# コンテンツの視聴傾向から属性データを推定してみた

瀧本 恭佑
日本テレビ放送網株式会社
ICT戦略本部 データアナリスト

春日井 健太
株式会社グラフ
開発部 データサイエンティスト

Google Cloud

# 自己紹介

# 自己紹介

瀧本 恭佑　日本テレビ放送網株式会社

略歴：
日本テレビ入社後、報道局で番組制作/記者業務。
2018 年より ICT 戦略本部、動画配信事業に携わる。
ビジネス課題/ニーズの発見・解決を通して、
データ側面からの事業支援を行う。

# 自己紹介

春日井 健太　株式会社グラフ

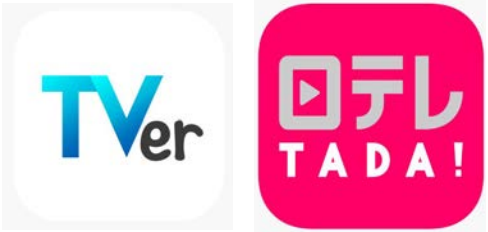
略歴:
2019 年 1 月、受託データ分析を主要事業とする株式会社グラフに入社。複数業界のクライアントに対して、データの利用価値を最大化させるための戦略立案や各種機械学習モデルの構築を担当。現在、日本テレビ様にて機械学習モデルの構築を行う。

# 背景

# 日本テレビの動画配信事業

- SVOD（Subscription VOD）定額制動画配信
  - 番組アーカイブの他、映画や海外ドラマ等多数配信

- AVOD（Advertising VOD）広告付き無料配信
  - ドラマ・バラエティ等テレビコンテンツを見逃し配信
  - TVer（民放公式テレビポータル）
    - MAU 1,000 万超　累計 DL 2,500 万超（2020 年 3 月）



# 課題

- 属性を推定してターゲティング可能な広告在庫を増やしたい
  - 生年月日/性別/郵便番号を入力してもらう仕様
  - 全体の約 10% のユーザーが未回答


- 機械学習やデータベースエンジニアのリソースも潤沢ではない
  - AutoML・AI Platform 等の各種 GCP サービスの導入により開発リソース・メンテナンスコストの最小化

