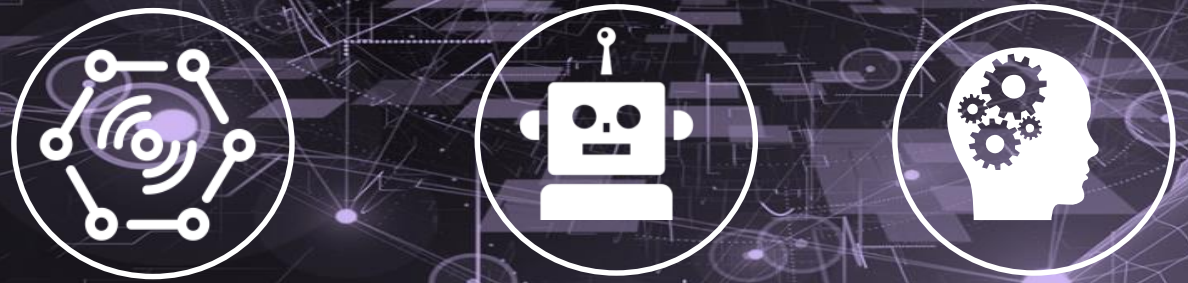
第３回 AI/Analyticsカンファレンス

開発者が語る！
AIによる自然言語処理（NLP） 5つのポイント



お客様の価値観を共有するパートナー
Value Engagement Partner
株式会社システム情報

目次 Contents

１．これだけは押さえておきたい！
自然言語処理の基礎知識

01 自然言語処理の基礎知識

自然言語

人が日常的に話し聞き、読み書きしている言語

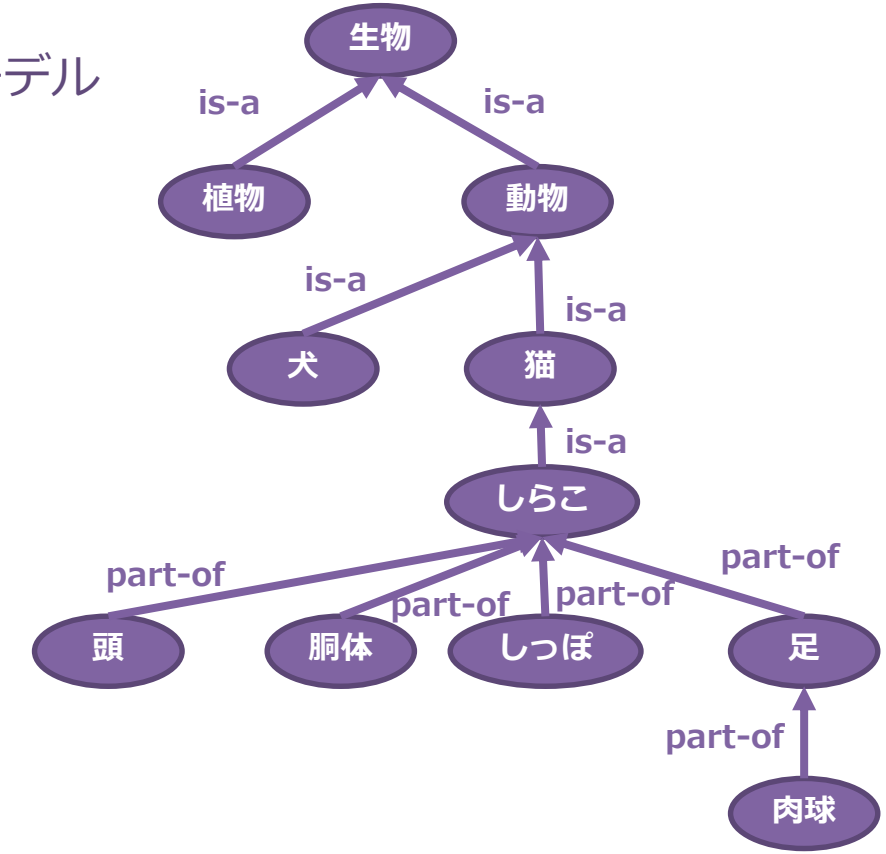
自然言語処理

自然言語をコンピュータ上で扱う技術

01 自然言語処理の基礎知識

オントロジー

情報の意味を定義するための概念・仕組み
個々の情報の繋がりを明らかにし、構造化し整理するモデル



☑ 世の中の様々な“モノごと”を
正確に整理できる

モノとモノの関係性(繋がり)、モノが持つ情報を表現できる

01 自然言語処理の基礎知識

コーパス

実際に使用されている例文を大量に集めたもの

テキストコーパス

文章を集めたもの
新聞記事、雑誌、小説、辞書etc..

