

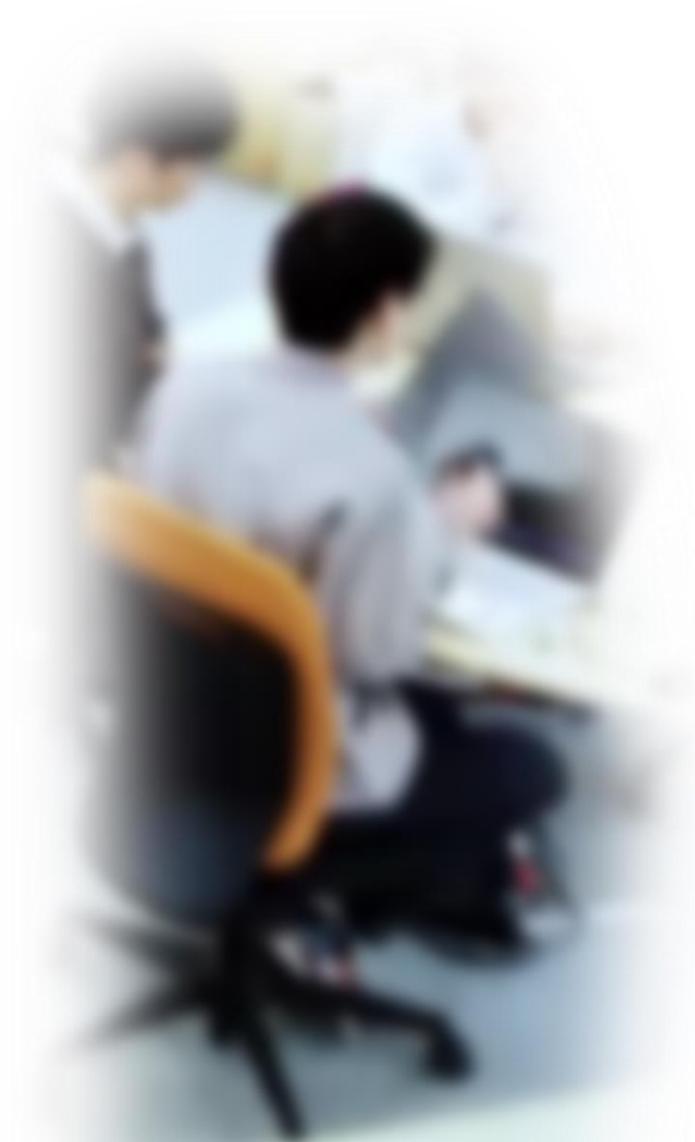
ae-12.現代のAI技術革新： TransformerからNoShot学習まで (画像理解，自然言語処理，時系 列データ処理の発展)

(AI 演習) (全15回)

<https://www.kkaneko.jp/ai/ae/index.html>

金子邦彦





アウトライン

1. イントロダクション
2. 時系列データ
3. 自然言語処理とチャットボット
4. 画像理解
5. Transformer と Vision Transformer (ViT)
6. 汎用性を持つニューラルネットワーク

人工知能

知的なITシステム

機械学習

データから**学習**し、知的能力を
向上

ディープラーニング

データから**学習**し、複雑なタスク
を実行。**多層のニューラル
ネットワーク**を使用

機械学習の基本



機械学習は、**コンピュータ**が**データ**を使用して**学習**することにより**知的能力を向上**させる技術

- **情報の抽出**：データからパターンや関係性を自動で見つけ出す能力を持つ
- **知的なタスクの実行**：予測，分類などの知的なタスクを実行



機械学習の3つの特徴



1. **情報の抽出**：データからパターンや関係性を自動で見つけ出す能力を持つ
2. **簡潔さ**：人間が設定していたルール等を，自動生成できる
3. **限界の超越**：従来の方法では困難だった課題も解決できる可能性がある

11-2. 画像理解と姿勢推定

画像理解の主な種類

① 画像分類

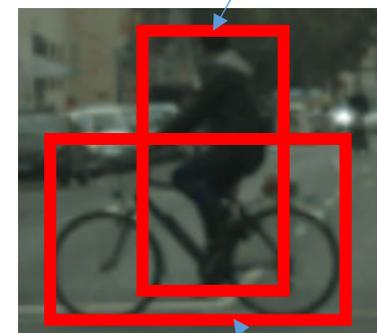
「何があるか」を理解



person
bicycle

② 物体検出

場所と大きさも理解



person

bicycle

③ セグメンテーション

画素単位で理解



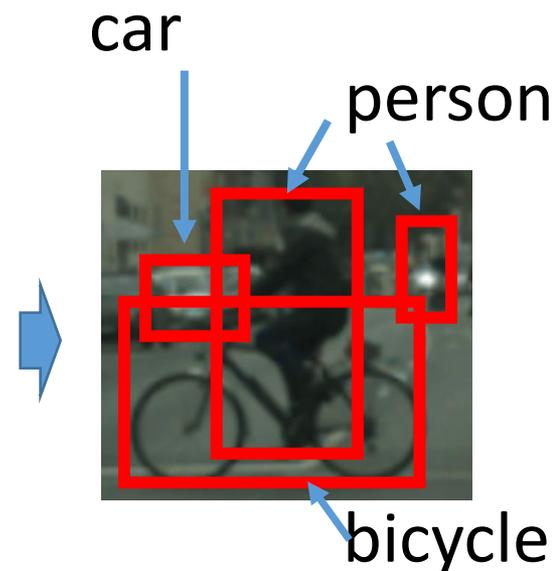
① 画像分類



```
Score 0.9827020168304443, Label lab_coat  
Score 0.0030872616916894913, Label syringe  
Score 0.0024311079178005457, Label beaker  
Score 0.0016609227750450373, Label stethoscope  
Score 0.00037950885598547757, Label plate
```

画像分類の結果は、ラベルと確率
※ 5つの候補 (top 5) が表示されている

② 物体検出



バウンディングボックス,
ラベルを得る

バウンディングボックスは、
物体を囲む最小のボックス（四角形）

③ セグメンテーション



物体の形を画素単位で抜き出し



ラベルを得ることもできる

画像の畳み込み

Input

0	1	1	0	1
0	1	1	0	1
0	1	1	0	1
0	1	1	0	1
0	1	1	0	1

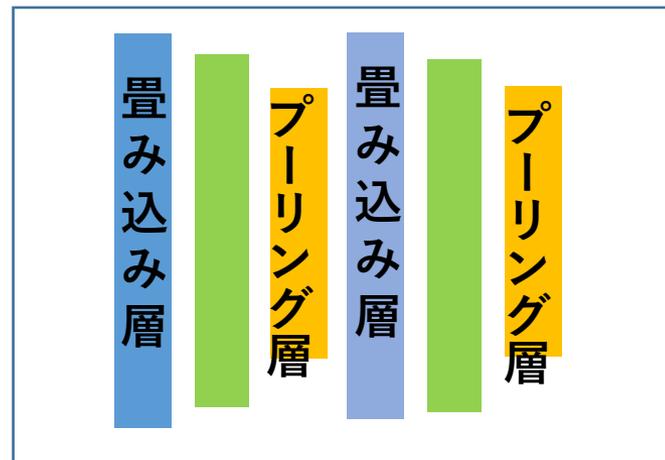
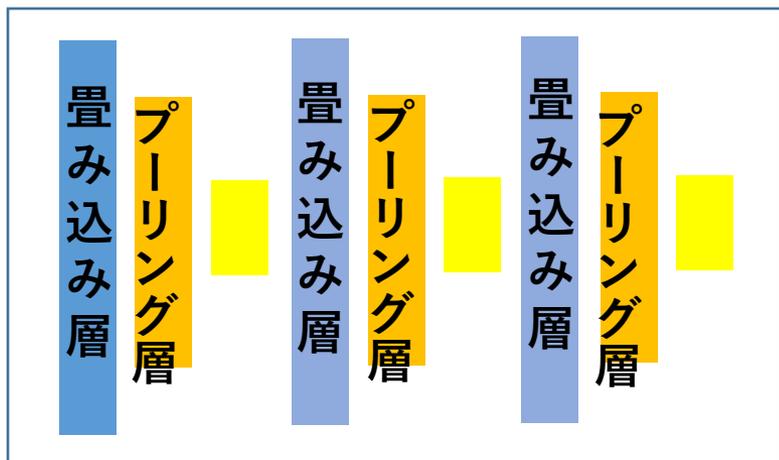
元画像 (5 × 5 マス)

Filter / Kernel

1	0	1
1	1	1
0	0	1

カーネル (3 × 3 マス)

畳み込みニューラルネットワーク (CNN) の例



さまざまなバリエーション

畳み込みニューラルネットワークでのパターン認識

「畳み込みニューラルネットワークの利用により、さまざまなレベルのパターンを抽出・認識できるようになる」という考える場合も

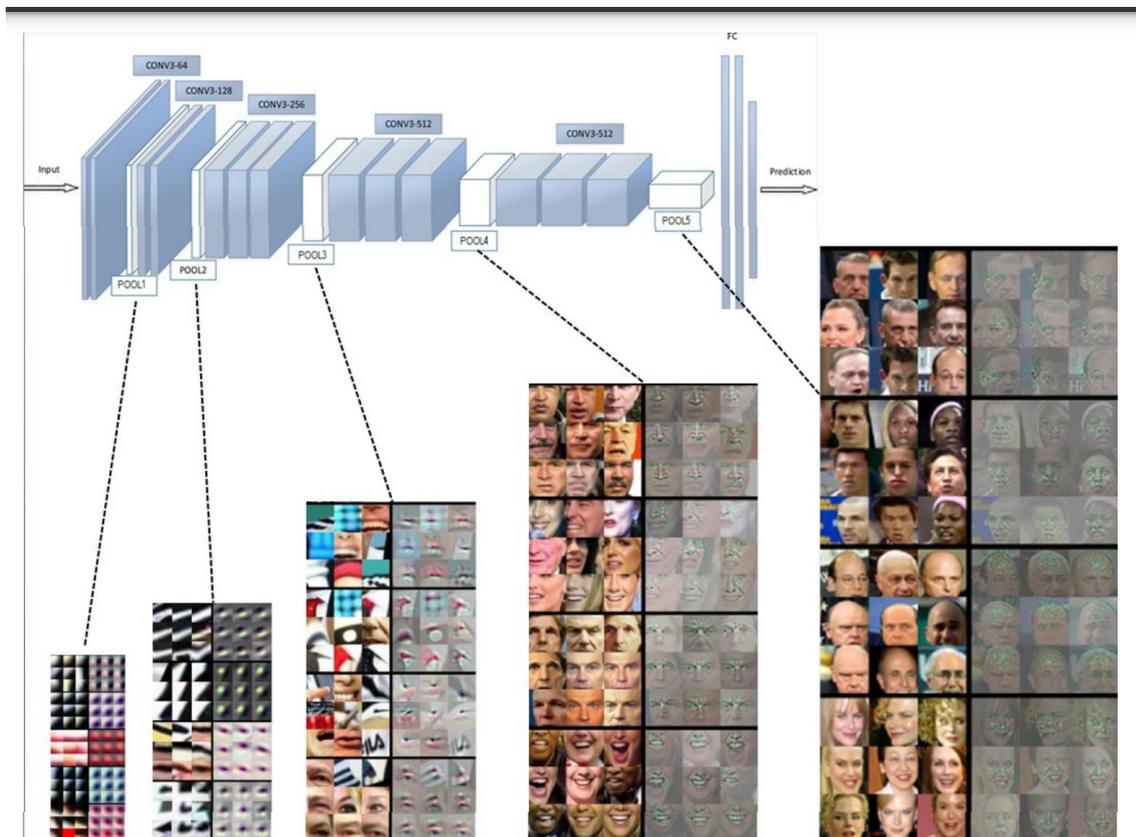


Fig. 2

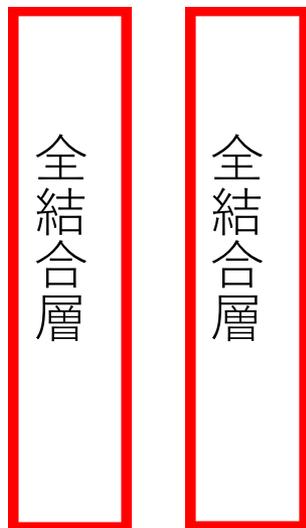
Mei Wang, Weihong Deng,

Deep Face Recognition: A Survey, arXiv:1804.06655, 2018.

