DATA MINING 1
Classification

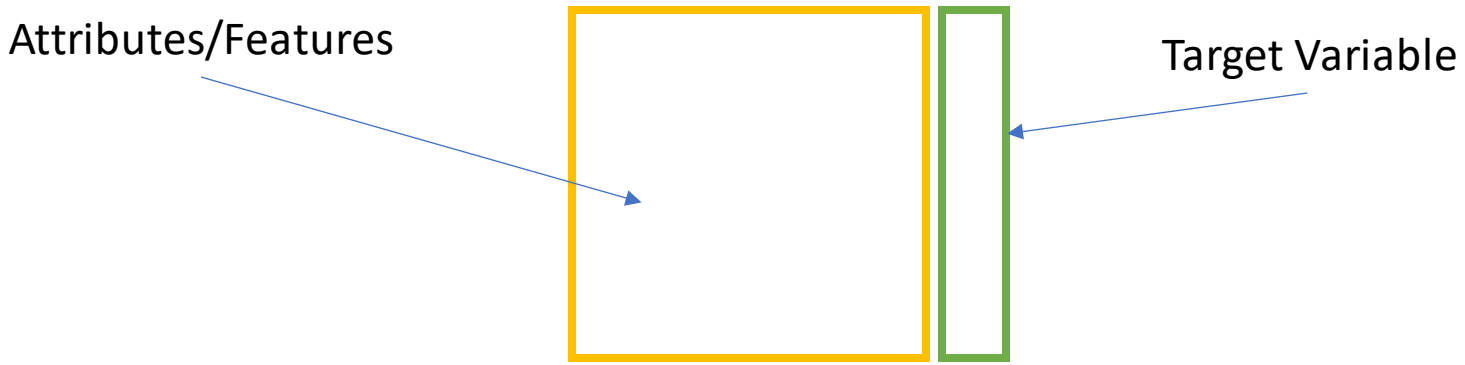
Dino Pedreschi, Riccardo Guidotti

Revisited slides from Lecture Notes for Chapter 3 "Introduction to Data Mining", 2nd Edition by Tan, Steinbach, Karpatne, Kumar

UNIVERSITÀ DI PISA

Supervised Learning

- Cluster analysis and association rules are not concerned with a specific target attribute.
- Supervised learning refers to problems where the value of a target attribute should be predicted based on the values of other attributes.
- Problems with a ***categorical target*** attribute are called **classification**, problems with a ***numerical target*** attribute are called **regression**.



What is Machine Learning?

- Machine Learning (ML) is the science (and art) of programming computers that can learn from data.

"ML is the field of study that gives computers the ability to learn without being explicitly programmed"
(Arthur Samuel, 1959)

What is Machine Learning?

- Machine Learning (ML) is the science (and art) of programming computers that can learn from data.

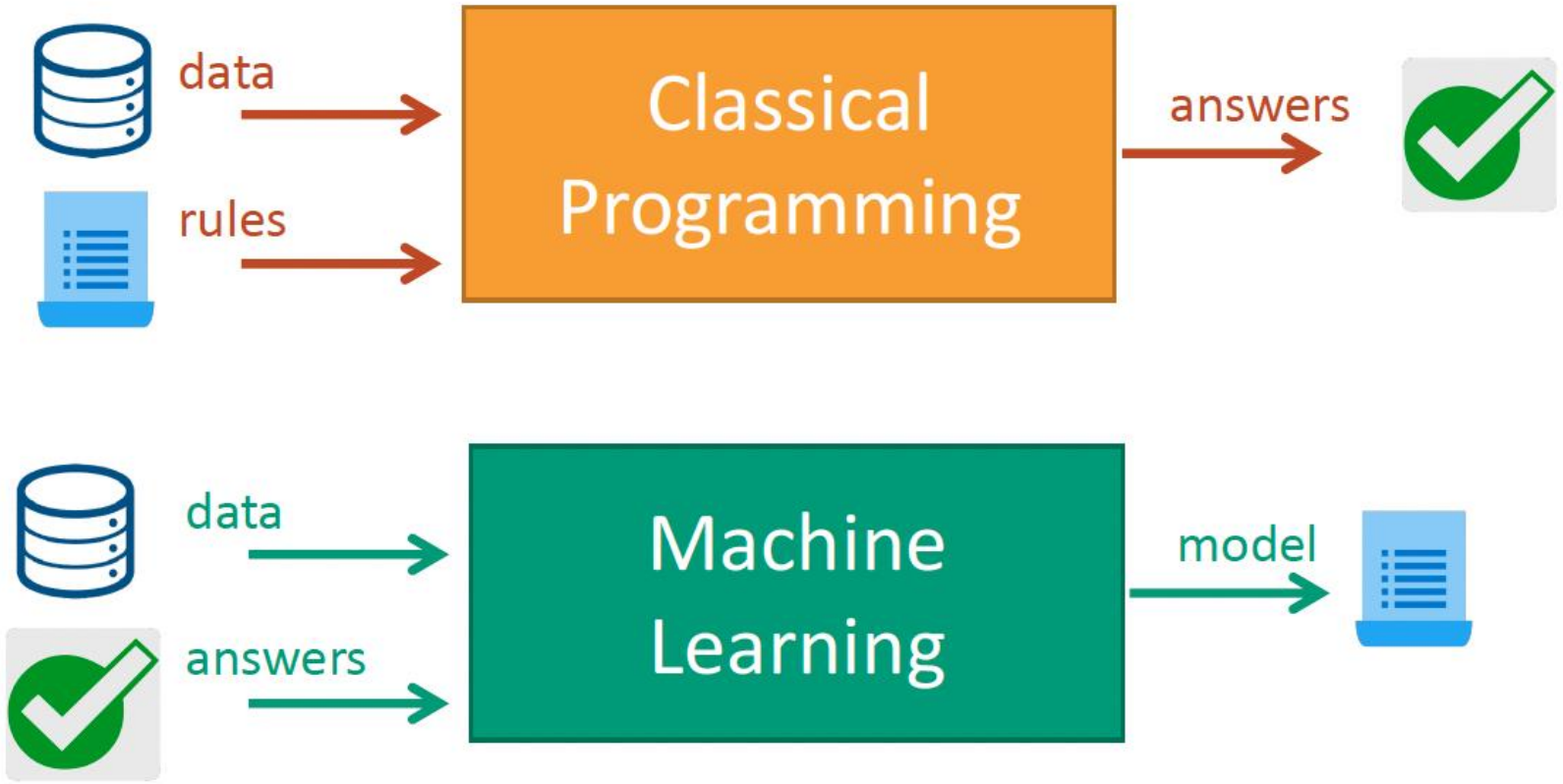
"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
(Tom Mitchell, 1997)

