千葉菌類談話会・スライド会　2024/1/6

# ChatGPTを利用した菌類の形質ベクトル特徴量のハイスループット生成：同定と菌類相調査への応用

中島淳志
Atsushi Nakajima

# 今日のお話のごく簡単な説明

## 菌類の名前

## 例：ベニテングタケ 

## を

## ↓この形で表現できるとうれしい↓

[ 0.01485809, -0.00965576, -0.00472758, ..., 0.00439323, -0.0176933 , -0.04111052]

# ベクトルには「形質データ」が詰まってる

[ 0.01485809, -0.00965576, -0.00472758, ..., 0.00439323, -0.0176933 , -0.04111052]


傘や柄のサイズ、色、形状、表面性状などの
**形態形質データ**

宿主・基質、栄養摂取様式、種間相互作用などの
**生態形質データ**

至適温度・pH、薬剤耐性、二次代謝産物などの
**生理形質データ**

# ベクトル、何がうれしいのか？？

## 機械可読

人間にはさっぱりわからないが

## 分類・予測

ディープラーニングの主目的

## 検索

ベクトル間の類似度から
迅速に候補を特定できる

## 可視化

3次元以下に削減することで
データの関係性が一目で分かる

# DNA塩基配列データとの比較

例えば、🧬塩基配列データは

- コンピュータにとって扱いやすい
- データ形式が標準化されている
- 数値化や自動化、拡張が容易

# 記載文は未開拓の宝の山

記載文の自然文データは塩基配列と違い

- **コンピュータにとって扱いにくい**

↓

しかし、

- **200年来の大量のデータが蓄積されている**
- **DNA以外の多様で重要な分類形質を含む**

# 記載文構造化とは？

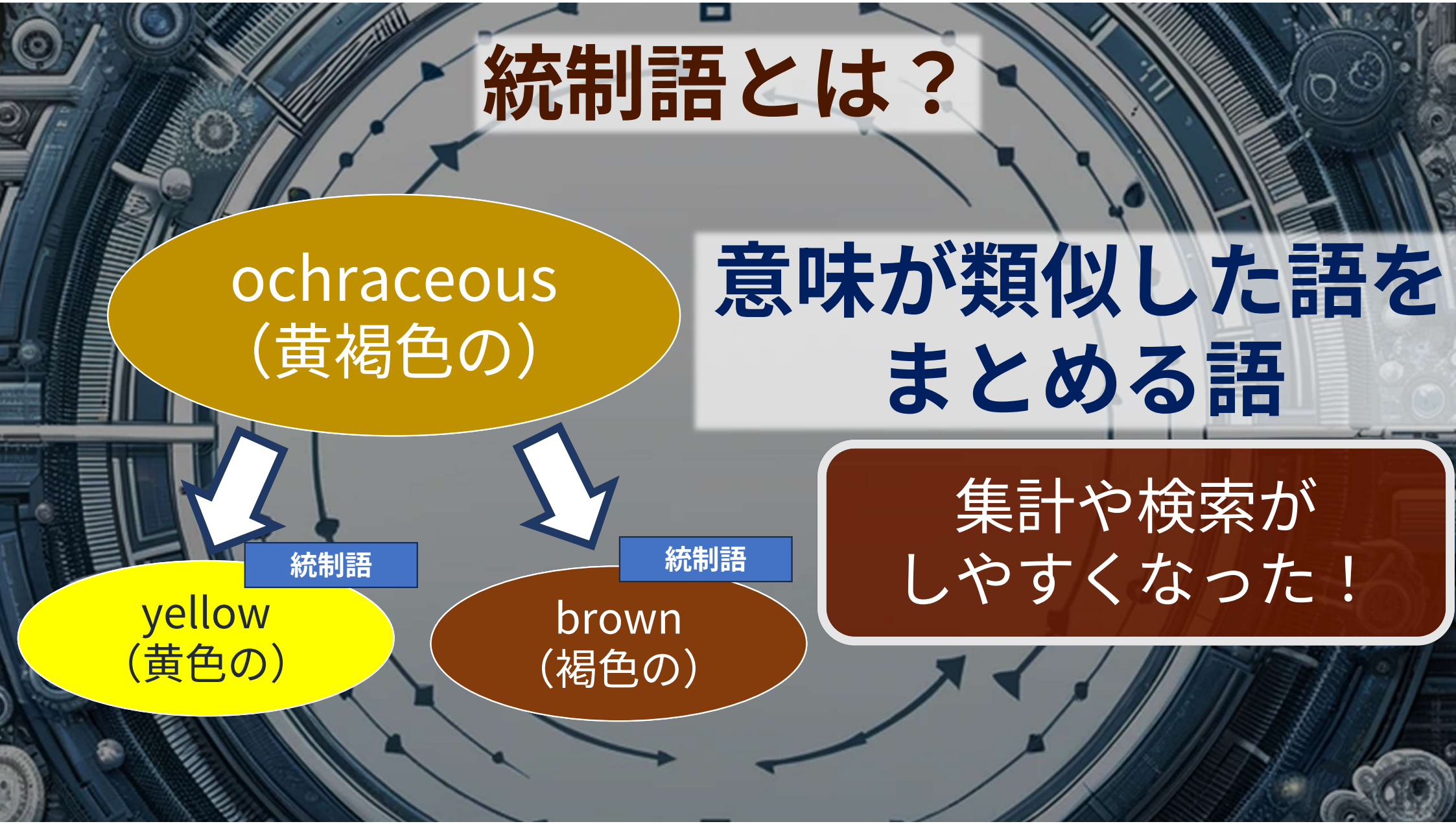
*Pileus* 20–35 mm wide, at first (sub)conical, later broadly convex or expanded, with or without low, large umbo, margin slightly decurved, later straight or even uplifted, and then pileus depressed around the umbo; young basidiomata with faint and fugacious remnants of a pale greyish velipellis; usually somewhat bicoloured, pale ochraceous brownish to pale nut-brown at the cen-

