THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

INTRODUCTION TO MACHINE LEARNING

DIT 45100

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Module 1
Introduction

Agenda

- Course Overview
- What is machine learning?
- How machine learning works

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Course Overview

Course Objectives

- This course introduces the fundamentals of machine learning theory and algorithm design that is necessary to adapt and apply AI to solve real-world problems.
- The course provides a foundational understanding of various machine learning and statistical pattern recognition algorithms and, using available machine learning platforms, teaches practical and applied design, testing, and implementation of machine learning models.
- Course Outcomes
  - Describe and apply machine learning algorithms and models, including rule and tree-based classifiers, uncertainty modeling, clustering and correlation techniques, and numeric prediction.
  - Design, construct, test, and implement supervised and unsupervised machine learning models using Python programming language and supporting libraries

Preparatory Activities

- The following preparatory activities prior to model building have already been covered in previous courses.
  - Business Understanding
  - Data Understanding
  - Data Preparation

Machine Learning Algorithms

- The following five families of supervised machine learning algorithms for building artificial intelligence solutions:

  Linear Regression Models

  Classification Models

  Non-Parametric Models

  Probabilistic Models

  Decision Trees

Machine Learning Algorithms

- We will also look at modelling approaches for:

  Unsupervised Learning

  Text Classification

- We also cover a range of approaches to evaluating & optimizing prediction models along the way.

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

What is Machine Learning?

Machine Learning

- Machine learning and pattern recognition can be viewed as two facets of the same field.
- Machine learning grew out of computer science, whereas pattern recognition has its origin in engineering.
- Both have undergone substantial development over the past two to three decades.
- Machine learning, now a days, is all around us

Machine Learning

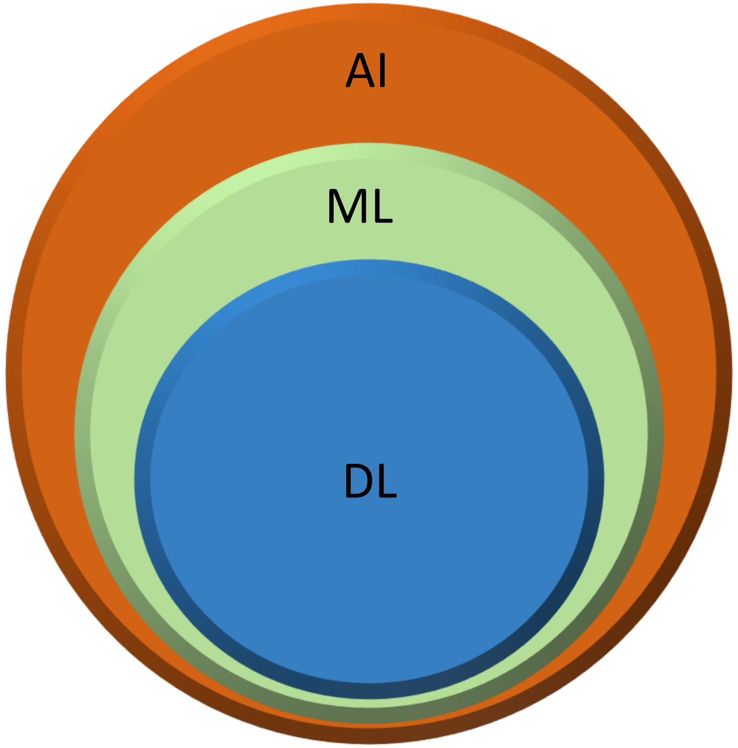
Formal definition:

Machine learning is the subfield of computer science that gives "computers the ability to *learn without being explicitly programmed*".

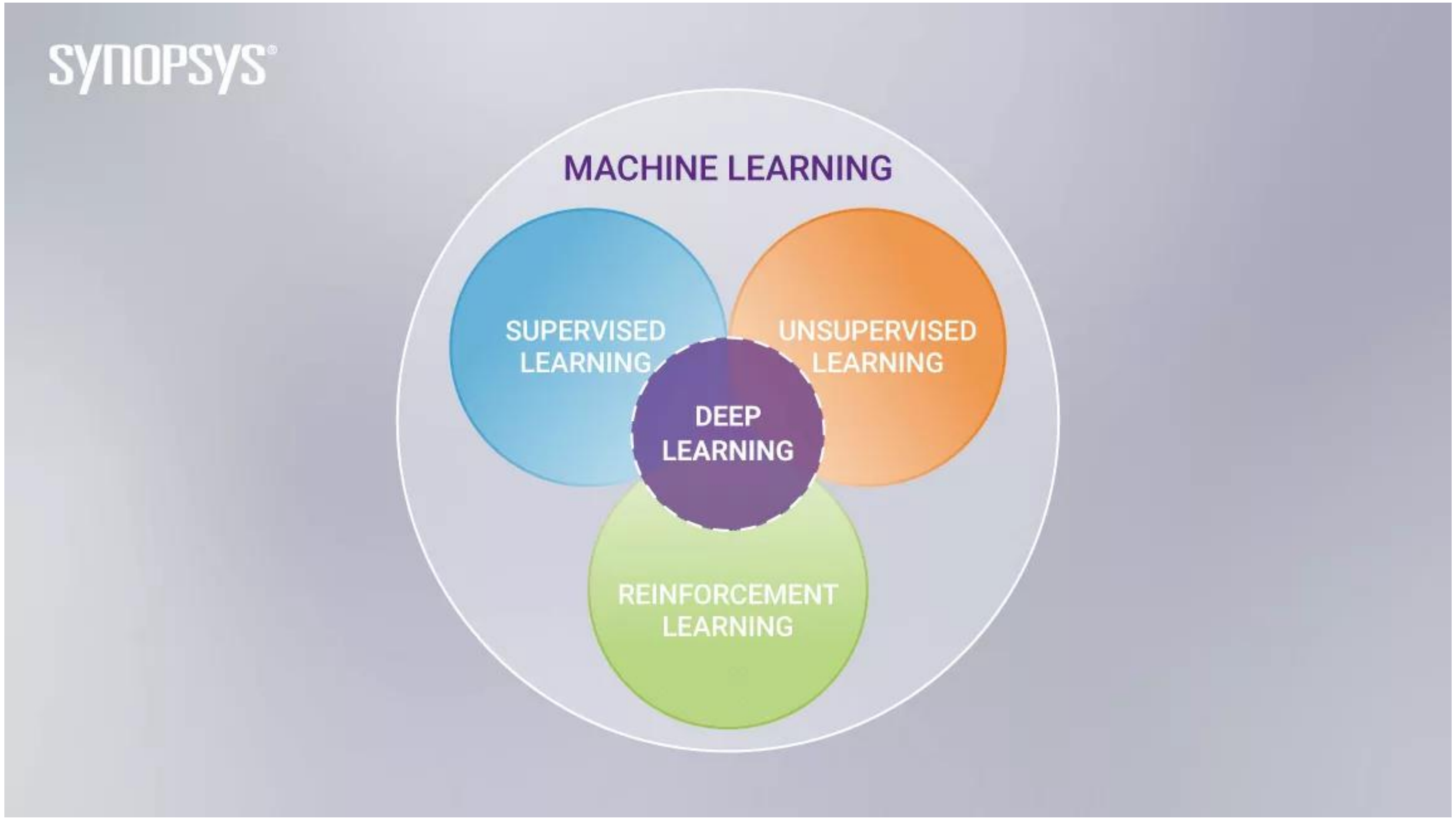
Arthur Samuel, 1959.

Machine Learning

- Machine Learning is a subfield of Artificial Intelligence that enables machines to improve at a given task with experience without being explicitly programmed.
- Deep Learning (DL) is a subset of ML that aims at imitating the human brain using mathematical equations.



