

データ分析によるデジタル・トランスフォーメーション

目次

● データ分析とは

- 意義
- 代表的手法
- 注意点

● 活用事例

- スポーツ分野
- 商業分野
- 建設業・製造業分野

● データ分析ツール

- 主なパッケージソフト
- Google Colaboratory
- データ分析ライブラリ

● オープンデータ活用

- 政府統計
- RESAS
- オープンデータ利活用事例

● 学び方

- オンライン講座
- 資格・試験
- コンテスト

データ分析とは

データ分析(data analysis)

●Wikipediaでは

「有用な情報を発見し、結論を報告し、意思決定を支援することを目的として、データを検査し、クリーニングや変換を経て、モデル化する一連のプロセス」

クリーニング

異常なデータ(異常値)を取り除いたり、欠落したデータを補う

モデル化

些末な細部に捕らわれず、重要な特徴のみを表現したモデル(模型)を作成する

データ分析の意義

- 過去データから将来を予測可能

例

過去の月次販売数の推移から来月の販売数を予測し、仕入数を決める。

- 複数のデータの関連性を推定可能

例

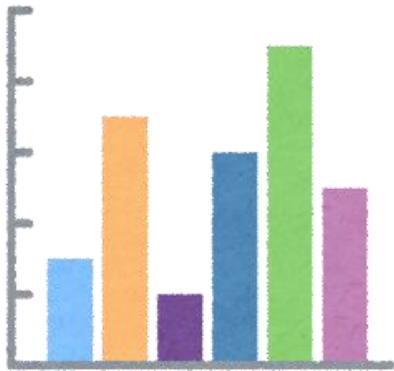
POSデータから、同時購入されることが多い商品を発見し、近隣に陳列することで、「ついで買い」を促す。

- まだラベルがわかっていない新たなデータのラベルを推定可能

例

自動車の加速度データと事故履歴(ラベル)から、事故発生の可能性が高い加速度を検知すると運転者に警告する。

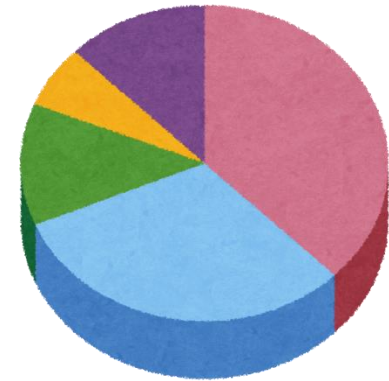
代表的なデータ分析手法



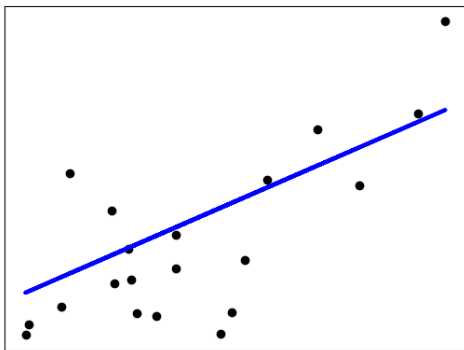
度数分析



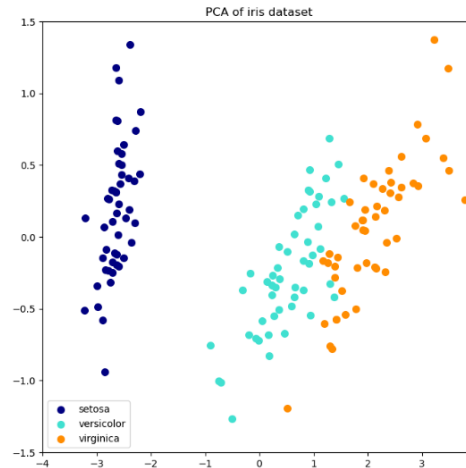
バスケット分析



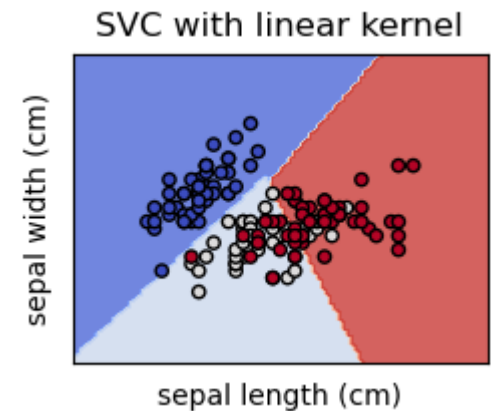
ABC分析



回帰分析



主成分分析

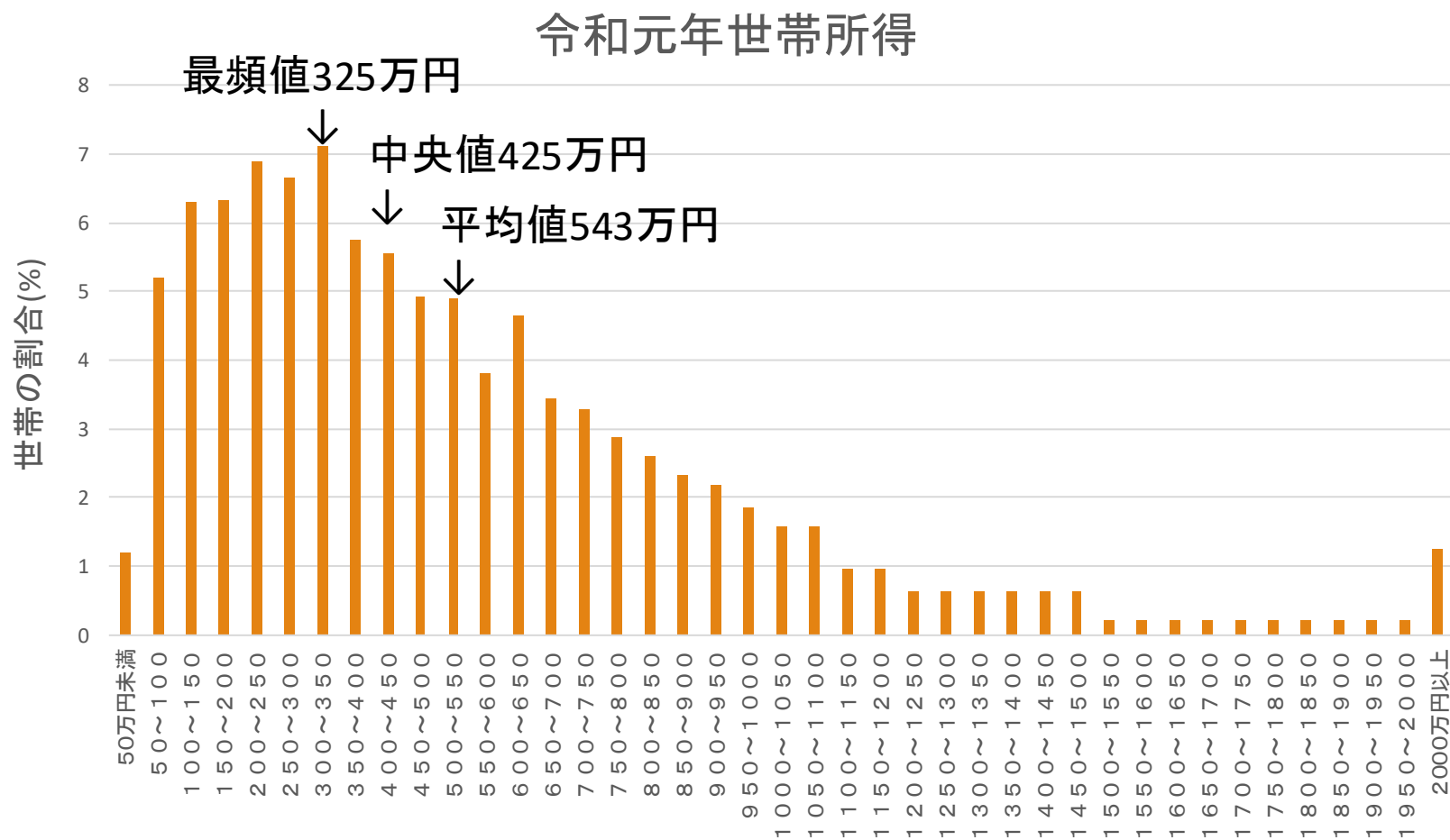


クラスター分析

下段の出典: scikit-learnホームページ

度数分析の一例

- 平均値がデータの特徴を表す最適指標とは限らない



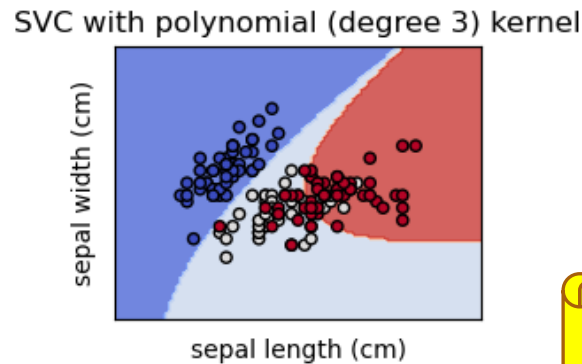
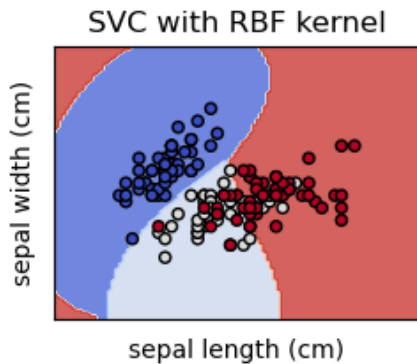
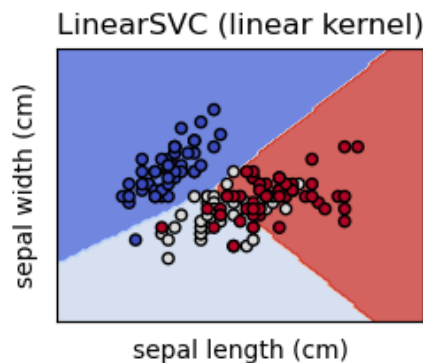
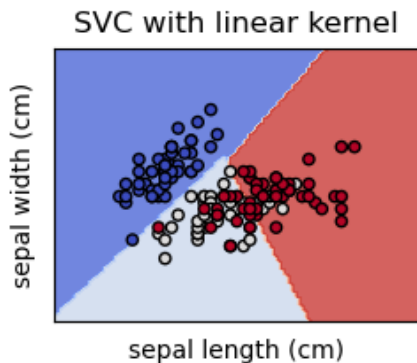
出典: 令和元年国民生活基礎調査のデータに基づき著者作成

