

## 統計的学習による意思決定システム

### -- 知識発見モデルへの実践 --

徐良為 (Xu Liangwei)

([liangweixu2205@gmail.com](mailto:liangweixu2205@gmail.com))

2023-11-18

# 自己紹介

- 名前： 徐 良為 (じょ りょうい)
- 出身： 中国上海
- 現在
  - NTTデータ数理システム 非常勤技術顧問
  - 個人事業主 (屋号 = Statistical Learning Workshop)
    - **機械学習の新しい技術、仕組への研究・実験**
    - **Youtubeに動画作成、統計・機械学習の普及・最新の研究動向など**

- 学歴
  - 学部：中国上海交通大学コンピュータサイエンス
  - 修士 & 博士：東京大学大学院情報工学研究科
- 職歴
  - NTTデータ数理システム データマイニング部 部長
    - 統計・機械学習・データ解析の受託・システム開発

# 機械学習の問題点

- モデル構築に、膨大な数のサンプルが必要

- ▶ 人間の子供に、数個の「机」を見せるだけで、「机」と認識 ← 「知識」の遺伝

- モデルの不可読性（ブラックボックス）

- ▶ 予測パフォーマンスの獲得に見合うモデルの複雑度合？

妨げ

- 自らの知識発見、発明などに必要な論理的な推論

- 知識の「遺伝」

- 人間とのコラボレーション

- 意思決定の道具としても限界

# 提案： ルールの線形結合モデル

## ■知識ルール： 条件（説明変数） ➡ 結論（目的変数）

BMI(肥満度) が 30 以上となると、高血圧になる確率は 0.7

塩分摂取量は 8g 以上の場合は、平均収縮期血圧は 150

## ■ルール

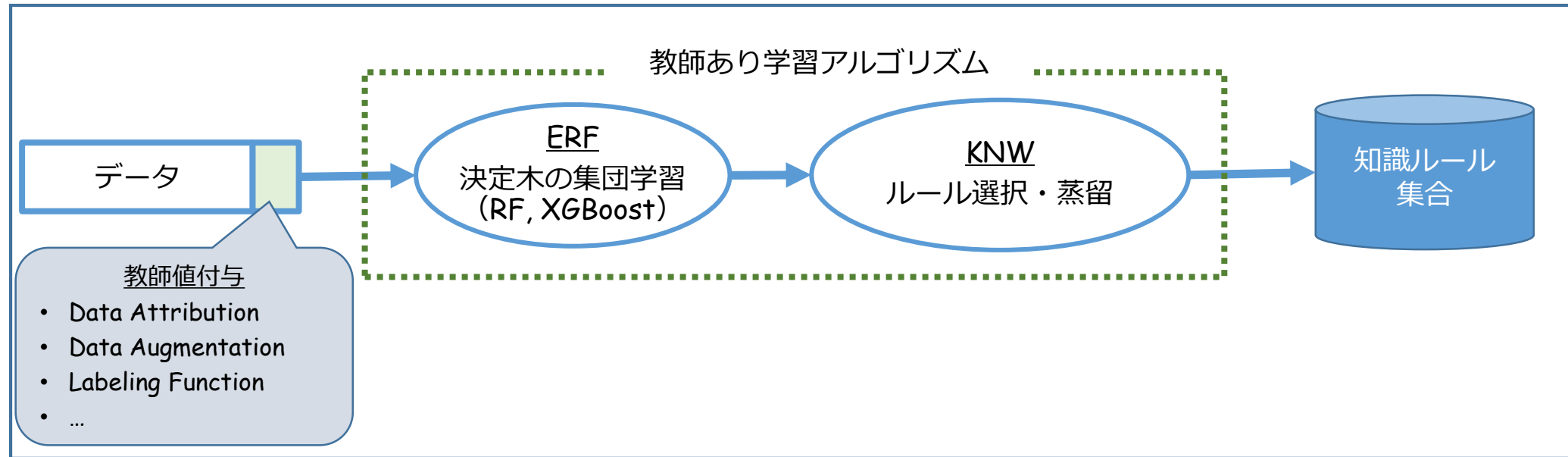
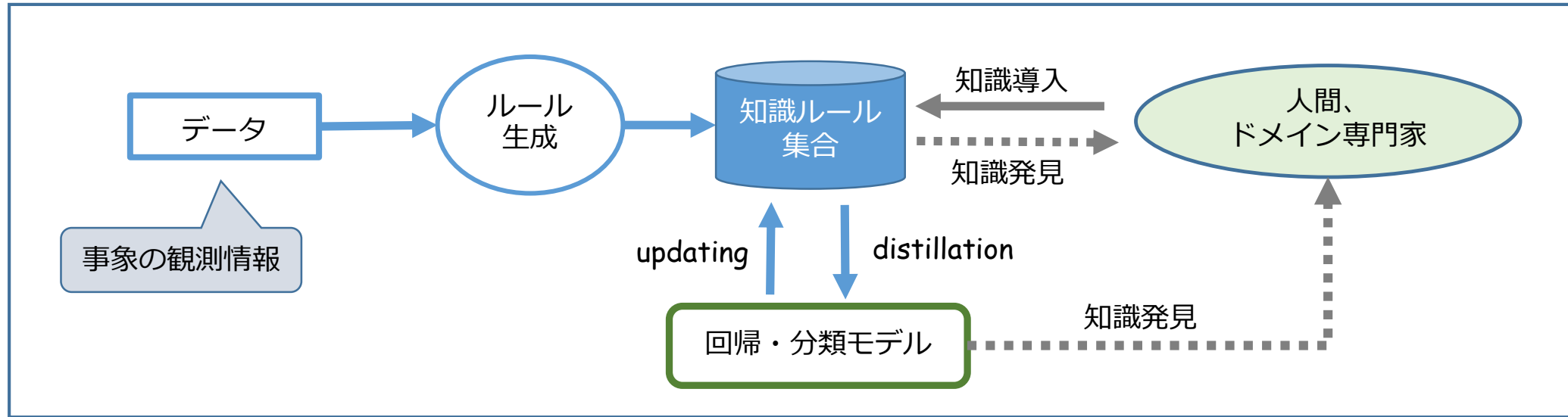
$$\begin{aligned} r(x) &= 1 \quad \text{if } cond(x) \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$x$  は説明変数を表し、 $cond(x)$ は $x$ に関する条件式を表す

