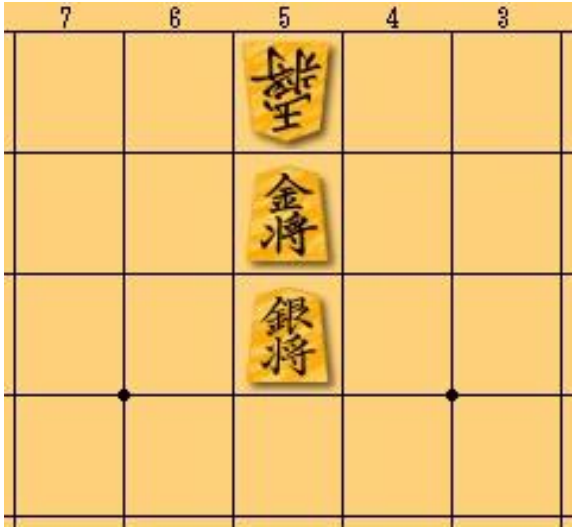
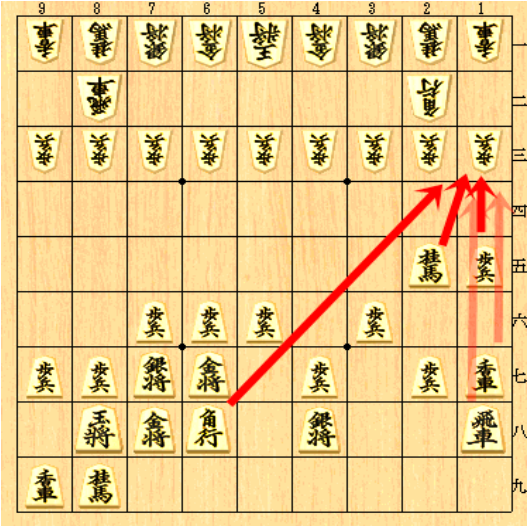
# 予兆を捉える機械学習モデル

