第20回図書館総合展　国立国会図書館主催フォーラム
「AIやクラウド技術は図書館をどう変えていくか
～国立国会図書館の次世代システム開発研究室の実験事業、関連研究から」

次世代システム開発研究室の取組

電子情報部電子情報企画課
次世代システム開発研究室
里見 航・青池 亨
2018/10/30

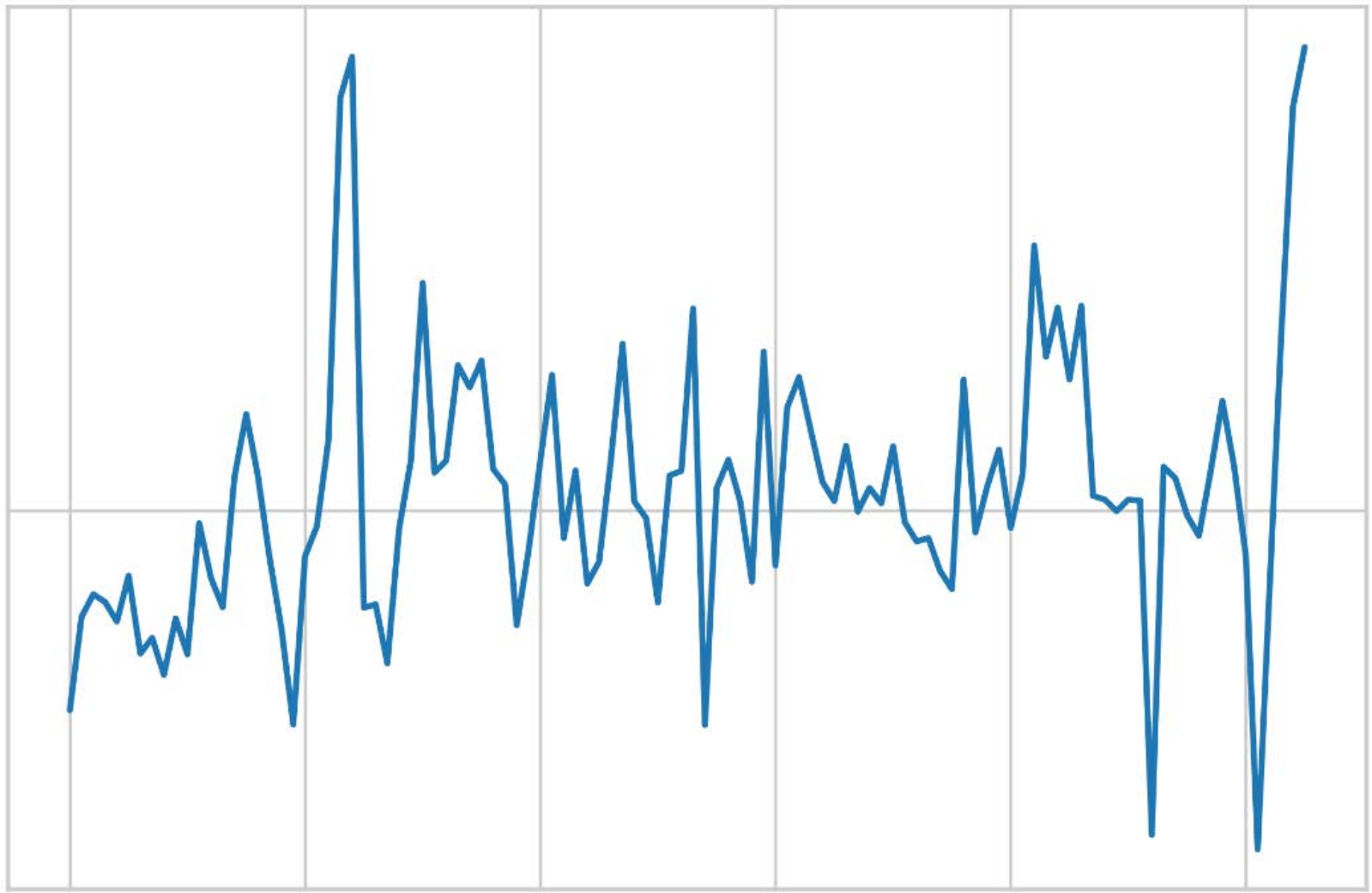
（１）図書館業務効率化への挑戦

取組1：来館者数の予測

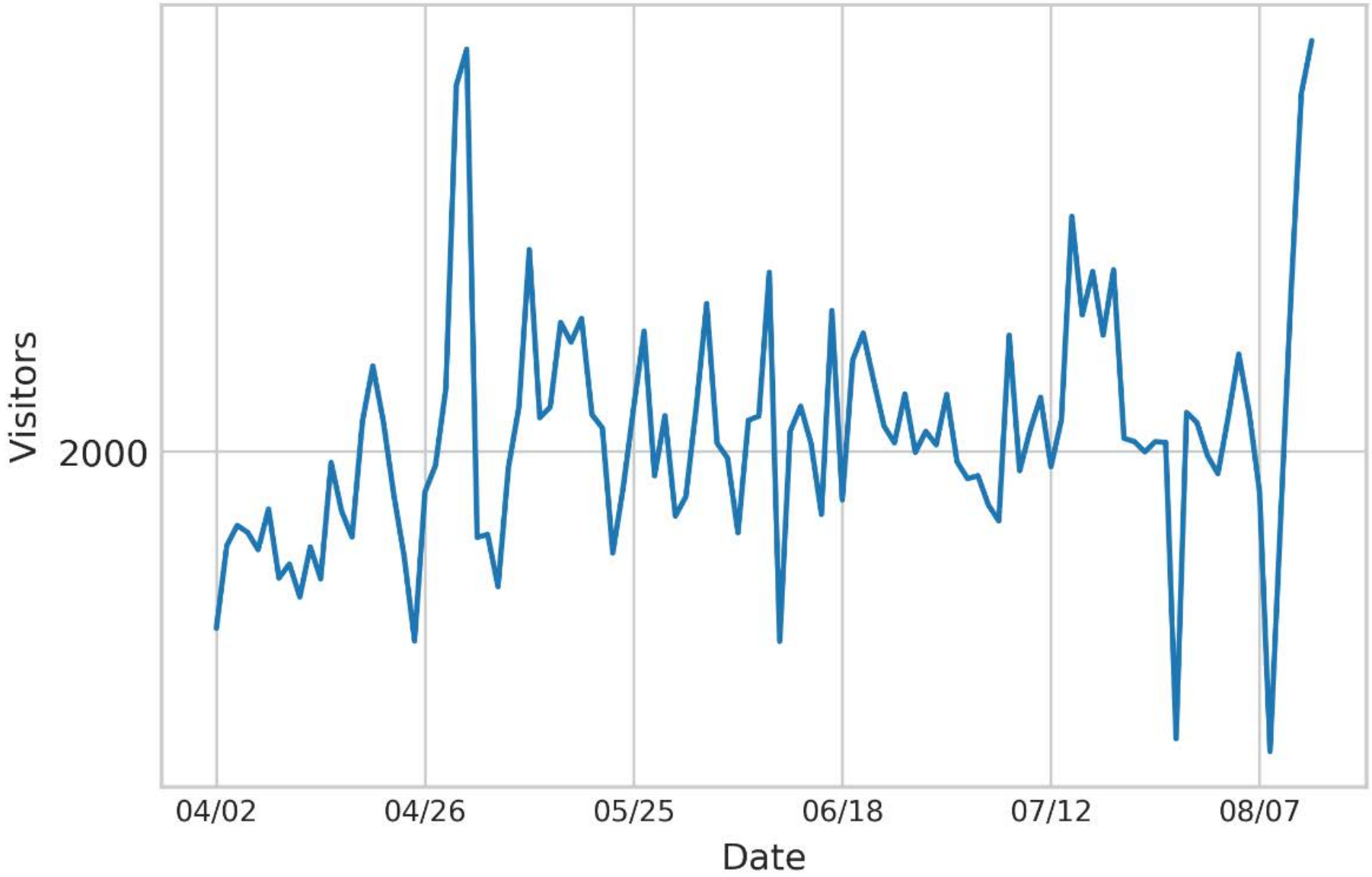
取組2：書架分類の推定

取組3：複写品質の向上

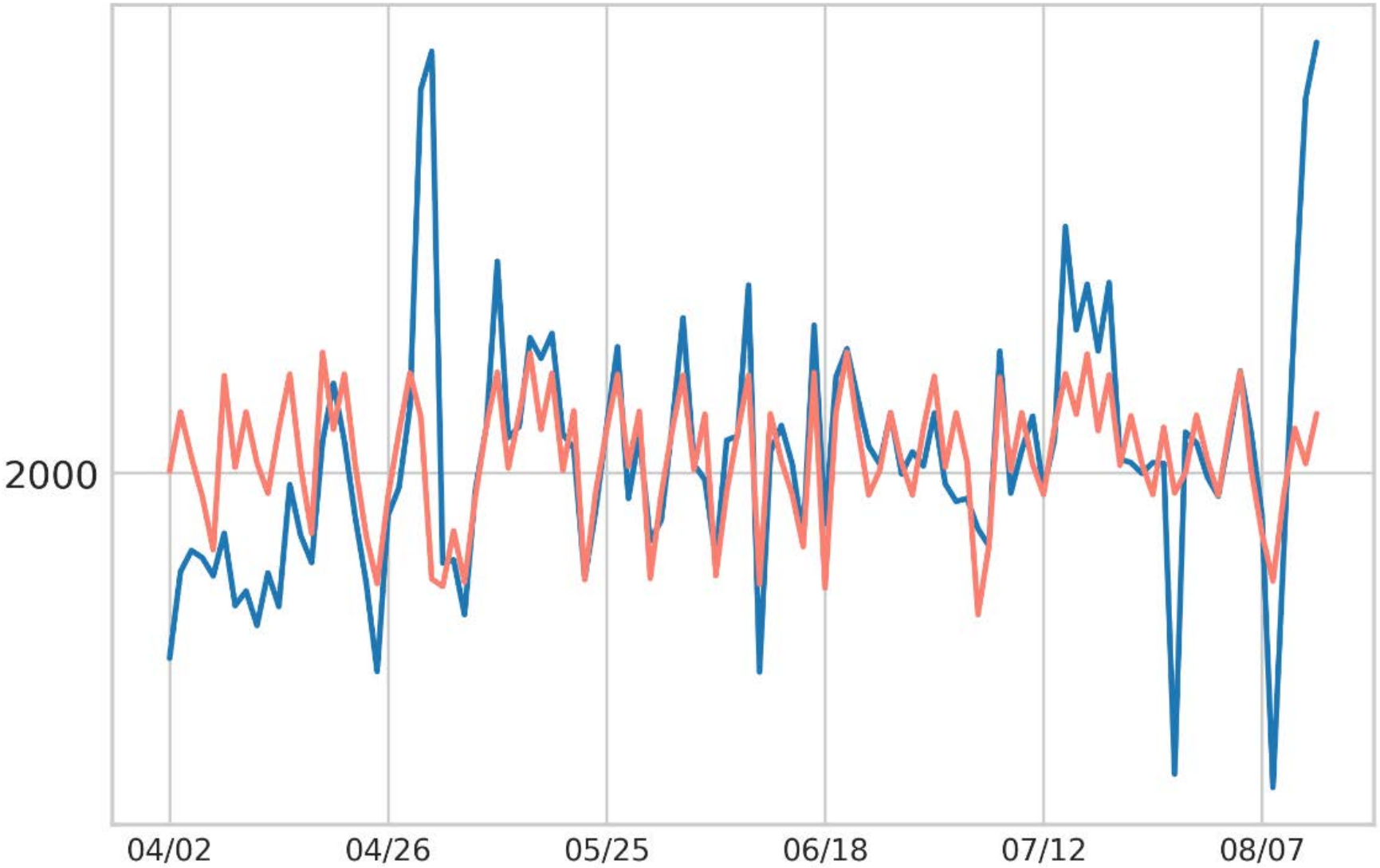
突然ですが、



