

統計数理 1

情報理工学研究科
数理・計算科学専攻
鈴木 大慈

s-taiji@is.titech.ac.jp
西八号館707号室

本講義の目的

- ・ **機械学習・データ解析**の手法を紹介
- ・ 各手法の理論と使い方を学ぶ
- ・ 実際にパターン認識などを解けるようにする

前提知識:

線形代数, 確率統計の基礎知識

機械学習とは

機械学賢くなるための方法論

Wikipedia 「人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとする技術・手法のことである。」

Arthur Samuel 「Field of study that gives Computers the ability to **learn without being explicitly programmed** 」 (1959)

コンピュータ将棋

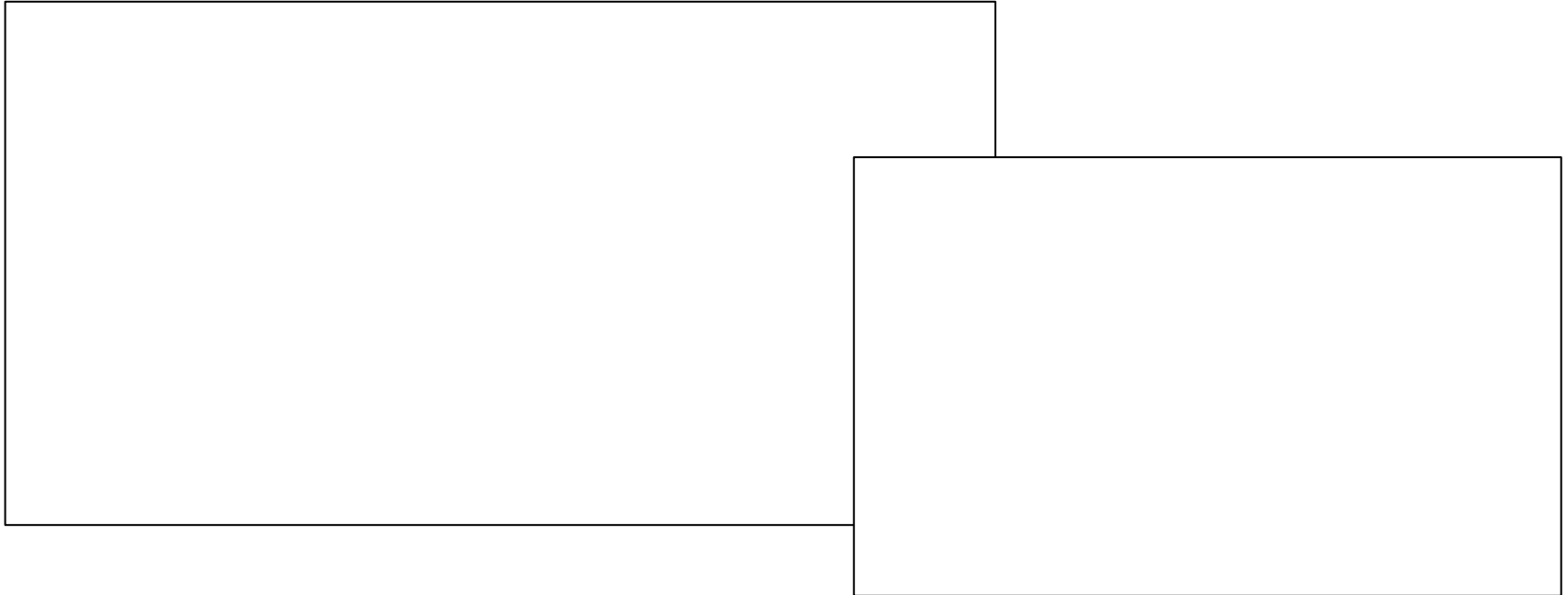


**Bonanza: 2005年6月 機械学習の利用
(棋譜データからの評価関数の学習)**

第2回将棋電王戦: 2013年3月～4月 3勝1敗1分

IBM Watson

クイズ番組「ジェパディ!」 2011年2月14-16日



機械学習・自然言語処理

- ・ 2880個のプロセッサ・コア
- ・ 2億ページ分のテキストデータ
(約100万冊の書籍に相当)

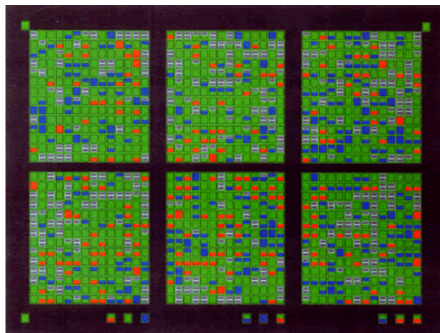
Q: 米国が外交関係を持たない世界の四力国のうち、最も北にある国

その他応用

検索エンジン



遺伝子データ解析



推薦システム

amazon.co.jp

マイストア | Amazonポイント | ギフト券 | タイムセール | 出品サービス | ヘルプ

カテゴリーからさがす

マイストア マイページ お客様へのおすすめ おすすめ商品を正確にする プロフィール 詳しくはこちら

マイストアのおすすめ商品
もしあなたが[ユーザー名]さんではない場合、サインインしてください

本日のみ
[おすすめ商品を見る](#)

