

機械学習超入門



講義をはじめる前に

下記データでIDを除く5項目を一枚のグラフにするにはどうすればいいでしょうか？

ID	country	continent	lifeExp	pop	gdpPercap
1	Argentina	Americas	75.32	40301927	12779.3796
2	Canada	Americas	80.653	33390141	36319.235
3	Cote d'Ivoire	Africa	48.328	18013409	1544.75011
4	Cuba	Americas	78.273	11416987	8948.10292
5	Czech Republic	Europe	76.486	10228744	22833.3085
6	Denmark	Europe	78.332	5468120	35278.4187
7	Eritrea	Africa	58.04	4906585	641.369524
8	Germany	Europe	79.406	82400996	32170.3744
9	Ghana	Africa	60.022	22873338	1327.60891
10	Greece	Europe	79.483	10706290	27538.4119
11	Guatemala	Americas	70.259	12572928	5186.05
12	Kenya	Africa	54.11	35610177	1463.24928
13	Mozambique	Africa	42.082	19951656	823.685621
14	Paraguay	Americas	71.752	6667147	4172.83846
15	Slovenia	Europe	77.926	2009245	25768.2576
16	Uganda	Africa	51.542	29170398	1056.38012
17	United States	Americas	78.242	301139947	42951.6531
18	Venezuela	Americas	73.747	26084662	11415.8057
19	Mauritania	Africa	62.664	3270065	1803.1515
20	Belgium	Europe	79.441	10392226	33692.6051

ID: 識別番号

Country: 国名

Continent: 大陸名

lifeExp: 平均寿命

Pop: 人口

gdpPercap: 一人当GDP

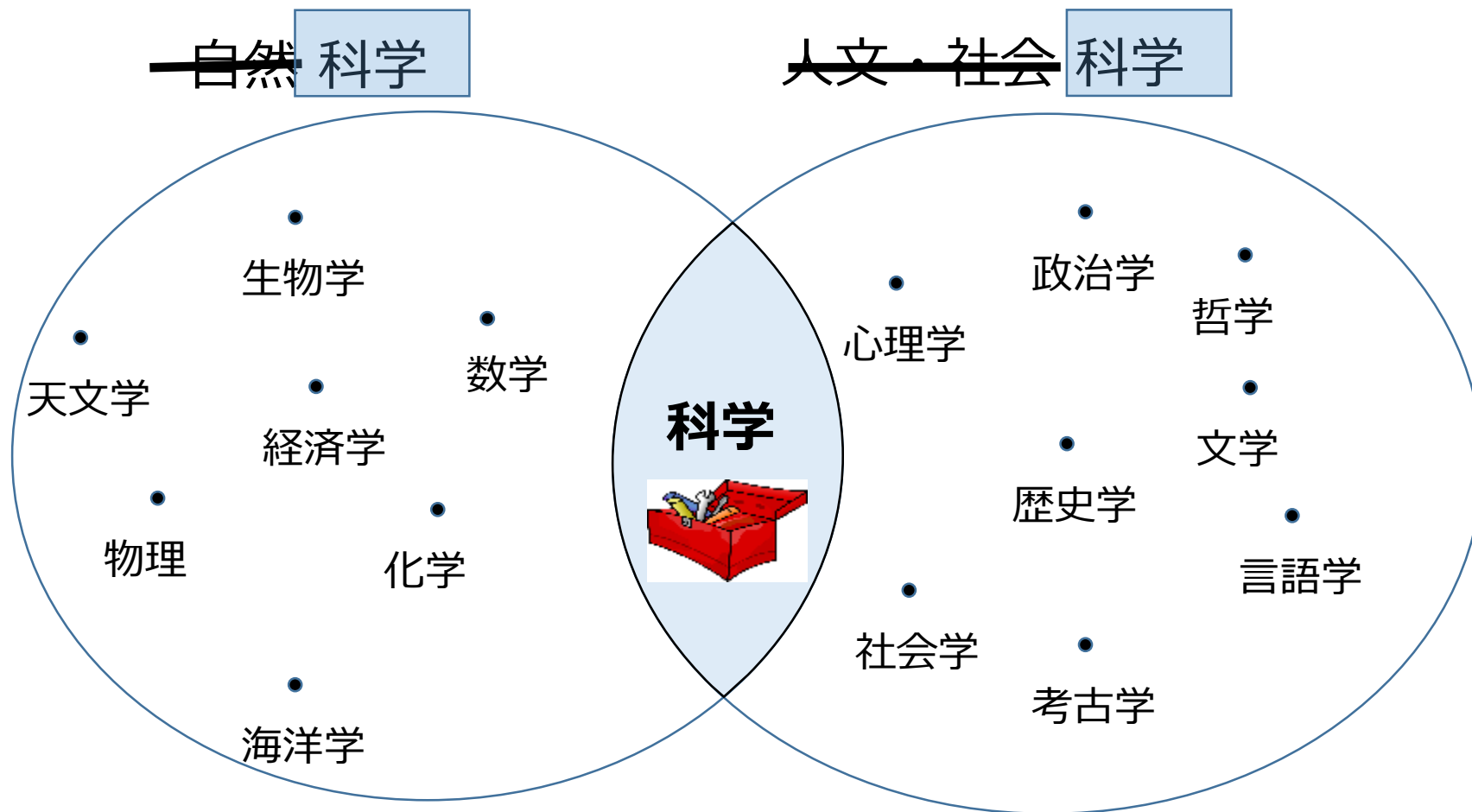
【出典】

gapminderより抜粋

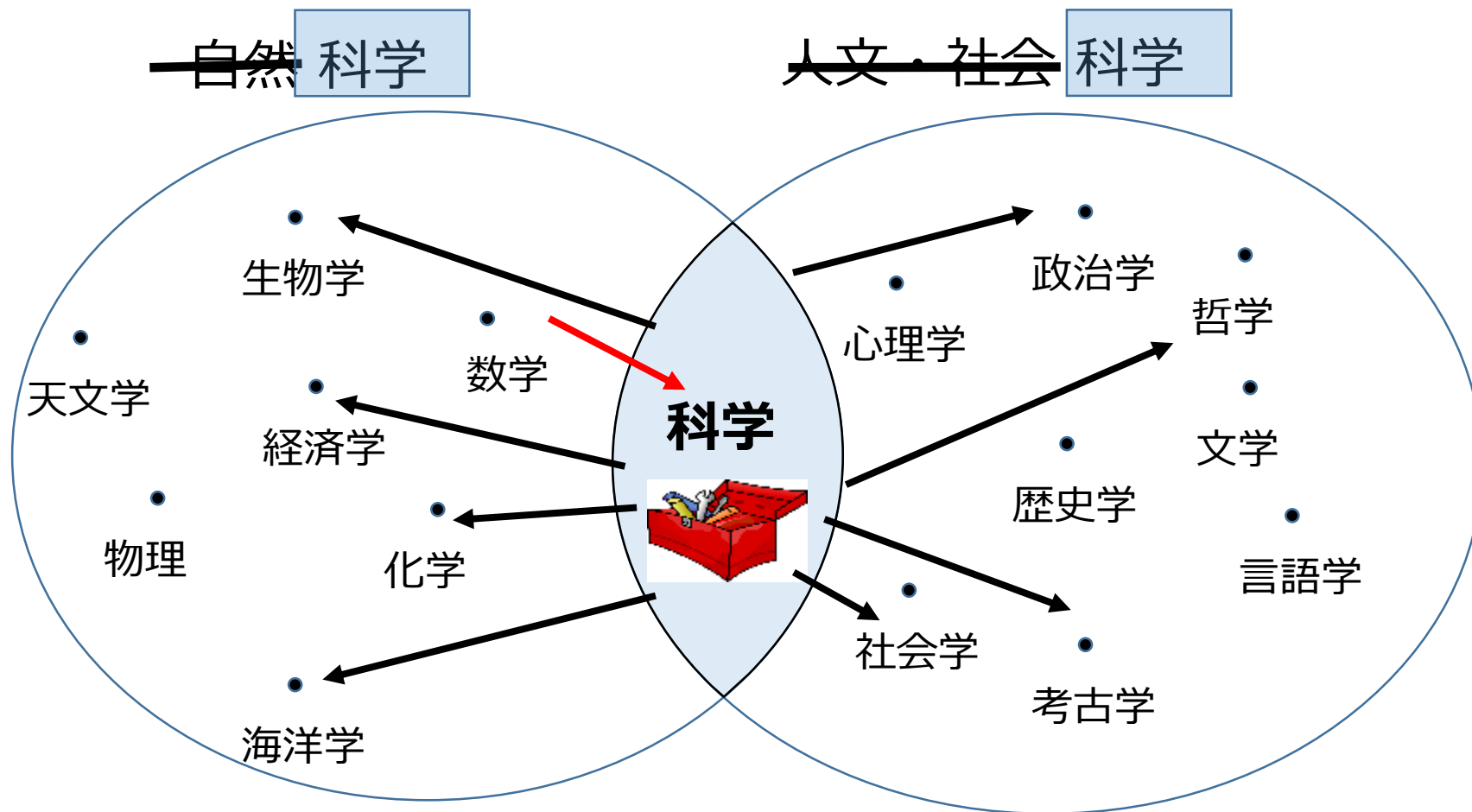
<https://www.gapminder.org/>

データサイエンスとは？

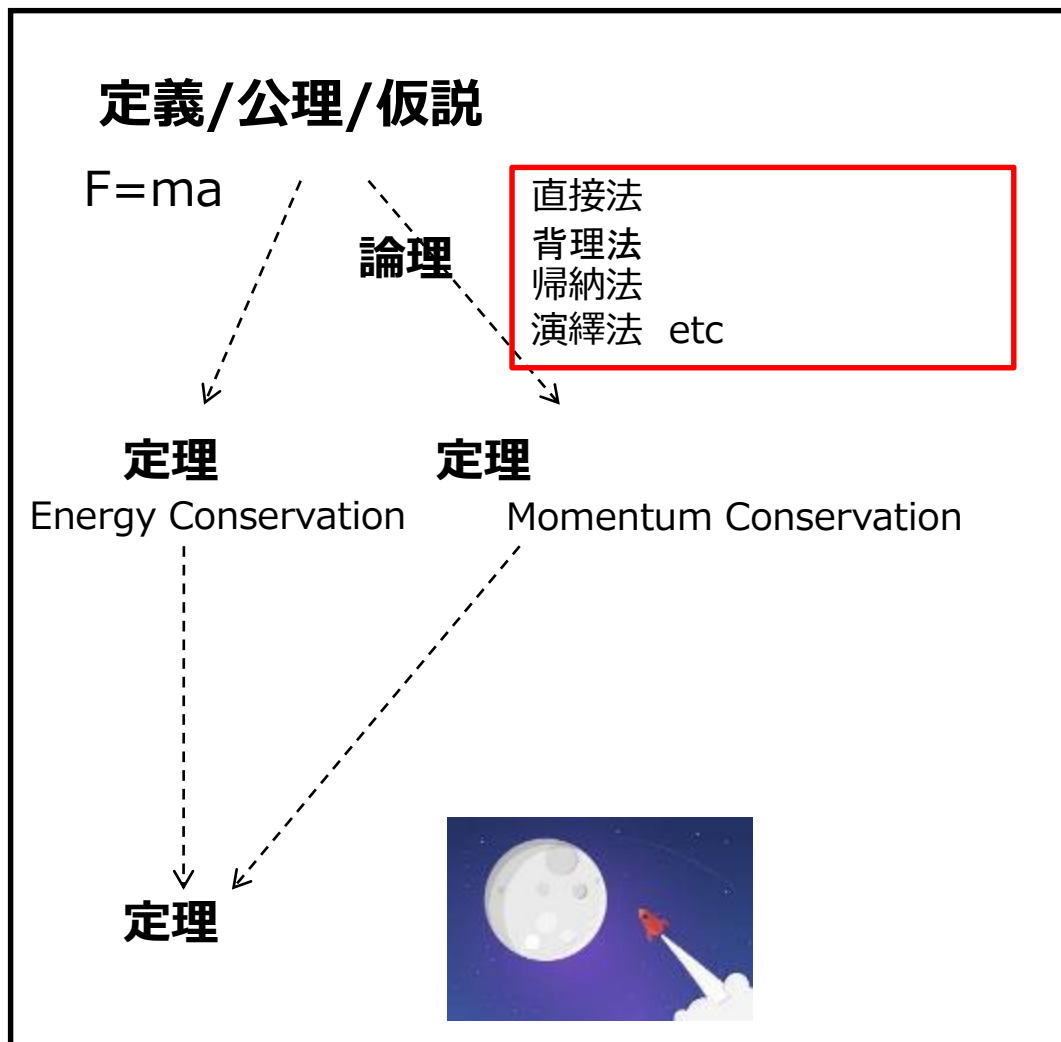
プログラムなしではじめる機械学習超入門



科学理論の構造

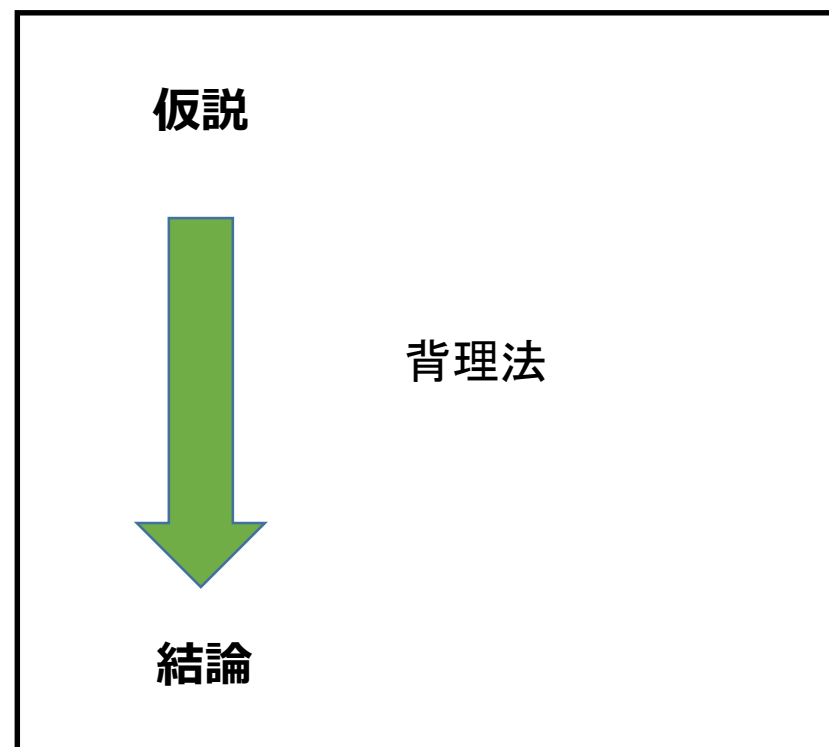


科学理論の構造



Theory

統計検定の論理構造



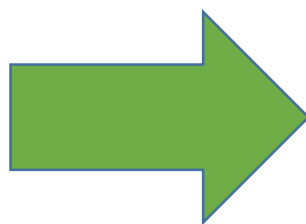
科学的思考による問題解決

化学

医療

数学

ビジネス



「分解と統合」
の哲学

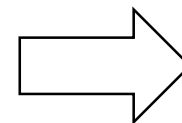
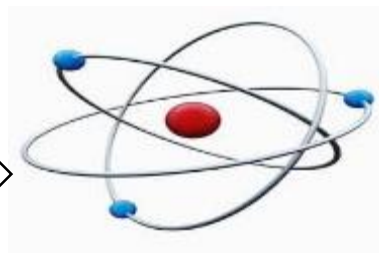
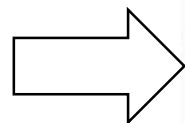


問題解決の為の
共通アプローチ？

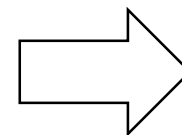
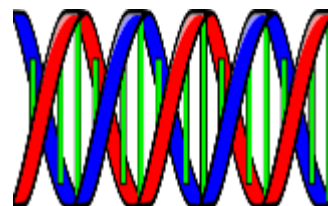
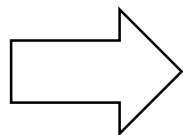
ルネ・デカルト
(1596-1650)

「分解と統合」の哲学

分子

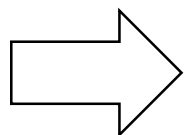


DNA

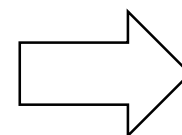


素因数分解

42



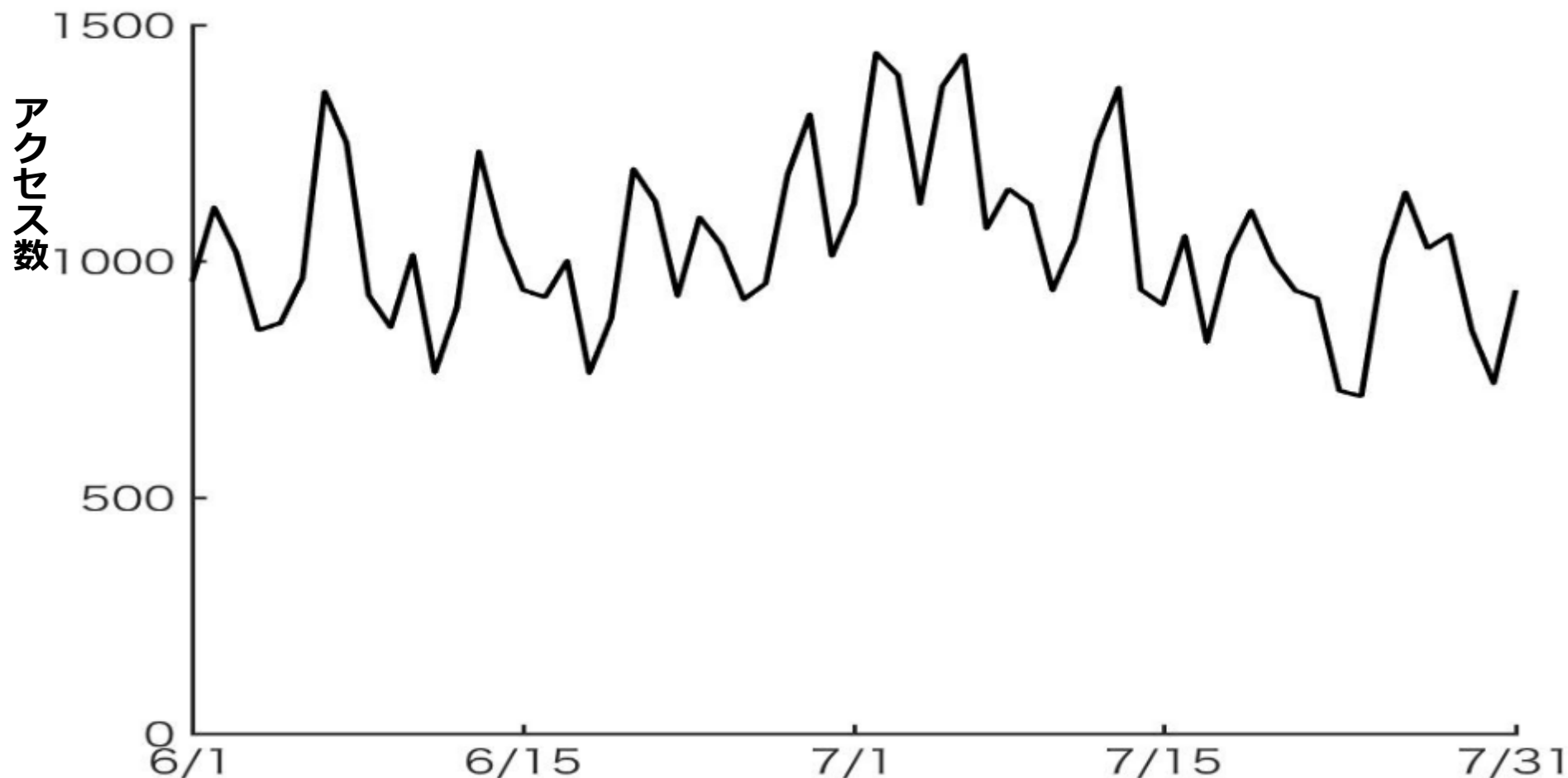
2, 3, 7



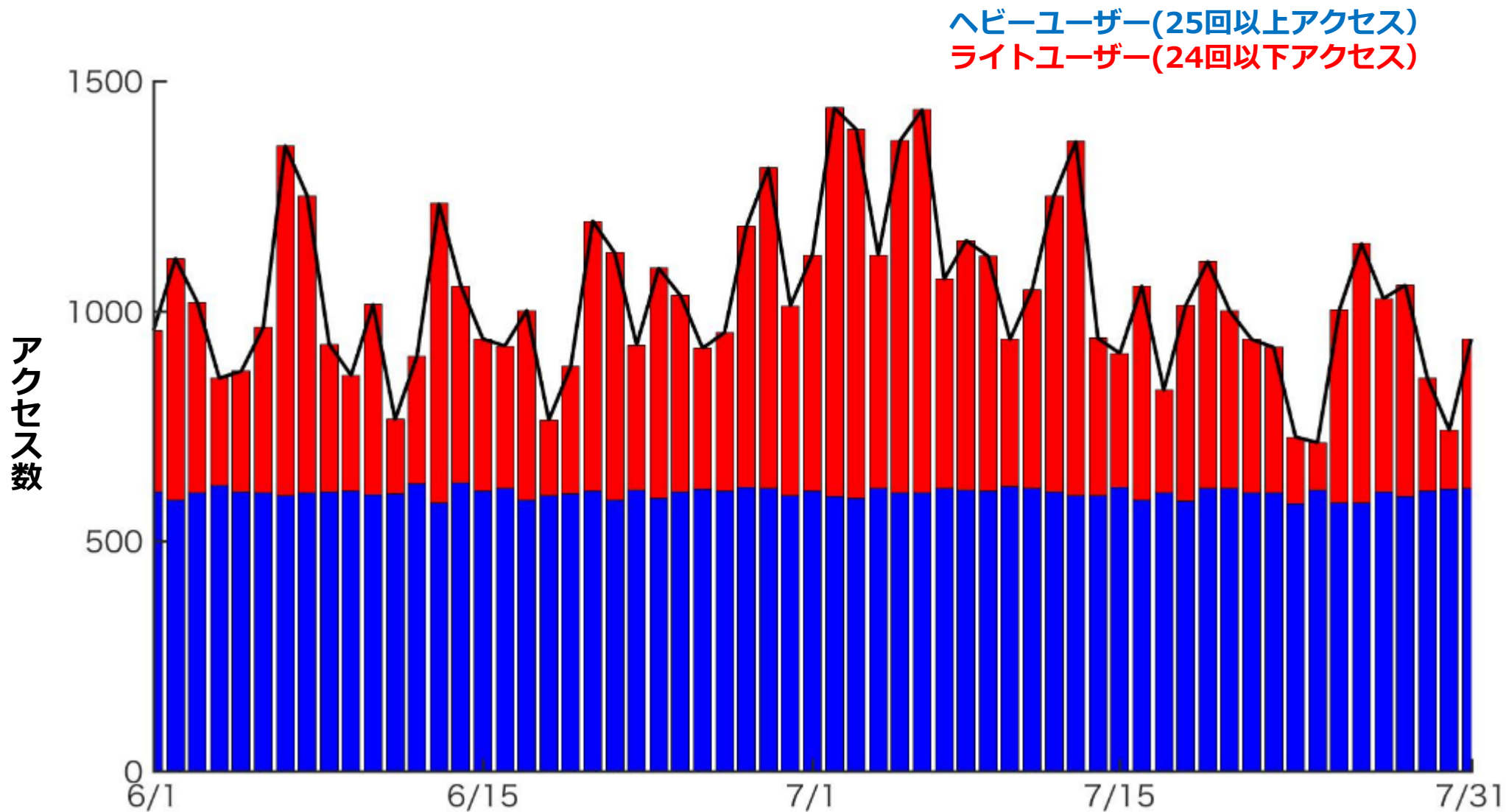
$42 = 2 \cdot 3 \cdot 7$

データ分析の考え方 1

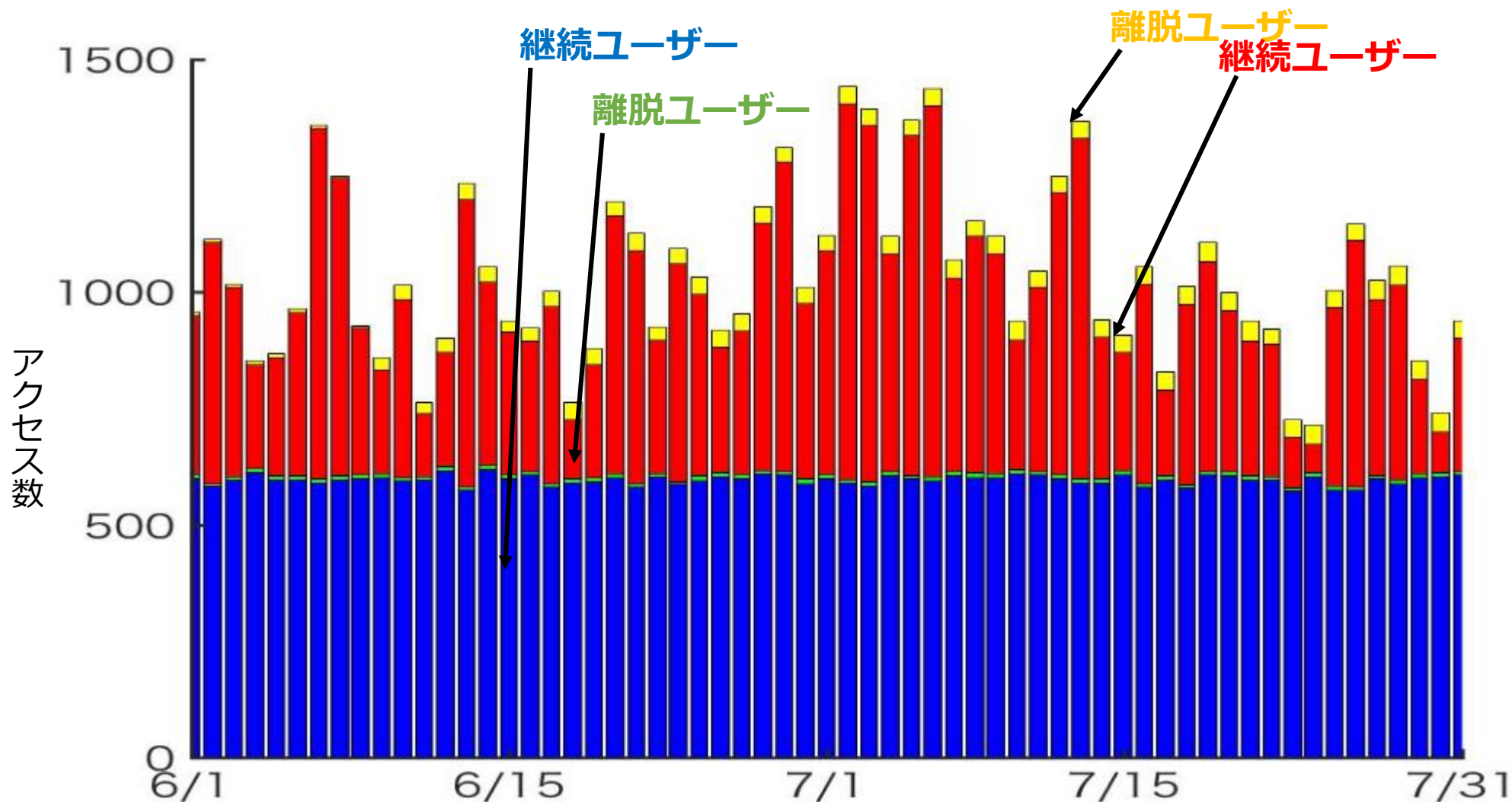
課題：「サイトへの登録者数が減少しているようだが、アクセス数からその原因を調査できないか？」



