IBM **Watson** Summit 2017
Cognitive and Cloud

【M4-2】

# 機械学習か？ルール定義か？
## 言葉の処理の２つの側面

日本アイ・ビー・エム株式会社
Watson開発　開発リード
村上 明子

主任デベロッパー
野村 有加

# 講師紹介


## 村上 明子

- 入社以来、東京基礎研究所テキストマイニングチームに所属、テキストマイニングツールTAKMI（現Watson Explorer）の研究開発に従事
- 現在はWatsonの言語処理関係のソフトウェア開発の開発リーダー

## 野村 有加

- ソフトウェア開発研究所にて、入社以来ソフトウェア製品開発に従事。製品開発の他、様々なIBM製品のデリバリープロジェクトも経験
- 現在はWatson Knowledge Studioのユーザーインターフェースの開発に従事するデベロッパー

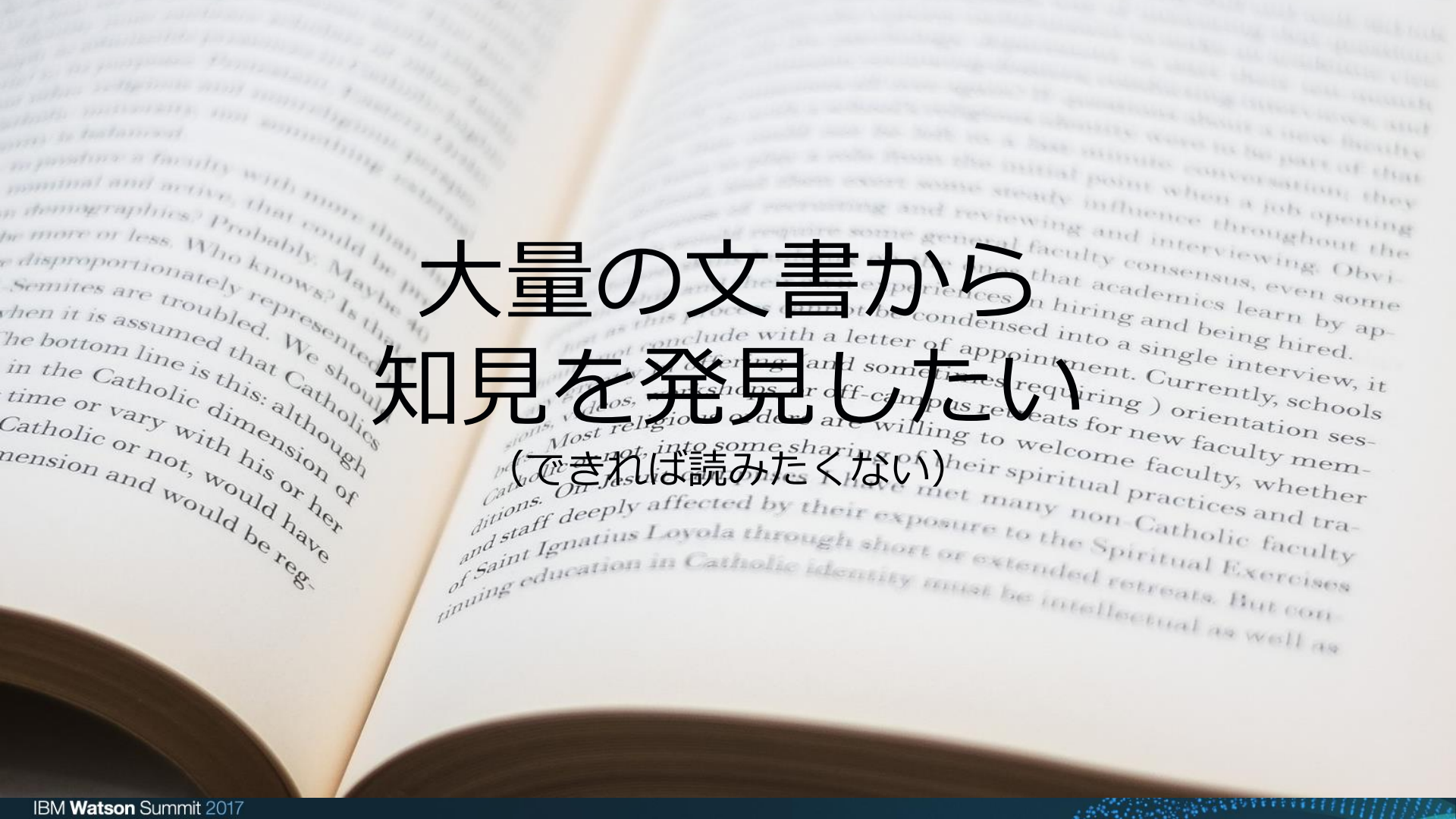
# 本日のお話の内容

- テキスト分析ユースケースのご紹介
- テキストからの情報抽出 - 機械学習とルール定義
- Watson Knowledge Studioご紹介デモ

