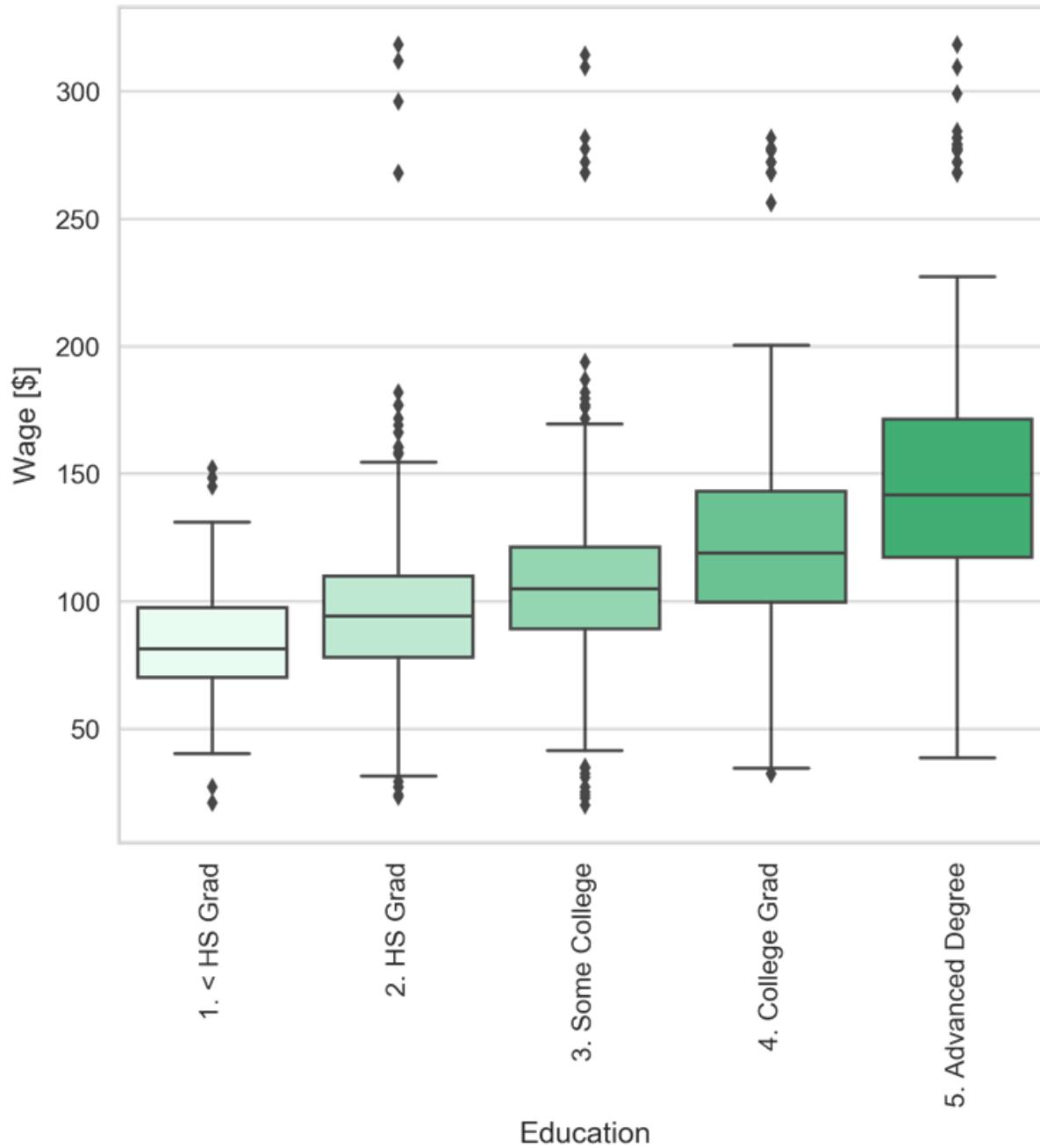


# What is machine learning?

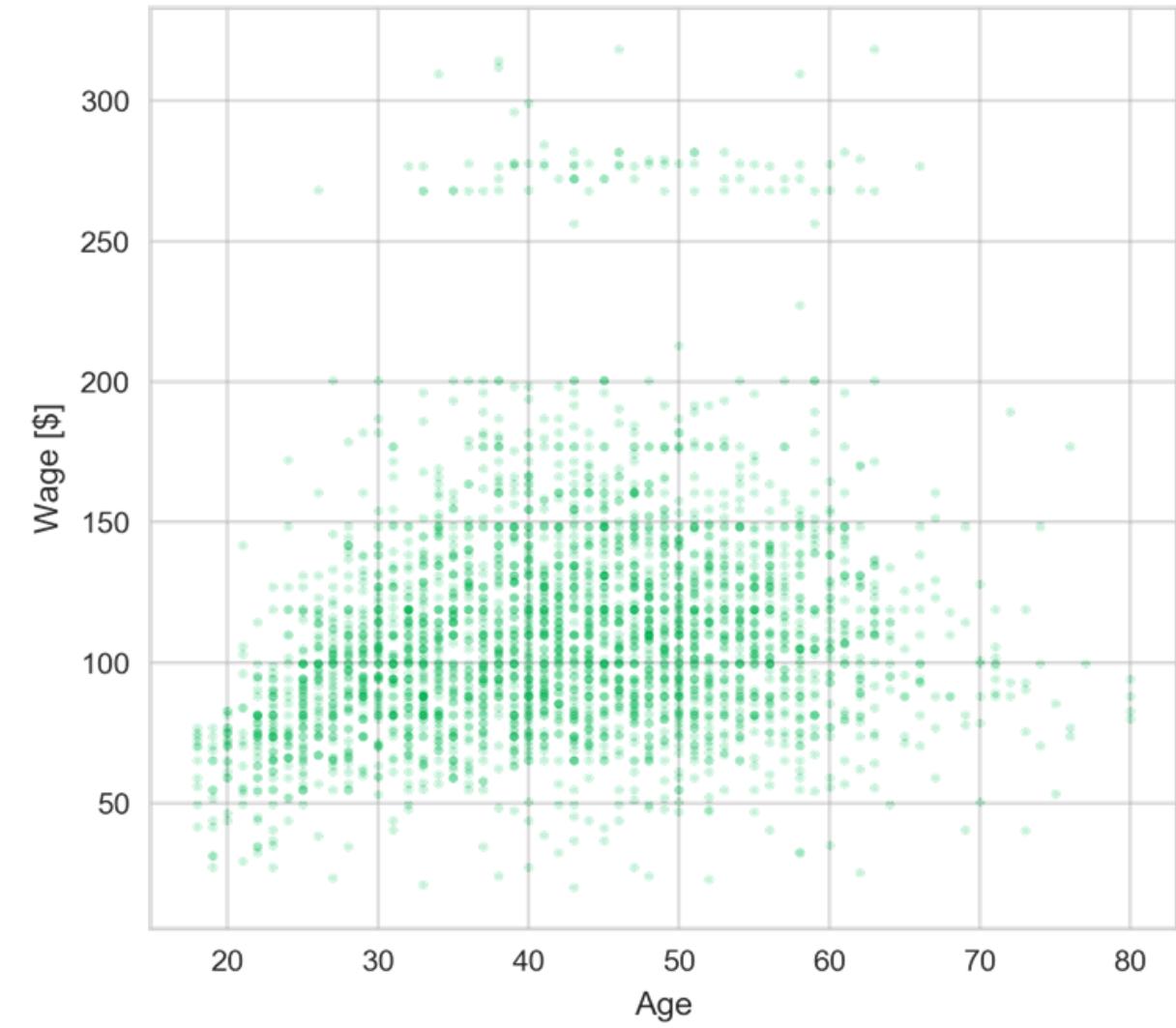
Lecture 01

# 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](mailto:mjw8@duke.edu))
2. Complete the poll at:  
**<https://pollev.com/kylebradbury>**



