シンポジウム「データサイエンスと人文学の協働による研究・教育の可能性 : 九州大学数理・データサイエンス教育研究センターの取り組み」

上山, あゆみ
九州大学大学院人文科学研究院 : 研究院長

太田, 真理
九州大学大学院人文科学研究院

内田, 誠一
九州大学大学院システム情報科学研究院 : 教授

川野, 秀一
九州大学大学院数理学研究院

他

https://doi.org/10.15017/6776430

出版情報 : 2023-03-15. Faculty of Humanities, Kyushu University
バージョン :
権利関係 :

令和5 年3 月15 日　@オンライン

シンポジウム「データサイエンスと人文学の協働による研究・教育の可能性」

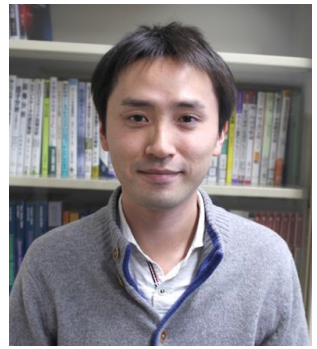
計量万葉学：データサイエンスによるアプローチ

九州大学大学院数理学研究院　川野 秀一

skawano@math.kyushu-u.ac.jp

https://sites.google.com/view/kawanolab

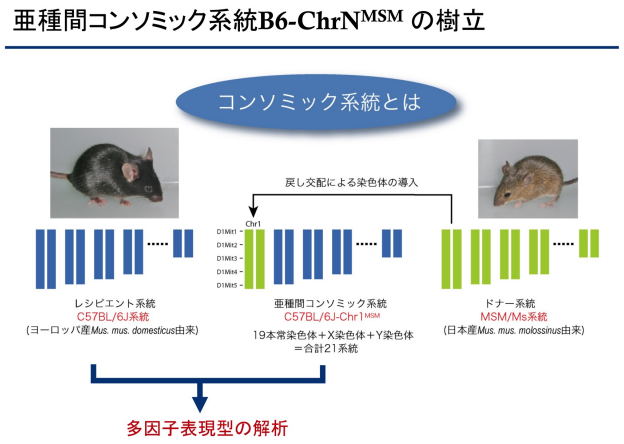
自己紹介

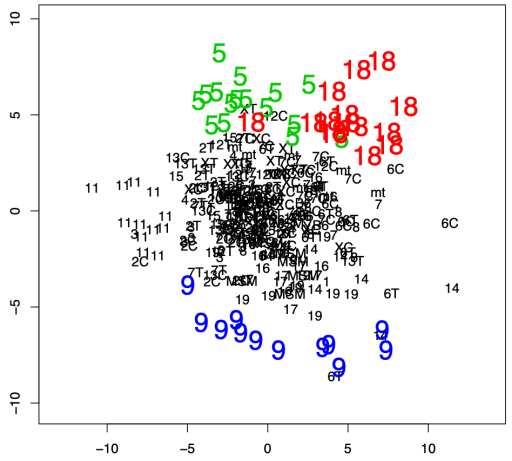


専門分野：統計科学，データサイエンス，機械学習

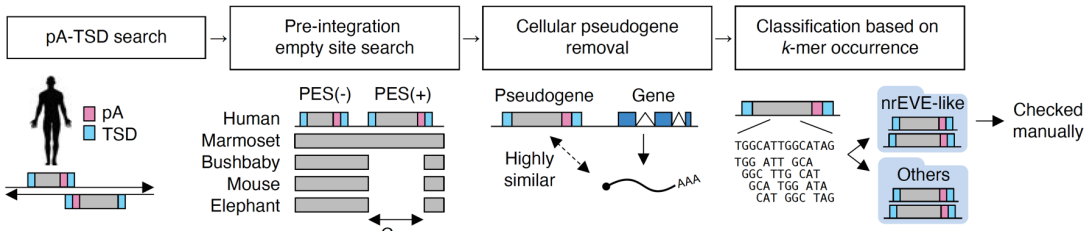
研究キーワード

- 統計学に基づくデータ解析手法の開発
- スパースモデリング
- バイオインフォマティクス
- 統計的因果推論





コンソミック系統マウスの統計モデリング (Kawano et al., 2018; CSDA)．数字はマウスの種類を表す．



新規ウイルス配列発見のためのバイオインフォマティクス手法の開発 (Kojima et al., 2021; PNAS)

