

産学連携

～企業データセットの公開意義とデータセットの更なる研究活用に向けて～

★ 田中克己(京都大学)
岡本 真(アカデミック・リソース・ガイド株式会社)

目次

1. 自己紹介
2. 企業データセットの学術利用～過去11年を振り返る～
3. 企業データセットの公開意義
4. 企業データセットの更なる研究活用に向けて
5. 質疑応答

自己紹介・田中克己(京都大学)

◆現在

1. 京都大学 名誉教授 情報学研究科(2017年～)
 - 京都大学 教授 情報学研究科社会情報学専攻(2001～2016年度)
2. 関西学院大学 社会情報学研究センター 顧問(2017年～)
3. ウィーン大学 客員教授(2017年)
4. 関西大学 客員教授(2018年～) 総合情報学部
5. 日本学術会議 連携会員 情報学委員会(2006年～)



自己紹介・田中克己(京都大学)



◆ 過去

- 1951年生まれ(66歳)
- 京大・工・情報工学科(1970年～1974年)
京大・院・工学研究科(修士:1974年～1976年、博士:1976年～1979年中退)、京大工博(1981年)
- 神戸大・教養部・助手 情報科学教室(1979年～1985年)
南カリフォルニア大・計算機科学科・客員研究員・講師(1982年～1984年)
神戸大・工・計測工学科(情報知能工学科)・助教授(1986年～1994年)、教授(1994年～2001年)
- 神戸マルチメディア・インターネット協議会会長(1998年～2000年)
通信・放送機構「次世代デジタル映像通信の研究開発」統括責任者(1996年～1998年度)
IPA未踏ソフトウェア創造事業プロジェクトマネージャー
情報通信研究機構(NICT)メディアインタラクショングループリーダー、情報信憑性検証技術研究代表者
- ACM TODS AREA EDITOR、情報処理学会 DBS研究会主査、理事、フェロー、日本データベース学会副会長など

自己紹介・岡本真 (アカデミック・リソース・ガイド株式会社)

◆ 現在

1. アカデミック・リソース・ガイド株式会社(2009年～)
 - 代表取締役／プロデューサー(2009年～)
 - ACADEMIC RESOURCE GUIDE編集長(1998年～)
2. オーマ株式会社(2008年～)
 - 取締役(2011年～)
3. SAVEMLAKプロジェクト(2011年～)
 - プロジェクトリーダー(2011年～)
4. NPO法人横浜コミュニティデザイン・ラボ(2002年～)
 - 理事(2010年～)



自己紹介・岡本真 (アカデミック・リソース・ガイド株式会社)

◆ 過去

- 1973年生まれ
 - 44歳
- 国際基督教大学(ICU)卒業(1997年)
 - 日本政治思想史を専攻
- 教育雑誌、学術書等の編集者等を経て、
- ヤフー株式会社に在籍(1999年～2009年)
 - YAHOO!カテゴリ、YAHOO!検索、YAHOO!知恵袋、YAHOO!検索ランキング、YAHOO!ラボ等の企画・設計・運用、また産学連携のコーディネートに従事
- 日本データベース学会理事(2012年～2016年)

2004年、総務省のデジタル資産活用戦略会議におけるウェブ情報利活用ワーキンググループで田中先生と出会う以降、気づけば2007年度に始まった情報信憑性研究に参画し、田中研究室の非常勤研究員に



