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


東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

# 自然言語処理の基礎
## (+人工知能、機械学習との関連)

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

# 講義内容

- 自然言語処理(+人工知能、機械学習との関連)
- 自然言語処理の応用

# 自然言語処理

- 「自然言語処理」とは？
  - 人間が日常的に使う言語をコンピュータで処理(理解)すること
  - 英語では Natural Language Processing (NLP)
  - 今流行の人工知能研究の一分野
- 「自然言語」とは？
  - 人間が日常的に使う、自然に発生した言語

    ⇔人工言語(プログラミング言語など)

# 自然言語処理で何ができる？

- 究極の目標

自然言語理解

人間のことばを理解する

人間のことばを聞く

人間のことばを話す

音声認識

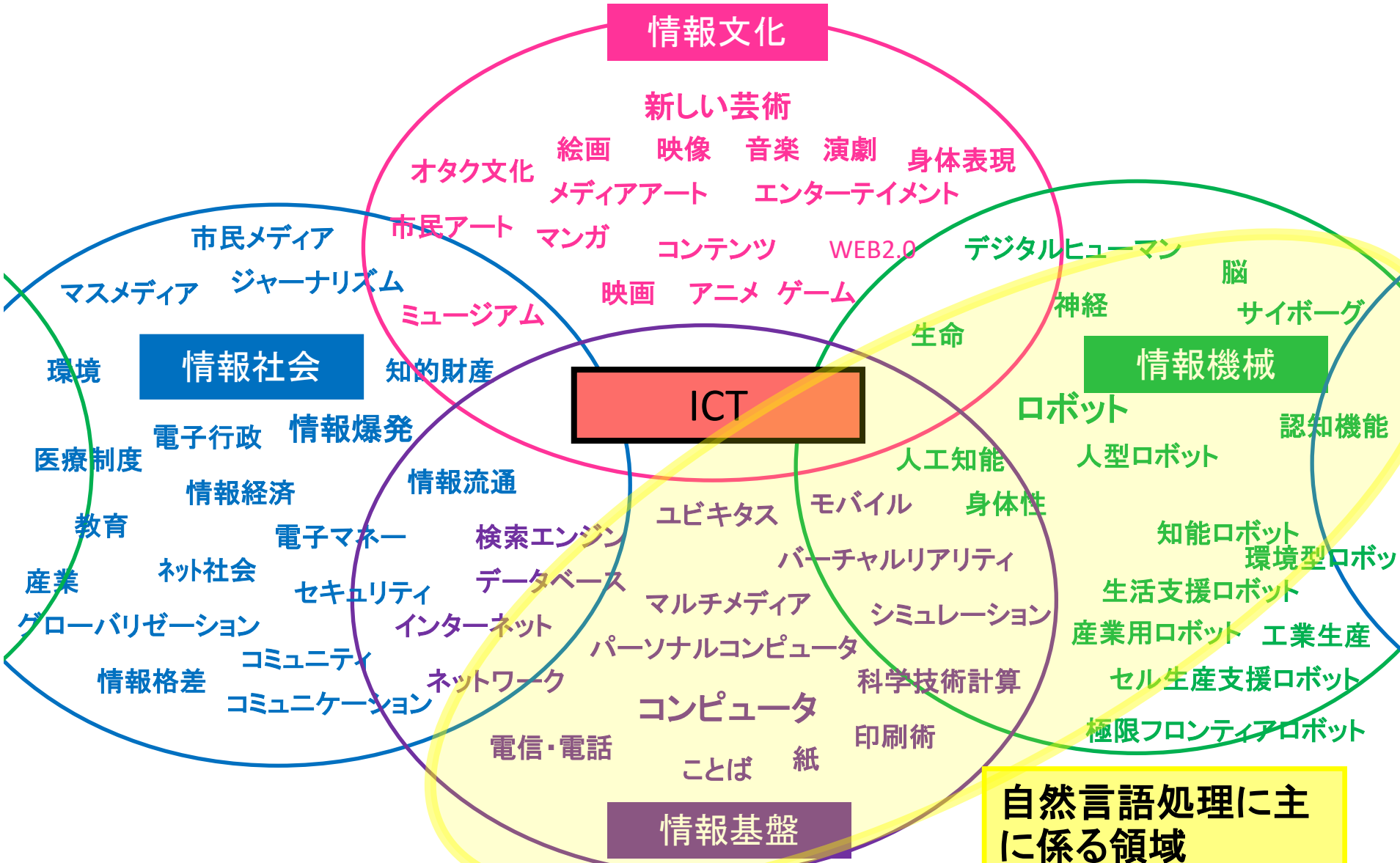
文章生成
音声合成

イラスト©いらすとや

言語に関する部分に限る

# 情報が世界を変える　－俯瞰図－



# 人工知能

- 「人工知能(AI: Artificial Intelligence)」とは
  - 明確な定義はなく、人によって定義が違う

| 研究者 | 所属 | 定義 |
|---|---|---|
| 中島秀之 | 公立はこだて未来大学 | 人工的につくられた、知能を持つ実態。あるいはそれをつくろうとすることによって知能自体を研究する分野である |
| 武田英明 | 国立情報学研究所 | |
| 西田豊明 | 京都大学 | 「知能を持つメカ」ないしは「心を持つメカ」である |
| 溝口理一郎 | 北陸先端科学技術大学院 | 人工的につくった知的な振る舞いをするためのもの（システム）である |
| 長尾真 | 京都大学 | 人間の頭脳活動を極限までシミュレートするシステムである |
| 堀浩一 | 東京大学 | 人工的に作る新しい知能の世界である |
| 浅田稔 | 大阪大学 | 知能の定義が明確でないので、人工知能を明確に定義できない |
| 松原仁 | 公立はこだて未来大学 | 究極には人間と区別が付かない人工的な知能のこと |
| 池上高志 | 東京大学 | 自然にわれわれがペットや人に接触するような、情動と冗談に満ちた相互作用を、物理法則に関係なく、あるいは逆らって、人工的につくり出せるシステム |
| 山口高平 | 慶應義塾大学 | 人の知的な振る舞いを模倣・支援・超越するための構成的システム |
| 栗原聡 | 電気通信大学 | 人工的につくられる知能であるが、その知能のレベルは人を超えているものを想像している |
| 山川宏 | ドワンゴ人工知能研究所 | 計算機知能のうちで、人間が直接・間接に設計する場合を人工知能と呼んで良いのではないかと思う |
| 松尾豊 | 東京大学 | 人工的につくられた人間のような知能、ないしはそれをつくる技術。人間のように知的であるとは、「気づくことのできる」コンピュータ、つまり、データの中から特徴量を生成し現象をモデル化することのできるコンピュータという意味である |

