Féidearthachtaí as Cuimse
Infinite Possibilities

# Machine Learning

## Lecture 1: Introduction

Bojan Božić & Bujar Raufi
School of Computer Science
TU Dublin, Grangegorman

bojan.bozic@tudublin.ie; bujar.raufi@tudublin.ie

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Slides adapted from Sarah Jane Delany and book slides from: Fundamentals of Machine Learning for Predictive Data Analytics. Kelleher, Mac Namee and D'Arcy

# Overview

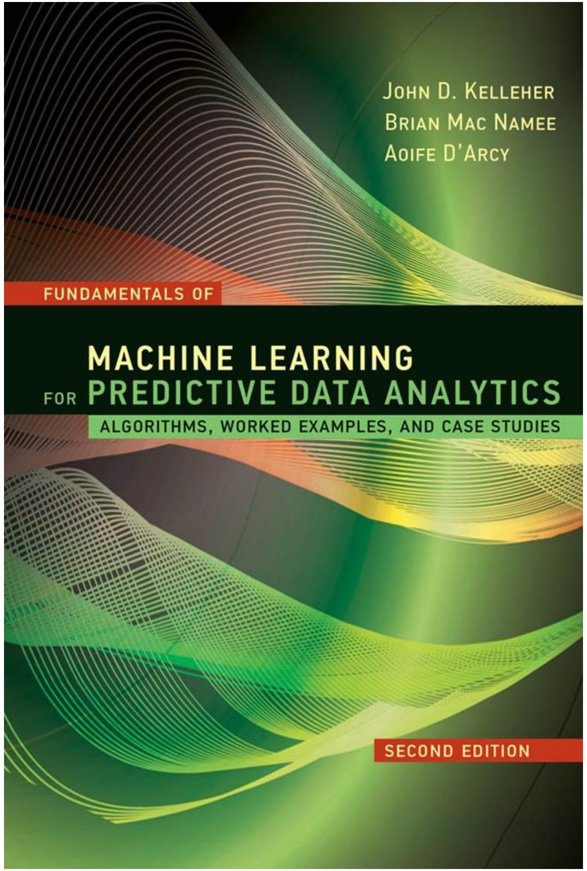
- Administrivia
- Module Outline
- Machine Learning Today - the good, the bad and the ugly
- Machine Learning Introduction
- What is (supervised) ML?
- Supervised v Unsupervised Learning
- Representing Data as Features

# Administrivia

- Lecture and Lab:
  - Thursday: 6pm to 10pm
  - Weeks 1, 7 and 13 will be in-person (but streamed), the rest will be online
  - No lecture on week 13 (07.05.2026)!
- All notes, lecture recordings, tutorials, lab work, and assignments will be available on Brightspace.

- For all module queries please contact: bojan.bozic@tudublin.ie

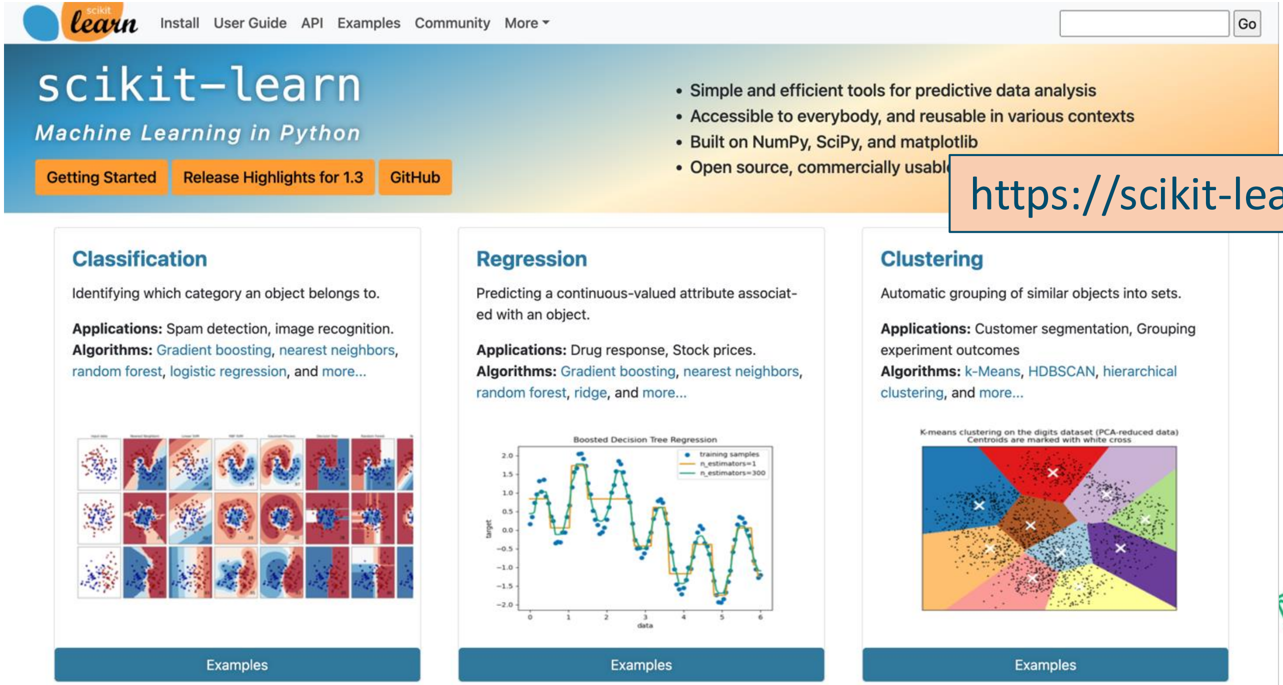
# Textbook



*Fundamentals of Machine Learning for Predictive Data Analytics*: John D. Kelleher, Brian Mac Namee, Aoife D'Arcy

# ML with Python

- Tutorials and lab work require a laptop with Jupyter and scikit-learn.



scikit-learn

*Machine Learning in Python*

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usabl

**Classification**

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.
**Algorithms:** Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

**Regression**

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.
**Algorithms:** Gradient boosting, nearest neighbors, random forest, ridge, and more...

**Clustering**

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes
**Algorithms:** k-Means, HDBSCAN, hierarchical clustering, and more...

https://scikit-learn.org/stable/

# Assessment

- Assessment is based on CA + final exam:

| Percentage | Activity |
|---|---|
| 30% | Assignment - due in week 13 |
| 20% | Lab Test - scheduled for week 9 |
| 50% | End of Semester Exam |

# Machine Learning - Overview

- Supervised Learning
- Classification: KNNs, Decision Trees, Naive Bayes
  - Neural Networks
  - Linear regression, Logistic Regression
- Dimensionality Reduction
  - Feature Selection, PCA
- The ML Process
  - Data Preprocessing, Missing Values, Scaling
  - Model Selection, Hyperparameters
- Evaluation
- Working with Text
- Unsupervised Learning

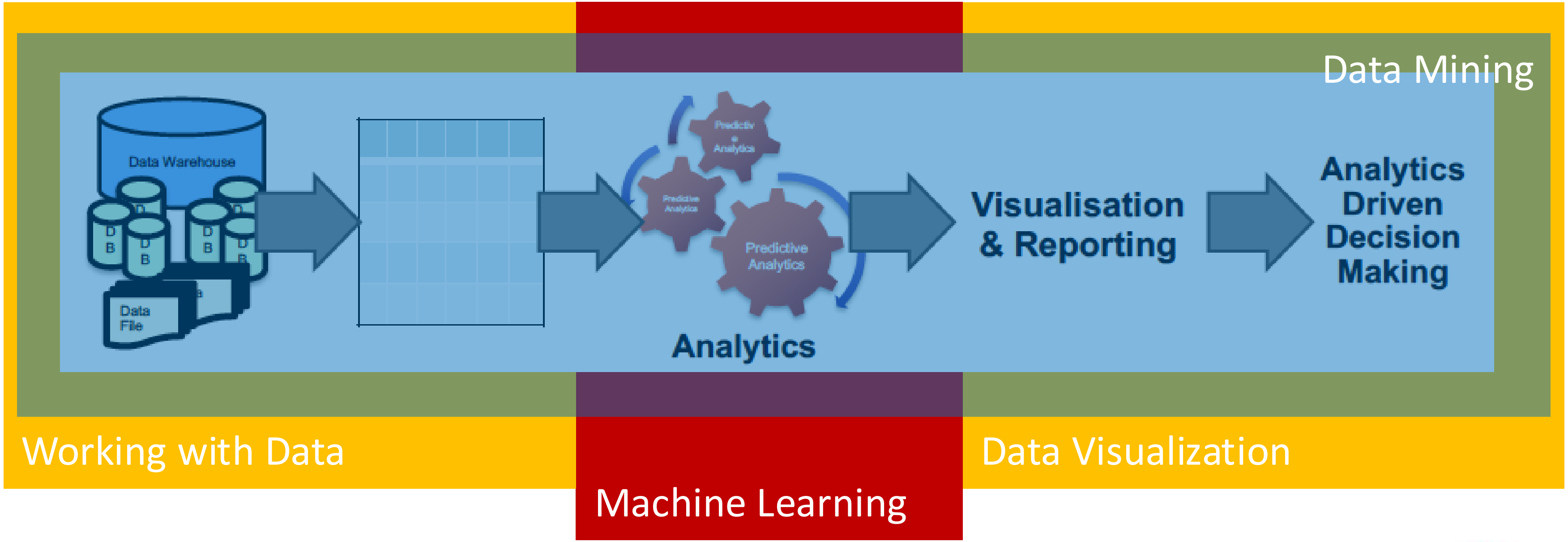
# Relevance of ML

- Explosion in rich, complex data to analyse - online and offline.
- Significant recent progress in algorithms and theory.
- Computational power is now available.
- Industry demand - Data scientists, Data engineers…
- New applications in many disciplines - medicine, engineering, humanities…