企業データセットの学術利用

～過去11年を振り返る～

企業データセットの学術利用

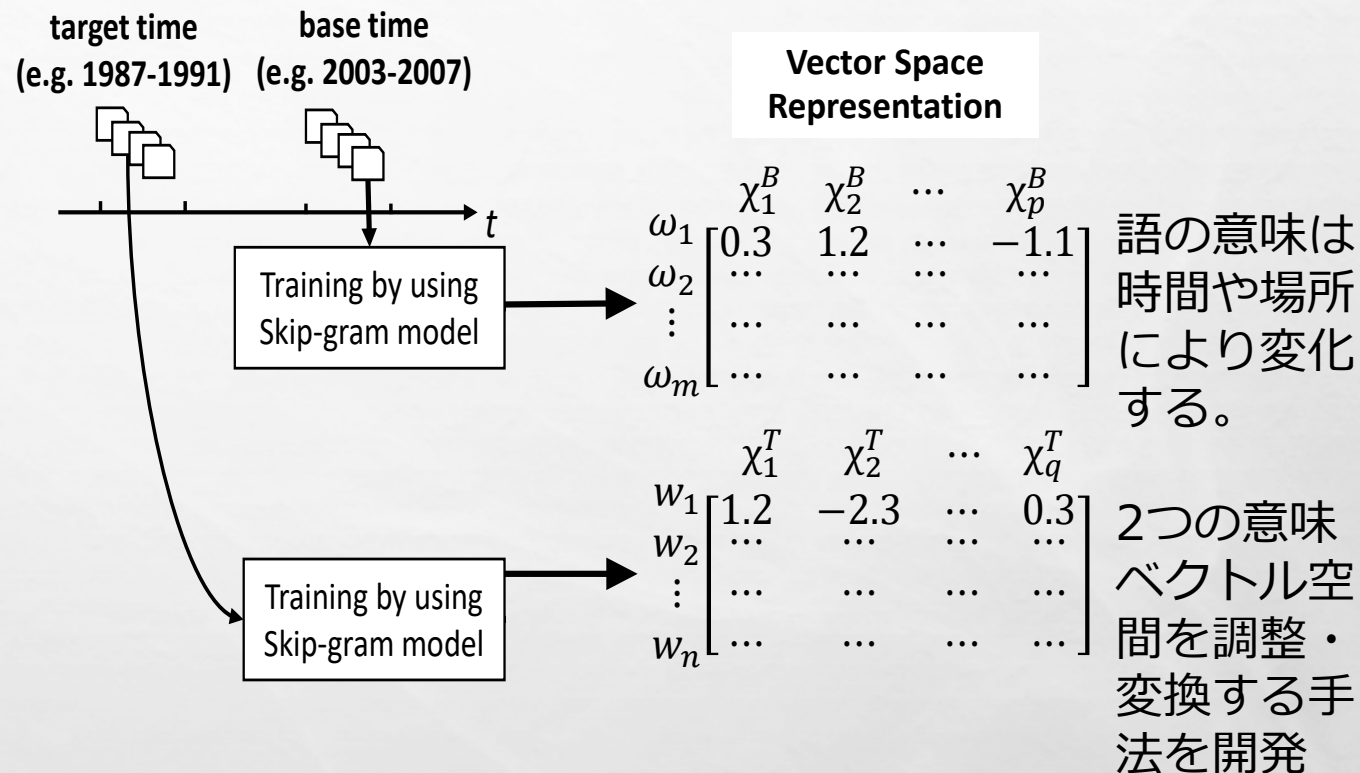
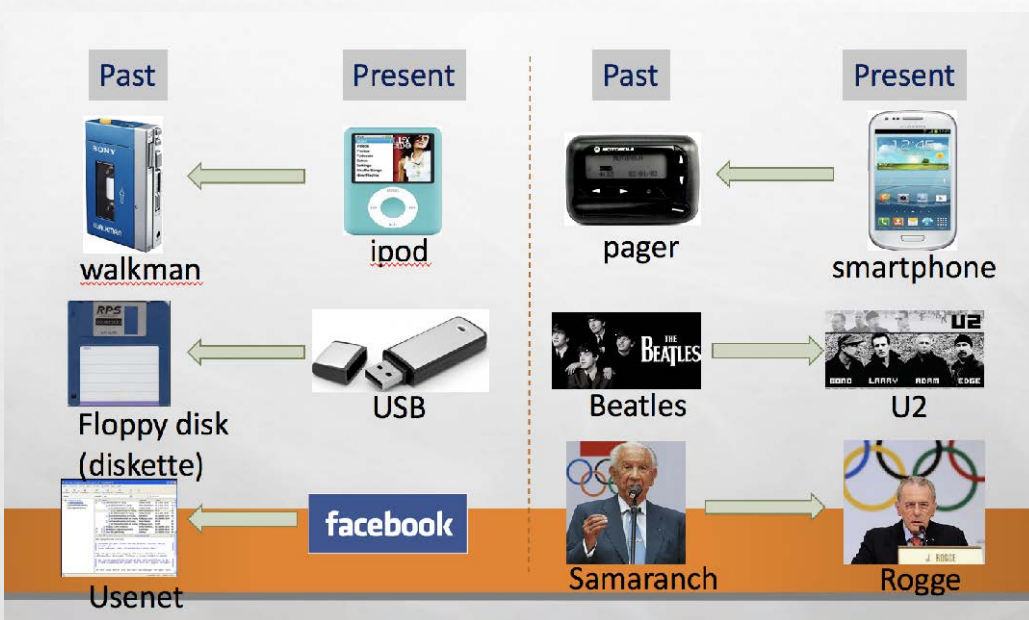
1. 企業データ公開前のアカデミアの現状

