

# プログラムなしではじめる 機械学習超入門

# 機械学習とは？

学習する？



人は経験から学習する



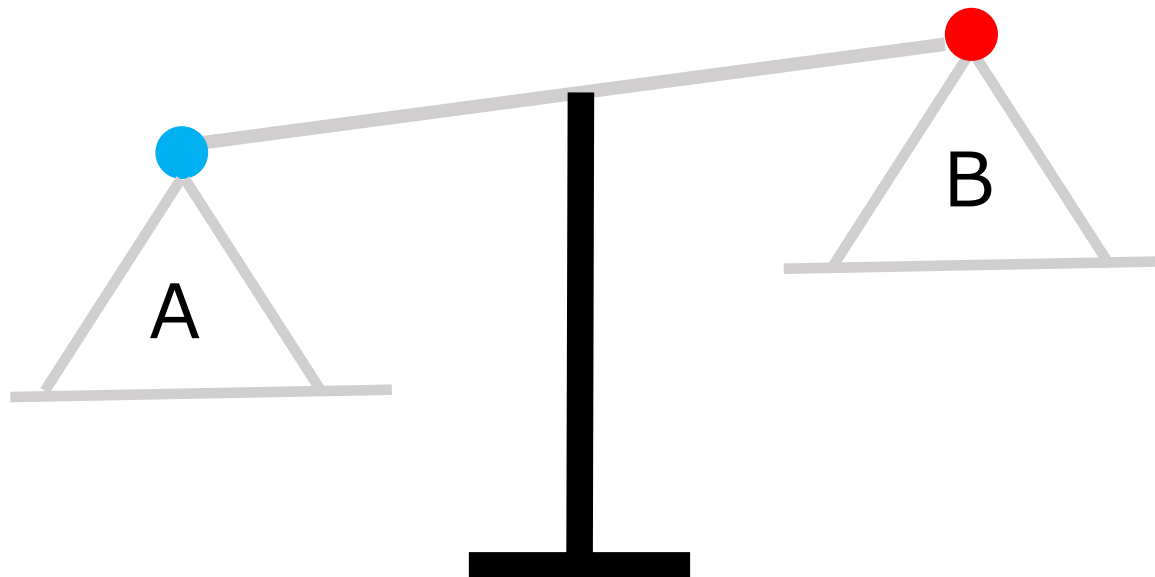
機械はデータから学習する

機械学習とは過去のデータから目的に応じて、コンピューターが知識や規則性を発見し、推論、識別、予測を行うこと

# 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

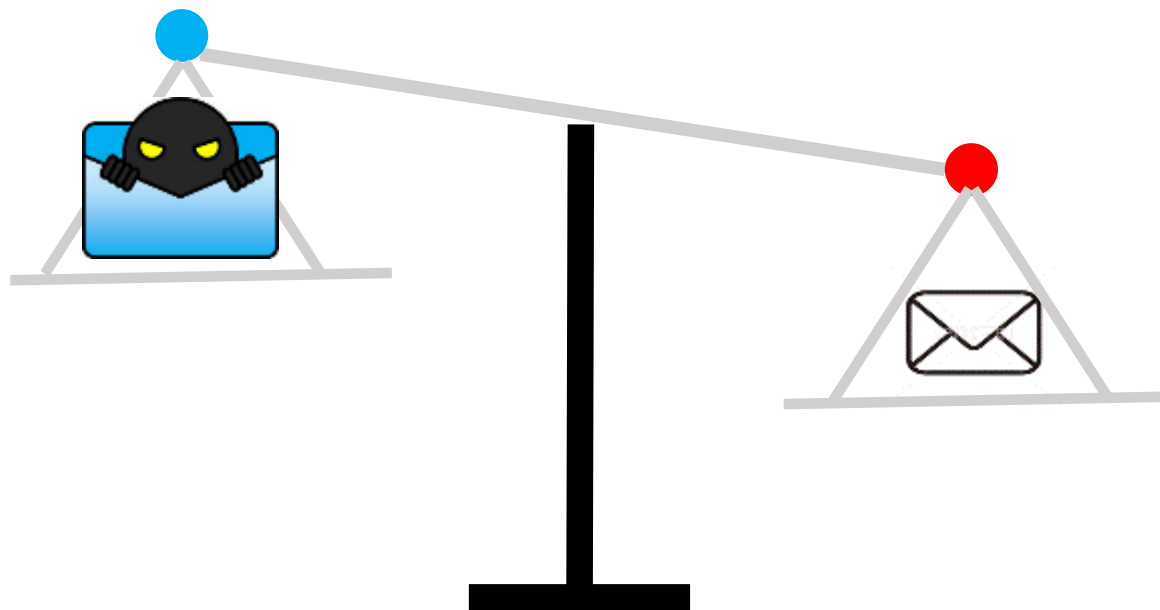


# 機械学習で問われる 3つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

ナイーブベイズ

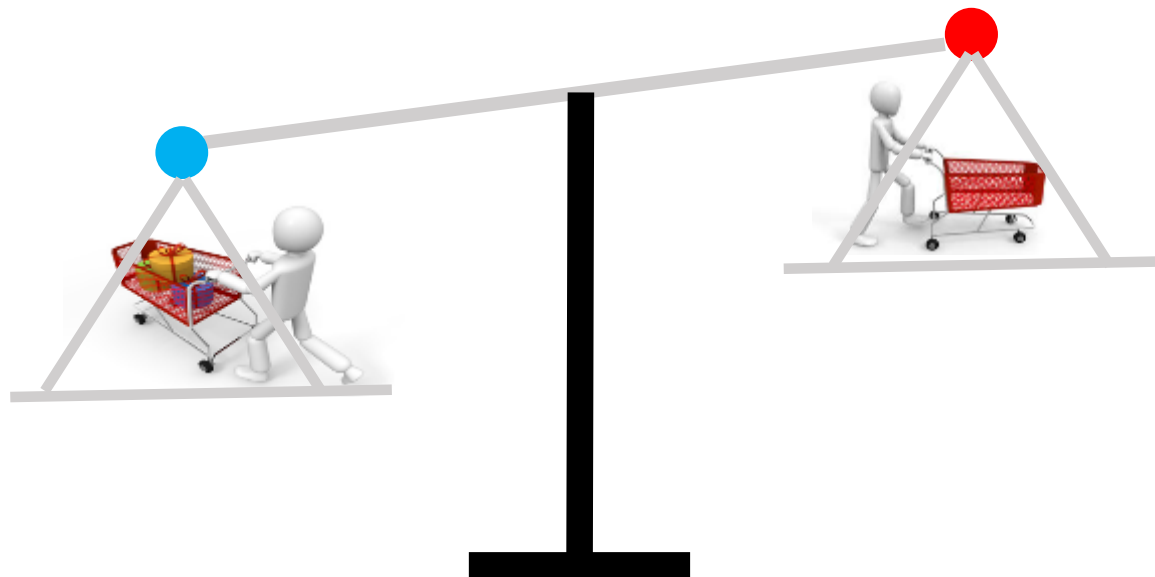


# 機械学習で問われる 3つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

決定木



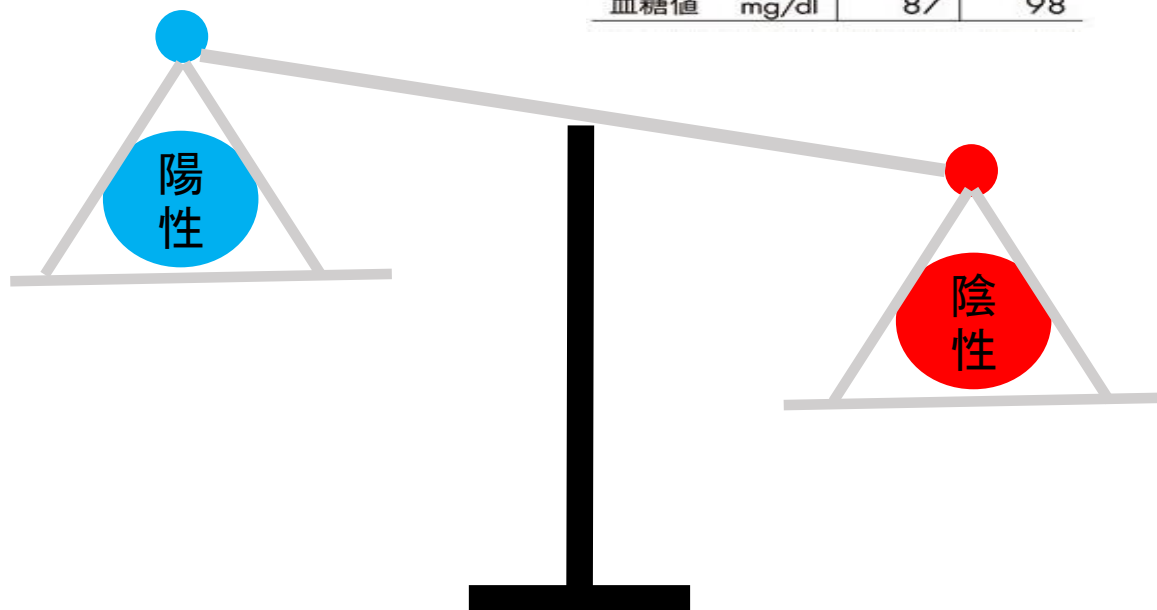
# 機械学習で問われる3つの質問

- 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



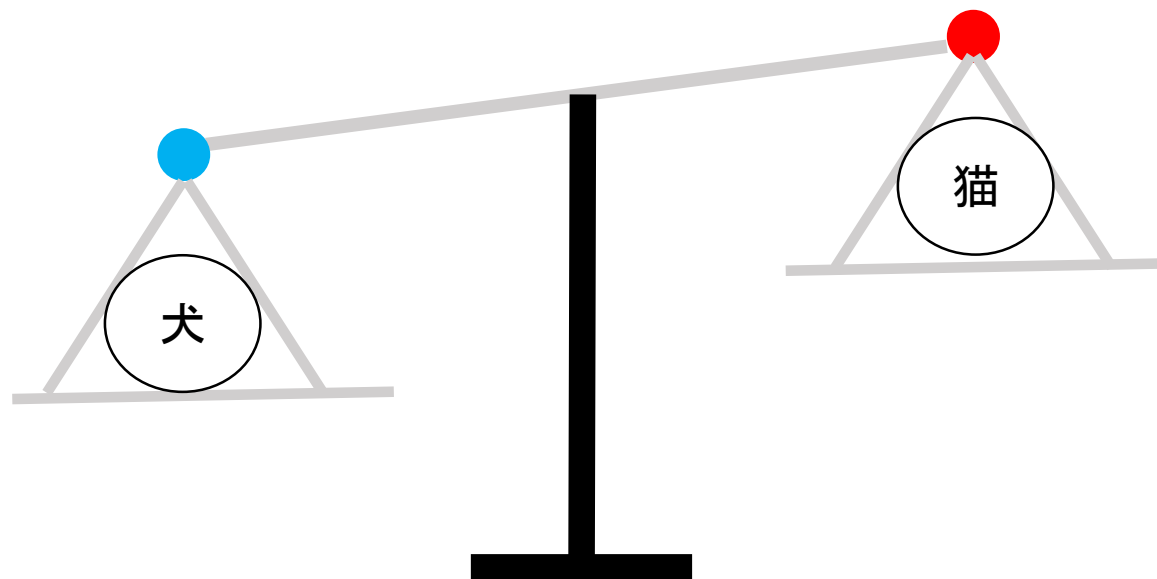
# 機械学習で問われる 3つの質問

- 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 1

「どのくらいの量または数か？」

回帰アルゴリズム



この物件の価格は？



800万円



2億5千万円

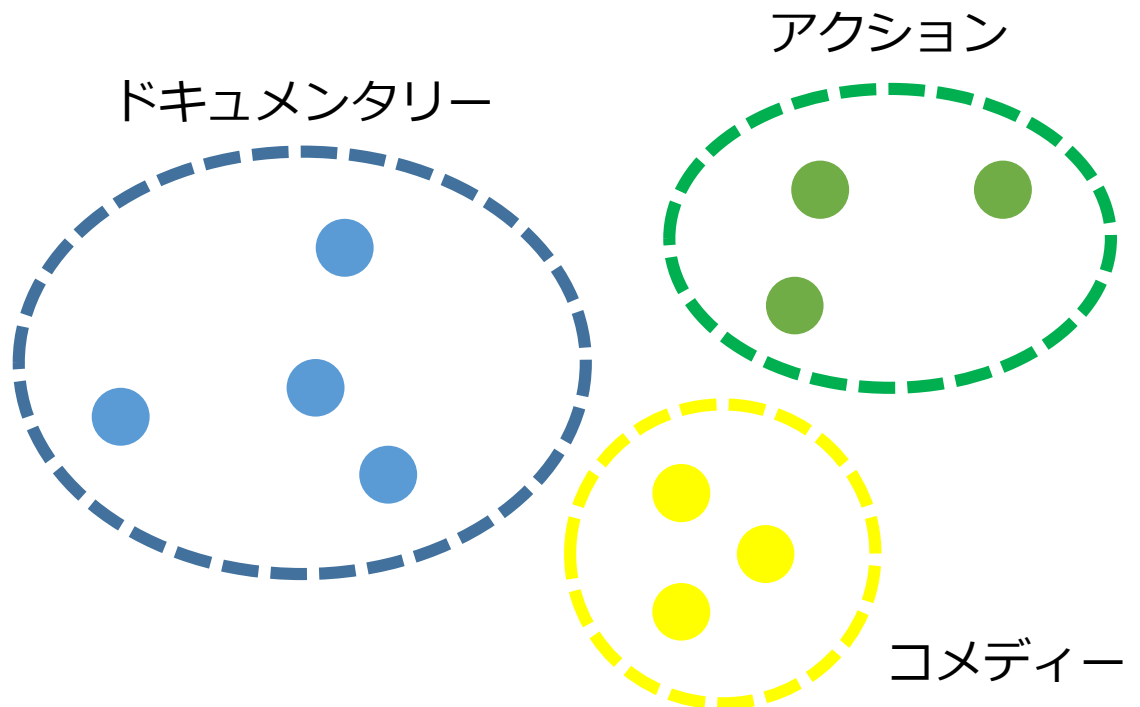
# 機械学習で問われる3つの質問

- 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

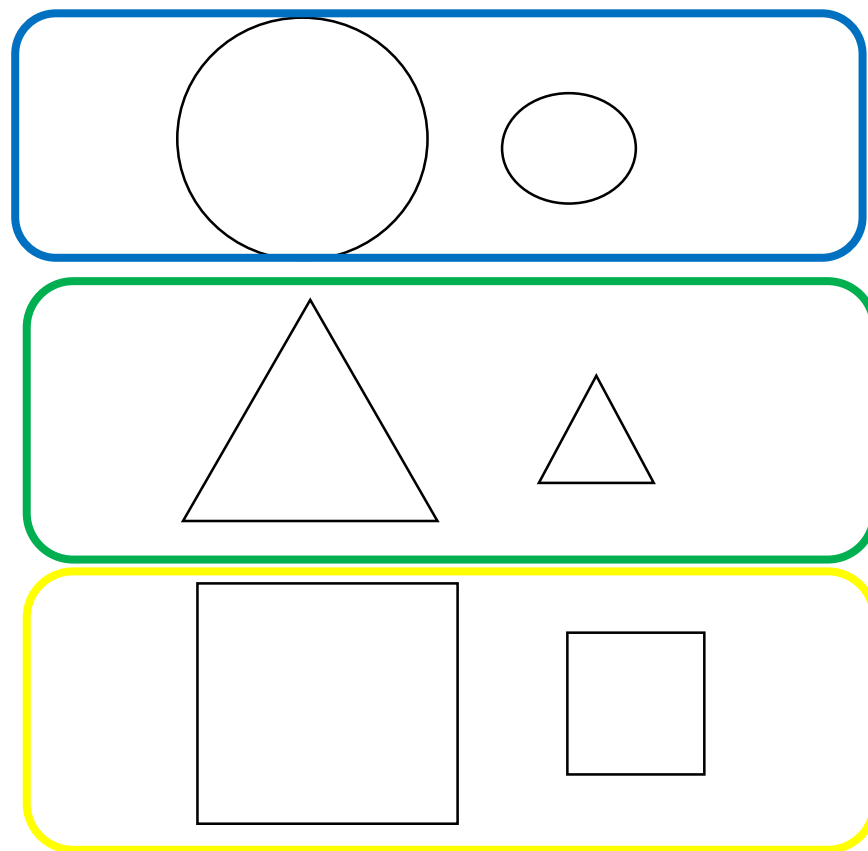
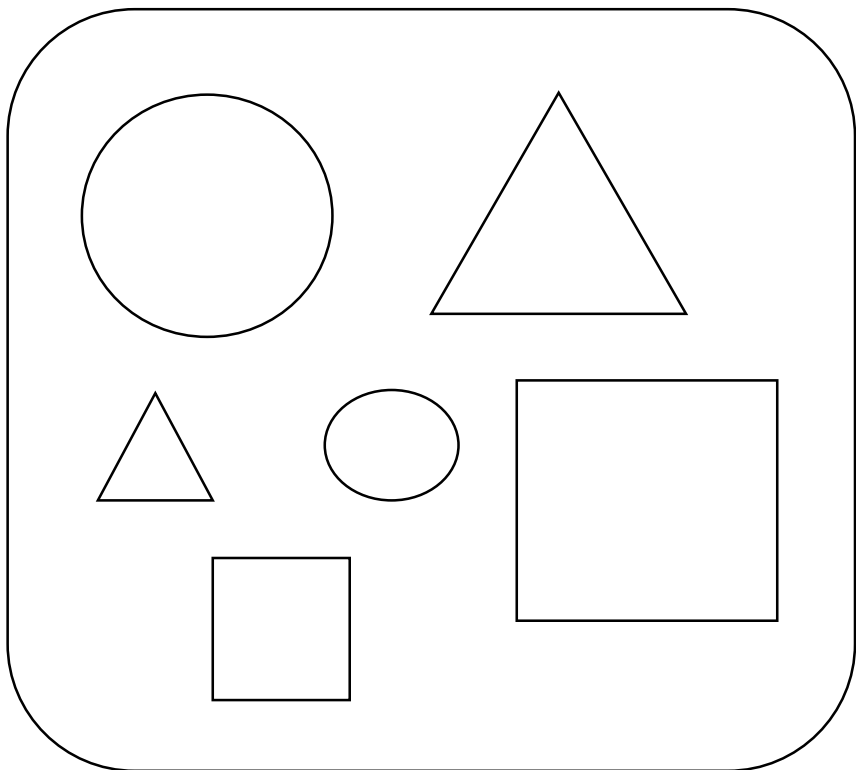
どの視聴者が同じ種類の映画を好むか？



# 機械学習で問われる 3つの質問

- 質問 3

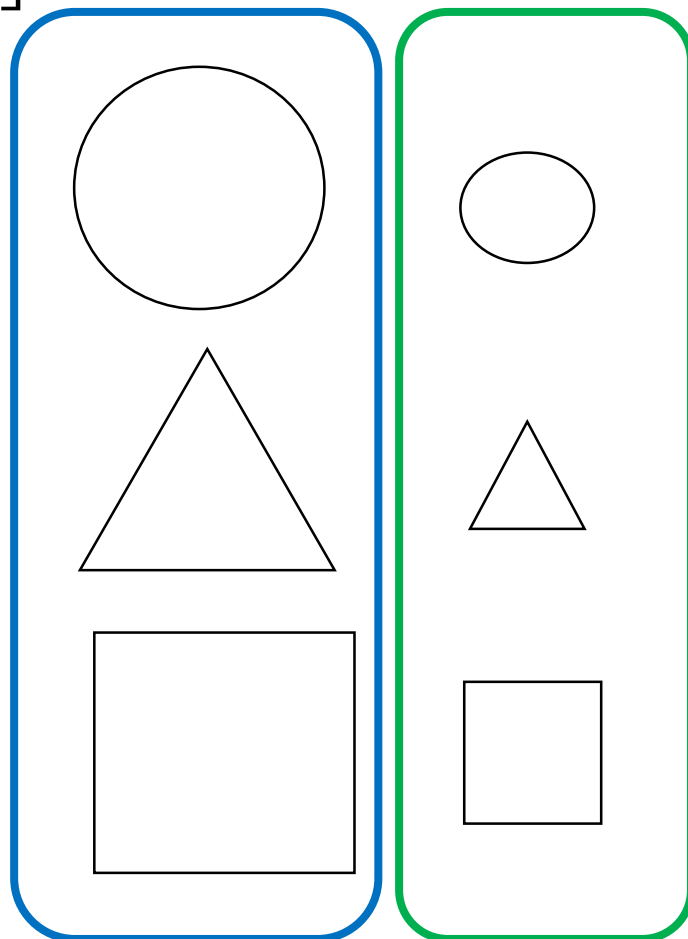
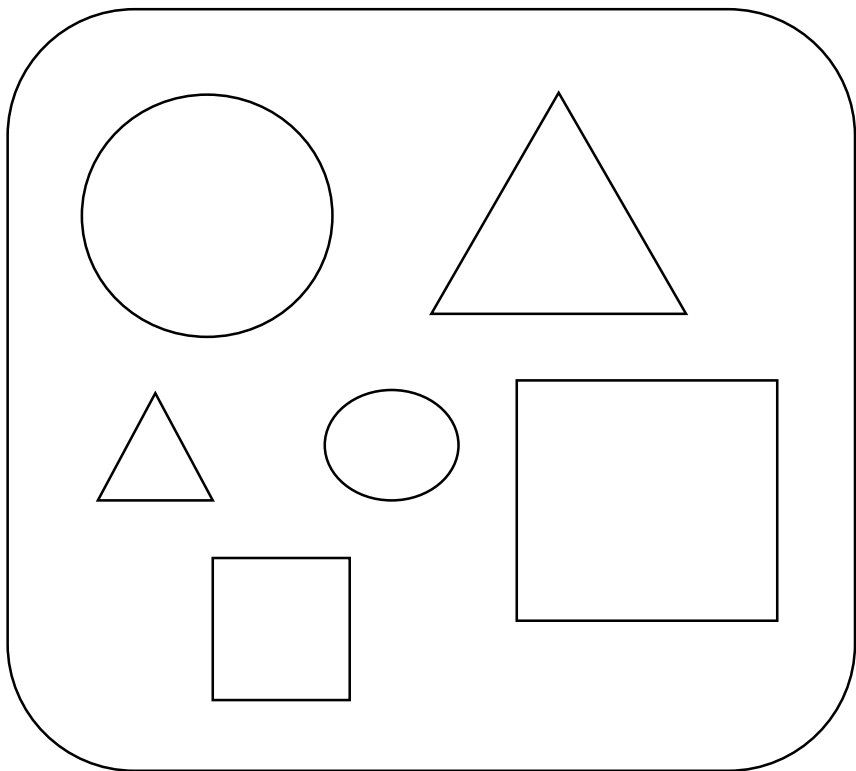
「どのような編成になっているのか？」



