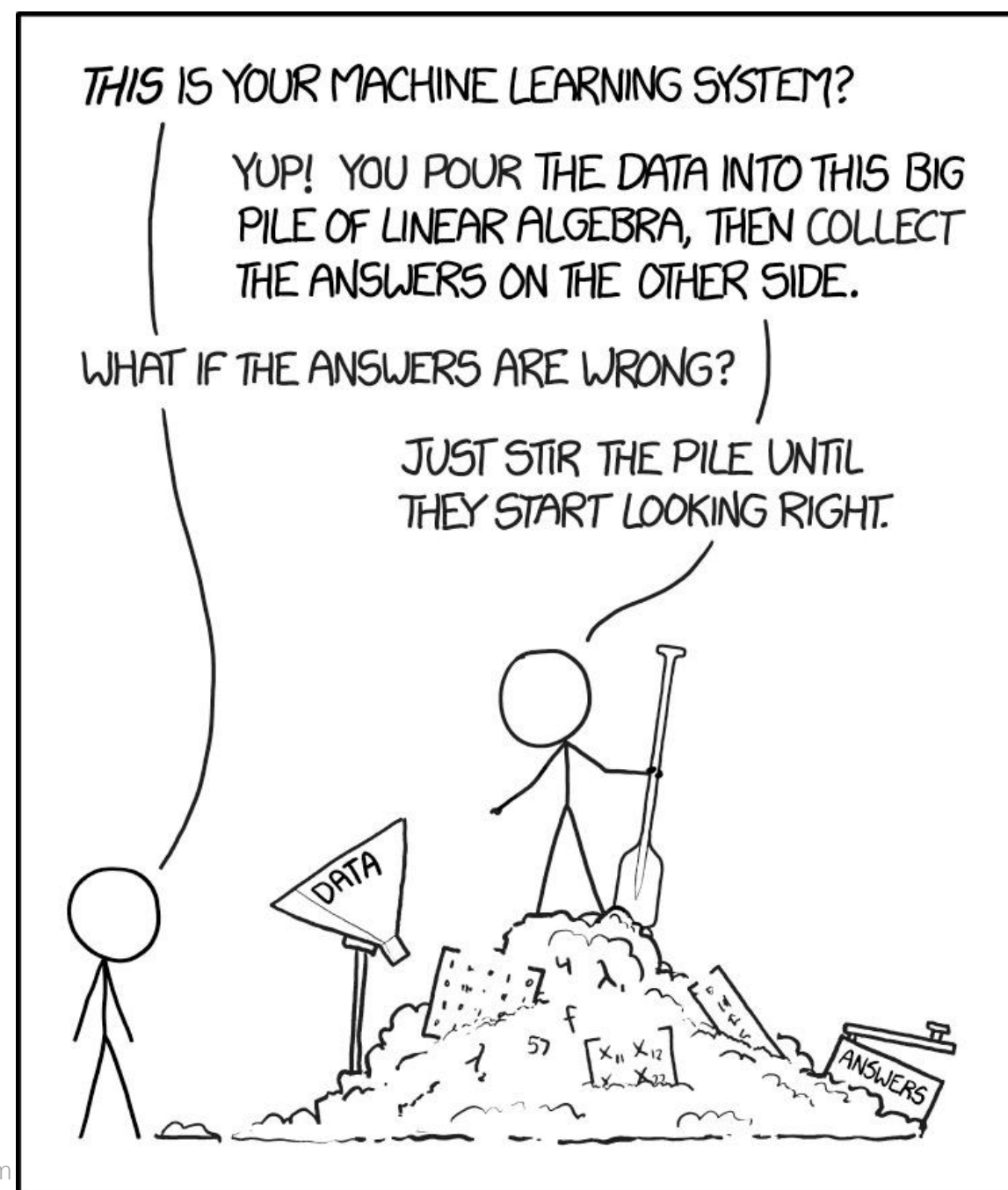


Welcome to Principles of Machine Learning IDS 705

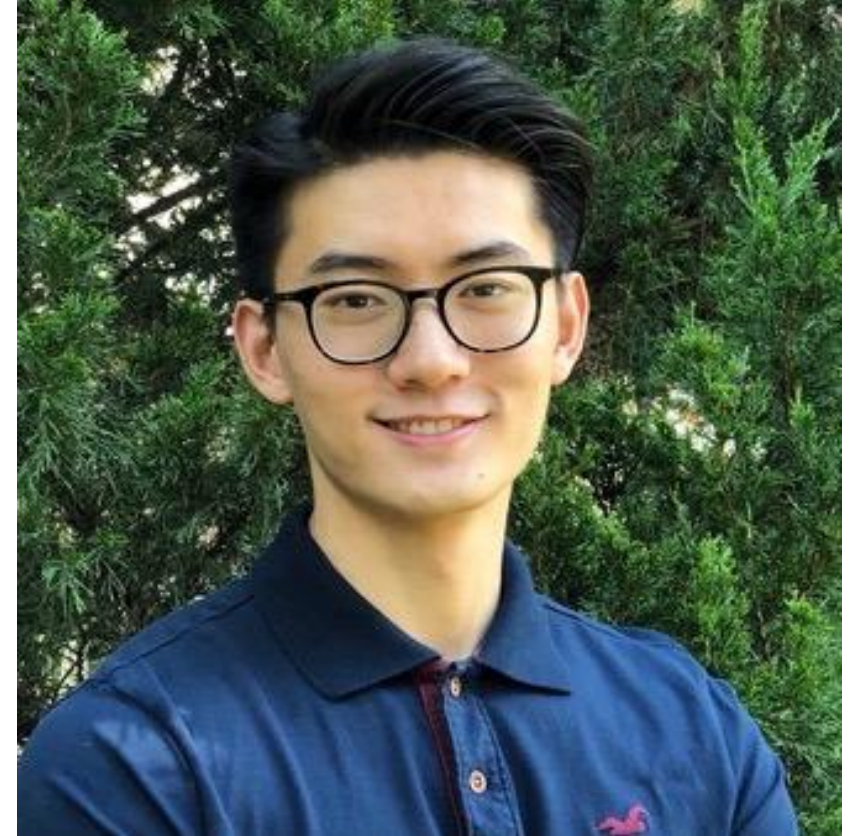
