

予測精度の高い、高速インシリコスクリーニングを目指した 「テーラーメイド・モデリング」

“Tailor made modeling”, “Personalized modeling”
for high speed and prediction of In Silico Screening

湯田 浩太郎
Kohtaro Yuta

In Silico Data, Ltd.

データ解析上での4種類の指標

分類率：既知データの再現性

予測率：未知データの予測性

予測信頼性：予測結果の信頼度

予測精度：予測の誤差レベルと傾向

◇分類率や予測率低下の原因となる要因

① サンプル数の問題

サンプル数増大



分類率、予測率低下

② サンプル母集団の特性上での問題

サンプルの分散／重複が大



分類率、予測率低下

③ 分類手法上での問題

分類に不適切な分類手法の利用



分類率、予測率低下

④ 分類パラメータ上での問題

分類能力の小さなパラメータ



分類率、予測率低下

◇分類率、予測率向上への4大アプローチ

① サンプル数の問題

サンプル数減少  サンプルのサブセット化

② サンプル母集団の特性上での問題

サンプルの分散の減少  サンプルのグルーピング

③ 分類手法上での問題

より強力な分類手法の利用
 線形から非線形分類へ

④ より強力なパラメータの利用

分類パラメータの変更  二次元から三次元へ

◇サンプル母集団のサブセット化上での留意点

□サブセット化の目的

サンプルセットを小さくすることで、予測モデルの信頼性と予測率を高く保つことが期待される。

□サブセット化実施上での留意点

サブセットにアサインする時の曖昧性の解決

- ・一元多項対応から、一元一項対応の実現

◇サンプルサブセット構築例

●一般的事例

◇化合物構造単位で分類

・例：脂肪族化合物群、芳香族化合物群、多環化合物群、その他



◆問題点

- ・分類が一義に定まらない化合物の処理が困難
分類基準の複数にまたがる化合物が多く存在する
分類を間違うと、不適切な予測モデルが適用されるために、
予測信頼性が低下する。

◇ECOSAR(EPA)による環境毒性予測

一元多項対応が原因となる予測信頼性の低下

同じ予測対象サンプルであっても、
選択されたサブセット単位で予測結果が大きく異なる

● Daphnid 96-hr LC50

Aromatic Amines: **95.502**

Esters: **41.180**

Phenols: **42.503**

Ecosar v0.99e

File Edit Functions BatchMode ShowStructure SpecialClasses Help

Previous Get User Save User CAS Input Calculate

Enter SMILES:

Enter NAME:

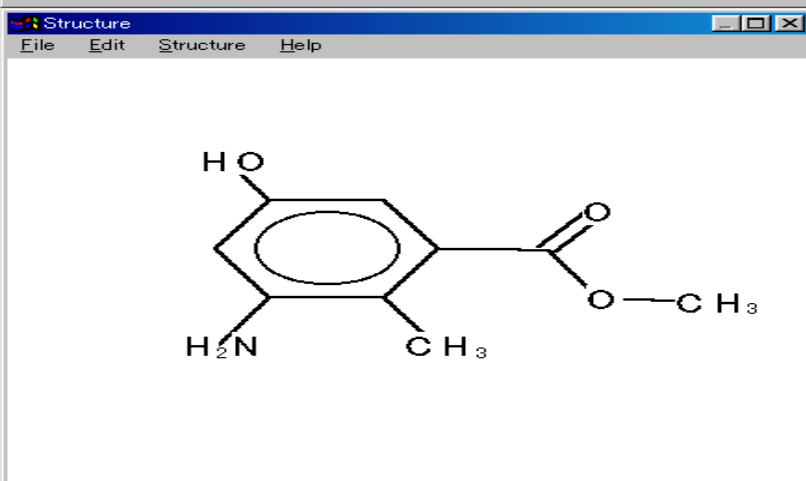
CAS Number:

Chemical ID 1: Measured Water Sol (mg/L):

Chemical ID 2: Melting Point (deg C):

Chemical ID 3:

Log Kow: Measured Log Kow:



Ecwin Results

Print Save Results Copy Remove Window Help

Neutral Organic SAR (Baseline Toxicity)	: Fish	14-day	LC50	471.597
Aromatic Amines	: Fish	96-hr	LC50	95.502
Aromatic Amines	: Fish	14-day	LC50	42.480
Aromatic Amines	: Daphnid	48-hr	LC50	1.523
Aromatic Amines	: Fish		ChV	0.499
Aromatic Amines	: Daphnid		ChV	0.034
Aromatic Amines	: Green Algae		ChV	7.331
Esters	: Fish	96-hr	LC50	41.180
Esters	: Daphnid	48-hr	LC50	287.370
Esters	: Green Algae	96-hr	EC50	3.239
Esters	: Green Algae		ChV	2.491
Esters	: Fish		ChV	19.787
Phenols	: Fish	96-hr	LC50	42.503
Phenols	: Daphnid	48-hr	LC50	13.308
Phenols	: Green Algae	96-hr	EC50	176.275
Phenols	: Fish	30-day	ChV	6.507
Phenols	: Fish	90-day	ChV	0.304
Phenols	: Daphnid	21-day	ChV	4.568
Phenols	: Green Algae	96-hr	ChV	14.567

