1-2　社会で活用されているデータ

東京大学 数理・情報教育研究センター
2020年5月11日

概要

- それではどういうデータがデータサイエンスでは用いられるのでしょうか？

- ここでは代表的なデータをいくつか見ていくことで、どういうデータが集められ、どう活用されているかを知ることを目標とします

本教材の目次

1-2-1　データ元の種類

調査（リサーチ）データ

- 研究やマーケティングなどで明確な意図をもって集められたデータのことです
  - 政府統計、マーケティング、財務状況、アンケートなどがあげられます

国勢調査
https://www.stat.go.jp/data/kokusei/2015/
で見ることができます

日銀短観
https://www.boj.or.jp/statistics/tk/index.htm/
で見ることができます

e-Stat：https://www.e-stat.go.jp/

大規模調査

- 近年ではインターネットをうまく使い大規模なアンケートを取る研究者もいます
  - Moral Machine：自動運転時に生じるかもしれないトロッコ問題を多くの人間　に生成・判断してもらう試みで、世界中から4千万人が参加しました[Awad et al.2018]
    （http://moralmachine.mit.edu/hl/ja）

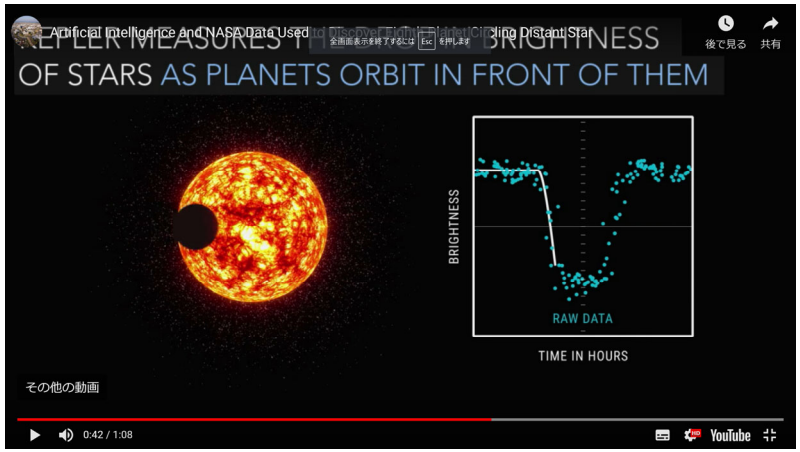
- Moral MachineはAIに道徳的な意思決定をさせるにはどうしたらよいかという問いに挑戦した野心的な研究でもあります

観測データ

- 天体観測や気象観測などのように探査機や気象レーダーを活用してデータを集めることもあります
  - このように現象を観測して得たデータを観測データと呼びます

- 天体観測の例は「1-1 社会におけるデータ・AI利活用」の物理学の所でも紹介しました

- 気象観測では気象レーダー、静止気象衛星、地上・地域気象観測など多面的にデータを収集します



Video credit: NASA/Ames Research Center

気象観測の概要は気象庁のホームページで確認することができます（jma.go.jp/jma/kishou/know/kansoku/weather_obs.html）

実験データ

- 原因の効果を測定するために使われます
  - ある原因の効果を測定するために、その他の条件は同じにしたサンプルを作成し比較します

- インターネット環境は実験状況を作りやすいため様々な実験が行われています
  - Facebook：オンライン広告の比較実験[Gordon et al.2018]
  - A/Bテストとも呼ばれます

- 社会科学では偶然実験状況になっているデータを分析することもあります。これを自然実験と呼びます
  - 自然実験の例：オランダ飢饉の影響 (http://economicspsychologypolicy.blogspot.com/2015/06/list-of-19-natural-experiments.html)

行動ログデータ

- インターネット行動ログやGPSデータはマーケティングで多用されています

- Valuesではインターネット行動ログ分析サービスを提供しています
  - 検索履歴などの行動ログをベースにマーケティングに有用な情報を提供しています
    （https://www.valuesccg.com/service/dmd/emarkplus/）

- ブログウォッチャーではGPSデータに基づいた商圏分析を行っています（https://www.blogwatcher.co.jp/）

マシンログデータ

- 機械やサイバーセキュリティーの現場ではマシンログデータがよく活用されます
  - 例えば自動車には大量のセンサーが設置され、制御やモニタリングに使われています（自動運転技術にも関連します）

- 他にも日本製鉄では製鉄工場における先進計測技術と制御技術を活用し製鉄作業の管理をしています[吉沢,中川2018]

- サイバーセキュリティの分野ではマシンログの異常を検知することで攻撃活動を監視する試みが盛んです[NICT2018]

