データサイエンスすいすい会

第11回「大規模データ分析基盤と自動機械学習（BigQuery + Matillion + ForecastFlow）」

すいすい会の紹介

- データサイエンスの領域に関して知見を共有しあい、みんなで実践的な理解度を高めたい

- 自分の知見などを自由に発言してください
  （呼び水のテーマ紹介をします）

- しばらくの期間は、自動機械学習を中心にお話をさせてください

- 資料：　GRIホームページ　https://gri.jp/news/12924
- Slack　ForecastFlowチャンネル
  https://join.slack.com/t/forecastflowusers/shared_invite/enQtNTgyMjcxOTg0NzcxLTBkOWEzYWMwNDJmNTkyMDQzYmIxYWU0YWI4ZmU3ZDU0ZTMxNDUwODAxMWFmYmU1YjJiZGI0MjRhYWYyYTNlZTQ
- YouTube: GRIチャネル　https://www.youtube.com/channel/UCDVGqf-dgczYHMfPzpQ0jNQ

©GRI Inc.

過去のすいすい会の動画（YouTube）

- 第１回「ビジネスでAIを活用するための問題設定」
  - https://www.youtube.com/watch?v=ppkgYD69NDU

- 第２回「アンケートと機械学習」
  - https://www.youtube.com/watch?v=-s1PcLQUBNI
  - https://www.youtube.com/watch?v=ylVT3OyayEk
  - https://www.youtube.com/watch?v=sFVabtfNvIk

- 第３回「分類問題の精度評価手法と不均衡データ」
  - https://www.youtube.com/watch?v=q6WJDTOgotA

- 第４回「自動機械学習での特徴量の作り方」
  - https://www.youtube.com/watch?v=Ms52EnCRk8g

- 第６回「予測スコアを用いた効果的な施策実施」
  - https://youtu.be/8Sm8ex6taE8

©GRI Inc.

本日のアジェンダ

- 列指向DB, ELTツール, 自動機械学習ツール, BIツールの役割
- 大規模データの扱いの悩みのポイント
- BigQuery-Matillion-ForecastFlow-Tableauを使った大規模データを扱う方法

©GRI Inc.

大規模データの例

- ウェブのアクセスログ、動画やライブ配信の視聴履歴、SNSデータ
- アプリログ
- 位置情報データ
- 大規模コールセンターの通話履歴
- IoTのセンサーデータ

©GRI Inc.

列指向DB（大規模データの分析に適している）

- 列指向DB：　列方向にデータを圧縮して格納
  - BigQuery, Redshift, Snowflake, TreasureData, Vertica, (Tableau Hyper)
  - 高速に集計できる
  - 圧縮時に時間が必要。１件のデータ修正は遅い
- 行指向DB
  - Oracle, MySQL, (CSV, Excel)

No	氏名	フリガナ	性別	生年月日	点数
1	板橋範久	イタバシノリヒサ	男	2001/08/19	94
2	猪俣清治	イノマタセイジ	男	2001/09/13	41
3	福島莉乃	フクシマリノ	女	1998/07/05	46
4	碓井桃歌	ウスイモモカ	女	2004/10/29	63
5	木田和島	キダワジマ	男	2006/11/01	60

https://www.slideshare.net/optsendai/ss-81845870

©GRI Inc.

列指向DB（大規模データの分析に適している）

	DB名	SELECT		
		1万	100万	1億
行指向	MySQL	0.06s	1.5s	55.5m
行指向	PostgreSQL	0.06s	0.7s	1.2h
列指向	TREASURE DATA	2s	2s	8s
列指向	Redshift	0.2s	0.3s	0.4s

