高次元統計モデリングユニット　山田　誠

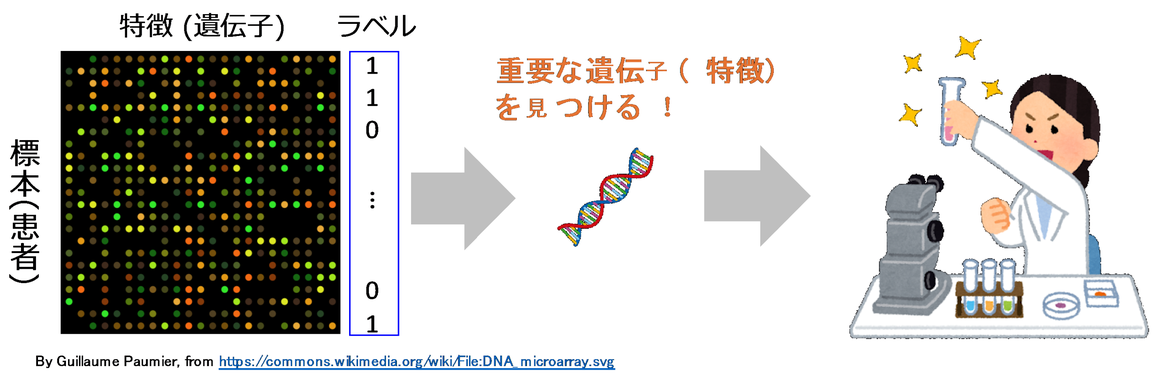
High-dimensional Statistical Modeling Unit

Makoto Yamada

高次元統計モデリングユニットの研究概要

目標

- 医療, 材料分野において**科学的発見をサポートする機械学習基盤**の構築
  - 重要な特徴をデータから精度よく容易に見つけられる手法の研究開発
  - 機械学習研究者以外でも容易に利用可能なソフトウェア開発
- 新規の科学的発見を容易にする基盤を確立し, **ヘルスケア, 材料, 農業等の分野で革新**を目指す
  - 医療費の削減 (個別化医療, 疾病予測)
  - 材料発見の大幅な効率化



By Guillaume Paumier, from https://commons.wikimedia.org/wiki/File:DNA_microarray.svg

2019年度成果

- **高次元非線形特徴選択**手法の確立

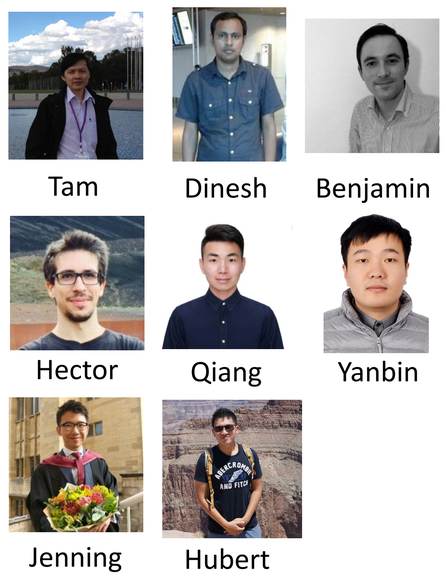
$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d_+} \frac{1}{2}\left\|\bar{\boldsymbol{L}} - \sum_{k=1}^{d}\alpha_k \bar{\boldsymbol{K}}^{(k)}\right\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

pyHSICLasso

github.com/riken-aip/pyHSICLasso

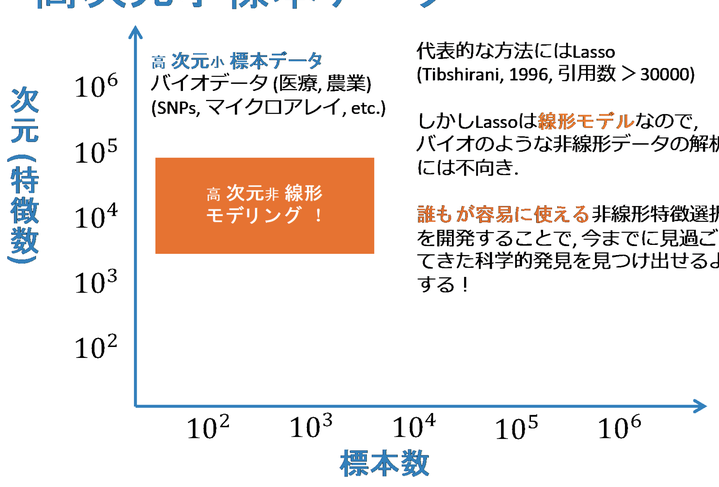
```
$ pip install pyHSICLasso
```

  - Block HSIC Lassoの提案 (ISMB 2019)
  - 教師無しHSIC Lassoの提案 (AAAI 2020)
  - HSIC Lassoの理論 (AISTATS 2020)
- **カーネル選択的推論**の確立
  - 選択的独立性検定 (AISTATS 2018)
  - 選択的二標本検定 (ICLR 2019)
  - 選択的適合度検定 (NeurIPS 2019)
  - 検出率の向上 (AISTATS 2020)
- 白血病・アトピー性皮膚炎の共同研究



高次元小標本データからの特徴選択

高次元小標本データ



代表的な方法にはLasso (Tibshirani, 1996, 引用数＞30000)

しかしLassoは**線形モデル**なので, バイオのような非線形データの解析には不向き.

