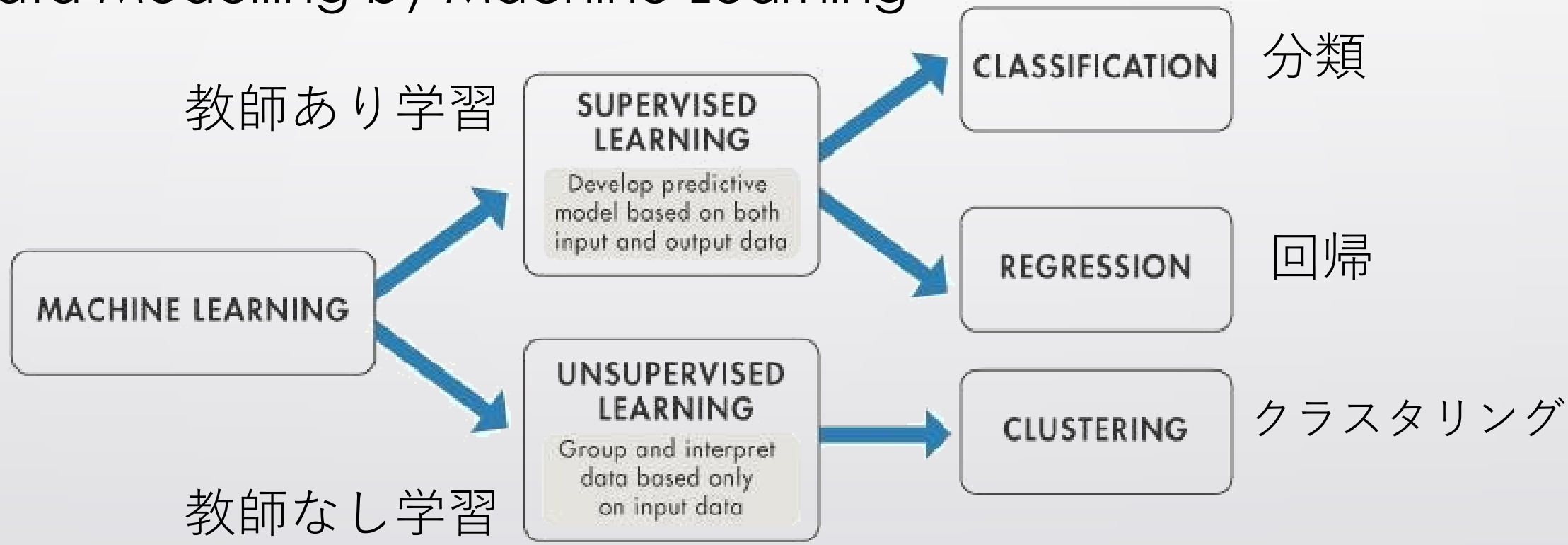
データマイニング

1: データマイニングの全体像
2: データセキュリティとデータサイエンスの倫理

土居 裕和 Hirokazu Doi

長岡技術科学大学 Nagaoka University of Technology

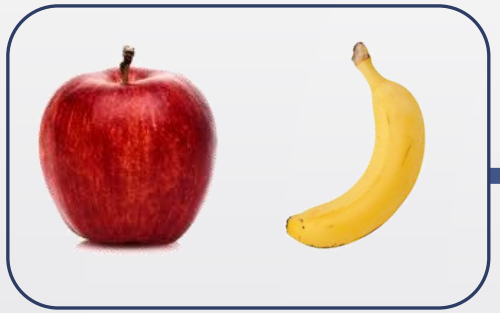
機械学習によるモデル化
Data Modelling by Machine Learning



Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)

