NTT DATA Mathematical Systems Inc.

アジア特許情報研究会 設立10周年記念講演会

テキストマイニングと深層学習を用いた英文技術文書の分析ソリューション

2018年11月30日

株式会社 NTTデータ数理システム
岩本 圭介

Trusted Global Innovator
NTT DATA Group

目次

1. NTTデータ数理システムのご紹介

- 会社紹介
- ソリューション紹介

2. 英文技術文書の分析

- テキストマイニングツール **Text Mining Studio**
- 英語解析における当社の取り組み と 分析例

3. データマイニング～深層学習の利用

- データマイニングツール **Visual Mining Studio**
  深層学習デザインツール **Deep Learner**
- 深層学習を用いての文書分類モデル構築

NTTデータ数理システムのご紹介

- 会社名： 株式会社NTTデータ数理システム
- 資本金： 5,600万円　（NTTデータ100%出資）
- 所在地： 東京都新宿区信濃町35 信濃町煉瓦館1階
- 従業員数： 約100名（80%が技術者）
- 沿革：　1982年4月　㈱数理システム設立
  2012年2月　NTTデータグループ入り
  2013年9月　㈱NTTデータ数理システムに社名変更

数理科学とコンピュータサイエンスを軸にして、社会のあらゆる分野に起こる問題解決のためのソリューションを提供する専門家集団です

NTTデータ数理システムのご紹介

■主な業務内容
・ビジネス・アナリティクス領域における、パッケージソフトウェアの開発・販売
・アプリケーション開発
・分析コンサルティング事業

■開発・分析対象領域
・AI、機械学習、深層学習
・数理計画、最適化
・統計解析、マイニング
・知識データベース、言語処理、パターン認識
・シミュレーションなどの科学計算

NTTデータ数理システムの強み

- パッケージ製品・受託コンサルティング導入実績数2,000社以上
- 数理科学のプロ集団としての豊富な経験・取引実績
- 自社パッケージによる開発・チューニング
- NTTデータグループとの事業連携の強化による包括的なサービス提供

主なパッケージ製品

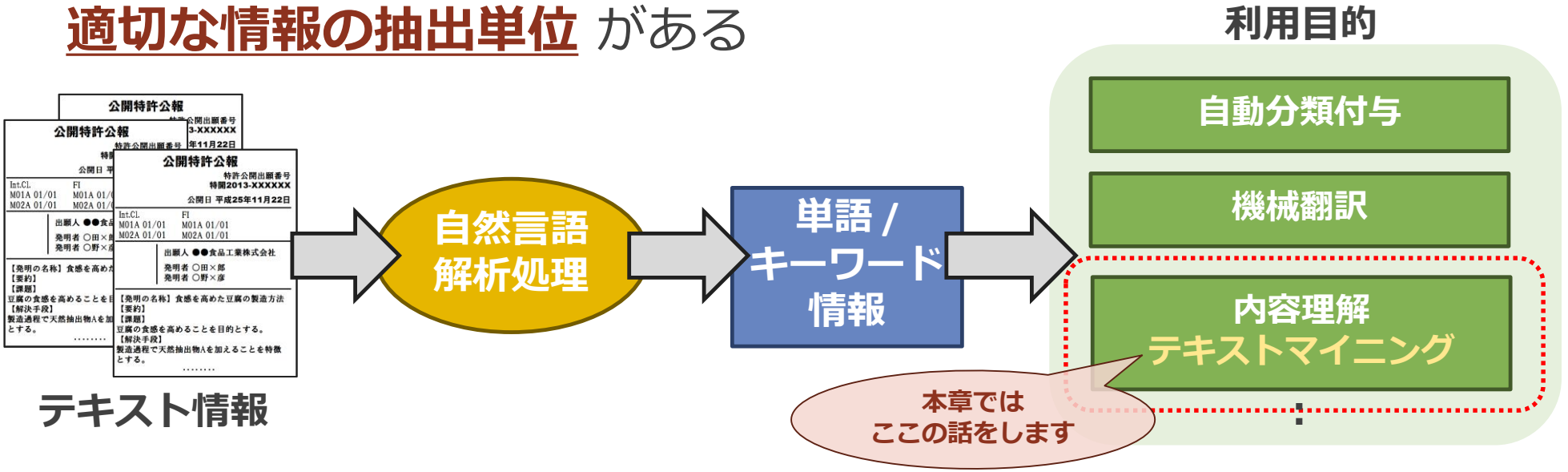
- データマイニング　Visual Mining Studio（需要予測・傾向分析・クラスタリング）
  Big Data Module（大規模データ分析・Hadoop）
- テキストマイニング　Text Mining Studio（ポジネガ分析・特徴分析・話題分析）
- 深層学習 デザインツール Deep Learner （ディープラーニング）
- ベイジアンネットワーク　BayoLinkS　（医療・故障診断）
- 最適化　Numerical Optimizer(スケジューリング・組み合わせ最適化)
- シミュレーション S^4 Simulation System（離散イベント・連続系・エージェント）
- 統計解析　S-PLUS　※　（回帰・検定・多変量解析）
- 統計解析　Visual R Platform　（Rユーザー向け分析プラットフォーム）
- CRM分析　CRM Insight　（購買傾向、ターゲティング）
- 特許分析　Patent Mining eXpress　（特許情報分析）
- 超高速シミュレーションMonaco
- 半導体TCADシミュレータ　Paradise World II
- 金融工学　Fiopt　（ポートフォリオ最適化・シナリオ発生）