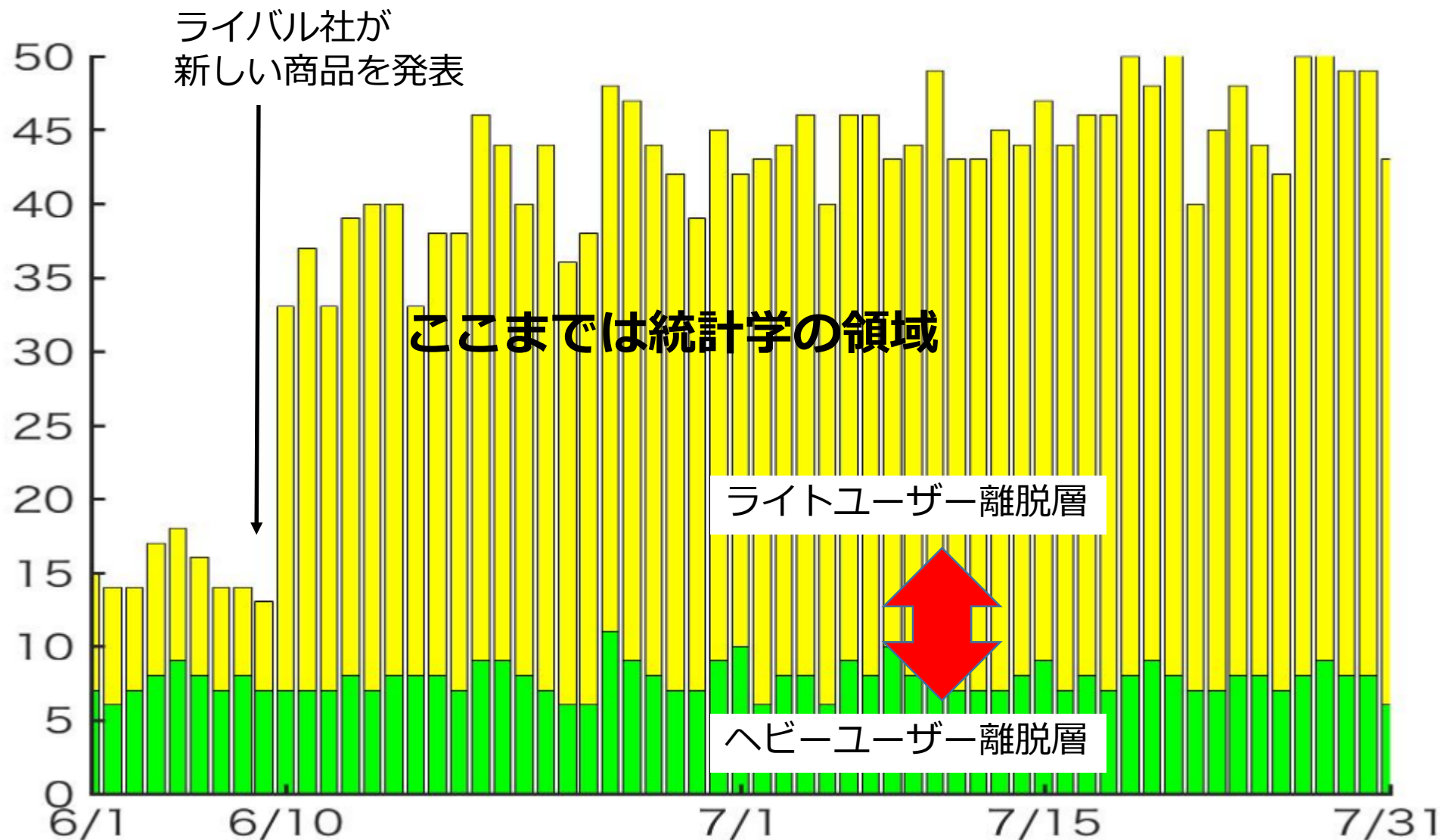
分解と統合



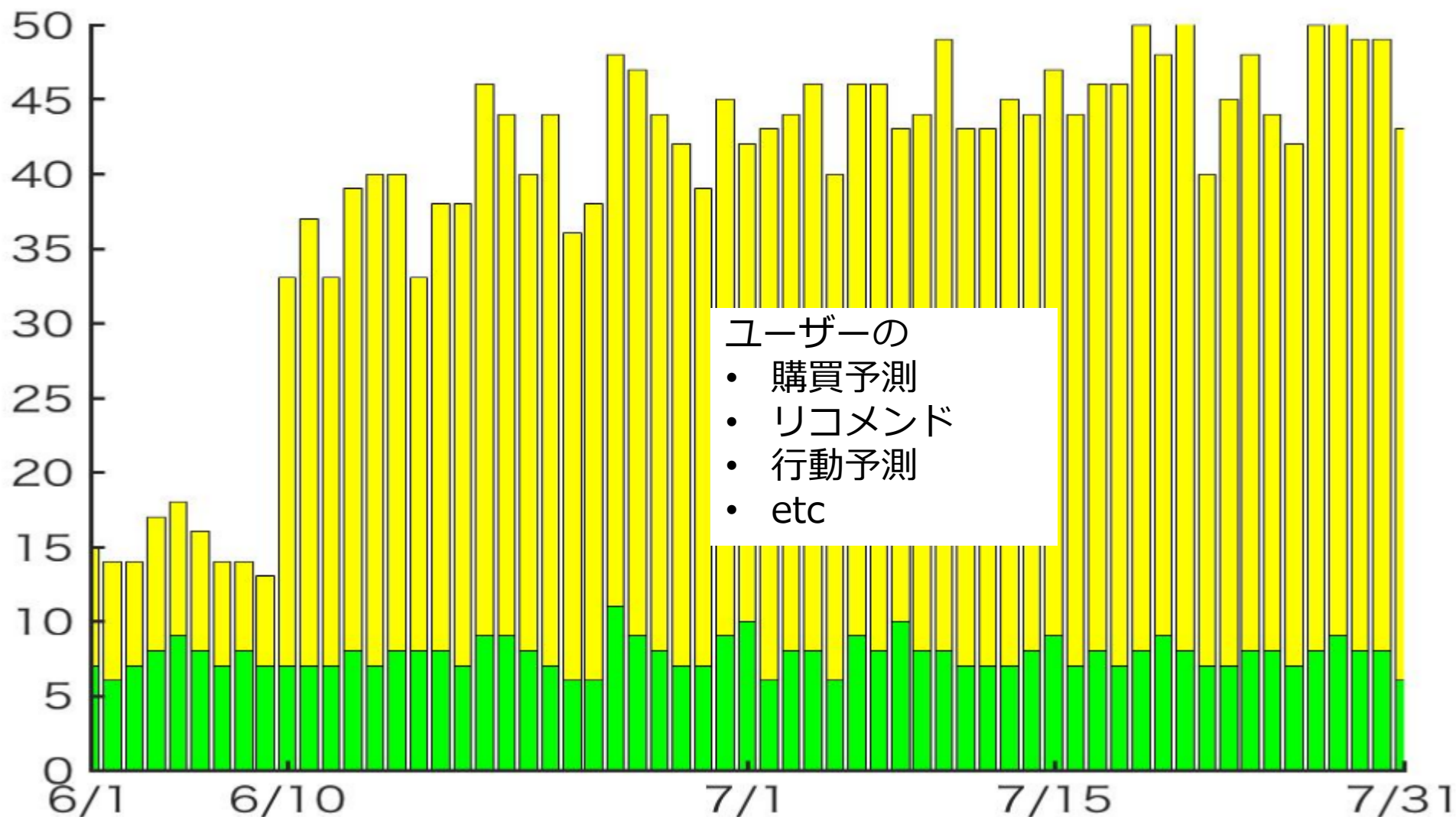
分解と統合



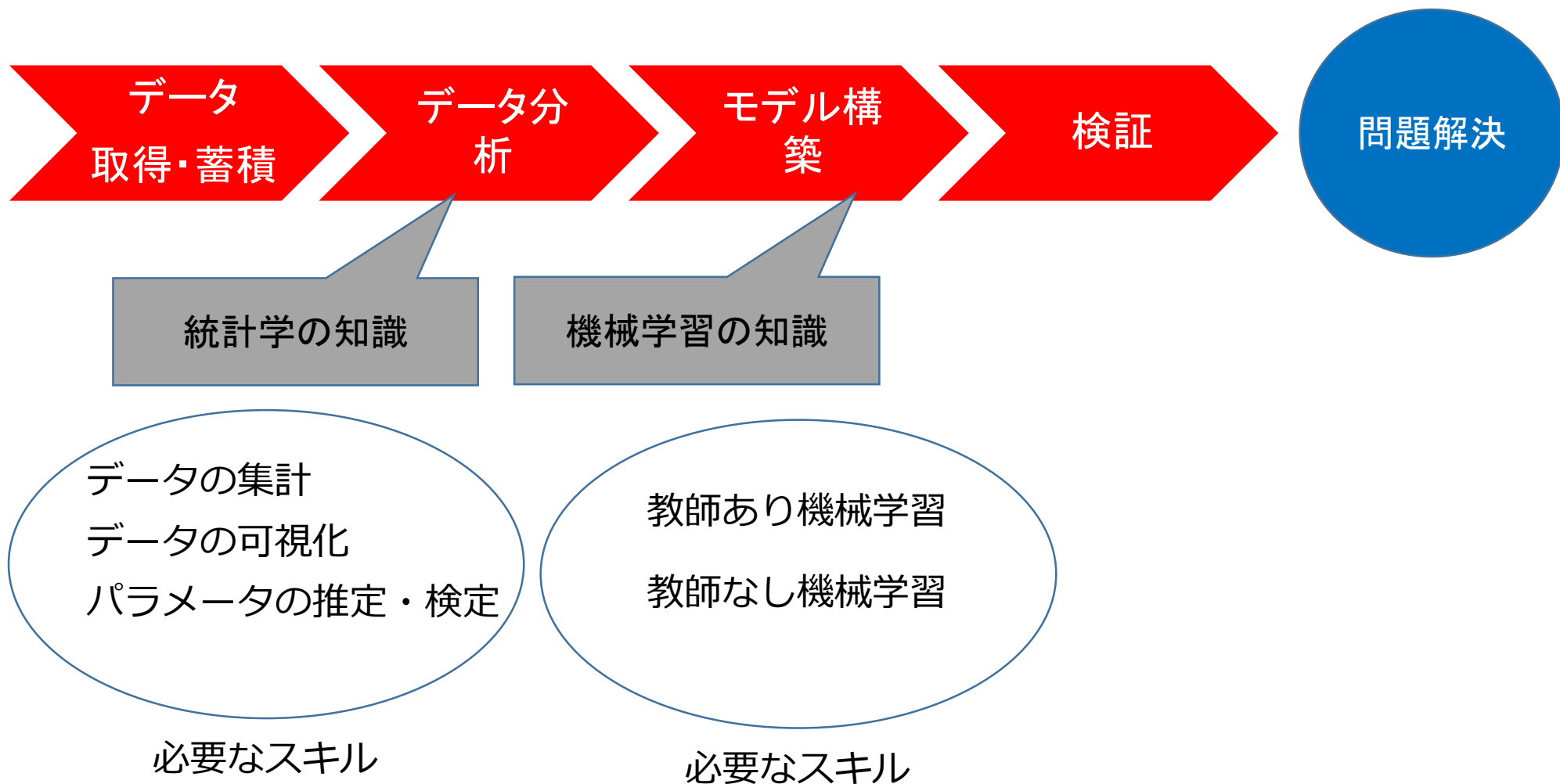
分解と統合



ここから先が機械学習



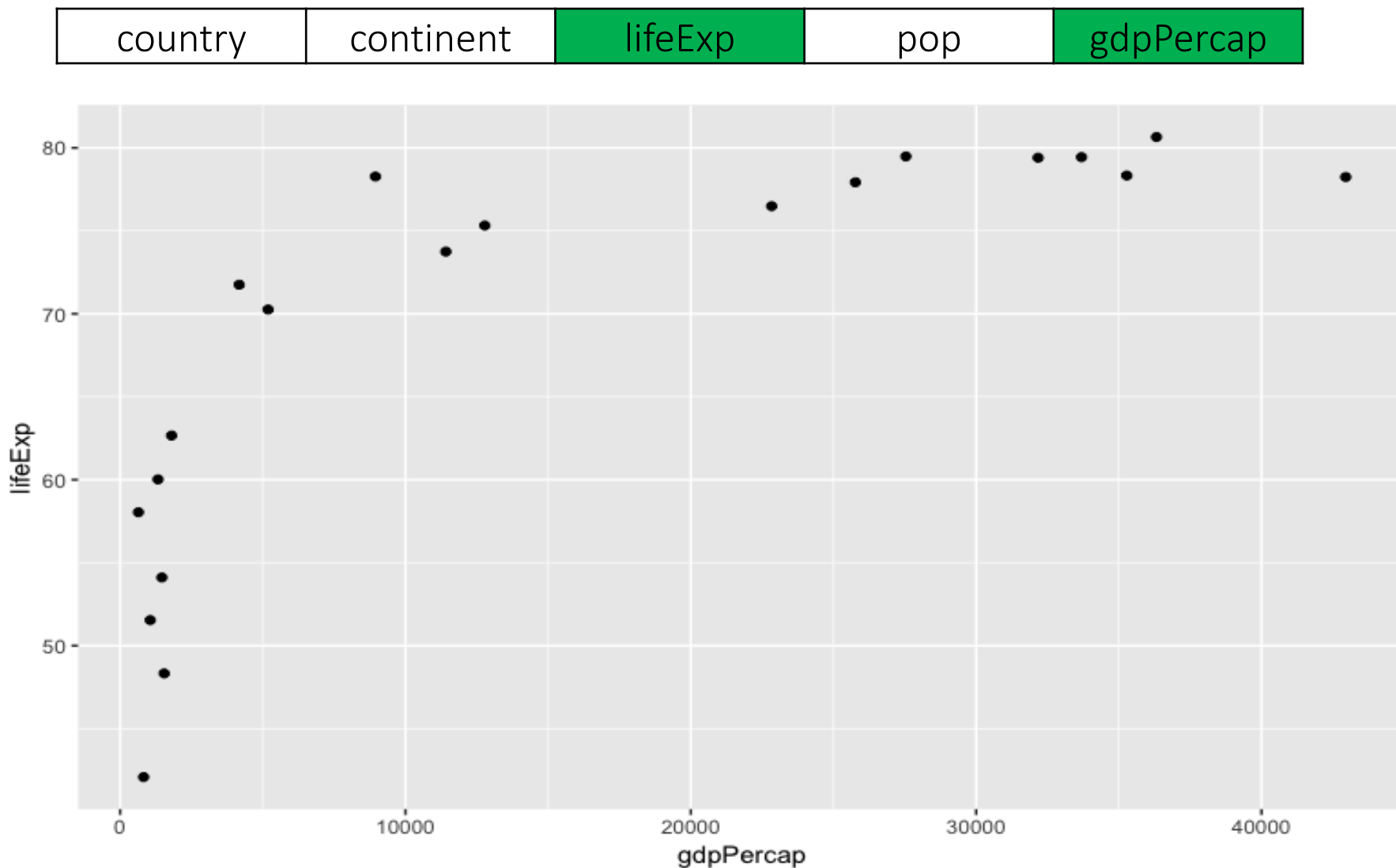
問題解決に必須なスキル



データの可視化

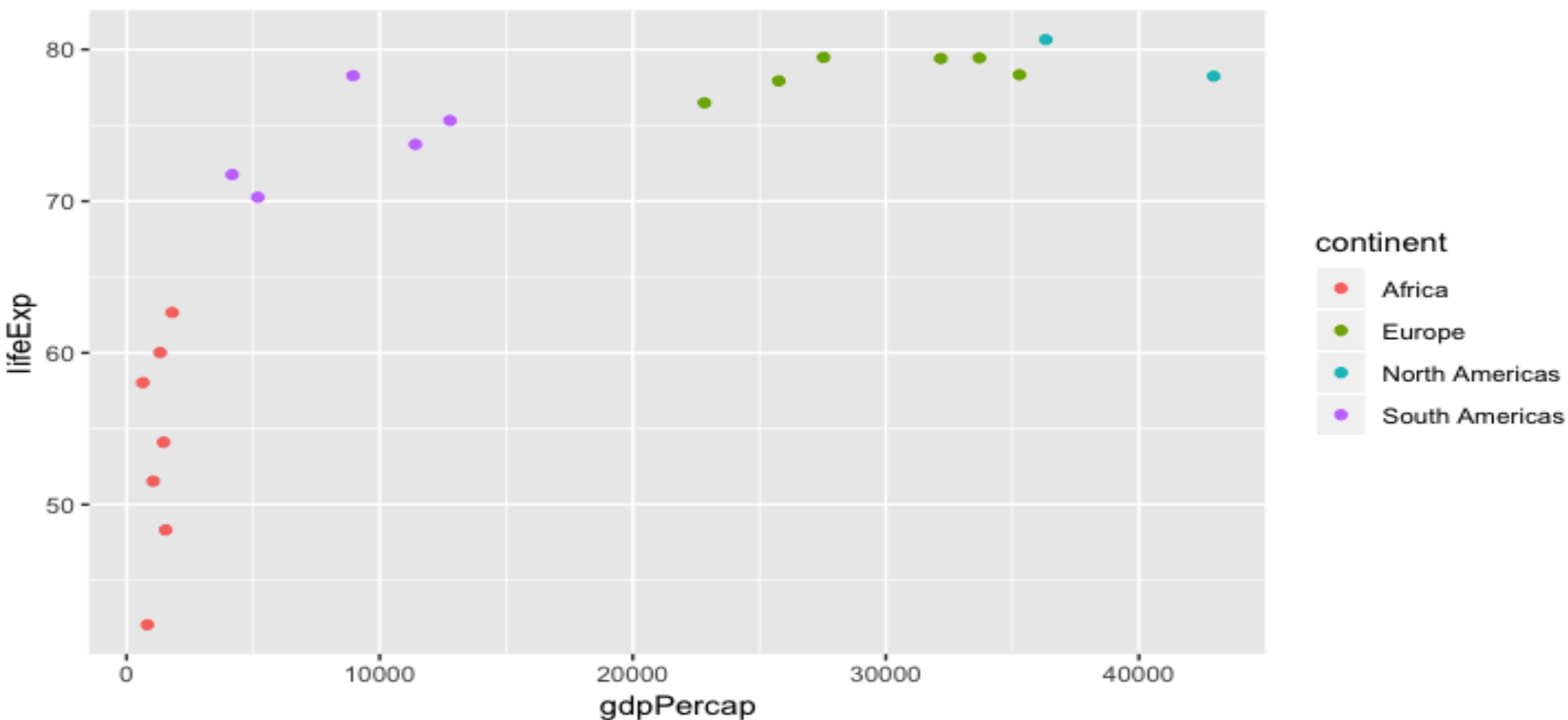
country	continent	lifeExp	pop	gdpPercap
Argentina	South Americas	75.32	40301927	12779.3796
Canada	North Americas	80.653	33390141	36319.235
Cote d'Ivoire	Africa	48.328	18013409	1544.75011
-----	-----	-----	-----	-----
Mauritania	Africa	62.664	3270065	1803.1515
Belgium	Europe	79.441	10392226	33692.6051

データの可視化



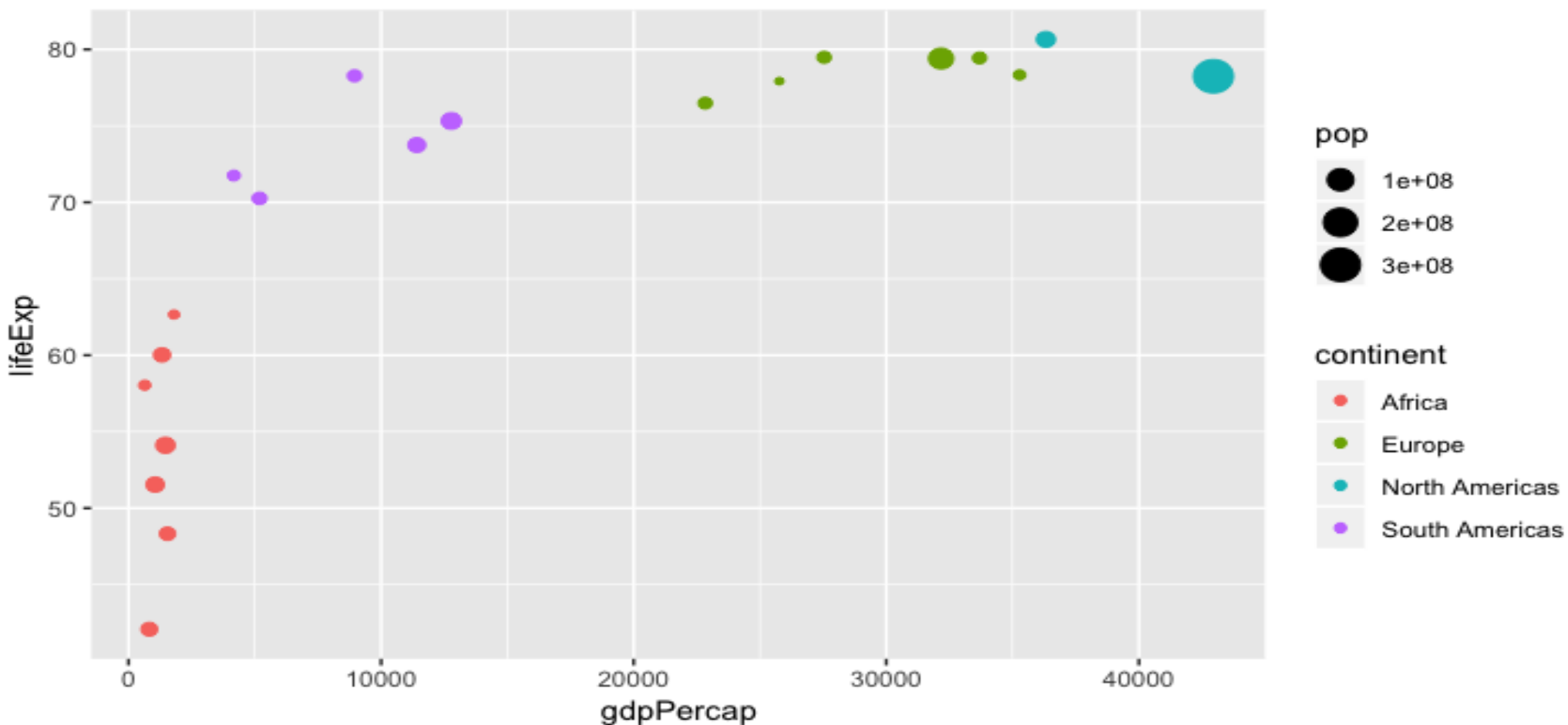
データの可視化

country	continent	lifeExp	pop	gdpPercap
---------	-----------	---------	-----	-----------

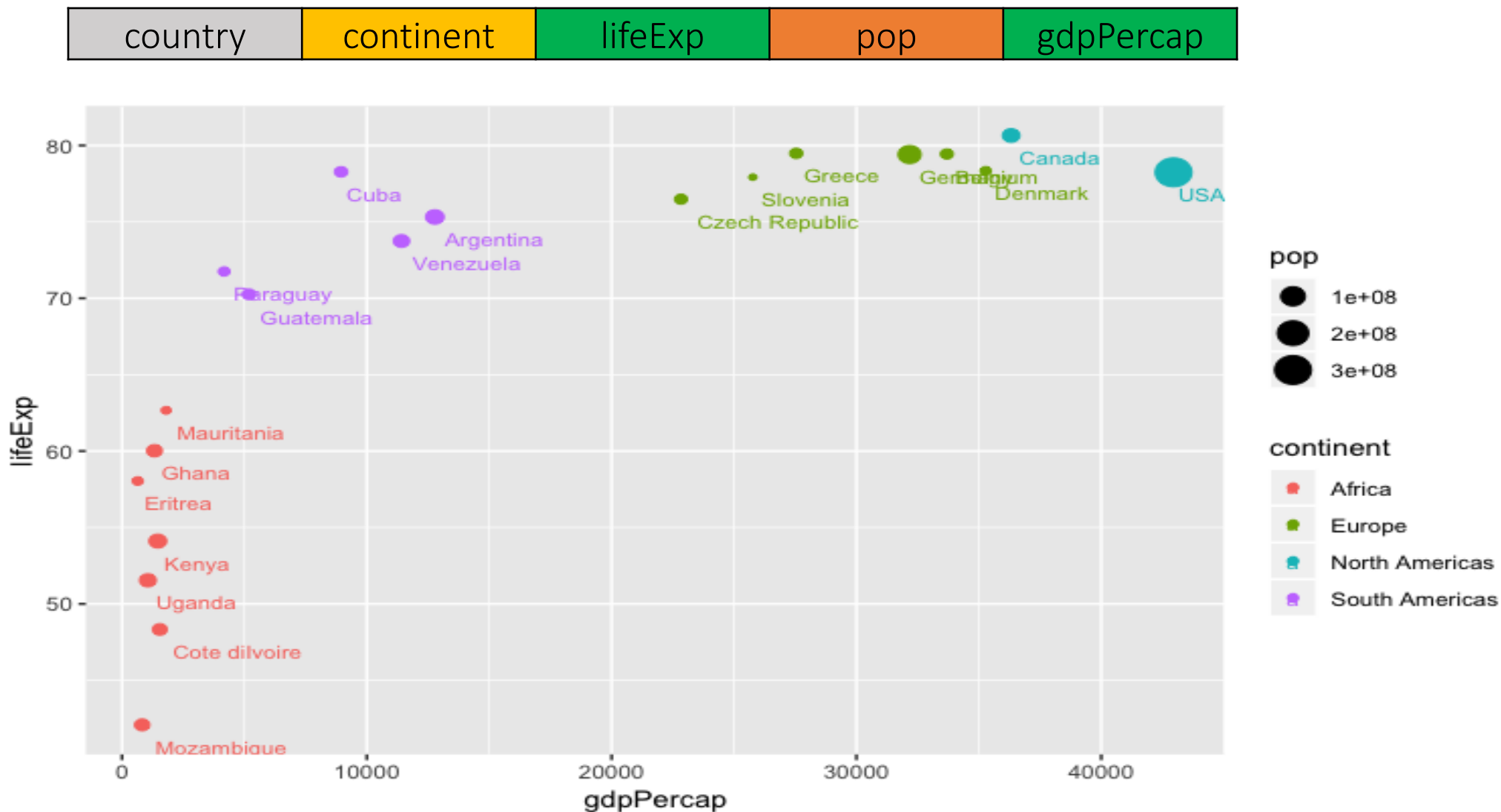


データの可視化

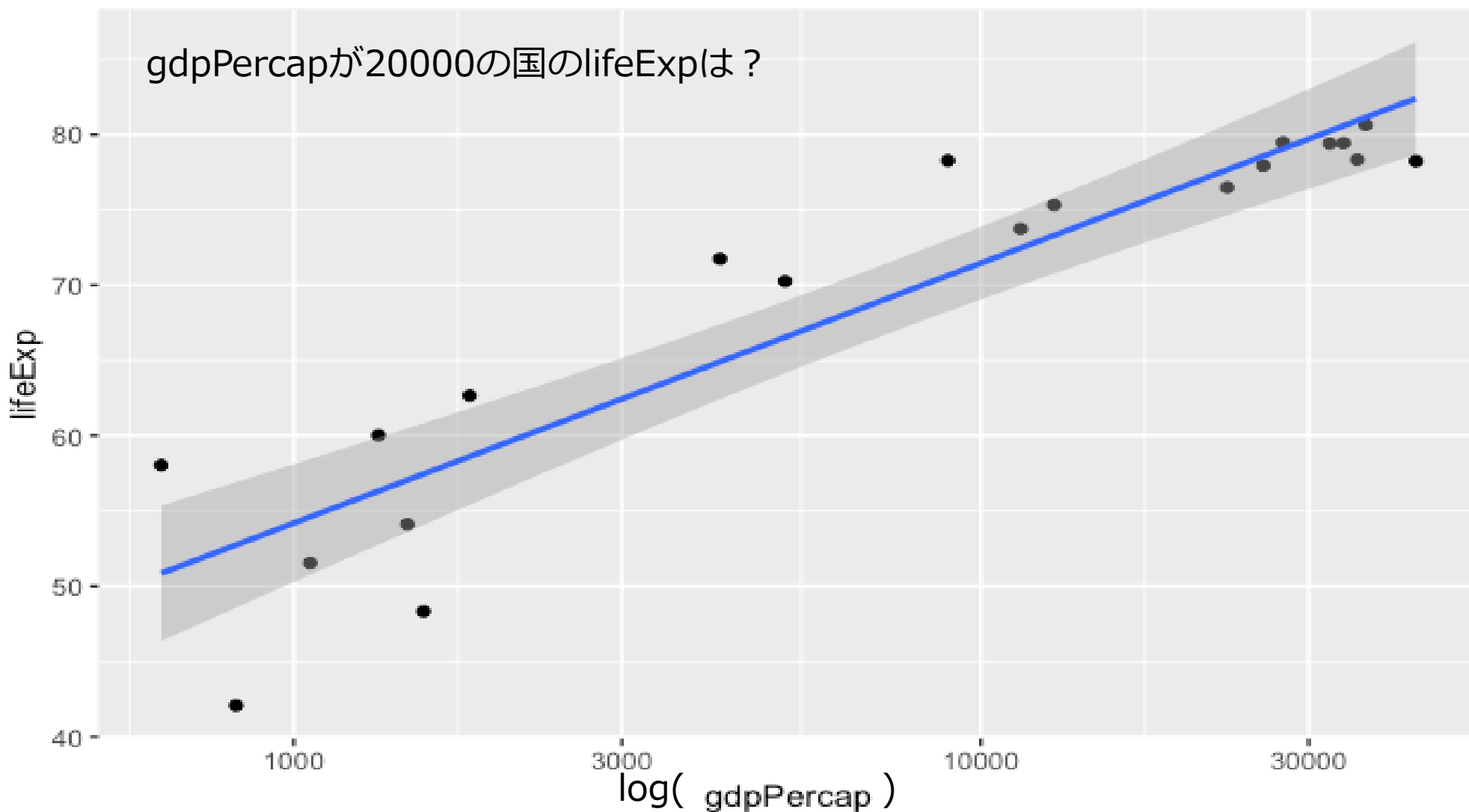
country	continent	lifeExp	pop	gdpPercap
---------	-----------	---------	-----	-----------



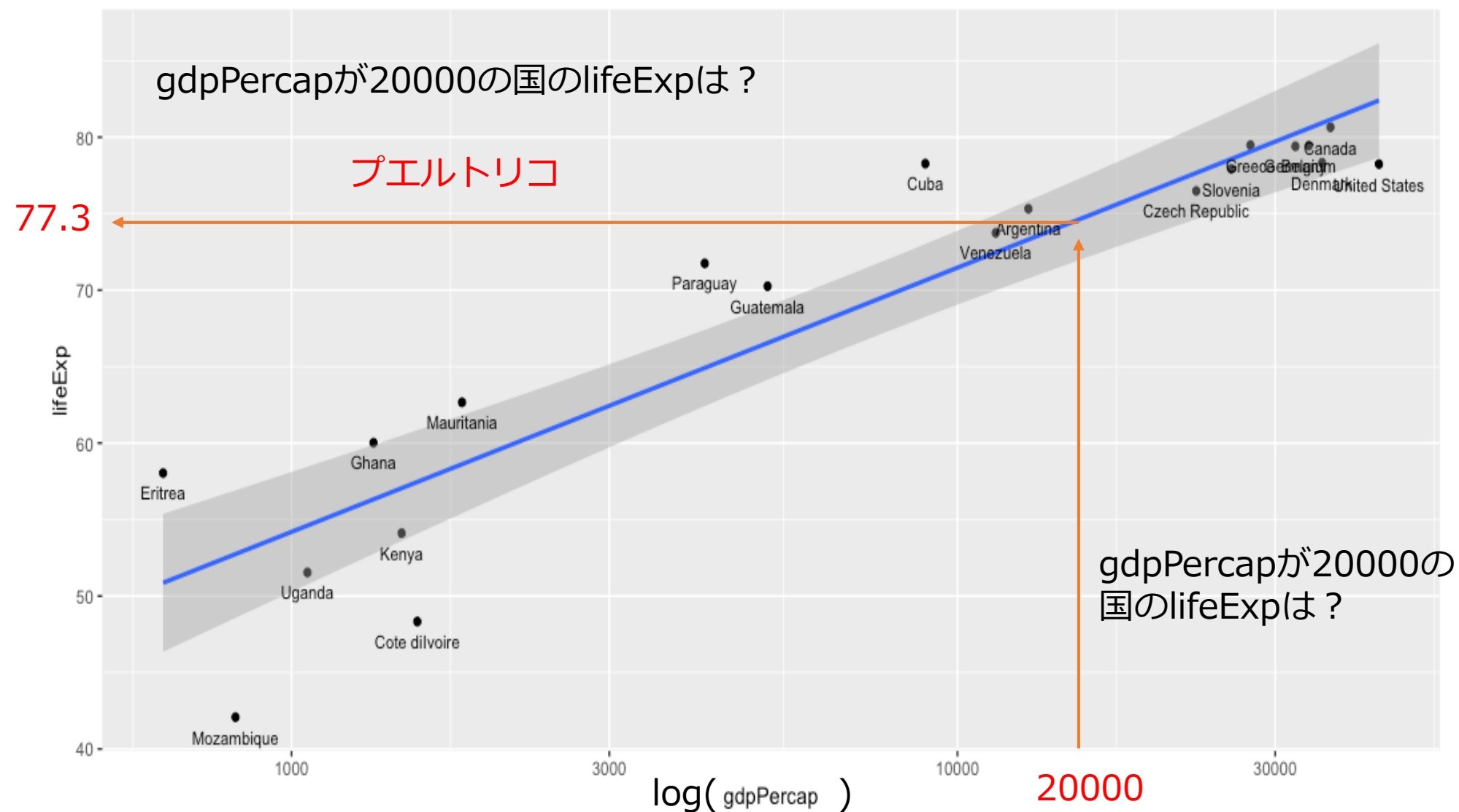
データの可視化



回帰モデル



予測モデルの設計



検定

オバマ大統領が簡単なテストで、6000万ドルもの収益を上げた方法



検定の応用

Join ABCSPORTS

Username:

Email:

Password:

☐ I accept the Terms and Conditions

Sign up +

Type A

Join ABCSPORTS

Username:

Email:

Password:

☐ I accept the Terms and Conditions

100% privacy. We will never spam you !