◇データ解析的観点でのサブセット構築

◆サブセット化(クラス分類時)実施上での留意点

1. サブセット間での問題

サブセット間でサンプル数に大きな偏りがない。

2. クラスポピュレーションの問題

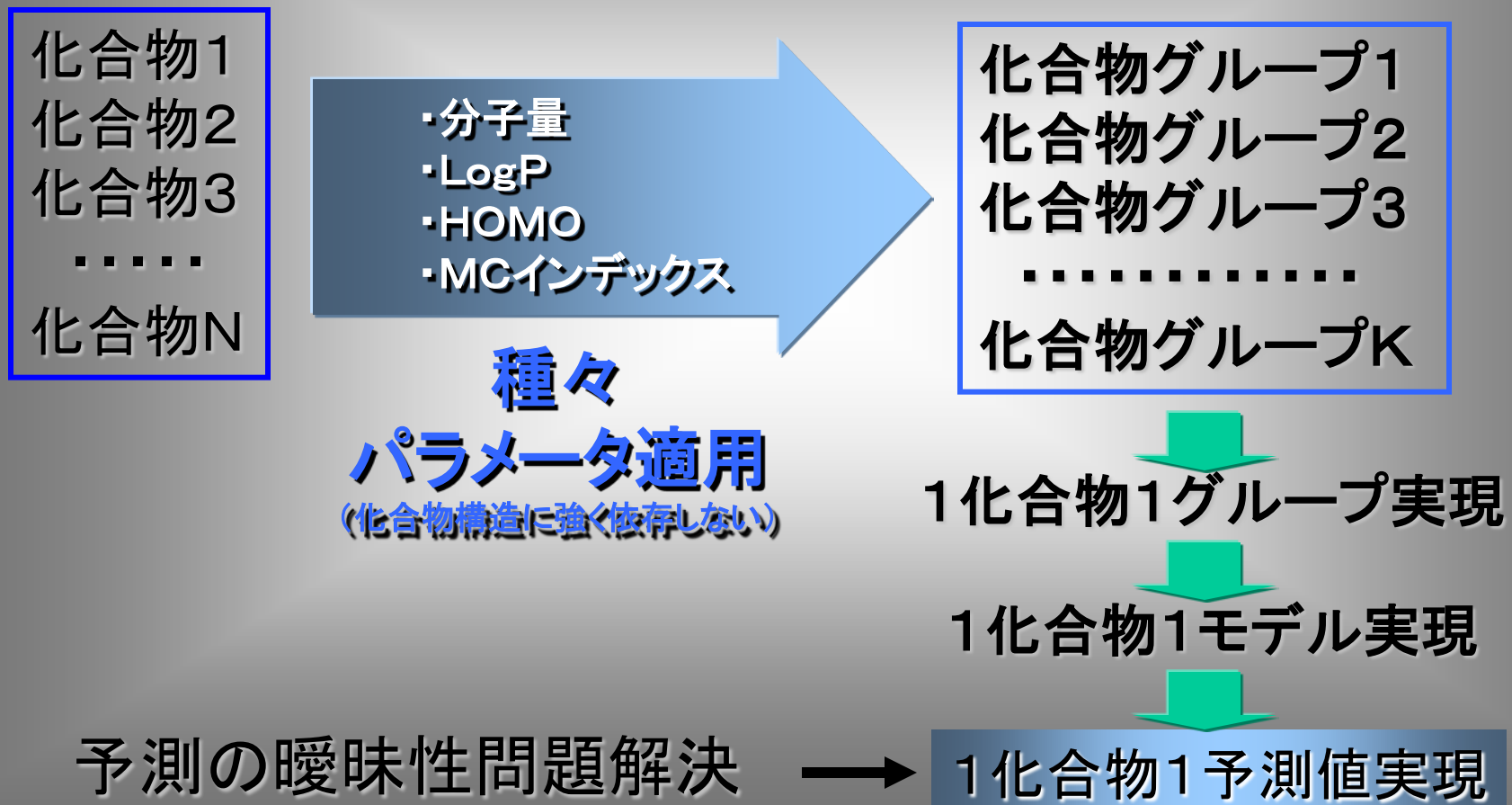
クラスポピュレーションのバランスが取れている。

3. サンプル上での問題

化合物構造式の変化性が均等に分散するものが良い。

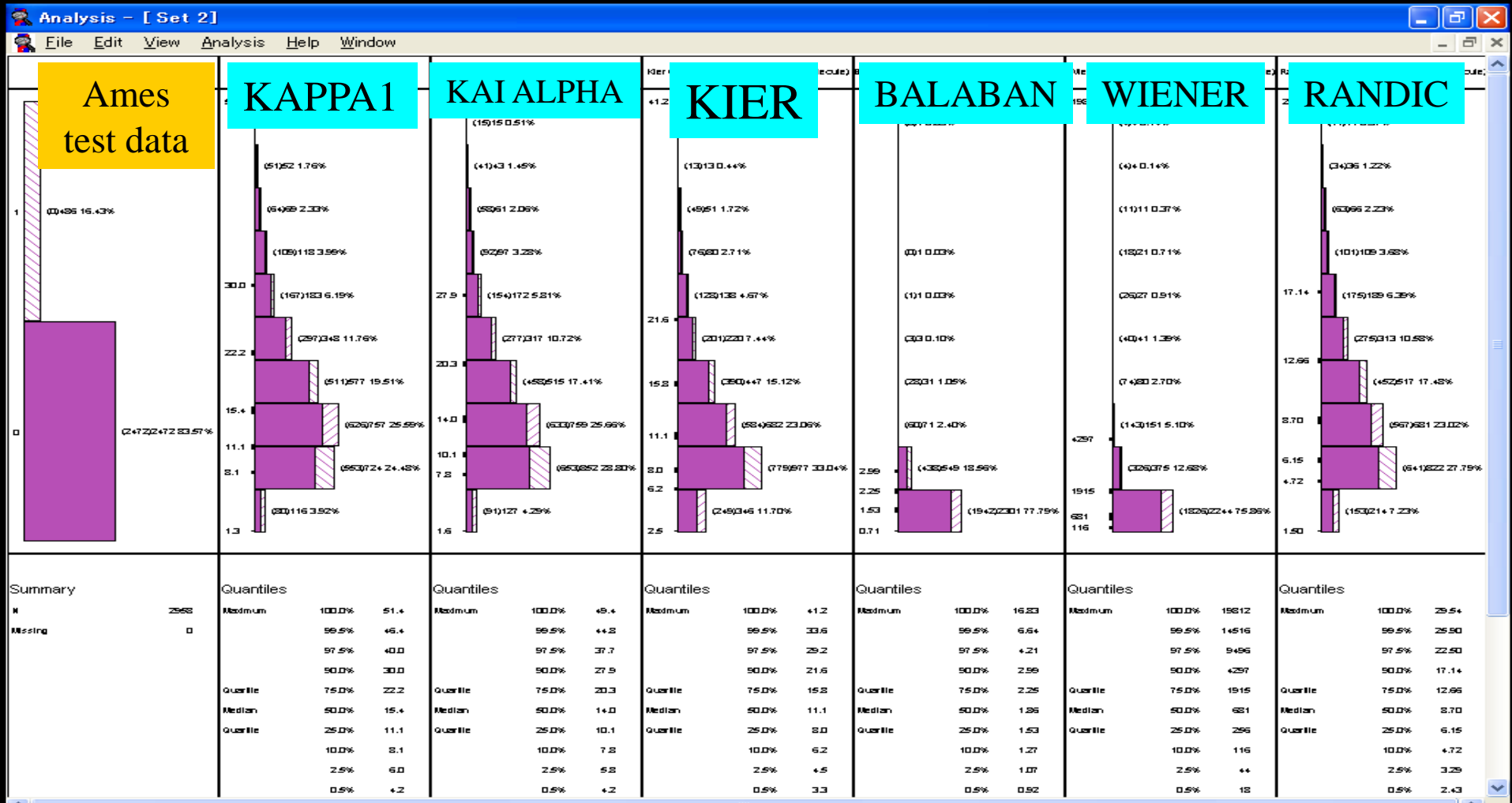
* クラスポピュレーション段階で分類知見を探索する時は
クラス間分布に偏りがあることが重要。

◇化合物の一元一項対応による予測信頼性の確立



湯田 浩太郎、猿渡 雄彦、中西 良文、松島 泰次郎、
パターン認識による、微生物変異原性予測におけるサンプリング手法に関する考察(II)
CBI学会2004年大会要旨集、東京、P2822、(2004)。

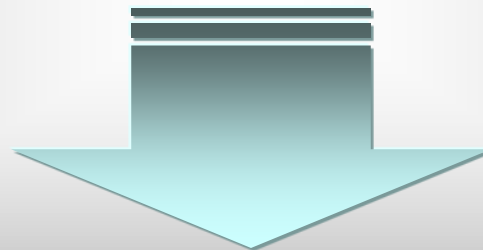
トポロジカル関連パラメータによるサブセット化



◇分類率、予測率向上を目指した提案(1)