※以外はすべて自社開発です

2. 英文技術文書の分析

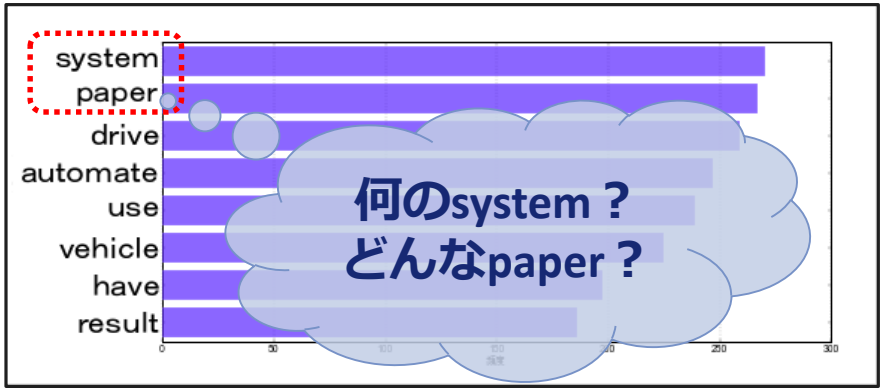
テキストマイニングに求められる情報抽出

- テキストの分析には、**利用目的**に応じた**適切な情報の抽出単位**がある



- 殊に、内容理解を目的とした**テキストマイニング**においては**ユーザーが直感的に理解できる情報の単位**であることが必要

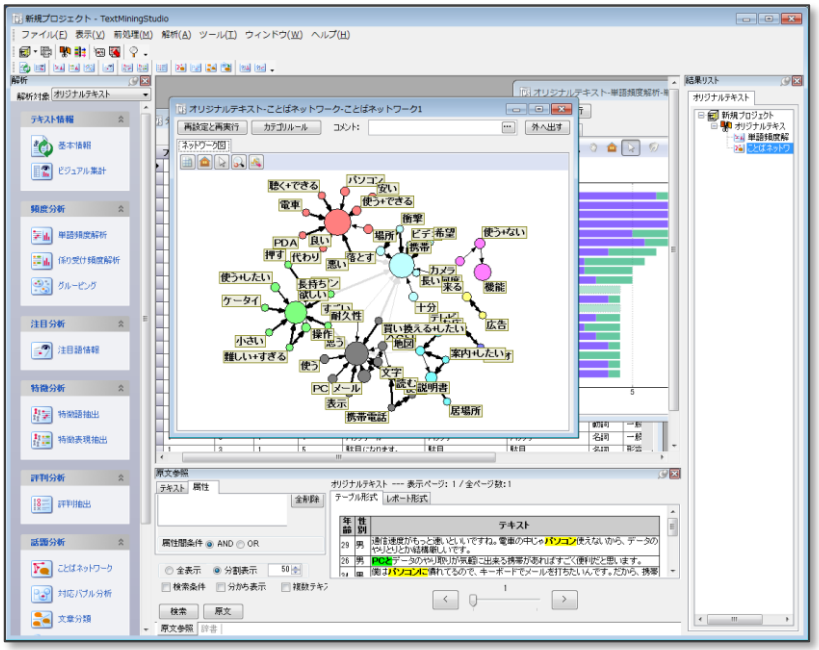
単語そのままをアウトプットにすると‥



Text Mining Studio

テキストデータから有益な情報を抽出するための
テキストマイニングツール

- **誰にでも**高度な
  テキストマイニングを
  - マウスの操作のみで分析が可能
  - 豊富な分析機能 と 強力なグラフ機能
- より**自由度**の高い分析を
  - カテゴリ機能等 「意味」 に着目した分析
  - データマイニングツール **Visual Mining Studio** とのシームレスな連係
  - アドオンツール **英語アドオン** の導入で
    英文の分析も可能



