# KuroNetくずし字認識と歴史ビッグデータ研究へのインパクト

北本 朝展　カラーヌワット タリン

ROIS-DS人文学オープンデータ共同利用センター

国立情報学研究所

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

# 大量の文書⇔少数の読者

**数億点以上**

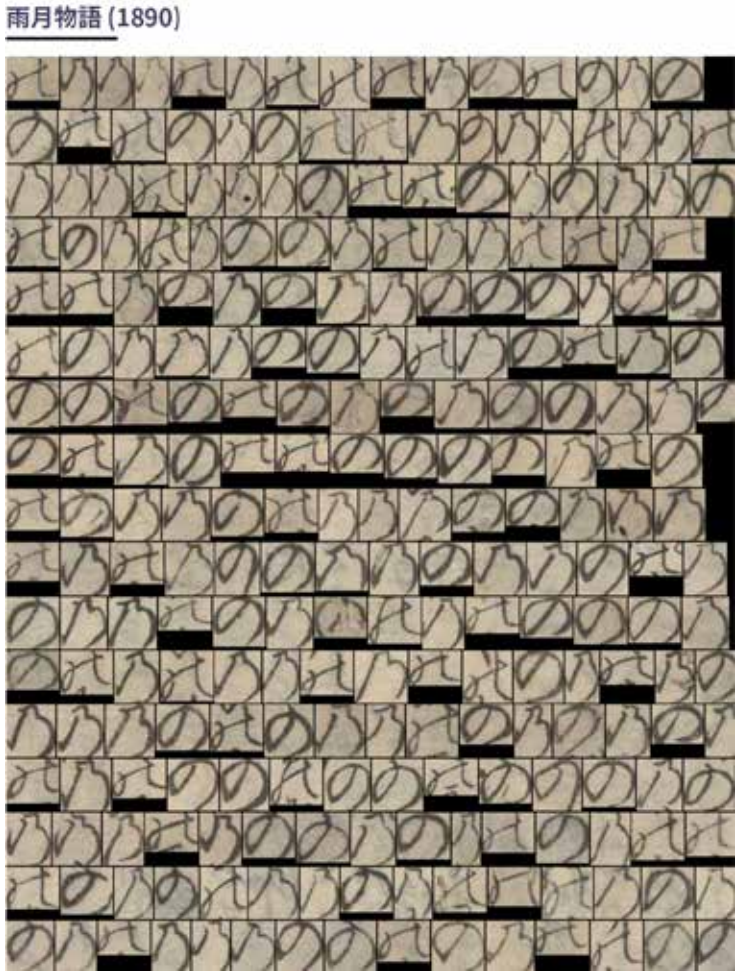
日本に残る古典籍・古文書等の推定点数

**数千～数万人**

くずし字がスラスラ読める推定人数

# くずし字データセット

http://codh.rois.ac.jp/char-shape/

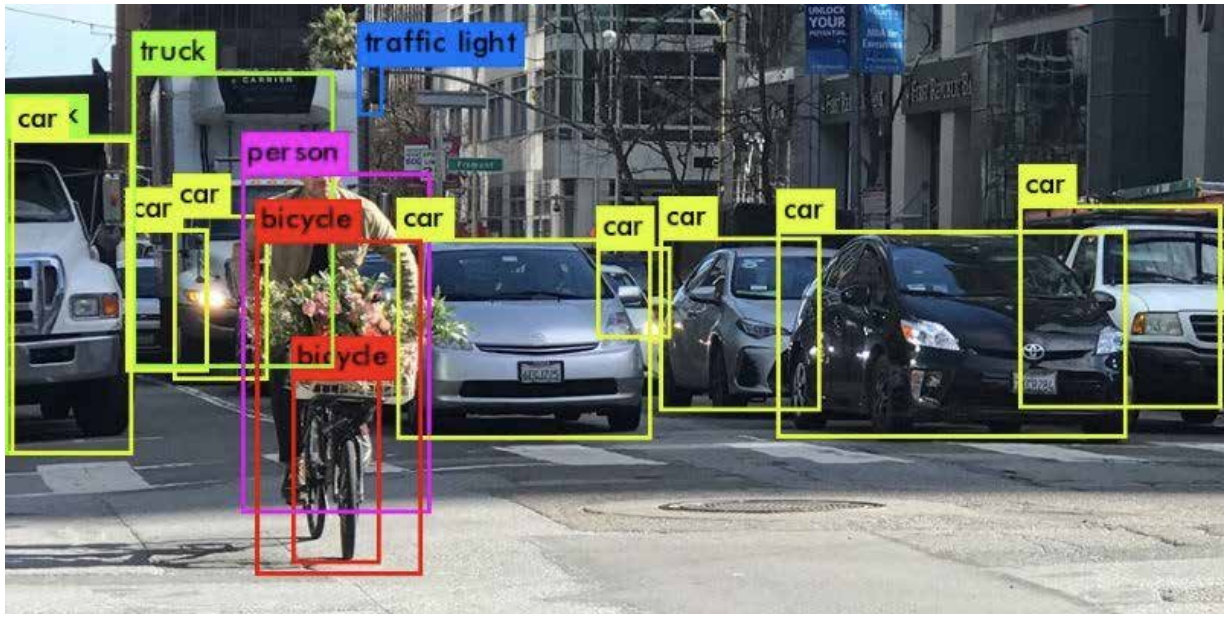



- 国文学研究資料館が作成、CODHが整理して公開するオープンデータ。
- 2020年7月現在
  - 文字種＝4,328
