

# プログラムなしではじめる 機械学習超入門



# 和から株式会社 会社概要

## Service



### 大人のための数学教室和（なごみ）

大人のための数学教室和（なごみ）は1対1の数学個別指導教室。仕事で数学が必要になった方、数字が得意になりたい方、数学的思考を身につけたい方だけでなく、電検・アクチュアリーなどの資格対策や、大学（院）授業対策など多様な目的で社会人の方が利用。算数の基礎から研究レベルの数学、物理学、ファイナンス数学など多様な数学が対応可能です。



### 大人のための統計教室和（なごみ）

基礎理論から実践的な統計学まで学ぶことができます。マーケティング担当、経営者の方、医療関係者、大学教授など月間400名以上が学びにきています。個別指導だけでなく、少人数講座も開催しております。これから統計をはじめめる方向けの「統計超入門講座」や、「医療統計基礎講座」、Excelを使ってその場で分析を楽しめる「Excel統計講座」など。



### 和（わ）からの企業研修

ビジネスシーンでも統計学を求められる時代。和からは、企業が統計リテラシーを底上げすることで統計学やデータ分析を“真に活かす”ことができると考えています。目指すのは企業内の誰もが統計学やデータ分析を使って、気軽に会話が出来ることです。統計学の初歩から、機学習などの先端スキルの研修をカスタマイズで提供させていただきます。

# 和から株式会社 会社概要

## Service



### ロマンティック数学ナイト

数学がとにかく好き、数学に興味がある、数学を共有したい、数学で繋がりたい、そんな人達のためのそんな人達による数学のショートプレゼン交流会を全国で開催しています。立場も、肩書も、年齢も、能力も、関係なく、自由に集い、共に活動できる参加型の数学コミュニティとして毎回200名以上の参加者で賑わっています。



### ロマンティック数学ナイト ゼミ

数学の”おもしろいところだけ”を学ぶことができる少人数制ゼミです。最先端の数学、未解決問題に挑戦するだけでなく、子どもの時にずっと疑問だった数学の定理について深く学んでいくのもこのゼミの特徴です。数学を楽しみと思う人と一緒にコミュニティーで学ぶことができます。

# 和から株式会社 会社概要

## About

- 設立：平成23年3月3日（事業開始平成22年1月）
  - 従業員数：15名（登録講師数35名）
  - WEB：<http://wakara.co.jp/>
  - 代表取締役：堀口智之
  - 資本金：3,141,592円
- \* 渋谷第一教室：東京都渋谷区渋谷3-6-19 第1矢木ビル4階B室（本社）
  - \* 渋谷第二教室：東京都渋谷区渋谷3-5-16 渋谷3丁目スクエアビル2F
  - \* 新橋教室：東京都港区東新橋2-10-10 東新橋ビル
  - \* 大阪教室：大阪市中央区伏見町4-4-9 淀屋橋東洋ビル3F
  - \* WakaLabo新宿：東京都新宿区西新宿7-9-6 寿ビル502



# 和から株式会社 会社概要

## 実績

### 大人のための統計教室和（なごみ）

2012年、統計ニーズが急増してきたことから開校しました。基礎から実践的な統計学・データ分析までを学ぶことができます。企業のマーケティング担当・データサイエンス担当、経営者の方、医療関係者、大学教授など月間400名以上の方が業務・研究活用のために学びにきています。

個別指導だけでなく、少人数講座も週1回程度開催、企業研修も実施しております。とくに、最近では企業におけるデータ分析導入のサポート（データ分析の入門研修からデータ分析導入の為の組織作りのアドバイスまで）を行っております。



2013年日本経営協会様講演  
「回帰分析からわかる統計実用基礎講座」



弊社主催「統計超入門講座」  
※月に複数回開催

### 講演会と企業研修

- ・一般社団法人日本経営協会
- ・練馬区生涯学習センター講演
- ・国分寺市光公民館講演
- ・大手広告系 R 社統計学研修
- ・大手広告系 R 社統計学OJT
- ・大手広告系 R 社組織導入サポート
- ・大手 I T 系 R 社統計学研修
- ・大手 I T 系 D 社統計学研修
- ・大手中古車販売 G 社統計学研修
- ・大手通信 S 社統計学研修
- ・大手損保会社 S 社
- ・資格合格率アップコンサル
- ・大手TV局 F 社数学番組制作補助
- ・大手TV局 T 社数学番組制作補助
- ・大手ゲーム制作会社 D 社  
～統計分析サポート～
- ・各社オペレーション業務効率化  
～データ分析補助～
- ・公益財団法人数学検定協会  
～統計講座共同開催など多数開催



# 和から株式会社 会社概要

## 実績

### News | ニュース

[> トップに戻る](#)

2018.02.01 | Press Release

## データビークル、和から株式会社の協力のもと「データ分析人材育成サービス」を開始～ビジネスパーソン向け統計・数学の入門講座を開講～

株式会社データビークル（本社：東京都港区、代表取締役社長 油野 達也）は、社会人向けの数学教室、統計教室などを運営する和から株式会社（本社：東京都渋谷区、代表取締役社長 堀口 智之）と業務提携を行い、統計家の西内 啓（データビークル共同創業者・最高製品責任者）が監修を行うビジネス統計学講座を開講することをお知らせいたします。

本リリースのPDF版はこちら



掲載サイトURL: <http://www.dtvcl.com/news/20180201/>

### ■ ビジネス統計学の第一人者が教えるプロと手を組んだ

このような背景のもと、データビークルは社会人向けの数学教室の運営実績があり「教えるプロ」である、和からと協力し、「ビジネスパーソン向け基礎統計学・数学講座」を新開講いたします。講座はデータビークルの共同創業者兼、最高製品責任者である統計家の西内 啓が監修をおこない、現場に必要とされる統計学に絞って斬新なカリキュラムを作成しました。

講座名：「ビジネス数学・統計学基礎講座」

開催予定日時：2018年2月から毎月開講（2月分は満席）

次回開講予定は3月7日（水）から開講予定

開講教室：東京、大阪（予定）

日程：2時間×4回

費用：15万円/人※ユーザー企業・パートナーには割引制度がございます。

定員：各会場10名講座の詳細情報、お申込みについてはデータビークルWEBサイト上で順次公開予定です。

※講座の内容・日程・費用などは予告なく変更の可能性があります。

### ■ 和から株式会社について

2010年に数学個別指導教室「大人のための数学教室和（なごみ）」の運営からスタートした和から株式会社は数学が苦手な大人から数学の業務・研究応用を目的としているマーケター、経営者、大学教授まで月間400名（2016年3月現在）を超える社会人に対して必要な数学の授業を日々提供しています。人に寄り添う「数学」をテーマに、近年は企業向けの統計学・数学の研修や数学の力を活かした社会問題解決コンサルティングなど様々な領域に活動を広げています。また、数学好き同士が熱く語り合う交流会「ロマンティック数学ナイト」も主催しています。

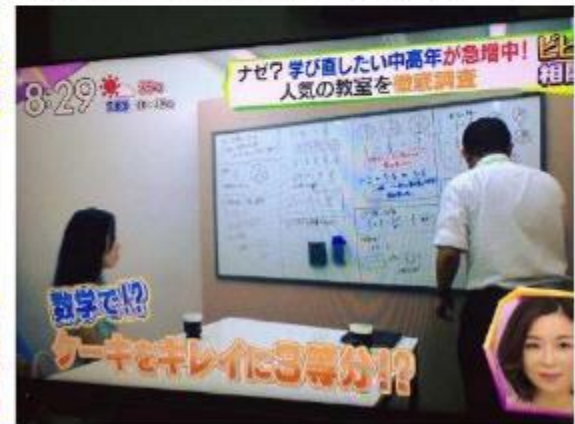
→ Webサイト <https://wakara.co.jp/>

# 和から株式会社 会社概要

## メディア掲載実績



番組名：白熱ライブ ビビット  
(月～金 朝8時～10時)  
出演者：国分太一・真矢ミキ ほか  
放送日：7月13日(月) 朝8時～10時



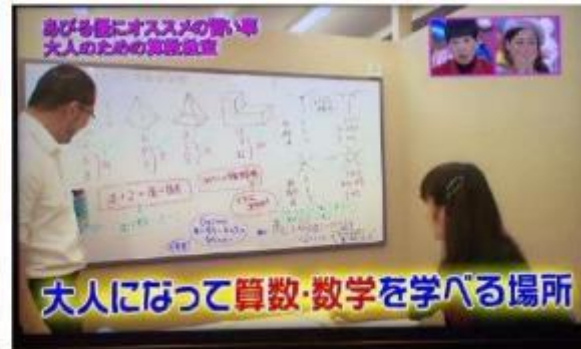


# 和から株式会社 会社概要

## メディア掲載実績



番組名：アッコにおまかせ！  
 （毎週日曜 朝11時45分～12時54分）  
 出演者：和田アキ子・峰竜太 ほか  
 放送日：11月08日（日）朝11時45分



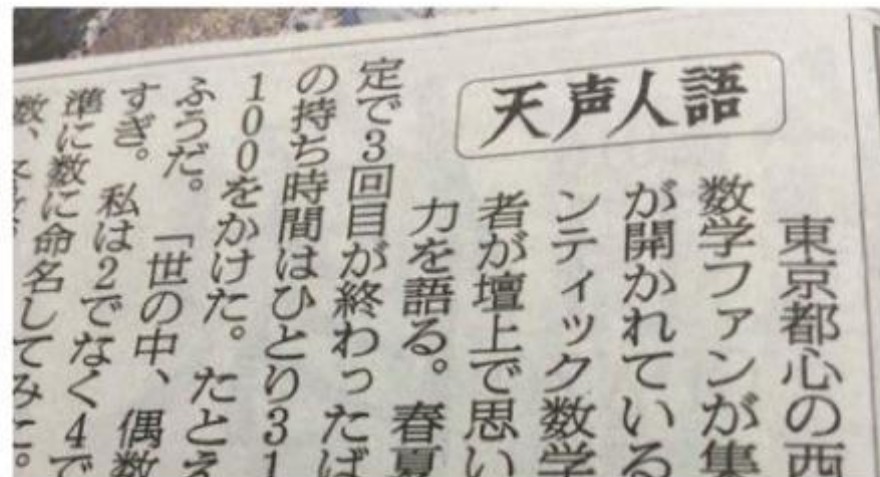


# 和から株式会社 会社概要

## メディア掲載実績

朝日新聞「天声人語」に掲載

朝日新聞「天声人語」にて「ロマンティック数学ナイト」が紹介されました



2016年10月7日の朝日新聞「天声人語」にて、当教室 及び イベント「ロマンティック数学ナイト」の様子が掲載されました。

# 機械学習とは？

コンビニに行け



# 機械学習とは？

コンビニに行け



# 機械学習とは？

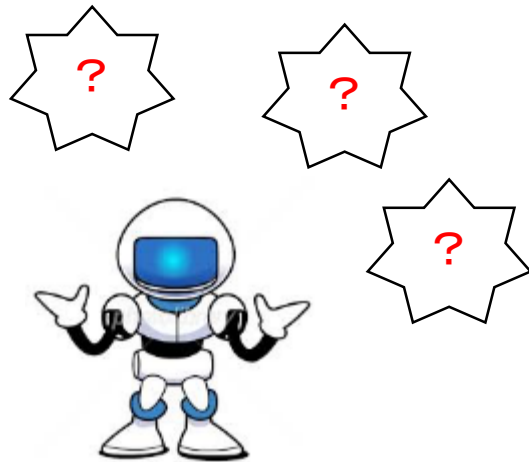
コンビニに行け





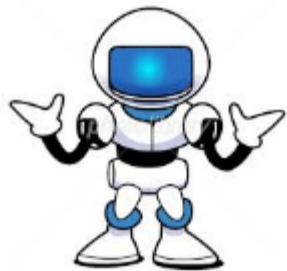
# 機械学習とは？

コンビニに行け



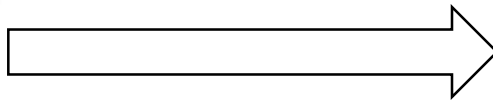
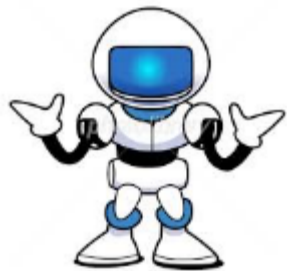
# 機械学習とは？

東に10歩



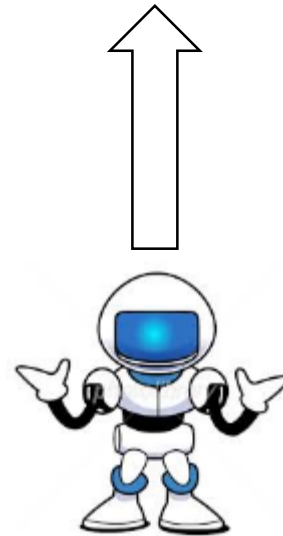
# 機械学習とは？

東に10歩



# 機械学習とは？

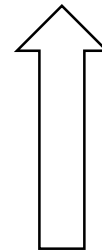
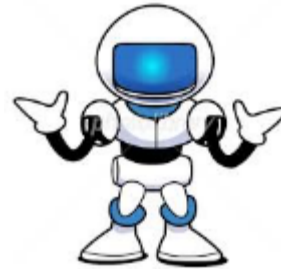
北に5歩





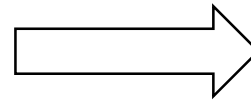
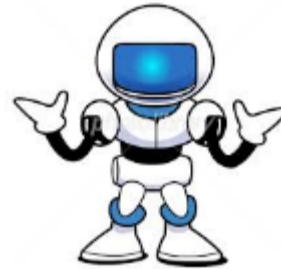
# 機械学習とは？

東に2歩



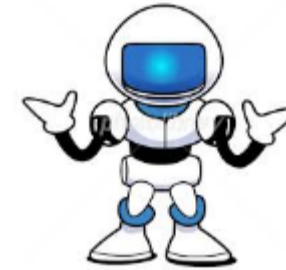
# 機械学習とは？

東に2歩



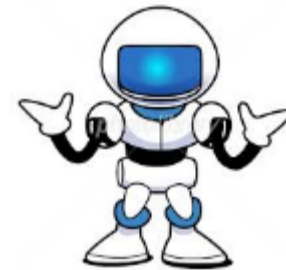
# 機械学習とは？

東に2歩



# 機械学習とは？

東に2歩



汎用性を持たない

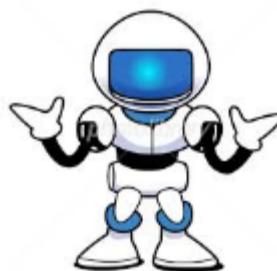


# 機械学習とは？

東に10歩  
北に5歩  
東に2歩



スタート点が違うと通用しない

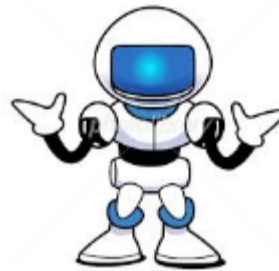


# 機械学習とは？

?

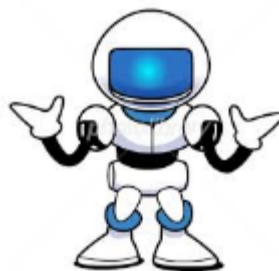


どんな質問をしたらいいか？



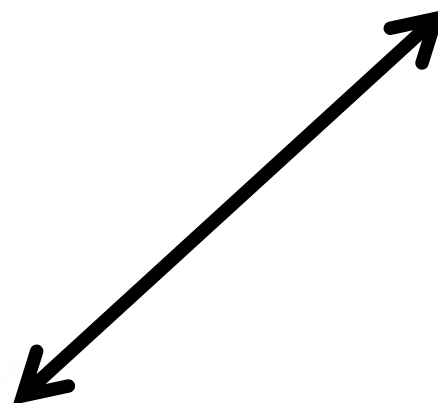
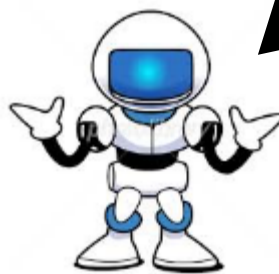
# 機械学習とは？

1 コンビニまでの距離を計算せよ



# 機械学習とは？

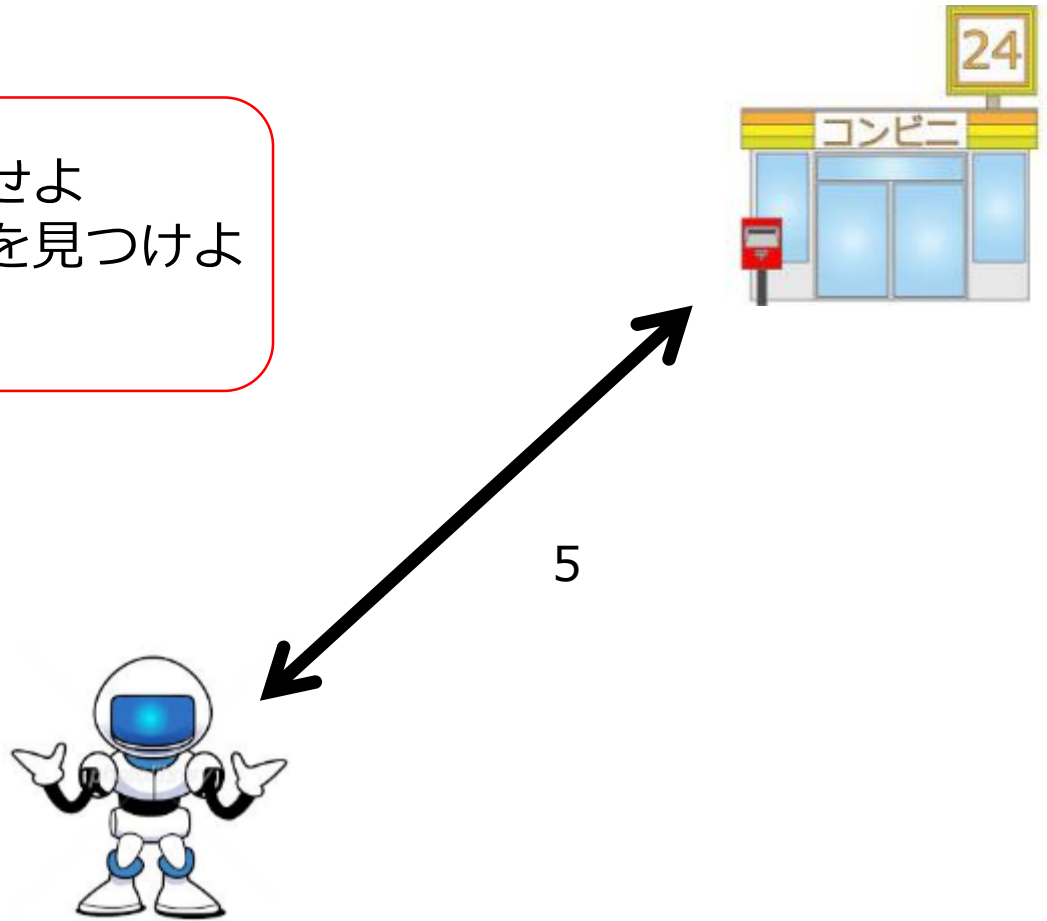
1 コンビニまでの距離を計算せよ





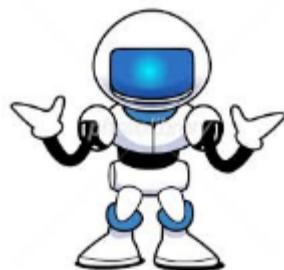
# 機械学習とは？

- 1 コンビニまでの距離を計算せよ  
2 動き回り、距離が減る方向を見つけよ



# 機械学習とは？

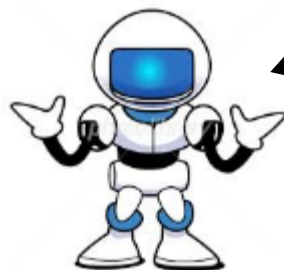
- 1 コンビニまでの距離を計算せよ  
2 動き回り、距離が減る方向を見つけよ



7.5

# 機械学習とは？

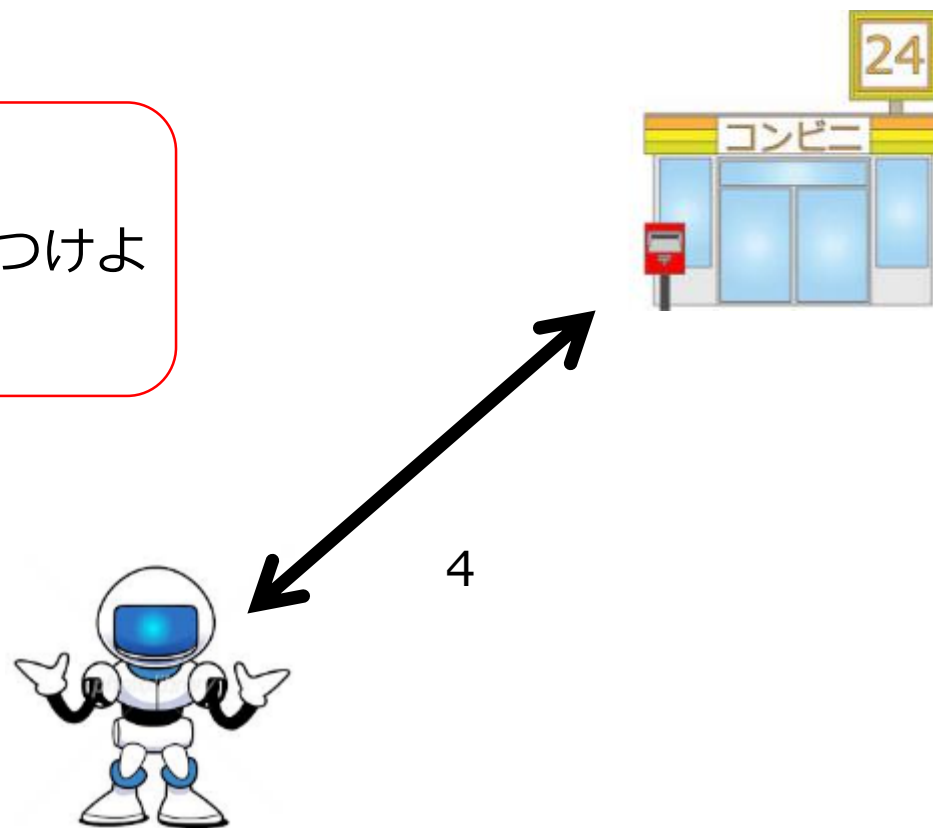
- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ



6.5

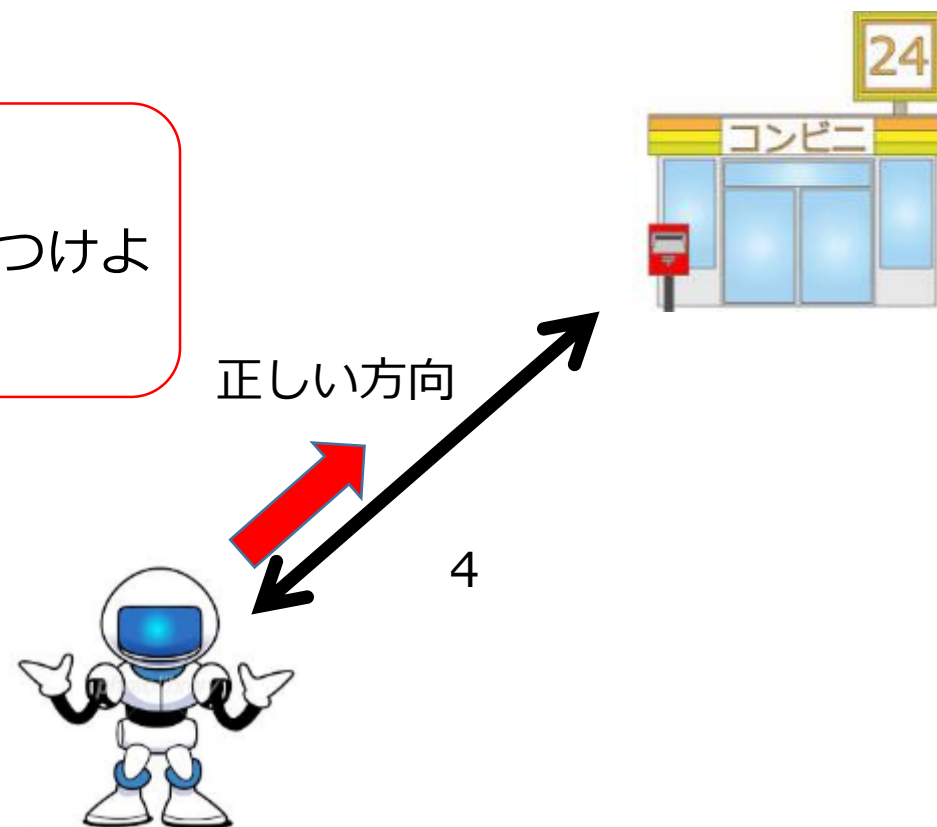
# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ



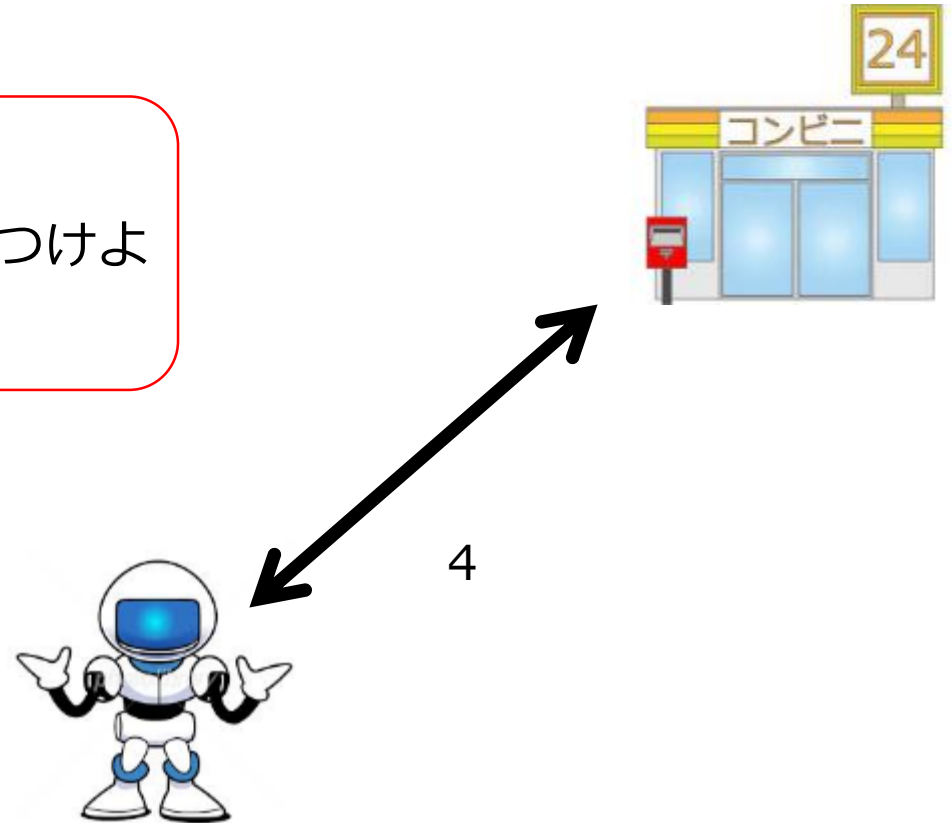
# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ



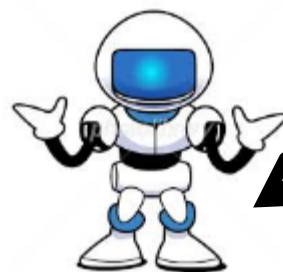
# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ
- 3 2を繰り返せ



# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ
- 3 2を繰り返せ

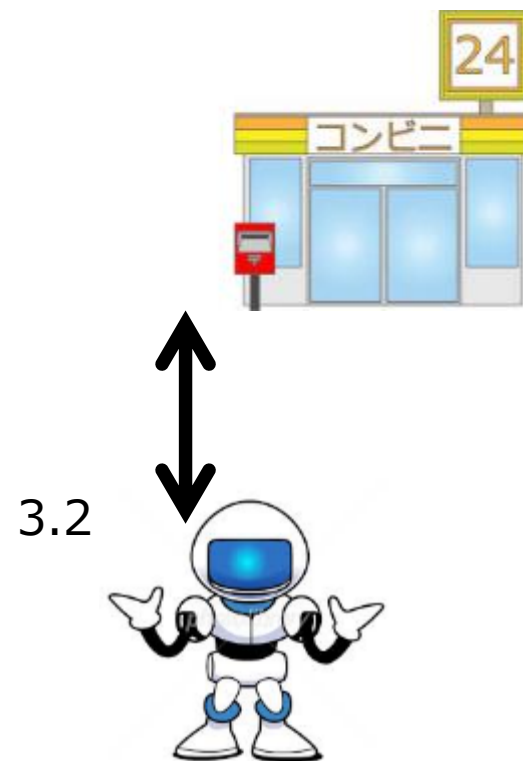


4.2



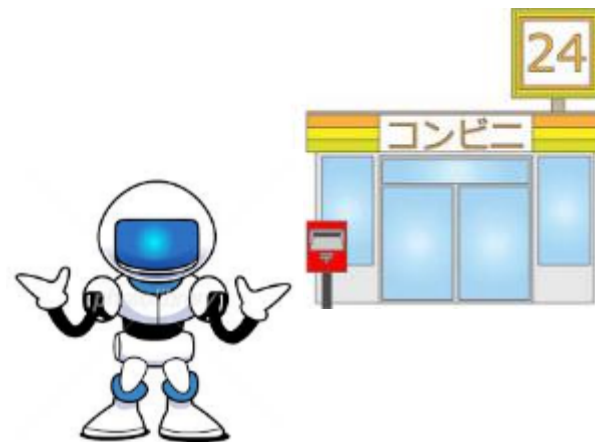
# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ
- 3 2を繰り返せ