**2.4 billion users** end of 2024
Average Daily Usage of Instagram is **33 Minutes**

**more than 1.8 billion** users worldwide. Around 22.22% of the global population uses Gmail. **121 billion** emails daily (as of 2022 data)

# How ML fits in this program



Data Mining

Working with Data

Machine Learning

Data Visualization

# ML Today – Advances in AI

- Dall-E: generating realistic images from a description in natural language [https://openai.com/dall-e-2/](https://openai.com/dall-e-2/)

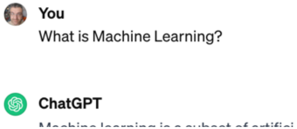
"a robot using a computer in the desert digital art”

“children playing football in a school year abstract art”

“photograph of an astronaut riding a horse”

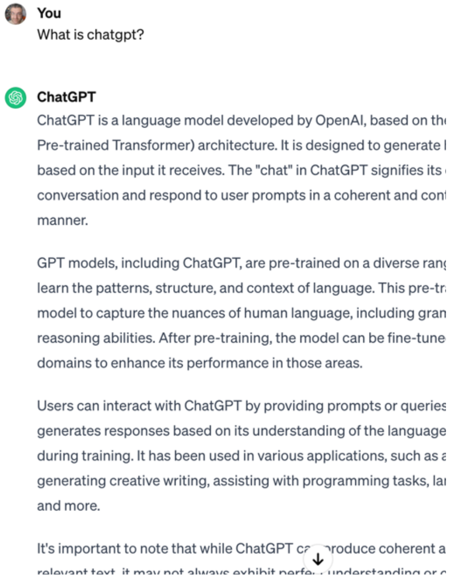
# Advances in AI

- ChatGPT: A chatbot built on large language models that interacts in a conversational way [https://openai.com/blog/chatgpt/](https://openai.com/blog/chatgpt/)



**You**
What is Machine Learning?

**ChatGPT**
Machine learning is a subset of artifici



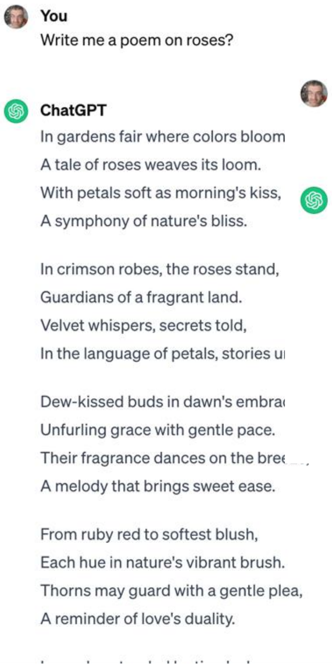
**You**
What is chatgpt?

**ChatGPT**
ChatGPT is a language model developed by OpenAI, based on th
Pre-trained Transformer) architecture. It is designed to generate
based on the input it receives. The "chat" in ChatGPT signifies its
conversation and respond to user prompts in a coherent and con
manner.

GPT models, including ChatGPT, are pre-trained on a diverse ran
learn the patterns, structure, and context of language. This pre-tr
model to capture the nuances of human language, including gra
reasoning abilities. After pre-training, the model can be fine-tune
domains to enhance its performance in those areas.

Users can interact with ChatGPT by providing prompts or queries
generates responses based on its understanding of the language
during training. It has been used in various applications, such as
generating creative writing, assisting with programming tasks, la
and more.

It's important to note that while ChatGPT c produce coherent a
relevant text, it may not always exhibit perfe understanding or c



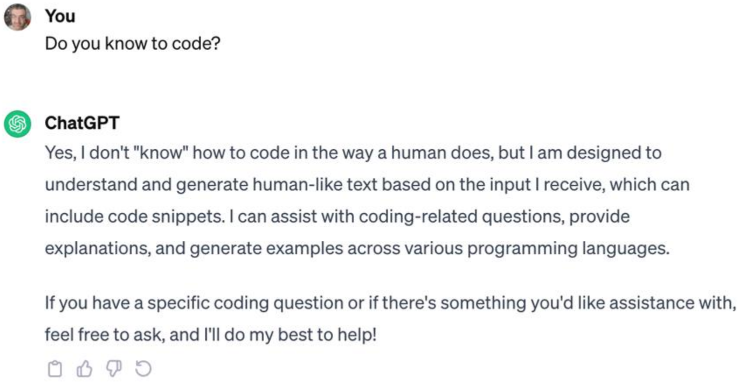
**You**
Write me a poem on roses?

**ChatGPT**
In gardens fair where colors bloom
A tale of roses weaves its loom.
With petals soft as morning's kiss,
A symphony of nature's bliss.

In crimson robes, the roses stand,
Guardians of a fragrant land.
Velvet whispers, secrets told,
In the language of petals, stories u

Dew-kissed buds in dawn's embra
Unfurling grace with gentle pace.
Their fragrance dances on the bre
A melody that brings sweet ease.

From ruby red to softest blush,
Each hue in nature's vibrant brush.
Thorns may guard with a gentle plea,
A reminder of love's duality.

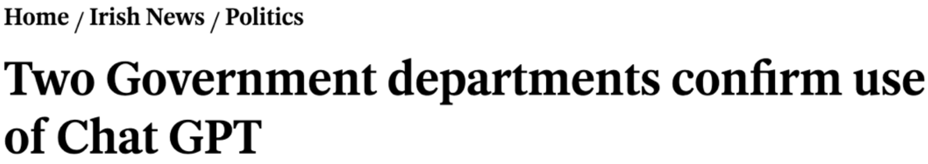


**You**
Do you know to code?

**ChatGPT**
Yes, I don't "know" how to code in the way a human does, but I am designed to understand and generate human-like text based on the input I receive, which can include code snippets. I can assist with coding-related questions, provide explanations, and generate examples across various programming languages.

If you have a specific coding question or if there's something you'd like assistance with, feel free to ask, and I'll do my best to help!

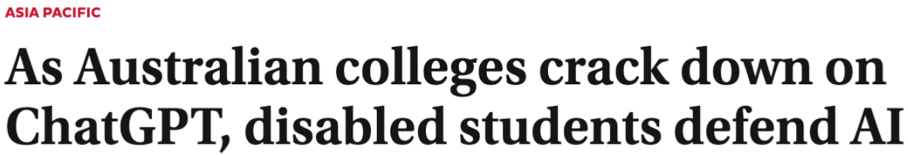
# Reactions



Home / Irish News / Politics

**Two Government departments confirm use of Chat GPT**

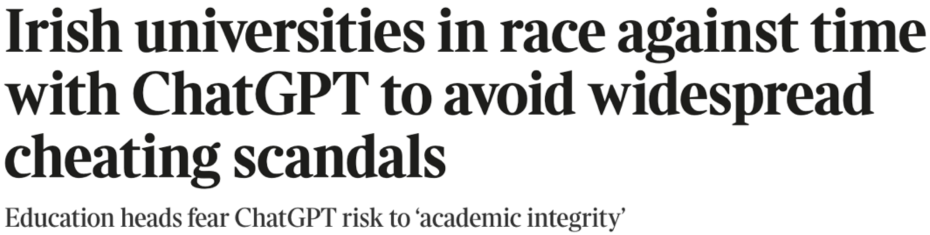
https://www.independent.ie/irish-news/politics/two-government-departments-confirm-use-of-chat-gpt/a1788591640.html



ASIA PACIFIC

**As Australian colleges crack down on ChatGPT, disabled students defend AI**

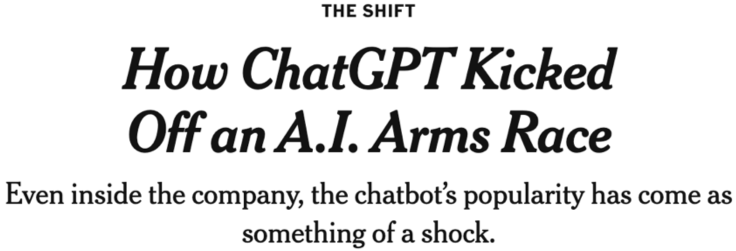
https://www.japantimes.co.jp/news/2023/01/24/asia-pacific/australian-colleges-chatgpt-disabled-students/



**Irish universities in race against time with ChatGPT to avoid widespread cheating scandals**

Education heads fear ChatGPT risk to 'academic integrity'

https://www.thetimes.co.uk/article/irish-universities-in-race-against-time-with-chatgpt-to-avoid-widespread-cheating-scandals-kbpmtqqw9