  - 文字数＝1,086,326
- ZIPファイルダウンロード→くずし字認識の訓練データとして利用。

# KuroNetくずし字認識

宇津保物語,『日本古典籍データセット』（国文学研究資料館蔵）

# 物体検出（Object Detection）




1. 画像中に存在する物体を認識する技術は、自動運転などビジネスへの応用範囲も広い。
2. くずし字を「オブジェクト」とみなせば、物体検出技術が使えるのではないか？

# AIくずし字OCRサービス

http://codh.rois.ac.jp/kuzushiji-ocr/

## KuroNet

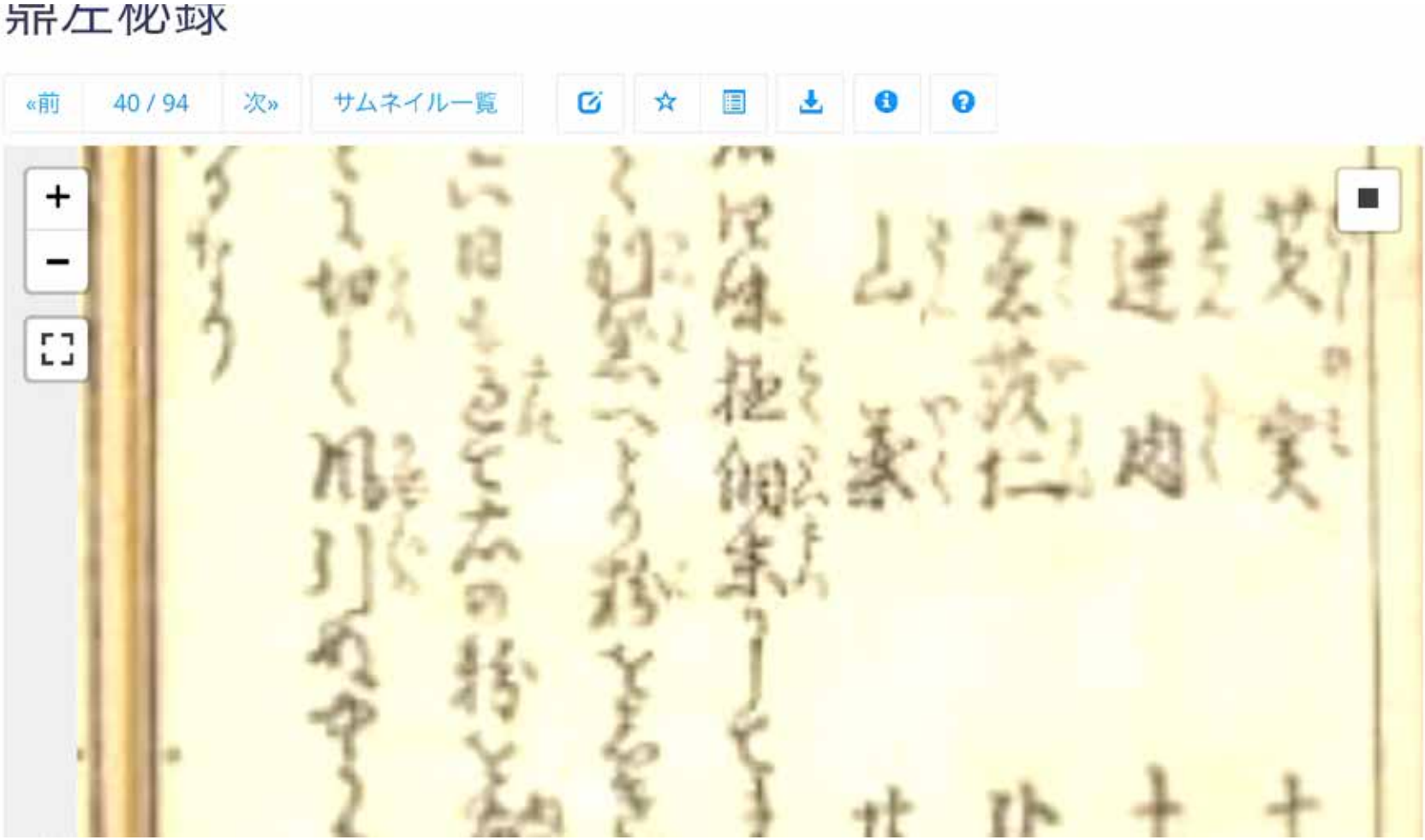
- IIIF準拠の画像を対象に多文字くずし字OCRを提供する、サーバAPIベースのサービス。
- 文字の切り出しも行うため処理が重い。
- **翻刻作業の下読み**などに使う。

