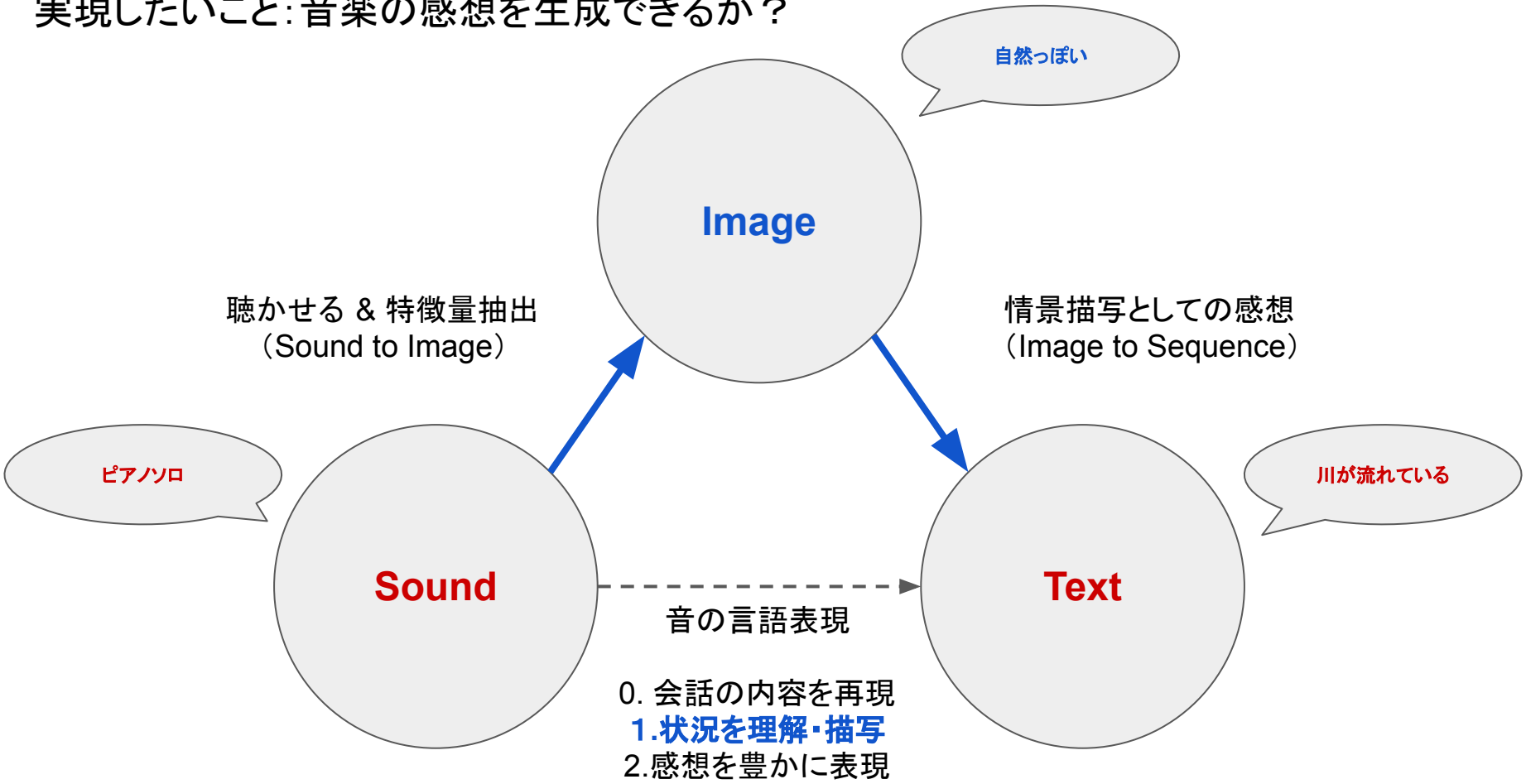
Oto no “Sommelier”

