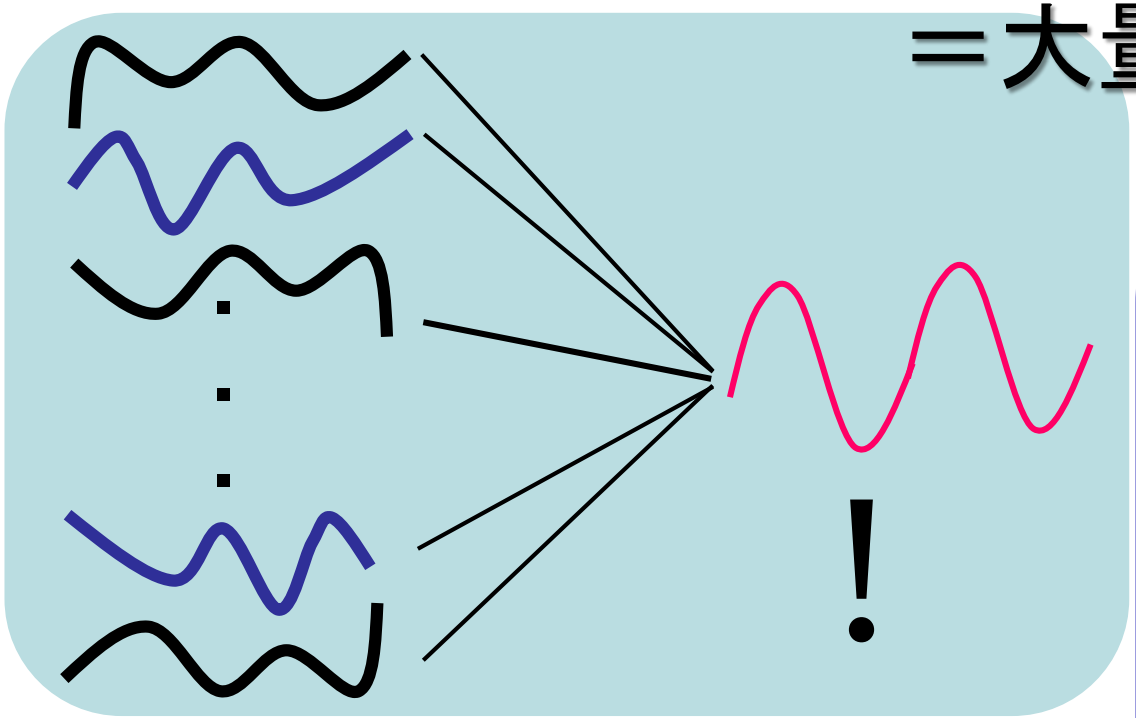
教師なしAIが拓く ゲノム解析の新時代

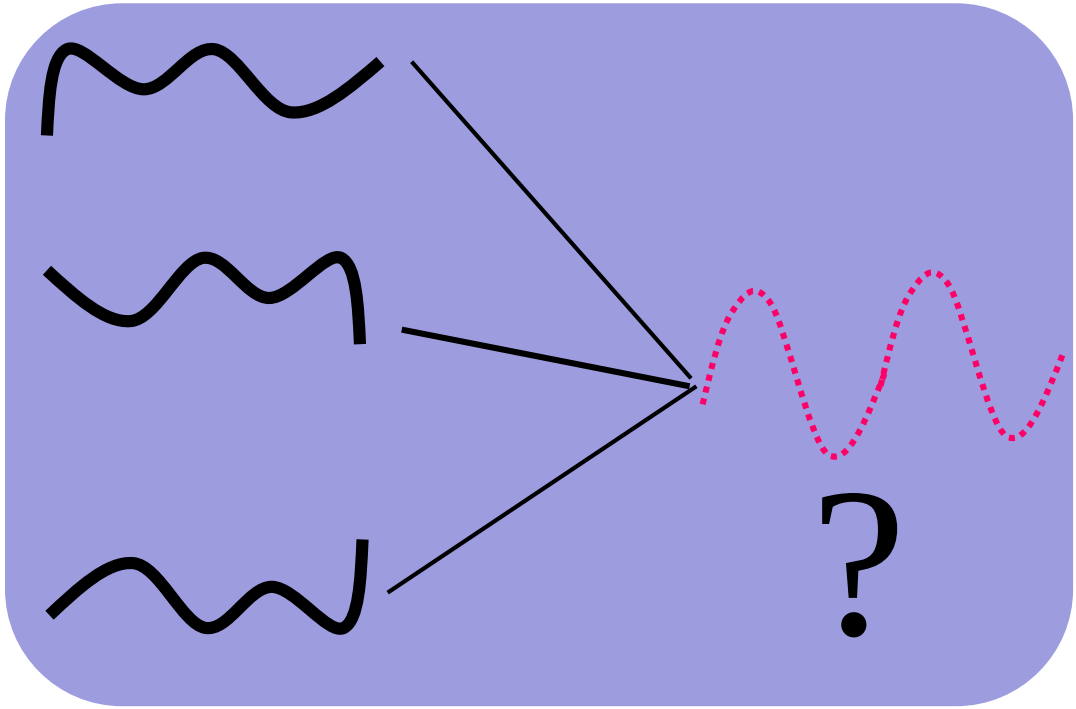
中央大学　理工学部　物理学科
教授　田口　善弘

2019年９月１２日

従来技術とその問題点

データサイエンス（機械学習）
＝大量の教師データ



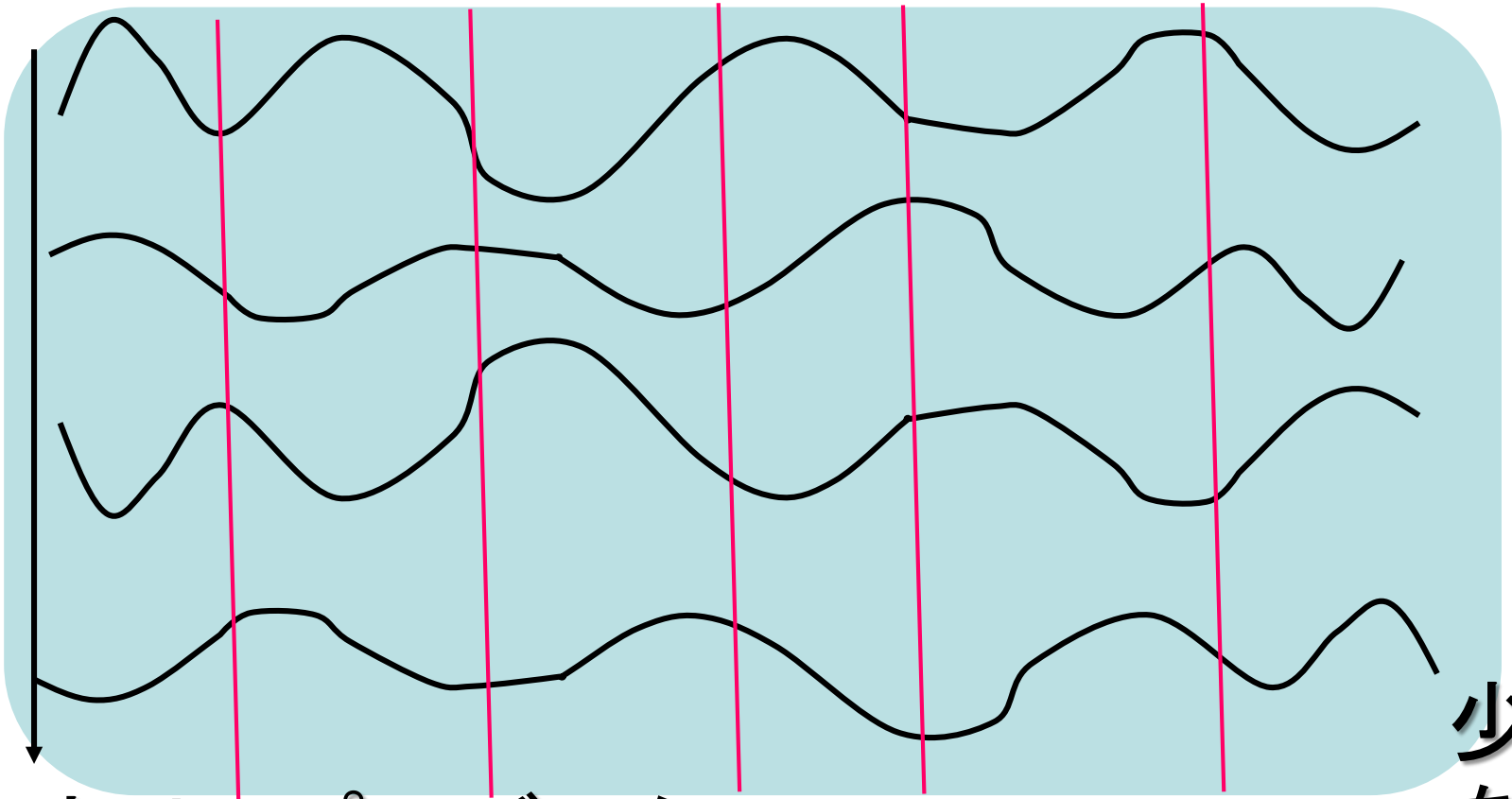


・計測に費用がかかる医学・生物学データ
・レアイベント（故障など）　　⇒適用不可

