システム入門（2）

データの種類と表現、特徴量

尾崎　遼

この講義の目標

様々な種類のデータが**コンピュータの中でどのように表現され、AIに入力されるか**について説明する。

個々のデータは複数の特徴量から成り、その特徴量をAIが何らかの形で解釈することで、推論・判別が行われる

特徴量の概念そのものと、特徴量がAIによってどのように扱われるかについて解説する

データ、タスク、モデル、特徴量

生データ ⇄ 特徴量 ⇄ モデル ⇄ タスク

データ前処理
データ整形

	p	alcgp	tobgp	ncases	ncontrols
	4	0-39g/day	0-9g/day	0	40
	4	0-39g/day	10-19	0	10
	4	0-39g/day	20-29	0	6
	4	0-39g/day	30+	0	5
5	25-34	40-79	0-9g/day	0	27
6	25-34	40-79	10-19	0	7
7	25-34	40-79	20-29	0	4
8	25-34	40-79	30+	0	7
9	25-34	80-119	0-9g/day	0	2
10	25-34	80-119	10-19	0	1
11	25-34	80-119	30+	0	2
12	25-34	120+	0-9g/day	0	1
13	25-34	120+	10-19	1	1

- 現実を計測・記録したもの
- いろんな形式がある
- ノイズや欠測もある
- 結合されている

- 生データを**数値として**表現したもの
- 少なすぎると表現力低下
- 多すぎるとモデルが複雑化

- データ同士の関係性を数式として定式化したもの
- 数値データが入力
- 評価が必要

- 答えたい問い
- やりたいこと

生データは特徴量（数値）に変換され、 AI（数理モデル）に入力される

スカラー

1

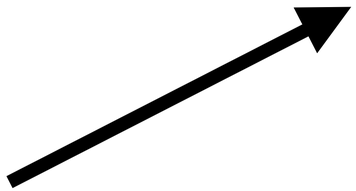
45

212948502348 39

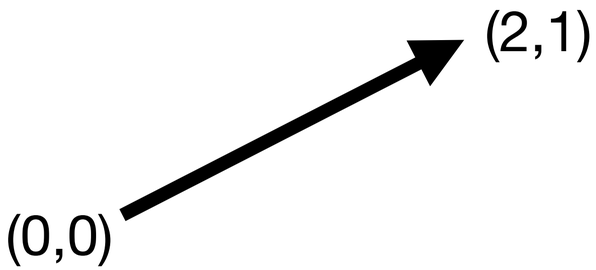
スカラーは単なる数

ベクトルって何（１）

