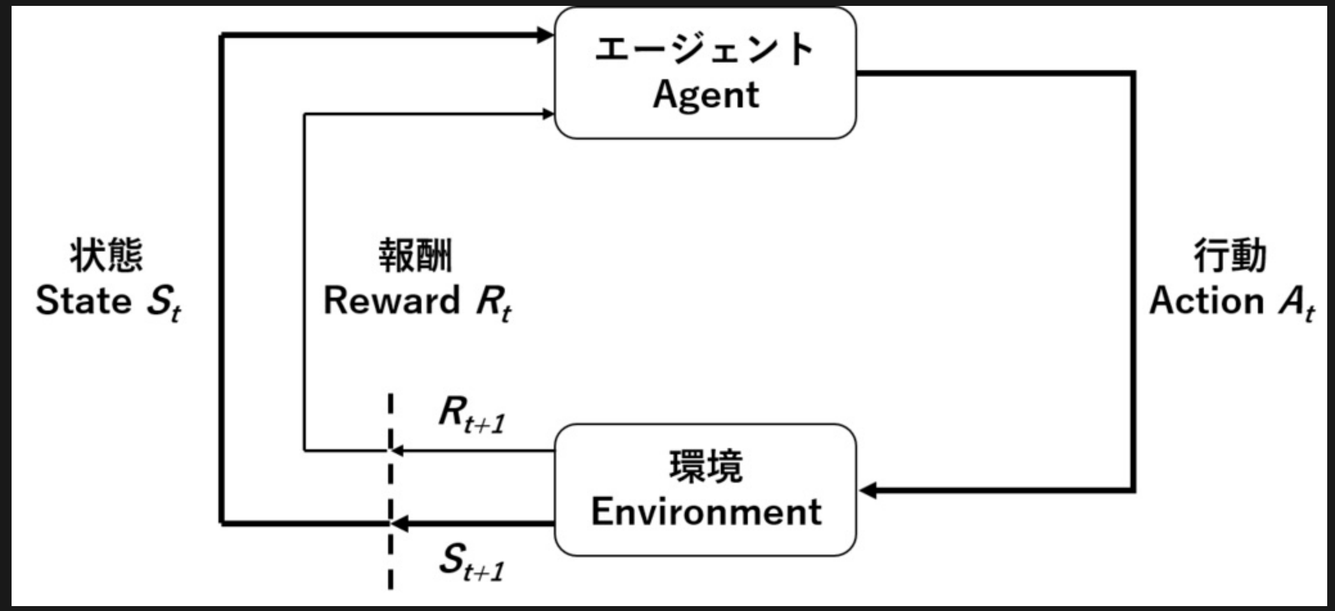


教師なし学習の代表的なモデル

- 異常値検出：Random cut forestなど
- クラスタリング：k-means, dbscanなど
- 次元削減：PCA, umap, t-sneなど
- 生成モデル：VAE, GANなど