Classical Programming vs Machine Learning

- A ML system is trained rather than programmed

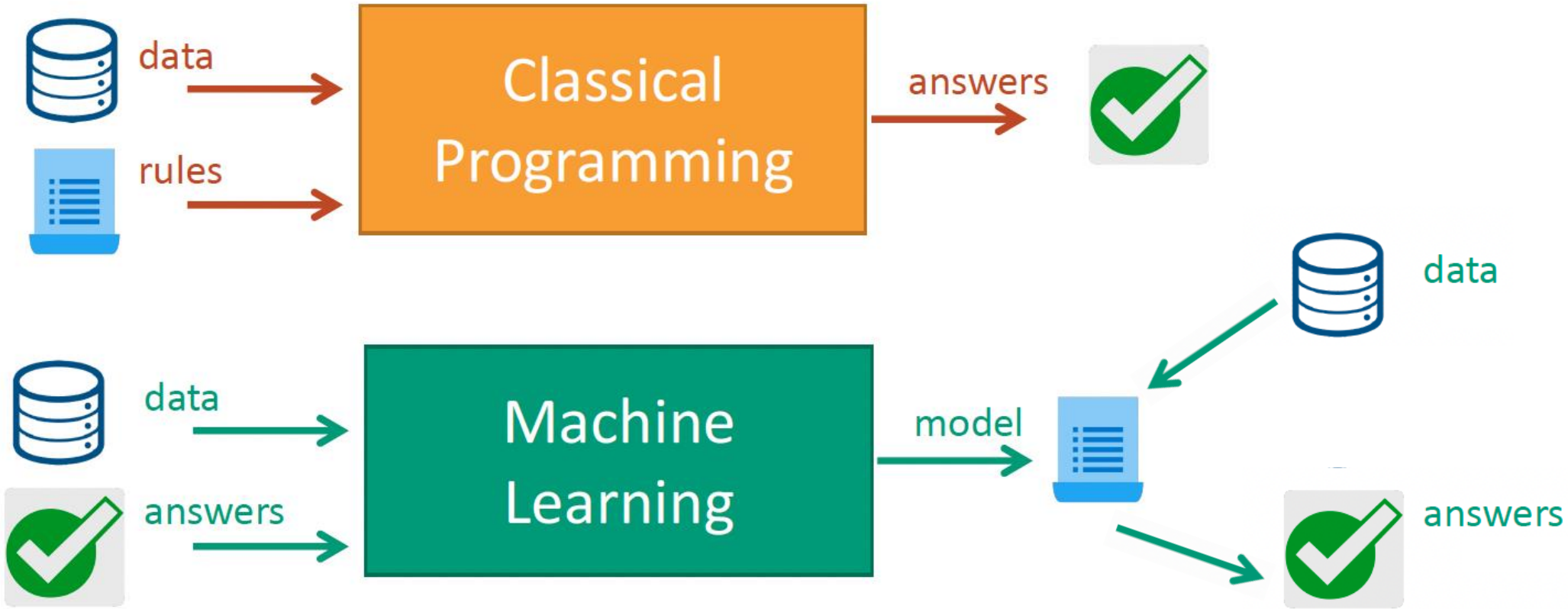
Classical Programming vs Machine Learning

