

Welcome to MAGIC's Big Data Bootcamp



Sam Schoberg



DEPARTMENT OF
COMPUTER SCIENCE



Premier
League

Goals

- Become comfortable using python
- To understand a data science workflow
- Understand ML topics/algos/issues at a high level
- Become a better Googler
- **Set you up to be able to continue learning after the class is over!**

What I'm hoping for

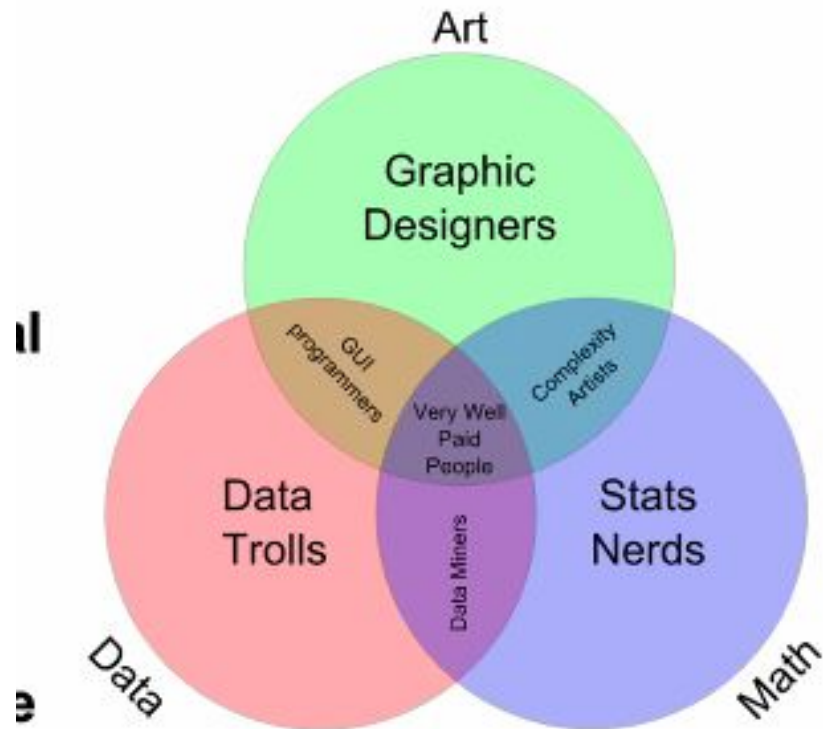
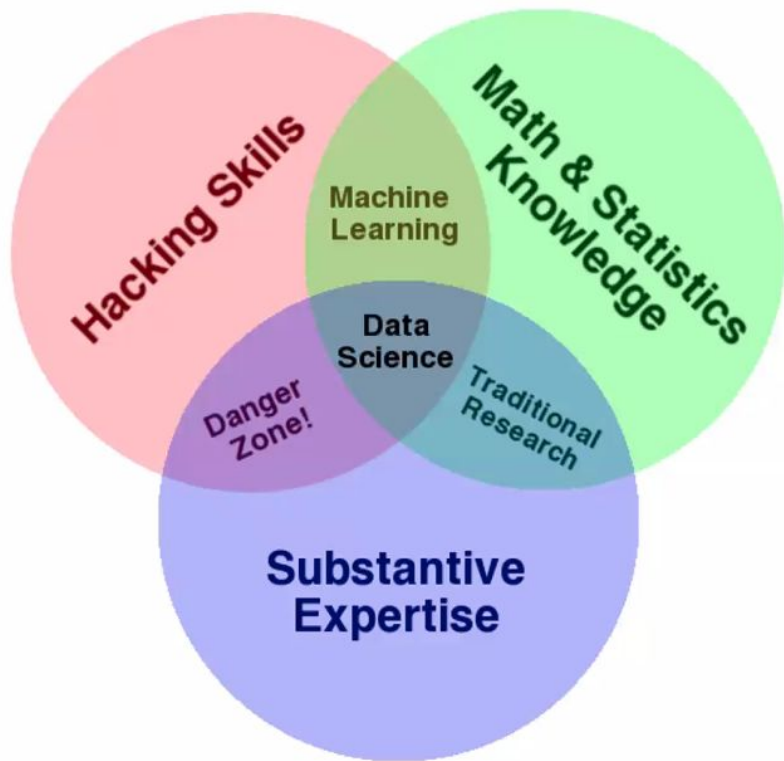
- Collaboration on slack
- Interrupt me with questions
- Answering questions
- Video on, if possible :)

What to expect

1. Intro to Python/Computational Thinking
2. Important libraries (matplotlib, pandas, numpy)
3. Web scraping
4. Overview of ML
5. An end to end ML example
6. Regression
7. Classification
8. ??? (Maybe some cool AWS stuff)

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

~ wikipedia



Why now?

Why now?

- Abundance of data (thanks, software engineers)
- Cheap data storage
- The cloud (cheap processing)



BLOG

Oil vs. Data – Which is more Valuable?

**PETER SILVA**

PUBLISHED APRIL 09, 2019

It depends who you ask.

In recent years there's been a volley of sorts about data replacing oil as the world's most valuable resource. The basic premise is that in this new digital economy, data and what you extract from that data is similar to oil a century ago. An untapped, massive asset that—depending on how you extract and use it—can have enormous rewards. The raw material's value comes from the refinement into a commodity. For oil, it's the energy extracted; for data, it's in the knowledge extracted.

Economists, professors and even CEOs are touting that data is the new oil in today's economy while others are saying

Last questions?



Week 2: Important Libraries



A quick recap...

- Storing data (single vals): integers, floats, strings, booleans
- Storing collections: lists (ordered), dicts/maps (unordered k->v pairs)
- Branching: Control flow of a program
- Loops: Iterates over numbers or a collection
- Functions: avoid rewriting code

<https://magicinc.org/big-data-bootcamp>

Week 2: Goals

1. Perform list operations using numpy
2. Load data into a pandas dataframe
3. Manipulate data in a pandas dataframe
4. Describe data using pandas functions
5. Plot data from dataframes using matplotlib
6. Explore correlation using pandas

Project

- COVID-19 Exploratory Data Analysis
 - Practice reading code that uses libraries

ML Project Checklist

1. Understand the problem
2. Get the data
3. Explore the data
4. Prepare the Data
5. Explore different models
6. Fine tune models
7. Present
8. Maintain

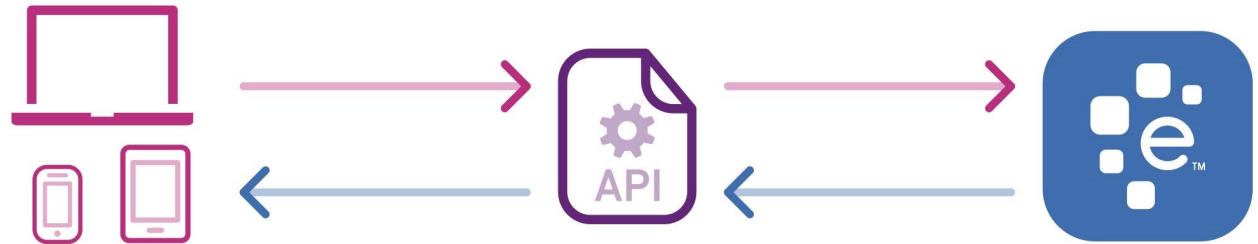
ML Project Checklist

1. Understand the problem
2. Get the data
3. Explore the data
4. Prepare the Data
5. Explore different models
6. Fine tune models
7. Present
8. Maintain



Week 3: Getting Data

BeautifulSoup



Last week:

- Storing data with pandas and dataframes
- Exploratory analysis with pandas
- Visualizations with matplotlib

ML Project Checklist

1. Understand the problem
2. Get the data
3. Explore the data
4. Prepare the Data
5. Explore different models
6. Fine tune models
7. Present
8. Maintain

ML Project Checklist

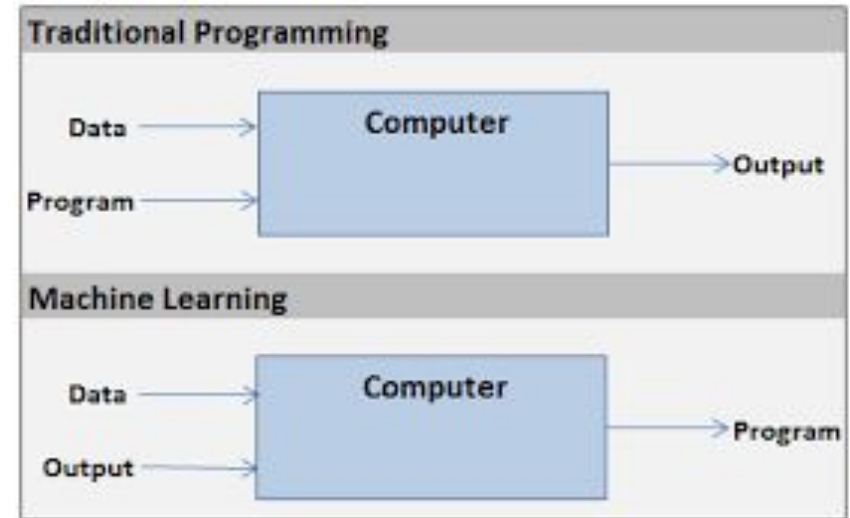
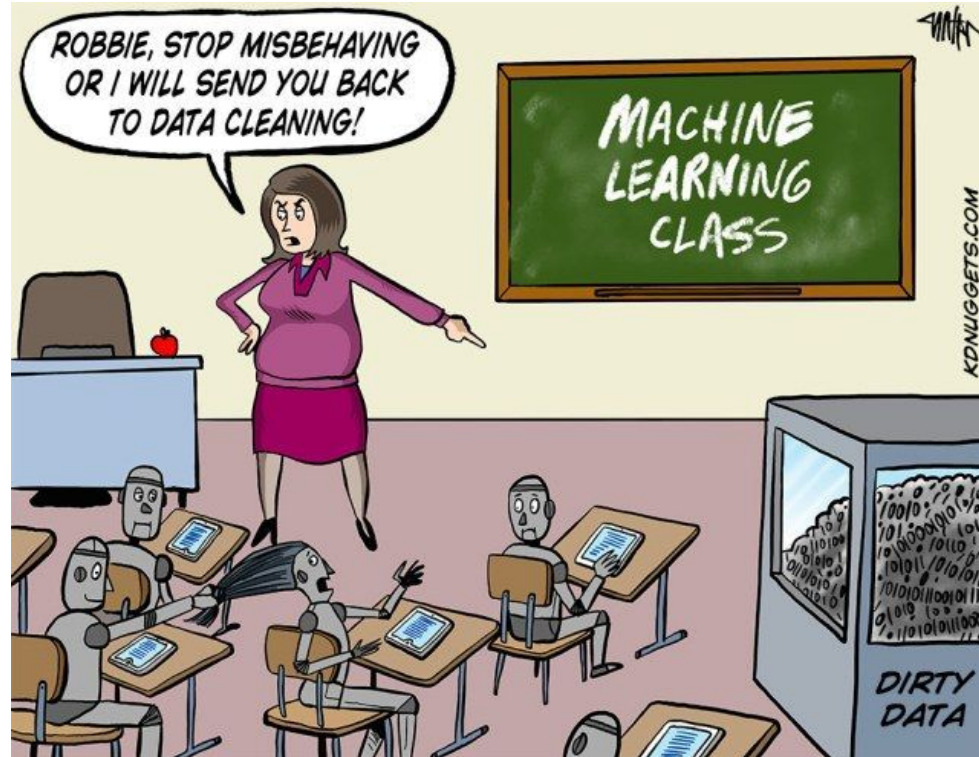
1. Understand the problem
2. Get the data
3. Explore the data
4. Prepare the Data
5. Explore different models
6. Fine tune models
7. Present
8. Maintain



Goals

- Understand structure of HTML sites
- Recognize patterns within HTML
- Use BeautifulSoup to scrape data from HTML sites
- Understand basics of APIs and request data
- Load data into dataframes

Week 4: Intro to ML

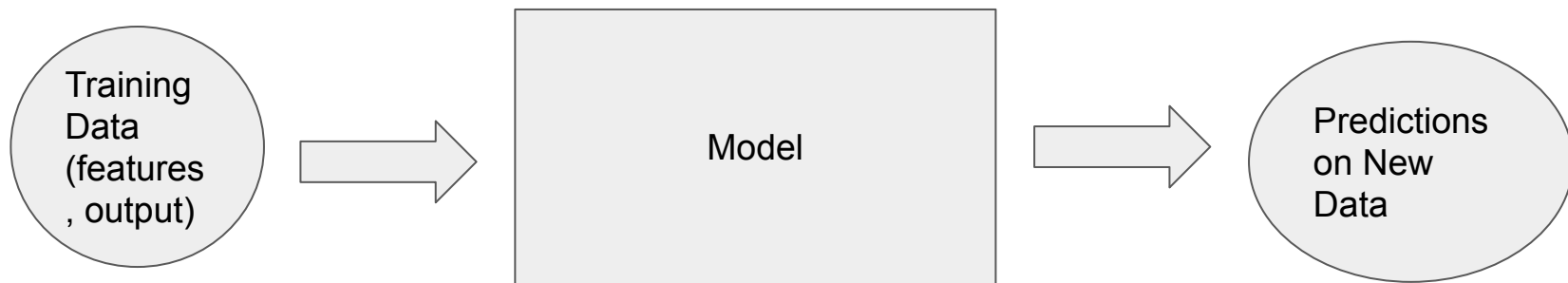


Goals

- Understand ML from a high level
 - What is ML?
 - Why/when to use ML? Which problems are suited for it?
 - Examples of applications
 - Vocabulary
- Understand the different types of ML systems
 - Supervised vs. Unsupervised
 - Online vs. Offline learning
- Identify challenges in ML systems

What is ML?

- The science (and art) of programming machines can learn from data
- Gives computers the ability to learn without being explicitly programmed
- Improve performance at a given task



Example: Spam Filter

- One of the first ML examples pushed to the masses - 90s

Data

Collection of emails labeled
spam or not spam

Why use ML?

- How would you write a spam filter using traditional programming techniques?

★ UNITED NATIONS

Junk - Google August 15, 2018 at 8:56 PM

UN

Your Payment Is Ready.

Reply-To: UNITED NATIONS

Attention Sir/Madam,

Sequel to United Nations public protection policy against fraudulent activities operating in Europe, US and various African banks. This council was set up to fight against scam and fraudulent activities worldwide, responsible for investigating the legitimacy of unpaid contract, inheritance and lotto winning claims by companies and individuals and directs the paying authorities worldwide to make immediate payment of verified claims to the beneficiaries without further delay.

It was resolved that all unpaid claims will be concluded via e-wire transfer through First Sunset Bank, which is very reliable and secure bank. Your beneficiary funds the sum of USD 4.8 million has been forwarded and deposited in First Sunset Bank for instant transfer to you once you contact them.

You are advised to contact First Sunset Bank via below email, to guide you further on the wire transfer procedures:

First Sunset Bank.

