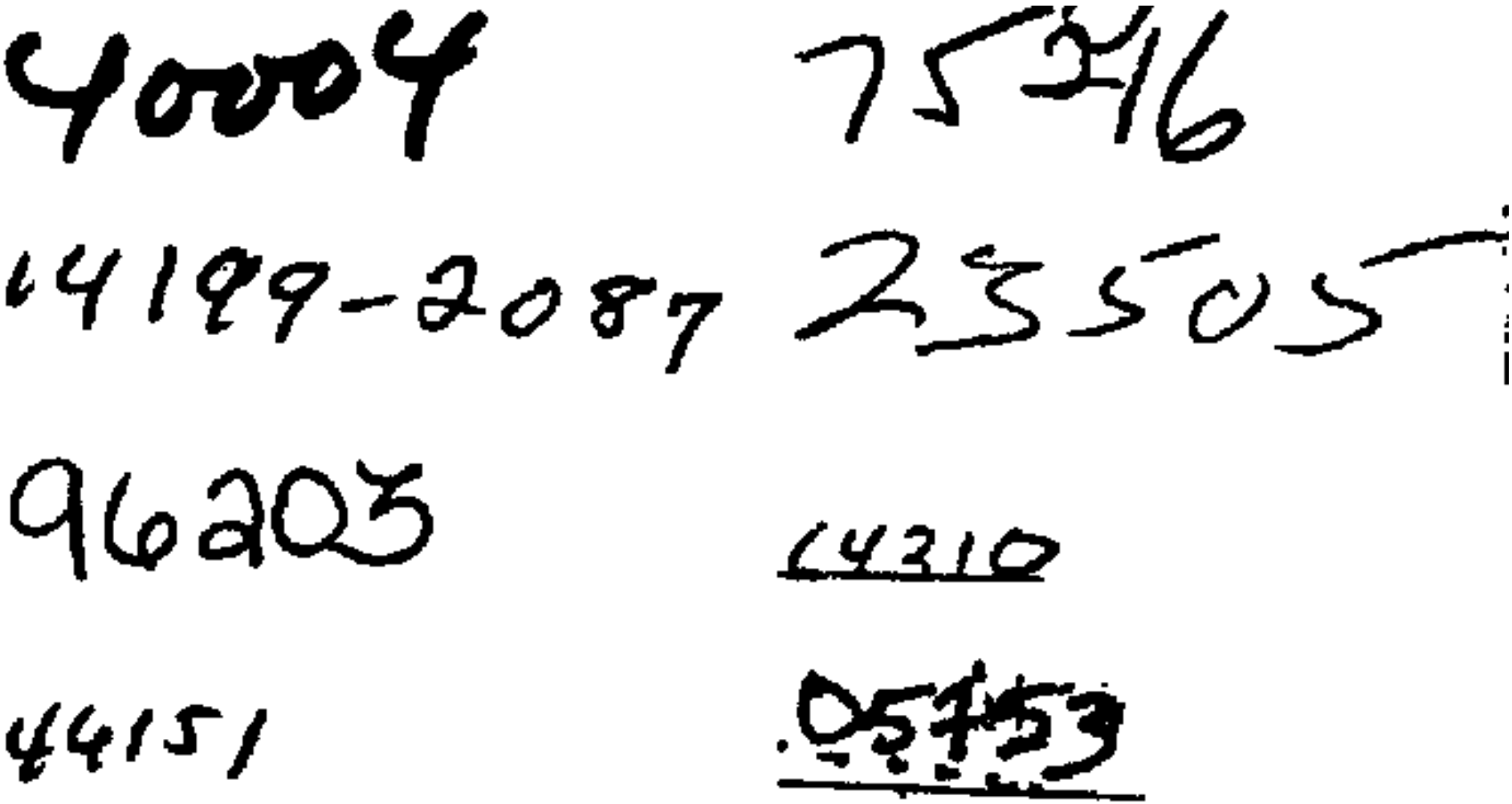
Part 8 Introduction to machine learning

Dr. Nguyen Quang Huy

May 16, 2020

What is machine learning

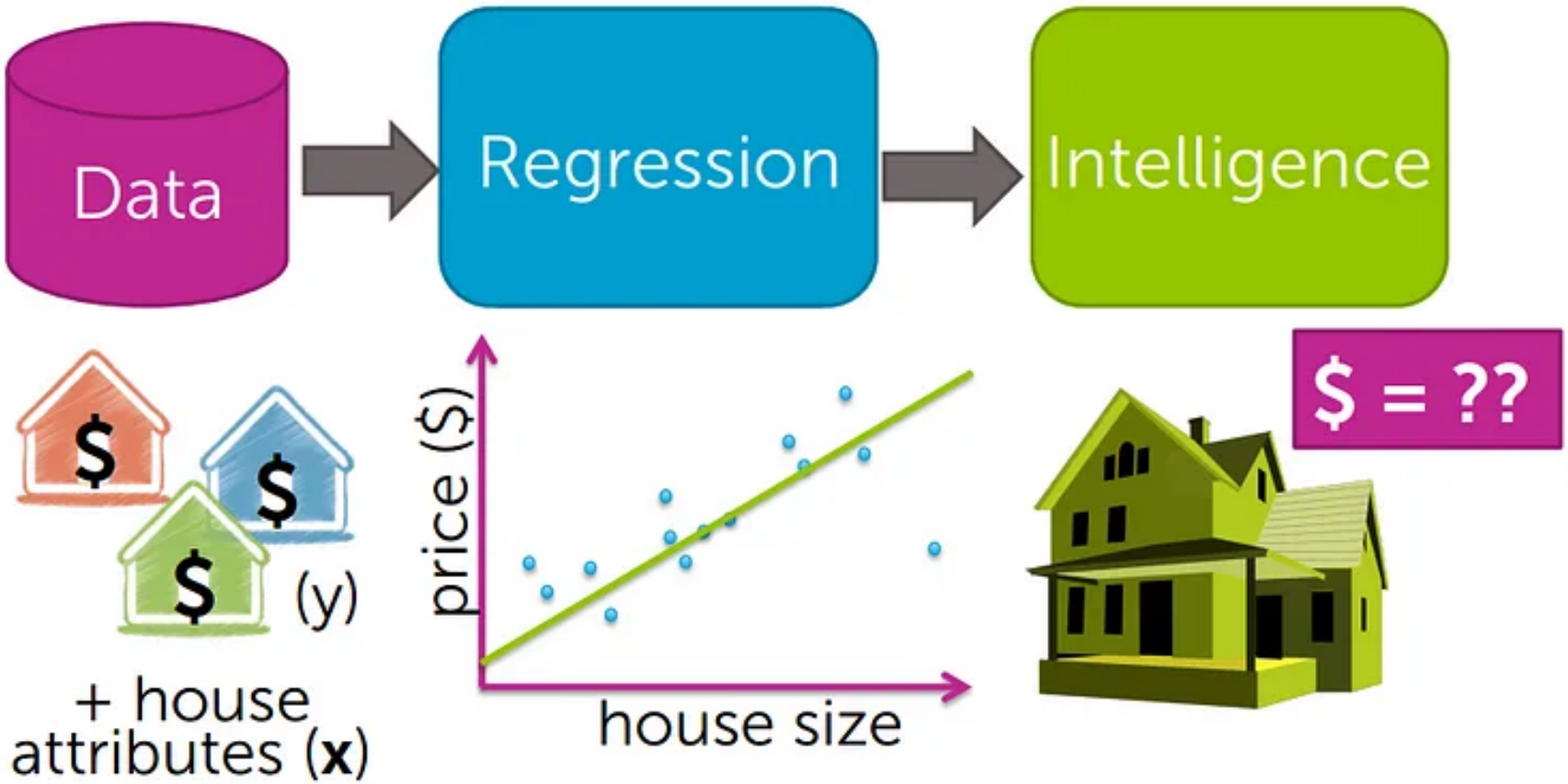
Handwritten zip code readers



What is machine learning

What is machine learning

