@島根大学人間科学部（オンライン）
2020.10.03

インターネットを使った研究データの探し方

二次分析とWebデータ収集

明治大学　研究・知財戦略機構
佐藤浩輔

Motivation

- 新型コロナウイルス(COVID-19)の社会的影響
  - 人と人との接触が制限される
  - 人を対象にした研究が制約される
    - 心理学実験
    - インタビュー等対面を必要とする調査

本セミナー：
インターネットを介したデータ収集について紹介

構成

1. 二次分析とWebデータ収集の概要
   - 二次資料・Webデータを利用した研究
   - ケーススタディ（時間があれば）
2. Web上のデータを使う際の注意点
   - データの特徴
   - 倫理
   - 技術的な話題
3. データ収集のデモ
   - Twitter APIを使ったデータ取得
   - wgetを使ったファイル取得
   - JSON/XML/HTMLの読み込み

- このセミナーで**扱う**こと
  - Web上で手に入るデータについて：
    - どのような種類があるか
    - どのように入手すればよいか
    - 扱う際の注意点はなにか
- このセミナーで**扱わない**こと
  - Web上での実験
  - オンライン調査
  - 分析手法の詳細(e.g. メタアナリシス)

※7月の「応用心理学Ⅰ」佐藤（浩）担当分をあわせて読むと理解が深まります
https://github.com/satocos135/lecture2020shimane

データをどう手に入れるか

①オンラインで
- オンライン実験
- Web調査/インターネットモニター
- メール・テレビ通話等によるインタビュー

⇒今回は扱わない

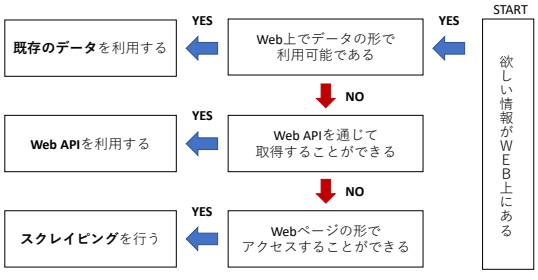
②既存のデータを分析する（二次分析）
- **既存の資料**を使う
- **オープンデータ**を使う

③Web上のデータを収集する
- **Web API**を用いてSNS上のデータを取得する
- Webページを**スクレイピング**する

④自分で生成する
- コンピュータシミュレーションを行う

⇒今回は扱わない

Web上のデータを入手するフロー



既存のデータの活用（二次分析）

一次データと二次データ

- 一次データ：研究者自身が集めたデータ
- 二次データ：研究者**自身ではない**主体が集めたデータ
  - 官公庁などの主体が収集する統計
  - 文書・書籍等の資料
  - 他の研究者等が公開しているデータ
    - 論文の実験・調査データ、あるいは論文そのもの
    - 研究用データセット
    - 各種アーカイブ、データベース

二次データを使うと何が嬉しいか

- Pros
  - 自分で集めていないので：
    - 個人では収集できないようなデータが使える
      - 地域や規模 e.g. 海外のデータ、国レベルのデータ
      - 専門性
    - 収集のコストを払わなくてよい
- Cons
  - 自分で集めていないせいで：
    - 自分の研究関心と合致した指標が取れているとは限らない
    - データ収集プロセスの詳細が不明瞭
      - 欲しい解像度でデータが取れているとは限らない
      - データの制約を理解しないと落とし穴にはまる
        - e.g. 基準が途中で変更された

Web上の二次データの例

- 各種オープンデータ
  - 国・地方公共団体・官公庁などのオープンデータ

DATA GO JP https://www.data.go.jp/

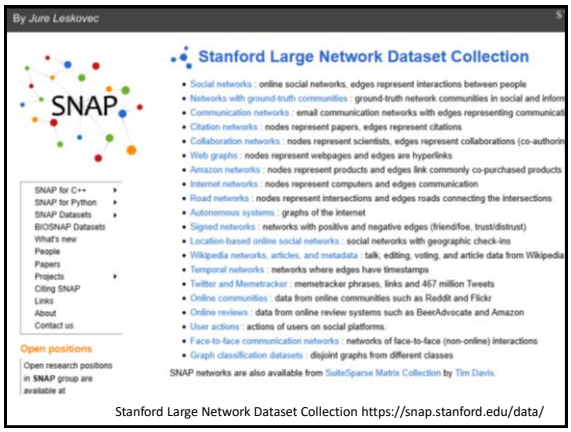
名称	組織名	ライセンス	API	主な分類	概要	更新日
大鰐町ホームページ	青森県大鰐町	CC-BY	無	人口・世帯、住民向け情報（暮らしの情報）、生活、子育て、健康、医療、農林水産業、観光情報、教育・文化・スポーツ、統計、その他	大鰐町では、オープンデータを公開しています。積次各種公共データをオープンデータとして公開していきますのでご活用ください。	2020/09/09
身延町オープンデータライブラリ	山梨県身延町	CC-BY	無	防災	身延町が保有する公共データを自由に利活用できるオープンデータとして随時公開しています。	2020/09/09
海陽町オープンデータ（徳島県オープンデータポータルサイト Our Open Data）	徳島県海陽町	CC-BY	無	交通、ごみ・環境保全	徳島県内各自治体のオープンデータを公開するサイト「徳島県オープンデータポータルサイト Our Open Data」に海陽町のオープンデータを公開しています。	2020/09/09
西海市オープンデータカタログサイト	長崎県西海市	CC-BY	無	観光、教育・文化・スポーツ、防災・安全・環境、その他	このサイトでは、誰でも自由に西海市が現在公開しているデータを検索したり、ダウンロードしたりすることができます。	2020/09/09