Text Mining Studio（TMS）

日本語解析における当社の取り組み

文章「従来手法よりも検出精度を向上させることができた。」

見出し語	原形
従来	従来
の	の
手法	手法
より	より
も	も
検出	検出
精度	精度
を	を
向上	向上
さ	する
せる	せる
こと	こと
が	が
でき	できる
た	た
。	。

一次処理である
形態素解析適用時の結果

自動連結機能

文節の単位をとらえ「**欲しい単語**」を適切に抽出する
- 連語の自動判定
- 文節内からキーワード相当部分のみを抽出

単語ID	見出し語	原形	品詞	係り先	態度表現
1	従来の	従来	名詞	2	なし
2	手法よりも	手法	名詞	4	なし
3	検出精度を	検出精度	名詞	4	なし
4	向上させることができた。	向上	名詞	-1	可能

態度表現機能

テキスト記述者の「**態度**」を適切に抽出する
- 否定、要望、疑問、可能/不可能、容易/困難 など
  9種類 の 態度表現ラベル を文節に自動的に付与

英語解析における当社の取り組み

- この「**欲しい単語**」を適切に抽出する機能を、
  英語 の 自然言語解析エンジン にも組み込み

文章 ”The system should be able to predict traffic jam.”

単語ID	見出し語	原形	品詞	係り先
1	The system	system	名詞	2
2	should be able	should able	形容詞	-1
3	to predict	predict	動詞	2
4	traffic jam	traffic jam	名詞	3
5	.	.	記号	2