- ① 検索エンジンの台頭に伴い、Web情報検索、Webマイニングがアカデミアでも重要な研究課題になってきていた(WWW、ACM SIGIR、ACM CIKM、ACM WSDMなど)。
- ② 検索エンジンの検索性能の向上を目指す研究では、実際規模のクエリログやクリックスルーデータが必要となっていた。
- ③ 情報検索の性能評価を行うのに必要となる実験データセットは作られてきたが、その規模や内容に問題があった(適合性に限定し過ぎた人工的な実験データセット)。クラウドソーシングで実験データセットを作成する試みも多く行われたが、やはり、規模・品質に問題があった。
- ④ Web検索エンジンの次のものとして「人に聞く」検索エンジンが注目された。そのベースになるデータはYahoo!知恵袋などのQAデータ。

企業データセットの学術利用

2. Yahoo!知恵袋のデータ公開をするに至った経緯
 - Web情報の信憑性分析プロジェクト
情報通信研究機構(NICT)情報信憑性検証技術
 - 民間企業(Yahoo!Japan)と大学との共同研究
研究費を出すのではなくデータを提供する共同研究
3. 企業データ公開に至るまでの障壁
 - クエリログデータ提供依頼(複数社)
クエリログに含まれる個人情報(「少年Aの実名」など)の排除が困難

時空間を超えた類似エンティティ検索



スキップグラムニューラルネットで学習

New York Times のコーパスデータ(1987-2007) <https://catalog.ldc.upenn.edu/Ldc2008t19>

180万記事(全体)の内の各45万記事データを学習。語数 36万語。

Amazon Review Dataset (SNAP Stanford) <https://snap.stanford.edu/data/web-Amazon.html>

250万件の商品 (1995~2013)の記述・レビュー・分類データ、3400万件の商品レビューデータ

UGC(ユーザ投稿型コンテンツ)の信憑性分析とユーザ分析

重要ながらも衰退しつつあるcQA(質問応答サイト)

日本国内最大の質問応答サイト

YAHOO! 知恵袋
JAPAN

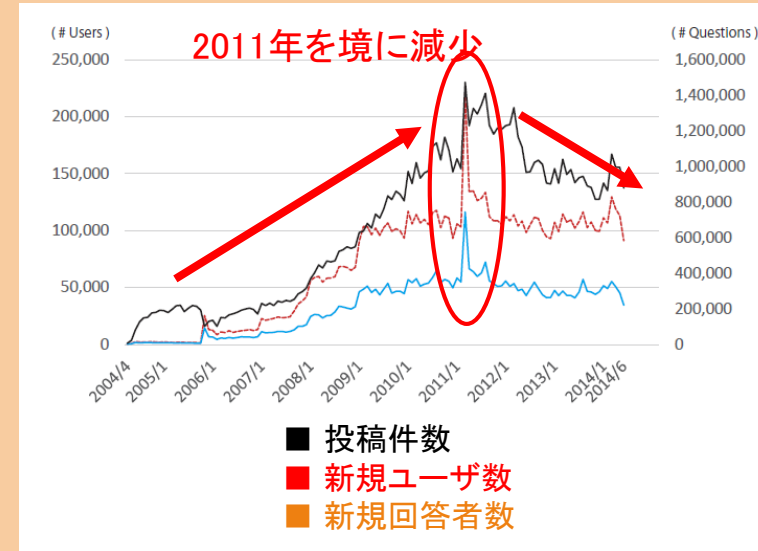
2014の年間利用件数

投稿数 35,000,000 件

訪問数 10,368,000,000 回

閲覧者 > 利用者

質問者・回答者のみならず
閲覧者にとって貴重な情報源



新規ユーザ数の減少
(特に回答者)

