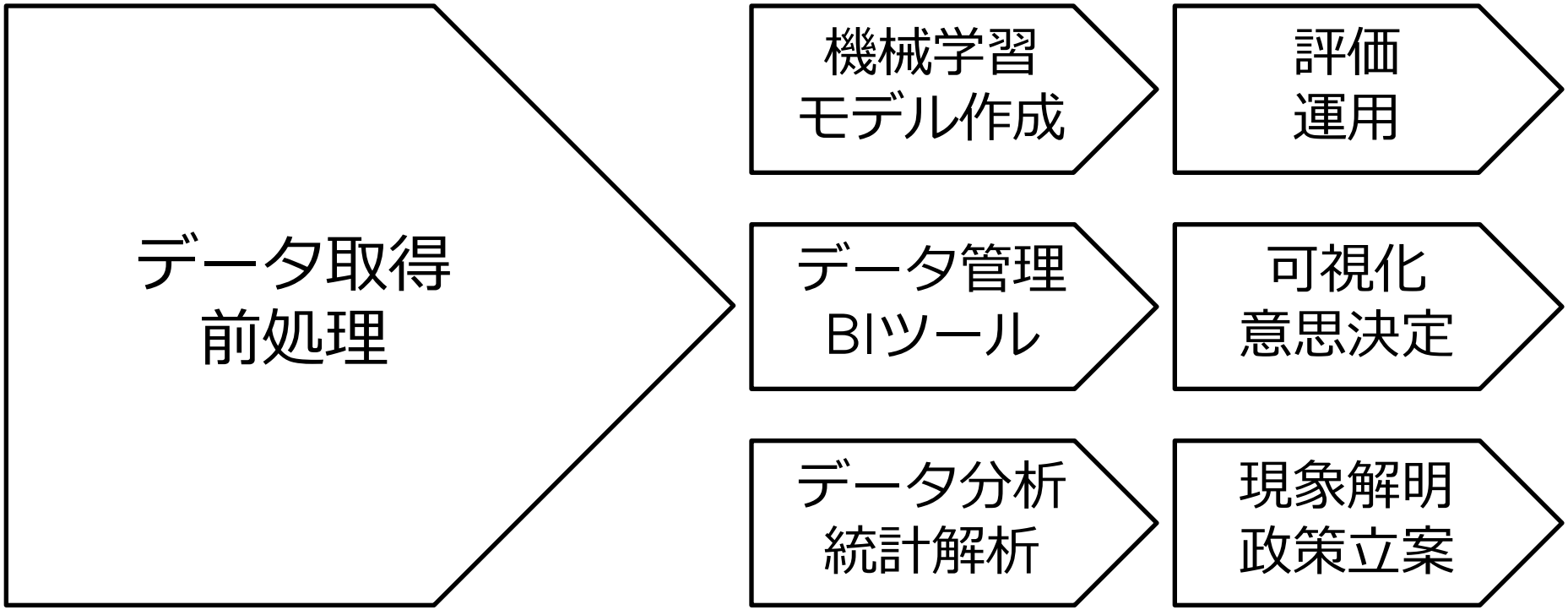
BigGorillaの活動紹介
～データ前処理ノウハウの共有化に向けて～

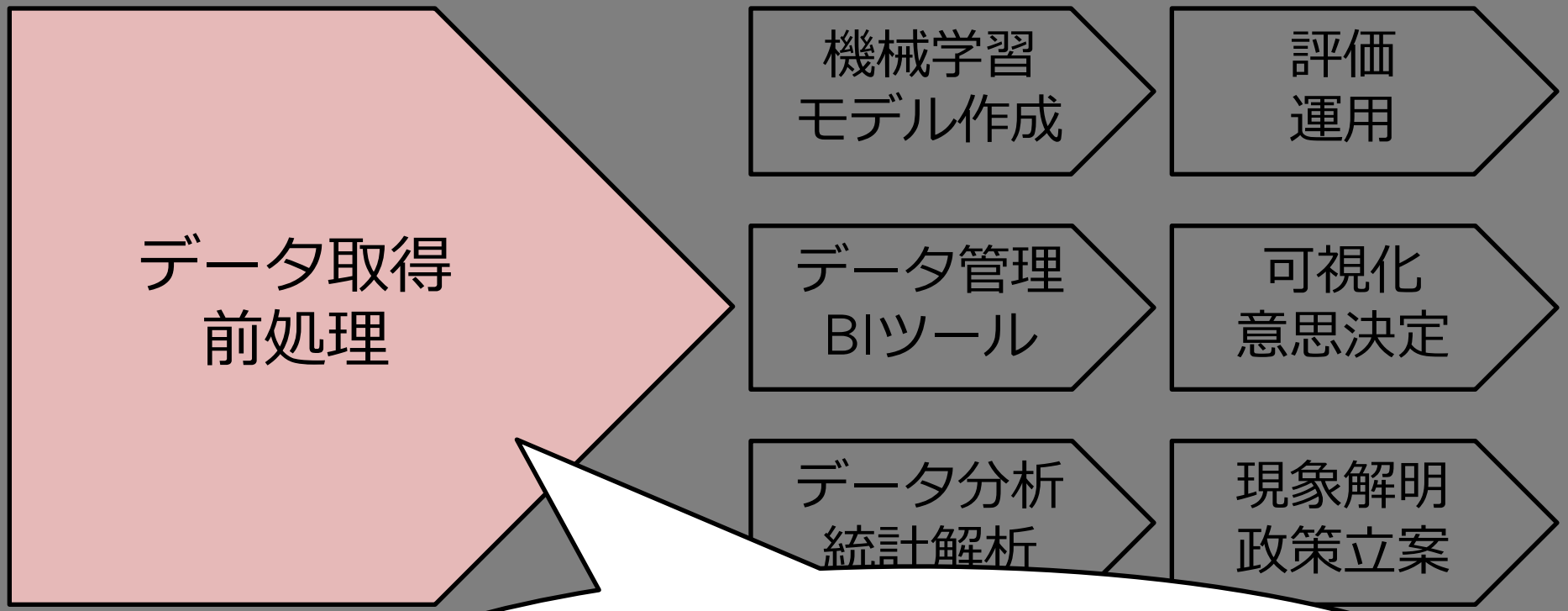