## ■モデル(ルールの線形結合)

$$f(x) = \beta_0 + \sum_{i=1}^m \beta_i r_i(x)$$

# AI電卓： データ➡モデル

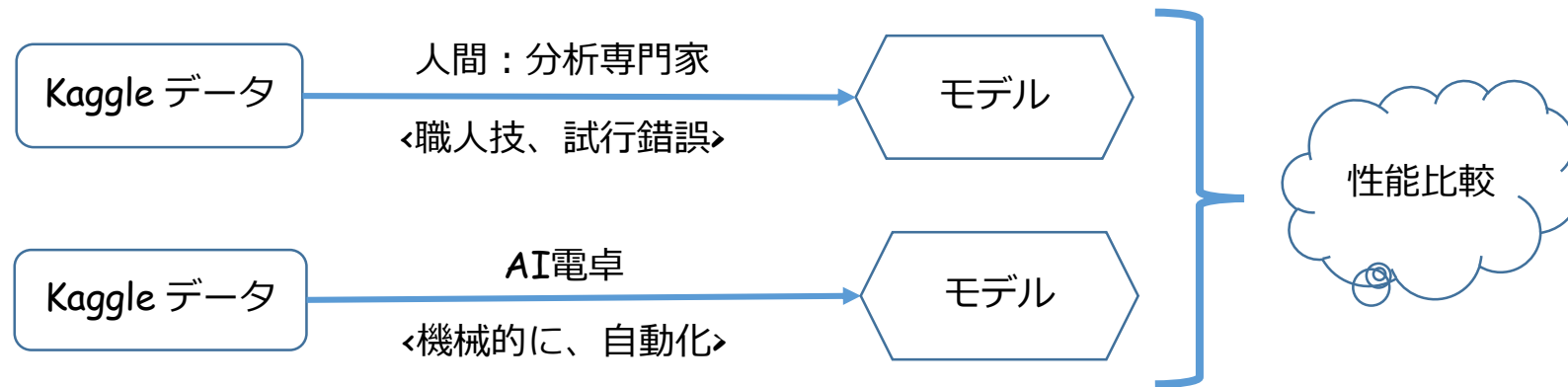


# AI電卓の精度検証： Kaggleデータ分析コンテスト

## ■ Kaggleとは

- Google社が運営、企業や、研究者がデータを投稿し、世界中の統計・機械学習・サイエンティストがデータに最適な予測モデルを競い合う
- 参加者
  - 世界最大規模（?）、約10万人参加
  - 情報科学者、統計学者、経済学者、数学者など
- 様々な業種から様々なテーマ
  - 金融、流通、生物・医療・バイオ、製造、メディア・広告、自然言語処理、有償テーマも多数
- 日本でも、Kaggleの攻略法、体験談などに関する書籍多数

## ■ 目的



# 検証1: House Price

- 不動産に関する属性から、不動産価格を予測するモデル

- 学習用データ（正解を含むデータ）

  - 属性数（列数） : 81

  - 行数 : 1460件

- 検証用データ（正解未知）

  - 属性数（列数） : 80

  - 行数 : 1459件

- モデル評価方法

学習データから構築したモデルを用いて、検証データに対して不動産価格を予測し、予測結果の精度

# データ特徴

SL Viewer - train.csv

	bsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
1	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	208500
2	0	0	0	NA	NA	NA	0	5	2007	WD	Normal	181500
3	0	0	0	NA	NA	NA	0	9	2008	WD	Normal	223500
4	0	0	0	NA	NA	NA	0					140000
5	0	0	0	NA	NA	NA	0	1				250000
6	320	0	0	NA	MnPrv	Shed	700	1				143000
7	0	0	0	NA	NA	NA	0	8	2007	WD	Normal	307000
8	0	0	0	NA	NA	Shed	350	11	2009	WD	Normal	200000
9	0	0	0	NA	NA	NA	0	4	2008	WD	Abnorml	129900
10	0	0	0	NA	NA	NA	0	1	2008	WD	Normal	118000
11	0	0	0	NA	NA	NA	0	2	2008	WD	Normal	129500

nrow=1,460, ncol=81

予測対象：販売価格

- 学習用、検証用両方のデータに、欠損（NA）が多数存在
