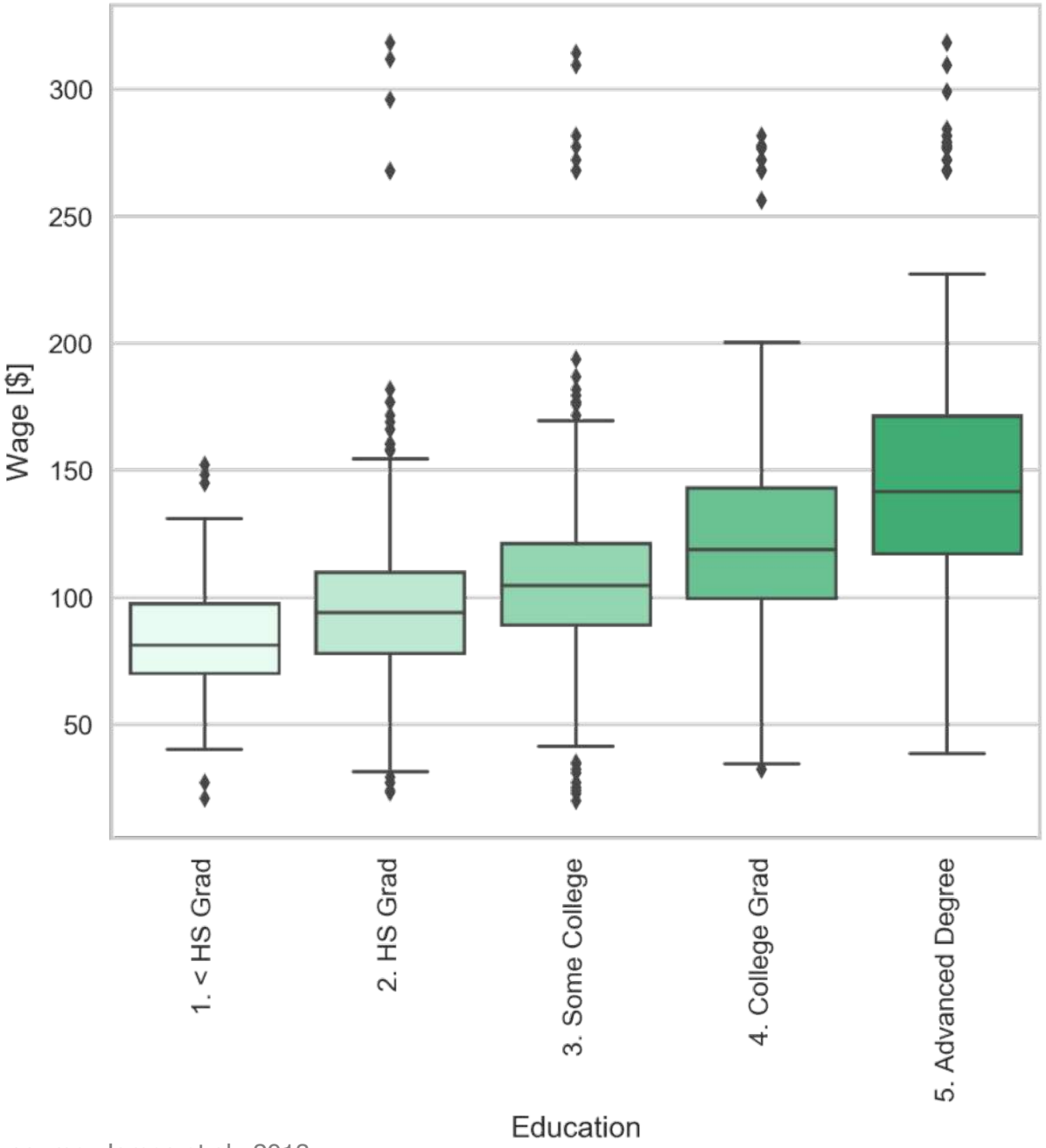
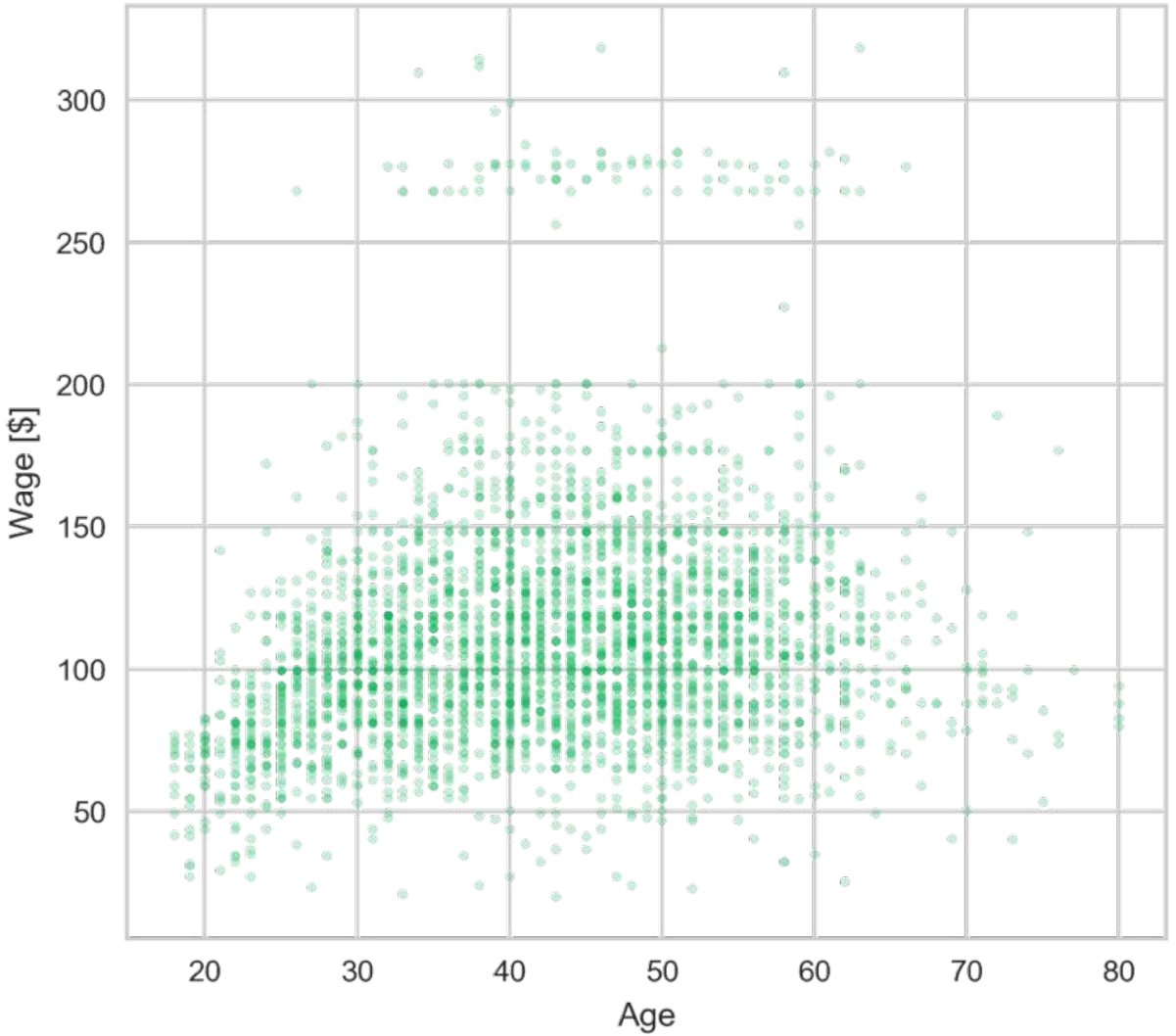
What is machine learning?

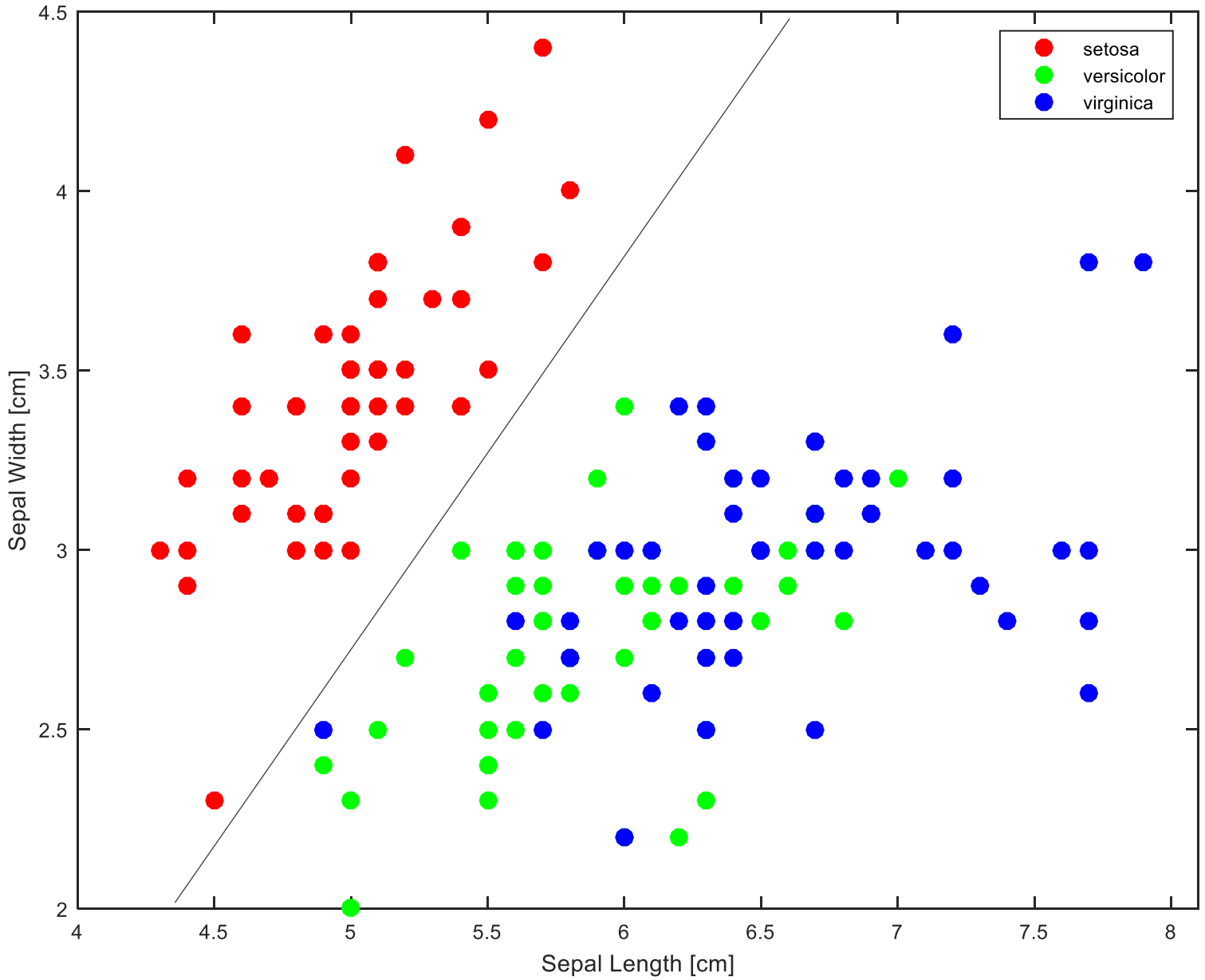
Machine Learning Day





Wage data from workers in the mid-Atlantic region

Data source: James et al., 2013



setosa

versicolor

virginica

Data Source: Fisher Iris Data

Challenges

What is this?

