Tech企業における A/B test（公開用）

2020/5/16
@日本評価学会社会実験分科会
Shota Yasui

自己紹介

名前：安井翔太（32）
職業：Economic Research Scientist
web: https://yasui-salmon.github.io/

経歴：
2011年 立教大学 経済学部卒業
2013年 Norwegian School of Economics MSc in Economics
2013年 Cyberagent 入社（総合職, 微妙な分析の量産）
2015年 アドテク部門へ異動（専門職, MLの応用）
2017年 AILabへ異動（研究職, ML + CI回りの応用）

良く使う言語: R, SQL, Python

Tech企業とA/Bテスト

テック企業とABテスト

テック企業で行われるABの量

Google +1000 test/day

Microsoft +200 test/day

CyberAgent +??? test/day

学会もある

CODE @MIT

Conference on **Digital Experiment**(Link)

Computer Scienceでの研究も盛ん

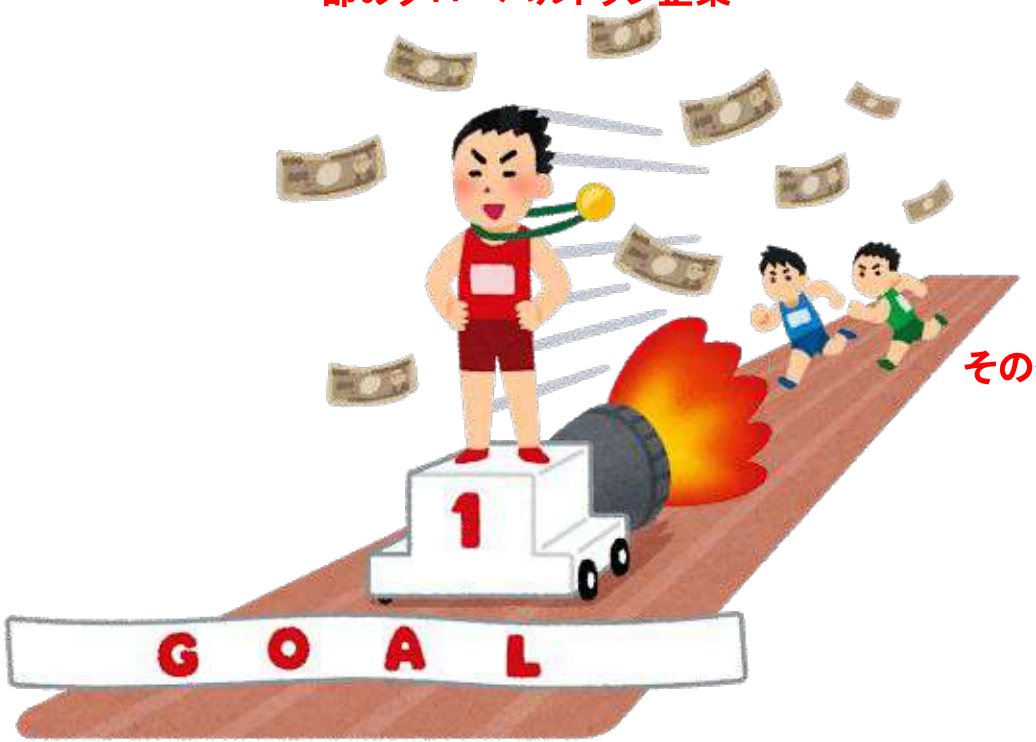
- sequential experiment
- efficient adaptive experiment
- best arm identification