- 予測対象（不動産価格）の分布にかなりの格差がある
- 不動産価格に関連性が薄い属性が多く見受けられる



# AI電卓でのモデル構築様子

## ■ 下記のような前処理を一切行わず

- 欠損の補填
- 予測対象（不動産価格）の分布への事前調整
- 特徴量抽出

## ■ 学習用 & 検証用データをそのまま導入、学習 & 予測

Variable	<selv>	<selt>	ncate	biggest block	NA	mean
1 Id			0	0.010274	0.000000	730.500000
2 MSSubClass	✓	✓	15	0.367123	0.000000	-inf
3 MSZoning	✓	✓	5	0.788356	0.000000	-inf
4 LotFrontage	✓	✓	0	0.193151	0.177397	70.049957
5 LotArea	✓	✓	0	0.295205	0.000000	10516.828125
6 Street	✓	✓	2	0.995890	0.000000	-inf
7 Alley	✓	✓	3	0.937671	0.937671	-inf
8 LotShape	✓	✓	4	0.633562	0.000000	-inf
9 LandContour	✓	✓	4	0.897945	0.000000	-inf
10 Utilities	✓	✓	2	0.999315	0.000000	-inf
11 LotConfig	✓	✓	5	0.720548	0.000000	-inf
12 LandSlope	✓	✓	3	0.946575	0.000000	-inf
13 Neighborhood	✓	✓	25	0.154110	0.000000	-inf
14 Condition1	✓	✓	9	0.863014	0.000000	-inf
15 Condition2	✓	✓	8	0.989726	0.000000	-inf
16 BldgType	✓	✓	5	0.835616	0.000000	-inf
17 HouseStyle	✓	✓	8	0.497260	0.000000	-inf
18 OverallQual	✓	✓	0	0.271918	0.000000	6.099315
19 OverallCond	✓	✓	0	0.562329	0.000000	5.575342
20 YearBuilt	✓	✓	0	0.079452	0.000000	1971.267822
21 YearRemodAdd	✓	✓	0	0.121918	0.000000	1984.865723
22 RoofStyle	✓	✓	6	0.781507	0.000000	-inf
23 RoofMatl	✓	✓	0	0.001100	0.000000	-inf