分類率 ⇒ **KY (K-step Yard sampling) 法**の
開発と実施

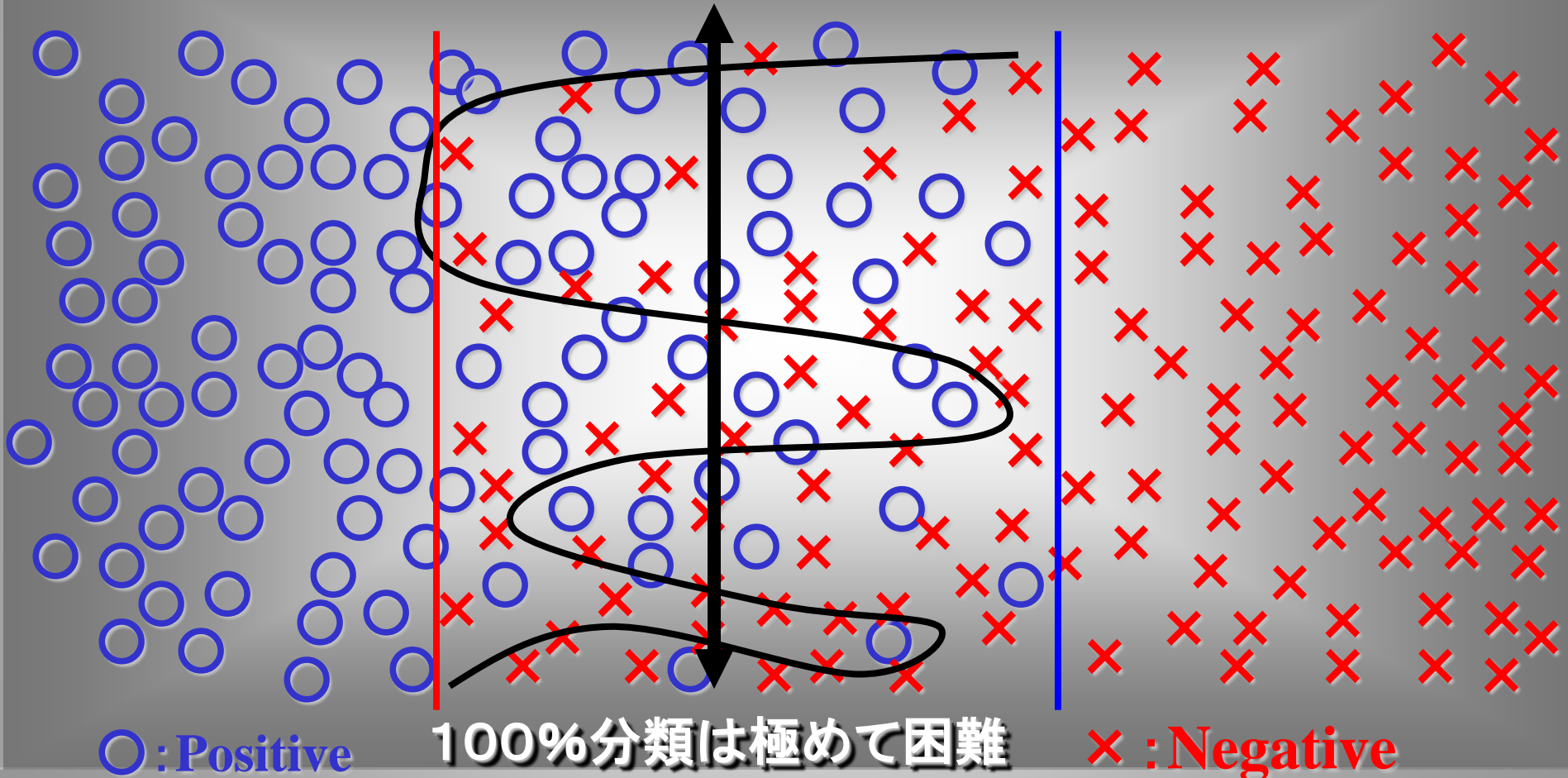
第34回構造-活性関連シンポジウム(2006)



予測率 ⇒ 「**テーラード・モデリング**」の
開発と実施

パターン分布とパターン分類：識別不可能な場合

100%分類不可能な識別線



グレーゾーンを取り出して再分類するアプローチ

“K-step Yard sampling (KY)法”



分類率／予測率向上への挑戦

K-step Yard sampling method

KY法



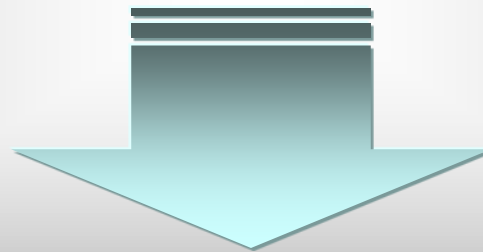
全く新規の考えに基づく、世界最初のアプローチ



分類率向上が極めて困難であった
Ames 試験データ 6、965 サンプルの、
100%分類に成功

◇分類率、予測率向上を目指した提案(2)