世帯所得

クラスター分析の一例

【問題】

- 3種類のアヤメ(setosa, versicolor, virginica)を対象に、多数のアヤメの花弁(sepal)の幅(sepal width)と長さ(sepal length)を測定したデータ集合がある。
- 新たに採取した種類不明のアヤメの花弁を測定し、その種類を推定したい。

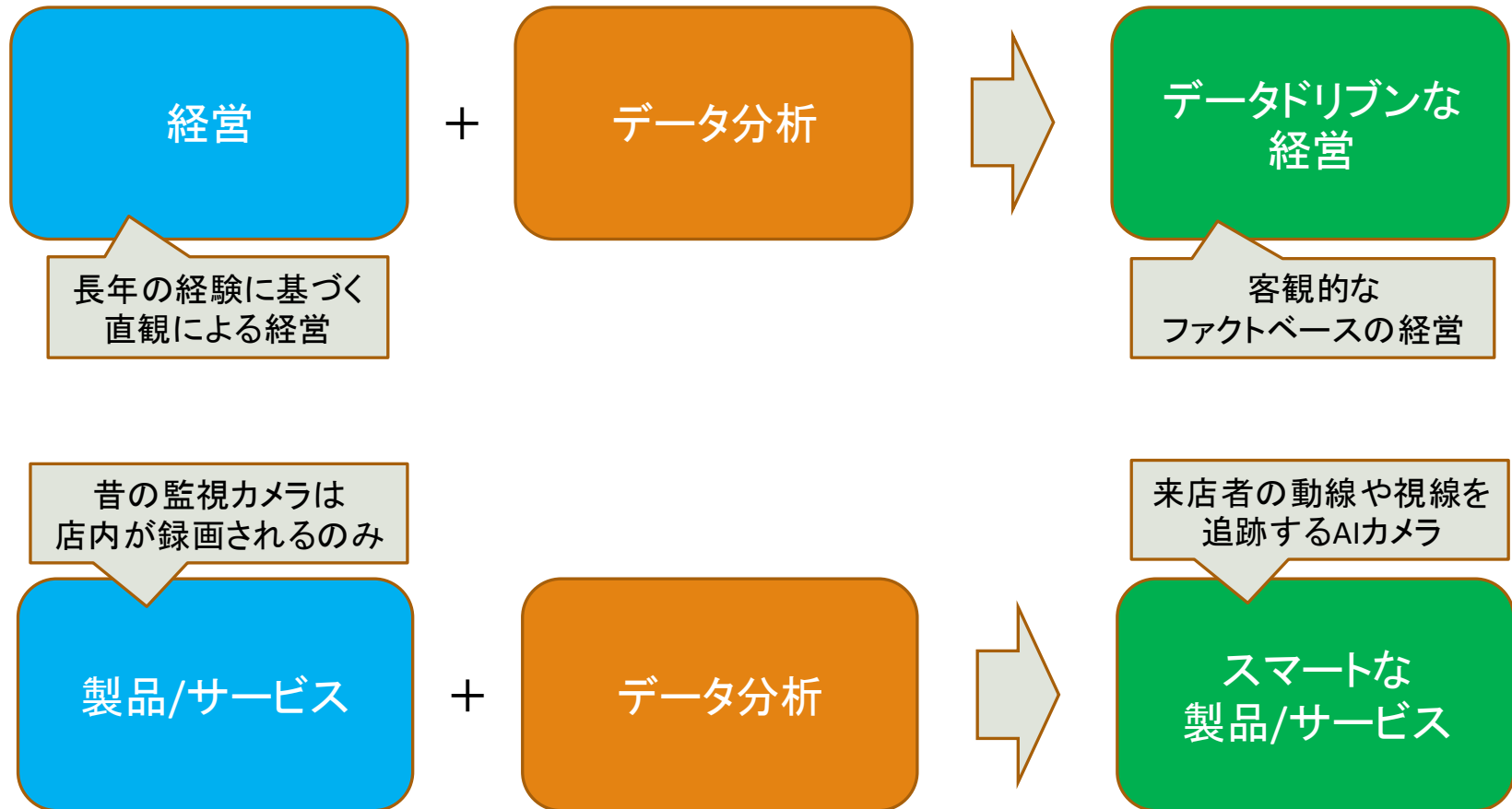


- ① アヤメのデータ集合を2次元平面にプロット。(アヤメの種類を○の色で表現)
- ② できるだけシンプルかつ精度高く3種類に分類する線を描く。
- ③ 左図では、4つの異なる手法で線を描いている。
- ④ 新たに採取したアヤメの花弁の測定値を左図にプロット。プロットした地点の背景色がアヤメの種類を表す。

出典: scikit-learnホームページの例題

- どの分類手法も、精度100%ではない
- 採取したアヤメの種類の推定も、精度100%というわけにはいかない

データ分析によるデジタル・トランスフォーメーション



データ分析の注意点

1. データ分析は100%正解を導く訳ではない
2. 目的に合ったデータを集めることが重要

データ収集のコストが課題

ポイントカードのデータで、顧客が誕生日近辺で購買しているのか分析したい



顧客データベース

誕生日のデータがない

無目的に手元に集まったデータ、あるいは異目的で集めたデータを流用しようとしても、データを集め直す羽目に成りがち

データ収集の事例

- RealNetworks社は、顔認識の製品・サービスSAFRを提供中
- 顔認識では、膨大な顔画像を収集し、機械学習させる必要がある
- 動画作成製品RealTimesを通じて顔画像を世界中から収集し、SAFRに利用（もちろんユーザーの了解の上で）
- 海外企業は戦略的なデータ収集が上手い

データサイエンティスト

●「データサイエンティスト」の定義

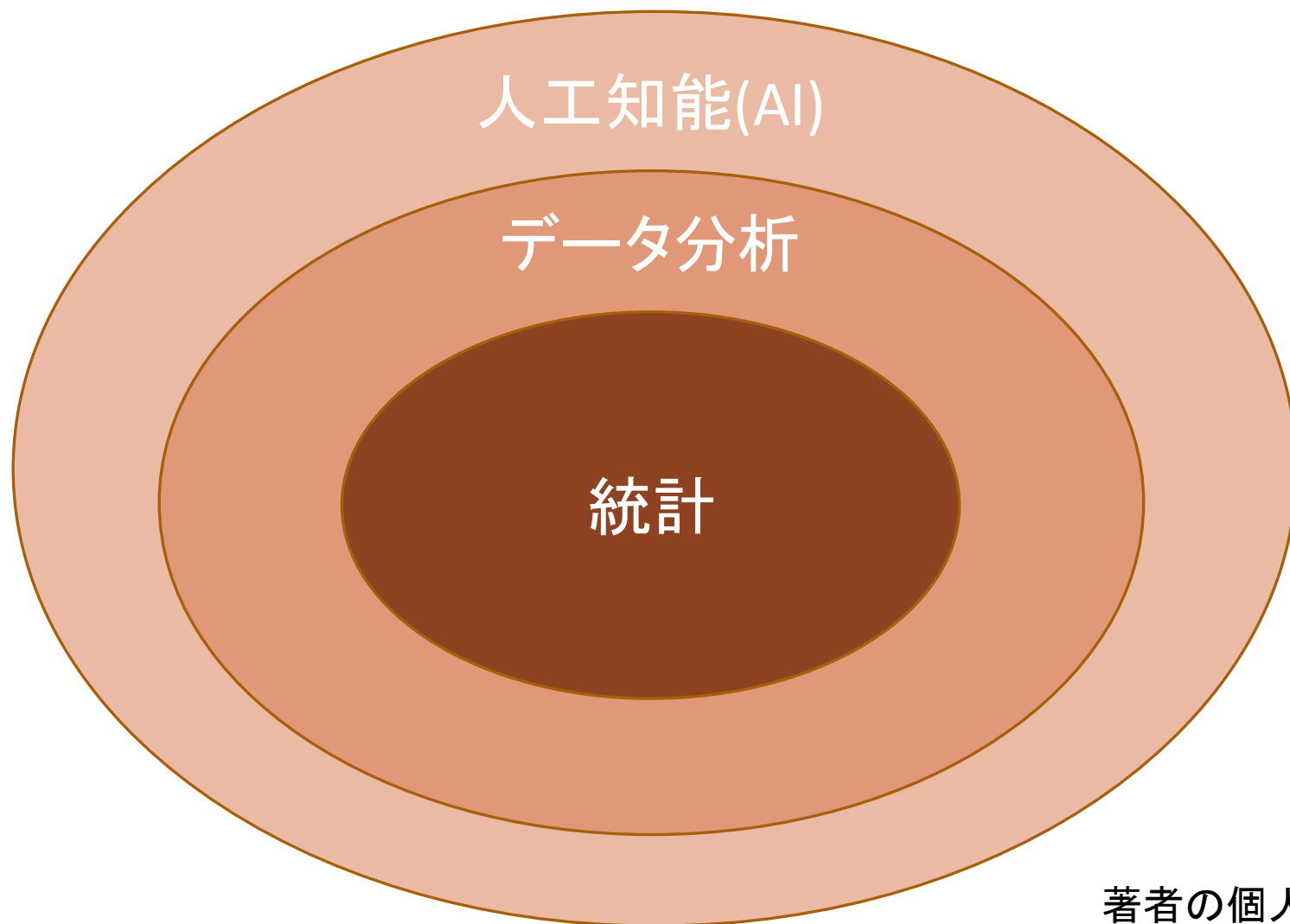
明確な定義はなく、「ビジネスに役立つようなデータ分析を行う人」と仮置きしておきます

●「データサイエンス」「データ分析」が流行っている理由

- ① インターネットの普及によってビッグデータが集まりやすく、かつアクセスしやすくなったため
- ② 各種センサやメディア処理の技術進歩により、我々の住む実世界のアナログ情報が計算機上のサイバー世界のデジタル情報へ変換できるようになったため(⇒デジタルツイン)