Variable	<selv>	<selt>	ncate	biggest block	NA	mean
1 OverallQual	✓	✓	0	0.271918	0.000000	0.000000
2 Neighborhood	✓	✓	0	0.154110	0.000000	0.000000
3 GrLivArea	✓	✓	0	0.051370	0.000000	0.000000
4 ExterQual	✓	✓	0	0.620548	0.000000	0.000000
5 BsmtQual	✓	✓	0	0.444521	0.025342	0.000000
6 KitchenQual	✓	✓	0	0.503425	0.000000	0.000000
7 GarageCars	✓	✓	0	0.564384	0.000000	0.000000
8 GarageArea	✓	✓	0	0.055479	0.000000	0.000000
9 TotalBsmtFt	✓	✓	0	0.093151	0.000000	0.000000
10 1stFlrSF	✓	✓	0	0.066438	0.000000	0.000000
11 FullBath	✓	✓	0	0.526027	0.000000	0.000000
12 Garage1	✓	✓	0	0.414384	0.055479	0.000000
13 Fireplaces	✓	✓	0	0.472603	0.472603	0.000000
14 TotRms	✓	✓	0	0.275342	0.000000	0.000000
15 YearBuilt	✓	✓	0	0.079452	0.000000	0.000000
16 YearRemodAdd	✓	✓	0	0.121918	0.000000	0.000000
17 Foundation	✓	✓	0	0.443151	0.000000	0.000000
18 GarageType	✓	✓	0	0.595890	0.055479	0.000000
19 MSSubClass	✓	✓	0	0.367123	0.000000	0.000000
20 Fireplaces	✓	✓	0	0.472603	0.000000	0.000000
21 BsmtFinType1	✓	✓	0	0.294521	0.025342	0.000000

結果  
モデル

# Kaggleからの成績評価(2023-5-25)

**House Prices - Advanced Regression Techniques**  
Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 4,728 teams · Ongoing

Overview Data Code Discussion **Leaderboard** Rules Team Submissions **Submit Predictions** ...

**Leaderboard** Raw Data Refresh

YOUR RECENT SUBMISSION

✓ **submission2.csv** Score: 0.11728  
Submitted by Xu Liangwei · Submitted 15 minutes ago

90	M.R.0024		0.11261	2	2mo
91	ningnujiel		0.11292	3	11d
92	<b>Xu Liangwei</b>		0.11385	271	16m

😊 Your Best Entry!  
Your submission scored 0.11728, which is not an improvement of your previous score. Keep trying!

参加チーム数 : 4728  
順位 : 92  
上位 : 1.94%

# 検証2: Store Sales = 時系列データ分析

- 食料品チェーン店 (grocery store) の日々の売上実績から、将来の売上を予測する
- 店舗数 (store\_nbr) = 54、商品分類 (family) 数=33
- 店舗毎、商品分類毎の売上に関する時系列数 =  $54 * 33 = 1782$
- 学習用データ (正解を含むデータ)
  - 売上履歴 (日数) : 1680日
  - 行数 : 294万件、属性数 (列数) : 17
- 検証用データ (正解未知)
  - 予測対象日数 : 16日
  - 行数 : 28512件 (= 店舗数 \* 商品分類数 \* 予測対象日数)
- モデル評価方法  
学習データから構築したモデルを用いて、検証データに対して売上を予測し、予測精度

