インシリコ(コン
毒性予測の問題と

株式会社　インシリコ

In Silico Data
Miracles by the KY-methods

ピュータ）による
解決に向けた提案

データ　湯田浩太郎

インシリコ(コンピュータ)による毒性予測を困難にする要因と問題点

◆毒性の特徴:

1. メカニズムが複雑で詳細不明
2. 多種多様な毒性

◇予測対象化合物の特徴:

1. 化合物の構造変化性が極めて高い
2. 予測対象化合物数が極めて多い

★予測における問題点

予測率が低い　！！

インシリコ(コンピュータ)予測へのアプローチ手法

1. データ解析によるアプローチ

基本技術:ケモメトリックス

- コンピュータケミストリー
- 統計／多変量解析／パターン認識等のデータ解析手法

2. 人工知能によるアプローチ

- コンピュータケミストリー
- 人工知能技術

3. 物理シミュレーションによるアプローチ

- 物理方程式
- シミュレーション技術

データ解析によるアプローチに必要な技術

1. コンピュータケミストリー関連技術

化学関連情報をコンピュータ上で扱う為の技術

◇化学データ解析を行う時前に必要な技術

① 化合物構造式関連技術

一次元、二次元、三次元化合物構造式の創出と取り扱い技術

② 化合物構造式情報を数値データ(パラメータ)に変換する技術

トポロジカルパラメータ、トポグラフィカルパラメータ、物理化学パラメータ、理論化学パラメータ、物性パラメータ、部分構造パラメータ、その他

2. データ解析関連技術(統計／多変量解析／パターン認識)

種々数値データを用いて解析目的に従ってデータ解析(予測等)を行う技術

◇化学データ解析を行う技術

①種々データ解析基本手法

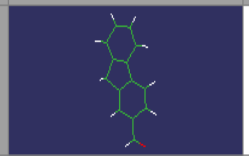
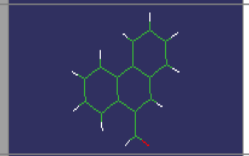
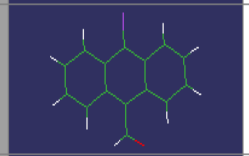
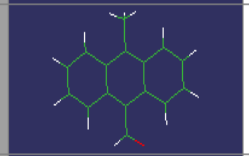
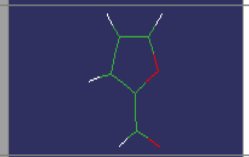
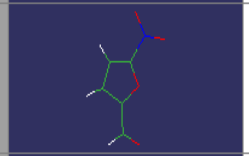
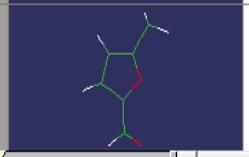
クラス分類、フィッティング、マッピング、クラスタリング、チャート、他

②解析目的に合わせたアプローチ

構造-活性／物性／毒性相関、インシリコスクリーニング、脱毒性デザイン、他

Tsar - ArCHO_Reagents.tsar

File Edit Display Structures Properties Analysis Tools Window Help

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
	Row ID	Whole Molecule	Name (Whole Molecule)	Molecular W (Whole Mol	#Bromir (Whole	#Carbo (Whole	#Chlorir (Whole	#Fluorin (Whole	#Haloge (Whole	#Hetero (Whole	#Hydroc (Whole	#Hydroc (Whole	#Nitroge (Whole	#Oxyger (Whole	#Sulphu (Whole	Chi 0 (a (Whole	Chi 1 (b (Whole	Chi 2 (p (Whole	Chi (Wh
5			2-FLUORENECARBOXALDEHYDE	194.23	0	14	0	0	0	1	1	0	0	1	0	10.251	7.3813	6.4554	0.80
6			PHENANTHRENE-9-CARBOXALDEHYDE	206.24	0	15	0	0	0	1	1	0	0	1	0	10.958	7.8982	6.7246	0.74
7			10-CHLORO-9-ANTHRALDEHYDE	240.69	0	15	1	0	1	2	1	0	0	1	0	11.828	8.3257	7.1556	0.87
8			10-METHYLANTHRACENE-9-CARBOXALDEHYD	220.27	0	16	0	0	0	1	1	0	0	1	0	11.828	8.3257	7.1556	0.87
9			FURFURAL	96.086	0	5	0	0	0	2	2	0	0	2	0	5.1129	3.4319	2.5587	0.20
10			5-NITRO-2-FURALDEHYDE	141.08	0	5	0	0	0	5	2	0	1	4	0	7.5605	4.7364	4.0914	0.70
11			5-METHYLFURFURAL	110.11	0	6	0	0	0	2	2	0	0	2	0	5.9831	3.8257	3.1925	0.49