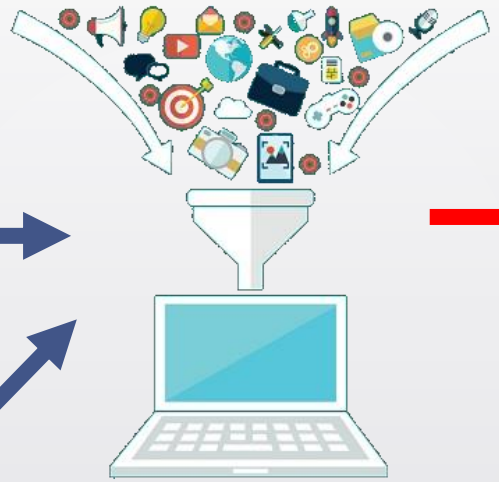
教師あり学習 Supervised Learning

特徴量 Feature



クラス Class

Apple Banana



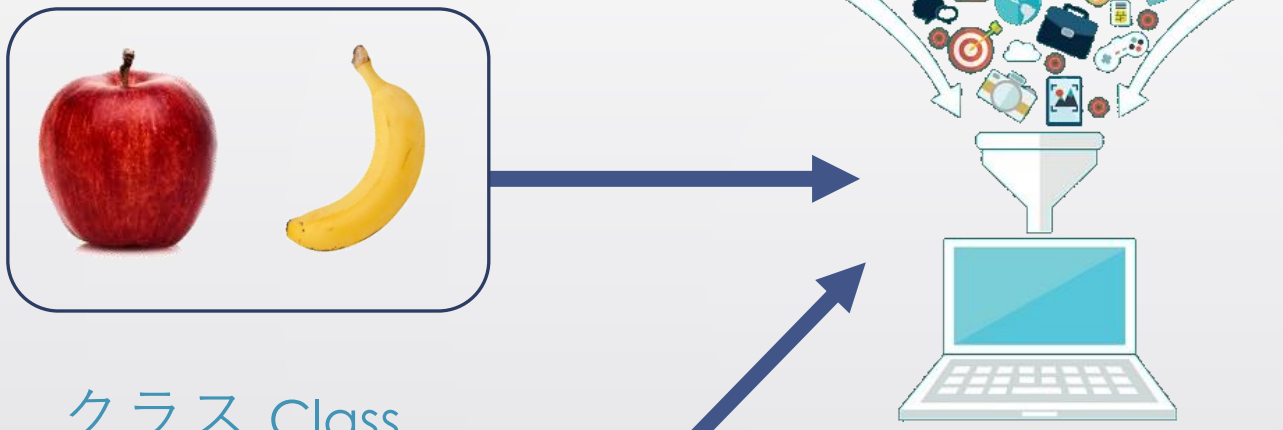
未知データに対するラベル
Label for the New Data

Apple

ラベルがついていない未知データ
Unknown Data without a Label

モデルの訓練（トレーニング）Model Training

特徴量 Feature



クラス Class

Apple Banana

予測 Prediction

Apple

ラベルがついていない未知データ
Unknown Data without a Label

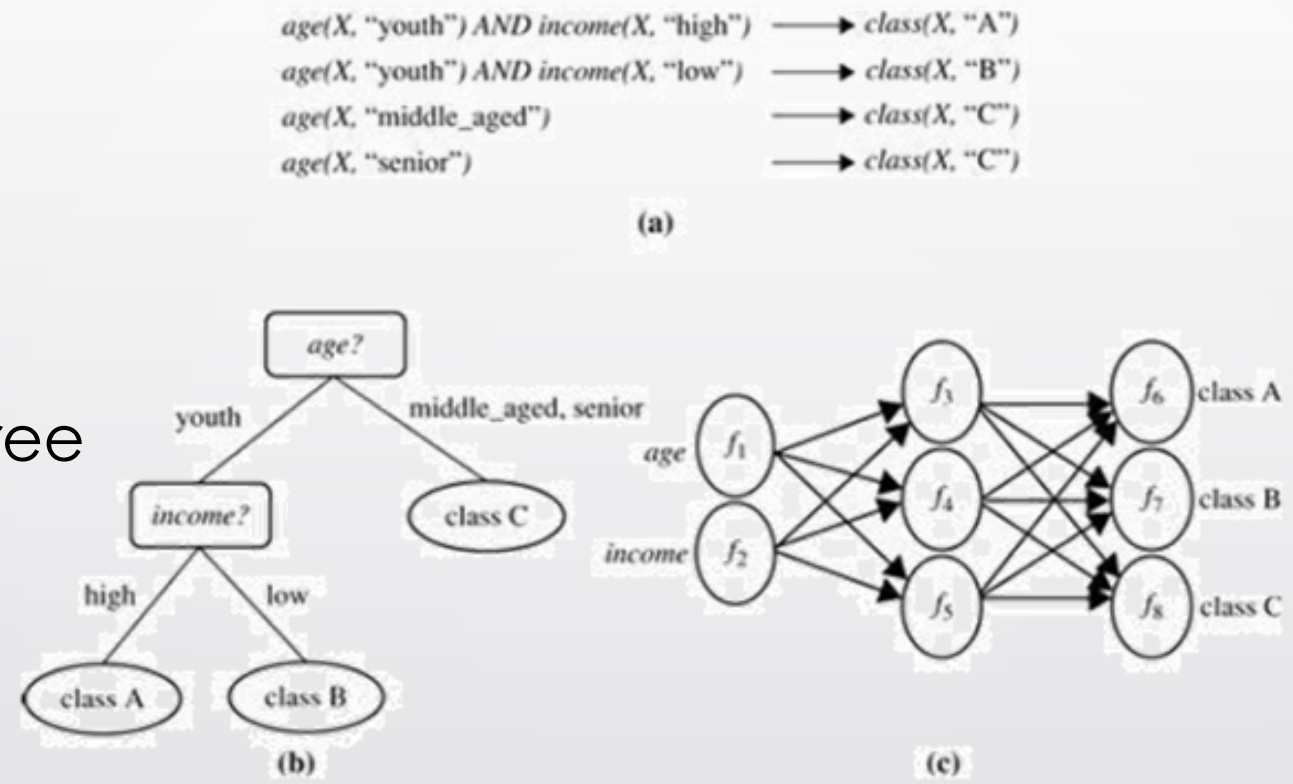
量を予測する回帰に対して、分類ではカテゴリー/クラスを予測する

Classification algorithms predict category/class while regression predicts quantity.

※量とクラスの予測の両方に対応したアルゴリズムも多い

There are many algorithms that can make predictions about both quantity and category.

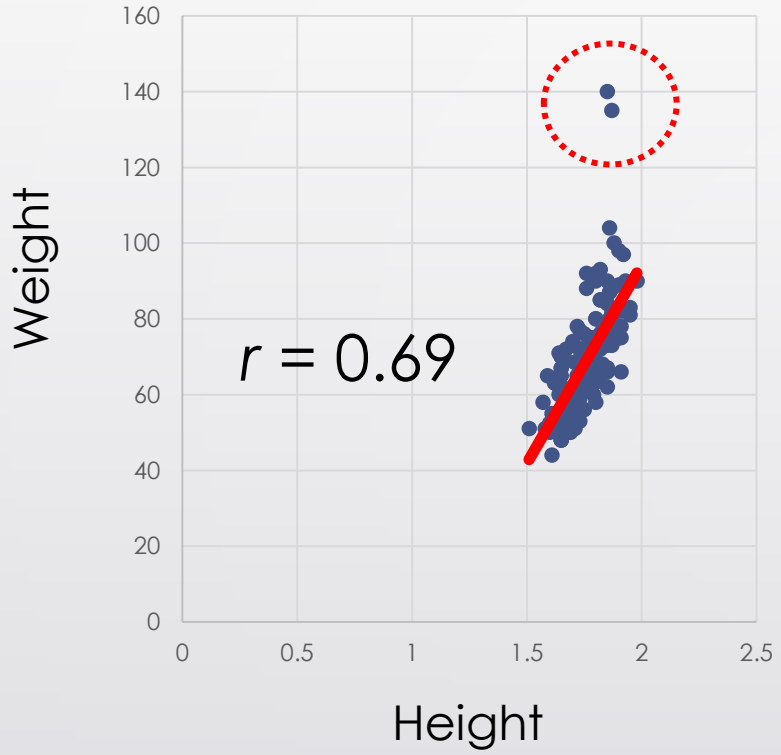
決定木
Decision Tree

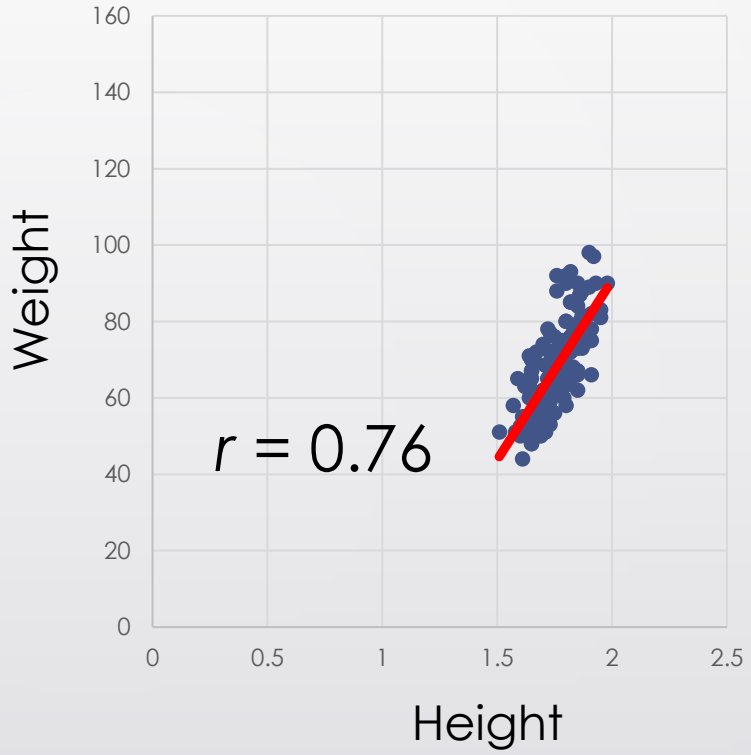


ニューラル
ネットワーク
Neural Network

FIGURE 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

外れ値 Outlier





結果に影響を与える可能性がある外れ値の存在をチェックする必要がある

Data set should be checked for the existence of outliers that can influence the results

外れ値分析による不正検知
Fraud Detection by Outlier Analysis

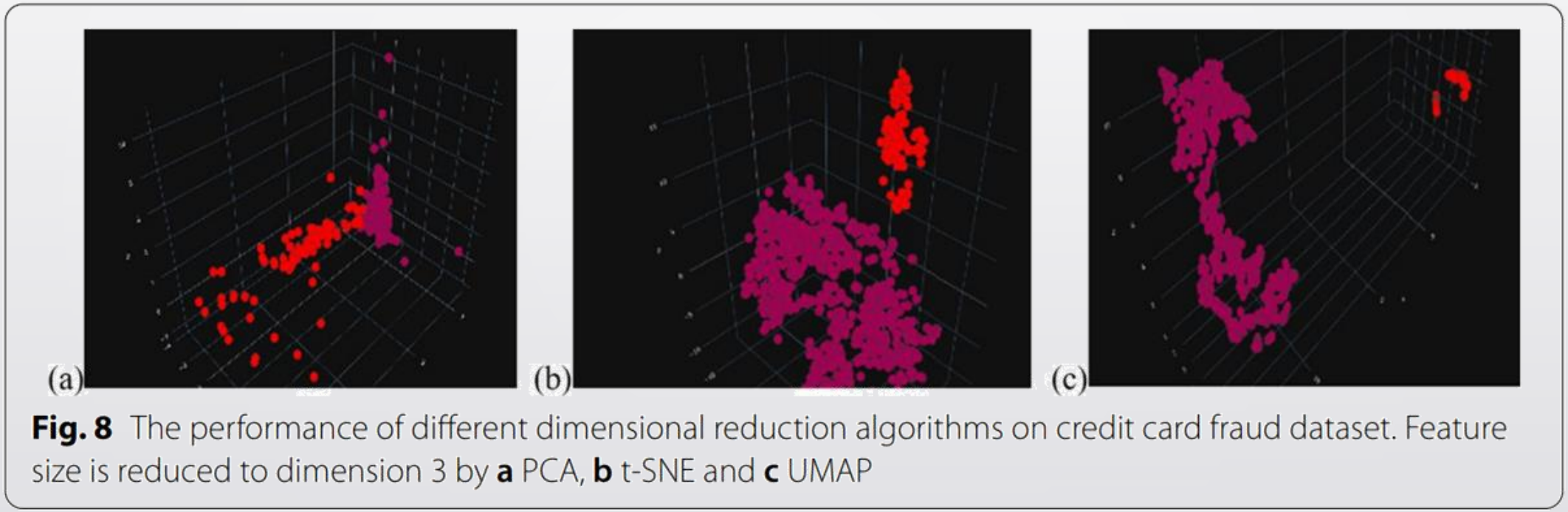


Fig. 8 The performance of different dimensional reduction algorithms on credit card fraud dataset. Feature size is reduced to dimension 3 by **a** PCA, **b** t-SNE and **c** UMAP

