Japan Open Science Summit 2018

民間企業データによる オープンサイエンスの可能性
- NTCIRでの経験から -

京都大学 国際高等教育院
附属データ科学イノベーション教育研究センター
特定講師 加藤 誠
Makoto P. Kato

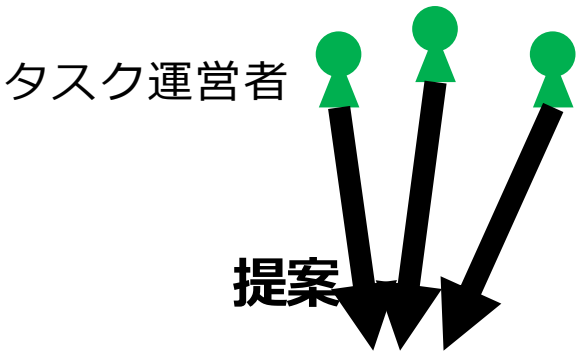
自己紹介

- 加藤 誠 (Makoto P. Kato)
- 研究
  - **情報検索**，データマイニング，機械学習
- 経験
  - **NTCIR：**
    研究者コミュニティでタスク設計・データセット構築・評価を行う試み
    - NTCIR-9 (2010-2011年): 1CLICK, INTENTタスク運営者
    - NTCIR-10 (2012-2013年): 1CLICK, INTENTタスク運営者
    - NTCIR-11 (2013-2014年): MobileClick, IMineタスク運営者
    - NTCIR-12 (2015-2016年): MobileClick, IMineタスク運営者
    - NTCIR-13 (2016-2017年): OpenLiveQタスク運営者
    - NTCIR-14 (2018年-): OpenLiveQタスク運営者

プログラム委員長

- **アジア版TREC**
  - ややNTCIRより
- **アカデミック版Kaggle**
  - 賞金はでません
- **より多様な (KDD|WSDM) Cup**
  - 1.5年で6-8タスクが提案

