歴史ビッグデータ：構造化ギャップを克服するワークフローの構築と過去世界の統合解析

北本 朝展、市野 美夏

ROIS-DS人文学オープンデータ共同利用センター CODH／国立情報学研究所

http://codh.rois.ac.jp/

@rois_codh

人文学オープンデータ共同利用センター

CODH http://codh.rois.ac.jp/

- **情報・システム研究機構 データサイエンス共同利用基盤施設内に、2017年4月1日に正式に発足。センター長：北本 朝展。**

1. **データ駆動型人文学：情報学・統計学の技術により、人文学の研究を革新。**
2. **人文学ビッグデータ：人文学のデータにより、非人文学の研究を革新。**

歴史ビッグデータセミナー

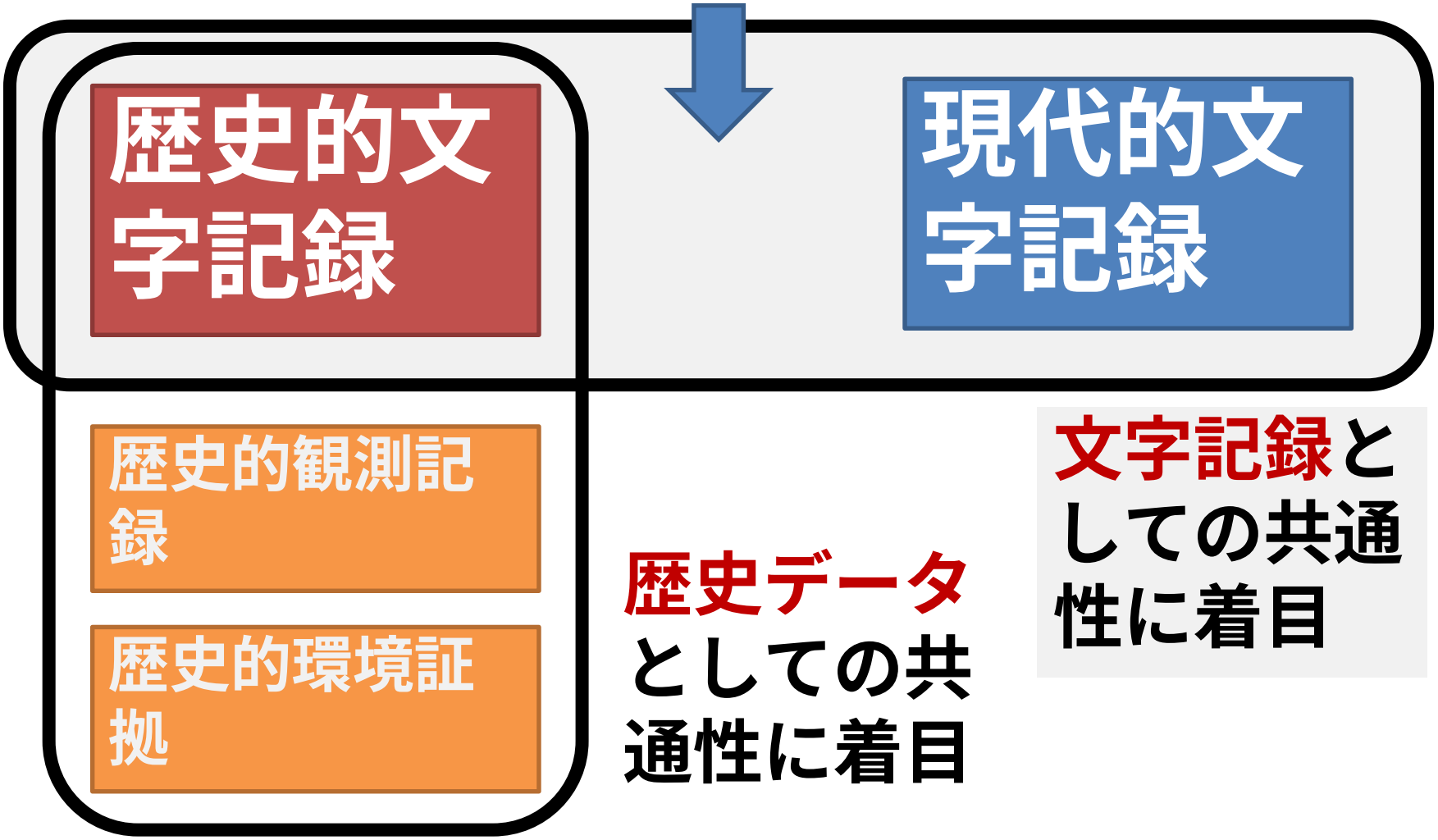
http://codh.rois.ac.jp/seminar/historical-big-data-20180312/