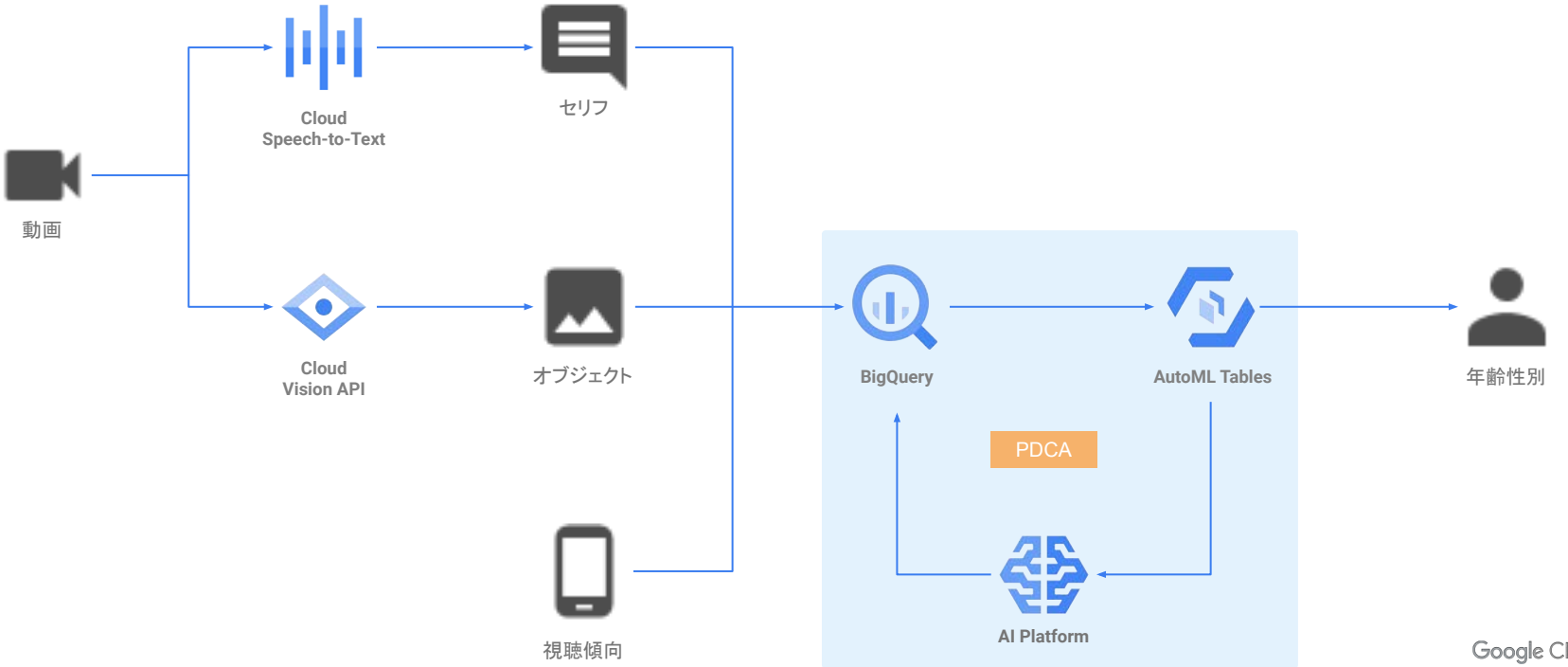
# 取り組み内容

# お話すること

- 年齢性別推定のために行ったこと
- 開発リソース・メンテナンスコストの最小化における、各種 GCP サービスの利点
- 基本的には Google さんにお任せの姿勢でいるものの、現場の人間（特にデータ分析者）がしなければいけないこと

# 全体フロー

各種 GCP サービスを利用し、視聴傾向等から年齢性別を推定



# Speech-to-Text による特徴量作成

- 実施内容
  - 動画中の音声データからセリフ情報を抽出し、キーワード等の動画を特徴付けるメタデータに変換
- 利点
  - BigQuery や GCS との連携が容易
  - 単語に分割されたデータを取得出来るため、そのまま特徴量として使用可能
  - 新たなアルゴリズムが世に出た際は、Google さんが実装してくれる

# Speech-to-Text による特徴量作成

| # | MOCO'S キッチン<br>2019 年 1 月 8 日回 | the-SOCIAL<br>2019 年 1 月 24 日回 | 3 年 A 組<br>―今から皆さんは、人質です―<br>第 1 話 |
|---|---|---|---|
| 1 | ソース | 富士 | 自殺 |
| 2 | トマト | 富士山 | 先生 |
| 3 | オリーブ | 山 | 生徒 |
| 4 | もてなし | 写真 | 自分 |
| 5 | ホワイト | 山中湖 | 友達 |
| 6 | 玉ねぎ | 撮影 | 本当 |
| 7 | オイル | シャッター | 教師 |
| 8 | 料理 | ロシア | 答え |
| 9 | 小麦粉 | スウェーデン | 人質 |
| 10 | 肘 | 東京 | ドーピング |

# Vision API による特徴量作成

- 実施内容
  - 動画中の画像データからオブジェクト情報を抽出し、キーワード等の動画を特徴付けるメタデータに変換
- 利点
  - BigQuery や GCS との連携が容易
  - 新たなアルゴリズムが世に出た際は、Google さんが実装してくれる

# Vision API による特徴量作成

| # | MOCO'S キッチン<br>2019 年 1 月 8 日回 | the-SOCIAL<br>2019 年 1 月 24 日回 | 3 年 A 組<br>—今から皆さんは、人質です—<br>第 1 話 |
|---|---|---|---|
| 1 | 料理 | 空 | 部屋 |
| 2 | 食品 | 写真撮影 | イベント |
| 3 | 材料 | 山 | 仕事 |
| 4 | レシピ | アイス | サラリーマン |
| 5 | 子供 | 北極圏 | 黒髪 |
| 6 | ベジタリアンフード | 成層火山 | スナップショット |
| 7 | 幼児 | 山岳地形 | 教室 |
| 8 | 娘 | カメラオペレーター | 写真撮影 |
| 9 | 部屋 | 海洋 | 生徒 |
| 10 | スープ | 観光 | 建物 |

# BigQuery によるデータ蓄積

- 実施内容
  - 動画のメタデータを始め、全体で 20 以上のデータソースからなる 60 TiB 以上のデータを保有
  - 各種ダッシュボード用 DM や機械学習用データセットを作成
- 利点
  - 管理に手間がかからない
  - 高速に大容量のデータ処理が可能

# AutoML Tables によるモデル構築

- 実施内容
  - ビジネス理解と機械学習知識を基に評価指標を設計
  - 当該ドメインの機械学習知識を基に精度向上のための仮説立案をし、データセットを作成
  - 特徴量エンジニアリングやハイパーパラメータチューニング等は AutoML に任せる

# AutoML Tables によるモデル構築

- 実施内容
  - AutoML の GUI 上で「欠損 %（カウント）」「固有の値」「詳細」を確認
    - 過学習を引き起こしそうな変数は除去
  - 公式ドキュメントを読み込み、AutoMLの特徴を理解
    - 日本語のテキストデータは過学習を避けるため除去する（トークン化は Unicode スクリプトの境界に基づく）等実施

# AutoML Tables によるモデル構築

- 実施内容
  - 適切に配列型を使用（列数は 1,000 列まで。配列の後方で現れる項目は、先頭より大きく重み付けされる）
    - 番組キーワード TOP10 や視聴回数 TOP10 番組等の変数から上手く学習してくれる
  - 要素数が多くなる配列は複数カラムに分割（配列内の最後の N 個（N = {1, 2, 4, 8, all}）の要素について、辞書ルックアップインデックスに変換し平均される）
    - 全期間の視聴番組 ID を 1 ヶ月毎でカラム分割