Machine Learning

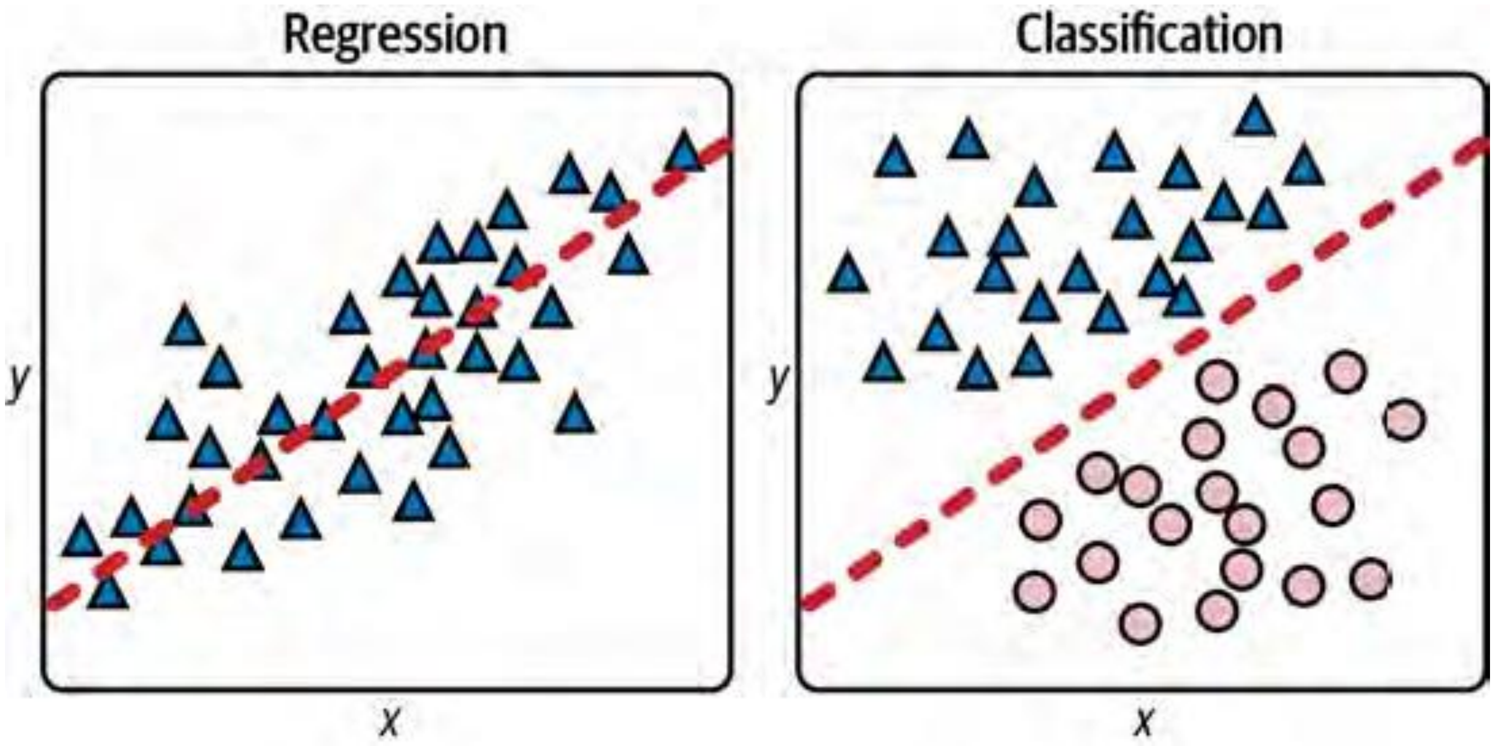


Flavors of Machine Learning

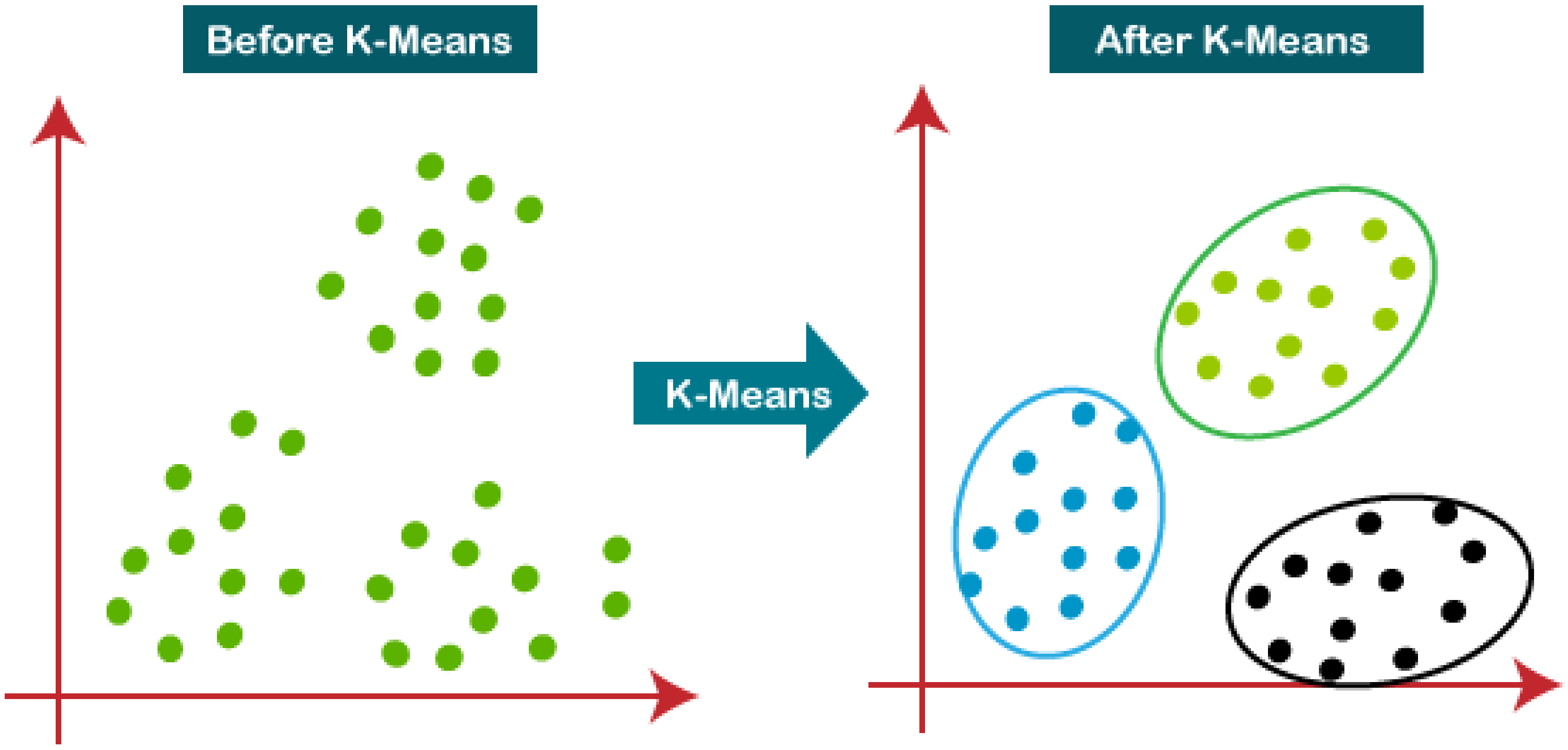
There are three major forms of machine learning

- Supervised Learning
  - Learning from examples
  - Labelled data
- Unsupervised Learning
  - Learning the internal structure of the data
  - No labels
- Reinforcement Learning
  - Learning from the environment
  - Actions / perceptions

