大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

IDR
情報学研究データリポジトリ

リアルデータの「共同利用」

あなたの情報が学術研究に!?
… でも大丈夫

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

情報学研究データリポジトリ

講師 大山 敬三

大学共同利用機関法人情報・システム研究機構
国立情報学研究所
コンテンツ科学研究系・教授
データセット共同利用研究開発センター長

このようなサイト，使っていますか？

Yahoo!知恵袋 https://chiebukuro.yahoo.co.jp/

わからないことを質問したり，
知っていることを答えたり，
溜まっている質問と回答を検索したり。

クックパッド https://cookpad.com/

今夜の献立を検索したり，
クリスマスの料理を選んでみたり，
作ってみたけど失敗したり，
自分なりにアレンジしたのを自慢してみたり。

楽天市場 https://www.rakuten.co.jp/

商品を検索したり，
欲しいものにチェックを付けたり，
注文したり，感想を書き込んだり。

楽天トラベル https://www.rakuten.co.jp/

ツアーを検索したり，予約したり，
ホテルの評判を見たり。

このようなサイト，使っていますか？

ホットペッパービューティー　https://beauty.hotpepper.jp/

お店を検索したり，
予約したり，
クーポンを使ったり，
できばえを書き込んだり。

ライフルホームズ　https://www.homes.co.jp/

アパートを探したり，間取りを見たり，
家賃を比べたり，
近くのコンビニや学校を確認したり，
内覧の予約をしたり。

ニコニコ動画　https://www.nicovideo.jp/

投稿したことありますか？
コメントしたことはありますか？

ニコニコ大百科　https://dic.nicovideo.jp/

投稿したことはありますか？

このようなサービスの会員の方は？

不満買取センター　http://fumankaitori.com/

家電で不便に感じたことや，
公共サービスを不満に思ったことや，
思い付いた食品のアイデアなどを投稿したり。

インテージシングルソースパネル　https://www.intage.co.jp/service/platform/issp/

テレビで番組やCMを見たり，
PCやスマホでウェブを見たり，
スマホでアプリをいじったり。

みんレポ　http://minrepo.com/

おすすめのお店を紹介したり，
購入した商品の使い勝手をレポートしたり，
楽しかった場所を投稿したり。

情報学の研究にはリアルデータが必要！

情報学は（ほぼ）道具を研究する学問
道具は実際に使ってこそ価値がある？

刀剣	美術品	戦の道具
情報学	理論の 美しさ	産業・社会 への応用

情報学の研究にはリアルデータが必要！

研究と実応用が接近：ビッグデータ，AI …

- 従来
  - 人手で集めたデータ
  - 実験室で作ったデータ
- 現在
  - 実際のサービスや社会で発生するリアルで大量のデータ

リアルデータを利用する研究の形

企業との共同研究や委託研究

- 企業のニーズを直接反映した研究

研究者にとって

- 閉鎖性が問題。
  - 研究の検証，再現ができない。
  - 参加できる人が限られる。

企業にとって

- 対象を広げようとすると手間がかさむ。

リアルデータを利用する研究の形

Do-It-Yourself（DIY）

- 研究者がインターネットなどから勝手にデータ収集

研究者にとって

- 日本の著作権法は研究に寛容。しかしデータの再配布はできない
- 他者の研究の検証，再現ができない
- 研究の準備段階の負荷が膨大

企業にとって

- 研究者が各自ばらばらにデータをもって行く
  - サービスの性能に悪影響
  - 利用状況が把握できず，フィードバックも得られない

※ 事業に影響が出れば損害賠償も

リアルデータを利用する研究の形

コンペティション

- 共通のデータと課題で技術を競う

  例: Kaggle, Deep Analytics

研究者にとって

- 学生の演習やスキルのアピールにはいいが
- 独自の研究課題の発掘には不向き

企業にとって

- コストがかかる。中小企業には企画・実施の体力がない
- ほどほどの難易度と多くの参加者を集められる面白さがある課題の設定が難しい

リアルデータを利用する研究の形

オープンアクセス型データ

- 誰でも自由に使えるデータ
- 行政や公共のデータが中心
- 個人に関わる情報は絶対に安全なところまで丸める

研究者にとって

- 自由に使える。検証性・再現性もある
- 企業が集めたリアルデータが提供されていない

企業にとって

- 誰がどう使うか想定できない
- そもそも企業の財産であるデータの提供は困難

リアルデータを利用する研究の形

第五の形：データセットの共同利用

- 共通のデータを多くの研究者が使う
- 利用のルールを決めて利用
- 利用者をしっかり管理
- 研究成果もしっかり共有

「共同利用」のメリット – 研究者には？

- 情報へのアクセスはオープンで公平
  - データについての情報は誰でも見られる（中身は契約者だけに）
  - 一定の基準を満たせば誰でも使える
    - 企業とのコネは必要ない
    - 他分野の研究者でも利用できる
- データ利用の自由度が高い（ただし研究目的）
  - 研究課題を自由に設定できる
  - 時間をかけてじっくり研究できる

「共同利用」のメリット – 研究者には？

- データの透明性が高い
  - 企業やユーザの権利を侵害する心配がない
    - ⇨ 大手を振って研究発表できる
  - 同一性が保証されている
    - 検証可能性や再現性がある
    - 他の研究との比較がしやすい
    - 研究のアピールがしやすい

「共同利用」のメリット
– 研究コミュニティには？

- 技術の比較評価のプラットフォーム
  - 共通の課題や評価尺度の定義
  - 研究成果の蓄積　…
  - ⇨ 研究者自身によるコンペの企画も可能
- 同一のデータや課題への多角的アプローチ
  - 共同研究の可能性
  - 異分野連携の核
- 参入障壁が低くなる
  - コミュニティの拡大
  - 大学院生によるチャレンジを促進

「共同利用」のメリット－企業には？

- 他の利用の形に比べてコストが最も小さい（ほぼデータを用意するだけ）
  - 共同研究のようなデータ準備や契約締結などの個別対応の手間が省ける
    ⇨より多くの研究者に使ってもらえる
  - DIY型のように自社システムに負荷をかけない
  - コンペのような企画や参加者募集，評価システム，賞品の準備などが不要
  - オープンデータ化のようなデータの徹底的なクリーニングが不要
    ⇨企業が使ってもらいたいデータを出しやすい

「共同利用」のメリット – 企業には？

- 利用者が管理されている
  - どういう研究者が興味を持っているかがわかる
  - どのような研究に使われているのかがわかる
  - 研究成果の把握ができる
    - 自社サービスの新規展開や課題解決につながる？
    - 共同研究や産学連携の候補を見つけられる