政府統計の総合窓口 https://www.e-stat.go.jp/

Web上の二次データの例

- 各種オープンデータ
  - 国・地方公共団体・官公庁のオープンデータ
- 研究用データセット
  - 各種研究用データ
  - 各種言語資料（コーパス）

WVS Database http://www.worldvaluessurvey.org/wvs.jsp



Stanford Large Network Dataset Collection https://snap.stanford.edu/data/

Social networks

Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter
soc-Epinions1	Directed	75,879	508,837	Who-trusts-whom network of Epinions.com
soc-LiveJournal1	Directed	4,847,571	68,993,773	LiveJournal online social network
soc-Pokec	Directed	1,632,803	30,622,564	Pokec online social network
soc-Slashdot0811	Directed	77,360	905,468	Slashdot social network from November 2008
soc-Slashdot0922	Directed	82,168	948,464	Slashdot social network from February 2009
wiki-Vote	Directed	7,115	103,689	Wikipedia who-votes-on-whom network
wiki-RfA	Directed, Signed	10,835	159,388	Wikipedia Requests for Adminship (with text)
gemsec-Deezer	Undirected	143,884	846,915	Gemsec Deezer dataset
gemsec-Facebook	Undirected	134,833	1,380,293	Gemsec Facebook dataset
soc-RedditHyperlinks	Directed, Signed, Temporal, Attributed	55,863	858,490	Hyperlinks between subreddits on Reddit
soc-sign-bitcoin-otc	Weighted, Signed, Directed, Temporal	5,881	35,592	Bitcoin OTC web of trust network
soc-sign-bitcoin-alpha	Weighted, Signed, Directed, Temporal	3,783	24,186	Bitcoin Alpha web of trust network
comm-f2f-Resistance	Weighted, Directed, Temporal	451	3,126,993	Dynamic face-to-face interaction network between group of people
musae-twitch	Undirected	34,118	429,113	Social networks of Twitch users.
musae-facebook	Undirected	22,470	171,002	Facebook page-page network with page names.
act-mooc	Bipartite, Directed, Attributed, Temporal	7,143	411,749	Student actions on a MOOC platform, with student drop-out binary labels.

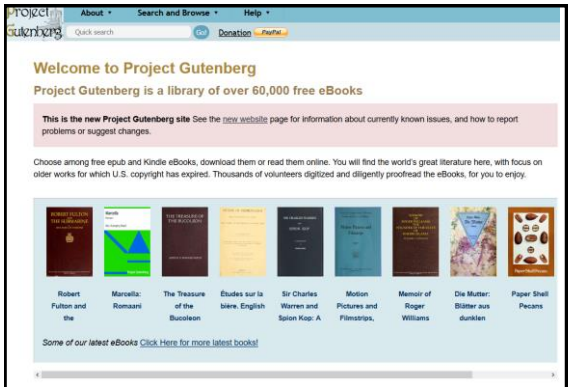
情報学研究データリポジトリ https://www.nii.ac.jp/dsc/idr/

情報学研究データリポジトリで提供されているデータ（一部）

企業提供データ	Yahoo!データセット	Yahoo!知恵袋データ（第3版）
	楽天データセット	楽天市場の全商品データ，レビューデータ
		楽天トラベルの施設データ，レビューデータ
		楽天GORAのゴルフ場データ，レビューデータ
		楽天レシピのレシピ情報，レシピ画像
		アノテーション付きデータ
	ニコニコデータセット	ニコニコ動画コメント等データ
		ニコニコ大百科データ
	リクルートデータセット	ホットペッパービューティーデータ
	クックパッドデータセット	レシピデータ，献立データ
	LIFULL HOME'Sデータセット	賃貸物件スナップショットデータ
		高精細度間取り図画像データ
		賃貸・売買物件月次データ
	不満調査データセット	投稿された不満データ，ユーザ情報
		カテゴリ別不満特徴語辞書
	Sansanデータセット	サンプル名刺データ
	インテージデータセット	インテージパネルデータ
	オリコンデータセット	顧客満足度調査データ
	ダイエット口コミデータセット	ダイエット商品口コミデータ
	弁護士ドットコムデータセット	法律相談データ
研究者提供データ	グループコミュニケーションコーパス	
	立命館ARC所蔵浮世絵データベース	
	理研記述問題採点データセット	
	大阪大学 マルチモーダル対話コーパス	

