Machine Learning

Introduction

Sarah Jane Delany

TU Dublin
School of Computer Science

Overview

- Administrivia
- Module Outline
- Machine Learning Introduction
  - What is (supervised) ML?
  - Supervised v Unsupervised Learning
  - Representing Data as Features

Administrivia

Lectures: Mondays 6pm to 8pm

Labs: PT Monday 8pm to 10pm

FT Wednesday 1pm to 3pm

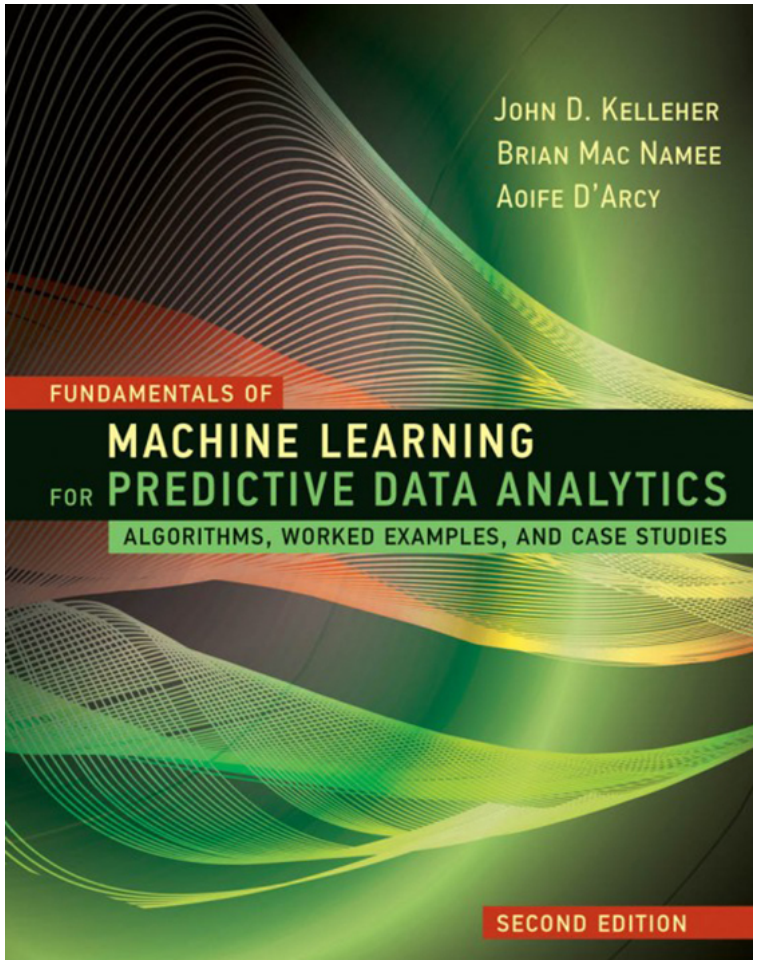
Bongo Virtual Classroom will be used for Lectures

MS Teams will be used for Labs

All notes, lecture recordings, tutorials, lab work, assignments, will be available on Brightspace

For all module queries
please contact sarahjane.delany@tudublin.ie

Textbook

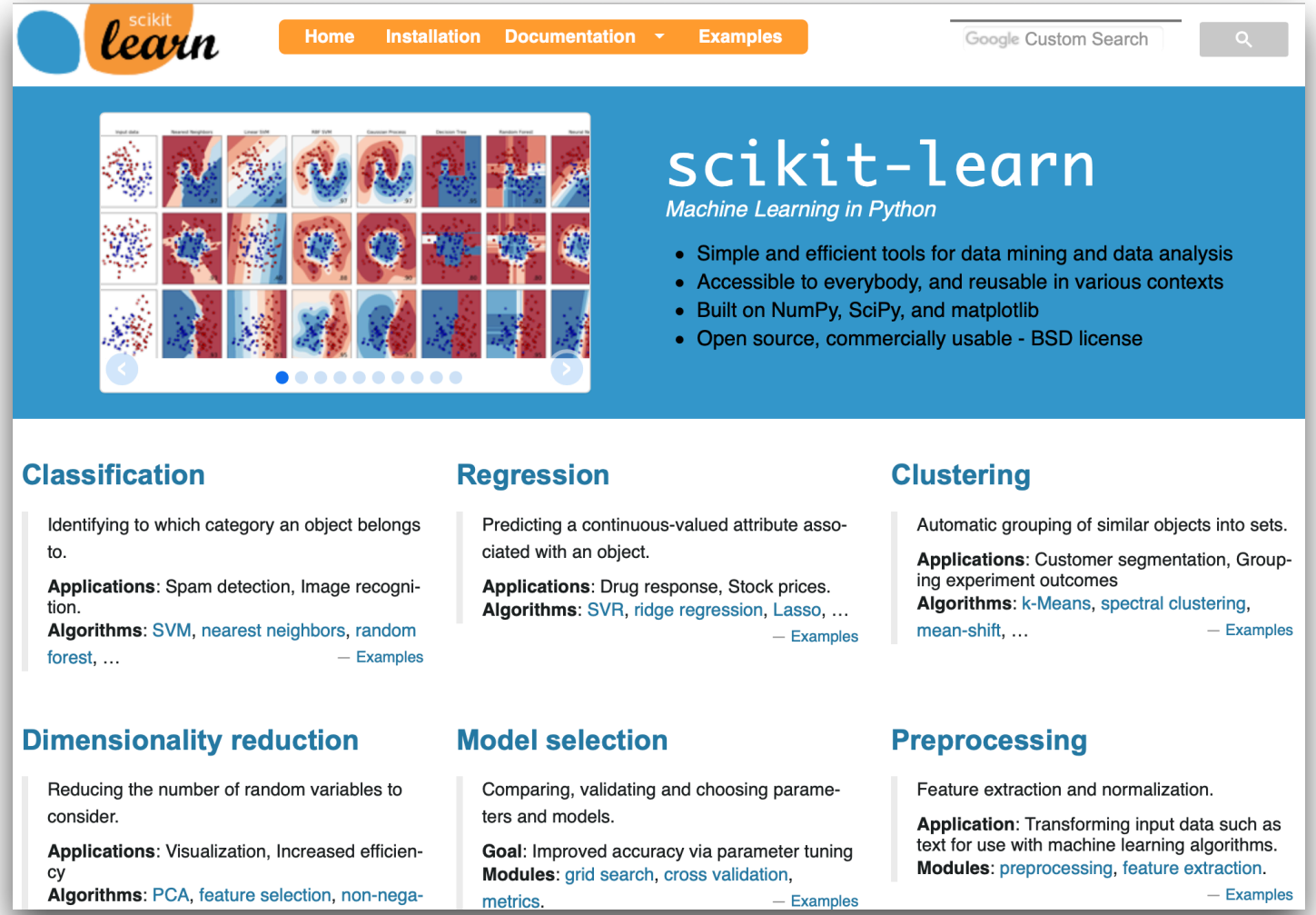


Fundamentals of Machine Learning for Predictive Data Analytics

John D. Kelleher, Brian Mac Namee, Aoife D'Arcy

ML with Python

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



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

Assessment

Assessment is based on CA + final exam:

30%	Assignment: practical application of ML due w/s April 26th
20%	Lab Test - scheduled for review week, week starting Mar 14th
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

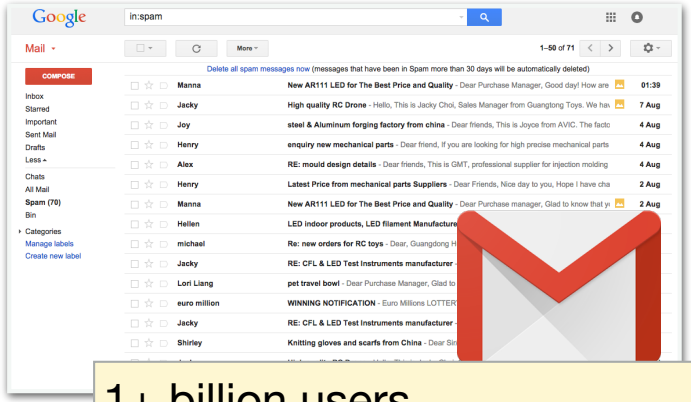
Unsupervised Learning

Reinforcement Learning

Relevance of ML

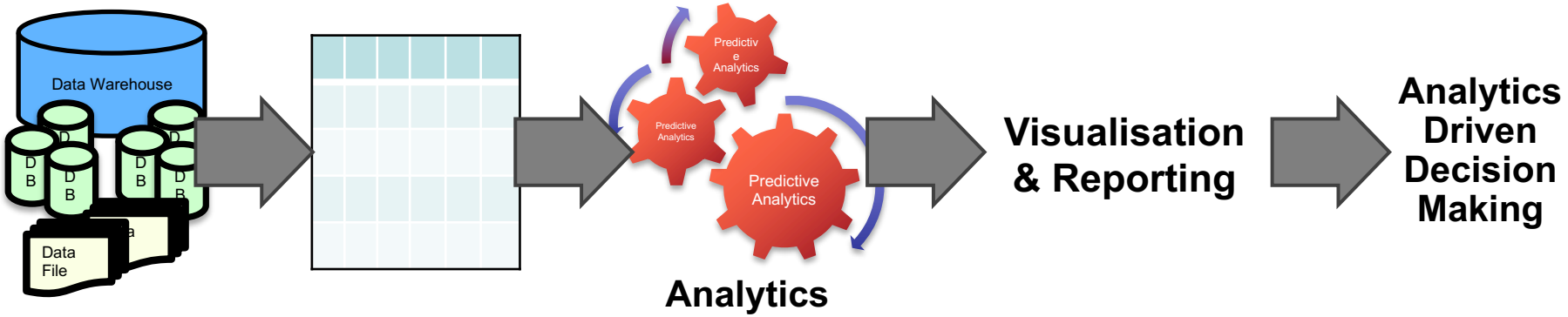
- Explosion in rich, complex data to analyse - online and offline.
- Significant recent progress in algorithms and theory.
- Computational power is now available.
- Industry demand - Data scientists, Data engineers…
- New applications in many disciplines - medicine, engineering, humanities…

330 million monthly active users
500 million tweets per day ~ 5,787 /sec
5 billion user sessions per day