共同研究者

村田右富実（関西大学 文学部 総合人文学科 国語国文学専修）

専門分野：日本古典文学

研究キーワード

- 『万葉集』を中心に奈良時代以前の韻文学
- 国語学
- 歴史学
- 考古学

共同研究における役割分担は？

- 研究の方針決め（二人で）
- データ提供（村田先生）
- データ解析（川野）
- 解析結果の解釈（村田先生）

万葉歌について

万葉集：現存する日本最古の歌集（8世紀中～後成立）

万葉歌は次の 4 つに分類される

- 短歌・・・約 4200 首　(5 7 5 7 7)
- 長歌・・・265 首　((5 7)×n＋7)
- 旋頭歌・・・62 首　(5 7 7 5 7 7)
- 仏足石歌・・・1 首　(5 7 5 7 7 7)

短歌の例　　玉剋春　内乃大野尓　馬數而　朝布麻須等六　其草深野

（たまきはる うちのおほのに うまなめて あさふますらむ そのくさふかの）

万葉歌の計量分析

1. 新井 (1998; 一橋論叢)

- それぞれの短歌と旋頭歌に含まれる母音 (a, i, u, e, o) と頭子音に着目
- 階層的クラスタリングと主成分分析により解析

2. 村田 (2009; 萬葉)

- それぞれの短歌に含まれる母音 (a, i, u, e, o) に着目
- 適合度検定により解析
- 山上憶良，東歌，防人歌の p 値が小さいという結果

万葉歌の計量分析

3. 村田・川野 (2014; 美夫君志)

- それぞれの短歌に含まれる母音 (a, i, u, e, o) に着目
- 音の羅列といった時系列情報を取り入れデータ化
- 混合効果モデルにより解析
- 山上憶良，東歌，防人歌の固定効果項の挙動が，他と比べて異なっているという結果

4. 村田・川野 (2016; 上代文学)

- 短歌内で使用されている文字に着目（特徴量 : 1,286）
- 1-class SVM を用いて解析
- 吉田宜，防人歌，巻十八の補修部が外れ値として検出

万葉歌の計量分析

5. 村田・川野 (2017; 文学・語学)

- 短歌内で使用されている音に着目（特徴量 : 31）
- 1-class SVM を用いて解析
- 巻 14 と巻 16 が外れ値として検出

6. 村田・川野 (2019; 萬葉)

- 万葉歌内での異伝注記の本伝，異伝の文字数に着目
- t 検定により解析
- 異伝の 2 つの種類間の p 値は小さいという結果

万葉歌の計量分析

7. 川野・村田 (2019; 応用統計学)

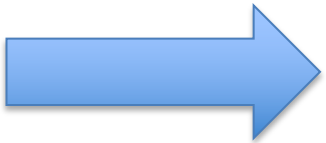
- 短歌内で使用されている音に着目
- スパース正準判別分析を用いて解析
- 柿本人麻呂，山上憶良，大伴旅人をうまく分類する音を検出

8. 村田・川野・吉川 (2021; 美夫君志)

- 短歌内で使用されている音節（2音節）に着目（特徴量 : 4,005）
- トピックモデル，特に潜在的ディリクレ配分法 (latent Dirichlet allocation: LDA) を用いて解析
- 9 個のトピック「心・思ふの恋歌」「反実仮想の恋の恋歌」「人言の恋歌」「山の歌」「海の旅の歌」「陸路の旅の歌」「川の歌」「夏の雑歌」「秋の雑歌」にわけることができた

本研究の目的

歌の調子からそれぞれの歌作者の特徴を捉えたい！
（「文字」ではなく「音」に着目する）



これまでは
- 単一の音のみに着目して主観的に判断
- 31 音に限定したデータ化
- 特徴量選択は行われていない

本研究で行ったこと

- 万葉歌内で用いられている音素パターンを特徴量として用いる（特徴量：68）
- 歌作者は，柿本人麻呂，山上憶良，大伴旅人に着目
- スパース正準判別分析により，判別分析と特徴量選択を同時に実行する