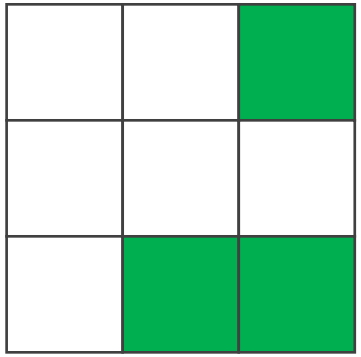
Image by artist Hikaru Cho

We **generalize** from past experiences

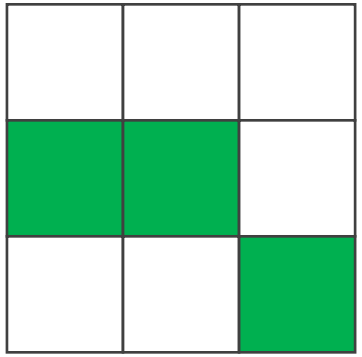
Image: “It’s not what it seems” by artist Hikaru Cho

Predict which class x_{new} belongs to…

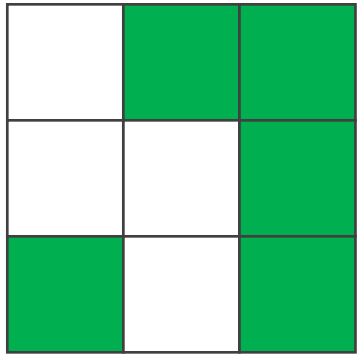
$f(\boldsymbol{x}) = 1$



$\boldsymbol{x}_0$

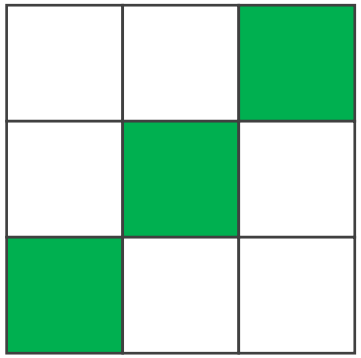


$\boldsymbol{x}_2$

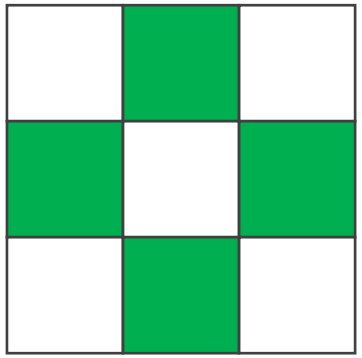


$\boldsymbol{x}_4$

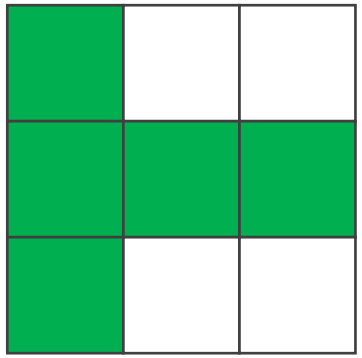
$f(\boldsymbol{x}) = 0$



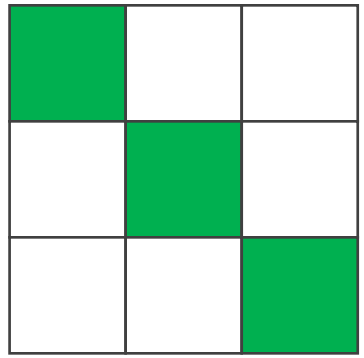
$\boldsymbol{x}_1$



$\boldsymbol{x}_3$



$\boldsymbol{x}_5$



$\boldsymbol{x}_{new}$

$f(\boldsymbol{x}_{new}) = ?$

Example credit: Yaser Abu-Mostafa, 2012

1 Machine learning is an **ill-posed problem**

There are often **many** models that will fit your data – how do we choose which to use?

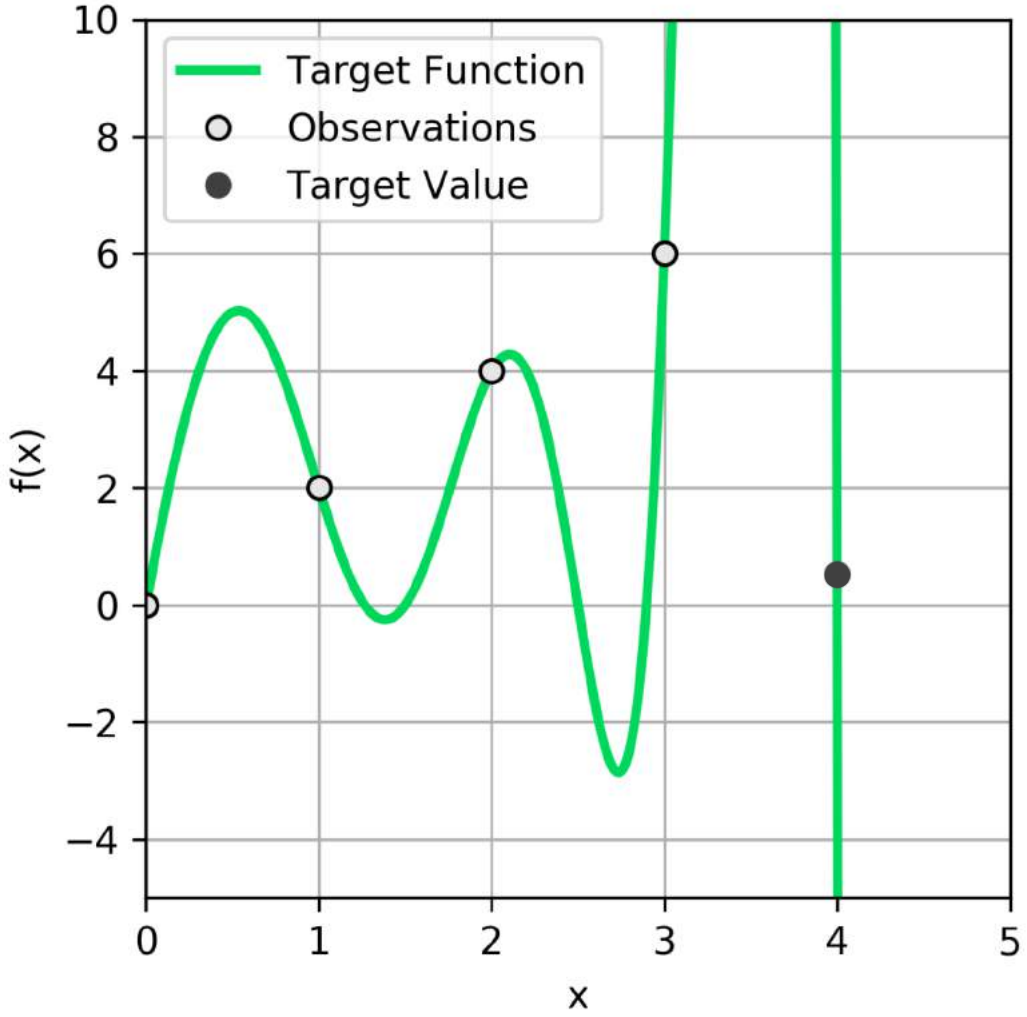
the best models

generalize well

Predict the next value in the sequence…

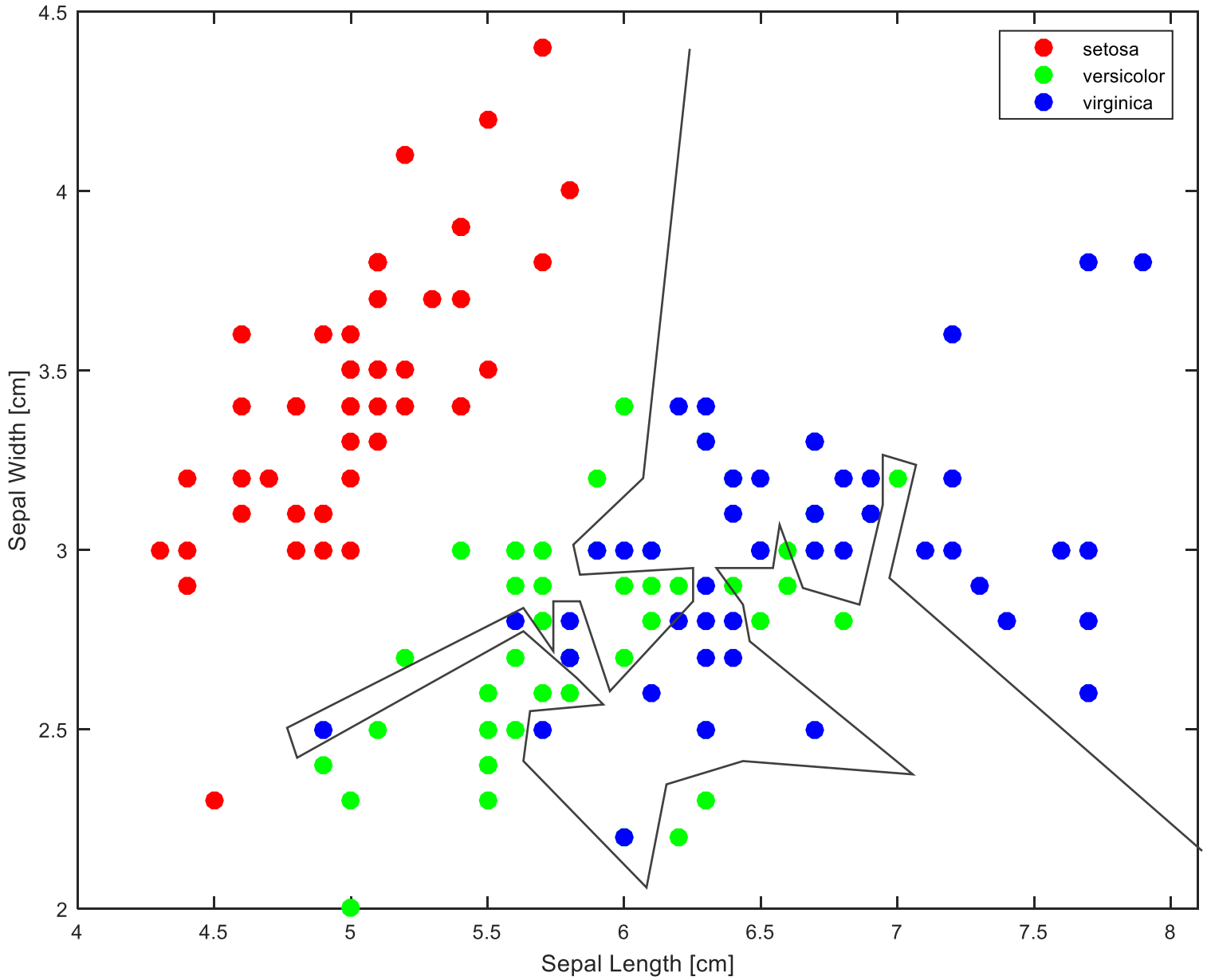
x	0	1	2	3	4
$f(x)$	**0**	**2**	**4**	**6**	**?**

$$f(4) = \mathbf{0.530}$$



Our guess:

$$f(x) = 16.2x - 6.36x^2 - 11.9x^3 - 4.77x^4 + 7.03x^5 + 8.32x^6 - 9.01x^7 + 2.75x^8 - 0.275x^9$$



setosa

versicolor

virginica

Data Source: Fisher Iris Data

2

Overfit works against generalization

We can use **regularization** to prevent overfit

What is machine learning?

A class of techniques where the **goal** is to **describe**, **predict**, and **strategize**…

…**based on** data, past experiences, and/or direct instruction…

…and do so **automatically**, with minimal human intervention.

Types of machine learning tools

Types of learning	Common use case
Unsupervised learning	**Describe**
Supervised learning	**Predict**
Reinforcement learning	**Strategize**