Data source: James et al., 2013



Wage data from workers in the mid-Atlantic region  
**How do you predict how much someone will make?**

# How can you tell these flowers apart?

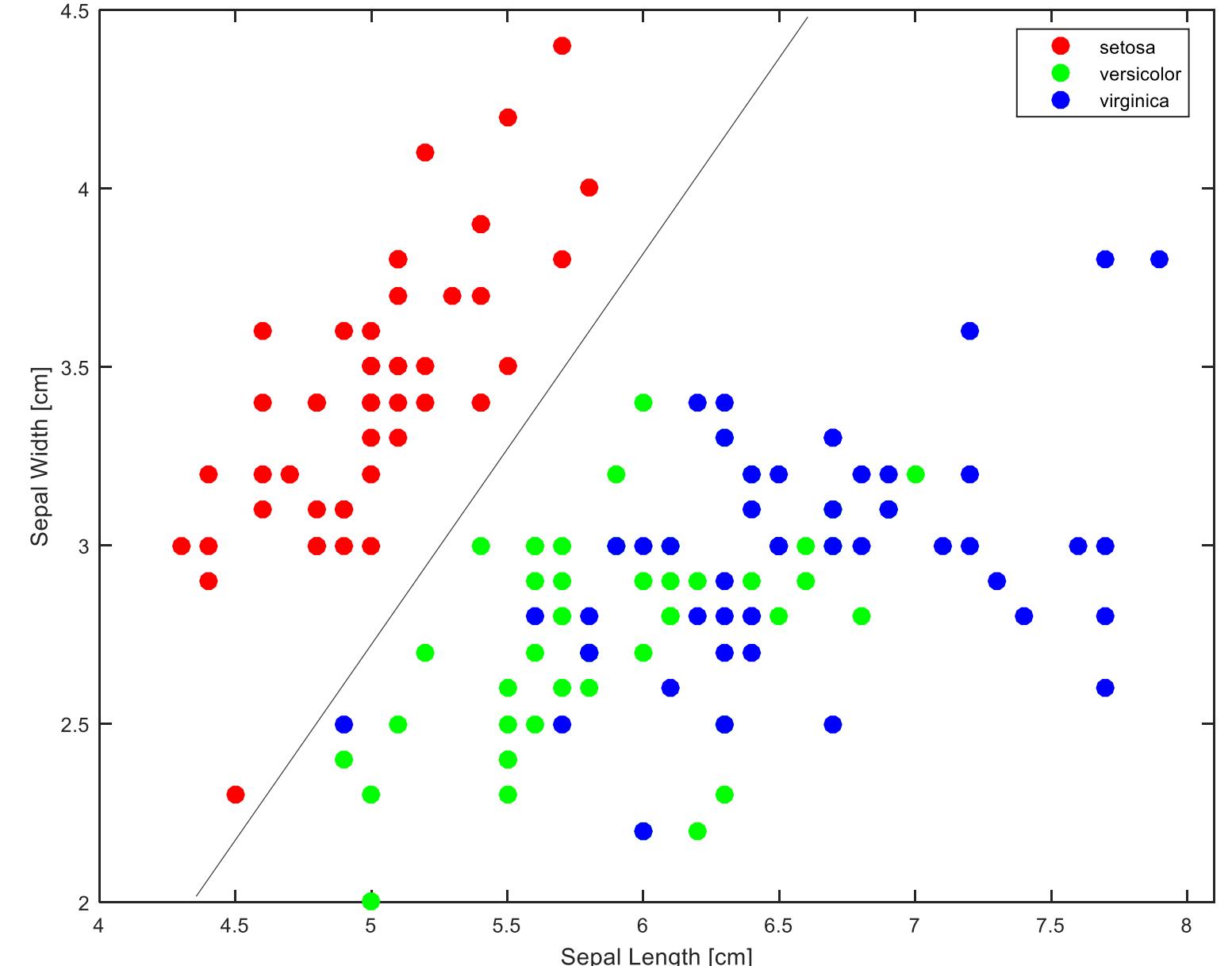


Iris setosa

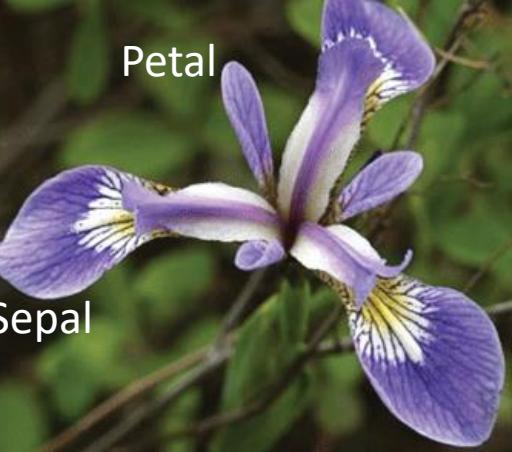


Iris virginica

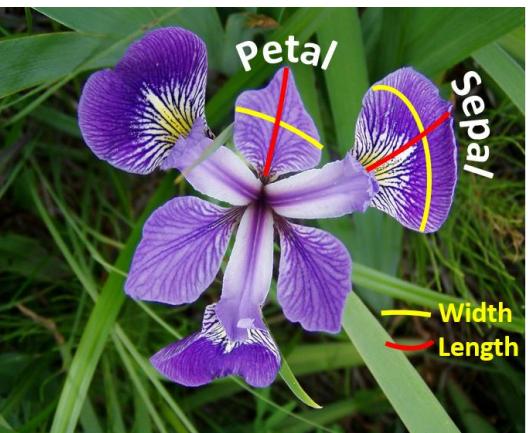
Image Sources: Srishti Sawla (setosa) and Ivo Dinov, University of Michigan SOCR (versicolor and virginica)



setosa



versicolor



virginica



Data Source: Fisher Iris Data

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

# Challenges

What  
is  
this?



Image by artist Hikaru Cho

Kyle Bradbury

What is machine learning?

Duke University | Lecture 01

# We generalize from past experiences

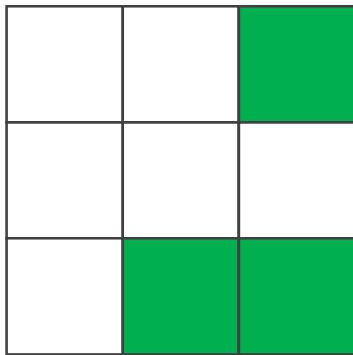
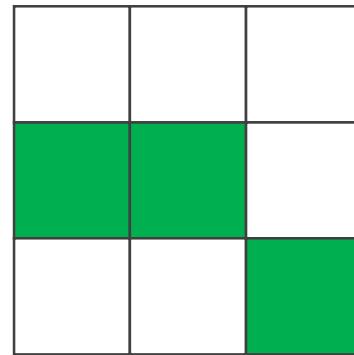
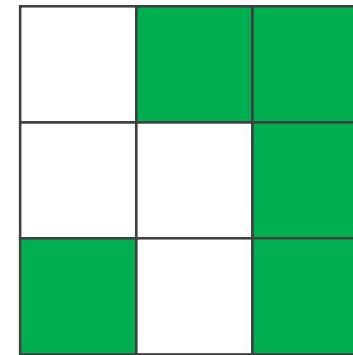


