

日本文化をAIとビッグデータ で読み解く

過去の日本文化を「見える化」する
デジタル技術の展開

北本 朝展 (KITAMOTO Asanobu)

国立情報学研究所 コンテンツ科学研究系

ROIS-DS人文学オープンデータ共同利用センター (CODH)

<https://researchmap.jp/kitamoto/>



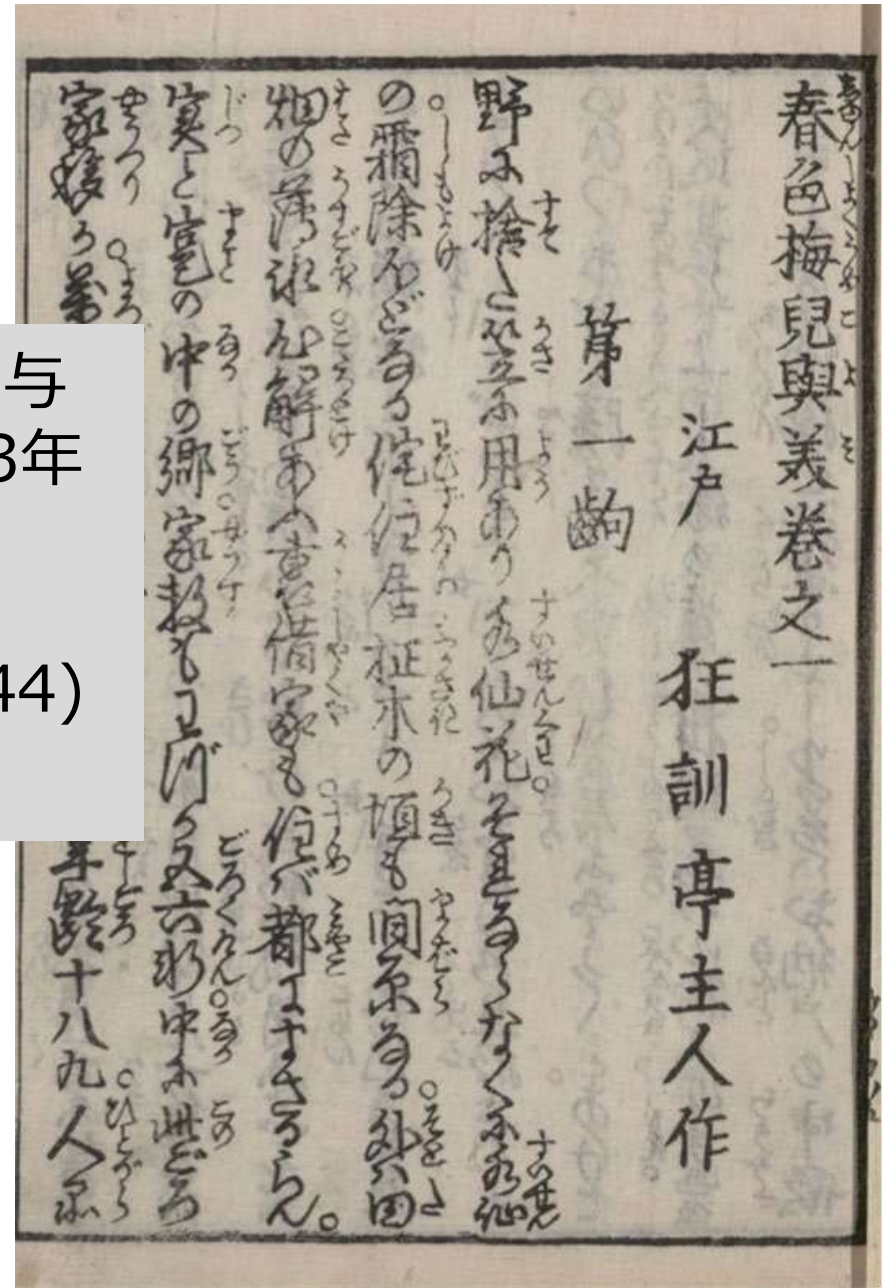
Alice's adventures in
wonderland (1894
edition)
by Carroll, Lewis,
1832-1898
(archive.org)

CHAPTER I.

DOWN THE RABBIT-HOLE.

ALICE was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, "and what is

『春色梅児与
美』 (1833年
版)
為永春水
(1790-1844)
(国語研蔵)



過去の文化遺産と「くずし字」の問題

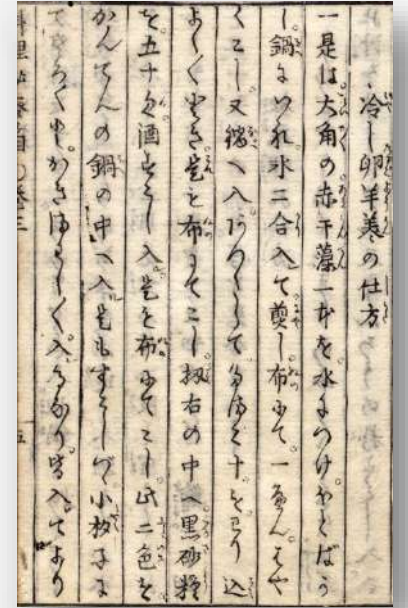


文書
数億点

日本に残された古典
籍や古文書の点数
(冊数) の推定

読者
数千人

くずし字をきちんと
読める人数の推定
(全人口の0.01%)



「見える」 ようになると戦略が変わる



眼の誕生, アンドリュー・パーカー著, 草思社, 2006.

機械の支援により「**新しいものが見える**」ようになると、そうでない人々に比べて優位に立つ

デジタル変革 (DX) : AIなどのデジタル技術を活用することで、これまでの仕事のやり方を変える。色々なものを「見える化」することは、新たなパワーを得ることにつながる。

大黒天



福德寿五色目鏡（ふくとく
じゅごしきめがね）
作：桜川 慈悲成, 1762-1833
画：歌川 豊国, 1769-1825
出版年：不明
所蔵：早稲田大学図書館古典
籍総合データベース