日常用語における
ベクトルのイメージ
（２次元）



幾何学における
ベクトルのイメージ
（２次元）



線形代数における
ベクトルのイメージ
（２次元）

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

線形代数における
ベクトルのイメージ
（４次元）

$$\begin{pmatrix} 2 \\ 1 \\ -3 \\ 5 \end{pmatrix}$$

三井生命/ベクトルくん　田中秀幸

ベクトルって何（2）

線形代数における
ベクトルのイメージ
（２次元）

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

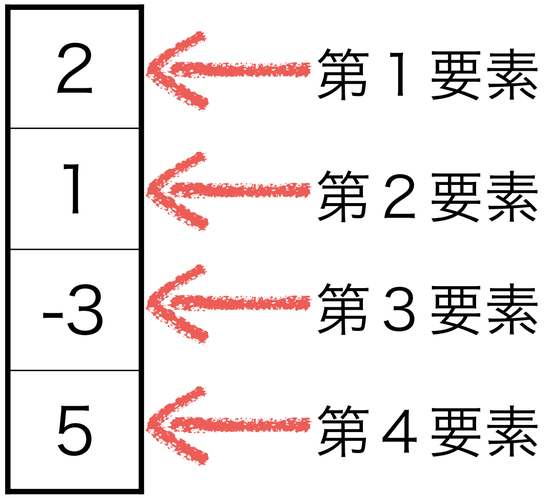
線形代数における
ベクトルのイメージ
（４次元）

$$\begin{pmatrix} 2 \\ 1 \\ -3 \\ 5 \end{pmatrix}$$

４次元のベクトル

2
1
-3
5

ベクトルの要素



d次元ベクトル　＝　d個の数字が並んでいるもの

各要素は１個の添字で指定できる

行列って何

行列
（4x3行列）

2	84	-10
1	61	-1
5	35	4
91	3	45

行

第１行	2	84	-10
第２行	1	61	-1
第３行	5	35	4
第４行	91	3	45

列

第１列	第２列	第３列
2	84	-10
1	61	-1
5	35	4
91	3	45

第（2,3）要素

	第３列
2	-10
第２行 1	-1
5	4
91	45

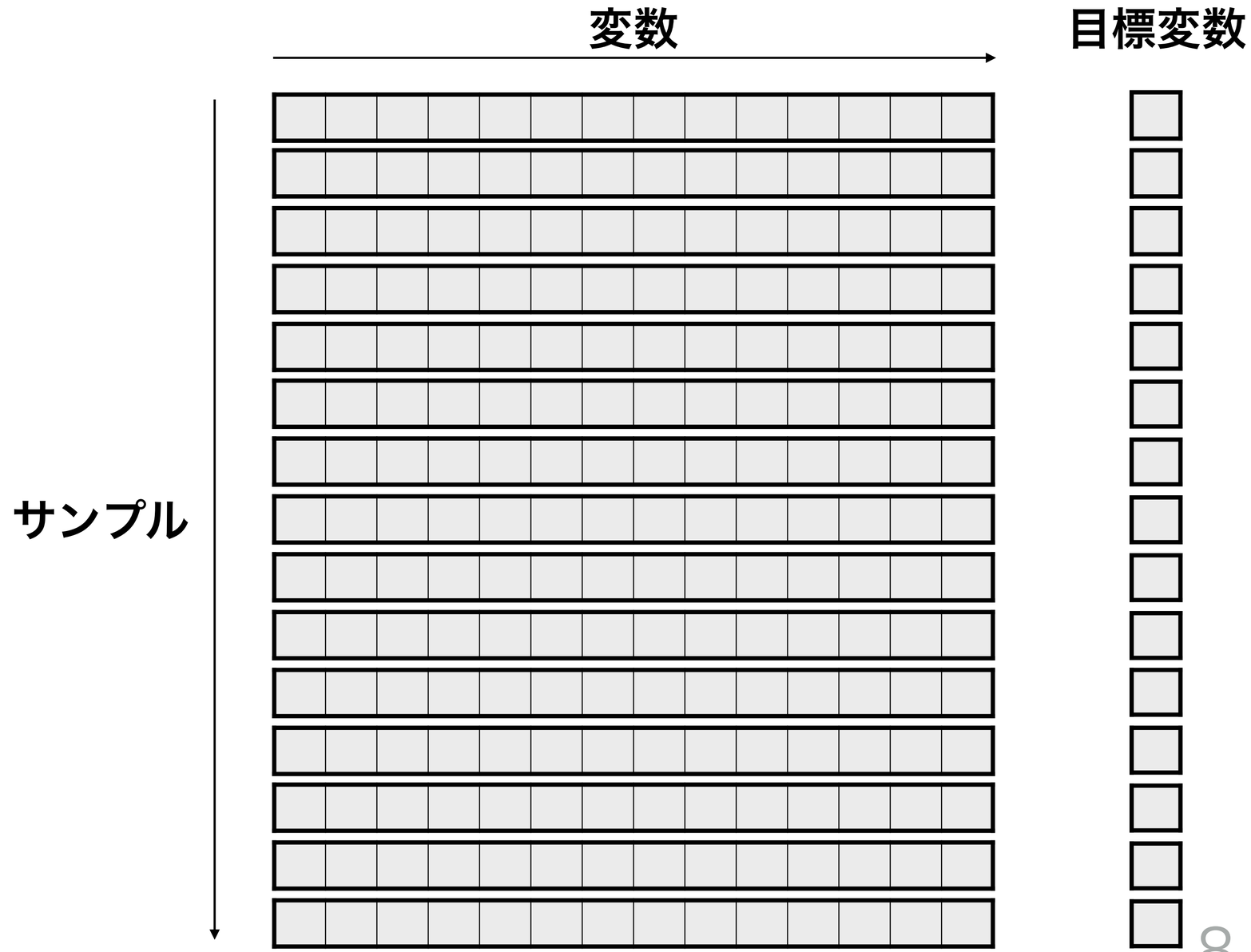
行と列から成る。各要素は２個の添字で指定できる。

ベクトルとしても表現可能

データ＝サンプル分の特徴量ベクトル

一つのデータセットは、特徴量と目標変数のペアが集まったもの（教師あり学習の場合）

特徴量（変数）は通常ベクトル



小まとめ：データの大きさと入れ物

スカラー

数値

ベクトル

数値がいくつか並んだもの。１個の添字で要素を指定できる。