おすすめ商品
[Amazonインスタント・ビデオ](#)
[Amazonファッション](#)
[Androidアプリストア](#)
[DIY・工具](#)
[DVD](#)
[Kindleストア](#)
[MP3ストア](#)
[PCソフト](#)
[TVゲーム](#)
[おもちゃ](#)
[カー用品](#)
[ジュエリー](#)
[スポーツ&アウトドア](#)
[ビデオ](#)
[ベビー&マタニティ](#)
[ペットフード・ペット用品](#)
[ホーム&キッチン](#)
[ミュージック](#)
[大型家電](#)
[家電・カメラ・AV機器](#)
[文房具・オフィス用品](#)

これらのおすすめ商品は、[あなたのお持ちの商品](#)などに基づいています。

表示: [すべて](#) | [ニューリリース情報](#) | [まもなく発売](#)

16. [トワイニング クオリティアールグレイ 100g](#)
トワイニング
おすすめ度: ★★★★★ (12)
在庫あり
価格: ￥864
新品の出品: ￥6498より

☐ 持っています ☐ 興味がありません ☒ ★★★★★ この商品を購入する
モンティース デイメジャー ゴールドなどを購入されたお客様におすすめ (おすすめの商品に反映させる商品)

17. [トワイニング クオリティレディグレイ 100g](#)
トワイニング
おすすめ度: ★★★★★ (21)
在庫あり
価格: ￥864
新品の出品: ￥6498より

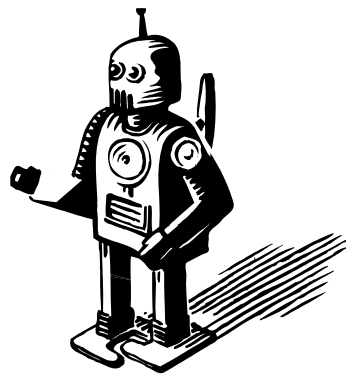
☐ 持っています ☐ 興味がありません ☒ ★★★★★ この商品を購入する
モンティース デイメジャー ゴールドなどを購入されたお客様におすすめ (おすすめの商品に反映させる商品)

18. [マッシュ デイメジャー](#)
高桑金属 (2011/8/22)
おすすめ度: ★★★★★ (3)
在庫あり

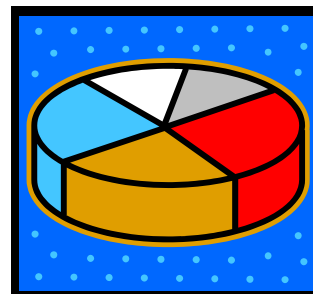
機械学習と統計学

機械学習と統計学には深い関係がある。

機械



統計



データからの情報抽出
乱雑なデータからその裏にある規則を見つける

機械学習の歴史

1946: ENIAC, 高い計算能力

フォン・ノイマン「俺の次に頭の良い奴ができた」

1952: A. Samuelによるチェッカーズプログラム

1957: Perceptron, ニューラルネットワークの先駆け

1960年代前半: ELIZA, 擬似心理療法士

1980年代: 人手による学習ルールの作りこみの限界
「膨大な数の例外」

1990年代: 計算機能力の向上, データ量の増大

→ データから自動的に規則性を見つける.

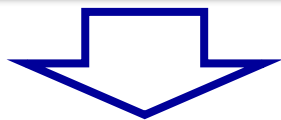
統計的学習への移行

機械学習の考え方

将棋の強いプログラムを作ろう！

強い手を指すようにプログラムする？

→ プログラマーが将棋に強くないといけない
そもそも盤面のパターンは 10^{220} 通り以上
無理...



具体的にどうやって指すかはプログラムしない！
プログラムするのは**学習の仕方**.
機械にデータから学習させる.

