

# 特許を対象とするスコアリングモデル および 大規模言語モデルの研究

2024/11/7

愛知工業大学 野中尋史

# 自己紹介

野中 尋史:hnonaka@aitech.ac.jp

愛知工業大学 経営学部 准教授

- 2015年 長岡技術科学大学 工学研究科 講師
- 2018年 長岡技術科学大学 工学研究科 准教授
- 2022年 愛知工業大学 経営学部 准教授  
(2006-2011年 豊橋技術科学大学知財連携コーディネーター)

委員歴

- 2020年4月 - 現在妙高市, 連携研究員
- 2021年4月 - 2022年3月電子情報通信学会, 会誌編集委員
- 2021年4月 - 2022年3月電子情報通信学会, 代議員
- 2019年4月 - 2022年3月長岡AIイノベーションハブ, 代表
- 2019年6月 - 2021年6月電子情報通信学会, 信越支部庶務幹事

# 特許・論文



## 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

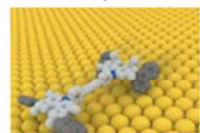
機械工学  
技術



融合



化学  
技術

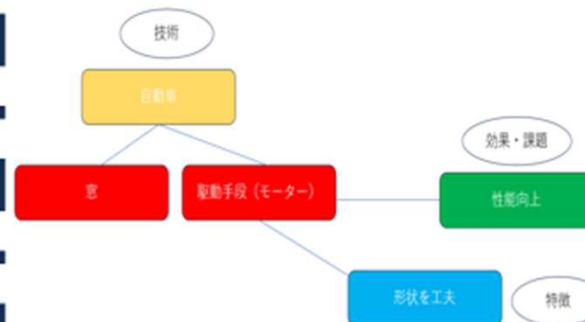


分子マシン

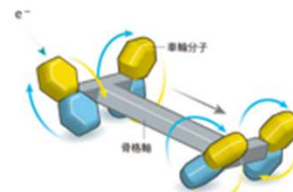
技術融合可能な  
技術群特定

## 創造された技術構成

### 知識グラフ



### 図面



LLMでの技術創造・推論  
(技術構成の具体化)

# 研究目的

- 技術を創造するAIを開発しよう！
  - 研究開発
  - 知財戦略
  - マーケティング
  - ...
- 大規模言語モデルだけでは不安...
- 特許スコアリングモデルと大規模言語モデルを組み合わせよう！

# 特許・論文



## 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

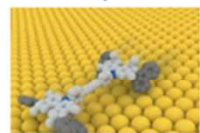
機械工学  
技術



化学  
技術



融合

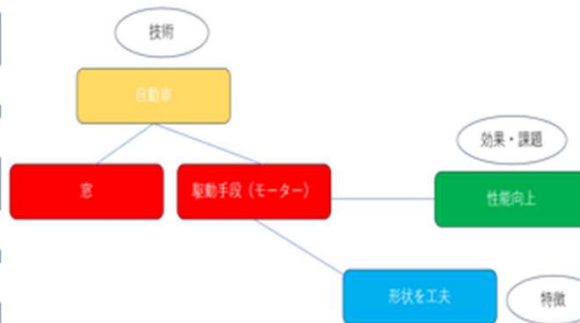


分子マシン

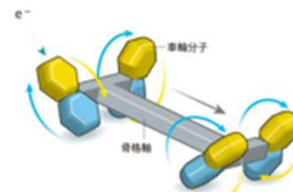
技術融合可能な  
技術群特定

## 創造された技術構成

### 知識グラフ



### 図面



LLMでの技術創造・推論  
(技術構成の具体化)

# 重要技術の特定とスコアリング

- 異分野にも波及する重要技術の定量評価の研究が盛ん
- 特許重要性に関するスコアリングの観点は複数ある
  - 権利期間
  - 引用情報
  - 特許権に関連するアクション
  - 特許文書の品質
  - ...

# 重要技術の特定とスコアリング

- 異分野にも波及する重要技術の定量評価の研究が盛ん
- 特許重要性に関するスコアリングの観点は複数ある
  - 権利期間
  - 引用情報

# 重要技術の特定とスコアリング

- 異分野にも波及する重要技術の定量評価の研究が盛ん
- 特許重要性に関するスコアリングの観点は複数ある
  - 権利期間：長期での予測が難しい
  - 引用情報



Technological Forecasting and Social Change

Volume 203, June 2024, 123390



## A study on patent term prediction by survival time analysis using neural hazard model

Koji Marusaki <sup>a</sup>  , Kensei Nakai <sup>b</sup>, Shotaro Kataoka <sup>c,f</sup>, Seiya Kawano <sup>d</sup>, Asahi Hentona <sup>b</sup>, Takeshi Sakumoto <sup>c</sup>, Yuta Yamamoto <sup>b</sup>, Kaede Mori <sup>g</sup>, Hirofumi Nonaka <sup>e,f</sup>

Show more 

 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.techfore.2024.123390>

[Get rights and content](#) 



# 背景と目的

- 特許情報を用いた技術の価値評価は企業での意思決定に有益[1]
  - 意思決定：企業の研究開発やM&A等
  - しかし、分析には費用・労力がかかる上、定量的な評価も難しい
    - 容易に特許技術を評価できる手法が求められる
- 特許権の生存分析による、各特許の価値評価
  - 日本の特許権は、出願日から最長20年間（一部例外あり）
    - 権利期間 = 権利者が「**特許料**」を何年間納付し続けるか
      - 価値のない特許にはお金を払わない
      - 仮説：権利期間の**長い特許 = 価値の高い特許**（権利）
  - 目的：特許権利期間の背景にある特許素性を用いた権利期間予測

# 先行研究

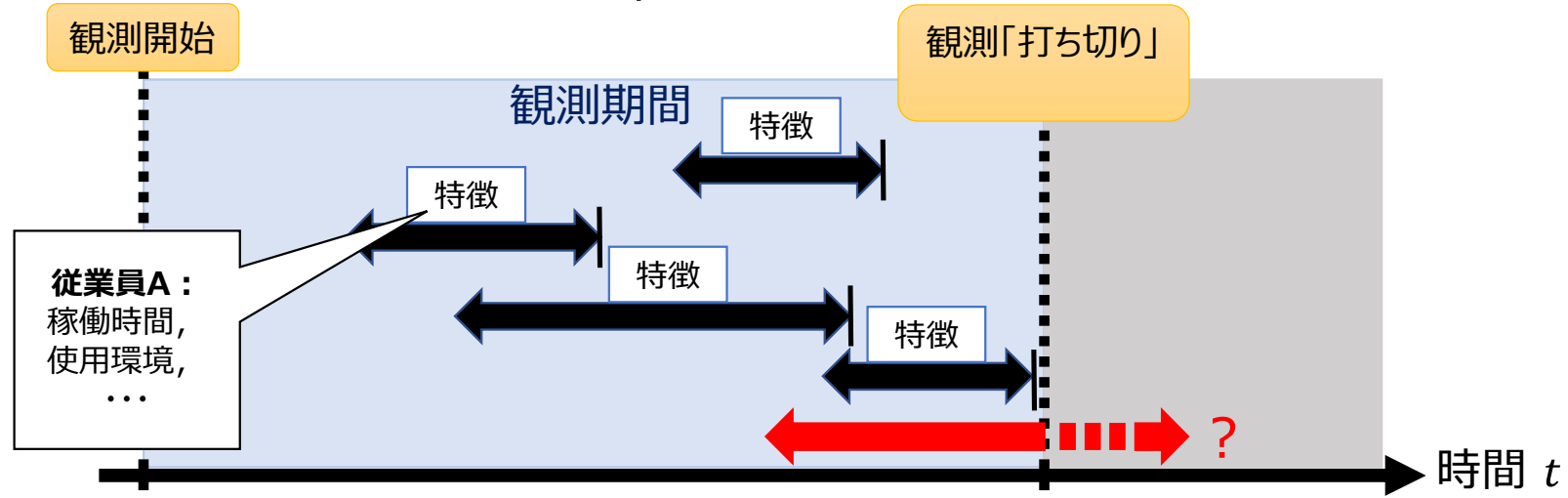
- 「**特許権利期間の長さが経済価値の指標となる性質**」
  - Bessen [2] : 予め算出した特許使用料収益推移を用いて、権利期間を推定
    - 権利存続中の特許から、正確な特許収入・特許権利期間の予測は困難
- **特許文書の素性から特許権利期間を生存分析**
  - 和田 [3] : 特許の引用情報に着目して、特許権利期間への影響を分析
  - Zeebroeck [4] : 欧州における特許権利期間を決定する要因を総合的に分析
- 2010年代より、ニューラルハザードモデルが提唱・適用[5,6]
  - 深層学習を生存分析に取り入れた手法
  - 説明変数－目的変数間の関係式における、非線形な作用も分析に適用可能ニューラルハザードモデルであるDeepSurv [7]を用いて特許素性から権利期間を分析

# 生存分析とは

- ある時系列に対して,  
「イベントが発生する**タイミング**」(目的変数)と  
「そのイベントが発生するための**条件**」(説明変数)との関係进行分析
  - →未知のケースに対して「いつイベントが発生するか」を予測可能
- 「イベント」の定義：
  - 同じ時系列において、「1回のみ」発生
  - 繰り返し発生する場合, それぞれを別々の事例と捉える
- 「イベント」の例：
  - 機械の故障, 定期課金サービスの解約, 借金の完済 等