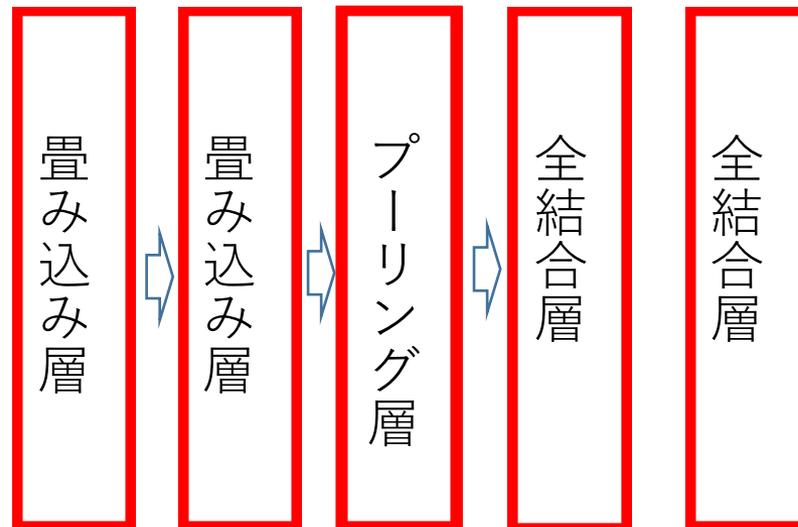
畳み込みニューラルネットワークの効果



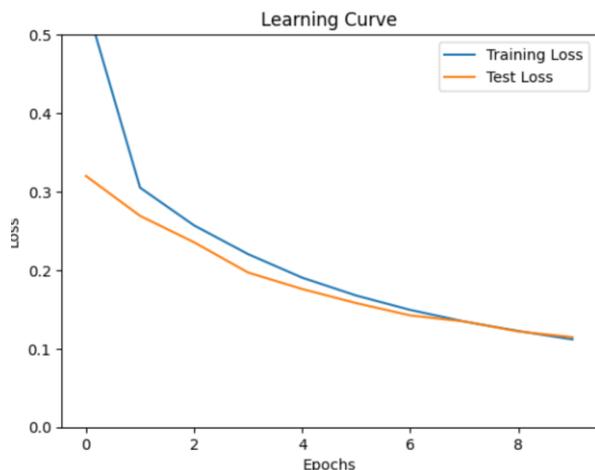
単純なニューラルネットワークの例



畳み込みニューラルネットワークの例

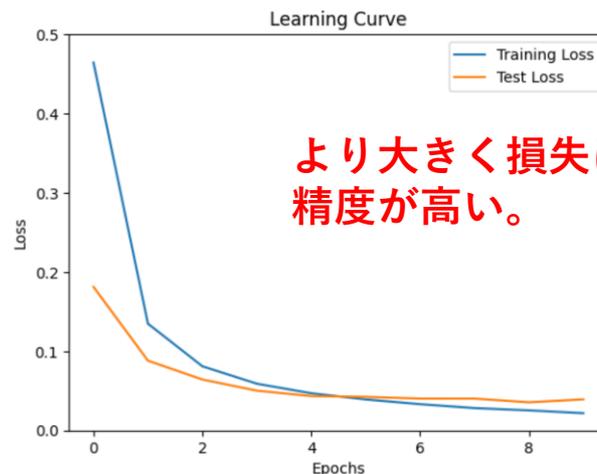


手書き文字 (0から9) MNIST の分類での損失



学習の繰り返しにより、損失は減少

手書き文字 (0から9) MNIST の分類での損失



学習の繰り返しにより、損失は減少

物体検出の基本的な考え方

元画像から
切り出す



最初の区切り

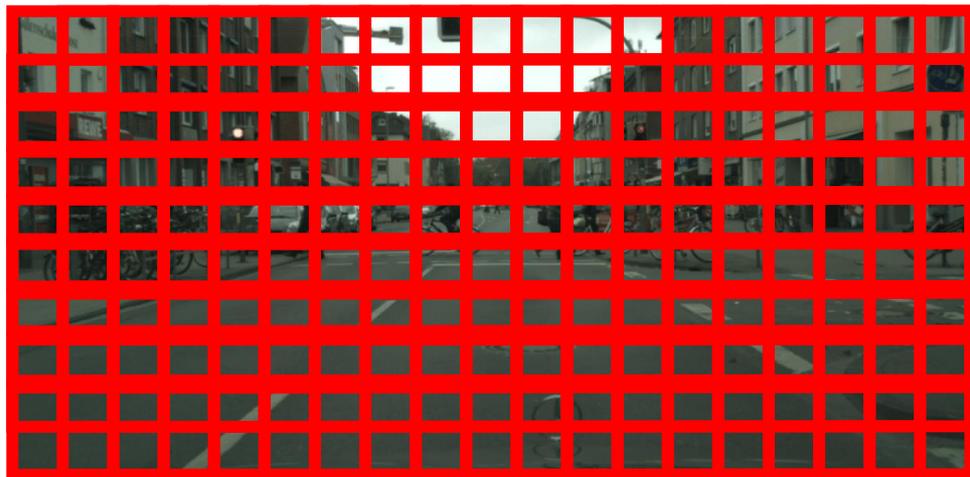


画像分類



結果

物体検出の基本的な考え方



区切りごとに画像分類を行う。

姿勢推定

姿勢推定は、人間の全身（人体）、頭部、およびその他のオブジェクトの**位置と方向**を推定する技術

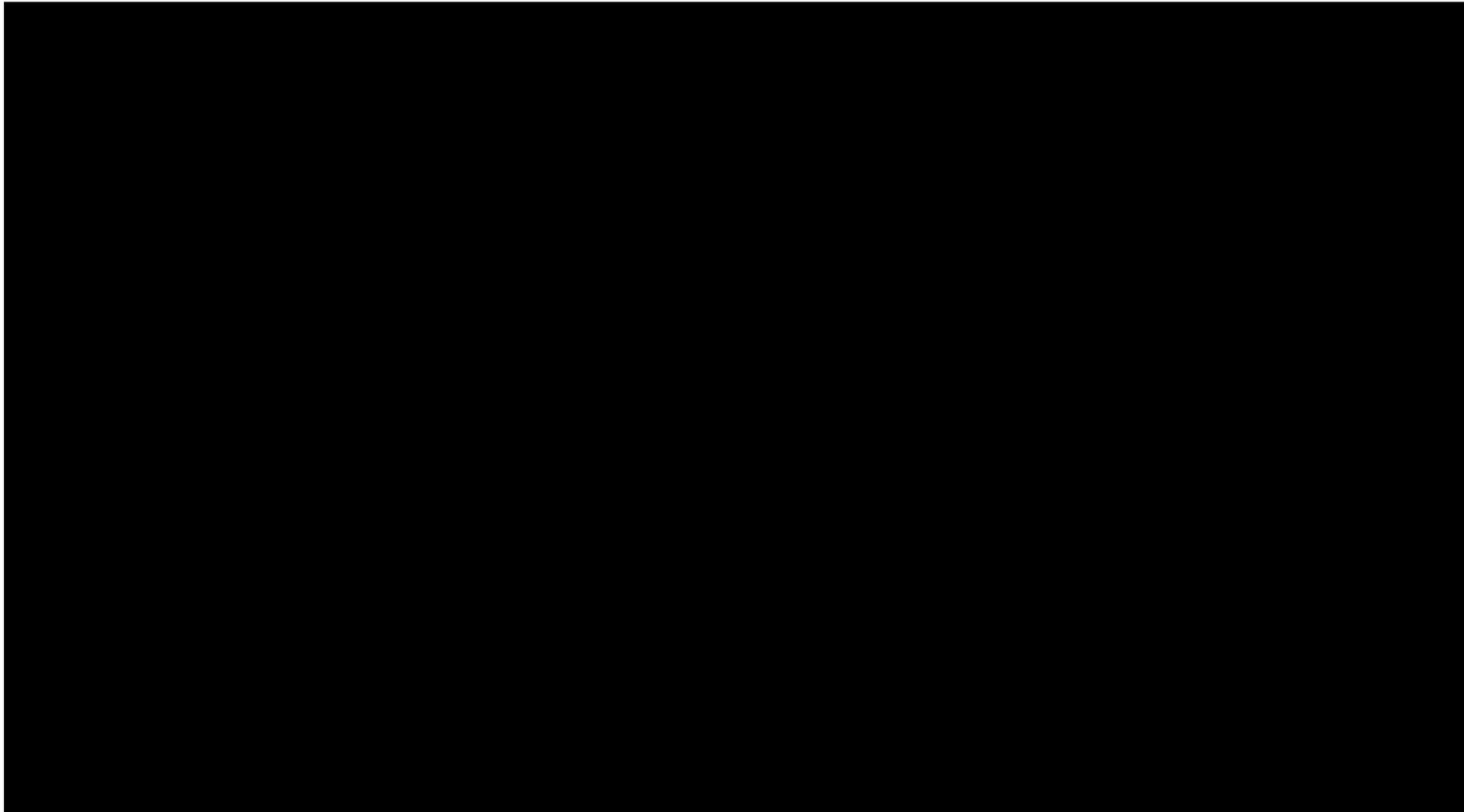


姿勢推定の結果

姿勢推定のビデオの例



姿勢推定のビデオの例



姿勢推定でのキーポイントの検出

元画像

①部位の
位置推定

②同一人物の
キーポイント特定



(a)



(b)



(c)