| | trait | source | scientificname |
|---|---|---|---|
| 39 | pileus_color_brown | 10.18476/2023.787646 | Inocybe centesima |
| 40 | pileus_color_dark | 10.18476/2023.787646 | Inocybe centesima |
| 41 | pileus_color_pale | 10.18476/2023.787646 | Inocybe centesima |
| 42 | pileus_color_purple to blue | 10.18476/2023.787646 | Inocybe centesima |
| 43 | pileus_color_yellow | 10.18476/2023.787646 | Inocybe centesima |
| 44 | pileus_presence_present | 10.18476/2023.787646 | Inocybe centesima |
| 45 | pileus_shape_conical | 10.18476/2023.787646 | Inocybe centesima |
| 46 | pileus_shape_convex | 10.18476/2023.787646 | Inocybe centesima |

## 文章からその種の形質を抽出してまとめる作業

少し機械にとって読みやすくなる！

Bandini, D. et al. 2023. https://doi.org/10.18476/202378764


統制語とは？
ochraceous
（黄褐色の）
意味が類似した語を
まとめる語
統制語
yellow
（黄色の）
統制語
brown
（褐色の）
集計や検索が
しやすくなった！

# 従来のいまいち😕な方法

十分実用に堪える結果が得られた。具体的には、「Natural Language Toolkit (NLTK)」の正規表現パーサー (nltk.RegexpParser) を利用して「要素」と「値」が対になるようにフレーズチャンクを抽出している。例えば、要素の抽出には「{^<JJ>*<NN|NNS|NNP|NNPS|CC|TO|\$>+<IN|NN|NNP|NNS|CD|JJ>*}」、値の抽出には「{<CC|DT|PDT|RBR|PRP\$|RB|TO|MD|WRB|JJ|JJS|JJR|IN|RB|RBS|VB|VBP|VBG|VBN|VBD|\(|\)|NN|NNS|NNP|NNPS>+}」という正規表現パターンを用いている。記載文特有の工夫としては、事前に要素の一覧を

