Machine Learning 101

What Is Important In Machine Learning

- **Business Objective**
- **Input**
- **Output**
- **Model Longevity, Adaptability and Maintenance**
- **Real-time or Batch**
- **Production Requirement & Management**
- **Metrics**

Converting Business Problems into Analytics Solutions

- Converting a business problem into an analytics solution involves answering the following key questions:
  1. What is the business problem?
  2. What are the goals that the business wants to achieve?
  3. How does the business currently work?
  4. In what ways could a predictive analytics model help to address the business problem?

Case Study: Motor Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, a motor insurance company is still losing too much money due to fraudulent claims.

- What predictive analytics solutions could be proposed to help address this business problem?

- Potential analytics solutions include:
    - Claim prediction
    - Member prediction
    - Application prediction
    - Payment prediction

Assessing Feasibility

- Evaluating the feasibility of a proposed analytics solution involves considering the following questions:
  1. Is the data required by the solution available, or could it be made available?
  2. What is the capacity of the business to utilize the insights that the analytics solution will provide?

- What are the data and capacity requirements for the proposed Claim Prediction analytics solution for the motor insurance fraud scenario?

- What are the data and capacity requirements for the proposed Claim Prediction analytics solution for the motor insurance fraud scenario?

Case Study: Motor Insurance Fraud

[Claim prediction]
Data Requirements: A large collection of historical claims marked as *'fraudulent'* and *'non-fraudulent'*. Also, the details of each claim, the related policy, and the related claimant would need to be available.
Capacity Requirements: The main requirement is that a mechanism could be put in place to inform claims investigators that some claims were prioritized above others. This would also require that information about claims become available in a suitably timely manner so that the claims investigation process would not be delayed by the model.

Designing the Analytics Base Table

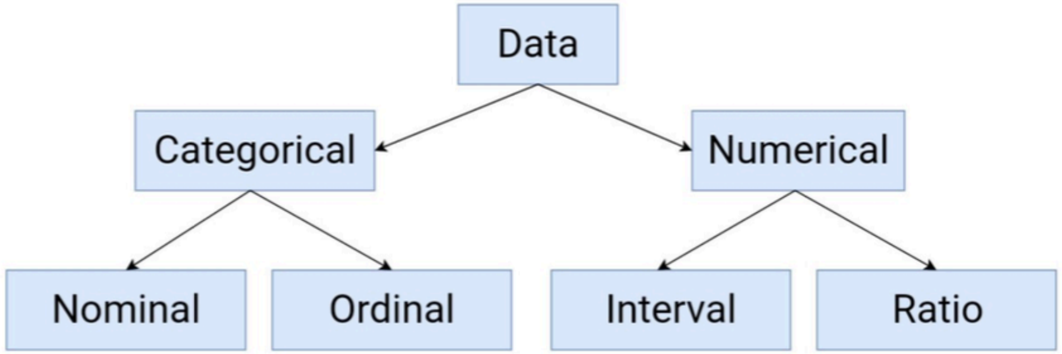
- The basic structure in which we capture historical datasets is the **analytics base table** (**ABT**)

Descriptive Features						Target Feature
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----

Figure: The general structure of an **analytics base table**—descriptive features and a target feature.

Data Science Basics

Structured Data Types



Data Pre-processing

- Data Quality Assessment
- Feature Aggregation
- Feature Sampling
- Dimensionality Reduction
- Feature Encoding