# 機械学習とは？

- 1 コンビニまでの距離を計算せよ
- 2 動き回り、距離が減る方向を見つけよ
- 3 2を繰り返せ



# 機械学習とは？

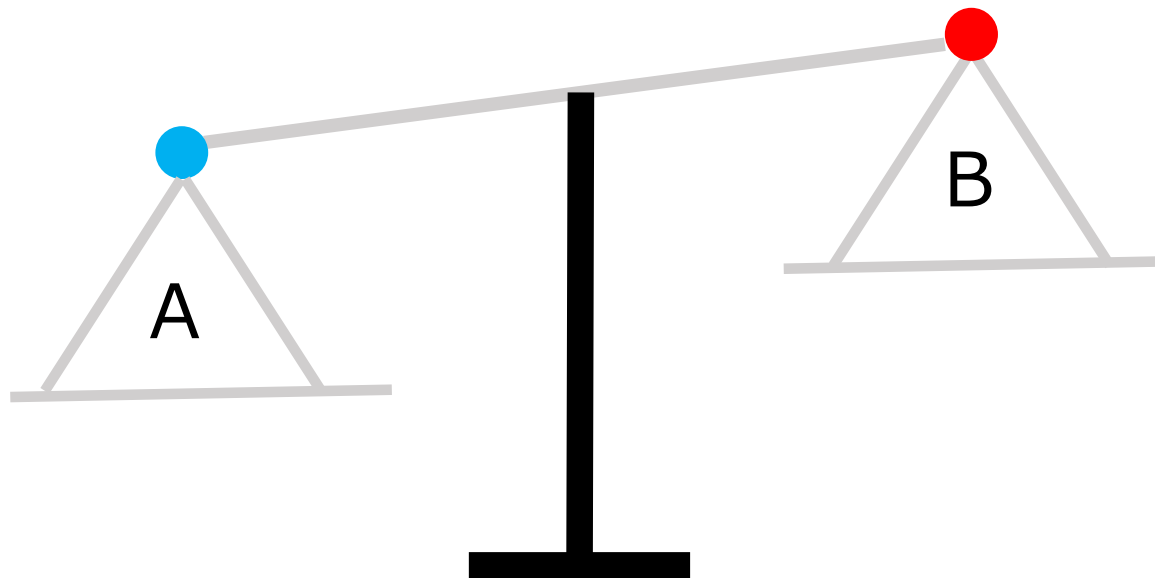


**問題を定式化し、その中で誤差を最小化する**

# 機械学習で問われる 3つの質問

- 質問 1  
「AかBか？」

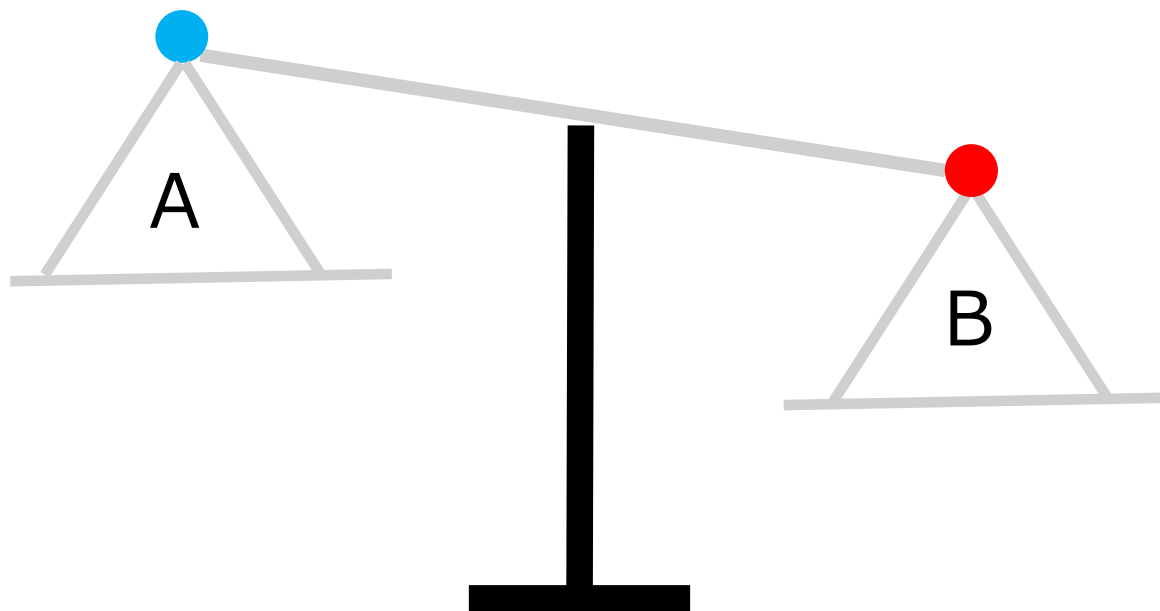
識別アルゴリズム



# 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

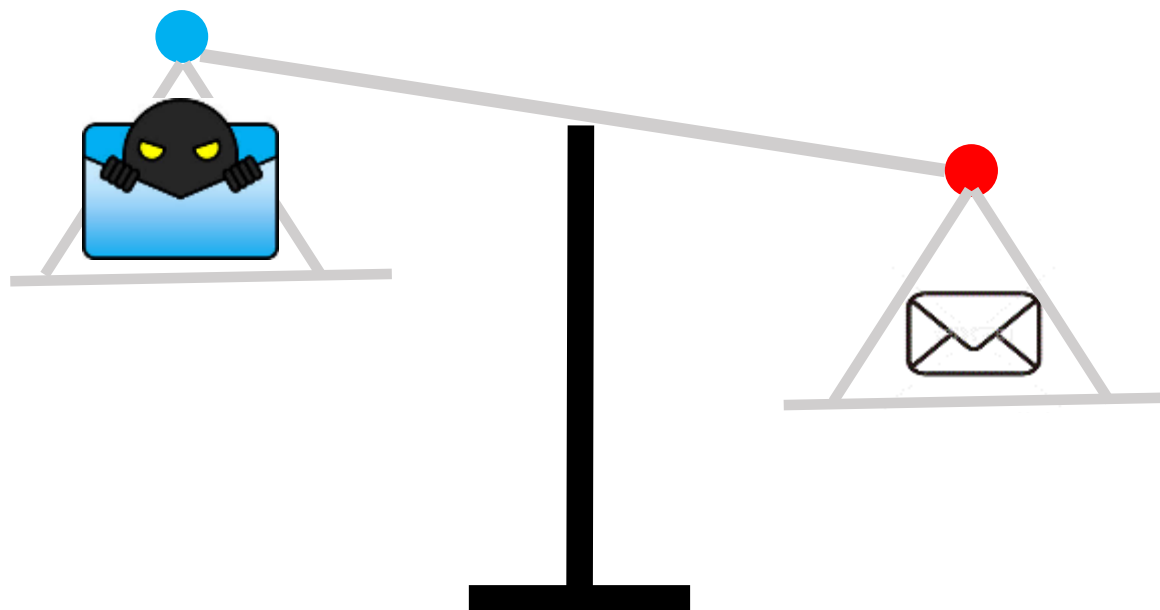


# 機械学習で問われる 3つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

ナイーブベイズ

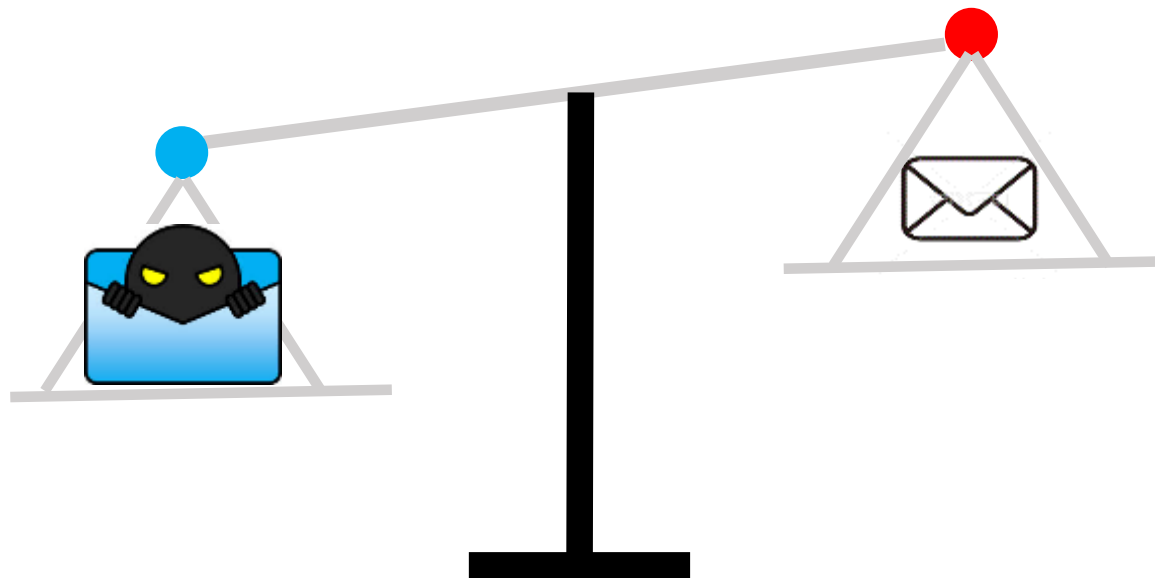


# 機械学習で問われる 3つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

ナイーブベイズ



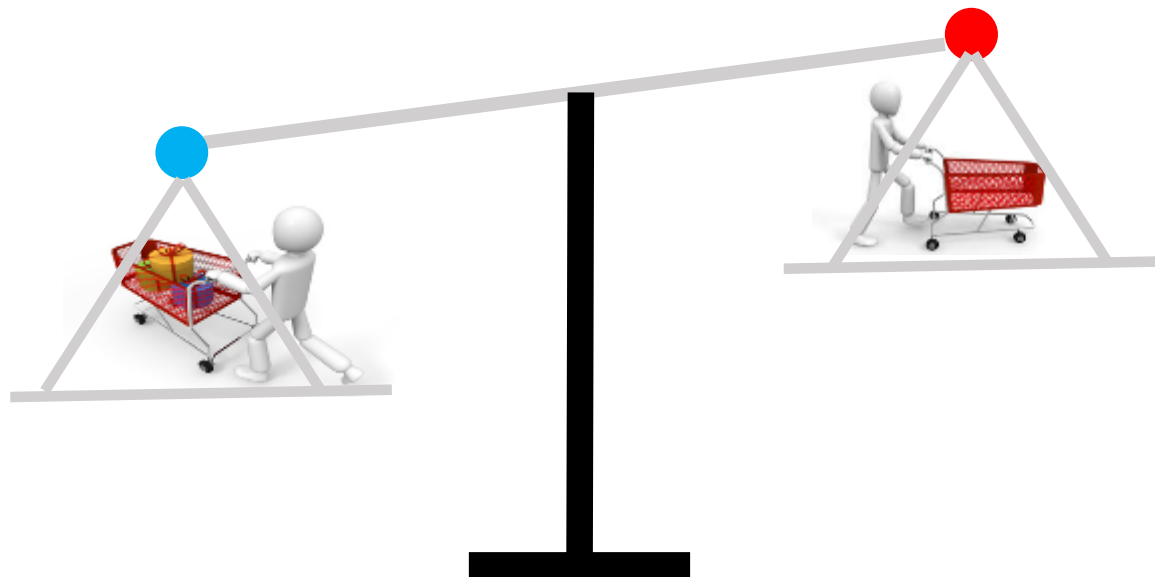


# 機械学習で問われる 3 つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

決定木

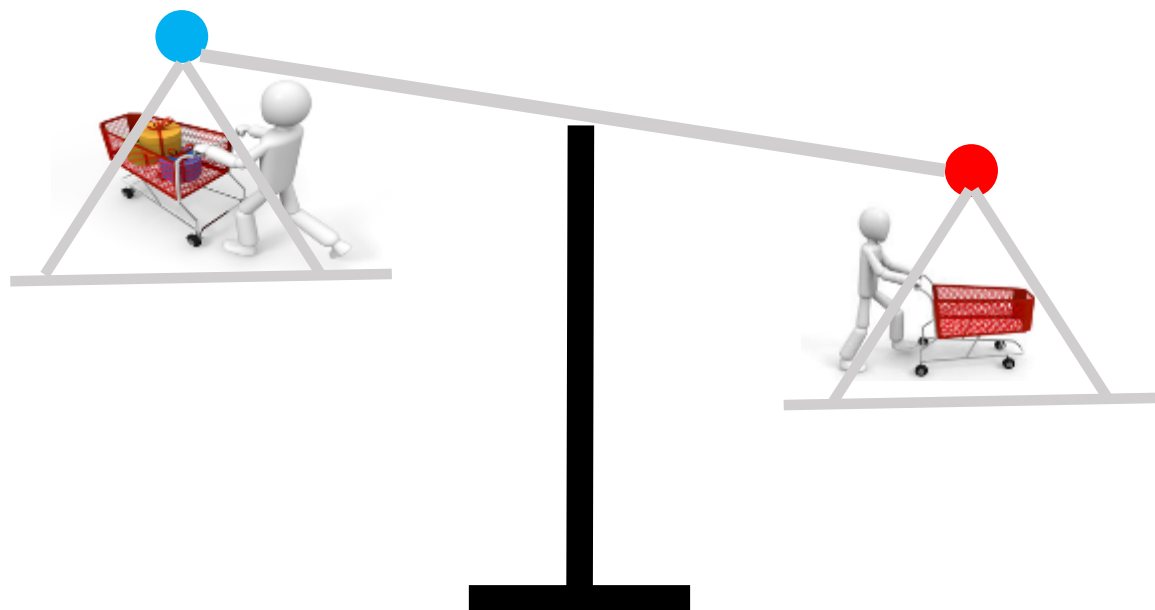


# 機械学習で問われる 3 つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

決定木



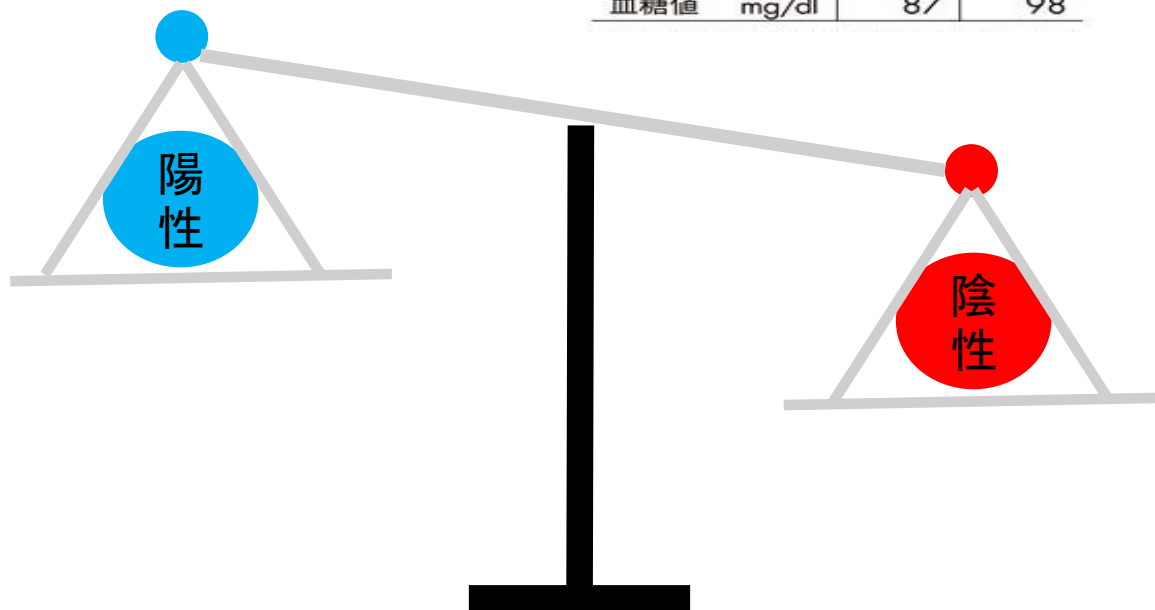
# 機械学習で問われる3つの質問

- 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



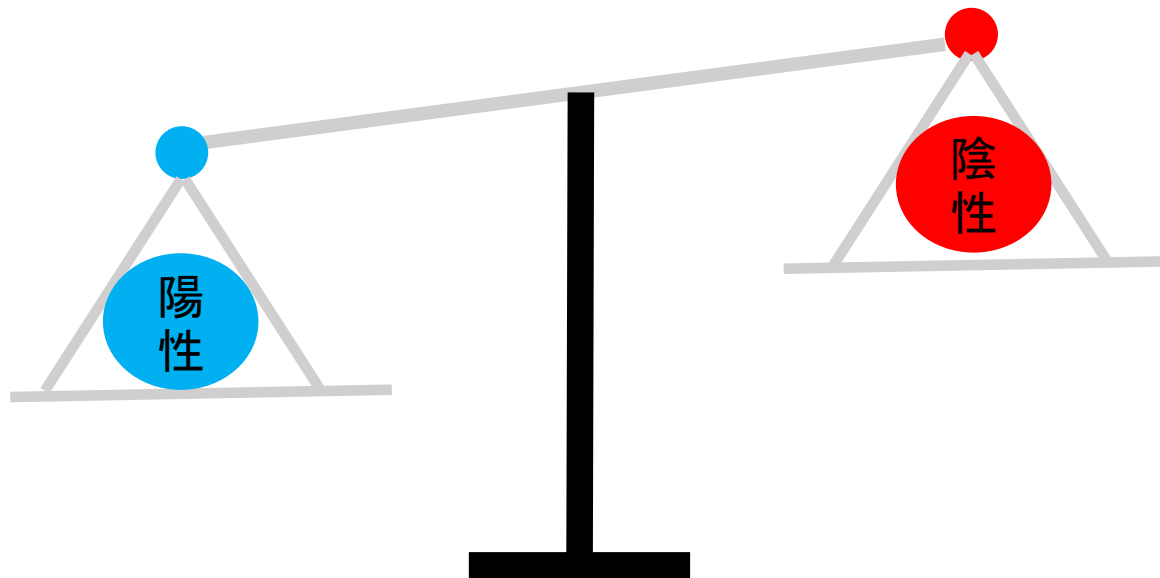
# 機械学習で問われる3つの質問

- 質問 1  
「AかBか？」

識別アルゴリズム

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



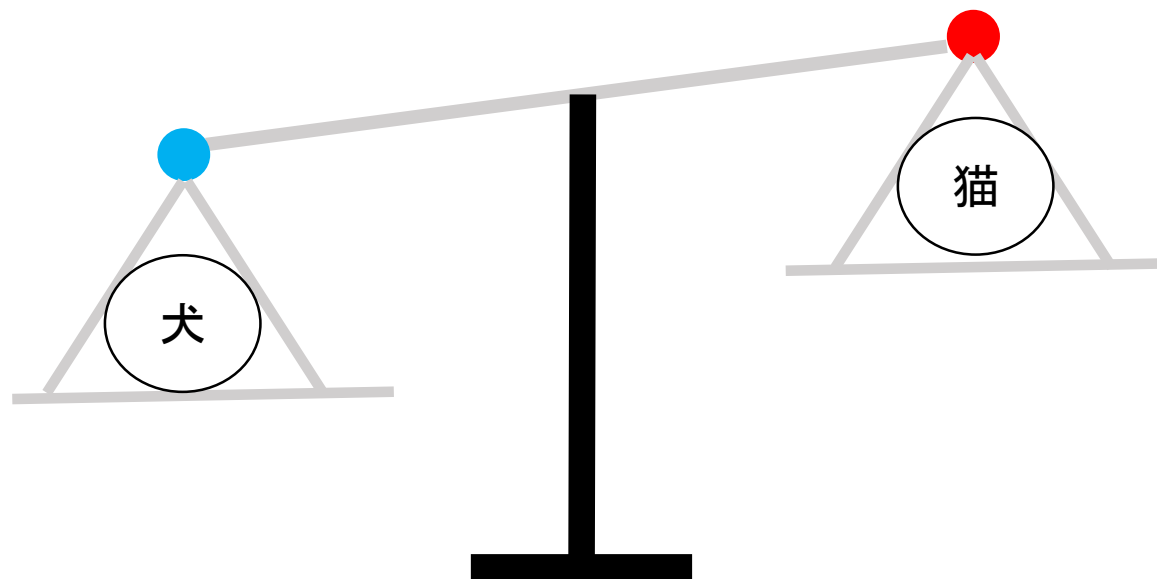
# 機械学習で問われる 3 つの質問

- 質問 1  
「AかBか？」



識別アルゴリズム

Deep learning

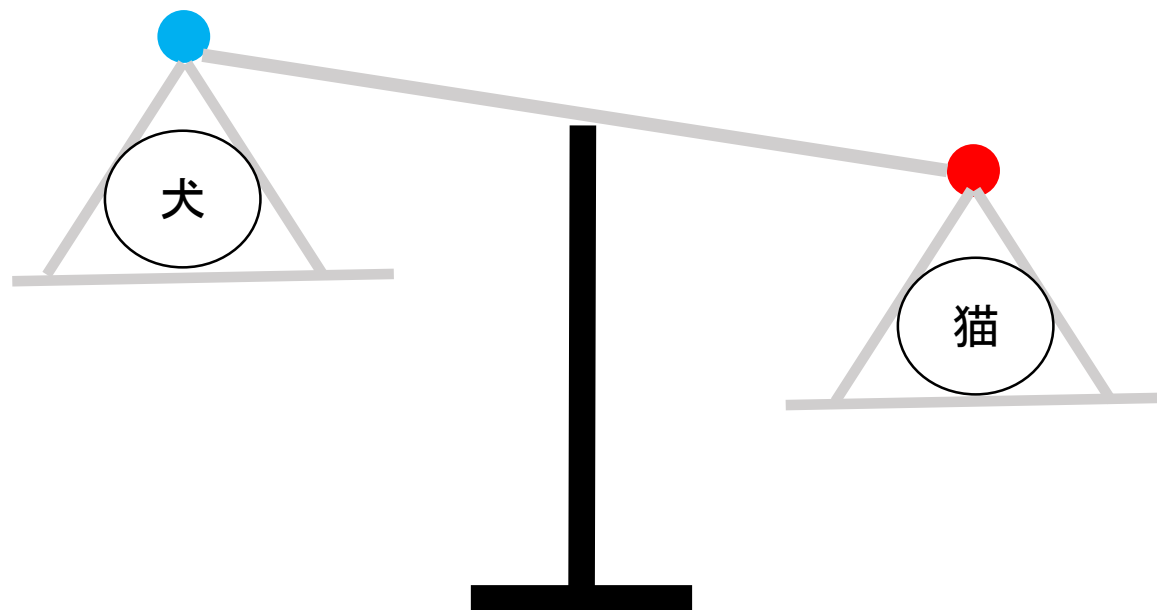


# 機械学習で問われる 3 つの質問

- 質問 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



この物件の価格は？



800万円



2億5千万円

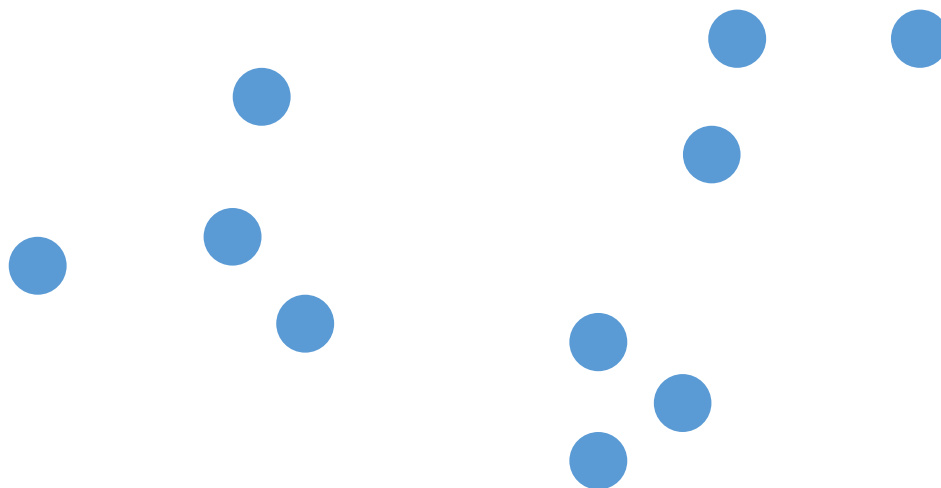
# 機械学習で問われる 3つの質問

- 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？



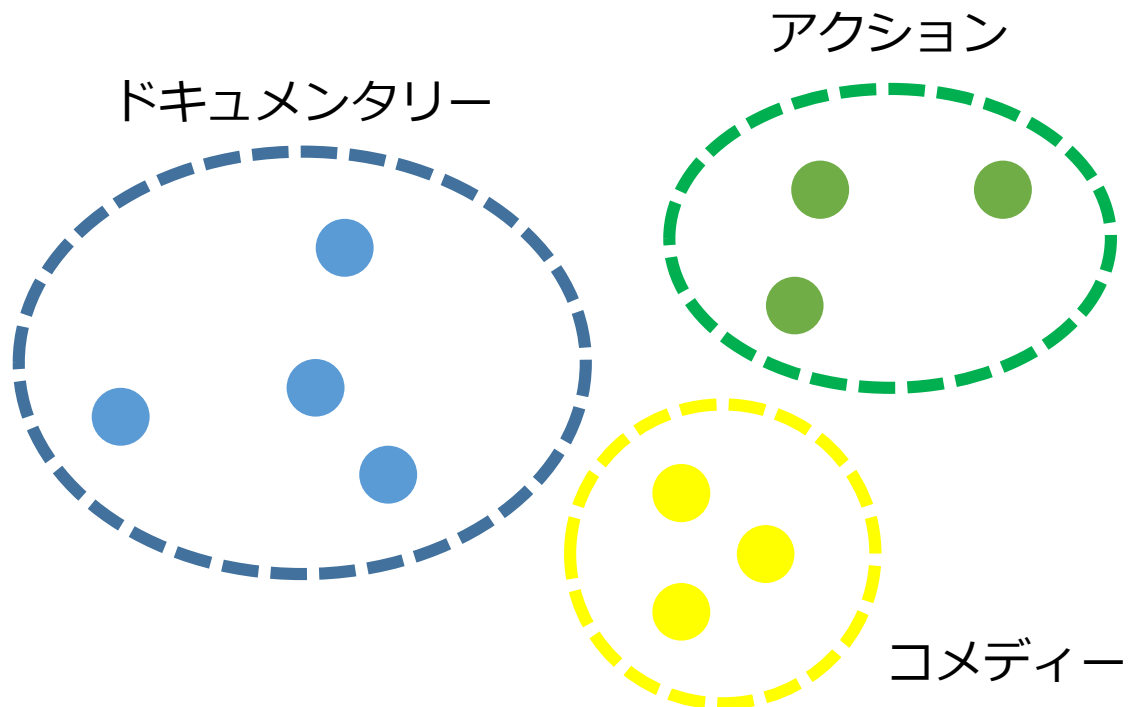
# 機械学習で問われる3つの質問

- 質問3

「どのような編成になっているのか？」

## 分類アルゴリズム

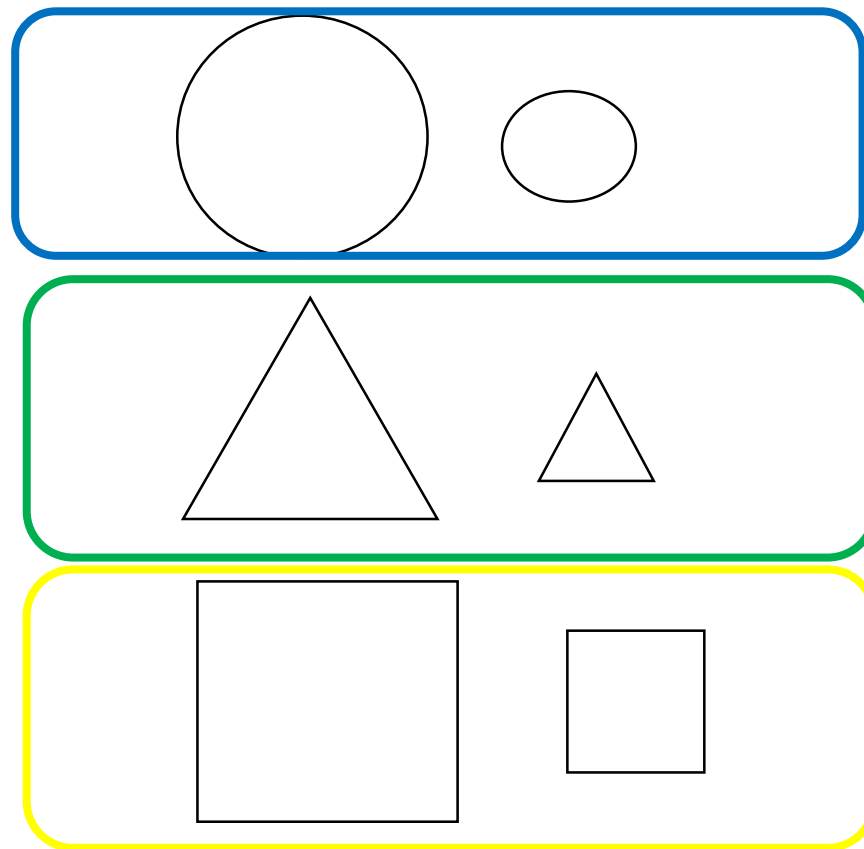
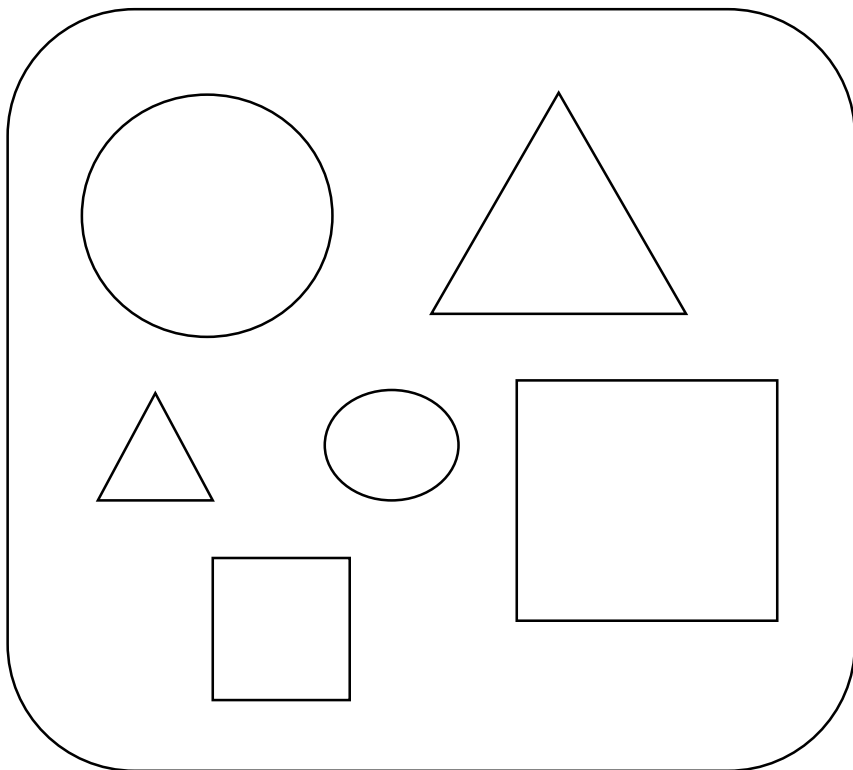
どの視聴者が同じ種類の映画を好むか？



# 機械学習で問われる 3つの質問

- 質問 3

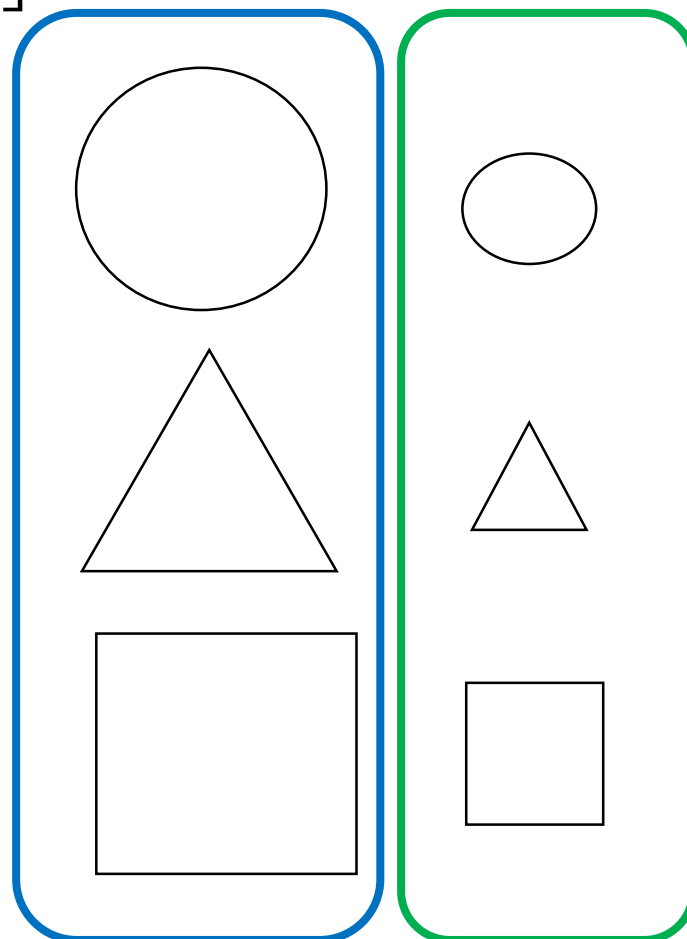
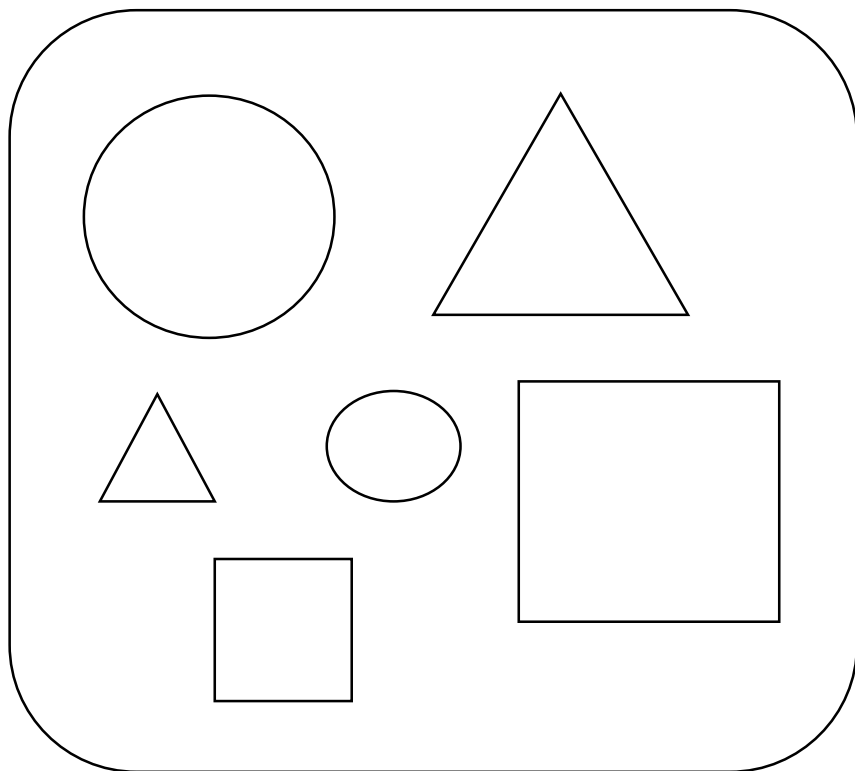
「どのような編成になっているのか？」



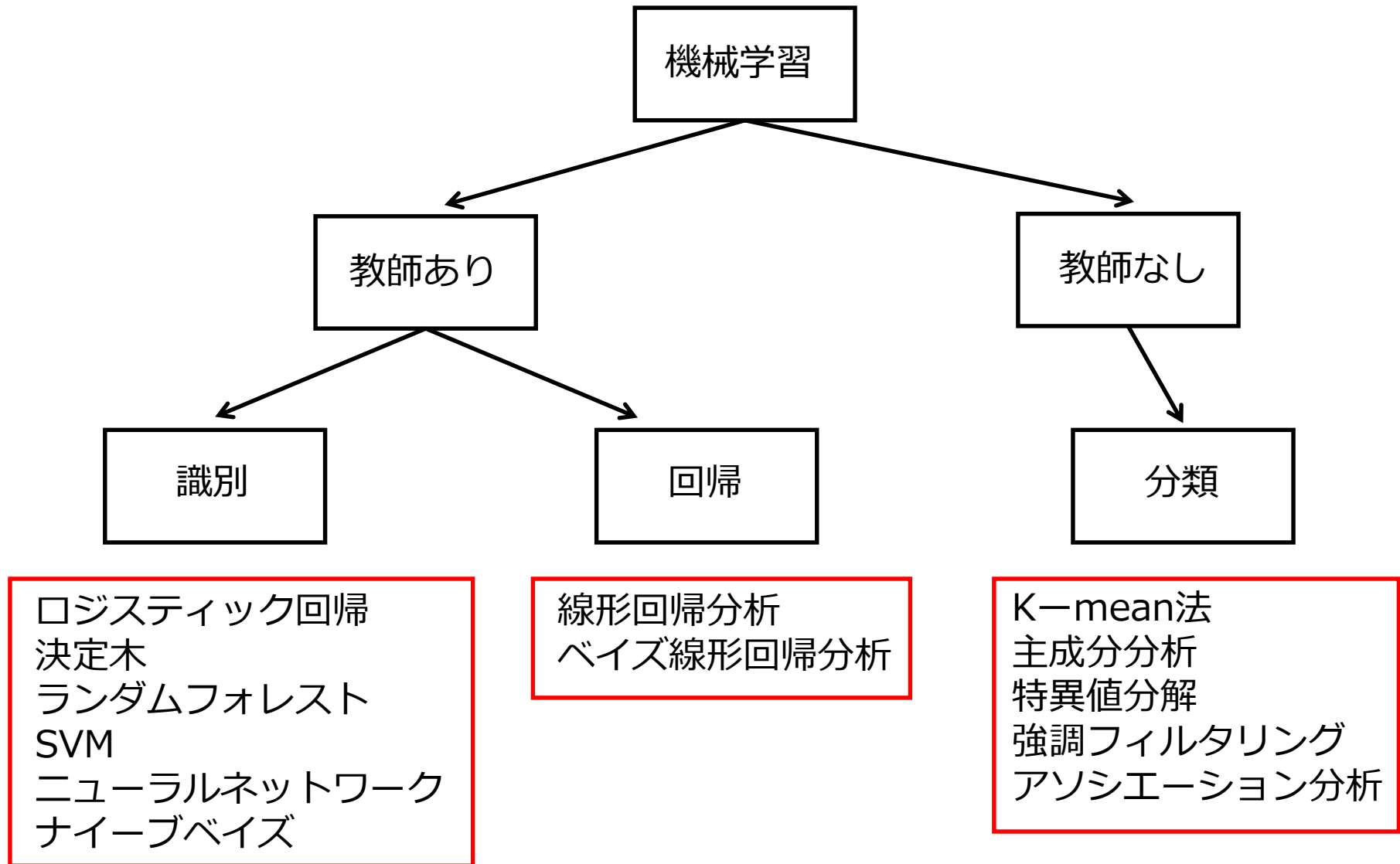
# 機械学習で問われる 3つの質問

- 質問 3

「どのような編成になっているのか？」



# 代表的な機械学習の手法



# 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



- 
- ・ 教師あり ・ 機械学習 ・ 識別



# 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

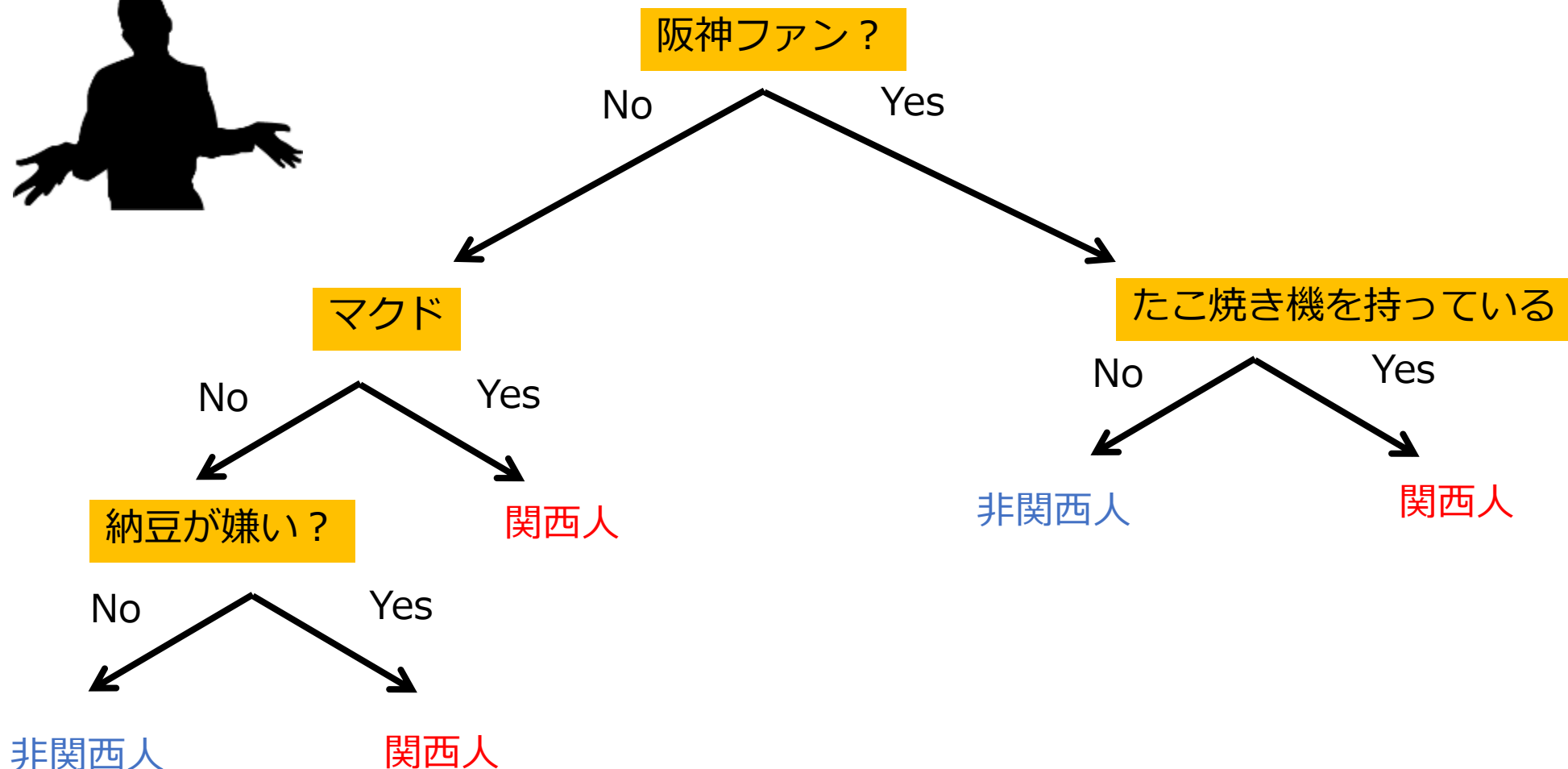


主成分分析



# 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>

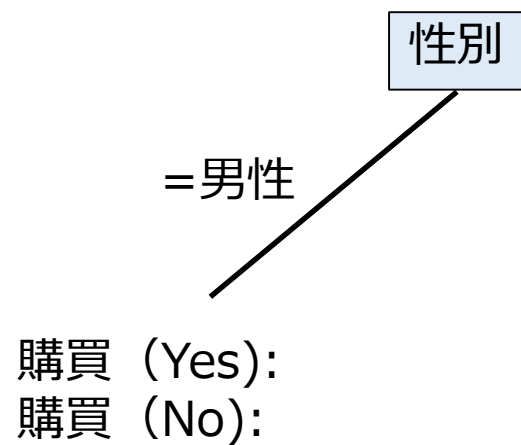


# 問題 決定木による識別

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
男性	No
女性	No
女性	No
男性	Yes
男性	Yes
男性	Yes
女性	Yes
男性	Yes
男性	Yes
女性	Yes

# 性別による分類

性別

=男性

購買 (Yes): 5人

購買 (No):

性別	購買
男性	No
女性	No
女性	No
男性	Yes
男性	Yes
男性	Yes
女性	Yes
男性	Yes
男性	Yes
女性	Yes

# 性別による分類

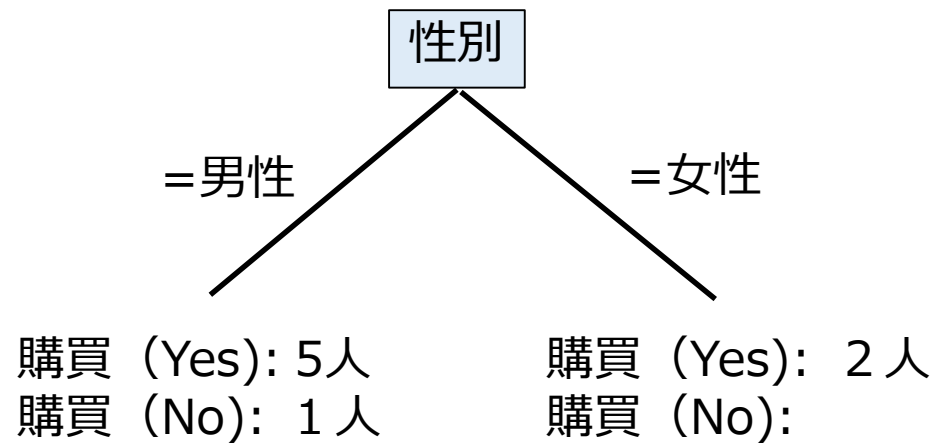
性別

=男性

購買 (Yes): 5人  
購買 (No): 1人

性別	購買
男性	No
女性	No
女性	No
女性	Yes
女性	Yes

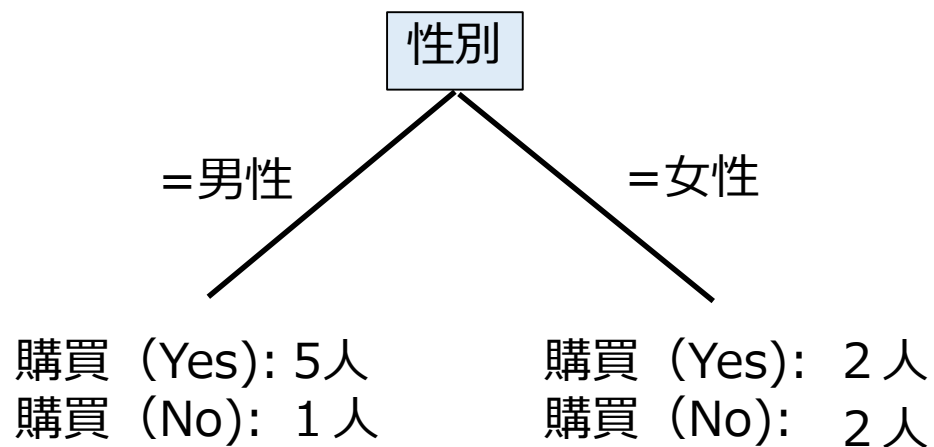
# 性別による分類



性別	購買
女性	No
女性	No
女性	Yes
女性	Yes

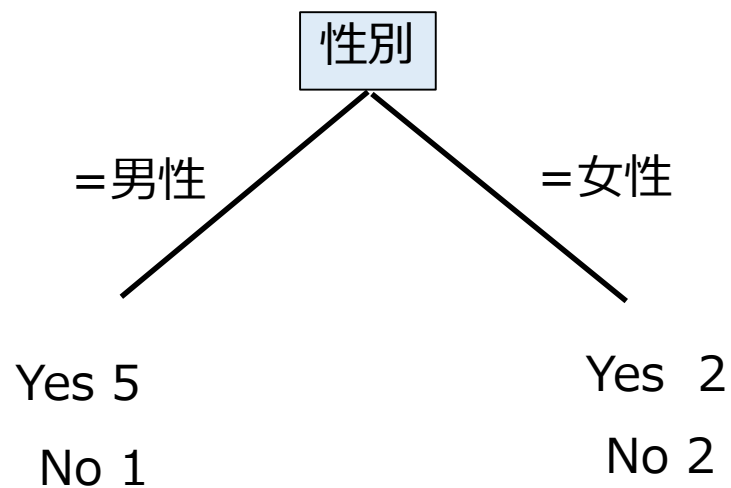


# 性別による分類



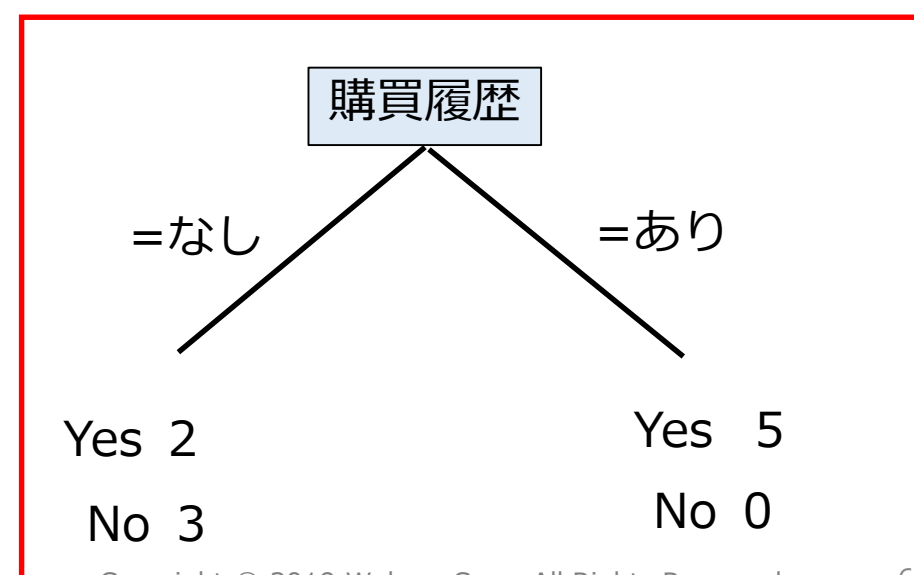
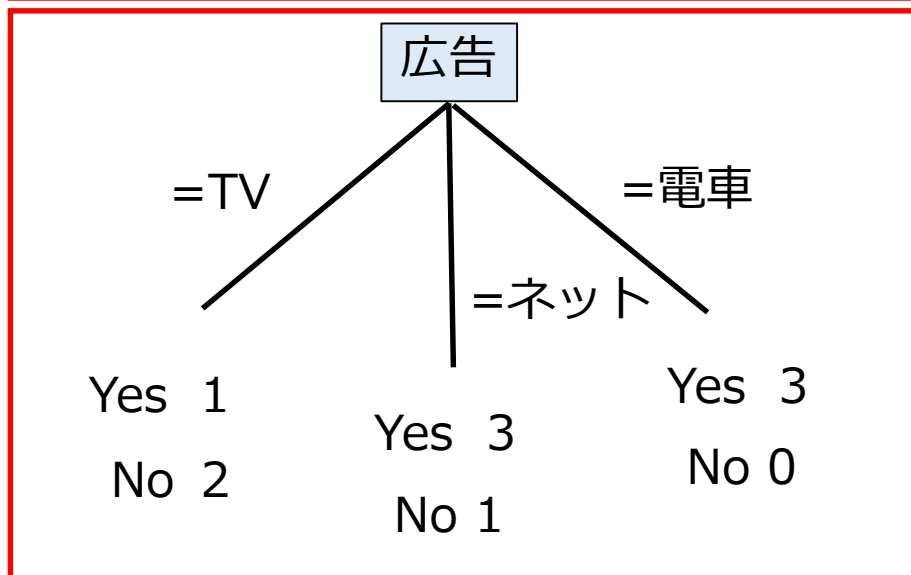
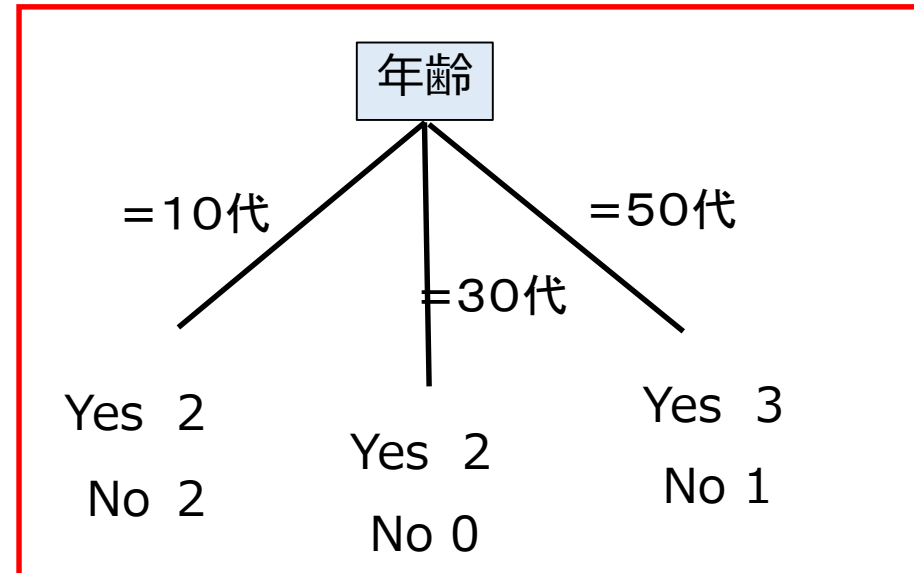
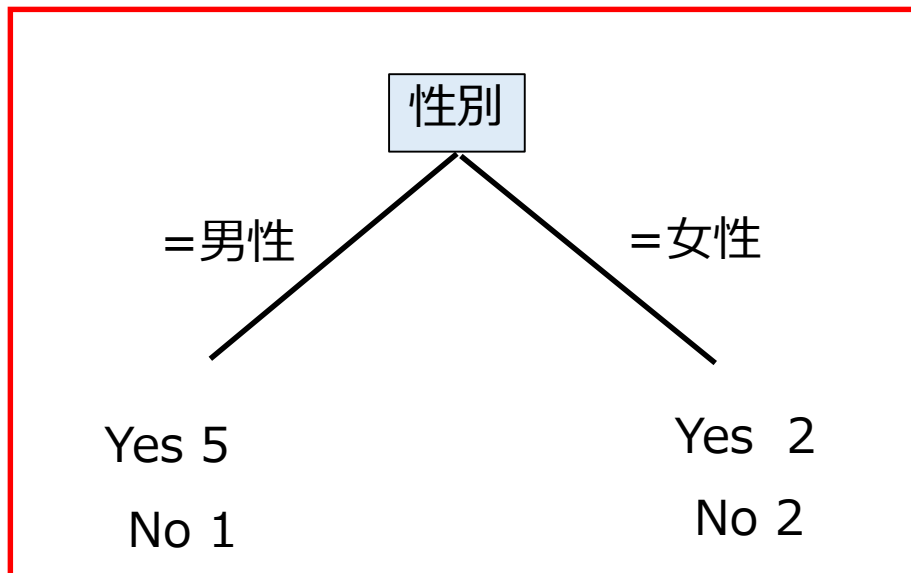
性別	購買
女性	No
女性	No