複雑なパターンを指定…
手作業での修正も多かった

中島淳志. 2020. https://media.niche-life.com/series/008/Niche008_06.pdf

# ChatGPTの登場

## OpenAI社が2022年11月に公開
## 「大規模言語モデル (LLM)」を基盤とする
## チャットボット

GPT-3.5　GPT-4

## あらゆる話題に自然に回答
## ⇒一躍世界的なトレンドに！

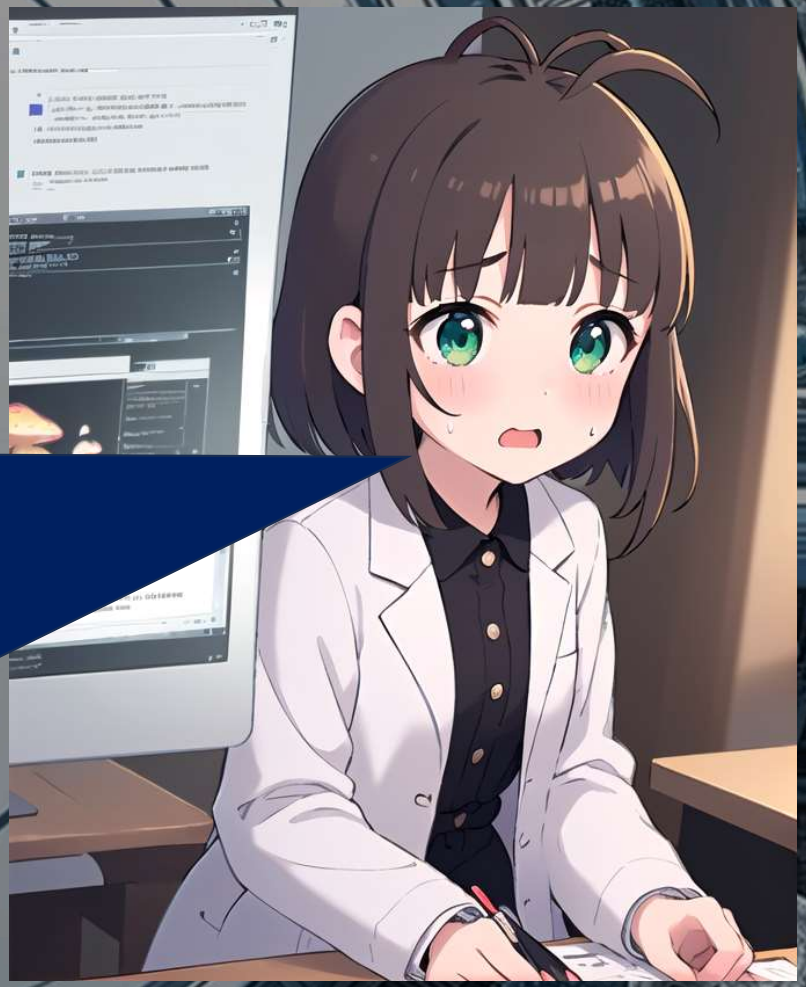
# 指示文（プロンプト）

この記載文を
JSON形式で
構造化してください♪

…のような
単純な指示でも
ある程度機能するが…

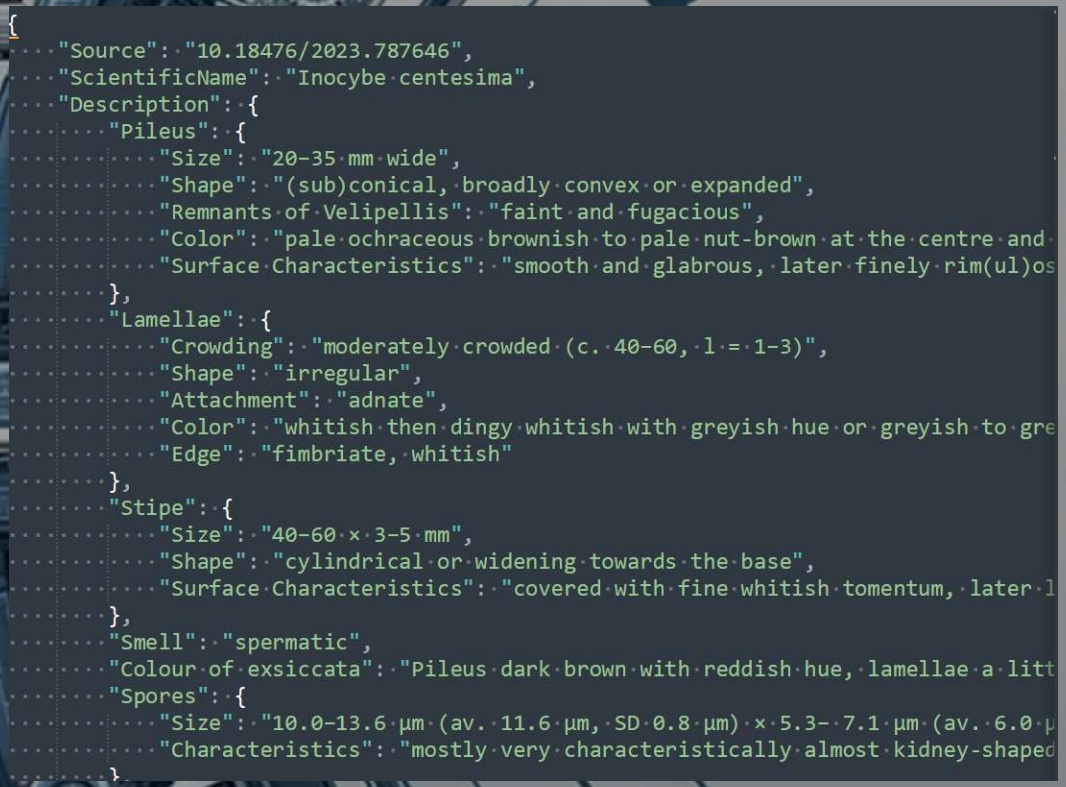
# 実際は「プロンプトエンジニアリング」が必要

'''As an experienced mycologist, your task is to meticulously parse descriptions of various fungi, ensuring that every feature of the original text is accurately transcribed. The end goal is to restructure these descriptions into a tightly organized JSON format, using as much identical terminology from the original text as possible. Note that the provided text may contain Optical Character Recognition (OCR) errors that must be corrected during your transcription process. The JSON output should always include both the cited source and the scientific name of the fungi described.While structuring your JSON, bear in mind that data should be organized in triplets composed of "elements", "attributes", and "values". The "elements" refer to aspects of the fungi, like "pileus", "stipe", "basidiospore", and so on. The "values" depict specific traits of these elements, such as "yellow", "cylindrical", or "3-5µm". The value should not be in the form "true" or "false," for example, "Structure": "septate" instead of "Separation": true. Finally, "attributes" should be selected, as much as possible, from this list: "amount", "amyloidity", "color", "development", "habitat", "position", "presence", "shape", "size", "structure", "surface", and "taste/odor". However, if none of the above apply, an arbitrary "attribute" can be assigned. Ensure that all components are well represented to maintain the accuracy of the database.''' + "#Example:" + '''{ "Source": "10.5586/asbp.2004.010", "ScientificName": "Ascocoryne turficola", "Description": { "Apothecia": { "Size": "0.5-4.0 cm", "Shape": "ceraceous", "Surface Characteristics": "turbinate with convex disc when young, then gelatinous, cup-shaped with central depression", "Color": "olivaceous, olivaceous brown to brownish-lilac" }, "Hymenium": { "Surface Characteristics": "smooth, then gibbous to cerebriform", "Color": "olivaceous" }, "Stalk": { "Surface Characteristics": "wrinkled", "Color": "hyaline, vinous-pink to lilac-pink", "Size": "3.0-8.0 cm long" }, "Position": "Solitary, rarely gregarious to cespitose", "Asci": { "Shape": "cylindric-clavate", "Structure": "amyloid pore", "Size": "up to 180.0 x 10.0 um", "Amount": "8-spored" }, "Ascospores": { "Shape": "ellipsoid", "Color": "hyaline", "Surface Characteristics": "with drops", "Size": "14.0-19.6 x 5.6-7.0 um", "Structure": "non-septate" }, "Paraphyses": { "Shape": "slightly blunt" } } }''' + "#Input: " + f"#Source: {source}#ScientificName: {scientific_name}#Description to reformat: {description}