  ⇨ 安心してデータを提供できる
- データや研究成果の情報がオープン
  - 社会・学術への貢献をアピールできる
  - 学生などに研究に取り組む姿勢の認知が進む

「共同利用」のメリット
－社会にとっては？

- 現実の課題に即応可能な研究開発が進む？
- データから社会や経済活動の理解が進む？
- 大規模リアルデータを扱える人材育成が進む！（データサイエンティストの育成）

⇨ 少し遠い話かも知れないが…

- 社会全体の効率化・高度化ができる
- 社会活動の適正化ができる
- 国全体の競争力が高まる

共同利用には中核となる機関が必要

機関に求められる条件

- 社会や学術界に認知されていること
- 安定的に業務が継続できること
- 企業と研究者の両方のマインドが理解できること
- 企業や研究者の信頼が得られること

共同利用には中核となる機関が必要

国立情報学研究所（NII）

- 情報学分野の大学共同利用機関

データセット共同利用研究開発センター（DSC）

- NIIに設置されたデータの共同利用のための組織

情報学研究データリポジトリ（IDR）

- DSCによる企業等のデータを共同利用するための事業

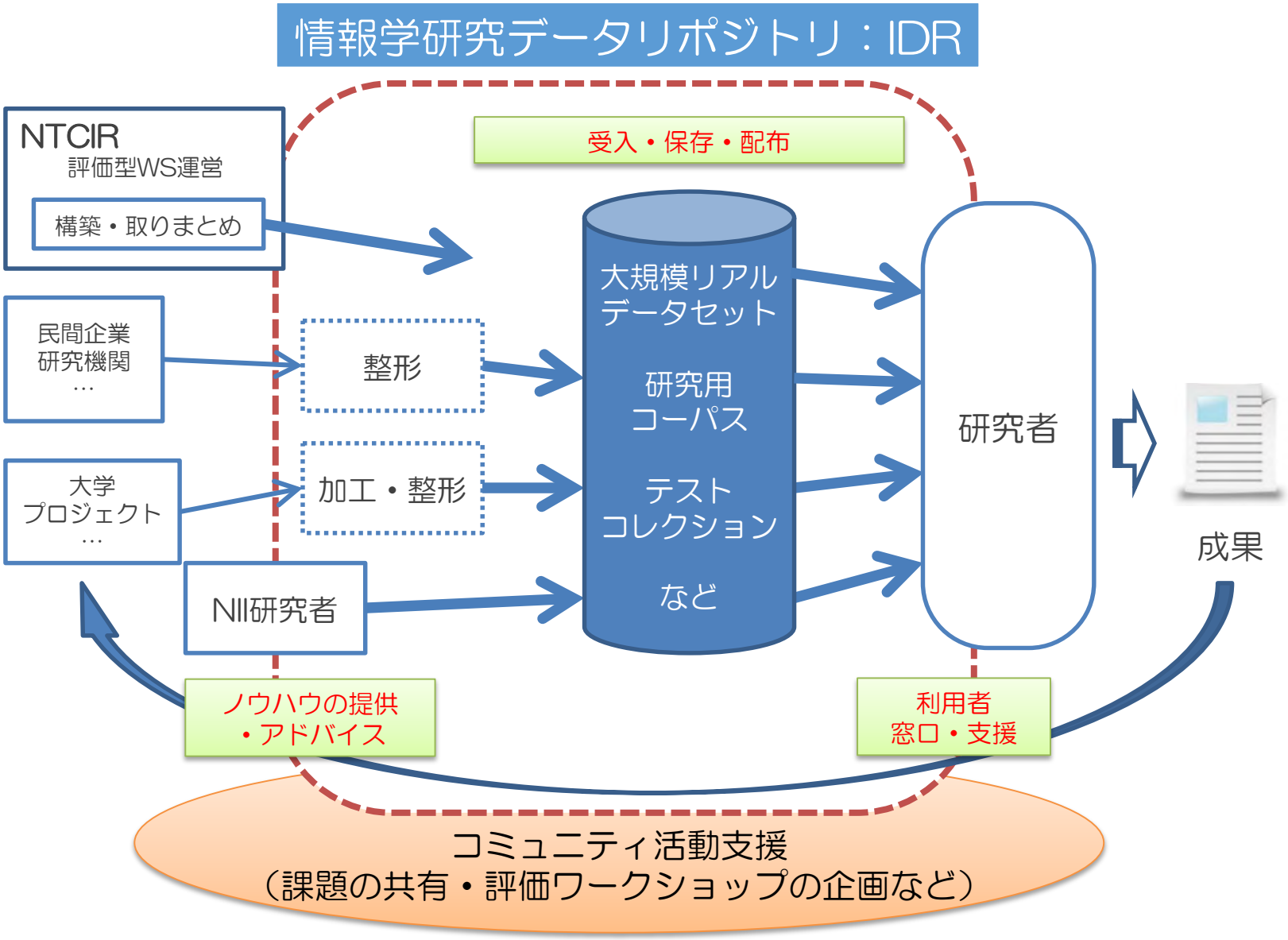
※大学共同利用機関とは？

- 個々の大学や研究者では整備することが難しい研究用の施設・設備や資料などを整備して多くの研究者の利用に供するために設置された学術研究機関。

IDRとはどんな事業？

情報学研究データリポジトリ：IDR

- 民間企業が持っているデータを受け入れて，多くの研究者に提供
- 研究者と企業をつなぐ窓口
- データセット構築から配布までのノウハウを集約して企業に提供
- 研究分野や産学の垣根を越えたコミュニティの形成や連携を促進


情報学研究データリポジトリ：IDR
NTCIR
評価型WS運営
構築・取りまとめ
受入・保存・配布
民間企業
研究機関
…
整形
大学
プロジェクト
…
加工・整形
NII研究者
大規模リアル
データセット
研究用
コーパス
テスト
コレクション
など
研究者
成果
ノウハウの提供
・アドバイス
利用者
窓口・支援
コミュニティ活動支援
（課題の共有・評価ワークショップの企画など）