音声コーパス

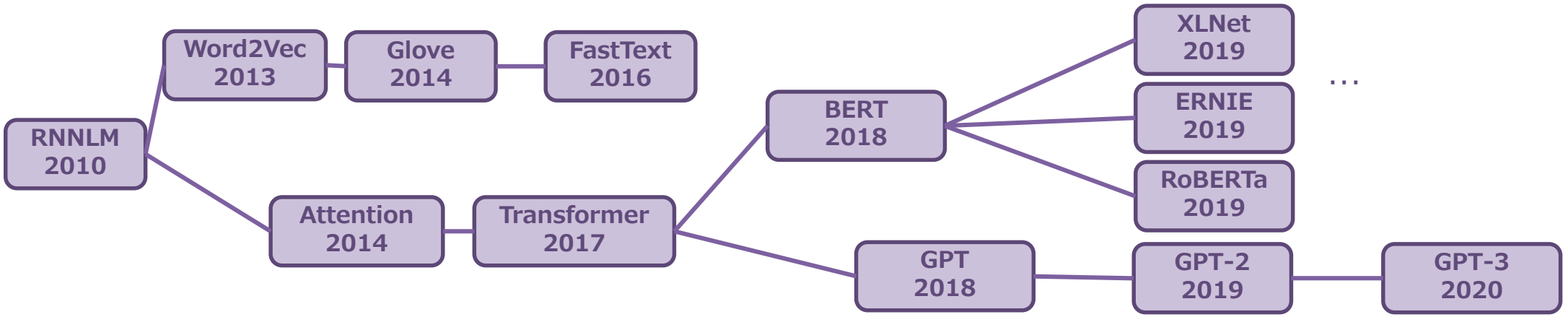
音声データを集めたもの
対話、インタビュー、講演etc..

自然言語処理の根幹！データの集まり

⇒コーパスの充実が成否を分けるポイント

基礎知識　深層学習を用いた手法

近年注目を浴びている、深層学習の手法を用いて解析・モデル作成を行う

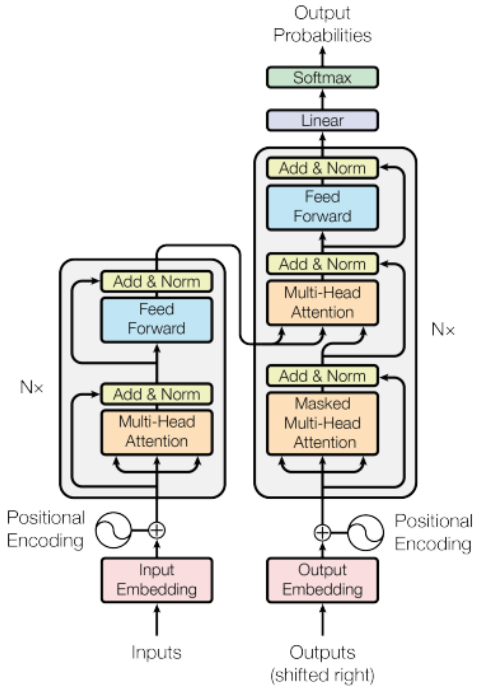


- 日々精度の高いTransformerベースのモデルが登場しているが、日本語で使用できるモデル(事前学習済みモデル)は限られる
  事前学習: 一般的なコーパス(Wikipedia)等であらかじめ学習しておく
- 現状活用できるのはBERT, RoBERTa, DistilBERT, ELECTRA, GPT2

01 自然言語処理の基礎知識

BERT

- BERT
  （Bidirectional Encoder Representation from Transformers）
  2018年10月にGoogleから発表された自然言語処理のための
  ディープラーニングモデル
- 日本語の事前学習モデルが多数公開されており、
  日本語での利用が最も盛んなモデル
- Googleの検索エンジン始め、世界的に普及が進んでいる