Types of machine learning

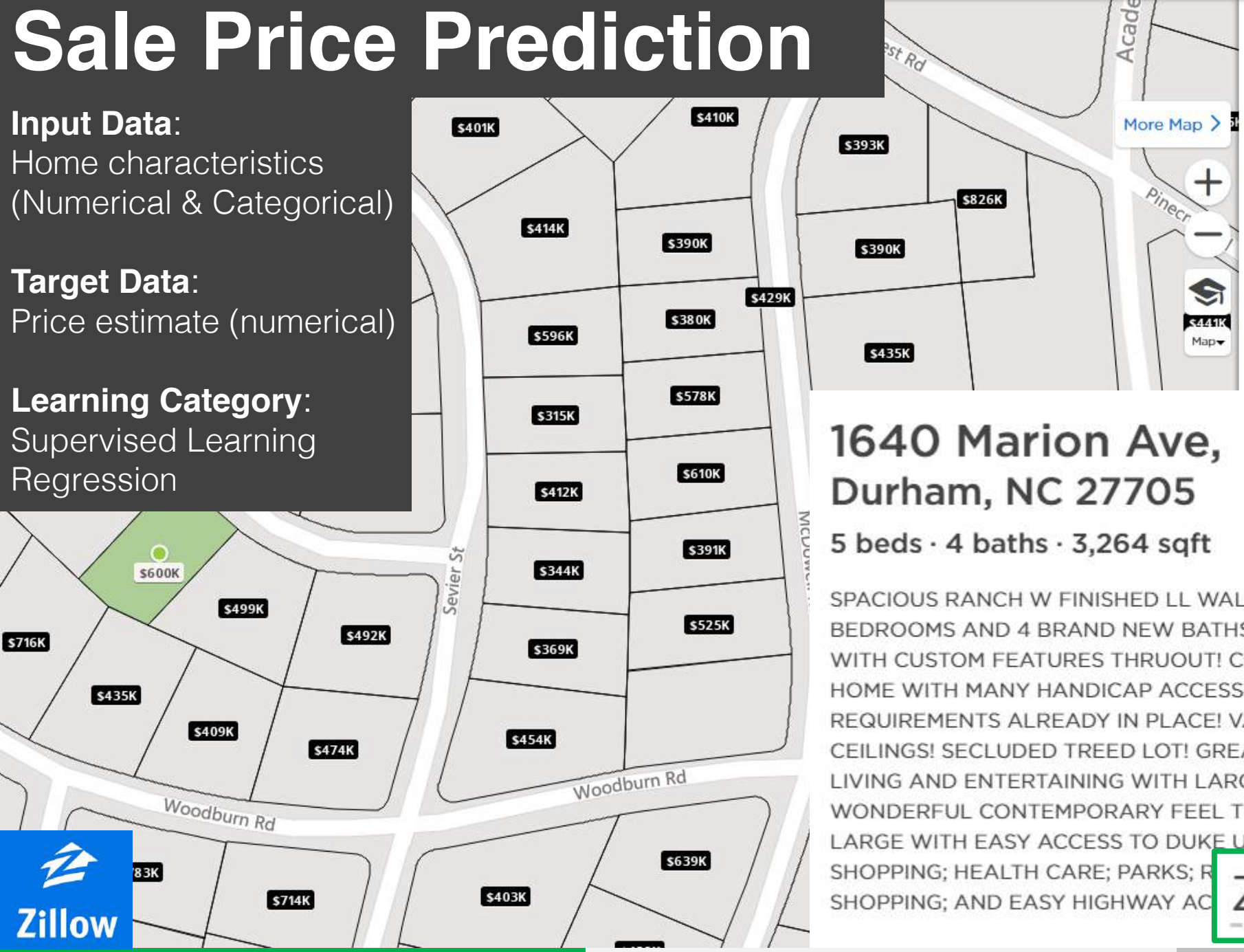
	Unsupervised Learning	Supervised Learning	Reinforcement Learning
Goal	**Describe** ...structure in data	**Predict** …from examples	**Strategize** Learn through interaction
Data available	predictors, $\mathbf{x}$	predictor and response pairs, $(\mathbf{x}, \mathbf{y})$	actions and delayed responses (called rewards)
Examples	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Classification • Regression	• Model-free learning • Model-based learning

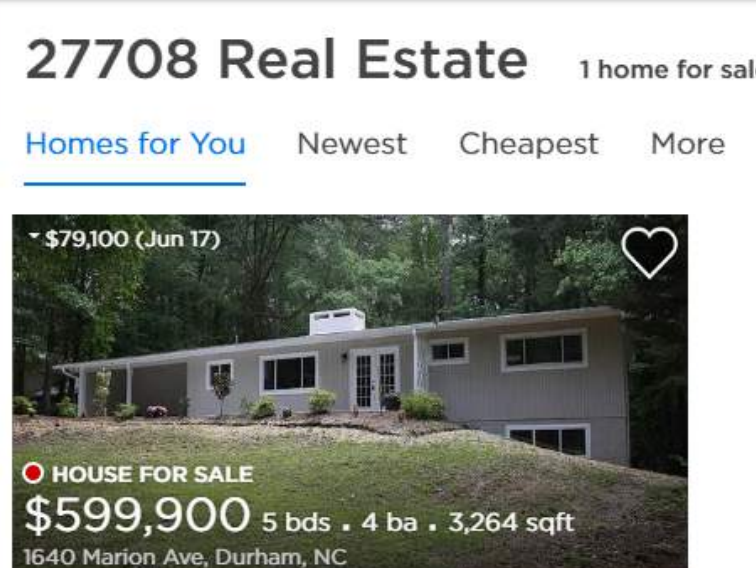
Sale Price Prediction

Input Data:
Home characteristics
(Numerical & Categorical)

Target Data:
Price estimate (numerical)

Learning Category:
Supervised Learning
Regression





1640 Marion Ave, Durham, NC 27705

5 beds · 4 baths · 3,264 sqft

SPACIOUS RANCH W FINISHED LL WALKOUT! 5 BEDROOMS AND 4 BRAND NEW BATHS! RENOVATED WITH CUSTOM FEATURES THRUOUT! CONTEMPORARY HOME WITH MANY HANDICAP ACCESSIBLE REQUIREMENTS ALREADY IN PLACE! VAULTED CEILINGS! SECLUDED TREED LOT! GREAT HOME FOR LIVING AND ENTERTAINING WITH LARGE REAR DECK! WONDERFUL CONTEMPORARY FEEL THAT LIVES LARGE WITH EASY ACCESS TO DUKE UNIVERSITY; SHOPPING; HEALTH CARE; PARKS; R
SHOPPING; AND EASY HIGHWAY AC

FOR SALE
$599,900
Price cut: -$79,100 (6/17)
Zestimate®: $619,585

EST. MORTGAGE
$2,284/mo
Get pre-qualified

Zillow

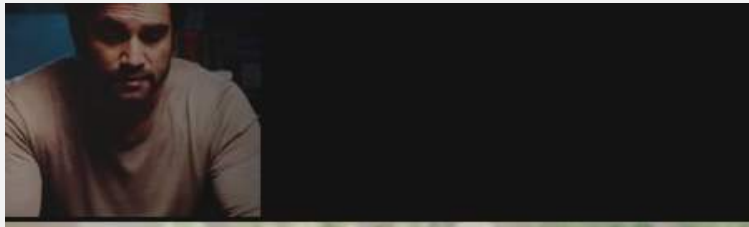
Zestimate®: $619,585

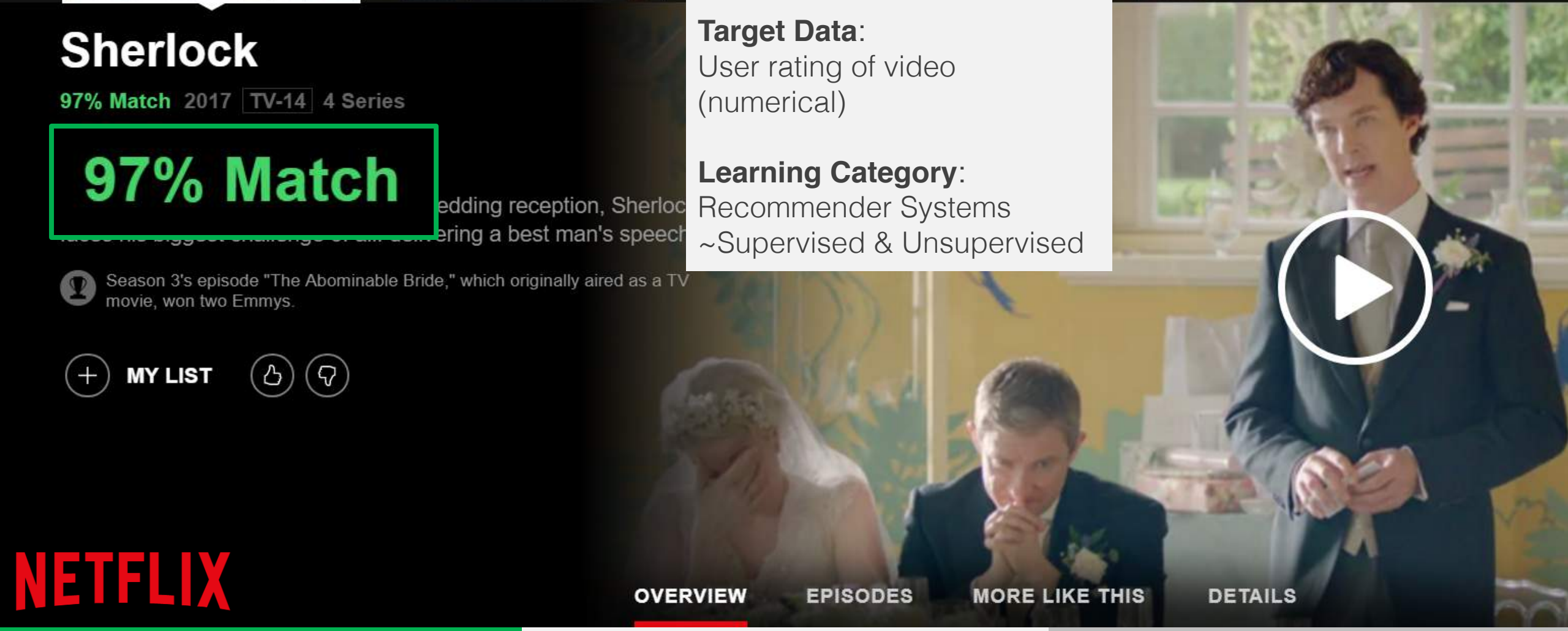
Video Recommendations

Input Data:
User video ratings
(numerical and categorical)

Target Data:
User rating of video
(numerical)

Learning Category:
Recommender Systems
~Supervised & Unsupervised





From: Internal Revenue Service [mailto:yourtaxrefund@InternalRevenueService.com]

Sent: Tuesday, July 22, 2008 9:47 AM

Subject: Get your tax refund now

Importance: High

After the last annual calculations of your account activity we have determined that you are eligible to receive a tax refund of $479.30 .

Please submit the tax refund request and allow us 2-6 days in order to process it.

A refund can be delayed for a variety of reasons. For example submitting invalid records or applying after the deadline.

To access the form for your tax refund, please click here (http://e-dlogs.rta.mi.th:84/www.irs.gov/)

Note: Deliberate wrong inputs will be prosecuted by law.

Regards,

Internal Revenue Service

Spam example source: itservices.uchicago.edu

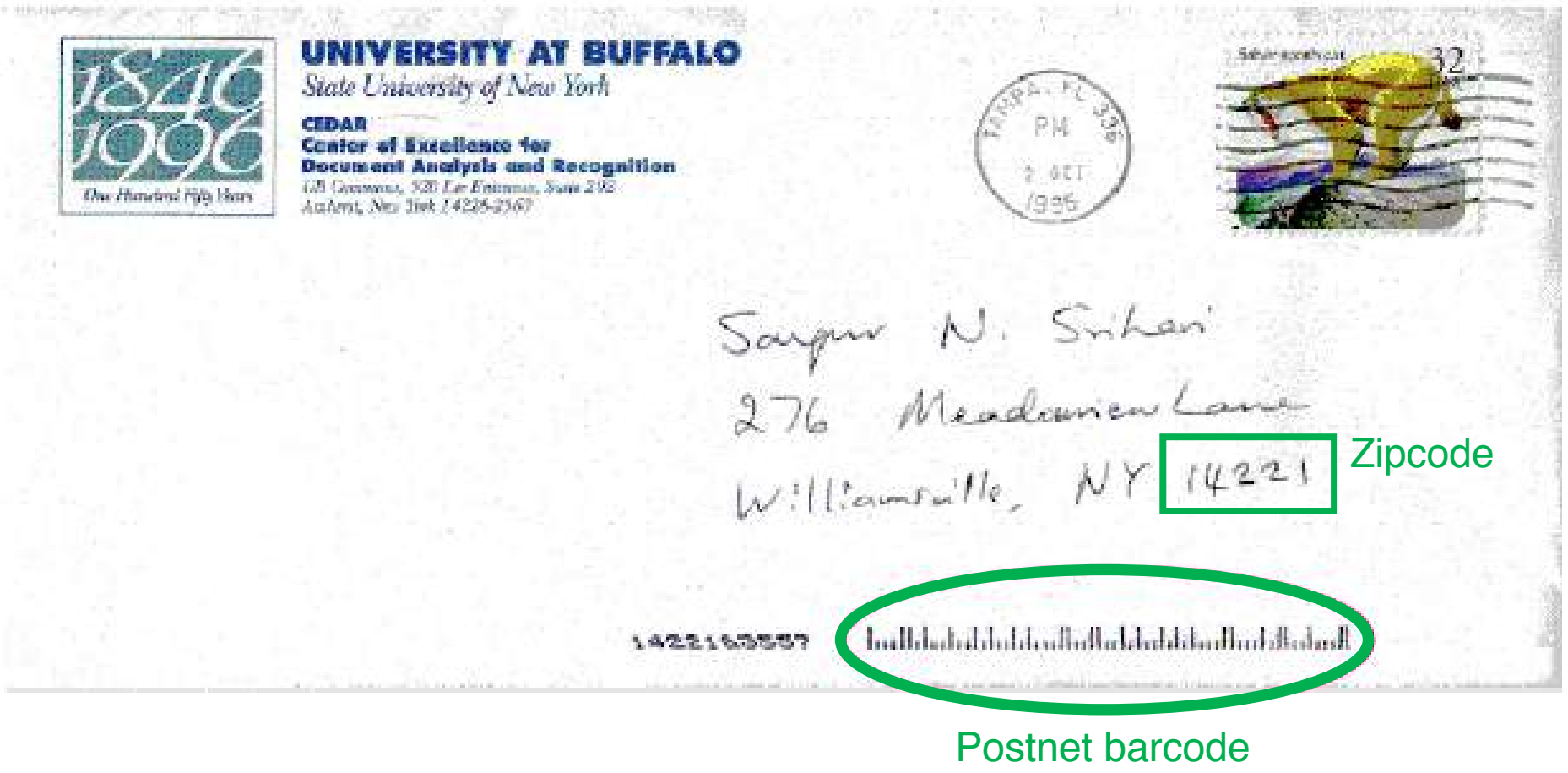
Spam Filters

Input Data:
Email text (text)