提供しているリアルデータ

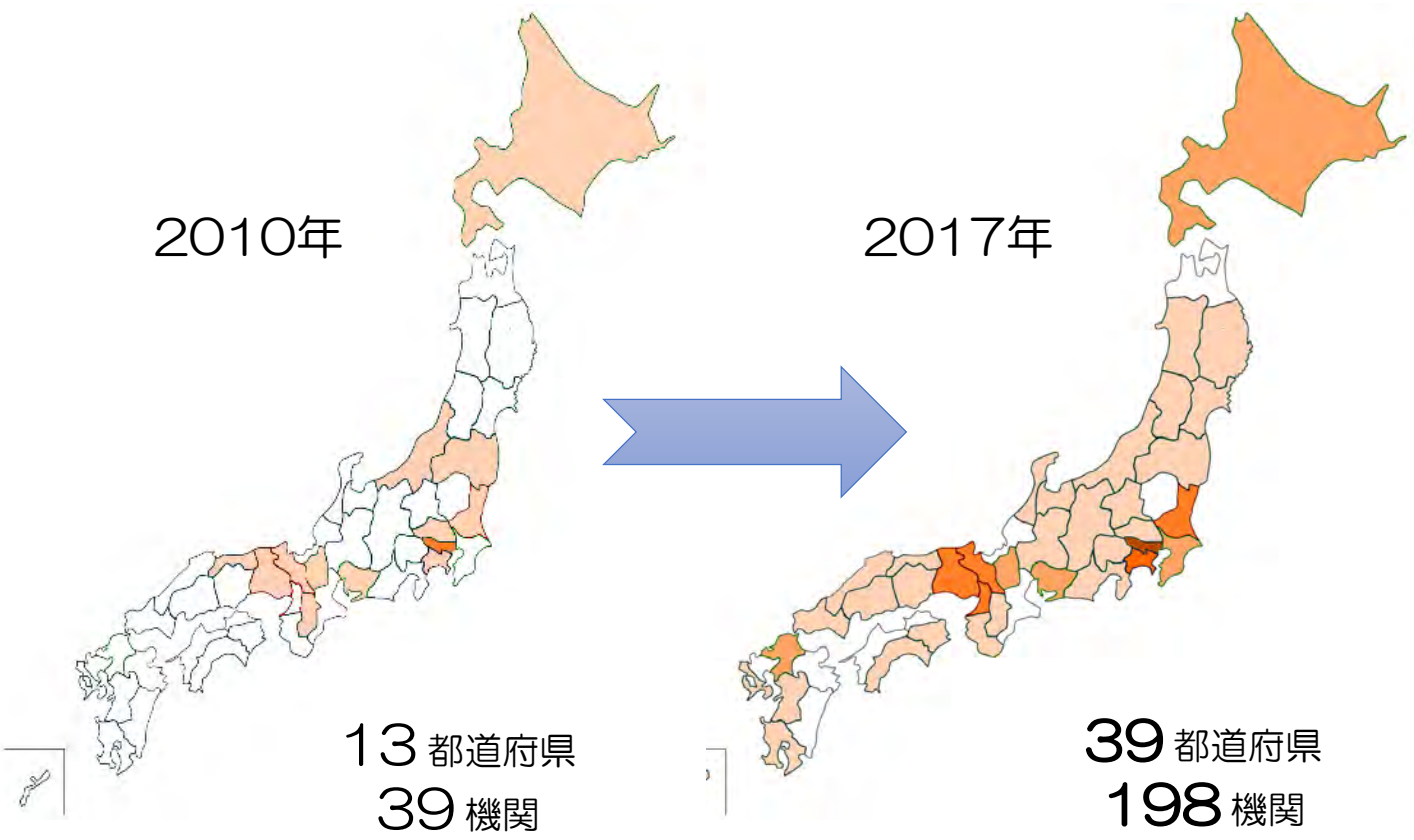
- Yahoo! データセット
- 楽天データセット
- ニコニコデータセット
- リクルートデータセット
- クックパッドデータセット
- LIFULL HOME'S データセット
- 不満調査データセット
- インテージデータセット

⇨これらサービスで取得されたデータのスナップショット

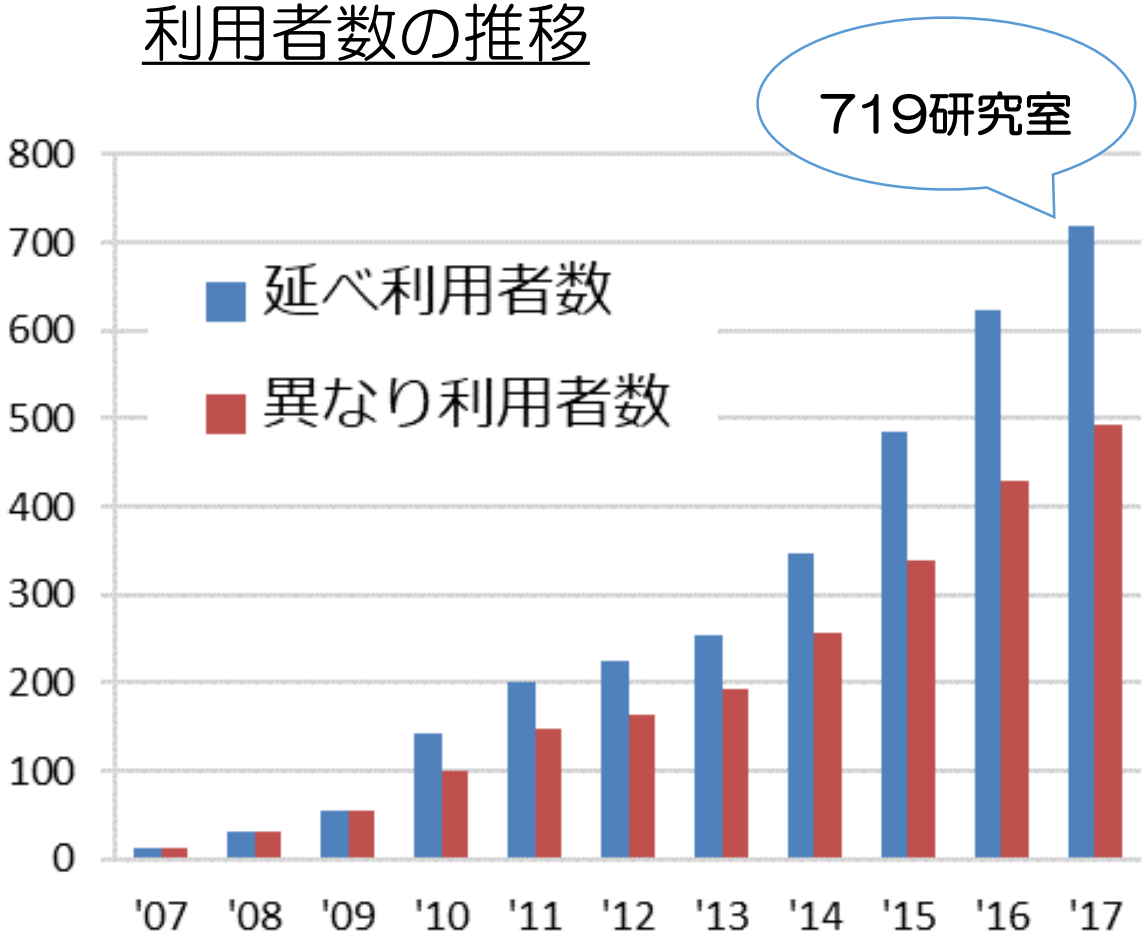
  - ユーザの登録個人情報は含まれていない
  - システムへのアクセスログや取引情報なども含まれていない