新技術の特徴･従来技術との比較

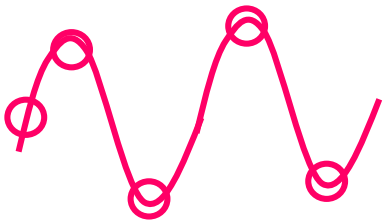
従来技術の問題点であった、大量のデータを必要とするという点を改良した。

高次元



少サンプルデータ

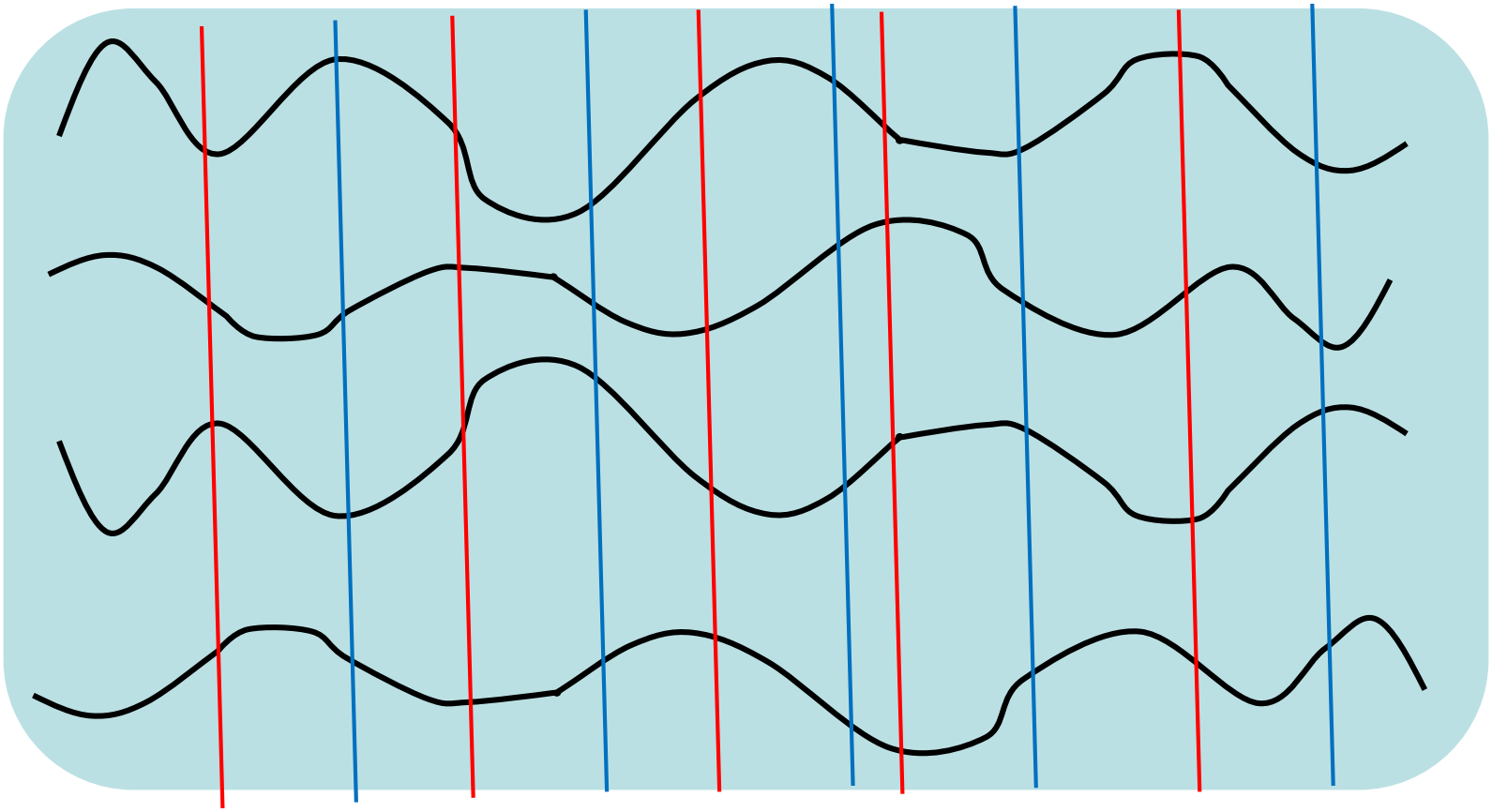
教師無し学習

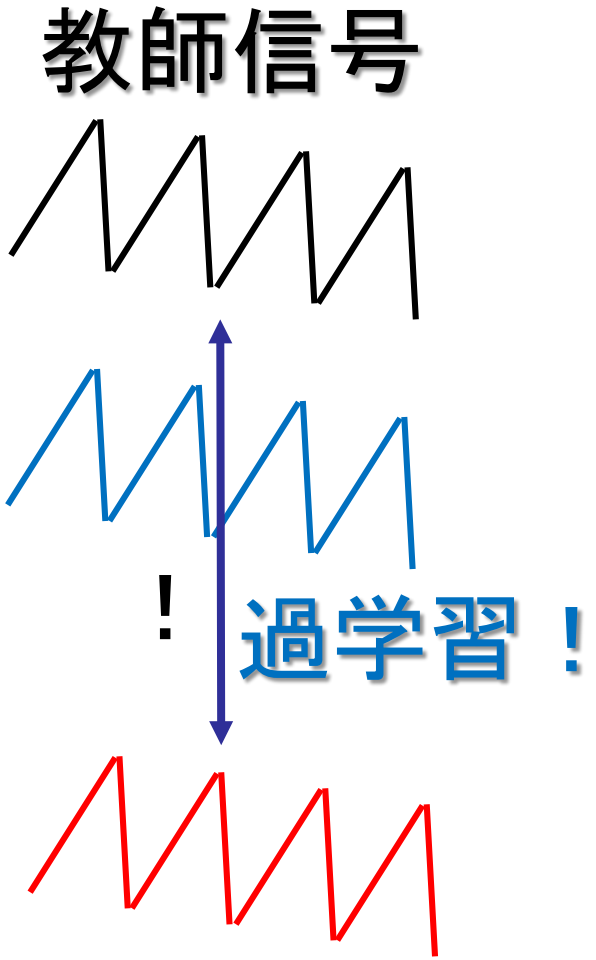


少数の秩序変数を発見的に選択

利点①：教師無し学習なので過学習が起きにくい、あるいは原理的に起きない。

