データサイエンス

すいすい会

第31回「【データ活用の実務担当者向け】クロス集計から自動機械学習へ ～ノーコードAI時代のインサイト発掘」

GRI
BIGDATA SCIENCE
&RESEARCH

すいすい会の紹介

■データサイエンスに関する知見を共有しあい
　みんなで実践的な理解度を高めたい

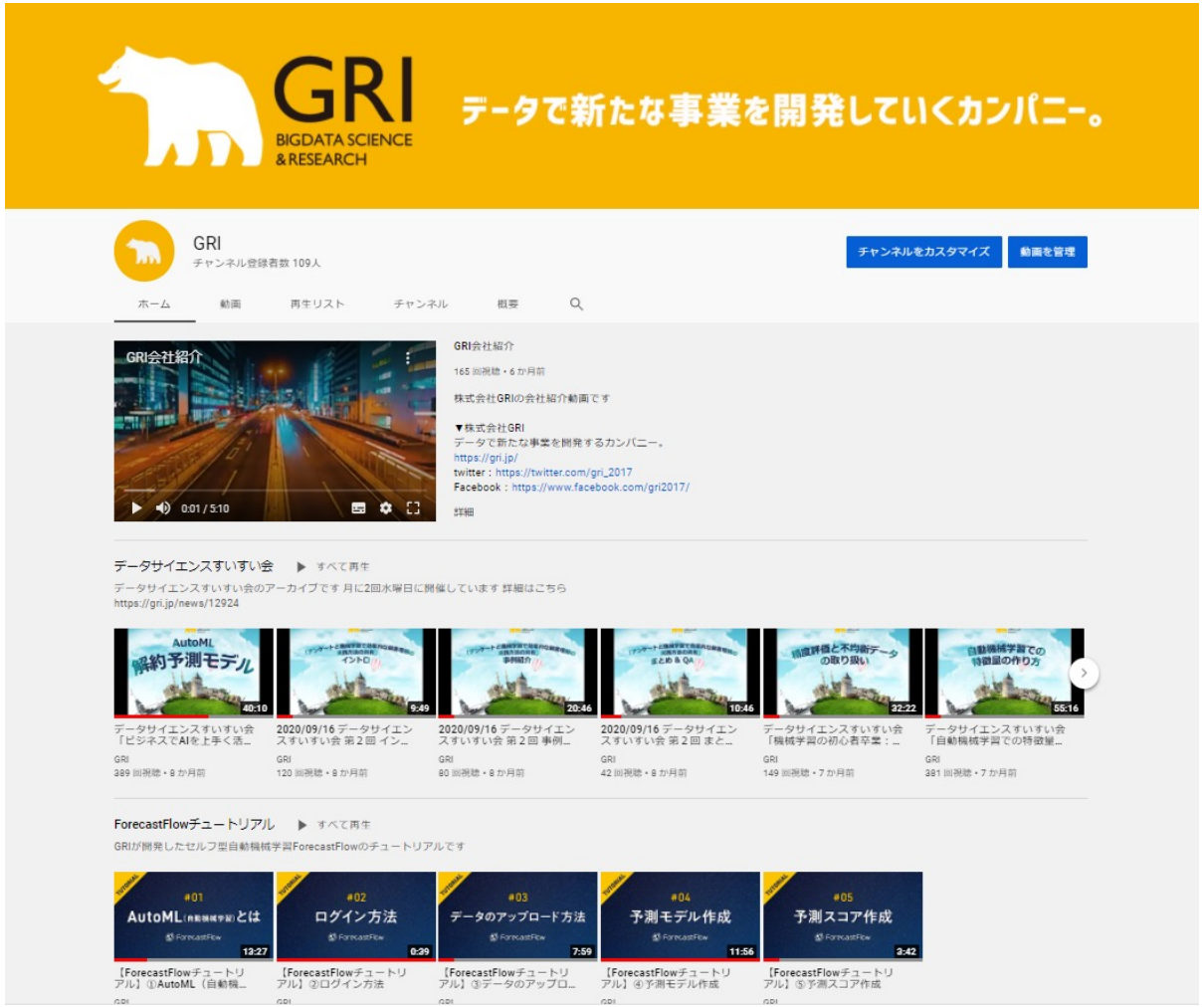
- 実践を重視しているGRIなりの考え方を共有します
- 集合知にするため、皆さんの知見をご共有いただきたいです

■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

過去のすいすい会の動画（YouTube）

https://www.youtube.com/c/GRIinc



ナビゲータの背景

古幡征史
株式会社GRI　取締役

- 株式会社GRI 取締役
- Ph.D in Computer Science
- GRIにて50以上のAI, BI, 分析基盤構築プロジェクトをリード
- KPMGコンサルティング、University of Southern California、ドワンゴを経て、2016年9月より現職

©GRI Inc.

データサイエンス

すいすい会

第31回「【データ活用の実務担当者向け】クロス集計から自動機械学習へ ～ノーコードAI時代のインサイト発掘」

GRI
BIGDATA SCIENCE
&RESEARCH

本日のアジェンダ

1．データ活用の実務担当者向けの自動機械学習の位置付け
　　株式会社GRI　古幡征史

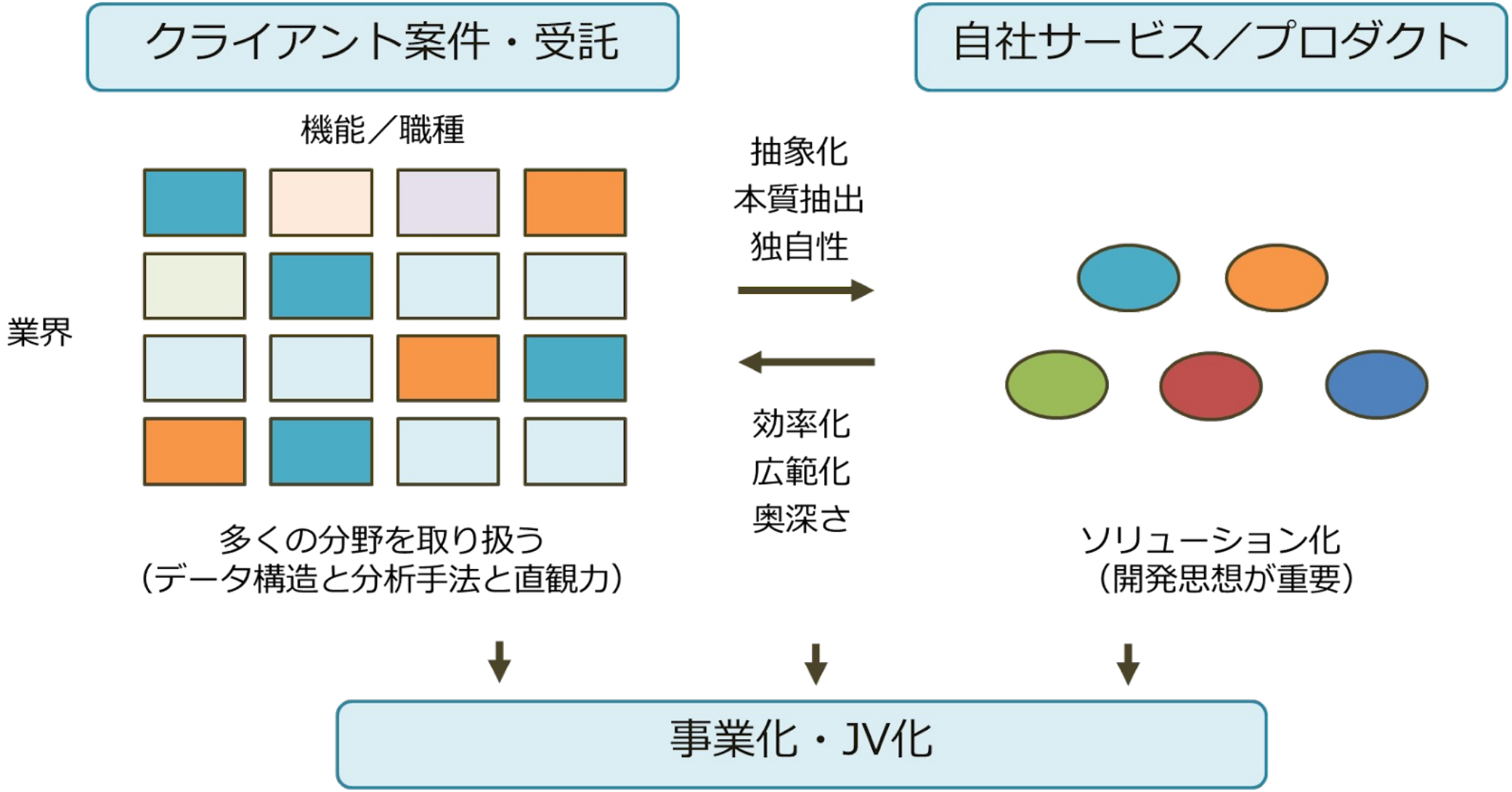
2．クロス集計から自動機械学習へ～ノーコードAI時代のインサイト発掘～
　　株式会社GRI　岡部泰三

©GRI Inc.

1. データ活用の実務担当者向けの自動機械学習の位置付け

©GRI Inc.

GRIでの受託案件とプロダクトの関係性の紹介



自動機械学習ツールは全自動で上手くいく訳ではない

- 「自動機械学習ツールを使えば、ボタン１つでAIが全部解決！」という営業トークに出会ったら怪しいと思っていただければ

©GRI Inc.

