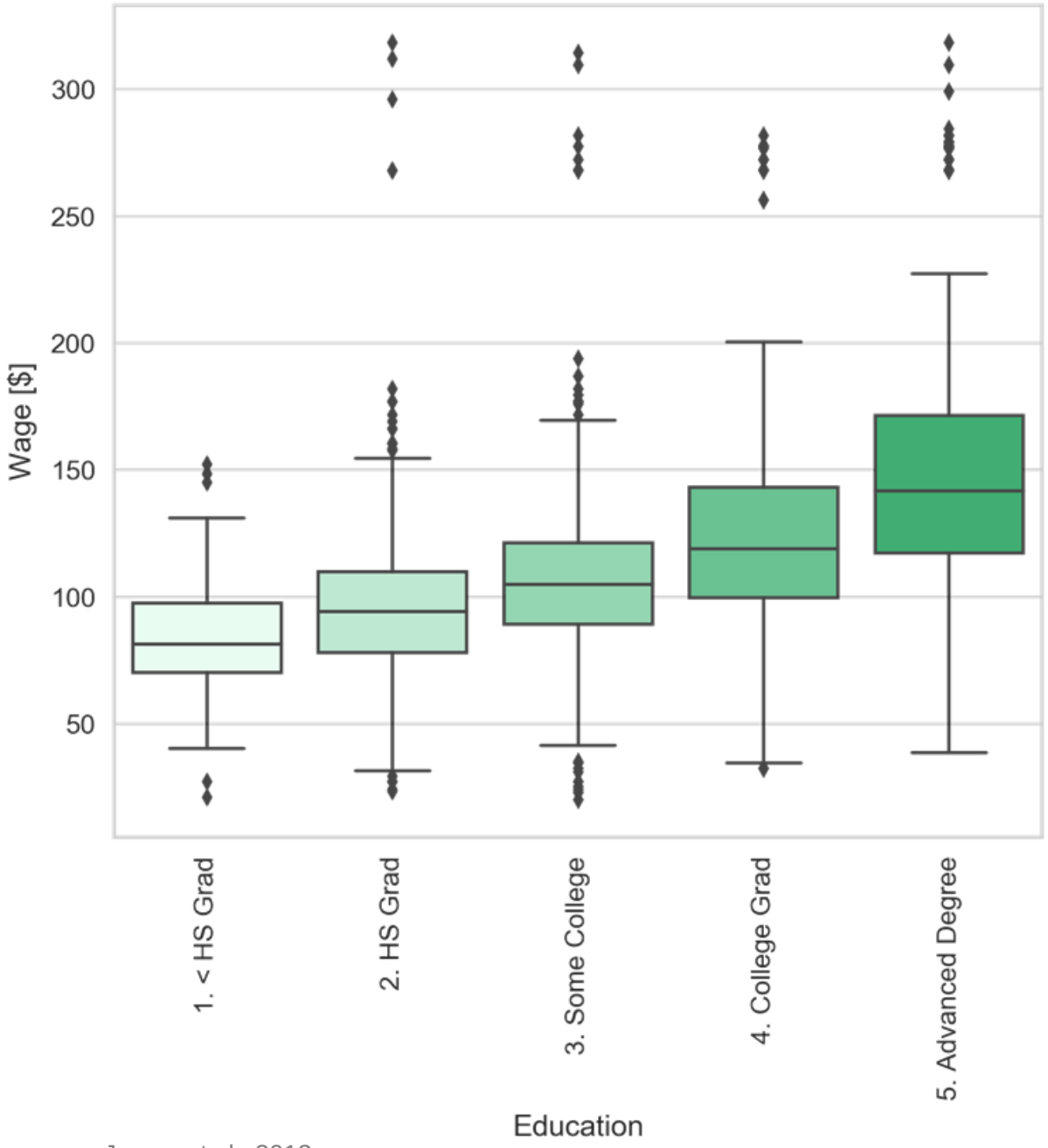
What is machine learning?

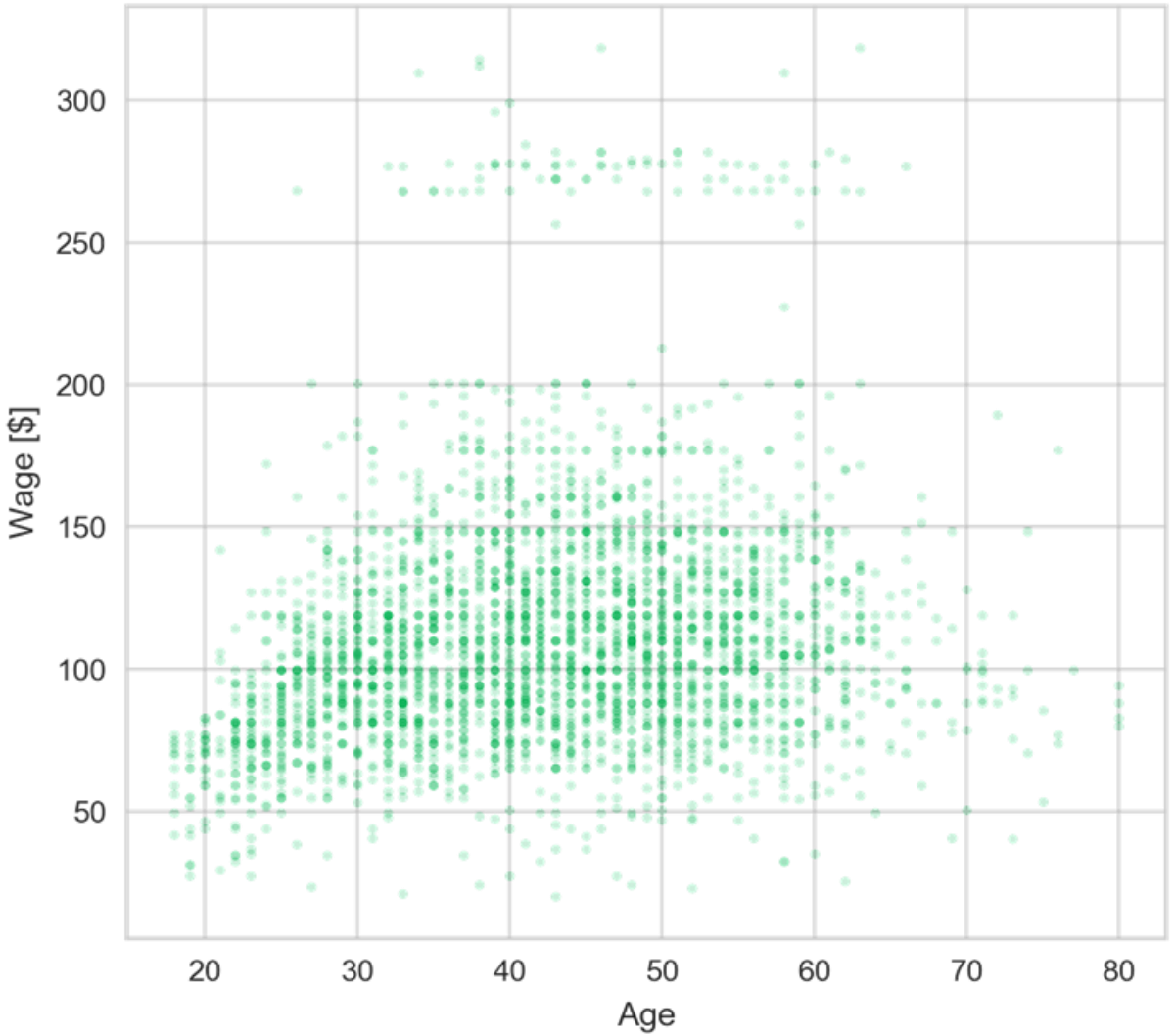
Lecture 01

Introductions

1. Register for Poll Everywhere here: https://bit.ly/2RBKLA3
   Be sure to use (a) your real name, AND (b) your Duke netid-based email address (e.g. mjw8@duke.edu)

2. Complete the poll at:
   https://pollev.com/kylebradbury





Wage data from workers in the mid-Atlantic region

How do you predict how much someone will make?