# （一度作れば使い回せるけど）

# JSON形式の出力を得る



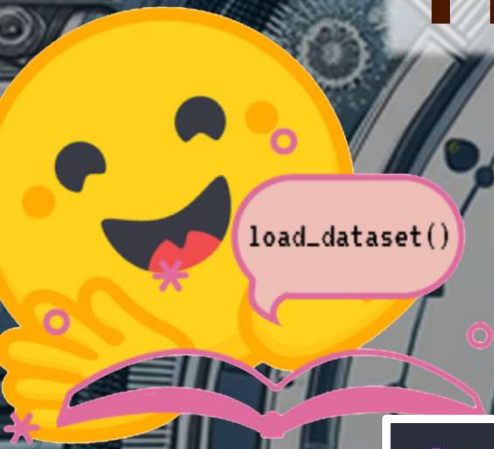
```
{
  "Source": "10.18476/2023.787646",
  "ScientificName": "Inocybe centesima",
  "Description": {
    "Pileus": {
      "Size": "20–35 mm wide",
      "Shape": "(sub)conical, broadly convex or expanded",
      "Remnants of Velipellis": "faint and fugacious",
      "Color": "pale ochraceous brownish to pale nut-brown at the centre and
      "Surface Characteristics": "smooth and glabrous, later finely rim(ul)os
    },
    "Lamellae": {
      "Crowding": "moderately crowded (c. 40–60, l = 1–3)",
      "Shape": "irregular",
      "Attachment": "adnate",
      "Color": "whitish then dingy whitish with greyish hue or greyish to gre
      "Edge": "fimbriate, whitish"
    },
    "Stipe": {
      "Size": "40–60 × 3–5 mm",
      "Shape": "cylindrical or widening towards the base",
      "Surface Characteristics": "covered with fine whitish tomentum, later l
    },
    "Smell": "spermatic",
    "Colour of exsiccata": "Pileus dark brown with reddish hue, lamellae a litt
    "Spores": {
      "Size": "10.0–13.6 µm (av. 11.6 µm, SD 0.8 µm) × 5.3–7.1 µm (av. 6.0 µ
      "Characteristics": "mostly very characteristically almost kidney-shaped
    }
```

| | trait | source | scientificname |
|---|---|---|---|
| 39 | pileus_color_brown | 10.18476/2023.787646 | Inocybe centesima |
| 40 | pileus_color_dark | 10.18476/2023.787646 | Inocybe centesima |
| 41 | pileus_color_pale | 10.18476/2023.787646 | Inocybe centesima |
| 42 | pileus_color_purple to blue | 10.18476/2023.787646 | Inocybe centesima |
| 43 | pileus_color_yellow | 10.18476/2023.787646 | Inocybe centesima |
| 44 | pileus_presence_present | 10.18476/2023.787646 | Inocybe centesima |
| 45 | pileus_shape_conical | 10.18476/2023.787646 | Inocybe centesima |
| 46 | pileus_shape_convex | 10.18476/2023.787646 | Inocybe centesima |

# ChatGPTの出力

# 統制語やルールベース処理適用

# TraitCircusデータセット




## Hugging Face Datasets🤗というサービスに投稿済み
## たった3行のコードで取得可能

```
from datasets import load_dataset
import pandas as pd
df = load_dataset("Atsushi/fungi_trait_circus_database")["train"].to_pandas()
```

2024年1月時点での形質データ数：

# 1,956,446

# ここからいよいよベクトル化

| | trait |
|---|---|
| 39 | pileus_color_brown |
| 40 | pileus_color_dark |
| 41 | pileus_color_pale |
| 42 | pileus_color_purple to blue |
| 43 | pileus_color_yellow |

'Abortiporus biennis: basidia amount 4-spored. basidia presence present, present or present. basidia shape clavate, clavate or clavate. basidia structure clamped. basidiospore amyloidity amyloid. basidiospore amyloidity no iodine reaction, reaction or reaction or reaction. basidiospore color hyaline, hyaline or hyaline or hyaline. basidiospore color yellow. basidiospore presence present, present or present or present or present. basidiospore shape ellipsoid, ellipsoid or ellipsoid or ellipsoid or ellipsoid or ellipsoid. …（中略）… tube color concolorous. tube presence present. '

41,558種を
約100円で処理

OpenAI
Embeddings

1536次元

[ 0.01485809, -0.00965576, -0.00472758, ...,
0.00439323, -0.0176933 , -0.04111052]