松尾豊『人工知能は人間を超えるか』(KADOKAWA、2015)p45をもとに講師作成

# 人工知能

- 「人工知能(AI: Artificial Intelligence)」とは
  - (コンピュータを使って)人間の知能の働きを人工的に実現したもの
    - 自然言語処理
    - ゲームAI: 将棋・囲碁でプロ棋士に勝利
    - 画像認識: 人間よりも高精度
    - 自動運転: 数年後の実用化に向けて実験中
  - 近年ビッグデータと機械学習により飛躍的に発展
    - データの増加と処理可能なコンピュータの発展

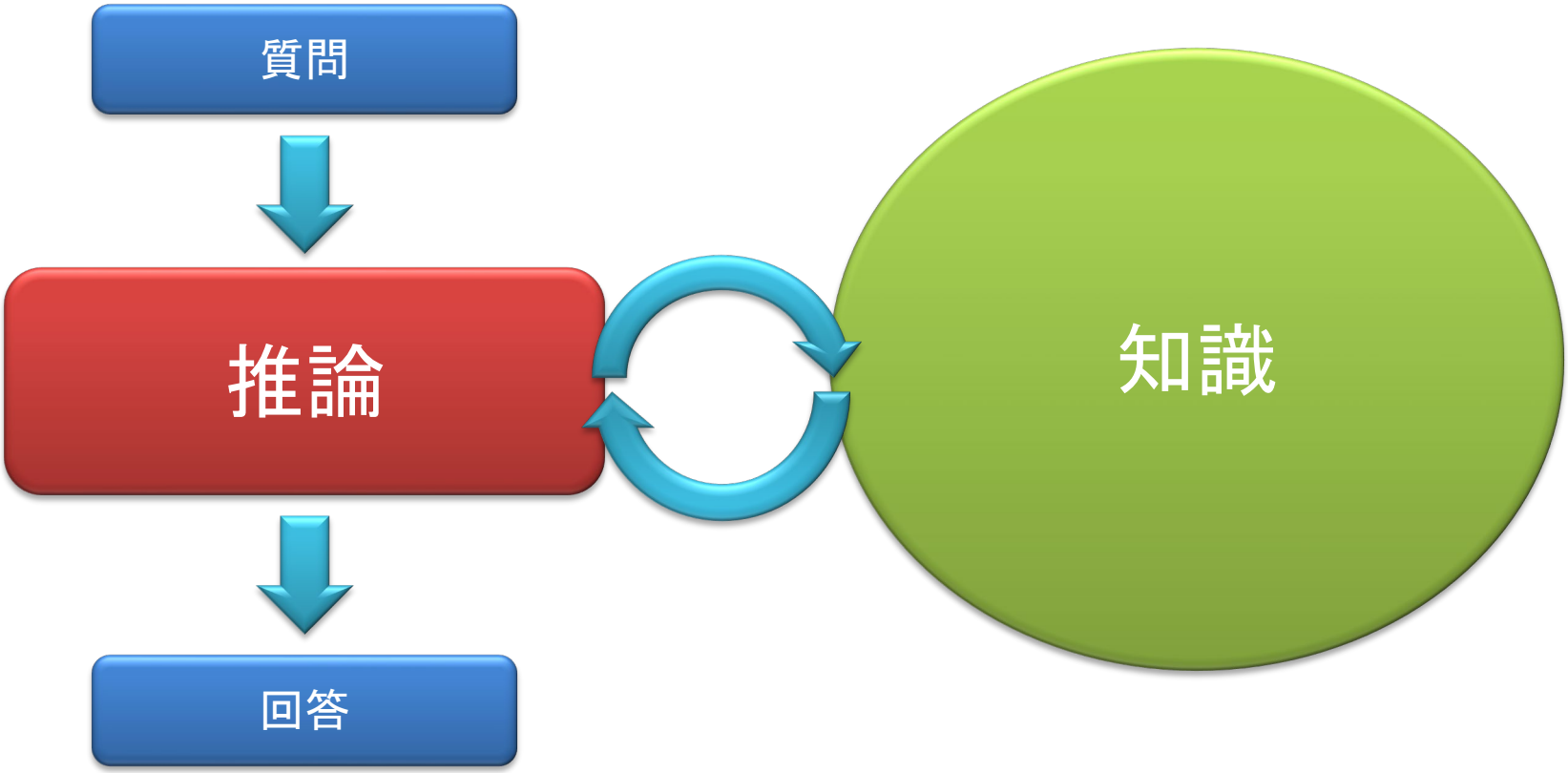
# 「強いAI」と「弱いAI」

- 「強いAI」
  - 汎用人工知能
  - 人間と同等かそれ以上の能力を持つ
    - 映画などに出てくる、世の中でイメージされる人工知能
  - 実現はまだまだ先(2045年？)
- 「弱いAI」
  - 「強いAI」の一部となる人工知能
  - 特定の問題に対してのみ処理可能
    - 例：将棋、囲碁、画像認識・・・
    - 問題を与えればそれを処理するが、それ以外はできない
  - 今流行っているのはこちら