**福德寿五色目鏡（ふくとく
 じゅごしきめがね）**
 作：桜川 慈悲成, 1762-1833
 画：歌川 豊国, 1769-1825
 出版年：不明
 所蔵：早稲田大学図書館古典
 籍総合データベース

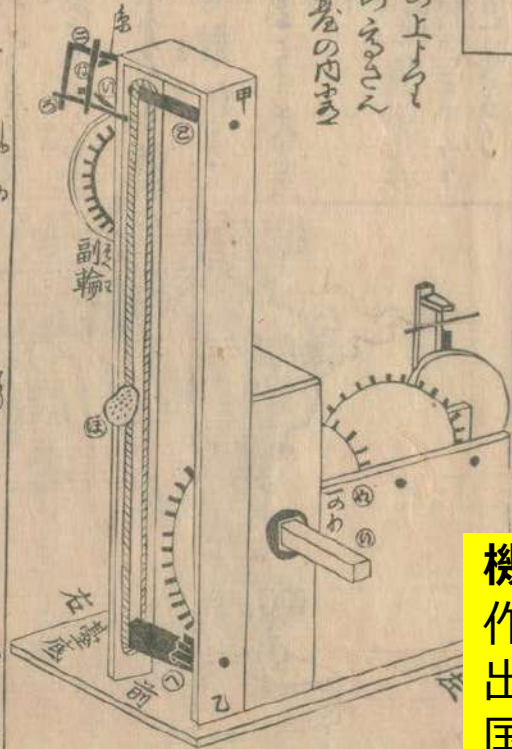


**福德寿五色目鏡（ふくとく
 じゅごしきめがね）**
 作：桜川 慈悲成, 1762-1833
 画：歌川 豊国, 1769-1825
 出版年：不明
 所蔵：早稲田大学図書館古典
 籍総合データベース

内かくくま

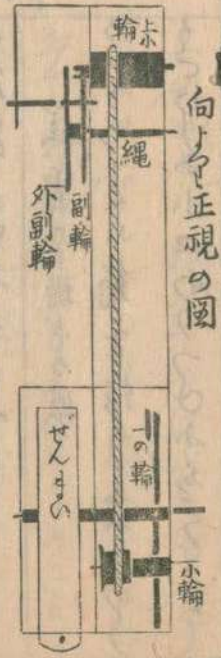
①②のちの山の上より
 壁乃底までのもろさ
 故に下半分の壁の内ま
 上半分の
 山の内まあり

糸の終とすへ
 引付めく



けかくくり一体乃趣意ハ鯉と張ぬきかして内乃腹の
 方ハ鉄と付けおくありさてまこせんまいひて磁石ハ
 びつて磁の内ハのちくくろく之の如くまをく内ハ登
 るゆ外乃磁のふりてと鯉磁石ふつひてのちくくろく

一の輪④よりとぬふ
 ひつて圓の故其脇の小輪
 ①よりとぬふひつてめづる
 ちつて磁石①よりとぬふ
 ひつてのちくくろく



向より正視の図

上下小輪の大小ハ先丈俵ハ定め鯉と登りて又その
 のちよりつめて後ハ龍おろくやうふけべー其登りつめ
 る所ハ草本とまげさせて見ゆやうふきんー登りつ
 めて龍のうしろくありのちくくろくめわ内草本あり
 うくまするのち乃ふひて龍のちくくろくけあ初のちく
 磁つが落ちて面白くべ上小輪の心とこふこつてそれ
 みてそと副輪とちくくろく

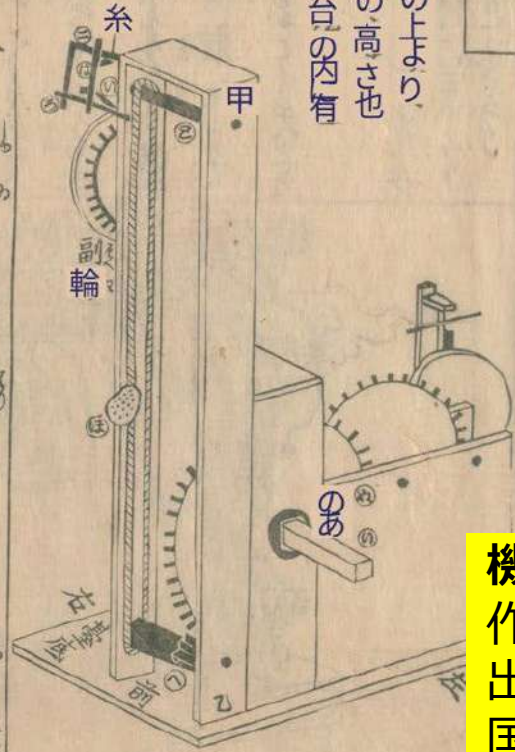
機巧図彙巻

機巧図彙 (きこうずい)
 作：細川頼直
 出版年：1796
 国会図書館蔵
 Doi: 10.11501/2568591

内がらくり

田○の高は山の上より
臺の底までの高さ也
故に下半分は台の内有
上半分は
山の内にあり

糸は龍を下へ
引付おく



此からくり一休の趣意は鯉を張ぬきにして内の腹の
方に鉄を付けおくなりさてまたぜんまいにて磯石を
もつて滝の内をのぼらす也国の如しじしやく内を登
るゆへ外の滝のおもてを鯉磁石につひてのぼるなり

一の輪○よりぬに

むつて周る故其脇の小輪
○よりよにむかつてめぐる
よつて磁石○より国に
むかつてのぼるなり



向より正視の図

上下小輪の大小は先大体に定め鯉を登らせて見るに
のぼりつめて後に竜出づるやうにすべし其登りつめ
る所は草木をしげらせて見へぬやうにすべし登りつ
めて滝のうしろへおつるやうのぼりつめぬ内草木に
かくれたるのミの所にて竜いづれはひゞけに初りのやう
滝つぼへ落ちて面白からず上小輪の心を二つにわつてそれ
にてそと副輪をめぐらすなり

機巧図彙 (きこうずい)
作：細川頼直
出版年：1796
国会図書館蔵
Doi: 10.11501/2568591

龍門滝
 世に傳て曰唐土の龍門の滝へ鯉登り得て後龍に化して天上すとけ説りよつてこれをにつくるものなり
 ○圖のどく瀧の下小鯉を置自然と瀧とのどくより瀧の本へのどくこの鯉化して龍となり雲をおこし天上を次乃圖のどく
 この鯉はなれたるものにて手にとりて見るに何のからくりも見へずはなれわざなり

龍門滝
 世に傳て曰唐土の龍門の滝へ鯉登り得て後龍に化して天上すと此説によつてこれをつくるものなり
 ○圖のどく瀧の下に鯉を置自然と瀧をのぼるなり瀧の本へのぼりこの鯉化して龍となり雲をおこし天上す次の圖のどし
 この鯉はなれたるものにて手にとりて見るに何のからくりも見へずはなれわざなり



