

研究データ利活用協議会シンポジウム

計算社会科学における データ利活用

データの取得・公開・分析とコミュニケーション

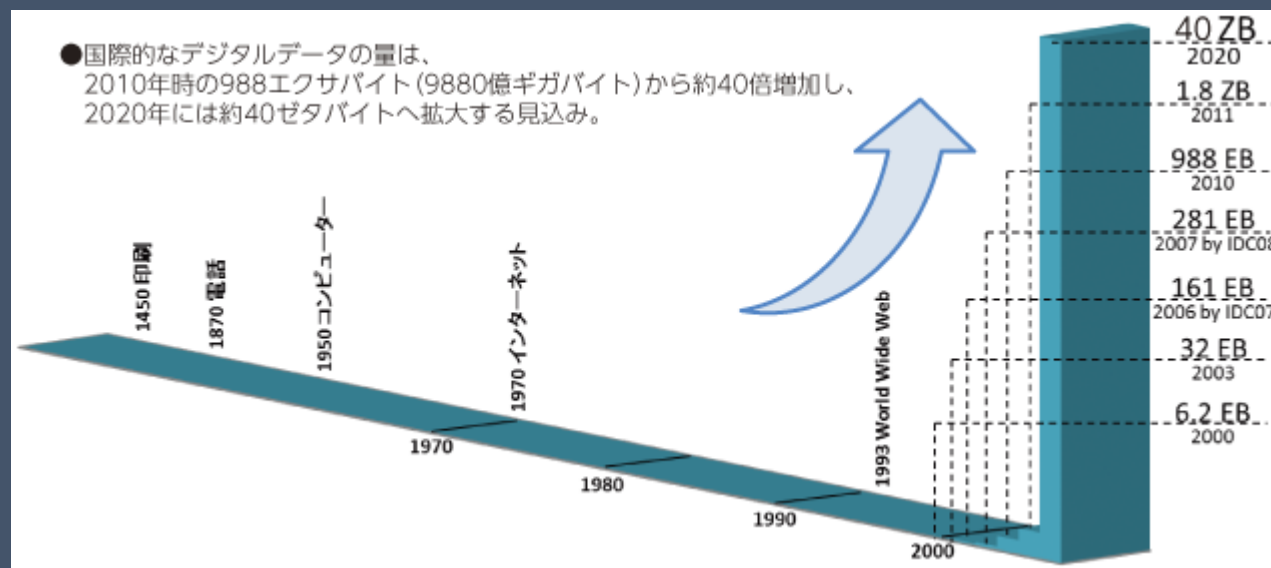
東京大学 大学院工学系研究科
鳥海不二夫

今日の内容

- 計算社会科学とデータ
- データの取得
 - 公開データの活用
 - オープンデータ
 - オープンソースソフトウェア
 - データの収集
 - 購入
 - 共同研究
 - 組織内データの活用
 - 開示請求
 - 構築
- データの公開

デジタルデータの世界

- コンピュータに保存されている大量のデータ
 - WEB, アクセスログ, ツイート, 動画, メール . . .
 - GPS位置情報, 監視カメラ, 電気使用量 . . .