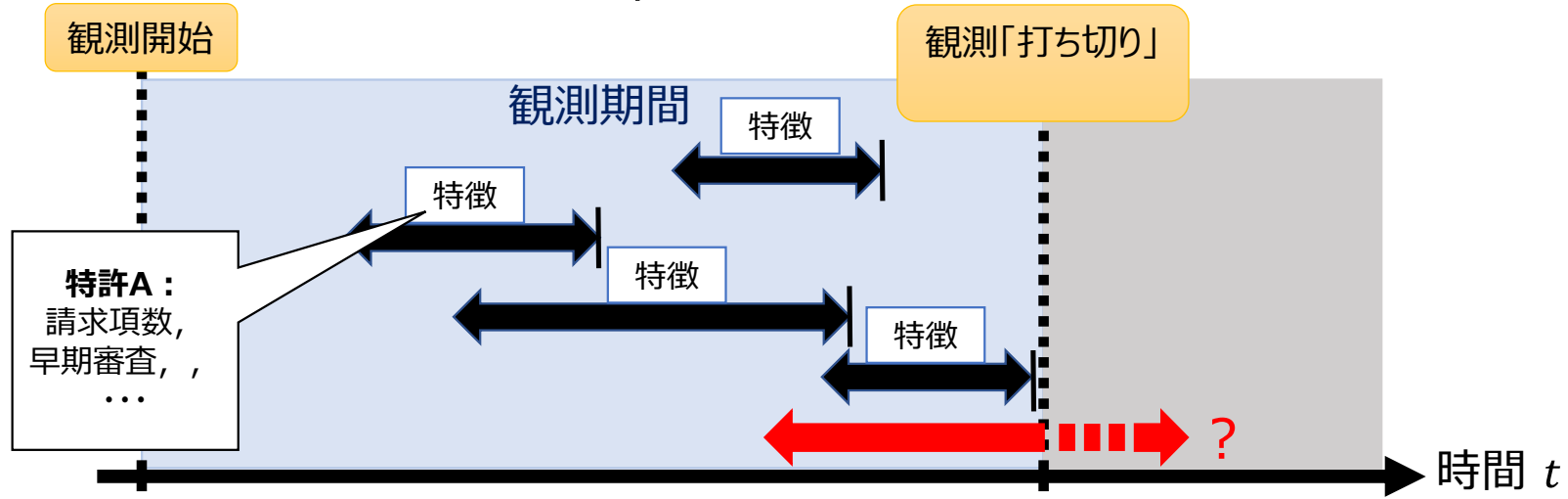
# 生存分析の例

- 「機械の故障時間予測」
  - イベント：機械の故障
  - 説明変数：機械の稼働時間，使用環境（標高，気温，湿度，・・・）
- 「説明変数がどのような場合に，故障する可能性が高いか」を分析



# 生存分析の例

- 「特許権利期間予測」
  - イベント：特許権の放棄（年金負支払）の判断
  - 説明変数：請求項数，早期審査の有無，
- 「説明変数がどのような場合に，権利を放棄する可能性が高いか」を分析



# 生存分析の概要

- 生存分析に必要なデータ :

- $x$ : ベースラインデータ (説明変数となる特徴ベクトル)
- $T$ : イベント発生時刻
- $E$ : イベント発生の有無 (0 or 1)

- 生存分析に必要な 2 つの関数 :

- 生存関数 Survival function :

- 時刻  $t$  の時点で生き延びる (イベントが発生しない) 確率

$$S(t) = \Pr[T > t]$$

- ハザード関数 Hazard function:

- $t$  時まで生存していた条件下で, 次の瞬間 ( $t + \Delta t$ ) に死亡 (イベントが発生) する確率

$$\lambda(t) = \lim_{dt \rightarrow 0} \frac{\Pr[t \leq T < (t + dt) | T \geq t]}{dt} = -\frac{d}{dt} \log[S(t)]$$

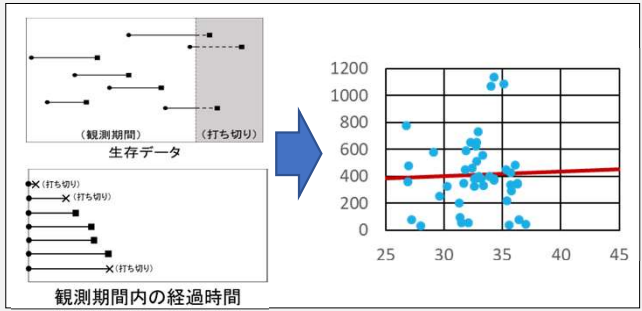
# 単純な回帰分析との違い

「打ち切り」: 生存データ特有の、「実験観測終了」によるイベント時系列の欠損

**×** 単純な回帰分析

- データの「打ち切り」を考慮しない
  - 「イベント発生」と「打ち切り」が区別できない

分析結果から、  
正確な生存時間の予測は難しい

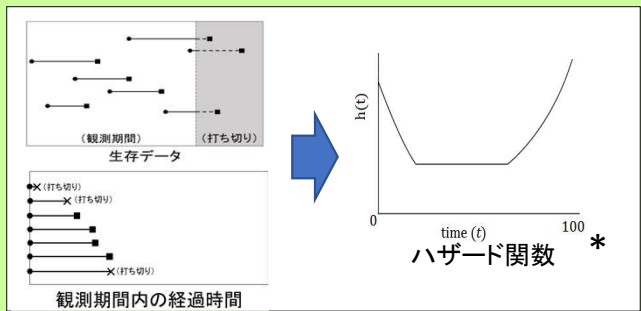


観測期間内の経過時間

生存分析

- データの「打ち切り」を考慮
  - 説明変数との関係を確率分布に反映

打ち切りがあっても、  
直前までの生存データを反映可能



観測期間内の経過時間

ハザード関数 \*

\*“Statistics How To” <<https://www.statisticshowto.datasciencecentral.com/hazard-function/>>より

# 分析手法：DeepSurv

Cox比例ハザードモデル[8]の深層学習版（多層パーセプトロン）  
従来手法との違い：「リスク関数 $h(x_i)$ および変数寄与度 $\theta$ の算定方法」

## Cox比例ハザードモデル（従来手法）

- リスク関数 $h(x_i)$ ：
  - 線形的に推定

$$\hat{h}_\theta(x_i) = \theta_1 x_{1,i} + \theta_2 x_{2,i} + \dots + \theta_n x_{n,i}$$

- 変数寄与度 $\theta$ ：
  - 対数尤度関数で最尤推定

$$l(\theta) := - \sum_{i:E_i=1} \left( \hat{h}_\theta(x_i) - \log \sum_{j \in \mathcal{R}(T_i)} \exp[\hat{h}_\theta(x_j)] \right)$$

## DeepSurv

- リスク関数 $h(x_i)$ ：
  - 変数寄与度 $\theta$ をネットワークの重みとして、損失関数で推定
- 変数寄与度は推定せず、直接リスク関数を算出



モデルのより柔軟な表現が期待できる



# 分析手法：DeepSurv

リスク関数  $\hat{h}_\theta(x_i)$  : 各特徴ベクトルを入力とし, ハザード関数  $\hat{h}_\theta(x_i)$ を推定して出力

$$\text{損失関数 } l(\theta) := - \sum_{i:E_i=1} \left( \hat{h}_\theta(x_i) - \log \sum_{j \in \mathfrak{R}(T_i)} \exp[\hat{h}_\theta(x_j)] \right)$$

$\hat{h}_\theta(x_i)$ : リスク関数  
 $\theta$ : ネットワーク (説明変数) の重み  
 $x$ : 説明変数  
 $\lambda_0(t)$ : ベースラインハザード関数  
 $\mathfrak{R}(T)$ : 時間  $T$  におけるリスク集合

ハザード関数  $\lambda(t|x)$  : 時間成分  $\lambda_0(t)$  と特徴成分が比例すると仮定

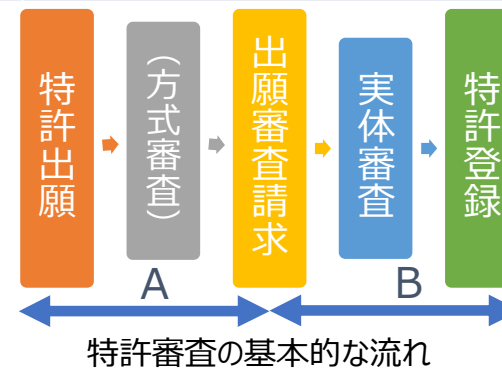
$$\lambda(t|x) = \underbrace{h_0(t)}_{\text{時間成分}} \cdot \underbrace{\exp[\hat{h}(x)]}_{\text{特徴成分 (リスク関数)}}$$

$\hat{h}(x)$ : リスク関数  
 $x$ : 説明変数  
 $h_0(t)$ : ベースラインハザード関数

# 入力する説明変数

Name	Content
Inventor	the number of inventors
Claim	the number of claims
IPC	(categorical variable) Whether each section of the IPC is assigned
Patent Owner Attributes	(categorical variable) Whether a patent owner belongs to "company" or "school" or "public_inst"
Citing	the number of examiner backward citations
Cited_std	the standardized value of the number of examiner forward citations
Reqstb_days	the term from the day of filing the examination request to the deadline of it (The earlier the request for examination, the higher the value)
Exam_days	the term from the day of requesting examination to the day of being granted

※審査請求期限：  
2001年9月30日出願分までは7年，それ以降は3年



2024/10/28

# 実験

- Pythonライブラリ「PySurvival」[10]のDeepSurvモデルを用いて分析
- 使用データ：
  - 知的財産研究所「IIPパテントデータベース」(2017) より
    - 1980～1999年出願の特許, かつ2004年末時点で**特許期間が満了**しているもの
    - 使用する項目においてデータに欠損・例外のない**予測対象特許の内訳**  
(権利失効時期別)
    - データ数：210,752件
- 対象特許：
  - 観測期間：1980～1999年 (20年間)
  - 予測期間：1980～2004年 (25年間)
  - 特許分類：すべて

権利失効時期	特許件数 (件)
観測期間中 (打ち切りなし)	81,051
観測期間外 (打ち切りあり)	129,701
<b>合計</b>	<b>210,752</b>

# 実験

Method	CPH	Deepsurv
C-index (CI)	0.680 (0.681, 0.679)	0.677 (0.683, 0.670)
RMS E (CI)	862.062 (922.285, 801.839)	568.641 (664.013, 473.270)

# 重要技術の特定とスコアリング

- 異分野にも波及する重要技術の定量評価の研究が盛ん
- 特許重要性に関するスコアリングの観点は複数ある
  - 権利期間
  - 引用情報：引用数は非階層的で過去の特許を過大評価.