当館の4月～8月の来館者数の推移です



取組１：来館者数の予測

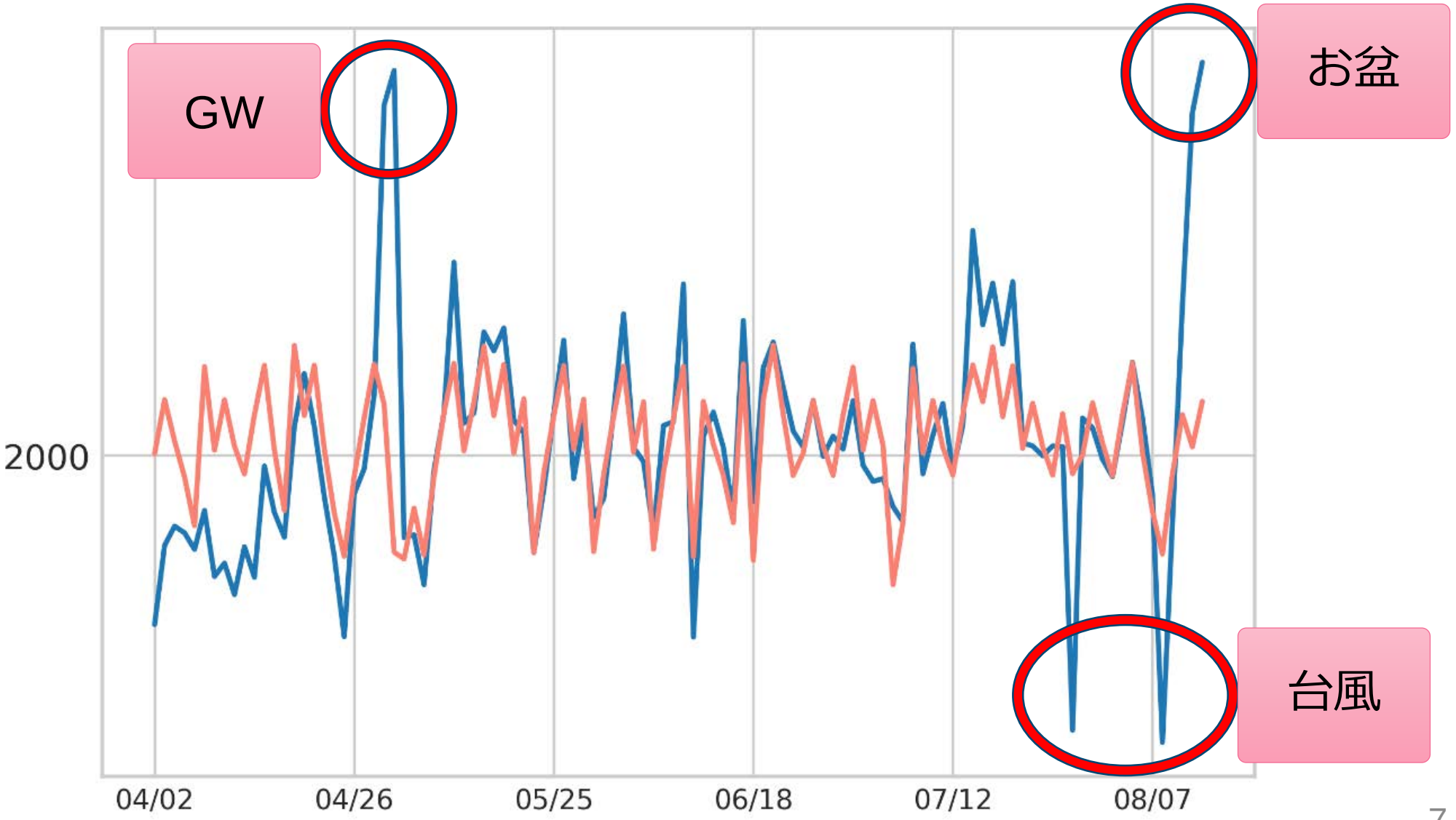


取組1：来館者数の予測

- 予測に使ったモデル

来館者数 ～ $normal($ 曜日ごとの平均値
（− 雨が降った日の影響）
（+ 資料整理休館日明けの調整）
, 分散)