教師あり学習による秩序変数選択

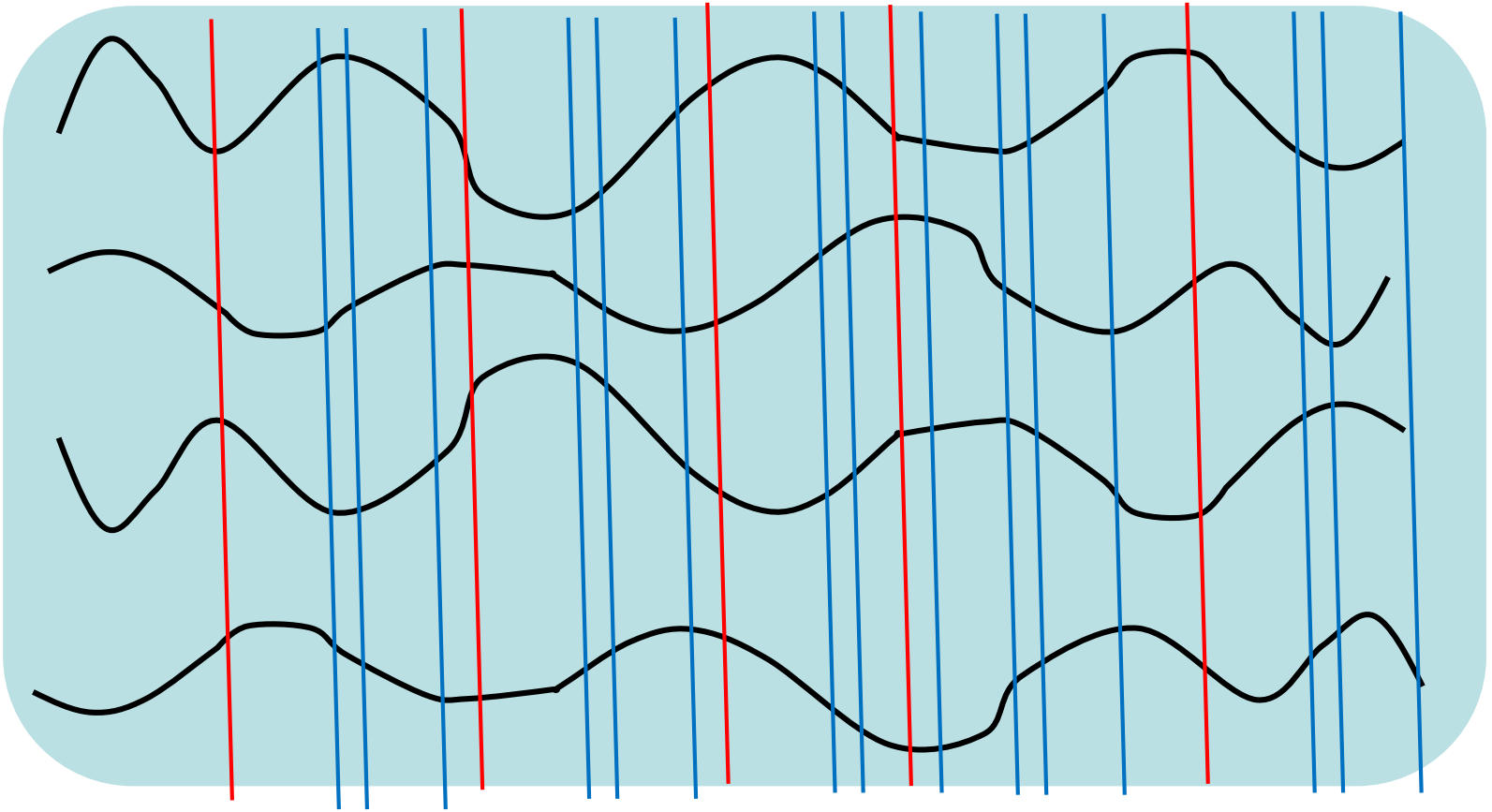




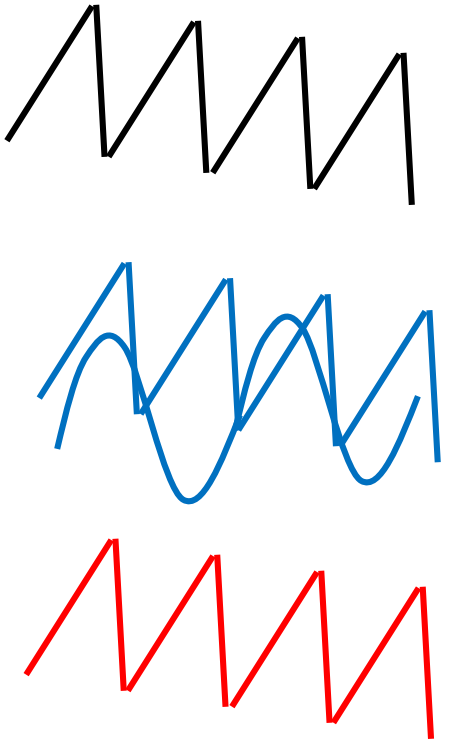
教師無し学習による秩序変数選択

利点②:
繰り返し学習のプロセスが無いので計算時間が短い。

教師あり学習による秩序変数選択



教師信号



教師無し学習による秩序変数選択

人工データによる例

新技術説明会
New Technology Presentation Meetings!

