ENSF 612: Fall 2021
Lecture - Machine Learning (ML) Basic Concepts

Dr. Gias Uddin, Assistant Professor
Electrical and Software Engineering
Schulich School of Engineering
University of Calgary
https://giasuddin.ca/

Topics

- Definition
- Example applications
- Terminology
- Types of Machine Learning
- Typical supervised learning pipeline

Machine Learning (ML) - Definition

Andriy Burkov (100-Page ML Book)

"Machine learning is a subfield of computer science that is concerned with building algorithms which, to be useful, rely on a collection of examples of some phenomenon. These examples can come from nature, be handcrafted by humans or generated by another algorithm."

"ML is also defined as the process of solving a practical problem by: 1) gathering a dataset and 2) algorithmically building a statistical model based on that dataset. That statistical model is assumed to be used somehow to solve the practical problem."

Integrates ideas from many disciplines

- Computer science
- Probability and statistics
- Optimization
- Linear algebra

Machine Learning (ML) - Examples

- Google's ranking of Web pages
- Automatic photo tagging via face recognition
- Spam filtering
- Games – IBM's Deep Blue computer
- Recipes – IBM's Watson invents new recipes with ML!

Machine Learning (ML) - Terminology

- Observations
  - Items or entities used for learning or evaluation
  - E.g., emails
- Features
  - Attributes (numeric) used to encode observation
  - E.g., length of email, date, presence of keywords
- Labels
  - Values/categories assigned to observation
  - E.g., spam or not spam
- Training, validation, and test data
  - Training – data given to algorithm for training
  - Validation – data used to select algorithm parameters
  - Test – data withheld during training and validation

Machine Learning (ML) - Types

- **Supervised Learning**
- **Unsupervised Learning**
- **Semi-Supervised Learning**
- **Transductive Learning**
- Deep Learning
- Reinforcement Learning

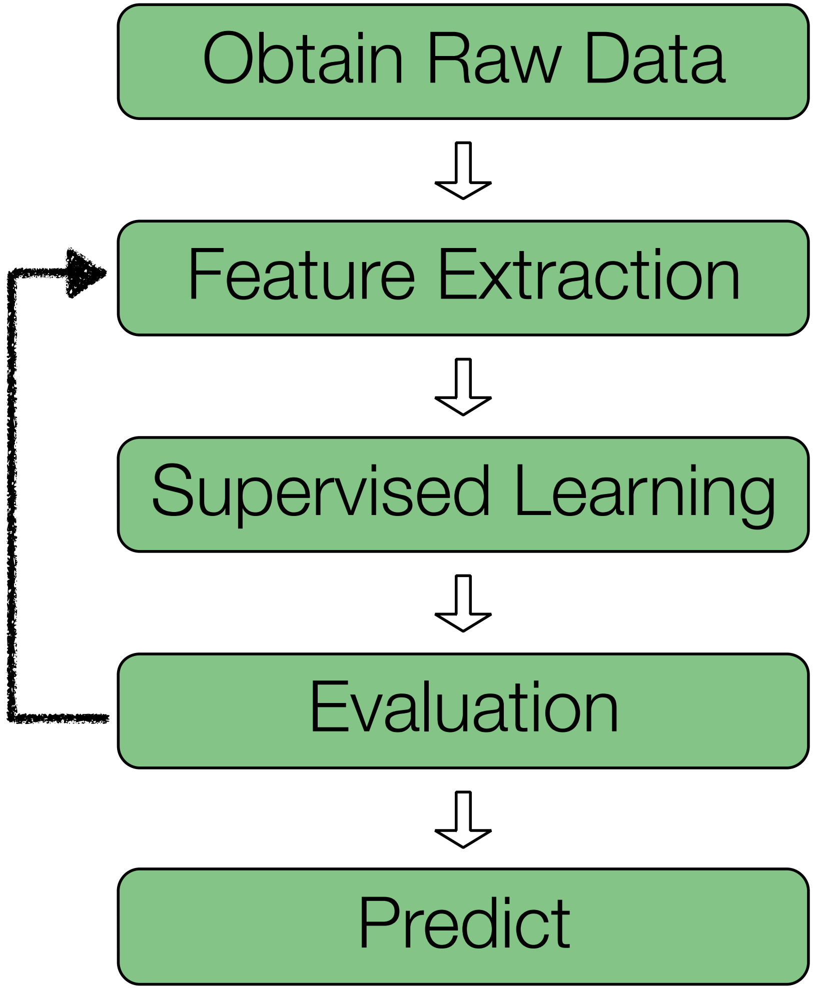
Supervised and Unsupervised ML

- Supervised – learning from labeled observations
  - Labels teach the algorithm how to map observations to labels
  - Examples
    - Classification – Assign a discrete category to each item
    - E.g., Spam filtering, alphabet recognition
    - Regression – Predict real value for each item
    - Labels are continuous
    - Can define closeness of prediction to label
    - E.g., Predicting Deerfoot trail commute times, stock prices

