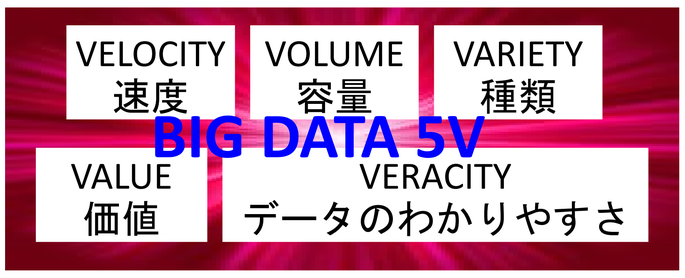
背景

■ビッグデータの利活用によるサービス提供

- ビッグデータを利用して新たな予測分析や価値創造を促進
- ビッグデータ解析からVALUEを引きだすVERACITYを発見すること



■創薬

現在、遺伝子と病気の関係が解明されつつあり、創薬は巨大なデータの中で 、病気の原因となる遺伝子（病因遺伝子）やその遺伝子が作るタンパク質と結合できる化合物を薬候補として捜し出すことができる。⇒ゲノム創薬

目的

1．ビッグデータを利活用した薬の副作用予測

問題点

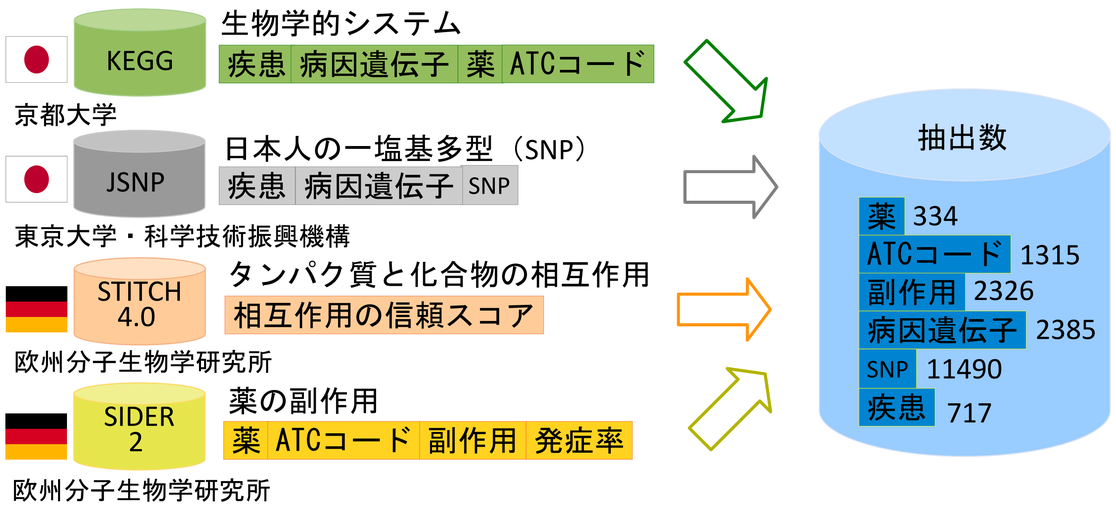
・臨床データは全て公開されていない
・創薬に関わるデータは完全には準備されていない