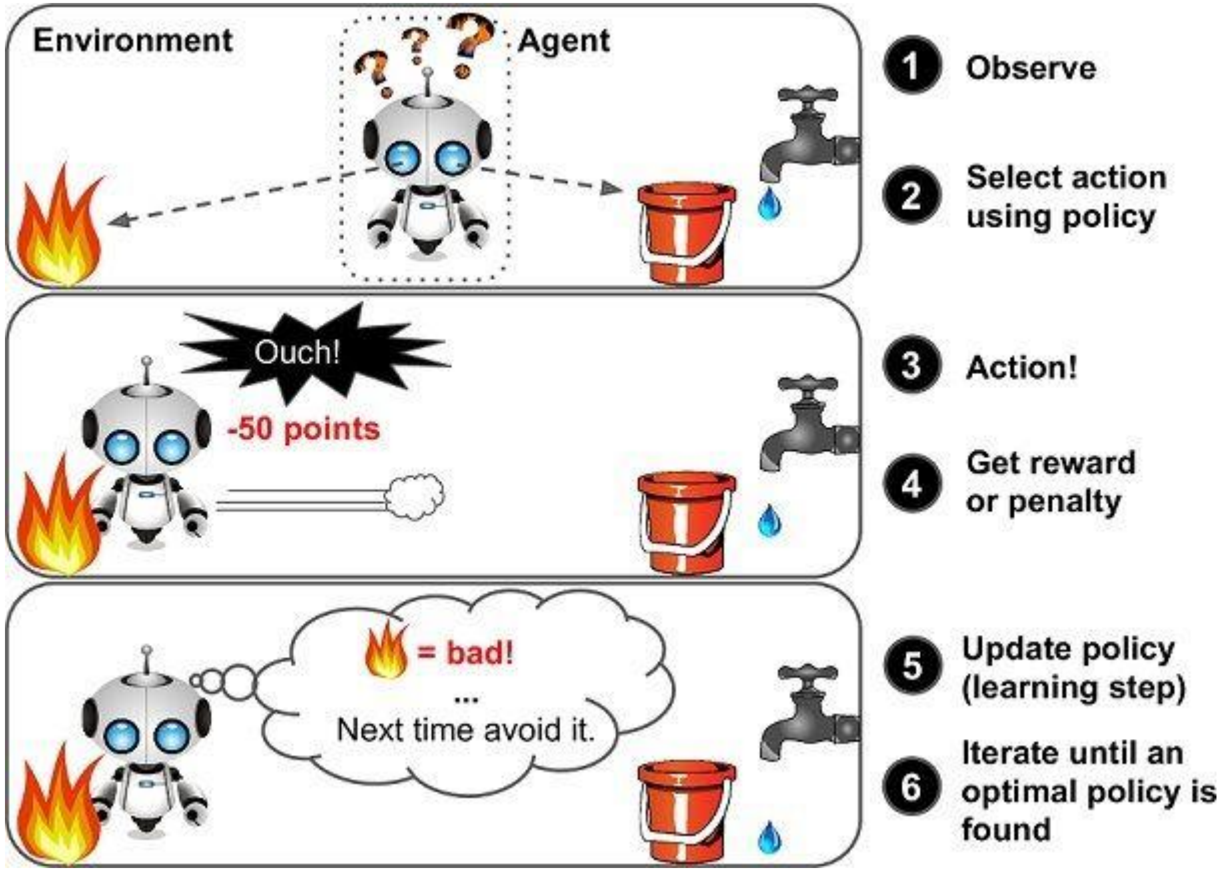
Supervised Machine Learning



Unsupervised ML: Clustering



Reinforcement Learning



Machine Learning

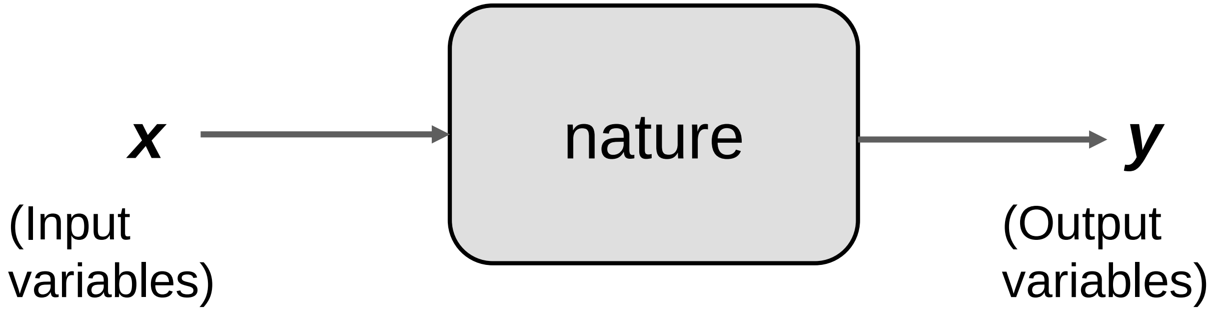
- (Supervised) Machine Learning is different from traditional approach to statistical modeling
  - Statistical modeling:
    - Model defining the rules → Answers
    - Objective is understanding the relationships
  - Machine Learning:
    - Answers → Model defining the rules
    - Objective is to make (accurate) predictions

Statistical Modeling: The Two Cultures

- There are two cultures in the use of statistical modeling to reach conclusions from data
  - Data Modeling
    - Traditionally, 98% of all statisticians use this approach
  - Algorithmic Modeling
    - 2% of statisticians, and many in other fields use this approach

https://www.jstor.org/stable/2676681

Statistical Modeling

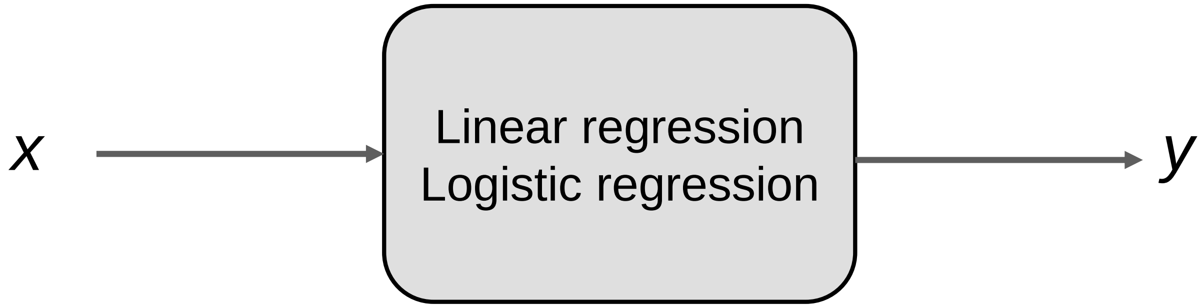


Goals of Data Analysis:

1. How **y** relates to **x**
2. Predict **y** from new **x**

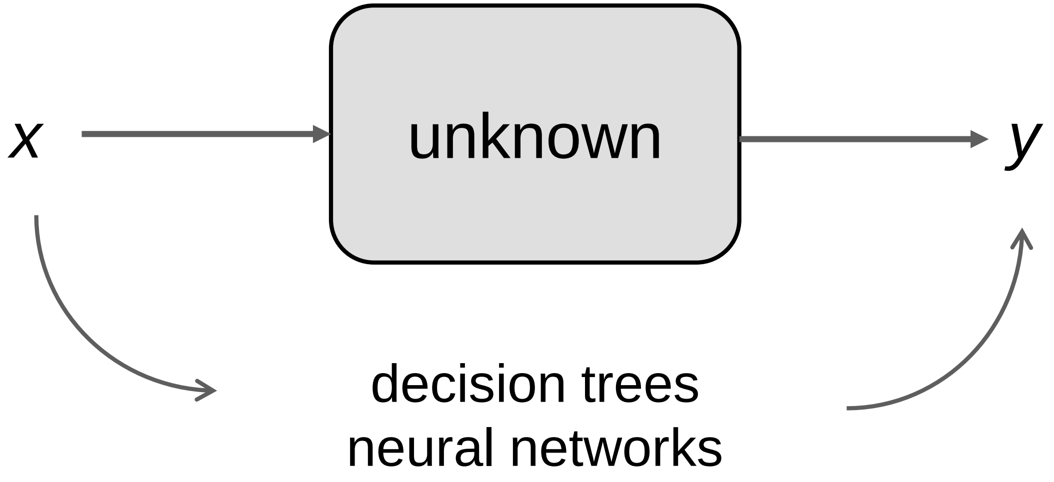
There are two approaches towards these goals

Data Modeling Culture



$\boldsymbol{y} = f(\boldsymbol{x}, \textit{parameters, random noise})$

Algorithmic Modeling Culture



$$f : \boldsymbol{x} \rightarrow \boldsymbol{y}$$

"If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools". That is, we need to use Machine Learning.

