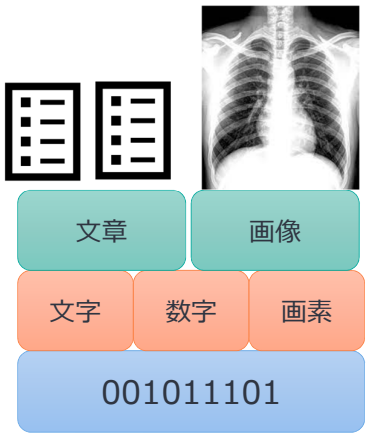
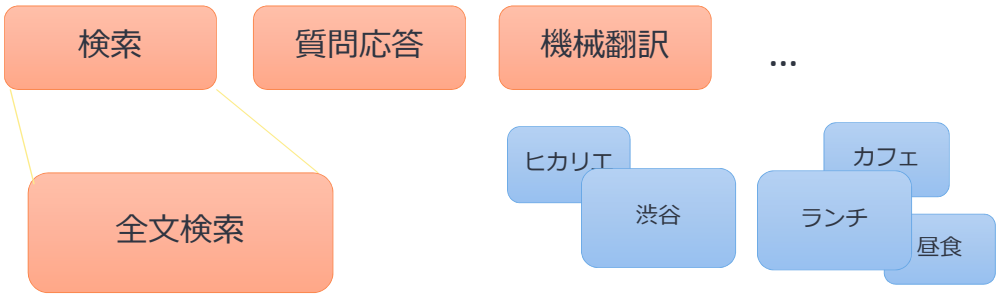
自然言語処理

- 人間が書いた文章をコンピュータで処理、解釈する技術
- 主な利用例
  - 目的の「言葉」が現れる文章を「検索する」
  - 「意味的に類似」している文章を探したり、文章群を分類したりする
  - 翻訳 (Google翻訳など)
  - 質問応答 (Siri, Pepper)
  - SNSデータの解析 (ブランドモニタリング、金融取引)
  - …
- 構造データ：コンピュータに処理しやすいように「設計」されたデータ
- 非構造データ：現実世界で発生した事象をそのまま記録したデータ
  - 音、画像、文章 (≒ 自然言語)
  - 近年、非構造データの利活用が大きなテーマ



単なる検索であっても…

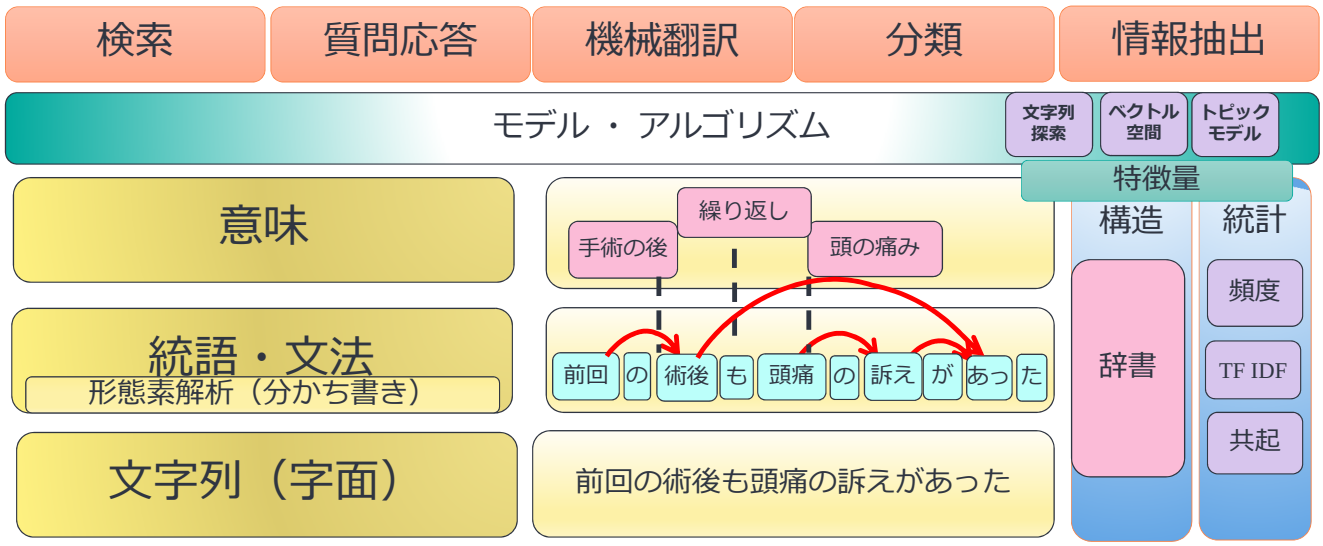


- 現在のコンテキスト・状況、意味や関連　曖昧さ、間違いを許容
- 理想的には、関連度の順に並んで欲しい
- 見落とし（ヒットすべきものがヒットしない）とノイズ（いらない情報がヒット）とのバランス