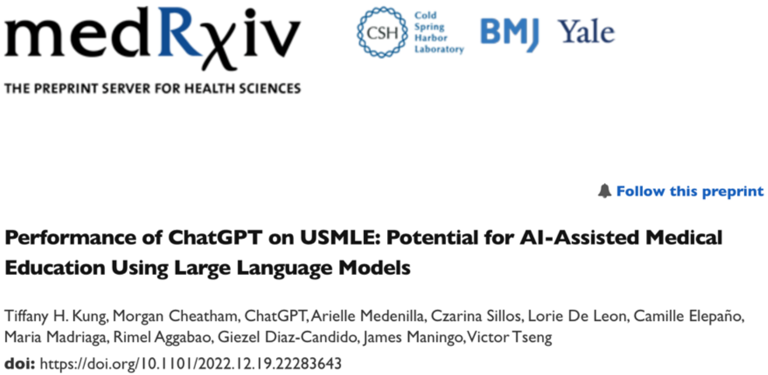
THE SHIFT

***How ChatGPT Kicked Off an A.I. Arms Race***

Even inside the company, the chatbot's popularity has come as something of a shock.

https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html

# Education


medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory BMJ Yale

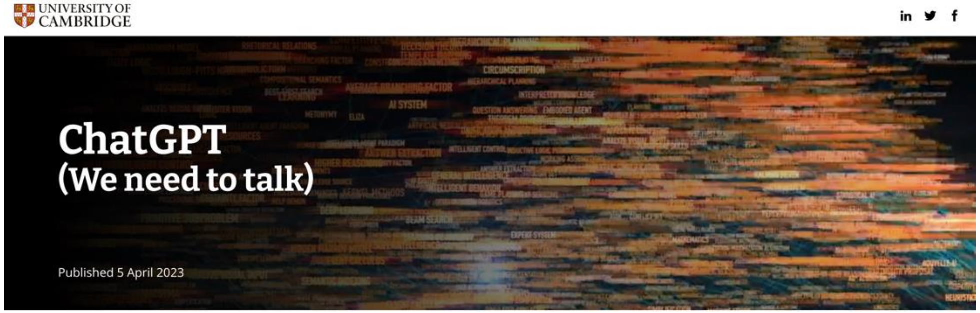
Follow this preprint

**Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models**

Tiffany H. Kung, Morgan Cheatham, ChatGPT, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, Victor Tseng

**doi:** https://doi.org/10.1101/2022.12.19.22283643

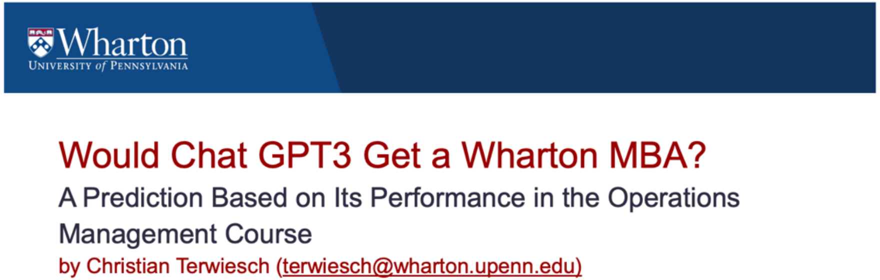
https://www.medrxiv.org/content/10.1101/2022.12.19.22283643v2.full


UNIVERSITY OF CAMBRIDGE

**ChatGPT**
**(We need to talk)**

Published 5 April 2023

https://www.cam.ac.uk/stories/ChatGPT-and-education


Wharton UNIVERSITY of PENNSYLVANIA

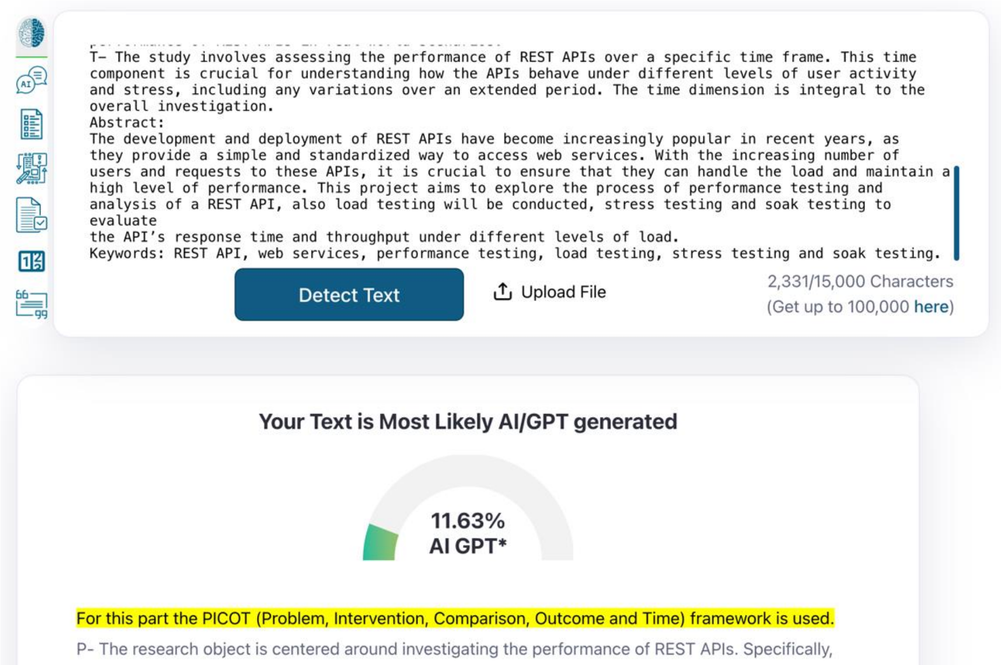
**Would Chat GPT3 Get a Wharton MBA?**

A Prediction Based on Its Performance in the Operations Management Course

by Christian Terwiesch (terwiesch@wharton.upenn.edu)

https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf

# GPTZero to detect AI



T- The study involves assessing the performance of REST APIs over a specific time frame. This time component is crucial for understanding how the APIs behave under different levels of user activity and stress, including any variations over an extended period. The time dimension is integral to the overall investigation.
Abstract:
The development and deployment of REST APIs have become increasingly popular in recent years, as they provide a simple and standardized way to access web services. With the increasing number of users and requests to these APIs, it is crucial to ensure that they can handle the load and maintain a high level of performance. This project aims to explore the process of performance testing and analysis of a REST API, also load testing will be conducted, stress testing and soak testing to evaluate
the API's response time and throughput under different levels of load.
Keywords: REST API, web services, performance testing, load testing, stress testing and soak testing.

2,331/15,000 Characters
(Get up to 100,000 **here**)

**Your Text is Most Likely AI/GPT generated**

**11.63%**
**AI GPT***

For this part the PICOT (Problem, Intervention, Comparison, Outcome and Time) framework is used.

P- The research object is centered around investigating the performance of REST APIs. Specifically,

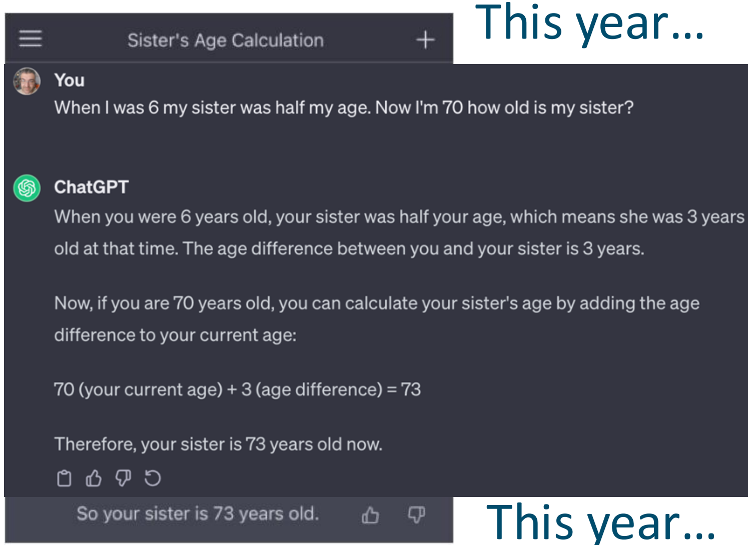
https://www.zerogpt.com/?gclid=CjwKCAiAkp6tBhB5EiwANTCx1AmcBLk3uO8ijqusyd1ngey8R2HZxG1WAwYgXoUuWFGDZ5xK7nVQlhoCad8QAvD_Bw

# But…

- ChatGPT can put together answers to questions but doesn’t “know” anything
- No comprehensive understanding of the physical and social world, no ability to reason about relationships between concepts and entities
- Examples of failures [https://github.com/giuven95/chatgpt-failures](https://github.com/giuven95/chatgpt-failures)

A year ago…

This year…



Sister's Age Calculation

**You**
When I was 6 my sister was half my age. Now I'm 70 how old is my sister?

**ChatGPT**
When you were 6 years old, your sister was half your age, which means she was 3 years old at that time. The age difference between you and your sister is 3 years.