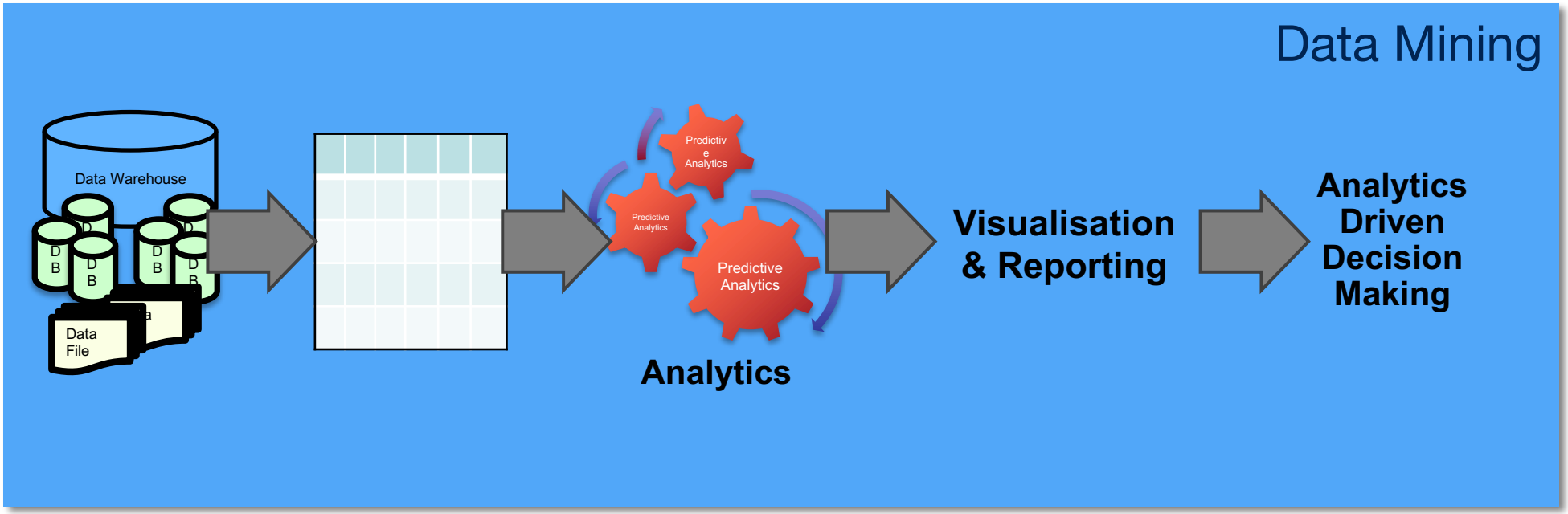


1+ billion users
Handles 2+ trillion mails per year

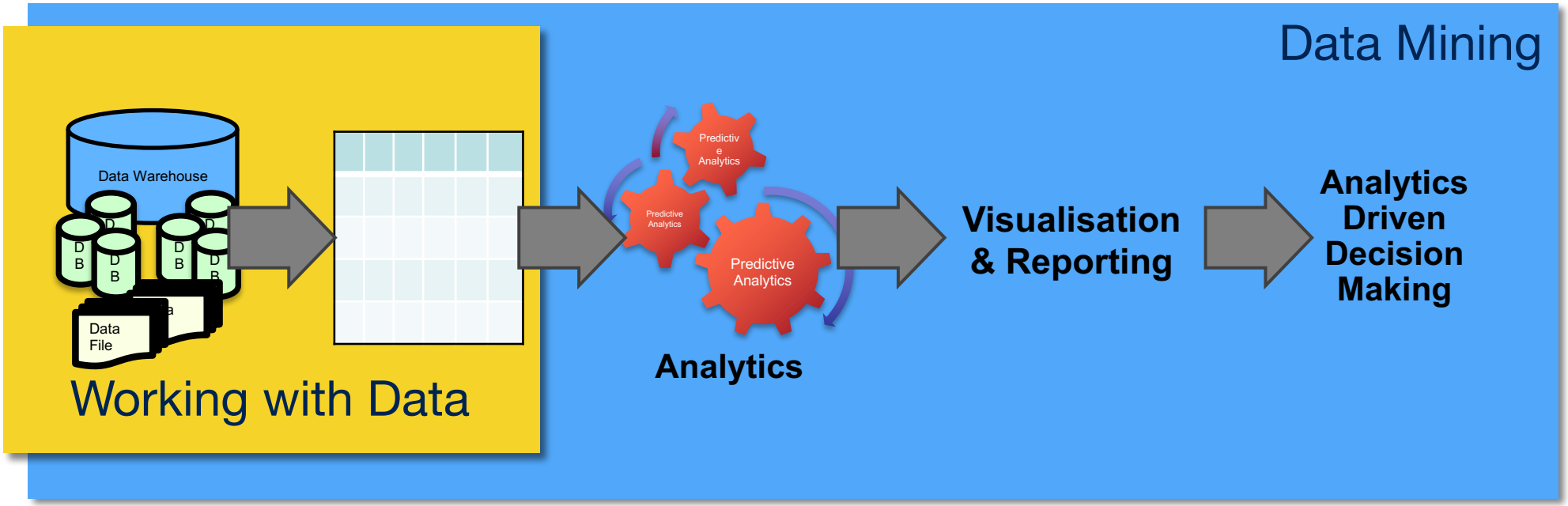
How ML fits into your programme…



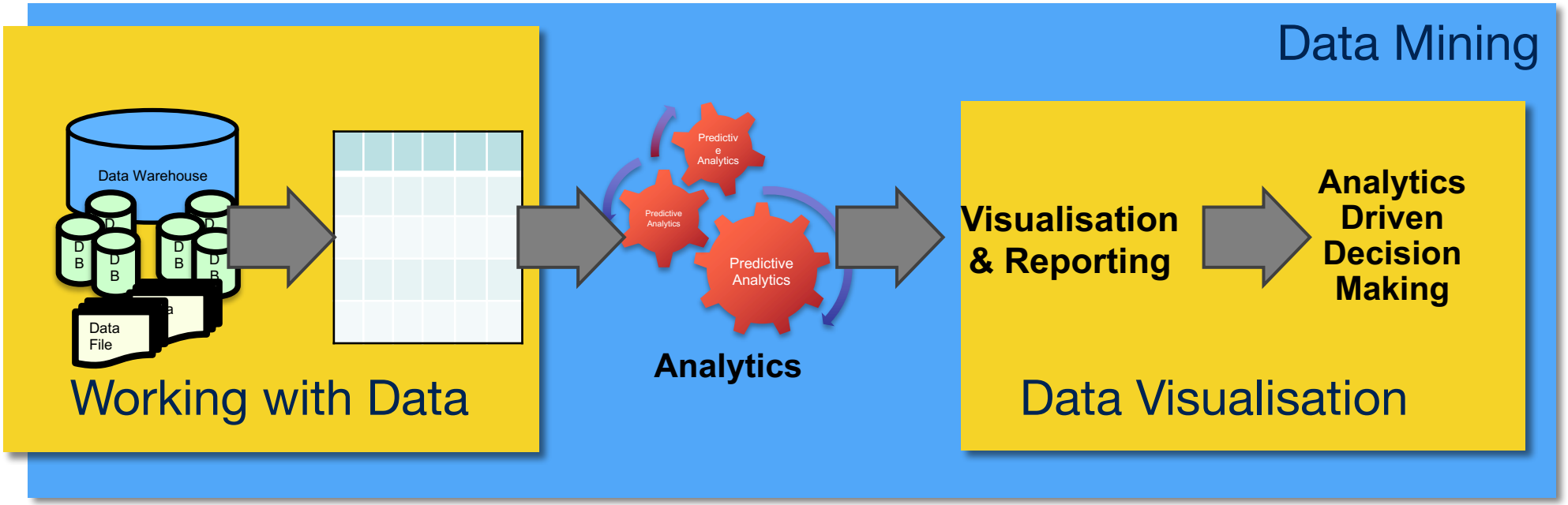
How ML fits into your programme…



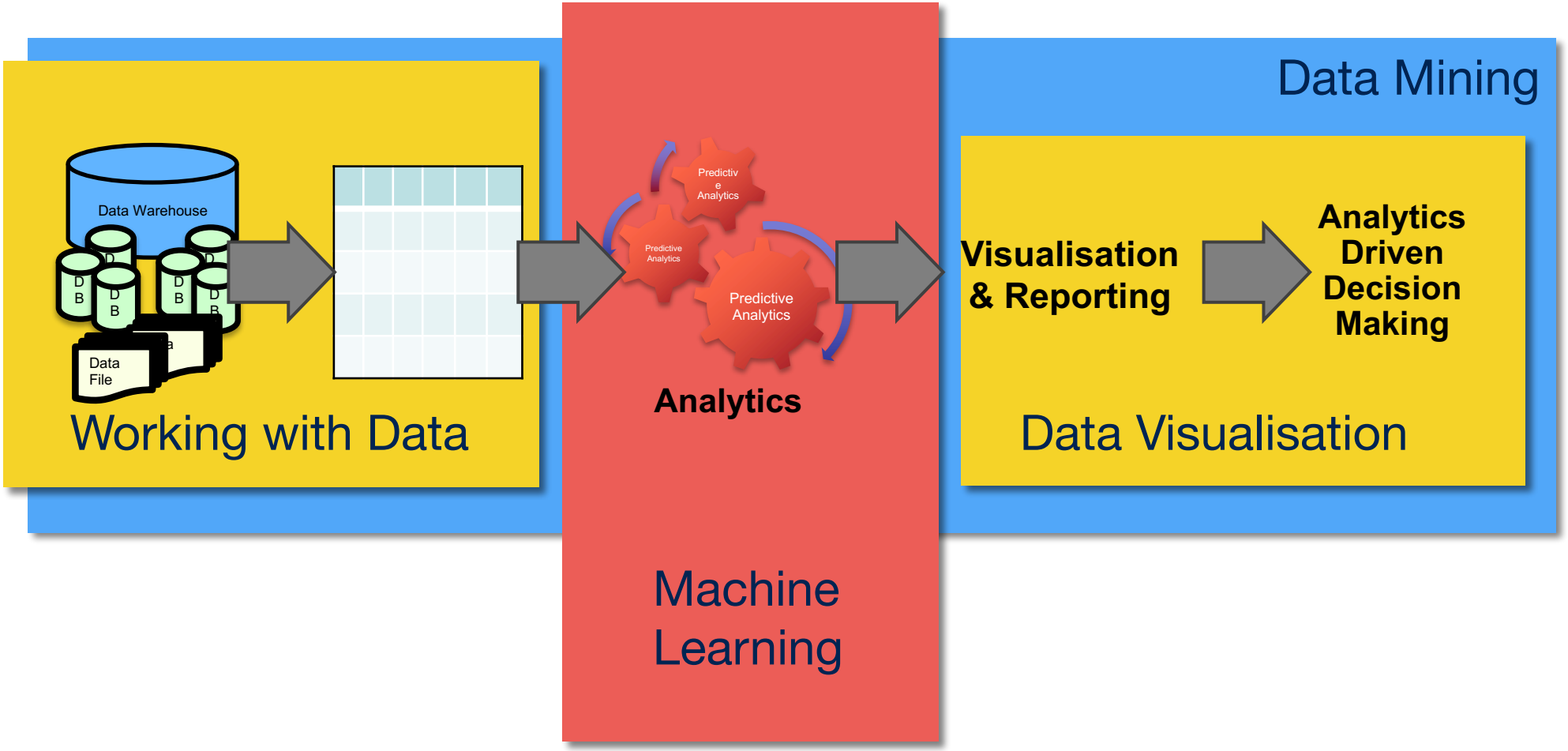