ギャラドス



コイキング

ポケモンずかん
<https://zukan.pokemon.co.jp/>

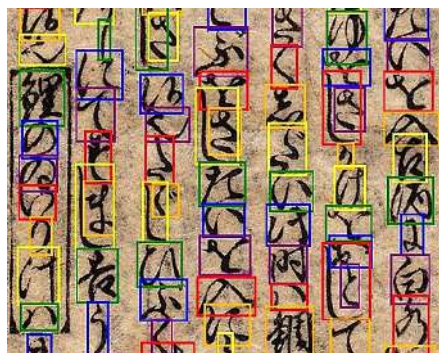
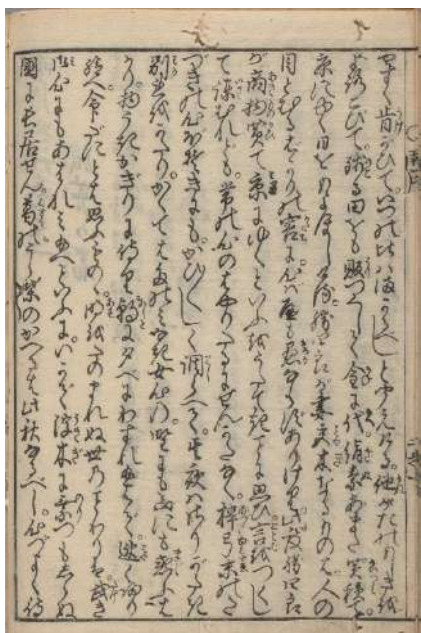
くずし字に関する技術

AIくずし字認識

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

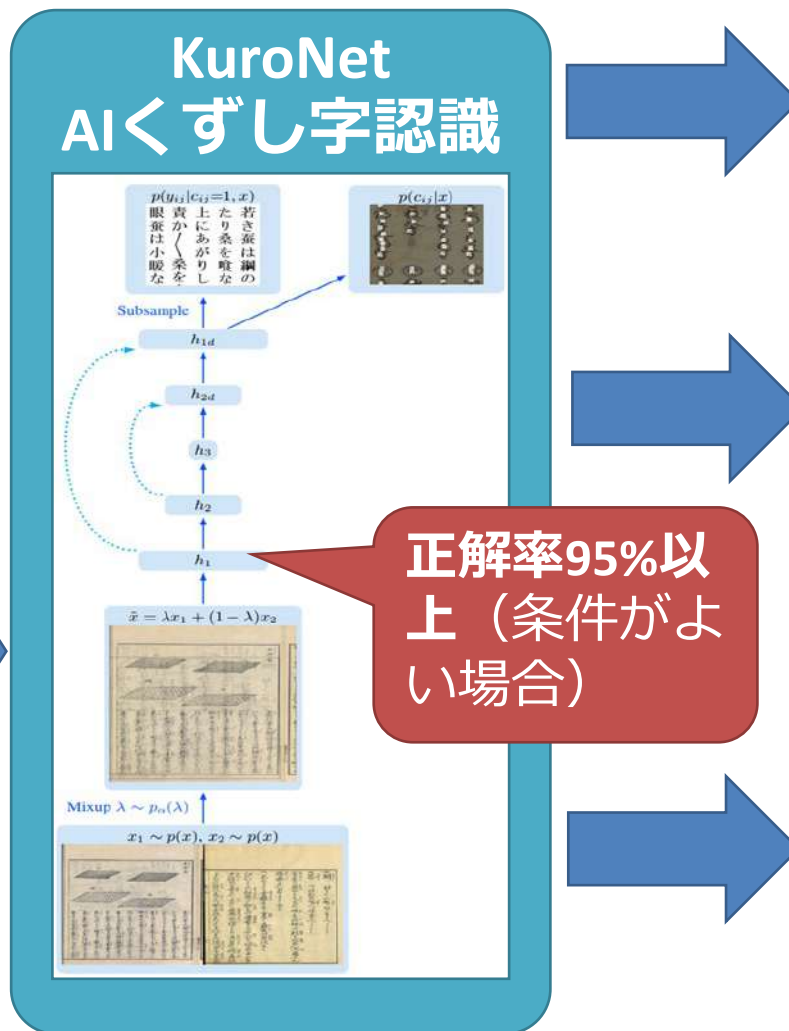
日本古典籍
データセット
(国文研蔵)

くずし字データ
セット (国文研・
CODH作成)



file	char	x	y
200003803_00024_2.jpg	U+3067	416	114
200003804_00024_2.jpg	U+3055	232	115
200003805_00024_2.jpg	U+304A	327	115
200003806_00024_2.jpg	U+3068	145	116
200003807_00024_2.jpg	U+3046	369	116
200003808_00024_2.jpg	U+305F	457	116
200003809_00024_2.jpg	U+5FA1	104	117
200003810_00024_2.jpg	U+3072	191	118
200003811_00024_2.jpg	U+540D	279	120
200003812_00024_2.jpg	U+3061	501	120