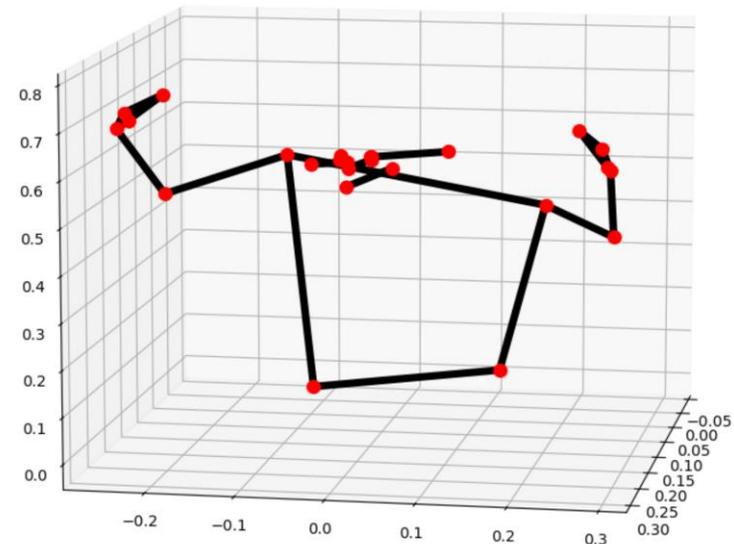
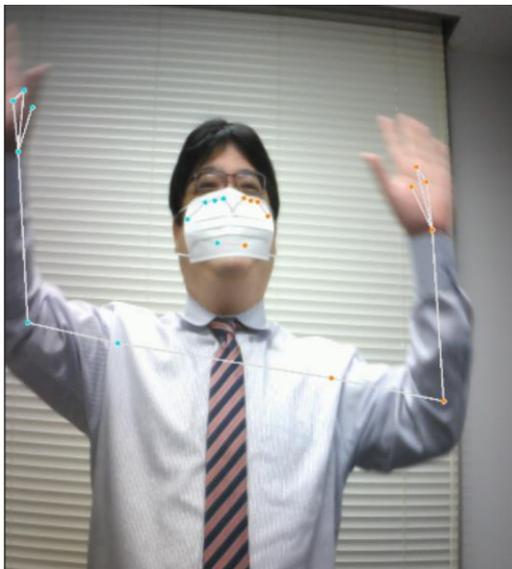
Fig. 6 An example of heatmap-based single-person pipeline with heatmap. (a) Original image, (b) heatmap generated by estimator, and (c) detection result.

発展演習 1 : 人体の 3 次元姿勢推定

- Google Colaboratory のページ

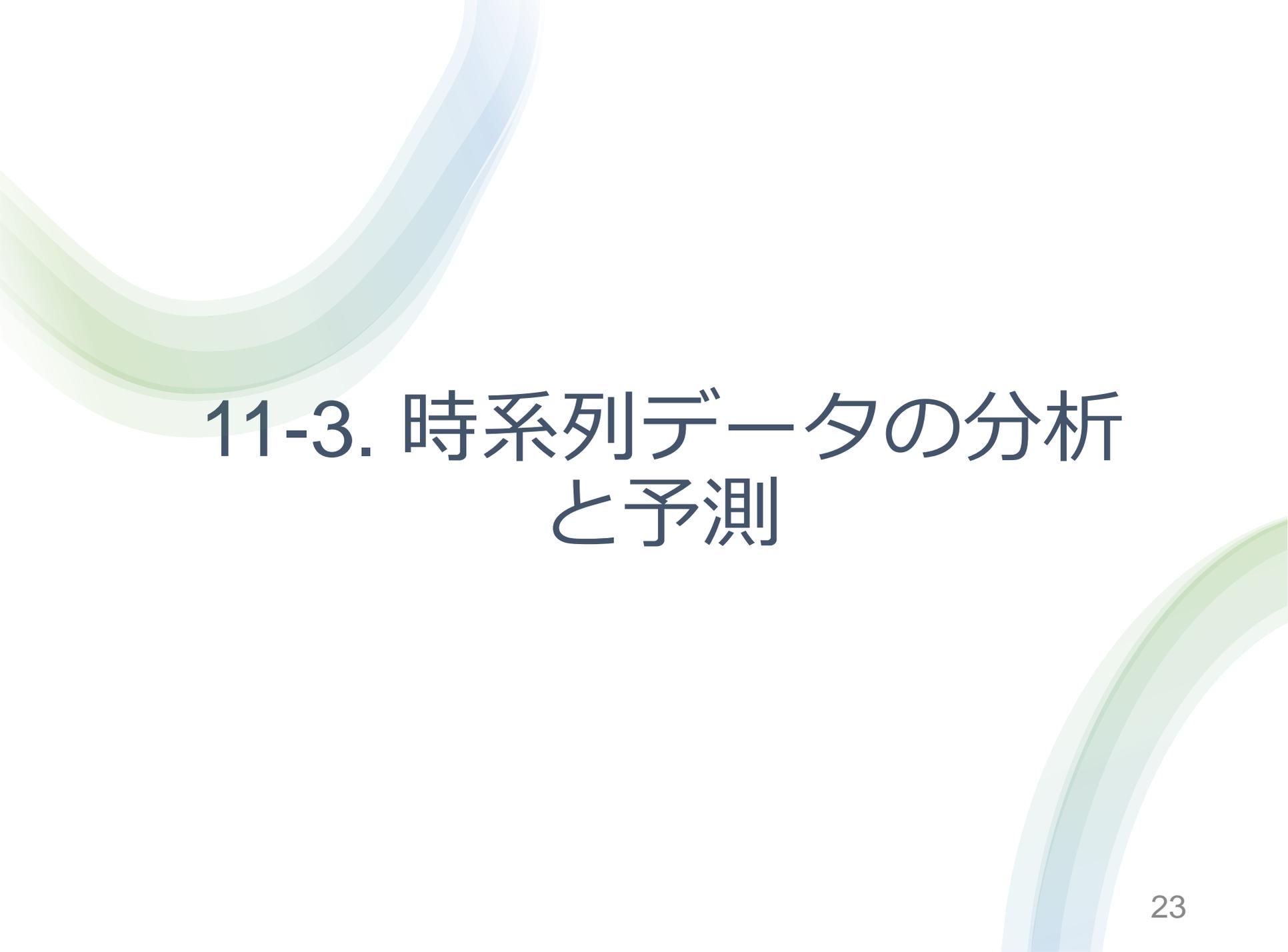
https://colab.research.google.com/drive/13nOMSWODzx_LjN9XEG99jtvGMAcl4m9V?usp=drive_link

- 人体の 3 次元姿勢推定のプログラム, 説明, 実行結果を記載
- 自分で画像を用意して試用可能 (Google アカウントが必要)



姿勢推定のまとめ

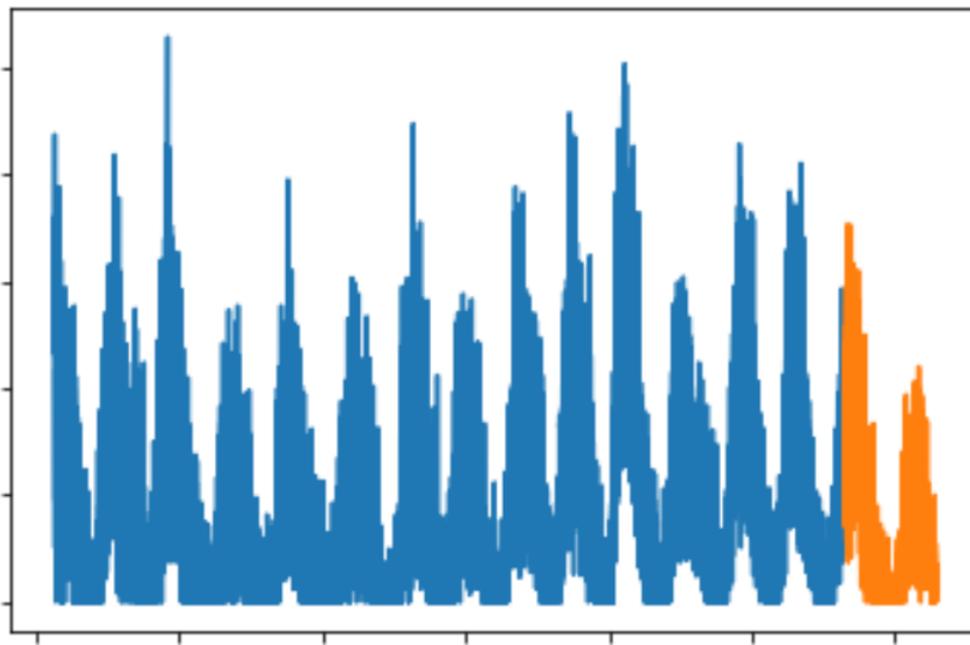
- **姿勢推定**：人物やオブジェクトの空間的な位置や向きを推定する技術である。人体の関節位置や頭部の向きなど、対象の姿勢に関する3次元的な情報を抽出することができる。
- **キーポイント検出**：姿勢推定において、人体の関節や特徴的な部位の位置を特定する技術である。検出されたキーポイントを接続することで、人体の骨格構造を推定することができる。
- **ヒートマップ**：姿勢推定において、キーポイントの位置確率をマップとして表現する手法である。



11-3. 時系列データの分析 と予測

時系列データの分析と予測は、機械学習の重要な応用分野である。

- 時系列データの例：太陽の黒点数の変化
- 予測の例：過去データをもとに未来を予測



太陽の黒点数
の変化

予測

時系列データ

時系列データは、時間の経過に伴って順序付けられたデータ
の並び

例

2024年12月3日の気温は15度

2024年12月4日の気温は13度

...

時系列データの例：太陽黒点観測データ

11314	1848	12	23	1848.977	353	23.8	1
11315	1848	12	24	1848.980	240	19.6	1
11316	1848	12	25	1848.982	275	21.0	1
11317	1848	12	26	1848.985	352	23.8	1
11318	1848	12	27	1848.988	268	20.8	1
11319	1848	12	28	1848.990	285	21.4	1
11320	1848	12	29	1848.993	343	23.5	1
11321	1848	12	30	1848.996	340	23.4	1
11322	1848	12	31	1848.999	238	19.6	1
11323	1849	1	1	1849.001	287	20.9	1
	year	month	day	dec_year	sn_value	sn_error	obs_num
72855	2017	6	21	2017.470	35	1.0	41
72856	2017	6	22	2017.473	24	0.8	39
72857	2017	6	23	2017.475	23	0.9	40
72858	2017	6	24	2017.478	26	2.3	15
72859	2017	6	25	2017.481	17	1.0	18
72860	2017	6	26	2017.484	21	1.1	25
72861	2017	6	27	2017.486	19	1.2	36
72862	2017	6	28	2017.489	17	1.1	22
72863	2017	6	29	2017.492	12	0.5	25
72864	2017	6	30	2017.495	11	0.5	30

