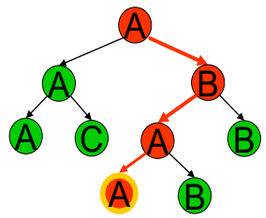
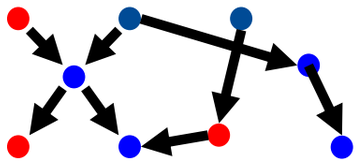
研究テーマ（有村）

- 北大大学院情報科学研究科，教授
- 専門：
  - データマイニング
  - 情報検索（とくに全文テキスト索引）
  - 計算学習理論（機械学習）

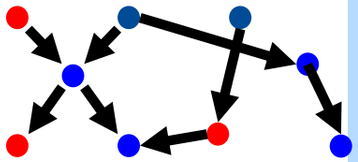




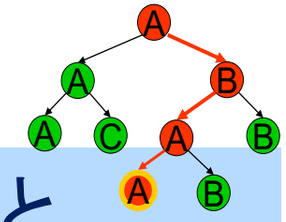
- 興味があること
  - 膨大なデータから，人間に役立つ情報と知識をとりだすこと
  - 高速なアルゴリズム（プログラム）を設計するこ
- 趣味
  - 読書・音楽鑑賞・ハイキング・家庭菜園の手伝い

Daisetsuzan, H. Arimura

- 情報検索(「文字列」-)
  - 文字列, 木, グラフ, 集合族, . . .
  - 検索・比較・発見・圧縮
  - アルゴリズムとデータ構造



- データマイニング
  - 列挙アルゴリズム: 多数の解をもれなく計算する
  - 構造・パターンマイニング
- 機械学習(計算学習理論)
  - 信頼されるAI(公平性・説明性)
  - 離散機械学習(決定木ほあk)
  - 効率良い学習アルゴリズム



やりたいこと

- 膨大なデータから, 人間に役立つ情報と知識をとりだす
- 高速なアルゴリズム(プログラム)を設計する

対象

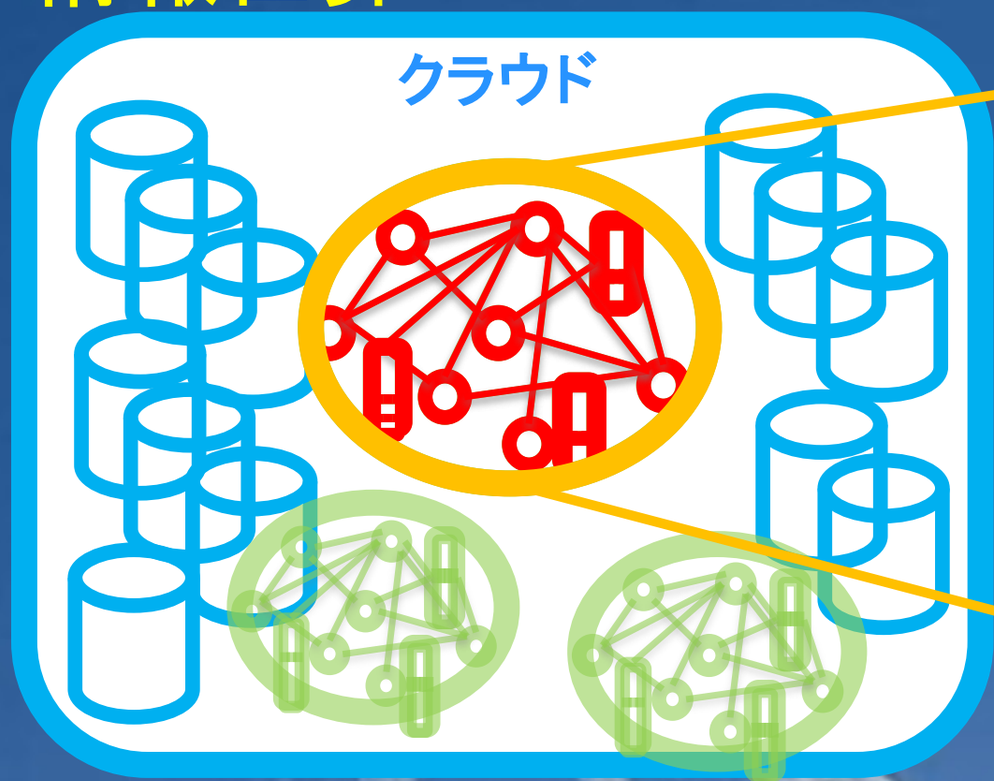
- 文字列・木・グラフ
  - 例: テキスト・遺伝情報, etc.
- 空間データ
  - 例: 地理・移動・音楽・??? データ

この研究室(IKN研)

- どちらかというと基礎
  - アルゴリズムが中心(道具)
  - 情報は「形」「道具」の学問
  - 先生方は二つの専門を持っている
    - 道具(How):
    - 対象(What):Web, Networks, 生命科学,
      機械学習, 検索, データ解析. . .
  - 現代は, いろいろなタイプの人が一緒にはたらくのがふつう:基礎・応用・人間/社会
- こんな人
  - 考えるのが好き and/or 手を動かす

21世紀の情報環境

情報世界



実世界

- 大量のデータ
- 多数のCPU
- 高速なネットワーク
- 膨大な計算

「集中」

- さまざまなデバイス
- 多様な人間活動と応用
- 多様で非均一な時空間
- 不完全で複雑なデータと情報

「分散」

大量データと次世代情報技術

ワトソン君

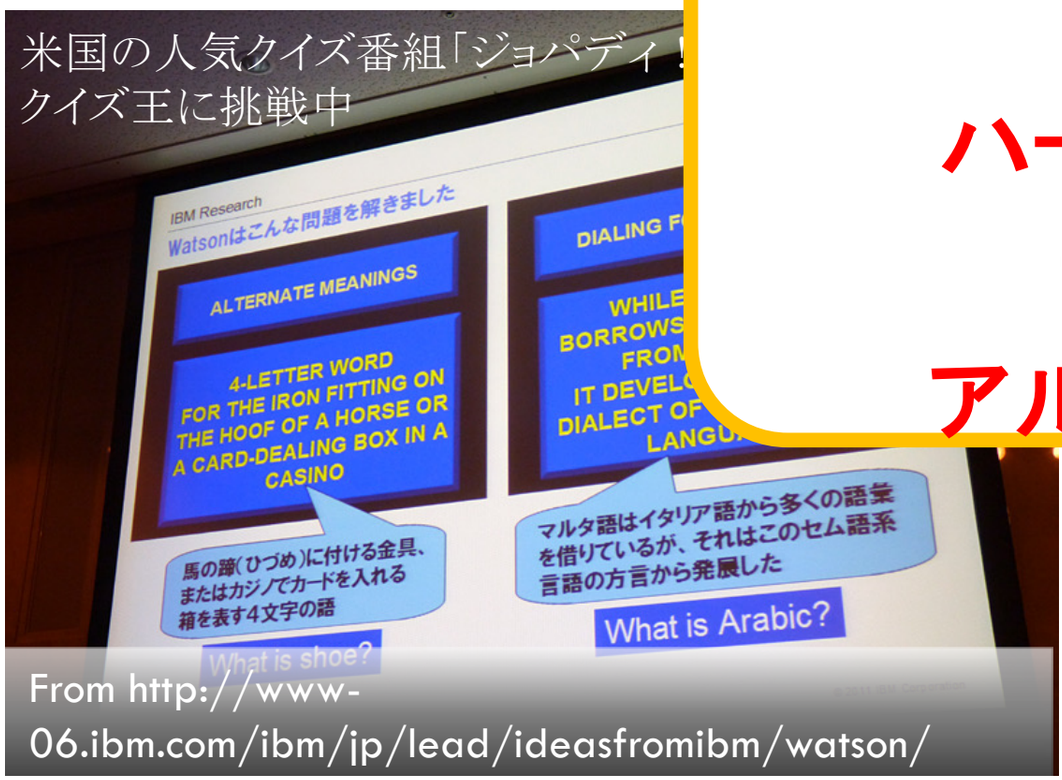
- IBMリサーチ (2011/02/16)
- クイズ番組で人間に勝利！
- １００万冊の本を読んで回答
- 人工知能と自然言語，アルゴリズム，検索の技

クラウド計算

世界中の情報を計算！

次世代IT（AI）の原動力

ハードウェア!
データ!
アルゴリズム!



