J06

FUJITSU

「テーラーメード・モデリング」による

化学データ解析手法（1）

富士通株式会社

大阪大学臨床医工学融合研究教育センター

湯田　浩太郎

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

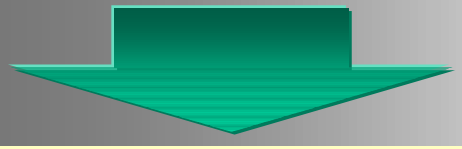
サブセットにアサインする時の曖昧性の解決

・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

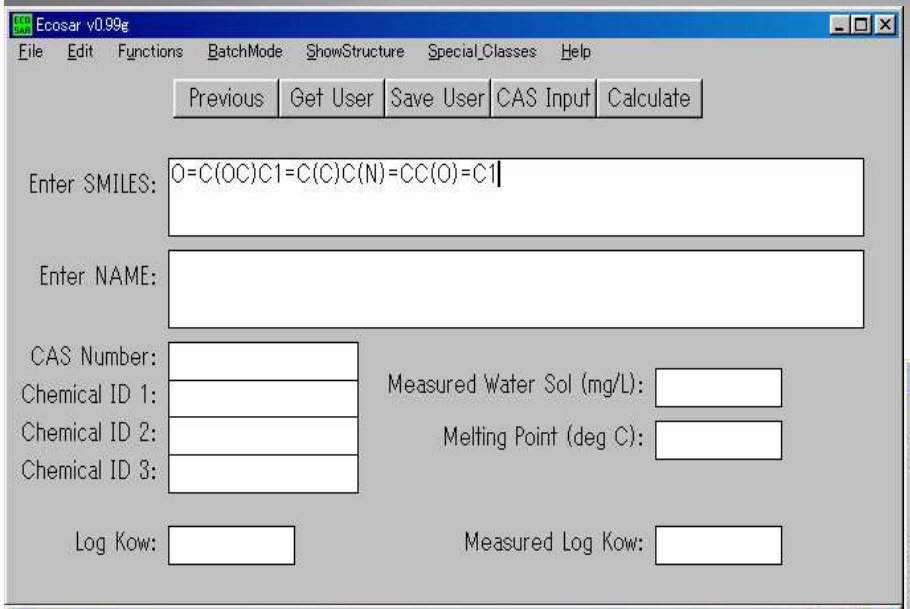
・分類が一義に定まらない化合物の処理が困難
分類基準の複数にまたがる化合物が多く存在する
分類を間違うと、不適切な予測モデルが適用されるために、
予測信頼性が低下する。

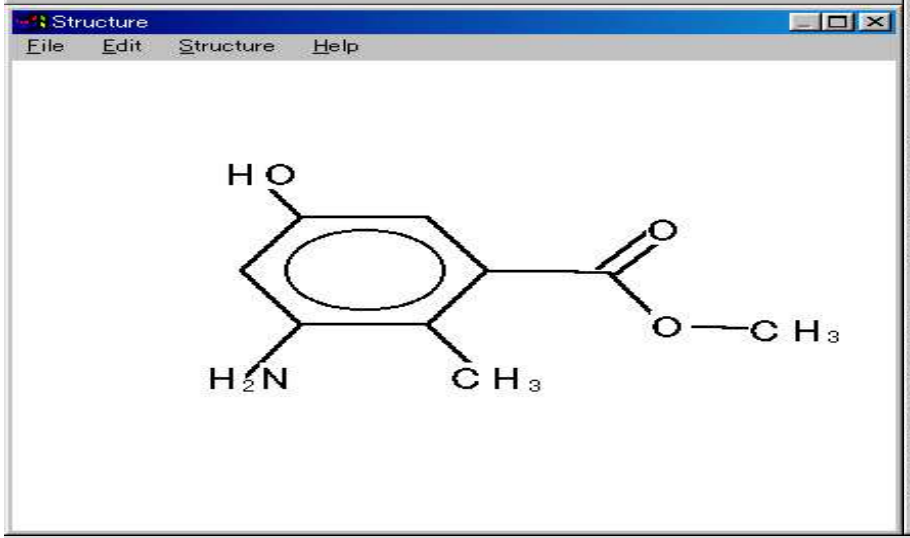
◇ECOSAR(EPA)による環境毒性予測

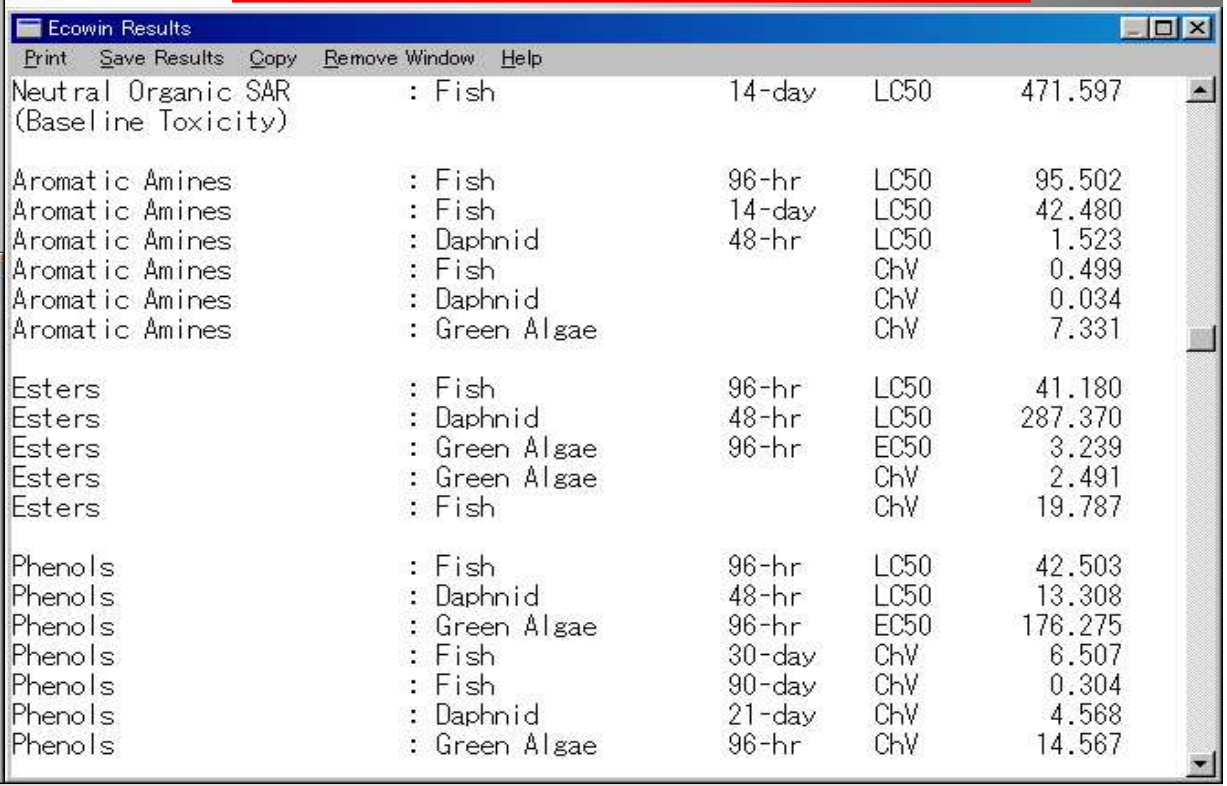
一元多項対応が原因となる予測信頼性の低下

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**







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。

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）.