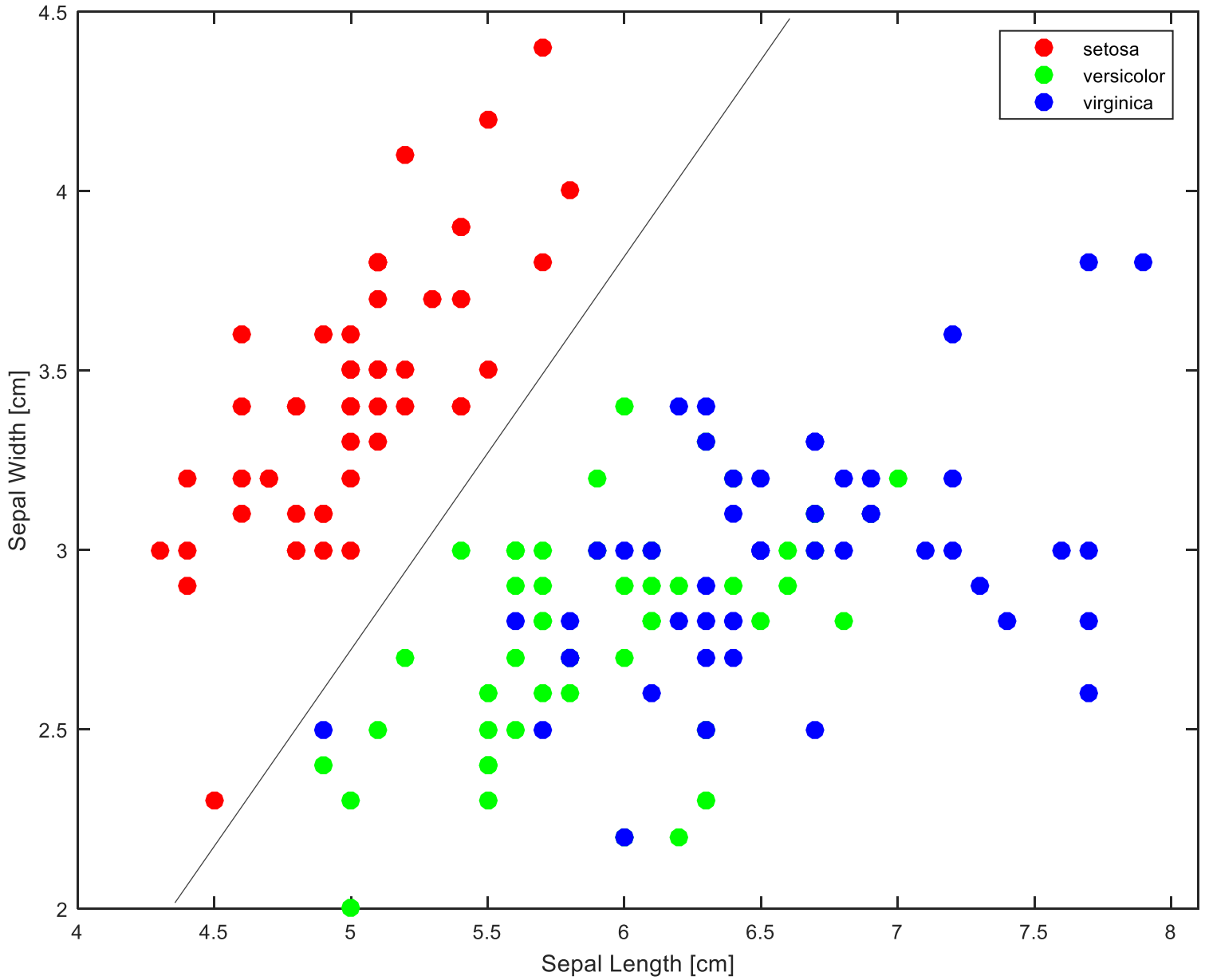
Data source: James et al., 2013

How can you tell these flowers apart?

Iris setosa

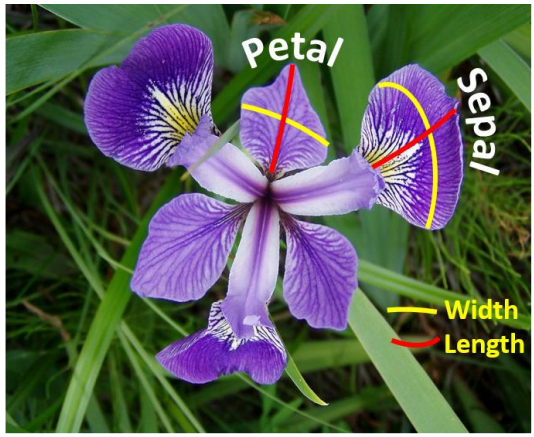
Iris virginica

Image Sources: Srishti Sawla (setosa) and Ivo Dinov, University of Michigan SOCR (versicolor and verginica)



setosa

versicolor



virginica

Data Source: Fisher Iris Data
Image Sources: Srishti Sawla (setosa) and Ivo Dinov, University of Michigan SOCR (versicolor and verginica)

Challenges

What is this?

Image by artist Hikaru Cho

We **generalize** from past experiences

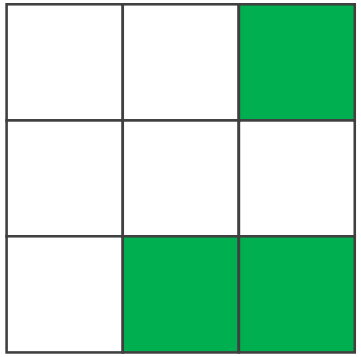
Image: “It’s not what it seems” by artist Hikaru Cho