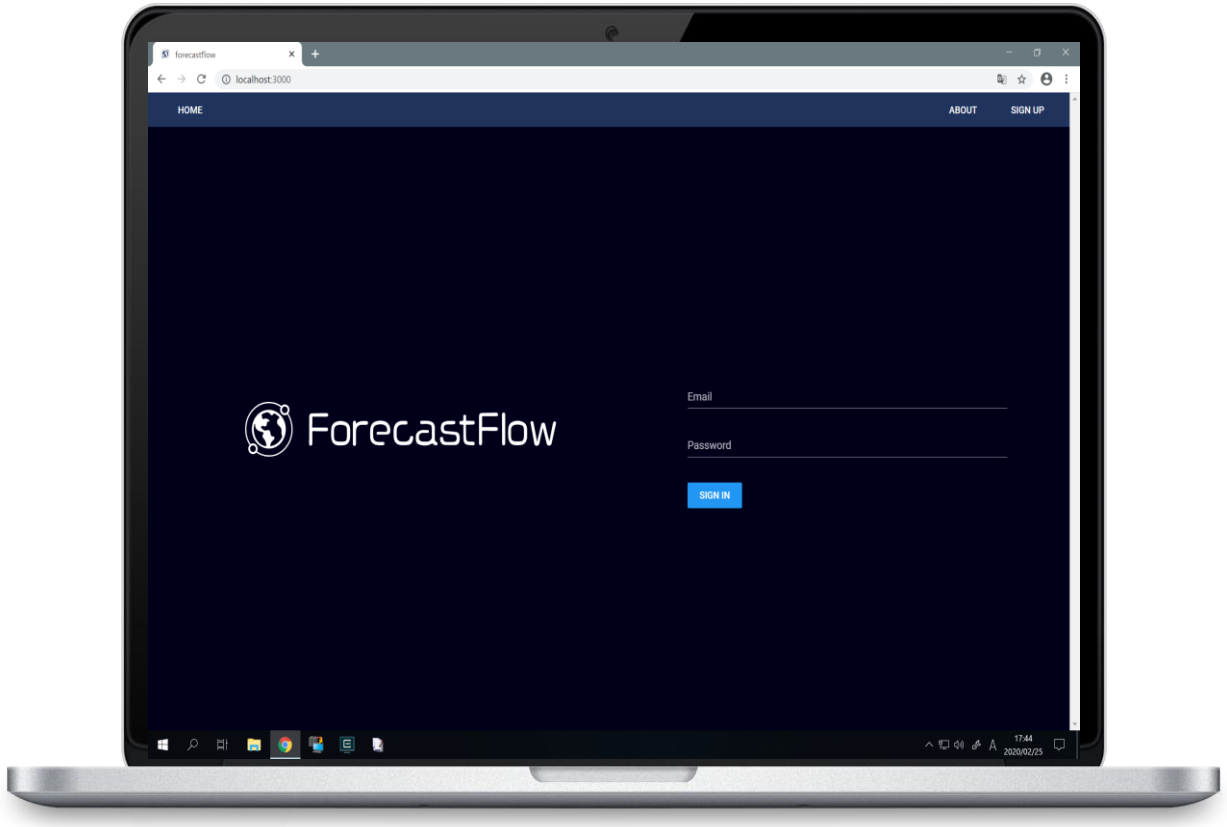
https://www.slideshare.net/optsendai/ss-81845870

©GRI Inc.

ForecastFlowとは

入力データに応じて最適な機械学習モデルを自動的に構築する
AutoML（自動機械学習） サービス

- 予測することが得意
- 誰でも予測が「かんたん」にできる



©GRI Inc.