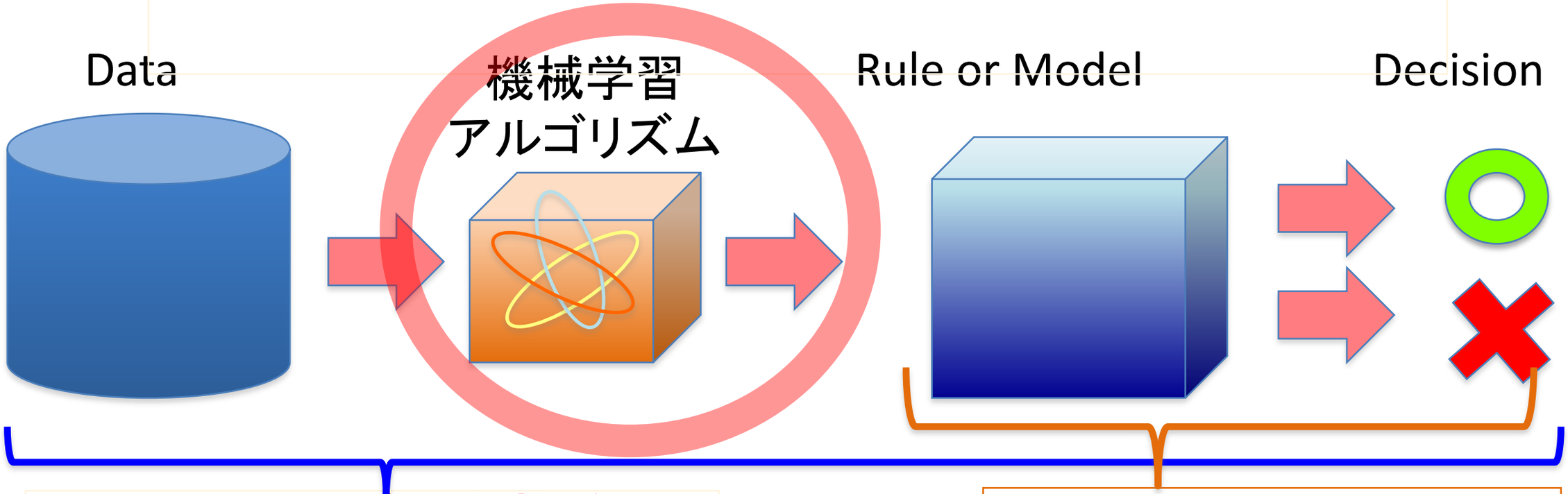
米国の人気クイズ番組「ジョパディ！」クイズ王に挑戦中

From http://www-06.ibm.com/ibm/jp/lead/ideasfromibm/watson/

From geek.com: Google server firm http://www.geek.com/articles/chips/up-next-for-google-enterprise-wars-2009078/

現代の人工知能

データから規則を学習する**機械学習技術**が中核技術になっている



現代のAI（データ指向AI）はデータに基づく意思決定のすべての局面をあつかう

伝統的なAI は規則を使った意思決定だけをあつかう

Modification of: Sara Hajian, Francesco Bonchi, Carlos Castillo, "Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining", KDD 2016, Tutorial

次世代の発見科学技術

* 最新の人工知能技術
* 21世紀はDeep Learningだけで

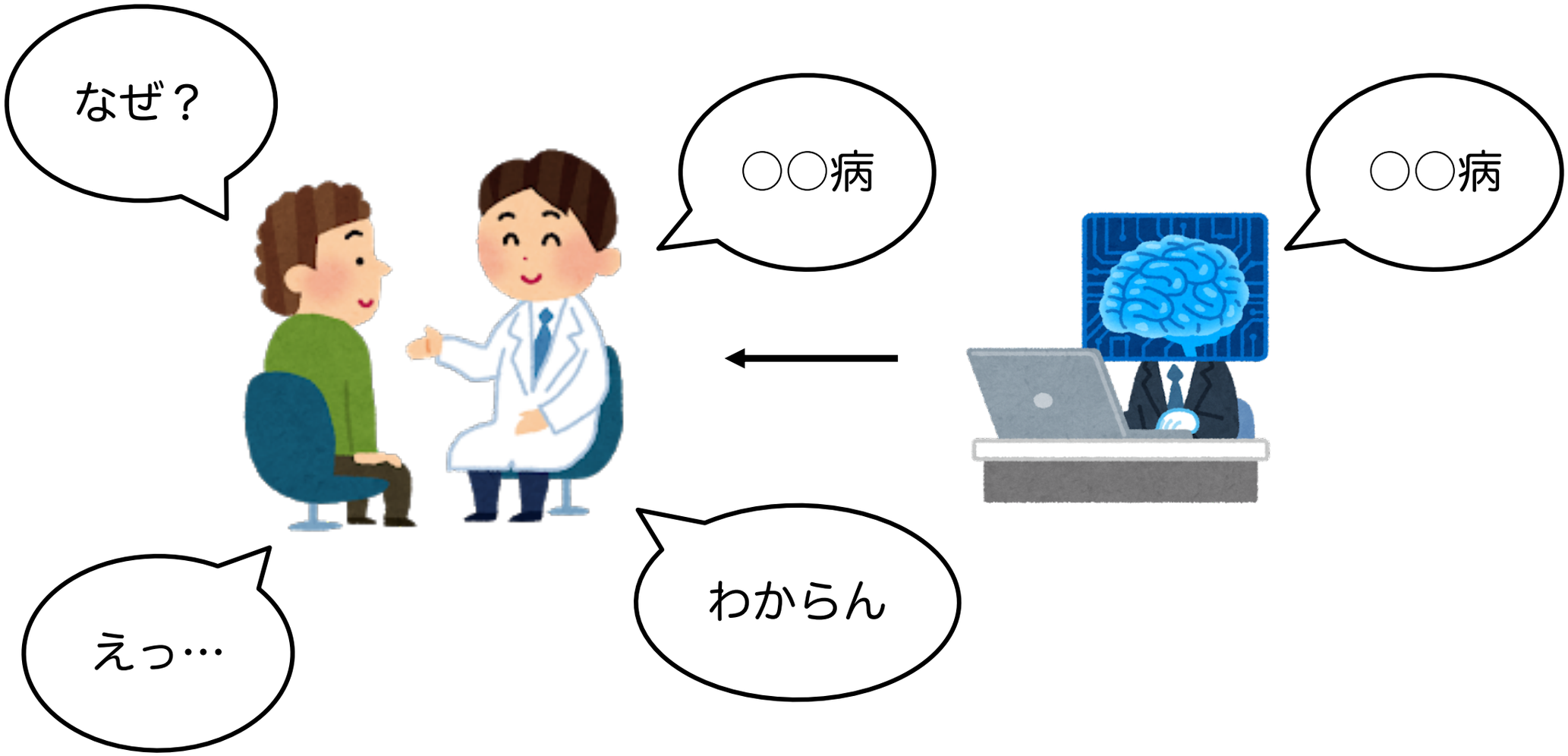
有村君、21世紀の知識獲得はDeep Learningで決まりかね？

- 21世紀の社会で、コンピュータが人間と働くために必要な人工知能技術とは何か？

有川節夫先生
（私の先生）

KSSの研究 ～その１～

信頼されるAIの基盤技術



どうやって AI の導きだす回答を
信頼させられるものにするか？

研究1：予測モデルに対する反事実説明
(Kanamori+, IJCAI2020)

Accepted at IJCAI 2020

DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization

○金森 憲太朗　（北海道大学）
高木 拓也　（富士通研究所 / 理研AIP）
小林 健　（富士通研究所 / 理研AIP / 東京工業大学）
有村 博紀　（北海道大学）

Take-home message: 機械学習と説明可能性

- 機械学習の目的の一つは, 入力空間 $\mathcal{X}$ と出力空間 $\mathcal{Y}$ を上手に対応づける関数 (**予測モデル**) $h : \mathcal{X} \to \mathcal{Y}$ を見つけること.
- **実応用上の要求**: 予測の**説明可能性** (解釈可能性)
  - **知識発見**: モデルの予測結果や傾向から, **新たな知見や仮説を見つけたい**.
  - **信頼性**: 医療や司法などの意思決定に応用するには, **予測根拠の提示が必要**.