優れた情報源であるcQAからのユーザ流出を防ぐため

➡ Yahoo! Japan と共同研究で**10年分**の実際の Yahoo! 知恵袋のログを解析
長くサービス利用を続けるユーザの特徴を分析

UGC(ユーザ投稿型コンテンツ)の信憑性分析

テキスト信憑性プラットフォーム - Windows Internet Explorer

http://localhost:53053/TextCredibilityPlatform/Chiebukuro.aspx?q=10671145&base_q=%E3%9C%9C%E3%9C%9C%E3%9C%9C

知恵袋検索語: テーブルマナー

質問

スパゲティを
スプーンを使って食べる
のはテーブルマナーか？

テーブルマナーのスパゲティとスプーンの関係について。イタリアを含む欧米で、ス...

質問日時: 2004/09/02 22:42:32 解決日時: 2004/09/05 0:45:44

回答数: 10 閲覧数: 189

解決済みの質問

1~10 件目 / 約 10 件 - 4.01 秒

1 ID非公開さん

スパゲティをスプーンも使って頂くのは、やはり日本的、日本特有の食べ方です。アメリカは分かりますが、ヨーロッパでは幼稚園に見られる光景かな。私個人的には、別スプーンの上で麺を巻かなくても皿の中でずりゃいやるかな。この人知らないんだなあ、と思います。スプーンがダシになって、自分の食べやすい形で良いとは思いますが、ライスをフォークの先にのせて頂くのは、最近同様の質問をみましたが、英国ではそれが通じません。日本でもその流儀が若い方より年配の方がやっているのをよく見かけます。これを実用性で欠けると思うので、マナー違反ではないものの、見ててどうかあ？って思います。楽しく食べるのが大切なので、食事の相手がこの2点をしていたら、私はちょっと恥ずかしいです...

2004/09/03 1:18:40

2 ID非公開さん

あの食べ方は欧米の子供たちに見られる食べ方です。大人はそんな食べ方しません。今もパスタを食べに行くときスプーンとフォークを使う人を見て違和感を感じています。でも確かにパスタをまきつける方向は決まっていたよね。左回ったような記憶がありますが間違っていたらごめんなさい。ライスは、背中に乗せて...はテーブルマナーの本に書いてありましたが普通に食べてもかまいません。豆類など乾かす/やすいものはフォークで潰して平たくしてからというふうなようですが、よほどマナーを無視した食べ方であれば特に気にしないでいいと思います。回答番号: 3535769さん。『スプーンと呼んでも間違いないんですよ？知りませんでしたか？』

2004/09/03 8:14:09

3 ID非公開さん

「スプーン」というのはあなたもどうかな、と.....