## KogumaNet

- 任意の画像を対象に一文字くずし字OCRを提供する、ブラウザベースのサービス。
- 文字の切り出しは人間が行うため、分類のみで処理が軽い。
- **わからない文字を調べる**場合に使う。

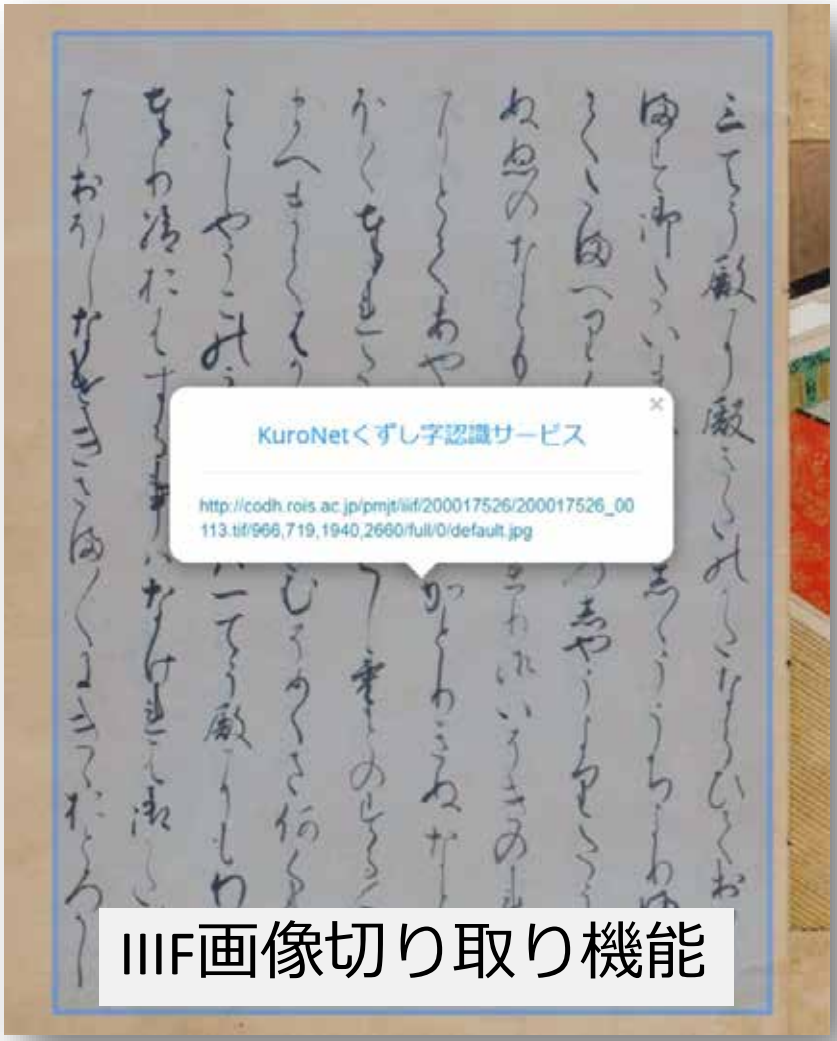
# KogumaNetくずし字認識

http://codh.rois.ac.jp/char-shape/app/single-mobilenet/

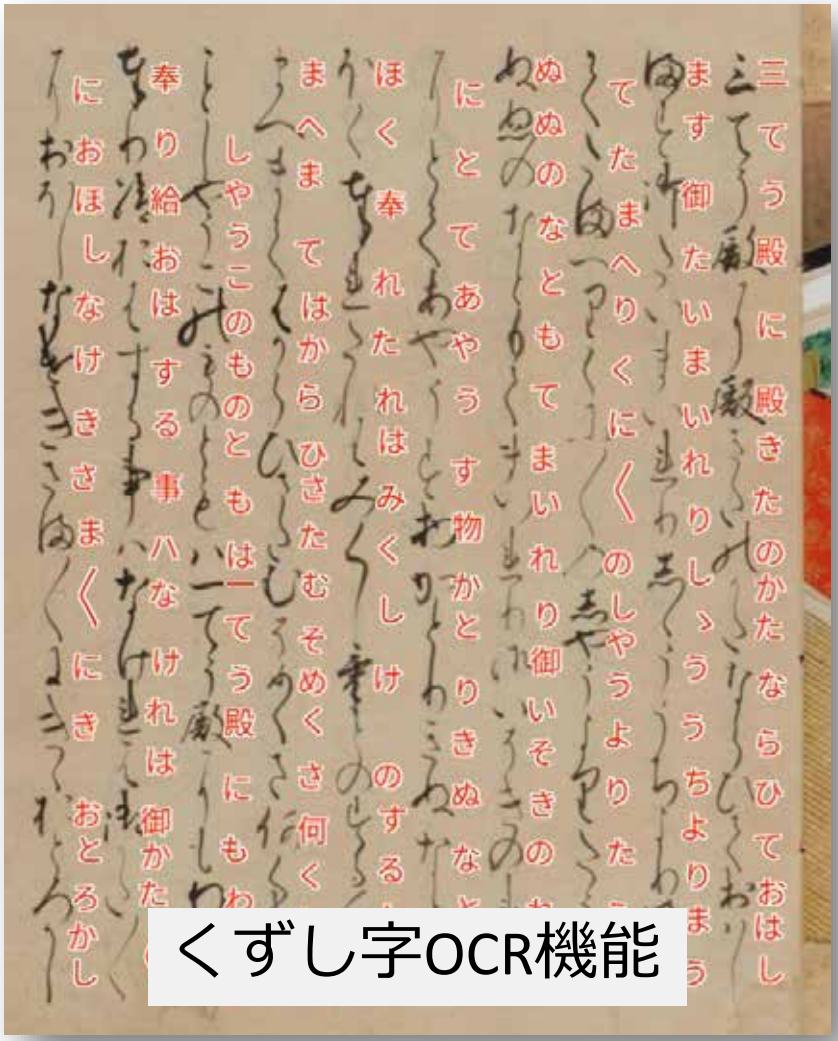


# KuroNetくずし字認識サービス

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