NTCIRの仕組み

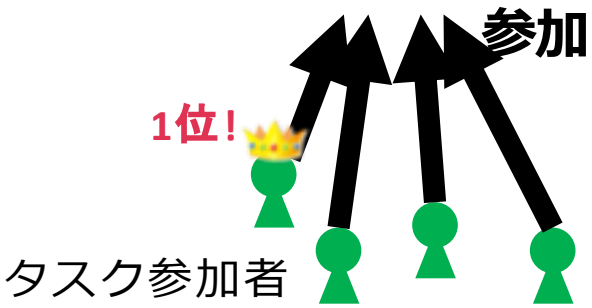


1. 自由にタスクを設定する
2. データセットを構築する
3. 参加者にシステムを作ってもらう
4. 参加者のシステムを評価する
5. 知見を得て論文を書く

OpenLiveQ タスク

WIN-WIN

運営者： 自分で問題を解かなくてもその分野が発展
参加者： 問題解決に注力できる



1. 好きなタスクを選ぶ
2. データセットをダウンロードしてタスクの要件にあったシステムを作る
3. タスク運営者に評価してもらう
4. 知見を得て論文を書く

NTCIRタスク

ソーシャルメディア

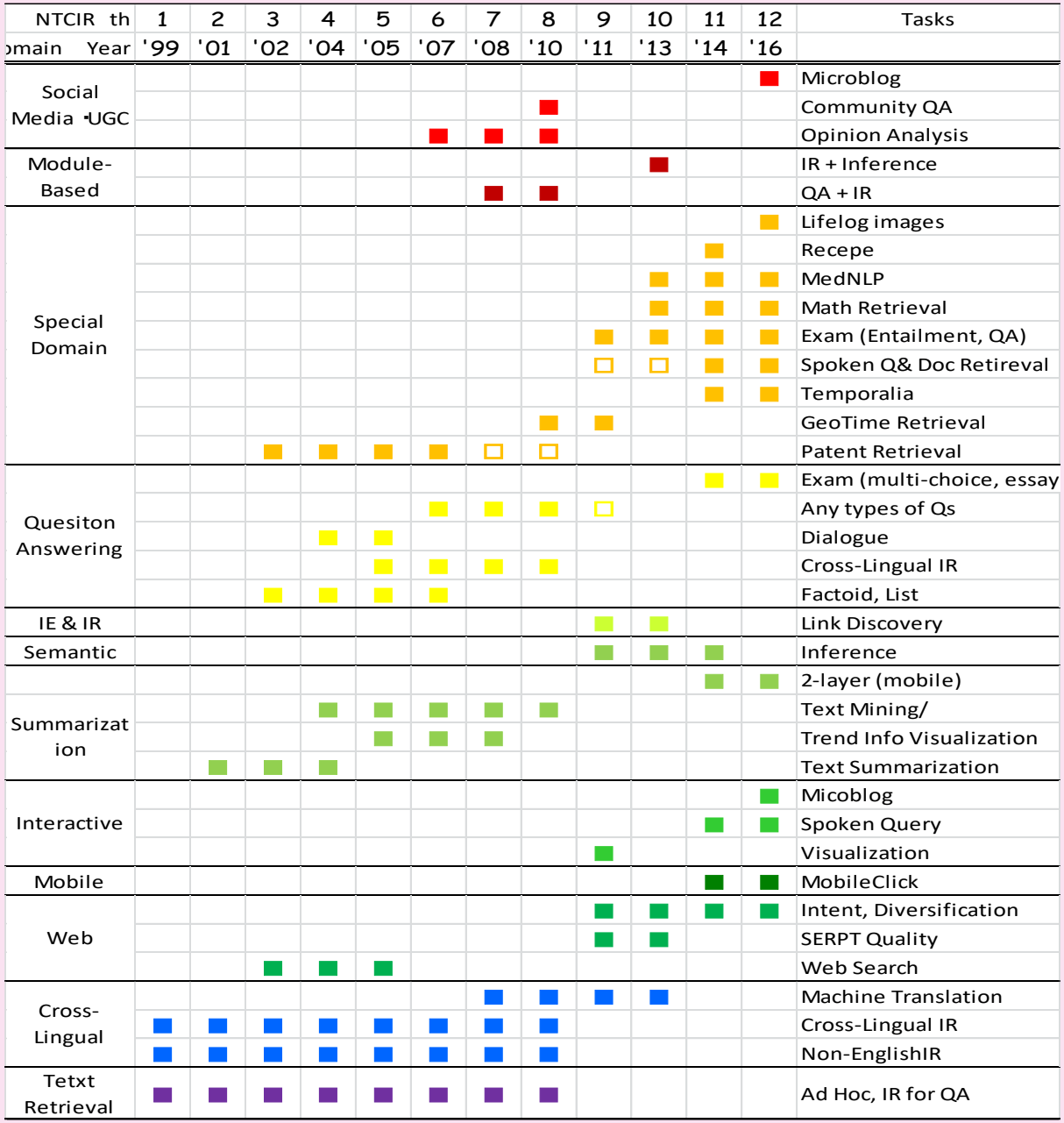