Sign up +

Type B

検定の応用

Type A

6/1	6/2	6/3	6/4	----	6/29	6/30
250	333	560	521	----	390	430

**1 日平均
445**

Type B (100% privacy. We will never spam you !)

6/1	6/2	6/3	6/4	----	6/29	6/30
159	253	462	412	-----	350	320

**1 日平均
405**

タイプAとBのサインアップ数の間に違いがあるのか？

検定の応用

Join ABCSPORTS

Username:

Email:

Password:

☐ I accept the Terms and Conditions

Sign up +

Join ABCSPORTS

Username:

Email:

Password:

☐ I accept the Terms and Conditions

100% privacy. We will never spam you !

Sign up +

Type B

Type A



18 % less signups

検定の応用

Join ABCESPORTS

Username:

Email:

Password:

☐ I accept the Terms and Conditions

Sign up +

Join ABCSPORTS

Username:

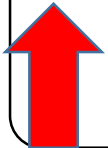
Email:

Password:

☐ I accept the Terms and Conditions

We guarantee 100% privacy. Your information will not be shared.

Sign up +



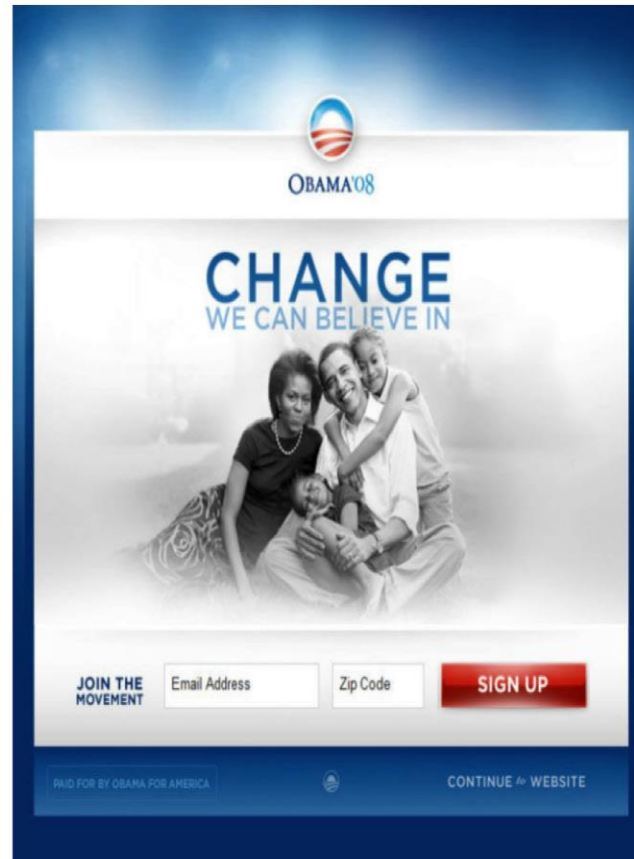
16 % more signups

ABテスト

パターン1（画像、オリジナル）：「Obama」の旗に囲まれる柔らかな

パターン2（画像）：家族と一緒に写っている写真

パターン3（画像）：正面からアップで撮影した凛々しい表情の写真



ABテスト

パターン1（オリジナル）：SIGN UP「会員登録」

パターン2：SIGN UP NOW「今すぐ会員登録」

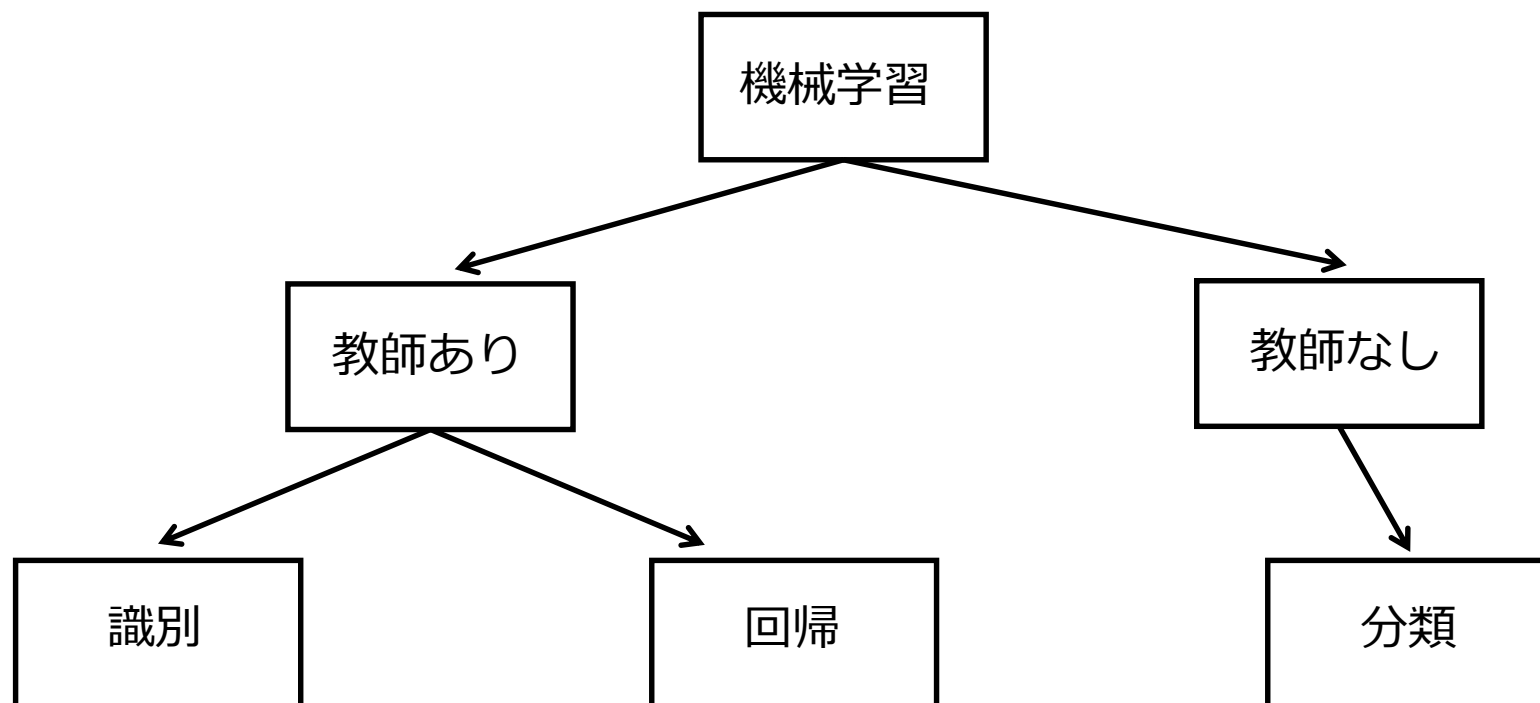
パターン3：JOIN US NOW「今すぐ参加する」

パターン4：LEARN MORE「もっと詳しく」

A red rounded rectangular button with a black border and a subtle drop shadow, containing the text "SIGN UP" in white, bold, uppercase letters.A red rounded rectangular button with a black border and a subtle drop shadow, containing the text "SIGN UP NOW" in white, bold, uppercase letters.A red rounded rectangular button with a black border and a subtle drop shadow, containing the text "JOIN US NOW" in white, bold, uppercase letters.A red rounded rectangular button with a black border and a subtle drop shadow, containing the text "LEARN MORE" in white, bold, uppercase letters.

機械学習

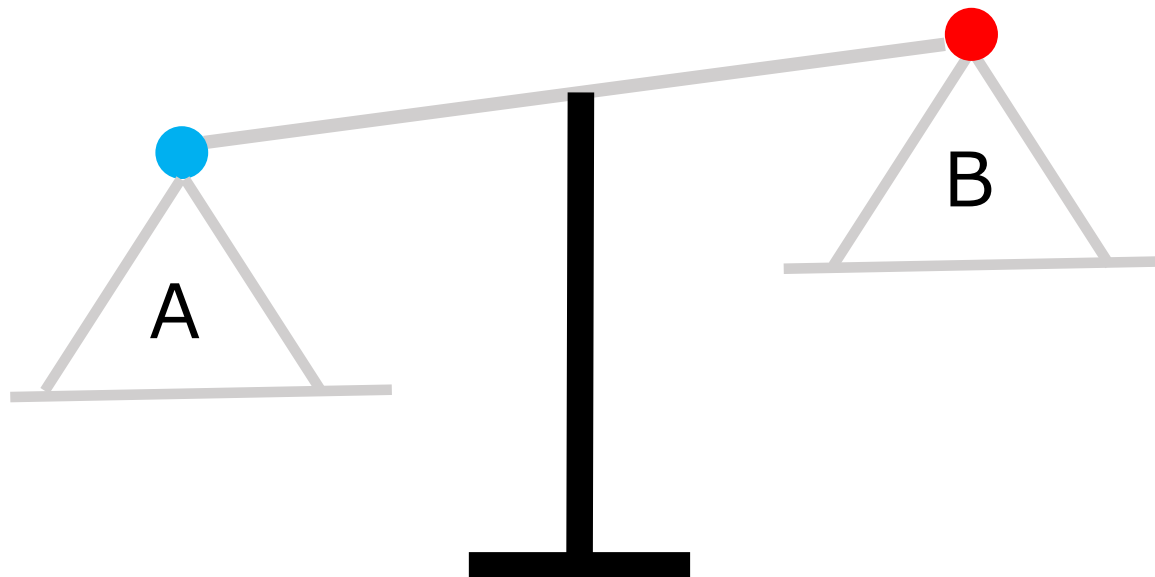
機械学習の区分



機械学習で問われる 3つの質問

- 質問 1
「AかBか？」

識別アルゴリズム

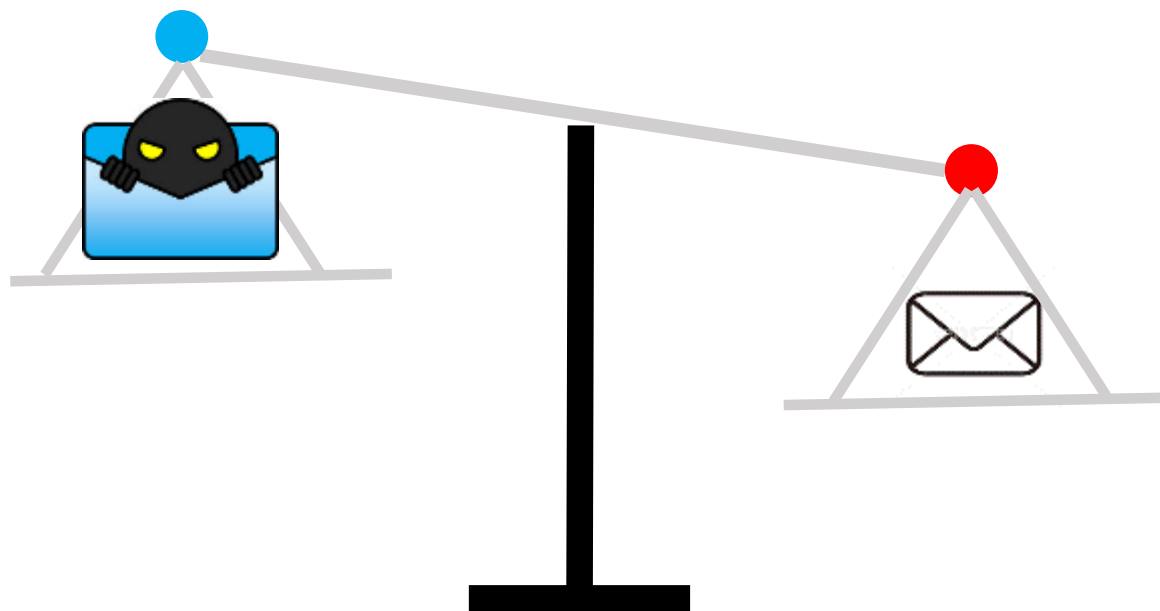


機械学習で問われる 3つの質問

- 質問 1
「AかBか？」

識別アルゴリズム

ナイーブベイズ

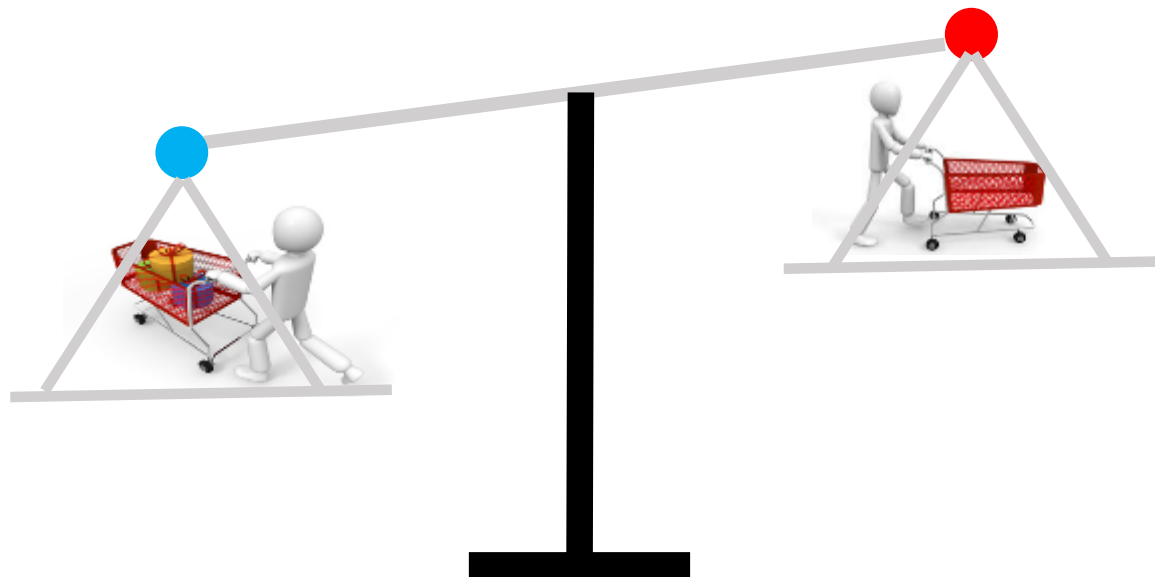


機械学習で問われる 3 つの質問

- 質問 1
「AかBか？」

識別アルゴリズム

決定木



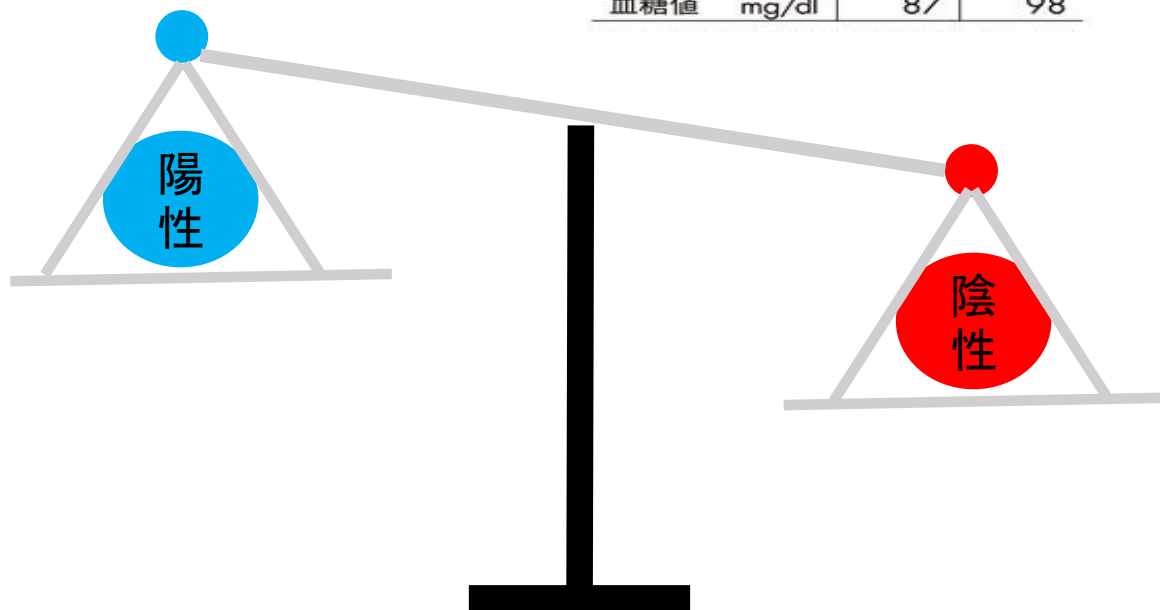
機械学習で問われる3つの質問

- 質問 1
「AかBか？」

識別アルゴリズム

ロジスティック回帰分析

		09年	→12年
身長	cm	170.5	170.5
体重	kg	68.9	65.5
腹囲	cm	92.5	83.5
中性脂肪	mg/dl	296	226
LDL	mg/dl	159	165
HDL	mg/dl	43	43
γ-GTP	IU/l	41	28
血糖値	mg/dl	87	98



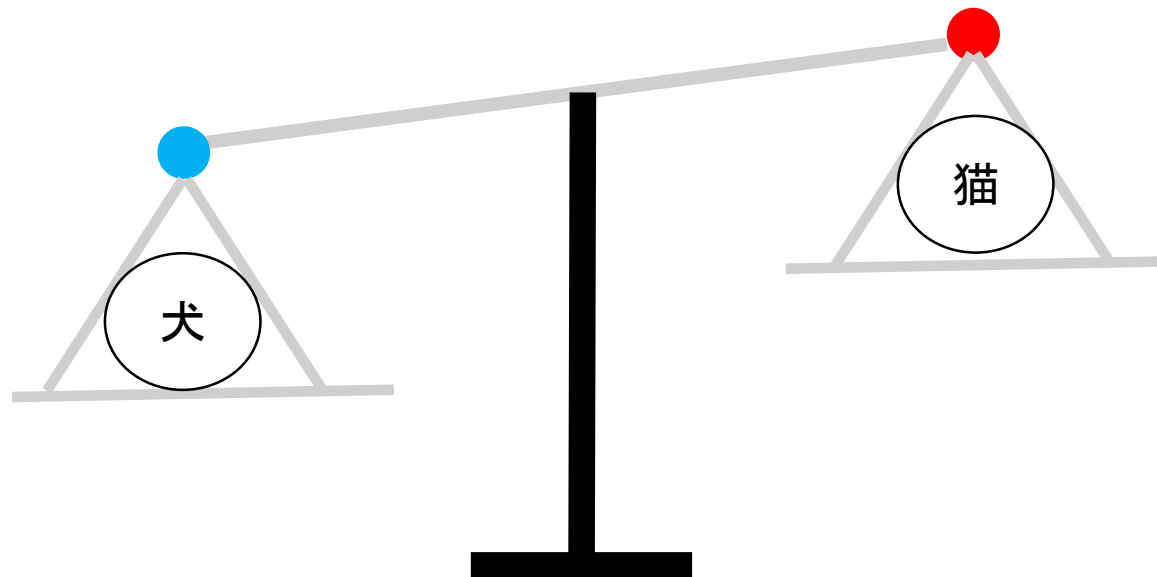
機械学習で問われる 3 つの質問

- 質問 1
「AかBか？」



識別アルゴリズム

Deep learning



機械学習で問われる 3つの質問

- ・ 質問 2

「どのくらいの量または数か？」

回帰アルゴリズム

次の火曜日の気温は何度か？

月曜日



32度

火曜日

何度？

機械学習で問われる 3つの質問

・質問 2

「どのくらいの量または数か？」

回帰アルゴリズム



800万円

2億5千万円

新しい物件価格を予測

過去のデータ



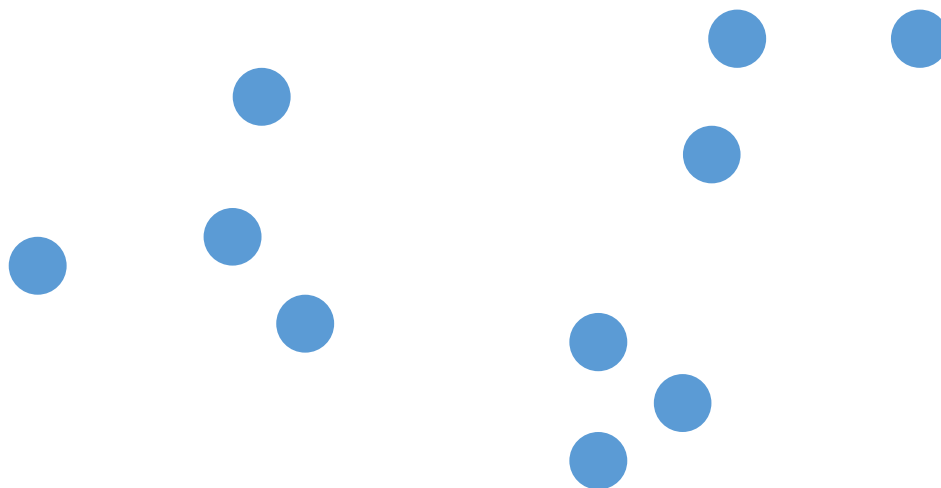
機械学習で問われる 3つの質問

- 質問 3

「どのような編成になっているのか？」

分類アルゴリズム

どの視聴者が同じ種類の
映画を好むか？



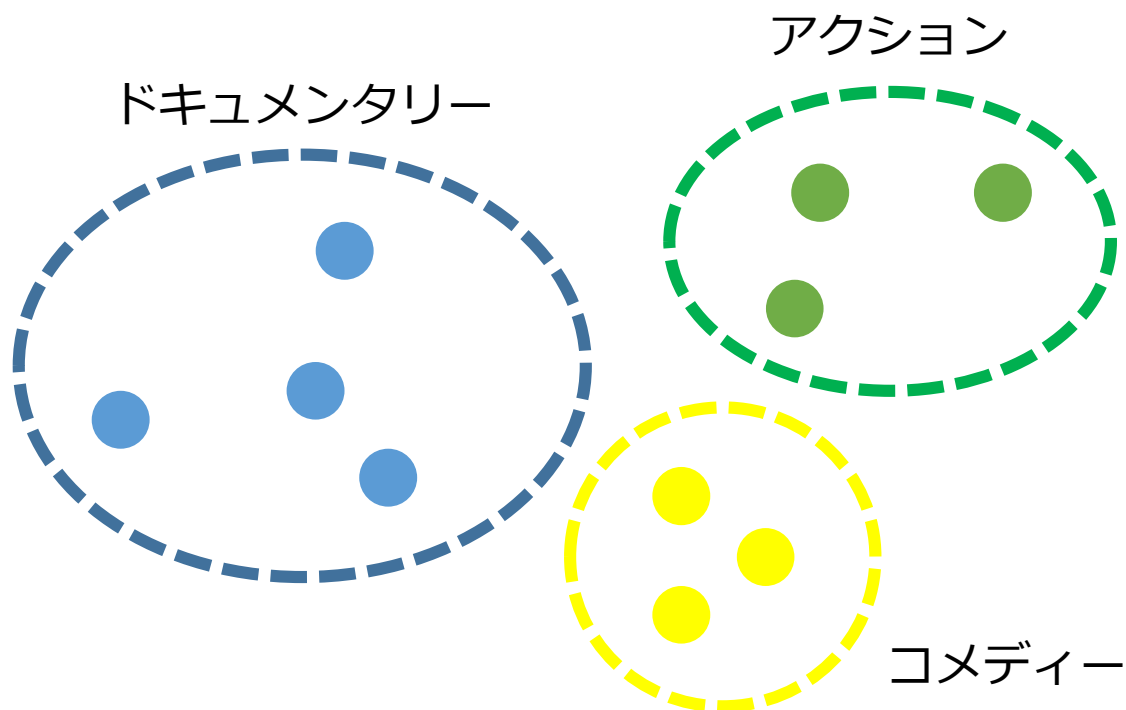
機械学習で問われる3つの質問

- 質問3

「どのような編成になっているのか？」

分類アルゴリズム

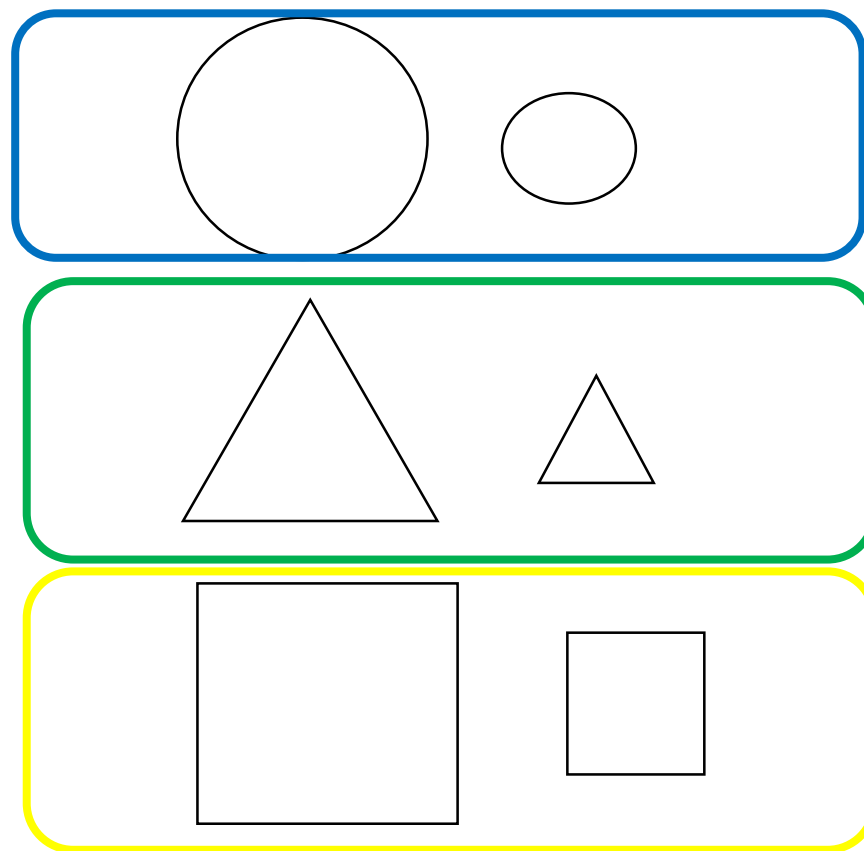
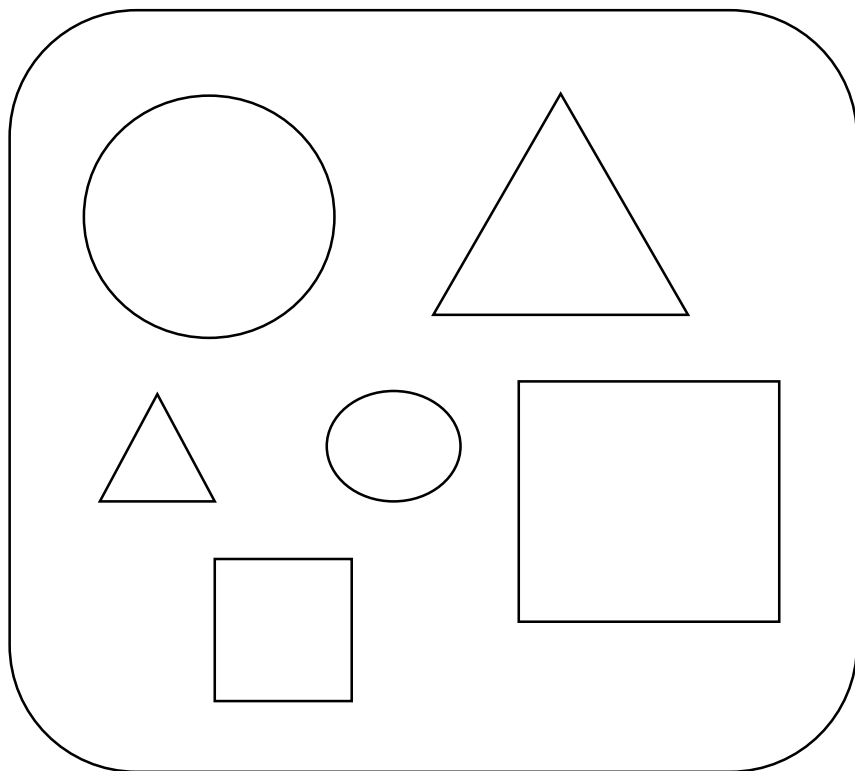
どの視聴者が同じ種類の映画を好むか？



