Data Mining and Machine Learning

Comp 3027J

Catherine Mooney

catherine.mooney@ucd.ie

Introduction

Lecturer and Timetable

- **Lecturer:** Catherine Mooney (catherine.mooney@ucd.ie)
- **Lecture Times:** Tuesdays @ 13:30 – 15:05
- **Lab Times:** Thursdays @ 13:30 – 15:05 (no lab Week 1)

Lectures and Text

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- Lectures based on sections of the book
- Lecture slides are based on the official slides that accompany this book, with some additional information
- I will recommend which section of the book you should read each week

Environment and Assessment

- **Moodle:** I will make lecture materials available on Moodle after class
- Material will be covered in lecture notes and tutorial sessions
- **Assessment:**
  - **Final Exam:** 50% (Two hour written exam)
    Candidates must complete question 1 worth 40 marks and any two other questions worth 30 marks each
  - **Assignment:** 50% Two part assignment, based on labs (more details later)

Moodle Enrolment

- https://csmoodle.ucd.ie
  Log in with your UCD username and password
- Find COMP3027J
- Enrolment Key: BeijingDublin

Any questions so far?

A little bit about me

- PhD Computer Science, University College Dublin
  Protein structure prediction using bidirectional recurrent neural networks (BRNN)
- 2009–2013: Post-doctoral researcher in Clinical Bioinformatics – School of Medicine, UCD
- 2013–2014: Senior post-doctoral researcher – School of Physics, Dublin Institute of Technology
- 2014–2016: Research Fellow – Department of Physiology & Medical Physics, Royal College of Surgeons in Ireland
- Assistant Professor, University College Dublin since 2016

Research Interests

- The application of **machine learning** to solve problems in **biology** and **medicine** is my research area
- Machine Learning – mostly neural networks
- Computational Biology/Bioinformatics/Medical Data Science/Health Informatics
- **Google Scholar:** `https://scholar.google.com/citations?user=C8F_xyoAAAAJ&hl=en`

Google Scholar

Catherine Mooney

FOLLOW

Assistant Professor, School of Computer Science, University College Dublin
Verified email at ucd.ie

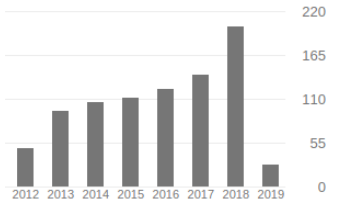
Machine learning medical data science computational biology bioinformatics biomarkers

GET MY OWN PROFILE

TITLE	CITED BY	YEAR
Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information G Pollastri, AJM Martin, C Mooney, A Vullo BMC bioinformatics 8 (1), 201	111	2007
Towards the improved discovery and design of functional peptides: common features of diverse classes permit generalized prediction of bioactivity C Mooney, NJ Haslam, G Pollastri, DC Shields PloS one 7 (10), e45012	100	2012
Distill: a suite of web servers for the prediction of one-, two-and three-dimensional structural features of proteins D Baú, AJM Martin, C Mooney, A Vullo, I Walsh, G Pollastri BMC bioinformatics 7 (1), 402	79	2006
Inhibition of dipeptidyl peptidase IV and xanthine oxidase by amino acids and dipeptides AB Nongonierma, C Mooney, DC Shields, RJ FitzGerald Food chemistry 141 (1), 644-653	64	2013

Cited by VIEW ALL

	All	Since 2014
Citations	994	711
h-index	18	17
i10-index	23	22



Co-authors

Example Research Project – Neonatal Seizure Prediction

- **Perinatal asphyxia:** interruption in blood flow and gas exchange to baby
- Some, but not all, of these babies will have seizures
- Can we predict which babies will have seizures?
- Why? So clinicians can decide as quickly as possible which babies to treat and how to treat them

Any questions?

Data mining and machine learning

- These days everyone seems to be talking about/doing data mining and machine learning
- “Easy” to do using R/Python/Weka...
- Very easy to do “wrong”!!

Data mining and machine learning

- The focus of this course is how to do data mining (DM) and machine learning (ML) “the right way”!
- What sort of questions can you answer with DM/ML?
- Understanding your data – when can you use DM/ML, when can you not use DM/ML?
- How to prepare your data correctly
- Which DM/ML algorithm to use
- How to train your DM/ML correctly
- How to evaluate your results
- The “art” machine learning!

Module Description

The objective of this module is to familiarise students with the fundamental **theoretical concepts** in data mining and machine learning, as well as to instruct students in the **practical aspects** of applying data mining and machine learning algorithms.

Key techniques in supervised machine learning will be covered, such as classification using decision trees and nearest neighbour algorithms, and regression analysis.

A particular emphasis will be placed on the **evaluation** of the performance of these algorithms. Further topics covered include data preparation and dimension reduction.

Learning Outcomes

On completion of this module, you will be able to:

- Distinguish between the different categories of data mining and machine learning algorithms
- Identify a suitable data mining/machine learning algorithm for a given application or task
- Run and evaluate the performance of a range of algorithms on real datasets using a standard machine learning toolkit

Feedback

Module Feedback for COMP3027J - Data Mining & Machine Learning

This should take approximately 3 minutes to complete. Click on the module code above to see details of this module. Your responses will remain anonymous and the results will not be made available to your lecturer or Head of School until after this semester's examination results have been issued.

Please complete all questions.

1	**I have a better understanding of the subject after completing this module.**	○Strongly Agree ○Agree ○Not Sure ○Disagree ○Strongly Disagree
2	**The assessments to date were relevant to the work of the module.**	○Strongly Agree ○Agree ○Not Sure ○Disagree ○Strongly Disagree
3	**I achieved the learning outcomes for this module.**	○Strongly Agree ○Agree ○Not Sure ○Disagree ○Strongly Disagree
4	**The teaching on this module supported my learning.**	○Strongly Agree ○Agree ○Not Sure ○Disagree ○Strongly Disagree
5	**Overall I am satisfied with this module.**	○Strongly Agree ○Agree ○Not Sure ○Disagree ○Strongly Disagree

Your comments are very important and valued by lecturers. Please ensure that neither the language nor content will cause personal offense to any individual lecturer.