データ活用の実務担当者が自動機械学習ツールと相性が良い

- データ活用の実務担当者
  - 営業戦略／マーケティング戦略／サプライチェーン戦略
  - ビジネス計数
  - デジタルイノベーション
  - DX推進
  - 社内コンサルタント
  - 業務コンサルタント

- ビジネスやデータのドメイン知識を持ち、データの扱いが上手い人！

Google Analytics　tableau　Power BI　alteryx　SQL

自動機械学習ツールは様々な使い道がある

	販売・ マーケティング	経営企画	R&D・製造	調達・物流	サービス	人事・総務・他
業界共通	売上予測	経営シナリオ別の 売上予測	IoTセンサーを利用した 次世代商品開発	出荷予測	故障予測 コールセンター入電予測 &難易度予測	離職予測 採用予測 従業員成果予測 従業員満足予測
サービス	SaaS会員解約予測 リードナーチャリング LTV予測 ユーザ属性推定 コンバージョン予測 TV視聴率予測 DM反応予測 消費者の好みの予測	売上予測 会員数予測 未払い退会予測			施設の利用予測	NPS予測
金融・保険	最適な訪問営業予測 与信予測	倒産予測			信用スコア 不正利用検知	
製造・ インフラ・建設	電気・ガス契約予測 不動産価格予測	電力・ガスの需要予測	土木での地質予測 設備効率予測 故障予測	発注商品価格の 妥当性判定	流量予測 故障地域予測	
流通・運輸・消費財	顧客の離反予測 RFM予測 店舗売上予測	出店用の来店予測	次世代自動車開発			
農業・畜産	商品市場での 取引価格予測		飼育牛の発情・分娩・疾 病予測			スタッフ 配置予測

©GRI Inc.

分析テーマをプロジェクトキャンバスで定義

- 大企業にように意思決定が複雑な場合、分析テーマの取りまとめが重要になる
- よくあるのが、そもそも何のためにこのプロジェクトやっていたんだっけ？
- 自部門以外の分析テーマの「やるやら」を上手く決めたい
- 三菱ケミカルさんの「プロジェクトキャンバス」をベースに、GRIのいくつかのプロジェクトで修正した例を共有
- 初期分析をした後にプロジェクトキャンバスが出来上がるイメージ

三菱ケミカル社のプロジェクトキャンバス　https://www.mitsubishichem-hd.co.jp/news_release/pdf/190718.pdf

©GRI Inc.

プロジェクトキャンバス（修正版）の書式例

ビジネスサイドの記載
データ活用の実務者の記載

分析テーマ名

分析内容が第三者でも分かりやすい名称

誰のため、何のため

この分析結果でメリットを受ける人と目的を記載（関与している人がワクワクするようにする）

ビジネスインパクト

数値的に見込まれる

想定分析期間

必要に応じてフェーズを分ける

着手判断

着手／待ち／キャンバス修正／却下

分析頻度

週次／隔週／月次／四半期／年次／ワンショット

精度測定法（ゴール）

数値の場合は計算式；質的な場合は完成の状態

成功基準

精度測定法に対応した成功条件を、どのタイミングで評価するか事前に決めておく

分析要件の整理（主担当）
担当者を記載

データ準備（主担当）
担当者を記載

データ分析（主担当）
担当者を記載

施策検討（主担当）
担当者を記載

データソース

使えそうなデータソースをリスト化

リスト化できない場合は、データソースを探したり、パイプライン構築に時間がかかりそうな旨を記載

分析上の課題点

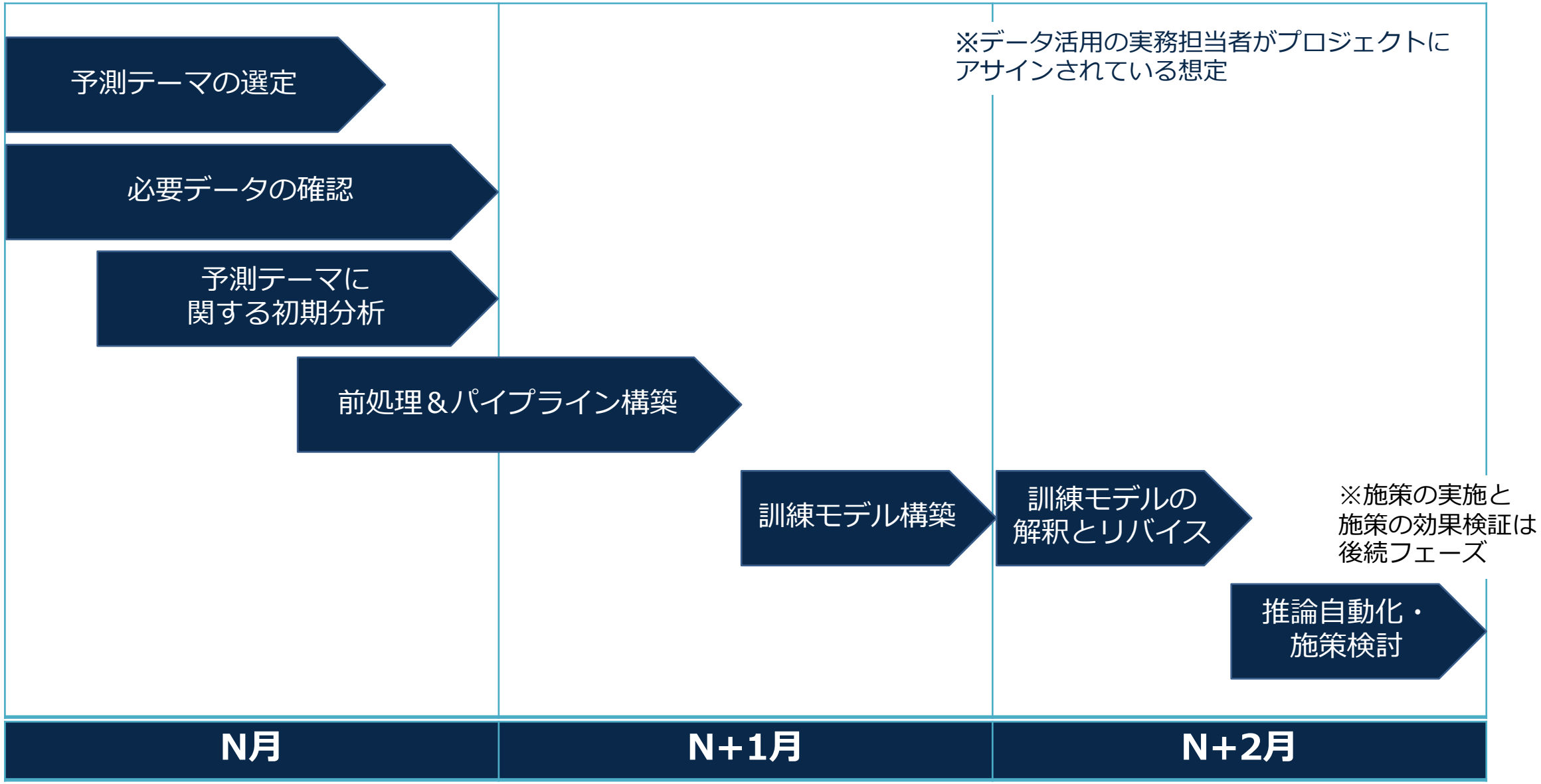
※キャンバスを埋める際に不確実な点を記載

分析の中で取りたい想定アクション （or 結果として取りたいアクション）

事前に分かっている場合に記載（分からない場合には、分析によって得られたアクションとの違いを比較）

©GRI Inc.

自動機械学習の導入に必要な期間（典型例）



自動機械学習ツールは多くあるので選定基準が重要

■ 選定基準の例

- クラウド基盤から自動で予測しやすい
- 訓練結果の解釈のしやすさ
- 大規模データ（特徴量の数とデータの行数）に対する速度と課金額
- チームメンバーとの機械学習結果の共有
- 分析エコシステムへの組み込みのしやすさ（疎結合）
- 予測精度（普通の自動機械学習ツールは、精度が似ている）

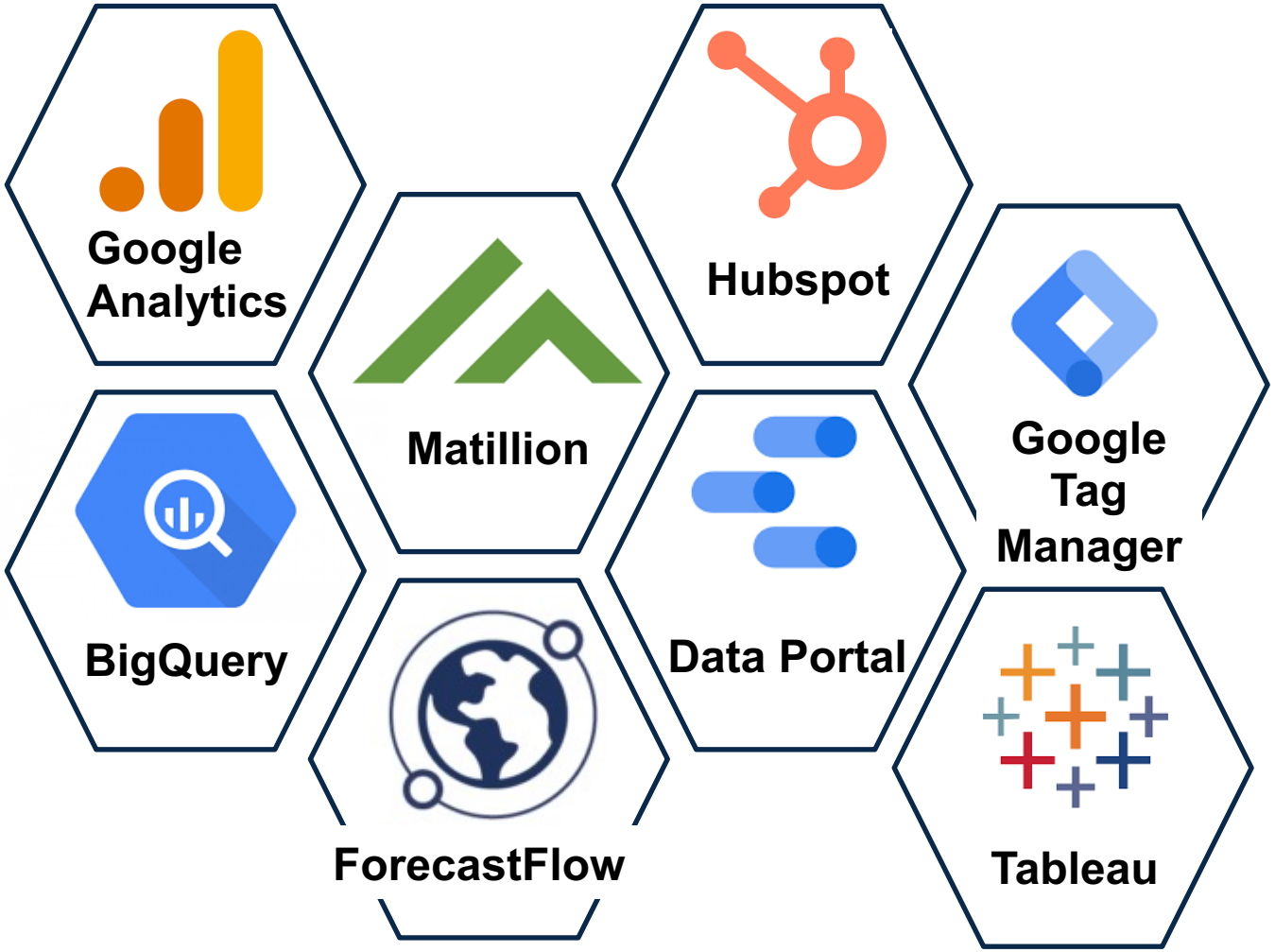
自動機械学習ツールの比較は簡単にできるので、解釈性の判断に使いやすい電話会社のデータセットをForecastFlowのサンプルデータセットを格納しています

- ビジネスデータとして分かりやすい
- 相関１の特徴量を含む
- カテゴリ特徴量と数値特徴量を含む
- 正解データのバランスが良い

©GRI Inc.