なぜそんなにABテストしているか？

長期的な利益につながるから

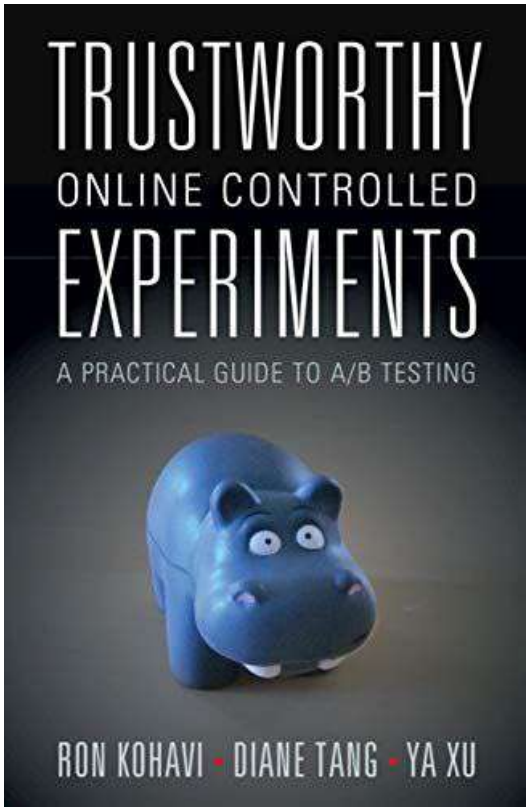
テック企業におけるRCT(A/Bテスト)の現実

一部のグローバルトップ企業



その他弱小ローカルテック企業

テック企業のABテストの考え方はこれ読むべき

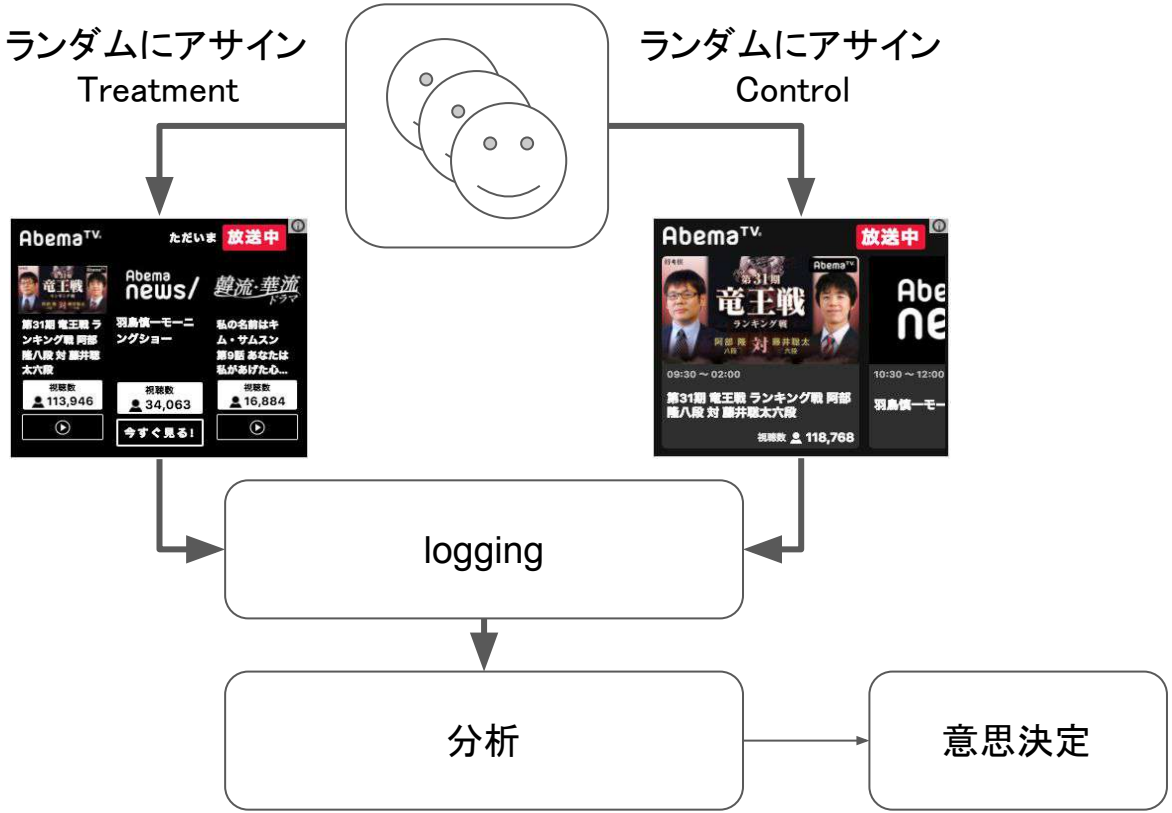


Ron kohavi

テック企業でABテスト文化を作り上げた人
Computer Scienceのトップ会議でABテストのチュートリアルや研究発表を続けて啓蒙活動を行ってきた人。

CyberAgentでのA/Bテスト

ABテストの基本的なプロセス

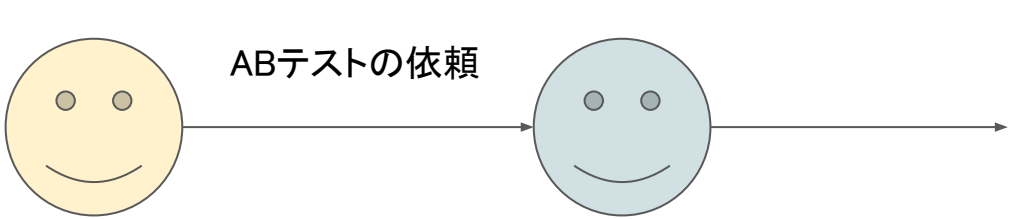


CAでどんな実験をしているか？

例：広告画像のABテスト

- どちらの画像の方がクリックされやすいだろうか？
- 数十～百程度のABテストが回っている
- 広告画像を作るクリエイターや営業の人が実施する

ABテストが始まるまでの流れ



バナーを作成

結果の分析

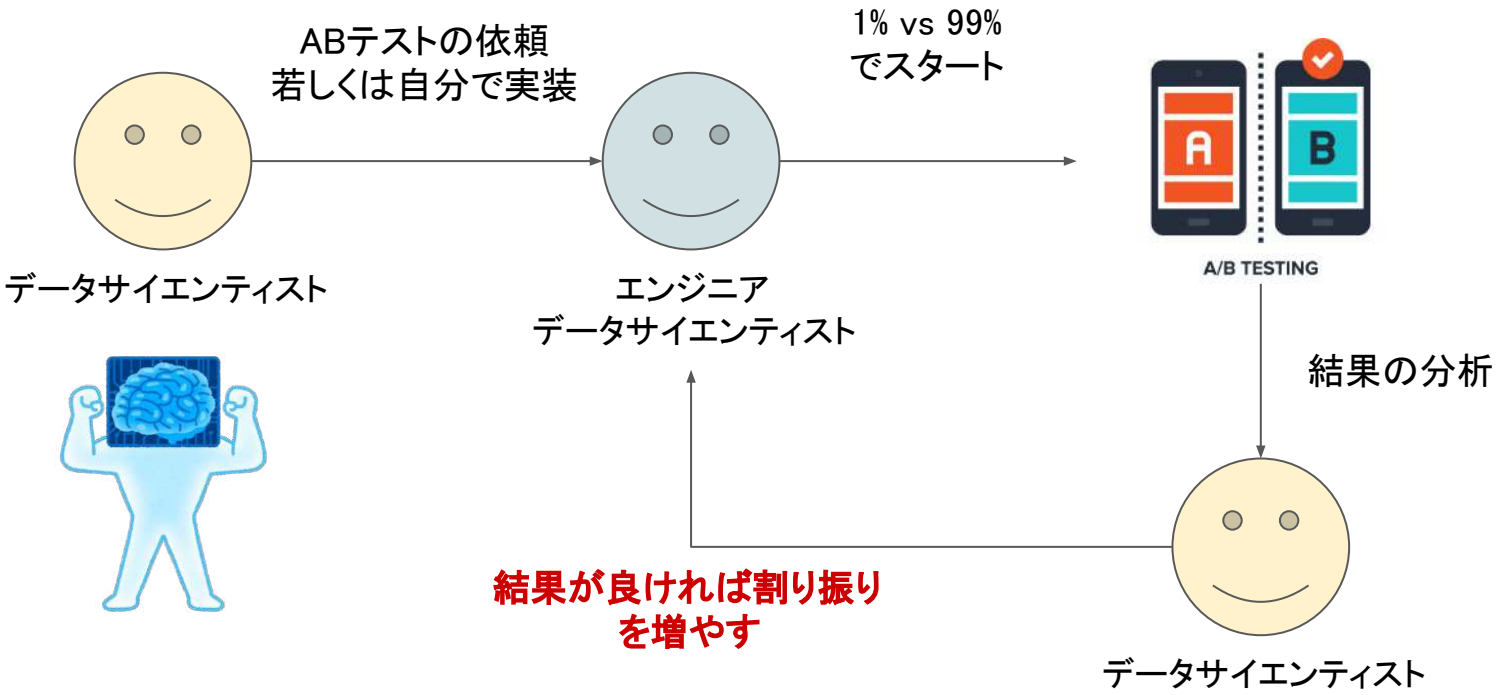
クリエイター

CAでどんな実験をしているか？

例：広告の配信アルゴリズムのABテスト

- どちらのアルゴリズムの方が売上に貢献するだろうか？
- 月10回程度

ABテストが始まるまでの流れ



社会科学的な設定と異なる点

- **意思決定者=分析者**
  - クリエイターがABを依頼して、自分で結果を見る。
  - データサイエンティストがABを始め、自分で結果をみる。