# 予兆を捉える機械学習モデル



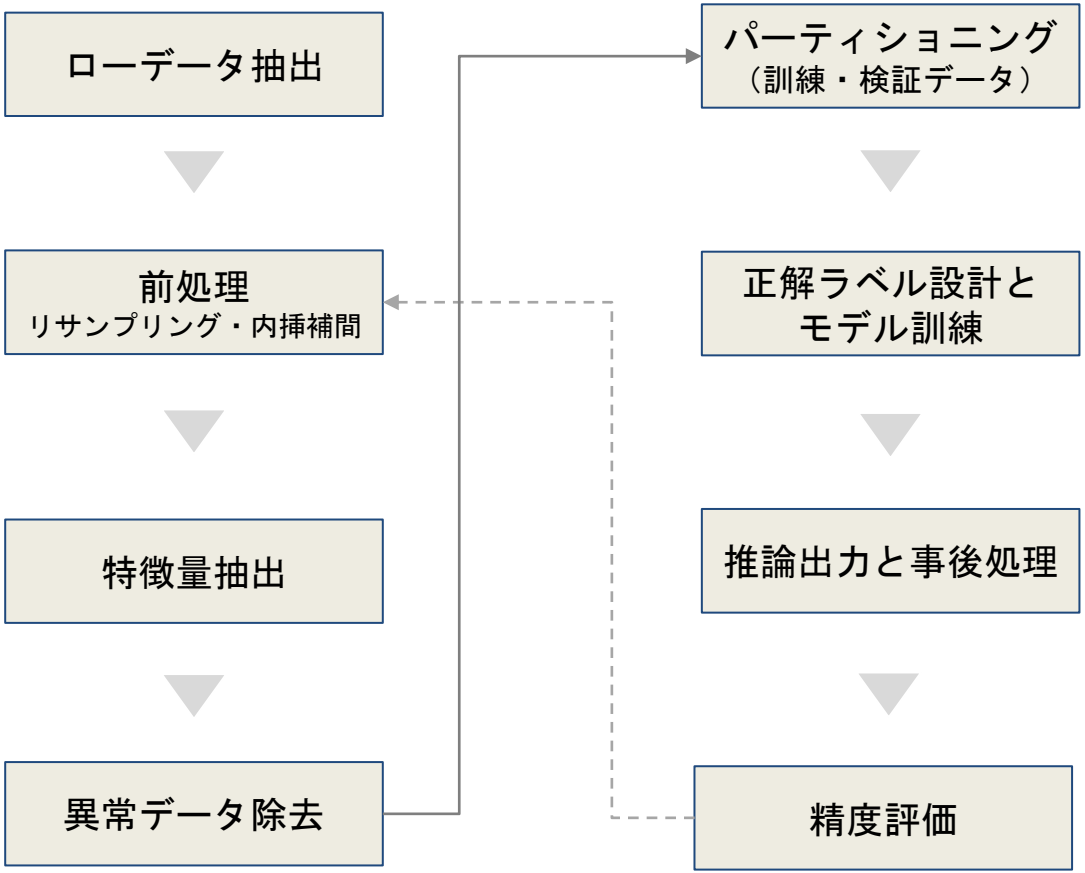

詰み手の状態を検知するモデル
（自動化による価値創出）




詰みそうな予兆を捉えるモデル
（事前対策による価値創出）


©GRI Inc.

# 典型的なパイプライン





©GRI Inc.

# プロジェクトとして見極めが難しいところ

- 分析人材やエンジニアリソースの割り振りの見極めは困難
  - IoTセンサーは時間経過にともなってデータをどんどん生成する
  - データが増えることで機械学習モデルの予測精度は改善される可能性がある
  - 一方で、問題設定や特徴量、モデル構築を改善することでも予測精度が改善される可能性もある
  - データの蓄積を待ちますか？それとも分析人材を積極的にアサインしますか？

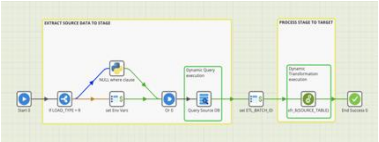

©GRI Inc.

# 現実的なソリューション

- 機械学習モデルは一度作ったら終わりではなく、継続的な訓練や監視システムがどうしても必要になる
  - 機械学習システムの初回導入は非常にハードルが高いので、初回は分析会社やシステムベンダーに依頼するのは全然ありだと思う
  - ただし、自分たちでお世話や改善ができる機械学習システムが理想

Google Big Query

snowflake



MATILLION

ForecastFlow

tableau


©GRI Inc.

# 予兆を捉える機械学習モデル構築のポイント

# 予兆を捉える機械学習モデルで気をつけたいところ

- **前処理と特徴量設計**
  - 少し先の未来に感度をもちそうな特徴量の設計
- **正解ラベル設計**
  - 検知したいイベント時刻を中心に重み付け
- **機械学習モデルの訓練/検証データのパーティショニング**
  - 「リーク」に気をつける


©GRI Inc.

# 予兆を捉える機械学習モデルで気をつけたいところ

- **前処理と特徴量設計**
  - 少し先の未来に感度をもちそうな特徴量の設計

- **正解ラベル設計**
  - 検知したいイベント時刻を中心に重み付け

- **機械学習モデルの訓練/検証データのパーティショニング**
  - 「リーク」に気をつける


©GRI Inc.

# 前処理と特徴量設計

# 特徴量抽出のステップとポイント

ローデータ抽出

時系列データの特徴量は、
時間ドメインの表現と周波数ドメインの表現がある

前処理
リサンプリング・内挿補間

周波数ドメインの特徴量を抽出するためには、
各データポイントが一定間隔で入ることが望ましいので、
ローデータの前処理としてリサンプリングと内挿補間が必要

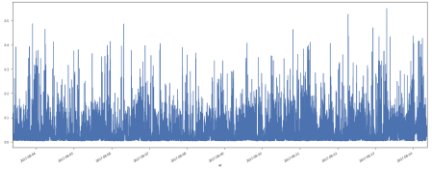
特徴量抽出

リサンプリングする際にはナイキスト周波数に注意

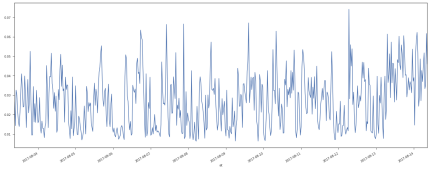

©GRI Inc.