本日覚えて帰って頂きたいこと

文書の山は宝の山！

道具は目的に合わせて選ぶ！

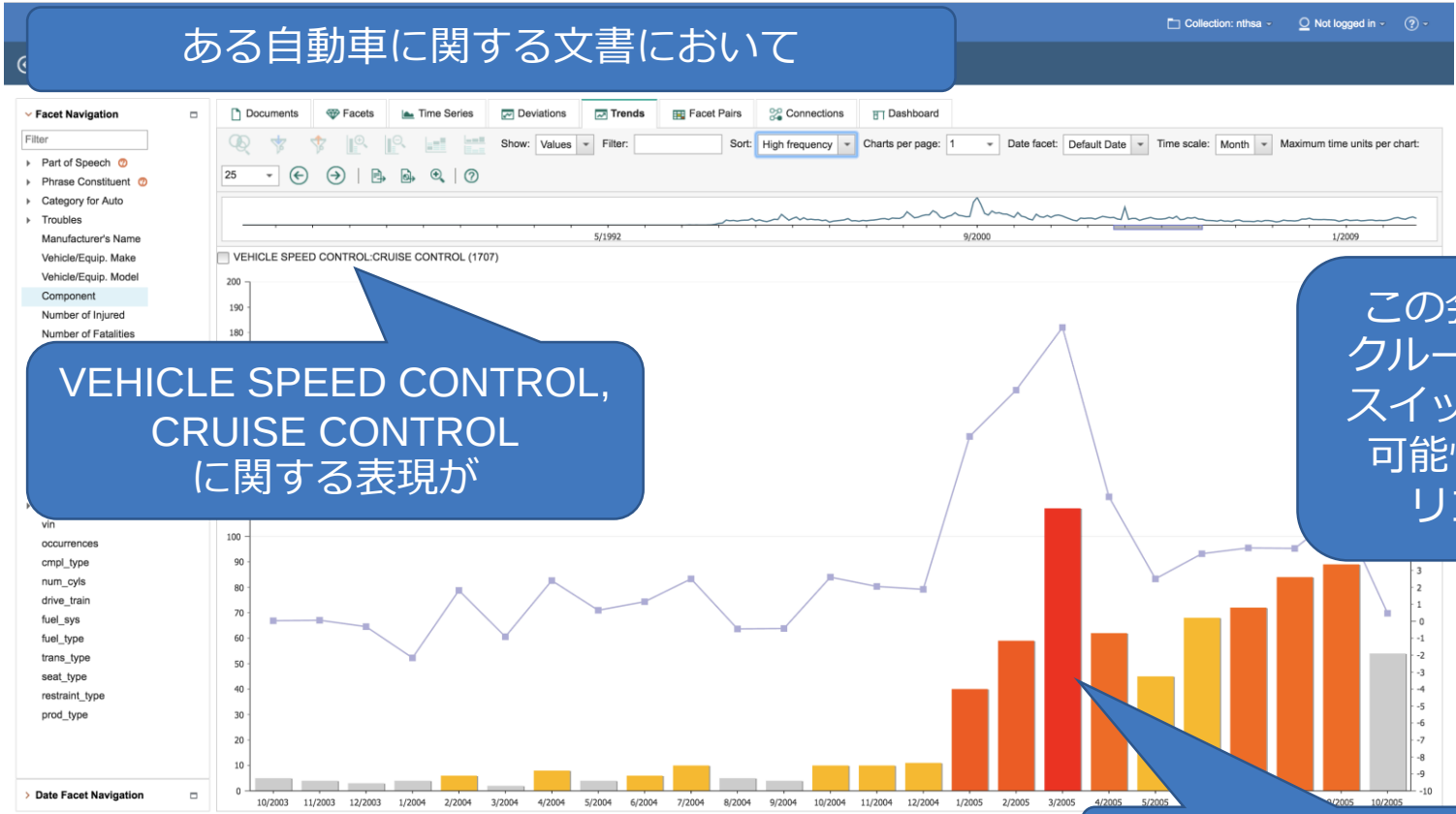


# 大量の文書から
# 知見を発見したい

（できれば読みたくない）

# 事例：自動車事故の報告書からの不具合の発見

ある自動車に関する文書において




VEHICLE SPEED CONTROL,
CRUISE CONTROL
に関する表現が

この会社は2009年に
クルーズコントロール
スイッチから火が出る
可能性があるとして
リコールを実施

2005年1月から急激に増加

# 事例：東日本大地震における「足りないもの」の理解

ガソリン

電池

水

ガソリンに関しては買占めが原因の品不足だったため、
被災地以外では震災後すぐに品切れ状態が解消

電池も同じく買占めが原因の品不足であったが、輪番
停電と余震への備えから4月半ばまで買占め状態が続く

震災直後は水（ミネラルウォーター）は品薄では
なかったが、水道汚染の報道とともに品不足が起こる

# テキストマイニングツール Watson Explorer

大量の文書から構造化データ、非構造化データを解析し、時系列や相関などを計算・可視化し、知見を発見するツール