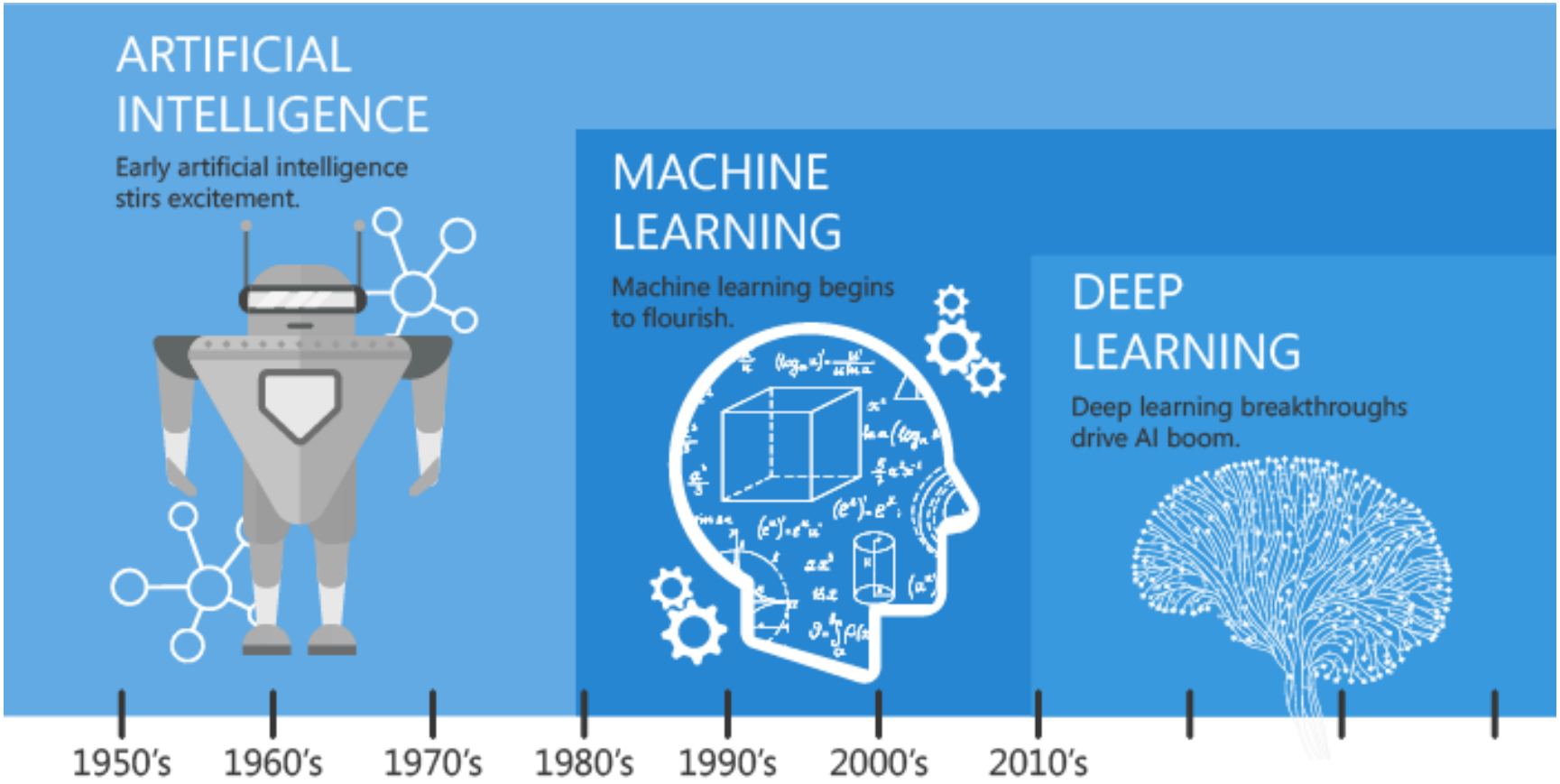
What is machine learning



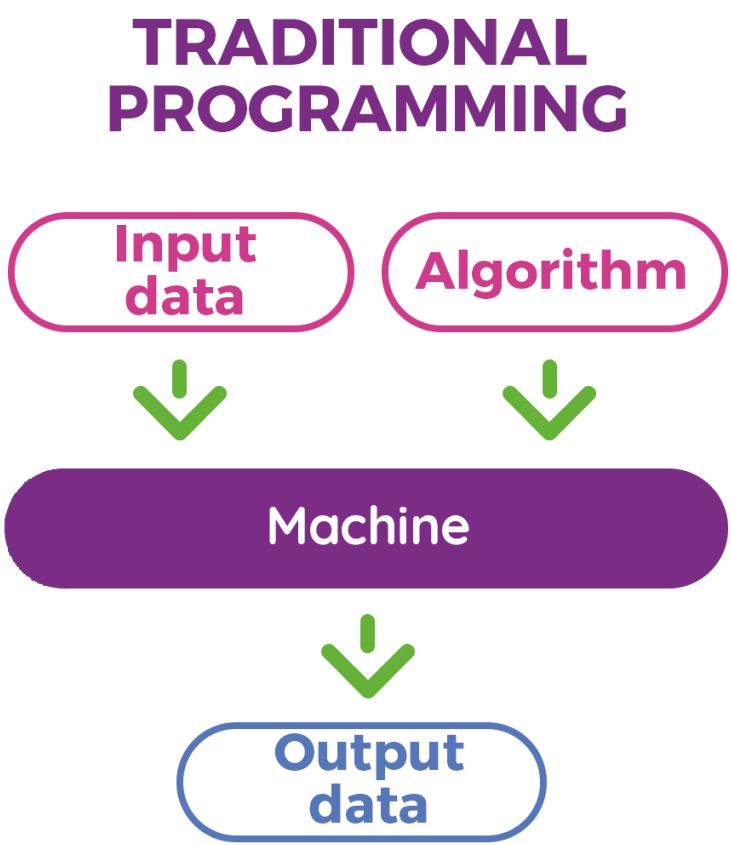
What is machine learning

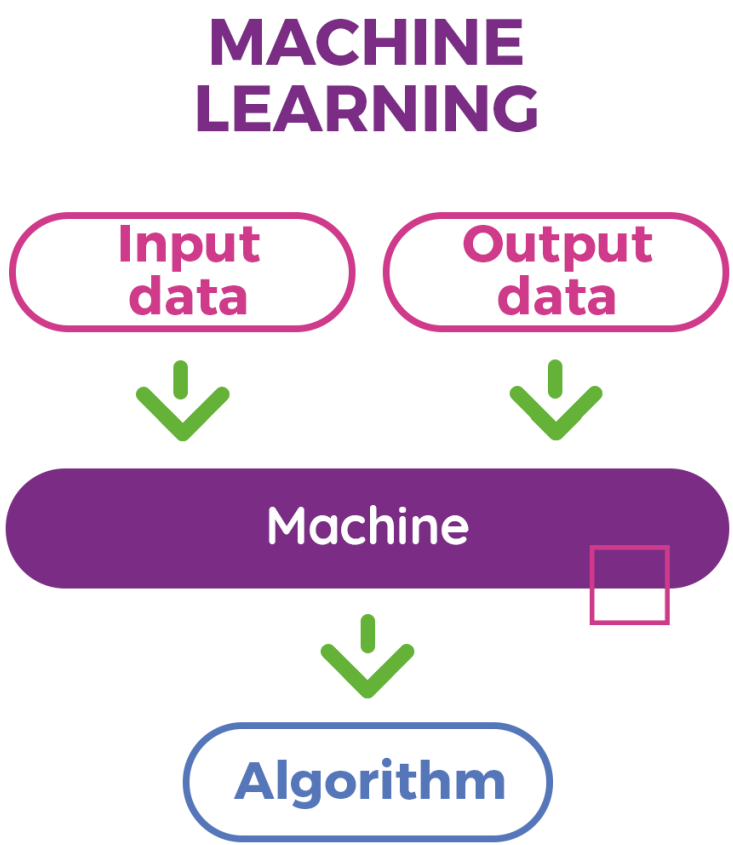
Machine learning versus Artificial Intelligence