機械学習で問われる 3つの質問

- 質問 3

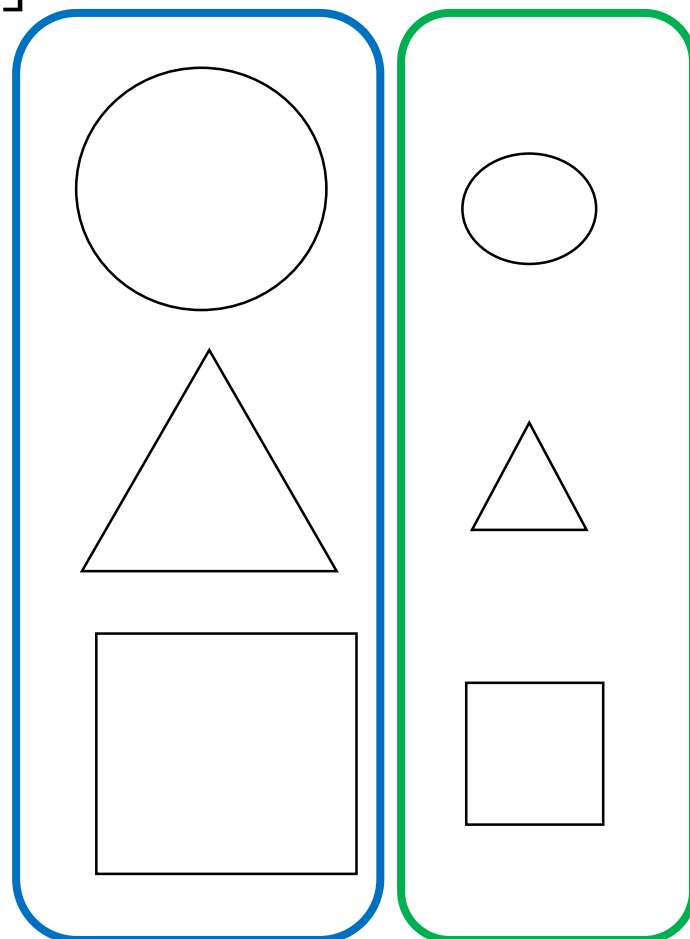
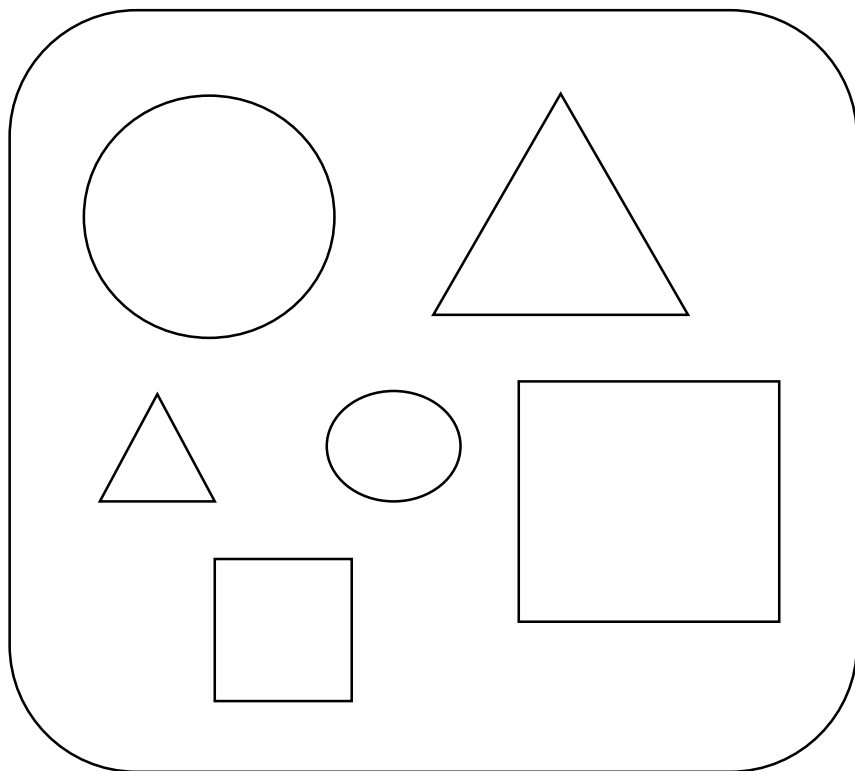
「どのような編成になっているのか？」



機械学習で問われる 3つの質問

- 質問 3

「どのような編成になっているのか？」



機械学習の区分

機械学習

教師あり

識別

AかBか

決定木



ナイーブベイズ



ニューラル
ネットワーク



SVM



ロジスティック回帰



教師あり

回帰

どのくらいの量か

回帰分析



教師なし

分類

どう分けるか

k-means法



主成分分析



・ 教師あり ・ 機械学習 ・ 回帰

機械学習の区分

機械学習

識別

AかBか

決定木



ナイーブベイズ



ニューラル
ネットワーク



SVM



ロジスティック回帰



回帰

どのくらいの量か

重回帰分析



分類

どう分けるか

k-means法



主成分分析



回帰分析

- ・ 予測したい変数を様々な要因から予測する方法

⇔ 住宅の価格を築年数・坪数から予測する方法



800万円

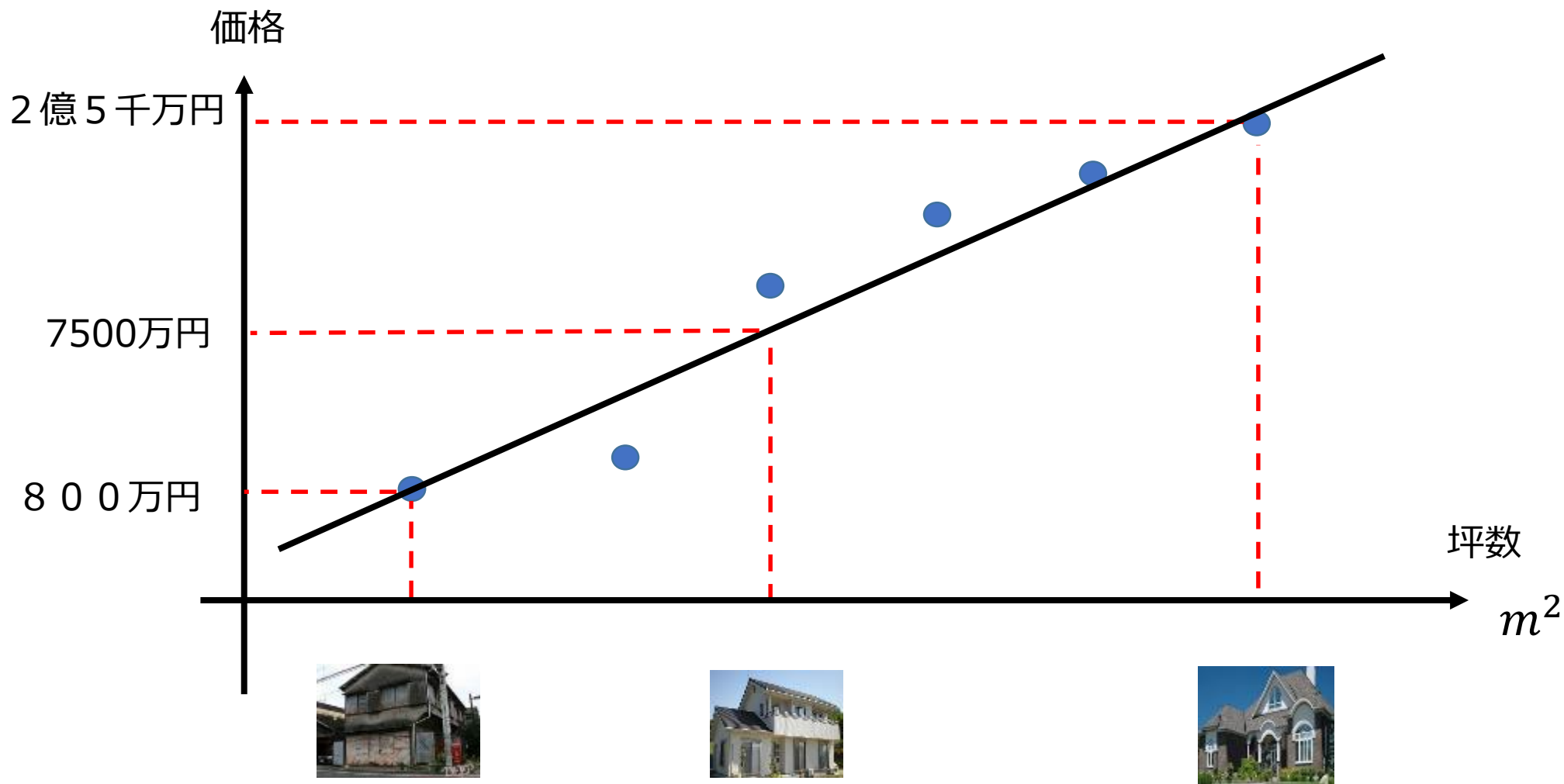


価格を予測する

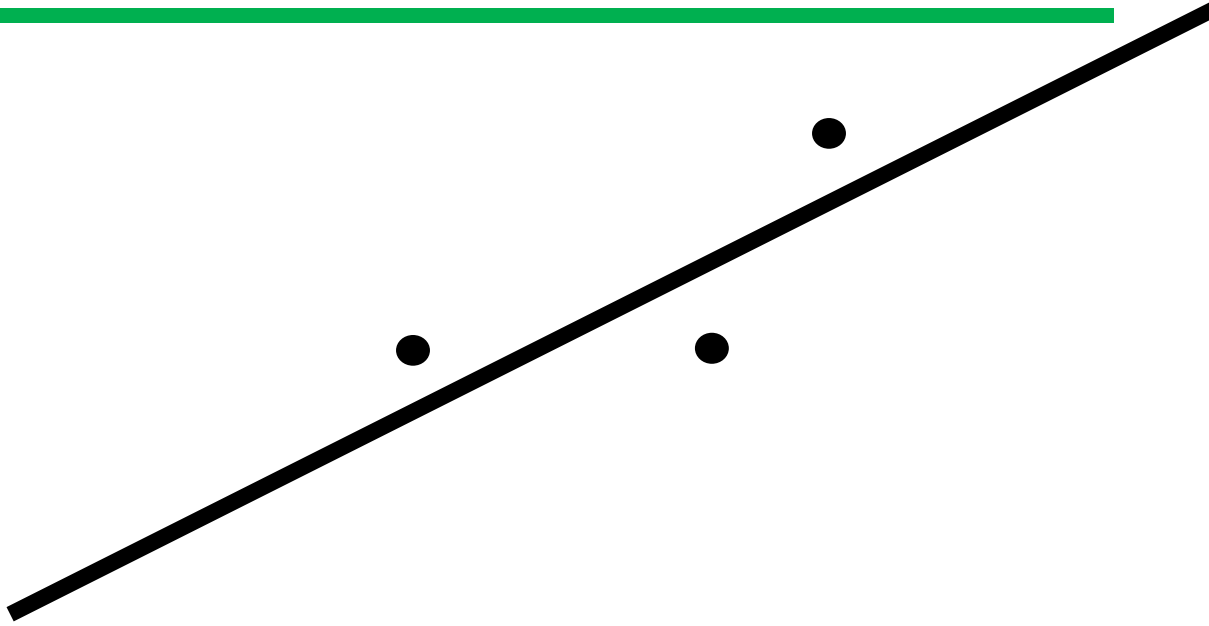


2億5千万円

回帰分析

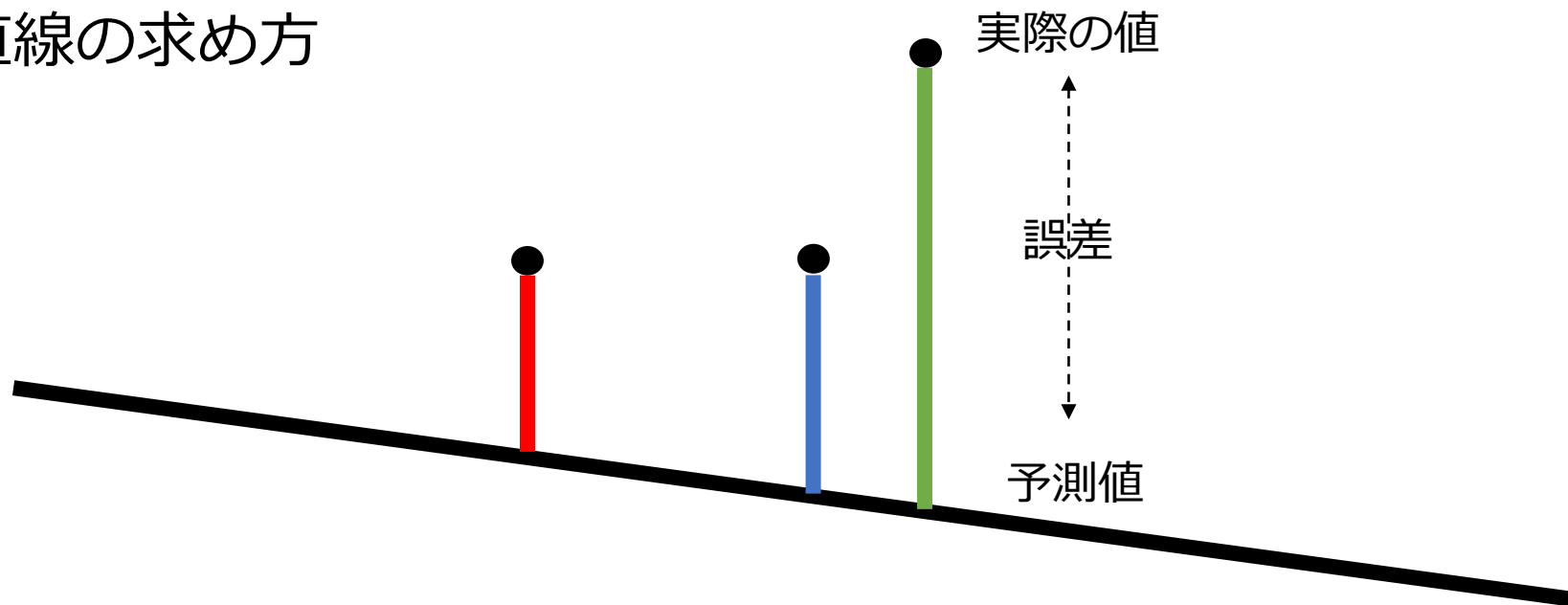


回帰モデル



回帰モデル

直線の求め方

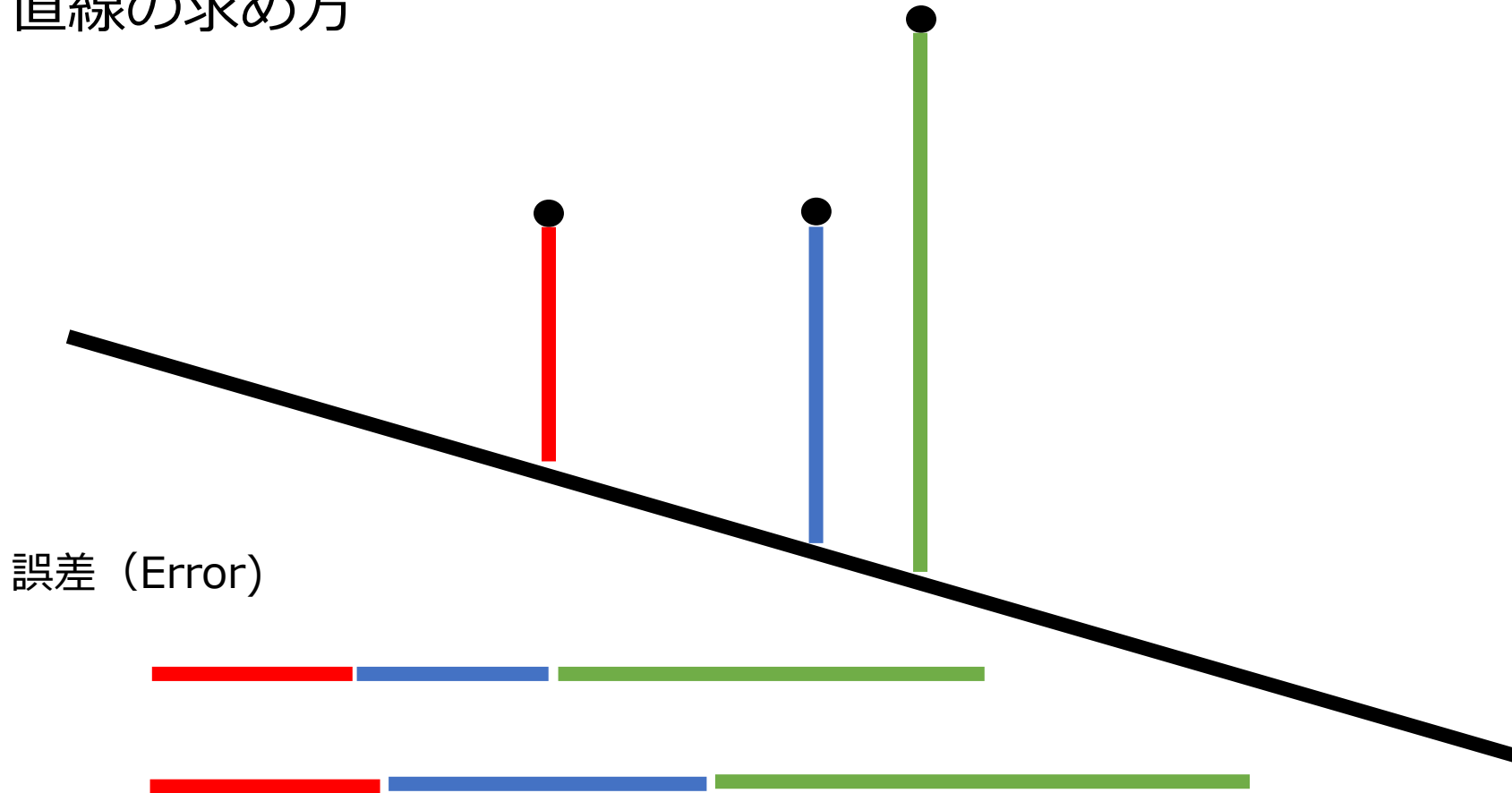


誤差 (Error)



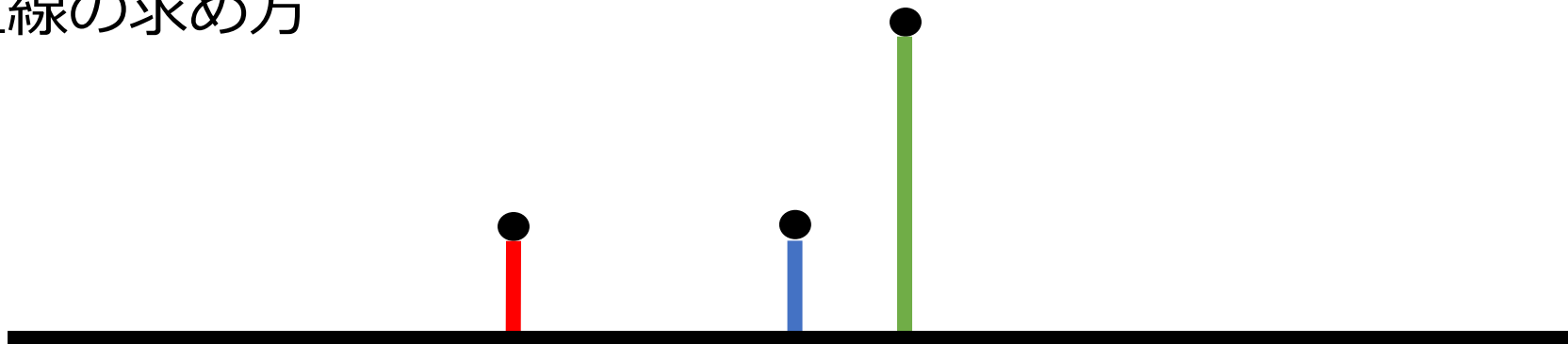
回帰モデル

直線の求め方

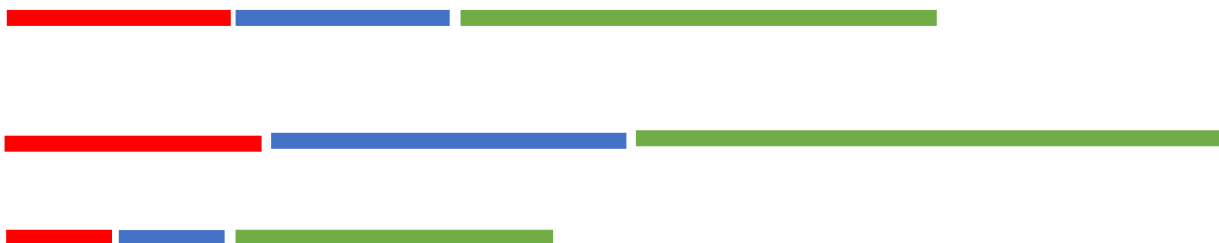


回帰モデル

直線の求め方

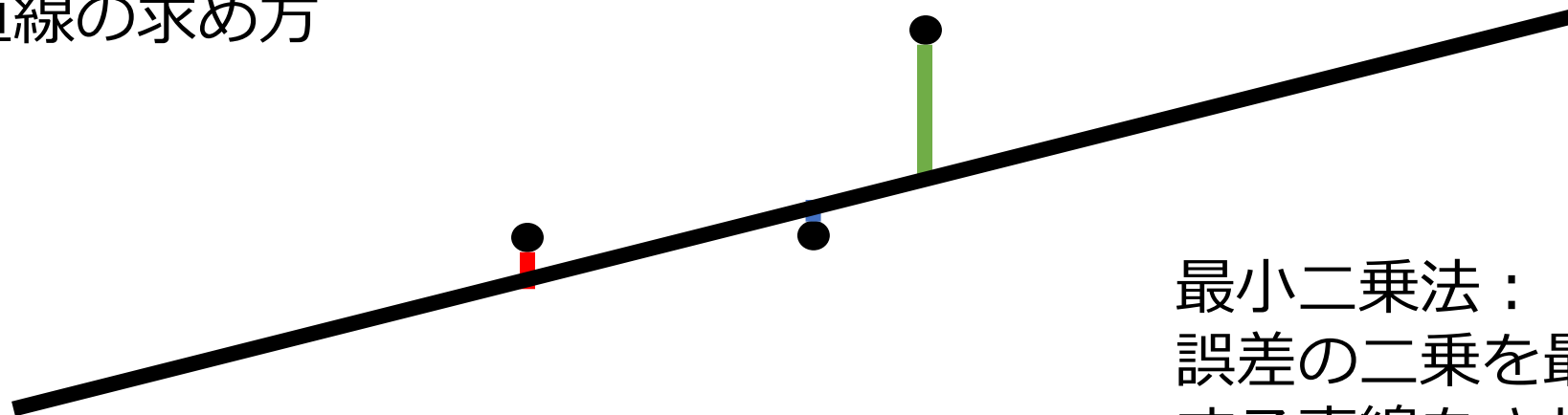


誤差 (Error)



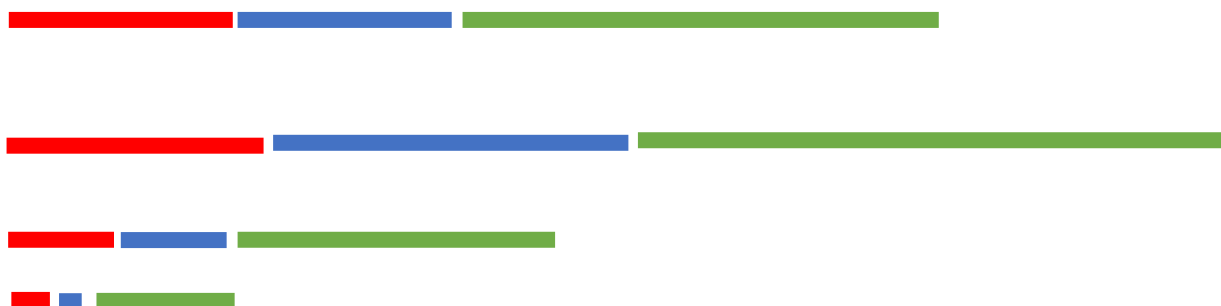
回帰モデル

直線の求め方

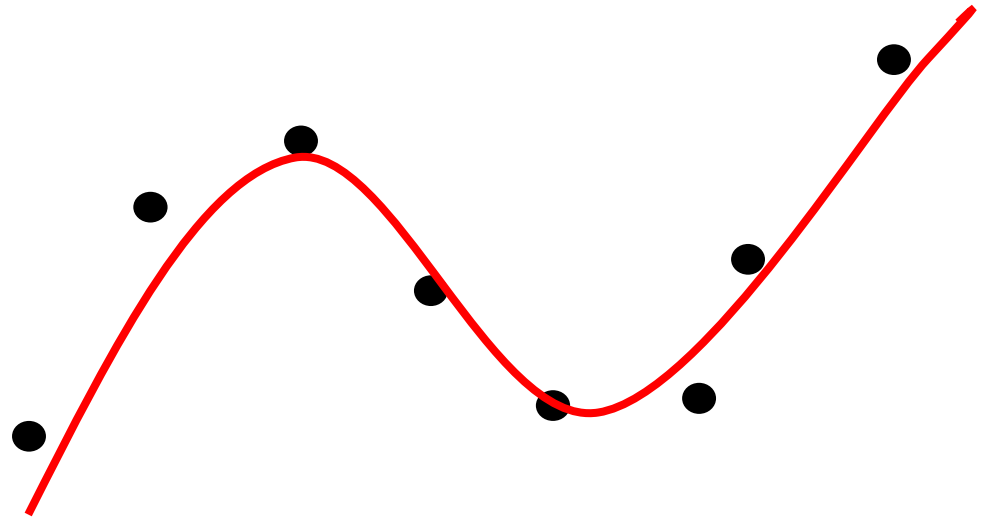
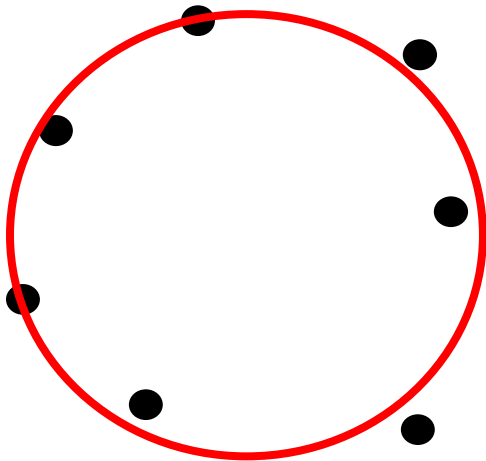