# 問題：どの変数で分類すべきか？

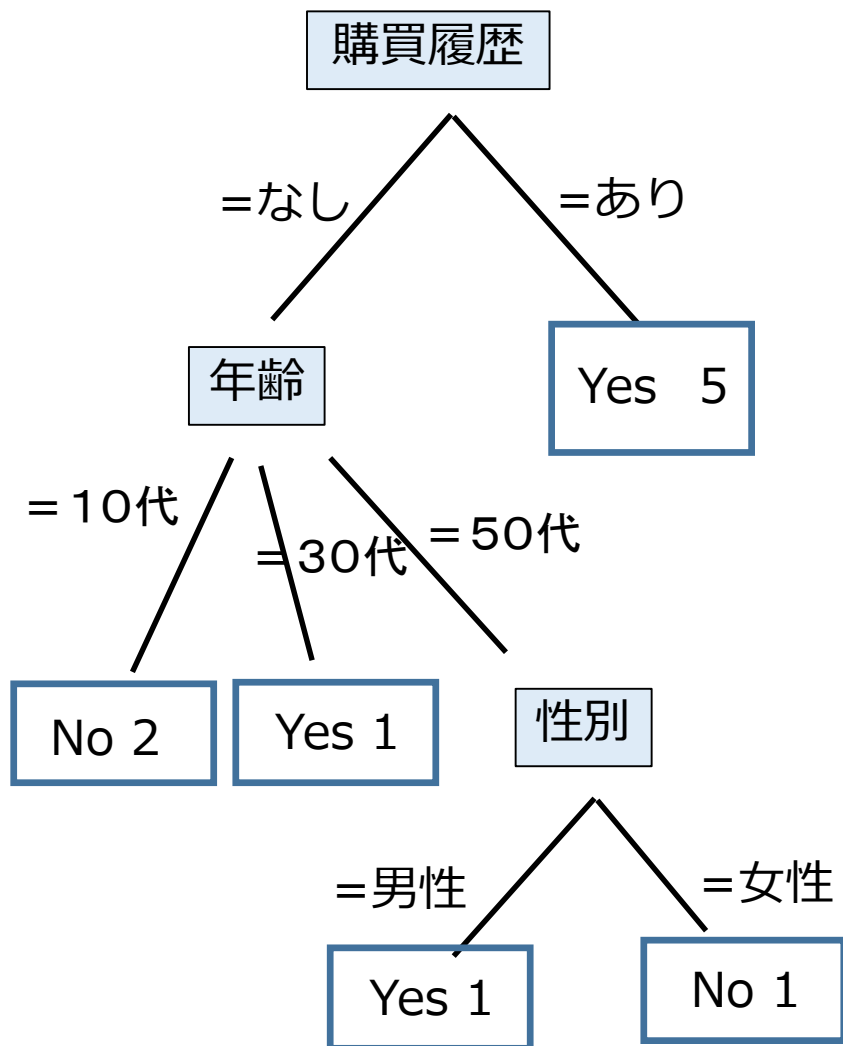


ID	性別	年齢	広告	購買歴	購買
A	男性	10代	TV	無	No
B	女性	10代	TV	無	No
C	女性	50代	ネット	無	No
D	男性	30代	TV	無	Yes
E	男性	50代	電車	有	Yes
F	男性	50代	ネット	無	Yes
G	女性	30代	電車	有	Yes
H	男性	10代	電車	有	Yes
I	男性	50代	ネット	有	Yes
J	女性	10代	ネット	有	Yes

# 問題：どの変数で分類すべきか？



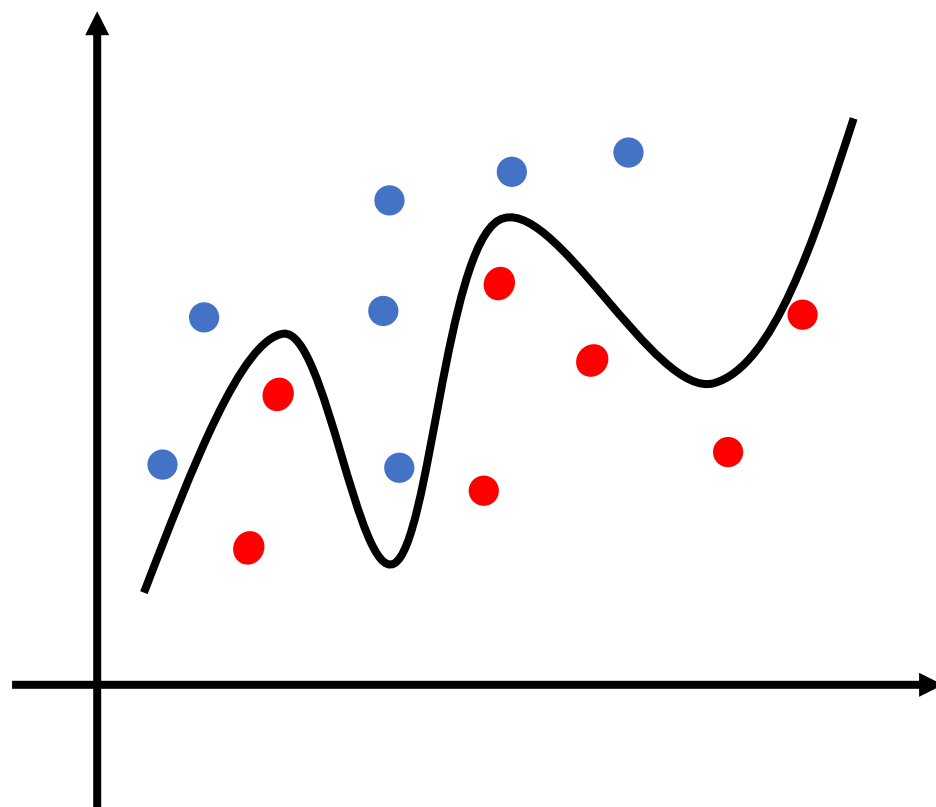
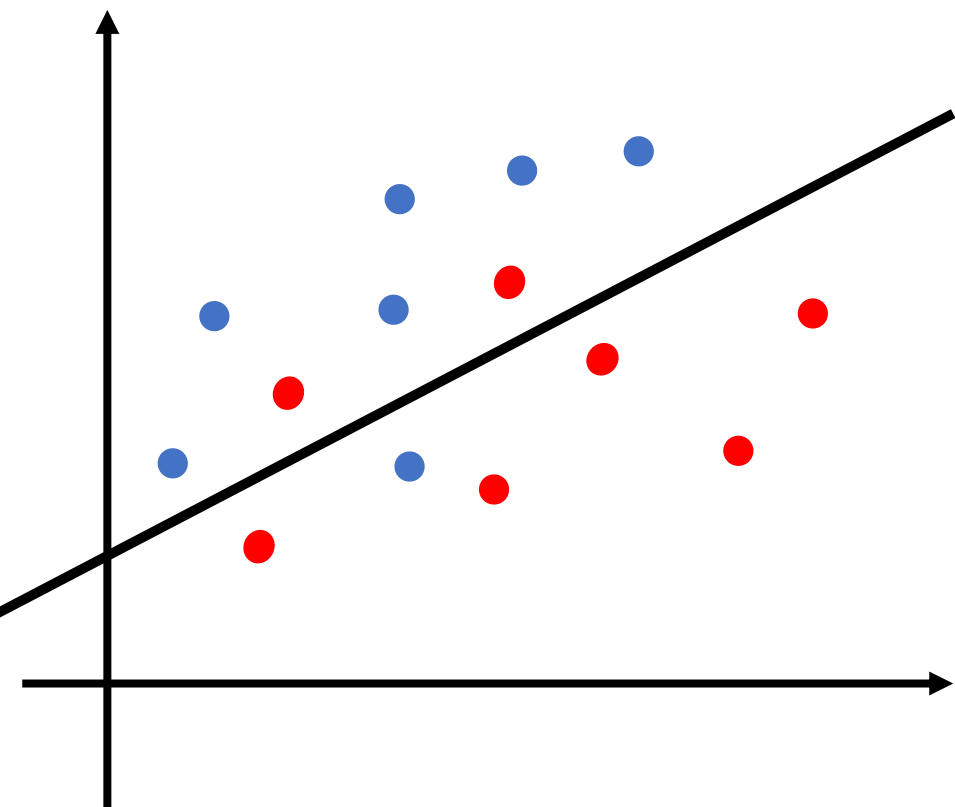
# 決定木による識別



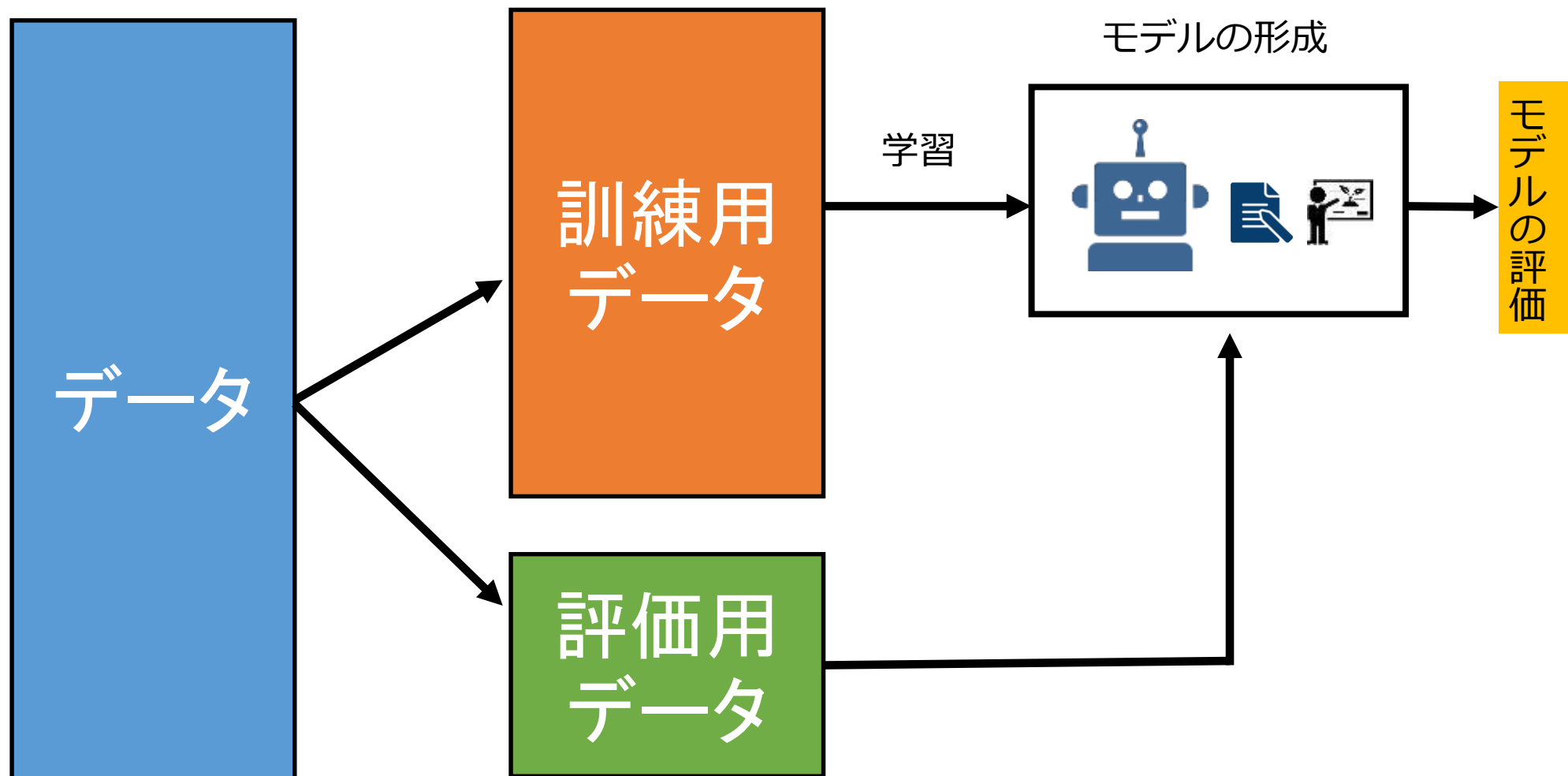
# ランダムフォレスト

- 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？



# 教師あり学習の枠組み

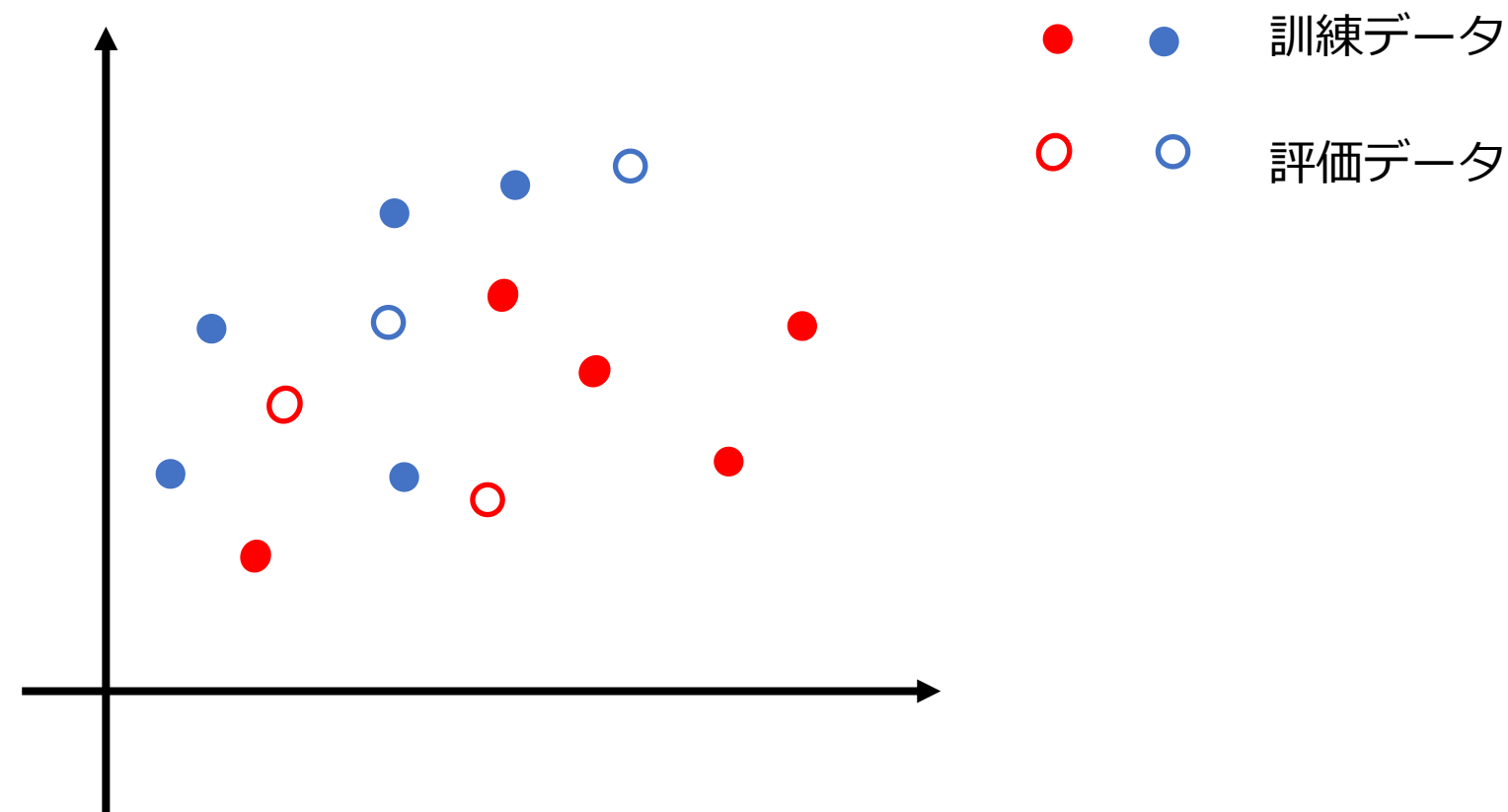


A scatter plot illustrating a linearly separable dataset. The plot shows two classes of data points (blue and red) distributed in a 2D space defined by a horizontal x-axis and a vertical y-axis. The blue points are generally clustered in the upper-left region, while the red points are clustered in the lower-right region, suggesting a linear decision boundary can separate the two classes.

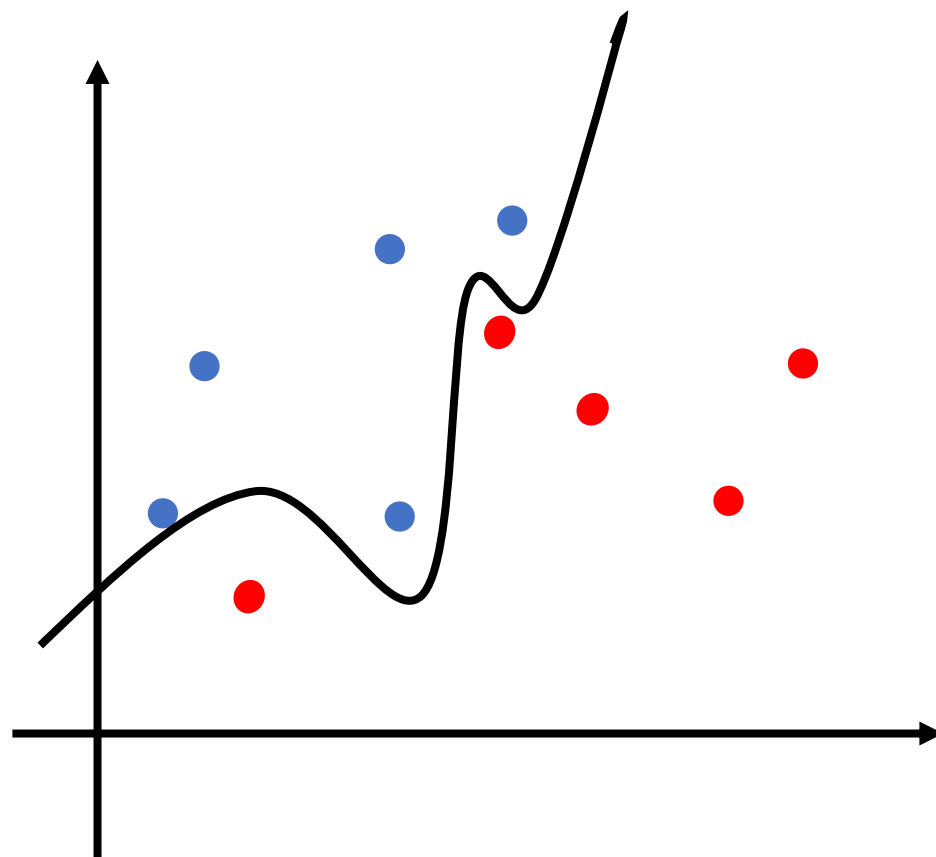
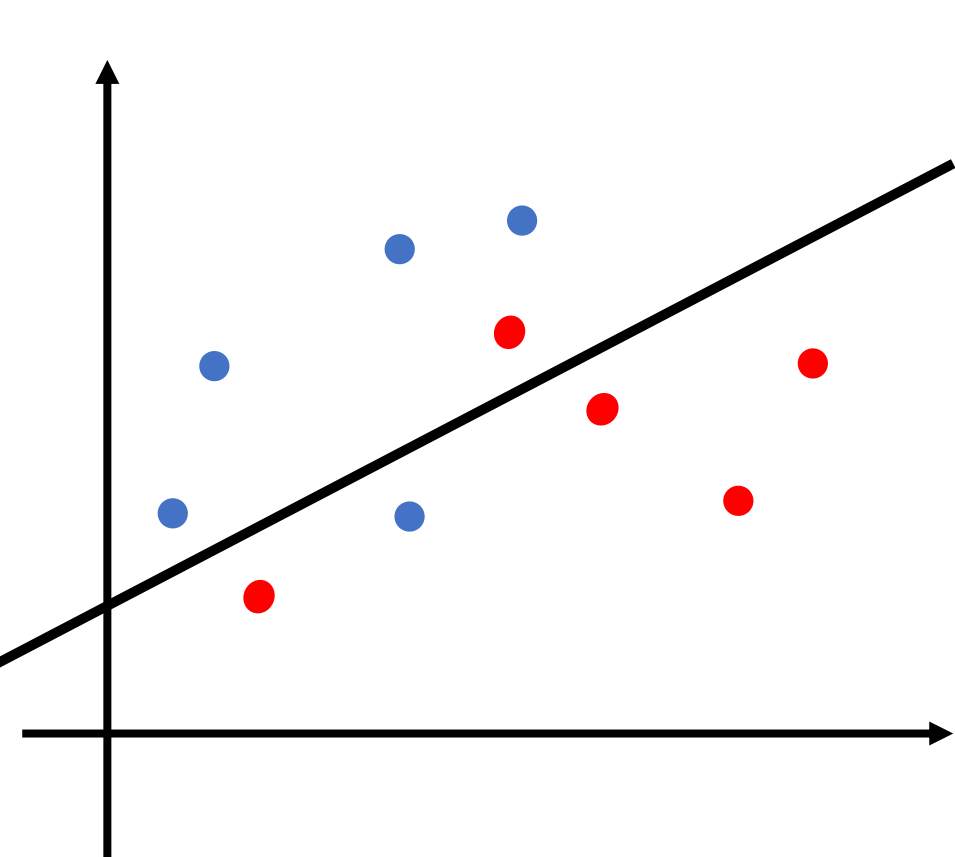
Class	X (approx)	Y (approx)
Blue	1.2	3.5
Blue	2.5	4.2
Blue	3.2	3.8
Blue	3.5	5.5
Blue	4.2	5.8
Blue	4.5	3.5
Red	2.2	2.5
Red	2.8	3.2
Red	3.8	2.8
Red	4.2	4.5
Red	5.2	3.8
Red	6.2	3.2
Red	6.8	4.2



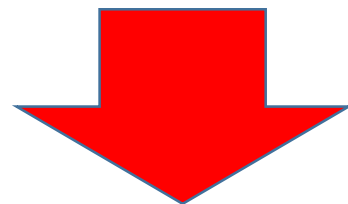
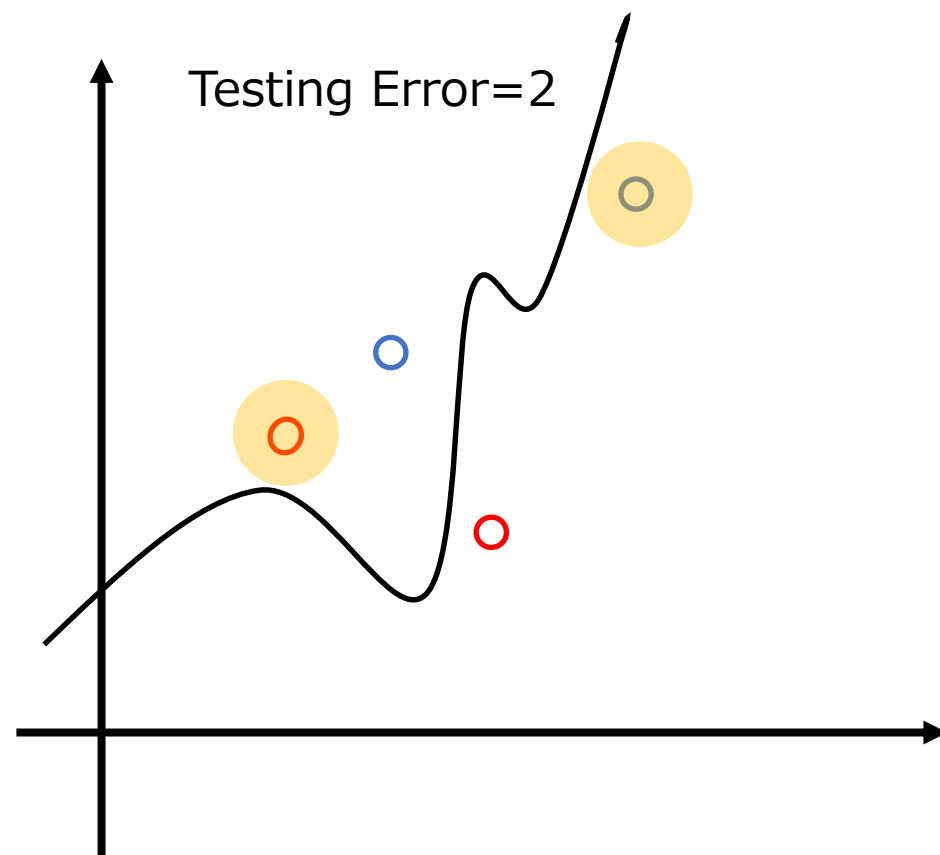
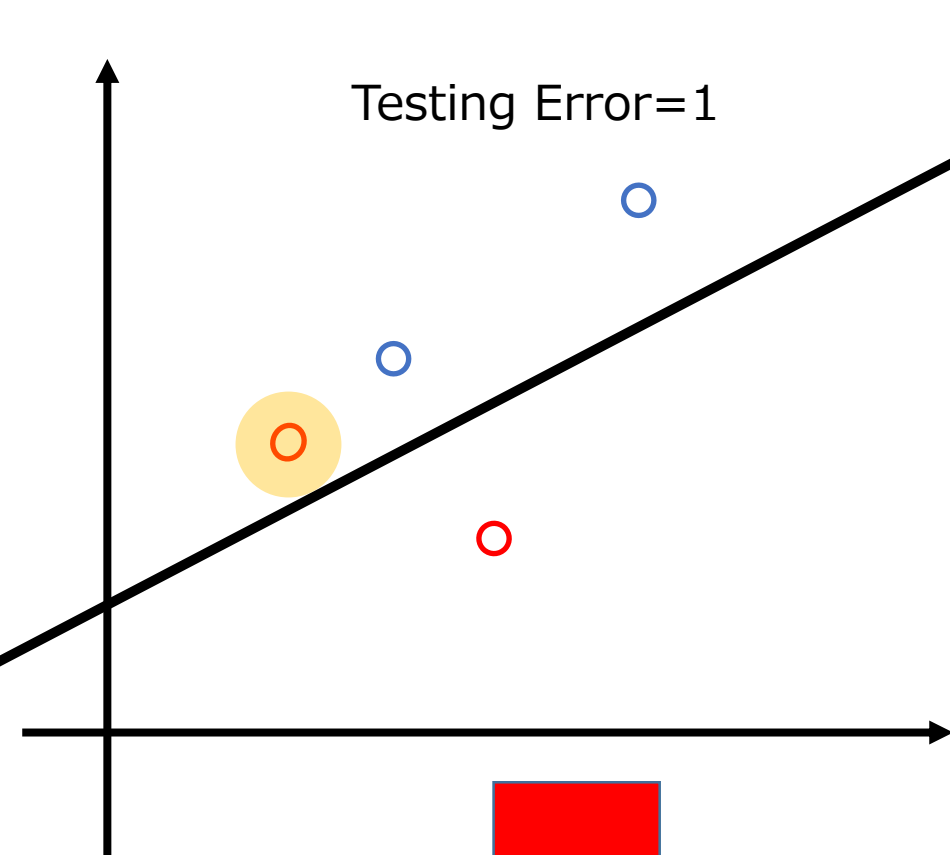
# どっちのモデルを選択すべきか？



# どっちのモデルを選択すべきか？

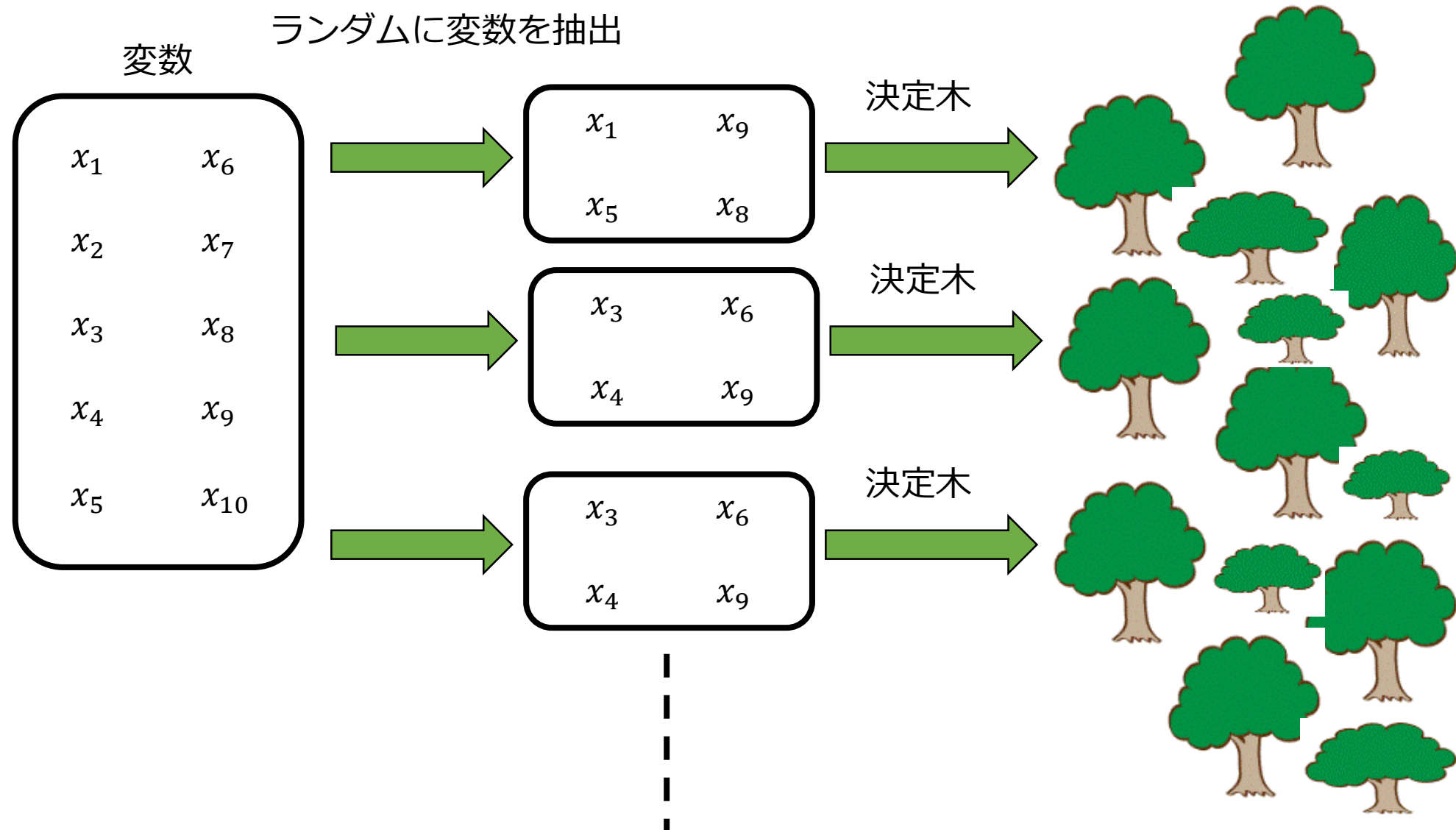


# どっちのモデルを選択すべきか？



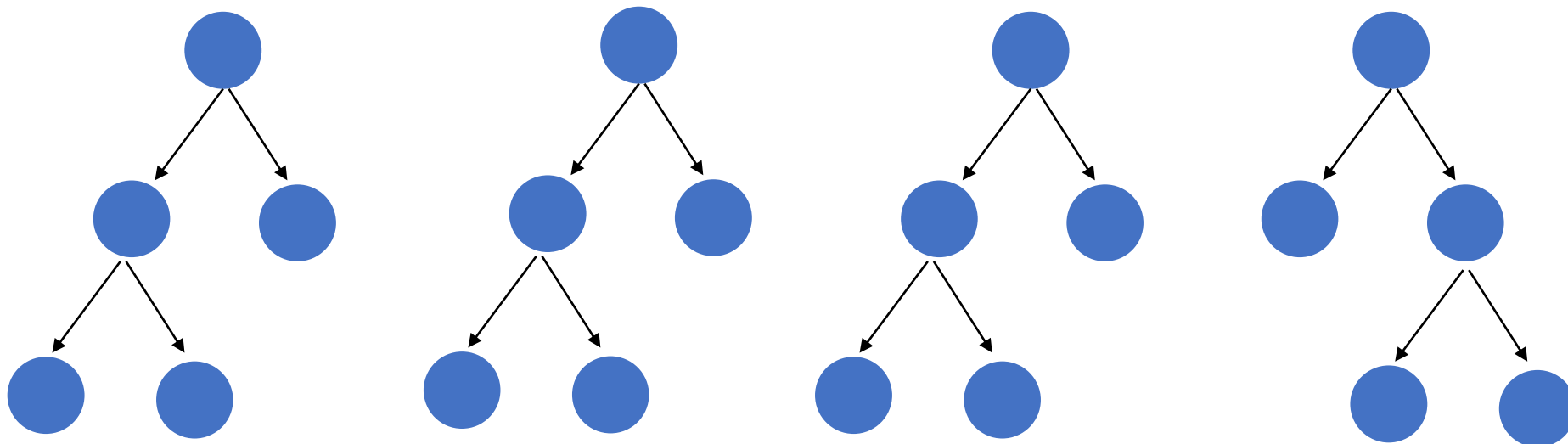
良いモデル

# ランダムフォレスト



# ランダムフォレスト

大抵の決定木は正解を提供している

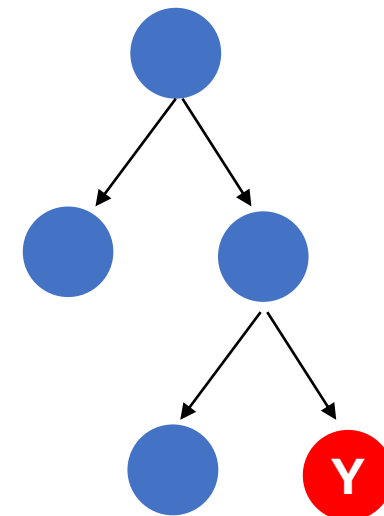
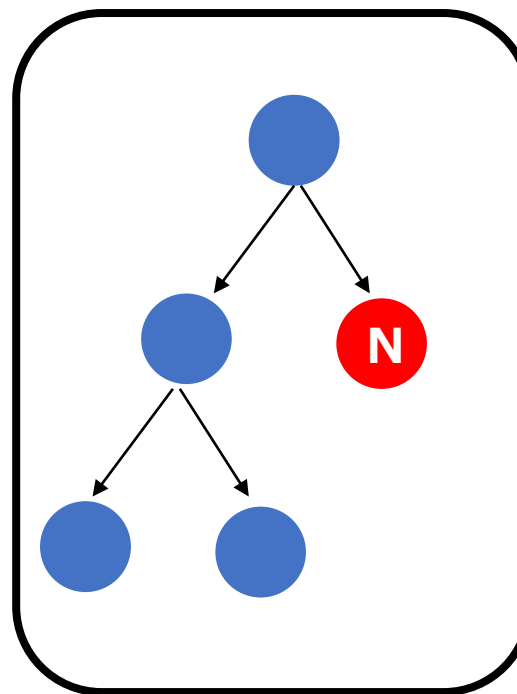
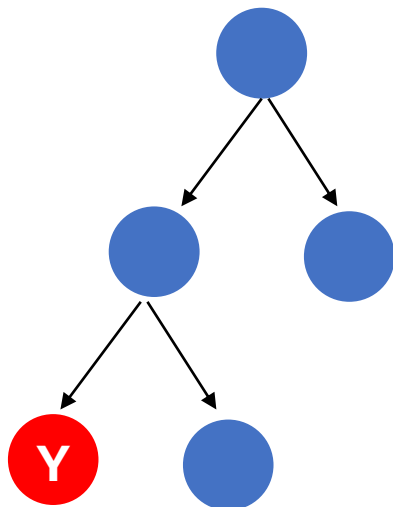
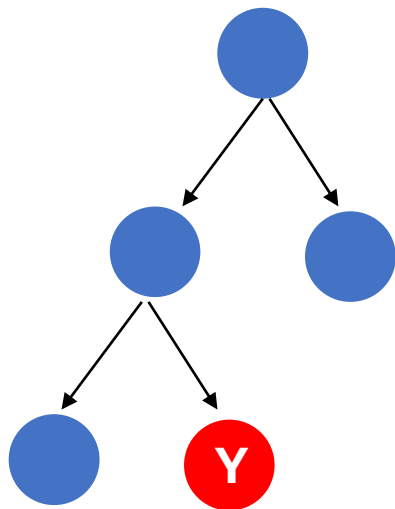