RECRUIT Institute of Technology　山下 宙元

データ活用のフロー



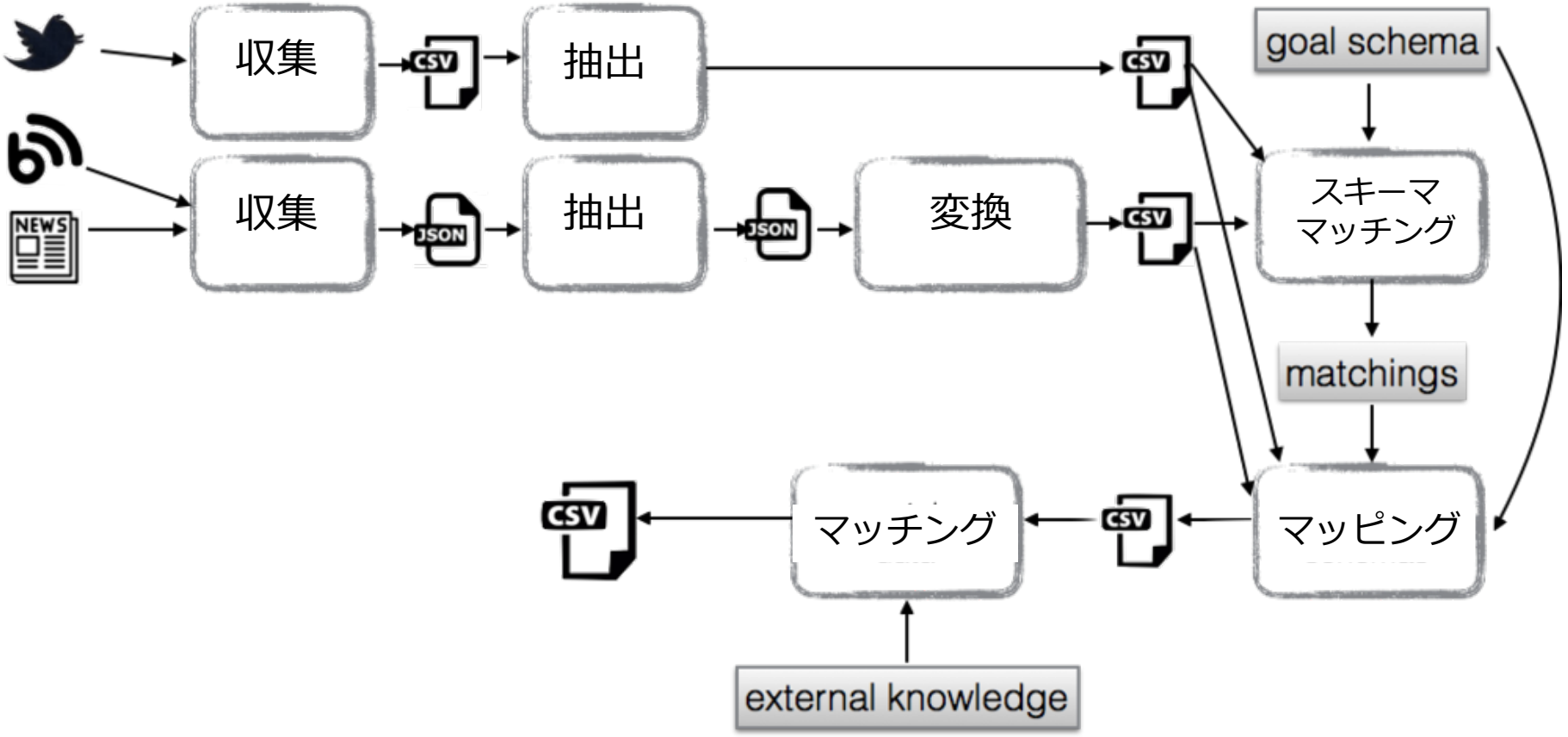
データ量の急速な増加、オープン分析ツールの増加により
データの利活用が高まっているが…