# データ特徴

date	store_nbr	Store_type	Store_cluster	Store_city	Store_state	family	sales	onpromotion	dcoilwtico	Holiday	SalaryDay	transact
2014-07-30	17	C	12	Quito	Pichincha	MAGAZINES	0.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	MEATS	226.317001	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PERSONAL CARE	195.000000	1	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PET SUPPLIES	0.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PLAYERS AND ELECTRO	5.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	POULTRY	0.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PREPARED FOODS	0.000000	0	104.290001		no	1098
2014-07-30	17	C	12	Quito	Pichincha	PRODUCE	0.000000	0	104.290001		no	1098

商品分類

予測対象：売上

プロモーションFG

店舗情報

振替休日・給料日

日ごとの石油価格




- 時系列データ（店舗、商品ごと、日々売上）以外の「外部要因」も含まれる
- 予測対象（売上金額）の分布にかなりの格差（桁違い）がある
- データの行数は比較的が多い（300万件前後）

# Kaggleからの成績評価(2023-9-26)

**Store Sales - Time Series Forecasting**  
Use machine learning to predict grocery sales

Kaggle · 695 teams · Ongoing

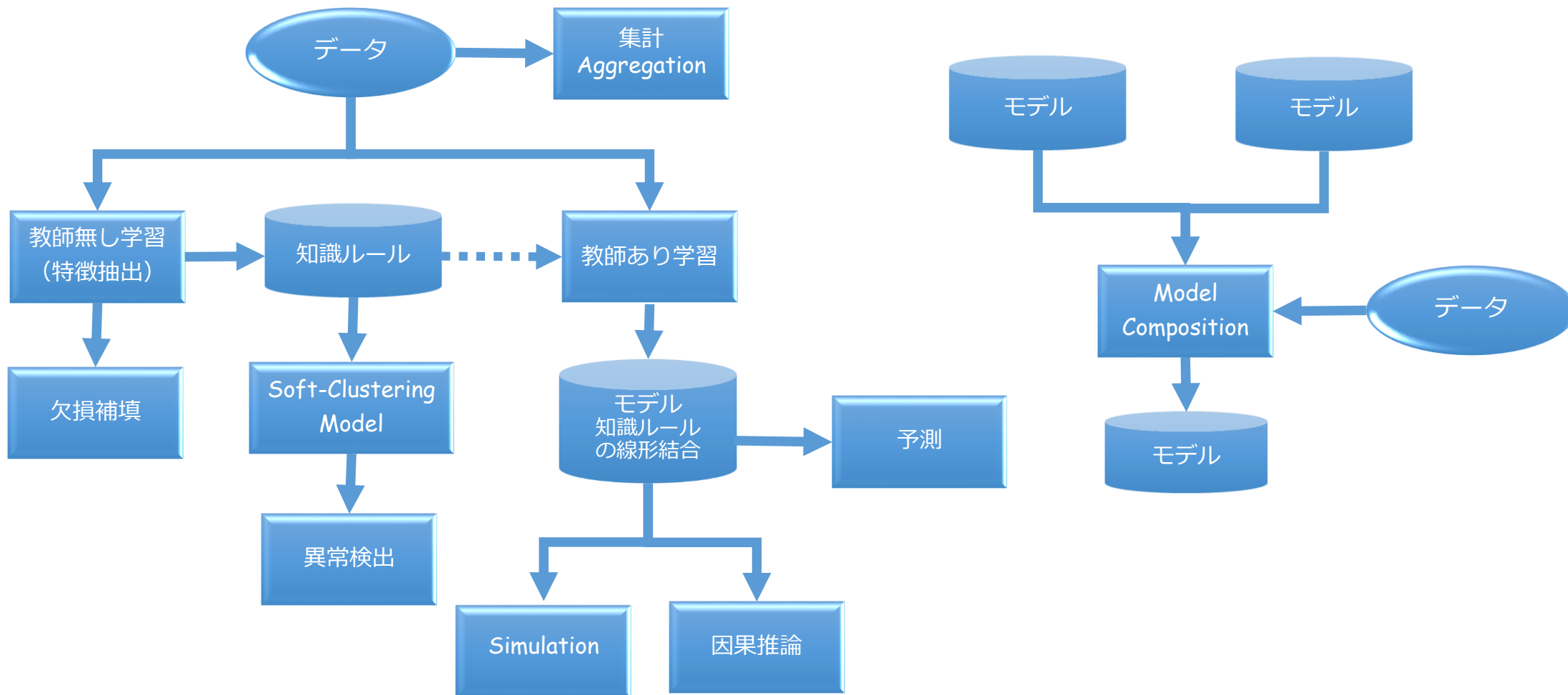
Overview Data Code Discussion **Leaderboard** Rules Team Submissions **Submit Predictions** ...