FUJITSU

◇化合物の一元一項対応による予測信頼性の確立

化合物1
化合物2
化合物3
・・・・・
化合物N

・分子量
・LogP
・HOMO
・MCインデックス

種々
パラメータ適用
（化合物構造に強く依存しない）

化合物グループ1
化合物グループ2
化合物グループ3
・・・・・・・・・・・・
化合物グループK

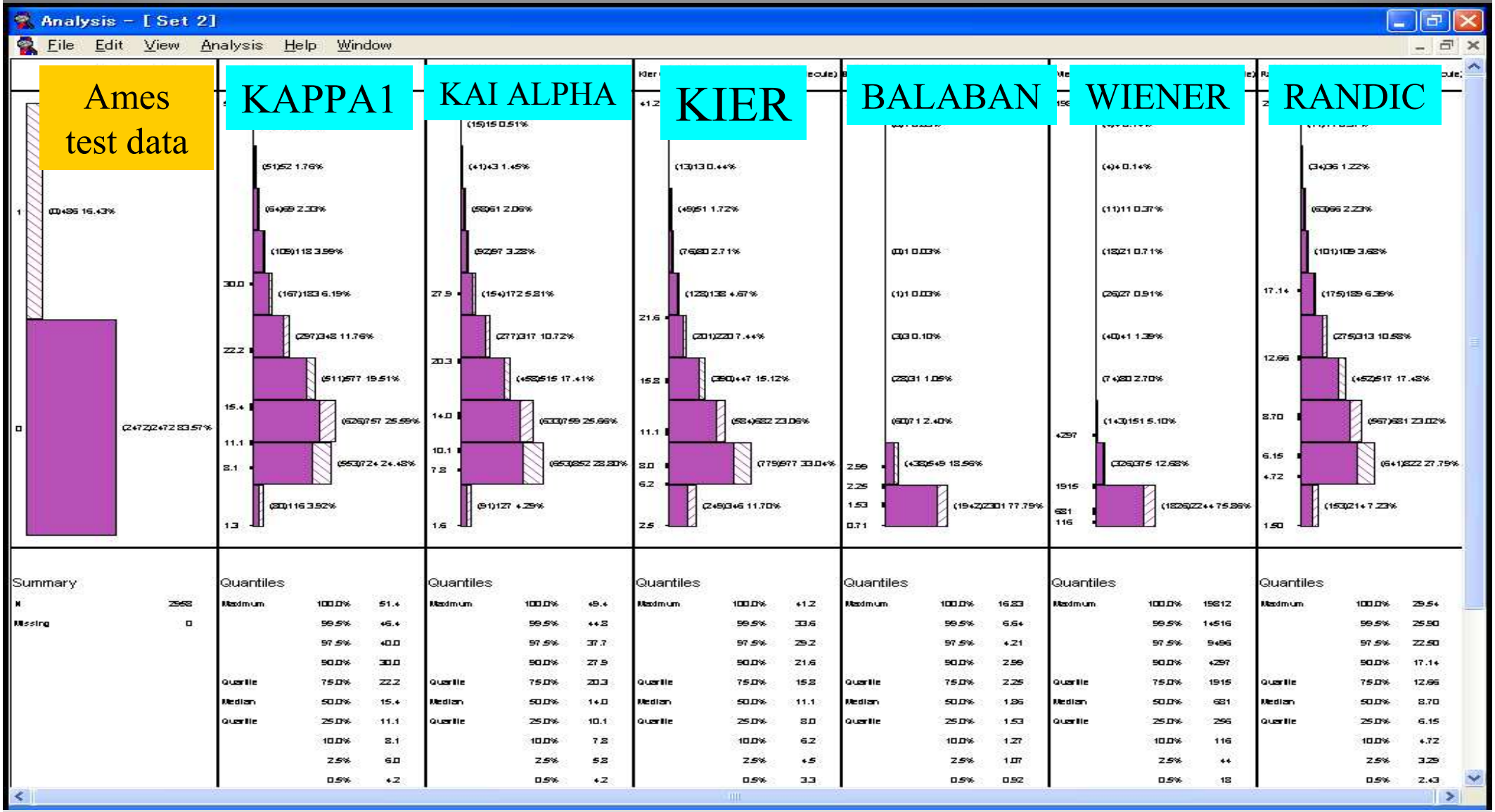
1化合物1グループ実現

1化合物1モデル実現

予測の曖昧性問題解決 → 1化合物1予測値実現

FUJITSU

■トポロジカル関連パラメータによるサブセット化

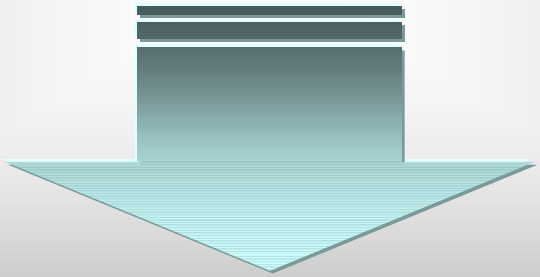