# 前処理

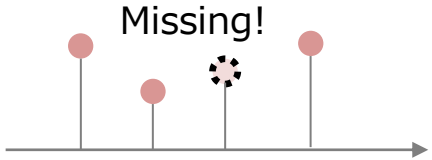
ローデータ



リサンプリング



線形補間 / 内挿
（Interpolation）



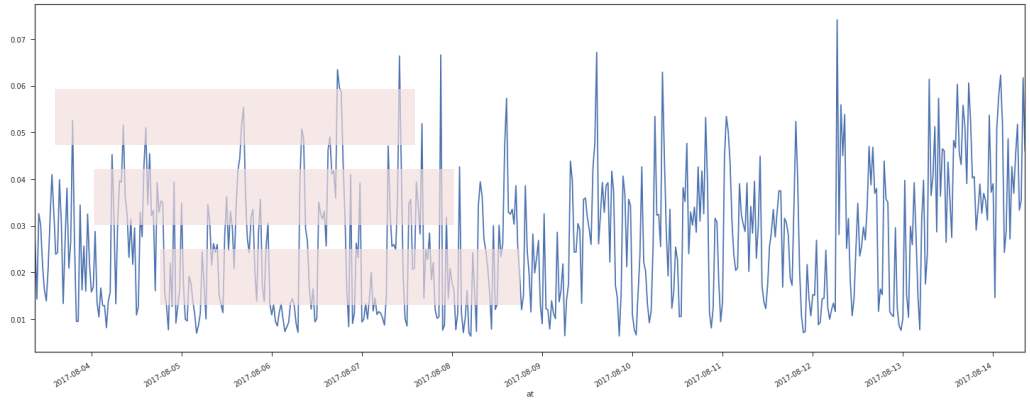

- ✓ ノイズ低減や長期的な構造を捉えるため、1分/10分/1時間などの適当な間隔内で平均値をとる
- ✓ 周波数ドメイン特徴量を計算するために欠損補間してデータ間の時差を等間隔に並べる


©GRI Inc.

# 時系列特徴量の２つの「ドメイン」

- 時間ドメイン特徴量
  - ある期間内の時系列データの平均値や分散などの特徴量

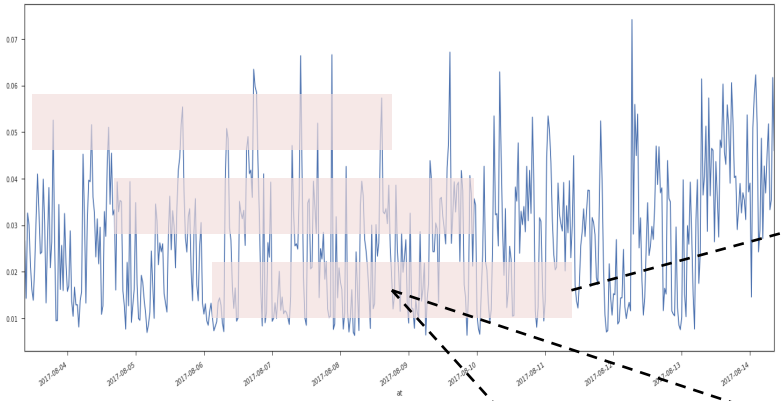
- 周波数ドメイン特徴量
  - フーリエ変換やウェーブレット変換を行うことで得られる特徴量




**いずれの特徴量も、**
**前処理されたデータの**
**スライディングウインドウから**
**計算されることは共通**


©GRI Inc.

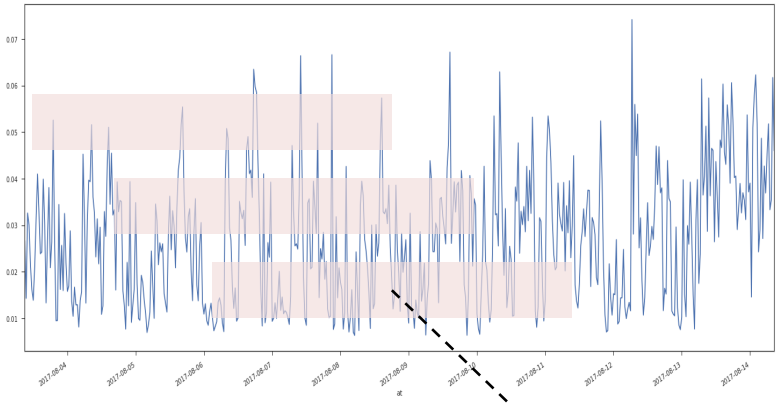
# 時間ドメインの特徴量の例

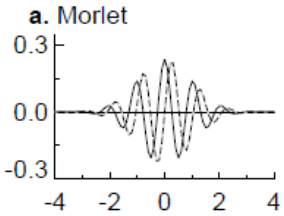



| | 意味 | | 特徴量 |
|---|---|---|---|
| ベーシック | ウインドウの先頭データ | | 観測値 (latest) |
| | | | 時間 (hour) |
| | | | ひとつ前の値からの変化量 (ch-percent) |
| | ウインドウ全体 | 値の大きさの特徴 | 平均値 (average) |
| | | | 最大値 (max) |
| | | | 最小値 (min) |
| | | 変化の傾向 | 線形トレンドの傾き (slope) |
| | | 変動の激しさ | 標準偏差 (std) |
| | | 前信号との関連性 | 自己相関係数 (autocorr) |
| | | 他信号との関連性 | 交差相関係数 (crosscorr) |
| アドバンスド | シフト差分 | 各特徴量の時間変化量 | 前2日と後2日の特徴量の差分(shiftdiff) |
| | 変化点・異常検知 | | 変化点数・間隔 |
| | | | 異常度スコア |
| | 時系列予測モデル | | 予測値からの誤差 |