正規 μ :平均
分布 ½ :標準偏差

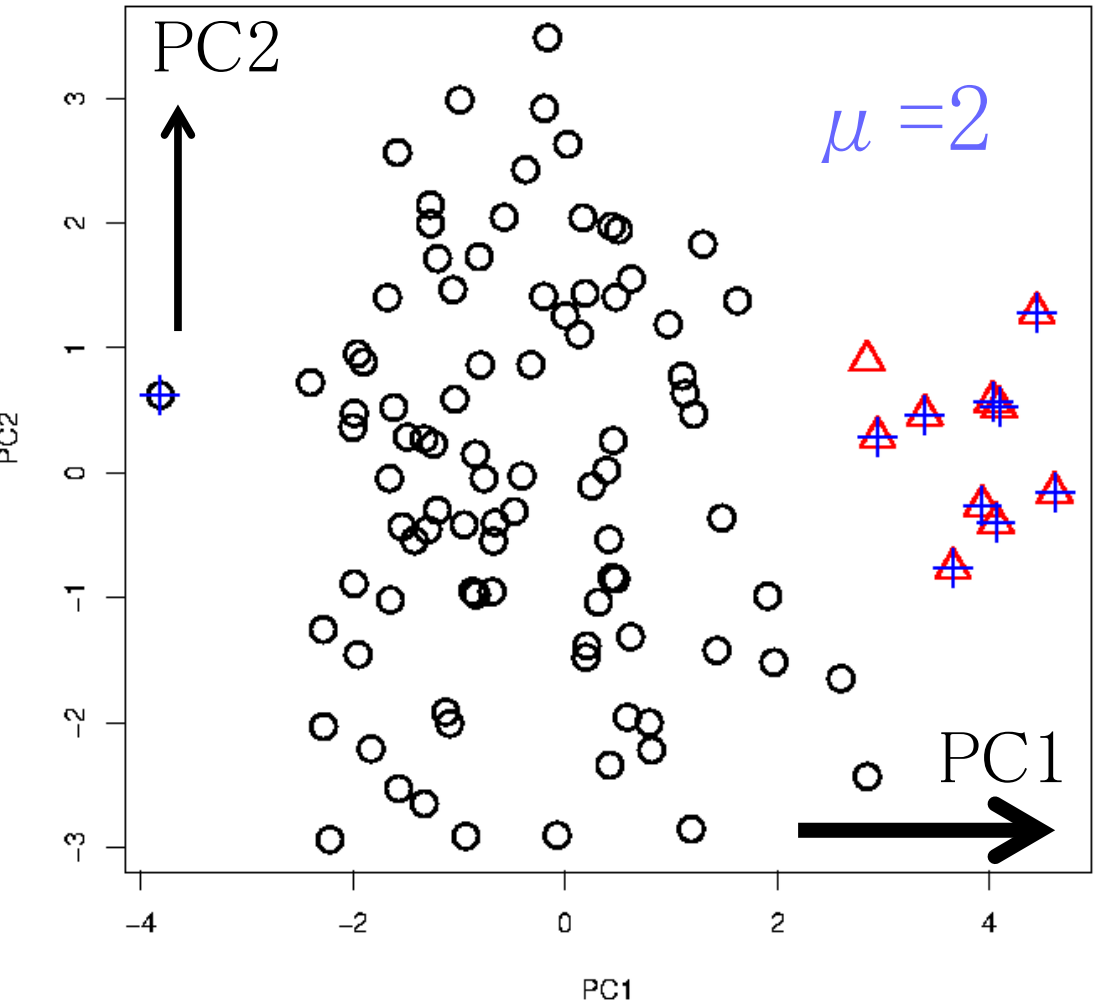
$[N(\mu,1/2)+N(0,1/2)]/2$

90 変数

10変数

10サンプル $N(\mu,1/2)$

10サンプル $N(0,1/2)$



+:上位 10外れ値

従って、教師無し学習で二群で差がある10変数を選べる
Accuracy:(100 trials)
89.5% (μ =2)
52.6% (μ =1)

データを10次元のベクトル100本とみて主成分分析で平面に射影。

想定される用途

- ゲノムデータの場合、変数の次元が遺伝子の数と同等で数千から数万である一方、サンプルする数は数個から数百個であり従来手法では無力だが、提案手法では有効に変数選択が可能。
- 上記以外に、計算時間が短かいことが期待できる。
- 過学習も起きにくい。

応用事例1：筋萎縮性側索硬化症（ALS）のバイオマーカー探索

典型的な高次元小サンプル問題！

新技術説明会
New Technology Presentation Meetings!

ALSの患者と健常者（計53名）の血清中の3000＋α個のマイクロRNAから100個以下（数%）選んで判別

予測	正解 変異保持者	家族性ALS	健常者	孤発性ALS
変異保持者	8	1	0	5
家族性ALS	2	5	0	0
健常者	1	0	16	8
孤発性ALS	1	0	1	5

教師あり学習（ANOVA）による選択

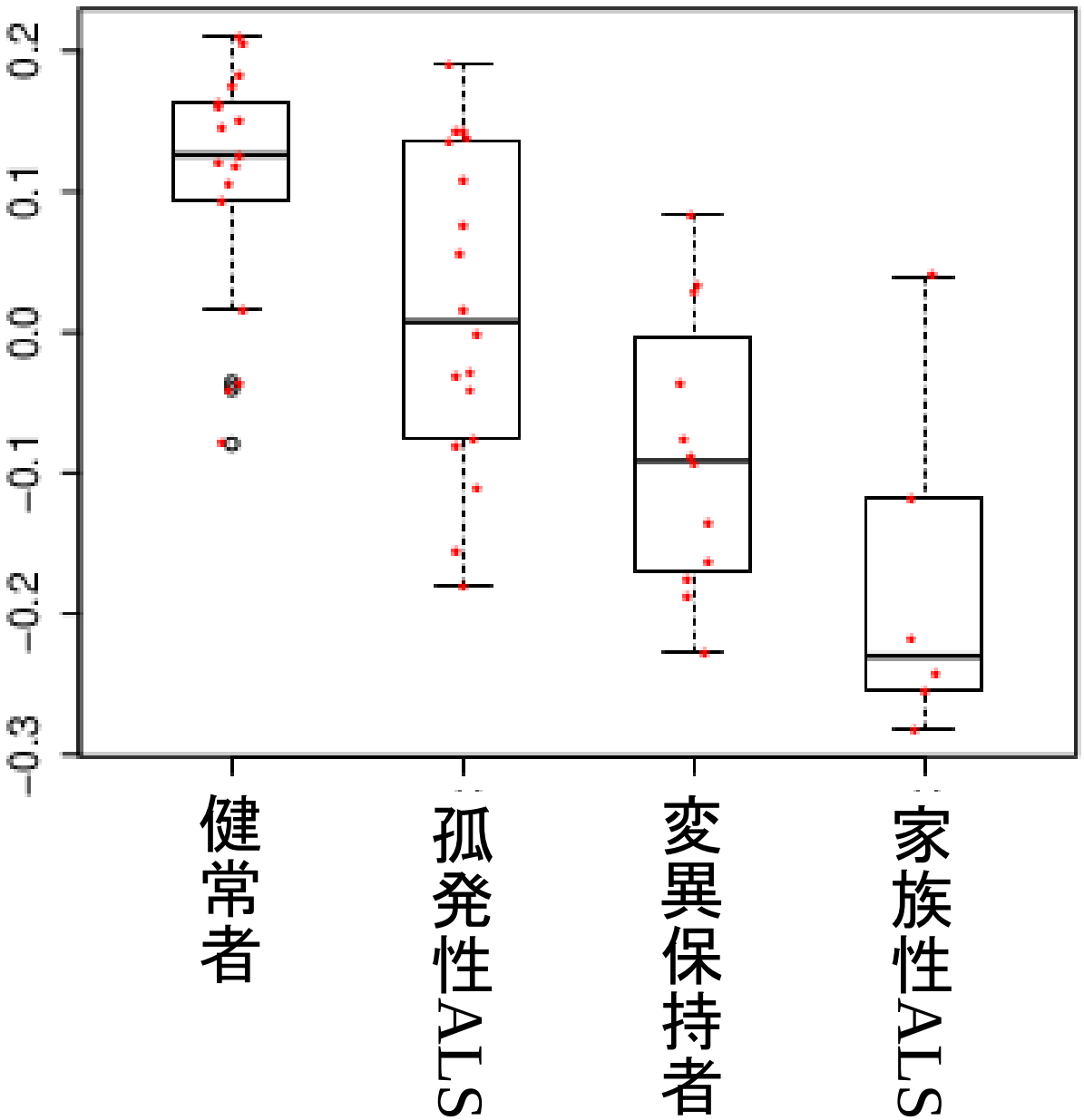
予測	正解 変異保持者	家族性ALS	健常者	孤発性ALS
変異保持者	8	1	0	2
家族性ALS	2	5	0	1
健常者	0	0	14	7
孤発性ALS	2	0	3	8