Leo Breiman, 2001

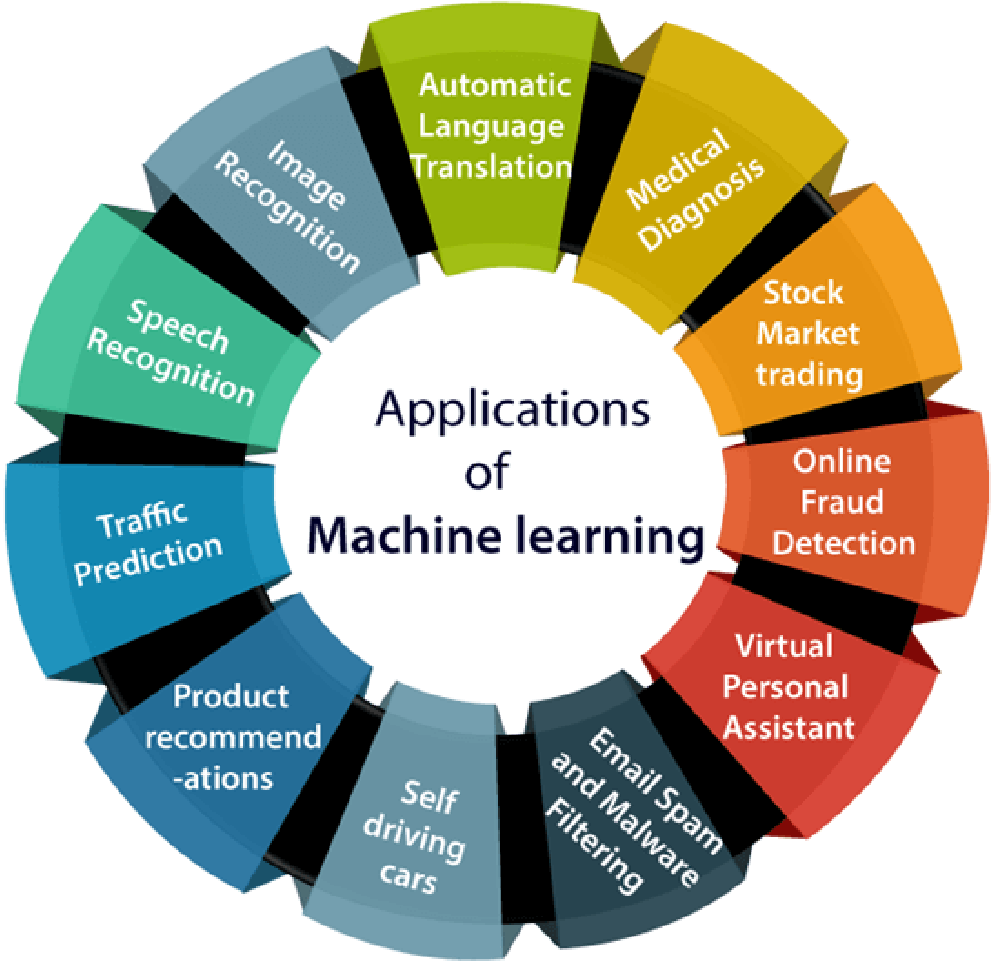
Major Machine Learning Techniques

- Regression / Estimation
  - Predicting continuous values
- Classification
  - Predicting the item class / category of a case
- Clustering
  - Finding the structure of data; summarization
- Associations
  - Associating frequent co-occurring items / events

Major Machine Learning Techniques

- Anomaly detection
  - Discovering abnormal and unusual cases
- Sequence mining
  - Predicting next events; click streams (Markov Models; HMM)
- Dimension reduction
  - Reducing the size of data (PCA)
- Recommendation systems
  - Recommending items

Applications of Machine Learning



Machine Learning

- Let’s look at a simple example of a machine learning model

Bank Loan Approval Scenario

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)?

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Bank Loan Approval Scenario

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

if Loan-Salary Ratio > 3 **then**
  Outcome='default'
Else
  Outcome='repay'
end if

- This is an example of a **prediction model**
- This is also an example of a **consistent** prediction model

Machine Learning

- Let’s now look at a relatively larger version of the Bank-Loan Dataset

Bank Loan (Extended) Dataset

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

Bank Loan (Extended) Dataset

```
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
```

How Does Machine Learning Work?

How does ML 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.

How does ML 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.

How does ML 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.
- However, because a training dataset is only a sample ML is an ill-posed problem.

A Simple Retail Scenario

ID	BBY	ALC	ORG	GRP
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

- Three binary descriptive features with yes/no entries
- Target feature (GRP) with 3 levels
- Data for 5 customers (instances)
- Objective:
  - to predict demographic group of each customer based on their shopping habits

A Simple Retail Scenario

Table: A full set of potential prediction models before any training data becomes available.

BBY	ALC	ORG	GRP	M_1	M_2	M_3	M_4	M_5	...	$M_{6\,561}$
no	no	no	?	couple	couple	single	couple	couple		couple
no	no	yes	?	single	couple	single	couple	couple		single
no	yes	no	?	family	family	single	single	single		family
no	yes	yes	?	single	single	single	single	single		couple
yes	no	no	?	couple	couple	family	family	family	...	family
yes	no	yes	?	couple	family	family	family	family		couple
yes	yes	no	?	single	family	family	family	family		single
yes	yes	yes	?	single	single	family	family	couple		family

2^3 = 8 combinations of feature values

3^8 = 6561 models consistent with the features

A Simple Retail Scenario

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

Notice that there is more than one candidate model left! It is because a single consistent model cannot be found based on a sample training dataset that ML is ill-posed.

Model Selection

- Consistency is akin to memorizing the dataset.
- Consistency with noise in the data isn't desirable.
- Goal: a model that **generalizes** 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

- Inductive bias is the set of assumptions that define the model selection criteria of an ML algorithm.
- There are two types of inductive bias that we can use:
  - restriction bias
  - preference bias
- Inductive bias is necessary for learning (beyond the dataset).

Summary of How ML Works

ML algorithms work by searching through sets of potential models.

There are two sources of information that guide this search:

- the training data
- the inductive bias of the algorithm.

THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

Inductive Bias Versus Sample Bias

Bias

- Inductive bias is necessary for machine learning, and in a sense, key goal of a data analyst is to find the correct inductive bias.
- Inductive bias is not the only type of bias that affects machine learning, a particularly important type of bias that we need to be aware of is **sampling bias**

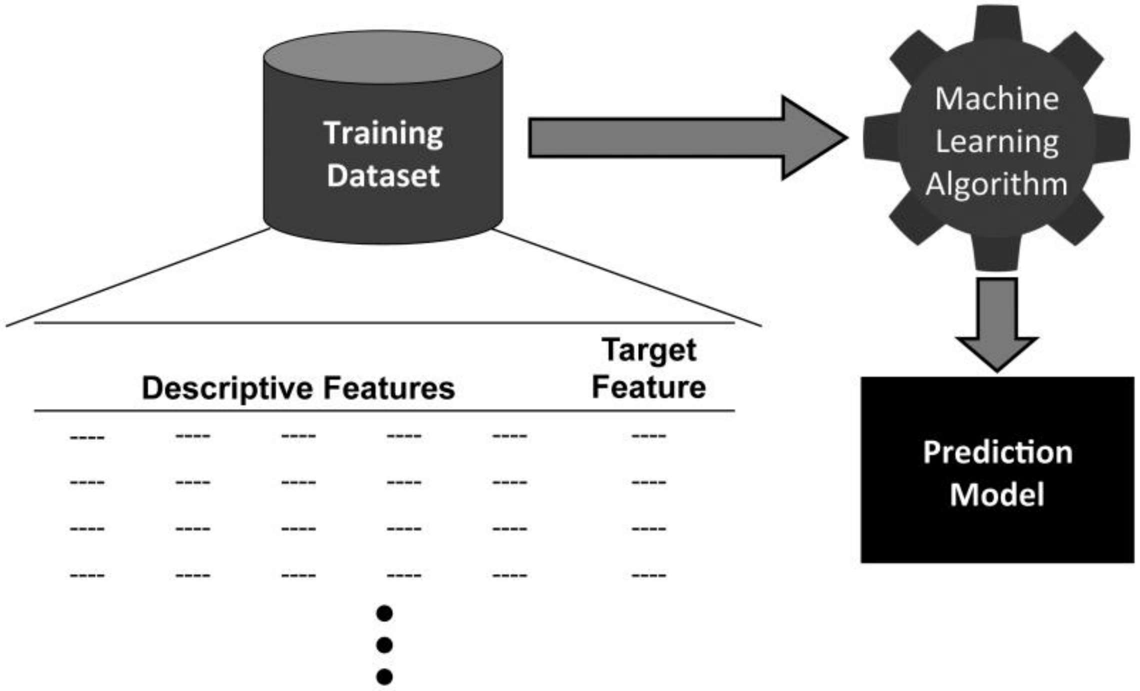
Bias

- Sampling bias arises when the sample of data used within a data-driven process is collected in such a way that the sample is not representative of the population the sample is used to represent.
- If a sample of data is not representative of a population, then inferences based on that sample will not generalize to the larger population.
- Sample bias is something that a data analyst should proactively work hard to remove from the data used in any data analytics or AI project.