2018年3月12日開催 参加者96名

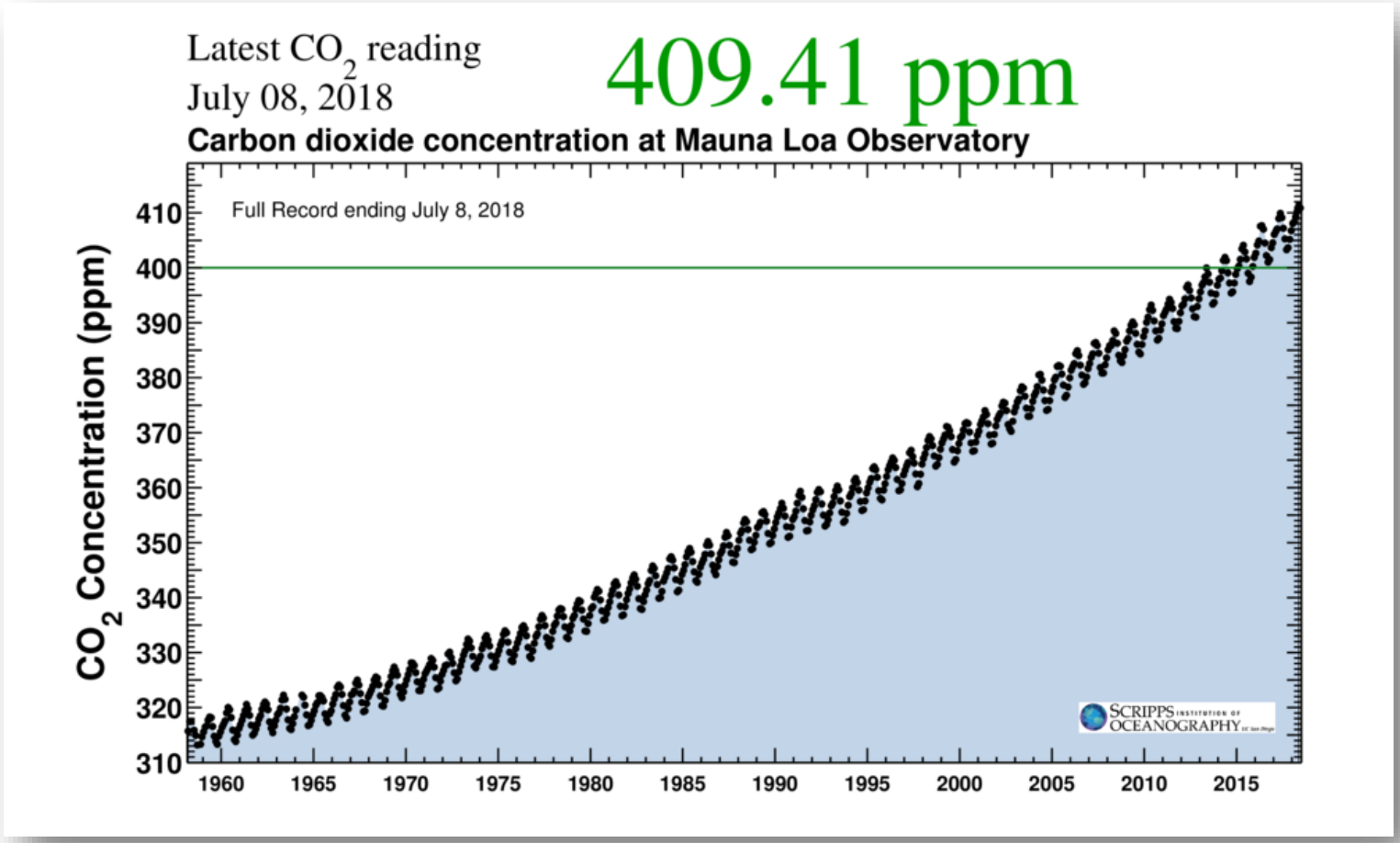
13:00	開場	
13:30-13:45	歴史ビッグデータと歴史的状況記録	北本 朝展（CODH/NII）、市野 美夏（CODH）
13:45-14:15	みんなで翻刻と古地震研究	加納 靖之（京都大学防災研究所）
14:15-14:45	古文書調査と自然現象	吉川 聡（奈良文化財研究所）
14:45-14:55	休憩	
14:55-15:25	古文書の気象・災害記録をどう活かすか―仙台・宮城での史料保全をふまえて	佐藤 大介（東北大学災害科学国際研究所）
15:25-15:55	株井戸－気候復元を活用した地下水管理制度の研究―	遠藤 崇浩（大阪府立大学）
15:55-16:25	近世日本の中央市場と気候変動	柴本 昌彦（神戸大学経済経営研究所）
16:25-16:35	休憩	
16:35-16:55	古気候復元のための日記天気記録の定量化に向けて	庄 建治朗（名古屋工業大学）
16:55-17:25	ミレニアム大気再解析への挑戦～古日記に記載された天気情報のデータ同化～	芳村 圭（東京大学生産技術研究所）
17:25-18:00	ディスカッション	講演者全員