our data must be

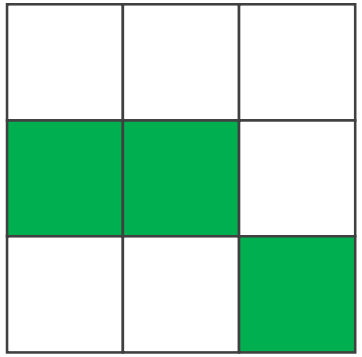
representative

Predict which class x_{new} belongs to…

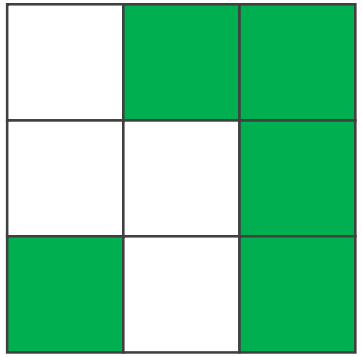
$f(\boldsymbol{x}) = 1$



$\boldsymbol{x}_0$

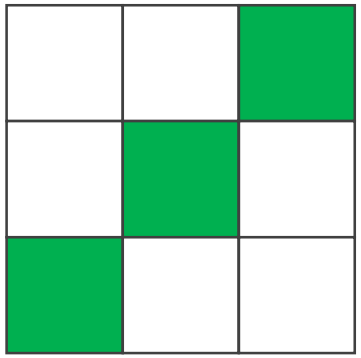


$\boldsymbol{x}_2$

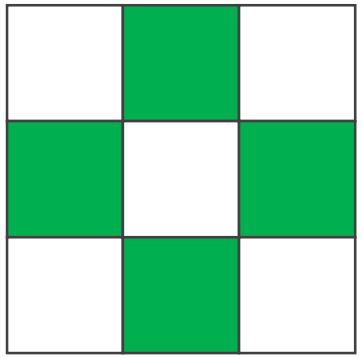


$\boldsymbol{x}_4$

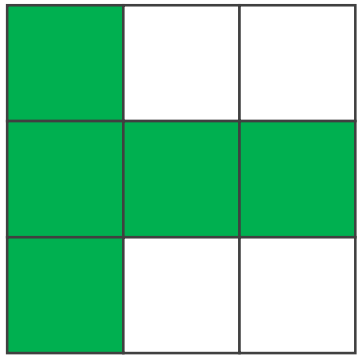
$f(\boldsymbol{x}) = 0$



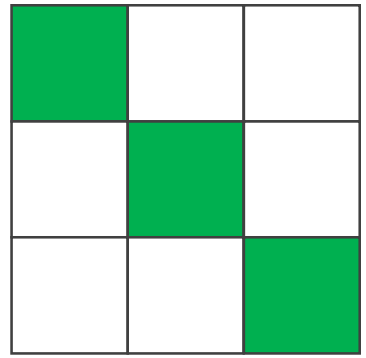
$\boldsymbol{x}_1$



$\boldsymbol{x}_3$



$\boldsymbol{x}_5$



$\boldsymbol{x}_{new}$

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

Example credit: Yaser Abu-Mostafa, 2012

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

There are often **many** models that fit your training data similarly well

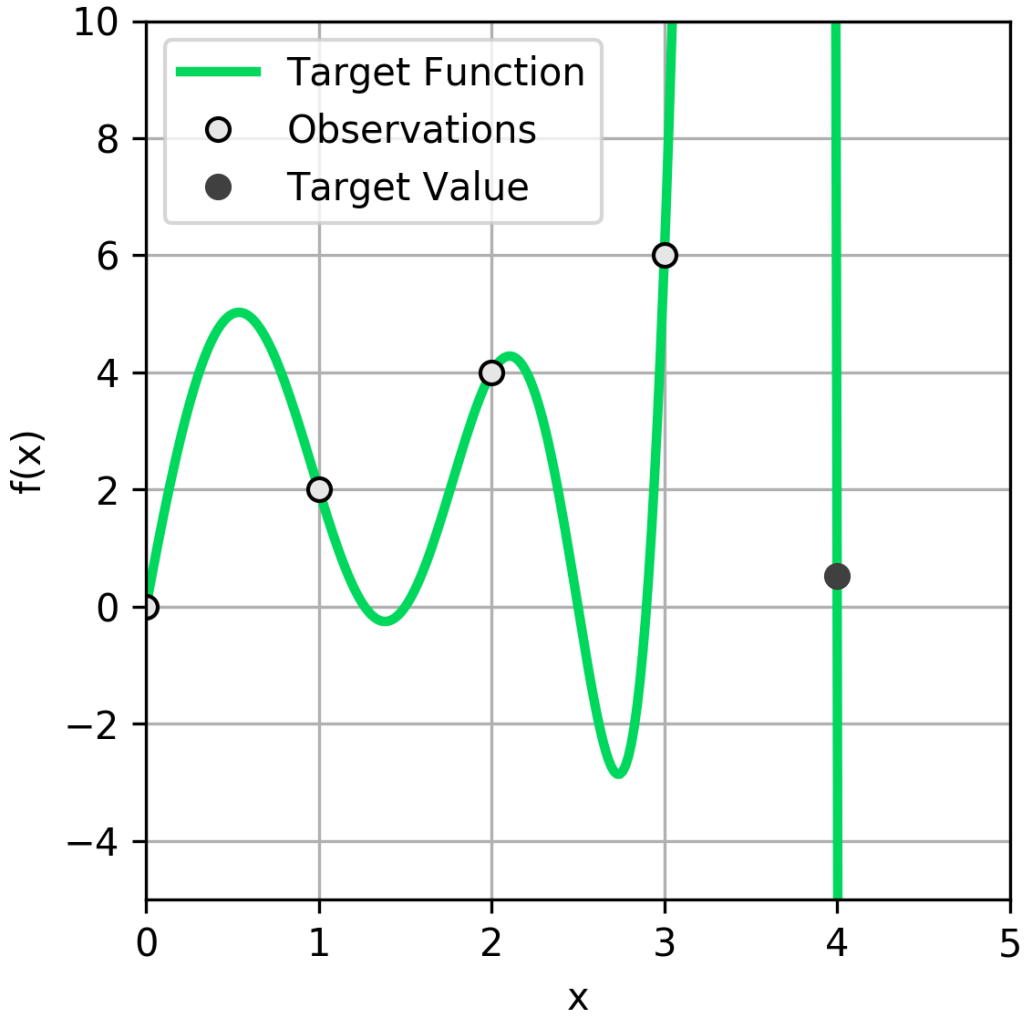
So how do we choose which to use?

the best models **generalize** well

Predict the next value in the sequence…

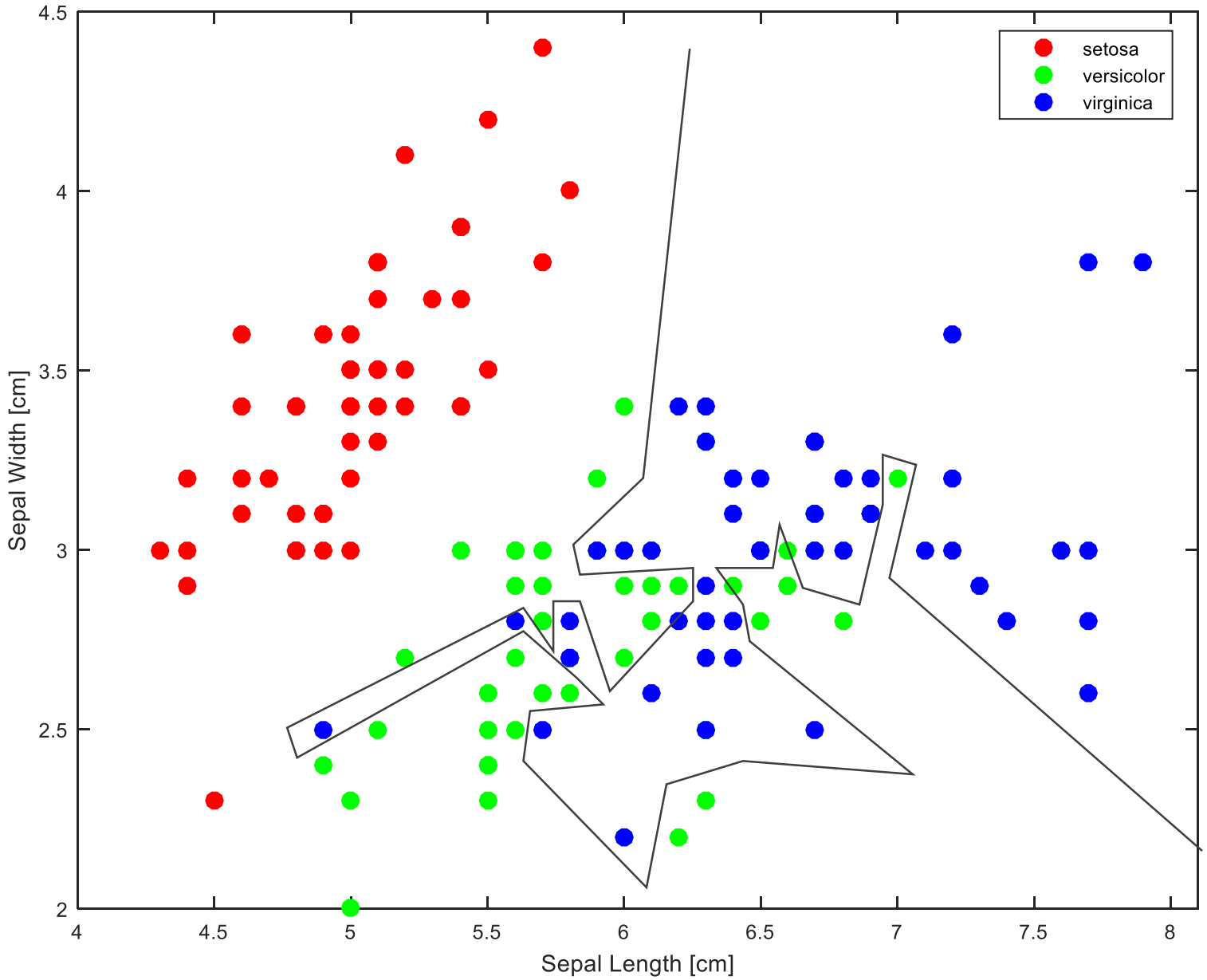
x	0	1	2	3	4
$f(x)$	**0**	**2**	**4**	**6**	**?**