# 形質特徴量ベクトルの活用①

## ベクトルを使うことで
## 種間の類似度を算出できる

近似最近傍探索ライブラリ

**Faiss**

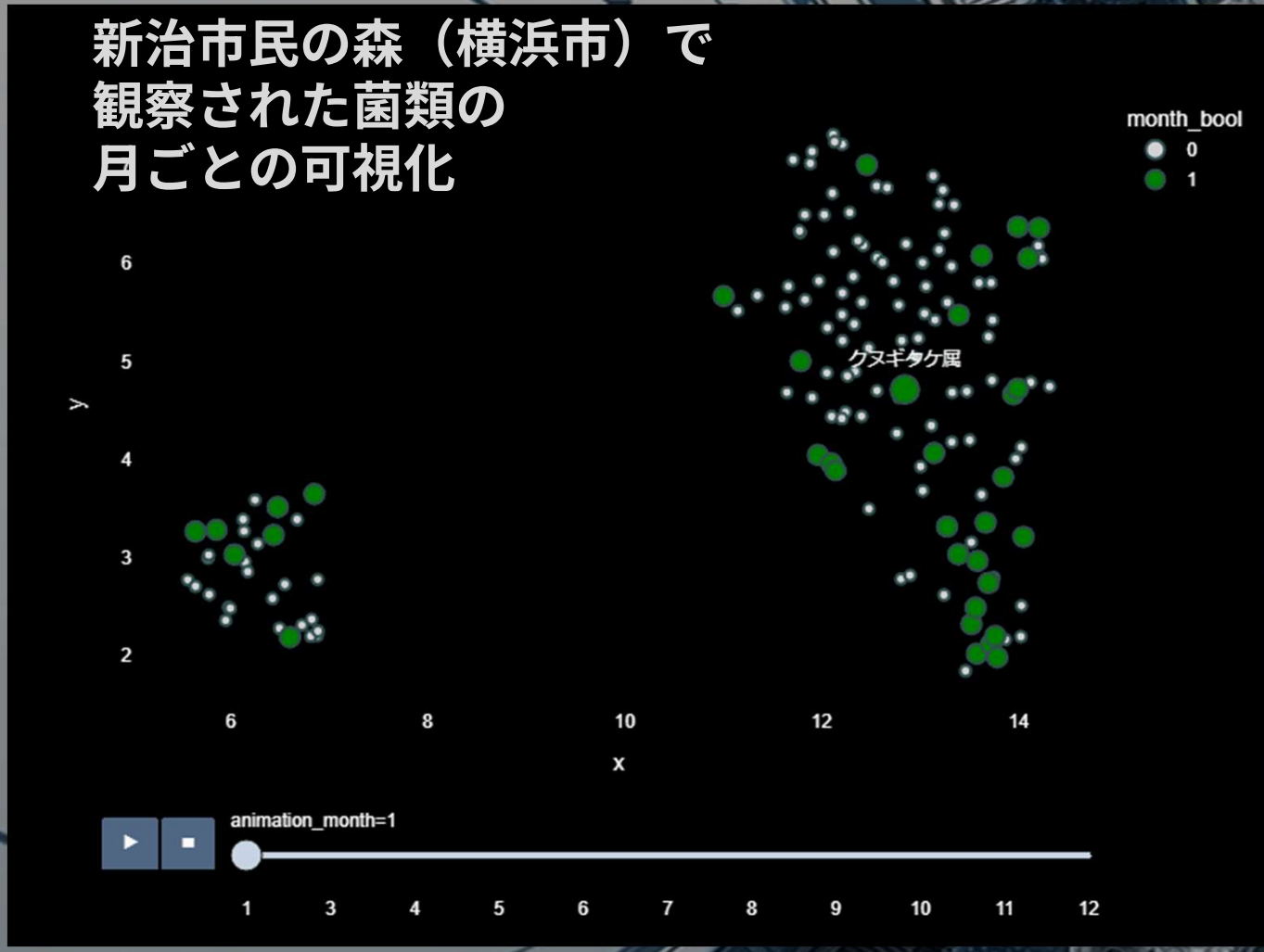
大量のベクトルデータから
一瞬で似たものを検索

【ウシグソコナヒトヨタケに近い種】

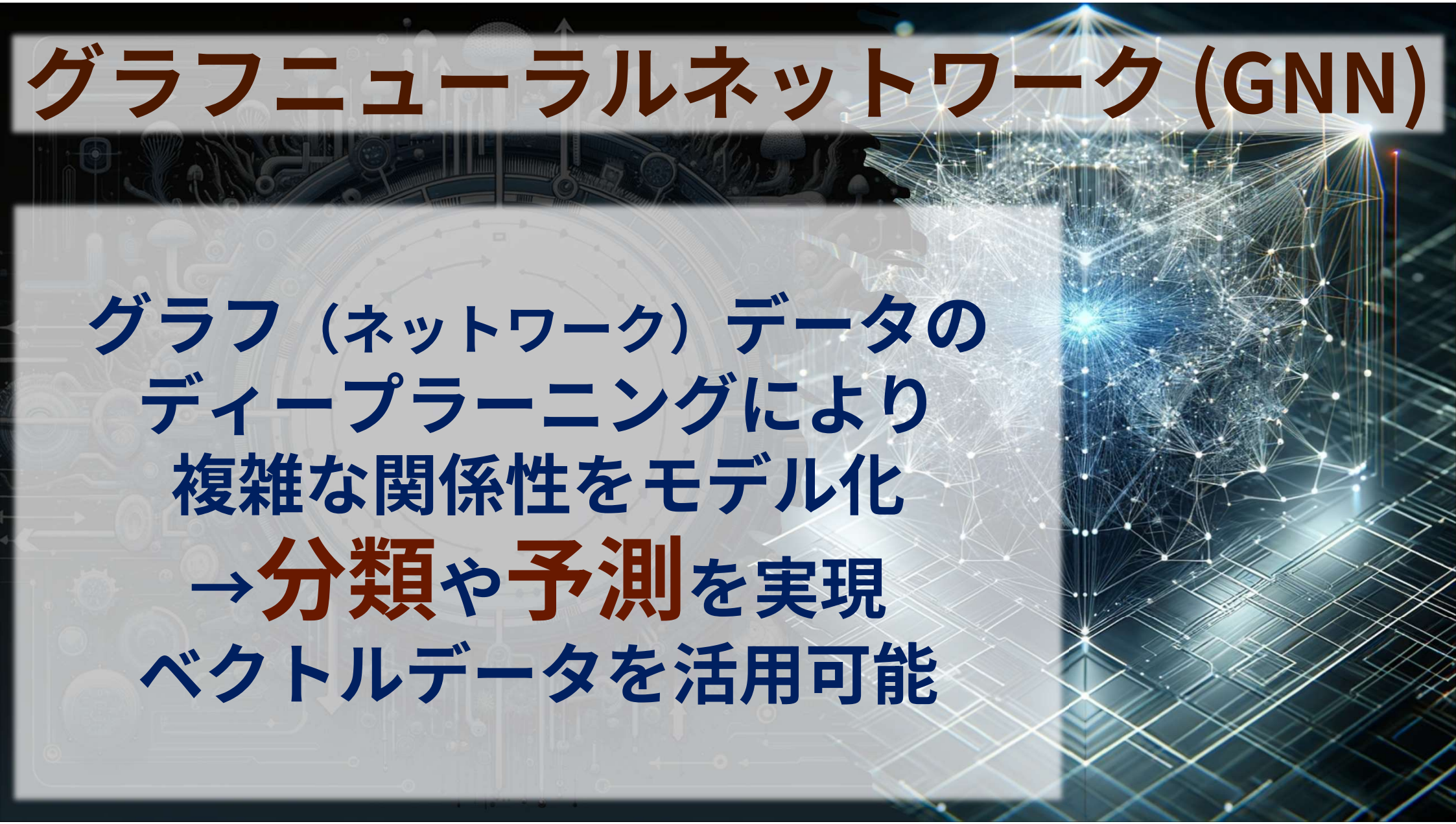
| | sci_name | distance |
|---|---|---|
| 1 | Coprinopsis indicifoetidella | 0.067199 |
| 2 | Coprinopsis geesterani | 0.069420 |
| 3 | Coprinopsis pseudomarcescibilis | 0.069422 |
| 4 | Coprinopsis ochraceolanata | 0.069539 |
| 6 | Coprinopsis arachnoidea | 0.070966 |