Email: firstsunsetbank@web.cg

Contact Person: Mrs. Agnes Scott

Please be informed that transfer time is limited sequel to policy, therefore you are advised to attend as soon as you read this email and also reconfirm your full details to them. We have copied all our co-ordinate security agencies for record purposes.

Thank you.

Your Faithfully,

Mrs. Ann Walter.

Director, Special Duties.

United Nations Security Council.

Why use ML?

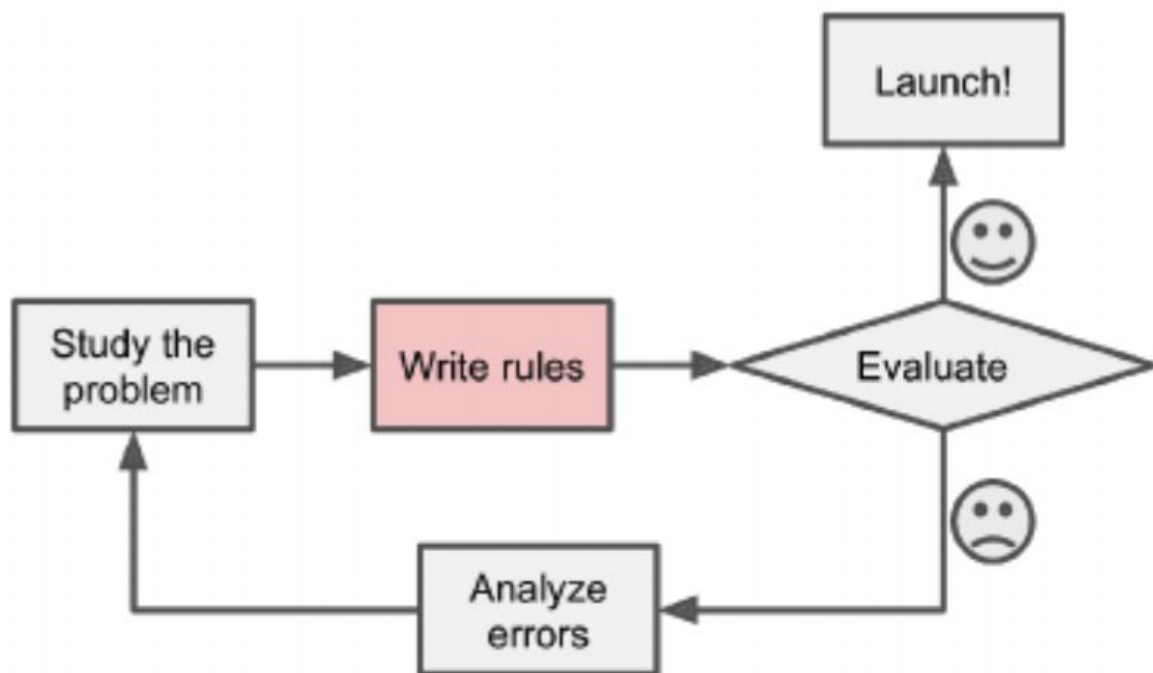
- How would you write a spam filter using traditional programming techniques?
 - Bag of words (4U, URGENT, !!!!!, Nigerian Prince, free, amazing)
 - Unrecognized email address
 - Urgent subject line
 - Typos

Why use ML?

- How would you write a spam filter using traditional programming techniques?
 - Bag of words (4U, URGENT, !!!!!, Nigerian Prince, free, amazing)
 - Unrecognized email address
 - Urgent subject line
- Detection Algorithm
 - Python “in”
 - If enough flags triggered, mark it as spam

Why use ML?

- How would you write a spam filter using traditional programming techniques?
 - Bag of words (4U, URGENT, !!!!!, Nigerian Prince, free, amazing)
 - Unrecognized email address
 - Urgent subject line
- Detection Algorithm
 - Python “in”
 - If enough flags triggered, mark it as spam
- Test and deploy



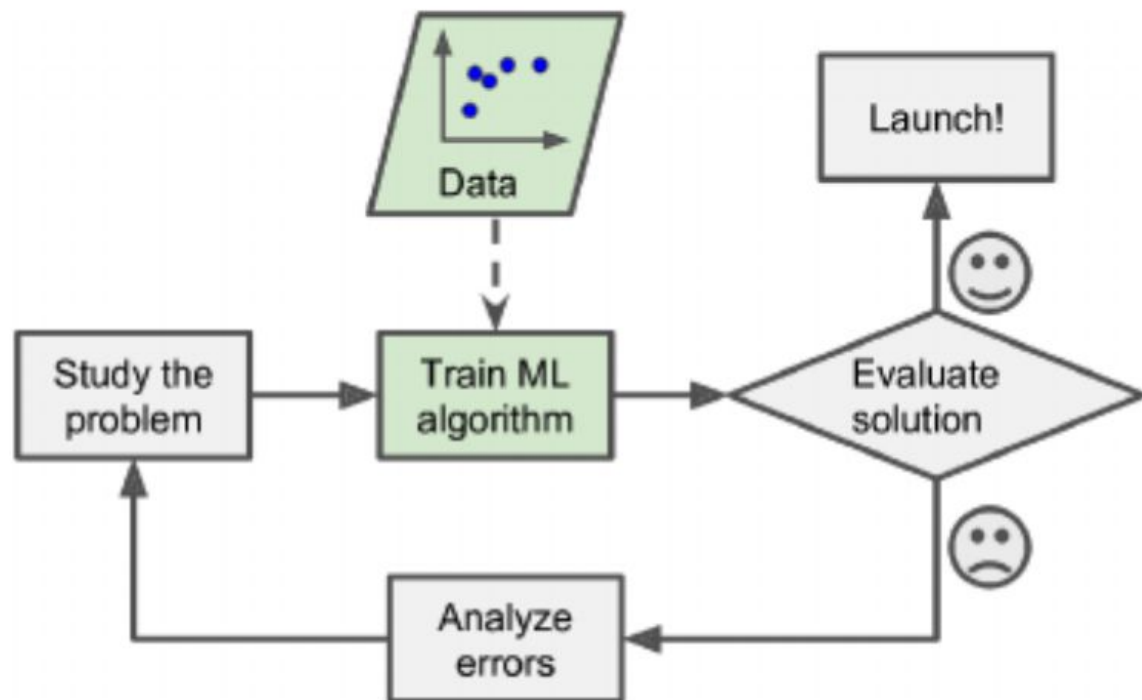
Problems?

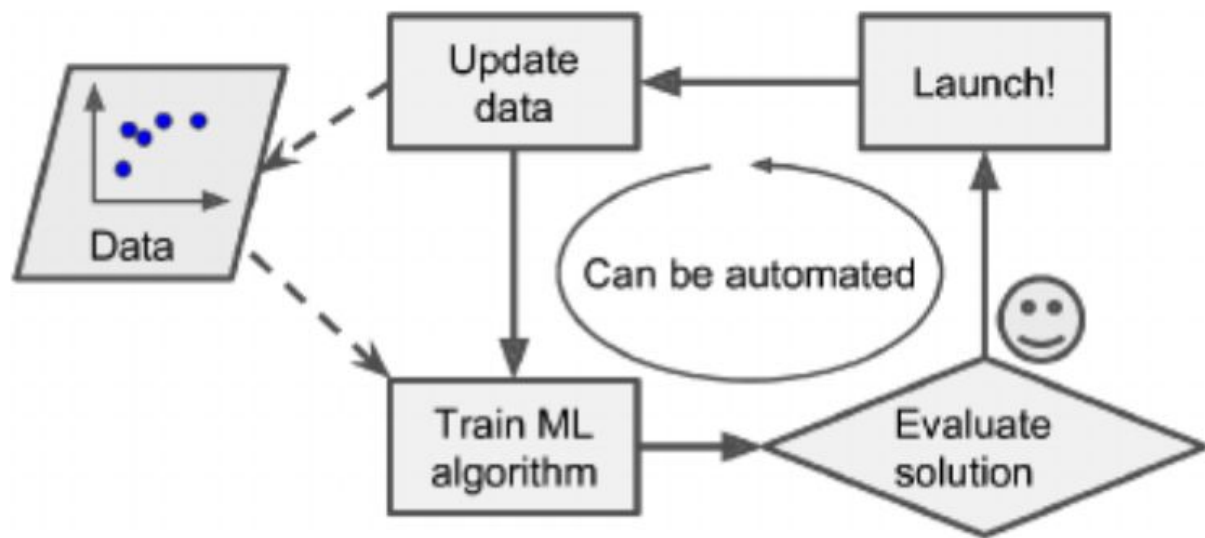
Problems?

- Long list of complex rules
- Doesn't update to live data
 - What if spammers start using different keywords or subject lines?

A ML algorithm can fix these issues...

- Automatically learn which words/phrases constitute spam
- Constantly updates the model to determine new spam techniques





Things ML is good for...

- Problems that require fine-tuning or long lists of rules
- Complex problems where there is no known algorithmic solution
- Fluctuating environments
- Lots o' data...

Key Terms: Label

A **label** is the thing we're predicting—the y variable in simple linear regression. The label could be the future price of wheat, the kind of animal shown in a picture, the meaning of an audio clip, or just about anything.

SPAM or NOT SPAM

Key Terms: Feature

A **feature** is an input variable—the x variable in simple linear regression. A simple machine learning project might use a single feature, while a more sophisticated machine learning project could use millions of features, specified as:

x_1, x_2, \dots, x_N

In the spam detector example, the features could include the following:

- words in the email text
- sender's address
- time of day the email was sent
- email contains the phrase "one weird trick."

Key Terms: Model

A model defines the relationship between features and label. For example, a spam detection model might associate certain features strongly with "spam". Let's highlight two phases of a model's life:

- **Training** means creating or **learning** the model. That is, you show the model labeled examples and enable the model to gradually learn the relationships between features and label.
- **Inference** means applying the trained model to unlabeled examples. That is, you use the trained model to make useful predictions (y'). For example, during inference, you can predict `medianHouseValue` for new unlabeled examples.

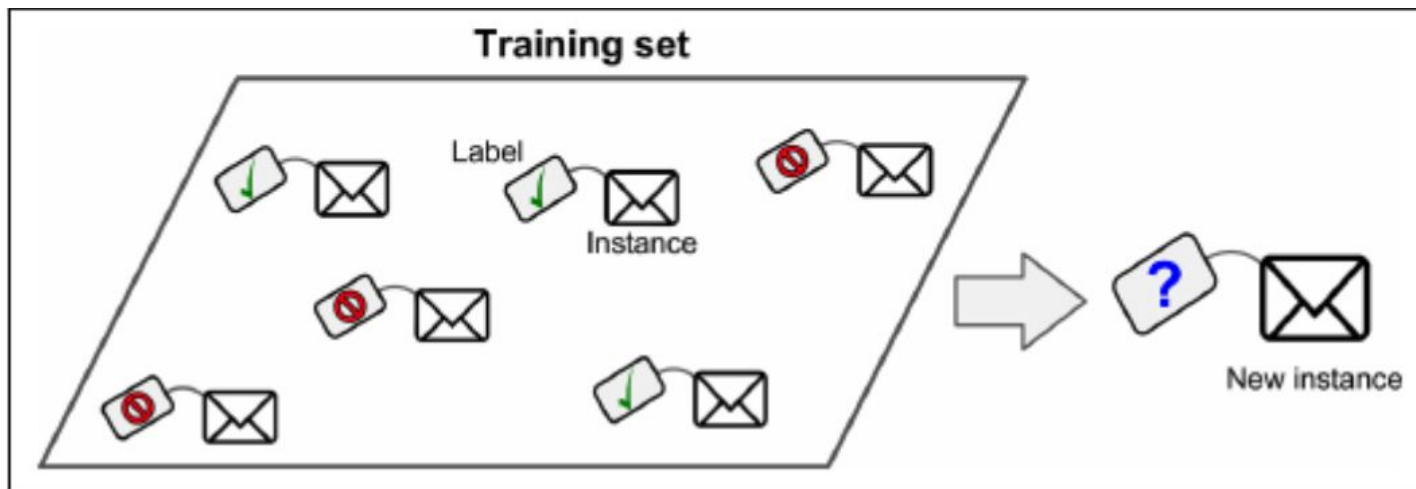
Types of ML Systems

Three main questions to ask...

1. Are they trained with human supervision? (supervised vs. unsupervised learning)
2. Can they learn on the fly? (online vs. batch learning)
3. Do they compare new data to old data points or do they use patterns learned from old data to make predictions? (instance based vs. model based learning)

Supervised Learning

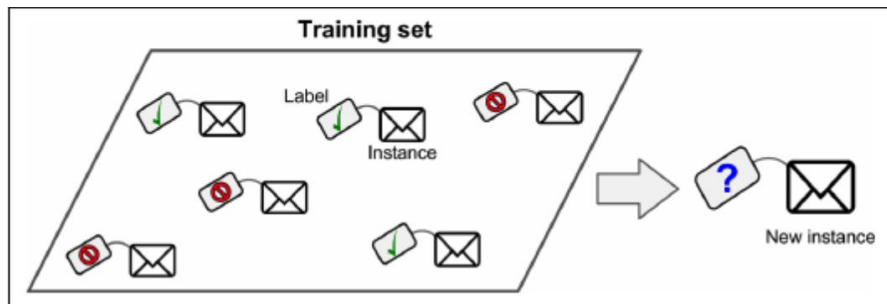
- In supervised learning, the training set you feed to the algorithm includes the desired solutions, called *labels*.



1. Supervised Learning: Regression vs. Classification

Classification

- Algorithm aims to predict a *class* based on *features*



Regression

- Algorithm aims to predict a *value* based on *features*



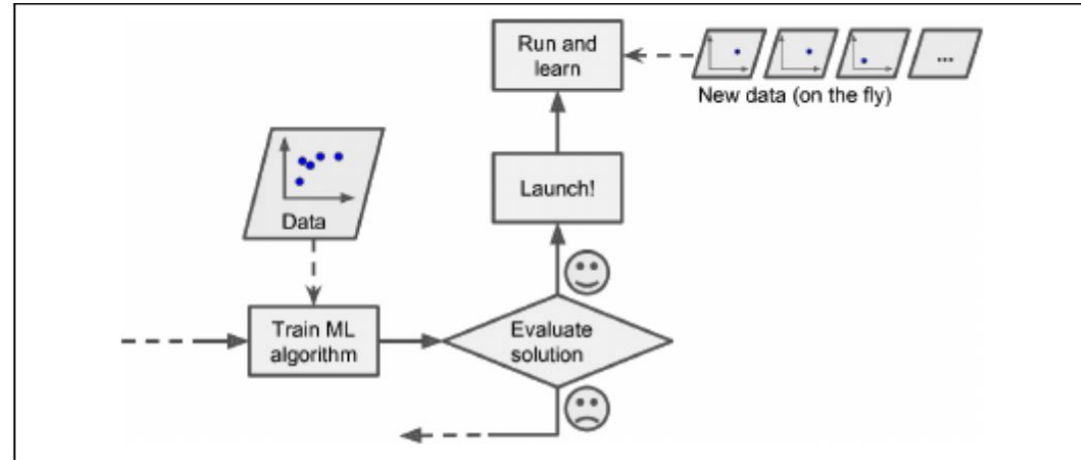
2. Can the algorithm learn incrementally?