39	Will Gilchrist		0.39261	2	12d
40	Mario Refoyo López		0.39495	4	1mo
41	<b>Xu Liangwei</b>		0.39778	110	5m
 Your Best Entry! Your most recent submission scored 0.39778, which is an improvement of your previous score of 0.40014. Great job!					
42	David Gilbertson		0.39842	20	1mo

参加チーム数 : 695  
順位 : 41  
上位 : 5.89%

**Tweet this**

# AI電卓モデル関連機能

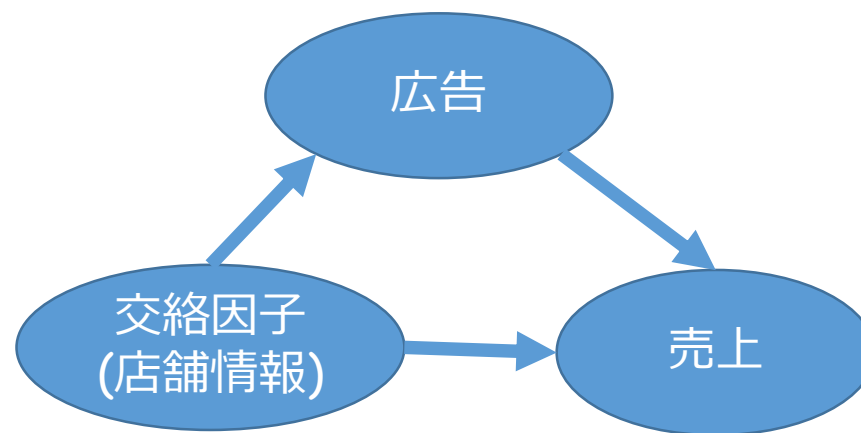




# 因果推論：店舗の選択(売上が広告から受ける影響)

店舗	競合店	人口構成	立地条件	...	売上	広告	売上.広告有	売上.広告無
s1	...	...	...	...	198	有	198	?
s2	...	...	...	...	55	無	?	55
s3	...	...	...	...	77	無	?	77
s4	...	...	...	...	160	有	160	?
s5	...	...	...	...	143	無	?	143
...	...	...	...	...	...	...	...	...

他の因子 = 店舗情報を考慮する



# AI電卓での使用方法(学習)

The screenshot shows the SL Viewer interface with a data table and a Causal Inference configuration window. A context menu is open over the data table, with 'Causal Inference' selected. A blue arrow points from the 'Train' option in the menu to the configuration window.

**medical\_data10K**

	年齢	性別	身長	体重	BMI_算出
1	50	女	156.453491	31.912710	22.422974
2	70	男	155.875275	68.428970	23.457228
3				70.703918	22.8938
4				58.539978	22.8938
5				69.74126	24.016193
6				58.79709	22.893896
7				50.37464	22.893896
8					22974
9					64598
10					22.422974
11					22.893896
12					23.064598
13					22.243906
14					23.852085
15	60	女	151.672333	67.700935	23.064598
16	70	女	154.918594	51.709755	22.893896
17	40	女	155.568054	31.305799	22.243906

**Causal Inference ( medical\_data10K )**

model_zip_file	E:/SLWorkshop/SLSystem/project/demo/Heal	...
treatment	性別	...
n_cluster	5	
nloop	1	
nitr	300	
eta	1.0	
nthread	7	

Close Run

Callouts:

- モデル選択 (Model Selection)
- 操作因子 (Treatment Factor)
- 潜在クラスタリングオプション (Latent Clustering Options)