●「データサイエンティストは21世紀で最もセクシーな職業である」 by 2012年ハーバードビジネスレビュー

データ分析、人工知能、統計



著者の個人的解釈

活用事例

スポーツ分野の活用事例

- スポーツのデータ分析市場は2026年までに51.1億ドル（約6,380億円）

出典：日本経済新聞電子版2022/4/17 いまどきの部活②

（原出典：モルドールインテリジェンス調査）

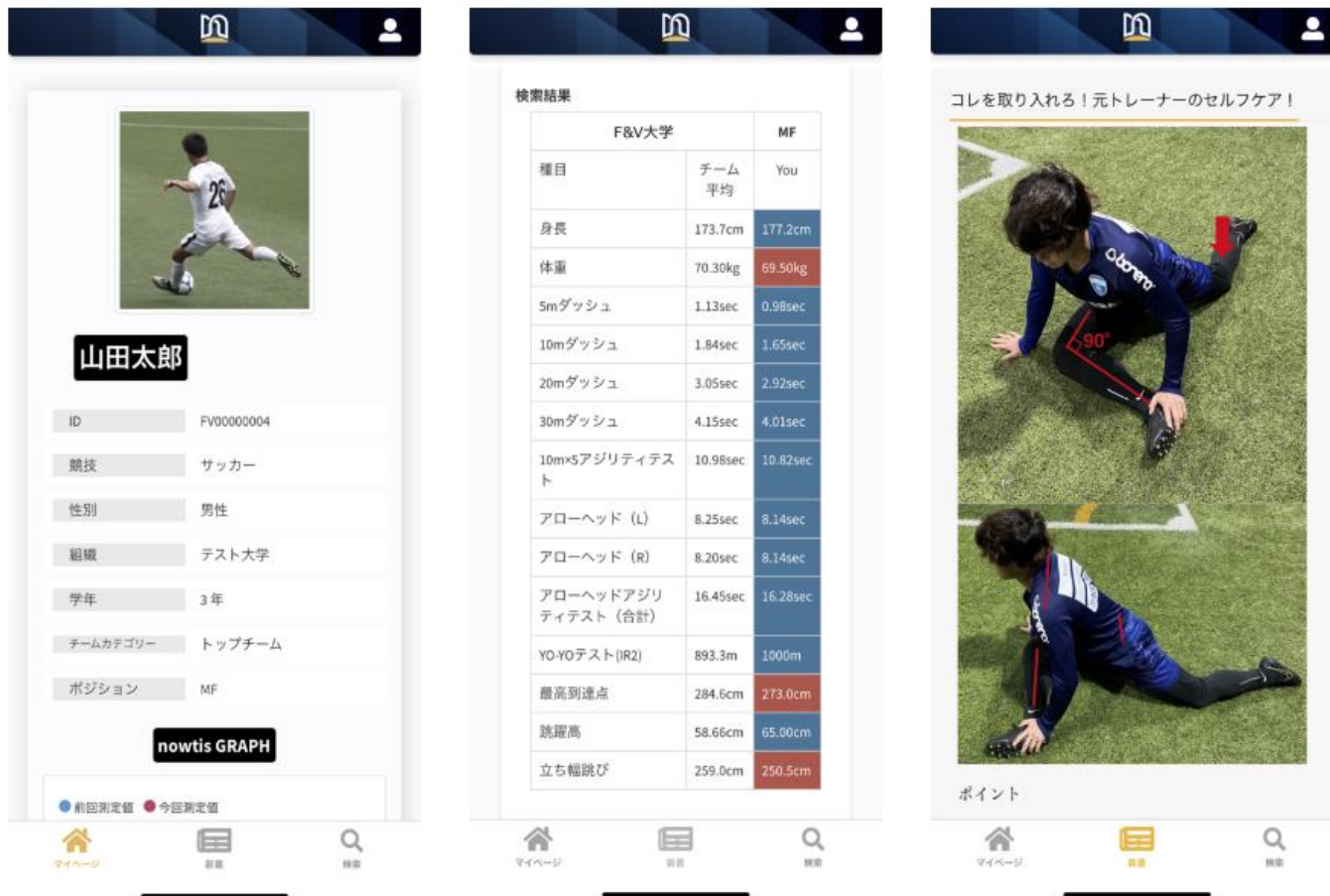
- 【従来】データ取得のために、特殊なデバイスやマーカの装着が必要



【現在】AI映像解析技術の進化によって、カメラ撮影だけでデータ取得が可能に

(株)F&V

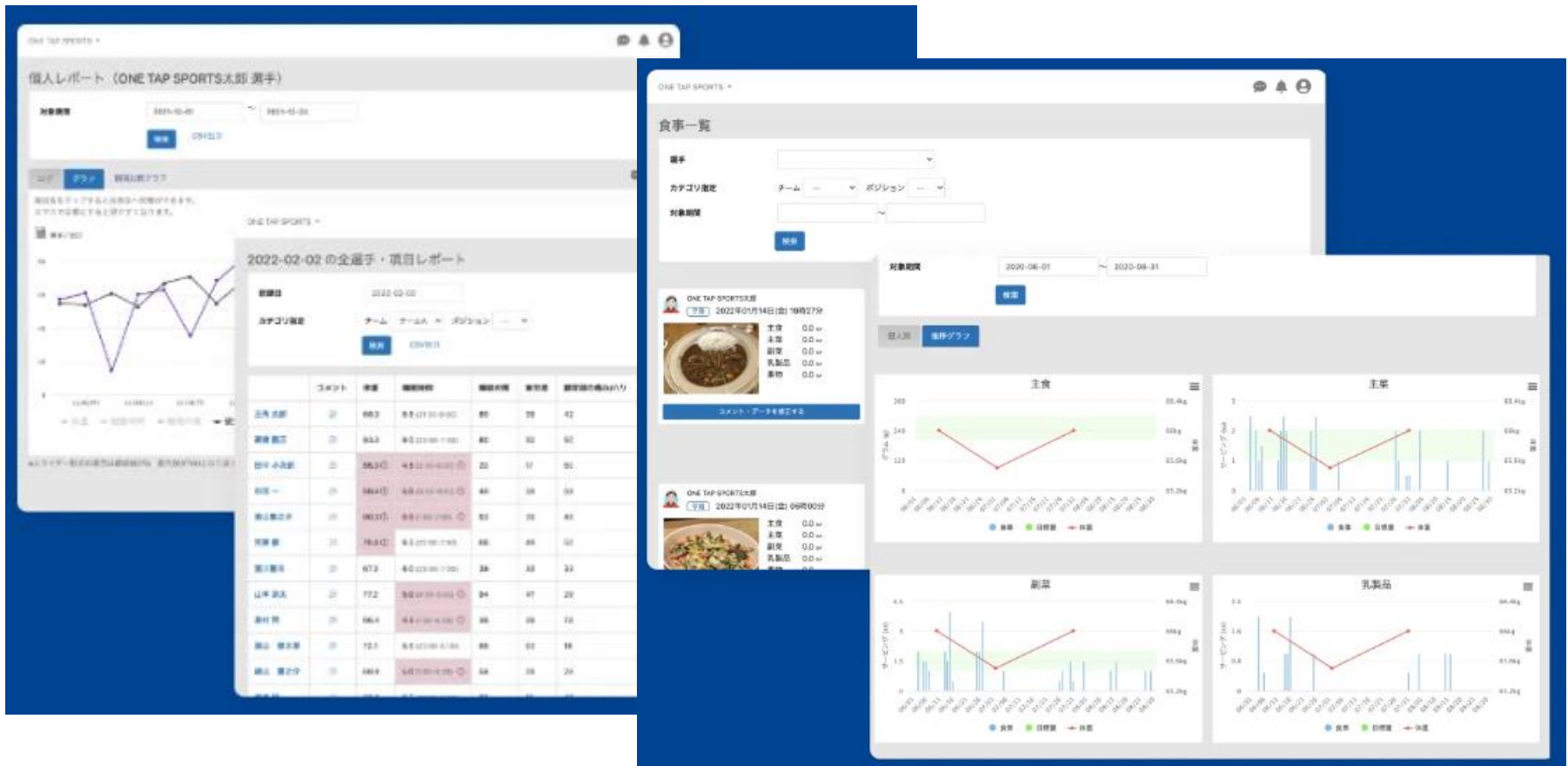
各競技に最適なテスト項目でアスリートの基礎能力を測定。プロになった選手やライバルチームの選手とのデータ比較が可能なアプリでアスリートの能力を可視化。



出典: <https://f-and-v.com/>

(株)ユーフォリア

選手のコンディションを見える化。疲労度・体温・睡眠時間などの体調データ、トレーニング内容や運動負荷、食事内容を日々記録し、監督・コーチはチーム全体と各個人の状況を把握可能。管理栄養士による栄養指導も可能。



出典: <https://one-tap.jp/>

(株)ネクストベース

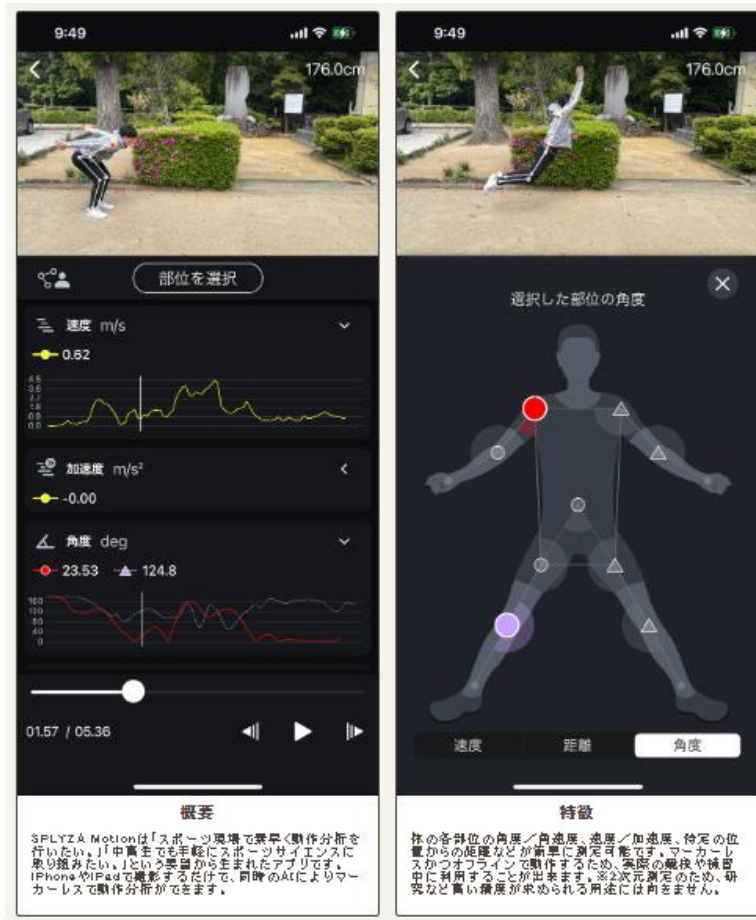
野球の投球・打球のスピードをスマホで計測。計測データのグラフ表示、打球のホームラン/ヒット/アウト推定、スローモーションでのフォームチェックが可能。