- 予測結果の説明 (のようなもの) を抽出する研究が注目されている.
  - 例) **特徴量重要度**: 各特徴量が予測にどのくらい貢献したかを示す指標.
  - いまのところ説明可能性 (解釈可能性) にちゃんとした定義は存在しない.

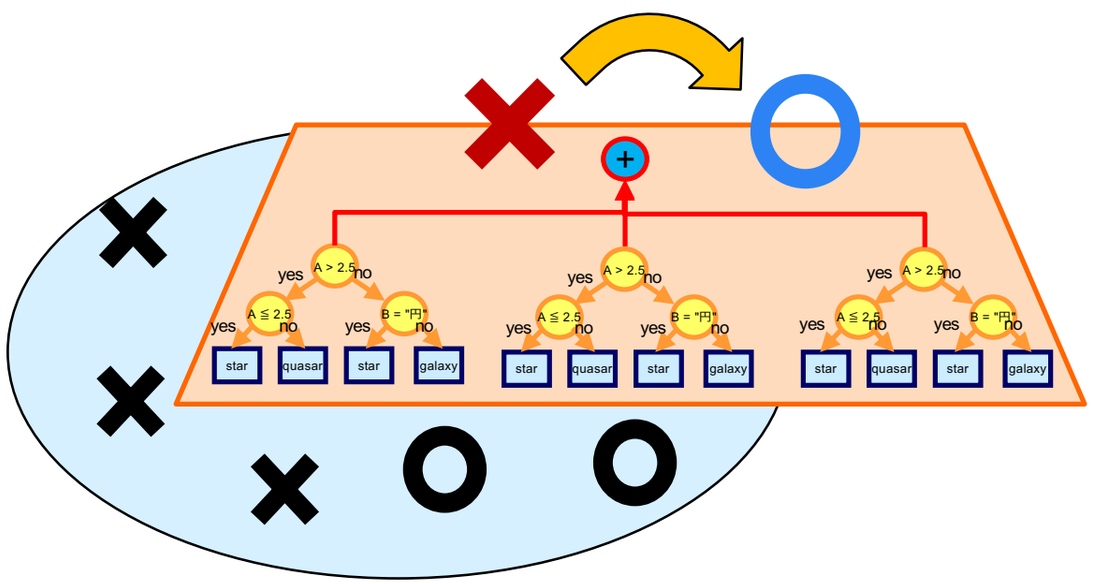
研究：人に優しい人工知能技術

K. Kanamori, T. Takagi, K. Kobayashi, H. Arimura, IJCAI-PRICAI 2020 (to appear)

機械学習の解釈性

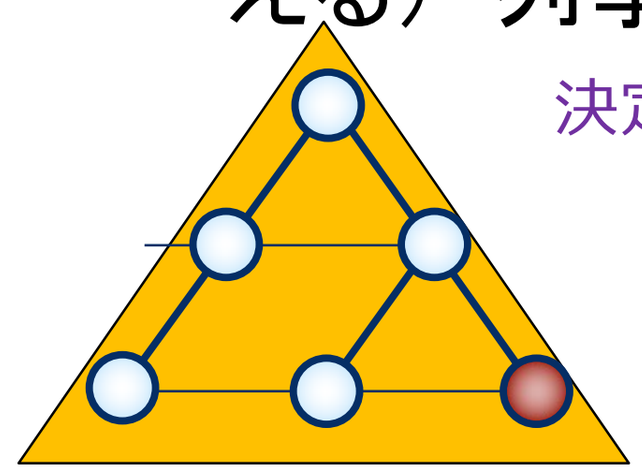
- 解釈性・説明性・公平性
- 反事実説明 (Kanamori+, IJCAI2020)

「予測を×から○へ改善するためには，状況をどう変るべきか？」



離散最適化に基づく学習

- 解釈しやすい離散的ルールを学習する
- トップK・計数（数を数える）・列挙・乱択



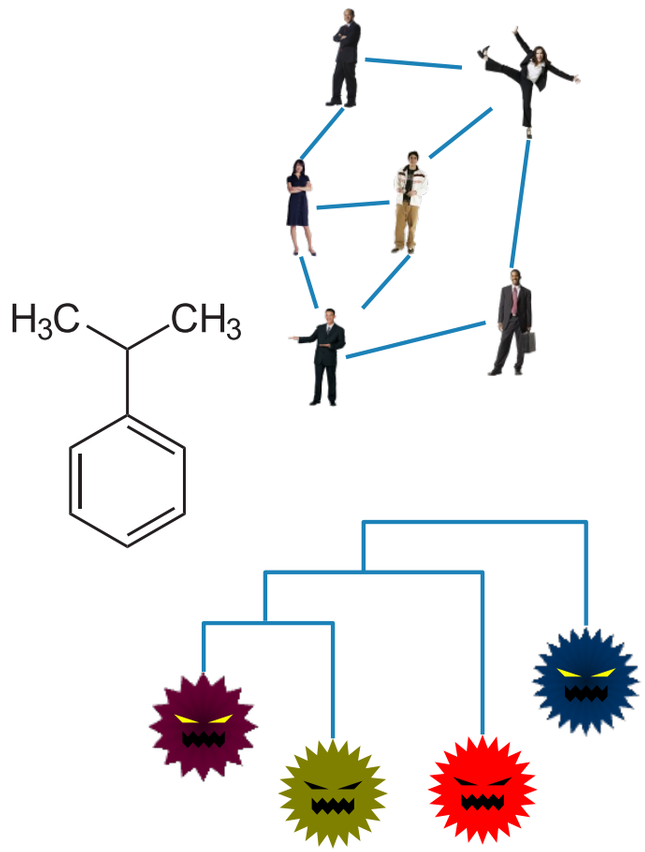
決定木を列挙!

制約
- サイズ
- 葉の例数
- 精度

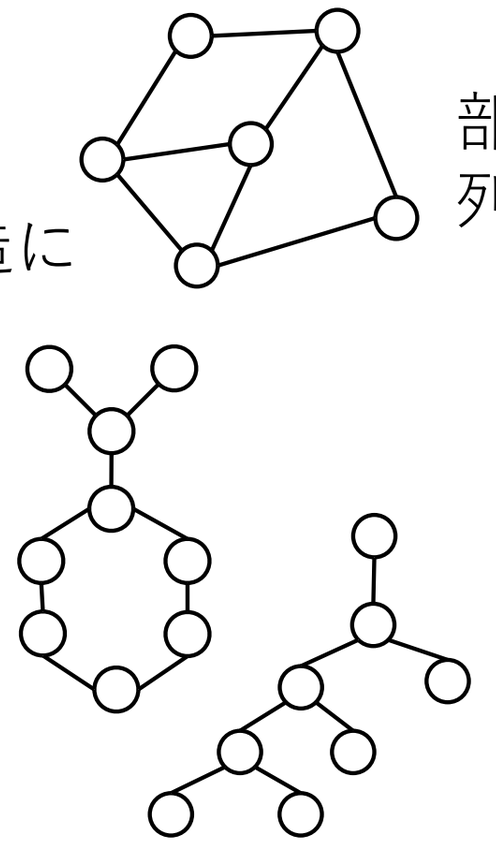
研究：グラフアルゴリズム & 列挙アルゴリズムによる知識発見

どんな構造が含まれているか，すべて挙げよ！

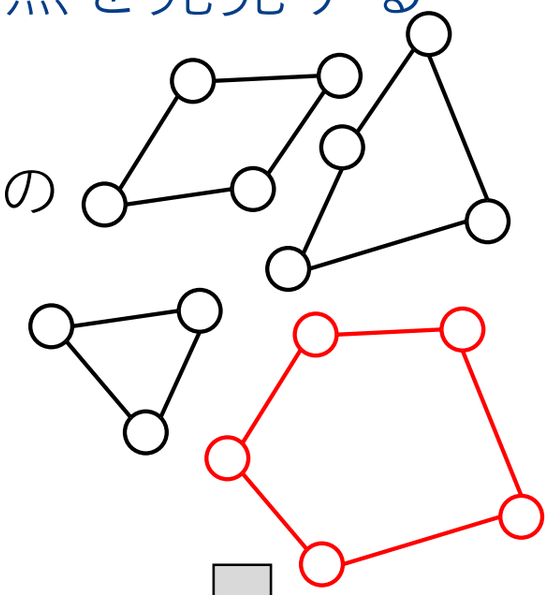
部分構造をすべてリストアップし，共通点や相違点を発見する



グラフ構造に抽象化



部分構造の列挙

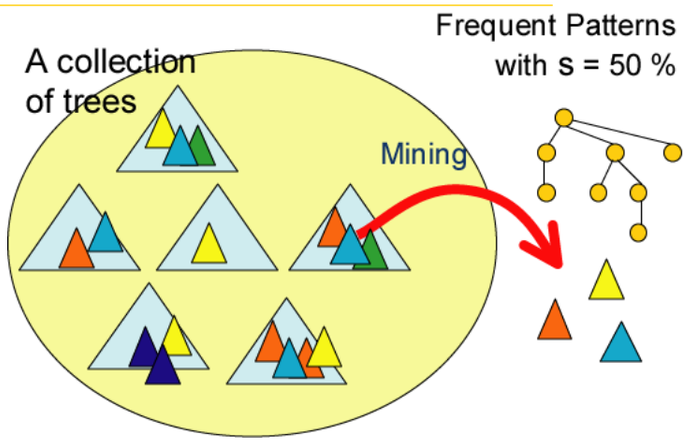


列挙した部分構造から知識発見

研究テーマ：半構造マイニング「最右拡張技法」(

高速半構造データマイニングエンジン

- Discovering all frequent sub-structures from a colection of labeled trees
- Extendible to most statistical functions