Artificial intelligence (AI) is wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence.



Machine learning versus Artificial Intelligence





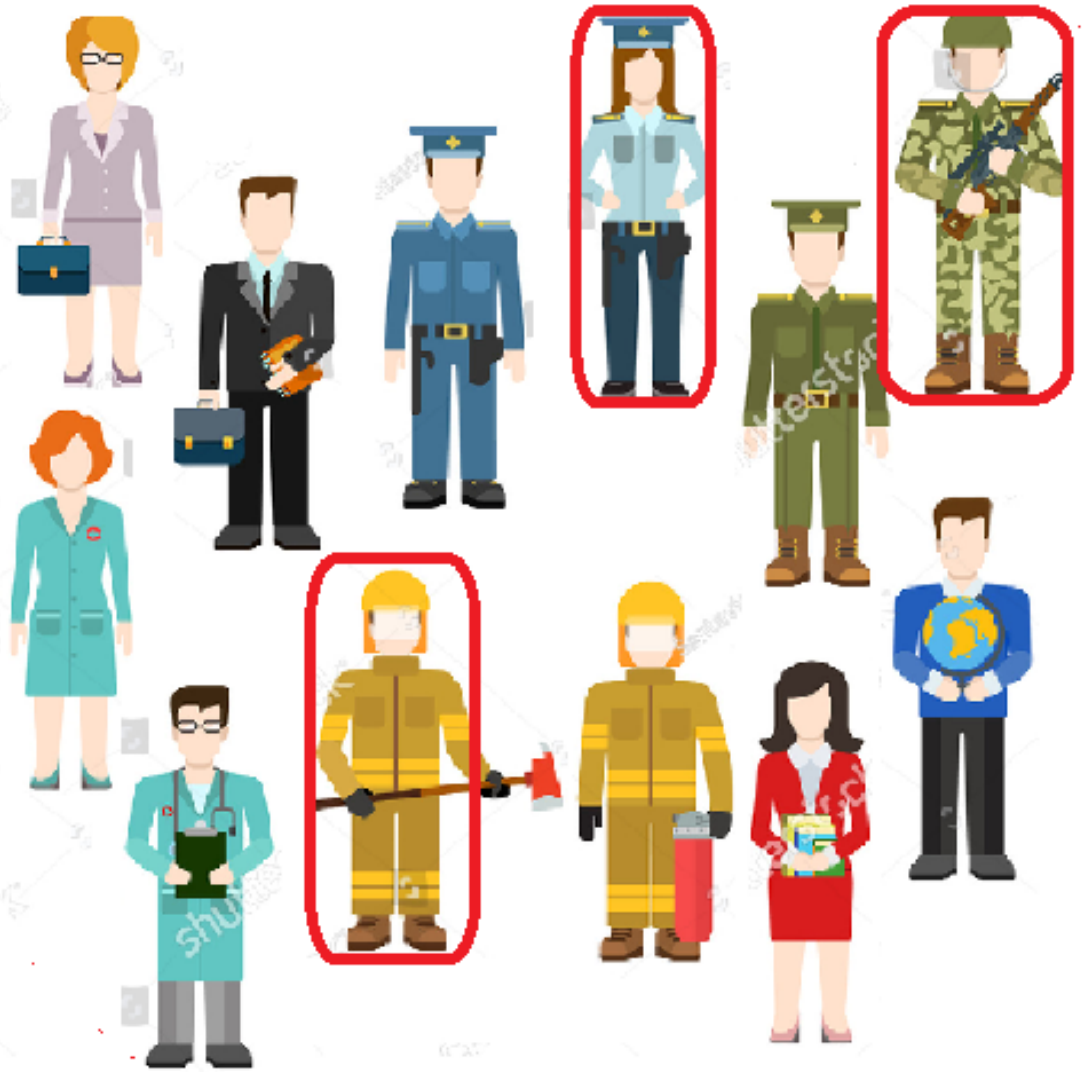
Machine learning versus Traditional Programming

These are rules to recognize our enemies!