Supervised and Unsupervised ML

- Unsupervised – learning from unlabeled observations
  - Algorithm should find latent structure from features alone
  - Examples
    - Clustering – partition observations into homogeneous regions
    - E.g., discover hidden traffic patterns on Deerfoot Trail
    - E.g., discover "communities" in Facebook
    - Dimensionality reduction –
      - Change initial feature representation to a more concise one
    - E.g., More concise representation of images

Supervised ML – Typical Pipeline



Supervised ML – Typical Pipeline

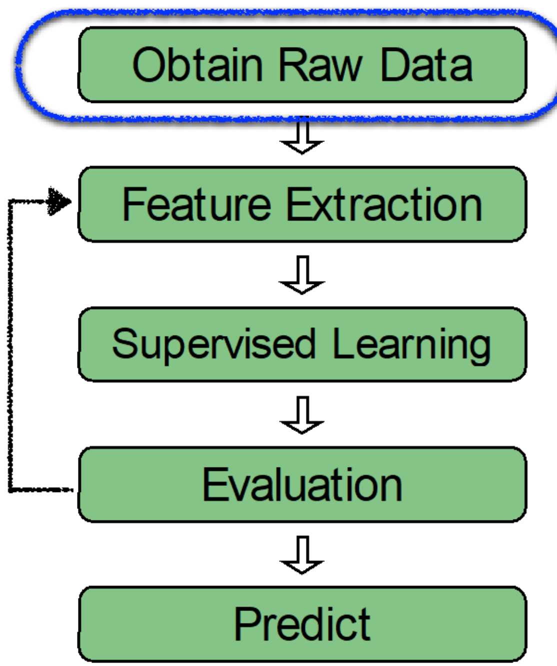
- **Obtain raw data**
  - E.g., emails, log files, sensor data, handwriting digital samples
- **Extract features**
  - Convert raw observation to numeric features
  - Needs domain knowledge and/or unsupervised learning
  - Effectiveness of pipeline depends heavily on this step!
- **Train** supervised learning algorithm using training data
- **Evaluate** learning algorithm using validation data
- Iterate till you're happy with model
- **Predict** using the trained model and test data

Supervised ML – Typical Pipeline

- **Example – spam detection**
- Observations – raw emails
- Labels – *spam* or *not spam*
- We are given a set of emails labeled as *spam* or *not spam*
- We want to predict if a new email is spam or not

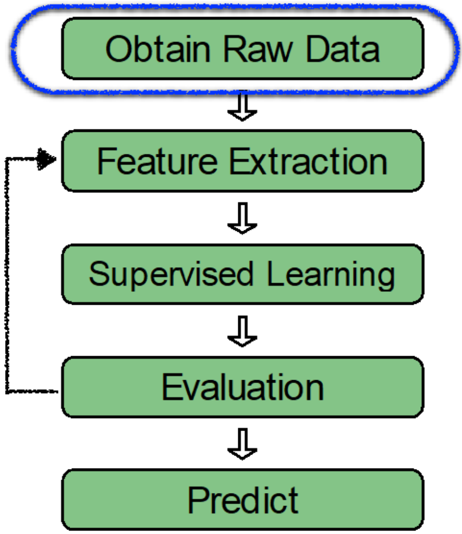
Supervised ML – Typical Pipeline

- Obtaining raw data, i.e., training set



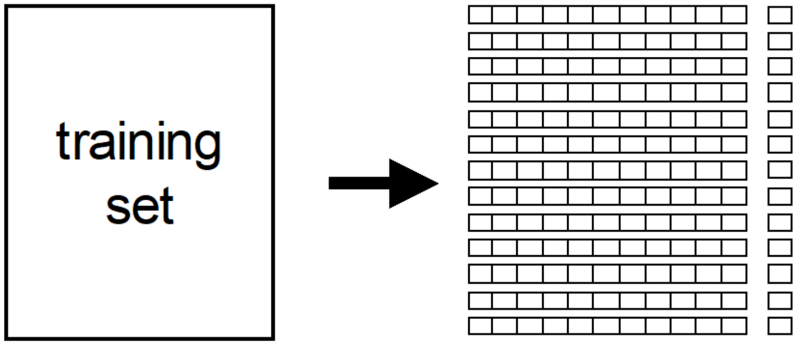
Supervised ML – Typical Pipeline

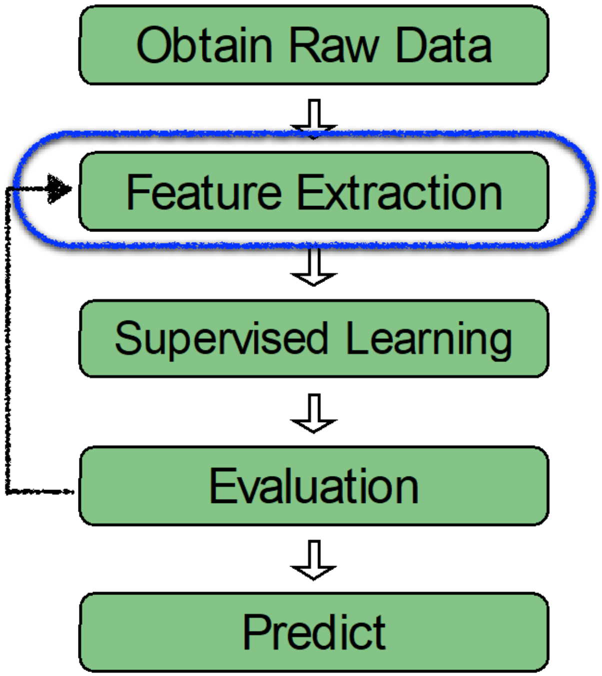
Observation	Label
From: illegitimate@bad.com "Eliminate your debt by giving us your money..."	spam
From: bob@good.com "Hi, it's been a while! How are you? ..."	not-spam