View 1

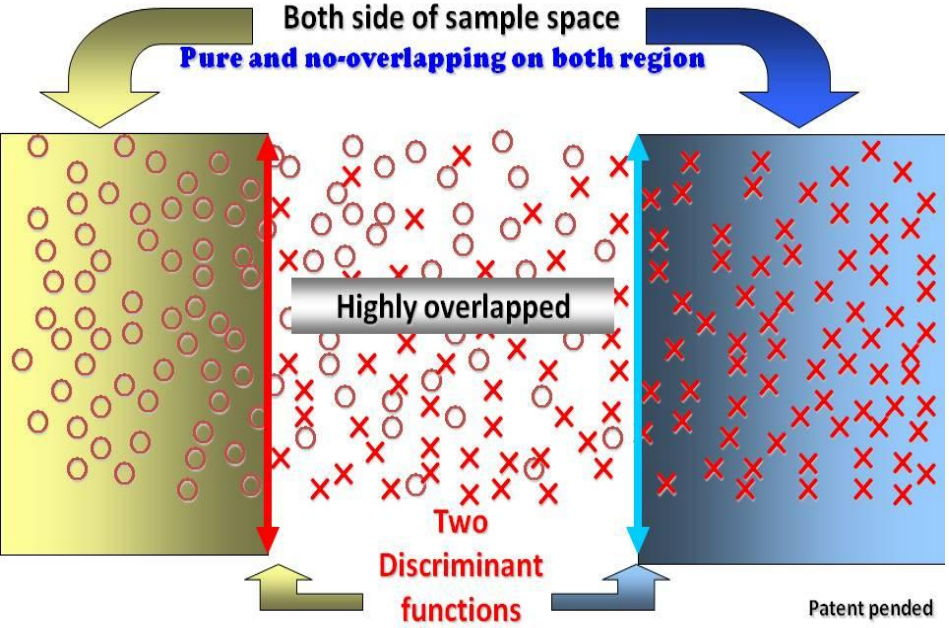
KY（K-step Yard sampling)法の提案

◇　究極のデータ解析手法（1）
◆2クラス分類:完全分類（100%）達成

Ames試験:　6965サンプルの100%分類実現

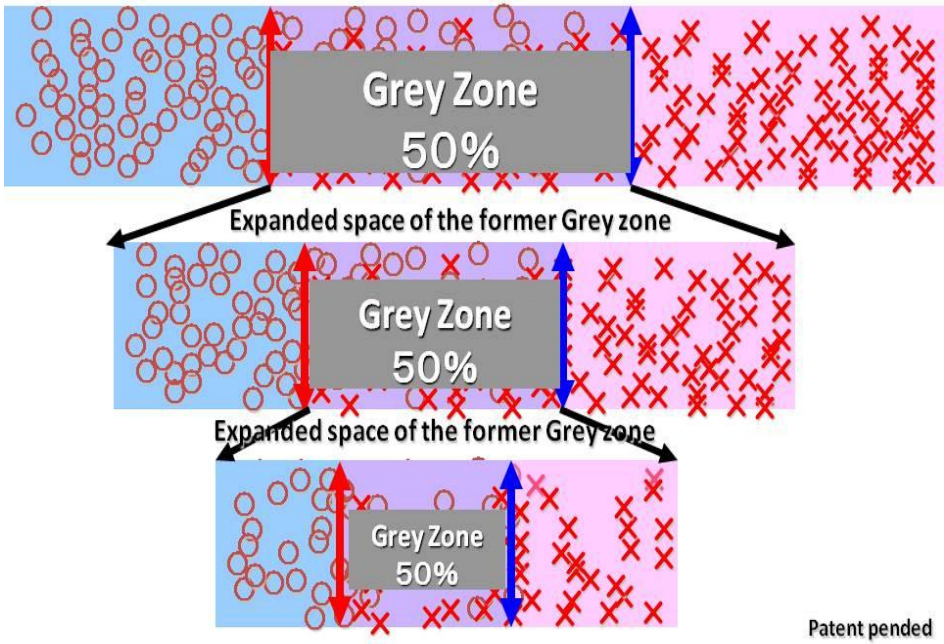