最小二乗法：
誤差の二乗を最小にする直線をさがす

誤差 (Error)



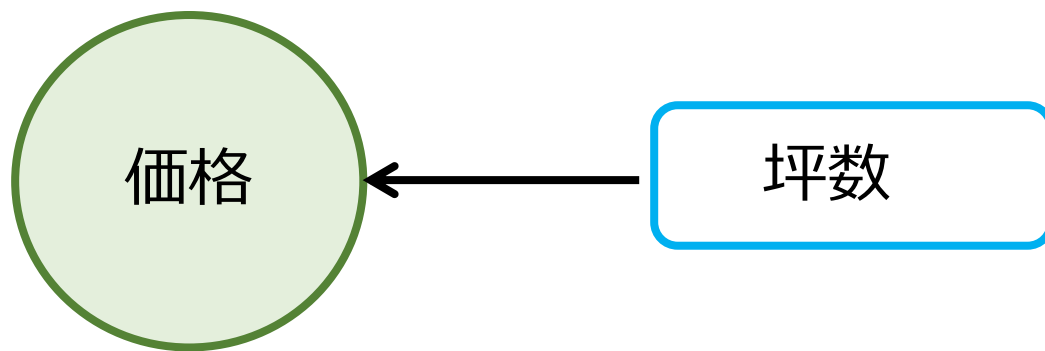
様々な回帰モデル



https://kwichmann.github.io/ml_sandbox/linear_regression_diagnostics/

単回帰モデル

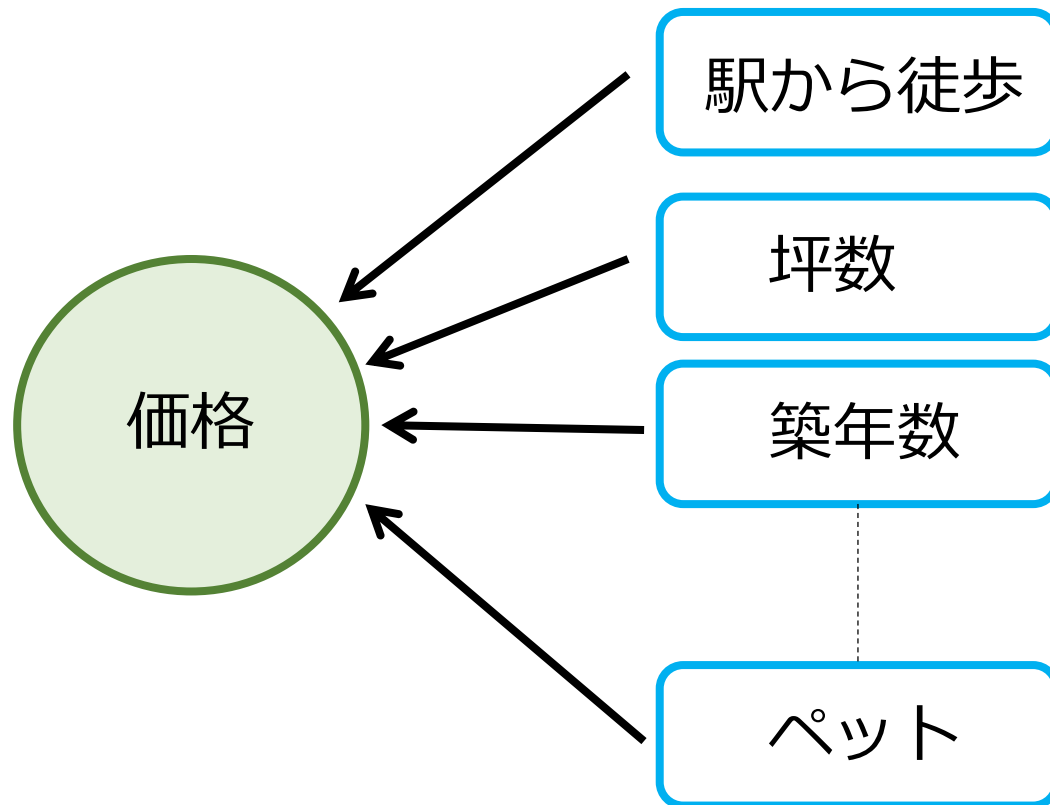
予測したい変数を要因となる変数の関係で予測・説明する手法



$$y = b_0 + b_1 x_1$$

重回帰モデル

予測したい変数をいくつかの要因となる変数の関係で予測・説明する手法



$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

ダミー変数を使った重回帰分析

家賃	徒歩	専有面積	築年数	階数	新築	2階以上	南向き	オート ロック	エアコン	バス トイレ別	追い焚き	フロー リング	ペット 相談可
88500	10	24.7	2	8	無	有	無	有	有	無	無	有	有
86700	12	13.1	3	7	無	有	有	無	有	有	無	有	有
87300	10	25.7	3	2	無	有	有	有	有	有	無	有	有

58500	16	20.7	3	8	無	有	無	有	有	有	無	有	有
126700	6	45.7	3	7	無	有	有	有	無	有	無	有	有
117000	3	32.5	3	2	無	有	無	無	無	有	無	有	有

家賃を予測する

問題「高価格の部屋の条件は？」

正解データ

評価用 データ

番号	正解の価格
1	92500
2	88000
3	88000
4	88000
5	88000

正解データ

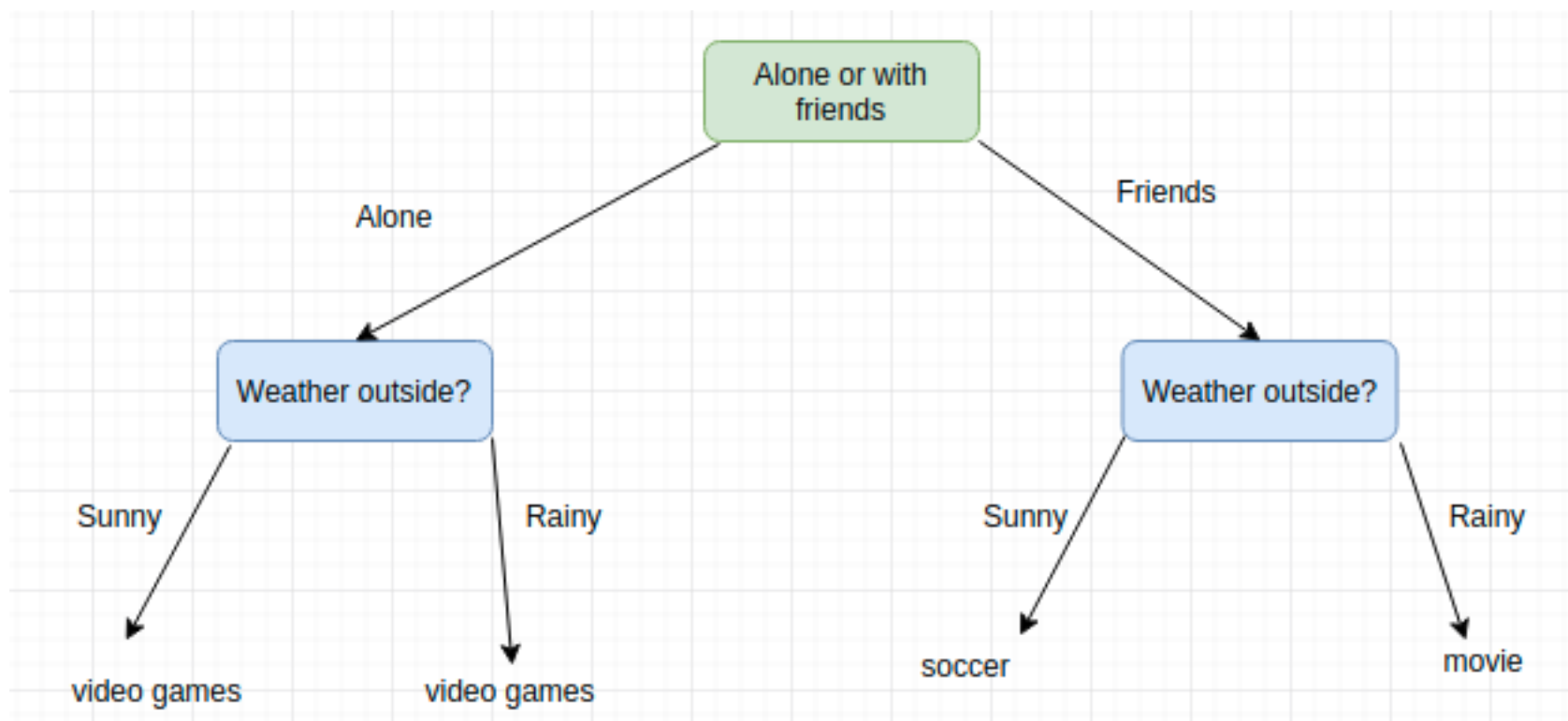
評価用 データ

番号	正解の価格
1	92500
2	88000
3	88000
4	88000
5	88000

予測価格
92025.715
89062.514
88457.078
93337.553
88465.588

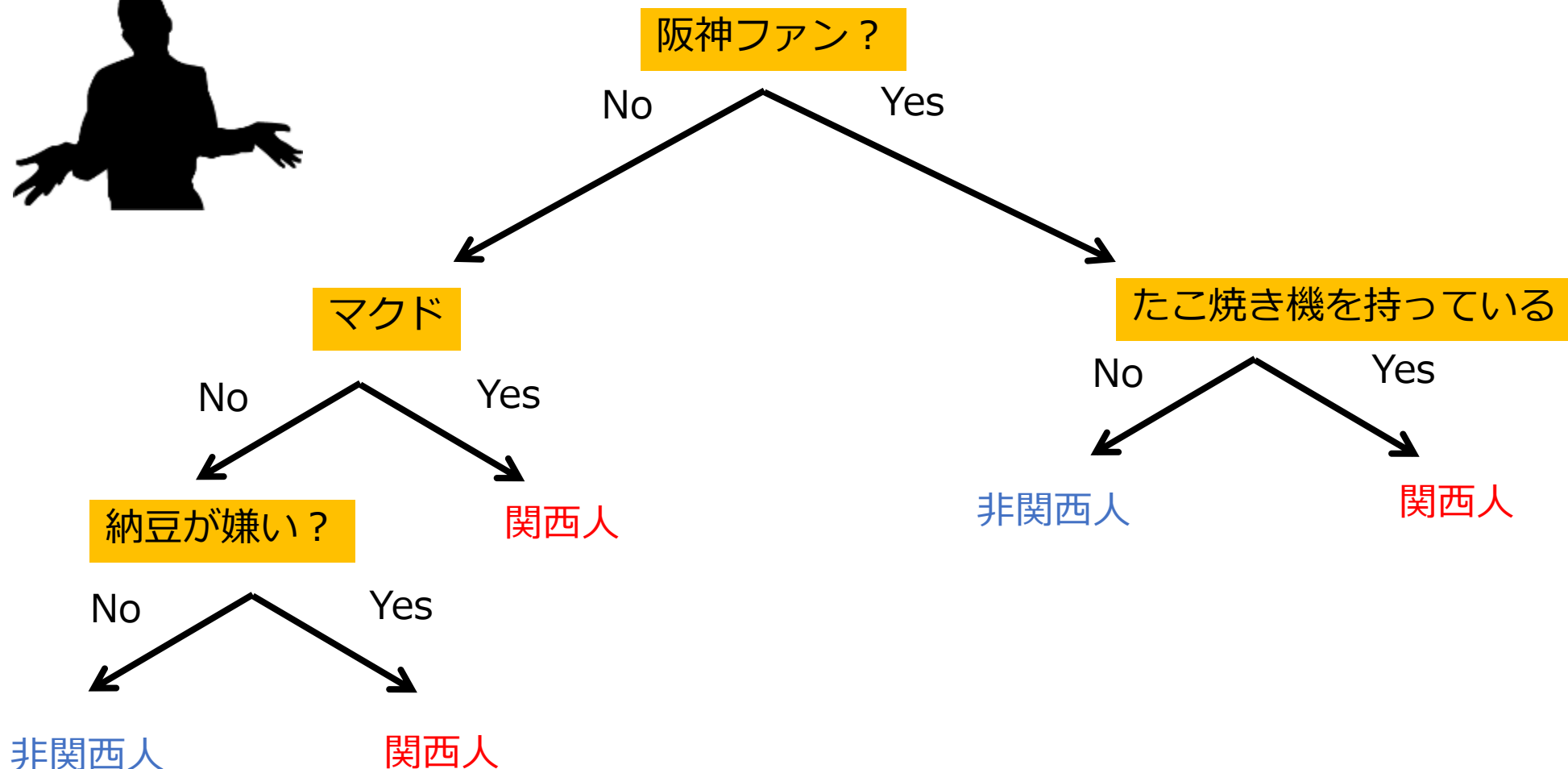
決定木

樹形図



決定木：識別能力の高い質問による分類

関西人なのか？



決定木の応用

- <http://jp.akinator.com>



問題「高価格の部屋の条件は？」

問題「大陸の識別」

country	continent	lifeExp	pop	gdpPercap
Argentina	Americas	75.32	40301927	12779.3796
Canada	Americas	80.653	33390141	36319.235
Cote d'Ivoire	Africa	48.328	18013409	1544.75011
Cuba	Americas	78.273	11416987	8948.10292
-----	-----	-----	-----	-----
Mauritania	Africa	62.664	3270065	1803.1515
Belgium	Europe	79.441	10392226	33692.6051

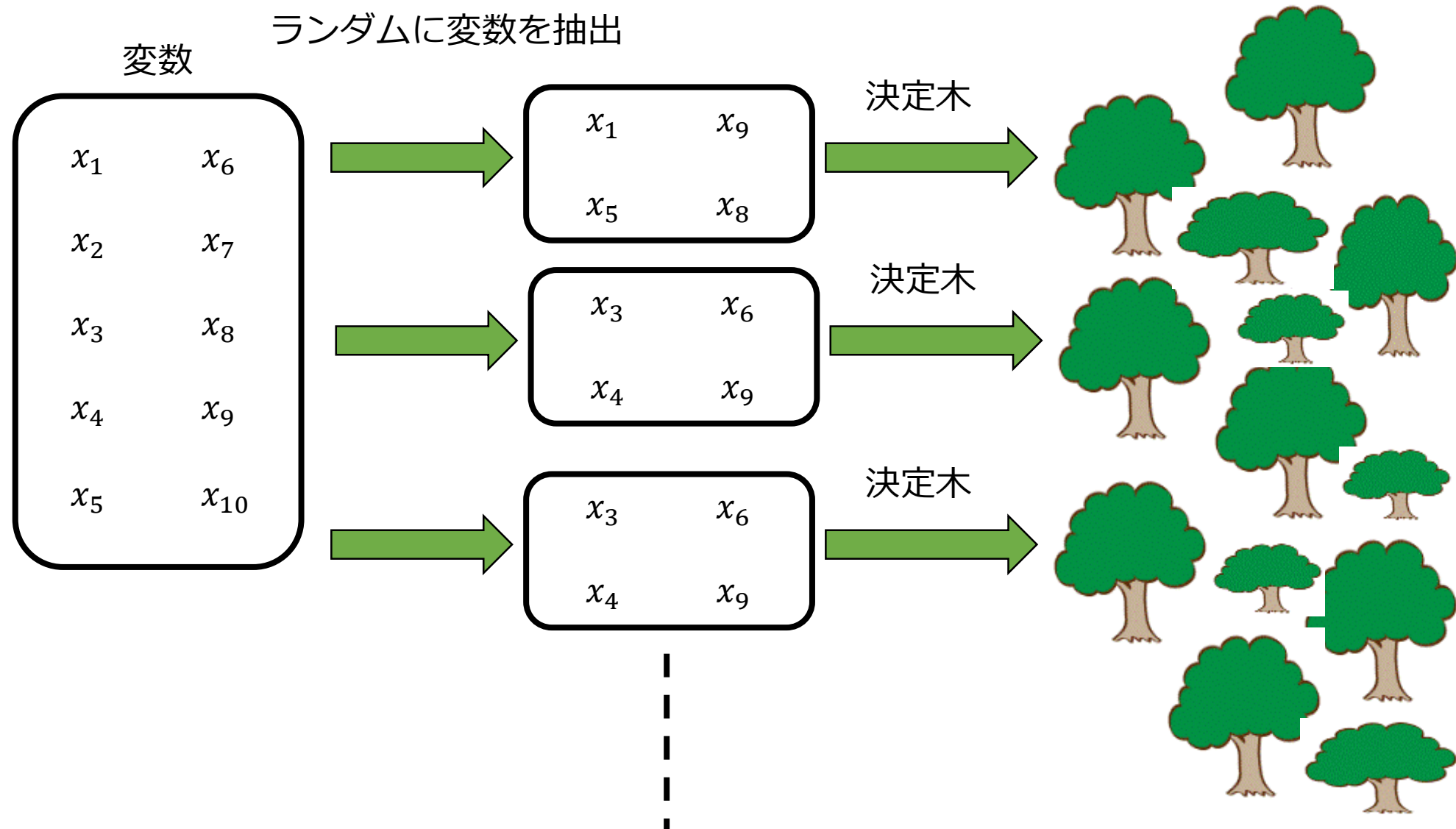
問題「各国が属する大陸を識別できるか？」

country	continent	lifeExp	pop	gdpPercap
Argentina	Americas	75.32	40301927	12779.3796
Canada	Americas	80.653	33390141	36319.235
Cote d'Ivoire	Africa	48.328	18013409	1544.75011
Cuba	Americas	78.273	11416987	8948.10292
-----	-----	-----	-----	-----
Mauritania	Africa	62.664	3270065	1803.1515
Belgium	Europe	79.441	10392226	33692.6051

ランダムフォレスト

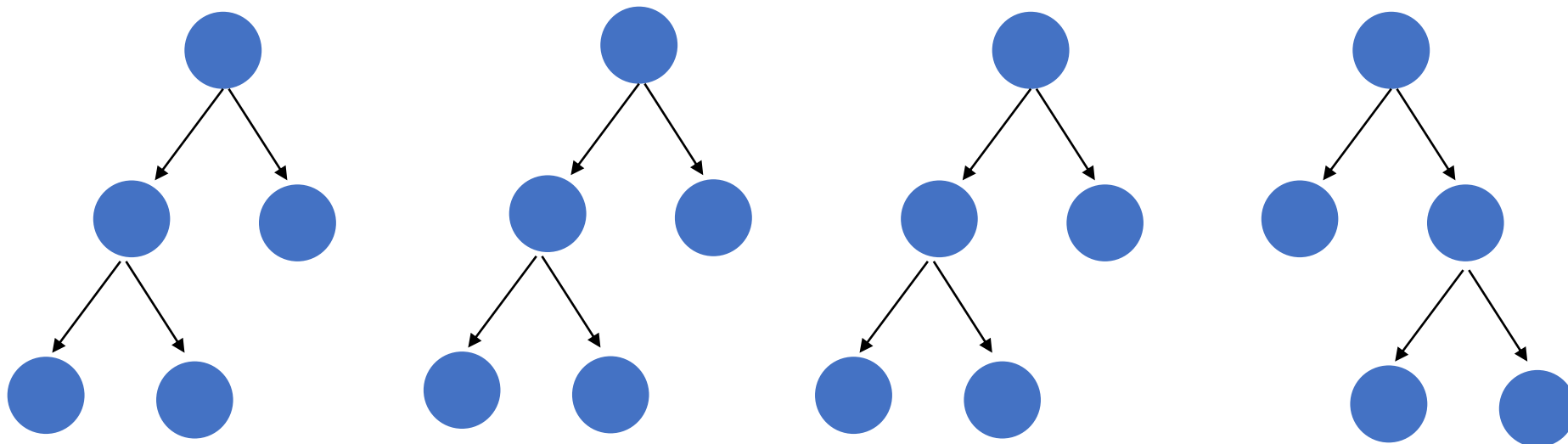
- 決定木を用いた集団学習を行うモデル
(過学習にならないように決定木を複数作って平均を取る)

ランダムフォレスト



ランダムフォレスト

大抵の決定木は正解を提供している



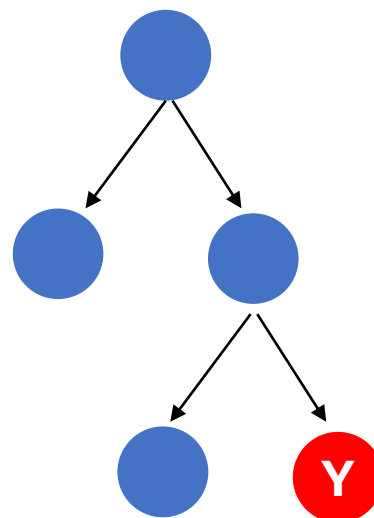
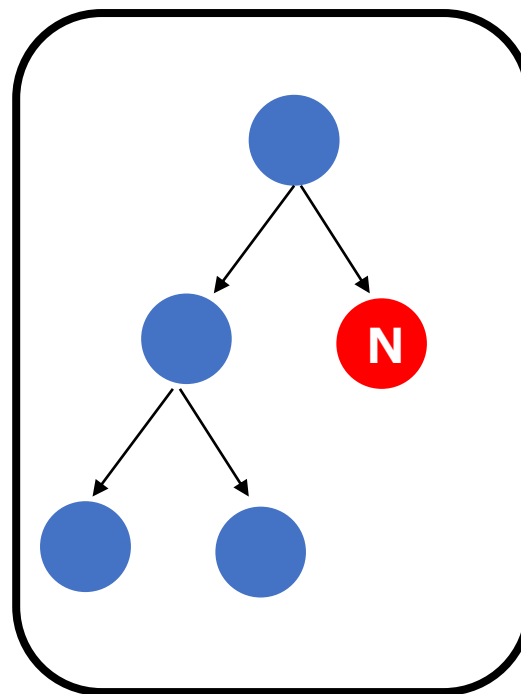
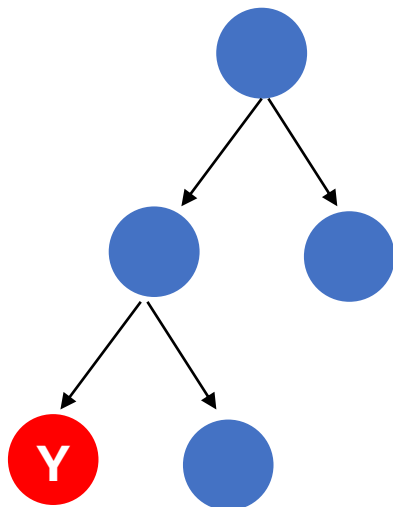
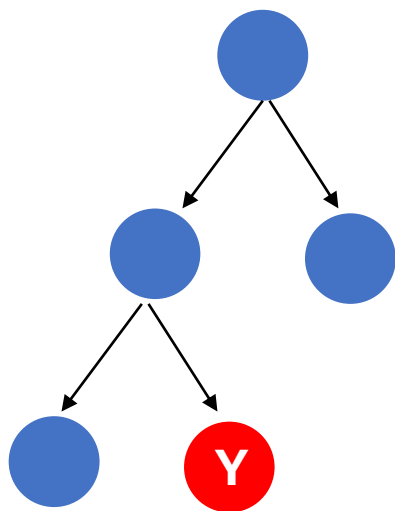
ランダムフォレスト

購入するかどうか？



「購入する」

間違い



多数決の原理

ざっくり分けるなら

機械学習

識別

AかBか

決定木



ナイーブベイズ



ニューラル
ネットワーク



SVM



ロジスティック回帰



回帰

どのくらいの量か

重回帰分析



分類

どう分けるか

k-means法



主成分分析

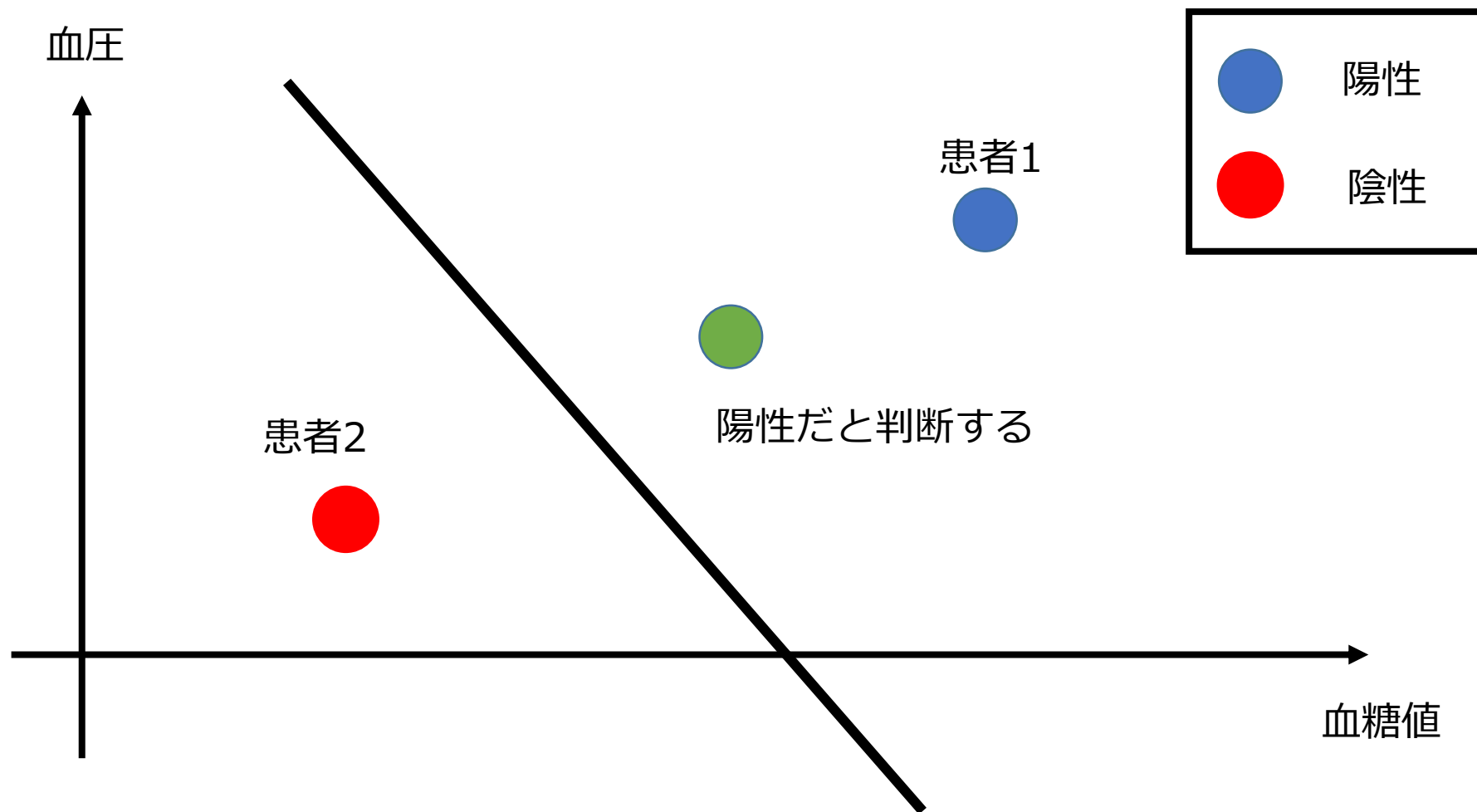


陽性・陰性？

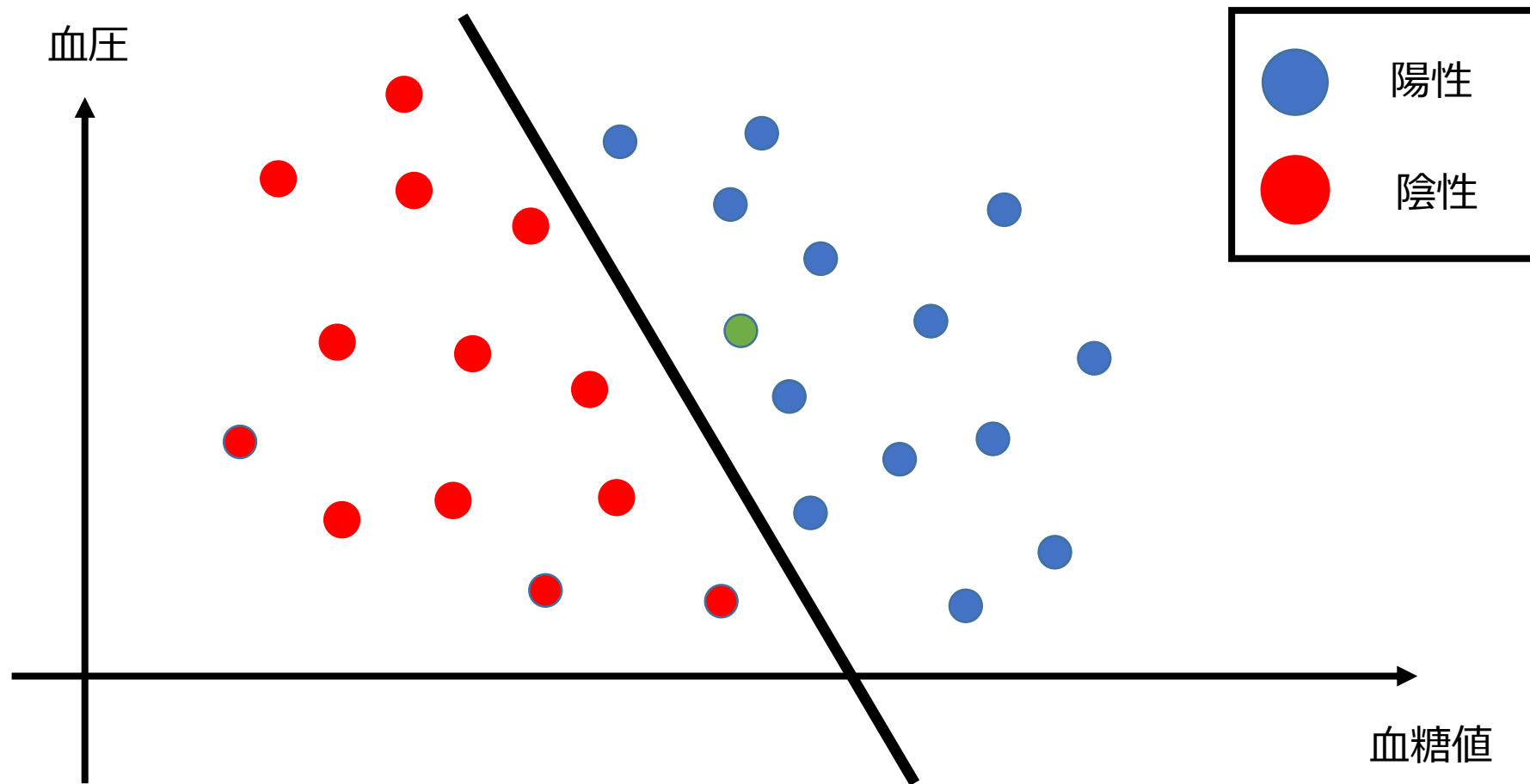


患者ID	血圧	血糖値	
1	120	200	陽性
2	80	130	陰性
3	110	190	?