Supervised ML – Typical Pipeline

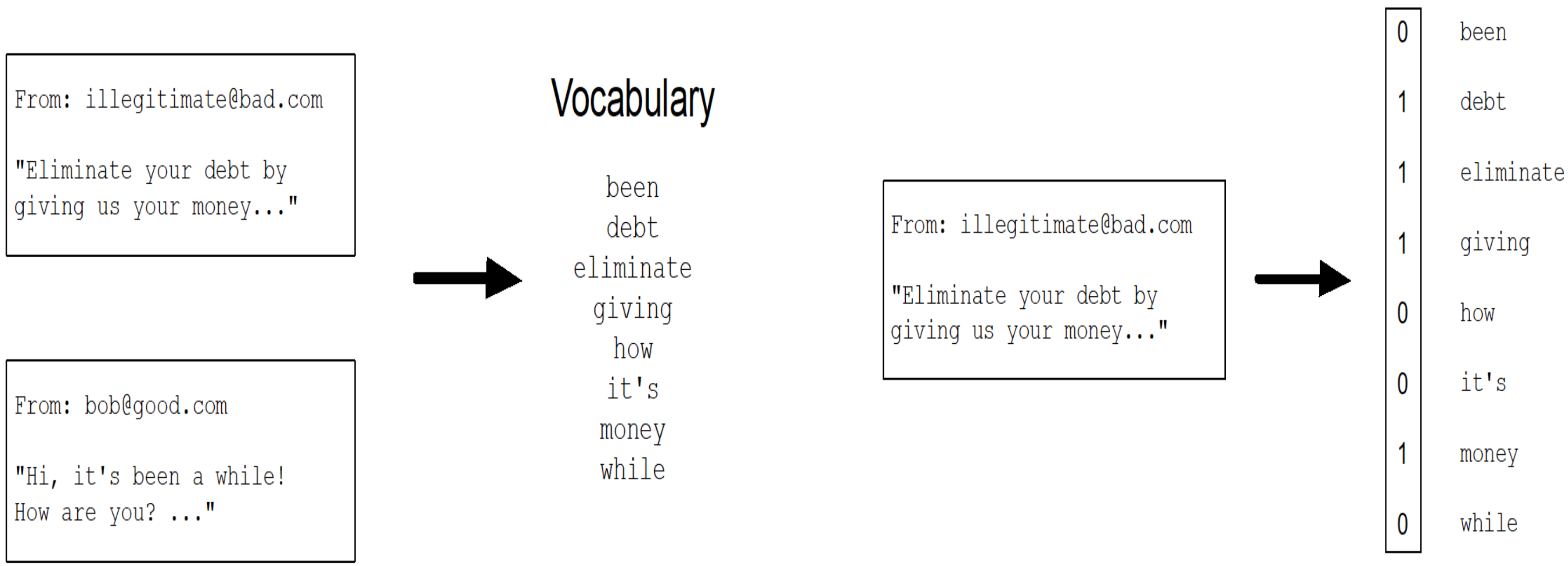
- Feature extraction – convert observation to numeric features