Web上の二次データの例

- 各種オープンデータ
  - 国・地方公共団体・官公庁のオープンデータ
- 研究用データセット
  - 各種研究用データ
  - 各種言語資料（コーパス）
- 各種アーカイブ
  - Project Gutenberg
  - 青空文庫
  - Wikipedia



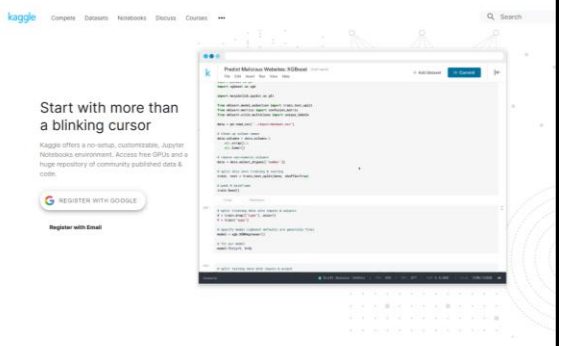
Project Gutenberg https://www.gutenberg.org/

青空文庫 https://www.aozora.gr.jp/

Wikimedia Downloads | https://dumps.wikimedia.org/

データを見つける

- Kaggle
  - 投稿されたデータに対して、分析・予測モデルの性能を競い合うプラットフォーム。多様なデータが公開されている
- Google Dataset Search
  - Web上で利用可能なデータセットの検索エンジン
- Social Science Japan Data Archive(SSJDA)
  - 日本国内で行われた統計調査・社会調査の個票データを収集・保管し二次利用に提供している
  - 学部生でも利用できるが申請が必要

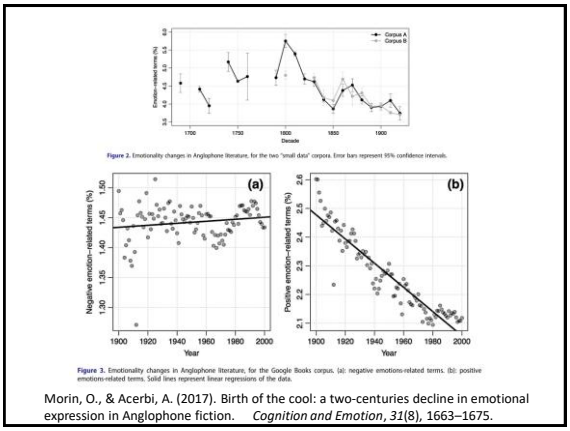


https://www.kaggle.com/

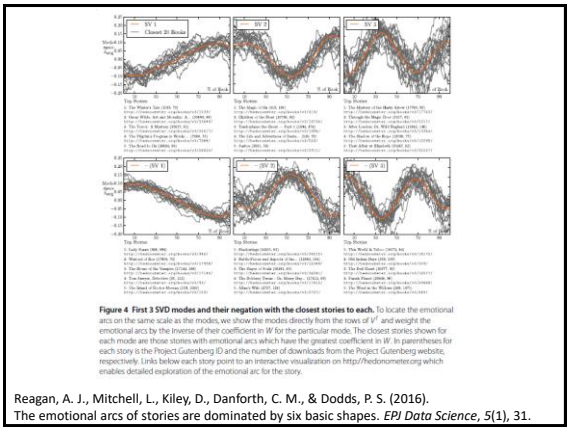
https://datasetsearch.research.google.com/

https://csrda.iss.u-tokyo.ac.jp/

二次データや資料を用いた研究の例



Morin, O., & Acerbi, A. (2017). Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. *Cognition and Emotion*, *31*(8), 1663–1675.



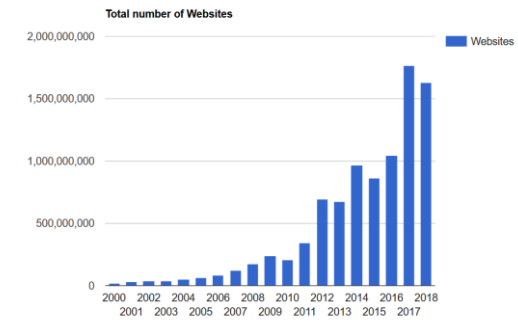
Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016).
The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*(1), 31.