川崎さんは神奈川県に住んでいます。

係り受け解析

形態素解析

| 川崎 | さん | は | 神奈川 | 県 | に | 住む | で | いる | ます |
|---|---|---|---|---|---|---|---|---|---|
| 固有名詞 | 接尾詞 | 助詞 | 固有名詞 | 名詞 | 助詞 | 動詞 | 助詞 | 動詞 | 助動詞 |

**定型情報**

**[保険商品]** 学資保険
**[性別]** 男性
**[年齢]** 70歳以上

**非定型情報[テキスト]**

【申出】 川崎さんは神奈川県に住んでいます。

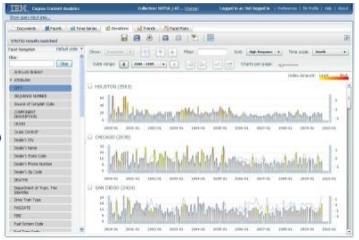
メタデータ抽出・作成部

定型部分取出し

テキストの形態素解析
キーワード抽出
係り受け解析
ファセット付与

[保険商品] 学資保険
[性別] 男性
[年齢] 70歳以上

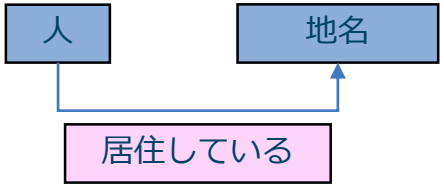
**[人名]** 川崎
**[地名]** 神奈川
**[動詞]** すむ
**[名詞-述語]** 川崎さん-住む