- **分析と意思決定のサイクルが短い**
  - 通常業務に加え、大量の分析と意思決定の日々。
- **検証したい介入の種類が多い**
  - 改善しうるものはなんでも試したい

ここから生まれる需要

- **分析として正しい ＜ 意思決定として正しい**
  - 最後の意思決定の質が重要
- **自動化の重要性**
  - 実験が多くなるとより面倒に・・・
- **複雑な設定の実験を簡素にする**
  - より多くの設定で実験を使える状態にしたい

→これにこたえたい

より効率的な意思決定の導入

意思決定する立場になると・・・

分析における
検出力の最大化？

意思決定における
Regretの最小化？

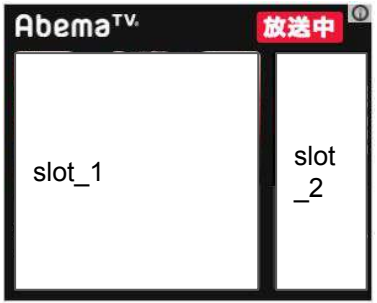
最適な選択との乖離＝Regret

一体どちらがゴールなのか？

考えるお題

あるユーザーに対してどちらの広告テンプレートを見せるべきか？

template_id: 26



template_id: 75

Regretという観点で実験を考え直す

1. **トータルで1000万回広告を表示する**
2. **より多くのクリックを集めたい**
3. **最初の200万回でABテストを行う**
4. **良かった選択肢を選び続ける**

実験では最適な選択との乖離が出来る＝Regret

→ビジネスにおける損失

Regretという観点で実験を考え直す

1. トータルで1000万回広告を表示する
2. より多くのクリックを集めたい
3. 最初の200万回でABテストを行う←ここに無駄がある
4. 良かった選択肢を選び続ける

