SASによるテキスト・アナリティクス入門

津田高治
アナリティクス推進グループ

Introduction of SAS Text Analytics

Takaharu Tsuda
Analytics business development

要旨：

以下に要旨を記載（100文字以内）

テキスト・アナリティクスの分類問題（教師付き・教師なし学習）やトピック抽出・次元圧縮などのテーマに対するSASの方法論の紹介を行う.

ベイジアンネット，ベクター空間モデル，SVDなど、ビジネス現場での事例紹介と共にその方法・適用例・結果を議論する.

キーワード：続けてキーワードを記載

テキスト・アナリティクス，ベイジアンネット，正規表現，SVD，活用事例

SASの考える先進アナリティクス

先進アナリティクス

テキストマイニング

意味ある話題の発見、ソーシャルメディア、コールセンターから知見を発見し、さらに深い分析へ結びつけるそのことでビジネスへの貢献をする

時系列予測

時系列を分析し、将来の意思決定に貢献する洞察を導く

最適化

ビジネスのダイナミクスを捉え、リソースの最適配置を提案する

データマイニング

データから関連性を発見し将来に関する精確な予測を実現する

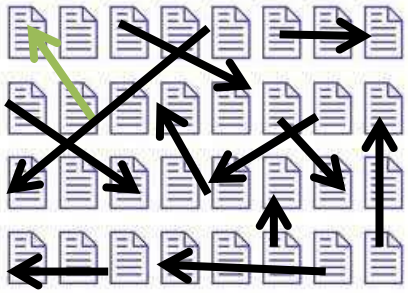
アナリティクス基盤

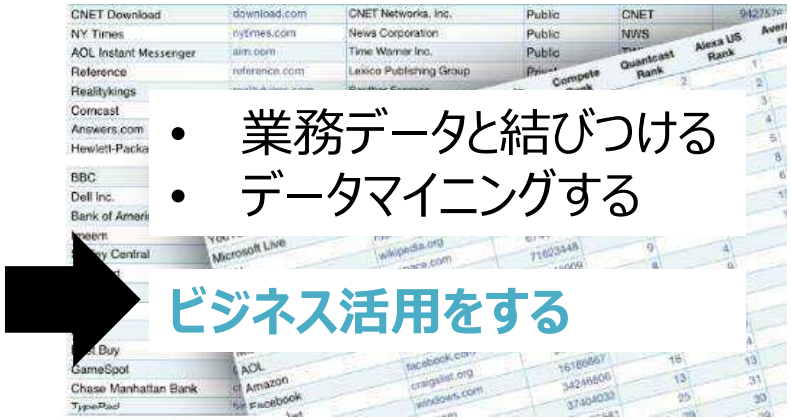
テキスト・アナリティクスとは

テキスト・アナリティクスとは

大量の文書から

隠れたパターン（知見）を発見





- 業務データと結びつける
- データマイニングする

ビジネス活用をする

でも知見で終わりではない

SASのテキスト・アナリティクスとは

SAS
テキストマイニング

データ準備			業務活用
アクセス	探索	情報追加	モデリング
ウェブ探索	語出現頻度	分類	予測モデル
ファイル情報検索	コンセプトリンク	固有表現抽出	時系列予測
SASのAPI	スペルミスの発見	感情分析	最適化
ファイル取り込み	ステミング	重複排除	マーケットバスケット
ドキュメント変換	トピックの抽出	言語識別	ネットワーク分析

エンタープライズ・レポート　ビジュアル化　アラート　サードパーティ・アプリと統合

SASのテキスト・アナリティクスとは

収益拡大 | マーケティングROI最大化 | 人件費削減 | 作業品質の標準化 | コスト削減 | イメージアップ

活用

マーケティング
顧客の興味エリアを理解し、リコメンデーションする（協調フィルタリング）

コールセンター
クレーム/要望などの感情・緊急度、製品区分等でテキストを自動振分け

セキュリティ・ディフェンス
事件の関係者や場所・モノ（クルマ、武器）などの関係性を描く、それと自社との関係をアラート

製品品質・製品開発
品質業務の高度化をテキストに基づく予測分析で行う、評判を製品開発に活かす

分析

SAS® Text Miner
- 自動分類：事前定義に従って自動振り分け
- 多様な分類器：ベイズ推定、K近傍等

SAS® Contextual Analysis
- 自動トピック：書かれている内容に従って文書群を自動トピック抽出
- 固有表現・事実の抽出：人名・地名・時間・モノなどを抽出する
- 文書レベルでの感情分析：ドキュメントレベルで感情分析を行う

SAS® Sentiment Analysis
- 詳細なレベルの感情分析：一つの文書の中で好評・不評が混在する、最も詳細レベルで評判分析
- 多言語対応：世界の27言語以上で自然言語処理を行う

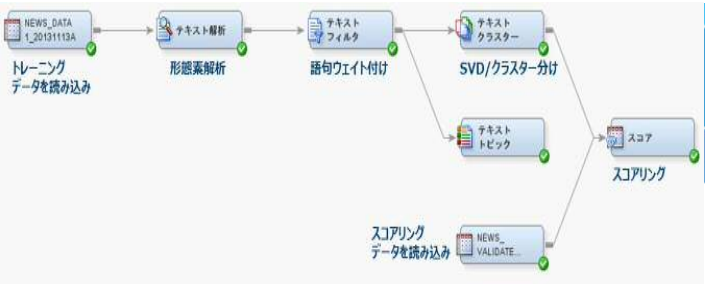
データ

多様なデータアクセス（各種データ形式対応、データソース接続、等）

SNS	フェイスブック	ブログ	ツイッター	ニュース
コールセンター	営業日報	論文・技術文書	社内文書	Eメール

統合メタデータ管理

事例と基本技術



導入事例A-声を発見する

①元の文章

行数＝文書数

「私はカフェで紅茶を飲むのが毎日です。」

毎朝エクセルシオールカフェでコーヒーを飲んでいる

私は紅茶は飲まないのです

②形態素解析の結果（列数が多い）

	語句の一致											
	私	は	カフェ	で	紅茶	を	飲む	のが	毎日	です	。	..
一致＝1	1	1	1	1	1	1	1	1	1	1	1	
			1			1	1					
	1	1			1		1			1		

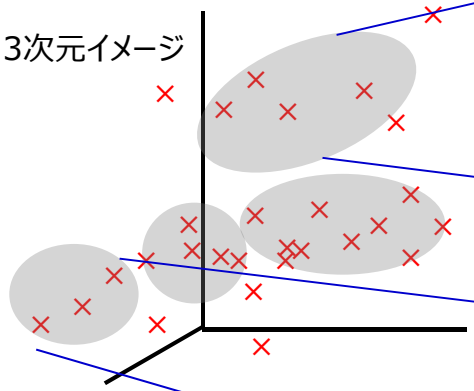
③次元数を落とす

	少ない列数（例.100）
行数＝文書数	1 0 1 0 0 1 1 0 ..
	0 0 0 1 1 0 0 0 ..
	0 0 1 0 0 1 1 0 ..
	0 1 0 0 1 0 0 0 ..