出典: <https://nextbase.co.jp/>

(株)SPLYZA

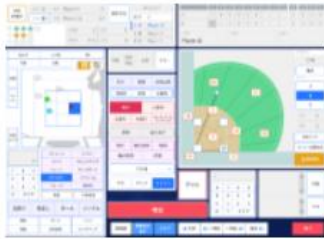
肩・肘・膝・足首といった特徴点をAI映像解析によって抽出し、選手の動作を分析。マーカを装着することなく、スマホ/タブレットのカメラ撮影で動作分析が可能。



出典: <https://www.splyza.com/>

データスタジアム(株)

サッカー・野球・バスケットボール・卓球など様々なスポーツ向けにデータ分析サービスを提供



Charlyze System

1球ごとの詳細なデータ入力と分析が可能な野球専用システム



Pitch Base

高速シーン再生技術を搭載した、野球映像の検索・分析アプリケーション



Rapsodo Baseball

Rapsodo社が開発した野球・ソフトボールの3Dトラッキングシステム

Football Analyzer
Data Analysis & Video Editor



Football Analyzer

1プレーごとのデータと映像を組み合わせた分析が可能なサッカー専用システム



Football BOX

プレーやさまざまな検索条件に合わせて映像検索が可能なサッカー専用システム

Rugby Analyzer
Data Analysis & Video Editor



Rugby Analyzer

1プレーごとのデータと映像を組み合わせた分析が可能なラグビー専用システム

出典: <https://www.datastadium.co.jp/>

商業分野の活用事例

- 陳列棚の最適化
- コールセンタのCS向上
- インバウンド顧客への接客

(株)サイバーリンクス

陳列棚のフェイスをスマホで撮影するだけで、棚割の現状を即座にデータ化。棚割データとPOSデータの分析により、棚割を最適化し、売上アップにつなぐ。



出典: <https://cyberlinks-portal.com/service/tanascan-ai/>

NTTテクノクロス(株)

コールセンタの通話内容を音声認識によりテキスト化。質問に対する回答をオペレータへレコメンドしたり、伝えるべき重要事項を伝えたかどうかを自動チェック。通話相手の感情を認識し、適切な対応によって顧客満足度を向上。



出典: https://www.ntt-tx.co.jp/products/foresight_vm/

情報通信研究機構(NICT)

31言語に対応した多言語音声翻訳アプリ。対訳データのビッグデータを統計処理することでリアルタイムの機械翻訳を可能とした。



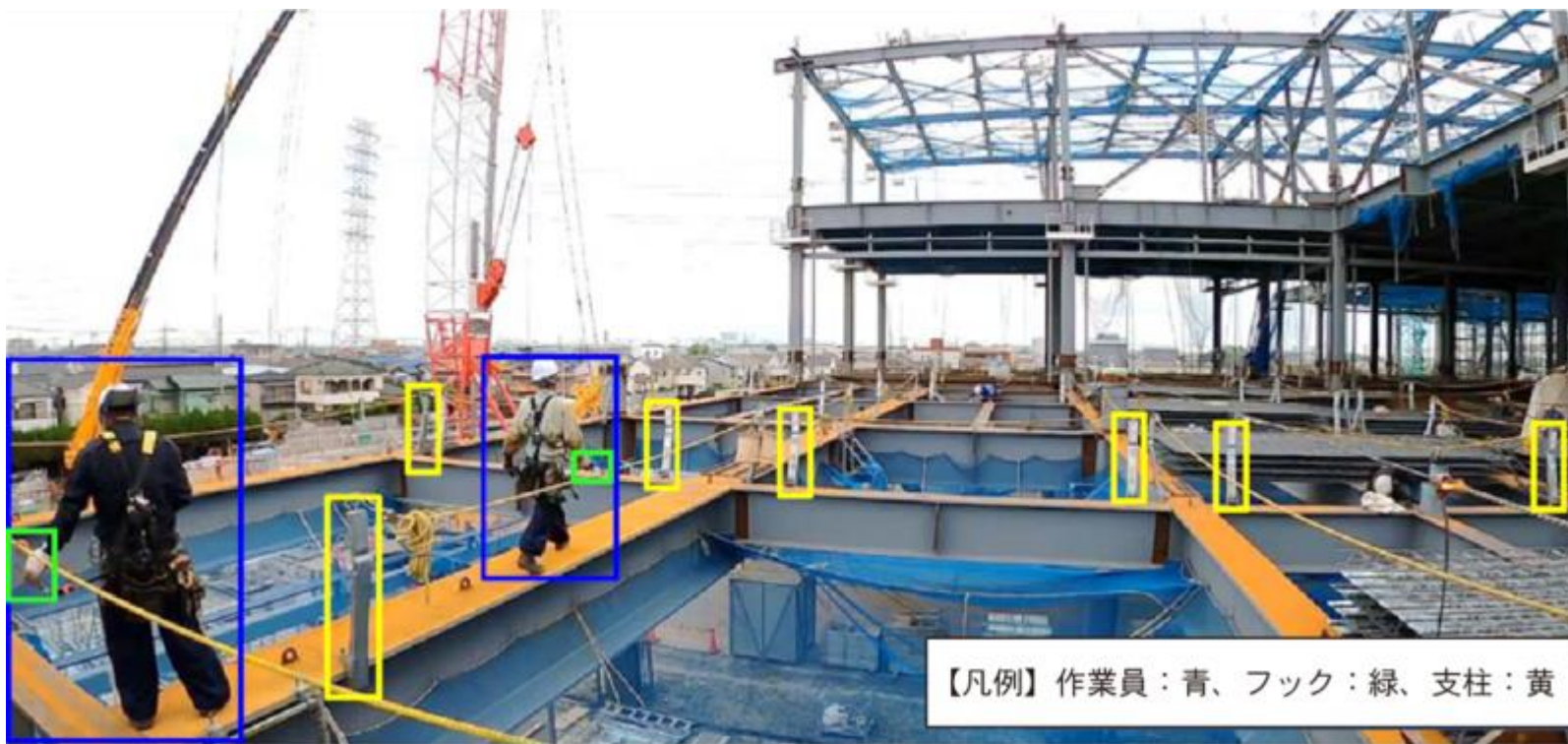
出典: <https://voicetra.nict.go.jp/index.html>

建設業・製造業分野の活用事例

- 作業員の安全管理
- 打音検査のDX化
- 既設設備のDX化

(株)奥村組

墜落制止用器具のフック掛け忘れをAIカメラで自動検知。
出典元のニュースリリースによれば、人や物体の重なり合いがなければ、15m以内で90%以上の精度で検知可能とのこと。



出典：<https://www.okumuragumi.co.jp/newsrelease/2022/ai.html>

福井工業高専

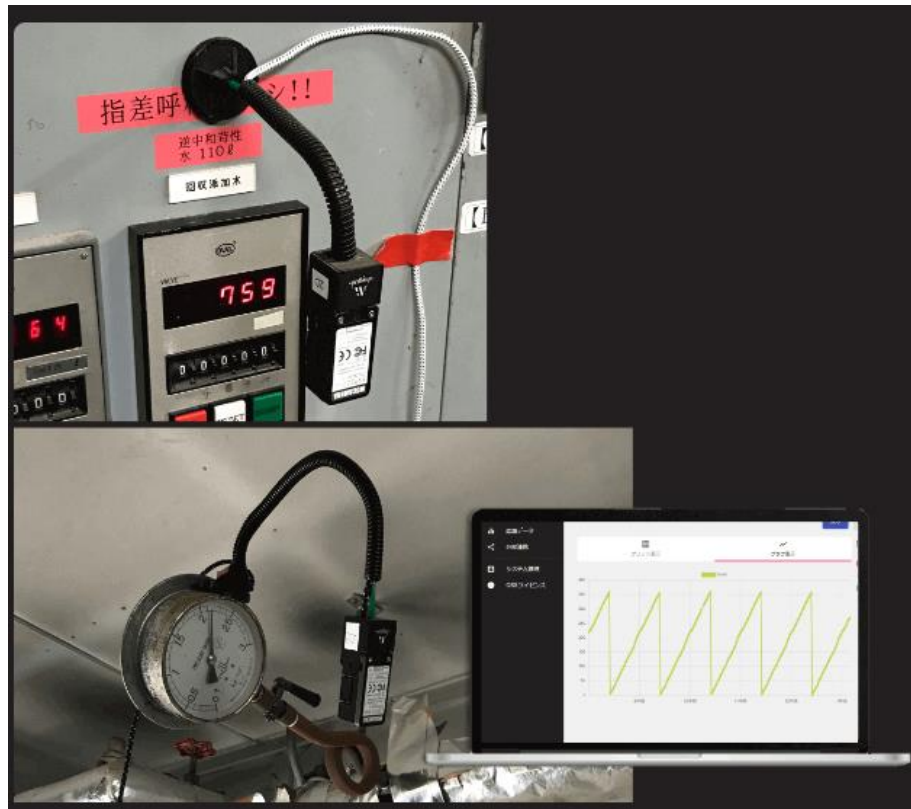
橋・トンネル・壁などの構造物の劣化判定で用いられている打音検査のAI化。
従来は、音の聞き分けに長年の熟練が必要であったが、センサデータに深層学習を適用することで、初心者でも打音検査を可能とした。
「全国高等専門学校ディープラーニングコンテスト2021」(高専DCON)で最優秀賞を獲得し、審査員のベンチャキャピタリストの企業評価額は6億円。



出典:<https://www.jdla.org/news/20210417001-2/>

(株)integrAI

既設設備のアナログメータ、デジタルメータ、ランプなどにAIカメラを外付けすることで、目視でのデータ取得から自動データ収集に業務プロセスを変革。
(株)integrAIは、DCON2019最優秀賞を受賞した長岡工業高専関係者の起業。



出典: <https://integr.ai/>

データ分析ツール

主なデータ分析パッケージソフト

- データ分析に使える主なパッケージソフト
 - Microsoft Excel
 - IBM SPSS
 - SAS
 - DataRobot
- Excelでも簡易なデータ分析は可能
- SPSS, SAS, DataRobotは高度な分析が可能だが、安価とは言えない



プログラミングは必要となるが、
無料のデータ分析ツールを使い倒してから
高度なパッケージソフトを導入するのが良さそう

RとPythonのどちらを選択するか

● R

統計分析に用いられるシンプルな言語。利用は無料。Web上に様々なライブラリが提供されており、ライブラリを利用して複雑なデータ分析をプログラミング可能。

総務省統計局「統計分析ソフトの使い方 R(アール)」

<https://www.stat.go.jp/teacher/r.html>

Rの配布元

<https://cran.r-project.org/index.html>

● Python

「読みやすさ」と「わかりやすさ」を重視したプログラミング言語。利用は無料。アプリケーション開発・Webシステム・科学技術計算など広く利用される。Web上には、Python対応の様々なライブラリが提供されており、データ分析においても昨今はRよりPythonが使われることが多い。

