クレジット:
UTokyo Online Education 知の構造化論　2020　美馬　秀樹

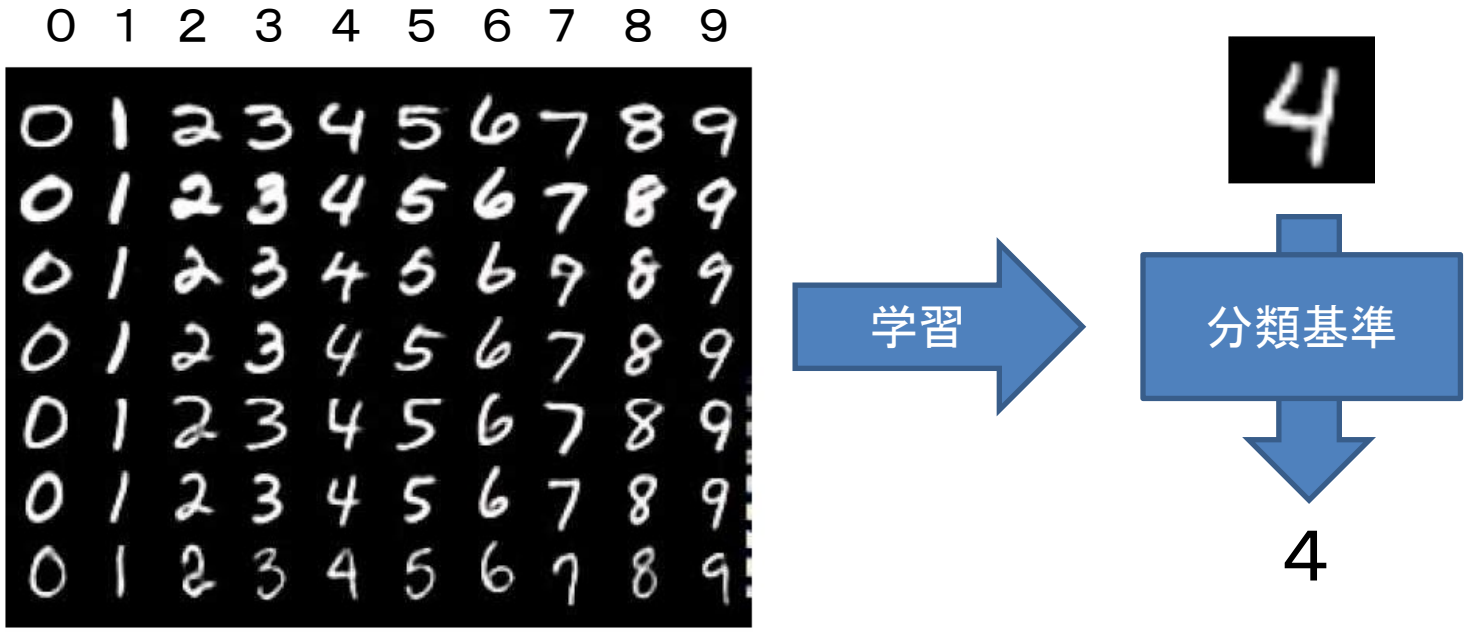
機械学習と演習

東京大学 工学系研究科／大学総合教育研究センター

美馬秀樹

機械学習とは

- 「機械学習」とは
  - **多くのデータから規則性・判断基準を抽出し、それを基に判断・予測を行う手法**
  - 例：手書き文字認識



機械学習

- 「機械学習」とは
  - 多くのデータから規則性・判断基準を抽出し、それを基に判断・予測を行う手法
  - 例：書籍の分類(テキスト分類)