教師無し学習による選択

孤発性ALSの判別が大きく向上！

教師無し学習の重要性

教師なし学習が示唆する選択基準



健常者＞孤発性ALS＞
変異保持者＞家族性ALS
という非自明な順に発現が
大きいマイクロRNAを選ぶ
という重要性。

孤発性ALSにおいては、
教師あり学習ではペナルティ
になる、「群内変異が大きい
マイクロRNA」を選ぶ必要性

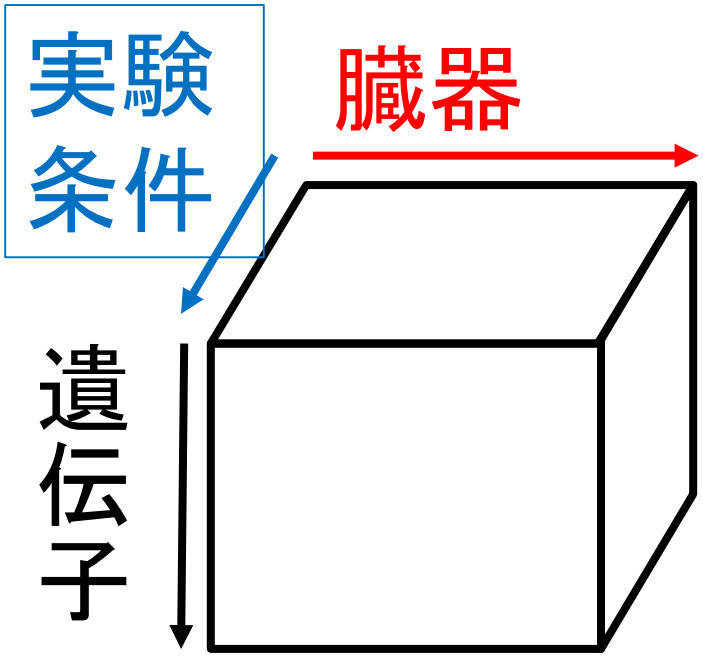
参考文献:

田口善弘,王 秀瑛「筋萎縮性側索硬化症のためのマイクロRNAバイオマーカーの探索」
情報処理学会研究報告バイオ情報学,
2018-BIO-56, No.2, PP.1 - 6.
http://id.nii.ac.jp/1001/00192710/

Y-h. Taguchi and Hsiuying Wang,
“Exploring microRNA Biomarker for Amyotrophic Lateral Sclerosis”, (2018)
Int. J. Mol. Sci. 2018, 19(5), 1318.
https://dx.doi.org/10.3390/ijms19051318

応用事例2:
心的外傷後ストレス障害（PTSD）由来の心臓疾患の原因遺伝子の推定

帰還兵などがPTSDと同時に心疾患を発症しやすいことはSoldier’s heartなどと言って有名。



コントロール群 → 5，5 ← ストレス群

ストレス期間（単位：日）	5		10			5		10	
休憩期間	24h	1.5 w	24h	6w		24h	1.5 w	24h	6w
扁桃体	3,2	5,4	3,4	3,4	海馬	3,5	4,5	5,4	4,5
内側前頭前皮質	4,5	5,5	3,4	4,4	中隔核	3,2	2,3	3,3	3,3
線条体	5,5	5,5	5,4	4,4	腹側線条体	5,5	5,5	3,4	5,4
血液	5,5	5,5	4,5	4,5	心臓	5,5	4,5	5,5	5,5
半脳	5,5	4,5	5,5	5,5	脾臓	5,5	5,5	5,4	5,5

典型的な高次元小サンプル問題！

ストレス期間	10日		5日	
休憩期間	24時間	6週	24時間	1．5週
扁桃体		○		○
海馬		○	○	○
内側前頭前皮質		○		
心臓	○			○
半脳			○	○
脾臓		○	○	○

実験条件特異的、かつ、臓器特異的にコントロール群とストレス群で発現差がある共通遺伝子群の特定に成功

教師あり学習ではそもそも、こういう複雑な臓器と実験条件の組み合わせで発現している遺伝子をさがす、ということを目指すことさえできない….。

参考文献：

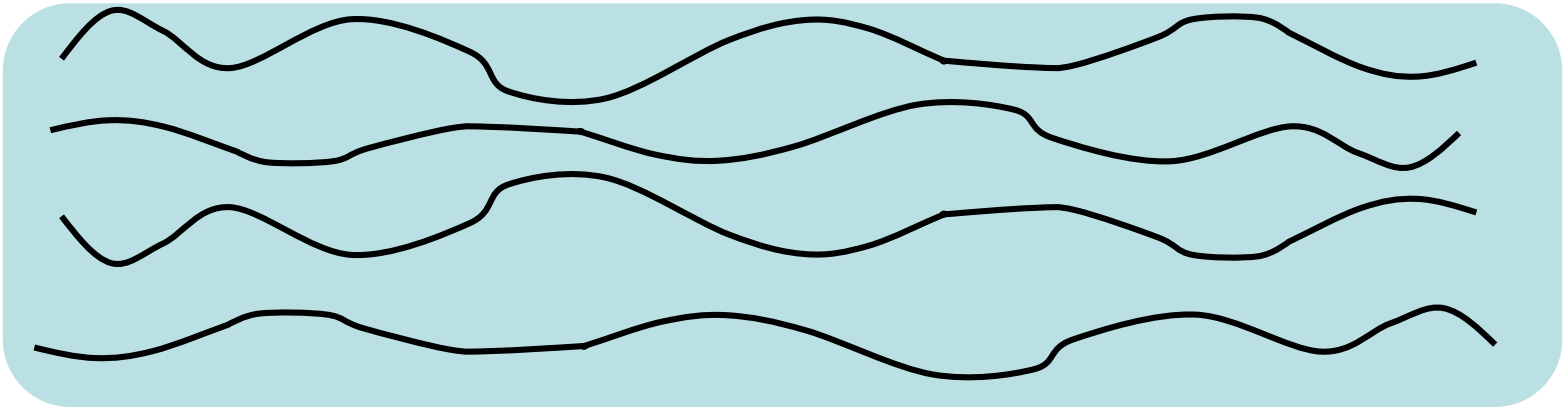
田口善弘,「テンソル分解を用いた教師なし学習による心的外傷後ストレス障害由来の心臓病原因遺伝子の同定」情報処理学会研究報告バイオ情報学,2017-BIO-51,No.1, PP.1 - 8
http://id.nii.ac.jp/1001/00183531/