- A ML system is trained rather than programmed

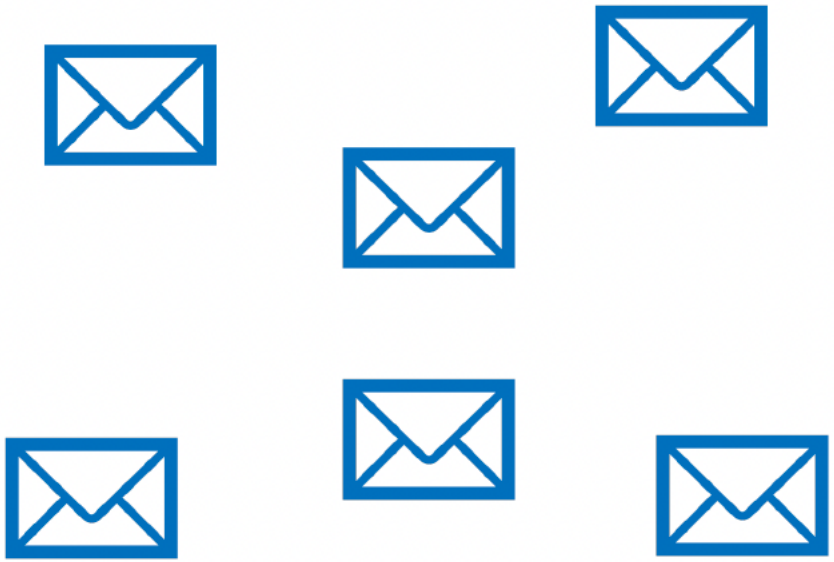


Classical Programming vs Machine Learning

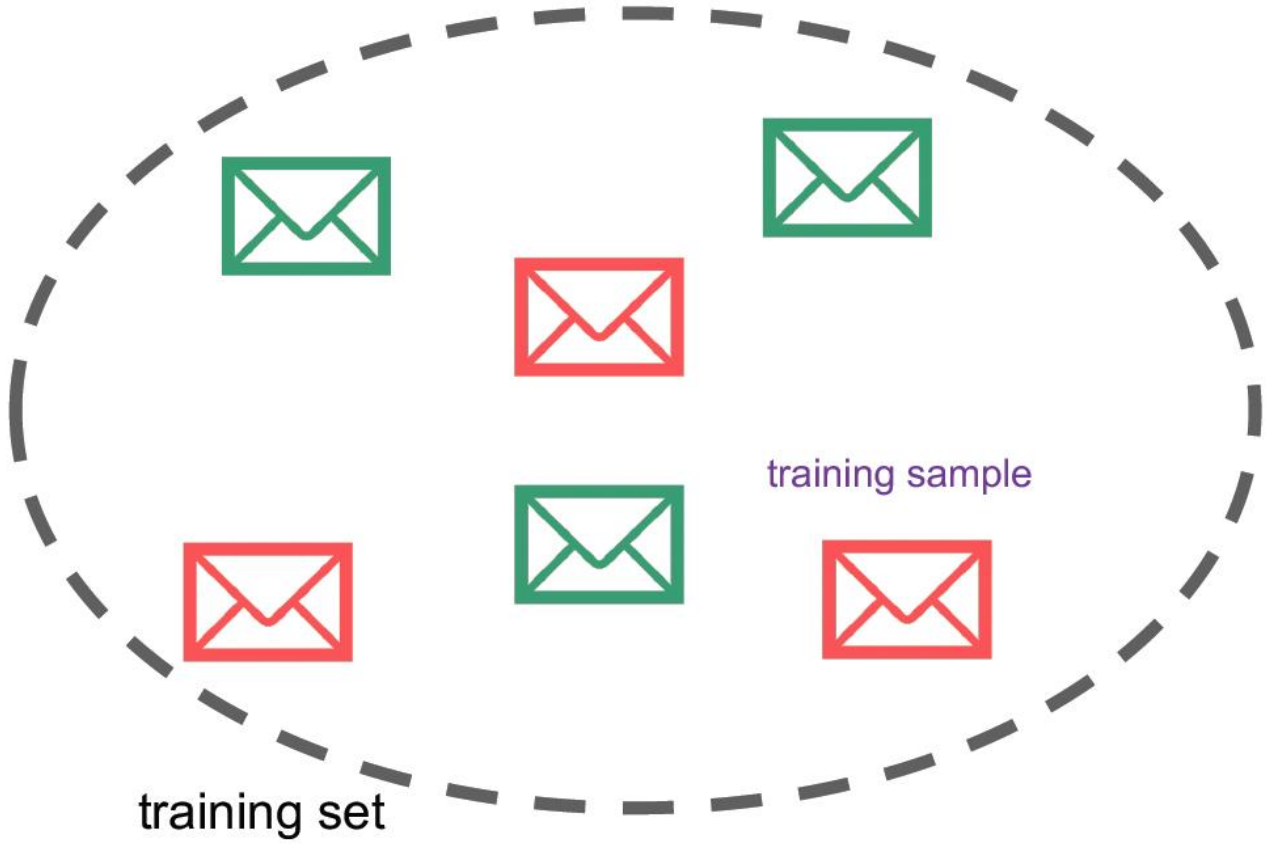
- A ML system is trained rather than programmed

