# インターネット上のデータからの人物の人気度及び知名度の推定

## Estimating Individuals' Popularity and Name Recognition from Internet Data

ソーシャルメディアユニット
平尾 喜洋 / Yoshihiro Hirao

## 研究目的

- テレビタレントの **知名度・人気度・それらの変化量** を推定
- データ：インターネット（Wikipedia, ニュースなど）
- 従来の質問調査法に依らない推定
- 応用先：キャスティング最適化、広告効果測定、学術的分析

# 定義

- **知名度 (Awareness)**

$$\text{知名度} = \frac{\text{そのタレントを知っている人数}}{\text{調査対象者数}}$$

- **人気度 (Popularity)**

$$\text{人気度} = \frac{\text{「非常に好き」または「やや好き」と回答した人数}}{\text{調査対象者数}}$$

# 使用データセット

- **タレントリサーチ (Video Research)**
  - 2022年7月, 2023年1月の人気度・知名度調査
- **Wikipedia Clickstream (2022/8〜2023/1)**
  - Wikipediaのページ間遷移数を記録
- **Wikipedia Page View数 (2022/8〜2023/1)**
  - 各タレントのWikipediaページの日次アクセス数
- **ニュースリスト (2022/8〜2023/1)**
  - タレント名を含むニュースをGNews APIで収集、 全152,261件

# 手法① Wikipedia Clickstreamを用いた推定

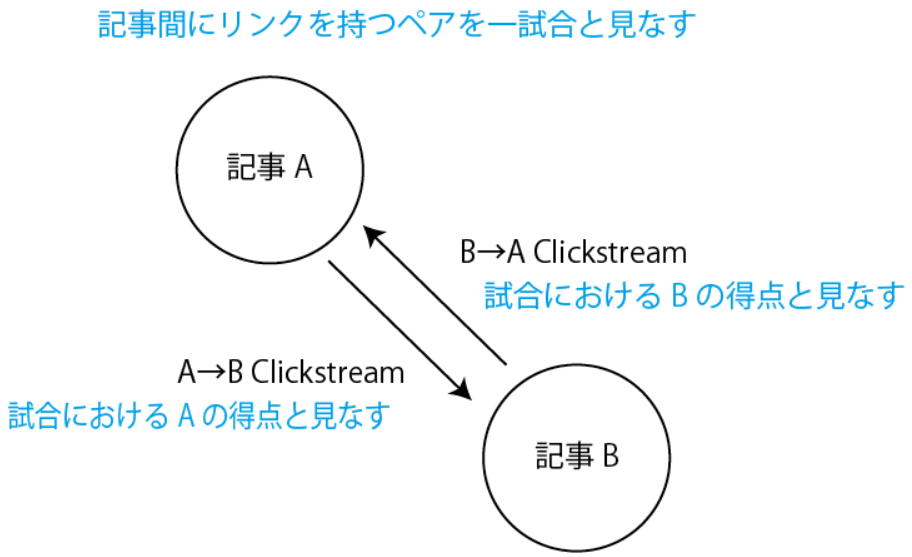
知名度（人気度）の高い人物から低い人物にアクセスが流れやすいという仮定のもと、複数の手法を適用

- **レイティング手法**
  - スポーツチームにおけるレイティング手法をWikipedia Clickstreamのデータに適用できるよう改良
- **PageRank**
  - ウェブページの重要度を測る指標PageRankをWikipedia Clickstreamのデータに適用できるよう改良
- **BrowseRank**
  - ウェブページの重要度を測る指標BrowseRankをWikipedia Clickstreamのデータに適用できるよう改良

# レイティング手法

Massey、Colley、KeenerのレイティングをWikipedia Clickstreamに適用できるよう改良

これらの手法は本来スポーツチームの順位付けで用いられるもので、Wikipediaにおけるリンク構造とClickstreamを右図のように見なす



## Masseyのレイティング手法（改良版）

レイティングベクトル$r$を求める

- $(m_{ij})$は本来各チーム間の試合数を表す行列
- $p$ は本来各チームの累積得点差を表すベクトル

$$m_{ij} = \begin{cases} \text{リンク数} & (i = j) \\ 0 & (i, j\text{にリンクなし}) \\ -1 & (i, j\text{にリンクあり}) \end{cases}$$

$$p_i = \sum_j \big(\text{アクセス数}(i \to j) - \text{アクセス数}(j \to i)\big)$$

$$(m_{ij})r = p$$

Massey (1997), Statistical models applied to the rating of sports teams, Bluefield College