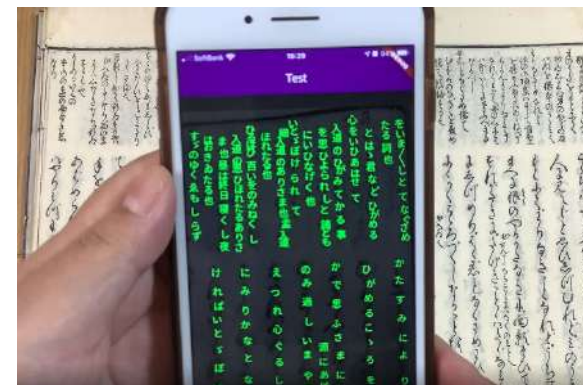
Google Researchカラーヌワット・タリンほか



くずし字認識サービス

kaggle

くずし字認識コンペ



AIくずし字認識アプリ
「みを」

日本古典籍くずし字データセット

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

国文学研究資料館と共同して作成・公開
文字種： **4,328** / 文字数： **1,086,326**



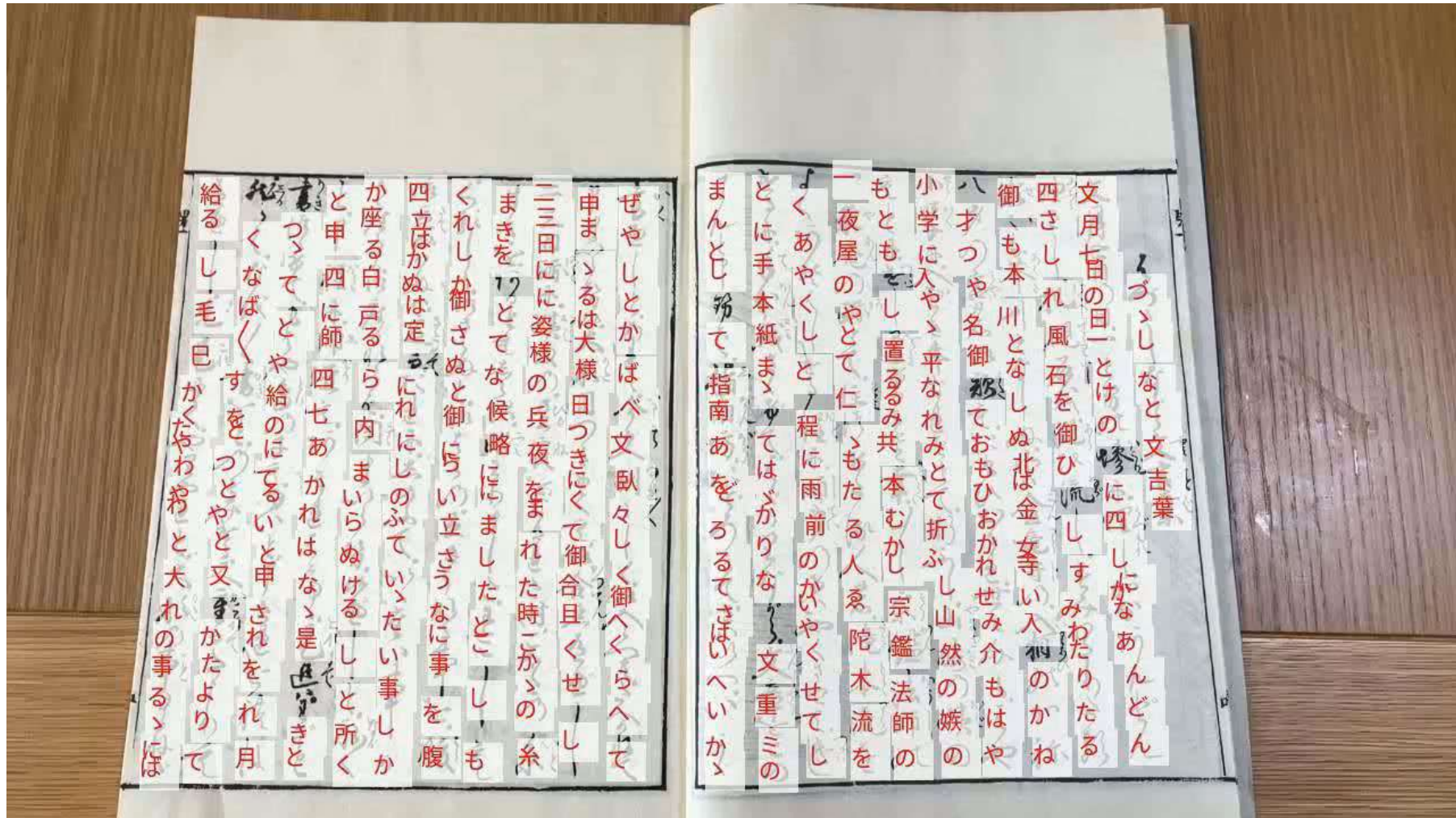
物体検出 (Object Detection)



物体検出技術：顔検出や自動運転など幅広い応用がある

くずし字 = オブジェクトとみなせば使える？

物体検出に基づくくくずし字認識



kaggle くずし字認識コンペ

<http://codh.rois.ac.jp/competition/kaggle/>



機械学習エンジニアが1500万人以上登録する、**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
- **参加チーム数**：293
- **参加者数**：338
- **結果提出回数**：2652

kaggle コンペの結果

上位の精度
は約95%

1. **さまざまな国の機械学習エンジニア、研究者が上位入賞**
2. **日本語やくずし字の知識がなくても開発可能**
3. **コンペの準備をきちんと進めるには、問題に詳しい人と一緒に取り組む必要がある**

#	Δpub	Team Name	Notebook	Team Members	Score	Entries
1	—	tascj			0.950	13
2	—	Konstantin Lopuhin			0.950	60
3	—	Kenji			0.944	161
4	▲1	YoudaoOCR			0.942	49
5	▼1	See--			0.940	42
6	—	abc			0.939	15
7	—	K_mat			0.934	20
8	—	t-hanya			0.920	21
9	—	Ollie, Nanashi, and Tom			0.910	35
10	—	Zenkei_R&D			0.903	144
11	—	masayai			0.903	12
12	▲5	Kirill Brodt (shad nsk)			0.901	4
13	▲1	James Day			0.901	33
14	▼1	NEU			0.900	54
15	▼3	s tatsuya			0.900	29