④n次元空間で近い文書をグループ化⇒「トピック」

低次元座標でクラスタリング

3次元イメージ



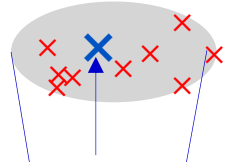
トピック＝携帯電話やスマートホン

1 よほど在庫が積み上がってるのか、ドコモ SC-04D Galaxy Nexus
2 私が持っている電話の中で、主に通話用に使っているのはウィル
3 スマートフォンじゃなくて、携帯電話をまだまだ使っている人って多い
4 うっかり買ってしまった iPhone 5s のおかげで余った iPhone 5 を、
5 先週やっと自宅のネット回線をWiMAXから固定回線に変更しました
6 『WiMAX 2+』のサービスが発表されたので、現在WiMAXサービスを
7 「わださん、このサービス使えば日本への電話が激安です！」短期

トピック＝B級グルメ

1 過去数回にわたって「こんな店ができますよ」的な紹介をしてきた板橋本
2 [image via Frederic Poirot's filckr] 世界でも広く知られている日本料理
3 ベトコンラーメンにハマって、近所のラーメン屋について色々と調べていた
4 昨日はムスメと地元パスタ屋ランチなどを食しノーラーメン。となれば今日
5 新潟の燕三条に行ったら何を食べたら良い！？という質問に、速攻で、し
6 気がつけば10月。そろそろ鍋の季節がやってまいりますね。恋人と、友
7 西台店に行った。今日のところは、特にメルマガなどで告知されてなかっ

⑤注目すべき記事を特定



おすすめ⇒メンズ108春もの柿落し

トピック＝渋谷ファッション

導入事例A-声を発見する～結果のイメージ

クラスターのイメージ

- トピック（クラスタリング）は一定の正確性がある
- 特徴が顕著な記事に高い注目度が付く傾向がある
- 情報量の多い記事に高い注目度が付く傾向がある

事件など

24	1	東京都三鷹市の女子高生殺害事件を受け、各メディアがこぞ
24	2	東京都三鷹市の女子高生殺害事件を受け、各メディアがこぞ
24	3	1:そーきそば Φ ★:2013/10/09(水) .97 ID????0東京都千代
24	4	(前略)…三鷹殺害 ()茨城・常陸太田市などで震度4 ()かつ
24	5	(前略)…、東京・三鷹市内で起きた女子高生刺殺事件で、被
24	6	(前略)…三鷹の女子高生のストーカー事件のアニメオタク報
24	7	神戸拓光 (kobe_takumi55) on Twitter より　不適切ツイート
24	8	東京都三鷹市の女子高生殺害事件を受け、各メディアがこぞ
24	9	千葉ロッテの選手を名乗る人物が2013年10月10日、三鷹市
24	10	8日夕方、東京・三鷹市の住宅街で、18歳の女子高校生が

パソコン 2013秋冬モデルについて

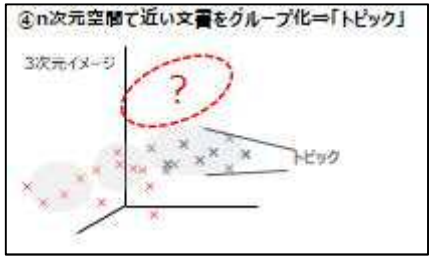
17060	VAIO 2013年秋冬モデルが発表されて、この夏に
524	VAIO 2013年秋冬モデルを発表！・ノートPCとして
3345	VAIO 2013年秋冬モデルが発表されて、この夏
9757	VAIO 2013年秋冬モデルを発表！・ノートPCとして妥協のない
11216	カスタマイズといっても実際にどれを選んだら良いのか迷います
10752	カスタマイズといっても実際にどれを選んだら良いのか迷います
13777	VAIO 2013年秋冬モデルが発表されて、この夏に登場した「VA
12223	VAIO 2013年秋冬モデ
13205	ますます高性能になっ
16053	VAIO 2013年秋冬モデ

グルメについて

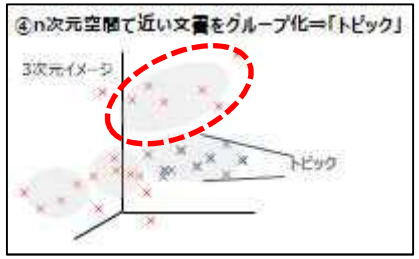
8153	ベトコンラーメンにハマって、近所のラーメン屋について色々と調べていたら
24820	過去数回にわたって「こんな店ができますよ」的な紹介をしてきた板橋本町
15490	[image via Frederic Poirot's filckr] 世界でも広く知られている日本料理。し
37310	ベトコンラーメンにハマって、近所のラーメン屋について色々と調べていたら
6941	昨日はムスメと地元パスタ屋ランチなどを食しノーラーメン。となれば今日は
15693	新潟の燕三条に行ったら何を食べたら良い！？という質問に、速攻で、しか
13212	気がつけば10月。そろそろ鍋の季節がやってまいりますね。恋人と、友人や
7045	西台店に行った。今日のところは、特にメルマガなどで告知されてなかった
24817	前回同様にまだまだ書かなきゃならない店もありますが、気になるネタがあ
19082	新橋駅(烏森口)から徒歩3分程にあるお店。今日の昼はガッツリ食べたい