Target Data :
Spam/not spam (category)

Learning Category:
Supervised Learning
Classification (binary)

Handwriting and Optical Character Recognition



Input Data:
Imagery

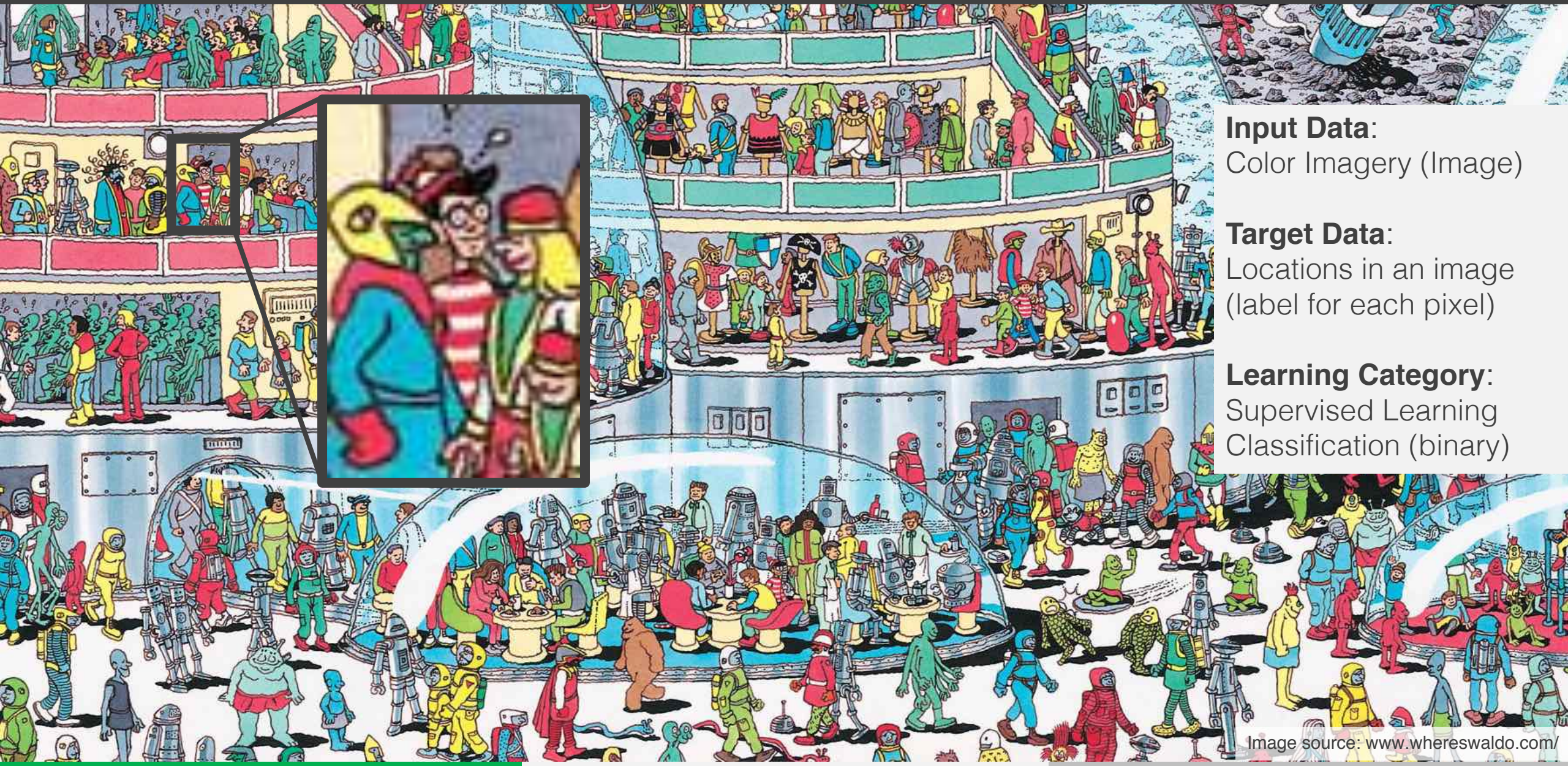
Target Data:
Text Characters

Learning Category:
Supervised Learning
Classification (multiclass)

Among the first handwritten addresses sorted automatically in October 1996

Image source: Sargur Srihari, SUNY

Where's Waldo = Computer Vision Problem



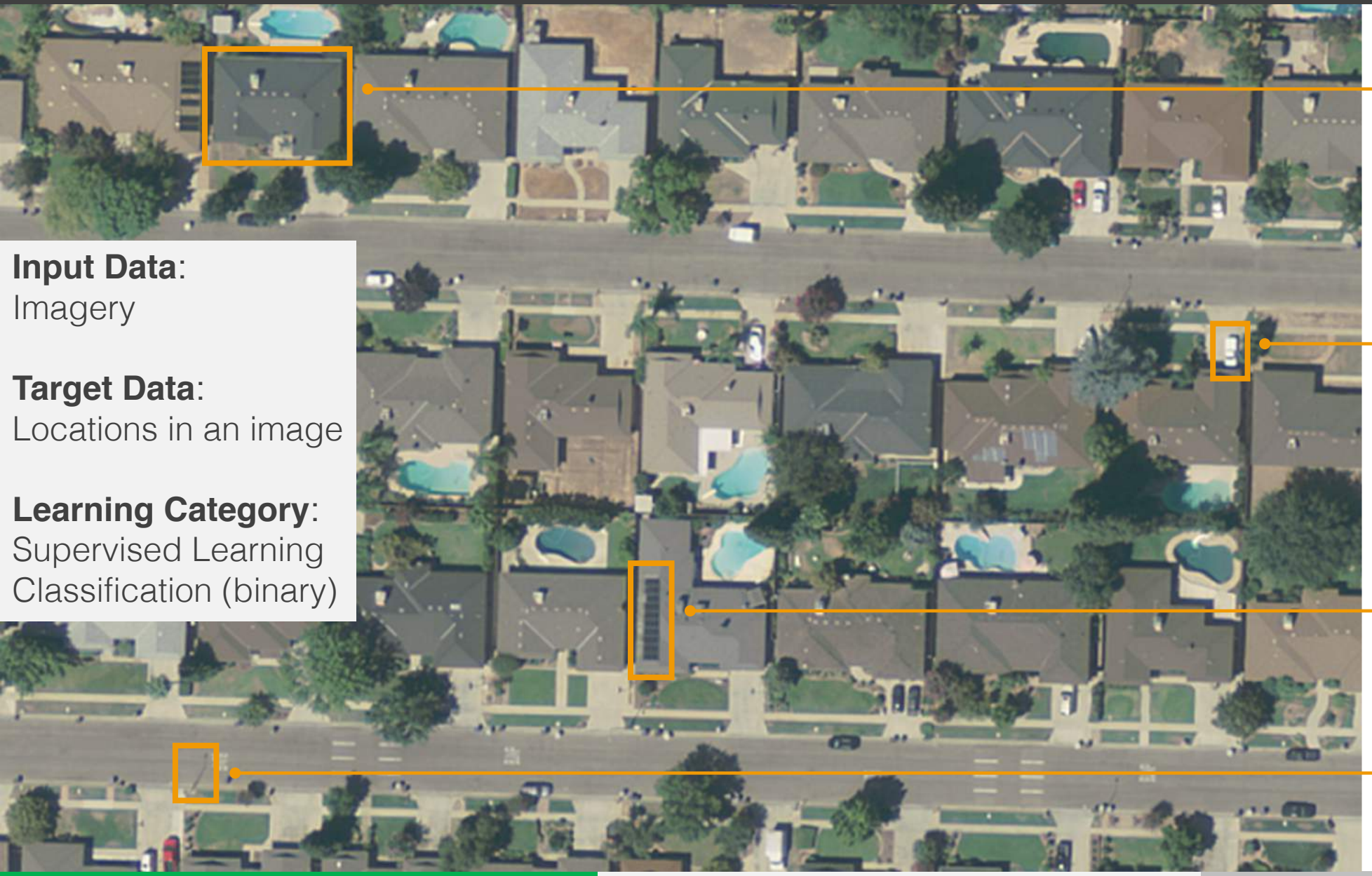
Input Data:
Color Imagery (Image)

Target Data:
Locations in an image
(label for each pixel)

Learning Category:
Supervised Learning
Classification (binary)

Image source: www.whereswaldo.com/

Object Recognition: Energy Systems



Input Data:
Imagery

Target Data:
Locations in an image

Learning Category:
Supervised Learning
Classification (binary)

Building
behind-the-meter
energy consumption

Car
transportation
energy consumption

Solar Array
distributed energy
resources

Light Pole
access to electricity

Types of machine learning

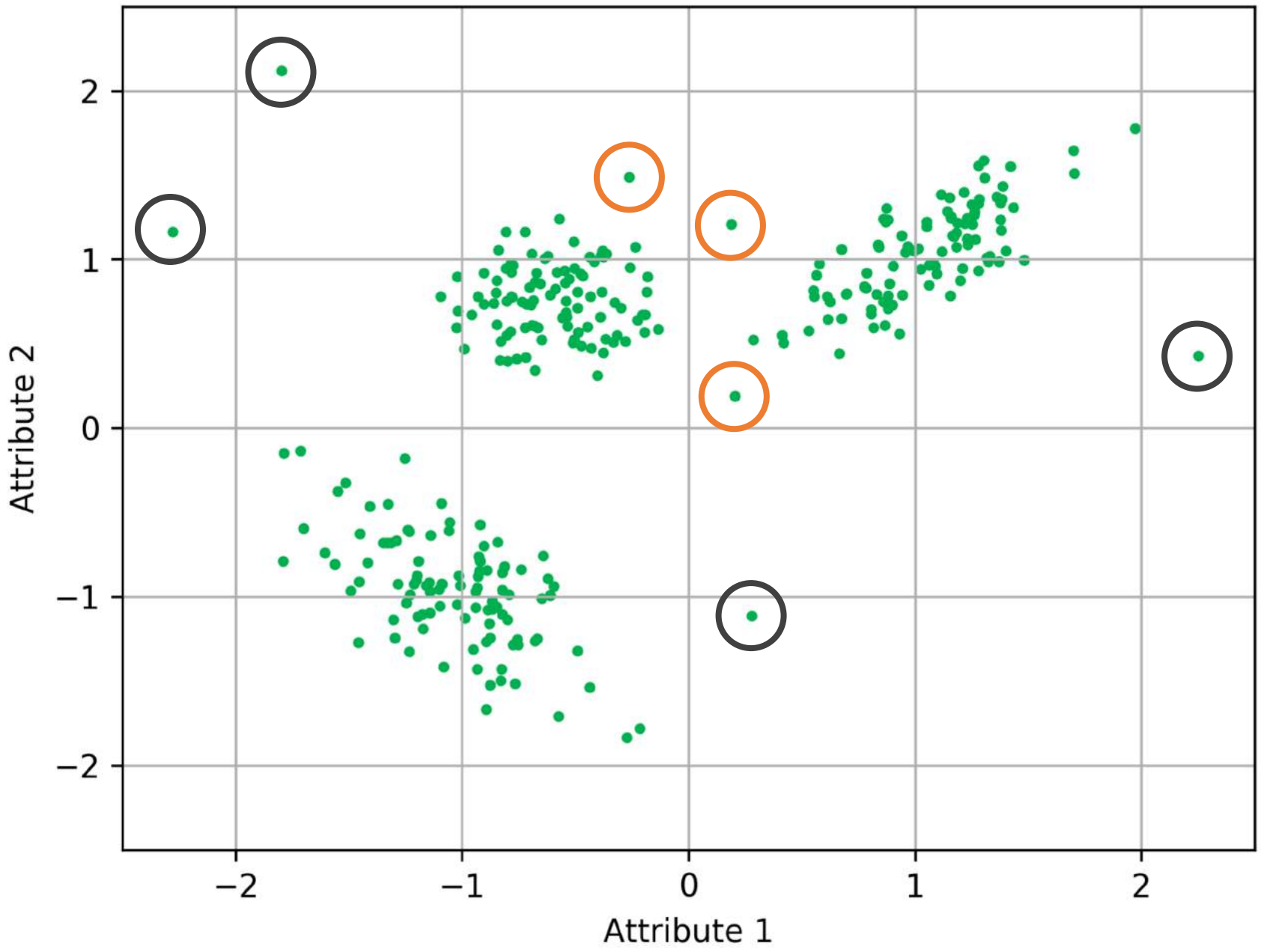
	Unsupervised Learning	Supervised Learning	Reinforcement Learning
Goal	**Describe** ...structure in data	**Predict** …from examples	**Strategize** Learn through interaction
Data available	predictors, $\mathbf{x}$	predictor and response pairs, $(\mathbf{x}, \mathbf{y})$	actions and delayed responses (called rewards)
Examples	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Classification • Regression	• Model-free learning • Model-based learning