どのくらい使われている？

データ提供先

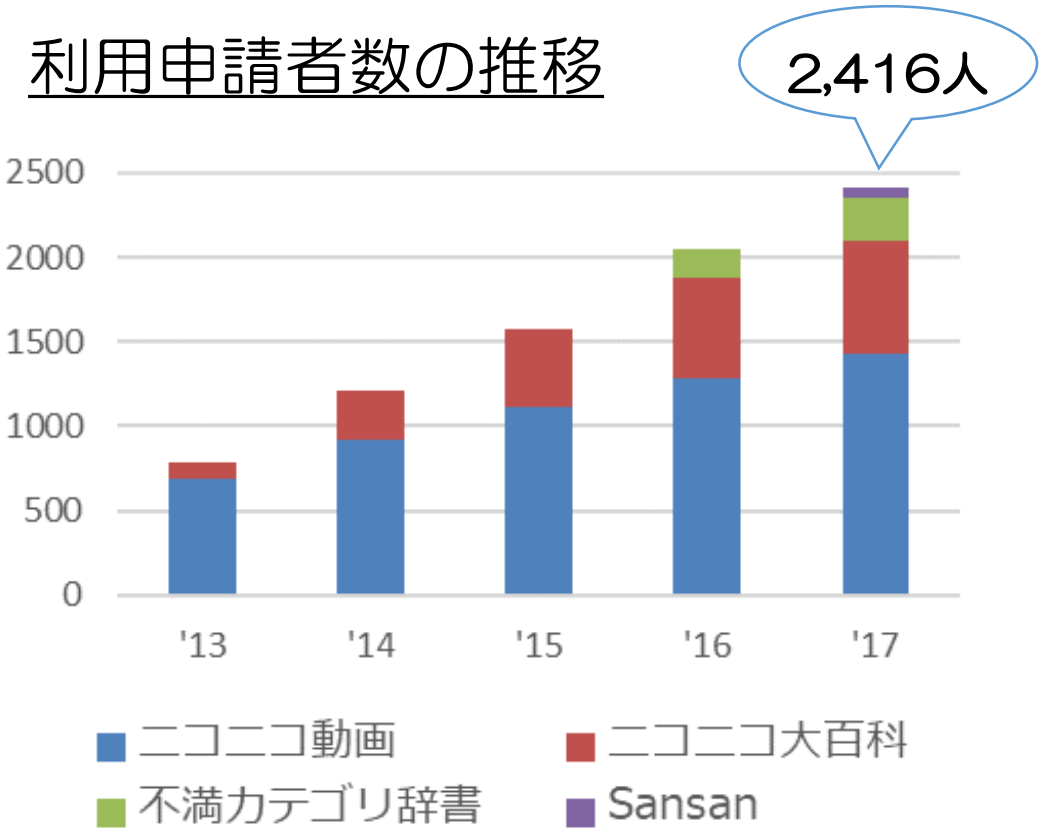


どのくらい使われている？

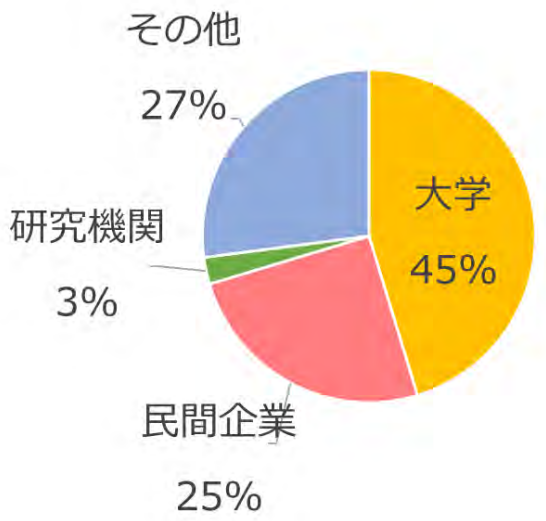


どのくらい使われている？

利用申請者数の推移

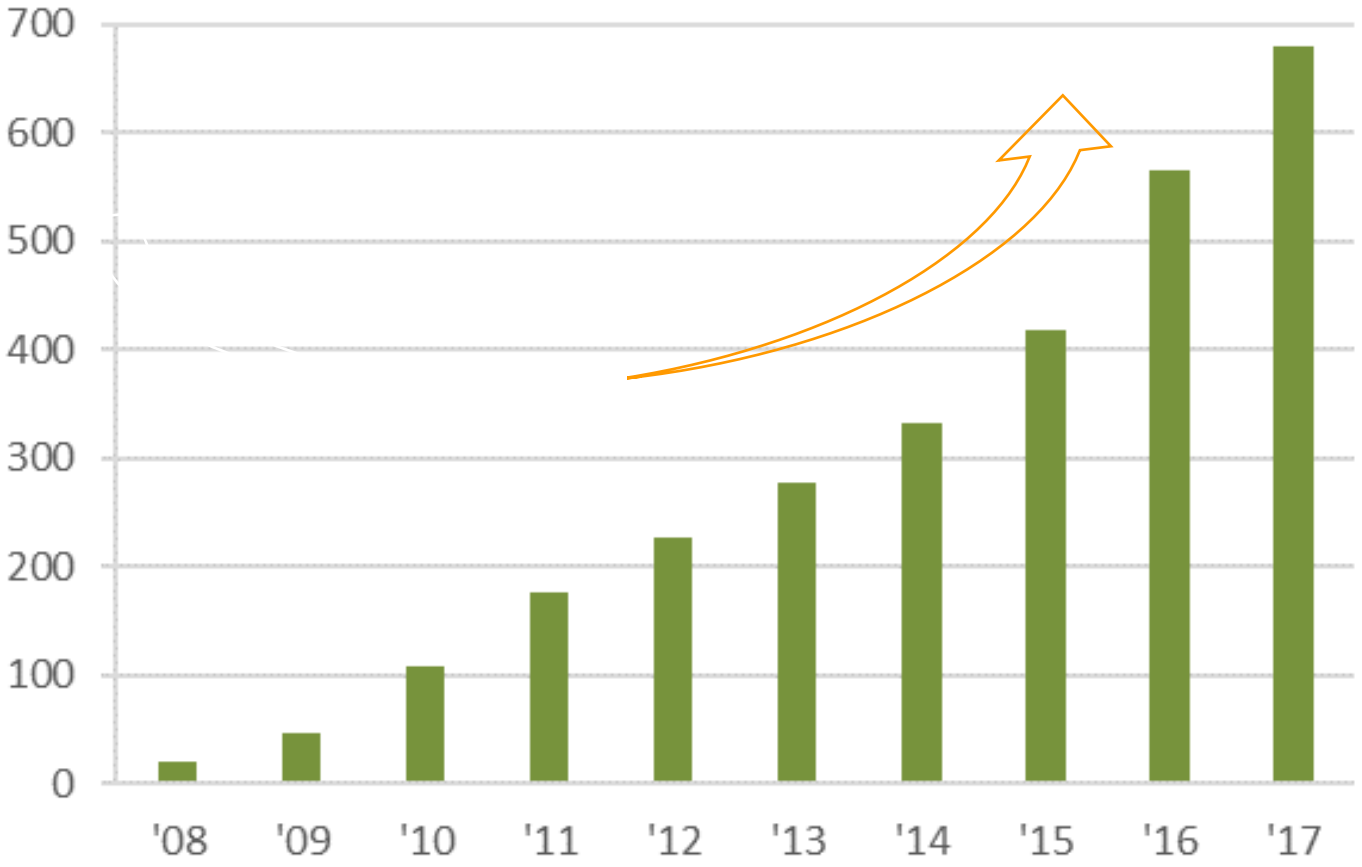


利用者の所属



どのくらい使われている？

研究成果発表数（累積）



研究成果リストは公開しています

https://dsc.repo.nii.ac.jp/

ここからが本題

ユーザのリスクの元は？

ユーザ ＝ データ提供元サービスの利用者本人

- サービス登録情報
  - 自分の個人情報，プロフィール，サービスIDなど
- 自分の投稿
  - 自分の個人情報やプライバシー
  - 公序良俗に反する情報
  - 第三者の権利侵害などの不正行為
- 他人の投稿
  - 自分の個人情報やプライバシー
  - 自分に対する誹謗・中傷
  - 自分の過去の不正行為などの情報

※ユーザでない人にとってのリスクはユーザのリスクに含まれる

ユーザのリスクが実害になるとき

例えばネット上で…

- 人に知られたくない個人情報やプライバシーが公表されたり不正利用されたりする
- 誹謗・中傷が本人に紐付けられて公表される
- 過去の不適切な投稿が本人に紐付けられて公表されたり不利な扱いを受けたりする（自業自得とはいえ）

ユーザのリスクが実害になるとき

思い浮かぶのは…

- Cambridge Analytica（CA）によるFacebookユーザーの個人情報不正利用
  - CAは，大学の研究者がfacebookのアプリにより収集した8700万人分の個人情報を、2016年の米大統領選でトランプ陣営勝利のための有権者操作目的で利用
  - アプリに同意した本人だけでなく友達の情報まで収集できていた
  - 学術研究目的から外れる目的外使用をした

⇨ IDRから提供されるデータで同じようなことはできないのか？

ユーザのリスクが実害になるとき

思い浮かぶのは…

- 企業が就活生の過去のSNSの書込みをチェックしている（らしい）
  - SNSのアカウントの提出を求める企業もあるとか。
  - 就活前にアカウントを削除したり過去の書込みを削除したりと学生も対策を講じているが。