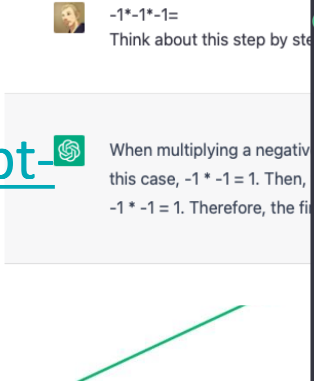
Now, if you are 70 years old, you can calculate your sister's age by adding the age difference to your current age:

70 (your current age) + 3 (age difference) = 73

Therefore, your sister is 73 years old now.

So your sister is 73 years old.

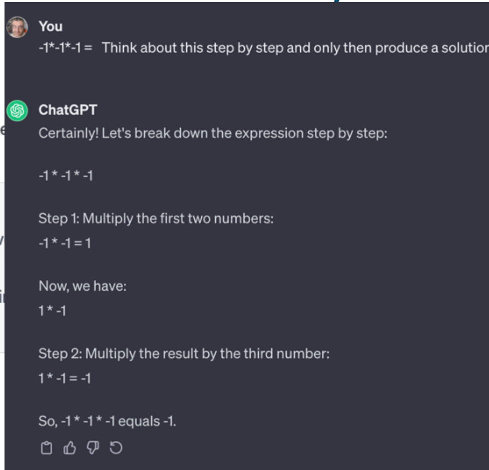
A year ago…



-1*-1*-1=
Think about this step by ste

When multiplying a negativ
this case, -1 * -1 = 1. Then,
-1 * -1 = 1. Therefore, the fi

This year…



**You**
-1*-1*-1 = Think about this step by step and only then produce a solution

**ChatGPT**
Certainly! Let's break down the expression step by step:

-1 * -1 * -1

Step 1: Multiply the first two numbers:
-1 * -1 = 1

Now, we have:
1 * -1

Step 2: Multiply the result by the third number:
1 * -1 = -1

So, -1 * -1 * -1 equals -1.

# DeepSeek

- Nvidia 2022 export ban
- DS built at a fraction of the cost of industry-leading models like OpenAI - because it uses fewer advanced chips.
- It is reportedly as powerful as OpenAI's o1 model - released at the end of last year - in tasks including mathematics and coding.
- Like o1, R1 is a "reasoning" model. These models produce responses incrementally, simulating a process similar to how humans reason through problems or ideas. It uses less memory than its rivals, ultimately reducing the cost to perform tasks.

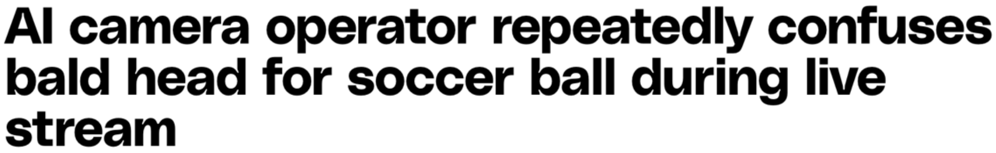
Source: https://www.bbc.com/news/articles/c5yv5976z9po

# Next…

- Applications will get more specific.
- Regulation
- EU AI Act

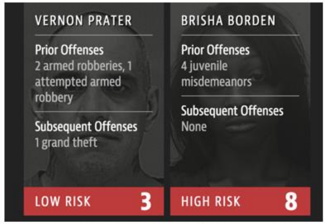
  https://artificialintelligenceact.eu/

# AI Systems are still brittle…



AI camera operator repeatedly confuses bald head for soccer ball during live stream

https://www.theverge.com/tldr/2020/11/3/21547392/ai-camera-operator-football-bald-head-soccer-mistakes



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

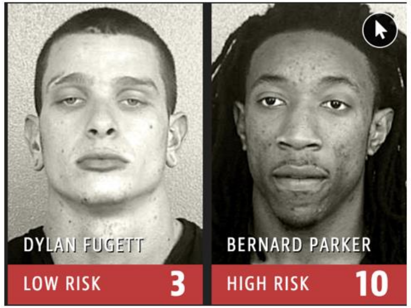
LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

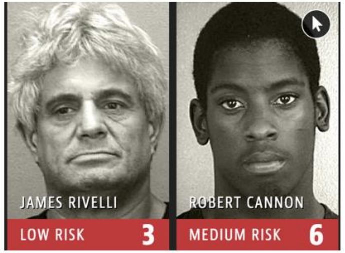


DYLAN FUGETT

LOW RISK 3

BERNARD PARKER

HIGH RISK 10

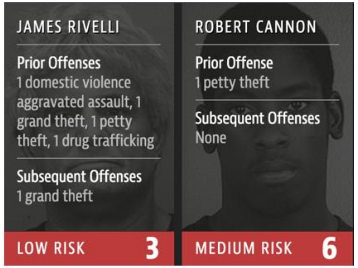


JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6



JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

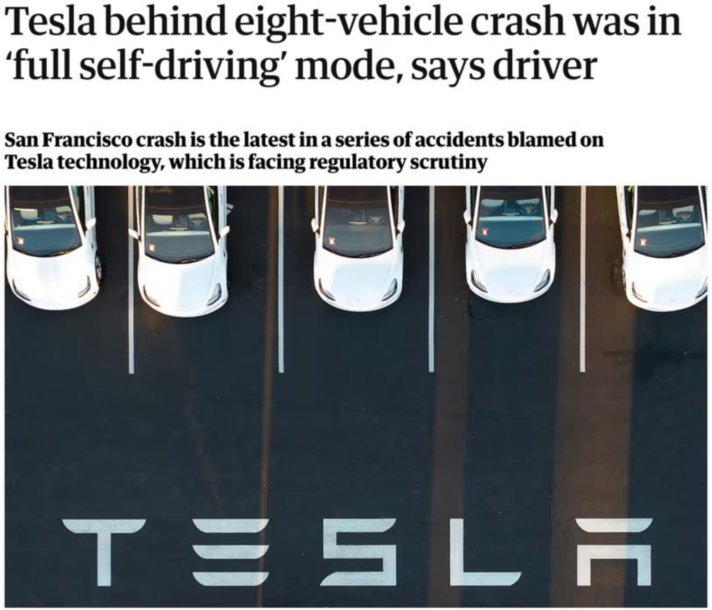
ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1



Tesla behind eight-vehicle crash was in 'full self-driving' mode, says driver

San Francisco crash is the latest in a series of accidents blamed on Tesla technology, which is facing regulatory scrutiny

https://www.theguardian.com/technology/2022/dec/22/tesla-crash-full-self-driving-mode-san-francisco

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

# Expert Systems (Rule-Based Systems)

- Rule:

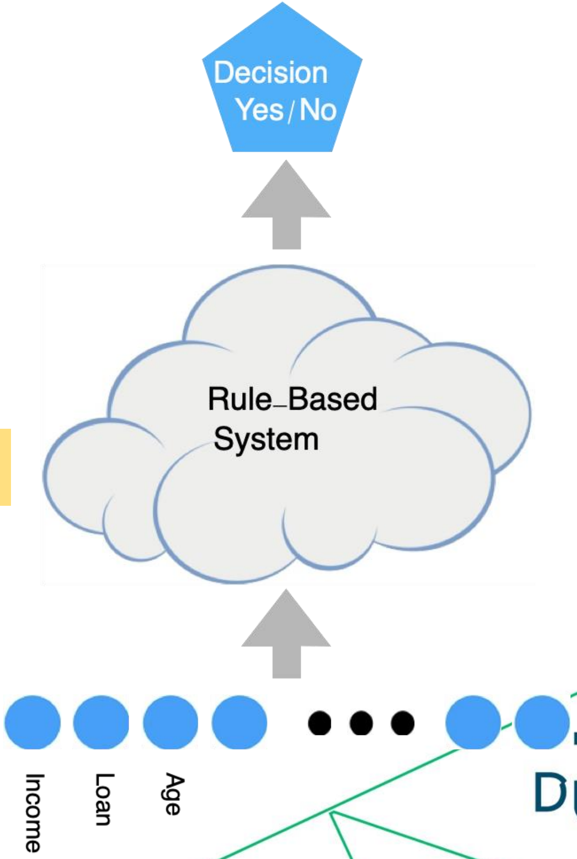
**If**

- Income > Expenditure &
- Collateral > Loan

**Then**

- Risk = Low

Knowledge engineering



# Learning from historical data

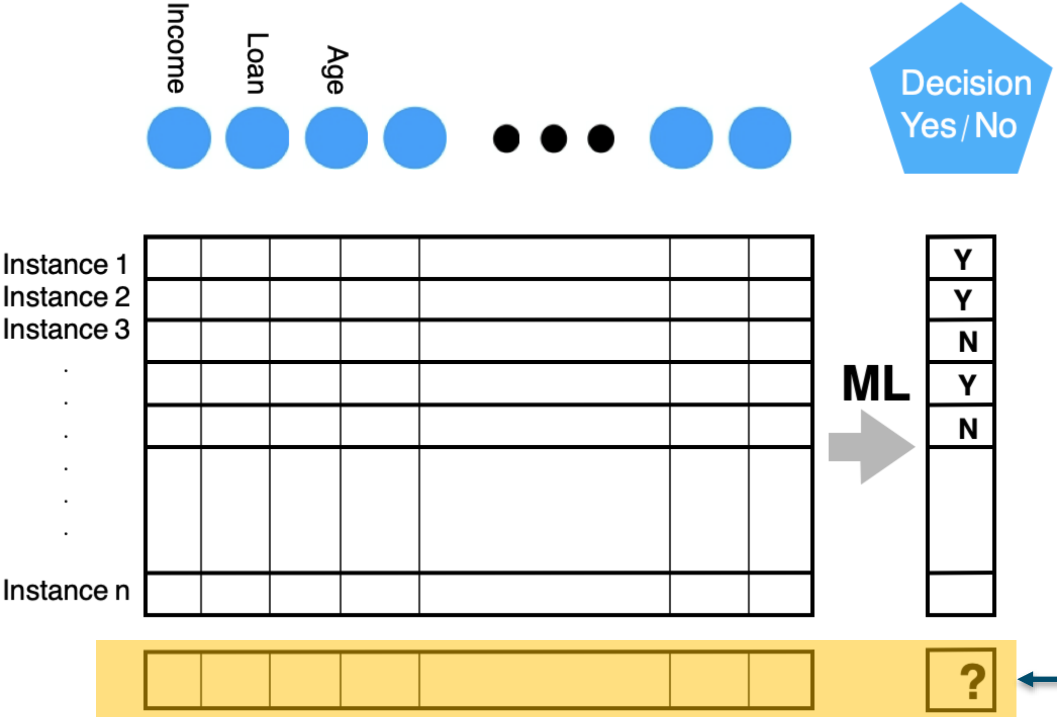


Table of historical data

Each row

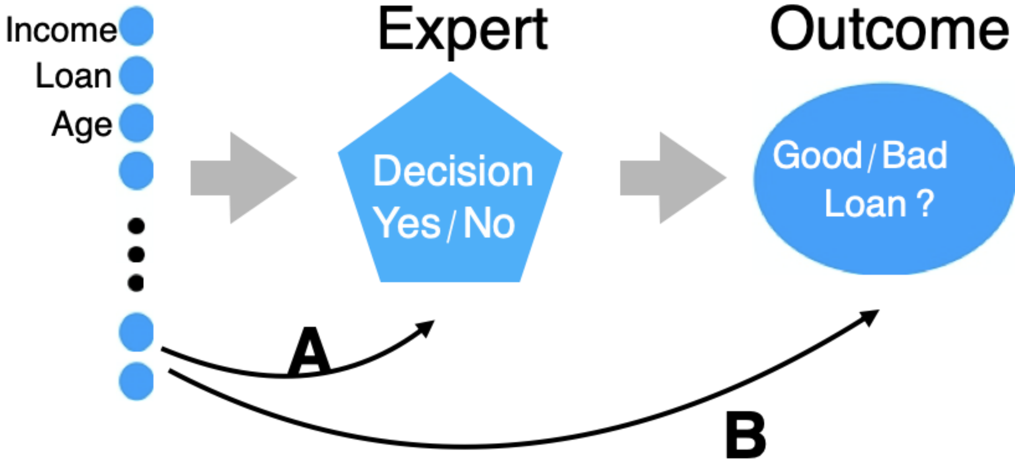
- ✓ description of instance
- ✓ decision (Yes/No)

Predict decision for new instance

# Learn what?

A. Learn expert decision-making
   - What if the expert gets it wrong sometimes?

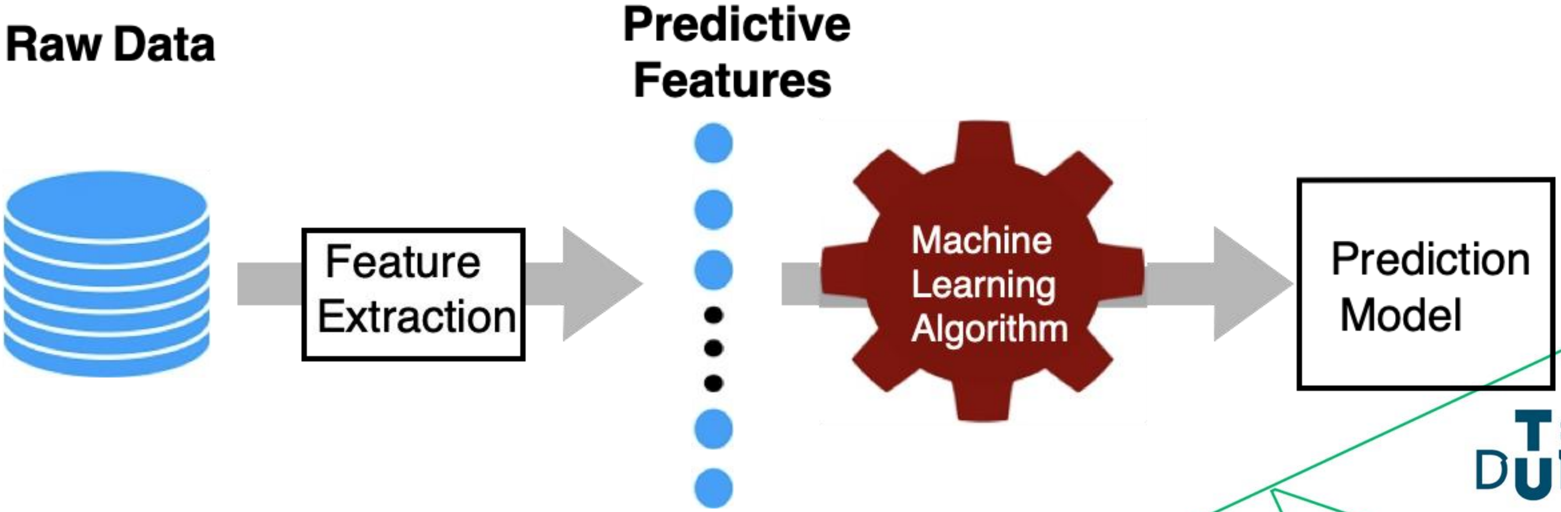
B. Learn from outcomes
   - Outperform experts



# Supervised ML (Predictive Analytics)

**Training Step:**

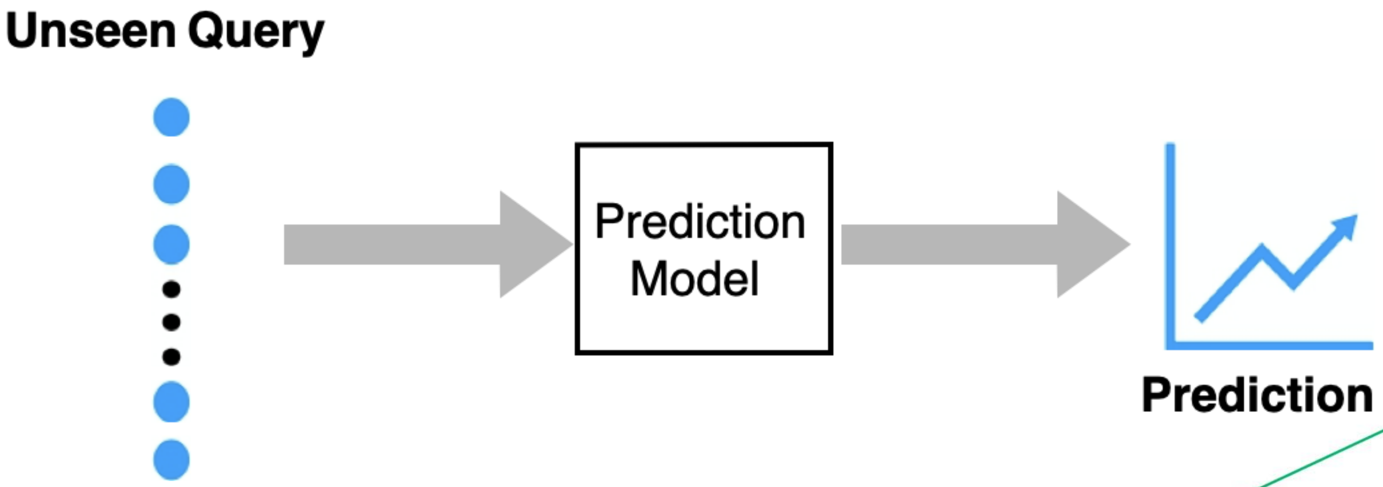
- Learning a model from a set of historical data instances