How ML fits into your programme…



How ML fits into your programme…



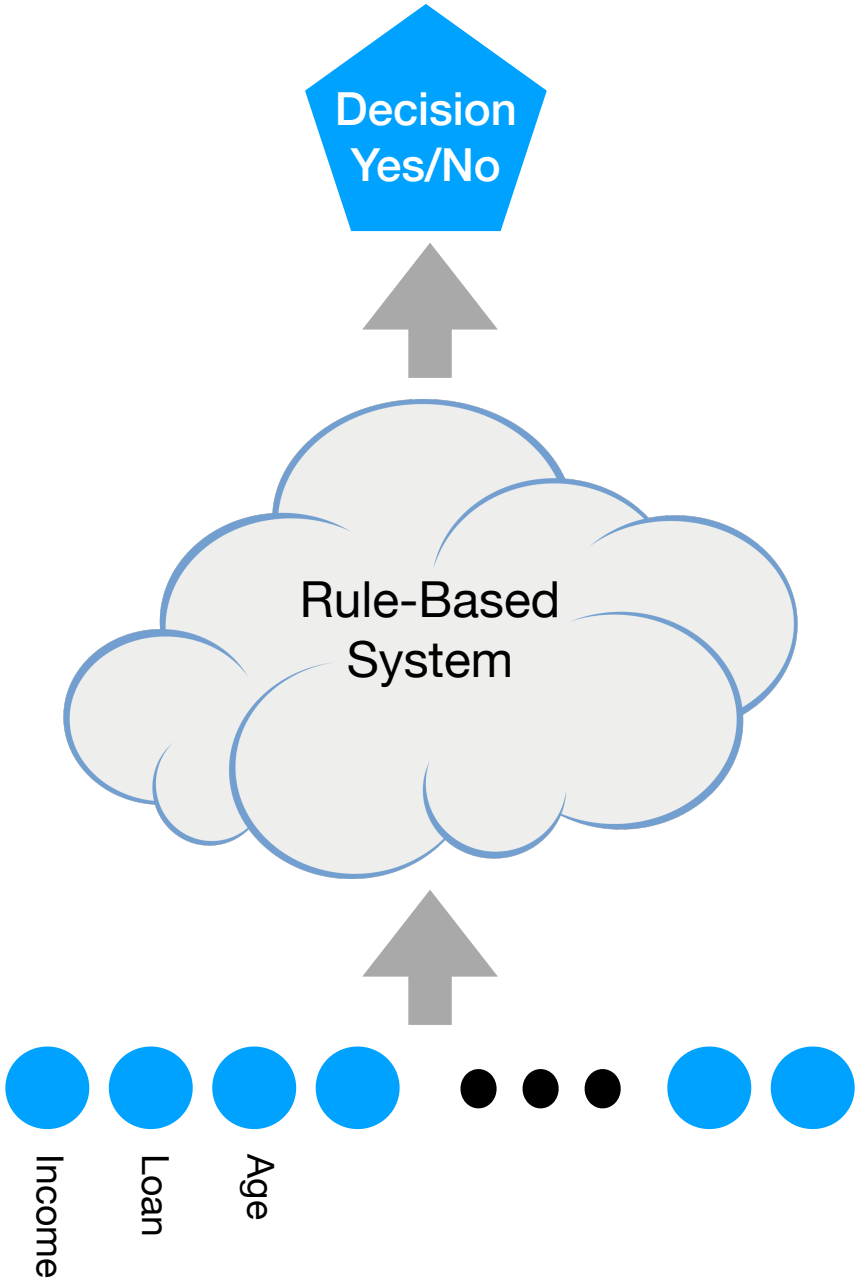
How ML fits into your programme…



Expert Systems (Rule-Based Systems)

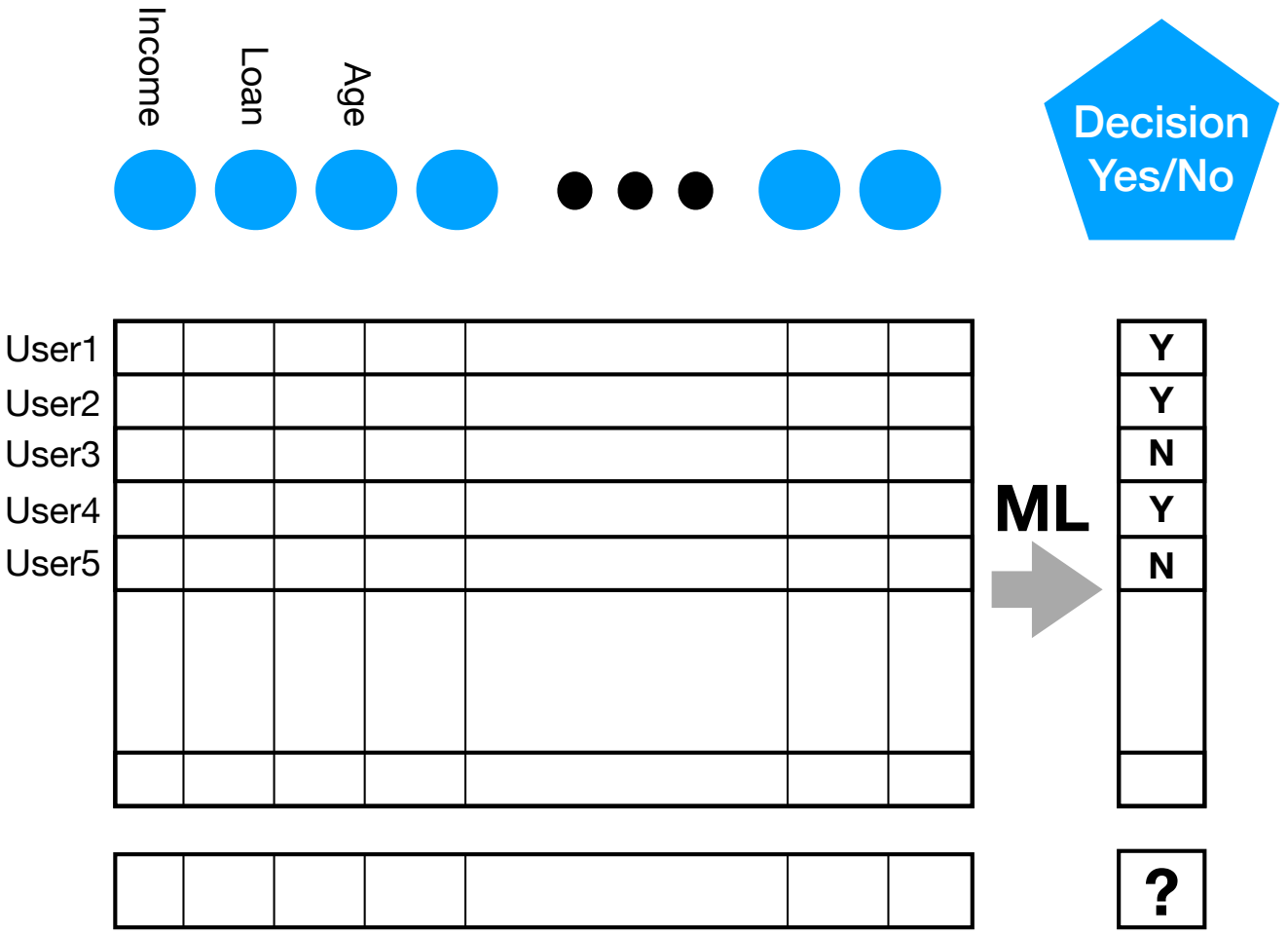
- Rule:
  - If
    - Income > Expenditure &
    - Collateral > Loan
  - Then
    - Risk = Low