## Development of the patent values evaluation method considering growth of technical community

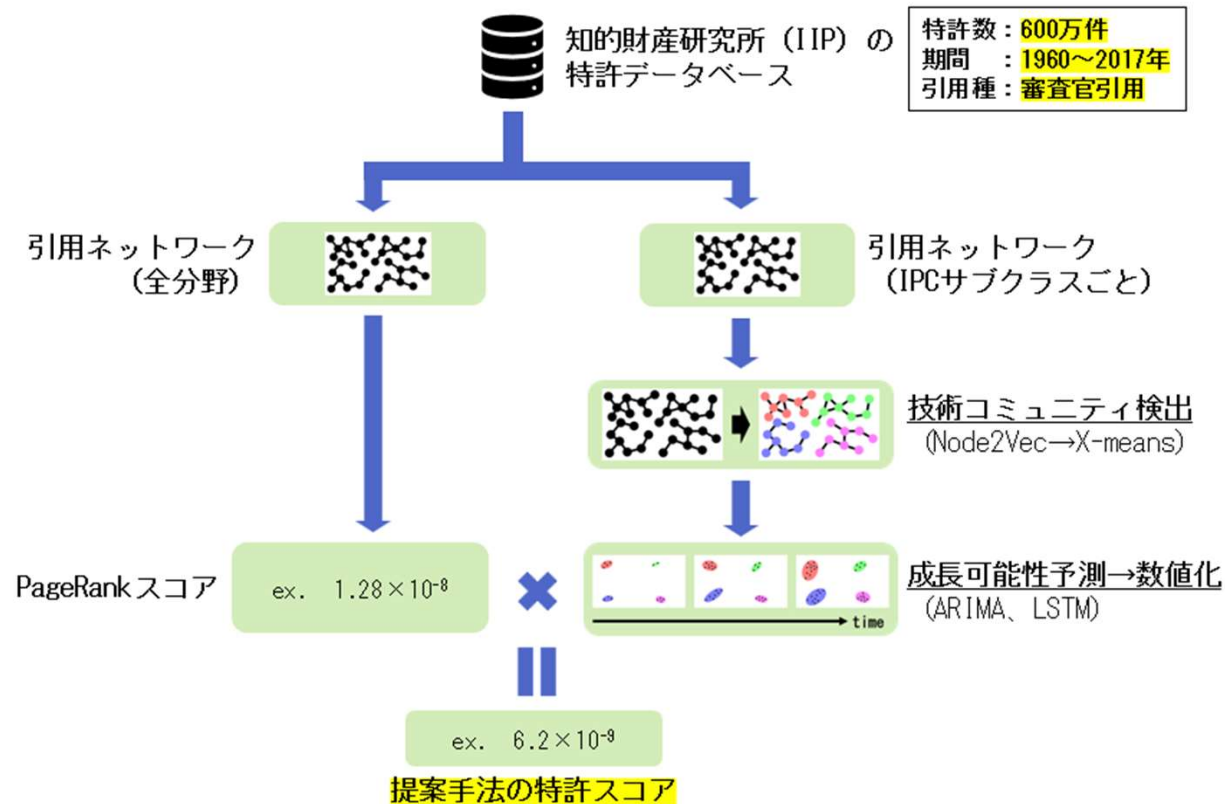
Publisher: **IEEE**

[Cite This](#)

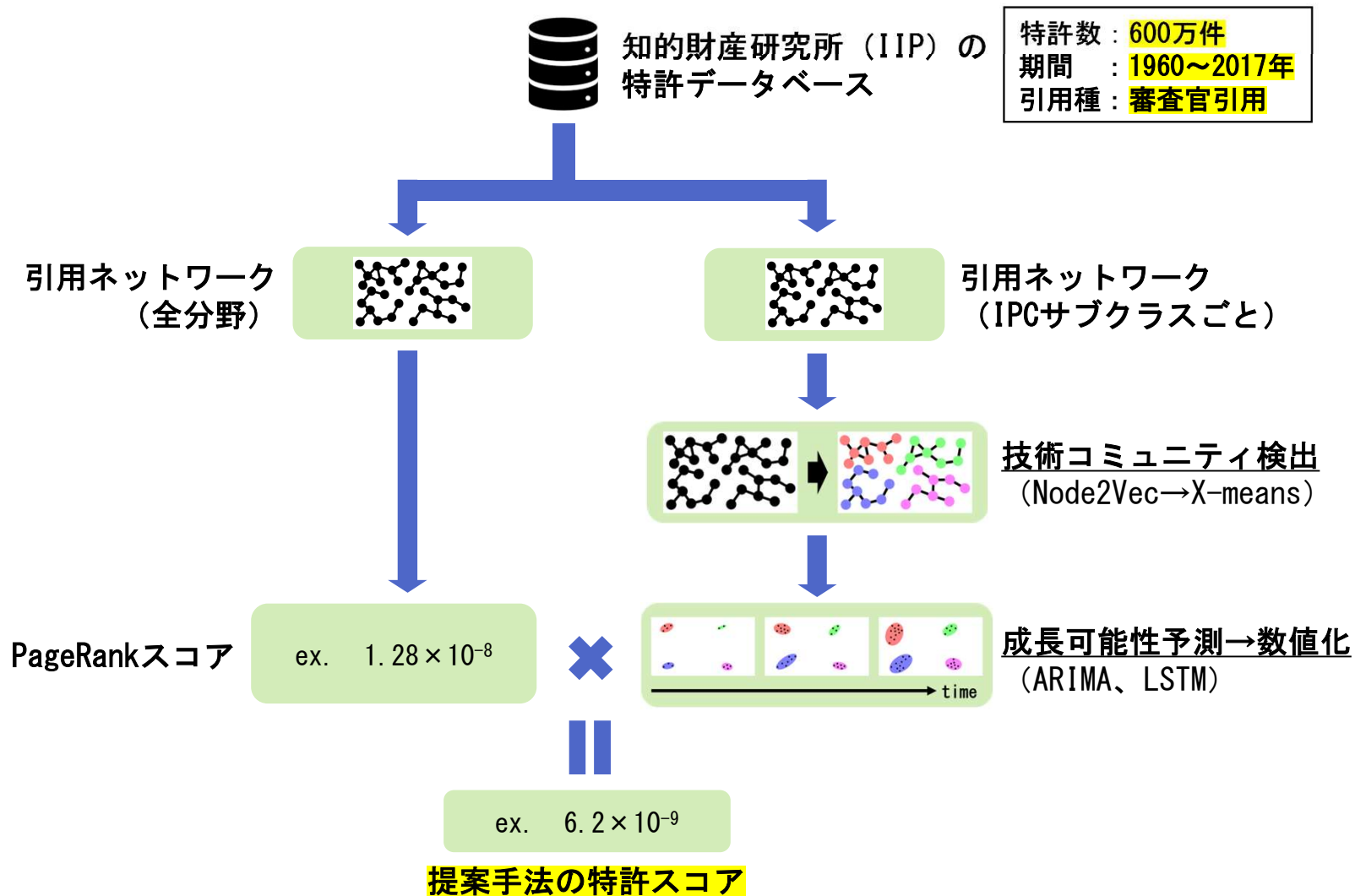
[PDF](#)

[Yuta Yamamoto](#) ; [Asahi Hentona](#) ; [Koji Marusaki](#) ; [Kohei Watabe](#) ; [Seiya Kawano](#) ; [Tokimasa Goto](#) **All Authors**