6 **Identify up to three aspects of the module that most helped your learning**

7 **Suggest up to three changes to the module that would enhance your learning.**

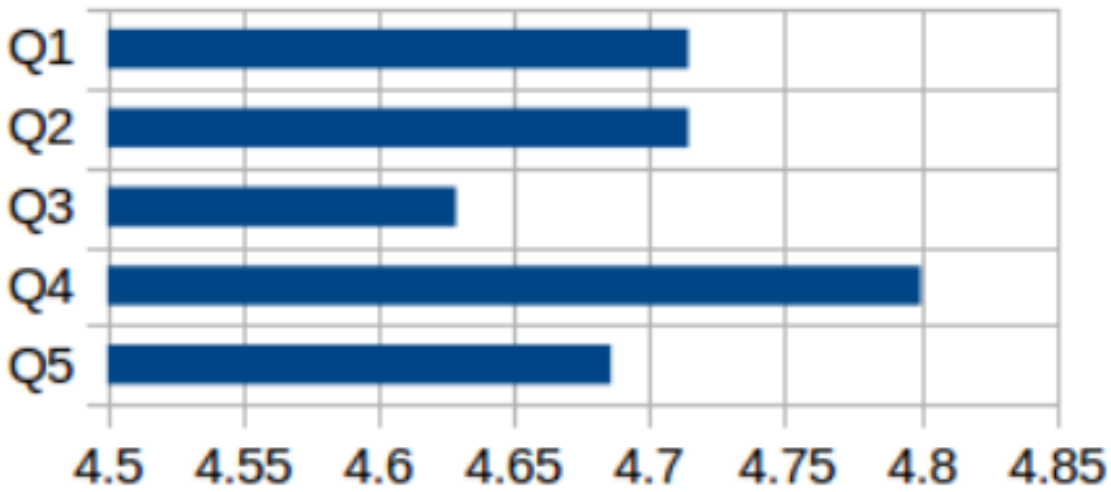
Thank you for your feedback. Click SUBMIT below if you are happy with your responses.

Feedback

Student Feedback for 2017/18 Semester 2

Click the module code below for detailed feedback about that module.

Q1 I have a better understanding of the subject after completing this module
Q2 The assessment was relevant to the work of the module.
Q3 I achieved the learning outcomes for this module
Q4 The teaching on this module supported my learning
Q5 Overall I am satisfied with this module



Learning Outcomes

On completion of this module, you will be able to:

- Distinguish between the different categories of data mining and machine learning algorithms
- Identify a suitable data mining/machine learning algorithm for a given application or task
- Run and evaluate the performance of a range of algorithms on real datasets using a standard machine learning toolkit

Module Feedback

- Identify up to three aspects of the module that most helped your learning
- Suggest up to three changes to the module that would enhance your learning

Module Feedback

- Labs too long/not long enough
- Too much math/not enough math
- Teaching building 3 – slow internet
- Assignment 1 – Grading not clear (see Module_grade_descriptors.pdf)

Any questions?

What is Data Mining/Machine Learning?

What is Data Mining/Machine Learning?

- The subfield of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).
- **Unsupervised Learning:** An algorithm that finds patterns in data when no manually labelled examples are available as inputs. More focused on data exploration and knowledge discovery. e.g. Clustering, Graph partitioning algorithms
- **Supervised Learning:** An algorithm that learns a function from examples of its inputs and outputs. It requires manually-labelled example data to learn the correct answer for a given query input. e.g. Classification, Regression algorithms

https://en.wikipedia.org/wiki/Machine_learning

What is the difference between Data Mining and Machine Learning?

What's the difference between data mining and machine learning?

- "The short answer is: None. They are ... concerned with the same question: how do we learn from data?"

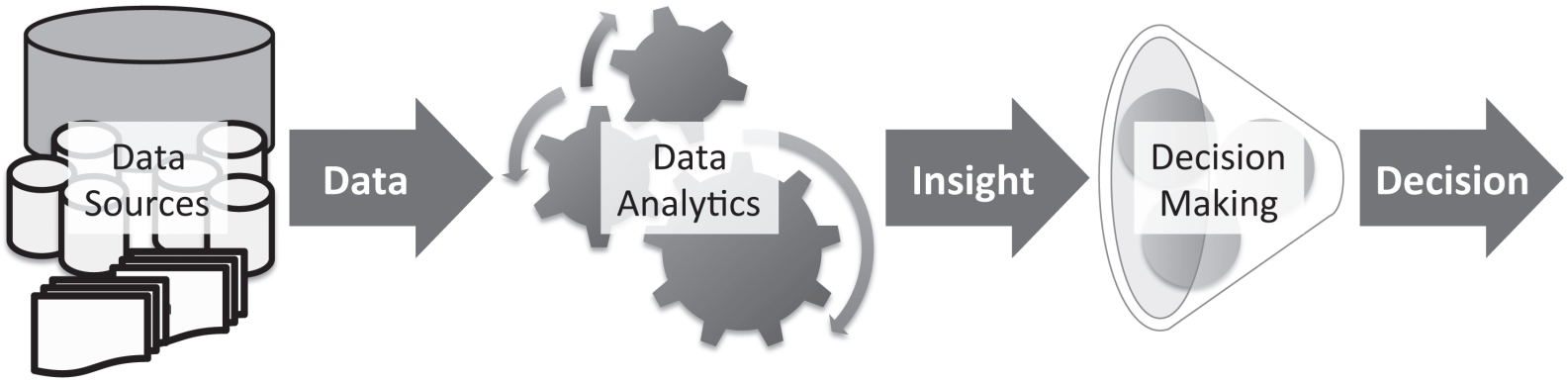
  Larry Wasserman, Professor in Statistics and Machine Learning, Carnegie Mellon
- They cover almost exactly the same material and use almost exactly the same techniques (take a look at the table of contents of some of the most popular text books...)
- They *emphasize* different things
  – The purpose of data mining is finding valuable insights in large databases
  – Machine learning is more focused on making accurate predictions