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



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

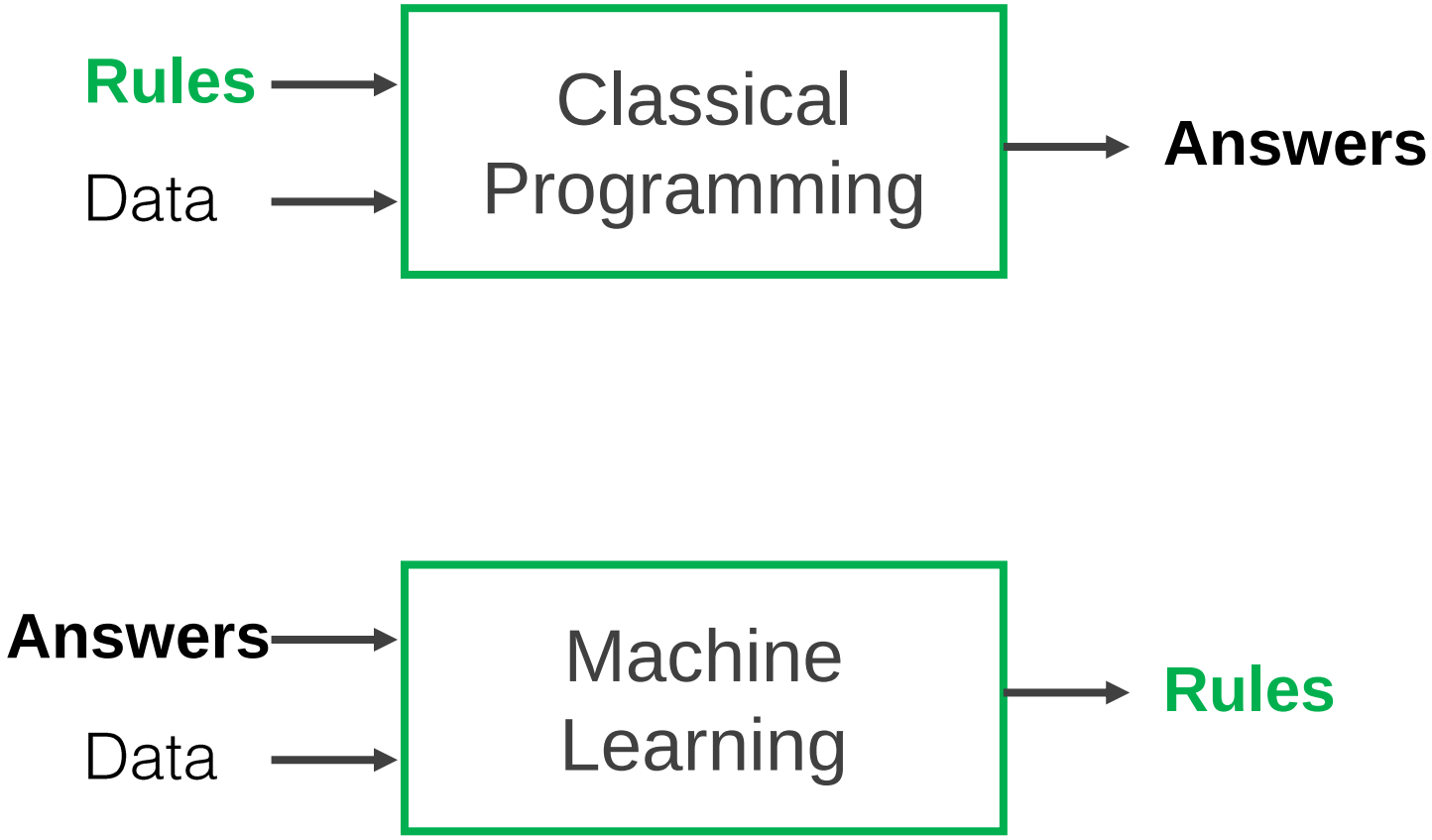
Data Source: Fisher Iris Data

Complex models overfit to the data

overfit works against generalization

Learning simpler representations of data enables learning

Machine learning suggests an alternative programming paradigm



François Chollet, *Deep Learning with Python*, 2017

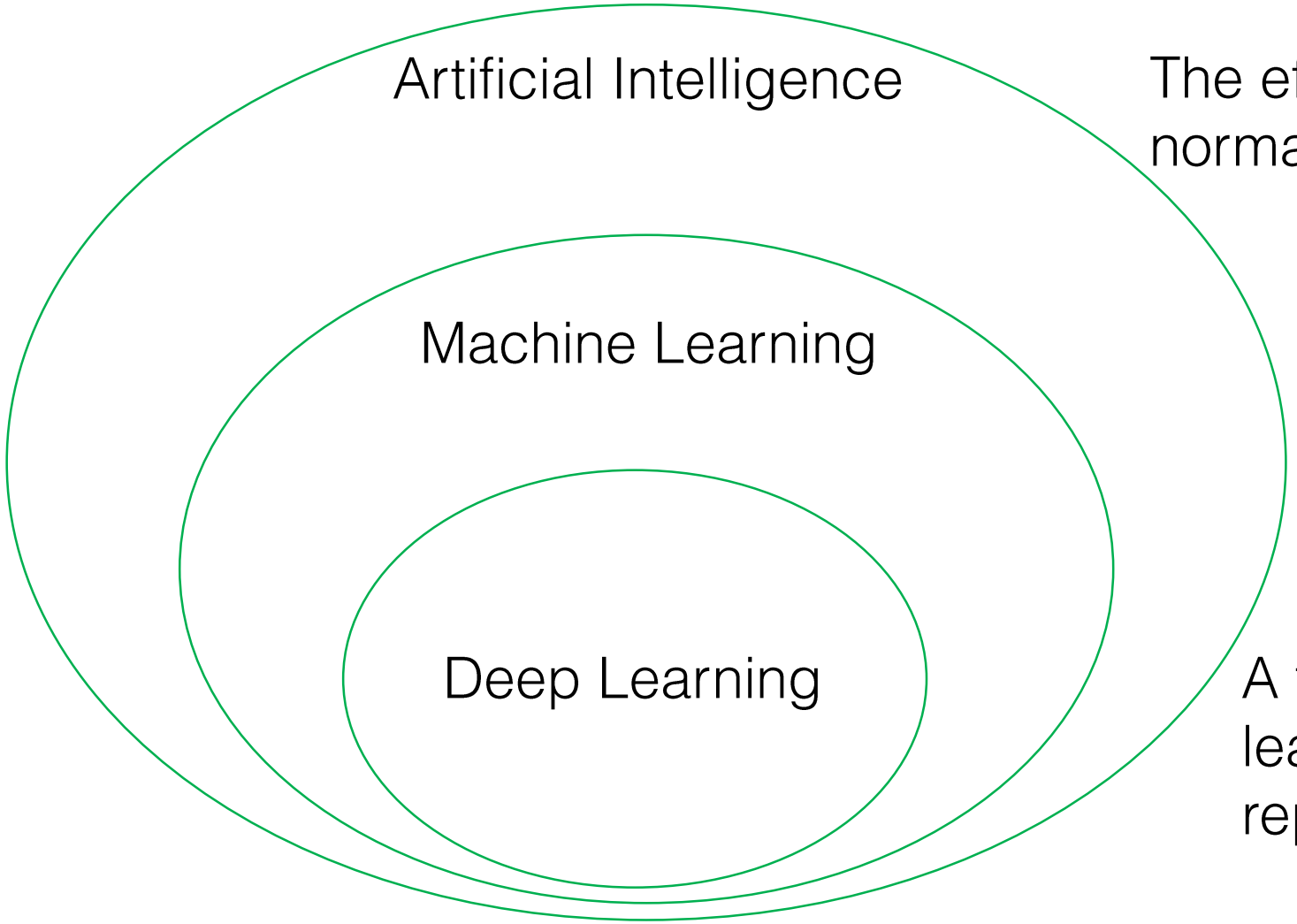
What is machine learning?