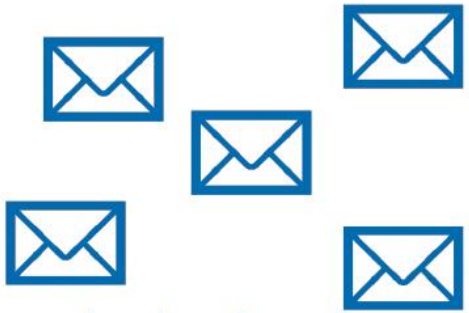


Example Spam Filter



Example Spam Filter



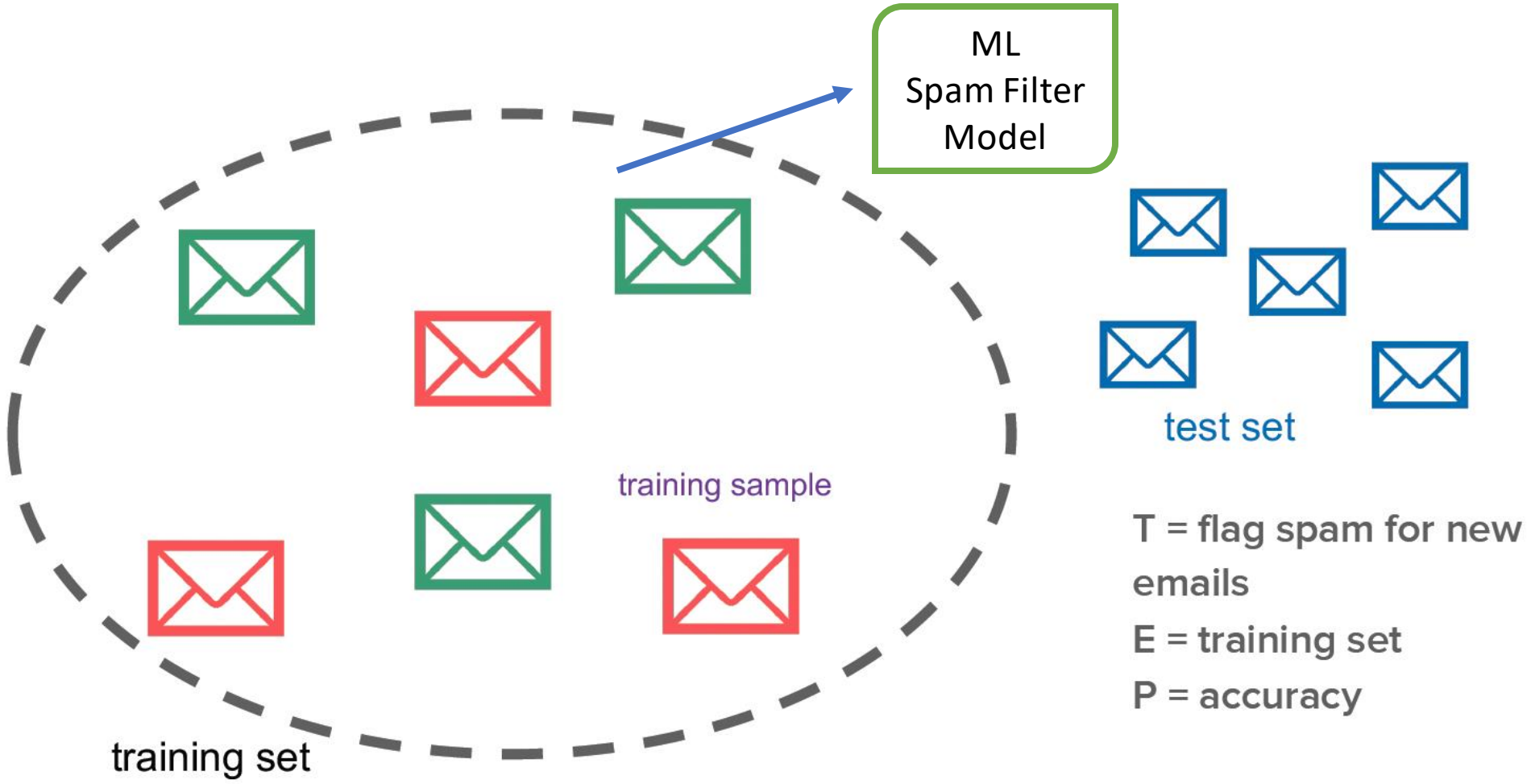


T = flag spam for new emails