導入事例A-声を発見する～変化を捉える

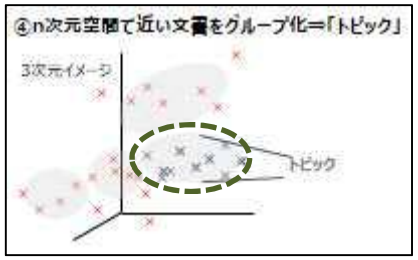
変化1



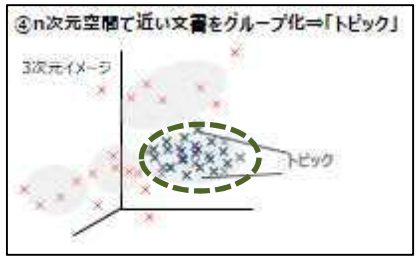
新しい話題の出現



変化2



話題の密度や件数の増加



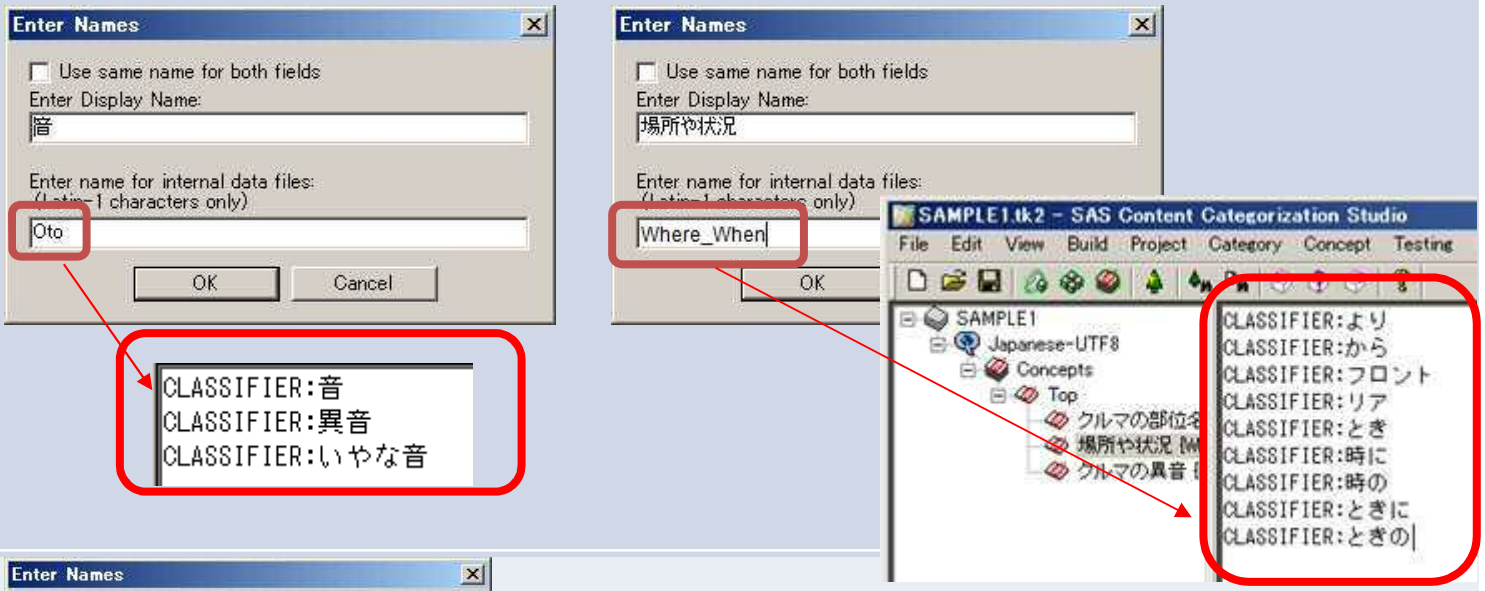
※密度は注目度、件数は盛り上がりを指すと考えられる

導入事例B-声を分類する

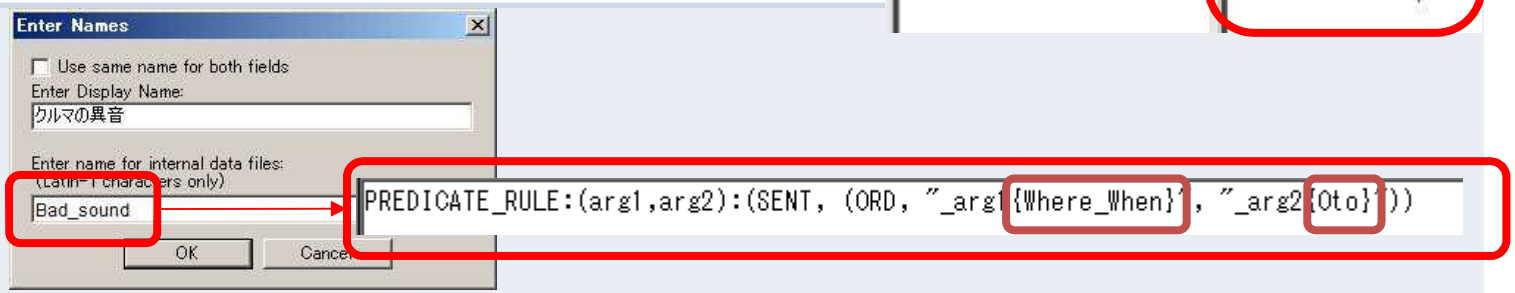
まず、固有表現を抽出するロジックと結果

「異音」
に関する固有表現

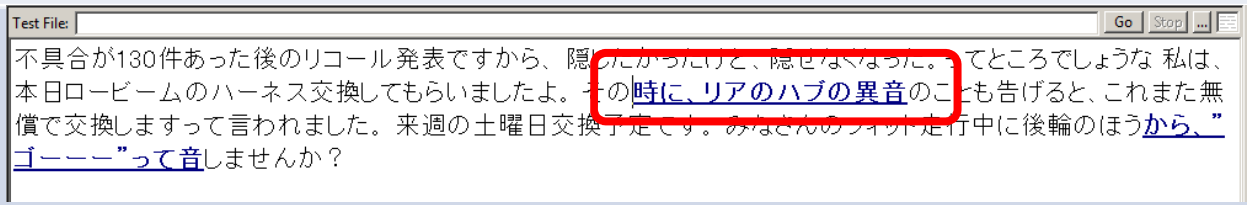
「場所や状況」
に関する固有表現



「クルマの異音」
に関する事実抽出



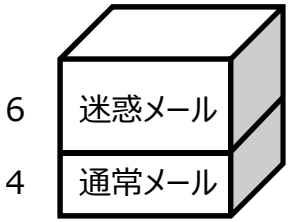
【テスト結果】
テキストの中から
「クルマの異音」
に該当する
事実の抽出が行われた



導入事例C-声を分類する

1)予め過去のデータで学習しルールを導いておく

過去のメール集



例．「アイドル」や「無料」など迷惑メールに特徴的な語句、「統計」や「科学」など通常メールに特徴的な語句を選択**(ツールで選択)**

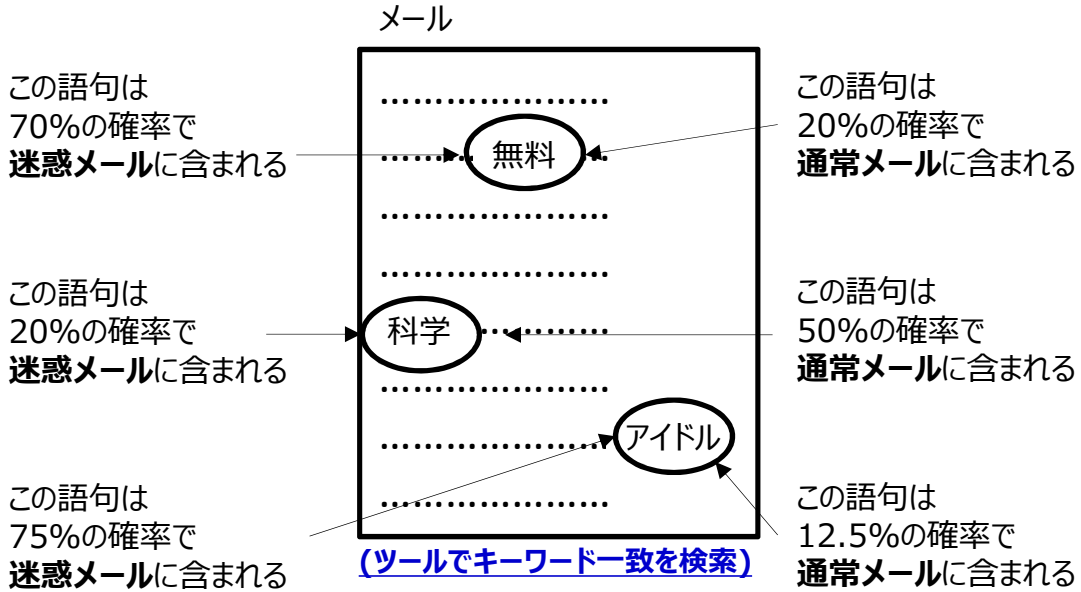
過去のデータ（メール）から以下の値を得る
(ツールで算出)
⇒通常メールと迷惑メールの比率は4:6
⇒キーワード別の
　通常メールと迷惑メールそれぞれに
　含まれる確率

語句	通常メール	迷惑メール
統計	0.8(*1)	0.1
科学	0.5	0.2
無料	0.2	0.7
アイドル	0.125	0.75

(*1)
通常メールが100通あれば、
その中の80通に「統計」という語句が含まれる

2) ルールを新しいデータに当てはめ、結果を導く

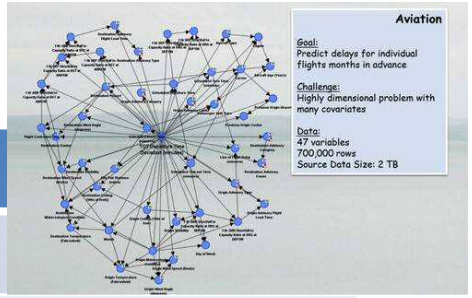
通常メール：迷惑メール ＝ 4:6



（このメールが）
通常メールである確率：迷惑メールである確率
＝ (0.2×0.4)× (0.5×0.4)× (0.125×0.4)：
　　(0.7×0.6)× (0.2×0.6)× (0.75×0.6)
＝0.0008 ：0.02268 **(ツールで算出)**
よって迷惑メールである確率の方が高い