AIくずし字認識アプリ「みを」

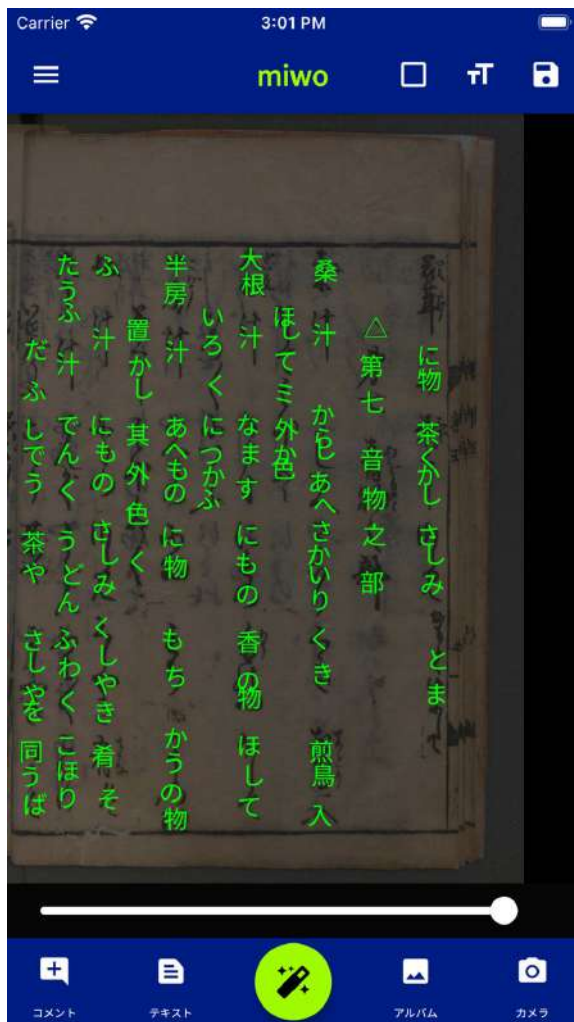
<http://codh.rois.ac.jp/miwo/>



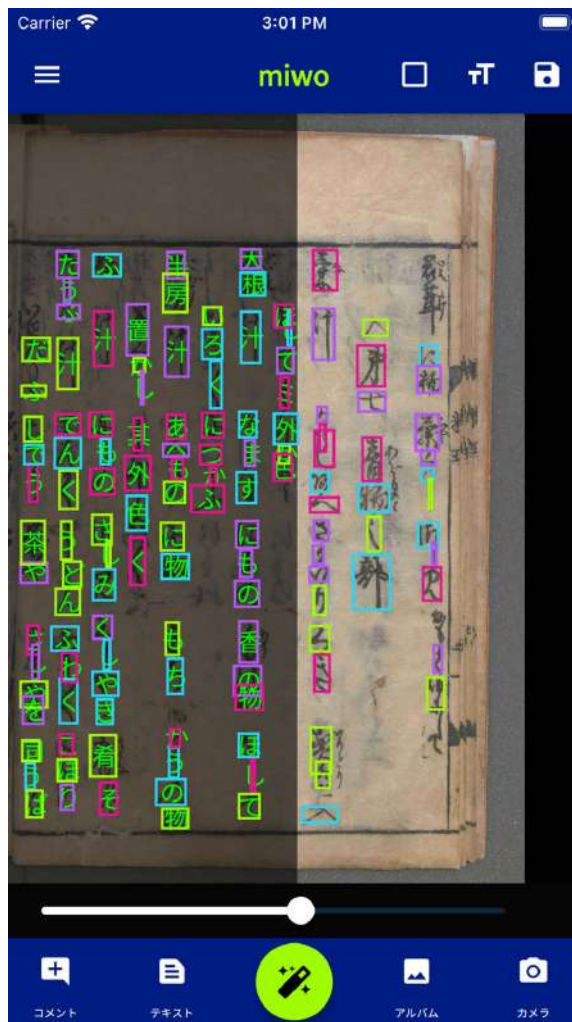
- 『源氏物語』第14巻「みをつくし」（漣標）＝「みを（船の水路）を示すために立ててある杭」の意、「身を尽くし」の掛詞でもある
- 「みをつくし」が人々の水先案内となるように、「みを」アプリが、くずし字資料を読むための道案内となることを目指す
- **2021年8月30日にiOS版とAndroid版を無料公開**
- アプリダウンロード数は約**15**万件、くずし字認識画像数は**170**万件以上
- 公開後1年間、1日2000-4000件程度の利用で安定



「みを」アプリのデザイン



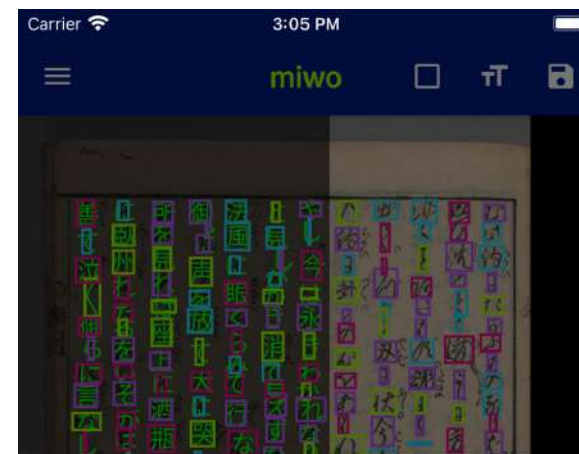
くずし字認識



Bounding Box 表示



認識結果修正機能



む此約にたがふものならば賢第吾を何ものとかせんとや
 たすら思ひ洗のども通りに方なしめしふの人のいふ人
 一百に千りをゆくをあたはず魂よく一日に千りをもゆく
 と此とわりを思ひやてみづから刃に伏今夜陰風に乗ては
 る〈来り菊花の約に世この心をあよれみ如へといひをは
 りて涙わき出るがやし今は永きわかれなり只母公によ
 つかへ治へとて座を立と見しがかき消て見えずなりにけ
 る左門院廿とゞめんとすれば法風に眼くらみて行なをし
 らず俯向につまづき倒れたる御に声を放て大に哭く老
 母同さめ驚き立て左門がある所を見れば座上に酒瓶魚盛
 たる皿ともあまた列へたるが中に臥州れたるをいそがは
 しく扶起していかにとどへとも只声を善て泣く御らに言
 なし老母問てふ伯氏亦穴が約にたがふ

テキスト出力

くずし字画像生成アプリ「そあん」

<http://codh.rois.ac.jp/soan/>

吾輩は猫である。名前はまだ無い。

どこで生れたかとうんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー泣いていた事だけは記憶している。吾輩はここで始めて人間というものを見た。しかもあとで聞くとそれは書生という人間中で一番獰悪な種族であったそうだ。この書生というのは時々我々を捕えて煮て食うという話である。しかしその当時は何という考もなかったから別段恐しいとも思わなかった。

くずし字画像を生成！

サンプル:

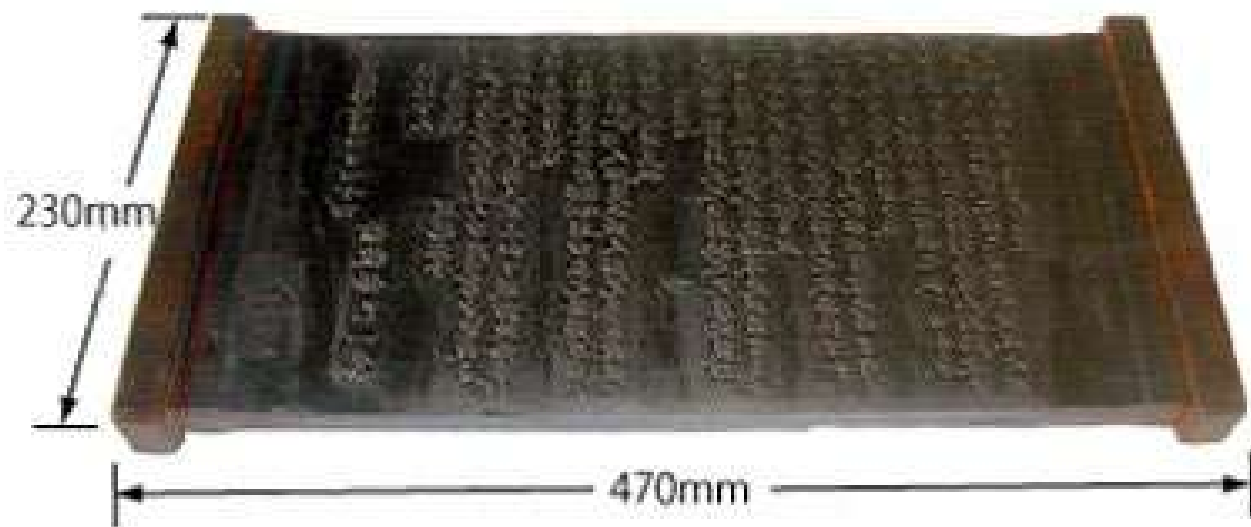
吾輩は猫である

日本国憲法第九条

日本の印刷技術

活版印刷

木版印刷



塙保己一史料館,
http://onkogakkai.com/aboutgunshoruijyu/gunshoruijyu_hangi/



印刷博物館, <https://www.printing-museum.org/collection/looking/15170.php>

歴史ビッグデータ

歴史ビッグデータの統合解析

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

過去のビッグデータを統合解析するための基盤技術の研究

地災撮要 巻11-12(地震之部)