ForecastFlowを使い倒している例

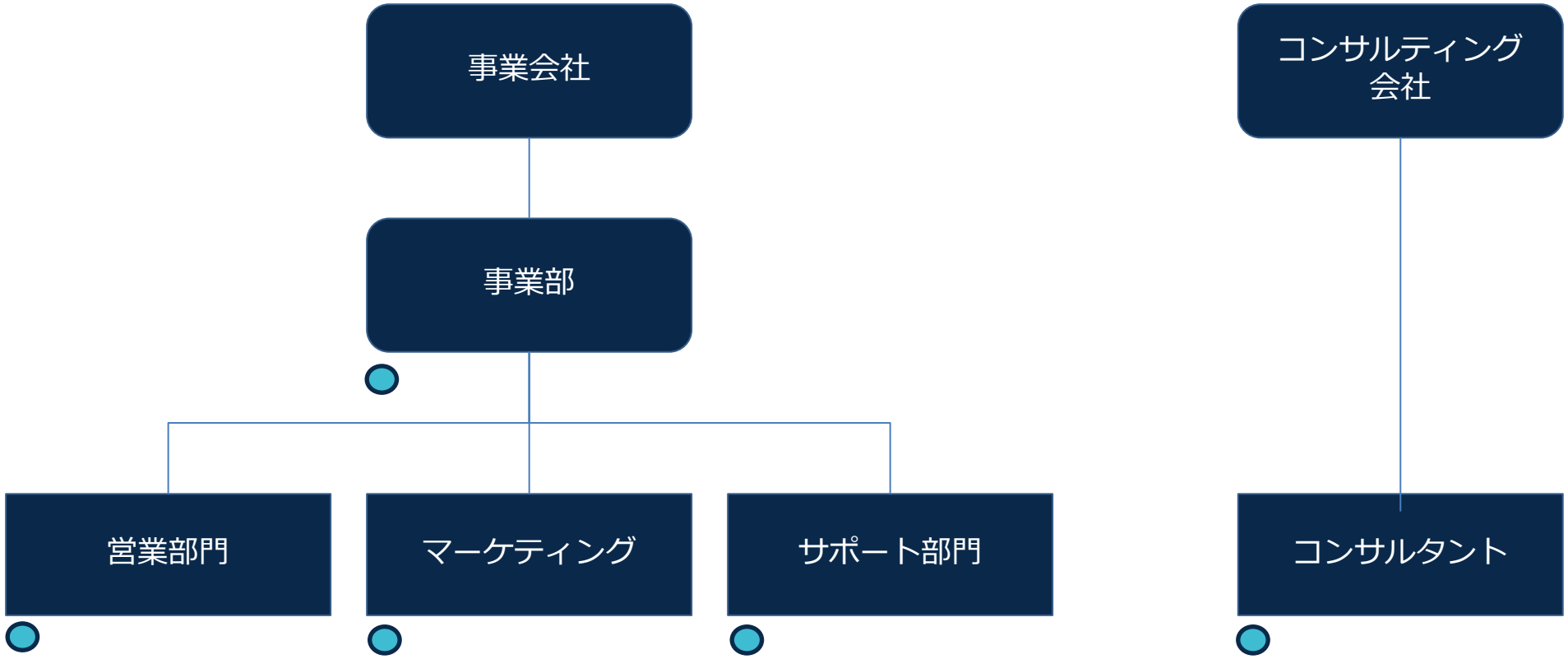
- 訓練モデルの高速処理
- 毎日、推論を高速処理（下記を数時間並列処理）
- 予測モデルの種類：数十
- 予測モデルの特徴量数百列
- 数千万行の推論データ



©GRI Inc.

ForecastFlowを使い倒している例

- 標準契約（月額10万円）では、１つの契約（チーム）につき10名（メンバー）まで利用可能（企業を超えて）
- ただし、１つのチームでは、データが完全共有される



©GRI Inc.

2. クロス集計から自動機械学習へ
～ノーコードAI時代のインサイト発掘～

©GRI Inc.

イントロダクション

こういうケースでどう分析を始めますか？

■ 米国の電話会社の例

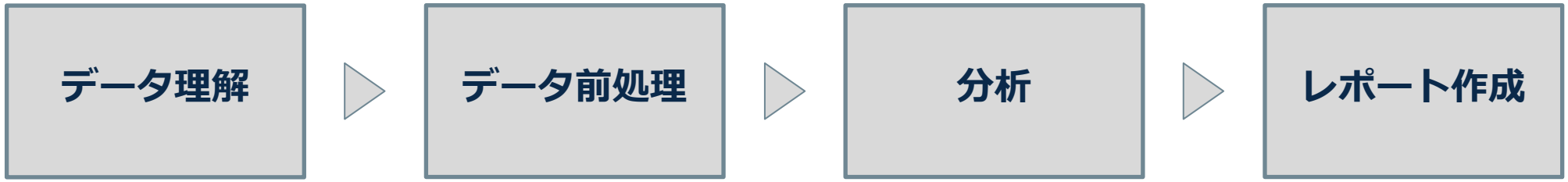
–解約者数が急増した

–以下のような顧客データがあるので原因を分析したい

©GRI Inc.

データ分析の典型的なプロセス

- 典型的には以下のようなプロセスで分析
  - ここではワンショットの分析を念頭に置いている



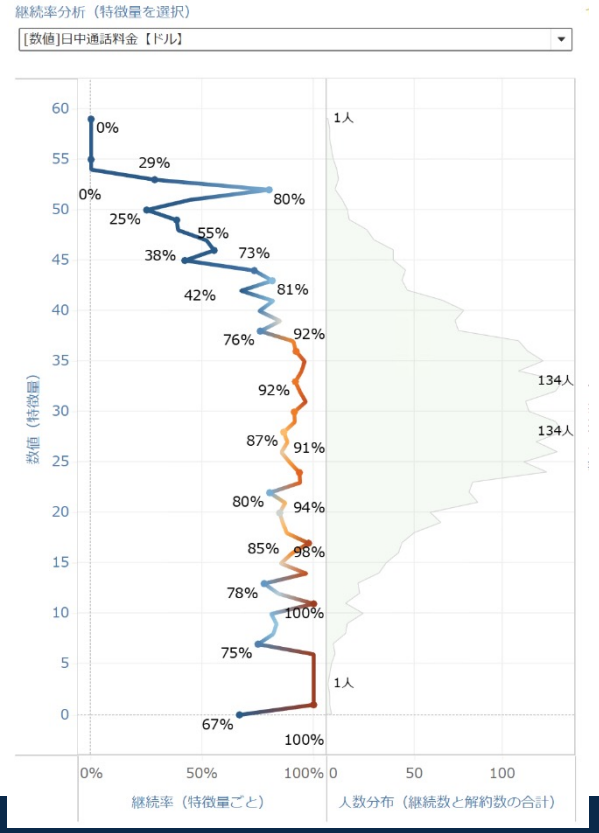
©GRI Inc.