⇨ IDRから提供されるデータは自分では削除できないし，不利な情報が集められると困る。

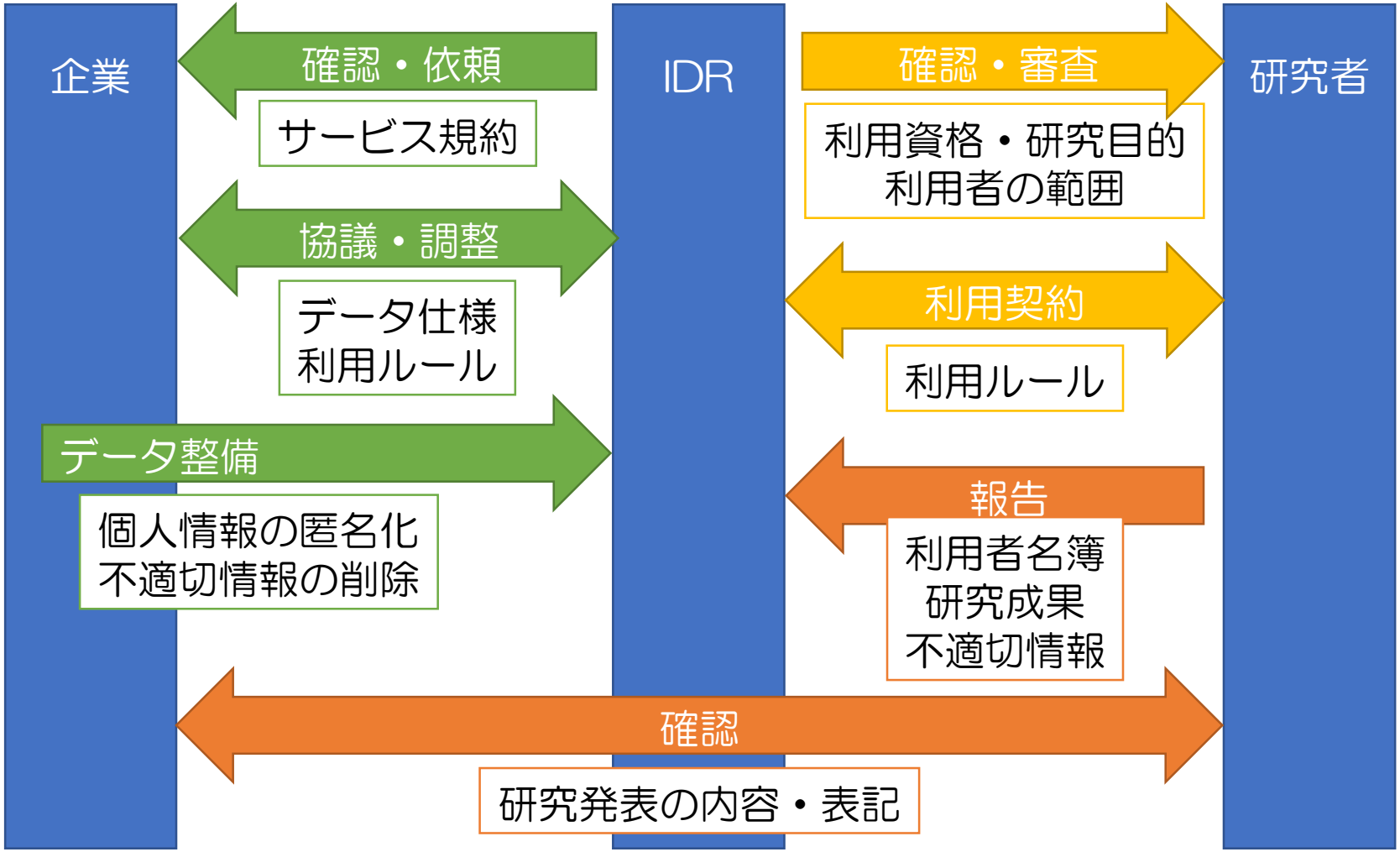
研究者が何をするか心配？

- 研究成果を公表する際，研究者はこのような行為をしかねない？
- このようなことを容易に行えるようにする技術やツールを研究者が開発する？
- そもそも勝手に自分のデータが集約されて自分に紐付けられるのはいやだ？

ユーザのリスクはみんなのリスク

- 企業 ⇨ 社会的な評判が重要
  - ユーザから批判を浴びるようなことは避けたい
  - ユーザに実害が起こらないように研究者の利用をコントロールしたい
- 研究者 ⇨ データが使えなくなるのが困る
  - 自分の行為で企業がデータ提供を止めることになると
    - 他のデータも使わせてもらえなくなるかも
    - 他の研究者にも迷惑がかかり，研究者コミュニティで生きていけない
    + 最近は研究倫理もうるさいので下手をすれば懲戒処分
- IDR ⇨ 他社に影響が及ぶのが怖い
  - 一人でも研究者が問題を起こせばIDRへの信頼が揺らぐ
  - 新たな企業がデータを提供してくれなくなる

そこで企業・IDR・研究者が連携して対策



対策１ – サービス利用規約

- 企業はサービス利用規約や取り組みをわかりやすく的確に
  - 研究目的で大学などの研究者に提供すること
  - 本人がサービスに登録した情報のどれを提供しどれを削除や匿名化するか
  - 不適切な情報の通報とそれに対する対応や，不適切な情報を積極的に除去する取り組みについて
- IDRによる規定の確認
  - NIIを通じて研究者に第三者提供できることになっているか
  - 技術的・コスト的に実現困難な条件となっていないか
  - 研究者にとって有用なデータが提供できるか
  - 不十分な場合は改訂を提案し，改訂後に取得したデータを受け入れることも

対策２ – データ整備

- 本人がサービスに登録した情報から，本人が特定できるデータを削除または匿名化
  - サービス利用規約の規定よりも厳し目にする
- 投稿データ中の個人情報や誹謗中傷などの不適切情報はできるだけ削除