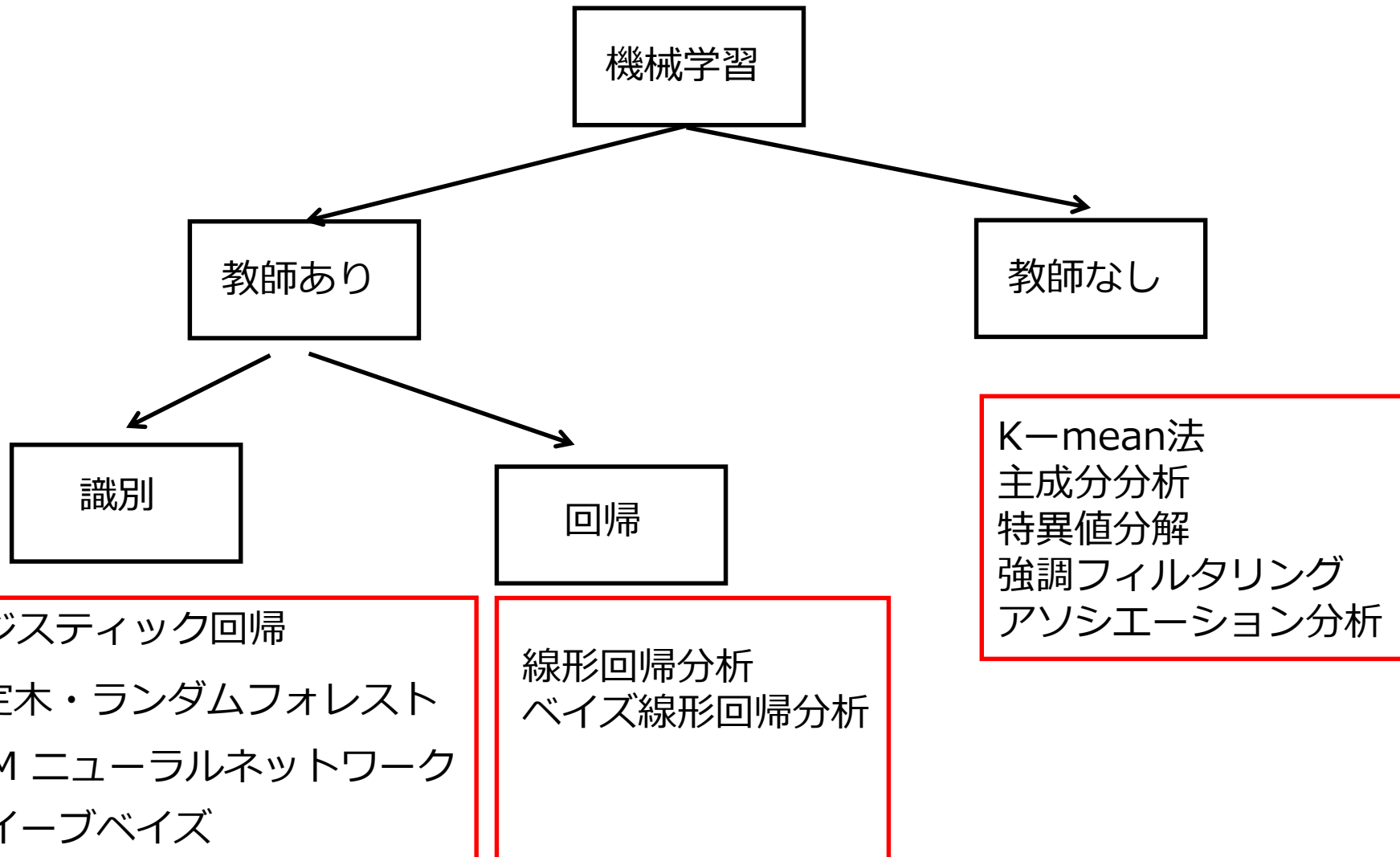
# 機械学習で問われる 3つの質問

- 質問 3

「どのような編成になっているのか？」

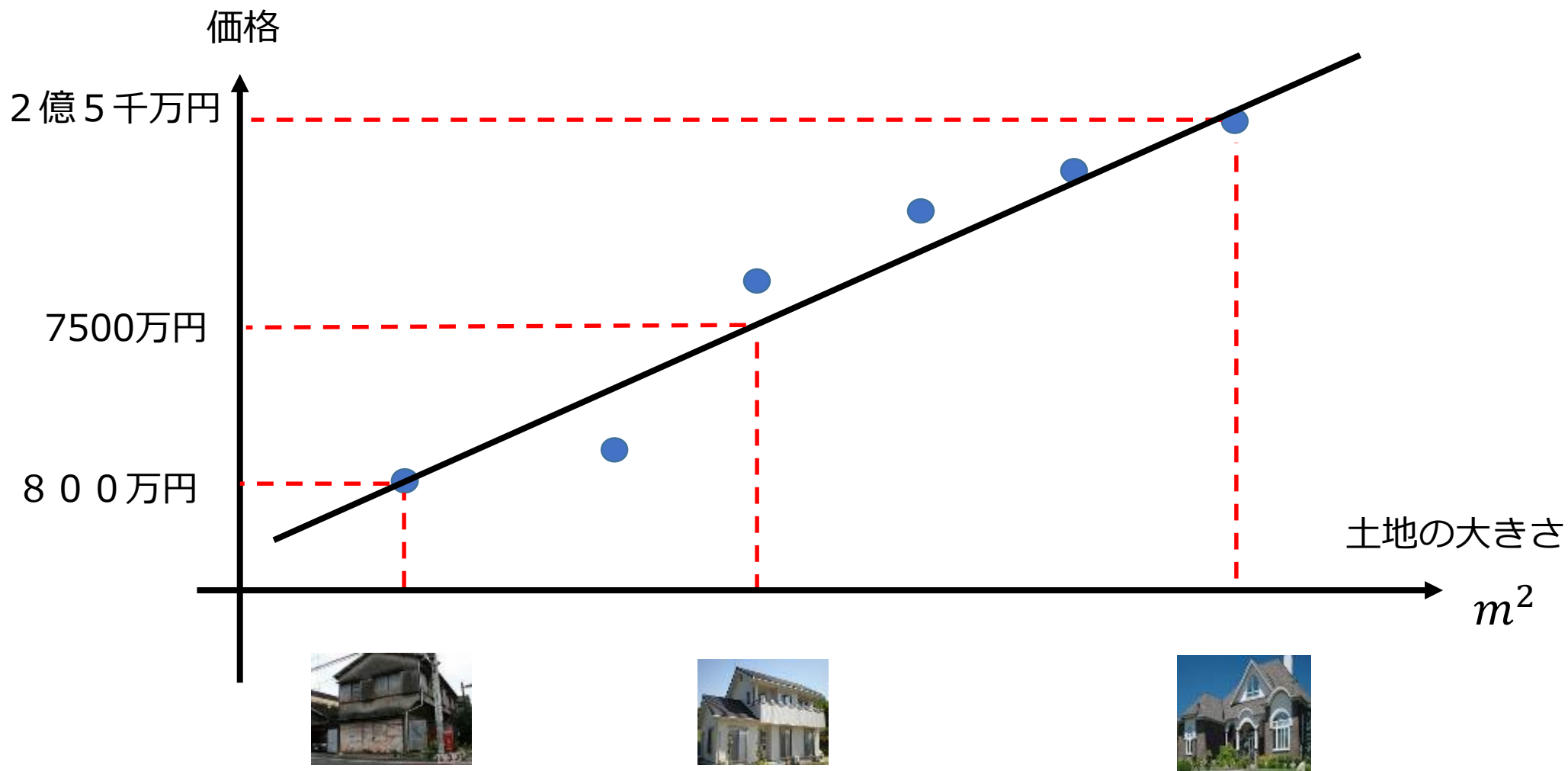


# 代表的な機会学習の手法



## ・ 教師あり ・ 機械学習 ・ 回帰

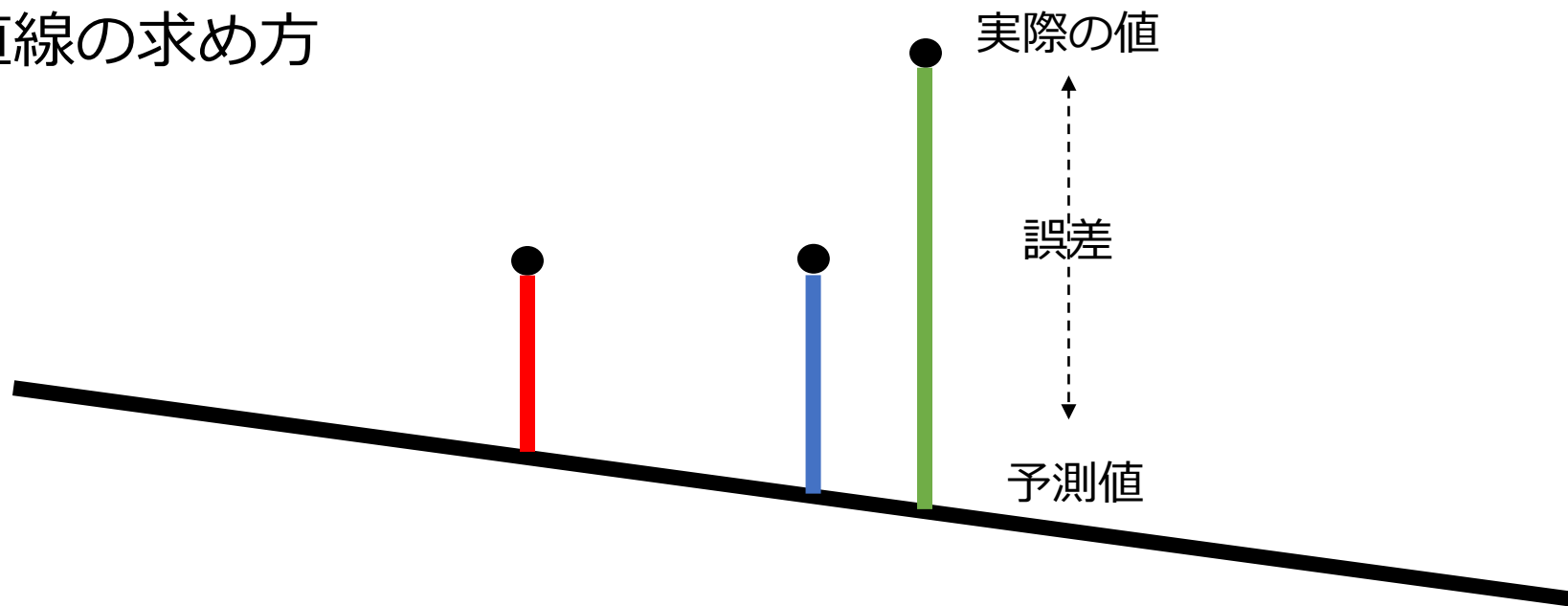
# 回帰分析



価格に影響を与える要因（築年数・広さ・・・）

# 回帰モデル

直線の求め方



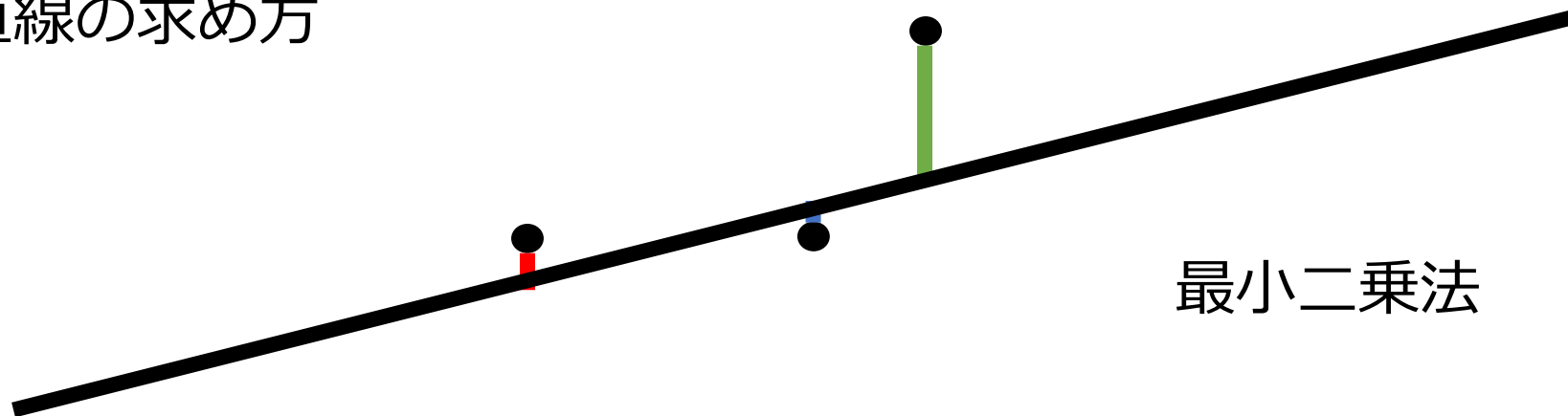
誤差 (Error)





# 回帰モデル

直線の求め方



最小二乗法

誤差 (Error)

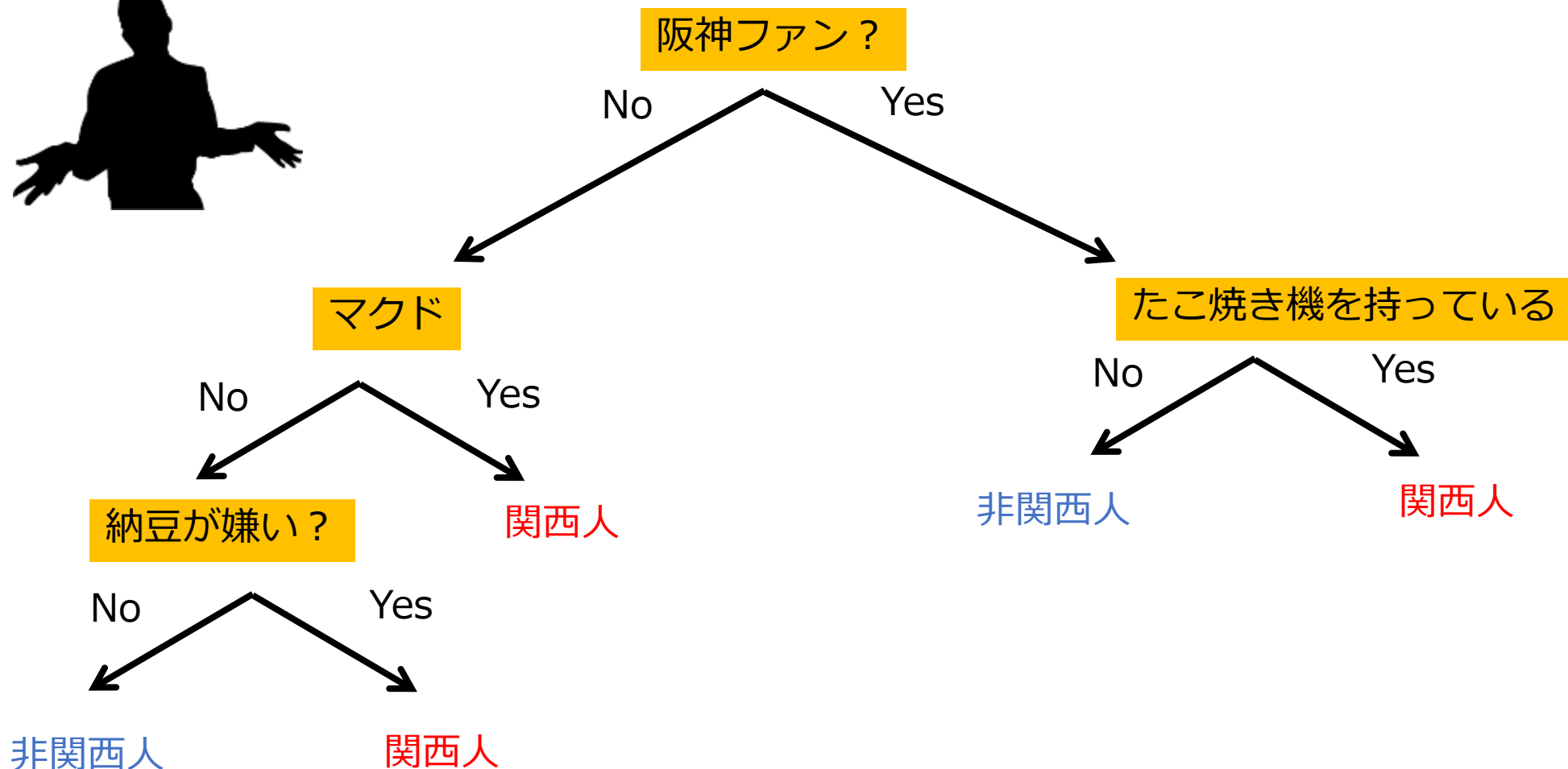


誤差 2 乗が最小の直線

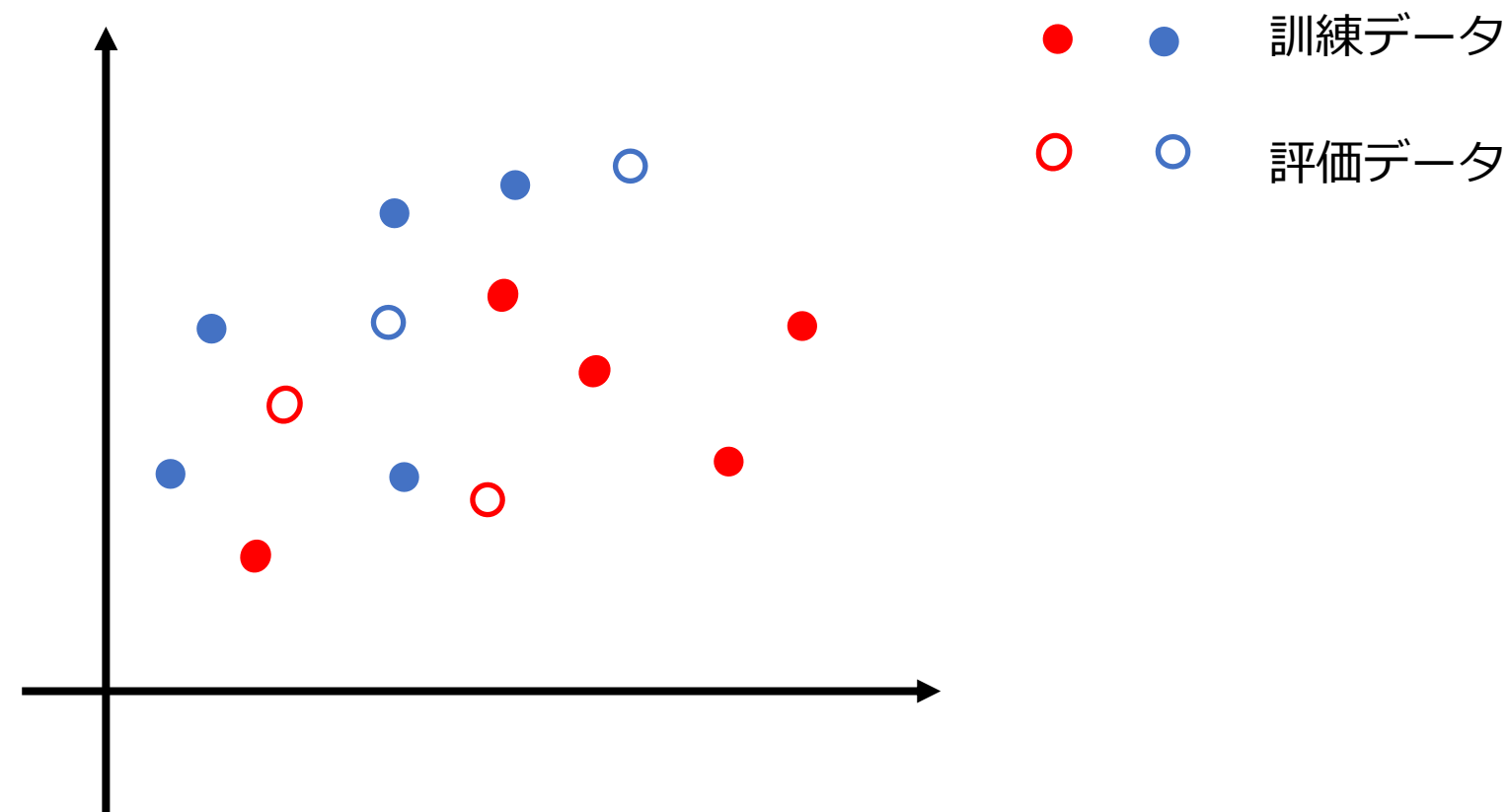
- 
- ・ 教師あり ・ 機械学習 ・ 識別

# 識別能力の高い質問による分類

関西人なのか？



# どっちのモデルを選択すべきか？

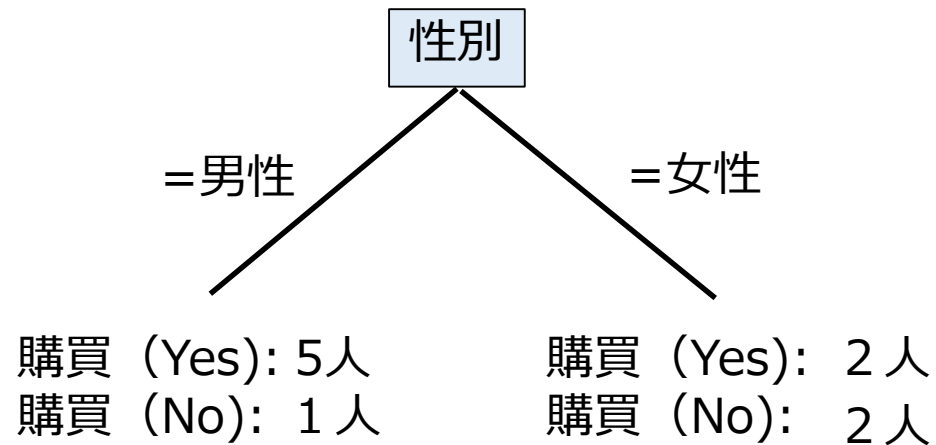


# 問題 決定木による識別

ある商品の顧客の属性として、性別、年齢、見た広告の種類、およびその商品の過去の購買履歴があたえられたとして、顧客が購買するかしないかに分類する決定木を考えよ。

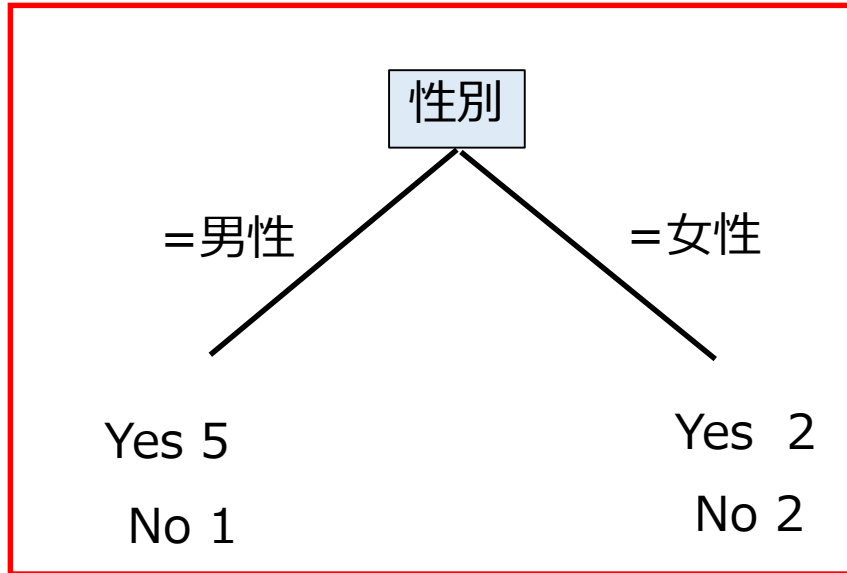
ID	性別	年齢	広告	購買歴	購買
A	男性	10代	TV	無	No
B	女性	10代	TV	無	No
C	女性	50代	ネット	無	No
D	男性	30代	TV	無	Yes
E	男性	50代	電車	有	Yes
F	男性	50代	ネット	無	Yes
G	女性	30代	電車	有	Yes
H	男性	10代	電車	有	Yes
I	男性	50代	ネット	有	Yes
J	女性	10代	ネット	有	Yes

# 性別による分類



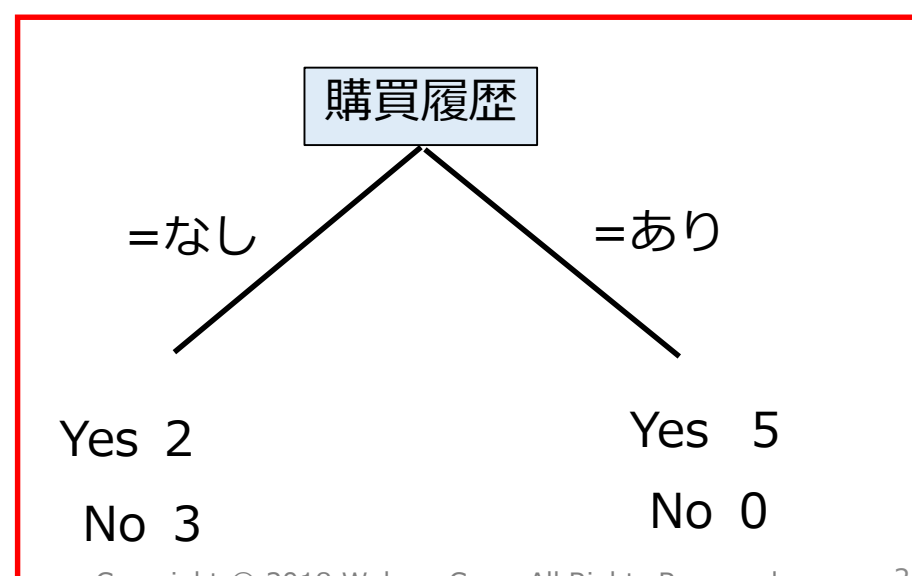
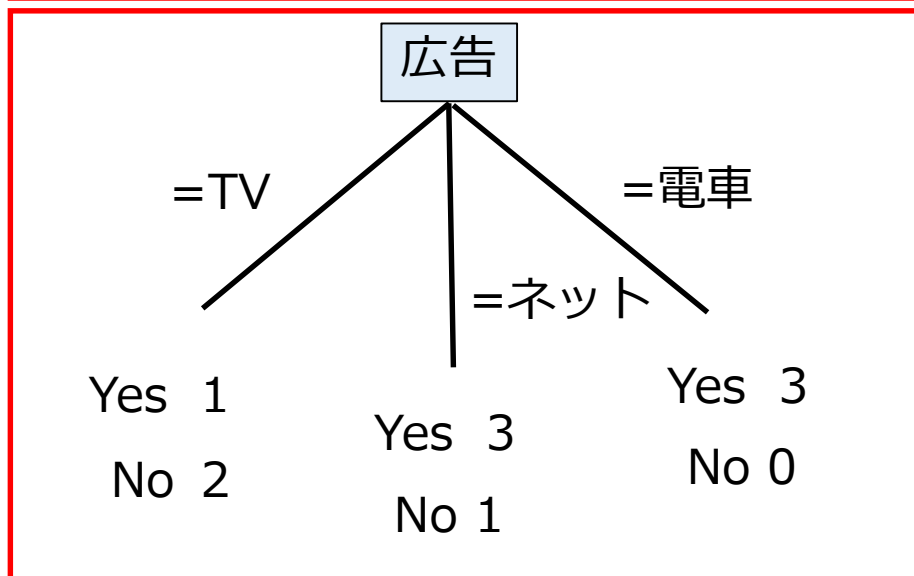
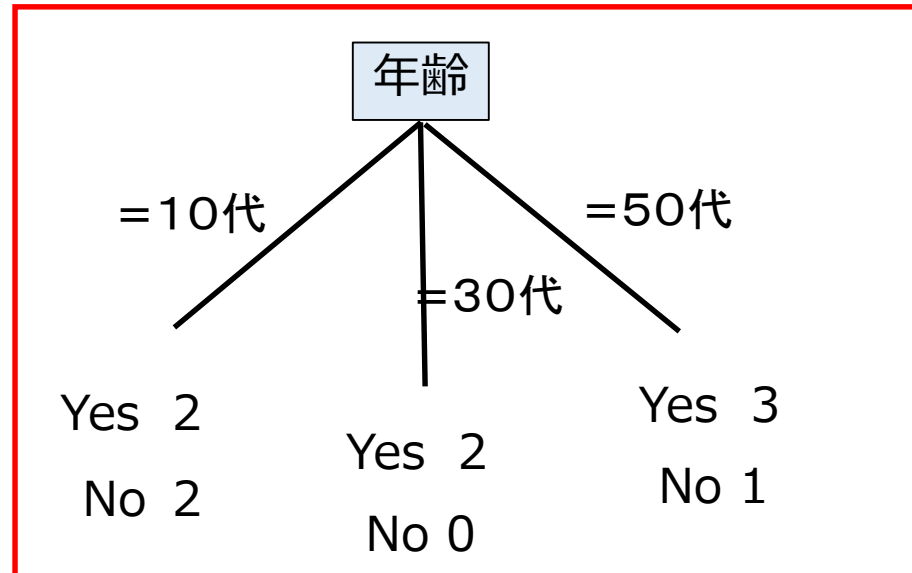
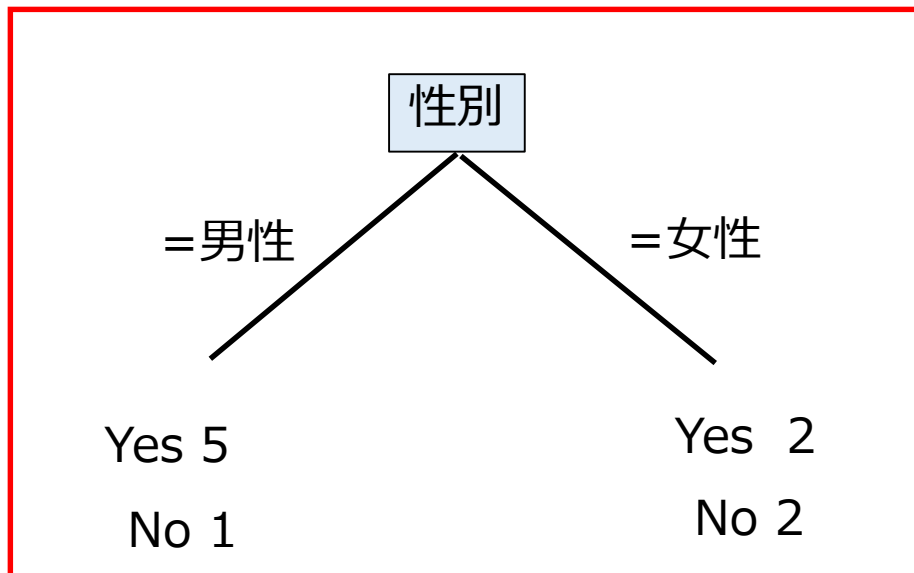
性別	購買
女性	No
女性	No