実験では最適な選択との乖離が出来る＝Regret

→ビジネスにおける損失

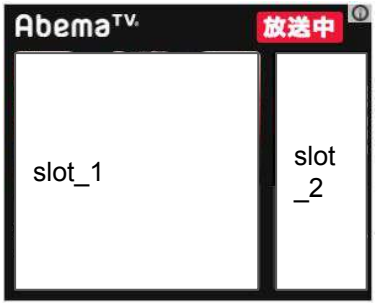
Adaptiveな実験としてのバンディット

あるユーザーに対してどちらの広告テンプレートを見せるべきか？

→クリックがより起きそうな方を**都度**選ぶべき

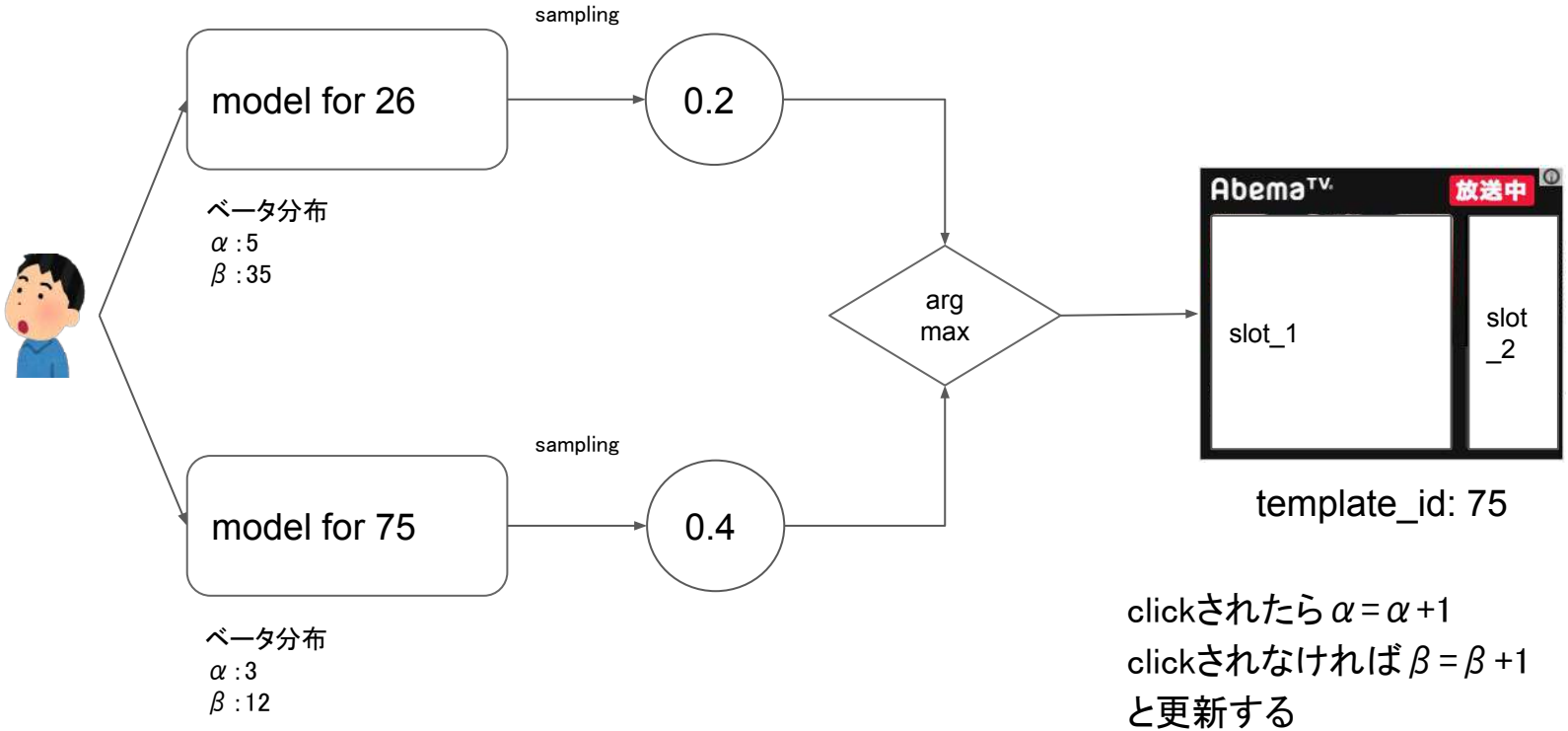
（クリックを増やしたいなら

template_id: 26



template_id: 75

Thompson Sampling MAB



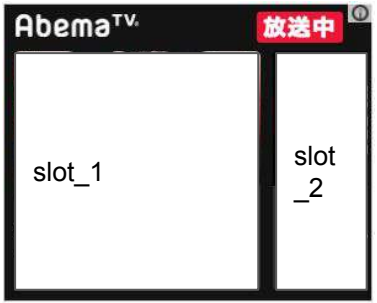
Adaptiveな実験としてのバンディット

あるユーザーに対してどちらの広告テンプレートを見せるべきか？

→クリックがより起きそうな方を選ぶべき（クリックを増やしたいなら