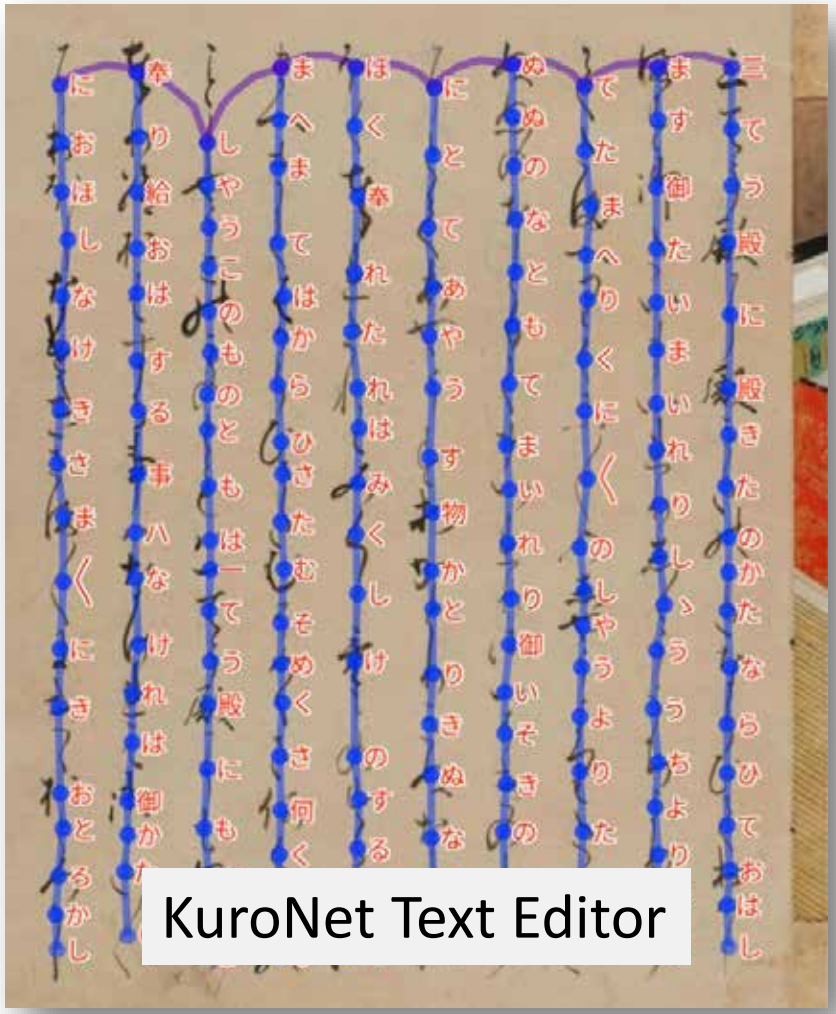

IIIF画像切り取り機能



くずし字OCR機能

# KuroNetくずし字認識サービス

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



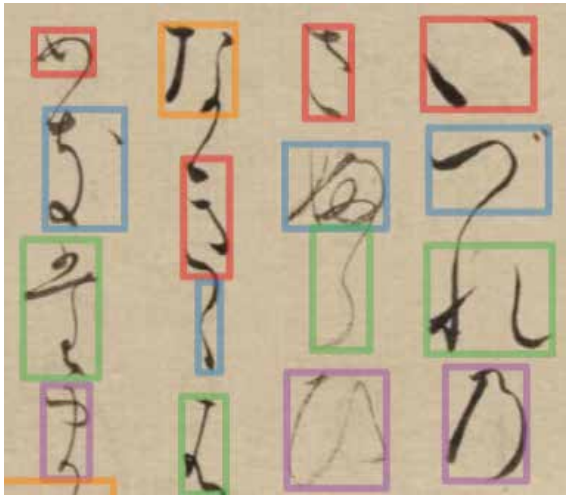

## テキスト化

三てう殿に殿きたのかたならひておはし
ます御たいまいれりしゝううちよりまう
てたまへりくにくのしやうよりたうき
ぬぬのなともてまいれり御いそきのれう
にとてあやうす物かとりきぬなとお
ほく奉れたれはみくしけのする人御
まへまてはからひさたむそめくさ何くれの
しやうこのものともは一てう殿にもわかち
奉り給おはする事ハなけれは御かたく
におほしなけきさまくにきおとろかし