First basic concept of KY method

Spatial region on sample space



Second basic concept of KY method

Multi-steps for 100% classification

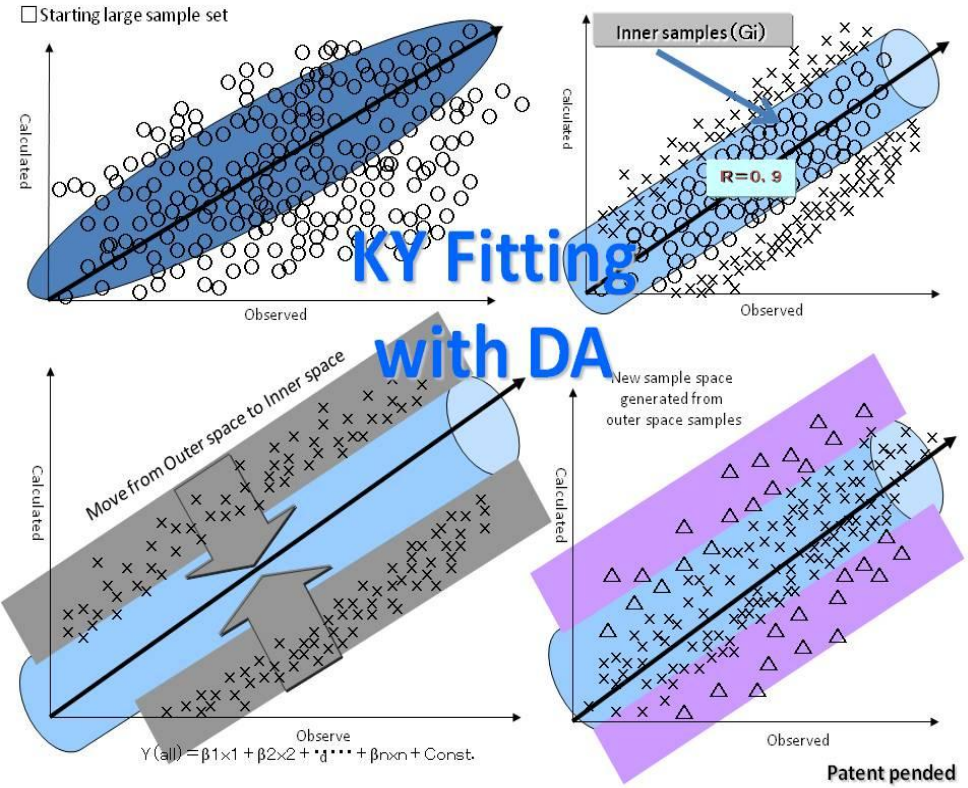