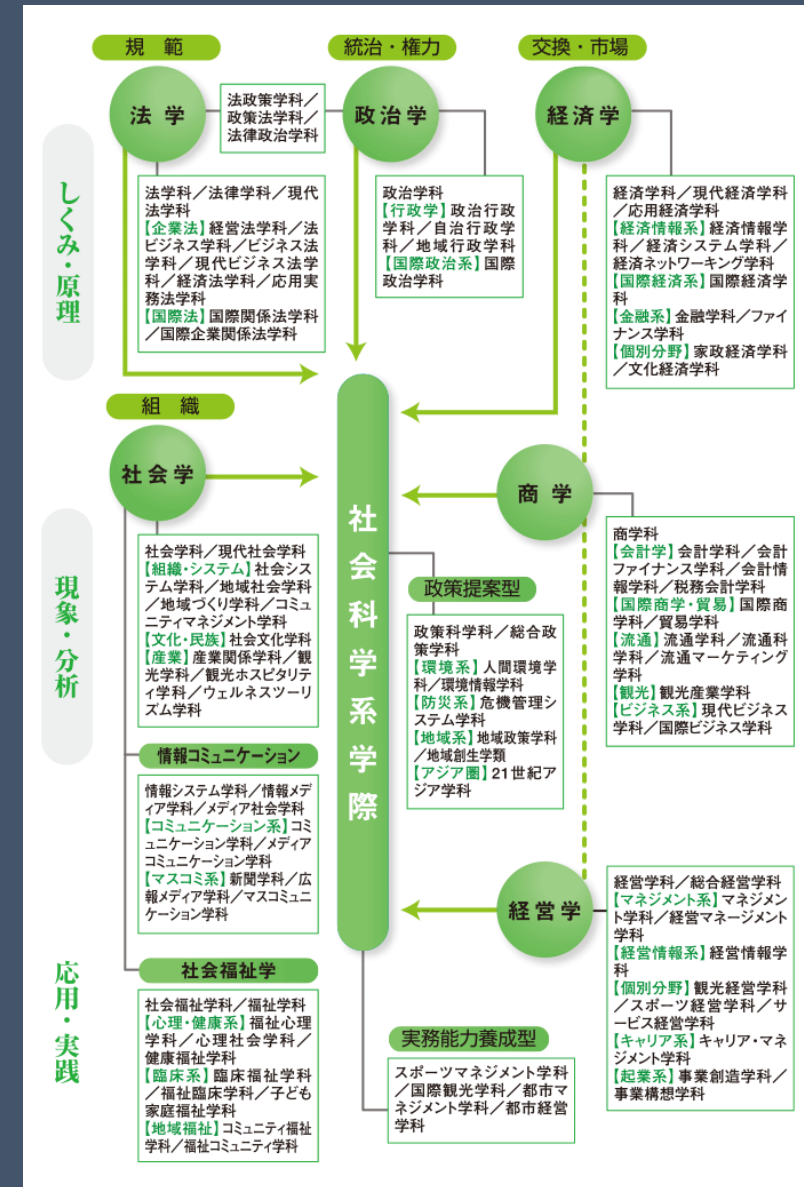
データと社会

- 従来の社会
 - 行動は記録に残らない
 - 定量的分析は不可能
 - アンケートや統計データによる分析
- デジタル社会
 - 人々の行動が日々デジタルデータ化
 - ネット上の行動, センサデータ
 - デジタルデータの分析→社会の理解
 - 人間の集団＝社会の理解へ



デジタル化社会の社会科学

- 人間社会を科学する社会科学
 - 社会学，経済学，法学，教育学，政治学，国際研究，コミュニケーション，メディア学
- 世の中に存在する様々な社会の問題を解決
 - 従来は文系の学問
 - デジタル化された社会→アップデートが必要
- 社会データ分析を用いた新しいアプローチ
 - データサイエンス
 - 人工知能



計算社会科学

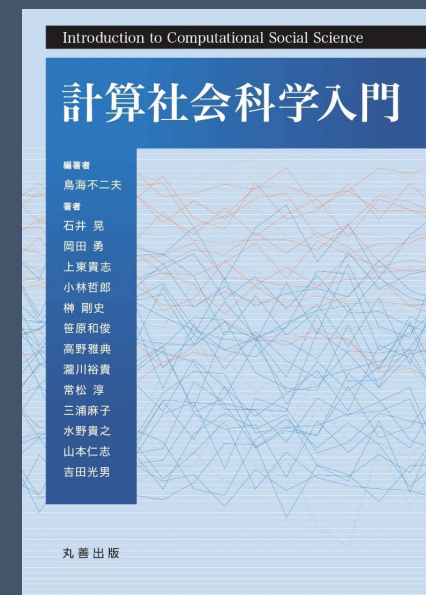
- 人間・社会と計算の組み合わせが可能になったことで実現した新しい学際科学
 - 現実世界を映し出すビッグデータ
 - 観察・実験の場としてのウェブ
 - 複雑な社会を扱う数理と情報技術

