CMSC 471

Machine Learning

What is AI?

- Building Intelligent Agents

What is AI?

- Building Intelligent Agents
- Agents that can:
  - Perceive
  - Reason
  - Act
  - (Autonomous)

Let’s try to build an Intelligent Agent

Bill applies for a credit card

Agent has to approve or reject Bill’s request

Let’s try to build an Intelligent Agent

Bill applies for a credit card

Agent has to approve or reject Bill’s request

Does this task require “Intelligence”?

Let’s try to build an Intelligent Agent

Bill applies for a credit card

Agent has to approve or reject Bill’s request

Does this task require “Intelligence”?

How can we automate this decision?

What does any agent need to make an Intelligent Decision?

What does any agent need to make an Intelligent Decision?

Answer : Information

Building an Automated Intelligent Agent (AI)

Step 1: **Gather Information**

In **AI terms**, this is like the agent “perceiving” its environment and ”remembering”
In **ML terms**, this step is called **data collection.**

Building an Automated Intelligent Agent (AI)

Step 1: **Gather Information**

What do information do we need to review Bill's request?

- Information about Bill:
  - Age, Zipcode, Income, Credit Score etc.
- Information about past customers:
  - What is the minimum age for approval?
  - What is the minimum income for approval?
  - …

Building an Automated Intelligent Agent (AI)

Step 1 Part 2: **Data Preparation**

We have the data now. **Can we process it?**

Building an Automated Intelligent Agent (AI)

Step 1 Part 2: **Data Preparation**

We have the data now. **Can we process it?**

Probably not straight away.
Represent the data in a way that is processable.

Building an Automated Intelligent Agent (AI)

Step 1 Part 2: Data Preparation

We have the data now. **Can we process it?**

Probably not straight away.
Represent the data in a way that is processable.

We worked on this before.
Environment States helped us fit any event into our problem paradigm. Think Graph Search Space, CSN, FOL etc.

For **ML**, data is typically represented as **feature vectors**.
Also called **attribute set**.

Building an Automated Intelligent Agent (AI)

Step 1 Part 2: Data Preparation

What do we need for reviewing credit card request?

- Age
- Income
- Credit Score
- Zipcode

Age	Income	Credit Score	Zipcode

Building an Automated Intelligent Agent (AI)

Step 1 Part 2: Data Preparation

What do we need for reviewing credit card request?

- Age
- Income
- Credit Score
- Zipcode

Age	Income	Credit Score	Zipcode

Feature Vector

Typical Representations: Tables, Arrays, Matrices etc.

Labeled vs Unlabeled data

Label/ Class known to us

Labeled Data

Age	Income	Credit Score	Zipcode	Review Decision

Unlabeled Data

Age	Income	Credit Score	Zipcode

Labeled vs Unlabeled data

Label/ Class known to us

Labeled Data

Age	Income	Credit Score	Zipcode	Review Decision

- Easier to work with
- Not always feasible to have
- Used for **Supervised Learning**

Unlabeled Data

Age	Income	Credit Score	Zipcode

- Harder to work with; but we have models specifically targeting unlabeled data
- Usually this is what we will have
- Used for **Unsupervised Learning**

Going back to our AI

- We have all the data we could possibly need
- We will consider the easier use case:
  - Everything is labeled!
- But now what?

Going back to our AI

- We have all the data we could possibly need
- We will consider the easier use case:
  - Everything is labeled!
- But now what?

We need to **understand** and **reason** over the data

ML approach is **Data Modeling**

Building an Automated Intelligent Agent (AI)

Step 2 : Data Modelling

We have a Feature Vector (X)
We have label/output (Y)

Building an Automated Intelligent Agent (AI)

Step 2 : Data Modelling

We have a Feature Vector (X)
We have label/output (Y)

We believe that Y is dependent on X i.e.
$Y = f(X)$

Building an Automated Intelligent Agent (AI)

Step 2 : Data Modelling

We have a Feature Vector (X)
We have label/output (Y)

We believe that Y is dependent on X i.e.
$Y = f(X)$

We don’t know what f is.
But if we can learn f, given any X we will be able to predict Y

Building an Automated Intelligent Agent (AI)

Step 2 : Data Modelling

We have a Feature Vector (X)
We have label/output (Y)

We believe that Y is dependent on X i.e.
$Y = f(X)$

We don't know what f is.
But if we can learn f, given any X we will be able to predict Y

This is what ML models do

Building an Automated Intelligent Agent (AI)

Step 2 : Data Modelling

We think that:

Review Decision = f (Age, Income, Credit Score, Zipcode, etc.)