- **FREQT**: Efficient ordered tree mining engine (SIAM DM'02)

Applications
Japanese Text Mining (Morinaga, Arimura et

Rightmost Expansion & Ordered Tree Enumeration Trees (SIAM DM'02, IEEE ICDM'02, DS'03))

FREQT: DEWS'02優秀論文賞（H14年6月）

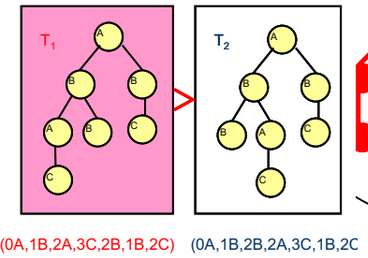
Unot: DEWS'04優秀論文賞（H16年6月）

「最右拡張技法」に基づく高速な木パターン発見エンジン

- 半構造データの特徴的部分構造の発見
- その後のグラフマイニングの基礎技術に
- 理論と実装：FREQT, StreamT, Unot
- SIAM DM'02, PKDD'02, IEEE ICDM'02), DS'03
- 公開＆応用

Google Scholar
引用数(2019.4)
[主要文献1] 596件

trees (Left-Biased Tree)
(Lexicographically largest trees over depth-label sequence encoding)



Difficulties in un-ordered tree mining

Exponentially m
isomorphic patte
■Unique represen
■Enumeration wit
duplicates in a un
■Efficient comput
occurrences

Awarded

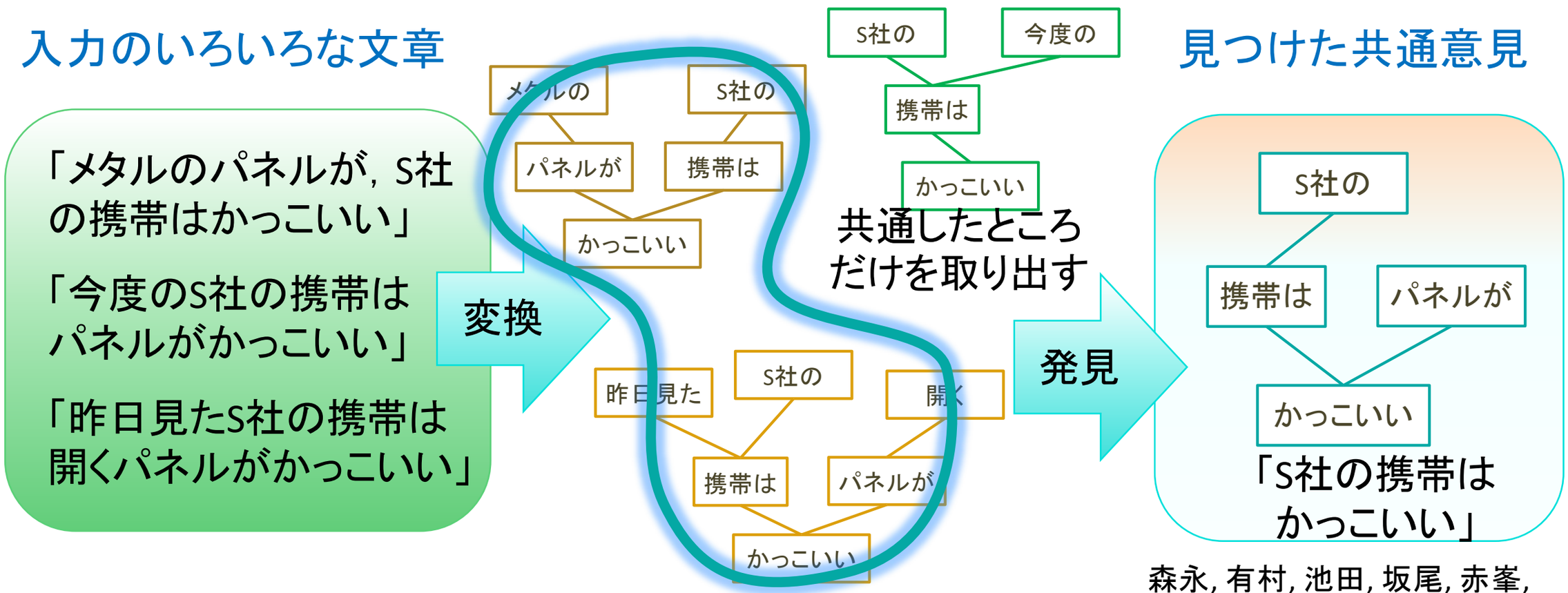
MaxMotif: Awarded DEWS2004（H17.6）

StreamT: '03AI学会大会優秀賞（H15.6）

コンピュータの学習の例

テキストマイニング：

* まず大量のテキスト（ブログでの評判や自由記述アンケート）を解析し，文の構造を「木」としてとりだす．
* 次に，共通する構造を発見（マイニング）すると，文書に隠れた共通の意見が取り出せる．



この仕組みは，企業でテキストデータの解析に使われた．

森永，有村，池田，坂尾，赤峯，FIT2004; ACM KDD2005.

人工知能技術の鍵！ = 機械学習技術

NIPS 2017（機械学習研究最大の学会）

- Thirty-first Conference on Neural Information Processing Systems, Dec. 4-9, Long Beach Convention Center
- Attracting over 8,000 registered attendees, up 2,000 from last year

Regist ration

exhibition booth

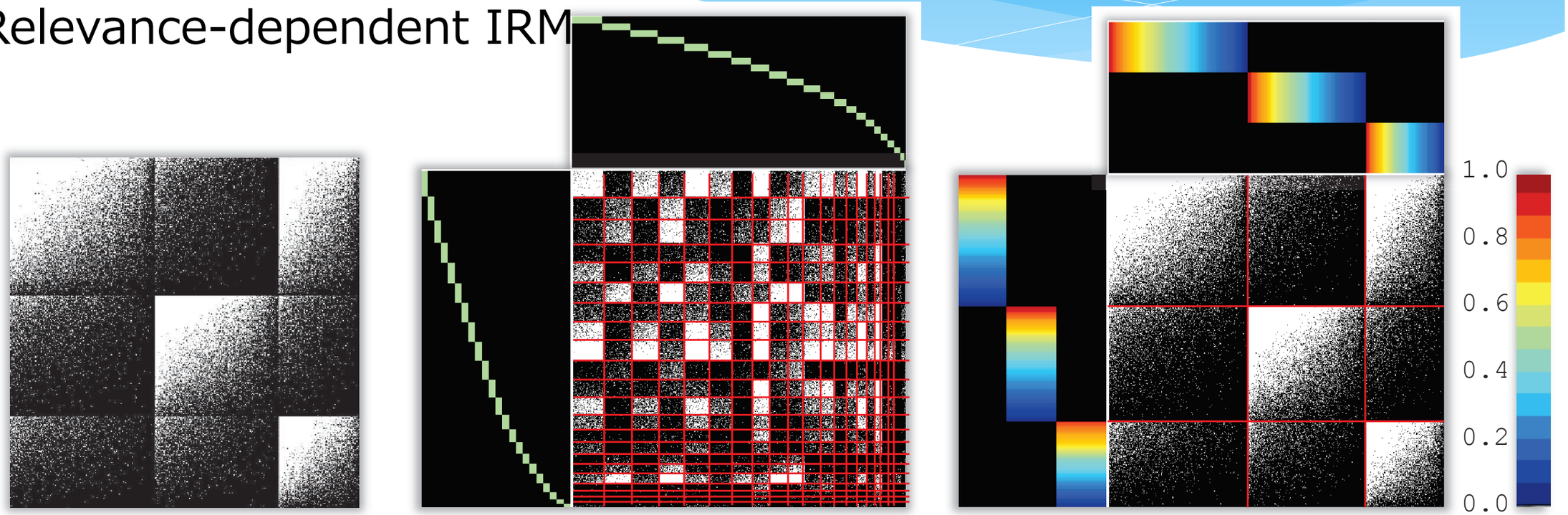
session

Poster site

研究：ベイズ学習・トピックモデル

(Ohama, Iida, kida, Arimura, IJCAI 2017, NIPS2017)