Benchaji et al. Journal of Big Data (2021) 8:151

データマイニングの流れ Steps in Data Mining

1. 目標設定 Goal Setting
2. データ収集 Data collection
3. 前処理 Preprocessing
4. 特徴量選択 Feature Selection (必要ないケースもある; can be skipped in some cases)
5. データ分析 Data Analysis・モデリング Modeling
6. 性能評価 Performance Evaluation
7. （ディプロイメント Deployment）

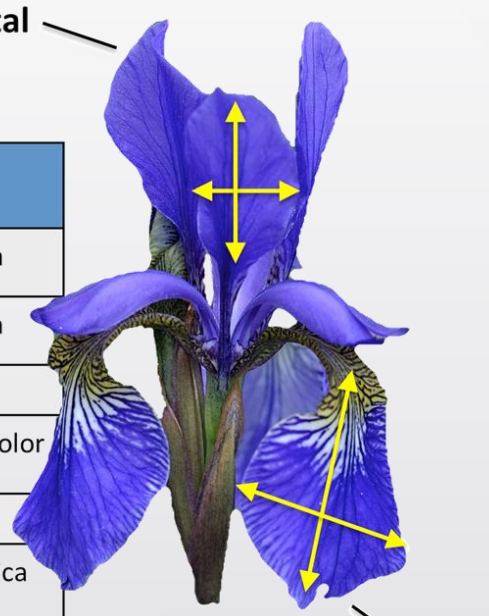
目標設定 Goal Setting

1. 調査の目的を設定する Verify the purpose of the survey
2. どのようなデータを集めるかを決める Decide what data to collect for our purposes
3. 目的を達成するには、どのようなデータ分析を行うのが適切かを、前もって考えておく Consider what type of data analysis is appropriate to achieve the goal

アヤメからの特徴量抽出 Feature Extraction from Iris

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



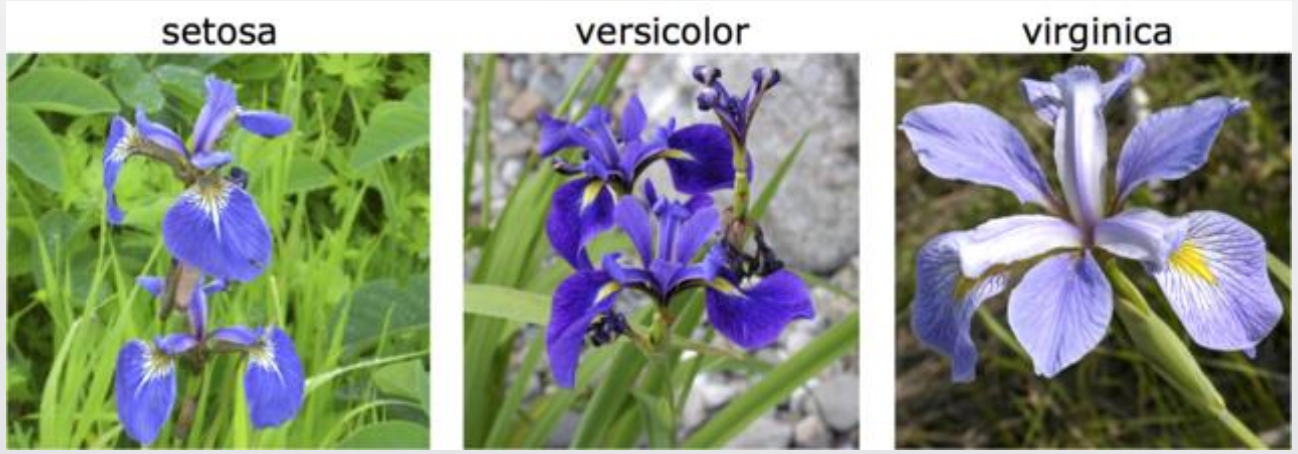
Features
(attributes, measurements, dimensions)

Class labels
(targets)

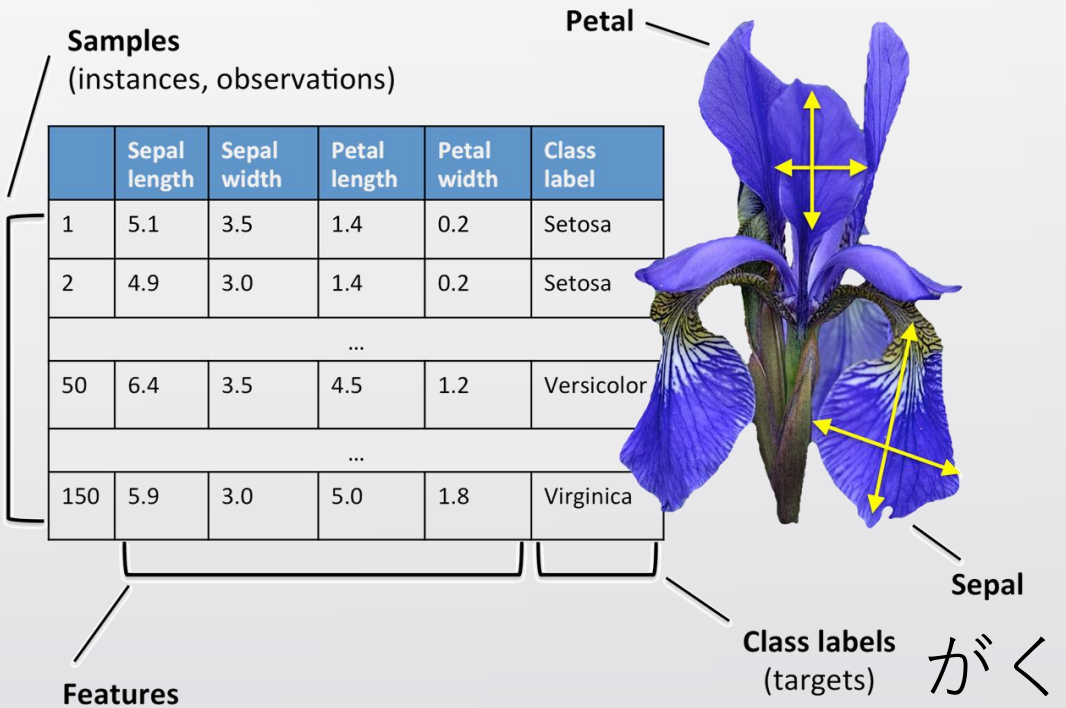
https://rindalog.blogspot.com/2017/09/pcairis.html

アヤメデータセット Iris Dataset

ヒオウギアヤメ　　ブルーフラッグ　アイリス・バージニカ



花弁



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

がく

データの読み込み Loading Data

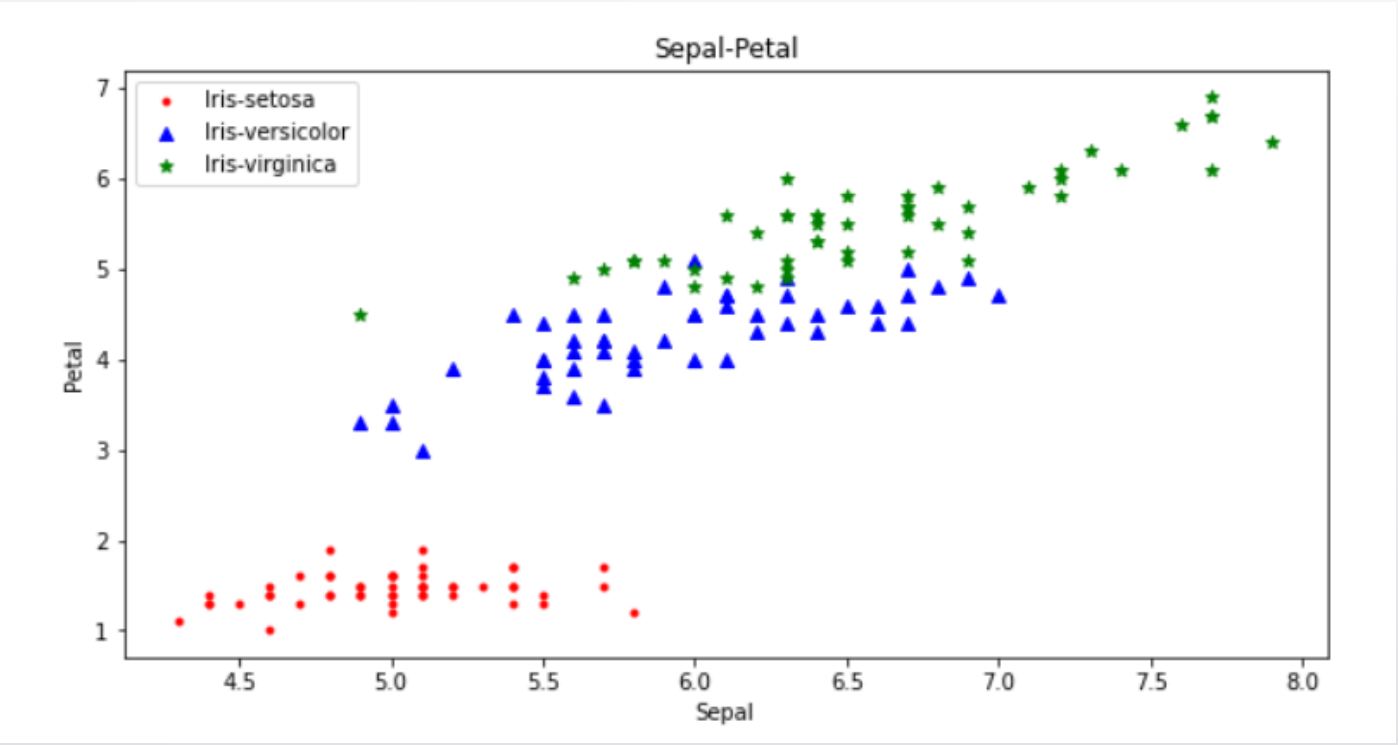
```
Iris plants dataset
--------------------

**Data Set Characteristics:**

    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica
```