行列

数値が縦横に並んだもの。２個の添字で要素を指定できる。ベクトル表現可能。

データ＝サンプル分の特徴量ベクトル

表形式のデータ

表計算ソフト（Excelなど）をイメージすればよい

プログラミングでよくつかうファイル形式

CSV、TSV

表形式のデータ

変数

サンプル

Name	Year of Birth	Year of Death
Kiyoshi Shiga	1871	1957
Shibasaburo Kitasato	1853	1957
Ogai Mori	1862	1922

Comma-separated values (CSV)

1 行の中で異なる列の要素をコンマでつなぐ

ヘッダー行（列名を記述した列）がある場合とない場合がある

CSV
（ヘッダー無し）

```
Kiyoshi Shiga,1871,1957
Shibasaburo Kitasato,1853,1931
Ogai Mori,1862,1922
```

CSV
（ヘッダーあり）

```
Name,Year of Birth,Year of Death
Kiyoshi Shiga,1871,1957
Shibasaburo Kitasato,1853,1931
Ogai Mori,1862,1922
```

Tab-separated values (TSV)

1 行の中で異なる列の要素をタブ（\t）でつなぐ

ヘッダー行（列名を記述した列）がある場合とない場合がある

TSV
（ヘッダー無し）

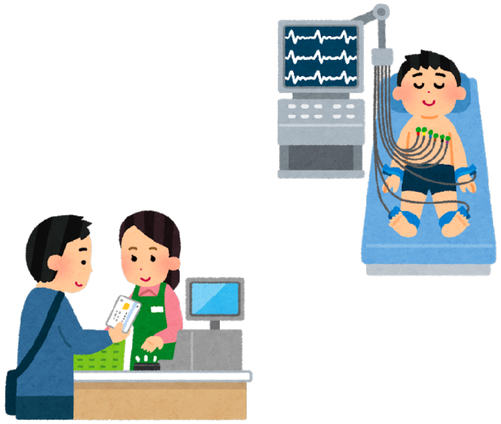
```
Kiyoshi Shiga	1871	1957
Shibasaburo Kitasato	1853	1931
Ogai Mori	1862	1922
```

TSV
（ヘッダーあり）

```
Name	Year of Birth	Year of Death
Kiyoshi Shiga	1871	1957
Shibasaburo Kitasato	1853	1931
Ogai Mori	1862	1922
```