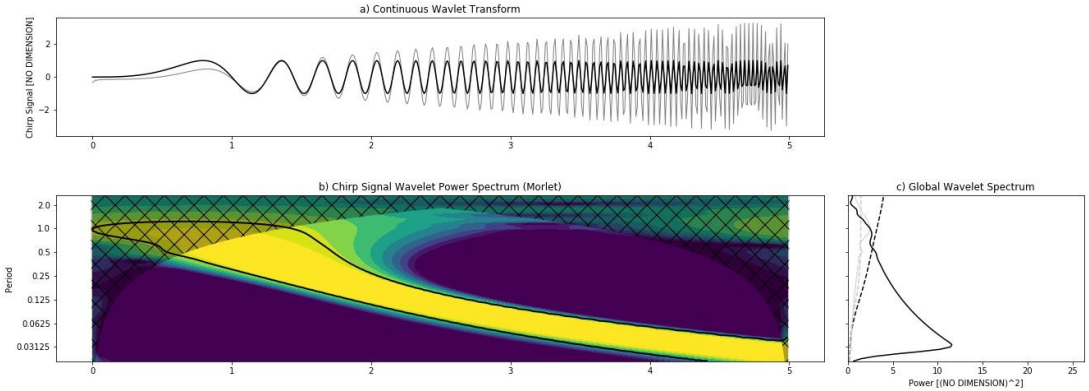

©GRI Inc.

# 周波数ドメイン特徴量






ウェーブレット変換を行い、
各周波数帯のスペクトルパワー
から統計量を算出

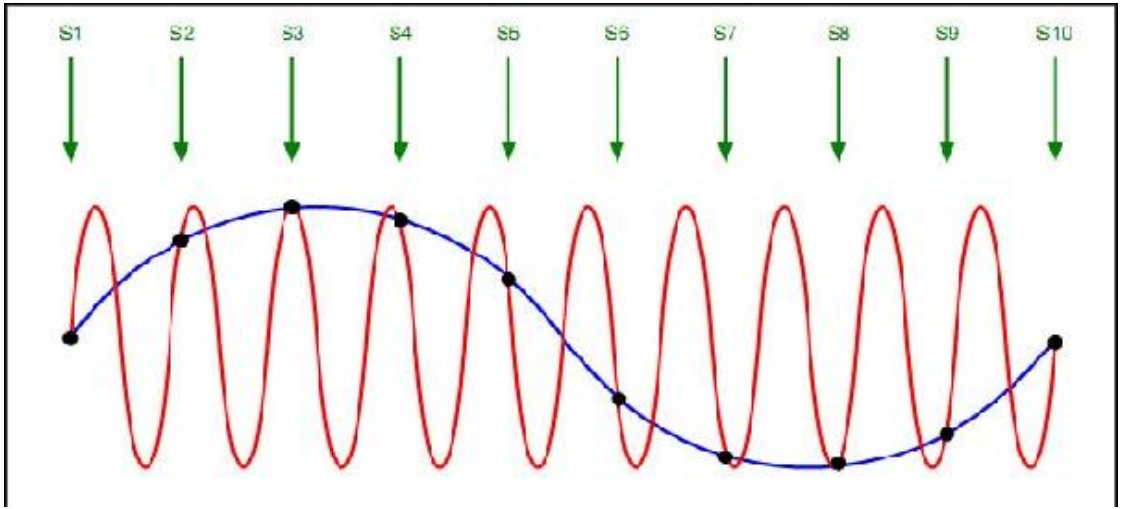




©GRI Inc.

# ナイキスト周波数

- リサンプリング前処理はノイズ低減などに有効だが、リサンプル時間が捉えたい予兆やイベントの時間スケールよりも十分短くとる必要がある




**[参考文献]**
**What happens to sounds above the Nyquist frequency in digital audio? Do they get aliased into some type of noise?**


©GRI Inc.