We know this

We must learn this

Given enough data, we can **train** an ML model to learn **f**

Building an Automated Intelligent Agent (AI)

Step 3 : Act

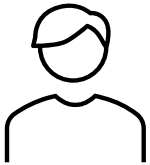
We know f and we know $Y = f(X)$

We have a new use case, X_{new}

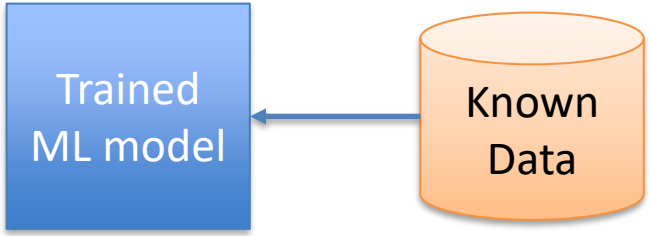
Can we decide Y_{new}?

This is **Class Prediction**

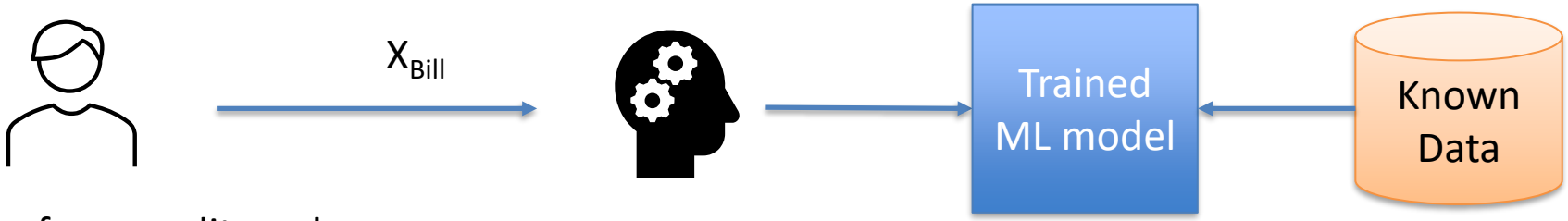
Building an Automated Intelligent Agent (AI)



Bill applies for a credit card

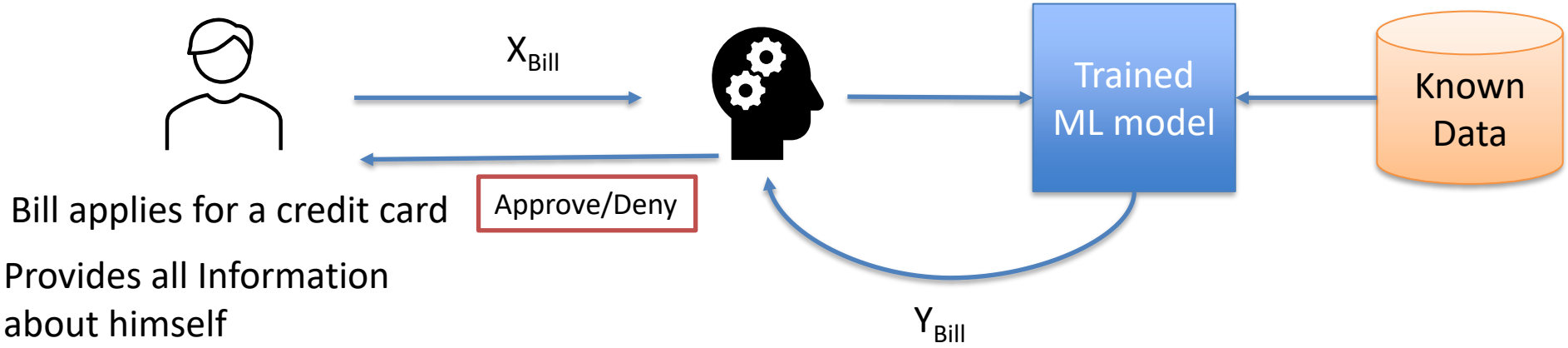


Building an Automated Intelligent Agent (AI)