自然言語処理の活用

- 経過記録、看護記録、退院時サマリ、各種レポートを”活用”できていますか？
- 活用の技術的難易度が高かった – なぜ難しいのか
  - コンピュータ技術の特性: 構造データを計算したり検索したりが得意
    - 構造データ「ある期間に「尿酸」の値が8.0以上の患者を抽出」→ 可能
    - 非構造データ「頭痛の訴えがある患者をカルテ（記事）から抽出」→ ？
    - 「頭痛あり ✓」のような構造化は可能（テンプレート入力）→ 自由度は失われる
  - 自然言語に内在する困難さ
    「昨日本屋で買った本を喫茶店で読んだ」→本を買ったのはいつ？ 本を読んだのはいつ？
- 機械学習、AIなどの技術的革新により、活用への道がひらけてきた
  - 大量のデータが収集可能になった
  - CPU / GPUなど計算速度の向上
  - 統計、最適化、シミュレーションなど数学的ツールの発展

自然言語処理 概念図



自然言語処理の課題 ～ InterSystems NLPのアプローチ

- 自然言語処理(構造解析アプローチ)の難しさ
  - 「特定の語句や文字が含まれるか」が分かるだけでは、「意味」が分かることにはならない
  - 「頭痛」「頭が痛い」「あたまが痛い」などの揺れ、多義語、同義語、類義語の存在
    - 辞書を作るにしても、「最新情報」に保つのが困難
  - 誤字脱字、新しい言葉、「辞書」にない言葉（専門用語など）
  - 構文の曖昧さ
  - 100%正確な把握は不可能
    - どんな問題でも完全に解決できるような技術は存在しない！
    - 数値化 → モデルを仮定して、統計的なアプローチ → 機械学習・AI

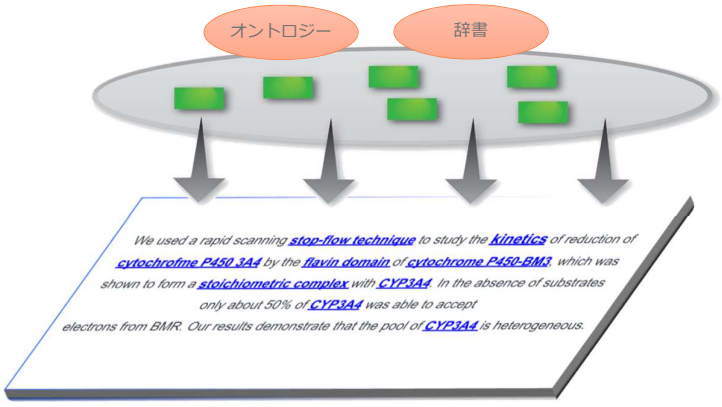
- InterSystemsがまず考えたこと →「単語」に区切ることは必須か
  - そもそも「単語」とは？
  - 「循環器」→ 循環 - 器？
  - 「単語」の境界 ≠「意味」の境界
  - 「単語」の境界にこだわるのではなく、「意味」の境界を特定してはどうか

単語ベースのテキスト分析: トップダウン

- 典型的なモデル： “Bag-of-words” モデル
- オントロジーや辞書がベース
  → 初出語、複合語への対応
  → メンテナンスが必要

※ オントロジー：概念体系

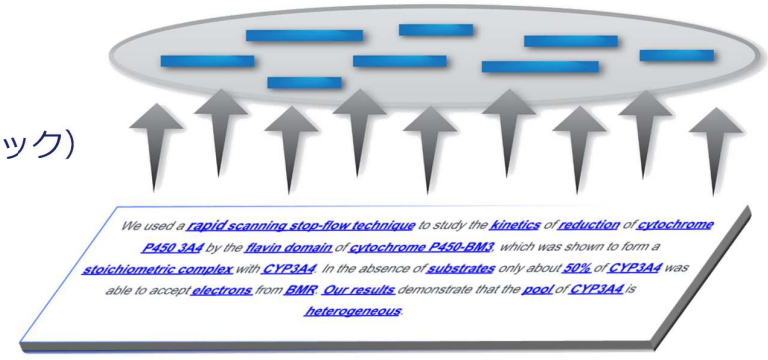
※ “Bag-of-words”：文の構造に関係なく、単語の集合を解析対象にする手法