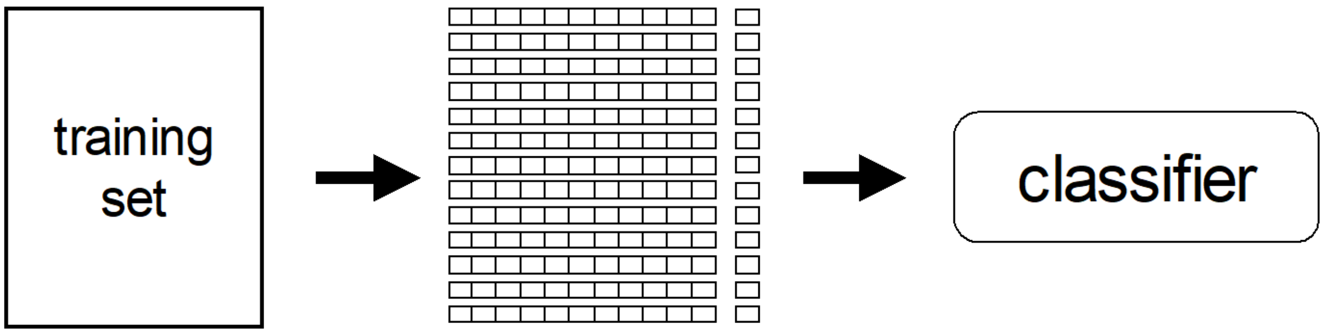
Supervised ML – Typical Pipeline

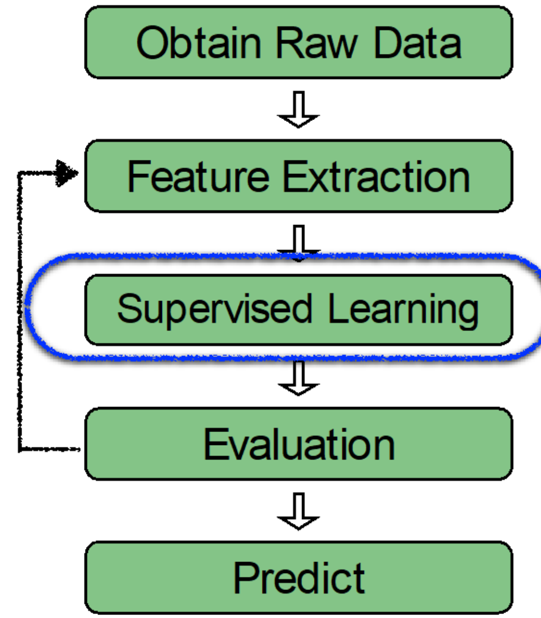
- Feature extraction – "bag of words" representation
- Observations are documents
- Build vocabulary
- Derive features from Vocabulary



Supervised ML – Typical Pipeline

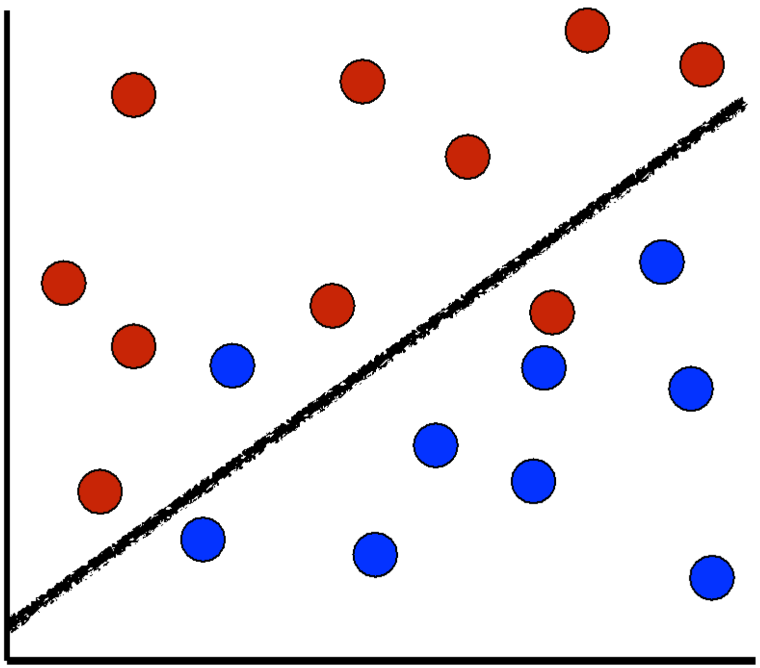
- Train classifier using training data
  - E.g., - Logistic regression, support vector machines, neural nets

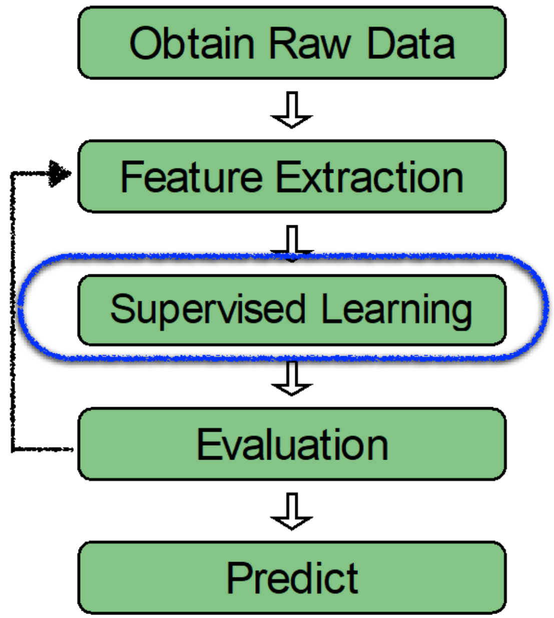




Supervised ML – Typical Pipeline

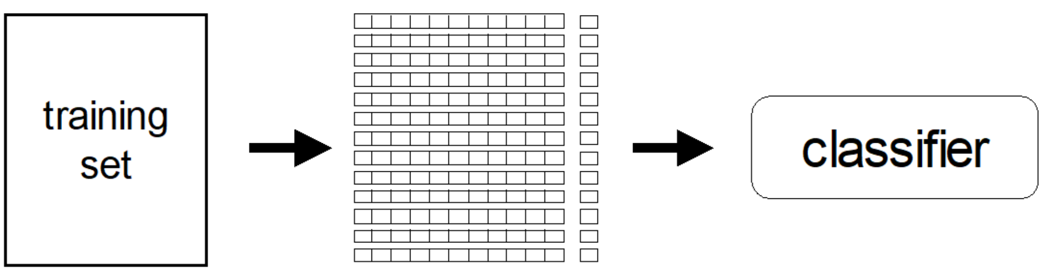
- Classifier - E.g., Logistic regression
  - Find linear decision boundary
  - Learning involves finding offset and feature weights