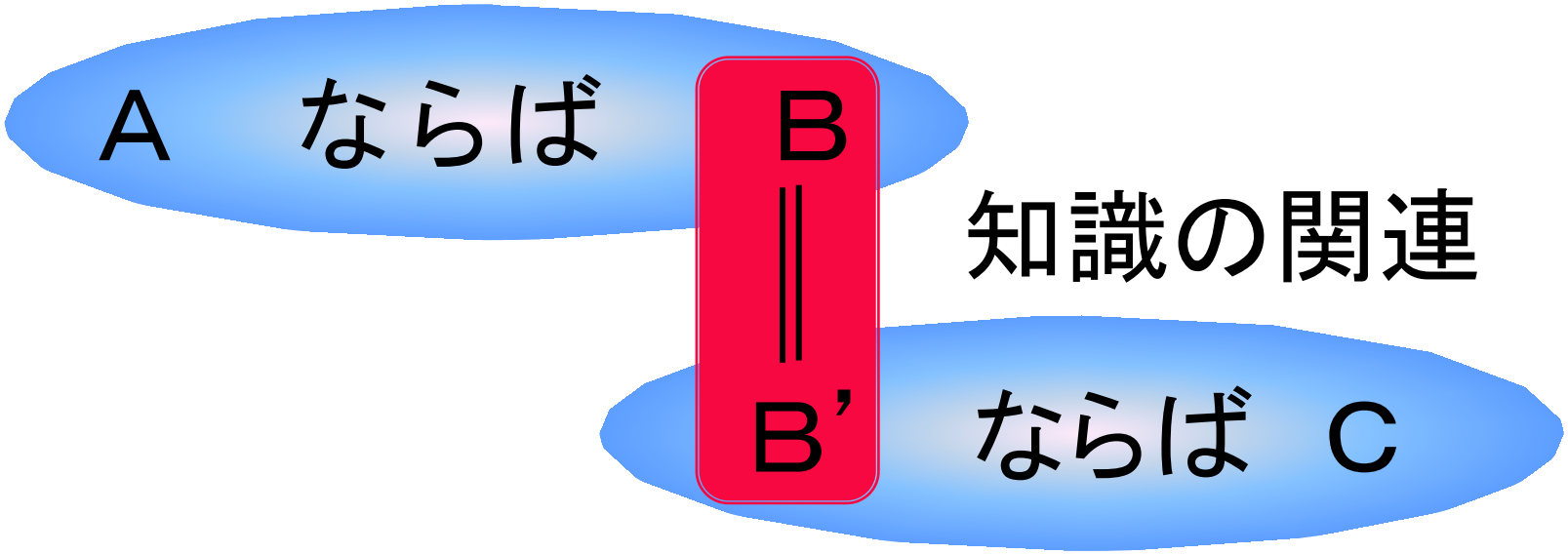
# 人工知能の基礎モデル

- エキスパートシステム



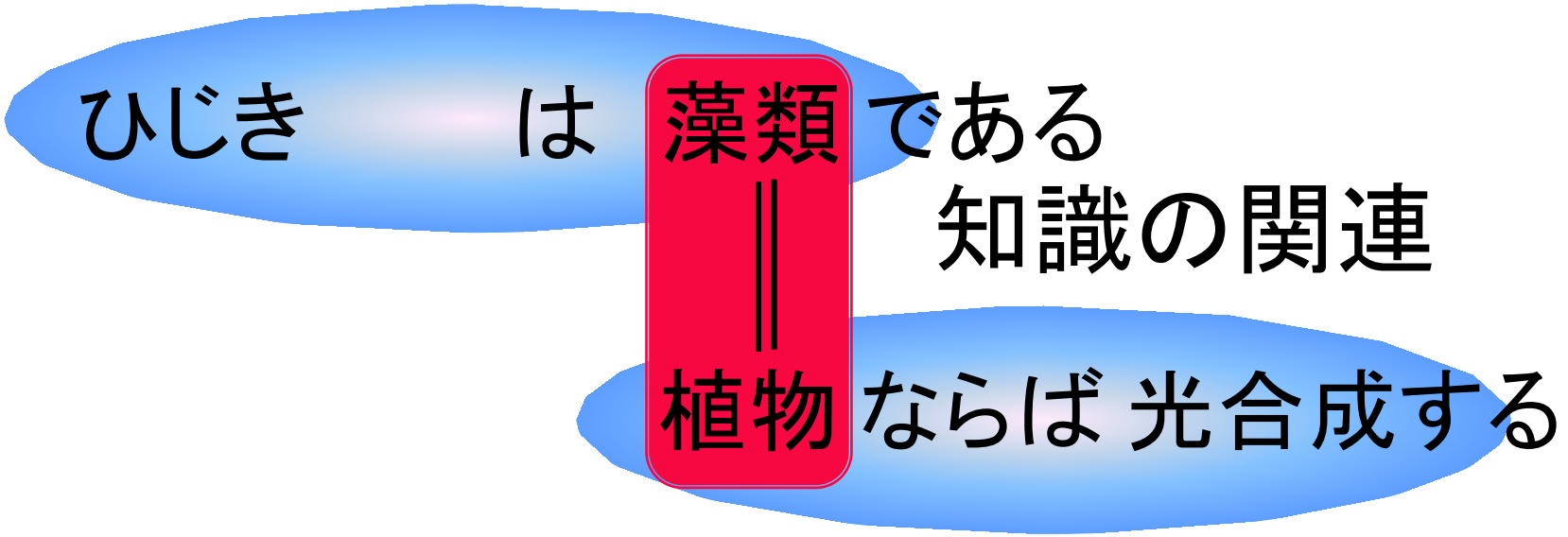
# 知識と推論－三段論法－




新たな知識

A　ならば　C

# 知識と推論の例




新たな知識

ひじき　ならば　光合成する

# 知識と推論による質問応答

- 「もし　A　ならば B」の集まり
  - もし 鳥である ならば 羽がある
  - もし 羽がある ならば 空を飛ぶ
  - もし カラスである ならば 鳥である
  - マギー は カラスである

