テキスト分類AIモデルの学習データ構築

RICOH
imagine. change.

株式会社リコー　デジタル戦略部
言語AI開発室
データ開発グループ
佐藤奈穂子

背景と目的

RICOH
imagine. change.

RICOH DIGITAL PROCESSING SERVICE
仕事のAI

2021年 「仕事のAI」シリーズを上梓

2023年　新サービス「ノーコード開発ツール」
（テキスト分類AI）

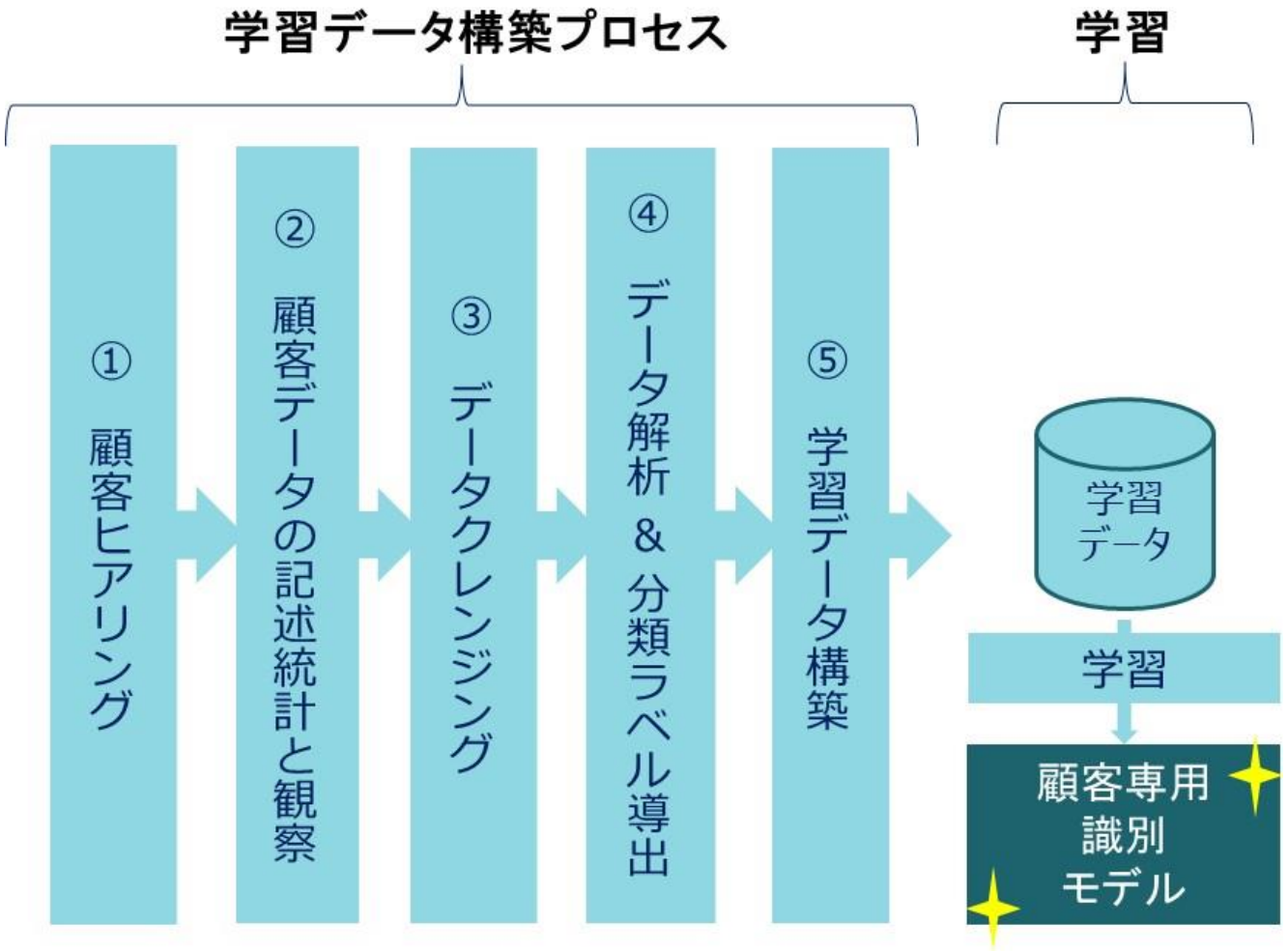
トライアルユーザーの保有データは様々

お客様にとって価値ある
分類AIをつくるために学習
データをどう作ればよいか？

目的　テキスト分類AI向け学習データ構築プロセス確立

実運用可能なプロセスを構築

RICOH
imagine. change.