Webデータの収集

インターネットとWeb

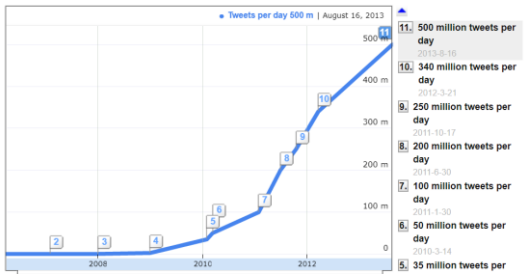
- インターネット：
  - 世界中の情報機器を接続するネットワーク
  - Webや電子メール、P2Pなどの様々なシステムの土台となる
- Web(World Wide Web: WWW)：
  - インターネット上での文書の公開・閲覧システム。ハイパーリンクによって様々な資源が結び付けられている
  - HTTPというプロトコルを用いて通信する
  - Webブラウザを用いて閲覧できる

インターネット上のWebサイトの数



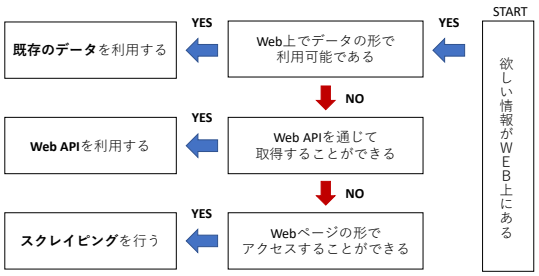
https://www.internetlivestats.com/total-number-of-websites/

Twitterにおける一日あたりのtweetの数



https://www.internetlivestats.com/twitter-statistics

Web上のデータを入手するフロー



Web APIを用いたデータ収集

- API: Application Programmable Interface
  - 他のプログラムからアクセスするために提供されているツール群
- Webサービスの中には(Web) APIを通じて様々な情報を取得できるものがある
  - Twitter
  - Instagram
  - Facebook
  - ....

Twitter APIの例(Twitter API v2)

- Sampled Stream
  `https://api.twitter.com/2/tweets/sample/stream`
  - 全ツイートの1%をリアルタイムにサンプリングして取得するAPI
- Filtered Stream
  `https://api.twitter.com/2/tweets/search/stream`
  - 事前に設定した条件に則ってフィルタした結果をリアルタイムに取得するAPI

HTTPとURI

- HTTP(HyperText Transfer Protocol)
  - クライアントとWebサーバが通信するためのプロトコル(取り決め)
  - クライアントは**URI**(URL)を用いてリソースを指定し、**メソッド**を指定して要求を送る
    - URI(Uniform Resource Identifier)
      - リソースの場所を指定する識別子
      - いわゆる「Webアドレス」
      - エンドポイントと呼ぶ
    - メソッド
      - GET: サーバからリソースに関するデータを受け取る
      - POST: ブラウザからサーバにデータを送信する
      - PUT: リソースに関するデータを変更する
  - クライアントからの要求（**リクエスト**）に対してサーバからの応答（**レスポンス**）がある

APIリクエスト：パラメータ

- APIにパラメータを指定できるものがある
- 例：ユーザーローカルWikipedia API
  http://wikipedia.simpleapi.net/
  - キーワードを指定すると、その言葉に関するwikipediaの記述のダイジェストを返す簡易API
  - APIのパラメータ
    - keyword: 検索語
    - output: 出力データ形式
  - 出力形式JSONで「YouTube」を検索するとすると：

`http://wikipedia.simpleapi.net/api?keyword=YouTube&output=json`

↑ここがパラメータ

`http://wikipedia.simpleapi.net/api?keyword=YouTube&output=json`


```
[
 - {
      language: "ja",
      id: "502914",
      url: "http://wikipedia.simpleapi.net/ja/502914/",
      title: "YouTube",
      body: "YouTube（ユーチューブ）は、アメリカ合衆国・カリフォルニア州サンブルーノのYouTube, LLCが運営する動画共
      length: 16448,
      redirect: 0,
      strict: "1",
      datetime: "2016-04-30T15:02:49+09:00"
  },
 - {
      language: "ja",
      id: "548286",
      url: "http://wikipedia.simpleapi.net/ja/548286/",
      title: "YouTube板",
      body: "ネット関係（ネットかんけい）は、匿名掲示板2ちゃんねるのカテゴリの1つである。インターネットおよびネット
      length: 3858,
      redirect: 0,
      strict: "0",
      datetime: "2016-02-11T23:20:30+09:00"
  },
```