誰もが容易に使える非線形特徴選択手法を開発することで, 今までに見過ごされてきた科学的発見を見つけ出せるようにする！

解釈性

- 少数特徴で高い予測性能
- 特徴の信頼度がわかる （仮説検定）

チャレンジ

- 非線形モデルは複雑　（誰も利用しない）
- 非凸最適化が利用される（最適化が難しい）

特徴選択（問題設定）

データ

●入力ベクトル・出力

$$\boldsymbol{x} \in \mathbb{R}^d,\ y \in \mathbb{R}$$

●学習データ

$$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$$

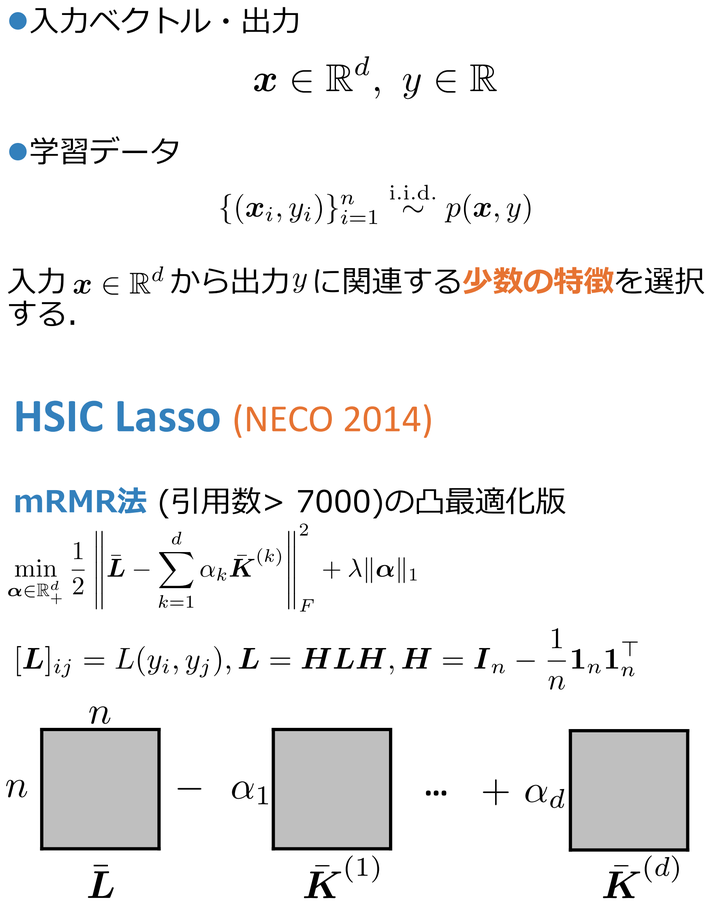
入力 $\boldsymbol{x} \in \mathbb{R}^d$ から出力 y に関連する**少数の特徴**を選択する.

HSIC Lasso (NECO 2014)

mRMR法 (引用数> 7000)の凸最適化版

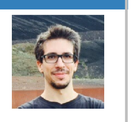
$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^d_+} \frac{1}{2}\left\|\bar{\boldsymbol{L}} - \sum_{k=1}^{d}\alpha_k \bar{\boldsymbol{K}}^{(k)}\right\|_F^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

$$[\boldsymbol{L}]_{ij} = L(y_i, y_j), \boldsymbol{L} = \boldsymbol{HLH}, \boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$



主要成果1：高次元非線形特徴選択

Block HSIC Lasso (ISMB 2019)

