MACHINE LEARNING

UC Berkeley Graduate School of Journalism

What is Machine Learning?

Teaching a computer to perform a specific task without using explicit instructions.

Understanding the terms

Classifier

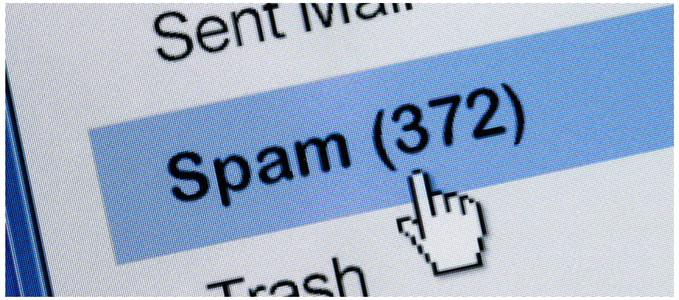
Determining what decision needs to be made based on existing data.

Model

The method used for determining the classification. In most machine learning applications, you're building a predictive model, sometimes a decision model.

Training Data

Data that you feed into your model to train it. The decisions or predictions will be made based on this training data. We will evaluate it's effectiveness.

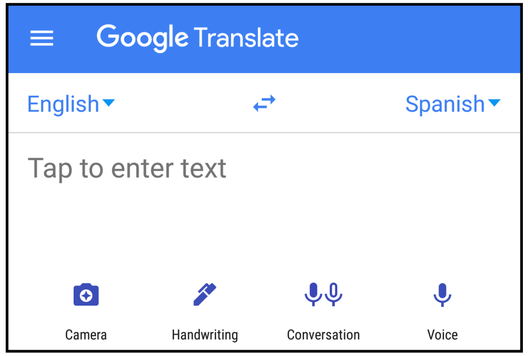


Spam Filtering

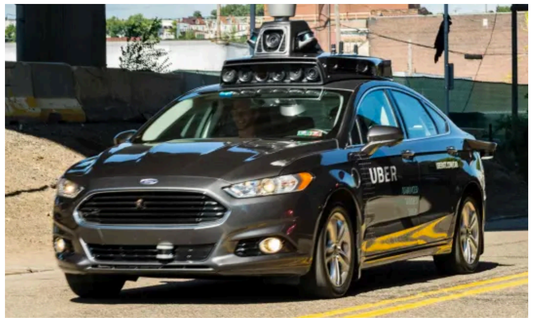
Facial Recognition

Advertising

Siri



Google Translate



Self-driving cars

Social media

Alexa

News assortment

Examples of Classifiers

Is this email spam?

Using certain attributes in previous emails to determine a binary (true or false) outcome. Is this spam?

Where can I find true love?

Some models are just about predicting a binary outcome, but rather finding many matches, like facial recognition or online dating. This is called multi-class classification.

Medical detection

Detecting a number of potential diseases that you might have based on examples previously seen. This is called multi-label classification.

Voting patterns and fraud

What attributes does a person have that predicts how they will vote? And can you find outliers when there are few examples? Called imbalanced classification.

Classifying Medical Data

Each row represents blood tests from a patient. Some have chronic kidney disease, and some do not. One column in the data specifies which have kidney disease.

click to scroll output; double click to hide

Age	Pressure	Gravity	Albumin	Sugar	Red Blood Cells	Pus Cell	Pus Cell clumps	Bacteria	Glucose	Blood Urea	Serum Creatinine	Sodium	Potassium	Hemoglobin	Packed Cell Volume	W Bl Cc
48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6
53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12
63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4
68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11
61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9
48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6
69	70	1.01	3	4	normal	abnormal	notpresent	notpresent	264	87	2.7	130	4	12.5	37	9
73	70	1.005	0	0	normal	normal	notpresent	notpresent	70	32	0.9	125	4	10	29	18
73	80	1.02	2	0	abnormal	abnormal	notpresent	notpresent	253	142	4.6	138	5.8	10.5	33	7
46	60	1.01	1	0	normal	normal	notpresent	notpresent	163	92	3.3	141	4	9.8	28	14