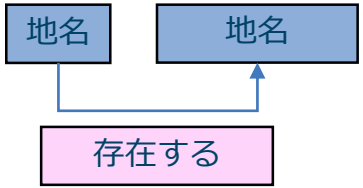
マイニングビュー

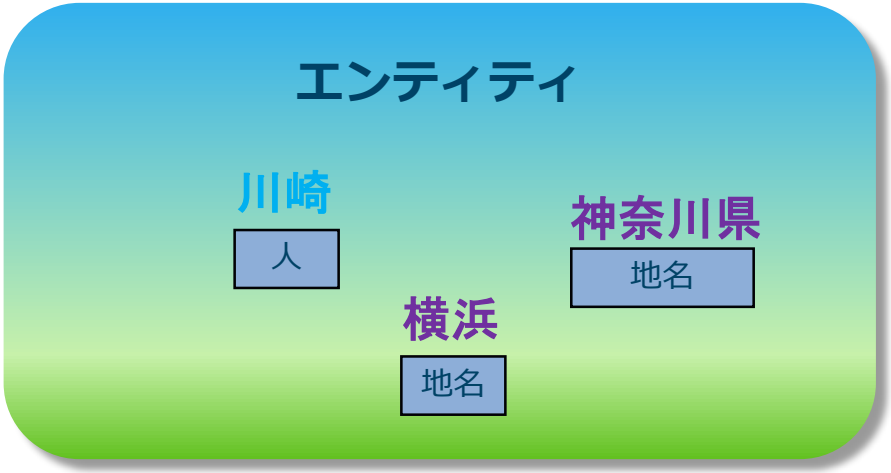
# テキストから抽出する構造化データ

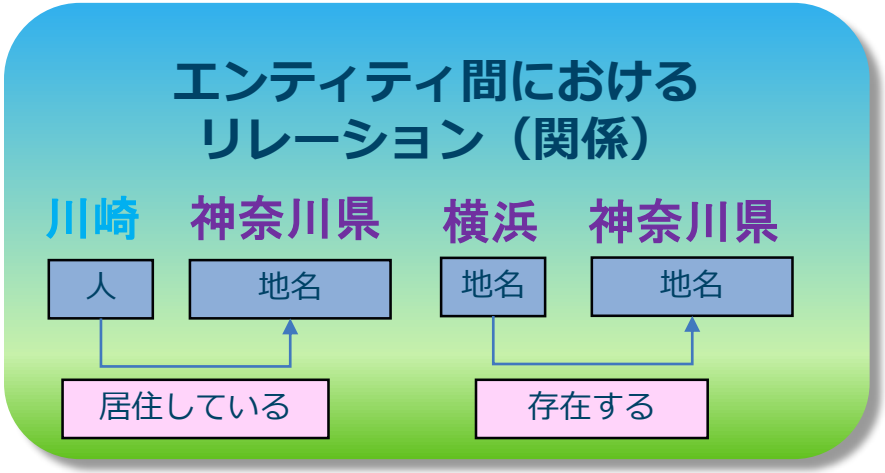
**川崎さんは神奈川県に住んでいます。**




**横浜は神奈川県にあります。**



# 分野専門知識の重要性

抗体・試薬等
操作

精製抗エボラウイルス核蛋白モノクローナル抗体（クローン 3-3D）を 1μg/ml に PBS (-)で希釈し，96 穴 ELISA プレートのレーン 1～6 の各ウェ ルに 100μlずつ分注する（図 1b）． 室温で 2 時間吸着させる（4℃で一夜 吸着させてもよい）

国立感染症研究所 クリミア・コンゴ出血熱診断マニュアルより引用
http://www.nih.go.jp/niid/images/lab-manual/ebora_2012.pdf

分野独自の情報抽出器が必要！

# テキストからの情報抽出　ー２つの手法ー

川崎さんが在庫についての質問した。

人

川崎に建設される倉庫に置きます。

地名

## 機械学習による情報抽出器

**多くの例を与えて、機械的に「モデル」を作る**

「川崎さんが在庫について質問した」
「川崎くんがご飯を食べた」
「川崎は明日来る予定です。」
…

「川崎に建設される倉庫に置きます」
「明日川崎に行きます」
「それは川崎にあると思います」
…

## ルール定義による情報抽出器

**人手によって「ルール」を生成する**

「○○さん」のように名詞に「さん」が続くものは“人”

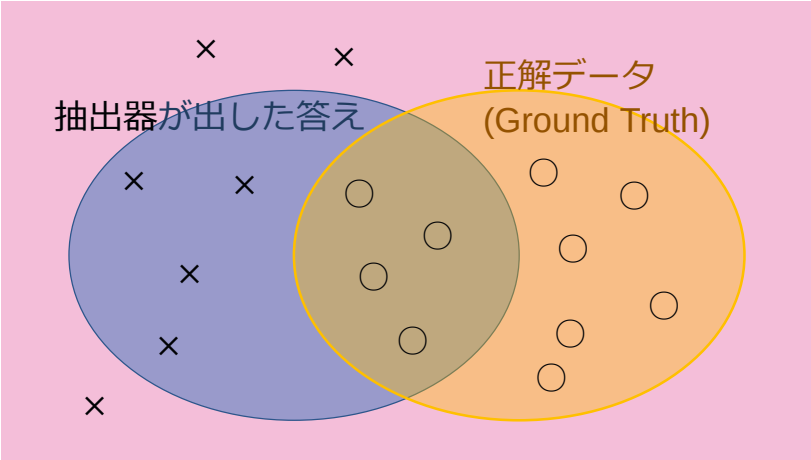
「○○に建設される」のように「建設される」と係り受けを持つものは“地名”