自動連結機能

単語がばらばらにならず、
日本語の「文節」と類似した感覚で単語をまとめ上げる。

その結果、ユーザーが辞書を整備することなく
“traffic jam” といった連語が抽出可能、
またより自然な係り受け関係を抽出できる

テキストマイニングの
アウトプットとしては不要な
前置詞・冠詞 などは
自動で削除され、現れない。

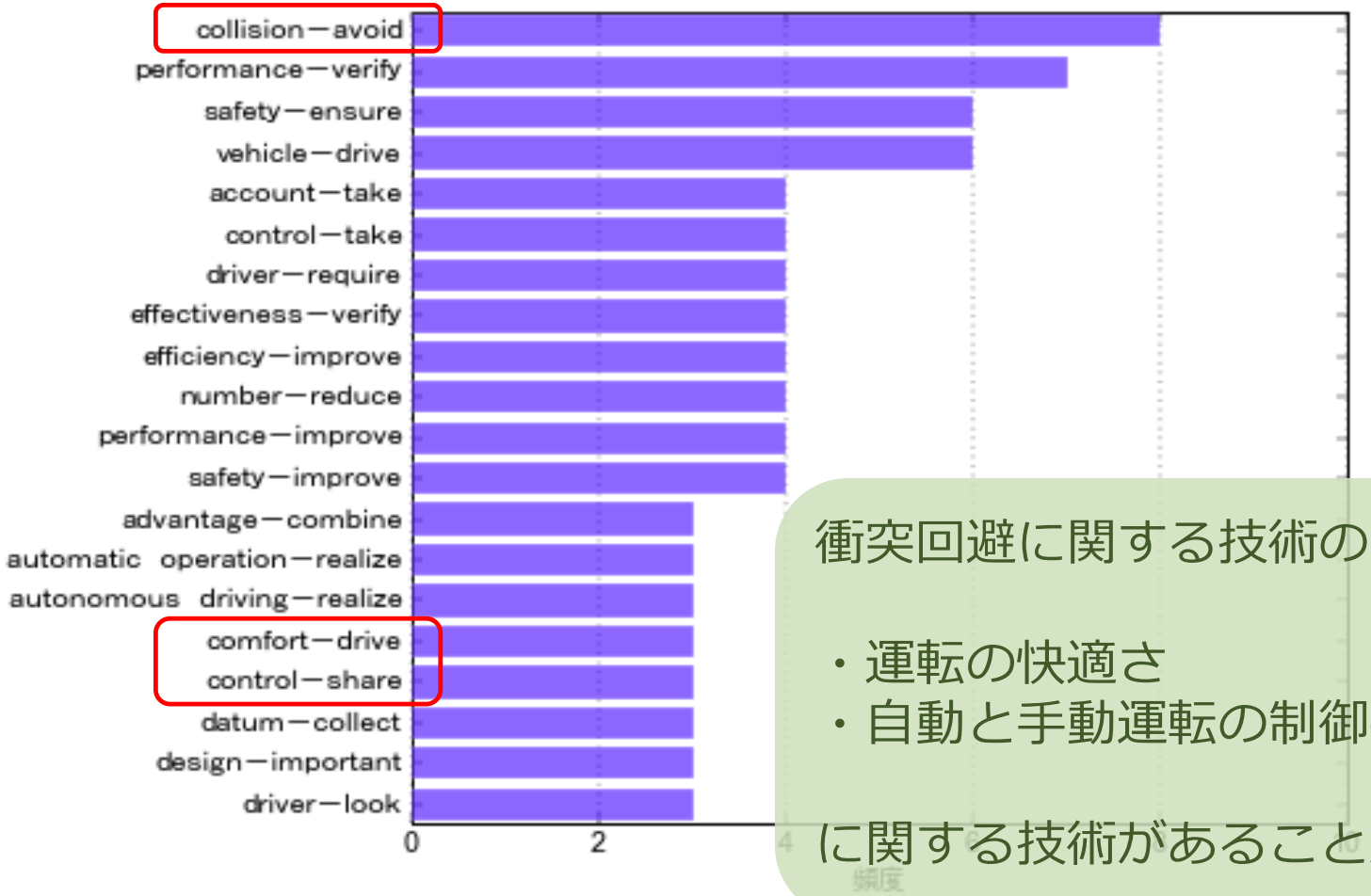
データクリーニングの手間が大幅に軽減

英語解析における当社の取り組み

- データから**連語の候補を自動的に抽出**する機能を備え、特別な技術分野依存の語彙にも対応可能
  - 連接語群の「**連語らしさ**」を統計的指標により評価