分類率 ⇒ 完全(100%)分類の目標達成

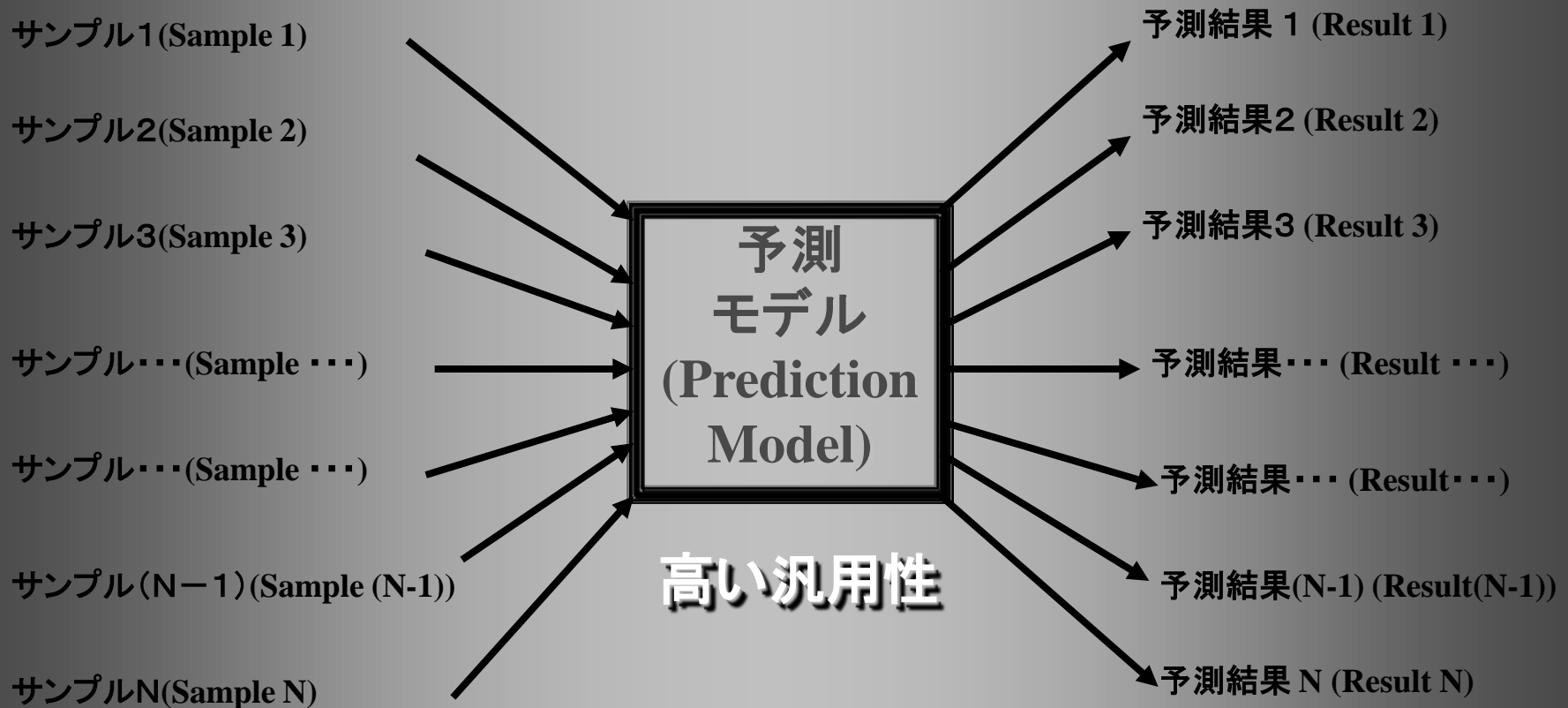


予測率 ⇒ 「テーラーメイド・モデリング」の
提案と実施

従来手法による予測アプローチ (Prediction approach by traditional method)

特徴: 全てのサンプルを対象とした予測モデルの構築

Features: Generate a prediction model which can handle all samples



利点 (Merit) : 少ない数の予測モデル作成で済む (Small number of prediction models are generated)

難点 (Weakness) : 予測率の向上が困難である (Difficult to achieve high prediction ratio)

