

# db tech showcase ONLINE 2020

27<sup>TUE</sup> OCTOBER - 10<sup>THU</sup> DECEMBER



## ようこそ！

## DB tech showcase ONLINE 2020へ

# AI/ML導入そして運用の成否を 分けるポイントとは ~教師なし異常検知の例を中心に~

株式会社インサイトテクノロジー  
プロダクト開発本部  
小浦方 寛太



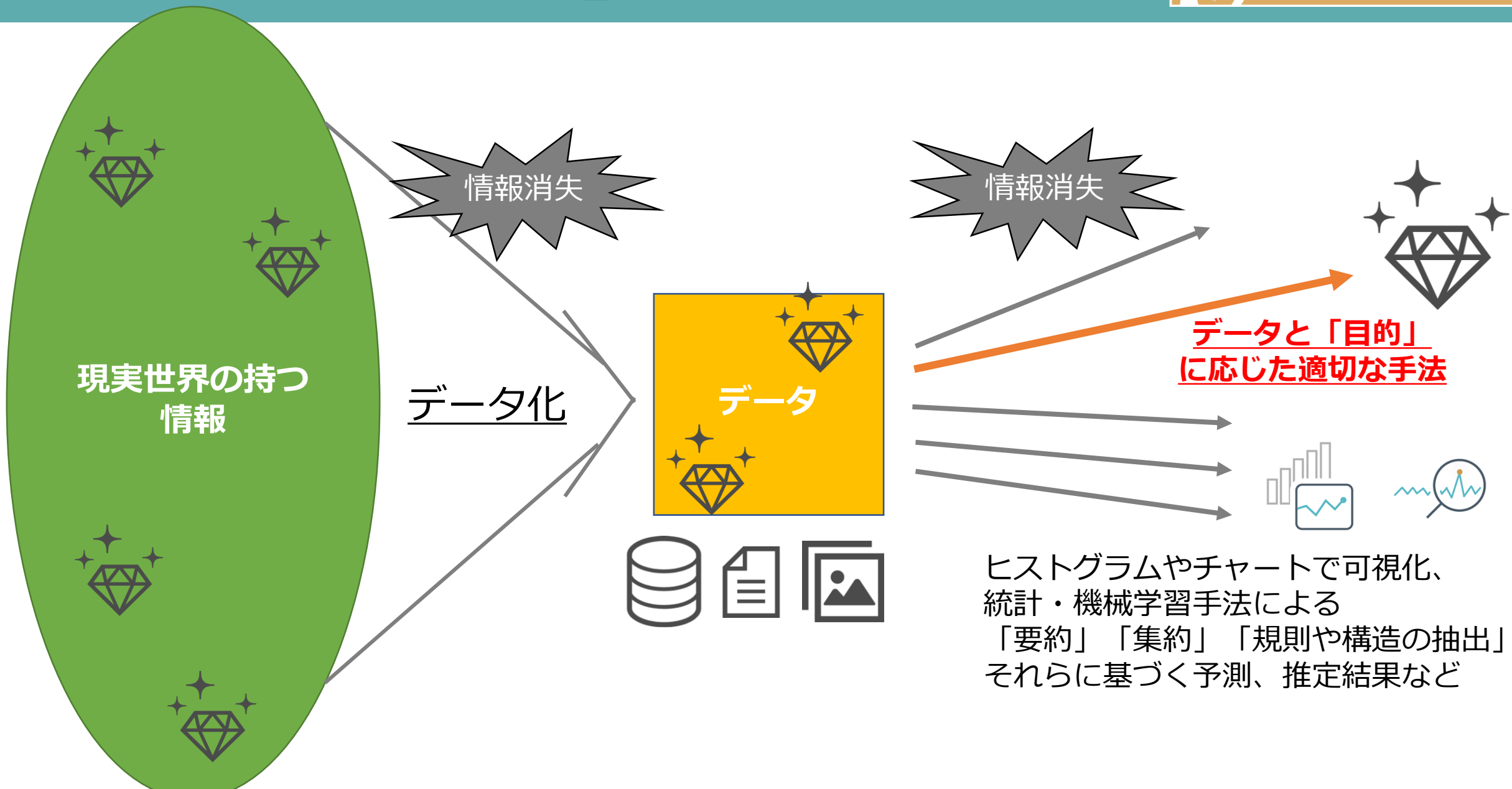
- 小浦方 寛太（こうらかた かんた）
- 株式会社インサイトテクノロジー プロダクト開発本部 所属
- 統計・機械学習を用いたシステムやプロダクトの開発に従事
- (教師なし)異常検知の案件を受け持つことが多い
- 好きな実問題解決へのアプローチはベイズ学習、モデリング