Pythonの配布元

(次頁に紹介するGoogle Colaboratoryを使えばインストール不要)

<https://www.python.org/>

今からデータ分析を学ぶならPythonがオススメ

Google Colaboratory (Colab)

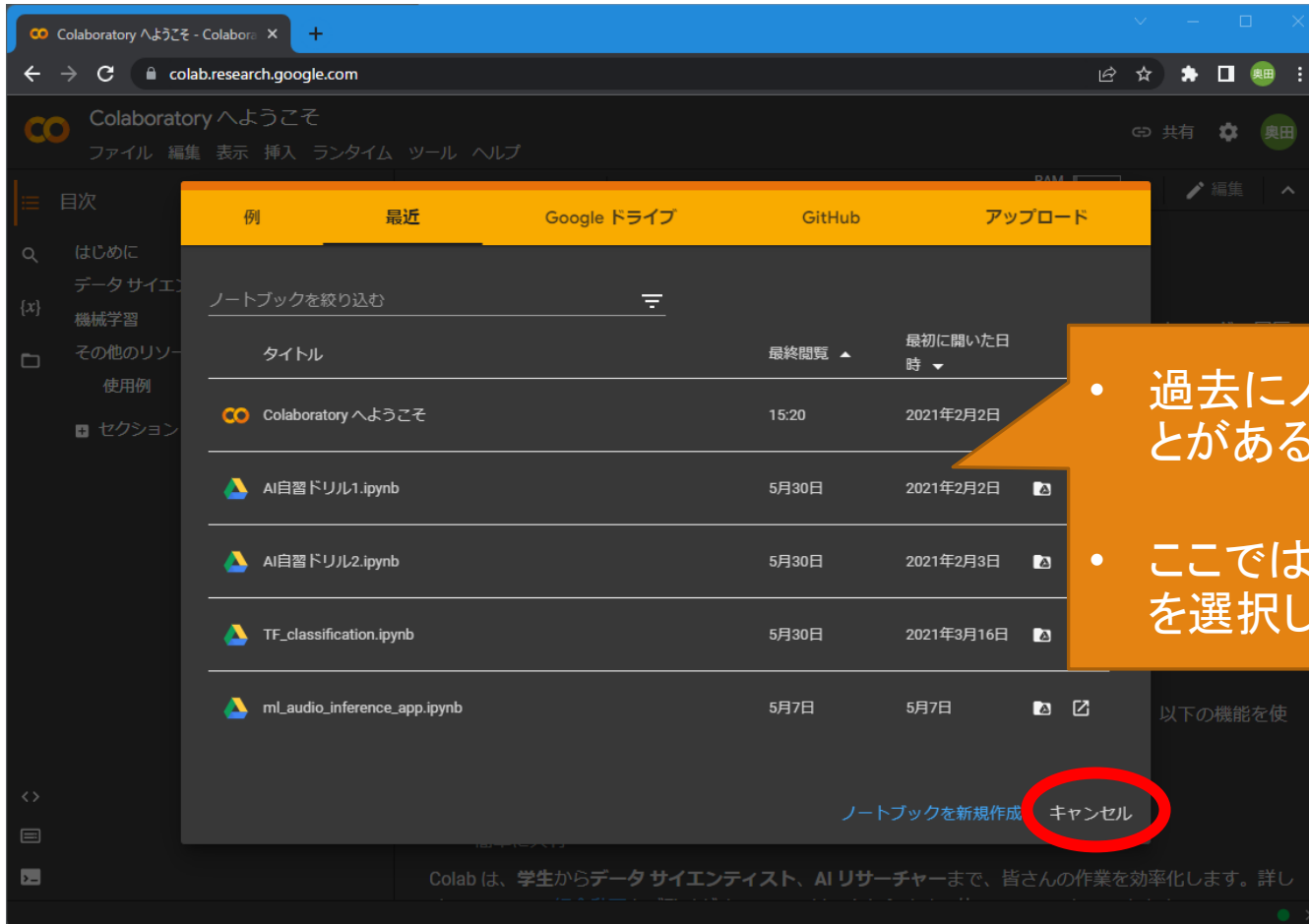
- Googleが提供している教育・研究用の開発環境。
- プログラミング言語Pythonを用いたデータ分析が可能。
- Googleアカウントを持っていれば、一般的な利用であれば無料。
- 自分のPCへのソフトウェアインストールは不要。WebブラウザからGoogle Colabのサイトへアクセスして利用。

<https://colab.research.google.com/>

- 見栄えの良い説明文とプログラムを同じ画面に並記できるため、プログラムを理解しやすい。
- データ分析に役立つPythonの主要なライブラリ/パッケージがあらかじめプレインストール。
- プレインストールされていないライブラリ/パッケージを、自分で追加インストールすることも可能。

Google Colabを試してみよう (1/5)

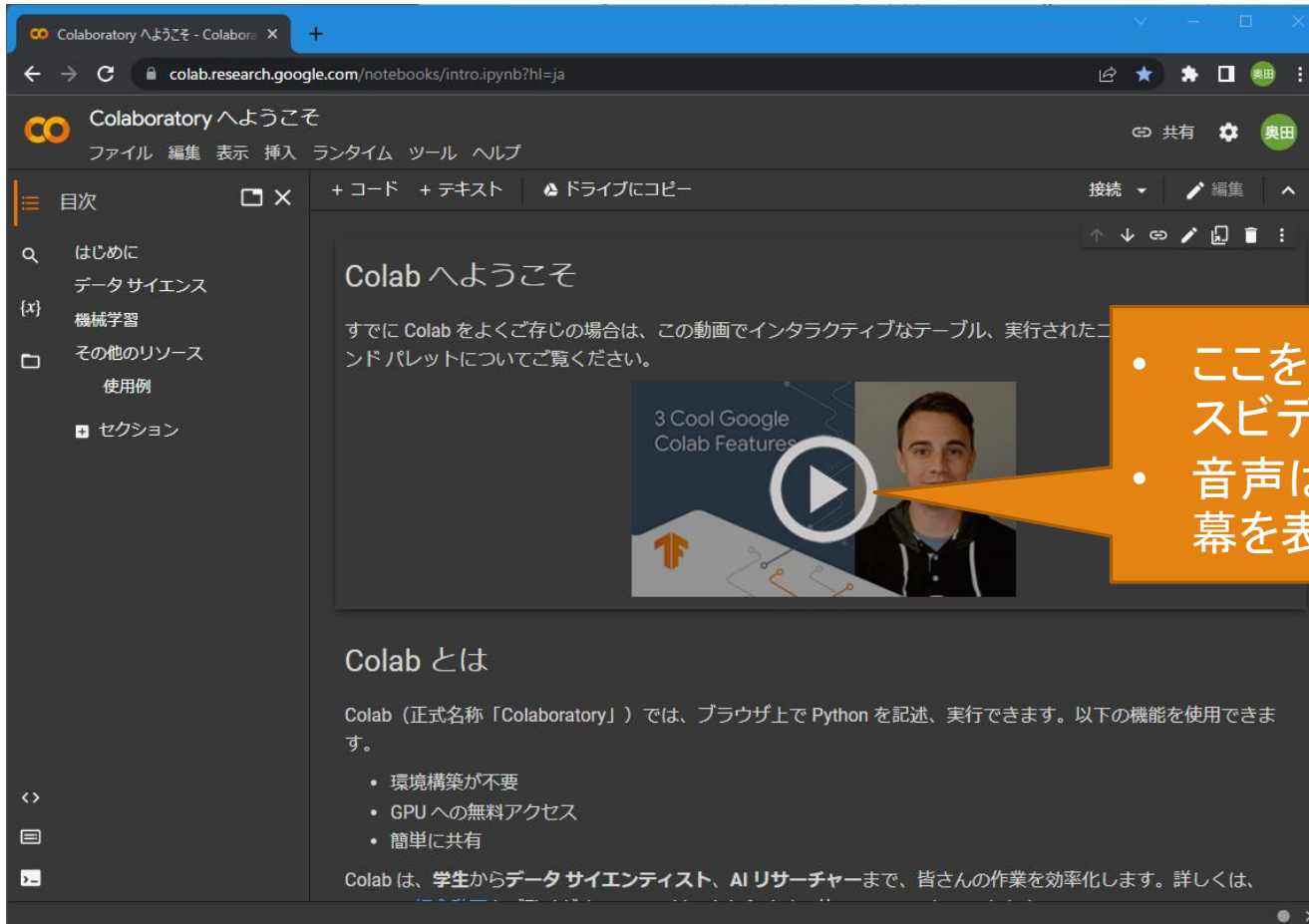
GoogleアカウントでGoogleにログインしてから、<https://colab.research.google.com/>へアクセスすると下記画面が表示される。



- 過去にノートブックを作ったことがあると履歴が表示される。
- ここでは下部の「キャンセル」を選択して、選択画面を消す。

Google Colabを試してみよう (2/5)