計算機科学
ビッグデータ
人工知能
機械学習
自然言語処理



社会心理学・社会学
社会情報学・政治学

統計数理
経済物理学，社会物理学
ネットワーク科学
ゲーム理論・ABS

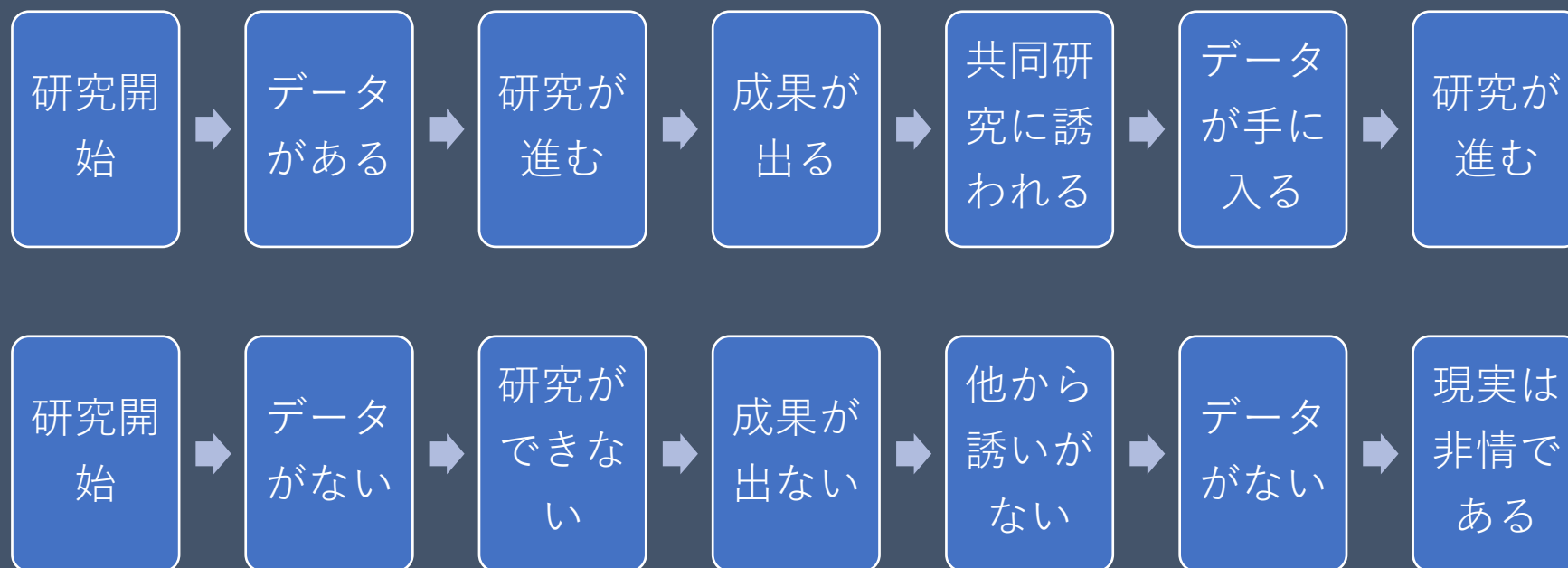


研究データとオープンサイエンス

- 研究データ
 - 研究実施過程で収集・生成・作成されるデータ
- オープンサイエンス
 - 研究論文のオープン化
 - 研究データのオープン化
- データとは
 - 観測データ, 派生データ, 標準データ, ソースコード

データの偏在性の法則

- データの偏在性の法則
 - データはデータが存在するところに集まるという計算社会科学の基本経験則



データをどのように手に入れるか

- 公開データの活用
 - オープンデータ
 - オープンソースソフトウェア
- データの収集
 - 購入
 - 共同研究
 - 組織内データの活用
 - 開示請求
 - 構築

公開データの探し方

- データポータルサイト
 - 公開データをまとめているサイト
 - データ所有者によるサイト
 - 政府機関
 - 公的機関
 - 公開データ集約サイト
 - SSJデータアーカイブ
 - Google Dataset Search
- 主なデータ提供者
 - 政府・自治体公開情報
 - 国際機関統計情報
 - 企業公開
 - 有償/無償/オープンコール
 - 研究機関
- オープンコール
 - コンペティション形式等
 - アルゴリズムの開発やデータ分析のためのデータ公開
 - Netflix Prize, Kaggleなど

公開データの例

組織	公開データ例
政府・自治体公開情報	地理・地域交通データ, 犯罪・防犯データ, 人口データ, 選挙・地方議会のデータ, 予算・税金データ, 金融・経済活動に関するデータ, エネルギーの生産・消費データ, 教育データ, 農業・漁業データ, 環境および汚染に関するデータ, 廃棄物および水管理データ, 地域・観光情報
国際機関統計情報	金融・経済活動に関するデータ, 人口データ, エネルギーの生産・消費, データ, 健康関連データ, 科学に関するデータ
企業公開・提供情報	企業情報, 金融・経済活動に関するデータ, メディア情報
研究機関公開データ	科学に関するデータ, 社会調査データ, 文化資料, 金融・経済活動に関するデータ

公開データセットの例

- WEB上に存在するデータセットの例
 - DATA GO JP
 - <https://www.data.go.jp/>
 - 情報学研究データリポジトリ (NII)
 - <http://www.nii.ac.jp/dsc/idr/>
 - Stanford Network Analysis Project(SNAP)
 - <http://snap.stanford.edu/>
 - ICWSM DATASET SHARING SERVICE
 - <http://www.icwsn.org/2018/datasets/datasets/>

公開データセット

- 利点

- 大量のデータを利用可能
 - もっとも効率的なデータ取得
- 誰でも利用可能
 - 多数の人と同一データを使うことで比較が容易

- 欠点

- データの独自性はない
- 欲しいデータがあるとは限らない
- データによっては利用に制限が存在

- 取得方法

- 公開サイトによって異なる
 - 自由にダウンロード可能な場合
 - 申請が必要な場合

- 注意点

- 商用利用は出来ないことが多い
- 研究利用が出来ないデータも
 - コンテスト用データセットなど
- 論文発表前に報告が必要
 - 30日前（！！）に報告など

データをどのように手に入れるか

- 公開データの活用
 - オープンデータ
 - オープンソースソフトウェア
- データの収集
 - 購入
 - 共同研究
 - 組織内データの活用
 - 開示請求
 - 構築

データの収集 (1/2)

- 購入
 - お金で解決できることはお金で解決
 - 新聞のテキスト, 商品の売り上げ情報, 人流データ, ...
 - 利用目的の制限
 - 商用利用の禁止, 二次配布の禁止, ...
 - 結果を公開できない場合がある
 - 事前に投稿論文を確認することを条件として許可される場合も
 - 購入時の契約を十分に確認
- 販売の周知
 - 公にされていない, 問い合わせれば出てくることも
 - わりと口コミが重要

データの収集 (2/2)

- 組織内データの活用

- 自社サービスのデータのほか、他社から購入したデータの活用
- 明確な目的無しで蓄積されたデータの大半はゴミ
 - 例えば、異常系ログ（エラーログ）のみを保存し、正常系ログを保存していない場合、エラー要因を多面的に分析することは困難
 - 自社サービス開発においては、データ蓄積方法の検討も大事
- 意外なデータから知見も
 - 日々の社食のデータからわかることもある