Y-h. Taguchi, “Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases”,
BMC Medical Genomics, 10 (Suppl 4) :67 (2017)
https://doi.org/10.1186/s12920-017-0302-1

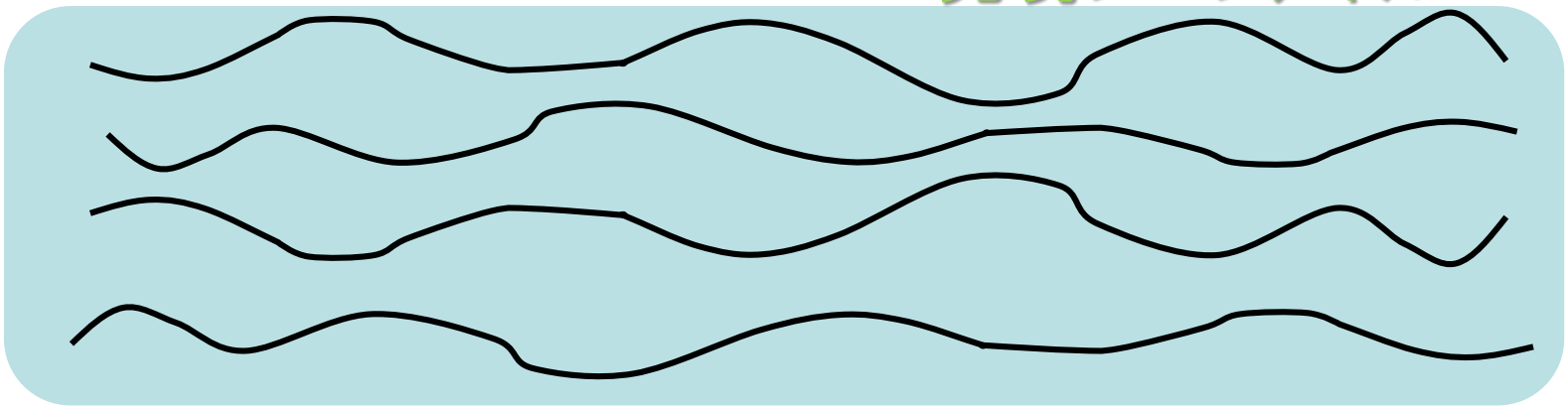
応用事例3：
遺伝子発現プロファイルからの創薬研究

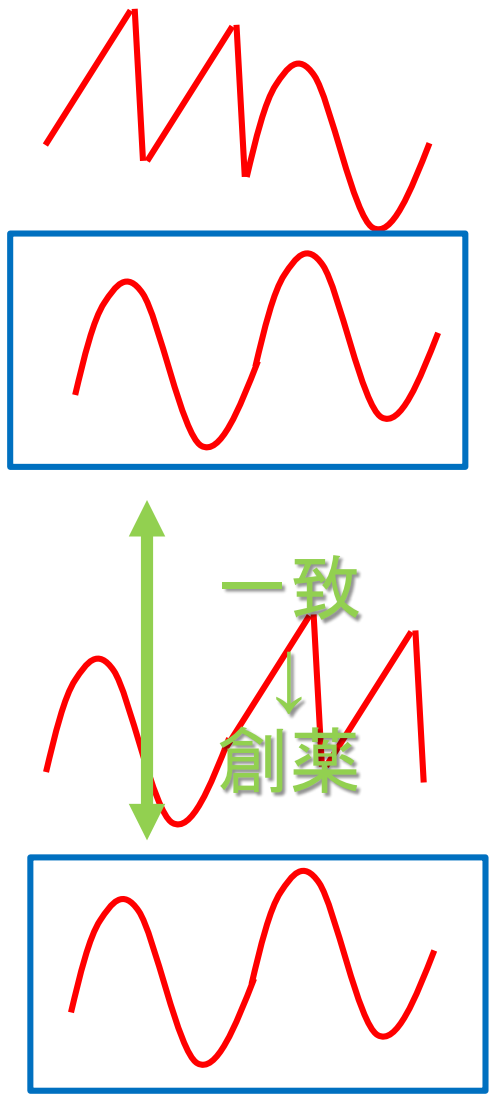
教師無し学習

データセットA → 患者対健常者遺伝子発現プロファイル



データセットB → 化合物投与後遺伝子発現プロファイル





教師無し学習で2つのデータセットから共通パターンを抽出可能

列:予測化合物
行:既知化合物

		Single Gene Perturbations from GEO up				
		F	T	P_F	P_{χ^2}	RO
心臓疾患	F	521	517	3.4×10^{-4}	3.9×10^{-4}	3.02
	T	13	39			
PTSD	F	500	560	3.8×10^{-2}	3.1×10^{-2}	2.67
	T	6	18			
急性白血病	F	979	89	2.7×10^{-1}	3.0×10^{-1}	2.19
	T	10	2			
糖尿病	F	889	177	1.2×10^{-2}	7.1×10^{-3}	3.00
	T	15	9			
腎臓がん	F	847	219	2.0×10^{-2}	1.2×10^{-2}	2.75
	T	14	10			
肝硬変	F	572	219	1.1×10^{-2}	8.1×10^{-3}	2.91
	T	8	10			