ライフログ・医療・特許

質問応答

要約

Web検索

言語横断



NTCIR-14 参加募集中

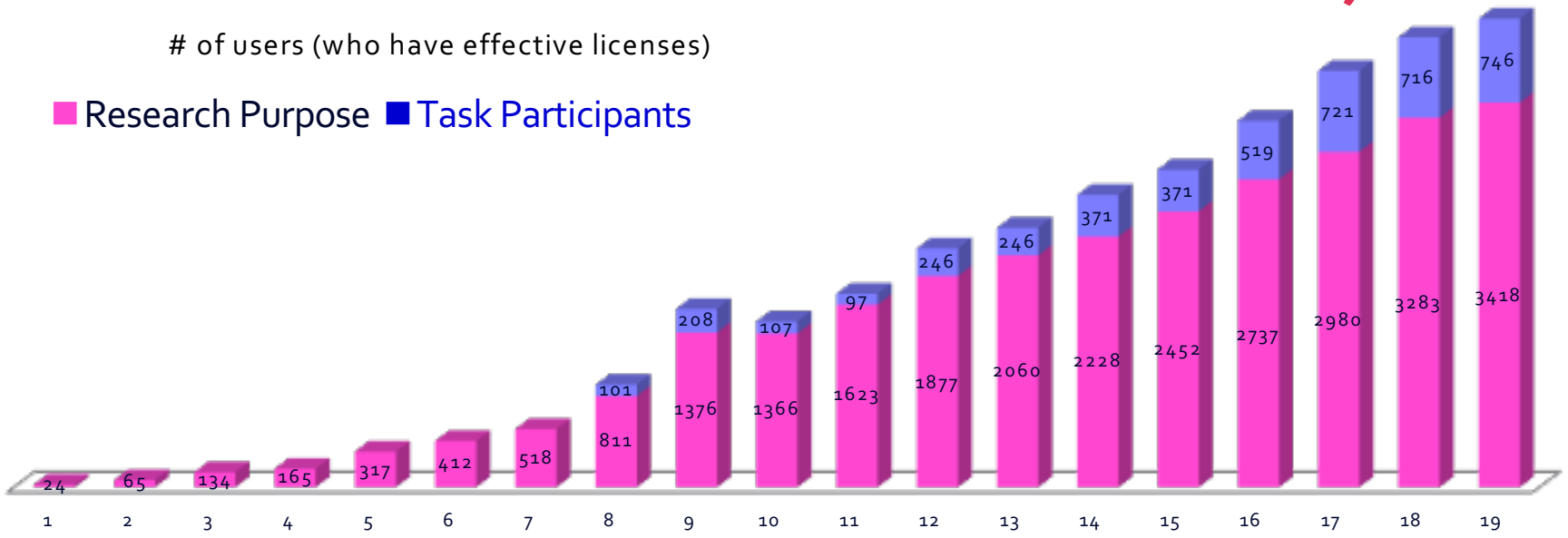
期間： 2018年1月 – 2019年6月

http://research.nii.ac.jp/ntcir/ntcir-14/

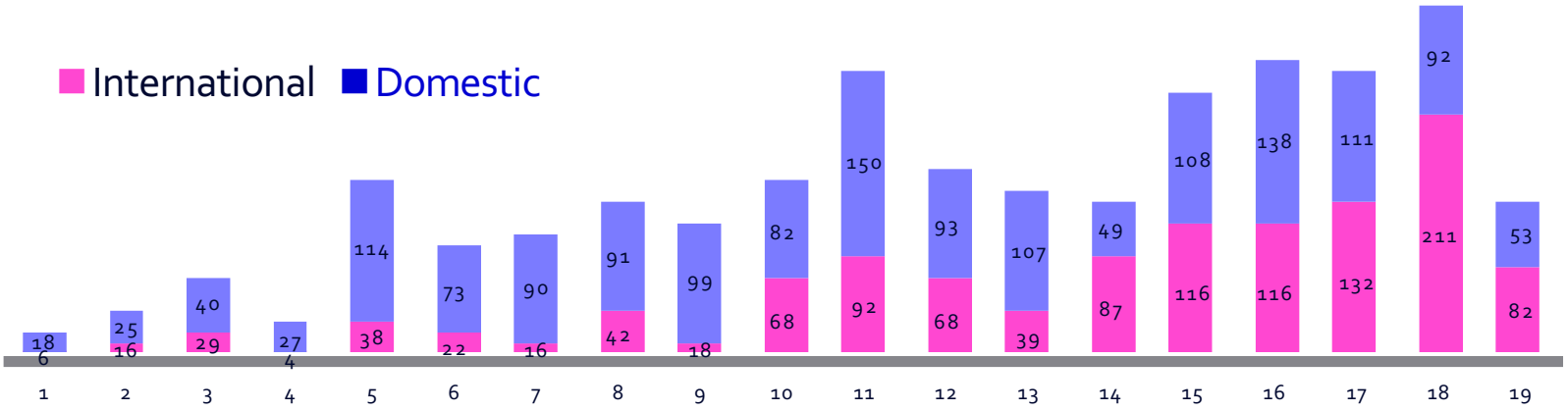
Lifelog-3	ライフログ検索
OpenLiveQ-2	ヤフー知恵袋での質問検索
QALab-PoliInfo	地方議会会議録を用いた質問応答タスク
STC-3	対話生成・対話の質評価
WWW-2	Web検索
CENTRE	CLEF/NTCIR/TRECの結果再現