Sound 2 Image 2 Sequence

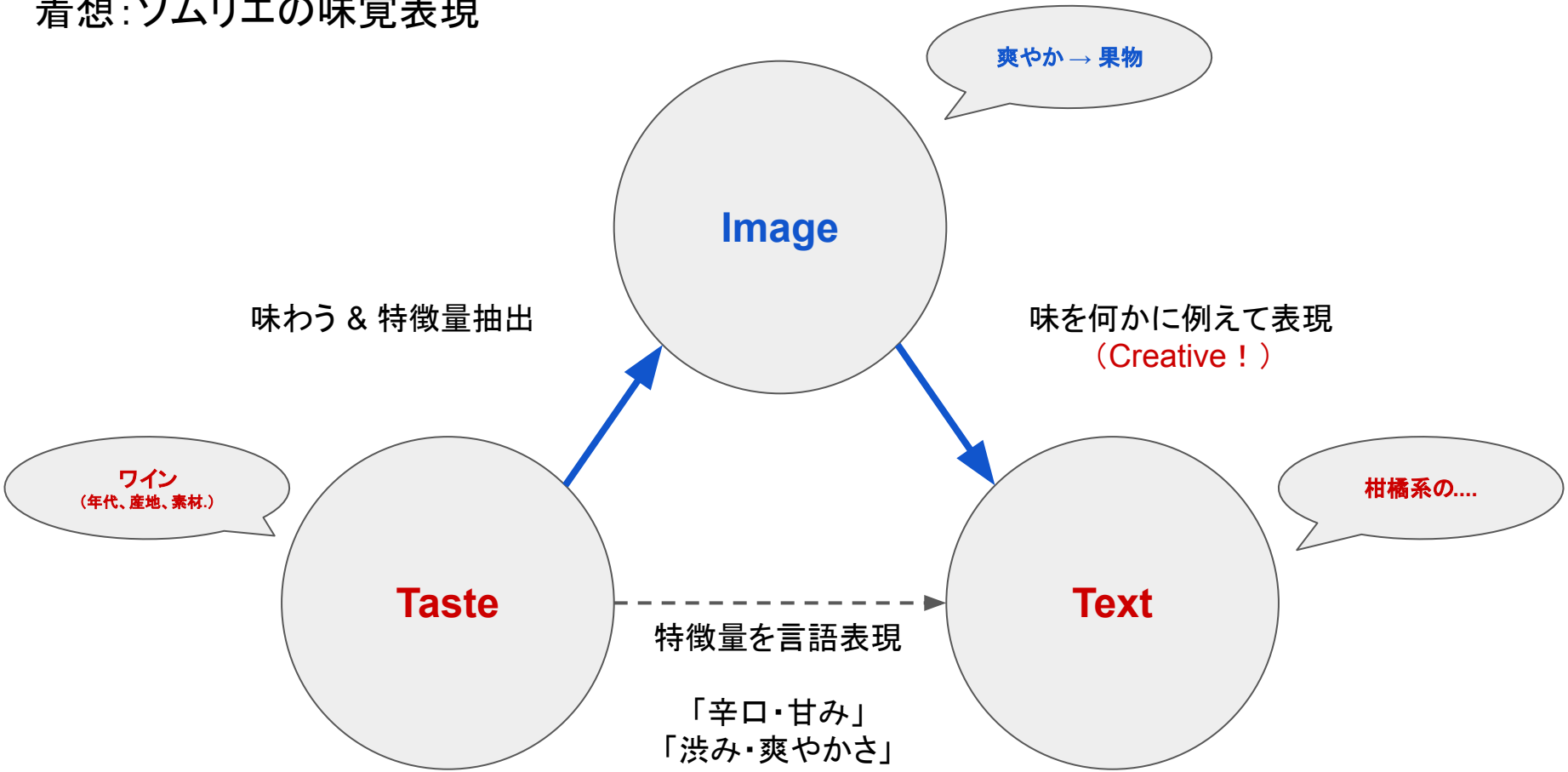
DIVE INTO CODE
Term 2019.4

Makoto Shirasu

実現したいこと：音楽の感想を生成できるか？



着想:ソムリエの味覚表現



なぜやるのか？

1. 音声処理 & 画像認識 & 自然言語処理をフルパッケージ（三位一体）で使えないか？

=> end2endで実装しない理由は？

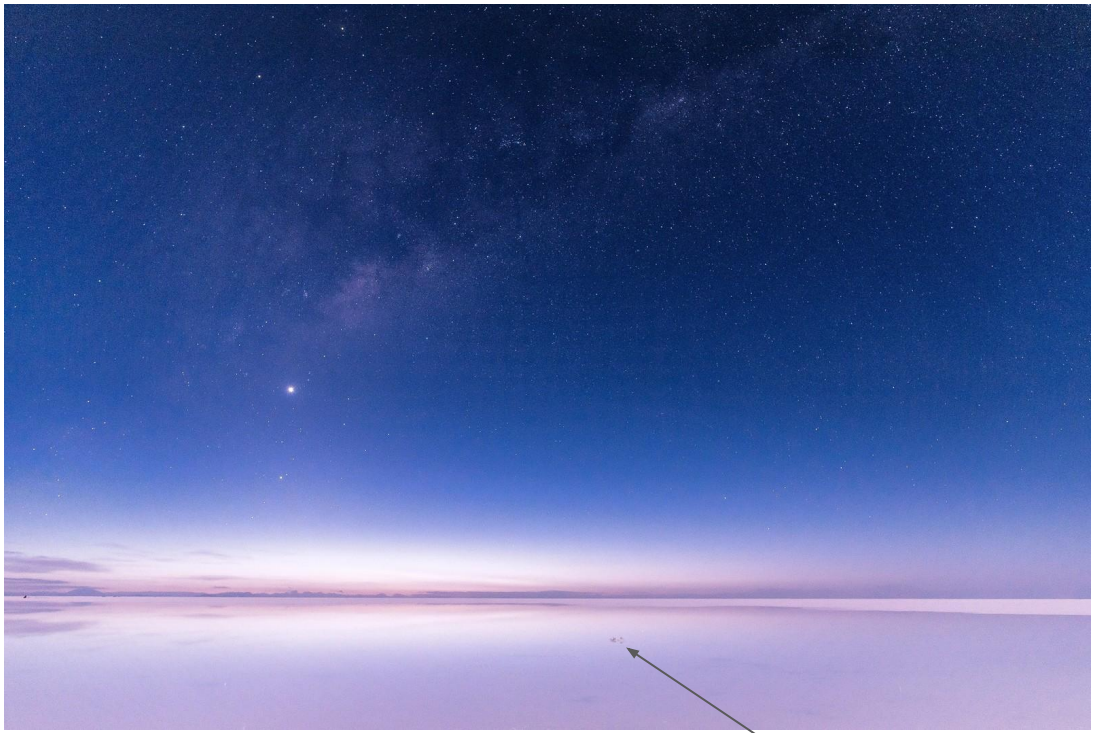
2. 音楽の感想を述べることは、なかなか難しい。

=> 音楽評論家のレビュー ：**職業柄、専門的な知識が多くて、ウンチクっぽい。**

=> 一般の人（TwitterやFacebookのSNS）：**「よかった！」「上手い！」 => もう少し豊かにできる？**

3. 感じたことを**「少しでも豊かな言葉で表現」**して、**「感想を共有する文化」**が生まれれば良いな。

Sprintで一番感動した瞬間



Image：

ウユニ塩湖

Caption：

A view of a lake with a boat in the background.

Tecnology：

Image to sequence

pytorch-tutorial/tutorials/03-advanced/image_captioning at master · yunjey/pytorch-tutorial