https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

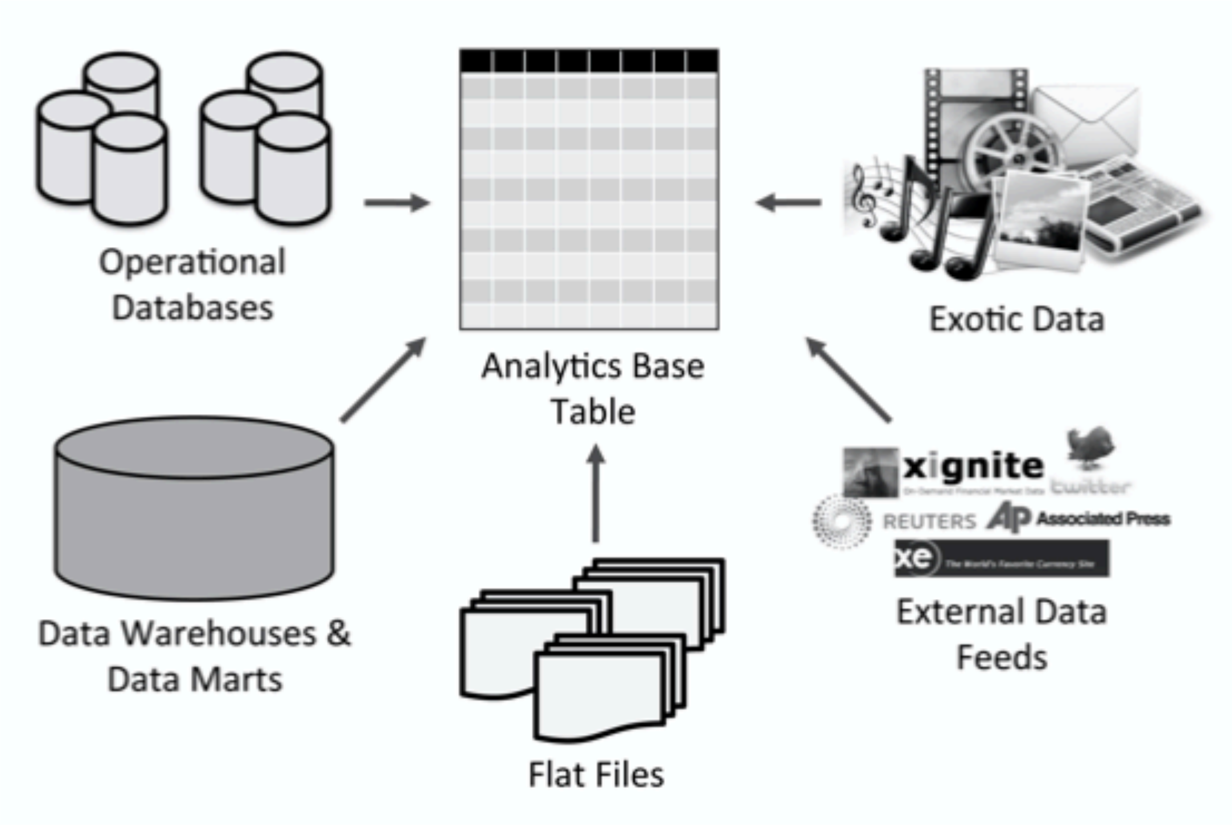
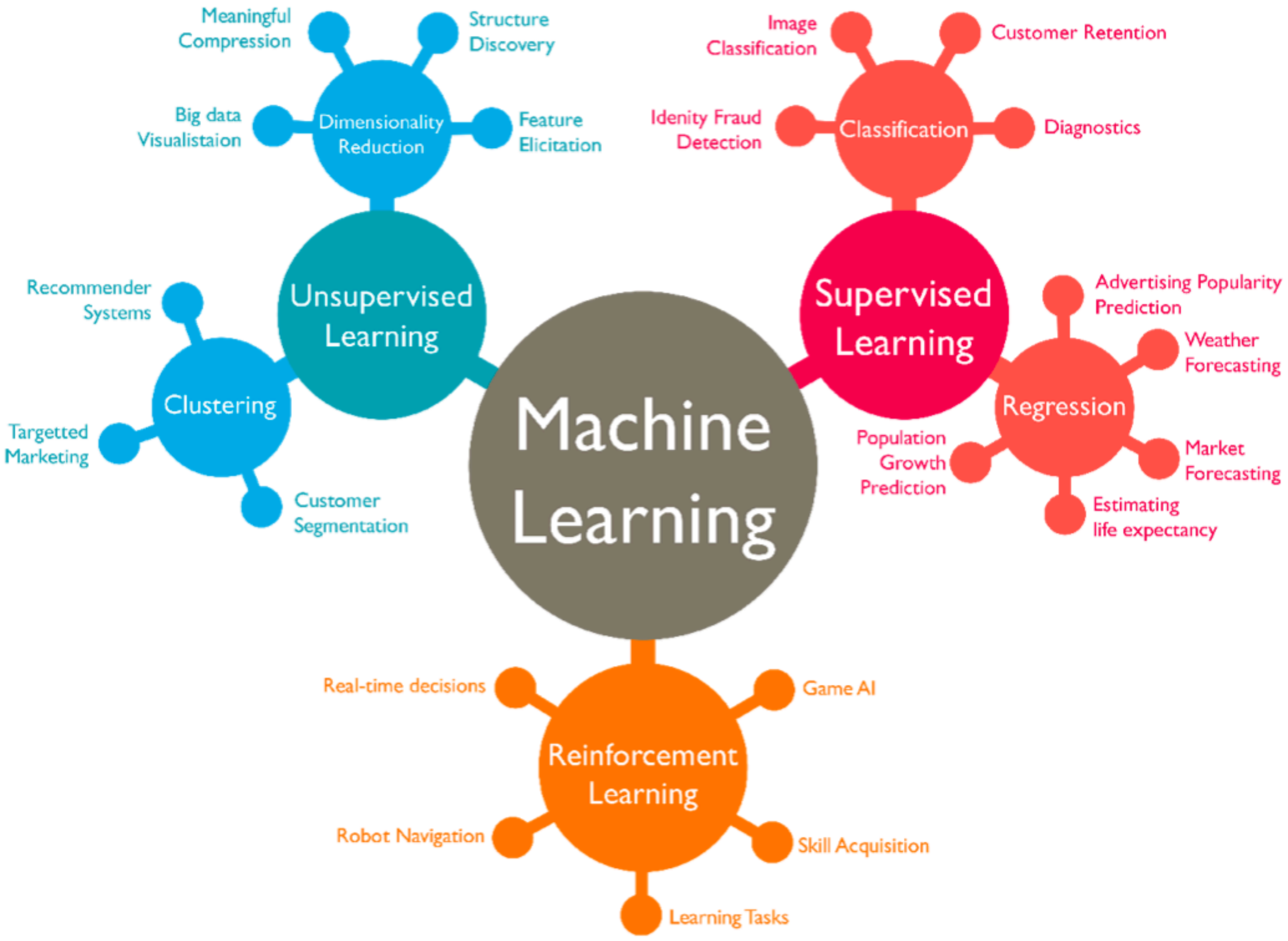