## Colleyのレイティング手法（改良版）

レイティングベクトル$r$を求める

- $(m_{ij})$は本来各チーム間の試合数を表す行列
- $b$ は本来各チームの累積勝敗差を表すベクトル

$$m_{ij} = \begin{cases} \text{リンク数} + 2 & (i = j) \\ 0 & (i, j\text{にリンクなし}) \\ -1 & (i, j\text{にリンクあり}) \end{cases}$$

$$b_i = \sum_j \begin{cases} 1 & \left(\text{アクセス数}(i \to j) > \text{アクセス数}(j \to i)\right) \\ 0 & \text{otherwise} \end{cases}$$

$$(m_{ij})r = b$$

Colley (2002), Colley's bias free college football ranking method, Princeton University

## Keenerのレイティング手法（改良版）

- $(a_{ij})$は本来得点率を表す行列

$$a_{ij} = \frac{\text{アクセス数}(i \to j) + 1}{\text{アクセス数}(i \to j) + \text{アクセス数}(j \to i) + 2}$$

行列$(a_{ij})$の固有値問題に帰着 → レイティング算出

## レイティング手法を適用する問題点

記事間にリンクが張られているタレント群でしか計算できず、単にWikipediaに記事が存在するタレント数よりも分析できる数が大幅に減ってしまう

Keener (1993), The Perron–Frobenius theorem and the ranking of football teams, SIAM Review

# PageRank（有向重み付き版）

- **基本モデル**
  - ランダムサーファーモデルを導入
  - ダンピング係数 $d = 0.85$

$$\pi^{(t+1)}(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{\pi^{(t)}(j)}{L(j)}$$

- **本研究での改良**
  - Wikipedia Clickstreamに対応
  - リンクに重みとしてClickstreamまたはその逆数を正規化したものを付与

$$\pi^{(t+1)}(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \pi^{(t)}(j) \cdot w_{ji}$$

Page et al. (1999), The PageRank citation ranking: Bringing order to the web, Stanford InfoLab

# BrowseRank（改良版）

- ユーザ行動をマルコフモデルで表現

## 滞在時間の仮定

$$f_i(t) = \theta_i e^{-\theta_i t}$$

## Q行列

$$q_{ij} = \theta_i P_{ij}, \quad (i \neq j)$$
$$q_{ii} = -\theta_i$$

## 定常分布

$$\pi Q = 0$$

Liu et al. (2008), BrowseRank: Letting web users vote for page importance, SIGIR

## BrowseRank：DMCへの変換

連続時間マルコフ過程は扱いが複雑 → **離散時間マルコフ連鎖（DMC）** へ変換

### ステップ1：ダミーノードの追加

- セッション終了やジャンプ先不明を吸収するノード

### ステップ2：遷移確率行列の構築

$$P_{ij} = \frac{N_{ij}}{N_i}$$

- $N_{ij}$：ページ $i$ から $j$ への遷移回数
- $N_i$：ページ $i$ からの遷移総数

### ステップ3：定常分布の計算

$$\tilde{\pi} = \tilde{\pi} P$$

Liu et al. (2008), BrowseRank: Letting web users vote for page importance, SIGIR

## ステップ4：BrowseRankの算出

$$\pi_i = \frac{1}{\theta_i}\tilde{\pi}_i \Big/ \sum_j \frac{1}{\theta_j}\tilde{\pi}_j$$