`http://wikipedia.simpleapi.net/api?keyword=YouTube&output=xml`


```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <result>
    <language>ja</language>
    <id>502914</id>
    <url>http://wikipedia.simpleapi.net/ja/502914/</url>
    <title>YouTube</title>
    <body>YouTube（ユーチューブ）は、アメリカ合衆国・カリフォルニア州サンブルーノのYouTube, LLCが
    <length>16448</length>
    <redirect>0</redirect>
    <strict>1</strict>
    <datetime>2016-04-30T15:02:49+09:00</datetime>
  </result>
  <result>
    <language>ja</language>
    <id>548286</id>
    <url>http://wikipedia.simpleapi.net/ja/548286/</url>
    <title>YouTube板</title>
    <body>ネット関係（ネットかんけい）は、匿名掲示板2ちゃんねるのカテゴリの1つである。インターネッ
    <length>3858</length>
    <redirect>0</redirect>
    <strict>0</strict>
    <datetime>2016-02-11T23:20:30+09:00</datetime>
  </result>
  <result>
    <language>ja</language>
    <id>2980412</id>
    <url>http://wikipedia.simpleapi.net/ja/2980412/</url>
    <title>YouTuber</title>
```


Web APIを使う流れ

①使いたいAPIを見つける
②必要な情報を調べる
- サービスによっては利用登録が必要
- 単位時間あたりのアクセス上限（レートリミット）が設定されていることもある

③データを取得する
- 使いたいAPIにリクエストを送り、レスポンスを結果として受け取る

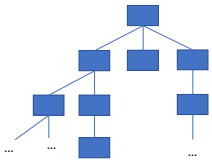
④加工して情報を抽出する
- そのままでは分析に扱いづらいので必要な部分だけ抽出する

APIのレスポンス:JSON

- JSON(JavaScript Object Notation)
  - テキストベースのデータ記述フォーマット
  - キー：値のペアの形式
    - 波括弧({})を使う
    - e.g. {"key1": "value1", "key2": "value2"}
  - リストを使って列挙することもできる
    - 角括弧([])を使う
    - e.g. {"key1": ["value1", "value2"]}
  - 要素を入れ子状にすることもできる
    - e.g. {"key1": {"key2": {"key3": "value1"}}}

APIのレスポンス: XML

- XML(EXtensible Markup Language)
  - テキストベースの汎用的なデータ記述言語
  - 山括弧(<>)で囲んだタグによって構造化されて記述される
    - 開始タグと終了タグによって要素を挟む
    - e.g. <tag>…</tag>
  - 木構造になっている



テキストデータの収集法

Webスクレイピング

- スクレイピング(scraping: こそげ落とす)
  - Webページを取得し、意味のある情報を抽出する
    - すべてのWebサービスにAPIが用意されているわけではない →Webページを直接ダウンロード・加工してデータに
  - Webページ：HTMLで記述されている
    - 様々なツールを用いて情報を抽出
      - 手動でダウンロードする
      - クローラなどを使って自動的にダウンロードする
        - クローラ(crawler): Webのデータを自動的に取得するプログラム。スパイダー(spider), ボット(bot)とも

テキストデータの収集法

データ取得時の注意

- 公式の取得法があればそれを使う
  - e.g. Wikipediaや青空文庫ではダウンロード用にダンプファイルを公開をしている
- 取得先に過度の負荷をかけないようにする
  - 短期間・高頻度にアクセスすると攻撃とみなされるかもしれない
  - 慣習的にはアクセス間隔を1秒以上あける
  - クローラによるアクセスを禁止しているサイトもあるので注意する

HTML

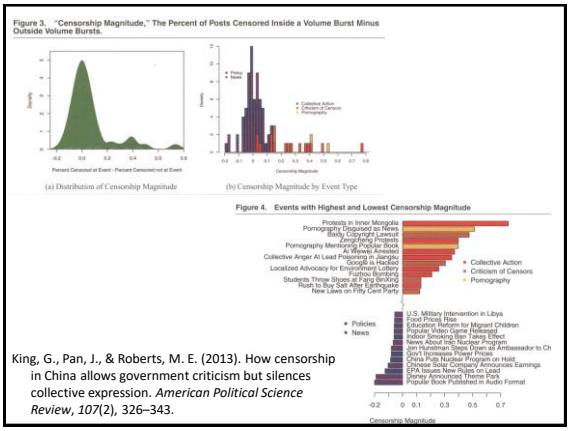
- HTML(HyperText Markup Language)
  - Webページを記述するための言語
  - 要素をタグで囲むことで構造化されている
    - 仕組みとしてはXMLとほぼ同じ
  - 木構造になっている

```
<!DOCTYPE html>
<html lang="ja">
 <head>
  <meta charset="UTF-8">
  <link rel="author" href="mailto:mail@example.com">
  <title lang="en">HyperText Markup Language - Wikipedia</title>
 </head>
 <body>
  <article>
   <h1 lang="en">HyperText Markup Language</h1>
   <p>HTMLは、<a href="http://ja.wikipedia.org/wiki/SGML">SGML</a>
      アプリケーションの一つで、ハイパーテキストを利用してワールド
      ワイドウェブ上で情報を発信するために作られ、
      ワールドワイドウェブの<strong>基幹的役割</strong>をなしている。
      情報を発信するための文書構造を定義するために使われ、
      ある程度機械が理解可能な言語で、
      写真の埋め込みや、フォームの作成、
      ハイパーテキストによるHTML間の連携が可能である。</p>
  </article>
 </body>
</html>
```