# 問題：どの変数で分類すべきか？



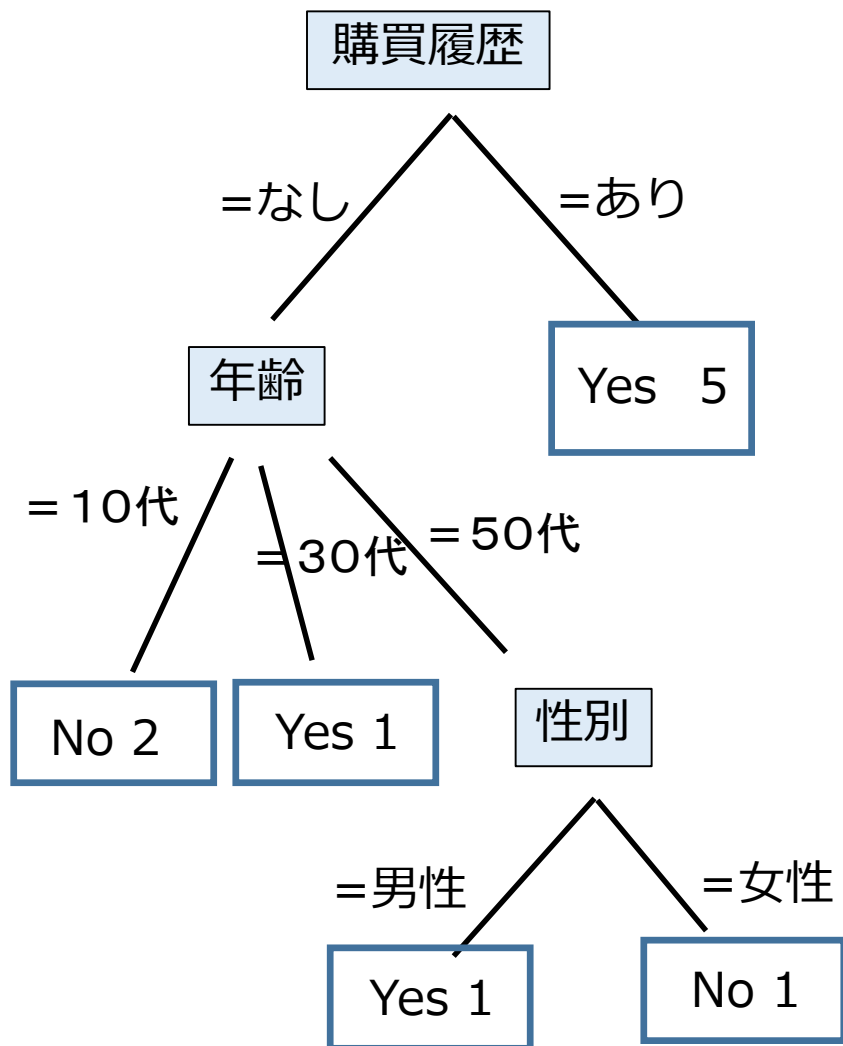
ID	性別	年齢	広告	購買歴	購買
A	男性	10代	TV	無	No
B	女性	10代	TV	無	No
C	女性	50代	ネット	無	No
D	男性	30代	TV	無	Yes
E	男性	50代	電車	有	Yes
F	男性	50代	ネット	無	Yes
G	女性	30代	電車	有	Yes
H	男性	10代	電車	有	Yes
I	男性	50代	ネット	有	Yes
J	女性	10代	ネット	有	Yes

# 問題：どの変数で分類すべきか？





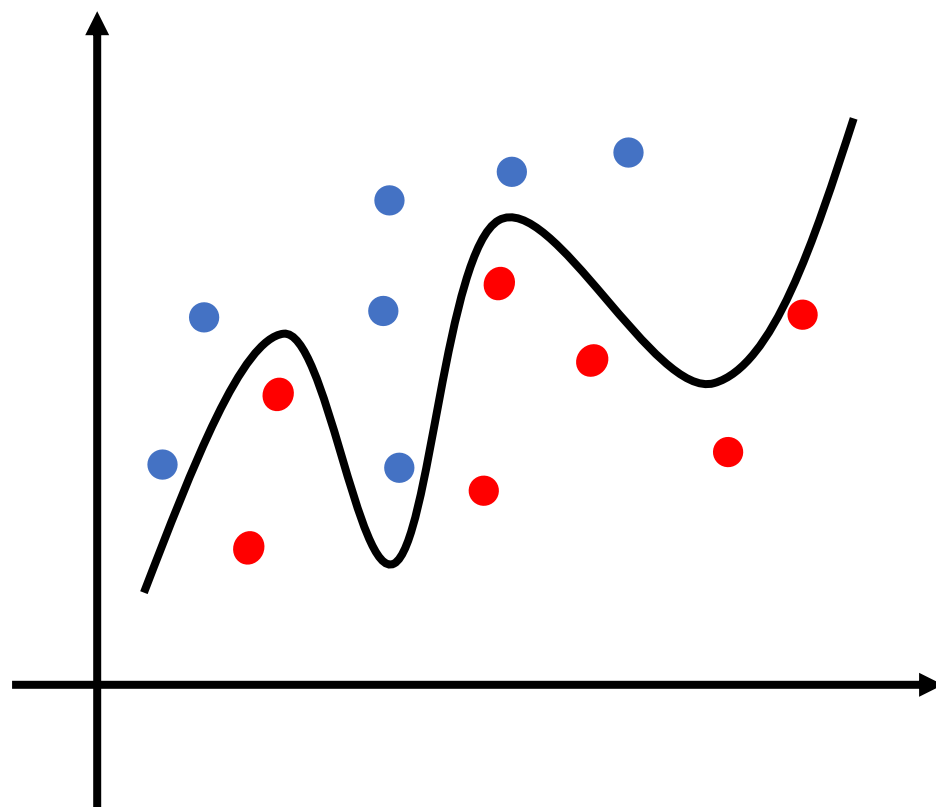
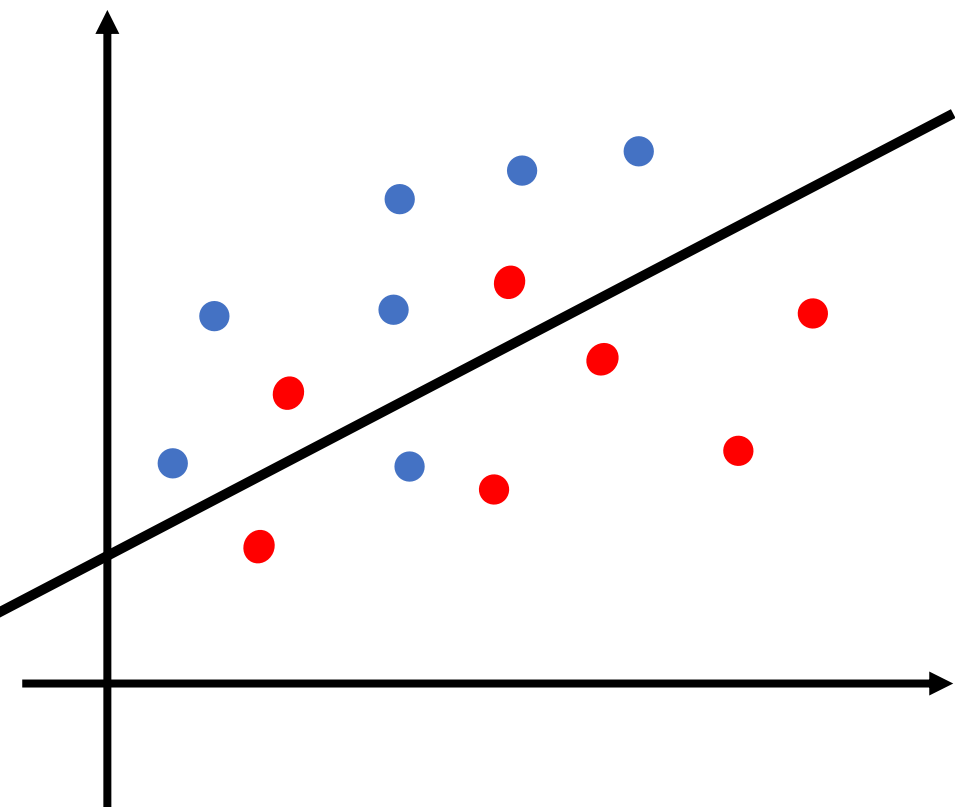
# 決定木による識別



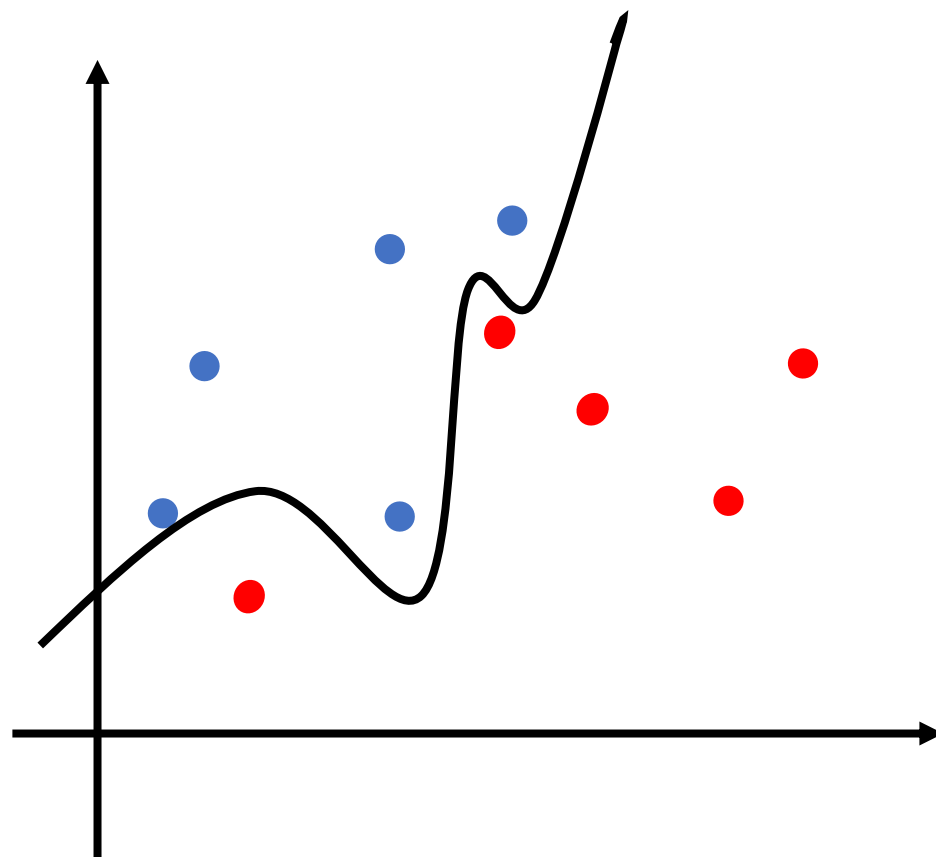
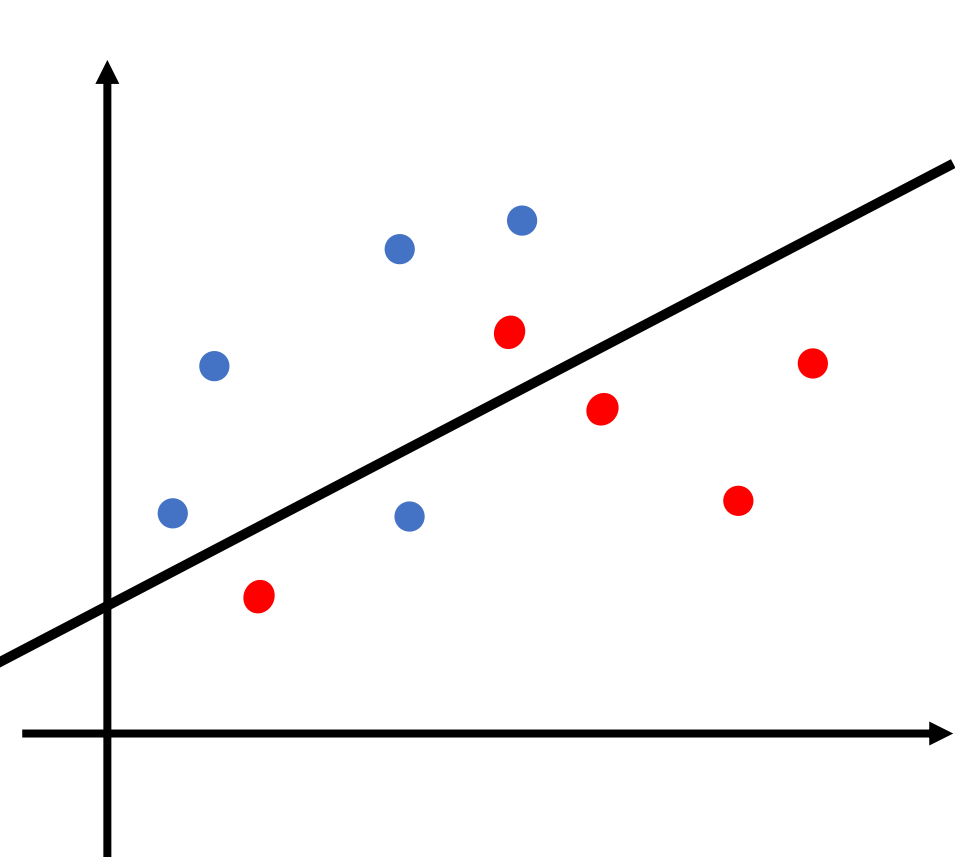
# ランダムフォレスト

- 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)

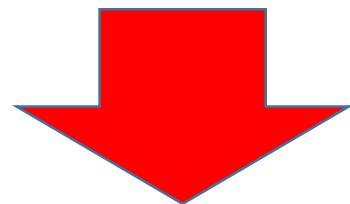
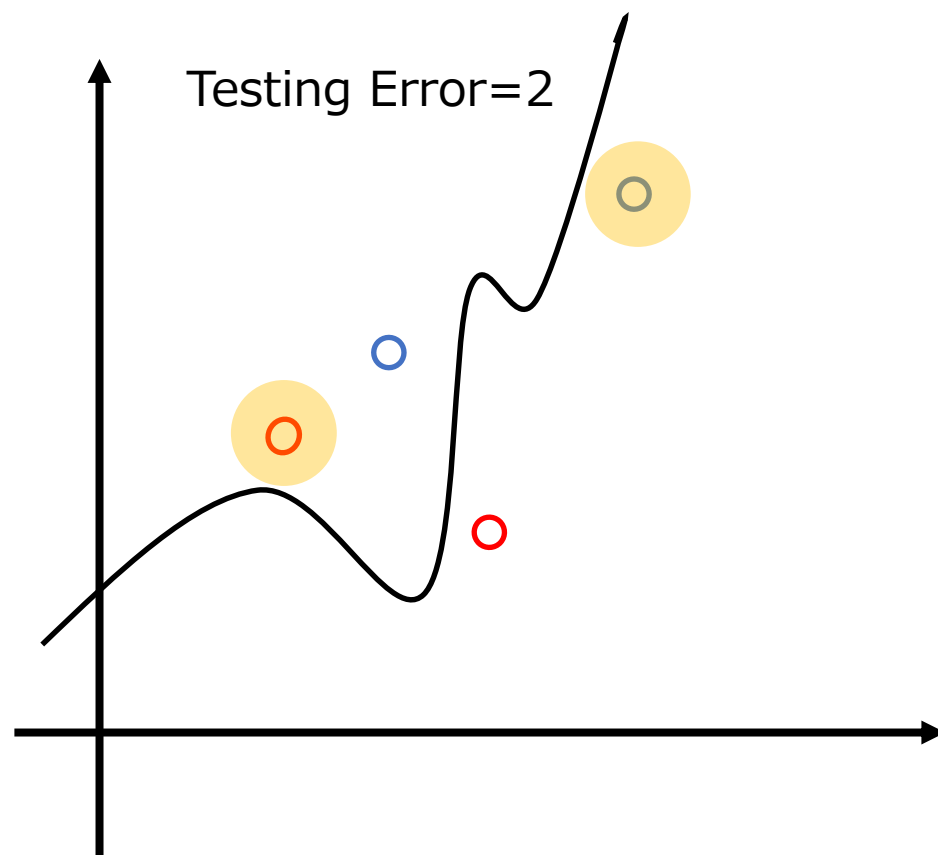
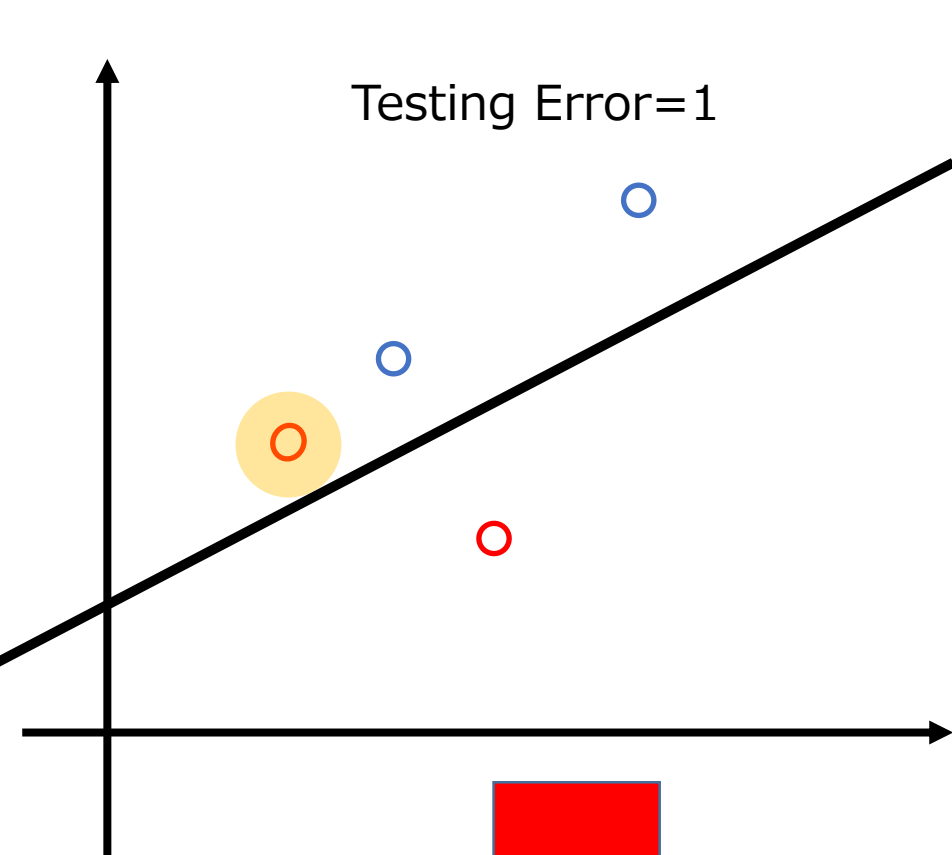
# どっちのモデルを選択すべきか？



# どっちのモデルを選択すべきか？

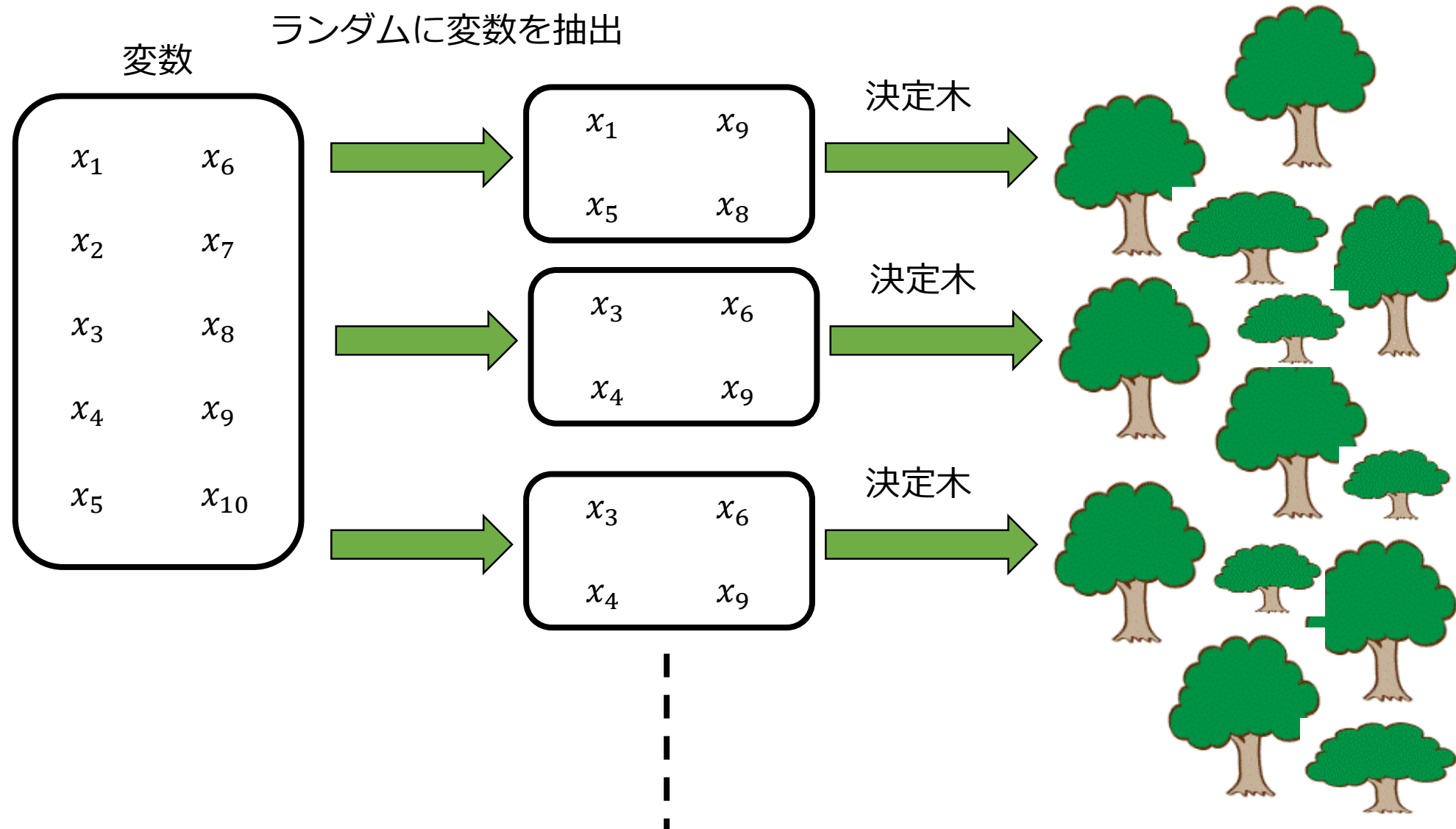


# どっちのモデルを選択すべきか？



良いモデル

# ランダムフォレスト



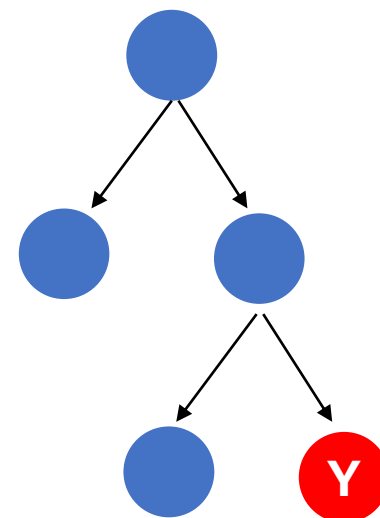
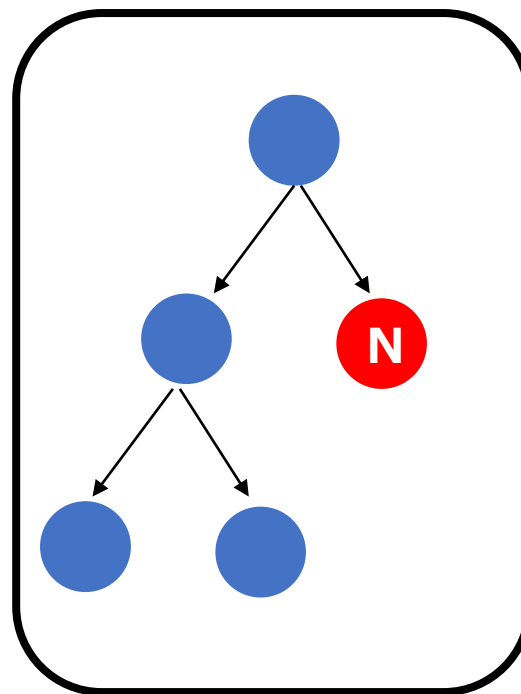
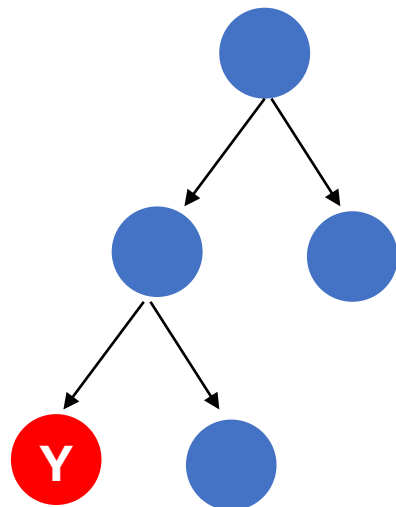
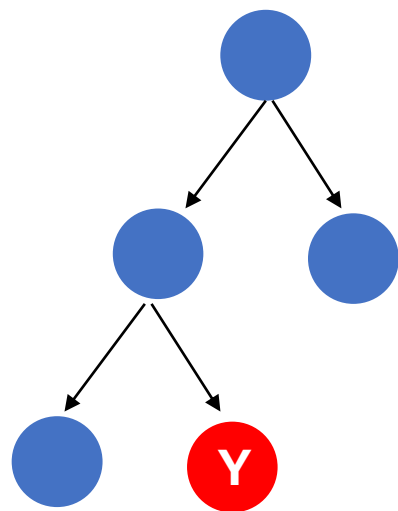
# ランダムフォレスト