- 開示請求

- 行政機関や独立行政法人に対する情報公開制度を活用
- 食品営業許可台帳は昔から開示請求されやすい
 - 開示請求されやすい情報は、オープンデータとして公開される流れ
- ただし、黒塗りオペレーションの関係で、電子データであっても、上で印刷したものをスキャンしてるケースがほとんど

企業からのデータ提供（共同研究）

- 共同研究の目的
 - 企業がやりたいが自社リソースをかけられない場合のアウトソーシング
 - 研究機関が持つ独自技術を使った分析のための共同研究
- 共同研究におけるリソース提供
 - 共同研究費+データの提供
 - 共同研究費を発生させない契約であればやりやすい場合も
 - 実際には研究にはお金がかかるので共同研究費ももらう方が良い
 - 研究発表可能な契約を締結することを忘れずに
- 共同研究の流れ
 - データのみを受け取り何が出来そうか確認（NDAの締結）
 - 簡単な分析から目的を設定（共同研究契約）
 - 研究の実施，発表，成果報告
 - サービス実装などはまた別の話

企業からのデータ提供（共同研究）

- 企業のWEBサービスデータの提供
 - ニュース記事
 - ソーシャルゲームのユーザ行動
 - フリマサイトの売買記録
 - ソーシャルメディアのコミュニケーションログ
 - etc...
- 直接大量データを利用可能
 - 場合によってはほぼDBそのものの提供も
 - もっとも効率的なデータ取得
- 取得方法
 - 企業からの相談・マッチング
 - 企業へお願い

企業からのデータ提供・注意点

- 契約問題のクリアが必要
 - データをもらうまでの契約が大変
 - 数ヶ月かかることもざらにある
- 個人情報保護・権利関係
 - 個人情報の削除は大きな手間
 - 企業にも多大な負荷がかかる
 - 研究室の管理体制
 - 情報漏洩は企業にも研究者にも致命的
- データの権限
 - 現場の人はデータを出したい
 - 経営陣はリスクを考えデータを出したくない

(鳥海的) データのもらい方

1. 学会等でデータを持っていそうな企業の人を探す
(企業ブース等)

2. お近づきになる

3.