取組1：来館者数の予測

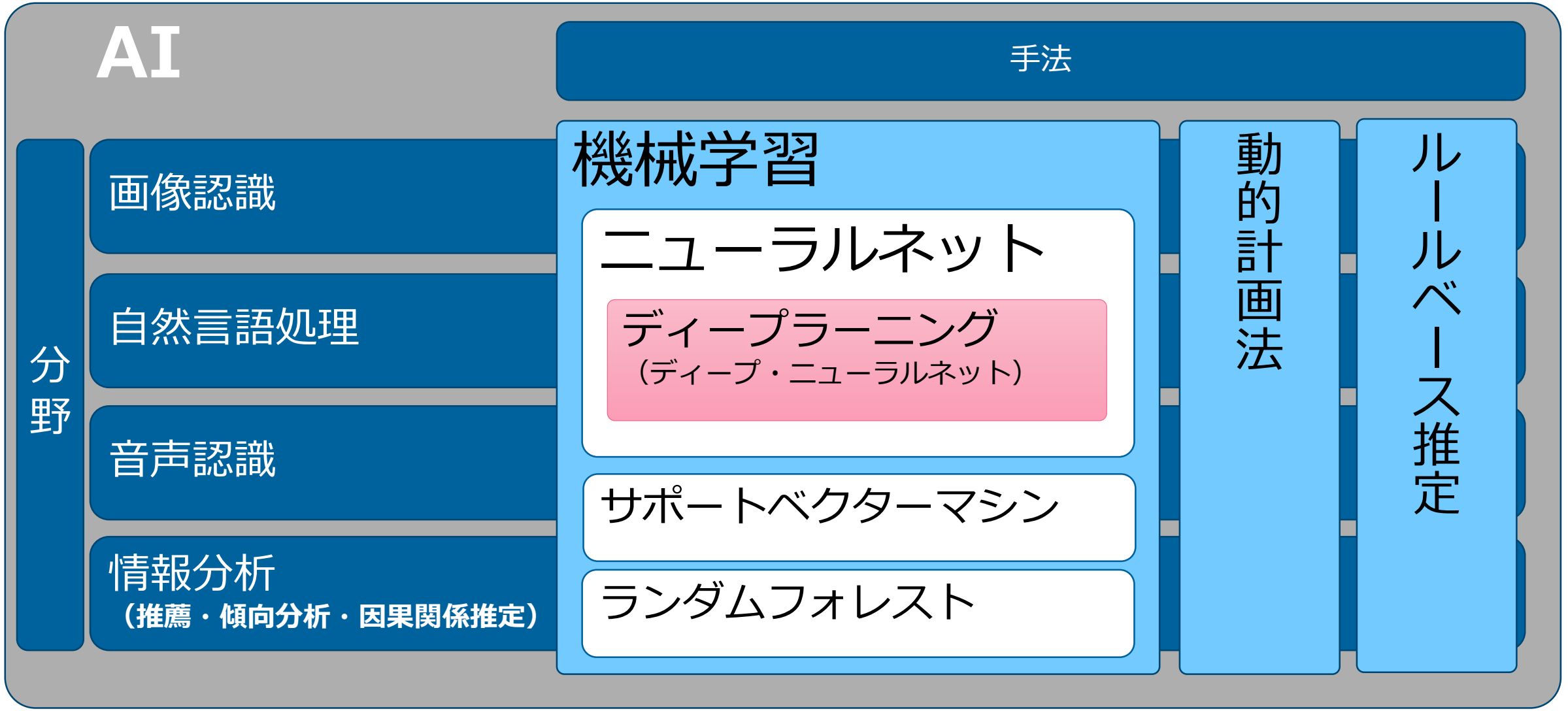


来館者数を推定するために「AI」を使う？

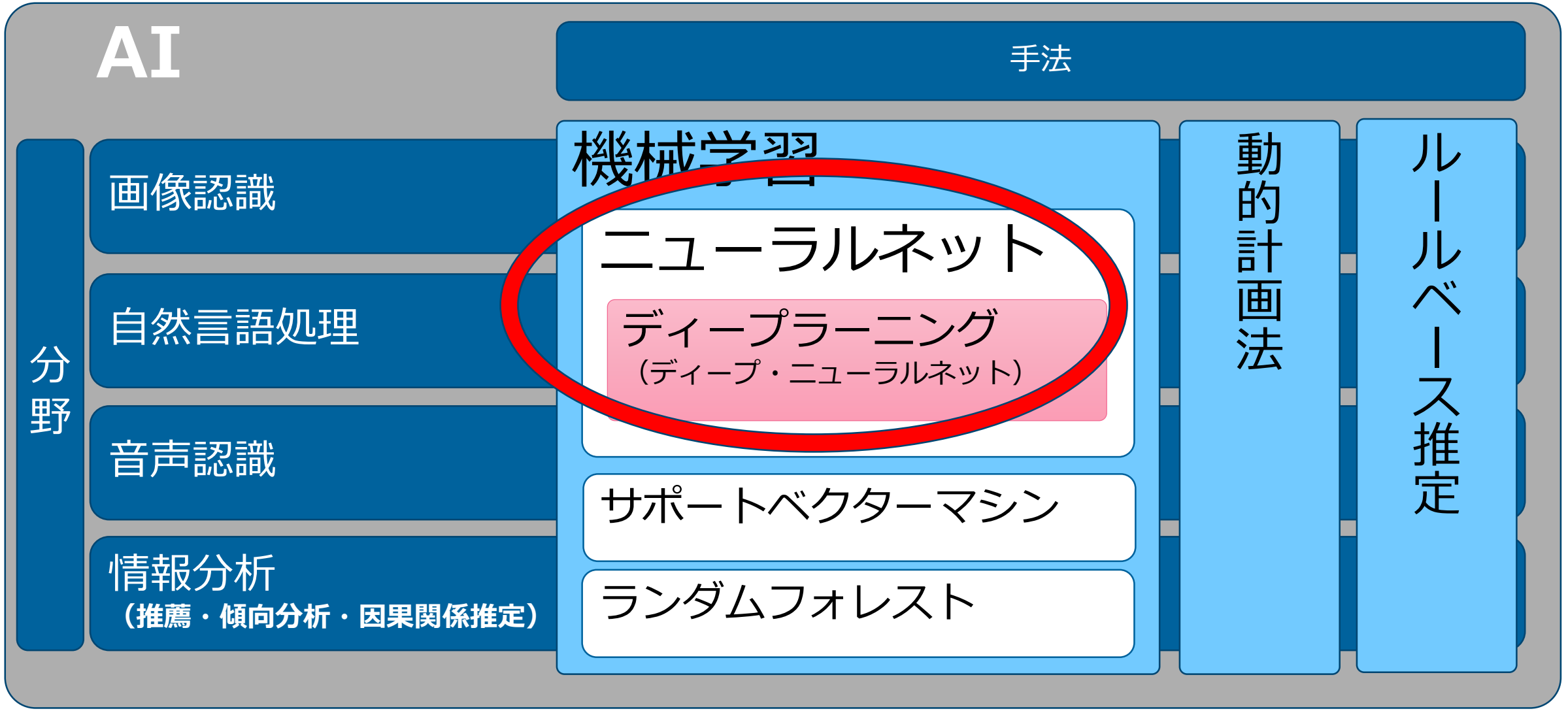
- 統計モデルを使った手法は解釈が容易
- 汎用性があり、計算も早い

- では、何にAIを使うのか？

まずは、AIの全体像


AI
手法
分野
画像認識
自然言語処理
音声認識
情報分析
（推薦・傾向分析・因果関係推定）
機械学習
ニューラルネット
ディープラーニング
（ディープ・ニューラルネット）
サポートベクターマシン
ランダムフォレスト
動的計画法
ルールベース推定