ペンギンは？

コウモリは？

マギーは空を飛ぶか？

YES

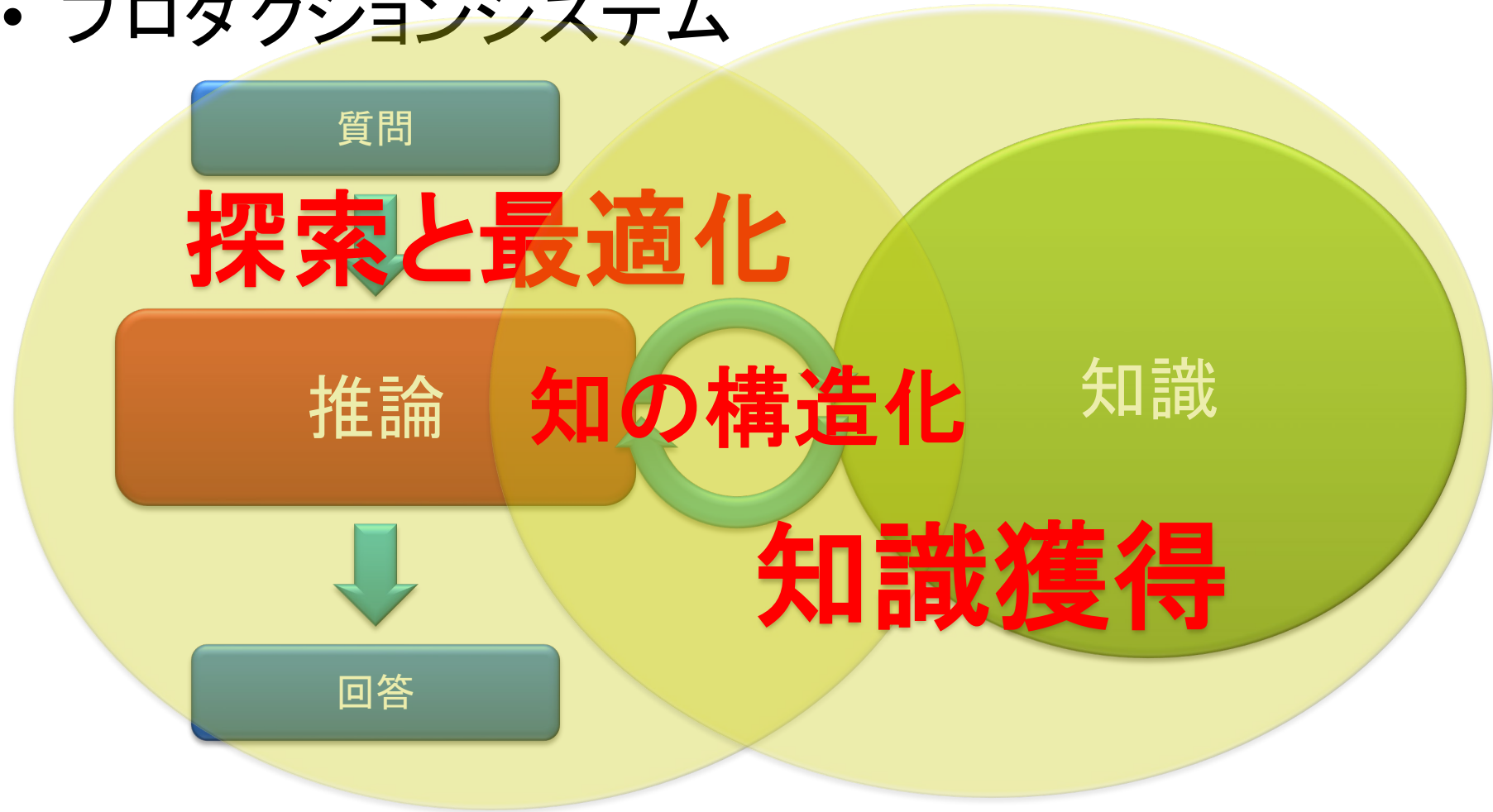
空を飛ぶのは何か？

マギー です

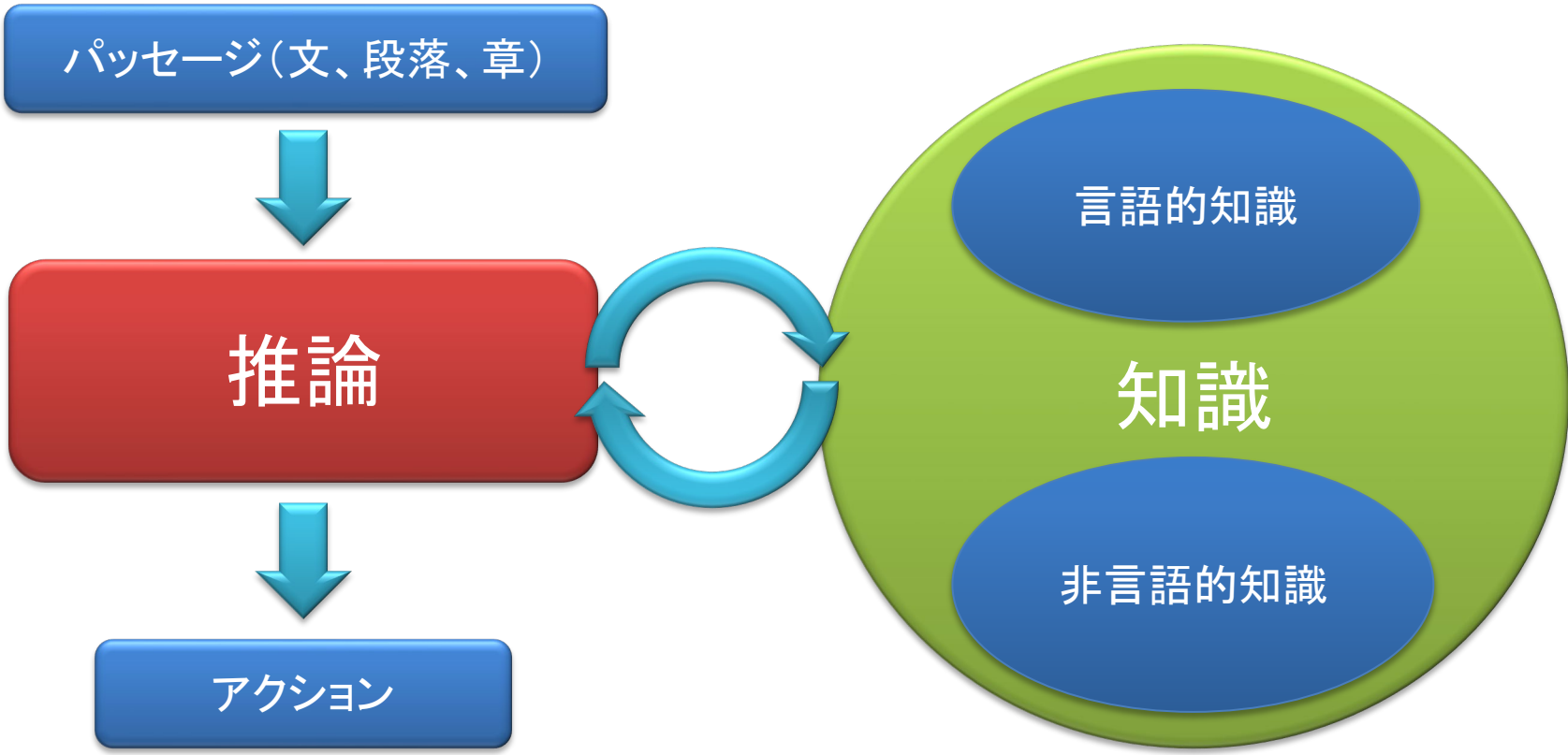
# 人工知能の基礎

- プロダクションシステム



# 自然言語処理の基礎

- プロダクションシステム



# データから知識・知へ

- データ：加工されていない生の記録