分類	書籍タイトル
情報科学	情報セキュリティ入門
情報科学	進化する情報社会
情報科学	情報社会学概論
情報科学	初めての情報理論
情報科学	情報社会のいま

©いらすとや

「情報科学」の本の共通点は・・・

人間が分類する場合

分類	書籍タイトル
？？？	情報システム入門

分類	書籍タイトル
情報科学	情報システム入門

機械学習

- 「機械学習」とは
  - 多くのデータから規則性・判断基準を抽出し、それを基に判断・予測を行う手法

分類	書籍タイトル
情報科学	情報セキュリティ入門
情報科学	進化する情報社会
情報科学	情報社会学概論
情報科学	初めての情報理論
情報科学	情報社会のいま

学習

分類基準:
もし「情報」という単語が入っていれば分類は「情報科学」

学習フェーズ

予測フェーズ

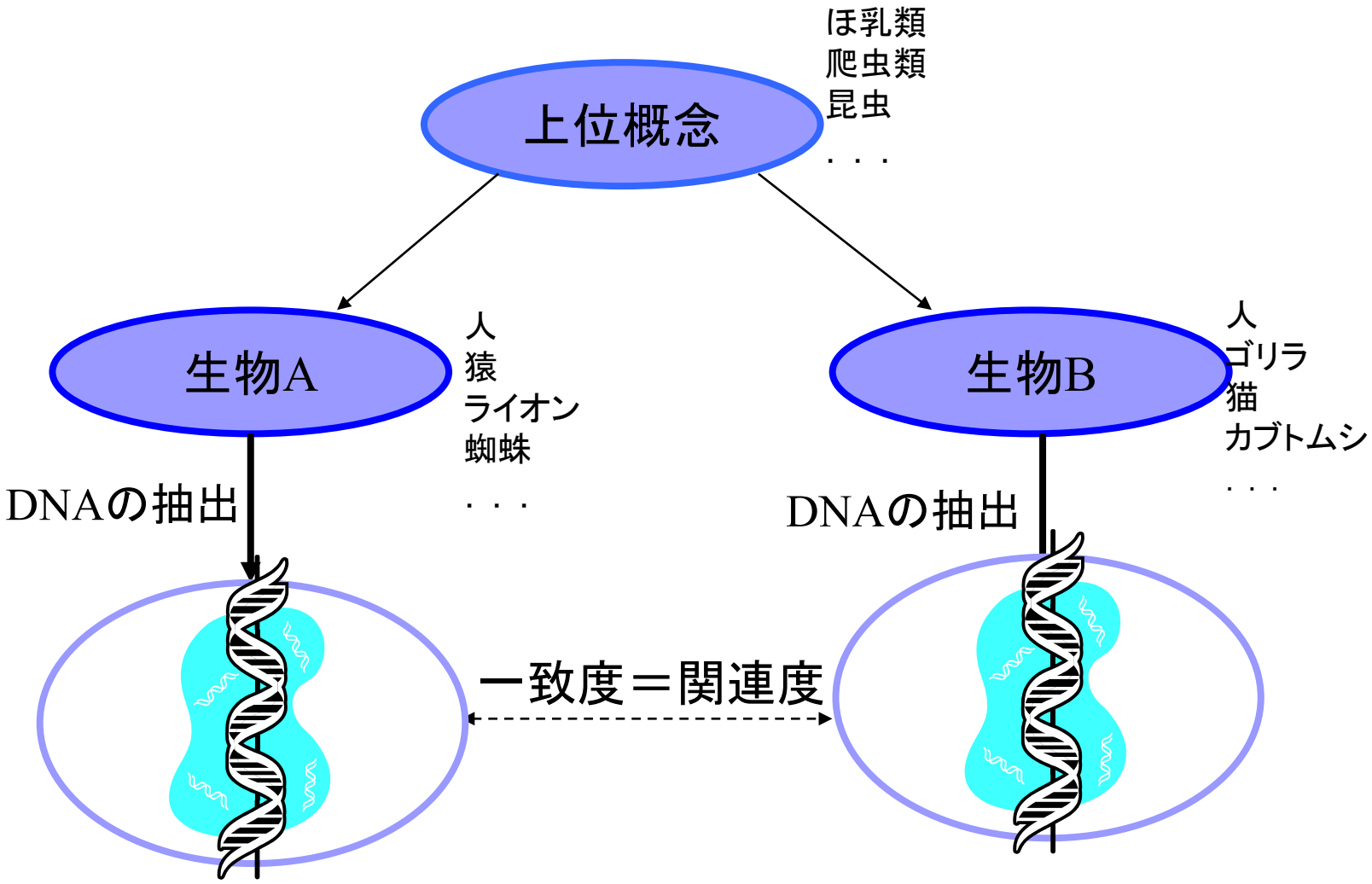
分類が未知のデータ

分類	書籍タイトル
？？？	情報システム入門

予測器

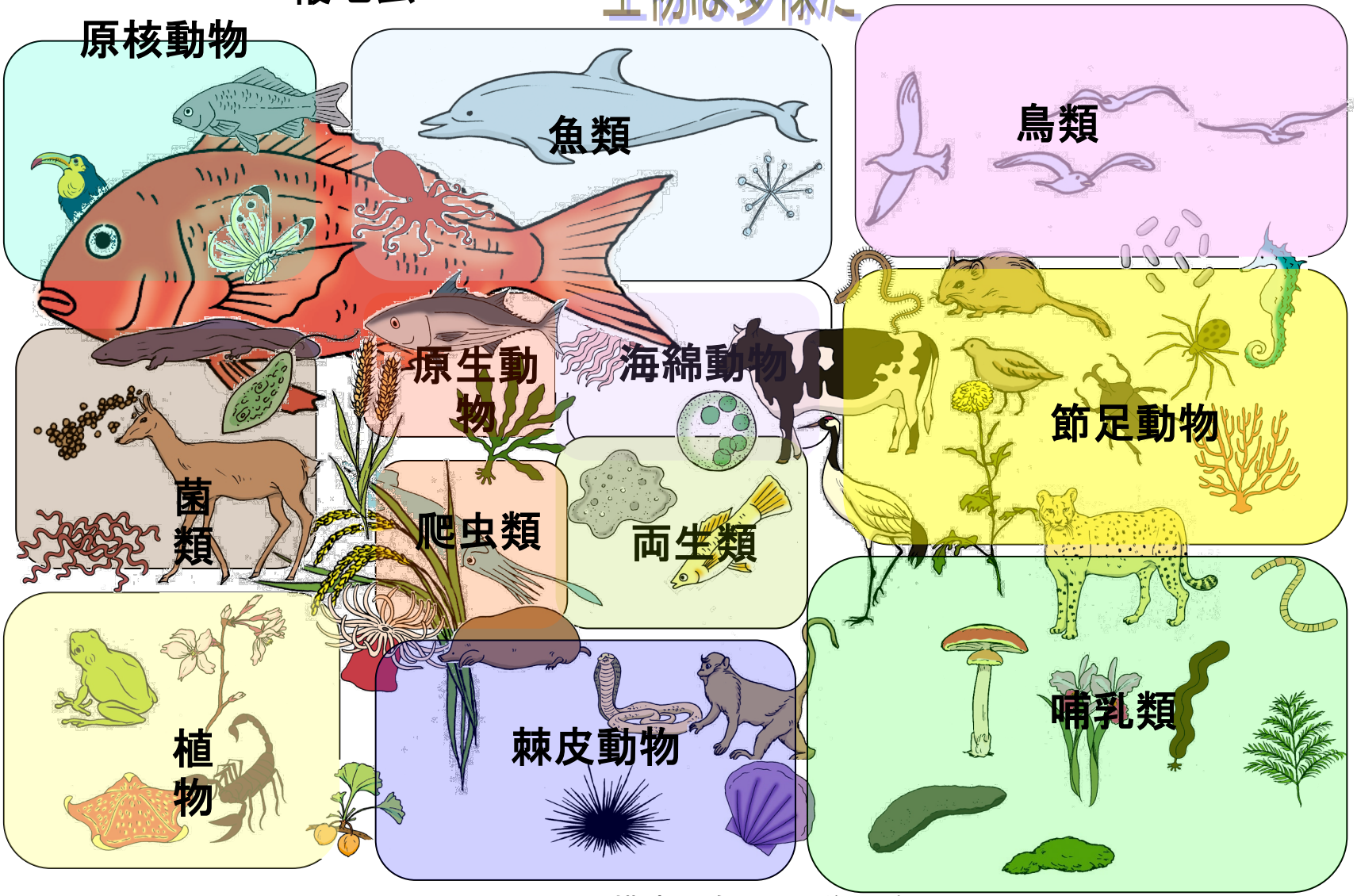
分類	書籍タイトル
情報科学	情報システム入門

バイオ・インフォマティクス



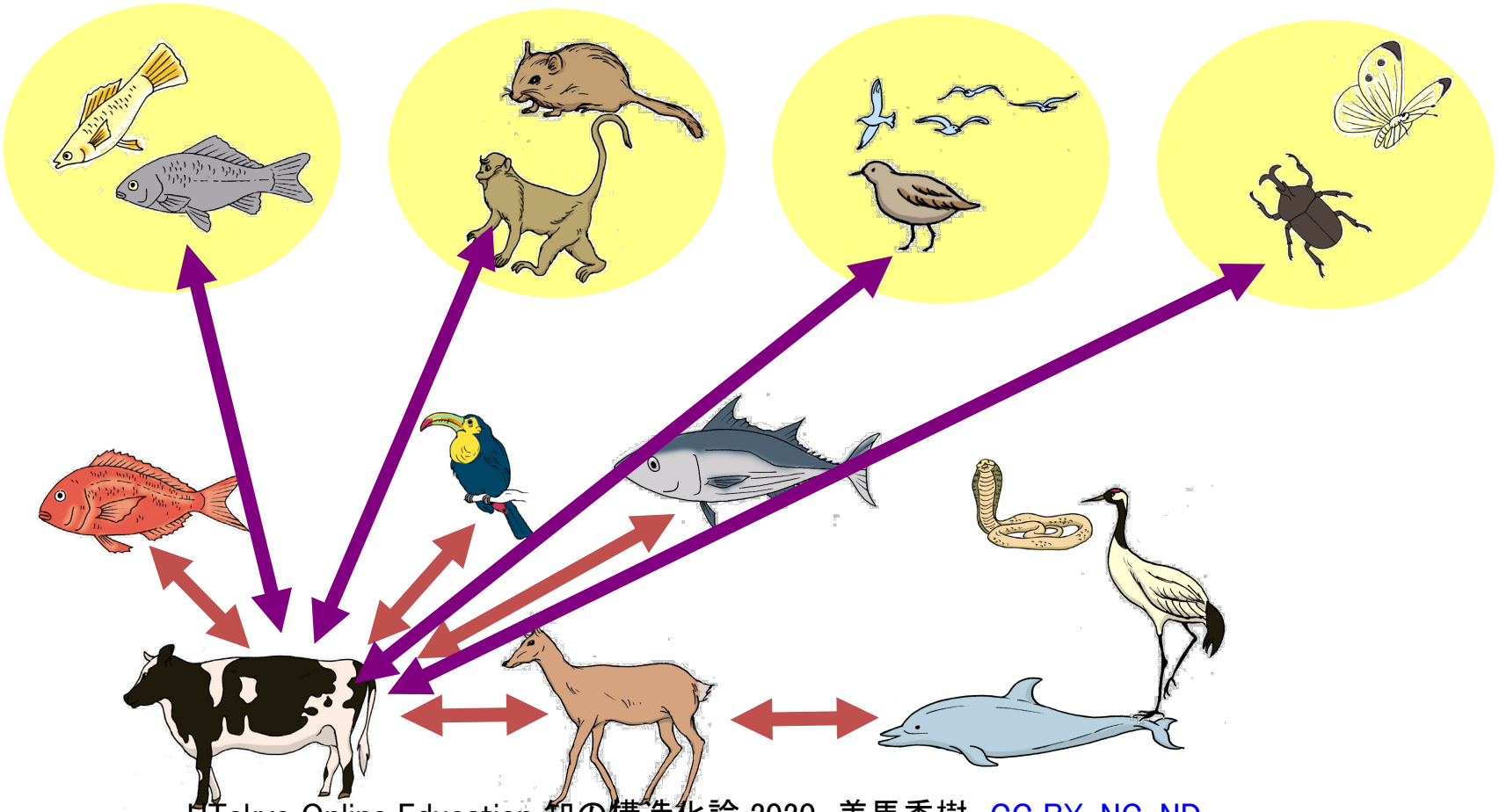
構造化されていない知識

-生物は多様だ-



クラスタリング、カテゴライジング

- 教師あり、教師無し



教師なし機械学習の応用

- 協調フィルタリング
  - 行動モデリングとリコメンデーション

年齢：xxx
性別：男
趣味：スポーツ、…
嗜好：ワイン
…

日時	品目	数量
06.05.22	ダビンチコード	1
06.05.22	ワイン	2
…	…	…

…　　…

同じものを買う可能性が高い

教師あり機械学習

- 学習フェーズ
  - 入力: 学習データ(特徴量(属性)と目的変数)
  - 出力: 分類器
    - 特徴量から目的変数を推定する分類器

分類	書籍タイトル
情報科学	情報セキュリティ入門
情報科学	進化する情報社会
情報科学	情報社会学概論
情報科学	初めての情報理論
情報科学	情報社会のいま

学習

分類器
もし「情報」という単語が入っているならば分類は「情報科学」

特徴量：書籍タイトル中の単語の頻度
目的変数：書籍の分類

教師あり機械学習

- 予測フェーズ
  - 入力: テストデータ(特徴量)
  - 出力: 目的変数
    - 学習フェーズで作成した分類器を用いて特徴量 から 目的変数の推定を行う