Knowledge Engineering



Learning from Historic Data

- Table of historic data
- Each row
  - description of instance
  - decision (Yes/No)



- Predict decision for new instance

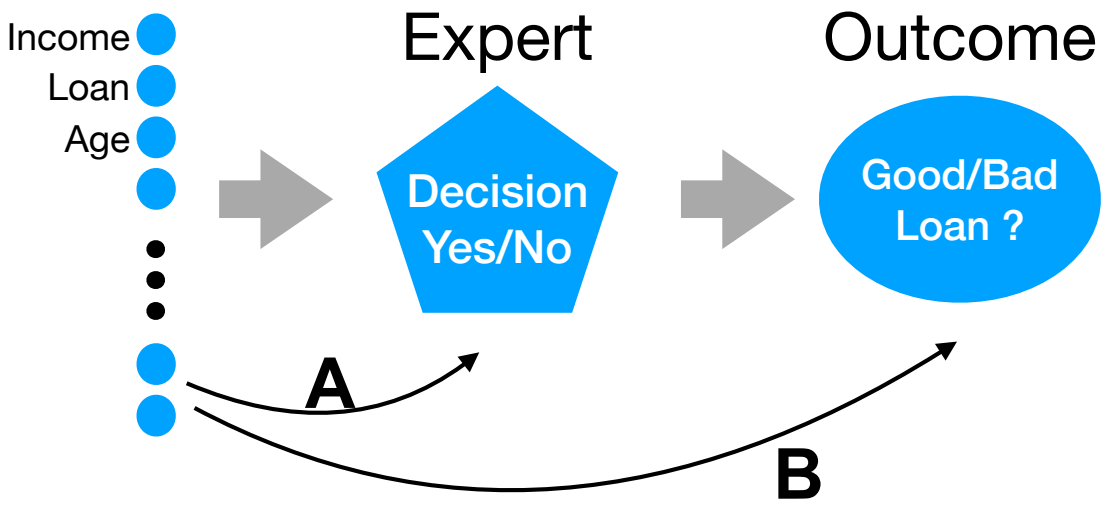
Learn What?

A Learn expert decision making

- What if the expert gets it wrong sometimes?

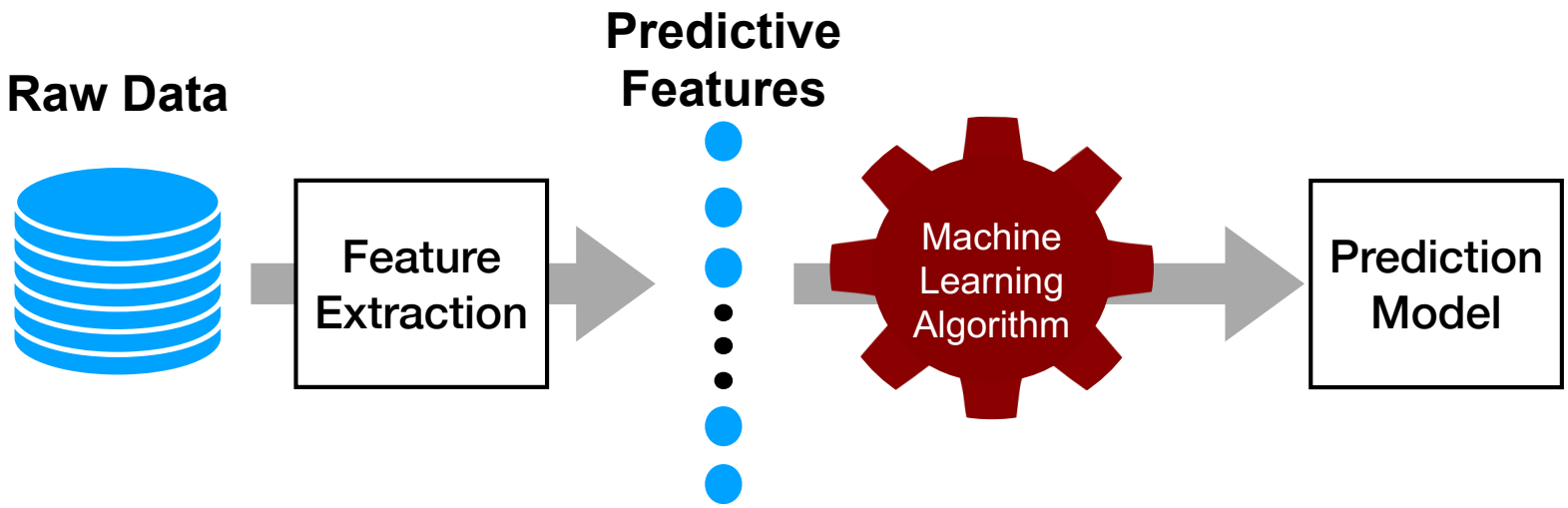
B Learn from outcomes

- Outperform experts



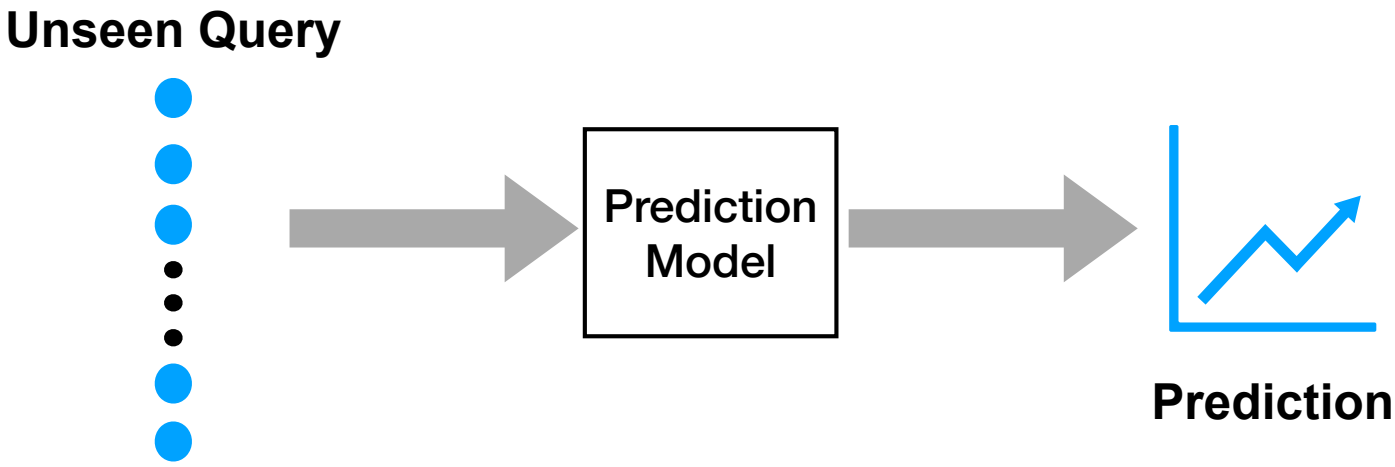
Supervised ML (aka Predictive Analytics)