購入するかどうか？



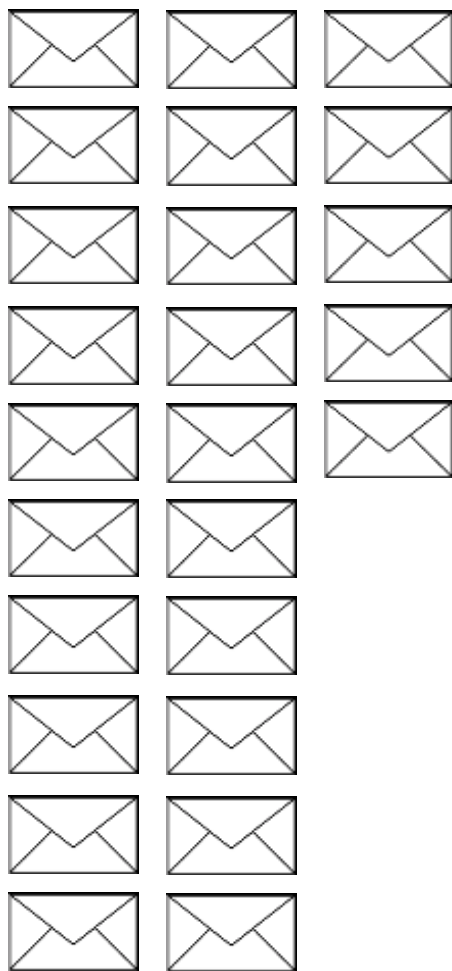
「購入する」

間違い

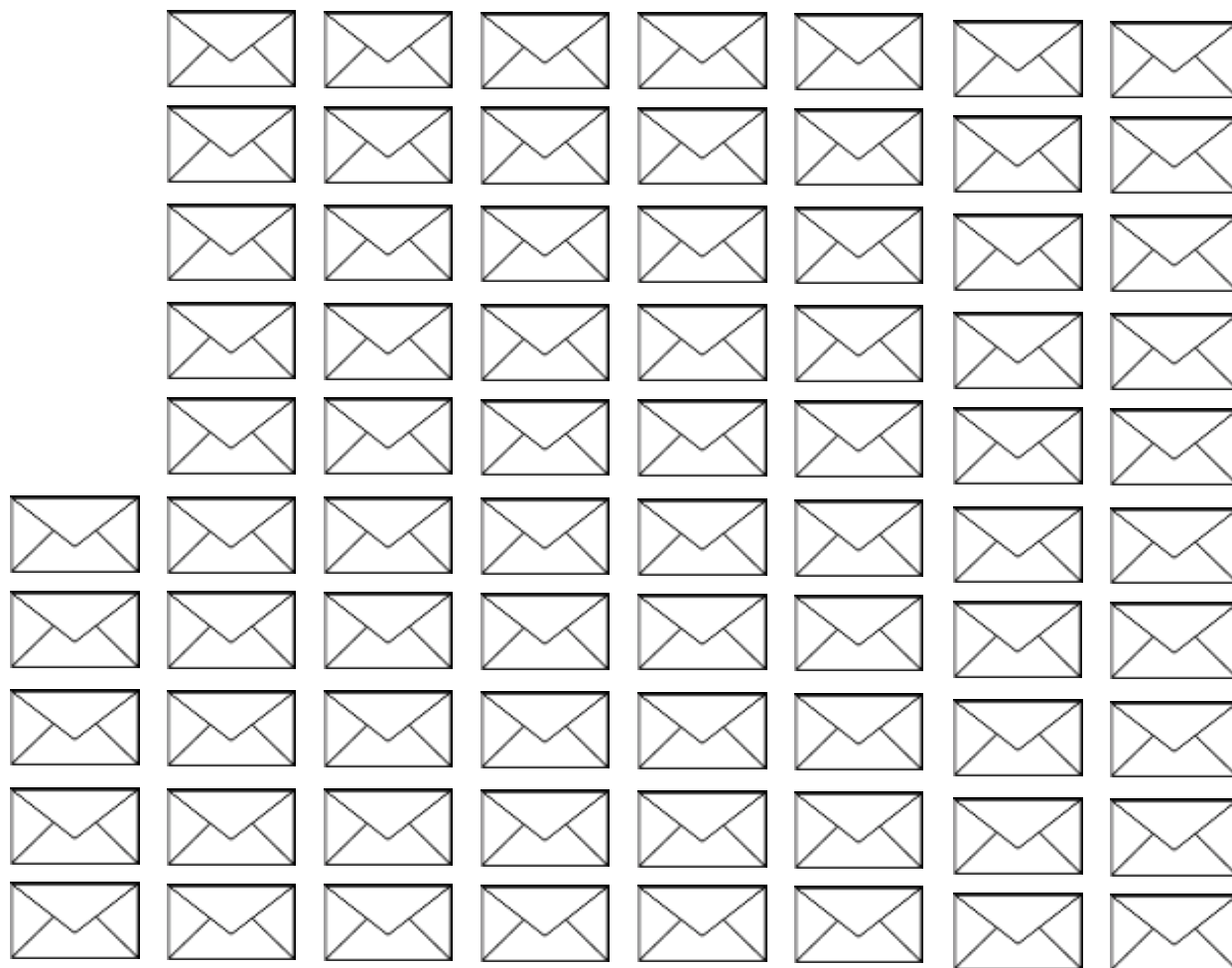


多数決の原理

# スパムメールの判別



**スパム(25)**

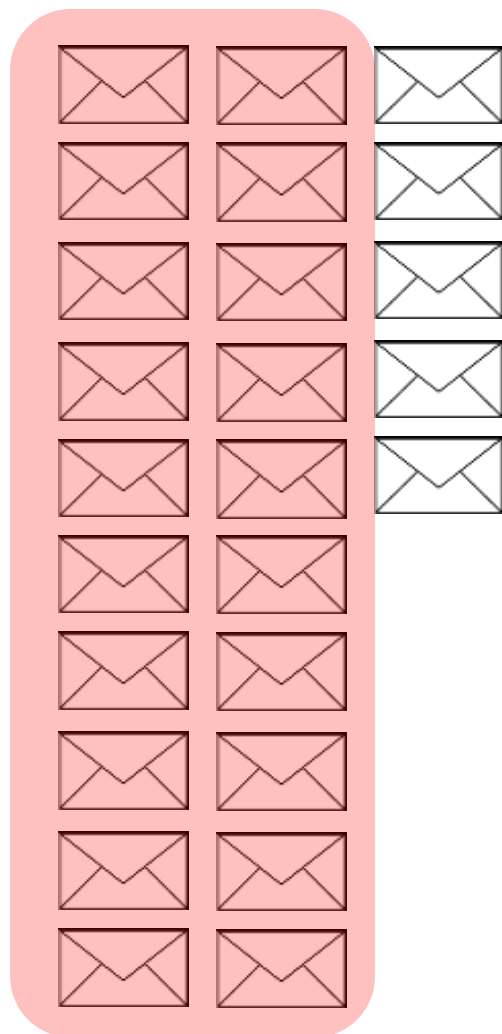


**通常メール(75)**

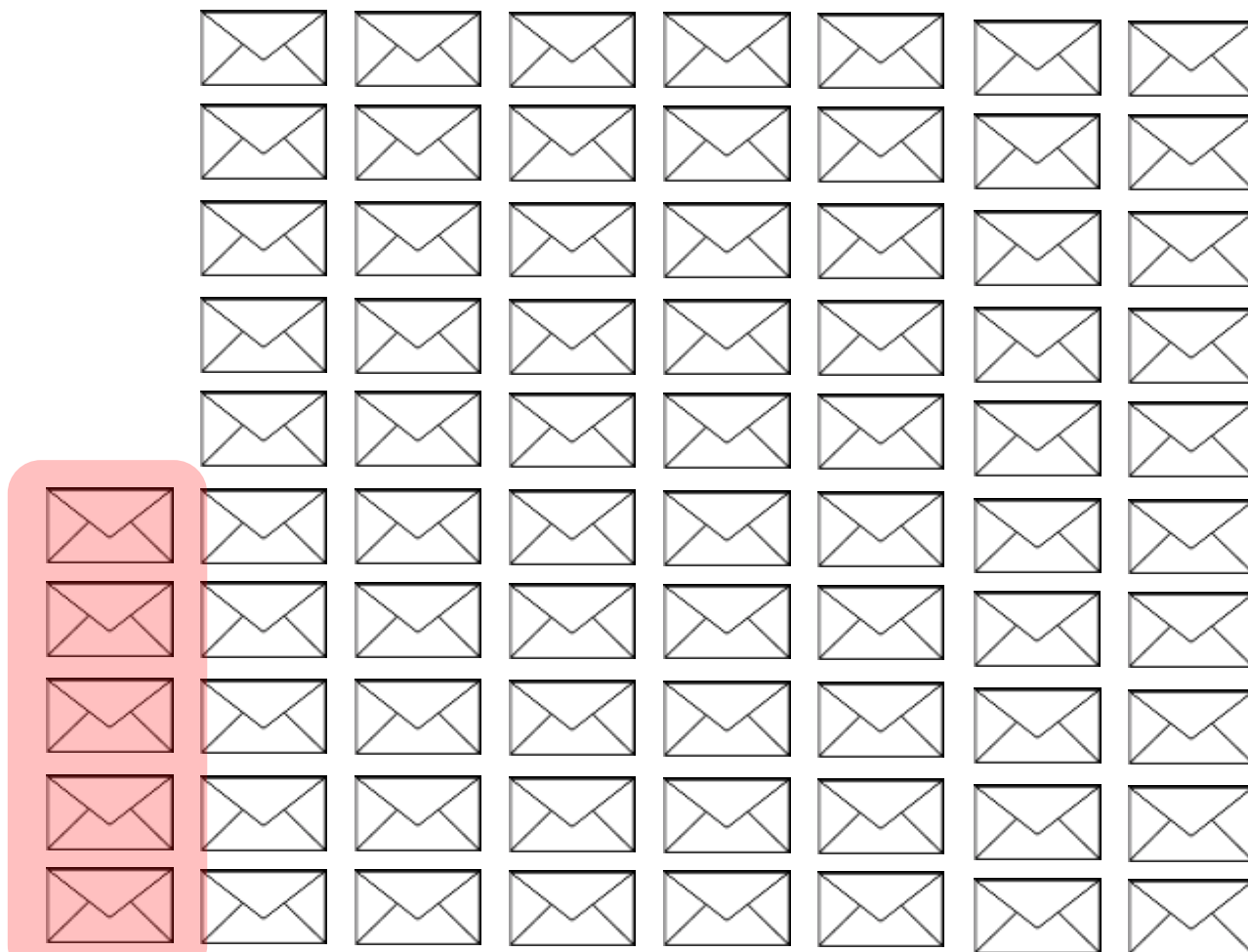


# スパムメールの判別

「出会い」

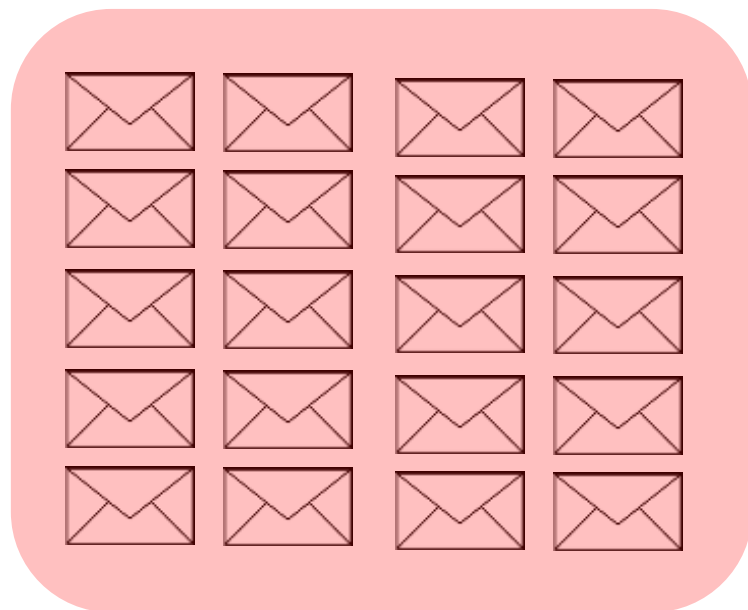


スパム(20/25)

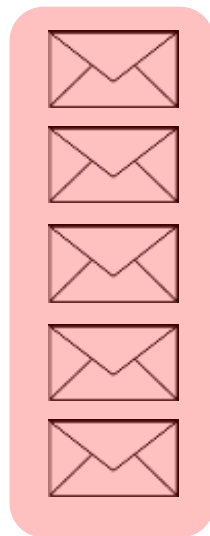


通常メール(5/75)

# スパムメールの判別



スパム(20)

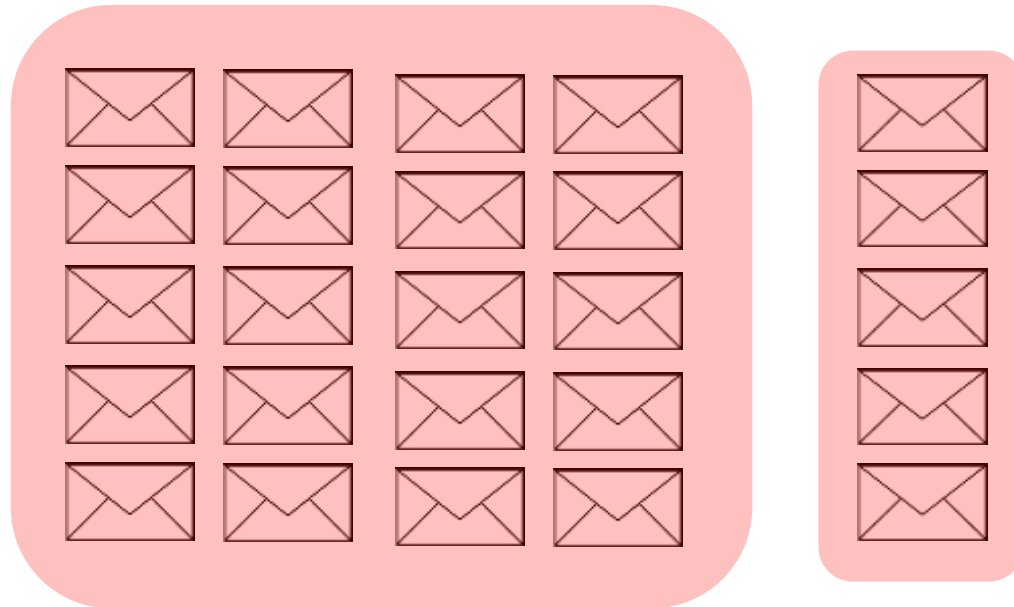


通常メール(5)

「出会い」

問題  
データによると、  
「出会い」という単語が含まれたメール  
がスパムメールである確率は？

# スパムメールの判別



スパム(80%)

通常メール(20%)

「出会い」

## 問題

データによると、  
「出会い」という単語が含まれたメールがスパムメールである確率は？

## 結論

もしメールに「出会い」という単語が含まれるとき、そのメールがスパムである確率は80%

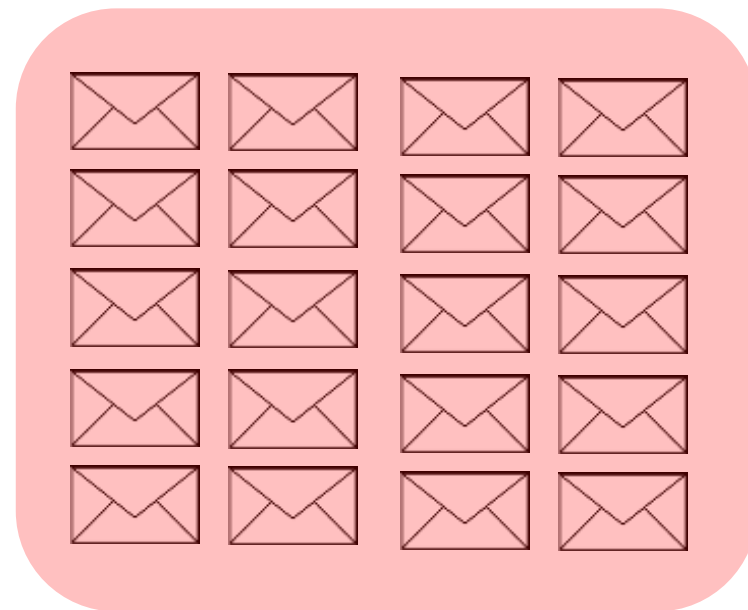
# スパムメールの判別

「出会い」 → 80%

「安い」 → 95%

「 題目がない」 → 70%

ナイーブベイズ



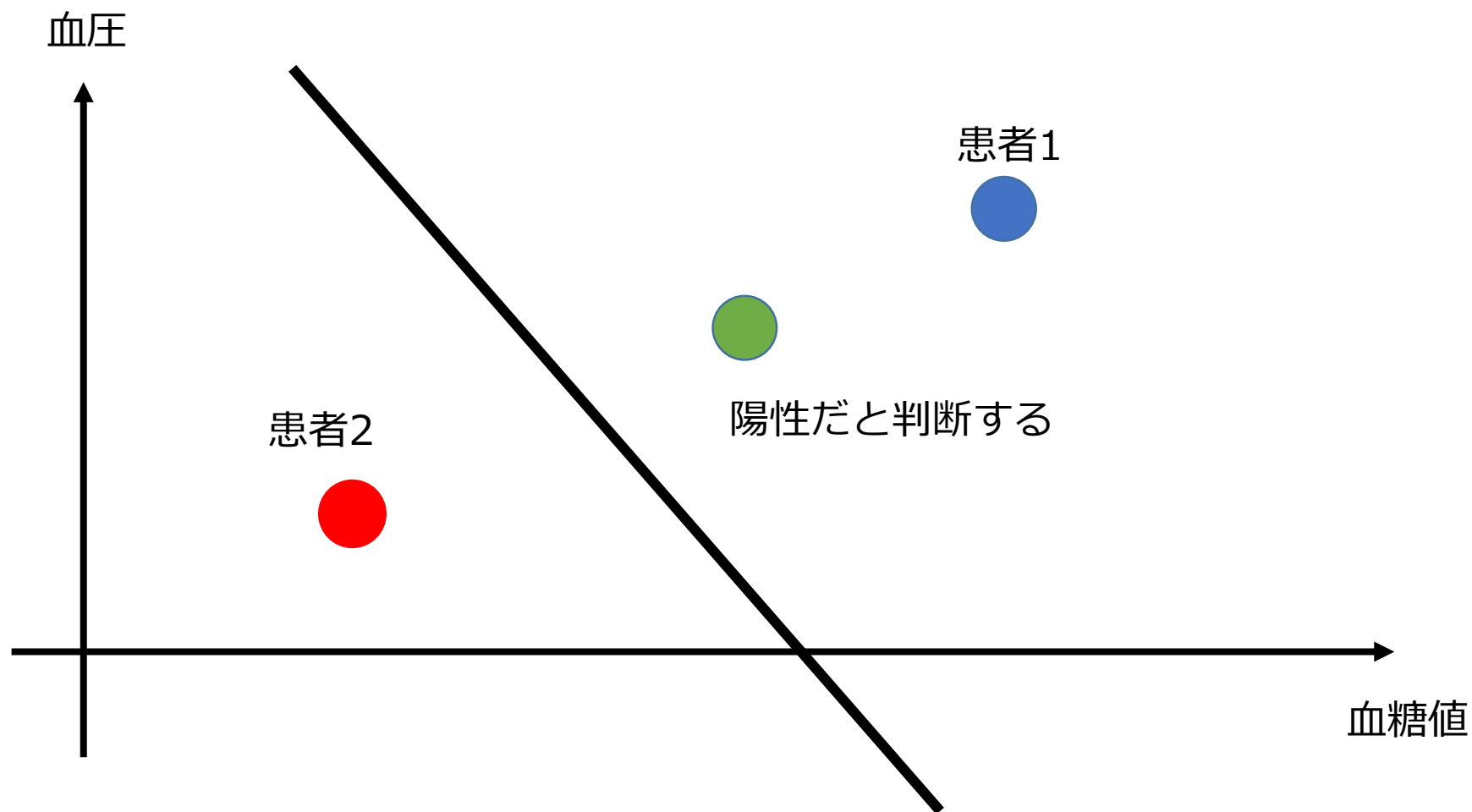
スパム

# 陽性・陰性？

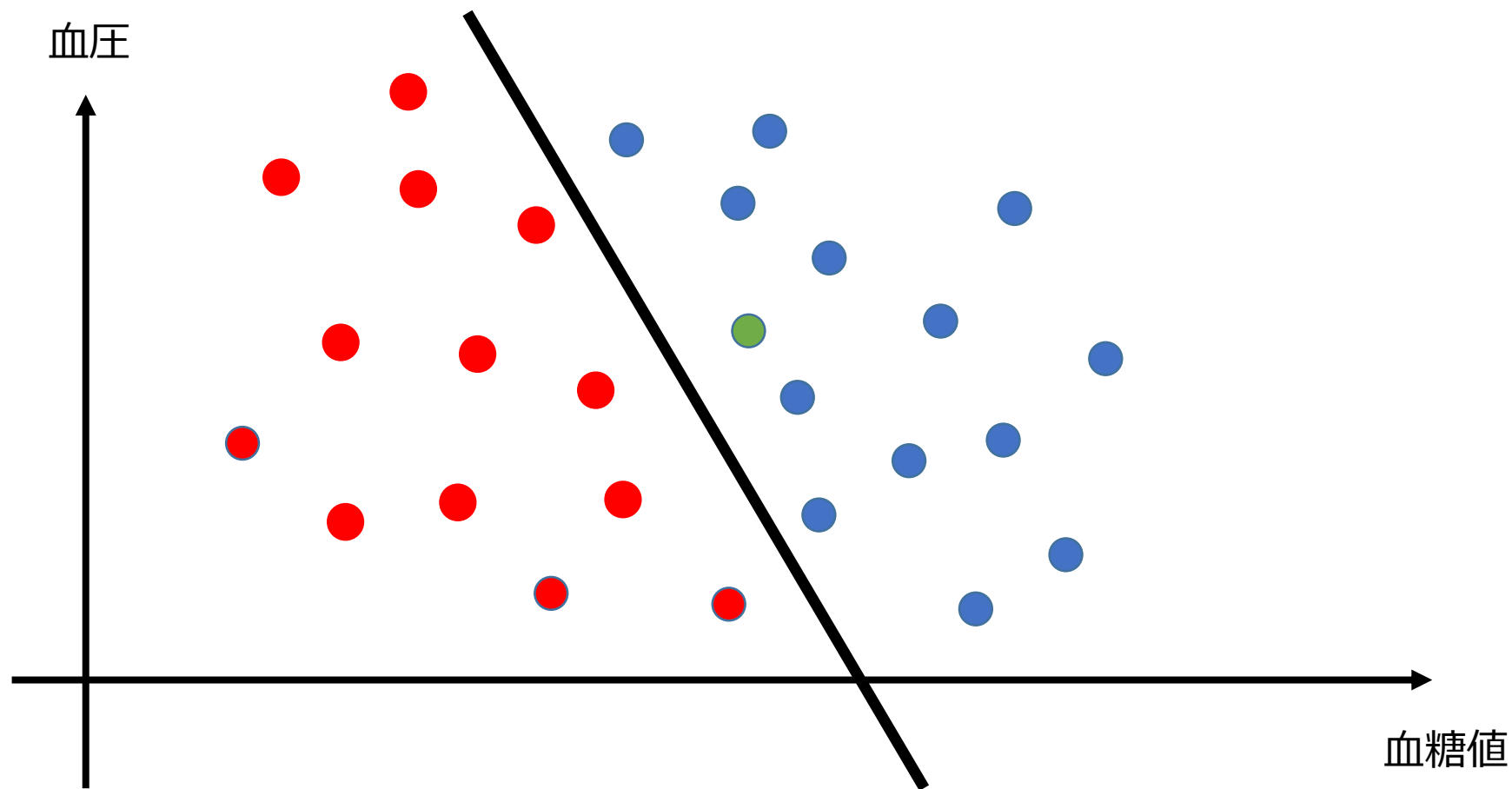


患者ID	血圧	血糖値	
1	120	200	陽性
2	80	130	陰性
3	110	190	?