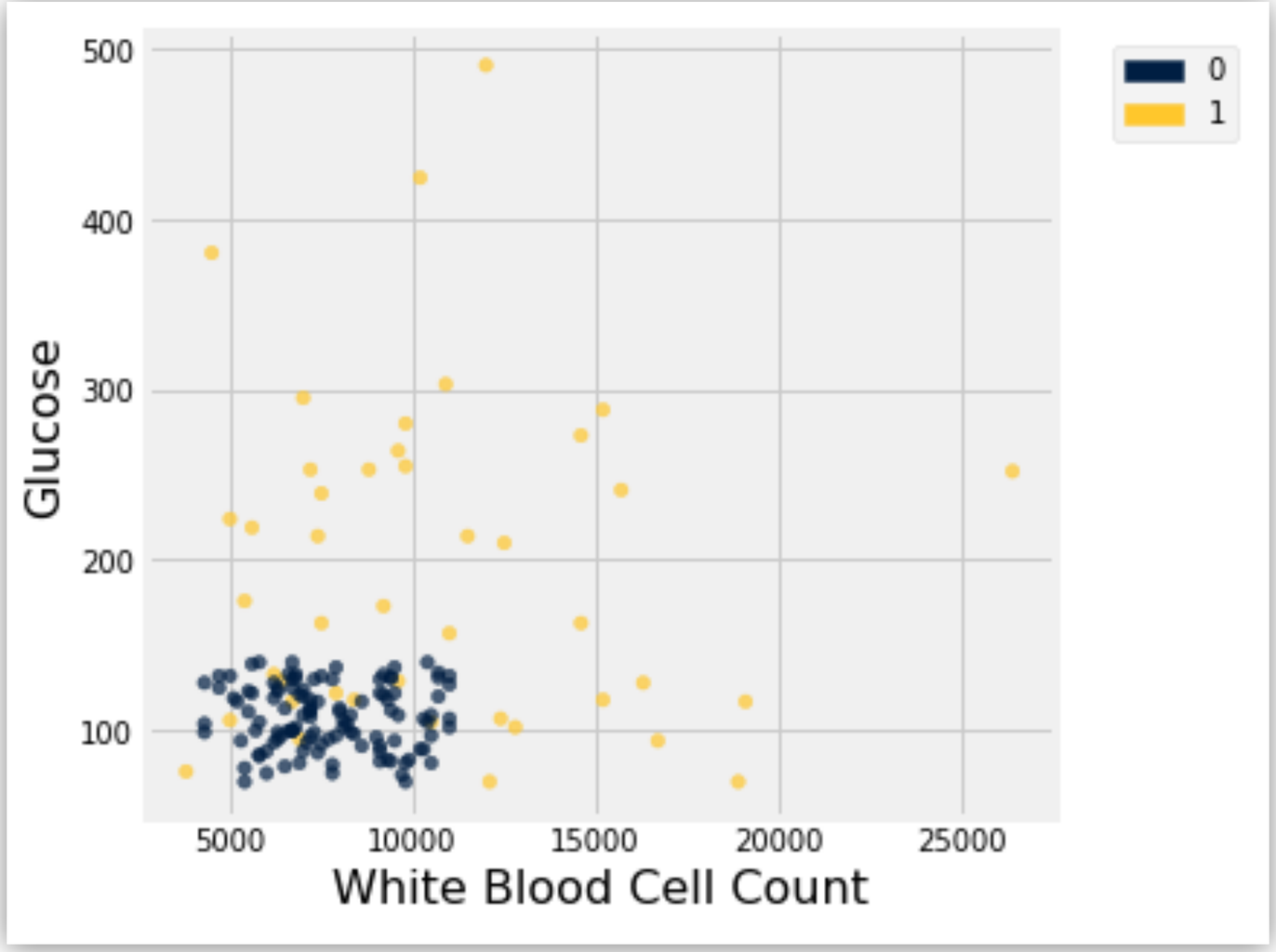
Class 0 means no kidney disease

Class 1 means they have kidney disease

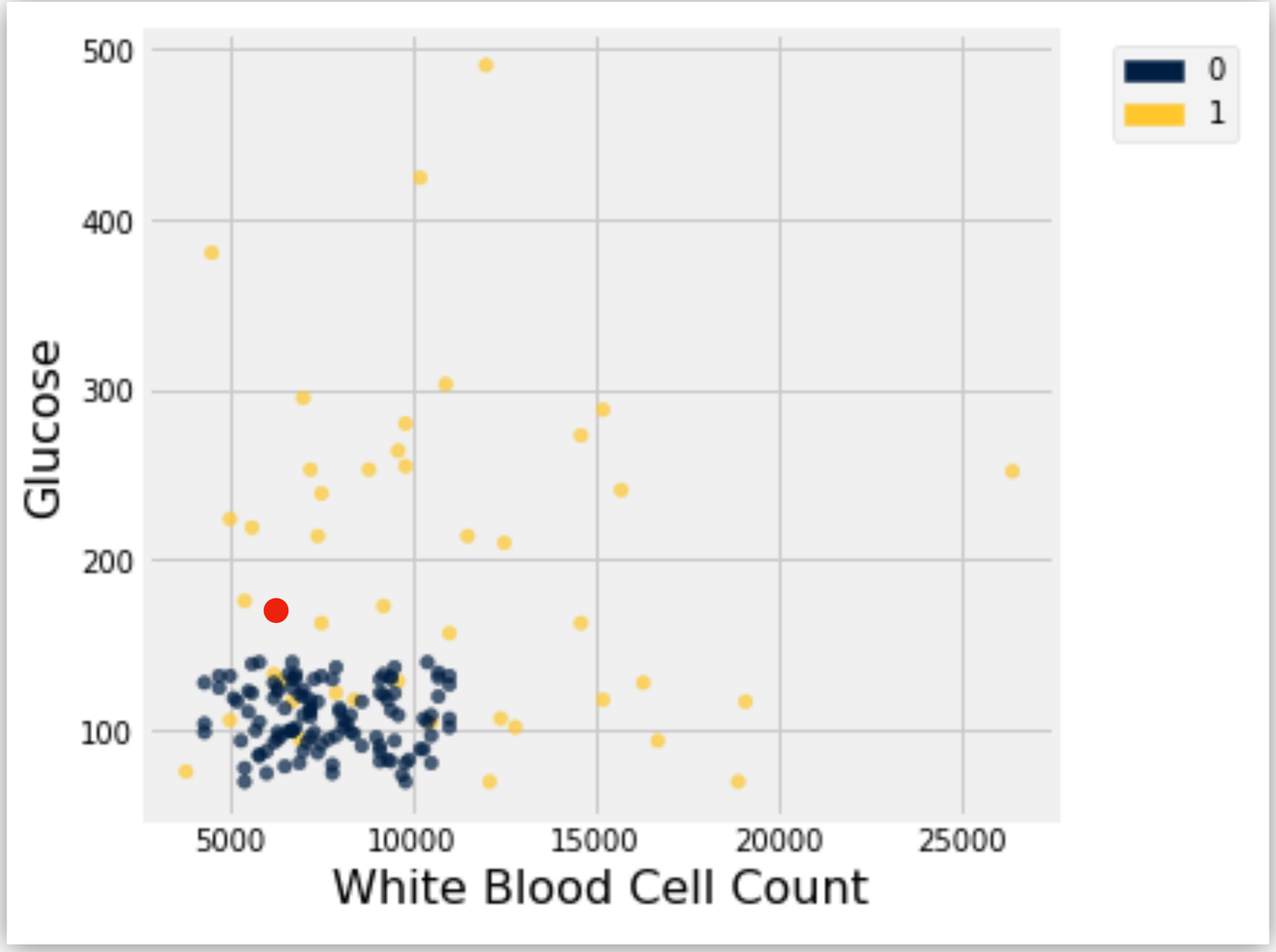
```
ckd.group('Class')