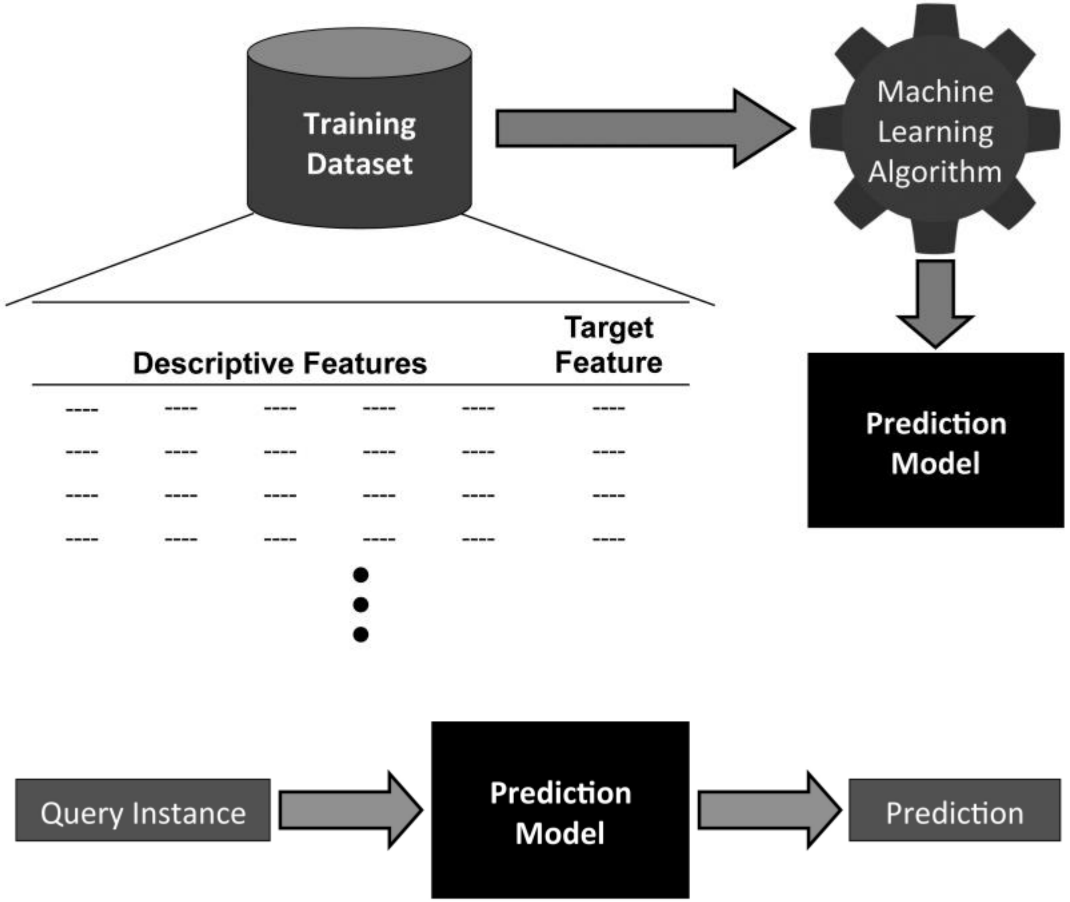
Supervised Machine Learning

- (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.
- The objective is to develop predictive models that can generalize well to new queries, i.e., unknown data that the models have not “seen” before during training
- It’s a two step process
  - Training
  - Prediction