2．失敗しない自然言語技術の選び方

02 失敗しない自然言語技術の選び方

深層学習 VS統計的手法・機械学習

	深層学習	統計的手法・機械学習
目的	精度を求める	解釈性を求める

深層学習

自動翻訳、チャットボット、音声対話システムなど
何よりも精度を求めるソリューションに適している

統計的手法・機械学習

テキストマイニング、データ分析、テキスト分析など
入出力に対して解釈(相関関係)を求めるソリューションに適している

02 失敗しない自然言語技術の選び方

自然言語タスクによる選定

自然言語タスク		
文章分類	文章生成	FAQ
ナレッジ検索	キーワード検索	チャットボット
真偽判定	Text to data	品詞タグ付け
…		

- ☑ タスクを知ることで、自然言語処理の応用範囲への勘所を掴むことができる
- ☑ 目的→どのタスクに適用できるか→どの技術を使うか という流れ
- ☑ 一つのタスクに固執するのはなく、都度適したタスクを
  時には複数組み合わせて「柔軟に選定」することが重要

⇒プロに相談しながら決めると吉

３．自然言語処理はデータが命

03 自然言語処理はデータが命

データの種類

辞書
単語のリスト

追加
- 商品名
- 社名
- 略語
- 社内用語
- 類義語
- 専門用語

基本
- IPA辞書:IPA 品詞体系日本語辞書
- Unidic:国語研短単位自動解析用辞書

コーパス
文章のリスト

タグ無しコーパス	文章のみがまとめられたコーパス ・検索ドキュメントなど
タグ付きコーパス	文章に正解ラベル(タグ)が付与されたコーパス ・FAQデータなど
対訳コーパス	対になる文章で対訳の形でまとめられたコーパス ・翻訳データなど

03 自然言語処理はデータが命

- ☑ 辞書にない単語は正しく解析できず答えられない
- ☑ コーパスに無い文章は探しても出てこず、答えられない

⇒データの充実が成否に“直結”

03 自然言語処理はデータが命

データ収集・作成する上でのポイント

- 自然言語は、「あいまいさ」があることが特徴。
  利用されるポイントに沿った学習データを作成することで、
  不要な「あいまいさ」を排除し品質の良いデータを作成することが重要
- 想定する利用シーンの明確化されているか
  - ・社内向け？コンシューマ向け？
  - ・利用者の知識レベルは？
- 学習データは利用シーンに沿っているか
  - ・入力データ(質問、検索文言etc.)は口語体か？文語体か？
  - ・専門用語での利用か？一般的な用語での利用か？
- 単語辞書(製品名・略語・社内用語etc.)は準備できているか