世間の注目を集めているのは、



ディープラーニングが性能を高めた研究課題

① 大量の学習データから特徴を取り出して分類する研究
→画像や文章の分類、画像から物体の検出

② 時系列(並び方の規則性)を学習する研究
→機械翻訳、質問応答、音声合成

③ 学習させたものと傾向の似たものを新たに生み出す研究
→白黒写真の着色、ラフスケッチの線画化

取組2：書誌からの書架分類推定

◆目的

- 書架分類（NDC/NDLC）及び件名標目を自動推定し候補として表示することで、書誌作成の支援を行う

◆方法

- タイトル（シリーズ名等を含む）・責任表示・出版者等の書誌情報を区分のないテキストデータとして単語に分解し、単語の組み合わせからの分類推測をニューラルネットで学習（文書やニュースを分類する技術の応用）
- 2016年以前の4,256,238件の当館書誌データから学習

評価結果とデモ

◆結果

- 約15万件の書誌データについての予測精度

予測内容	精度
NDC3桁（版の区別なし）	75%（候補Top3でよければ88%）
NDC2桁	82%
NDLC	63%
件名標目	43%

- 書誌情報ではなく、キーワード等を入力しても対応する分類が出力される
- レファ協の質問文と回答に付けられたNDC分類をデータに足すと発展する可能性も？

取組3：複写品質向上のための背景白色化

◆目的

- デジタル化された書籍の画像を対象に、文字以外の背景部分を白くすることで可読性と複写（プリントアウト）時の品質を向上させる

◆方法

- コントラスト調整などを人手で行った画像と元画像のペア用意し、生成系のディープラーニングを利用することで、元画像と補正画像の対応関係を学習
- 著作権切れの資料26点からなる画像4,555点とそれらを人手で修正した画像を学習データとして利用

Before

をここでのぞいて見よう。

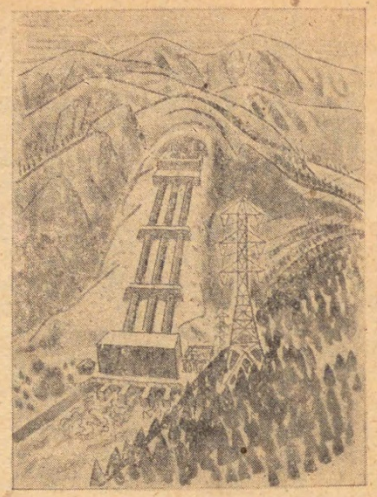
七　雷災防止

昭和十五年、日本學術振興會の中に、雷災防止委員會といふものが出來た。日本中のこの方面に關係した學者が數十人も集つて、互に協力して雷の研究を始めたのである。その研究は雷の害を避けること、卽ち雷災防止を目的としてゐるのである。

そんなことを聞くと、フランクリン以來、避雷針といふものが出來てゐるのに、今頃になつて何を研究するのか、と不思議に思ふ人があるかもしれない。しかし、今頃になつてかういふ委員會を作らねばならないほど、まだまだわからないことがいくらも殘されてゐるといふのが、いなむことのできない事實なのである。

例へば、避雷針にしても、絶對に大丈夫で安全だといふものはまだないのである。

それを研究してゐる人々の間でさへも、これが良い、あれが良いといろいろ違つた意見が出てゐる有樣である。その他にも、無線通信における空電の妨害は雷と密接な關係があつて、無線電送寫眞の繪が汚くなつたり、受信が出來なくなつたりするのも、雷の惡戲のせゐだといふ場合がたくさんある。また高壓送電線へ雷が落ちて生じる被害の問題なども、時局柄益〻重要になつて來てゐる。それで研究しなければならないことは、あとからあとからといくらでも盡きないのである。



第4圖

科學が進歩すると、いろいろの自然現象の本性が段々とわかつて來る。それで暴

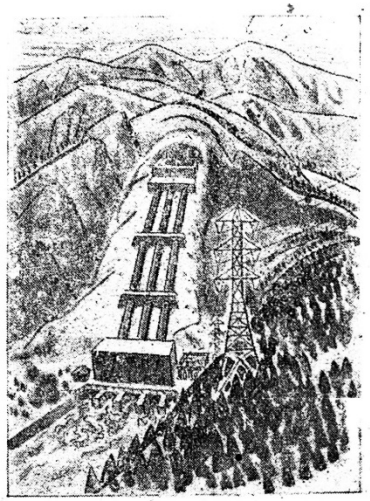
After

をここでのぞいて見よう。

七　雷災防止

昭和十五年、日本學術振興會の中に、雷災防止委員會といふものが出來た。日本中のこの方面に關係した學者が數十人も集つて、互に協力して雷の研究を始めたのである。その研究は雷の害を避けること、卽ち雷災防止を目的としてゐるのである。そんなことを聞くと、フランクリン以來、避雷針といふものが出來てゐるのに、今頃になつて何を研究するのか、と不思議に思ふ人があるかもしれない。しかし、今頃になつてかういふ委員會を作らねばならないほど、まだまだわからないことがいくらも殘されてゐるといふのが、いなむことのできない事實なのである。

例へば、避雷針にしても、絶對に大丈夫で安全だといふものはまだないのである。それを研究してゐる人々の間でさへも、これが良い、あれが良いといろいろ違つた意見が出てゐる有樣である。その他にも、無線通信における空電の妨害は雷と密接な關係があつて、無線電送寫眞の繪が汚くなつたり、受信が出來なくなつたりするのも、雷の惡戲のせゐだといふ場合がたくさんある。また高壓送電線へ雷が落ちて生じる被害の問題なども、時局柄益〻重要になつて來てゐる。それで研究しなければならないことは、あとからあとからといくらでも盡きないのである。



第4圖

科學が進歩すると、いろいろの自然現象の本性が段々とわかつて來る。それで暴

（２）デジタル化画像の活用事例

取組4：OCRの精度向上

取組5：挿絵検索

国立国会図書館デジタルコレクション

- 国立国会図書館デジタルコレクション(デジコレ)でインターネット公開されている著作権保護期間満了資料は、図書資料だけでも**40万点**以上(2018/11現在)
- コマ数に換算すると**3000万コマ**以上！