テキスト化サービス

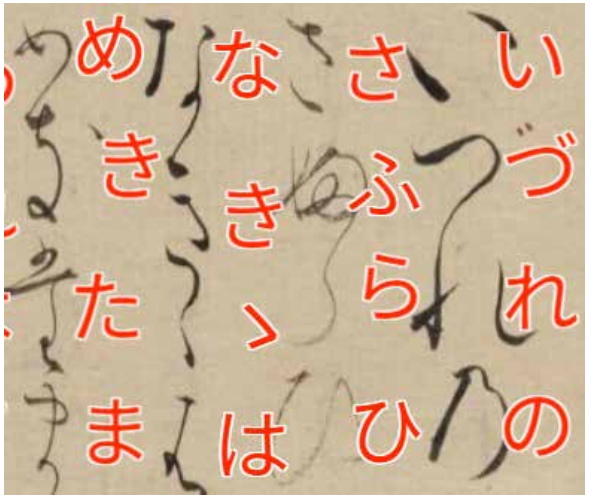
# IIIF Curation Viewer文字表示

http://codh.rois.ac.jp/software/iiif-curation-viewer/

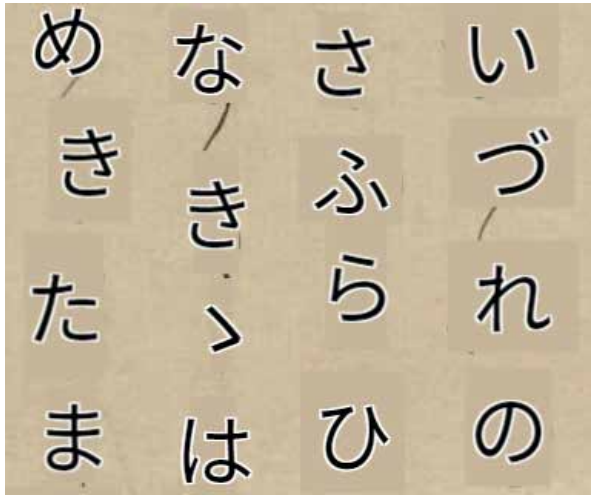


Box frame marker
矩形枠線マーカー

出典：『源氏物語』（国文研蔵）



Character marker (standard mode)
文字マーカー（標準モード）



Character marker (overwrite mode)
文字マーカー（上書きモード）

アノテーションビューモードは文字マーカー（2種類）、矩形枠線マーカー、画像マーカーという4種の表示方法を提供

# kaggle くずし字認識コンペ


上位は精度約95%

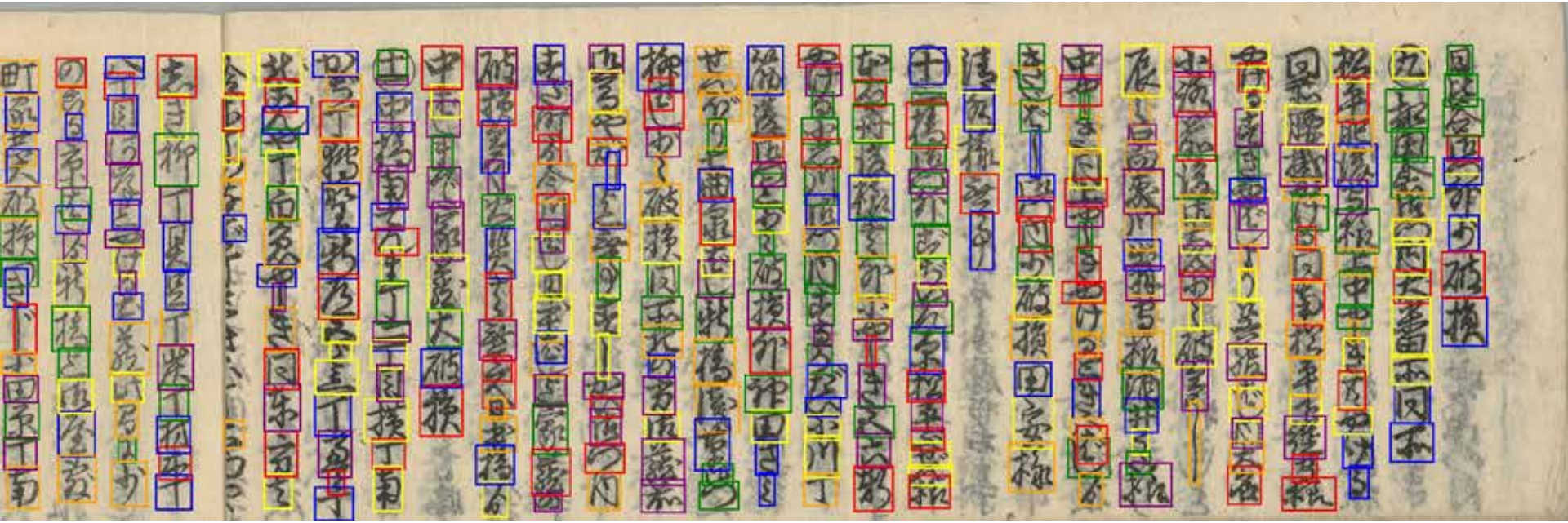
機械学習エンジニアが300万人以上登録する、世界最大のデータサイエンスプラットフォームKaggleにて、くずし字認識コンペを開催。