Batch Learning

- Must be trained every time using all available data.
- Tradeoffs?
- Examples: data only available every so often.

Online Learning

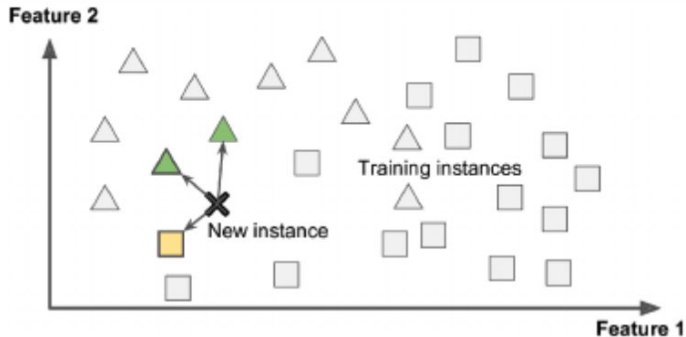
- Can be trained incrementally.
- Tradeoffs?
- Examples: Stock prices, live data.



3. How does the algorithm make predictions?

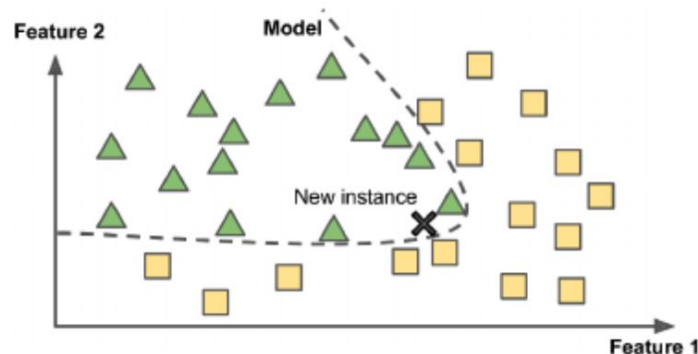
Instance Based Learning

- Make predictions on new data based on known similar values.
- Requires a *measure of similarity*
- Learn training data by heart, generalize new data by *closeness* to old data



Model Based Learning

- Make a model from training data and make predictions based on that model.



3. How does the algorithm make predictions?

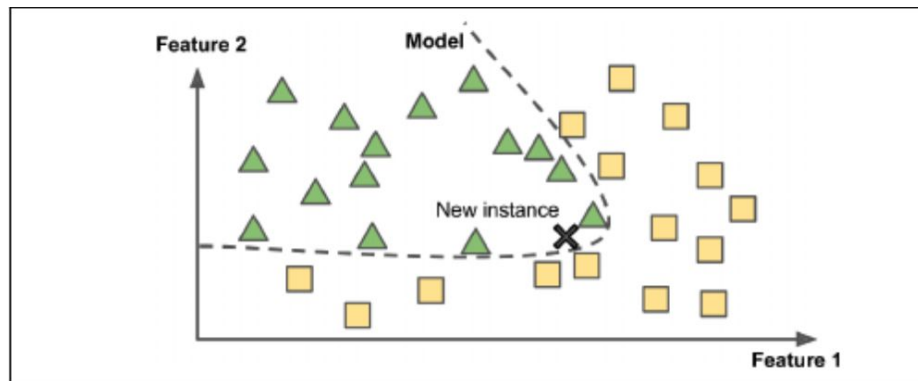
Instance Based Learning

MINIMIZE A COST FUNCTION (WRONG ANSWERS)



Model Based Learning

MINIMIZE A COST FUNCTION (WRONG ANSWERS)



To recap...

- Supervised learning: data has labels
 - Regression vs. Classification
- Batch Learning vs. Online Learning
- Instance Based Learning vs. Model Based Learning

To recap...

- You studied the data
- You selected a model
- You trained the model on the training data
- You applied the model to make predictions on new data

Challenges: The Data

1. Insufficient quantity of training data
2. Non-representative training data
3. Poor-Quality Data
4. Irrelevant Features

1. Insufficient Data

Mail thinks this message is Junk Mail.

Move to Inbox

★ UNITED NATIONS

Junk - Google August 15, 2018 at 8:56 PM

UN

Your Payment Is Ready.

Reply-To: UNITED NATIONS

Attention Sir/Madam,

Sequel to United Nations public protection policy against fraudulent activities operating in Europe, US and various African bank council was set up to fight against scam and fraudulent activities worldwide, responsible for investigating the legitimacy of unpaid contract, inheritance and lotto winning claims by companies and individuals and directs the paying authorities worldwide to make immediate payment of verified claims to the beneficiaries without further delay.

It was resolved that all unpaid claims will be concluded via e-wire transfer through First Sunset Bank, which is very reliable and secure bank. Your beneficiary funds the sum of USD 4.8 million has been forwarded and deposited in First Sunset Bank for immediate transfer to you once you contact them.

You are advised to contact First Sunset Bank via below email, to guide you further on the wire transfer procedures:

First Sunset Bank.

Email: firstsunsetbank@web.cg

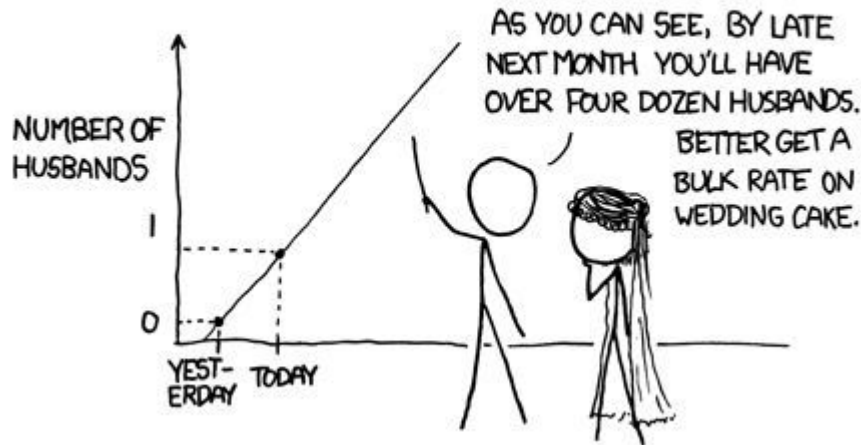
Contact Person: Mrs. Agnes Scott

Please be informed that transfer time is limited sequel to policy, therefore you are advised to attend as soon as you read this email and also reconfirm your full details to them. We have copied all our co-ordinate security agencies for record purposes.

Thank you.

Your Faithfully,
Mrs. Ann Walter.
Director, Special Duties.
United Nations Security Council.

MY HOBBY: EXTRAPOLATING



2. Nonrepresentative Training Data

- Your training data be representative of the new cases you want to generalize to.

EURO ONLINE LOTTO DE NAVIDAD S.A

ANWALTSKANZLEI CARLOS PEDRO & CO

HAUPTBÜRO: CL. DE AYALA 11 28001 MADRID SPANIEN

TEL/FAX: 0034 666-717-858 E-MAIL:

pedrocarloslawfirm@gmail.com

REFERENZ NR: PBA/0809/EU; Übersetzer Kopie

Sehr Geehrter Begünstigter;

22/09/2020

Abschließende Mitteilung für die Zahlung des nicht beanspruchten Preisgeldes

Wir möchten Sie informieren, dass das Büro des nicht Beanspruchten Lotteriede Preisgeldes in Spanien, unsere Anwaltskanzlei ernannt hat, als gesetzliche Berater zu handeln, in der Verarbeitung und der Zahlung eines Preisgeldes, das auf Ihrem Namen gutgeschrieben wurde seit über zwei Jahren nicht beansprucht wurde.

Der Gesamtbetrag der ihnen zusteht beträgt momentan €1.340.200,00 EUROS

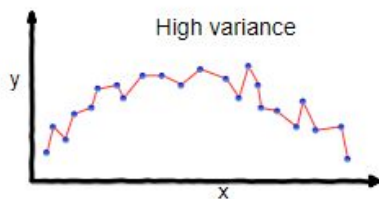
Das ursprüngliche Preisgeld betrug €1.250.000,00 EUROS. Diese Summe würde für nun mehr als zwei Jahre Gewinnbringend angelegt, daher die aufstockung auf die oben benannte Gesamtsumme. Entsprechend dem Büros des nicht Beanspruchten Preisgeldes, würde dieses Geld als nicht beanspruchten Gewinn einer Lotteriefirma bei ihnen zum verwalten niedergelegt und in ihrem namen versichert. Nach Ansicht der Lotteriefirma würde ihnen das Geld nach einer Weihnachtsförderunglotterie zugesprochen. Die Kupons würden von einer Investmentgesellschaft gekauft.Nach Ansicht der Lotteriefirma wurden sie damals Angeschrieben um Sie über dieses Geld zu informieren, es hat sich aber leider bis zum Ablauf der gesetzten Frist keiner gemeldet um den Gewinn zu Beanspruchen. Dieses war der Grund weshalb das Geld zum verwalten niedergelegt wurde. Gemäß des Spanischen Gesetzes muss der inhaber Oder inhaberin alle zwei Jahre über seinen vorhanden Gewinn informiert werden. Sollte dass Geld wieder nicht beansprucht werden, wird der Gewinn abermals über eine Investmentgesellschaft für eine weitere Periode von zwei Jahren angelegt werden.Wir sind daher, durch das Büro des nicht Beanspruchten Preisgelds beauftragt worden sie anzuschreiben. Dies ist eine Notifikation für das Beanspruchen dieses Gelds.

3. Poor Quality Data/Irrelevant Features

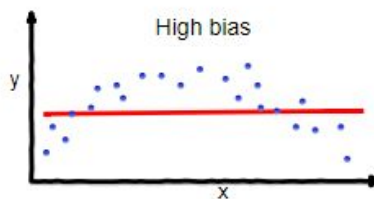
- Some observations are clearly outliers
- If observations are missing features (someone didn't specify their age on a form).
- Features aren't valuable to the prediction (time email is sent).

Pitfalls

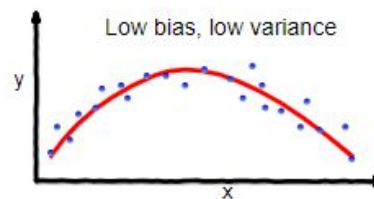
- Overfitting: Performs well on training data, but doesn't generalize to new data well
 - Taxi driver
- Underfitting: Model too simple to find the patterns in the data



overfitting

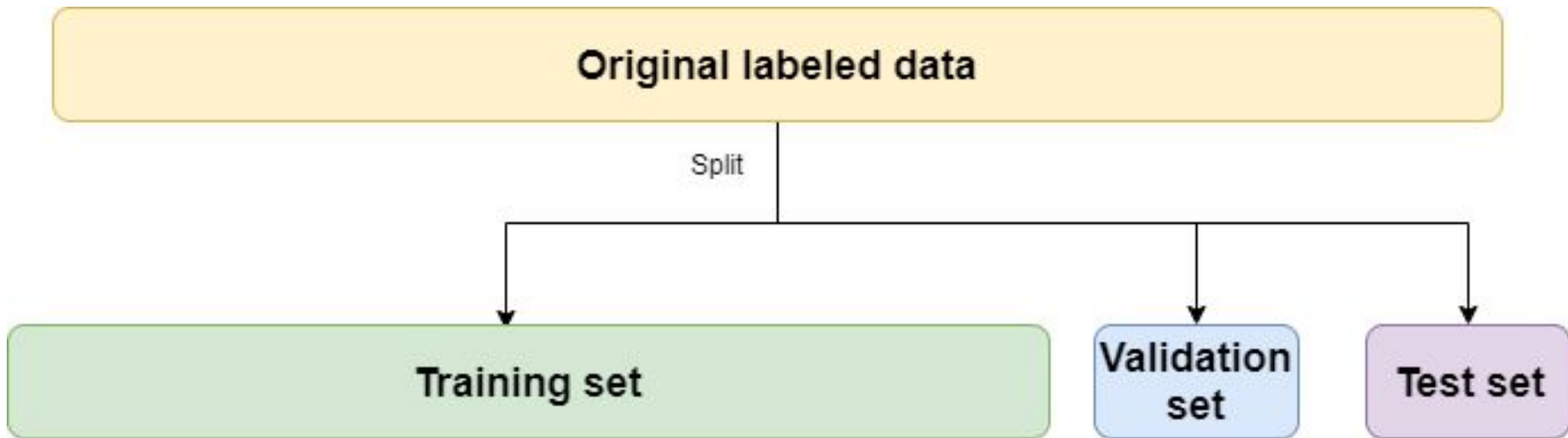


underfitting



Good balance

Solutions



Resources

<https://developers.google.com/machine-learning/glossary>

<https://www.youtube.com/watch?v=zPG4NjlkCjc>

Week 5: Intro to Regression

High school math + algorithms! :)

Last weeks

- Tonight: Intro to Regression (lecturing)
- 10/19: Implementing regression (coding)
- 10/26: Intro to classification (lecturing)
- 11/2: Implementing classification (coding)

Goals

- Understand linear regression from a high level
- Measure the accuracy of a model with mean squared error (MSE)
- Understand the idea of loss for a function
- Understand how to use loss to find a better model
- Learn about hyperparameters and how they affect training

In a nutshell...

ML systems learn

How to combine input

To produce useful predictions

On never-before-seen-data

Our task...

What is the value of a house in California? Supervised regression.