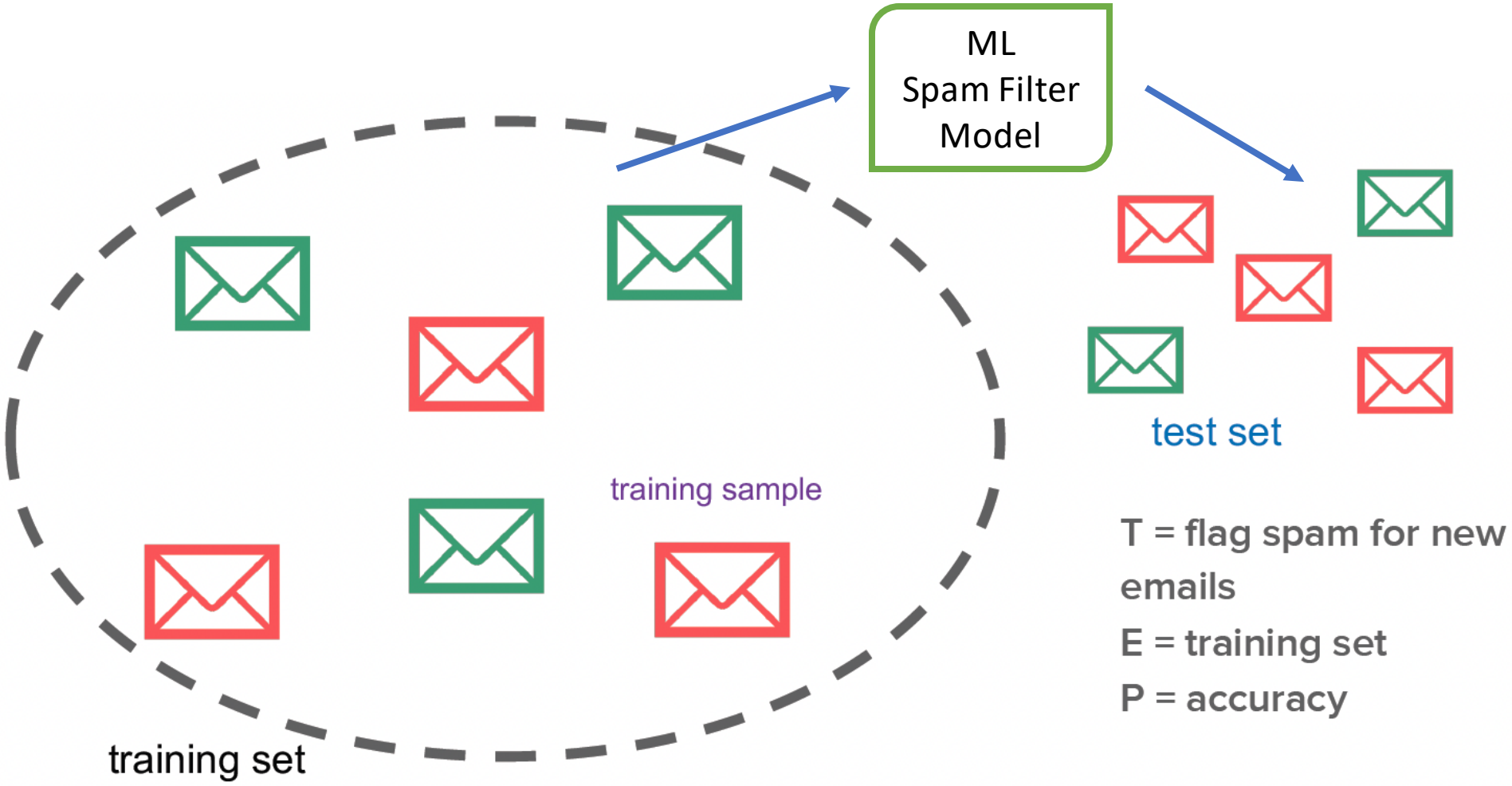
E = training set

P = accuracy

Example Spam Filter



Example Spam Filter

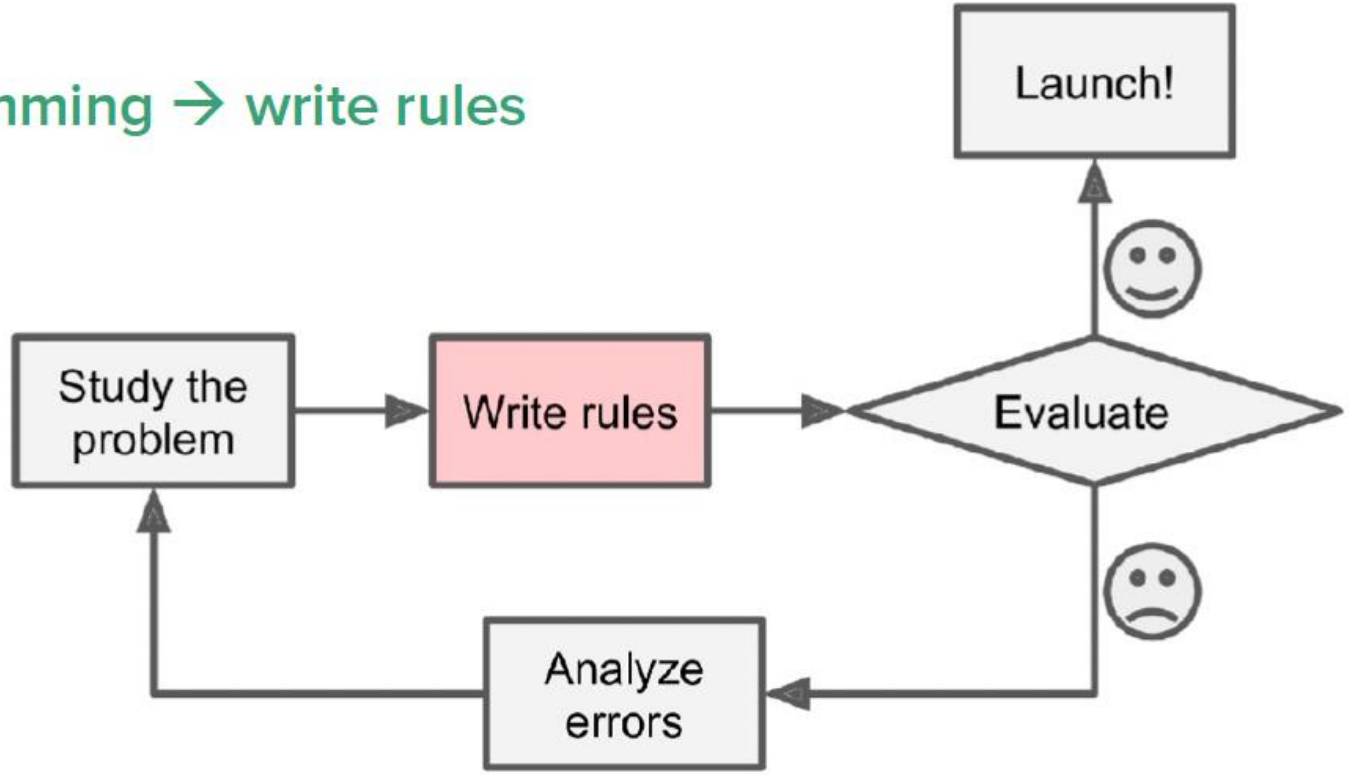


Examples of Classification Task