分析例：係り受け解析による技術傾向の把握

- 対象データ：「自動運転」に関する論文の英語抄録部分 438件

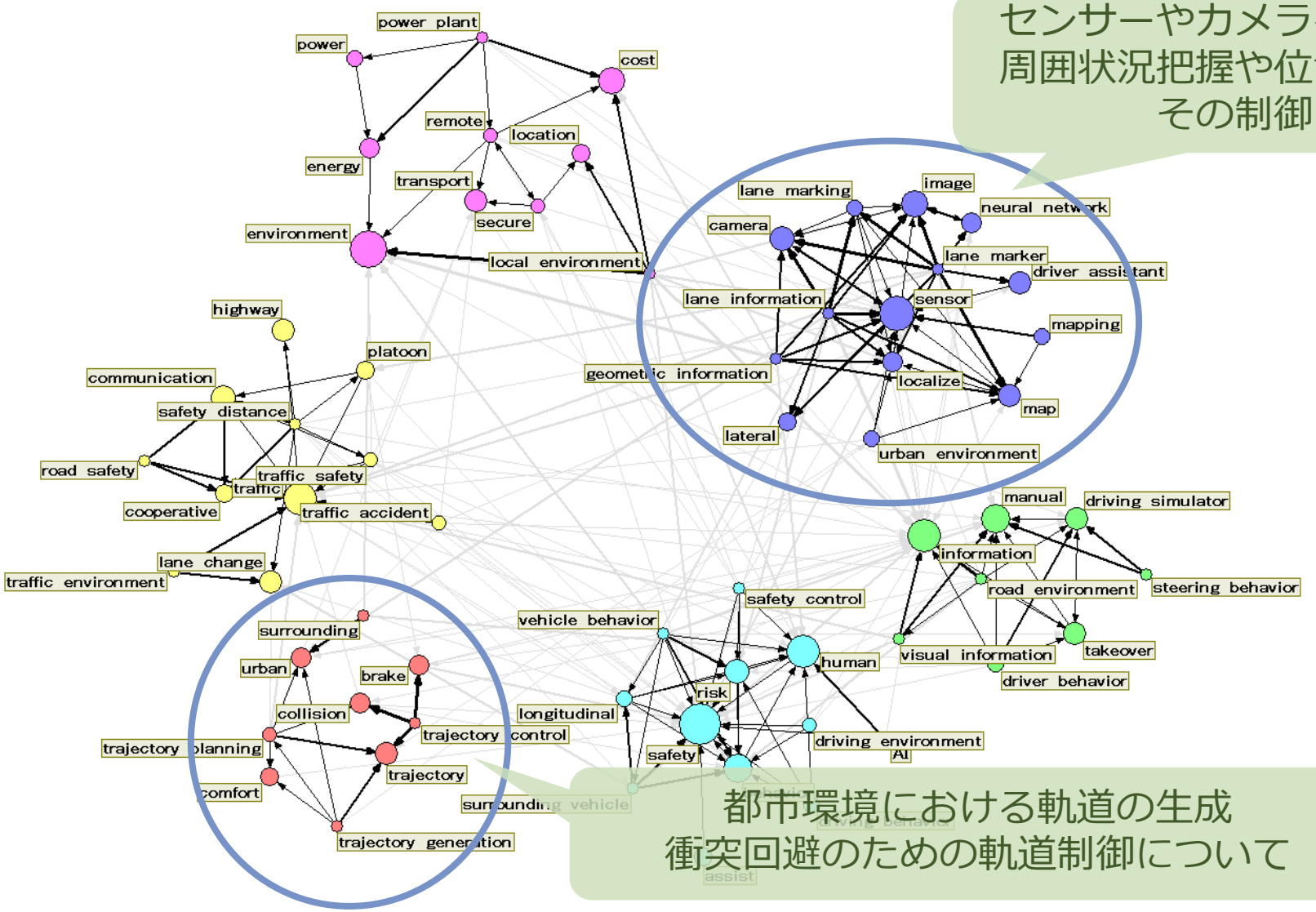


衝突回避に関する技術のほか

・運転の快適さ
・自動と手動運転の制御の割り当て

に関する技術があることがわかる

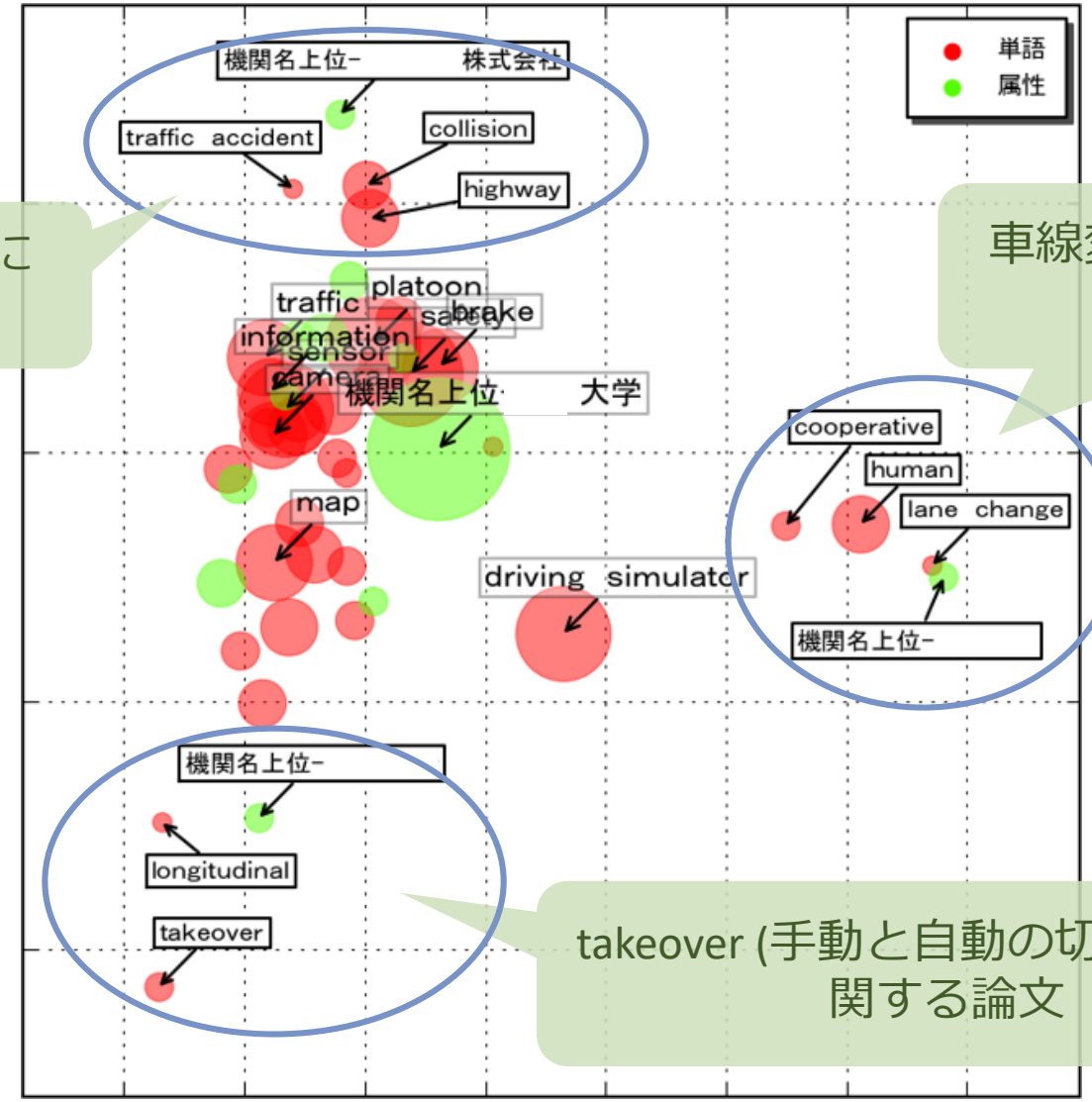
分析例：ネットワーク図による技術間の関係図示



センサーやカメラを用いた
周囲状況把握や位置確認と
その制御

都市環境における軌道の生成
衝突回避のための軌道制御について

分析例：文献発行機関と技術とのマッピング



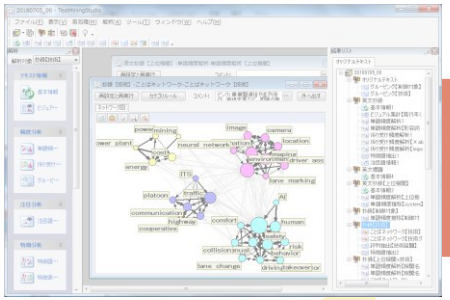
事故の際の技術に関する論文

車線変更や協調運転に関する論文

takeover (手動と自動の切り替え) に関する論文

3. データマイニング～深層学習の利用

テキストマイニング×データマイニング でできること



自然言語処理
内容理解・把握
ルールベース分類
（グルーピング）

文書の教師あり分類

- 機械学習による
  分類モデル構築・ラベル付与

この論文は
どういう分類？

文書

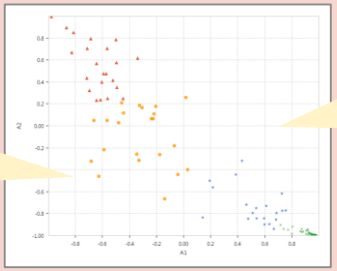
モデル

技術分類は
○○です

文書を数値情報化して利用

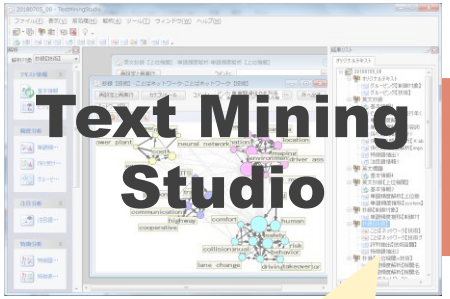
– クラスタリング・可視化

どんな文書が
近い文書？



どんな分布、
どんなかたまりを
なしている？