Credit Fraud

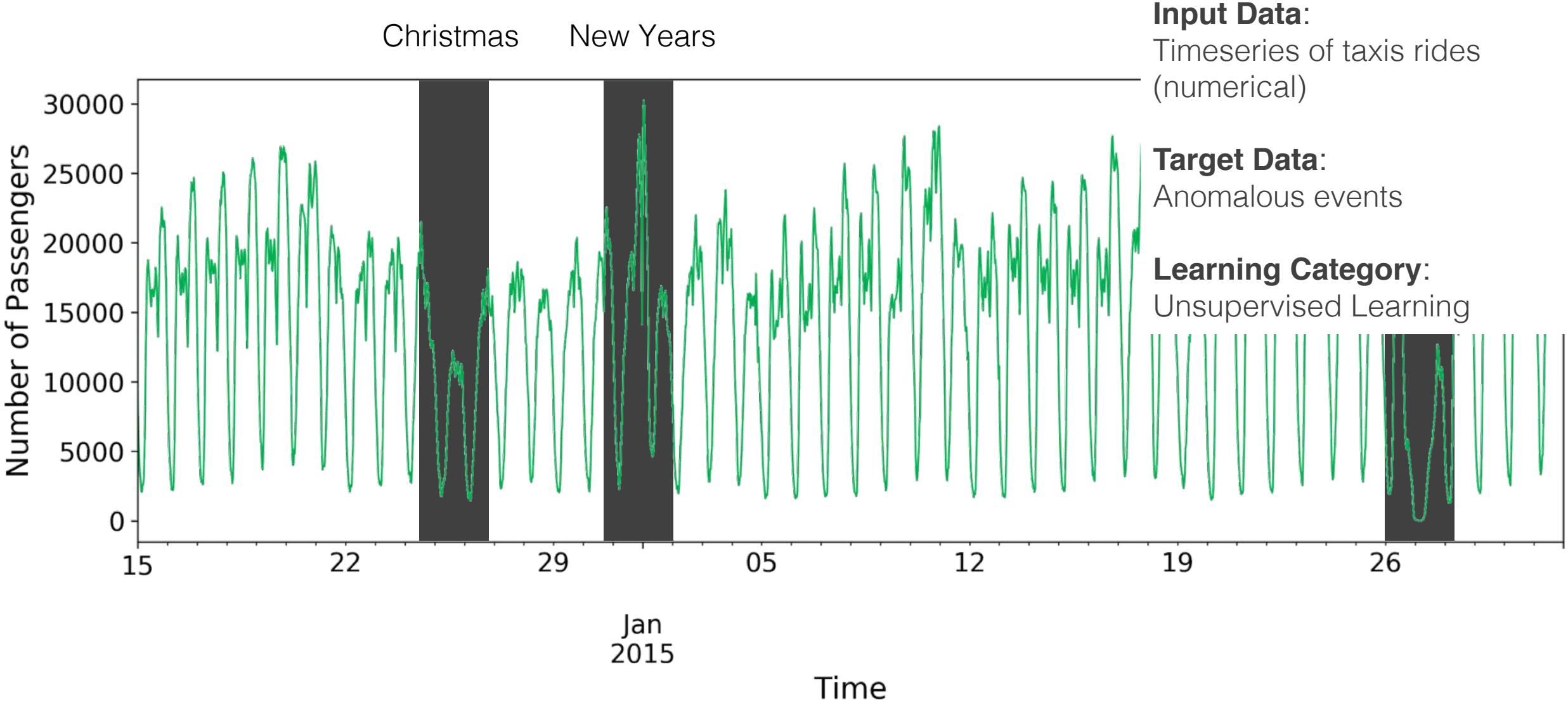
Input Data:
Account transactions, dates, locations, demographic information
(Numerical and categorical)

Target Data:
Anomalous transactions

Learning Category:
Unsupervised Learning
Clustering, Density Estimation



Anomalous Event Detection: NYC Taxis



Input Data:
Timeseries of taxis rides (numerical)

Target Data:
Anomalous events

Learning Category:
Unsupervised Learning

Data source: Numenta Anomaly Benchmark (NAB), from kaggle.com

Types of machine learning

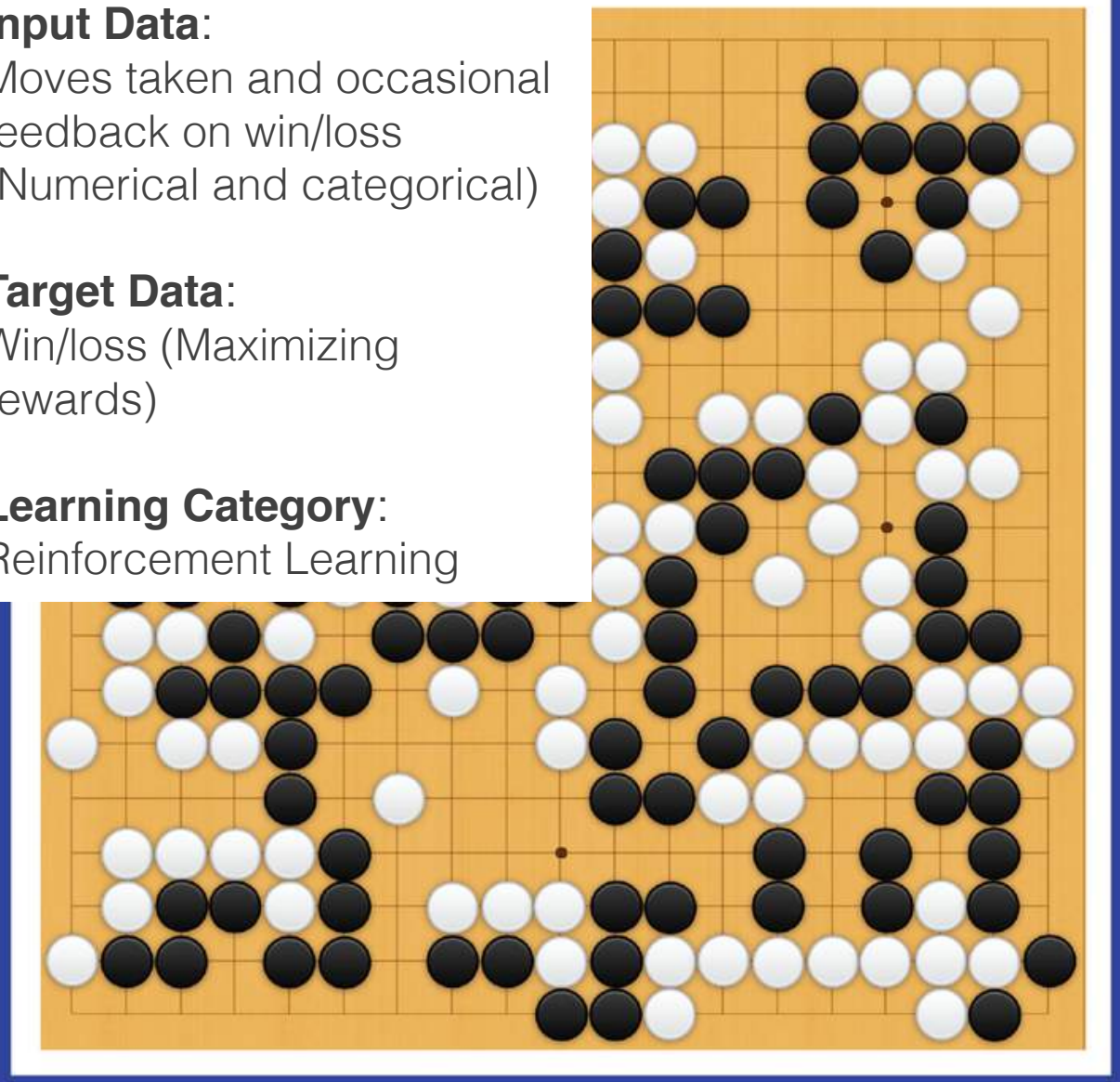
	Unsupervised Learning	Supervised Learning	Reinforcement Learning
Goal	**Describe** ...structure in data	**Predict** …from examples	**Strategize** Learn through interaction
Data available	predictors, $\mathbf{x}$	predictor and response pairs, $(\mathbf{x}, \mathbf{y})$	actions and delayed responses (called rewards)
Examples	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Classification • Regression	• Model-free learning • Model-based learning

Learning a strategy to master games

Input Data:
Moves taken and occasional feedback on win/loss (Numerical and categorical)

Target Data:
Win/loss (Maximizing rewards)

Learning Category:
Reinforcement Learning



Manufacturing – learn to pick up iron cylinders

Input Data:
Actions taken and occasional feedback on success/failure (Numerical and categorical)

Target Data:
Success/failure (Maximizing rewards)

Learning Category:
Reinforcement Learning

Source: MIT Technology Review; Company: FANUC

Types of machine learning

	Unsupervised Learning	Supervised Learning	Reinforcement Learning
Goal	**Describe** ...structure in data	**Predict** …from examples	**Strategize** Learn through interaction
Data available	predictors, $\mathbf{x}$	predictor and response pairs, $(\mathbf{x}, \mathbf{y})$	actions and delayed responses (called rewards)
Examples	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Classification • Regression	• Model-free learning • Model-based learning