ForecastFlowの狙い

予測ができれば世界は変えられる

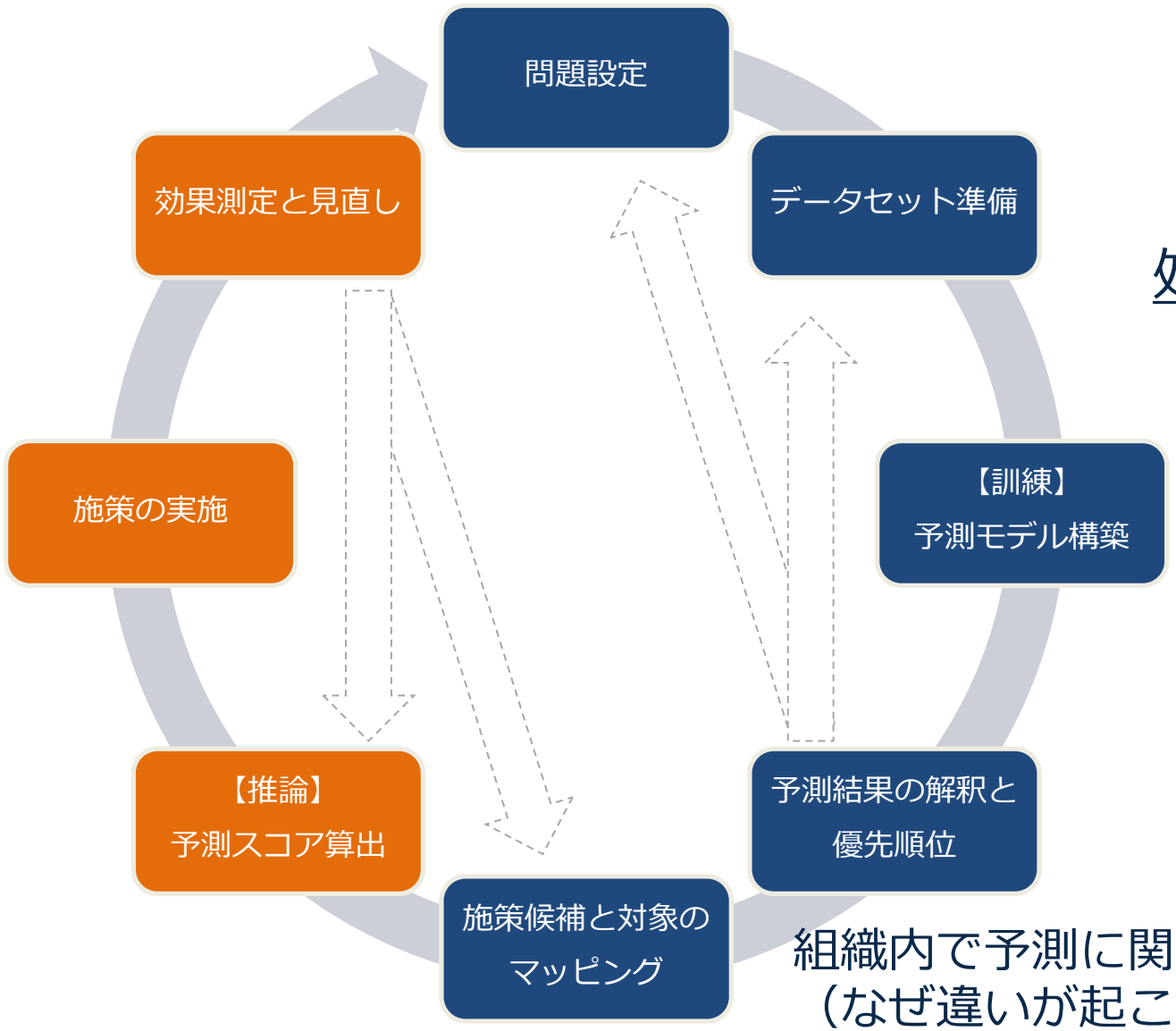
データから事業インパクトを出す成果を挙げる

©GRI Inc.

ForecastFlowを利用した現代的な予測の業務活用

予期的推論

- 短期サイクル
- 効果のある戦術的な施策の実施



処方的アナリティクス

- 長期サイクル
- 事業インパクトを出す戦略的な仕組み作り

組織内で予測に関する共通認識形成
（なぜ違いが起こるのかを理解）

©GRI Inc.

