さまざまなデータ

広島大学 AI・データイノベーション教育研究センター
稲垣知宏

目標

さまざまなデータとその特徴を理解する。

この授業で紹介すること

- さまざまデータ
- 統計データのポータルサイト

キーワード

調査データ、観測データ、実験データ、ログデータ、1次データ、2次データ

こんなことはありませんか？

待ち合わせの場所に行くのに、普段なら
10分前には着く時間に自宅を出たのに、
渋滞に巻き込まれてしまい遅れそうです。

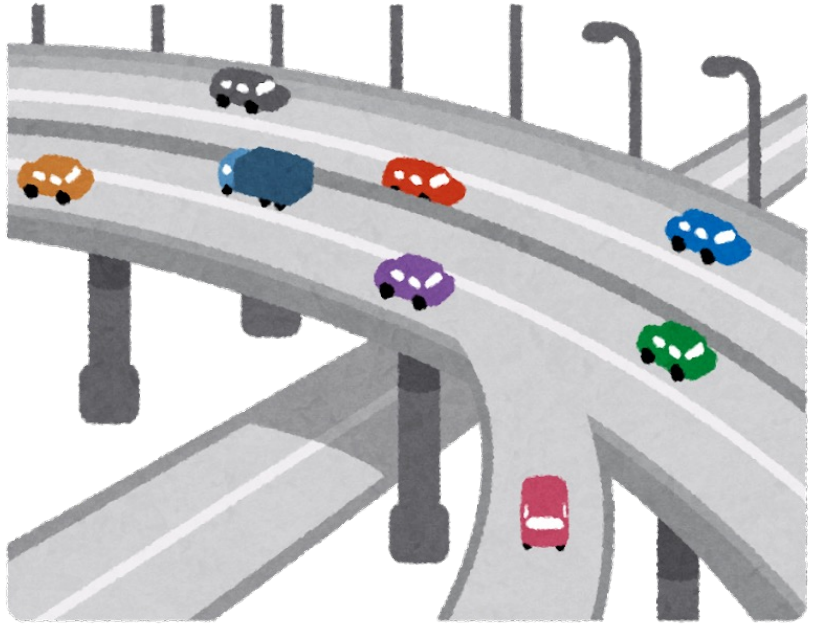
渋滞に巻き込まれないようにするには、
どうすれば良いでしょうか？

渋滞に巻き込まれないために

渋滞に巻き込まれないには、あらかじめ道路情報を確認し、早めに出発する、混雑しそうな経路を避ける、待ち合わせ時間を変更するなど、できそうです。

問題を根本的に解決し、渋滞が起き難くするためには、渋滞の原因を探る必要があります。

渋滞の原因を探るには、どのようなデータが有用でしょうか。



調査データ

渋滞の原因を調べるために、渋滞に巻き込まれている人に尋ねることから始めましょう。

対象の道路をいつも利用している運転手にインタビューすることで、どうのようなときに渋滞が発生するかのヒントが得られるかもしれません。

目的を持って集められたデータを調査データと呼びます。

調査では、誰を対象にするのか、何をどのように聞くのかを検討し、明確にしておくことが重要です。

観測データ

曜日、時間帯、あるいは天候毎に、実際の交通量がどうなっているかを調べましょう。

通勤時間、天候など、渋滞が発生すると思われる状況とそうでない状況での交通量を調べることで、渋滞発生の条件が明らかになるかもしれません。

現象を観測することで得られるデータを観測データと呼びます。

観測では、時期、時間、何をどの様にして観測するのかを検討し、また、記録していくことが重要です。

実験データ

原因と考えられる要素のみを抜き出した模型を用意して、渋滞が起きるかどうかを確認することで、渋滞の原因を確定しましょう。

渋滞発生の原因と条件を再現した実験を行い、その他の条件に関わらず、渋滞が発生すれば、渋滞の原因が判明するでしょう。

実験により得られるデータを実験データと呼びます。

実験では、どの要素を抜き出し、その要素は考慮しなくて良いとしたのか、また、揃えた条件と変更した条件を明確にすることが重要です。なお、計算機上に構築した仮想的な模型で実験することもあります。