Figure: The different data sources typically combined to create an analytics base table.

Machine Learning



Supervised vs Unsupervised Learning

- Supervised Learning: (x, y)
  - Y are sometimes called labels
  - Objective: P(Y|X)
- Unsupervised Learning: x
  - Objective: f(x)
- Semi-supervised Learning: [X, (x,y)]
  - Large amount of X
  - Learn labels to update/improve model training
  - A few (x,y)
- Active Learning: https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

Supervised Learning

- Regression
- Classification
  - Binary
  - Multi-class
  - Multi-label
  - Multi-task: https://en.wikipedia.org/wiki/Multi-task_learning

Unsupervised Learning

- K-Means
- Hierarchical Clustering
- PCA
- Auto-encoder
- tSNE: https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

Predictive Model Evaluation

- The most important part of the design of an evaluation experiment for a predictive model is ensuring that the data used to evaluate the model is not the same as the data used to train the model.

- The purpose of evaluation is threefold:
  1. to determine which model is the most suitable for a task
  2. to estimate how the model will perform
  3. to convince users that the model will meet their needs

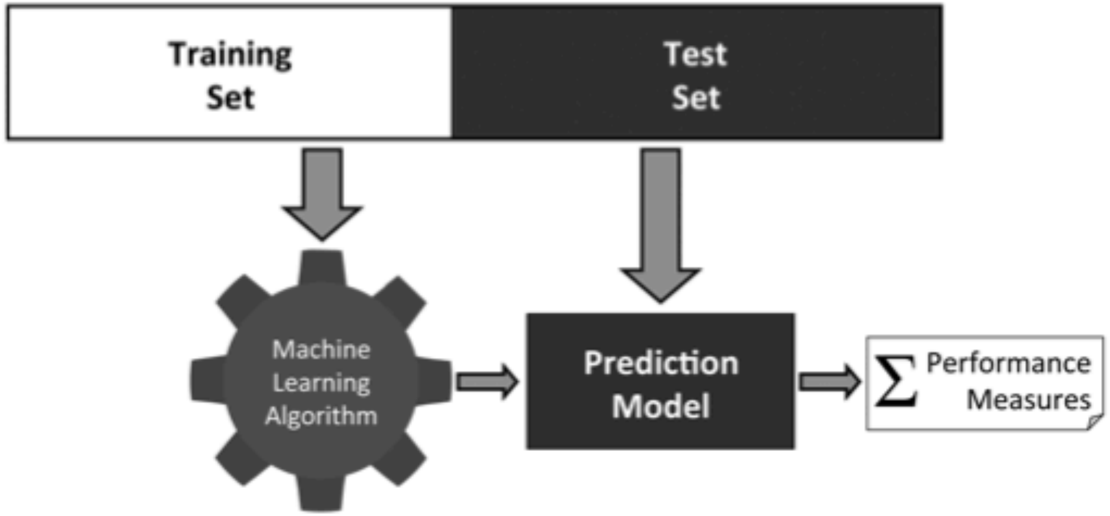


Figure: The process of building and evaluating a model using a **hold-out test set**.

Validation Metrics

- Remember confusion matrix the way you remember your name
- Consider when to use which one(s) and why:
  - Precision
  - Recall
  - Accuracy

Table: A sample test set with model predictions.

ID	Target	Pred.	Outcome	ID	Target	Pred.	Outcome
1	spam	ham	FN	11	ham	ham	TN
2	spam	ham	FN	12	spam	ham	FN
3	ham	ham	TN	13	ham	ham	TN
4	spam	spam	TP	14	ham	ham	TN
5	ham	ham	TN	15	ham	ham	TN
6	spam	spam	TP	16	ham	ham	TN
7	ham	ham	TN	17	ham	spam	FP
8	spam	spam	TP	18	spam	spam	TP
9	spam	spam	TP	19	ham	ham	TN
10	spam	spam	TP	20	ham	spam	FP

- For binary prediction problems there are 4 possible outcomes:
  1. True Positive (TP)
  2. True Negative (TN)
  3. False Positive (FP)
  4. False Negative (FN)

Table: The structure of a confusion matrix.

		Prediction	
		positive	**negative**
Target	**positive**	*TP*	*FN*
	negative	*FP*	*TN*

Table: A confusion matrix for the set of predictions shown in Table 1 [7].

		Prediction	
		'spam'	*'ham'*
Target	*'spam'*	6	3
	'ham'	2	9

Model Validation

Confusion Matrix

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

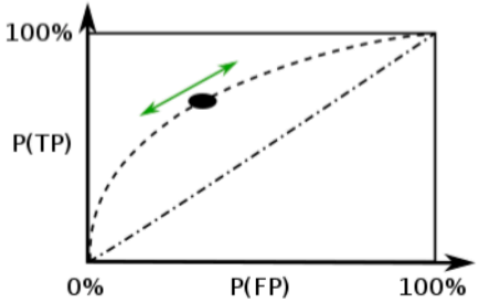
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

ROC and AUC

- Receiver Operating Characteristic
- Area Under the Curve



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