ディープラーニングの技術は何が凄いのか

を理解するために、まずは

特徴量

という概念を正しく知りましょう

「特徴量」(英：feature) とは
分析対象を表現する
予測の手掛りとなる変数

列の項目に相当

年齢	性別	年収	家族人数	勤務年数	…
45	1	700	2	2	…
32	0	450	4	3	…
:	:	:	:	:	:

【例①】物件条件から家賃の予測

特徴量の例①

面積	最寄駅(分)	築年数	管理会社Tel	正解
45	10	20	xxx	15
32	7	45	xxx	10

予測
14
9

面積,最寄駅,築年数
は特徴量として
効きそう

管理会社Tel
は効かない

【例②】個人データから保険契約するかを予測

年収	既婚	身長	生年月日
450	0	176	YYMMDD
600	1	155	YYMMDD

正解
0
1

予測
0
1

年収,既婚は効きそう

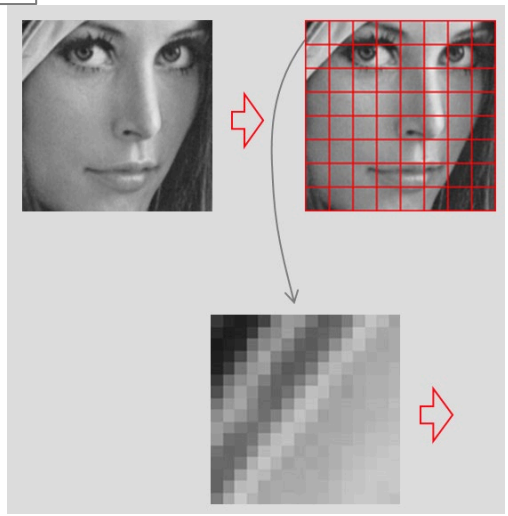
身長は効かない

生年月日は現日付との差をとって**年齢**に変換すれば効きそう

（参考）非構造化データの場合の特徴量

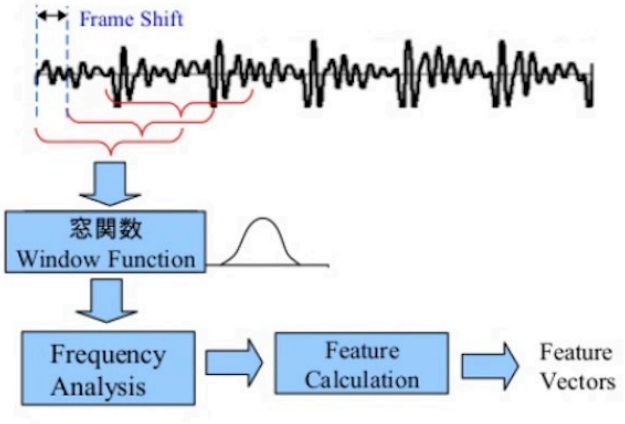
画像認識

特徴量 ＝ 画像のピクセル



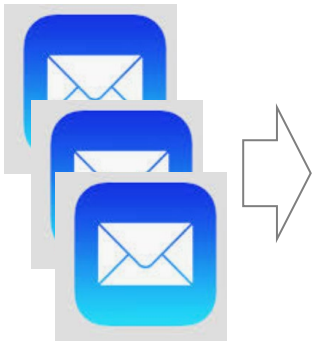
音声認識

特徴量 ＝ 音波の波形を処理した結果



自然言語処理

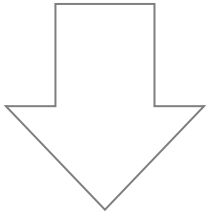
特徴量 ＝ 注目文言の出現頻度



	単語１	単語２	単語３	…	単語M
文章１	4	8	0	…	2
文章２	2	0	1	…	6
文章３	7	0	8	…	4
文章４	3	4	3	…	1
⋮	⋮	⋮	⋮	…	⋮
文章N	0	2	5	…	6