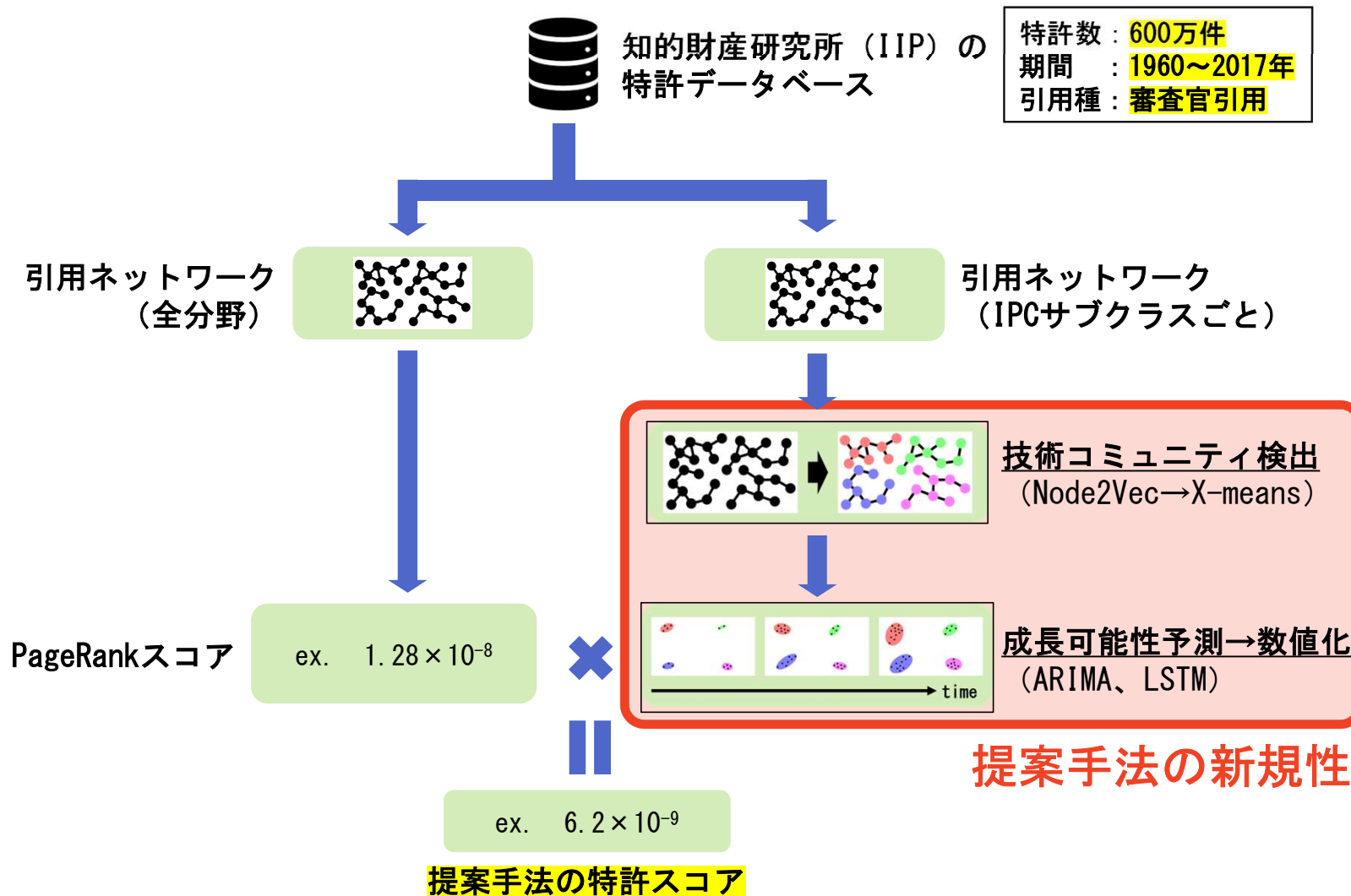
# 引用ネットワークの成長性も加味したスコアリング



# 引用ネットワークの成長性も加味したスコアリング

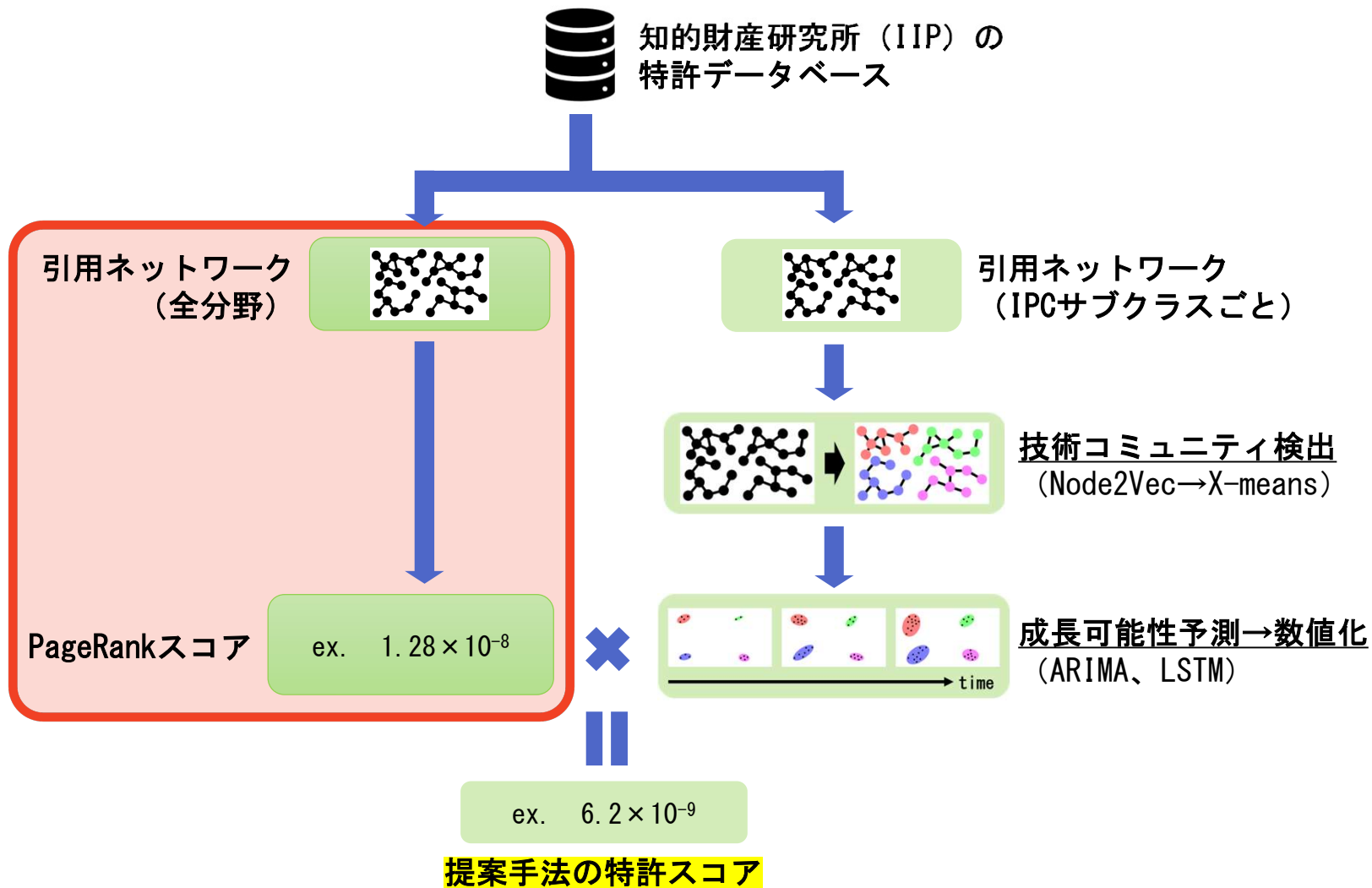


# 引用ネットワークの成長性も加味したスコアリング

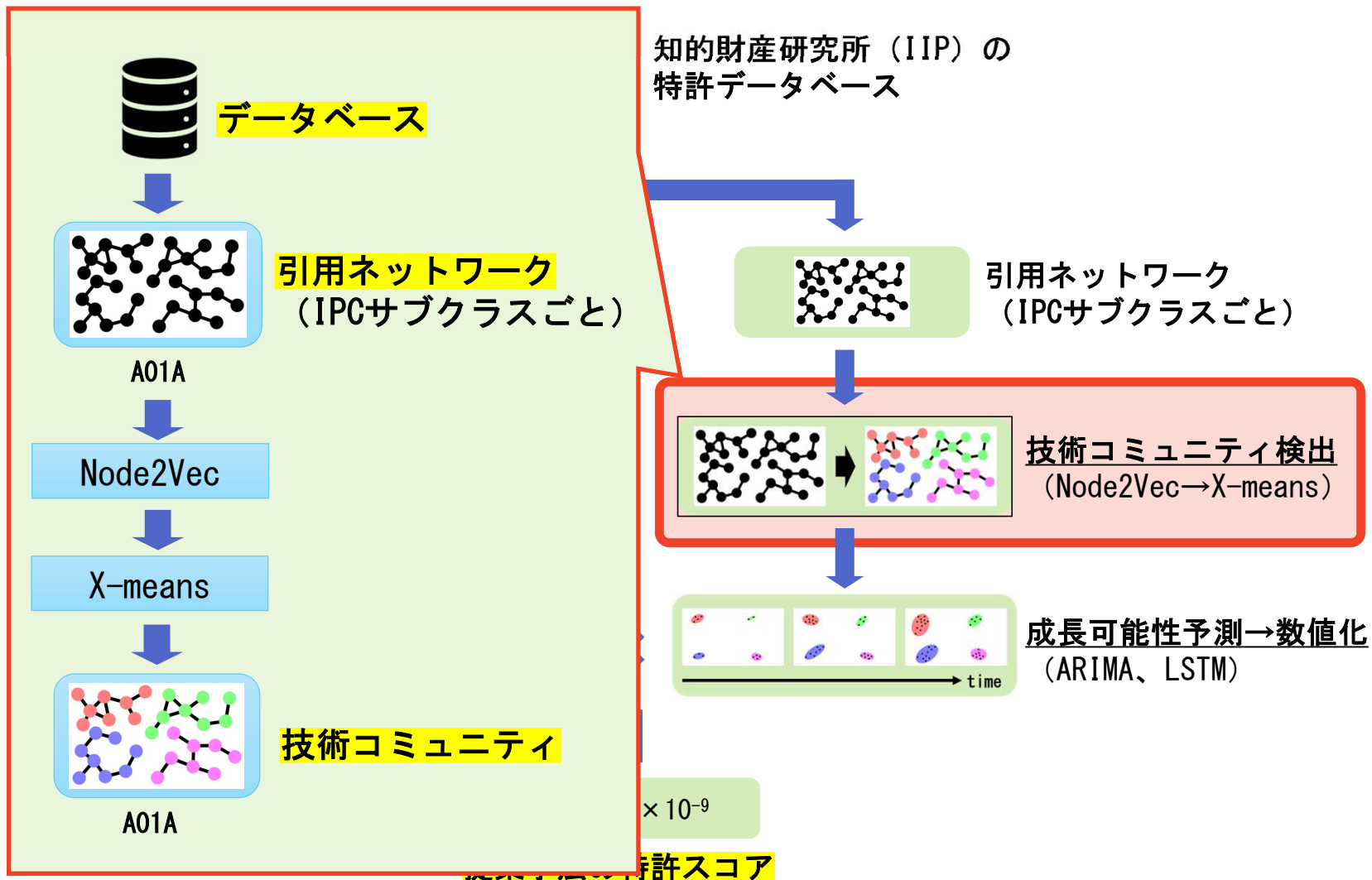




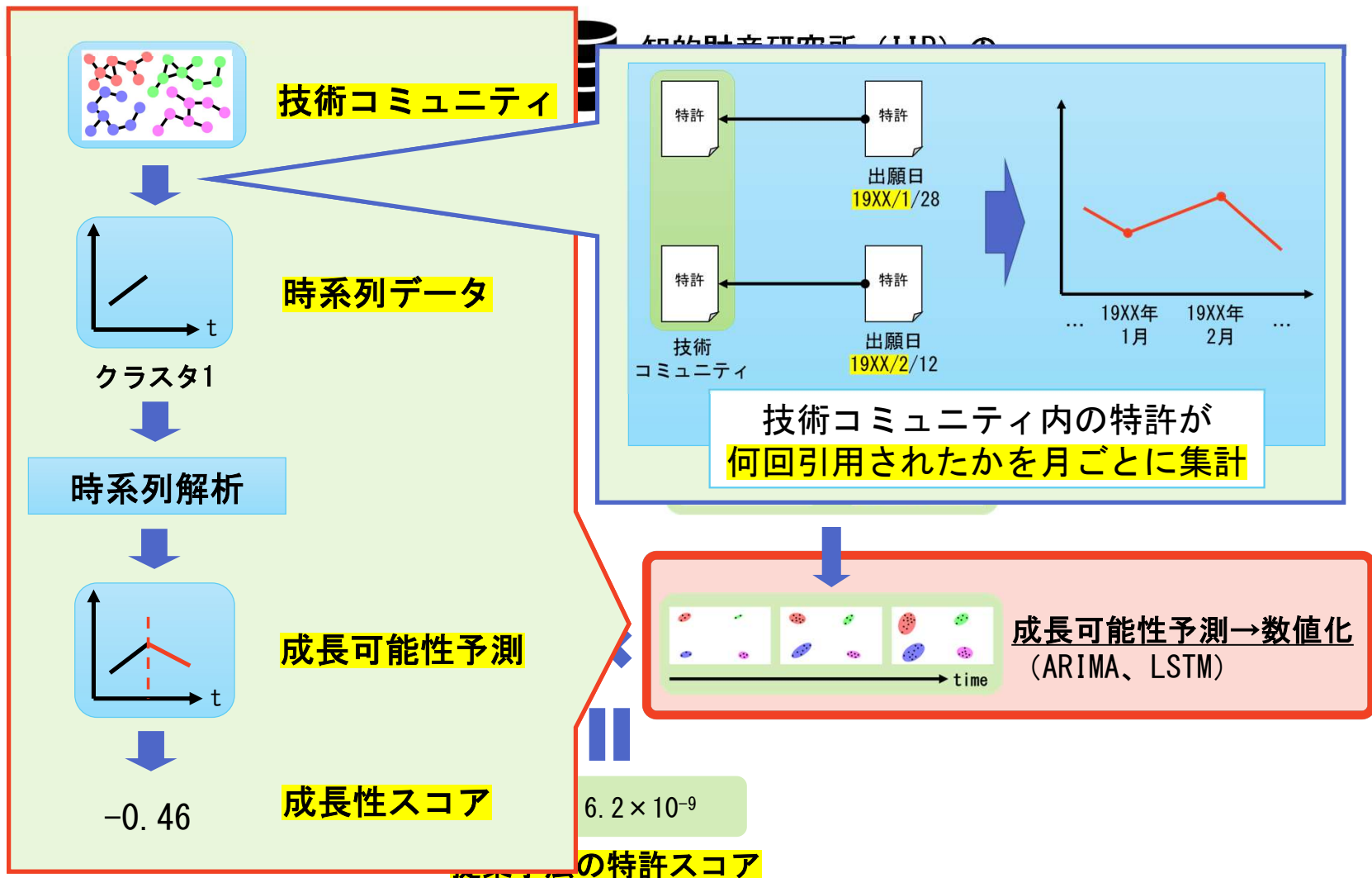
# 引用ネットワークの成長性も加味したスコアリング



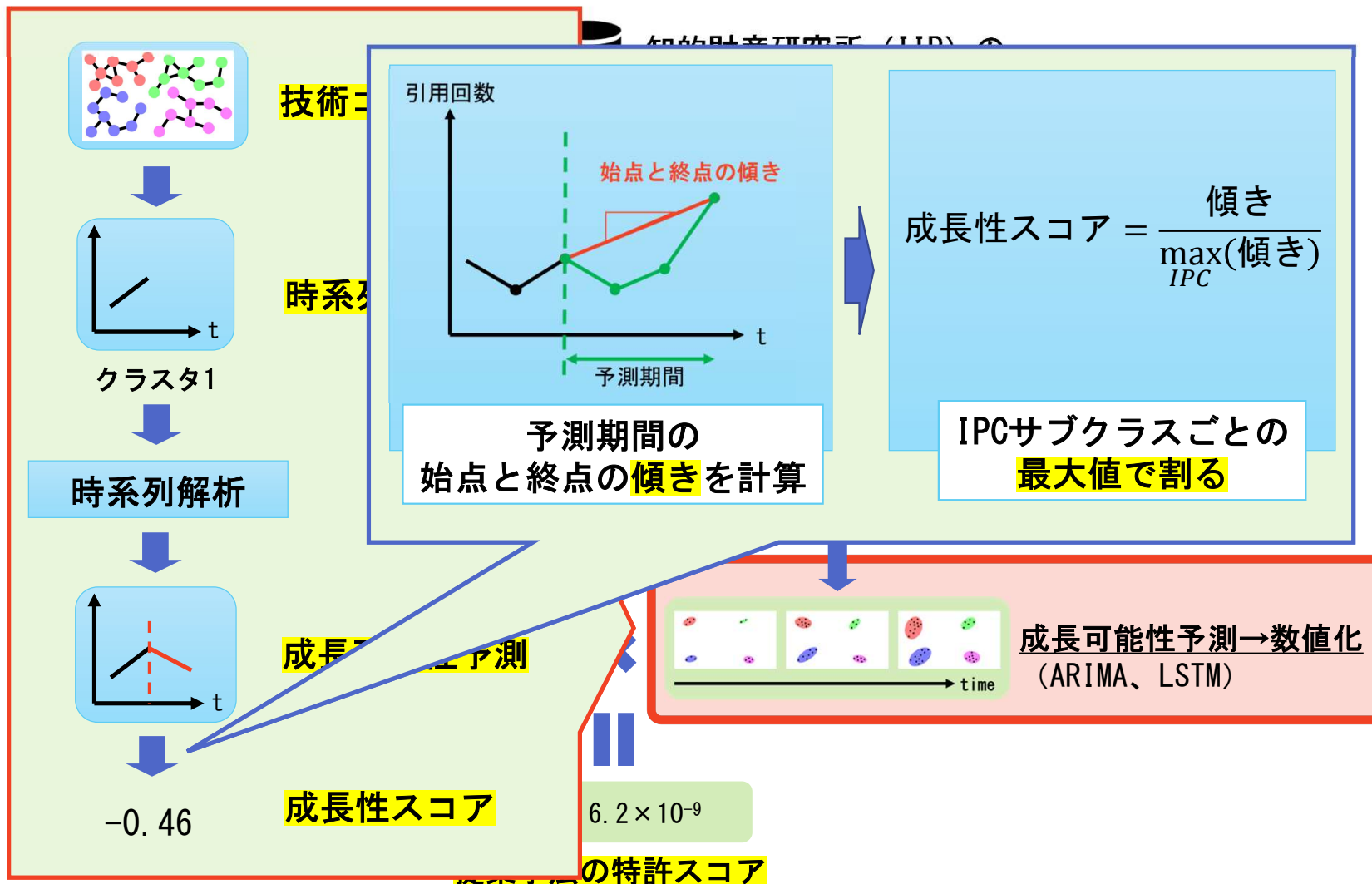
# 引用ネットワークの成長性も加味したスコアリング



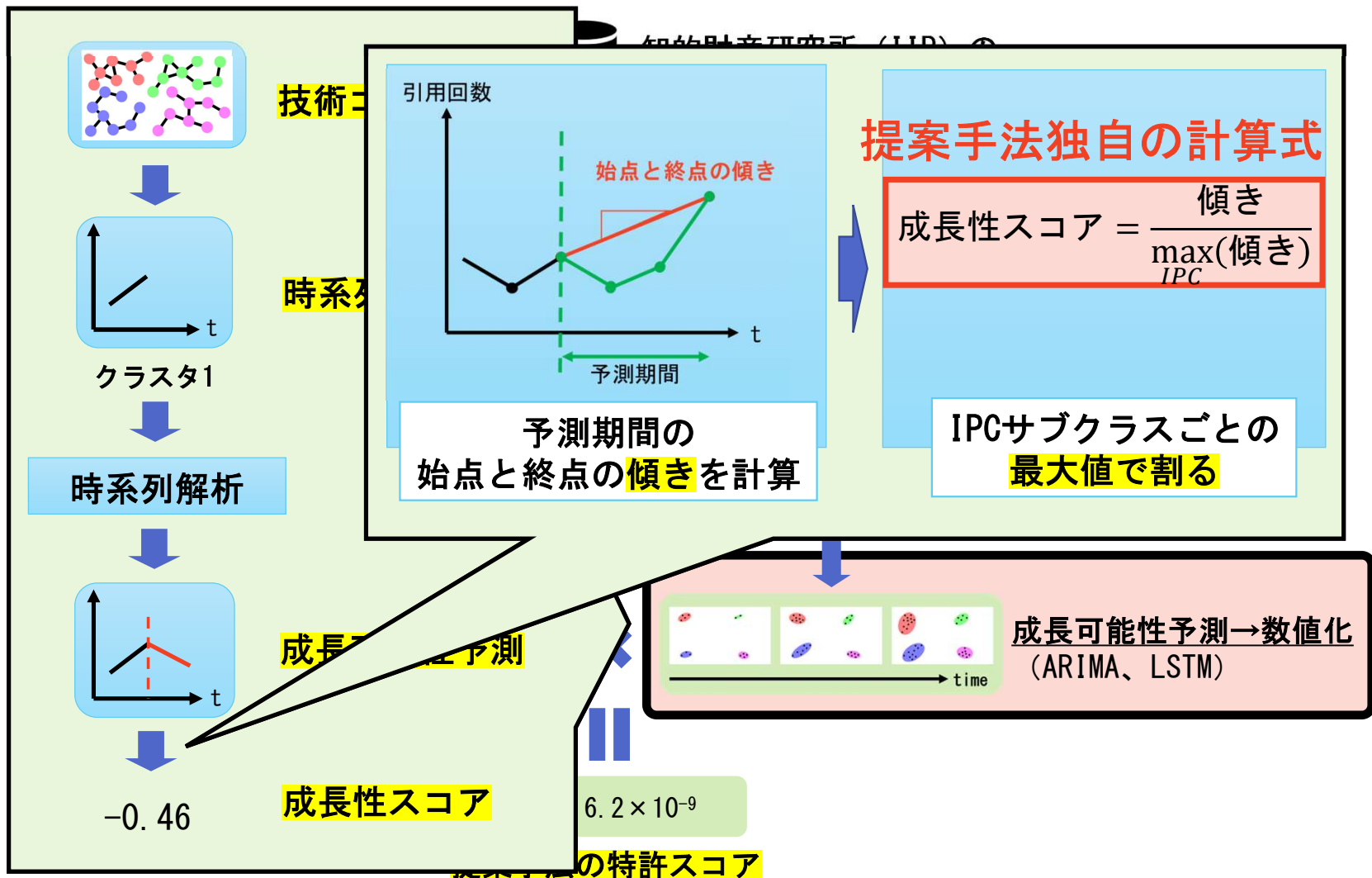
# 引用ネットワークの成長性も加味したスコアリング



# 引用ネットワークの成長性も加味したスコアリング



# 引用ネットワークの成長性も加味したスコアリング



# 業種ごとに結果が異なる

## 結果 業種分類：電気機器

スコア	特許の内容
PageRank	レーザープリンタ、DVD、DVDレコーダ
提案手法	広告配信システム、ネット投票システム、情報共有支援システム

- PageRankスコアでは、プリンタ、DVDなど、昔からある重要な製品に関する特許を検出
  - 昔からある&重要な特許＝後発特許から多く引用されている
- 提案手法のスコアでは、情報系など、今後成長が期待される分野の特許を検出
  - 成長性可能性予測が特許の価値評価に寄与



- ✓ 提案手法の方が正確に特許価値を評価
- ✓ 同様の知財戦略を取る他分野でも提案手法のスコアが有効

## 結果 業種分類：医薬品

スコア	特許の内容
PageRank	企業の主力製品となる薬
提案手法	(PageRankほどの特徴なし)

- 業界の性質上、医薬品系特許は物質そのものの特許が多い
  - 類似・代替技術の出願は拒絶されやすい
  - 参入障壁が大きく、成長分野でも技術コミュニティが成長しない
- 発売から時間がたっても、製品の価値が下がりにくい
  - 1つ新薬を出せば、その後長期間にわたって利用される
    - その新薬の関連特許は技術分野の成長性に関わらず価値を維持続ける