```

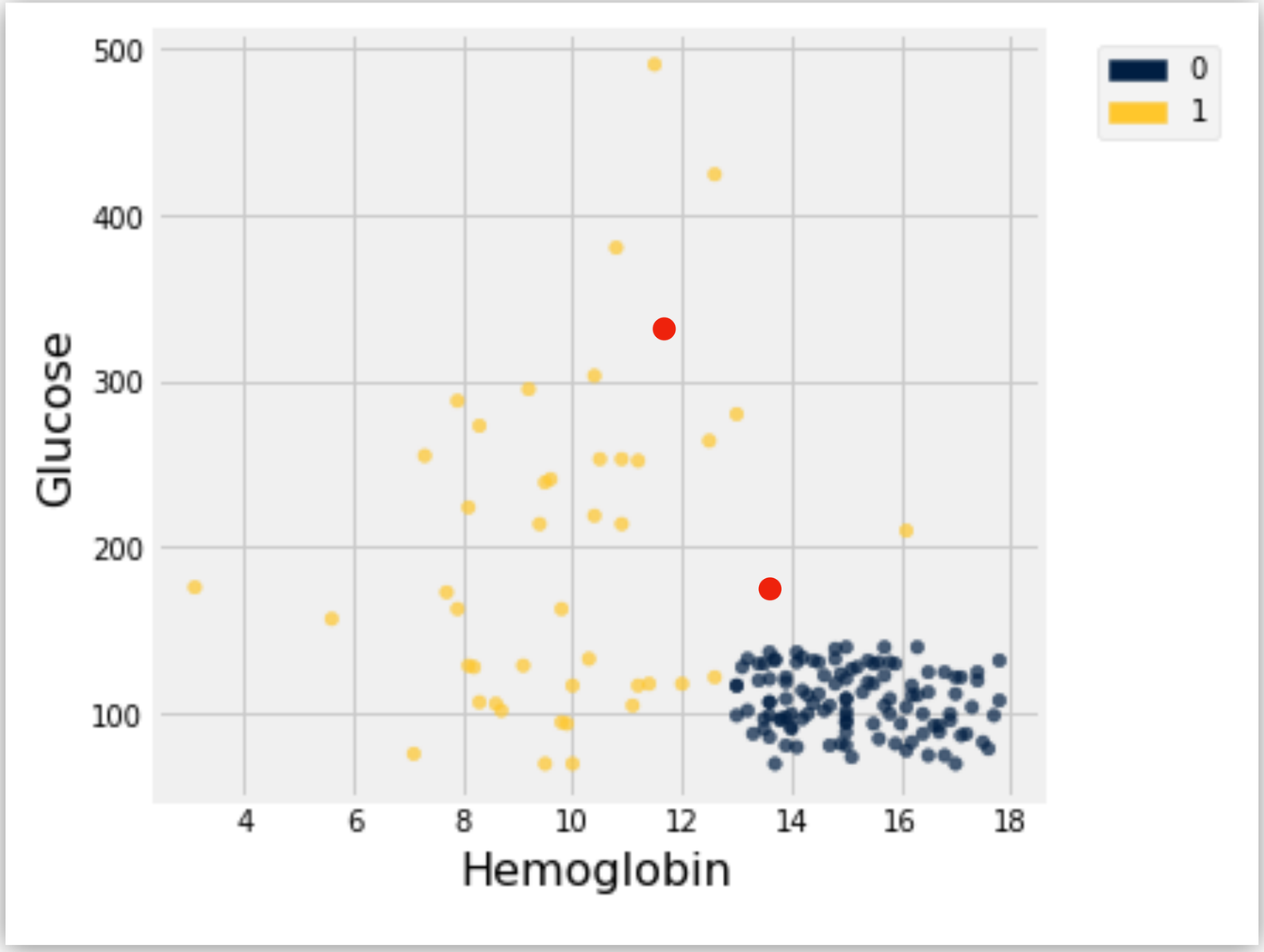
Class	count
0	115
1	43

Look for clustering in the data



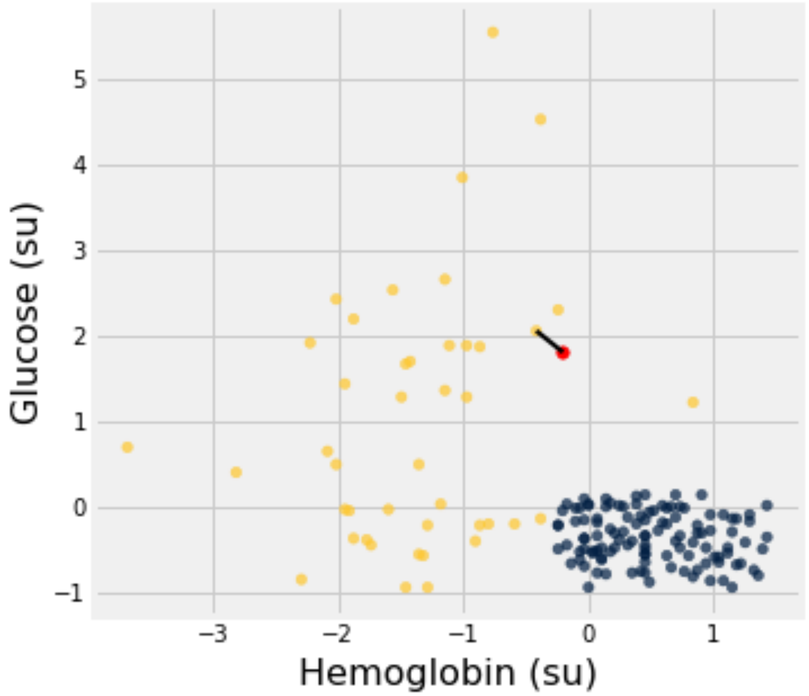
Look for clustering in the data