Machine learning versus Traditional Programming

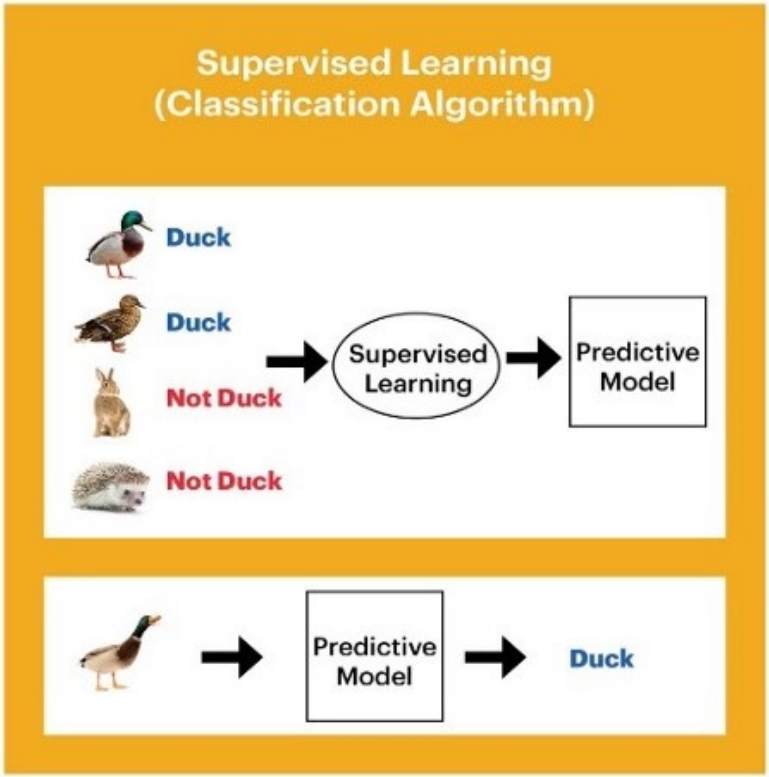
They are our enemies

recognize them by yourself!



What is machine learning

- Machine learning refers to a vast set of tools for **understanding data**.
- These tools can be classified as **supervised** or **unsupervised**



Supervised learning.

Several examples of supervised learning

- Design an automatic spam detector that could filter out spam before clogging the users' mailboxes.
- Predict the number of specific antigen from a number of clinical measures.
- Handwritten digit recognition.
- House price prediction.
- Automaticly detect a customer's commment is positive or negative.
- Early warning of changes in client creditworthiness in a bank.

Unsupervised learning

Several examples of unsupervised learning

- Customer segmentation to understanding different customer groups to build marketing and business strategies.
- Clustering DNA patterns to analyze evolutionary biology.
- Recommender systems, which involve grouping together users with similar viewing patterns in order to recommend similar content.
- Anomaly detection, including fraud detection or detecting defective mechanical parts.

Basic concepts: Features and Outcome

Input variables:

- Input variables can go by different names such as *predictors, independent variables, features* or just *variables*.
- The *input variables* are typically denoted using the symbol X, with a subscript to distinguish them: X_1, X_2 $\cdots$

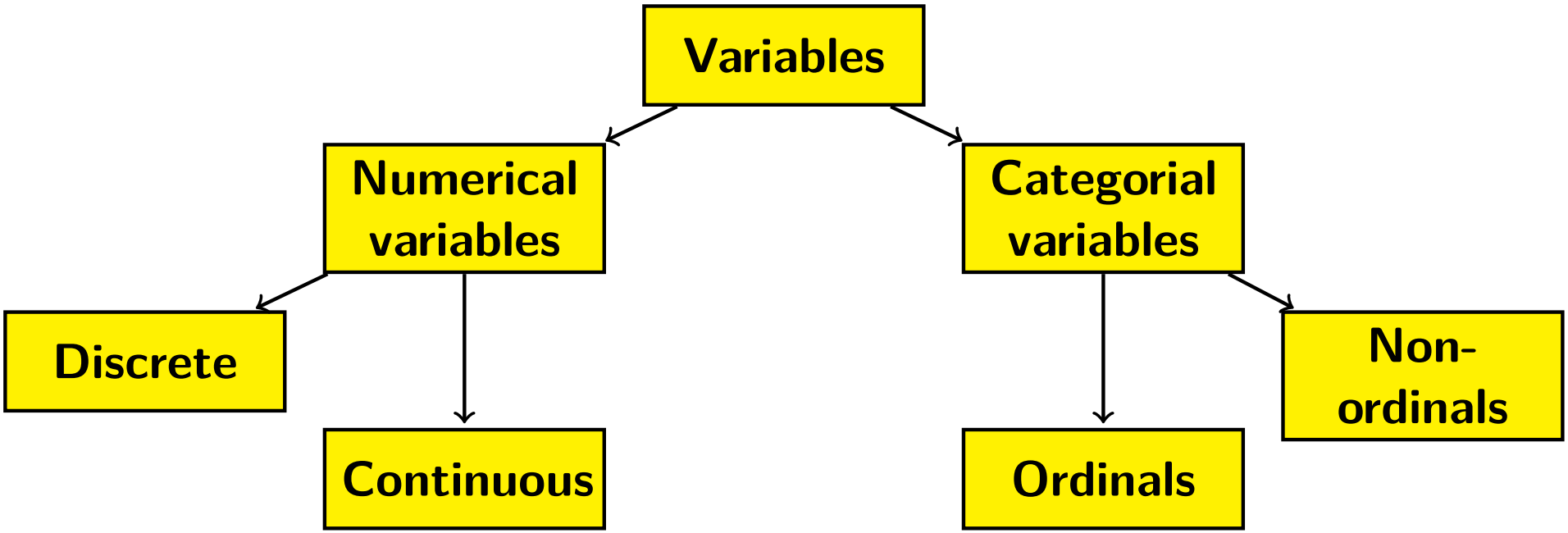
Output variables: typically denoted using the symbol Y.