前頁の画面で「キャンセル」をクリックすると、「Colabへようこそ」画面を表示。



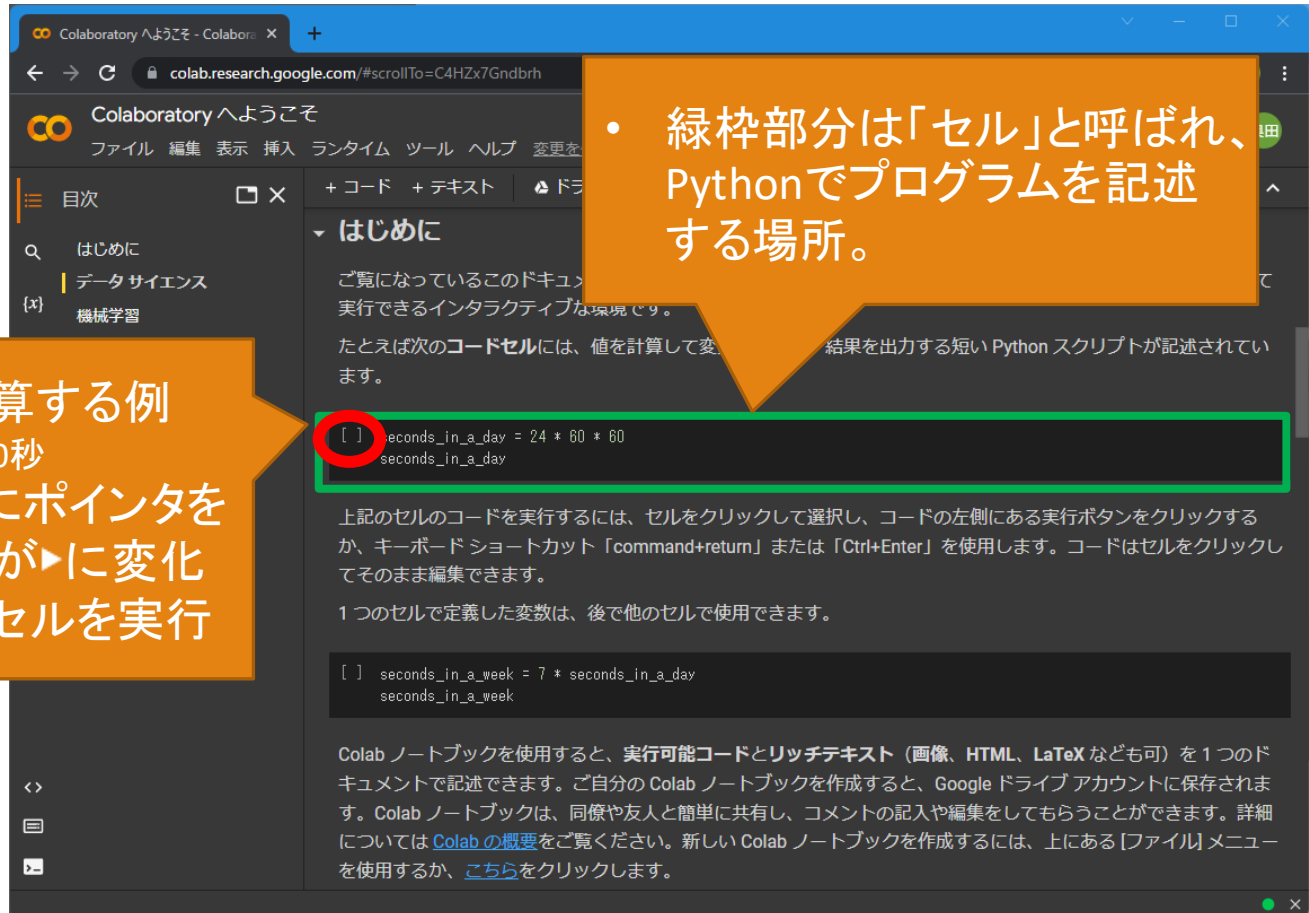
- ここをクリックするとガイダンスビデオが開始。
- 音声は英語だが、日本語字幕を表示可能。

Google Colabを試してみよう (3/5)

前頁の画面をスクロールして「はじめに」を表示。

- 一日の秒数を計算する例
24時間 × 60分 × 60秒
- 赤丸で囲んだ[]にポインタをかざすと、マークが▶に変化
- ▶をクリックして、セルを実行

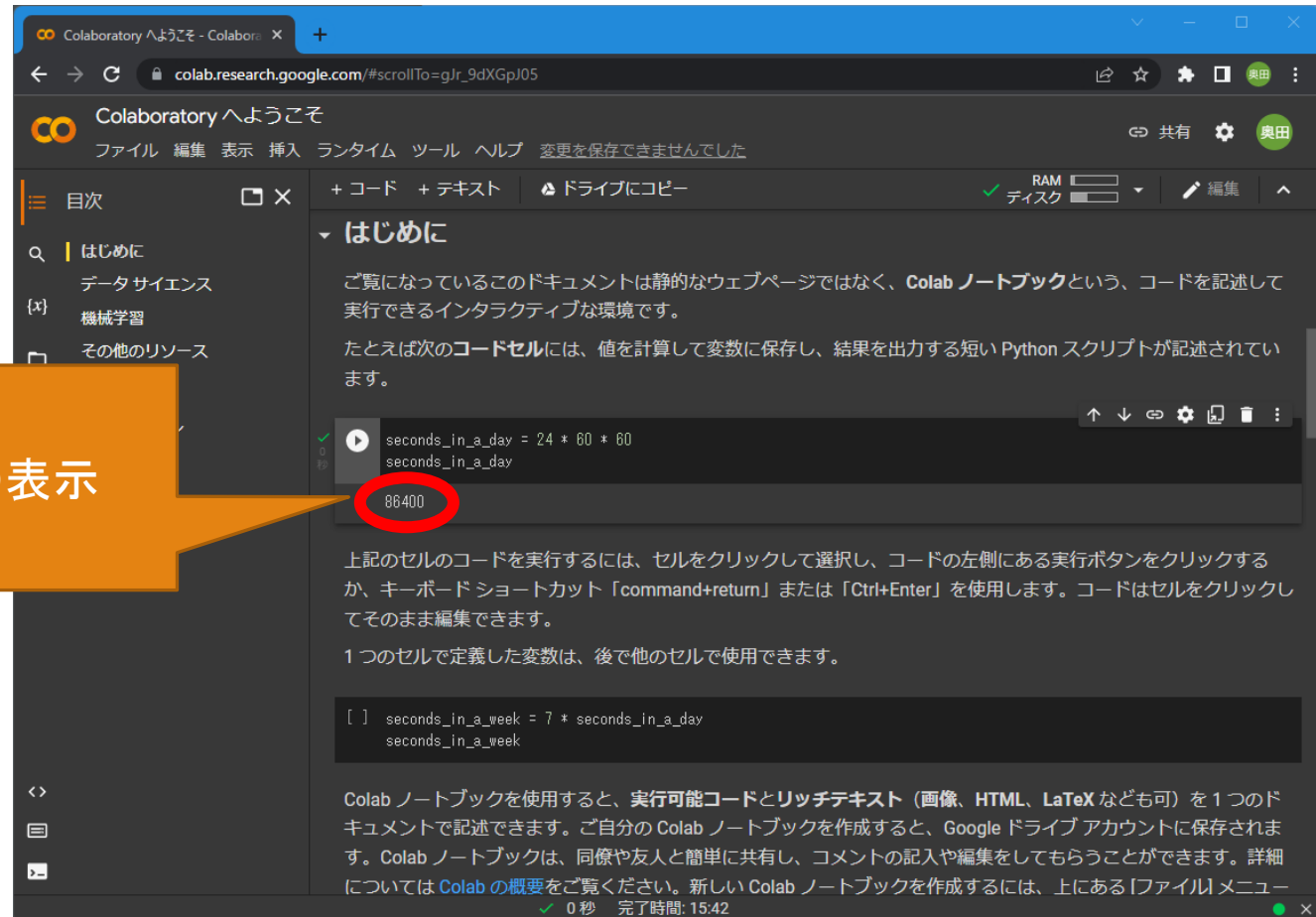
• 緑枠部分は「セル」と呼ばれ、Pythonでプログラムを記述する場所。



Google Colabを試してみよう (4/5)

一日の秒数の計算結果を表示

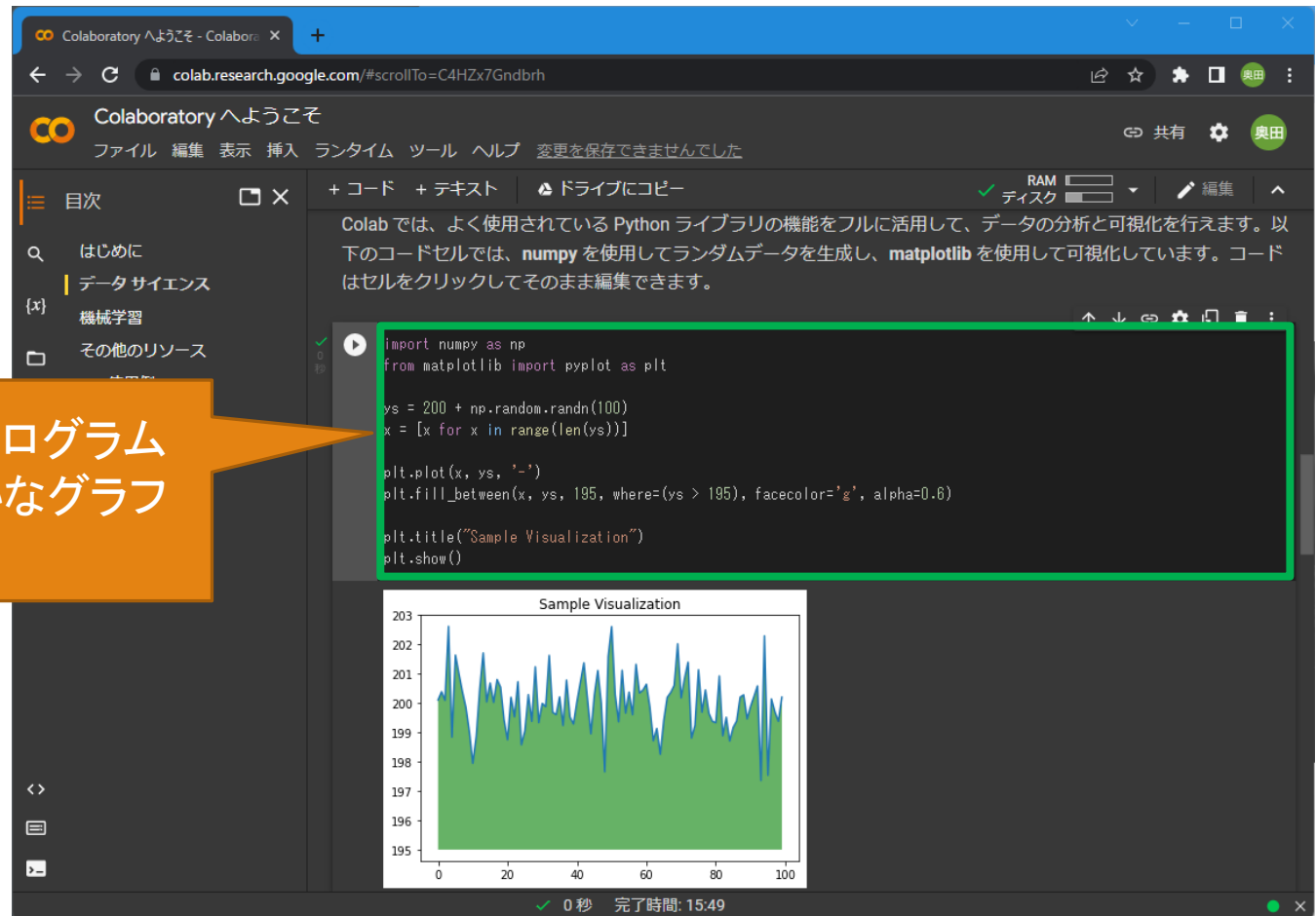
- 計算結果86,400の表示



Google Colabを試してみよう (5/5)