目的変数が未知のデータ

分類	書籍タイトル
？？？	情報システム入門

特徴量:書籍タイトル中の単語の頻度

単語	頻度
情報	1
システム	1
入門	1

分類	書籍タイトル
情報科学	情報システム入門

分類基準:
もし「情報」という単語が入っていれば分類は「情報科学」

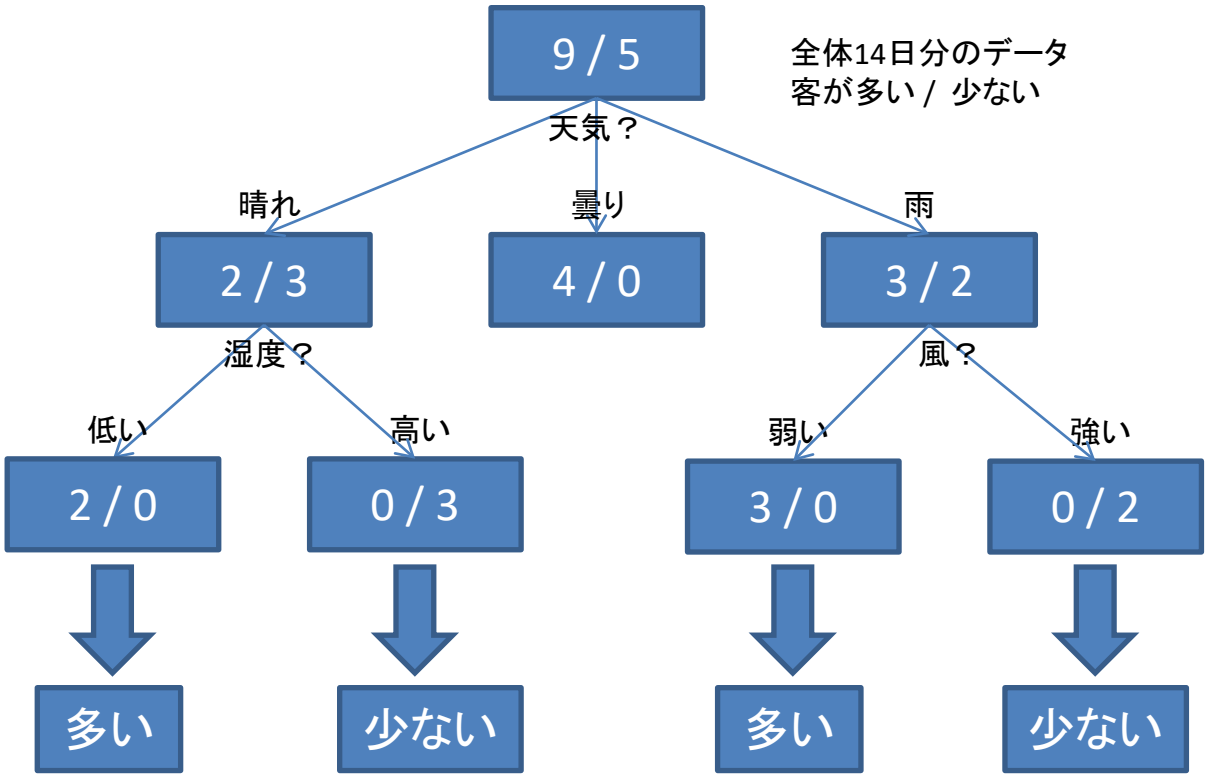
機械学習手法：決定木

- 特徴量の条件を繰り返し適用し分類を行う
  - 条件適用の順番を学習データから決定する

例：気象条件からゴルフ場の客数の予測

日	天気	湿度	風	客数
1	晴れ	高い	弱い	少ない
2	晴れ	高い	強い	少ない
3	曇り	低い	弱い	多い
4	雨	高い	弱い	多い
⋮	⋮	⋮	⋮	⋮

特徴量: 天気、湿度、風
目的変数: 客数(多 or 少)