# 情報抽出 - 機械学習とルール定義

| | 良い点 | 悪い点 |
|---|---|---|
| 機械学習<br>(Maximum Entropy Model, etc.) | ▪文脈に即した抽出ができる<br>▪全体最適なモデルを作成できる<br>▪ビッグデータを活用できる | ▪抽出された理由が不透明である<br>▪見えている表現が抽出できないことがある<br>▪過学習する可能性がある<br>▪十分な学習データが必要であり、データ作成にコスト(時間)がかかる |
| ルール定義 | ▪抽出した理由が説明できる<br>▪メンテナンスや拡張が容易<br>▪小さく始められる<br>▪意図したものを取りこぼしなく抽出できる | ▪見えていないデータを抽出できない危険がある<br>▪ルール作成にある程度の習熟が必要<br>▪作成者によってばらつきがある |

# 情報抽出器精度の指標　ー適合率と再現率ー

- 情報抽出器の精度は以下の２つで判定
  - 「抽出器が出した答えがどれだけ合っていたか（適合率）」
  - 「抽出器がどれだけ正解を抽出できたか（再現率）」
  - 全体の精度は「F値」という再現率と適合率の調和平均で見る

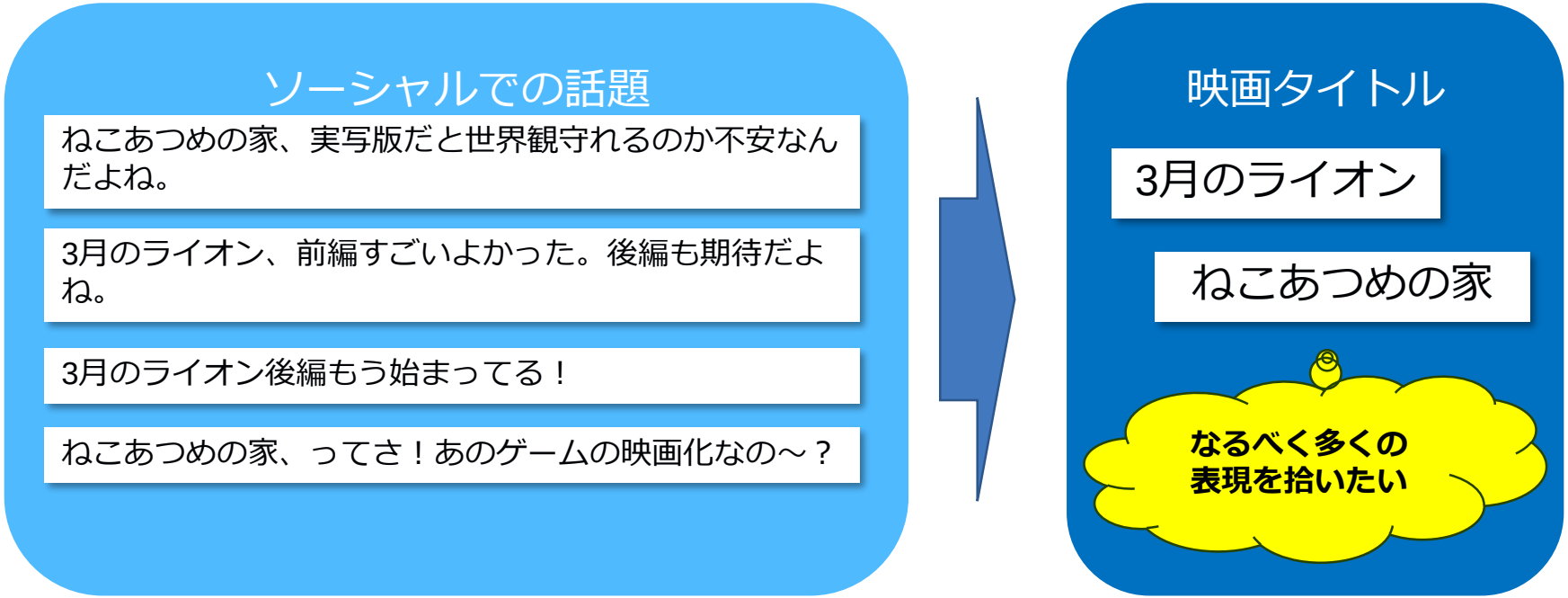



$$\text{適合率} = \frac{\text{抽出器が出した答え} \cap \text{正解データ}}{\text{抽出器が出した答え}} = \frac{4}{8} = 0.5$$