Relevance-dependent IRM



(a) Synthetic Data　(b) IRM solution　(c) R-IB solution

Figure 3: Synthetic example: (a) 500×500 relational data; (b) IRM solution; (c) R-IB solution. In (b) and (c), the left and top matrices indicate $\boldsymbol{z}_1$ and $\boldsymbol{z}_2^\top$ in a 1-of-K representation, respectively. Colored areas on matrices $\boldsymbol{z}_1$ and $\boldsymbol{z}_2^\top$ indicate relevance parameters for corresponding objects. For intuitive understanding, a relevance parameter $\theta \in [0, +\infty]$ is transformed into a probability as $1 - \exp(-\log(2) \times \theta)$, so that the probability is 0.5 when $\theta = 1$.

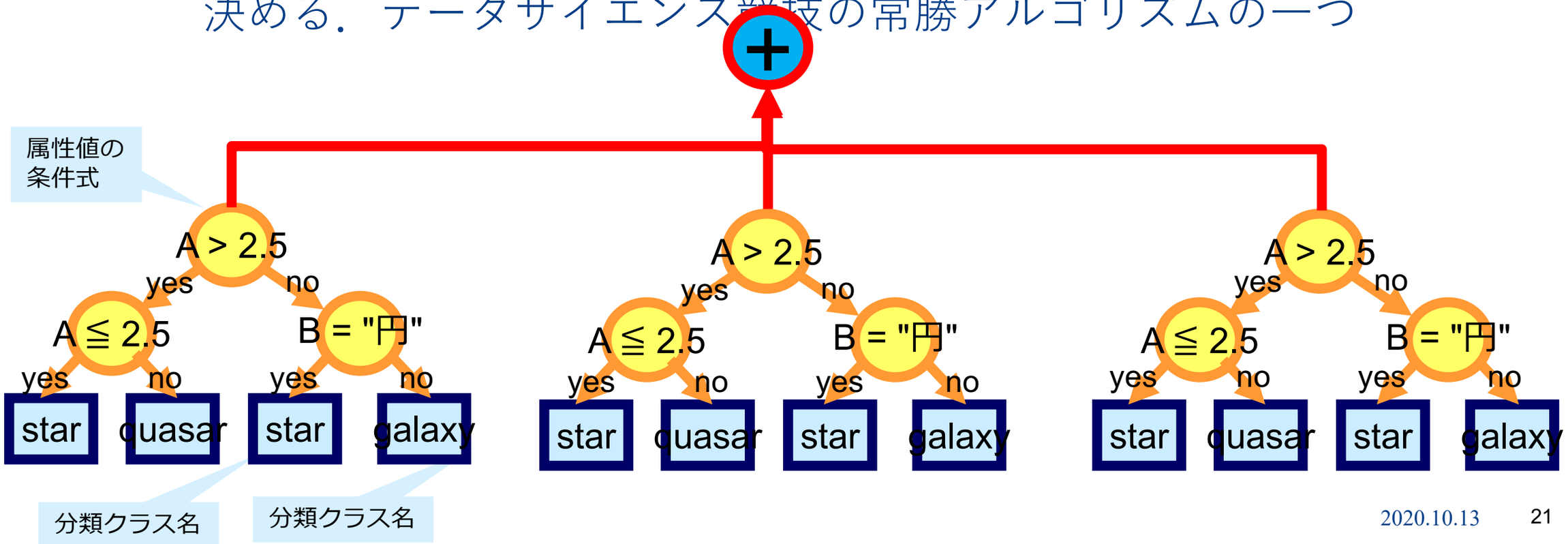
- 従来手法（IRM）は、個々の人や商品の「活動度」に影響されて本来存在しない、小さなクラスに分かれている。
- 提案手法 (R-IB) は、個々の活動度に影響されずに、大きな（正しい）クラスタ構造を見つけている

研究：機械学習の研究

* **決定木(decision tree)と木アンサンブル（勾配ブースティング，ランダムフォレスト）による自動分類**

▼決定木：対象を，その属性の値に従って分類する．基本は「すごろく」と同じ．条件式の値（YES/NO）にしたがって，根から葉までたどる．ゴールが分類クラス．

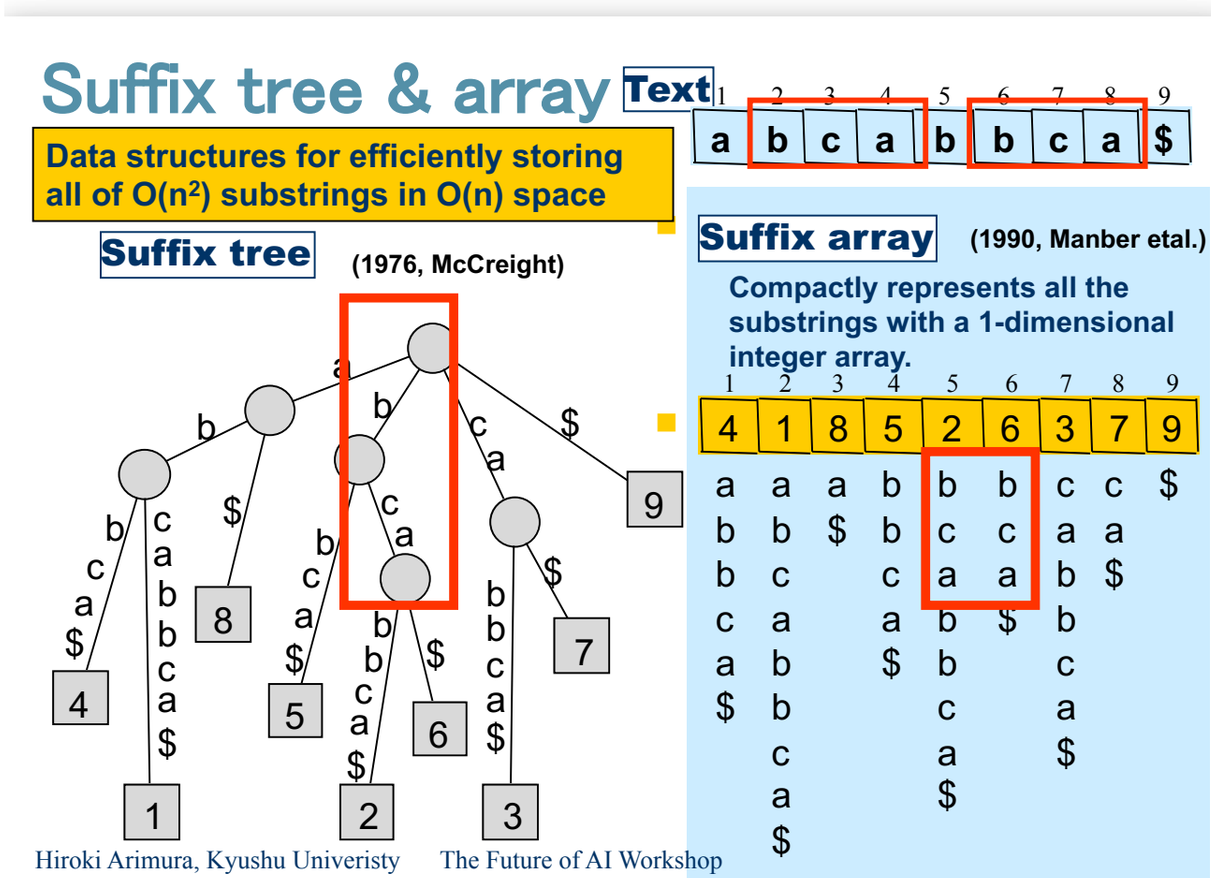
▼木アンサンブル：多数の決定木の予測を，多数決して全体の予測を決める．データサイエンス競技の常勝アルゴリズムの一つ