# 陽性・陰性を識別する

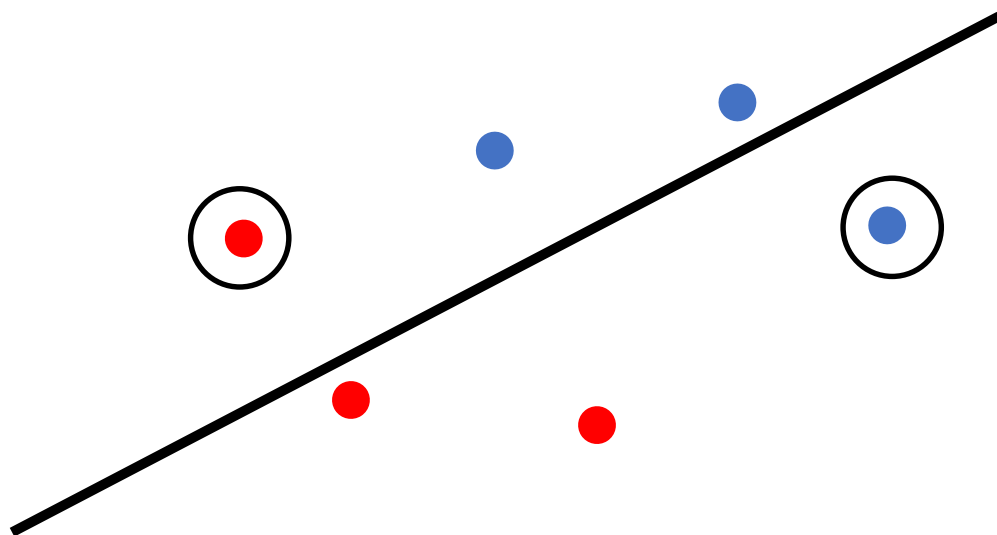


# 陽性・陰性を識別する



ロジスティック回帰

# ロジスティック曲線の求め方

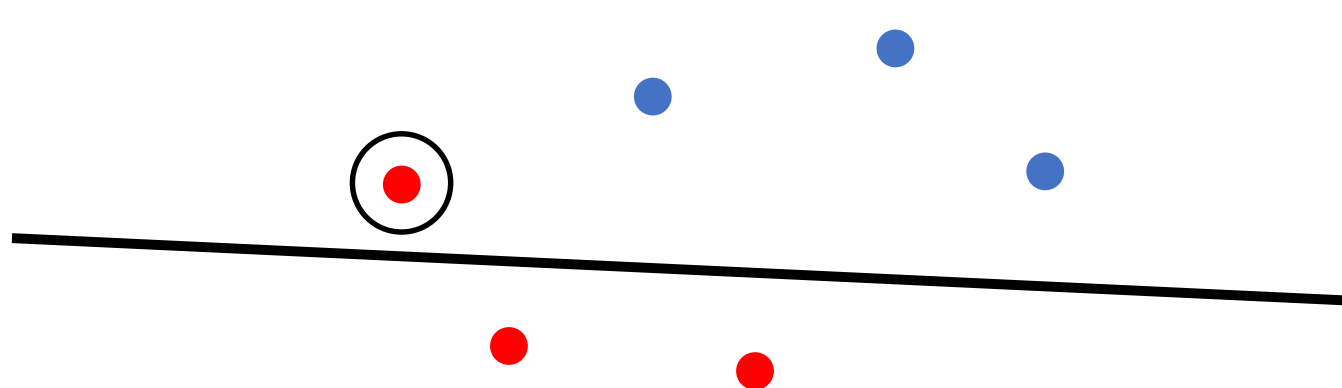


間違えの数

2



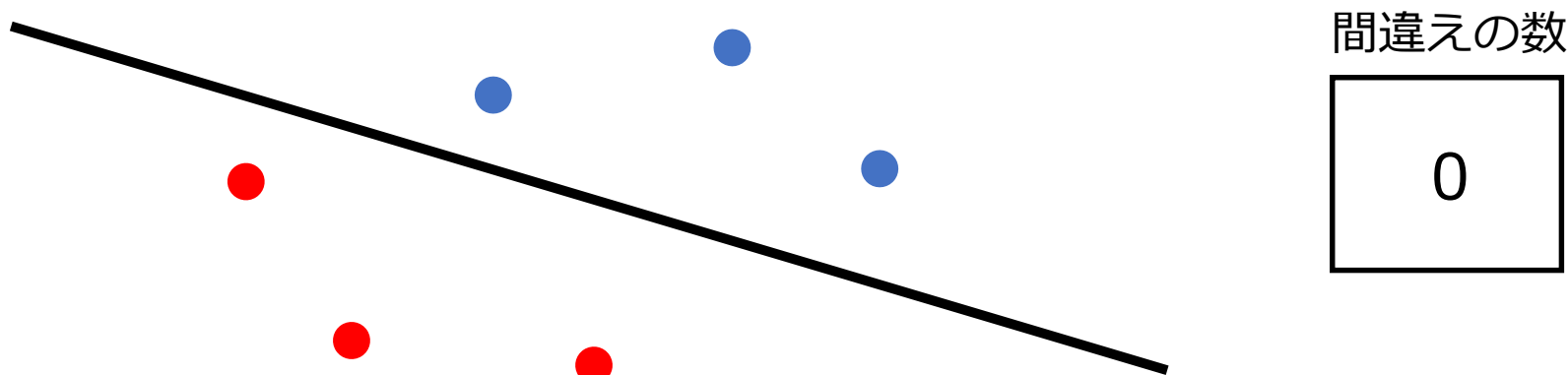
# ロジスティック曲線の求め方



間違えの数

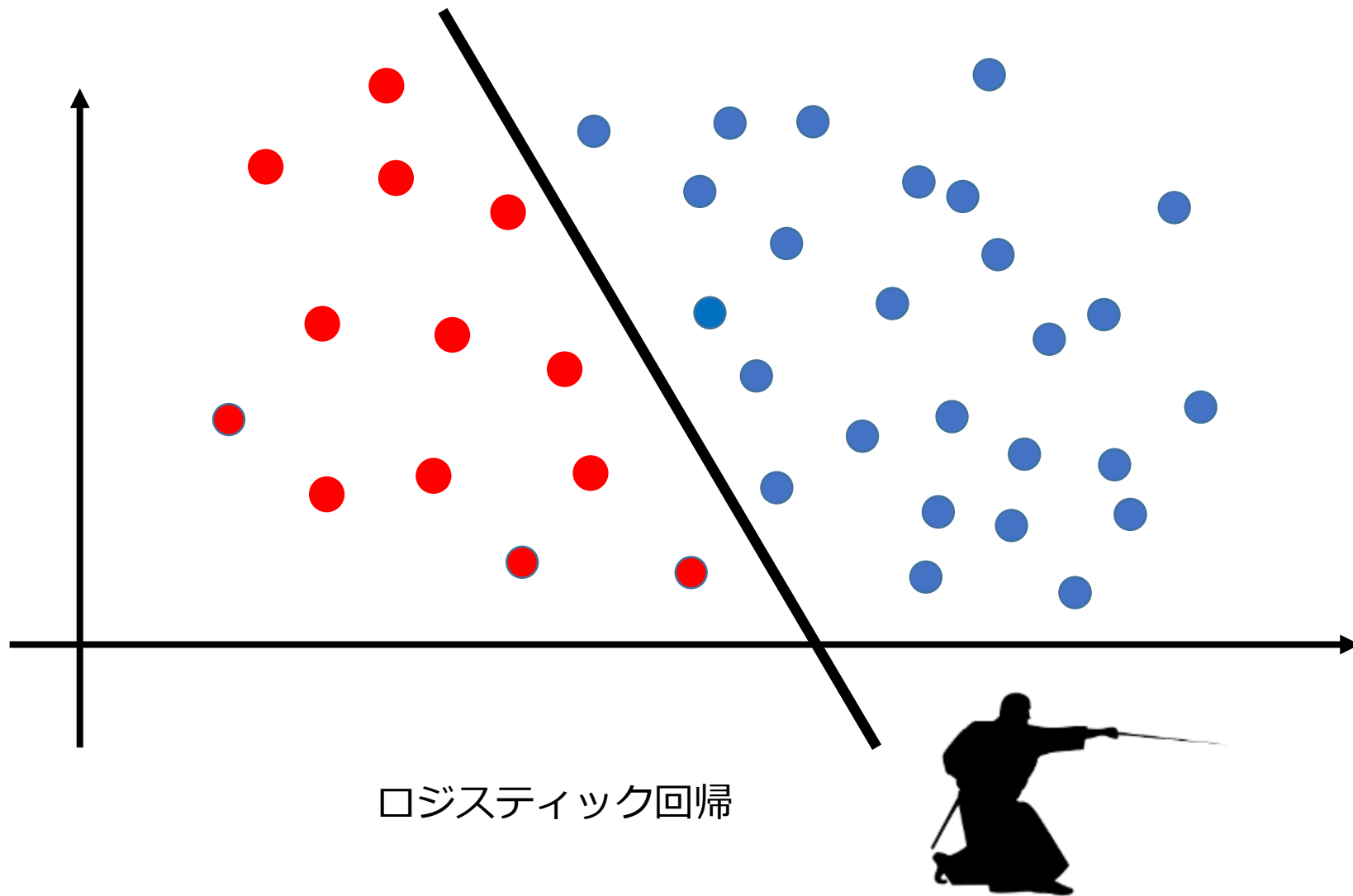
1

# ロジスティック曲線の求め方

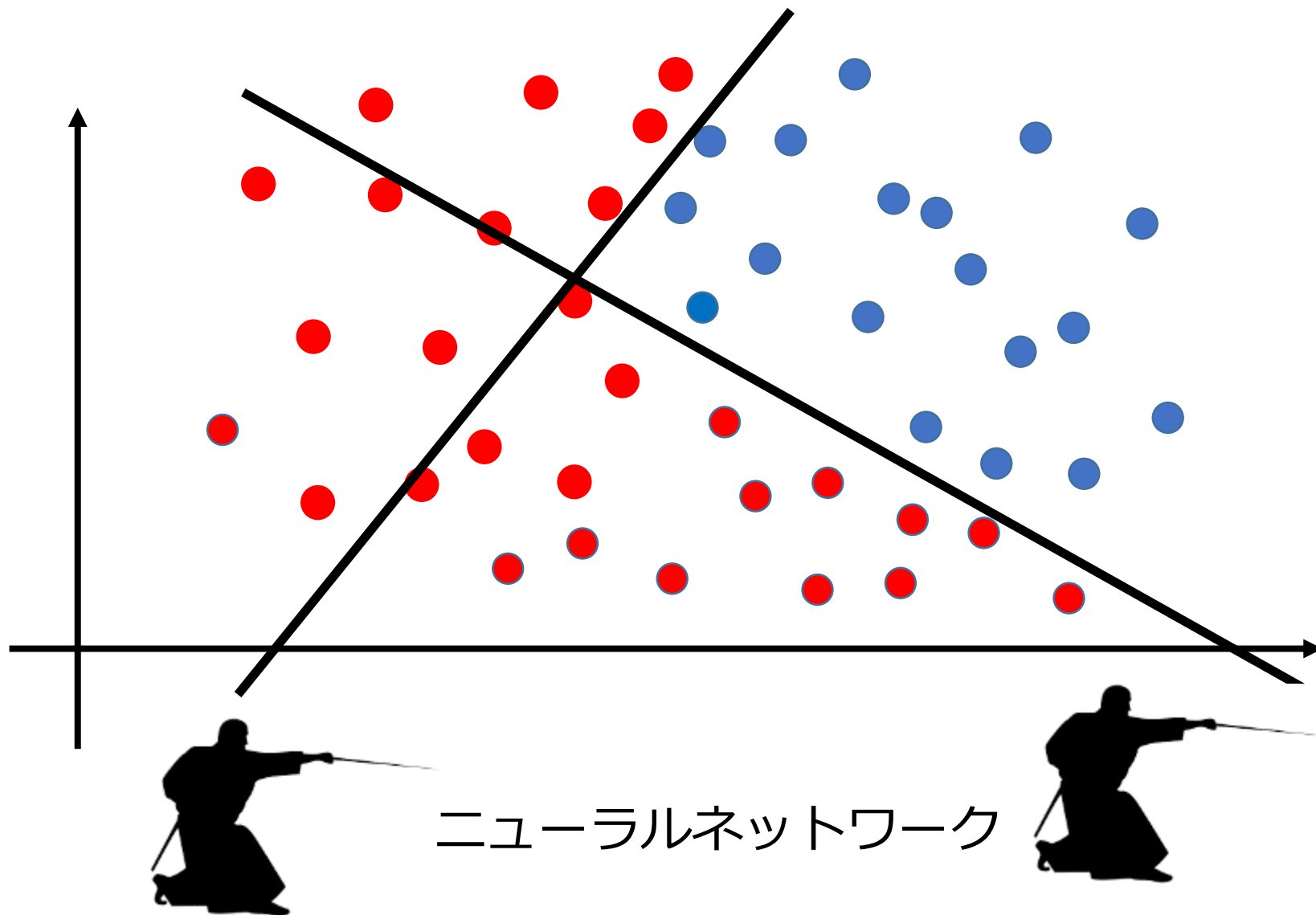


最尤法によってロジスティック曲線を求める

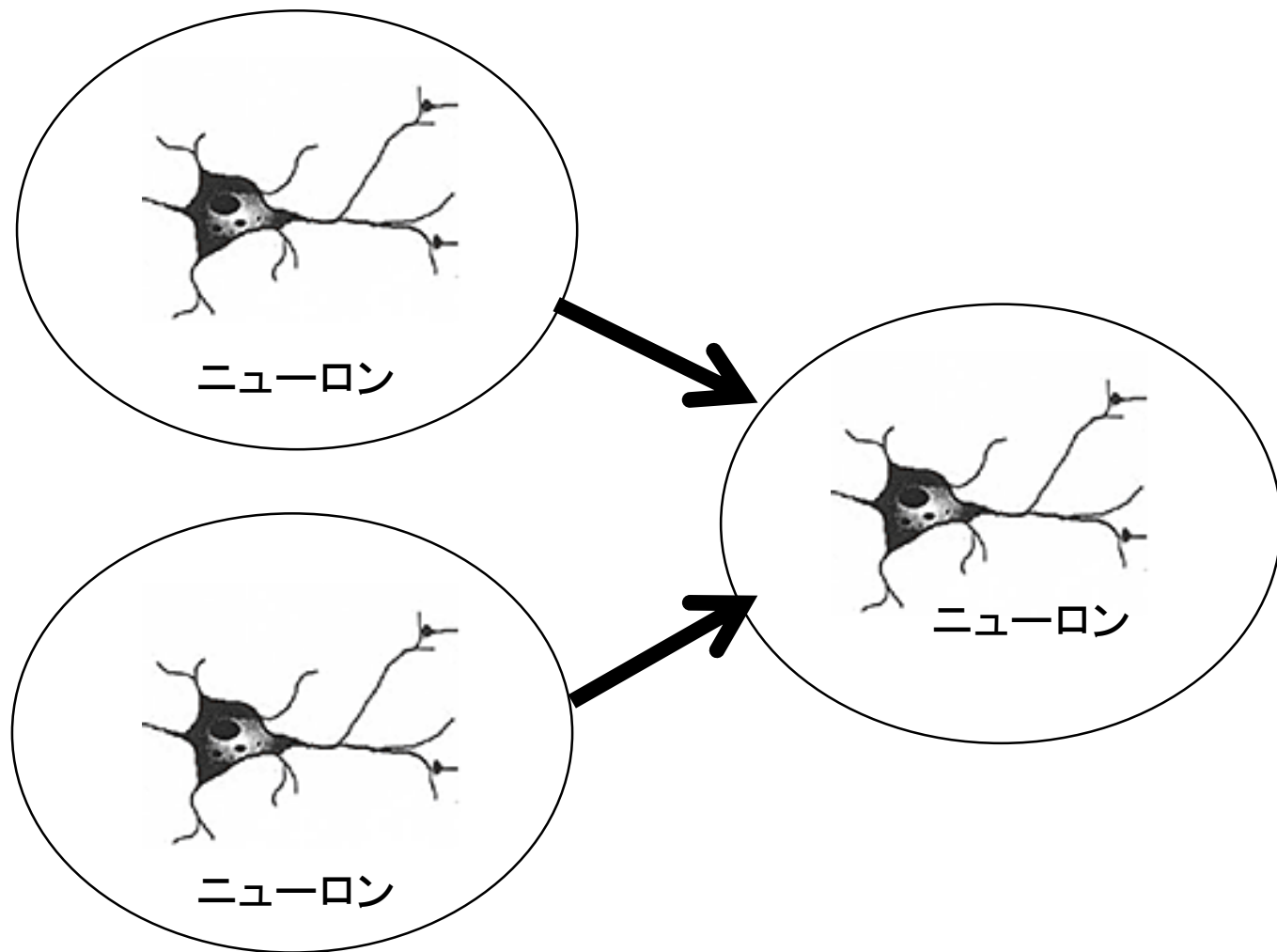
# いろいろな識別方法



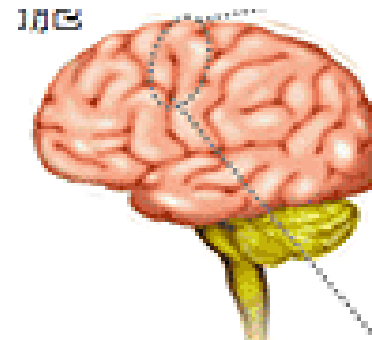
# いろいろな識別方法

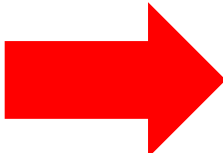


# ニューラルネットワーク

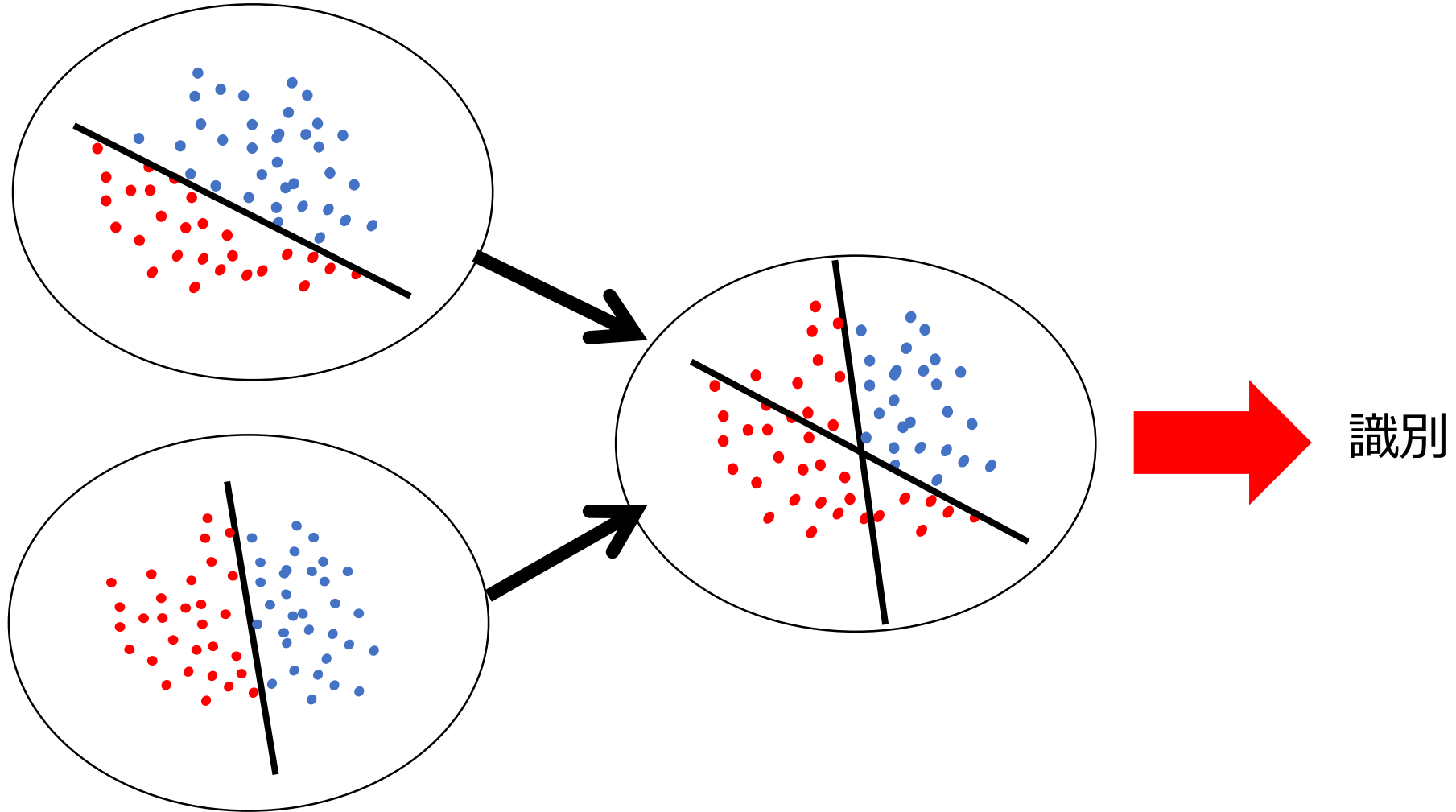