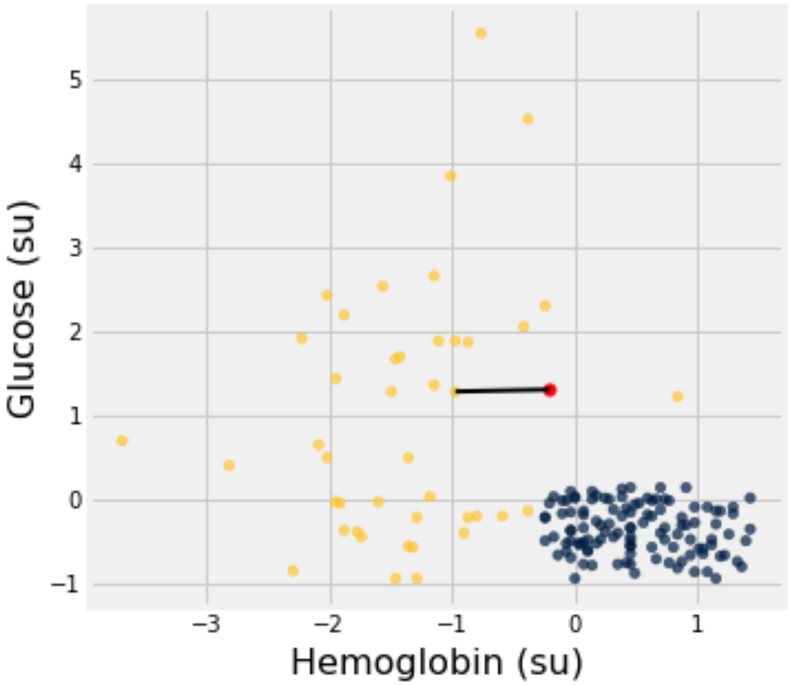


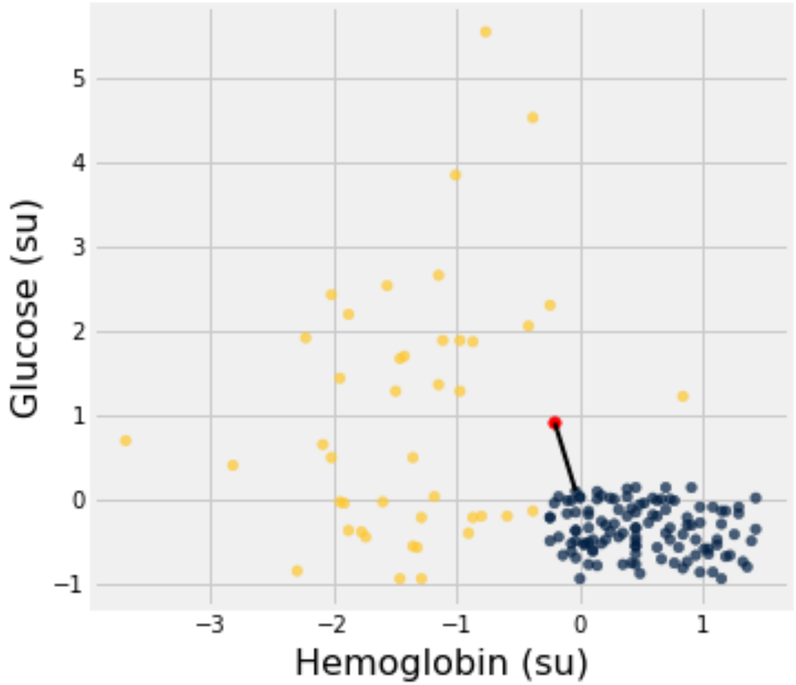

Glucose
Hemoglobin
500
400
300
200
100
4
6
8
10
12
14
16
18
0
1

Nearest Neighbor Classifier