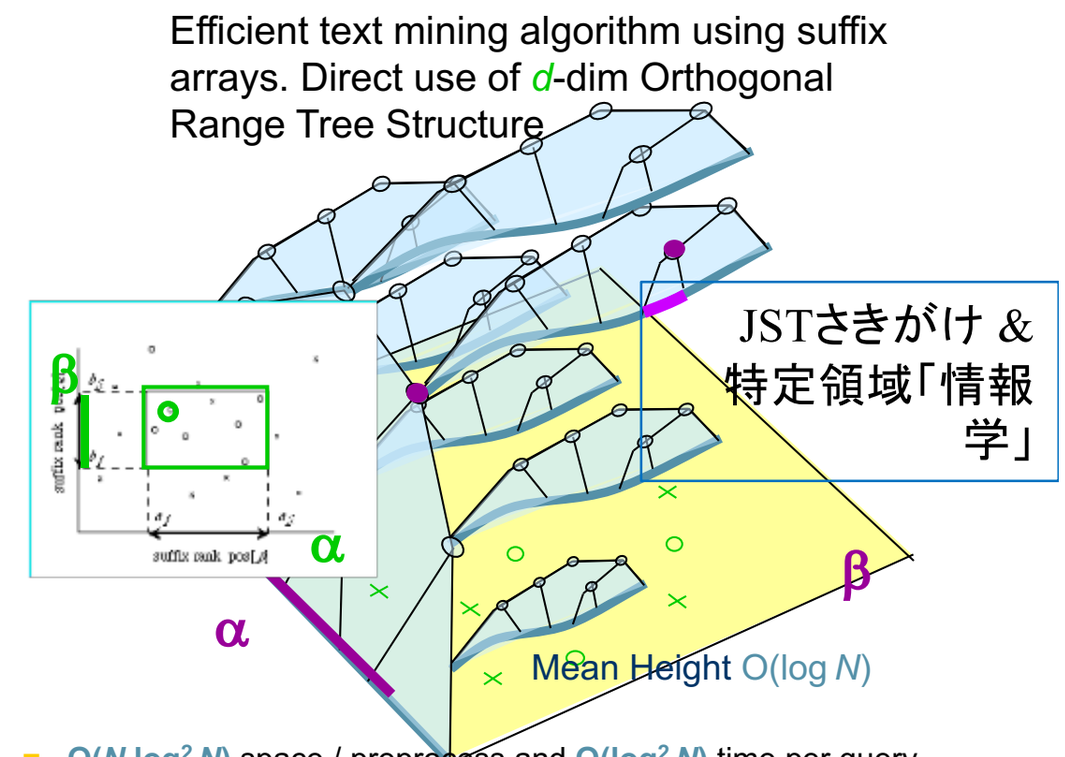


研究テーマ：ビッグデータの検索

- 世界初の線形LCP配列計算アルゴリズム (CPM2001)
- 研究分野ベスト10引用論文として表彰：2010年に分野トップ会議(CPM) での「接尾辞木データ構造40周年記念」集会で表彰
- CACM紙の「接尾辞木データ構造40周年記念」特集で言及された：「数十年以上の未解決問題を肯定的に解決し，接尾辞配列と接尾辞木の相互変換のMissing Linkをつないだ」.

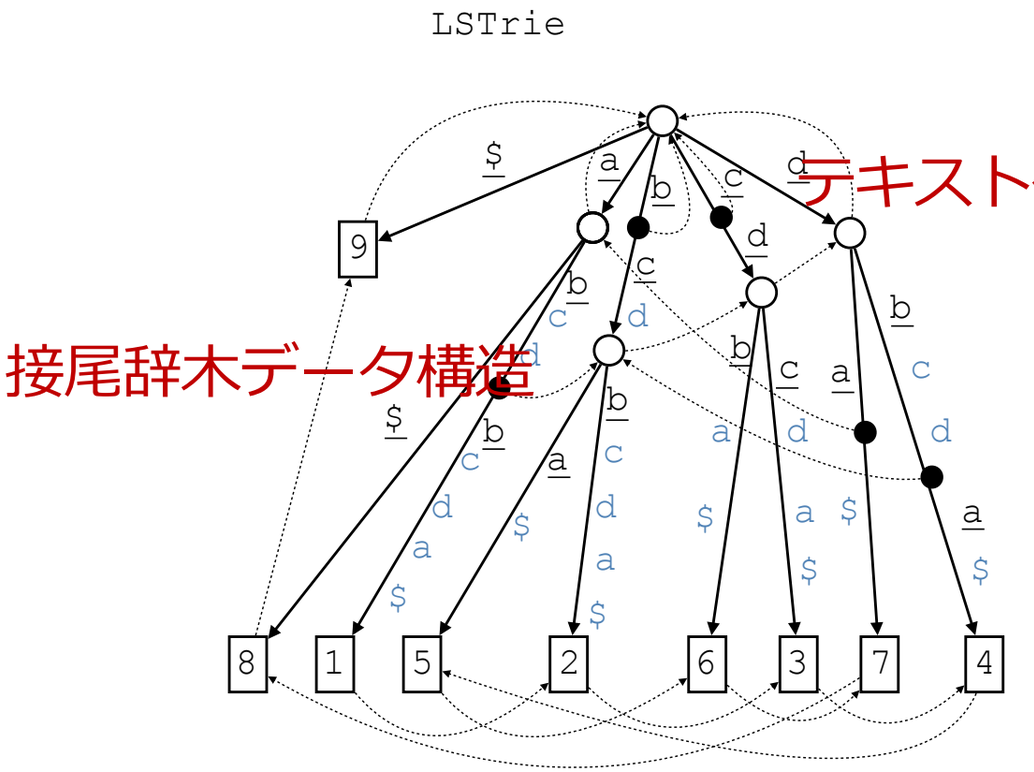
Google Scholar
引用数(2019.4)
[主要文献2] 535件

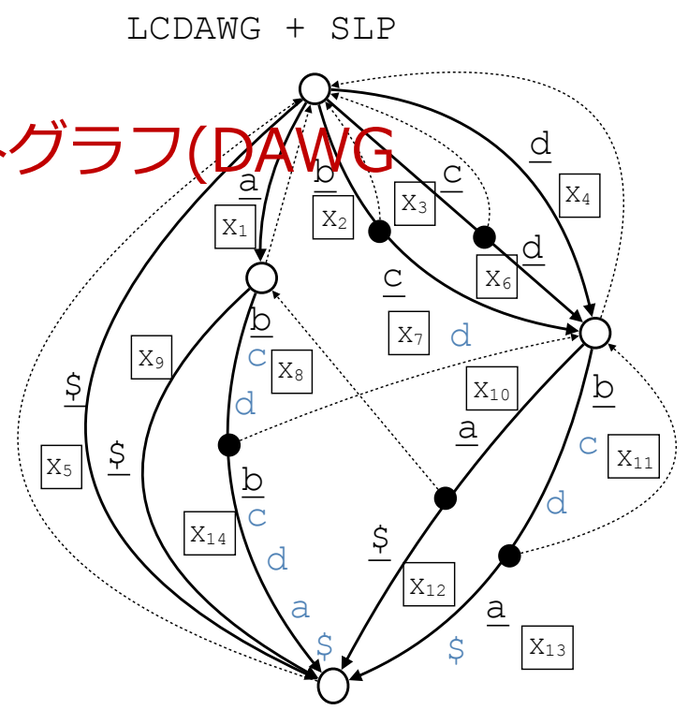




研究：文字列アルゴリズムとデータ構造

- **大規模テキスト・系列データを効率よく圧縮・検索・解析するためのアルゴリズムとデータ構造を研究する**
- 文字列・正規表現・木構造・NoSQLの検索アルゴリズム
- 接尾辞木と接尾辞配列，BWT，テキストグラフ(DAWG)（二分探索木や，ハッシュ表，領域検索，RMQなどのテキストへの拡張）





