Demystifying Data Science: What the Heck is Machine Learning?

•••

August 23, 2017

Z. W. Miller

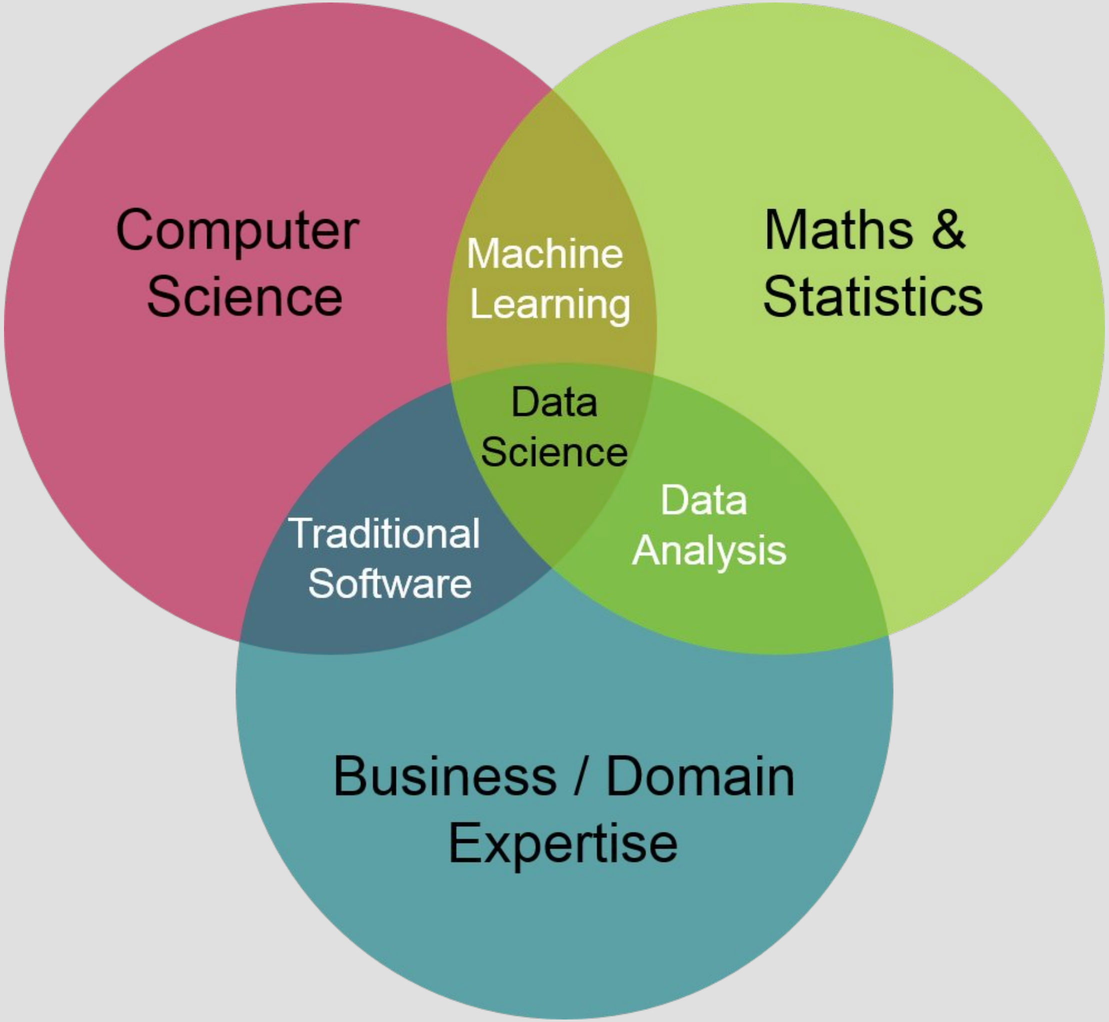
Who am I?

- Recovering nuclear physicist (PhD)
- Data management and analysis junkie
- Educator: physics, math, computer science, astronomy, data science
- Senior data scientist at Metis

- Data science and machine learning educators
- 12 week intensive bootcamps
- Part-time live-online courses
- Corporate training

METIS

What is a data scientist?


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise

Data scientists are generally...

- Organizing and aggregating data
- Analyzing that data to try to find patterns
- Building pipelines to handle incoming data
- Converting data into insights

Data scientists are generally...

- Organizing and aggregating data
- Analyzing that data to try to find patterns
- Building pipelines to handle incoming data
- Converting data into insights

We’ll focus on these sections today.

Why is big data a problem?

Which number is missing?

0 1 2 3 4 5 7 8 9

Which number is missing?

0 1 2 3 4 5 7 8 9

6

Which number is missing?

3 0 2 8 5 1 6 9 7

Which number is missing?

3 0 2 8 5 1 6 9 7

4

Which number is missing?

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

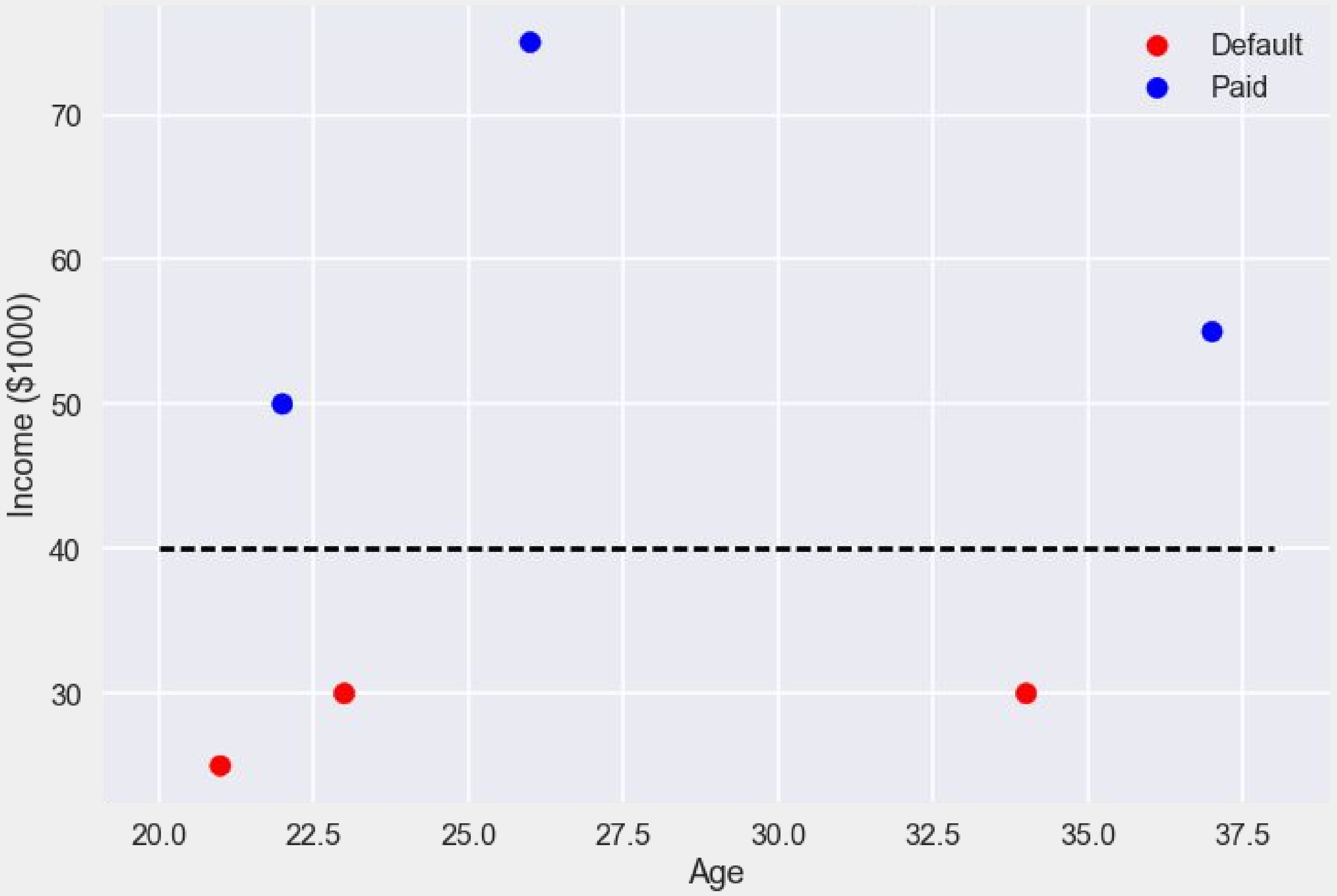
Which number is missing?