# Supervised ML (Predictive Analytics)

**Prediction Step:**

- Using the model to make predictions



**Unseen Query**

Prediction Model

**Prediction**

# Classification Task

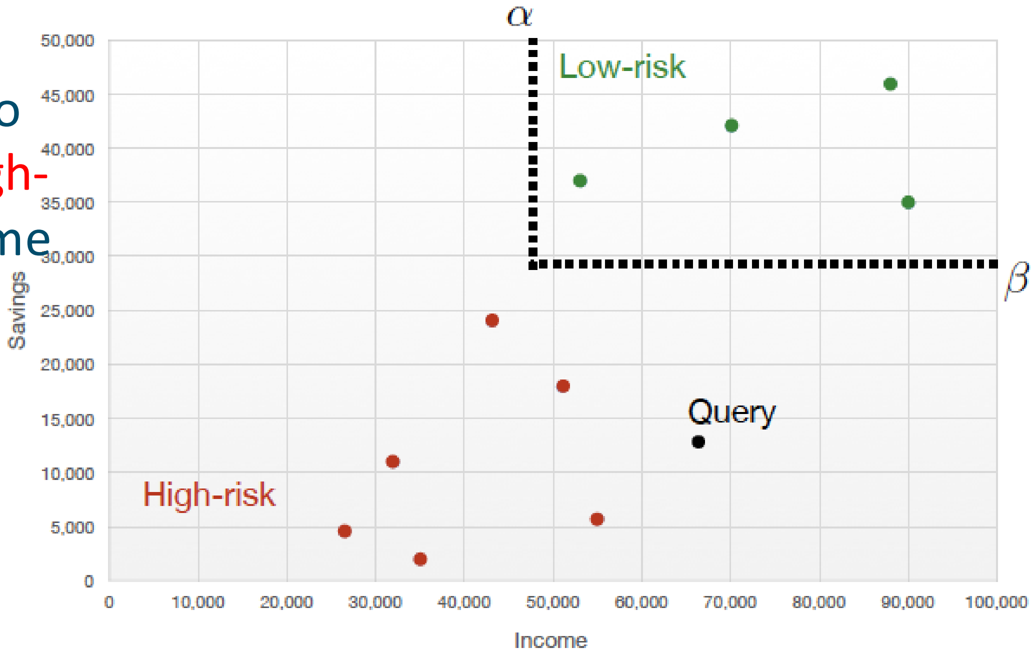
Example: Credit scoring

- A training set with 10 examples (customers)
- Each example has one of two class labels = {High-risk, Low-risk}

| Example | Income | Savings | Married | Gender | Age | Class |
|---|---|---|---|---|---|---|
| **1** | 35,000 | 2,000 | Y | M | 32 | High-risk |
| **2** | 51,000 | 18,000 | N | M | 34 | High-risk |
| **3** | 70,000 | 42,000 | Y | F | 41 | Low-risk |
| **4** | 26,500 | 4,500 | N | M | 22 | High-risk |
| **5** | 32,000 | 11,000 | N | F | 25 | High-risk |
| **6** | 53,000 | 37,000 | N | F | 39 | Low-risk |
| **7** | 88,000 | 46,000 | Y | M | 48 | Low-risk |
| **8** | 55,000 | 5,700 | N | M | 55 | High-risk |
| **9** | 90,000 | 35,000 | Y | F | 61 | Low-risk |
| **10** | 43,000 | 24,000 | Y | M | 33 | High-risk |

# Classification task

- Manually classify customers into two categories (low-risk and high-risk) based on savings and income data.



**If**

income > α & savings > β

**Then**

Risk - low

# Classification Task

| Example | Income | Savings | Married | Gender | Age | Class |
|---|---|---|---|---|---|---|
| **1** | 35,000 | 2,000 | Y | M | 32 | High–risk |
| **2** | 51,000 | 18,000 | N | M | 34 | High–risk |
| **3** | 70,000 | 42,000 | Y | F | 41 | Low–risk |
| **4** | 26,500 | 4,500 | N | M | 22 | High–risk |
| **5** | 32,000 | 11,000 | N | F | 25 | High–risk |
| **6** | 53,000 | 37,000 | N | F | 39 | Low–risk |
| **7** | 88,000 | 46,000 | Y | M | 48 | Low–risk |
| **8** | 55,000 | 5,700 | N | M | 55 | High–risk |
| **9** | 90,000 | 35,000 | Y | F | 61 | Low–risk |
| **10** | 43,000 | 24,000 | Y | M | 33 | High–risk |

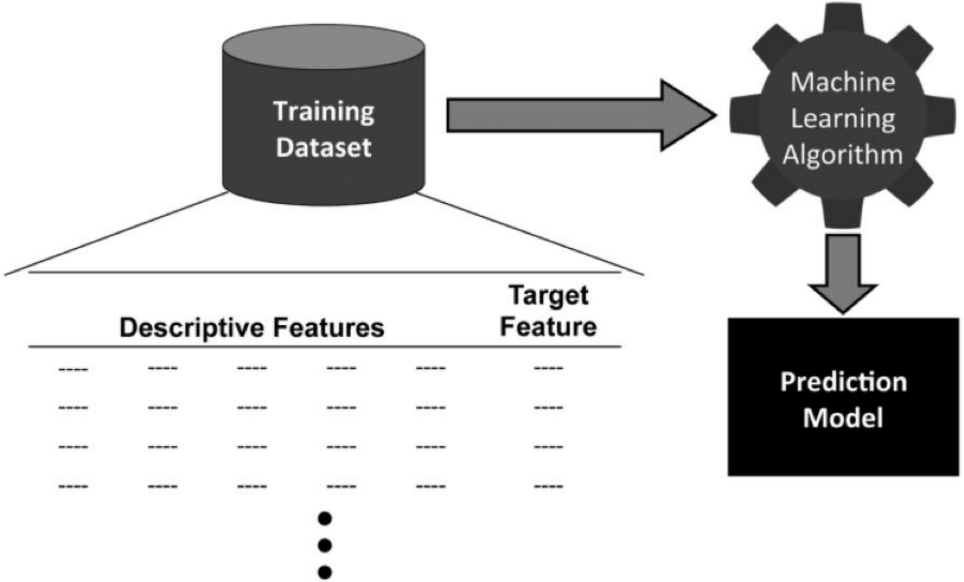
- Q. To which class does this new customer belong?

| Example | Income | Savings | Married | Gender | Age | Class |
|---|---|---|---|---|---|---|
| **X** | 66,000 | 13,000 | Y | M | 44 | ??? |

- Q. Can we train an algorithm to learn to automatically classify new customers as either low-risk or high-risk?

# Supervised Learning

- Supervised Machine Learning algorithms automate the process of learning a model that captures the relationship between the **descriptive features** and the **target feature** in a **training dataset.**

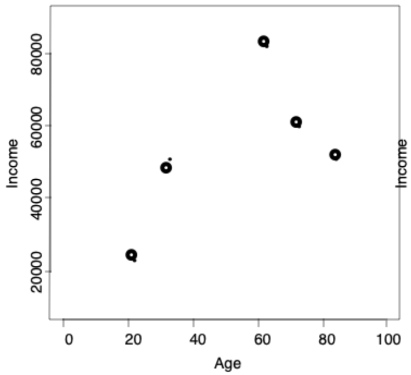


# Supervised Learning

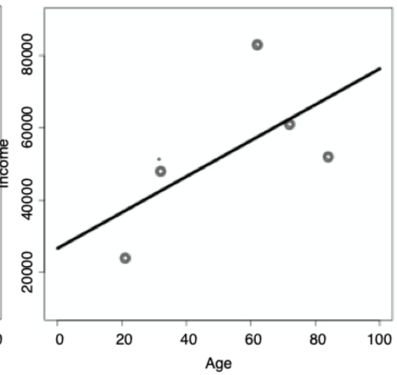
- ML algorithms search through all possible patterns that exist between a set of descriptive features and a target feature to find the best model that is **consistent** with the training data (i.e. agrees with all the training instances).
- Useful predictive models must be able to **generalise** well, i.e. make predictions for queries that are not present in the training data.

# What can go wrong?

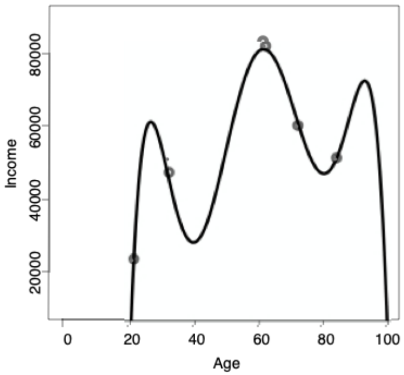
- **Underfitting** occurs when the prediction model is too simplistic to represent the underlying relationship between the descriptive and target features.
- **Overfitting** occurs when the model is so complex that it fits the data too closely and becomes sensitive to noise (e.g. mislabelled feature values).