Image: "It's not what it seems" by artist Hikaru Cho

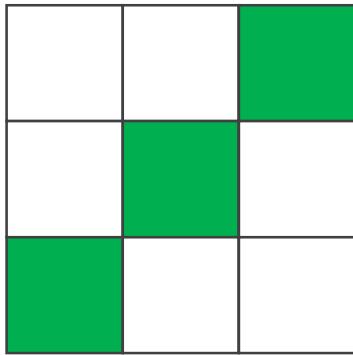
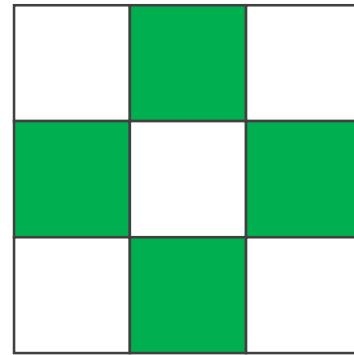
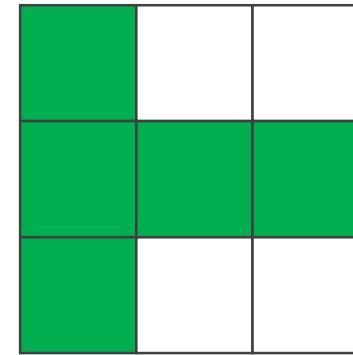
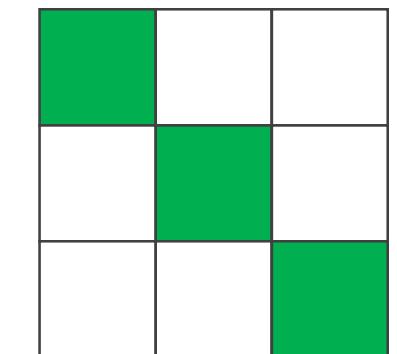
our data must be  
**representative**

# Predict which class $x_{\text{new}}$ belongs to...

$$f(x) = 1$$

 $x_0$  $x_2$  $x_4$ 

$$f(x) = 0$$

 $x_1$  $x_3$  $x_5$  $x_{\text{new}}$ 

$$f(x_{\text{new}}) = ?$$

Example credit: Yaser Abu-Mostafa, 2012

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

There are often **many** models that  
will fit your training data

So how do we choose which to use?

the best models  
**generalize well**

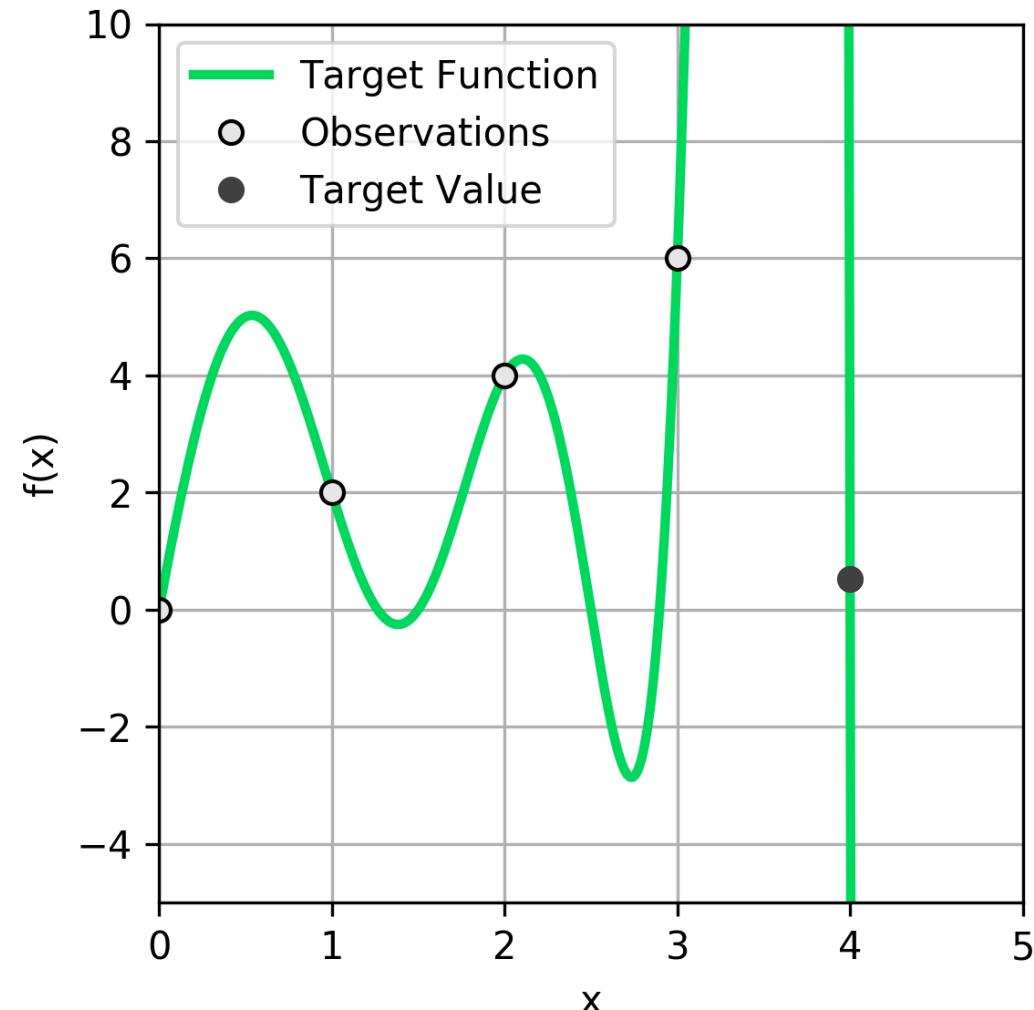
# Predict the next value in the sequence...

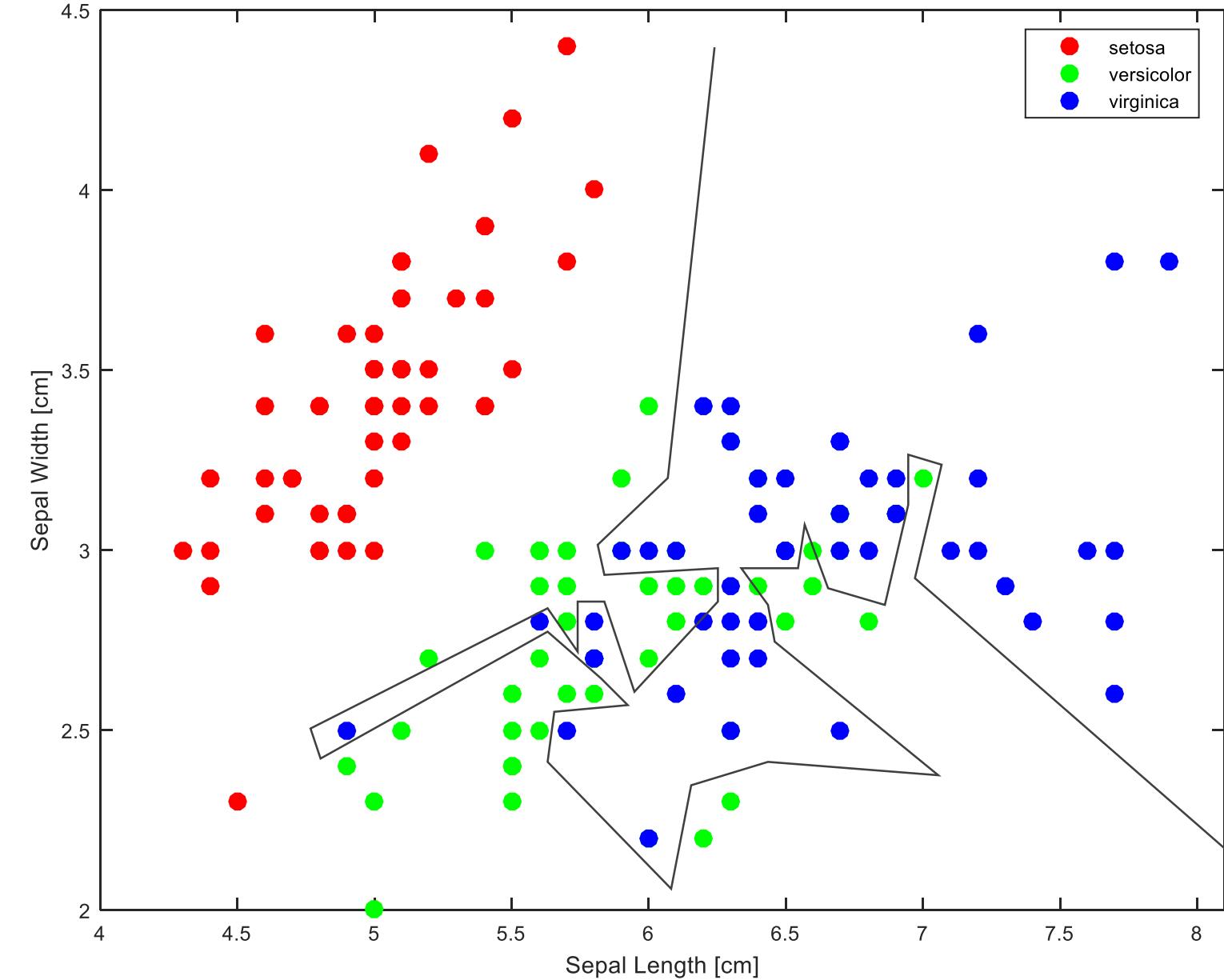
$x$	0	1	2	3	4
$f(x)$	0	2	4	6	?