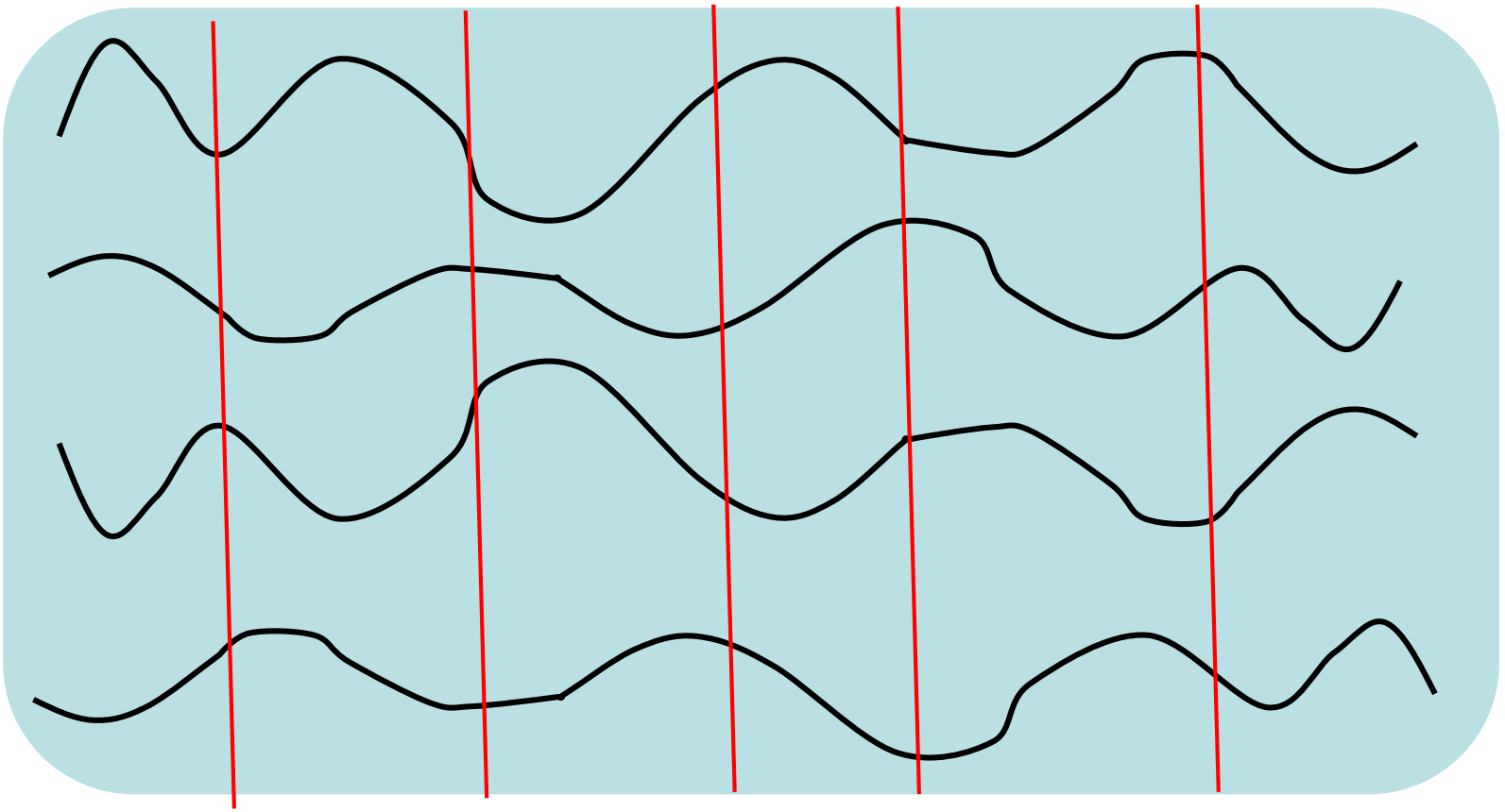
参考文献:

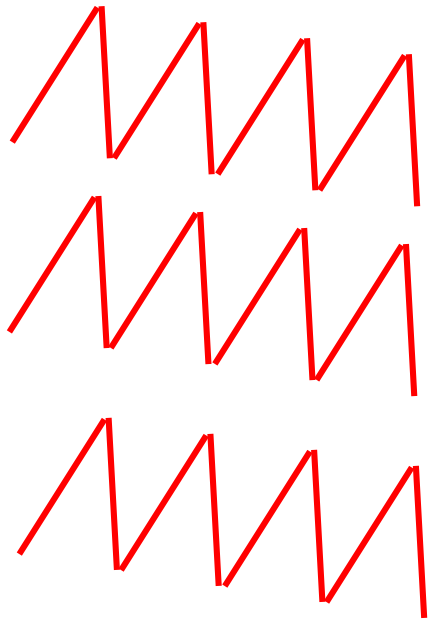
田口善弘「疾患とDrugMatrixデータセットとの間の遺伝子発現の統合解析におけるテンソル分解を用いた教師なし学習による変数選択を用いた候補薬物の同定」
情報処理学会研究報告バイオ情報学, 2018-BIO-55,No.1,PP.1 - 6. http://id.nii.ac.jp/1001/00191249/

Y-h. Taguchi, "Identification of candidate drugs using tensor-decomposition-based unsupervised feature extraction in integrated analysis of gene expression between diseases and DrugMatrix datasets"
Scientific Reports,7,13733 (2017)
https://www.nature.com/articles/s41598-017-13003-0