NTCIRテストコレクション利用者数

4,164 ユーザ



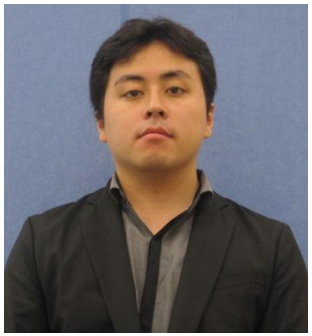
International Users vs Domestic Users



NTCIR-13 OpenLiveQの事例

タスク運営者: 加藤 誠，山本 岳洋 (京都大学),
藤田 澄男，西田 成臣，真鍋 知博 (ヤフー株式会社)




- **タスク： ヤフー知恵袋の検索システム構築**
  - 1,000キーワードに対して検索結果を出力
- **データ： 知恵袋のクエリ・クリックログ**
  - どの検索結果を誰がどのくらいクリックしたか
- **評価： ヤフー知恵袋におけるオンライン評価**

OpenLiveQデータ

	Training	Testing
Queries	1,000	1,000
Documents (or questions)	984,576	982,698
Clickthrough data (with user demographics*)	Data collected for 3 months	Data collected for 3 months
Relevance judges	N/A	For 100 queries

OpenLiveQデータ

answers & # views

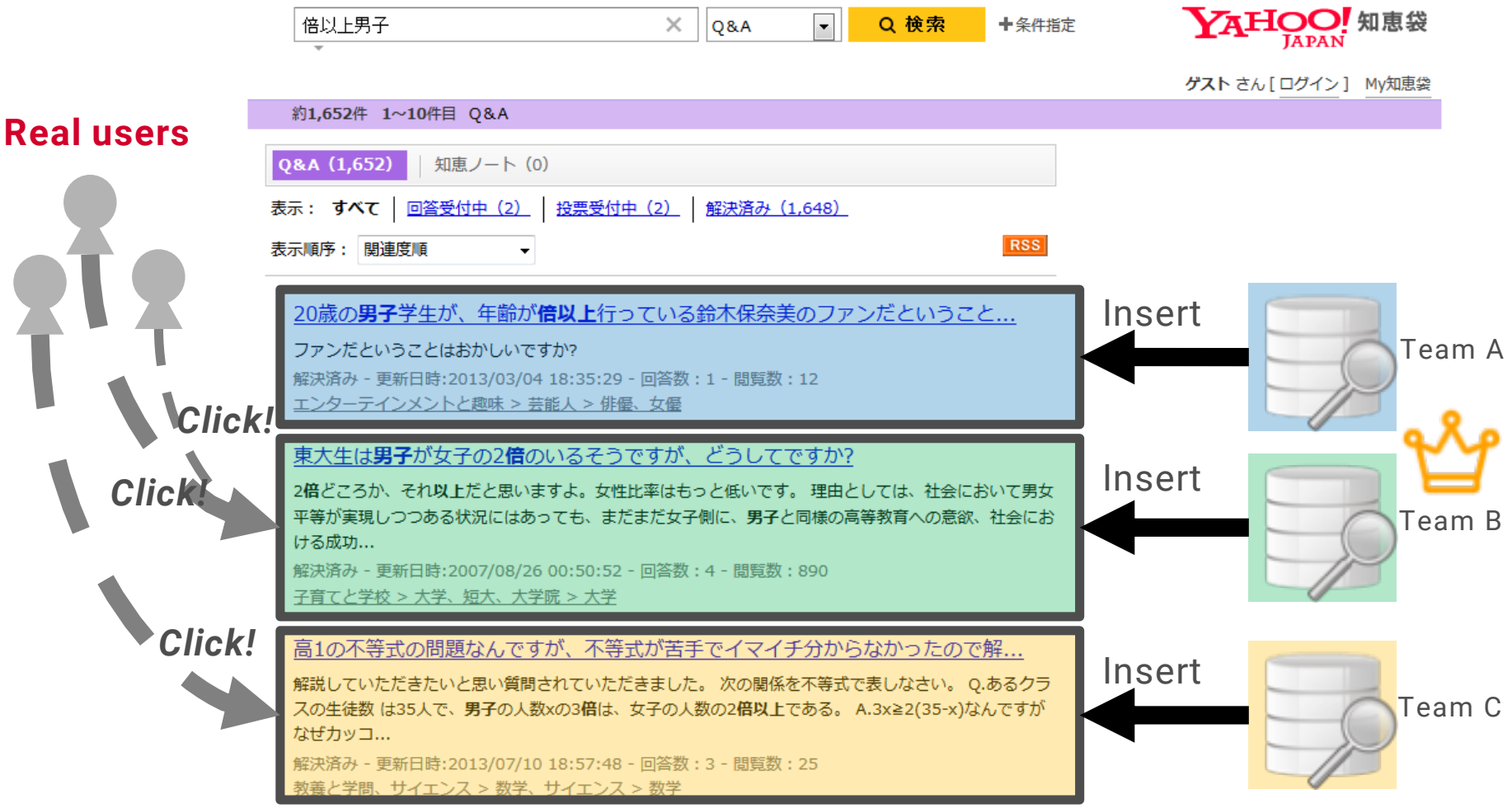
Query ID	Rank	Question ID	Title	Snippet	Status	Timestamp	# answers	# views	Category	Body	Best answer
OLQ-0001	1	q13166161098	バイオハザードって設定に無…	弾いたけどピアノ弾いたことない人は楽譜…	Solved	2016/11/13 3:35	1	42	エンターテインメントと趣味 > ゲーム	バイオハザードって…	レベッカもジルも弾けなかった場合
OLQ-0001	2	q14166076254	バイオハザードって設定にむ…	タックルしたり足で蹴りまくればこわせる…	Solved	2016/11/10 3:47	1	18	エンターテインメントと趣味 > ゲーム	バイオハザードって…	なので、バイオハザードアウトブレイクシリーズ…
OLQ-0001	3	q11166238681	バイオハザードのゲームの…	バイオハザード4が好きで30週くらいして…	Solved	2016/11/21 3:29	3	19	エンターテインメントと趣味 > ゲーム	バイオハザードのゲーム…	個人的には… BIOHAZARD REVELATIONS UNVEILED EDITION …
…	…	…	…	…	…	…	…	…	…	…	…
OLQ-2000	998	q11137434581	夫婦生活で一番疲れる…	自分が相手に隠していることがあった…	Solved	2014/10/28 15:14	6	0	生き方と恋愛、人間関係の悩み >…	夫婦生活で一番疲れる事は何ですか…	自分が相手に隠していることがあったら、どんなこと…
OLQ-2000	999	q1292632642	夫婦生活について教えて下さい…	主人とセックスをしておらず二年半に…	Solved	2012/9/3 9:51	5701	0	生き方と恋愛、人間関係の悩み > …	夫婦生活について教えて下さい。…	
OLQ-2000	1000	q1097950260	旦那との今後の夫婦生活、旦那…	結婚して来年の1月で2年になります。…	Solved	2012/12/5 10:01	4640	0	生き方と恋愛、人間関係の悩み > …	旦那との今後の夫婦生活、…	