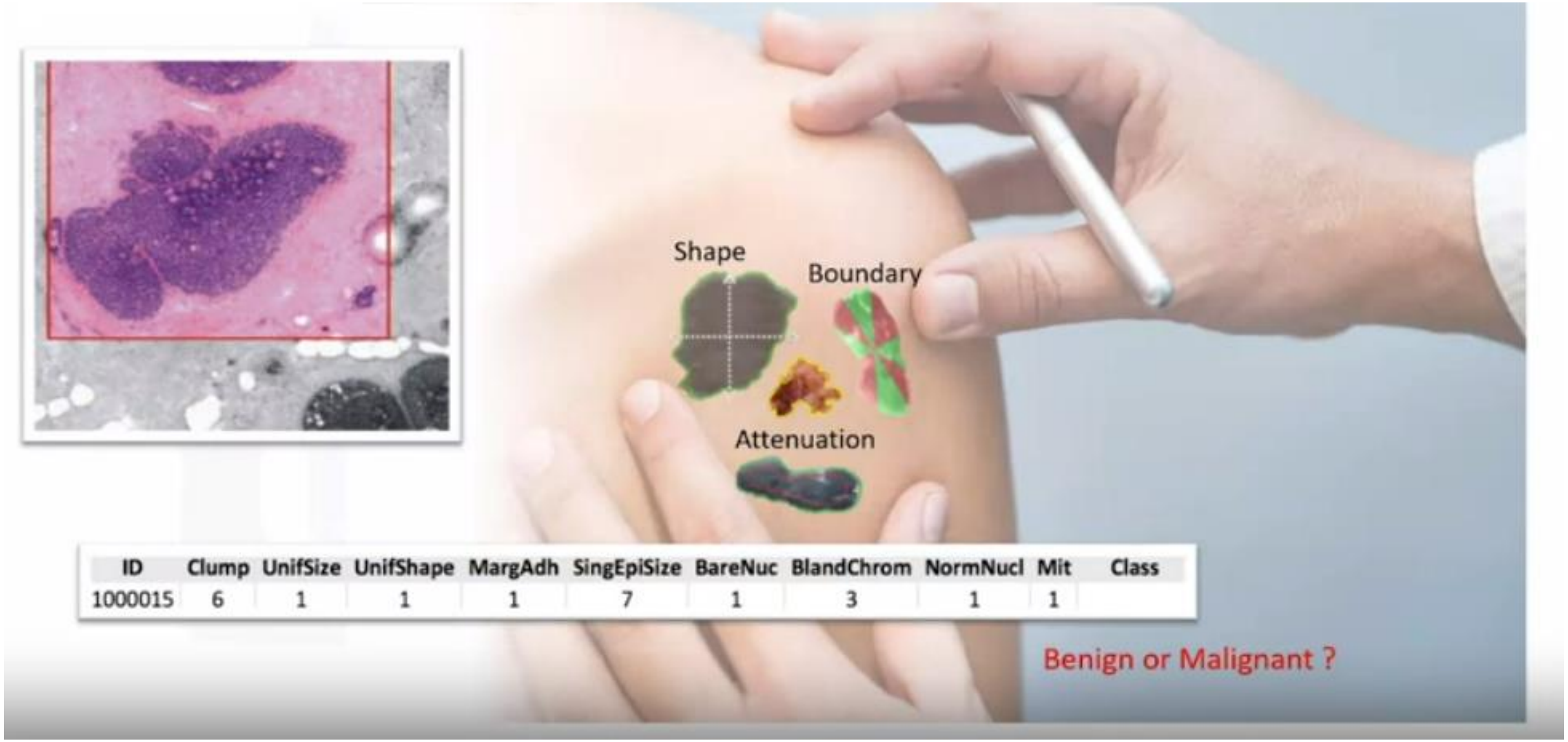
Step 1: Model Training



Step 2: Prediction

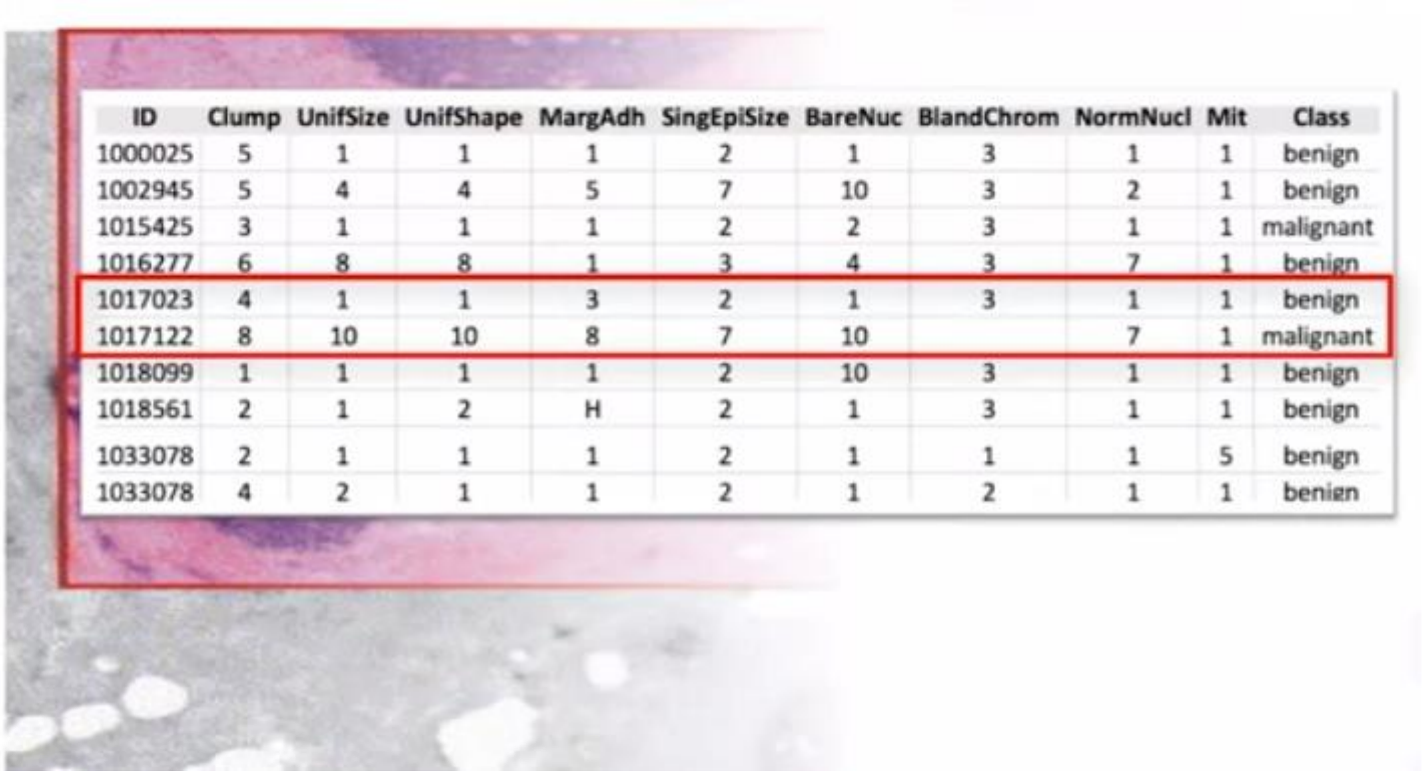


Is This a Benign or Malignant Cell?



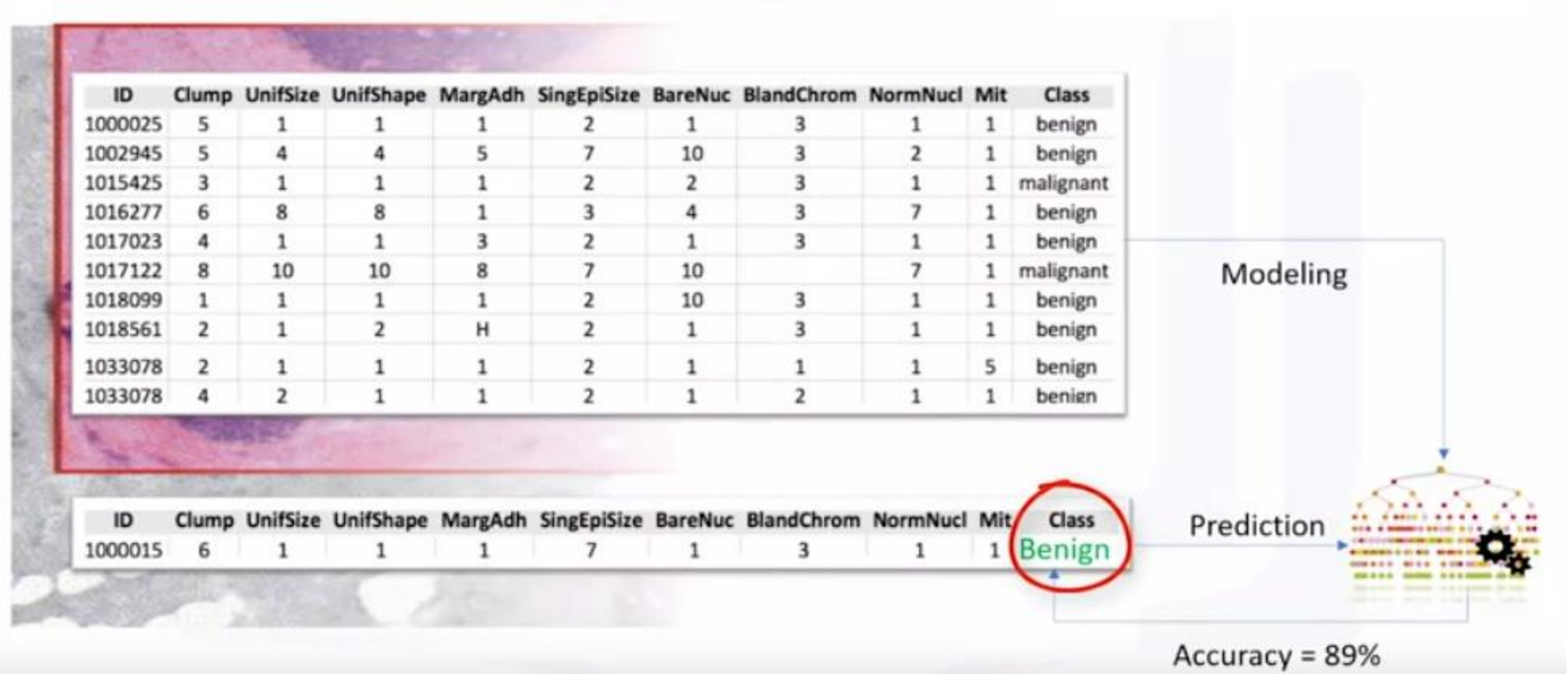
IBM Developer

Machine learning helps with predictions!



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Machine learning helps with predictions!