2．ビッグデータ解析から薬候補の発見

方法

■データベースの構築

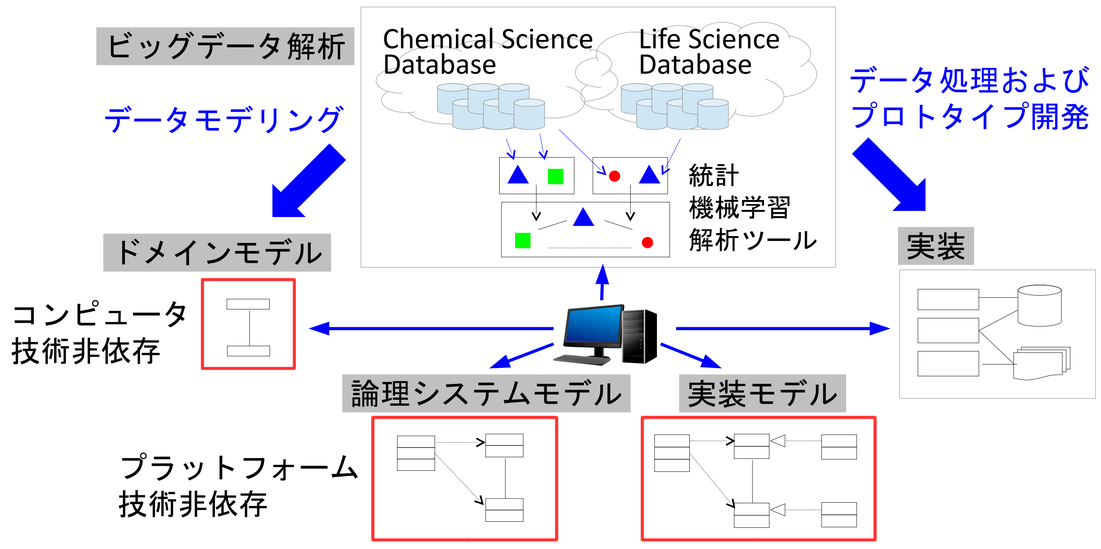
公開されているデータベースからのデータ抽出



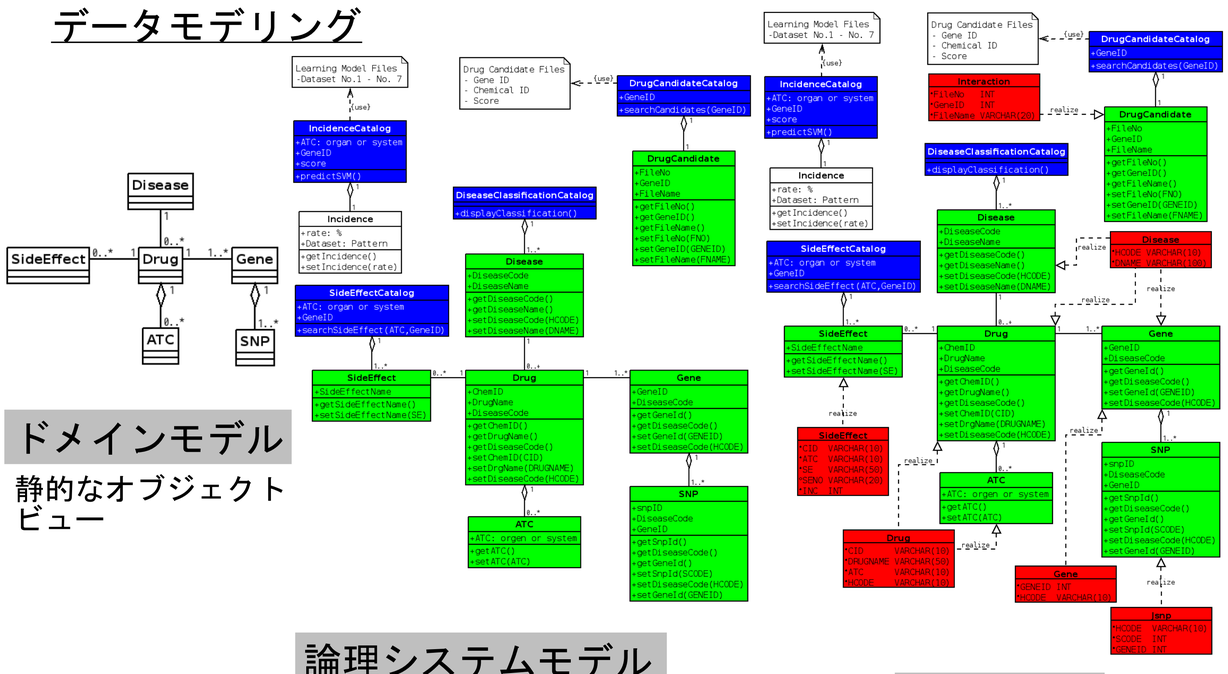
■データを中心にしたシステム開発

ソフトウェア開発手法

ビッグデータ解析とモデル駆動型アーキテクチャ（MDA）との連携によるデータモデリング



データモデリング



ドメインモデル

静的なオブジェクトビュー

論理システムモデル

クラス設計および静的なオブジェクトビュー

実装モデル

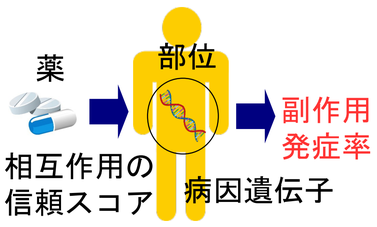
論理システムモデルにより設計されたモデルを実装するためのビュー

結果

副作用の予測

■薬の特性を表すピンポイントデータの抽出

薬理学やゲノム薬理学の知識を基に副作用モデルを構築して薬の特性を表すピンポイントデータを抽出



薬の特性	ピンポイントデータ	
生物学的 解剖学的	スコア	相互作用の信頼スコア
	部　位	ATCコード第1レベル 効果をもたらす部位・器官
遺伝的 臨床的	遺伝子	病因遺伝子ID
	副作用	発症する副作用
	発症率	副作用の発症率