FUJITSU

◇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

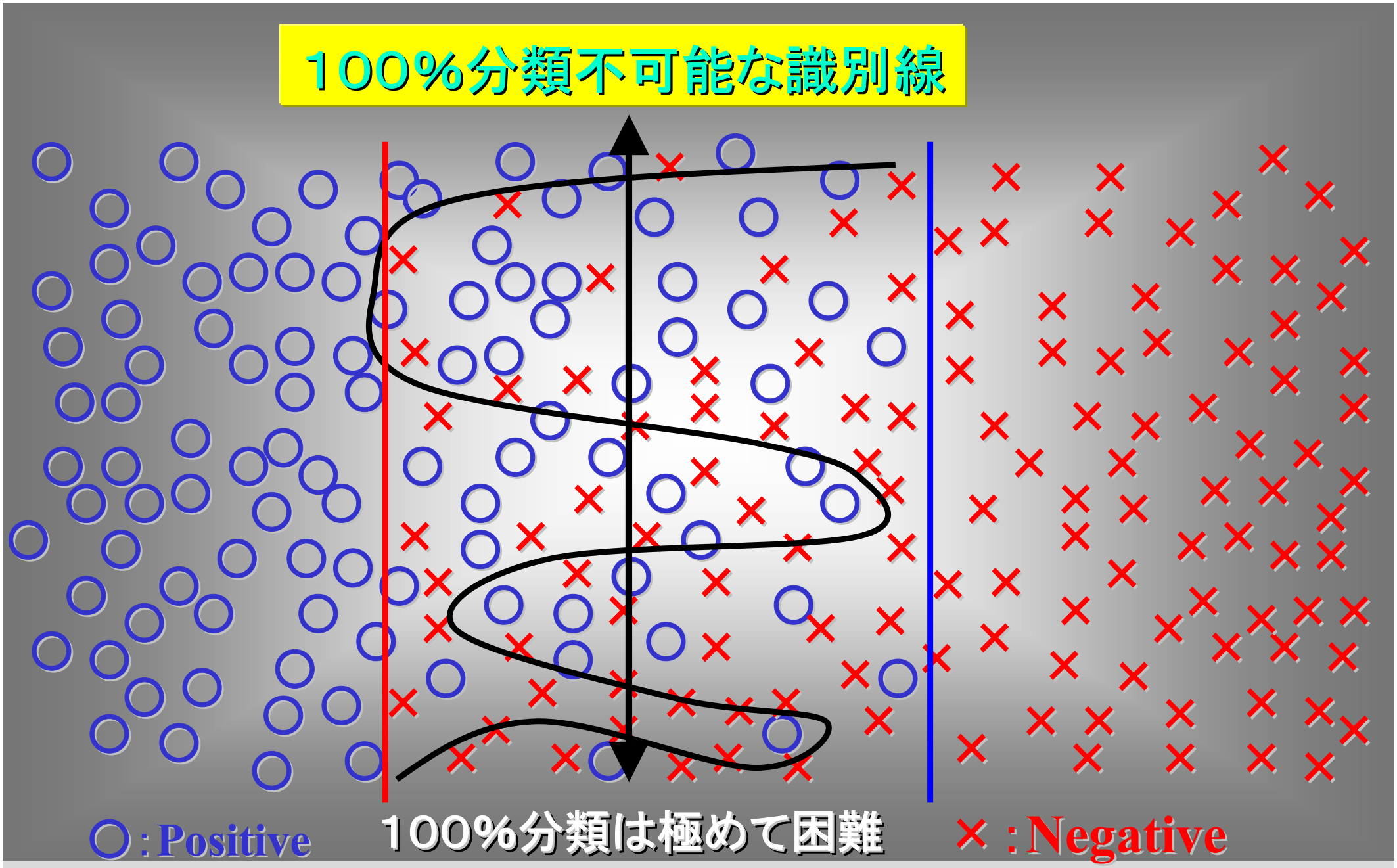
第34回構造－活性相関シンポジウム(2006)



予測率　⇒　「テーラーメード・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

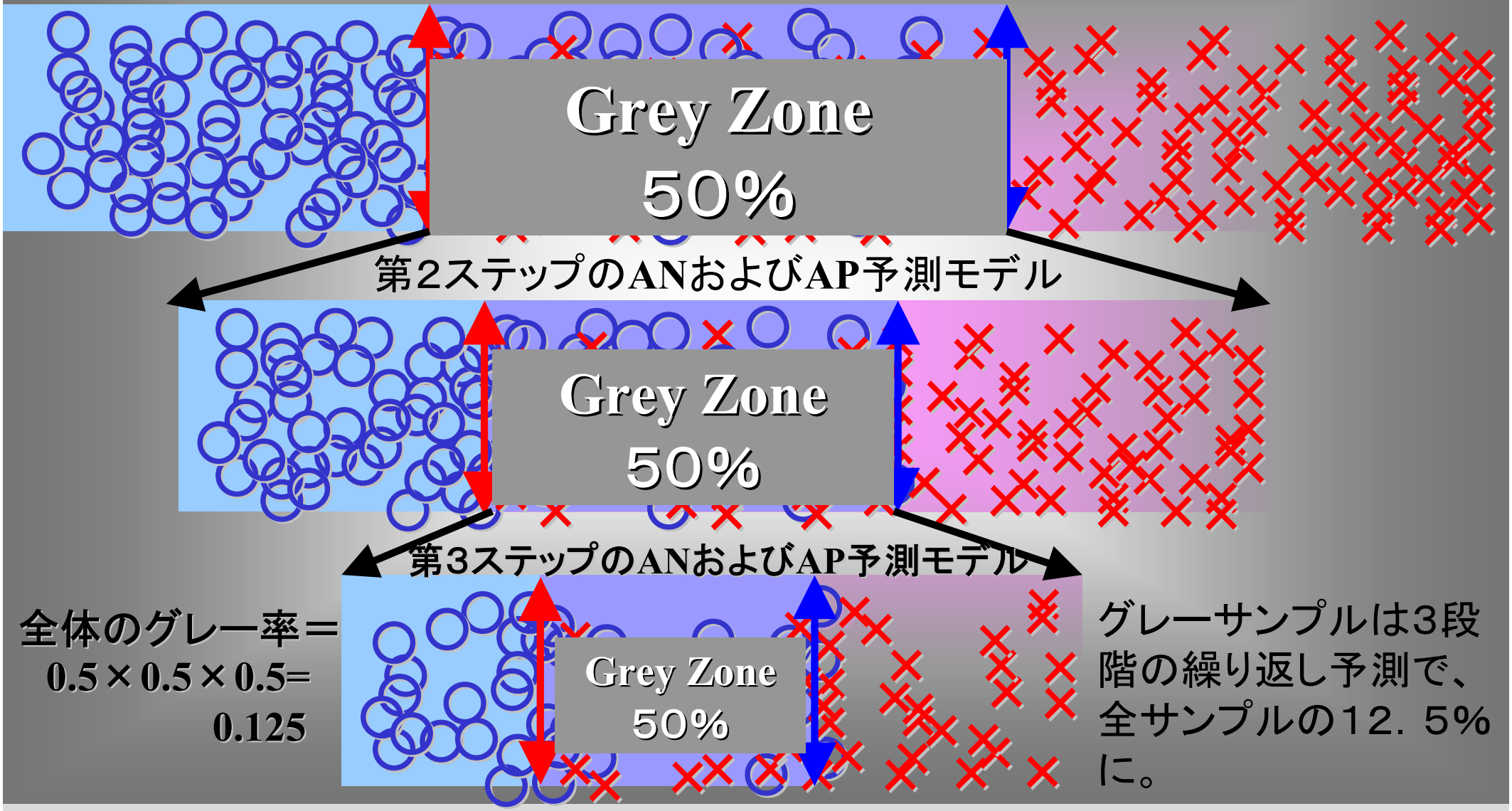
FUJITSU



グレーゾーンを取り出して再分類するアプローチ

FUJITSU

“K-step Yard sampling（KY）法”