OpenLiveQデータ

CTR　Gender　Age

Query ID	Question ID	Rank	CTR	Male	Female	0s	10s	20s	30s	40s	50s	60s
OLQ-0001	q10165187300	1	0.059	1	0	0	0	0	0	0	1	0
OLQ-0001	q11164148731	1	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q11166231691	1	0.023	1	0	0	0	0	0	1	0	0
OLQ-0001	q11166372256	1	0.036	1	0	0	0	1	0	0	0	0
OLQ-0001	q13161212253	1	0.051	0.909	0.091	0	0.091	0.364	0.182	0.182	0.091	0.091
OLQ-0001	q13166161098	1	0.021	0	0	0	0	0	0	0	0	0
OLQ-0001	q14164350104	1	0.1	0	0	0	0	0	0	0	0	0
OLQ-0001	q14164384744	1	0.188	1	0	0	0	0.5	0	0	0.5	0
OLQ-0001	q14165651359	1	0.048	1	0	0	0	0	0	1	0	0
OLQ-0001	q11166278091	2	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q11166476886	2	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q12164569302	2	0.037	0	0	0	0	0	0	0	0	0
OLQ-0001	q12165573687	2	0.083	0	0	0	0	0	0	0	0	0
OLQ-0001	q10162841855	3	0.036	1	0	0	0.5	0.5	0	0	0	0
OLQ-0001	q12164050757	3	0.06	1	0	0	0	0	0	1	0	0
OLQ-0001	q12164687517	3	0.049	0.833	0.167	0	0	0	0.167	0.667	0.167	0
OLQ-0001	q12165837862	3	0	0	0	0	0	0	0	0	0	0
OLQ-0001	q14158395769	3	0.027	0	0	0	0	0	0	0	0	0