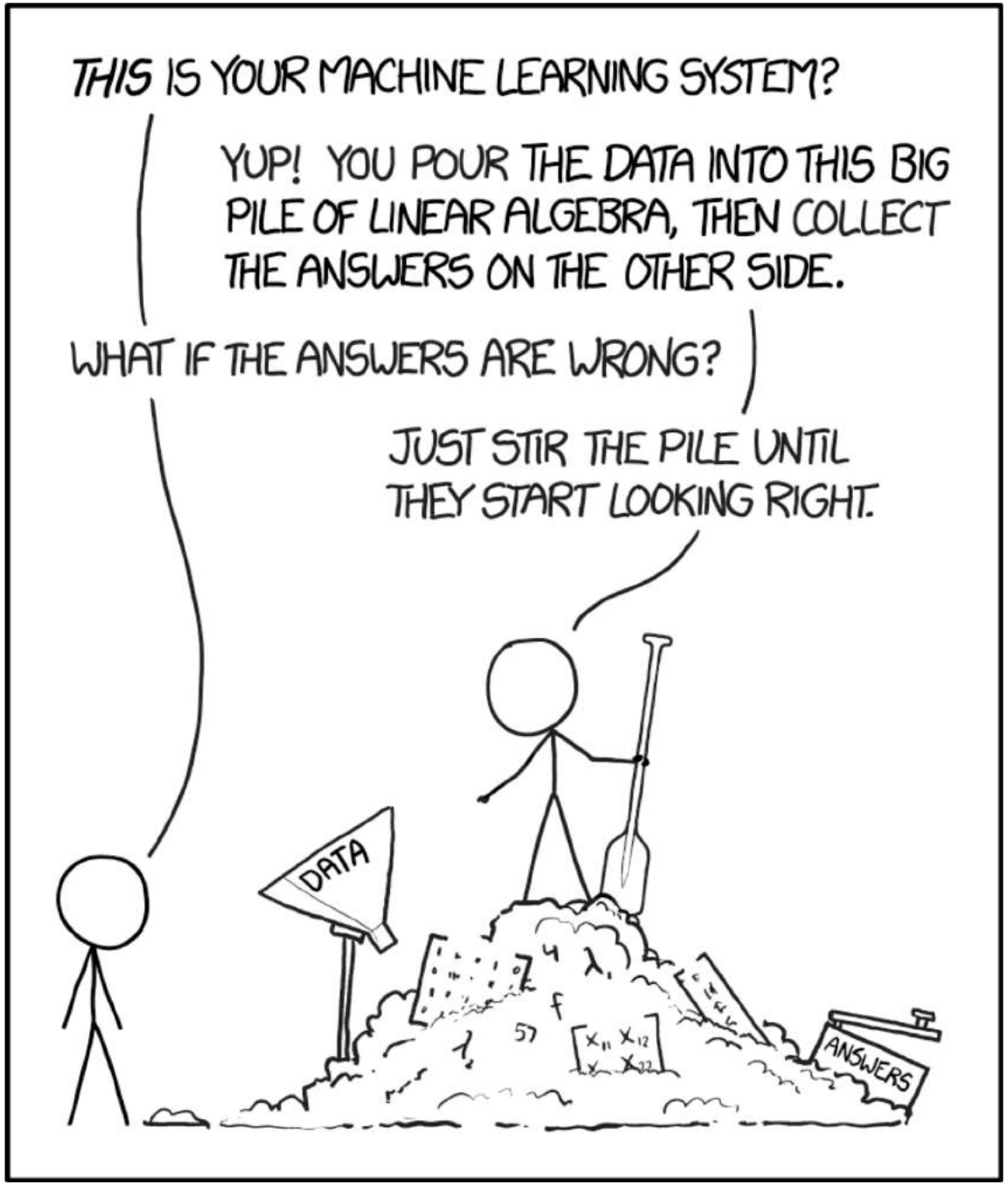
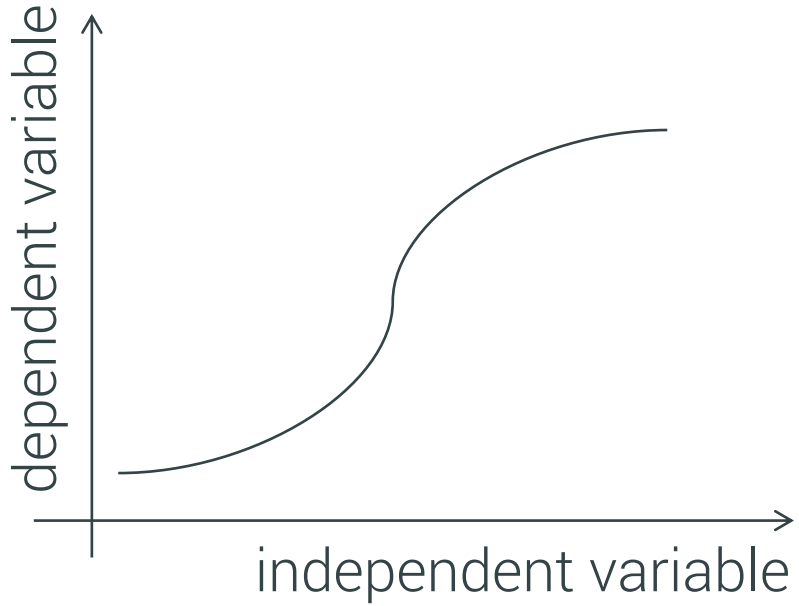


Image: xkcd.com

A Taste of Supervised Learning

Classification and Regression

Common language



independent variable

input

predictor

feature

x

dependent variable

output

response

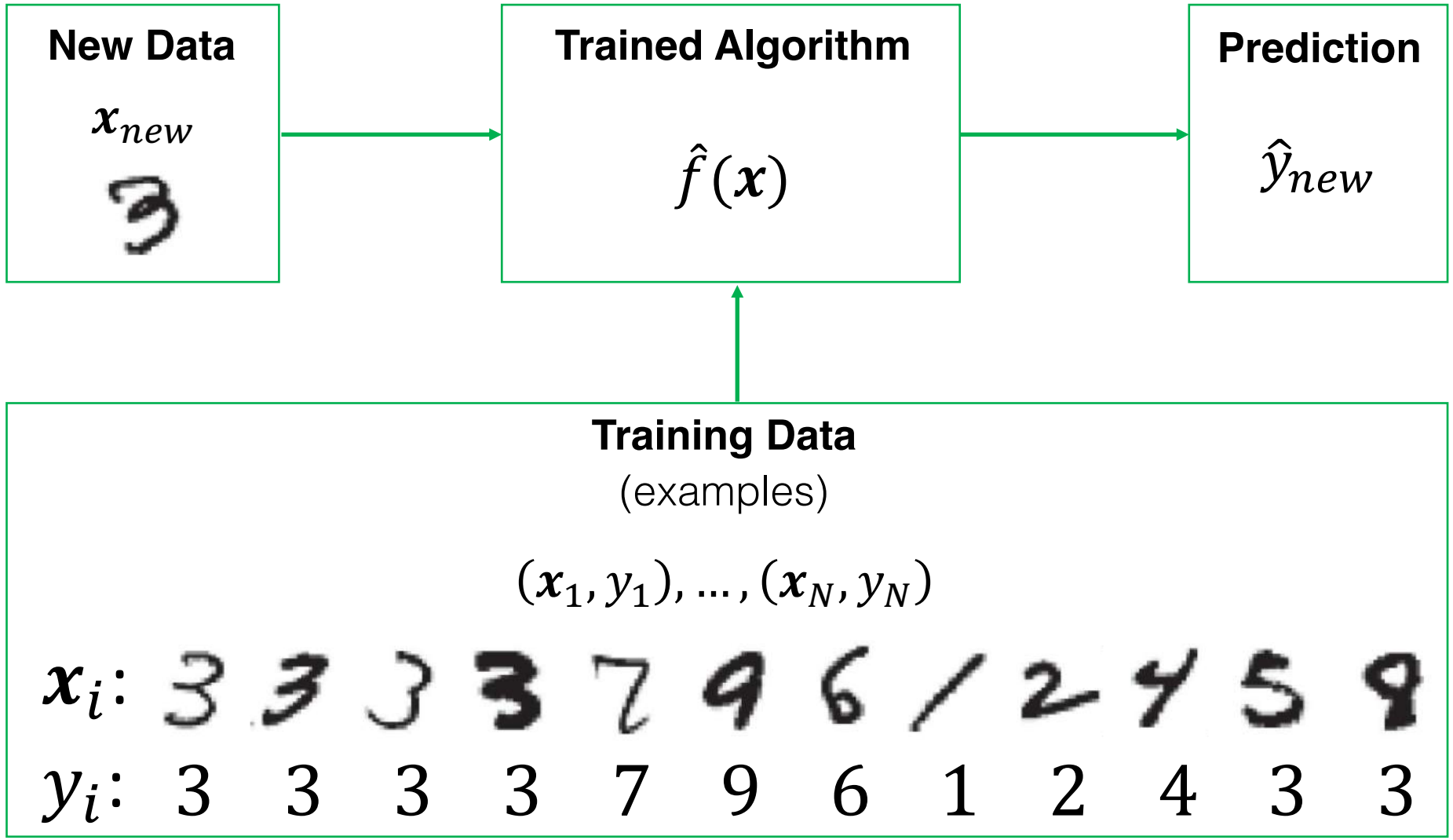
target

y

Supervised learning

Example:
Digits classification

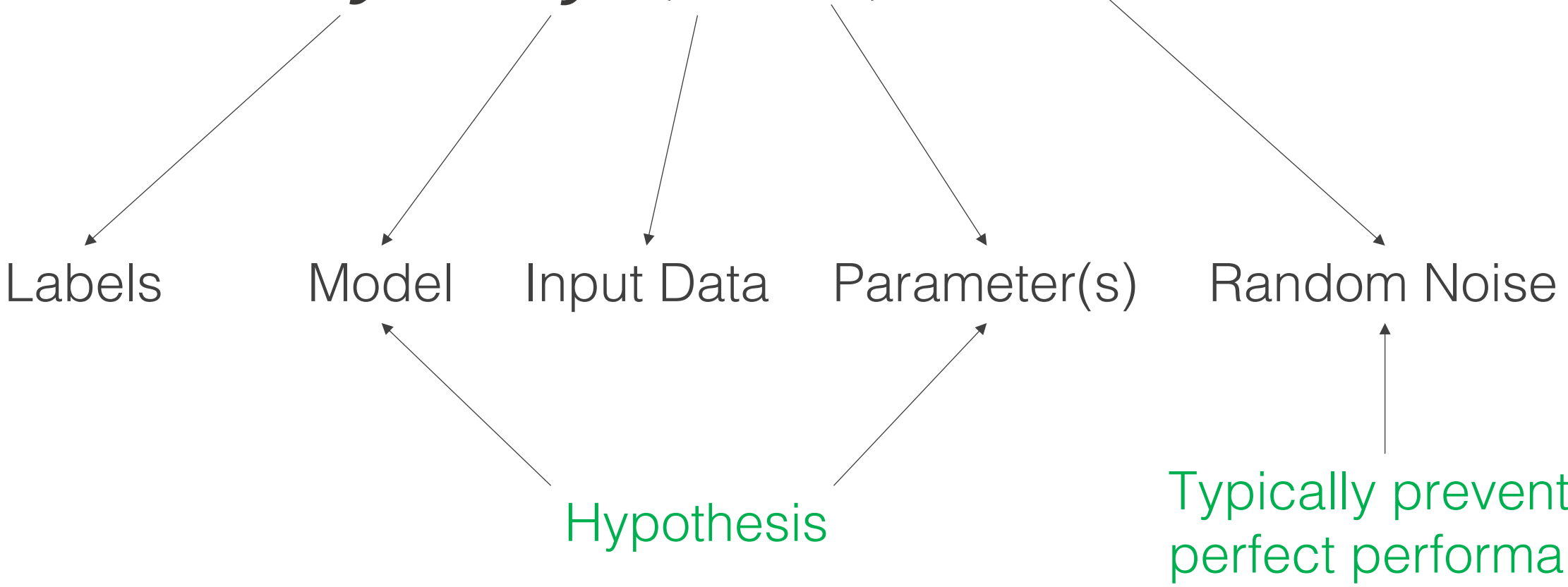
Objective: create an algorithm that predicts well