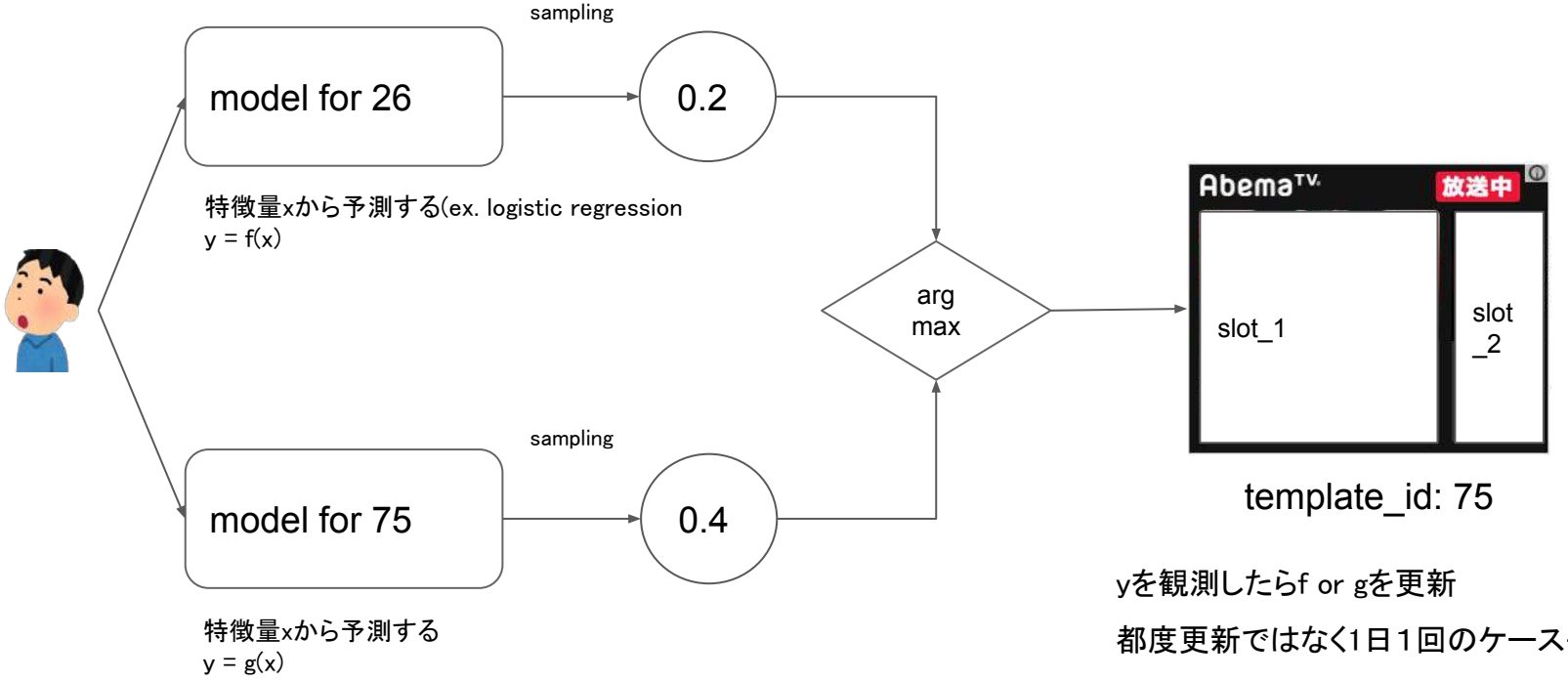
→機械学習で予測して、予測値が大き方を選べば良いのでは？

template_id: 26



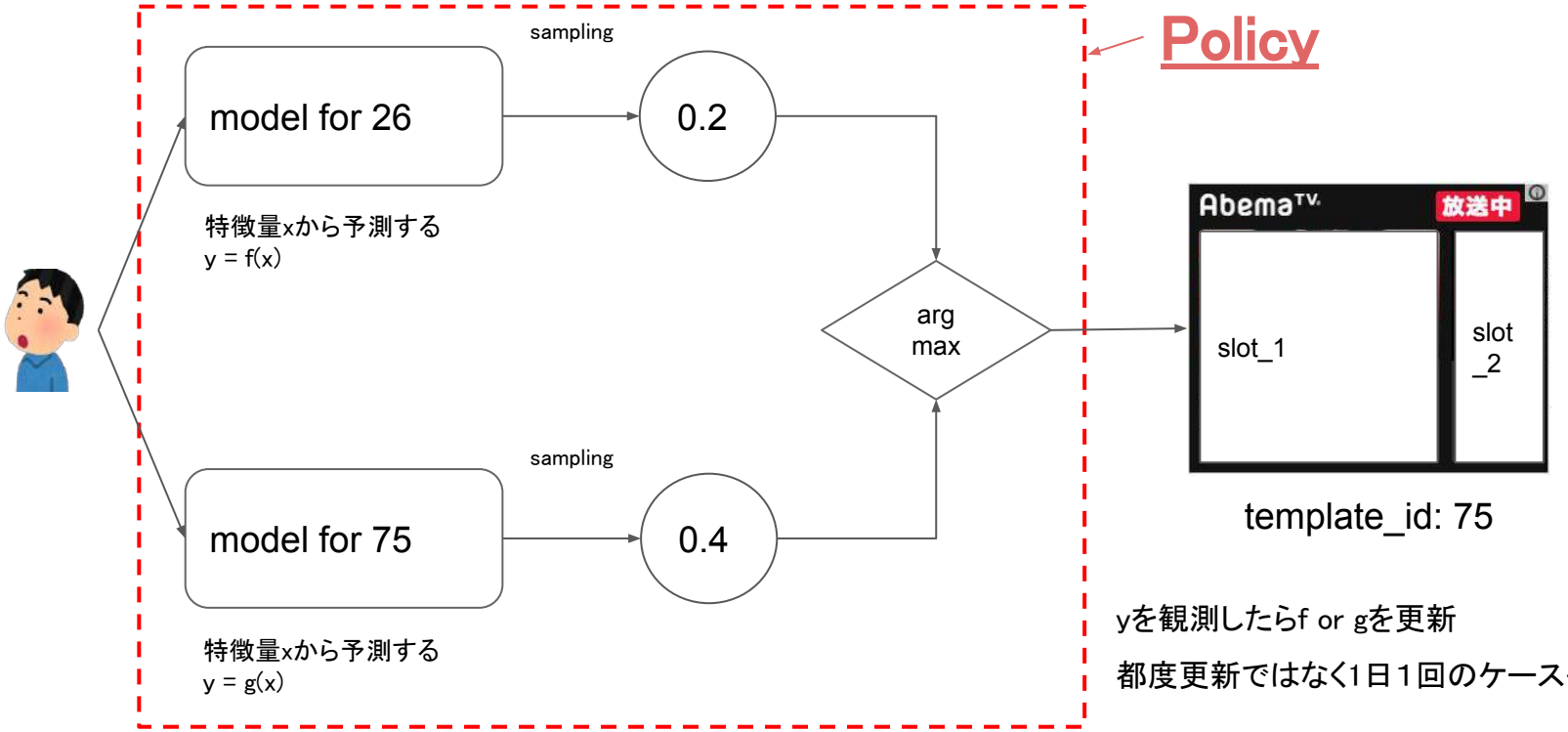
template_id: 75

Thompson Sampling Contextual Bandit



男性には26だけど女性には75が良いといった傾向が汲み取れる

Thompson Sampling Contextual Bandit



男性には26だけど女性には75が良いといった傾向が汲み取れる

バンディットの利点

- **Regretが一定のバウンドに収まる**
  - 意思決定の質がある程度担保される

- **分析→意思決定のフローが自動**
  - 意思決定者としては楽
  - ちゃんと動作しているか？という運用コストが発生