  取得における条件が明らかであることが大切

- 情報：データが何らかの文脈で解釈（理解）されたもの。それぞれの集団によって共通の意味を汲みとられる。

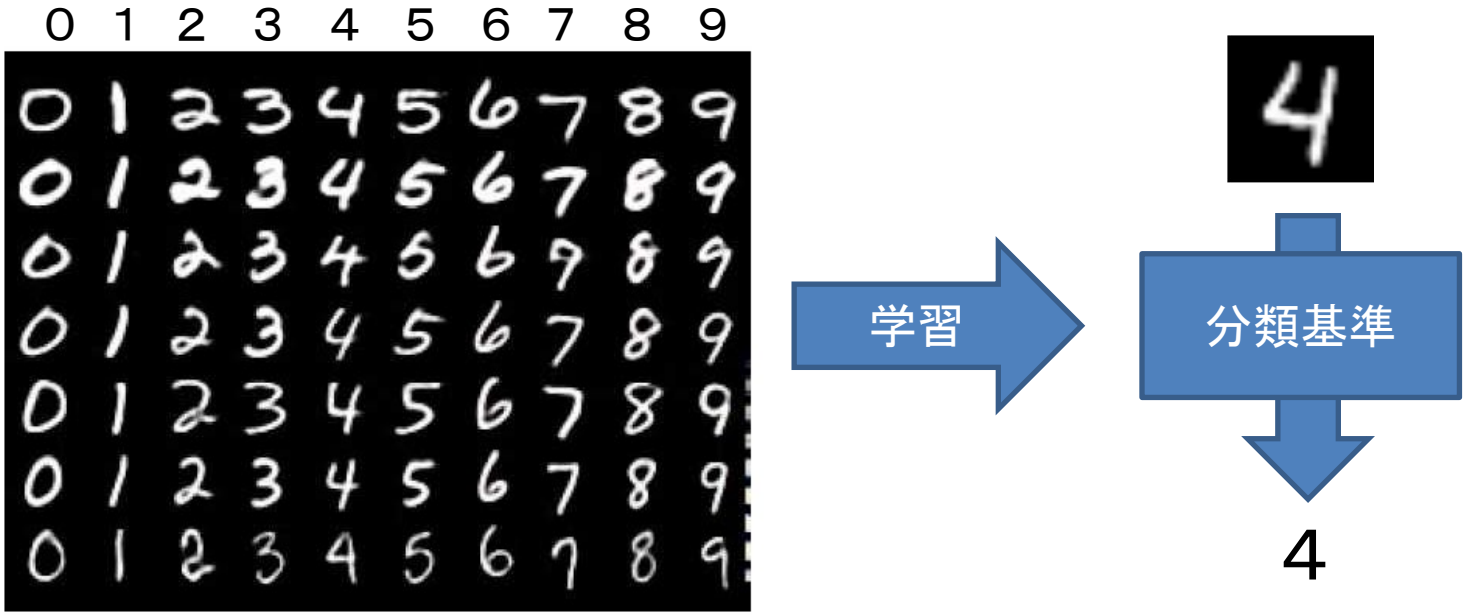
出展：元国立国会図書館長
長尾真先生スライドより引用

- 知識：情報を秩序化、体系化、抽象化し、他の知識との関係性を付けたもの。
- データや情報の解釈（理解）に必要なもの
- 知：知識を超えた、慣習や善・徳に支えられた判断をともなう何ものか（？）