Bill applies for a credit card
Provides all Information about himself

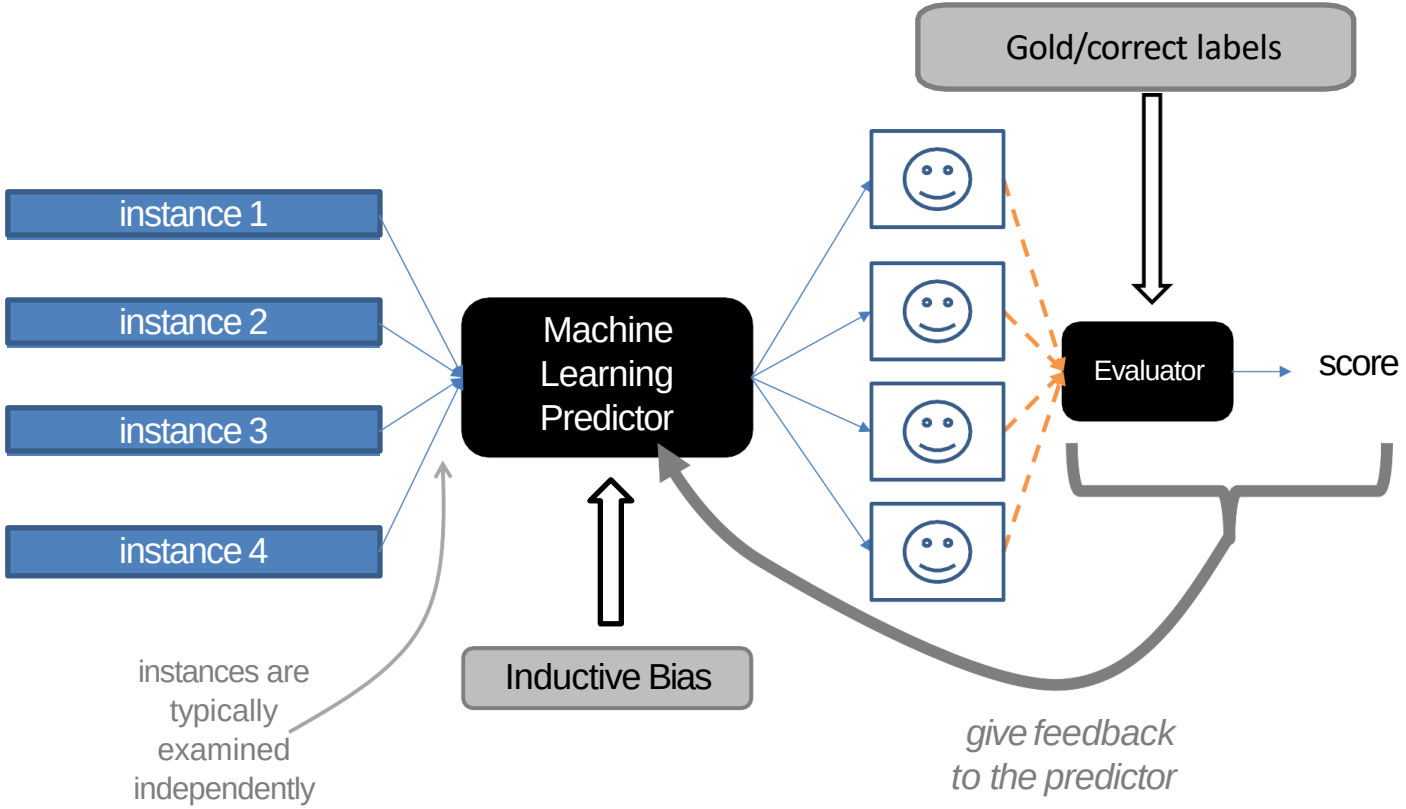
Building an Automated Intelligent Agent (AI)