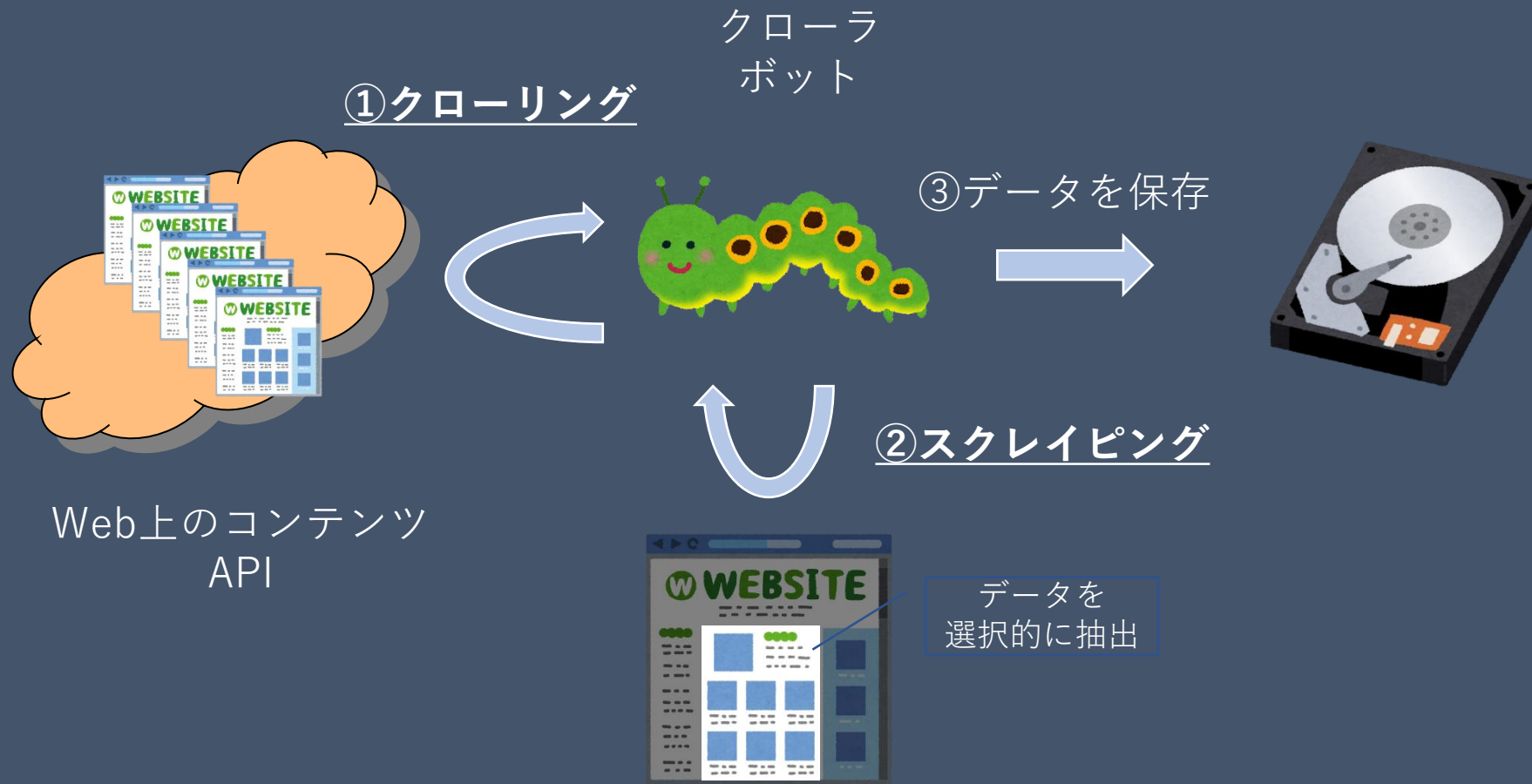
4. 企業内の調整を待つ

5. 断られる (1へ戻る)

公開データの自動収集

- APIを利用した自動収集
 - 機械可読なデータの提供
 - 昔はXML, 今はJSON
 - 昔はAmazon, 今はTwitterが有名
 - Twitterはアカデミック用のAPIも公開
 - 1000万/月でSearchAPIを使えるのは破格
- クローリングとスクレイピングによる自動収集
 - クローリング (crawling)
 - ウェブページの自動巡回
 - スクレイピング (scraping)
 - ウェブページから必要な部分のみを抽出
 - APIを利用した場合でも, APIをクローリングし, 取得データから必要な部分をスクレイピングする

クローリング・スクレイピング



自動収集における注意事項

- 自動収集の禁止

- 利用規約で自動収集を禁止しているケースがある

- Instagramの例

- <https://www.facebook.com/help/instagram/581066165581870>

不正な方法を用いて、アカウントの作成、情報へのアクセス、または情報の取得を試みることは禁止されています。

これには、弊社から明示的な許可を得ることなく、自動化された手段を用いてアカウントを作成したり、情報を取得したりする行為が含まれます。

- ケンブリッジ・アナリティカ（CA）による選挙介入疑惑の後、FacebookやInstagramのデータ取得は困難になっている

- 利用規約を遵守するかどうか

- 人工知能学会 倫理指針

人工知能学会会員は専門家として、研究開発に関わる法規制、知的財産、他者との契約や合意を尊重しなければならない。人工知能学会会員は他者の情報や財産の侵害や損失といった危害を加えてはならず、直接的のみならず間接的にも他者に危害を加えるような意図をもって人工知能を利用しない。

- 収集先サービスに負荷をかけない

- 過剰なアクセスにより、サービス障害を引き起こす可能性

- 岡崎市立中央図書館事件

- 貸出情報を取得するクローラを稼働させたところ、図書館のサーバがダウンし、偽計業務妨害容疑で逮捕

サービスプロバイダとの良好な関係

- クックパッドのデータを研究者に公開します
 - <http://techlife.cookpad.com/entry/2015/02/24/161915>
 - 「残念ながら、研究者の中には、クックパッドのデータを使用するため、悪質なクロールを行う方がいます。」（2015年2月24日 クックパッド）
- Wikipediaデータベース
 - <https://ja.wikipedia.org/wiki/Wikipedia:%E3%83%87%E3%83%BC%E3%82%BF%E3%83%99%E3%83%BC%E3%82%B9%E3%83%80%E3%82%A6%E3%83%B3%E3%83%AD%E3%83%BC%E3%83%89>
 - 「ウィキペディアのコンテンツなどのデータは、再配布や再利用のために利用できる一元化されたデータベース・ダンプでの提供が行われています。クローラを用いてコンテンツを収集しないでください。」

Twitterデータの収集 (1/2)

- Twitter APIの学術研究利用
 - 学術機関所属者（大学院生含む）による非商用目的の利用を緩和
 - フルアーカイブが利用可
 - 1000万/月のSearchAPI利用
- データの保存例
 - YYYY-MM-DD-HH-CrawlerID.txt
 - テキストファイルにデータを保存
 - ツイートIDとJSONをタブ区切り
 - データの重複除去
 - 1カラム目のツイートID（数値）によるユニーク処理
 - `sort -n -u YYYY-MM-DD-HH-*.txt -o YYYY-MM-DD-HH.txt`
 - データの圧縮
 - `gzip YYYY-MM-DD-HH.txt`
 - JSONの圧縮効率は良い（12%程度）
 - 最近だと lz4 もお勧め（圧縮効率はgzipの半分、圧縮・伸張は10倍高速）
 - データの読み込み
 - `zcat YYYY-MM-DD-HH.txt.gz | cut -f 2`
 - 2カラム目を取り出し、JSON Lines (.jsonl) 形式での取り扱い
 - 破損すると伸張エラーが出るので、エラー検知もしやすい

Twitterデータの収集 (2/2)

- ツイートは割と消える [吉田 20]
 - [鳥海 20]が公開しているデータセットのうち、感情分析に用いられた28,547,492ツイートについて、データセットのツイートIDをもとに再取得したところ、ツイートが削除されたまたは非公開にされたことにより、4,370,020ツイート（15.3%）がアクセス不能であった。
 - 調査開始時にフルアーカイブを利用してデータを収集したとしても、当時の状況を再現するのは困難
- Twitter Search APIを定期的に起動して常時収集
 - 1時間おきにcronでクローラを定期起動
 - Dropboxの共有ファイル機能を利用し、そのファイルに収集条件（キーワード）を記述しておく
 - クローラは起動時にこのファイルをロードする
 - 共同研究者等もこのファイルを書き換えることで、容易に条件の追加ができる

[鳥海 20] 鳥海不二夫, 榎剛史, 吉田光男. ソーシャルメディアを用いた新型コロナ禍における感情変化の分析. 人工知能学会論文誌. 2020, vol. 35, no. 4, p. F-K45_1-7.