統計学の出番

統計と機械学習の対応

機械学習

統計

学習

推定

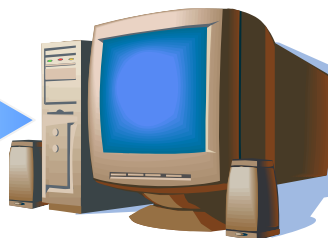
学習の仕方

統計モデリング
＋推定手法

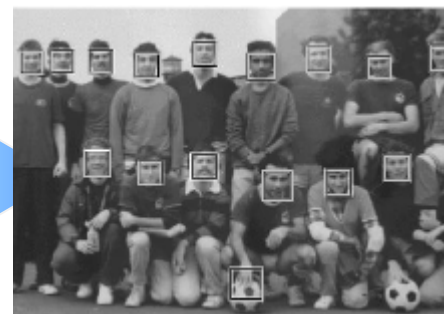
大量データ



学習(推定)



判別



[Viola&Jones 2001]

一方, 統計学では. . .

Harold Jeffreys: *Theory of Probability* (1960).

The fundamental problem of scientific progress, and a fundamental one of everyday life, is that of **learning from experience**. Knowledge obtained in this way is partly merely description of what we have already observed, but part consists of making **inferences from past experience to predict future experience**. This part may be called generalization or induction.

統計と機械学習の相違点

機械学習:

機械が賢くなれば良い.

パフォーマンス(正解率や予測精度)重視.

統計学:

人間が有用な知見を得ることが大事.

信頼区間や検定を用いた統計的推論(inference).

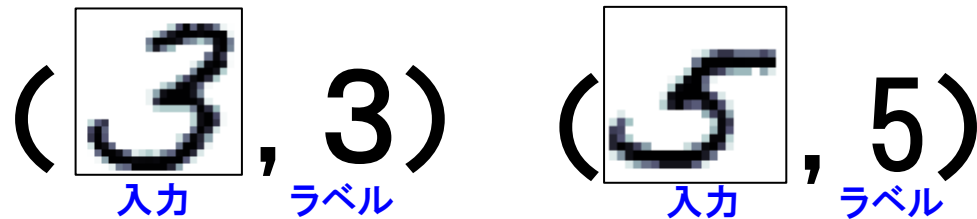
解釈可能性に関する態度には大きな違いがある.

機械学習の問題設定

教師有りと教師なし

教師有り学習:

データ(x,y) ある入力xとそれに対するラベルyの組
問題の例: 回帰, 判別



教師なし学習:

データ(x) ラベルがない

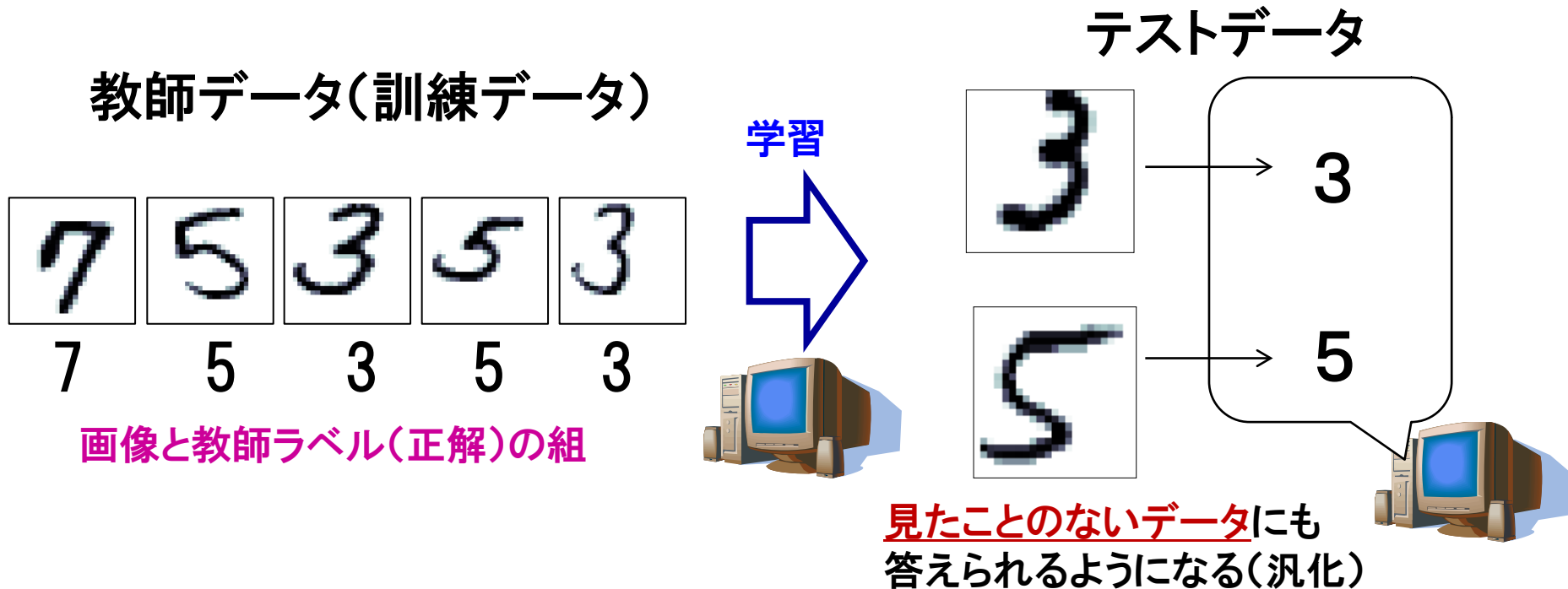
問題の例: クラスタリング, 音源分離, 異常検知



半教師有り学習:

ラベルの付いているデータと付いてないデータが混在

教師有り学習



1. スпамフィルター

メール文面 → スпам or まともなメール

2. 遺伝子データによる疾病診断

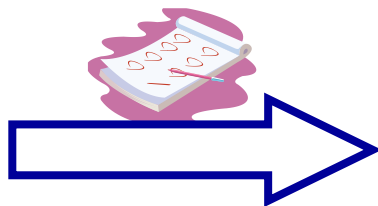
遺伝子発現データ → ある疾病のリスクのありなし

3. 物体認識

画像 → なにが写っているか

教師有り学習のイメージ

学習(訓練)



さまざまな例題の
解答を教える
(教師データ)



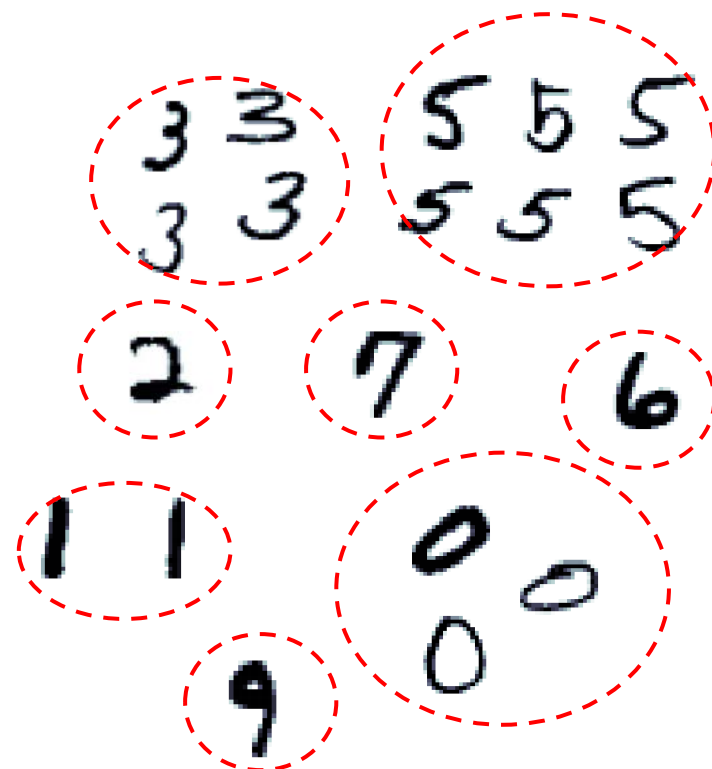
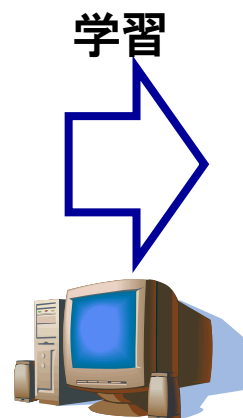
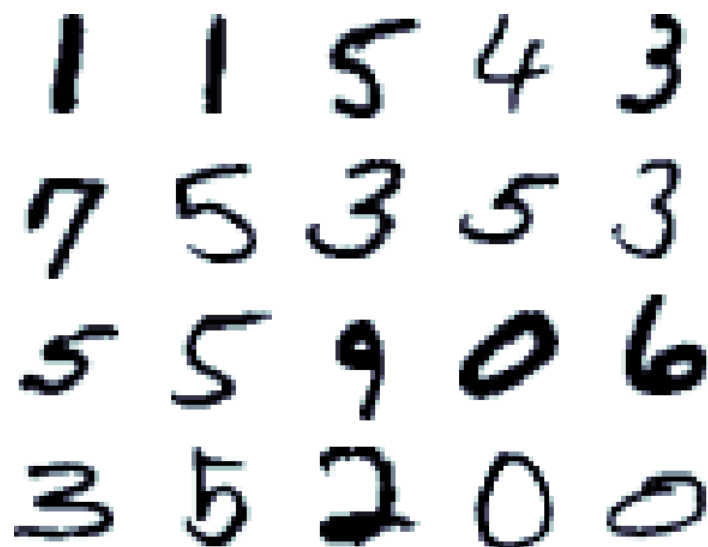
テスト