万葉歌のデータ化

- 各歌人の各歌を音素パターンに分ける
  （音素パターン：「あ」から「を」までの **68** 種類）
- 今回は短歌（字余り・字足らず含む）のみを対象

例えば，柿本人麻呂の 1 巻 30 の歌は下記のようになる

	あ	い	う	え	お	か	き	～	を
1 巻 30	1	0	0	0	1	2	2	～	0

左散難弥乃　思賀乃辛碕　雖幸有　大宮人之　船麻知兼津
（ささなみの しがのからさき さきくあれど おほみやひとの ふねまちかねつ）

各歌人の歌数

柿本人麻呂：70　　山上憶良：62　　大伴旅人：55

解析手法（スパース正準判別分析）

スパース正準判別分析 (Witten and Tibshirani, 2011; JRSS-B)

- フィッシャーの正準判別分析にスパース推定を組み込んだもの
- 高次元データにも対応

以下の最大化問題を解くことによって第 k 判別軸を得る

$$\max_{\boldsymbol{\beta}_k} \left\{ \boldsymbol{\beta}_k^T \hat{\Sigma}_b^k \boldsymbol{\beta}_k - \lambda_k \sum_{j=1}^{p} |\hat{\sigma}_j \beta_{kj}| \right\} \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \tilde{\Sigma}_w \boldsymbol{\beta}_k \leq 1$$

$\tilde{\Sigma}_w = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$：群内分散の推定量

$\hat{\Sigma}_b^k = \frac{1}{n} X^T Y (Y^T Y)^{-1/2} P_k^{\perp} (Y^T Y)^{-1/2} Y^T X$：群間分散の推定量

$\lambda_k = \lambda \| \tilde{\Sigma}^{-1/2} \hat{\Sigma}_b^k \tilde{\Sigma}^{-1/2} \|$

解析手法（スパース正準判別分析）

Witten and Tibshirani (2011) の貢献

従来の正準判別分析の定式化

① $\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^T \hat{\Sigma}_b \boldsymbol{\beta}_k \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \hat{\Sigma}_w \boldsymbol{\beta}_k = 1,\ \boldsymbol{\beta}_k^T \hat{\Sigma}_w \boldsymbol{\beta}_i = 0 \ (i < k)$

Witten and Tibshirani (2011) の正準判別分析の定式化

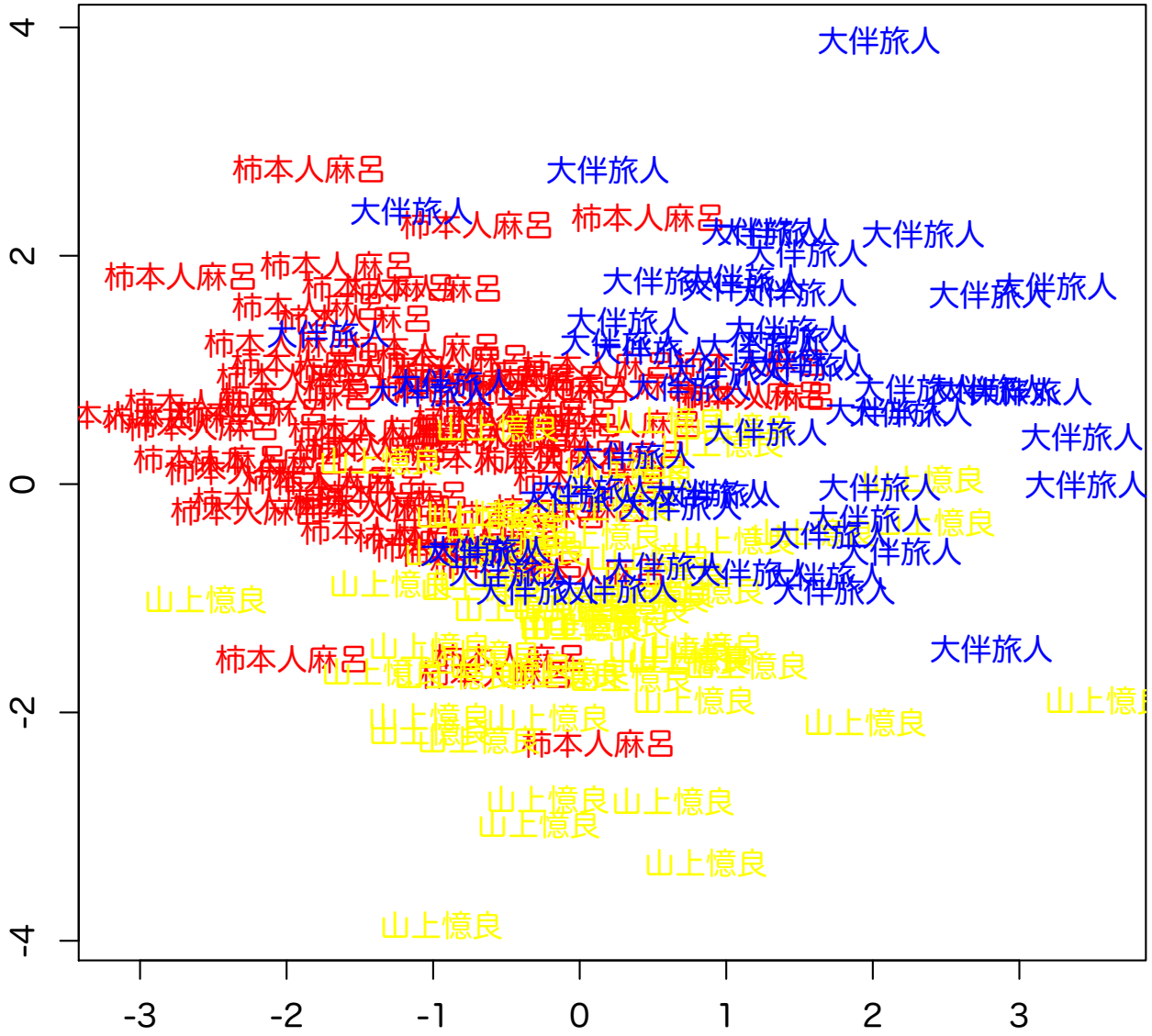
② $\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^T \hat{\Sigma}_b \boldsymbol{\beta}_k \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \hat{\Sigma}_w \boldsymbol{\beta}_k \leq 1,\ \boldsymbol{\beta}_k^T \hat{\Sigma}_w \boldsymbol{\beta}_i = 0 \ (i < k)$