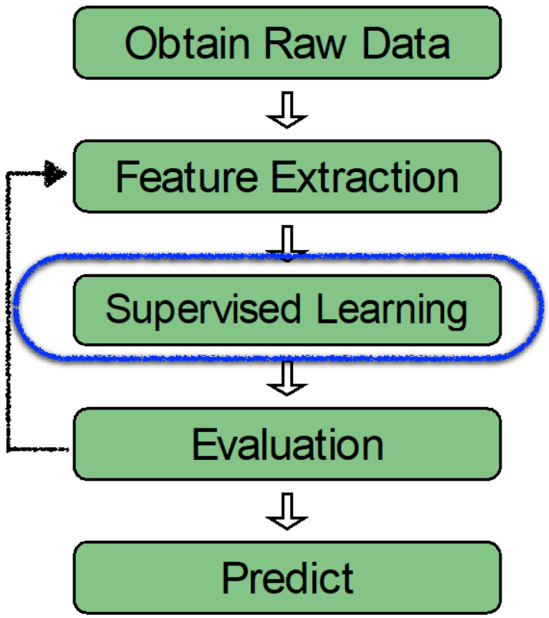




Supervised ML – Typical Pipeline

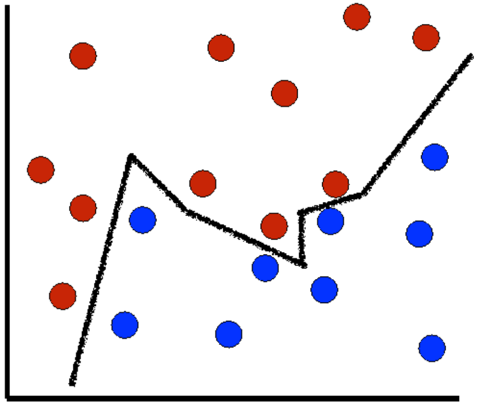
- How can we evaluate quality of classifier?
- Need good predictions on unobserved data
  - Good "generalization" capability
- Want to avoid "overfitting
  - Classifier fits training data very well but fails on other data

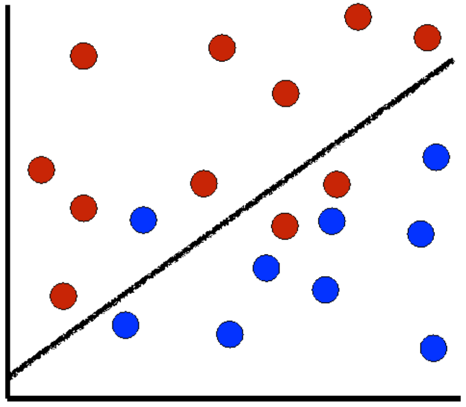


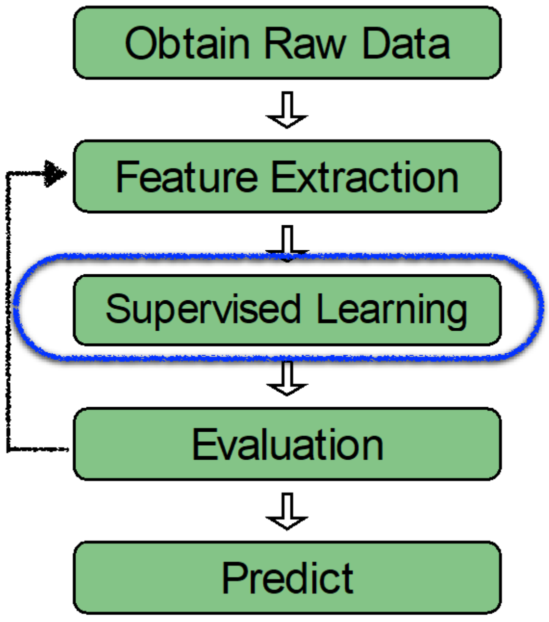


Supervised ML – Typical Pipeline

- Fitting training data does not guarantee generalization
- Left figure – perfect fit but complex model/overfitting
- Right figure – a few training errors but simple/general