データ分析屋・機械学習エンジニアの立場から、  
実問題を解決する機械学習プロジェクトにこれまで携わって  
きた経験を基に、以下のことについて話します

- 機械学習プロジェクトを成功させるために（と言うより、**避けられる失敗を避ける**ために）気を付けるべきこと
- 私がデータ分析屋・機械学習エンジニアとして、**機械学習のプロジェクトに取り組むときに日頃考えていること**

- データから「価値」を得る流れ
- 機械学習とは
  - 機械学習に向いてること
  - 機械学習に(も)できないこと
- 事例: 膨大な機器からのサイレント故障検知
- 機械学習を含めたシステムを成功させるには
- 機械学習プロジェクトの失敗と学び
- 教師なし異常検知の困難
- 事例: 膨大な機器からのサイレント故障検知(再)
- 事例: 大量のセンサーデータからの未知の障害検知
- まとめ

# データから「価値」を得る流れ



# データから得る「価値」の例

- 飲食店の過去の来客数や天気予報のデータなどから明日の来客数が高い精度で予測できた結果、廃棄ロスを減らすことに成功
- 設備機器のセンサーデータから故障の予兆を検知できた結果、予知保全が可能に
- 過去のクレジットカードの取引履歴をもとにクレジットカードの不正使用を検知
- 手書き数字のデータを学習することで、封筒の手書きの郵便番号を画像認識で読み取り地域毎の仕分けを自動化
- . . .



# データから「価値」を得る流れ



- 現実世界からの「データ化」、データからの情報操作・抽出、それぞれで**情報が失われる**
- 現実世界はデータ化したとしても、人間の情報処理能力では扱えないほど膨大で複雑であることがしばしばであり、**人間が活用するには整理や要約をする必要**がある
- データから得た情報に**「価値」**があるか（役に立つか）は、その情報を**何に・どう活用したいのかという「目的」に依る**
- **データの性質と「目的」に応じて適切な統計・機械学習手法を用いる**ことで、はじめてデータから**価値**ある情報を得ることができる



# データから「価値」を得る流れ(補足)



- 先の図の流れを実際の現場で実現するためには、機械学習による**データ活用の基盤を構築し、運用監視・保守**をする必要がある
- 実際の現場で価値を提供できるか否かとしては、使いたい**データ処理に必要な計算コスト**（マシンリソース・計算時間）が「割に合うか」といった観点も重要

- 「統計学との違いは何か？」という問いがよくあるが、個人的にはあまり区別する必要を感じていない
- 統計学は目的としてデータの「説明」を重視する傾向があるのに対し、機械学習の方は「**予測**」重視の傾向がある
- **機械学習とは、データから規則や構造を抽出することにより、未知の現象に対する予測やそれに基づく判断を行うための計算技術の総称である**  
(※引用元： 須山 敦志. バイズ推論による機械学習入門. 講談社, 2017.)

- データが人間の手で処理できないほど「**大量**」である
- データ間の関係が人の手に余るほど「**複雑**」である
- データを「**高速**」に判断して処理する必要がある
- データを「**安定**」して処理し続ける必要がある

**人間では(容易に)できないことを実現してこそ**

## 「機械」学習と言えど、 人間の主観を完全に排除することはできない

- 未知の状況に対しても予測をする必要があるが、何かしらの仮定を導入しない限りは「解」が定まらない（**不良設定問題**）
- 万能なアルゴリズムは原理上存在しない（**ノーフリーランチ定理**）
- 客観的な分類はありえない（**醜いアヒルの子の定理**）