機械学習用のデータセットの例

■自動機械学習では、データセットを準備しておく必要がある

ID
（予測する最小粒度）

予測ターゲット
（正解データ）

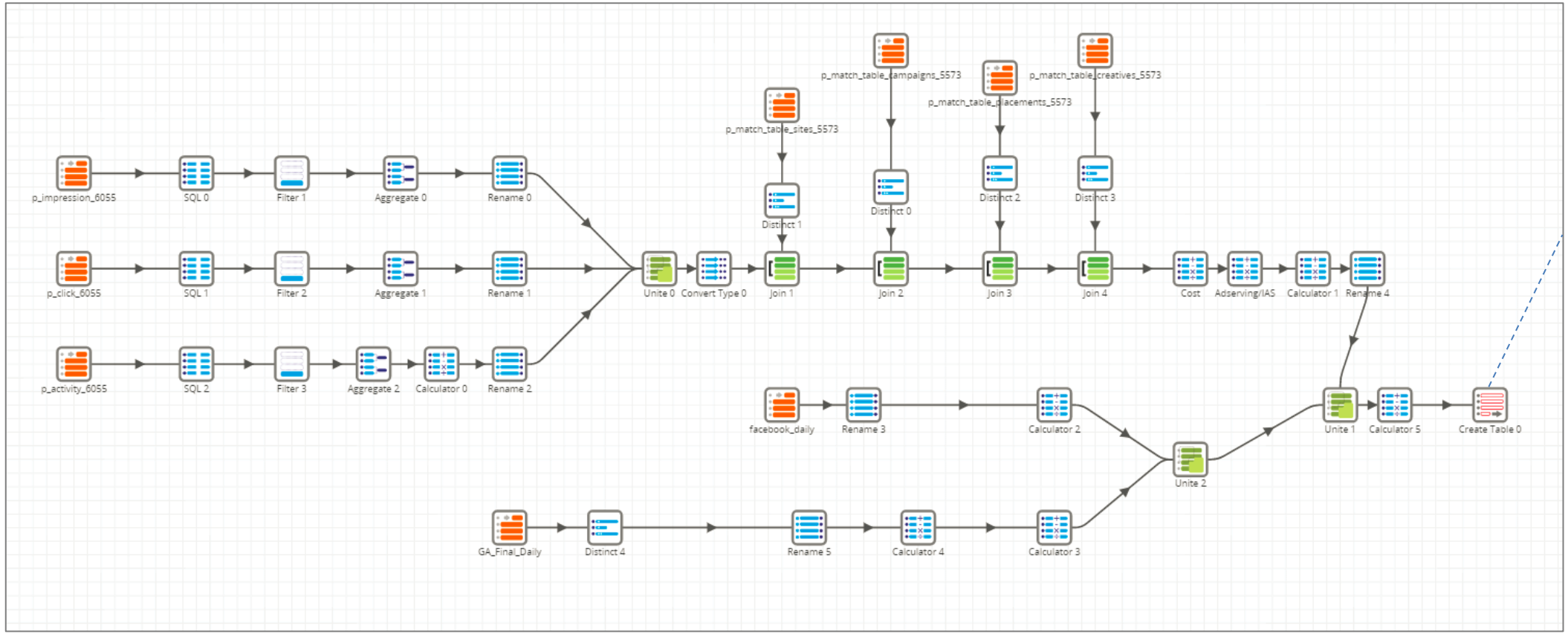
特徴量
（予測ターゲットを説明する変数）

回線契約ID	解約フラグ	STATE（州）	契約月数	国際通話契約	ボイスメールメッセージ数	日中通話時間【分】
4153824657	解約	CA	5	Yes	4	265.1
9814569871	継続	WA	36	No	12	48.2
3914835598	継続	TX	18	No	6	120.1
…						

©GRI Inc.

機械学習用のデータセットの準備の例

- 多くのテーブルから結合や集計を繰り返し、機械学習用のデータセットを作る
- 大規模だと時間がかかる
- 非エンジニアにとって、機械学習のデータセットを作るための特徴量エンジニアリングがつらい（列が大量にあるため）