決定木の応用例

- Akinator

  質問を繰り返して思い 浮かべているものを 当てる

  http://jp.akinator.com/

著作権等の都合により、
ここに挿入されていた画像を削除しました

Akinatorの画面
http://jp.akinator.com/

自然言語処理・人工知能・機械学習の関係

- 自然言語処理は人工知能分野の一部

- 人工知能 ≠ 機械学習
  - 機械学習は人工知能分野の技術の一つ
  - 機械学習を使わない人工知能もある

- 人工知能 ≠ ディープラーニング
  機械学習 ≠ ディープラーニング
  - ディープラーニングは機械学習手法の一つ

自然言語処理と機械学習
―大学間でのカリキュラムの比較への応用―

- 大学間での学生の移動を促進するため、大学間でカリキュラムを比較したい

- 電子シラバスを標準的な分野体系に分類できれば比較ができるのではないか？

  ⇒シラバスを自然言語処理を用いて自動分類する

シラバスの自動分類

- 対象データ: シラバステキスト
- 分類体系: NDC(日本十進分類)
  - 図書に対する分類法
  - 分類記号として数字を用いた階層的な分野分類

10区分		100区分		1000区分	
000	総記	300	社会科学一般	320	法律
100	哲学, 思想	310	政治	321	法学
200	歴史, 地理	320	法律	322	法制史
300	社会科学	330	経済	323	憲法
400	自然科学, 医学	340	財政	324	民法
500	技術･工学	350	統計	325	商法
600	産業	360	社会	326	刑法, 刑事法
700	芸術･美術, スポーツ	370	教育	327	司法, 訴訟手続法
800	言語	380	風俗, 民俗学	328	諸法
900	文学	390	国防, 軍事	329	国際法

西南学院大学図書館　資料の探し方
日本十進分類法（NDC）の一例
http://www.seinan-gu.ac.jp/library/old/guide/search.html

分類処理の流れ

前準備

Wikipedia
テキストデータ

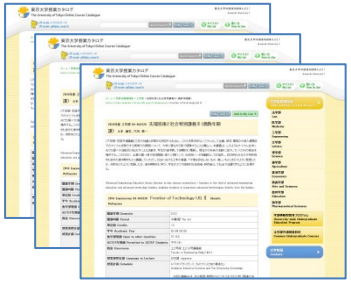
Word2Vec

工学: (-0.28987, 2.20560,
-0.13070　0.67409, …)

単語の
ベクトル表現

学習フェーズ

NDC付きシラバス



ベクトル化

(0.72628,　0.84896,
1.94840,　0.66509, …)

NDC +
文書ベクトル
表現

Random Forest

NDC分類器

分類フェーズ

シラバス

ベクトル化

文書
ベクトル表現

NDC分類器

NDC

word2vec[Mikolov 2013]

- ニューラルネットワークを用いて単語をベクトル表現化する手法
- テキスト中の各単語に対しその周辺に出現する単語の情報を基に計算
- ベクトル表現はある種の単語の意味を表現する
  - ベクトル間の演算や意味の演算が可能
    - 例: v(king) - v(man) + v(woman) = v(queen)
- 今回は Wikipedia のテキストから学習

工学: (-0.28987, 2.20560, -0.13070　0.67409, …)
分野: (1.17059, 1.94050, 1.00932, 1.04591, …)
必要: (-0.76447, 1.06354, 2.38880, -0.42196, …)
・・・

対象文書のベクトル化

「数学ＩＡ」シラバス
工学全分野で必要不可欠な道具である、常微分方程式、ベクトル解析、変分法について学ぶ。実践的な理解を目指す。…

単語抽出

工学、全、分野、で、必要、不可欠、だ、道具、だ、ある、・・・

各単語のベクトル化

工学: (-0.28987, 2.20560, -0.13070　0.67409, …)
全:　(0.72628, 0.84896, 1.94840, 0.66509, …)
分野: (1.17059, 1.94050, 1.00932, 1.04591, …)
で:　(1.89374, 2.01249, -0.65686, -2.03772, …)
必要: (-0.76447, 1.06354, 2.38880, -0.42196, …)
・・・