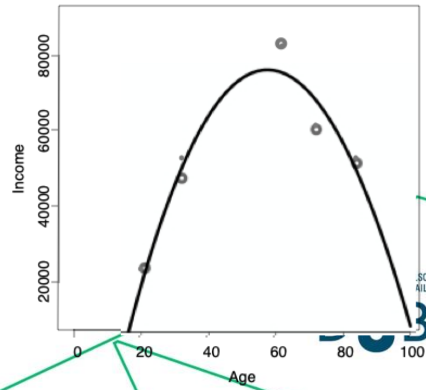
Dataset



Underfitting



Overfitting

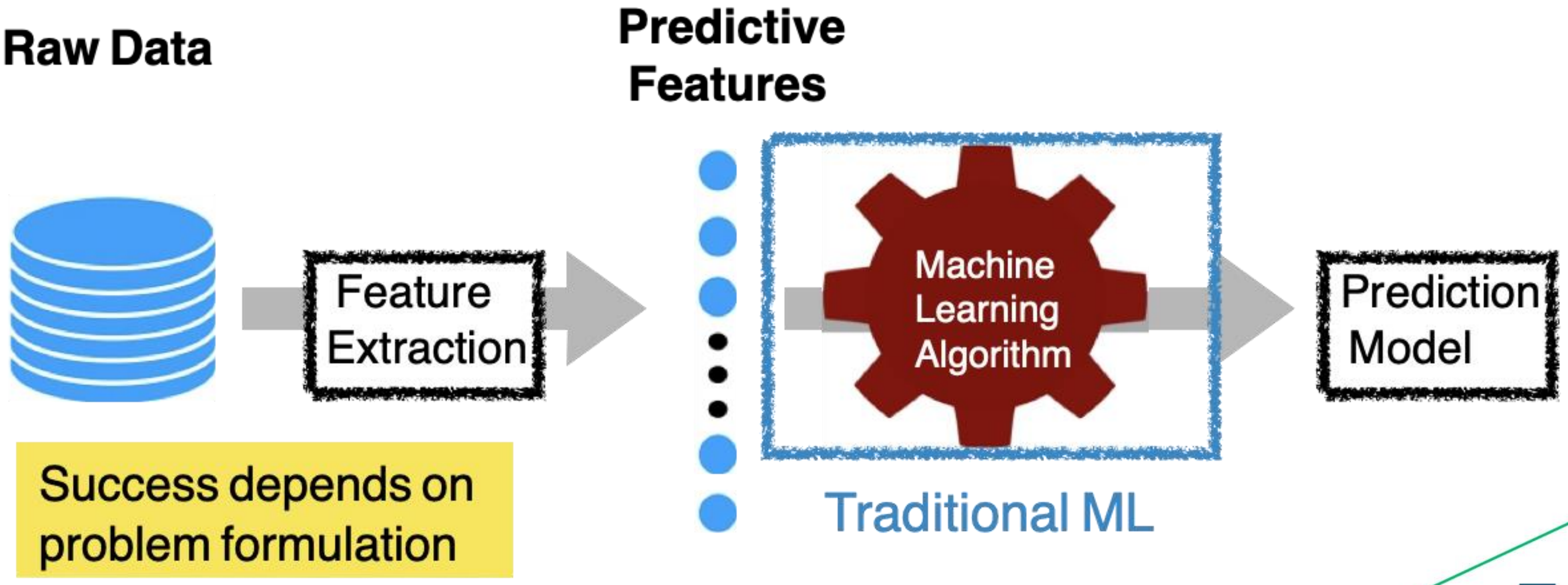


Just Right

# Supervised Learning Algorithms

- Typically, training data does not have enough information to choose a single best model, so additional assumptions are needed to drive the model selection, known as the **inductive bias.**
- **Restrictive bias** constrains the set of models that will be considered (e.g. linear regression considers models that produce predictions based on a linear combination of descriptive features)
- **Preference bias** guides the algorithm to prefer certain models over others (e.g., the decision tree prefers shallower trees)
- Different algorithms have a different inductive bias,
- It is important to identify the machine algorithm that will fit the predictive task most appropriately.

# Supervised ML (aka Predictive Analytics)



**Raw Data**

Feature Extraction

**Predictive Features**

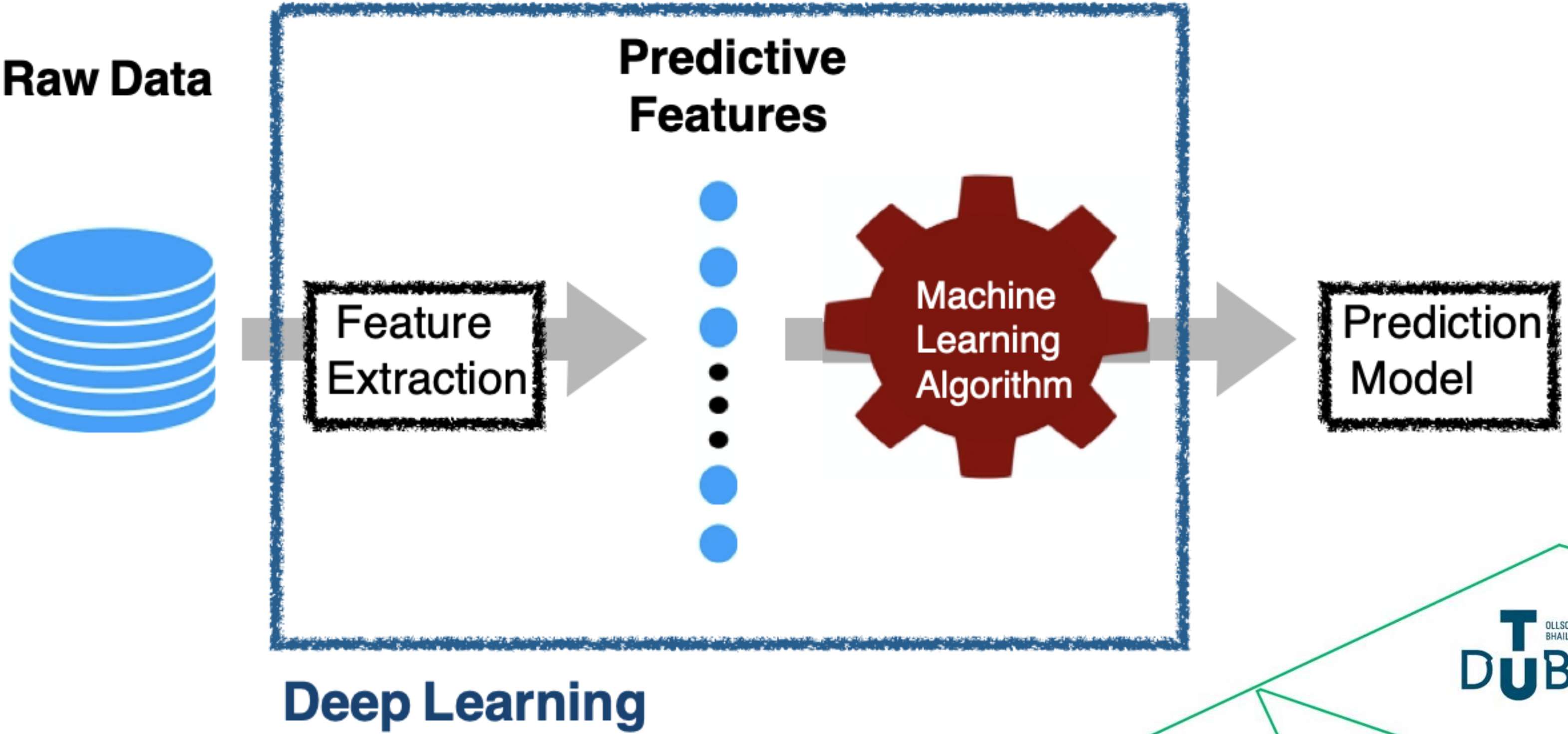
Machine Learning Algorithm

Traditional ML

Prediction Model

Success depends on problem formulation

# Supervised ML (aka Predictive Analytics)



# Supervised vs. Unsupervised Learning

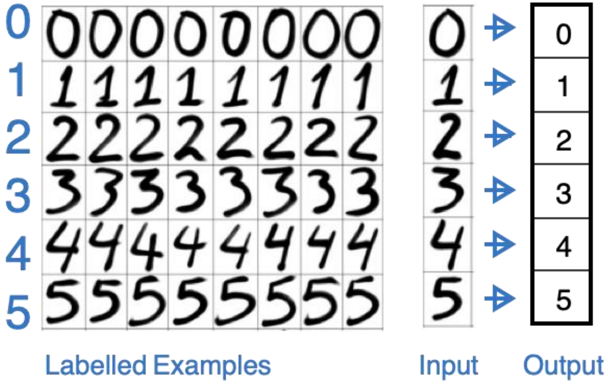
- **Supervised Learning:**
  - An algorithm that learns a function from examples of its inputs and outputs. It uses manually labelled example data (i.e. a training set ) to predict the correct answer for new unseen query inputs.

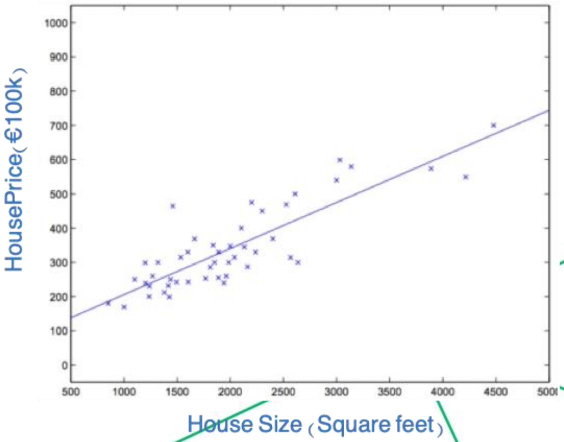
    e.g. Classification, regression algorithms

- **Unsupervised Learning:**
  - An algorithm that finds structure in data where no manually labelled examples are available as inputs - i.e. there is no training set. These algorithms are more focused on data exploration and knowledge discovery.
  - e.g. Clustering, topic modelling algorithms

# Supervised Learning

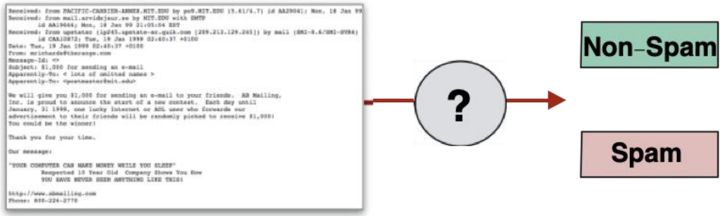
- **Classification:**
  - Examples are represented by a set of features, which help decide the target class to which a new query input belongs (i.e. the output is a class label or target feature).
- **Regression:**
  - Examples are characterized by a set of features which help decide the value of a continuous output variable (i.e. the output is a number).