■手法

◆発症する副作用

部位と病因遺伝子による副作用の分類

◆作用の発症率

スコア・部位・遺伝子・発症率間の関係を表す回帰式を用いた判別分析によるクラスタリングとサポートベクタマシン(SVM)による学習予測

不完全なデータへ対応するためにデータセットを準備

データセット	スコア	部位	遺伝子	発症率
1	•	•		•
2	•			•
3		•		•
4	•	•	•	•
5	•		•	•
6		•	•	•
7			•	•

■評価

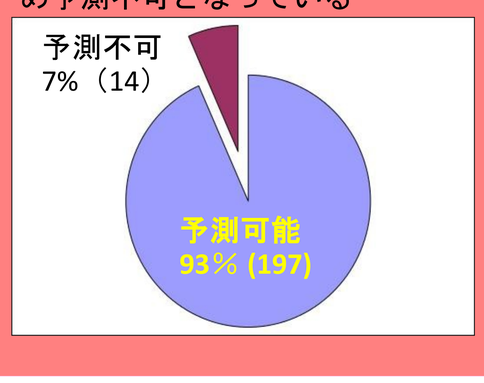
ほぼ100％の予測が可能

RxList
WebMD

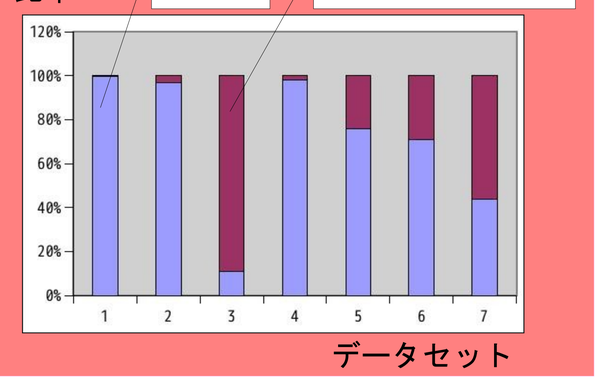
SIDER 2データベースに登録されていない薬9件による評価

◆発症する副作用

該当副作用が登録されていないため予測不可となっている



◆副作用の発症率



薬候補の発見

■手法

ドラッグ・リポジショニングの考え方を応用

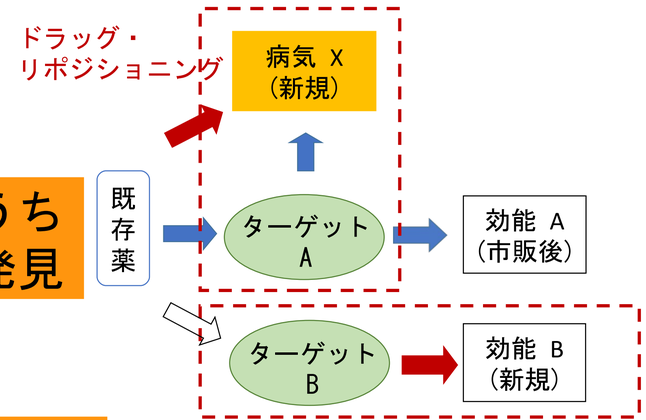
■評価

◆SNPや病因遺伝子による疾患グループ化

薬のない543件の疾患のうち328件の疾患は薬候補を発見

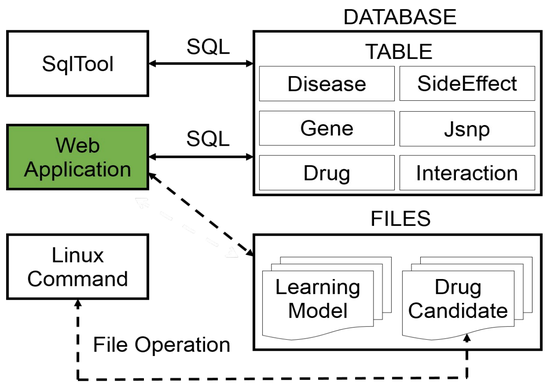
◆既存薬と病因遺伝子の新規な相互作用

32件の新規な相互作用を発見

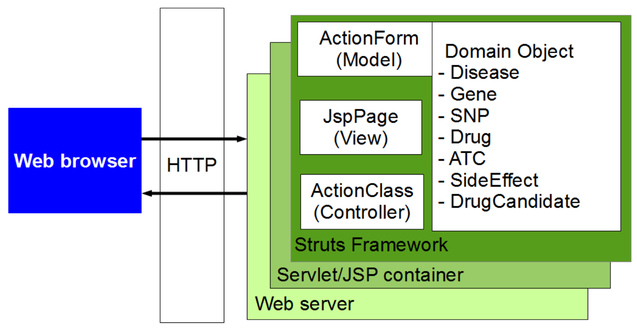


アプリケーション開発