# 予兆を捉える機械学習モデルで気をつけたいところ

- 前処理と特徴量設計
  - 少し先の未来に感度をもちそうな特徴量の設計
- 正解ラベル設計
  - 検知したいイベント時刻を中心に重み付け
- 機械学習モデルの訓練/検証データのパーティショニング
  - 「リーク」に気をつける


©GRI Inc.

# 正解ラベル設計

# 正解ラベルを定義する上でのポイント

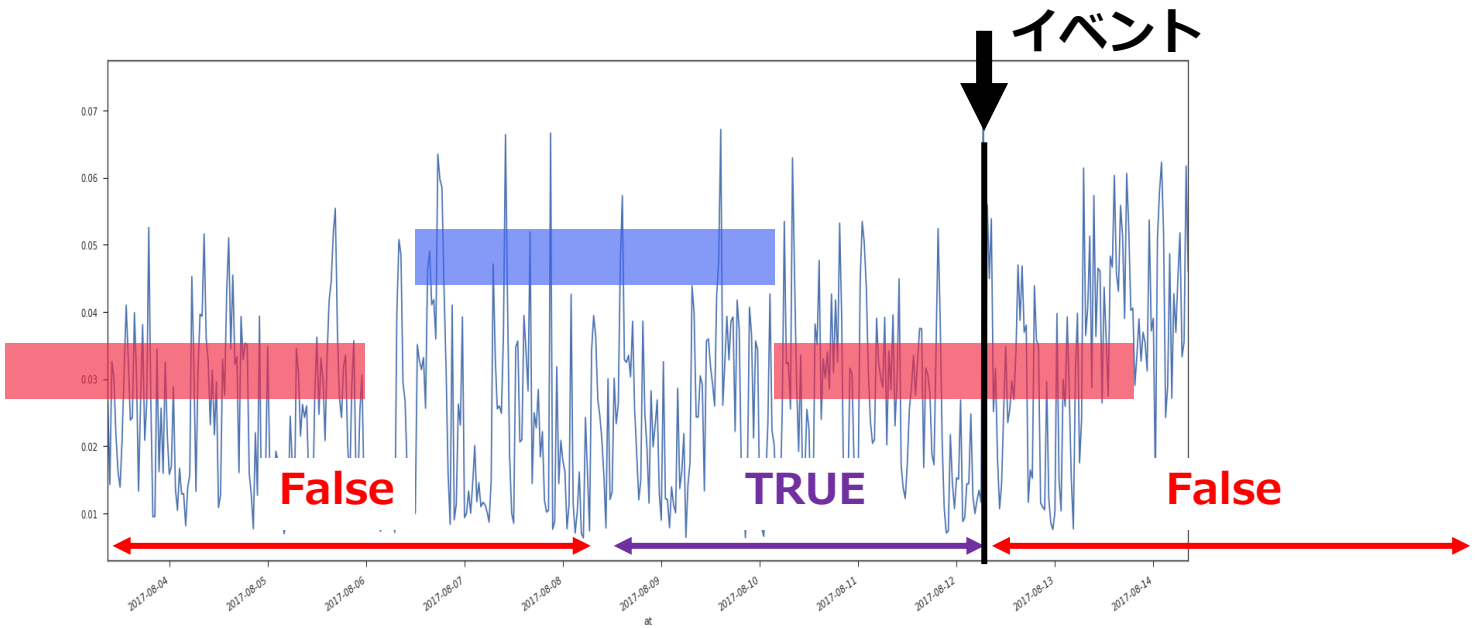
- イベントよりもどのくらい前から予兆は発生しそうか、あるいは、ビジネス要請として何時間前に捉えたいかによって定義する

- 正解ラベルの付け方を工夫することで、予測精度や運用でメリットがあるケースも


©GRI Inc.