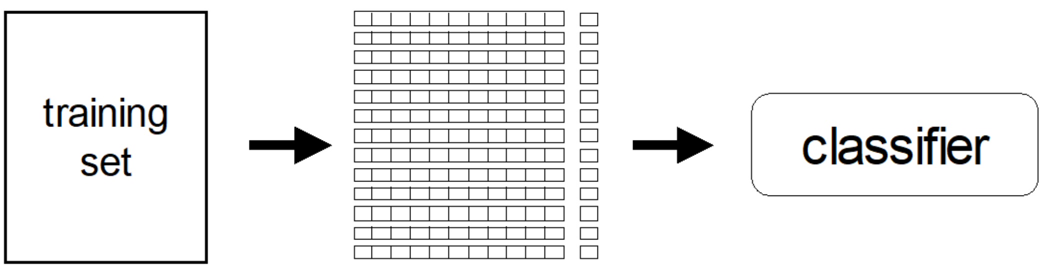


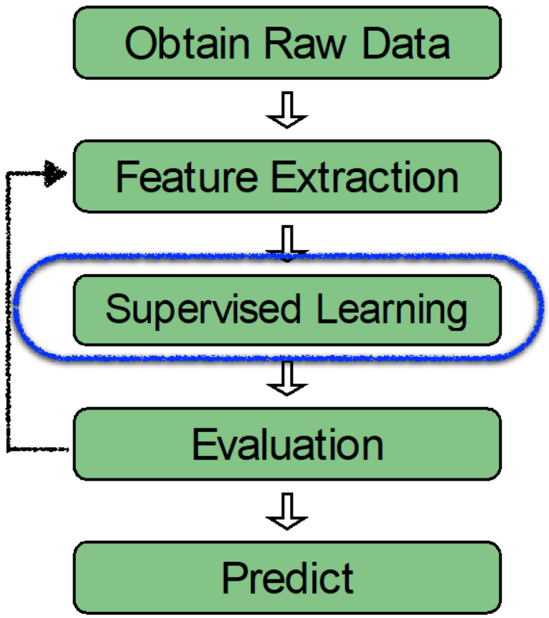




Supervised ML – Typical Pipeline

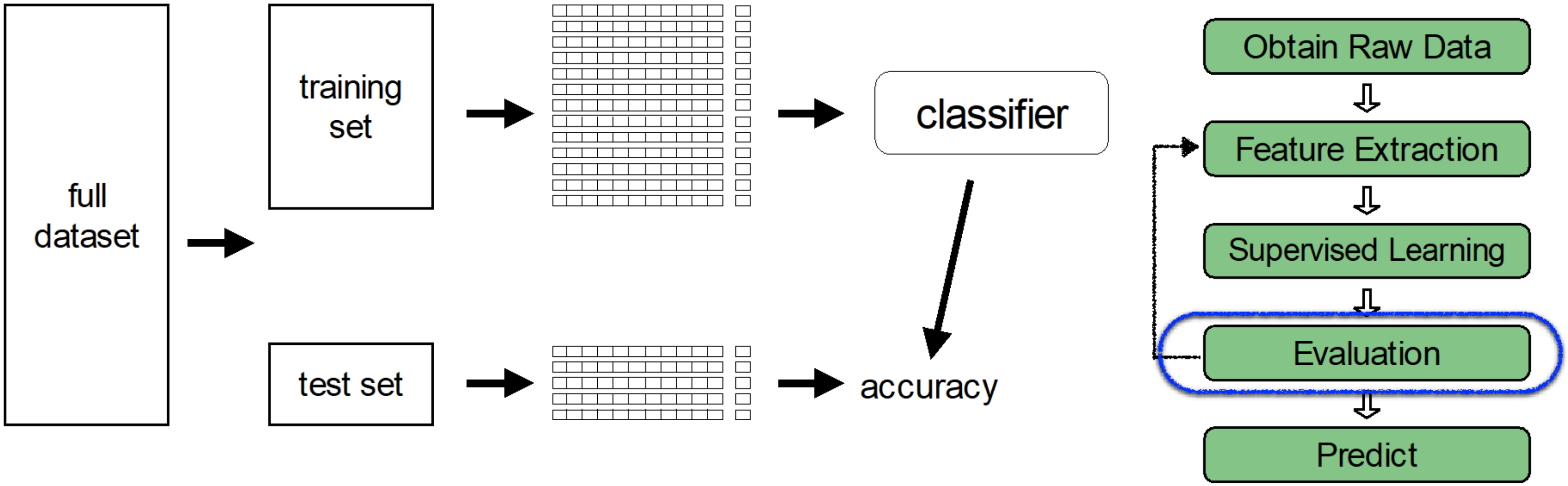
- How can we evaluate quality of classifier?
- **Use a test set to simulate unobserved data**