■システム構成



■Webソフトウェア構成



Tomcatのポート番号は8080（デフォルト設定）

クラスタリング

●部分的最小二乗法（PLS）による回帰分析

データセットNo.4の場合

$y' = a1 * スコア + a2 * 部位 + a3 * 遺伝子 + b$

(y‘:予測値, a1 : 説明変数*スコア*の係数, a2 : 説明変数*部位*の係数, a3 : 説明変数*遺伝子* の係数, b:intercept)

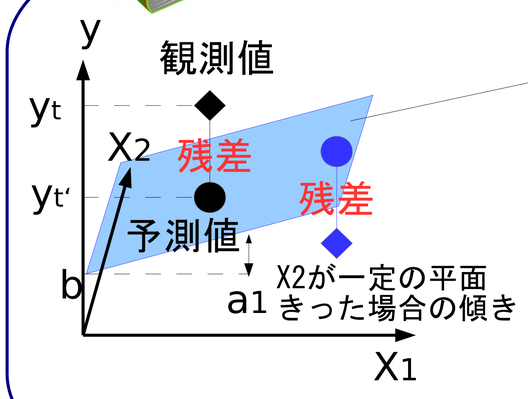
●PLS回帰式を用いた判別分式による2値化

判別分式　$f(x) = y - y'$　(y: 観測値, y‘: 予測値)

判別ルール　If $f(x) \geq 0$ Then $sgn[f(x)] = 1$ &

If $f(x) < 0$ Then $sgn[f(x)] = -1$

最小二乗法の考え方（重回帰の場合）



説明変数が2つ（X_1,X_2）の場合

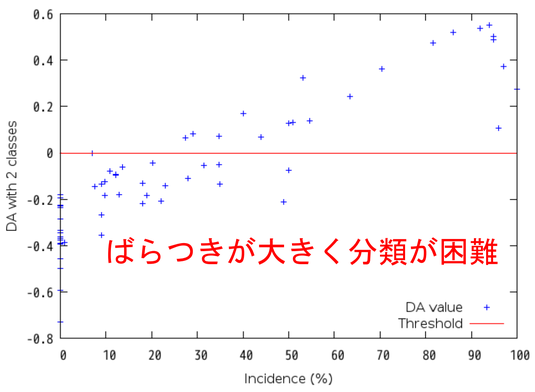
回帰平面

$y_{t'} = a1 * X_{1t} + a2 * X_{2t} + b$

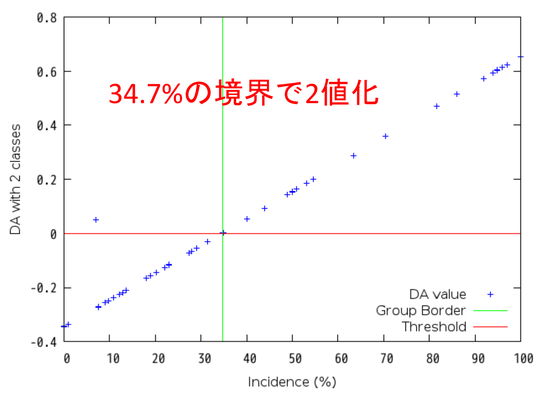
- 回帰平面を標本（観測値）のなるべく「近く」に通す
- 「近く」を観測値のy軸方向の距離（残差）で図る
- 数学的に扱いやすいように残差の2乗の和（残差平方和）を最小にする→最小二乗法