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

…**based on** data and past experience…

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

What is machine learning?



The effort to automate intellectual tasks normally performed by humans

Techniques to automatically describe, predict, and strategize based on data, without being explicitly programmed

A type of machine learning that learns successive layers of data representations

François Chollet, *Deep Learning with Python*, 2017

Types of machine learning tools

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

Types of machine learning

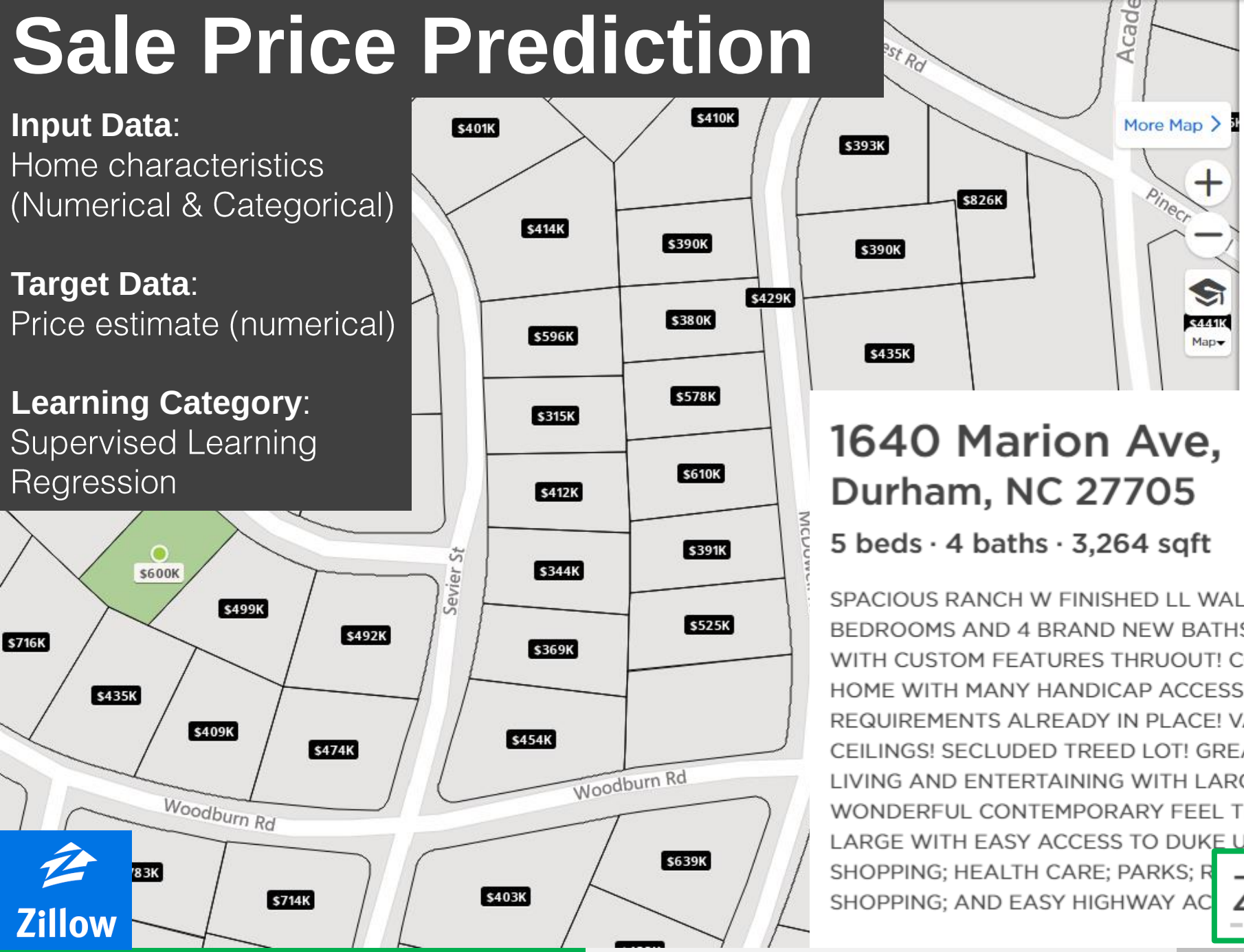
	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Goal	**Predict** …from examples	**Describe** ...structure in data	**Strategize** learn by trial and error
Data	(x, y)	x	delayed feedback
Types	• Classification • Regression	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Model-free learning • Model-based learning