Task	Attribute set, $\boldsymbol{x}$	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

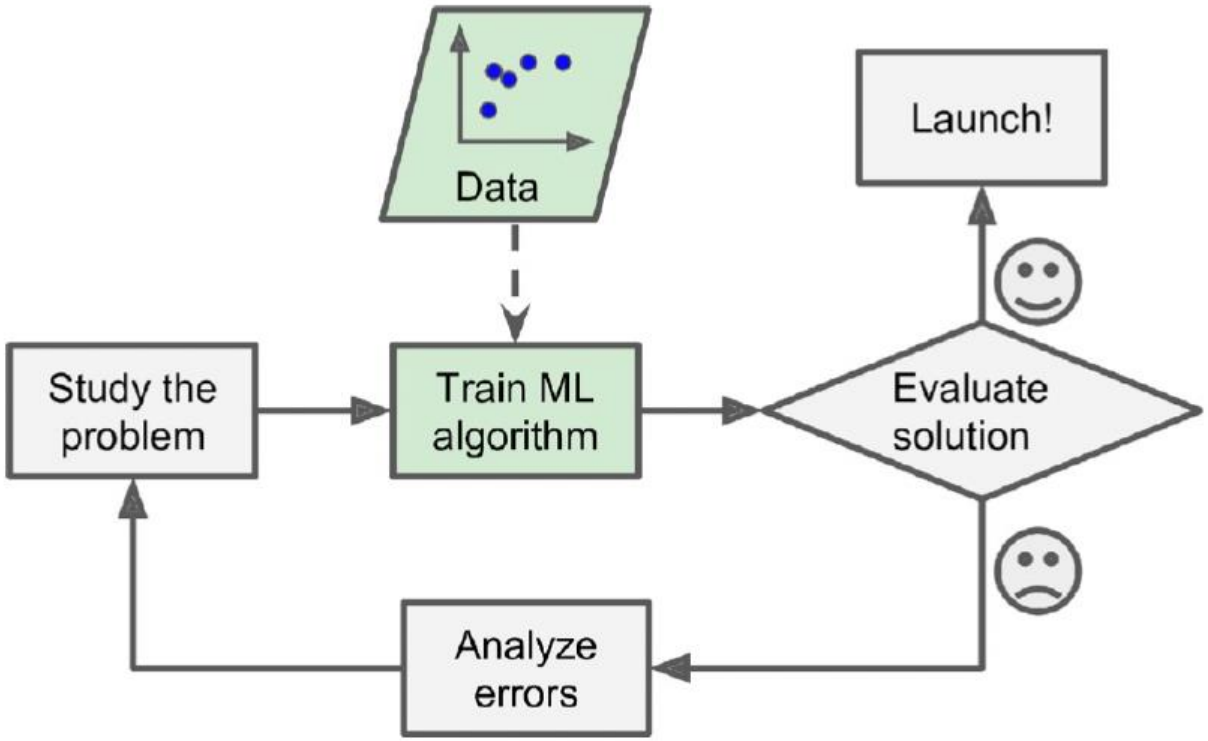
Why do we want to use Machine Learning?