OpenLiveQの評価

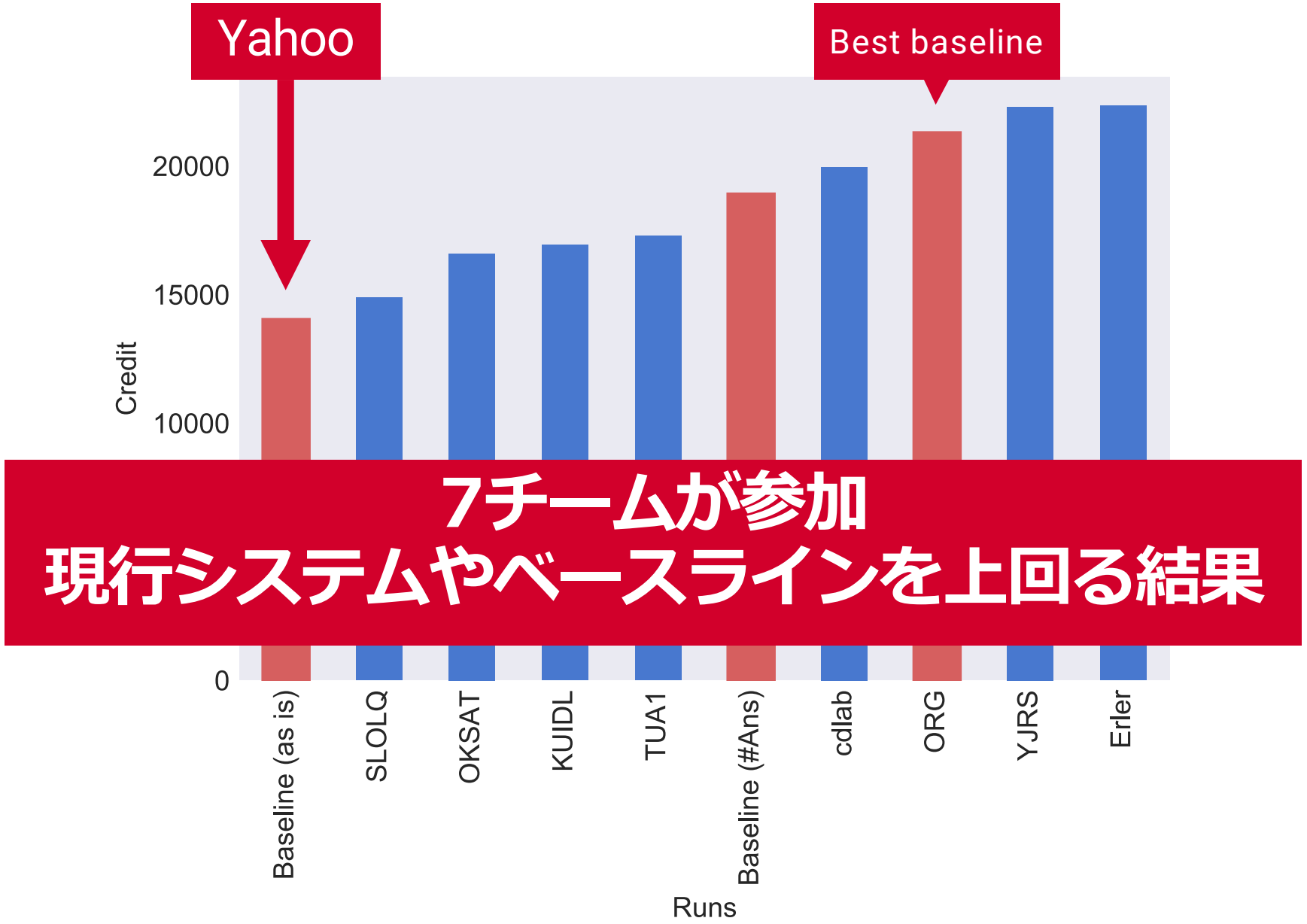


参加チームのシステム出力を

実際のヤフー知恵袋上で実際のユーザに提示し

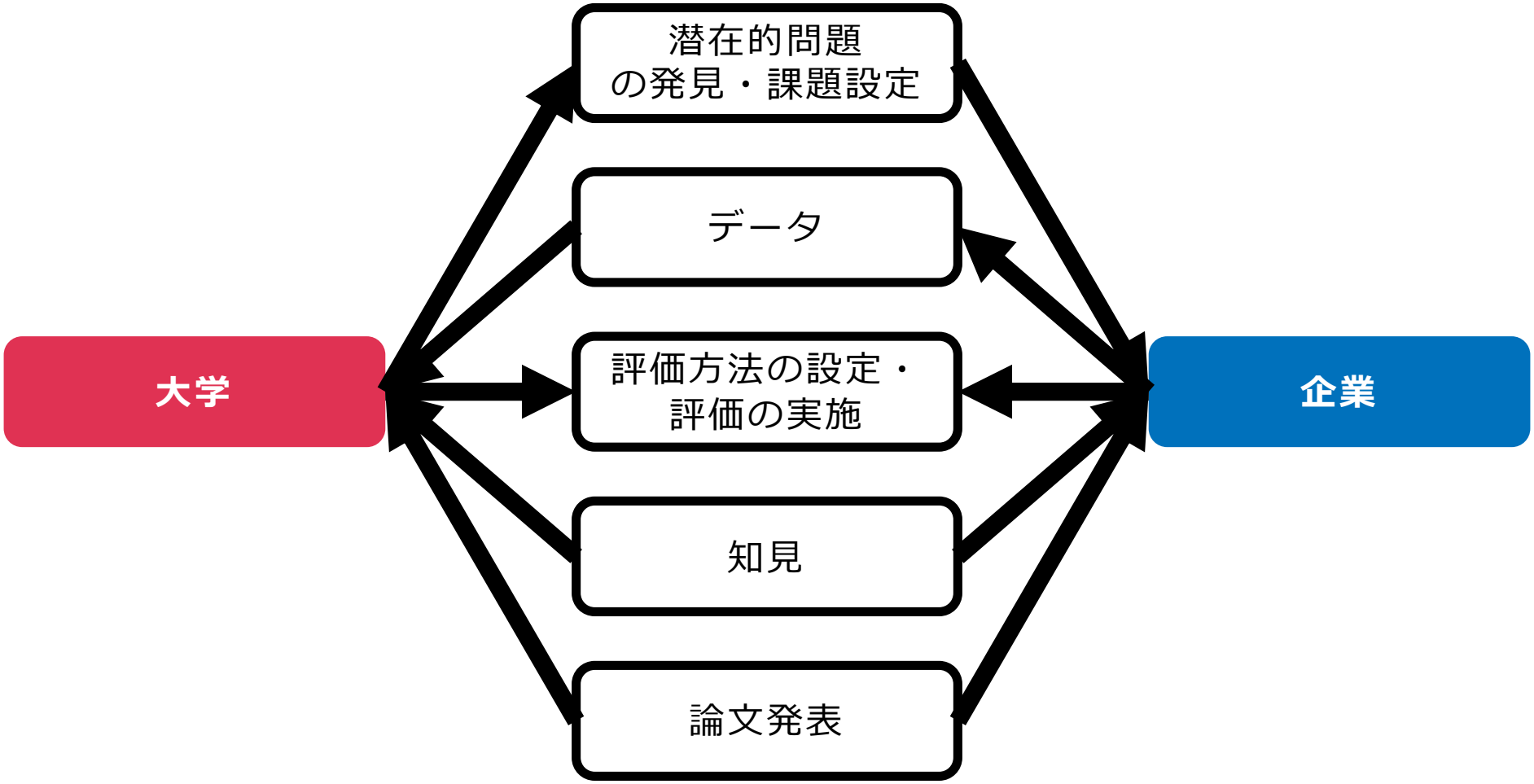
ユーザのクリック回数でシステムを評価

OpenLiveQの結果



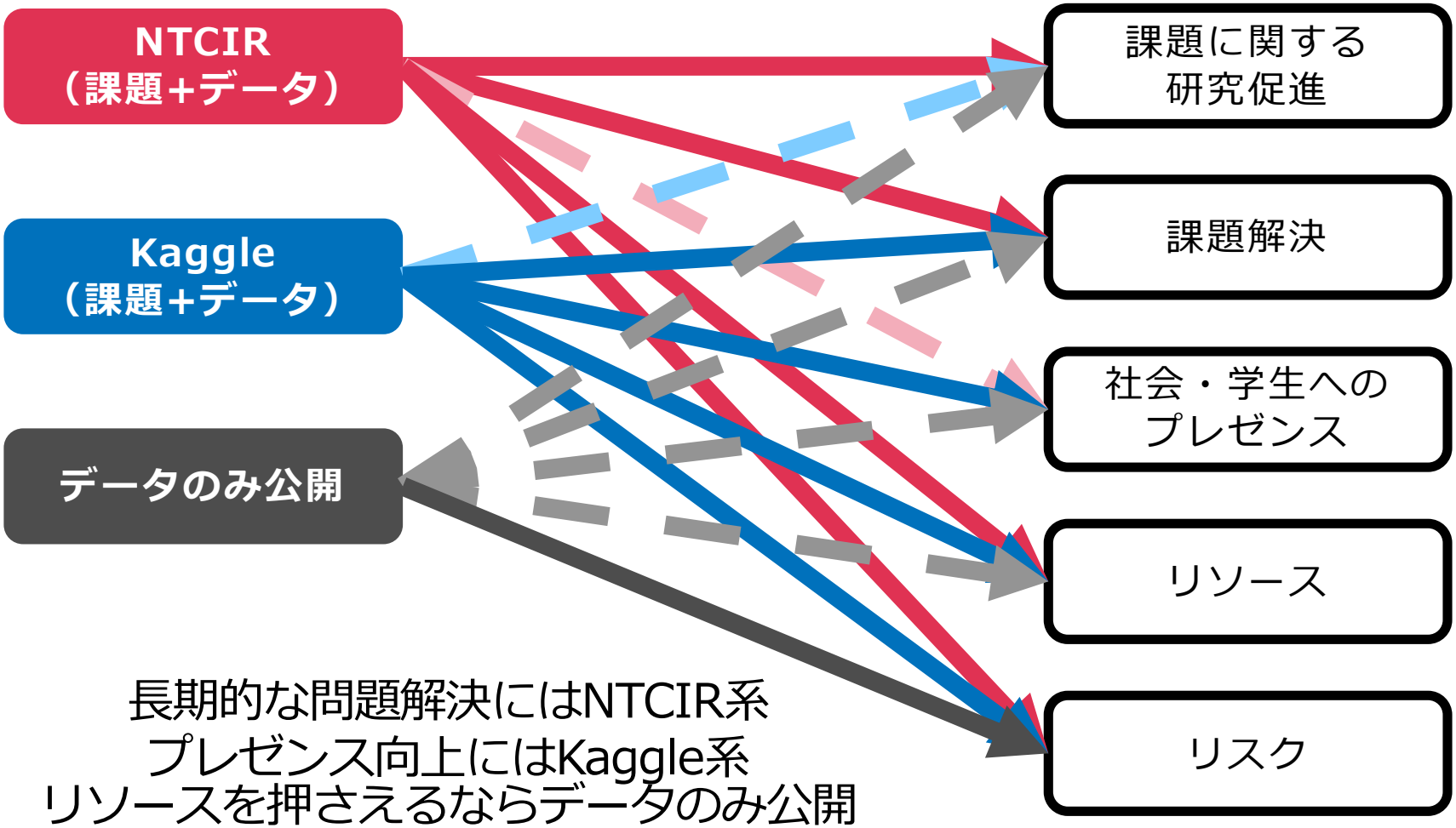
7チームが参加
現行システムやベースラインを上回る結果

OpenLiveQモデル



- **評価型ワークショップの意義や限界**
  - ○ 研究志向（がむしゃらに精度を求めない）
  - ○ 取り組んだ内容を発表議論する場がある
  - ○ データ・論文のアーカイブ・公開
  - × 縮小傾向（評価データ作成のコスト低下、他の取り組みとの競合）
- **企業との連携のメリット・難しさ**
  - ○ 現実的なタスク・貴重なデータ
  - ○ （たぶん）提供側のメリット
    - （たぶん）リクルーティング、共同研究開発
  - × （たぶん）実応用への課題は未解決・リソース
  - × オンライン評価の再現性

- （私が勝手に考えた）
  企業の方がデータを公開する方法とその効果



長期的な問題解決にはNTCIR系
プレゼンス向上にはKaggle系
リソースを押さえるならデータのみ公開