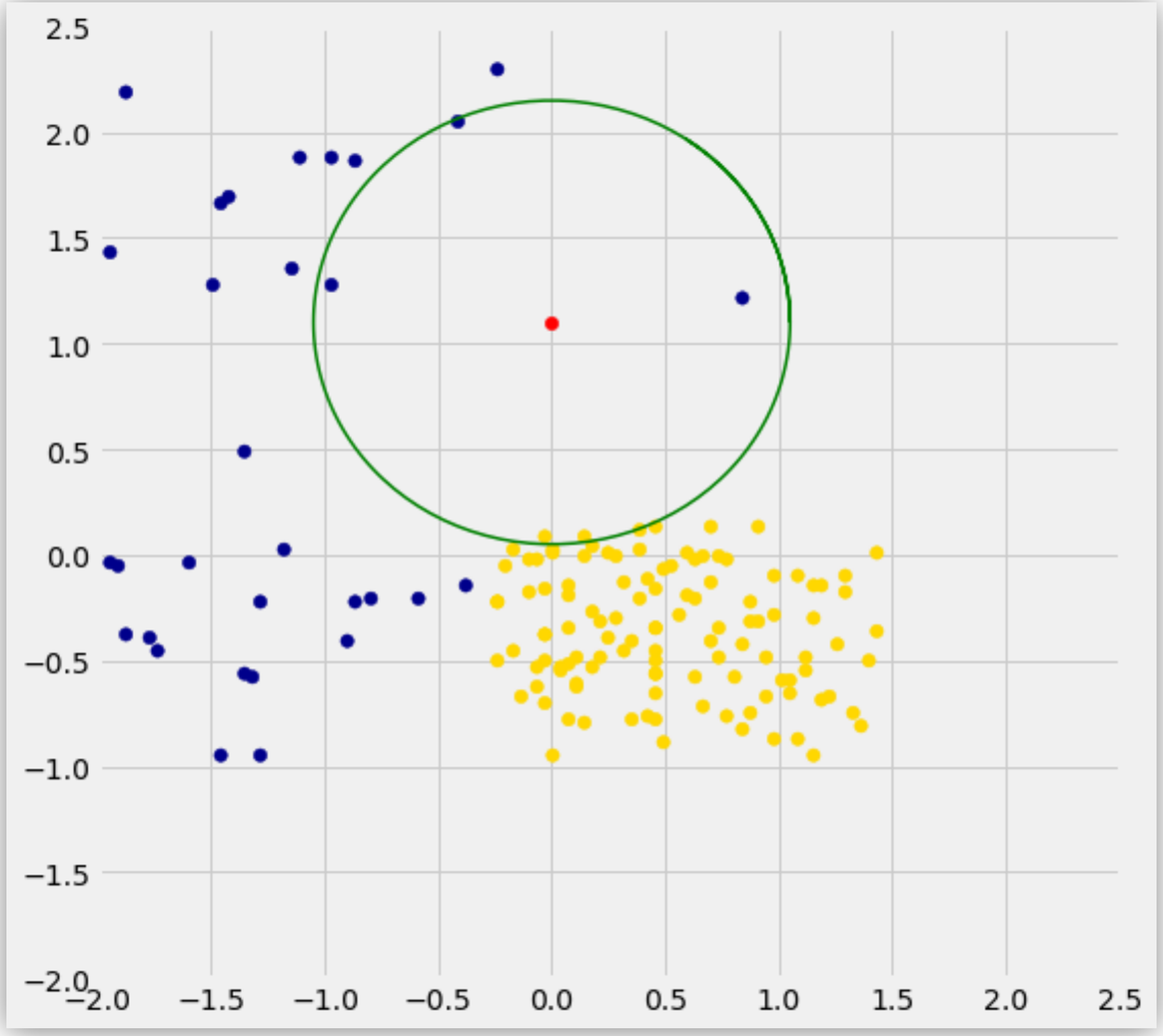
Find the points that are closest to the new data point to make a determination of whether it’s a good candidate for either classification.