③ $\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^T \hat{\Sigma}_b^k \boldsymbol{\beta}_k \quad \text{subject to} \quad \boldsymbol{\beta}_k^T \hat{\Sigma}_w \boldsymbol{\beta}_k \leq 1$

上の①から③までの最大化問題はすべて同値

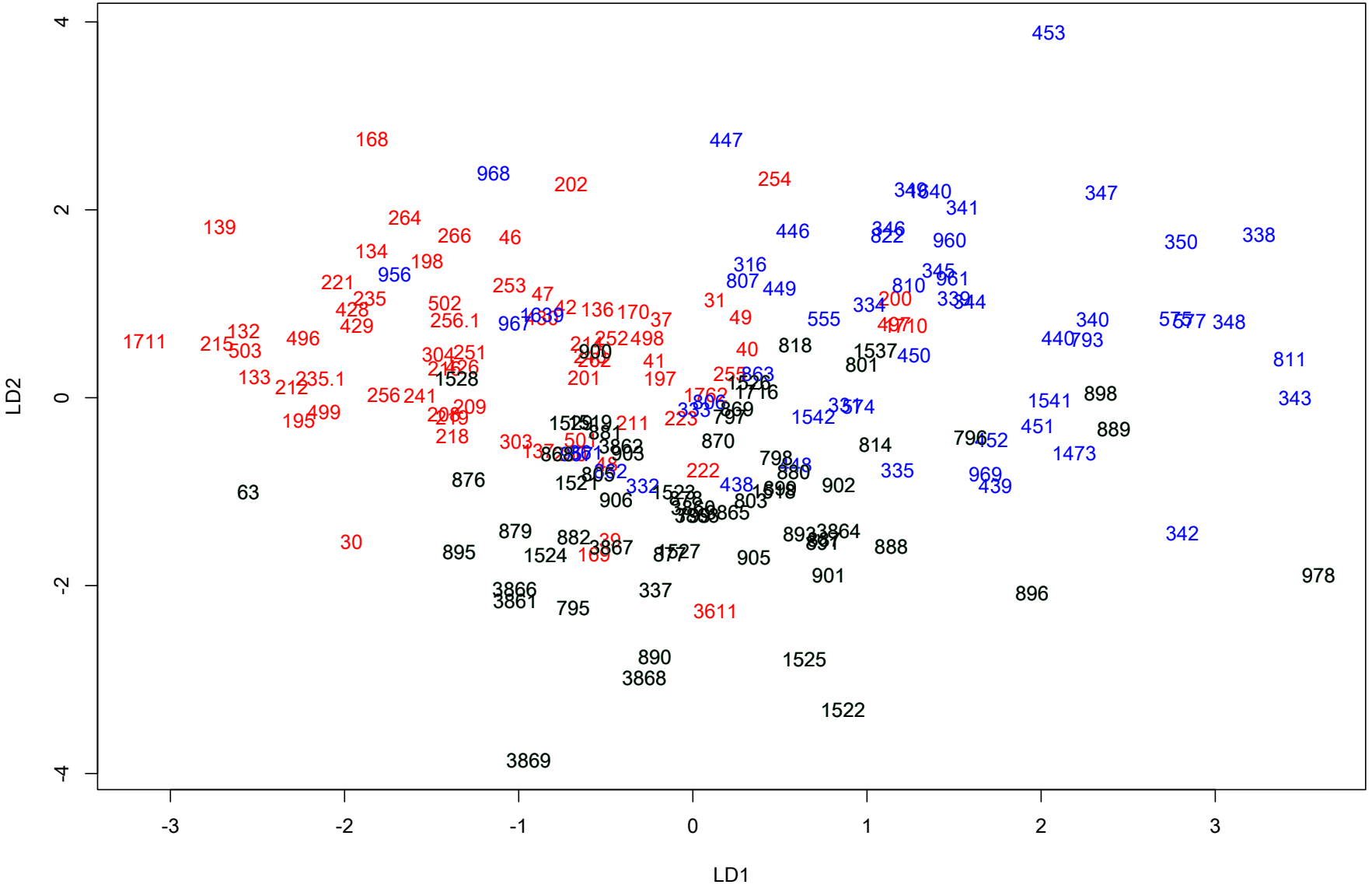
解析結果（歌人）

憶良とその他を分ける軸 →



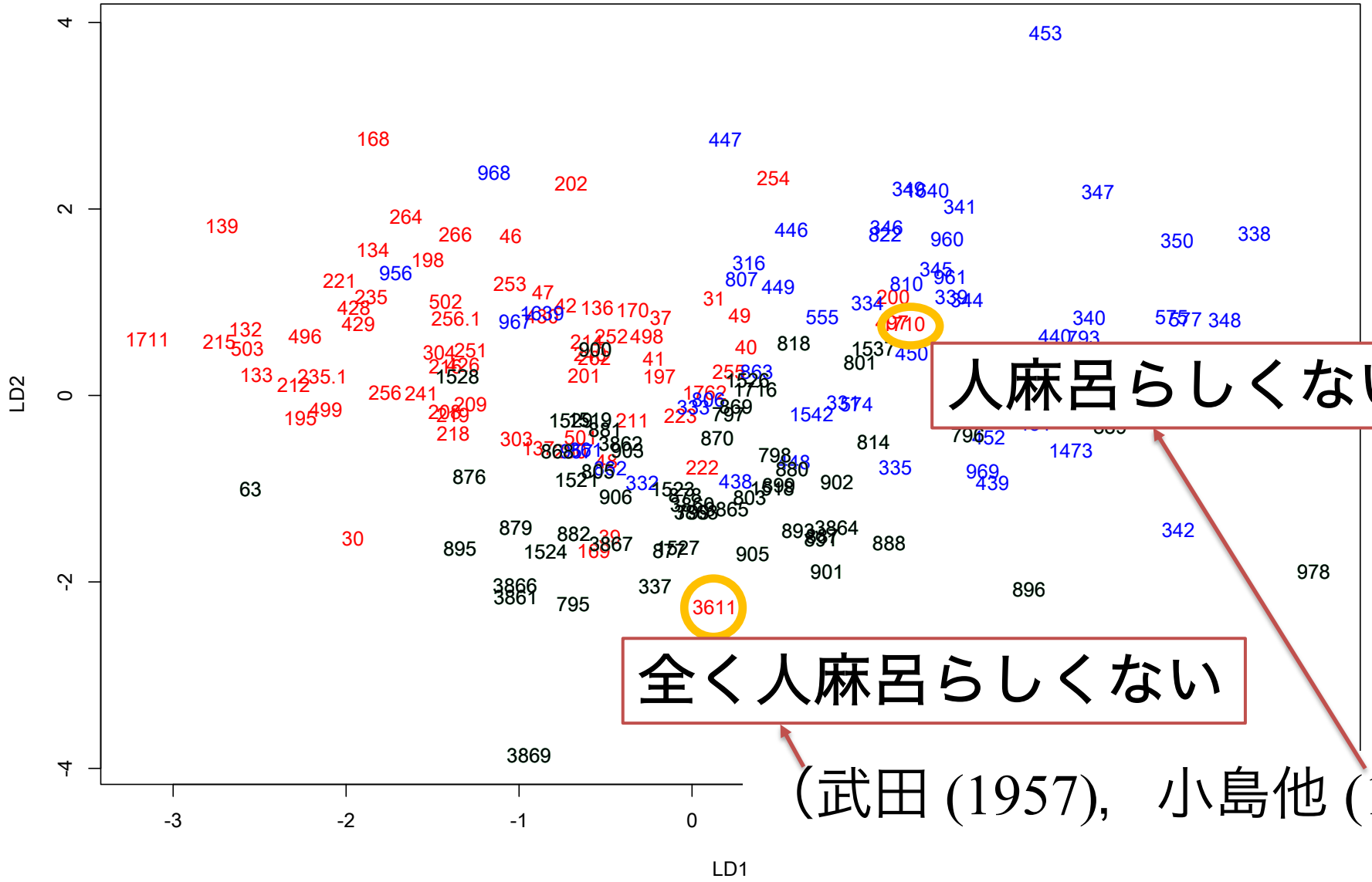
LD1 ← 人麻呂と旅人を分ける軸

解析結果（歌番号）



赤：柿本人麻呂　青：大伴旅人　黒：山上憶良

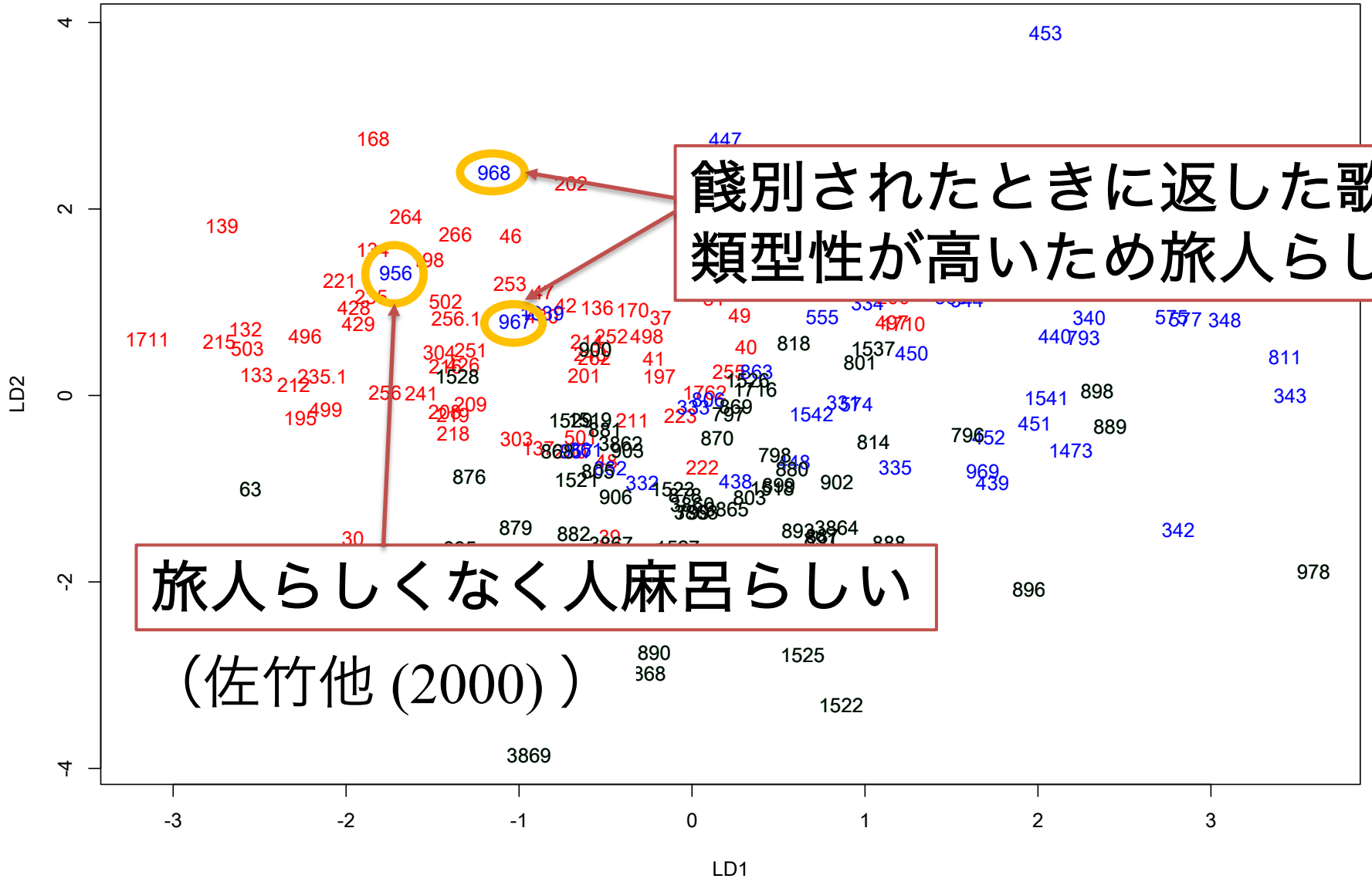