資料画像を用いた研究を行うモチベーション

- 先進的な技術を取り入れて、検索性と発見性を高め、より良いサービス提供につなげたい
- デジタル化した資料の利活用促進策として、機械学習用データセットの公開も進めたい

取組4：OCRの精度向上

◆課題

1. 認識精度の向上
2. 古い字形の認識
3. 文字以外の部分を無理やり文字として認識してしまう/文字領域が飛ばされる

◆取組

- 3.について、文字の書かれた領域(本文や表)とその他の画像領域(写真や挿絵)の分離抽出をDeep Learningで実現。OCRのプロセスに組み込んで評価中。

◆副産物

- 抽出された画像領域についても、類似画像の検索によるサービス提供が行えるのではないか？

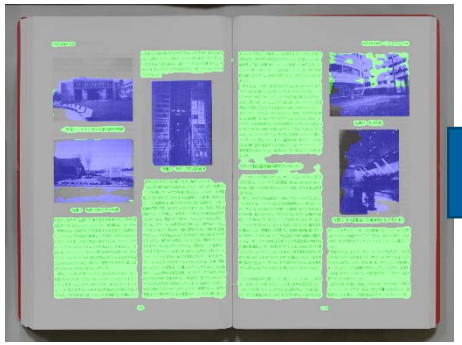
⇒挿絵検索

取組5：挿絵検索

自動抽出～検索までのフローチャート

元画像

文字領域と画像領域を自動認識
(Deep Learningを利用した機械学習)



緑色の文字領域はOCR精度向上に利用

切り出された画像領域

各画像の特徴を
ベクトル表現に変換

画像同士の特徴の近さ
を画像検索に利用

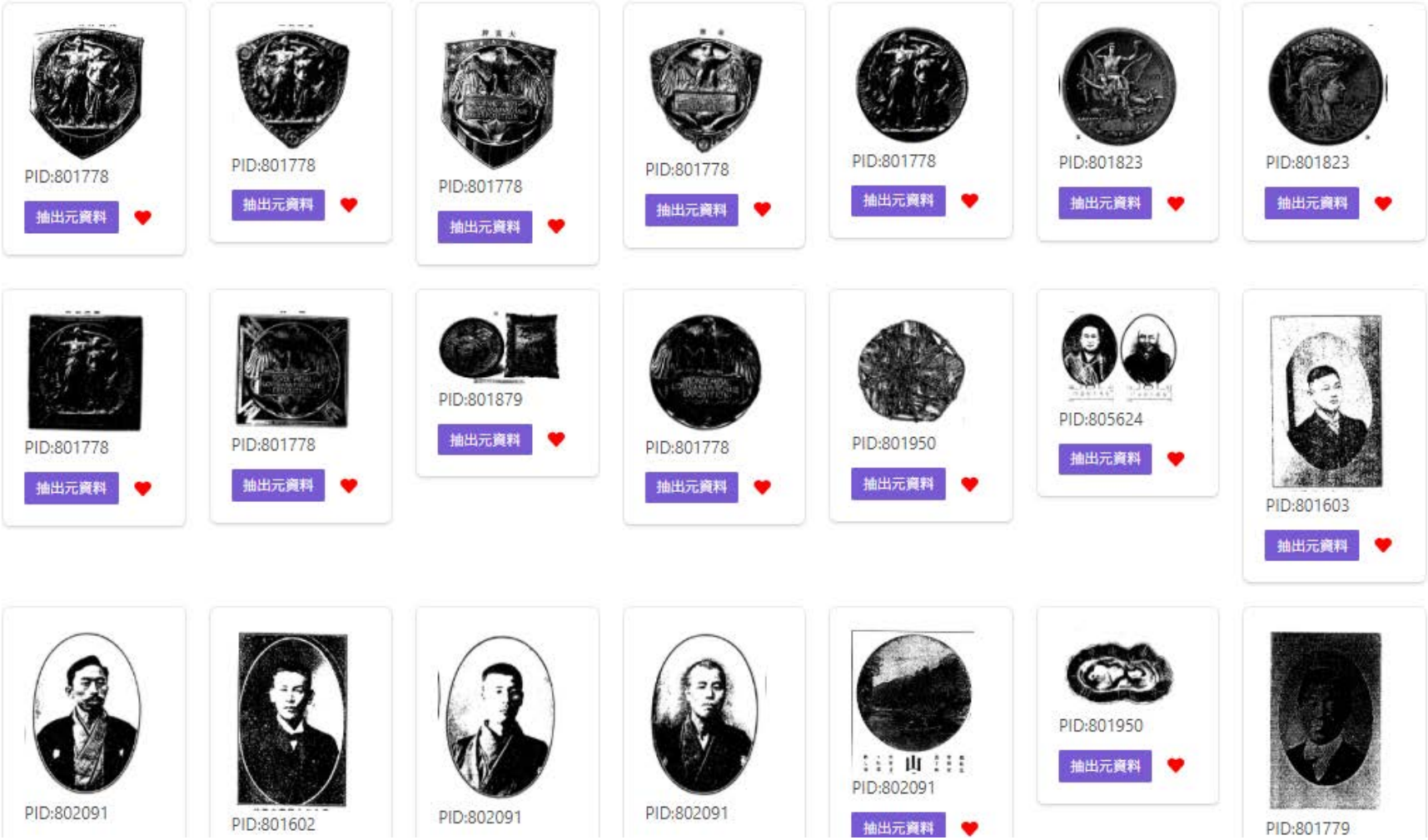
実験データ

- NDCの大分類が6(産業)の著作権保護期間満了資料が対象
- 資料1,481点、194,916コマから切り出された、計284,134か所*の挿絵・写真領域の画像

*毛筆による手書き文字等が「絵」と認識される場合があるため、実際の挿絵・写真領域よりも多くの領域が抽出される。が、検索上は上位に来ないので影響は少ない。

検索例(万博記念メダル)