予測率向上に関する現時点での限界

■ 予測の一般的な実施形態

一つの判別関数で 多様性の高い複数サンプルの予測実施

メタン、エタンレベルの予測から、ステロイドやマクロライド等までの予測が要求される
予測に無理がある

判別関数は高い汎用性を持つ、
このために予測の切れが悪くなる。

複数サンプルの予測を保証。
この目的のために、余分な
情報を含み、オーバースペック。

予測率向上のための発想の転換

現状でのアプローチ

少ない判別関数で多数のサンプルを予測

予測対象 サンプル特異性の無い 判別関数による予測

発想転換

今回の提案によるアプローチ

サンプル特異性の高い 判別関数を構築

提案

「テーラーメイド・モデリング」

「テーラーメイド・モデリング」の究極の形

特徴: サンプル単位での予測モデルの構築

Features: Generate a prediction model which is designed for only 1 samples



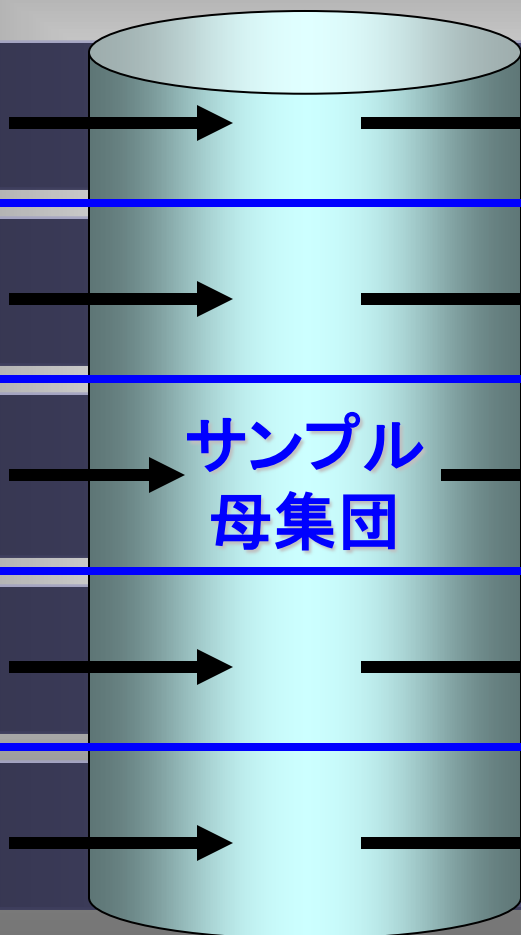
利点 (Merit) : 予測率が大幅に向上する (High prediction ratio will be achieved)

難点 (Weakness): 計算時間がかかる (Need large calculation time)

「テーラード・モデリング」による予測の流れ

予測対象化合物	類似性 スクリーニング	サンプル サブセット	予測モデル 構築	予測結果
---------	----------------	---------------	-------------	------

クラス未知
サンプルA



サブサンプル
セットA

予測モデルA

ポジ

クラス未知
サンプルB

サブサンプル
セットB

予測モデルB

ネガ

クラス未知
サンプルC

サンプル
母集団

サブサンプル
セットC

予測モデルC

ポジ

クラス未知
サンプ
ル...

サブサンプル
セット...

予測モデ
ル...

ポジ

クラス未知
サンプルZ

サブサンプル
セットZ

予測モデルZ

ネガ

サンプル母集団からの予測用サンプルの取り出し

サンプル母集団

類似サンプル空間



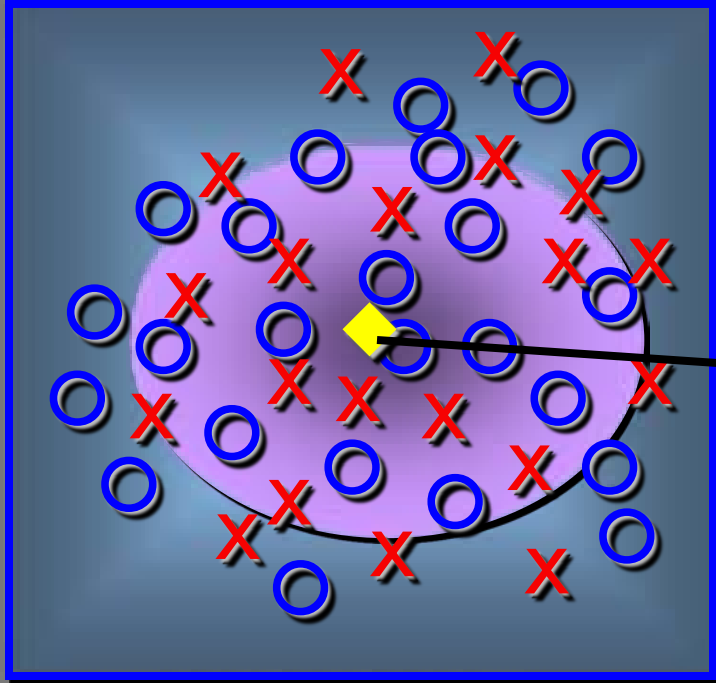
◆ 予測対象サンプル

“似た化合物は似た活性を示す”