人間の脳はどやって識別  
を行なっているのか？



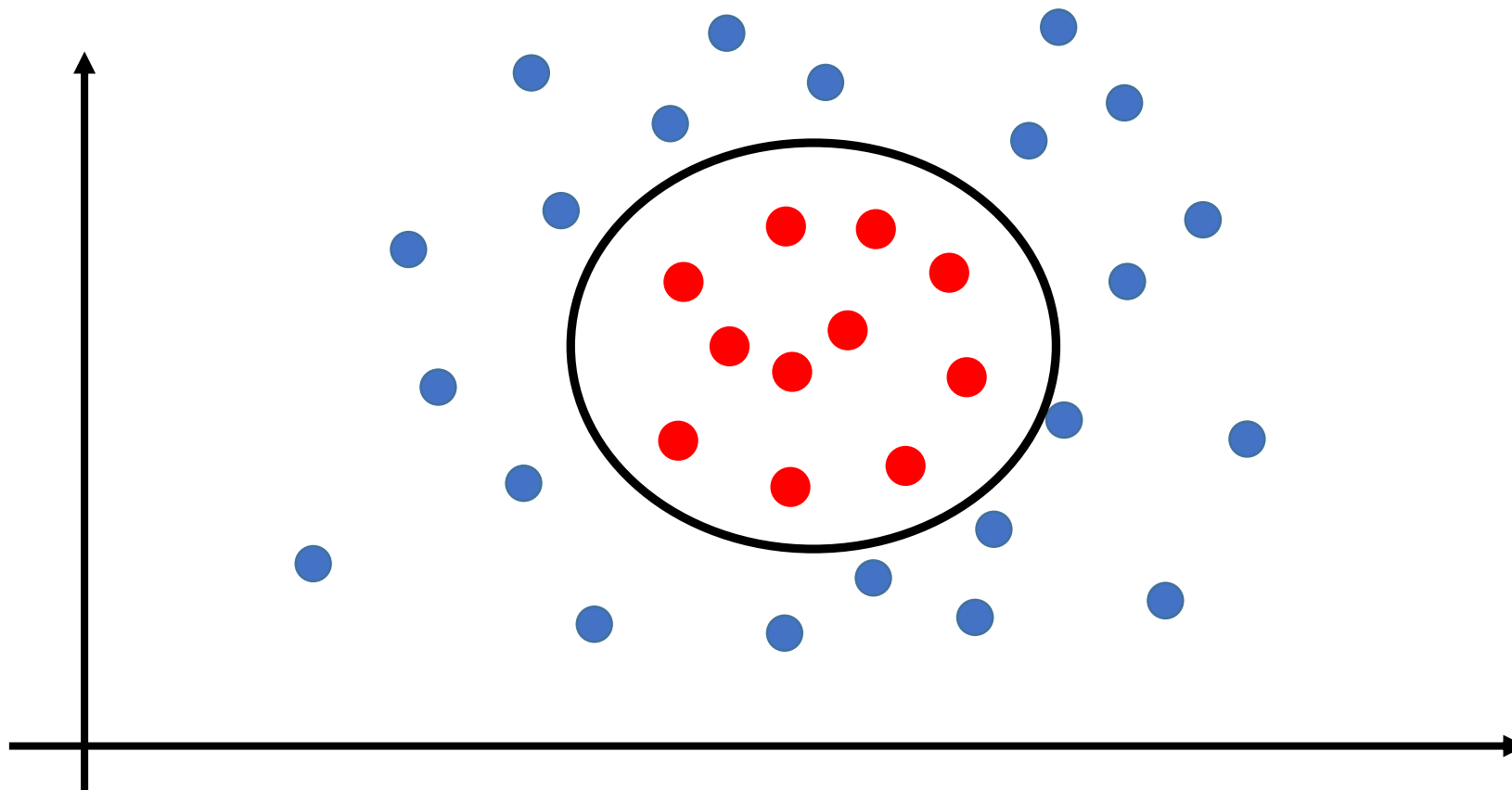
 識別

# ニューラルネットワーク



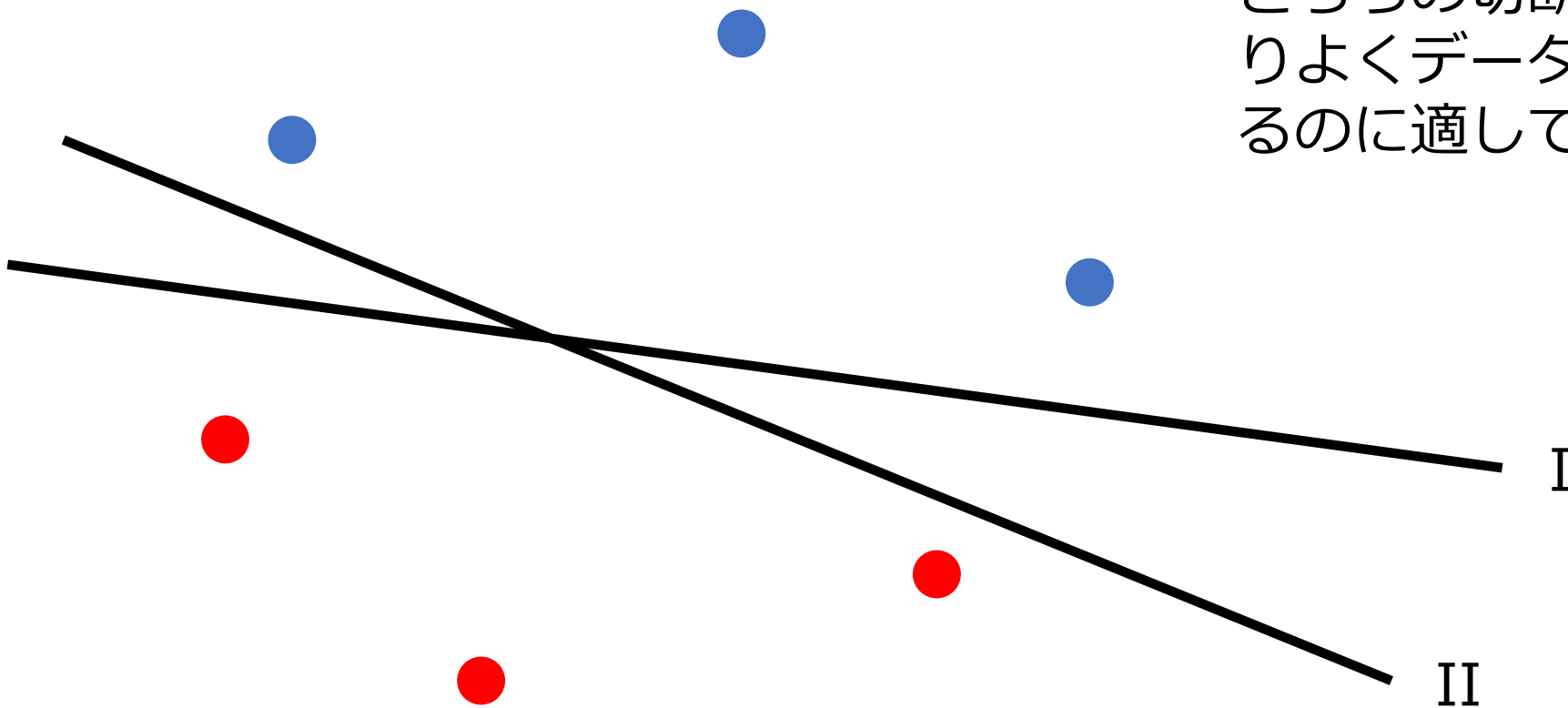
# いろいろな識別方法

## サポートベクターマシーン



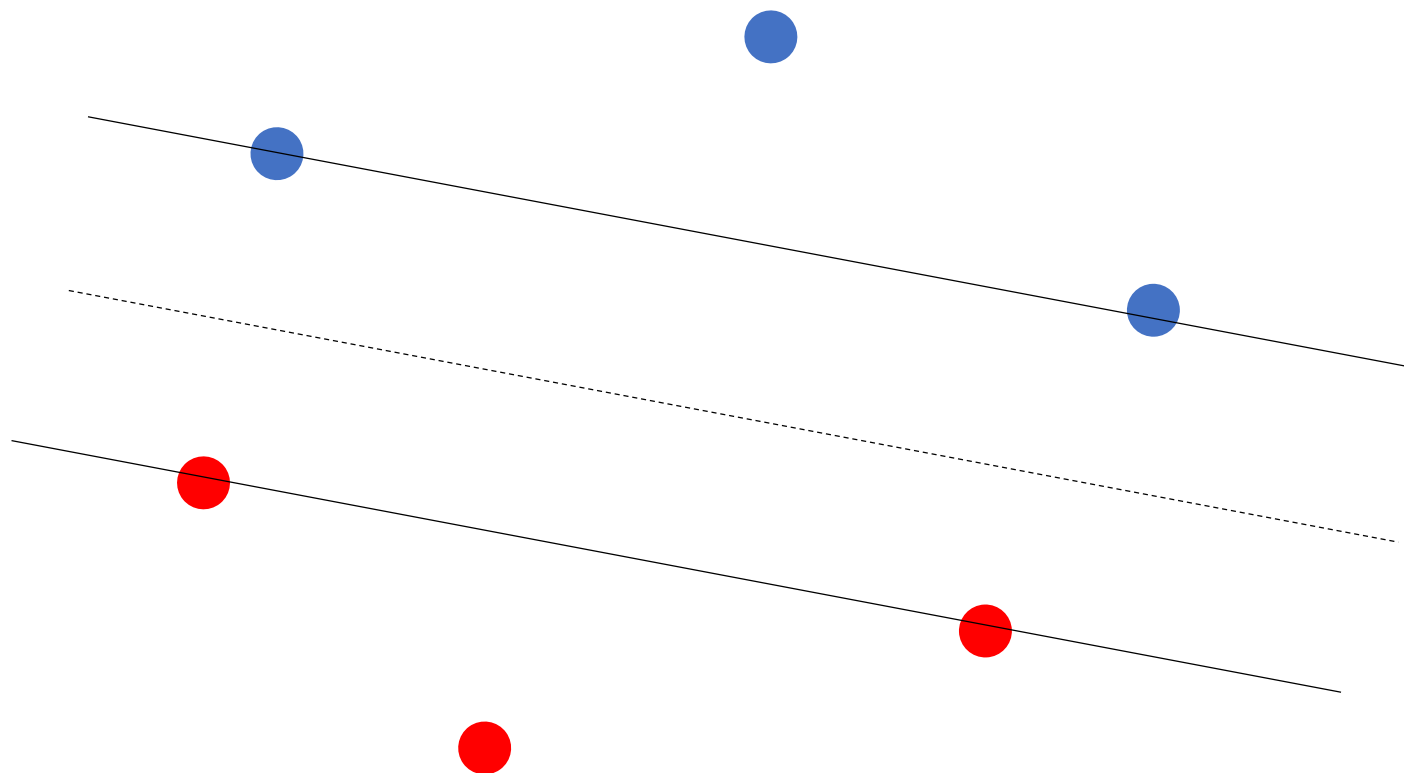
# いろいろな識別方法

どちらの切断の方がよりよくデータを分断するのに適しているか？





# いろいろな識別方法

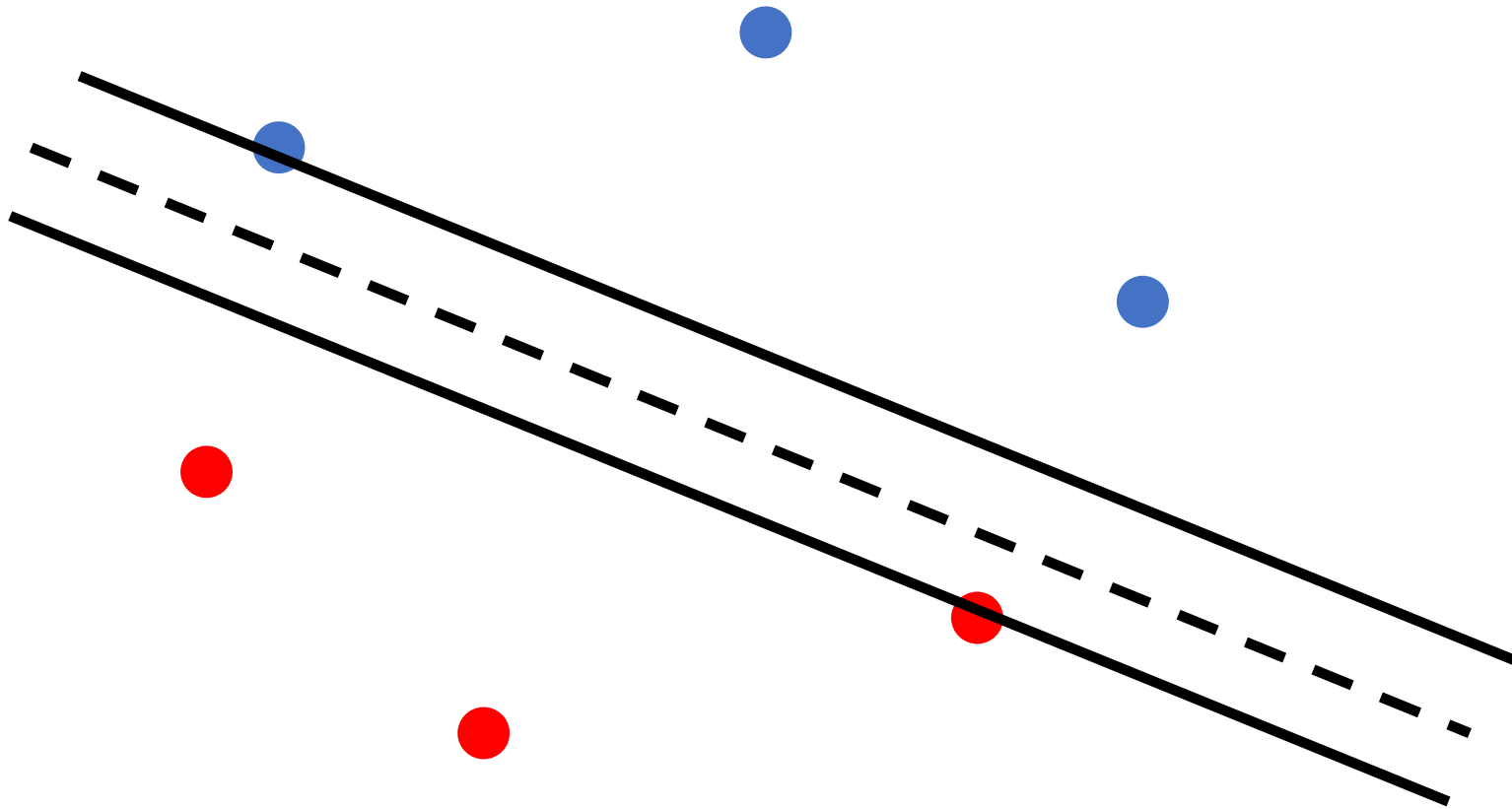


マージンの最大化

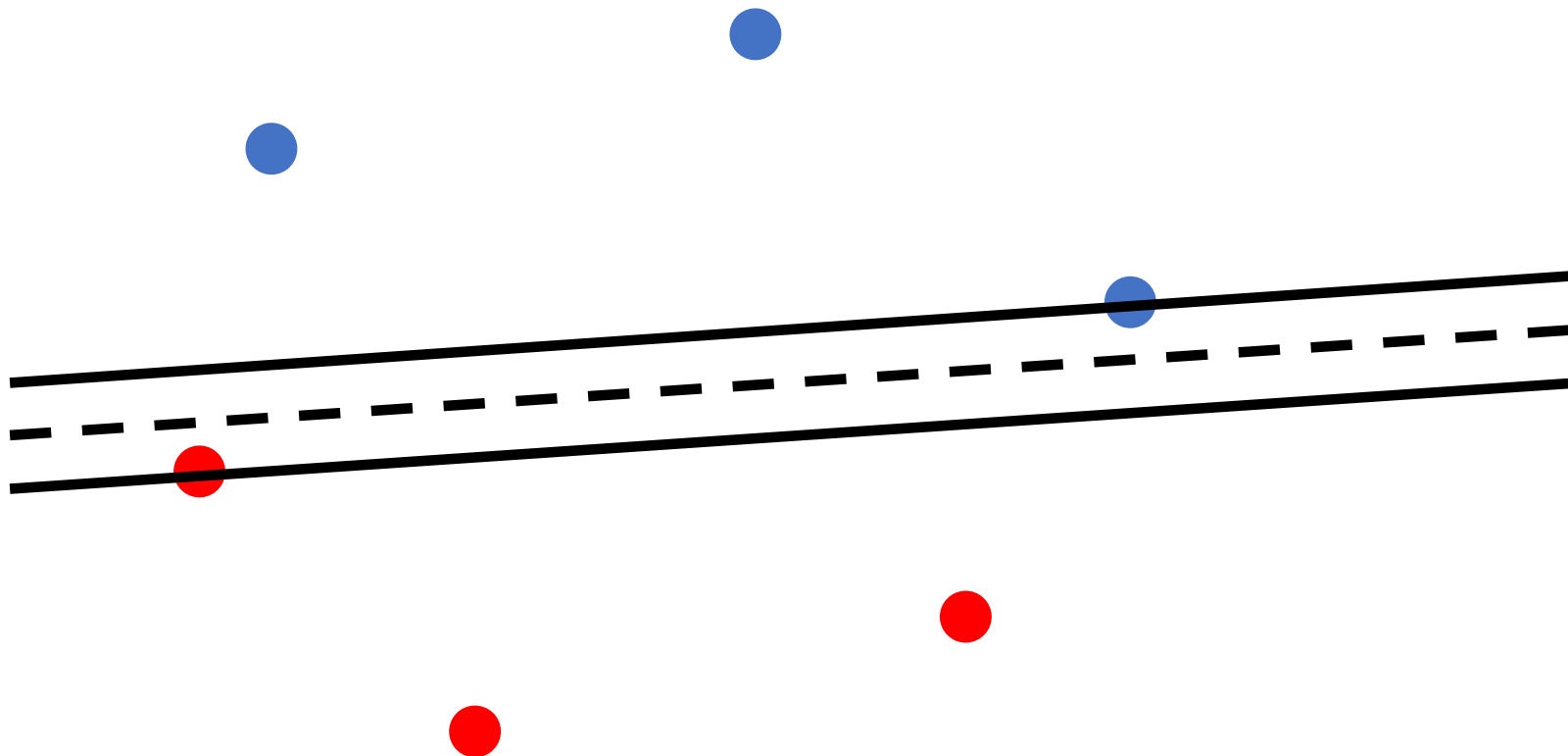
境界データからの  
距離を最大化する

2つのグループ  
の間に一番広い  
道路を設計する  
のと同じ

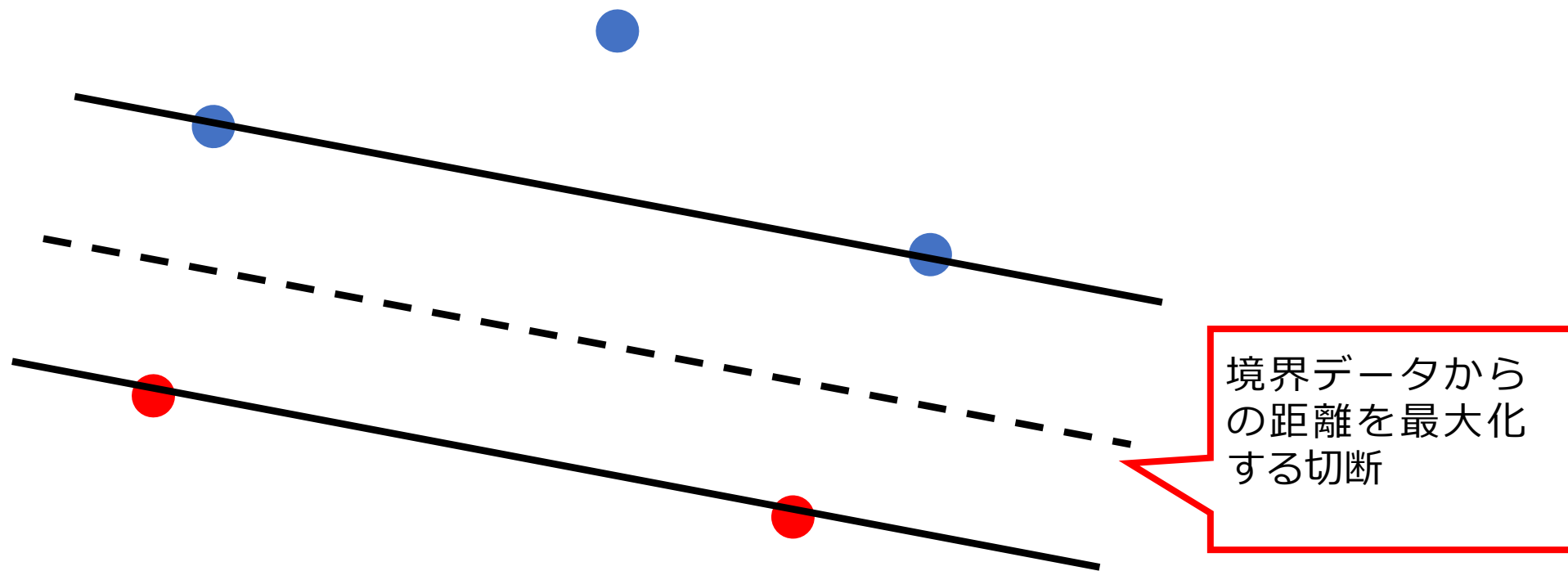
# いろいろな識別方法



# いろいろな識別方法



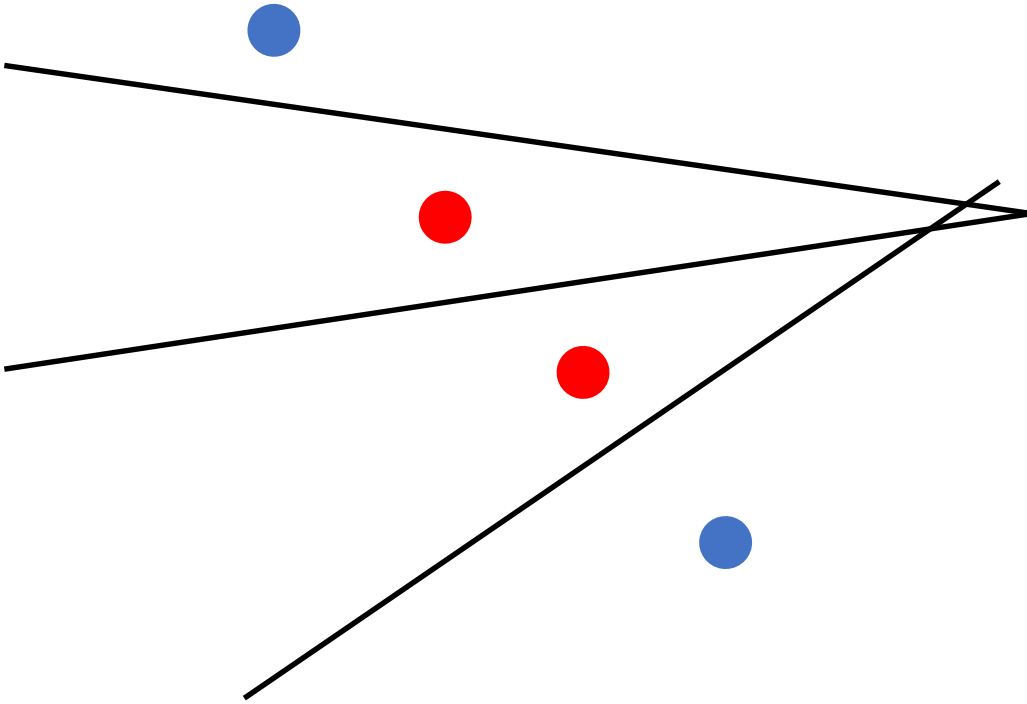
# いろいろな識別方法



サポートマシーンベクトル  
(SVM)