- Often called *outcomes*, *responses* or *dependent variables*.

X_1	X_2	X_3	$\cdots$	X_p	Y
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,p}$	y_1
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,p}$	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$\cdots$	$x_{n,p}$	y_n

Variable types

The first step to build a machine learning algorithm is to know what type of variables.

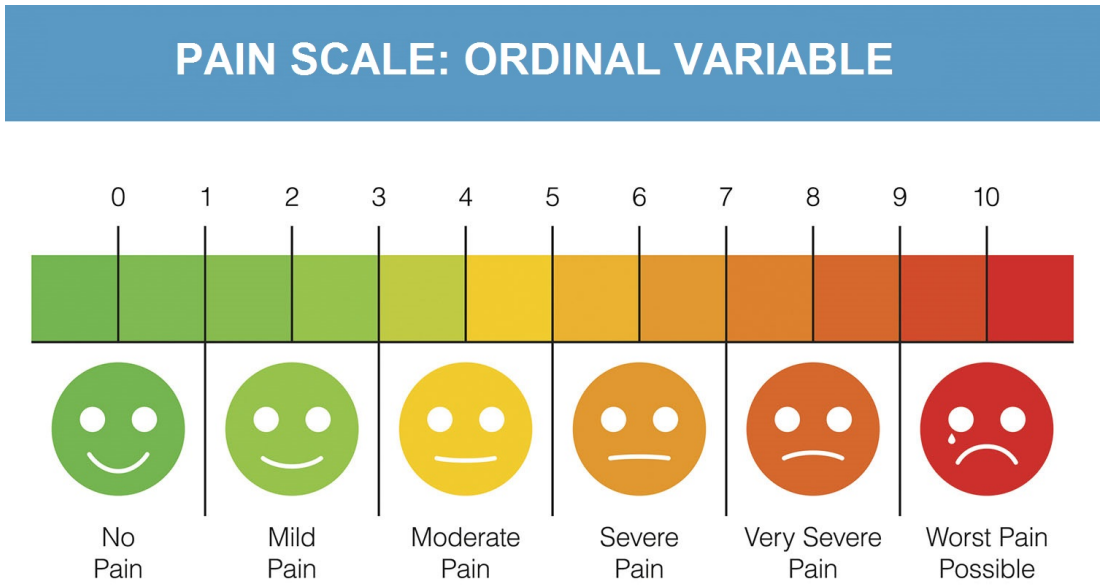


Categorical data

Variables that are defined by a small number of groups.

- A simple examples are sex, male or female.
- Regions: North, Central or South.

Non-ordinal or **nominal variable** is defined as variable that is used for namming or labelling while **ordinal variable** is a type of categorical variables **with an order**.



Numerical variables

Continuous variables can takes any value such as heights (measured with enough precision) while discrete variables must be counted such as heights rounded to the nearest inch.

There is a confusion between discrete variables and ordinal variables:

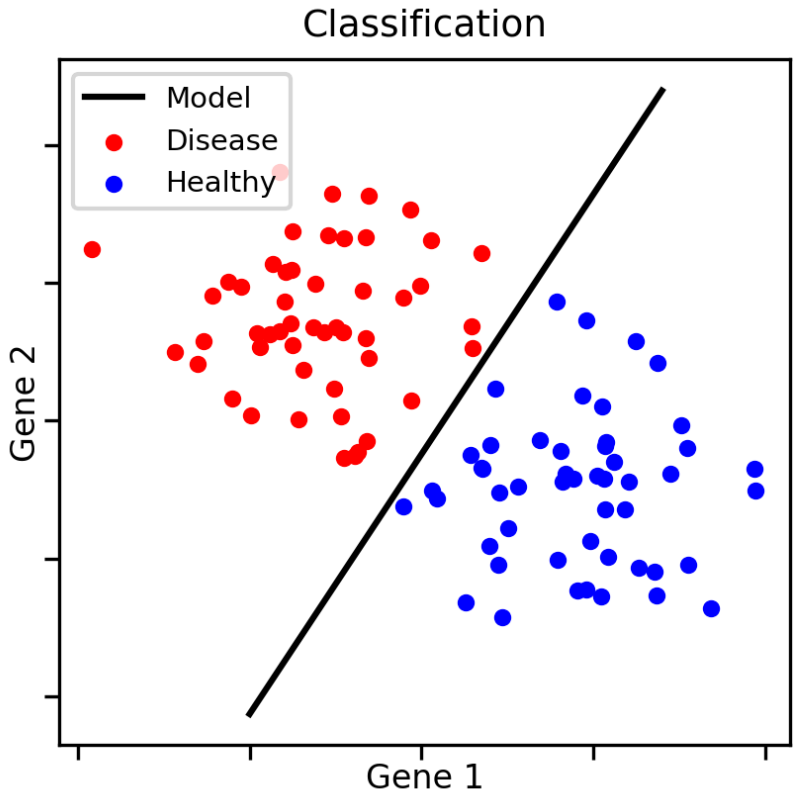
- Small number of groups with each group having many members $\rightarrow$ ordinal variable.
- Many groups with few cases in each group $\rightarrow$ discrete variable.

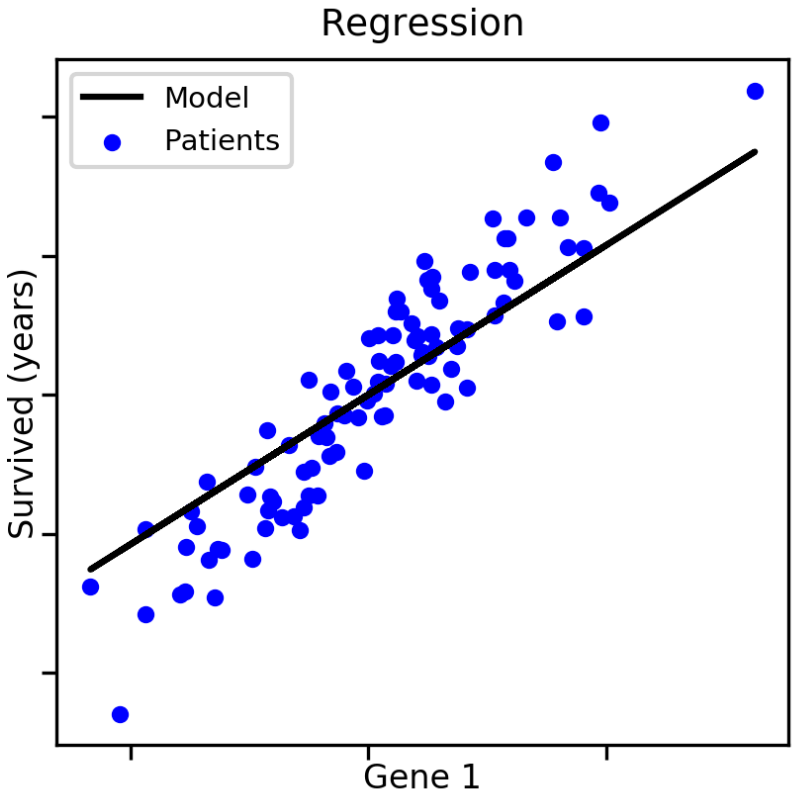
Example:

- Number of packs of cigarettes a person smokes a day: 0,1, or 2 would be considered ordinal
- Number of cigarettes the person smokes a days: 0, 1, 2, ... 50, ... would be considered a numerical variable.