データ活用のフロー

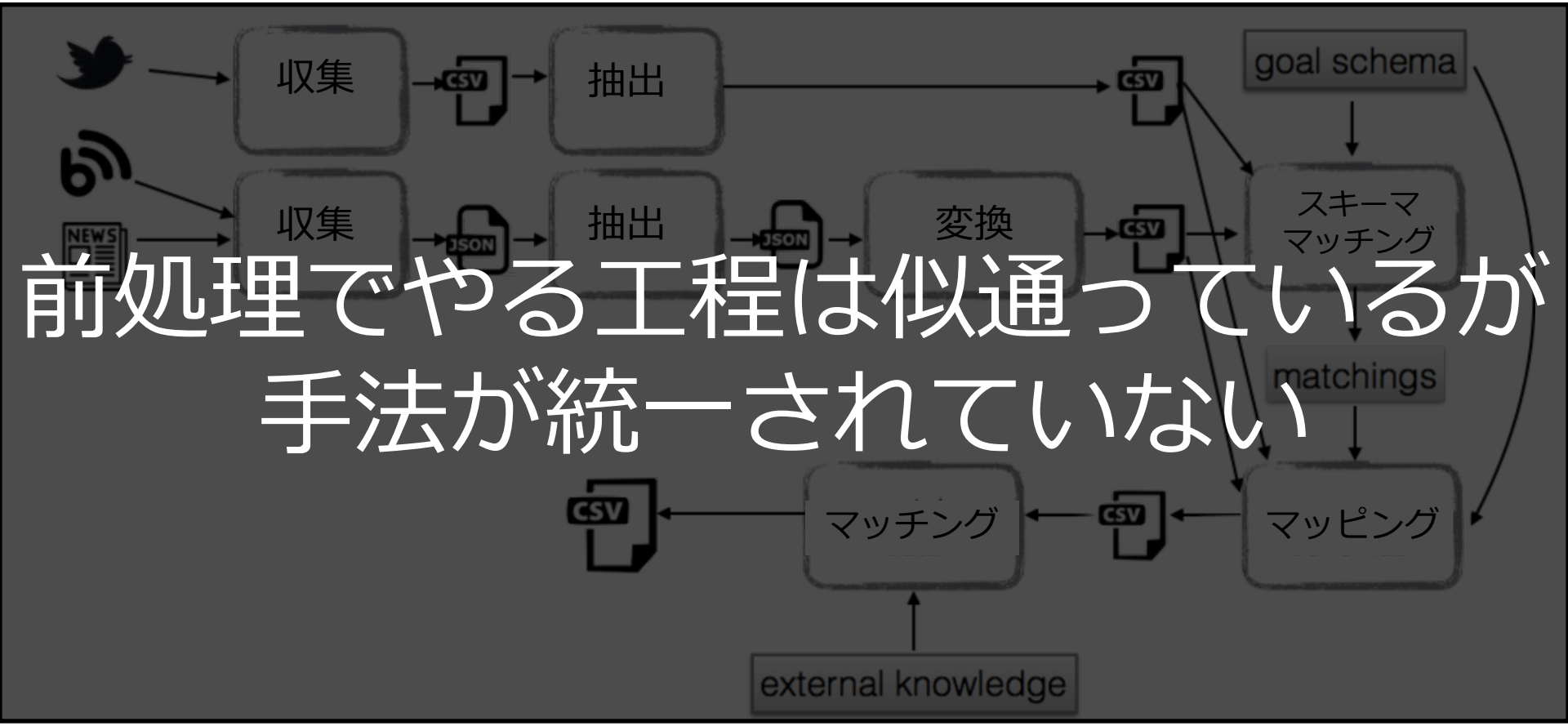


データ利活用において
工数の8割は前処理

前処理には複数のタスクがある



前処理には複数のタスクがある



前処理でやる工程は似通っているが
手法が統一されていない

BigGorillaのミッション

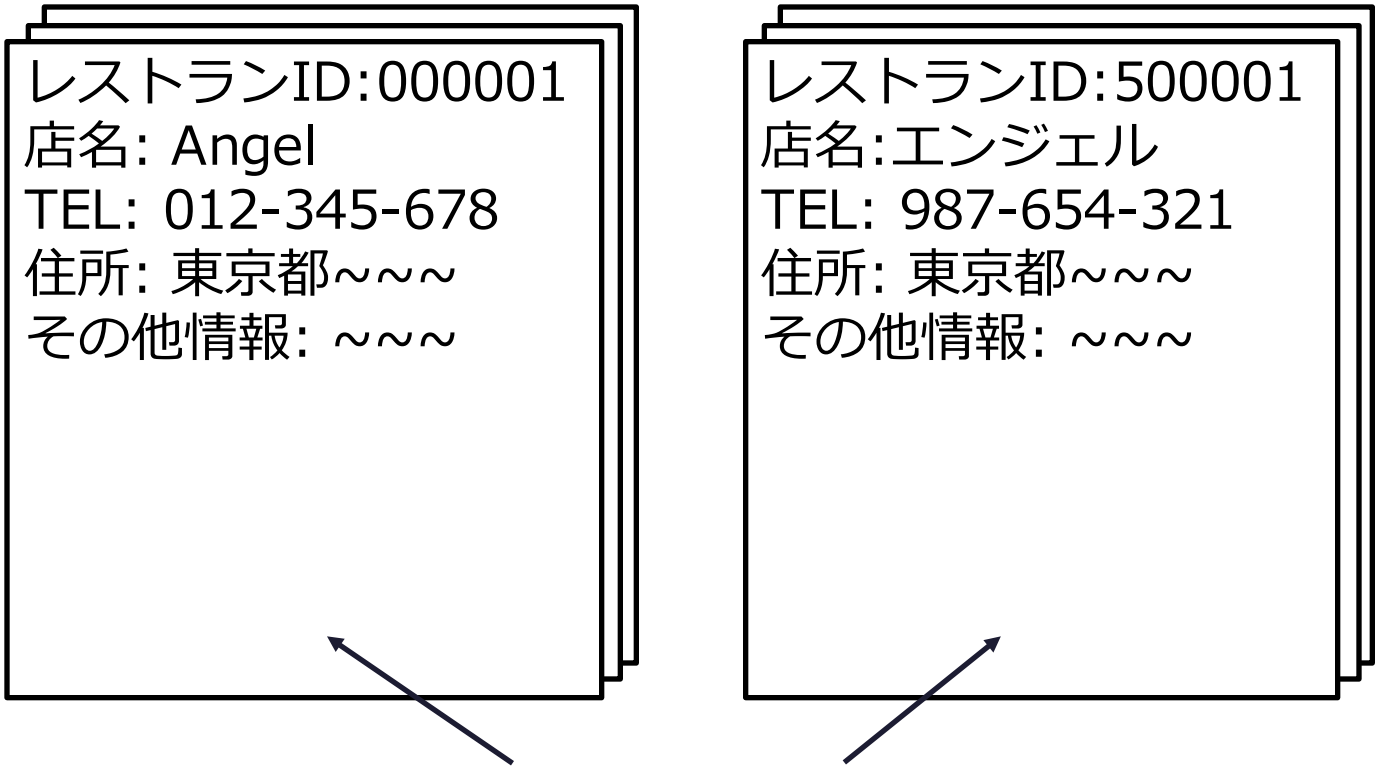
■  BigGorilla とは？

データ準備とデータ統合のためのオープンソース・
エコシステムを開発/啓蒙する**オープンコミュニティ**
（RITとWisconsin大学によりプロジェクト開始）