  - 実際は，AI等の技術や専門スタッフによるパトロール等によって，可能な範囲で既に削除されている。
  - 削除したものは研究者には提供しない。
    （研究者としては研究ネタとして欲しくて仕方ないのだが…）

⇨ 提供データからだけでは個人の特定はできない（はず）

- 投稿データを完全にきれいにするのは不可能 → リスクが残る

対策３ – 研究者の管理

- 大学の正規の教員など身元の確かな研究者にのみ提供
  - IDRが利用申請書に基づいて申請者の身元を確認
  - いくつかのステップで本人確認も行っている
- 大学などの機関と利用契約を締結
  - 契約は研究室などの単位で
  - 機関をまたがる共同研究などの場合は機関ごとに契約
  - 利用ルールに違反した場合は所属機関が責任を負う
- 研究代表者が研究グループの名簿を管理
  - 研究グループは研究代表者の監督の下でデータを利用
  - 研究代表者がデータの利用ルールを守らせる

  ※ 特に学生が問題を起こしやすいので注意

対策４ – データの利用ルール

- 企業とIDRが相談してデータの利用ルールを決める
  - 無断で第三者にデータの提供や開示をしない
    - デモであっても第三者が操作してデータをそのまま見られるものは禁止
  - 個人を特定できる情報などは研究成果の公表であっても開示しない
  - データから個人を特定する行為や特定につながる行為を行わない

対策５ – フォローアップ

- 利用者は利用報告を毎年提出
  - 利用状況，研究成果（論文リスト）
  - 研究代表者の異動，研究グループ名簿

  ※　研究代表者が大学などを異動したら再契約

  ※　データ利用を終了する場合はデータ削除確認書を提出

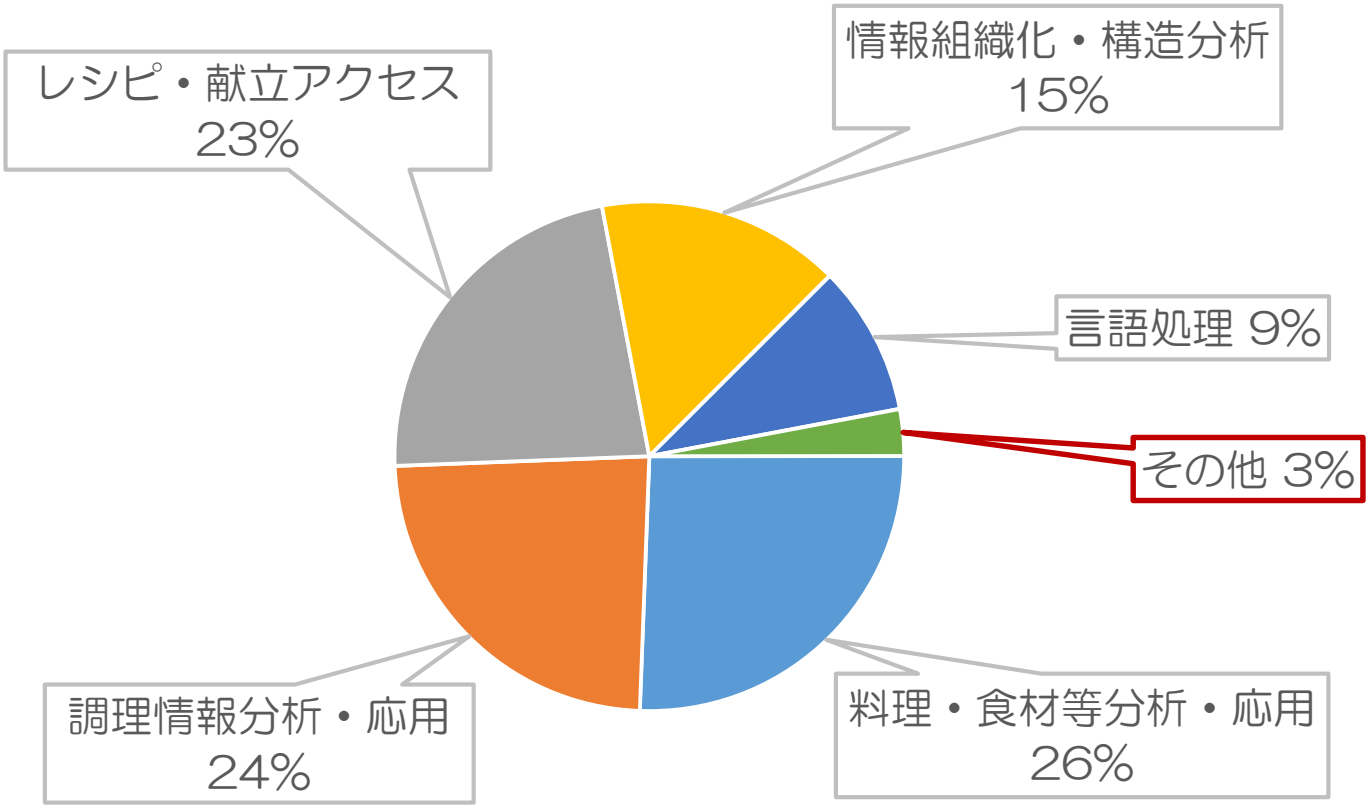
  ⇨ データの所在を最後まで追跡
- 研究発表の際は原稿等を予め企業に提出
  - 不適切なデータ利用や情報開示がないか企業がチェック
- 利用者は不適切な情報を見つけたらIDRに報告
  - その情報は見つけたこと自体も含めて秘密扱い
  - 企業がデータ差し替えなどの措置をし，利用者はその指示に従う