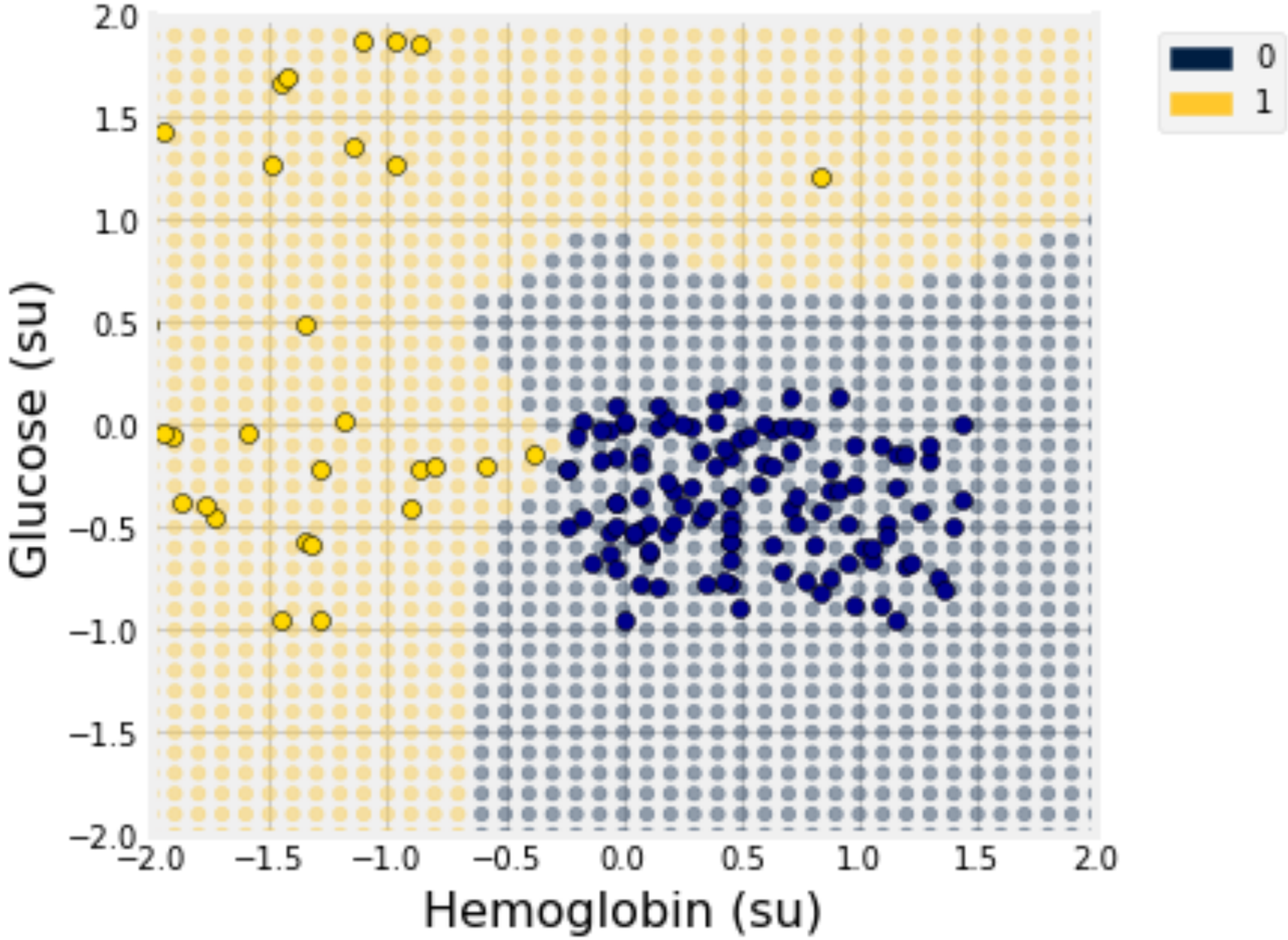


Find the five nearest neighbors



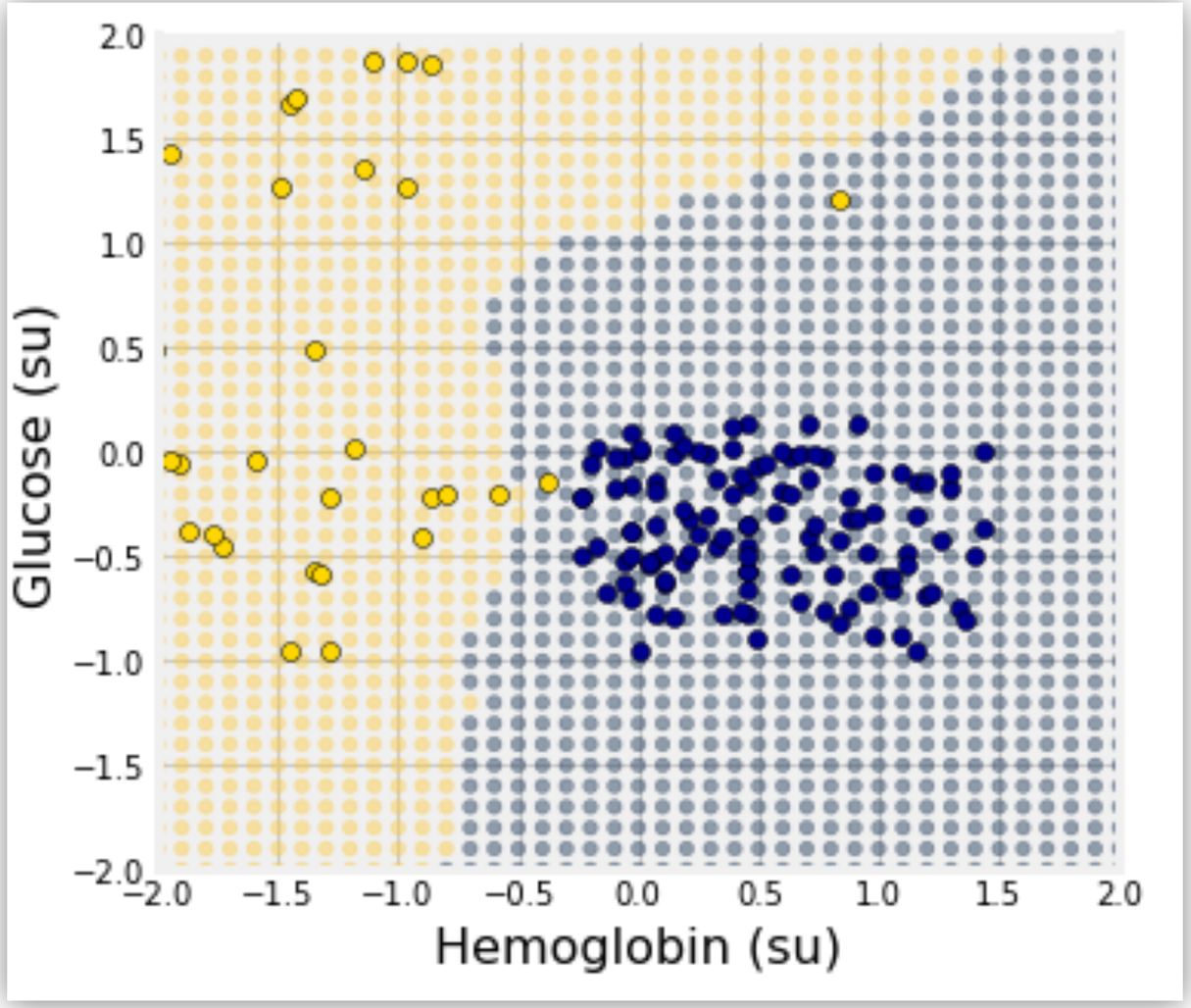
Decision Boundaries

Locate the nearest points classification and create a boundary.