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

5

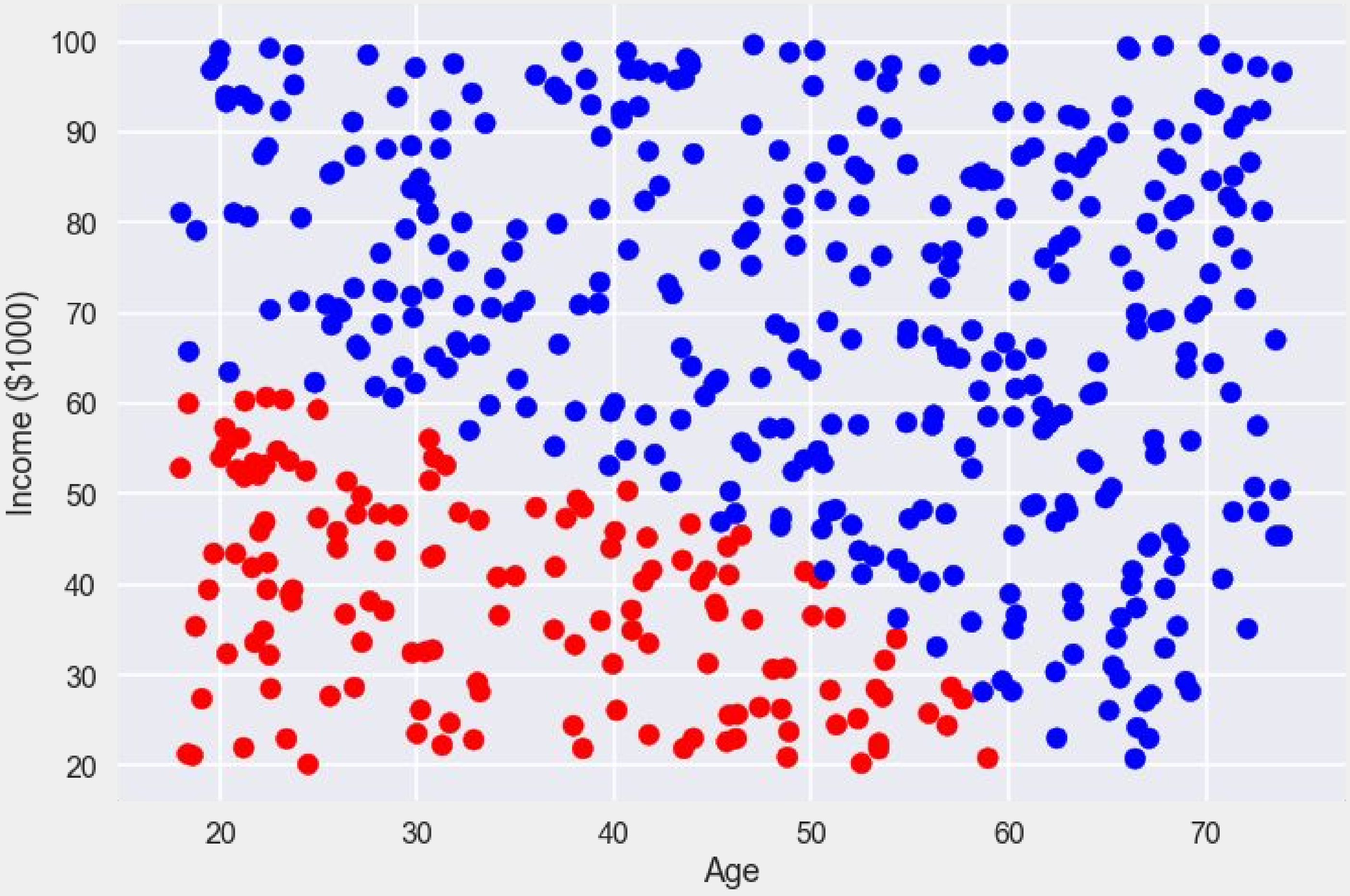
Let’s look at some more realistic examples

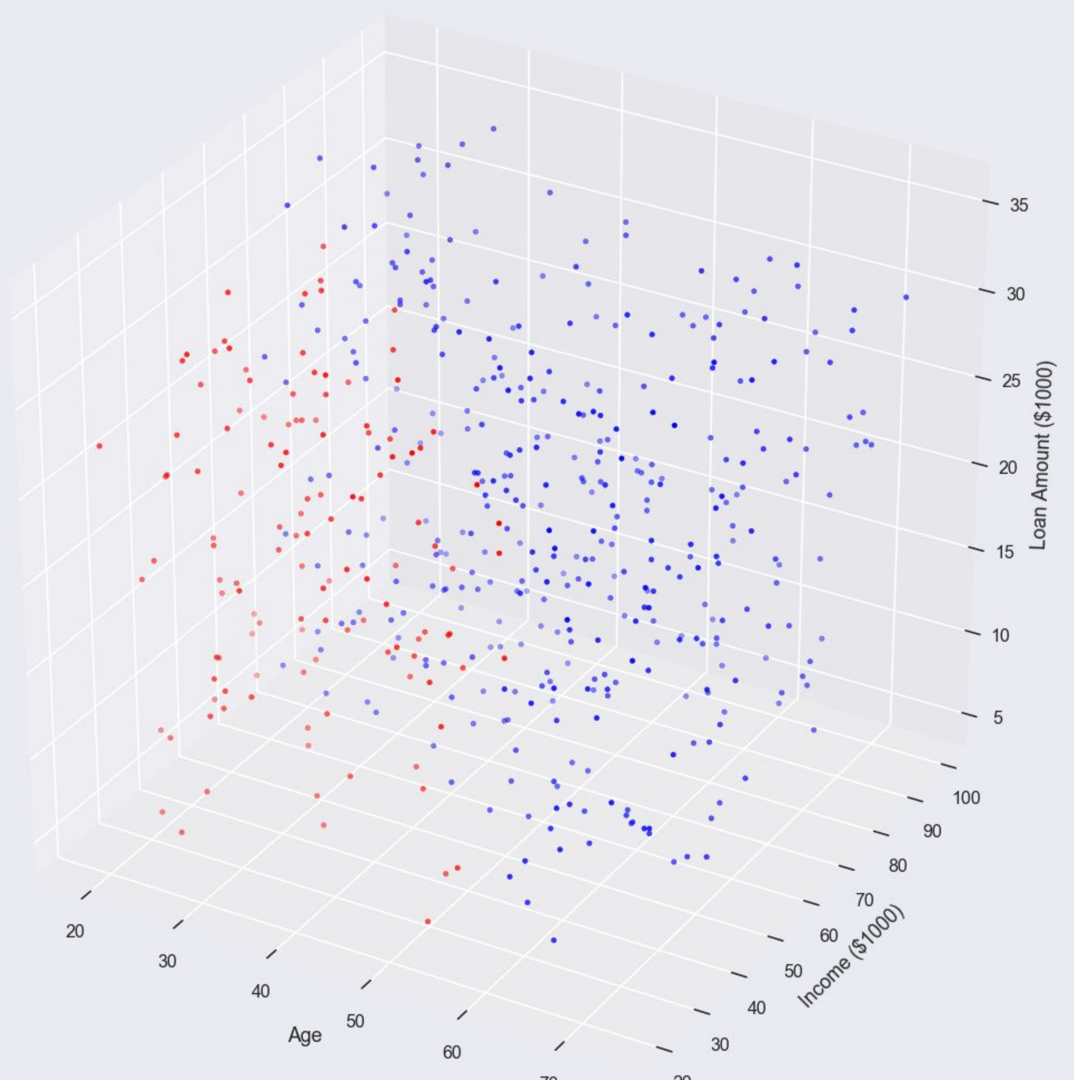
Age	Salary	Loan Amount	Paid Back
21	25K	5K	N
23	30K	5K	N
34	30K	5K	N
22	50K	5K	Y
37	55K	5K	Y
26	75K	5K	Y



Default
Paid
70
60
50
40
30
Income ($1000)
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5
Age

Age	Salary (K)	Loan Amount (K)	Paid Back
21	25	5	N
23	30	5	N
34	30	5	N
22	50	5	Y
37	55	5	Y
26	75	5	Y
43	51	10	Y
57	43	20	N
21	23	3	N
23	35	7	Y
22	16	65	N
39	110	25	Y
36	93	17	N
45	130	10	Y


Income ($1000)
100
90
80
70
60
50
40
30
20
20
30
40
50
60
70
Age


Loan Amount ($1000)
5
10
15
20
25
30
35
Age
20
30
40
50
60
Income ($1000)
30
40
50
60
70
80
90
100

Our data can get too wide as well

If we're making a decision about a loan applicant, we aren't going to look at just age, income, and loan amount. We'll also want to look at things like:

Number of previous loans,
percentage of previous loans re-payed,
credit score,
savings account size,
current job,
how long at current job,
married or single,
value of assets,
number of current loans,
own a house,

Our data can get too wide as well

If we're making a decision about a loan applicant, we aren't going to look at just age, income, and loan amount. We'll also want to look at things like:

Number of previous loans,
percentage of previous loans re-payed,
credit score,
savings account size,
current job,
how long at current job,
married or single,
value of assets,
current loans,
number of current loans
own a house,
cosigner,
spouse income,
education level,
current debt level

Our data can get too wide as well

If we’re making a decision about a loan applicant, we aren’t going to look at just age, income, and loan amount. We’ll also want to look at things like:

Number of previous loans,
percentage of previous loans re-payed,
credit score,
savings account size,
current job,
how long at current job,
married or single,
value of assets,
current loans,
number of current loans
own a house,
cosigner,
spouse income,
education level,
current debt level,
value of current loans,
children,
geographic location,
employment history,
age,
Income,
loan amount

What is learning?

What is machine learning?

Let’s define learning for our purposes…

Learning is not about memorization and recollection. It is about generalizing conclusions to previously unseen examples.

Supervised

vs

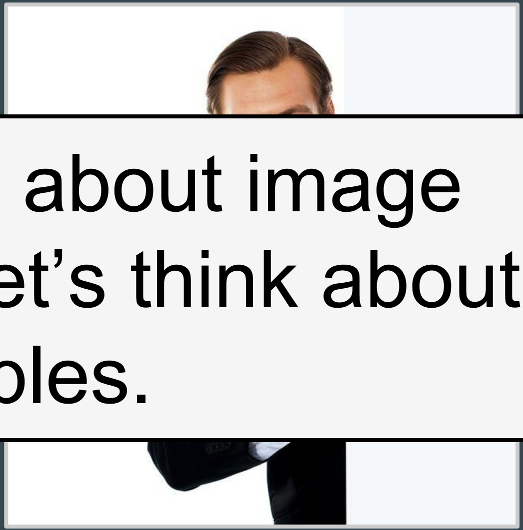
Unsupervised

Supervised Learning

DOG

NOT DOG

Supervised Learning



Don’t worry, we’ll talk a little about image recognition later. For now, let’s think about more “standard” data examples.