回帰式を用いた判別分析の比較

データセットNo.4の場合（閾値=0）



重回帰式による判別分析



PLS回帰式による判別分析

◆クラスタリング性能評価

データセット	クラスタリング	正答率(%)	データセット	クラスタリング	正答率(%)	データセット	クラスタリング	正答率(%)
1	41%-100%	99	4	34.7%-100%	100	6	50%-100%	100
2	55%-100%	96		27.4%-100%	100		34.7%-100%	100
	29%-100%	100		12.1%-100%	100		27.4%-100%	100
	9%-100%	100		94%-100%	100		18%-100%	100
3	70.3%-100%	100	5	34.7%-100%	100	7	12.1%-100%	100
	36%-100%	99		27.4%-100%	100		81.7%-100%	100
4	18%-100%	100	6	12.1%-100%	100		49%-100%	100
	81.7%-100%	100		53.2%-100%	95		18%-100%	100
	63.5%-100%	100		51%-100%	100			

学習予測

●SVMによるクラスタリングパターンの学習と予測

データセットNo.4の場合

入力空間　$X = \{(スコア1, 部位1, 遺伝子1), \ldots, (スコアn, 部位n, 遺伝子n)\}$

出力ドメイン　$Y = \{1, -1\}$

クラス分類式　$f(x) = \langle w \cdot x \rangle + b$　(w: weight vector, b: bias)

分類ルール　If $f(x) \geq 0$ Then $sgn[f(x)] = 1$　(positive class)

If $f(x) < 0$ Then $sgn[f(x)] = -1$ (negative class)

◆SVM学習モデル性能評価

データセット	クラスタリング	正答率(%)	データセット	クラスタリング	正答率(%)	データセット	クラスタリング	正答率(%)
1	41%-100%	100	4	81.7%-100%	100	6	53.2%-100%	100
	0.1-100%	100		63.5%-100%	100		51%-100%	100
2	55%-100%	23		34.7%-100%	100		50%-100%	100
	29%-100%	95		27.4%-100%	100		34.7%-100%	100
	9%-100%	40		12.1%-100%	100		27.4%-100%	98
3	0.1%-100%	100	5	0.1%-100%	100		18%-100%	98
	70.3%-100%	100		94%-100%	100		12.1%-100%	99
	36%-100%	90		34.7%-100%	100	7	0.1%-100%	100
	18%-100%	20		27.4%-100%	99.5		81.7%-100%	100
	0.1%-100%	86		12.1%-100%	99.6		49%-100%	100
				0.1%-100%	100		18%-100%	99
							0.1%-100%	100

まとめ

1．副作用予測はほぼ100%可能

データ同士の相関関係からの予測とリスク回避

2．ビッグデータ解析から薬候補の発見を提案

3．ビッグデータ解析とMDAの連携

データモデリングによるアプリケーション開発

展望

■ビッグデータによる未来予測

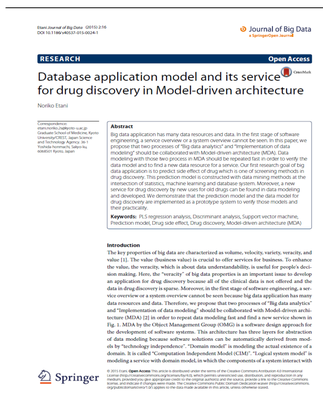
データ間にある相関関係から予測を行い、将来のリスクを回避

ビッグデータのセキュリティやプライバシー保護は重要
お金になるデータはネットワーク犯罪者が狙っている

■個別化医療に向けた「個別化創薬」への取り組み

SNP解析により、体質や病気へのかかりやすさなどの個人差が、どの遺伝子と関係しているかを科学的に解明され、病気に関連する遺伝子と作用する薬の候補品が見つかれば、治療効果が高く、副作用の少ない薬の登場が期待される。また、一人一人に最適な治療方法や薬の調合・投与量などを提供することができる。

成果


Journal of Big Data

RESEARCH　Open Access

Database application model and its service for drug discovery in Model-driven architecture

Noriko Etani

2015年8月7日、本研究論文はジャーナルに採録されました

NORIKO ETANI

Database application model and its service for drug discovery in Model-driven architecture.

Journal of Big Data.2015, 2:16, DOI:10.1186/s40537-015-0024-1.

http://www.journalofbigdata.com/content/2/1/16

開発環境

■オープンソース

OS　Windows8

Java　1.8.0_60

統合開発環境　Eclipse 4.5.0 Mars

Eclipse 日本語化プラグイン　Pleiades 1.6.0

Eclipse Tomcatプラグイン　Sysdeo 3.3.1

Webコンテナ/HTTP/Webサーバ　Tomcat 8.0.26

アプリケーションフレームワーク　Struts 1.3.10

データベース　HSQLDB 2.3.3

機械学習　LibSVM 3.20

■フリーウェア

ドメインオブジェクト　図書管理システム（bcat）