`www.r-bloggers.com/whats-the-difference-between-machine-learning-statistics-and-data-mining/`

Example Applications:

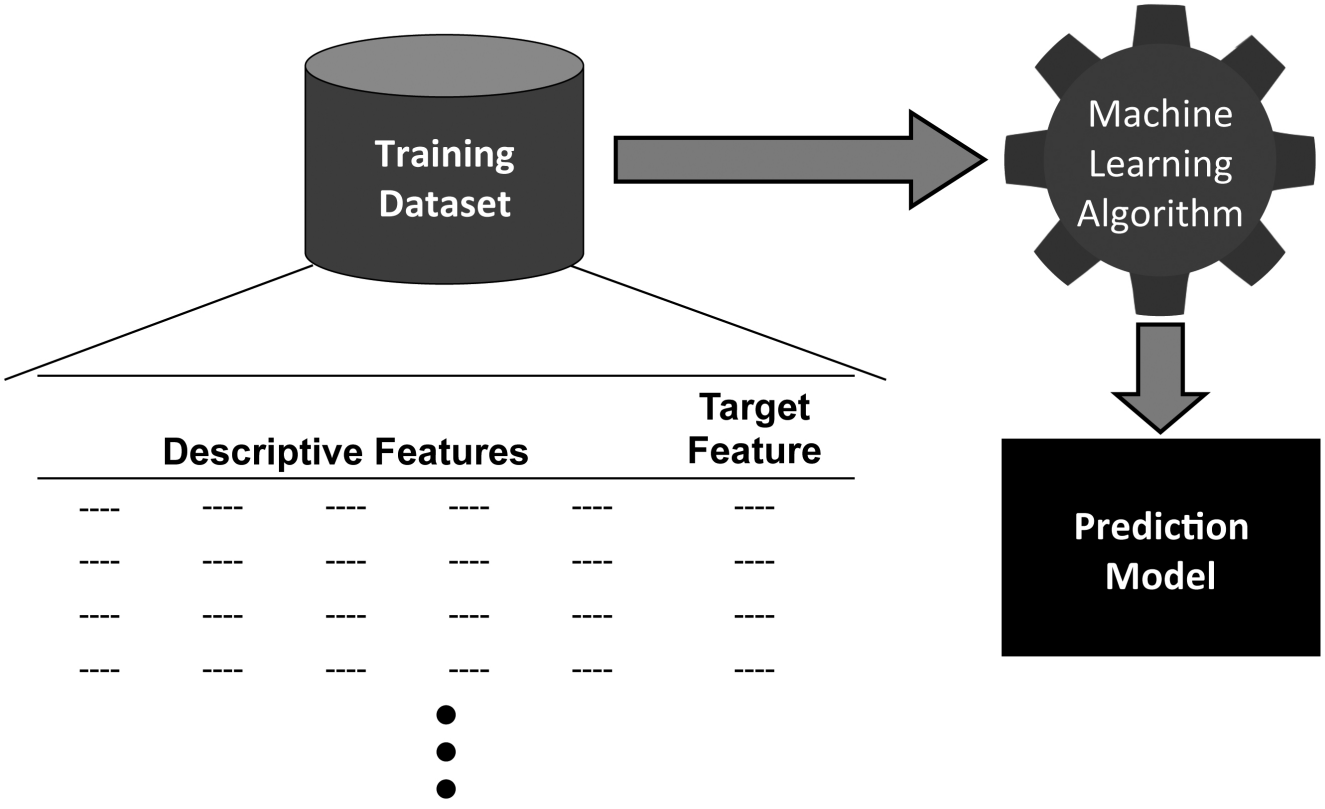
- Data Security – cybersecurity, Malware
- Personal Security – screenings at airports, stadiums, concerts
- Financial Trading – predict the stock markets
- Healthcare – improve medical outcomes, diagnosis
- Fraud Detection – ecommerce fraud prevention, money laundering
- Smart Cars – "A smart car would not only integrate into the Internet of Things, but also learn about its owner and its environment"

http://www.informationweek.com/strategic-cio/executive-insights-and-innovation/11-cool-ways-to-use-machine-learning/d/d-id/1323375
http://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai/#56fcf34e10cf



What is Data Mining/Machine Learning?

Moving from **data** to **insights** to **decisions**



Step 1

Supervised machine learning techniques automatically learn a **model** of the relationship between a set of **descriptive features** and a **target feature** based on a set of historical examples, or **instances**

Step 2

We can then use this **model** to make predictions for new instances

Any questions?

A very simple example using mortgages that a bank has granted in the past

- **Descriptive features** (OCCUPATION, AGE, LOAN-SALARY RATIO)
- **Target feature** defaulted or paid back in full (OUTCOME)
- Each row is referred to as a **training instance**
- The dataset is referred to as a **training dataset**.

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

- 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 very simple **prediction model**

- 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 very simple **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)

- 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 very simple **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)
- **feature design** and **feature selection** are two important topics that we will cover in detail

- What is the relationship between the **descriptive features** and the **target feature** (OUTCOME) in the following 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	1.20	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	2.10	37	professional	house	ftb	repaid
19	213,300	61,100	3.49	21	industrial	apartment	ftb	default
20	284,100	32,300	8.80	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

ftb – first-time buyer; stb – second-time buyer.