テスト問題が
解けるようになる



教師なし学習



ラベルなしで画像データのみ
が与えられる

似たものの同士をまとめる

1. ニュース記事を話題ごとに分類（クラスタリング）
2. 音源分離（独立成分分析）
3. 故障の検知（異常検知）

本講義の内容

機械学習手法の諸手法を統計的観点から解説

- ・ 教師有り学習

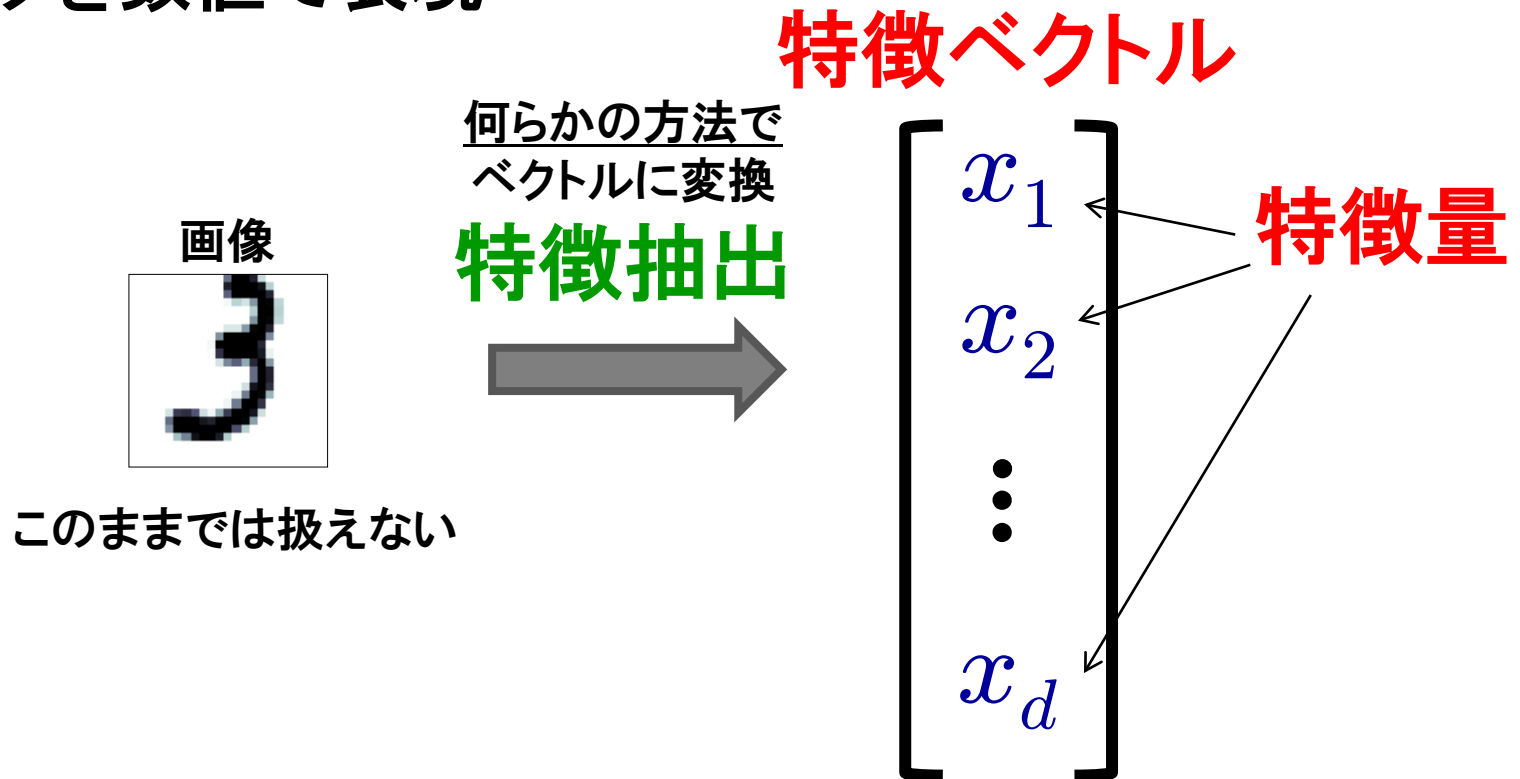
- 統計的決定理論
- 過学習と正則化: バイアス-バリエアンスのトレードオフ
- 高次元スパース推定
 - ・ Lasso, 低ランク行列推定, グループ正則化, 共分散選択
- 正則化学習法の最適化技法
- オンライン学習
- カーネル法
- 学習理論