歴史ビッグデータ

http://codh.rois.ac.jp/historical-big-data/

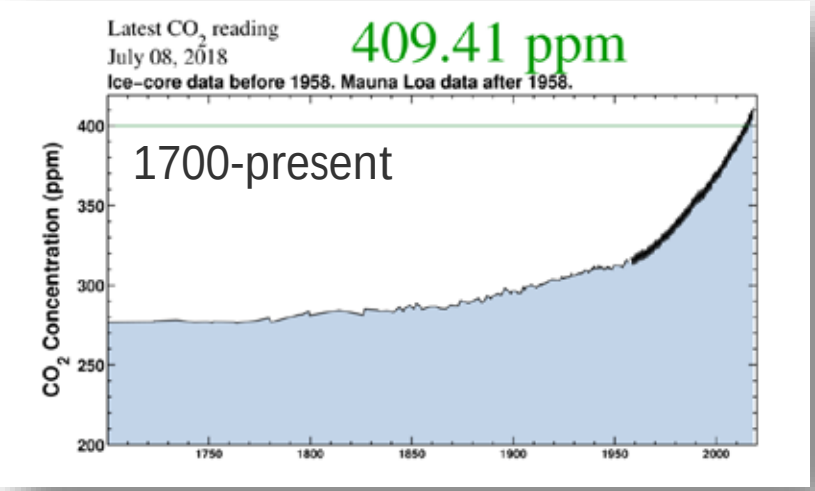


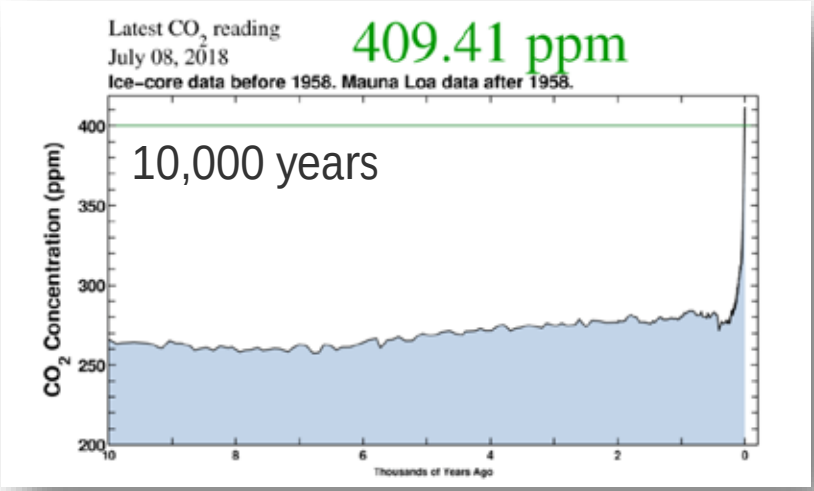
歴史的観測記録

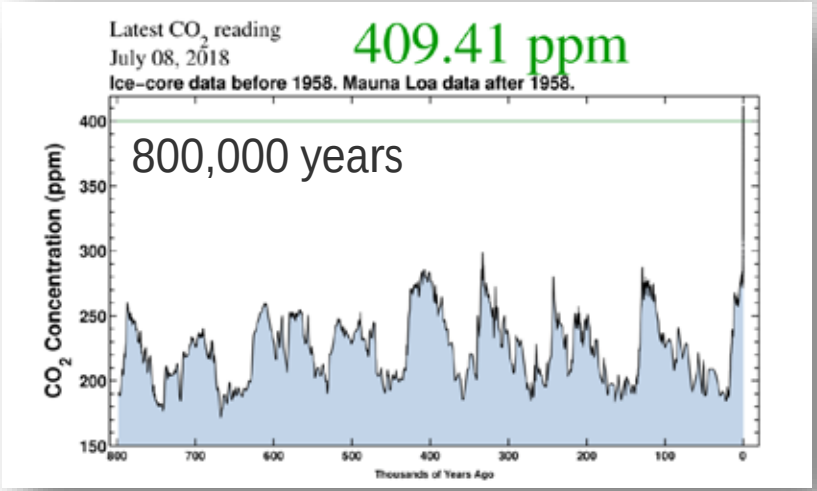


THE KEELING CURVE: https://scripps.ucsd.edu/programs/keelingcurve/

歴史的環境証拠



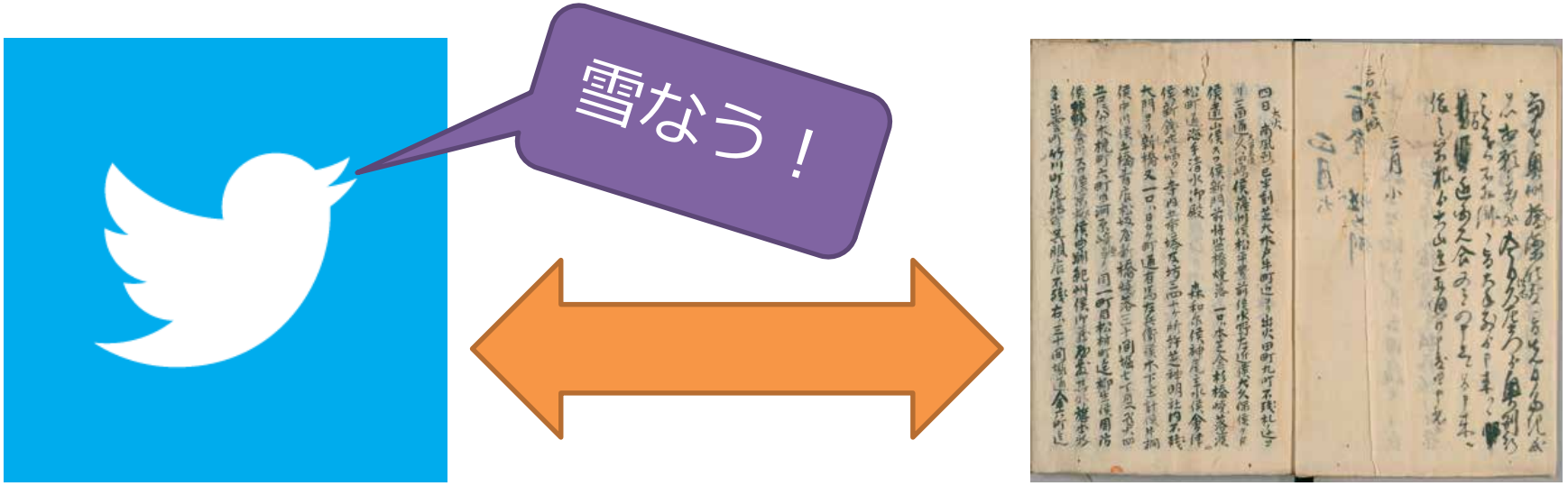




After 1958, observation at Mauna Loa; before 1958 ice-core data.