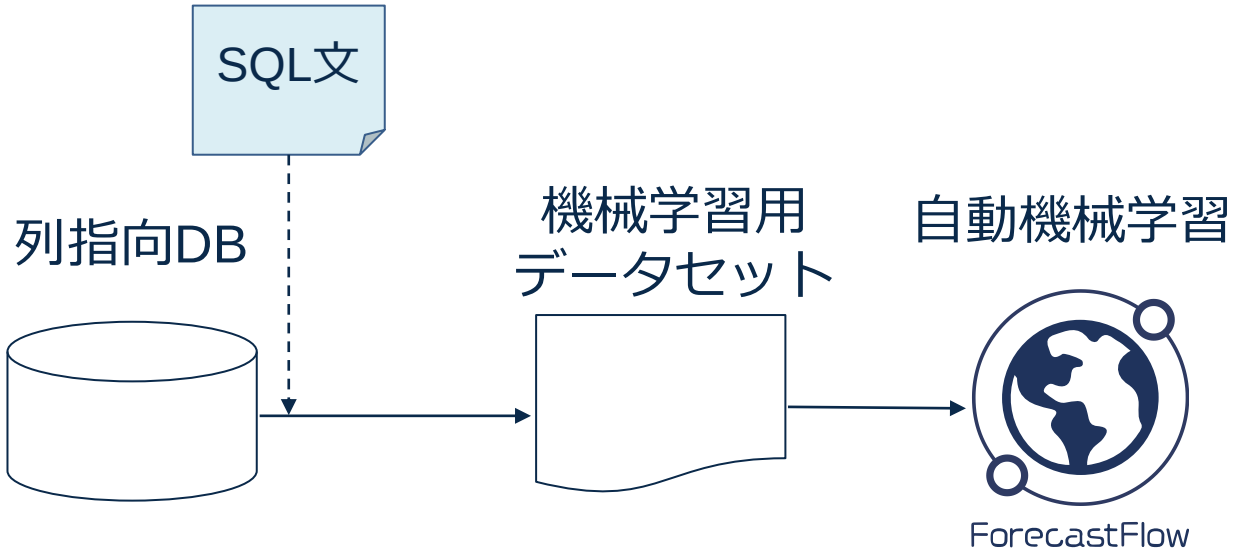


機械学習用
データセット

©GRI Inc.

列指向DBでの特徴量エンジニアリングの悩み１

- 機械学習のデータセットは列を多く作る必要があり、列指向DB上に直接SQL文を書くと、スパゲッティのように複雑にからまり、他の人には可読性が悪い
- データサイエンティストはSQLを記述できた方が良いが、もう少し楽をするべき

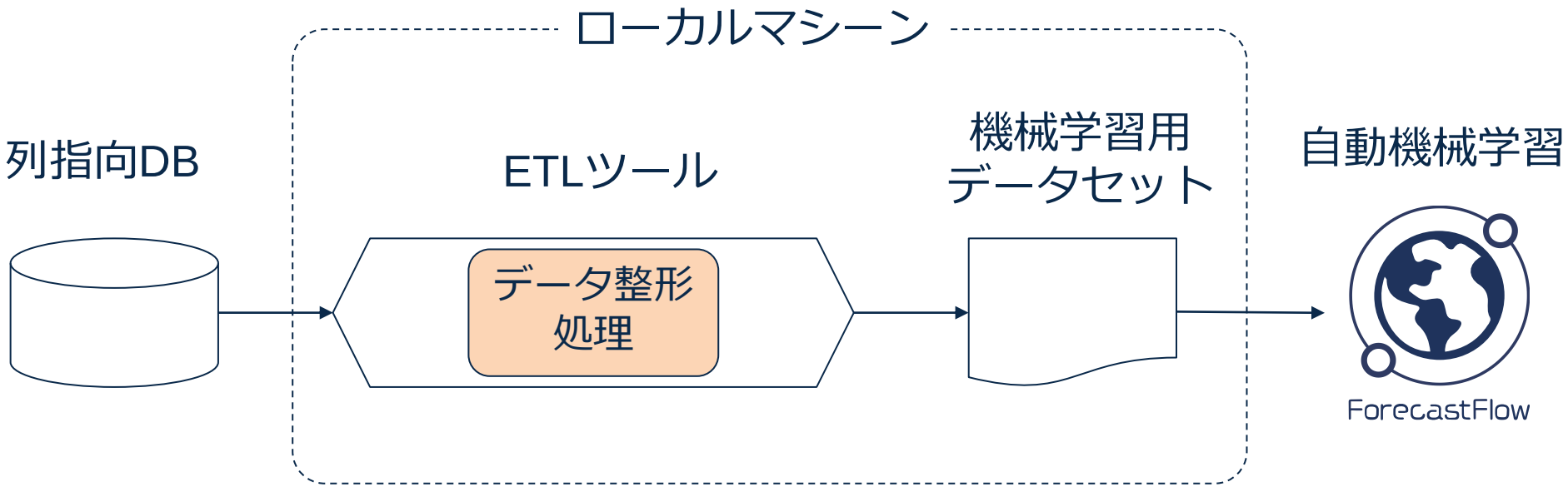


©GRI Inc.