https://ja.wikipedia.org/wiki/HyperText_Markup_Language

Webデータを使った研究の例

- King, Pan & Roberts(2013)
  - 中国のオンライン検閲に関する研究
    - 1000以上のソーシャルメディアをクローリング
    - 投稿内容および投稿後削除されたかチェック
    - 批判や賛同といった国家への支持の違いを考慮するため、投稿のセンチメントを機械学習を用いてラベル付け
  - 結果
    - 検閲と国家への支持は無関係
    - ポルノ・検閲批判・集合行動に関する投稿が検閲されやすいことを明らかに



King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343.

Webで手に入るデータの注意点

- ビッグデータの10の特徴(Salganik, 2017)
  - 研究にとって有益な特徴
    - 巨大さ
    - 常時オン
    - 非反応性
  - 問題となる特徴
    - 不完全性
    - アクセス不能性
    - 非代表性
    - ドリフト
    - アルゴリズムによる交絡
    - 汚染
    - センシティブ

Salganik (2017) *Bit by bit*

ビッグデータの特徴①

巨大さ

- データセットが巨大であること有益な研究
  - まれなできごとの研究
  - 不均質性 heterogenityの研究
    - 実験の処理の効果の違い
    - 地域の特性の違い(e.g. Chetty et al., 2014)
  - 微小な差異の検出
    - 1%の差異が意味を持つ分野もある
- 落とし穴
  - 系統誤差に注意しなければならない
    - 系統誤差：偶然でないデータの偏り
    - 偶然誤差は減っても系統誤差は減らない

ビッグデータの特徴②

常時オン

- 絶えずデータを収集
  - 時系列データを取ることができる
  - 予期せぬ出来事の研究ができる
    - 歴史的なできごと・事件
    - アラブの春
    - 9.11テロ
  - リアルタイム推定が可能になる

ビッグデータの特徴③
非反応性

- 社会科学における「反応性 reactivity」
  - 人は観察されると行動を変える(Webb, 1966)
  - 実験者効果
- オンラインのデータ
  - データをとられることを人々が通常意識していないという意味で、非反応的
  - 落とし穴
    - 非反応的であるからといって、**人々のそのままの態度や行動を表しているわけではない**
    - 社会的望ましさなどといった要因の影響はなお残る

ビッグデータの特徴④
不完全性

- 欲しい情報が入っているとは限らない
- 研究上の構成概念と対応するか
  - 構成概念妥当性：測りたいものを測れている？

ビッグデータの特徴⑤
アクセス不能性

- データが存在しても研究者がアクセスできるとは限らない
  - 政府や自治体、企業の中にあるデータ

ビッグデータの特徴⑥
非代表性

- ビッグデータの多くは非代表的
  →母集団を代表してはいない
  - 研究結果をどの程度一般化できるか？

ビッグデータの特徴⑦
ドリフト

- ドリフト(浮動)
  - 時間にともなうシステムの変化
    - どのようなシステムか
    - 誰が使うのか
    - どのように使うのか

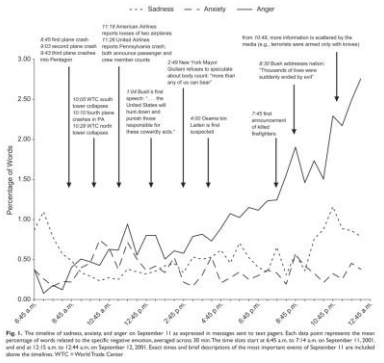
ビッグデータの特徴⑧
アルゴリズムによる交絡

- システム上の行動：人間のありのままの行動ではない
  - システム設計者の企図によって人工的な結果(artifact)が生じる
  - Ugander(2011): Facebookにおけるネットワーク
    - 友達の人数は「20」が突出して多い
      - 友人を20人になるまで増やすようシステムがうながす仕組みがある
    - 「友達の友達」同士は友達になりやすい
      - 社会ネットワークにおいては推移性 transitivity として知られる現象
      - 社会理論を知っている設計者がシステムに理論を組み込んでいる(遂行性 performativity)