DOG

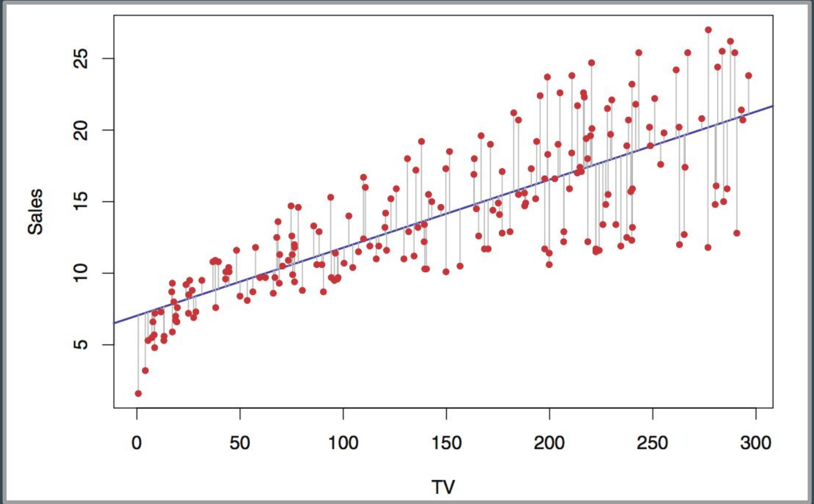
NOT DOG

Supervised Learning

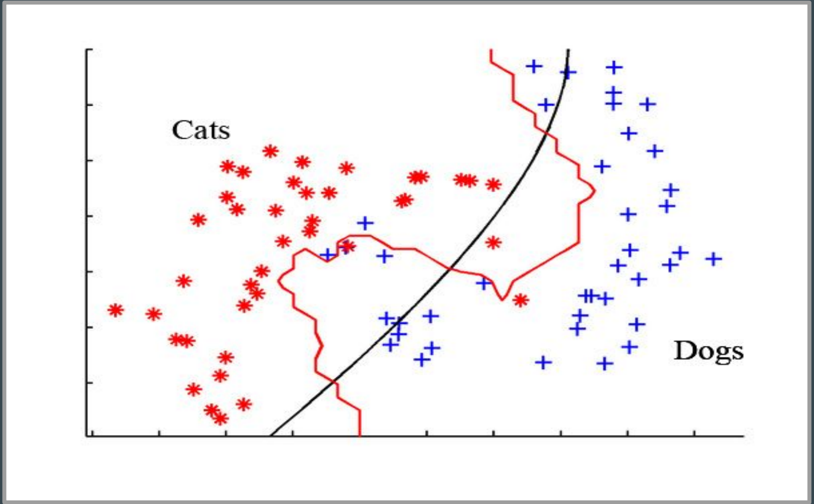
- We have “truth” about each datapoint - and some sort of history to build from.
- We want to build a model that learns about all the history.
- We want to make **predictions** going forward.

2 Most Common Flavors for Supervised Learning

Regression



Classification



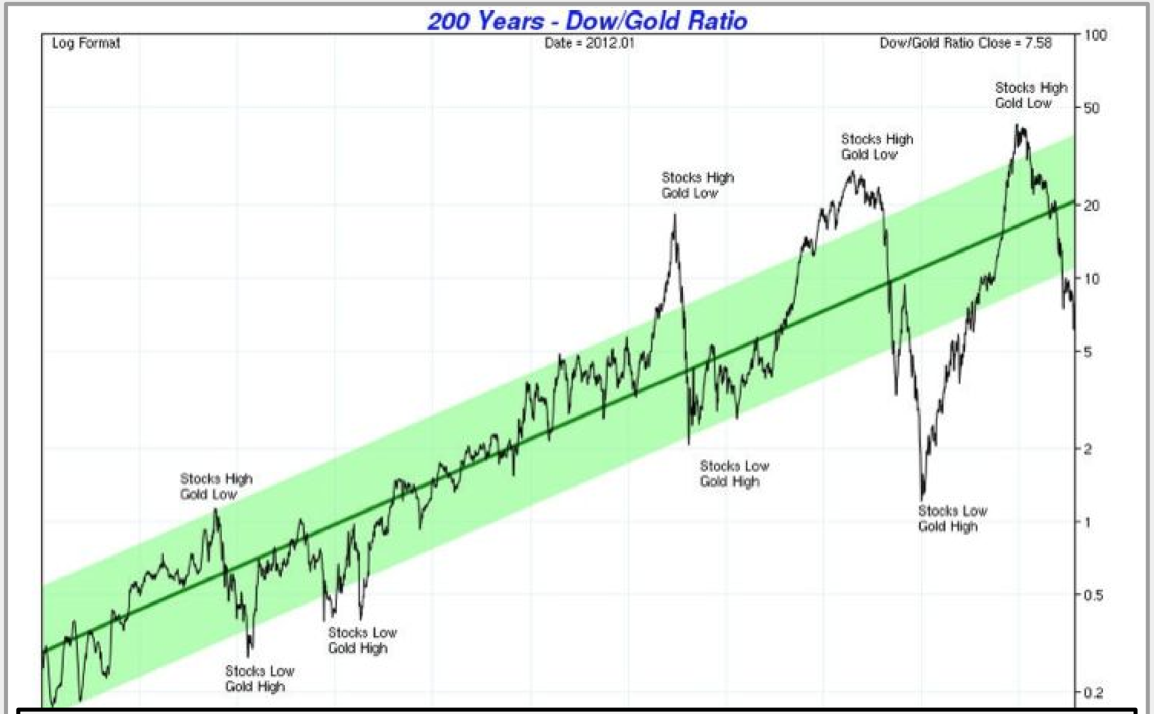
Regression

- We feed in numeric data and want the machine to understand the trends in the data.
- We ask the machine to build a model that accounts for those trends
- We use that trend to make predictions

Predicting Housing Markets

Predicting the Stock Market

Predicting the Stock Market

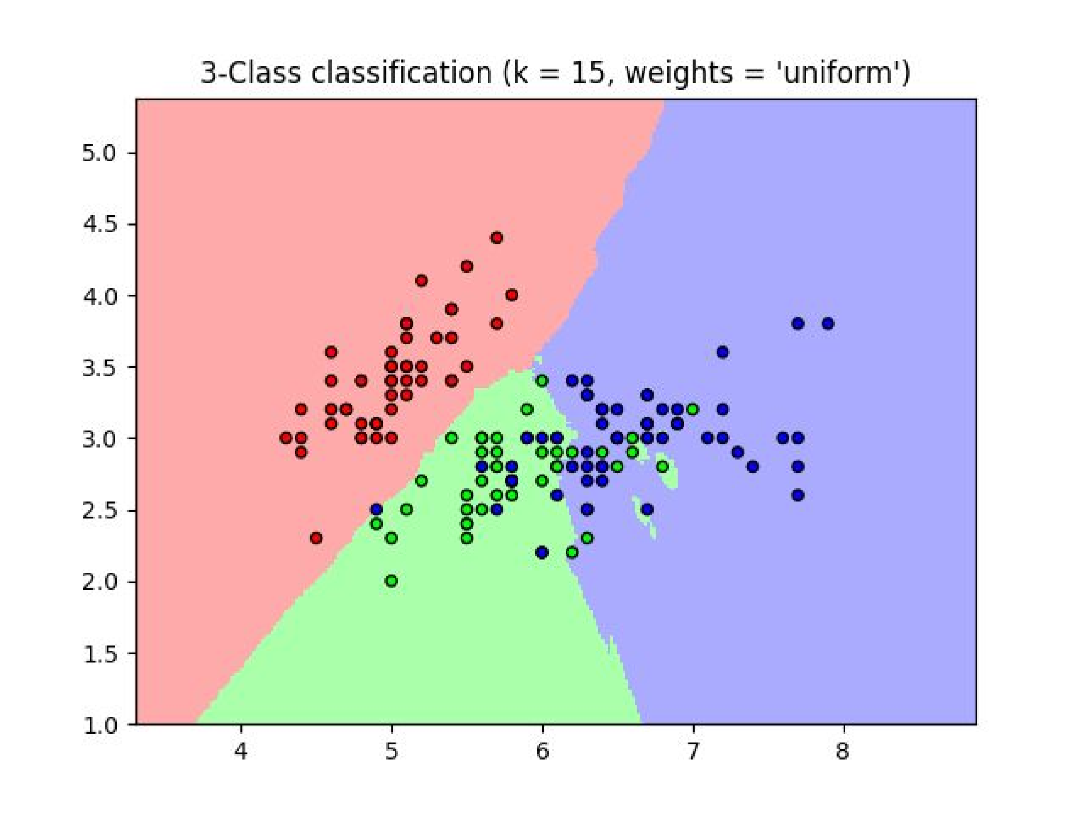


Please don't try this. You'll fail

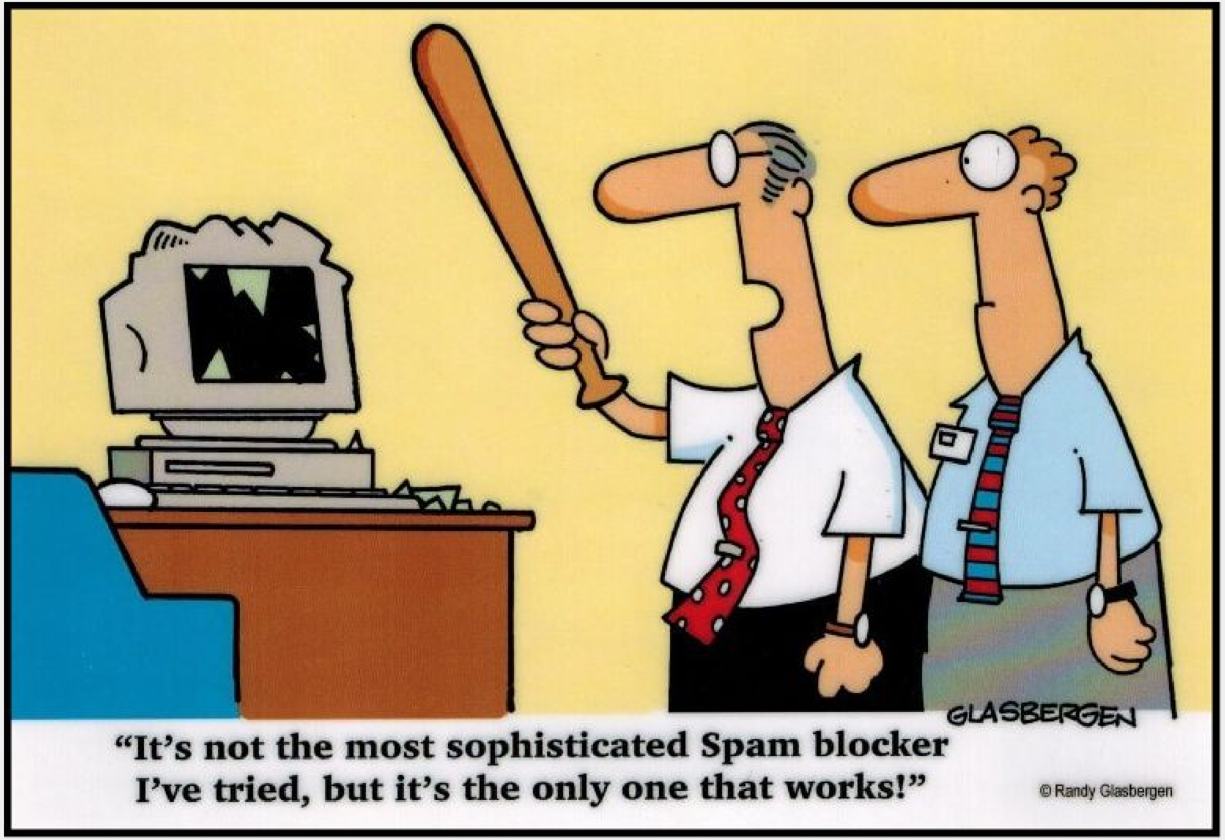
Classification

- We feed in data to the machine and ask it, which type of thing do you think this is?
- We ask the machine to draw lines in the sand that decide, “left of here, you’re type A. Down from here is type B. etc…”
- We can add new data points and decide based on those lines what type of thing we have.