年月日の値

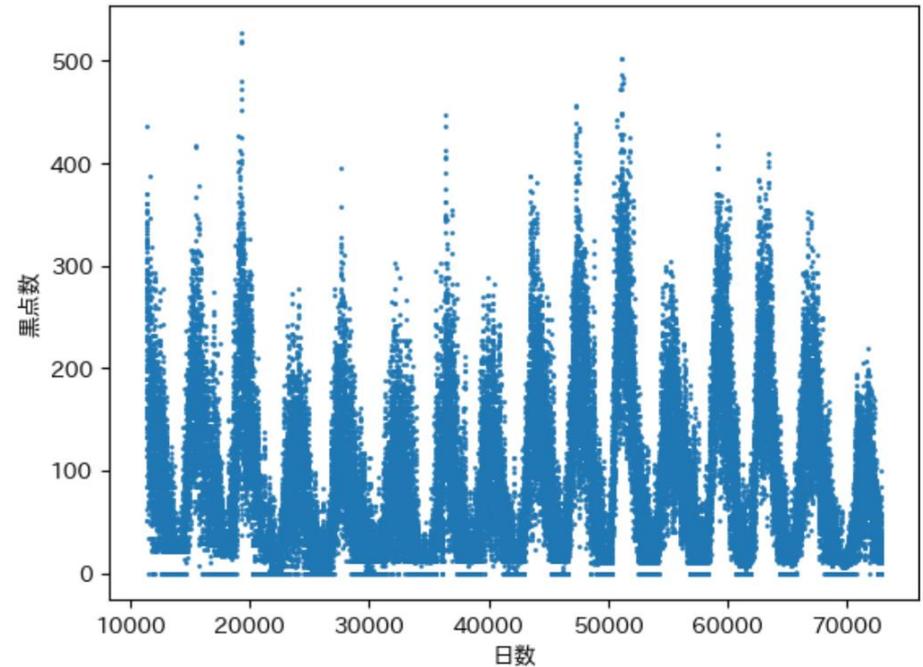
黒点数

黒点数の観測データ

1848年12月23日～2017年6月30日の
毎日の黒点数データが公開されている

https://data.heatonresearch.com/data/t81-558/SN_d_tot_V2.0.csv

太陽黒点数の散布図



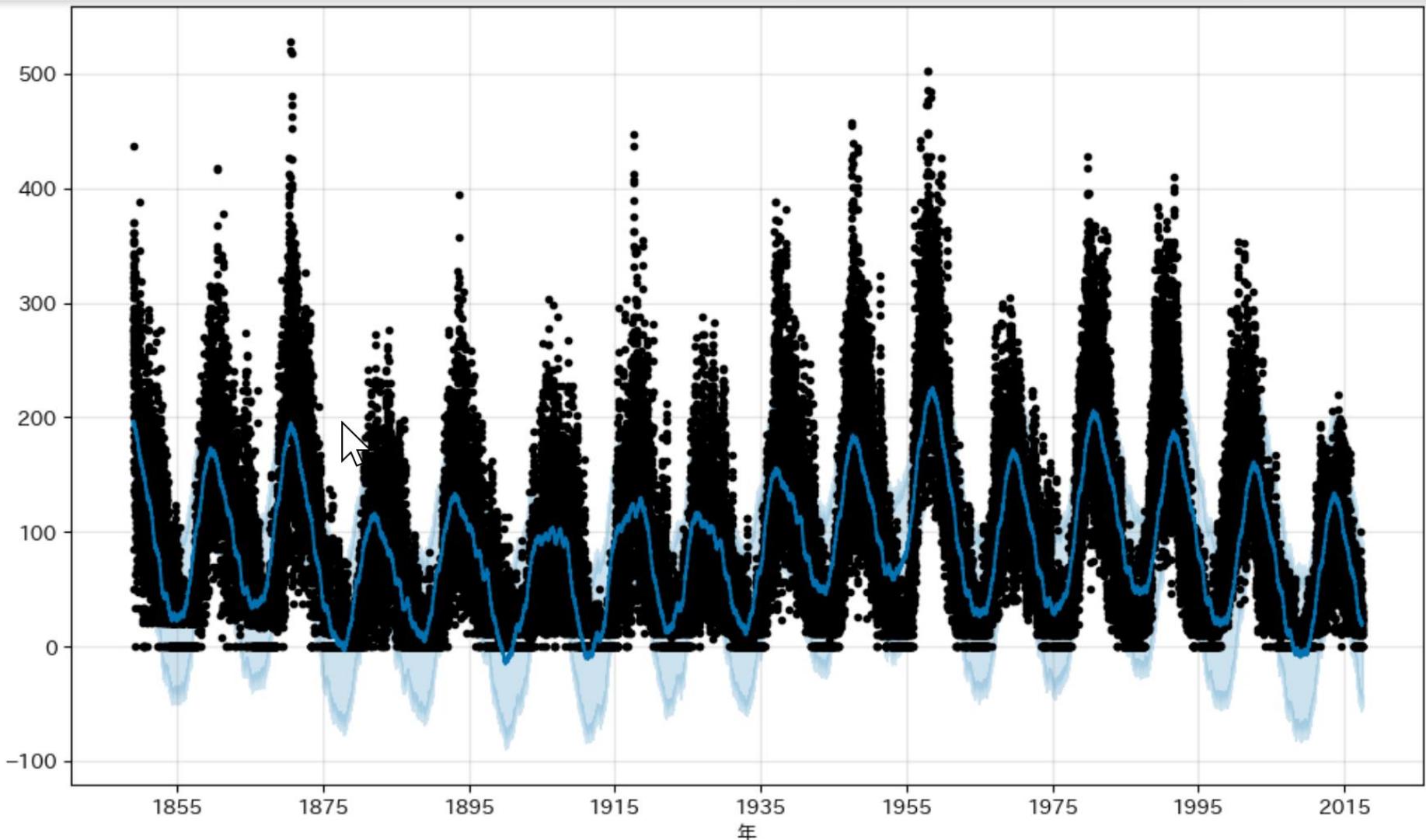
散布図でプロットすることで、
時間的な変化を視覚化

時系列データの特徴

- **周期性**：日、週、月、季節などの**定期的な間隔で繰り返されるパターン**
例：自然現象、社会的活動
- **トレンド**：時間の経過による**増加、減少、一定レベルの維持などの方向性**
- **特定のイベントや時期**（例えば正月、学校の学期開始時期など）との**関連性**

分析の例（太陽黒点データ） Prophet を使用

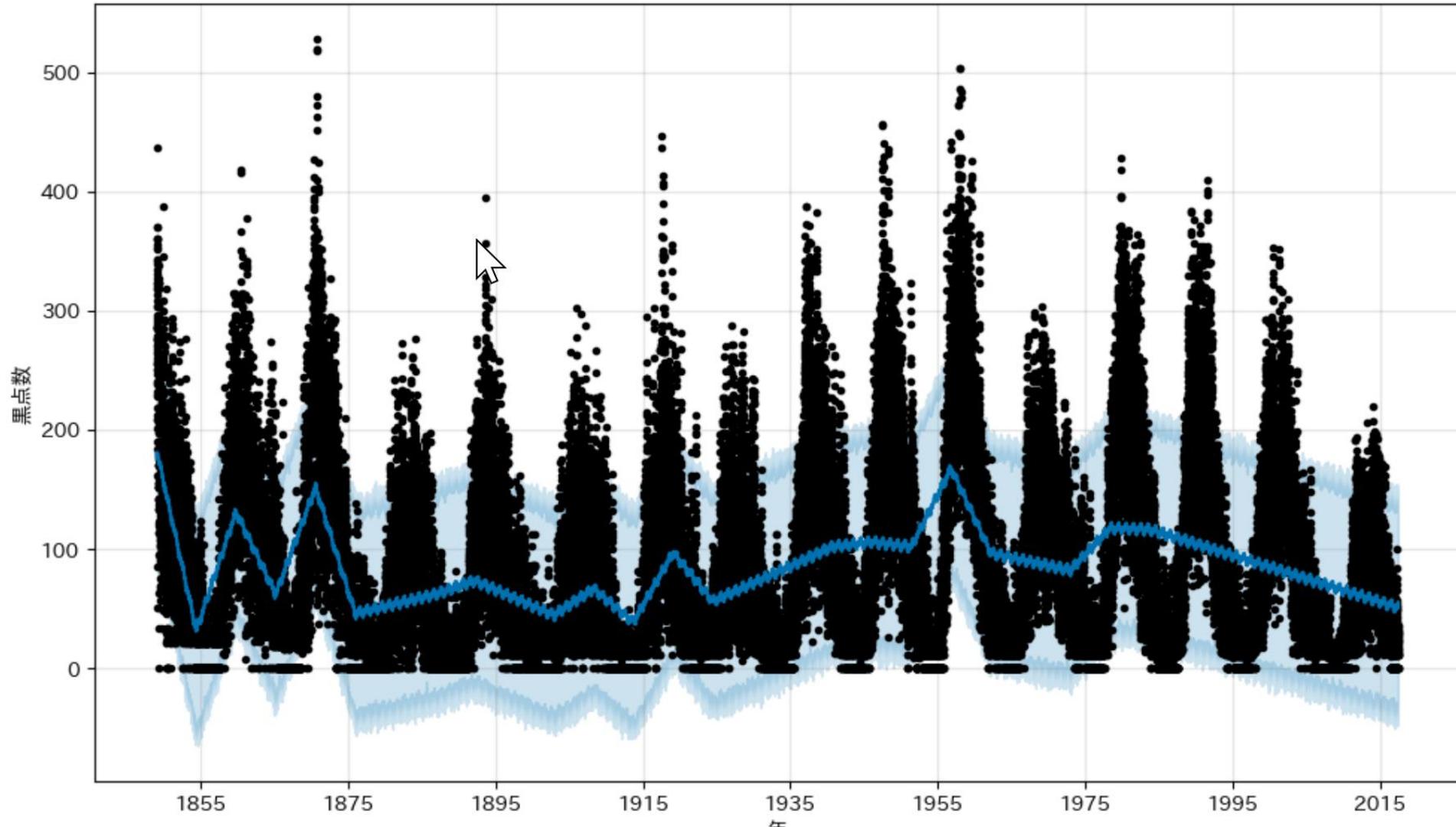
時系列データの分析には、Prophetというツールが有効である。
黒：元データ 水色：分析結果（短期的な変動を除去）