Image: xkcd.com



TA's



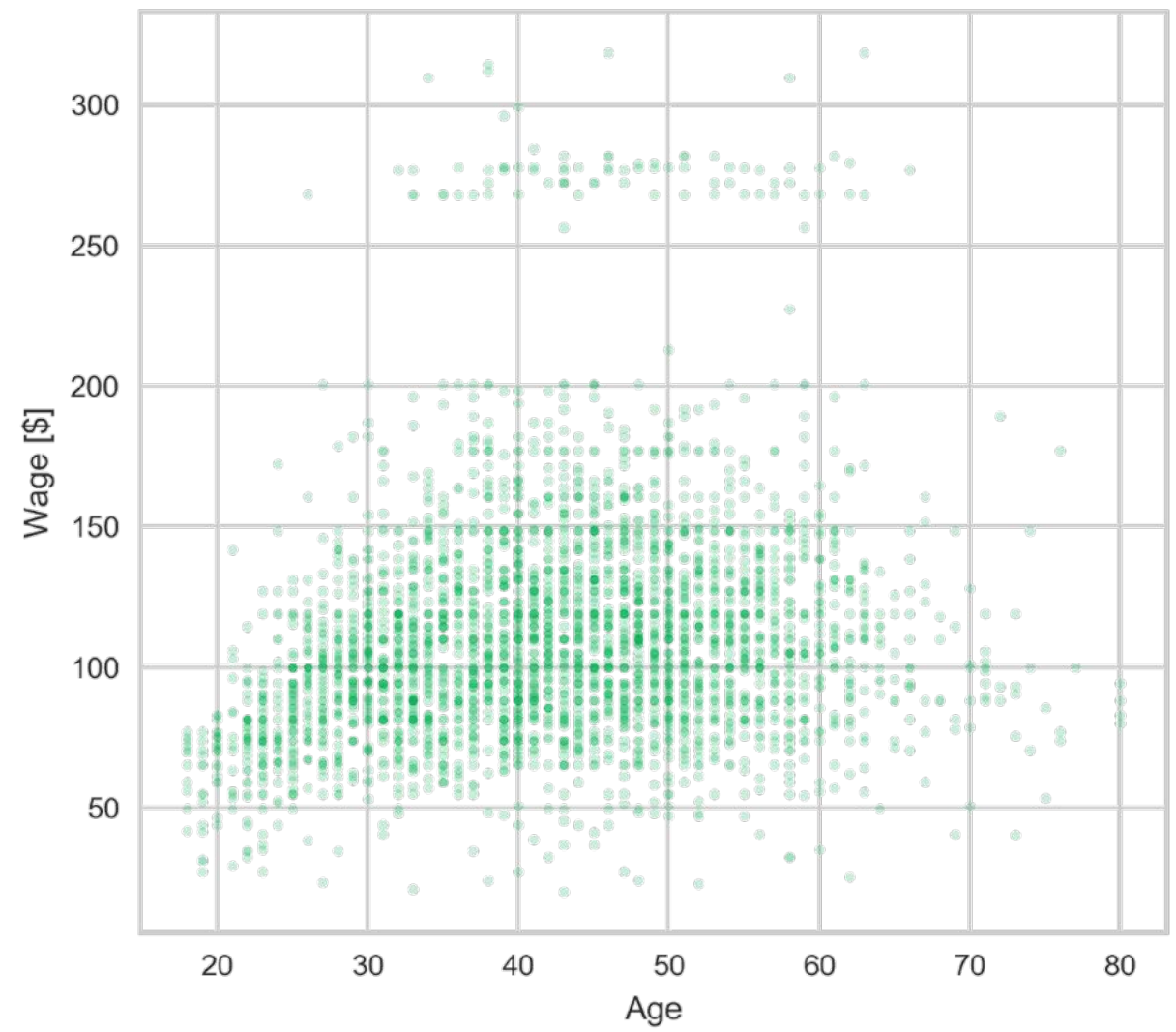
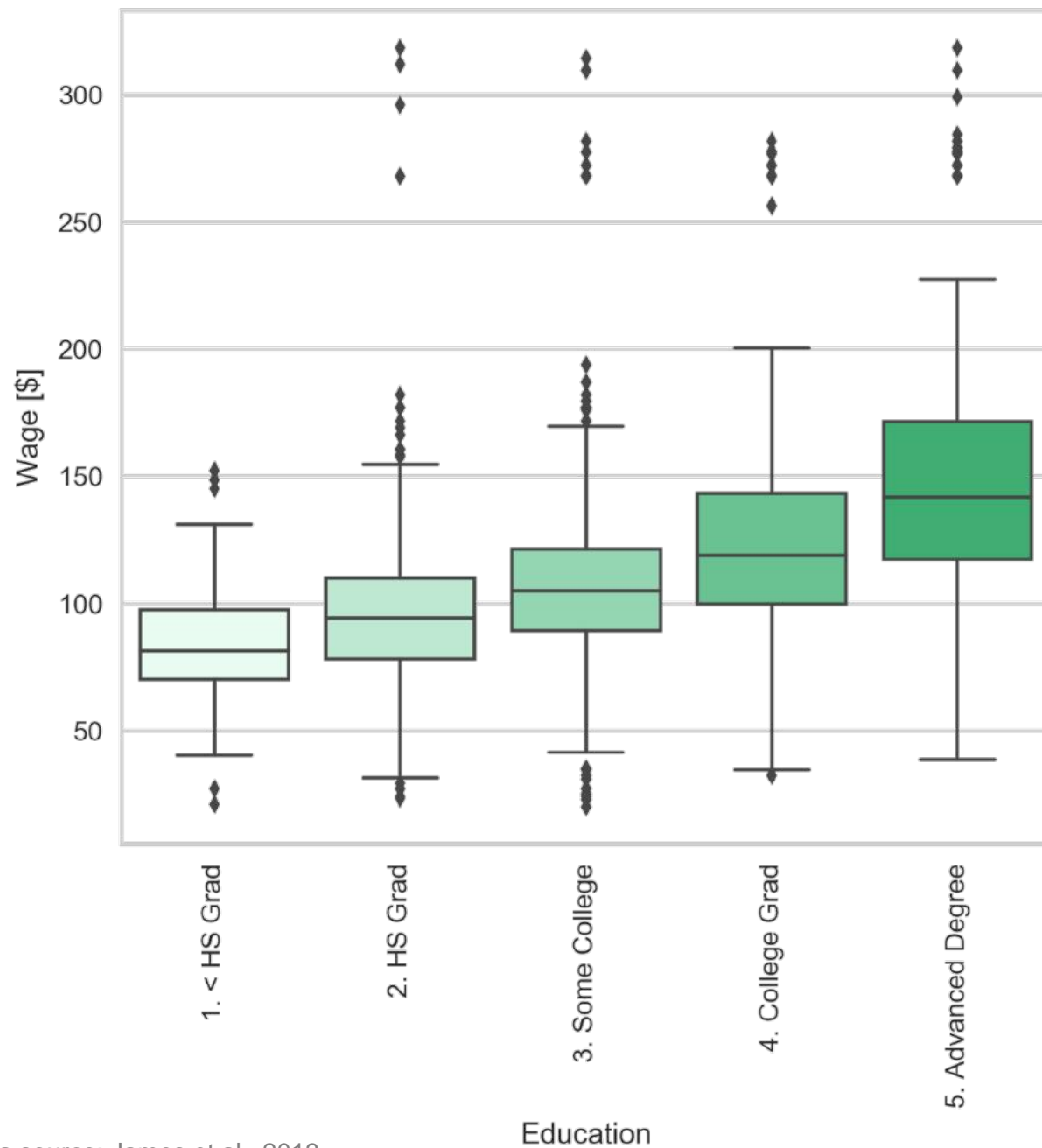
Dean Huang