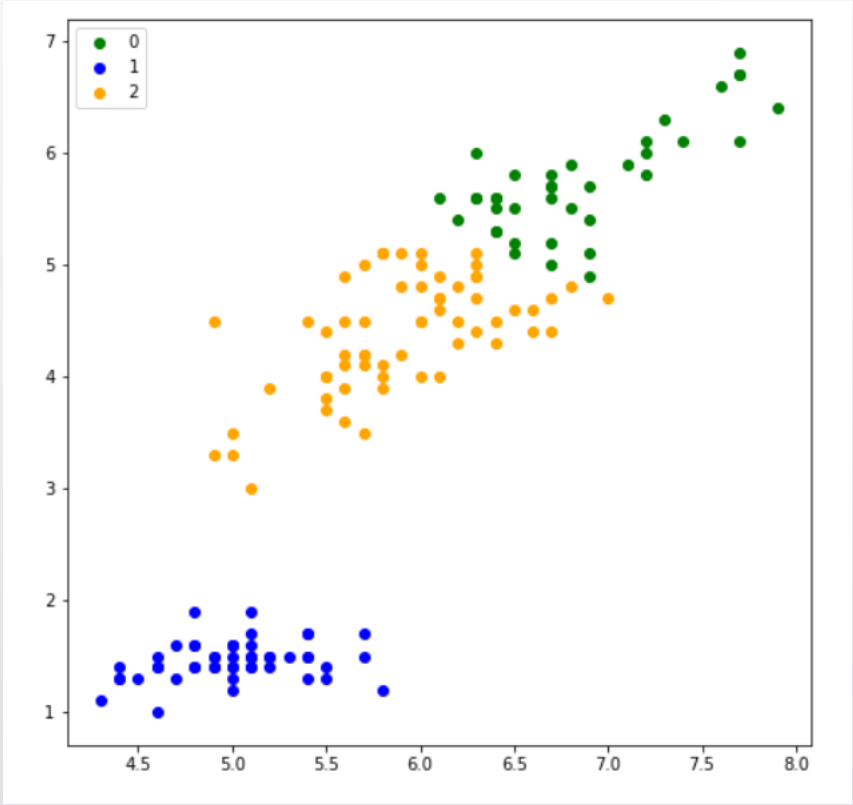
記述統計 Descriptive Statistics

```
In [15]: df.describe()
```

Out[15]:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000


Sepal-Petal
Iris-setosa
Iris-versicolor
Iris-virginica
Petal
Sepal
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0


0
1
2
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2
...	...	...	...	...
145	6.7	3.0	5.2	2.3
146	6.3	2.5	5.0	1.9
147	6.5	3.0	5.2	2.0
148	6.2	3.4	5.4	2.3
149	5.9	3.0	5.1	1.8

