# Chapter 1: Introduction

## (Presentation Slides Adapted from Book's Website)

MATH2319

1. **What is Predictive Data Analytics?**
2. **What is Machine Learning?**
3. **How Does Machine Learning Work?**
4. **What Can Go Wrong With ML?**
5. **PDA Project Lifecycle: Crisp-DM**

# What is Predictive Data Analytics?

- Predictive Data Analytics (PDA) encompasses
  1. the business
  2. the data processes (a.k.a. data pipeline)
  3. mathematical/ statistical/ computational models (a.k.a. machine learning)

  that enable a business to make **data-driven decisions**.

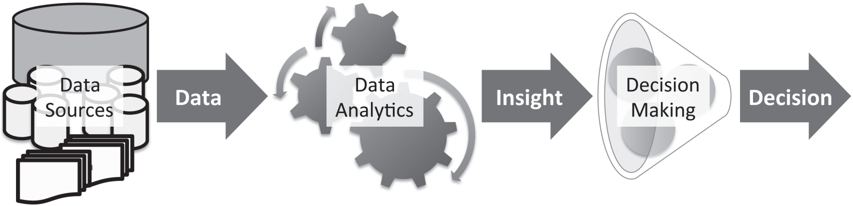



**Figure:** Predictive data analytics moving from **data** to **insights** to **decisions**.

**Example Applications:**

- Loan default prediction
- Fraud detection
- Medical diagnosis
- Document Classification
- Predict selling price of a diamond
- Can you predict when a volcano will erupt?

**Two Types of Predictions:**

- Predicting a categorical target variable: **Classification**
- Predicting a numerical target variable: **Regression**

# What is Machine Learning?

- (Supervised) Machine Learning techniques automatically learn a model of the relationship between a set of **descriptive features** and a **target feature** from a set of historical examples.

# Summary of the Learning Setup

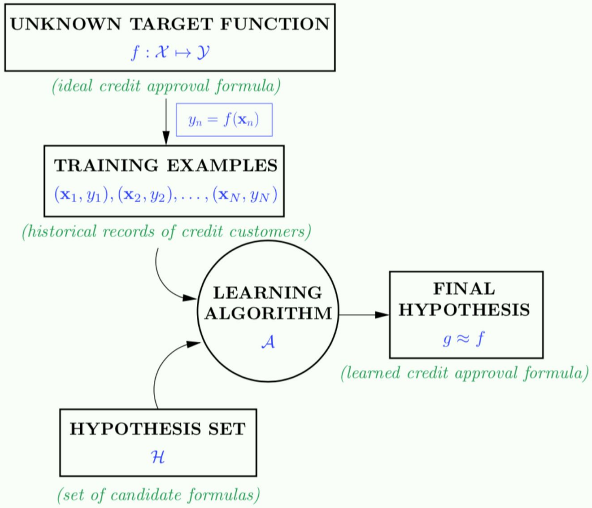



**Figure:** (from "Learning from Data" textbook)

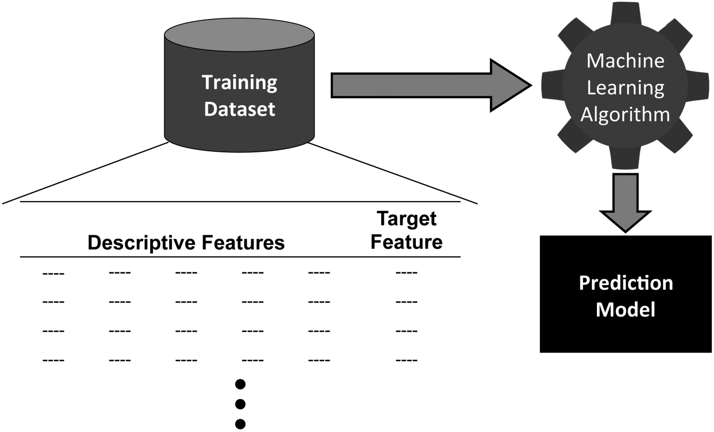



**Figure:** Using machine learning to induce a prediction model from a training dataset.


**Figure:** Using the model to make predictions for new query instances.

## Many names for the same things

Some other names for **features:**

- attributes
- variables

Some other names for a **target feature:**

- response variable
- dependent variable

Some other names for **descriptive features:**

- independent variables
- explanatory variables

Some names for **rows** in a dataset:

