
発表日：2025/05/23

ローカル環境で サービスを実現したい！

- 必要な技術要素と今後のタスク -

ML/DS 白川桃子

01

ローカル環境でのサービスに必要な技術要素

音声
to
文字
起こし

pyaudio

GCS

Speech-to-Text

：音声データ

：文字データ

：ラベル/スコア

：マッチング結果

：プロフィール

データ
分析
to
表示

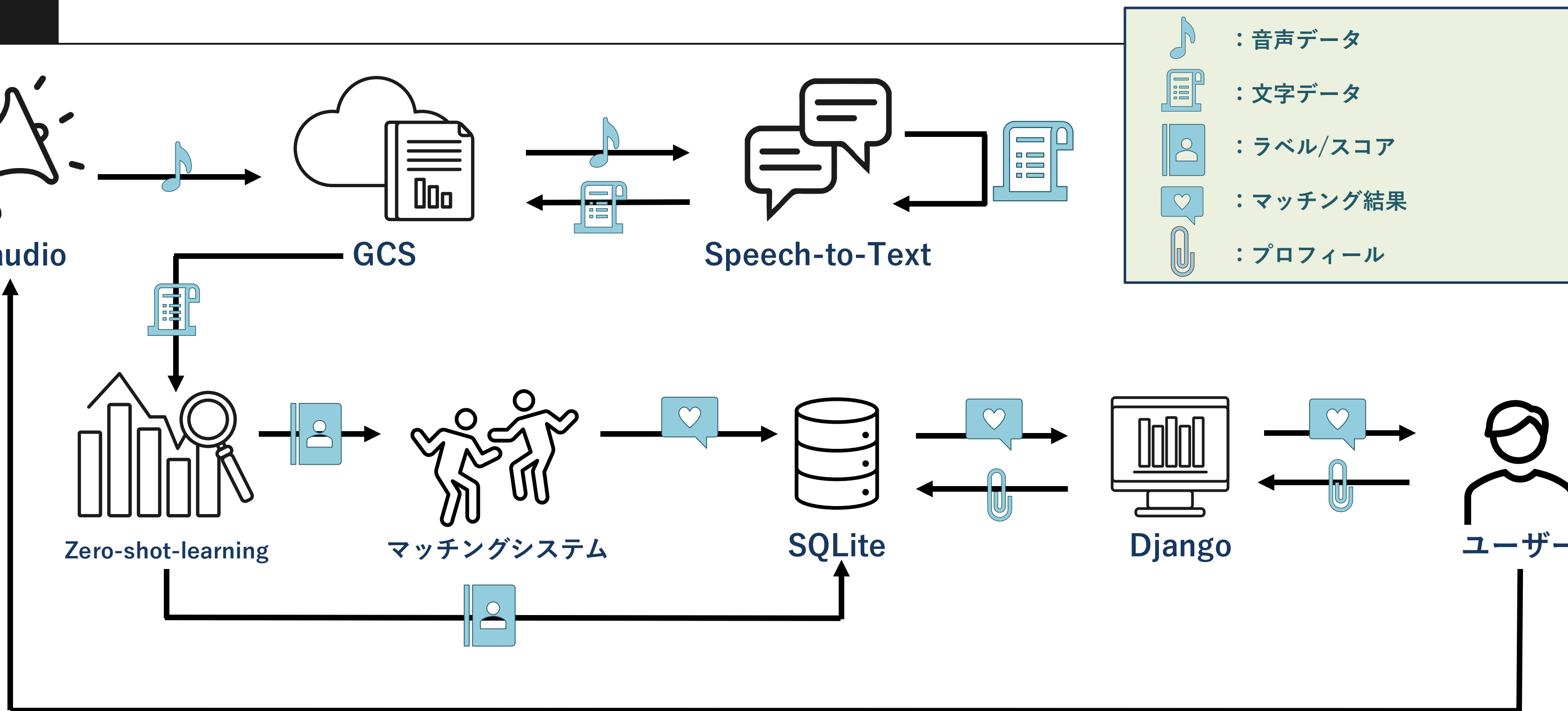
Zero-shot-learning

マッチングシステム

SQLite

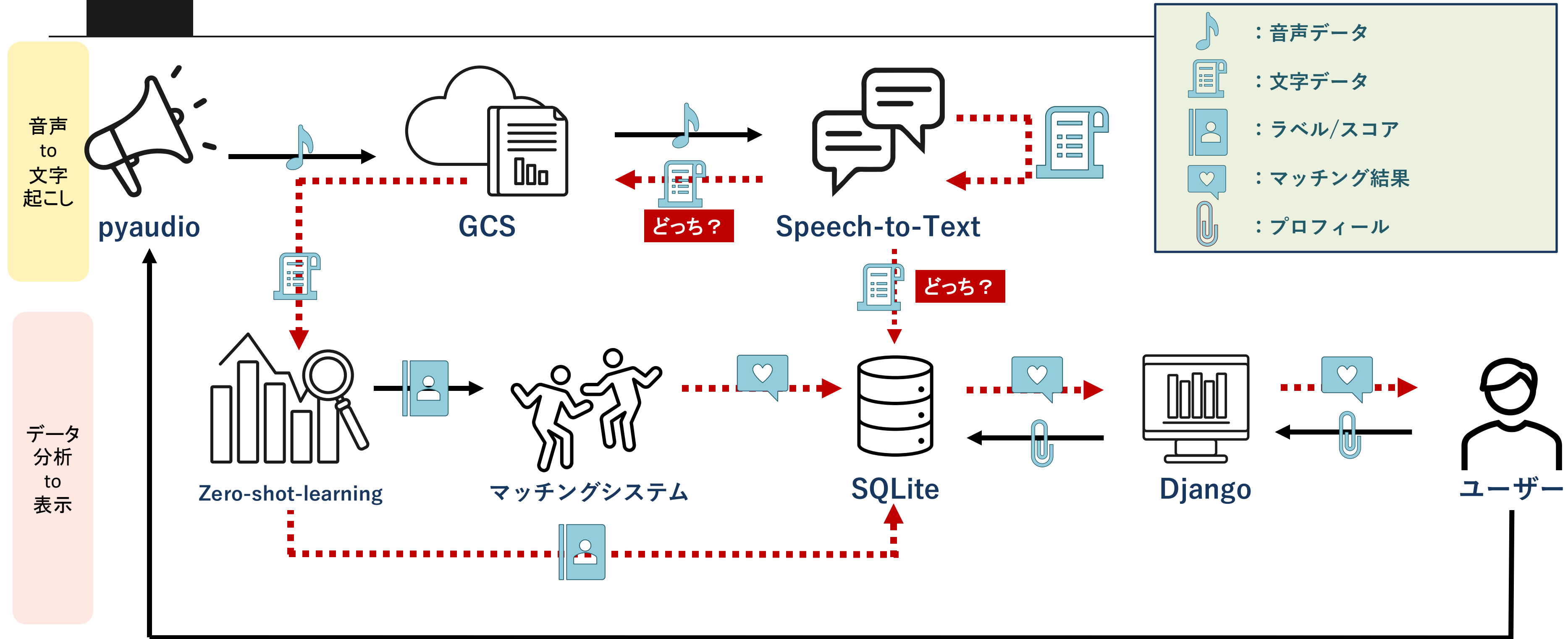
Django

ユーザー



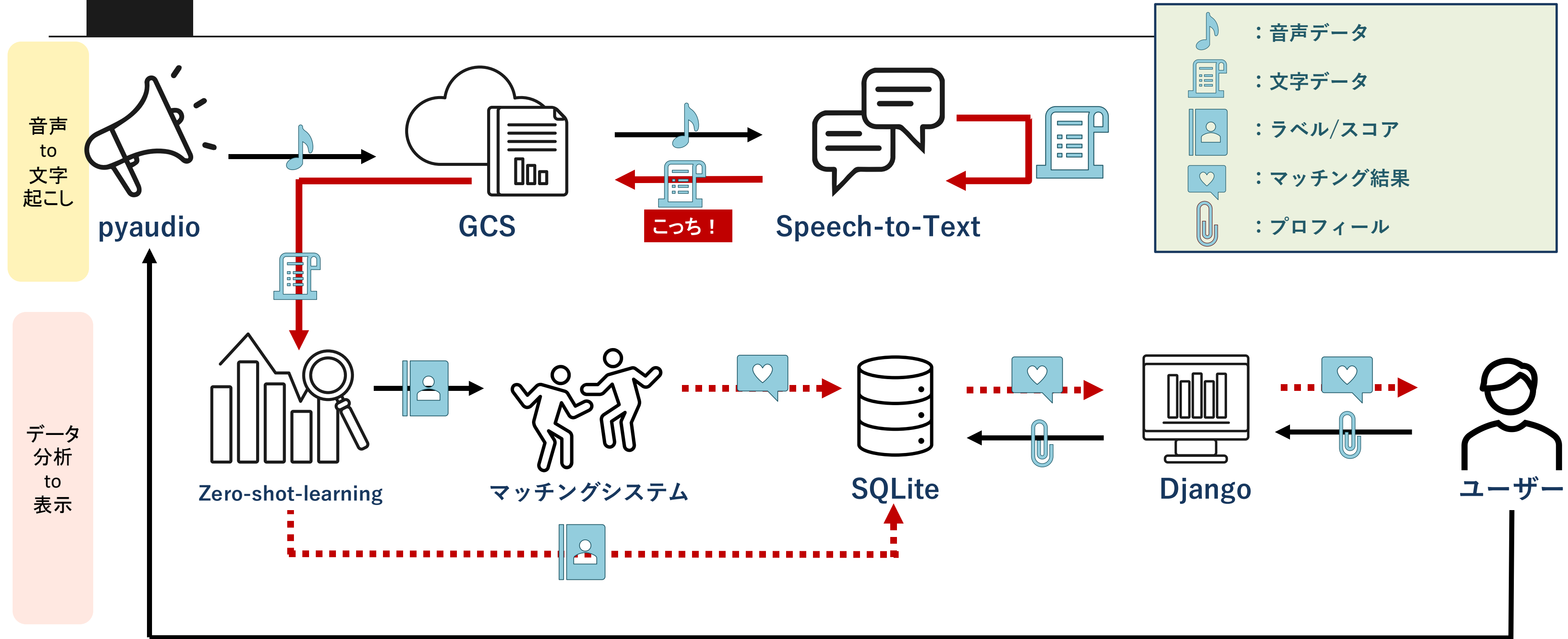
01

ローカル環境でのサービスに必要な技術要素



01

ローカル環境でのサービスに必要な技術要素



全ての発話内容が話者タグだけ変化して格納している

【本来の出力イメージ】

話者 0: 明日はミーティングです

話者 1: 識別できないね

話者 0: 悲しいです



```
=== Welcome to KOKORONIA ===  