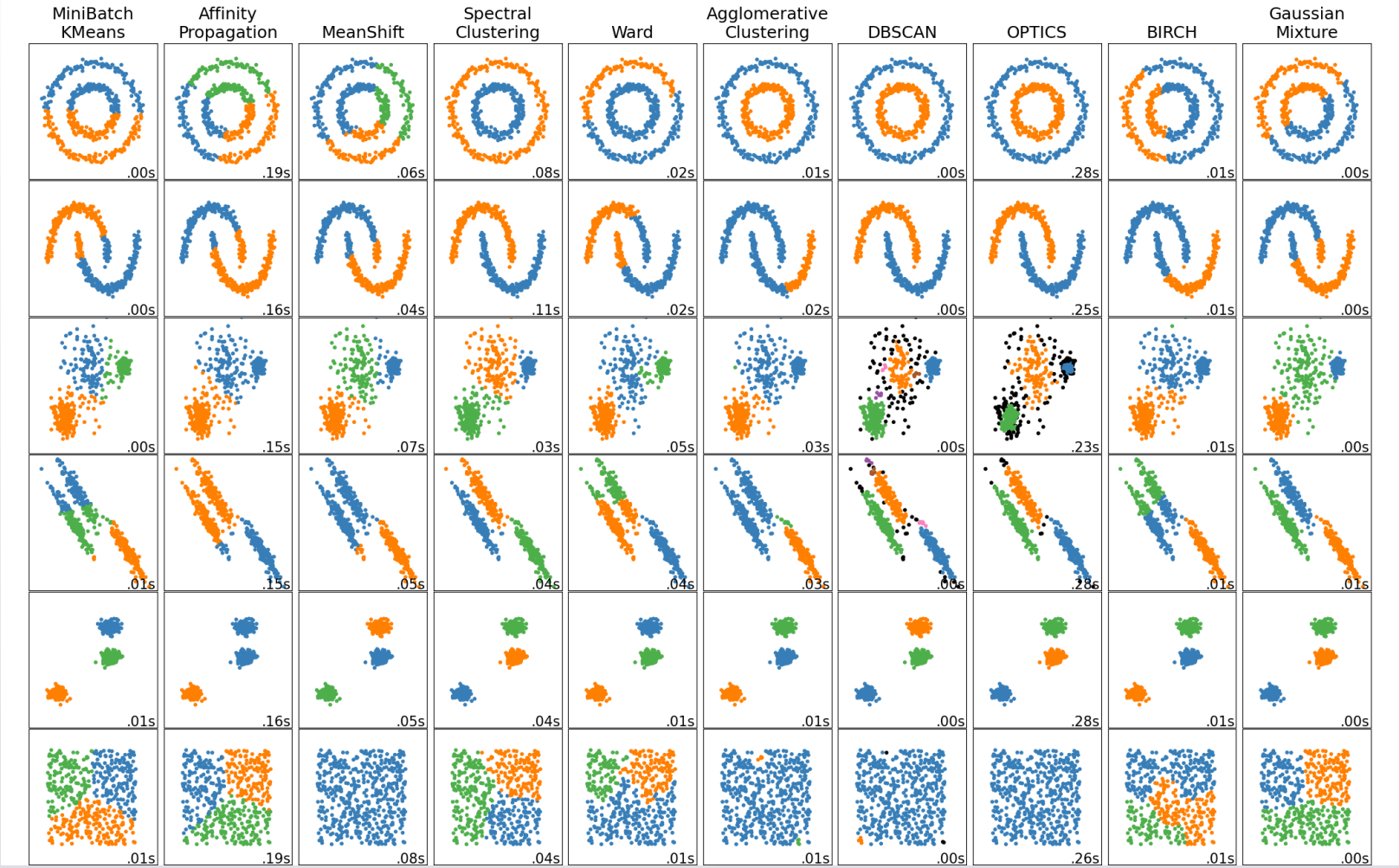
クラスタリング

Final result of clustering depends on

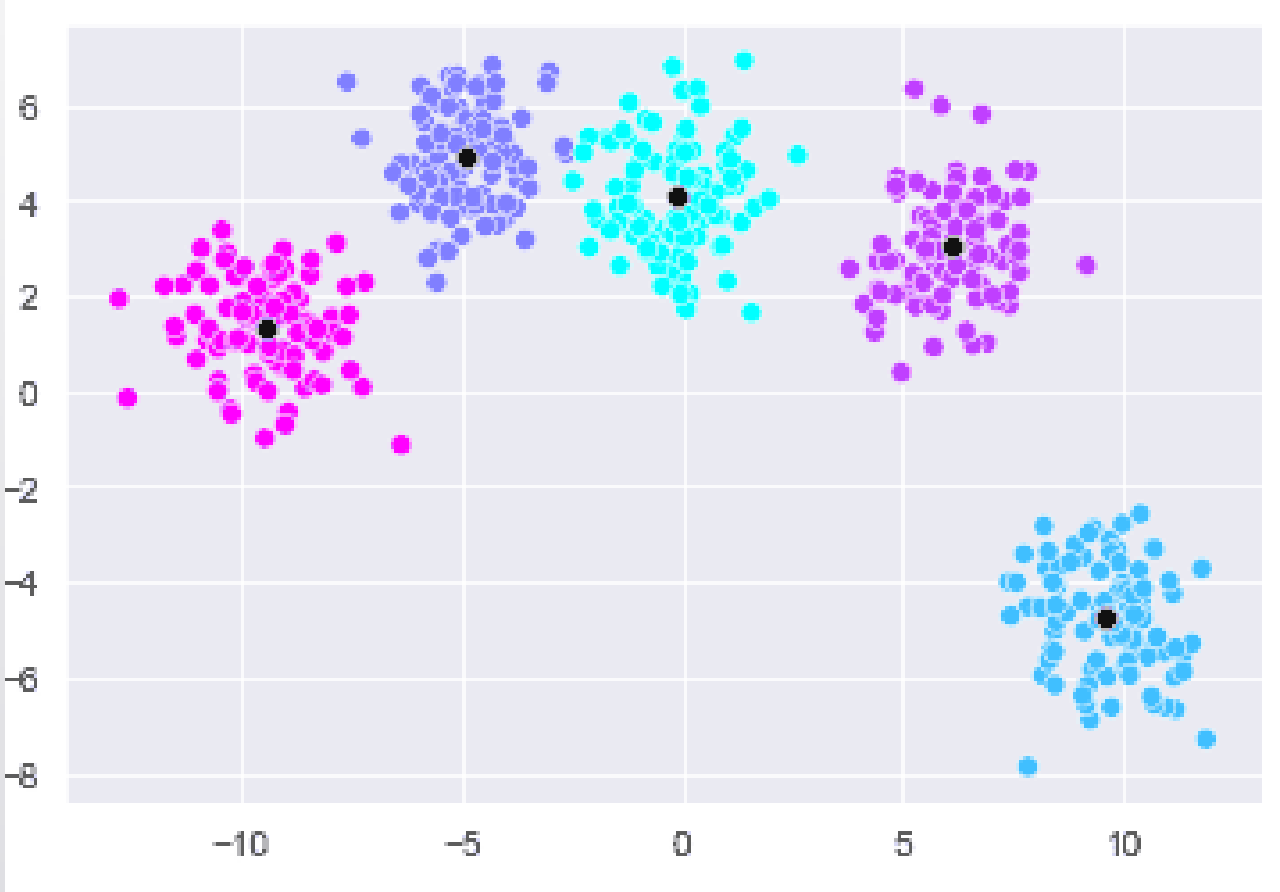
Type of algorithm
アルゴリズムの種類

Parameter Setting
パラメータ設定

https://scikit-learn.org/stable/modules/clustering.html



K-Means Clustering



クラスターの数を指定しなくてはいけない

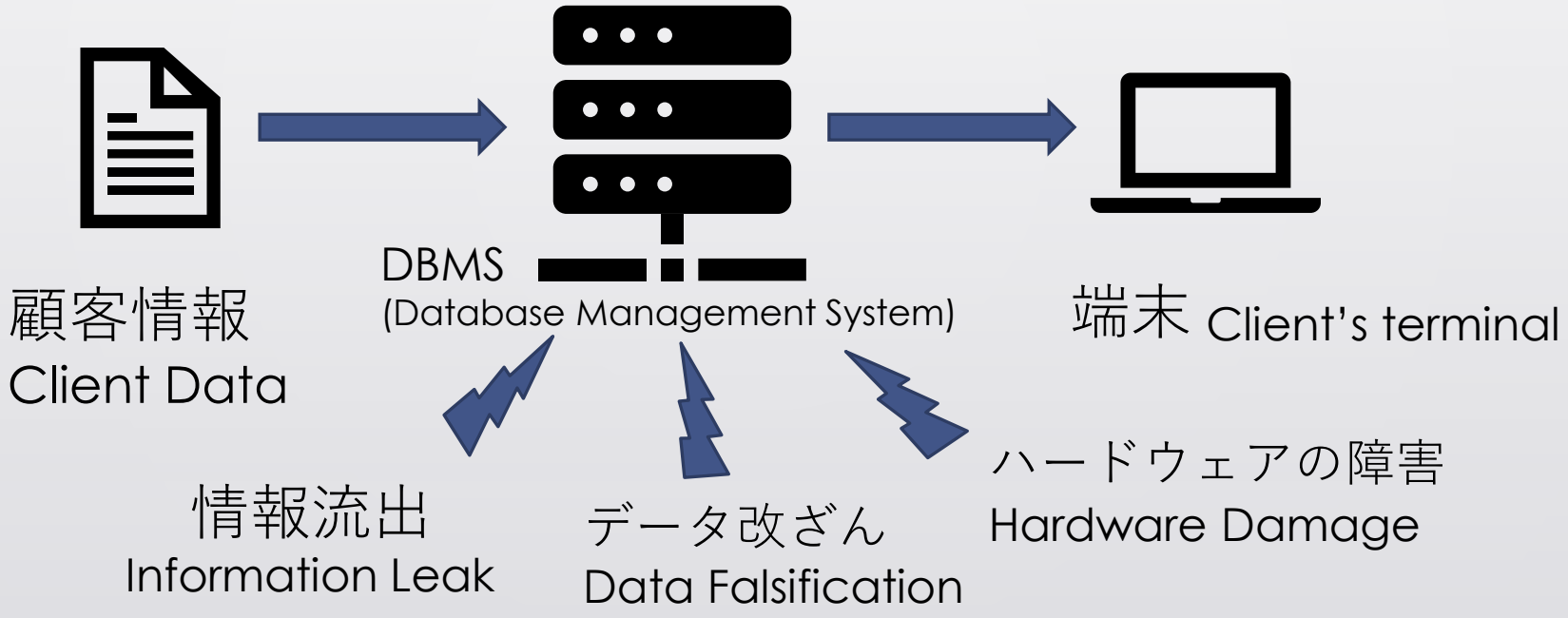
You have to specify the number of clusters, *k*.