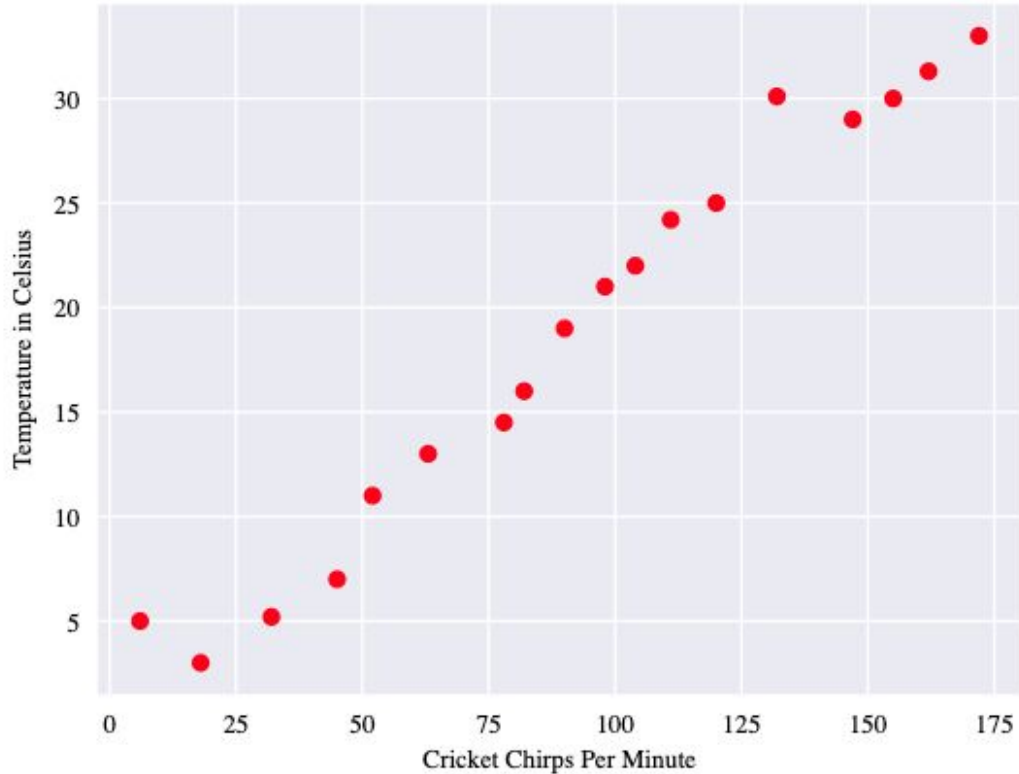
Features ($x_0, x_1, x_2, \dots, x_n$)

Label (y)

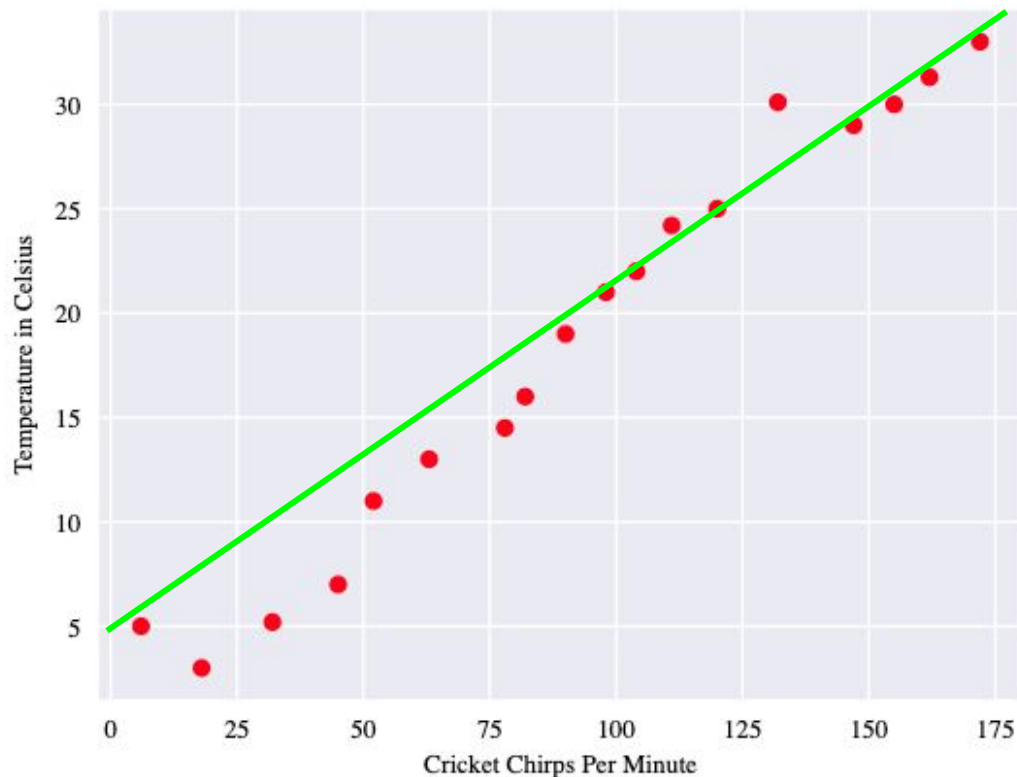
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-114.31	34.19	15.0	5612.0	1283.0	1015.0	472.0	1.4936	66900.0
1	-114.47	34.40	19.0	7650.0	1901.0	1129.0	463.0	1.8200	80100.0
2	-114.56	33.69	17.0	720.0	174.0	333.0	117.0	1.6509	85700.0
3	-114.57	33.64	14.0	1501.0	337.0	515.0	226.0	3.1917	73400.0
4	-114.57	33.57	20.0	1454.0	326.0	624.0	262.0	1.9250	65500.0

$$y' = b + w_1x_1 + w_2x_2 + w_3x_3$$

Building linear regression intuition...



Building linear regression intuition...



$$Y = mx + b$$

y : Temperature in celcius

m: is the slope of the line

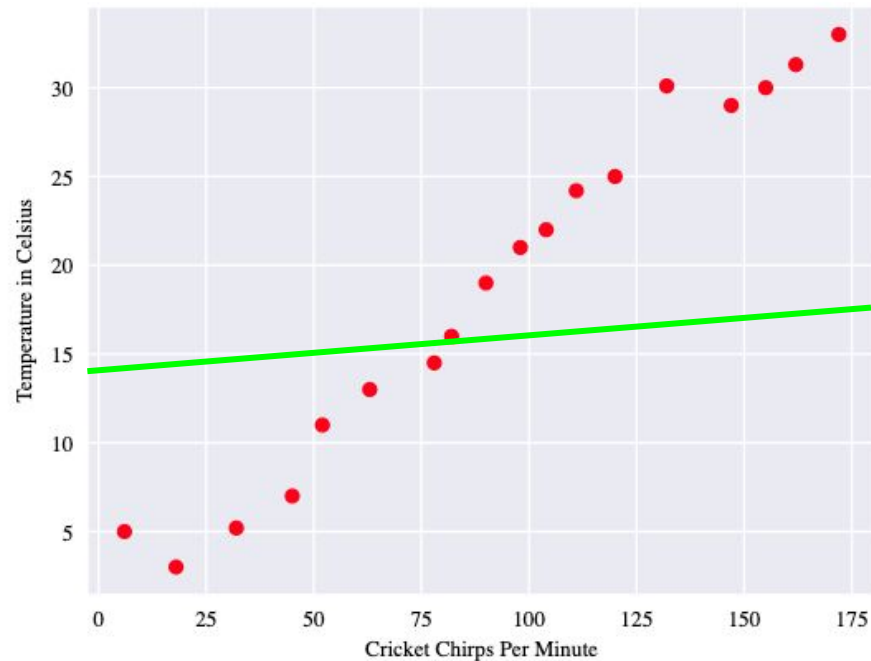
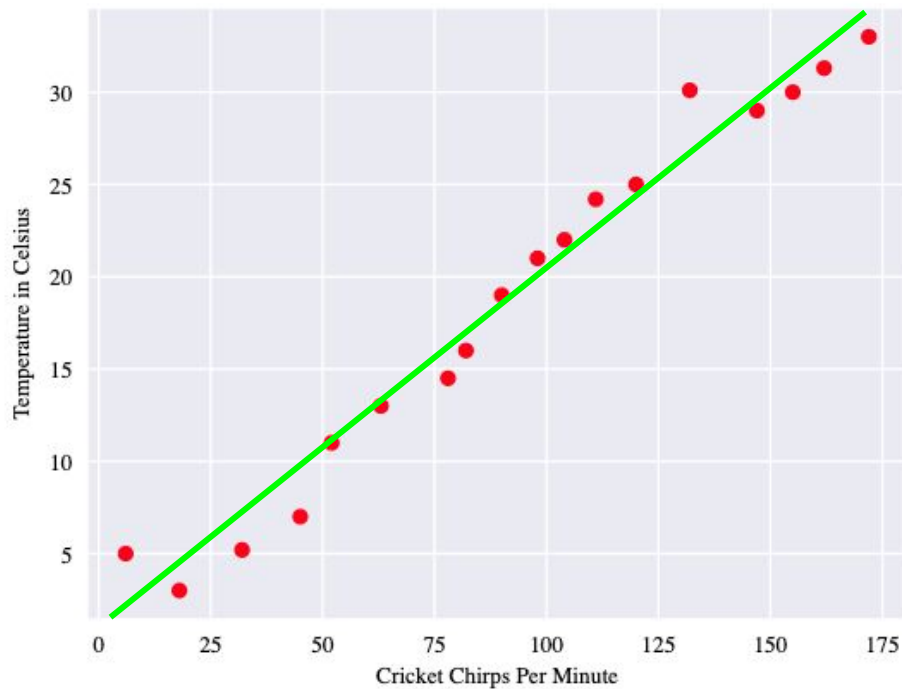
x: is the number of chirps per minute

b : is the y intercept

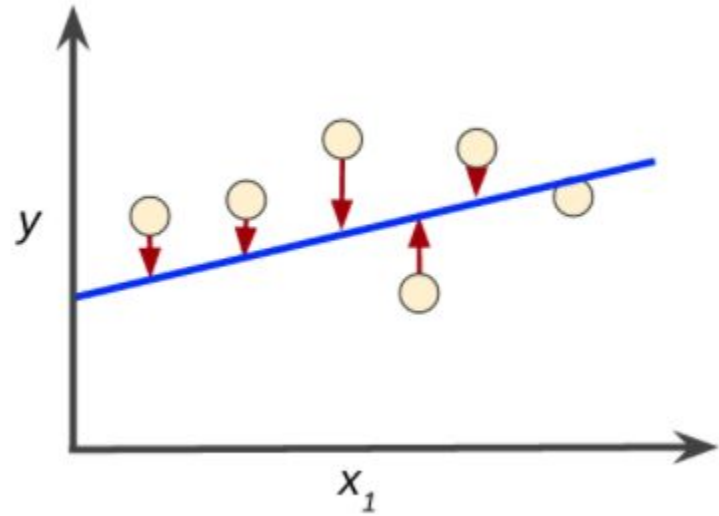
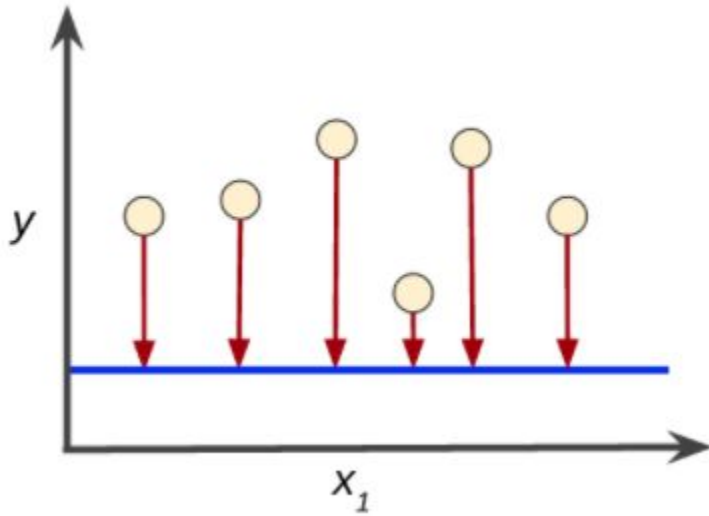
$$y' = b + w_1 x_1$$

$$Y = 5 + (30/150)x_1$$

Building linear regression intuition...

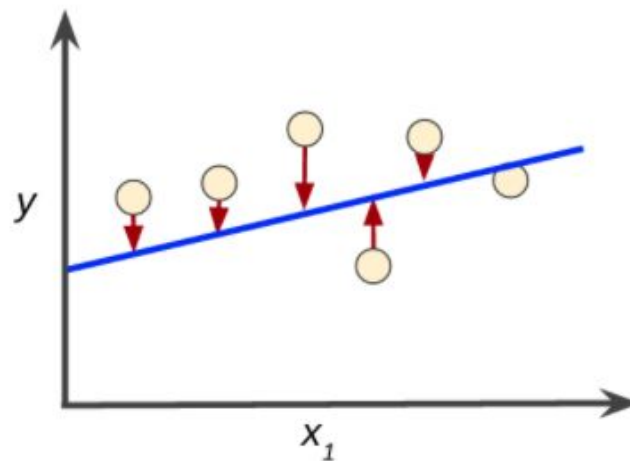
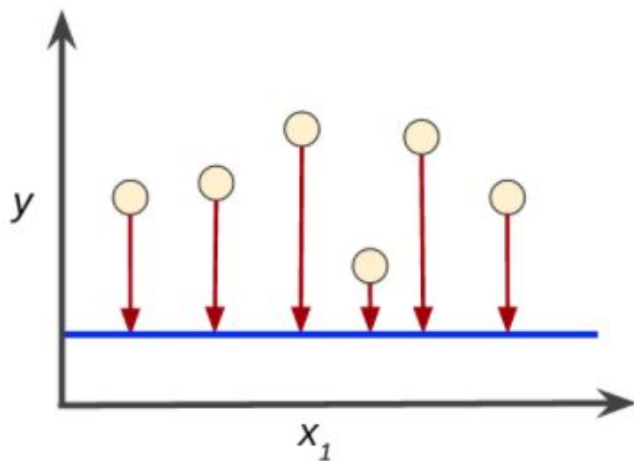


Building linear regression intuition...



Building linear regression intuition...

- **Training** a model simply means learning (determining) good values for all the weights and the bias from labeled examples.
- The goal of training a model is to find a set of weights and biases that have *low* loss, on average, across all examples

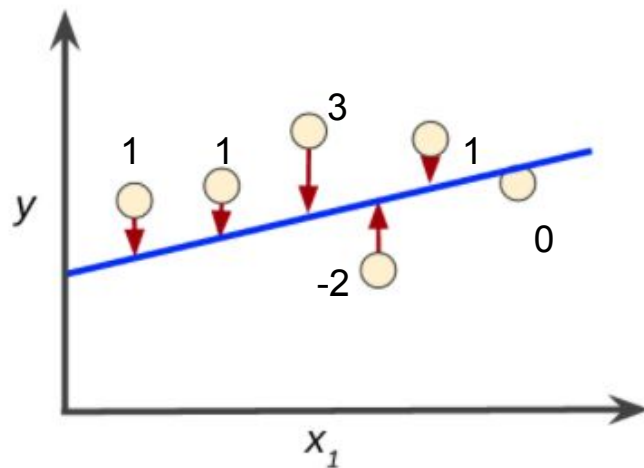
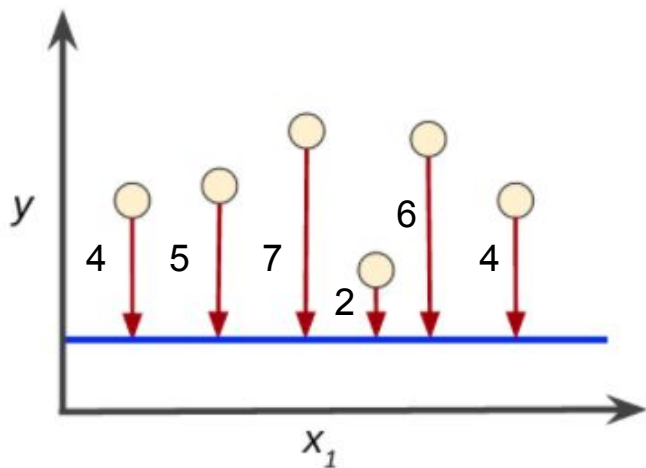


Building linear regression intuition...