# ランダムフォレスト

購入するかどうか？



「購入する」



多数決の原理

# ざっくり分けるなら

## 機械学習

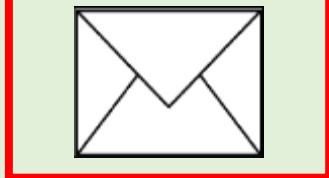
### 識別

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

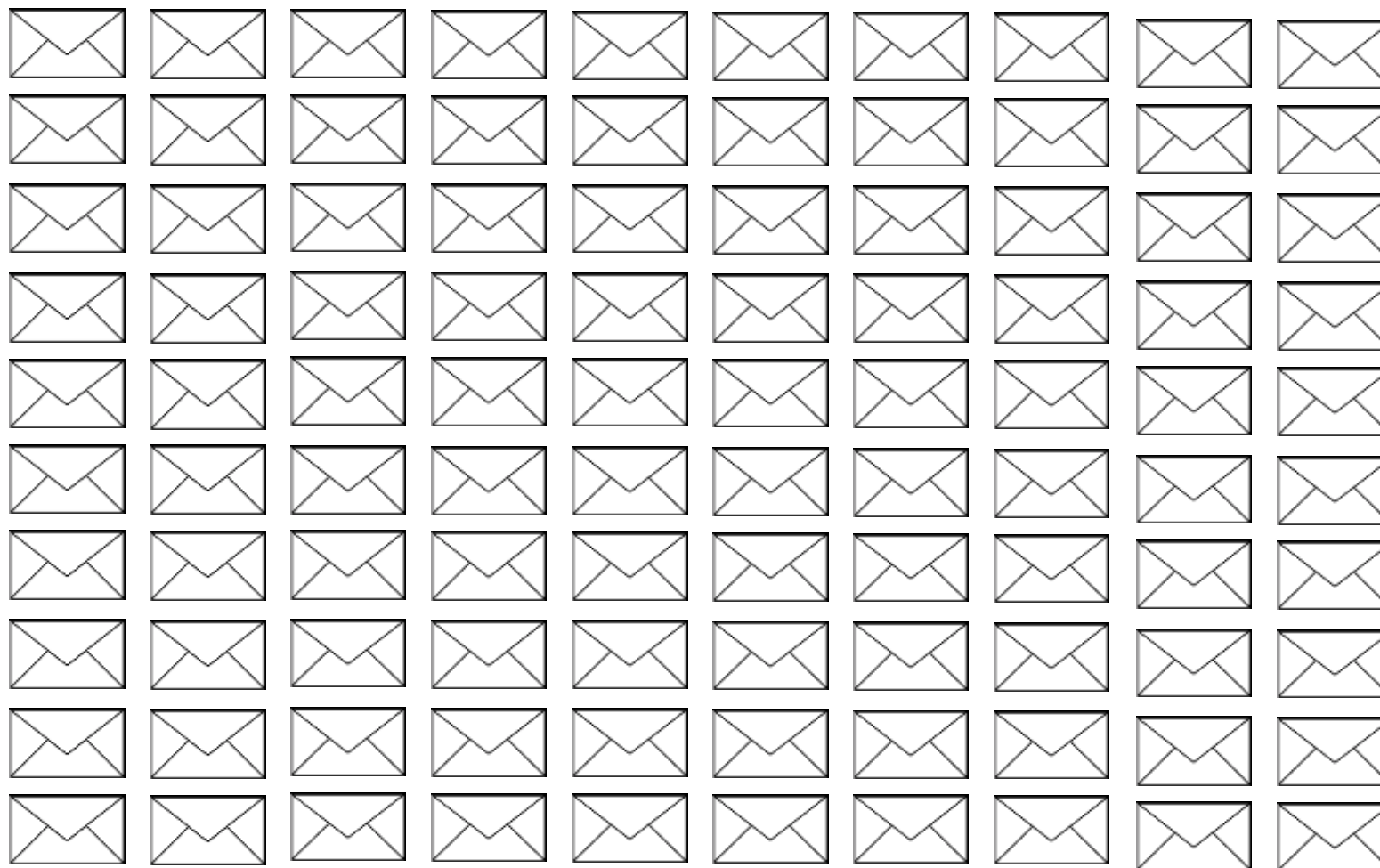
k-means法



主成分分析

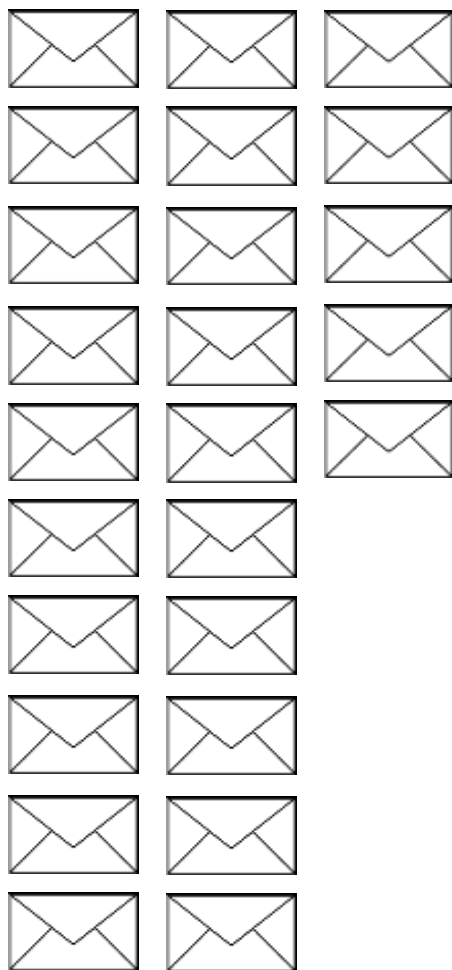


# スパムメールの判別

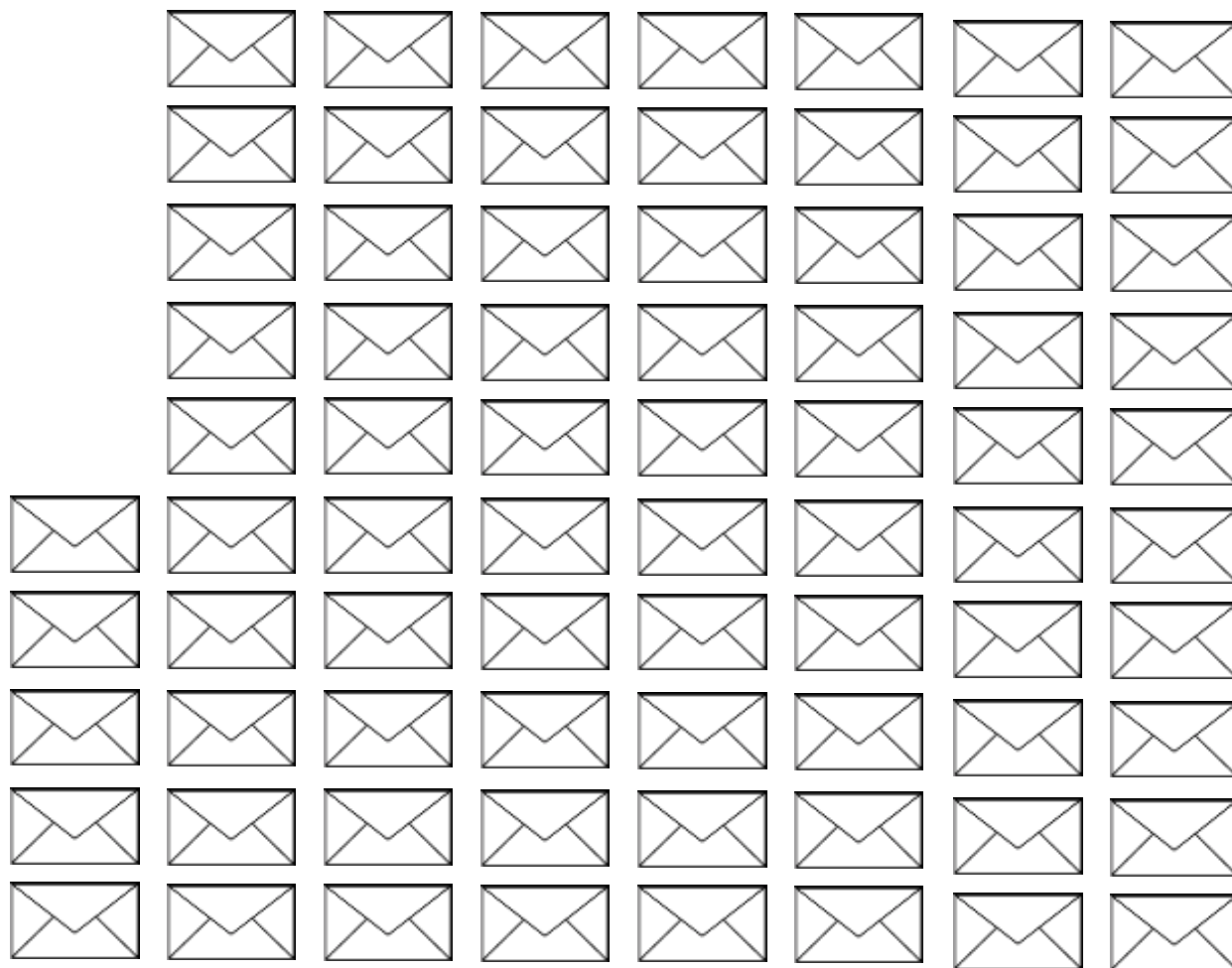




# スパムメールの判別



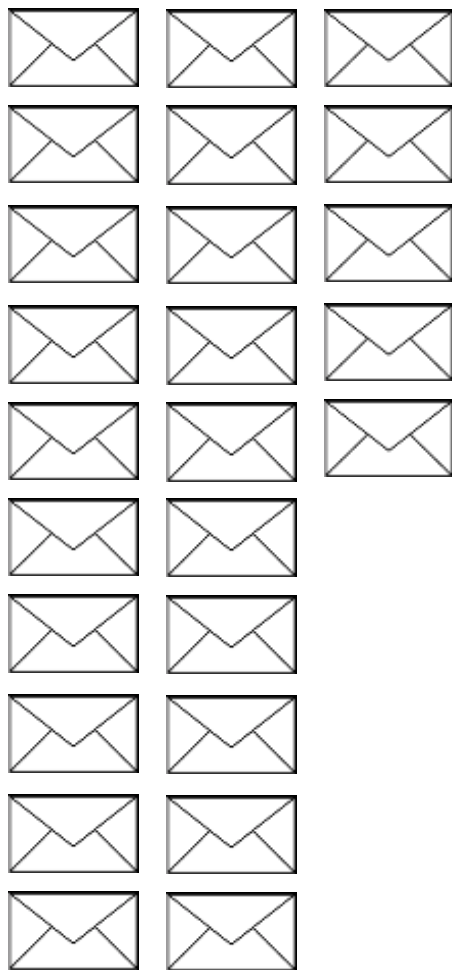
**スパム(25)**



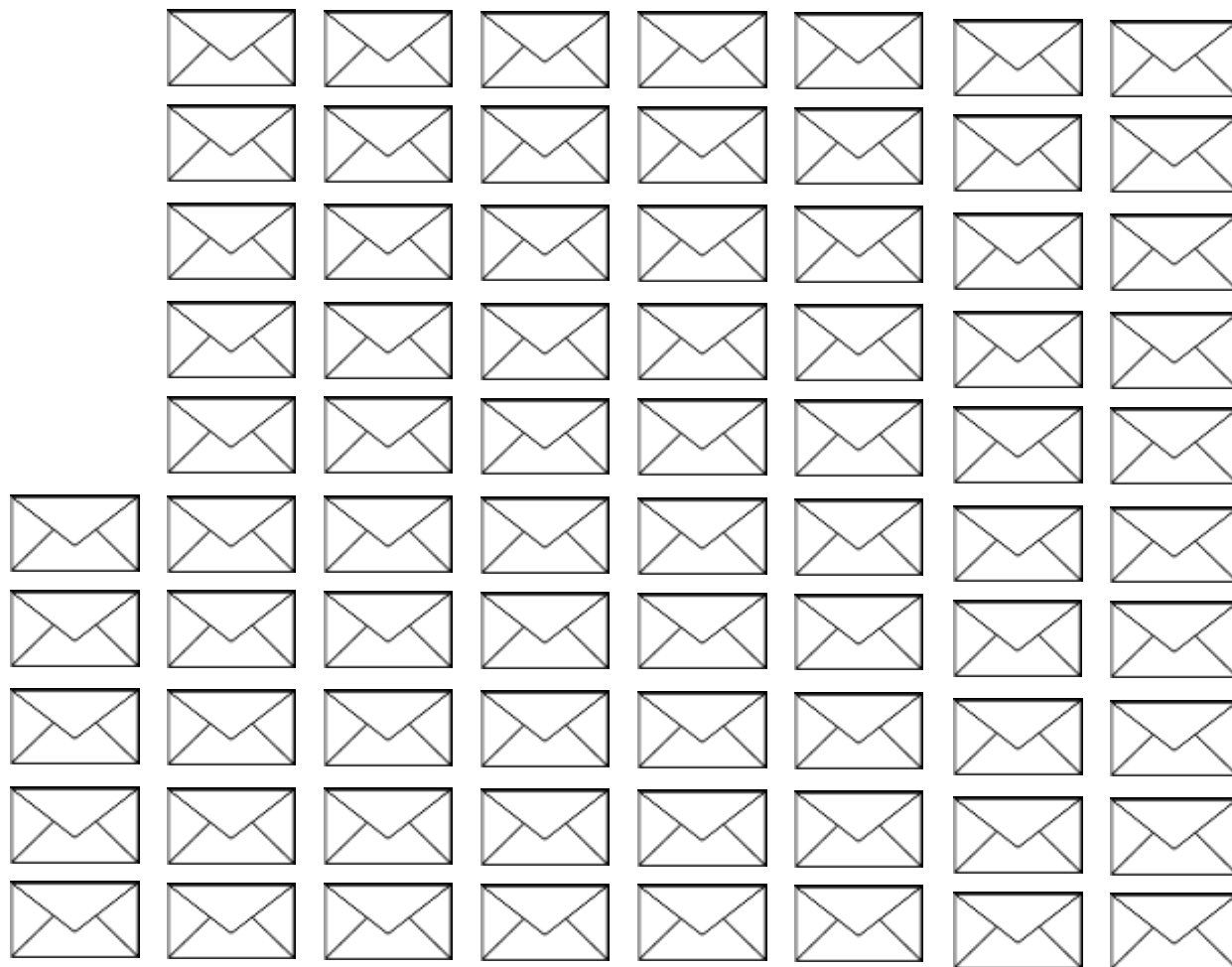
**通常メール(75)**

# スパムメールの判別

「出会い」



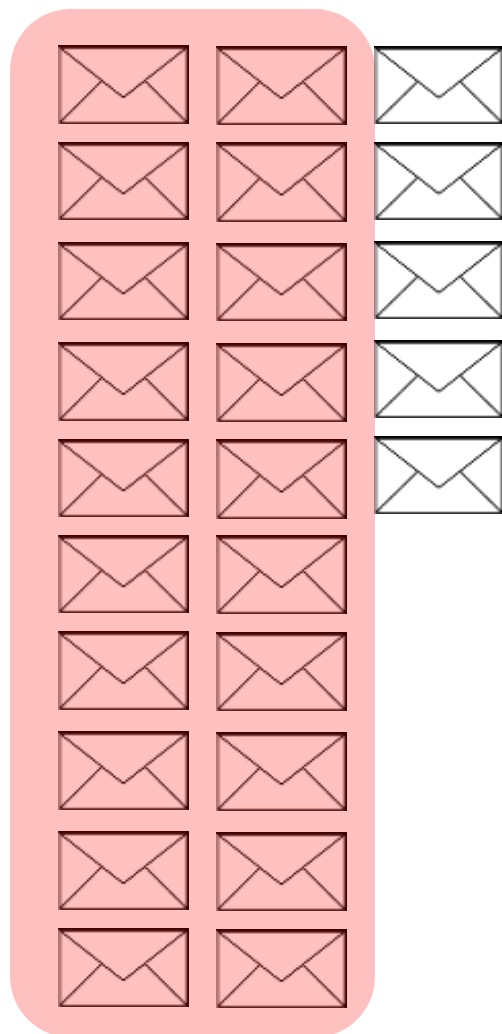
スパム(25)



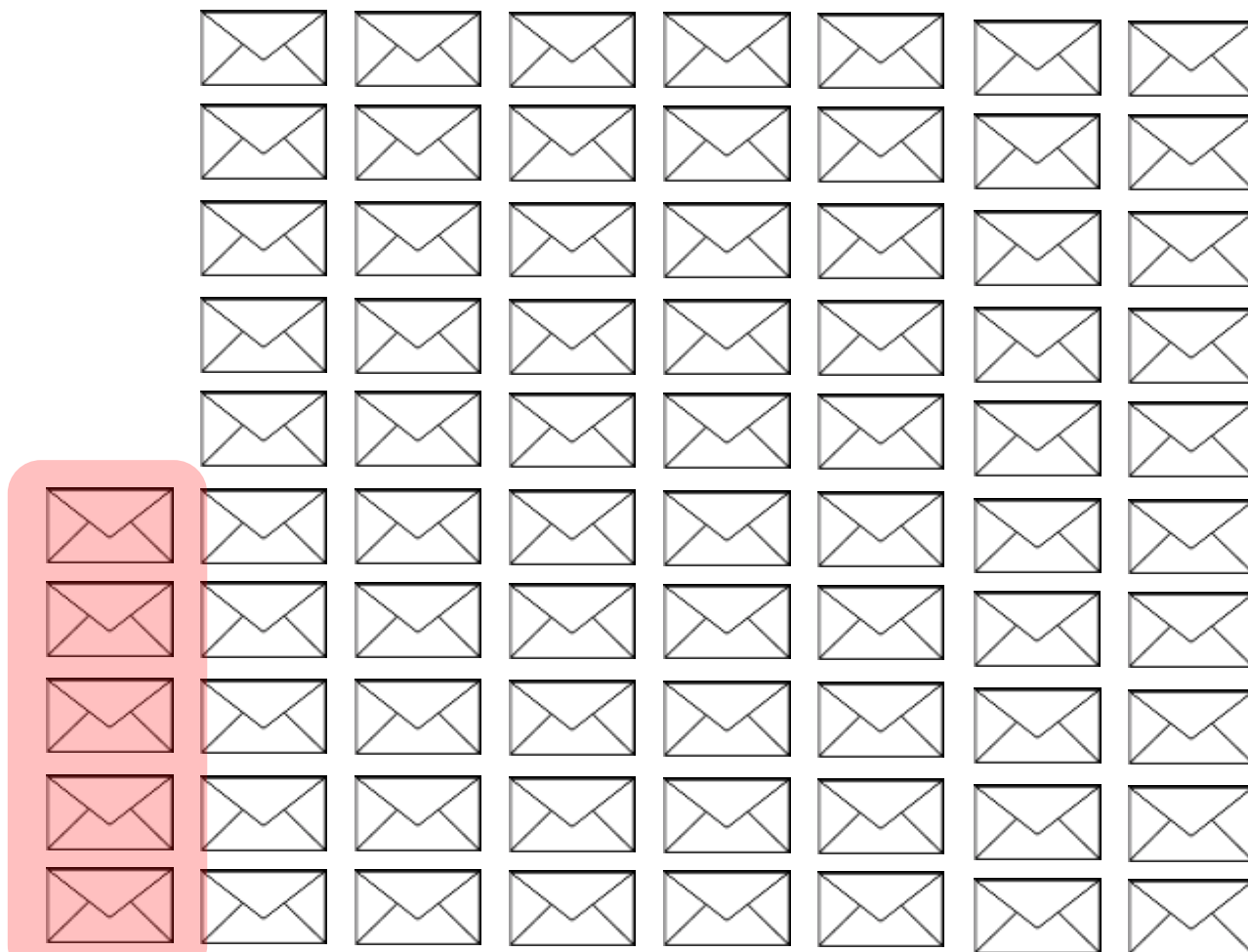
通常メール(75)

# スパムメールの判別

「出会い」

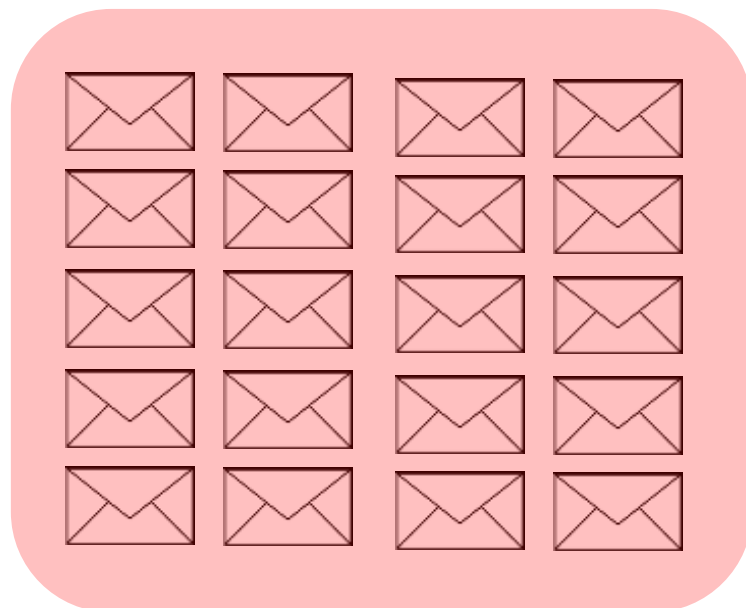


スパム(20/25)

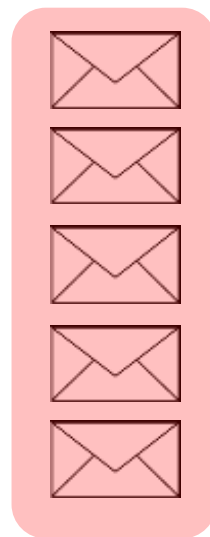


通常メール(5/75)

# スパムメールの判別



スパム(20)

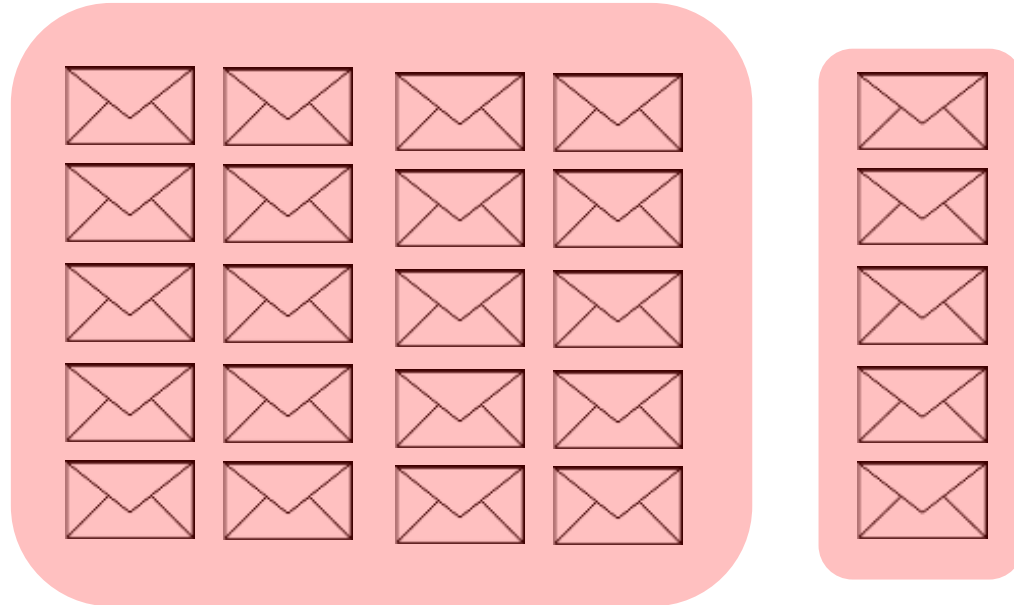


通常メール(5)

「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%

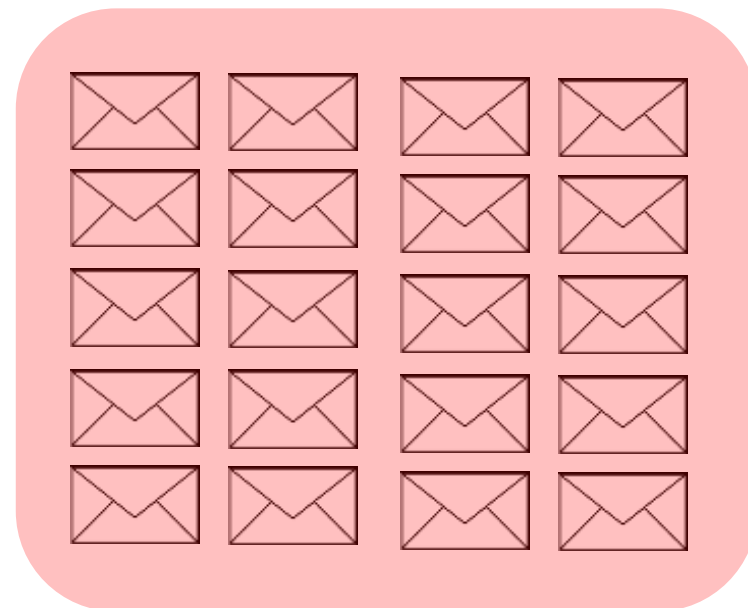
# スパムメールの判別

「出会い」 → 80%

「安い」 → 95%

「 題目がない」 → 70%

ナイーブベイズ



スパム

# 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



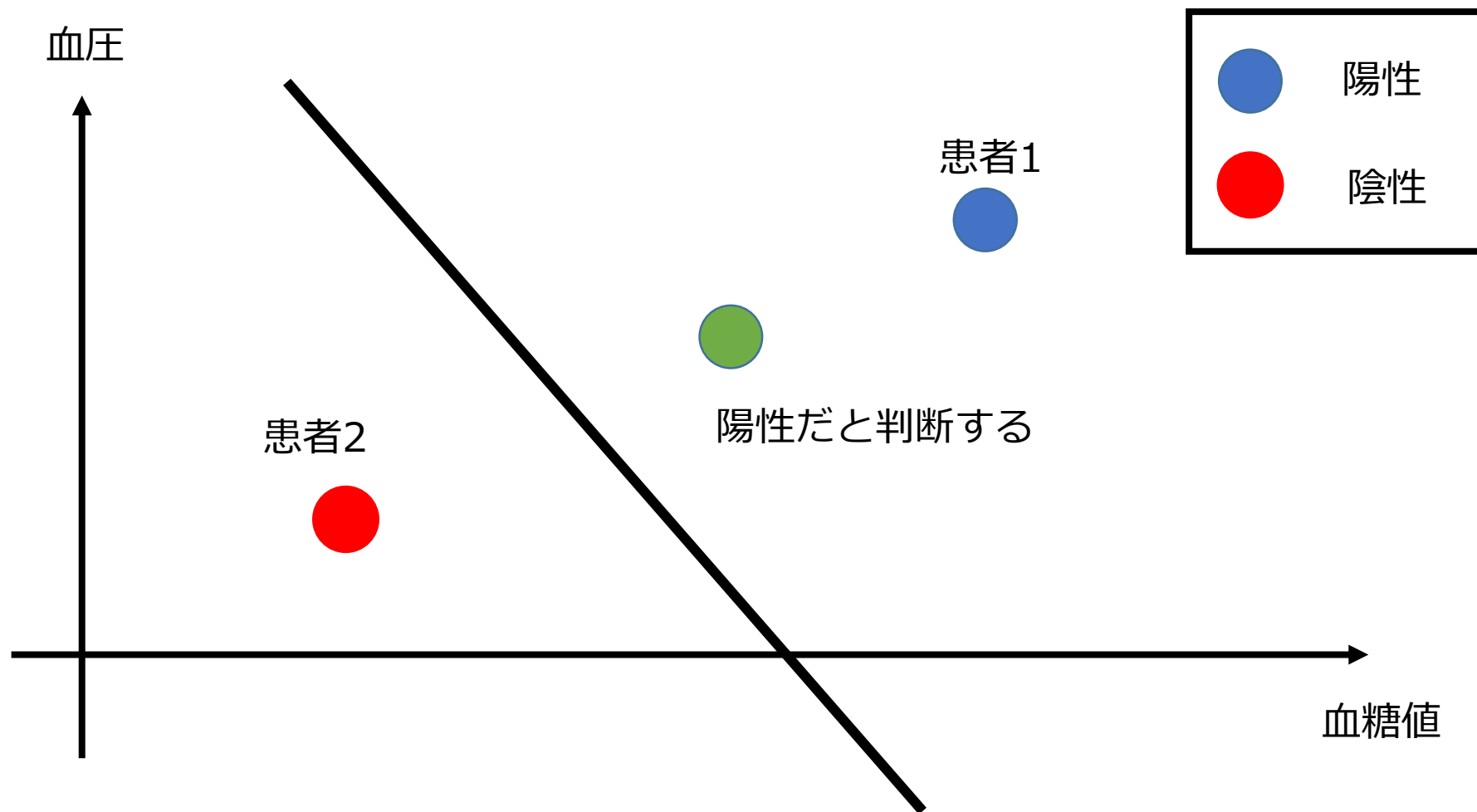
# 陽性・陰性？



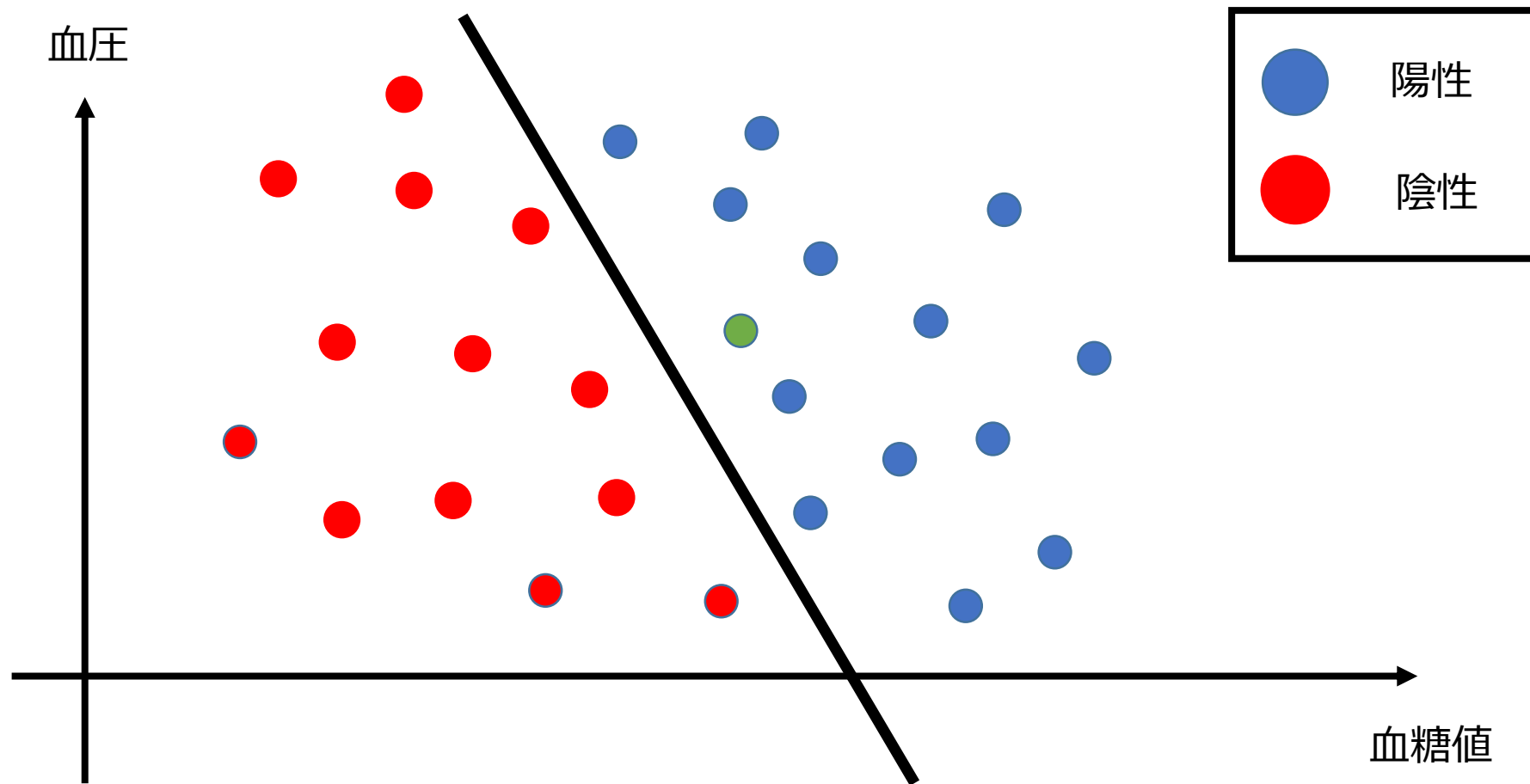
患者ID	血圧	血糖値	
1	120	200	陽性
2	80	130	陰性
3	110	190	?