Supervised machine learning model

We search for the model that best fits our data

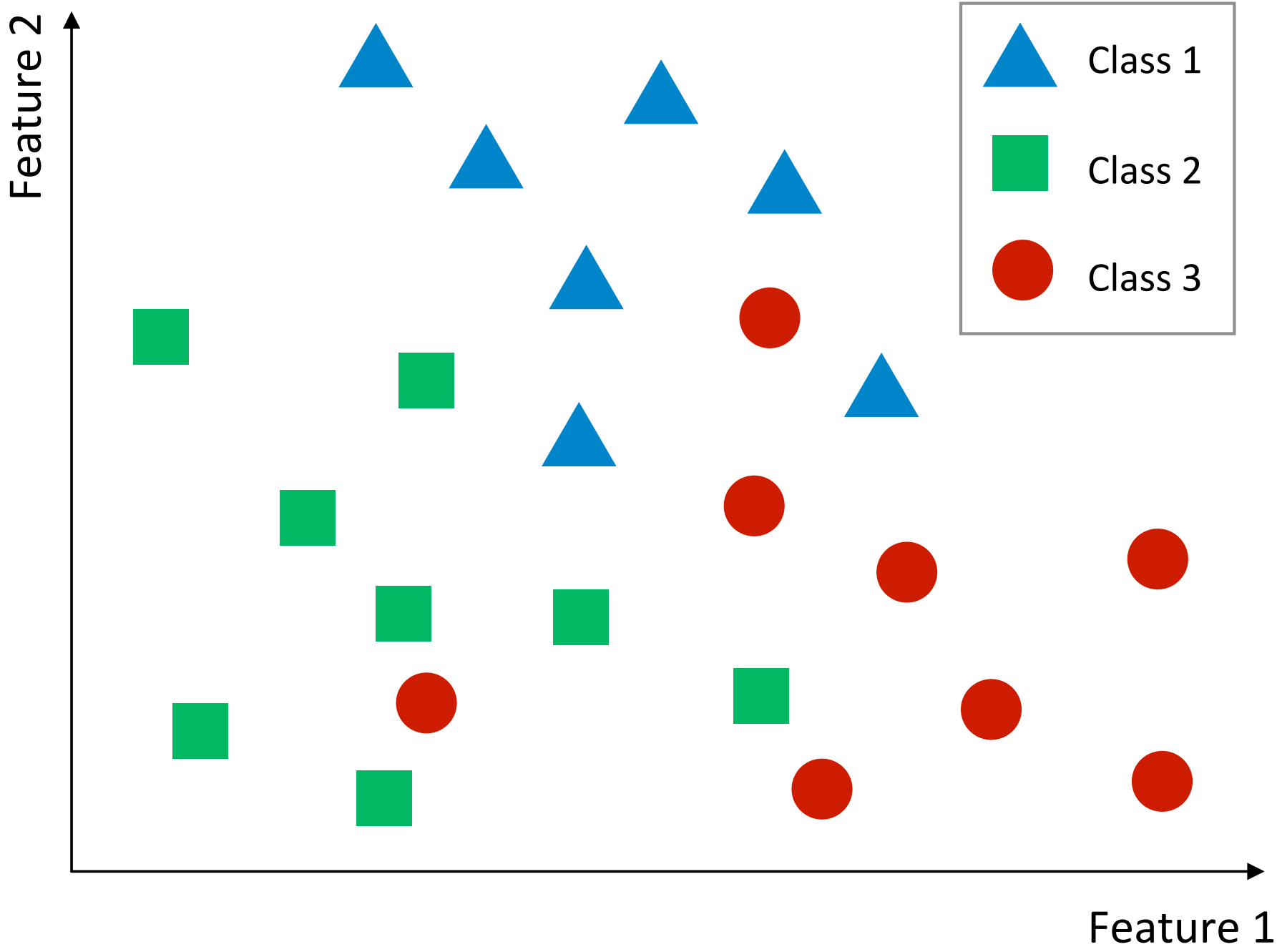
$$y = f(x, \boldsymbol{w}) + \epsilon$$



K-Nearest Neighbors

Classification and Regression

K Nearest Neighbor Classifier

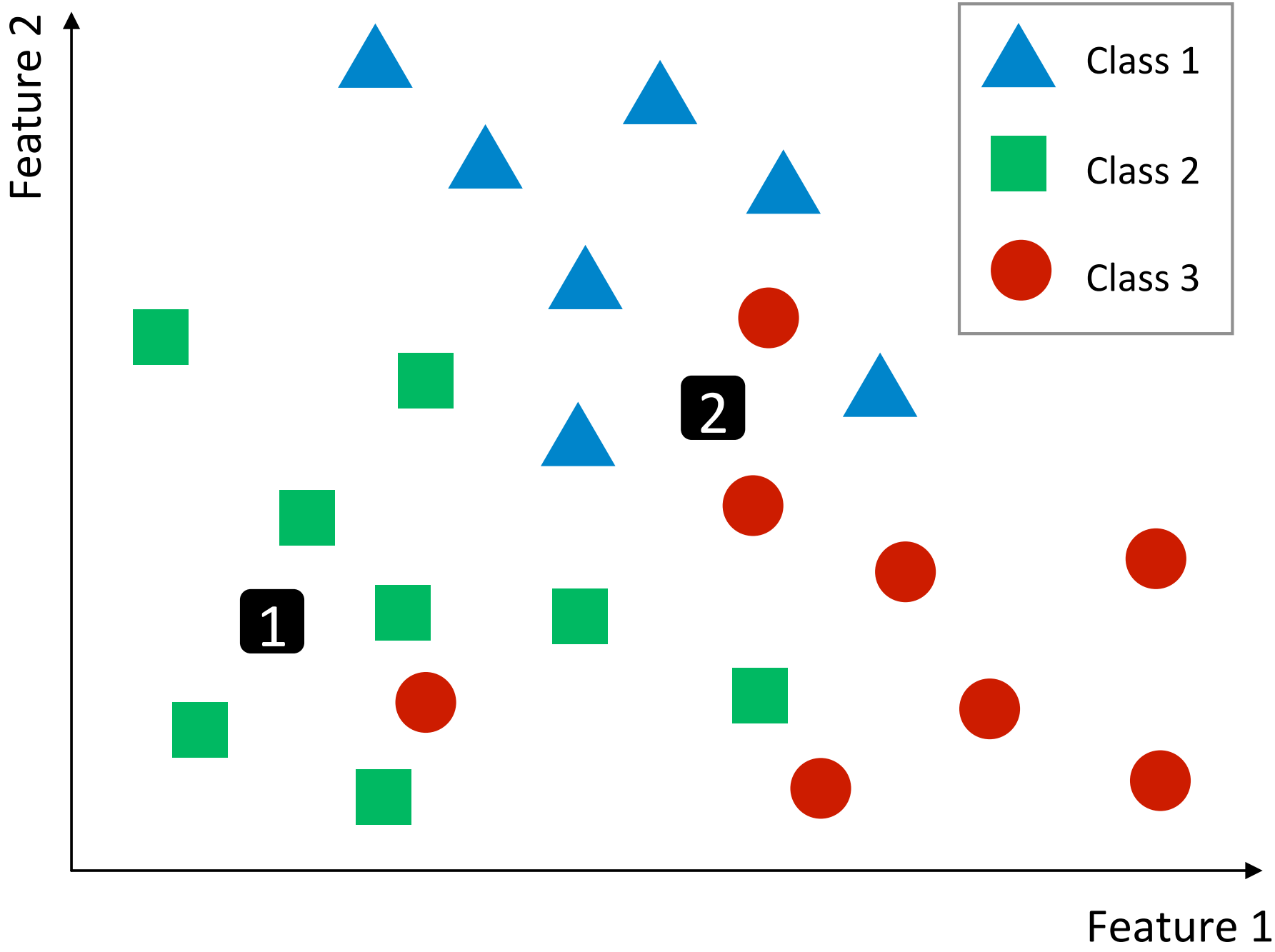


Step 1:
Training

Every new data point is a model parameter

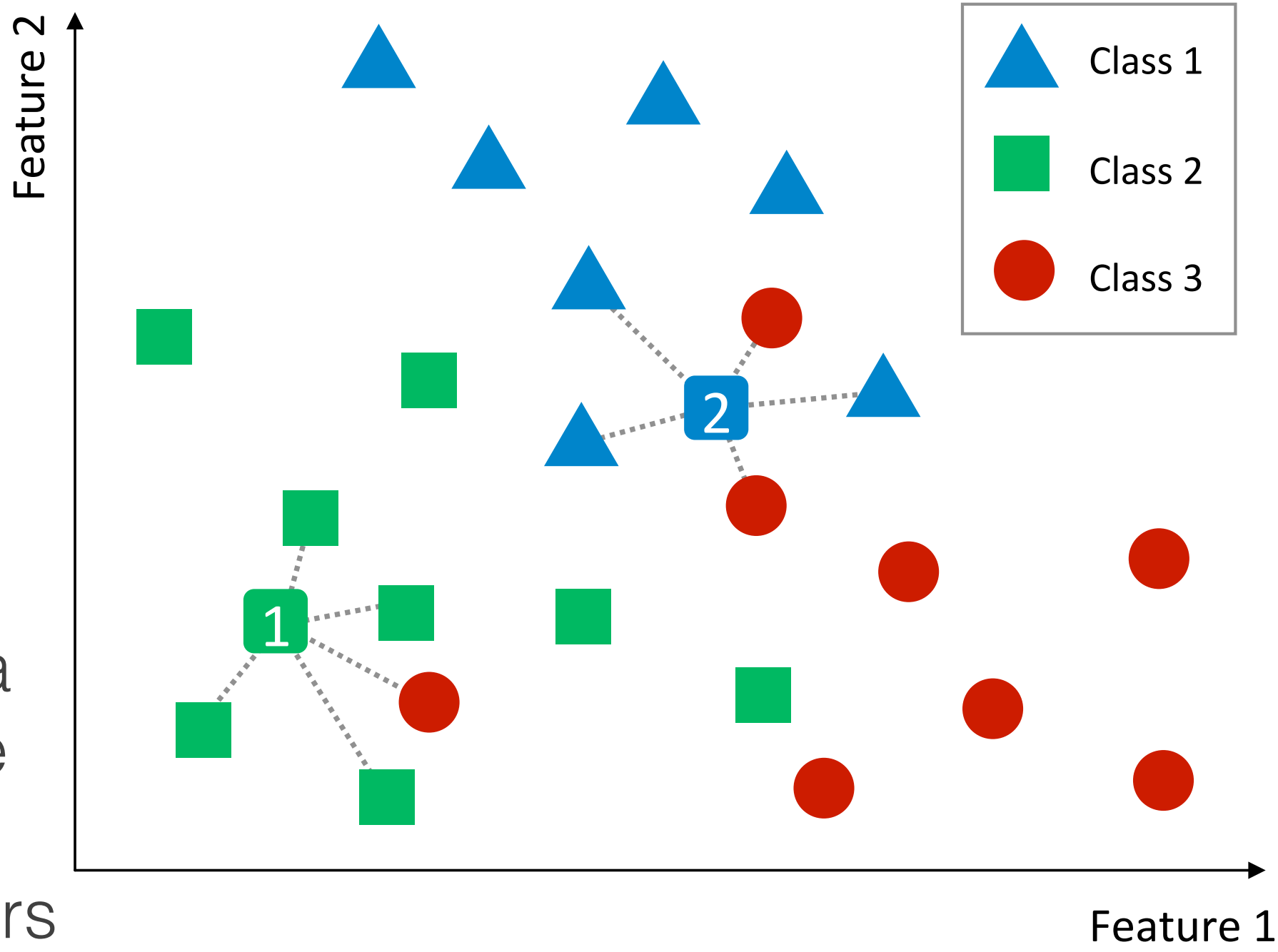
K Nearest Neighbor Classifier

Step 2:
Place new (unseen) examples in the feature space

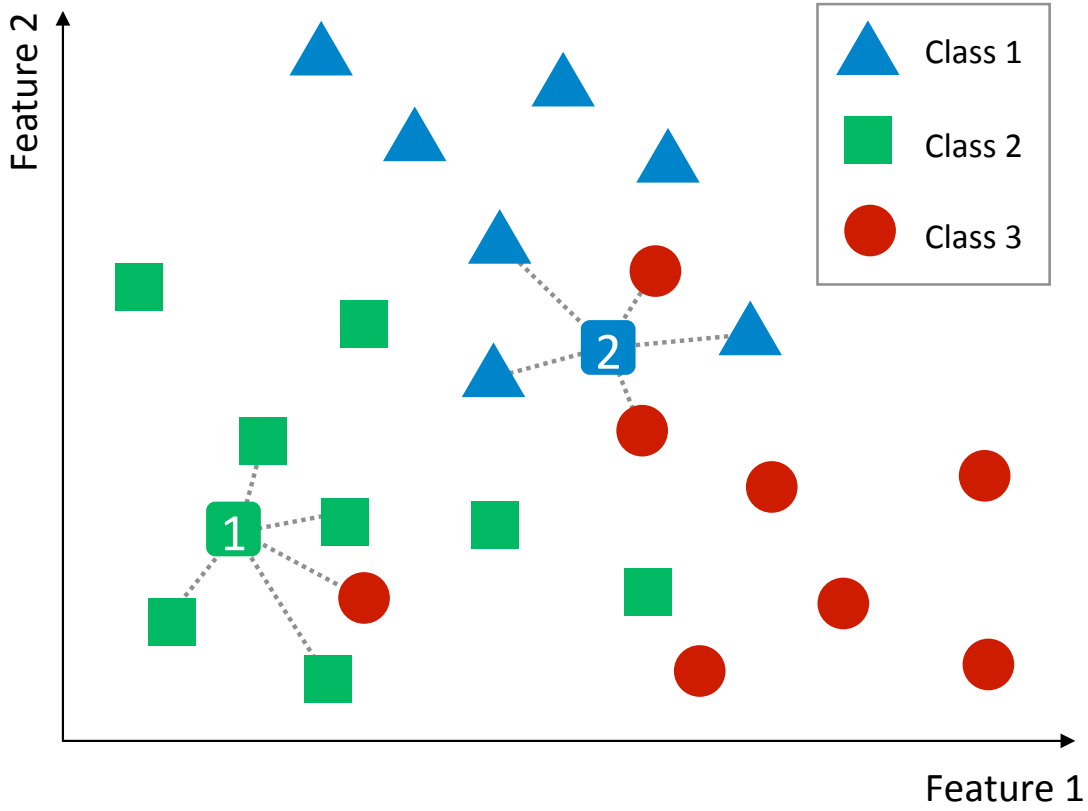


K Nearest Neighbor Classifier

Step 3:
Classify the data by assigning the class of the k nearest neighbors



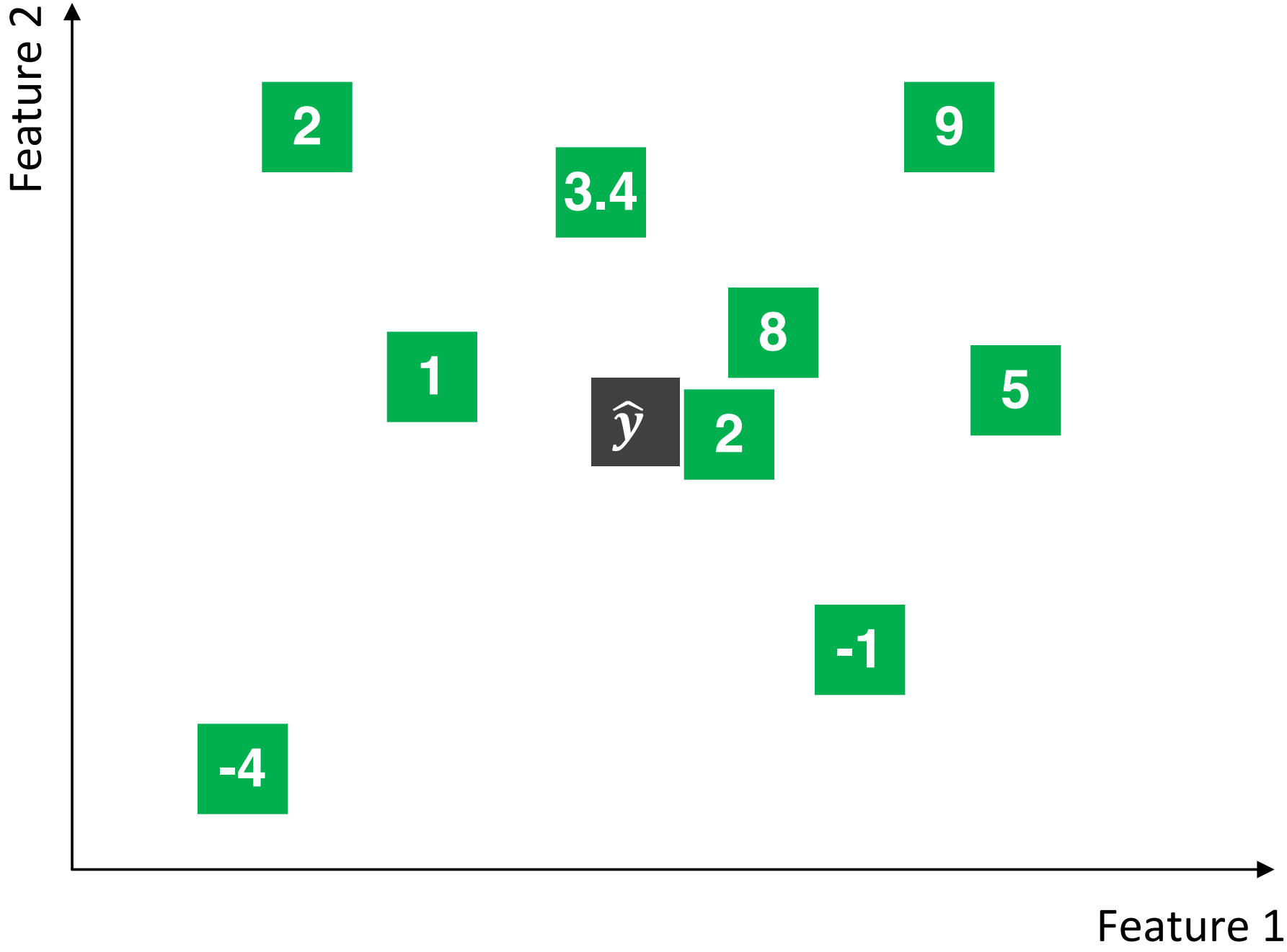
K Nearest Neighbor Classifier



Score vs Decision :
For 5-NN, the confidence score that a sample belongs to a class could be: {0,1/5,2/5,3/5,4/5,1}

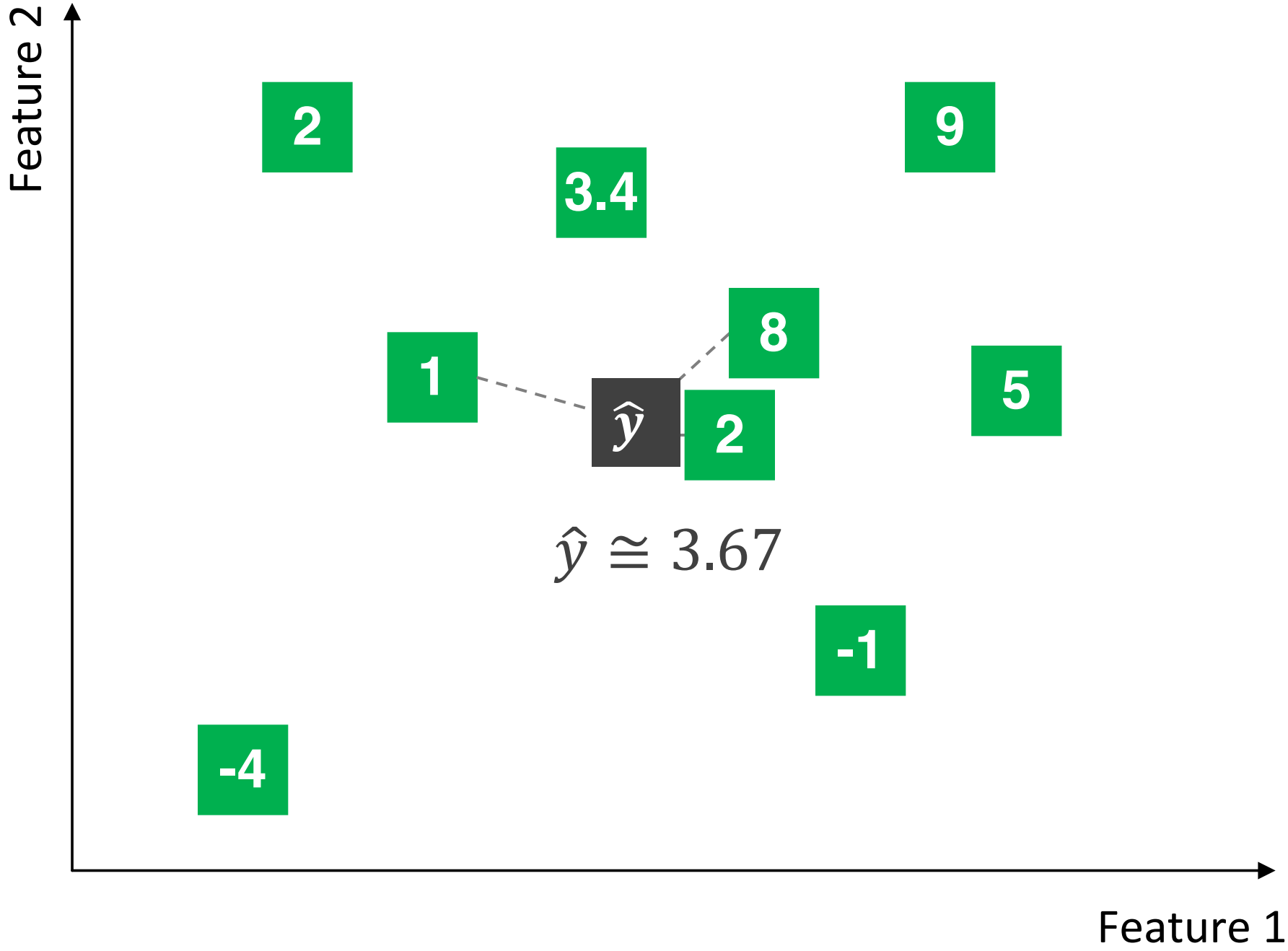
Decision Rule:
If the confidence score for a class > threshold, predict that class

K Nearest Neighbor Regression



K Nearest Neighbor Regression

$$\hat{y} = \frac{1}{k} \sum_{y_i \in \{\text{k nearest}\}} y_i$$



Summarizing supervised learning

Let's review and bring it all together

Components of supervised learning

Input	$\boldsymbol{x}$	
Output	y	