- 期間：2019年7月19日～10月14日（約3ヶ月）
- 参加チーム数：293
- 参加者数：338
- 結果提出回数：2652

# くずし字認識の現在

「みんなで翻刻」で扱った地震史料に対して、
Kaggleの1位モデルを適用したくずし字認識結果

江戸大地震并二出火場所細見録,石本文庫,東京大学地震研究所図書室蔵
http://wwweic.eri.u-tokyo.ac.jp/dl/meta_pub/G0000002erilib_L000008



日比谷御門外少破損
九和田倉御門内大番所同所
松平肥後守様上中やしき共やける
腰掛やける同南松平下従守様
やけるすきやばしより呉服ばし内大名
小路前後御大名少々破多し
辰之口向森川出羽守様酒井うた様
中やしき同上やしきやけるときはばし
きじばし御門内少破損田安様
清水様無事
十一橋御門外ごぢいん原松平ぶぜん様
本心舟後様其外にやしき五六軒
やける小石川御門内するがだい小川丁
筋違御門迄少々破損外神田さみ
せんぼり七曲泉ばし新橋浅草御門
柳ずし少々破損同所北の方御蔵前
御馬やかし迄無事すしかい御門内
すだ町ゟ今川ばし日本ばし上家蔵の
破損多く火火者垂二大日本橋ゟ
中ばしまで家蔵大破損
十中橋南てんま丁二丁目横丁南
かぢ丁転野新道五郎三丁過々半
北こんや丁白急やしき同東方其
令ちこいなば
しき柳丁具足丁炭丁村木丁
八丁目河岸上やける土蔵此間に少
のこる京ばしゟ新橋上御屋敷
町家共大破損つぎじ小田原丁南