  - この辺りはデータサイエンティストが頑張る？

バンディットのログから評価を行う

分析・評価の必要性

- **広告主に対するレポート**
  - 次にどんな広告画像を作るべきなのか？
  - 配信した広告画像は何が良かったのか？

- **バンディットの操作には分析が必要**
  - バンディットの選択肢を増減させたい
  - 今ある選択肢の中でいらないものはどれか？

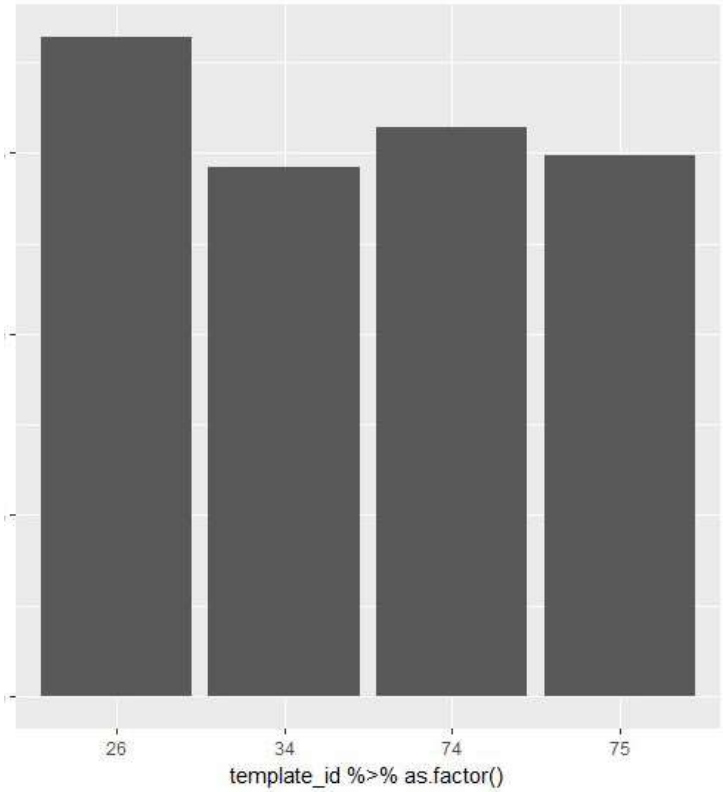
→効果の推定が必要になる

バンディットフィードバック

- **Thompson Samplingの特徴**
  - ある確率にしたがって腕を選択する＝ABテスト
  - X毎に選択確率が異なる＝X毎に異なる確率でAB
- **因果推論からの観点**
  - 腕の選択確率＝真の傾向スコア
  - Xは全て既知