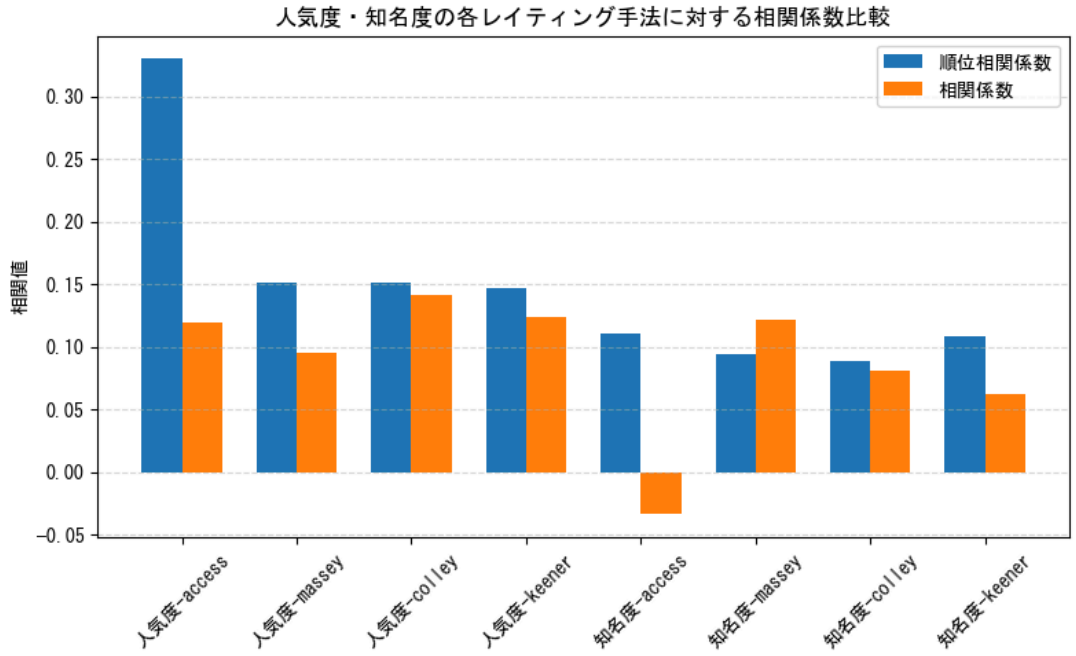
- $\theta_i$: ページ $i$ の滞在率（逆数が平均滞在時間）

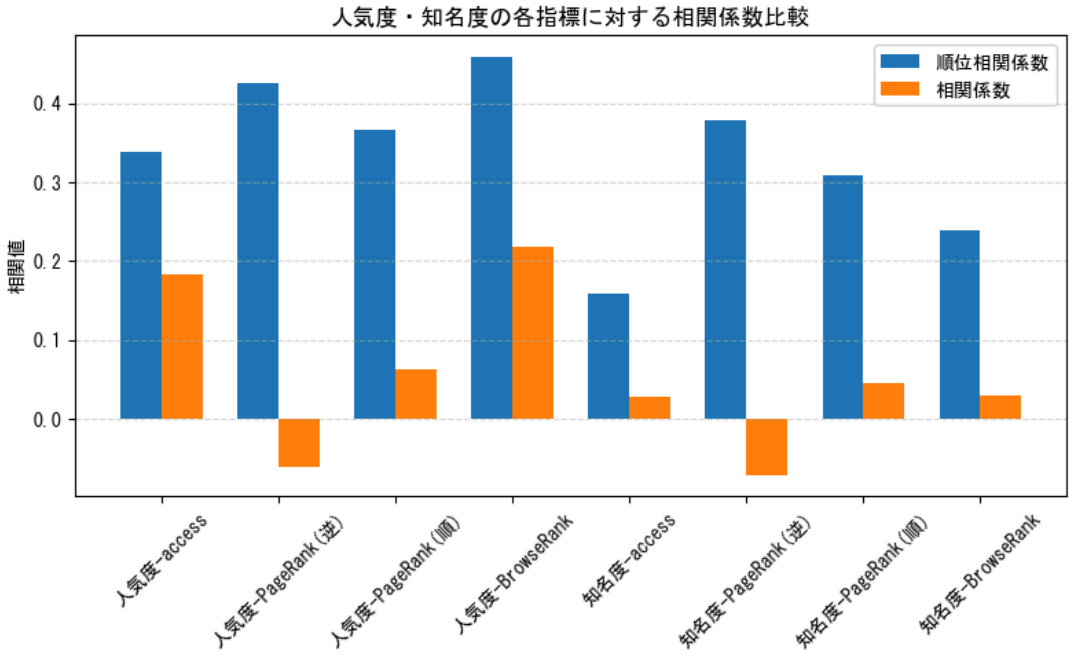
- **本研究での改良**
  - Wikipedia Clickstreamに対応
  - 遷移率をClickstreamの逆数で正規化
  - 疑似ノード導入
  - DirectAccess数を滞在時間として利用

Liu et al. (2008), BrowseRank: Letting web users vote for page importance, SIGIR

# 実験結果（Clickstream）

- レイティング手法：全体的に精度が低い
- PageRank：順位相関に関してはDirectAccessより精度が高い
- BrowseRank：順位相関、相関共にDirectAccessより精度が高い