# 形質特徴量ベクトルの活用②

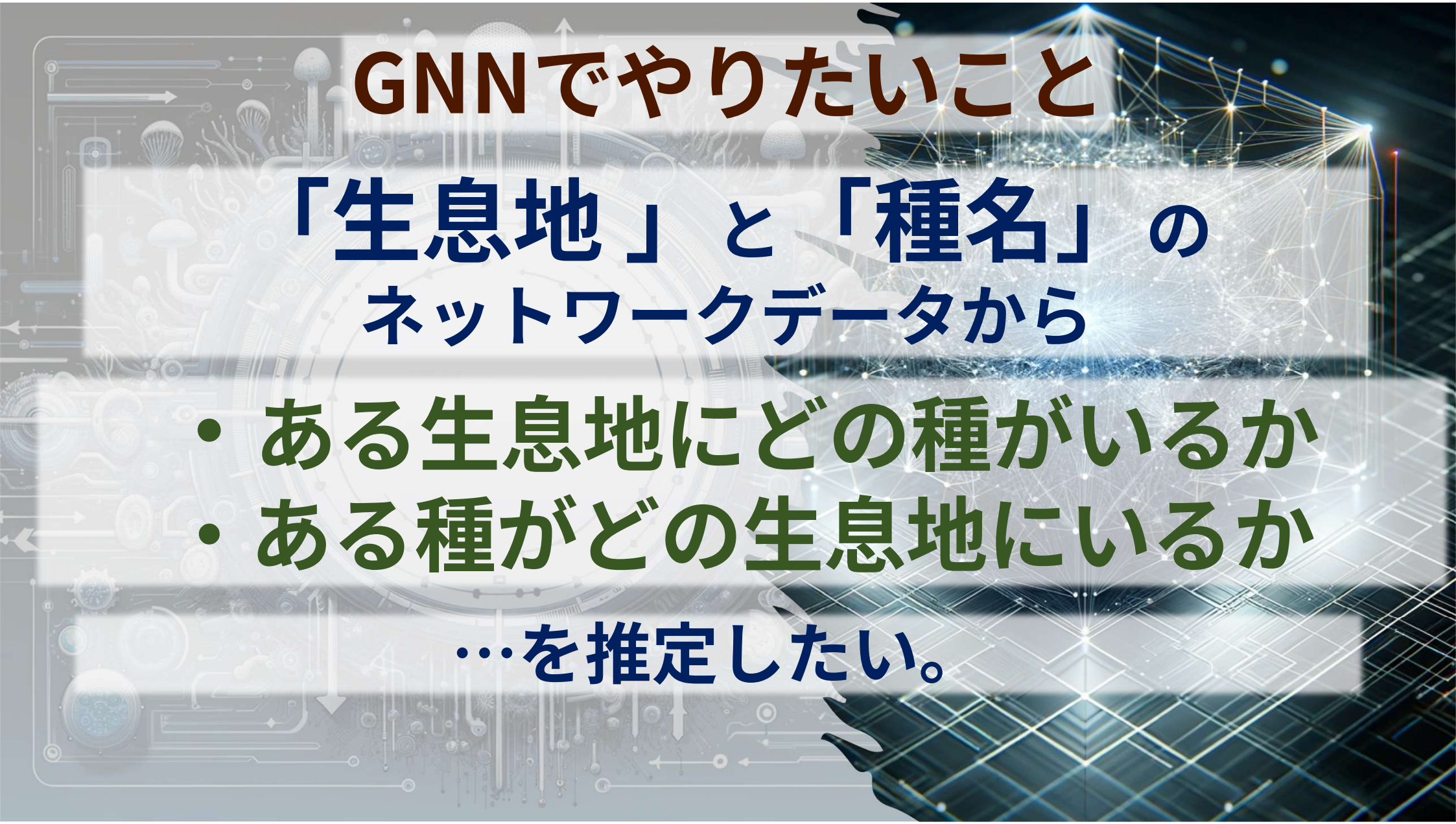
次元削減手法
との併用で
種間の距離を
可視化できる



# グラフニューラルネットワーク (GNN)



グラフ（ネットワーク）データの
ディープラーニングにより
複雑な関係性をモデル化
→**分類**や**予測**を実現
ベクトルデータを活用可能



# GNNでやりたいこと

## 「生息地」と「種名」の
## ネットワークデータから

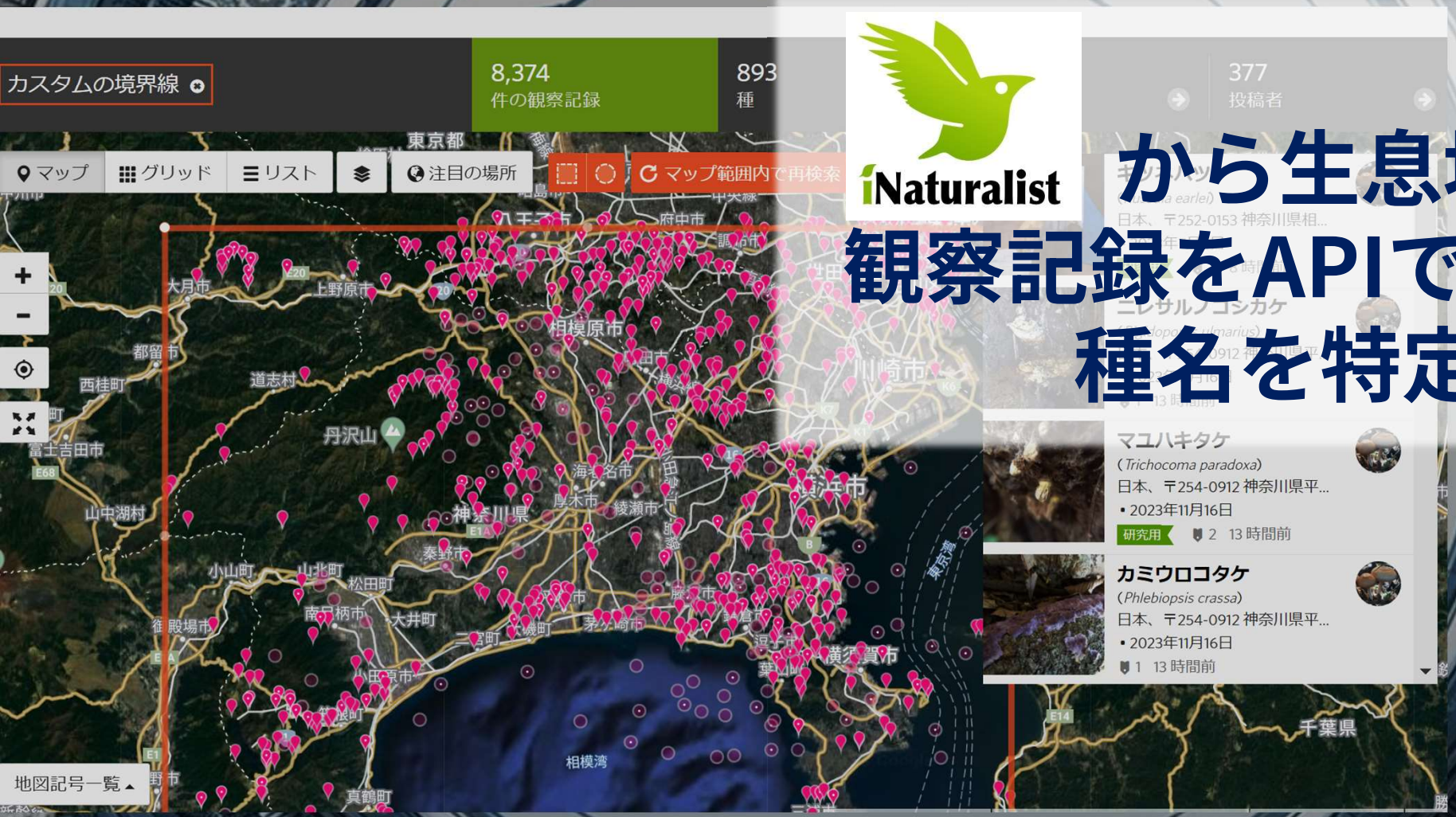