Michael Tang

What is machine learning?

Lecture 01



Wage data from workers in the mid-Atlantic region in one field

How much will someone with some college education make? What if they're 20 years old?

We use heuristic rules from past experiences and data to make decisions and understand the world daily

Machine learning seeks to enhance that process

**We don't just stir
the pile...**



How can you tell these flowers apart?

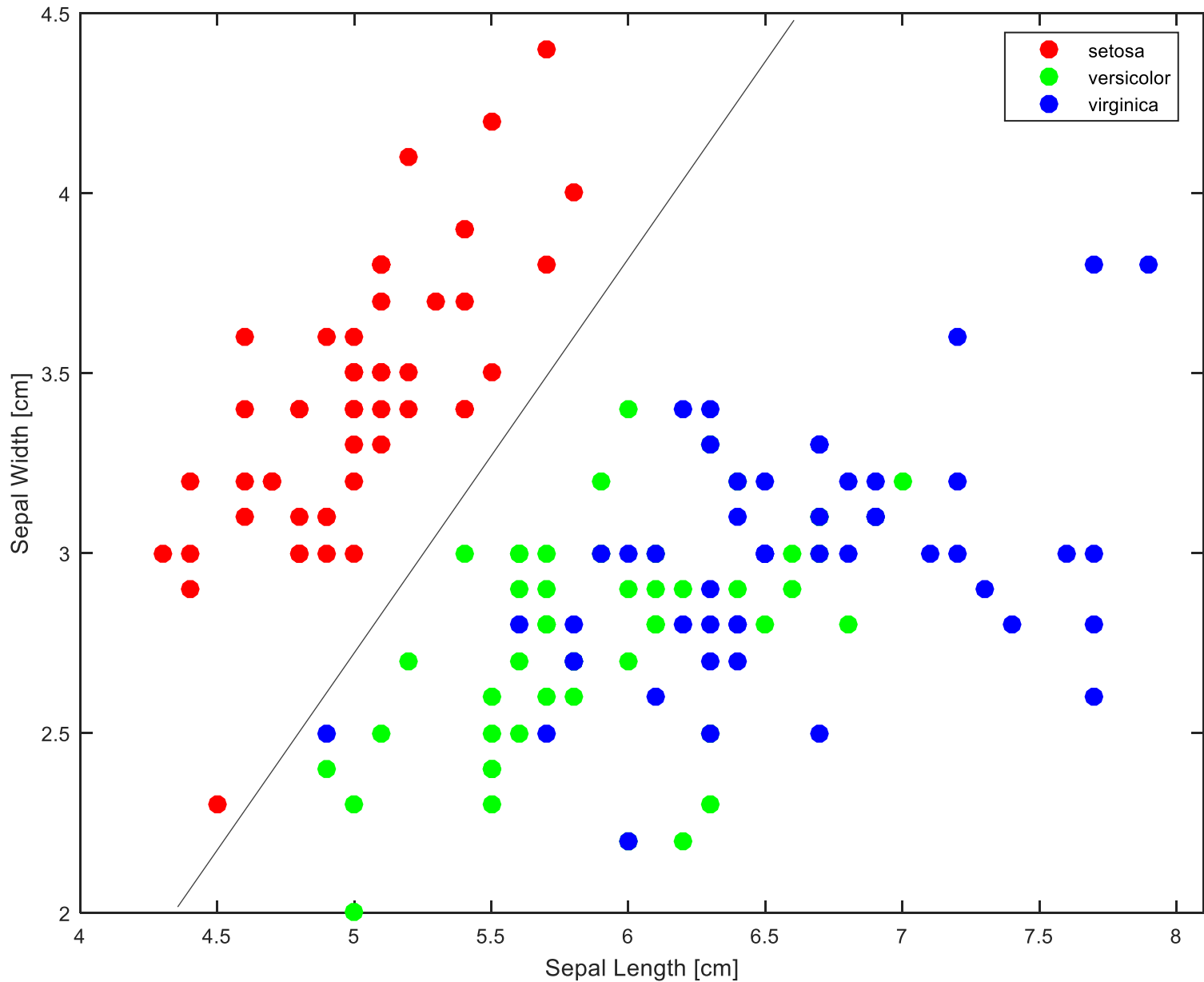


Iris setosa



Iris virginica

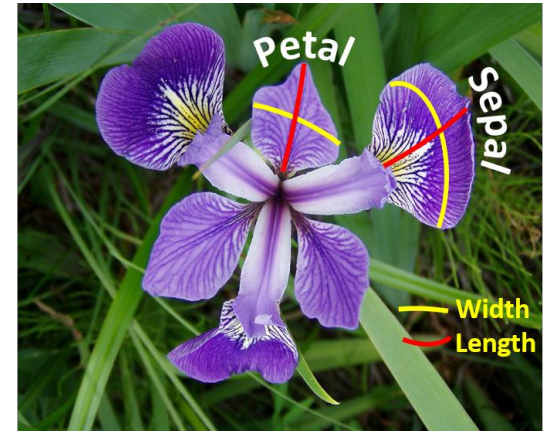
Image Sources: Srishti Sawla (setosa) and Ivo Dinov, University of Michigan SOCR (versicolor and virginica)



setosa



versicolor



virginica



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

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.