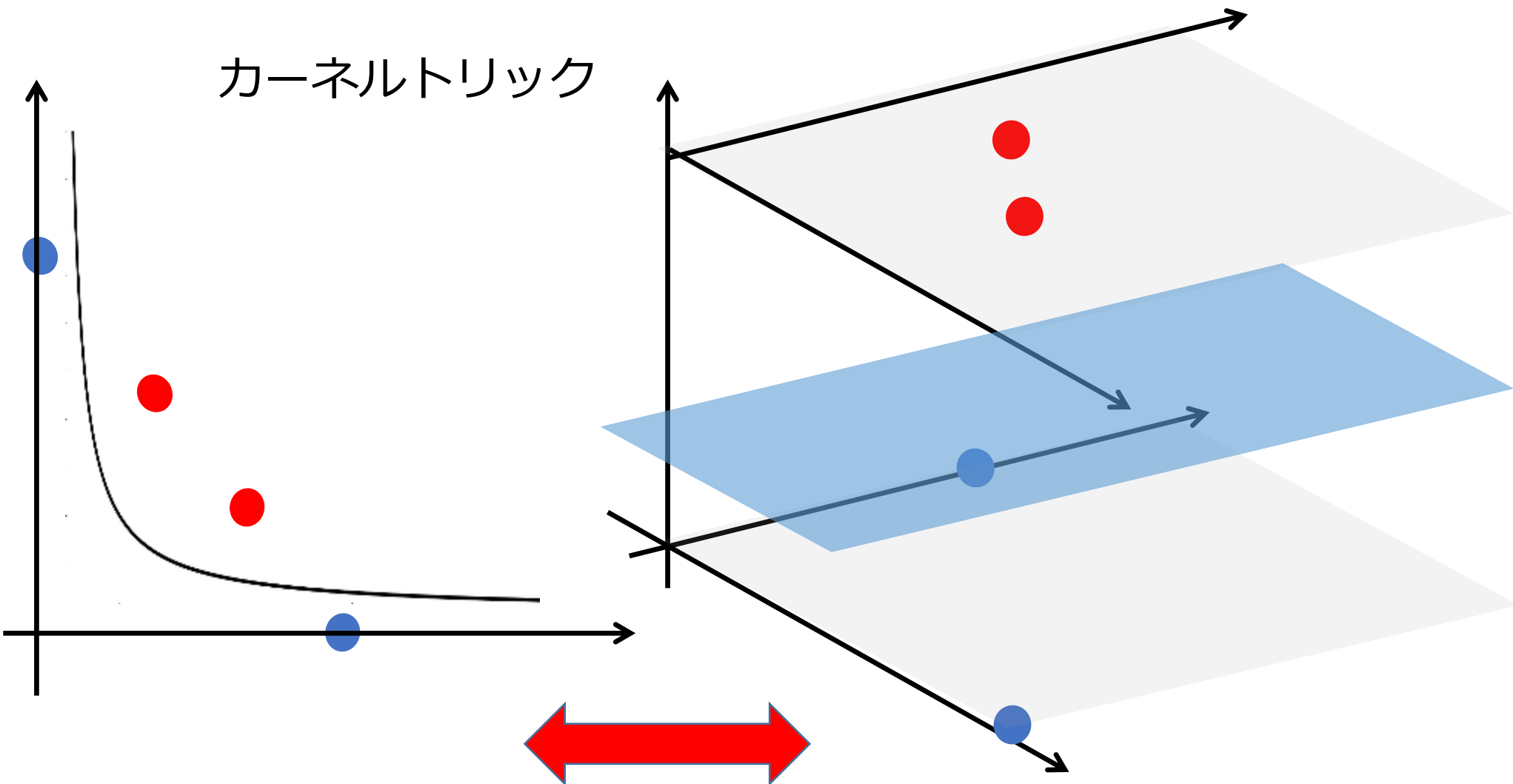
# いろいろな識別方法

直線では分類できない？



# いろいろな識別方法

カーネルトリック



同じこと

## 教師なし・機械学習・データの分類

# クラスター分析

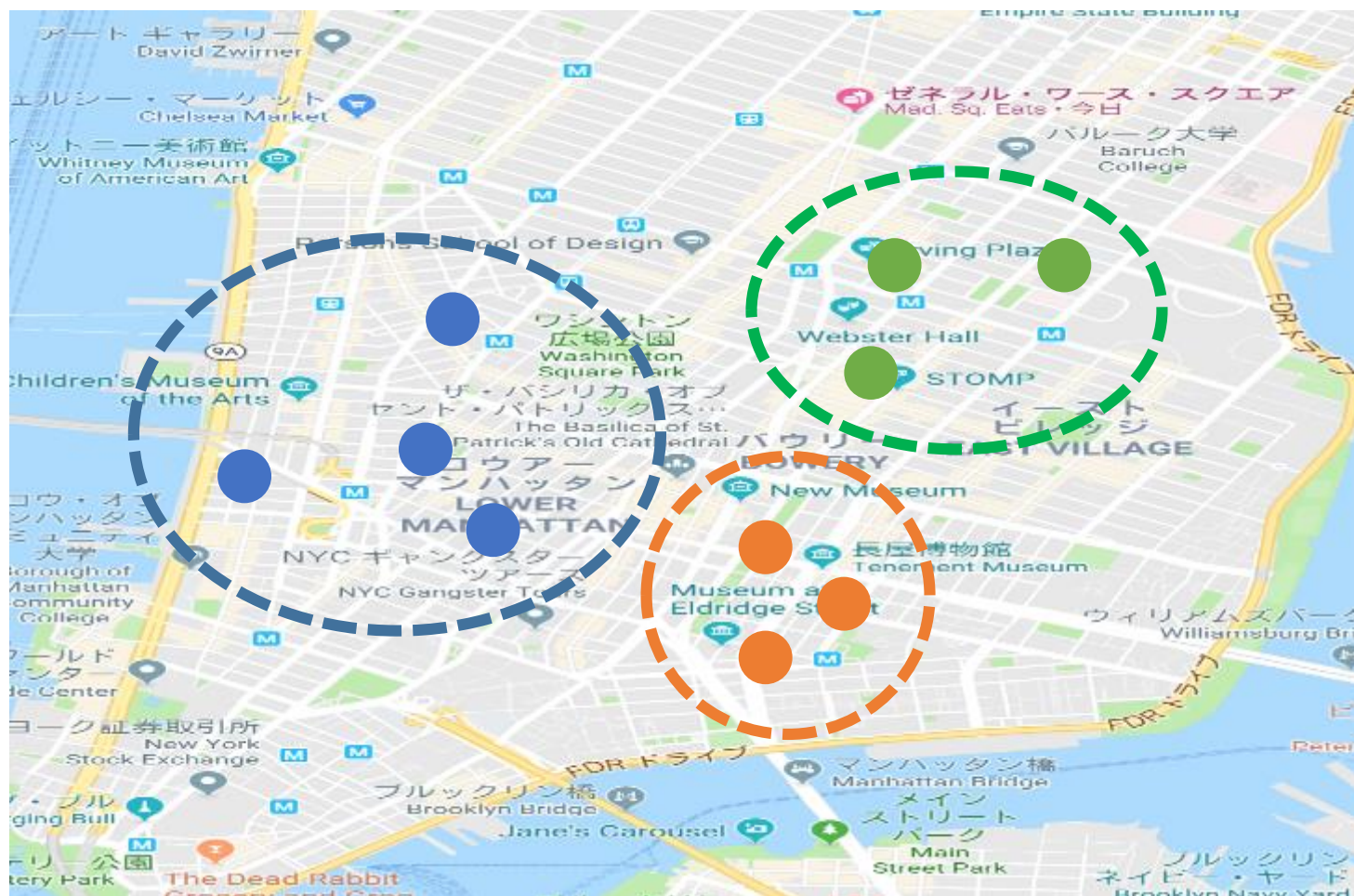
---

- K-mean法



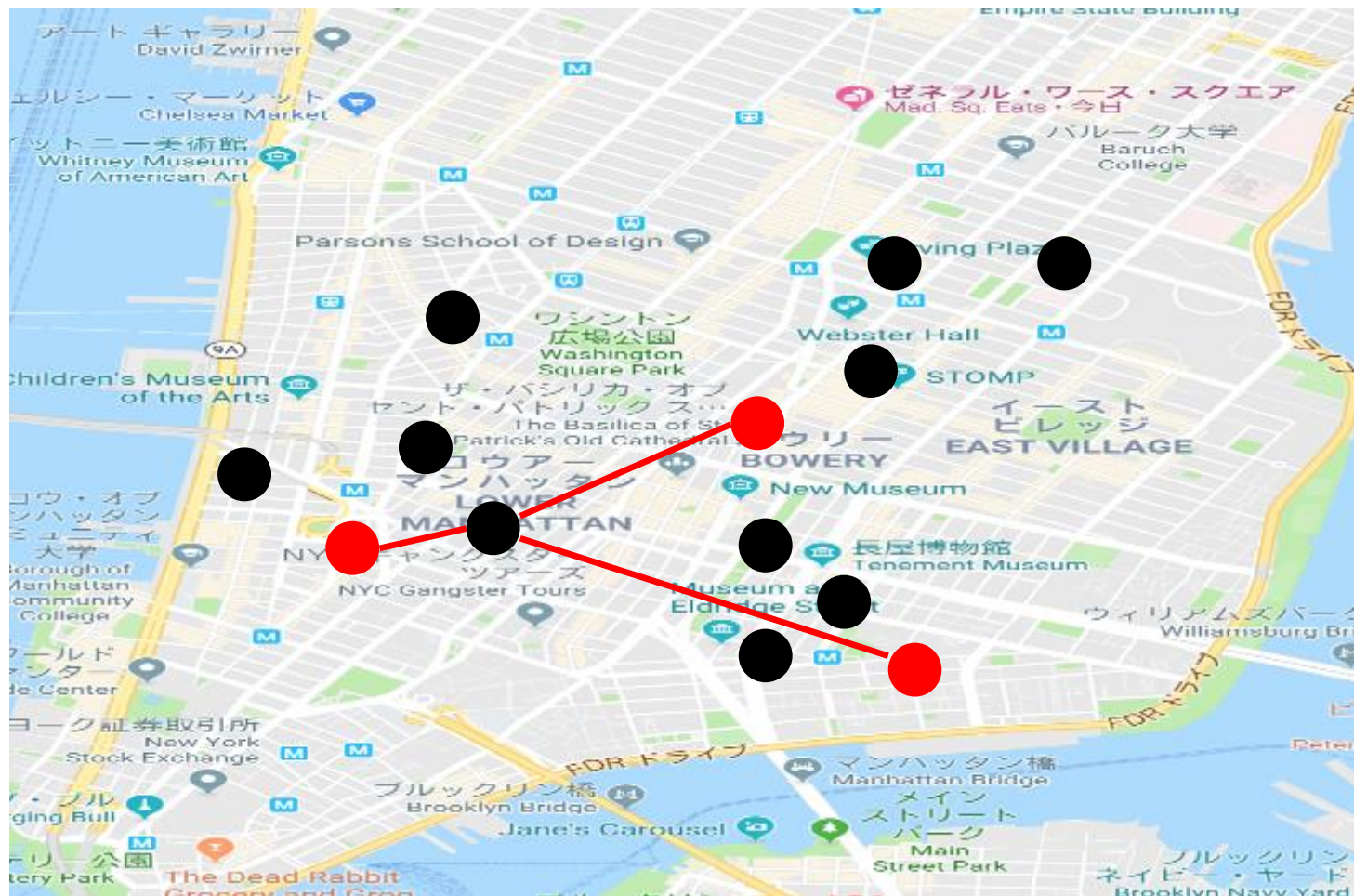
# どこにピザ屋を出店するか？

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？



# 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





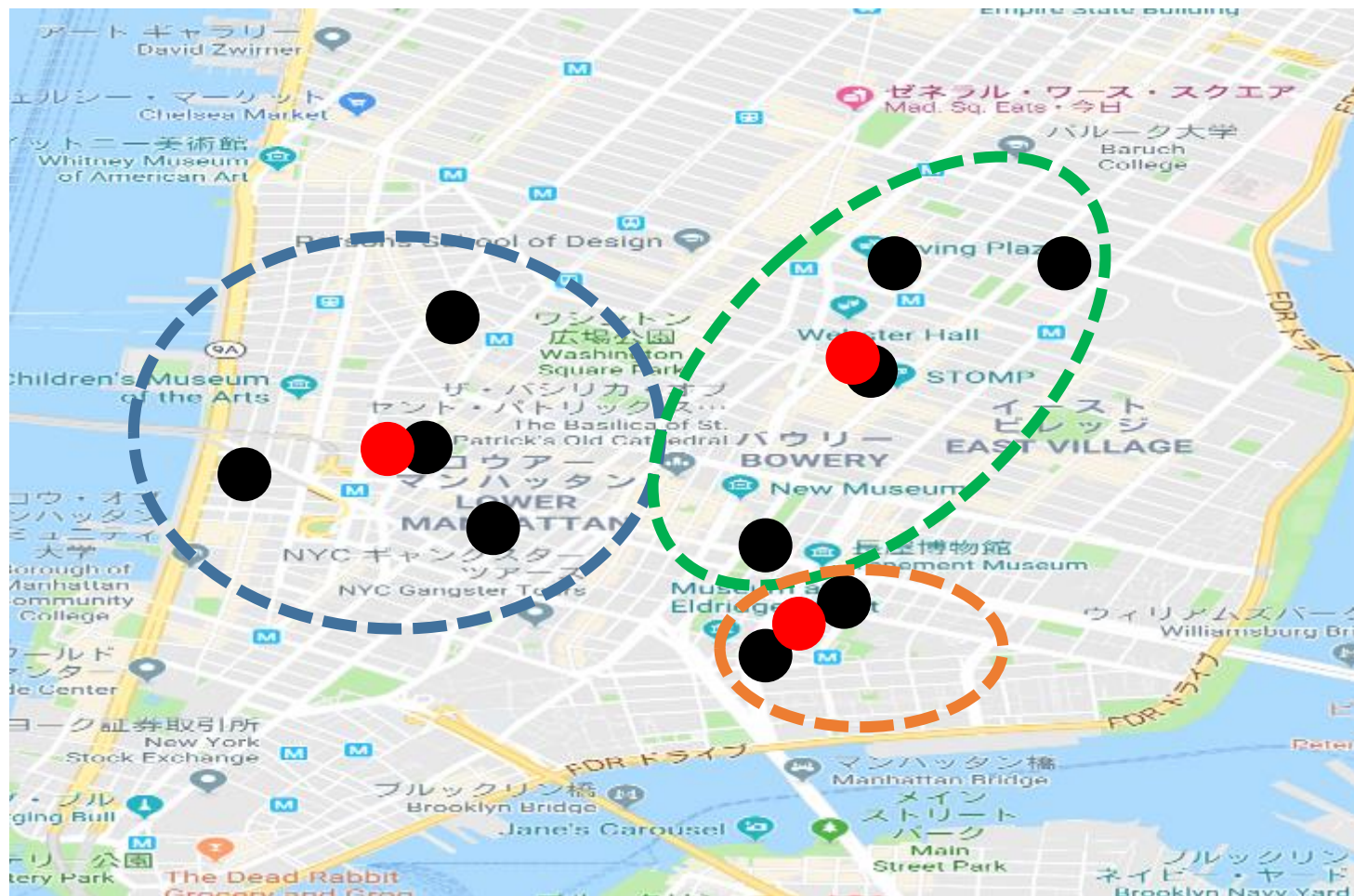
# K-mean法

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

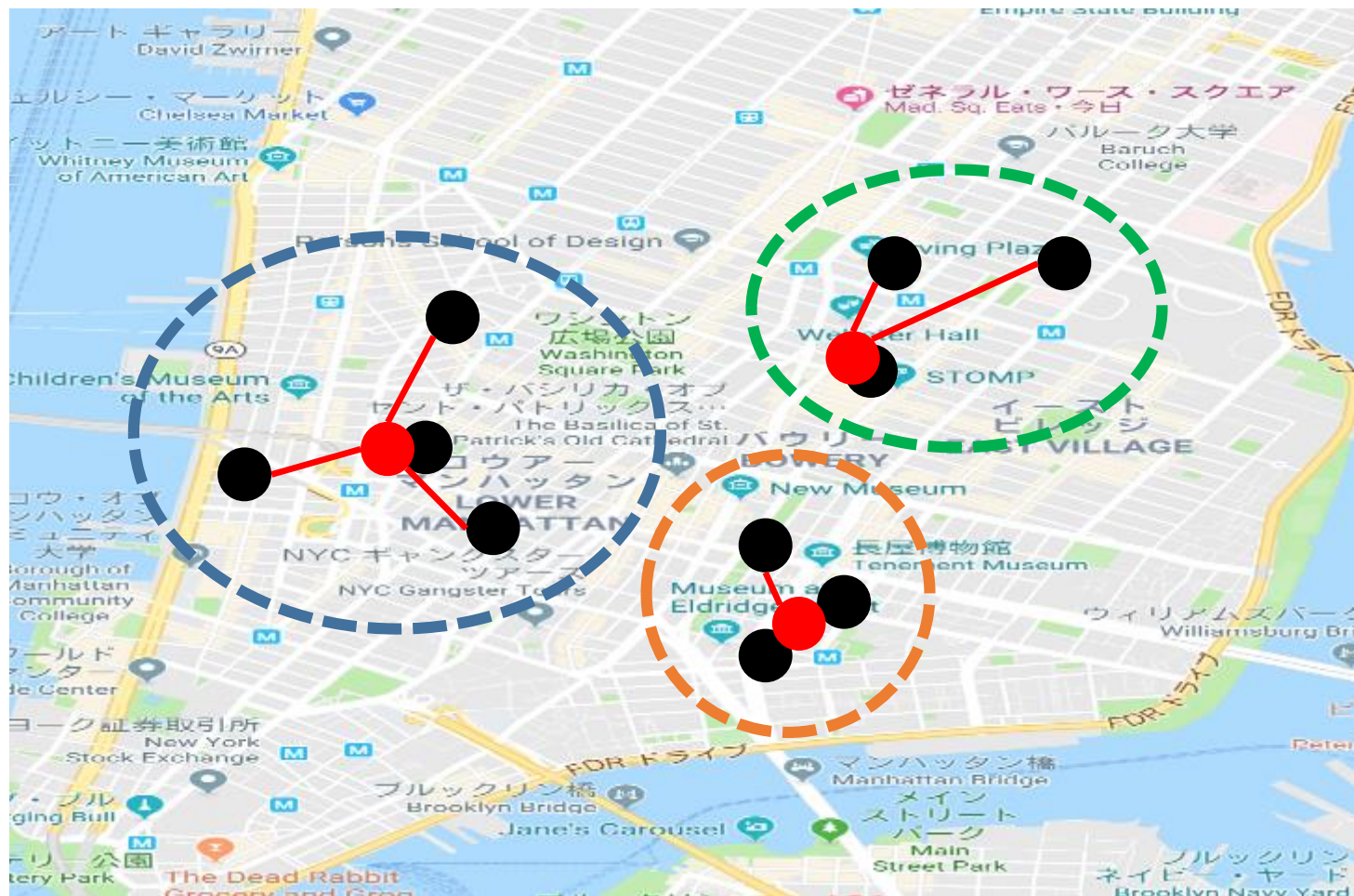
各クラスター内のデータの平均点(重心)を新たなシードとする。





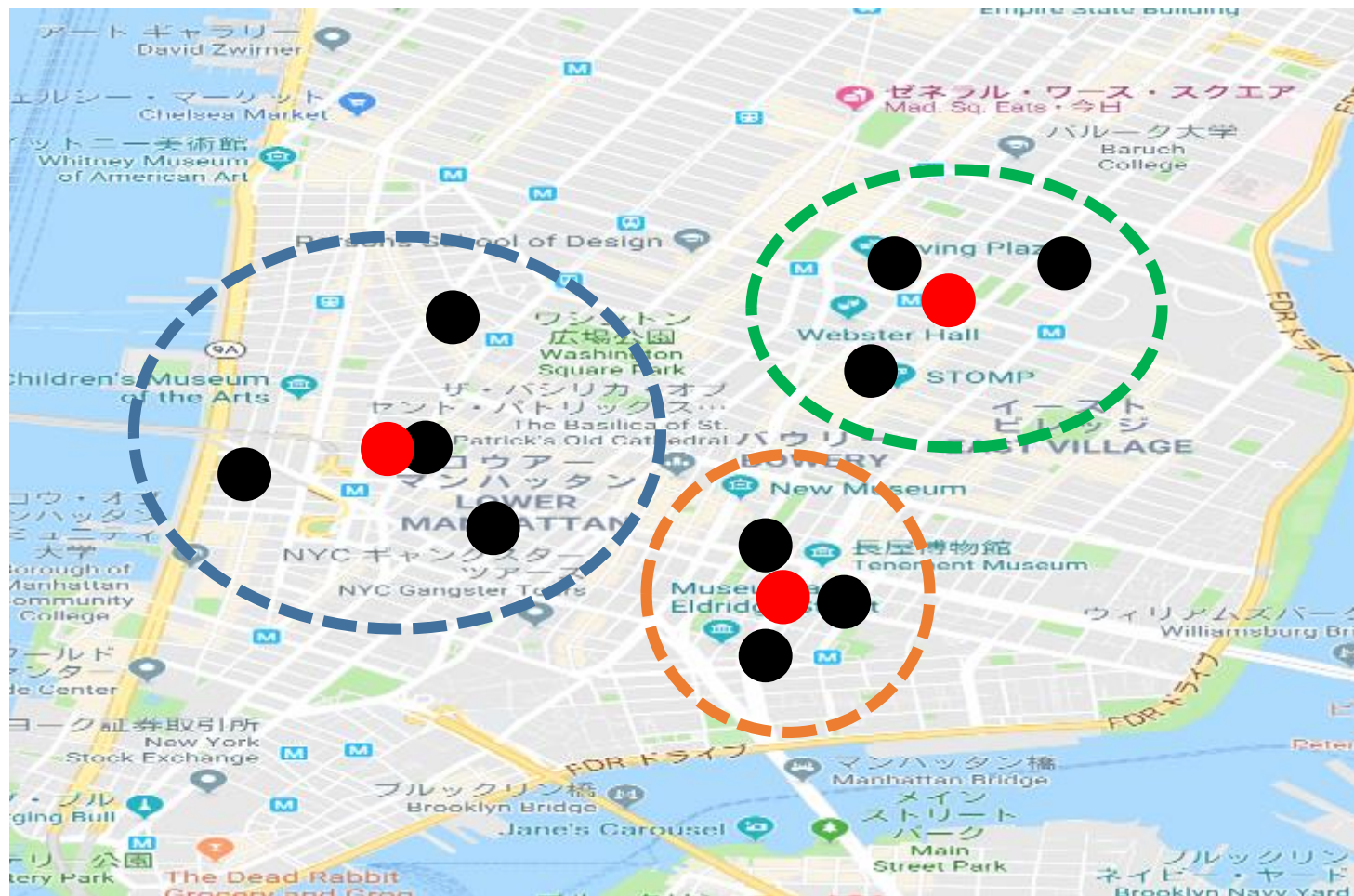
# K-mean法

各データと最も近いシードを紐づける。



# K-mean法

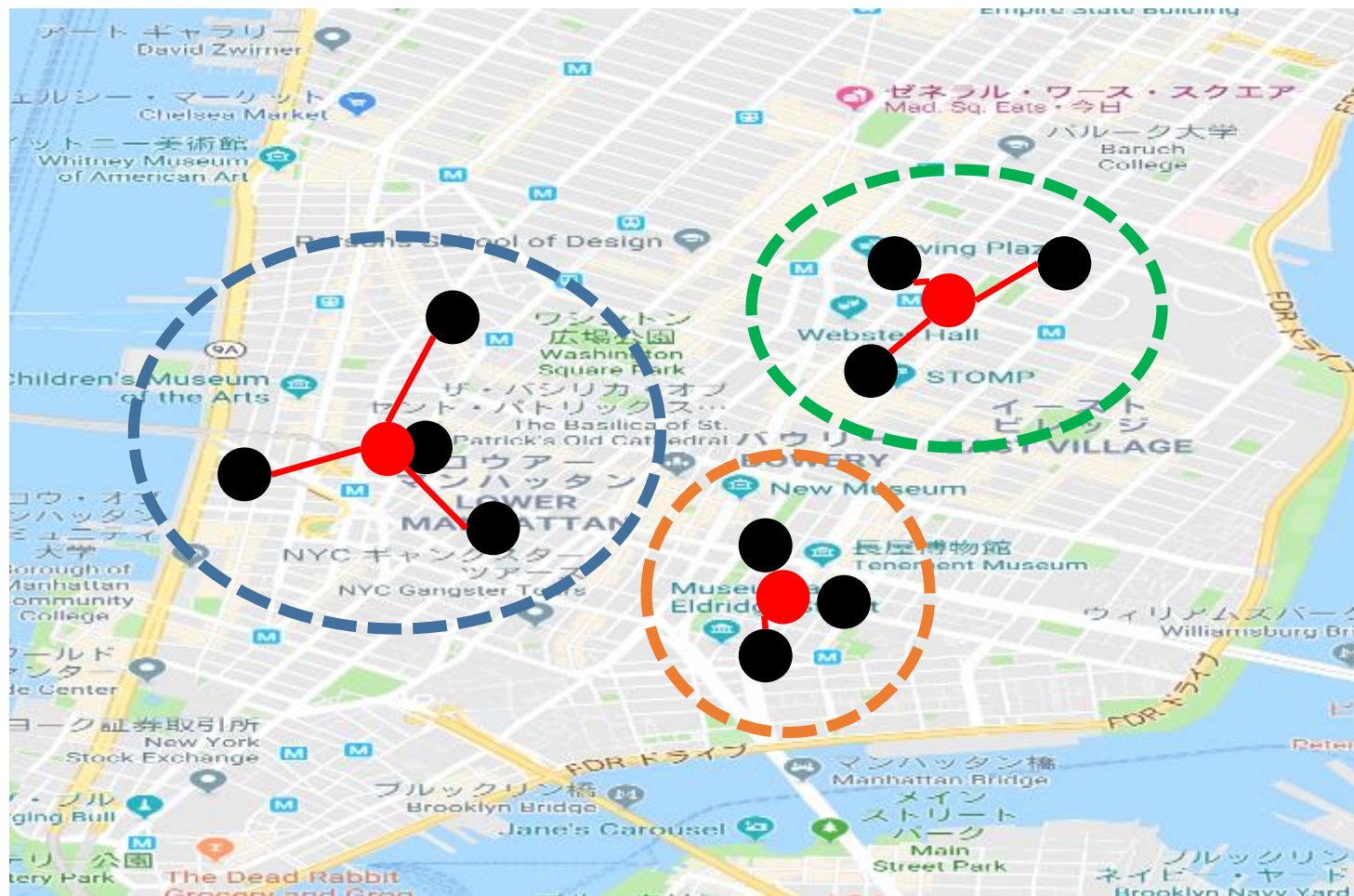
各クラスター内のデータの平均点(重心)を新たなシードとする。





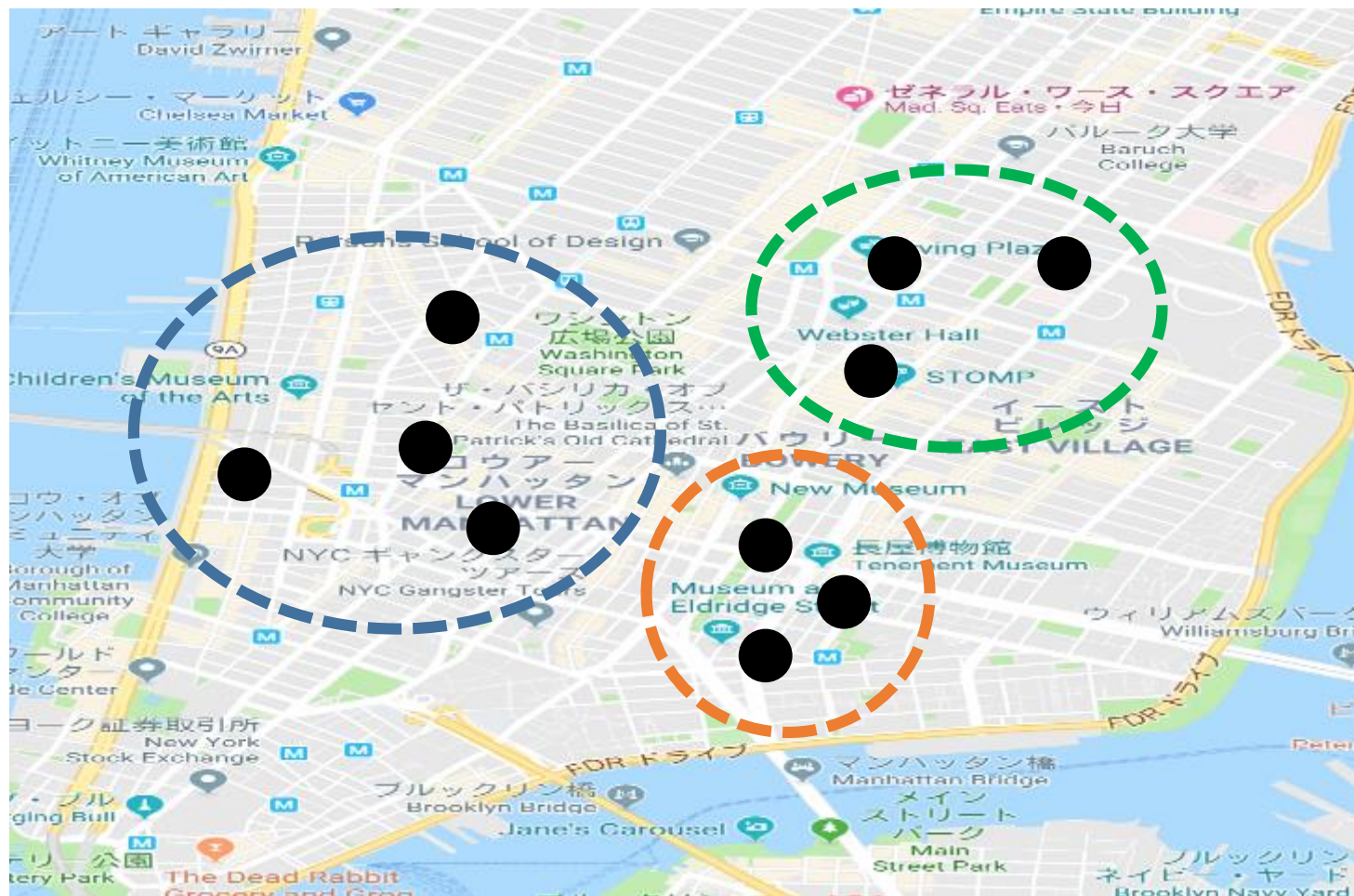
# K-mean法

各データと最も近いシードを紐づける。



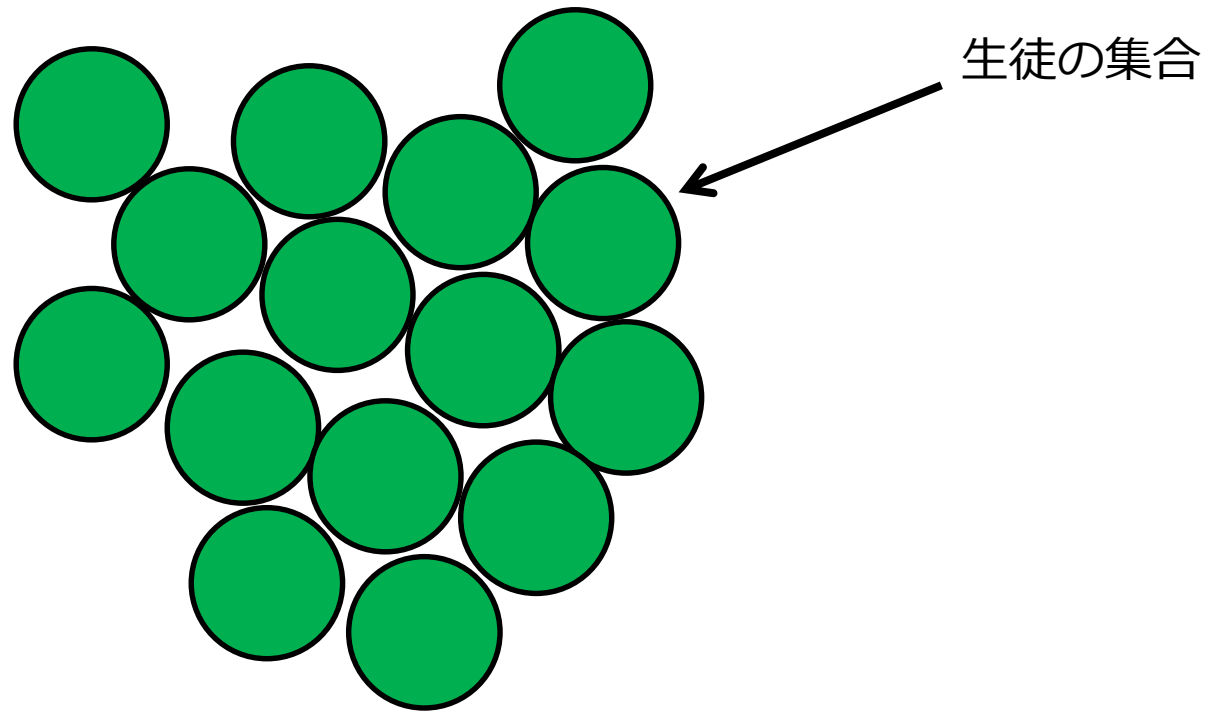
# K-mean法

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。

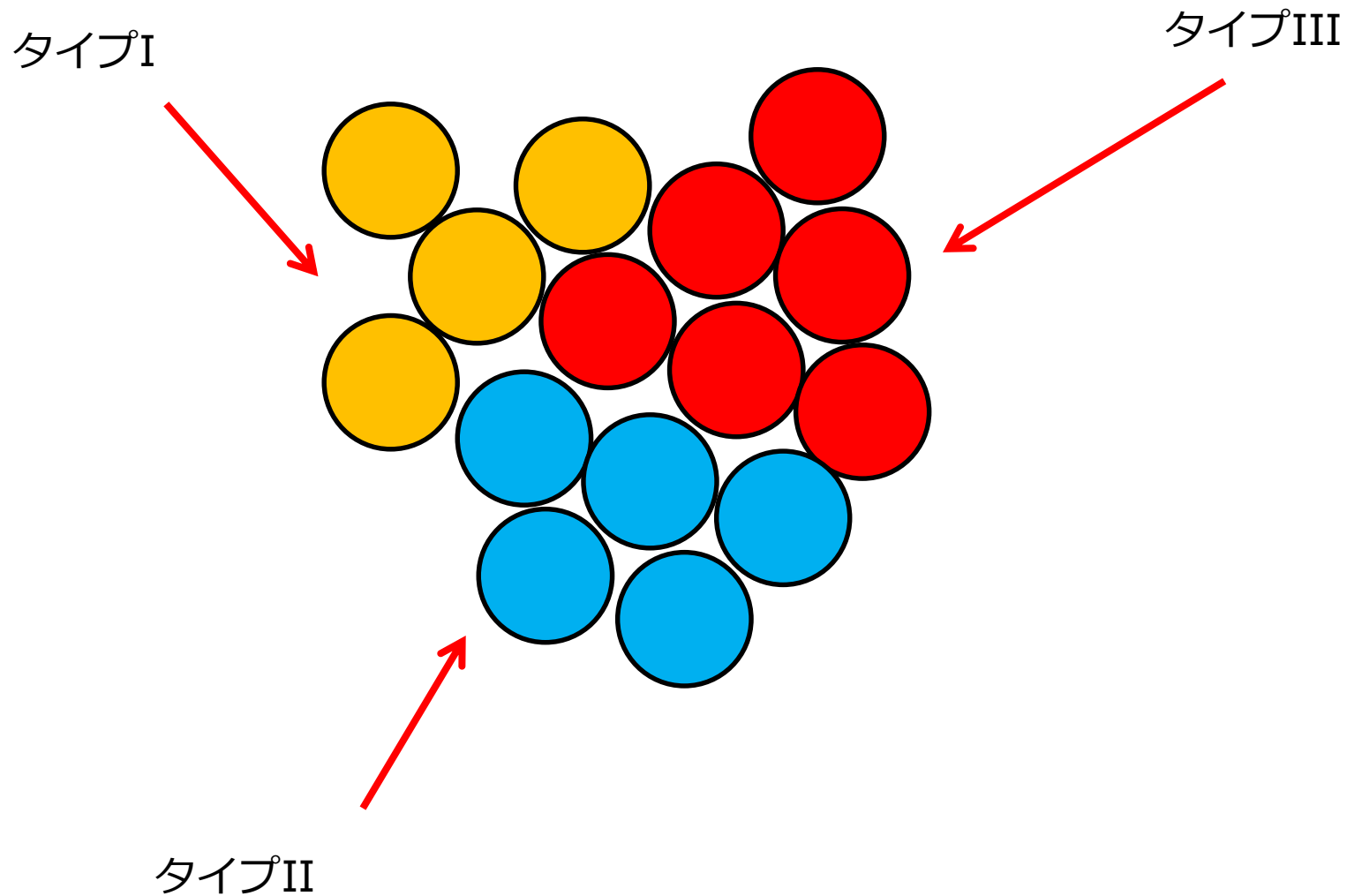




# 主成分分析？



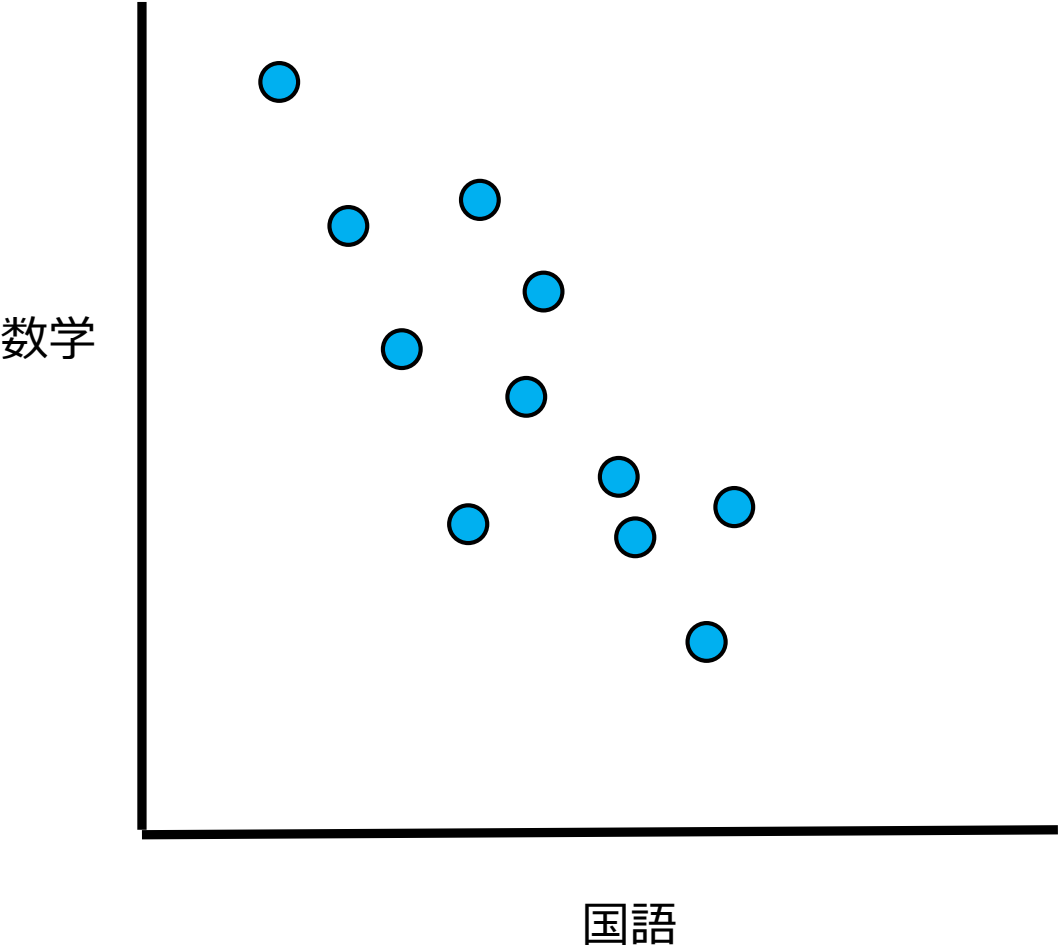
# 主成分分析？



# データ

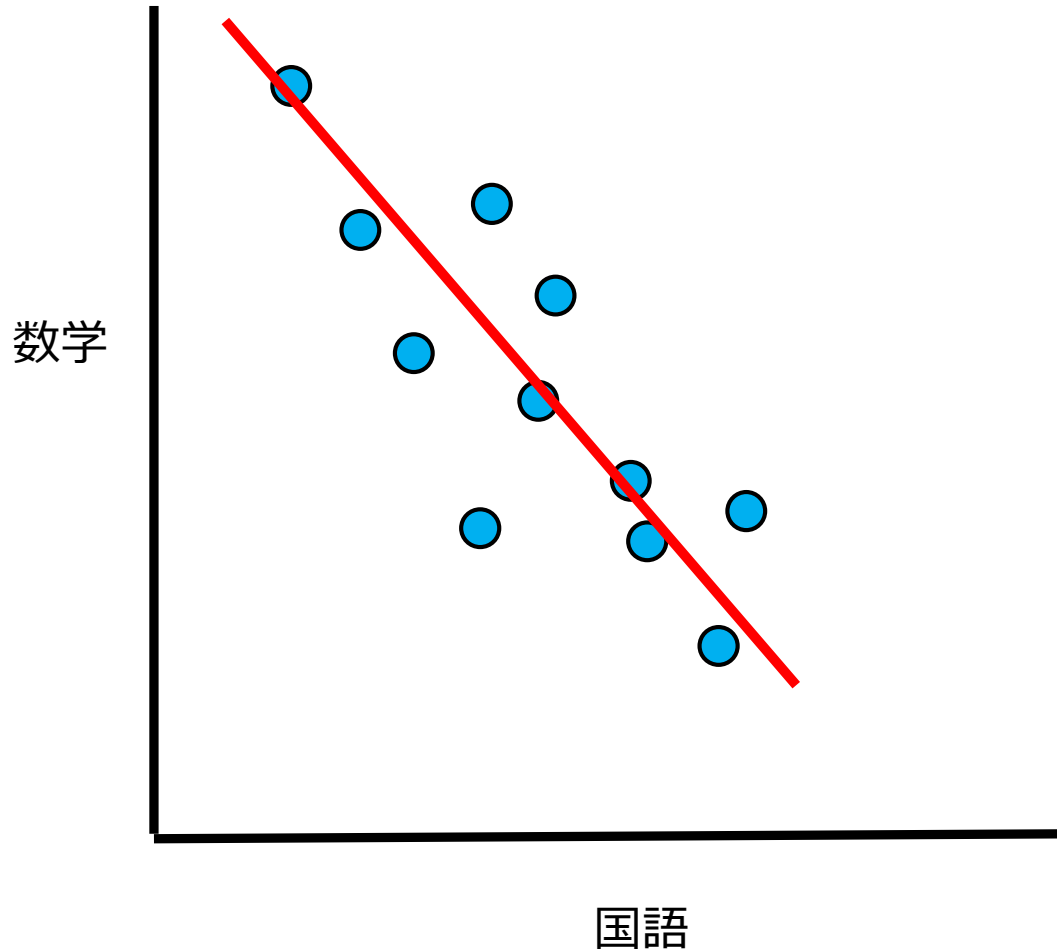