テキストマイニング×データマイニング でできること



自然言語処理
内容理解・把握
ルールベース分類
（グルーピング）

データマイニングツール
Visual Mining Studio

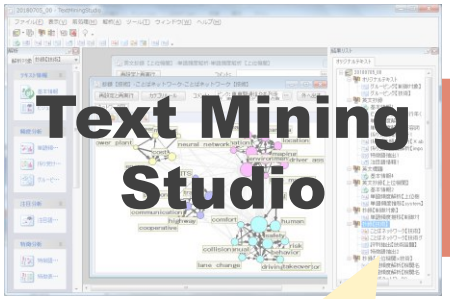
＋

深層学習デザインツール
Deep Learner

※Deep Learner は、当社製品
- Visual Mining Studio
- Visual R Platform
- Big Data Module
- BayoLinkS

のアドオンです。
上記4製品のいずれかが必要です。

テキストマイニング×データマイニング でできること



自然言語処理
内容理解・把握
ルールベース分類
（グルーピング）

	Visual Mining Studio	Deep Learner
機械学習による分類手法	**実用的な手法を多数搭載** 決定木 Support Vector Machine Neural Network（3層） ….等々	**より複雑な Neural Network モデル** 多層パーセプトロン **テキストの扱いに適したモデル** RNN, LSTM
テキストの数値情報化(ベクトル化)及び活用手法	**文書ベクトルの作成次元圧縮** 主成分分析 対応分析　…等々 **クラスタリング** 階層型クラスタリング k-Means　…等々	**文書をベクトル化する Neural Network モデル** RNN + AutoEncoder
可視化手法	**共通分析基盤 Visual Analytics Platform** が備える強力な可視化機能を利用可能	