= the square of the difference between the label and the prediction

$$= (\text{observation} - \text{prediction}(\mathbf{x}))^2$$

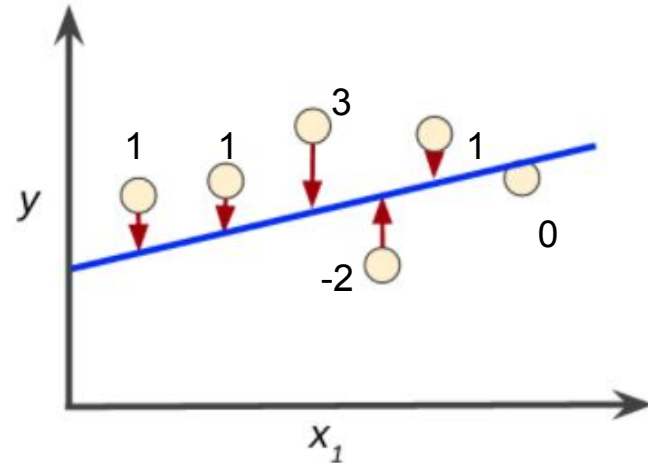
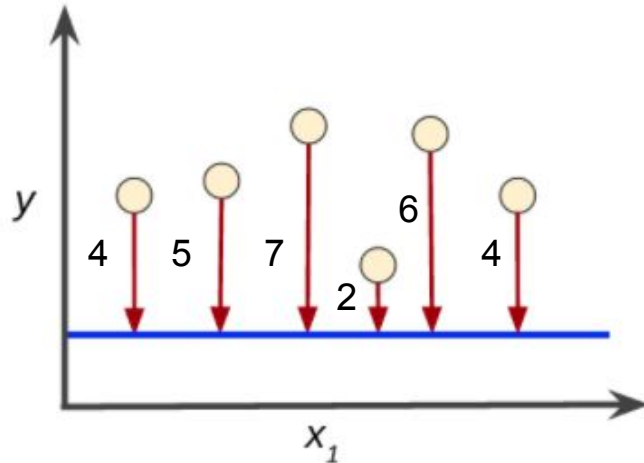
$$= (y - y')^2$$



Building linear regression intuition...

$$= 4^2 + 5^2 + 7^2 + 2^2 + 6^2 + 4^2 = 146$$

$$1^2 + 1^2 + 3^2 + (-2)^2 + 1^2 + 0^2 = 16$$



Recap

- Linear regression: fitting a line to a collection of data points to then make predictions
- We want to find values for w_1 and b that make a “good” model
- A “good” model is one that has a low total **loss**
- A popular way to measure loss is squared loss

$$y' = b + w_1 x_1$$

Recap

- Linear regression: fitting a line to a collection of data points to then make predictions
- We want to find values for w_1 and b that make a “good” model
- A “good” model is one that has a low total **loss**
- A popular way to measure loss is squared loss

$$y' = b + w_1 x_1$$

HOW DO WE FIND A GOOD MODEL???

Training a Linear Regression Model



Training a Linear Regression Model

$$y' = b + w_1 x_1$$

Very low loss

W_1, B

Training a Linear Regression Model

Pick random values for slope (w), bias (b)

Make an initial prediction on the training data and compute loss

While you haven't found the best model

- Change slope and bias slightly and recompute loss

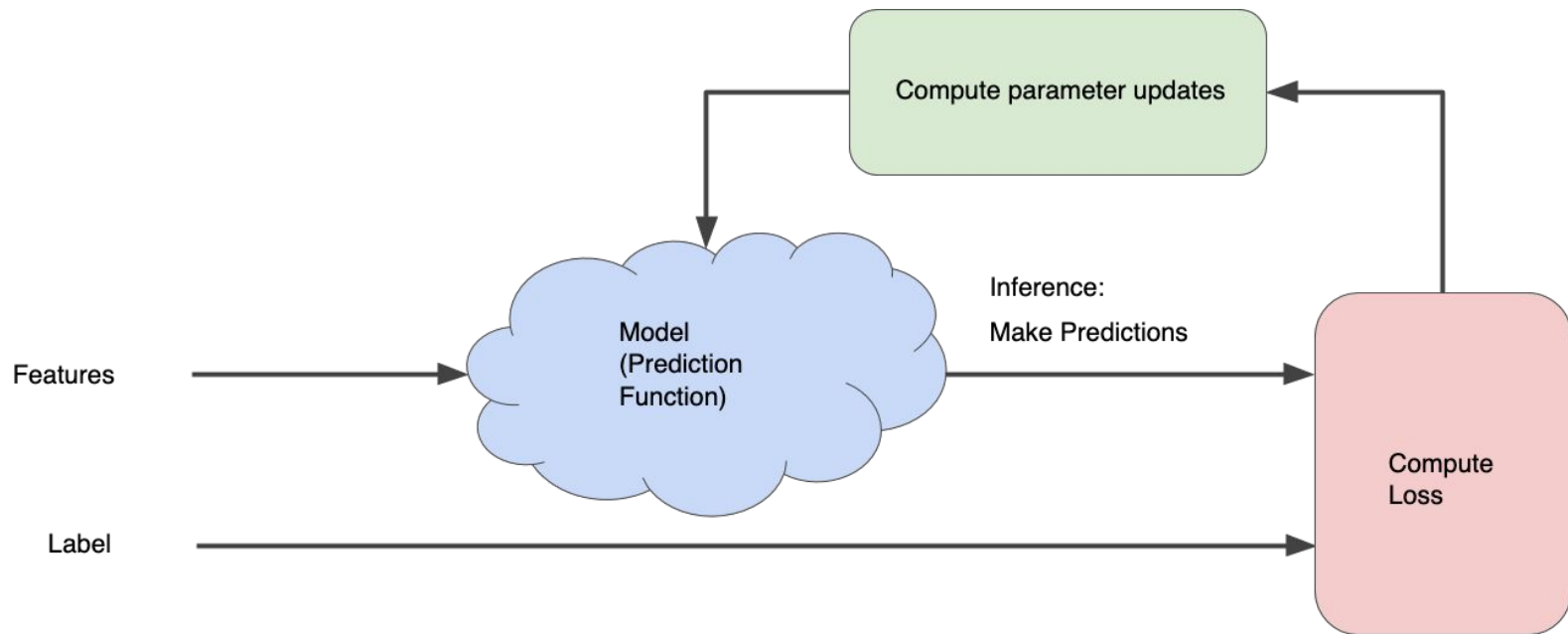
- If the loss is less than before:

 - Keep moving w , b in that direction

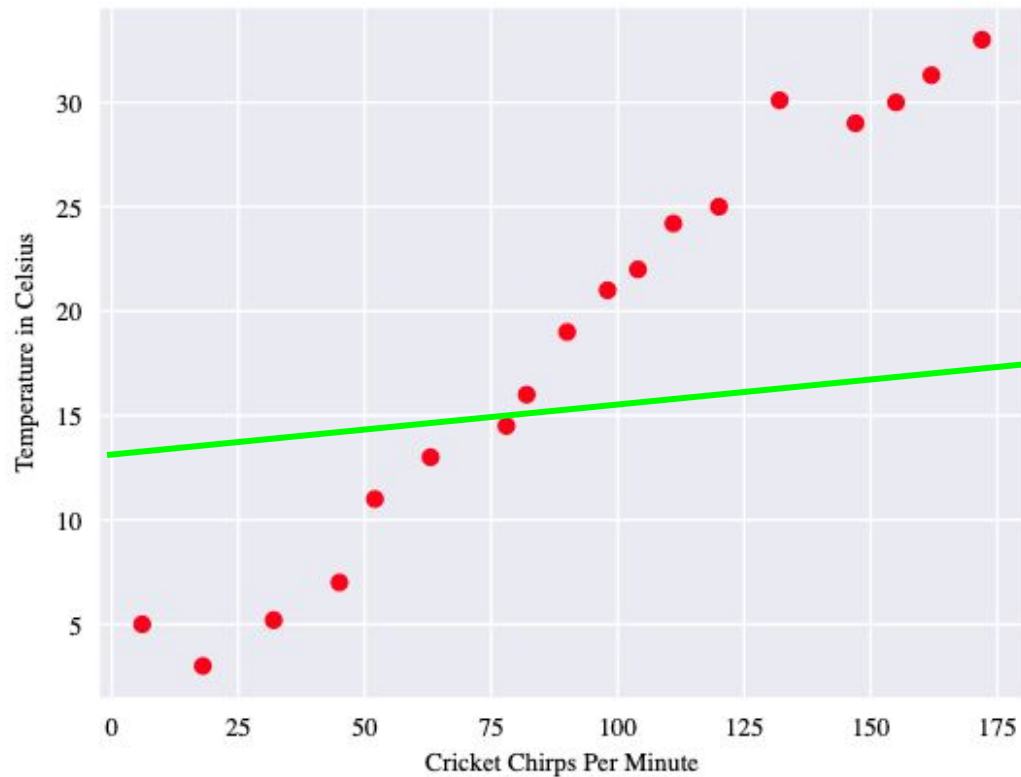
- Else:

 - Move w , b in another direction

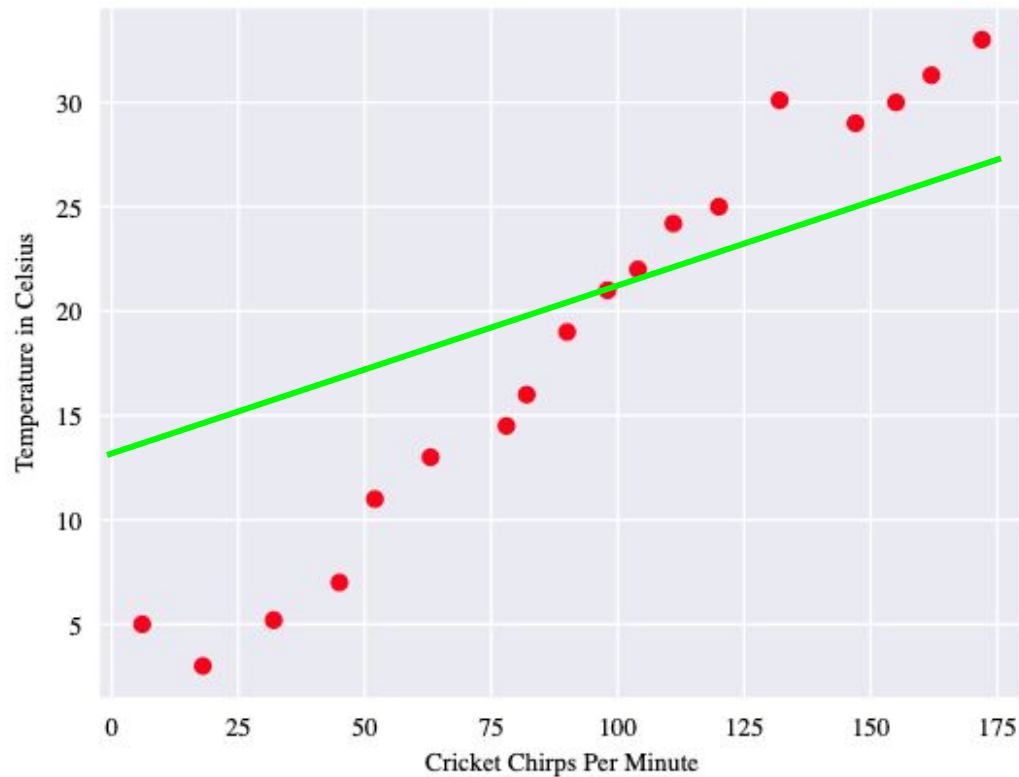
Training a Linear Regression Model



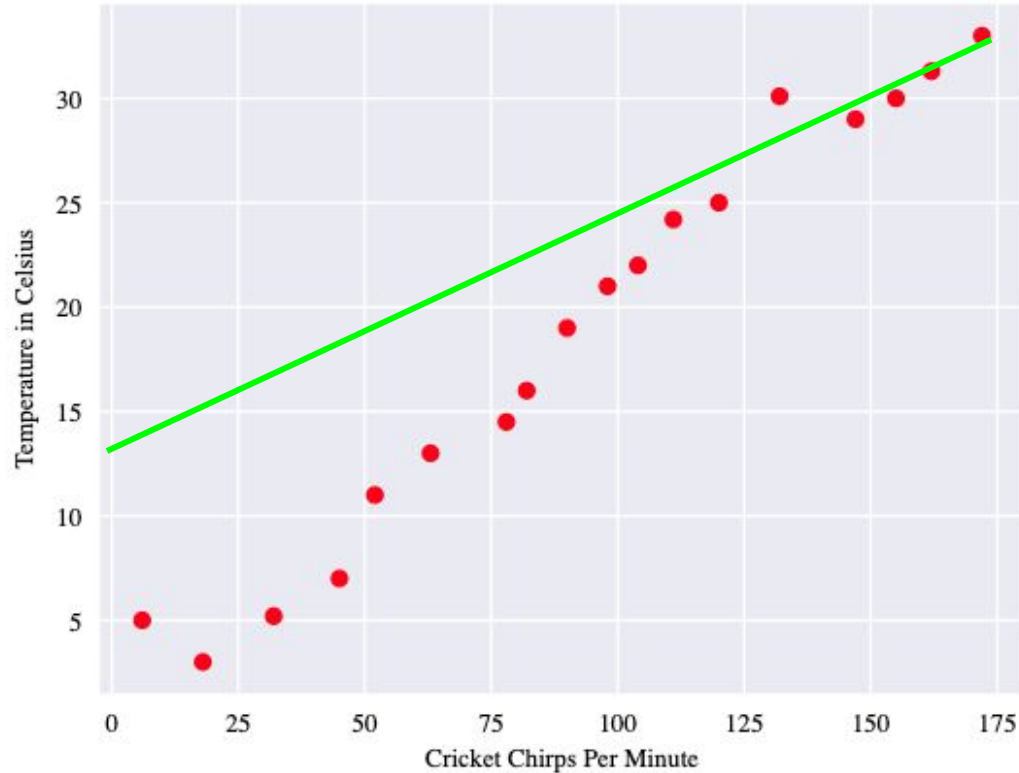
Training a Linear Regression Model



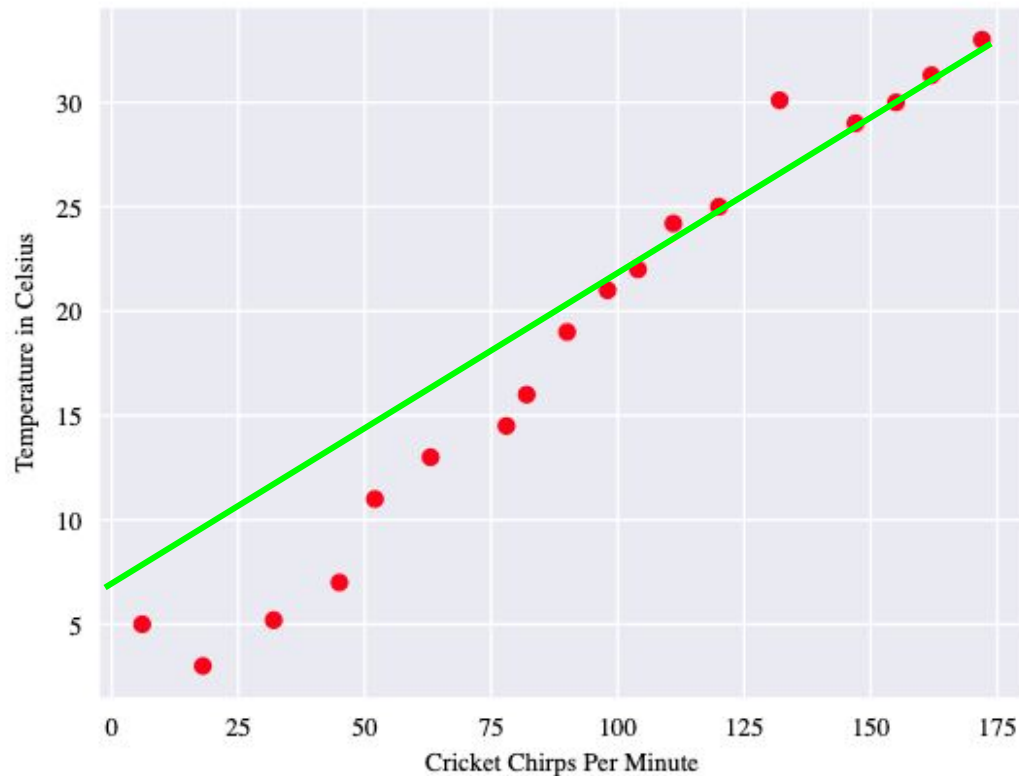
Training a Linear Regression Model



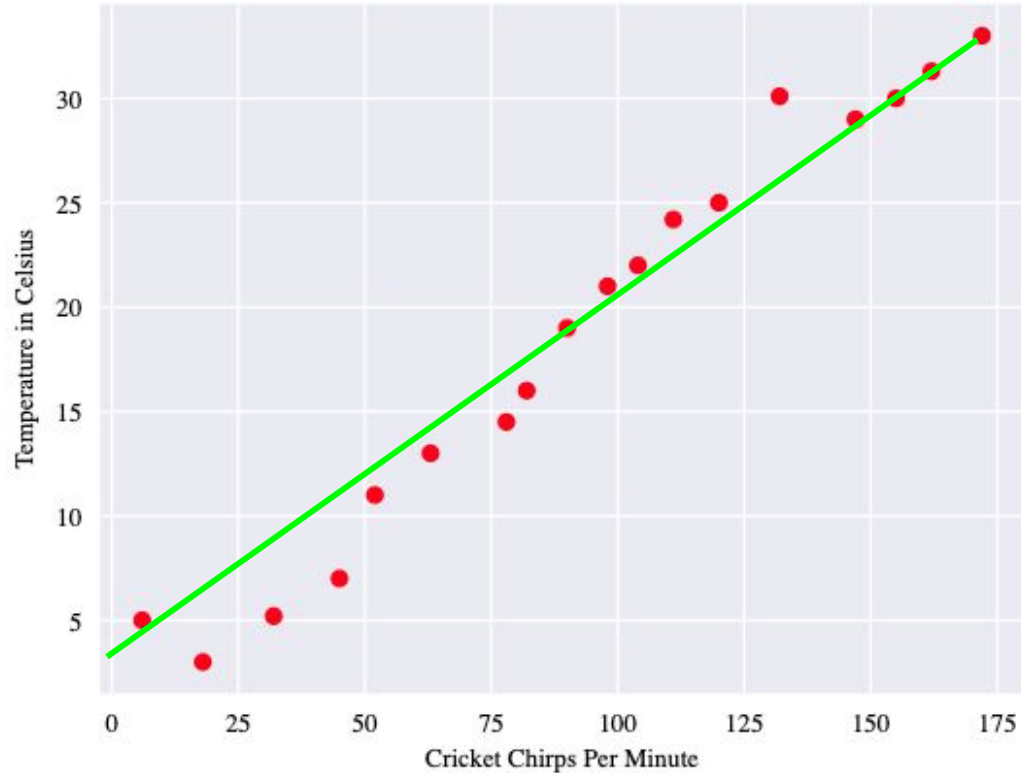
Training a Linear Regression Model



Training a Linear Regression Model



Training a Linear Regression Model



Hyperparameters: Learning Rate

- How fast should the model learn?
- How much should you change your weights by at each step?
- Ex: Predicting housing price from median income

W_1, B

Very low loss

0

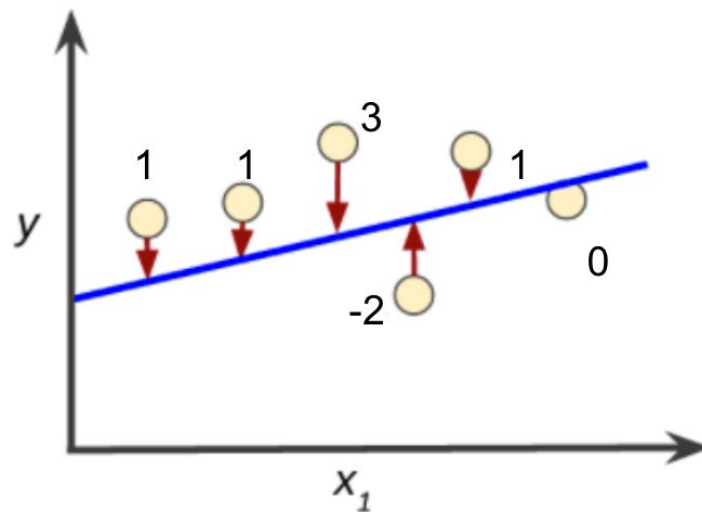
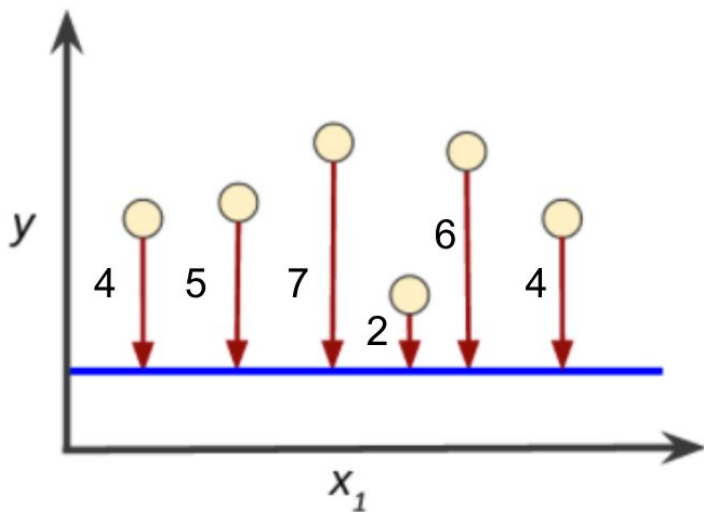
1000

Hyperparameters: Learning Rate

- Learning rate too small
 - Finding the right model takes a long time
- Learning rate too large
 - You might not find the right model because it keeps jumping over the ideal values

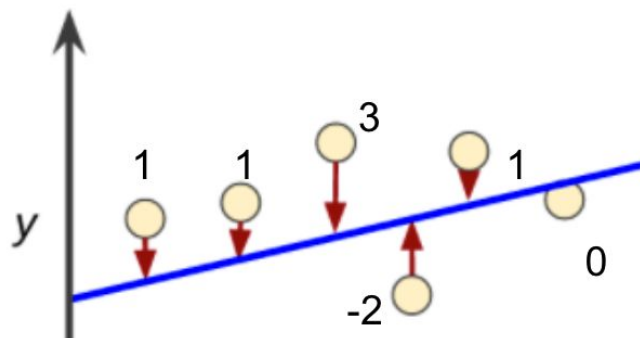
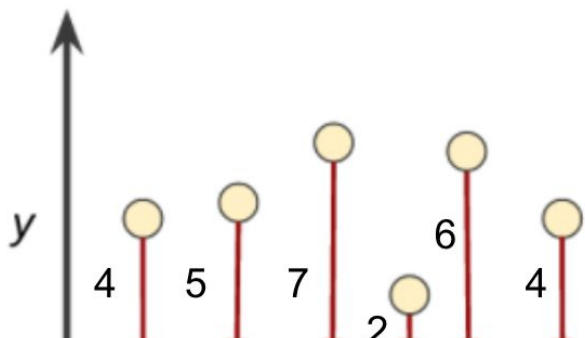
Hyperparameters: Batch size

- How many data points are we going to use to compute the loss of the model?
 - Low batch size -> Compute loss faster, less accurate result for loss
 - High batch size -> Longer to compute loss, more accurate loss result



Hyperparameters: Epoch Number

- How many times are we going to show the model all the data?
 - Low epoch size -> Compute loss faster, less accurate result for loss
 - High epoch size -> Longer to compute loss, more accurate loss result
-
- 100 rows, batch size to 25
 - 4 iterations



Summary...

- In linear regression, we want to find a line that “best fits” the data points
 - By a line we mean a bias and weight values
- A good line/model is one with a low loss
- We find a good line by iteratively tweaking our line and recalculating loss until we're happy with how low our loss is
- Hyperparameters
 - Learning rate: How much should we change the weight values at each step?
 - Batch size: How many points should we use to calculate the loss?
 - Epoch: How many times should we train the model on all the data?

Coming up next week...

The code!

Offline support: [playing with hyperparameters](#)

Who can get the lowest MSE based on tuning the hyperparameters?

Week 6: End to end regression

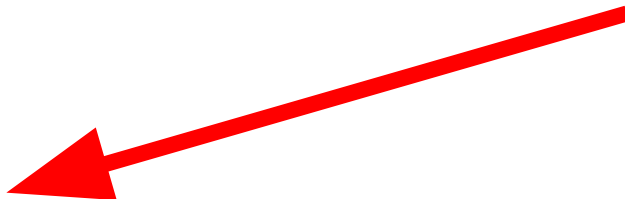


Goals

- Perform an entire ML workflow on a dataset :D

ML Project Checklist

1. Understand the problem
2. Get the data
3. Explore the data
4. Prepare the Data
5. Explore different models
6. Fine tune models
7. Present
8. Maintain



In a nutshell...

ML systems learn

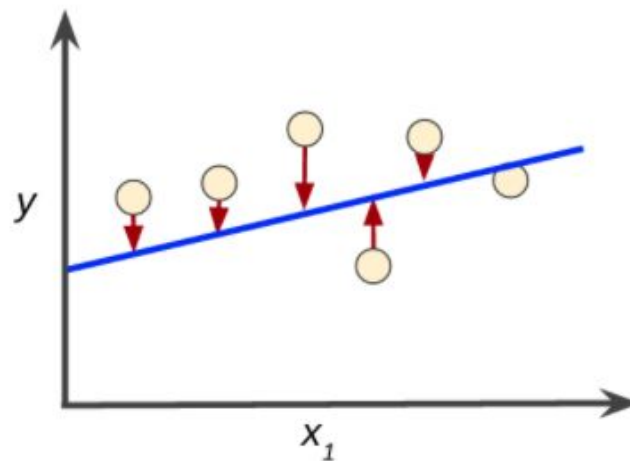
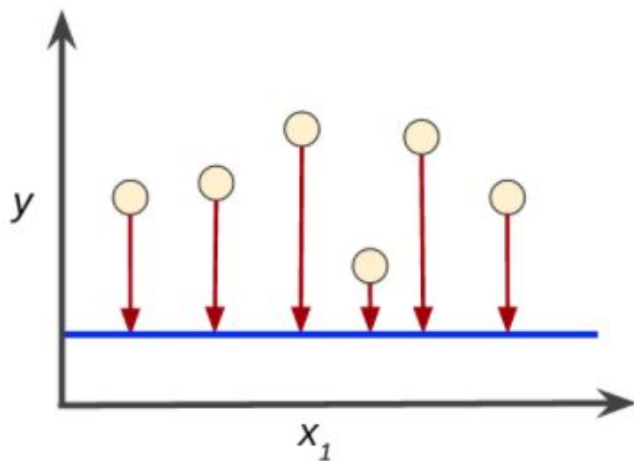
How to combine input

To produce useful predictions

On never-before-seen-data

Building linear regression intuition...

- **Training** a model simply means learning (determining) good values for all the weights and the bias from labeled examples.
- The goal of training a model is to find a set of weights and biases that have *low* loss, on average, across all examples



Improvement

- More data (preferable)
- More complex model (decision tree, neural net)
- Better feature selection

Looking ahead

- Next week: classification lecture
- 2 weeks: End to end classification

Week 7: Intro to Classification



Goals

- Understand classification problems from a high level
- Extend the ideas from regression problems to classification problems.
- Gain familiarity with the MNIST dataset
- Measure accuracy of a binary classifier (accuracy vs. precision/recall)
- Describe the transition between binary classification and multiclass classification

In a nutshell...