分析の例（太陽黒点データ） Prophet を使用

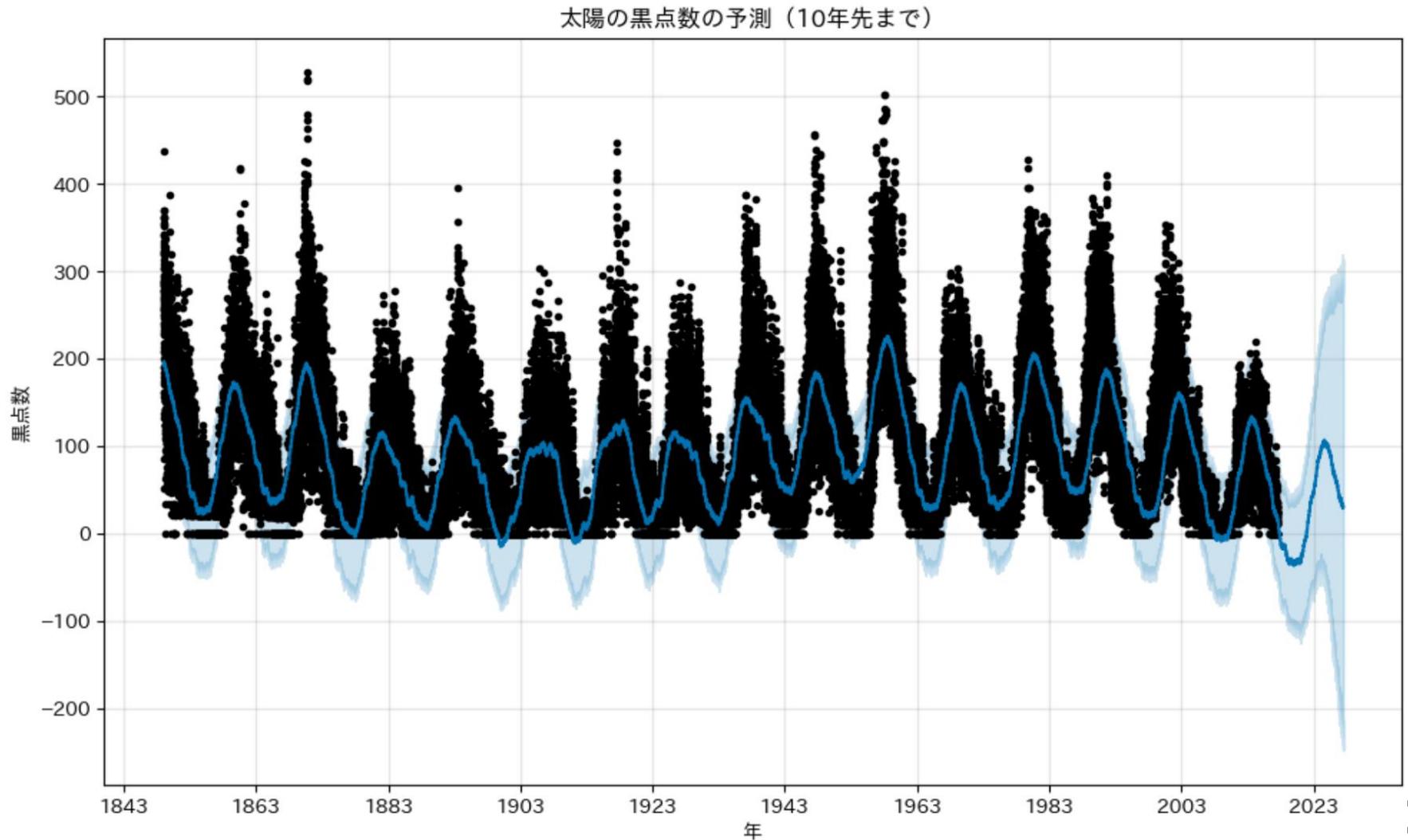
黒色：元データ 水色：元データから算出されたトレンド

太陽の黒点数の分析



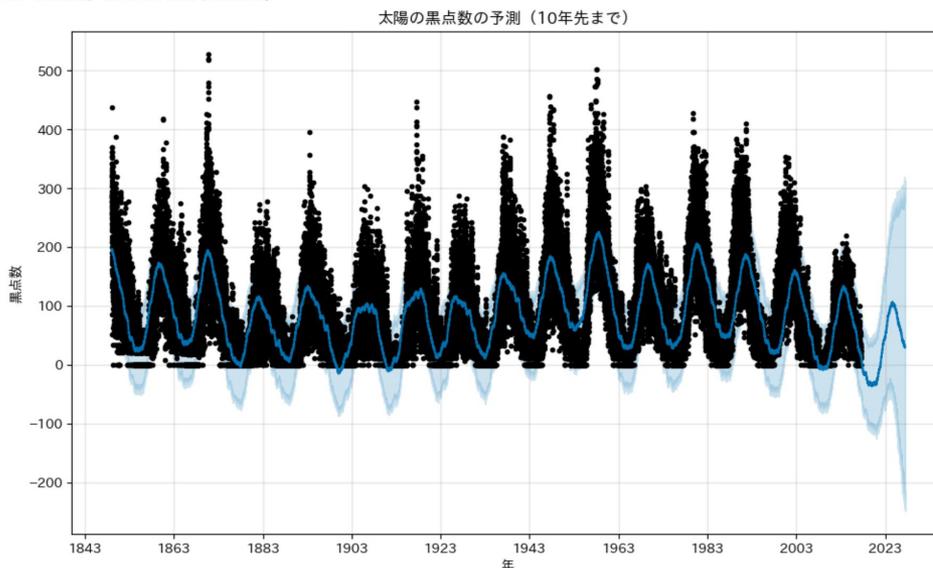
予測の例（太陽黒点データ） Prophet を使用

黒：元データ 水色：分析結果（短期的な変動を除去、将来予測）



- Prophet は Python のライブラリ
- **時系列データ**に対して、**周期性**、**トレンド**、**特定のイベント**や**時期との関連性**を分析する機能を持つ
- **機械学習**の一手法（ただし、ニューラルネットワークではない）

過去データの学習により，周期性，トレンドなどのパターンが抽出され，未来予測に適用



- **Prophet**
 - 機械学習のために、統計技術を使用（学習の結果を確認しやすい）
 - 複雑なパターンでの予測が苦手
- **ニューラルネットワークの一種：リカレントニューラルネットワーク**
 - 複雑なパターンでの学習，予測が可能
 - 学習のために大量のデータを必要とする

リカレントニューラルネットワーク (RNN)



仕組み

- **回帰構造を持つニューラルネットワーク（出力の一部を次の入力に使用）**
- 過去の情報を考慮した処理が可能

特徴

- **時系列データ**の扱いに適している
- **複雑なパターン**の学習が可能．データの長さに制約はない
- LSTM等の**改良版**では長期記憶が可能（**長期的なパターン**の学習が可能）