解析結果（歌番号）



（武田 (1957)，小島他 (1996) 等々）

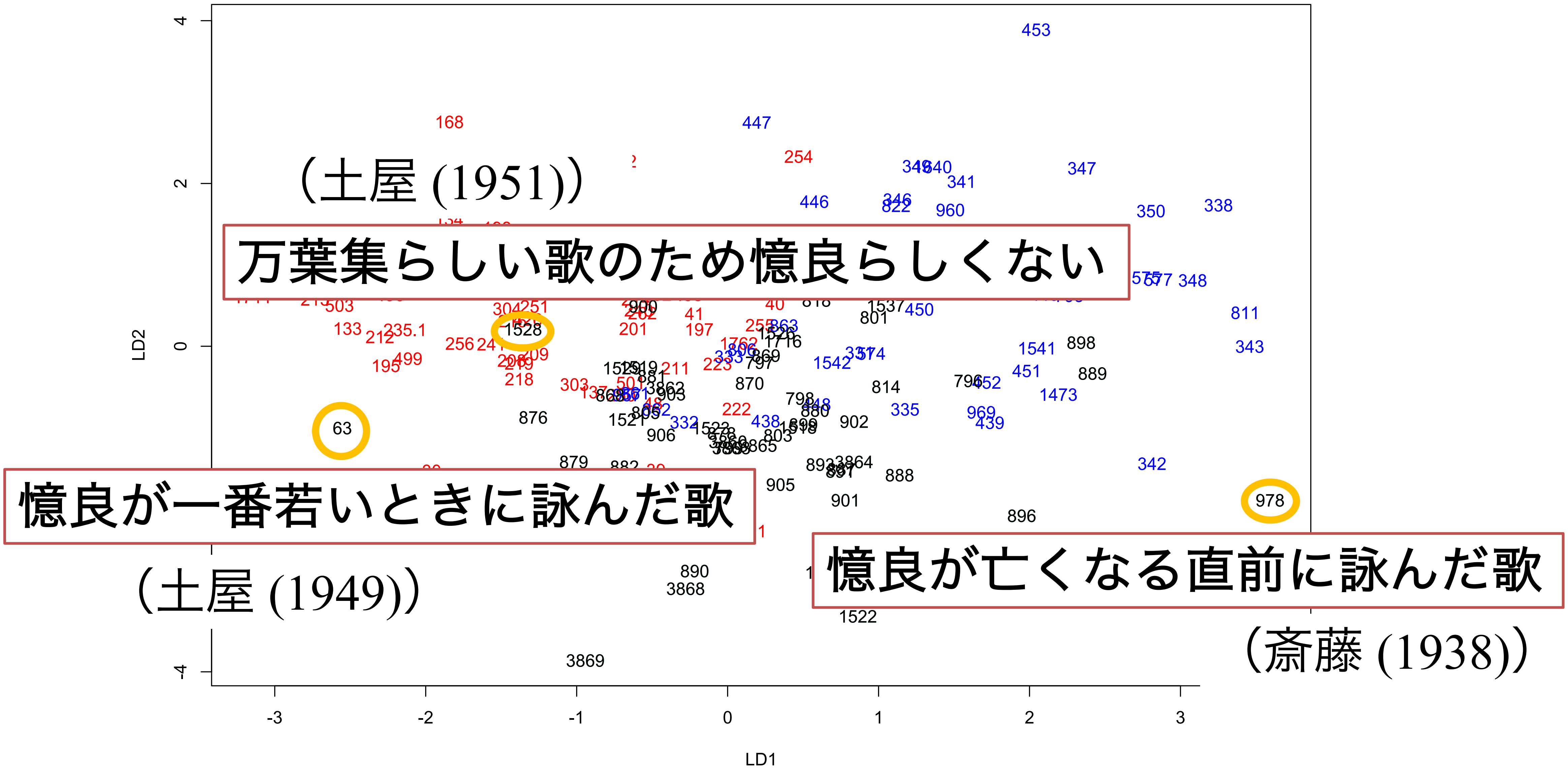
赤：柿本人麻呂　青：大伴旅人　黒：山上憶良

解析結果（歌番号）



赤：柿本人麻呂　青：大伴旅人　黒：山上憶良

解析結果（歌番号）



赤：柿本人麻呂　青：大伴旅人　黒：山上憶良

解析結果（判別軸）

大伴旅人を特徴付ける

	第一判別軸
と	0.358
な	0.289
る	0.268
し	0.242
べ	0.181
む	0.172
に	0.165

- 大久間他 (1982)
- 井村 (1984)

柿本人麻呂を特徴付ける

	第一判別軸
も	-0.120
で	-0.131
の	-0.135
ぢ	-0.148
ふ	-0.149
や	-0.161
い	-0.178
お	-0.210
ど	-0.224
ま	-0.305