THE KEELING CURVE:
https://scripps.ucsd.edu/programs/keelingcurve/

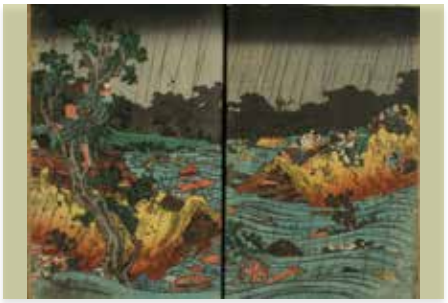
歴史的文字記録



1. ツイートと日記はどのように違うのか？
2. 記述の信頼性はどのくらい違うのか？
3. **構造化プロセスに現在向けのアルゴリズム**（NLP等）がどのくらい利用できるのか？

歴史的文書に基づく過去復元


＊1


＊1


＊2


＊1

地震

天気

季節

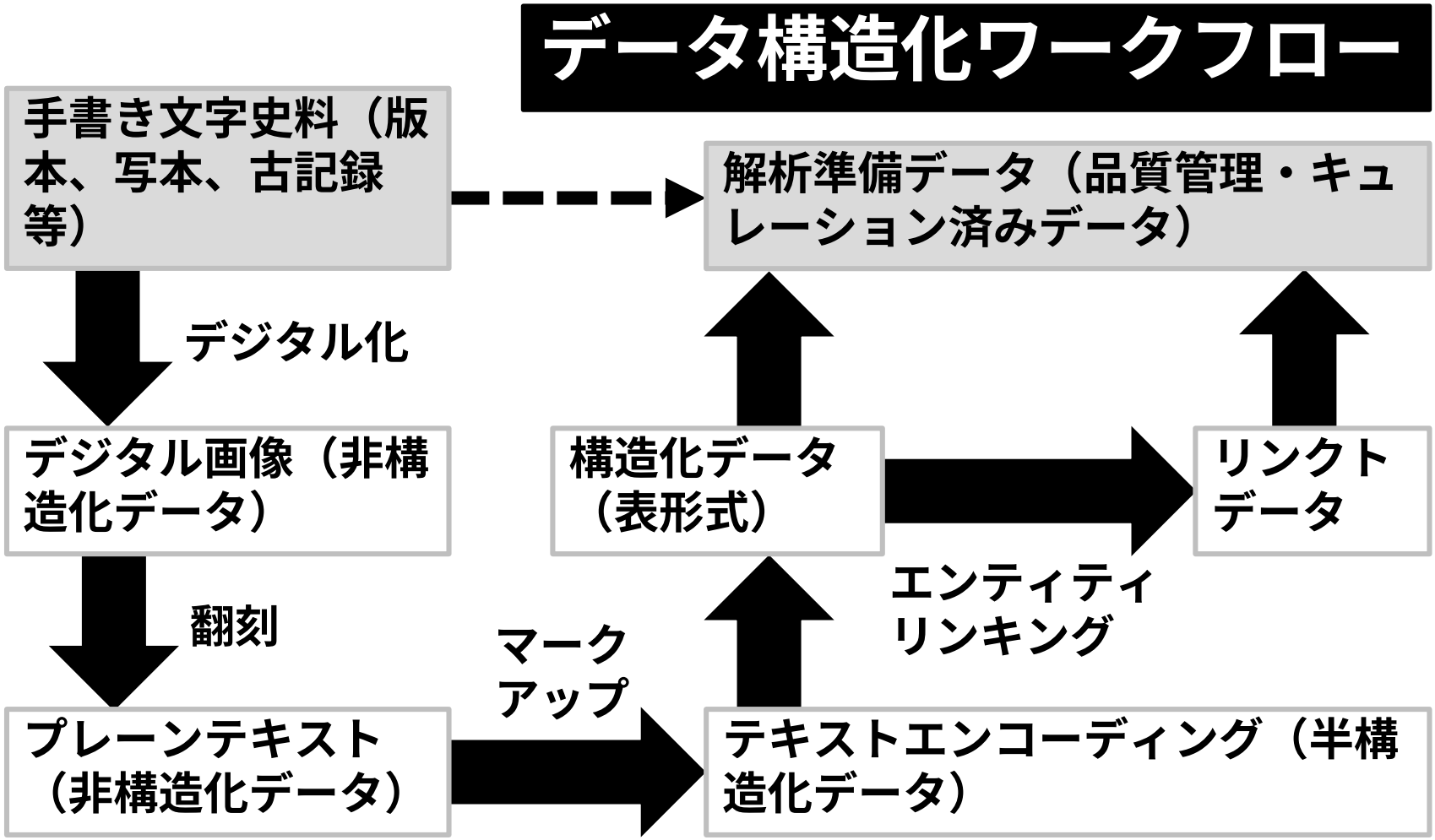
火山噴火

歴史ビッグデータ

歴史資料から得られた情報を統合解析する

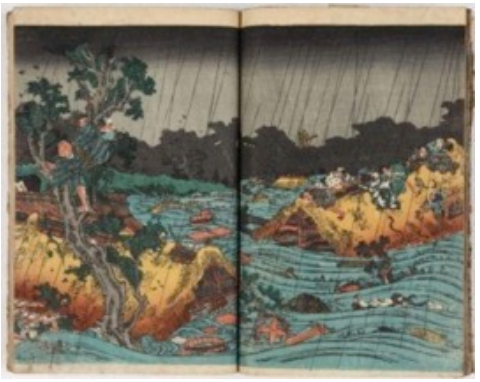
出典：＊1早稲田大学図書館古典籍総合データベース・＊2国立国会図書館デジタルコレクション

課題：データ構造化ギャップ



史料探査の改善

?