ログデータ

道路情報アプリのログ（履歴）を確認することで、渋滞発生時をの情報を集めましょう。

過去の履歴であるログデータと実験結果を付き合わせることで、実験の正当性を示せるかもしれません。

ログを確認することで得られるデータをログデータと呼びます。

道路情報アプリのログは膨大なデータ量になっていることが多いです。ログから有用な情報を得るには、データを抽出するための工夫が必要になります。

身近なデータの例

気象	気温、気圧、湿度、降水量など
医療	検査結果、画像、処方箋記録など
交通	走行履歴、挙動履歴、渋滞状況など
携帯	位置情報、通信履歴、検索履歴など
SNS	呟き、画像、動画など

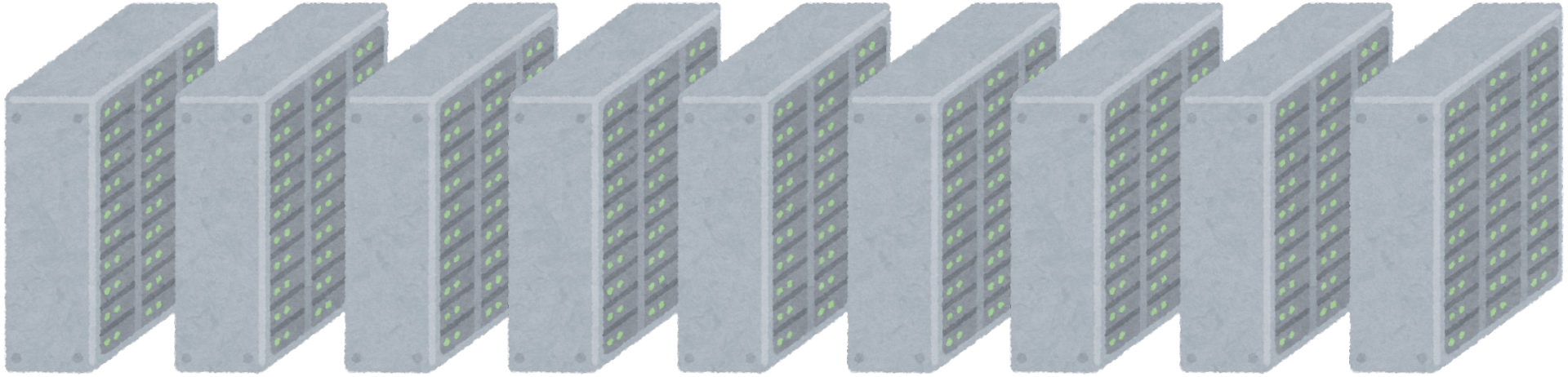
データの利用

気象	天気予報、災害予測など
医療	各種診療、医薬品開発など
交通	到着時刻予測、経路検索、道路整備など
携帯	人口動態把握、感染症対策など
SNS	トレンド予測、商品開発など

データとデータサイエンス

データサイエンスとは、数学、統計学、情報学等を利用してデータから情報を得ることを目的とした学問分野です。

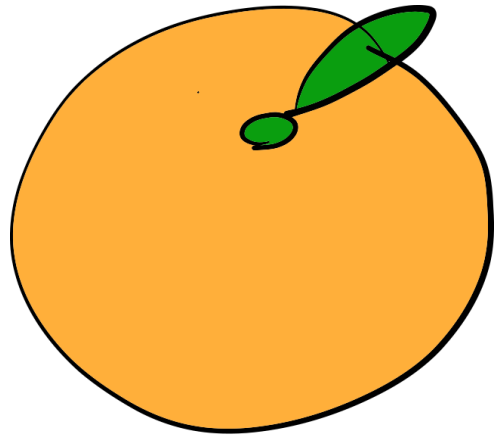
コンピュータとネットワーク、及びセンサー技術の発展と普及により、膨大なデータが蓄積されるようになり、さまざまなデータから情報を得ることの重要性が認識され、注目されています。



例題

Aさんは、大学生が美味しいと感じるみかんを育てたいと考えています。 みかんの美味しさは、みかん畑の降雨量によっていると考えて、降雨量が異なる地方のみかんを集めてきました。また、みかんの美味しさは保存方法によっても変化するのではないかと考えています。