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

- To manually learn this model by examining the data is almost impossible.
- For a machine learning algorithm, however, this is simple.
- When we want to build prediction models from large datasets with multiple features, machine learning is the solution.

Any questions? Break for 5 mins?

How Does It 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.

- 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.

- 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 of the data, machine learning is called an ill-posed problem.
- An ill-posed problem is a problem for which a unique solution cannot be determined using only the information that is available.

Another very simple example...

- A supermarket chain wants to be able to classify customer households into the demographic groups: single, couple, or family, based solely on their shopping habits.
- The table shows 3 **descriptive features** describing the shopping habits of 5 customers (**training instances**) – did they buy baby food, alcohol, or organic products? The **target feature** is GROUP (single, couple, or family).

ID	BABY FOOD	ALCOHOL	ORGANIC	GROUP
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

The full set of potential prediction models before any training data becomes available

- There are three binary **descriptive features**, so there are $2^3 = 8$ possible combinations
- For each of these 8 possible combinations, there are 3 possible **target features** i.e. there are $3^8 = 6,561$ possible prediction **models**!!

Baby Food	Alcohol	Organic	Group	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
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

A sample of the models that are consistent with the training data

- Blanked out columns in the table indicate the **models** that are not consistent with the **training data**.
- 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.

BABY FOOD	ALCOHOL	ORGANIC	GROUP	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	...	$\mathbb{M}_{6\,561}$
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

So what criteria should we use for choosing between models?

- What if a new customer starts shopping at the supermarket and buys baby food, alcohol, and organic vegetables?
- In the previous example there are 3 **models** that are consistent with our training data:
  - $\mathbb{M}_2$ will return GRP = single
  - $\mathbb{M}_4$ will return GRP = family
  - $\mathbb{M}_5$ will return GRP = couple
- No learning is taking place here because the set of consistent models tells us nothing about the underlying relationship between the **descriptive features** and **target features** beyond what a simple look-up of the **training dataset** would provide.
- Consistency $\approx$ memorizing the dataset.

- If a **predictive model** is to be useful, it must be able to make predictions for queries that are not present in the data.
- A prediction model that makes the correct predictions for these queries captures the underlying relationship between the descriptive and target features and is said to **generalise** well.
- The goal of machine learning is to find the predictive model that **generalises** best.
- To find the best model, a machine learning algorithm must use some criteria for choosing among the candidate models it considers during its search.

What criteria should we use?

- Lots of different machine learning algorithms.
- Each machine learning algorithm uses different model selection criteria to drive its search for the best **predictive model**.
- The set of assumptions that defines the model selection criteria of a machine learning algorithm is known as the **inductive bias** of the machine learning algorithm.

- It has been shown that there is no particular **inductive bias** that on average is the best one to use.
- In general, there is no way of knowing for a given predictive task which inductive bias will work best.
- **The ability to select the appropriate machine learning algorithm (and hence inductive bias) to use for a given predictive task is one of the core skills that a data analyst must develop!!**

How machine learning works (Summary)

- Machine learning 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

- There are many different types of machine learning algorithms.
- In this course we will cover thee of the four families of machine learning algorithms that are in your book:
  1. **Information based learning (Chapter 4)**
  2. **Similarity based learning (Chapter 5)**
  3. **Probability based learning (Chapter 6 – no)**
  4. **Error based learning (Chapter 7)**

Any questions?

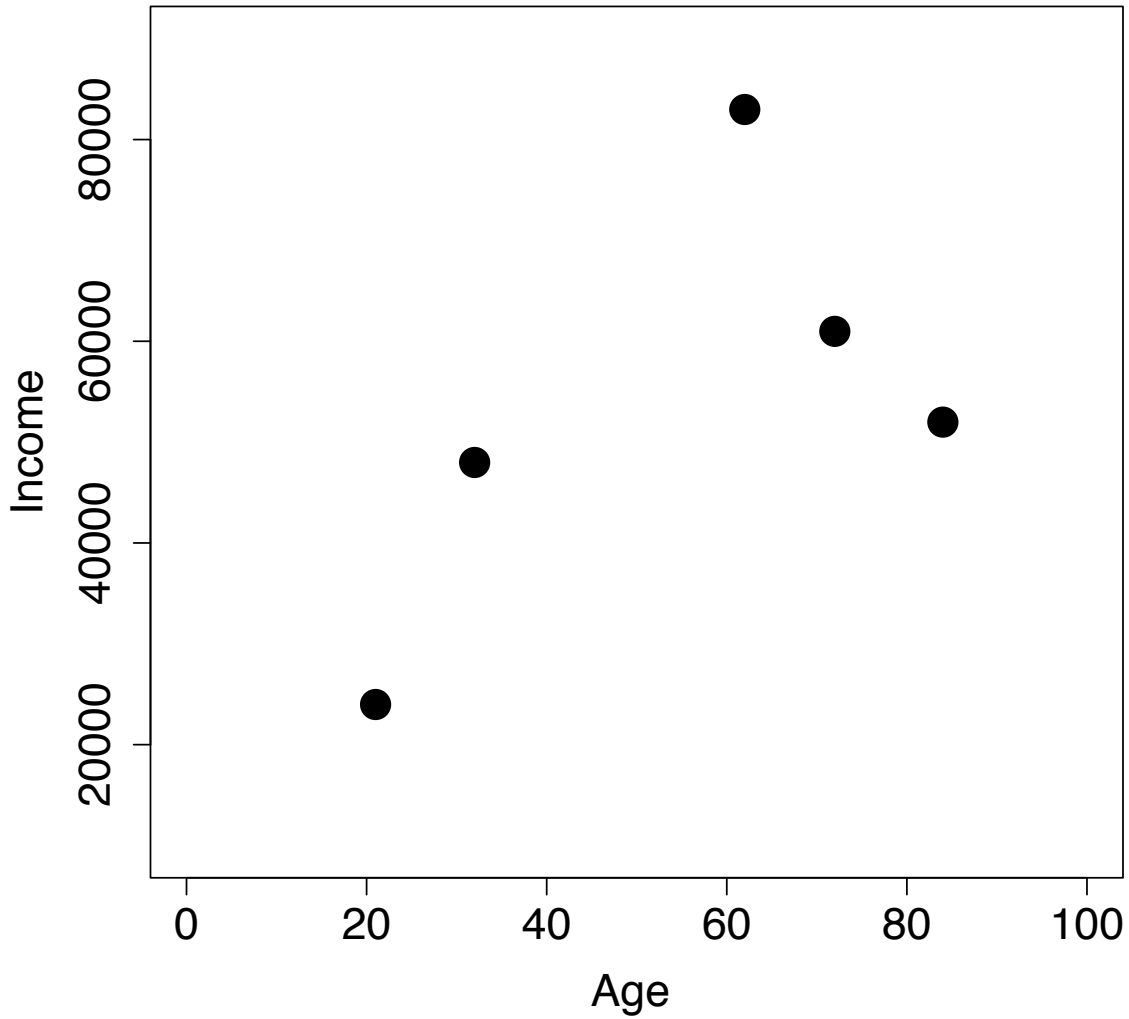
What Can Go Wrong?