- ✓ PageRankの方が正確に特許価値を評価
- ✓ 手法的に医薬品の分野では提案手法のスコアは有効でない

# 特許・論文



# 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

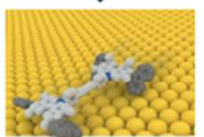
機械工学  
技術



化学  
技術



融合



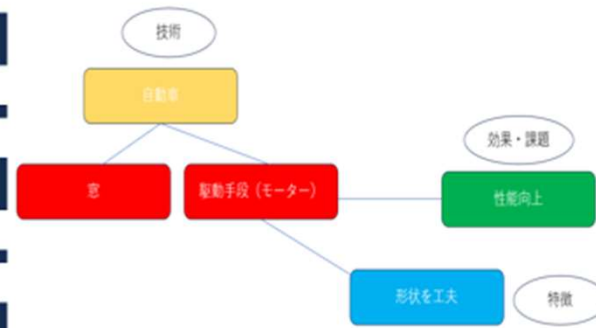
分子マシン

技術融合可能な  
技術群特定

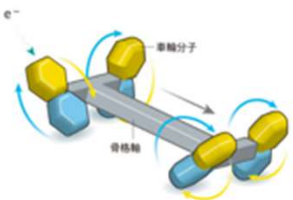
LLMでの技術創造・推論  
(技術構成の具体化)

# 創造された技術構成

## 知識グラフ



## 図面



# 特許・論文



# 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

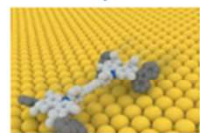
機械工学  
技術



化学  
技術



融合

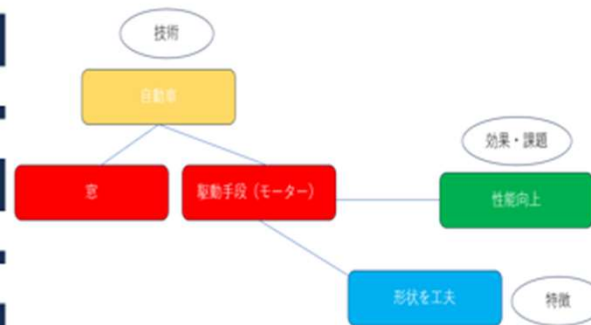


分子マシン

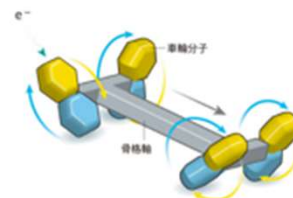
技術融合可能な  
技術群特定

# 創造された技術構成

## 知識グラフ



## 図面



LLMでの技術創造・推論  
(技術構成の具体化)



# 技術の融合可能性の評価

- 技術が融合するかどうかは類似性尺度や課題との共起ネットワークで判断
  - 技術要素としてデータセットに着目した類似性尺度
  - 課題と共起する技術のネットワーククラスタリング

# 技術の融合可能性の評価

- 技術が融合するかどうかは類似性尺度や課題との共起ネットワークで判断
  - 技術要素としてデータセットに着目した類似性尺度
  - 課題と共起する技術のネットワーククラスタリング

IEEE Access  
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 December 2023, accepted 28 February 2024, date of publication 11 March 2024, date of current version 21 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375750