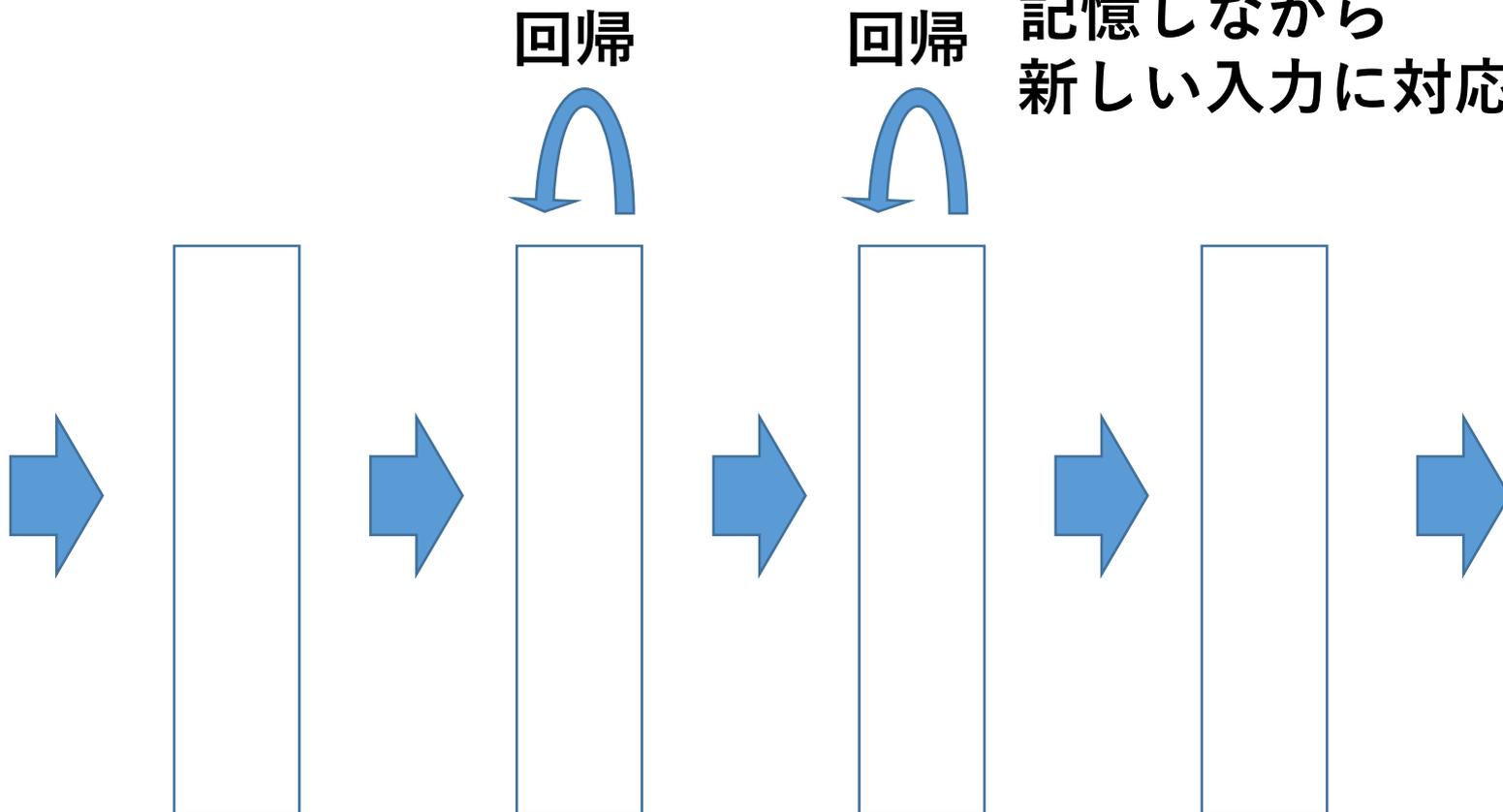
主な応用分野

- 時系列データの将来予測

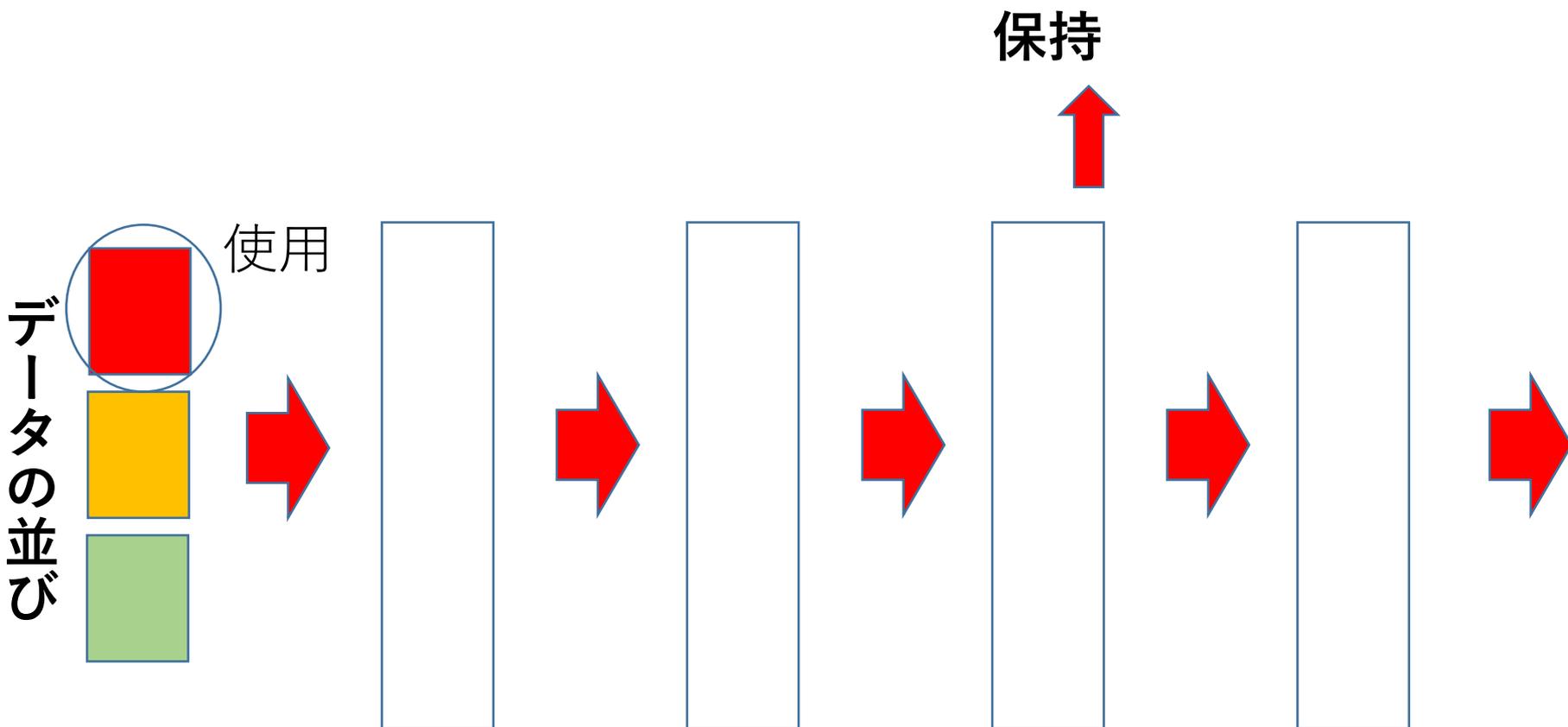
リカレントニューラルネットワーク

回帰のしくみ

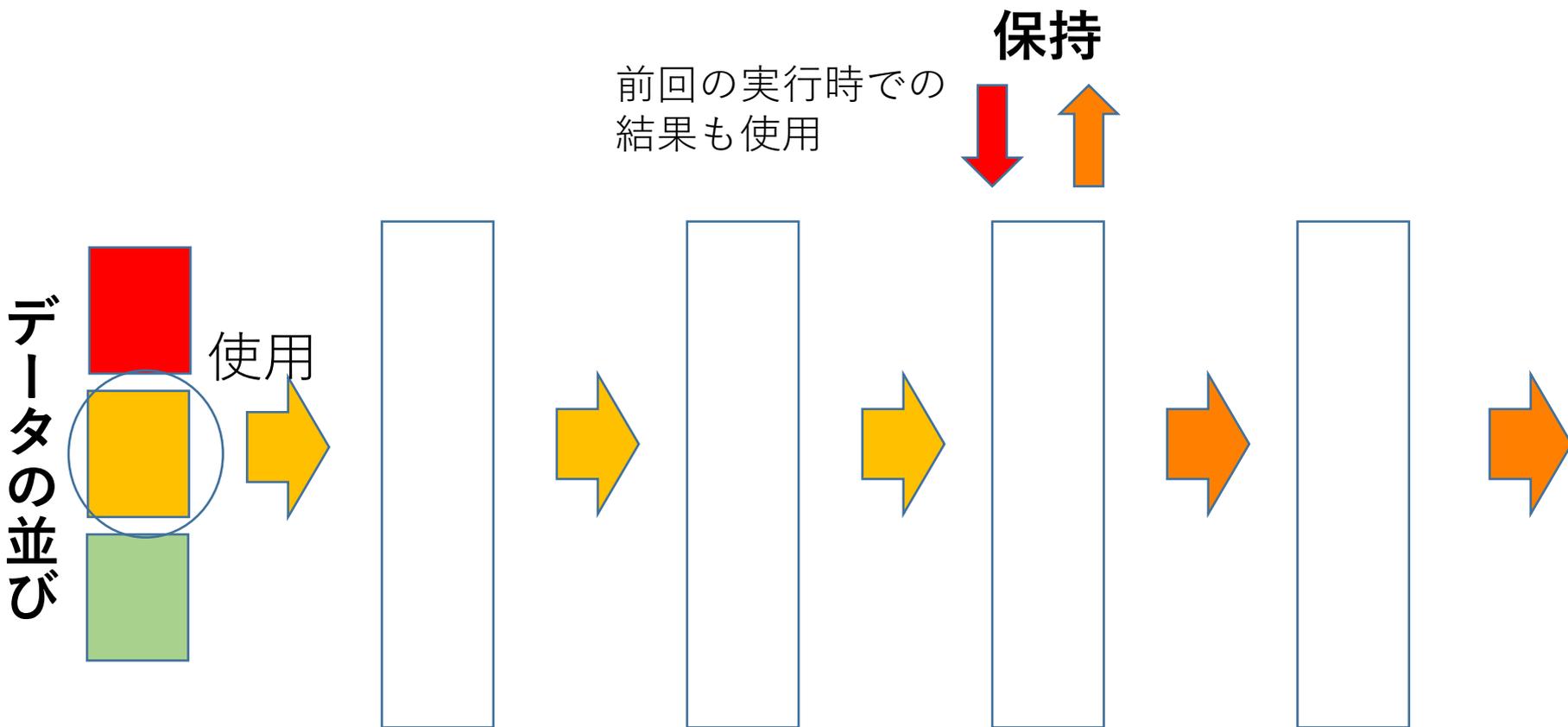
前回の実行時での結果を
記憶しながら
新しい入力に対応



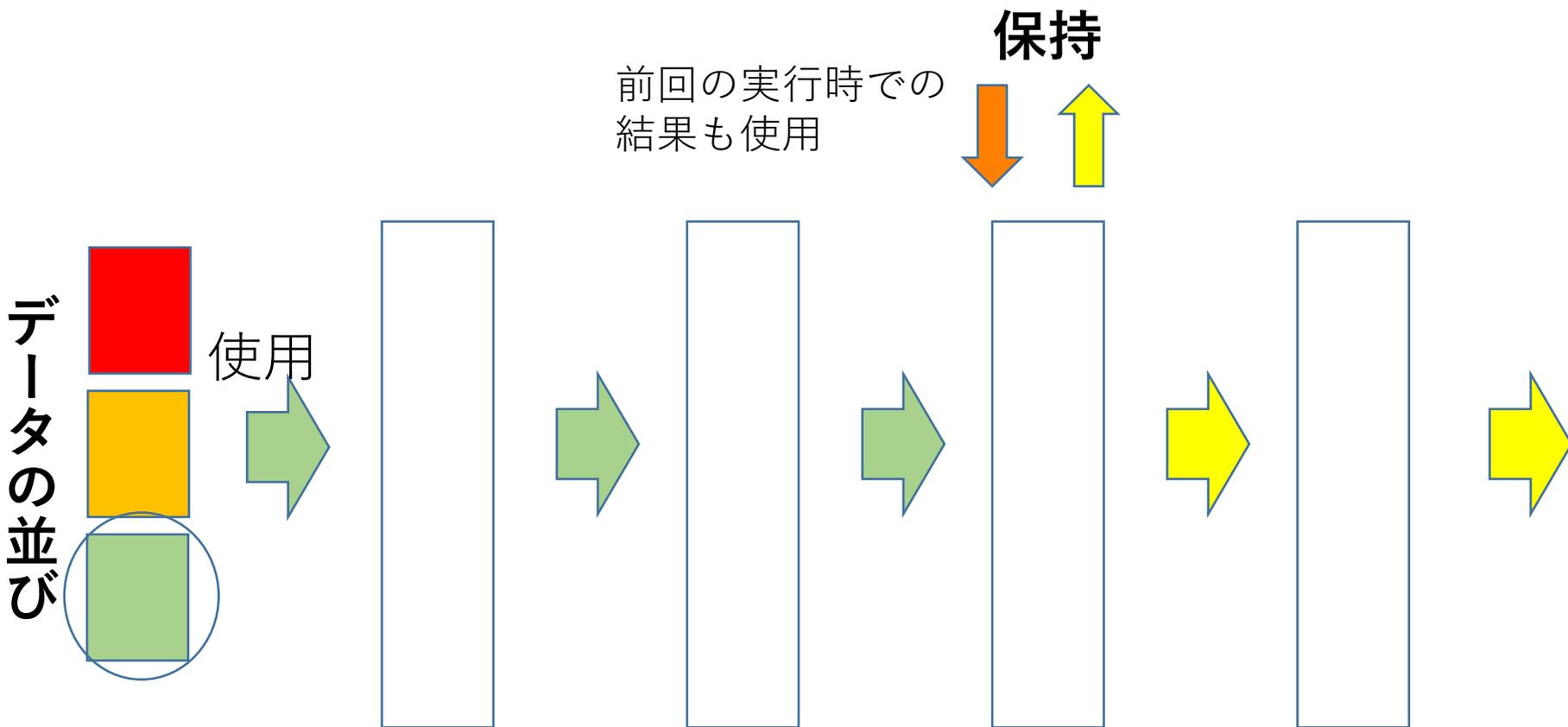
リカレントニューラルネットワークの動作イメージ ①



リカレントニューラルネットワークの動作イメージ ②



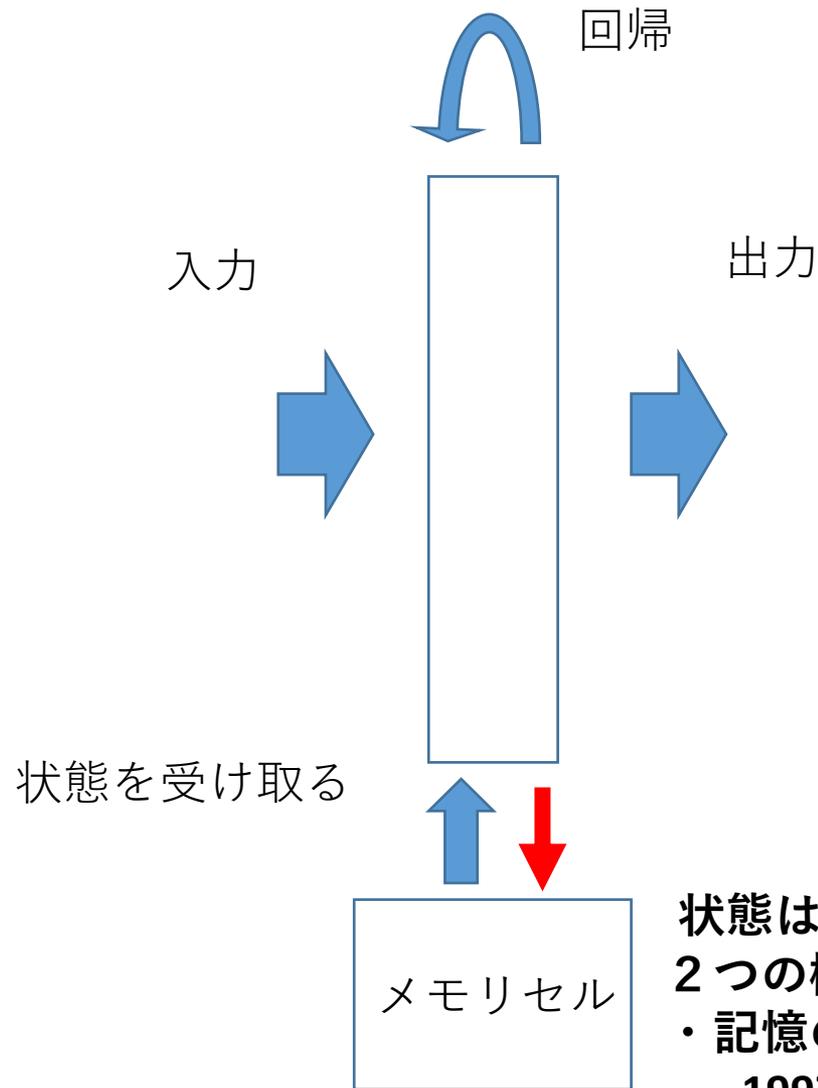
リカレントニューラルネットワークの動作イメージ ③



LSTM の特質

- リカレントニューラルネットワーク（RNN）の改良版として 1997年に提案された。
- RNNの弱点：「**長期的な依存関係の学習が困難**」
を克服
- LSTM は、**メモリセルにより長期の記憶を保持し、**
複雑なデータを扱うことができる。