2004/09/02 23:45:24

メジャー度 92

ですます/である判定 ですます調

話題網羅度 45

話題専門度 14

メジャー度 86

ですます/である判定 ですます調

話題網羅度 45

話題専門度 20

メジャー度 43

ですます/である判定 判定不能

話題網羅度 0

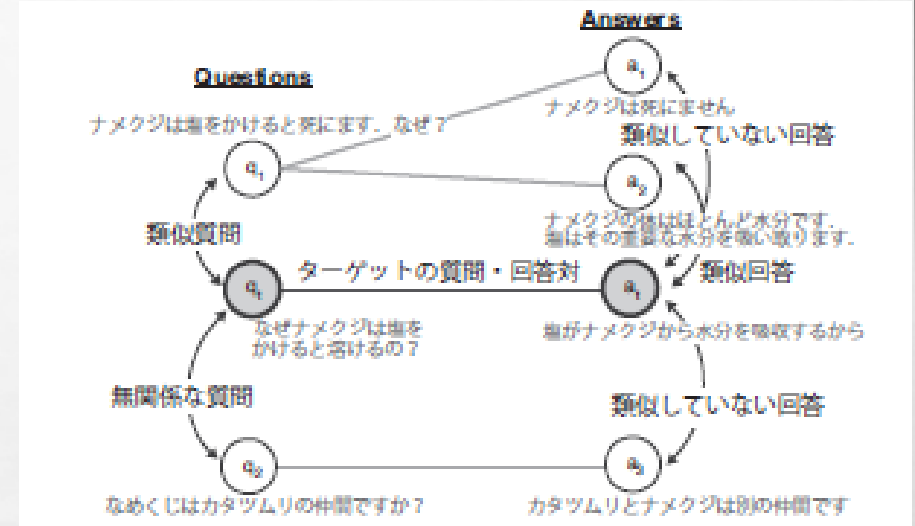
話題専門度 0

回答内容が
多数派か

信頼出来る
文体か

回答文が重要
な話題を網羅
しているか

専門語を多く
含んでいるか



類似質問に対する回答が似ているものが多ければ、その質問・回答のペアは信頼できる

質問「なぜナメクジは塩をかけると溶けるのか」

回答「塩がナメクジから水を吸収するため」

類似質問

「ナメクジは塩をかけると死ぬのはなぜ」

類似回答

「ナメクジの体のほとんどは水。塩がその水を吸収するから」

すぐやめた人と長く続けた人で何が違う？

2004 - 2014 (10年分)のYahoo!知恵袋の全QA・ユーザデータ

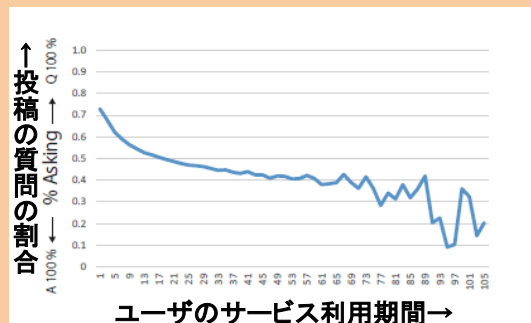
質問: 84,123,965 件
タイトル, 本文, カテゴリ, 日時, BA, など

回答: 224,969,887 件
対応する質問ID, 本文, 日時, BA, etc...

ユーザ: 10,391,194 人
匿名化されたID, 任意入力のプロフィール項目類, etc...

知見1 投稿内容

QA比による比較



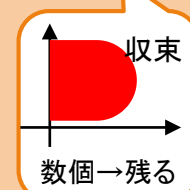
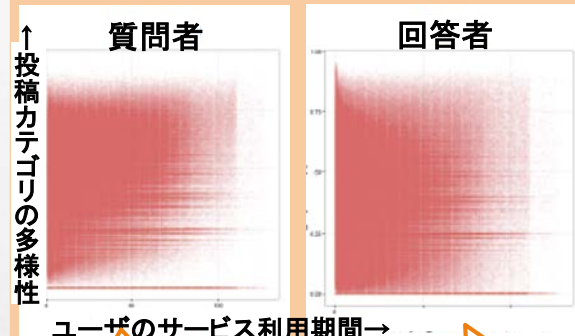
すぐやめたユーザ
(利用期間1~5ヶ月)
は投稿の7割が質問

長く残ったユーザ
(利用期間2年~)
は半数以上が回答

長く残るユーザは
質問より回答を好む

知見2 カテゴリ

エントロピーによる比較



長く残るユーザは
多様なカテゴリで質問し
いくつかのカテゴリで回答

知見3 報酬・満足度

BAの付与・獲得による比較
(ベストアンサー)