- instances
- observations
- records

| ID | Occupation | Age | Loan-Salary Ratio | Outcome |
|---|---|---|---|---|
| 1 | industrial | 34 | 2.96 | repaid |
| 2 | professional | 41 | 4.64 | default |
| 3 | professional | 36 | 3.22 | default |
| 4 | professional | 41 | 3.11 | default |
| 5 | industrial | 48 | 3.80 | default |
| 6 | industrial | 61 | 2.52 | repaid |
| 7 | professional | 37 | 1.50 | repaid |
| 8 | professional | 40 | 1.93 | repaid |
| 9 | industrial | 33 | 5.25 | default |
| 10 | industrial | 32 | 4.15 | default |

- What is the relationship between the **descriptive features** (Occupation, Age, Loan-Salary Ratio) and the **target feature** (Outcome)?

**if** LOAN-SALARY RATIO $>$ 3 **then**
  OUTCOME=*'default'*
**else**
  OUTCOME=*'repay'*
**end if**

- This is an example of a **prediction model**
- Notice that this model does not use all the features and the feature that it uses is a derived feature (in this case a ratio). Two important topics that we will return to again and again:
  - **feature design**
  - **feature selection**

- What is the relationship between the **descriptive features** and the **target feature** (OUTCOME) in the following dataset?
- In the table below, Type "stb" refers to second-time buyer and "ftb" refers to first-time buyer.

| ID | Amount | Salary | Loan-Salary Ratio | Age | Occupation | House | Type | Outcome |
|---|---|---|---|---|---|---|---|---|
| 1 | 245,100 | 66,400 | 3.69 | 44 | industrial | farm | stb | repaid |
| 2 | 90,600 | 75,300 | 1.2 | 41 | industrial | farm | stb | repaid |
| 3 | 195,600 | 52,100 | 3.75 | 37 | industrial | farm | ftb | default |
| 4 | 157,800 | 67,600 | 2.33 | 44 | industrial | apartment | ftb | repaid |
| 5 | 150,800 | 35,800 | 4.21 | 39 | professional | apartment | stb | default |
| 6 | 133,000 | 45,300 | 2.94 | 29 | industrial | farm | ftb | default |
| 7 | 193,100 | 73,200 | 2.64 | 38 | professional | house | ftb | repaid |
| 8 | 215,000 | 77,600 | 2.77 | 17 | professional | farm | ftb | repaid |
| 9 | 83,000 | 62,500 | 1.33 | 30 | professional | house | ftb | repaid |
| 10 | 186,100 | 49,200 | 3.78 | 30 | industrial | house | ftb | default |
| 11 | 161,500 | 53,300 | 3.03 | 28 | professional | apartment | stb | repaid |
| 12 | 157,400 | 63,900 | 2.46 | 30 | professional | farm | stb | repaid |
| 13 | 210,000 | 54,200 | 3.87 | 43 | professional | apartment | ftb | repaid |
| 14 | 209,700 | 53,000 | 3.96 | 39 | industrial | farm | ftb | default |
| 15 | 143,200 | 65,300 | 2.19 | 32 | industrial | apartment | ftb | default |
| 16 | 203,000 | 64,400 | 3.15 | 44 | industrial | farm | ftb | repaid |
| 17 | 247,800 | 63,800 | 3.88 | 46 | industrial | house | stb | repaid |
| 18 | 162,700 | 77,400 | 2.1 | 37 | professional | house | ftb | repaid |
| 19 | 213,300 | 61,100 | 3.49 | 21 | industrial | apartment | ftb | default |
| 20 | 284,100 | 32,300 | 8.8 | 51 | industrial | farm | ftb | default |
| 21 | 154,000 | 48,900 | 3.15 | 49 | professional | house | stb | repaid |
| 22 | 112,800 | 79,700 | 1.42 | 41 | professional | house | ftb | repaid |
| 23 | 252,000 | 59,700 | 4.22 | 27 | professional | house | stb | default |
| 24 | 175,200 | 39,900 | 4.39 | 37 | professional | apartment | stb | default |
| 25 | 149,700 | 58,600 | 2.55 | 35 | industrial | farm | stb | default |