- ・ 教師なし学習

- クラスタリング
 - ・ K-means法
 - ・ Gaussian Mixture Model, EMアルゴリズム
- 独立成分分析

特徴量の概念

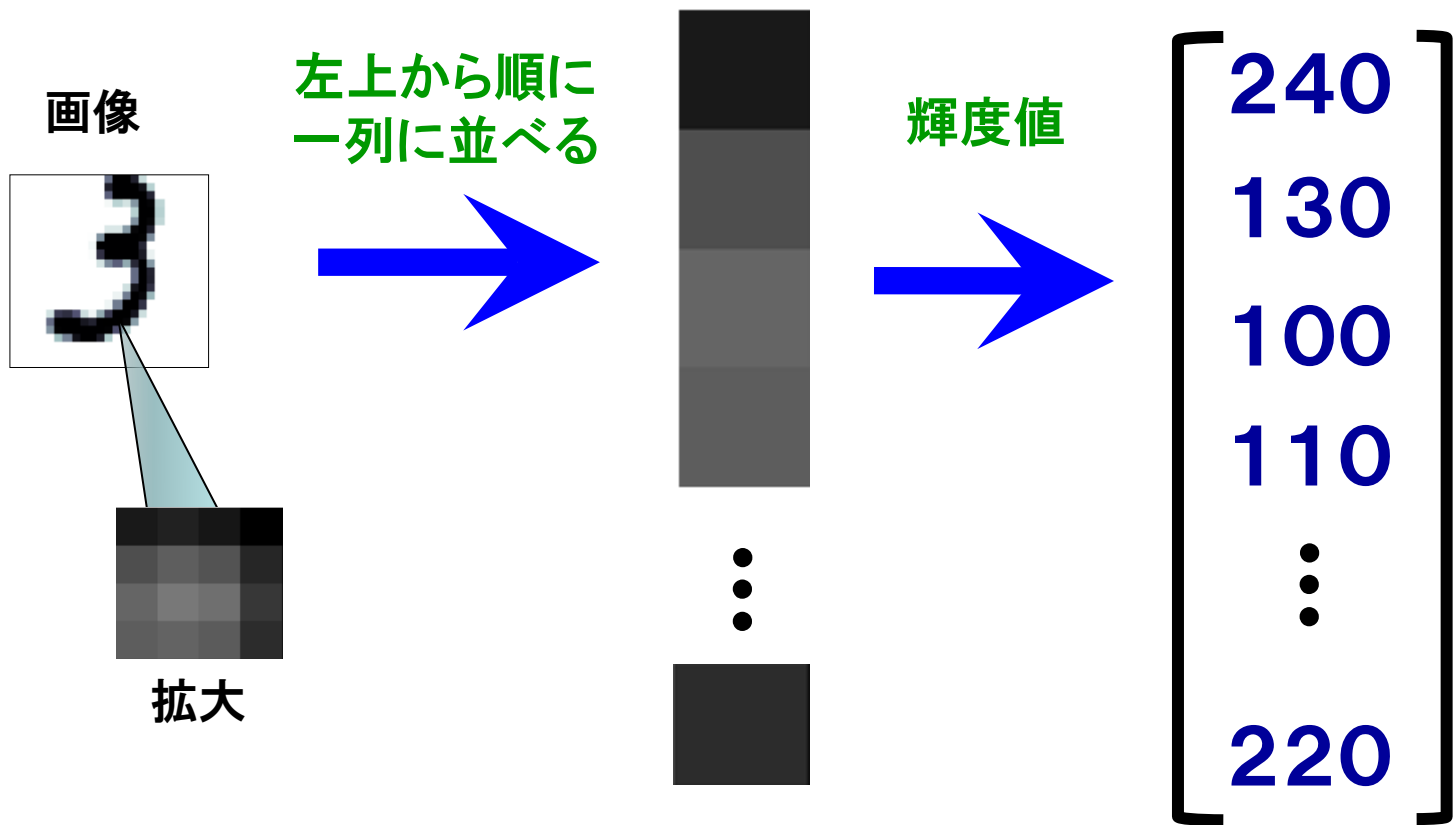
データを数値で表現



- データに応じて様々な特徴量が提案されている。
- 画像ならSIFT,SURF,GIST,勾配など。
- Deep Learning: 特徴量を自動的に学習できる。
- 実はこの手順はとっても大事

ナイーブな方法

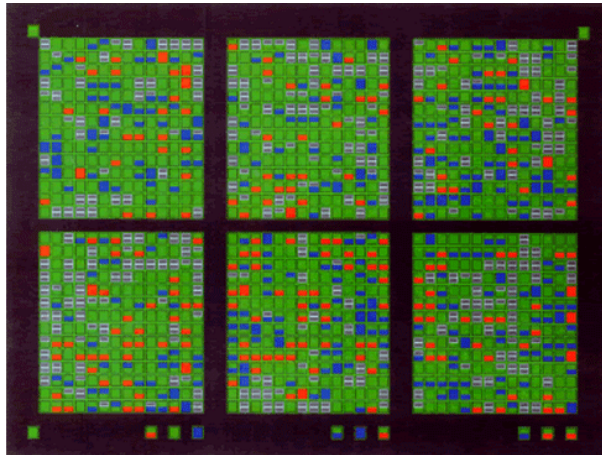
ピクセルごとの輝度を並べる



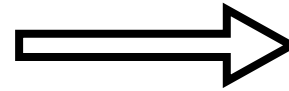
画像データでなくとも同様

特徴ベクトル

DNAデータ



ベクトル化



教師有り学習の手続き

- 特徴ベクトルの概念によりデータによらず共通の手法が適用可能.