51 / 75

自然科学的データ

人文社会的データ

歴史的資料 (史料)

このスクリーンショットは、地災撮要のウェブページと、その中の歴史的資料の画像を示しています。ウェブページには「自然科学的データ」と「人文社会的データ」の分類があり、右側には「気候」「地震」「噴火」「疫病」「経済」「人口」「政治」「文化」の項目がリストアップされています。また、「歴史的資料 (史料)」という赤い文字で強調された部分があります。

自然科学的データ

人文社会的データ

気候

地震

噴火

疫病

経済

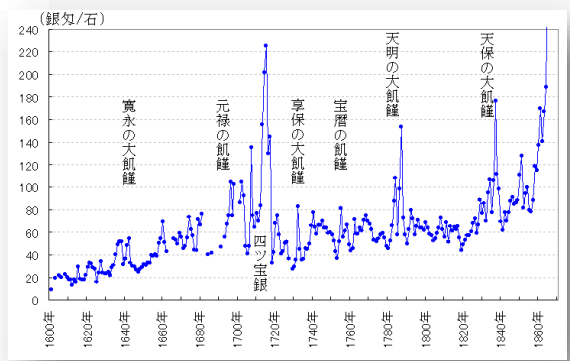
人口

政治

文化

データ
構造化
ワーク
フロー

歴史ビッグデータ研究基盤 (機械可読データ)



江戸マップβ版

<http://codh.rois.ac.jp/edo-maps/>

番号	分類	現代語訳	翻刻	地図
2-001	施設	幸橋御門	幸橋御門	拡大図
2-002	施設	山下御門	山下御門	拡大図
2-003	施設	数寄屋橋御門	数寄屋橋御門	拡大図
2-004	施設	鍛冶橋御門	鍛冶橋御門	拡大図
2-005	施設	呉服橋御門	呉服橋御門	拡大図
2-006	地名	一石橋	一石橋	拡大図
2-007	地名	出橋	出橋	拡大図
2-008	町名	丸屋町	丸屋丁	拡大図