$$\text{再現率} = \frac{\text{抽出器が出した答え} \cap \text{正解データ}}{\text{正解データ}} = \frac{4}{10} = 0.4$$

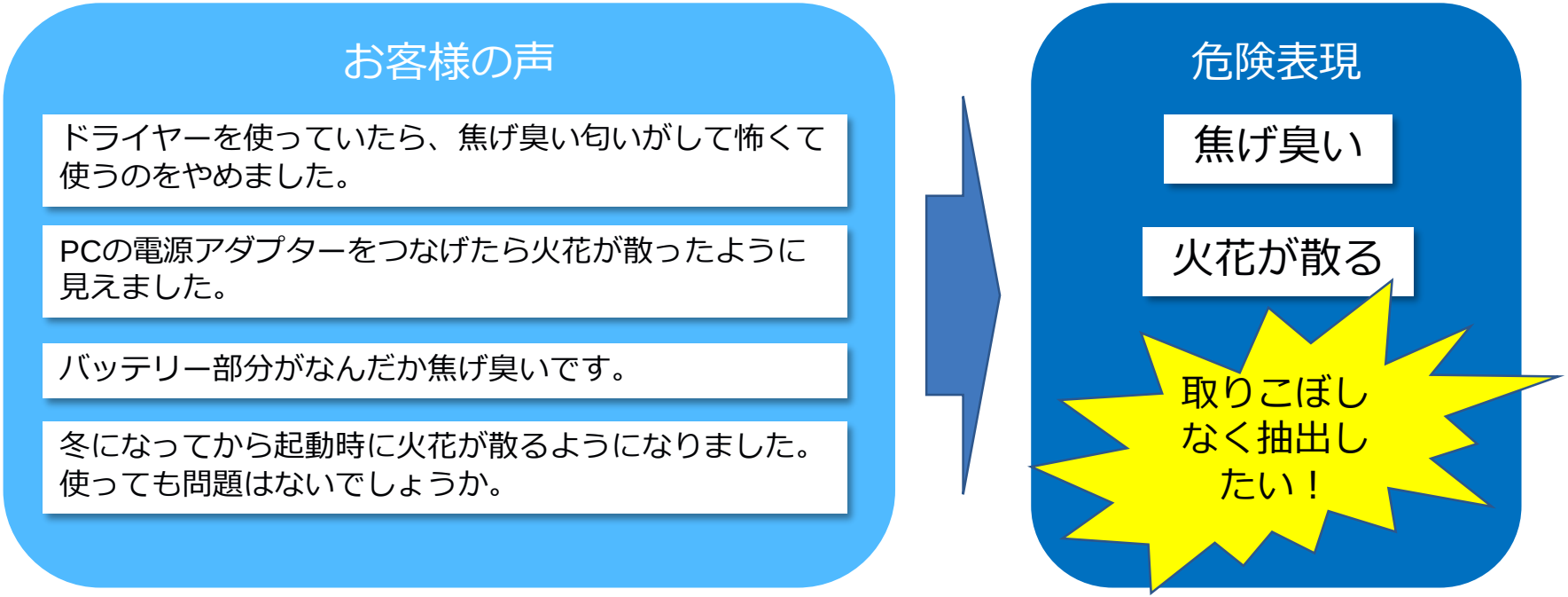
$$\text{F値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} = 0.44$$