解きたい問題やデータに関する情報（**ドメイン知識**）を用いて、人間が適切なモデルやアルゴリズムを設計・選択する必要がある。

# 機械学習に(も)できないこと(補足)



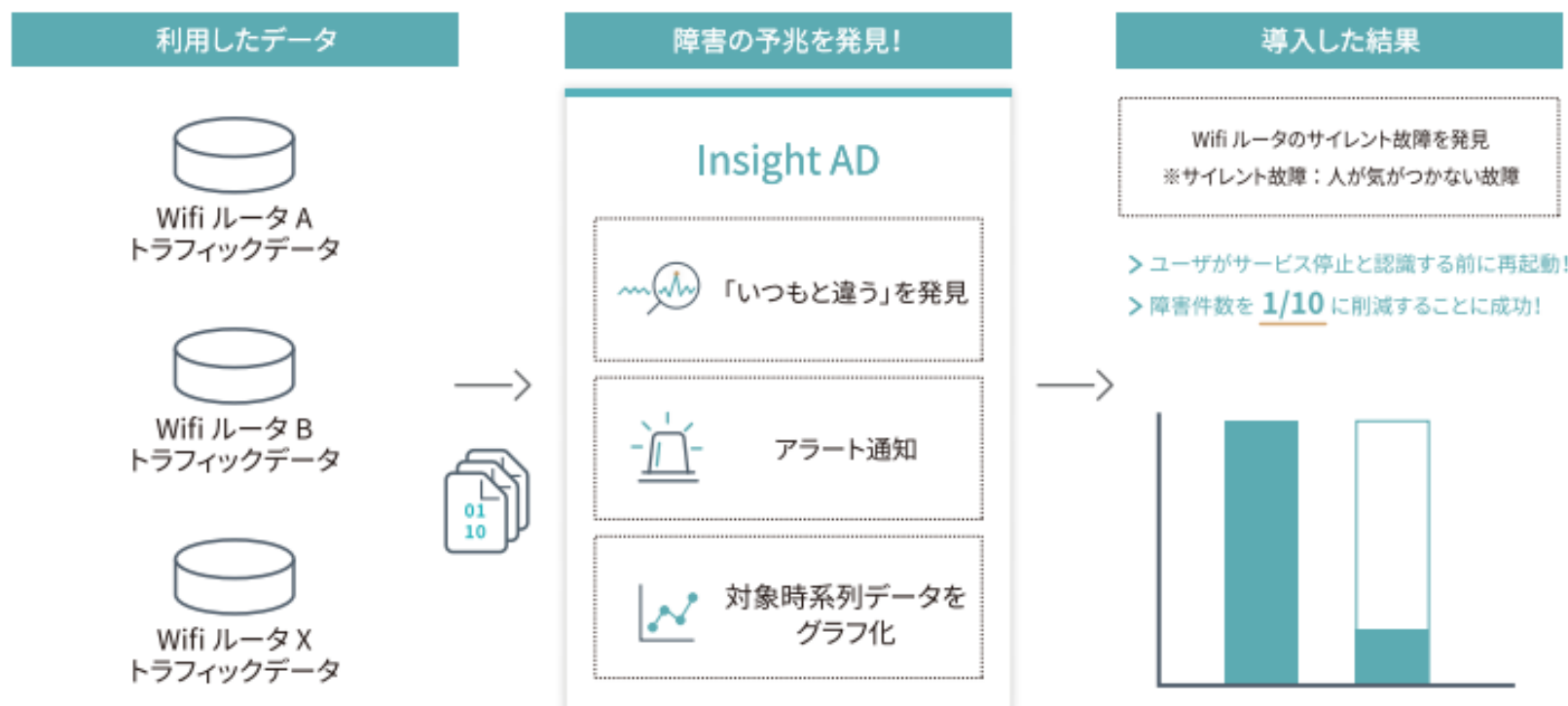
- 「機械」学習と言えど、使うためには**データや問題に関する専門的な知識（ドメイン知識）は必要**
- ただし、それを IF-THEN のようにルールを直接書き下すのではなく、**「データの前処理や整備」「データから適切に情報を抜き出す方法の選定」「機械学習の結果の解釈」などに用いる**

# データ分析屋として思うこと

- 先のページで機械学習に不可能なこととして述べたことは、形式的に証明される不可能性であり、つまり**人間でも不可能**
- しかしそれらは「機械学習やAIの限界」を示すものとして語られがちであり、人間の心理として、**統計・機械学習・AIに対して「客観性」や「真実」といったものを期待してしまいがちなのだと思う**
- 「とにかくデータをたくさん学習させればよい」は基本的に誤り。  
**データに関する情報、ドメイン知識を基に適切な手法を用いることが重要**
- 機械学習は、己の意図を持った上で使いこなすべき「道具」である