# AutoML Tables によるモデル構築

- 実施内容
  - ユーザー単位の予測に加えて、動画視聴ログ単位での予測も実施
    - 結果をユーザー単位にまとめる際、可視化して確認しながら実施するために AI Platform を使用

# AutoML Tables によるモデル構築

- 利点
  - 特徴量エンジニアリングやハイパーパラメータ調整不要でも高精度
  - 短期間でモデル構築可能
  - 新たなアルゴリズムが世に出た際は、Google さんが実装してくれる
  - 機械学習の知見が深くない人でも操作・理解しやすい

# AI Platform によるモデル検証

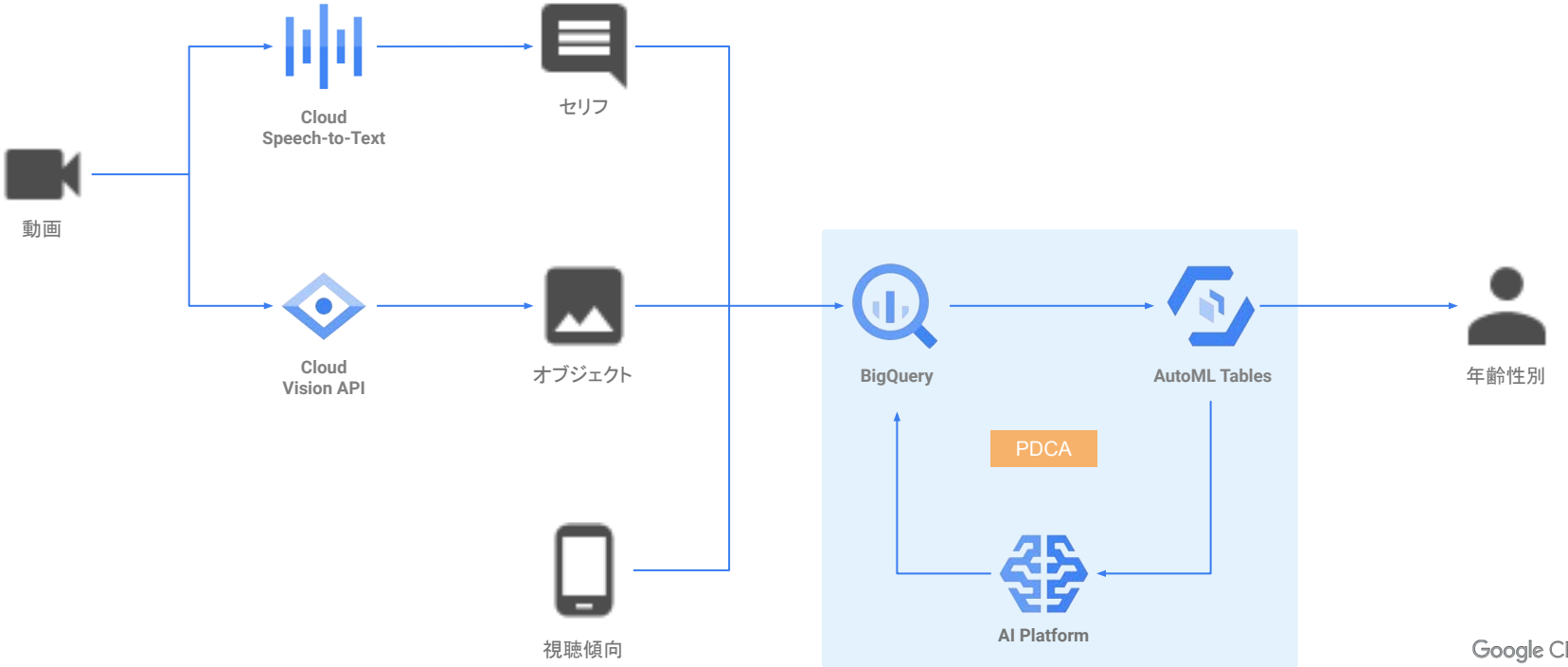
- 実施内容
  - 過学習やリークが発生していないかデータ確認
    - AutoML は手軽であるがゆえに気付きにくい
  - 予測が上手くいった・上手くいってない対象の特徴を分析
    - 分析結果を基に、拡充すべきデータや除去すべきデータを考察し、データセット作成に反映
  - 動画視聴ログ単位の予測結果をユーザー単位にまとめる
    - 複数通りのまとめ方の中でどの方法が適切か、可視化して確認しながら検討

# AI Platform によるモデル検証

- 利点
  - 実施内容に応じた開発環境の構築が容易
  - ノートブック形式で情報共有が容易
  - BigQuery や Data Portal では難しい柔軟な分析が可能
  - ブラックボックスになりがちな AutoML に対して、データセットの不具合確認や PDCA による精度改善を行いやすい

# 取り組みまとめ

各種 GCP サービスを利用することでメンテナンスコストを抑えた



# 今後の展開

# 目指すところ

- コンテンツ制作からユーザーサービスまで
  データインテグレーション
  - 視聴データの分析深化
  - データに基づくコンテンツ制作

- クラウドを適切に利用してコスト最適化

# Thank you