■ 活動の目的

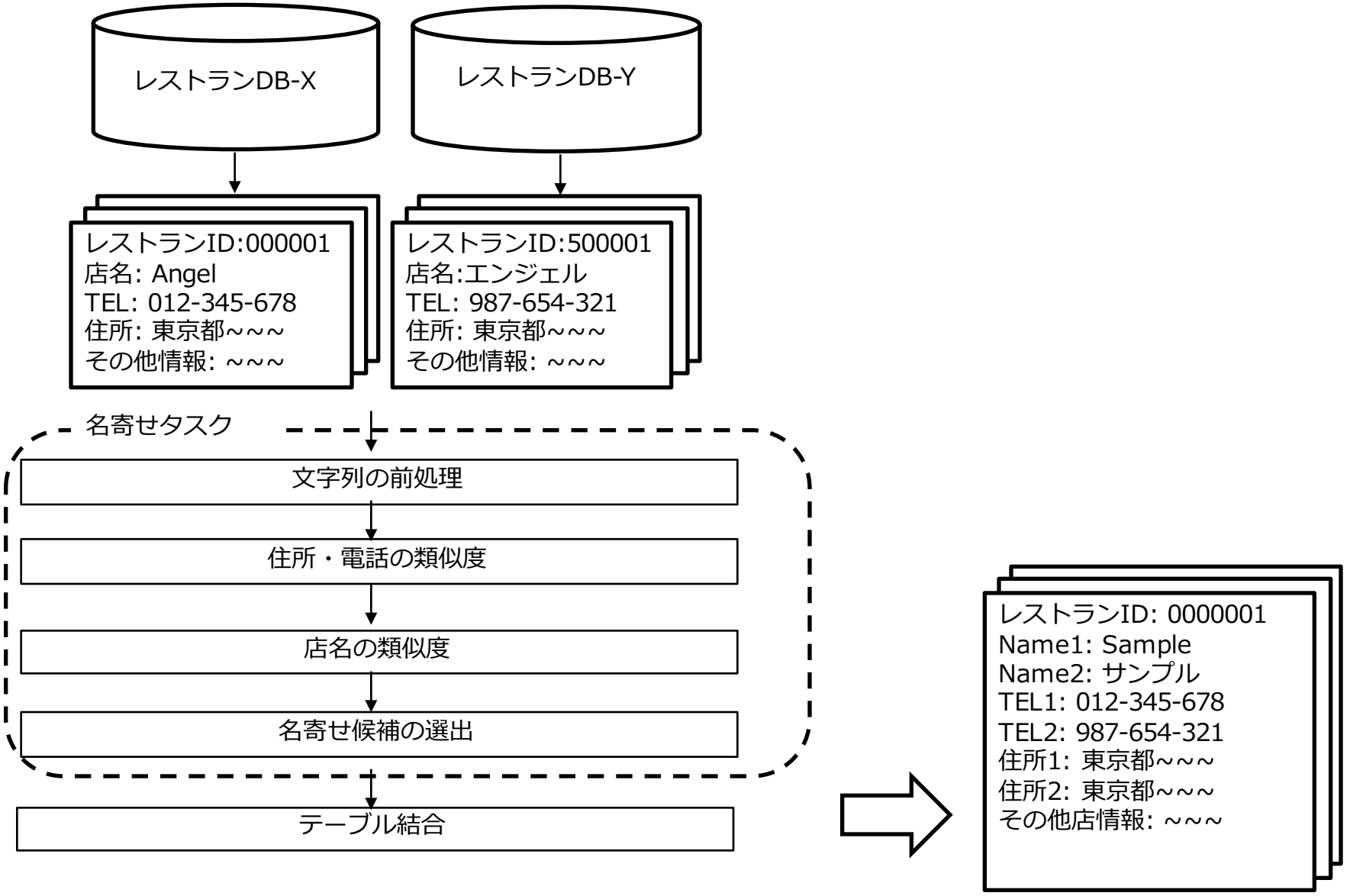
データに向き合う人が直面する困難を解決するための
知識共有の場（ソフトウェア/ワークフロー/データセット）
となること