データ分析の典型的なプロセス

■ 分析パートでは実務者はクイックに結果を確認できるExcelやBIを好むことが多い

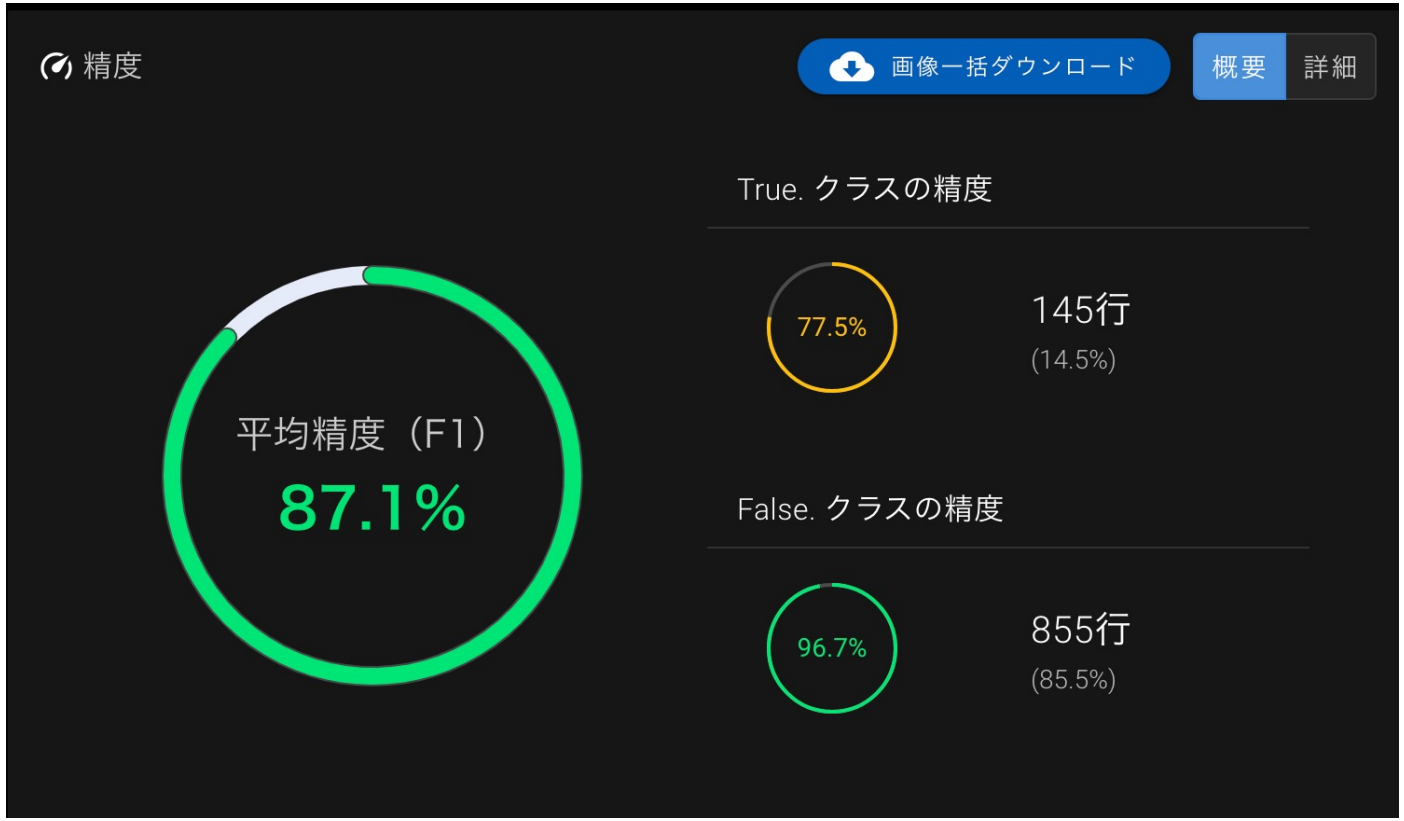
データ理解 ▶ データ前処理 ▶ **分析** ▶ レポート作成

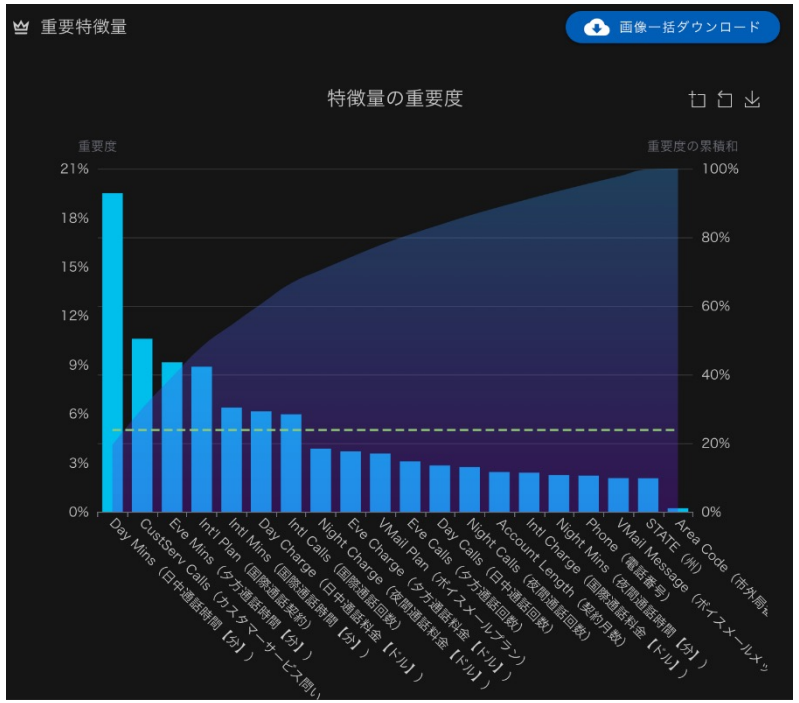
STATE（州）	Account Length（契約月数）	ID	Area Code（市外局番）	Phone（電話番号）	Int'l Plan（国際通話契約）	VMail Plan（ボイスメールプラン）
KS	128	4153824657	415	382-4657	no	yes
OH	107	4153717191	415	371-7191	no	yes
NJ	137	4153581921	415	358-1921	no	no
OH	84	4083759999	408	375-9999	yes	no
OK	75	4153306626	415	330-6626	yes	no
AL	118	5103918027	510	391-8027	yes	no
MA	121	5103559993	510	355-9993	no	yes
MO	147	4153299001	415	329-9001	yes	no
LA	117	4083354719	408	335-4719	no	no
WV	141	4153308173	415	330-8173	yes	yes
IN	65	4153296603	415	329-6603	no	no
RI	74	4153449403	415	344-9403	no	no

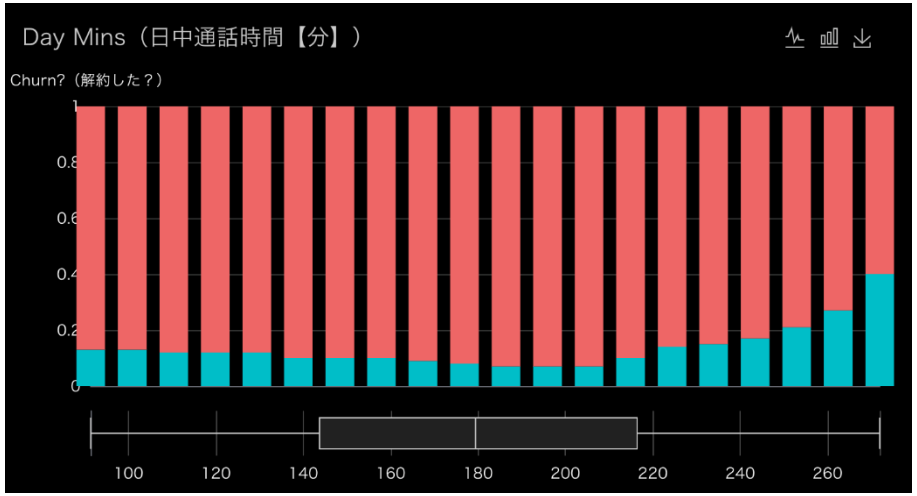


実務家をサポートする自動機械学習ツールという選択肢

■ +αで自動機械学習という選択肢





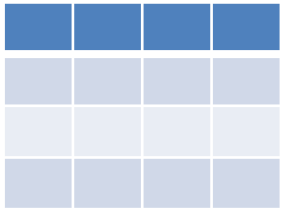


なぜ機械学習か

表形式データの機械学習

実務で活用事例の多い機械学習は大きく以下の3パターンで本日は特に表形式データに焦点を当てる

表形式



1行ごとに対し予測を実行、ビジネスデータで有用なケースが多い
分類問題と回帰問題の2種類
例）サブスクビジネスの解約予測、ユーザーごと購買金額予測
本日の対象：実務で頻出の形式で、機械学習によるメリットを享受しやすいため

時系列

時系列の数値予測
例）需要予測、株価予測、感染者数予測

深層学習

膨大なデータから機械的にパターンを見つけ出す
分類から回帰、生成まで幅広い応用
例）犬と猫の画像を分類、音声合成、機械翻訳

©GRI Inc.

機械学習活用の3つのフェーズ

機械学習活用は大きく以下の3つのフェーズに分かれる

インサイト発掘 （訓練フェーズ）	未知の予測・推定 （推論フェーズ）	施策の効果検証 （検証フェーズ）
・原因追求型（過去指向）	・将来予測型（未来指向）	・軌道確認型（現在指向）
・過去の結果からユーザーの態度変容（解約・ロイヤル化など）を予測するモデルを作成	・訓練フェーズで作成したモデルをもとに正解のわからない未知の変数を予測	・施策の効果は意味があるものだったのか検証し、次のアクションの意思決定に活かす
・施策を念頭に置いて問題設定・特徴量作成	・アウトプットであるスコアを元に施策の出し分けを検討	・スコアによる施策出し分けでフェアな比較が可能に

まずはインサイト発掘を念頭に話し、その後施策と効果検証での活用について触れる

©GRI Inc.

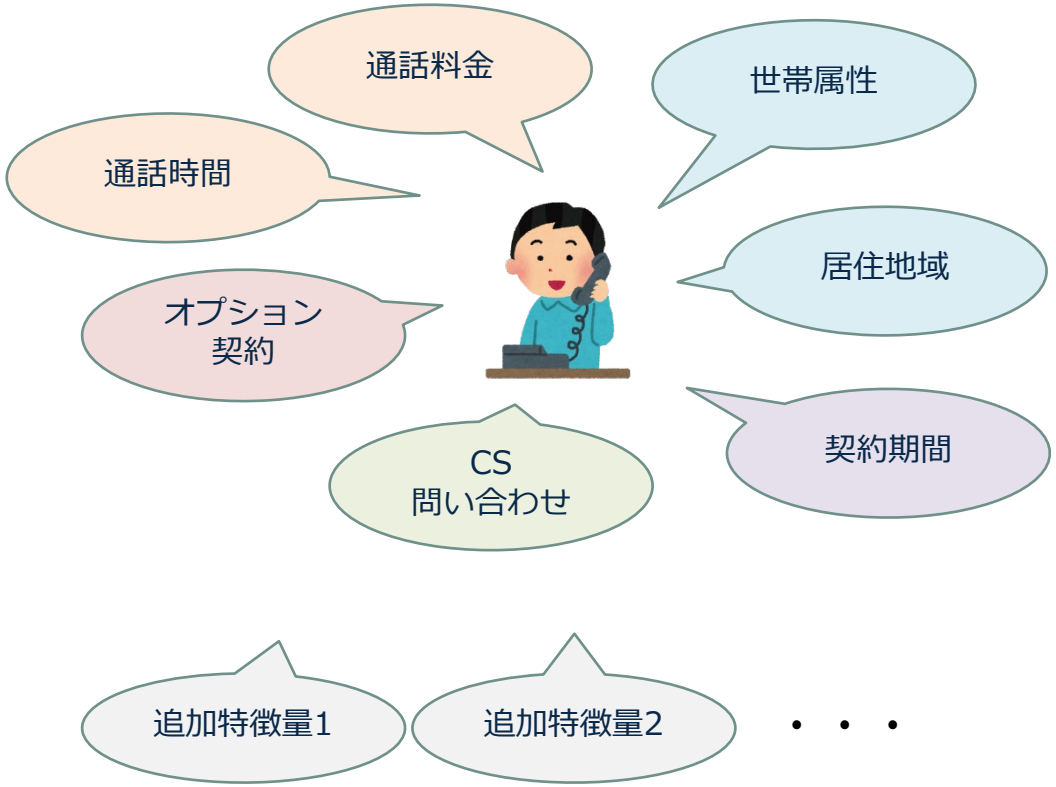
なぜ機械学習か？

1. 組み合わせ爆発の問題
2. 交絡因子の問題
3. 分析軸探索時の問題

©GRI Inc.