[2-296]
地名：磯辺大神宮（イソベ大神宮）
分類：寺社仏閣

29枚の江戸切絵
図から8722ヶ
所の地名を抽出
してデータベー
ス化



千代田区

© 2020 ZENRIN

Google Earth

国立情報学研究所 市民講座 2023

1997

35° 40'52.26" N 139° 45'26.42" E 標高 2 m 高度 2.00 km

江戸観光案内 – 観光ビッグデータ

<http://codh.rois.ac.jp/edo-spots/>

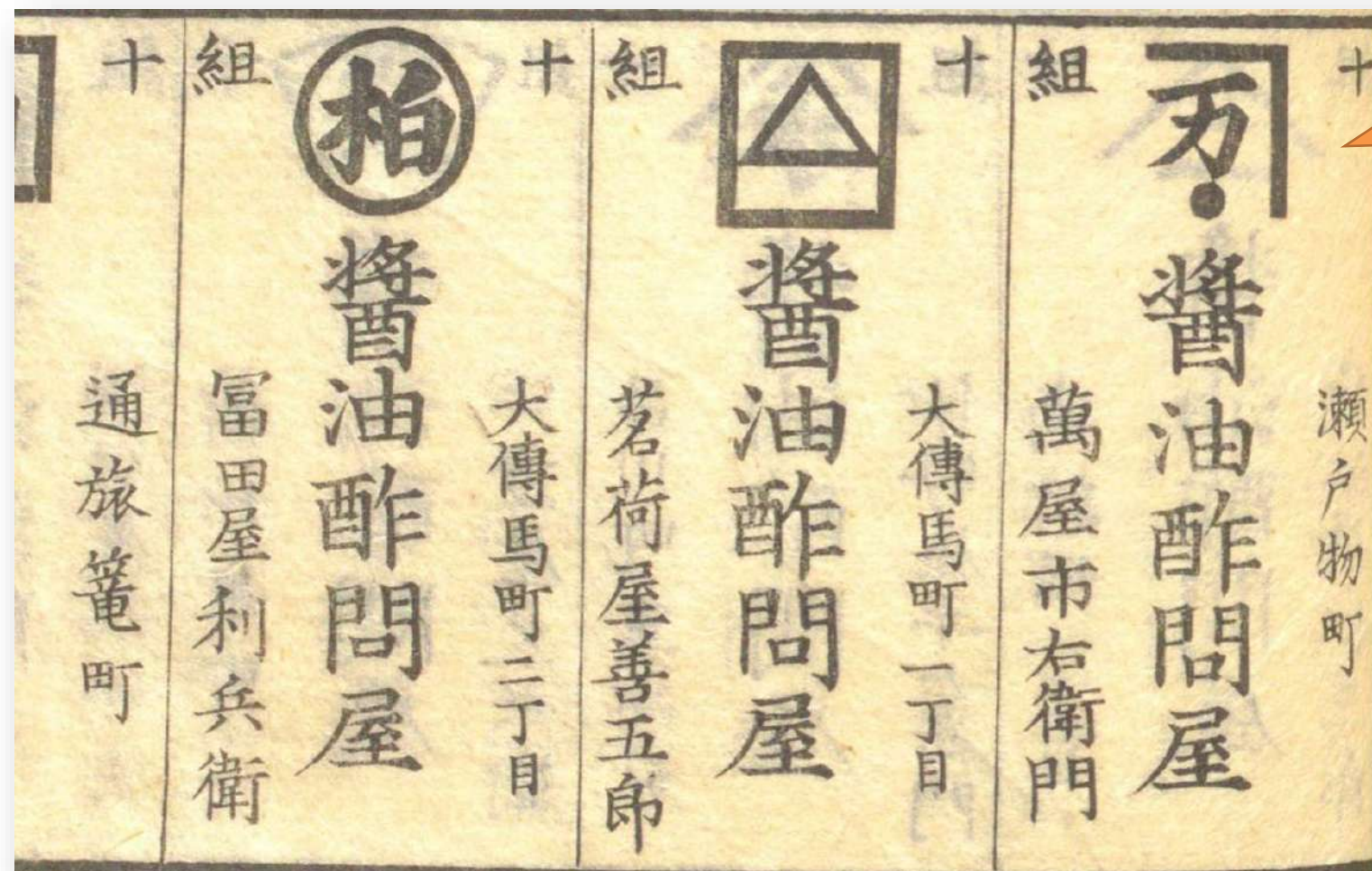


江戸の人々はどんなところを観光していたの？

江戸時代の名所記
や名所案内を各世紀2点ずつ選択し、
挿絵部分の地名を
データベース化

江戸買物案内 – 商業ビッグデータ

<http://codh.rois.ac.jp/edo-shops/>



江戸の人々はど
んなところで
ショッピングし
ていたの？

『江戸買物独案内』（1824）から、商人名や居所などを抽出しデータベース化

過去の人々の移動履歴

<http://codh.rois.ac.jp/edomi/route/1/>

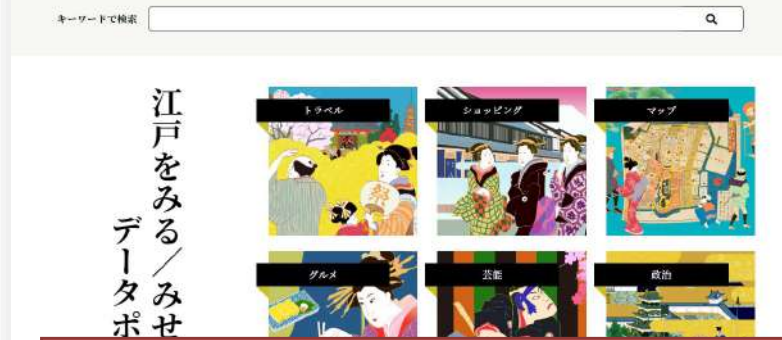


江戸の人々はどのように移動していたの？

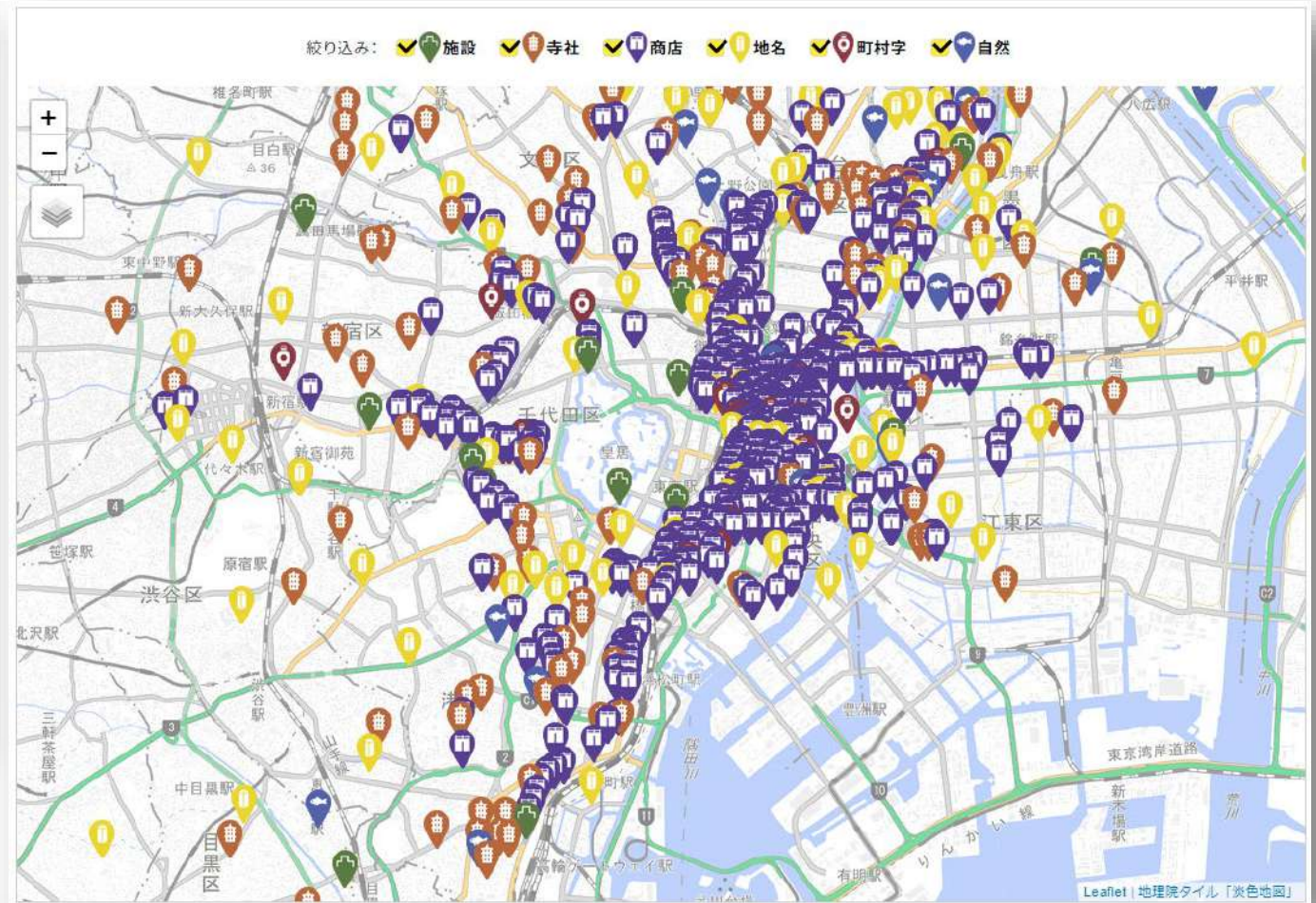
清河八郎が安政2年（1855）に江戸を訪問した日記『西遊草』を分析し、日時と地名をデータベース化

edomi - 江戸をみる／みせるデータポータル

<http://codh.rois.ac.jp/edomi/>



もし江戸時代にYahoo!
があったら？



これからのAI活用（大規模言語モデル）

「つくし」プロジェクト

<http://codh.rois.ac.jp/tsukushi/>



AIくずし字認識

大規模言語モデル (LLM) = ChatGPTやBard など

自動テキスト化／翻訳実験の結果はAIで自動生成したものであり、人間による確認や修正は行っていませんのでご注意ください。

★AIくずし字認識 (RURI)

両国橋納涼図武蔵国と下総にかゝるおらんゆへに此名ありとかや橋の長さ九十六間ありて川上はすみだ川なり川下は海にちしか故にすな取のたのしみあり又三ふくの夏の日舟をうかべて暑をさくるものおひしたく舟と陸のへたてなくにきはひをなす目ざましき風情なり