Bill applies for a credit card

Provides all Information about himself

Machine Learning Framework: Learning

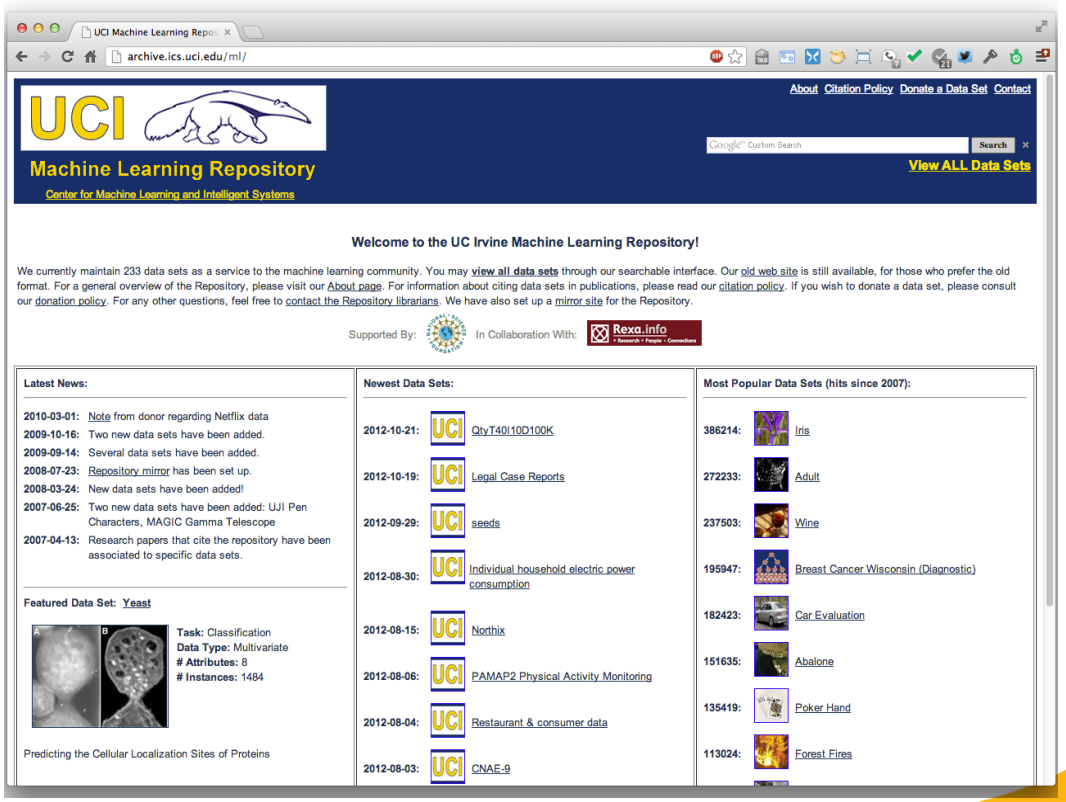


What will we learn in ML?

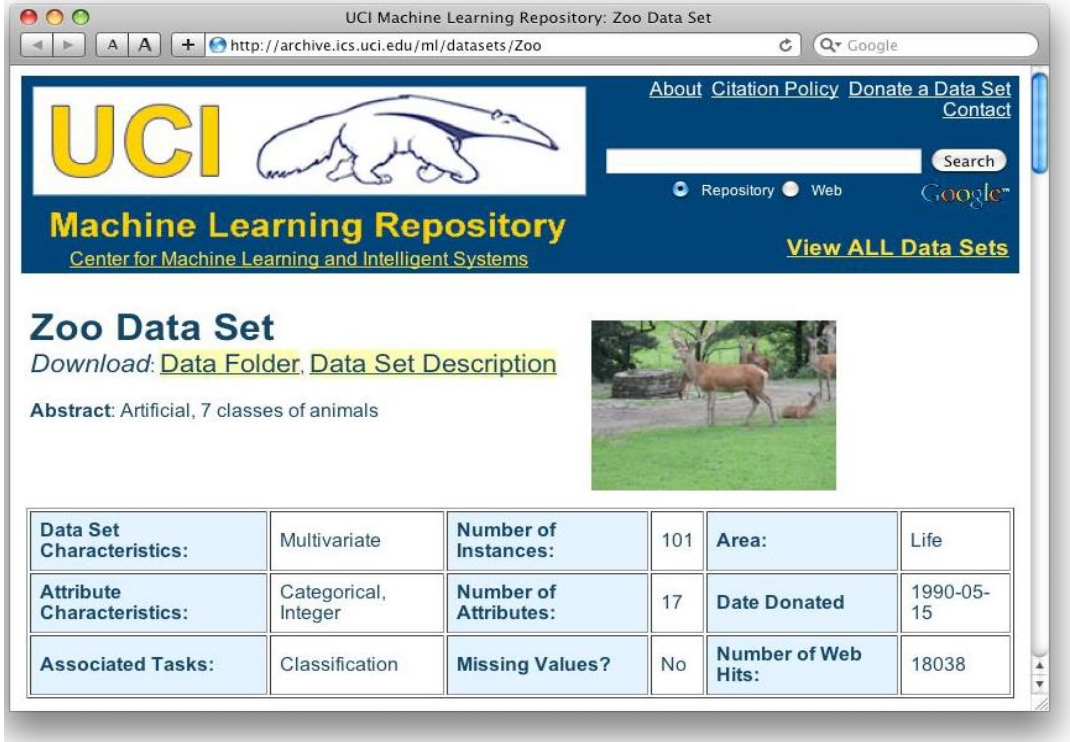
- Building Machine Learning Framework:
  - **Preparing** Data for ML
  - **Training** a model
  - **Evaluating** Model Performance
- Different types of ML models