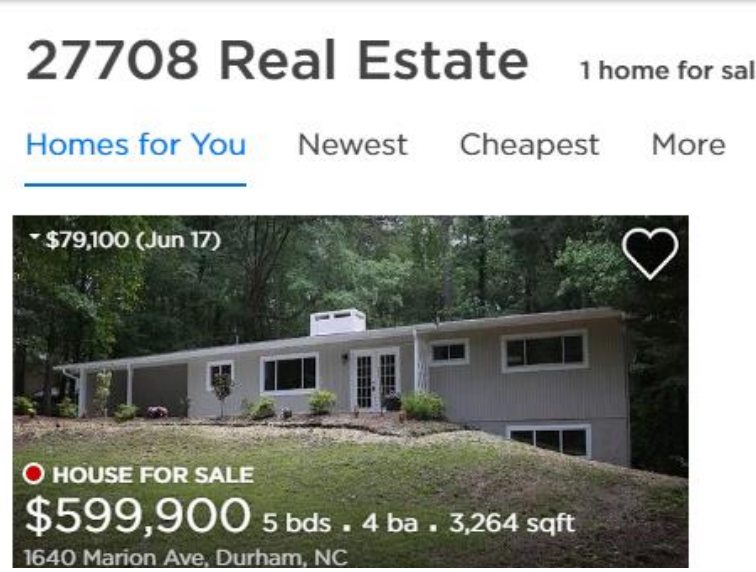
Sale Price Prediction

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

Zestimate®: $619,585

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

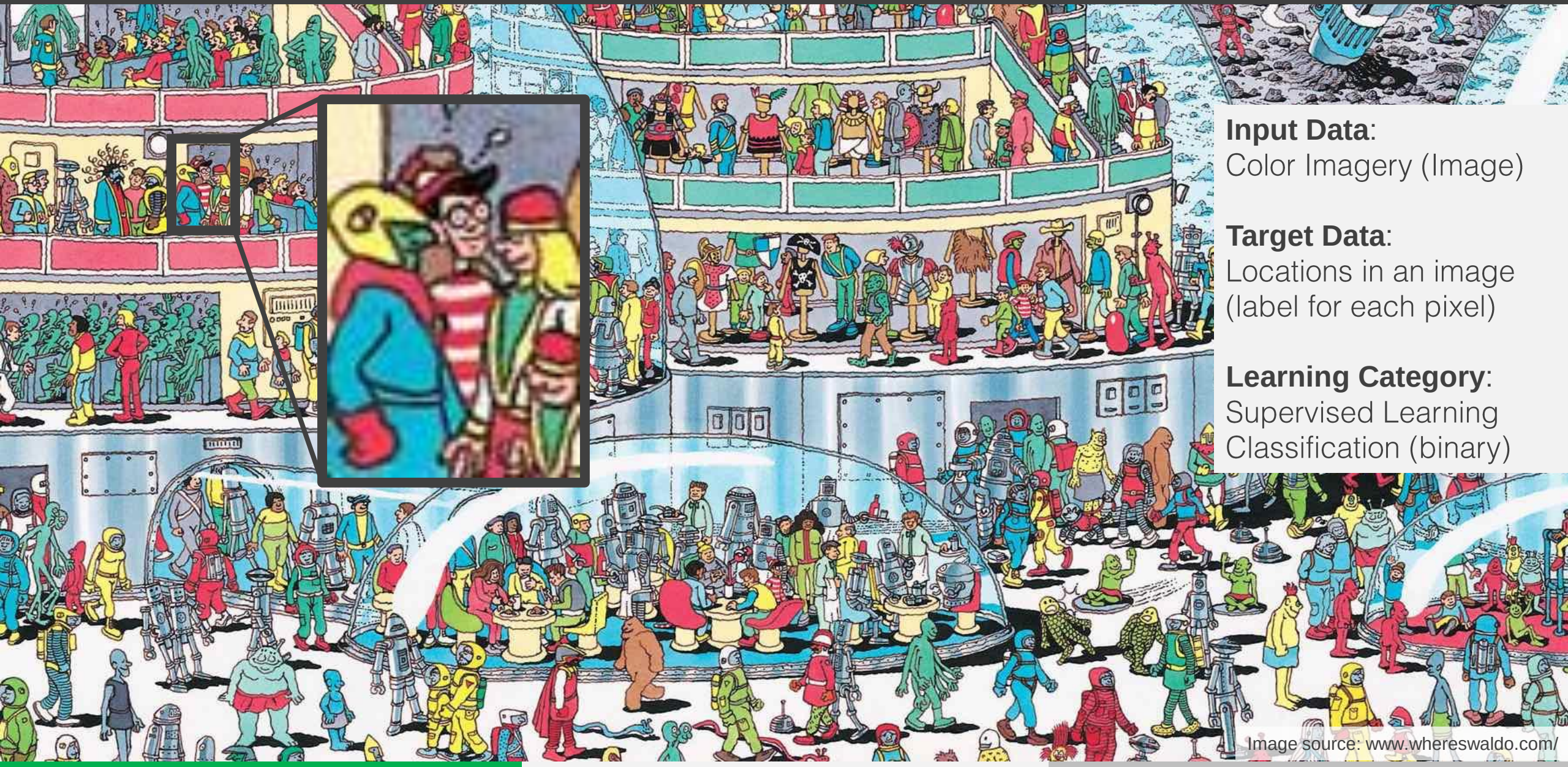
Spam Filters

Input Data:
Email text (text)

Target Data :
Spam/not spam (category)

Learning Category:
Supervised Learning
Classification (binary)