# 陽性・陰性を識別する

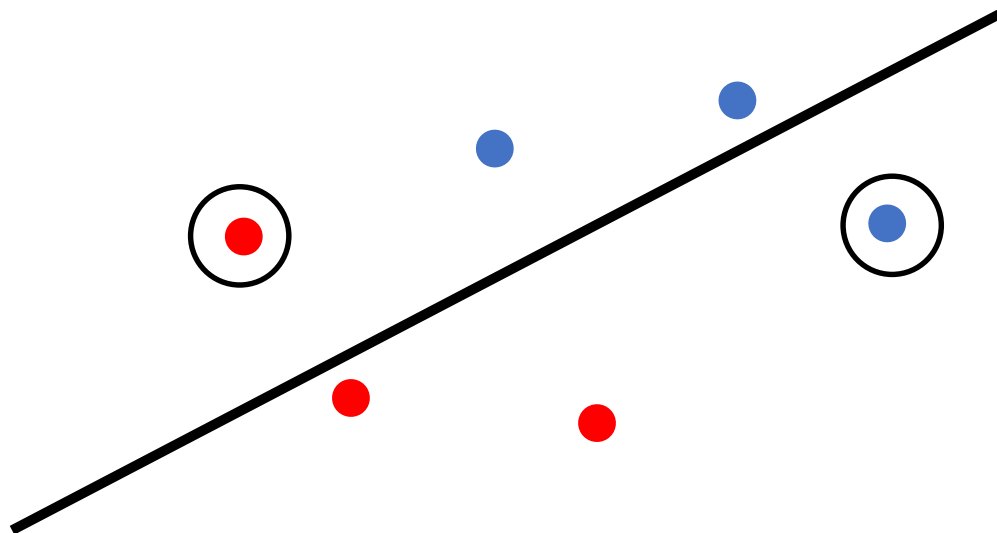


# 陽性・陰性を識別する



ロジスティック回帰

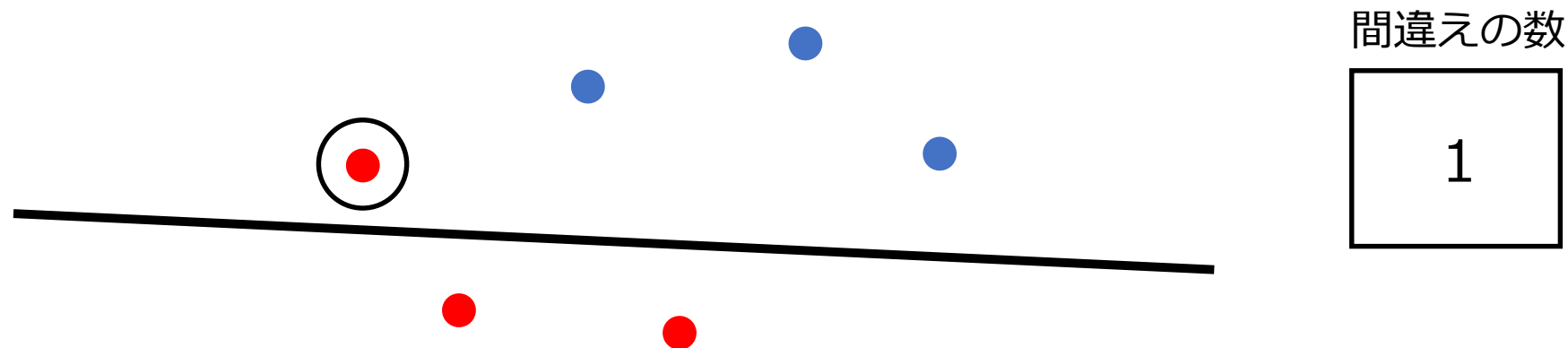
# ロジスティック曲線の求め方



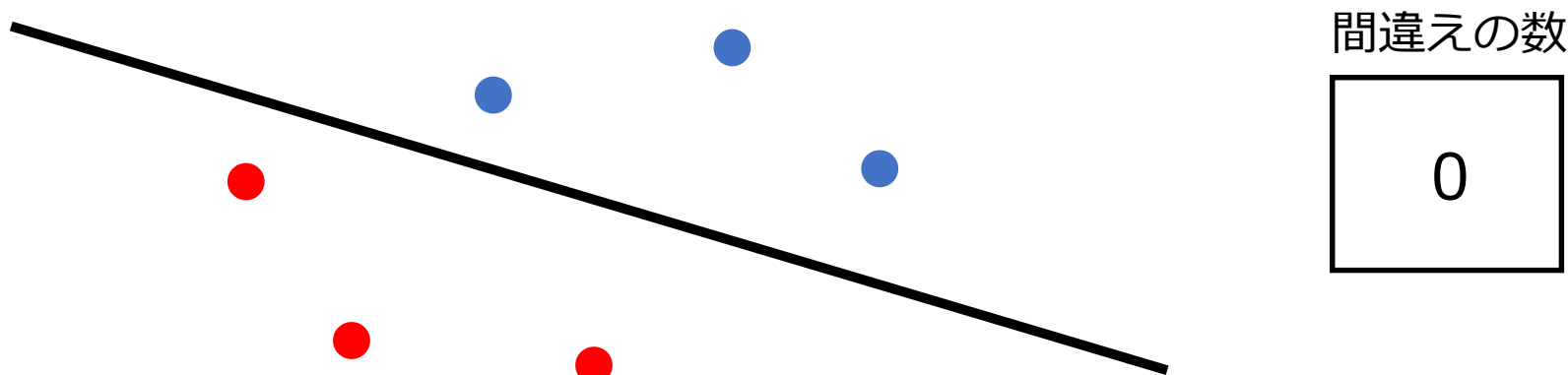
間違えの数

2

# ロジスティック曲線の求め方

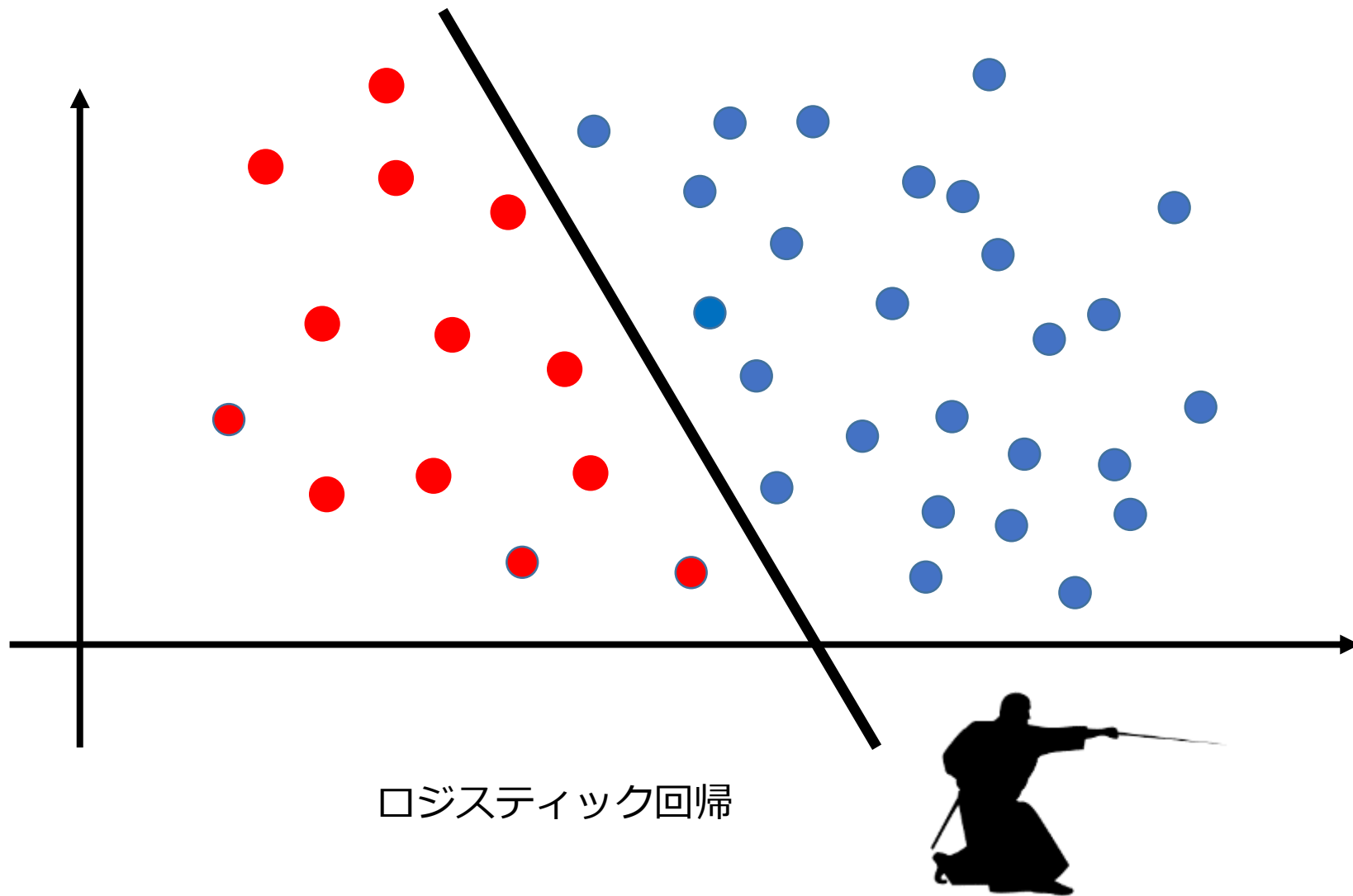


# ロジスティック曲線の求め方

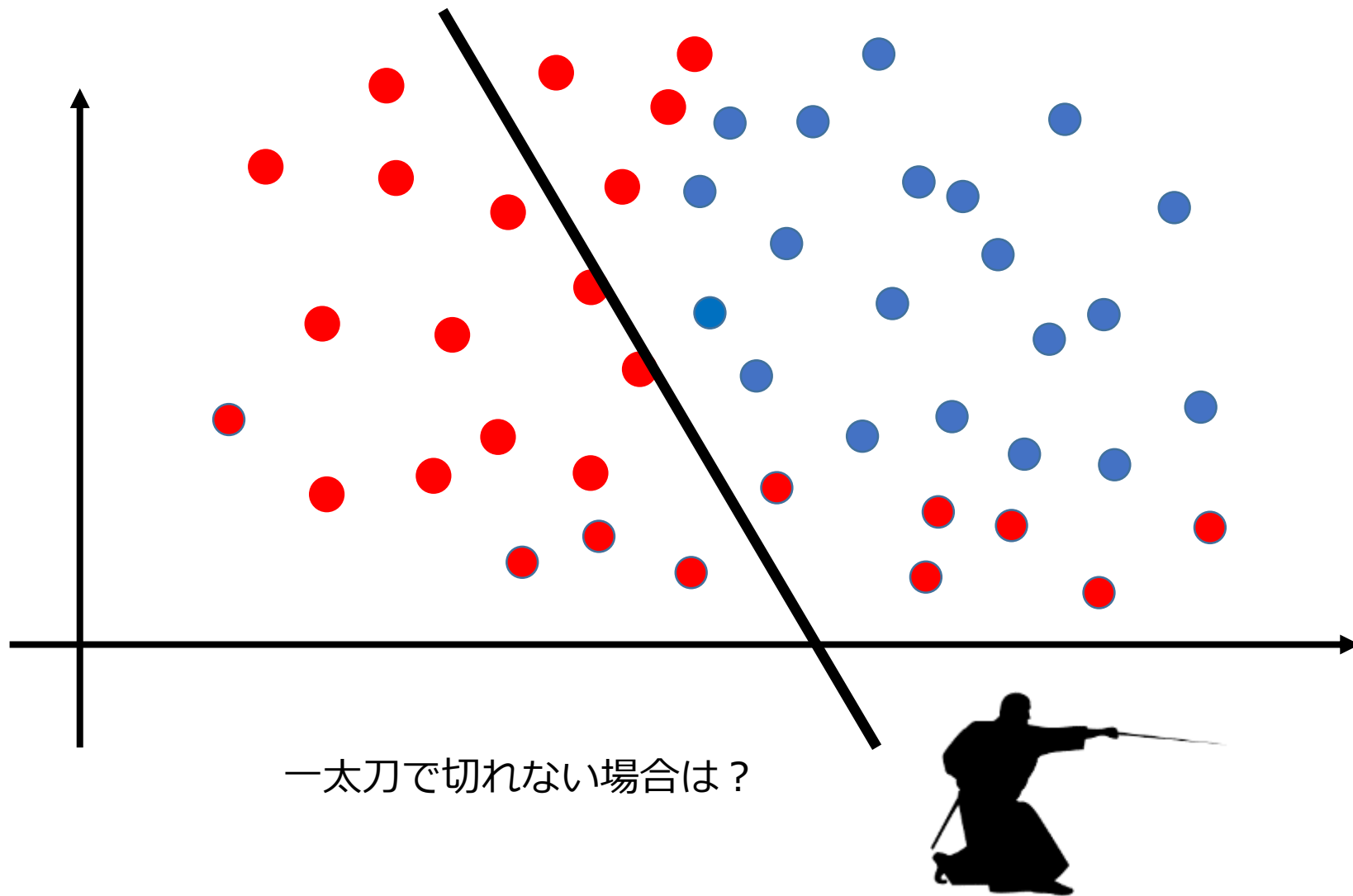


最尤法によってロジスティック曲線を求める

# いろいろな識別方法



# いろいろな識別方法



# 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

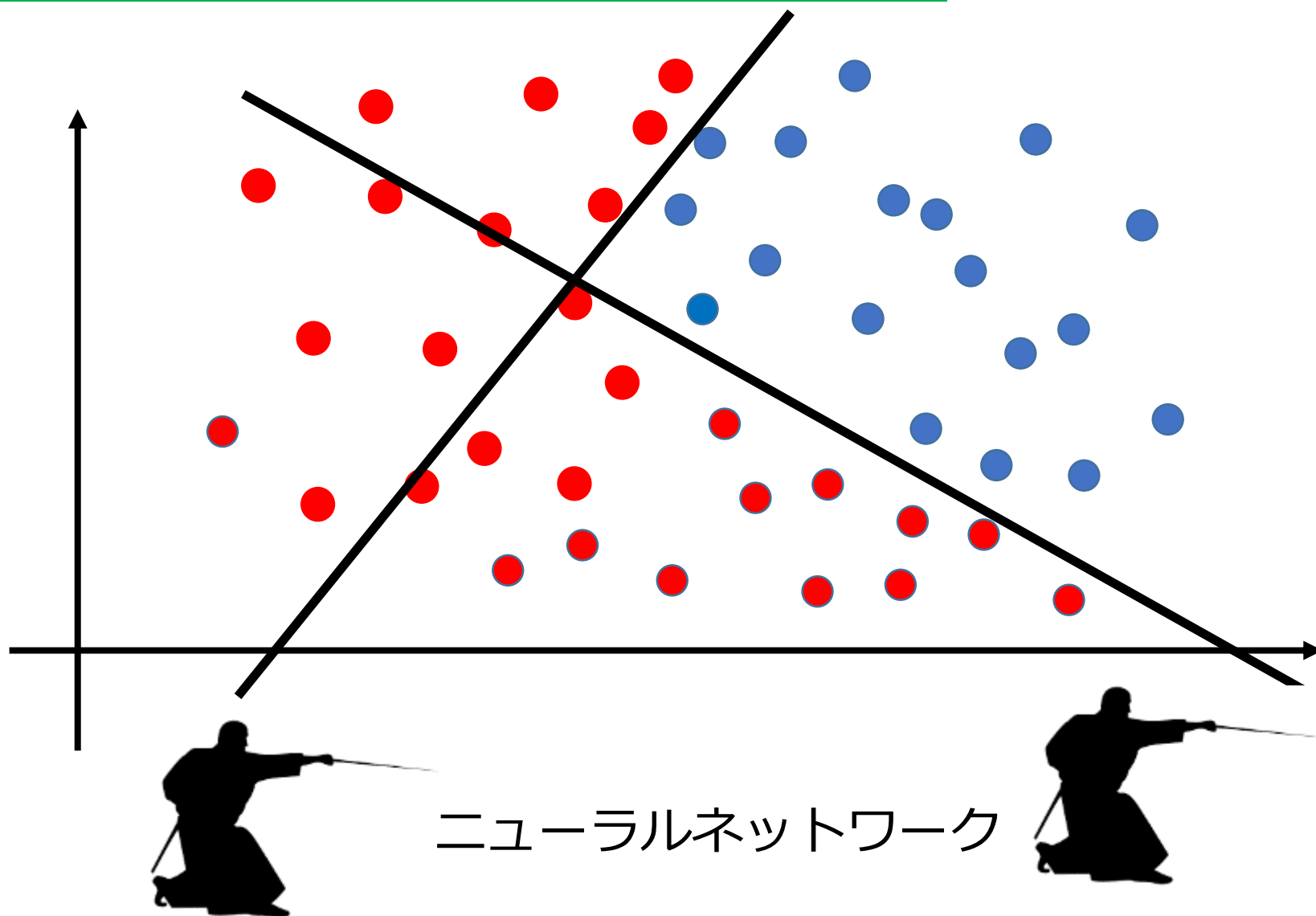


主成分分析

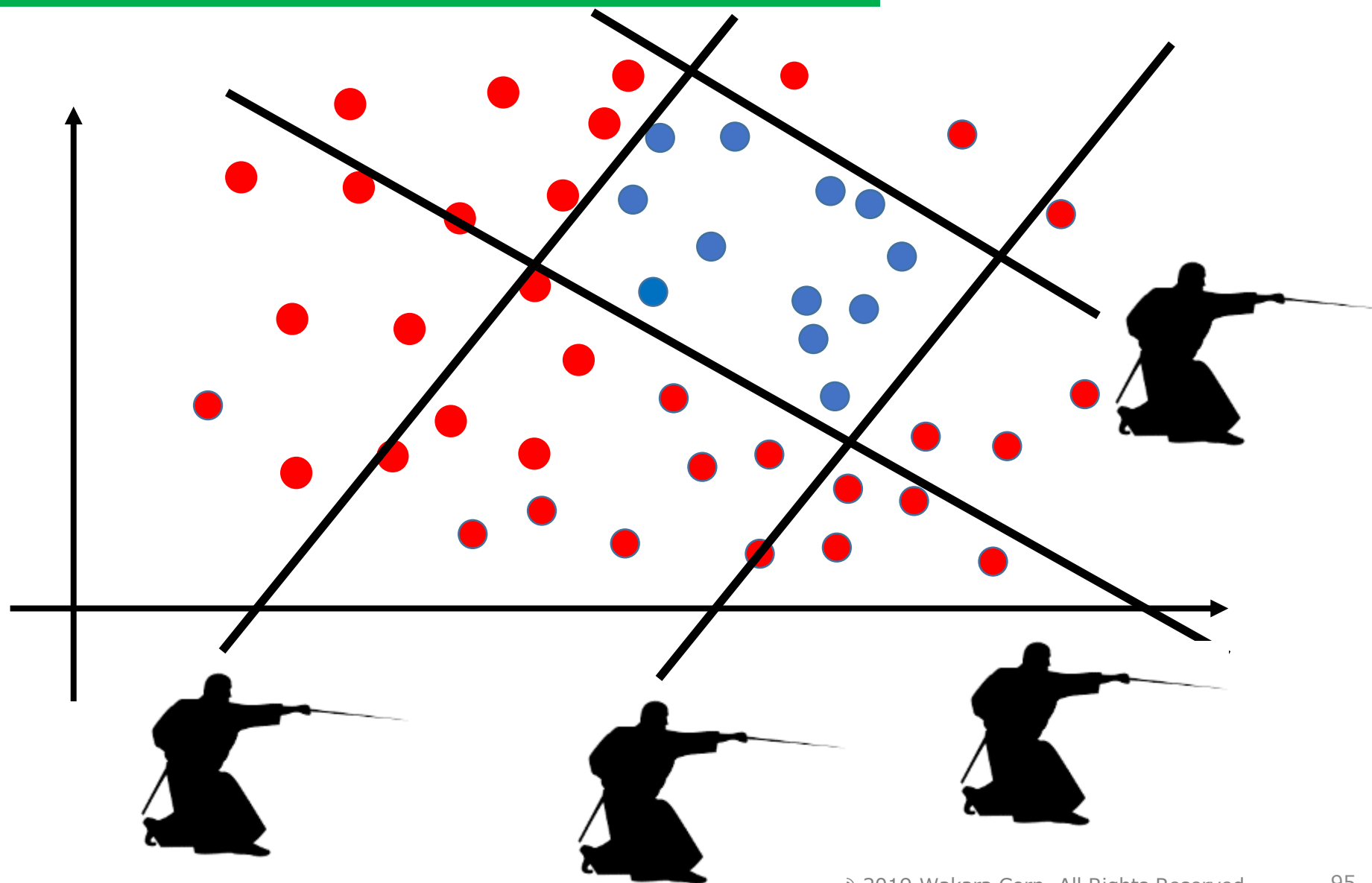




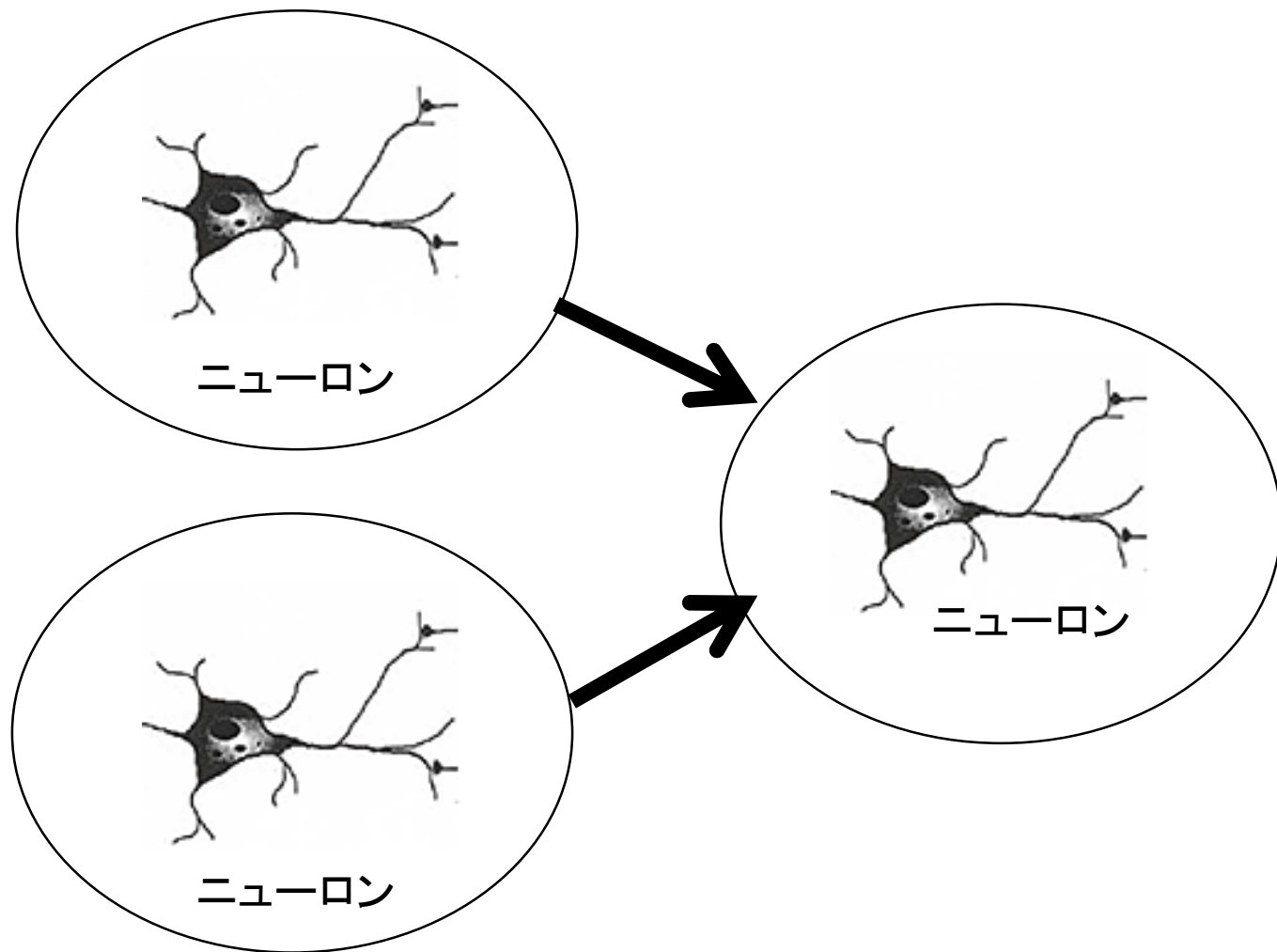
# いろいろな識別方法



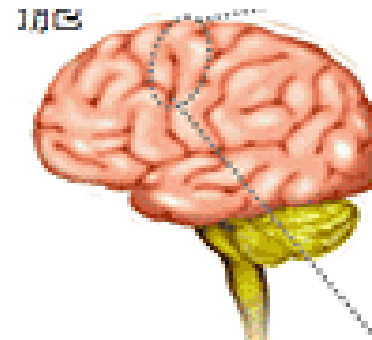
# いろいろな識別方法

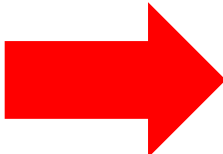


# ニューラルネットワーク

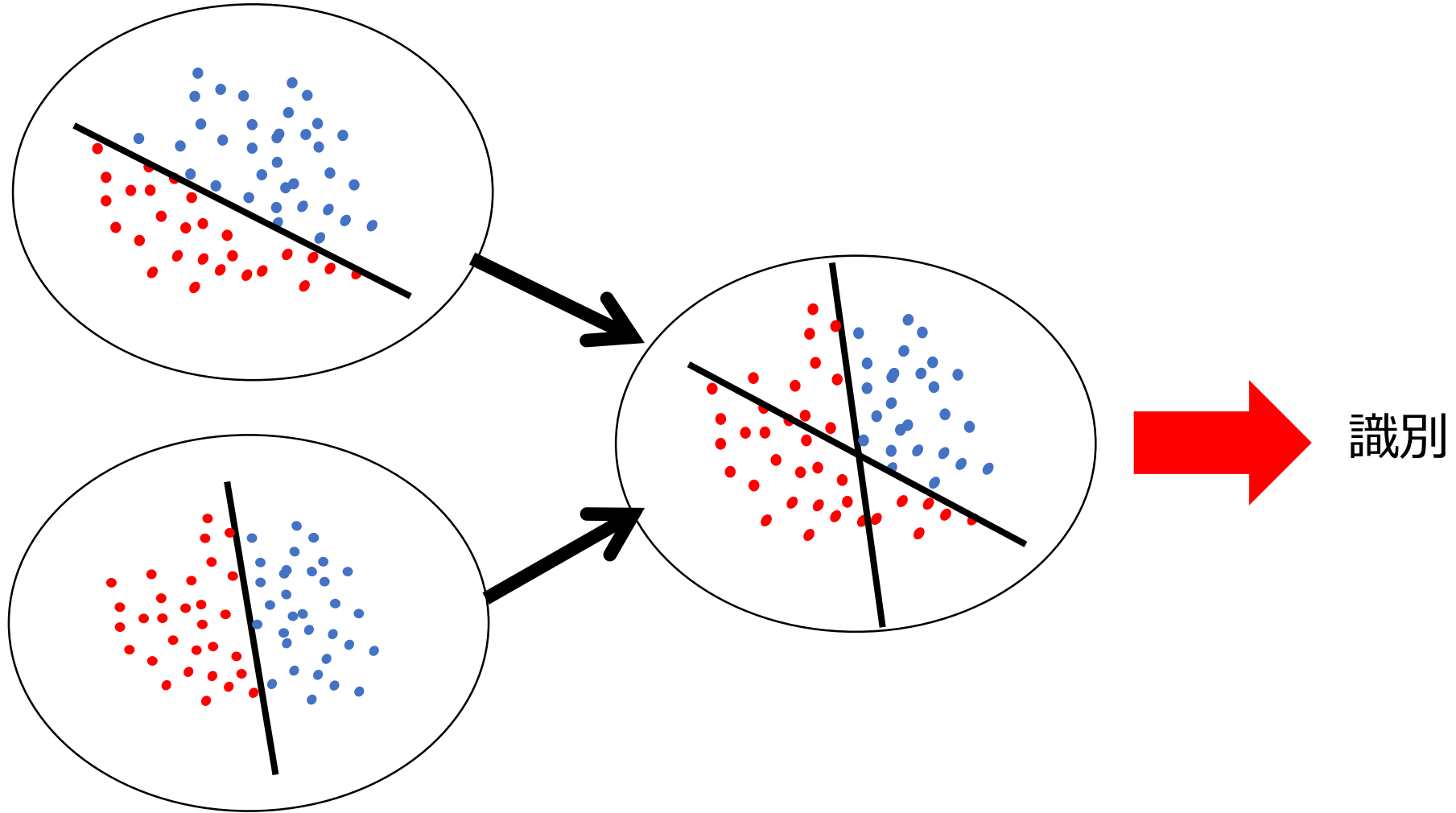


人間の脳はどやって識別  
を行なっているのか？

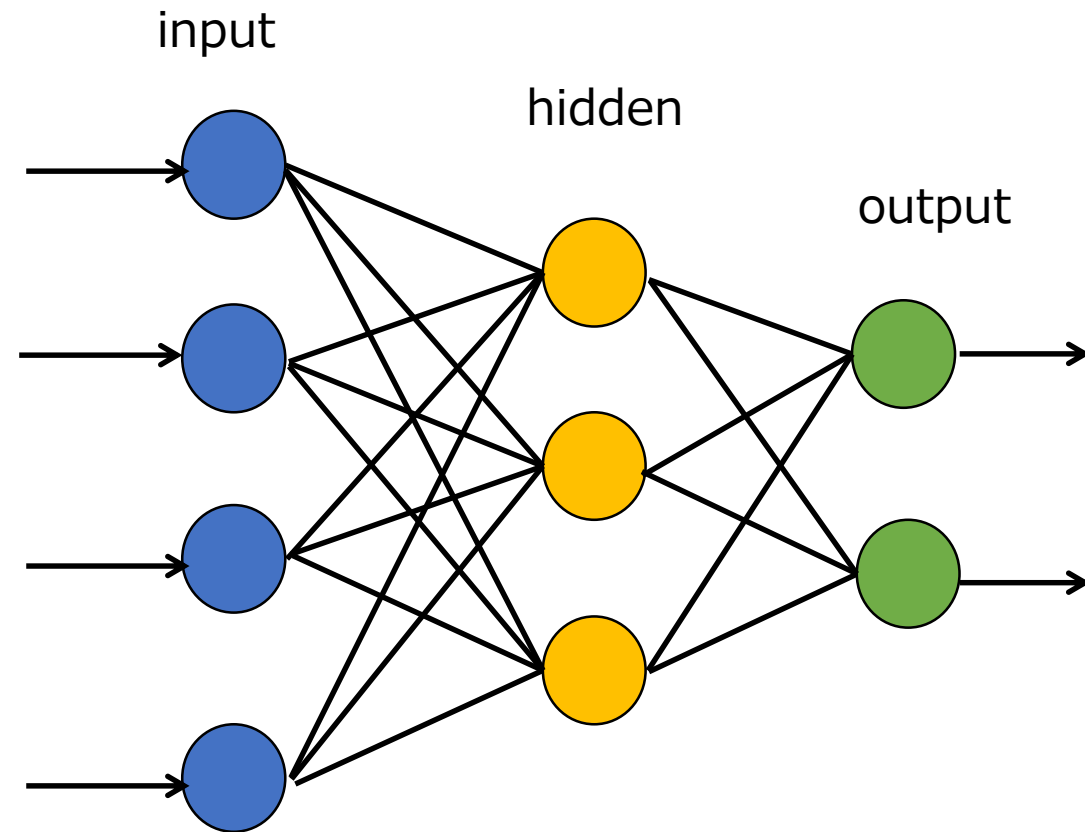


 識別

# ニューラルネットワーク

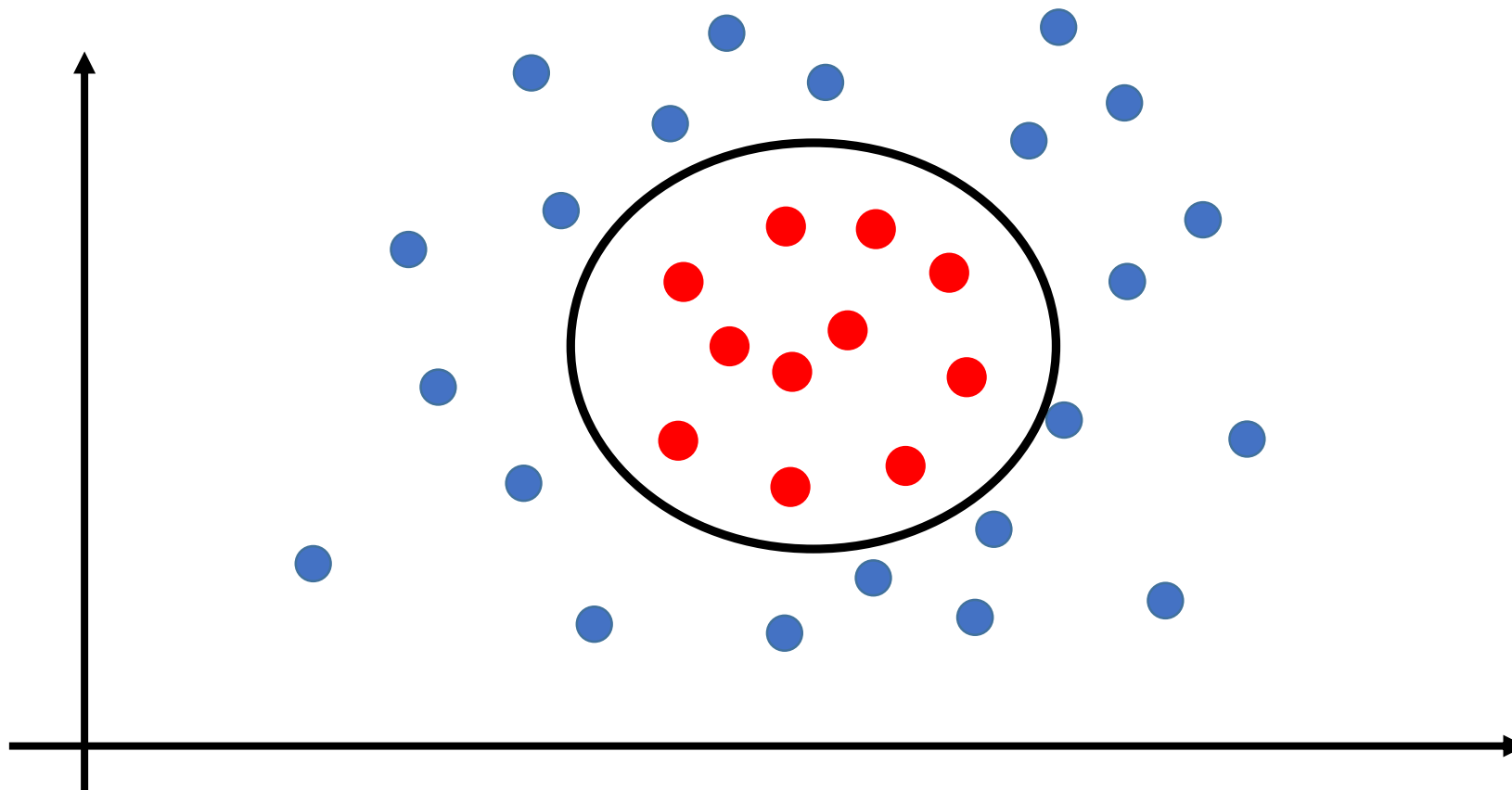


# 神経細胞のモデル



# いろいろな識別方法

直線で分けられない場合は？



# 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

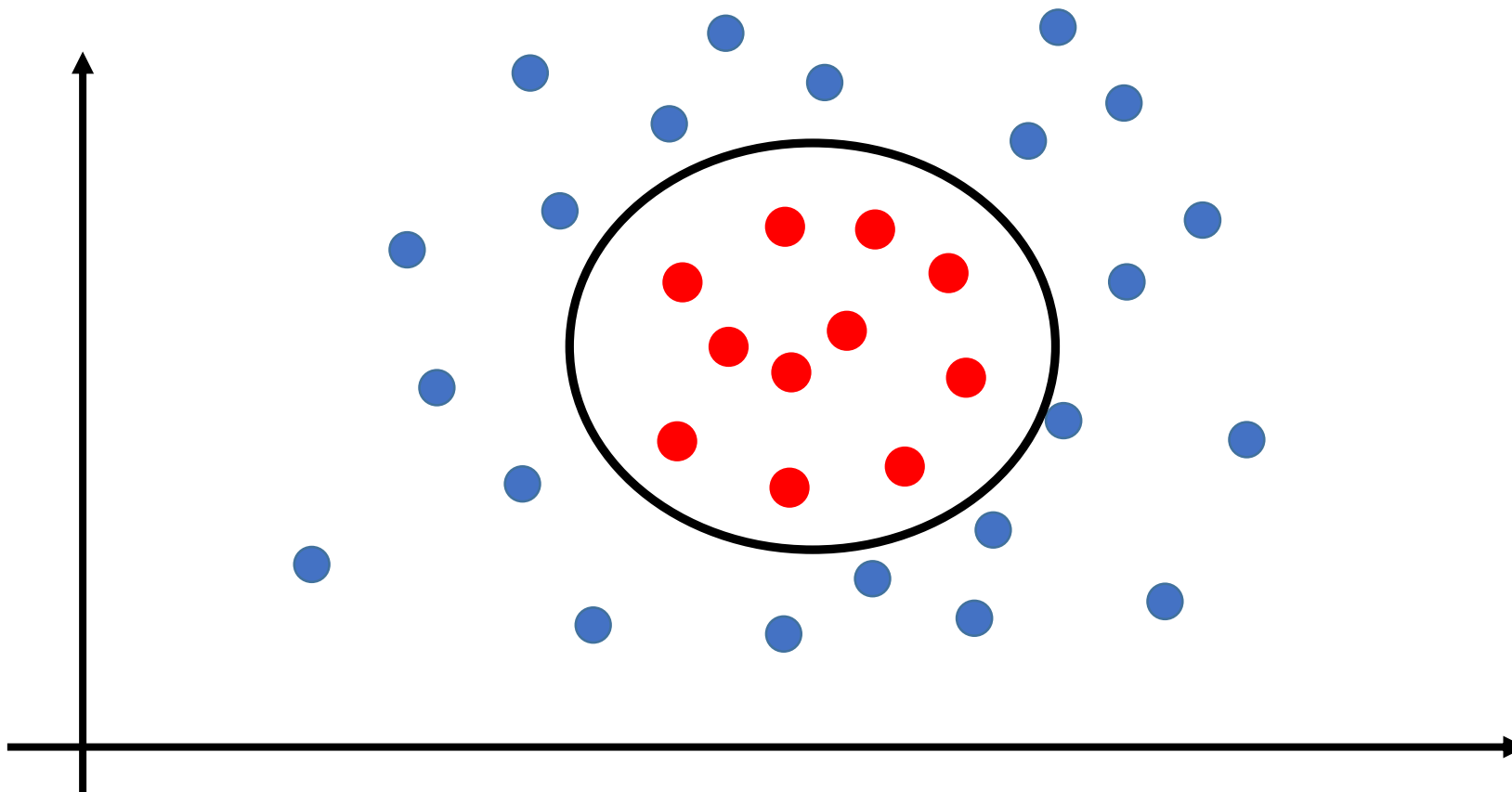


主成分分析



# いろいろな識別方法

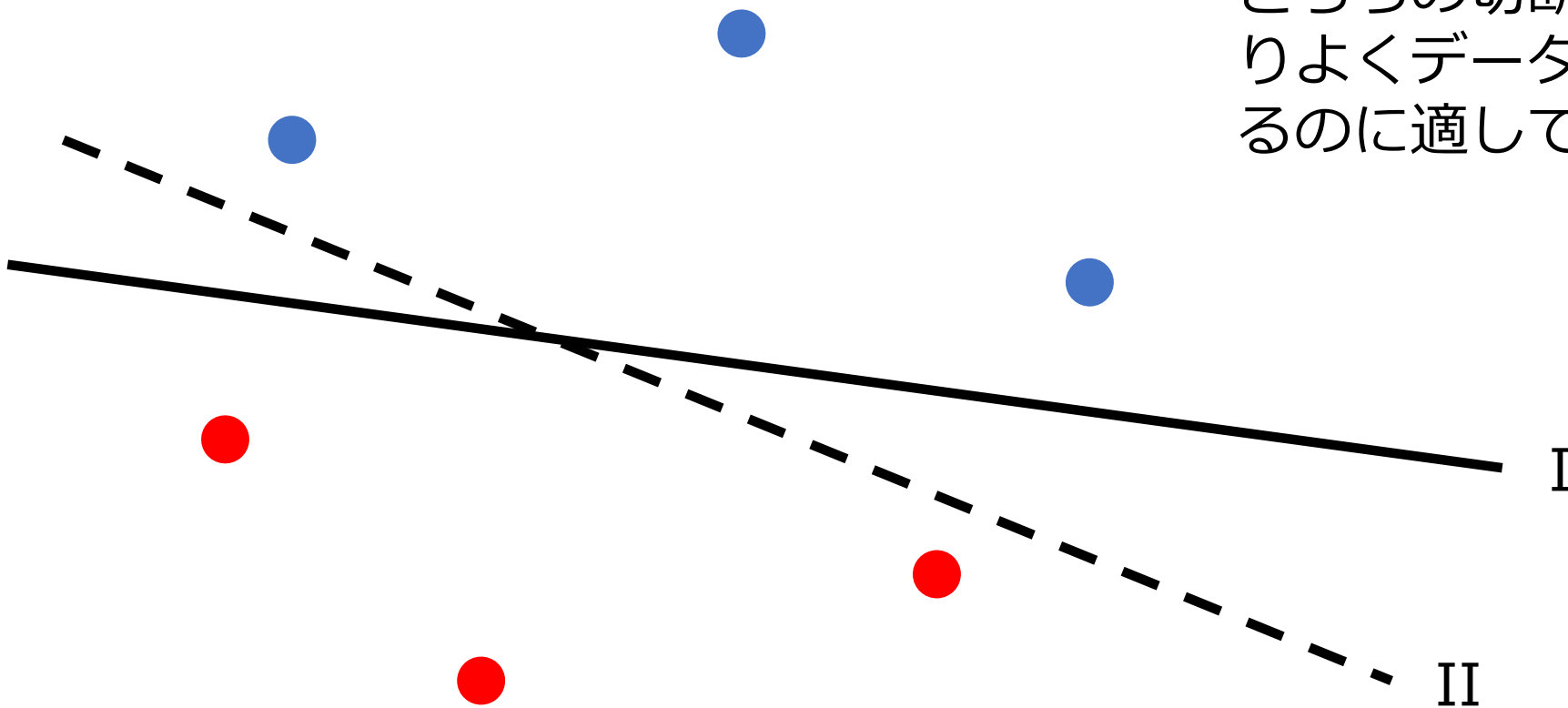
## サポートベクターマシーン





# いろいろな識別方法

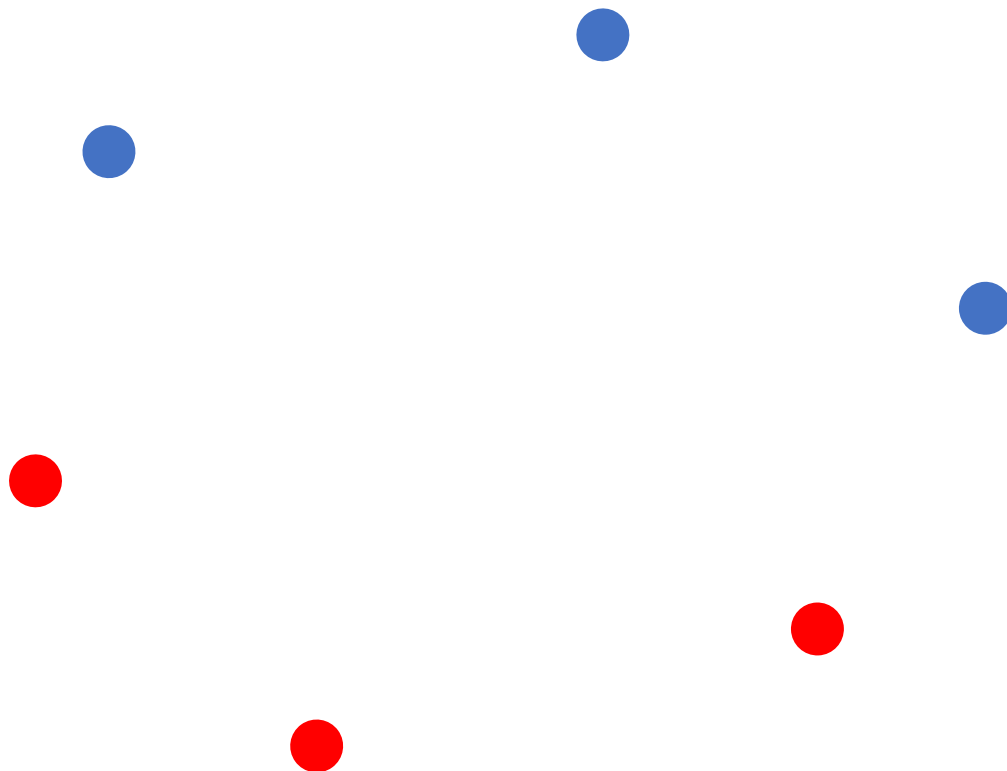
どちらの切断の方がよりよくデータを分断するのに適しているか？



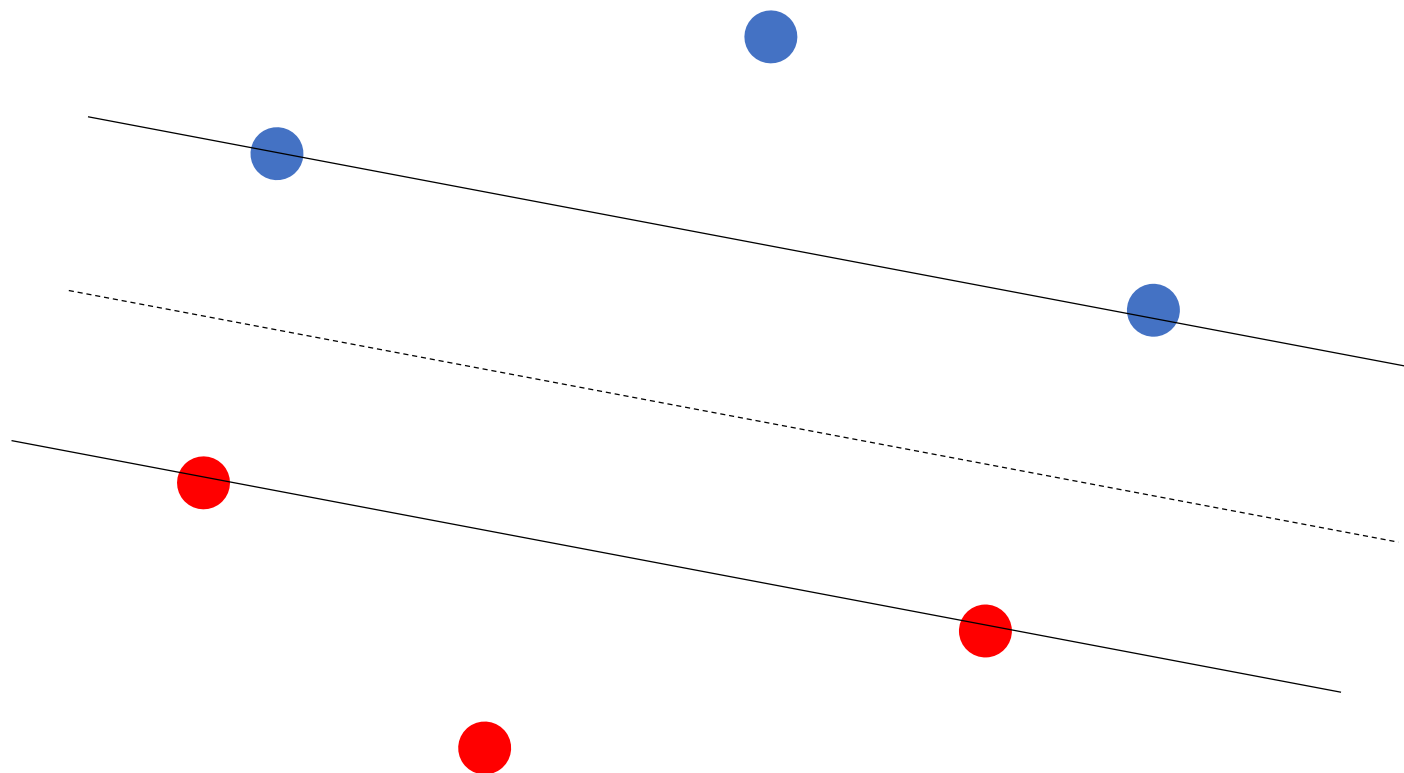