# 特徴量と正解ラベルの関係性

- 特徴量算出ウインドウの先頭時点とイベント発生時点との関係性から正解ラベルを定義
- 例えばイベント時点からN時点前までをTrueとすると、以下のイメージ

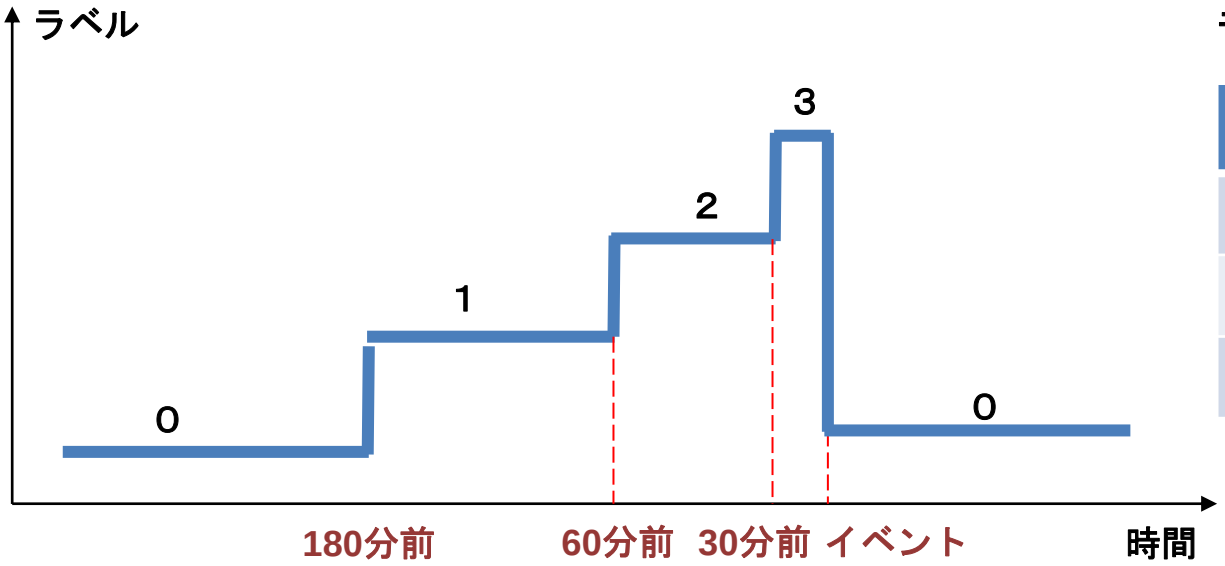




©GRI Inc.

# 正解ラベルをマルチラベルにする場合

- イベント発生時刻からの時差に応じてマルチクラスとして正解ラベルをアサイン
- モデル推論結果は各クラスのスコアになるため、各クラス値でスコアの重み付けなどが可能に
- 各クラスにて予兆の特徴が異なる場合にも有効




**モデル推論結果**

| TIME | クラス 0 | クラス 1 | クラス 2 | クラス 3 |
|---|---|---|---|---|
| 20:00 | 0.2 | 0.34 | 0.26 | 0.2 |
| 20:30 | 0.1 | 0.5 | 0.2 | 0.2 |
| 21:00 | 0.8 | 0.1 | 0 | 0.1 |


©GRI Inc.

# 予兆を捉える機械学習モデルで気をつけたいところ

- 前処理と特徴量設計
  - 少し先の未来に感度をもちそうな特徴量の設計
- 正解ラベル設計
  - 検知したいイベント時刻を中心に重み付け
- 機械学習モデルの訓練/検証データのパーティショニング
  - 「リーク」に気をつける


©GRI Inc.