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

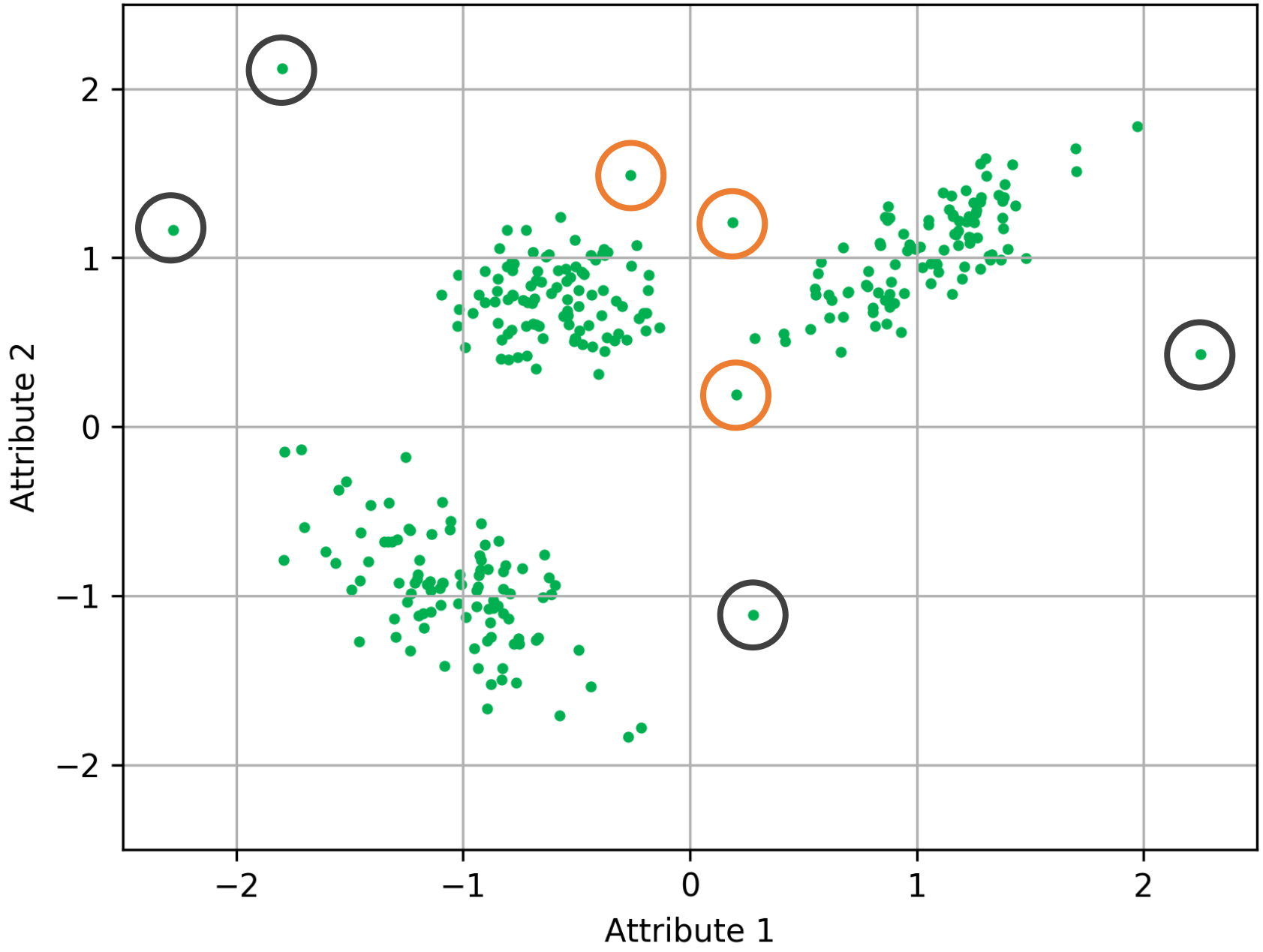
Light Pole
access to electricity

Credit Fraud

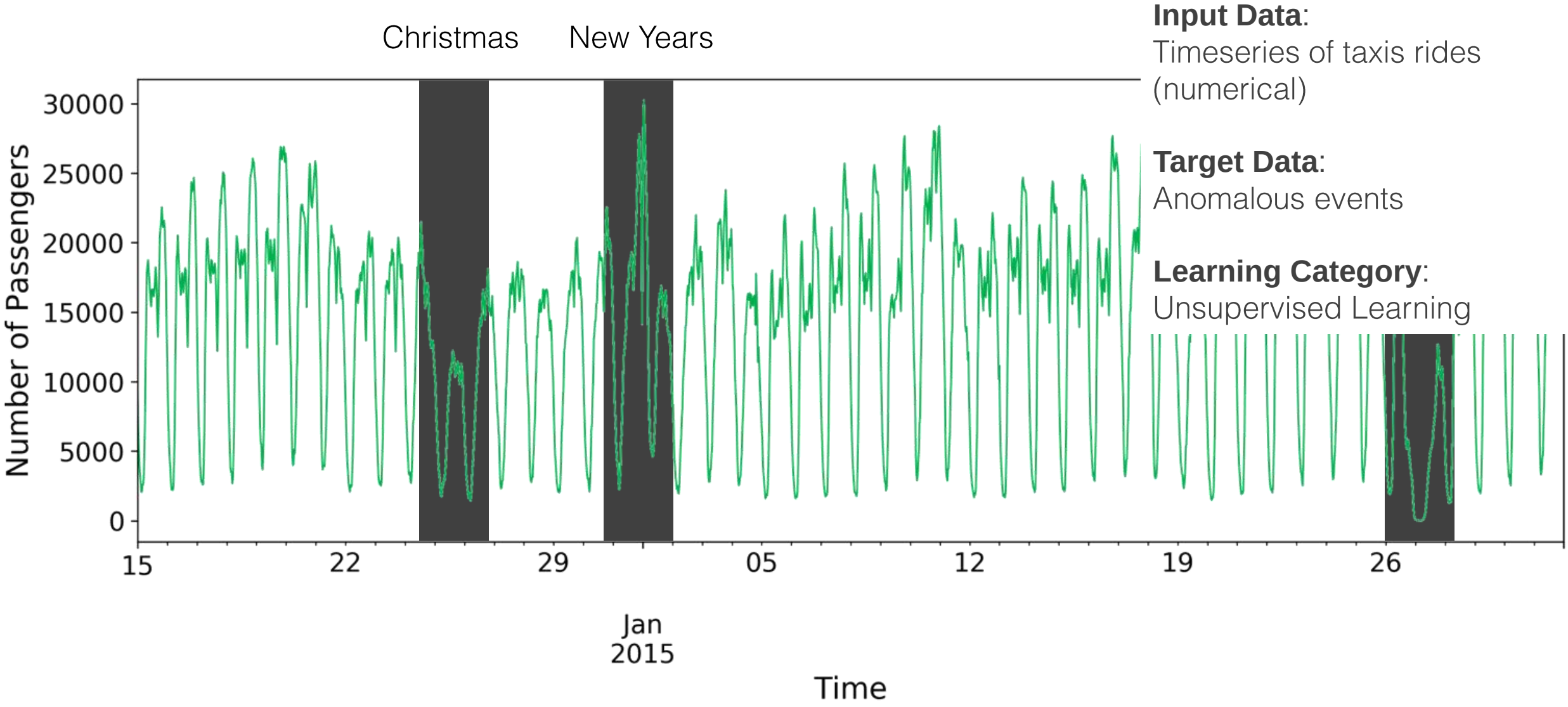
Input Data:
Account transactions, dates, locations, demographic information
(Numerical and categorical)

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

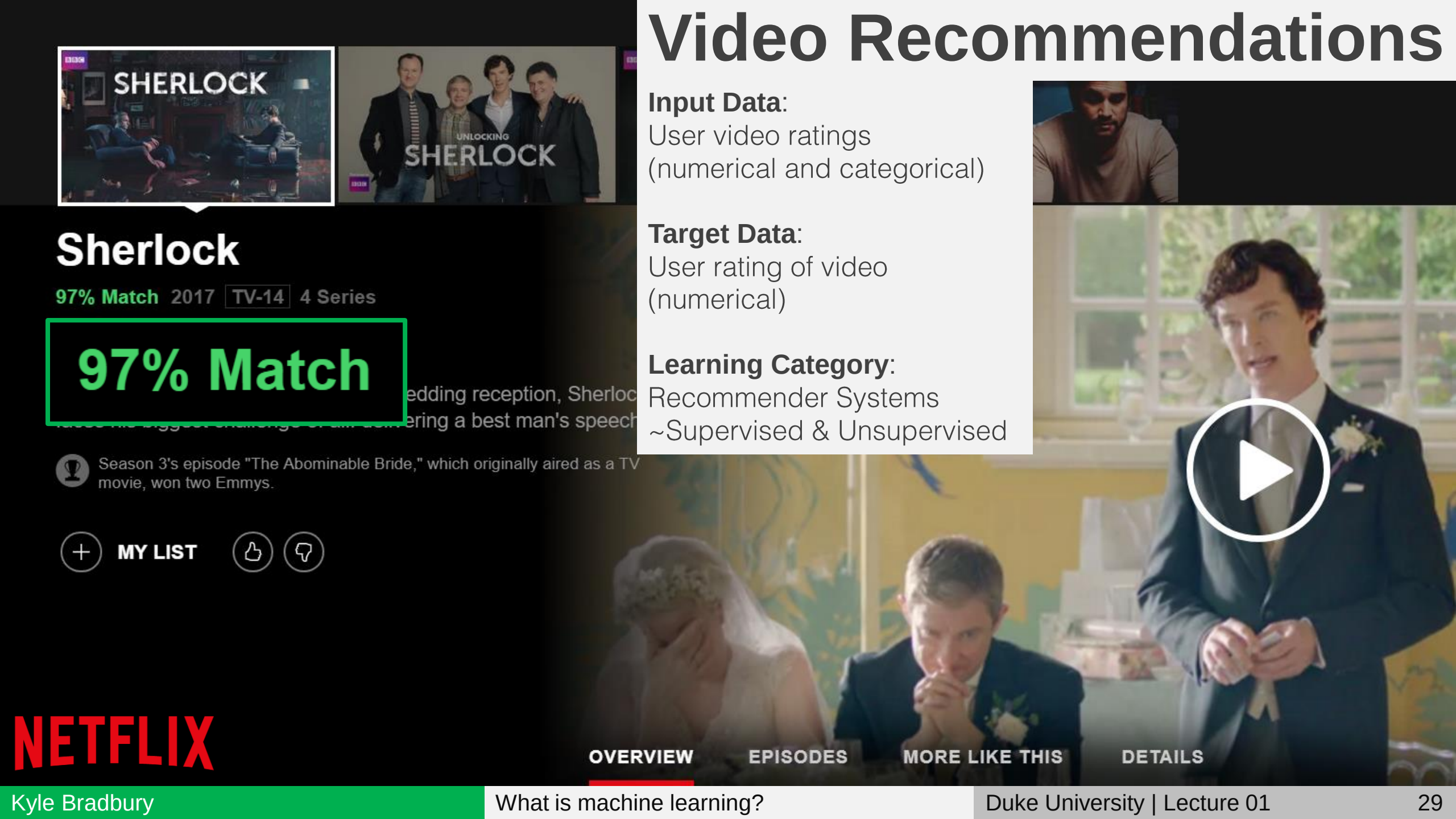
Data source: Numenta Anomaly Benchmark (NAB), from kaggle.com

Video Recommendations

Input Data:
User video ratings
(numerical and categorical)

Target Data:
User rating of video
(numerical)