# 手法② ニュースを用いた推定

## バースト検出

- 判定条件

$$\text{その日のPV} > 2 \times \text{直近7日平均PV}$$

## ニュースの影響度（インパクトスコア）

- 各記事タイトルに対して $-5 \sim +5$ のスコア
- Gemini 2.0 Flashで算出

## この手法の問題点

- 内容が全く分からないようなタイトルのニュースが存在
- LLMは知名度の低いタレントに関するスコアリングの精度が低い

## Publisherスコア

ニュース媒体ごとの影響力を定量化

- **Publisherスコア1**：バースト日数の合計を発行ニュース数で正規化した値
  （1 記事あたり，どの程度長期間にわたって注目を集める傾向があるかを表す）
- **Publisherスコア2**：バースト日の Page View 数合計を発行ニュース数で正規化した値
  （1 記事あたりの注目度の大きさを表す）
- **Publisherスコア3**：バースト日の当日と前日に発行されたニュース数合計を発行ニュース数で正規化した値
  （バーストを引き起こしやすい媒体かどうかを表す）

# 特徴量抽出

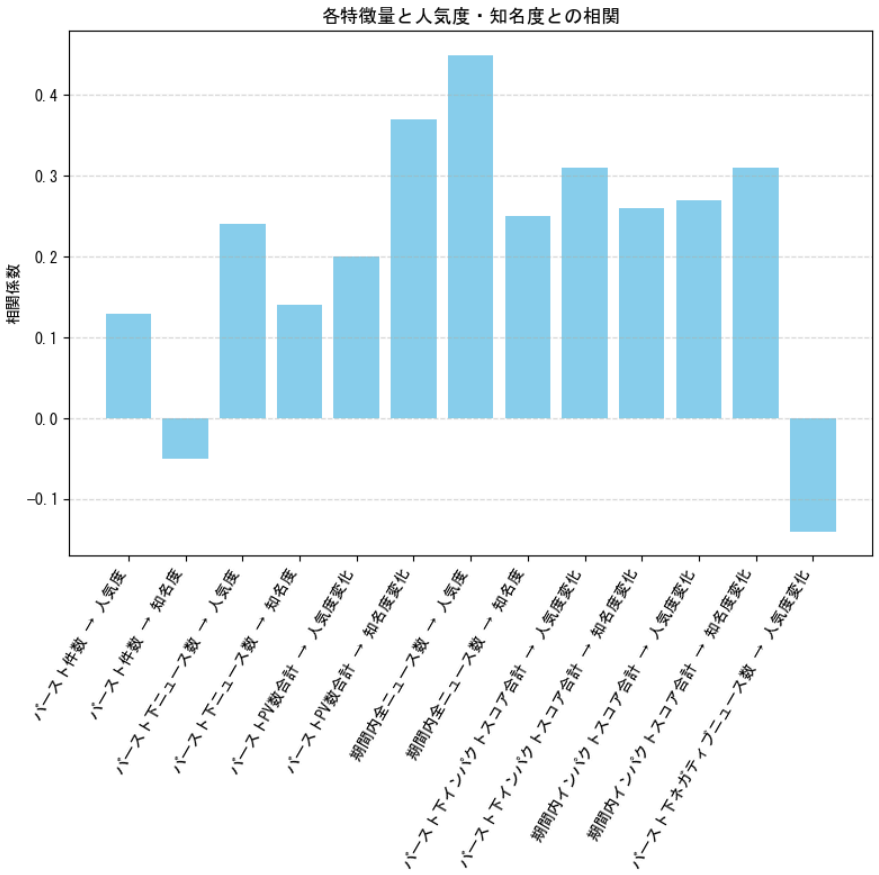
ここでのニュースとは、各タレントの名前を含むニュース

- 各調査時点の人気度/知名度
- 人気度変化/知名度変化
- 期間内に起きたバーストの件数
- バーストした日のPage View数合計
- バースト当日と前日に発行されたニュース数合計
- 期間内に発行された全ニュース数合計
- バースト下ニュースのインパクトスコア合計
- 期間内全ニュースのインパクトスコア合計
- インパクトスコア正のニュース数/インパクトスコア負のニュース数
- 各Publisherスコア合計