Training Step:
Learning a model from a set of historical data instances



Supervised ML (aka Predictive Analytics)

Prediction Step:
Using the model to make predictions



Classification Task

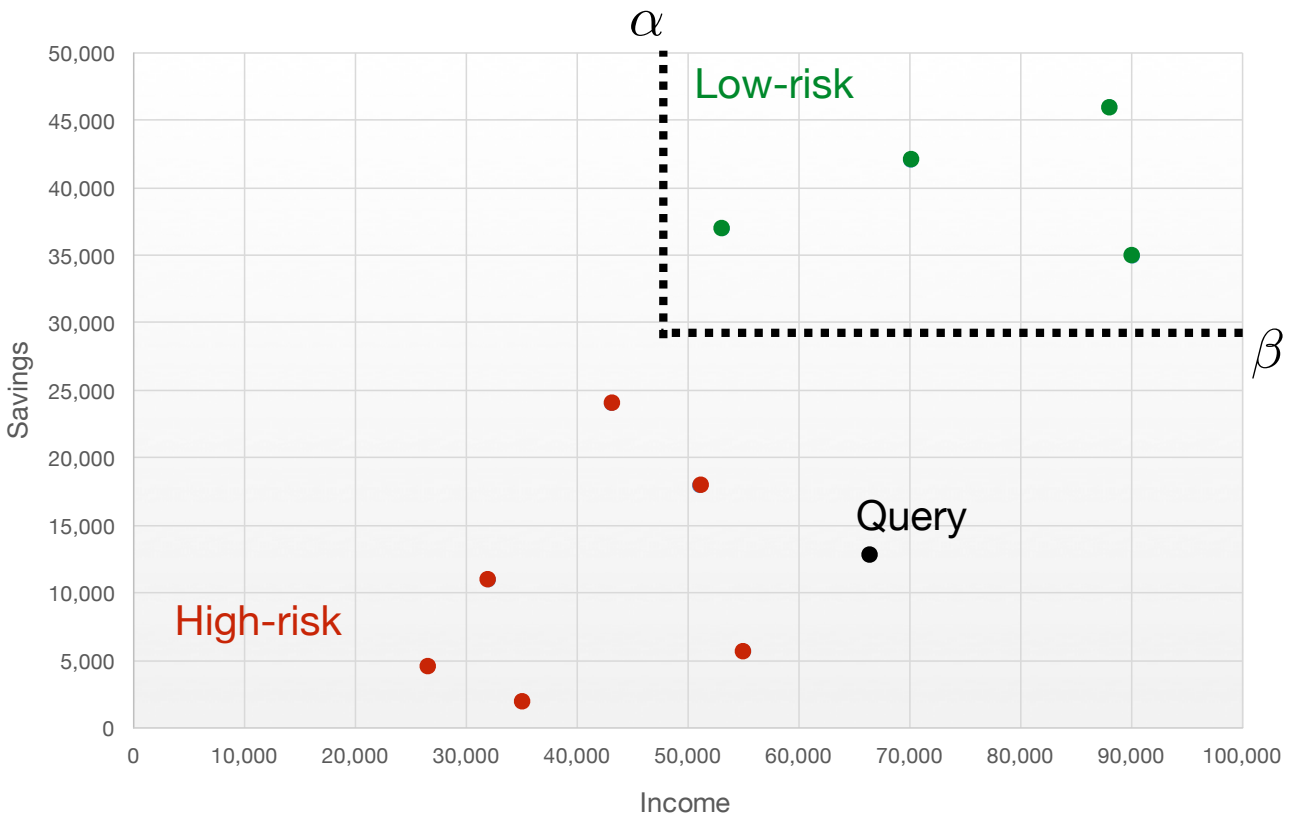
Example:
Credit scoring

- 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

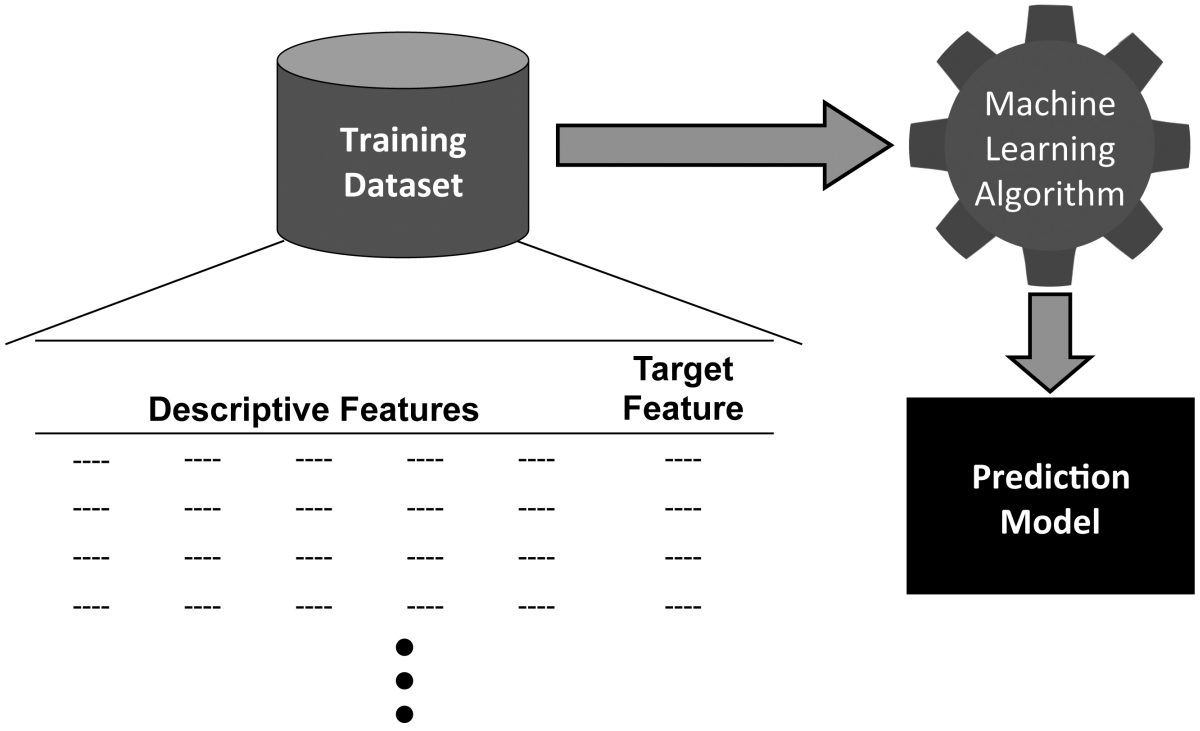
Q. To which class does this new customer belong?