$X_1 = a$
$X_2 = b$
$X_3 = c$
$X_4 = d$
$X_5 = \$$
$X_6 = d$
$X_7 = X_3X_6$
$X_8 = X_2X_7$
$X_9 = \$$
$X_{10} = a$
$X_{11} = X_2X_7$
$X_{12} = \$$
$X_{13} = X_{10}X_{12}$
$X_{14} = X_{11}X_{13}$
$X_{15} = X_8X_{14}$
$X_{16} = X_1X_{15}$

Optimally Computing Compressed Indexing Arrays Based on the Compact Directed Acyclic Word Graph

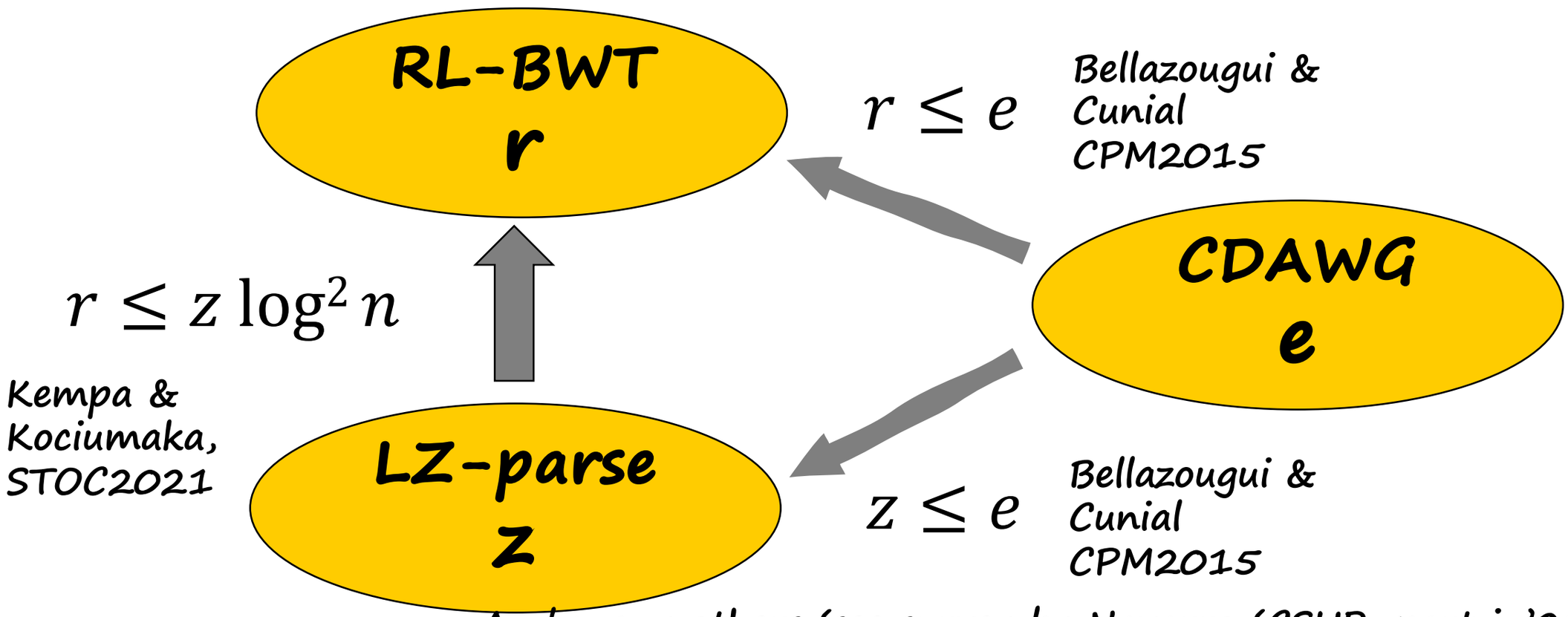
Hiroki Arimura[1]
Shunsuke Inenaga[2]
Yasuaki Kobayashi[1]
Yuto Nakashima[2]
Mizuki Sue[1]

1) Graduate School of IST, Hokkaido University, Japan
2) Department of Informatics, Kyushu University, Japan

For details and proofs, visits the arXIv site of this talk:
- Manuscript at https://arxiv.org/abs/2308.02269 ;
- Slide pdf at "Code section" or Github https://ikndeva.github.io)

This work is partly supported by MEXT Grant-in-Aid for Basic Research A, 2000-2004, Japan

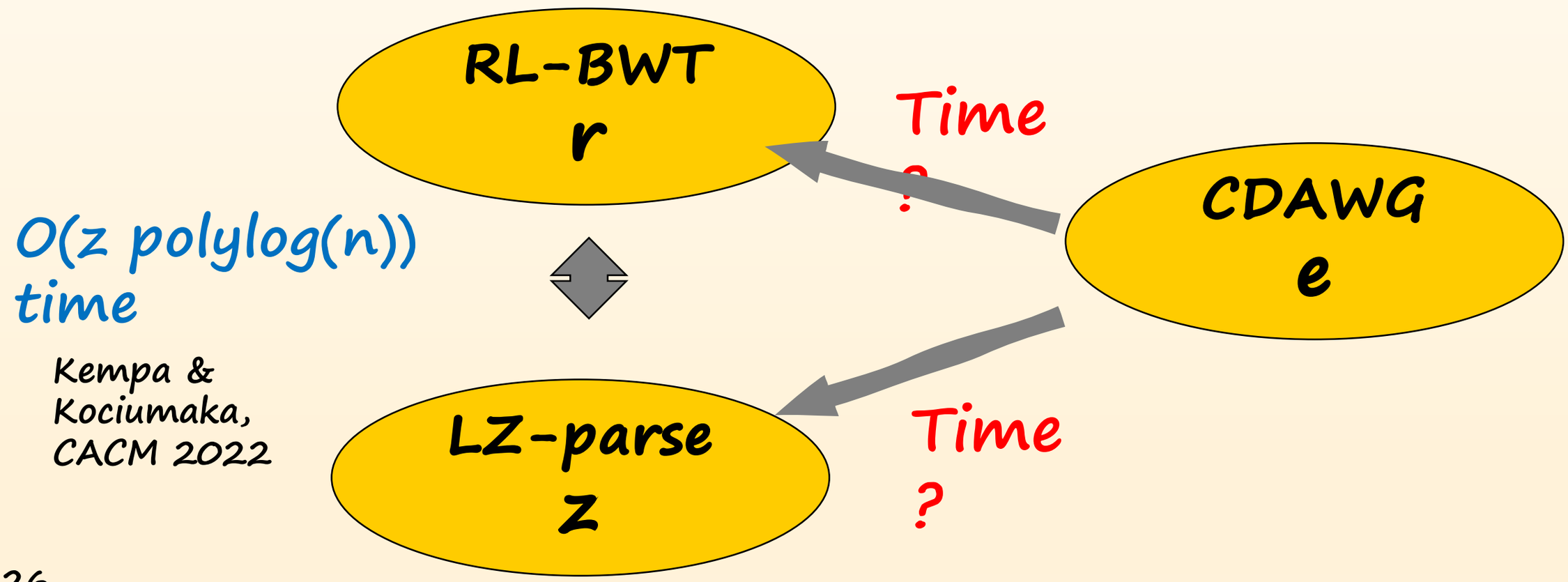
- Consider the relationship between their sizes
  - has been studied so far.