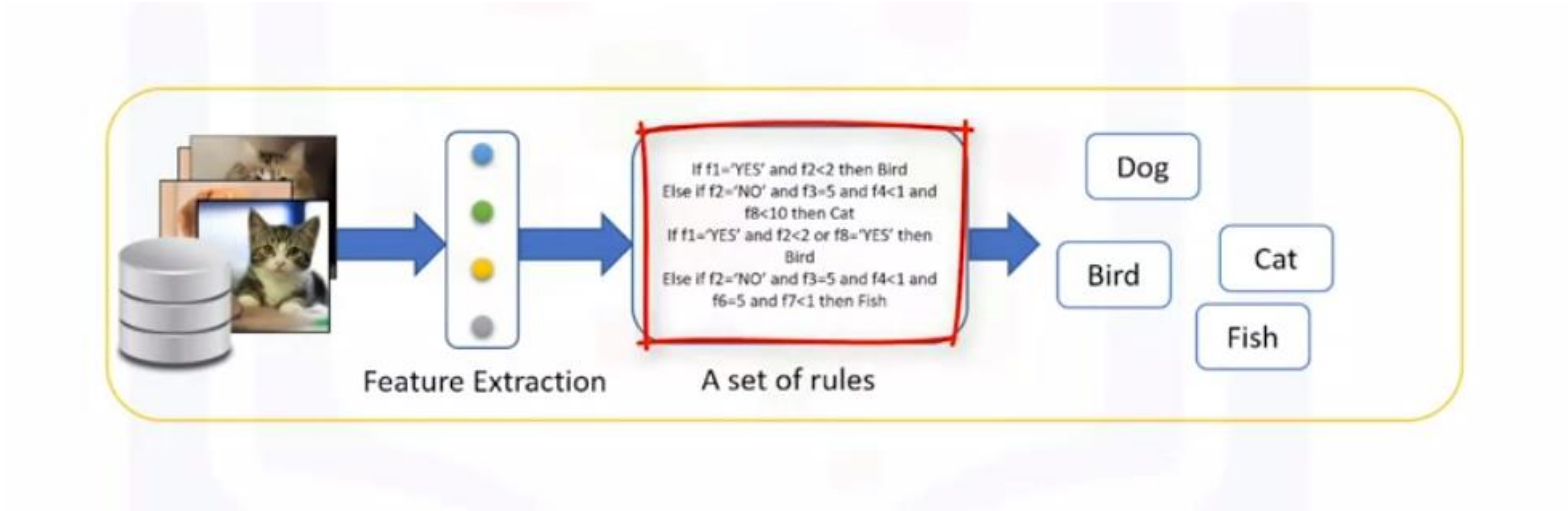


ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

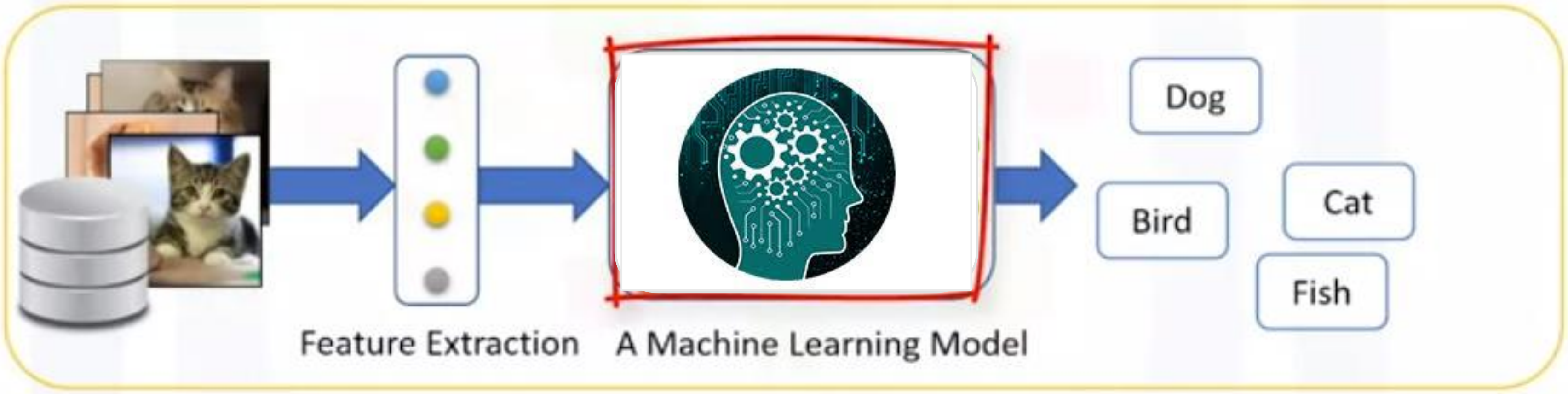
ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	Benign

IBM Developer

How Machine Learning Works



How Machine Learning Works



THE UNIVERSITY OF WINNIPEG | Professional, Applied and Continuing Education

What Can Go Wrong with Machine Learning?

What Can Go Wrong with ML?

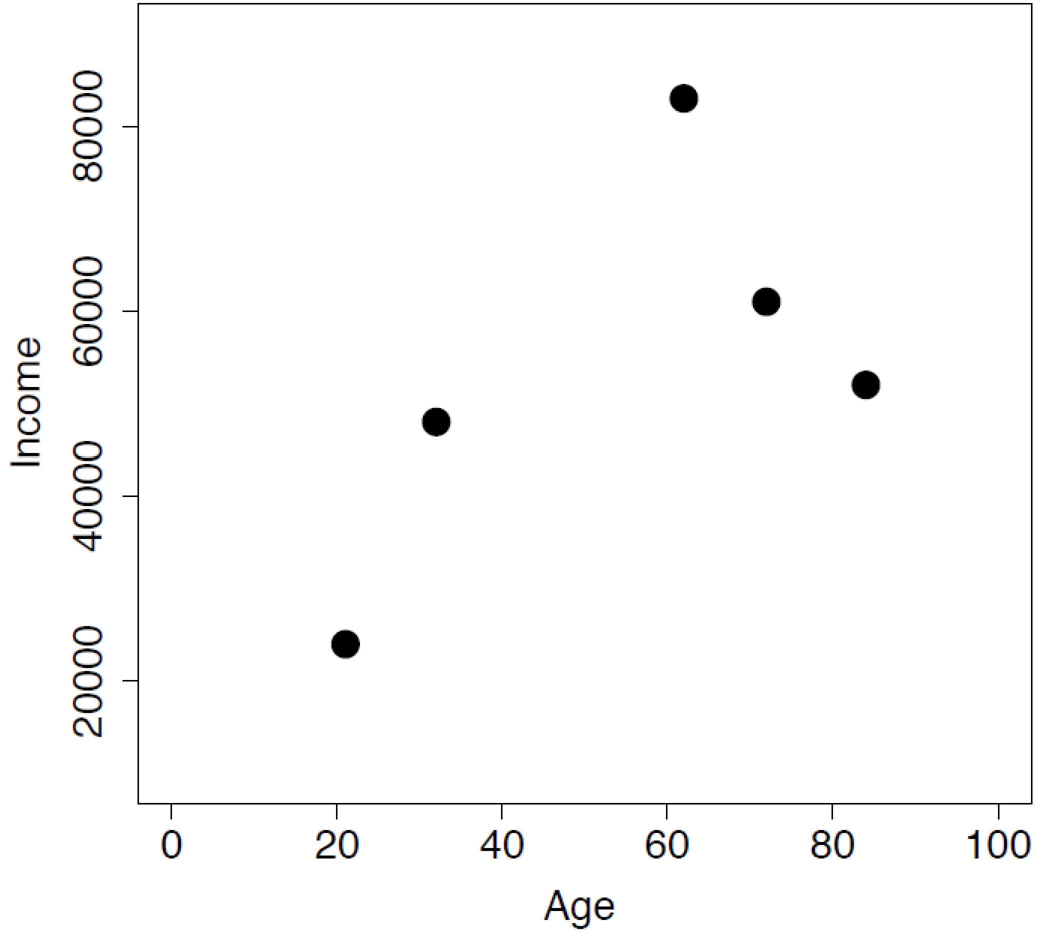
- No free lunch!
- What happens if we use wrong inductive bias?
  - Underfitting
  - Overfitting

Age-Income Dataset

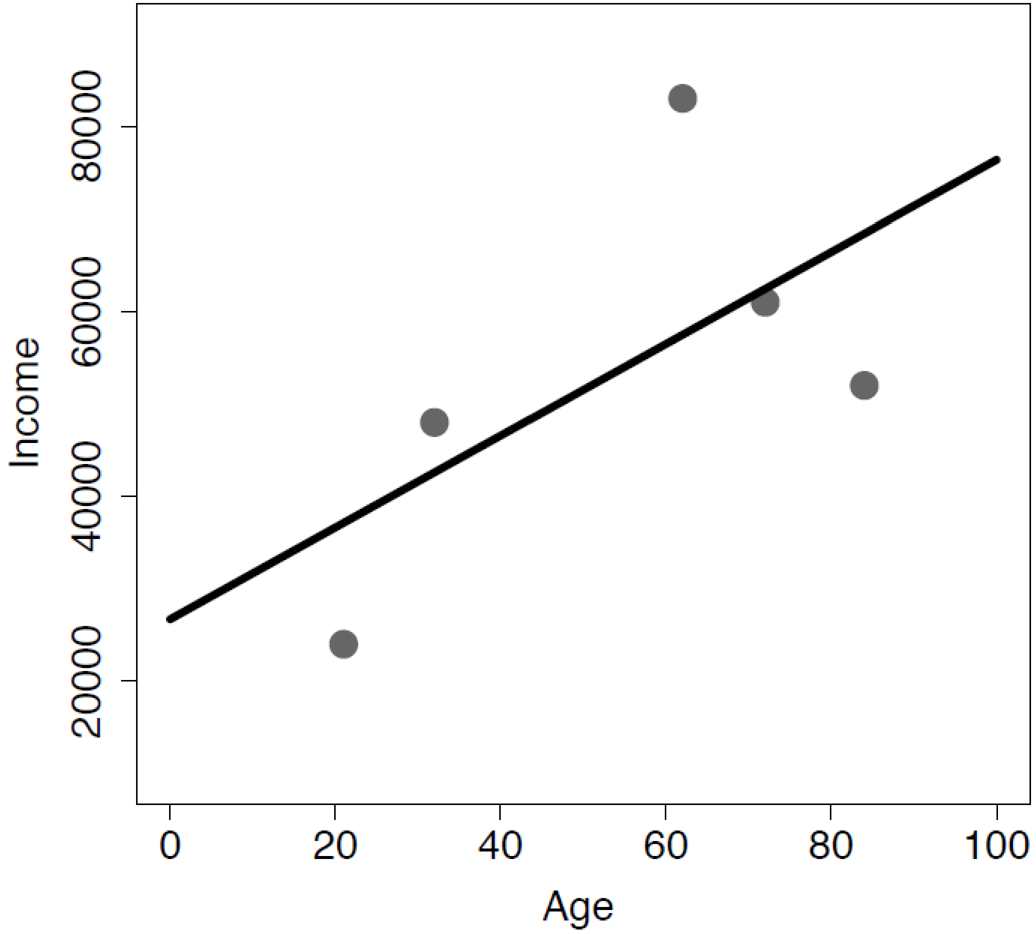
ID	Age	Income
1	21	24000
2	32	48000
3	62	83000
4	72	61000
5	84	52000