LSTM の仕組み

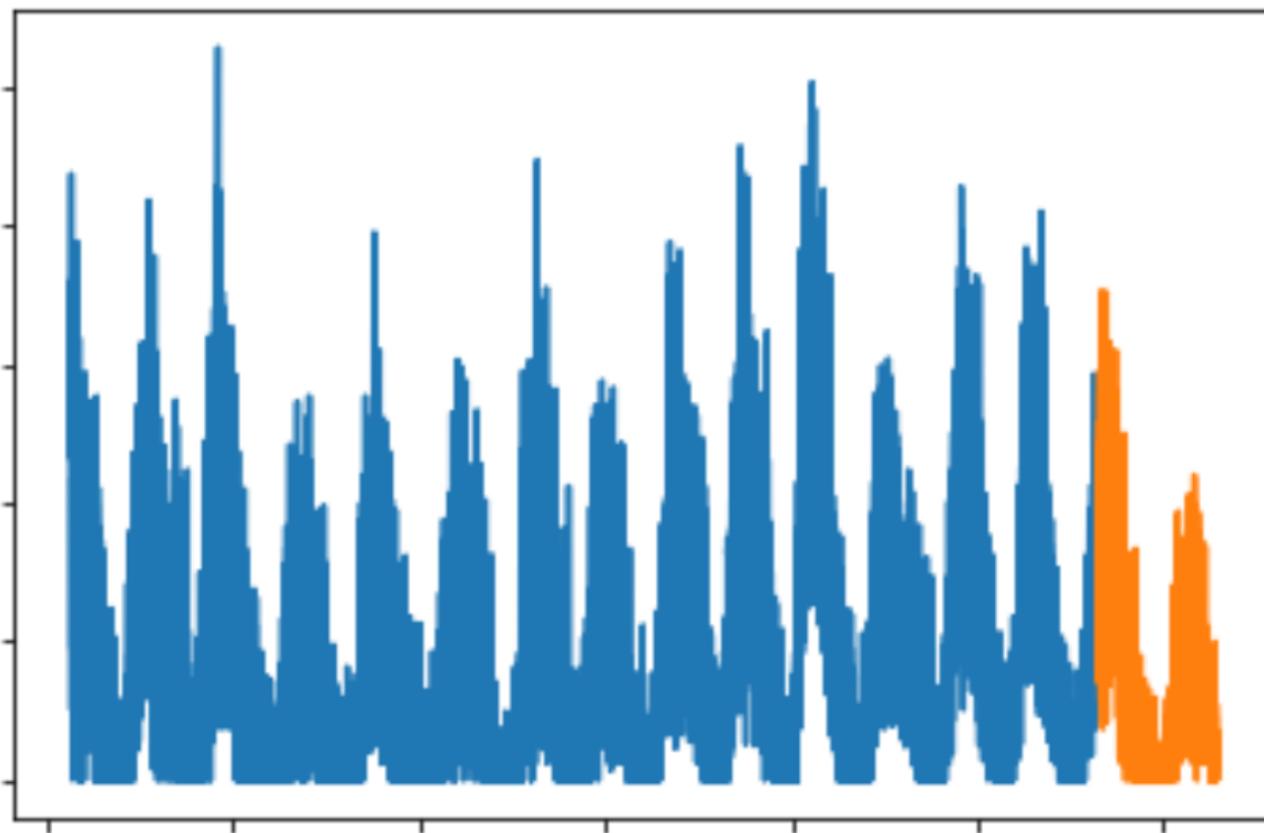


LSTM のメモリセルは、長期の記憶の保持を可能とする

状態は「メモリセル」に記憶されている
2つの機能

- ・ 記憶の持続 (constant error carousel)
1997年発表
- ・ 記憶の忘却 (forget gate)
1999年発表

LSTM を用いた太陽の黒点数の変化予測



予測結果は、複雑な
パターンを再現

太陽の黒点数の変化

予測

1848年～1999年のデータを用いて、**2000年以降を予測**
(ディープニューラルネットワークによる予測)

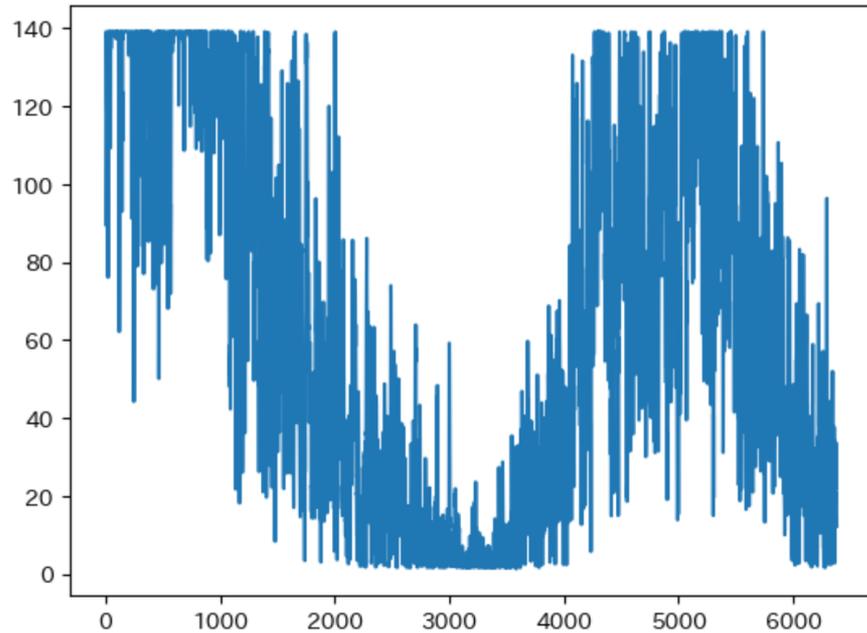
発展演習 2 : Prophet, LSTMを用いた太陽の黒点数の分析, 将来予測

- Google Colaboratory のページ

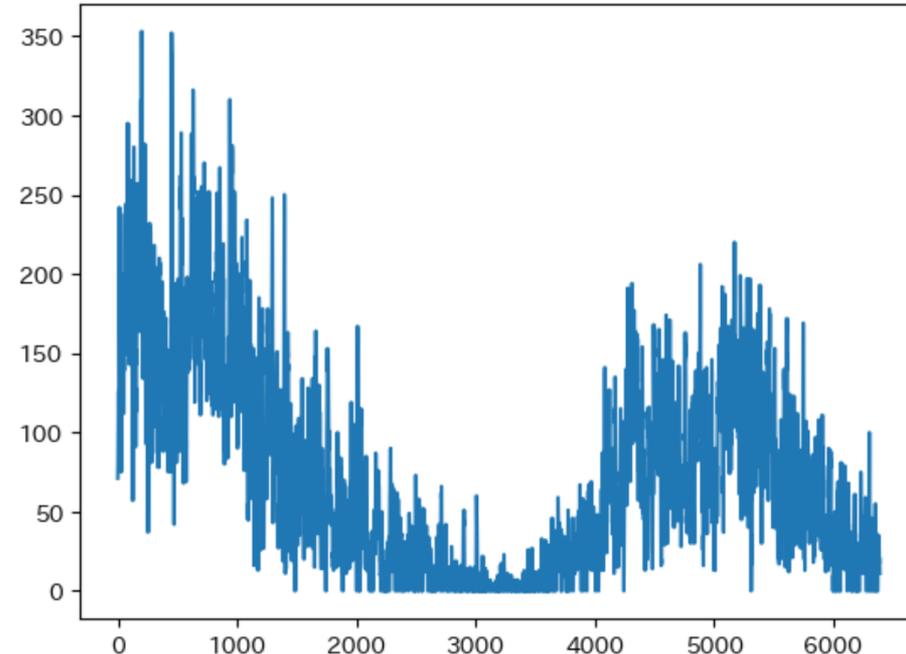
https://colab.research.google.com/drive/1qhx5l0iEPUm-QRTuEBSqBLd3V9KvqltK?usp=drive_link

- プログラム, 説明, 実行結果を記載

1848年から1999年までの黒点数をもとに AI が予測した 2000年以降の黒点数



2000年以降の黒点数 (実測値)



時系列データの分析と予測まとめ

- **時系列データ**：時間の経過に伴って順序付けられたデータの並びである
- **LSTM (Long Short-Term Memory)**：リカレントニューラルネットワークの改良版。メモリセルにより長期の記憶を保持し、複雑なパターンの学習が可能である。
- **リカレントニューラルネットワーク**：時系列データを扱うための特殊なニューラルネットワーク構造である。出力の一部を次の入力として使用する回帰構造を持ち、過去の情報を考慮した処理が可能である。
- **Prophet**：Pythonのライブラリ。周期性、トレンド、特定のイベントや時期との関連性を分析する機能を持つ。

11-4. Transformerを基盤とする現代の大規模言語モデル

自然言語処理と人工知能

自然言語処理は、人間が普段使う言語（日本語、英語など）を**コンピュータが理解**したり、**生成**したりする技術

- AI による**人間の言語の理解**
- AI と人間の**自然な対話**
- **言語間の翻訳**
- **文書類**（人間の言葉によるデータ）の**分析と活用**

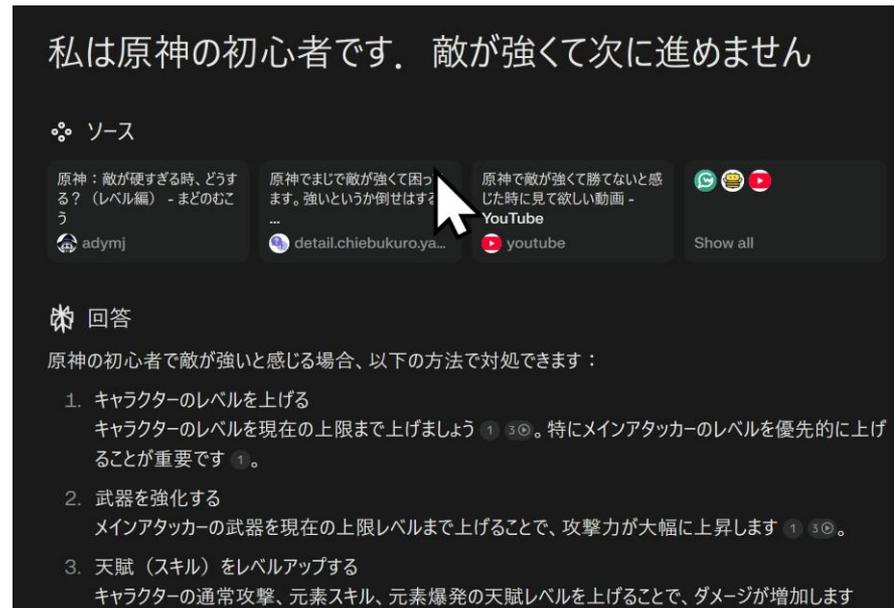
① Perplexity AI (検索型AI)

- 自然な会話で人間の質問を理解
- 最新の情報を文章で回答.
- 情報源を明示し, 画像を含む質問にも対応可能である.
- 対話を通じて必要な情報を段階的に絞り込むことができる.