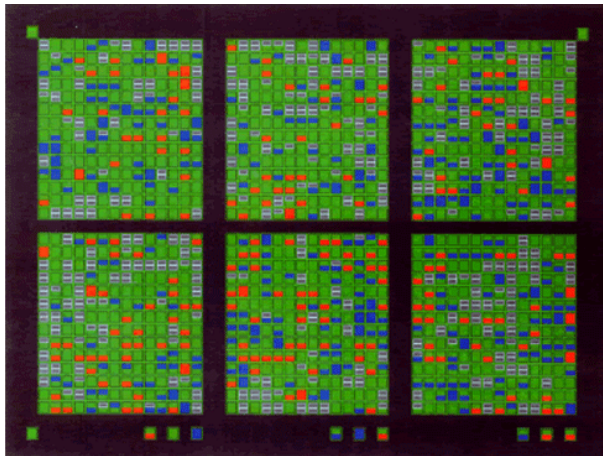


特徴ベクトル x からラベルへの写像 $f(x)$ を
推定する問題に落ちる。
→ 統計の問題！

高次元性の問題

- 特徴ベクトルは高次元になりがち
→ 無駄な情報を落とそう: **スパース推定**

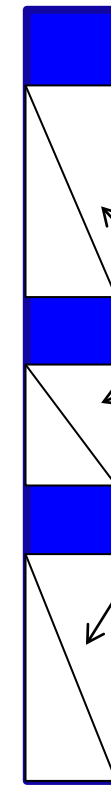
DNAデータ



ベクトル化



特徴ベクトル



多くの特徴量は特定の
疾病に無関係

高次元

スパース推定を用いることで次元がサンプル数より
高くても推定可能に。

スパース推定の応用

- 推薦システム

	Movie A	Movie B	Movie C	...	Movie X
User 1	4	8	*	...	2
User 2	2	*	2	...	*
User 3	2	4	*	...	*
⋮					

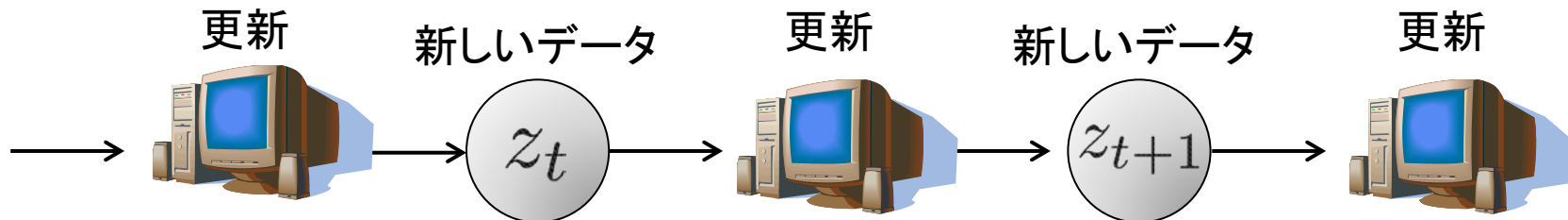
各ユーザーが各映画をどれだけ好むかという部分的情報がある。

→ 残りの部分(*の部分)を埋めたい。

低ランク行列補完で可能。

オンライン学習

- **ビッグデータ時代**
 - 沢山のサンプル: データが到着するごとに学習.
- 一度に全部のデータを使いたくない.
- 今到着したデータのみを用いて情報を更新.
- 過去のデータを見返す必要がない分, 計算が楽.
- しかも, 収束の保証あり.