出展：元国立国会図書館長
長尾真先生スライドより引用

# 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



# 機械学習と知識獲得

- 「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 |


人間が分類する場合

©いらすとや

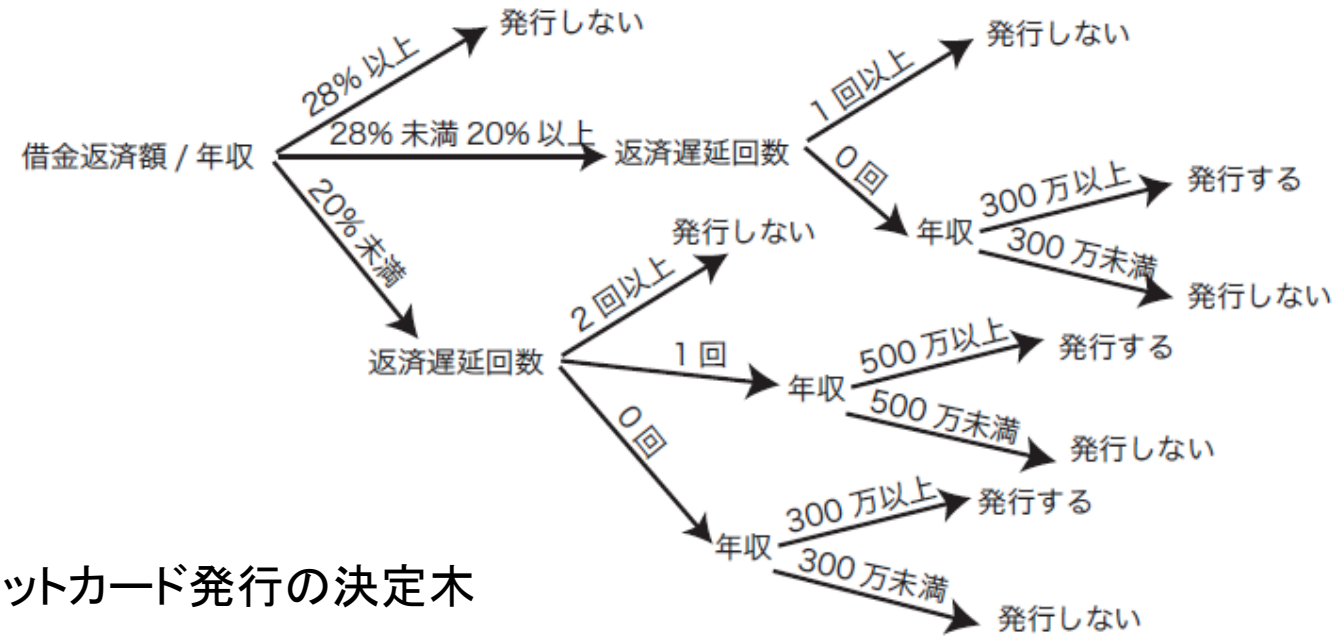
| 分類 | 書籍タイトル |
|---|---|
| ？？？ | 情報システム入門 |

| 分類 | 書籍タイトル |
|---|---|
| 情報科学 | 情報システム入門 |

# 色々な木 – 決定木

- 決定木
  - 枝に条件判断が書いてあり，その結果に従っていくと何らかの判断ができる木




クレジットカード発行の決定木

# 機械学習の処理

- 「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 |

学習

**分類基準**:
「情報」という単語が入っていれば分類は「情報科学」

©いらすとや

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 |

# 機械学習の例

- 「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：画像認識

猫

Photo by 松岡明芳,from Wikipedia
CC BY-SA 4.0

©ばくたそ Photo by Gaku0318

犬

Photo by Stefan Bauer,from Wikipedia
CC BY-SA 2.5

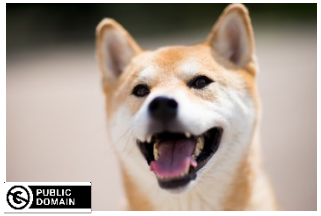
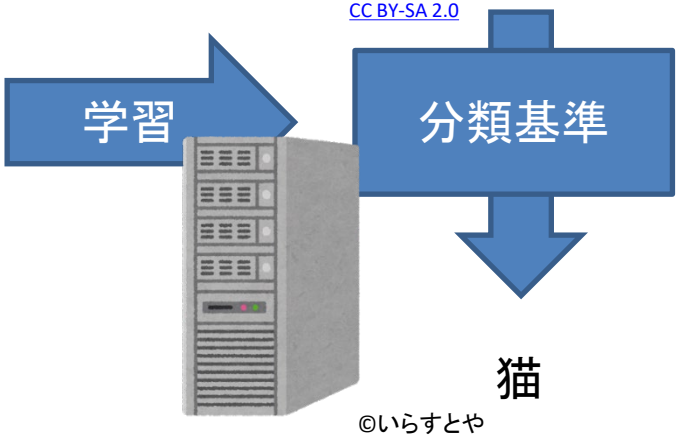