→IPWで因果効果が推定できる状況

Biased Result

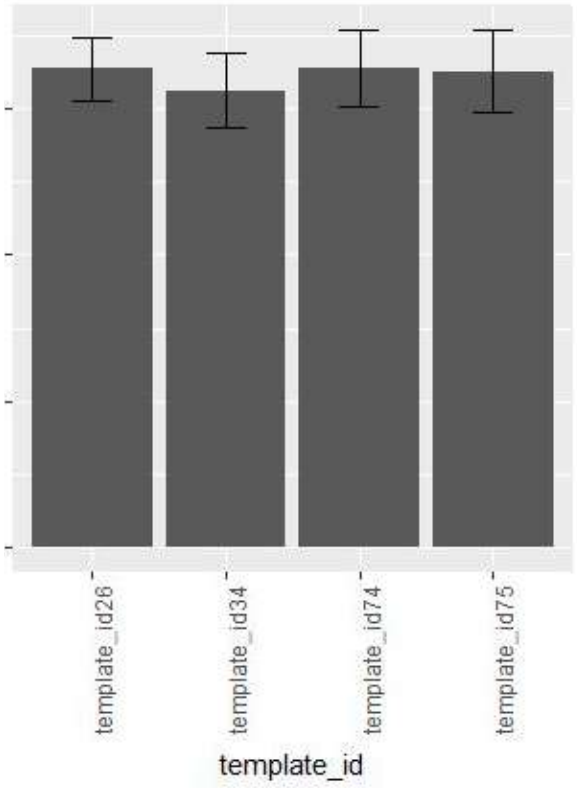


- template_id毎にCTRを計算する
- template_id:26のCTRが高そう

→Biasを含んだ結果

営業や事業責任者の方が見るデータ

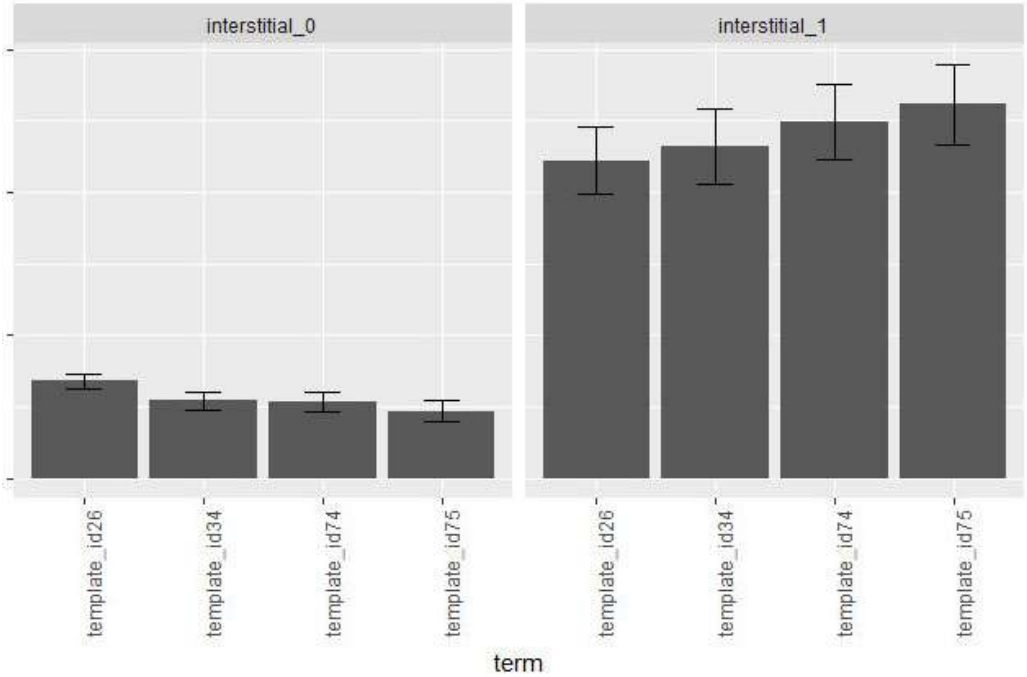
IPW result



- バイアスがある程度減っているはず。
- 26がよかったというのは幻想だった。
- CTRはどれも大差ないという結果。

ATEベースで意思決定して良いか？

Heterogeneity



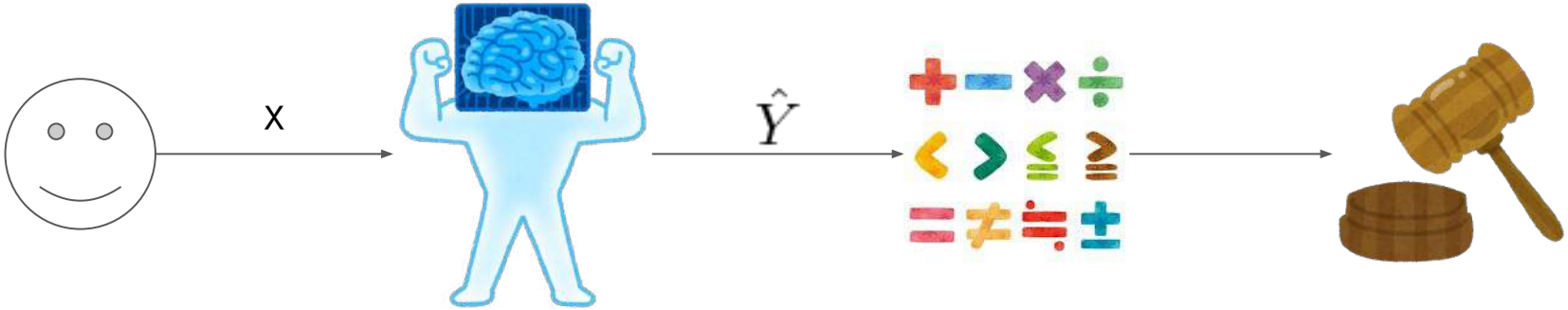
- interstitial
  - 1: 全画面で見せる
  - 0: 記事の中で見せる
- 非インステ枠では26が良い
- インステ枠では75が良い