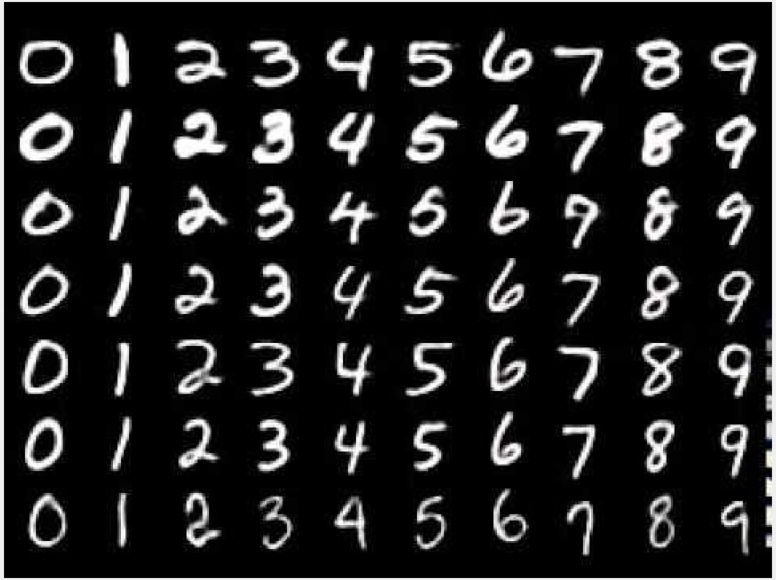
What type of flower?



Spam Classification



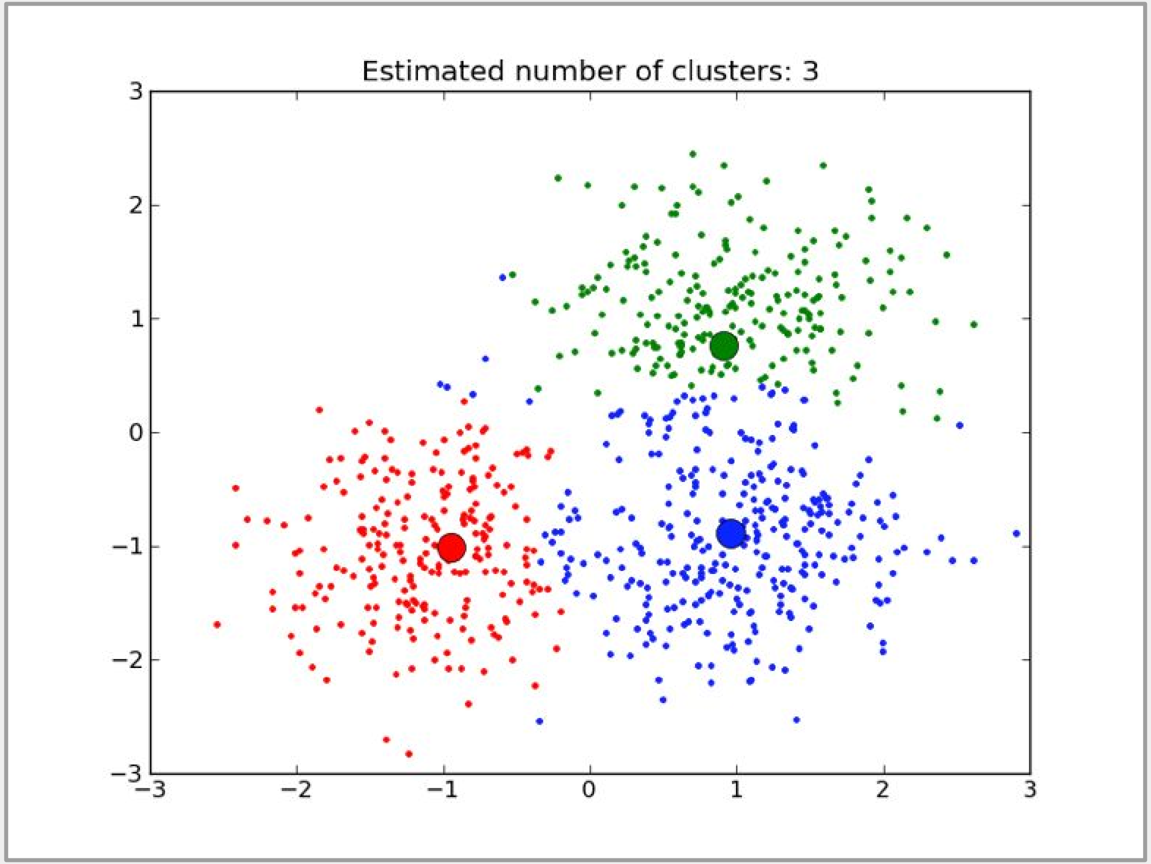
Optical Character Recognition



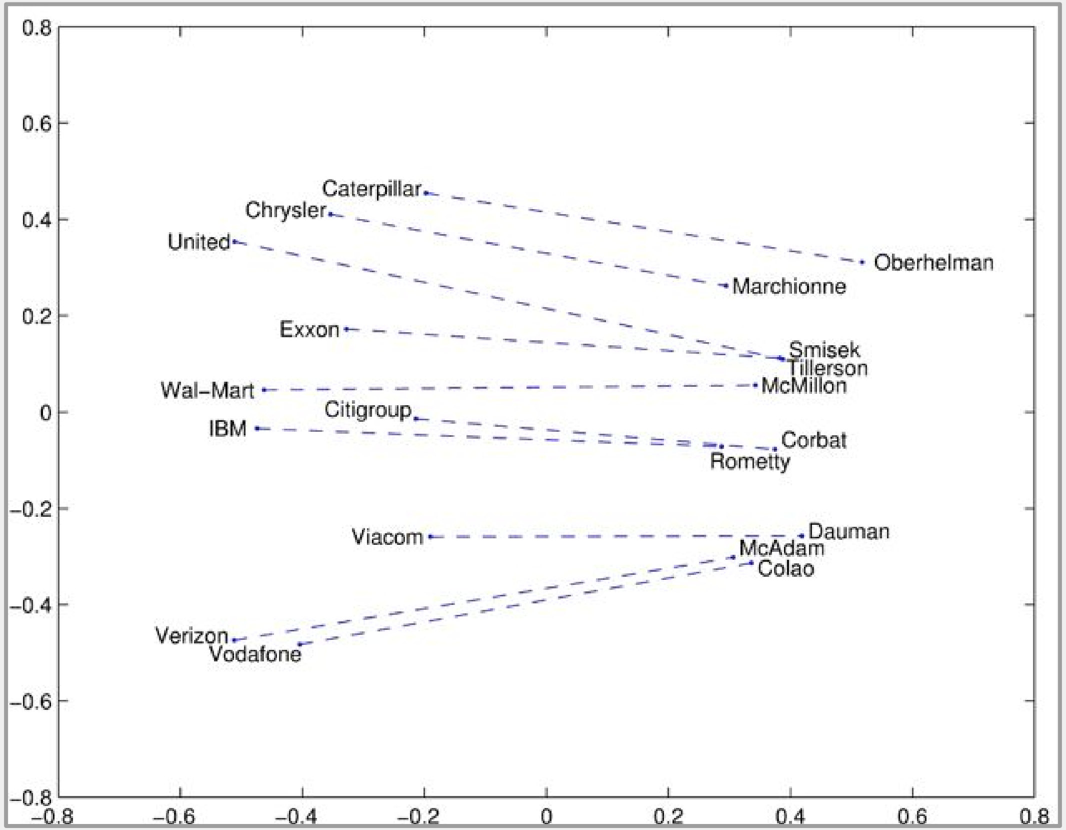
Unsupervised Learning

- There is no truth. We just want to find structure in the data.
- Our models don't make predictions, they look for **patterns.**

Clustering



Natural Language Processing



How does machine learning work, at it’s core?

- At the end of the day, we choose some value to optimize
- We write some clever code to allow the machine to do it for us
- We use the resulting output to make future decisions

So wait, is regression really “machine learning?”