安政江戸台風のデータを集めたい。でも、どの資料に、どんなデータがあるのだろう？

1. **検索サイトが存在しない。**
2. **くずし字は全文検索もできない。**
3. **図書館などを一つずつ回り、一冊ずつ読んで探すしかない。。**

れきすけ

「れきすけ」の提案

利用してほしい人	探している人

歴史資料に〇〇に関する記録を見つけた
情報を伝えてもっと
資料を利用してほしい

図書館員など

利用実績

研究に利用できそうな
資料を探す
〇〇に関する記録が
ある資料を見つけたい

研究者など

資料の情報を
登録

資料の情報を
検索・取得

山脇弁治日記
秋田県立公文書館

別所万右衛門記録
東北大学佐藤大介氏

Time Information System HuTime

国際日本文化研究センター
関野樹氏

みんなで翻刻
https://honkoku.org/

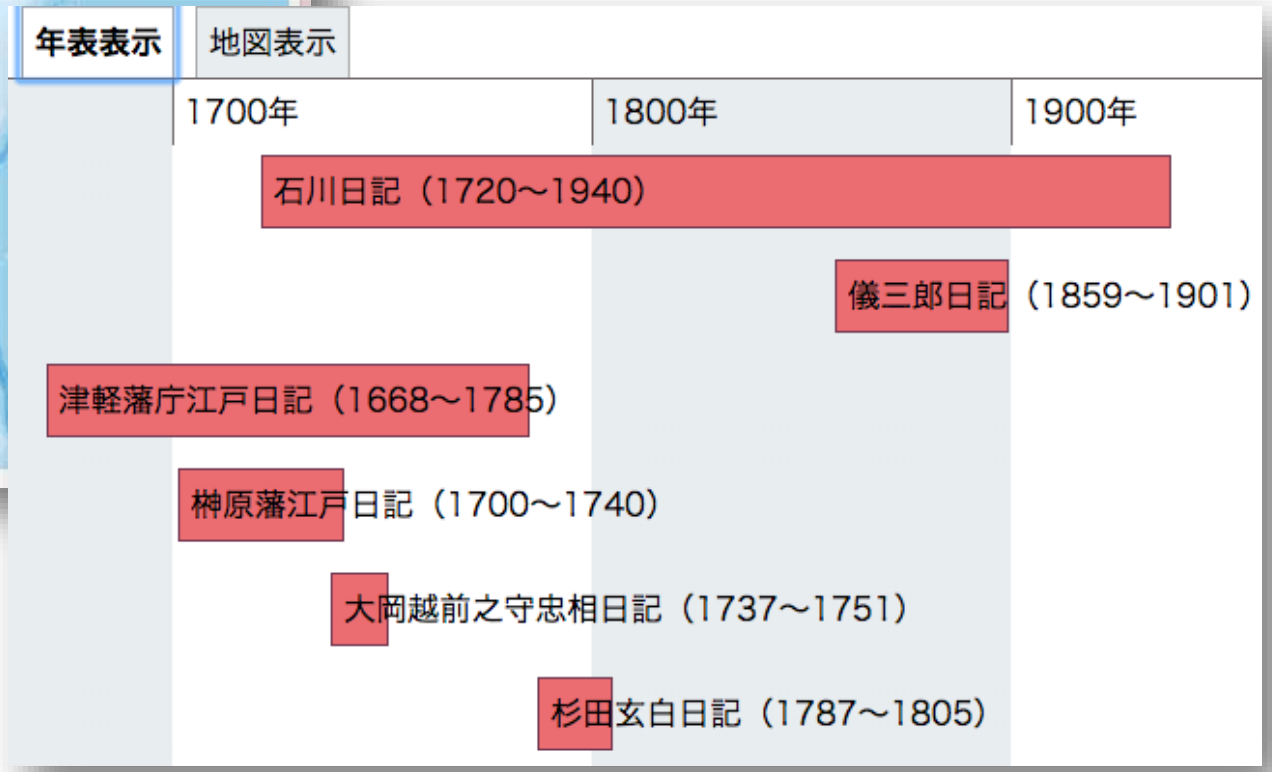
Koji
国立歴史民俗博物館
橋本雄太氏

さまざまな機関やプロジェクトと連携

れきすけ

地図・年表表示


検索結果リスト
年表表示
地図表示
+
−


年表表示
地図表示
1700年
1800年
1900年
石川日記（1720～1940）
儀三郎日記（1859～1901）
津軽藩庁江戸日記（1668～1785）
榊原藩江戸日記（1700～1740）
大岡越前之守忠相日記（1737～1751）
杉田玄白日記（1787～1805）