一日の秒数の計算結果を表示

- 緑枠の数行のプログラムによって、きれいなグラフを表示



データ分析ライブラリ

● Pythonで利用可能な主な無料ライブラリ

名称	概要	URL
NumPy	科学計算の基本パッケージ。多次元配列・行列の操作を高速に演算。	https://numpy.org/
SciPy	科学計算・技術計算のパッケージ。FFT、信号処理など高度な演算をサポート。	https://scipy.org/
pandas	データフレームなどのデータ構造を持ち、効率的な前処理が可能。	https://pandas.pydata.org/
scikit-learn	機械学習ライブラリ。様々な分類、回帰、クラスタリングが可能。	https://scikit-learn.org/stable/
matplotlib	可視化パッケージ。プロット、散布図、棒グラフ、円グラフなど様々な表現が可能。	https://matplotlib.org/

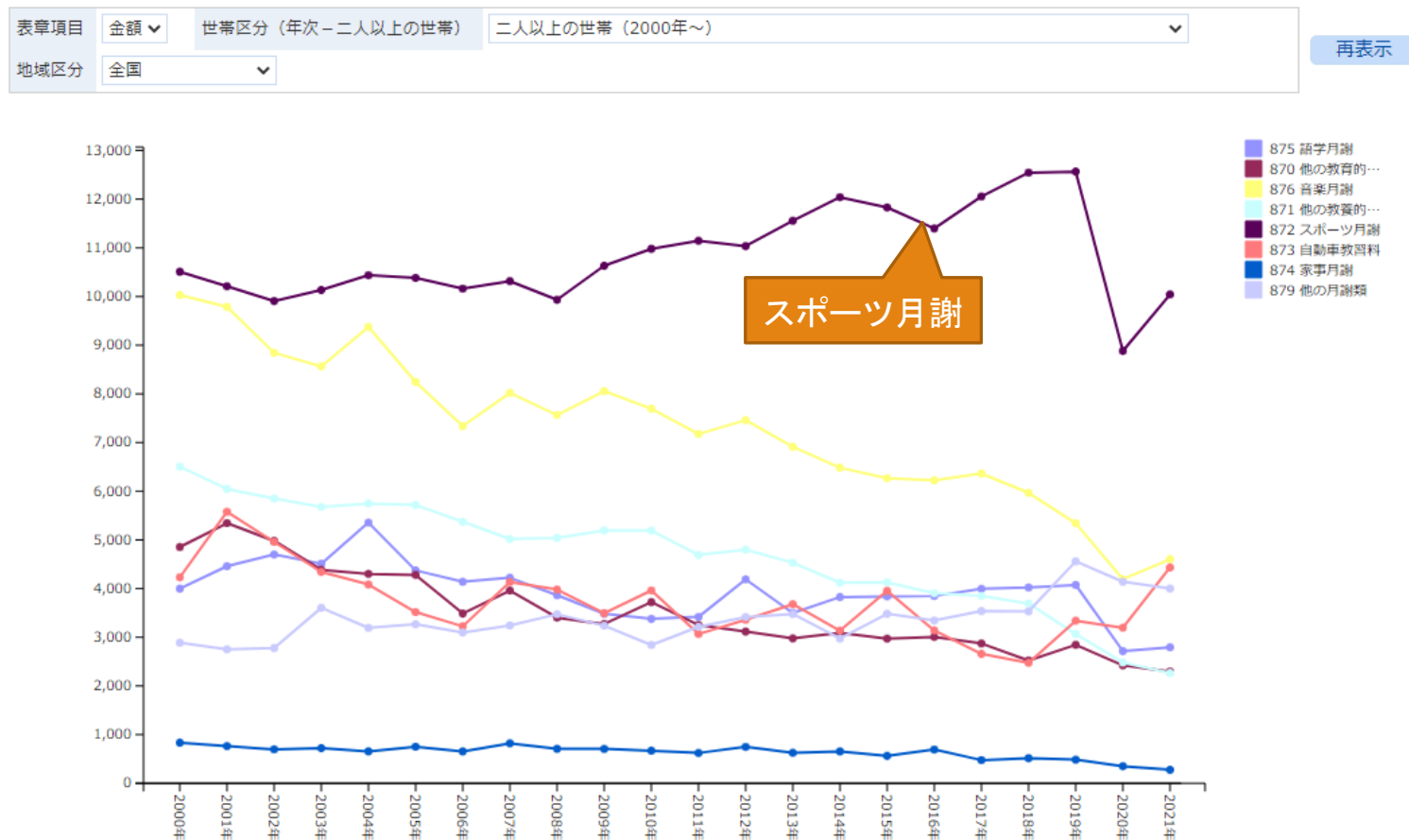
オープンデータ活用

政府統計の総合窓口 e-Stat

- 政府が公表する統計データのポータルサイト。総務省統計局/統計センターが運営。
- 統計データにフィルタをかけた結果を、Excel形式やcsv形式でダウンロードできる。
- ホームページ <https://www.e-stat.go.jp/>

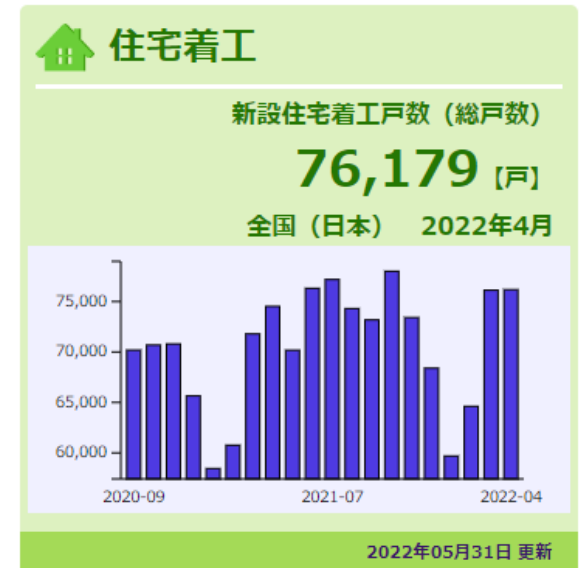
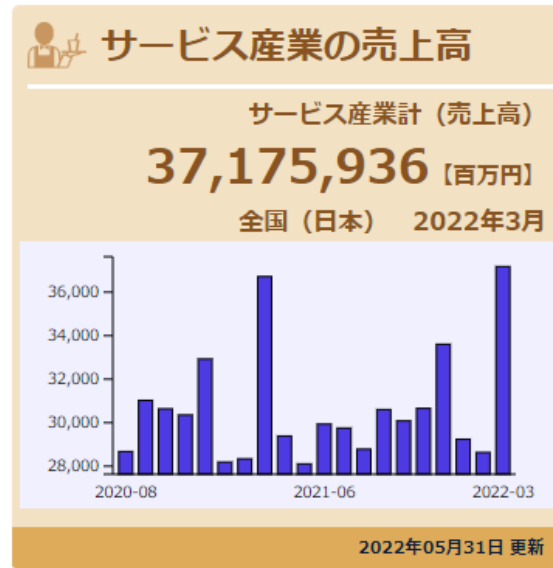
e-Stat 使用例

- 家計調査から、2000年以降の各種月謝の時系列変化を抽出。
- スポーツ月謝はコロナ前まで堅調に増加。他の月謝は長期減少傾向。
- e-Statだけでグラフ化が可能だが、高度なデータ分析用にファイルダウンロードも可能。



総務省統計局 統計ダッシュボード

- 総務省統計局が整備し、独立行政法人統計センターが運営しているサイト。
- 主要な統計データを加工し、グラフ等を用いて見やすく表示。
- ホームページ <https://dashboard.e-stat.go.jp/>



地域経済分析システム RESAS

- 経済産業省地域経済産業グループ地域経済産業調査室・内閣府地方創生推進室が運営しているサイト。
- 産業構造や人口動態、人の流れなどの官民ビッグデータを集約・可視化。
- ホームページ <https://resas.go.jp>



RESAS 使用例 (1/2)

- 横浜市中心部の産業割合を抽出。

地図上で調査地域を指定

The screenshot displays the RESAS web application interface. The main map shows a detailed view of Yokohama City Center, with a yellow rectangular selection box highlighting a specific area. The sidebar on the right contains several filters and options:

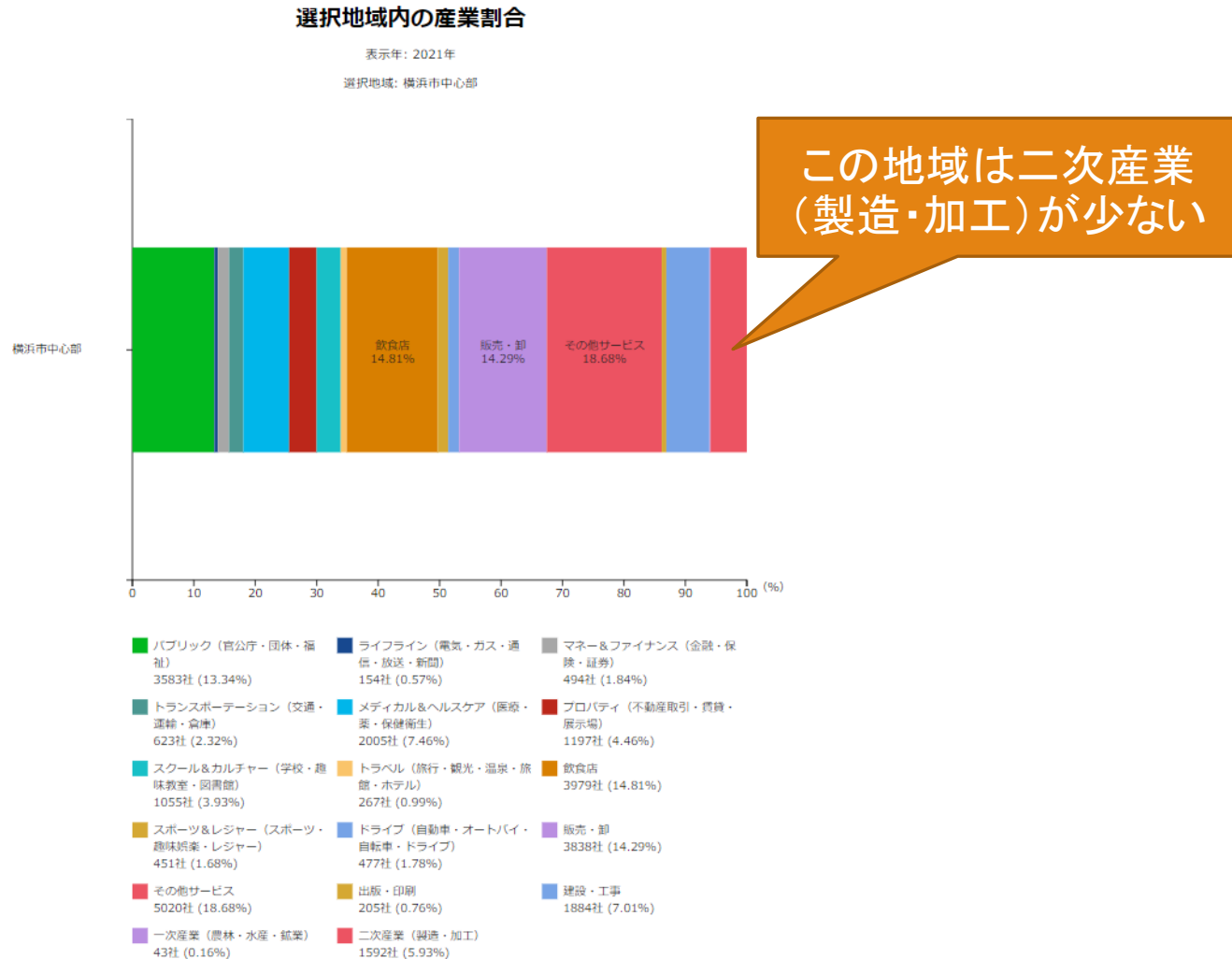
- 表示年を指定する** (Specify display year): 2021年
- 表示産業を指定する** (Specify display industry):
 - すべての大分類 (All major categories)
 - すべての中分類 (All sub-categories)
 - すべての小分類 (All minor categories)
- 事業所情報を読み込む** (Load business information)
- 地域選択モード** (Area selection mode):
 - 任意の地域を選択する (Select any area)
 - 選択地域内の産業割合を見る (View industry ratio in selected area)
 - の推移を見る (View trend)