## RESEARCH ARTICLE

### Metadata-Based Clustering and Selection of Metadata Items for Similar Dataset Discovery and Data Combination Tasks

TAKESHI SAKUMOTO<sup>1</sup>, TERUAKI HAYASHI<sup>2</sup>, HIROKI SAKAJI<sup>3</sup>, AND HIROFUMI NONAKA<sup>4,5</sup>

<sup>1</sup>Department of Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan

<sup>2</sup>Department of Engineering, The University of Tokyo, Bunkyo, Tokyo 113-8656, Japan

<sup>3</sup>Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan

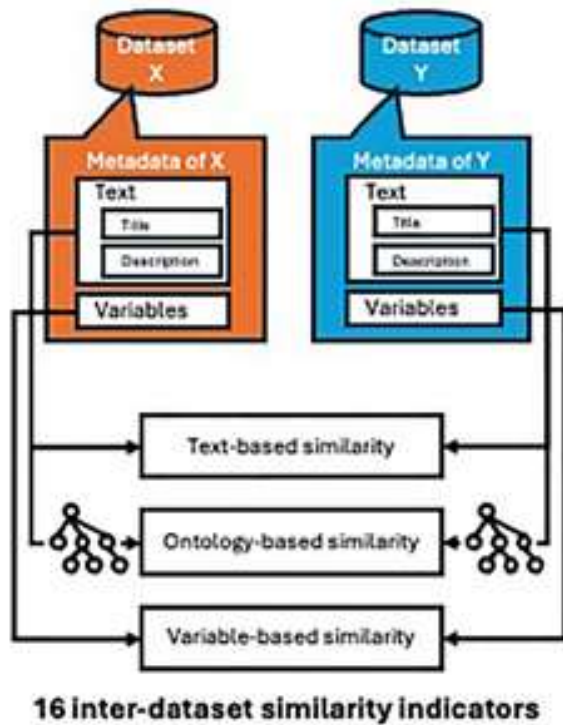
<sup>4</sup>Faculty of Business Administration, Aichi Institute of Technology, Toyota, Aichi 470-0392, Japan

<sup>5</sup>Mayolab Company Ltd., Nagaoka, Niigata 940-2137, Japan

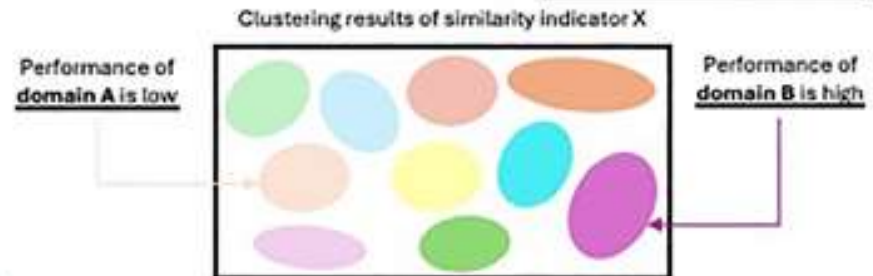
Corresponding author: Takeshi Sakumoto (s183353@stn.nagaokaut.ac.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP20H02384 and Grant JP19K12116.