Challenges



What
is
this?

We **generalize** from past experiences

...and training data



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

Challenge #1

our data must be
representative

How about this one?



Image by artist Hikaru Cho



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



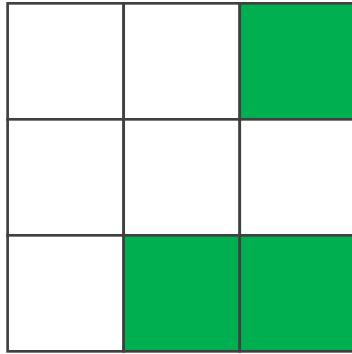
Developing **representations of data** to enable learning

To the poll!

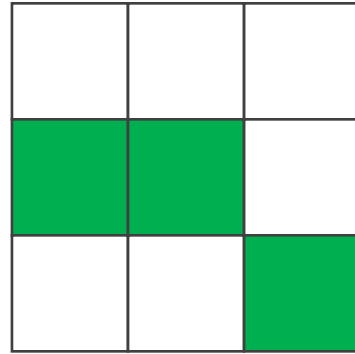
bit.ly/3Hwmba9

Predict which class x_{new} belongs to...

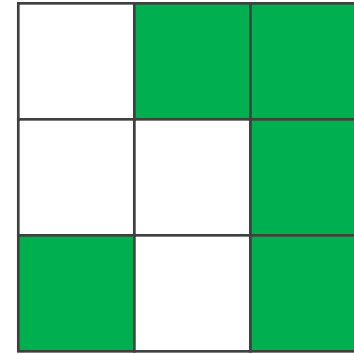
$$f(x) = 1$$



x_0

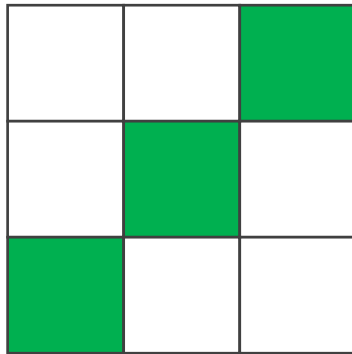


x_2

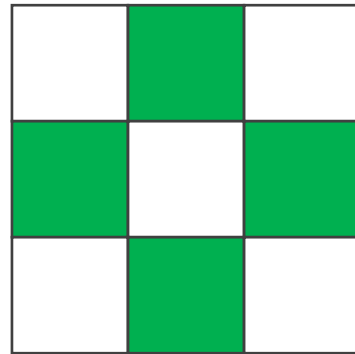


x_4

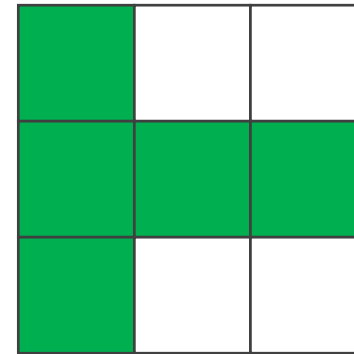
$$f(x) = 0$$



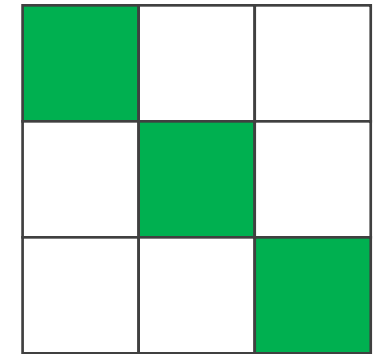
x_1



x_3



x_5



x_{new}

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

Challenge #3

Machine learning is an
ill-posed problem

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

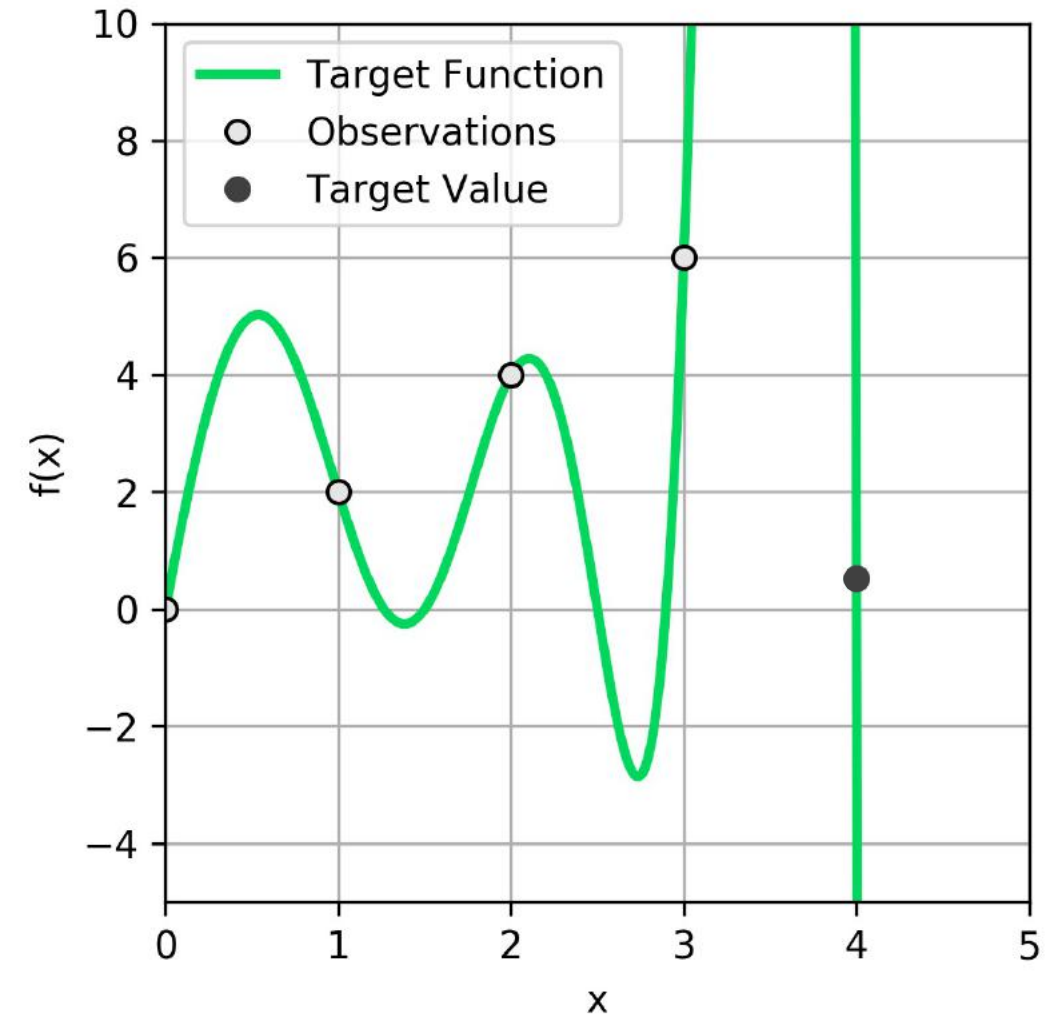
So how do we choose which model to use?

the best models
generalize well

Predict the next value in the sequence...