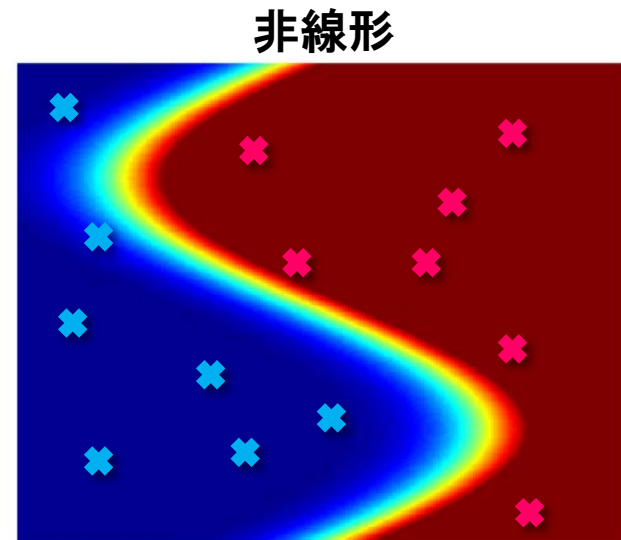
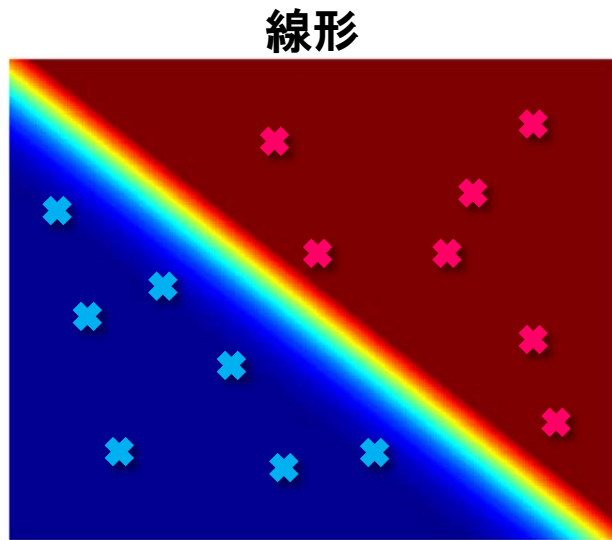


カーネル法

- 非線形な学習

赤と青の領域の間に境界(判別平面)を引きたい。

例: 赤は顔画像の領域, 青は顔でない画像の領域



線形な処理は簡単(線形代数の領域)

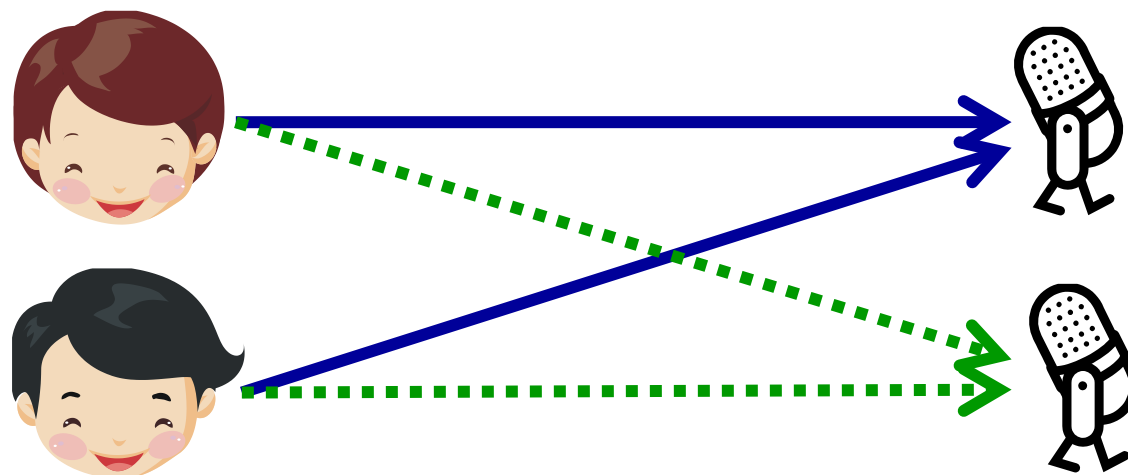
非線形な処理は難しい

→ **カーネル法で解決**

独立成分分析

- カクテルパーティー問題

人は複数人が話しているなかでも話し相手の声だけ抽出可能



独立成分分析:混ざった声を元通りに分離可能

本講義の内容

機械学習手法の諸手法を統計的観点から解説, 基本的に黒板を使用

- ・ 教師有り学習

- 統計的決定理論
- 過学習と正則化: バイアス-バリエアンスのトレードオフ
- 高次元スパース推定
 - ・ Lasso, 低ランク行列推定, グループ正則化, 共分散選択
- 正則化学習法の最適化技法
- オンライン学習
- カーネル法
- 学習理論

- ・ 教師なし学習

- クラスタリング
 - ・ K-means法
 - ・ Gaussian Mixture Model, EMアルゴリズム
- 独立成分分析

成績評価

出席30%＋レポート70%

質問などの連絡先

数理計算科学専攻 鈴木大慈

西八号館W707号室

s-taiji@is.titech.ac.jp