み	-0.344

- 稲岡 (1985)

山上憶良を特徴付ける

	第二判別軸
げ	-0.123
き	-0.129
て	-0.134
つ	-0.135
で	-0.162
べ	-0.176
ち	-0.203
ら	-0.269
ね	-0.294
ぶ	-0.313

- 高木 (1956)

解析結果（判別軸）

	第一判別軸
き	0
く	0
こ	0
す	0
せ	0
そ	0
た	0
つ	0
ほ	0
よ	0
が	0
げ	0
ざ	0
ず	0
ぜ	0
づ	0
ば	0
び	0

	第二判別軸
や	0
へ	0
あ	0
ぞ	0
だ	0
こ	0
ほ	0
よ	0
が	0
ざ	0
ず	0
づ	0
ひ	0
	0
	0
ぐ	0
ろ	0

り	0
え	0
む	0

赤は共通している部分

歌を詠む際にごく
ありふれた音素が
0 となっている

ほとんど出てこ
ない音素も 0 と
なっている

本共同研究の始まり

2012年4月頃

- 村田先生が万葉歌をデータ化していた（整形，アノテーションもつけている状態）が，その活用方法に困っていた.
- 学内の共同研究を企画する先生に村田先生が相談.
- 企画する先生から川野に連絡があり，データ解析で困っている文学部の先生がいると説明を受ける.
- 後日，川野，村田先生，企画する先生の 3 人で集まって話し合う.
- 学内の以下の研究費を獲得して共同研究開始！

平成24-25年度 大阪府立大学異分野研究シーズ発掘・連携促進・融合領域創成支援事業「多変量解析を利用した万葉短歌の声調外在化についての研究」

本共同研究を通して学んだこと

- 共同研究開始当初は相手の言っていることが全くわからない
  - 基本的な用語のすり合わせに一年かかった
  - いまでも打ち合わせでは用語を確認している
- 研究成果を発表してもなかなか理解が得られない
  - 村田先生所属の学会ではデータと結果，川野所属の学会では統計手法だけ理解される
- 科研費の申請区分がわからない

本共同研究を通して学んだこと

- 研究で一番苦労するところは特徴量作成
  - いつもかなり長い時間議論している（打ち合わせの 8 割以上？）
  - 特徴量は粗く作ってもダメだし，細かく作りすぎてもダメ
  - 良い特徴量を作るためには良いデータが必要
- 誰も参入していない分野なので成果は出し放題
- 双方が非常にハッピーとなる成果を出すのはかなり難しい（いまはこの部分を頑張っている）
- 一番大切なことは対等な関係でたくさん議論すること

まとめと課題

- 日本古典文学分野，特に万葉歌研究に対するデータサイエンスによるアプローチを紹介した
- 万葉歌の「音」に着目し特徴量を作成した
- 作成したデータに対してスパース正準判別分析を用い判別分析と特徴選択を試みた

- 長歌の取り扱い
- 時系列情報を保持したまま特徴量を作成する方法