なぜ機械学習か？ 1.組合せ爆発の問題

特徴量の数が増えると指数関数的にその組み合わせが増加し検証不可能に



一般に特徴量は相互に関係しているためその組み合わせまで考慮して検証することが必須
もし仮に特徴量が100種類あるとすると、、、

・100種類の特徴量が仮に全てBOOL型として**2^{100}通り**
→1秒間に10G=10^{10}通り処理したとしても**約10兆年**
宇宙誕生から現在までの時間のおよそ1000倍

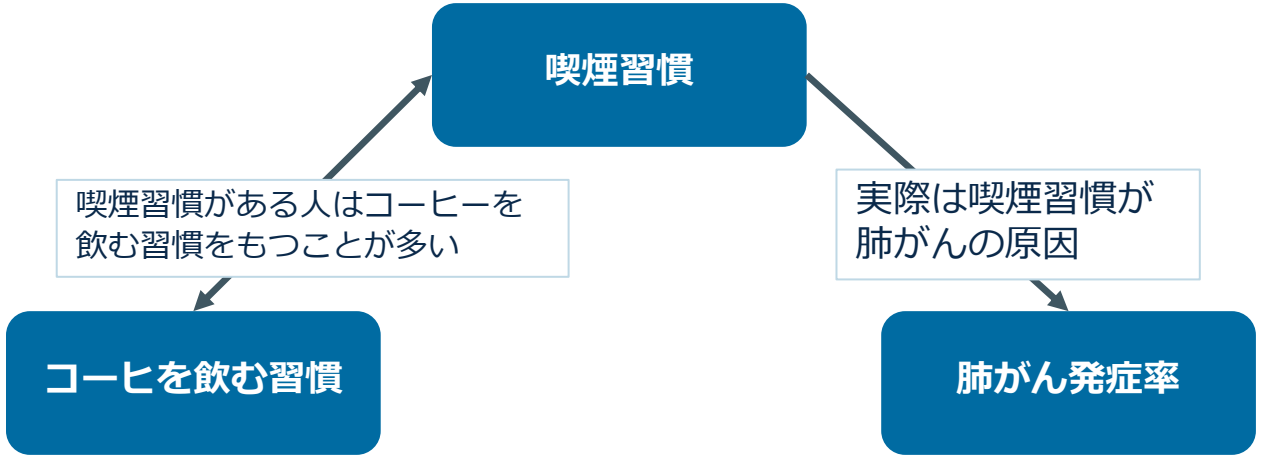
→アルゴリズムが必要

©GRI Inc.

なぜ機械学習か？ 2.交絡因子の問題

重要な仮説から順番に見ていけばいい？
→単相関だけで見ていると交絡因子の存在に気づかず、間違った結論を導く恐れがある

コーヒーを飲むほど肺がんになりやすい？



×コーヒーを飲むのはやめましょう

○タバコを吸うのはやめましょう

機械学習は100種類の特徴量を一斉に考慮するため、（完全とは言えないものの）各特徴量の相関を取り除いた結果が確認可能

©GRI Inc.

なぜ機械学習か？ 3.分析軸探索時の問題

大量に分析軸がある時でも人間の脳で同時に検討できるのは2-3軸が限界
→ドメイン知識に基づいて限られた範囲を探索することに

性別	年代
家族構成	居住地域
職種・職位	世帯・個人年収
ライフスタイル	価値観
購買履歴	ポイント履歴
Webアクセス	契約履歴

・
・
・

・例えばクロス集計では2-3軸を同時に検討していき特徴的なセグメントを洗い出す（この時の分析軸はドメイン知識に基づき当たりをつけて選定）

・結果として膨大にある分析軸から限られた探索範囲のみを選んで結果を出すことになり、特徴的なセグメントの発見を見逃してしまう恐れがある

・かといって広範囲の探索は一般的に時間がかる、、、

→上記のようなドメイン知識に基づく集計分析は基本所作なので必要不可欠なのだが、一方で自動機械学習による広範囲の探索結果があると結果をレポートするときに自信が持てる

©GRI Inc.

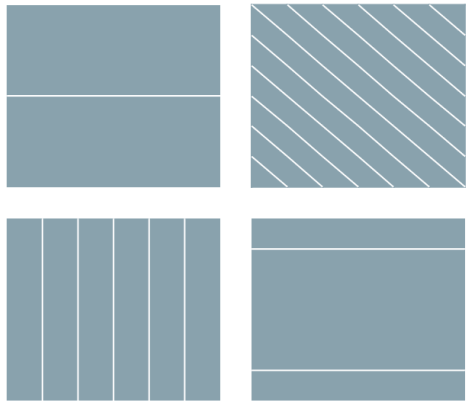
なぜ機械学習か？

分析の目的はパターンを見出すことで、その本質は要素分解し意味のあるまとまりで統合すること
→このうち要素ごとの差異が大きくなるような**分析軸の探索**は機械学習が得意なポイント

分析対象母集団

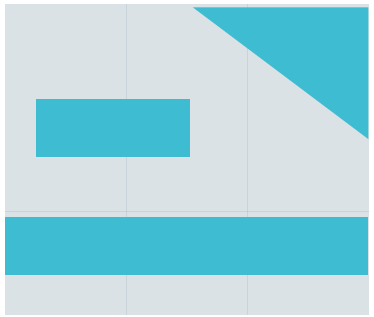
分析軸で要素分解

分析軸（ディメンション）で分解された各要素ごとに集計値（メジャー）を計算



要素を統合・選定

意味のあるまとまりで要素を統合、施策などのアクションにつなげる



ちなみに、新たな分析軸を作成することは
「特徴量エンジニアリング」と呼ばれ
それはそれで重要なトピック　　（物理では座標変換）

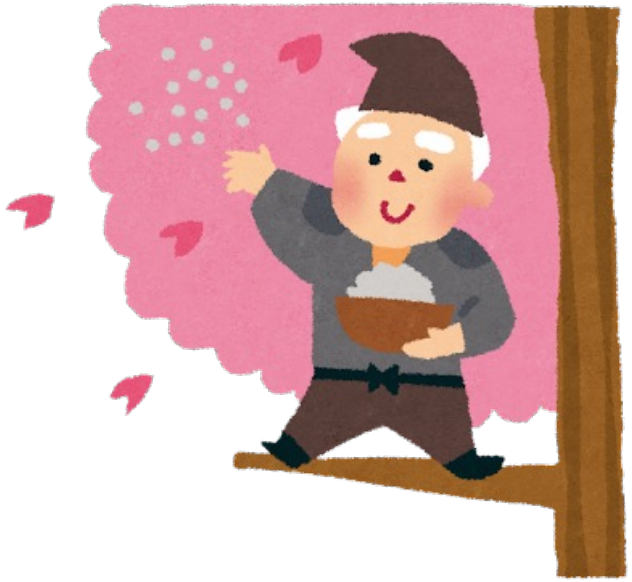
©GRI Inc.

なぜ機械学習か？

分析の目的はパターンを見出すことで、その本質は要素分解し意味のあるまとまりで統合することにある
→このうち要素ごとの差異が大きくなるような分析軸の探索は機械学習が得意なポイント

要するに、掘るべき場所を教えてもらう

機械学習



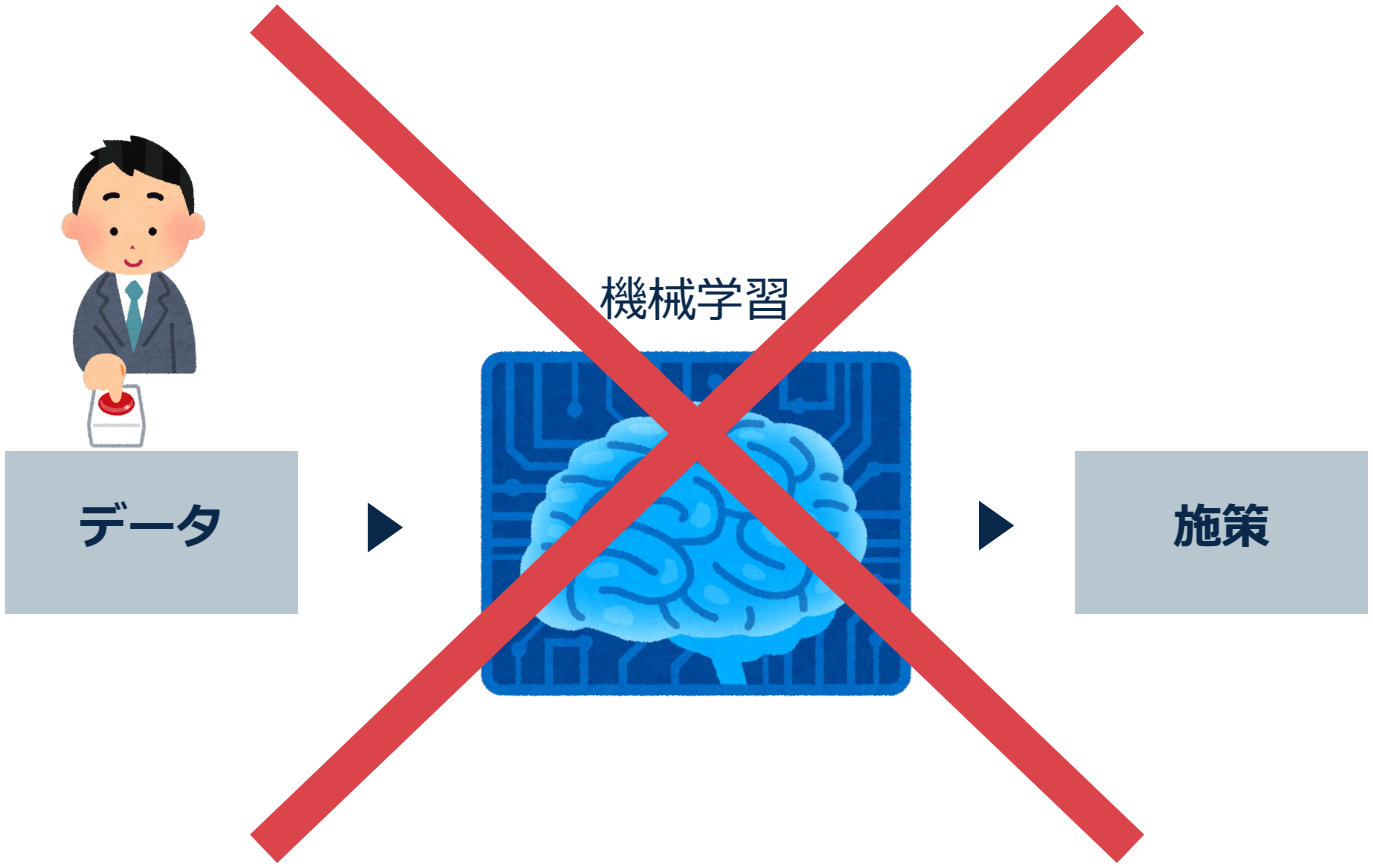
分析担当者

©GRI Inc.