Traditional Programming → write rules



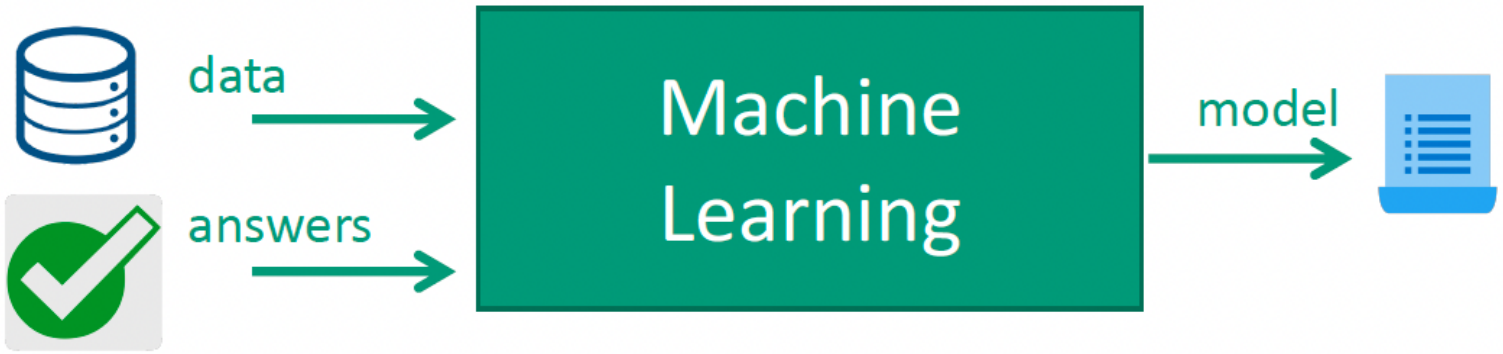
Why do we want to use Machine Learning?

Machine Learning: train based on data (examples)



Why do we want to use Machine Learning?

- Problems for which existing solutions require a lot of finetuning or a long list of rules
- Complex problems for which a traditional approach yields no good solution
- Changing environments
- Getting insights about complex problems and large amount of data



What is Classification?

- Classification consists in learning a **model/function *f*** that maps each attribute set ***x*** into one of the predefined class labels ***y***: *f(x) = y*

What is Classification?

- Classification consists in learning a **model/function *f*** that maps each attribute set ***x*** into one of the predefined class labels ***y***: *f(x) = y*
- The ***learning*** is performed on a given a collection of records named ***training set*** where each record is by characterized by a tuple ***(x, y)***, where ***x*** is the attribute set and ***y*** is the class label
  - x: attribute, predictor, independent variable, input
  - y: class, response, dependent variable, output

What is Classification?

- Classification consists in learning a **model/function *f*** that maps each attribute set ***x*** into one of the predefined class labels ***y***: *f(x) = y*
- The ***learning*** is performed on a given a collection of records named ***training set*** where each record is by characterized by a tuple ***(x, y)***, where ***x*** is the attribute set and ***y*** is the class label
  - x: attribute, predictor, independent variable, input
  - y: class, response, dependent variable, output
- **Goal**: previously unseen records should be assigned a class as accurately as possible. A ***test set*** is used to determine the accuracy of the model ***f***.

What is Classification?

- Classification consists in learning a **model/function *f*** that maps each attribute set ***x*** into one of the predefined class labels ***y***: *f(x) = y*
- The ***learning*** is performed on a given a collection of records named ***training set*** where each record is by characterized by a tuple ***(x, y)***, where ***x*** is the attribute set and ***y*** is the class label
  - x: attribute, predictor, independent variable, input
  - y: class, response, dependent variable, output
- **Goal**: previously unseen records should be assigned a class as accurately as possible. A ***test set*** is used to determine the accuracy of the model ***f***.
- Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to evaluate it.

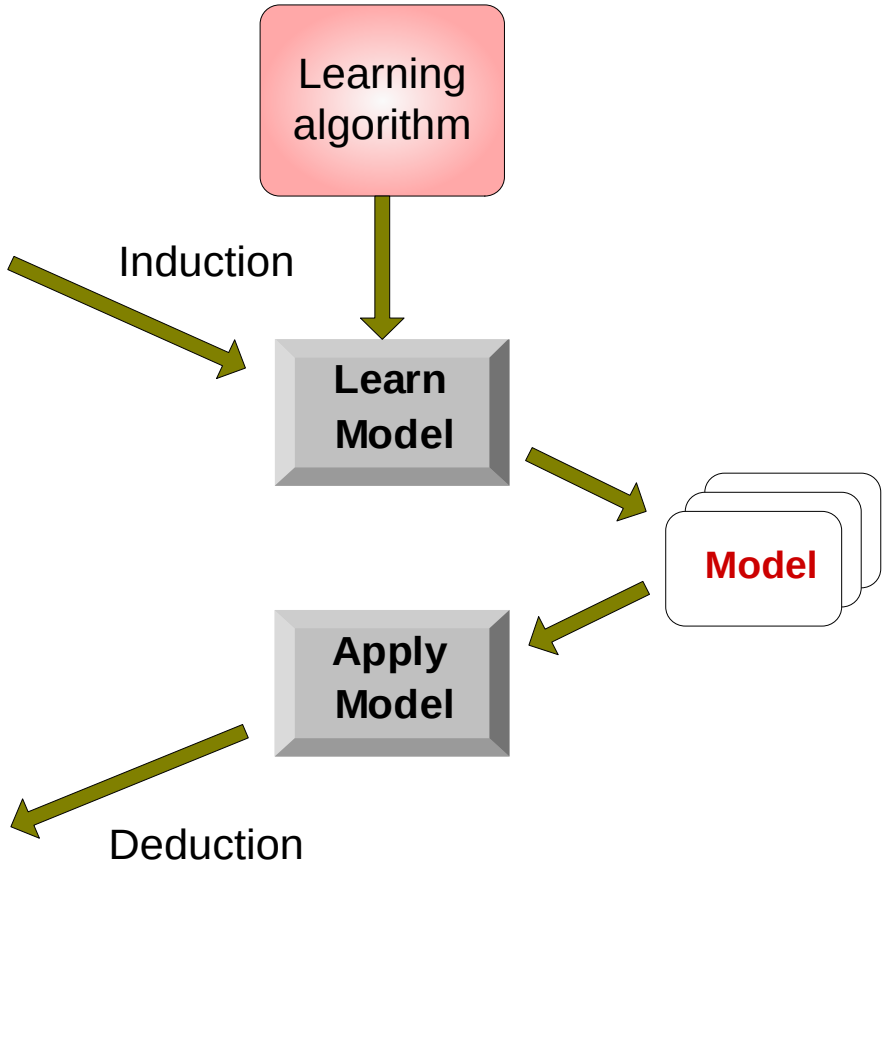
General Approach for Building Classification Model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