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

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

Kyle Bradbury

What is machine learning?

Duke University | Lecture 01

14

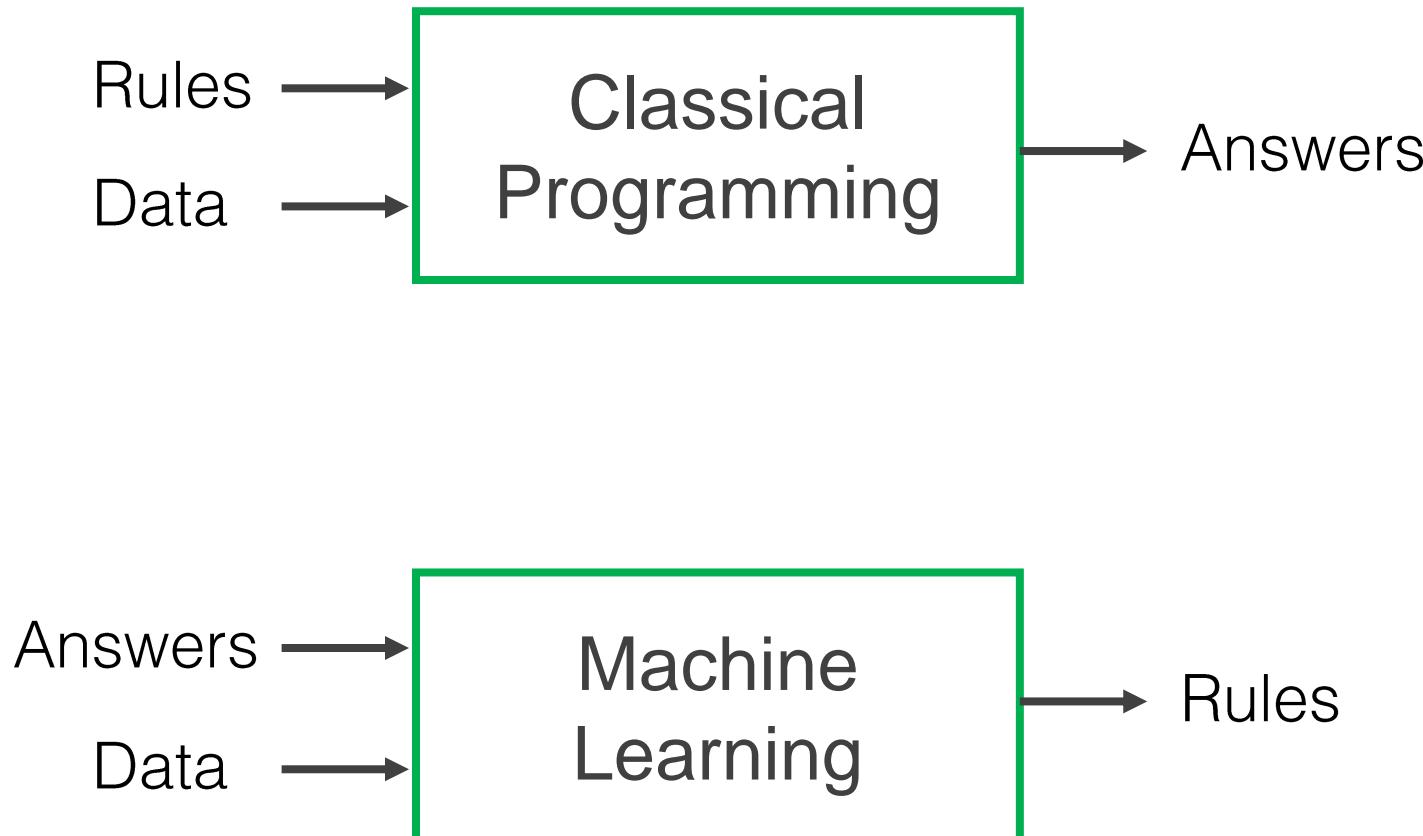
Complex models overfit to the data

**Overfit** works against  
generalization



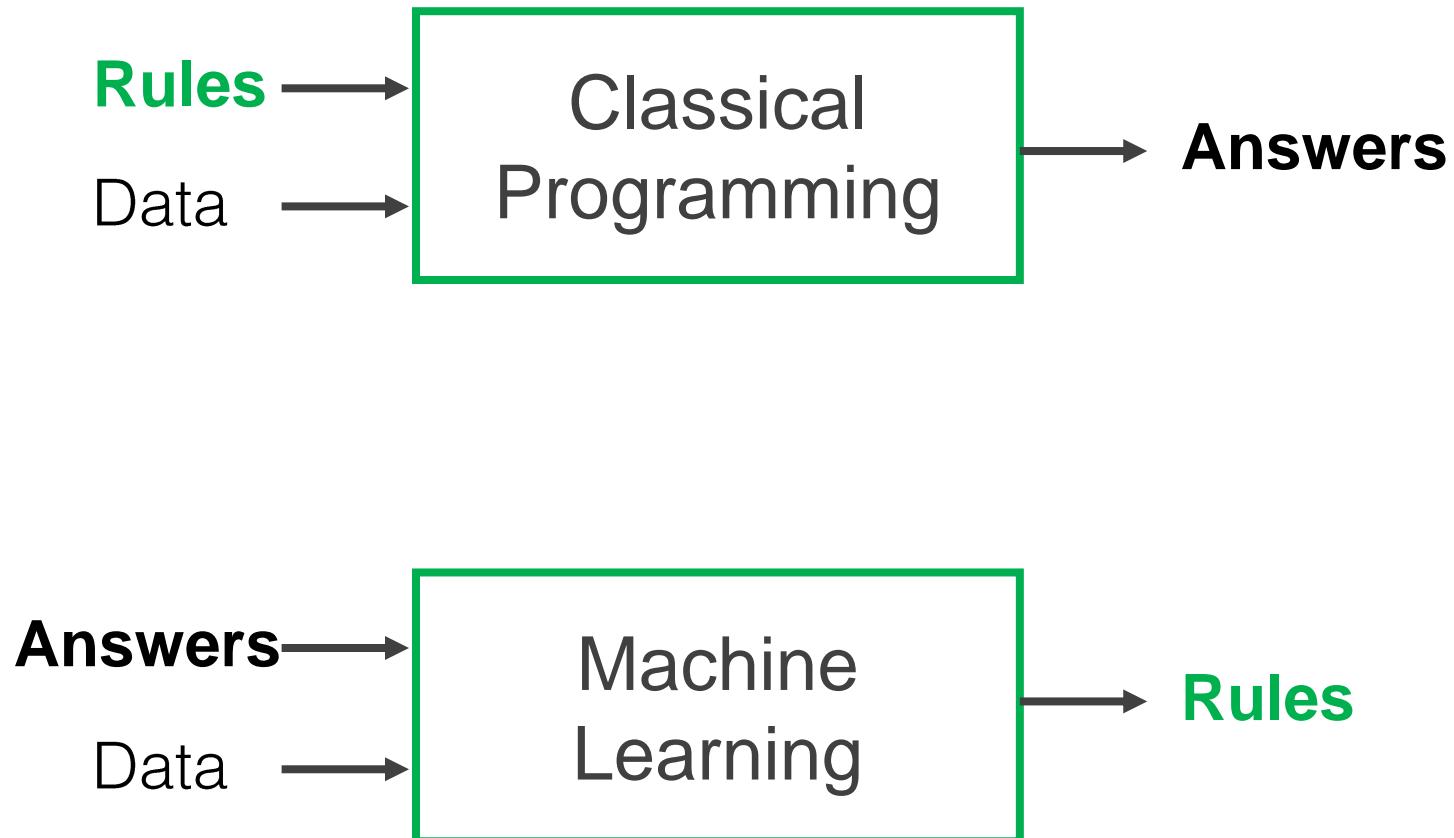
**Learning simpler representations  
of data enables learning**