**if** LOAN-SALARY RATIO $< 1.5$ **then**
  OUTCOME=*'repay'*
**else if** LOAN-SALARY RATIO $> 4$ **then**
  OUTCOME=*'default'*
**else if** AGE $< 40$ **and** OCCUPATION $=$*'industrial'* **then**
  OUTCOME=*'default'*
**else**
  OUTCOME=*'repay'*
**end if**

- The real value of machine learning becomes apparent in situations like this when we want to build prediction models from large datasets with multiple features.

# How Does Machine Learning Work?

- Machine learning algorithms work by searching through a set of possible prediction models for the model that best captures the relationship between the descriptive features and the target feature.
- An obvious search criteria to drive this search is to look for models that are consistent with the data
  - **Consistent Model:** A model that makes no mistakes on the training data
- However, because a training dataset is only a sample, ML is an ill-posed problem.

**Table:** A simple retail dataset

| ID | BABY FOOD | ALCOHOL | ORGANIC | GROUP TYPE (TARGET) |
|---|---|---|---|---|
| 1 | no | no | no | couple |
| 2 | yes | no | yes | family |
| 3 | yes | yes | no | family |
| 4 | no | no | yes | couple |
| 5 | no | yes | yes | single |

- Here it is assumed that there is no noise in the data.
- Notice there are $2^3 = 8$ possible combinations, and we already know the target value for 5 of them (3 descriptive features all with 2 possible values).
- The goal here is find a ML model to predict the remaining 3 combinations.

**Table:** A sample of the models that are consistent with the training data

| BBY | ALC | ORG | GRP | $\mathbb{M}_1$ | $\mathbb{M}_2$ | $\mathbb{M}_3$ | $\mathbb{M}_4$ | $\mathbb{M}_5$ | ... | $\mathbb{M}_{6\,561}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| no | no | no | couple | couple | couple | single | couple | couple | | couple |
| no | no | yes | couple | single | couple | single | couple | couple | | single |
| no | yes | no | ? | family | family | single | single | single | | family |
| no | yes | yes | single | single | single | single | single | single | | couple |
| yes | no | no | ? | couple | couple | family | family | family | ... | family |
| yes | no | yes | family | couple | family | family | family | family | | couple |
| yes | yes | no | family | single | family | family | family | family | | single |
| yes | yes | yes | ? | single | single | family | family | couple | | family |

- There are $3^3 = 27$ models that are all consistent with the training data (3 missing combinations, 3 different group possibilities).
- They all agree on the sample data, but they disagree on the predictions denoted by ?.
- Since a single consistent model cannot be found based on a sample training dataset, ML is said to be ill-posed.

- Consistency $\approx$ memorizing the dataset.
- Consistency with noise in the data isn't desirable.
- Goal: a model that **generalises** beyond the dataset and that isn't influenced by the noise in the dataset.
- So what criteria should we use for choosing between models?

- Inductive bias is the set of assumptions that define the model selection criteria of an ML algorithm.
- There are two types of bias that we can use:
  1. Restriction bias (required)
  2. Preference bias (optional, usually a good idea to avoid overfitting)
- Inductive bias is necessary for learning (beyond the dataset).

## Examples of Inductive Bias

- Example 1:
  - Restrictive bias: a tree model
  - Preference bias: shallow trees
- Example 2:
  - Restrictive bias: a nearest neighbour model
  - Preference bias: higher degrees of $K$ (the number of neighbours to examine)
- Example 3:
  - Restrictive bias: a linear model
  - Preference bias: fewer number of terms