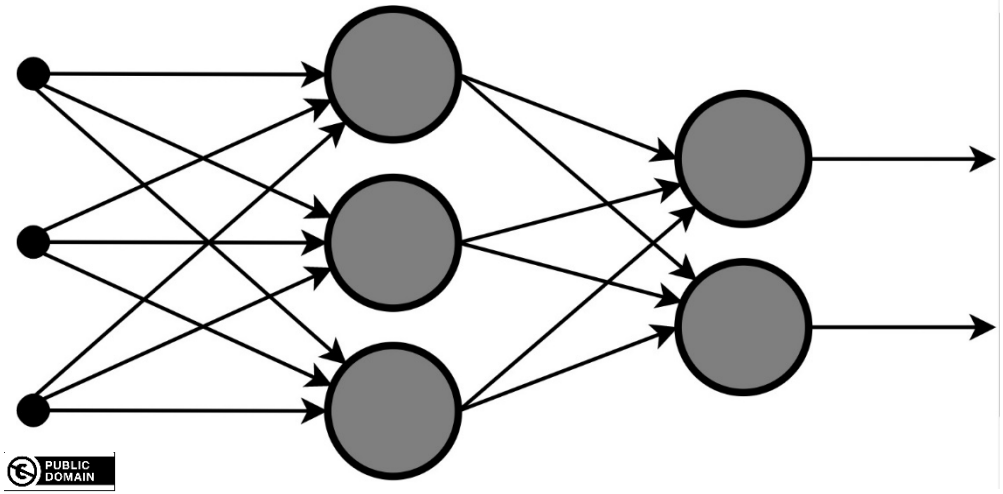

Photo by Hisashi ,from Wikipedia
CC BY-SA 2.0




©いらすとや

# ディープラーニング

- 機械学習手法の一つ
- 「分類基準として何を使うか」も自動的に学習




- 画像認識では人間を超える正解率

# 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

# 自然言語処理で何ができる？

- IBM Watson

著作権等の都合により 省略しました

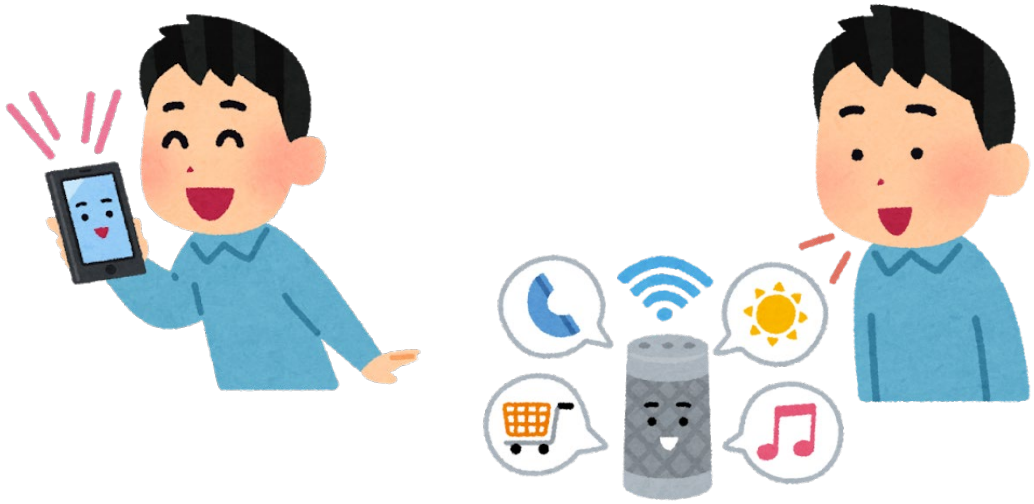
IBMワトソンの画像

クイズ番組で人間に勝利(2011)

# 自然言語処理で何ができる？

- Siri(Apple), Googleアシスタント, Alexa(Amazon)
  - スマートフォンやスマートスピーカーに話しかけ、会話、操作を行う
  - 人間の言葉を聞き、理解し、応答をする



イラスト©いらすとや

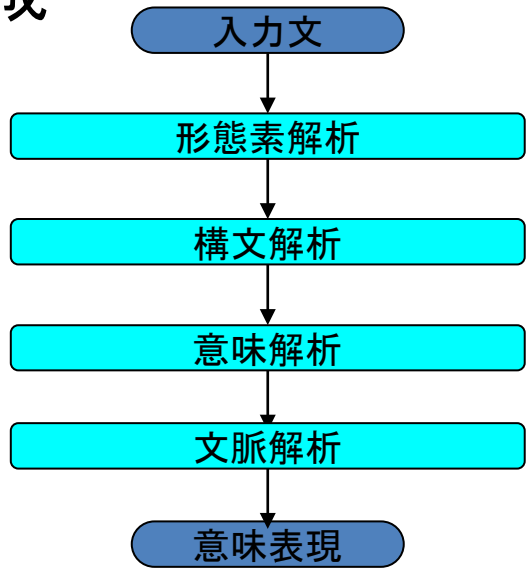
 

# 自然言語処理

# 自然言語処理(NLP)

- **計算機を用いて言語の理解を行う**
  - 形態素解析　－単語(形態素)に区切る
  - 構文解析　－語構成、文の構成(主語、述語等)
  - 意味解析　－意味表現の生成
  - 文脈解析　－文脈の理解