大量の文書⇔少数の読者

数億点以上

日本に残る古典籍・古文書等の推定点数

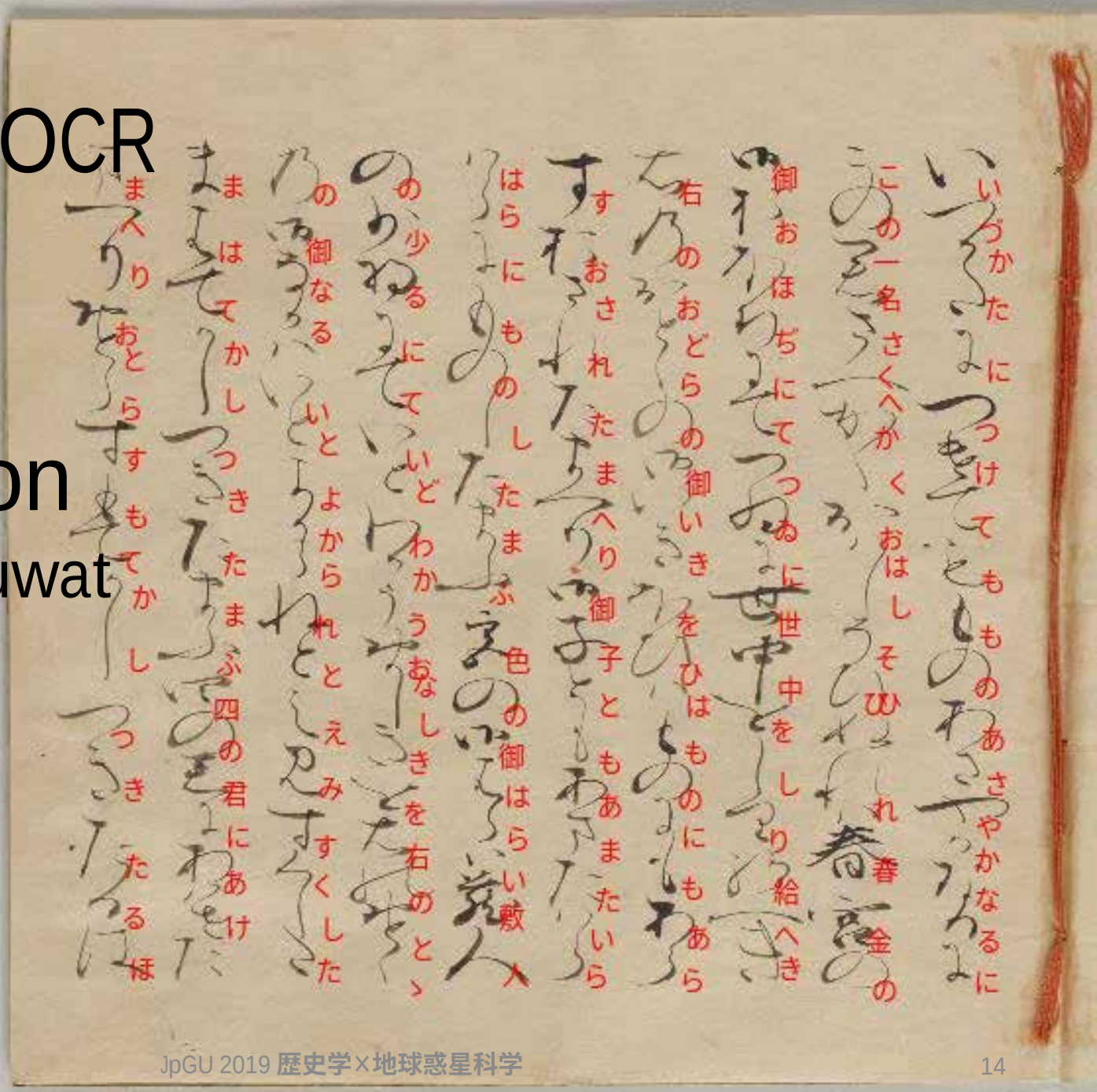
数千～数万人

くずし字がスラスラ読める推定人数

くずし字OCR

Optical Character Recognition

by Tarin Clanuwat



AIによるくずし字認識

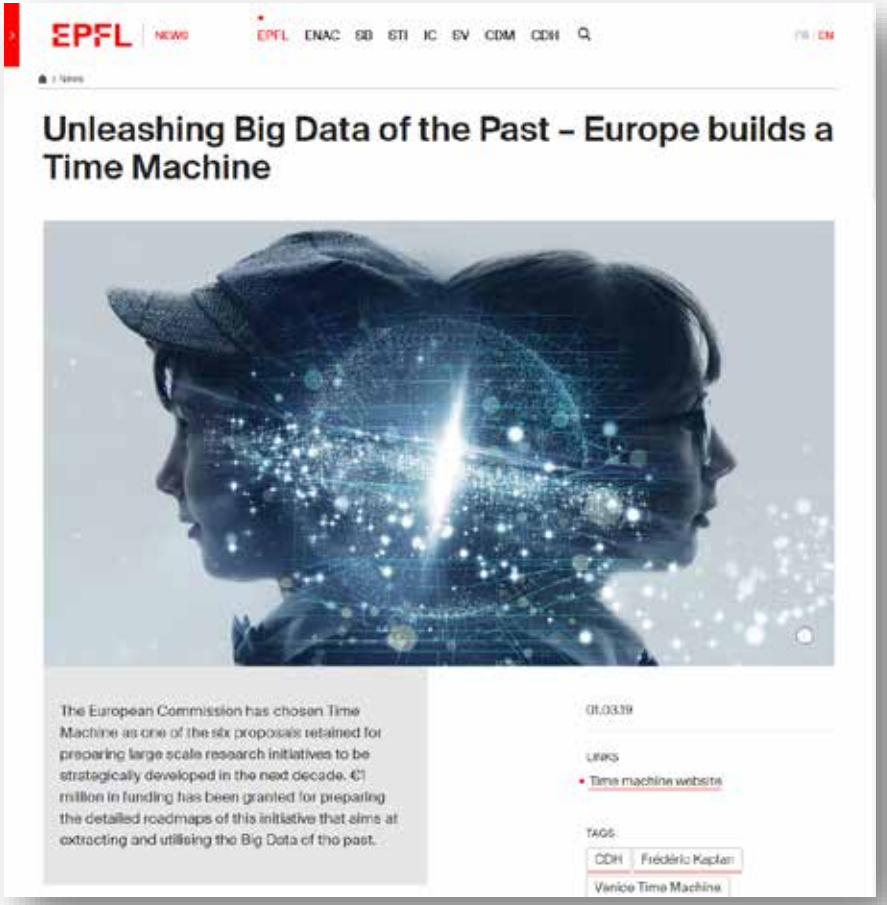
AIの利点

1. **人間が勉強するより、機械が学習したほうが早い？**
2. **今後も性能の向上が期待できる。**
3. **文字がどれだけ多くても、AIは疲れを知らない！**

AIの欠点

1. **決して100%の性能は達成できない。**
2. **学習用のデータセットを増やすのが大変。**
3. **人間が読まないと、意味を把握することはできない！**

AIと歴史ビッグデータ