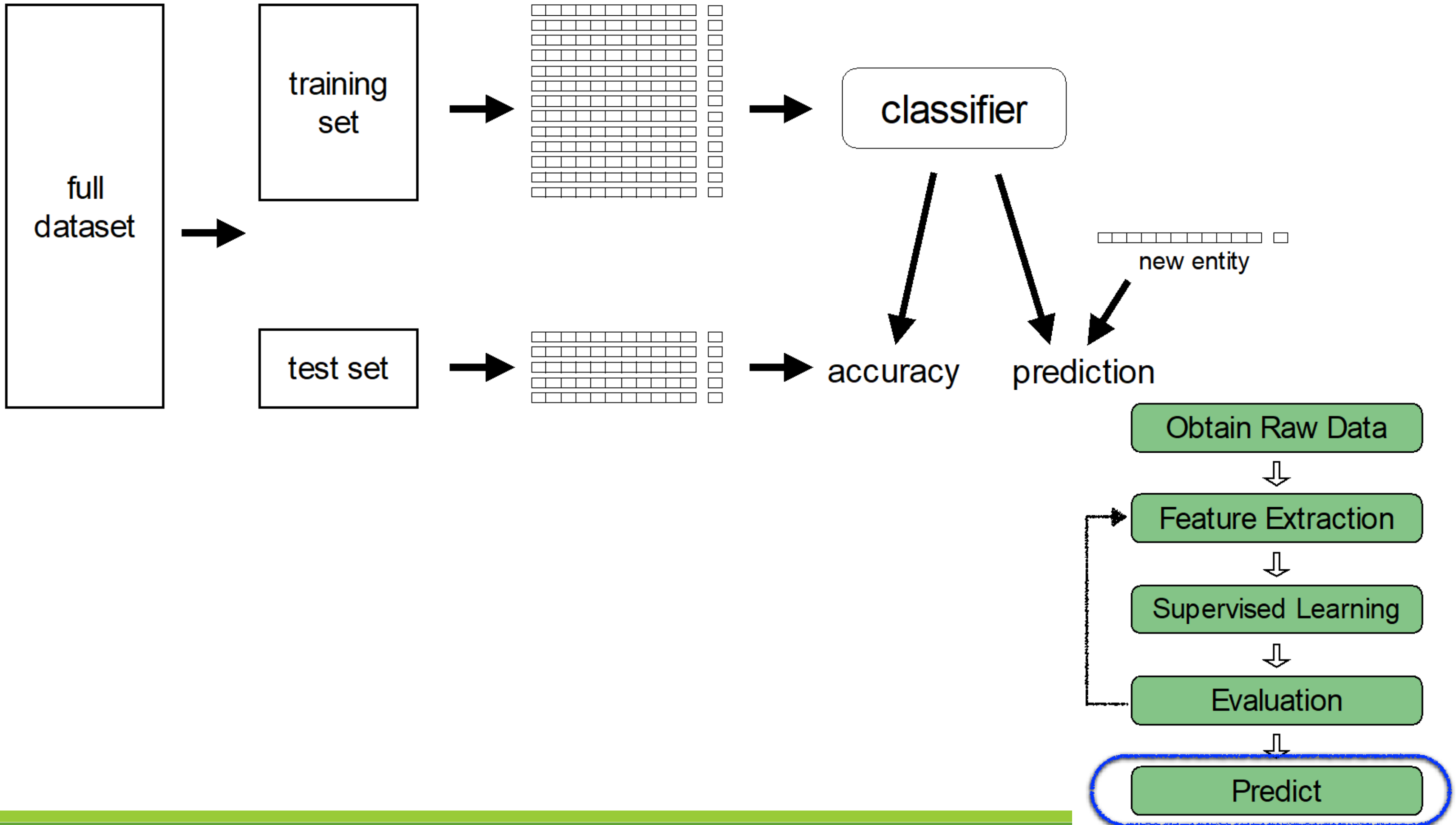
Supervised ML – Typical Pipeline

- Evaluation: split dataset into "training" and "test" set
- Train on training set – (don't expose test set)
- Predict on test set –(ignore labels)
- Compare test predictions with underlying test labels



Supervised ML – Typical Pipeline

- Use final classifier to predict labels for future observations



Semi-Supervised ML

- Goal of the semi-supervised learner is the same as the supervised learner
- Dataset contains both labeled and unlabeled examples
- Quantity of unlabeled examples is much higher than the labeled examples
- Types of Semi-Supervised Learning
  - Self-Training
    - Build supervised classifier on labeled dataset
    - Apply the classifier on unlabeled dataset. Promote records with high confidence from unlabeled to labeled datasets. Repeat.
  - Co-Training

Transductive ML

- It's about reasoning from observed (e.g., training) cases to specific (test) cases
- Example:
  - Only 5 labeled datapoints with labels
  - Label the unlabelled according to the clusters they belong to. Clustering can be done using standard algorithms.
  - The algorithm needs to repeated, if more unseen data are added

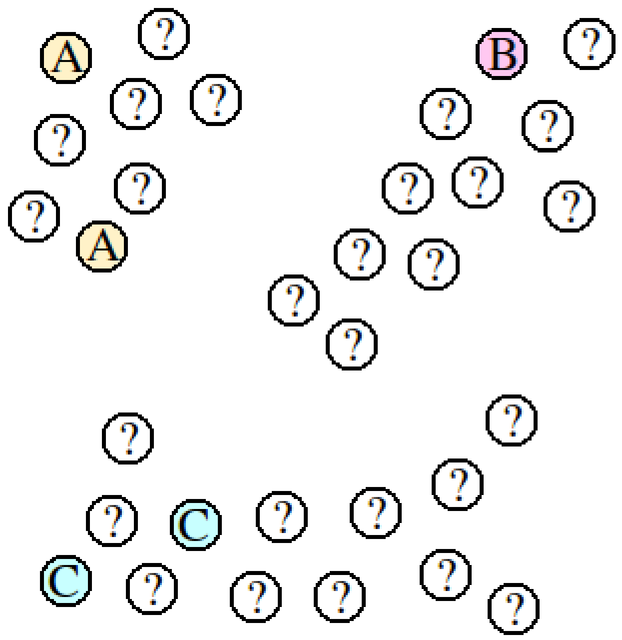


Image: https://en.wikipedia.org/wiki/Transduction_(machine_learning)