データセキュリティ
Data Security

情報資産 Information Asset

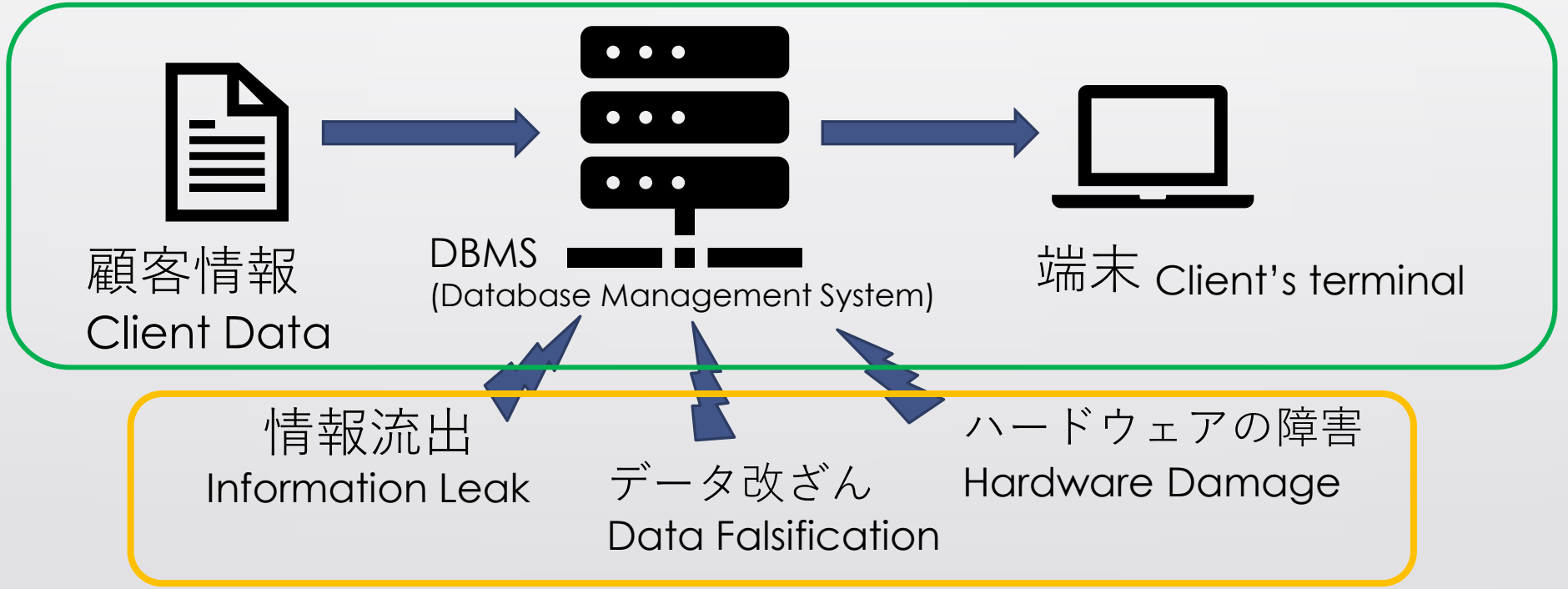
個人や企業にとって、価値のある情報や情報システム

Information or the integrity of information system valuable for an individual or corporation.



リスクと脆弱性 Risk and Vulnerability

脆弱性：情報システムの弱点 Weak point of an information system



リスク：情報システムが損なわれる可能性 Possibility that an information system is damaged

機密性 Confidentiality

許可された人のみが情報にアクセスできる

Only persons granted permission can access information

Confidentiality

アクセス制限 Access Control

サーバールームの施錠 Locked Server Room

暗号化 Encryption

https://www.simplilearn.com/what-is-information-security-article

完全性 Integrity

改ざんなどされることなく、情報が完全に保たれている
Information is being kept intact without falsification etc

Integrity

アクセスログ・操作 Logging Access and Operation

直感的なユーザー・インタフェース
Intuitive User Interface

https://www.simplilearn.com/what-is-information-security-article

可用性 Availability

許可された人が、必要な時に、いつでも情報にアクセスできる

Any persons grated permission can access information whenever necessary

Availability

複数箇所におけるデータ保存

Data storage in multiple sites

定期的なバックアップ

Periodical BackUp

https://www.simplilearn.com/what-is-information-security-article

AUTO & TRUCK MANUFACTURERS 2011年4月28日 / 6:00 午前 / 12年前更新

UPDATE 1-Sony may face global legal scrutiny over breach

Tom Hals, Leigh Jones

5 分で読む

* U.S. lawyers focused on data breaches eye legal action

* British government watchdog launches investigation

* U.S. state attorneys general also discussing incident (Recasts with comments from U.S. legislators)

WILMINGTON, Delaware/NEW YORK, April 27 (Reuters) - Sony Corp 6758.T could face legal action across the globe after it belatedly revealed one of the biggest online data breaches ever.

In the United States, several members of Congress seized on the breach, in which hackers stole names, addresses and possibly credit card details from users of Sony's PlayStation Network, to push for tougher laws protecting personal information.

AP Photo/Shizuo Kambayashi

個人情報 Personal Information

Information that identifies particular individual

Data extracted from individual's body

Official number assigned to each individual

「個人情報保護法」

Personal Information Protection Law

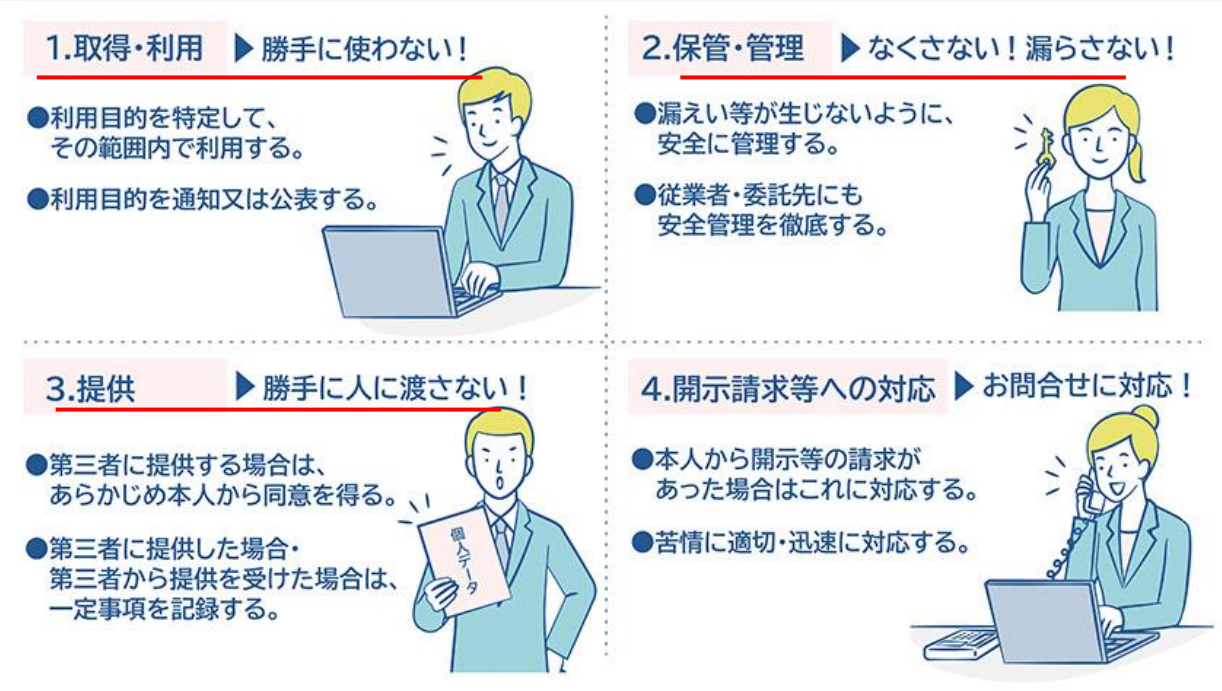
生存する個人に関する情報で、氏名、生年月日、住所、顔写真などにより特定の個人を識別できる情報

Information concerning living individual…
Information by which a particular individual is identifiable

政府広報オンライン

データはあらかじめ決まった目的でしか利用しない
Use information only for pre-defined purposes

流出を防止するため、データを安全に管理する
Should be managed securely to avoid data leakage



第三者とデータを共有する場合は、関係者の同意を得る
Consent must be obtained from relevant individuals when their data is shared with the third party.

請求があった場合には、関連情報を開示する
Should disclose any relevant information whenever requested

Samarati & Sweeney, 1998

マサチューセッツ州のGIC (Group Insurance Commission)が、氏名を削除

したうえで、医療保険に関連する情報を民間企業に販売

GIC in Massachusetts sold to private companies the data relevant to medical insurance after "anonymyzation", i.e. deleting individuals' names

- Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Samarati & Sweeney, 1998

マサチューセッツ州ケンブリッジの選挙人名簿は20ドルで購入可能だった

Voters' list in Cambridge, Massachusetts could be purchased for 20 dollars

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

Samarati & Sweeney, 1998

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
		asian	09/27/64	female	02139	divorced	hypertension
		asian	09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	.....
.....	.....	.....	.....	.....	.....	.....	
.....	.....	.....	.....	.....	.....	.....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	.....
.....	.....	.....	.....	.....	.....	.....	

個人情報の例: 表1
Example of Personal Information: Table 1

Identification Number	User ID	Name	Gender	Age	Income
識別ナンバー	ユーザーID	氏名	性別	年齢	年収
339829Q	sanapon	真田昌幸	男	26	411万円
905473R	oggi1985	荻野吟子	女	33	536万円
099878L	murachan	紫式部	女	39	681万円
013214H	shozan.s	佐久間象山	男	23	309万円

個人情報の例: 表2
Example of Personal Information: Table 2

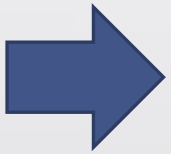
User ID	Product	Price	Date	Store
ユーザーID	**購買物品**	**購買価格**	**購買日時**	**購買店舗**
murachan	人参	100円	2021/2/3 18:09	Cマート 代田2丁目店
oggi1985	バナナ	150円	2021/2/3 21:13	Cマート 小石川店
oggi1985	ダイエットコーク	160円	2021/2/4 21:15	Cマート 小石川店
murachan	粉ミルク	980円	2021/2/4 21:16	Cマート 代田2丁目店
murachan	紙おむつ	1700円	2021/2/4 21:16	Cマート 代田2丁目店

名前が削除されていても、表に含まれた情報がら、個人を特定できる場合がある

Even if name is deleted, sometimes, an individual can be identified from information contained in a table.