特徴量ベクトルの作成

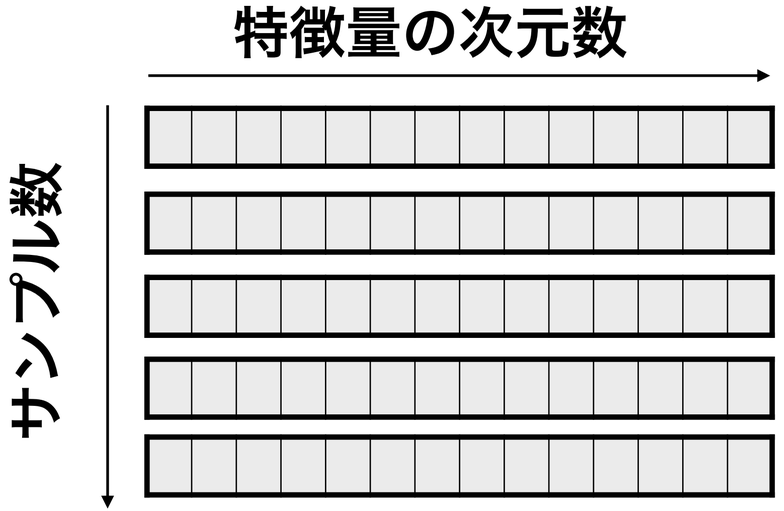
生データ



カウント、画像、テキスト etc.
スケールもバラバラ

どうやって変換すればいいのか？

特徴量ベクトル



AI（モデル）に入力する**特徴量ベクトルを作る**ステップが実は技術的に奥深いところ

数値データの変換、スケーリング

数値データの値に意味があるか

どんなスケール（値の範囲）か？

どんな分布か？

説明変数と目標変数はどのような関係か？

対数変換、べき変換

スケーリング

Min-Maxスケーリング：元の数値を最大値と最小値の間の範囲になるように変換

標準化：元の数値を平均０、分散１となるように変換

数値データの二値化・離散化

二値化

ある閾値の前後で0と1に変換（例：音楽の再生回数）

離散化

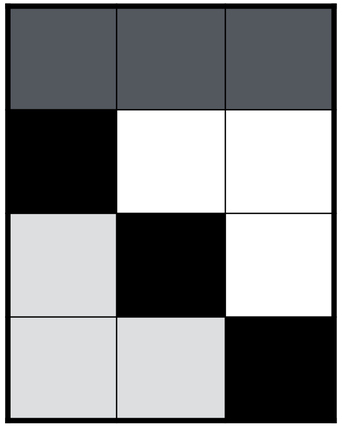
固定幅による離散化（例：年齢区分）

分位数による離散化（例：食べログのスコア）

数値のどのような特徴に意味があると仮定するかに依存

画像データの数値表現

白黒画像
（行列）



数値表現
（行列）

170	170	170
255	0	0
80	255	0