### How ML works (Summary)

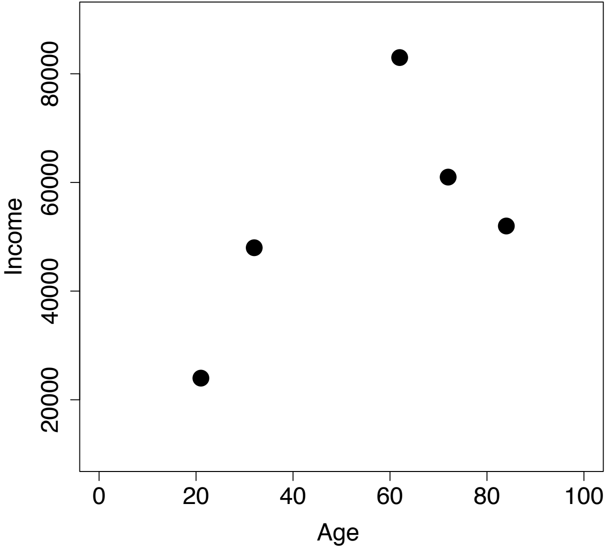
- ML algorithms work by searching through sets of potential models.
- There are two sources of information that guide this search:
  1. the training data,
  2. the inductive bias of the algorithm.

# What Can Go Wrong With ML?

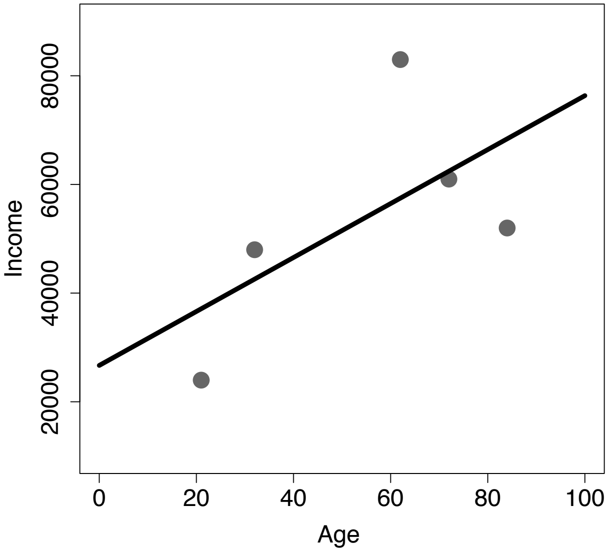
- No free lunch!
- What happens if we choose the wrong inductive bias:
  1. **underfitting**
  2. **overfitting**

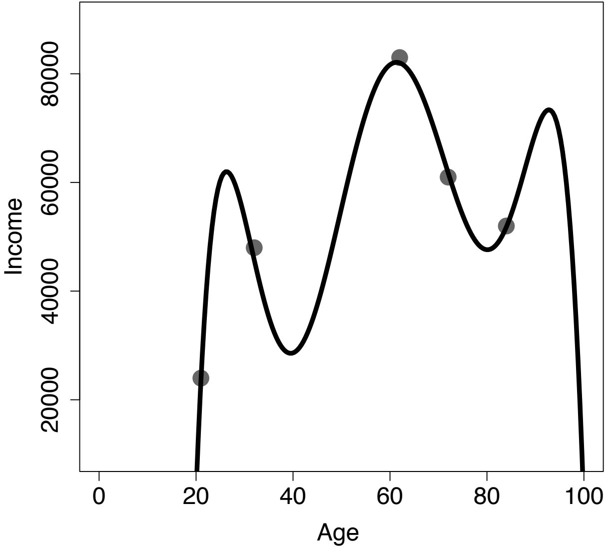
**Table:** The age-income dataset.

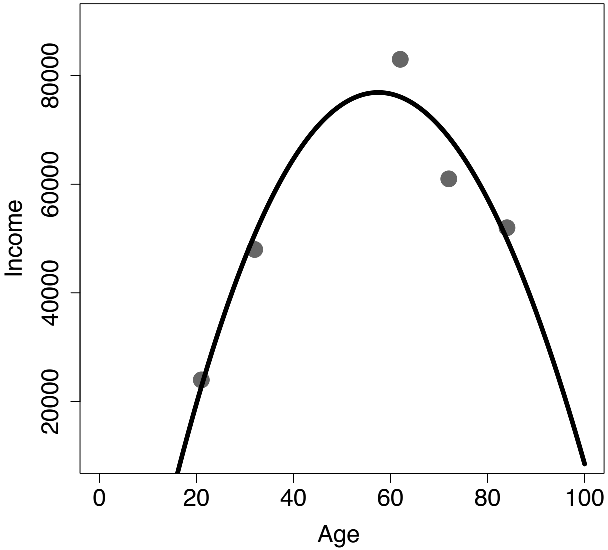
| ID | AGE | INCOME |
|---|---|---|
| 1 | 21 | 24,000 |
| 2 | 32 | 48,000 |
| 3 | 62 | 83,000 |
| 4 | 72 | 61,000 |
| 5 | 84 | 52,000 |