# 目的に依存した情報抽出の手法 -広くカバーしたい場合-




再現率高　→　機械学習による情報抽出

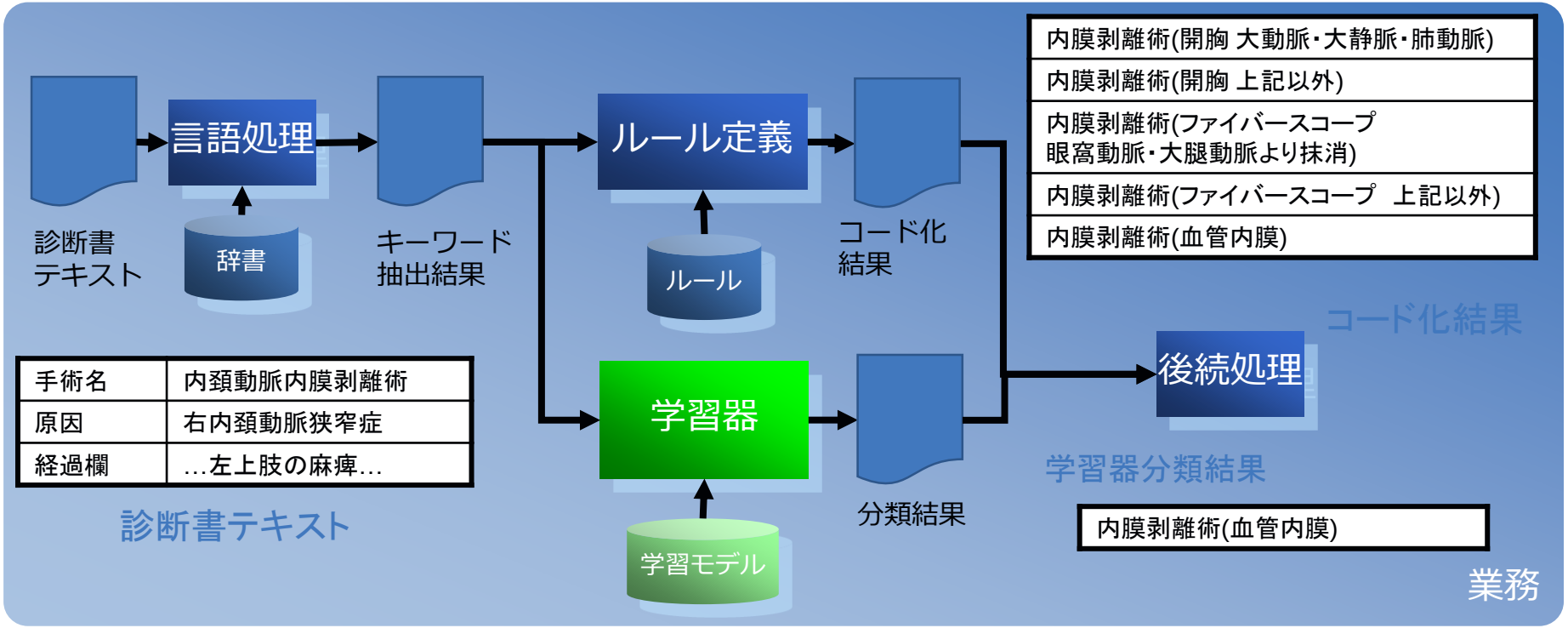
# 目的に依存した情報抽出の手法 -取りこぼしたくない場合-




適合率高　→　ルール定義による情報抽出

# 事例: 大手保険会社様

- ルール定義で抽出した手術コードを元に、機械学習を含む後続処理を行う



# ルール定義結果の正解データへの利用

- 機械学習には多くの「正解データ(Ground Truth)」が必要
  - しかし、正解データの作成には時間がかかる・・・！

ルール定義である程度の表現を抽出し、
それを正解データとできないか？

# IBM Watson Knowledge Studio

機械学習モデルやルール定義の作成により、業界や分野ごとの知識だけでなく、
各分野の言葉の使われ方の微妙な違いまでWatsonに教えることが可能になります


IBM創立者のトーマス・J・ワトソンSr.の第一子であるワトソンJr.は
1937年にブラウン大学を卒業、営業販売員としてIBMに入社し、
1956年にその会社の最高経営責任者になりました。