列指向DBでの特徴量エンジニアリングの悩み２

- 列指向DBとETLツール（AlteryxやTableau Prep）を接続して、機械学習のデータセットをETLツール上で記述する方法（データ整形処理の可視化）
- コンポーネントがあっても、実質的にデータを抽出してから、ローカルマシーンの上でデータ整形処理を行っているため、PCのスペックに依存し、処理時間がものすごくかかる（列指向DBの良さを消している）

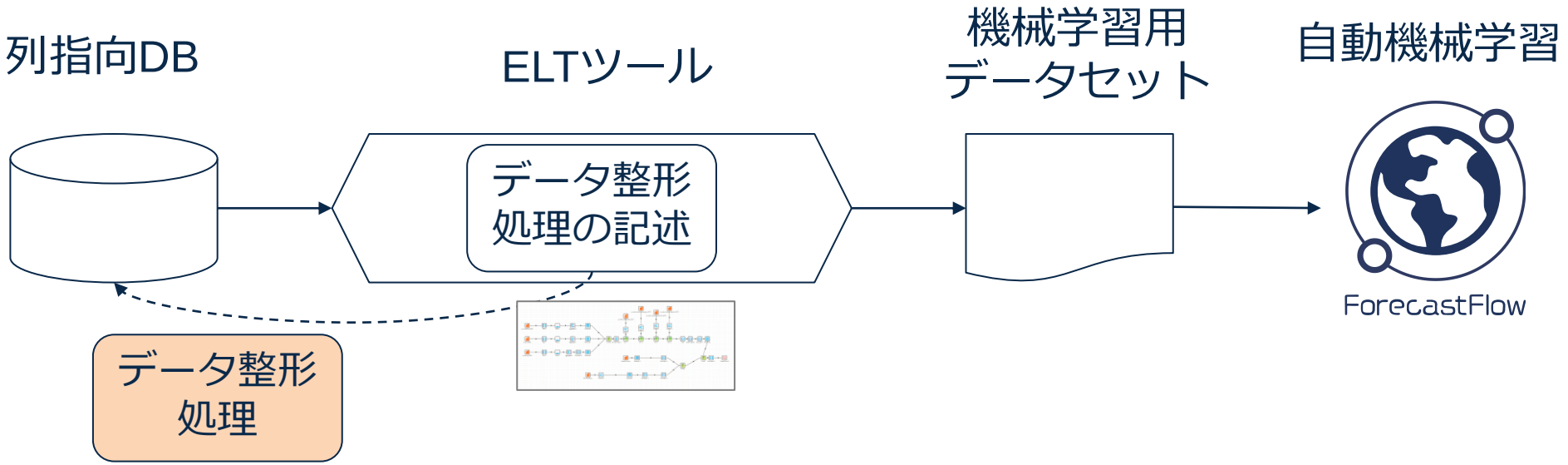
※AlteryxのIn-DBが使える場合、この心配は無用（Redshift-Alteryxなど）



©GRI Inc.

Matillionでの解決方法

- データ整形処理をグラフィカルに記述するが、データ整形処理自体を列指向DB側に任せる（BigQueryのパワーを使える）
- クラウド処理でアカウントを使いまわせる（named accountではなく、ローカルマシーンにインストール不要）



©GRI Inc.