生データのままでは、必ずしも理想的な特徴量を得られない

予測に影響を及ぼす因子を過不足なく含むデータを作りたい



特徴量設計（特徴量エンジニアリング）のプロセスが重要

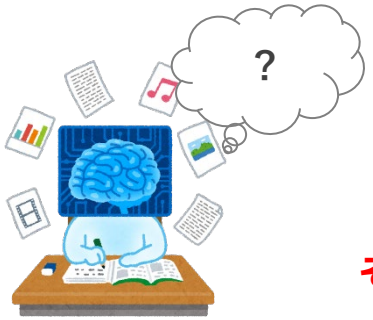
- ✓予測変数として採用する列を選別
- ✓元データに**前処理**を施す

従来の機械学習では特徴量設計（前処理）が大変

カテゴリカルデータの処理

コンピュータは数値しか処理できない為、
文字列データを数値に変換してから機械学習モデルに入力

初回/リピート
初回
3回目
5回以上
…



「初回」なら1
それ以外は0に変換

初回フラグ
1
0
0
…

欠損値処理

欠損値(歯抜け)の多い場合は良い精度を期待できない ➡ 適切な値で補充

性別	年齢	身長	体重
男	35	170	
男		165	60
		159	
男	12	155	40
男		165	62
女		145	35

特徴量の変換・追加

予測に効きそうな情報をより効果的な形に編集し、新しい価値を持たせる

- 集計、カウント、データを結合・分割など

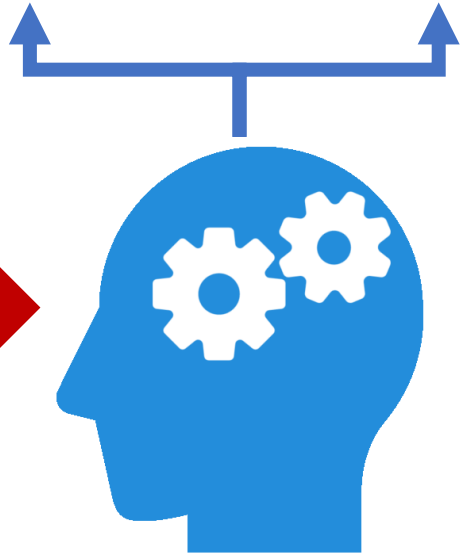
会員ID	年月日	購入個数
1	2019/10/11	2
2	2019/10/12	1
1	2019/10/12	1
3	2019/10/12	2
2	2019/10/13	1

集計

会員ID	総数
1	3
2	2
3	2

ディープラーニングは、
特徴量をデータから自動的に抽出できる

顔・身体の
特徴を悩む



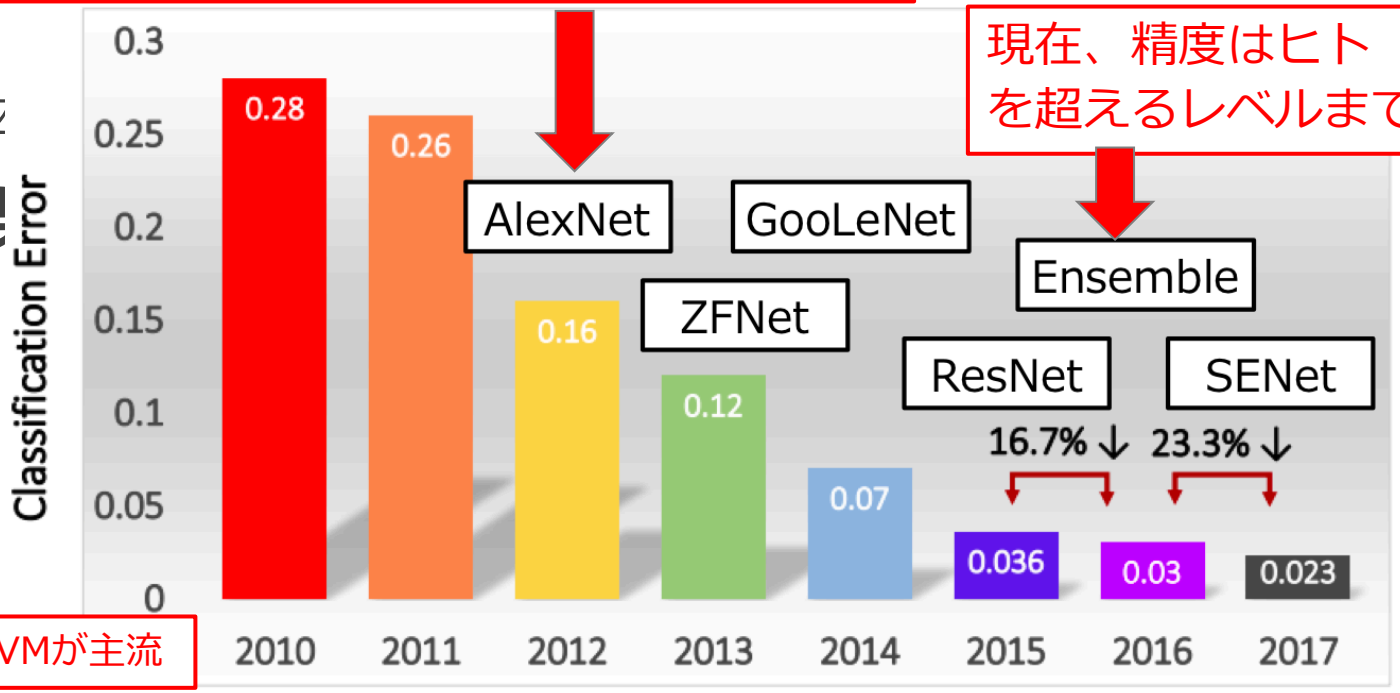
大量なデータから
自動的に特徴を抽出

画像認識技術の進歩

- ディープラーニングが大きく注目されるきっかけは、2012年に開催された国際コンペ **ILSVRC（ImageNet Large Scale Visual Recognition Challenge）**
- 大[illegible]を学習し、テストデ[illegible]
- 具体[illegible]
  「[illegible]

ディープラーニングを利用したモデルが
前年までの誤差率を10%以上改善

現在、精度はヒト（約5%）
を超えるレベルまで到達



従来はSVMが主流

http://image-net.org/challenges/talks_2017/ILSVRC2017_overview.pdf