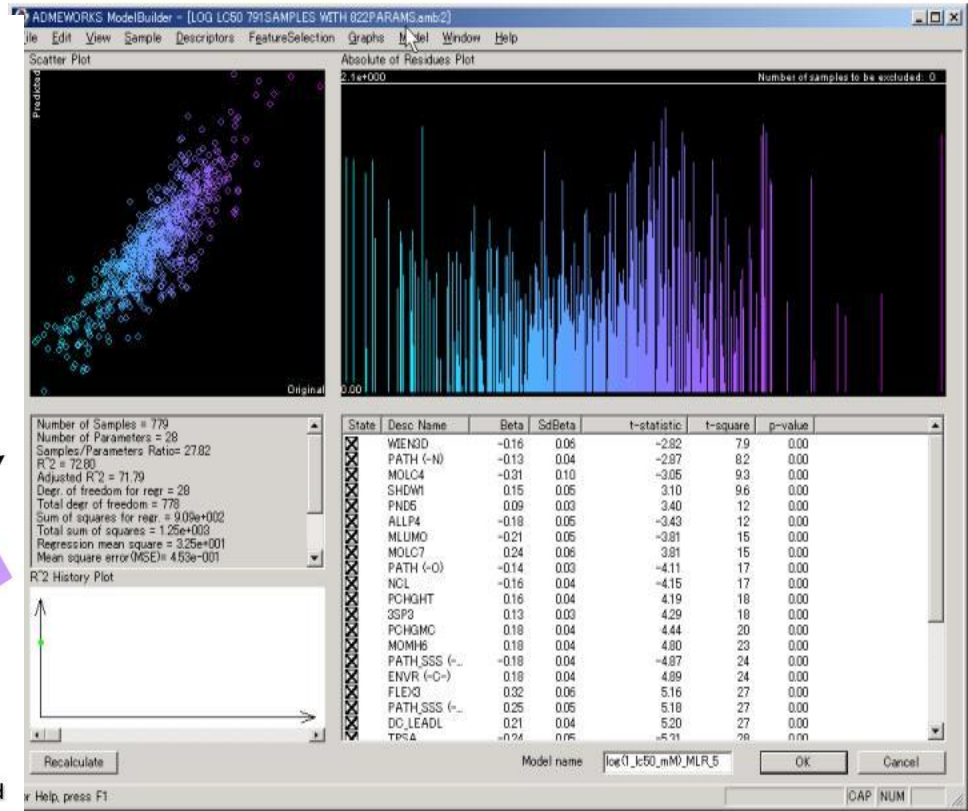
KY（K-step Yard sampling)法の提案

◇ 究極のデータ解析手法（2）
◆フィッティング：極めて高い相関および決定係数の実現

Fish: 96 hours LC50、 Number of samples: 791、Log(1/LC50_Mm) (Max/Min) : 6.376 / -2.963