テキストからの情報抽出器の作成

**エンティティ**
IBM(会社)
ブラウン大学（大学）
トーマス・J・ワトソンSr.（人）
ワトソンJr.（人）

**関係**
トーマス・J・ワトソンSr.：IBM(創立した)
ワトソンJr.：IBM（入社した）
ワトソンJr.：IBM（最高責任者）

**照応関係**
IBM創立者＝トーマス・J・ワトソンSr.
その会社＝IBM

**直感的なUIでの機械学習正解データやルールの作成**

**共同作業と統計情報の可視化による機械学習モデルの改善**

**多くのワトソンソリューションとの連携**

**分野ごとのテキストアノテーター（情報抽出器）を作成・再利用・共有することが可能となります**
**これにより、Watsonソリューションをより強力なものにすることができます**

# Watson Knowledge Studio デモ

- **本日のデモで対象とした文書**
  - 医用画像診断装置の発明に関する特許文書

- **デモでお見せする内容**
  - ルール定義の作成方法
    - Rule Editor
  - 機械学習の正解データ作成
    - Ground Truth Editor
  - 両者を組み合わせて意味抽出を行うモデルの作成と精度評価

# IBM Watson Knowledge Studio

機械学習モデルやルール定義の作成により、業界や分野ごとの知識だけでなく、
各分野の言葉の使われ方の微妙な違いまでWatsonに教えることが可能になります


**「言語は生きている」**

**継続的な情報抽出器の
メンテナンスが重要
メンテナンスの必要性の判断も含め
情報抽出器の作成・再利用・共有が
ワンストップで可能**

**直感的なUIでの
機械学習正解データや
ルールの作成**

**共同作業と
統計情報の可視化
による機械学習
モデルの改善**

**多くのワトソン
ソリューション
との連携**

英語・日本語を含む9ヶ国語対応

無償トライアルをご用意

# 本日のお話の内容

本日覚えて帰って頂きたいこと

文書の山は宝の山！

道具は目的に合わせて選ぶ！

「コグニティブ・インフラストラクチャー」61番ブースにてデモンストレーション中です。お待ちしております

「Watson Knowledge Studio」
「ワトソンナレッジスタジオ」
で検索！
ぜひフリートライアルを
ご利用ください

ワークショップ、セッション、および資料は、IBMまたはセッション発表者によって準備され、それぞれ独自の見解を反映したものです。それらは情報提供の目的のみで提供されており、いかなる参加者に対しても法律的またはその他の指導や助言を意図したものではなく、またそのような結果を生むものでもありません。本講演資料に含まれている情報については、完全性と正確性を期するよう努力しましたが、「現状のまま」提供され、明示または暗示にかかわらずいかなる保証も伴わないものとします。本講演資料またはその他の資料の使用によって、あるいはその他の関連によって、いかなる損害が生じた場合も、IBMは責任を負わないものとします。 本講演資料に含まれている内容は、IBMまたはそのサプライヤーやライセンス交付者からいかなる保証または表明を引きだすことを意図したものでも、IBMソフトウェアの使用を規定する適用ライセンス契約の条項を変更することを意図したものでもなく、またそのような結果を生むものでもありません。

本講演資料でIBM製品、プログラム、またはサービスに言及していても、IBMが営業活動を行っているすべての国でそれらが使用可能であることを暗示するものではありません。本講演資料で言及している製品リリース日付や製品機能は、市場機会またはその他の要因に基づいてIBM独自の決定権をもっていつでも変更できるものとし、いかなる方法においても将来の製品または機能が使用可能になると確約することを意図したものではありません。本講演資料に含まれている内容は、参加者が開始する活動によって特定の販売、売上高の向上、またはその他の結果が生じると述べる、または暗示することを意図したものでも、またそのような結果を生むものでもありません。 パフォーマンスは、管理された環境において標準的なIBMベンチマークを使用した測定と予測に基づいています。ユーザーが経験する実際のスループットやパフォーマンスは、ユーザーのジョブ・ストリームにおけるマルチプログラミングの量、入出力構成、ストレージ構成、および処理されるワークロードなどの考慮事項を含む、数多くの要因に応じて変化します。したがって、個々のユーザーがここで述べられているものと同様の結果を得られると確約するものではありません。

記述されているすべてのお客様事例は、それらのお客様がどのようにIBM製品を使用したか、またそれらのお客様が達成した結果の実例として示されたものです。実際の環境コストおよびパフォーマンス特性は、お客様ごとに異なる場合があります。