# いろいろな識別方法

サポートベクターマシーン

マージンの最大化



# いろいろな識別方法

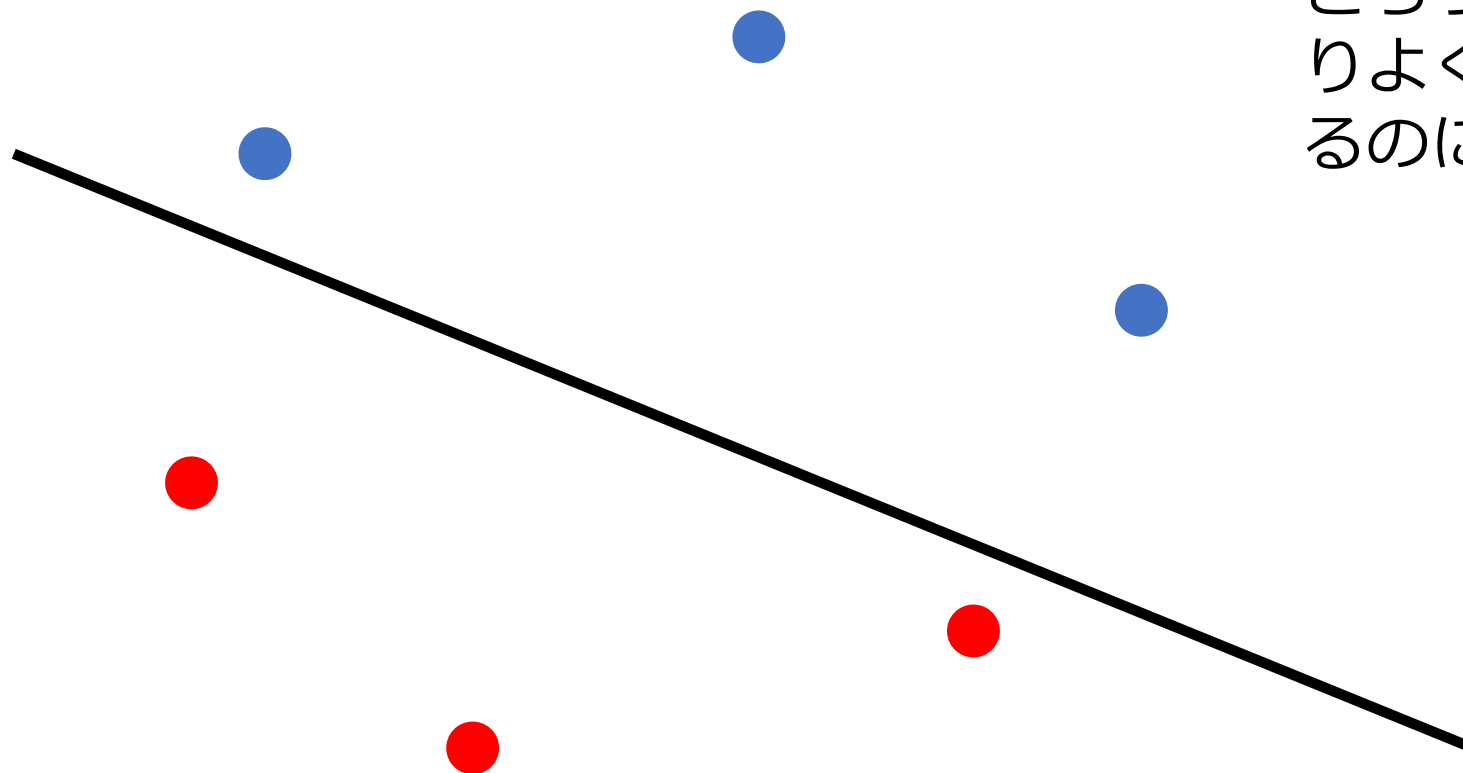


マージンの最大化

境界データからの  
距離を最大化する

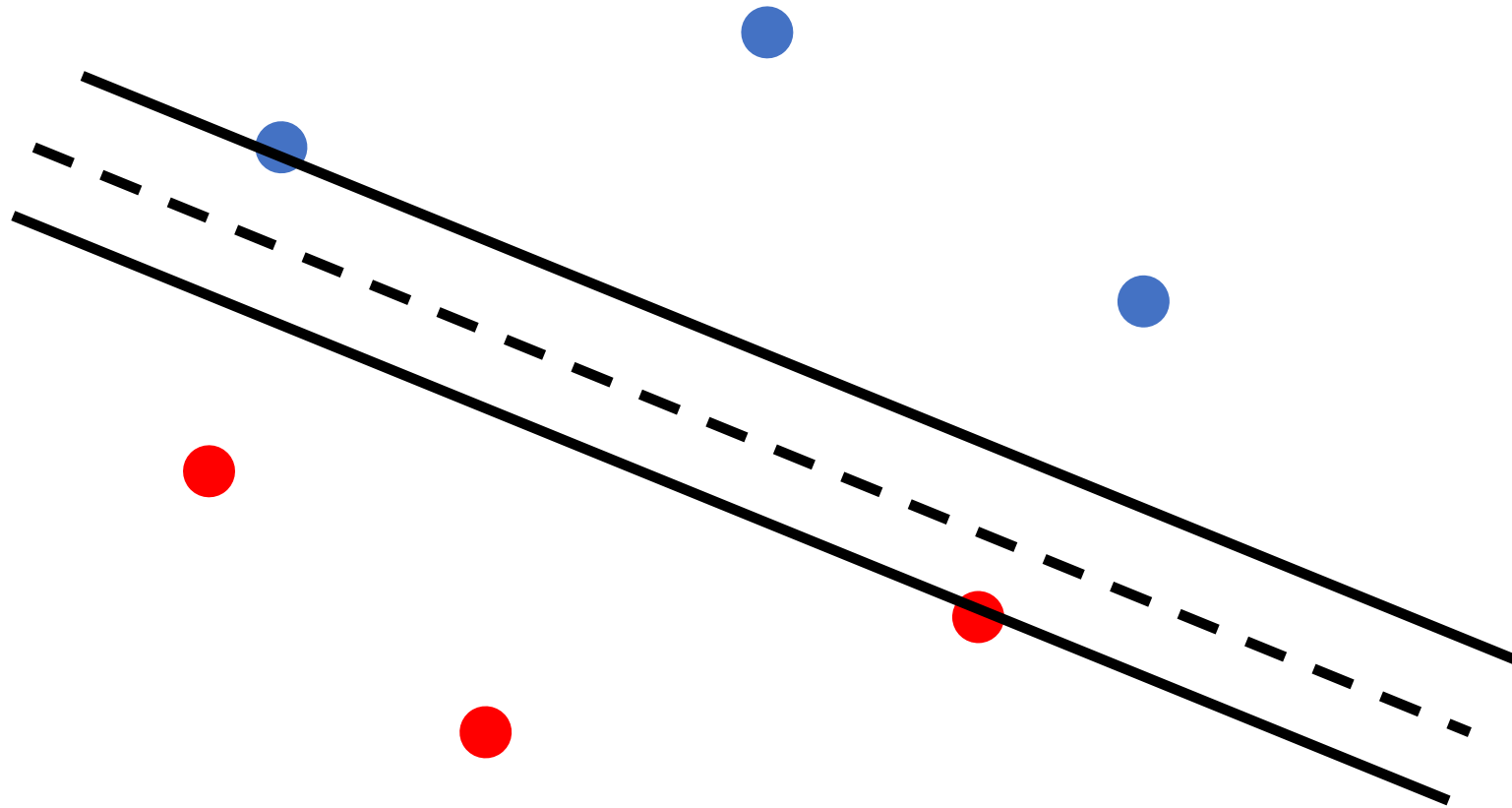
2つのグループ  
の間に一番広い  
道路を設計する  
のと同じ

# いろいろな識別方法

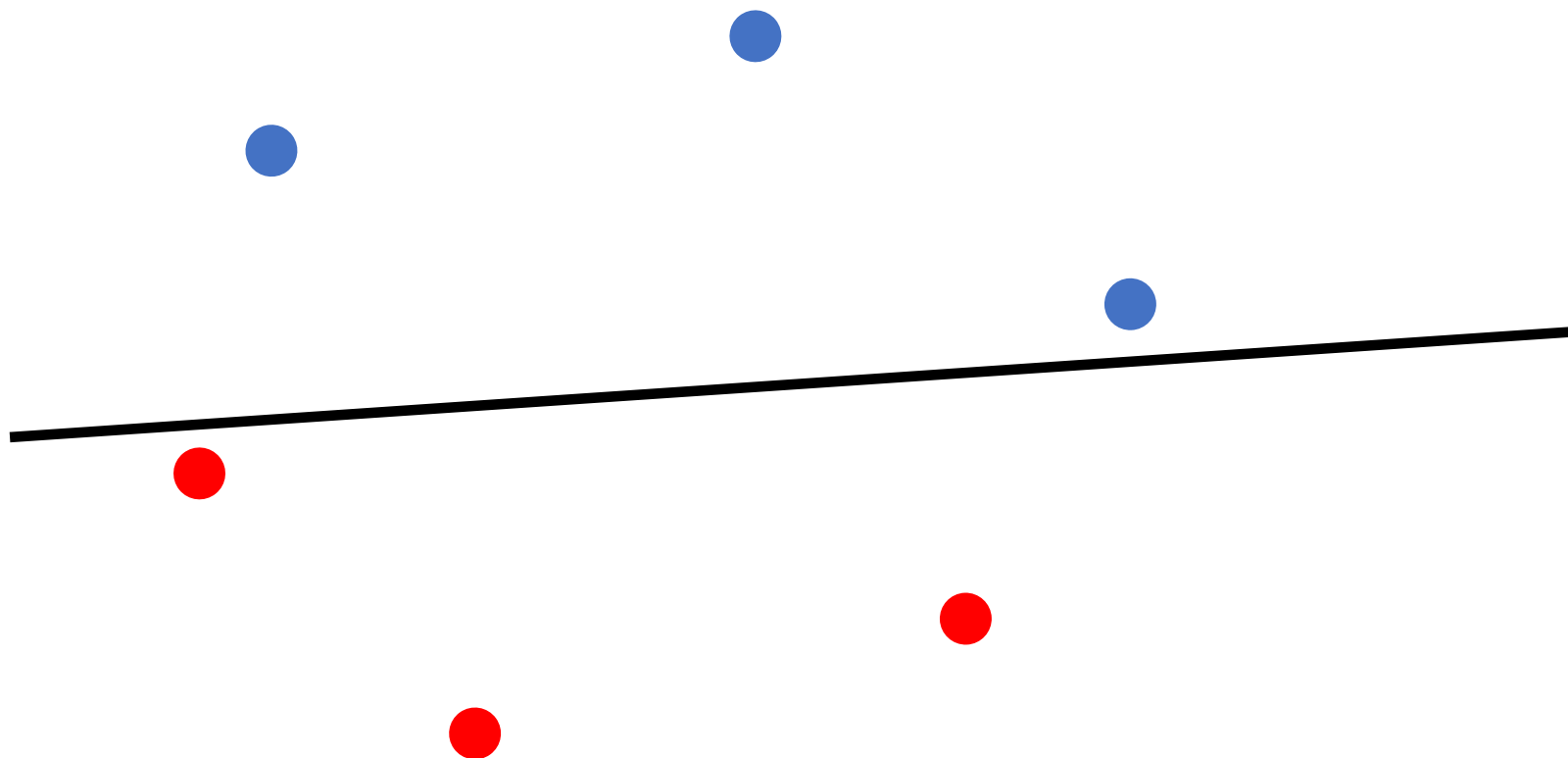


どちらの切断の方がよりよくデータを分断するのに適しているか？

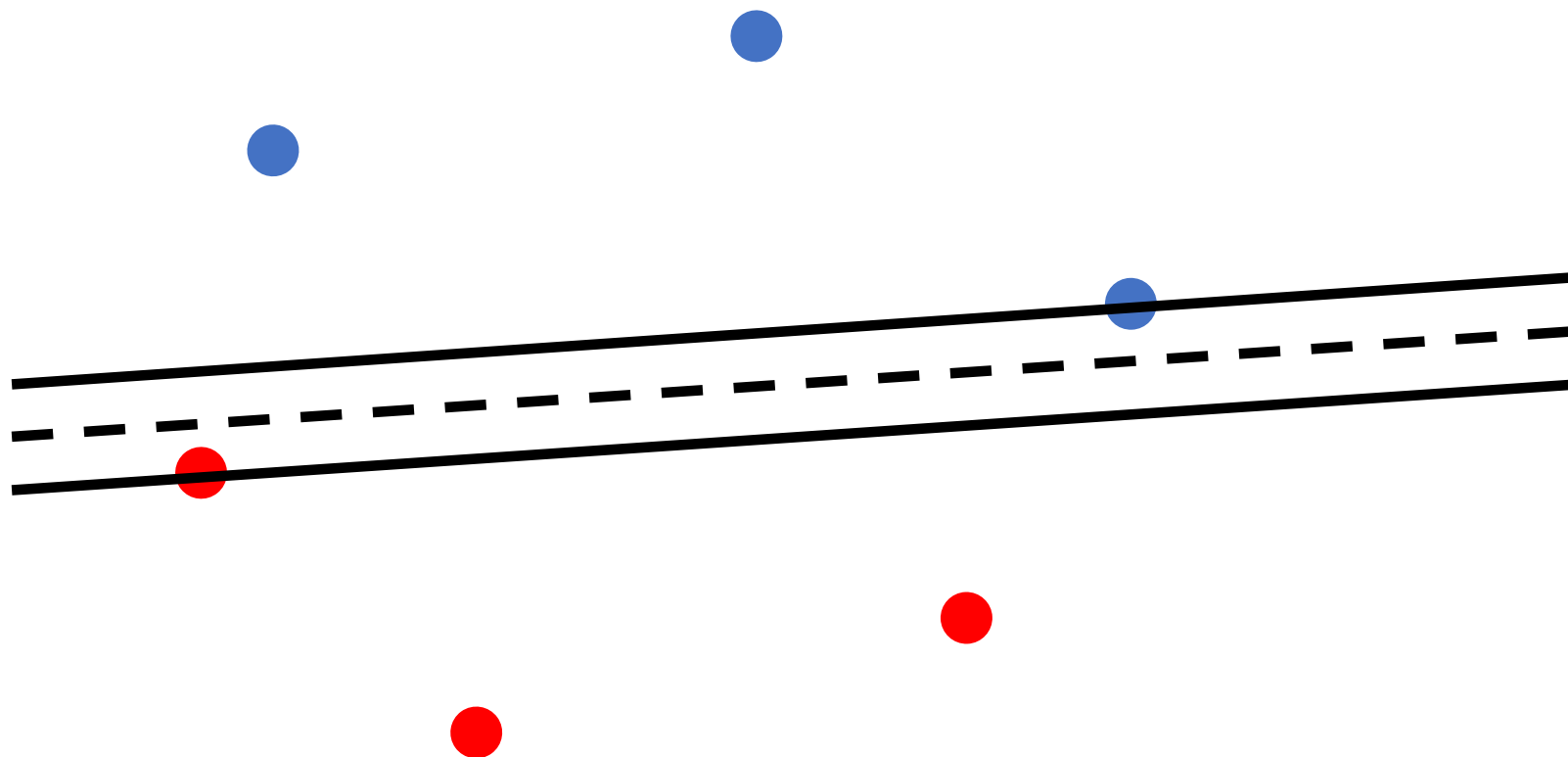
# いろいろな識別方法



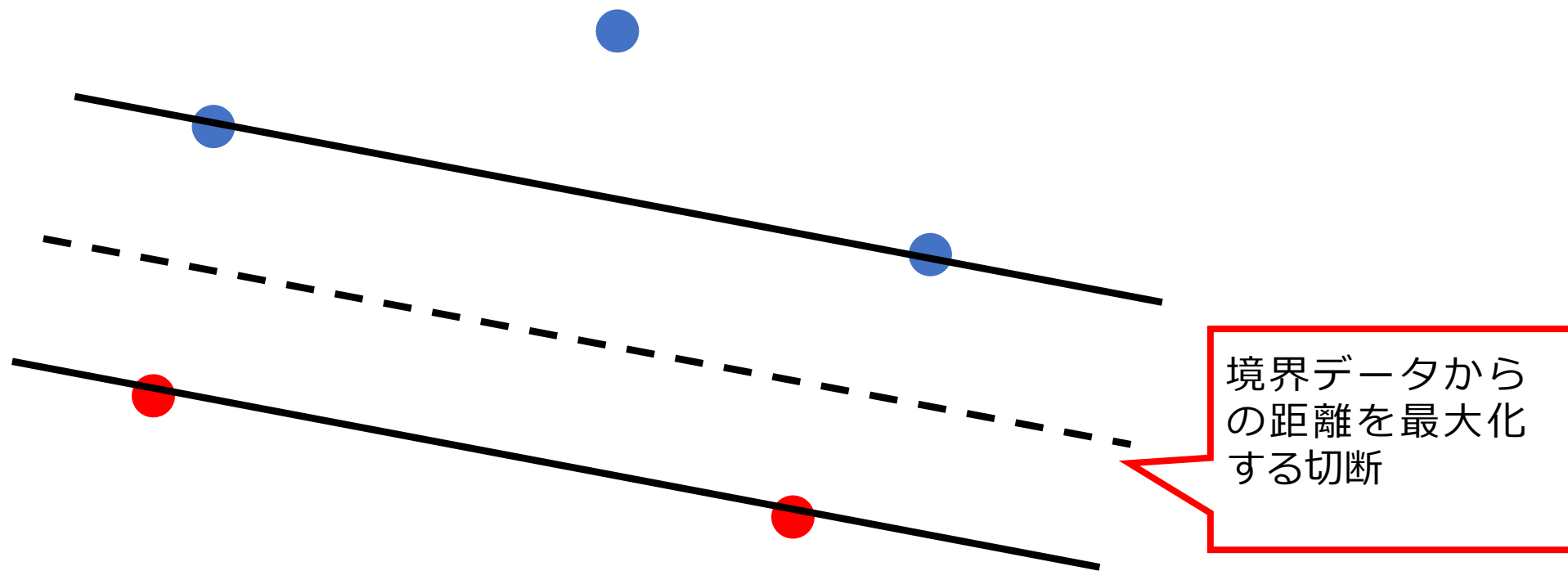
# いろいろな識別方法



# いろいろな識別方法



# いろいろな識別方法

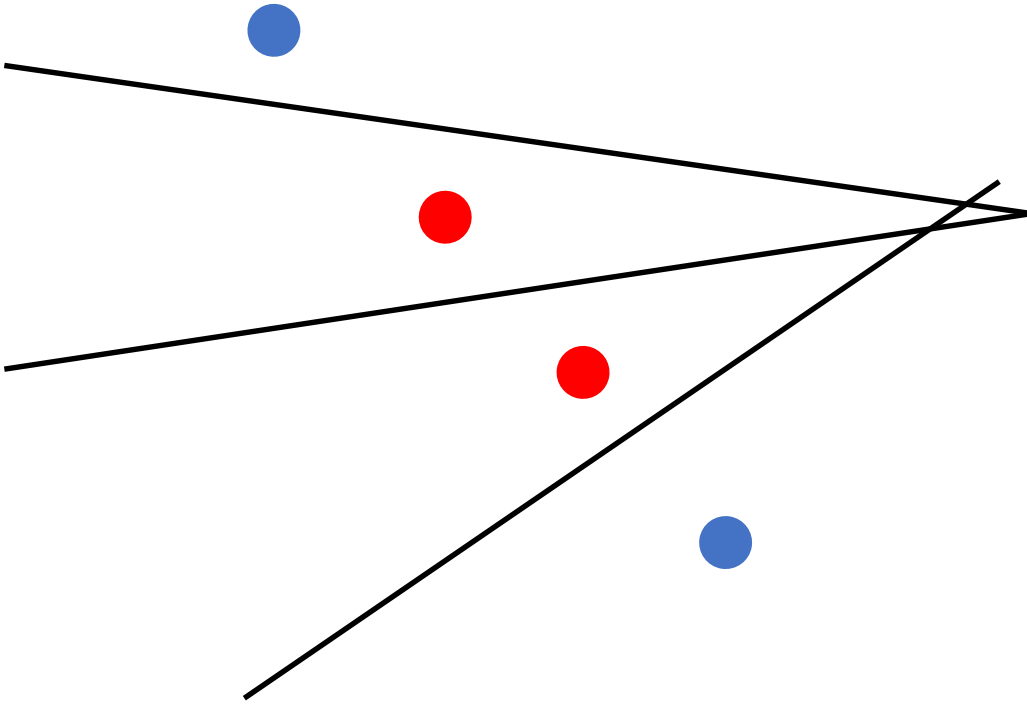


サポートマシーンベクトル  
(SVM)

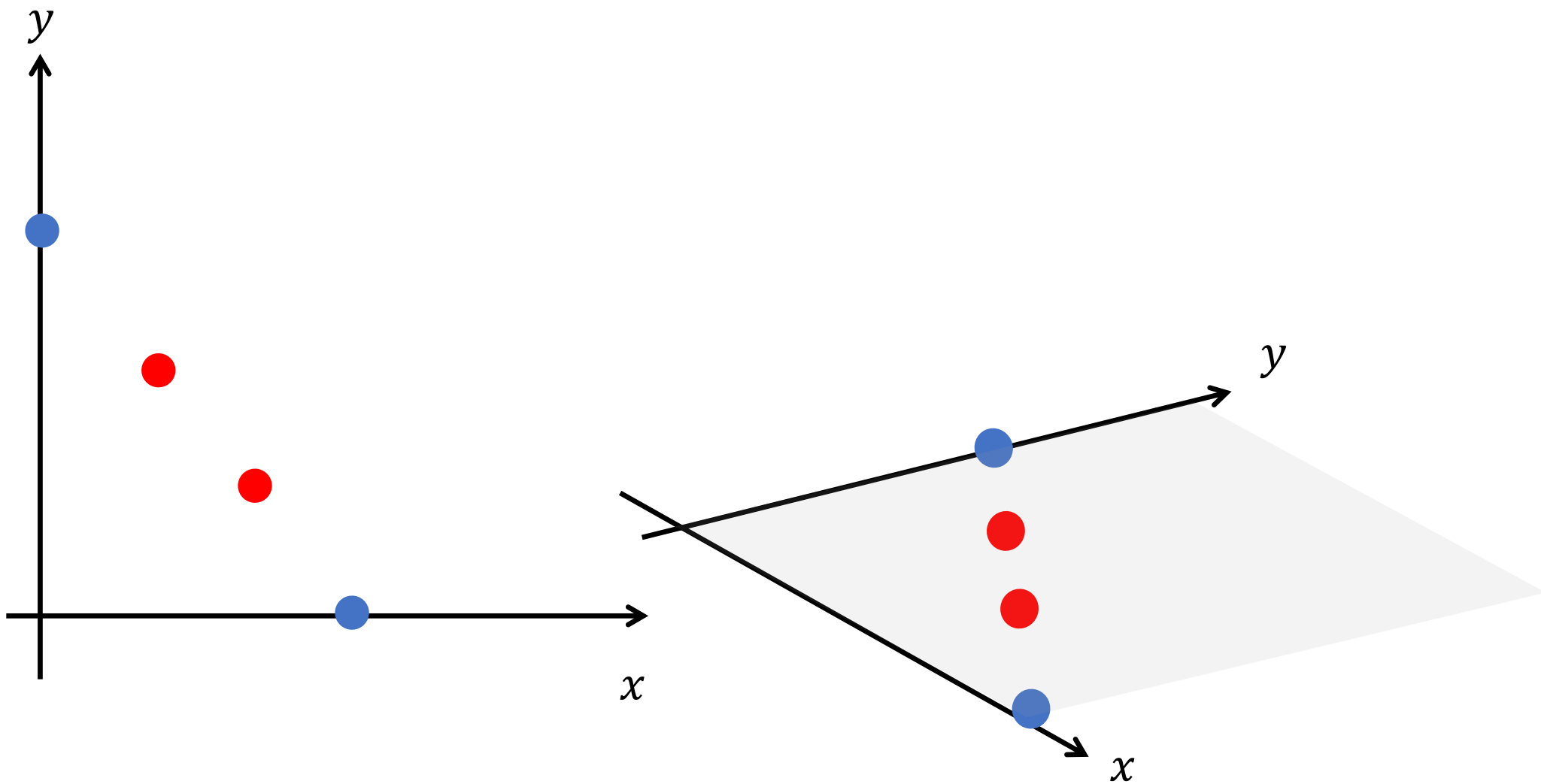


# いろいろな識別方法

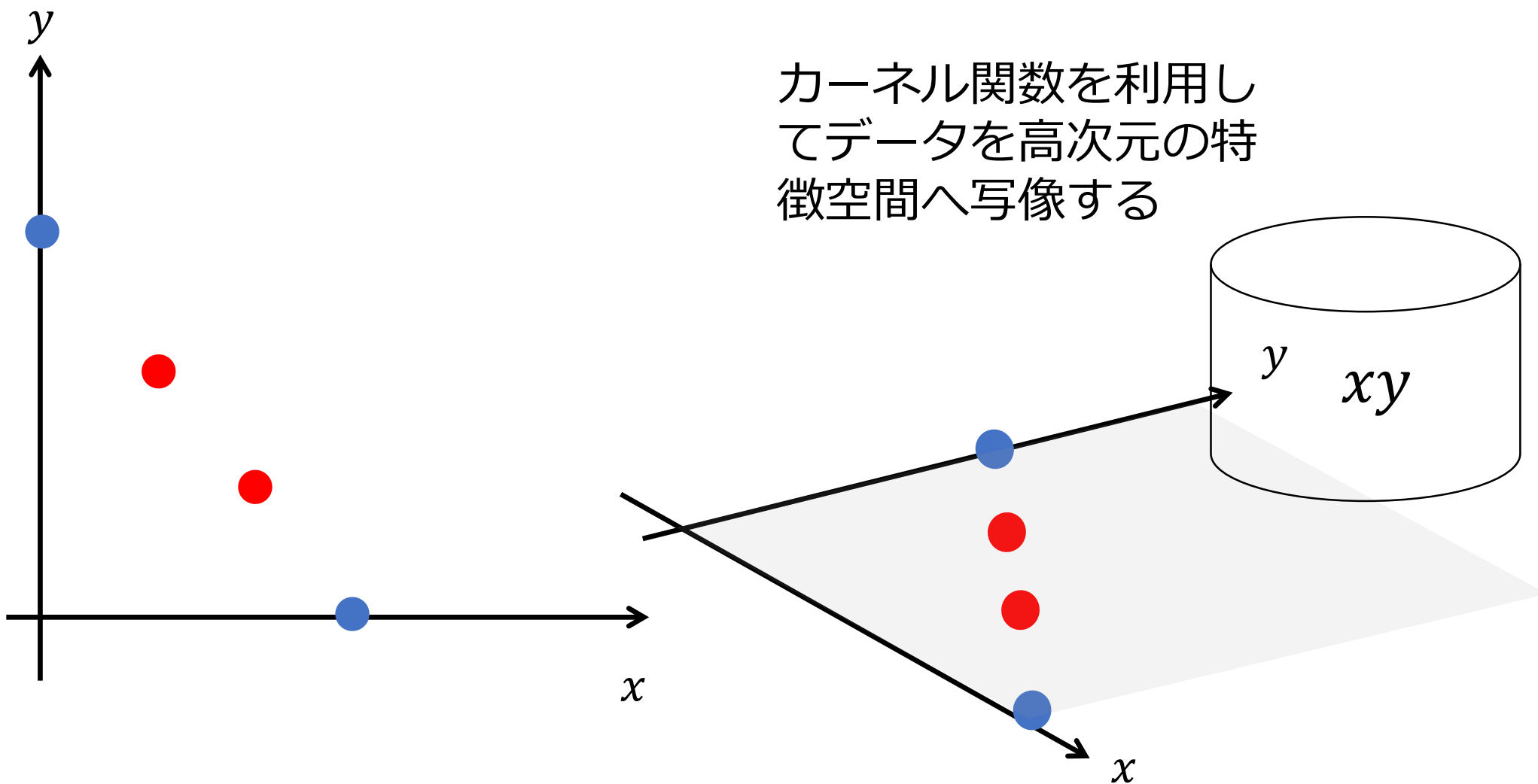
直線では分類できない？



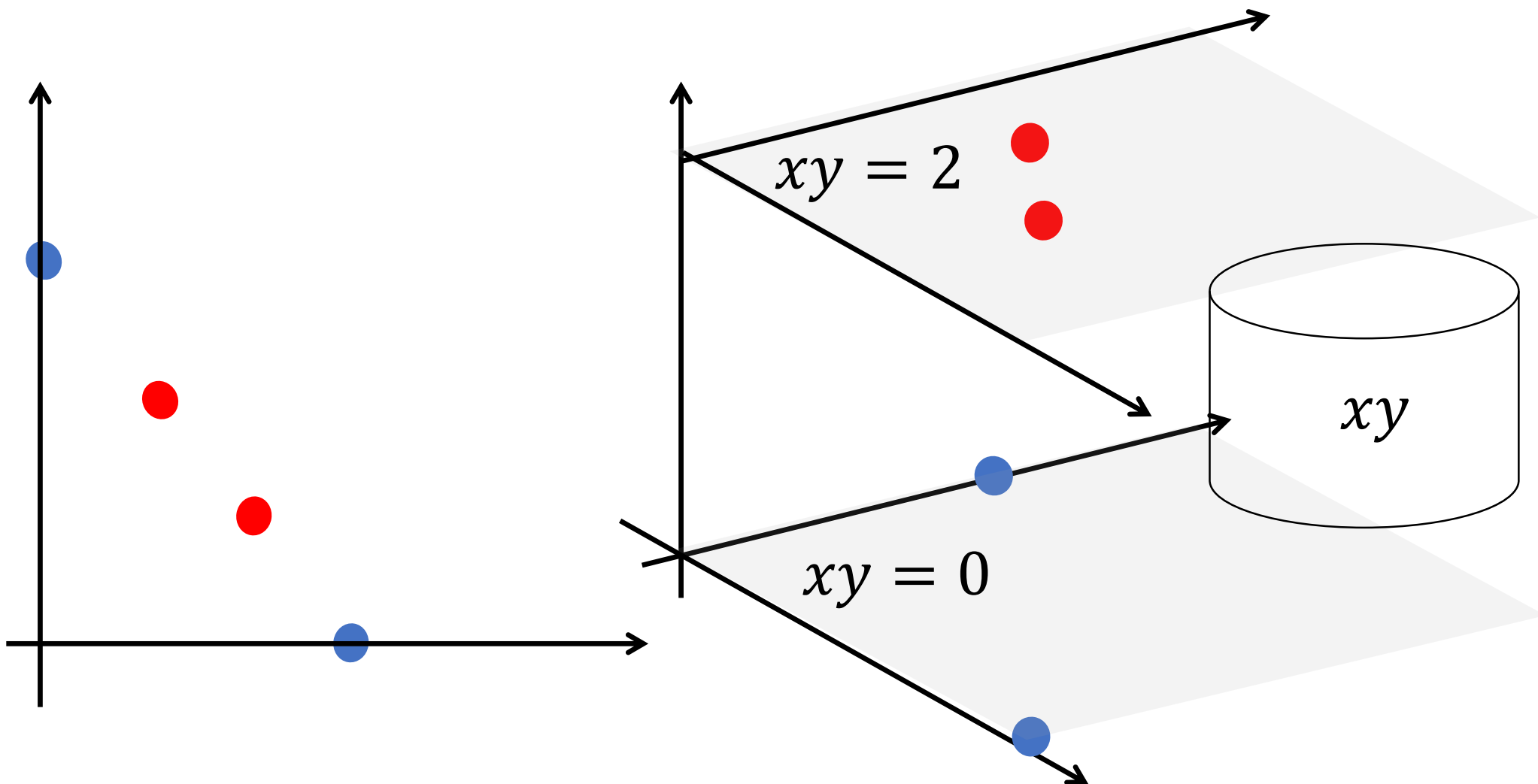
# カーネル・トリック



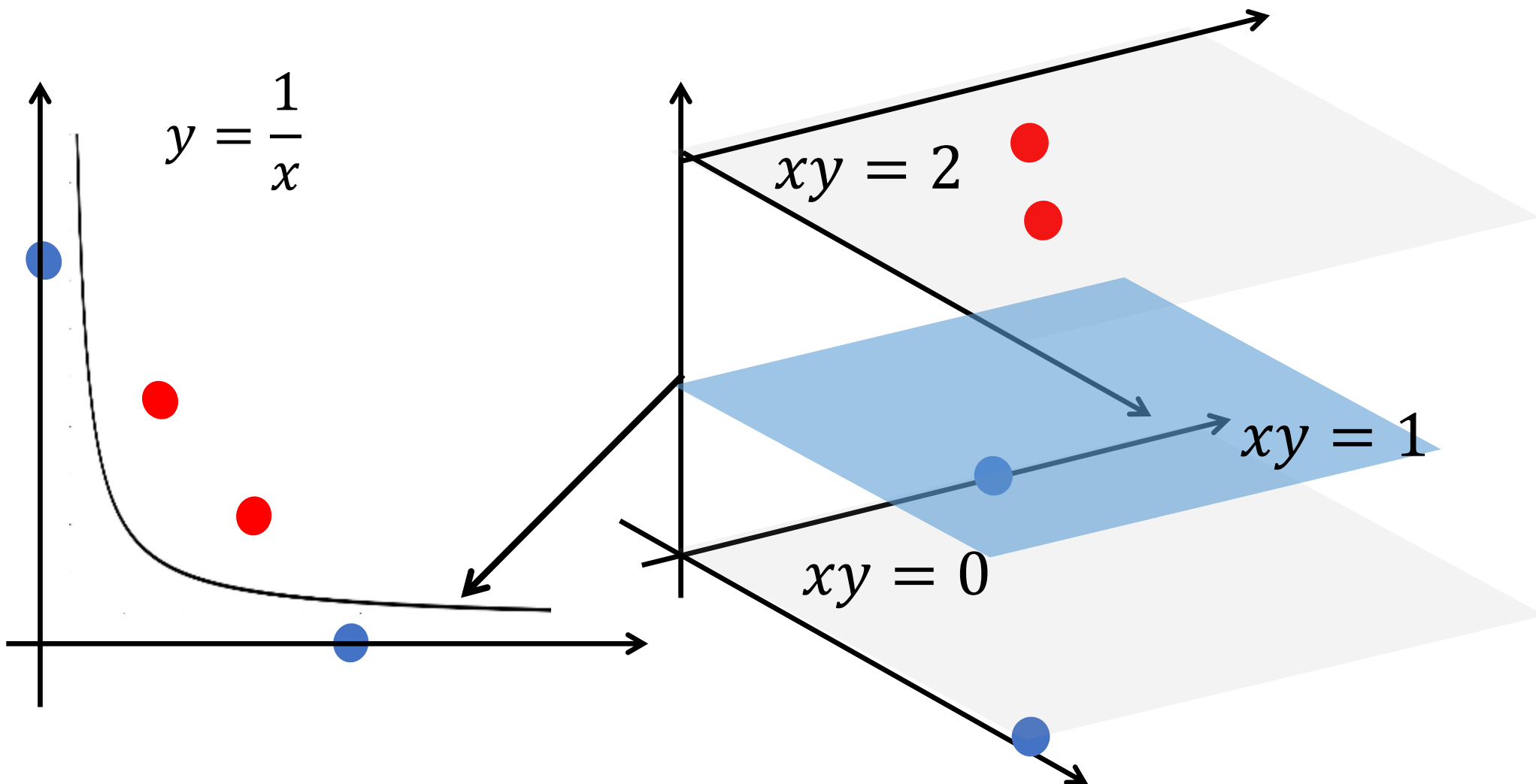
# カーネル・トリック



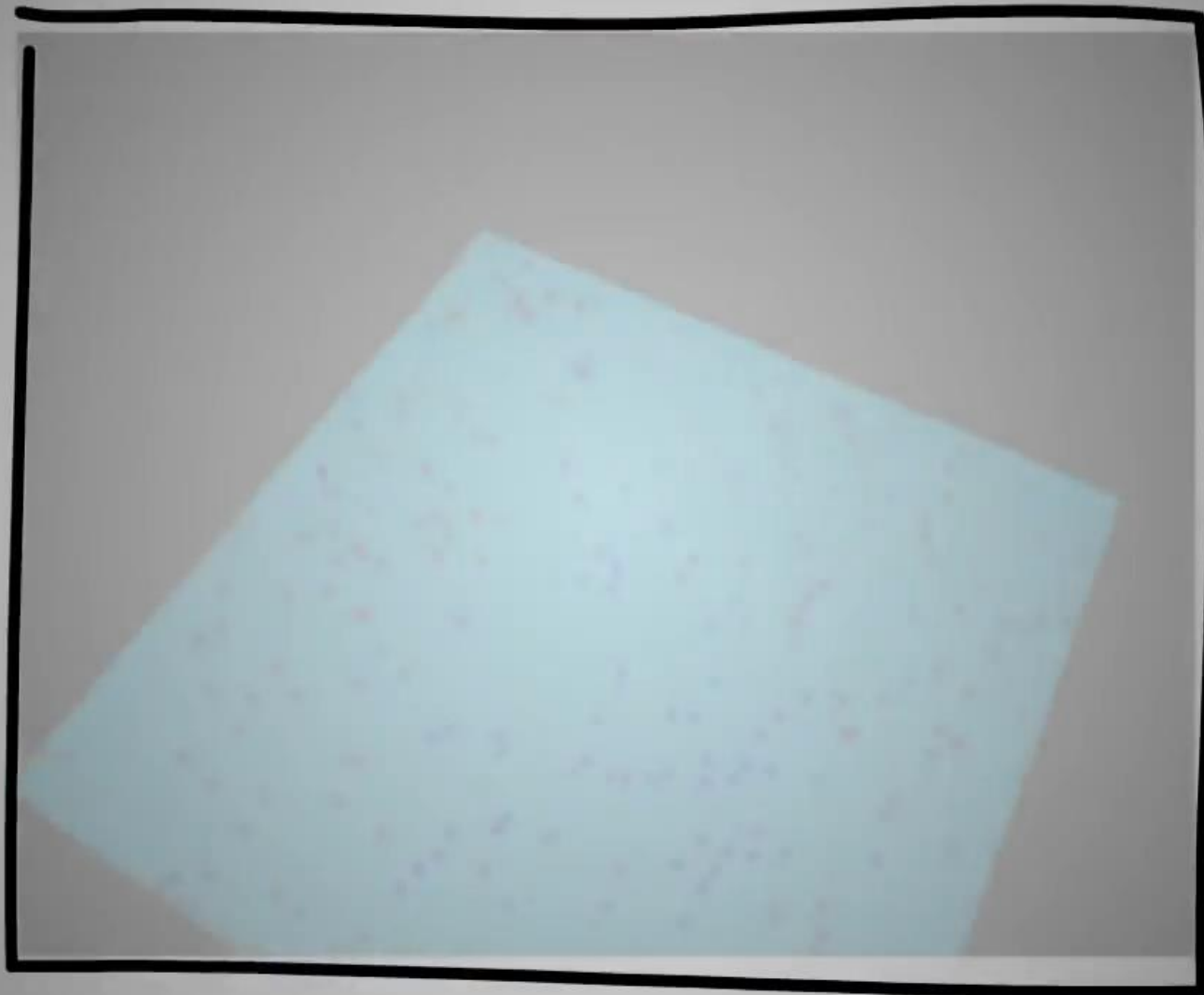
# カーネル・トリック



# カーネル・トリック



# カーネルトリック



## ・ 教師あり ・ 機械学習 ・ 回帰

# 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





# 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

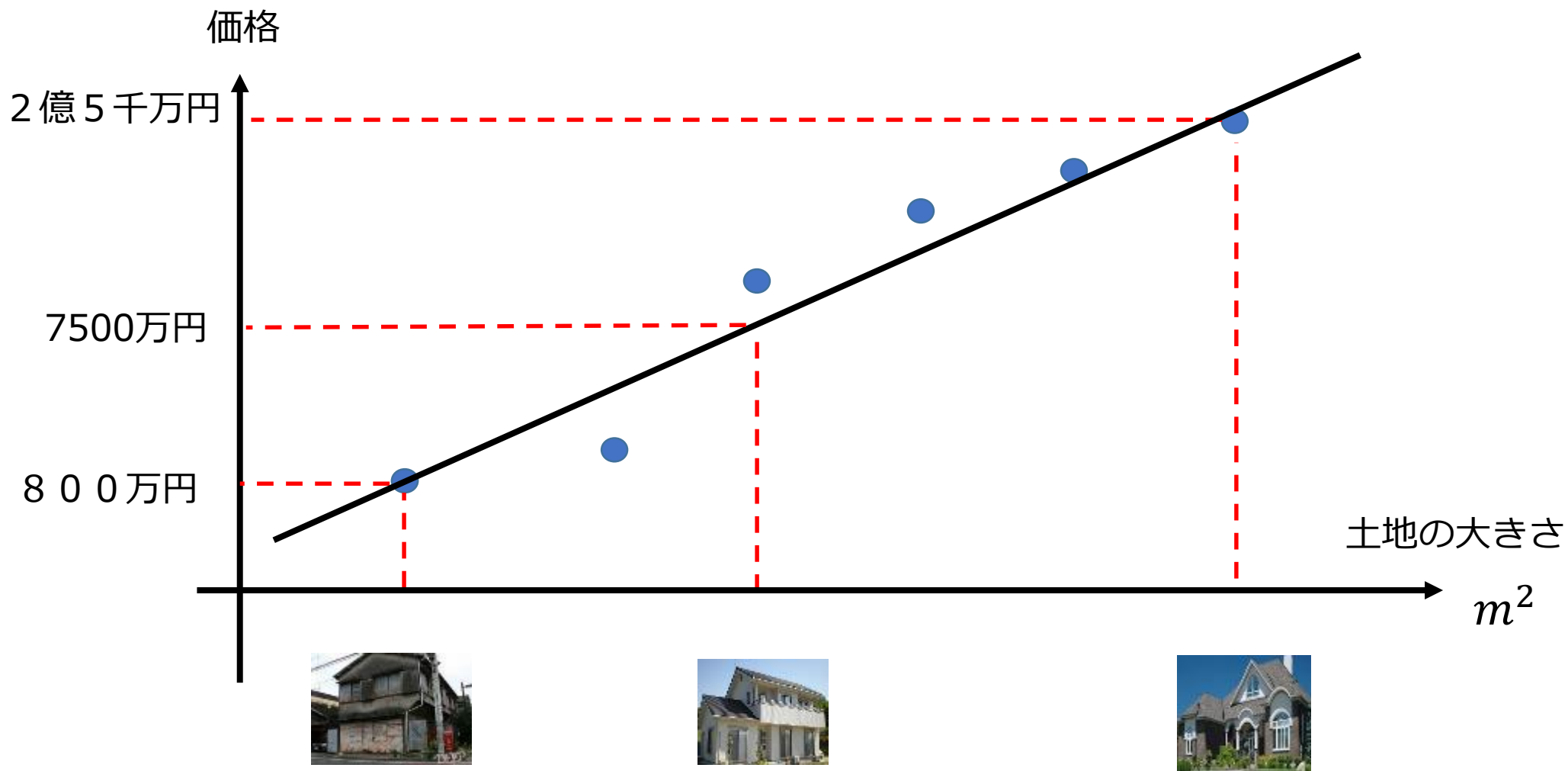


価格を予測する



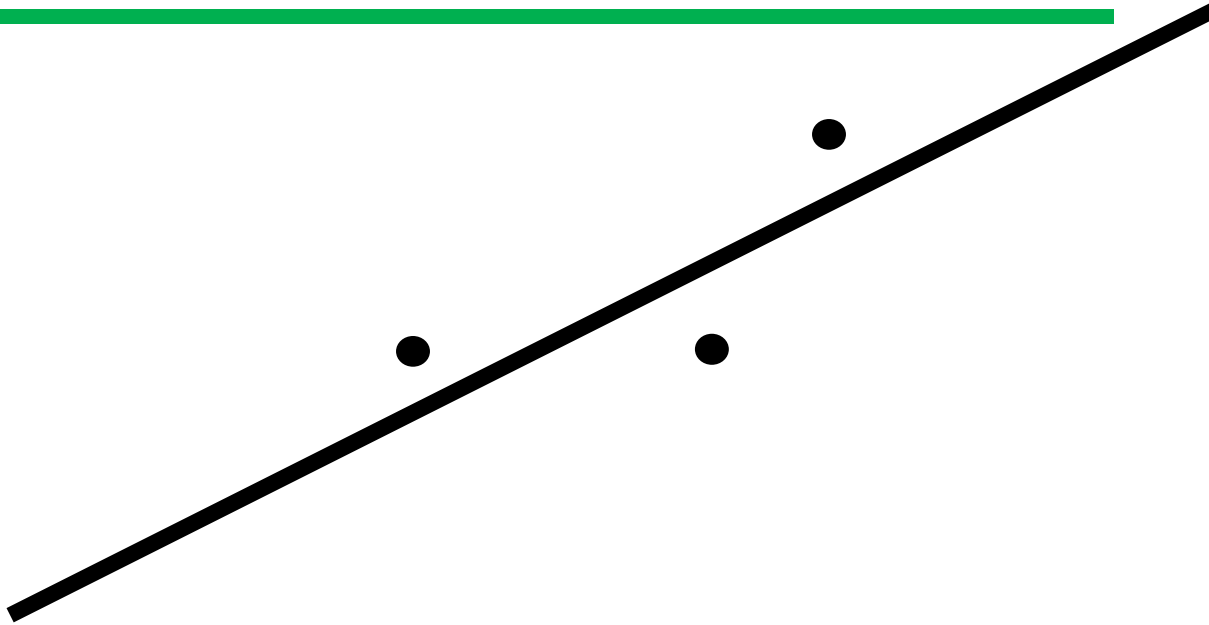
2億5千万円

# 回帰分析



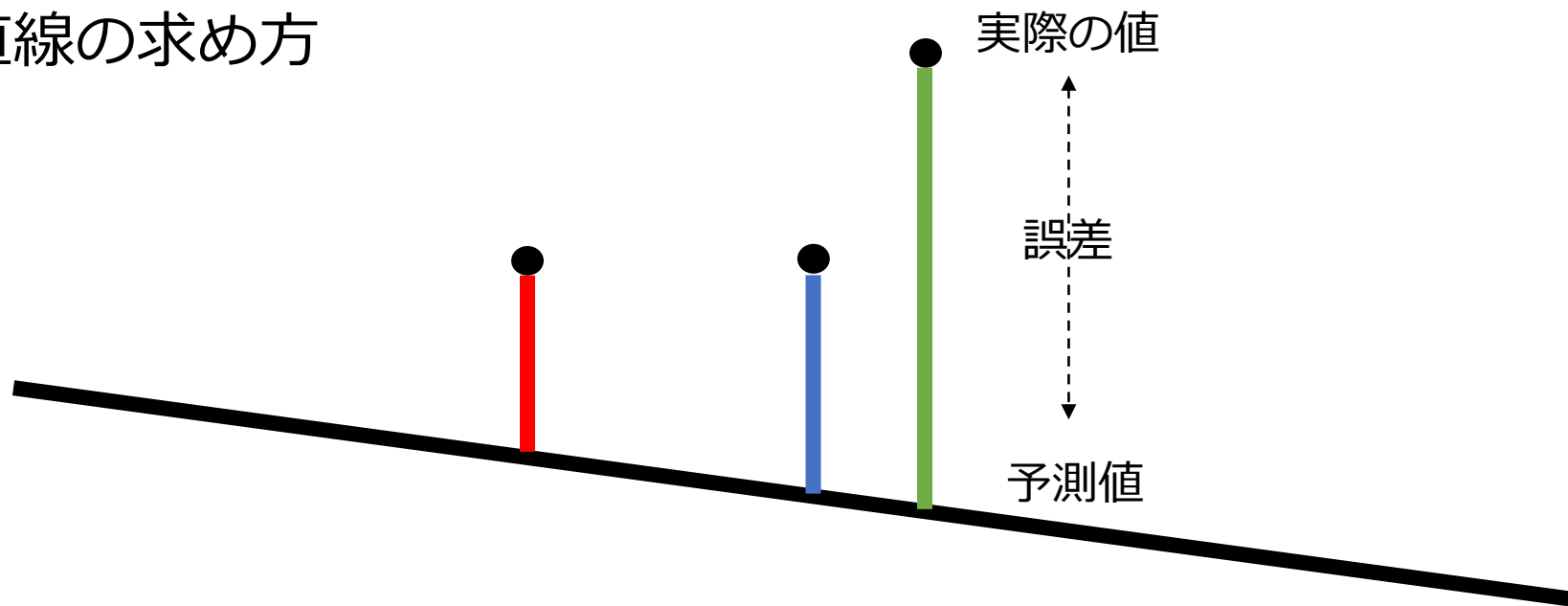
価格に影響を与える要因（築年数・広さ・・・）

# 回帰モデル



# 回帰モデル

直線の求め方

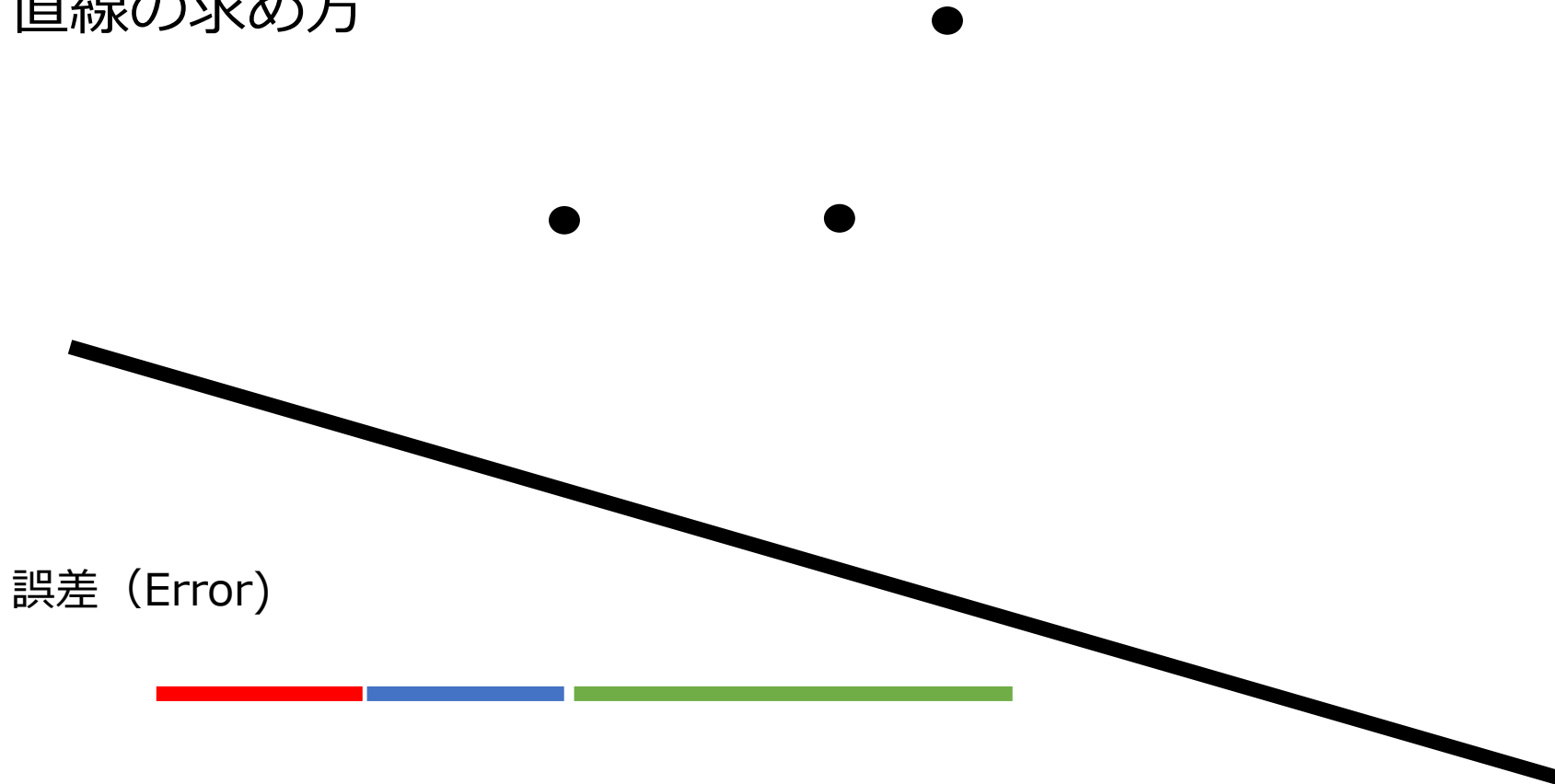


誤差 (Error)