これまで問題があった？

研究成果を分析してみました。

料理レシピデータを使った研究

- 楽天レシピデータ+クックパッドデータ

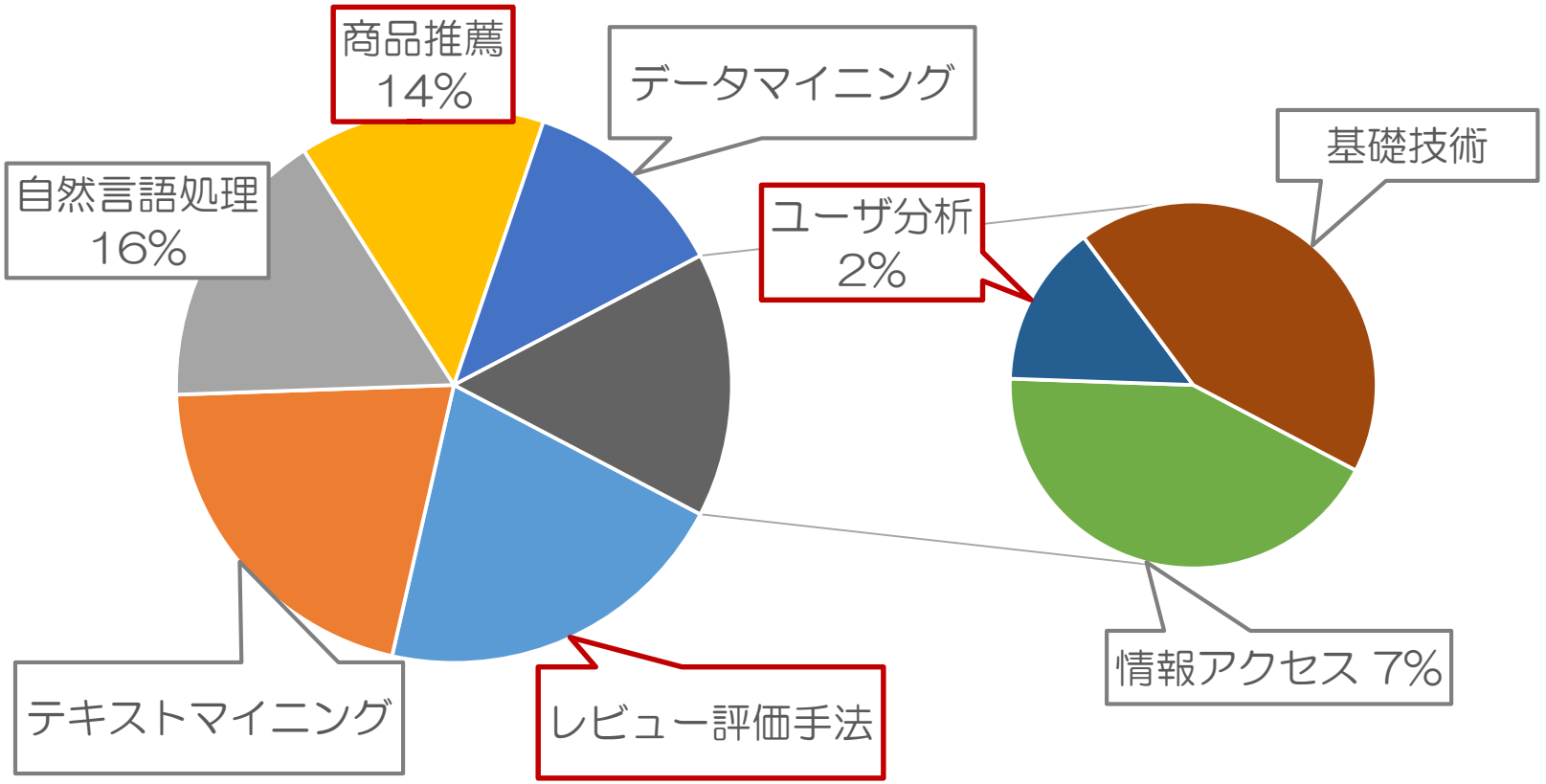


料理レシピデータを使った研究

- 大部分の研究はユーザ個人とは関係ない
- 「その他」が気になる？
  - １件はシステム技術，１件は料理画像生成
  - １件は全国の食品消費量分析への応用
  - ２件は料理の嗜好の分析 – もしかしたら？
- タイトルを見るとちょっと気になる研究も
  - 「レシピサイトにおける提供者と使用者の嗜好抽出と可視化」