- ある生息地にどの種がいるか
- ある種がどの生息地にいるか

…を推定したい。

# GNNを構築する – 生息地データ


国土交通省がメッシュごとの
気温、降水量、平均標高などの
データを公開している

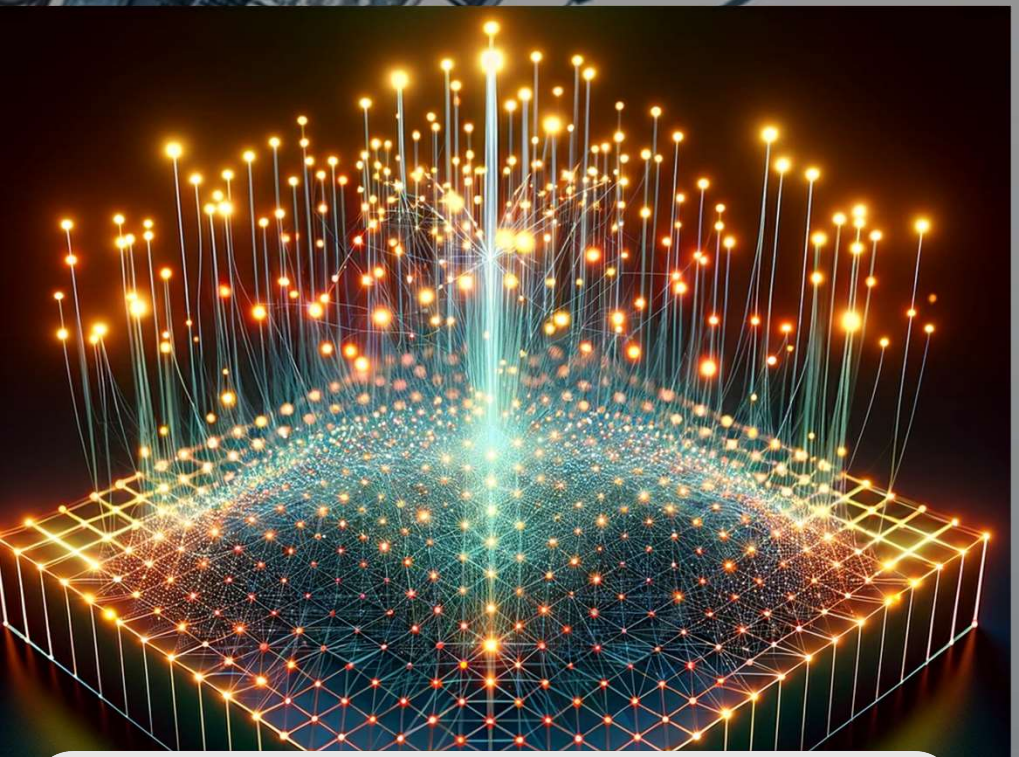
# GNNを構築する – 種名データ




iNaturalist

## から生息地ごとの観察記録をAPIで取得し種名を特定

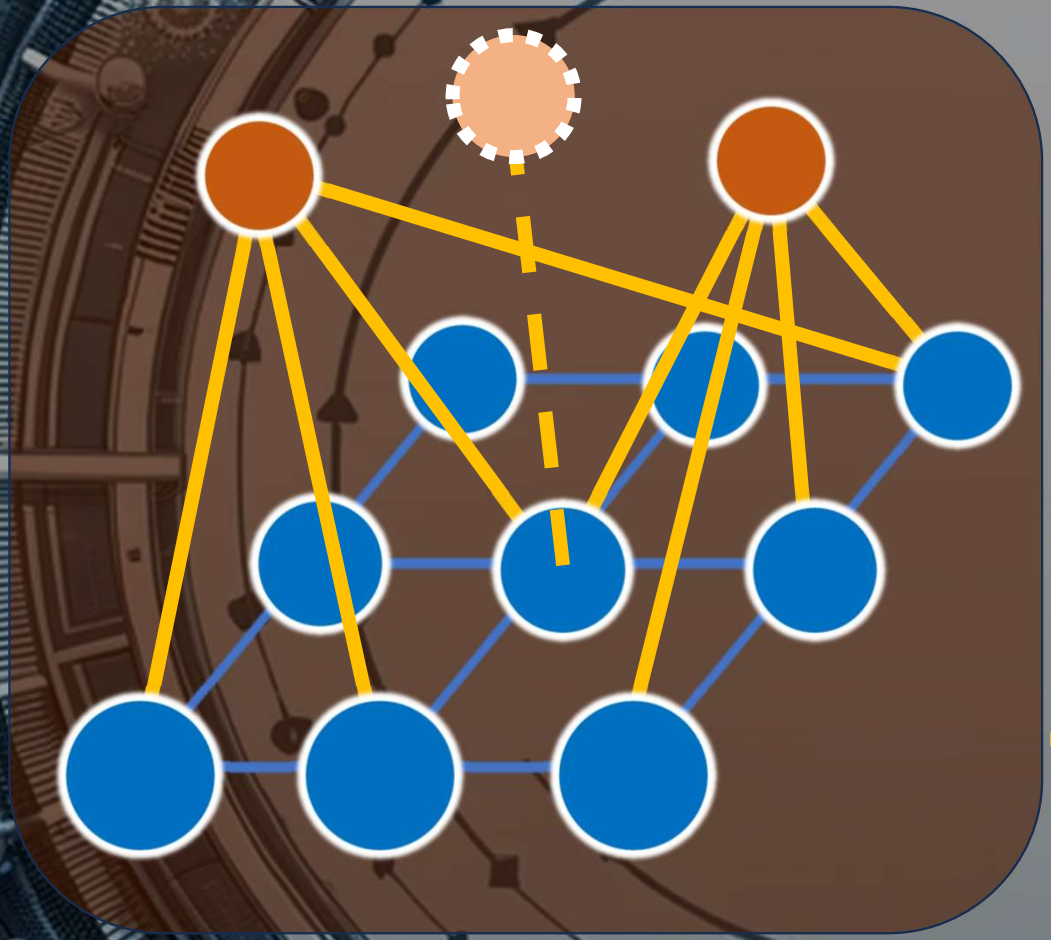
# GNNを構築する – 種名データ



生息地と種名の繋がりを
ネットワークで表現

種名の特徴量として
今回作った
ベクトルデータ
を用いる

# まとめると…



**生息地ノード**

気温、降水量など

**種名ノード**

形質特徴量ベクトル

予測したい
ノードやエッジ

# GNNモデルができるとどうなる？

- 調査不能/不十分な地点の菌類相予測
- ある地点で未発見の種の予測
- 経時的変化の検出（気候変動、ナラ枯れ）

今後もデータの拡充と
GNNの構築を目指していきたい


おわり