Enterを押して会話を開始^-^  
録音中です！会話が終わったらEnterを押してね  
録音が完了しました  
音声データのアップロード完了: gs://kokoronia/media/audio/r
```

```
=== 文字起こしを開始します ===  
文字起こし中...
```

```
=== 話者分離された生データ ===
```

```
[無音] 0.00s - 3.20s  
話者 0: 明日|アシタ,アス,ミヨーニチ (3.20s - 3.70s)  
話者 0: は|ワ (3.70s - 3.80s)  
話者 0: ミーティング|ミーティング (3.80s - 4.40s)  
話者 0: です|デス (4.40s - 4.80s)  
話者 0: 識別|シキベツ (4.80s - 6.20s)  
話者 0: でき|デキ (6.20s - 6.60s)  
話者 0: ない|ナイ (6.60s - 6.60s)  
話者 0: ね|ネ (6.60s - 7.00s)  
話者 0: 話|ハナシ,バナシ (7.00s - 9.00s)  
話者 1: 明日|アシタ,アス,ミヨーニチ (3.20s - 3.70s)  
話者 1: は|ワ (3.70s - 3.80s)  
話者 1: ミーティング|ミーティング (3.80s - 4.40s)  
話者 1: です|デス (4.40s - 4.80s)  
話者 1: 識別|シキベツ (4.80s - 6.20s)  
話者 1: でき|デキ (6.20s - 6.60s)  
話者 1: ない|ナイ (6.60s - 6.60s)  
話者 1: ね|ネ (6.60s - 7.00s)  
話者 1: 話|ハナシ,バナシ (7.00s - 9.00s)
```

```
=====
```

Geminiに原因の要因を聞いてみた

①最小or最大の発話者数に合わせようとしている

→最小発話者数=最大発話者数=2に設定しても失敗...

②機械音声から特徴を抽出できない

→人が発話した音声データでも失敗...

③音声データの詳細設定のミス(e.g.)サンプリングレート, チャンネル数など

→問題がないことを確認済

③Google Cloud Speech-to-Text APIのダイアライゼーション機能は不適切

Gemini「上記の方法でも話者分離がうまくいかない場合、Google Cloud Speech-to-Text APIの話者分離機能が、
あなたの目的や音声データの特性に対して、現状では十分な精度を提供できない
と判断せざるを得ません。」