★現代文翻訳 (GPT-4)

両国橋の納涼図は、武蔵国と下総に架かるため、この名前があるとされています。橋の長さは96間あり、川上は墨田川です。川下は海に至るため、砂取りの楽しみがあります。また、三伏の夏の日には船を浮かべて暑さを凌ぐものが多く、船と陸の間にはほとんど区別がなく、賑わいを成す光景は非常に鮮やかです。

★英語翻訳 (GPT-3.5)

Ryogoku Bridge's cool summer scenery is named because it spans Musashi and Shimousa. The bridge is 96 ken long, and the river upstream is Sumida River. The river downstream leads to the sea, so there is the pleasure of taking sand. Also, many people float boats to escape the heat on the hottest days of summer, and there is little distinction between boats and land. and the sight

ChatGPTはすべてを知るわけではない



明日、
雨降る？



ChatGPT
GPT-4

GPT-4の場合
「私のトレーニングデータは2021年9月までの情報で更新されています。それ以降の具体的な出来事や情報については知ることはできません。」

外部情報なしに答える場合



『安政風聞集』について教えて

安政風聞集（あんせいふうぶんしゅう）
編集：金屯道人
出版年：1856
所蔵：早稲田大学図書館
古典籍総合データベース



Q: 安政風聞集は誰が書いた？

A: 民俗学者である柳田國男 ✕

Q: いつ地震があった？

A: 「安政江戸地震」のことなら
1855年11月11日（Wikipediaと同じ）

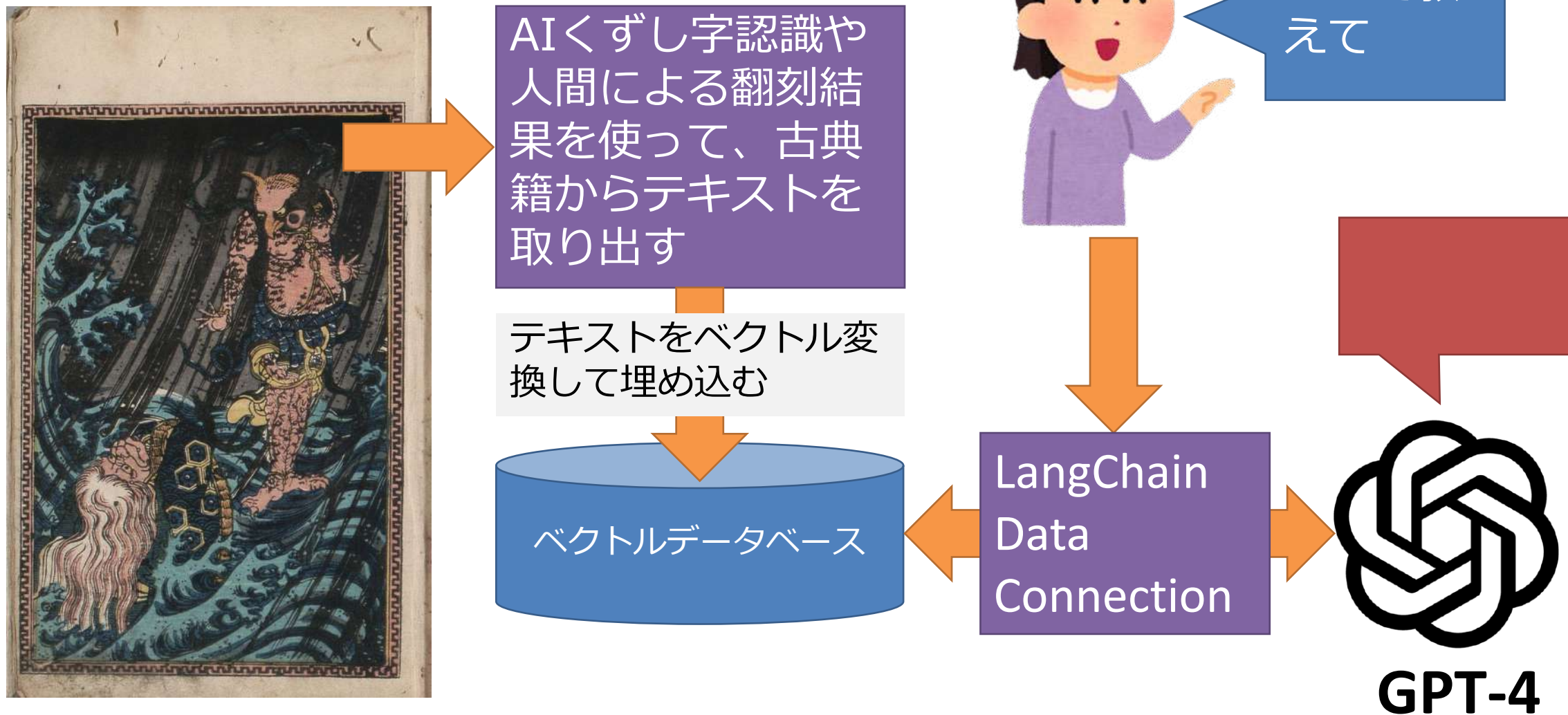
Q: 被害が少ないところは？

A: 具体的な地域についての被害の程度を指摘するのは困難です。

Q: いつ台風があった？

A: 情報を具体的に提供することは困難です。

外部情報を与える方法



外部情報も使って答える場合



『安政風聞集』について教えて

安政風聞集（あんせいふうぶんしゅう）
編集：金屯道人
出版年：1856
所蔵：早稲田大学図書館
古典籍総合データベース



Q: 安政風聞集は誰が書いた？

A: 金屯道人が編輯した ✓

Q: いつ地震があった？

A: 去年の10月2日に大地震がありました。（序文にある情報） ✓

Q: 被害が少ないところは？

A: 小川町、福井町、鳥越新堀、駒形町、駒込追分、王子権現、上板橋、吾妻橋、神田橋平岡家、護持院原杉 (Fact Checkが必要)

Q: いつ台風があった？

A: 8月25日の夜に大風雨がありました ✓

人文情報学（デジタル・ヒューマニティーズ）への誘い

1. **人文学**：歴史学、文学、言語学、美術など、人間の文化（知的活動）に関する研究を展開する
2. **精読**：これまでは、**人間が書物や作品を「読む」**という方法で、研究がおこなわれてきた
3. **遠読**：これからは、**機械も書物や作品を「読む」**ようになり、大規模な新しい知識が生まれてくる
4. **日本文化を「見える化」**することで、過去の文化と現代人のギャップを縮め、親しみやすくする！

より詳しく



ROIS-DS人文学オープンデータ共同利用
センター (CODH)

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