1. まず Perplexity AI のページにアクセス

<https://www.perplexity.ai/>

2. ログインの必要はない



② ChatGPT (チャットボット)

1. まず ChatGPT のページにアクセス

ChatGPT

<https://chatgpt.com/>

2. ログインの必要はない

※ログインは必須ではありませんが、**Googleアカウントによるログイン**を行うことで、履歴の保存や個人設定などが可能になります。

チャットボットのベストプラクティス



実践ポイント

プロンプト：

- 明確で具体的な質問・要求
- 追加要求による回答の改善

データや例の活用：

- 関連データの提供
- 回答として欲しい内容の例示

回答の確認：

- 正確性の検証
- 根拠の確認

Transformerとは？



- 2017年にGoogleが提案した**自然言語処理モデル**（論文：
"Attention is All You Need"）
- 行列演算と並列処理で、**文章全体を一度に計算**（RNNや
LSTMの順次処理に比べ数十倍の処理速度向上）

Transformerの主要技術：アテンション



- 効果：長文でも性能低下が少ない（従来モデルでの数百単語の制限を超えて処理可能）
- 与えられた文章（入力）の各単語について次の処理を行う
 - 各単語について3種類の情報を計算
 1. **その単語が何を表すか**（意味ベクトル）
 2. **その単語がどのような文脈で使われるか**（文脈ベクトル）
 3. **その単語が他の単語とどう組み合わせるか**（結合ベクトル）
 - これら3つの情報の積により、**単語間の関連の強さを計算**（例：「みかん」という単語の意味は、「りんご」との関連が強いか弱いかで変化）
 - **学習データ中の多様な文章から、この3つの情報の適切な値を学習**（語順、文法、慣用句などの言語の規則性を反映）

- 並列GPUでの行列演算により、**数千億単語以上のデータを学習**
- 学習のあとの追加的な学習（ファインチューニング）の2段階学習により、**様々なタスクに適応できる仕組み**（特定タスクへの最適化）
- 対話型AIに応用。**ChatGPT**（数千億パラメータ）など、**大規模言語モデルの基盤技術**として普及

- **Transformer** : 2017年にGoogle社が提案した自然言語処理モデル。行列演算と並列処理により**文章全体を一度に計算し、従来モデルと比べて処理速度を向上**させた技術。
- **アテンション機構** : Transformerの核となる技術で、**長文処理における性能低下を抑制**する。単語の意味、文脈、他語との関連性を計算し、言語の規則性を学習する仕組み。
- **大規模言語モデル** : 並列GPUでの行列演算により、**数千億単語以上のデータを学習した言語処理システム**。様々なタスクに対応可能な汎用的な言語理解・生成能力を持つ。



11-5. Vision Transformer (ViT) による画像処理

Vision Transformer

- 自然言語処理において高い性能を示す Transformer を画像に応用したもの
- 画像全体を小さな「パッチ」（例：16×16ピクセル）に分解
- 各「パッチ」を単語のように扱い、アテンション機構で処理
- AIは、パッチ間の関係性を学習することで、画像全体の理解を実現

セグメンテーションの種類



セマンティック・セグメンテーション

画像内のすべての画素にラベルを付ける。

「個々の物体を検出する」という概念はない



インスタンス・セグメンテーション

物体を検出する。同じ種類の複数の物体がある場合は、別々のものとして認識。各物体に対してセグメンテーションを実施。

「物体として識別できない部分は結果がない」ということもある



パノプティック・セグメンテーション

セマンティック・セグメンテーションとインスタンス・セグメンテーションの同時実行



演習 1 . Vision Transformer による画像セグメンテー ション

画像セグメンテーション

【トピックス】

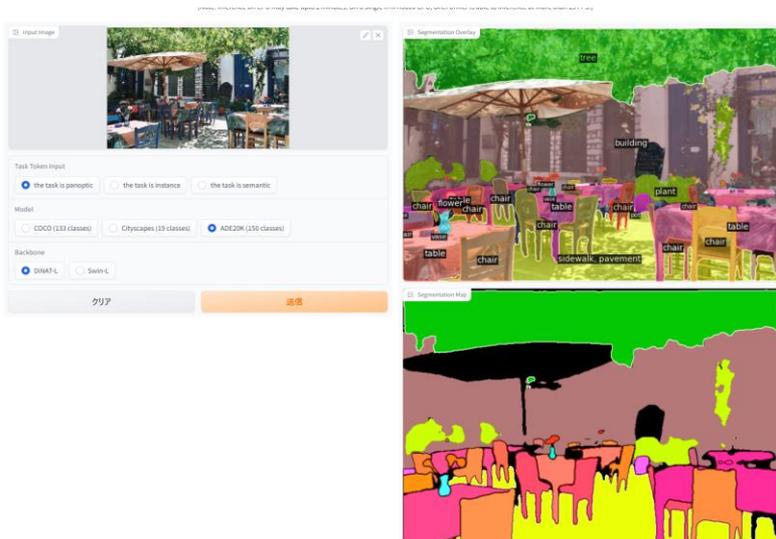
- 画像セグメンテーション
- OneFormer のデモ
- Vision Transformer

OneFormer

OneFormer は、**Vision Transformer の技術**をベースに、パノプティック、インスタンス、セマンティックの**セグメンテーション**を行うもの（**2022年発表**）

<https://huggingface.co/spaces/shi-labs/OneFormer>

- セグメンテーションの種類：パノプティック、インスタンス、セマンティック
- 訓練データ：COCO（133 クラス）、Cityscapes（19 クラス）、ADE20K（150クラス）



文献: Jitesh Jain, Jiachen Li, MangTik Chiu, Ali Hassani, Nikita Orlov, Humphrey Shi, OneFormer: One Transformer to Rule Universal Image Segmentation, arXiv:2211.06220, 2022.

① OneFormer のデモページを開く

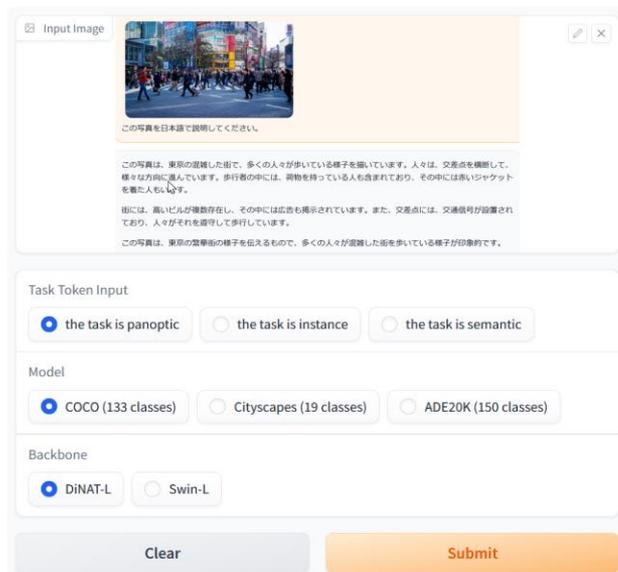


<https://huggingface.co/spaces/shi-labs/OneFormer>

オンラインサービスであり、混雑時などは動かない場合がある。授業中で動かなかった場合には、後日試してほしい

② 画面の「Input Image」で画像ファイルを設定するか、下の「Examples」で画像を選ぶ

③ 「Submit」 をクリック。結果を確認。



次のような結果が得られる



元画像



インスタンス・セグメンテーション



パノプティック・セグメンテーション

- OneFormer のデモサイトを使用
- URL: <https://huggingface.co/spaces/shi-labs/OneFormer>
- 訓練データは COCO
- バックボーンは DiNAT-L を使用



セマンティック・セグメンテーション

Vision Transformer (ViT) によるセグメンテーションまとめ



- **Vision Transformer**は、自然言語処理で実績のある**Transformerを画像処理に応用したモデル**。画像を小さなパッチに分割し、アテンション機構を用いてパッチ間の関係性を学習することで、画像全体の効果的な理解を実現する。
- **OneFormer** : Vision Transformer の応用。パノプティック、インスタンス、セマンティックの**3種類のセグメンテーションを統合的に実行**できるモデルである。



11-6. Vision Transformer による NoShot 学習

NoShot の考え方

Transformerや**Vision Transformer**は**自然言語処理**や**画像処理**において**汎用性が高い**

例：SAM (Segment Anything Model)

- セグメンテーションの手法
- Vision Transformerをベースにしている
- **学習時の訓練データに無かったラベル等**に対しても回答できる NoShot 学習の能力を持つ（汎用性を持つ）
- 画像中のポイントを指定してのセグメンテーションなどの応用

従来の画像認識

学習時に使用されたラベル（例：人間、自動車、信号機）の
範囲内でのみ識別・分類が可能

NoShot での画像認識

- 汎用性
- **学習時のラベルにはない新しいラベルに対しても、識別・分類が可能**

NoShot のセグメンテーションの例



Text Prompt

hard object

画像と英語のプロンプトを
AI に与える。

プロンプトは自由

汎用性

- **学習時のラベルにはない新しいラベルに対しても、識別・分類が可能**

多様な応用

- 学習済みのモデルは、さまざまな種類のものに対応する汎用性を持つ。

演習 2 SAM

NoShot の画像セグメンテーション



① Segment Anythingデモサイトを開く

<https://segment-anything.com/demo>

最初の画面では、**チェック** 次の画面で「**Try Demo**」

②基本的な使い方 次の2通り

- ・ **既存の画像の選択（画面で選択）**
- ・ **自分のコンピュータから画像をアップロードする。（画面の「Upload an image」をクリック**

③操作方法

- ・ **画像上でクリックすることで、物体の自動セグメンテーション（切り抜き）**が開始される。
- ・ 以下の3つの操作モードを切り替えて使用する：
 - ・ **点による指定**：画像上の物体をクリックして選択（Hover & Click）
 - ・ **ボックスによる指定**：物体を四角形で囲んで選択（Box）
 - ・ **全物体の自動検出**：画像内のすべての物体を自動検出（Everything）



- **NoShot学習**：事前学習済みモデルが，学習時には見たことのない新しいタスクや概念に対しても対応できる能力のことである。
- **SAM (Segment Anything Model)**：Vision Transformerをベースとした高度なセグメンテーションモデルである。**学習時に使用していないラベルに対しても柔軟に対応**できる。画像中の任意の点を指定することで直感的な操作が可能となる（例：クリックによる対象物の選択）。