万国博覧会の記念メダル(左上)で検索した結果

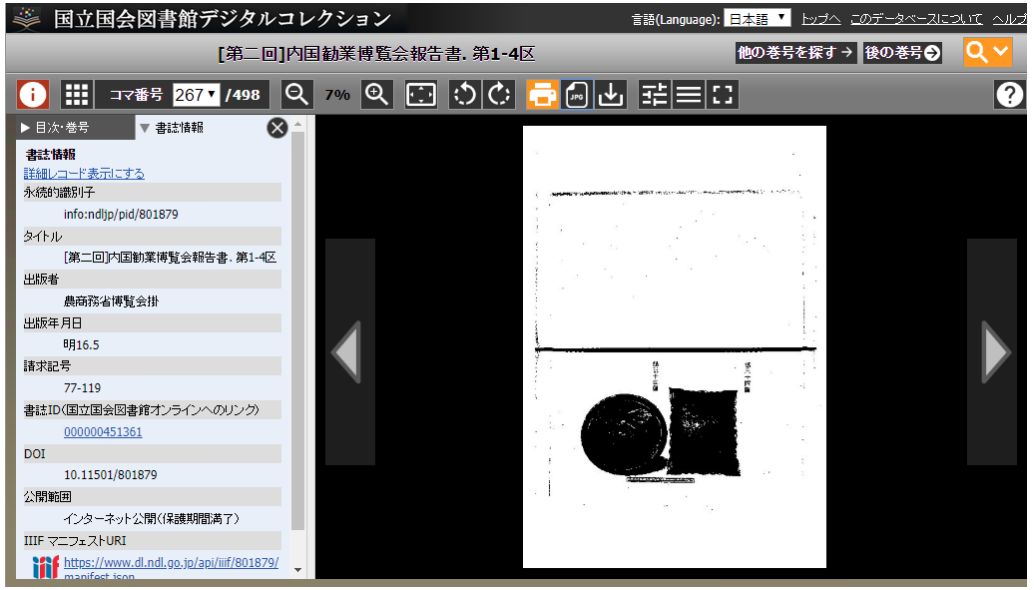


←メダルが見つからなくなると
構図の似た肖像画がヒットする

提供元のデジコレではどう見える？

聖路易万国博覧会本邦参同事業報告. 第1編　323コマ目

[第二回]内国勧業博覧会報告書. 第1-4区 267コマ目

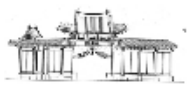


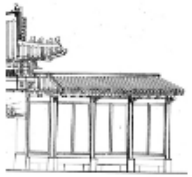
千九百年巴里万国博覧会臨時博覧会事務局報告. 下 457コマ目

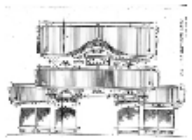
→画像の領域だけを抽出して検索できている

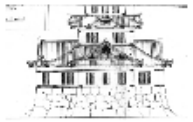
検索例(図面)

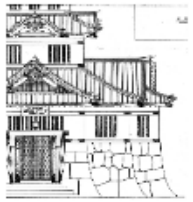
検索画像


PID:801797
抽出元資料


PID:801797
抽出元資料


PID:801779
抽出元資料

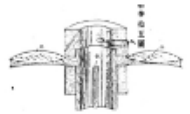

PID:801796
抽出元資料


PID:801796
抽出元資料


PID:801797
抽出元資料


PID:801779
抽出元資料


PID:801797
抽出元資料


PID:801789
抽出元資料


PID:801779
抽出元資料


PID:801797
抽出元資料


PID:801779
抽出元資料


PID:801779
抽出元資料

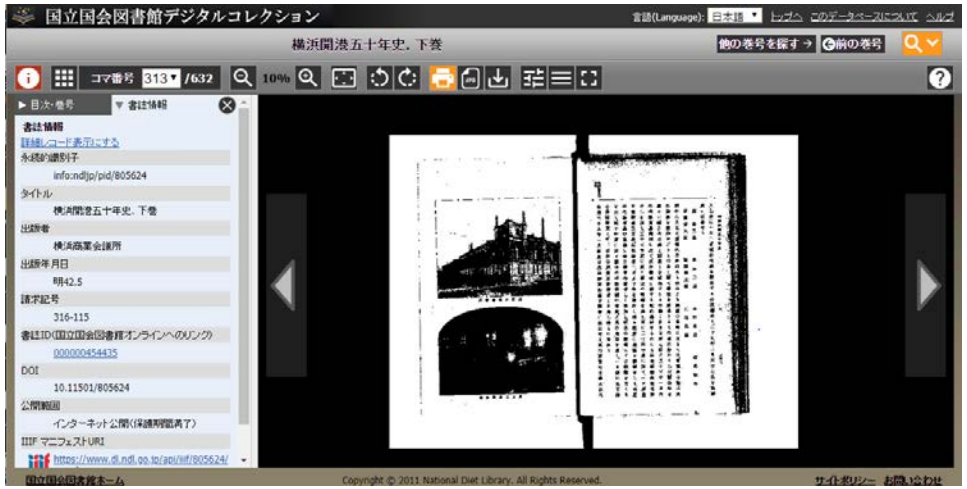

PID:801797
抽出元資料

検索例(建造物の写真)