Yep. Let’s talk about regression…

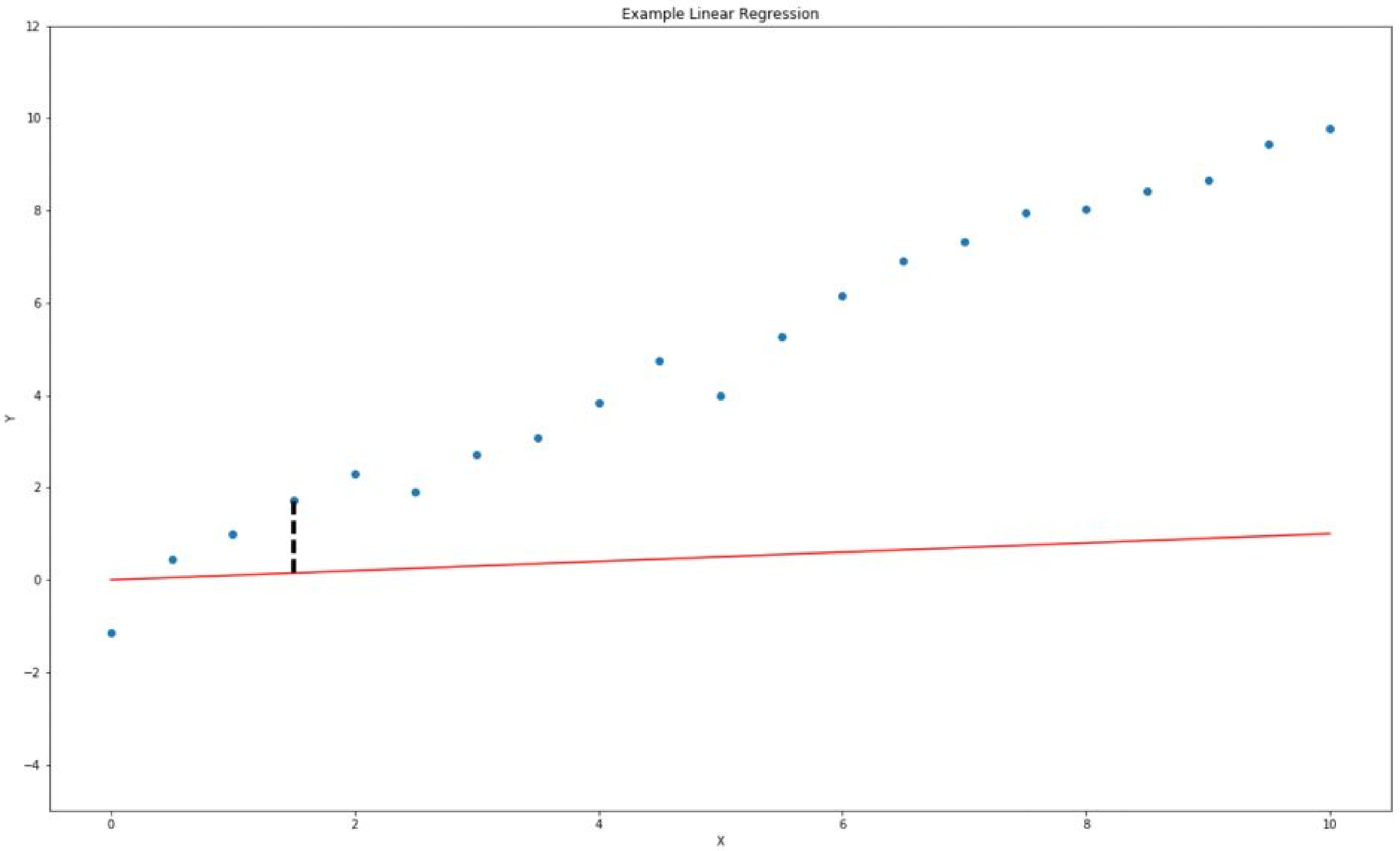
Regression

- What are we optimizing:

  Being as close to all the points as possible (smallest "errors")
- How are we clever:

  Guess and check in a smart way
- What do we get:

  A model we can use for prediction


Example Linear Regression
12
10
8
6
4
2
0
−2
−4
Y
0
2
4
6
8
10
X

How about clustering...?

Clustering

- What are we optimizing:

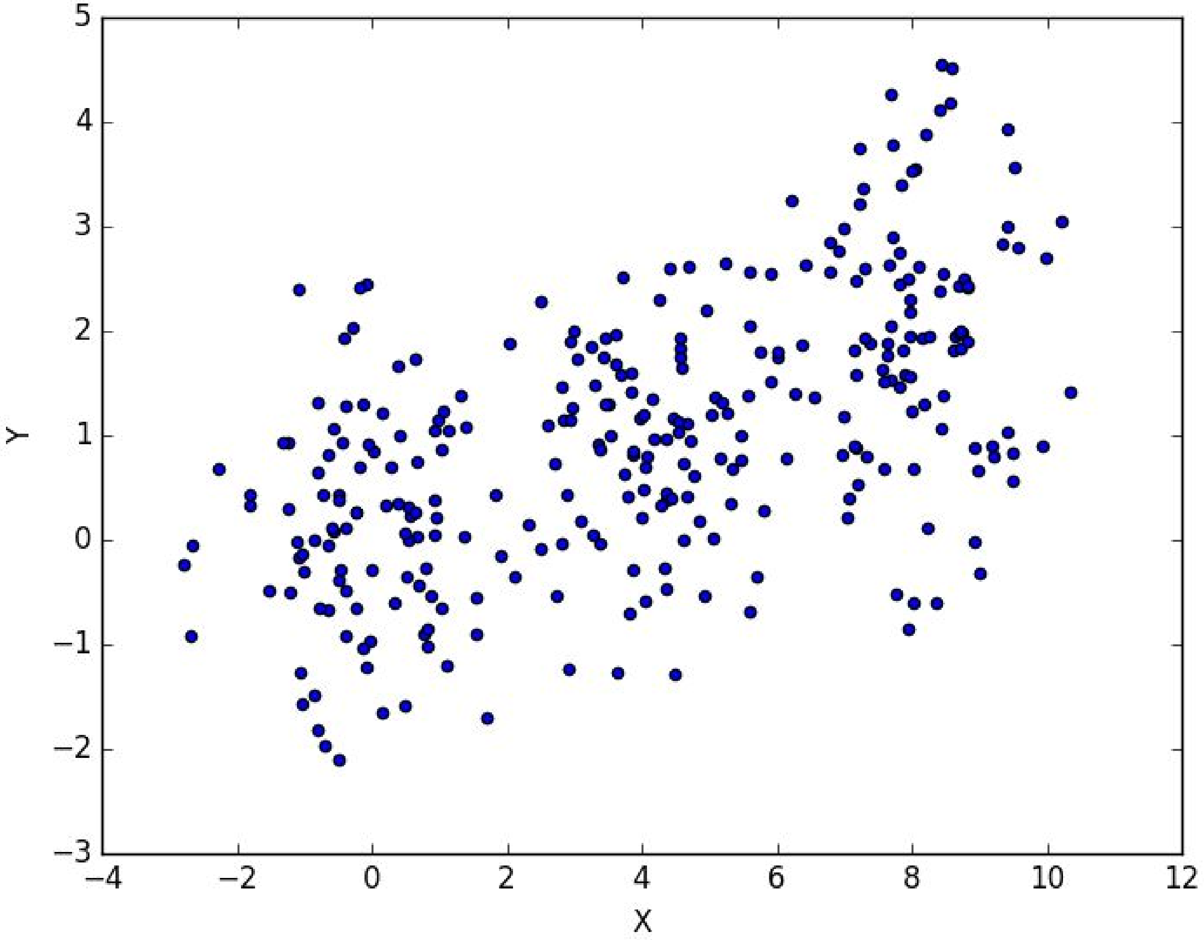
  How “grouped up” are all the points

- How are we clever:

  We always move towards more grouped points (high density)

- What do we get:

  A divided space we can use for segmentation



5
4
3
2
1
0
−1
−2
−3
Y
−4
−2
0
2
4
6
8
10
12
X

And classification?

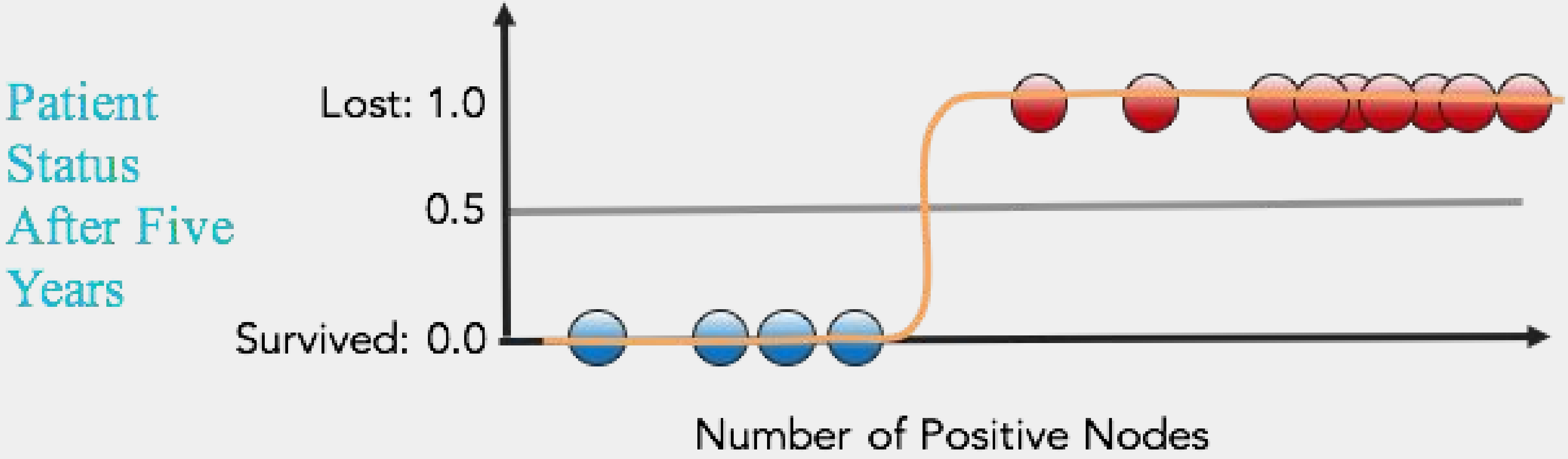
Classification

- What are we optimizing:

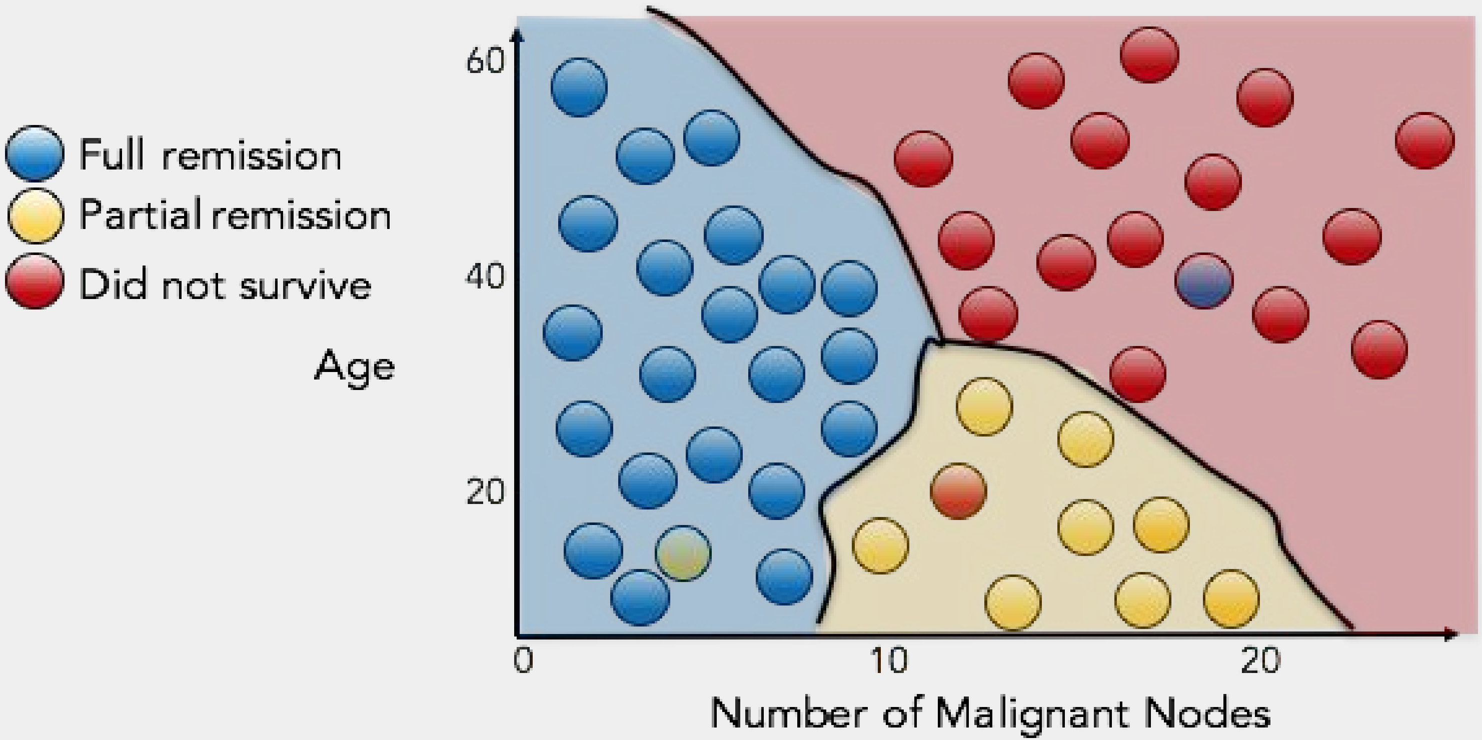
    Some measure of accuracy… am I identifying the classes correctly?
- How are we clever:

    Depends on the type of model - could be building decision trees, could be guess and check intelligently
- What do we get:

    Segmented space and probabilities for class ID

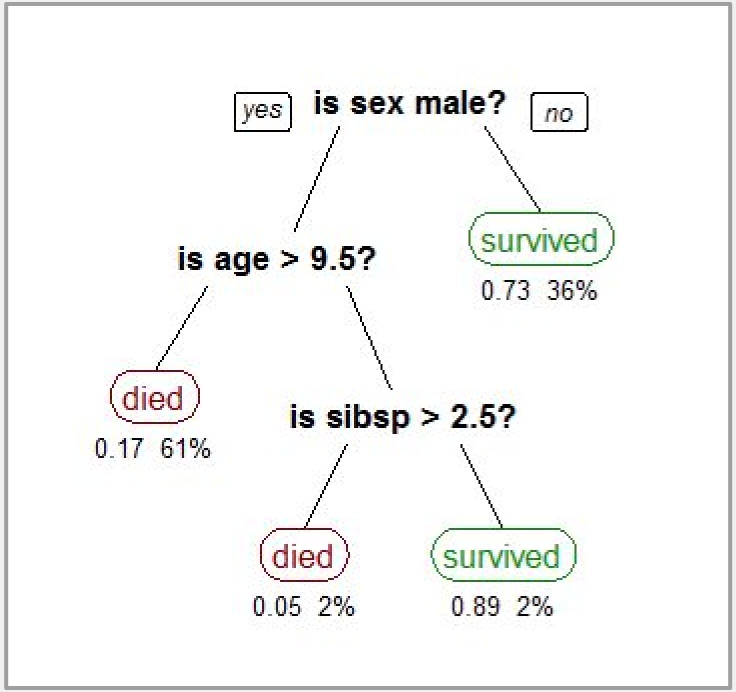


A simple way to handle classification is to just map a “regression” into a decision by using a mathematical trick.



Full remission
Partial remission
Did not survive
Age
60
40
20
0
10
20
Number of Malignant Nodes

Another way is to build these trees by optimizing how much we learn at each step.

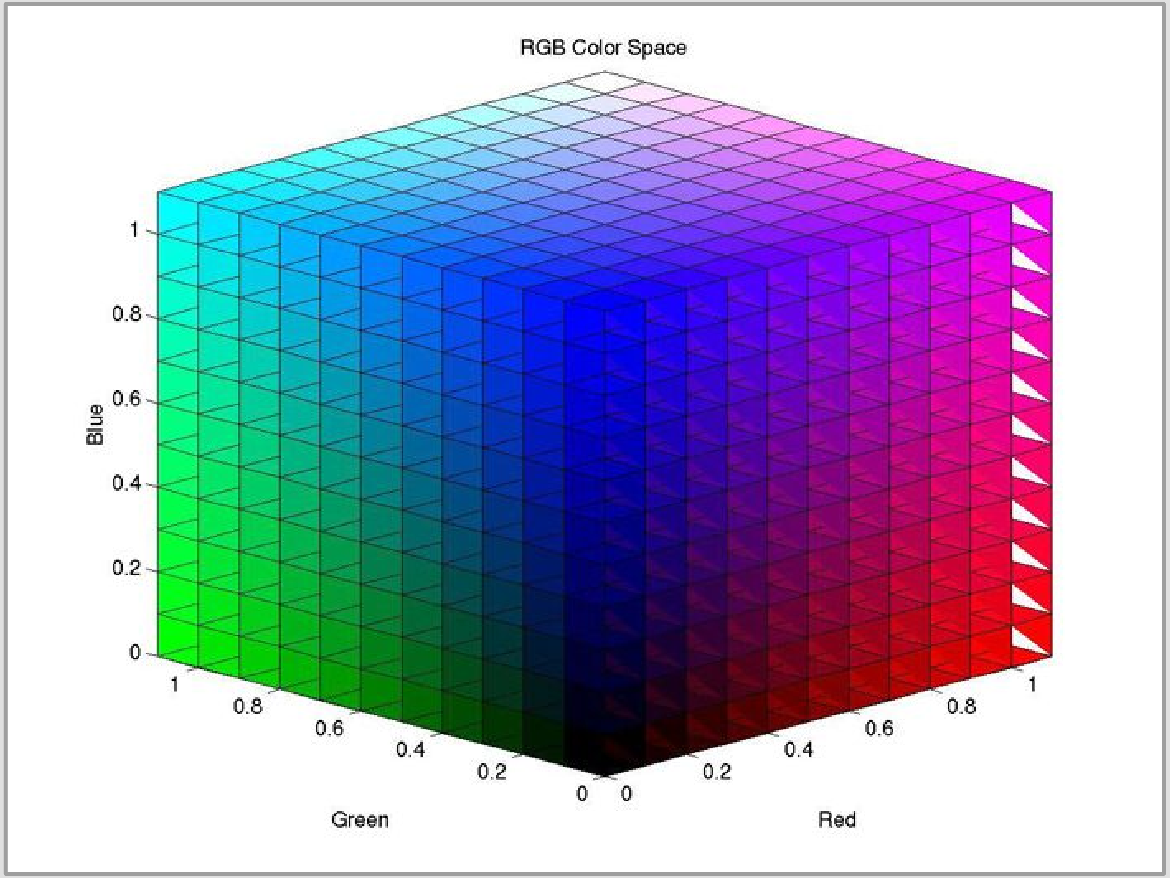


Ok, great.

Could you actually show us some examples in practice?

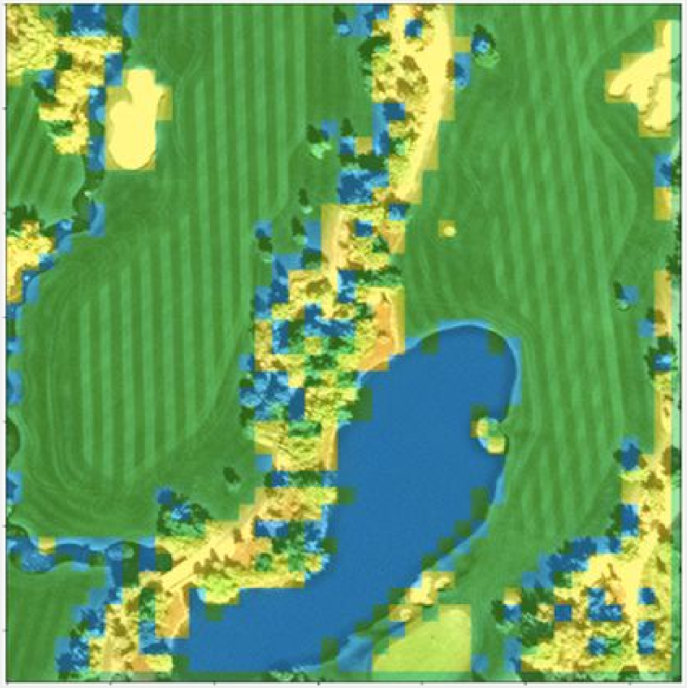
Let's start with a slightly silly example to build up an understanding.