サンプル数:779、 パラメータ数:28、 信頼性指標:27.8
R2:72.8、 R:85.3、 F値:71.7、 クロスバリデーション:69.6



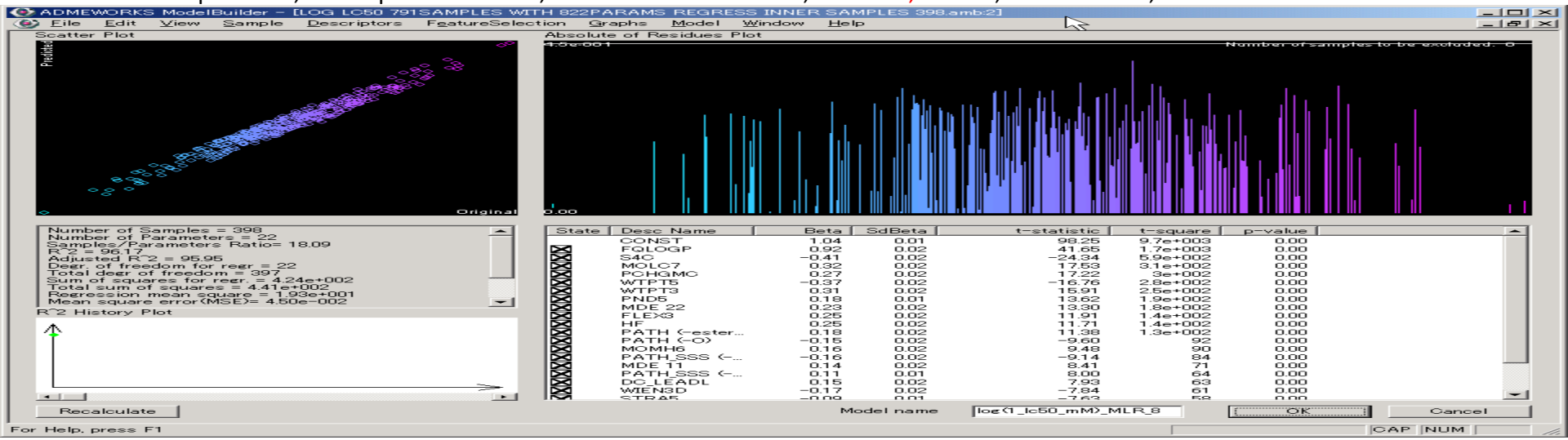


◇Fitting KY method Step1 (Inner sample set)

Patent pended

Step1 : Inner sample set

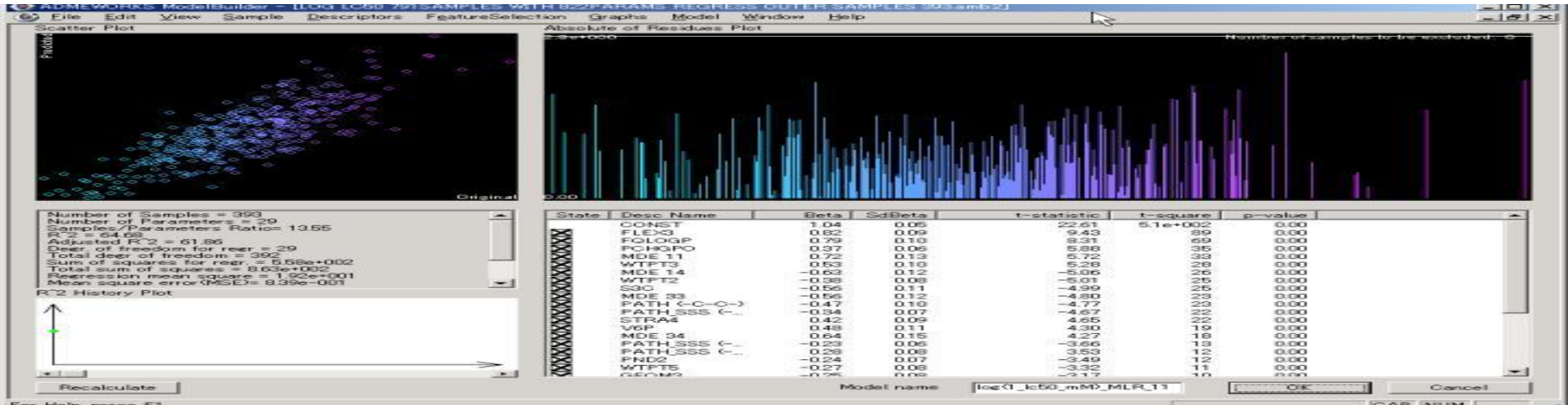
Number of samples : 398, Used parameters : 22, Confidance ratio : 18.1, R2 : 96.2, R : 98.1, F-value : 428, CV : 94.4



◇Fitting KY method Step1 (Outer sample set)