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

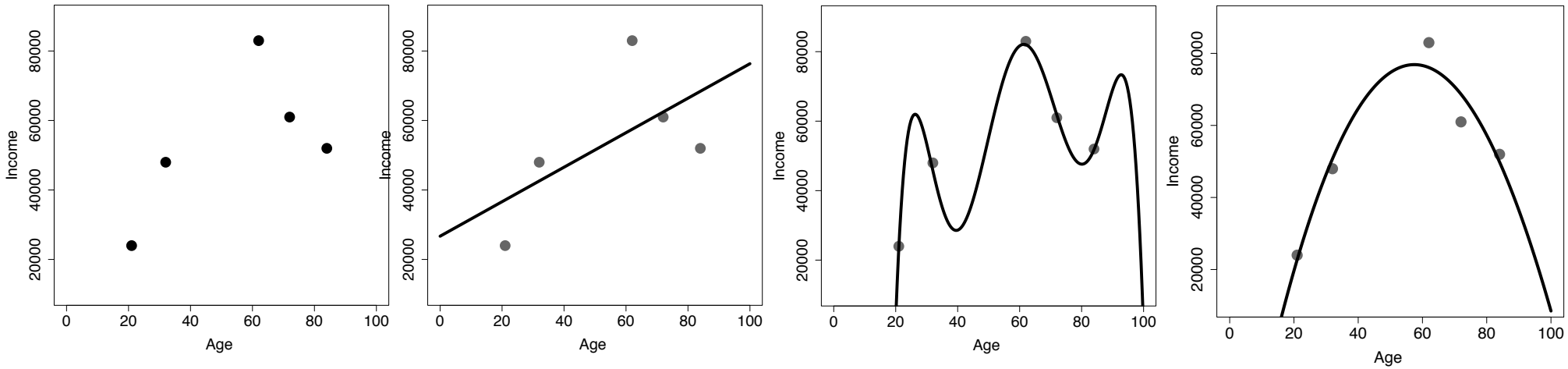


Supervised Learning

- 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 features and the target feature
- Overfitting occurs when the model is so complex that it fits to the data too closely and becomes sensitive to noise (e.g. mislabelled feature values)