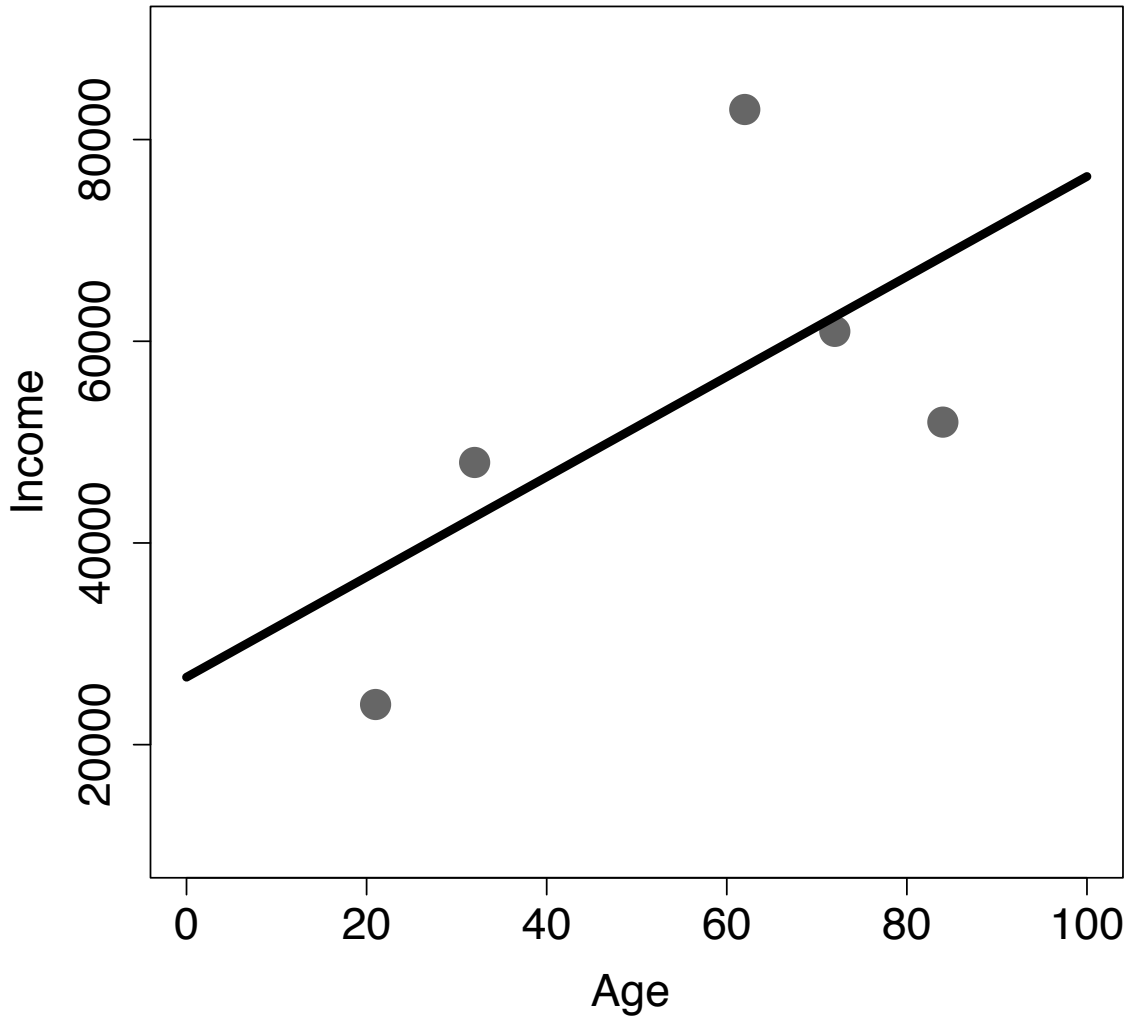
What happens if we choose the wrong inductive bias:

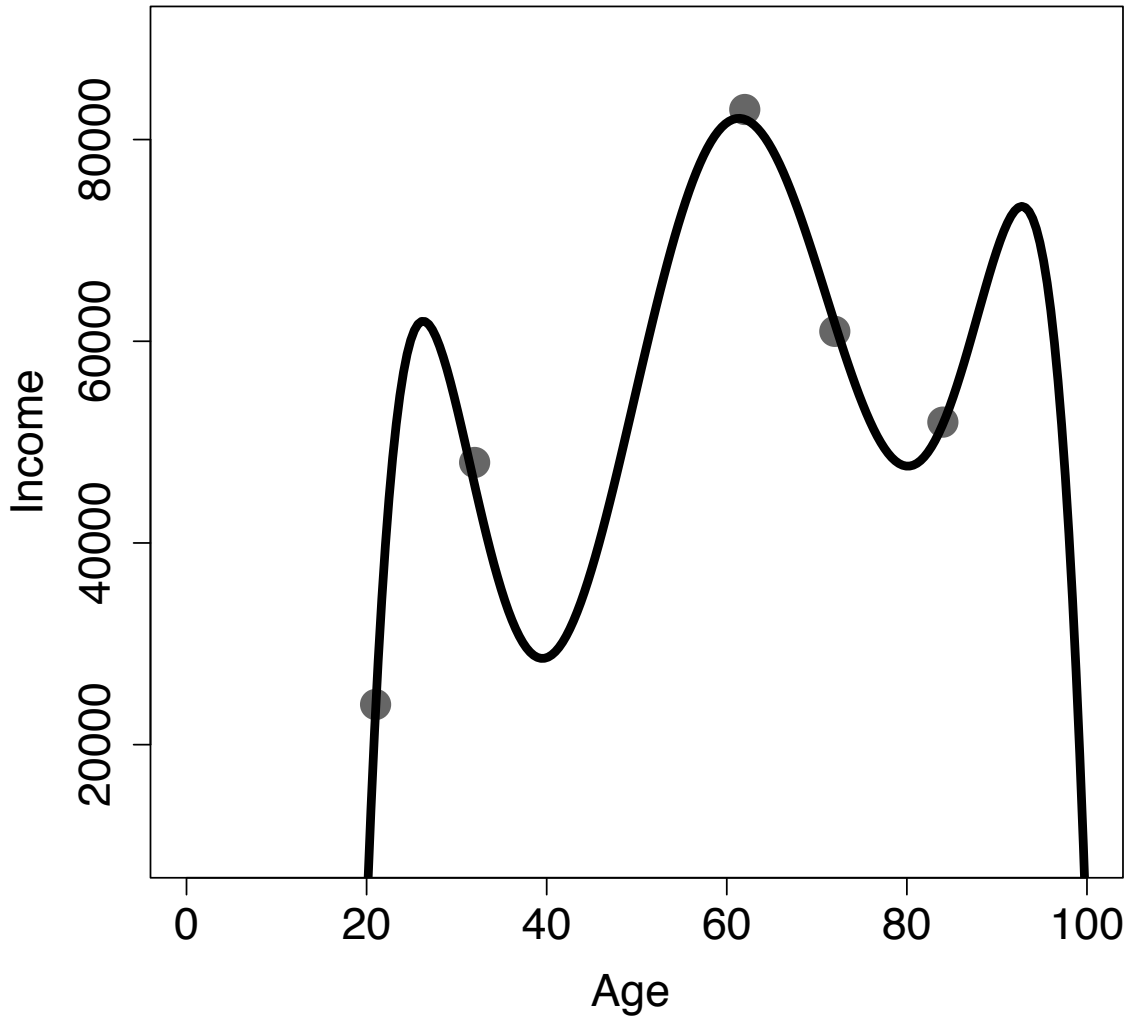
1. **underfitting**
2. **overfitting**

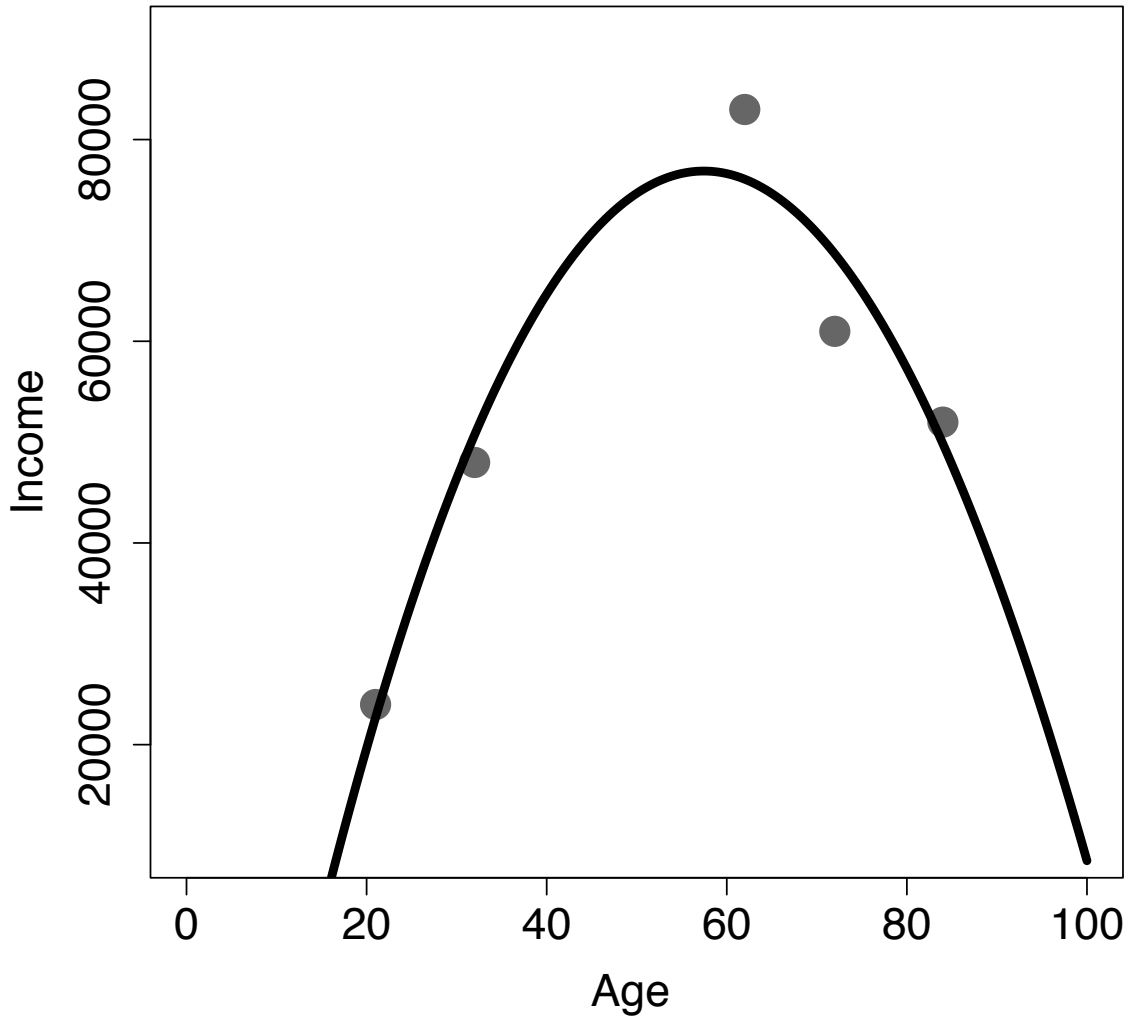
Another very simple example – The age-income dataset.