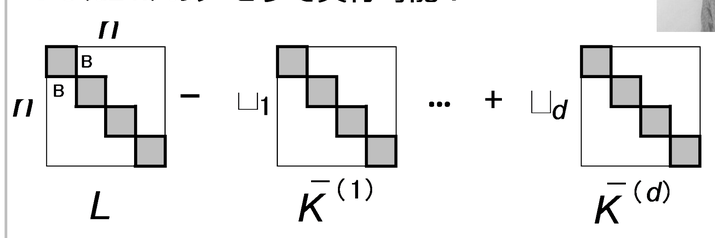


大規模HSIC Lasso

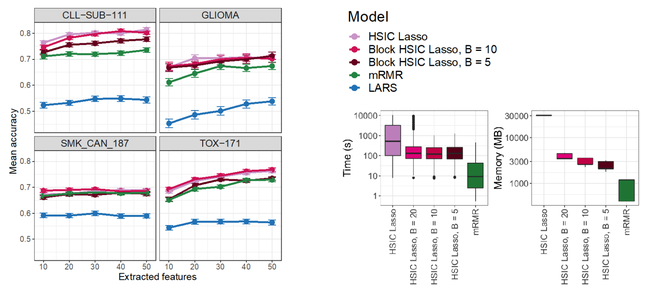
$$\max_{\alpha\in\mathbb{R}^d_+} \sum_{k=1}^{d}\alpha_k \mathrm{HSIC}_b(f_k, y) - \frac{1}{2}\sum_{k,k'=1}^{d}\alpha_k\alpha_{k'}\mathrm{HSIC}_b(f_k, f_{k'}) - \lambda\|\alpha\|_1$$

●理論保証が可能 (AISTATS 2020)

●$O(dBn)$のメモリで実行可能！



●HSIC Lasso vs. Block HSIC Lasso



Block HSIC LassoはHSIC Lassoと**同精度かつ高速**！

●**WTCCC1 datasets（ 超大規模GWASデータ ）**

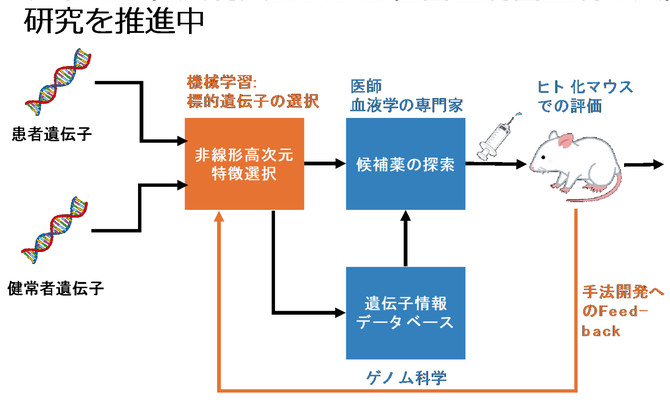
- ●RA: **352,773** markers x 3,451 cells x 2 classes
- ●T1D: **352,853** markers x 3,443 cells x 2 classes
- ●T2D: **353,046** markers x 3,456 cells x 2 classes

●**実験結果（ 2 値分類）**

データ	特徴数	サンプル数	Raw	LARS	HSIC Lasso
RA	352,773	3,451	0.671 ± 0.002	0.572 ± 0.002	**0.767 ± 0.004**
T1D	352,853	3,443	0.671 ± 0.006	0.569 ± 0.004	**0.788 ± 0.002**
T2D	353,046	3,456	0.609 ± 0.004	0.565 ± 0.005	**0.675 ± 0.003**

共同研究

●アトピー性皮膚炎および急性白血病白血病の共同研究を推進中



主要成果2：カーネル選択的推論

選択的推論

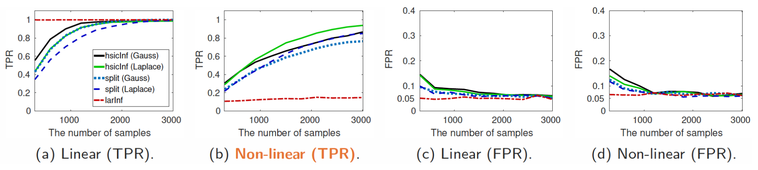
検定する項目を選択した後で, 選択した項目に関する検定を行う枠組み.

カーネル法に基づいた選択的推論の確立

- ●選択的独立性検定　(AISTATS 2018)
- ●選択的二標本検定　(ICLR 2019)
- ●選択的適合度検定　(NeruIPS 2019)
- ●検出力の向上　(AISTATS 2020)

HSIC Inference (hsicInf)

$$H_{0,m}: \ \mathrm{HSIC}(X_m, Y) = 0 \mid \mathcal{S} \text{ was selected},$$
$$H_{1,m}: \ \mathrm{HSIC}(X_m, Y) \neq 0 \mid \mathcal{S} \text{ was selected}.$$



(a) Linear (TPR). (b) Non-linear (TPR). (c) Linear (FPR). (d) Non-linear (FPR).

線形手法では見つけられない特徴を選択可能！

その他成果

機械学習基礎

- ベイズ最適化（ECAI 2020）
- 木構造Wasserstein（NeurIPS 2019）
- GNN近似度（NeurIPS 2019）
- 相互情報量×最適輸送（投稿中）
- GraphLIME (投稿中)
- Feature Selection Network (投稿中)

機械学習応用

- グラフェン層数識別（npj Comp. Materials）
- 自然言語処理（EMNLP 2019）

まとめ

●高次元非線形特徴選択

●非線形選択的推論アルゴリズム

●共同研究（アトピー性皮膚炎・白血病）

今後の予定