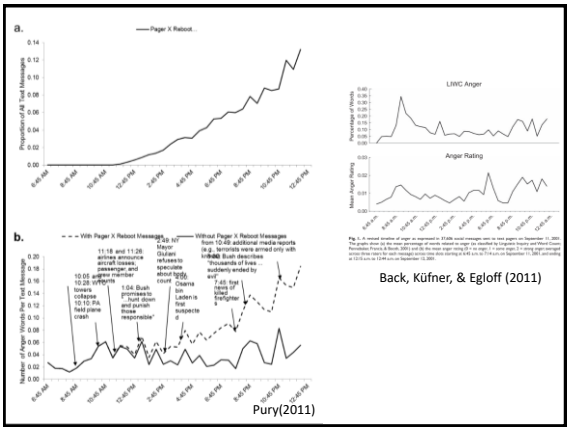
ビッグデータの特徴⑨

汚染

- スパムやボットなど、人間の行動を反映しないデータが紛れ込んでいる
  - Back, Küfner, & Egloff(2010): 9.11後のSNS上のメッセージを分析
    →「9.11後に怒りの感情がSNS上で増加している」
  - Pury(2011):「Backらの結果は誤り」
    - Backらの結果はBotの仕業
    - Botの投稿を取り除くとBackらの結果は再現されない
      →人工的結果(artifact)
  - その後、Backら自身の再集計後の分析でも結果は再現されず(Back, Küfner, & Egloff, 2011)



Back, Küfner, & Egloff(2010)



Pury(2011)

Back, Küfner, & Egloff (2011)

ビッグデータの特徴⑩

センシティブ

- 個人のセンシティブな情報が含まれる
  - 複数のデータをつなげることで個人が特定できてしまうかもしれない

Webデータに関わる倫理的事項

- データはネットに「落ちている」わけではない
  - 公開したり投稿した人がその背後にいる
    →被害が生じないようにする配慮が必要
- Webデータの取り扱い
  - 統一的なガイドラインがあるわけではない
    - 学問分野によっても取り扱いが異なる
  - 「Web上で公開されている」からといって好きに使ってよいというわけではない

ケース①

2017年に立命館大学の研究者らが「ドメインにより意味が変化する単語に着目した猥褻な表現のフィルタリング」として、ユーザ投稿型サイトであるPixivから取得した年齢制限のある(R-18)文章作品を対象にテキストマイニングを行い、学会発表を行った。
対象となった作品を「有害」と論文内でラベル付けたこと、作品名とURLを公開したこと、倫理委員会の審査を経ていなかったことなどが批判された(松谷, 2017)。現在、予稿PDFは非公開となっている。

ケース②

官報に掲載された破産者らの氏名や住所などの個人情報を、インターネット上の地図上にまとめたという「破産者マップ」が公開された。名誉やプライバシーを侵害するという批判が相次ぎ、政府の個人情報保護委員会がサイトを閉鎖するよう運営者に行政指導。１９日にサイトは閉鎖された。

被害対策弁護団によると、少なくとも過去約３年分の全国の破産申し立て事件についての債務者らの名前、住所、事件番号などがグーグルマップ上に掲載されていた。一時期は名前を元に検索できる機能もあったという。破産者マップでは「官報に掲載された破産者を地図上に可視化しました」と説明していた。

これに対し個人情報保護委は、「本人同意を得ずに、個人データを第三者に提供してはならない」「個人情報を取得時に利用目的を、本人に通知または公表しなければならない」などと、個人情報取扱事業者に対して定める個人情報保護法に照らして問題がある、と指摘した。運営者から１９日未明に「サイトを閉鎖する」との連絡があったという。

破産者マップに行政指導　プライバシー侵害の批判相次ぐ：朝日新聞デジタル　2019年3月20日
https://www.asahi.com/articles/ASM3N3T12M3NULFA008.html　(2020.10.02アクセス)

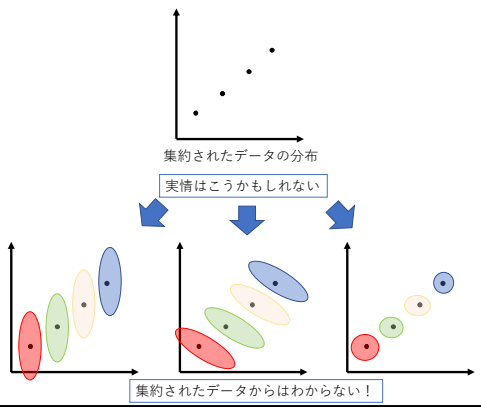
参考：インターネットを介した心理学研究を実施する際の倫理的原則(British Psychological Society, 2017)

- **個人・コミュニティの自由意志、プライバシー、尊厳の尊重**
  - 他者に観察されることが想定される、公的な場における公的な行動でない限り、同意が必要
    - 何が「公開」情報であるかは文脈に依存する
  - 個人の研究に参加しない権利を尊重する
    - 削除の申し入れがあったら従う
- **科学的誠実さ**
  - 研究対象者の状況や心情、行動を理解し、研究の実施状況およびデータの品質について配慮する
- **社会的責任**
  - 社会構造を混乱させることを回避し、研究のもたらしうる結果それぞれについて慎重に配慮する
    - 特に既存の社会集団への影響に注意
- **利益の最大化および害の最小化**
  - 研究から得られる科学的な価値を最大化し、研究から生じうる害から研究対象者を守る
    - 機密性・匿名性の保持など