# Classification Tasks

- **Binary classification:** Assign a new query input to one of two possible target class labels.



```
Received: from PACIFIC-CARRIER-ANNEX.MIT.EDU by po9.MIT.EDU (5.61/4.7) id AA29041; Mon, 18 Jan 99
Received: from mail.arvidsjaur.se by MIT.EDU with SMTP
        id AA19666; Mon, 18 Jan 99 21:05:04 EST
Received: from upstate (ip245.upstate-sc.quik.com [209.213.129.245]) by mail (SMI-8.6/SMI-SVR4)
        id CAA10872; Tue, 19 Jan 1999 02:40:37 +0100
Date: Tue, 19 Jan 1999 02:40:37 +0100
From: mrichards@therange.com
Message-Id: <>
Subject: $1,000 for sending an e-mail
Apparently-To: < lots of omitted names >
Apparently-To: <postmaster@mit.edu>

We will give you $1,000 for sending an e-mail to your friends.  AB Mailing,
Inc. is proud to anounce the start of a new contest.  Each day until
January, 31 1999, one lucky Internet or AOL user who forwards our
advertisement to their friends will be randomly picked to receive $1,000!
You could be the winner!

Thank you for your time.

Our message:

"YOUR COMPUTER CAN MAKE MONEY WHILE YOU SLEEP"
          Respected 10 Year Old  Company Shows You How
          YOU HAVE NEVER SEEN ANYTHING LIKE THIS!

http://www.abmailing.com
Phone: 800-226-2770
```

?

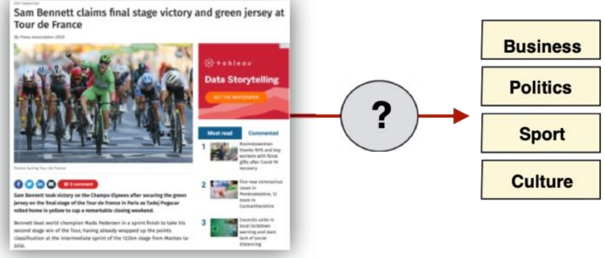
Non-Spam

Spam

Input

2 classes

- **Multiclass classification:** Assign a new query input to one of  M > 2 different target class labels.



Sam Bennett claims final stage victory and green jersey at Tour de France

By Press Association 2020

**Sam Bennett took victory on the Champs-Elysees after securing the green jersey on the final stage of the Tour de France in Paris as Tadej Pogacar rolled home in yellow to cap a remarkable closing weekend.**

Bennett beat world champion Mads Pedersen in a sprint finish to take his second stage win of the Tour, having already wrapped up the points classification at the intermediate sprint of the 122km stage from Mantes-la-Jolie.

?

Business

Politics

Sport

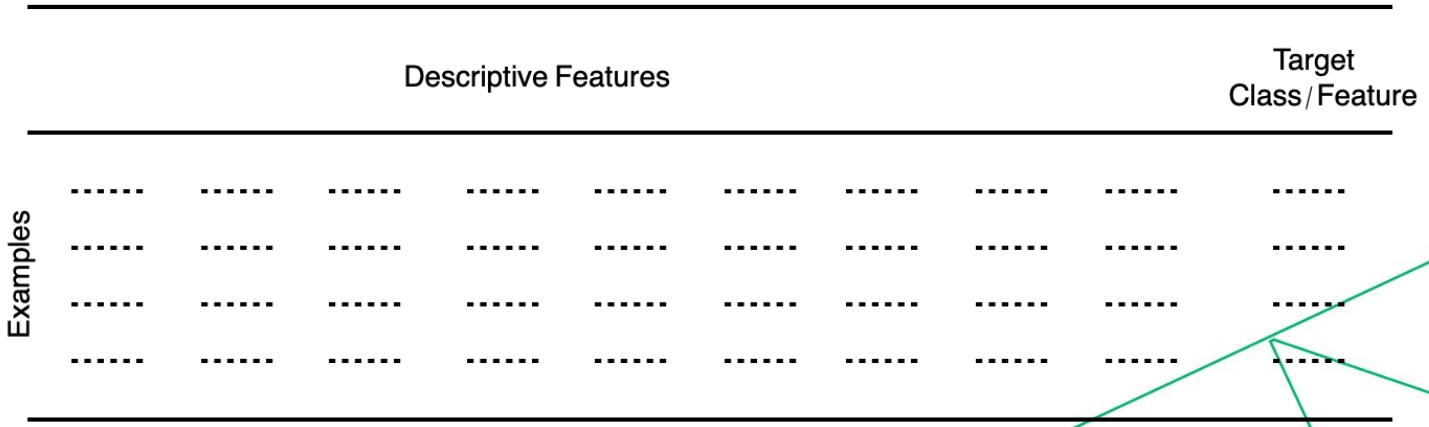
Culture

Input

4 classes

# Representing Data

- Commonly, we use a tabular structure to represent a dataset, often referred to as the **analytics base table (ABT)**.
- Each row represents a different example and comprises a set of **descriptive features**.
- For prediction, each row also has a **target class label** or **target feature** for classification and regression, respectively - i.e. the “correct answer".



# Representing Data

The descriptive features used to represent examples can be distinguished by the type and number of values they can take.

- **Binary:** Takes only two values - a boolean True/False decision
  e.g. married={True, False}, test_result={Pass, Fail}
- **Categorical (Nominal):** A feature that takes values from a finite set of values, with no intrinsic ordering to the values
  e.g. blood_group={A,B,AB,O}, nationality={French, Irish, Italian}
- **Ordinal:** A categorical variable with a clear ordering of the variables.
  e.g. grade={A, B, C, D, E, F}, dosage={Low, Medium, High}
- **Interval:** Values that allow ordering and subtraction but do not allow other arithmetic operations
  e.g date, time
- **Continuous:** Numeric measurements, with or without a fixed range for the values.
  e.g. temperature, price, age, weight, height, latitude, longitude etc.

# Typical Classification Task

- The training set with N=10 examples (customers). Each is described by D=5 features: 3 continuous, 2 categorical
- Each example has one of two class labels = {High-risk, Low-risk}

| Example | Income | Savings | Married | Gender | Age | Class |
|---|---|---|---|---|---|---|
| **1** | 35,000 | 2,000 | Y | M | 32 | High–risk |
| **2** | 51,000 | 18,000 | N | M | 34 | High–risk |
| **3** | 70,000 | 42,000 | Y | F | 41 | Low–risk |
| **4** | 26,500 | 4,500 | N | M | 22 | High–risk |
| **5** | 32,000 | 11,000 | N | F | 25 | High–risk |
| **6** | 53,000 | 37,000 | N | F | 39 | Low–risk |
| **7** | 88,000 | 46,000 | Y | M | 48 | Low–risk |
| **8** | 55,000 | 5,700 | N | M | 55 | High–risk |
| **9** | 90,000 | 35,000 | Y | F | 61 | Low–risk |
| **10** | 43,000 | 24,000 | Y | M | 33 | High–risk |

Q. To which class does this new customer belong?

| Example | Income | Savings | Married | Gender | Age | Class |
|---|---|---|---|---|---|---|
| **X** | 66,000 | 13,000 | Y | M | 44 | ??? |

# Algorithms

- Many different learning algorithms exist for prediction (e.g. k -nearest neighbour, decision tree, neural network, support vector machine).
- Due to processing, memory, and storage constraints, problem dimensions will often determine which algorithm will be practically applicable.

1. Number of input examples N.
   ➡ Sometimes millions of input examples.
2. Number of features (dimensions) D representing each input example.
   ➡ Often 10-1000, but sometimes far higher.
3. For classification, the number of target classes M.
   ➡ Often small (binary), but sometimes far higher.

# Machine Learning - Overview

- Supervised Learning
- Classification: KNNs, Decision Trees, Naive Bayes
  - Neural Networks
  - Linear regression, Logistic Regression
- Dimensionality Reduction
  - Feature Selection, PCA
- The ML Process
  - Data Preprocessing, Missing Values, Scaling
  - Model Selection, Hyperparameters
- Evaluation
- Working with Text
- Unsupervised Learning

# Questions?

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH

TU DUBLIN

TECHNOLOGICAL
UNIVERSITY DUBLIN