Prediction: regression and classification

- When the output $\hat{Y}$ is continuous we refer to the machine learning task as regression;
- When the outcome is categorical, we refer to the machine learning task as classification

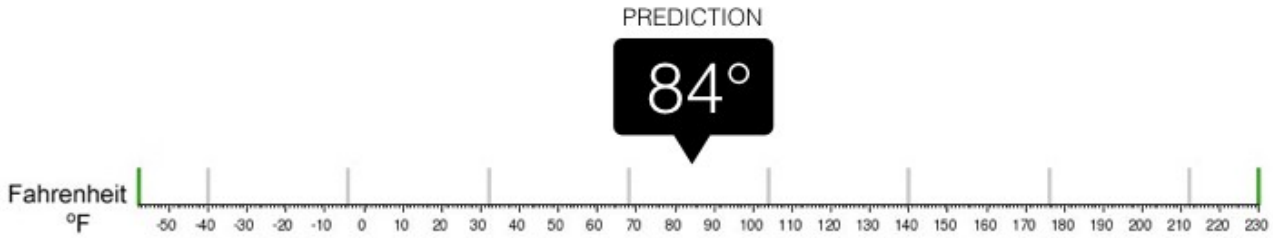




Prediction: regression and classification

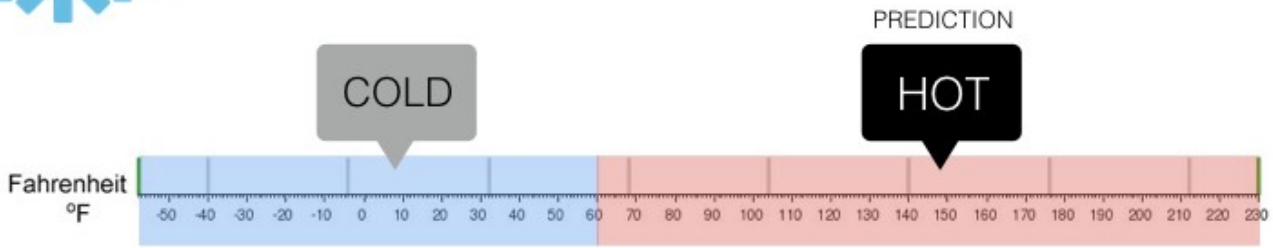
Regression

What is the temperature going to be tomorrow?



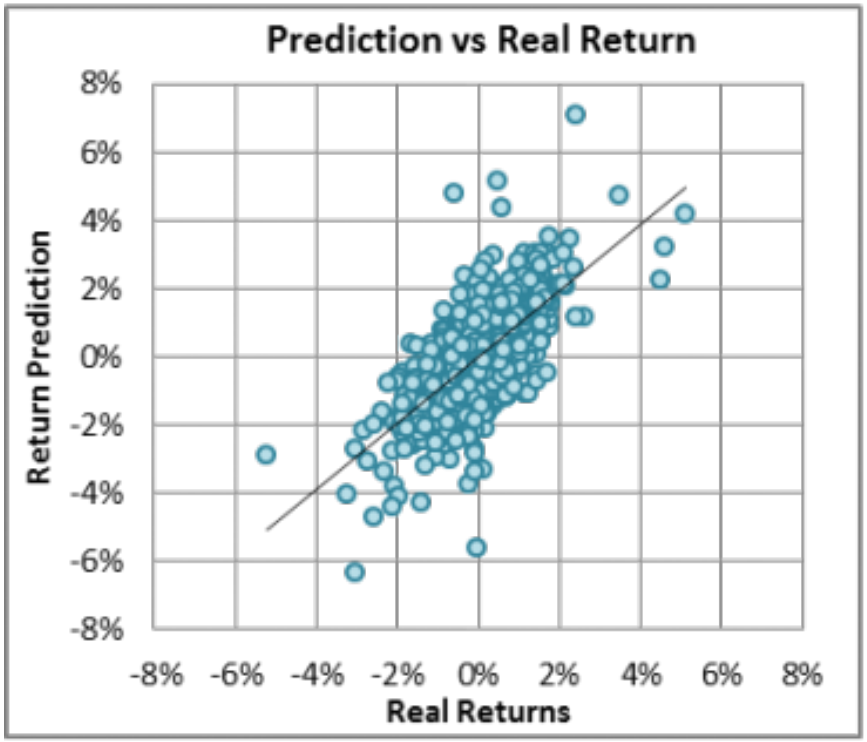
Classification

Will it be Cold or Hot tomorrow?



Prediction: regression and classification

Regression vs Classification



Basic concepts: Inference or prediction

The main output of all machine learning models is a function **f** such that

$$f(X) \approx Y$$

- We are interested in understanding the way that Y is affected as $X_1, \cdots, X_p$ change.
- f cannot be treated as a black box, we need to know its exact form.

We are often interested in the following questions:

- Which features are associated with the outcome?
- What is the relationship between the outcome and each feature?
- Can the relationship between outcome and each predictor be adequately summarized in a model?

Inference case study

```
library(tidyverse)
str(mpg)
```

```
## tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
##  $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
##  $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
##  $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8
##  $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999
##  $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr [1:234] "f" "f" "f" "f" ...
##  $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr [1:234] "p" "p" "p" "p" ...
##  $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
```

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

  Answer: *displ*, *year*, *cyl*, *trans*, *drv*, *class*

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

  Answer: *displ*, *year*, *cyl*, *trans*, *drv*, *class*

- What is the relationship between the outcome and each feature?