導入事例C-声を分類する～分類精度

ベイジアンネットのイメージ図



下記の精度報告は特定データを用いた例示であり、一般的な精度保証をするものではありません

	カテゴリー判別	感情判別
分類精度	95%	91%
分類対象	ニュース、SNS	ニュース、SNS
活用した手法	• Naive Bayes • Tree augmented network (TAN), • Bayesian network-augmented naïve (BAN), その他	左記に加え 係り受けを考慮した分析
分類精度の算出手法・根拠	• 教師データから分類ルールを機械学習で導く • 検証データを分類ルールに則って自動分類 • 検証データの分類結果を検証	
誤分類の原因	同一キーワードが複数の意味で使われることに対応できていない	皮肉表現やスラング，抑制表現に対応できていないため
誤分類の改善	ニュースとSNSの2種類のデータそれぞれにモデル構築するなどで、表現の多様性に対応する	多様なニュアンスに対応できる教師データを用意するなど機械学習の学習材料を豊富にする

ビジネス活用～アプローチ別

導入事例	技術	分析のタイプ	手法の例	目的	活用例
導入事例A	機械学習	教師なし	SVD	未知のトピックを発見	ネットの声を把握
導入事例B	ルールベース	教師あり	ベイジアンネット	サブジェクト・エキスパートが高い精度で分類ルール作成	製品品質など基幹の業務
導入事例C	機械学習	教師あり	正規表現	一定品質で標準的な分類	顧客の声からリコメンド

参考～テキストのサマリー

方法

①語に重みをつける（以下はエントロピーのイメージ）

エントロピー大きい

ジェリービーン1つを語とイメージ
（語が内容を示唆しない）

エントロピー小さい

語が内容（この例では位置）を示唆する

【語の重み】とは文書内における重要度（内容を示唆する力）のこと

②文の重みを算出、特に重要度の高い文をピックアップする

③話し手のクセ（言いたいことを言う前のキュー）などを考慮（プロジェクト別でチューニング）

④テキストサマリーとする

結果

ミーティングのテキスト起こし（あー、まぁ、などもそのまま収録。）

サンプルインプット

メンバー#	コメント
S	おはようございます
S	Zさんはどうした
S	Zさん、と
S	わからない
B	わからないことは考えない
S	P社か
C	おはようございます
S	おはようございます
D	今日こんだけかな
C	あとは聞いてないです
C	Vさんま
S	ちょっと見ましょうか
S	リンクで呼ぶか
B	いない
S	いや、[illegible]の上にいない
S	席はず[illegible]
D	おはようございます