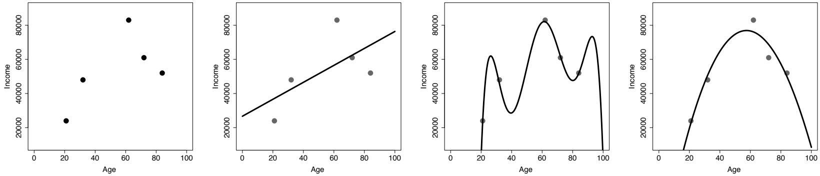


80000
60000
40000
20000
Income
0
20
40
60
80
100
Age


Income
80000
60000
40000
20000
0
20
40
60
80
100
Age


Income
Age
20000
40000
60000
80000
0
20
40
60
80
100


Income
80000
60000
40000
20000
0
20
40
60
80
100
Age



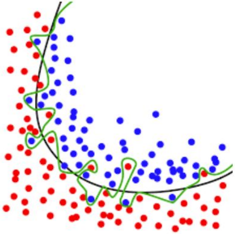

(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

**Figure:** Trying to achieve a balance when trying to predict income from age

- Striking the right balance between model complexity and simplicity (between underfitting and overfitting) is the **hardest part** of machine learning.

## Bias-Variance Trade-off Illustration

In a 2-dimensional feature space:



- The green boundary gives 100% accuracy on the (training) dataset whereas the black boundary makes some mistakes.
- The inexperienced ML practitioner will go for the green boundary.
- However, what you should really use in the real world (for making predictions) is the black boundary and **not** the green boundary.
- Why?

## Concept of Bias-Variance Trade-off

- The green boundary corresponds to a very complex model and it has
  - **High Variance:** the boundary is likely to change (a lot!) for a new sample from the population of all possible training datasets.
  - **Low Bias:** the boundary is likely to have very little difference from the true boundary for this training dataset.
- The black boundary corresponds to a very simple model and it has
  - **Low Variance**
  - **High Bias**
- In general, all models will trade some bias with some variance.
- Figuring out the optimal model complexity (that is, the optimal bias-variance trade-off) is at the **heart of machine learning**.
- However, figuring out the optimal model complexity is more difficult than it looks and this is why machine learning is **both art and science at the same time**.

# Types of machine learning algorithms

- Four fundamental learning paradigms:
  1. **Information based learning**
  2. **Similarity based learning**
  3. **Probability based learning**
  4. **Error based learning**
- Basic ML algorithms are sometimes combined via boosting and ensembles, giving rise to more complex ML algorithms:
  - Gradient boosting
  - Random forests
  - Etc.

# PDA Project Lifecycle: Crisp-DM

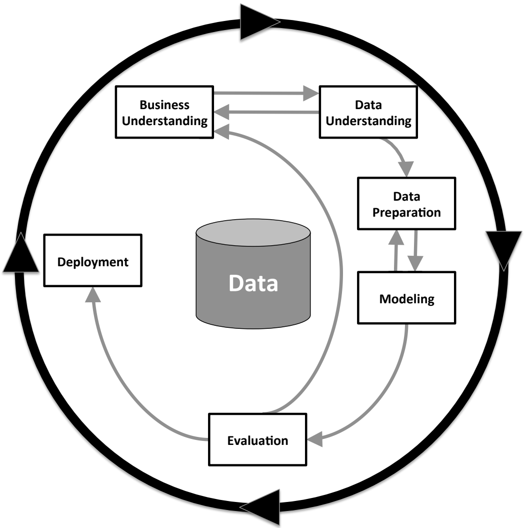



**Figure:** A diagram of the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** process which shows the six key phases and indicates the important relationships between them.