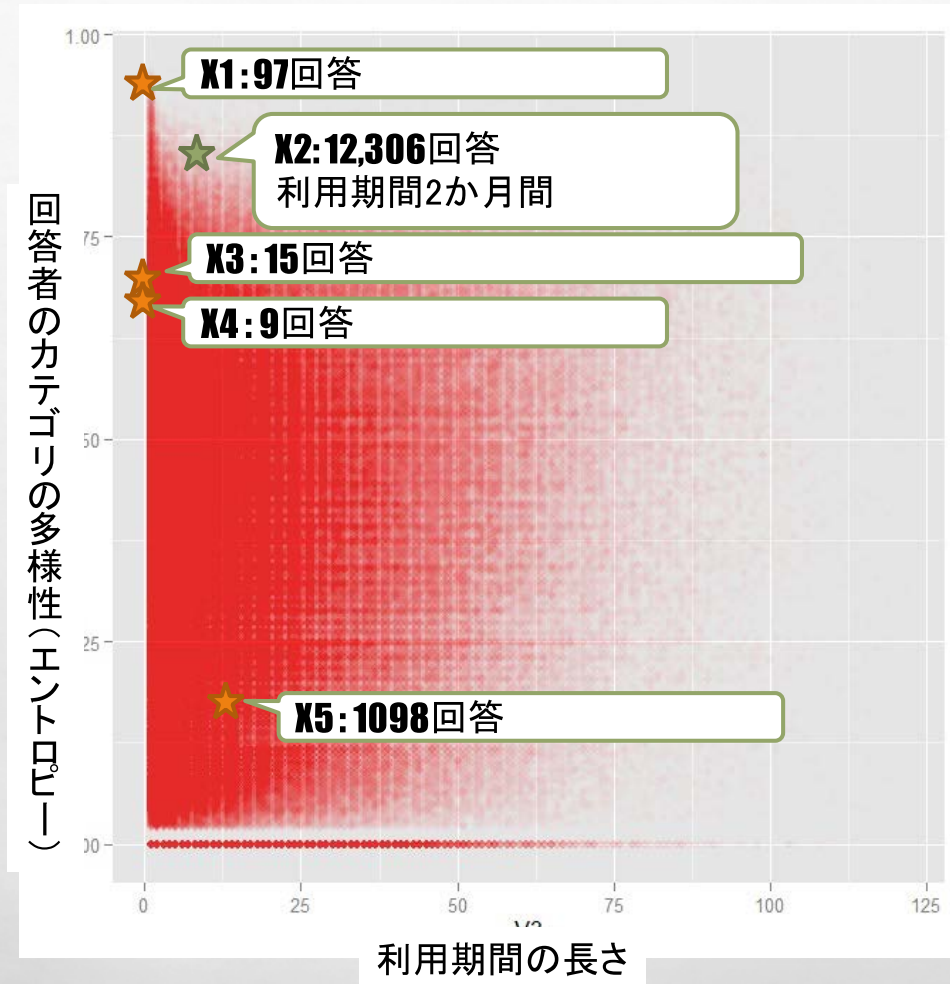
よい回答を多く
もらった質問者は
長く残る



長く残る回答者は
報酬に釣られない

回答中心で、多くのカテゴリで質問、いくつかのカテゴリで回答をし、
BA率をあまり気にしないユーザがサービスに長く残る

回答者の品質分析: 悪質(スパマー)回答者Xの発見



ユーザレビューデータの信憑性分析

国内最大の
映画レビュー
コミュニティ

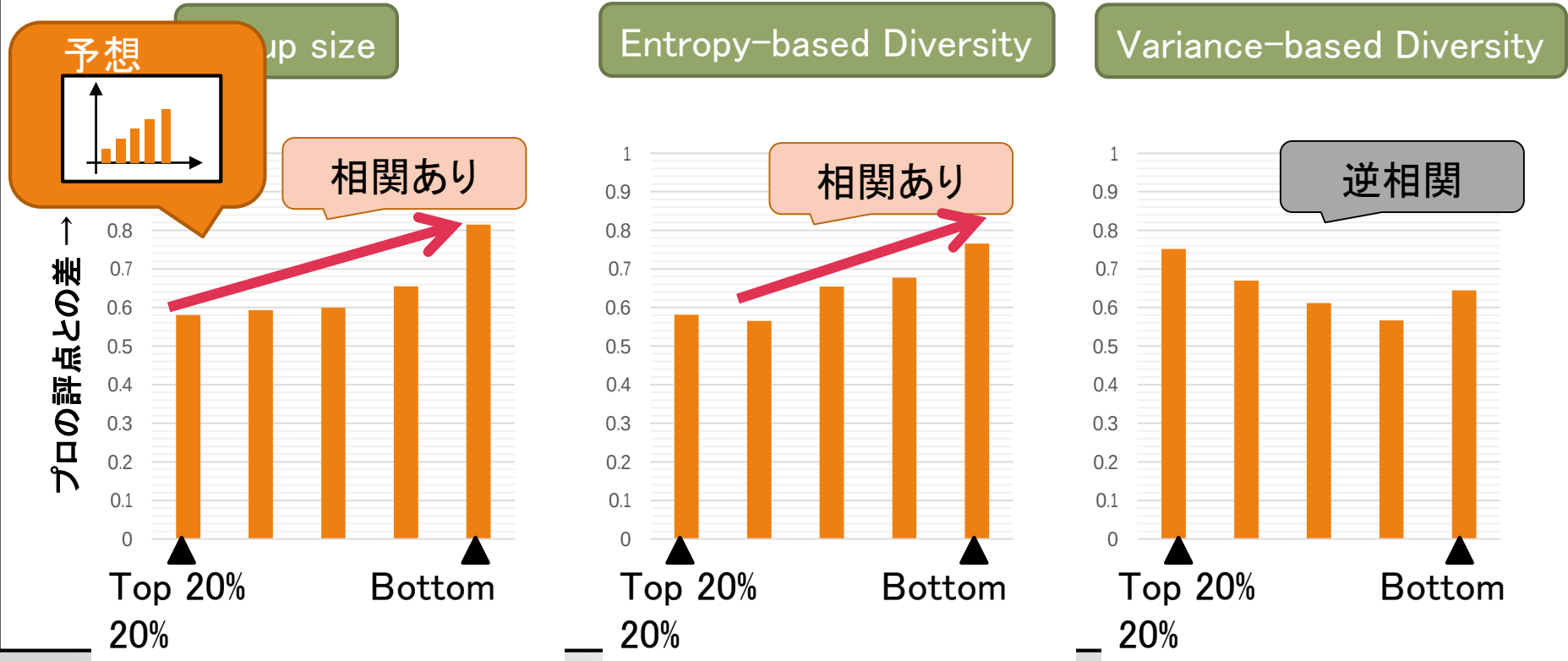
YAHOO! 映画
JAPAN

データ

映画

27,516 本
レビュアー 数
158,385 人
レビュー件数
1,124,555 件

100人以下のグループが評価した映画をプロット



グループの人数が十分に大きくない場合、そもそもグループの信頼度は高くないが、
グループの人数と *Entropy-based Diversity* はグループの信頼度と相関を持つ。
⇒グループが小さく、かつ *Entropy* 的に非多様な際、そのグループは信頼できない

UGC(ユーザ投稿型コンテンツ)の信憑性分析

「本格」お店気分な我が家のカレー

名前に付けられた修飾表現が、内容を表すのに相応しくないものも多い(=誇張)



内容:肉、ジャガイモ、人参、玉ねぎ、カレールー、...

あまり本格的ではない?

名前に“本格”を含むカレー

ターメリック
0.25
ガラムマサラ
0.41

名前に“本格”を含まないカレー

ターメリック
0.07
ガラムマサラ
0.15

クックパッドデータ
料理レシピで頻出する
修飾表現(全43万件)

誇張の割合

修飾表現	割合	修飾表現	割合
甘い	0.4	和風	0.11
簡単	0.29	ジューシー	0.10
定番	0.26	夏	0.10
アッサリ	0.23	シットリ	0.10
節約	0.20	中華	0.07
本格	0.17	辛い	0.07
シンプル	0.17	シャキシャキ	0.04
コッテリ	0.13	サッパリ	0.03
ヘルシー	0.11	カラフル	0.00