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

  Answer: *displ*, *year*, *cyl*, *trans*, *drv*, *class*

- What is the relationship between the outcome and each feature?

  Answer: Larger displ $\rightarrow$ less fuel economy, year 2008 $\rightarrow$ better fuel economy , more cyl $\rightarrow$ less fuel economy $\cdots$

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

  Answer: *displ*, *year*, *cyl*, *trans*, *drv*, *class*

- What is the relationship between the outcome and each feature?

  Answer: Larger displ $\rightarrow$ less fuel economy, year 2008 $\rightarrow$ better fuel economy , more cyl $\rightarrow$ less fuel economy $\cdots$

- Can the relationship between outcome and each predictor be adequately summarized in a model?

Inference case study

- Which features are associated with the car fuel economy (hwy or cty)?

  Answer: *displ*, *year*, *cyl*, *trans*, *drv*, *class*

- What is the relationship between the outcome and each feature?

  Answer: Larger displ $\rightarrow$ less fuel economy, year 2008 $\rightarrow$ better fuel economy , more cyl $\rightarrow$ less fuel economy $\cdots$

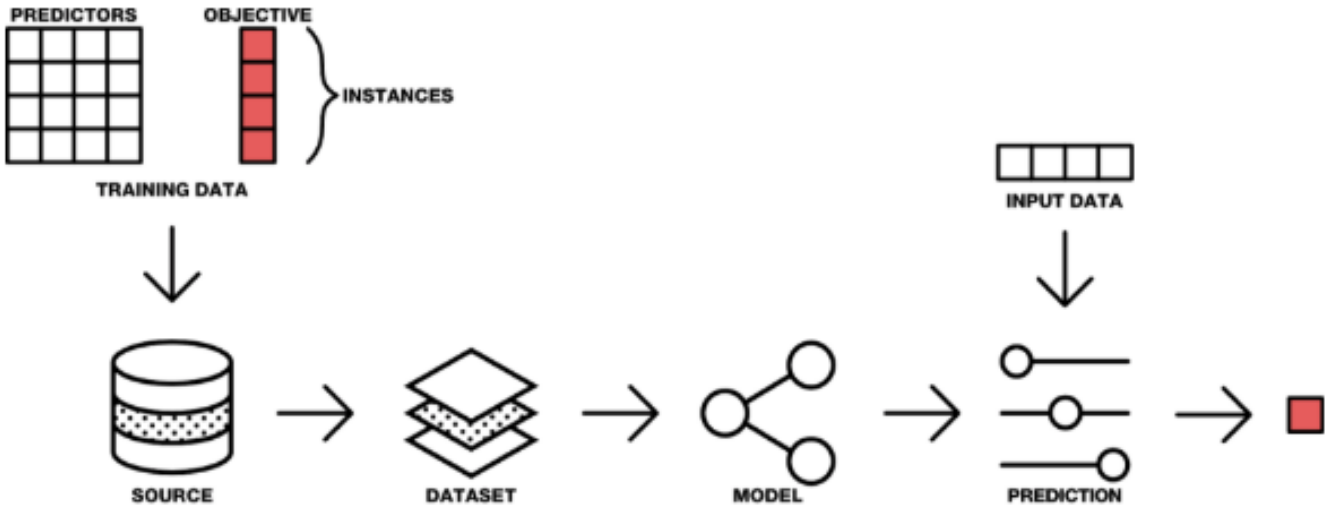
- Can the relationship between outcome and each predictor be adequately summarized in a model?

  Find a function f (in a closed form) such that

$$f(displ, year, cyl, trans, drv, class) \approx hwy$$

Basic concepts: Prediction

- When the set of inputs X are readily available but the output Y cannot be obtained.
- We can predict Y using $\hat{Y} = \hat{f}(X)$ where $\hat{f}$ is an estimation of function f.
- $\hat{f}$ could be treated as a black box, provided that it yields accurate predictions for Y.



Parametric vs non parametric models

Parametric models involve a two-step model-based approach

- First, we make an assumption about the functional form of f.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

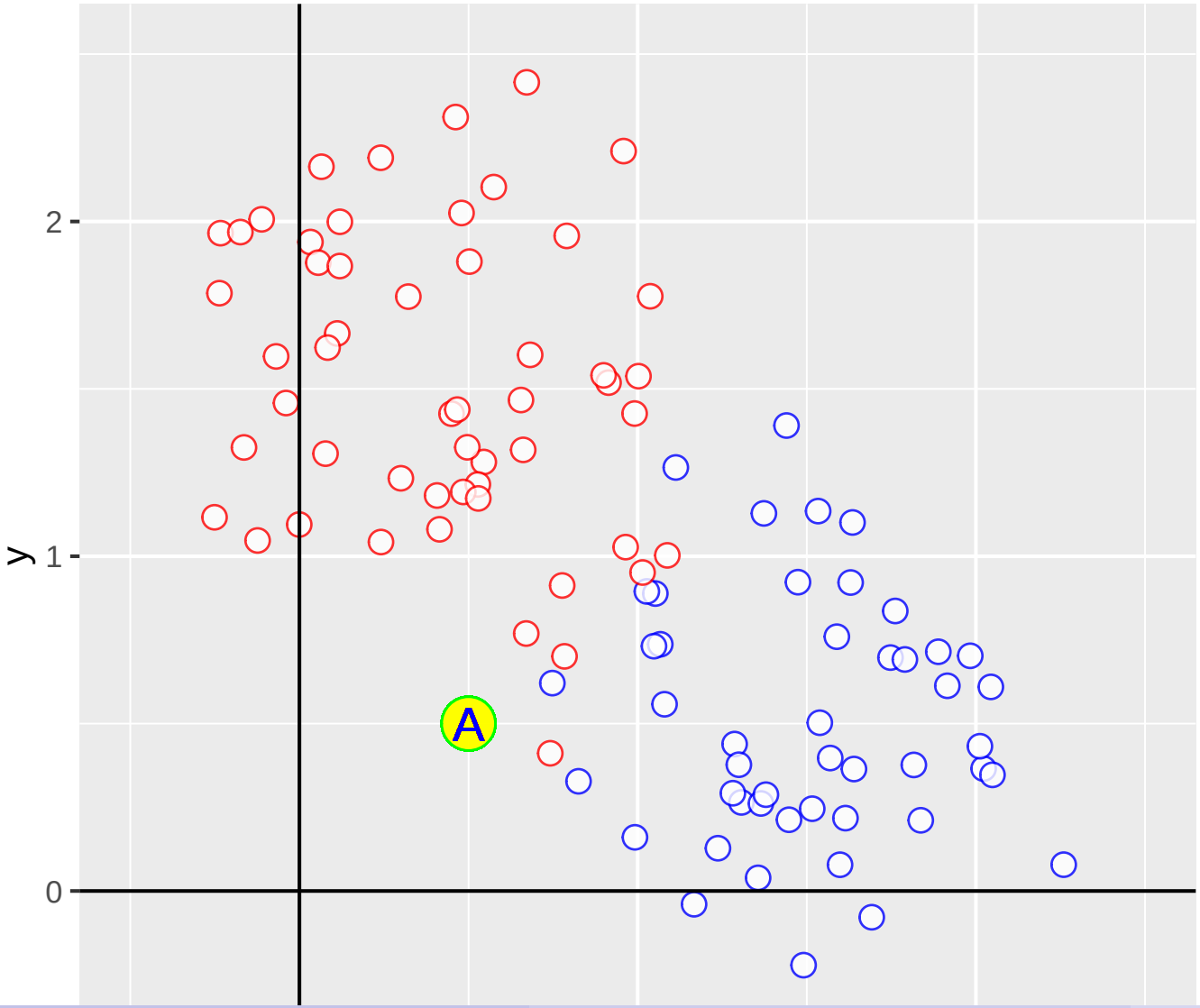
- Then, we need a procedure that uses the data to fit or train the model. In the case of the linear model we need to estimate the parameters $\beta_0, \beta_1, \cdots, \beta_p$