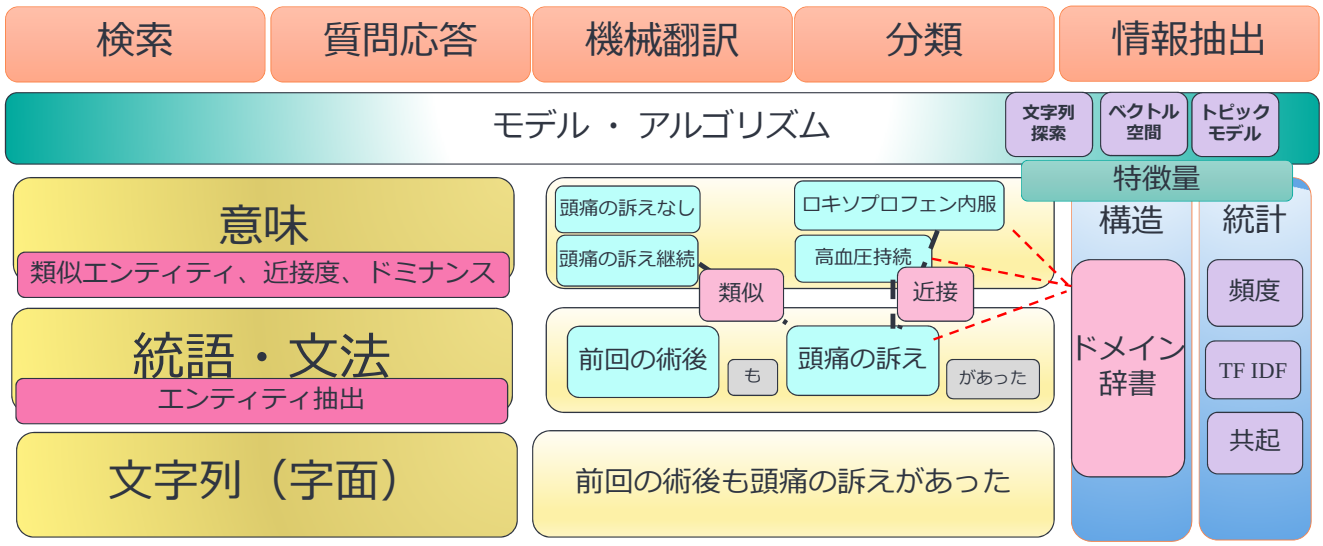
InterSystems NLP： “コンセプトレベル”からのボトムアップ

InterSystems NLPは、文の構造そのものから、
意味を持つ文字列の連なり（エンティティ）を抽出する

未定義語や新たな複合語、言い回しを発見
→ オープン・エンド
→ 発見的探索
→ 知識
（オントロジー・辞書へフィードバック）



InterSystems NLP による自然言語処理 概念図



例： InterSystems NLPと形態素解析の比較

既往にｱﾙﾂﾊｲﾏｰ型 認知症が あり、指示動作は入らず、見当識障害認めている。

形態素解析)　**既往 - に - ｱﾙﾂﾊｲﾏｰ - 型 - 認知 - 症 - が - あり - 、- 指示 - 動作 - は - 入ら - ず - 、- 見当 - 識 - 障害 - 認め -て - いる - 。**

IS NLP)　既往にｱﾙﾂﾊｲﾏｰ型認知症があり、指示動作は入らず、見当識障害認めている。

転倒転落危険度Ⅲであり、活動性の向上も認めるため、転倒予防と所在確認のために、うーご君常時使用、 ３点柵＋オーバーテーブル固定、入り口に赤外線センサー使用にて対応中。

形態素解析)　**転倒 - 転落 - 危険 - 度 - Ⅲ - で - あり - 、- 活動 - 性 - の - 向上 - も - 認める - ため - 、- 転倒 - 予防 - と - 所在 - 確認 - の - ため - に - 、- う - ー - ご - 君 - 常時 - 使用 - 、- ３ - 点 - 柵 - ＋ - オーバー - テーブル - 固定 - 、- 入り口 - に - 赤外線 - センサー - 使用 - にて - 対応 - 中 - 。**

IS NLP)　転倒転落危険度Ⅲであり、活動性の向上も認めるため、転倒予防と所在確認のために、うーご君常時使用、 ３点柵＋オーバーテーブル固定、入り口に赤外線センサー使用にて対応中。

InterSystems NLPが標準で計算する数値

- 頻度（Frequency） ： エンティティが何回出現するか
- スプレッド(Spread) ： エンティティが出現するソースの数
- ドミナンス(Dominance)： エンティティの意味的な重要度
- 近接度(Proximity)：2つのエンティティの関連度

類似エンティティの例

エンティティ 1	エンティティ 2	コサイン類似度	類似の種類
粉塵暴露歴	粉塵曝露歴	0.937	誤字
脂質異常症	高脂血症	0.933	言い換え
喘息	気管支喘息	0.815	言い換え
尿検査	検尿	0.937	言い換え
呼吸困難感	呼吸苦	0.909	言い換え
特記事項なし	特記事項無し	0.986	表記の揺れ
著変ありません	著変なし	0.926	表記の揺れ
二人暮らし	2 人暮らし	0.934	表記の揺れ
術後の経過	術後経過	0.974	表記の揺れ
PET−CT	PET/CT	0.697	表記の揺れ
大腸癌	肝臓癌	0.937	関連語
主婦	会社員	0.870	関連語
当科紹介	当院紹介	0.889	関連語


The power behind what matters.
InterSystems
Health | Business | Government
Thank you.