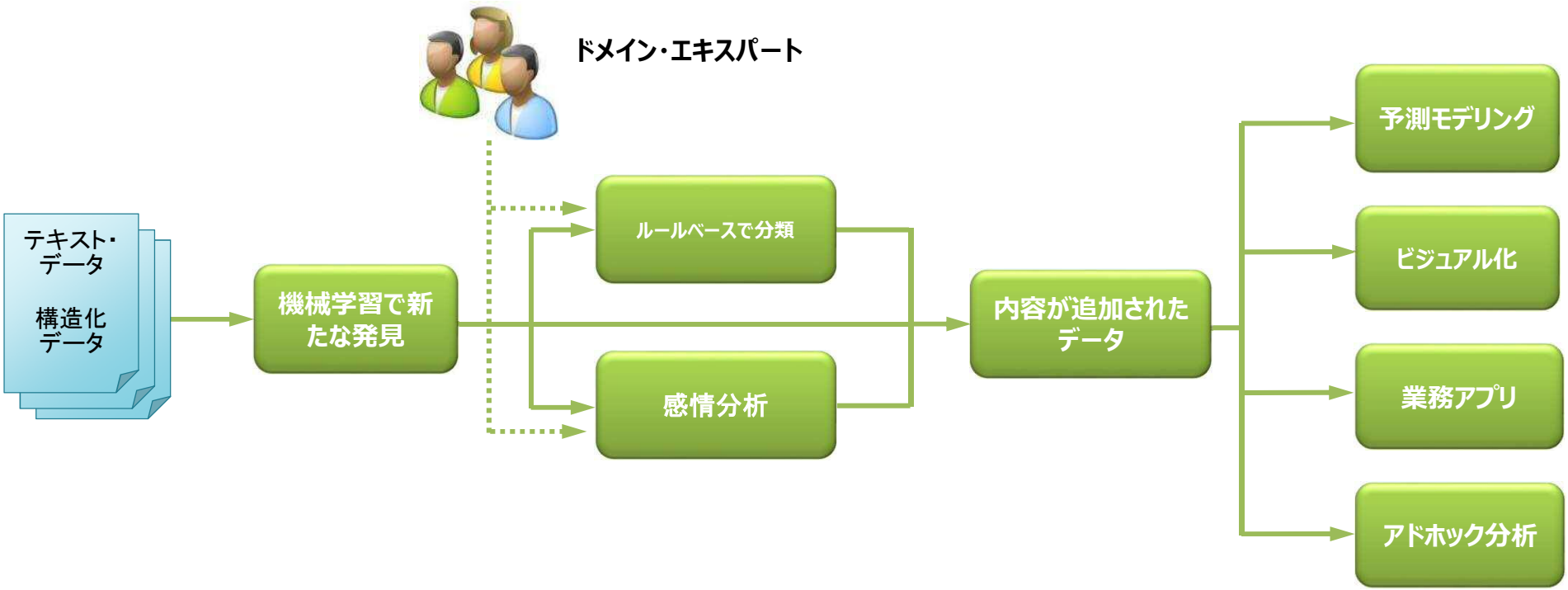
15分程度の会話ログを9行にサマリー（おおよその内容を示唆）

サンプルアウトプット

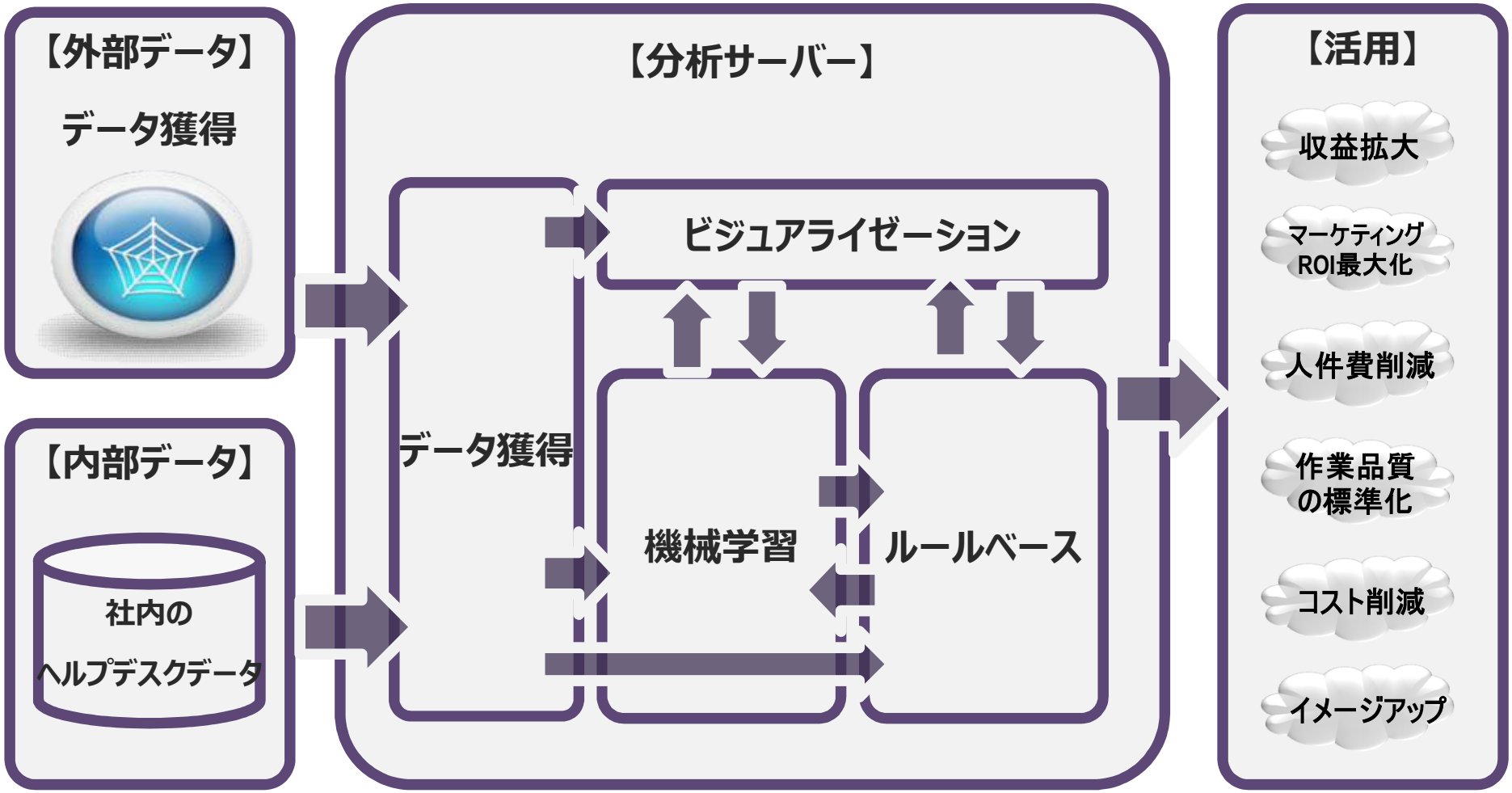
SAS システム

OBS	person	comment
76	B	分析者育てた2年間J社のサービスを行うことを前提に何かいくらかで[illegible]ングコース持っているらしい
92	B	15人がマックスで別に15人じゃなくても構わないんだけども、[illegible]の技術者育てているのついでに
96	B	データ周りは詳しいのだけれどそういう人たちに分析のスキルをつけさせたい、っていうのがH社さんの
111	B	いやただ、リソースとしてやるんじゃなくて、[illegible]やるのその半年か一年間いくばくかでお預かりし
124	S	夏の発表シリーズがいくつかあって[illegible]一緒にやっているのが8月25日のP3新製品発表会と
129	S	リストが出てくる、今週、[illegible]をベースにデモを作れないか考えて、具体的にやることはデモと
133	S	これは一般的な話[illegible]ですが、ただマーケティングとしては弊社のリスト集客だと広がらない
153	S	D社F社、と[illegible]に今入り込もうしていて、D社は製品は売れたがユースケースを売り込もう、F社
157	S	まだ[illegible]ばっかり現時点はデータの整理と分析の準備をなさっていてそれに難航している、

テキスト・アナリティクスのプロセス



システムの概念図


【外部データ】
データ獲得
【内部データ】
社内の
ヘルプデスクデータ
【分析サーバー】
データ獲得
ビジュアライゼーション
機械学習
ルールベース
【活用】
収益拡大
マーケティング
ROI最大化
人件費削減
作業品質
の標準化
コスト削減
イメージアップ

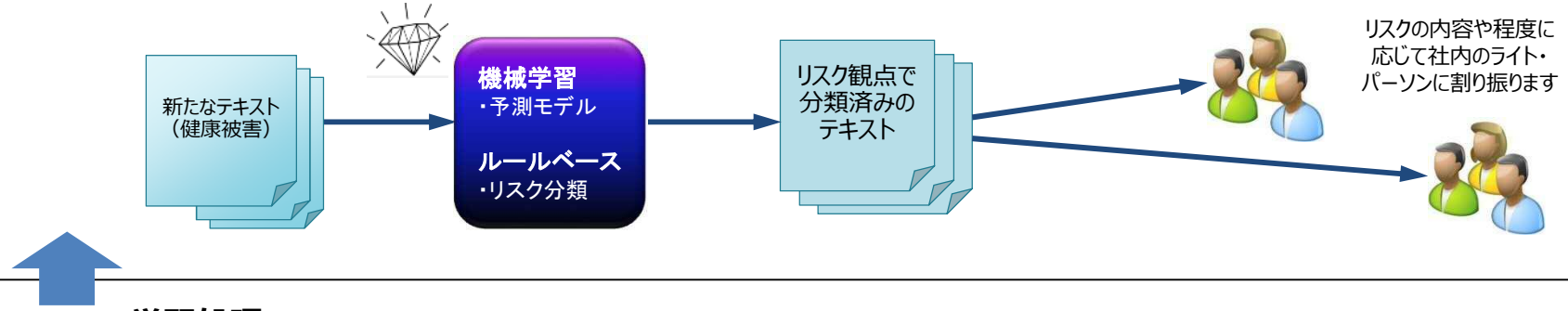
テキスト・アナリティクスの発展

テキストを分析する	テキストから知見を得る	テキストからアクションする	アクションを精緻化する
・頻度 **・分類** **・インデックス付け**	**・自動トピック抽出** **・エンティティの抽出** **・要約**	**・ビジネスを予測する** **・リコメンドする** **・プロファイリングする**	**・パワフルな機械学習** **・精緻なルールベース** **・2つの統合と発展**
例．顧客を知るために •テキストを読む •人海戦術	例．顧客を知るために •興味エリア・マップ •関係者のマップ	例．顧客を知るために •未来のビジネスを知る •個人の興味を知る	例．顧客を知るために •未知の声を知る •既知のルールにする •リコメンド精度の向上