# 事例: 膨大な機器からのサイレント故障検知

- 日本各地の膨大な数のネットワーク機器の故障検知
- pingの返答はあっても実際には障害が発生している「サイレント故障」
- 今まで故障検知をしてきた方と協議を繰り返し、**得られたドメイン知識を異常検知モデルに反映**できたことが成功につながった



## 機械学習を含めたプロダクトをビジネスとして成功させるために必要な4者

1. プロダクトに関するドメイン知識を持った人
2. 統計や機械学習に明るい人
3. データ分析基盤を作れるエンジニアリング能力のある人
4. 失敗しても構わないとリスクを取ってくれる責任者

(※引用元：有賀 康顕, 中山 心太, 西林 孝. 仕事ではじめる機械学習. オライリー・ジャパン, 2017.)

**上記4者の役割を果たす人がいて（1人が複数の役割を果たすこともある）、**

**かつ互いに連携が取れていると、**

**プロジェクトの成功確率がグンと上がる**



## 課題・失敗

### 目的や課題が曖昧なままとりあえずAI/機械学習

- 「このデータ使ってAI/機械学習で何かできない？」
- 区切りのないままPoCを始めてグダグダに...

## 解決策・学び

- ✓ 目的・課題・PoCのゴールを明確にする
- ✓ ...とは言うものの、目的・課題を最初から適切に設定するのは難しい。第0ステップとして、**今あるデータを可視化して、何ができるかをドメイン知識を持つ人とデータ分析をする人とで話し合う**
- ✓ 課題を明確にした結果、**機械学習じゃなくていいよね**ということも結構ある

## 課題・失敗

### データおよびドメイン知識をいただくことにハードルがある

- お客様の方で、結果が出るか分からないPoCにいきなりデータを渡すことに抵抗がある
- セキュリティ等の事情で、外部の分析者に提供できる情報に制限がある

## 解決策・学び

- ✓ お客様で簡易的にでも検証できる仕組みをご用意する
- ✓ クラウドでモデルの検証用APIを公開する
- ✓ データを外に出せないお客様向けの簡易検証プログラム提供する
- ✓ お持ちのデータで機械学習によって何ができるか、イメージしていただく第一歩に

## 課題・失敗

### PoCで精度がよかったモデルが本番環境では精度がイマイチ

- PoC時に交差検証して過学習していないことは確かめたはずなのに…
- 実は、本番時に適用される環境には個別に差異があった

## 解決策・学び

- ✓ 本番時の環境にできる限り近いデータでテストしておく
- ✓ PoC時のモデルの前提条件（仮定）を明確にしておく（**どういう場合にうまくいかなくなるかを認識しておく**）
- ✓ 適用する環境で「共通の性質」を表現したモデルを用いつつ、「個別の性質」を学習できるようなアルゴリズムを設計する

## 課題・失敗

PoCで「結果」が出ても

**お客様に本番環境で使っていただければ意味がない**

- お客様が実際のシステム運用のイメージを持つことができず、PoCや導入段階で頓挫してしまう
- 機械学習モデルの精度をいくら上げて、ビジネス的な価値に結び付かないケースもある
- 複雑なモデルになるほど、その可視化や解釈が難しくなる