Clustering on Images



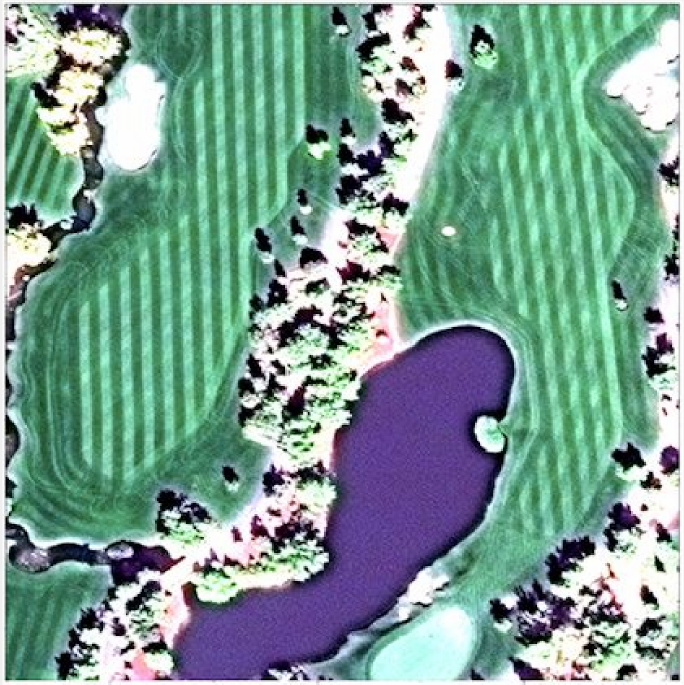
Clustering on Images

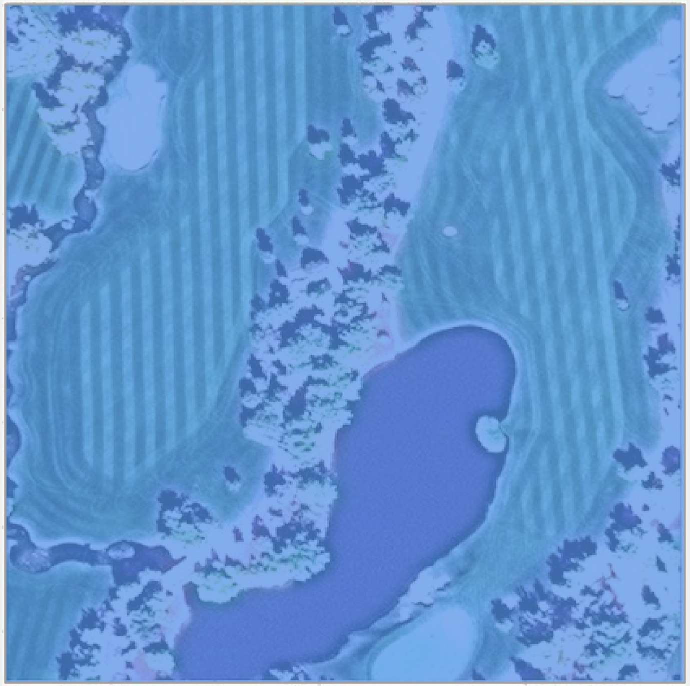
Now let's save some lives!



Images: Susmita Datta (https://ssmtdatta.github.io/satellite-imagery/)

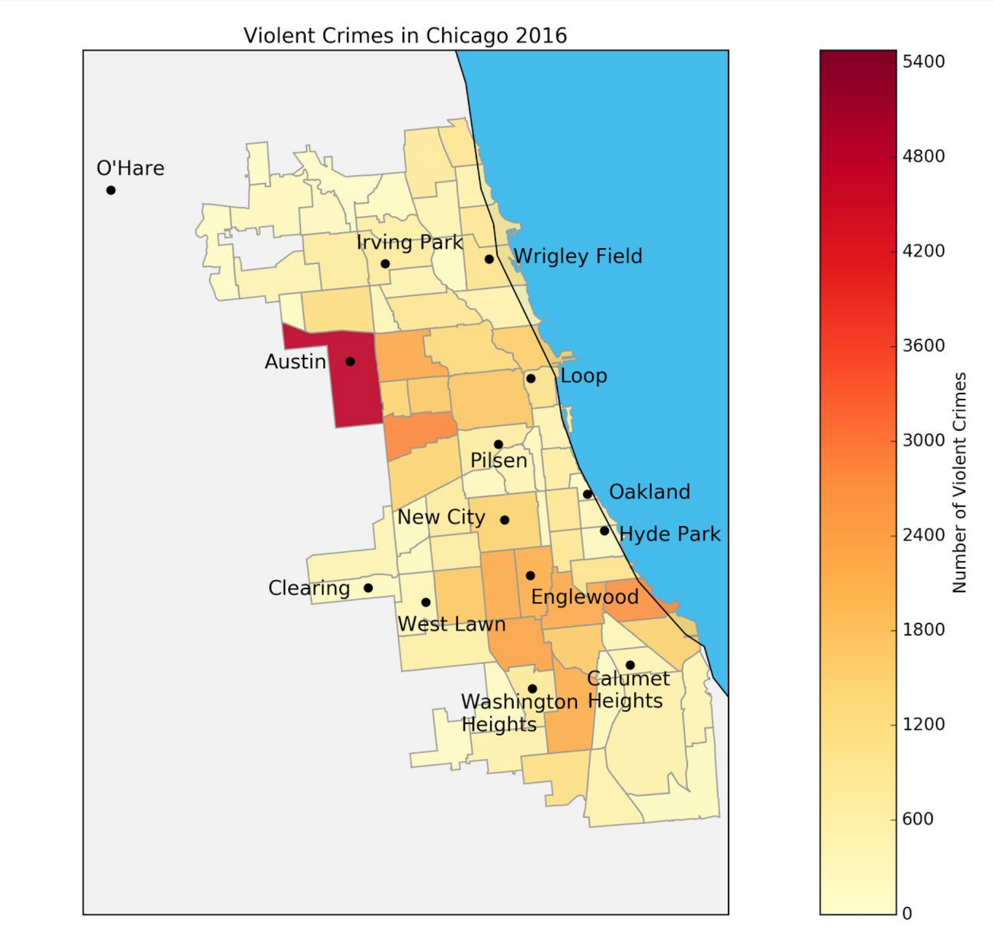
Now let's save some lives!



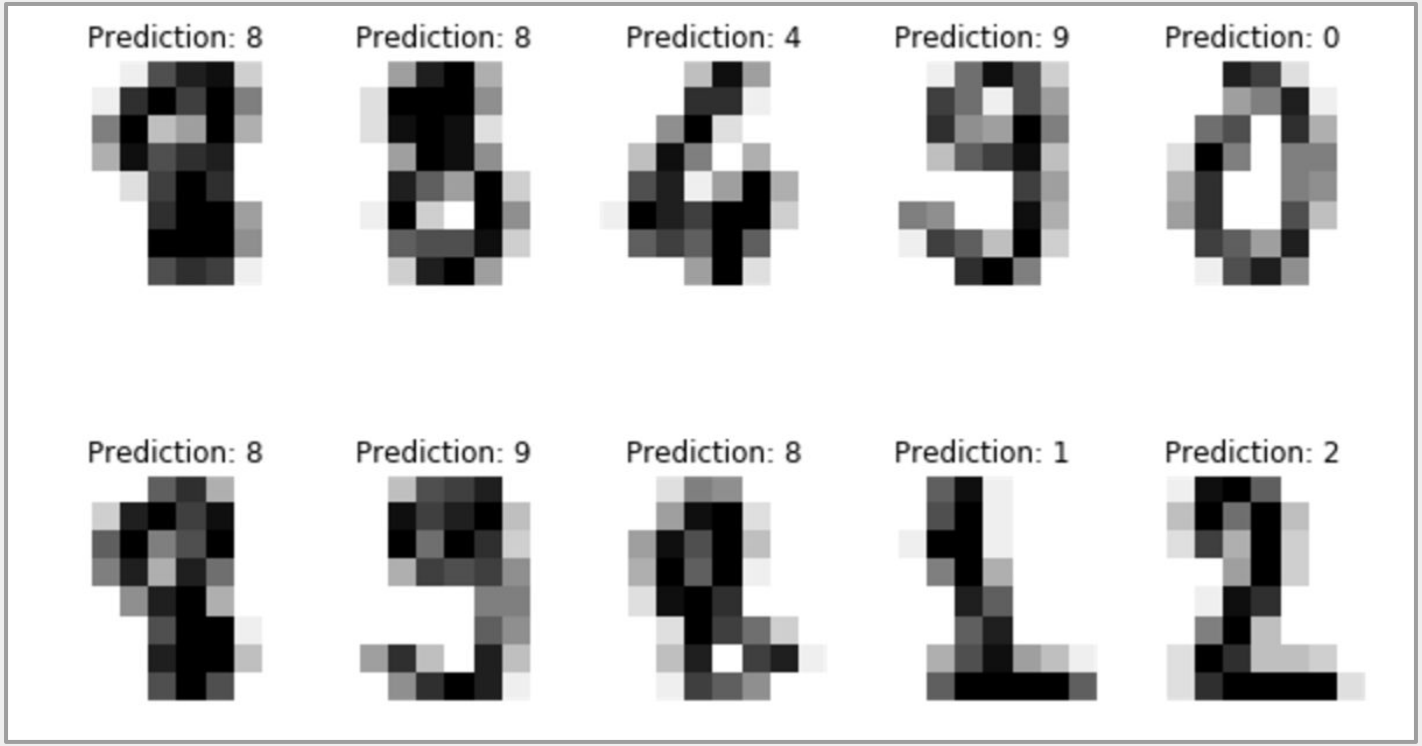


Images: Susmita Datta (https://ssmtdatta.github.io/satellite-imagery/)