- **アプリケーション**
  - 変換系
  - 探す系
  - 分析系
  - 上記の統合系




計算機の発展 → 大量のテキストを高速に処理

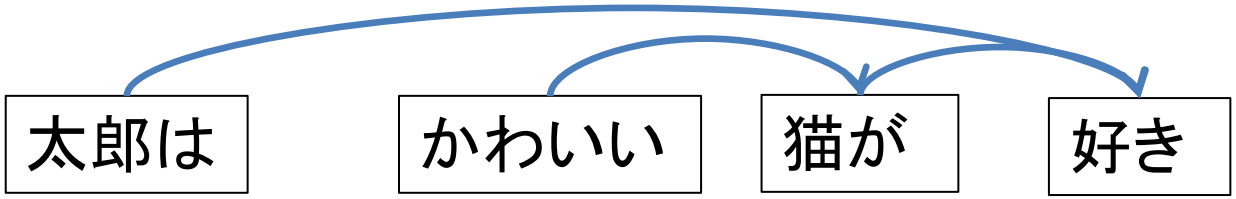
# 自然言語処理の基礎技術

太郎はかわいい猫が好き

単語に分割　**形態素解析**

太郎　は　かわいい　猫　が　好き

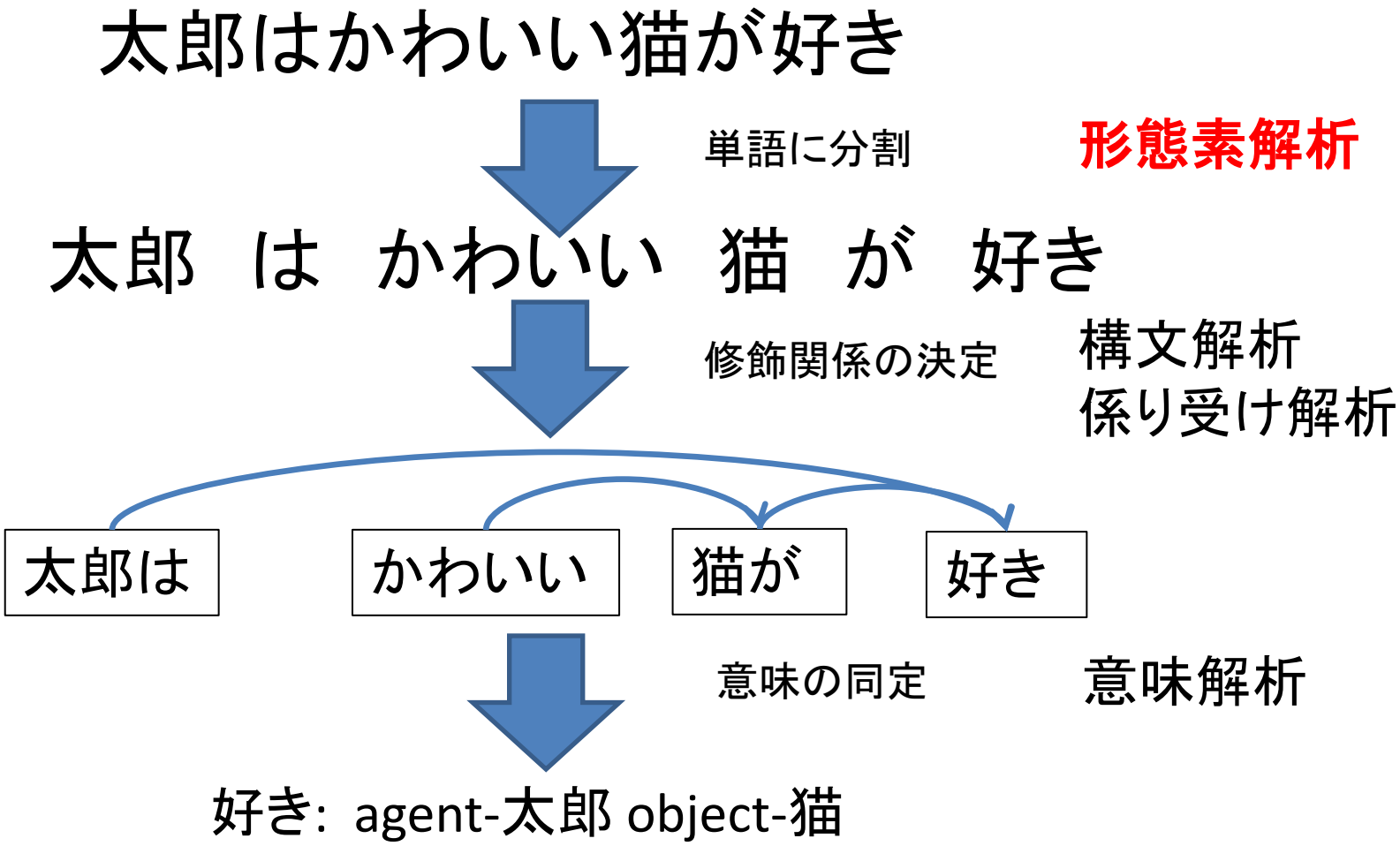
修飾関係の決定　**構文解析**
**係り受け解析**




意味の同定　**意味解析**

好き: agent-太郎 object-猫

# 自然言語処理の基礎技術



# 形態素解析

- 文を形態素(単語)に分割し、品詞などの属性情報を同定する

例: 構造改革を推進する

| 表層 | 品詞 | 読み |
|---|---|---|
| 改革 | 名詞-サ変接続 | カイカク |
| 構造 | 名詞-一般 | コウゾウ |
| 推進 | 名詞-サ変接続 | スイシン |
| する | 動詞-自立 | スル |
| を | 助詞-書く助詞 | ヲ |
| ⋮ | ⋮ | ⋮ |

構造／改革／を／推進／する

名詞／名詞／助詞／サ変名詞／サ変動詞

# 形態素解析

- 文を形態素(単語)に分割し、品詞などの属性情報を同定する

  例: この先生きのこるには

  × この／先生／きのこる／に／は

  「きのこる」という単語は辞書にない

  ◯ この／先／生き／のこる／に／は

  連体詞／名詞／動詞／動詞／助詞／助詞

# 形態素解析演習

- http://chamame.ninjal.ac.jp/
- または「形態素解析　茶まめ」で検索
- 青空文庫https://www.aozora.gr.jp/からテキストを選ぶ
- 課題１：形態素解析を行い、区切りや品詞の誤りを見つけ、どうすれば解決するかを考察する
- 課題２：CSV形式で出力し、エクセルやR等で開いたあと、名詞の頻度を集計し、上位１０件を抽出
- ITC-LMSより課題提出

# エクセルでのデータ集計

- 「データ」→「フィルター」を利用しデータをフィルタリング
  - “名詞”のみのデータを作成
- 「データ」→「ピボットテーブル」によりデータを集計する