４．忘れてはいけない！前処理の重要性

04 忘れてはいけない前処理の重要性

自然言語処理 ４つの前処理

01
テキストのクリーニング

テキスト内に含まれるノイズを除去
- htmlタグ等のテキスト属性を除去
- 無駄な行やスペースの削除
- 特殊文字、絵文字，顔文字，ハッシュタグ等の除去

02
単語の分割

辞書データを元に単語を分割
- 形態素解析器による単語分割

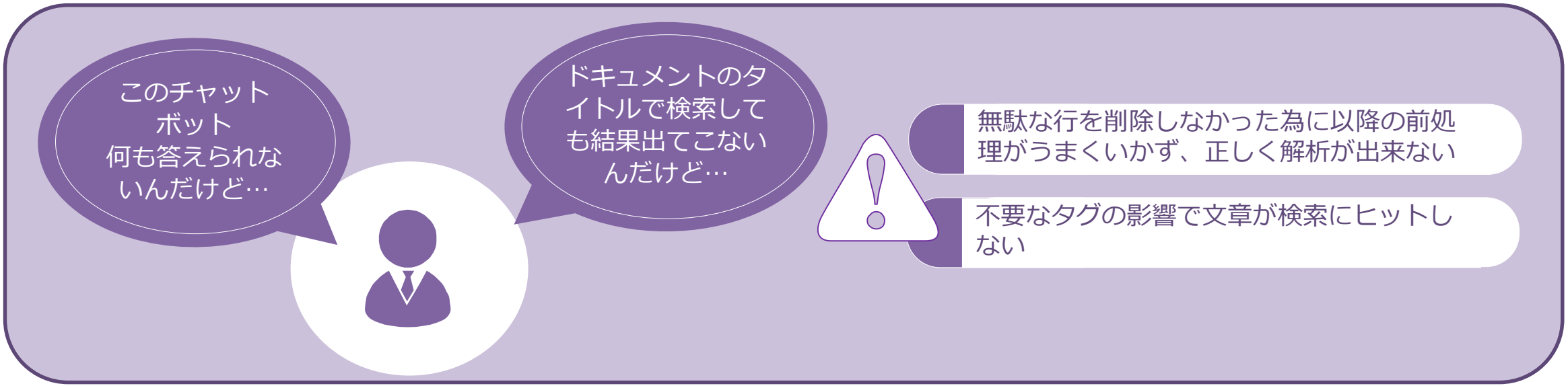
03
単語の正規化、
ストップワードの除去

単語の文字種の統一、つづりや表記揺れの吸収
- 用語（同義語）の統一
- 「です」「ます」の除去

04
単語のベクトル化

単語を機械が理解できる形に変換

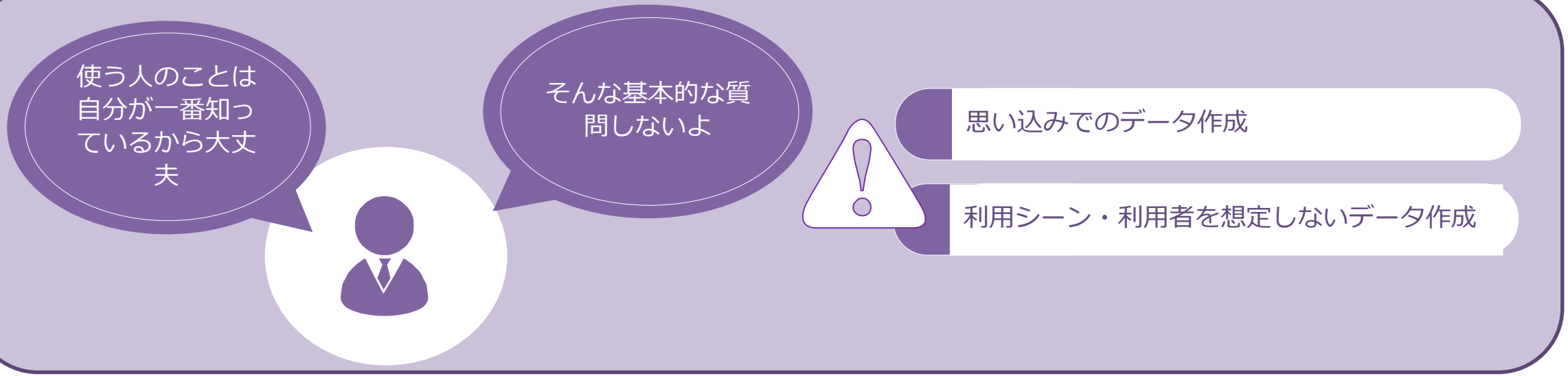
04 忘れてはいけない前処理の重要性



- ☑ 集めた辞書もコーパスも正しく前処理できなければ結果は出ない
- ☑ 作成時点では高い精度、実際に利用すると低い精度となる要因の一つに前処理の不備によることも多い