陽性・陰性を識別する

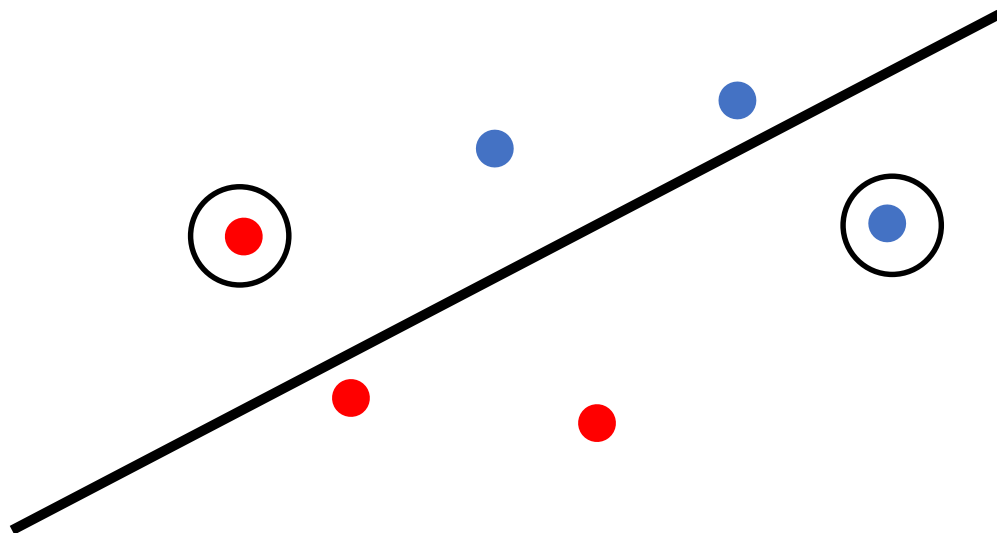


陽性・陰性を識別する



ロジスティック回帰

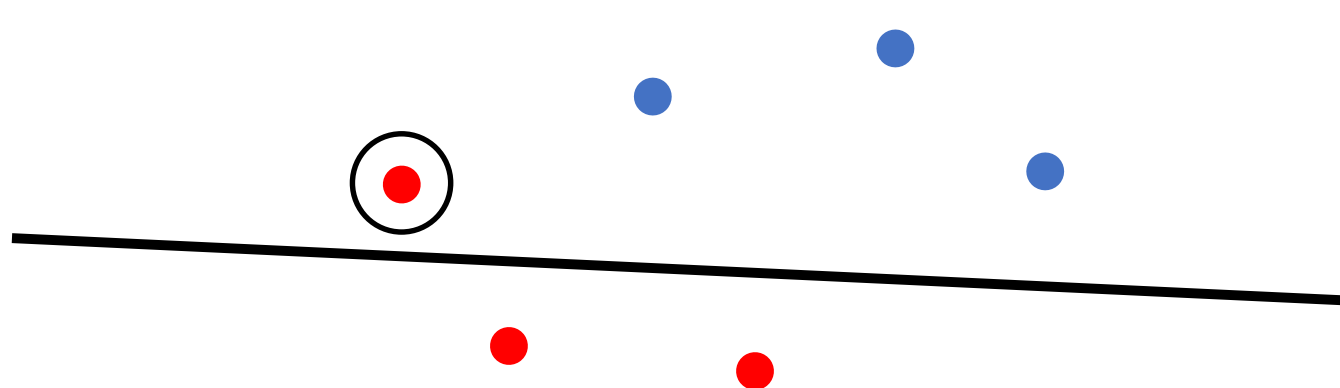
ロジスティック曲線の求め方



間違えの数

2

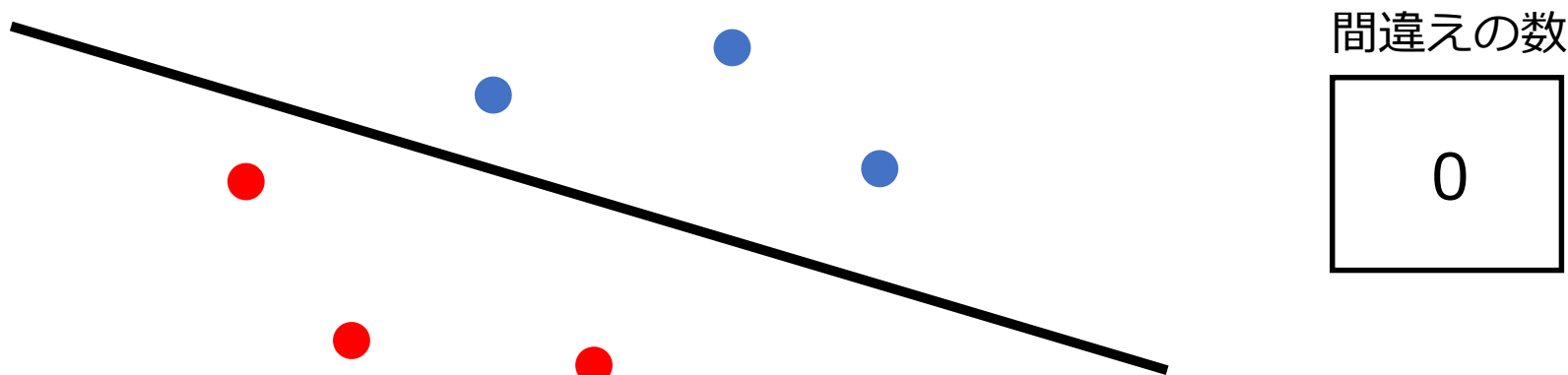
ロジスティック曲線の求め方



間違えの数

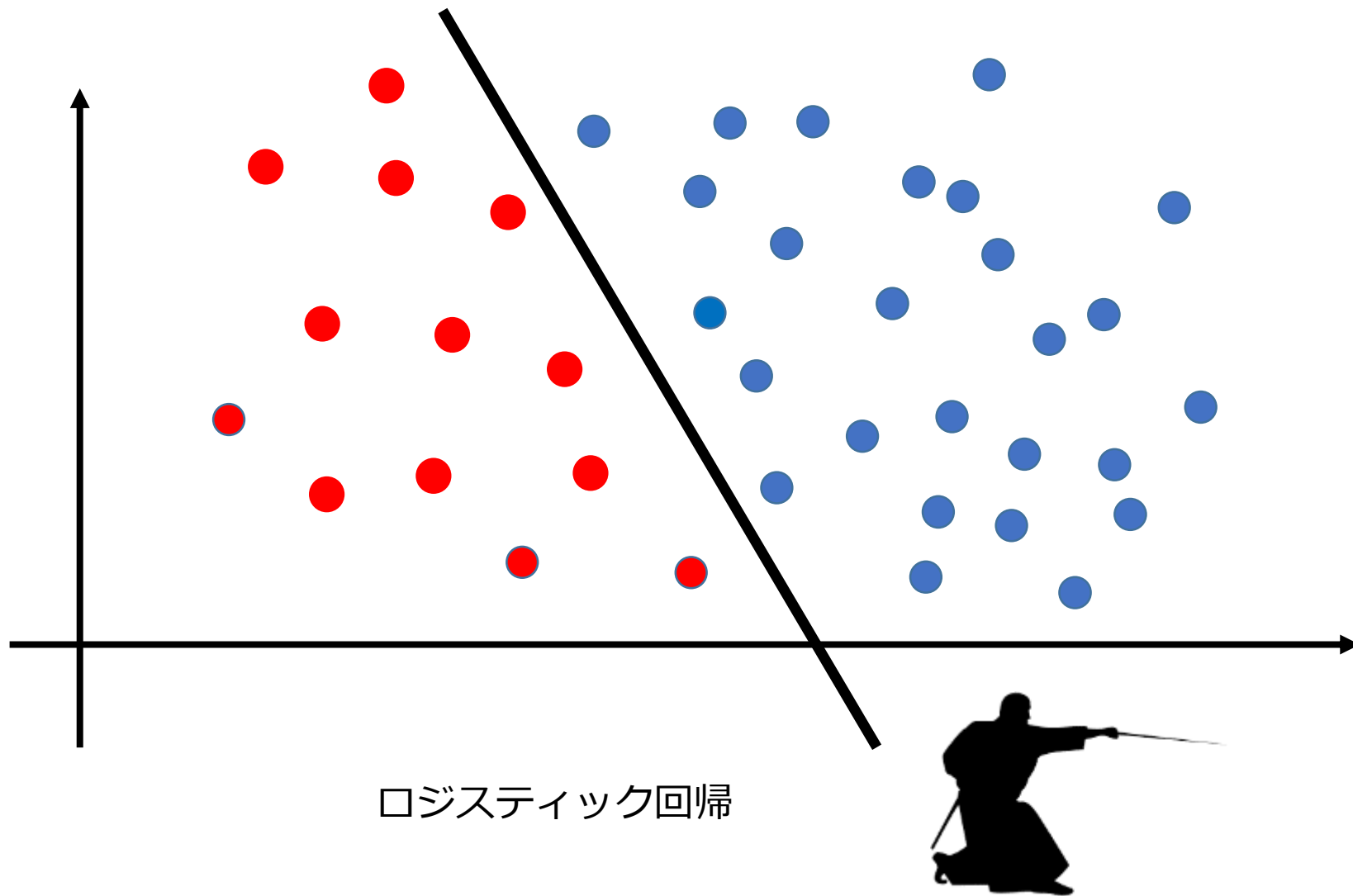
1

ロジスティック曲線の求め方

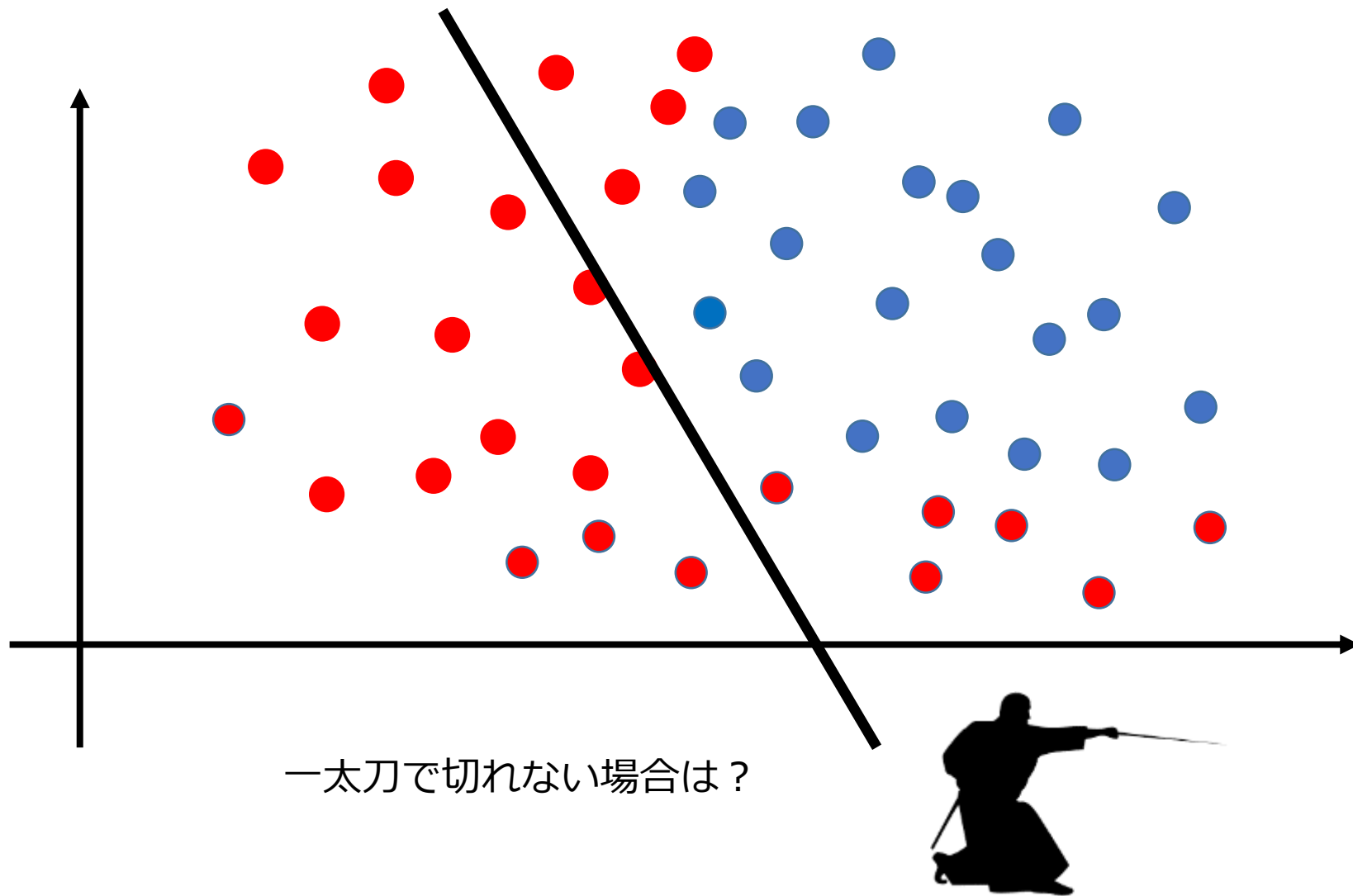


最尤法によってロジスティック曲線を求める

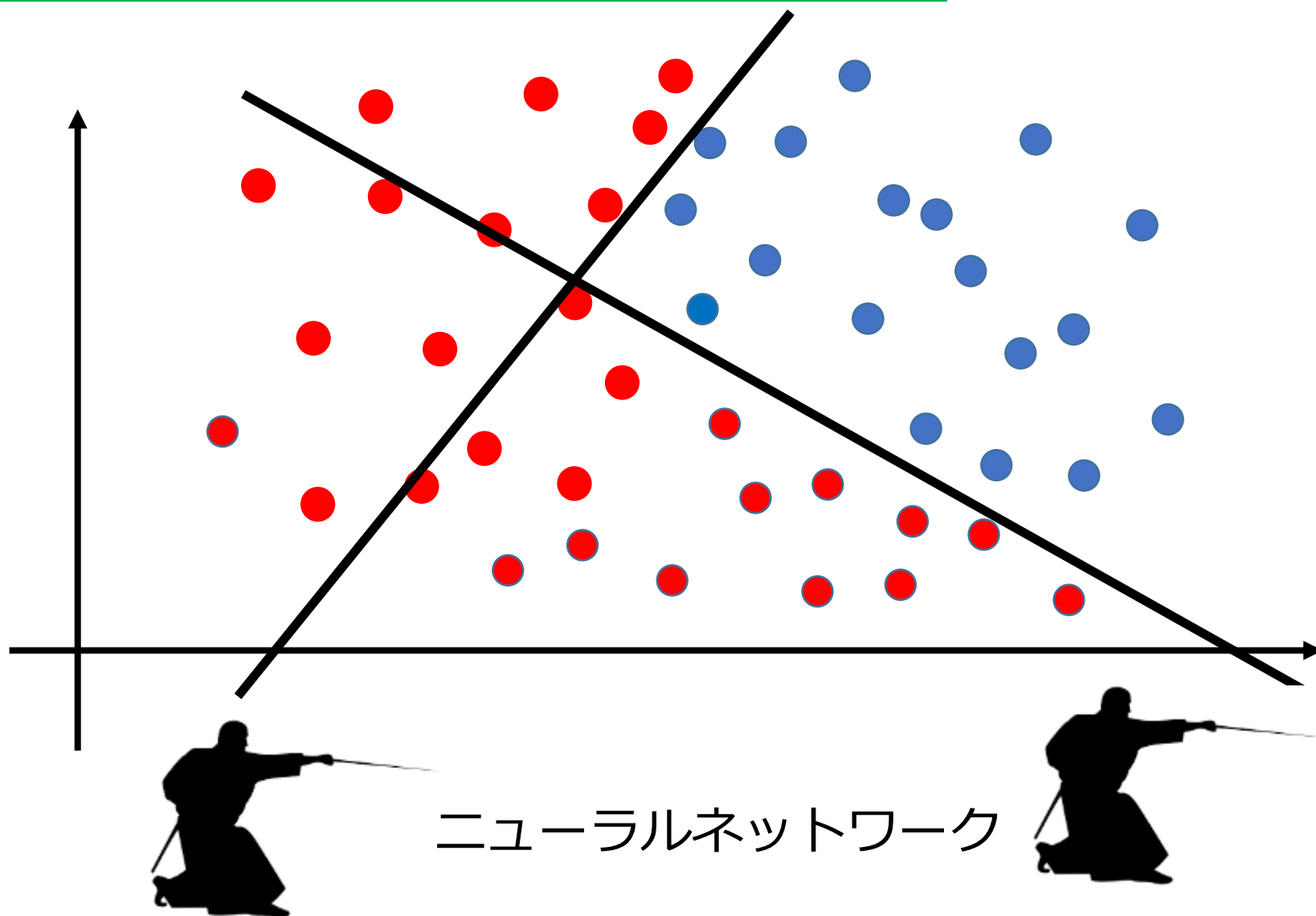
ロジスティック回帰分析



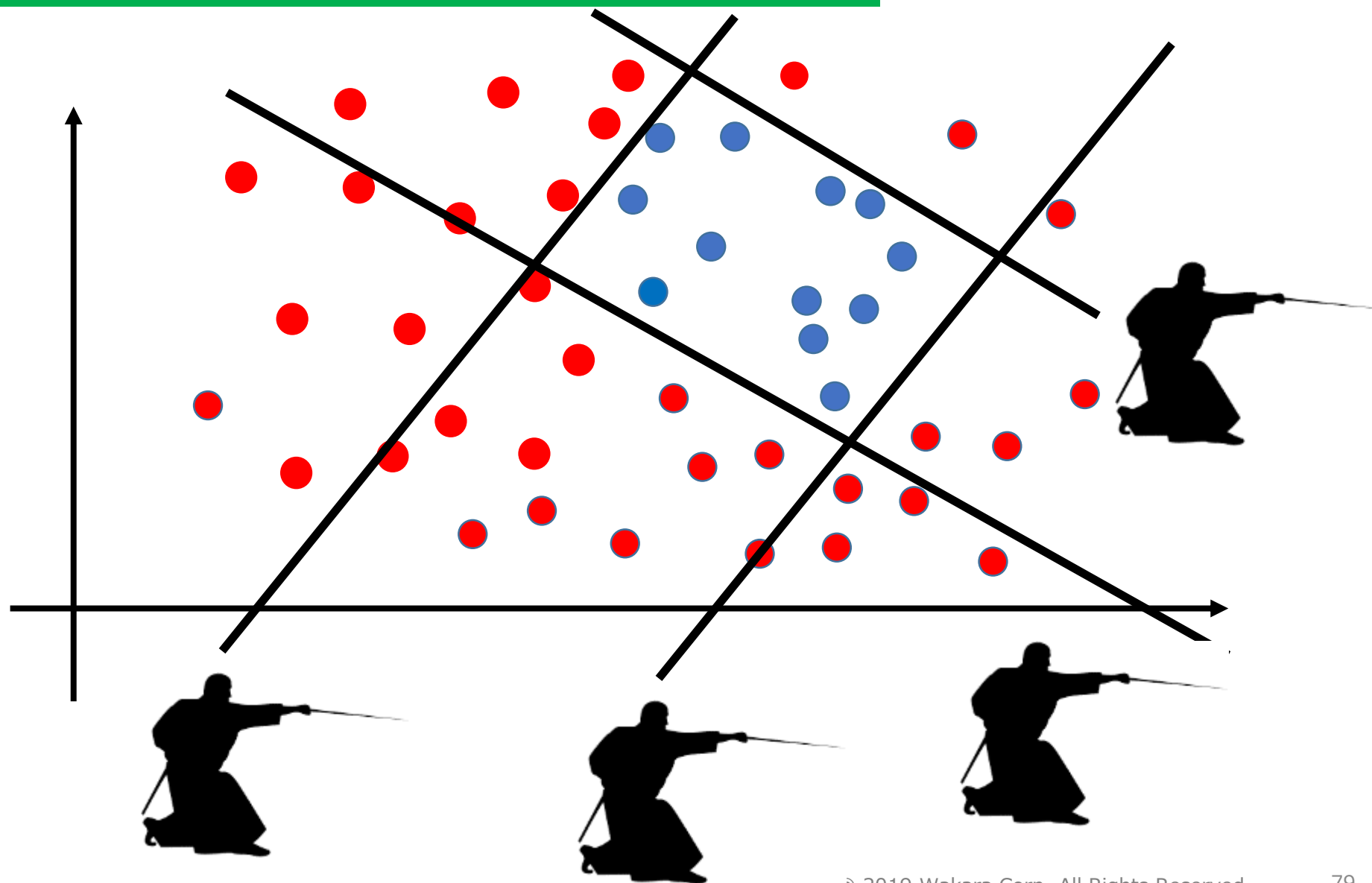
いろいろな識別方法



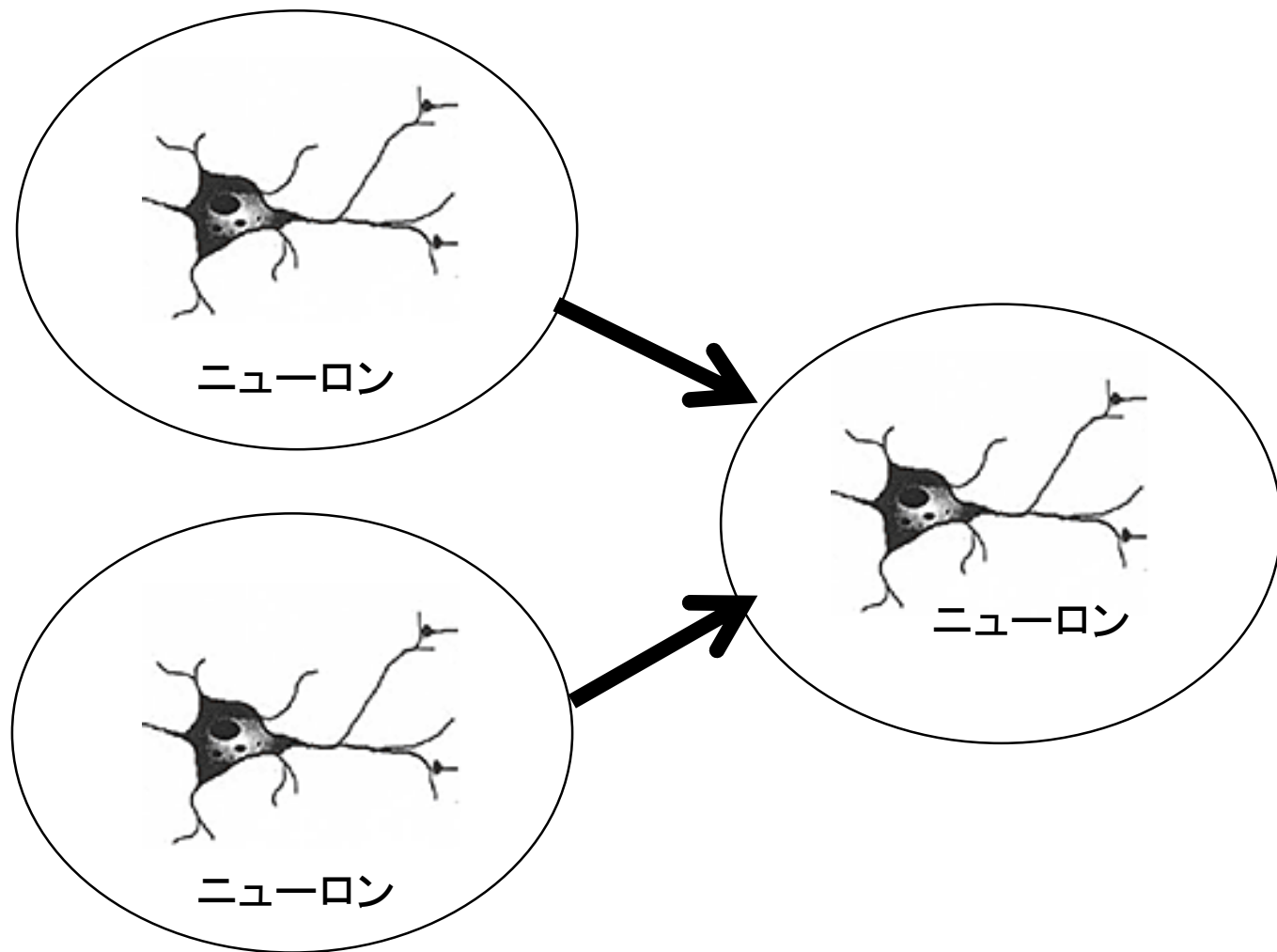
いろいろな識別方法



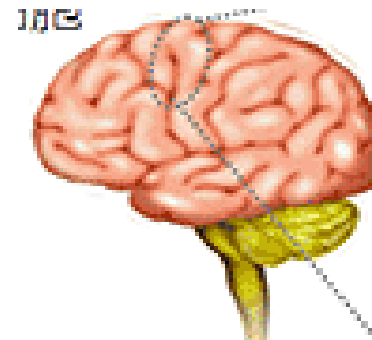
いろいろな識別方法

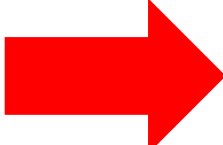


ニューラルネットワーク

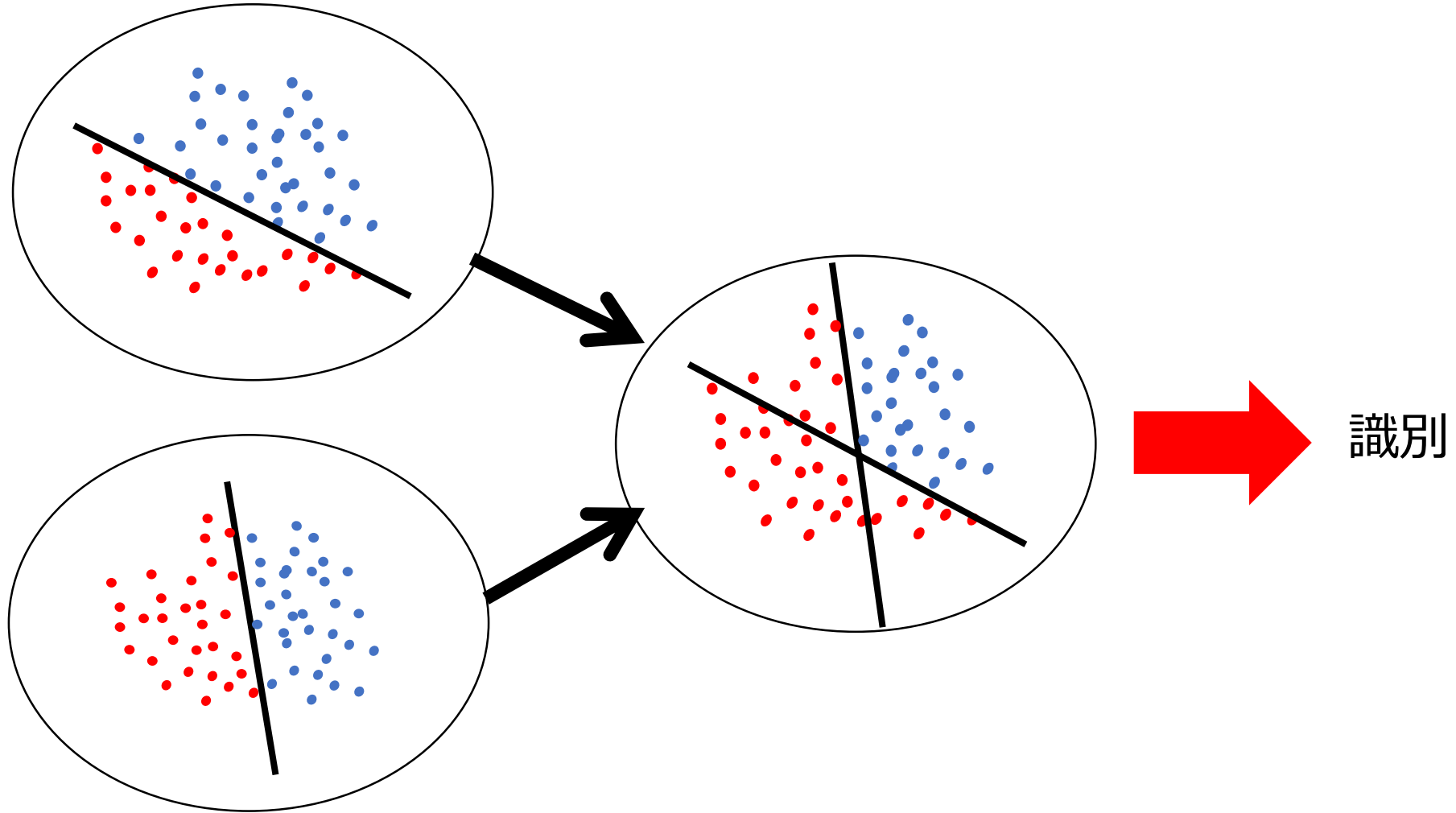


人間の脳はどやって識別
を行なっているのか？

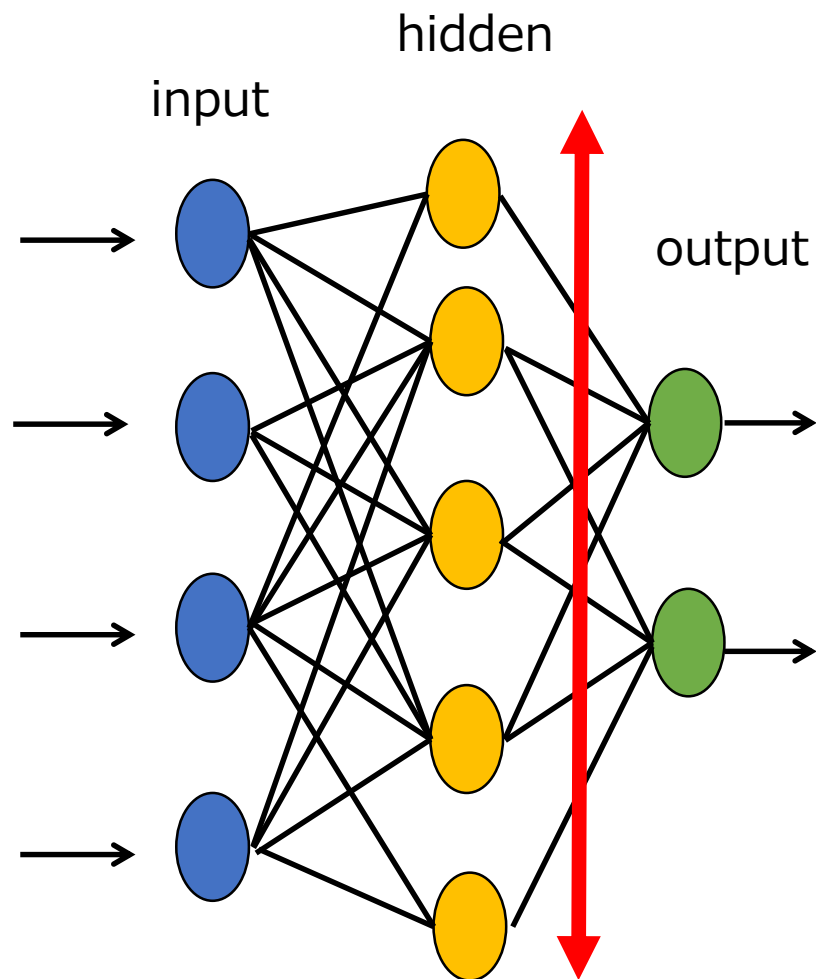


 識別

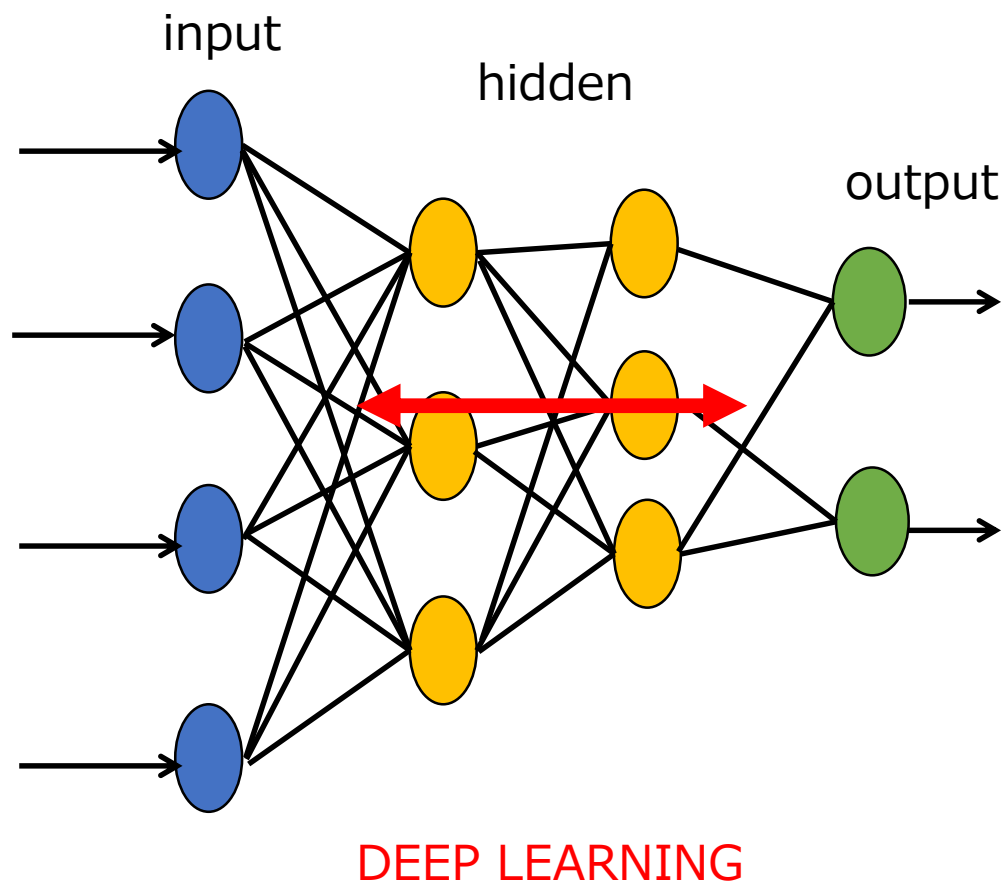
ニューラルネットワーク



ニューラルネットワークとディーププランニング



ニューラルネットワーク



はじめてDeep Learningを学ぶ方に



機械学習の区分

機械学習

識別

AかBか

決定木



ナイーブベイズ



ニューラル
ネットワーク



SVM



ロジスティック回帰