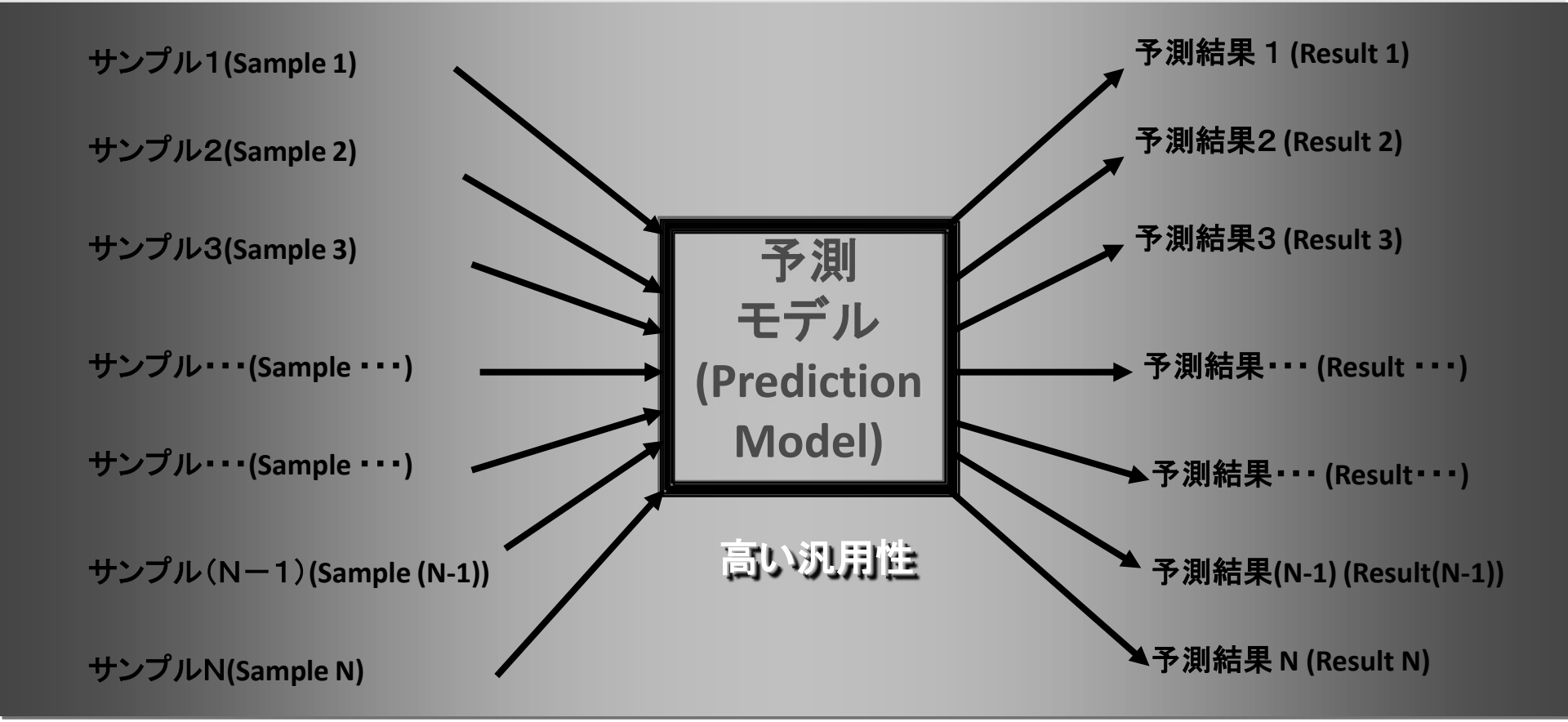
Step1 : Outer sample set

Number of samples : 393, Used parameters : 29, Confidance ratio : 13.6, R2 : 64.7, R : 80.4, F-value : 22.9, CV : 57.5