Learning Category:
Recommender Systems
~Supervised & Unsupervised

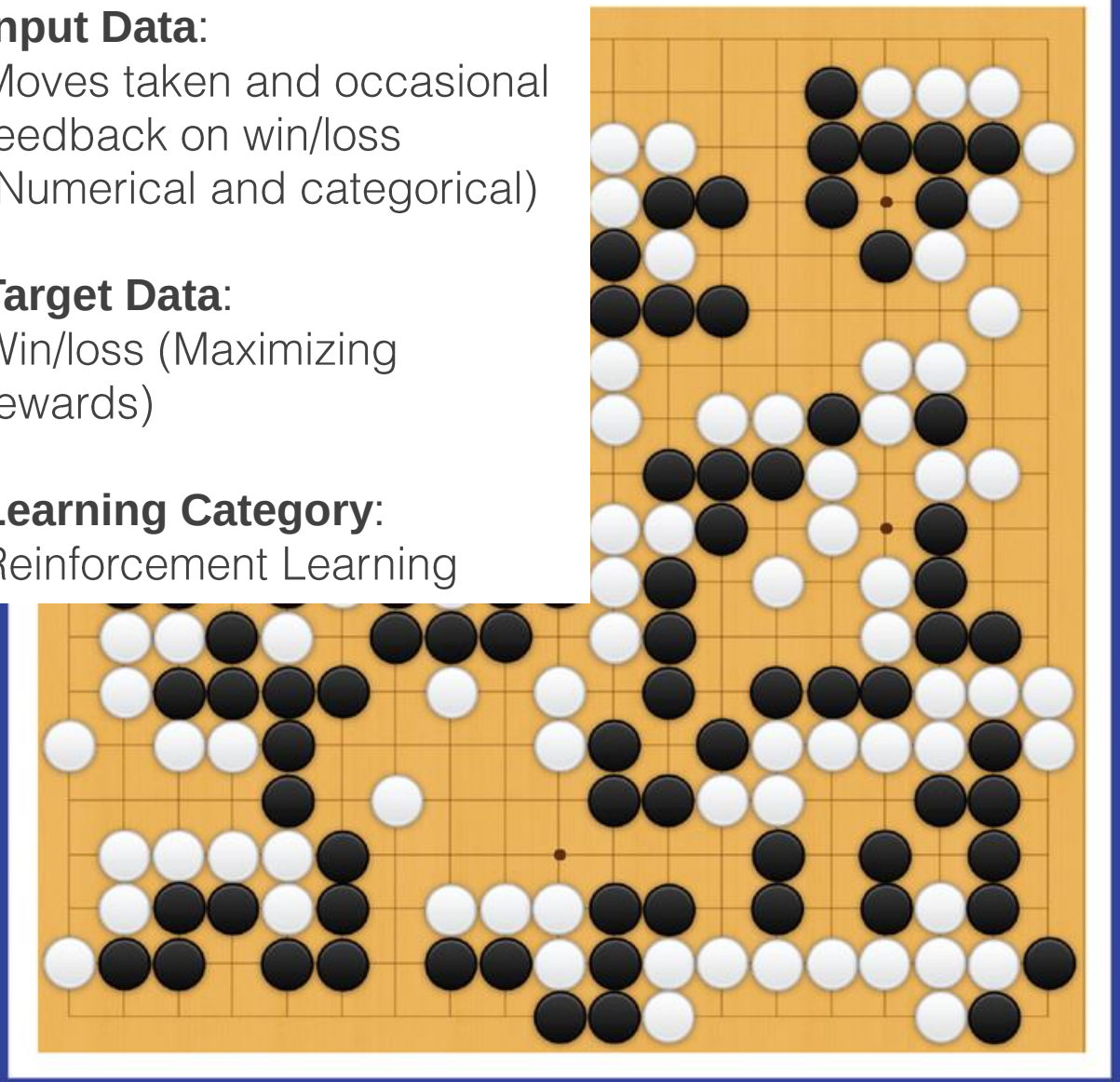


Learning a strategy to master games

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

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Goal	**Predict** …from examples	**Describe** ...structure in data	**Strategize** learn by trial and error
Data	(x, y)	x	delayed feedback
Types	• Classification • Regression	• Density estimation • Clustering • Dimensionality reduction • Anomaly detection	• Model-free learning • Model-based learning

Course logistics

Learning objectives

- Fundamentals of **machine learning**
- **Structure** a machine learning problem
- Automatically make **decisions** from data
- Understand the techniques/algorithms and when to use them
- **Communicate** and effectively **interpret** machine learning decisions

Pedagogy

- Desirable difficulties and practice
- Mental models for independent learning
- Interpretation and communication
- Reading, reflection, and recall

Through this course you will…

- **Apply** machine learning techniques and **communicate** findings
- Compete in a **Kaggle** **machine learning competition**
- Create a **data science portfolio** of the work you complete
- Implement your own **end-to-end machine learning project**

Course website

kylebradbury.github.io/ids705

Graded Components

Class participation	5%
In-class quizzes	10%
Assignments (5 assignments)	40%
Kaggle competition	20%
Final project	25%

Action items

1. Log onto Piazza (make sure you can access it and ask questions)
2. Get the textbooks (available free online)
3. Complete the first reading
4. (if you haven't already) Create an account at Poll Everywhere using this link: https://bit.ly/2RBKLA3
   Be sure to use (a) your real name, AND (b) your Duke netid-based email address (e.g. mjw8@duke.edu)
5. Begin working on Assignment #1

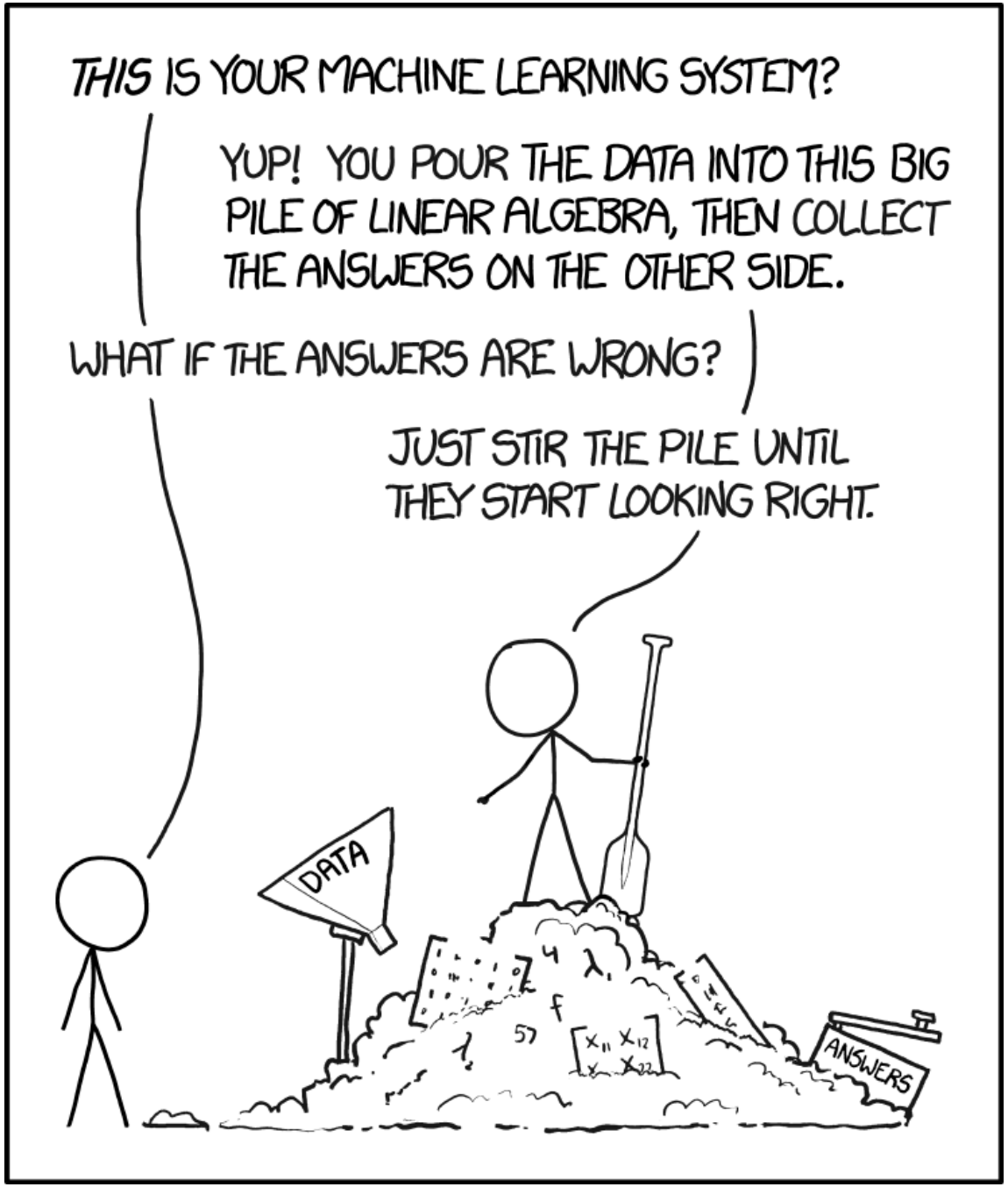


Image: xkcd.com