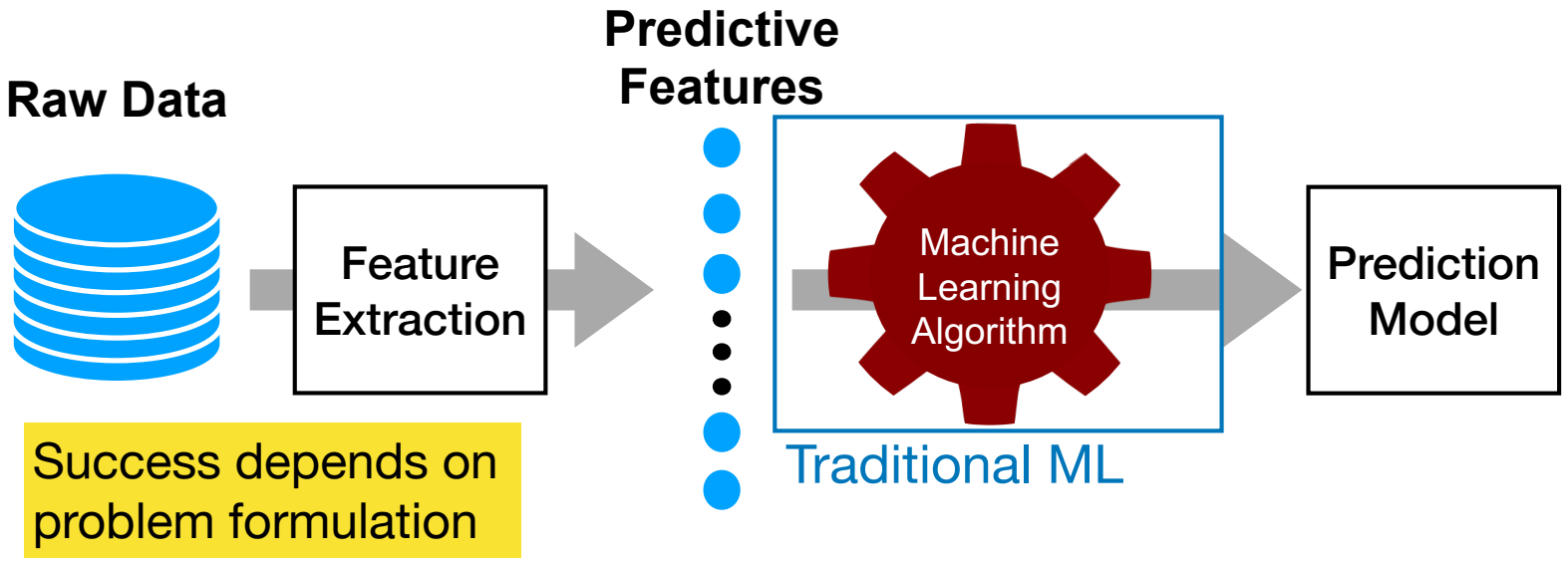


Dataset | Underfitting | Overfitting | Just Right

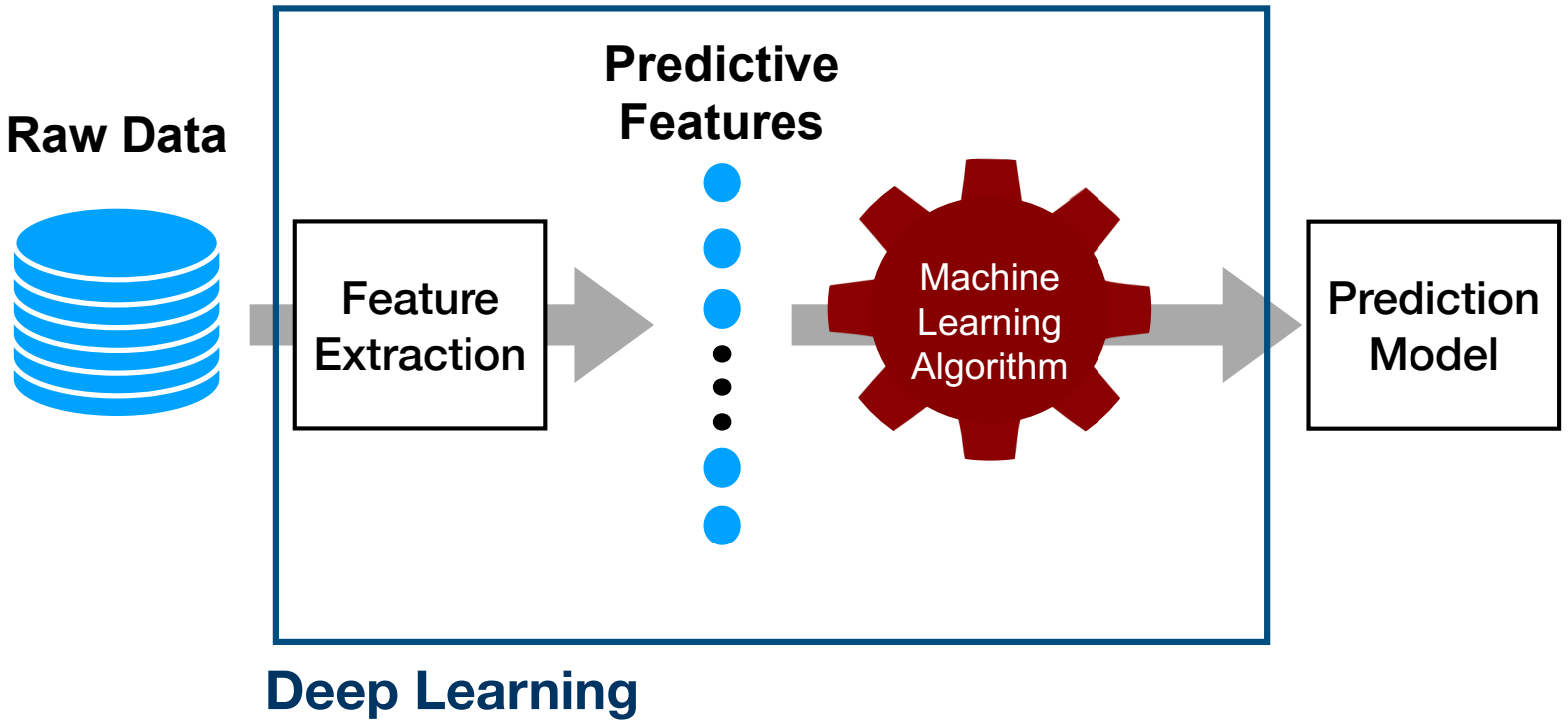
Supervised Learning Algorithms

- Typically not enough information in training data 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. decision tree prefers shallower trees)
- Different algorithms have different inductive bias
- Important to identify the machine algorithm which will fit most appropriately to the predictive task

Supervised ML (aka Predictive Analytics)



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.

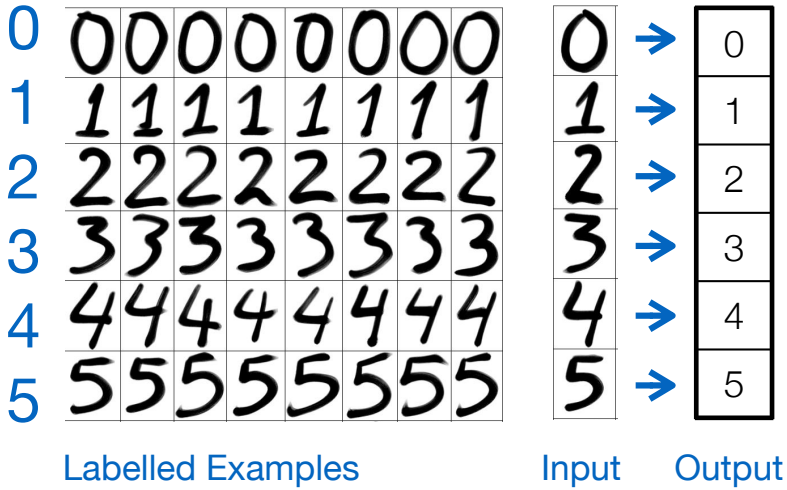
  e.g. Classification, regression algorithms

- 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.

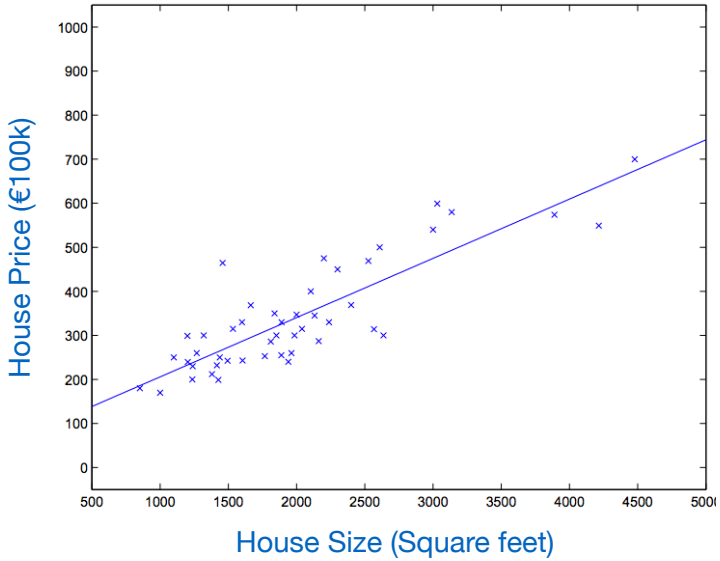
  e.g. Clustering, topic modelling algorithms

Supervised Learning

- **Classification:**
  Examples 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).

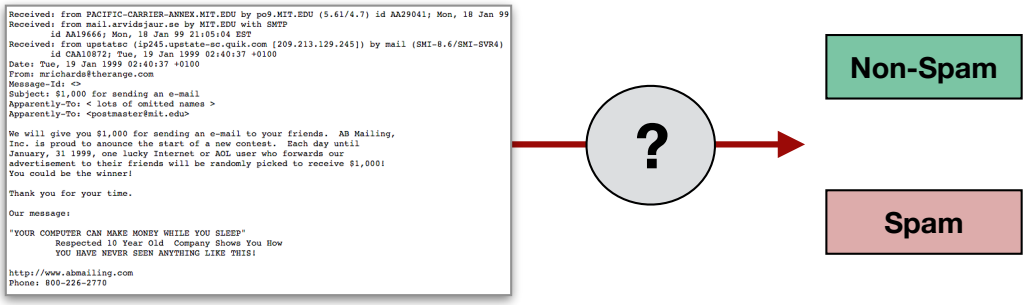


- **Regression:**
  Examples characterised by a set of features, which help decide the value of a continuous output variable
  (i.e. the output is a number).



Classification Tasks

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



Input

2 classes

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

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 is composed of a set of descriptive features.
- For classification, we have training data where each row also has a target class label - i.e. the "correct answer".

	Descriptive Features									Target Class
Examples	------	------	------	------	------	------	------	------	------	------
	------	------	------	------	------	------	------	------	------	------
	------	------	------	------	------	------	------	------	------	------
	------	------	------	------	------	------	------	------	------	------

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

- 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).
- Problem dimensions will often determine which algorithm will be practically applicable, due to processing, memory, and storage constraints.

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, 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

Unsupervised Learning Algorithms
k-Means, Hierarchical clustering
Spectral Clustering

Ensembles
Boosting, Bagging

Dimensionality Reduction
Feature Selection, PCA

The ML Process
Data Preprocessing, Missing Values, Scaling
Model Selection, Hyperparameters
Evaluation

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

Unsupervised Learning

Reinforcement Learning