美味しいみかんを育てるには、どのようなデータが有用でしょうか。



解説

さまざまなデータが考えられると思います。例えば、
調査データ：大学生を対象に、美味しいと感じるみかんを調査
観測データ：みかん畑の降雨量を測定したデータを確認
実験データ：保存時の温度などを変えて、糖度の時間変化を測定
などが考えられるでしょうか。

１次データと2次データ

例題の解説で、以下の回答例を示しました。

調査データ：大学生を対象に、美味しいと感じるみかんを調査

観測データ：みかん畑の降雨量を測定したデータを確認

実験データ：保存時の温度などを変えて、糖度の時間変化を測定

ここで、調査データと実験データはAさんが集めると想定していますが、降雨量はAさんが測定するとは考えていません。

調査目的に合わせて自分たちで集めたデータを１次データ、他の目的で集められたデータや他者が公開しているデータを２次データと呼びます。

統計データのポータルサイト

２次データとして利用できるデータはいろいろな形で、公開、販売されています。

規模の大きいものとしては、政府統計のポータルサイト

e-Stat：https://www.e-stat.go.jp/

があります。

政府統計の総合窓口(e-Stat)
（https://www.e-stat.go.jp/）

広島の統計情報

地方自治体等が公開している地方毎のデータもあります。

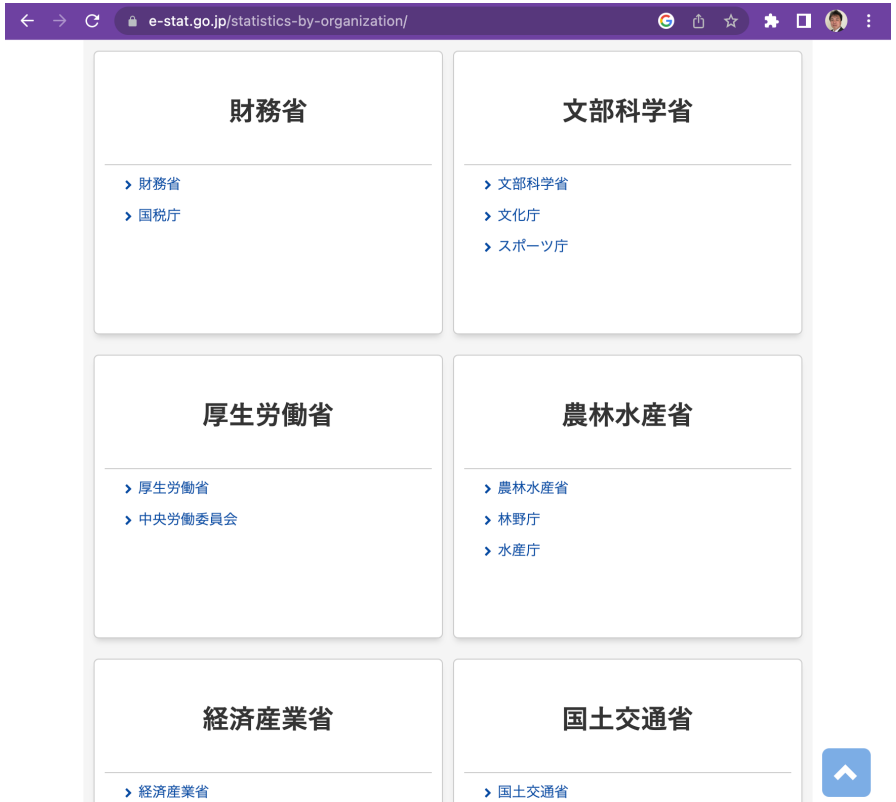
広島の場合、広島広域都市圏と広島県のオープンデータポータルサイト

https://hiroshima-opendata.dataeye.jp/

があります。

データの分類

e-Statのデータは、分野、組織毎にまとめられています。



メタデータ

各データに付けられた付帯情報のデータをメタデータと呼びます。

e-Statで、学校基本調査を検索し、令和4年度の高等教育機関の総括データのページを探してみましょう。

付帯情報として政府統計名、提供統計名、提供分類１－４といったメタデータが付けられていることが分かります。

問題

1. e-Statから「青少年のインターネット利用環境実態調査」の最新の結果をダウンロードして、青少年のインターネット利用率を確認しなさい。
2. 内閣府のページで、「青少年のインターネット利用環境実態調査」の調査対象と調査内容（調査票）を確認しなさい。