利点③:
安定した変数選択が可能

教師無し学習

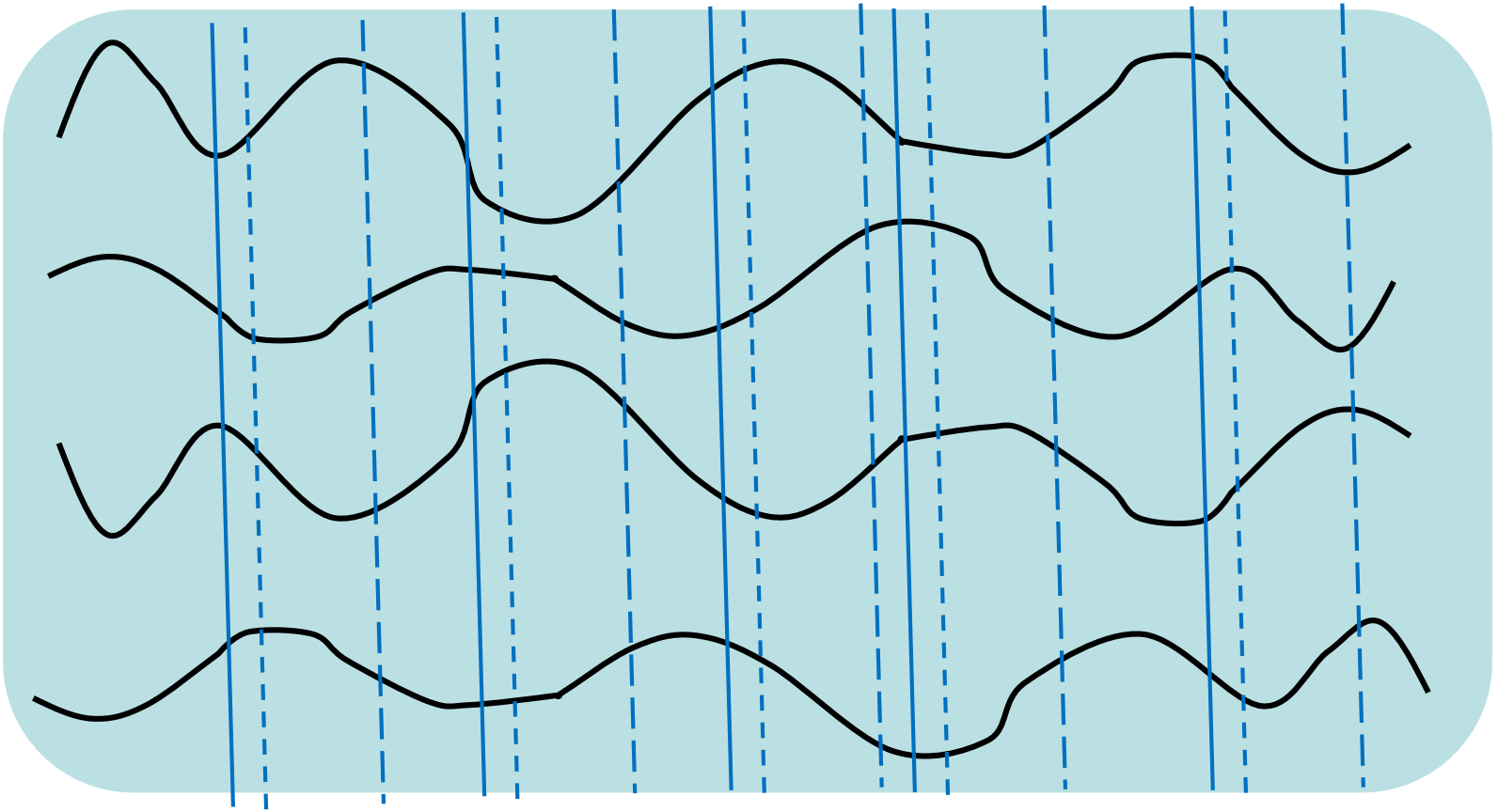


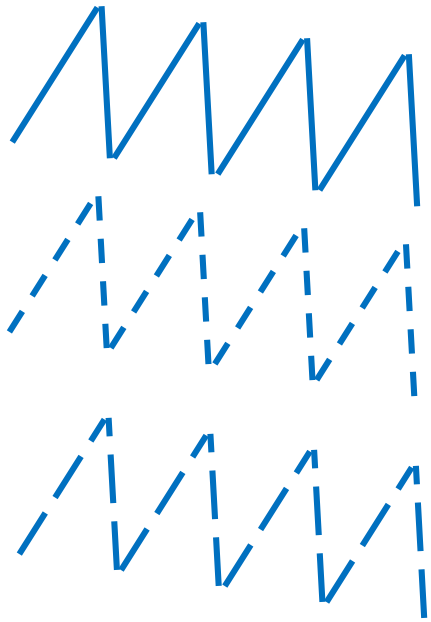


どのサンプルの組み合わせでも同じ変数選択がなされる。

教師あり学習では不安定な変数選択

教師あり学習





サンプルの組み合わせで異なった変数選択＝不安定

応用事例4：
血清中マイクロRNAを用いた安定なバイオマーカー選択

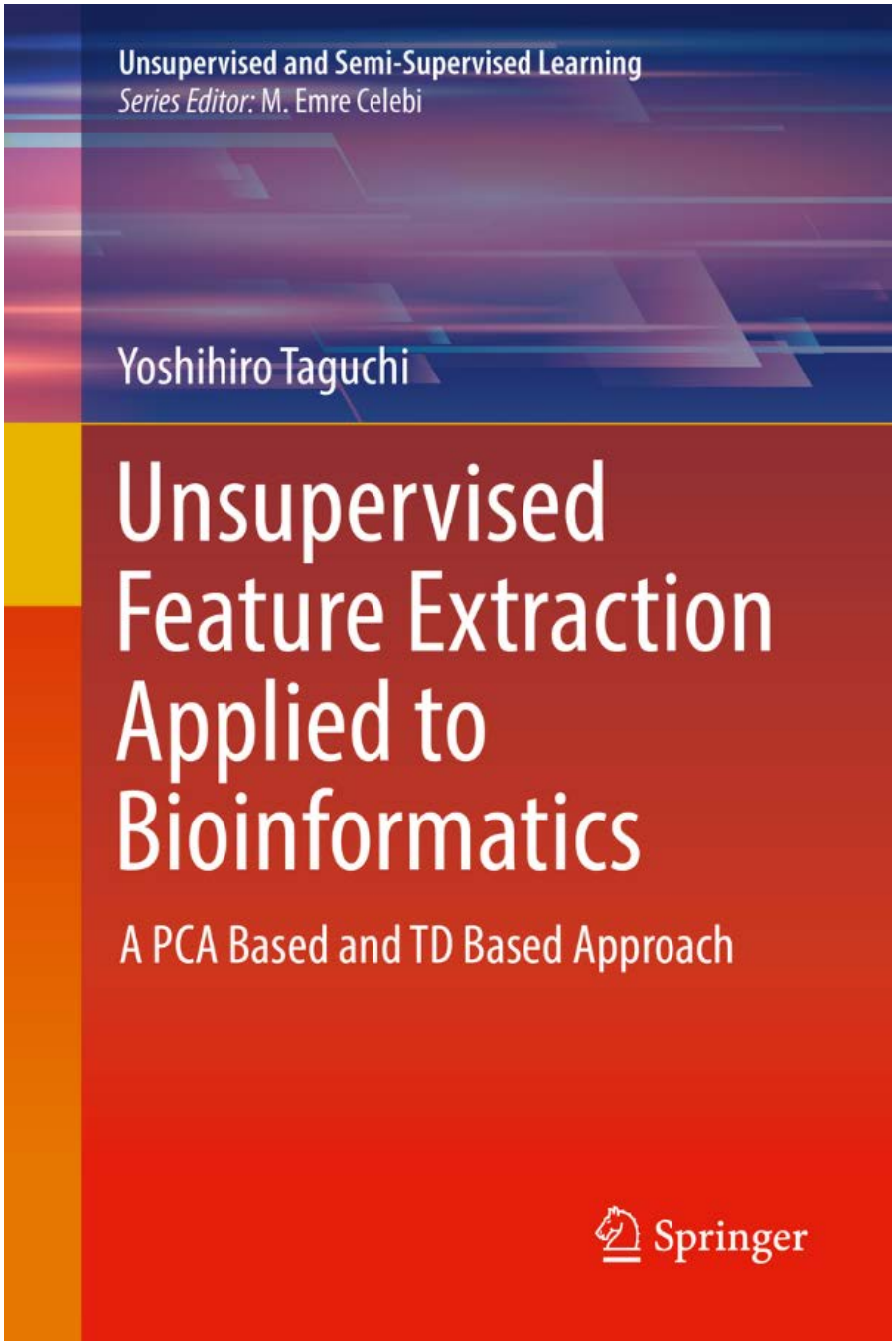
14種類の疾患を10個の血清マイクロRNAで判別
全サンプルの90%を選択した時の10種類のマイクロRNAの選択安定性をチェック

10マイクロRNA×14疾患＝140マイクロRNA
100回トライ
提案手法：140個中129個が100%選択
教師あり学習：0～40個しか100%選ばれない。
他の教師無し学習：140個中111個が100%選択

教師無し学習の変数選択安定性は圧倒的！

参考文献:

Y-h. Taguchi, Y. Murakami (2013) Principal Component Analysis Based Feature Extraction Approach to Identify Circulating microRNA Biomarkers. PLoS ONE 8(6): e66714. https://doi.org/10.1371/journal.pone.0066714



Springerから教科書を出版予定
2019/9/13刊行予定

https://www.springer.com/jp/book/9783030224554

実用化に向けた課題

- 検証するための実験データが不足しているために、成果がアピールできない。
- 研究成果をアピールできる実験データの準備

企業への期待

- 解析できるデータの提供をお願いしたい。社外にデータを持ち出すことに対する抵抗が大きく、僕自身がデータ解析に携われることが少なく、協業が難しい。
- 僕自身がデータを触れない場合は、社内に僕の技術を習得できる人材（数理やコンピュータがある程度できる人材）を準備してその人に教えることで協業が可能になる。
- 対価として研究費の提供

本技術に関する知的財産権

- 現在、準備中です。

産学連携の経歴

- 2018年　A社（国内大手製薬企業）と共同研究実施
- 2019年　B社（某企業製薬部門）と個人コンサル契約、技術移転。
- 2019年　某製薬ベンチャーと技術移転交渉中

お問い合わせ先（必須）

中央大学　研究推進支援本部

ＴＥＬ　０３－３８１７－１６０３
ＦＡＸ　０３－３８１７－１６７７
e-mail　clip@tamajs.chuo-u.ac.jp