回帰

どのくらいの量か

重回帰分析



分類

どう分けるか

k-means法



主成分分析

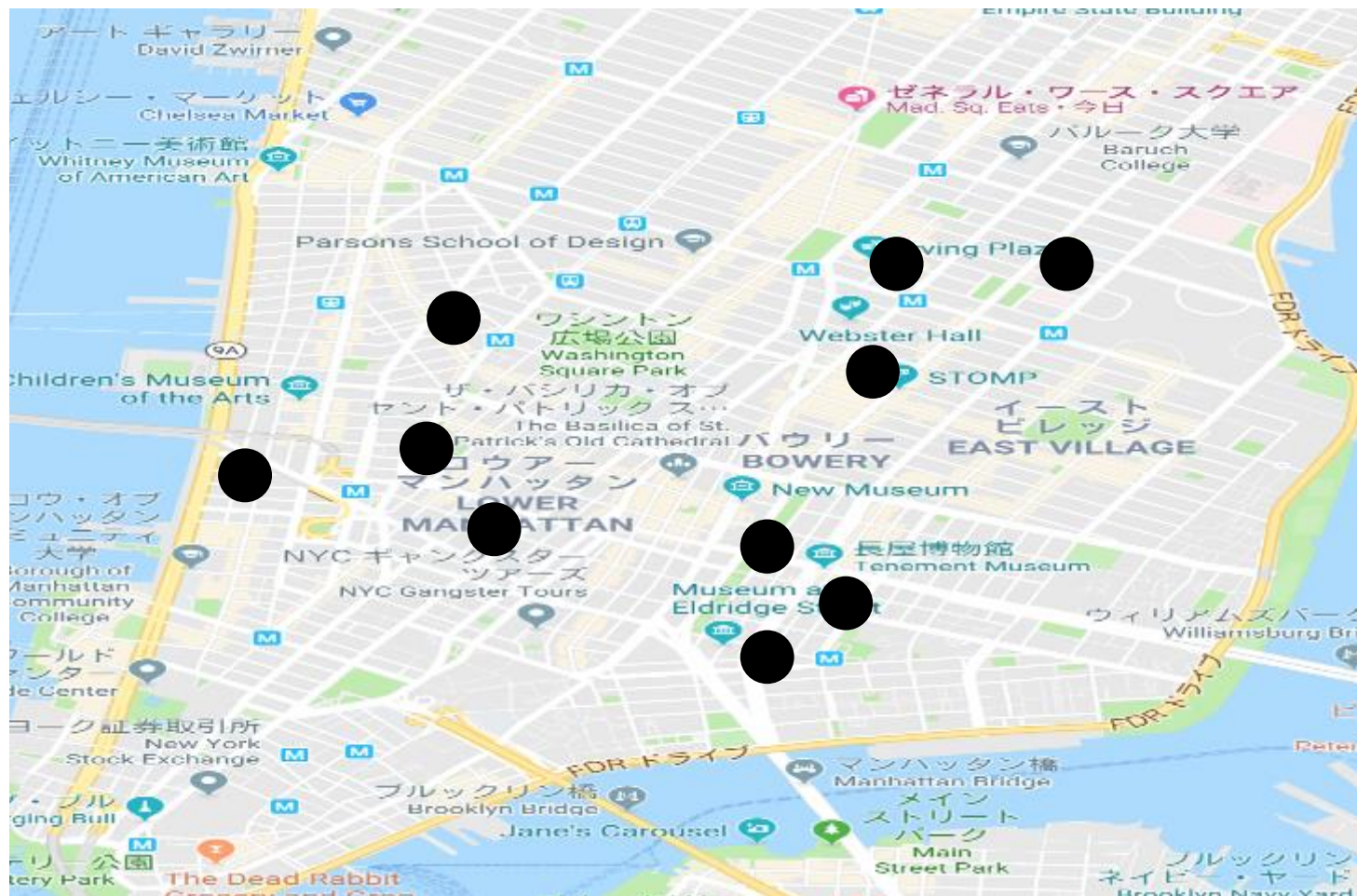


クラスター分析

- K-mean法

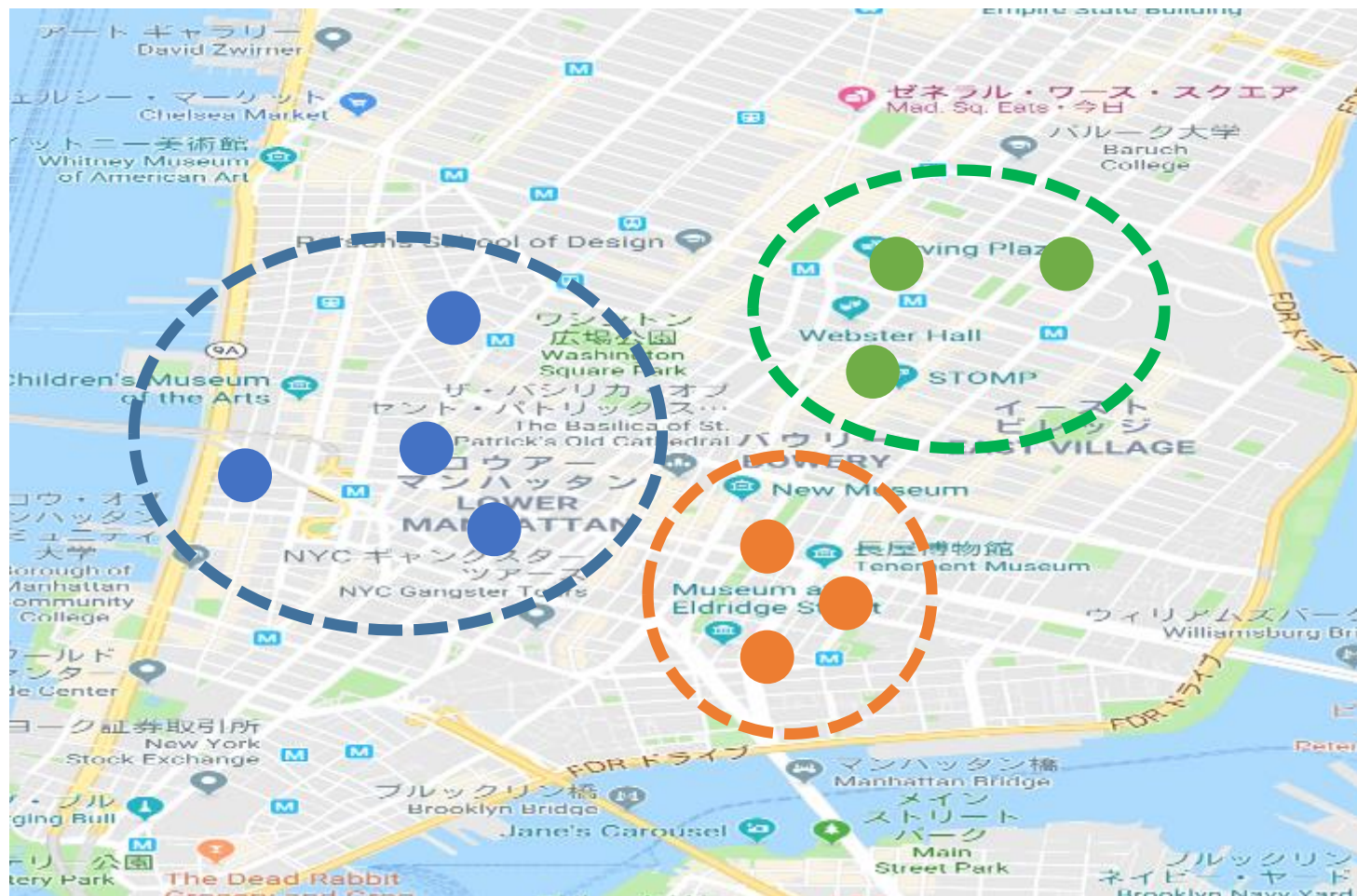
どこにピザ屋を出店するか？

3つのグループに分けるとしたら、どのようなグループ分けを行うか？



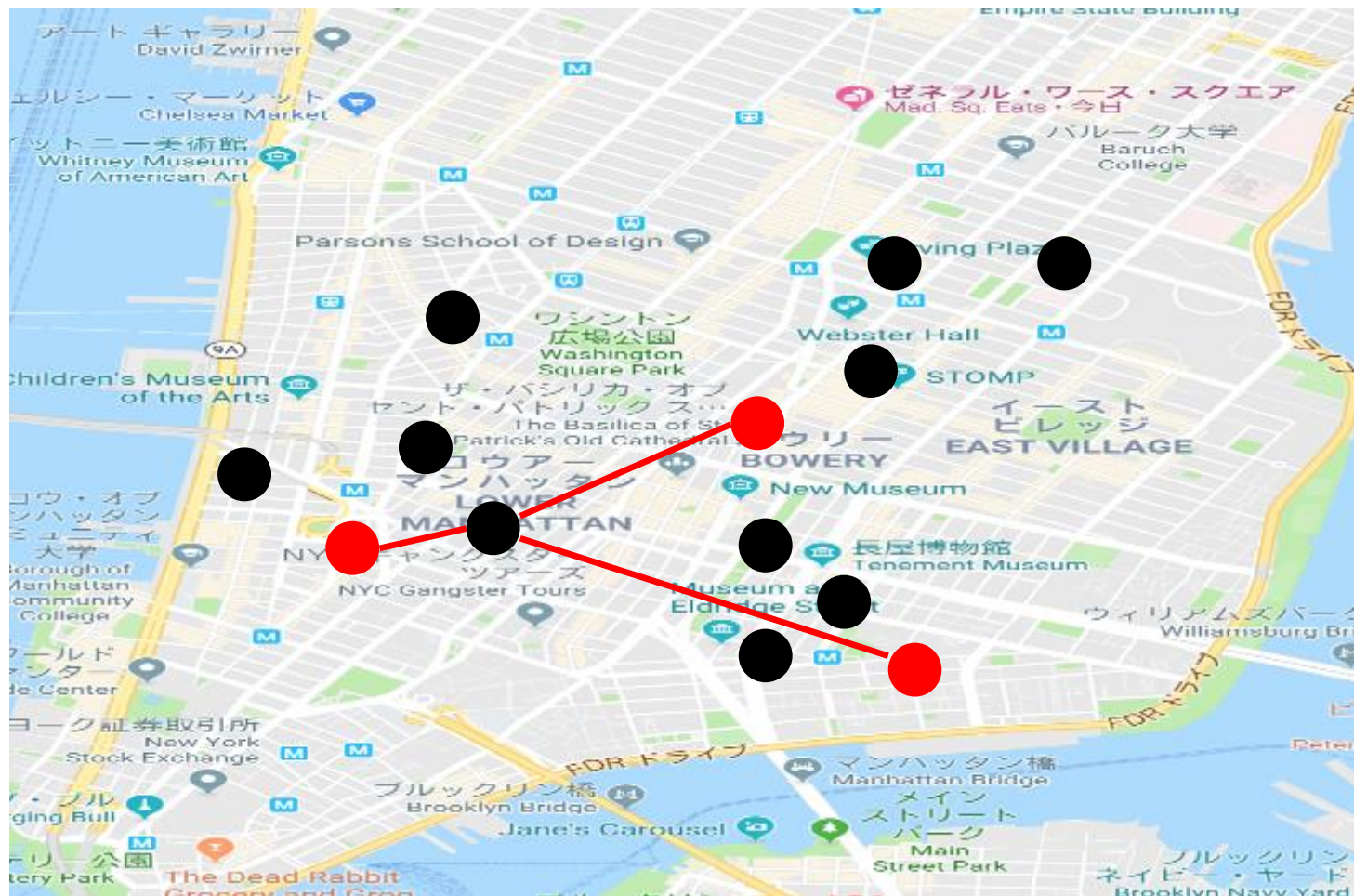
どこにピザ屋を出店するか？

3つのグループに分けるとしたら、どのようなグループ分けを行うか？



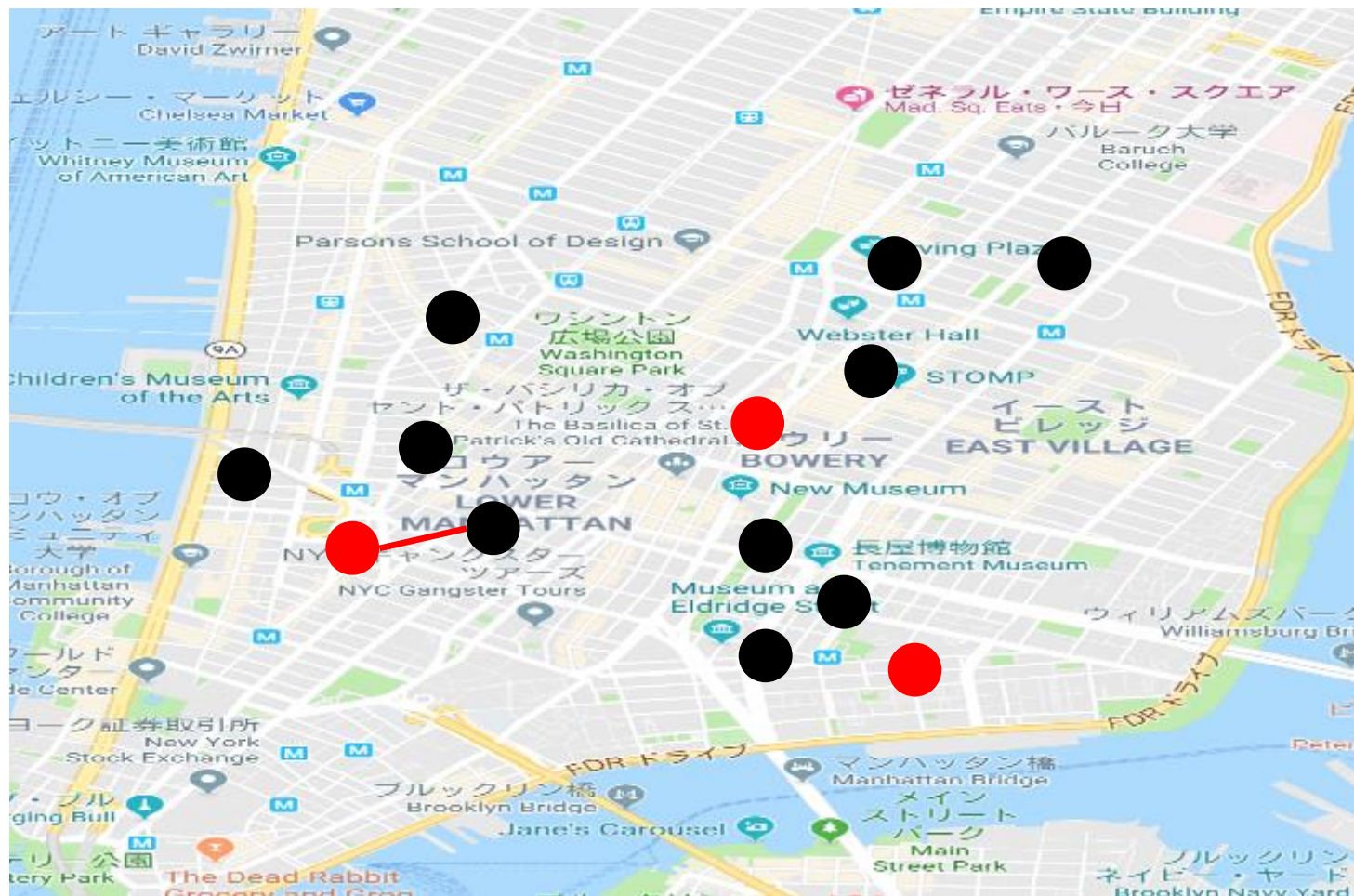
K-mean法

シード(seed)と呼ばれる点●を適当に配置する。各点からシードまでの距離を測る



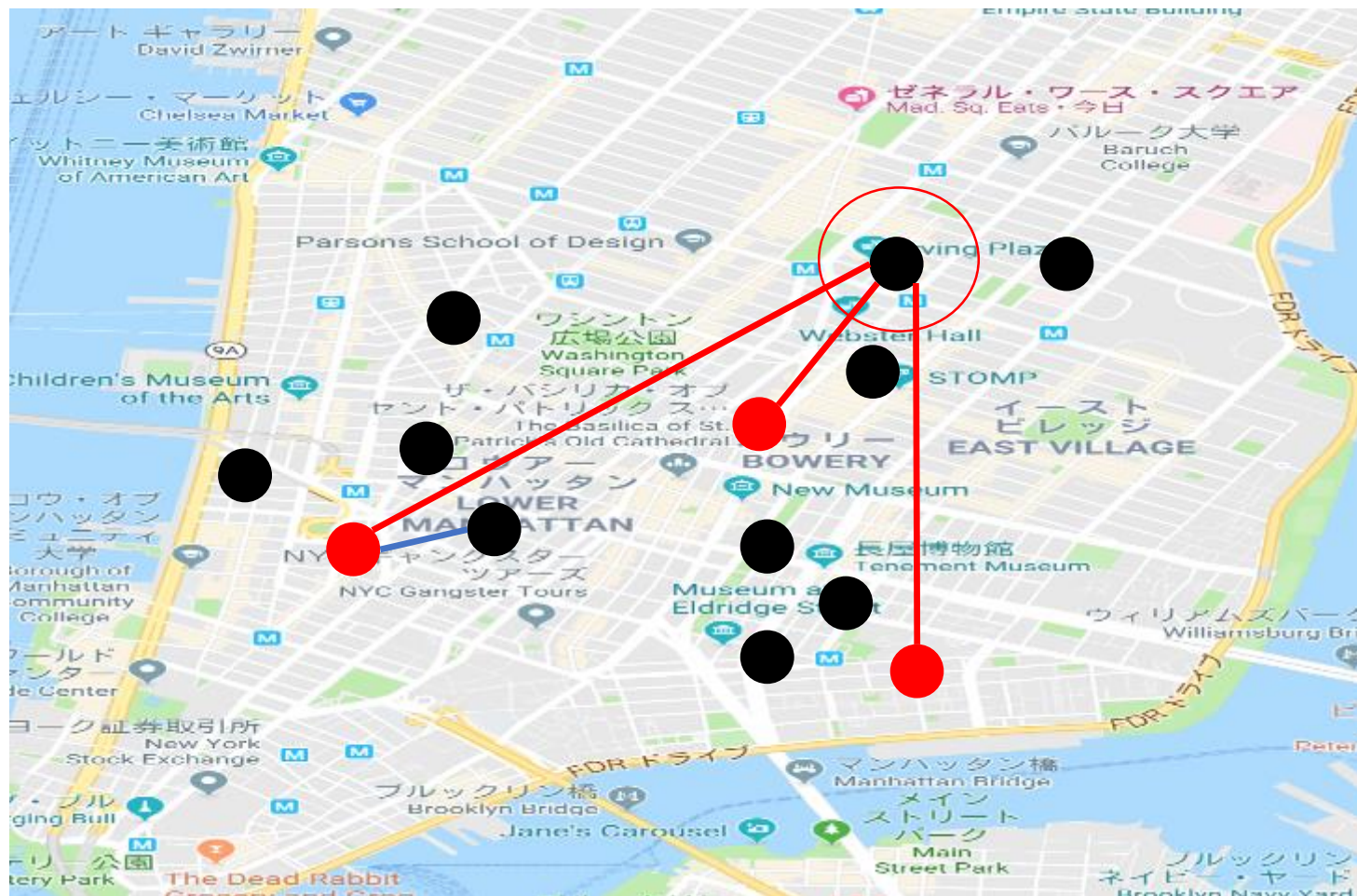
K-mean法

各点からシードまでの距離を測り、一番距離の短いシードと紐づける



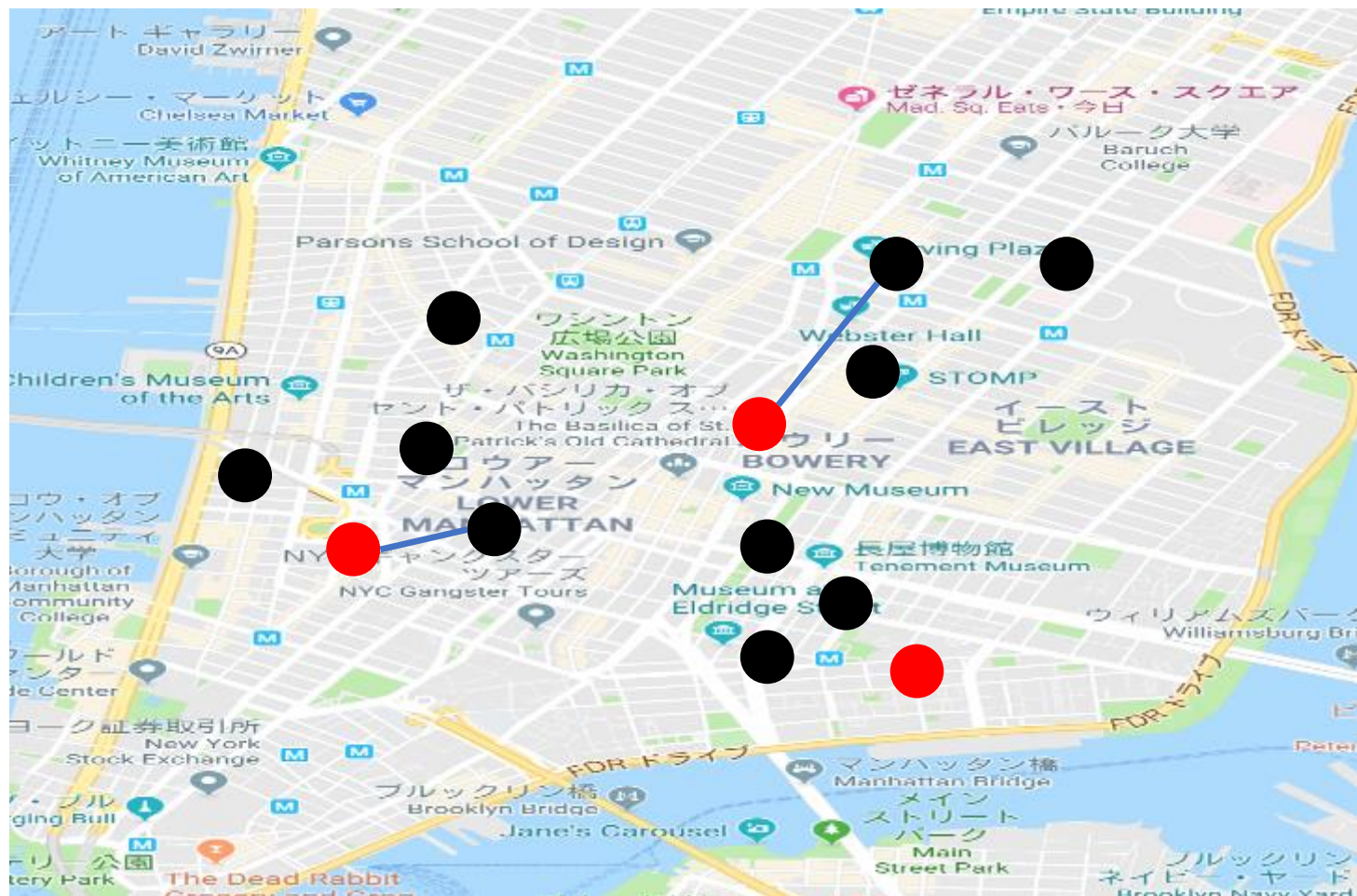
K-mean法

各点からシードまでの距離を測り、一番距離の短いシードと紐づける



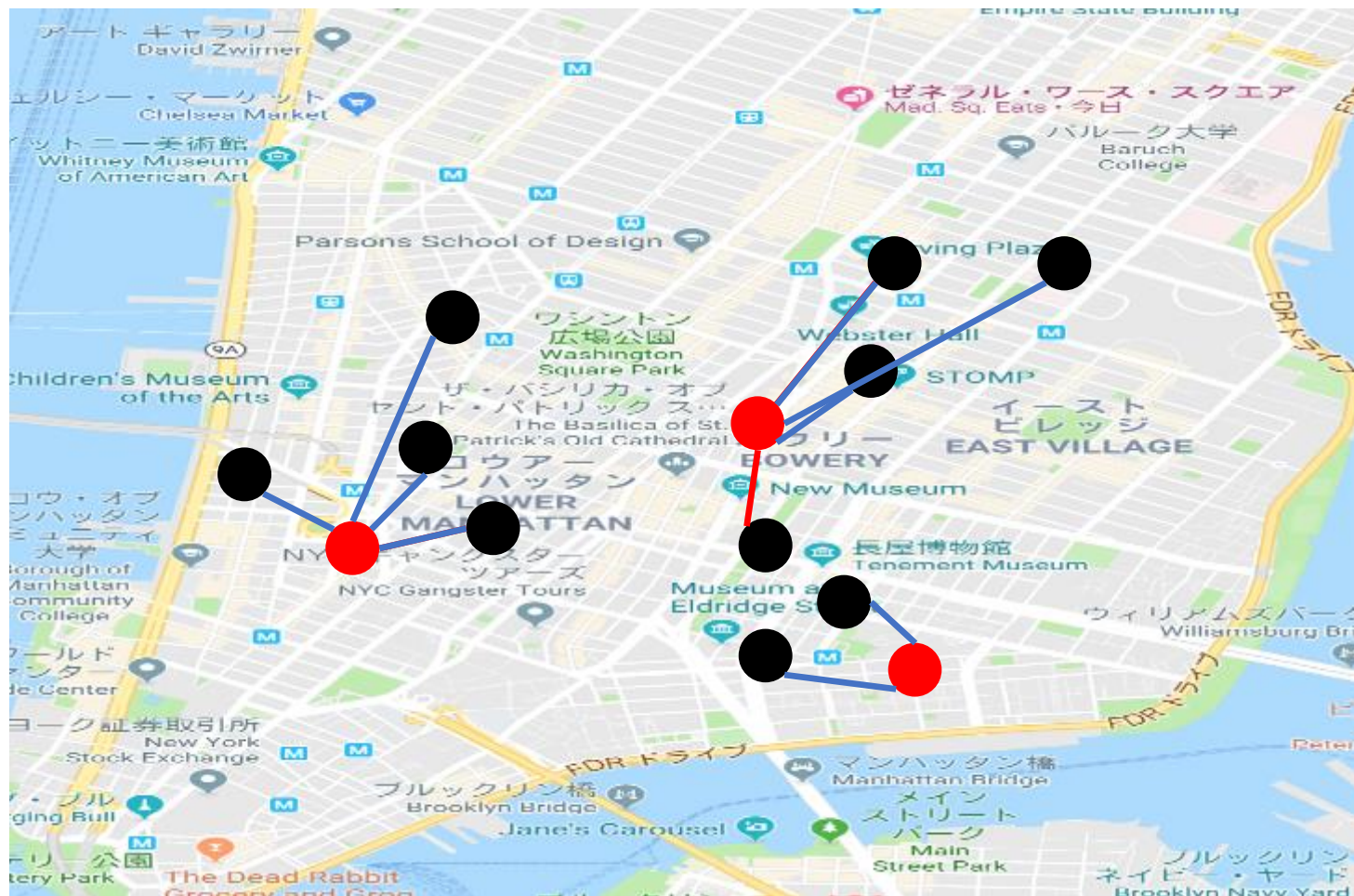
K-mean法

各点からシードまでの距離を測り、一番距離の短いシードと紐づける



K-mean法

各点からシードまでの距離を測り、一番距離の短いシードと紐づける



K-mean法

シードごとにグルーピングを行う。この時各グループを**クラスター**と呼ぶ。



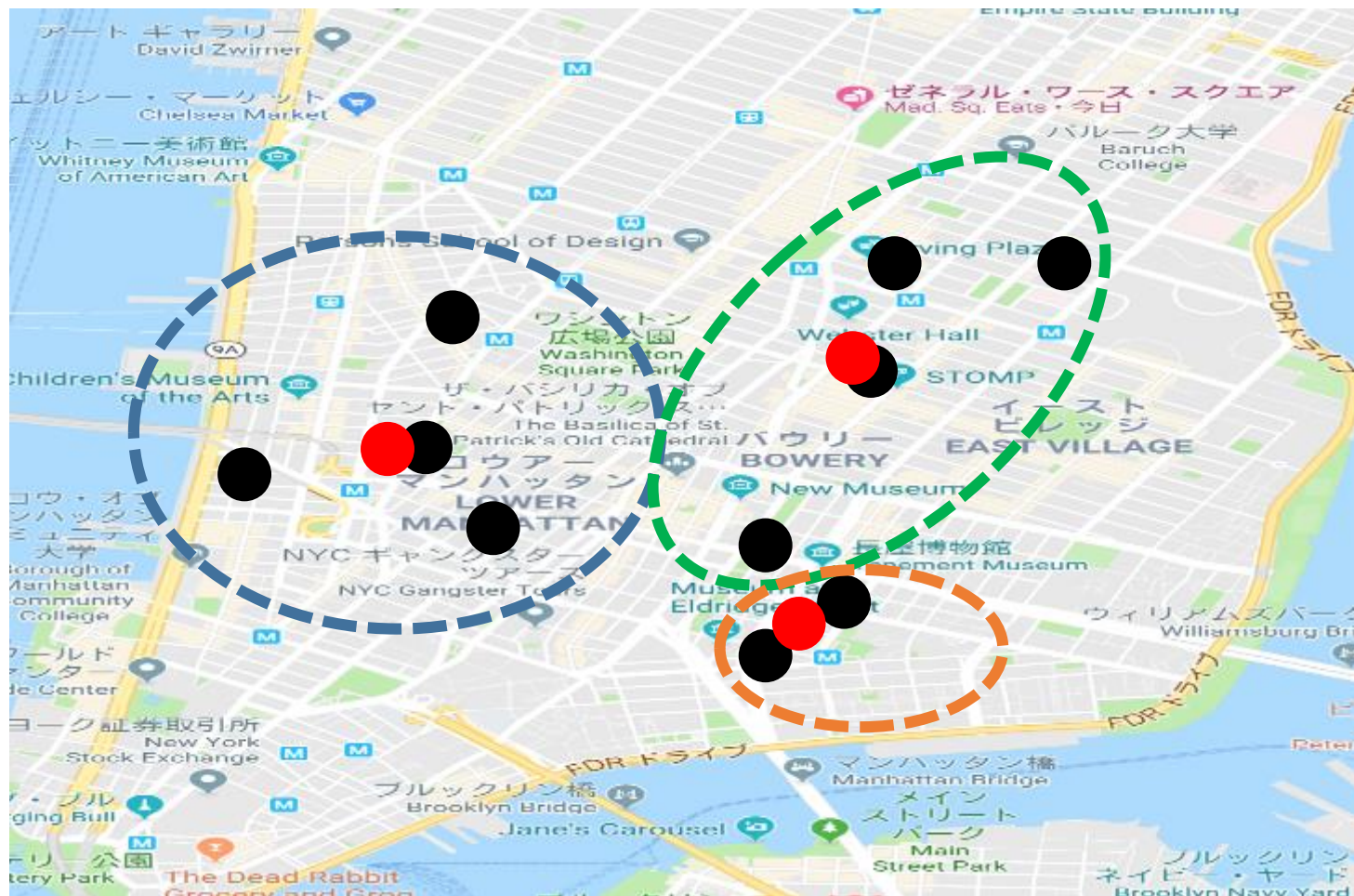
K-mean法

クラスターが出来たらシードを除去する。



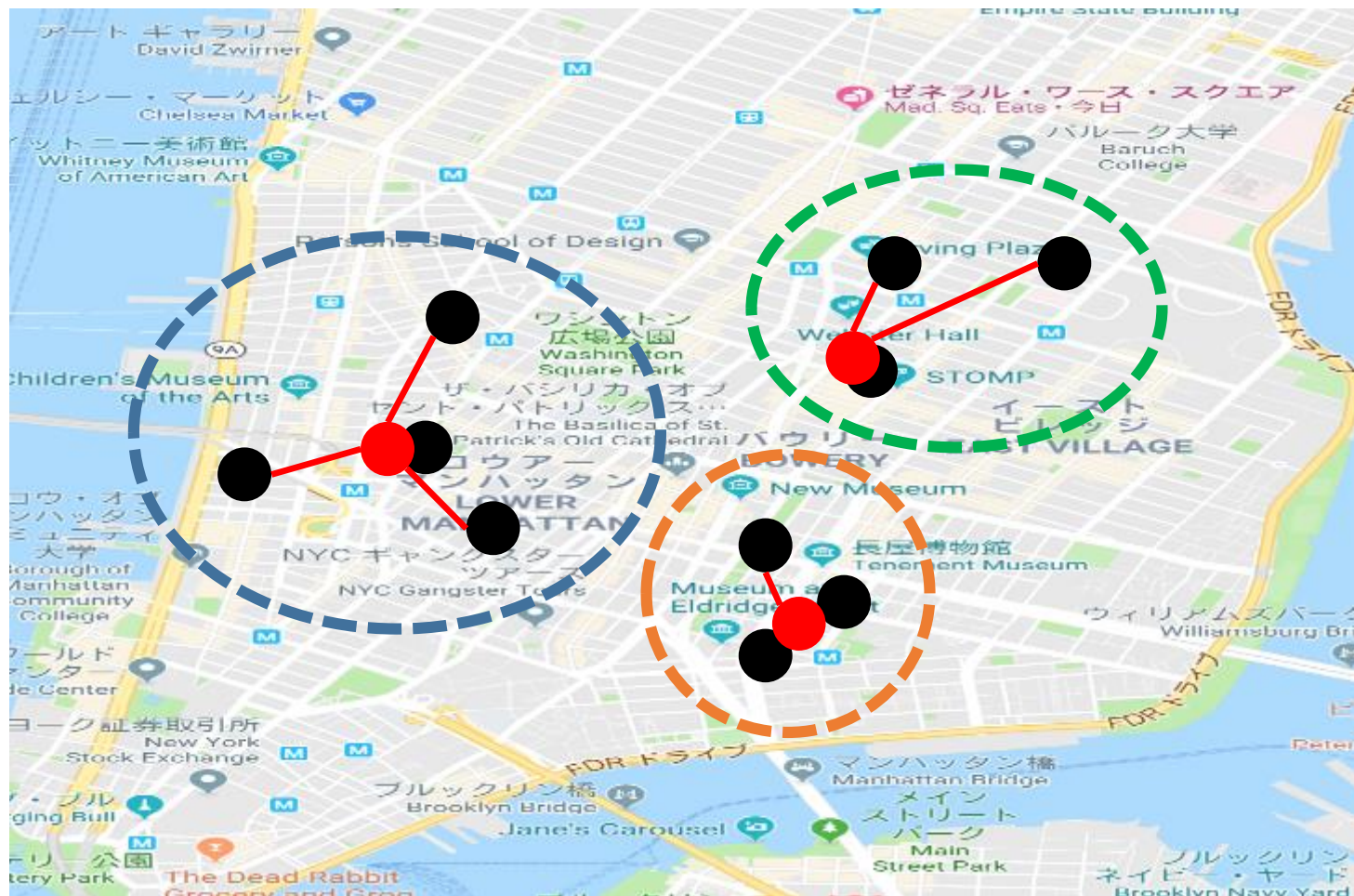
K-mean法

各クラスター内のデータの平均点(重心)を新たなシードとする。



K-mean法

各データと最も近いシードを紐づける。