ケーススタディ
～表記揺れのあるデータセットの名寄せ～

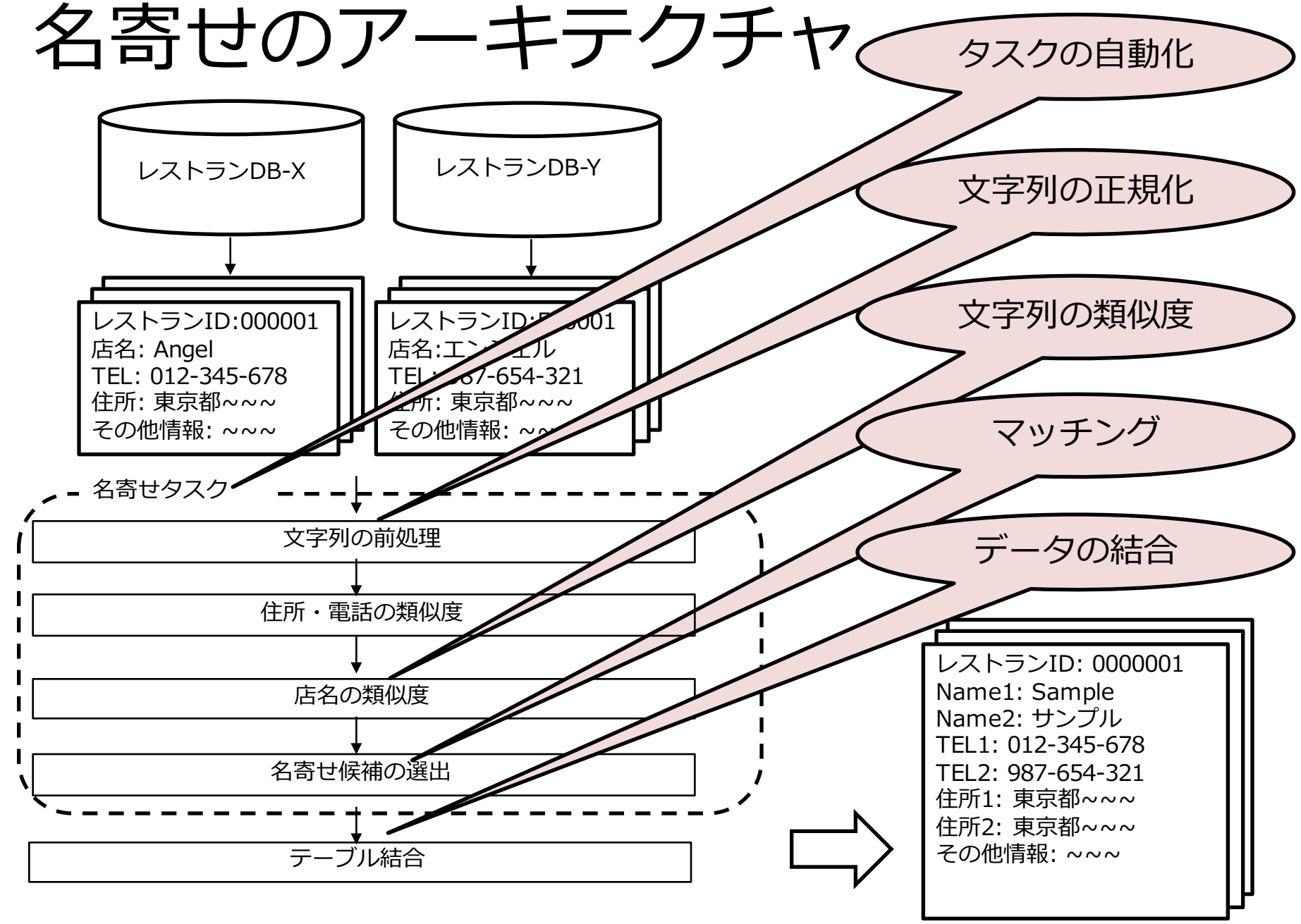


名寄せしたいが…
人手で辞書をつくるのは面倒…

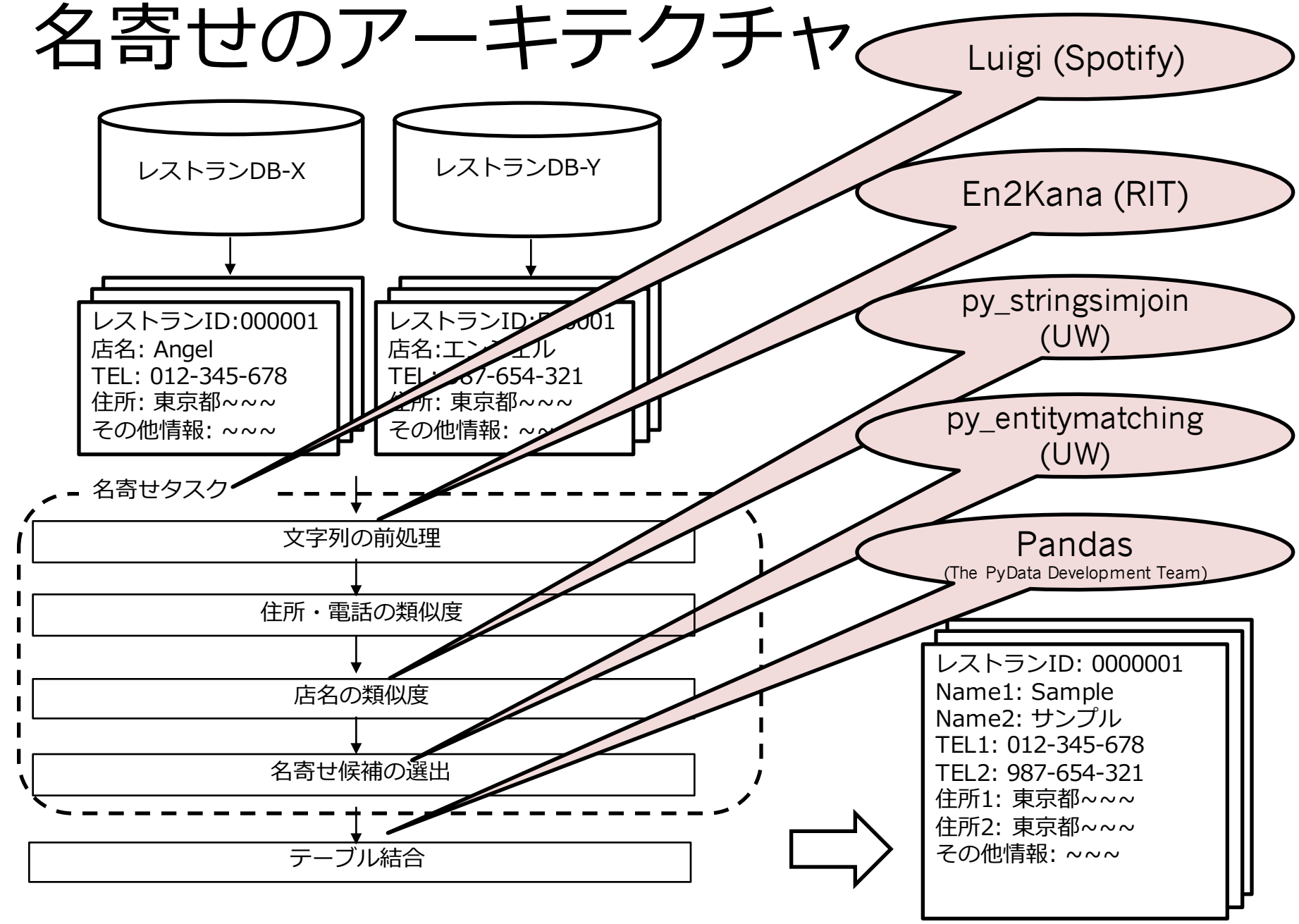
名寄せのアーキテクチャ



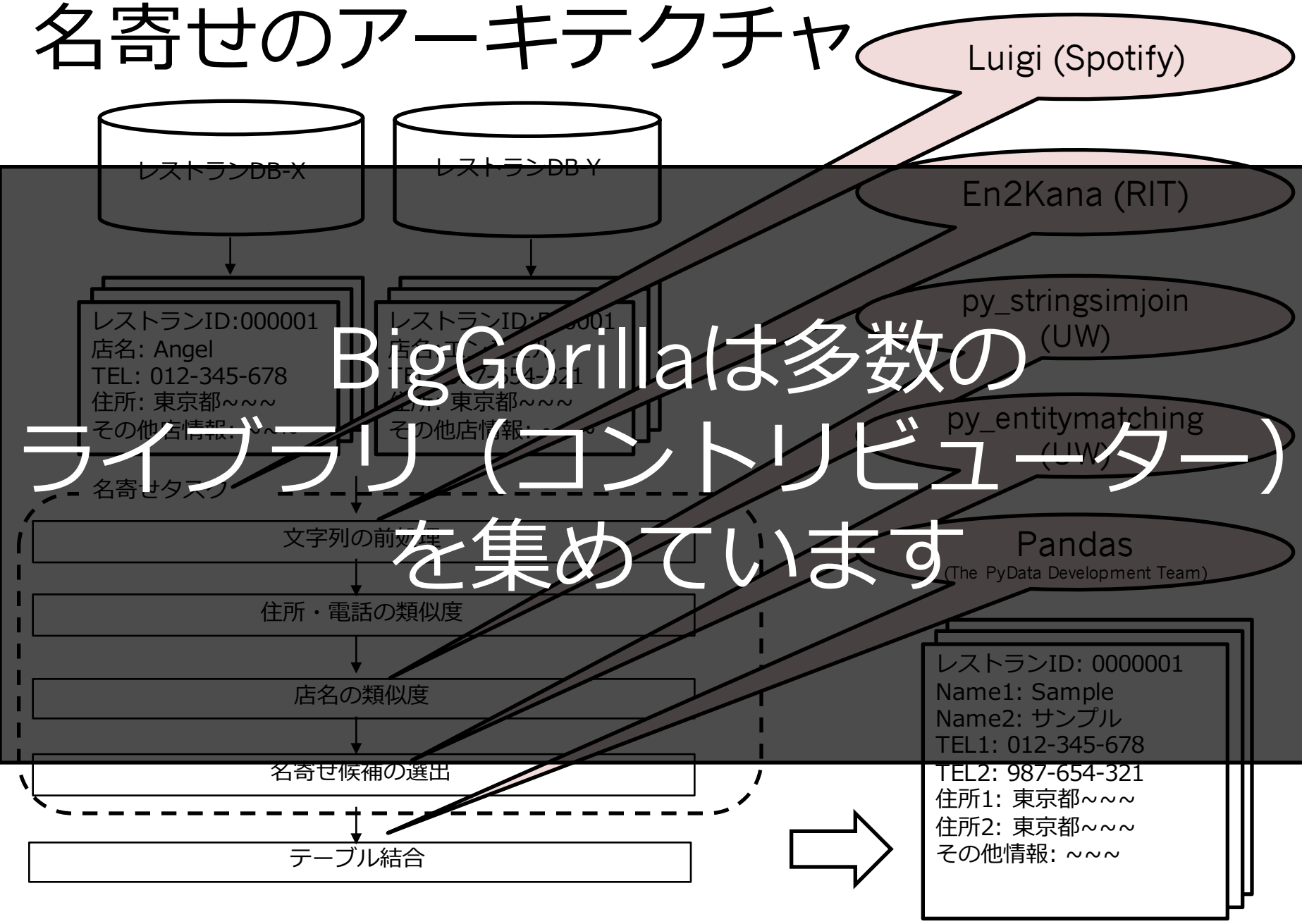
名寄せのアーキテクチャ



名寄せのアーキテクチャ

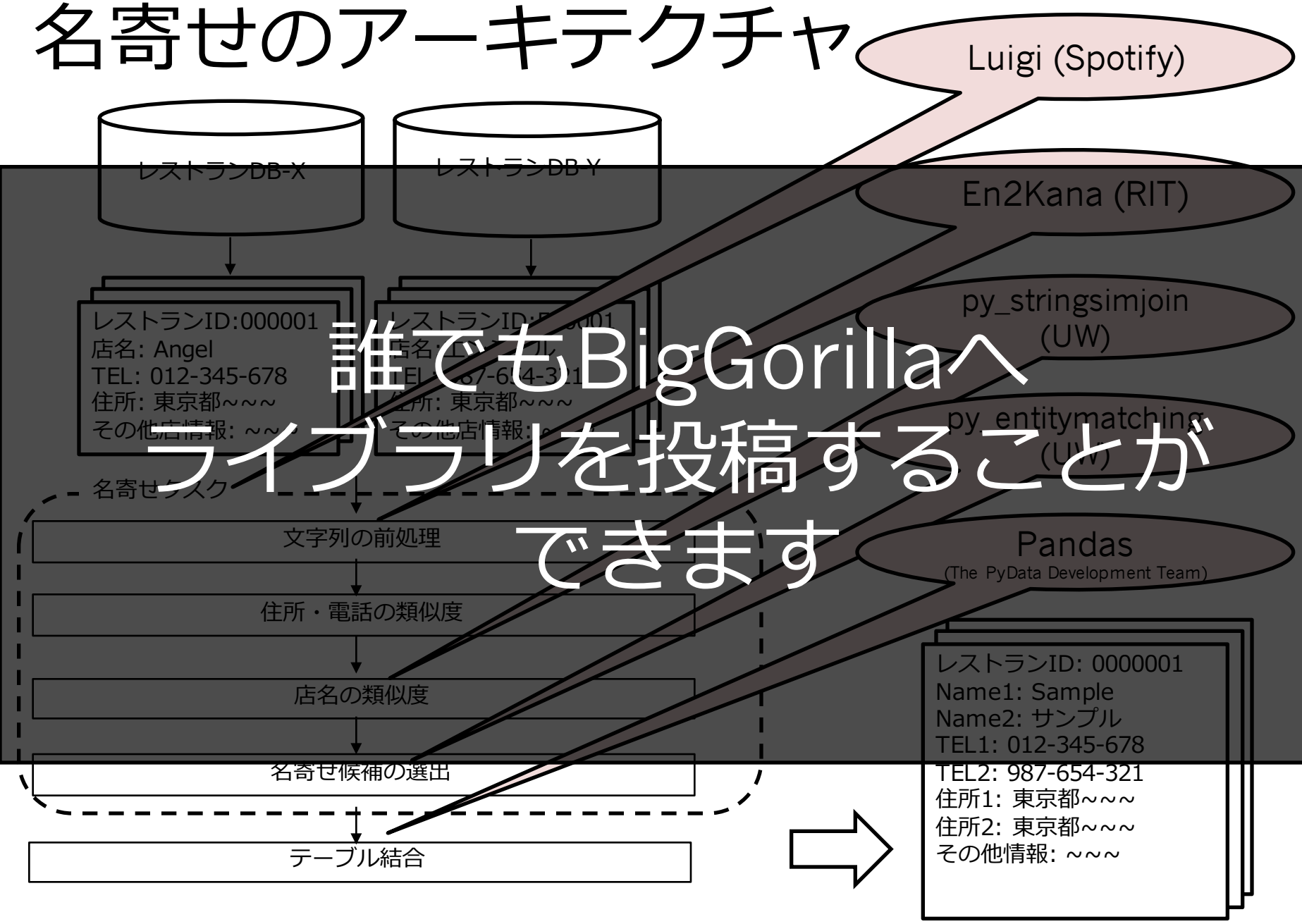


名寄せのアーキテクチャ



BigGorillaは多数のライブラリ（コントリビューター）を集めています

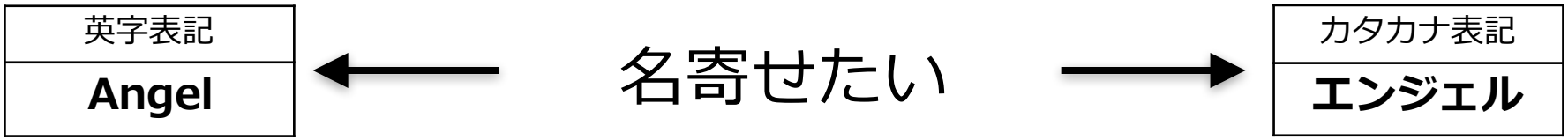
名寄せのアーキテクチャ