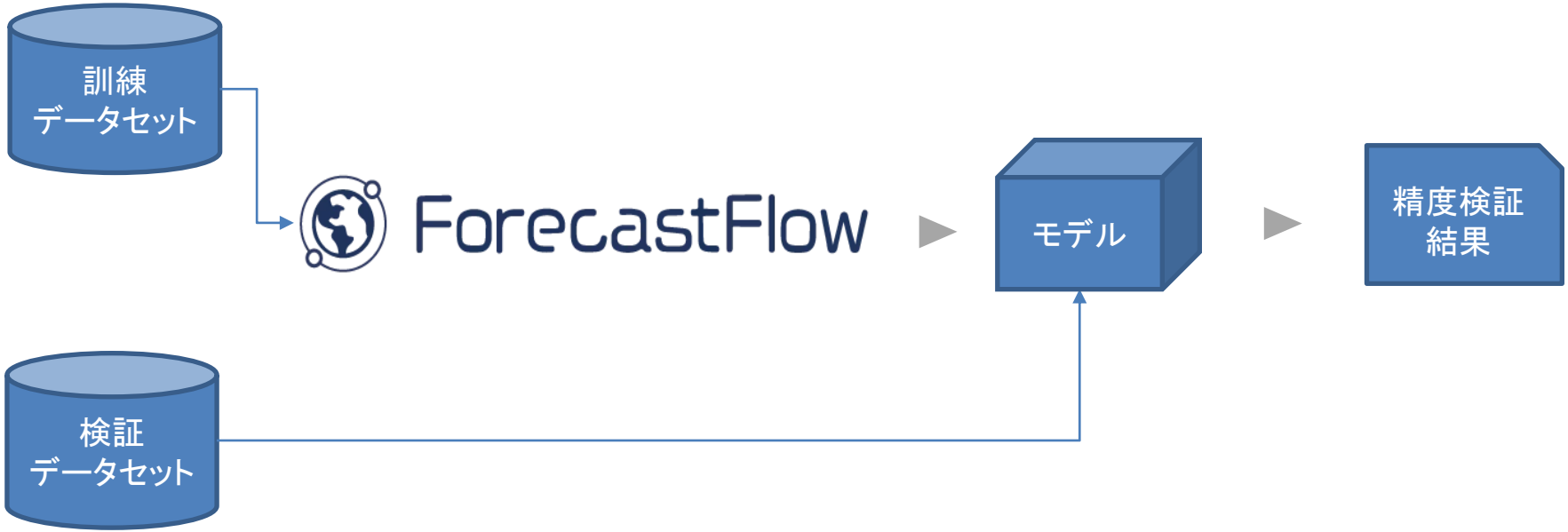
# 訓練データと検証データのパーティショニング

# モデルの訓練で気をつけたいポイント

- モデルの訓練フェーズに未来の情報を含めないように訓練/精度検証データセットのパーティショニングを行う
  - これを怠ると、いわゆる「リーク」の原因になり、実際より高い精度に見えてしまう場合がある


©GRI Inc.

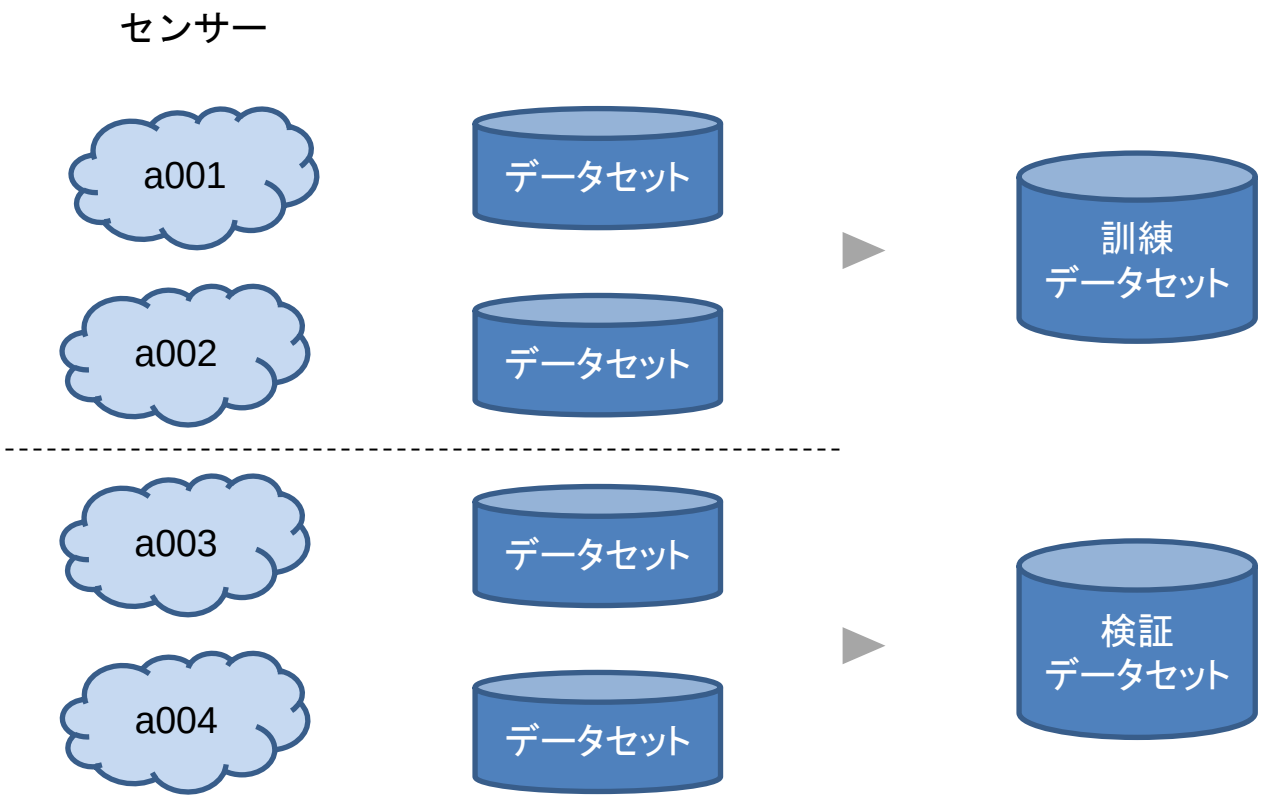
# 機械学習システムの入出力





©GRI Inc.