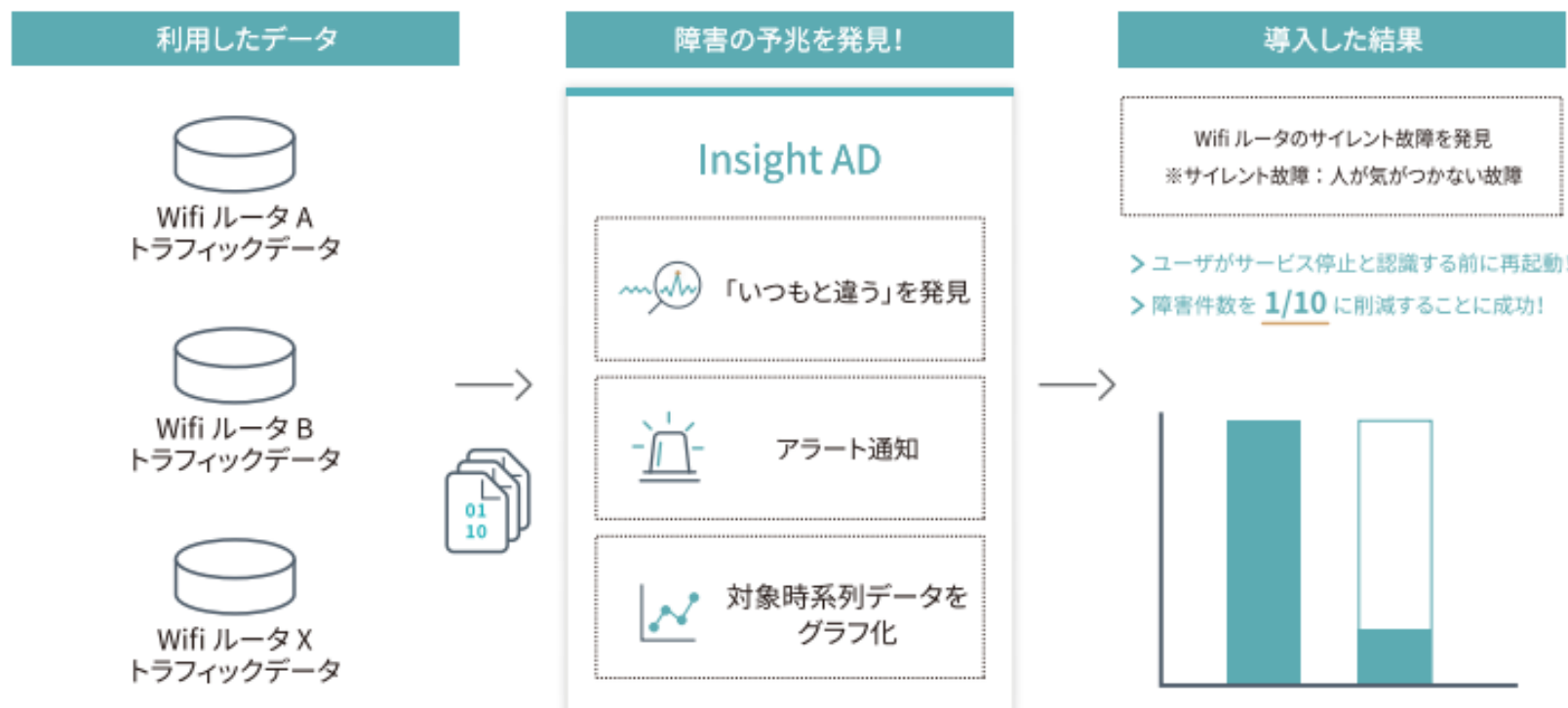
## 解決策・学び

- ✓ ビジネス的な効果を測るKPIを早い段階で合意する
- ✓ お客様の課題を詳細にインタビューした上で、実現すべき優先順位を明確にする
- ✓ **機械学習の結果を具体的にどう活用するか、本番運用イメージをお客様と共有する**
- ✓ 例えば説明や解釈のし易いモデルを用いて、お客様が「使える」かたちで提供する

- 異常検知の案件は、異常時のデータが僅かでもあればいい方で、**基本的にはまず異常の教師データはない**
- 教師なしでも使える手法を用いれば、何かしらの異常検知のアルゴリズムは作れるが、**それが本当に使えるのかの評価が問題**
- 基本的には、正常時のデータを加工して**人工的に「異常データ」を作り、それを検知できるかで評価**する
- 検証のために作る「異常データ」は、ドメイン知識を持つ人としてしっかり協議して作成する