First thing's First: Get the Data!

http://archive.ics.uci.edu/ml



Zoo Data Set



Data Set Characteristics:	Multivariate	**Number of Instances:**	101	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	17	**Date Donated**	1990-05-15
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	18038

http://archive.ics.uci.edu/ml/datasets/Zoo

Zoo data

Feature/ Attributes

animal name: string
hair: Boolean
feathers: Boolean
eggs: Boolean
milk: Boolean
airborne: Boolean
aquatic: Boolean
predator: Boolean
toothed: Boolean
backbone: Boolean
breathes: Boolean
venomous: Boolean
fins: Boolean
legs: {0,2,4,5,6,8}
tail: Boolean
domestic: Boolean
catsize: Boolean

Label/ Output

type: {mammal, fish, bird, shellfish, insect, reptile, amphibian}

101 examples

aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,mammal
antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,mammal
bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,fish
bear,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,mammal
boar,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,mammal
buffalo,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,mammal
calf,1,0,0,1,0,0,0,1,1,1,0,0,4,1,1,1,mammal
carp,0,0,1,0,0,1,0,1,1,0,0,1,0,1,1,0,fish
catfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,fish
cavy,1,0,0,1,0,0,0,1,1,1,0,0,4,0,1,0,mammal
cheetah,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,mammal
chicken,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,bird
chub,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,fish
clam,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,shellfish
crab,0,0,1,0,0,1,1,0,0,0,0,0,4,0,0,0,shellfish
…

Data Preprocessing

- Feature functions help extract useful features (characteristics) of the data

- They turn *data* into *numbers*

Data Preprocessing

- Feature functions help extract useful features (characteristics) of the data
- They turn *data* into *numbers*
  - Features that are not 0 are said to have fired
  - Often binary-valued (0 or 1), but can be real-valued

Data Preprocessing

Type = shellfish

Type_mammal	Type_fish	Type_bird	Type_shellfish	Type_insect	Type_reptile	Type_amphibian
0	0	0	1	0	0	0

Types of ML task

Output \ Data	Labeled	Unlabeled
Discrete	✓ classification or categorization	✓ clustering
Continuous	✓ regression	dimensionality reduction

Choosing an ML model: Depends on your task

- Classification:
  - Decision Trees
  - Naïve Bayes
  - Logistic Regression etc.
- Regression:
  - Random Forest
  - Linear Regression etc.
- Clustering:
  - K-Means Clustering
  - Mean-Shift Clustering
  - DBSCAN etc.

Evaluation: How Well Are We Doing?

Classification

- Precision, Recall, F1
- Accuracy
- Log-loss
- ROC-AUC
- …

Regression

- (Root) Mean Square Error
- Mean Absolute Error
- …

Clustering

- Mutual Information
- V-score
- …

Evaluation methodology (1)

Standard methodology:

1. Collect large set of examples with correct classifications (aka ground truth data)
2. Randomly divide collection into two disjoint sets: ***training*** and ***test*** *(e.g., via a 90-10% split)*
3. Apply learning algorithm to **training** set giving hypothesis H
4. Measure performance of H on the held-out **test** set