全体のグレー率＝
0.5×0.5×0.5=
0.125

グレーサンプルは3段階の繰り返し予測で、全サンプルの12．5%に。

分類率／予測率向上への挑戦

FUJITSU

K-step Yard sampling method

KY法

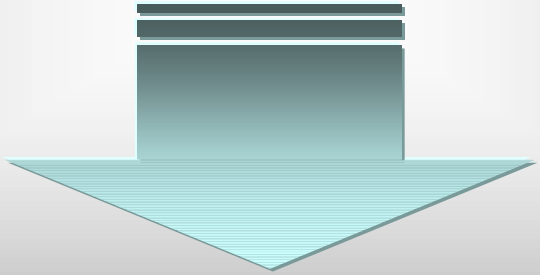
全く新規の考えに基づく、世界最初のアプローチ

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

特許出願済

◇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

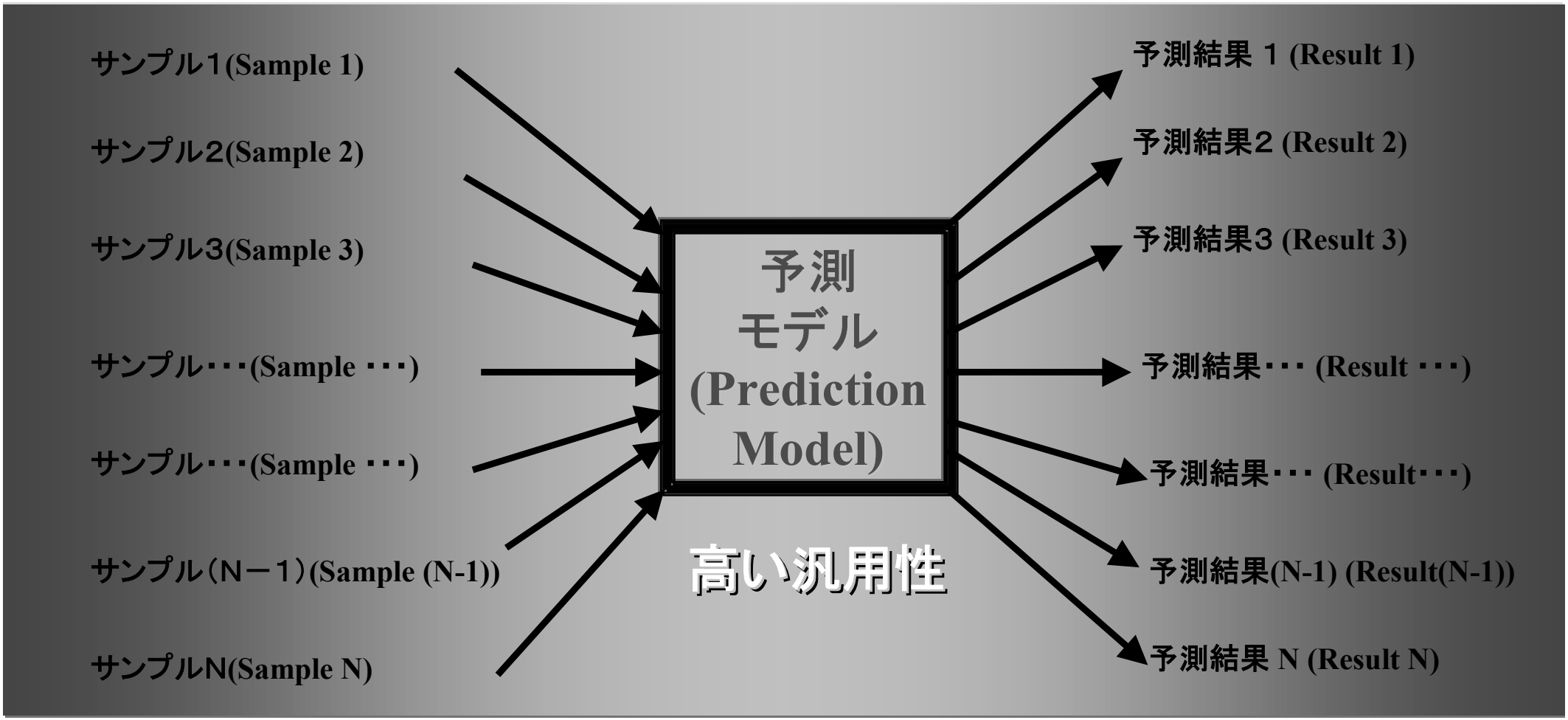


予測率　⇒　「テーラーメード・モデリング」の提案と実施

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

FUJITSU