誰でもBigGorillaへ
ライブラリを投稿することが
できます

En2Kana：英語とカタカナの表記揺れを解決するモジュール

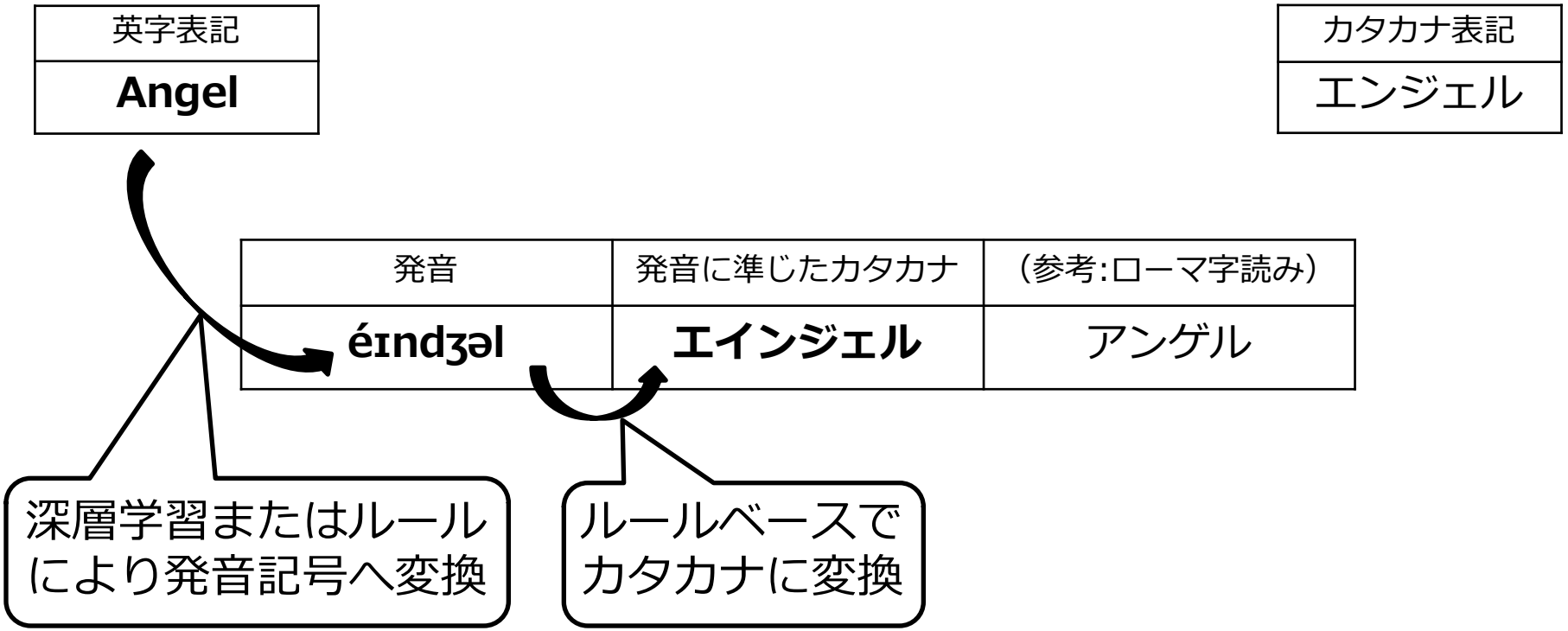
発音データを追加で用いることで
文字情報のみでも名寄せが可能に



発音	発音に準じたカタカナ	（参考：ローマ字読み）
éɪndʒəl	エインジェル	アンゲル

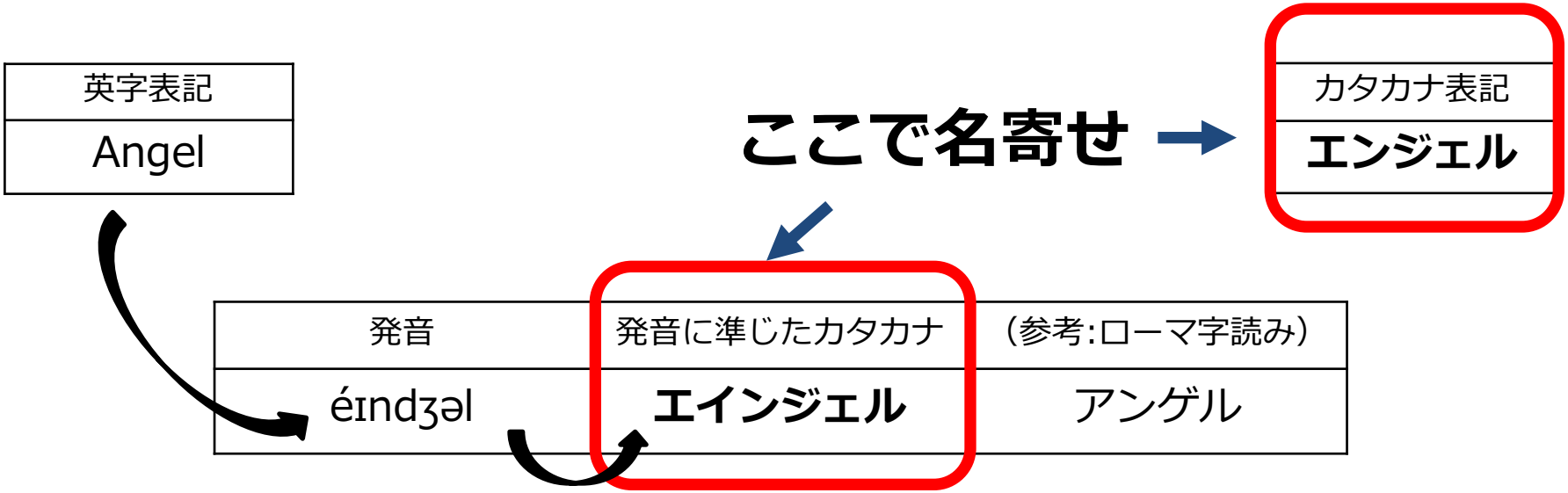
En2Kana：英語とカタカナの表記揺れを解決するモジュール

発音データを追加で用いることで
文字情報のみでも名寄せが可能に



En2Kana：英語とカタカナの表記揺れを解決するモジュール

発音データを追加で用いることで
文字情報のみでも名寄せが可能に



テーブル名寄せの効果・影響