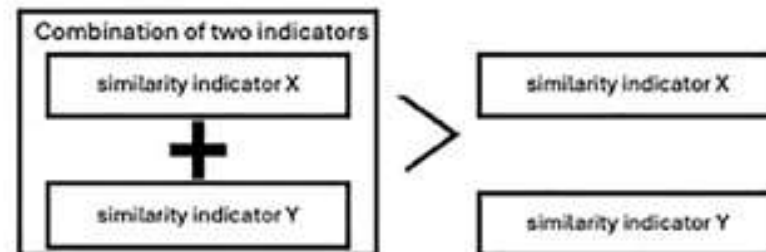
# データセット融合の際の類似度尺度



① Comparison of similarity indicators for 15 dataset domains through the clustering task Covid-19, Investing, SNS, ...



② Evaluation of the combination of two different indicators



# ドメインごとに類似性尺度は異なる

TABLE 5 A Comparison of Clustering Performances for Each Medical Domain

Model	Covid-19	Cancer	Heart Disease
Jaccard(T)	0.288	0.569	0.170
Jaccard(T+D)	0.348	0.471	0.211
Cosine(TF-IDF(T))	<b>0.611</b>	0.582	0.280
Cosine(TF-IDF(T+D))	0.391	0.557	0.330
Cosine(BERT(T))	0.290	0.500	<b>0.406</b>
Cosine(BERT(T+D))	0.173	0.707	0.189
Cosine(Word2Vec(T))	0.205	0.691	0.400
Cosine(Word2Vec(T+D))	0.158	0.626	0.177
DT(Cosine(TF-IDF(T)))	0.417	0.559	0.307
DT(Cosine(TF-IDF(T+D)))	0.504	0.489	0.252
Wu-Palmer(T)	0.220	<b>0.763</b>	0.128
Wu-Palmer(T+D)	0.146	0.738	0.167
Navigational(T)	0.139	0.658	0.169
Navigational(T+D)	0.145	0.494	0.174
Dice(V)	0.238	0.181	0.150
Cosine(TF-IDF(V))	0.292	0.262	0.177

TABLE 6 A Comparison of Clustering Performances for Each Financial Domain

Model	Investing	Currency	Banking
Jaccard(T)	0.414	0.163	0.542
Jaccard(T+D)	0.148	0.168	0.434
Cosine(TF-IDF(T))	0.208	<b>0.703</b>	0.601
Cosine(TF-IDF(T+D))	0.562	0.169	<b>0.646</b>
Cosine(BERT(T))	0.373	0.252	0.331
Cosine(BERT(T+D))	0.208	0.206	0.281
Cosine(Word2Vec(T))	0.538	0.470	0.464
Cosine(Word2Vec(T+D))	0.265	0.212	0.333
DT(Cosine(TF-IDF(T)))	0.500	0.592	0.318
DT(Cosine(TF-IDF(T+D)))	<b>0.578</b>	0.201	0.565
Wu-Palmer(T)	0.370	0.232	0.527
Wu-Palmer(T+D)	0.465	0.210	0.517
Navigational(T)	0.484	0.202	0.390
Navigational(T+D)	0.497	0.180	0.451
Dice(Variables)	0.429	0.311	0.308
Cosine(TF-IDF(Variables))	0.419	0.273	0.211

# 技術の融合可能性の評価

- 技術が融合するかどうかは類似性尺度や課題との共起ネットワークで判断
  - 技術要素としてデータセットに着目した類似性尺度
  - 課題と共起する技術のネットワーククラスタリング

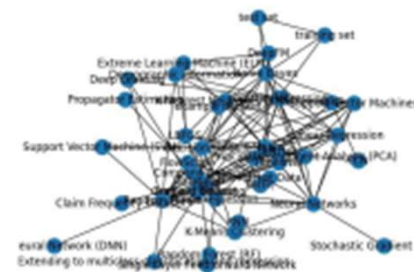
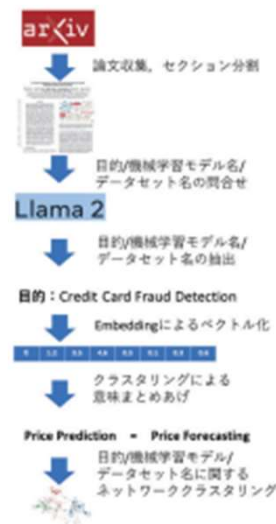
The 38th Annual Conference of the Japanese Society for Artificial Intelligence, 2024

LLM およびネットワーク解析を利用した学術論文からの研究目的・機械学習モデル名・データセット名の抽出と相互の関連性分析  
Extraction of research objectives, machine learning model names, and dataset names from academic papers using large-scale language models and graph network analysis, and a method for analyzing their inter-relationships

西尾 紗也香 \*1    野中 尋史 \*1    早矢仕 晃章 \*2    坂地 泰紀 \*3    作本 猛 \*4  
Sayaka Nishio    Hirofumi Nonaka    Teruaki Hayashi    Hiroki Sakaji    Takeshi Sakumoto

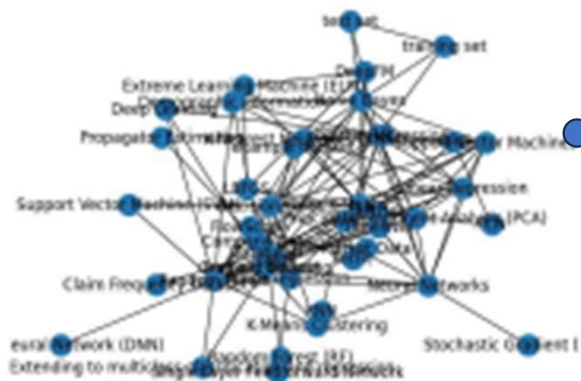
\*1 愛知工業大学    \*2 東京大学    \*3 北海道大学    \*4 長岡技術科学大学  
Aichi Institute of Technology    University of Tokyo    Hokkaido University    Nagaoka University of Technology

In this study, we propose a methodology extracting tasks, machine learning methods, and dataset names from scientific papers and analyzing the relationships between these information.



# 技術の融合可能性の評価

- 技術が融合するかどうかは類似性尺度や課題との共起ネットワークで判断
  - 技術要素としてデータセットに着目した類似性尺度
  - 課題と共起する技術のネットワーククラスタリング



計量ファイナンス分野では予測モデル開発のためにオルタナティブデータの融合が行われようとしている！？

# 特許・論文



## 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

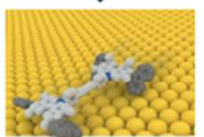
機械工学  
技術



化学  
技術

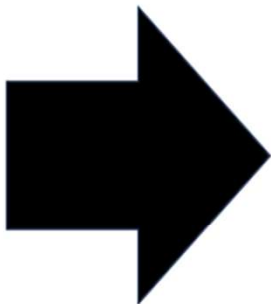


融合



分子マシン

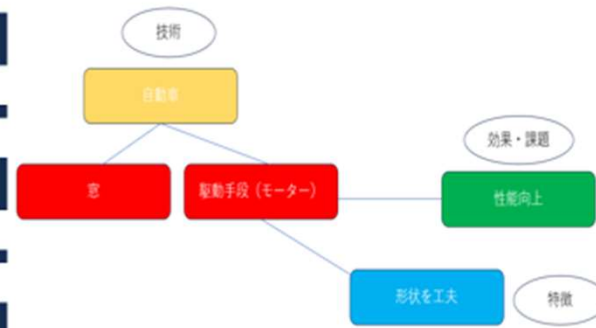
技術融合可能な  
技術群特定



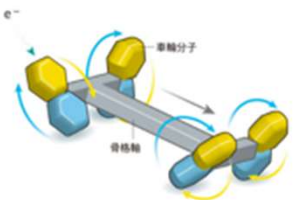
LLMでの技術創造・推論  
(技術構成の具体化)

## 創造された技術構成

### 知識グラフ



### 図面



# 特許・論文



## 異分野横断 技術創造手法

重要  
技術  
特定



機械工学技術

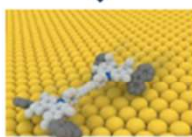
機械工学  
技術



化学  
技術



融合

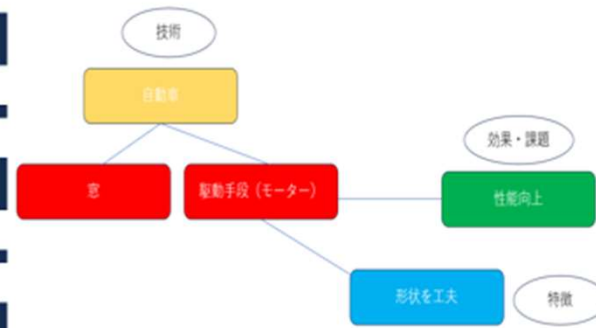


分子マシン

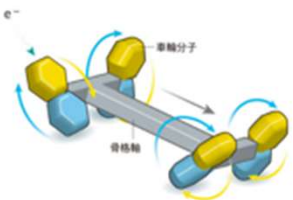
技術融合可能な  
技術群特定

## 創造された技術構成

### 知識グラフ



### 図面



LLMでの技術創造・推論  
(技術構成の具体化)