[吉田 20] 吉田光男. COVID-19流行下におけるソーシャルメディア --日本での状況と研究動向・公開データセット-- . 人工知能. 2020, vol. 35, no. 5, p. 644-653.

Twitter開発者ポリシー

- Twitter利用者の以下に関する情報を取得したり推定することや、取得もしくは推定した情報の保存は行わないでください。
 - 健康状態 (妊娠を含む)・財務状況の悪化・政治的所属もしくは政治理念・人種・宗教、哲学的な信仰もしくは信念
 - 性生活もしくは性的指向・労働組合への加盟の有無・犯罪容疑もしくは実際の犯罪行為
- Twitter利用者をTwitter外の個人識別子と関連付けようとする場合、関連付けを行う前に利用者から明示的なオプトインの承諾を得る必要があります。
- Twitterは第三者に対するTwitterコンテンツの再頒布に制限を設けています。
- Twitterコンテンツを第三者に提供する場合（ダウンロード可能なデータセットまたはAPIを通じて提供する場合を含む）、ツイートID、ダイレクトメッセージID、アカウントIDのみを配布できます（下記の例外を除きます）。
- Twitterは、学術研究者が非営利の研究目的でツイートIDおよびアカウントIDを共有することを特別に許可しています。
- 学術研究者は、学術機関を代表して非営利の研究目的で使用する場合に限り、ツイートIDやアカウントIDを無制限に配布することが許可されます。たとえば、ピアレビューまたは研究の検証を可能にする目的でツイートIDを無制限に共有することが認められます。

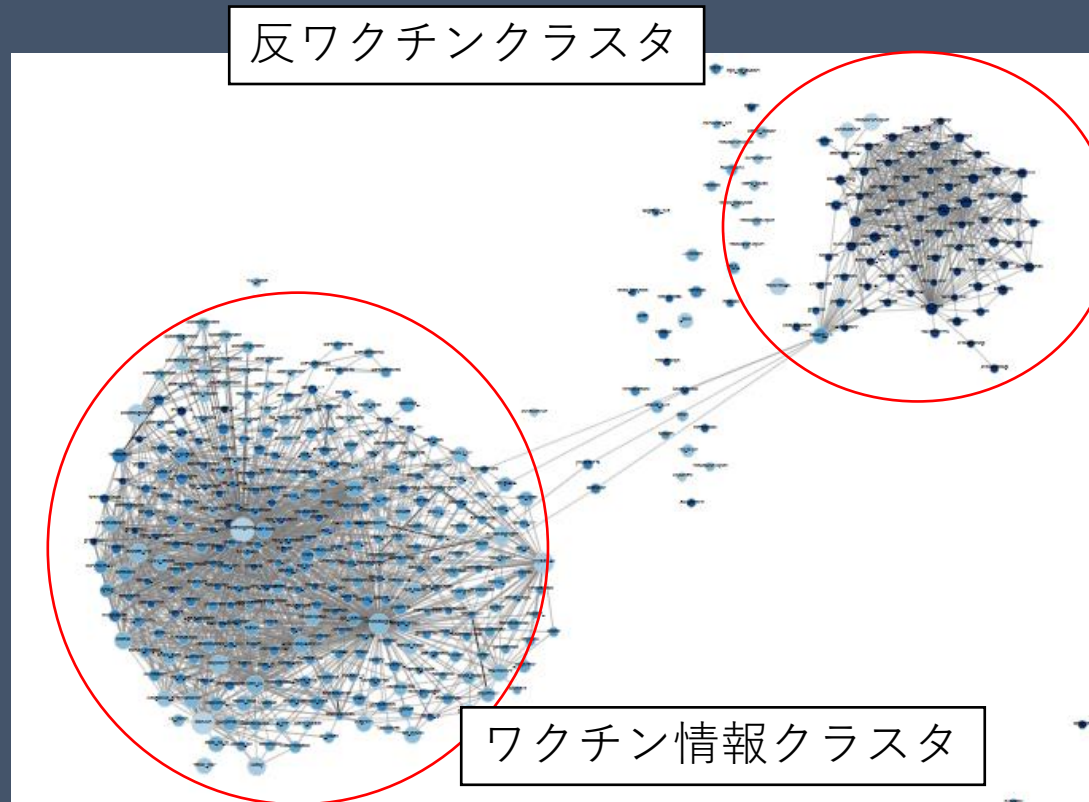
簡単ツイート収集アプリ

- Tiny Tweet Crawler 
 - 検索キーワードを含むTweets を収集するフリーウェア
<http://mtmr.jp/ttc/>
- Web Tweet Crawler
 - 検索キーワードを含むTweets を収集するWEBアプリ
<http://torix.sakura.ne.jp/twitter/>