データ分析の実務担当者が検討すべきこと

自動機械学習に関してのよくある勘違い

自動機械学習はデータを入れてボタン一つ押せば全て解決な万能ツールではない



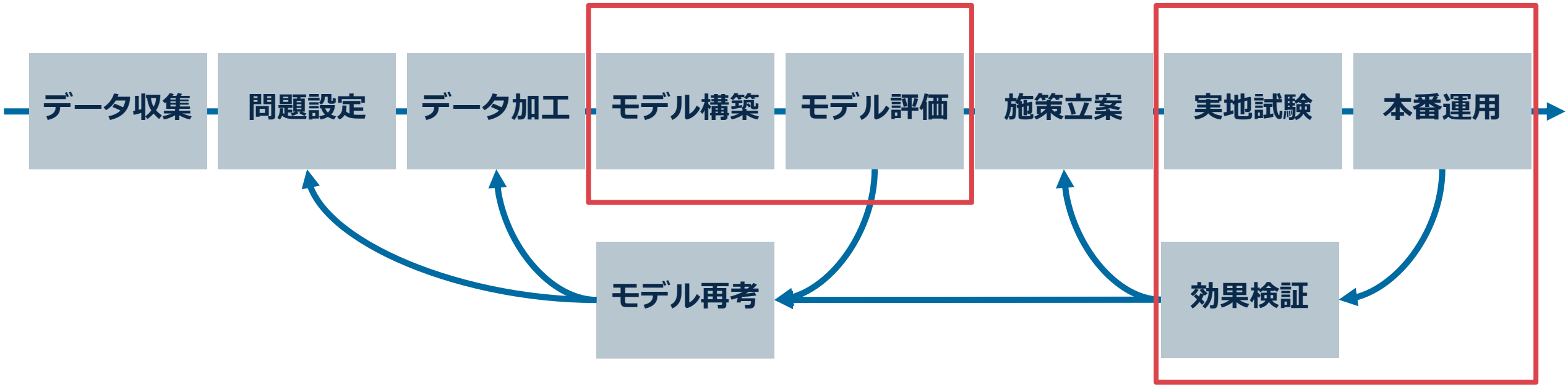
自動機械学習の正しい活用イメージ

実際は以下のような限られた範囲で有用性を発揮するもの

→何の問題を解かせるのか？そのためにどんなデータを準備するか？結果からどんな施策を打つのか？といった項目は引き続き実務担当者が検討すべき事項

→機械学習を使いこなして成果を出せる人材（=データ活用の実務担当者）は引き続き活躍

むしろ自動機械学習という新たな武器を得ることで分析スピードを加速できる



©GRI Inc.

トランザクションデータ活用検討時の例

例えばトランザクションデータの要素をモデル化すると以下のような構造

ひと	×	もの	×	時間	×	行動
行為主体全般 例えば以下を想定		行為対象全般 例えば以下を想定		事象が発生したタイミング・順序		ひとがものに対してとったアクション
・ユーザー ・社員 ・会社 ・動物（IoT） ・車・ロボット ・計測機器（IoT）		・商品 ・サービス ・動画/音楽/映画 ・記事				・購買/返品 ・閲読/読了 ・契約/解約 ・入職/離職/転職

©GRI Inc.

トランザクションデータ活用検討時の例

ひと × もの × 時間 × 行動

具体例では以下のようなイメージ

ユーザーが動画視聴サイトで動画を視聴

ユーザーがWebメディアの記事を閲読

コンビニの利用客が商品を購入

©GRI Inc.

トランザクションデータ活用検討時の例

ただし「データが沢山あるから」といって下のような表をこのまま機械学習に入れても何も出てこない

ひと × **もの** × **時間** × **行動**

ひとID	ものID	Timestamp	行動
1	2	18:24:01	購買
1	3	18:24:05	購買
2	1	18:23:52	購買
2	1	18:24:10	返品
2	3	18:24:13	購買
3	5	18:24:15	購買
3	7	18:24:16	購買

⋮

©GRI Inc.

機械学習のための一枚表

目的に合わせて以下のような表を用意する必要がある

ID
予測を行う粒度

ターゲット
予測対象

特徴量
ターゲットを予測するための変数

ID	解約	来訪日数	購買金額	ブラウザ	OS	都道府県
1	解約	5	593	Chrome	Windows	Aichi
2	継続	61	189	Edge	Windows	Tokyo
3	継続	39	105	Safari	Macintosh	Nara
4	継続	57	38	Chrome	Windows	Tokyo
5	解約	90	79	Safari	Macintosh	Hyogo

©GRI Inc.

分析時にデータ活用の実務担当者が検討すべきこと

何を目的として、どの粒度で何をどのように集計して結果を出すか、は実務担当者の腕の見せ所

ひと × **もの** × **時間** × **行動**

集約要素
- 性別
- 年齢
- 婚姻ステータス
- 子供の有無
- 年収
- 職種・役職
- 会員/非会員
- ロイヤル/新規
- ・・・

集約要素
- 価格帯
- カテゴリ/ジャンル
- 製造元
- ブランド
- コンテンツの内容
- ・・・

集約要素
- 時間帯
- 朝昼夜
- 通勤時間帯
- 昼休憩中
- 季節性
- イベント期間
- 施策実施期間

集約対象
- 購買金額
- 購買回数
- 訪問日数
- 閲読/視聴時間

集約手法
- カウント
- 平均/中央値
- 最大/最小

©GRI Inc.

分析時にデータ活用の実務担当者が検討すべきこと

例えば分析時の「主語」を決める際はビジネス上の目的とデータ構造、双方の理解が必要

「ひと」が主語

ひとID	会員	性別	購買金額
1	会員	男性	3,000
2	非会員	男性	1,200
3	非会員	男性	1,000
4	会員	女性	4,000
5	非会員	女性	1,500
6	会員	女性	4,500
7	会員	男性	2,900
8	会員	男性	3,300

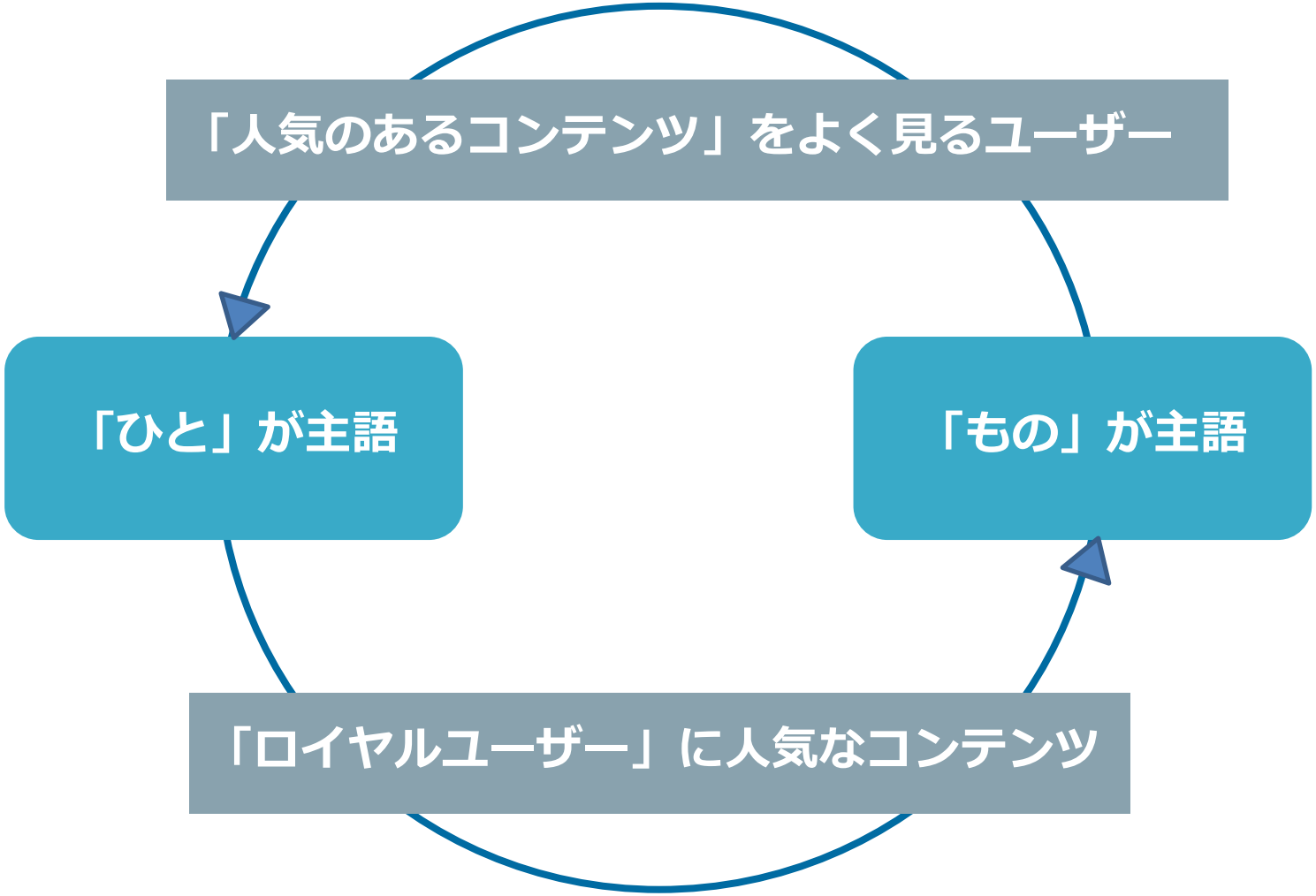
「もの」が主語

ものID	カテゴリ	メーカー	価格
1	家具	A社	3,000
2	雑貨	A社	1,200
3	雑貨	B社	1,000
4	家具	C社	4,000
5	雑貨	C社	1,500
6	家具	A社	4,500
7	家具	B社	2,900
8	家具	C社	3,300

©GRI Inc.