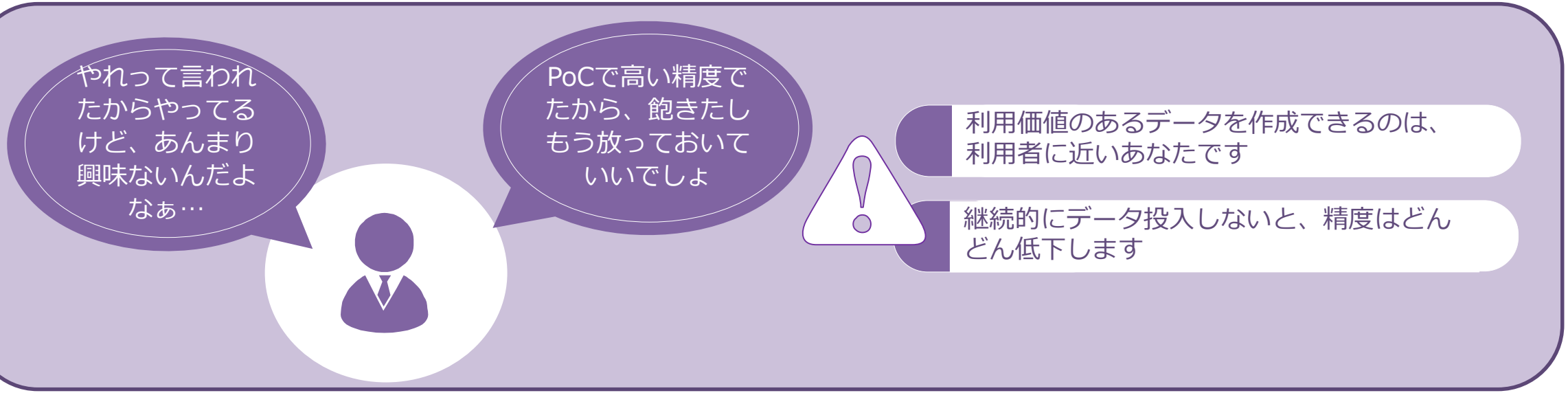
５．成否を分けるのは、最後は人

05 成否を分ける、最後は人



- ☑ ターゲットとなるペルソナは定義できているか
- ☑ 自然言語は「あいまい」
  自然言語処理は「あいまい」なデータを活用する技術
  ・誰が
  ・どこで
  ・何を
  を明確にし、そのペルソナが満足するモデルを目指すことが、成功への近道

05 成否を分ける、最後は人



- ☑ 興味を持ち取り組める人を巻き込む
  発注側もチャレンジ精神を持った「人」がいることが重要

- ☑ 自然言語処理の学習は専門知識を持った「人」のコピーを育てること、
  放っておくと知らないことが増えていきます
  ・学習データの品質は「人」に依存
  ・モデルを育てるのは「人」
  ・モデルを成長させていくのは「人」

06 最後に・・・

システム情報では自然言語処理に関して
アセスメント～PoC～サービスインまで
トータルでご支援させていただいております

お困りのことがありましたらご相談ください

https://www.sysj.co.jp/solution/service/ai-iot-utilization-support

COMPANY
企業情報

●社　　名	**株式会社システム情報** 英字名 SYSTEM INFORMATION CO.,LTD.
●本　　社	東京都中央区勝どき1-7-3 勝どきサンスクエア7F
●T　E　L	03-5547-5700（代表）
●F　A　X	03-5547-5820
●U　R　L	https://www.sysj.co.jp
●創　　業	1980年1月12日

都営地下鉄大江戸線「勝どき駅」A1出口 徒歩0分（駅直結）

●役　　員					
	代表取締役社長	鈴木　隆司		取締役（監査等委員）	師橋　卓久
	取締役　上席執行役員	河野　逸人		取締役（監査等委員）	鷲﨑　弘宜
	取締役　上席執行役員	増田　航太		取締役（監査等委員）	足立　伸男
	上席執行役員	石川　勝雄		取締役（監査等委員）	山内　玲
	執行役員	梅原　隆		フェロー	小林　浩
	執行役員	新谷　忍			
	執行役員	平出　浩太郎			
	執行役員	髙橋　陽二郎			

●資　本　金	5億02百万円
●年　　商	[連結] 127億71百万円（2020年9月）
●社　員　数	[連結] 805名（2020年10月）
●関 係 会 社	株式会社エーエスエル 株式会社SICデジタル 株式会社シンクスクエア
●ビジネスパートナー契約	株式会社ソフトバンク（SynchRoid/AutomationAnywhere販売代理店契約） 株式会社NTTデータ（ビジネスパートナー） 株式会社NTTデータ（WinActor販売特約店契約） 日本アイ・ビー・エム株式会社（コアパートナー） 日本アイ・ビー・エム株式会社（IBM Cloud Partner League） 日本アイ・ビー・エム株式会社（IBM PartnerWorld Silver Business Partner） 三菱電機インフォメーションシステムズ株式会社（MDISソフトウェアパートナーズ） アマゾン ウェブ サービス ジャパン株式会社（APNセレクトコンサルティングパートナー） UiPath株式会社（開発リソースパートナー認定） Agile Performance Holarchy（Agile CxO Transformation Partner） Scaled Agile, Inc.（Scaled Agile Partner Network Bronze Partner）

CMMI DEV /5
Exp. 2018-11-06 / Appraisal #25452

PMP保有者 150名
2018年3月 現在在籍数