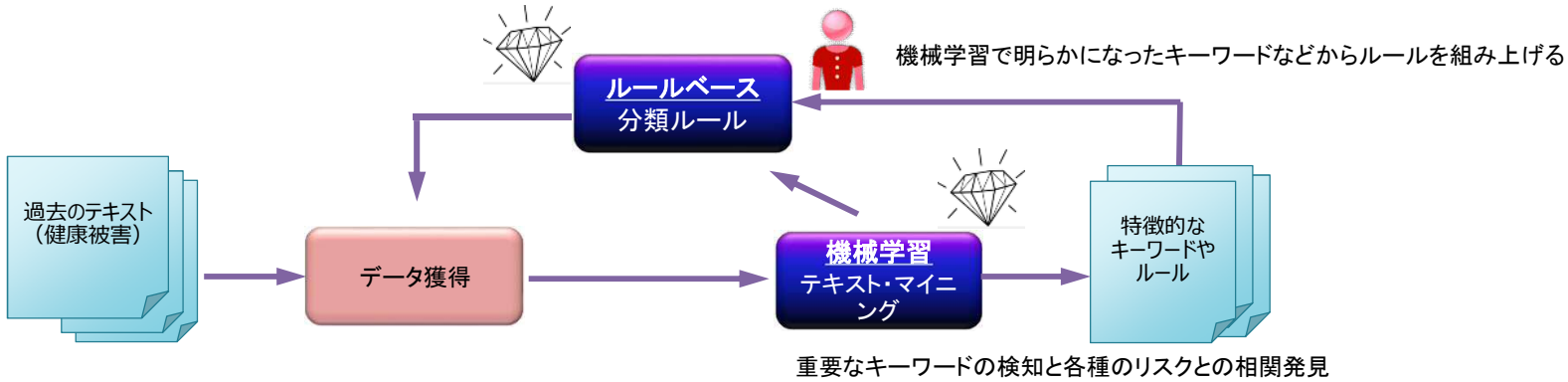
事例紹介

事例～製薬業界

オンライン処理（顧客の声などテキストが発生した際にリスクスコアを付与）



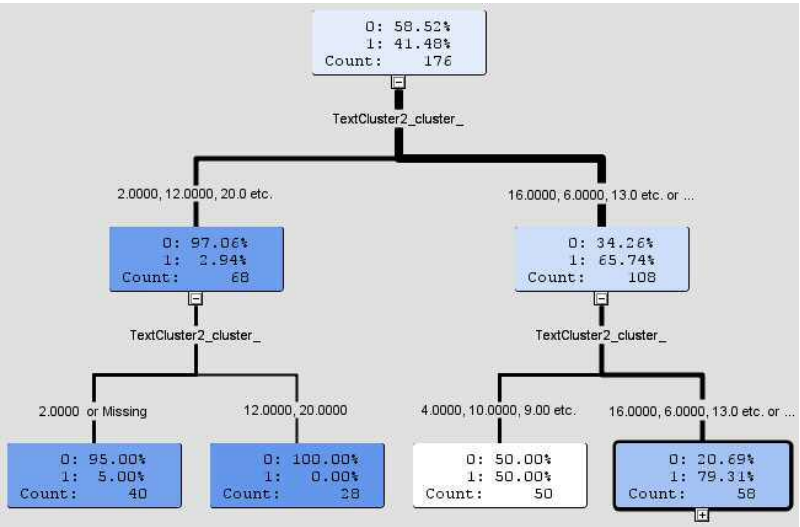
学習処理（機械学習、ルールベースでリスクとの関連付けを構築する）



事例～製薬業界

機械学習のイメージ

機械学習	**特定リスクの有無**をテキストの内容で説明する
活用	新たなテキストに対して、そのテキストの内容から**特定リスクの可能性**を算出する



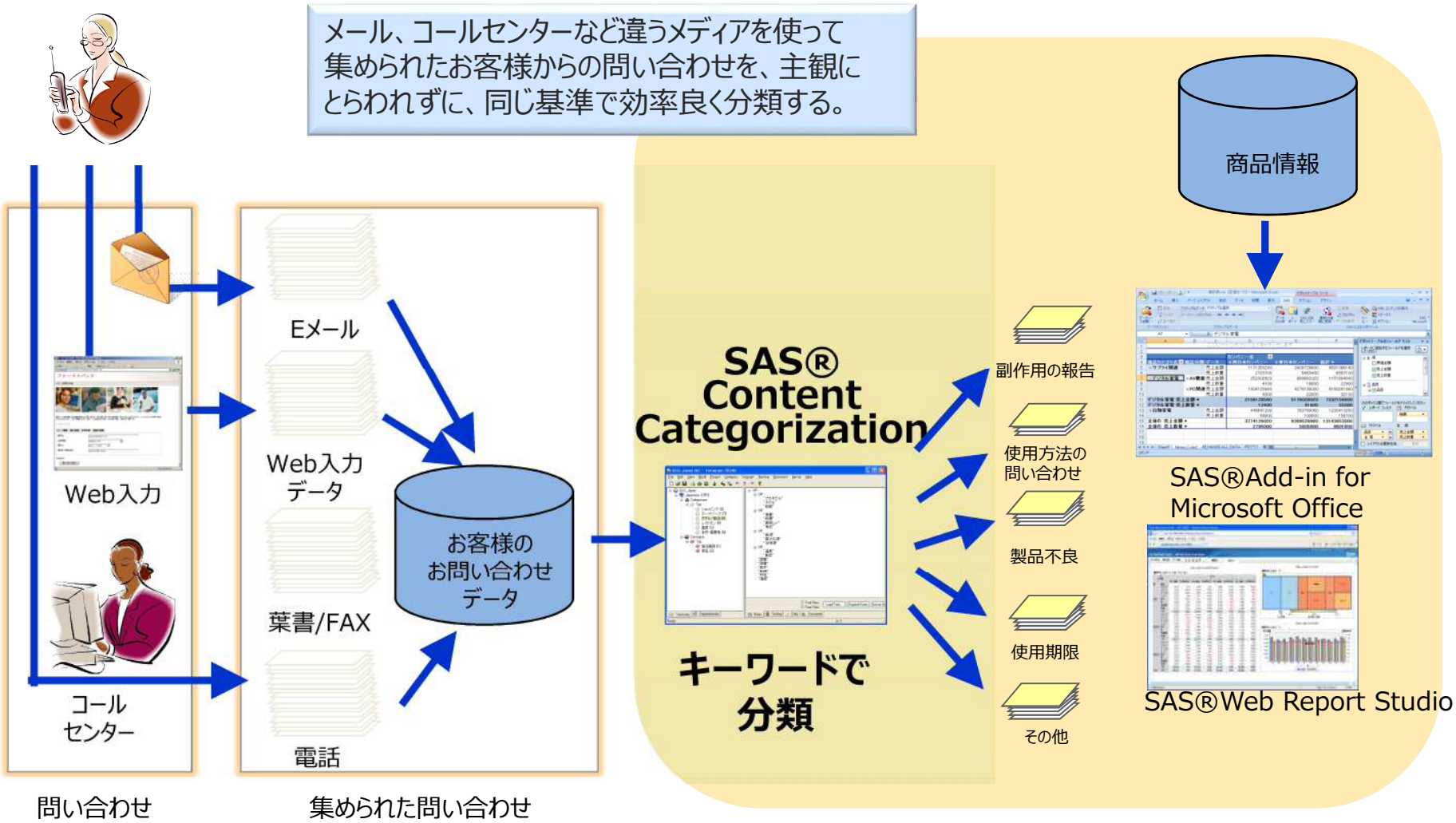
リスク関連キーワードのイメージ

ハイ・リスク

ロー・リスク

事例～製薬業界

メール、コールセンターなど違うメディアを使って
集められたお客様からの問い合わせを、主観に
とらわれずに、同じ基準で効率良く分類する。

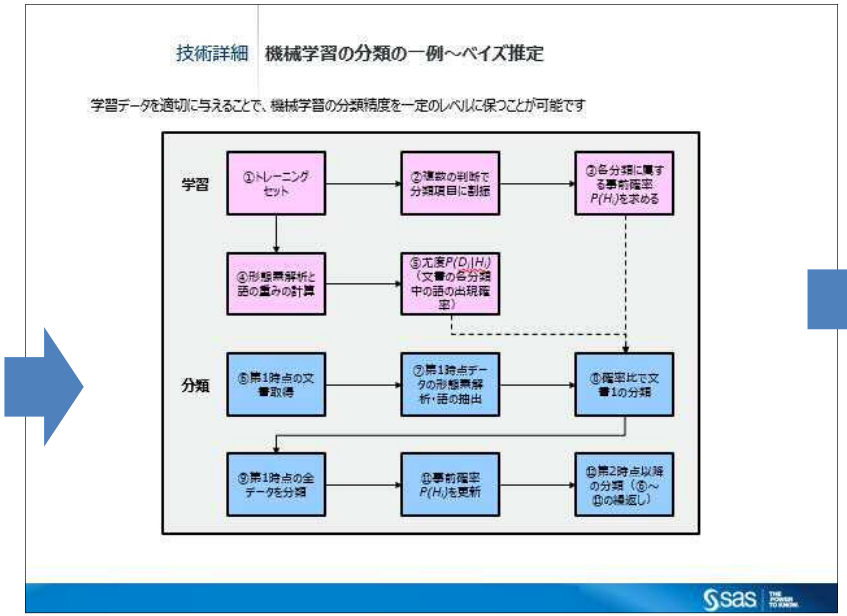


事例～製造業

機械学習のイメージ（ベイジアンネット）

テキスト情報
・ニュース
・SNS

データ獲得



6	内容	企業の評価 価格 製品の品質 使い勝手 販売活動 その他
×		
3	感情	ポジティブ ニュートラル ネガティブ

キーワードの例など（クルマ）

内容の分類
【企業の評価】（例．会社）