https://actu.epfl.ch/news/unleashing-big-data-of-the-past-europe-builds-a-ti/

- 10年10億Euroの欧州超大型プロジェクトFET Flagshipの選考が進行中。
- Time Machineプロジェクトは、第二次選考を通過し、6グループに残った。

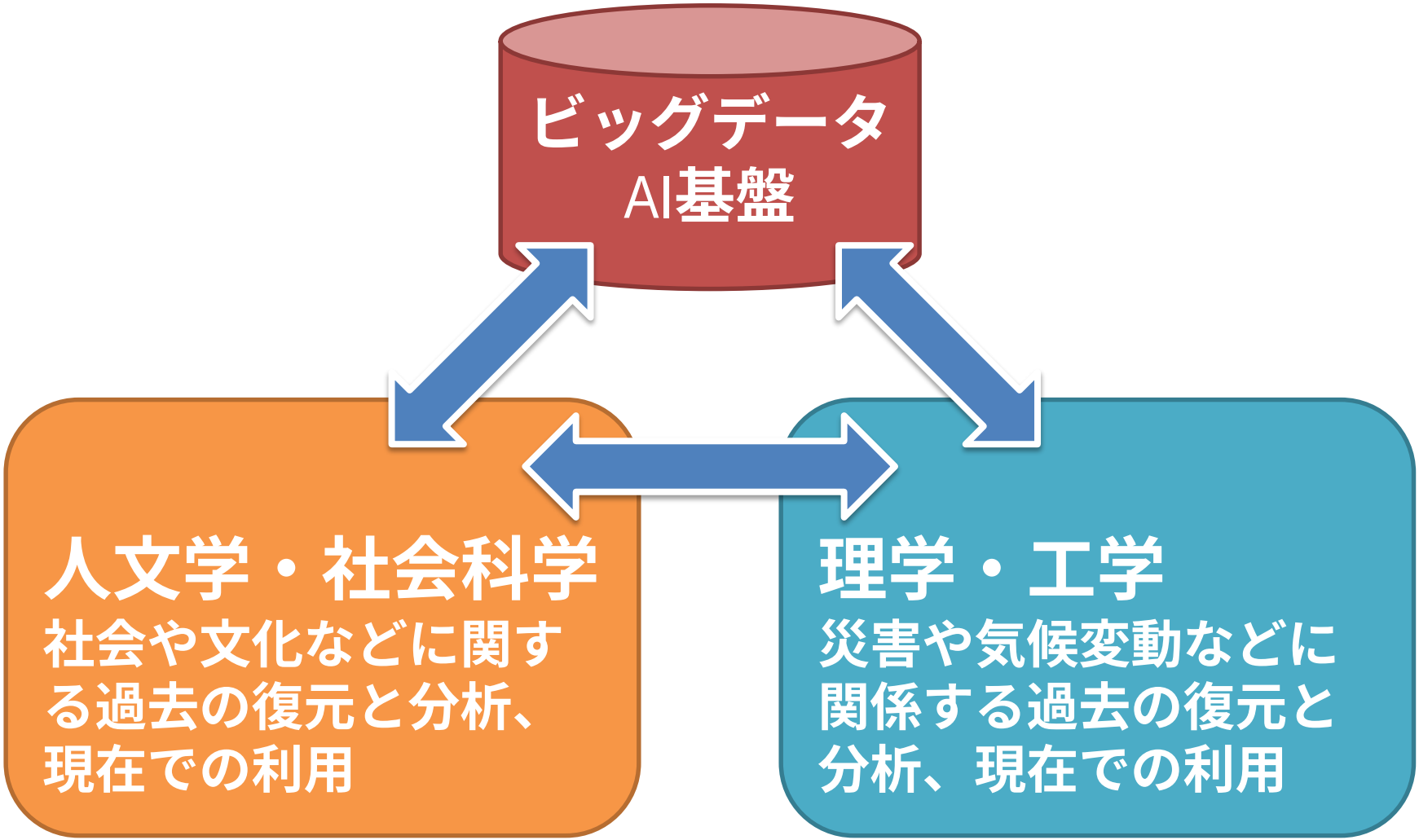
Time Machineのアピールポイント

1. One of the most advanced Artificial Intelligence systems ever built.
2. Cultural Heritage as a valuable economic asset.
3. A new age for Social Sciences and Humanities.
4. Transforming education.
5. A unique alliance and a network of cities.