隠れ適合の割合

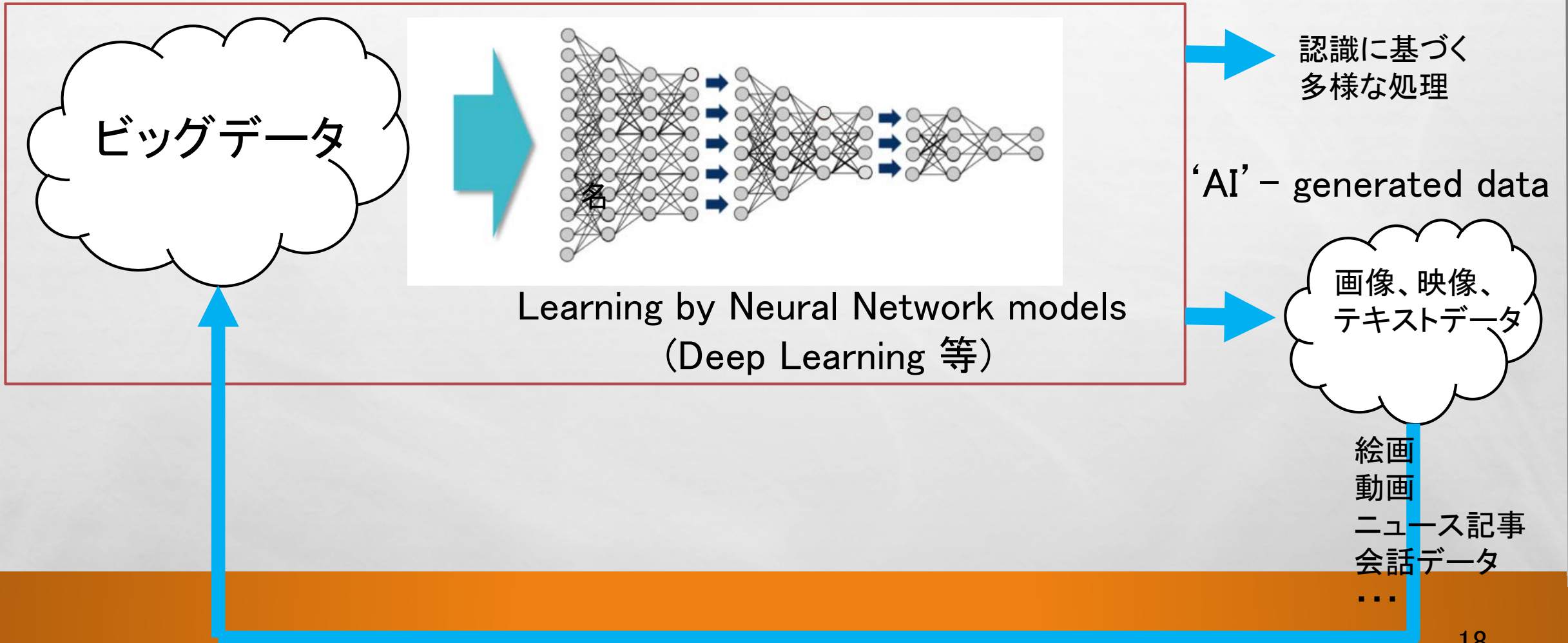
修飾表現	割合	修飾表現	割合
シットリ	0.70	シンプル	0.51
コッテリ	0.67	カラフル	0.50
ヘルシー	0.66	和風	0.49
節約	0.64	定番	0.46
アッサリ	0.63	ジューシー	0.40
サッパリ	0.63	夏	0.40
簡単	0.57	中華	0.37
シャキシャキ	0.56	辛い	0.37
本格	0.51	甘い	0.36

企業データセットの公開意義

- 「公開」ではなく「提供」?!
- 企業が提供している「情報サービス」の品質改善や新しい情報サービスの実現
- ビッグデータの機械学習による新しい情報サービスの実現
 - 公開・提供しなくても企業が自前で行う可能性も
 - アルゴリズムは公開、データは公開しない

‘AI’ が新たなビッグデータを産み出す

‘AI’



企業データセットの更なる 研究活用に向けて

- 産学連携2.0 OR MORE
 - データセット提供にもとづく新しい産学連携研究
 - 企業と大学の間での人材の双方向の派遣
 - 大型研究プロジェクトへの企業の参画
- 大学における「民間との共同研究」「委託研究」の枠組みを変える必要

質疑応答