WTCで収集可能なデータ

列名	説明
id	Tweet ID
screen_name	ツイートユーザのScreenName
name	ツイートユーザの名前
created_at	ツイート作成日
text	ツイート本文
retweeted_id	リツイート元のID
retweeted_user	リツイート元のツイートを投稿したユーザ
retweeted_create_at	リツイート元の投稿時間
in_reply_to_screen_name	リプライ先のScreenName
in_reply_to_status_id	リプライ先のTweetID
in_reply_to_user_id	リプライ先のUserID
source	ツイートを投稿したアプリ
retweet_count	リツイートされた回数
favorite_count	Favられた回数
user_id	ツイートユーザのユーザID
statuses_count	ツイートユーザの総ツイート数
followers_count	ツイートしたユーザのフォロワー数
friends_count	ツイートしたユーザがフォローしている数
favourites_count	ツイートしたユーザの総ふぁぼり数
keyword	検索に使ったワード

ワクチンにおける分断

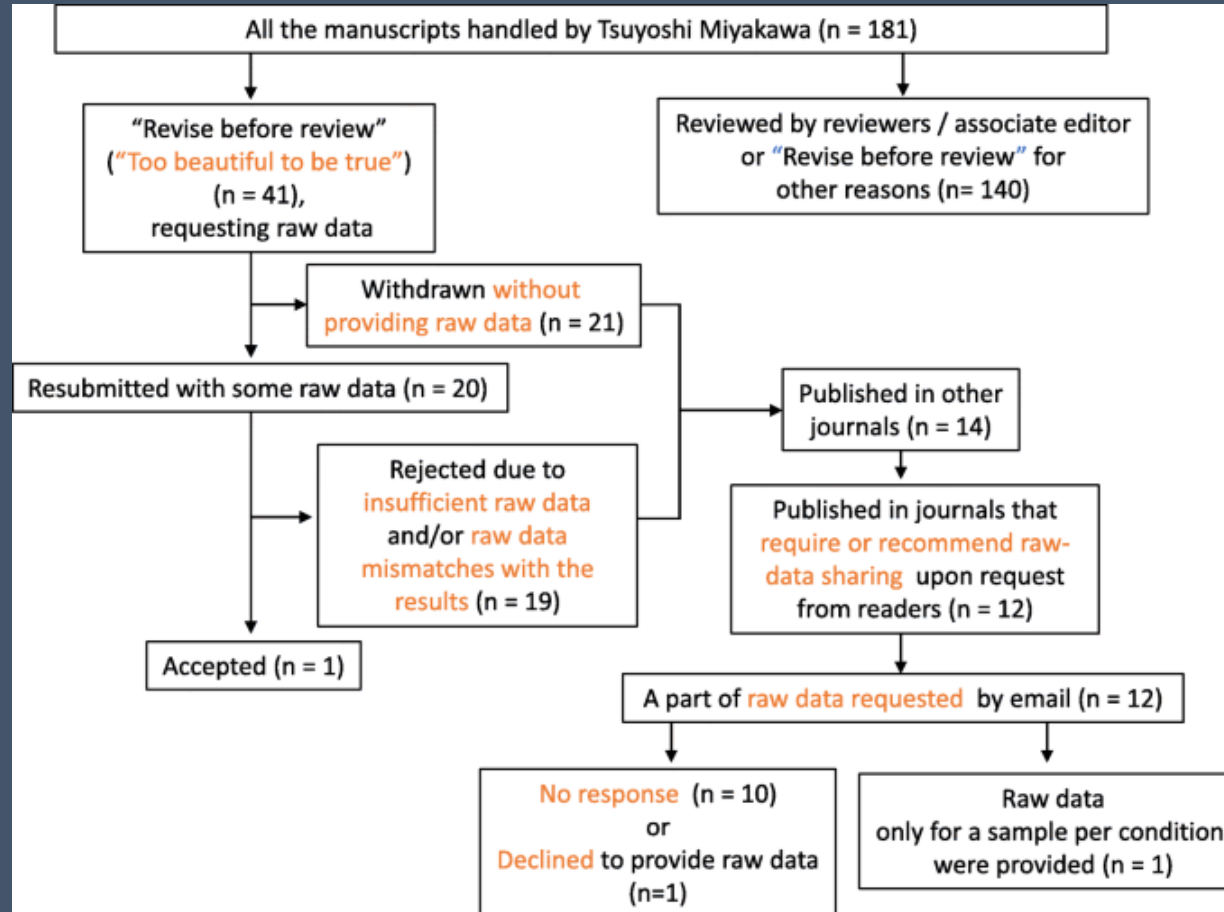


- 大きく二つのクラスタを抽出
 - ワクチンにポジティブ
 - ワクチンにネガティブ
- 反ワクチンの偏りが大きい
 - 偏りの算出
 - あらかじめ抽出していたコミュニティと比較してどの程度偏ったアカウントによるクラスタか計算
 - ポジティブ：0.760
 - ネガティブ：1.593
- ポジティブな方が偏りが少ない
 - ネガティブ派は偏ったアカウントによる拡散の可能性