80	80	255

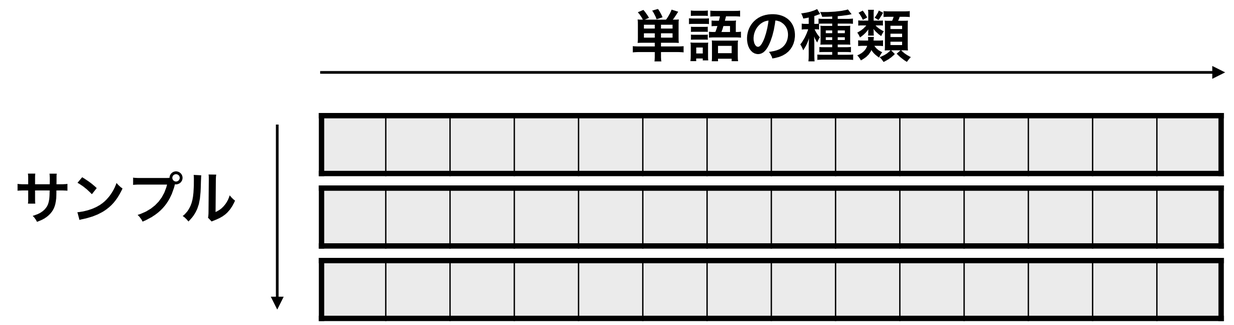
画像のベクトル表現

画像はどのピクセルの色がどうだったかを表すベクトル

テキストデータの数値表現

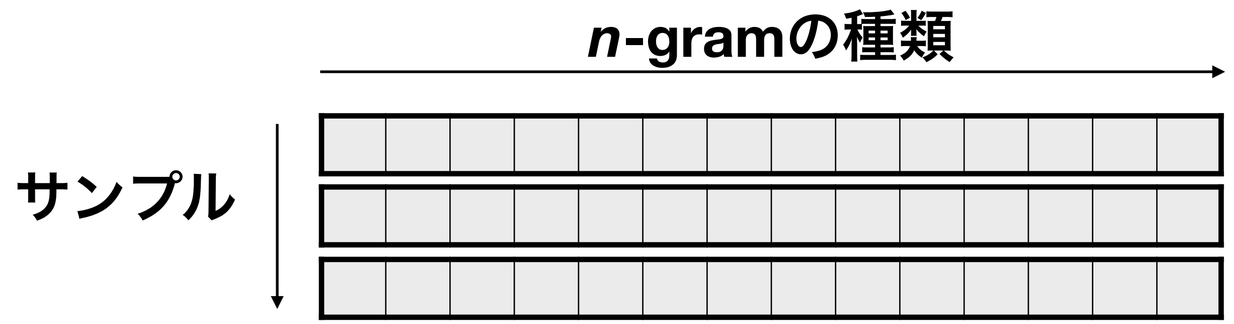
Bag-of-words

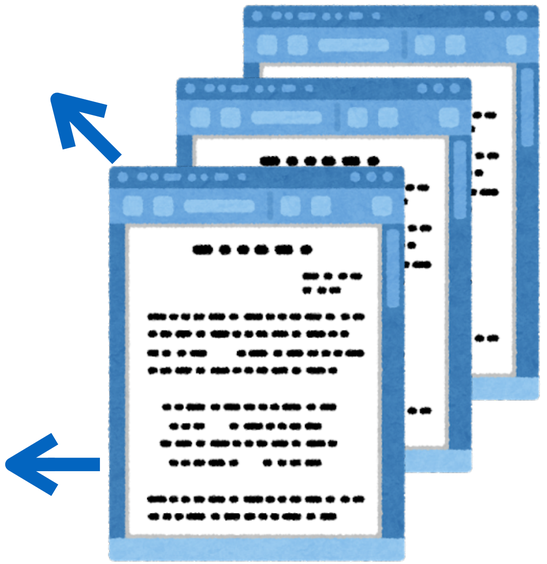
・単語をカウント



Bag-of-*n*-gram

・n個の単語の組を
カウント





テキストは単語などが何個あったかを表すベクトル

カテゴリ変数の数値表現

One-hot encoding

・分かりやすい
・冗長な表現

Dummy coding

・余分な次元が
取り除かれている

カテゴリ変数はどのカテゴリかを表すベクトル

カテゴリ変数の例（10種）

区
千代田区
港区
世田谷区
中央区
練馬区
杉並区
板橋区
足立区
葛飾区
北区

小まとめ：特徴量ベクトルの作り方

離散化・二値化、スケーリング

画像

どのピクセルの色がどうだったかを表すベクトル

テキスト

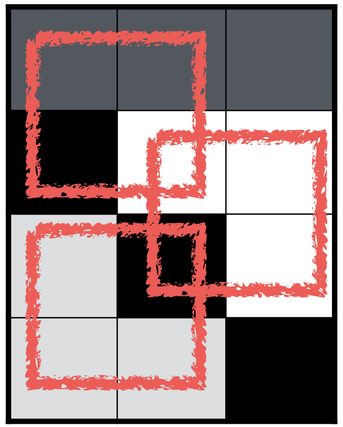
単語などが何個あったかを表すベクトル