強化学習

- あまり詳しくないのと, 専門用語多めなので説明割愛
  - 概念だけなら難しくない
  - alfa goとか, 将棋のAIとか



機械学習の種類

機械学習

教師あり学習

分類問題　　回帰問題

説明変数 + 目的変数の組み合わせを使って
学習するモデル

教師なし学習

クラスタリング　　次元削減

自動生成

説明変数のみを使用して
学習するモデル

強化学習

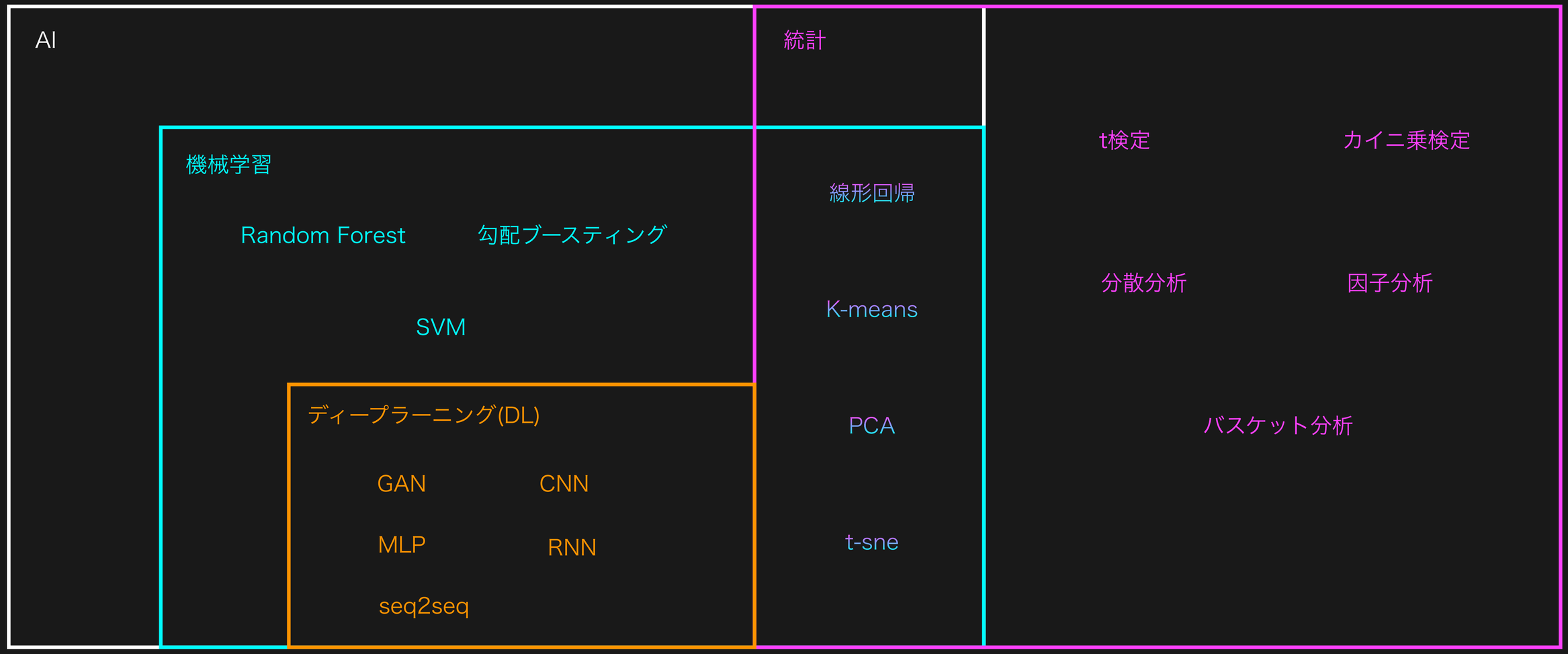
詳しくないです…

次の行動が複数パターンある状態において
最適となる手を予測できるよう学習するモデル

分野(AI・機械学習・DL)


AI
機械学習
Random Forest
勾配ブースティング
線形回帰
t-sne
dbscan
SVM
K-means
PCA
knn
ディープラーニング(DL)
GAN
CNN
seq2seq
VAE
MLP
RNN

分野(統計・AI・機械学習・DL)



Jupyter入門

Jupyter入門

- Jupyter入門資料を使う

Python入門

Python入門

- Jupyter上で動かす
  - AWS Sagemakerのノートブックインスタンスにアクセスしてもらう
- Pythonは人気なので, いくらでも勉強用の資料やサイトがある
  - https://sites.google.com/view/ut-python/resource/%E6%95%99%E6%9D%90%E8%AC%9B%E7%BE%A9%E5%8B%95%E7%94%BB

データ分析コンテストの紹介

データ分析コンテストの紹介

- 部会発表資料を使う

データ分析コンテストに挑戦

データ分析コンテストに挑戦

- Jupyter上で手を動かす
  - AWS Sagemakerのノートブックインスタンスにアクセスしてもらう

発表者用：!aws sagemaker create-presigned-notebook-instance-url --notebook-instance-name "data-analytics-training-2021-" --session-expiration-duration-in-seconds 28800