# 回帰モデル

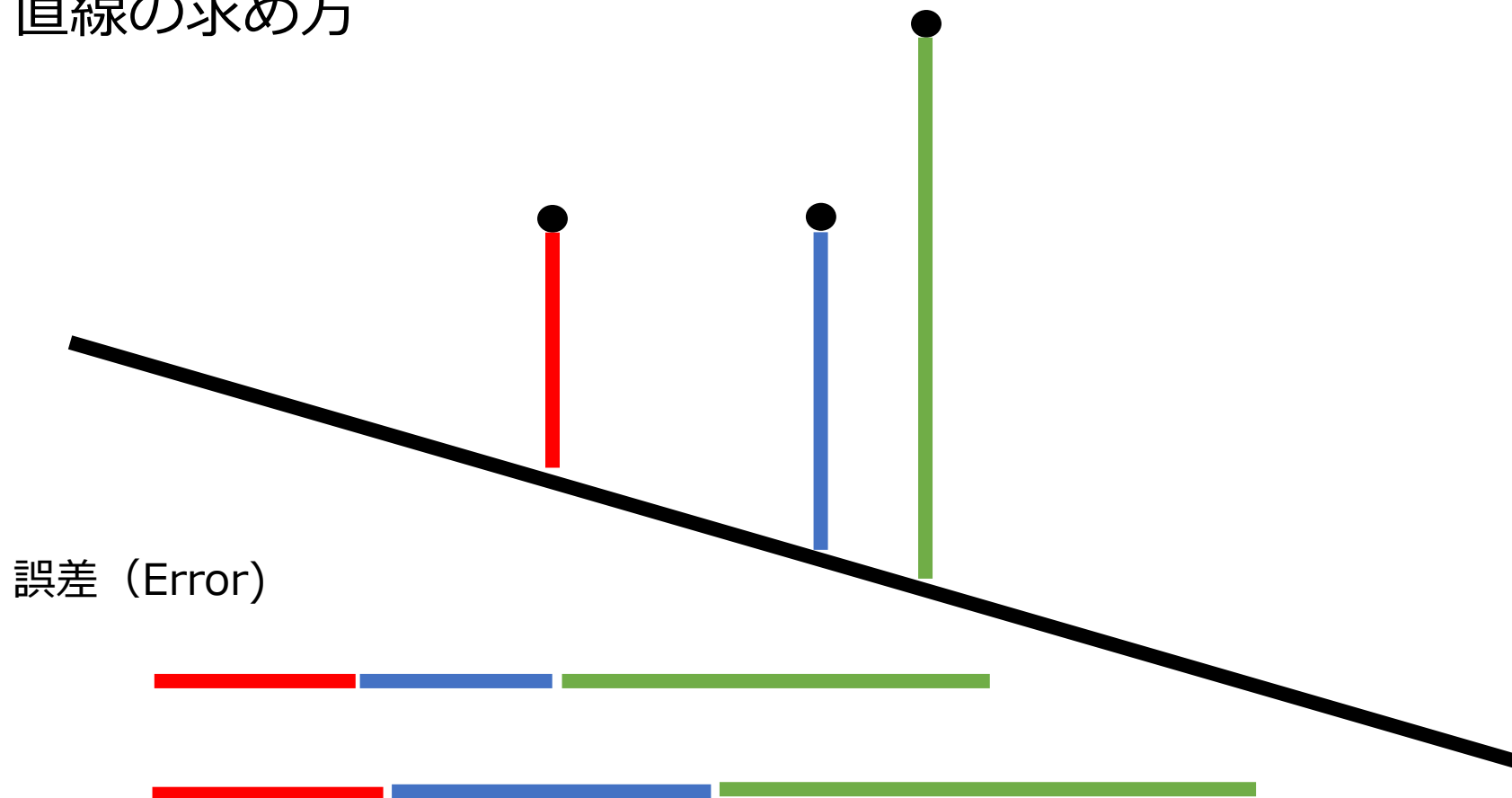
直線の求め方



誤差 (Error)

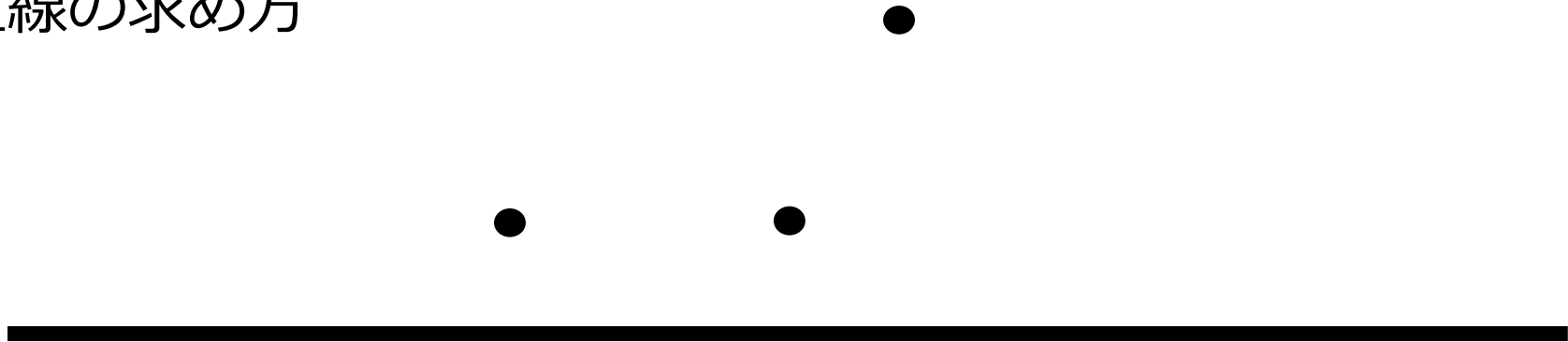
# 回帰モデル

直線の求め方



# 回帰モデル

直線の求め方

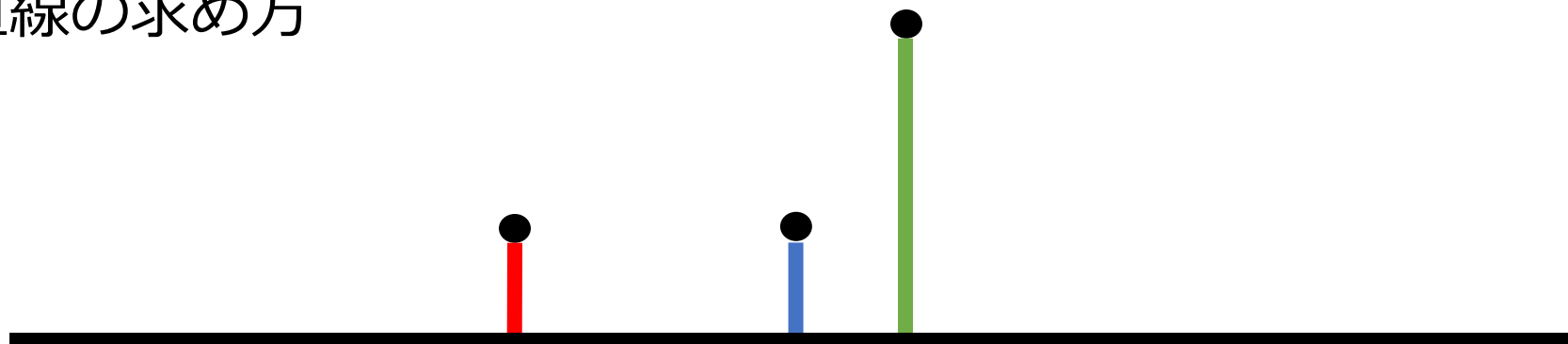


誤差 (Error)



# 回帰モデル

## 直線の求め方



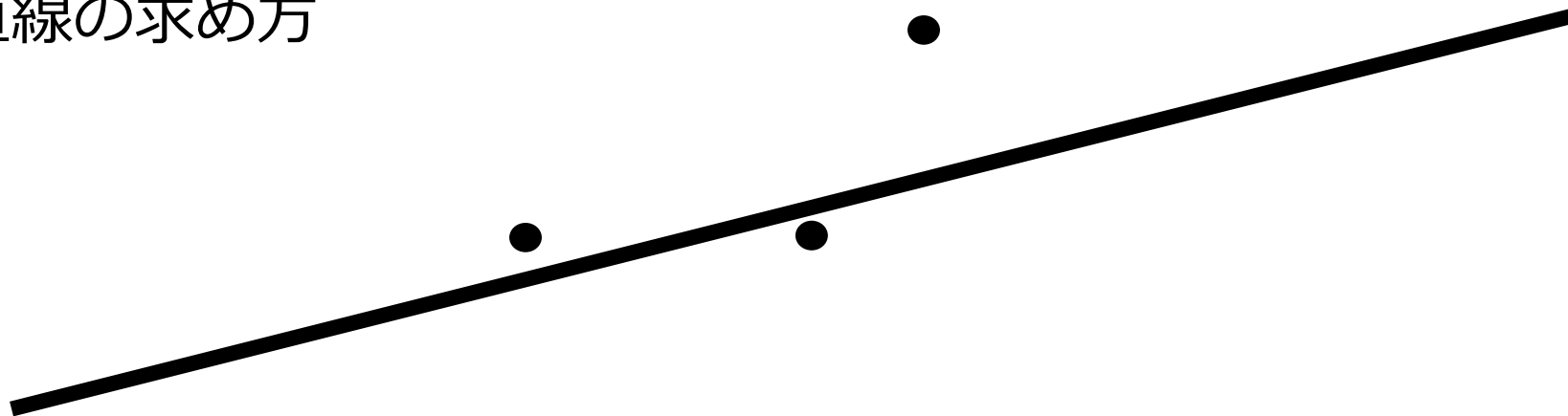
## 誤差 (Error)



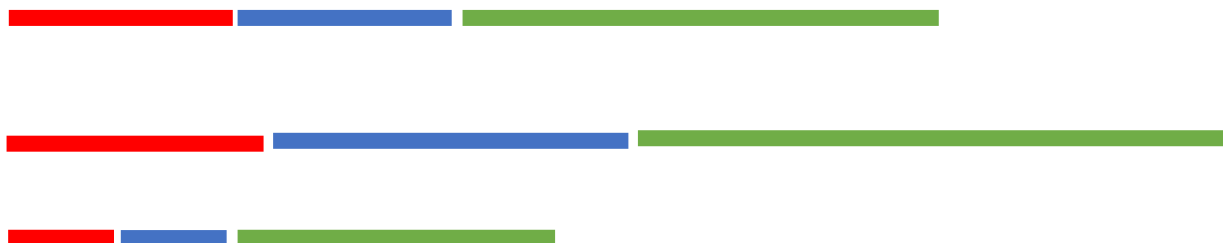


# 回帰モデル

直線の求め方

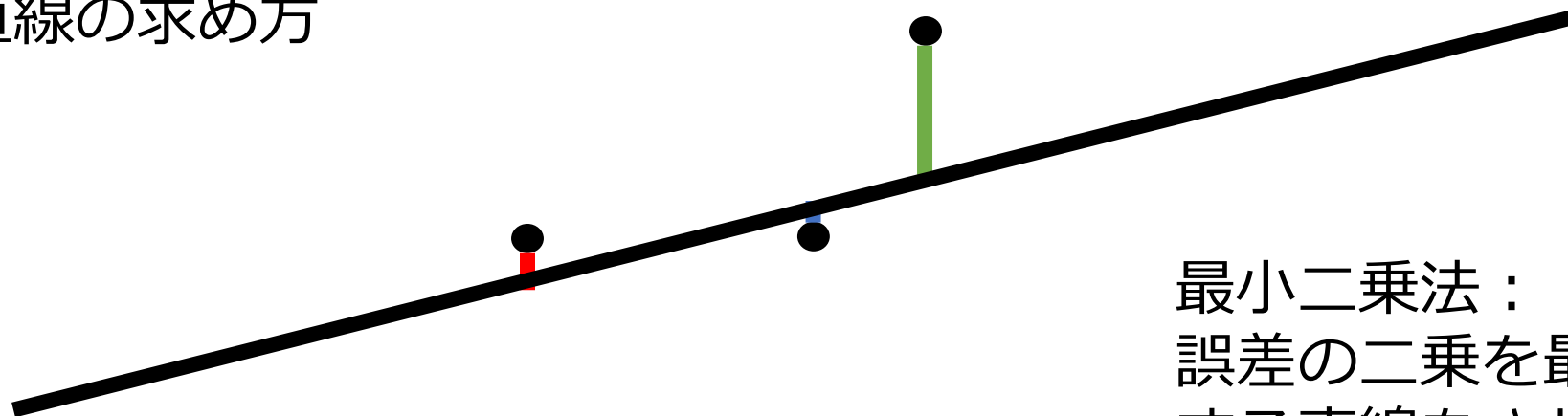


誤差 (Error)



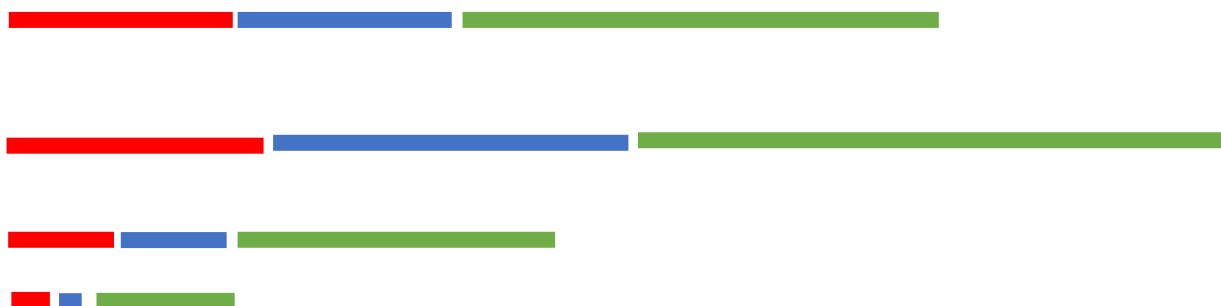
# 回帰モデル

直線の求め方

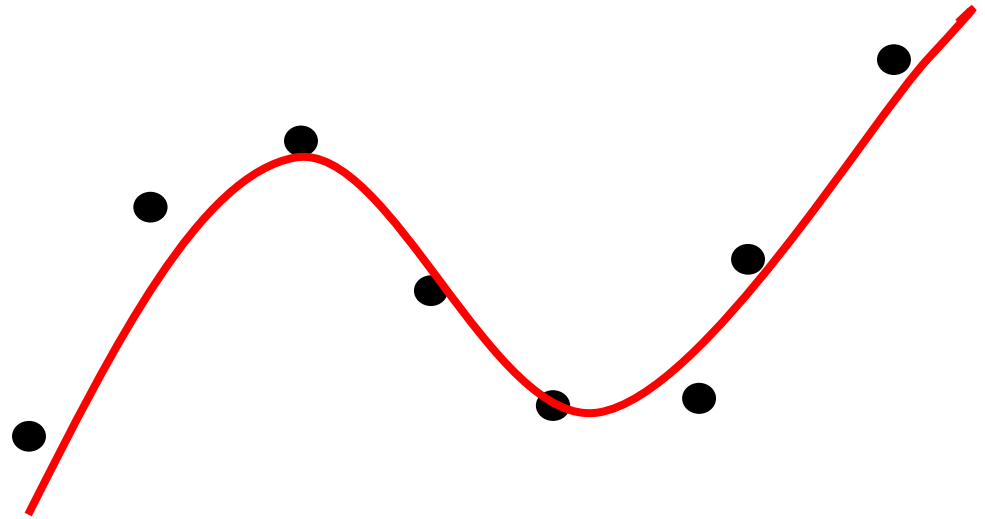
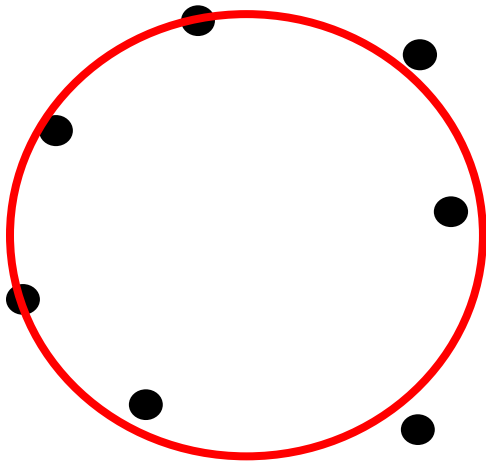


最小二乗法：  
誤差の二乗を最小にする直線をさがす

誤差 (Error)



# 様々な回帰モデル



[https://kwichmann.github.io/ml\\_sandbox/linear\\_regression\\_diagnostics/](https://kwichmann.github.io/ml_sandbox/linear_regression_diagnostics/)