【価格】（例．キロ）
【製品の品質】（例．リコール）
【使い勝手】（例．初めて）
【販売活動】（例．人気）

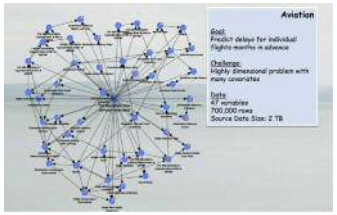
感情の分類
【ポジティブ】（例．思う、良い）
【ネガティブ】（例．事故、走行）

クレームにつながるキーワード

クレーム・テキストから機械学習で抽出することが可能です。

上述のキーワードがその例です。

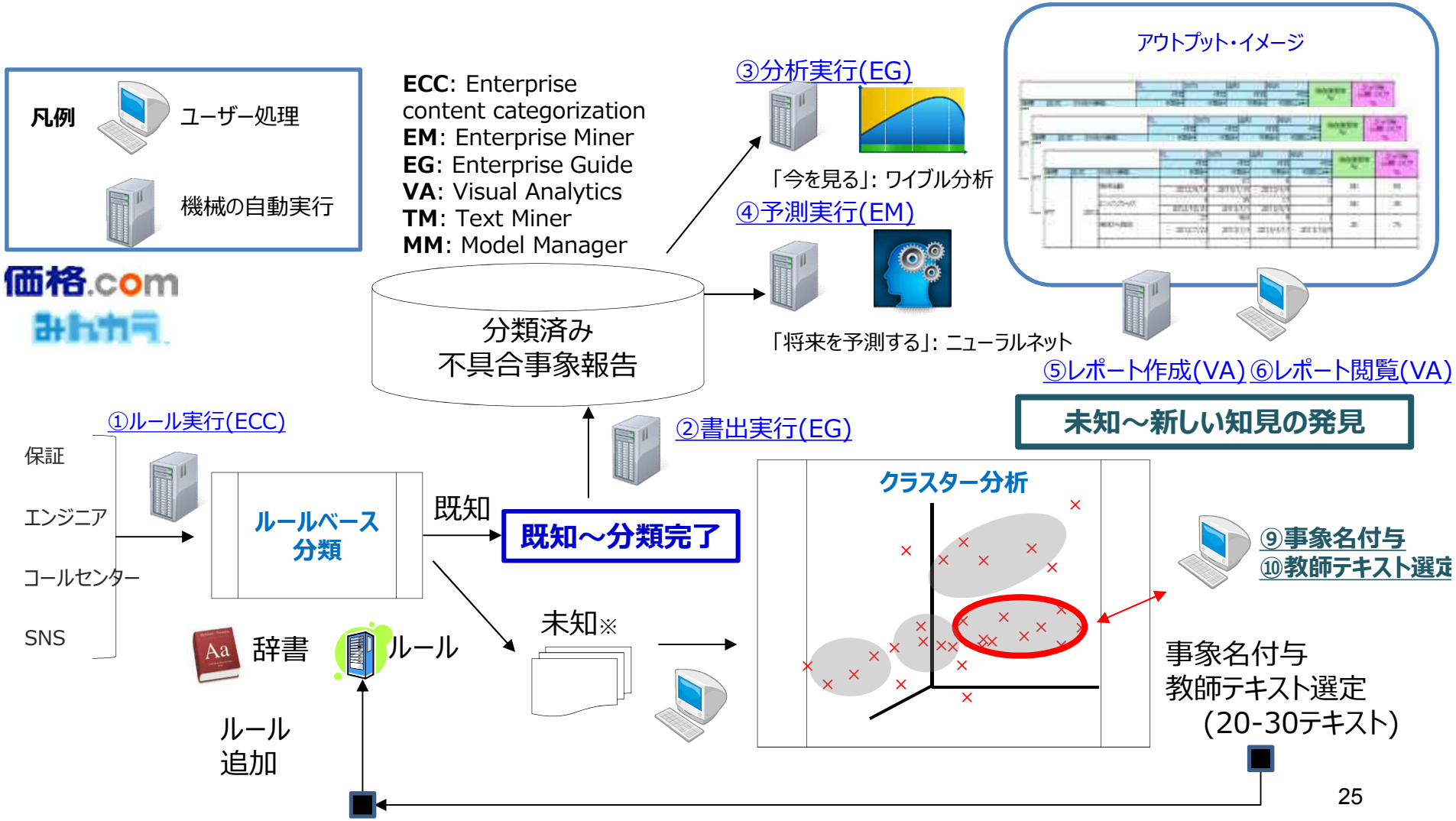
ベイジアンネットのイメージ図



分類	正解率（例）
内容	86% (86/100)
感情	80% (80/100)

事例～製造業

青字：既知の不具合事象に対する処理
緑字：未知の不具合事象に対する処理



事例～通信

ジャンル

網頁分類層級一	網頁分類層級二
藝術文化	
	導覽
	活動
	攝影
	歌手歌星
	評論
	繪畫
	建築
	傳統藝術
運動體育	
	職棒
	職籃
	活動
	運動明星
	滑板
	武術
	足球
	生存遊戲
科學科技	
	颱風
	地震
	天文學

コンテンツ

ユーザーのサイトアクセス

ユーザーの興味を知る

モデル

ジャンルと特徴的な語句との関係を統計的に導く

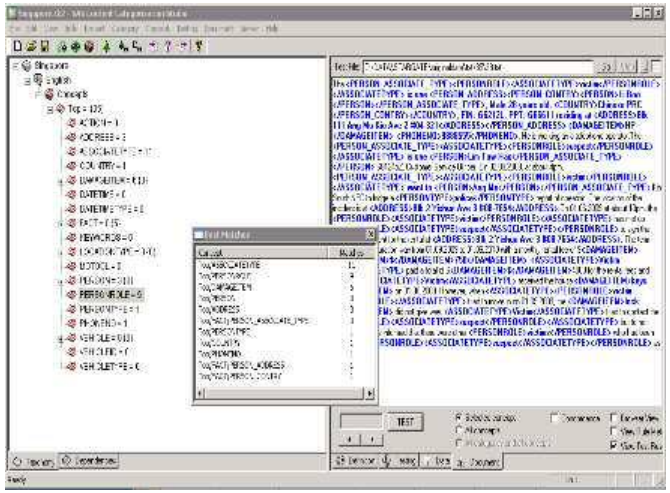
sas THE POWER TO KNOW.