従来、顧客データの一次解析～分類ラベル提案は担当者の力量に依っていた。

が！

このプロセス化、および顧客データの形式別に各プロセスでやるべきことを記載した定義書の作成により、初心者でも一定のレベル（品質）で捌けるようになった。

各プロセスにおける課題

RICOH
imagine. change.

① 顧客ヒアリング（自社データでやりたいこと確認）

内容把握していない
何かできないか？

② 顧客データの記述統計と観察

多様なデータ形式
申告？回答？対話？日報？
アンケート回答？長文？短文？
既存分類有り？

③ データクレンジング

データ毎にクレンジング内容
が異なる

④ データ解析結果からの分類ラベル導出

専門用語
ストップワード

⑤ 分類のための学習データ構築

カテゴリ定義
アノテーションルール
例外対応
データ品質チェック

プロセス②

顧客データの観察例

RICOH
imagine. change.

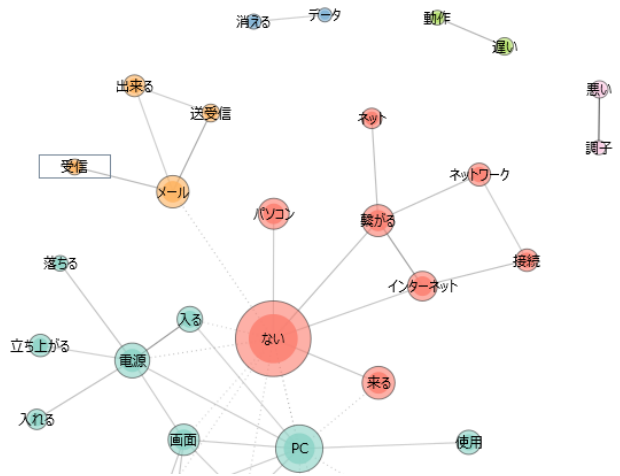
顧客データの記述統計と観察によるプロファイリングをして、次行程でどのようなクレンジングが必要かを判断する

顧客	A社	B社	C社	D社	E社	F社	G社	H社	I社	J社
データ内容	問合せVOC アンケート自由記述のテキスト	コールセンターのVOC オペレータの回答テキスト	コールセンターのVOC 企業サイトへの問い合わせテキスト 監視レポートテキスト	アンケート自由記述のテキスト	自社への提言テキスト	企業動向記事テキスト	コールセンターのVOC 電話音声の書き起こし、オペレータの回答無し	問合せ（電話、メール、Web）テキスト	意識調査回答テキスト	日報テキスト
1レコードのテキスト長	長文/短文混じり	長文	長文/短文混じり 監視レポートは短文 問い合わせは長文	長文	100文字～150文	長文（300文字前後、複数文）	長文	長文	長文	長文/短文混じり
データ形式	顧客の申告テキストとアンケート自由記述テキストが混在	顧客の申告とオペレータの回答の対話 複数日間にわたる対話記録あり	監視レポートの定型表現と問い合わせの自由記述テキストが混在 オペレータの回答も混在	アンケート自由記述テキストのみ	自由記述テキスト	自由記述テキスト	顧客の申告テキストのみ	顧客の申告テキストのみ	自由記述テキスト	活動報告テキストとその他情報が混在
独自分類有無	独自分類有り 商品情報(3)/性別(2)	独自分類有り（本文中への埋め込み情報は無し） 大分類（４） 中分類（10） 小分類（多数）	独自分類体系はなく、オペレータが判断して本文先頭に【】()で独自ラベルを付与(入っていないレコードもある)	独自分類有り(2)	独自分類有り(5)	独自分類有り(1)	独自分類有り(2)	独自分類有り(6)	無し	独自分類有り(3)

プロセス④ データ解析～分類ラベル導出例

RICOH imagine. change.

クレンジングで分割・カテゴライズしたデータ群に対し、単語統計量による話題傾向、文脈傾向を確認し、分類ラベル案を導出する



KH Coder出力例

パソコンの不具合事例を分類できそう！

PCの電源が入らない/立ち上がらない/落ちる
ネット/インターネット/ネットワークに繋がらない
メール送受信ができない
データが消える
動作が遅い
調子が悪い

分類ラベル案

- 電源不具合
- ネットワーク接続
- データ消失
- その他
- 動作遅延
- メール送受信

RICOH

imagine. change.