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

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



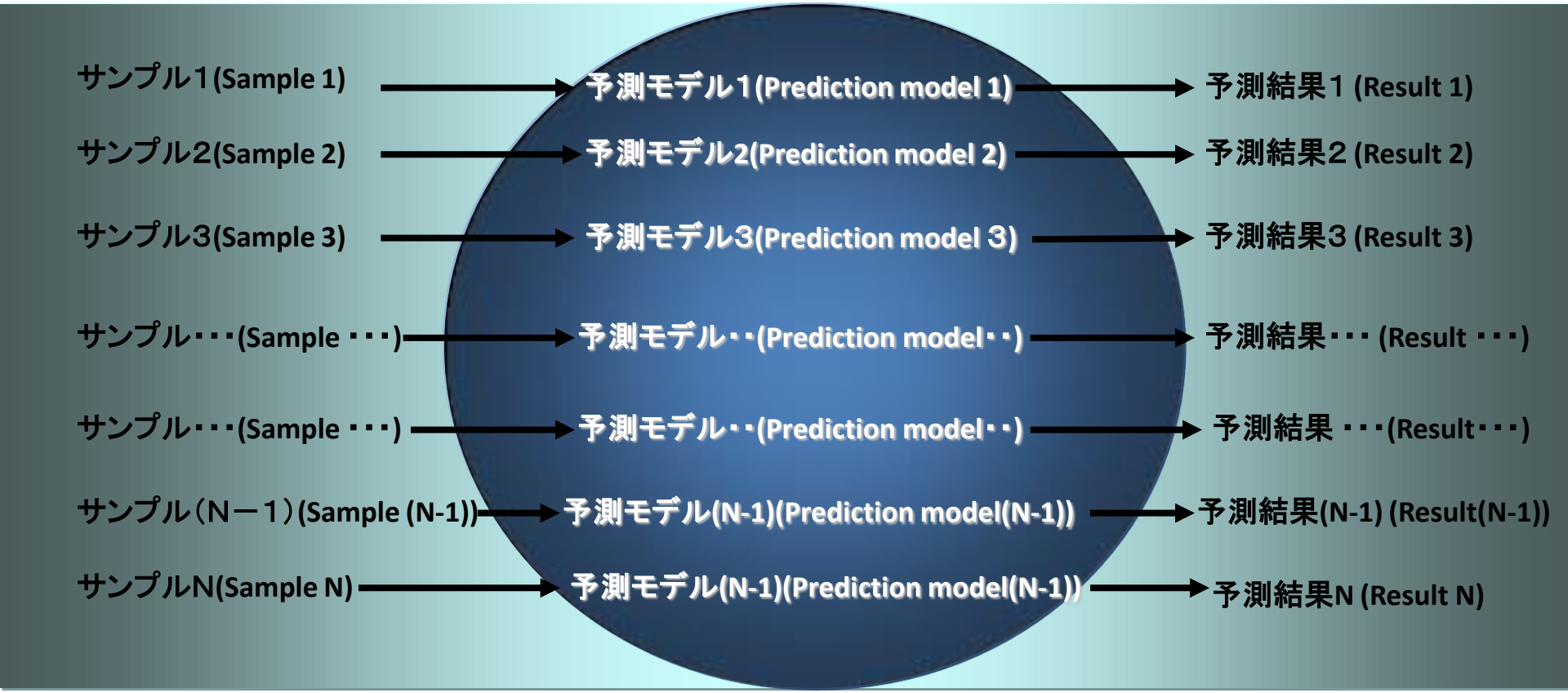
利点 (Merit)　：　少ない数の予測モデル作成で済む（Small number of prediction models are generated ）

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

「テーラーメード・モデリング」の提案

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



インシリコ(コンピュータ)による予測の提案

◆ 最大の問題点:予測率が低い

◇ 解決手法の提案:「KY法」や「テーラーメードモデリング」の導入による
分類率や予測率の総合的な向上を目指す

＊2クラス(バイナリ)データ(ポジ／ネガ)

1. 分類率は100%を実現可能(KY法)
予測率は分類率を超えることは無い。従って分類率を最大まで高める。
2. サンプル数フリーなので、どんなにサンプル数が増えても解析精度は落ちず、
予測信頼性はサンプル数の増大に伴い向上する。
3. KY法独自のステップ情報により、予測信頼性の向上がさらに促進される。
4. 「テーラーメードモデリング」は予測率の向上が期待される手法となる。

＊連続データ

1. 極めて高い相関および決定係数値を実現
2. サンプル数が増えても、高い相関／決定係数を実現
3. KY法独自のステップ情報により、予測信頼性の向上がさらに促進される。