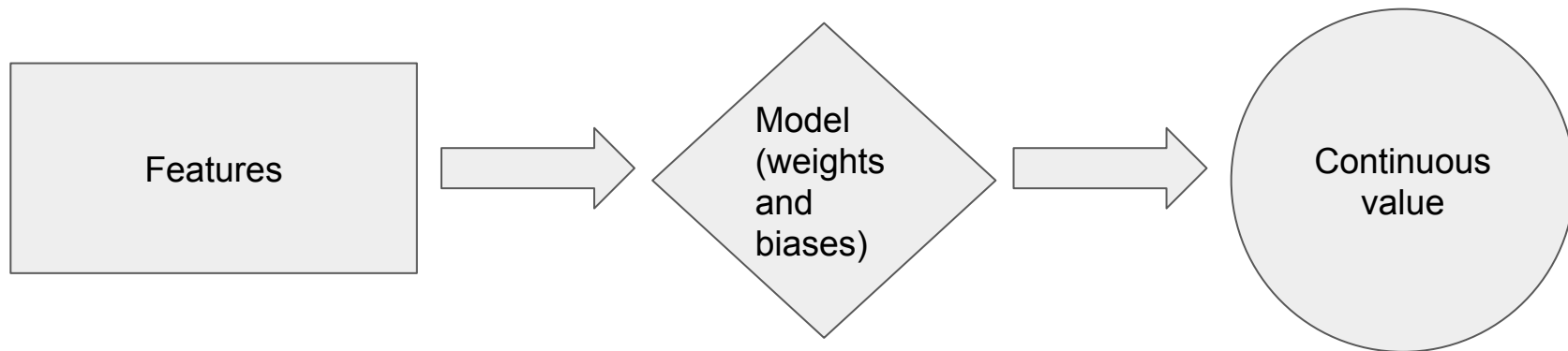
ML systems learn

How to combine input

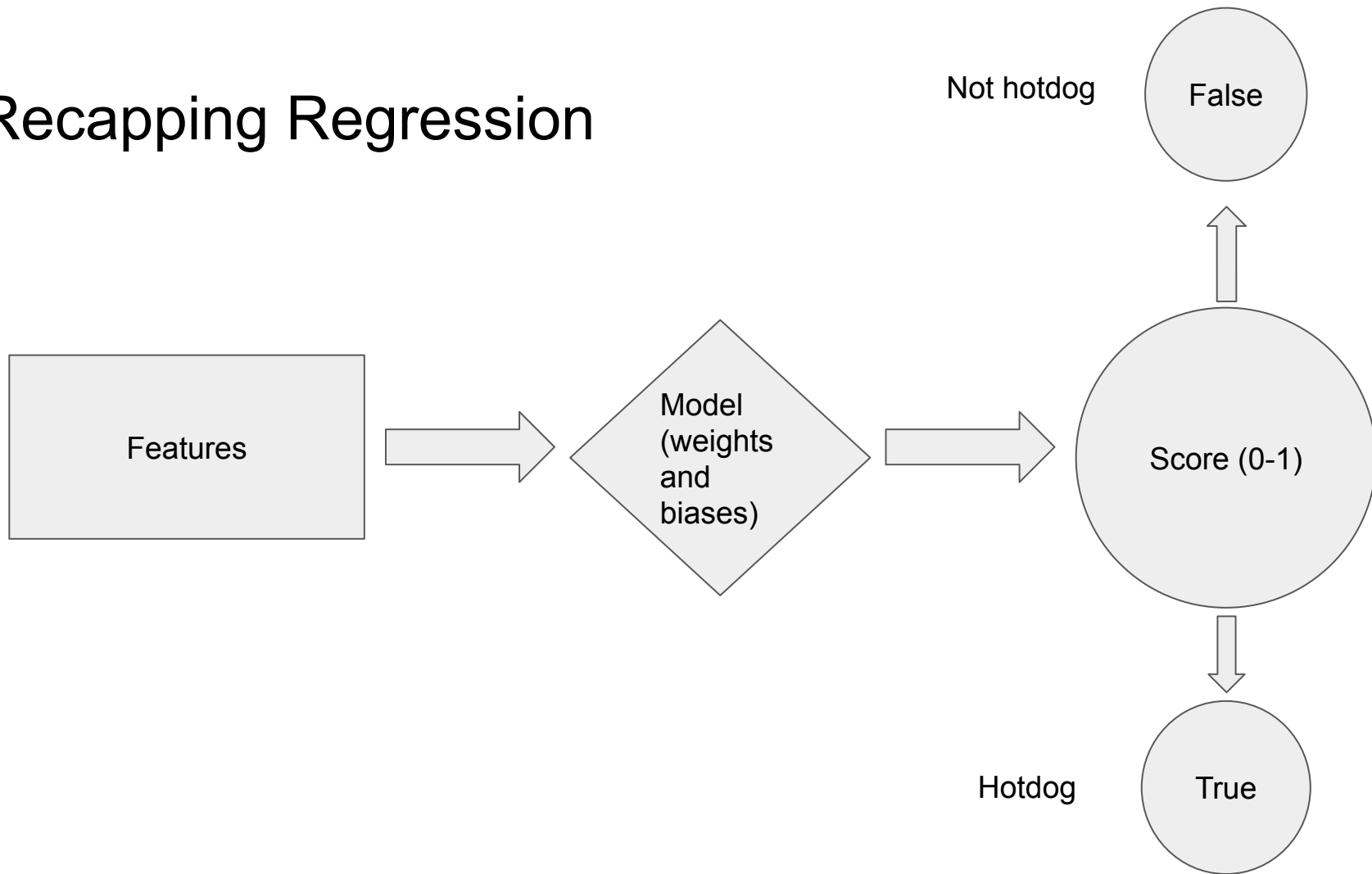
To produce useful predictions

On never-before-seen-data

Recapping Regression



Recapping Regression



Classification Score

Confident

Less confident

Confident

0

0.5

1



Vocab: Classification

- Class: output value we're trying to predict (1 or 0, True or False)
- Score: output of a classifier (on a scale of 0 to 1).
 - If <0.5 then 0
 - If >0.5 then 1
- Binary Classifier: Task of classifying elements of a dataset into two groups based on a classification rule.
- Multiclass Classifier: Task of classifying elements of a dataset into three or more groups based on a classification rule.

MNIST Dataset

- Collection of handwritten digits (0-9) that we want to classify to integers
- Hello world ML example
- When new algorithm developed, often times test performance on this classic dataset
- 70,000 images.
- 784 features
 - Image is 28x28pxl
 - Each feature has value 0-255 that represents intensity



MNIST Dataset

5	0	4	1	9	2	1	3	1	4
3	5	3	6	1	7	2	8	6	9
4	0	9	1	1	2	4	3	2	7
3	8	6	9	0	5	6	0	7	6
1	8	7	9	3	9	8	5	9	3
3	0	7	4	9	8	0	9	4	1
4	4	6	0	4	5	6	7	0	0
1	7	1	6	3	0	2	1	1	7
8	0	2	6	7	8	3	9	0	4
6	7	4	6	8	0	7	8	3	1

Classification: Training

- We can use the same linear model as regression tasks to train a classifier
- `sklearn.linear_model.SGDRegressor` vs. `sklearn.linear_model.SGDClassifier`
 - `fit(features, classes)`
 - `predict (features)`

Training a Linear Regression Model

Pick random values for slope (w), bias (b)

Make an initial prediction on the training data and compute loss

While you haven't found the best model

- Change slope and bias slightly and recompute loss

- If the loss is less than before:

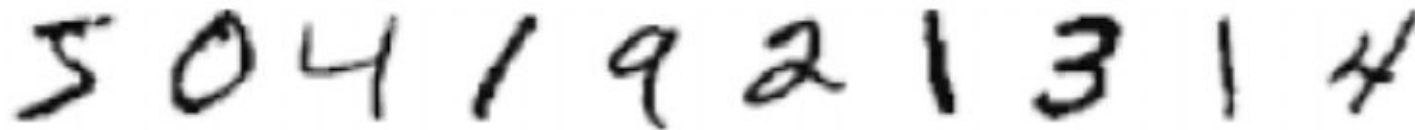
 - Keep moving w , b in that direction

- Else:

 - Move w , b in another direction

Classifier: Measuring Accuracy

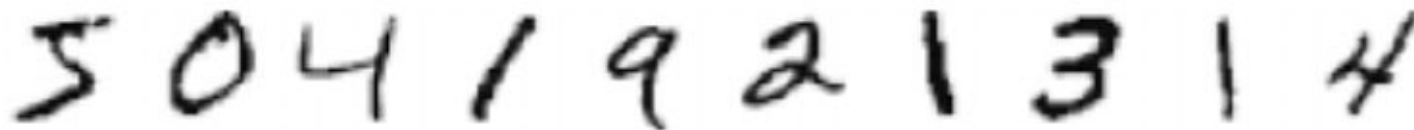
- 5 Classifier
 - Label data as either 5 or not 5
 - Train the model on the features and labels from the training set using LinearModel
 - Make predictions
 - Measure accuracy: (number of correct predictions/total predictions)

A row of ten handwritten digits: 5, 0, 4, 1, 9, 2, 1, 3, 1, 4. Each digit is on a separate background patch.

T F F F F F F F F T = 90%

Classifier: Measuring Accuracy

- Never 5 Classifier
 - Create a model that says no matter what, the image is not a 5
 - Measure accuracy: (number of correct predictions/total predictions)

A row of ten handwritten digits: 5, 0, 4, 1, 9, 2, 1, 3, 1, 4. Each digit is written in a dark, slightly noisy font on a light background.

F

F

F

F

F

F

F

F

F

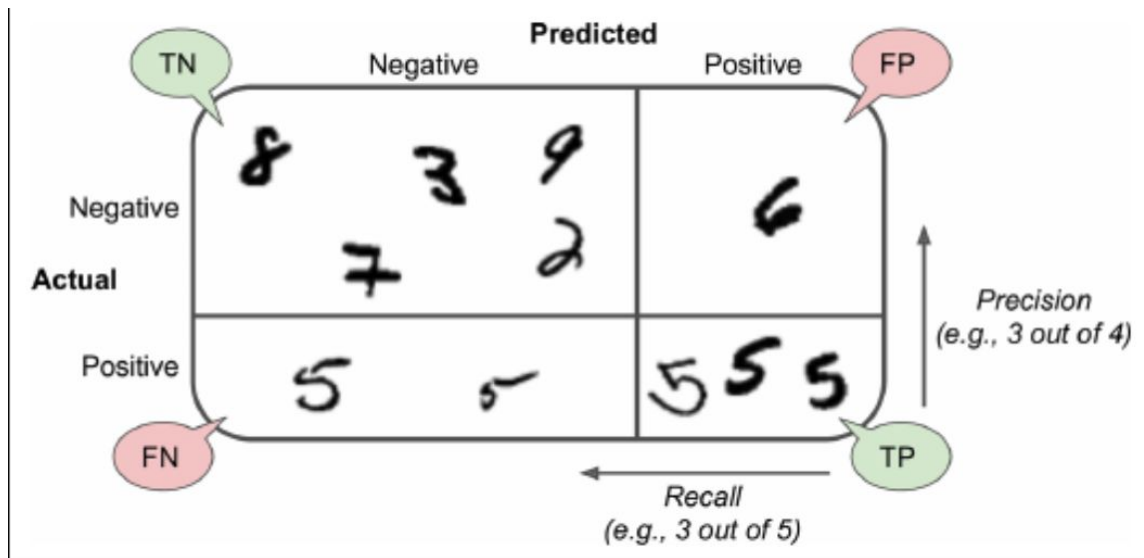
F = 90%

Classifiers: Measuring Accuracy

- Accuracy is a bad measure of performance when some classes are much more frequent than others in a dataset!!!

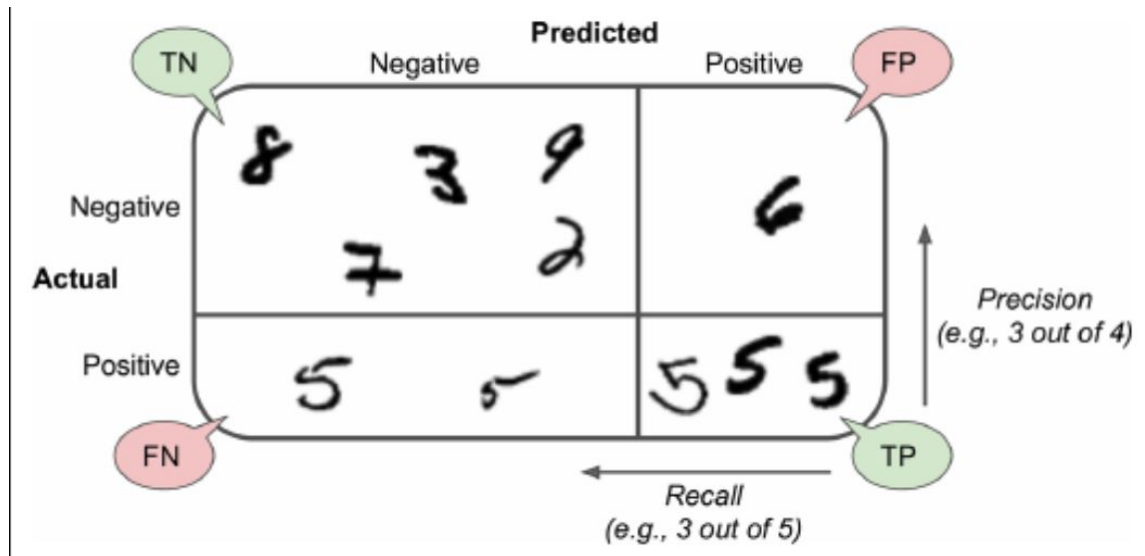
Classifiers: Measuring Accuracy

	Predicted Not 5	Predicted 5
Actually not 5	5, true negative	1, false positive
Actually 5	2, false negative	3, true positive



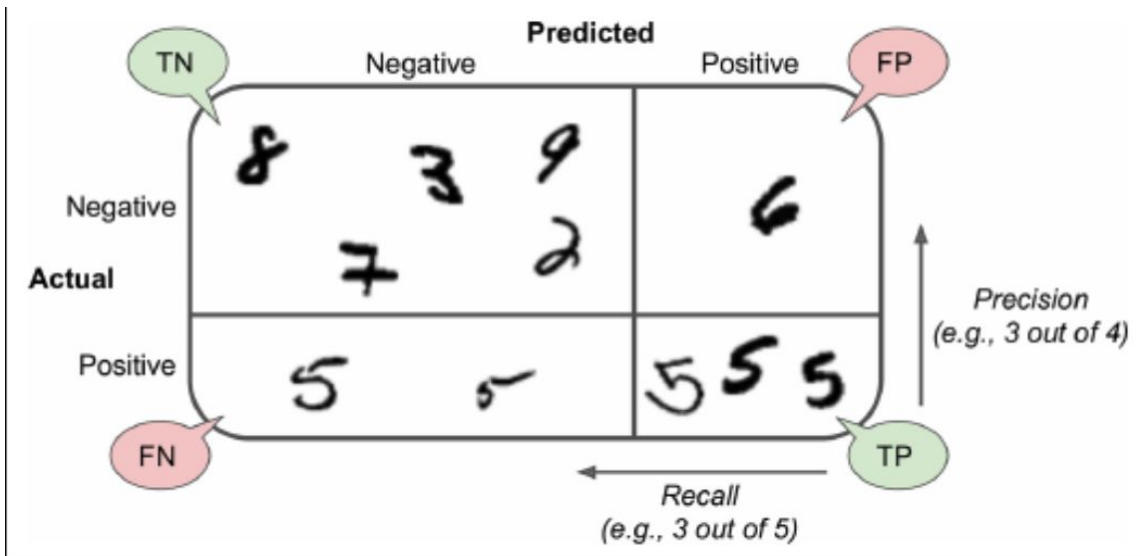
Classifiers: Measuring Accuracy

- Precision:
 - Accuracy of positive predictions. Ratio of predicted 5s that are actually 5s
 - $(TP)/(TP+FP)$