- ATEで悪くとも、あるXでは良い腕も存在しているかもしれない。
- ATEにしたがった意思決定をしても、改善出来ない可能性がある。

複雑な状況での効果検証

CyberAgent

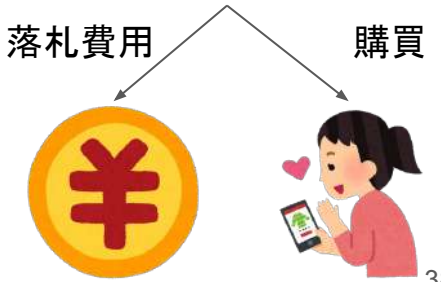
広告オークションにおける機械学習の役割



何かしらの
機械学習モデル

予測値を元に
入札額を計算

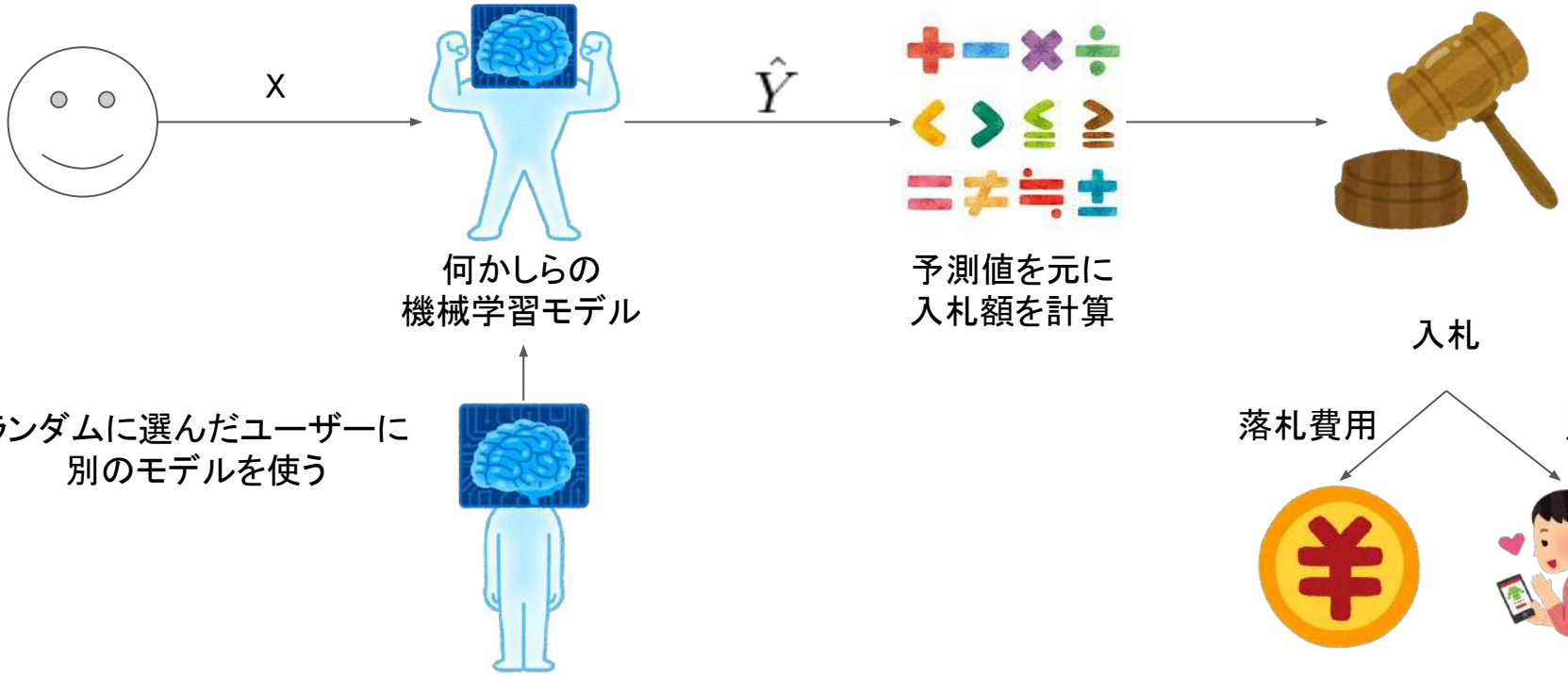
入札



- オークションで落札すると広告が表示出来る
- 落札するとコストが発生する
- 広告をクリックして物が買われると嬉しい

→機械学習は購買確率やクリック確率を予測している

機械学習のABテスト


X
$\hat{Y}$
何かしらの
機械学習モデル
予測値を元に
入札額を計算
入札
落札費用
購買
ランダムに選んだユーザーに
別のモデルを使う

何が難しいか？

- **費用の側面と収益の側面がある**
  - 薬であれば効果と副作用？
- **どの様に意思決定すればよいか？**
  - 収益と費用が同じ単位ではない場合が多い
    - 収益 vs ユーザー体験など
  - 収益も増えるけど、費用も増えるモデルは良いモデルなのか？

→実験が出来ても意思決定が難しい

取り敢えずの対処

- **以下の結果以外では結論が出る**
  - 費用が変わらないけど、収益が増える。
  - 収益が変わらないけど、費用が減る。

- **何かしらの重みを決めてスコアを出す**
  - Overall Evaluation Criteria（OEC）と呼ばれるもの
    - Google, Amazon, Microsoft, etcはOECを決めるチームが存在する
  - スコアに差があれば結論を出す

ある実験結果

CV	Cost	CPA
+31%*	+28%*	-2%

Table 3: Online relative comparison of FFM and FFMIW. The values shown are the relative change in the FFMIW against the FFM. * denotes statistical significance.

- 機械学習におけるデータのバイアスを除去してABテスト
- CPA: 購買を1件獲得するあたりの平均コスト
- Cost: 広告費用

→平均コストは有意な差が無いが、広告費用は有意に増加。
広告配信ビジネス的には嬉しい結果。

Yasui, Shota, Gota Morishita, Komei Fujita, Masashi Shibata. “A Feedback Shift Correction in Predicting Conversion Rates under Delayed Feedback.” The Web Conference (WWW), Taipei, Apr. 2020

おしまい