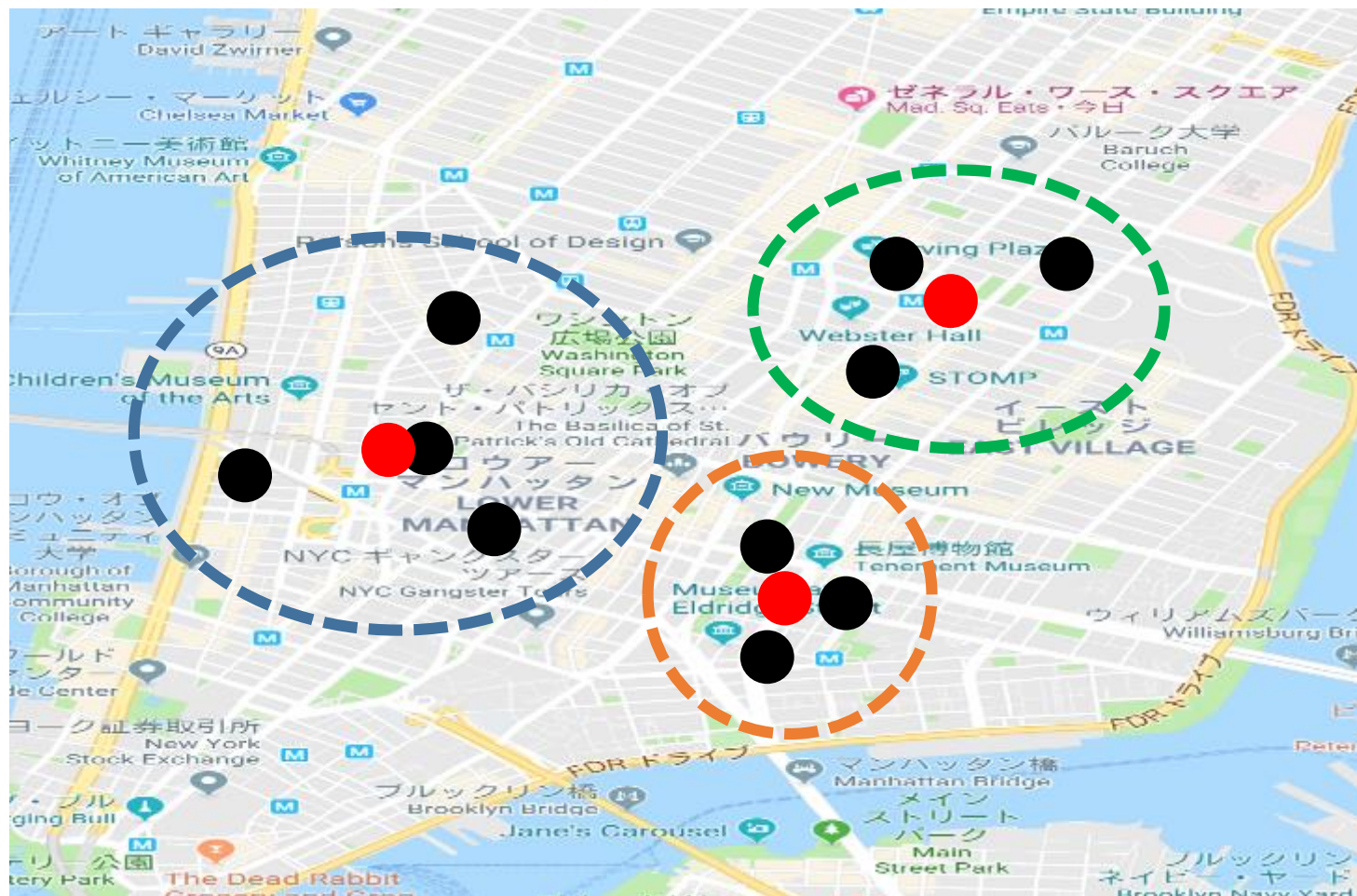
K-mean法

いったんクラスター分類を外し、



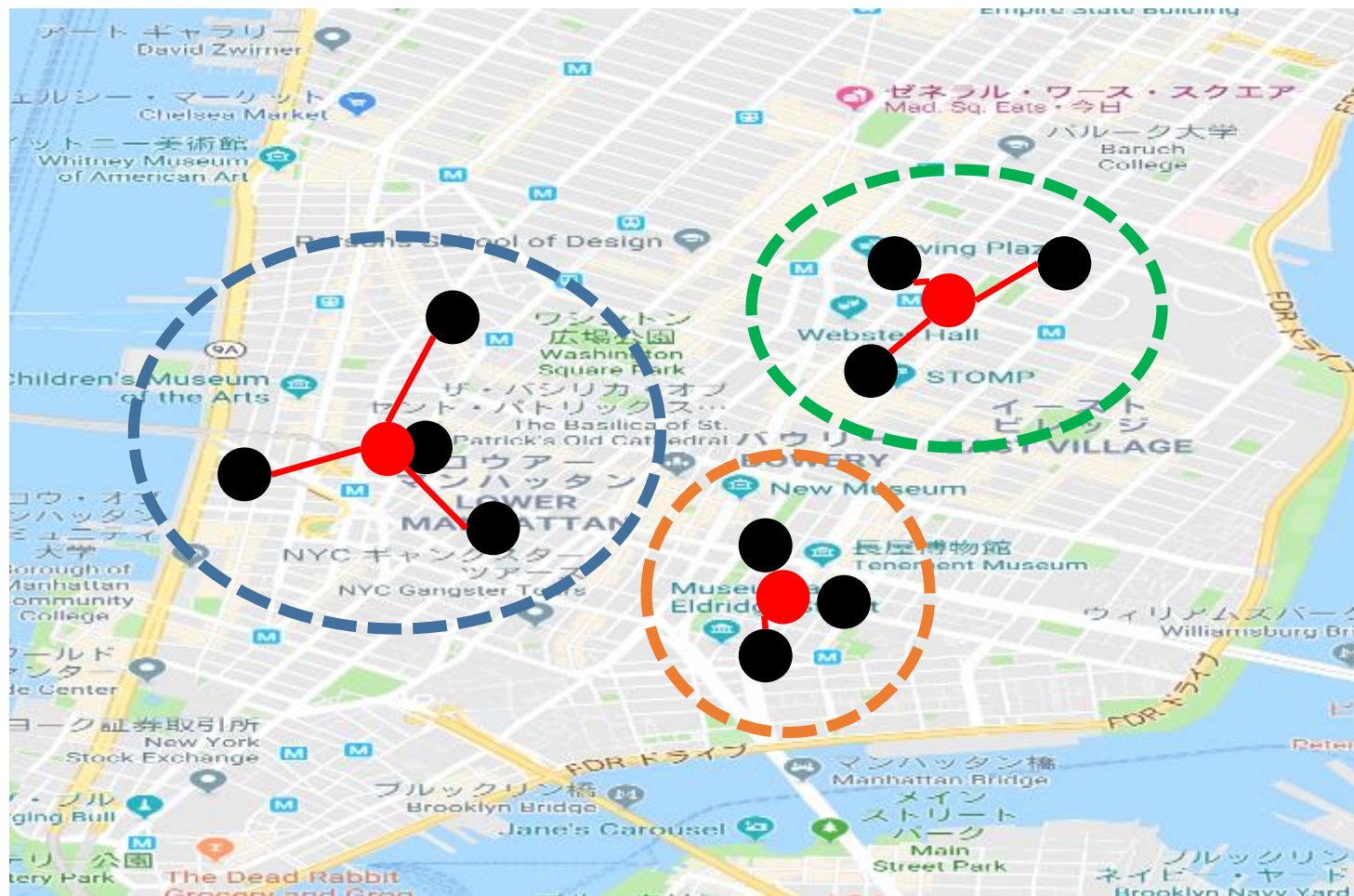
K-mean法

各クラスター内のデータの平均点(重心)を新たなシードとする。



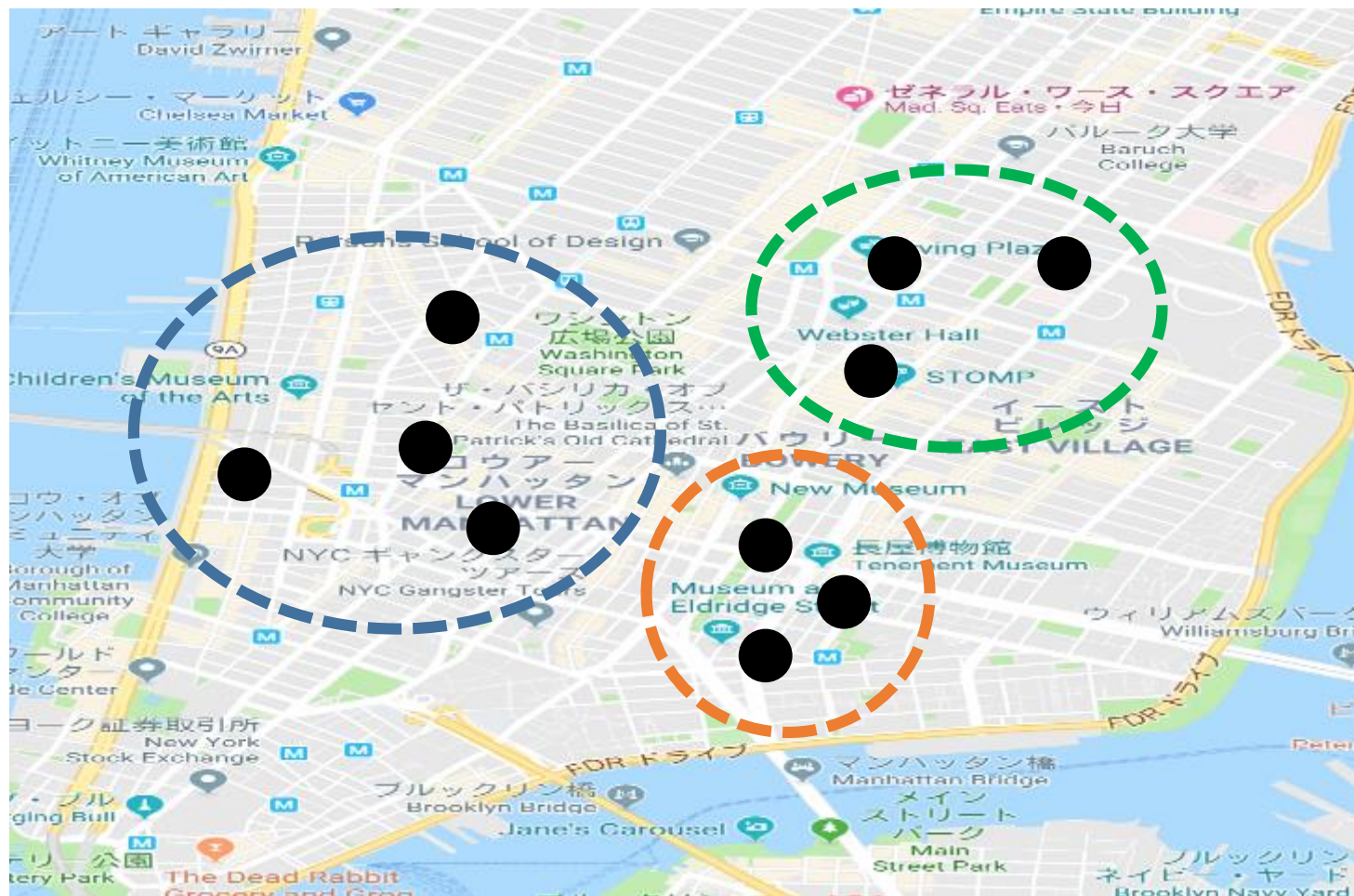
K-mean法

各データと最も近いシードを紐づける。



K-mean法

以上の過程を繰り返し、クラスターに変動がなくなれば終了。



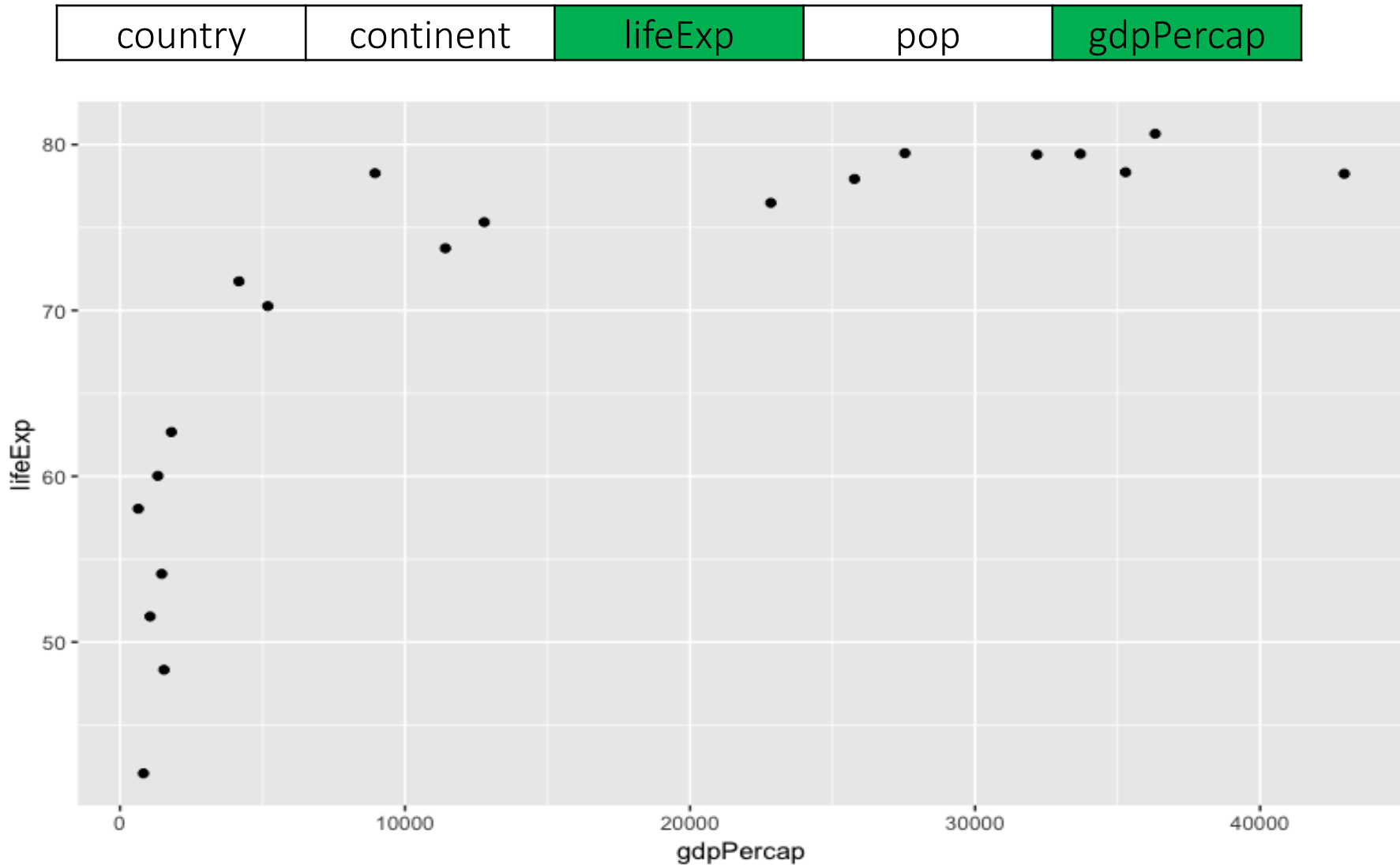
K-mean法の応用

country	continent	lifeExp	pop	gdpPercap
Argentina	Americas	75.32	40301927	12779.3796
Canada	Americas	80.653	33390141	36319.235
Cote d'Ivoire	Africa	48.328	18013409	1544.75011
Cuba	Americas	78.273	11416987	8948.10292
-----	-----	-----	-----	-----
Mauritania	Africa	62.664	3270065	1803.1515
Belgium	Europe	79.441	10392226	33692.6051

データの可視化

country	continent	lifeExp	pop	gdpPercap
		75.32	40301927	12779.3796
		80.653	33390141	36319.235
		48.328	18013409	1544.75011
		78.273	11416987	8948.10292
		-----	-----	-----
		62.664	3270065	1803.1515
		79.441	10392226	33692.6051

データの可視化



4つのクラスターに分類