## 教師なし・機械学習・データの分類

# 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



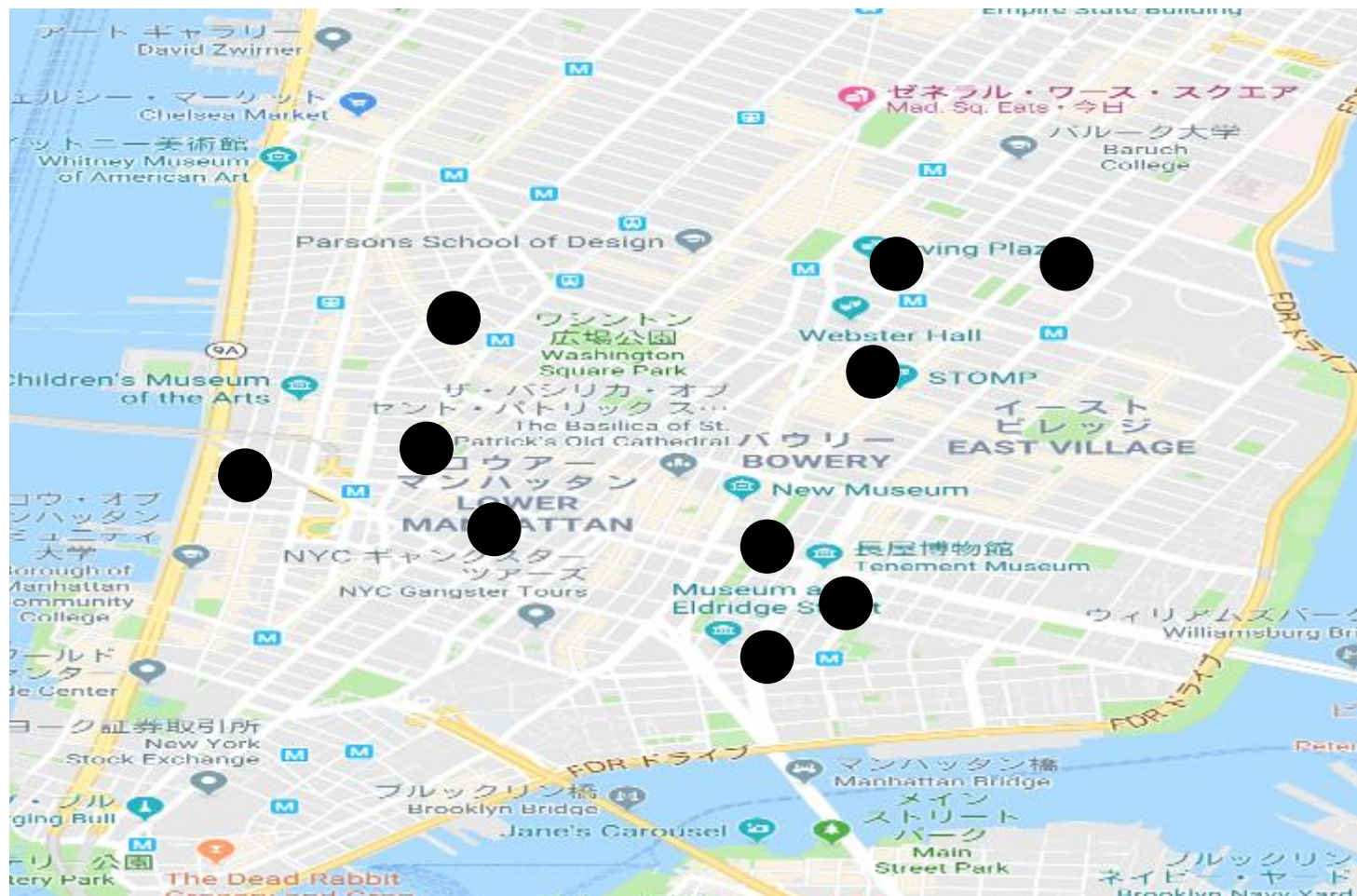
# クラスター分析

---

- K-mean法

# どこにピザ屋を出店するか？

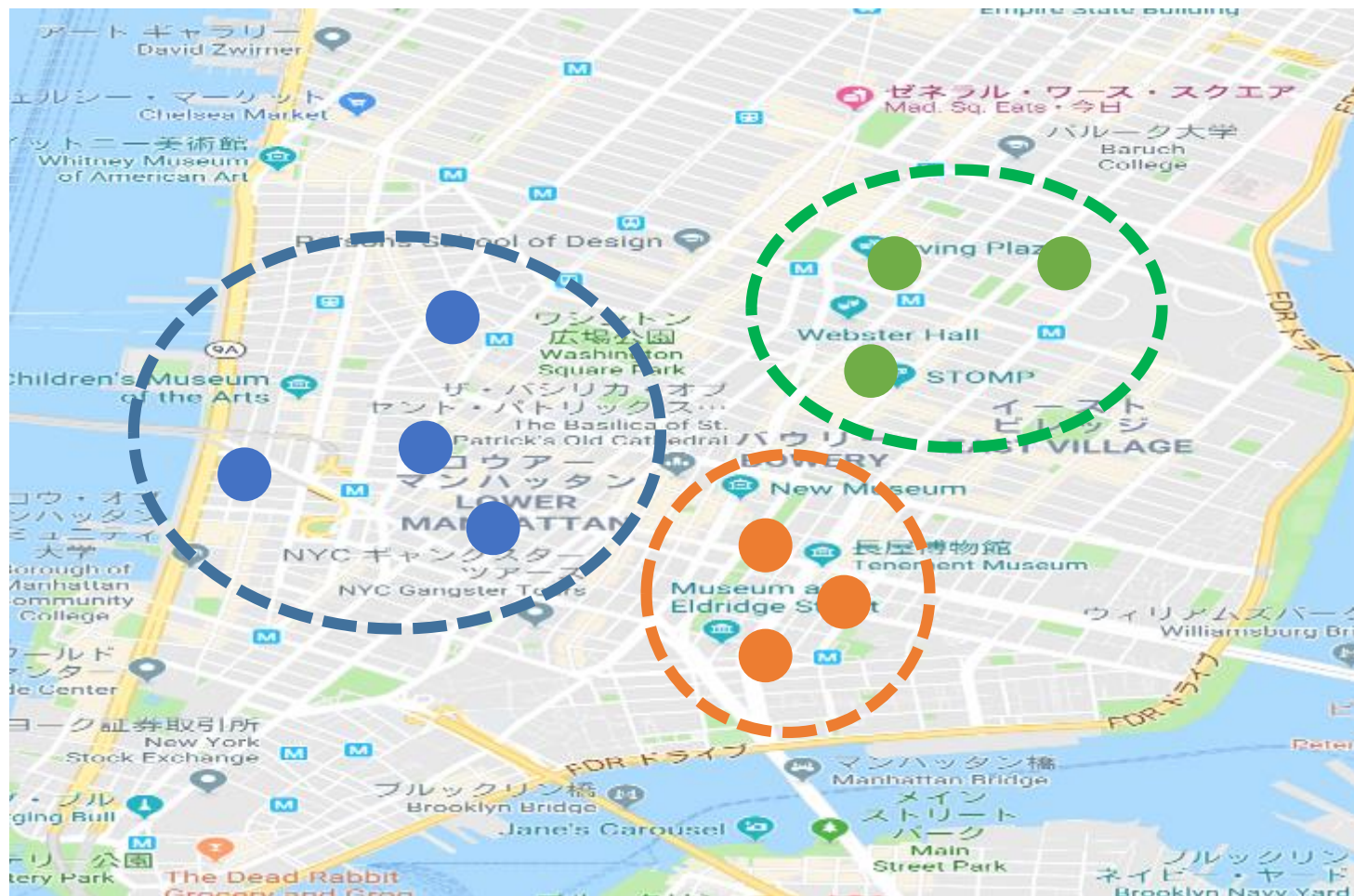
3つのグループに分けるとしたら、どのようなグループ分けを行うか？





# どこにピザ屋を出店するか？

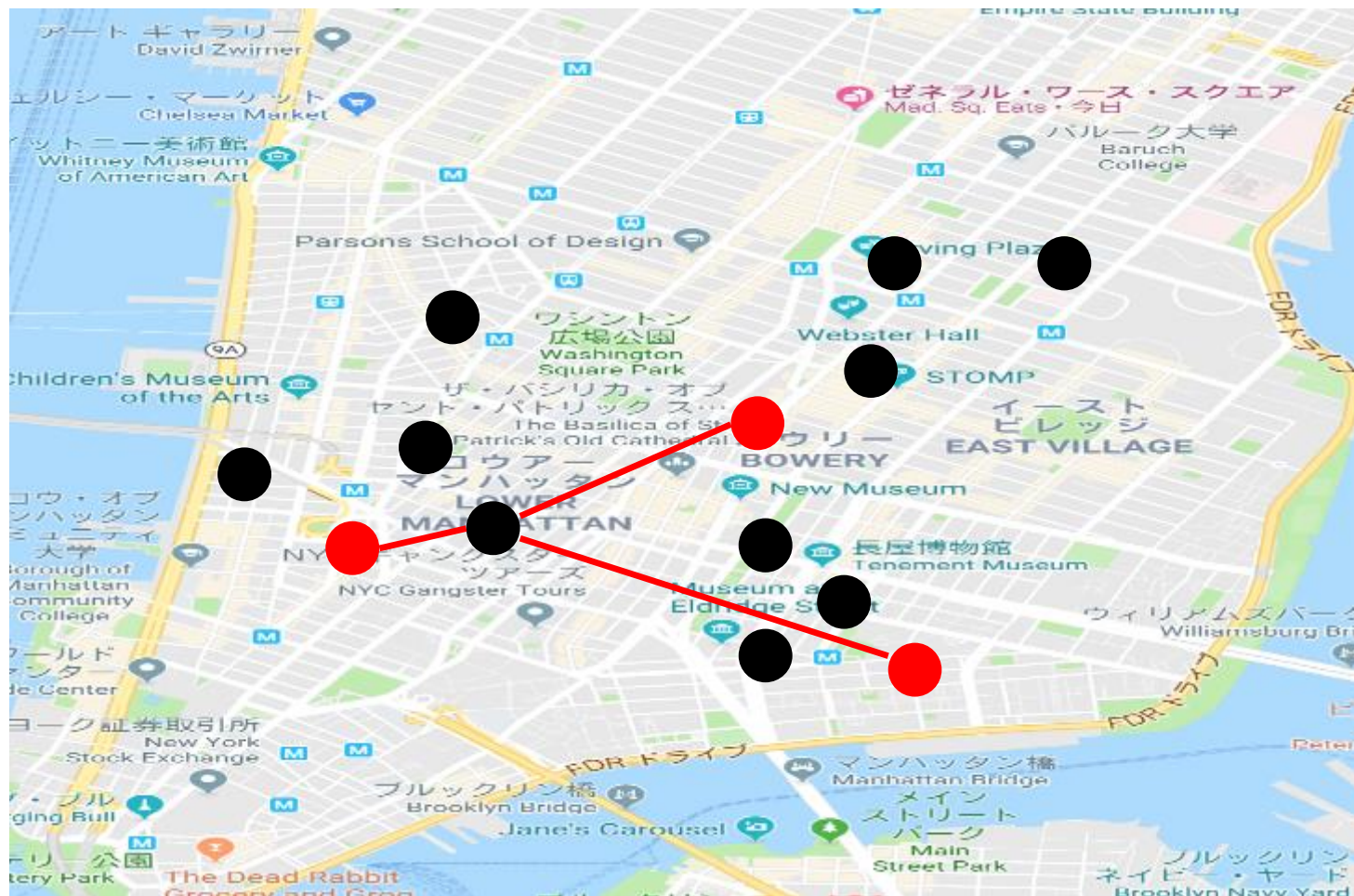
3つのグループに分けるとしたら、どのようなグループ分けを行うか？





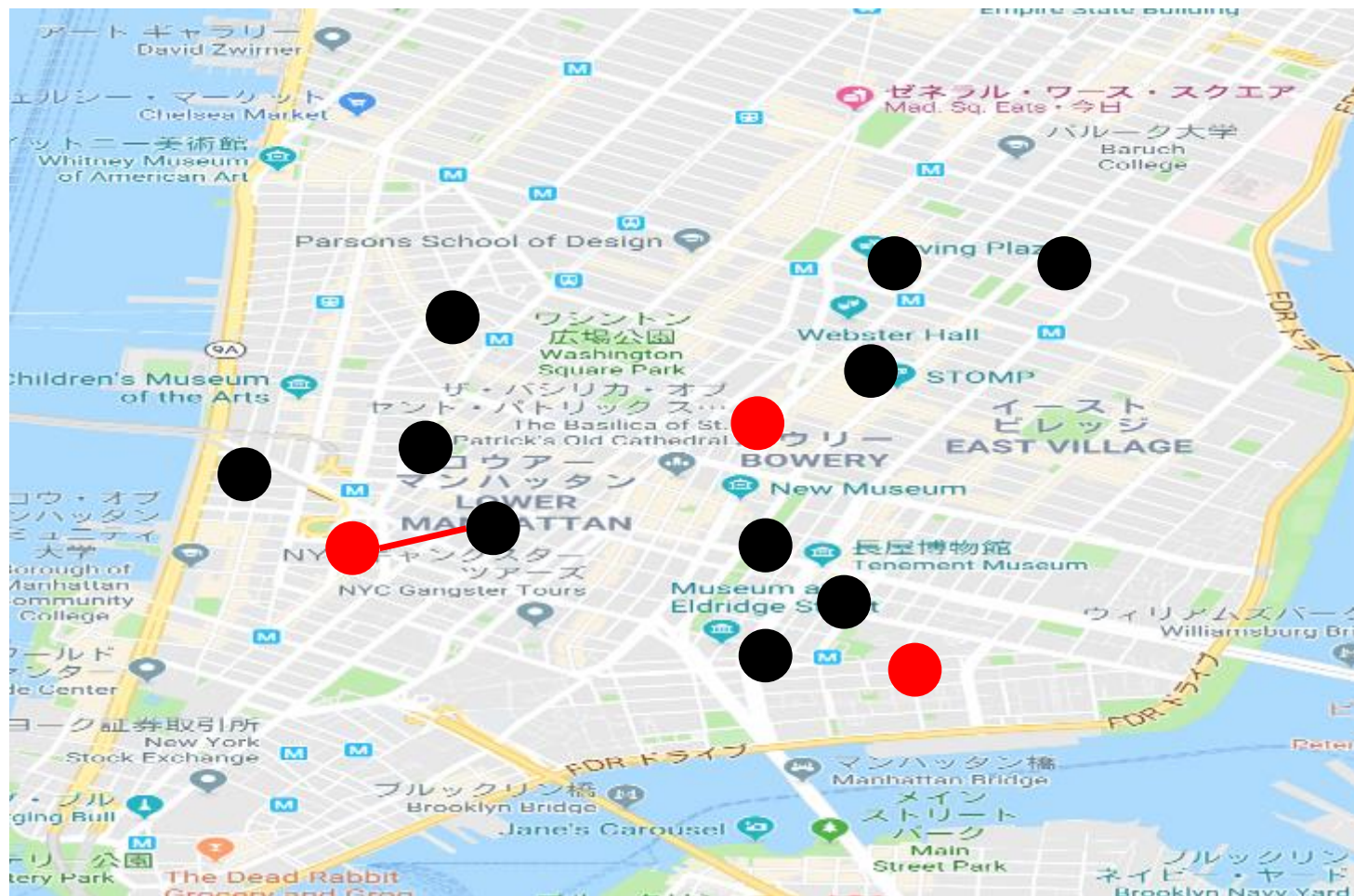
# K-mean法

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

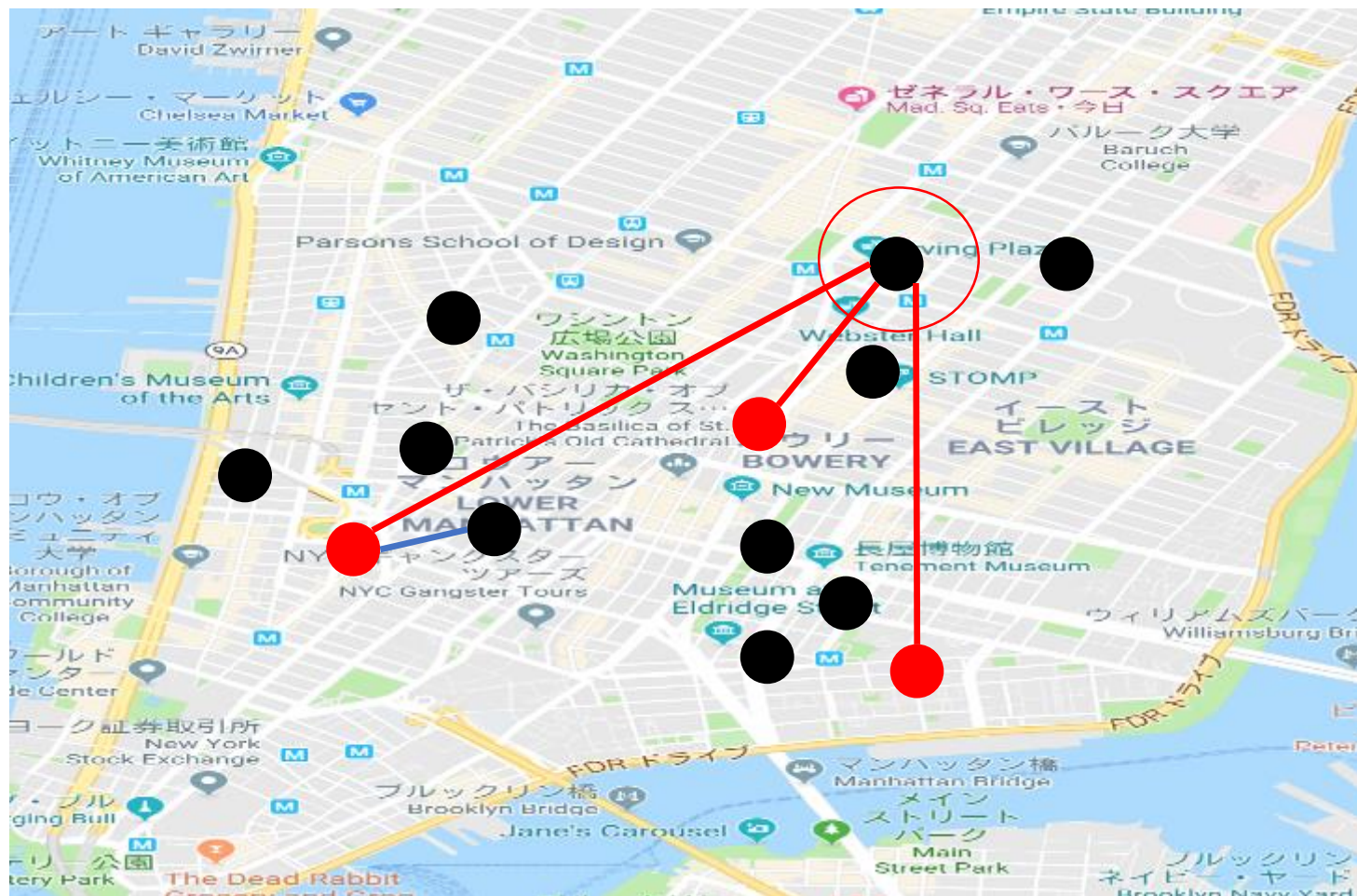
各点からシードまでの距離を測り、一番距離の短いシードと紐づける





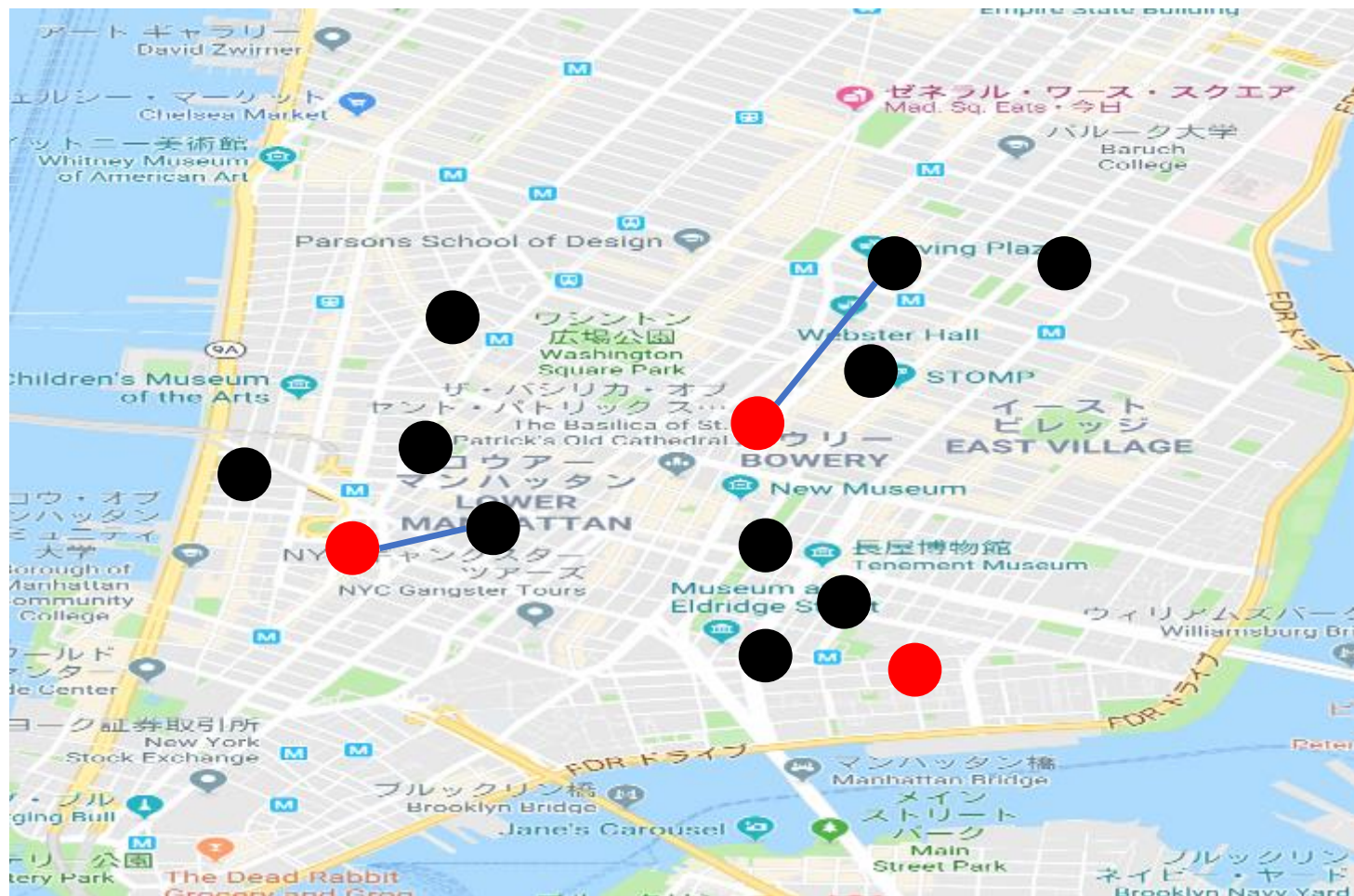
# K-mean法

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

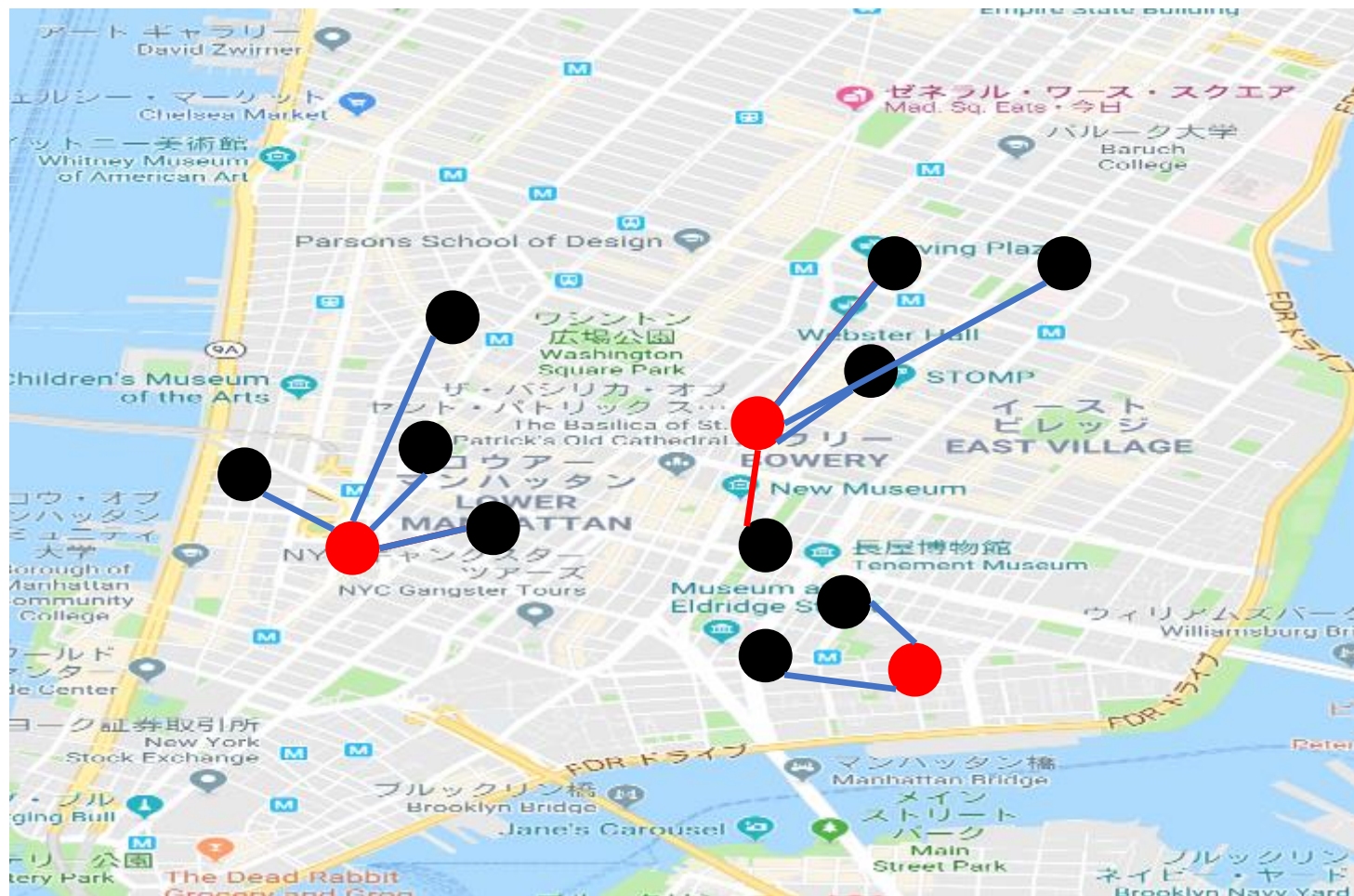
各点からシードまでの距離を測り、一番距離の短いシードと紐づける





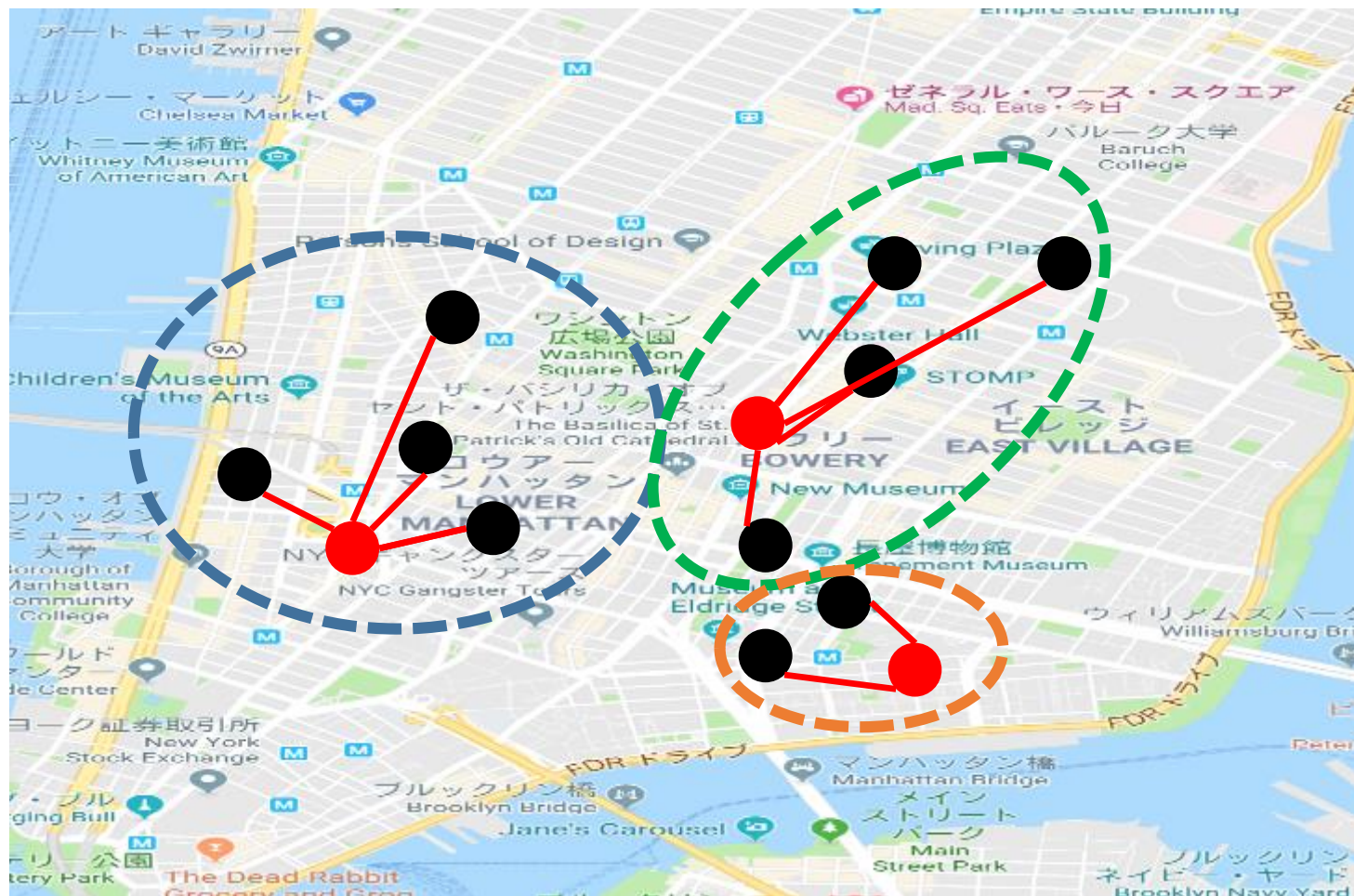
# K-mean法

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。





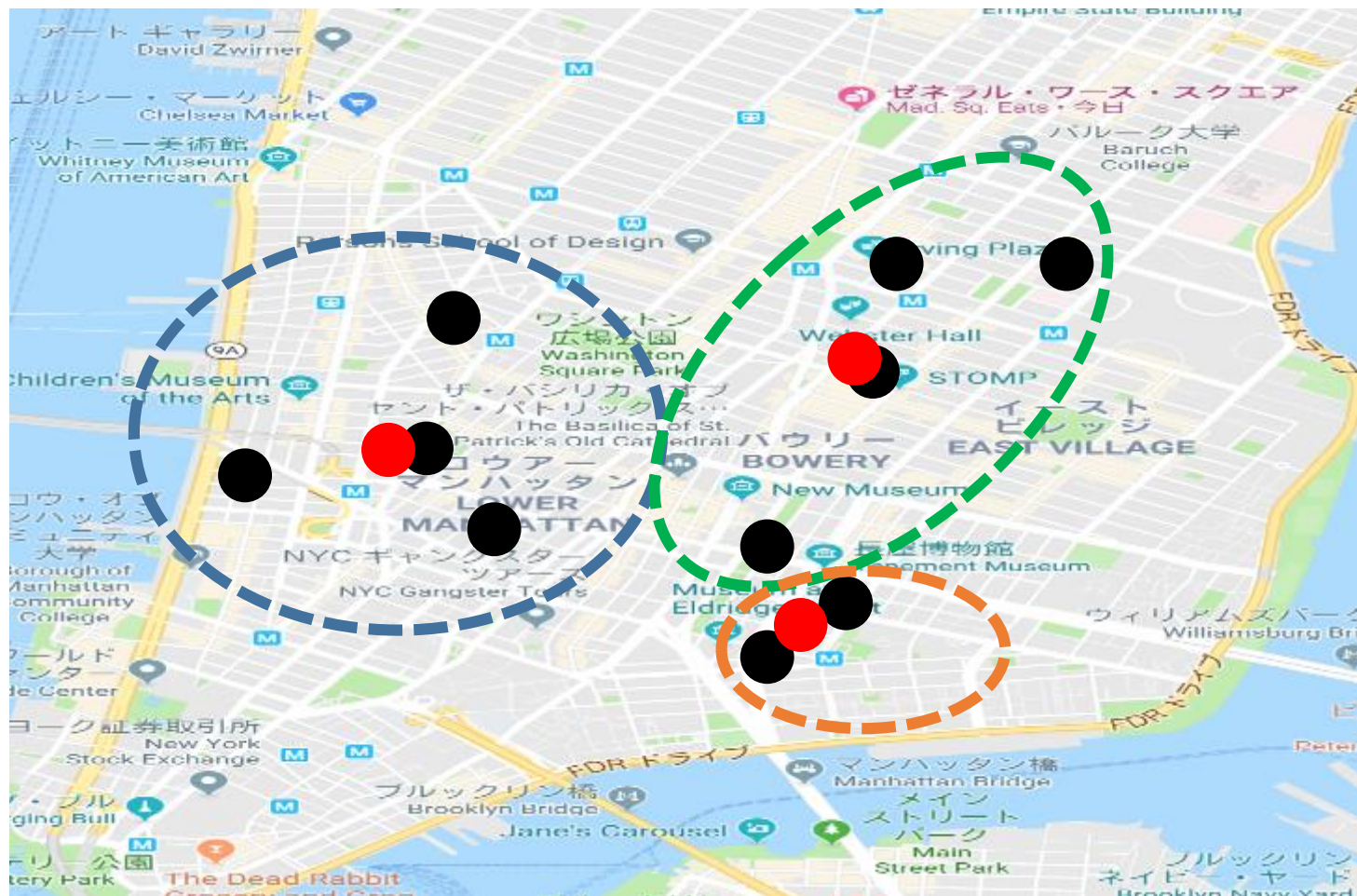
# K-mean法

クラスターが出来たらシードを除去する。



# K-mean法

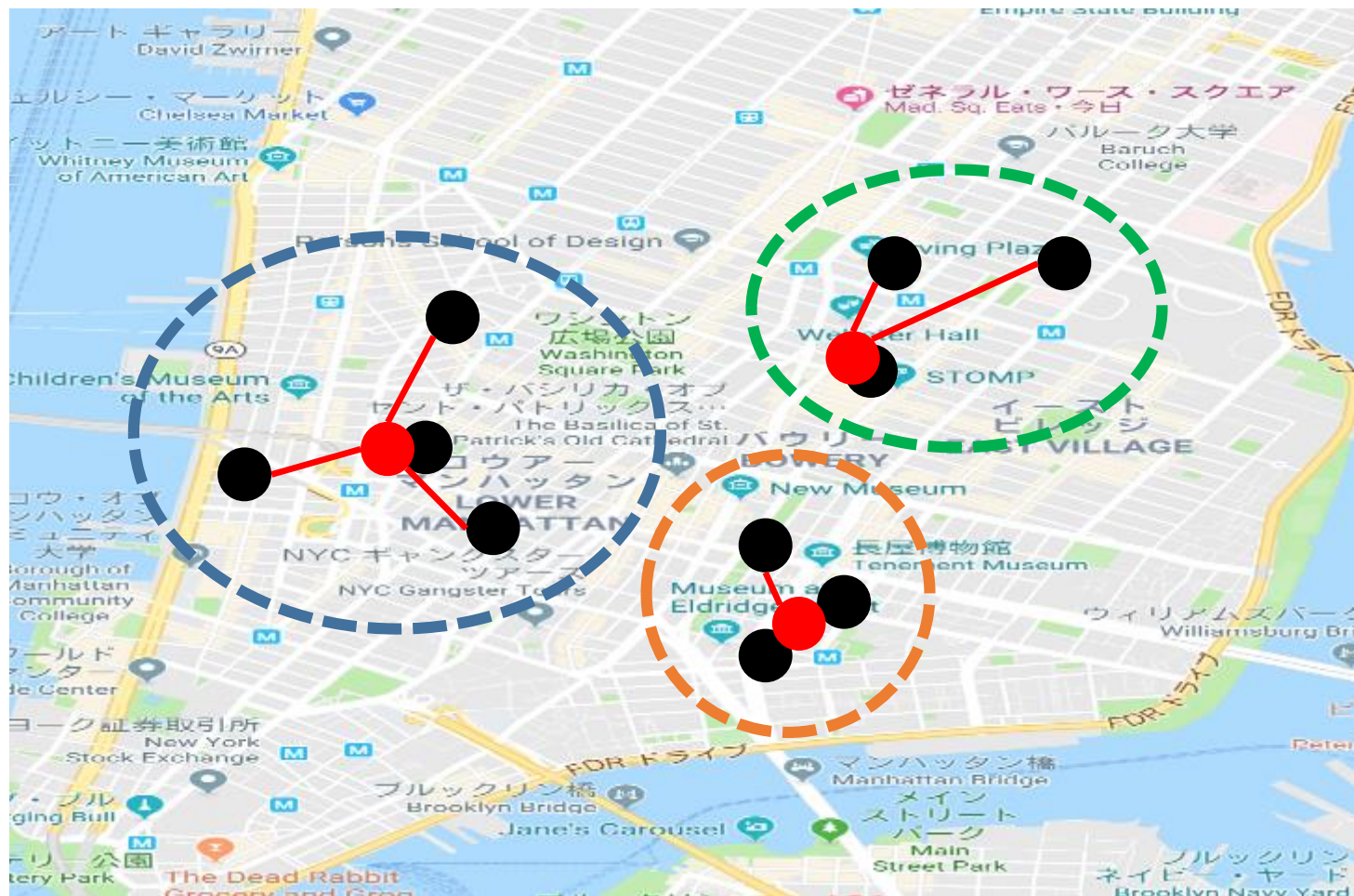
各クラスター内のデータの平均点(重心)を新たなシードとする。





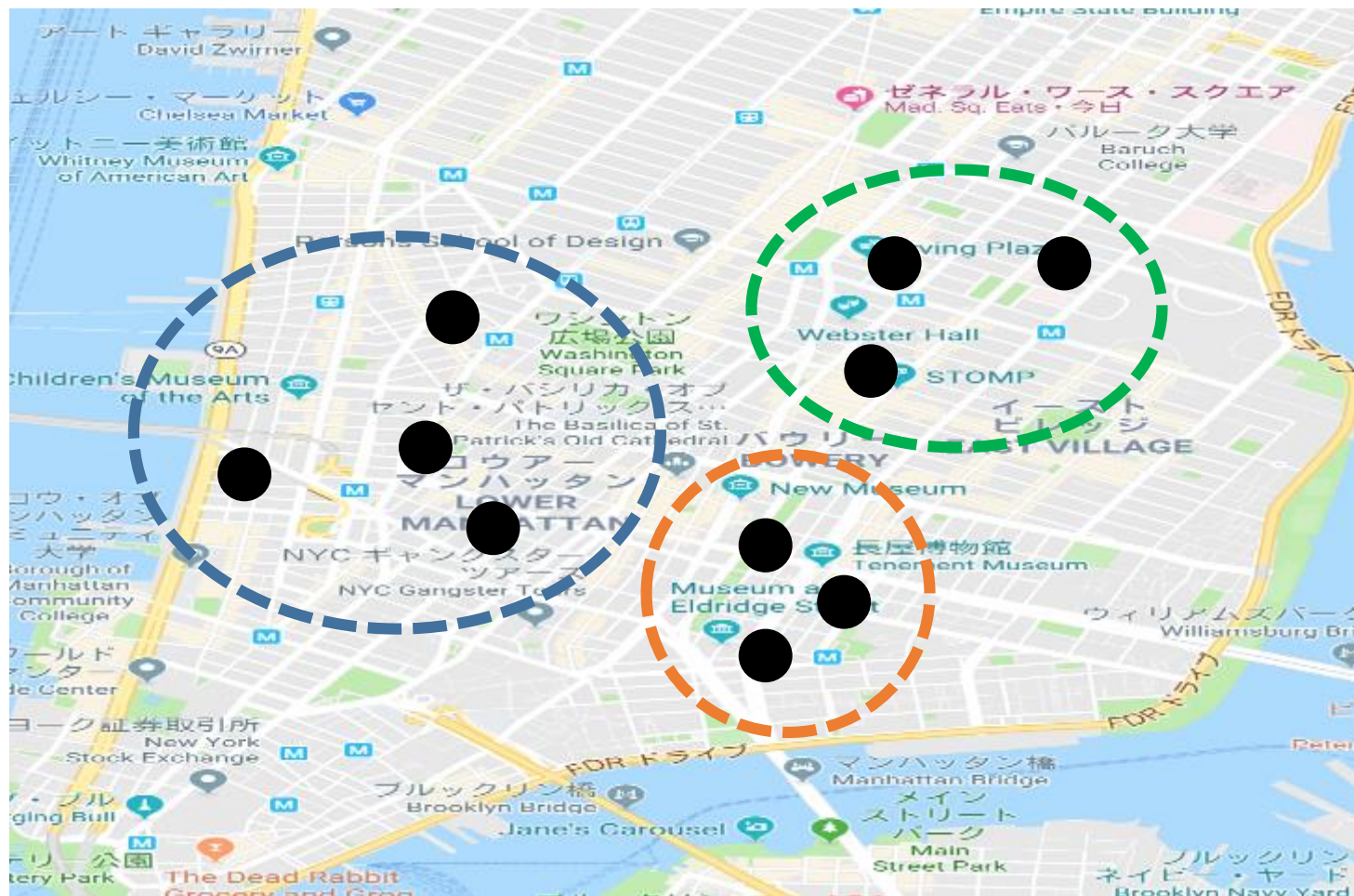
# K-mean法

各データと最も近いシードを紐づける。



# K-mean法

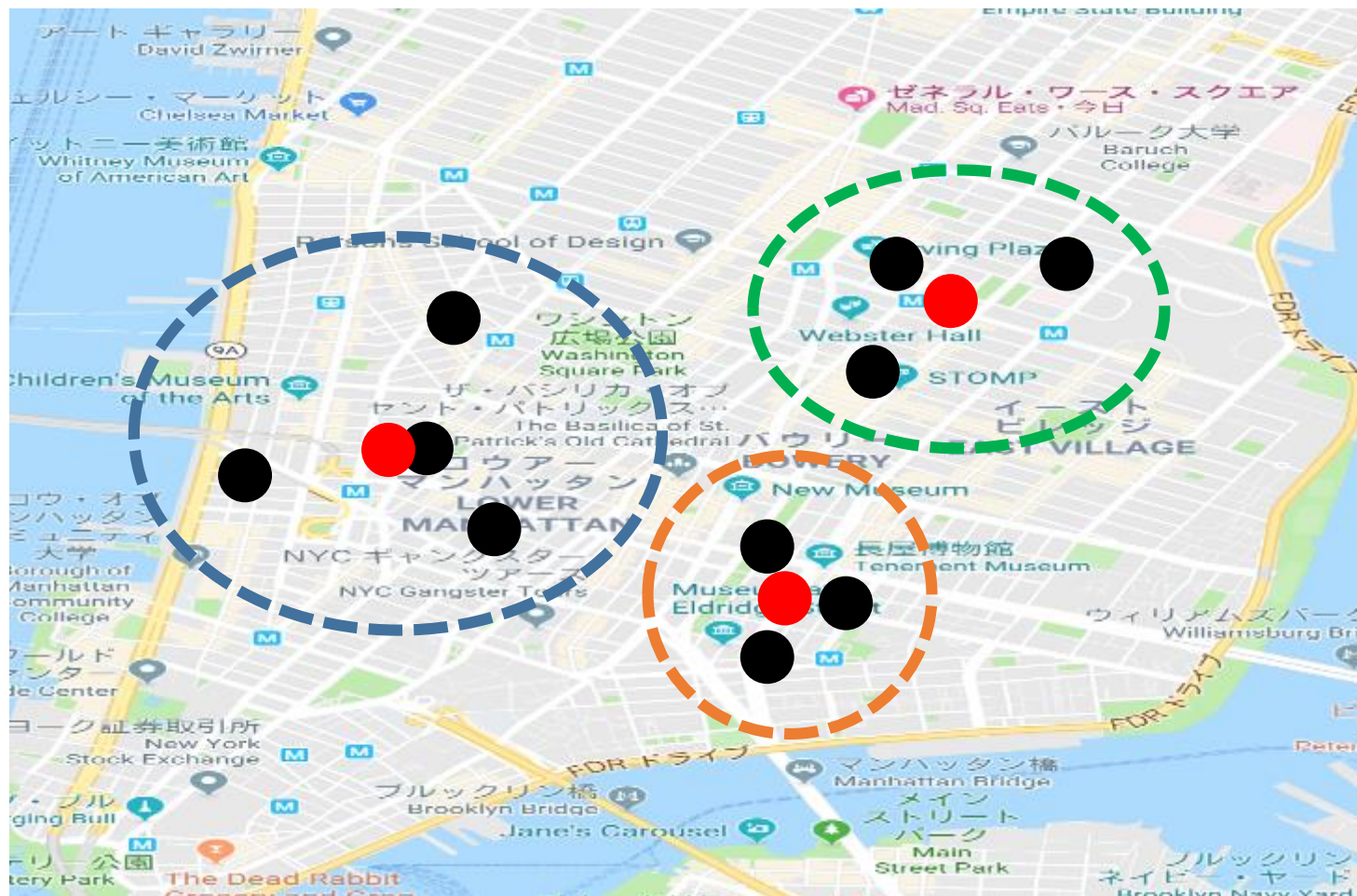
いったんクラスター分類を外し、





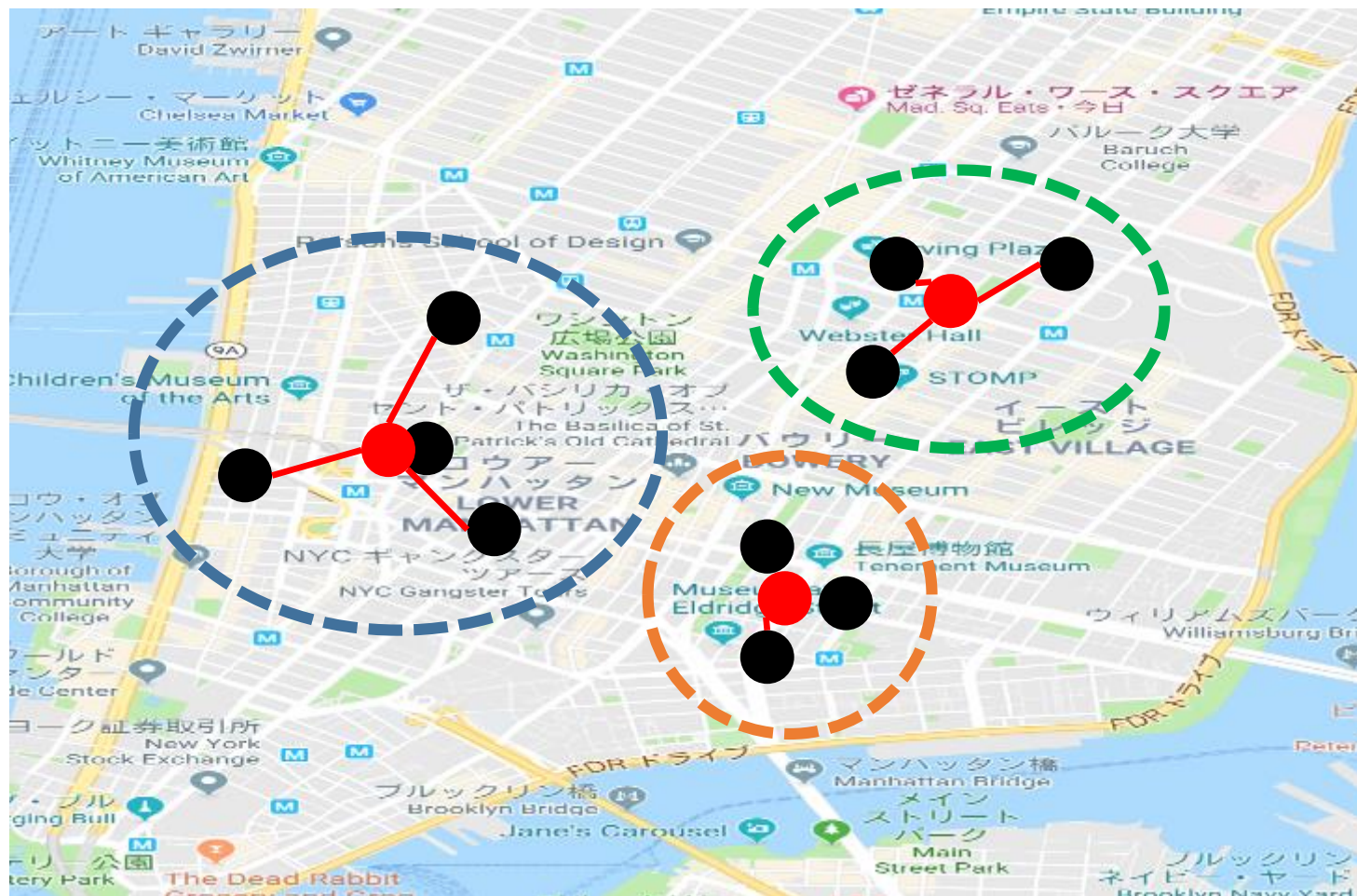
# K-mean法

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

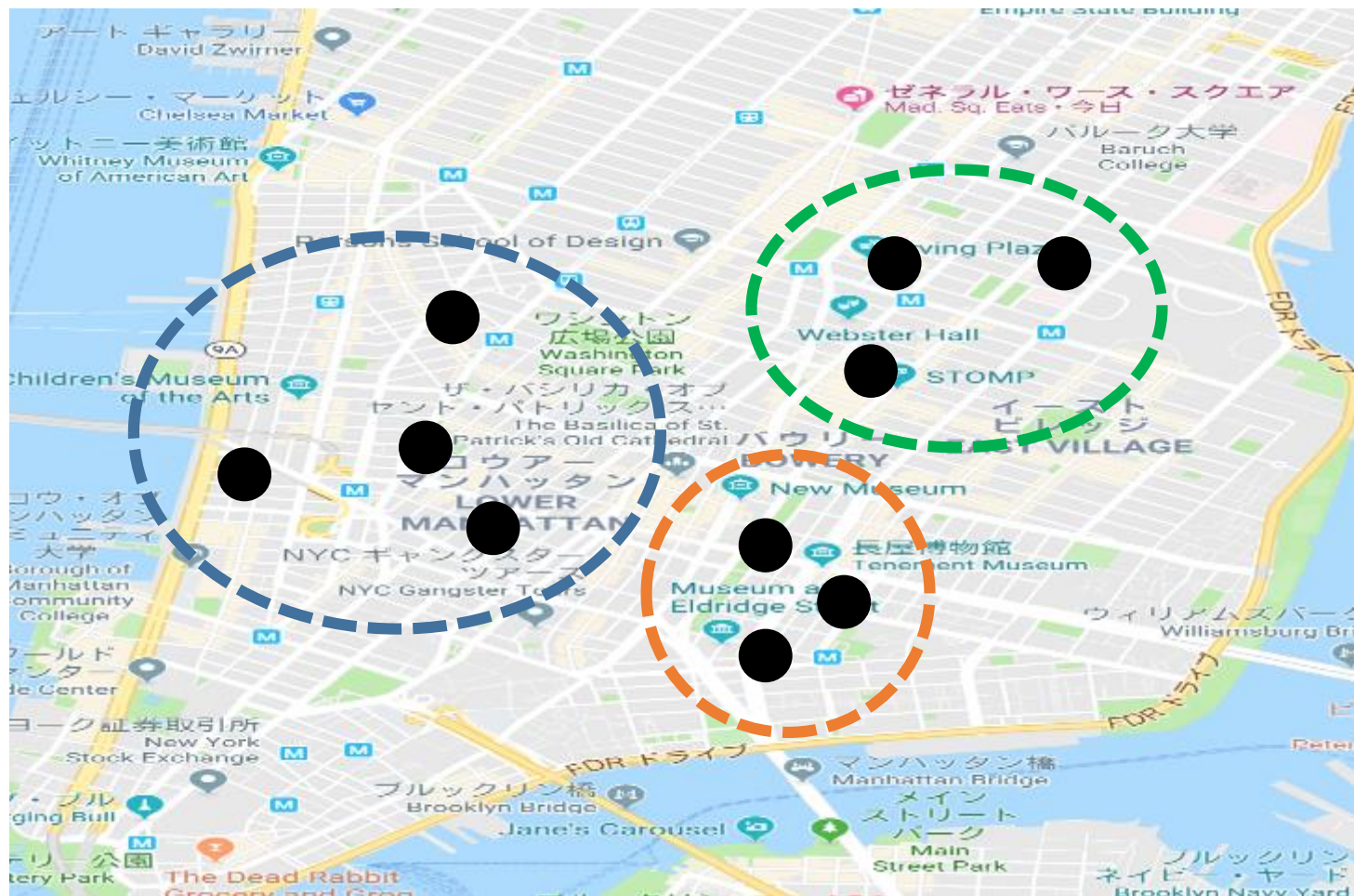
各データと最も近いシードを紐づける。





# 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。



# 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



# 主成分分析

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

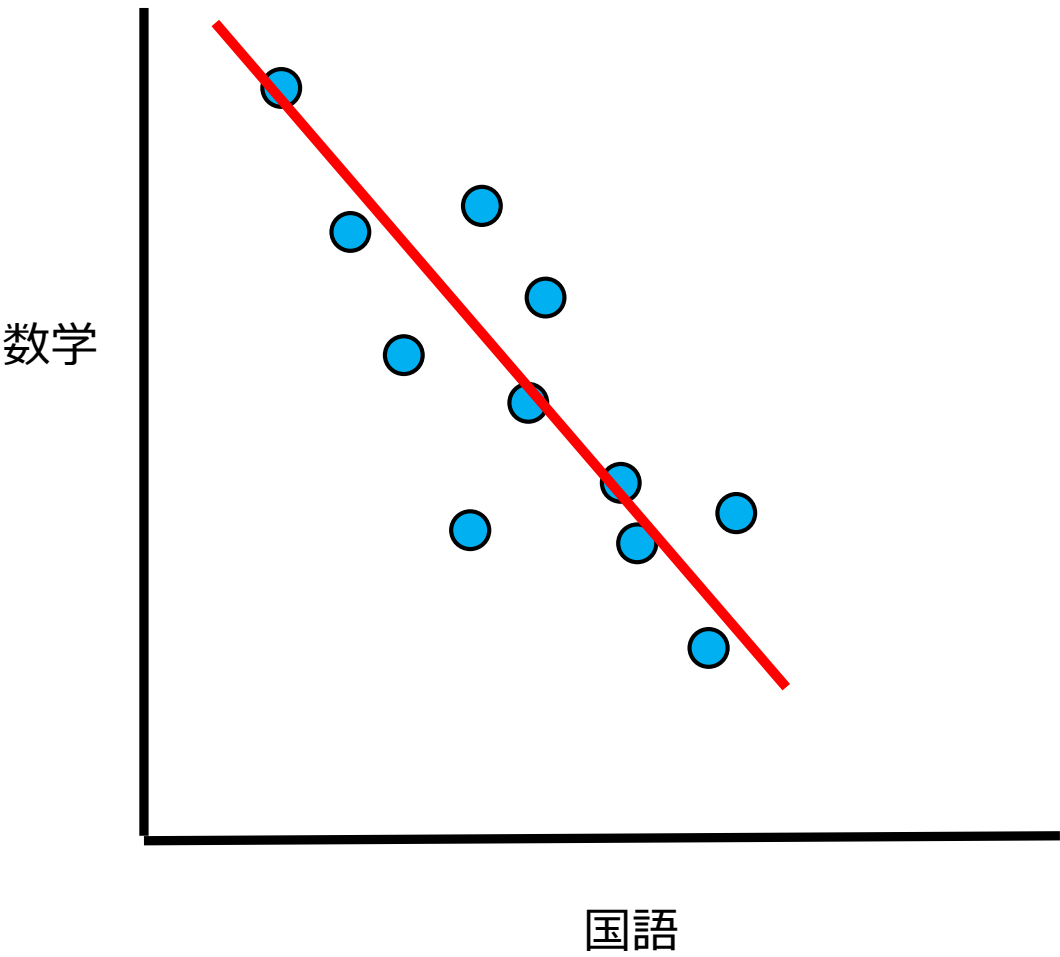
# 2Dデータの可視化



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

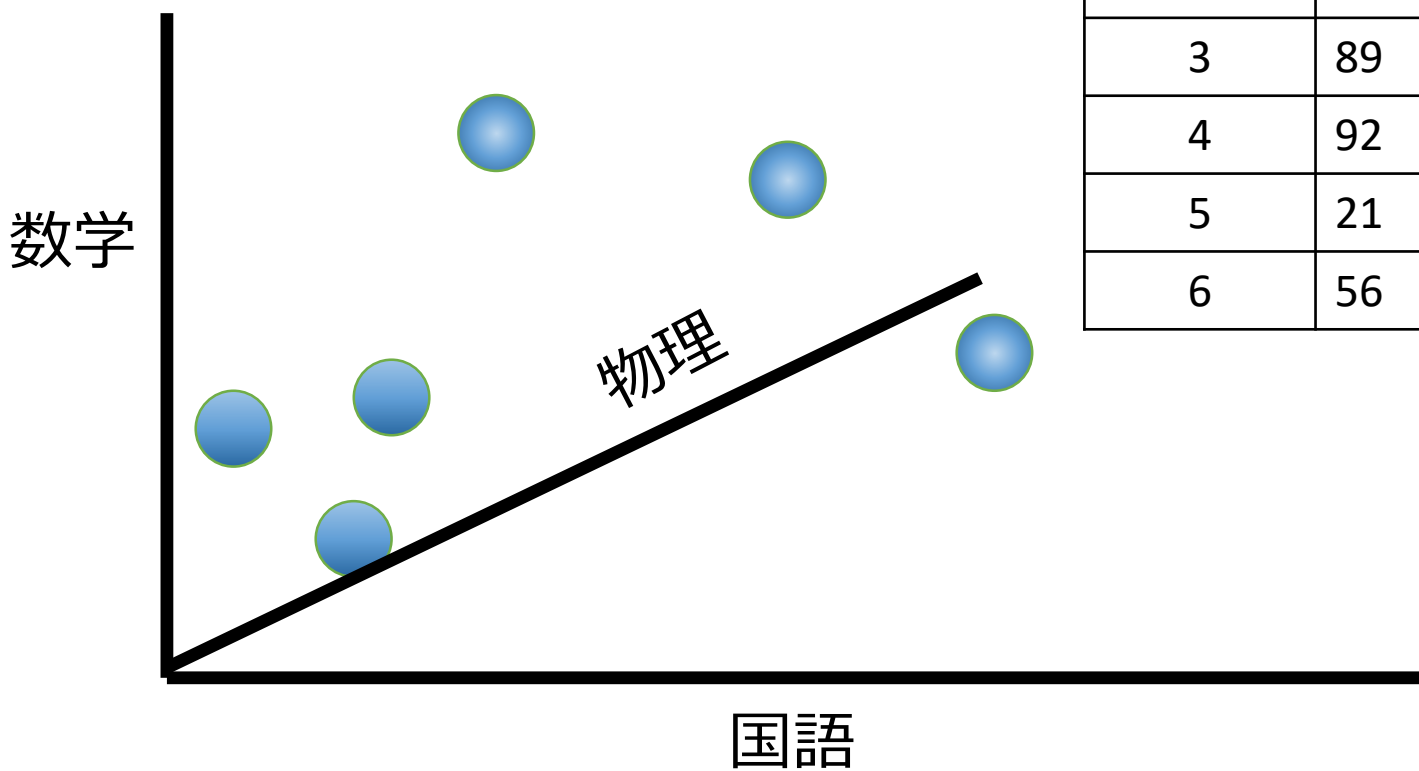


学生ID	数学	国語
1	23	89
2	45	52
3	89	65
4	92	34
5	21	84
6	56	76

# 3Dの可視化する

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

# 3Dの可視化する

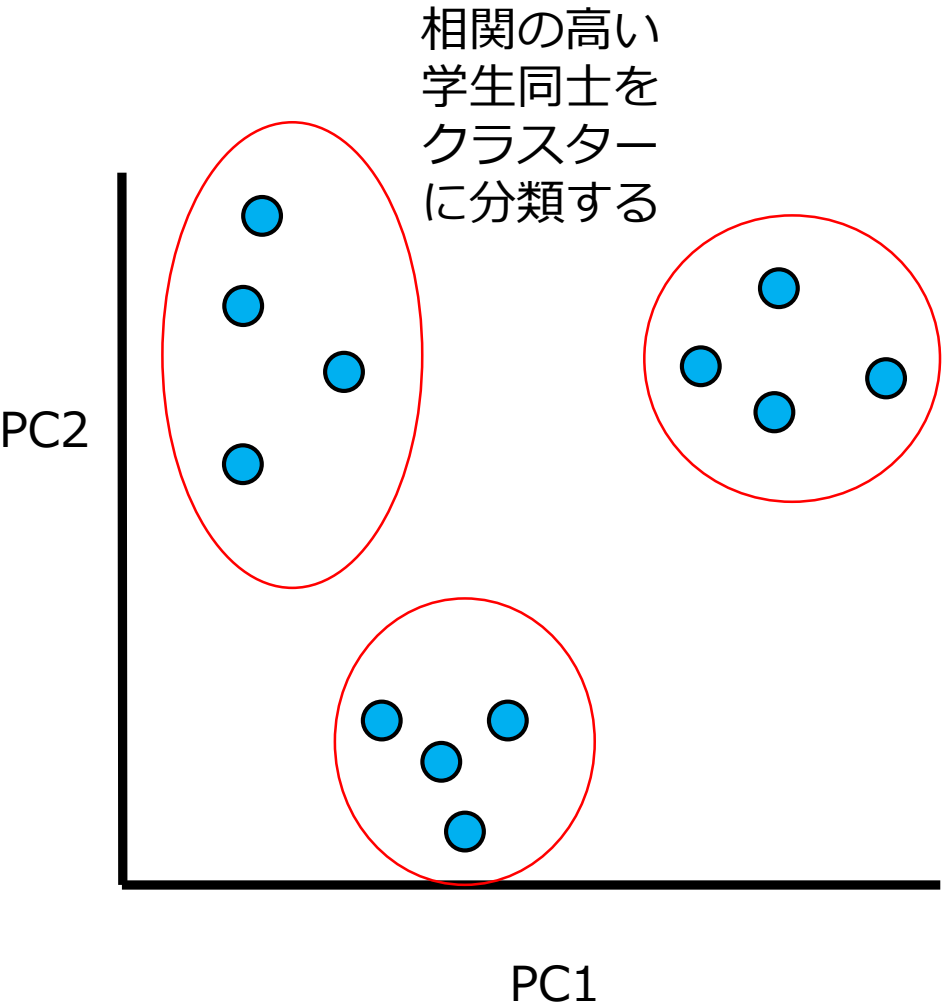


学生ID	数学	国語	物理
1	23	89	34
2	45	52	32
3	89	65	87
4	92	34	95
5	21	84	21
6	56	76	34

# 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

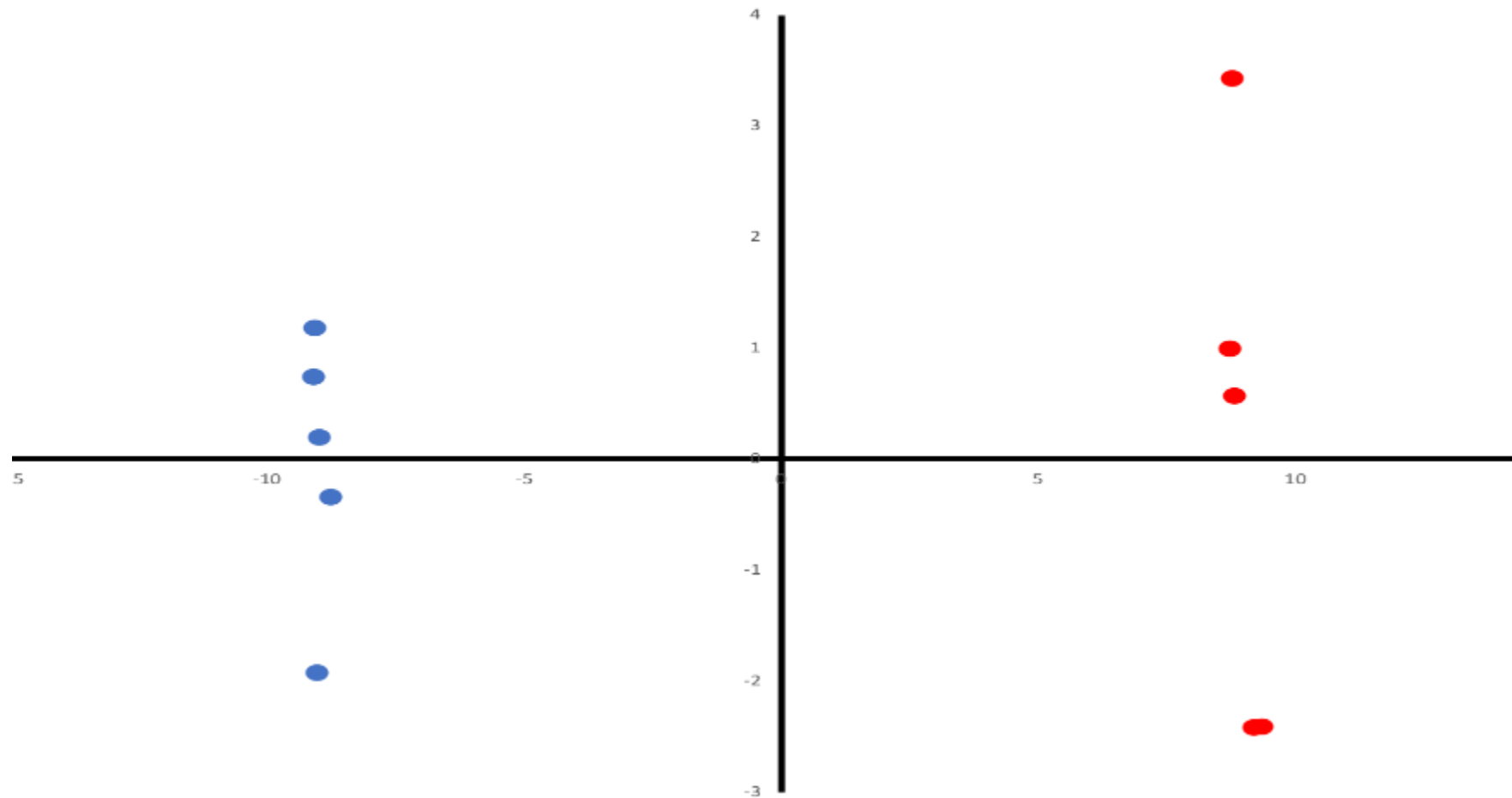
# 主成分による可視化



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

## 主成分分析



# 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



# 機械学習分類

手法名	目的	活用例	メリット
決定木	識別	顧客情報から購買/非購買を予測する	識別要因を解釈しやすい
ランダムフォレスト	識別	購買データ分析	決定木より精度が高い
SVM	識別	成分からワインの品種予測	少ないデータ数でも精度が高い
ナイーブベイズ	識別	迷惑メールのフィルタリング	テキストデータなどを扱いやすい
ロジスティック回帰	識別	医療診断 (陽性／陰性)	各要因が結果にどの程度影響を与えているかがわかる
ニューラルネットワーク	識別	画像認識	複雑な識別も可能 (画像、音声、テキストなど)
重回帰	回帰	家賃の予測	予測モデルとして使える
K-means法	分類	購買行動の傾向から 全顧客のグループ化を行う	データ数が多い時にいくつかのグループにまとめることができる
主成分分析	分類	アンケート項目を集約する	項目数が多い時に項目を集約 (次元削減)できる