Matillionを活用して自動化する箇所の例

列指向DB

訓練

ForecastFlow

列指向DB

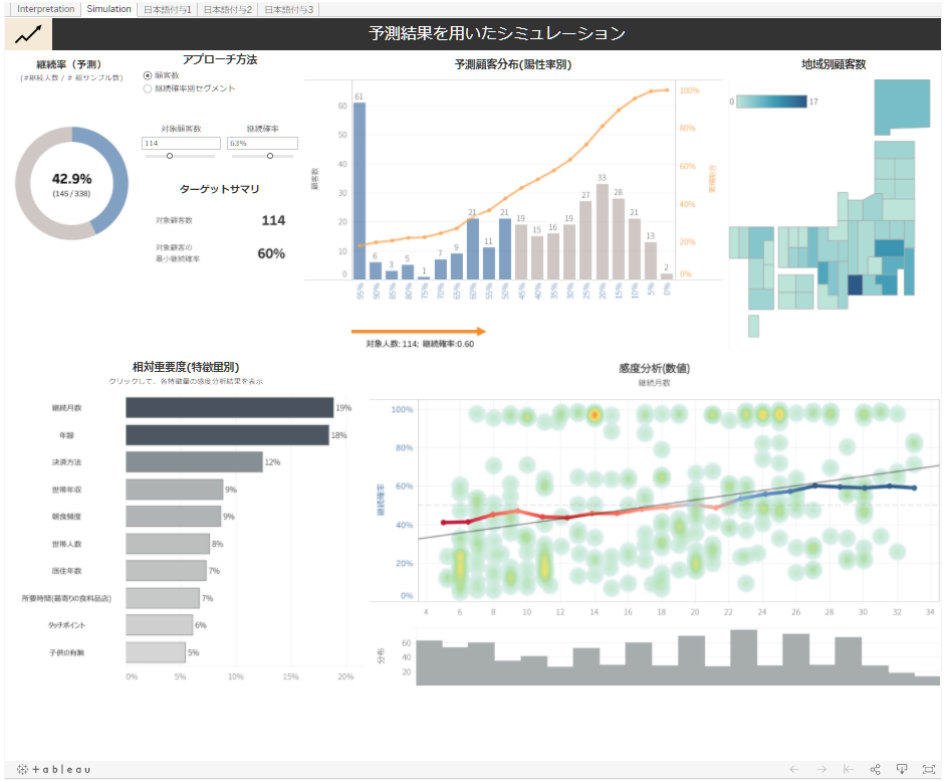
推論

ForecastFlow

可視化ツール

+ableau SOFTWARE

施策の検討



©GRI Inc.