大規模言語モデルの発展のためには  
良質のデータセット整備が重要

## 科学知識発見を目的とした特許のアノテーション

日浦 隆博<sup>1,2</sup>, 吉田 奈央<sup>3</sup>, 松井 陽子<sup>2</sup>, 河野 誠也<sup>2,1</sup>, 野中 尋史<sup>4</sup>, 吉野 幸一郎<sup>2,1</sup>  
<sup>1</sup> 奈良先端科学技術大学院大学, <sup>2</sup> 理化学研究所 GRP, <sup>3</sup> ANNOTAN, <sup>4</sup> 愛知工業大学  
hiura.takahiro.hu6@is.naist.jp  
{yoko.matsui, seiya.kawano, koichiro.yoshino}@riken.jp  
annotan.tsukuba@gmail.com, hnonaka@aitech.ac.jp

- 特許中の知識を知識グラフとして整理：知識推論へ



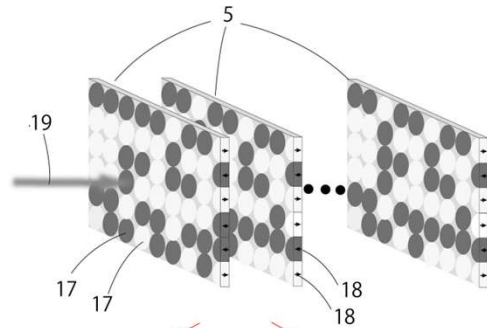
## 特許を対象とする画像言語モデル開発のための データセットの構築

Dataset Development of Vision-Language Model for Patent Data

安藤 七哉<sup>\*1</sup> 溝口 月斗<sup>\*1</sup> 石川 治樹<sup>\*1</sup> 伊豫田 彬<sup>\*1</sup> 河野 誠也<sup>\*2</sup> 吉野 幸一郎<sup>\*2</sup>  
Kazuya Ando Tsukito Mizoguchi Haruki Ishikawa Akira Iyoda Seiya Kawano Koichiro Yoshino

野中 尋史<sup>\*1</sup>  
Hirofumi Nonaka

<sup>\*1</sup>愛知工業大学 Aichi Institute of Technology <sup>\*2</sup>理化学研究所 GRP RIKEN GRP



【課題を解決するための手段】光入力信号が入力する磁性層を備え、該磁性層から光出力信号が出力される磁気光学演算素子であって、光入力信号が直線偏光、円偏光もしくは楕円偏光のいずれかであり、磁性層は磁気光学効果により光出力信号の偏光状態を変える磁気光学材料を含み、磁性層が、光入力信号を推定可能な磁気光学回折ニューラルネットワークを形成する。

図9は磁場印加下におけるレーザー照射による光磁気書き込みの概要図である。

### ○データセットの概要

- ・ 公開特許公報410103件 ( 補正公報除く,2022年-2023年 )
- ・ ケース1(解決手段)とケース2(図のキャプション)を作成
  - 解決手段は特許中のタグ情報から取得
  - 図のキャプションは正規表現を利用

### ○データセット全体の統計情報

- ・ 特許公報当たりの総図面数：12890111件
- ・ 特許公報当たりの平均図面数：31.4枚

### ○ケース1データセット固有の統計情報

- ・ 「課題を解決するための手段」の平均文字数：1506.2文字
- ・ 要約書「解決手段」の平均文字数：236.5文字

### ○ケース2データセット固有の統計情報

- ・ 図のキャプション中の平均文字数：337.5文字

大規模言語モデルの発展のためには  
良質のデータセット整備が重要

# 特許オープンデータ

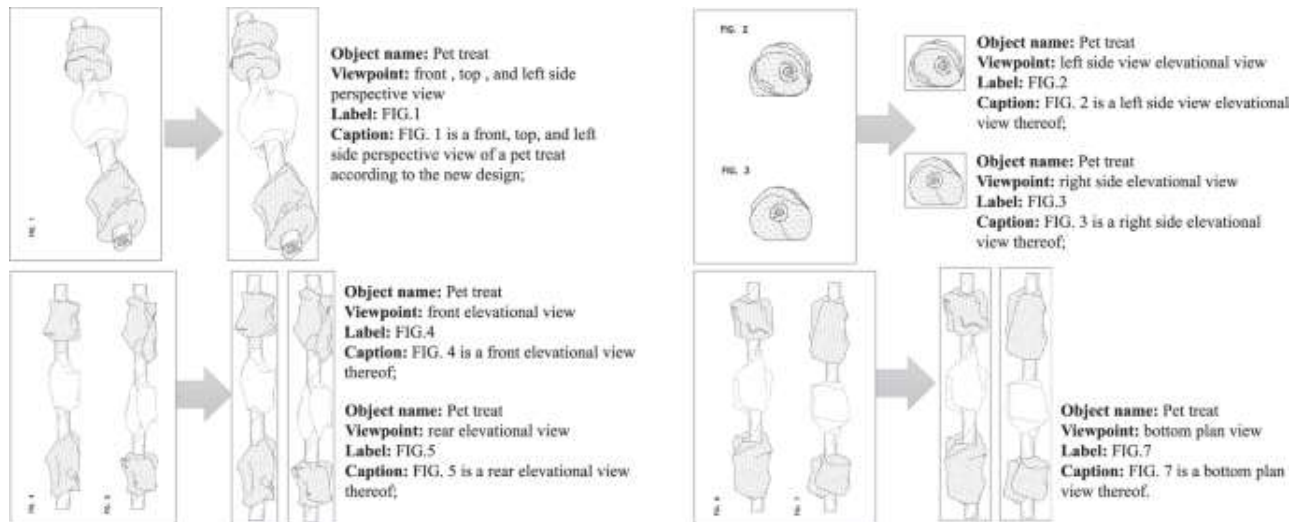
データセット	概要	図面タスクの有無	図面タスクの概要
CLEF-IP 2010年代	英語の特許対象. 検索, 情報抽出など	○	フローチャートとUML の対応
NTCIR PATMN 2000年代	日本 (とUSPTO) の特許対象. 検索, 情報抽出など	×	
TREC-CHEM 2010年代	化学特許対象	○	化学式とテキスト
DeepPatent2 2020年代	米国意匠特許30 万件	○	図と図のキャプション
HUPD	USPTOのテキスト 情報の構造化デー タ	×	

# 特許オープンデータ（図面のみ）

データセット	概要	図面タスクの有無	図面タスクの概要
CLEF-IP 2010年代	英語の特許対象. 検索, 情報抽出など	○	フローチャートとUMLの対応
DeepPatent2 2020年代	米国意匠特許30万件	○	図と図のキャプション
Ours	日本特許40万件	○	図と図に対応する各種 説明文※

# DeepPatent2

<https://www.nature.com/articles/s41597-023-02653-7>



図面とキャプション対応

# 特許関連研究のデータセットの整備について

- 2010年代をピークに減少傾向
- 特許実務者と情報科学分野の研究者が交流し、共同で良質なデータセットを整備する必要がある
  - 特許実務者と情報科学分野の研究者で断絶がある？
  - 官・民・学で連携の必要性



# 知財に関する産学連携

- 牛久先生（ムーンショットプロジェクトPM）、吉野先生、河野先生ら立ち上げた論文・特許情報処理に関する企業「NexaScience」
  - <https://www.nexascience.com/>
- 野崎篤志氏が運営するEパテントチャンネル出演（本日の研究の詳細）
  - <https://www.youtube.com/watch?v=P4iSXebFW7Q>
- 野中が立ち上げた特許解析・モノづくりデータ解析を行う企業「フリーヒルズラボ」
  - [hnonaka@aitech.ac.jp](mailto:hnonaka@aitech.ac.jp) まで（当該企業のことのみならずお気軽にお問合せください）

# 参考文献

- [1] Marusaki, K., Nakai, K., Kataoka, S., Kawano, S., Hentona, A., Sakumoto, T., ... & Nonaka, H. (2024). A study on patent term prediction by survival time analysis using neural hazard model. *Technological Forecasting and Social Change*, 203, 123390.
- [2] Sakumoto, T., Hayashi, T., Sakaji, H., & Nonaka, H. (2024). Metadata-Based Clustering and Selection of Metadata Items for Similar Dataset Discovery and Data Combination Tasks. *IEEE Access*.
- [3] Yamamoto, Y., Hentona, A., Marusaki, K., Watabe, K., Kawano, S., Goto, T., ... & Nonaka, H. (2021, December). Development of the patent values evaluation method considering growth of technical community. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-6). IEEE.
- [4] 日浦隆博, 日浦隆博, 吉田奈央, 松井陽子, 河野誠也, 河野誠也, 野中尋史, 吉野幸一郎, 吉野幸一郎. (2024). 科学知識発見を目的とした特許のアノテーション. *言語処理学会年次大会発表論文集(Web) 30th*.
- [5] 安藤一哉, 溝口月斗, 石川治樹, 伊豫田彬, 河野誠也, 吉野幸一郎, 野中尋史. (2024). 特許を対象とする画像言語モデル開発のためのデータセットの構築. In *人工知能学会全国大会論文集 第38回 (2024)* (pp. 3Xin2103-3Xin2103). 一般社団法人 人工知能学会.
- [6] 西尾紗也香, 野中尋史, 早矢仕晃章, 坂地泰紀, & 作本猛. (2024). LLM およびネットワーク解析を利用した学術論文からの研究目的・機械学習モデル名・データセット名の抽出と相互の関連性分析. In *人工知能学会全国大会論文集 第38回 (2024)* (pp. 3Xin2110-3Xin2110). 一般社団法人 人工知能学会.

リサーチマップのURL

※<https://researchmap.jp/7000022337>

ご清聴ありがとうございました