At the bottom right, there are buttons for **追加する** (Add), **シェア** (Share), and **ツイート** (Tweet). A legend at the bottom left indicates that blue dots represent **単独事業所** (Individual business) and red dots represent **複数事業所** (Multiple businesses).

産業割合をクリック

RESAS 使用例 (2/2)

- 指定地域の産業割合がグラフ表示される。



オープンデータカタログサイト

- 二次利用が可能な公共データの検索サイト。デジタル庁が運営。
- ホームページ <https://dashboard.e-stat.go.jp/>

The screenshot shows the homepage of the DATA GO.JP Open Data Catalog Site. At the top, there is a dark blue header with the logo "DATA GO.JP" and the text "データカタログサイト". Below the header is a navigation bar with links: "お知らせ", "データ", "データベースサイト一覧", "公共データ活用事例", "コミュニケーション", and "開発者向け情報". A search bar is located below the navigation bar, with the placeholder text "データセットを検索...". Below the search bar is a link to the "意見受付コーナー" (Opinion Reception Corner) with the text "にて、オープンデータの掲載に関する御要望等を受け付けております（匿名可）". The main content area is divided into six sections: 1. "データ" (Data) with an illustration of a hand holding a magnifying glass over various file formats (HTML, PDF, XLSX, CSV, XML, JPG) and the text "データセット 組織 グループ タグ". 2. "オープンデータの取組（リンク集）" (Open Data Initiatives (Link Collection)) with an illustration of a computer monitor and a clipboard, and the text "オープンデータに関する方針・決定 公共データ活用事例一覧 データベースサイト一覧". 3. "コミュニケーション" (Communication) with an illustration of hands and speech bubbles, and the text "意見受付コーナー 意見・回答公開コーナー 掲載データ利用の御連絡". 4. "一億総活躍社会の実現" (Achievement of a Society where 100 million people are fully engaged) with the text "子育て支援 安心につながる社会保障 希望を生み出す強い経済". 5. "2020年東京オリンピック・パラリンピック競技大会" (2020 Tokyo Olympic and Paralympic Games) with the text "大会の円滑な準備及び運営 大会を通じた新しい日本の創造". 6. "地方公共団体データベースサイト一覧" (List of Local Government Database Sites) with an illustration of a bar chart.

DATA GO.JP データカタログサイト

お知らせ ▾ データ ▾ データベースサイト一覧 ▾ 公共データ活用事例 ▾ コミュニケーション ▾ 開発者向け情報 ▾

データセットを検索...

[意見受付コーナー](#) にて、オープンデータの掲載に関する御要望等を受け付けております（匿名可）

データ

データセット 組織 グループ タグ

オープンデータの取組（リンク集）

オープンデータに関する方針・決定
公共データ活用事例一覧
データベースサイト一覧

コミュニケーション

意見受付コーナー 意見・回答公開コーナー
掲載データ利用の御連絡

一億総活躍社会の実現

子育て支援 安心につながる社会保障 希望を生み出す強い経済

2020年東京オリンピック・パラリンピック競技大会

大会の円滑な準備及び運営 大会を通じた新しい日本の創造

地方公共団体データベースサイト一覧

オープンデータ利活用事例

- デジタル庁オープンデータ100で紹介している利活用事例。
- ホームページ https://www.digital.go.jp/resources/data_case_study/

No事業名称	概要・テーマ	事業者名	分類
1 アグリノート	農業は、記憶から記録へ	ウォーターセル株式会社	産業創出
2 イーグルバス	センサー乗降データでバス路線の最適化！	イーグルバス株式会社	産業創出
3 カーリル	借りたい一冊、見つかる！	株式会社カーリル	産業創出
4 家計簿・会計アプリ Zaim	もっと、お金に、楽しさを！	株式会社Zaim	防犯・医療・教育等
5 花粉くん	花粉症の日常を、ちょっと新しい体験に。	株式会社 博報堂アイ・スタジオ	防犯・医療・教育等
6 ココゆれ	住宅を建てる前に、地震のリスクがすぐ分かる！	大和ハウス工業株式会社	産業創出
7 さばえぶらり	鯖江の新たな魅力、発見！	ATR Creative	産業創出
8 GEEO（ジュー）	日本国内の不動産の相場感が分かる！	おたに	産業創出
9 周辺環境スカウター	あなたの住みたい街の戦闘力はいくつ？	30min.（株式会社イード）	産業創出
10 全国避難所データベース	いざという時。すべての人に、避難のための情報を	電通・ゼンリンデータコム	防災・減災
11 ミルモ	介護を支える、新しいカタチ	株式会社ウエルモ	少子高齢、産業創出
12 HalexDream!	気象庁が発表するあらゆるデータを活用	株式会社ハレックス	産業創出
13 セーフティマップ	カーナビデータを活用した事故多発箇所、危険箇所の提供	本田技研工業株式会社	防犯・医療・教育等

出典：上記HPより一部抜粋

AI技術で良く使われるオープンデータ(一部抜粋)

- Kaggle
 - 最適なモデル作成を競い合うコンペティションプラットフォーム。題材として様々なデータセットが提供されている。
- MNIST
 - 手書き文字(数字)の画像データセット。
- Fashion-MNIST
 - 10種類に分類されたファッションの画像データセット。
- CIFAR-10/CIFAR-100
 - CIFAR-10は10種類、CIFAR-100は100種類に分類された画像データセット。
- ImageNet
 - 1,400万枚の画像データセット。
- COCO
 - 33万枚のキャプション付き画像データセット。

学び方

総務省統計局 データサイエンス・オンライン講座

- 総務省統計局がデータサイエンス力の高い人材育成を目的として開講。
- 今年度は、次のオンライン講座を開講。いずれも受講料は無料。オンライン講座gaccoに登録するだけで受講可能。
 - 社会人のためのデータサイエンス入門
 - 社会人のためのデータサイエンス演習
 - 誰でも使える統計オープンデータ

「データサイエンス・オンライン講座」2022年度開講スケジュール

2022年3月時点の予定のため、変更の可能性があります。

講座名	2022年											2023年		
	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月		
社会人のための データサイエンス入門			開講予定											
社会人のための データサイエンス演習							開講予定							
誰でも使える 統計オープンデータ										開講予定				

出典：総務省統計局HP

- 今年度の受講タイミングを逃した場合は、総務省統計局ホームページを定期的にウォッチすることで今後の開講を把握できる。（毎年、同様の講座が開講されている）
- 総務省統計局ホームページ <https://www.stat.go.jp/>

無料オンライン講座 gacco

- gaccoは世界で利用されるMOOC(Massive Open Online Courses)の日本版。
- (株)ドコモgaccoが運営。
- gaccoでは、データサイエンスやAIを学べる講座が提供されている。(下記リストは2022/5/23時点)
 - AI活用人材育成講座
 - ✓ AIがビジネス現場で活用されている事例を紹介する事例集
 - ✓ AIの実装に必要な知識を学ぶ講座
 - 社会の中のAI ～人工知能の技術と人間社会の未来展望～
 - 高校生のためのデータサイエンス入門
 - 統計学 I : データ分析の基礎
 - はじめてのAI
 - など
- 講座一覧は随時アップデートされている。
- gaccoホームページ <https://gacco.org/index.html>

データ分析スキルに関わる資格・試験

- 日本統計学会認定 統計検定
 - 統計に関する知識や活用力を評価
 - 統計検定1級、統計検定準1級、統計検定2級、統計検定3級、統計検定4級、統計検定統計調査士、統計検定専門統計調査士、統計検定データサイエンス基礎、統計検定データサイエンス発展、データサイエンスエキスパート
 - ホームページ <https://www.toukei-kentei.jp/>
- 日本ディープラーニング協会
 - ディープラーニングに関する能力や知識を検定
 - E資格(エンジニア資格)、G検定(ジェネラリスト検定)
 - ホームページ <https://www.jdla.org/>

コンペティション Kaggle

- データサイエンスのコンペティションが開催されるプラットフォーム。
- 40万人のデータサイエンティストが集まっており、スキルアップのためにデータ分析の腕試しを行っている。
- 企業はデータと課題をKaggleに提供し、データサイエンティストはそれに対する最適なモデルを競い合う。
- 優秀なモデルを作成したデータサイエンティストは、企業からKaggleを通じて賞金を獲得。
- 企業の立場で見ると、課題とデータを公開して、世界中のデータサイエンティストから優秀なモデルを調達するようなイメージ。
- Googleが提示した課題例
 - 【課題】Python Notebooksのコードとコメントの関連を予測
 - 【コンペ期間】2022/5/11～2022/8/11
 - 【賞金】1位:5万USD～5位:1万USD
- Kaggleホームページ <https://www.kaggle.com/>