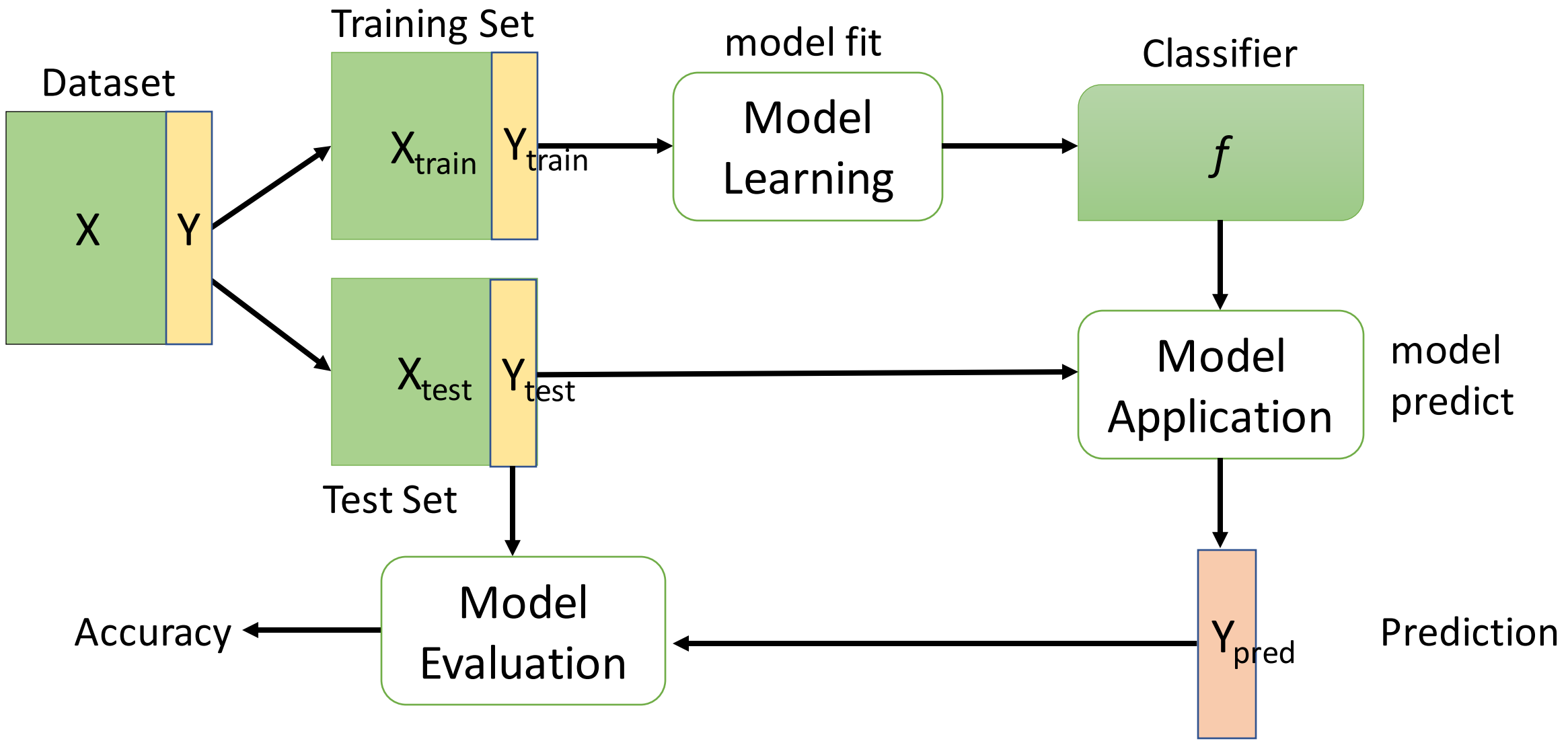
Test Set



Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines
- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

What is Classification?



References

- Chapter 3. Classification: Basic Concepts and Techniques.