# 実験結果（ニュース）

代表的な相関関係をピックアップ

- **バースト件数 - 人気度・知名度：**
  - 相関ほぼなし
- **バースト下ニュース数 - 人気度：**
  - 知名度よりも高い相関
- **バーストPV数合計 - 人気度変化・知名度変化：**
  - 特に知名度変化と正の相関
- **期間内全ニュース数 - 人気度・知名度：**
  - 特に知名度と正の相関
- **インパクトスコア合計 - 人気度変化・知名度変化：**
  - 正の相関
- **ネガティブニュース数 - 人気度変化：**
  - 弱い負の相関



# 回帰分析

- 説明変数3つの組合せで線形回帰及びランダムフォレスト回帰
- $R^2$スコア上位6組を比較

結果：

- 知名度変化以外の指標においてランダムフォレスト回帰の方が精度が高い
- 人気度推定 → BrowseRank、Publisherスコアが重要
- 知名度推定 → ニュースインパクトスコア、Publisherスコアが重要
- 人気度変化 → バースト規模、BrowseRank、Publisherスコアが重要
- 知名度変化 → Publisherスコア、PageView数合計が重要

# 分析結果を踏まえた知名度の時間的変化モデル構築

- 回帰分析の結果より
  - 人気度変化：予測困難
  - 知名度変化：一定の説明力
- 本実験では
  **知名度変化の数理モデル化**に焦点

## 知名度変化モデルの基本仮定

- 知名度は最大値 100 の指標
- $(100 - \text{元の知名度})$ を「伸びしろ」と解釈
- 各時点の元の知名度は既知とする

## 知名度変化モデル（基本式）

$$\text{知名度変化量} = (100 - \text{元の知名度})(a \cdot \mathrm{PV} + b \cdot \mathrm{Publisher} + c) + d$$

- PV：期間内 Page View 数
- Publisher：Publisher スコア
- $a, b, c, d$：推定係数

# 各知名度変化モデルにおける決定係数

| モデル | LOO $R^2$ | 10-Fold $R^2$ |
|---|---|---|
| $(100 - \text{元の知名度})(a\,\mathrm{PV} + b\,\mathrm{Publisher} + c) + d$ | **0.3250** | **0.2828** |
| $(100 - \text{元の知名度})(a\,\mathrm{PV} + b) + c$ | 0.3220 | 0.2771 |
| $(100 - \text{元の知名度})(a\,\mathrm{Publisher} + b) + c$ | 0.2132 | 0.1888 |
| $(100 - \text{元の知名度})(a\,\mathrm{PV} + b) + c\,\mathrm{Publisher} + d$ | 0.3209 | 0.2763 |
| $(100 - \text{元の知名度})(a\,\mathrm{Publisher} + b) + c\,\mathrm{PV} + d$ | 0.2639 | 0.2240 |
| $a(100 - \text{元の知名度}) + b\,\mathrm{PV} + c\,\mathrm{Publisher} + d$ | 0.2395 | 0.1903 |
| $a(100 - \text{元の知名度}) + b\,\mathrm{PV} + c$ | 0.2334 | 0.1859 |
| $a(100 - \text{元の知名度}) + b\,\mathrm{Publisher} + c$ | 0.0836 | 0.0653 |
| $a\,\mathrm{PV} + b\,\mathrm{Publisher} + c$ | 0.2206 | 0.1693 |

## 基本モデル

$$\text{知名度変化量} = (100 - \text{元の知名度})(a \cdot \mathrm{PV} + b \cdot \mathrm{Publisher} + c) + d$$