カテゴリ変数

どのカテゴリかを表すベクトル

特徴量の自動作成

手動の特徴量生成
（SIFT、HOG）

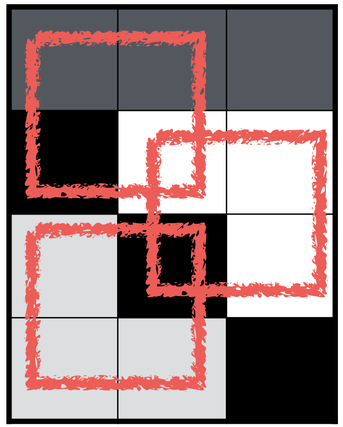


固定のルール・前処理
（画像勾配、エッジ検出、
スムージング、正規化）

特徴量ベクトル

170	1	80	0	170	255	1	33	170

自動の特徴量生成
（深層学習（CNN）など）



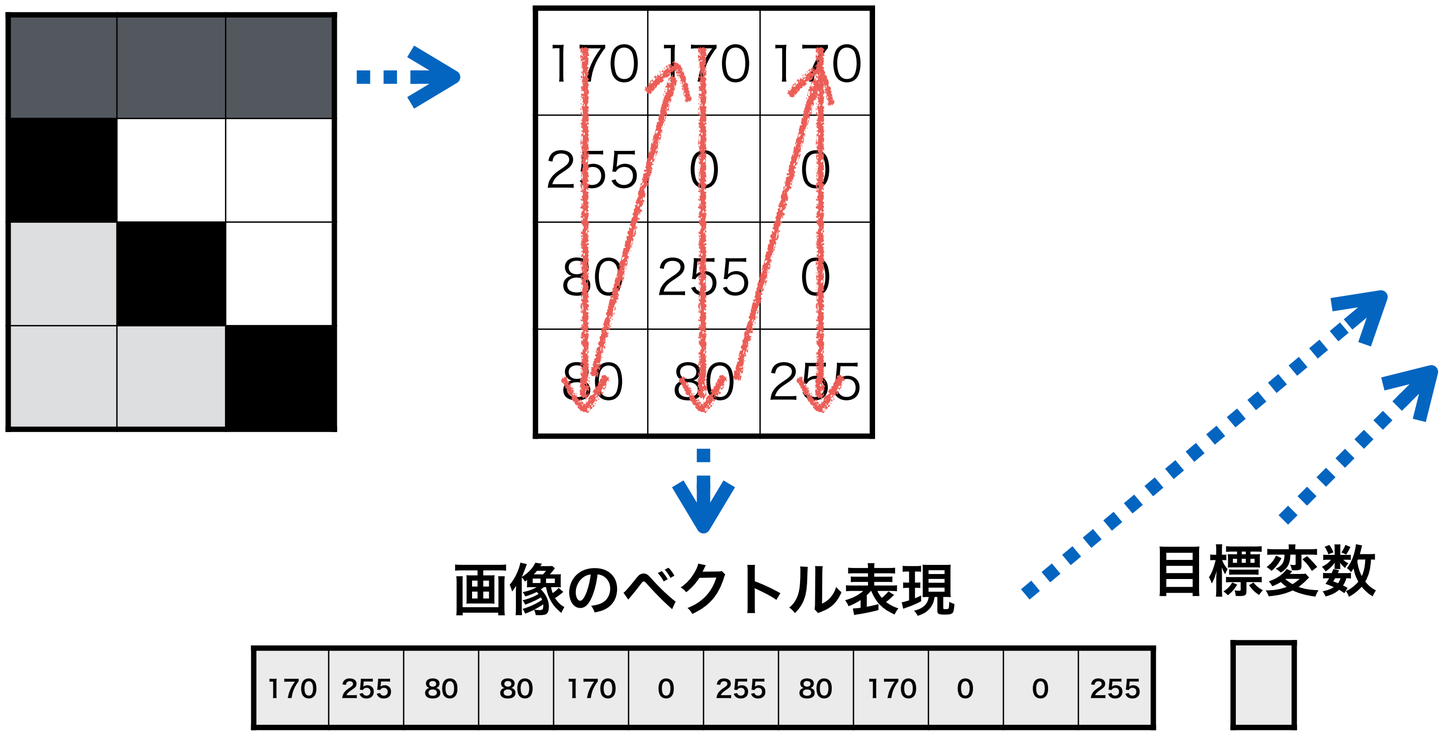
CNN
（畳み込み層、
プーリング層）

特徴量ベクトル？

170	1	80	0	170	255	1	33	170

特徴量をどう抽出するかはドメイン知識に依存

特徴量の自動作成（深層学習（CNN））



深層学習は特徴量作成のステップ（ドメイン知識）をカプセル化し、学習データからパラメータを自動的に学習

まとめ

AIへの入力は特徴量ベクトルである

- 計測と生データを特徴量ベクトルに変換するプロセスがある

データの種類によって特徴量ベクトルへの変換方法は様々

人間がタスクを達成するために、データの間の関係をどう仮定してモデリングするかが重要

- 特徴量を手動で作るにせよ自動で作るにせよ