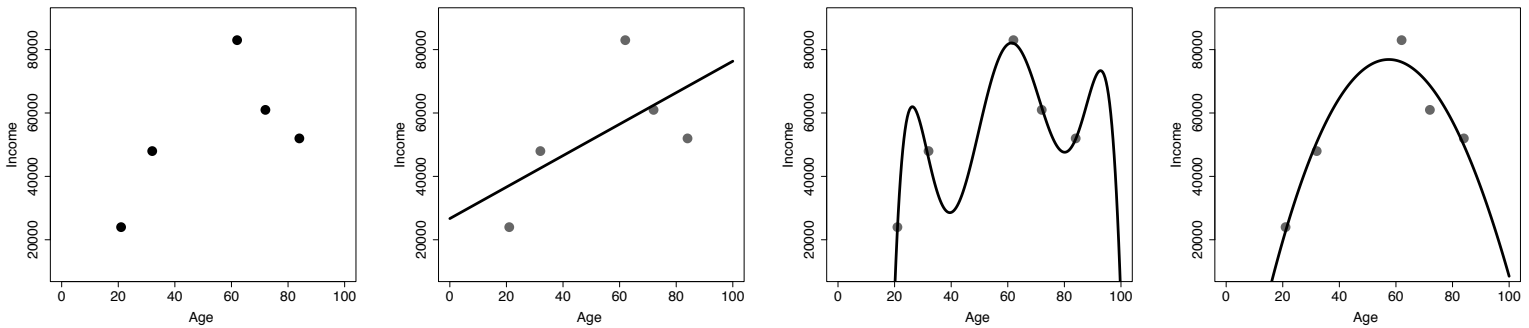
ID	AGE	INCOME
1	21	24,000
2	32	48,000
3	62	83,000
4	72	61,000
5	84	52,000


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


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


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


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



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

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

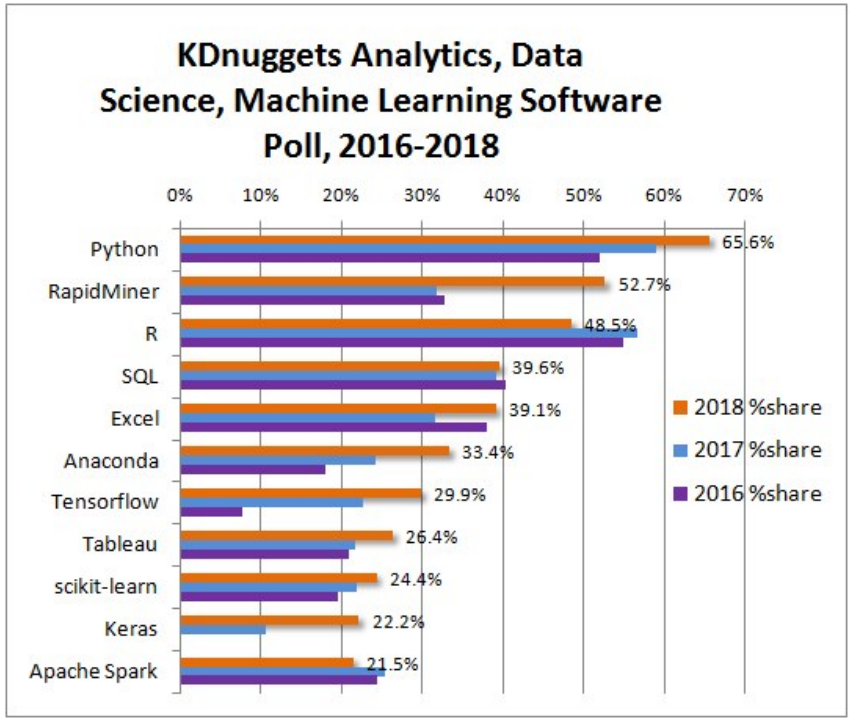
Summary

- Machine Learning techniques automatically learn (**model**) the relationship between a set of **descriptive features** and a **target feature** from a set of historical examples (**instances**).
- Machine Learning is an **ill-posed** problem:
  i.e. there is not only one solution to any problem
  1. **generalize**,
  2. **inductive bias**,
  3. **underfitting**,
  4. **overfitting**.
- Striking the right balance between **model** complexity and simplicity (between underfitting and overfitting) is the hardest part of machine learning.