特徴:総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

FUJITSU

■ 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

FUJITSU

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

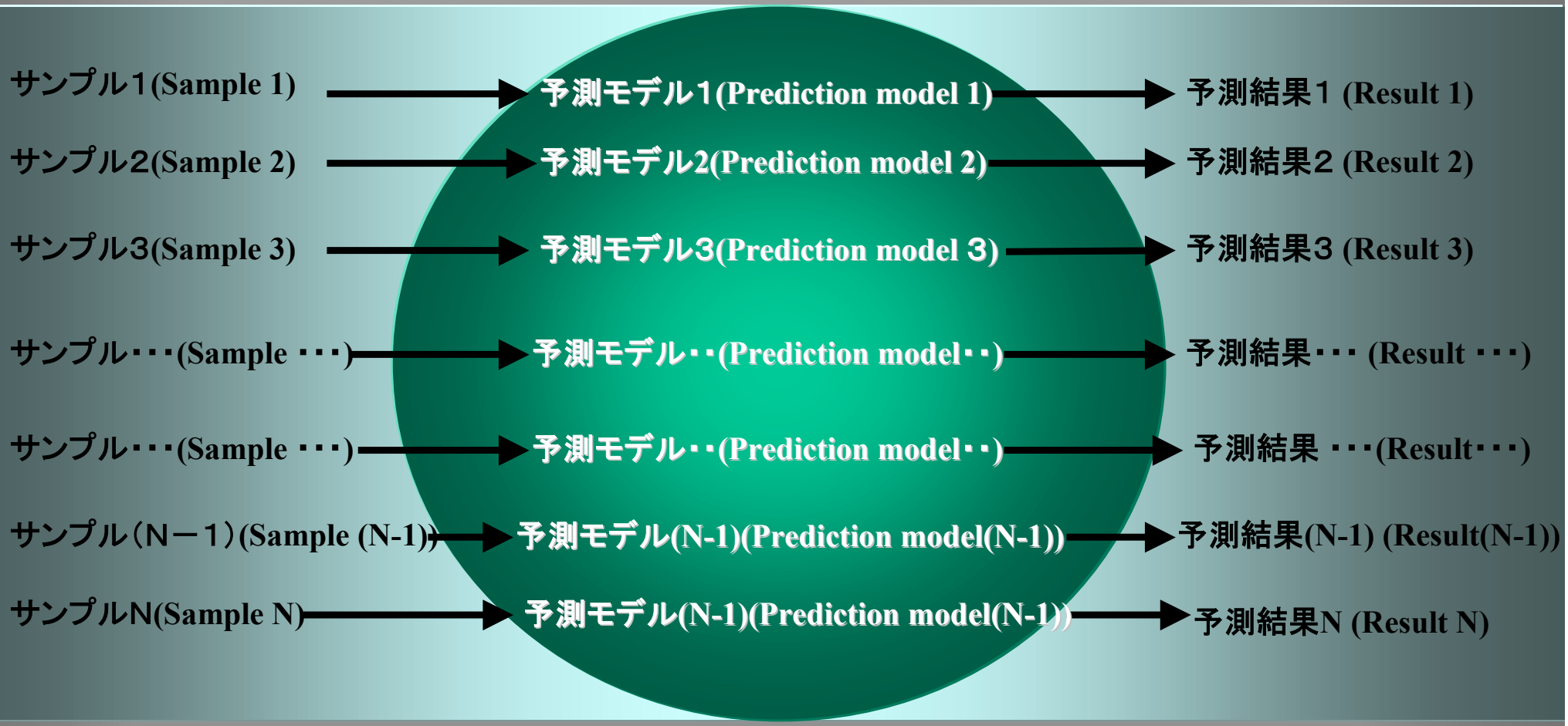
「テーラーメード・モデリング」

「テーラーメード・モデリング」の究極の形

FUJITSU

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



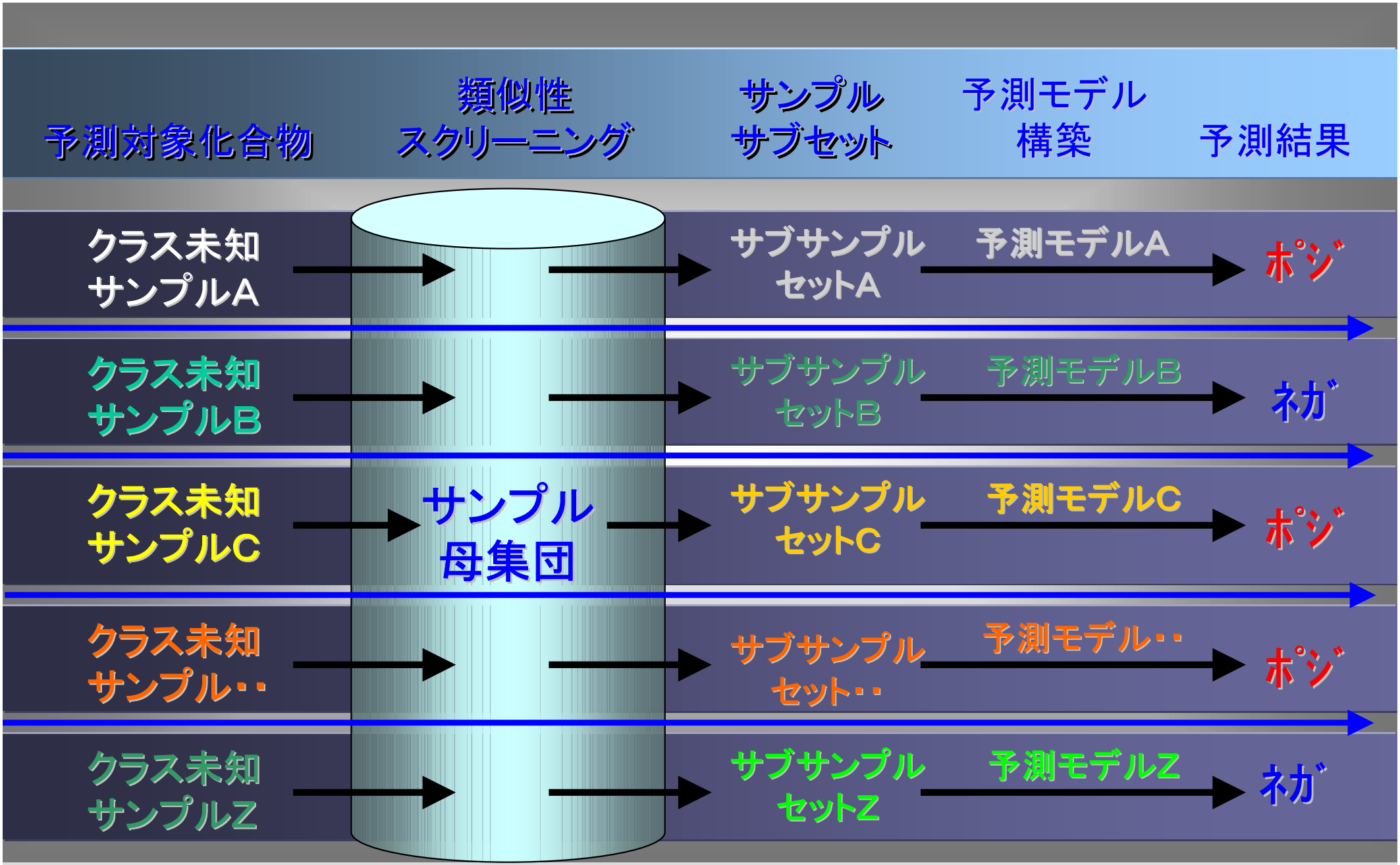
利点 (Merit)　：　予測率が大幅に向上する（High prediction ratio will be achieved）