分析時にデータ活用の実務担当者が検討すべきこと

実際は「ひと」と「もの」を行ったり来たりすることで分析を進める



©GRI Inc.

自動機械学習（AutoML）とは

表形式の自動機械学習とは

一枚表を用意すれば、特徴量から目的変数を予測するパターンを自動で検出するシステム

目的変数　　特徴量

ID	Churn?	STATE	Int'l Plan	Day Calls	Day Charge
4153824657	False.	KS	No	110	45.07
4153717191	False.	OH	No	123	27.47
4153581921	False.	NJ	No	114	41.38
4083759999	False.	OH	Yes	71	50.9
4153306626	False.	OK	Yes	113	28.34
5103918027	False.	AL	Yes	98	37.98
5103559993	False.	MA	No	88	37.09
4153299001	False.	MO	Yes	79	26.69
4083354719	False.	LA	No	97	31.37
4153308173	False.	WV	Yes	84	43.96
4153296603	True.	IN	No	137	21.95
4153449403	False.	RI	No	127	31.91
4083631107	False.	IA	no	96	21.9
5103948006	False.	MT	no	88	26.67

・例えば先ほどの電話会社の例で言うと、解約しそうな顧客を事前に見出すためのパターンを検出

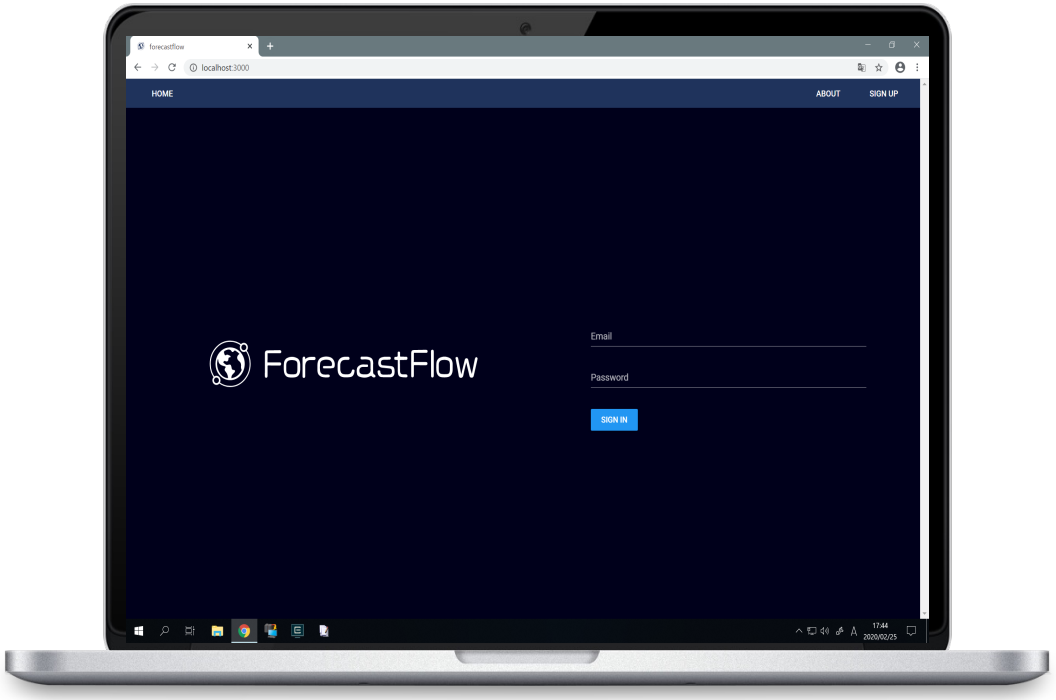
・特徴量（＝分析軸）だけから目的変数の予測を試みる

©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



サービスサイト： https://forecastflow.jp/
ログイン画面： https://forecastflow.org/

デモ1

デモ１ ～米国の電話会社での解約予測～

■ 米国の電話会社の例

–解約者数が急増した

–以下のような顧客データがあるので原因を分析したい

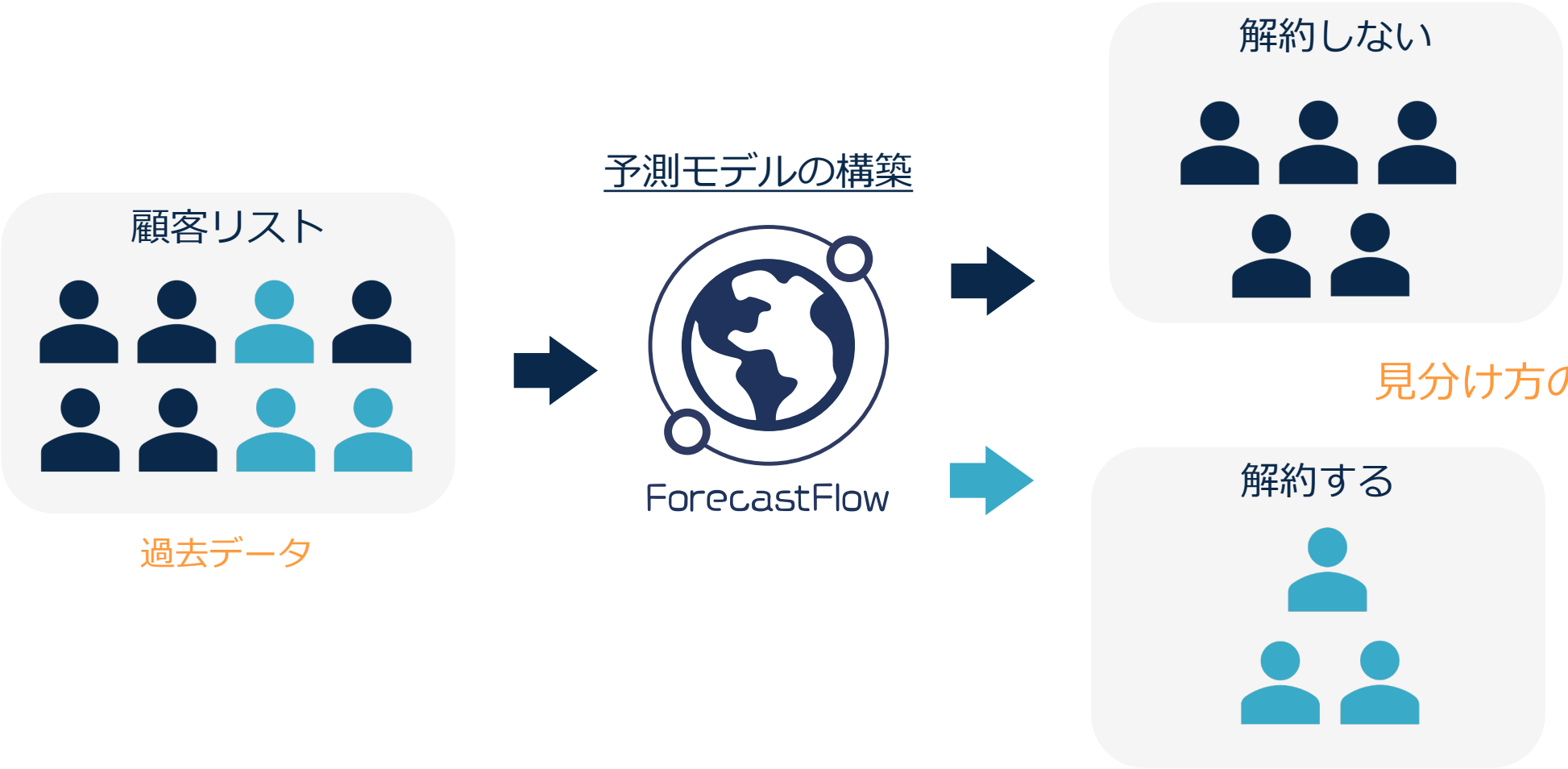
©GRI Inc.

施策出し分けと効果検証

訓練フェーズで行っていることとその解釈

インサイト発掘（訓練フェーズ）

過去データから復帰する顧客と復帰しない顧客を見分ける法則を学習
→実務担当者はこの法則を理解する



見分け方の理解をする

推論フェーズで行っていることとその解釈

未知の予測・推定
（推論フェーズ）

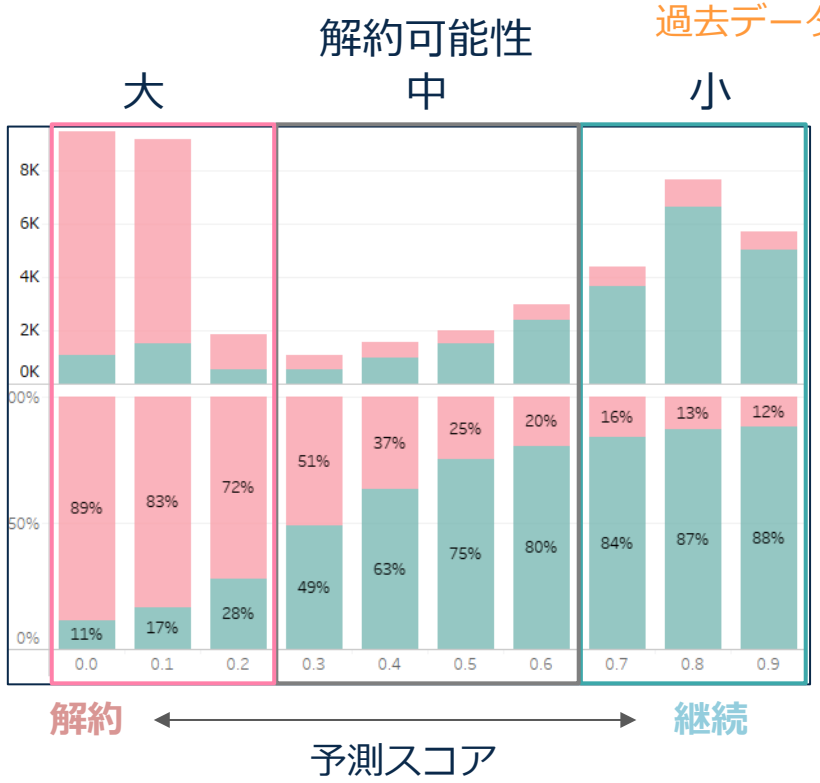
最新データを予測モデルに投入し顧客ごとの解約予測スコアを算出
→実務担当者は予測スコアに応じて施策対象を選定

©GRI Inc.