# Machine learning suggests an alternative programming paradigm



François Chollet, *Deep Learning with Python*, 2017

# Machine learning suggests an alternative programming paradigm



François Chollet, *Deep Learning with Python*, 2017

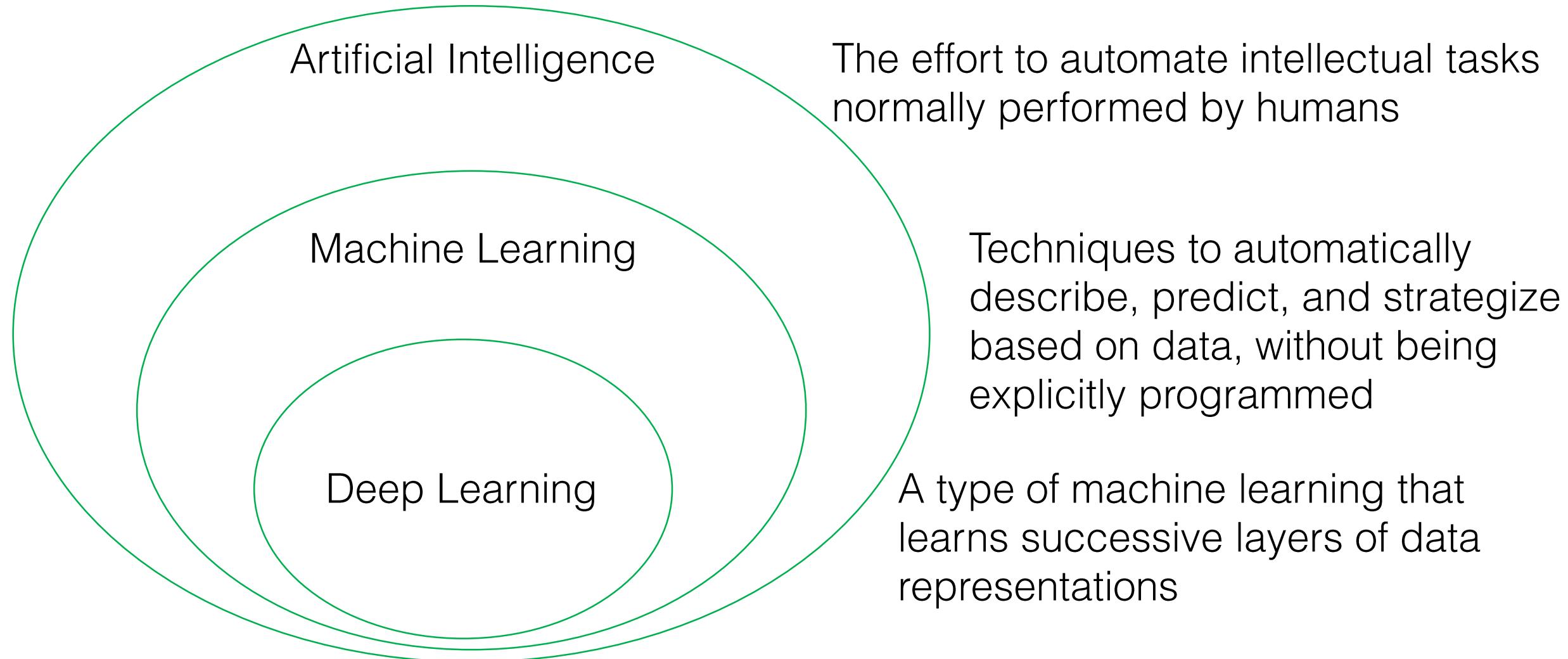
# What is machine learning?

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

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

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

# What is machine learning?



François Chollet, *Deep Learning with Python*, 2017

# Types of machine learning tools

## Types of learning

**Unsupervised** learning

**Supervised** learning

**Reinforcement** learning

## Common use case

**Describe**

**Predict**

**Strategize**

# Types of machine learning

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Goal	<b>Predict</b> ...from examples	<b>Describe</b> ...structure in data	<b>Strategize</b> learn by trial and error
Data	$(x, y)$	$x$	delayed feedback
Types	<ul style="list-style-type: none"><li>Classification</li><li>Regression</li></ul>	<ul style="list-style-type: none"><li>Density estimation</li><li>Clustering</li><li>Dimensionality reduction</li><li>Anomaly detection</li></ul>	<ul style="list-style-type: none"><li>Model-free learning</li><li>Model-based learning</li></ul>

# Sale Price Prediction

## Input Data:

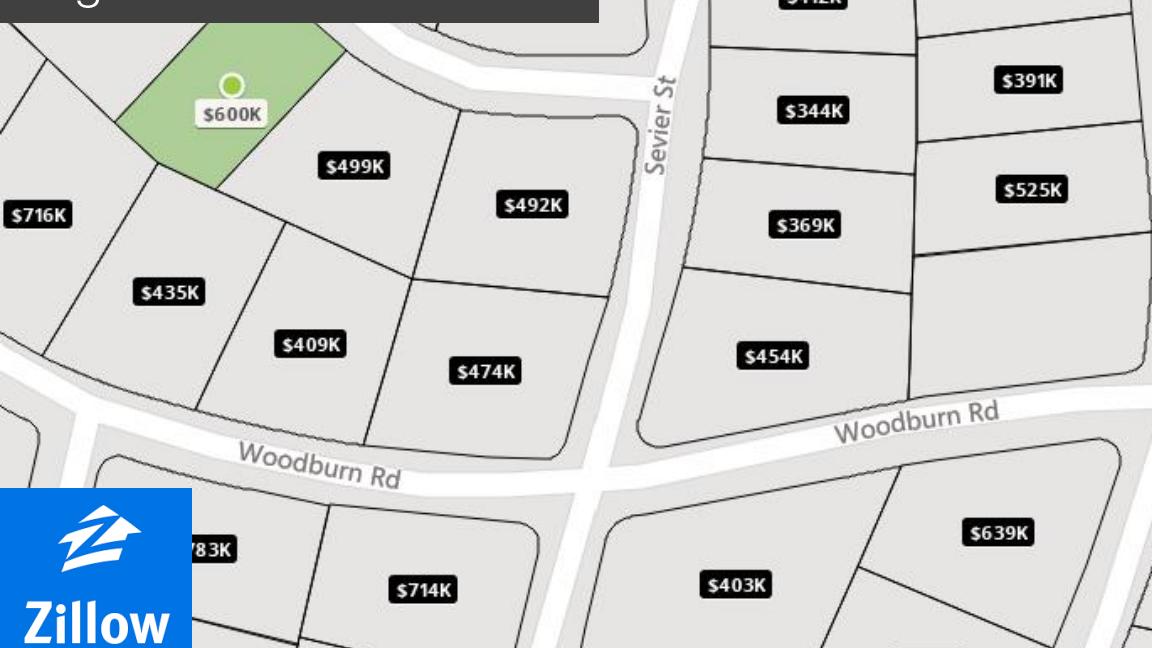
Home characteristics  
(Numerical & Categorical)