- 発音情報がないときよりも**10%以上の精度改善**
- 英語とカタカナの表記揺れの多い
  10万件以上のデータテーブルを結合可能に
- レストラン名だけでなく
  美容院名・会社名・人名・商品名など
  他ドメインへも適用可能
- **サーチエンジンのクエリ改善**にも

BigGorilla その他のライブラリ

BigGorilla:

データの前処理問題を解決するための
知識共有を目指したオープンコミュニティ

ライブラリ例:

Usagi: メタデータの検索、管理

KOKO: テキストデータから欲しい情報を抽出

他にも現在約**30**個のライブラリが集積
（いつでも投稿お待ちしております）

BigGorillaへのコントリビューション

1. ライブラリの紹介

既存のものでも自作でもOK
データの前処理に必要なものをご共有ください

2. 活用事例・ユースケースの紹介

どのような方法でどのような前処理問題を解決したか
このような事例があったのでシェアしたいなど
資料、ドキュメント、コード等をご共有ください

3. フォーラムへの投稿（意見・質問）

こういう前処理の問題で困っている
こんな活用方法もあったなど

BigGorillaへのコントリビューション

1. ライブラリの紹介

既存のものでも自作でもOK
データの前処理に必要なものをご共有ください

2. 活用事例（データ前処理）の紹介

どのようなライブラリを活用したか
どのような前処理問題を解決したか等の
事例やコードをご共有ください

データ前処理の問題を
一緒に解決しませんか？

3. フォーラムへの投稿（意見・質問）

こういう前処理の問題で困っている
こんな解決方法もあったなど