類似サンプル群

予測用サンプルの取り出しと、テーラーメイド予測

サンプル **類似空間**

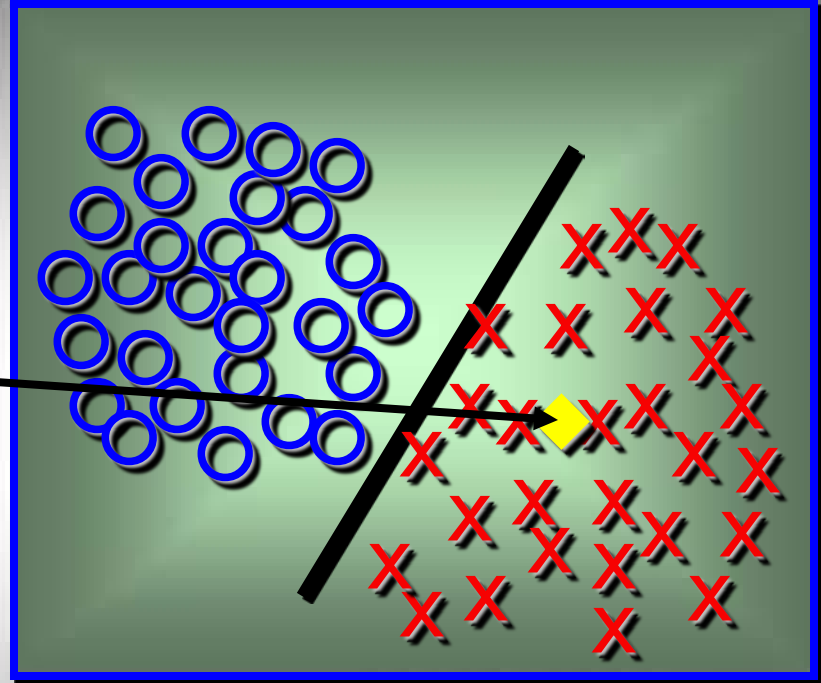


予測対象サンプルを中心とし、
サンプル母集団からの
類似サンプル群の取り出し

◆ 予測対象サンプル

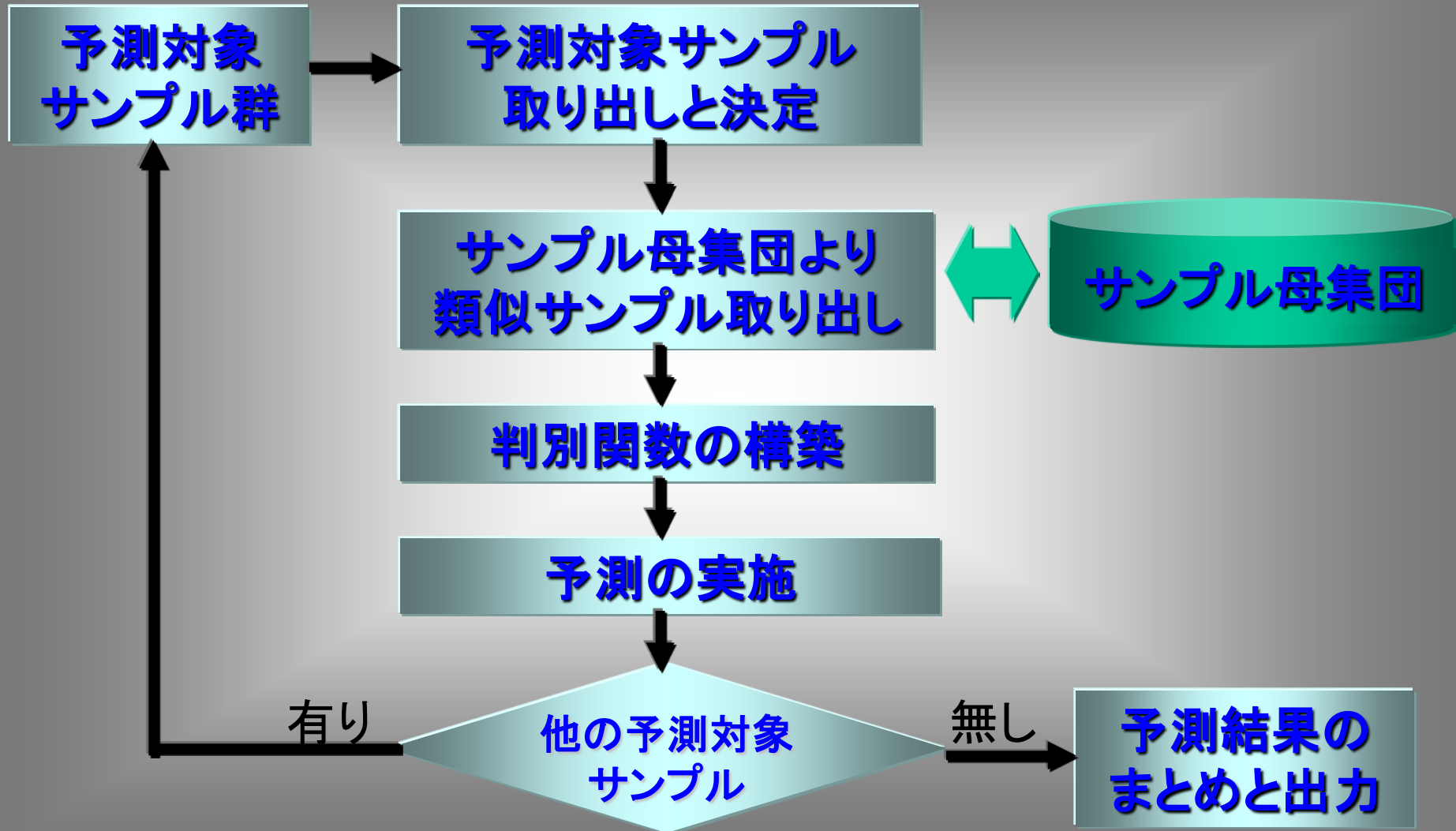


サンプル **分類空間**

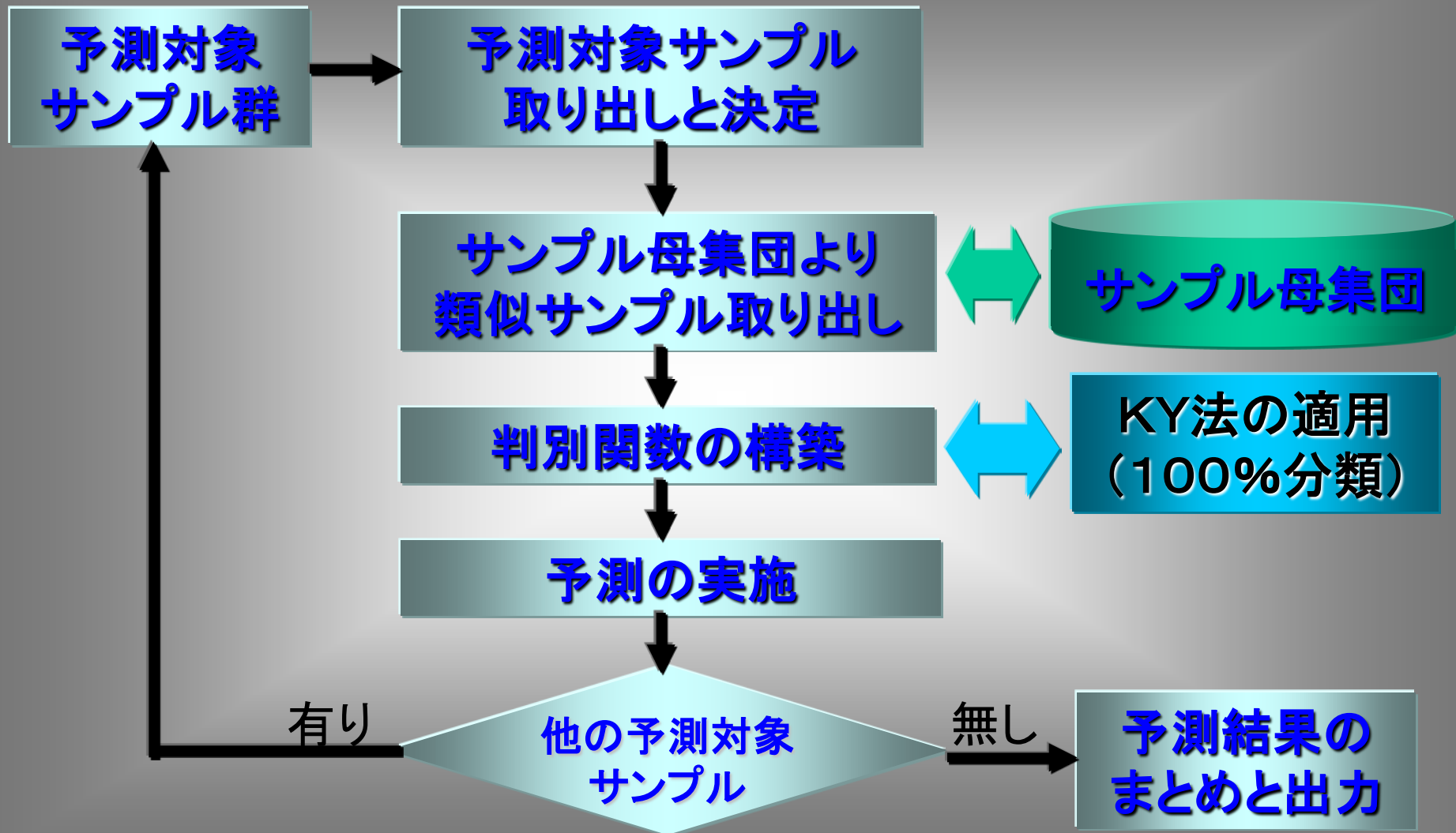


取り出されたサブセットの
サンプル空間再構成と、
テーラーメイド予測の実施

「テーラーメイド・モデリング」のブロックダイアグラム



「テーラーメイド・モデリング」+「KY法」



「テーラード・モデリング」の特徴と今後

長所:

1. 予測率

基本原理から **予測率の向上**が期待される(要検証)。

2. サンプル数の問題

個々のサンプル単位で予測モデルを構築する。

従って、**サンプル数の多少にかかわらず**限界に近く、高い予測率を得る。

3. KY法との連携でより高い予測率が得られる

分類率で100%を実現する**KY法との連携**により、

テーラード・モデリング単体での実施よりも更に高い予測率を達成する可能性がある。

欠点:

1. 分類／予測の実施に計算時間がかかる(スパコン主体?)



湯田 浩太郎
Kohtaro Yuta

In Silico Data, Ltd.