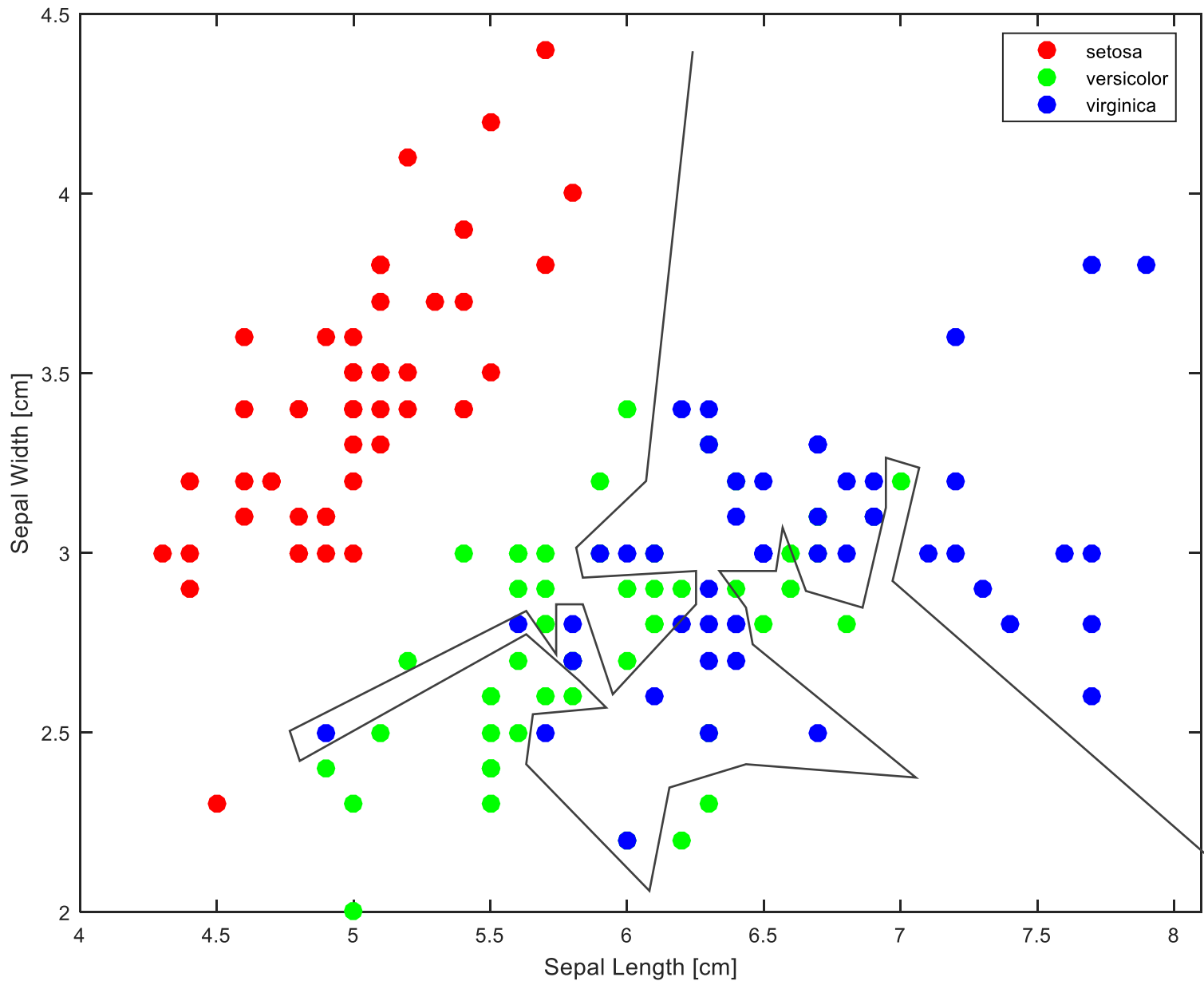
x	0	1	2	3	4
$f(x)$	0	2	4	6	?

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



A model:

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



setosa



versicolor



virginica



Complex models overfit to the data

overfit works against
generalization

Summary of the key challenges of ML

1. **Data** must be representative

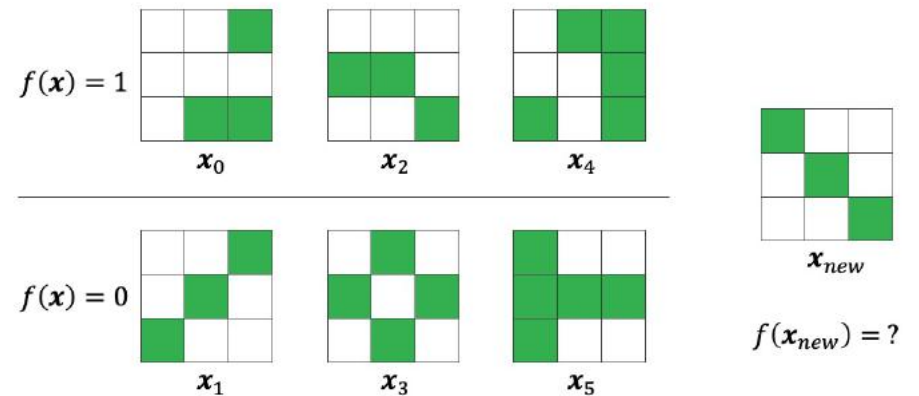


Image by artist Hikaru Cho

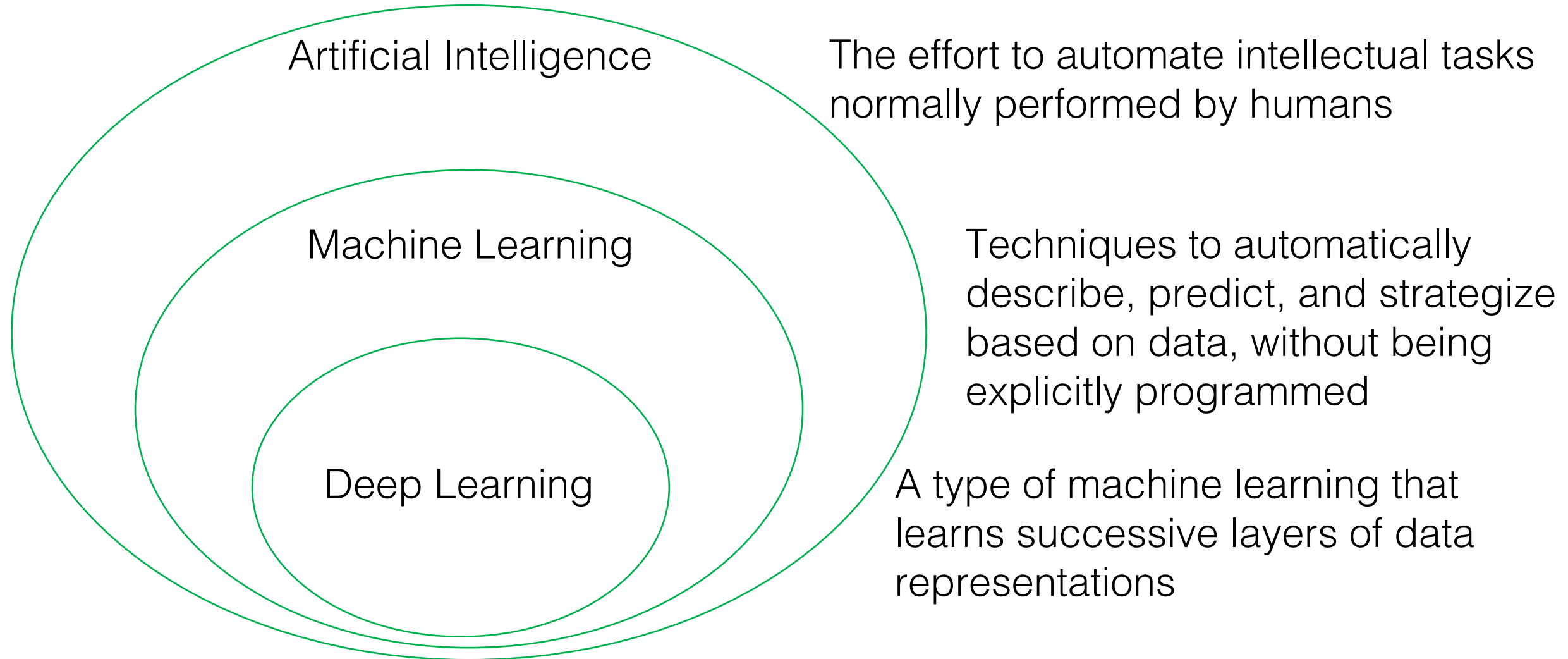
2. Create efficient **representations** of data enable learning
(e.g. feature engineering or use algorithms that automatically extract features)



3. Models must **generalize** well
(pick the simplest model that achieves your goal)



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



27708 Real Estate

1 home for sale

[Homes for You](#) [Newest](#) [Cheapest](#) [More](#)

HOUSE FOR SALE
\$599,900 5 bds · 4 ba · 3,264 sqft
1640 Marion Ave, Durham, NC

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

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



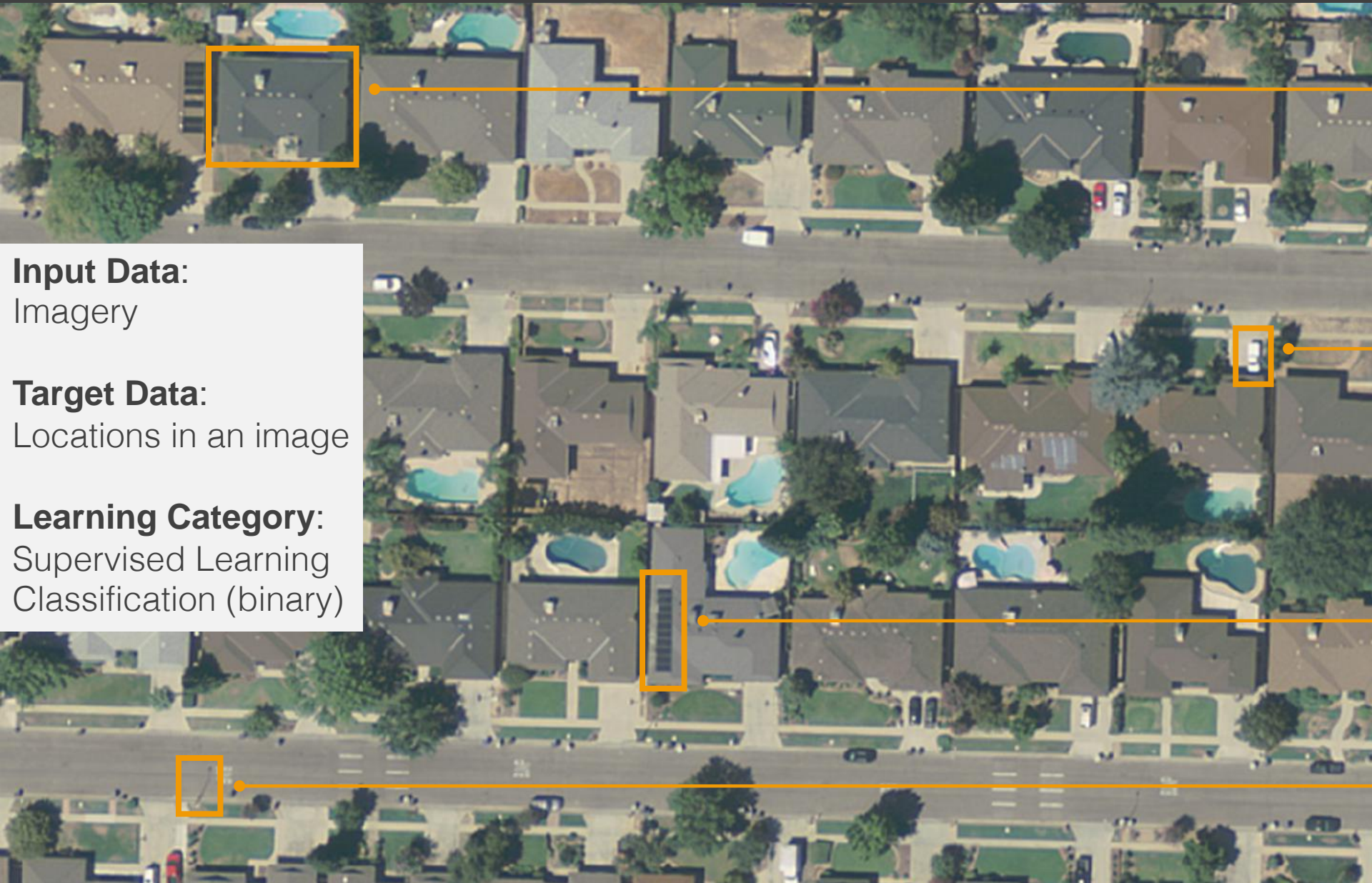
Input Data:
Color Imagery (Image)