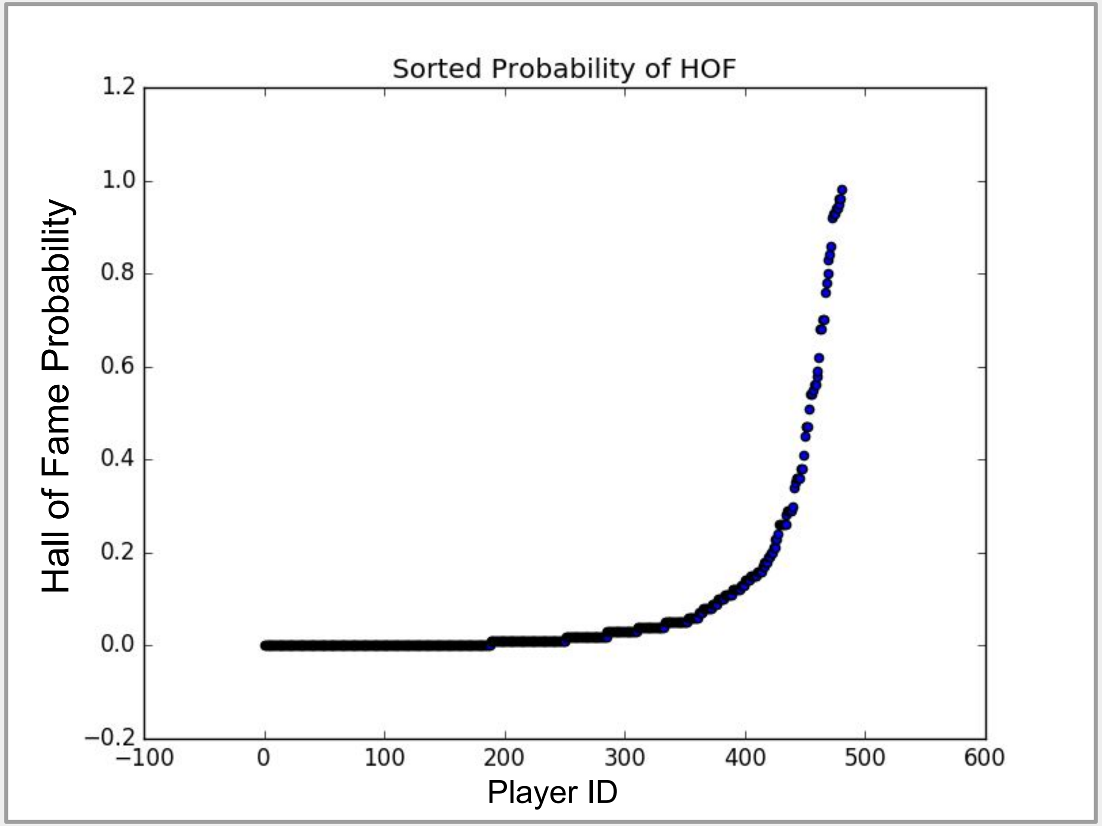
Now let's swiftify justice...



Now let's save some money!



Now let's win some bets

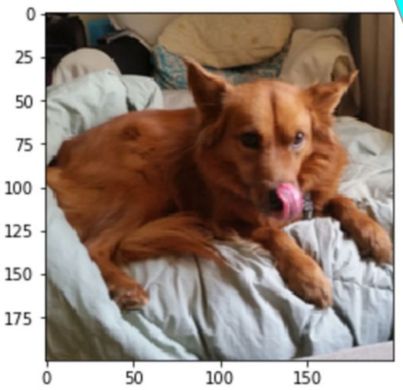


Now let's learn how to see pictures

In [265]:

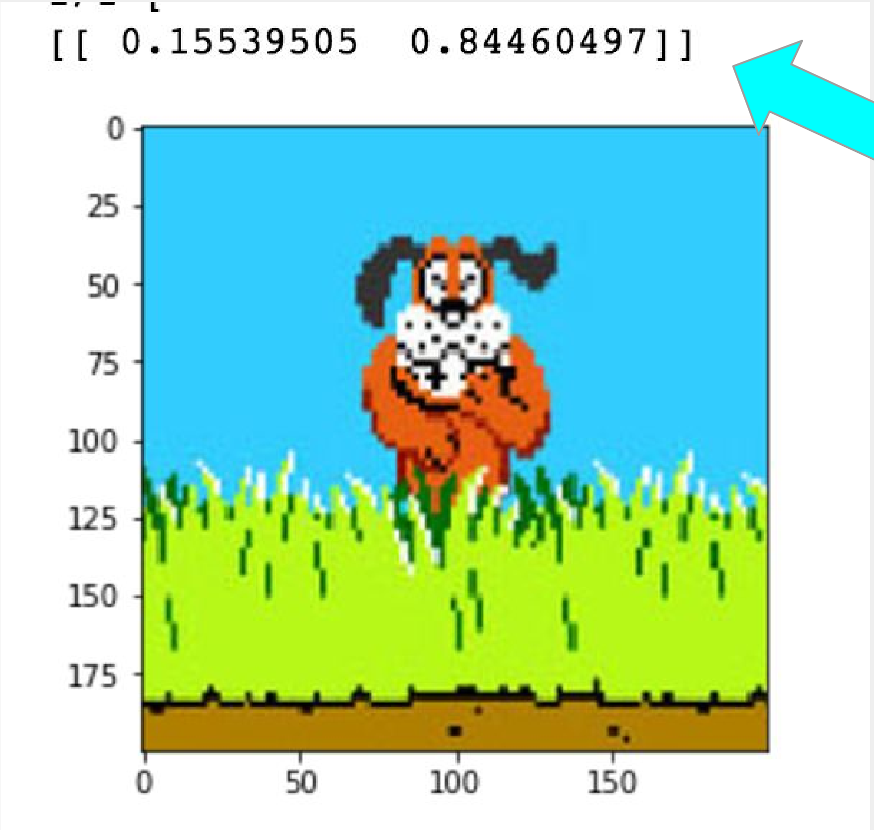
```
def image_to_predict(imgpath):
    img = Image.open(imgpath)
    imgr = img.resize((width,height),resample=Image.ANTIALIAS)
    img_data = np.array(imgr)/255.
    return img_data

img_to_predict = np.empty(shape=(1,200,200,3))
pic = image_to_predict('topredict/zachdog.jpg')
img_to_predict[0] = pic
imshow(img_to_predict[0])
prediction = model.predict(img_to_predict, batch_size=1, verbose=1)
print prediction
```

```
1/1 [==============================] - 0s
[[ 0.00377928  0.99622077]]
```



Catness vs Dogness

Now let's learn how to see pictures



Catness vs Dogness

I've shown some visually appealing results, but what else is machine learning being used for?

- Fraud detection on your credit card
- Deciding who gets loans
- Weather prediction
- Detecting cancer from scans
- Diagnosing diseases
- Stock markets
- Self-driving cars
- Image recognition
- Alexa/Siri/OK Google
- Marketing
- Monitoring big machines
- Chat bots
- Recommender systems on text
- Spotify Recommendation
- Pandora Recommendation
- Reading documents to determine classification
- Creating word associations
- Cell system optimization
- Business Intelligence for HR
- Social network building
- SO MANY MORE

Let's Q&A

Thanks for listening to me chat.

zach@thisismetis.com

https://www.linkedin.com/in/zachariah-miller/

Do we still have time? Then, how does a spam filter work?

Spam

“My name is Prince Abdullah, one of the Nigerian royal family. My father has just passed away and I’ve inherited a great sum of money. Due to the laws of Nigeria, I need somewhere to store the money temporarily. If you can help me, I’ll give you a 20% share of the inheritance. Please response immediately, as I don’t have much time.”

Ham

Hey Mom,

I totally didn’t forget your birthday, Amazon is just out of stock for the item I’m sending you. It will arrive a few days late. Really sorry, and I hope you had a happy birthday yesterday.

Love,
Zach

Spam

"My name is Prince Abdullah, one of the Nigerian royal family. My father has just passed away and I've inherited a great sum of money. Due to the laws of Nigeria, I need somewhere to store the money temporarily. If you can help me, I'll give you a 20% share of the inheritance. Please response immediately, as I don't have much time."

Ham

Hey Mom,

I totally didn't forget your birthday, Amazon is just out of stock for the item I'm sending you. It will arrive a few days late. Really sorry, and I hope you had a happy birthday yesterday.

Love,
Zach

What question should we ask?

- How many times did a SPAM email have the words “Nigerian royal family” and “inheritance” and “please response immediately”?
- How many times did a NON-SPAM email have those?
- What if we devised a way to find all the combinations of words in the email, then asked that question about ALL of those combinations… then we could learn about the spam vs non-spam ratios from historical data.

Spam

“My name is Prince Abdullah, one of the Nigerian royal family. My father has just passed away and I’ve inherited a great sum of money. Due to the laws of Nigeria, I need somewhere to store the money temporarily. If you can help me, I’ll give you a 20% share of the inheritance. Please response immediately, as I don’t have much time.”

Spam

“My name is 53% SPAM, one of the 75% SPAM. My father has just passed away and I’ve inherited a 95% SPAM. Due to the 60% SPAM, I need somewhere to store the money temporarily. If you can help me, I’ll give you a 55% SPAM. 99% SPAM, as I don’t have much time.”

Spam

“My name is 53% SPAM, one of the 75% SPAM. My father has just passed away and I’ve inherited a 95% SPAM. Due to the 60% SPAM, I need somewhere to store the money temporarily. If you can help me, I’ll give you a 55% SPAM. 99% SPAM, as I don’t have much time.”

Spam

“My name is 53% SPAM, one of the 75% SPAM. My father has just passed away and I’ve inherited a 95% SPAM. Due to the 60% SPAM, I need somewhere to store the money temporarily. 50% SPAM, I’ll give you a 55% SPAM. 99% SPAM, as I don’t have much time.”

So Spam or not?

- We can go through and analyze each combination of words and let them vote spam or not spam, with a power equal to its sureness
- We label as spam if we have stronger votes for Spam than Ham
- This is called a "Naive Bayes" approach (though I've oversimplified just a touch)