テキストを扱う上での **Deep Learner** の利点

文書の教師あり分類

- 従来手法では 単語の有無 で文書を特徴付ける
  Bag of Words という考えを用いることが多いが
  一般にこれは語順・文脈の情報が反映されない

区分	TCP－IP	ポート	インストール	停止	Win. 98	XP	CD	カスタム	用紙サイズ	設定	…
インストール	1	1	1	1	0	0	0	0	0	0	
インストール	0	0	1	0	1	1	1	0	0	0	
設定	0	0	0	0	1	0	1	1	1	1	

- RNN では可変長の系列情報を自然に扱える
  - 含まれる単語数は各文書で異なるため、テキストは**可変長情報**
  - さらに、その中の**単語出現の順序も考慮**される

文書を数値情報化して利用

– 語順・文脈を考慮して文書を**数値（ベクトル）化**できる
  - ベクトル化することで、
    互いがどれだけ似ているかを定量的に評価できる

Deep Learner が前提とするデータ

テキスト	分類
視線情報を用いた自動運転制御について	注目
プラント自動運転のための油圧自動制御方式	非注目
協調運転のための視覚情報検出技術	注目

行ID	単語ID	単語	分類
1	1	視線情報	注目
1	2	用いる	注目
1	…	…	注目
2	1	プラント	非注目
2	2	自動運転	非注目
2	…		非注目