# AI電卓での利用方法(結果)

SL Causal Inference Result - individual\_effect

average\_effect ci\_hst ci\_th sc\_expectation psh prh ph phs phr rules\_model rule\_table

	Name	mean	count	min	max	sigma
1	<収縮期(最高)血圧>	132.384079	10000	88.299660	207.211533	16.816507
2	mu1<性別=女>	132.220947	10000	83.349670	209.396530	16.253786
3	mu1<性別=男>	133.024246	10000	91.381126	207.553207	14.441153
4	tau1	0.803544	10000	-47.736763	48.174957	9.732852
5	mu2<性別=女>	132.279709	10000	79.666718	220.124939	17.387304
6	mu2<性別=男>	133.155167	10000	34.939934	258.937927	16.838097
7	tau2	0.875645	10000	-105.077797	128.051178	13.289062
8	mu3<性別=女>	132.627853	10000	118.338524	142.589996	7.114850
9	mu3<性別=男>	132.904633	10000	118.338524	142.589996	7.215300
10	tau3	0.276636	10000	-1.510979	4.196259	0.977100
11	mu4<性別=女>	131.837585	10000	80.017227	238.455536	19.745895
12	mu4<性別=男>	133.040131	10000	-40.408798	372.751404	26.644506
13	tau4	1.203473	10000	-170.872681	244.774338	30.977966

ATE  
平均効果

CATE  
条件付き平均効果

individual\_effect

	高血圧>	収縮期(最高)血圧.<predict>	e<性別=女>	e<性別=男>	mu1<性別=女>	mu1<性別=男>	tau1	mu2<性別=女>
1	08.398438	108.538887	0.775850	0.224150	108.538887	116.813217	8.274330	108.538887
2	33.853821	133.527267	0.630454	0.369545	115.211761	133.527267	18.315506	133.527267
3	29.557693	132.439453	0.640542	0.359457	132.439453	127.535881	-4.903572	132.439453
4	38.642776	136.744492	0.640960	0.359039	136.744492	135.460800	-1.283691	136.744492
5	22.598335	123.422363	0.635454	0.364547	160.543076	123.422363	-37.120712	123.422363
6	30.111465	131.263199	0.629460	0.370539	131.263199	132.257385	0.994186	131.263199
7	18.740143	118.352270	0.620460	0.370531	118.352270	121.718681	3.365407	118.352270

nrow=10,000, ncol=63



# 参考動画（検索=「AI電卓」）

Youtube チャンネル

[https://www.youtube.com/@ai\\_dentaku](https://www.youtube.com/@ai_dentaku)

動画（時間順）

1. [背景紹介](#)
2. [背景紹介\(コア技術\)](#)
3. [データ可視化（見る）](#)
4. [分析1：説明変数重要度計算、知識ルールを入力、編集、ラベリング](#)
5. [分析2：モデル構築、評価、知識ルール抽出](#)
6. [潜在クラスタ、モデルの予測根拠、知識発見](#)
7. [因果推論\(入門編\)](#)
8. [因果推論\(AI電卓編\)](#)
9. [Kaggleデータ分析：不動産価格予測](#)
10. [Conformal Prediction による予測信頼区間の推定](#)

# 時系列予測に必要なラグ情報

- 機械学習のモデルは関数  $y = f(x)$  を求める過程である
- 時系列の場合は  $y$  は時間  $t$  に依存することになり、通常  $y_t$  と表現される
- 時間に依存することを利用、モデル ( $f$ ) を構築するときに使える情報は、 $x$  に加え、次の二種類の情報も利用可能
  - 時間軸 (日付) から算出可能な情報 ( $\tau$ )、例えば、週、月、旬、季節など
  - 歴史を表す (lag=ラグ) 情報: 前日の売上 ( $y_{t-1}$ )、2日前の売上 ( $y_{t-2}$ )、 $\dots$ 、 $k$ 日前の売上 ( $y_{t-k}$ )  
 $k$ は、ユーザ指定 1 より大きい自然数となります。
- 上記まとめると、 $y_t$  を予測するモデル ( $f$ ) としては、次のようになる

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-k}, \tau, x)$$