Any questions?

Labs...

We will be using R and RStudio during the labs.



The 19th annual KDnuggets Software Poll (2018)

- What software you used for Analytics, Data Mining, Data Science, Machine Learning projects in the past 12 months?
- Over 2,300 voters – analytics and data science community and vendors

https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html

What is R?

- R is a language and environment for statistical computing and graphics.
- R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible (via packages).
- `https://www.r-project.org/about.html`
- RStudio `https://www.rstudio.com/`

A brief introduction to R

- How to install R and a Brief Introduction to R (Avril Coghlan, old but good) `http://a-little-book-of-r-for-bioinformatics.readthedocs.io/en/latest/src/installr.html`
- An Introduction to R `https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`
- R/RStudio Tutorial For Beginners (First 30 mins or so... 1.25 long in total) `https://www.youtube.com/watch?v=qEJHYIa-EhI`

A quick introduction to machine learning in R with caret

- http://topepo.github.io/caret/index.html
- https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf

Plagiarism

- Plagiarism is a serious academic offence
- Our staff and demonstrators are proactive in looking for possible plagiarism in all submitted work
- Suspected plagiarism is reported to the CS Plagiarism subcommittee for investigation
- Usually includes an interview with student(s) involved
  1st offence: usually 0 or NG in the affected components
  2nd offence: referred to the University disciplinary committee
  Student who enables plagiarism is equally responsible
- `http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf`
  `http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf`
  `http://libguides.ucd.ie/academicintegrity`

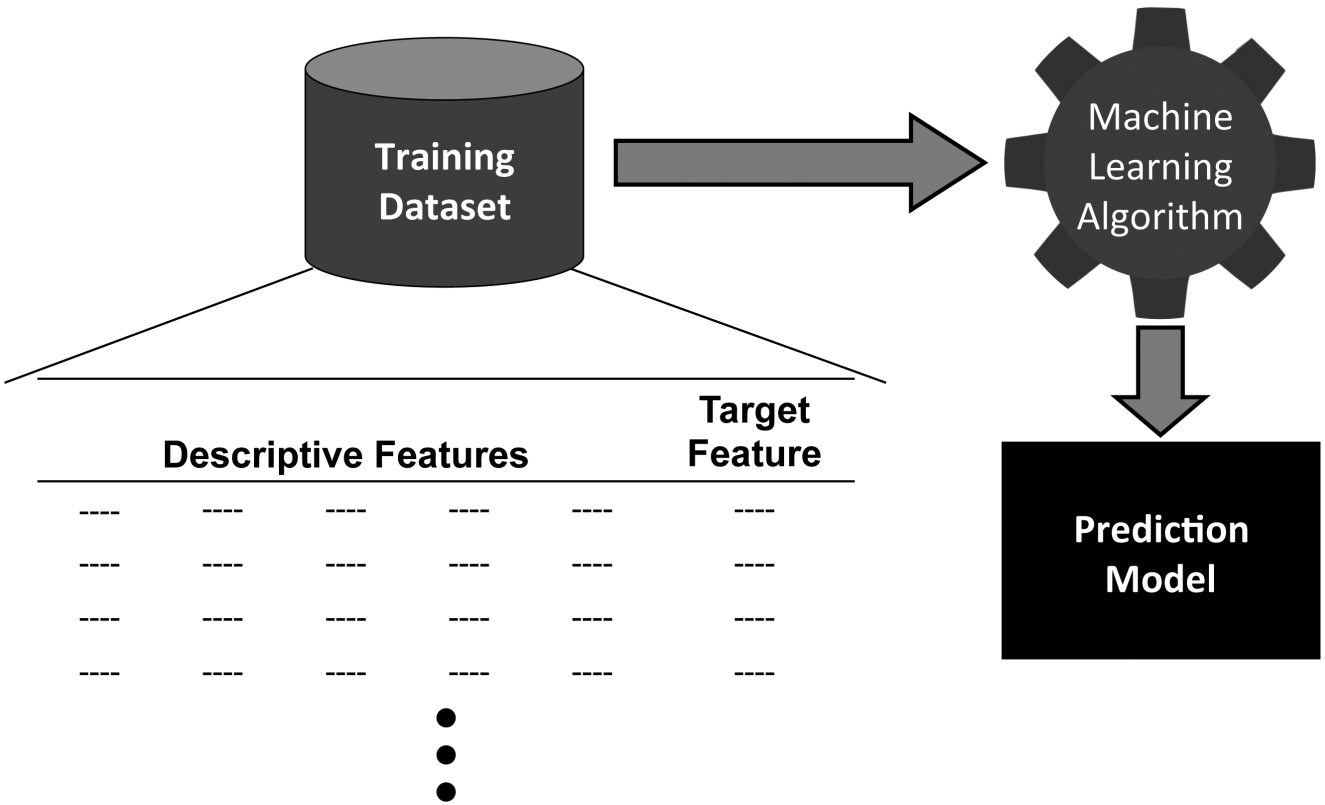
Recommended Reading

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- This week we covered Chapter 1, sections 1.1 – 1.4 in class
- I would suggest that you would read over these sections again
- Email me if you have any questions and I will cover them at the beginning of class next week
- Next week we will cover Chapter 2

Any questions?

- What is supervised machine learning?

- What is supervised machine learning?
- Supervised machine learning techniques automatically learn the relationship between a set of **descriptive features** and a **target feature** from a set of historical **instances** to build a **prediction model**.



Thanks for listening... see you next week!!