## 簡略モデル

$$\text{知名度変化量} = (100 - \text{元の知名度})(a \cdot \mathrm{PV} + b) + c$$

## PageView に時系列減衰を導入

- 単純な PV 合計では時間構造を捉えきれない
- 時系列的減衰を導入したWeightedPV を設計
- 3種類の減衰モデルを比較

## 忘却曲線に基づく減衰モデル（WeightedPV1）

$$w(d) = \frac{1.84}{\log_{10}(d \times 24 \times 60) + 1.25 + 1.84}$$

- $d$：イベント発生日からの経過日数
- Ebbinghaus の忘却曲線に基づく

## 指数減衰モデル（WeightedPV2）

$$w(d) = e^{-0.4d}$$

- イベント直後の急激な関心低下を表現
- 短期的話題性を強調

## 二相集団記憶減衰モデル（WeightedPV3）

$$w(d) = e^{-0.4d} + d^{-0.3}$$

- 第1項：短期記憶（指数減衰）
- 第2項：長期記憶（冪則減衰）

# WeightedPV を用いたモデル精度

| モデル | LOO $R^2$ | 10-Fold $R^2$ |
|---|---|---|
| 基本モデル | 0.3250 | 0.2828 |
| WeightedPV1 基本 | 0.3250 | 0.2823 |
| WeightedPV2 基本 | 0.2132 | 0.1840 |
| WeightedPV3 基本 | 0.3168 | 0.2719 |
| 簡略モデル | 0.3220 | 0.2771 |
| WeightedPV1 簡略 | 0.3218 | 0.2765 |
| WeightedPV2 簡略 | 0.0713 | 0.0358 |
| WeightedPV3 簡略 | 0.3113 | 0.2633 |

# 更に過去のデータを用いた比較

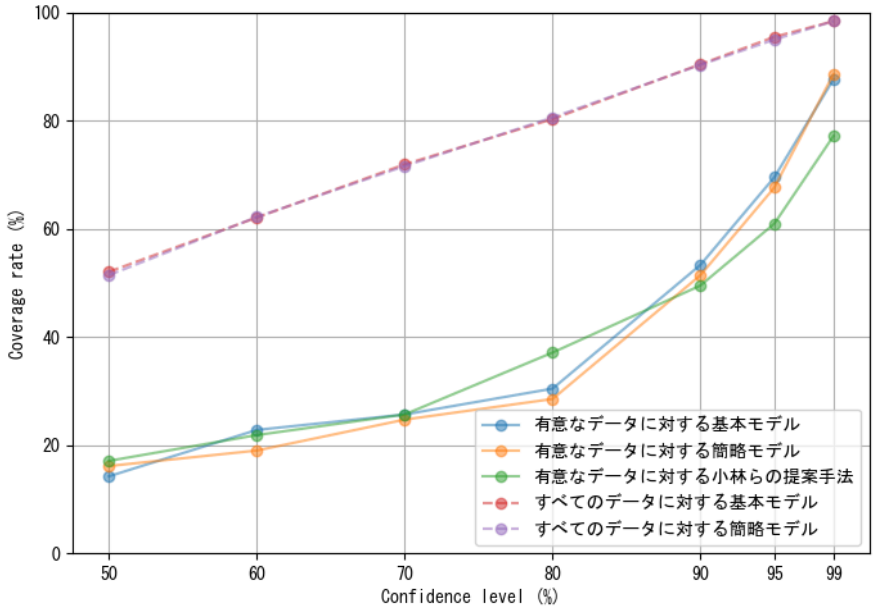
3年半のデータで検証

| **観測期間** | LOO $R^2$ | 10-Fold $R^2$ |
|---|---|---|
| 基本 | **0.6775** | **0.6526** |
| 簡略 | **0.6490** | **0.6150** |

- 観測期間を長くすると精度が大幅向上
- 基本モデルと簡略モデルの精度差が拡大
  - 期間が長期になるほど Publisher スコアの寄与が大きい
- 知名度変化は**中長期的指標**であることを示唆

# 先行研究との比較

- ニュース解析中心の先行研究
- 本研究：
  - PageView に基づくモデル
  - 数式による時間推移表現
- 母比率の差の検定を有意水準5%で実施し，有意なデータを判別
- 予測値が正解変化量の信頼区間に含まれる割合を算出
- 同等以上の精度＋高い再現性



## 本手法の考察

- 知名度変化モデルを構築
- 時系列減衰を導入した WeightedPV を比較
- 結果：
  - 複雑な減衰モデルは必ずしも有効でない
  - **単純な累積 PageView × 長期間観測が最も安定**

## 今後の展望

- 線形モデルの制約を超えた**非線形・時系列モデル**の導入
- SNS や記事内容など，**追加データを統合した分析**
- 人物評価の**継続的モニタリング**への応用

# まとめ

- Wikipedia Clickstream
  - レイティング・PageRank・BrowseRankを適用
  - しかし全体的に精度は限定的
- ニュース＋バースト＋LLMスコア
  - 人気度・知名度推定：ランダムフォレスト回帰分析において一定の精度
  - 知名度変化推定：PageView・Publisherスコアが重要，線形回帰分析において一定の精度
- 知名度変化の数理モデル構築
  - PageView・Publisher 指標を用いて線形回帰モデルにより分析を実施
  - **単純な累積 PageView と長期間観測が有効**であることを示した