Objective: We would like to predict income based on the age of an individual

Scatter plot of the age-income dataset

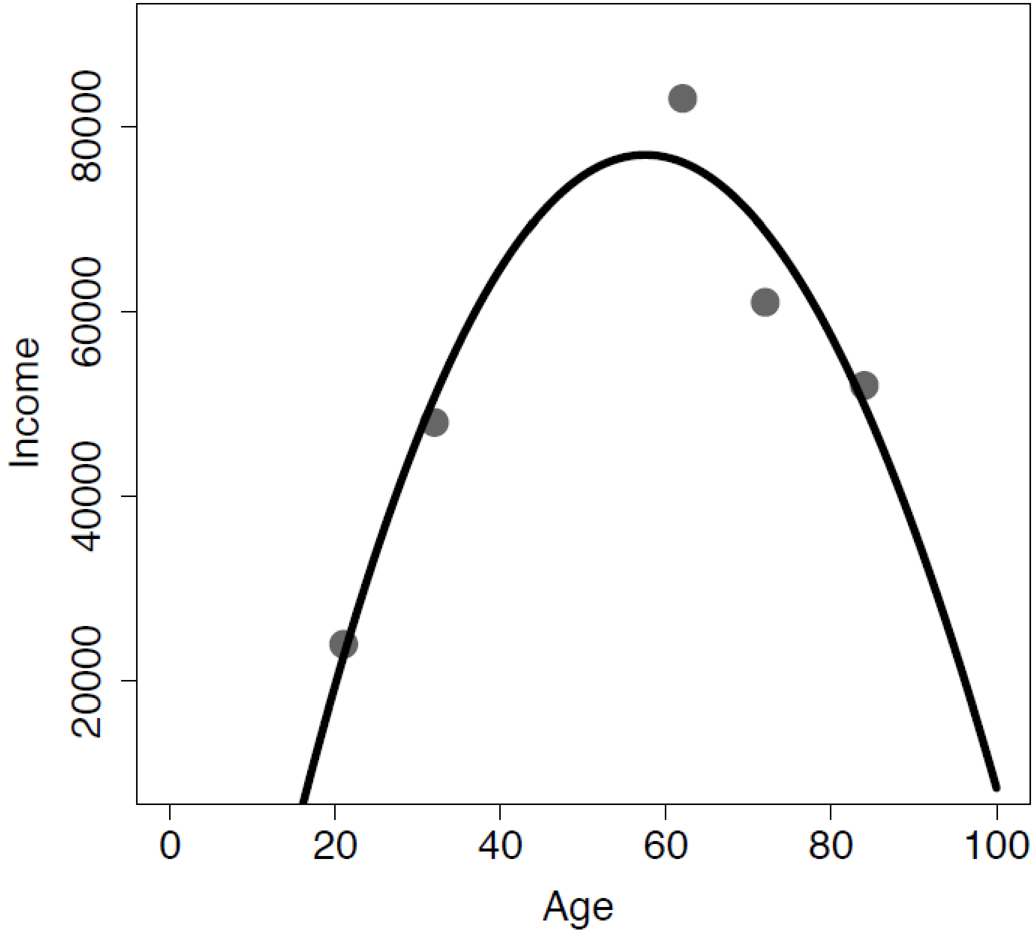


A Straight Line Fit

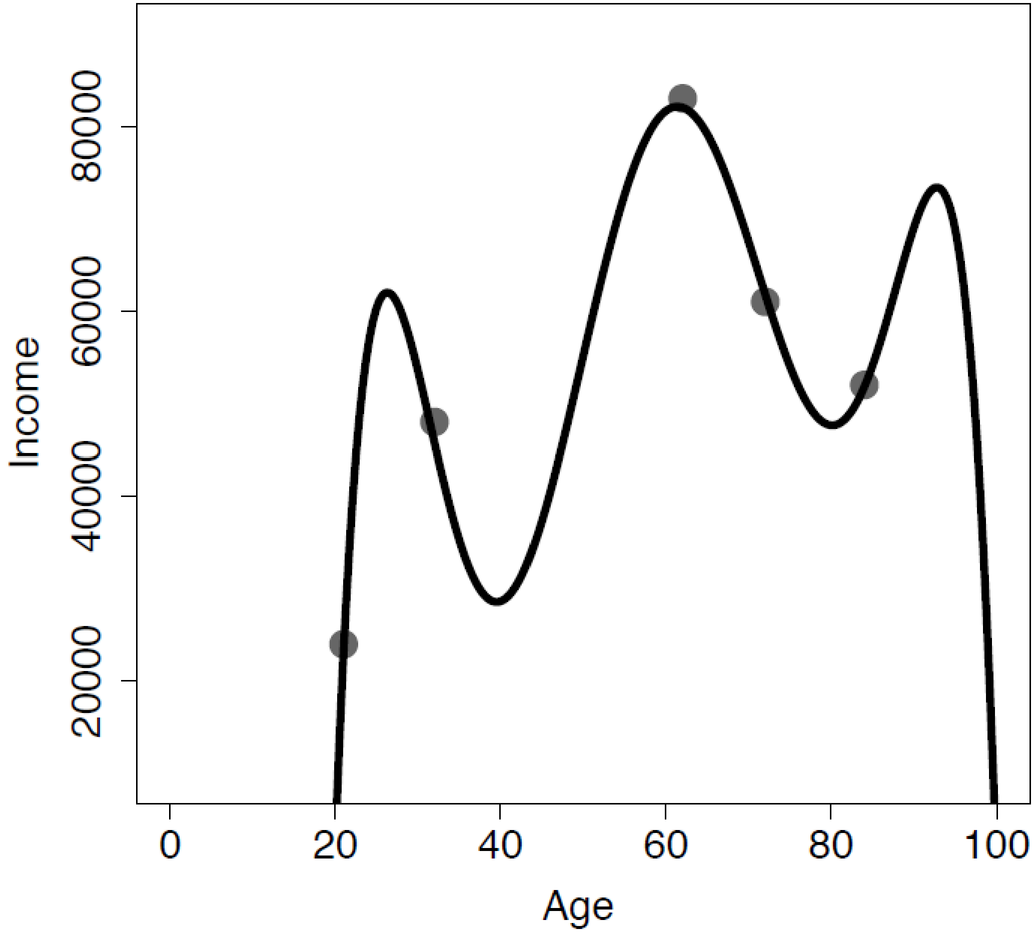


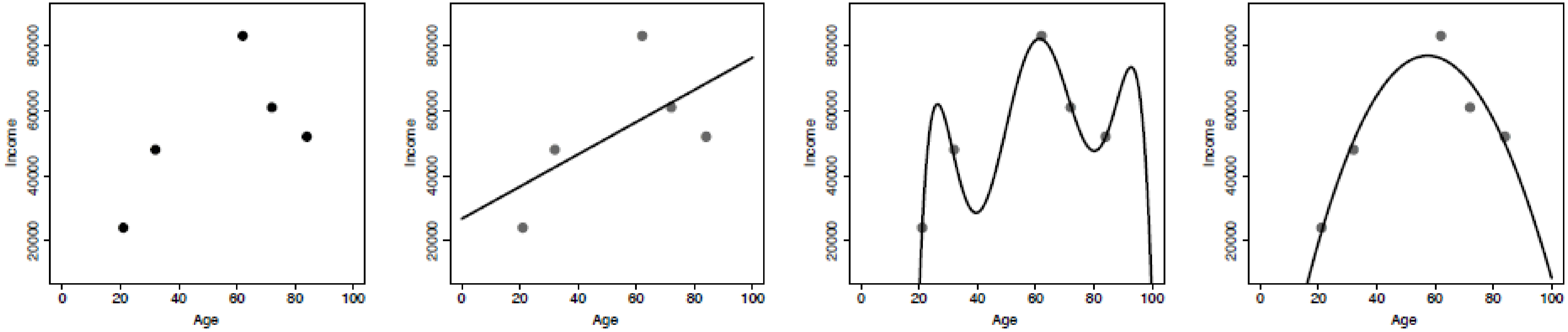
The University of Winnipeg | Professional, Applied and Continuing Education

2nd Order Polynomial Fit



Higher Order Polynomial Fit





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

Figure: Striking a balance between overfitting and underfitting when trying to predict age from income.

Preventing Model Overfitting

- Bigger datasets
- Regularization
- Validation dataset