ショッピングサイトデータを使った研究

- 楽天市場データ



ショッピングサイトデータを使った研究

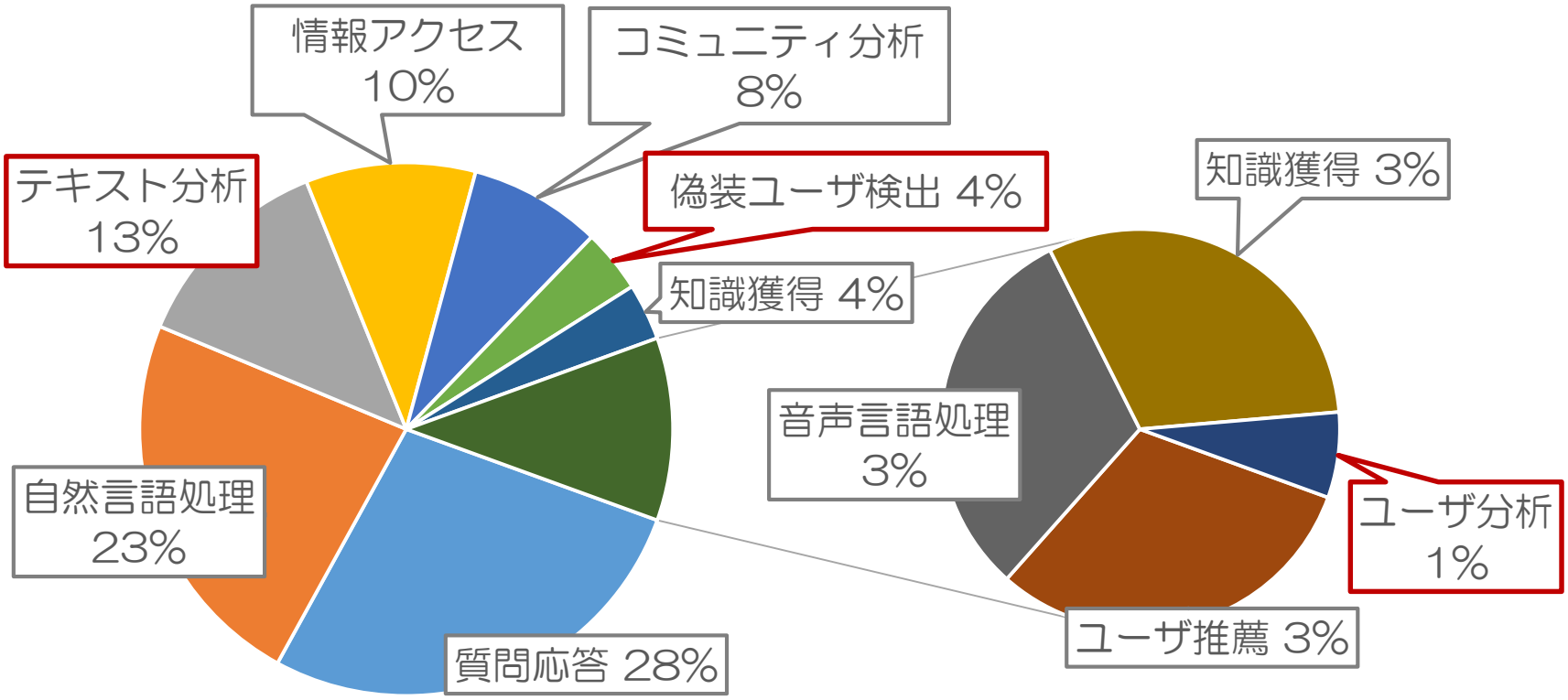
- レビュー（口コミ）を使った研究が気になる？
  - レビュー評価手法やテキストマイニング
    - 統計的な手法が多い … 安心
  - 自然言語処理
    - 技術が中心 … 個々のユーザには関係ない
- レビューの投稿が広告の表示に使われない？
  - 商品推薦
    - 楽天市場データは商品の分析に使われるだけ
- レビュー以外のデータと組み合わせたら？
  - データマイニング
    - ほとんどは統計的な手法なので問題ない

実際はどんな研究？ – 楽天市場の場合

- タイトルだけ見るとちょっと気になる研究も
  - 商品推薦
    - 「マイクロブログの分析に基づくユーザの嗜好とタイミングを考慮した情報推薦手法の提案」など
  - ユーザ分析（手法としてはデータマイニング）
    - 「レビュアー属性・時系列要因とレビュー行動」
    - 「レビュー順序グラフに基づく購買行動パターンの分析」

Q&Aサイトデータを使った研究

- ヤフー知恵袋データ



Q&Aサイトデータを使った研究

- 気になる研究がある？
  - テキスト分析
    - 「うつ病についての関心の推移」
    - 「民事紛争のウェブ相談における感情」
    - 「『契約・解約』に関する消費者トラブル相談事例の分類と分析」
    - 「『Yahoo!知恵袋』に見る夫婦間葛藤解決方略」
    - 「インターネットに求める子育ての悩み」

    など

    ⇨ 多くは特定のトピックについての統計的な分析で，個別の質問内容の記載や具体的なユーザへの言及はない

Q&Aサイトデータを使った研究

- 気になる研究が結構ある？
  - ユーザ分析
    - 「Q&Aサイトにおけるユーザの要求・関心の時空間的な推移の可視化」
    - 「An Analysis of Users in a Q&A Site Submitted Many Answers Where First Polar Words are Negative Words」

    ⇨ 多数のユーザに関する統計的な分析で，個別の質問内容の記載や具体的なユーザへの言及はない

Q&Aサイトデータを使った研究

- 気になる研究が結構ある？
  - 偽装ユーザ検出
    - 「Q&Aサイトで複数のアカウントを不正に用いるユーザの検出」

    など（全て同一の研究グループ）

    ⇨ ユーザ番号は匿名化されているため個人は特定できない

これまで問題があった？

幸いなことに，ユーザに不安を与えるような問題は，これまで一度も起こっていません。

これからも大丈夫なの？

- 研究の深化にはよりディープなデータが必要
  - 時系列データ，トランザクション，ユーザログ
  - 応用分野の需要に応えられる詳細なデータ
    - 栄養学，保健学，経済学，環境科学…

⇨ 利用目的や利用方法，データセキュリティなどに，より厳密な管理が必要

- 制度的・技術的な解決策を一緒に考えていくことが必要
- ユーザへの情報開示と理解の増進が重要になる

まとめ

- 実際に役立ち，あるいは実社会を理解する研究には大規模リアルデータが必要
- ユーザも企業も安心できるデータの研究利用には「共同利用」が有効
- 企業，IDR，研究者が連携して対策することによって安全を確保
- 今後，研究を深化するためにはリアルデータの共同利用がますます重要に