Transductive SVM

Unlabelled data guides the linear boundary away from the dense regions

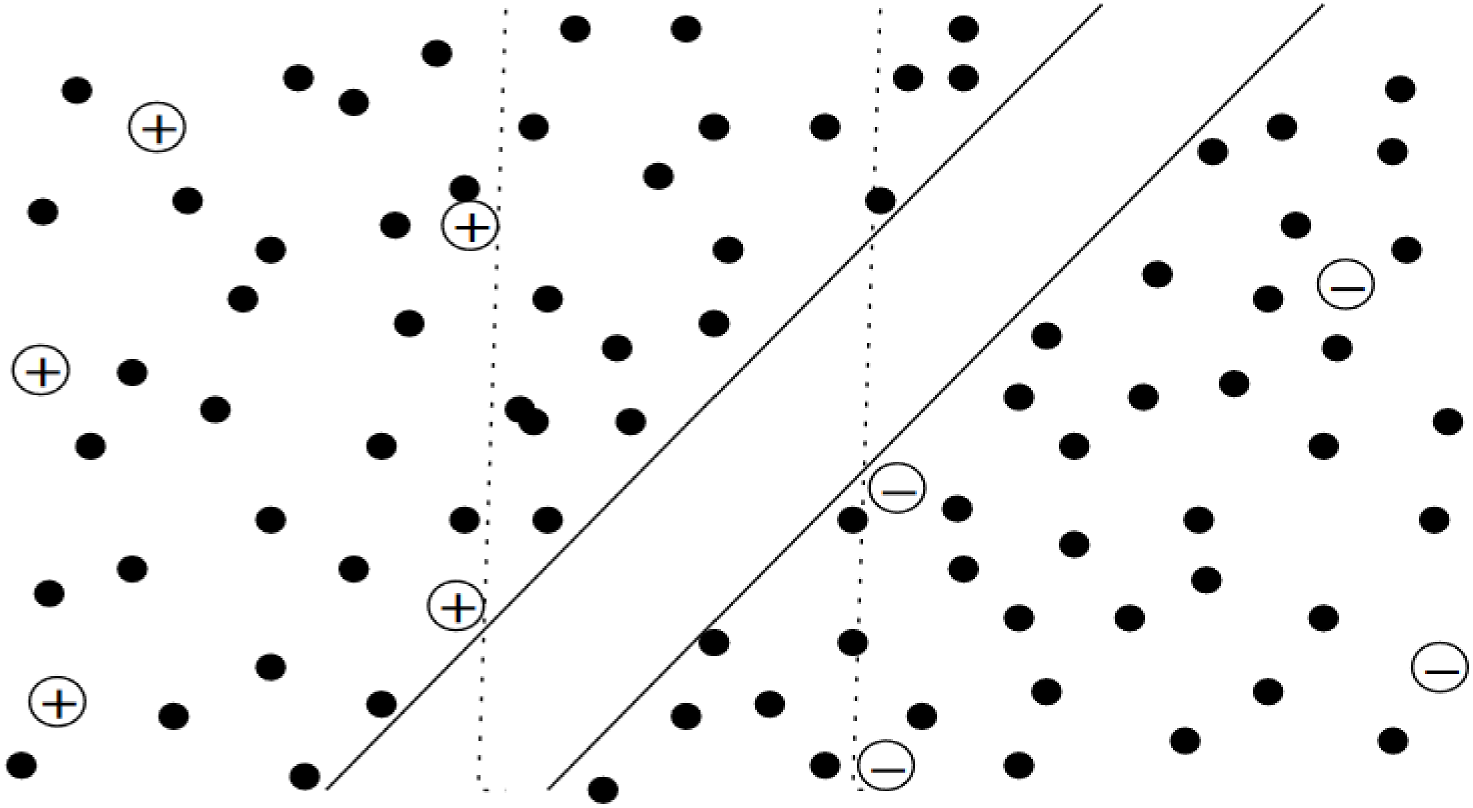


Image credit: https://www.cs.ubc.ca/~schmidtm/MLRG/Semi-Supervised%20Learning.pdf

Topics

- Definition
- Example applications
- Terminology
- Types of Machine Learning
- Typical supervised learning pipeline

Acknowledgements

Portions of these slides were adapted from external material available under creative commons license CC-BY-NC-SA 4.0. This license grants the ability to share and adapt the material for non-commercial purposes.

Name of the creator: Dr. Amit Talwalkar, University of Berkeley and team

License notice: https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

Copyright notice: CC-BY-NC-SA 4.0

Link to material: https://courses.edx.org/courses/BerkeleyX/CS190.1x/1T2015/info