仮名化 Pseudonymization 匿名化　Anonymization

識別ナンバー	ユーザーID	氏名	性別	年齢	年収	住所
339829Q	sanapon	真田昌幸	男	26	411万円	・・・
905473R	oggi1985	荻野吟子	女	33	536万円	・・・
099878L	murachan	紫式部	女	39	681万円	・・・
013214H	shozan.s	佐久間象山	男	23	309万円	・・・

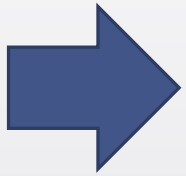


ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円

個人識別情報を削除する　Deleting person identifiable information

k-匿名化 *k*-Anonymization

ユーザーID	性別	年齢	年収
sanapon	男	26	411万円
oggi1985	女	33	536万円
murachan	女	39	681万円
shozan.s	男	23	309万円



性別	年齢	年収	
男	[20-29]	[300-499]	万円
女	[30-39]	[500-699]	万円
女	[30-39]	[500-699]	万円
男	[20-29]	[300-499]	万円

データの中に、同じ属性を持つ人が、少なくともk (>=2)人含まれる

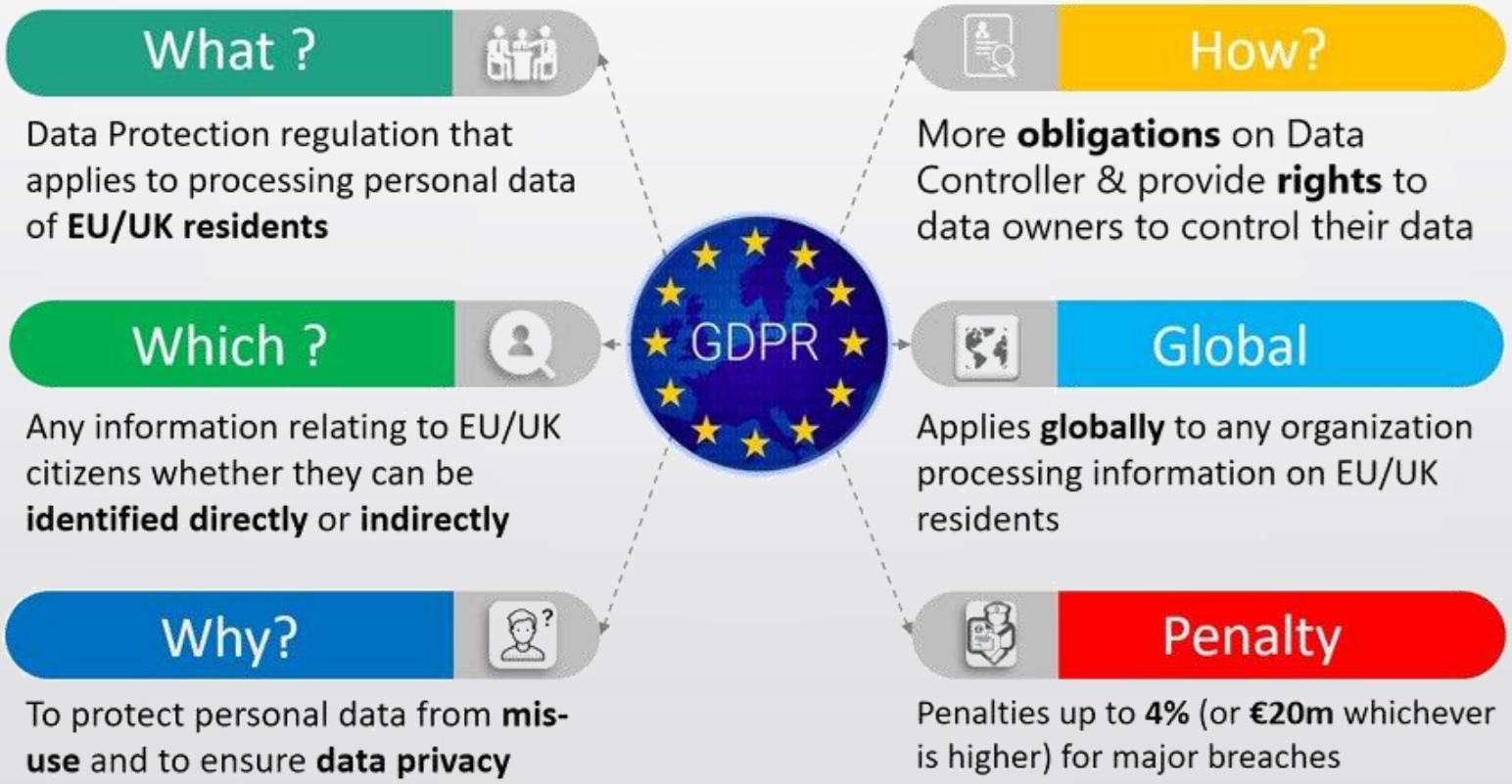
At least *k* (>= 2) persons with the same attributes are included in a table.

データサイエンスの倫理
Ethics in Data Science

一般データ保護規則 General Data Protection Rule (GDPR)

EU/UKに住む人々の個人情報が対象

個人情報を保護し、プライバシーを守る

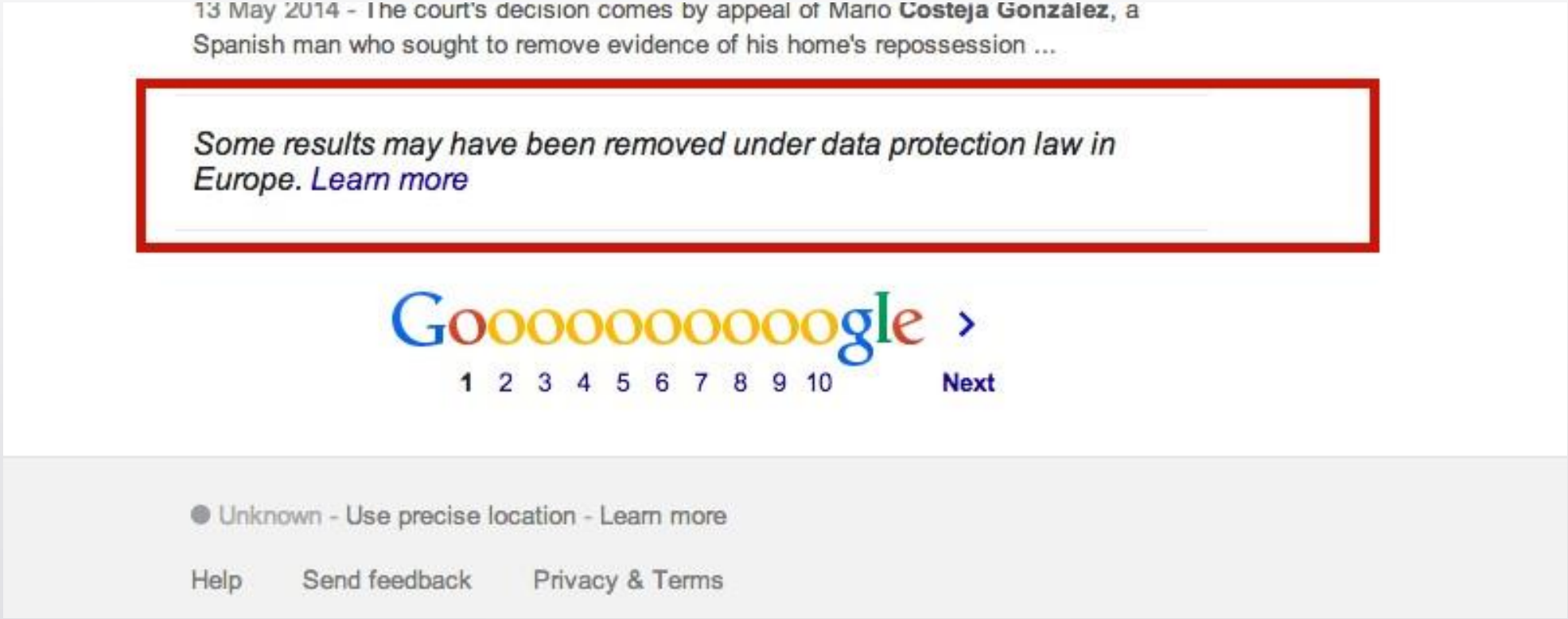


個人が、自分の個人情報をコントロールできるようにする

EU/UKに住む人々のデータを扱う、**世界中の機関**に適用される

Privacy Preservation in Federated Learning: Insights from the GDPR Perspective

忘れられる権利 The Right to be Forgotten



Retrieved on 2023/01/28 from
https://www.cima.ned.org/publication/right-to-be-forgotten-threat-press-freedom-digital-age/