# 安全なパーティショニング

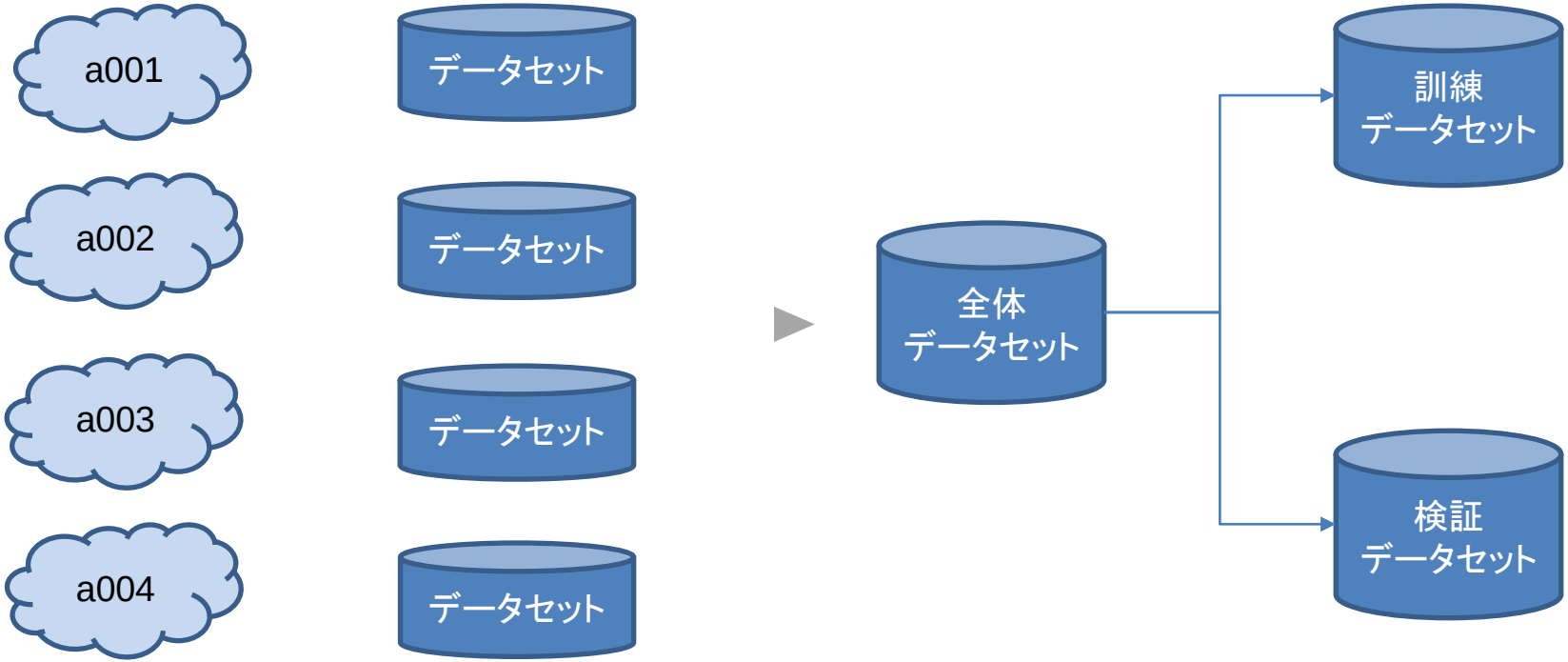




©GRI Inc.

# リーク発生の危険性があるパーティショニング

全体データセットをランダムに振り分けると、
同じセンサーの未来情報が訓練データセット
に含まれる可能性がある





©GRI Inc.

# ForecastFlowにおける精度検証データセットの設定方法

- モデル構築画面の詳細設定に、精度検証用データセットの準備方法から、ランダム分割ではなく、他のデータセットも選択可能

©GRI Inc.

# まとめ

# 今回の内容

- 予兆を捉える機械学習モデルの重要ポイントをいくつかピックアップしてご紹介

- 予測タスクとしては難しい部類に入りますが、予兆を捉えることは未来の情報を事前に知れることにつながるので、情報価値として非常に高く、チャレンジする価値もあると思います

- 今回はご紹介しきれなかった内容もまだ沢山ありますので、ご興味ある方は是非ご連絡ください


©GRI Inc.