検索画像

著作権保護期間満了資料にはマイクロ資料が多く、風景写真による検索は難しいが、横浜の生糸検査所の外観の写真を、異なる資料から発見している。

生糸検査所の写真のある2資料について



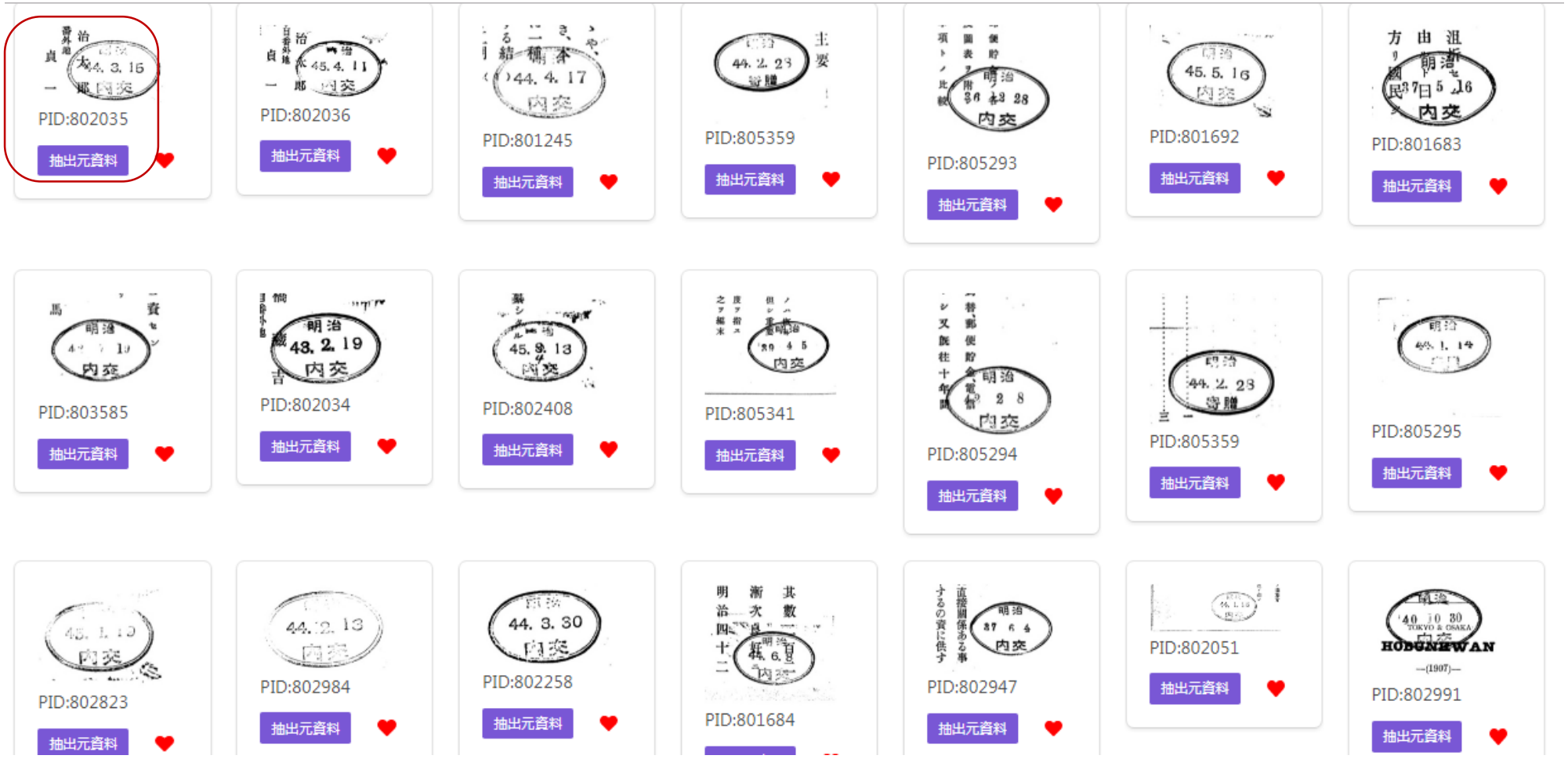
横浜商業会議所 (1909)『横浜開港五十年史・下巻』
(http://dl.ndl.go.jp/info:ndljp/pid/805624)
313コマ目

神奈川県内務部(1911)『神奈川県産業要覧』
(http://dl.ndl.go.jp/info:ndljp/pid/801245)
42コマ目

タイトルや目次には言及がなく、マイクロ資料のためOCRも難しいが、画像検索によって同一の施設に言及していることがわかった。

こんな使い方も可能(内務省交付本を印影から検索)

検索画像



東京書籍館の蔵書印を検索


PID:801760

抽出元資料


PID:801764

抽出元資料


PID:801741

抽出元資料


PID:801761

抽出元資料


PID:801751

抽出元資料


PID:801750

抽出元資料


PID:801759

抽出元資料


PID:801743

抽出元資料


PID:841911

抽出元資料


PID:801757

抽出元資料


PID:801745

抽出元資料


PID:841910

抽出元資料

ただし……
似た蔵書印は判別困難(例：東京図書館と帝国図書館)

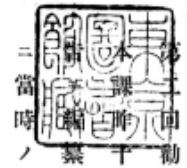
検索画像

帝国

PID:802079

抽出元資料

東京



PID:801305

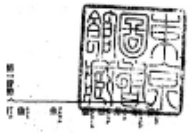
抽出元資料

帝国

PID:804175

抽出元資料

東京



PID:803099

抽出元資料

帝国

PID:805292

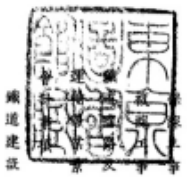
抽出元資料

帝国

PID:801398

抽出元資料

東京



PID:805398

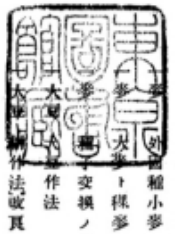
抽出元資料

帝国

PID:805471

抽出元資料

東京



PID:802494

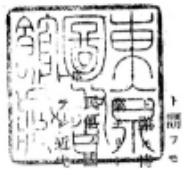
抽出元資料

帝国

PID:805293

抽出元資料

東京



PID:804986

抽出元資料

東京

PID:802428

抽出元資料

東京

PID:804286

抽出元資料

帝国

PID:801183

抽出元資料

デモ

次世代デジタルライブラリー（仮称）

- 今回ご紹介したNDCの大分類が6の著作権保護期間満了資料の画像検索機能について、NDLラボのページから実験サービスとして今年度内の公開を目指している

これまでの研究成果を活用した他の機能としては、
「複写品質向上のための背景白色化」
「OCRテキストによる全文検索機能」
「見開き資料画像の自動分割・背景除去機能」
を搭載予定

共同研究やデータ利活用事例の募集

今回紹介したような技術、及び当館データの利活用に関心のある方のご連絡お待ちしております。

lab@ndl.go.jp