ボート？？（そうは見えないけど ...）

要件定義:

・選択肢は？:①入力した音から状況を推測する、②豊かな感想を生成する

=> 今回は、まずシンプルに①の実現を目指す。

・重要な評価指標は？

=> 感想なので、正確性よりも**類似度（＝〇〇っぽい）**が重要

・どこに力を入れる？

①モデルのつなぎこみ（音声 => 画像 & 画像 => シーケンス）

②音声データの取り扱い（前処理、モデル）

③データ収集

プロジェクトの全体感（構想）

1. **データ収集**：どのようなデータ（音声・画像・動画）をどこから取得する？

   => Youtube？Kaggle？Free Sample？

2. **前処理**：音声データの前処理、学習データ（画像・音声の対応関係）の生成

   => Kaggle（FreeSound主催）のコンペから学ぶ。

3. **モデル構築・予測・評価**：特に、どのようなモデルを構築する？

   => Image to Sequenceをアレンジして、”Sound => Image => Sequence”

4. **その他**：参考になりそうな論文を読む＆まとめて、ヒントを得る。

Expansion（Not to Do）

1. 言語表現の豊かさの追求

   => 言葉の表現力を高めようとすると、生成モデルを用いる必要があるかも。

2. リアルタイム性の追求（マイクで拾った音声をリアルタイムで変換する）

   => まずは、音声ファイルを読み込ませて出力すること。

おわり