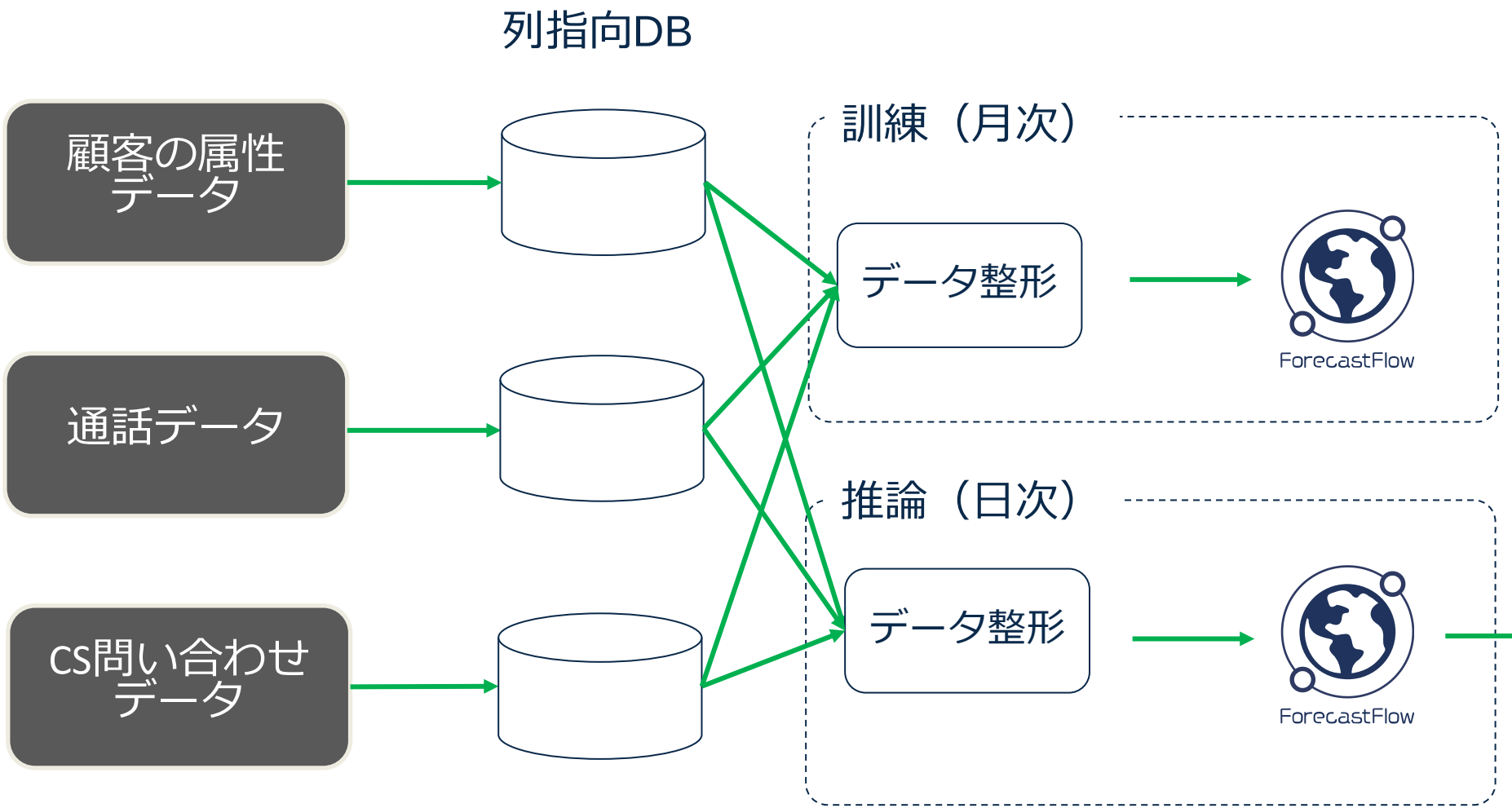
デモ

デモのシナリオ～ある米国の電話会社～

- 解約が増えていることが悩み
  - 解約しそうな顧客を見つけ出し、アプローチすることに
  - 機械学習モデルを作ったところ、解約しやすさ（解約スコア）は通話時間やCS問い合わせ回数などによって主に決まっていることが判明
  - 日々データが更新されるので、当然顧客ごとの解約スコアは毎日変動する
  - 毎日その日のアプローチ対象となる顧客を選定したい

©GRI Inc.

データフロー全体



©GRI Inc.

デモ

©GRI Inc.

デモのまとめ（BigQuery-Matillion-ForecastFlowのメリット）

■データ整形のフローがグラフィカルでわかりやすく、非エンジニアの人でも保守がしやすい

■解約スコアの算出といった日次処理を自動化

■大規模データに対する前処理を列志向DBのパワーを活かしてできる

■参考文献

–http://forecastflow.jp/datascicene/report3/

©GRI Inc.

まとめ

まとめ

- 大規模データの場合、列指向DBの力を活用するのがポイント
- 列指向DBで機械学習のデータセットをSQLのみで管理するのは、非エンジニア（ビジネス現場の人材）にとって非効率
- MatillionなどのELTツールを使うと、列指向DBの力を使いつつ、データフローの可視化ができ、属人化を防ぐことができる
- さらに、ジョブのスケジューリング設定が簡単にできるため、自動機械学習の推論結果を可視化する所まで、自動化できる

©GRI Inc.

質問タイム

- 音声でもチャットでも大丈夫です

- はじめに今回のトピックに関係する質問を受け付け、その後、今回のトピックに関係ない質問をお願いします

©GRI Inc.

ForecastFlowのアナウンス

- モデル分割機能でのバグ修正
  - bool型の特徴量が選択できなかった課題を解決

©GRI Inc.

次回すいすい会の内容予定

- 21/2/24（水）12:00~
- ECサイトでの自動機械学習の利用例
- 特徴量エンジニアリング（特徴量のNullと0の違いの意味）

©GRI Inc.

残っている質問（今後、説明していきたい内容）

- その他のビジネスの実践例を知りたい
- どのくらい正解データの数があるべきか
- 未来の施策と機械学習の関係性
- ダウンサンプリングとProbability Calibration
- 効果検証
- 様々な自動機械学習ツールの比較
- 様々なETL/ELTツールの比較
- Feature Store、分析基盤、CDPの説明

©GRI Inc.

体験アカウントの配布

- 訓練する際、最大1,000行になります
- トライアル・アカウントの詳細ページ
  - https://forecastflow.jp/free/

- 申し込みメールを下記へ
  - forecastflow@gri.jp

  - ForecastFlow無料アカウント申請フォーム
    - お名前：
    - 所属名：
    - 部署名：
    - ForecastFlowで利用するメールアドレス：
    - イベントパスワード： 2020

©GRI Inc.