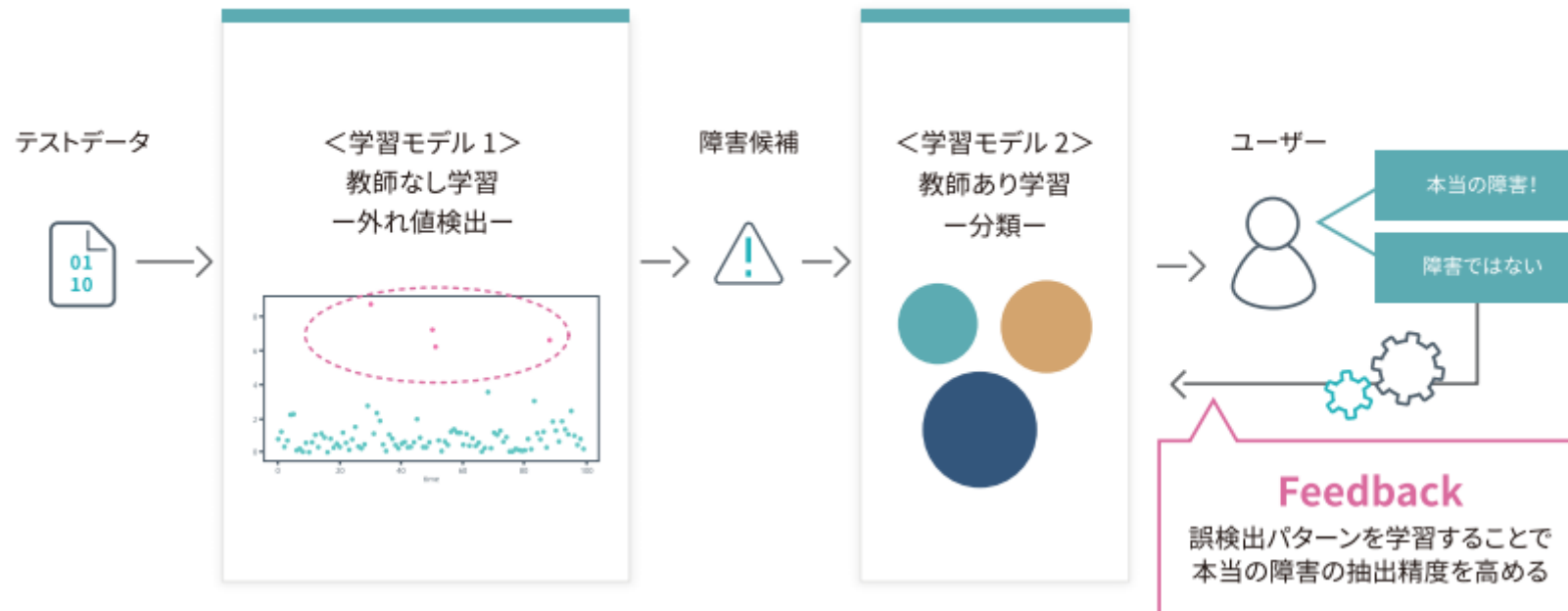
# 事例: 膨大な機器からのサイレント故障検知(再)

- 日本各地の膨大な数のネットワーク機器の故障検知
- 検証時に、**本番の運用を具体的に想定したシミュレーション**（一日に発砲されるアラート数を見積もり、それが現実に対処できるか、その具体的なフローまで確認）を行ったため、スムーズに導入ができた
- **時間経過によるデータの性質の変化も考慮**し、適応的に学習できる方法を採用することで、長期的に安定して精度を維持



# 事例: 大量のセンサーデータからの未知の障害検知

- 数十～数百のセンサーが取り付けられた機器から、センサー間の関連を考慮して異常検知（今までは、ルールベースで既知の障害パターンしか検知できず）
- **ドメイン知識を事前に十分に得ることが困難なケースであったため、教師なしモデルで一旦検出した異常候補をユーザにアノテーションしてもらったことで、その不足を補うようなシステムを構築**



- 「とにかくデータをたくさん学習させればよい」は基本的に誤り。それよりも**ドメイン知識の方がはるかに重要**
- **現実の綺麗でないデータと単純でないビジネス課題を、統計・機械学習で解けるかたちに持っていくのが、特に難しく、特に面白い**

機械学習プロダクトを提供する人間として、以下の心構えが必要と思う

- 提供する**モデルやアルゴリズムをどのような意図（仮定や前提）で設計したのか、その適用範囲はどこまでか（こういった場合にうまくいかないか）**、提供先のお客様にご納得いただけるまでご説明する
- データを分析するだけでなく、お客様のビジネスを把握すること**もまた重要であり、その上で現場のお客様が実際に「使える」ような機械学習のプロダクトをご提供する**



# アンケートのご協力をお願いいたします



[https://docs.google.com/forms/d/e/1FAIpQLSetwbx49vWbirGI-F4zEFDXBfoyeIVAVLO4As7Q4CgKhqrGMQ/viewform?usp=sf\\_link](https://docs.google.com/forms/d/e/1FAIpQLSetwbx49vWbirGI-F4zEFDXBfoyeIVAVLO4As7Q4CgKhqrGMQ/viewform?usp=sf_link)

# db tech showcase ONLINE 2020

27<sup>TUE</sup> OCTOBER - 10<sup>THU</sup> DECEMBER



# Thank You!