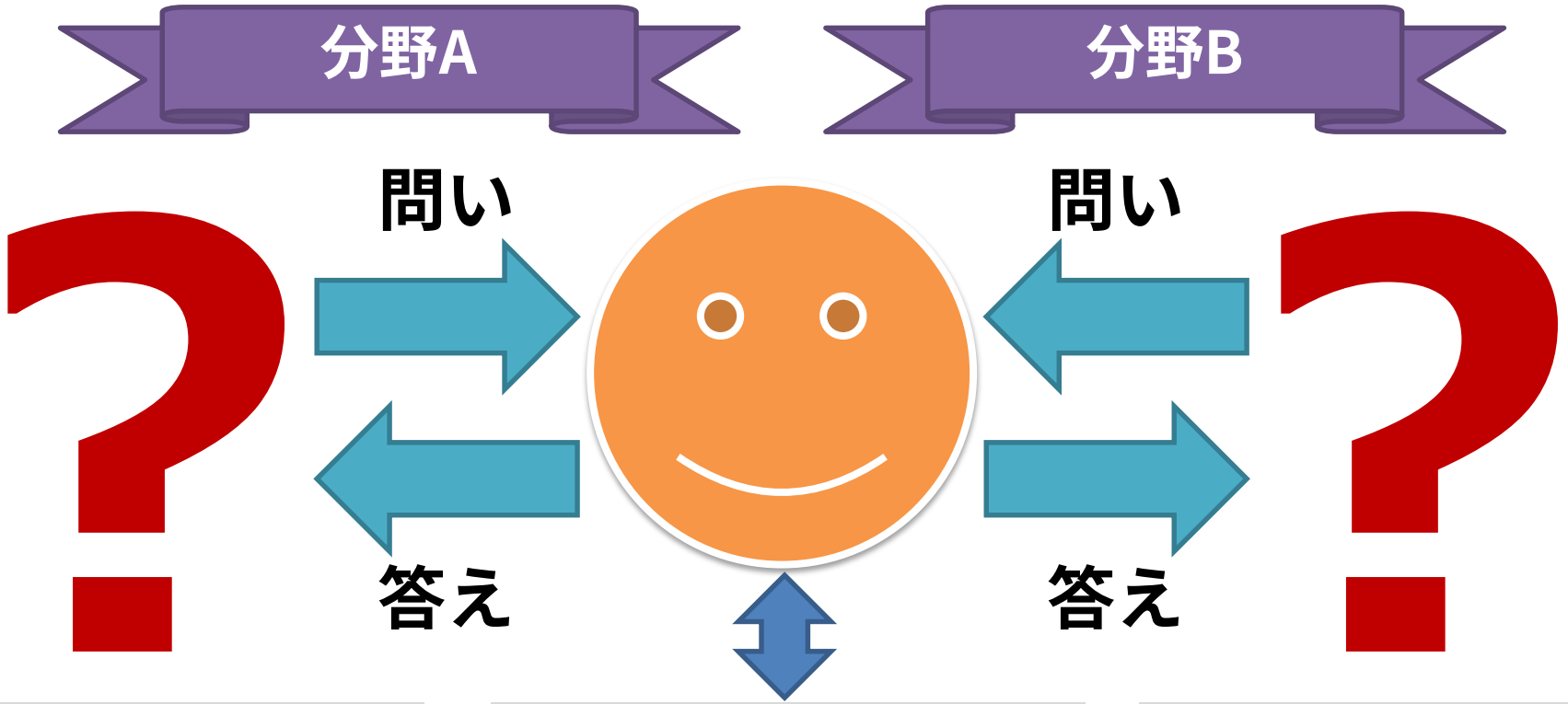
日本版タイムマシンとは？

1. **欧州では「タイムマシン」のアイデアが高く評価されている。**
2. **欧州文化遺産の<u>デジタル化とデータ駆動型研究</u>に加え、<u>教育や観光</u>なども視野。**
3. **<u>日本の文化</u>においては、<u>自然環境や災害（とその復興）</u>なども欠かせない。**
4. **日本の歴史ビッグデータは、<u>社会と自然との関係</u>**を扱うことが大事では？

社会と自然の関係


ビッグデータ
AI基盤
人文学・社会科学
社会や文化などに関する過去の復元と分析、現在での利用
理学・工学
災害や気候変動などに関係する過去の復元と分析、現在での利用

同床同夢？同床異夢？



分野Aと分野Bが相互に貢献しあう。

データ・ツール・メソッドの共有

それぞれの夢に挑める基盤を作る。

おわりに

1. 歴史ビッグデータは、**人間による文字記録を主な対象**とし、人間による観測や自然から得られる証拠も統合する。
2. 歴史ビッグデータでは、**非構造化データを半構造化データ、構造化データへと変換**する構造化ギャップの解消を目指す。
3. **欧州の**Time Machine**プロジェクト**の動向を受けて、日本でも可能性を探りたい。

関連情報

- **歴史ビッグデータ**
  - http://codh.rois.ac.jp/historical-big-data/
- **支援**
  - **人文学ビッグデータにおける構造化ギャップの克服と分野横断的利用の検証、機構間連携・文理融合プロジェクト**
  - **歴史ビッグデータ研究基盤による過去世界のデータ駆動型復元と統合解析、科研費基盤 (A)**
  - **「データ構造化ギャップ」に挑む特任研究員の募集を、今週から開始予定！**