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

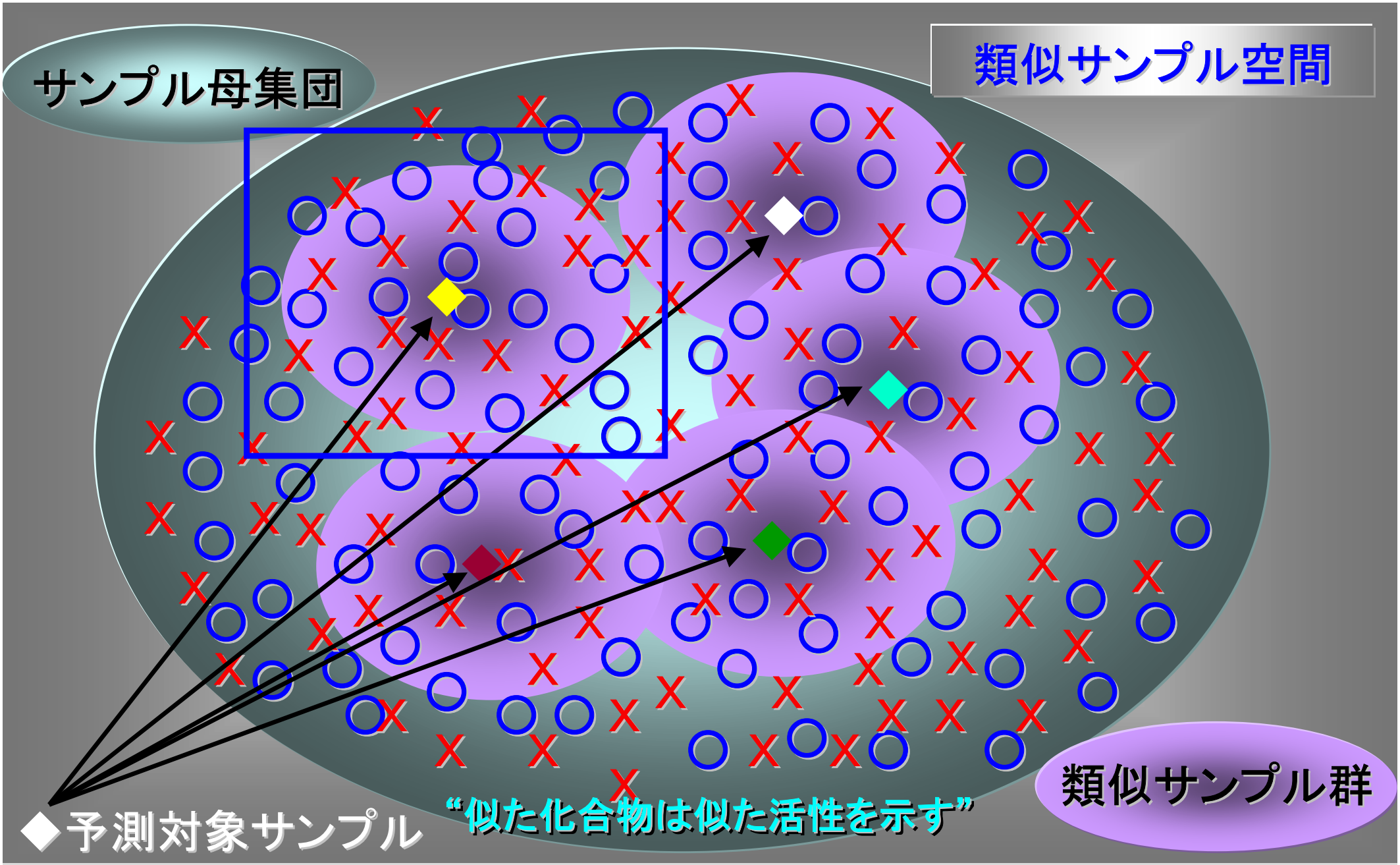
特許出願済

「テーラーメード・モデリング」による予測の流れ

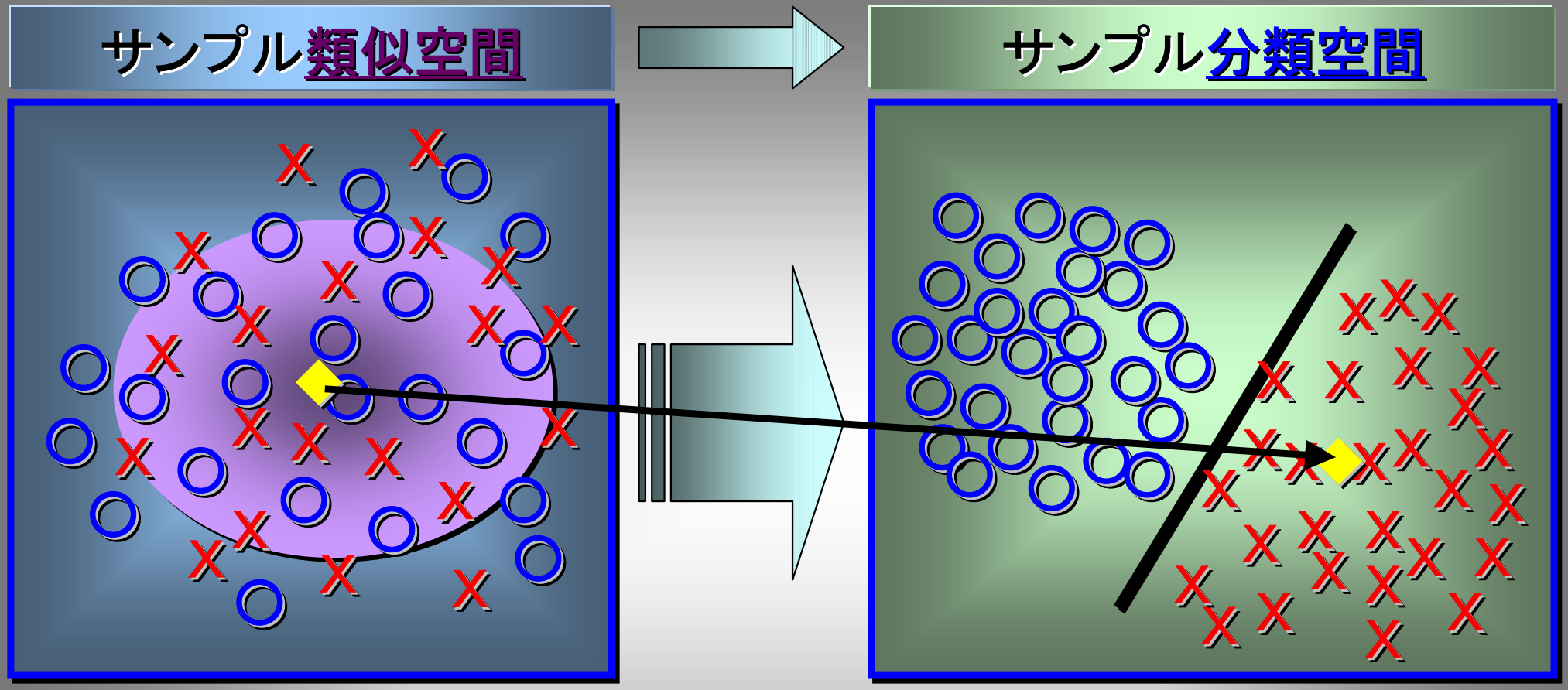
FUJITSU



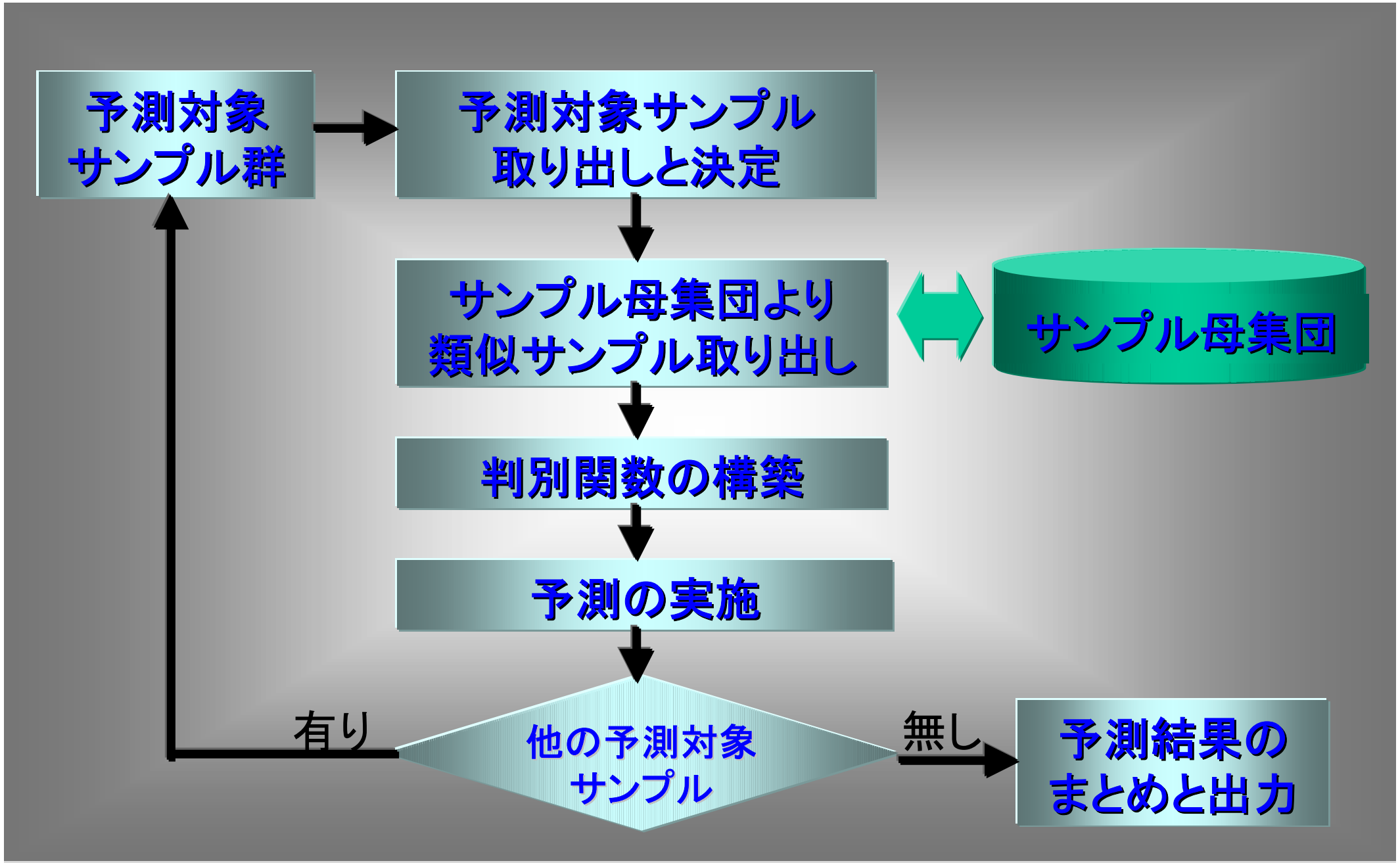
サンプル母集団からの予測用サンプルの取り出し