サイバーセキュリティー
における攻撃活動の例

1	systeminfo
2	whoami
3	netview
4	dir c:¥users¥rh¥desktop
5	dir
6	dir d:¥
7	netstat -an
8	ipconfig /all
9	whoami /groups
10	net view
11	ping 8.8.8.8
12	tasklist

1-2-2　データの所有者

1次データ、2次データ、メタデータ

- 1次データとは自社データのことを指します
  - 自社内の業務の中で集めたデータや調査目的に従って集めたものです
  - 調査の自由度が高く競合他社は手に入りません

- 2次データとは外部のデータのことを指します
  - 気象情報や政府が公開する統計などがここに入ります
  - 調査の自由度は低く競合他社も手に入りますが、企業レベルで手に入らない情報を含んでいる可能性があります

- メタデータとはデータに付随している情報（作成者、作成日時等）を指します
  - データ管理の際に役立てます

オープンデータ

- オープンデータとは広く社会に利用してもらうことを目的に公開されたデータのことです

- 自社データは競合他社が利用できない分強みになりえます
  - あまりにも強みになるためデータ独占を是正しなければいけないという指摘もあります[Economist2017]

（https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data）

- そのためオープンデータに関する様々な試みが行われています
  - 日本の公共データに関しては次のサイトがあります

（https://www.data.go.jp/）
  - 他にも農業、生物、地球科学、経済、物理など多岐にわたって様々なデータが公開されています

（https://github.com/awesomedata/awesome-public-datasets）

1-2-3　構造化データと非構造化データ

構造化データ

- 事前に定めたデータモデル（一定の規則・構造に従って記述されるものです）で管理できるものを指します
  - SQLと呼ばれる問い合わせ言語を使用します

構造化データの例

職員			
ID	氏名	住所	メール
123	多摩川渡	東京都　立川	***1@hit.org
124	玉剛	東京都　新宿区	***2@hit.org
125	荻窪勝彦	神奈川県	***3@hit.org
126	丹生義弘	東京都　日野市	***4@hit.org

メール送信履歴		
送り手	受け手	日時
***1@hit.org	***2@hit.org	2043/2/1
***1@hit.org	***3@hit.org	2043/2/1
***1@hit.org	***4@hit.org	2043/2/1
***2@hit.org	***4@hit.org	2043/2/1

注文					
ID	日付	メニュー	金額	個数	総計
123	2043/2/3	かつ丼	520	1	520
123	2043/2/3	コーラ	160	1	160
123	2043/2/3	杏仁豆腐	220	1	220
125	2043/2/3	カツカレー	630	1	630
125	2043/2/3	激辛	100	1	100

非構造化データ

- テキストや画像、音声、動画は必ずしも事前に定めたデータモデルで管理することが困難なことがあります
  - 非構造化データと呼ばれます

- 例えば文章はセンチメント（肯定的か否定的か）分析をすることで数字に変換し分析することがあります
  （https://guides.lib.uoguelph.ca/TextSentimentVisualizer）

- 衛星写真（夜間に撮影すると光量で経済活動が活発かどうかわかります）を用いて貧困地域のGDPを推定することもあります
  （https://news.stanford.edu/2016/08/18/combining-satellite-data-machine-learning-to-map-poverty/）

ビッグデータとアノテーション

- 文章や画像などのデータは人間が見ればすぐにラベルづけ（アノテーション）できるものが多いです
  - 「あの映画はつまらなかった」＝負のセンチメントだとわかります
  - 右下は画像のアノテーション例

- こうしたラベルを全ての文書や画像につけるのは困難かつ高価です（ラベルづけにはAmazon Mechanical Turkなど業者を使うこともあります https://aws.amazon.com/jp/mturk/faqs/）

- そのため限られたラベルつきの訓練データを有効活用することが鍵になります

「夕暮れ時の街並みの写真」

1-2-4　参考文献

参考文献

[Awad et al.2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Friedemann Schulz, "The Moral Machine Experiment", Nature 563(7729) , November 2018.

[Economist2017] The Economist, May 6th 2017, "Regulating the internet giants The world’s most valuable resource is no longer oil, but data The data economy demands a new approach to antitrust rules"

[Gordon et al.2018] Gordon B.R. and Zettelmeyer F., “A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Face book”, https://www.kellogg.northwestern.edu/faculty/gordon_b/files/fb_comparison.pdf

[NICT2018] NICT NEWS 2018 No.6 サイバーセキュリティの研究開発最前線

[吉沢,中川2018]吉沢一郎、中川繁政、製鉄設備におけるシステム・計測制御技術の進歩と展望、新日鉄住金技法第411号、2018