技術的な話題

二次分析で特に注意すべきこと

- 政府統計などのデータ：集約されたデータ
  - 交絡(confounding)の影響を受けやすい
    - 交絡：第三の変数の影響を受けること
  - 生態学的相関(ecological correlation; Robinson, 1950)
    - 地域（集団）レベルであてはまることが**個人レベルで当てはまるとは限らない** (ecological fallacy:生態学的誤謬)



Webスクレイピングの戦略

- クローラをどうするか？
  - 自分で作る
    - 多少プログラミングが必要…ハードルが高い？
  - 既存のツールを使う
    - ダウンローダ(e.g. wget)
      - URIのリストを読み込ませて収集する

データの整形

- 取得したデータ
  - CSV / TSV: 表形式
    - 扱いやすい
  - HTML / XML / JSON：入れ子構造
    - そのままでは扱いづらい→情報を抽出する必要がある

データ処理のフロー

データ収集デモ

Twitter APIを使ったデータの取得

- Twitter APIを使うには開発者登録が必要
  - rtweetパッケージを使えば、開発者登録なしにAPIを使うことが可能(要Twitter アカウント)
- できること（一部）
  - キーワード検索
  - ユーザのつぶやきを取得
  - 全つぶやきの1%をリアルタイムで取得

ダウンローダを使ったファイル収集

- wget
  - 指定したURLをダウンロードするコマンド
  - リンクをたどって再帰的にダウンロードすることもできる
  - オプション
    - --wait
      - アクセス間隔（秒）
    - --input-file
      - ファイルからURLリストを読み込む
    - --no-crobber
      - ファイルを上書きしない
    - --user-agent
      - アクセスする側の情報を記載する
    - -P
      - 保存先フォルダを指定する

実行例

```
wget --input-file=url_list.txt --wait=5 --user-agent="Example Crawler(自分のメールアドレス)"
```

クローラを使った収集時は

- 連絡先を明示する
  - ユーザエージェントに「ボット名(連絡先)」の書式で
- アクセス間隔を(1秒以上)開ける
  - 間隔なしでアクセスすると相手先に負荷がかかる

JSONの読み込み

- jsonliteパッケージを使って読み込む(Rの場合)
- 構造は失われるが、表形式で読み込むこともできる(fromJSON())

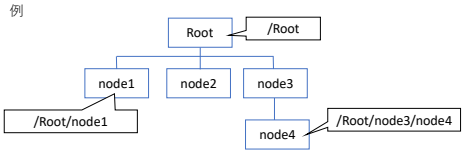
XMLの読み込み

- xml2パッケージを使う(Rの場合)
- データフレームで読み込もうとすると構造が失われる
- XPath記法を用いて構造を保持したまま情報を取得することができる

XPath記法

- XPath(XML Path Language)
  - XMLの特定要素を指定するための記法
  - ノード位置の指定や属性を指定できる
    - つながっているノード同士で
      - 「根」に近い方のノード：親(parent)ノード
      - 「葉」に近い方のノード：子(child)ノード
    - 同じ階層のノード：兄弟(siblings)ノード
  - HTMLにも使える

例



- Rootは根ノード
- Rootの子ノードはnode1, node2 , node3
- Rootはnode1, node2,node3の親ノード
- node1, node2, node3は兄弟ノード

…

HTMLの読み込み

- xml2パッケージを使う(Rの場合)
  - XMLと同じようにXPath記法が使える
- 一般にHTMLの構造は複雑
  - ブラウザのdeveloper toolを使うなどして対象の構造を調べるとよい

HTMLの初歩

- 基本構造
  - ヘッダー (<head>)
    - Webサイトに関する情報（メタデータ）
  - ボディ(<body>)
    - Webサイト本体の記述
- 要素の例
  - リンク(<a>)
    - href属性にリンク先が指定される
  - リストの項目(<li>)
    - 順序なしリストタグ(<ul>),順序付きリストタグ(<ol>)でグループ化されている
  - 要素をグループ化( )
- 属性とクラス、ID
  - 属性：要素の動作を指定したり調整したりする
  - クラス：同じように扱いたい要素に名前をつける
  - ID：要素に固有の名前をつける

まとめ

- 対面で収集する以外にも選択肢はある
  - 既存データ（二次分析）
  - Webデータ
- ただし、これらの方法にも制約がある
  - データの不完全性
  - 交絡・生態学的相関, etc.
- 制約を理解した上で利用すれば、研究の幅を広げ、実りあるものにできる