学生ID	数学	国語	物理	社会	化学
1	23	89	34	74	36
2	45	52	32	87	54
3	89	65	87	78	75
4	92	34	95	43	89
5	21	84	21	98	43
6	56	76	34	31	56

# 2Dデータ

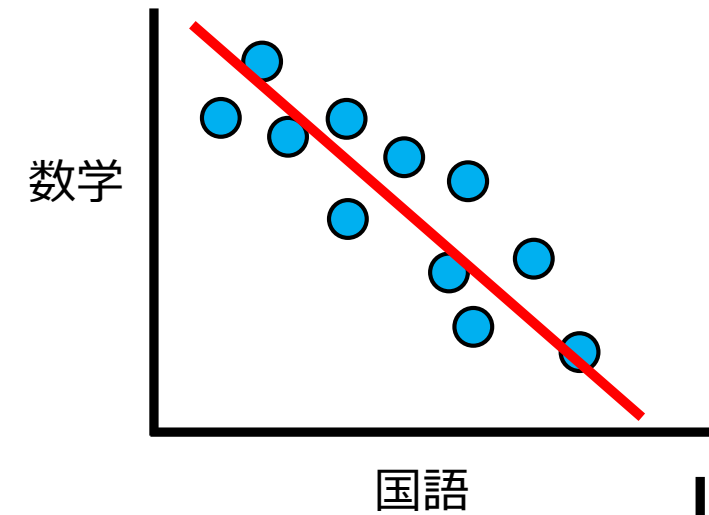


学生ID	数学	国語
1	23	89
2	45	52
3	89	65
4	92	34
5	21	84
6	56	76

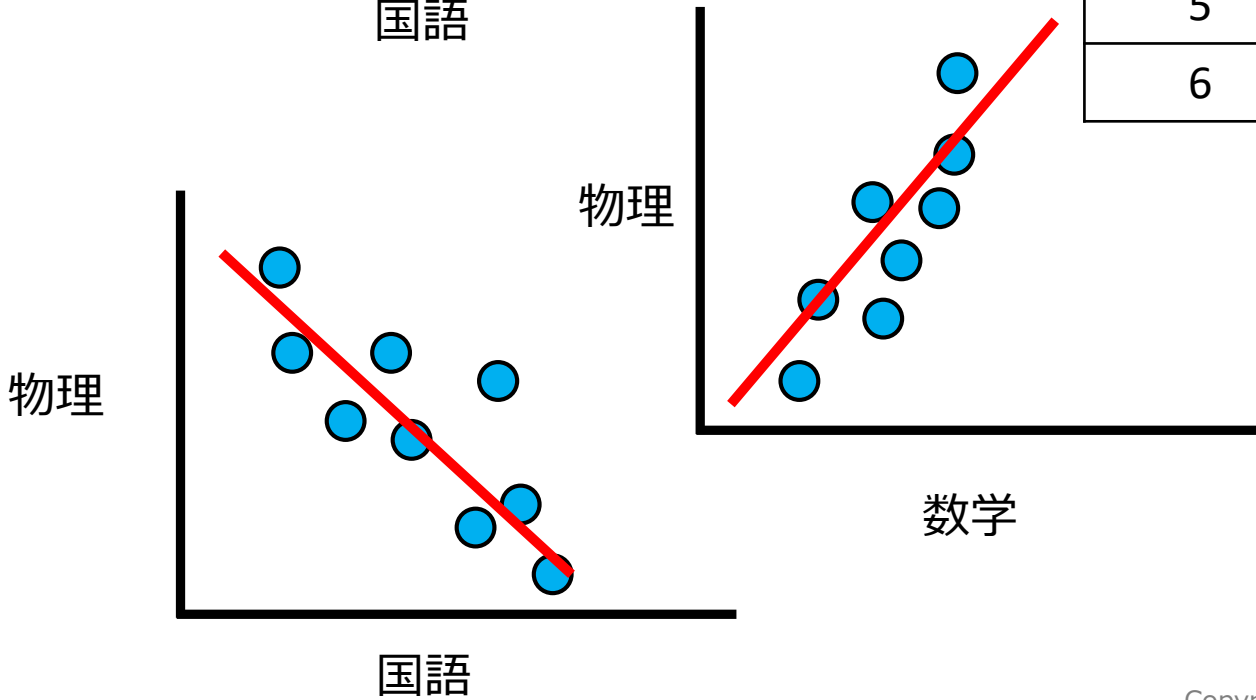
# 2Dデータ



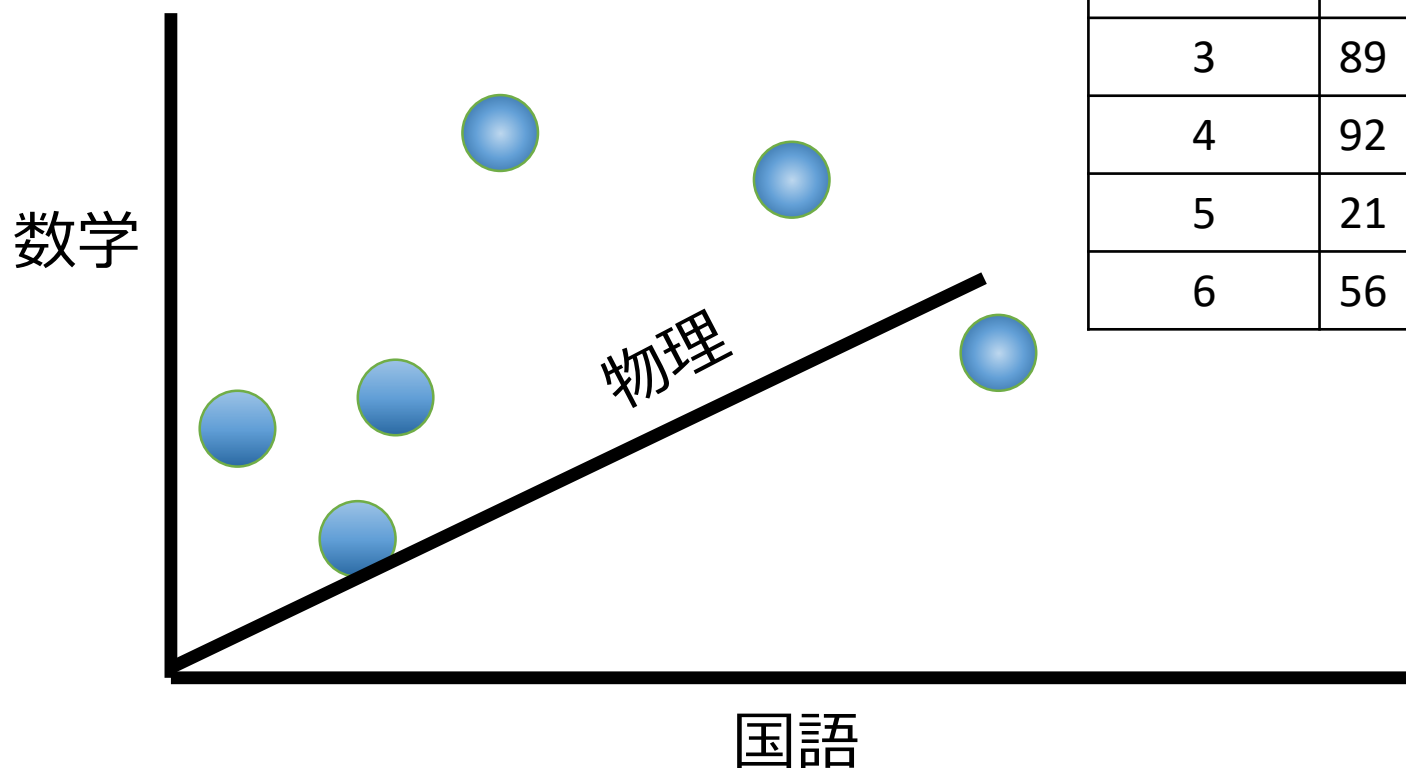
学生ID	数学	国語
1	23	89
2	45	52
3	89	65
4	92	34
5	21	84
6	56	76



学生ID	数学	国語	物理
1	23	89	34
2	45	52	32
3	89	65	87
4	92	34	95
5	21	84	21
6	56	76	34



# 3Dで可視化する



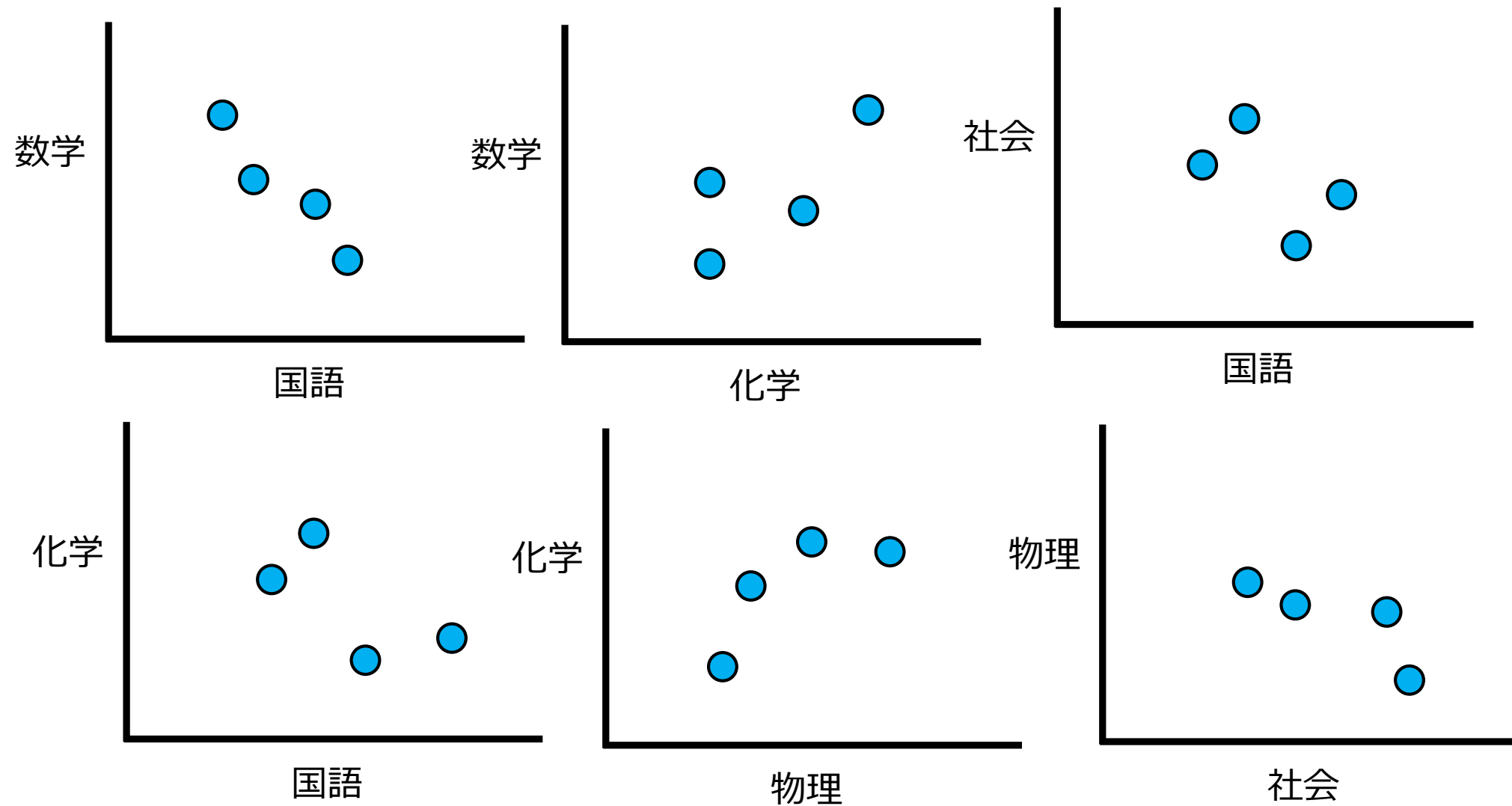
学生ID	数学	国語	物理
1	23	89	34
2	45	52	32
3	89	65	87
4	92	34	95
5	21	84	21
6	56	76	34

# 多次元データの可視化？

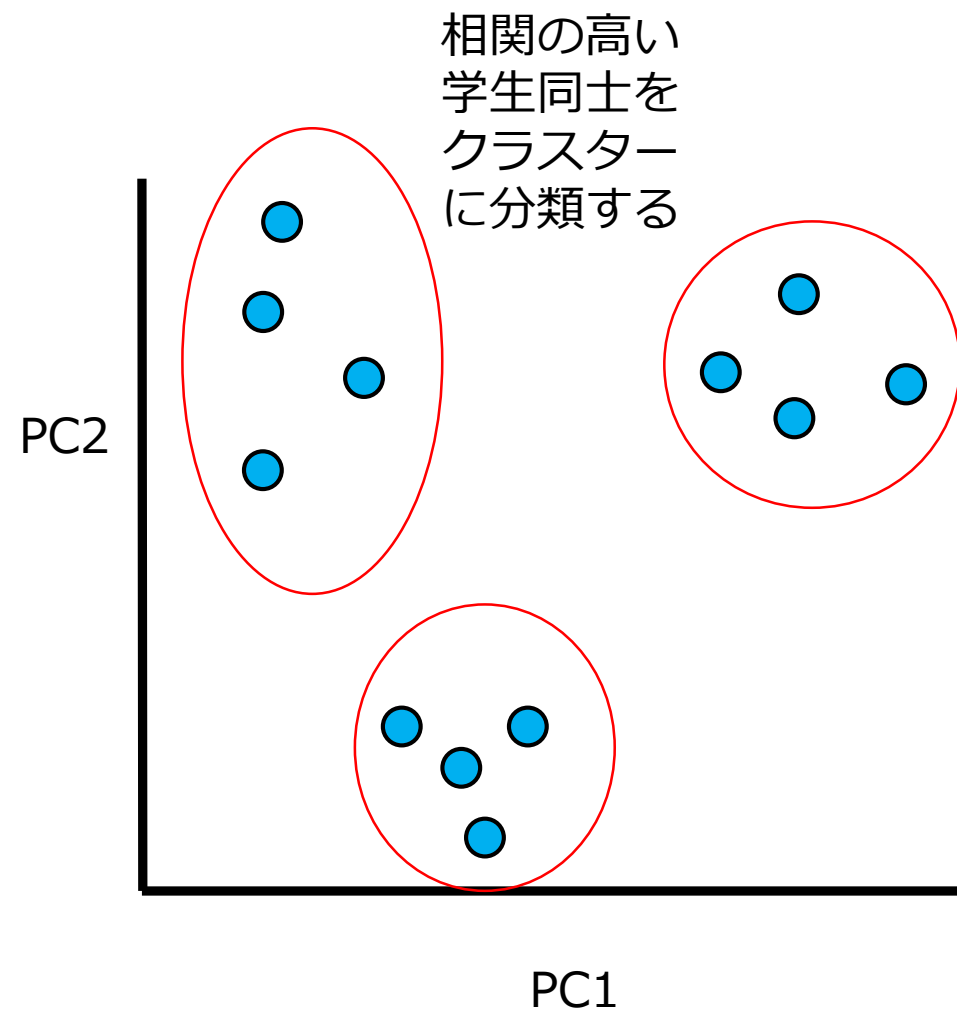
学生ID	数学	国語	物理	社会	化学
1	23	89	34	74	36
2	45	52	32	87	54
3	89	65	87	78	75
4	92	34	95	43	89
5	21	84	21	98	43
6	56	76	34	31	56



# 多次元データの可視化？



# 主成分による可視化



学生ID	数学	国語	物理	社会	化学
1	23	89	34	74	36
2	45	52	32	87	54
3	89	65	87	78	75
4	92	34	95	43	89
5	21	84	21	98	43
6	56	76	34	31	56

主成分分析はデータ間の相関を集約して多次元データを2次元空間でグラフ化することを可能にしてくれる

# 主成分による可視化

2Dグラフでクラスターを識別したら元データに戻る