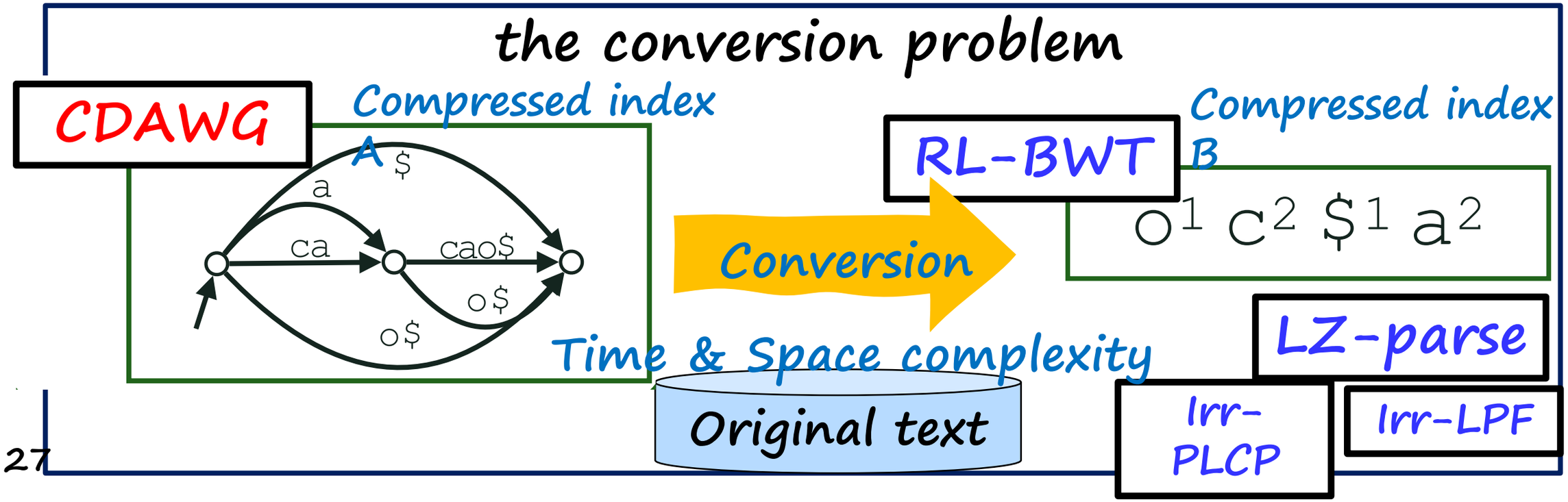
And many others (see survey by Navarro (CSUR, part i, '21)

Backgrounds: Previous work

- The time and space complexities of conversions
  - have not been studied very much

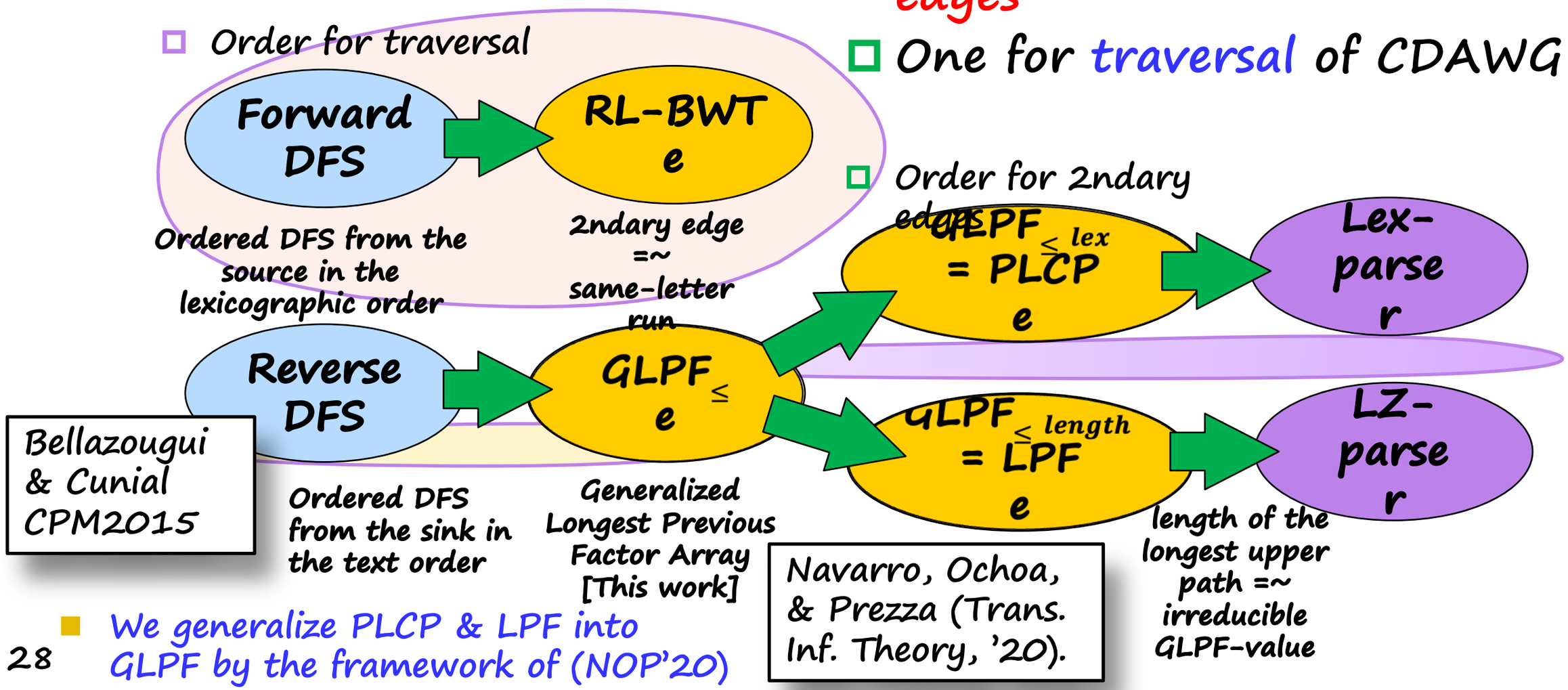


We devise efficient algorithms that solves the conversion problem from the CDAWG for a text T into various compressed indexes for T in linear time and space in the combined input/output sizes

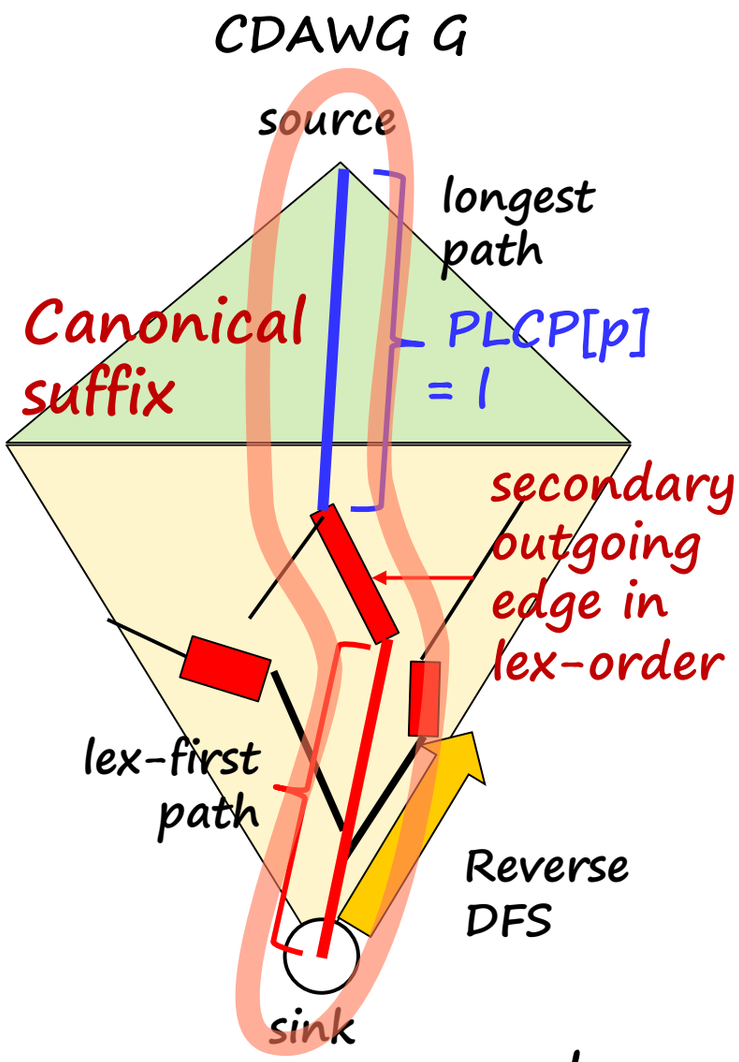


Our approach

- We use two orders of paths
  - Order for traversal
  - Order for defining 2ndary edges
  - One for traversal of CDAWG



- We generalize PLCP & LPF into GLPF by the framework of (NOP'20)



- **Observation A1**: O(e) secondary outgoing edge of CDAWG(T) under the length-order determines the irreducible value PLCP[p] = l by the length l of the longest path from the source to the corresponding branching node

- **Observation A2**: O(e) secondary outgoing edges can be enumerated in the text order of its "canonical suffix" by the reverse DFS from the sink.

We can extend the above result from PLCP to PLPF by employing the definition of 2ndary outgoing edges in length-order

おまちしています

情報知識ネットワーク研究室

有村・
{arim}@ist.hokudai.ac.jp
ex. 7678, 7679