透明性とトラスト Transparency and Trust

透明性 Transparency :

データマイニングシステムやAIが、結論を導くプロセスが、人々に説明可能になっていること

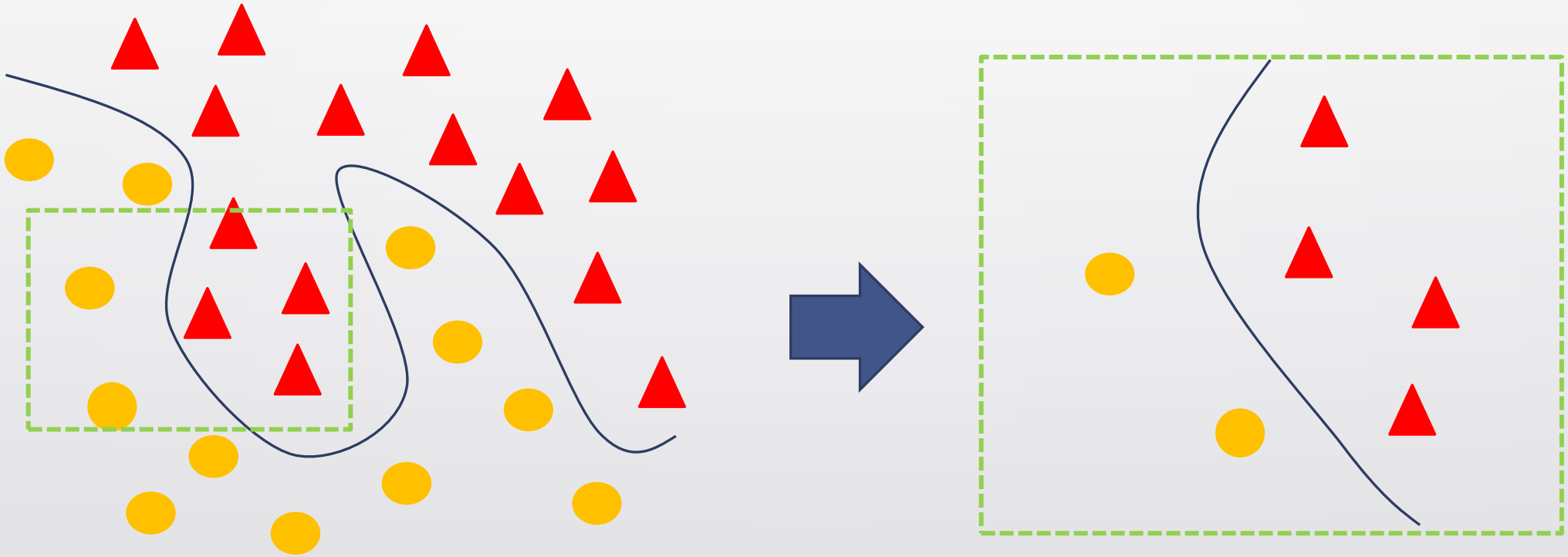
To make the process of reaching conclusions by data-mining system and AI understandable and accessible to ordinary people

トラスト Trust :

人々が、そのデータ処理プロセスを理解したうえで、データマイニングシステムやAIを信頼すること

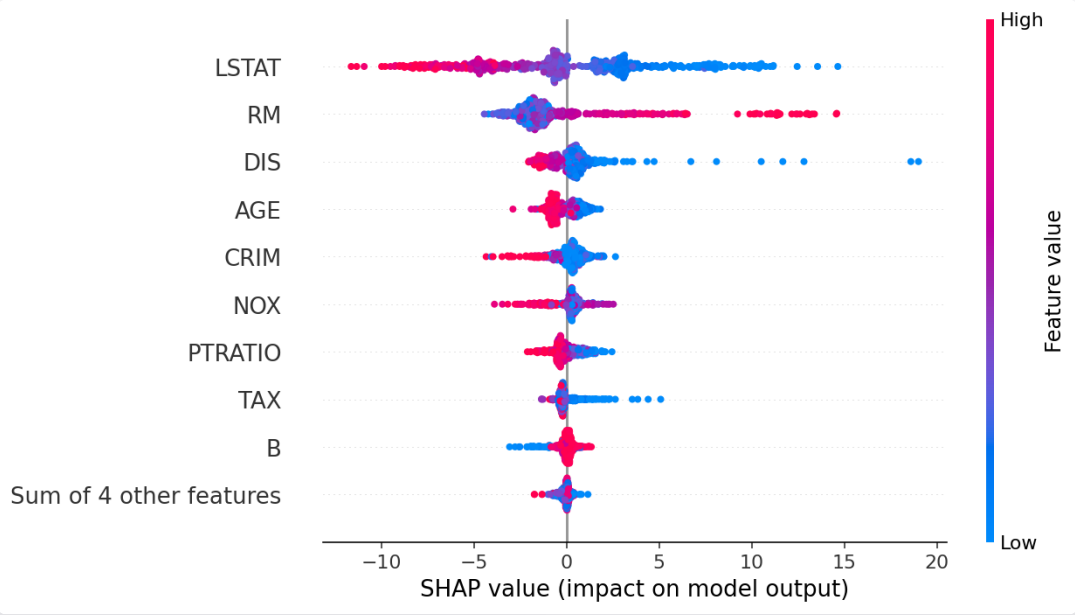
People trust data-mining system and AI after understanding their process of data processing

説明可能なAI Explainable AI



説明可能なAI Explainable AI

SHAP value



https://github.com/slundberg/shap

Grad CAM (Gradient-weighted Class Activation Mapping)

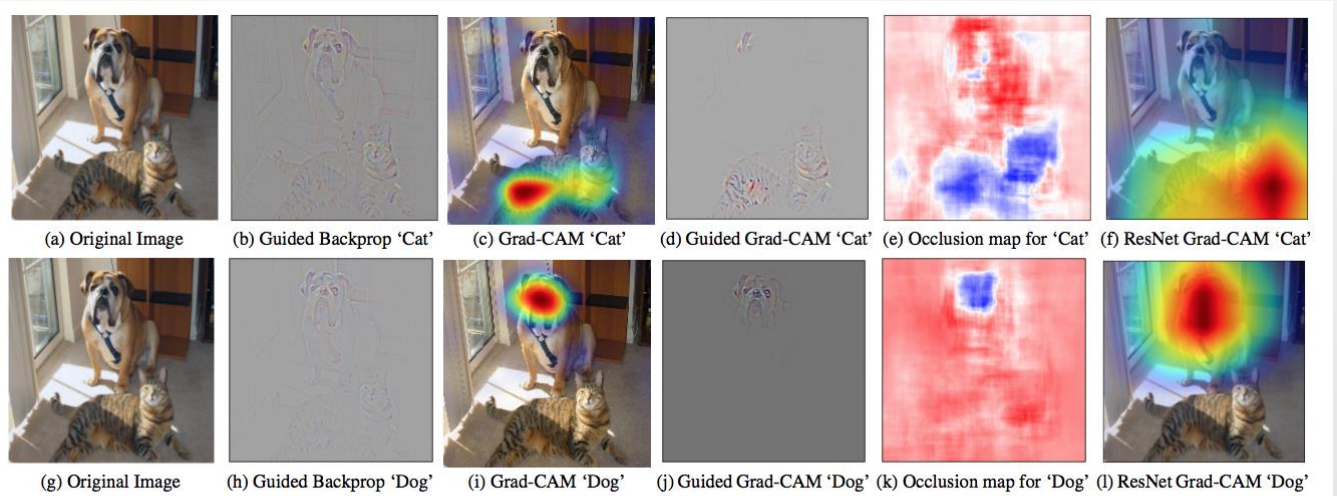


Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG and ResNet. (b) Guided Backpropagation [46]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations.Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (d, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.