Non-parametric models

- do not make explicit assumptions about the functional form of f.
- We seek an estimate of f that gets as close to the data points.

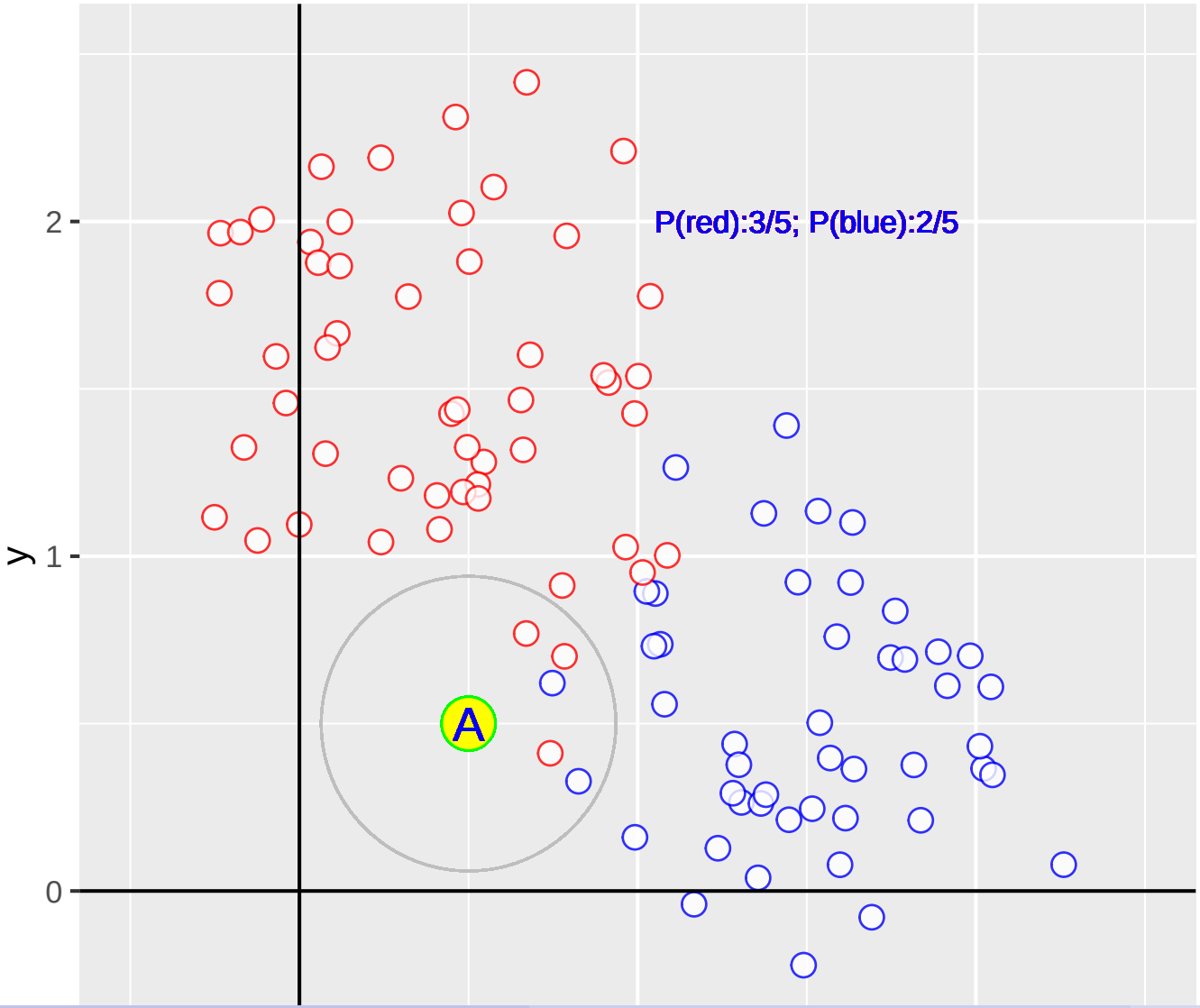
Parametric vs non parametric models

Which group does point A belong to ?



Parametric vs non parametric models

Non-parametric KNN method with parameter $k = 5$



Parametric vs non parametric models

Parametric method: logistic regression

- Step 1. Logistic model

$$\mathbb{P}(Colour = "Blue") = \frac{exp(\beta_0 + \beta_1\ x + \beta_2\ y)}{1 + exp(\beta_0 + \beta_1\ x + \beta_2\ y)}$$

- Step 2. Based on the dataset, using maximum likelihood approach we obtain:

$$(\beta_0, \beta_1, \beta_2) = (-8.668, 15.612, -7.286)$$

Thus:

$$\begin{aligned} \mathbb{P}(\mathbf{A} = "Blue") &= \frac{exp(\beta_0 + \beta_1 \times 0.5 + \beta_2 \times 0.5)}{1 + exp(\beta_0 + \beta_1 \times 0.5 + \beta_2 \times 0.5)} \\ &= 0.01 \end{aligned}$$

–> –> –> –> –> –> –> –> –> –> –> –> –> –> –> –> –> –> –> –>