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

Learning Category:
Supervised Learning
Classification (binary)

Image source: www.whereswaldo.com/

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:

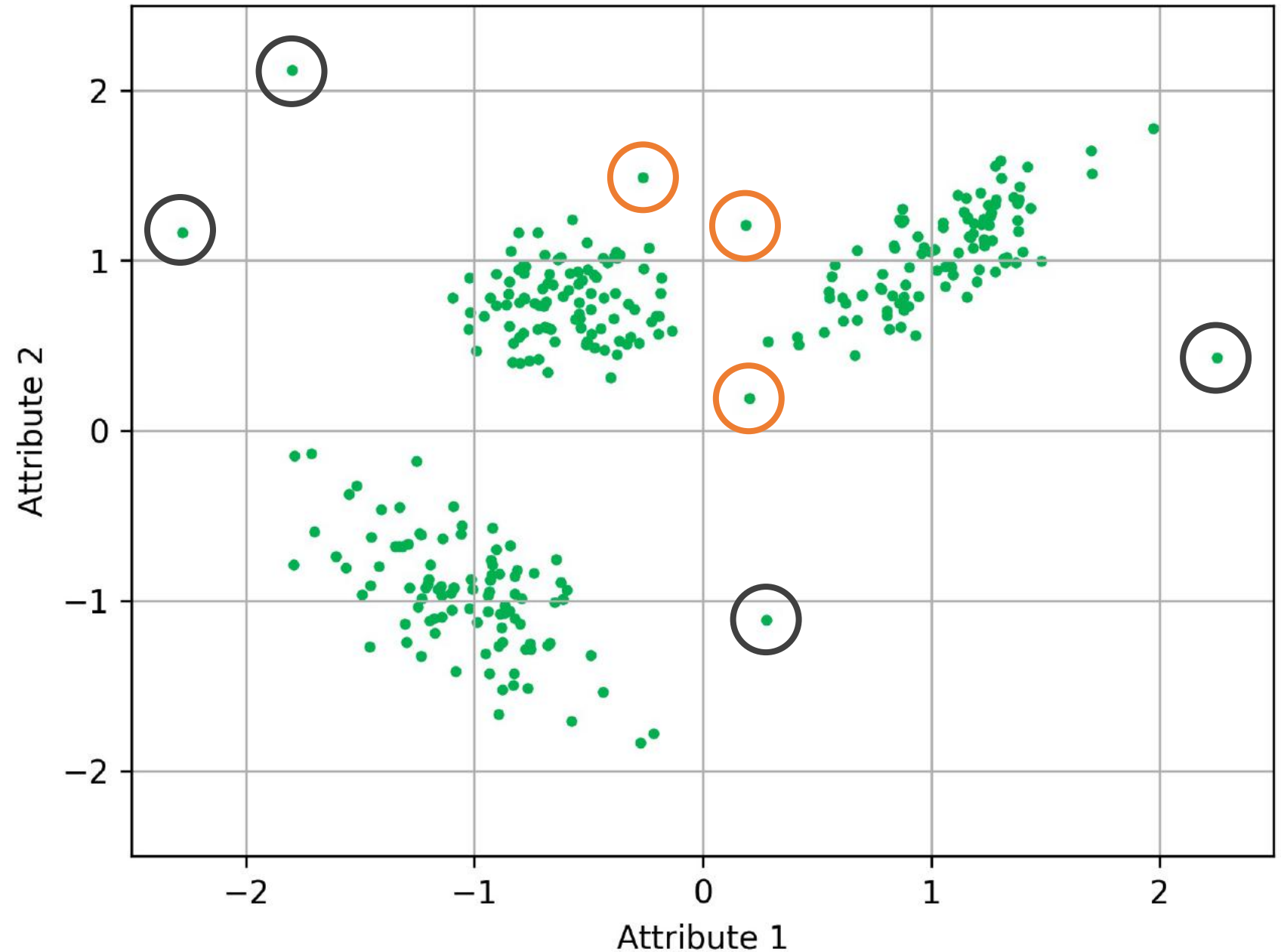
Account transactions, dates,
locations, demographic
information
(Numerical and categorical)

Target Data:

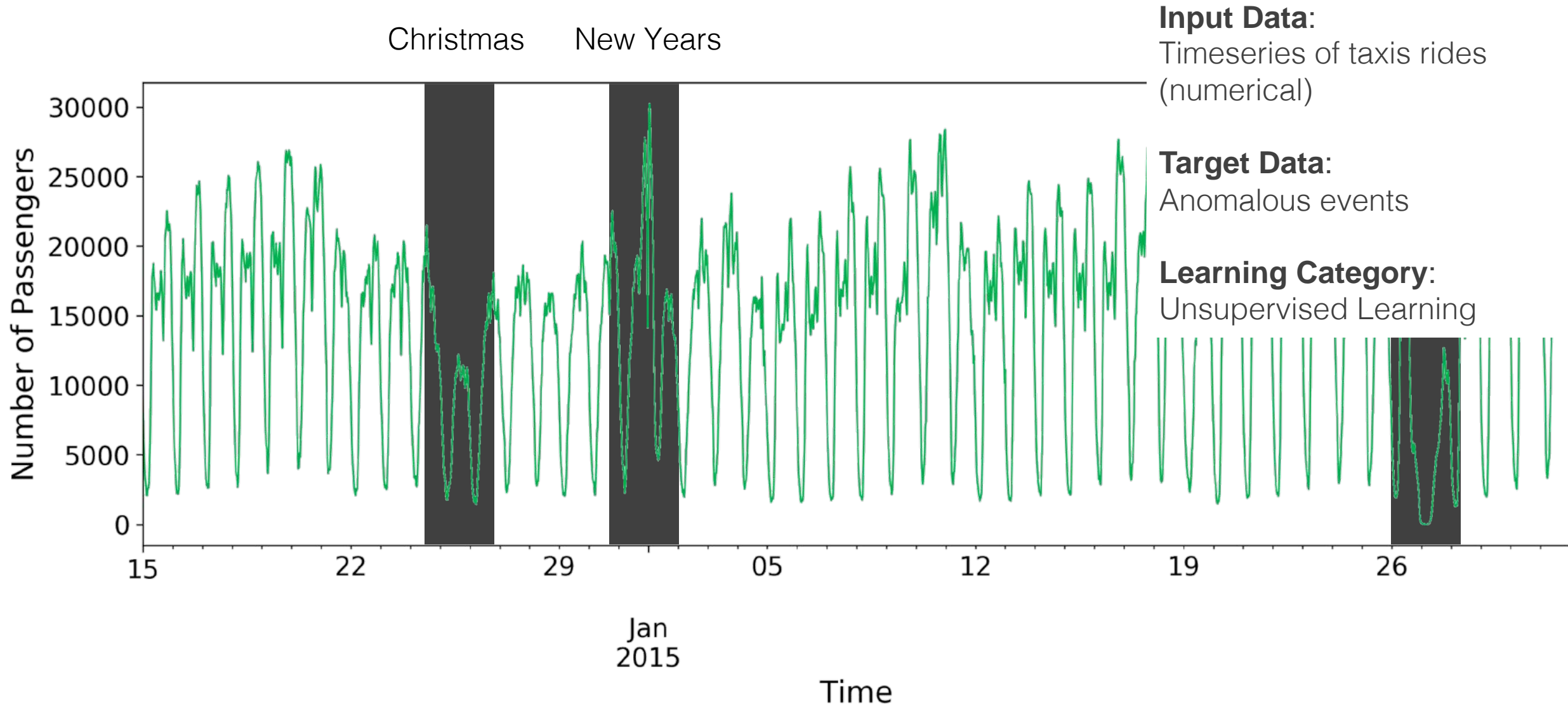
Anomalous transactions

Learning Category:

Unsupervised Learning
Clustering, Density
Estimation



Anomalous Event Detection: NYC Taxis



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

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



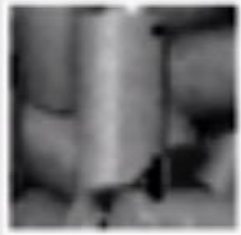
RESULT B + Res



Google DeepMind

Manufacturing – learn to pick up iron cylinders

Success Failure



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	<ul style="list-style-type: none">• Classification• Regression	<ul style="list-style-type: none">• Density estimation• Clustering• Dimensionality reduction• Anomaly detection	<ul style="list-style-type: none">• Model-free learning• Model-based learning

Cautionary advice for aspiring data scientists on machine learning

1



Machine Learning
should **NOT** be the
first tool you reach for
in **all** cases

When NOT to use machine learning

Your objective doesn't require machine learning
(e.g. a data visualization or hypothesis test answers your question)

If you already have the answer
(e.g. domain knowledge provides an answer, simple heuristics work)

Don't have sufficient training data and can't get it
(e.g. highly nonstationary problems, unique cases)



Communicating your findings effectively is often **more important** than squeezing out a tiny bit more performance

Pitch 1

Includes superfluous information that can be easily summarized

Mod_conf_covid	Pred	GT
0.9385243634	1	1
0.6457346346	1	0
0.3184929258	0	0
0.0282716155	0	0
0.8332211769	1	1
0.7290847238	1	1
0.7239875235	1	1
0.3495938237	0	1
0.1929357283	0	0
0.6399959583	1	0

Jumps into the methods without mentioning context

Our deep neural network has a precision of 0.67 and recall of 0.8

Uses metrics that people outside the field may not understand

Pitch 2

Conveys the meaning of the metrics in easy-to-understand language

Our COVID detection model **successfully detects 4 out of 5** instances of COVID-19.

False COVID diagnoses in patients without COVID occur in only **2 of every 6 patients**.

This technique is 30% cheaper than the alternative rapid test and equally effective

Provides context as to why this approach is better than the alternatives

*Note all data/facts on this slide are invented for illustration and not valid

Course logistics

Learning objectives

Through this course you will learn...

- To describe fundamental concepts in **machine learning**
- How to **structure** a machine learning problem
- To automatically make **decisions** from data
- Understand how ML algorithms work and when to use them
- To communicate and effectively **interpret** machine learning output
- To implement your own **end-to-end machine learning project**

Pedagogy

- **Good learning is active learning**
 - Our instructional team provides guidance and structure through the material
 - Most learning is through your work on the assignments, project, and quizzes
- **Desirable difficulty leads to meaningful learning**
 - Creates reusable mental models for independent, lifelong learning
 - Enhances abilities to interpret machine learning results effectively
- **Reading, reflection, and recall is a pattern for effective learning**
 - You'll interact with each concept 4 times typically (lectures, readings, quizzes, and assignment)
 - We will work to avoid the illusion of knowledge

Course website

kylebradbury.github.io/ids705

Course communications

<https://edstem.org/us/courses/16087/discussion/>

(A link is available on Sakai)

Graded Components

Quizzes

(No Collaboration)

20% (One before each class, <1% each)

Assignments

(Some collaboration)

50% (6 assignments, ~8.3% each)

Final project

(Team-based project)

30%

Action items

1. Complete the first set of readings
2. Take Lecture 2 Quiz on Sakai before the start of our next class (can be found on Sakai under “Tests & Quizzes”)
3. Log into each of the course sites (Sakai, Ed Discussions, and Gradescope)
4. Begin working on Assignment #1