データの公開

- なぜ公開するのか？
 - 研究内容の再現性の担保（研究公正）
 - 多くの学術誌がデータ公開を促している
 - <https://journals.plos.org/plosone/s/data-availability>
 - *PLOS journals **require authors to make all data necessary** to replicate their study's findings **publicly available without restriction** at the time of publication. When specific legal or ethical restrictions prohibit public sharing of a data set, **authors must indicate how others may obtain access to the data.***
 - 研究者自身のブランディング
 - 論文の被引用数が伸びる [Piwowar 13]
 - 学術コミュニティへの貢献
 - オープンアクセス・オープンデータは分野の振興に繋がる（はず）
- どのように公開するか
 - 「オープンデータ」として公開するのが望ましいが…
 - Open Knowledge International, 総務省による定義
 - 利用, アクセス, 再利用, 再配布が誰にでも認められるデータ
 - 無償かつ機会判読に適した誰もが二次利用可能なデータ

研究データの公開の現状



公開可能なデータ

- 自前で収集したデータなら自由に公開できる？
 - パネル調査サービスやクラウドソーシング
 - どのような許諾の元の調査かを確認する必要がある
 - 自前のアンケート
 - 個人情報保護を含む倫理的観点の配慮
 - ウェブ上のデータ
 - 著作権法第三十条の四により、加工済みデータ（モデル）の公開や販売が可能
 - 利用規約等による制約
 - Twitterであれば、公開できるデータはIDに限定
 - <https://developer.twitter.com/ja/developer-terms/agreement-and-policy>
 - Twitterコンテンツ（ダウンロード可能なデータセット、API経由を含む）を第三者に提供する場合、ツイートID、ダイレクトメッセージIDおよびアカウントIDのみを配布することができます（下記の例外を除く）。
- ウェブのデータは国を横断する
 - 日本の法令のみに縛られるわけではない
 - EU一般データ保護規則（GDPR）

データの公開場所

- 論文の補足資料
 - 少量のデータ（エクセル）を論文とセットで公開
 - 公開が容易，参照されやすい
- 研究者個人・所属組織のWebサイト
 - 自由度の高い公開方法
 - 永続性やバージョン管理に難点がある
 - 論文で使われたデータと改良されたデータの双方にアクセスしたいが…
- コード・データ共有サービス
 - GitHub, Zenodo, J-STAGE Data, …
 - **Zenodoをお勧めしています**
 - 非営利機関運営，無償，分野限定なし，DOI，バージョン管理，50GB
 - 機能に制約があるものの，バージョン管理などが標準で使える
- データ公開・提供サービス
 - 情報学研究リポジトリ（NII）など
 - データ公開に関する多くの作業を事業者に委託できる

データの倫理

- 従来の研究倫理にとどまらない
 - データを捏造・改ざんしてはならない
 - 計算社会科学・ソーシャルメディア研究は個人情報と密接に関係する
 - データの裏側には人間（データ主体）が居ることに注意
 - 行動データであれ、テキストデータであれ、ウェブで入手可能なデータは、人間が作成したもの
- 倫理規定
 - インフォームド・コンセント（説明を受けたうえでの同意）が要求されるが…
 - ウェブサイトにアクセスするユーザに許可取りますか？
 - Twitterのデータ収集でユーザに許可取りますか？
 - 研究倫理審査
- データの保存・活用
 - 長期間の安定的な保存，アクセスコントロール
 - 匿名化した活用

流出データは公開データ？

- 不倫するCEOは仕事で不正する可能性が2倍 [Griffin 19]
 - Ashley Madison（不倫目的のSNS）のデータと公的なデータを突き合わせた調査
 - 公的なデータは標準データであるものの、Ashley Madisonのデータは流出データ
 - 著者はPublic Domainであると主張している
*We have discussed the use of the data with many people, including attorneys, who confirm that the data are permissible to use for research purposes because **the data are now in the public domain and available for research use in the same way that they are available to and used by the press.** We believe it is also ethical to use the data, and **the use of hacked data has become common both by the press and in academia.***

まとめ

- 大規模データは計算社会科学では必要不可欠
- データの入手方法はさまざま
 - データの偏在性に負けるな！
- データは自由に使えるわけではない
 - データの入手法，使い方には注意
- 大きなデータには大きな力
 - 扱うにあたって倫理的観点の配慮が必要
- 計算社会学会にデータセットのリンク集



<https://css-japan.com/bookicss/>

詳しく知りたい方は

- まえがき
- 第1章 計算社会科学とは（笹原和俊）
- 第2章 Web調査（三浦麻子）
- 第3章 デジタル実験（小林哲郎）
- **第4章 データ収集・公開データセット（吉田光男，榊剛史）**
- 第5章 ネットワーク（鳥海不二夫）
- 第6章 テキスト分析：データとしてのテキスト（瀧川裕貴）
- 第7章 ソーシャルデータ分析のための教師あり機械学習（水野貴之）
- 第8章 社会シミュレーション（岡田勇，山本仁志）
- 第9章 統計モデリング（高野雅典）
- 第10章 社会物理（石井晃）
- **第11章 計算社会科学における倫理（常松淳）**
- 第12章 計算社会科学の今後の展望と課題（上東貴志）
- 索引