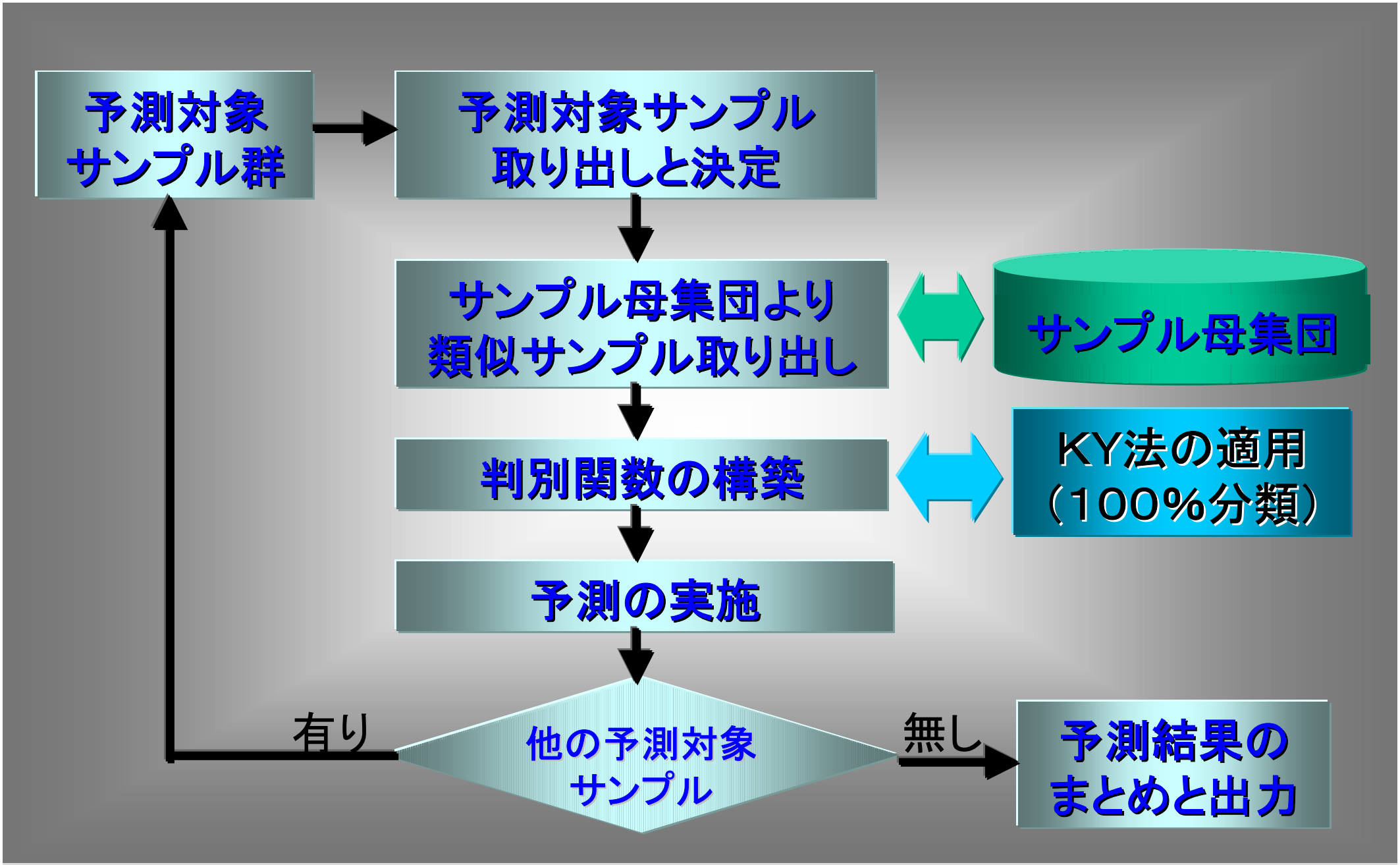
予測用サンプルの取り出しと、テーラーメード予測



「テーラーメード・モデリング」のブロックダイアグラム



「テーラーメード・モデリング」＋「KY法」



「テーラーメード・モデリング」の特徴と今後

FUJITSU

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、
テーラーメード・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？）