## Target Data:

Price estimate (numerical)

## Learning Category:

Supervised Learning  
Regression



## 27708 Real Estate

1 home for sal

Homes for You

Newest

Cheapest

More

**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

**Zestimate®: \$619,585**

# Spam Filters

**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

**Input Data:**  
Email text (text)

**Target Data :**  
Spam/not spam  
(category)

**Learning Category:**  
Supervised Learning  
Classification (binary)

Spam example source: itservices.uchicago.edu

# 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/](http://www.whereswaldo.com/)

# Object Recognition: Energy Systems



- **Building**  
behind-the-meter  
energy consumption

**Input Data:**

Imagery

**Target Data:**

Locations in an image

**Learning Category:**

Supervised Learning  
Classification (binary)



- **Car**  
transportation  
energy consumption

- **Solar Array**  
distributed energy  
resources

- **Light Pole**  
access to electricity

# Credit Fraud

## **Input Data:**

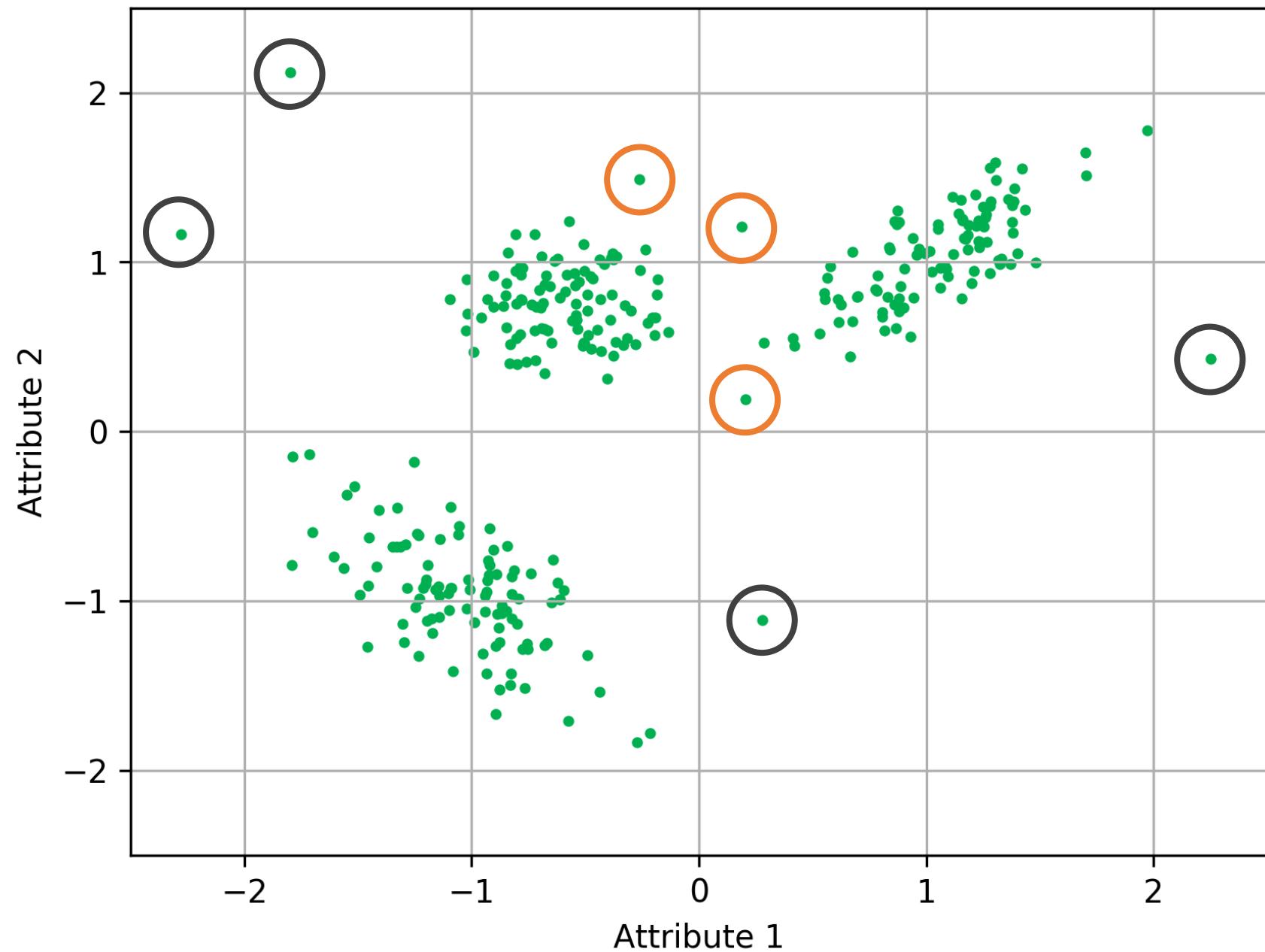
Account transactions, dates,  
locations, demographic  
information  
(Numerical and categorical)