# KuroNetの成功と課題

- KuroNet成功の要因は「処理順の逆転」。それを可能としたのは機械学習の進歩。

1. **通常のOCR：レイアウト解析→文字認識**
2. **KuroNet：文字認識→レイアウト解析**

- 文字認識が先だと、人間可読な出力が先に得られ、レイアウト解析をパスできる。
- コピーペーストや全文検索に必要なレイアウト解析→テキスト化処理は研究中。

# 歴史資料に基づく過去復元


＊1


＊1


＊2


＊1

地震

天気

季節

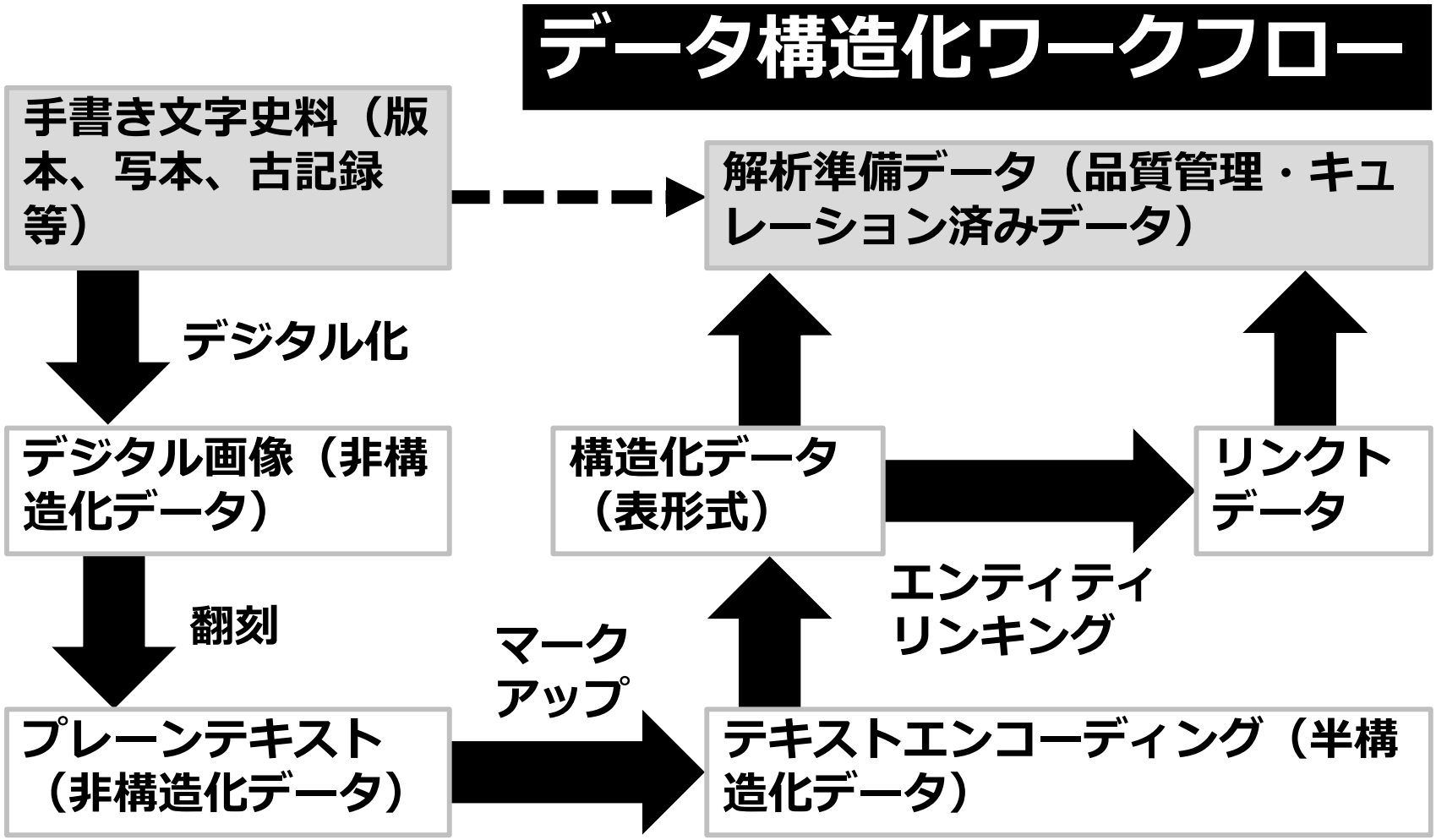
火山噴火

**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

# データ構造化ギャップ


データ構造化ワークフロー
手書き文字史料（版本、写本、古記録等）
解析準備データ（品質管理・キュレーション済みデータ）
デジタル化
デジタル画像（非構造化データ）
構造化データ（表形式）
リンクトデータ
翻刻
マークアップ
エンティティリンキング
プレーンテキスト（非構造化データ）
テキストエンコーディング（半構造化データ）

# 歴史ビッグデータ研究へのインパクト

- 歴史ビッグデータ：「いつどこで何が起こった」に関する索引を多数の史資料を対象に構築し、その内容を統合解析する。

1. 全文テキスト化：KuroNet等により史資料を翻字し、全文検索を活用。
2. スポッティング：史資料から重要な文字・単語を検出し、その周辺を翻刻。

- 研究者が手を出しづらい、データ的文書の翻刻の効率化に期待。

# 今後の課題

1. KuroNetくずし字認識の精度を向上させ、より多くの文字種を認識可能とする。
2. 古文書などにも対応できるよう、**学習データの規模と多様性を拡大**する。そのために**「みんなで翻刻」などのデータを活用**できる仕組みを構築する。
3. 歴史ビッグデータの調査・収集ワークフローにくずし字認識を組み込む。

# おわりに


- 謝辞：本研究は、科研費「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」（研究代表者：北本 朝展）、「ディープラーニングによるEnd-to-End日本古典籍くずし字認識の研究」（研究代表者：カラーヌワット タリン）、国文学研究資料館開発系共同研究などの支援を受けました。
- 謝辞：石本文庫の画像の取得には、橋本雄太氏や加納靖之氏の協力を得ました。


- **KuroNetくずし字認識サービス**
  - http://codh.rois.ac.jp/kuronet/
- **歴史ビッグデータ**
  - http://codh.rois.ac.jp/historical-big-data/