FUJITSU

「テーラーメード・モデリング」による

化学データ解析手法（1）

2007. 11. 16

□Ames試験サンプル（6，965）を用いた、代表的な判別分析法 の適用結果

１．線形最少二乗アルゴリズムによる判別分析

分類率（全体）：73.50(6965)　ポジティブ：73.02(2932)　ネガティブ：73.84(4033)
誤分類サンプル数　(1846)　(791)　(1055)
予測率（L 100 OUT)　72.58%　乖離度（0.92%）
(L 500 OUT)　73.32%　乖離度（0.18%）

２．ＳＶＭ（サポートベクターマシーン）

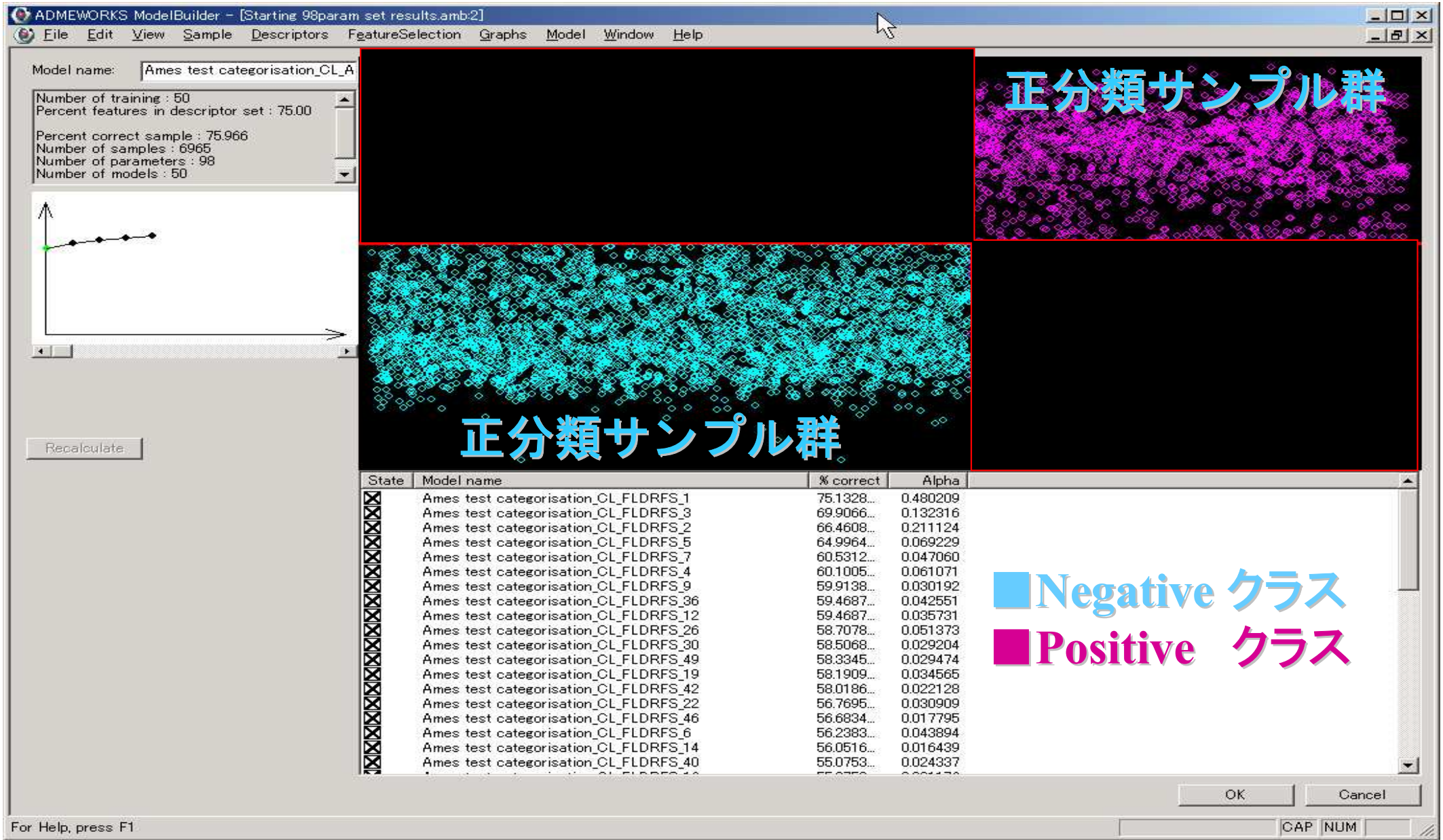
分類率（全体）：90.87(6965)　ポジティブ：86.83(2932)　ネガティブ：93.80(4033)
誤分類サンプル数　(636)　(386)　(250)
予測率（L 500 OUT)　80.99%　乖離度（9.88%）

３．ＡｄａＢｏｏｓｔ

分類率（全体）：77.24(6965)　ポジティブ：66.13(2932)　ネガティブ：85.32(4033)
誤分類サンプル数　(1585)　(993)　(592)
予測率（L 500 OUT)　75.16%　乖離度（2.08%）

ＡｄａＢｏｏｓｔによる分類結果

77.24%分類率の場合の6,965サンプルの分離状態図



分類／予測率向上のための問題点

FUJITSU

■　サンプル数と分類／予測率の問題

サンプル数を増やすと予測性が向上する。

解析母集団のサンプル数が増えると、
分類／予測率は一般的に下がる。

■　分類率と予測率の基本的な関係

分類率は予測率よりも常に高い値となる。

分類率　≧　予測率

予測率の向上を議論する前に分類率の向上が先。

KY（K-step Yard sampling）法の特徴

FUJITSU

◇分類率を極限（100%）まで高める手法

■　開発目標（1）

分類／予測の困難性にかかわらず、
常に完全分類（100%）を実現。

■　開発目標（2）

サンプル数がどんなに大きくなっても、
常に完全分類（100%）を実現。

■　開発目標（3）

分類率と予測率の乖離が少ない手法を目指す。

特許出願済み