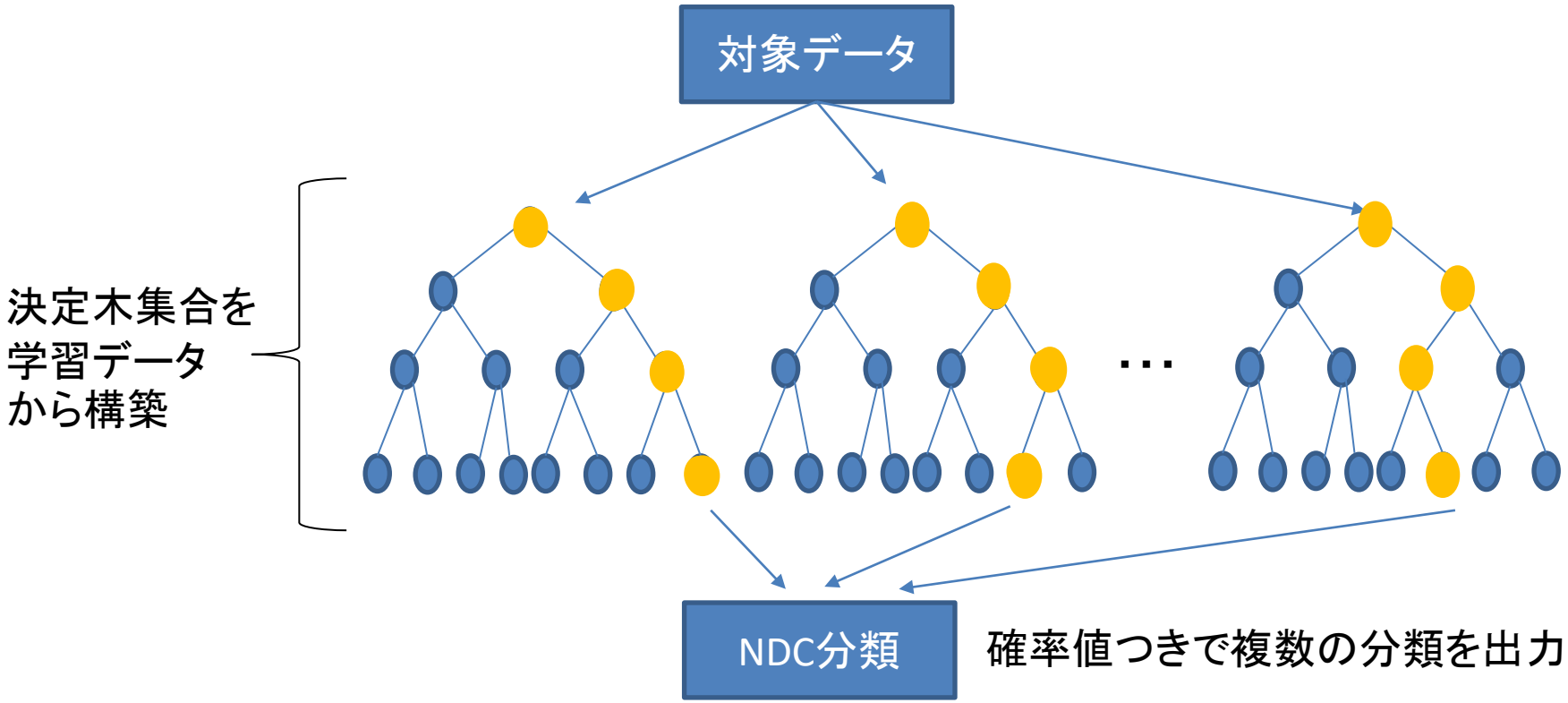
平均

(0.77248 1.13985 -0.11331 -1.13872, …)