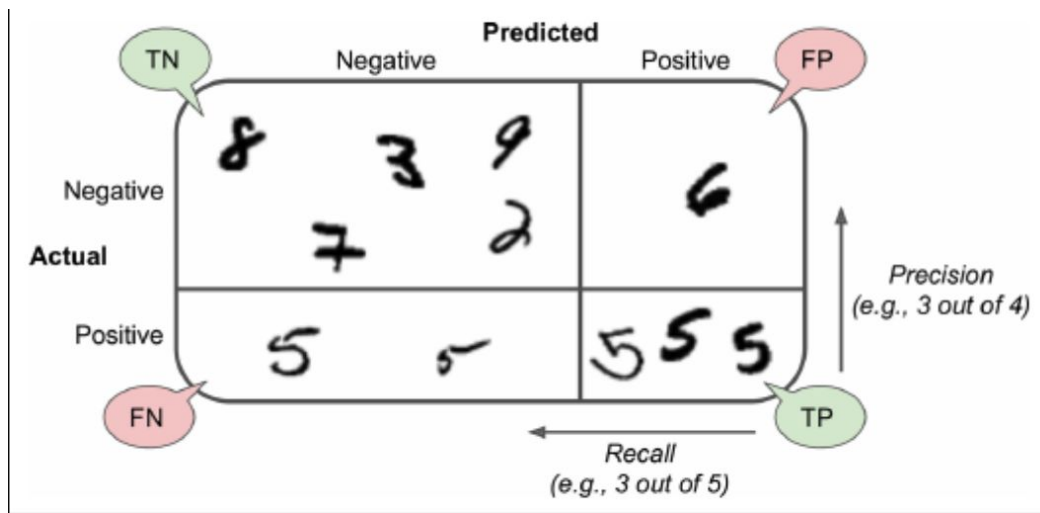
Classifiers: Measuring Accuracy

- Recall:
 - Ratio of positive instances that are correctly classified. How many true 5s are classified correctly?
 - $(TP)/(TP+FN)$

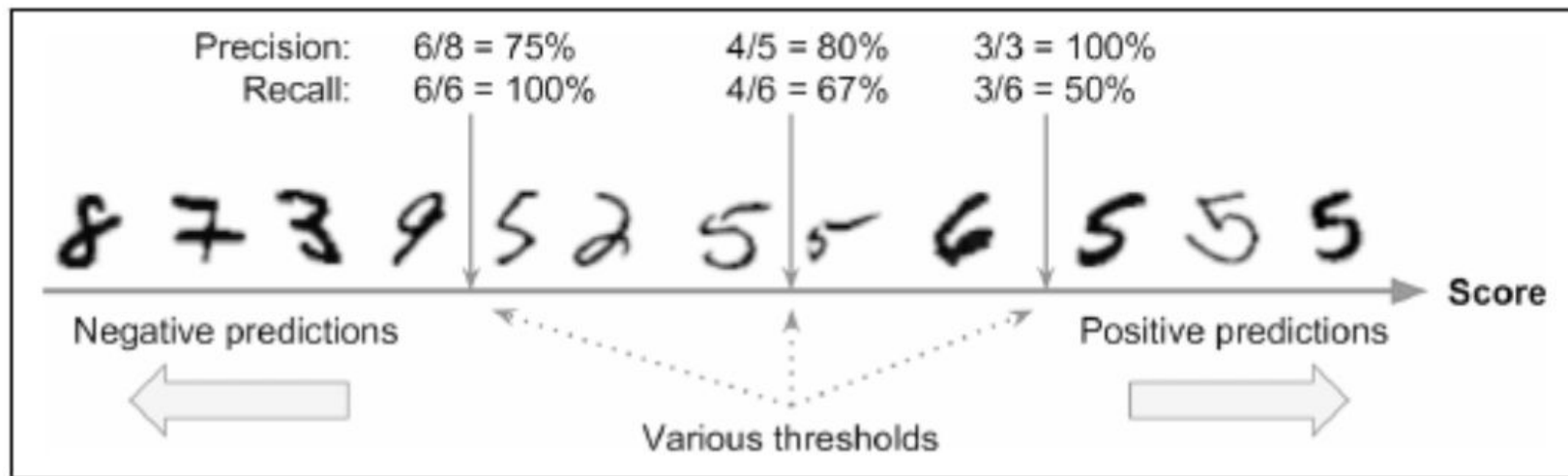


Precision vs. Recall

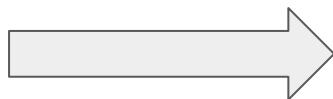
- Accuracy of positive predictions
- How many predicted 5s were actually 5s?
- $TP/(TP+FP) = 3/4$
- Ratio of positive instances classified correctly.
- How many true 5s were classified correctly?
- $TP/(TP+FN) = 3/5$



Precision vs. Recall



Multiclass Output



5	0	4	1	9	2	1	3	1	4
3	5	3	6	1	7	2	8	6	9
4	0	9	1	1	2	4	3	2	7
3	8	6	9	0	5	6	0	7	6
1	8	7	9	3	9	8	5	9	3
3	0	7	4	9	8	0	9	4	1
4	4	6	0	4	5	6	1	0	0
1	7	1	6	3	0	2	1	1	7
9	0	2	6	7	8	3	9	0	4
6	7	4	6	8	0	7	8	3	1

Multiclass Output: OvR

- One versus the rest
- Create a separate 5 or not 5 classifier for every class
 - 0 or not 0, 1 or not 1, ... , 9 or not 9
- For each prediction, pick the class whose classifier has highest positive score
- Must train each classifier on entire dataset (can be slow)

Multiclass Output: OvO

- One versus one duel for each class!
- Train a binary classifier for each pair of classes
 - 0 vs 1, 0 vs 2, 0 vs 3, ... 9 vs. 8, ... etc
 - $N \times (N - 1) / 2$ classifiers
- To classify, run an observation through all classifiers to see which class wins the most duels
- Don't have to train each classifier on all the data.
- Sometimes faster to train many classifiers on small sets than few classifiers on big sets.

Recap

- Features -> model -> score -> binary class
- We can train the model using the the same linear regression
- We measure the accuracy of a model using a confusion matrix
 - Precision vs. recall tradeoff
- We can train multiclass classifiers using many binary classifiers
 - OvO and OvR

Next week

- Coding a classification model
- Class recap and resources to keep learning

Big Data Bootcamp: Last week!

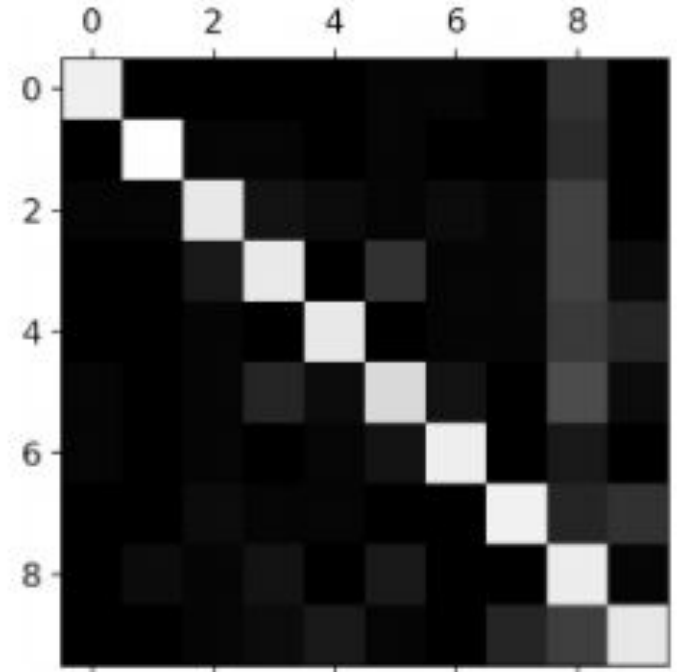
5	0	4	1	9	2	1	3	1	4
3	5	3	6	1	7	2	8	6	9
4	0	9	1	1	2	4	3	2	7
3	8	6	9	0	5	6	0	7	6
1	8	1	9	3	9	8	5	9	3
3	0	7	4	9	8	0	9	4	1
4	4	6	0	4	5	6	1	0	0
1	7	1	6	3	0	2	1	1	7
8	0	2	6	7	8	3	9	0	4
6	7	4	6	8	0	7	8	3	1

Goals

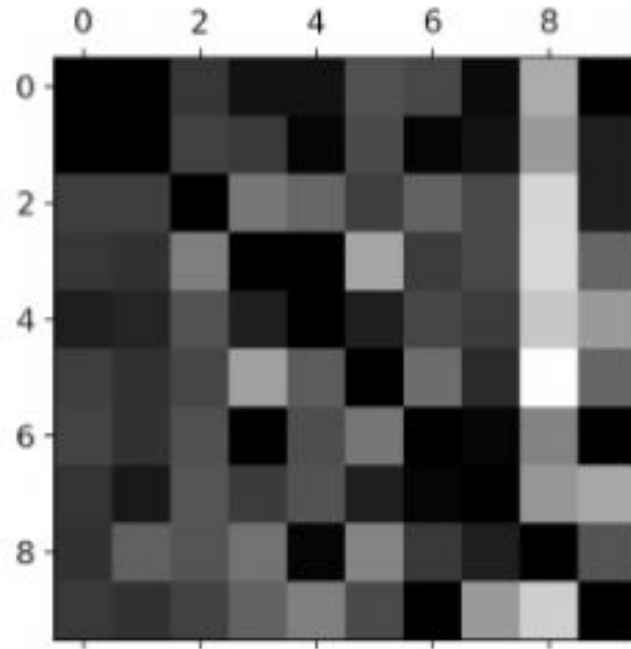
- Train a model to classify the handwritten digits in the MNIST dataset.
- Recap and next steps :)

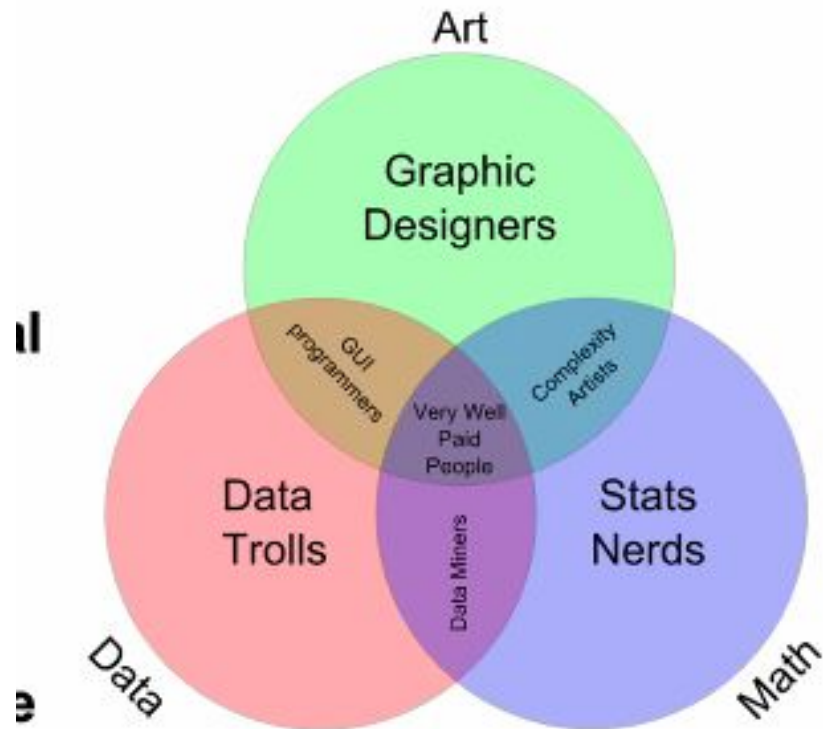
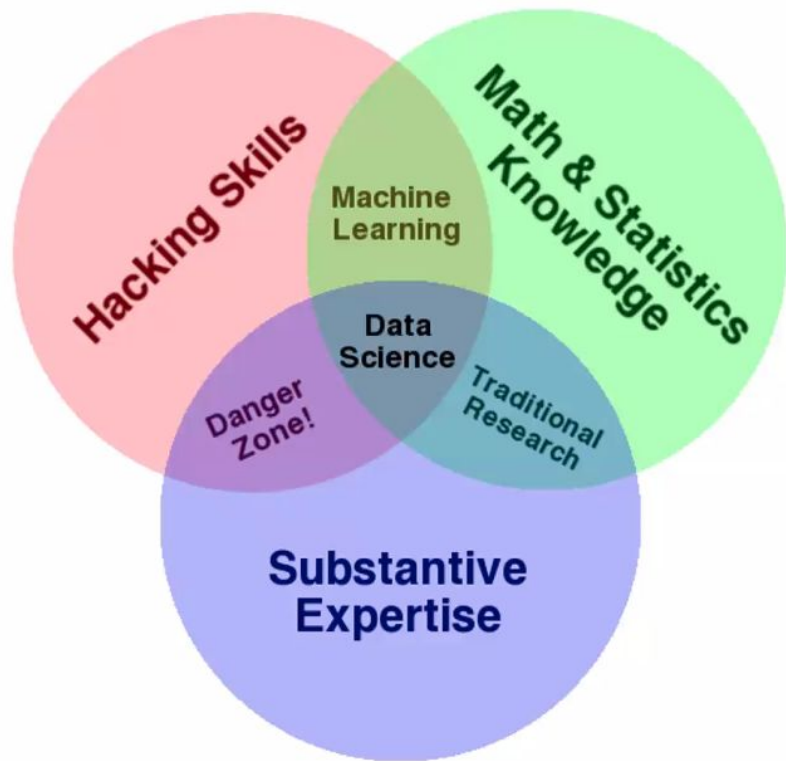
Rows: Actual classes, Cols: Predicted Classes

```
array([[5578, 0, 22, 7, 8, 45, 35, 5, 222, 1],  
       [ 0, 6410, 35, 26, 4, 44, 4, 8, 198, 13],  
       [ 28, 27, 5232, 100, 74, 27, 68, 37, 354, 11],  
       [ 23, 18, 115, 5254, 2, 209, 26, 38, 373, 73],  
       [ 11, 14, 45, 12, 5219, 11, 33, 26, 299, 172],  
       [ 26, 16, 31, 173, 54, 4484, 76, 14, 482, 65],  
       [ 31, 17, 45, 2, 42, 98, 5556, 3, 123, 1],  
       [ 20, 10, 53, 27, 50, 13, 3, 5696, 173, 220],  
       [ 17, 64, 47, 91, 3, 125, 24, 11, 5421, 48],  
       [ 24, 18, 29, 67, 116, 39, 1, 174, 329, 5152]])
```



Rows: Actual Classes, Cols: Predicted Classes





Looking Back

1. Intro to Python/Computational Thinking
2. Important libraries (matplotlib, pandas, numpy)
3. Web scraping
4. Overview of ML
5. Regression Theory
6. Regression Ex.
7. Classification Theory
8. Classification Example

Communities

- [Towards Data Science](#)
- [r/learnpython](#)
- [r/MachineLearning](#)
- [r/LearnMachineLearning](#)
- [r/DataScience](#)
- MAGIC!

Resources

- [Andrew Ng Stanford ML Course Youtube](#)
- [Andrew Ng ML Course Coursera](#)
- [Hands on ML with Sklearn, Keras, and Tensorflow](#)
- [Learn Python the Hard Way](#)
- [Deep Learning with Python](#)

That's all folks :)

<https://www.linkedin.com/in/stschoberg>

stschoberg@gmail.com

Slack, etc.

Stay tuned for post class survey!