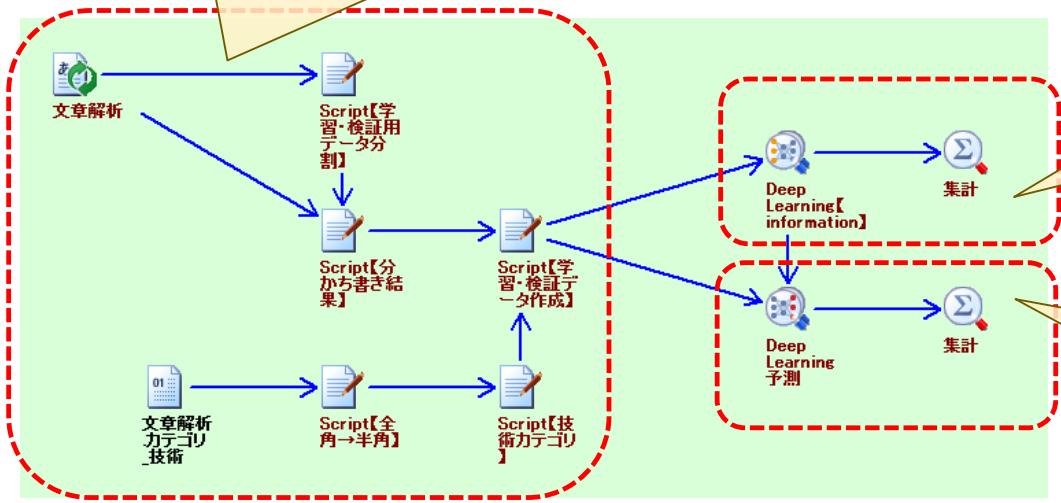
キー列：「１件のデータ」として扱う単位

説明変数列：予測の根拠とする情報

目的変数列：予測対象の情報

テキストマイニング×深層学習 連係のイメージ

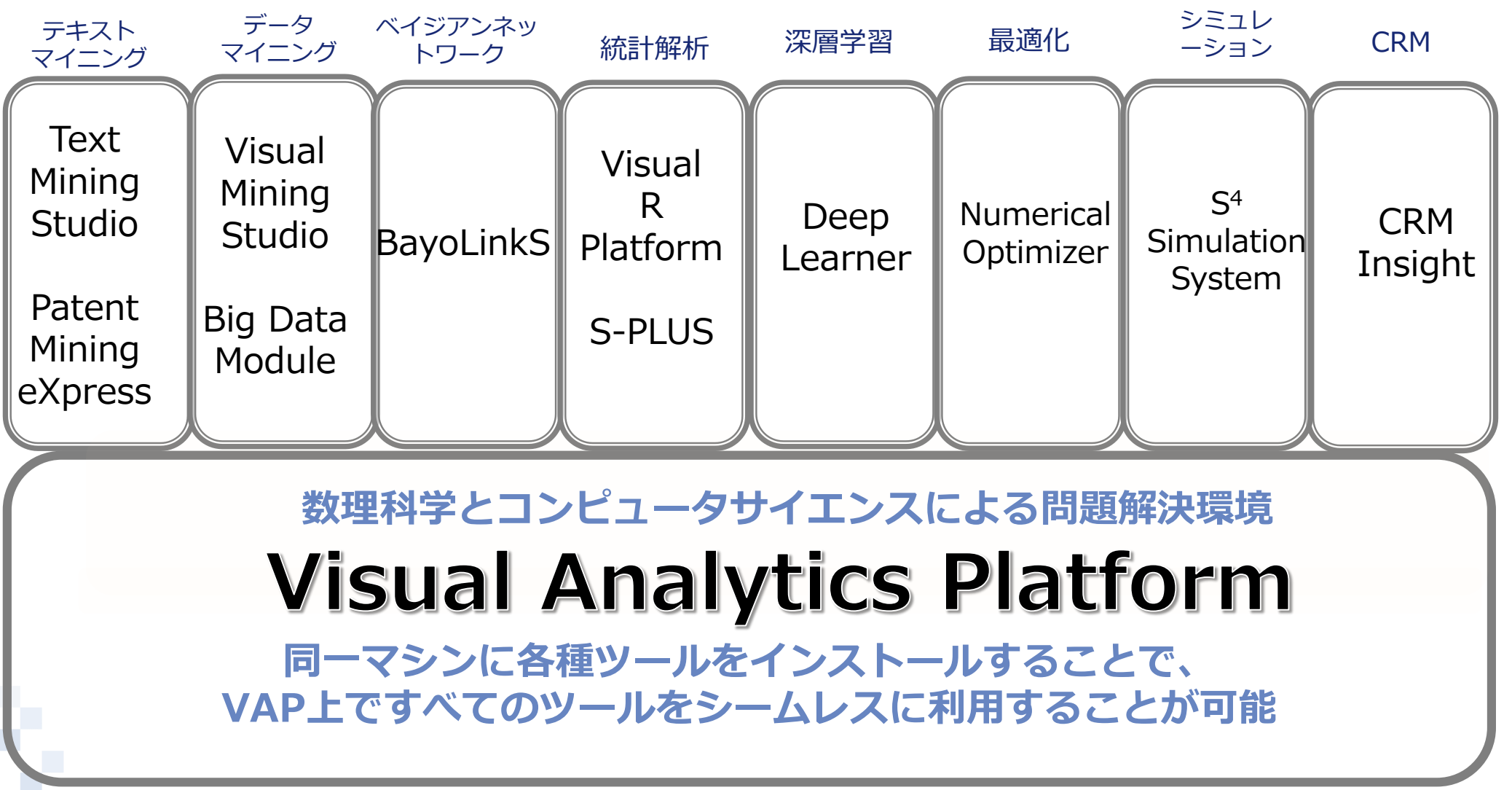
前処理
- **TMS出力結果の分かち書き結果と属性情報をマージし Deep Learner への入力データを作成**
- **学習データ と 検証データ を分割**



学習データ を用いて予測モデル作成

検証データ を用いて予測モデル評価

業界初！シームレス連係



数理科学とコンピュータサイエンスによる問題解決環境

Visual Analytics Platform

同一マシンに各種ツールをインストールすることで、
VAP上ですべてのツールをシームレスに利用することが可能

数理システムのソフトウェア　充実サポート

導入前から導入後まで、完全サポート！

導入前　無料サポート

体験セミナー
基本的な操作方法を
実習を通して理解

分析個別相談会
体験セミナー後、
分析の進め方を
技術コンサルタントと相談

テスト使用
（30日間）
安心の
操作サポートつき
（メール対応）

分析個別
コンサルティング
（1h）
VMSの疑問点や、
今後の分析について
を相談

導入後　保守サポート

スキルアップセミナー
一歩進んだ分析活用法を体得
※開催有無は各製品に
よって異なります

分析個別相談会
スキルアップセミナー後
日頃のお悩みを
技術コンサルタントに相談

いつでも操作サポート（メール対応）

ユーザー様のご予算・スケジュールにあわせた コンサルティング・教育支援の充実！

導入後　各種サポート

マンツーマンで
一緒に分析を
してほしい
⇒ **スポット**
コンサルティング

複数人向けに
分析スキルがつく
講習をしてほしい
⇒ **教育支援**

自分で分析を進める
上で、定期的なアド
バイスや作業援助を
してほしい
⇒ **定期個別サポート**

分析作業を任せて作
業結果を出してほし
い
⇒ **分析受託**

NTT DATA

NTT DATA Mathematical Systems Inc.

Trusted Global Innovator

NTT DATA Group

NTT DATA