対象テキストを200次元の特徴ベクトルに変換

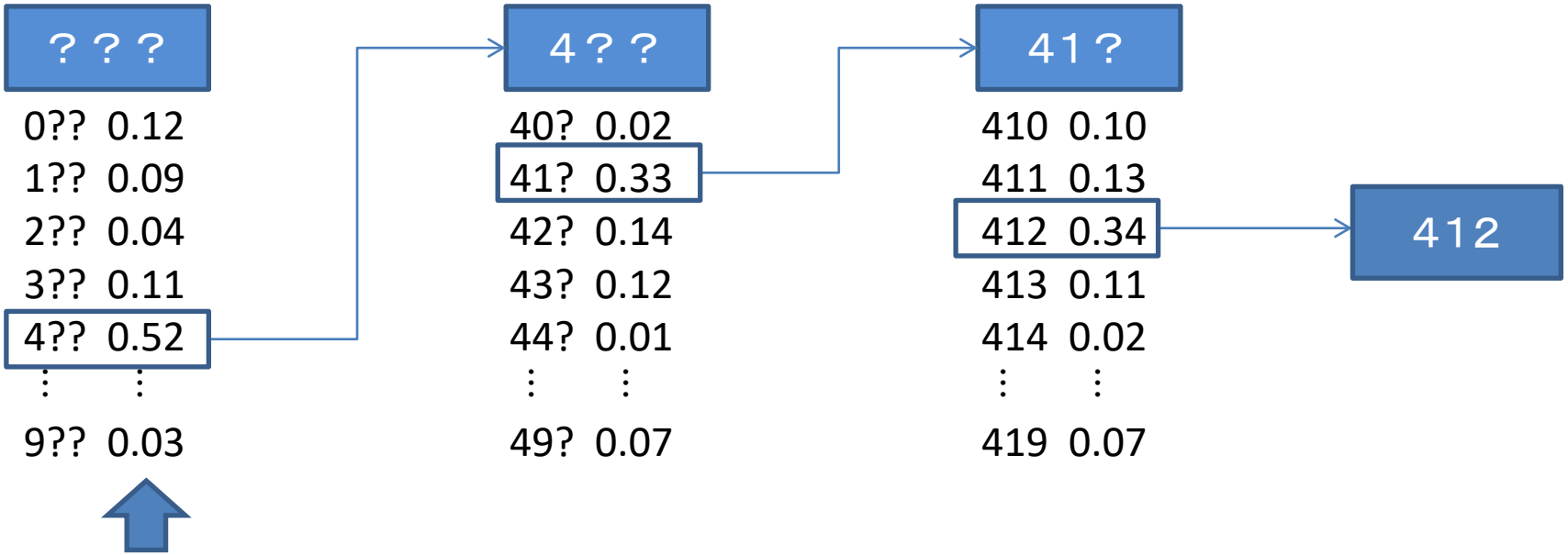
Random Forest

- 決定木を大量に作成し多数決で分類を決定



分類手法: 階層的分類

- NDCの3桁までを対象とし1桁ずつ分類を行う



分類実験

- データ
  - 東京大学授業カタログの2015年度シラバス
    - 人手でNDCを付与
    - 一講義に複数のNDCも可
- 実験結果(分類精度)

学部・研究科	講義数
工学部(A1A2のみ)	420
文学部	741
人文社会学研究科	440
教育学部	110
合計	**1,711**

	一桁目	二桁目	三桁目
TOP1	0.707	0.452	0.253
TOP2	0.856	0.676	0.466
TOP3	0.909	0.816	0.673

自動分類結果例

講義名	人手によるNDC	自動分類結果(TOP3)
環境エネルギーシステム	501(工業基礎学), 543(発電)	541(電気回路・計測・材料) 531(機械力学・材料・設計) 501(工業基礎学)
基礎情報学	007(情報科学)	531(機械力学・材料・設計) 007(情報科学) 417(確率論．数理統計学)
哲学演習(2)	133(近代哲学)	892(ラテン語) 134(ドイツ・オーストリア哲学) 829(その他東洋の諸言語)
日本語音韻の諸問題	811(音声．音韻．文字)	810(日本語) 811(音声．音韻．文字) 837(読本．解釈．会話)
博物館教育論	069(博物館), 370(教育)	375(教育課程．学習指導．教科別教育) 370(教育) 361(社会学)

他大学シラバスの分類実験

- 東北大学シラバスの自動分類実験

講義名	自動分類結果(TOP3)
材料力学Ⅰ	531(機械力学・材料・設計) 413(解析学) 427(電磁気学)
流体力学Ⅰ	531(機械力学・材料・設計) 534(流体機械．流体工学) 417(確率論．数理統計学)
ロボット工学	801(言語学) 413(解析学) 830(英語)

講義名	自動分類結果(TOP3)
考古学概論	202(歴史補助学) 210(日本史) 207(研究法．指導法．歴史教育)
倫理思想概論	134(ドイツ・オーストリア哲学) 150(倫理学．道徳) 131(古代哲学)
宗教学概論	167(イスラム) 182(仏教史) 181(仏教教理．仏教哲学)

カリキュラムの比較

http://mimasearch.ut-catalog.iit.jp/