Training Data	$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)$	
Target function	$f(\boldsymbol{x}) \rightarrow y$	This is unknown, but the best you could ever do
Hypothesis set	$f_i(\boldsymbol{x}) \rightarrow \hat{y}$	Functions to consider in trying to approximate $f(x)$
Learning algorithm	Optimization technique that searches the hypothesis set for the function f_i that best approximates f (typically by choosing parameters in a model)	

Supervised Learning

Unobservable | Observable

Data Generating Process

$p(X, Y)$

Training Data

$(x_1, y_1), \ldots, (x_N, y_N)$

Learning Algorithm

Chooses a hypothesis, $\hat{f} = f_i$ based on the training data such that $\hat{f}(x) \approx f(x)$

Final Hypothesis

$\hat{f}(x) \rightarrow \hat{y}$

predictions

Target Function

The best function predicting y from x

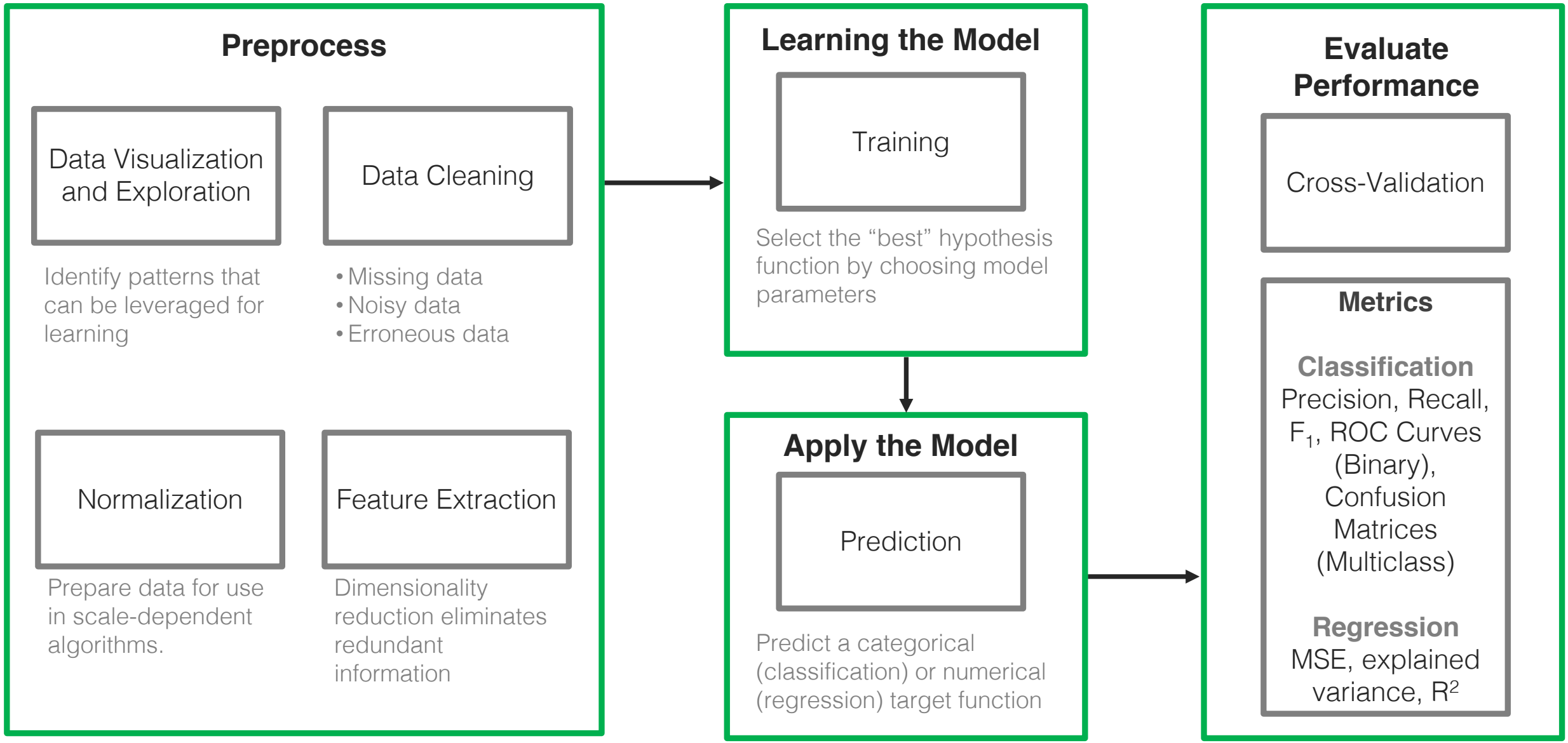
$f(x) \rightarrow y$

Hypothesis Functions Set

$f_1, f_2, f_3, \ldots$

- Need to select the hypothesis functions (models to train)
- Need to select the learning algorithm (for fitting the models to the data)

Supervised learning in practice



Want to learn more?

List of additional resources:

http://www.kylebradbury.org/datascience.html

ENERGY

data analytics lab

Duke ENERGY INITIATIVE

iiD SOCIAL SCIENCE RESEARCH INSTITUTE