## **Target Data:**

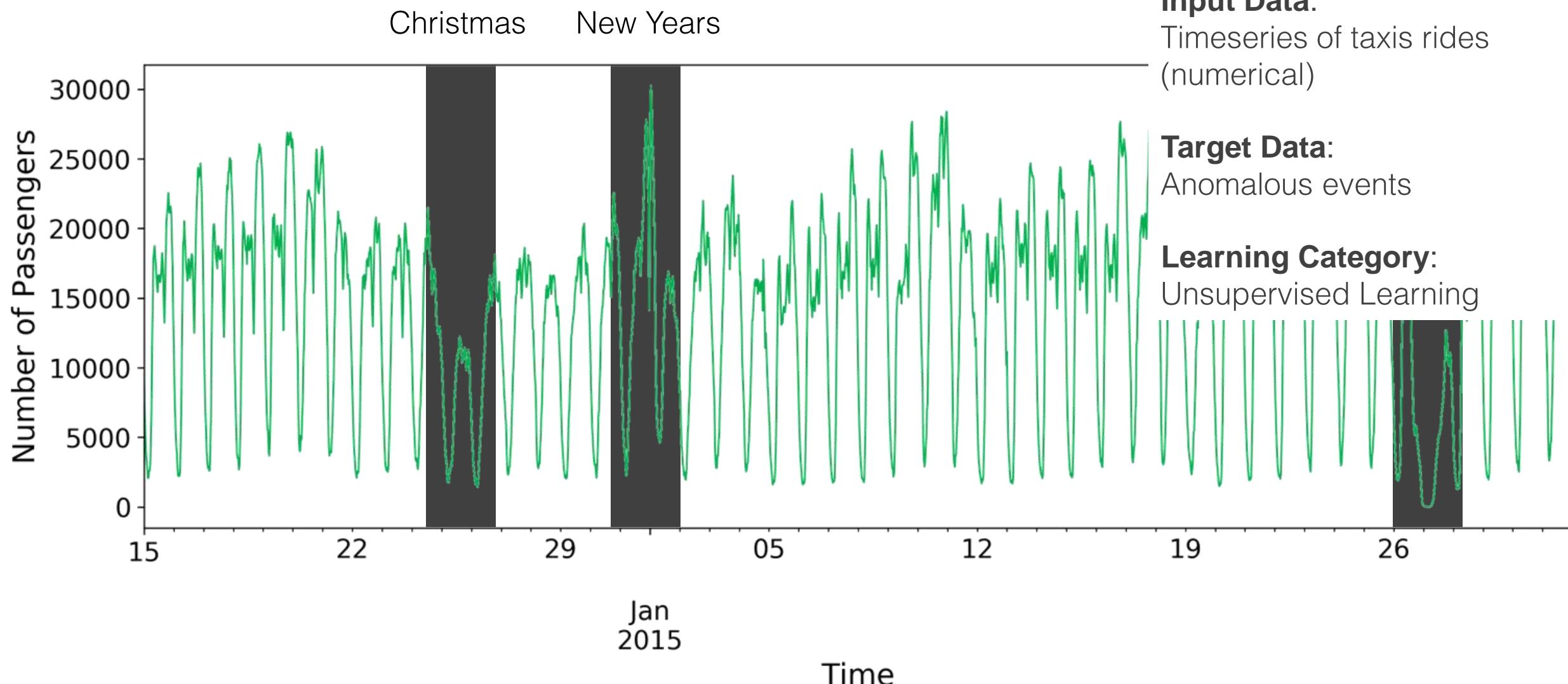
Anomalous transactions

## **Learning Category:**

Unsupervised Learning  
Clustering, Density  
Estimation



# Anomalous Event Detection: NYC Taxis



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

# Video Recommendations



## Sherlock

97% Match 2017 TV-14 4 Series

**97% Match**

Wedding reception, Sherlock...  
...and his biggest challenge of all: delivering a best man's speech



Season 3's episode "The Abominable Bride," which originally aired as a TV movie, won two Emmys.

+ MY LIST



**NETFLIX**

OVERVIEW

EPISODES

MORE LIKE THIS

DETAILS

What is machine learning?

Duke University | Lecture 01

### Input Data:

User video ratings  
(numerical and categorical)



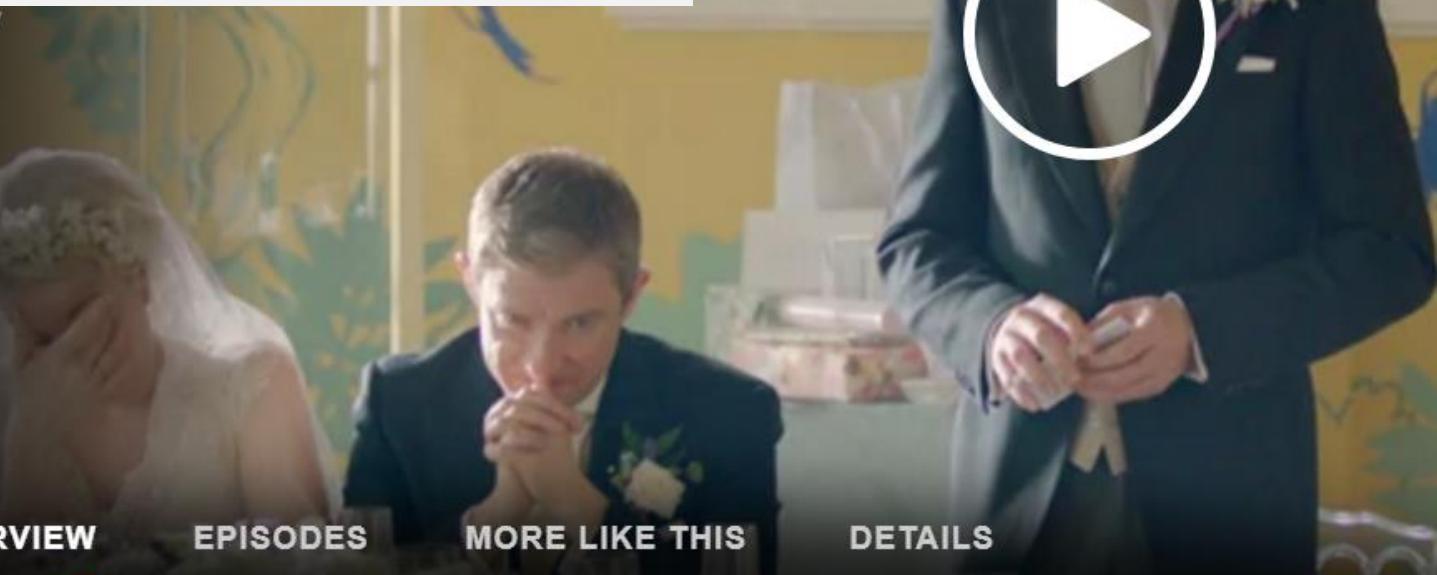
### Target Data:

User rating of video  
(numerical)



### Learning Category:

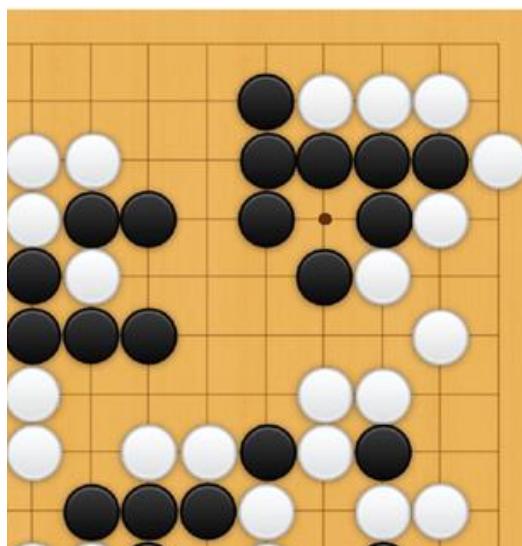
Recommender Systems  
~Supervised & Unsupervised



# Learning a strategy to master games

## Input Data:

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

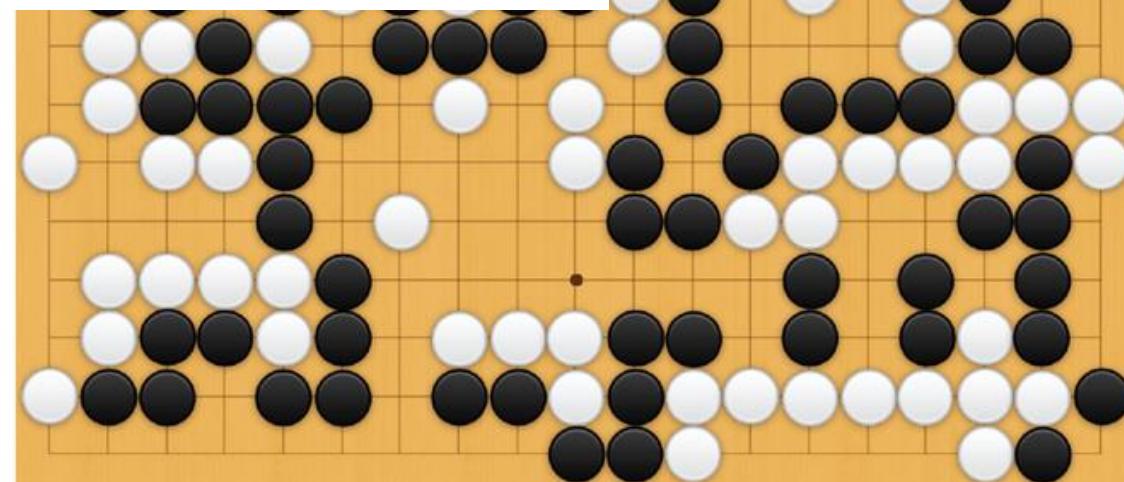


## Target Data:

Win/loss (Maximizing rewards)