訓練フェーズと推論フェーズの比較

インサイト発掘（訓練フェーズ）

過去データ（正解がわかる）

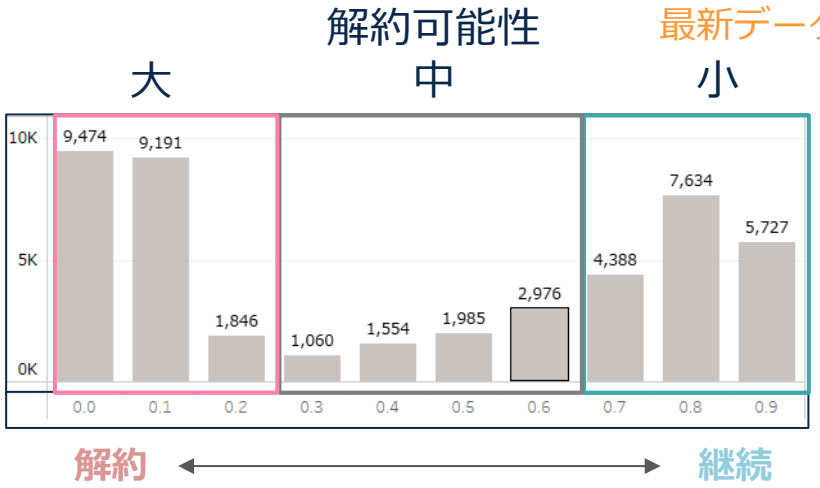


スコアごとのボリューム

解約者割合

未知の予測・推定（推論フェーズ）

最新データ（正解は不明）



施策A　施策B　施策C

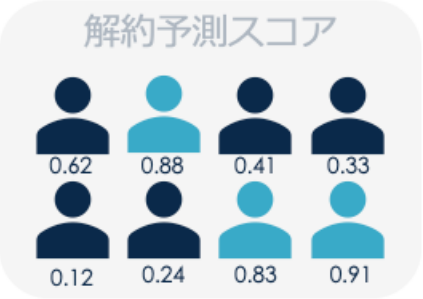
解約スコアに応じて施策や優先度を検討

©GRI Inc.

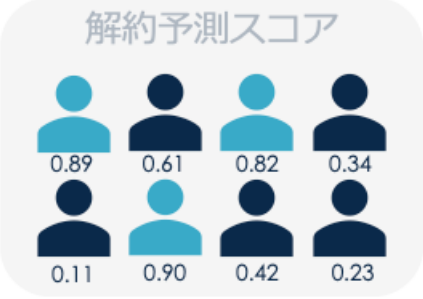
A/Bテストとの考え方の違い

A/Bテスト

A



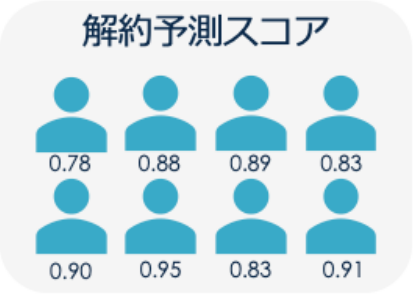
B



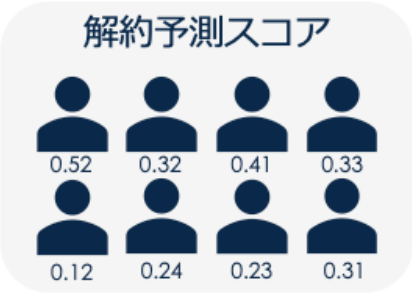
・AとBにはそれぞれ多種多様なユーザーが入る
・サンプルが少ないと分布が偏ることも
・層化を行うとしても軸の選定が難しい
・一律での施策は機会損失につながる

予測スコア

解約可能性大



解約可能性小



・AとBのサンプルが検証したい指標によって適切に層化される
・同一サンプル内でのA/Bテスト実施によってフェアな比較が可能
・ユーザー属性ごとに属性にあった施策を出し分けることができる

©GRI Inc.

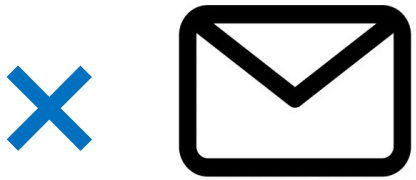
予測スコアを用いたWhat-if分析

さらにWhat-if分析によってパーソナライズされた施策検討も可能に

解約スコア：**0.95**
どの施策を打つ？

解約スコアが最も低くなる「クーポン配布」を選択

その他特徴量はそのまま予測スコアを計算

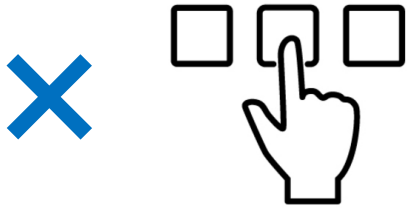


現実
if

ID	メルマガ	性別	購買金額	・・・	解約予測スコア
1	未受信	男性	3,000		**0.95**
1	受信	男性	3,000		**0.90**

現実
if

ID	クーポン	性別	購買金額	・・・	解約予測スコア
1	未配布	男性	3,000		**0.95**
1	配布	男性	3,000		**0.28**



現実
if

ID	オプション	性別	購買金額	・・・	解約予測スコア
1	未契約	男性	3,000		**0.95**
1	契約	男性	3,000		**0.65**

©GRI Inc.

機械学習活用時の悩みと考え方

機械学習活用時のお悩みポイントと考え方

■ そもそも何ができるかわからない（問題設定ができない）

- 経験上以下のようなケースで有用
- （１）質の良い正解ラベルが付与可能かつある程度均衡（2値分類の場合10%程度まで）
- （２）手元のデータと正解ラベルの間に何らかの関係がある（≠相関、線形関係）
- （３）データボリュームがある程度確保可能（数千程度～）

■ どのようなデータセットを準備すればいいかわからない

- フレームワークは前述した「ID・推定対象・特徴量」
- まずは詳細に入り込みすぎず最小限のデータセットから始めるのがベター

■ 必要なデータがない/あっても（技術的・計算リソース的に）加工ができない

- まずはデータ基盤の整備から始める（過去のすいすい会参照：11回,22回,27回など）

■ 精度が上がらない

- 特徴量の追加、問題設定の見直し（特に対象母集団）、データの追加などを順次検討

©GRI Inc.

まとめと今後

©GRI Inc.

まとめ

- データ活用の実務担当者こそが自動機械学習の利用を推進できる
- 自動機械学習を利用すると複雑な組み合わせの効果を勘案できる
- 予測スコアを利用して施策立案に役立たせることができる
- ただし、データを正しく準備する必要がある

©GRI Inc.

ForecastFlowの無料トライアル

- 90日間、無料でForecastFlowを試すことができます

- 下記より、お申込み
- https://forecastflow.jp/

©GRI Inc.

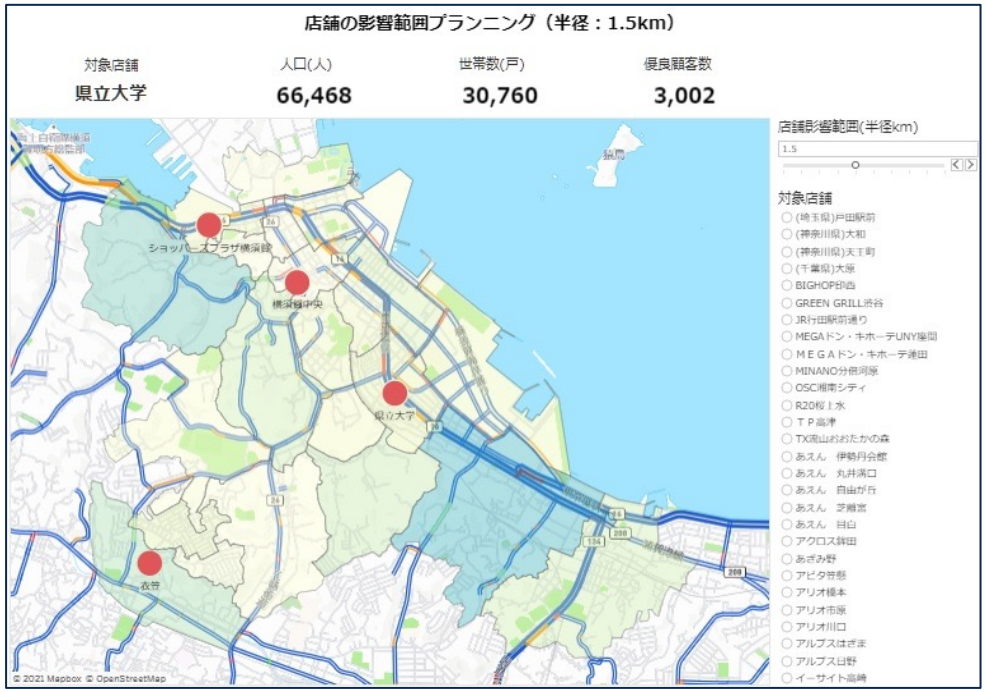
次回のすいすい会

■2022年8月24日（水）18:30-19:30

■ロケーションインテリジェンスを加速化させるLLocoのご紹介

- 地理情報データプラットフォーム

LLoco



残っている質問（今後、説明していきたい内容）

- 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実行タイミングの関係性
- ダウンサンプリングとProbability Calibration
- ForecastFlowのアルゴリズムを知りたい
- 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.