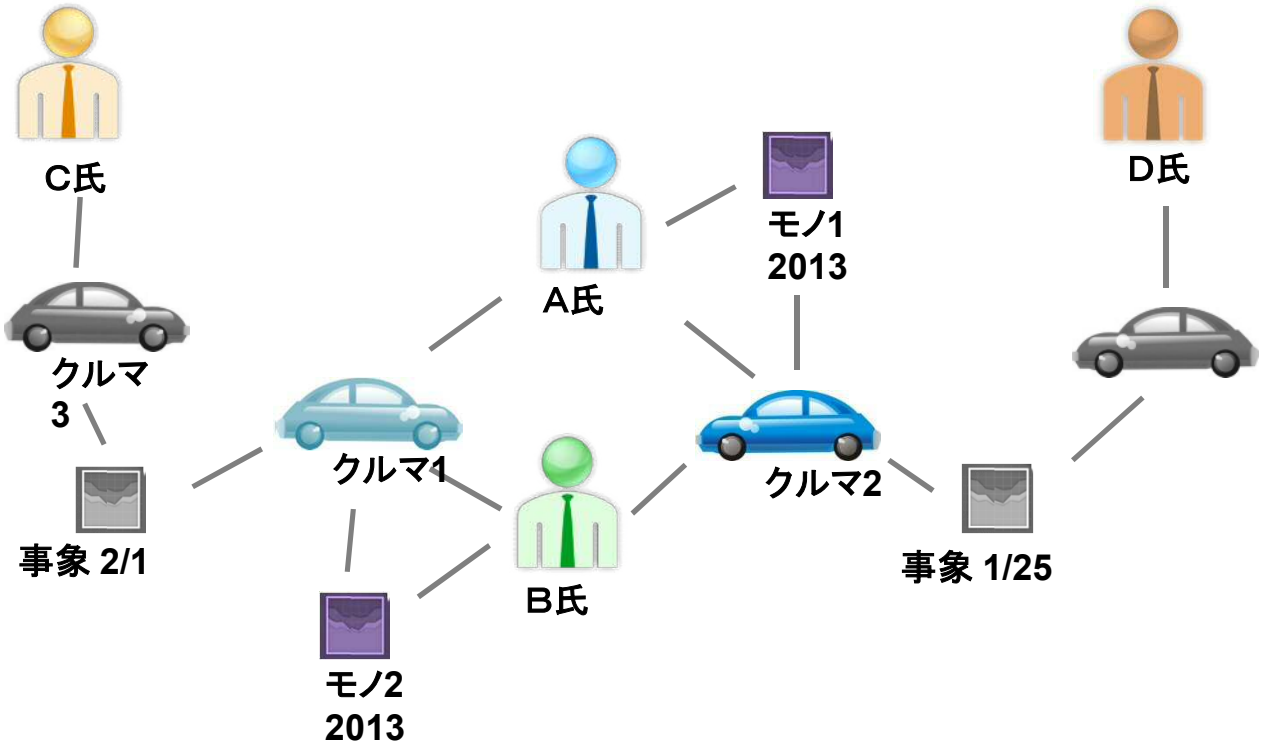
カテゴリー類推

例．マーケティング
「地域Ａに住む映画好きの男性一覧」

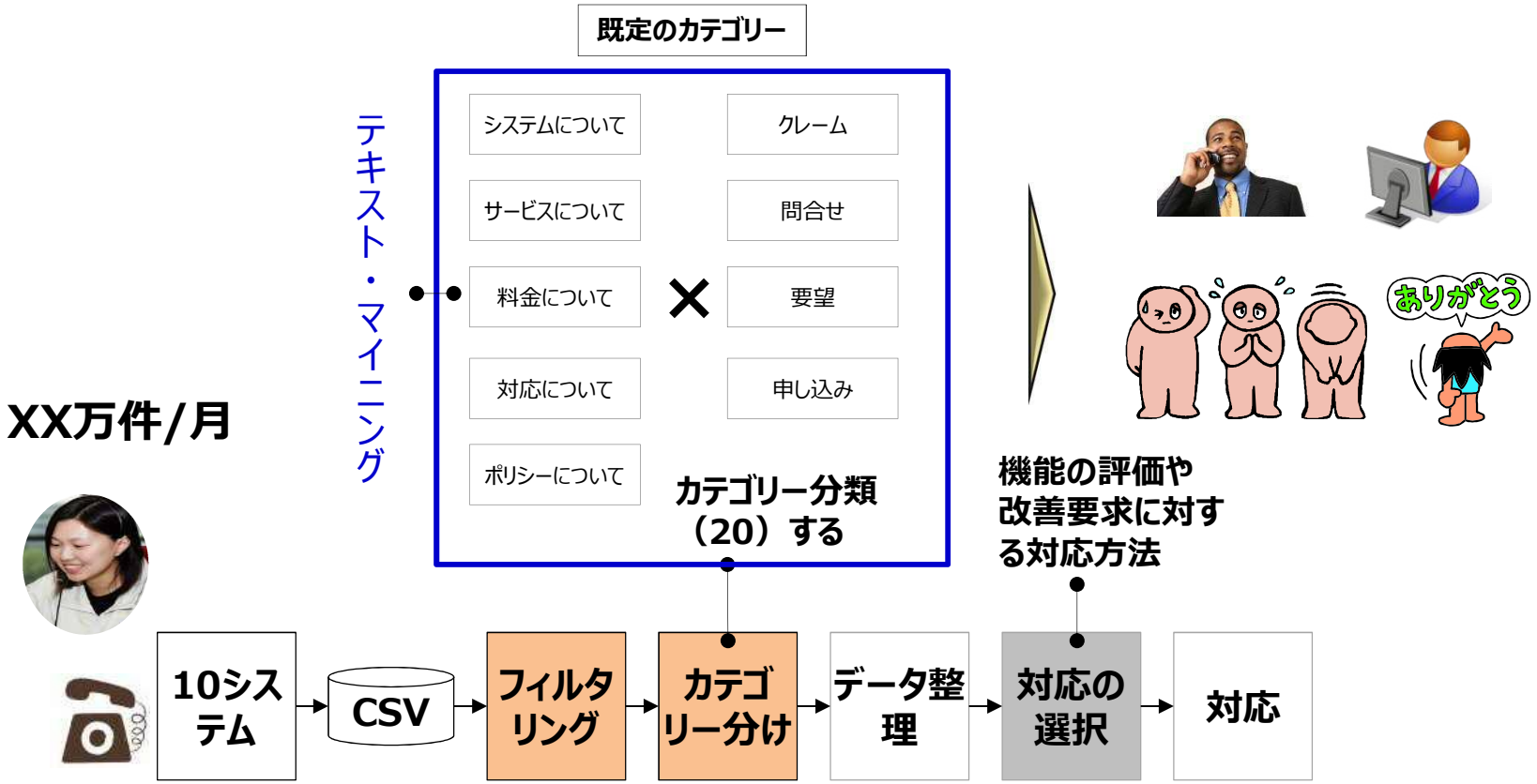
事例～公共

- エンティティ- 人物 / 時間/ 場所 / 行動 / モノ（車両）
- 例. A氏が2014年4月9日B氏所有のクルマに乗りXYZショッピング・モールで薬品を買った





事例～公共


既定のカテゴリー
システムについて
クレーム
サービスについて
問合せ
料金について
×
要望
対応について
申し込み
ポリシーについて
テキスト・マイニング
カテゴリー分類（20）する
XX万件/月
機能の評価や改善要求に対する対応方法
ありがとう
10システム
CSV
フィルタリング
カテゴリー分け
データ整理
対応の選択
対応

事例まとめ

業種	会社名	業務課題	SASによる解決
製造業	A社（国内）	テキストデータを用いた顧客へのリコメンドを実施していたが、データ量や海外展開などを考慮すると生産性を大きく向上する必要があった	SAS® Text MinerによりGUIベースのアナリティクスを可能にすることで生産性を大きく向上させることができた。将来の海外展開もより容易な点を評価。
製造業	B社（国内）	新規顧客の顧客像を明確に描くことができず、マスに対する画一的な訴求しかできていなかった	SAS® Enterprise Miner とText Minerによりインターネット上の消費者の感想と受注データの相関分析をし、新規顧客のプロファイルを浮かび上がらせ、施策に結びつける
製造業	C社（国内）	製品品質のさらなる向上のために社内データ・インターネットデータを活用したい	SAS® Enterprise Miner とText Miner, SAS® Enterprise Content categorizationにより製品品質と社内データ・SNSデータの相関を導き、よりプロアクティブな活動を容易にした
金融業	D社（国内）	マーケティングでアナリティクスを活用しているが、さらに詳細・幅広い顧客の理解やマーケティングの施策を実施したい	顧客の取引履歴のテキストデータをEnterprise Content Categorization とText Minerで分析して顧客の興味やセグメントをより深く・広く理解、マーケティングなどの施策に活用
金融業	E社（国内）	コールセンターのお客様との膨大な会話ログをテキスト化して、ビッグデータ活用につなげたい	テキスト化された音声データを大量データストアに格納した上で、顧客の要望や興味を分析し、営業的顧客対応やマーケティング等に活用する
パブリックセキュリティ	F 国家警察（海外）	事件の関係者・場所・時間・もの（クルマ・武器・薬物等々）とそれらの関係性を取り調べ記録からマップとして抽出することに多大な時間とワークがかかっていた	SAS® Enterprise Content Categorizationにより取り調べの電子データから左記のエンティティ（関係者・場所・等々）を自動抽出し、関係者マップを描き出すことを可能に。また過去の類似事件の関係者一覧等。