## Learning Category:

Reinforcement Learning



## THE ULTIMATE GO CHALLENGE

GAME 3 OF 3

27 MAY 2017



vs



AlphaGo

*Winner of Match 3*

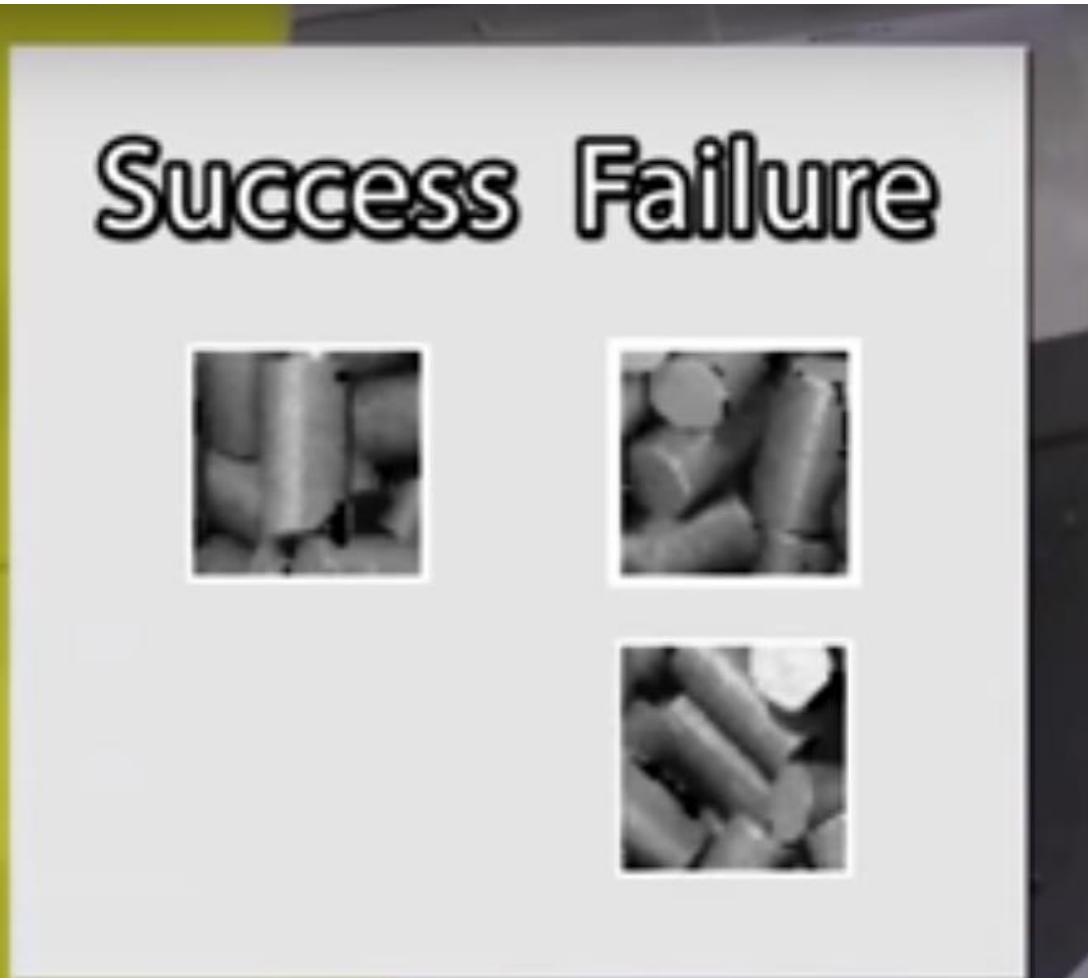
Ke Jie

**RESULT B + Res**



Google DeepMind

# 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	<b>Predict</b> ...from examples	<b>Describe</b> ...structure in data	<b>Strategize</b> learn by trial and error
Data	$(x, y)$	$x$	delayed feedback
Types	<ul style="list-style-type: none"><li>Classification</li><li>Regression</li></ul>	<ul style="list-style-type: none"><li>Density estimation</li><li>Clustering</li><li>Dimensionality reduction</li><li>Anomaly detection</li></ul>	<ul style="list-style-type: none"><li>Model-free learning</li><li>Model-based learning</li></ul>

# 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

- Interpretation and communication
- Reading, reflection, and recall
- Desirable difficulties and practice
- Learn how to learn independently

# 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**
- Create a **video** communicating your final project

# Course website

[kylebradbury.github.io/ids705](https://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. Get the textbooks (available free online)
2. Log onto Piazza (make sure you can access it and ask questions)
3. Complete the first reading
4. 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. mjh8@duke.edu)**
5. Begin working on Assignment #1

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG  
PILE OF LINEAR ALGEBRA, THEN COLLECT  
THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL  
THEY START LOOKING RIGHT.

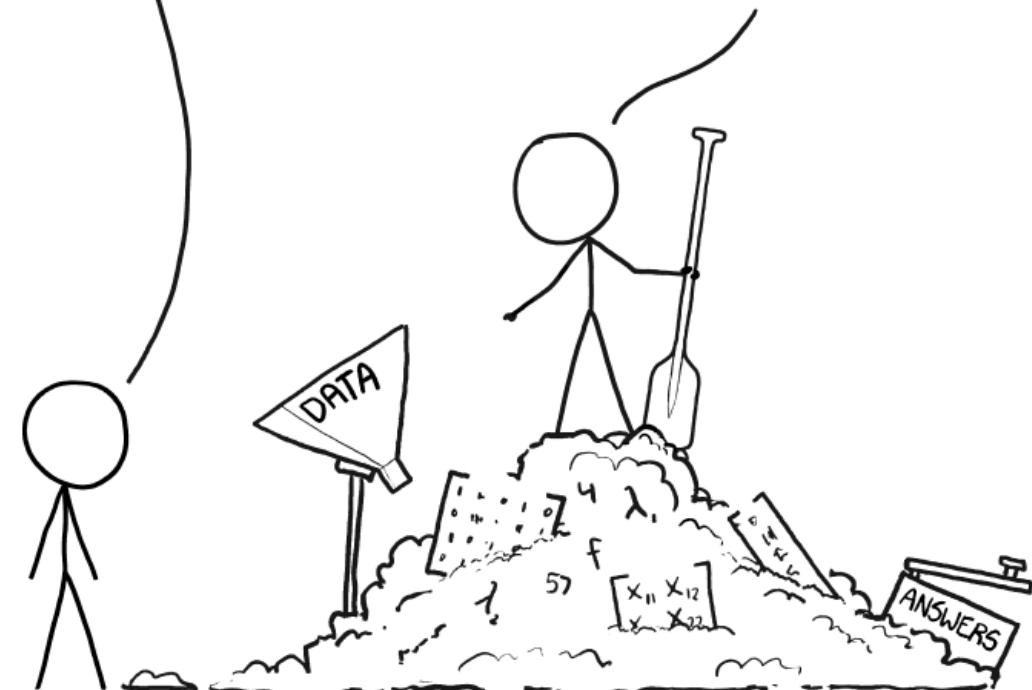


Image: xkcd.com