Decision Boundaries

Take an average of the five nearest points to smooth out the boundary.

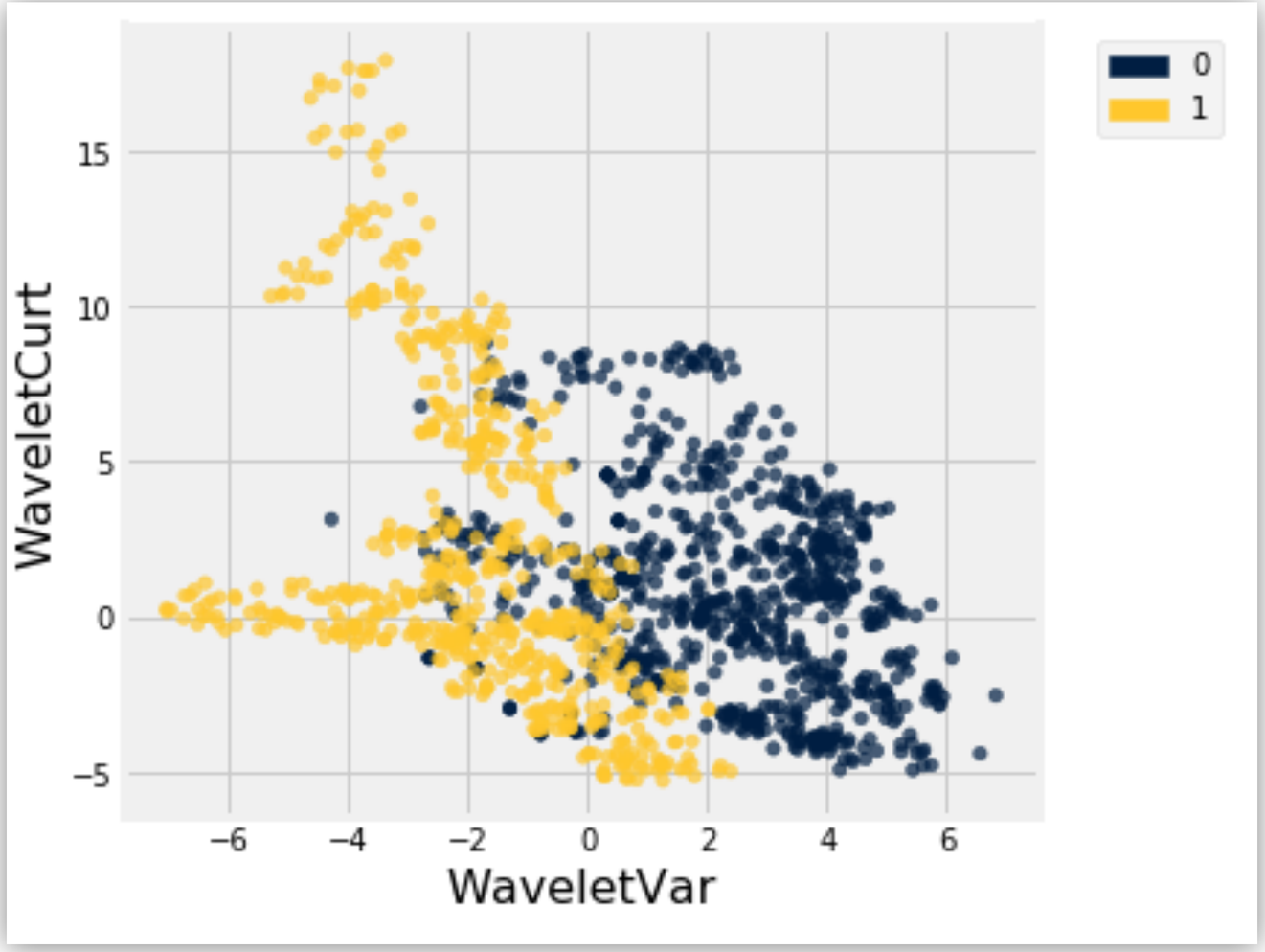


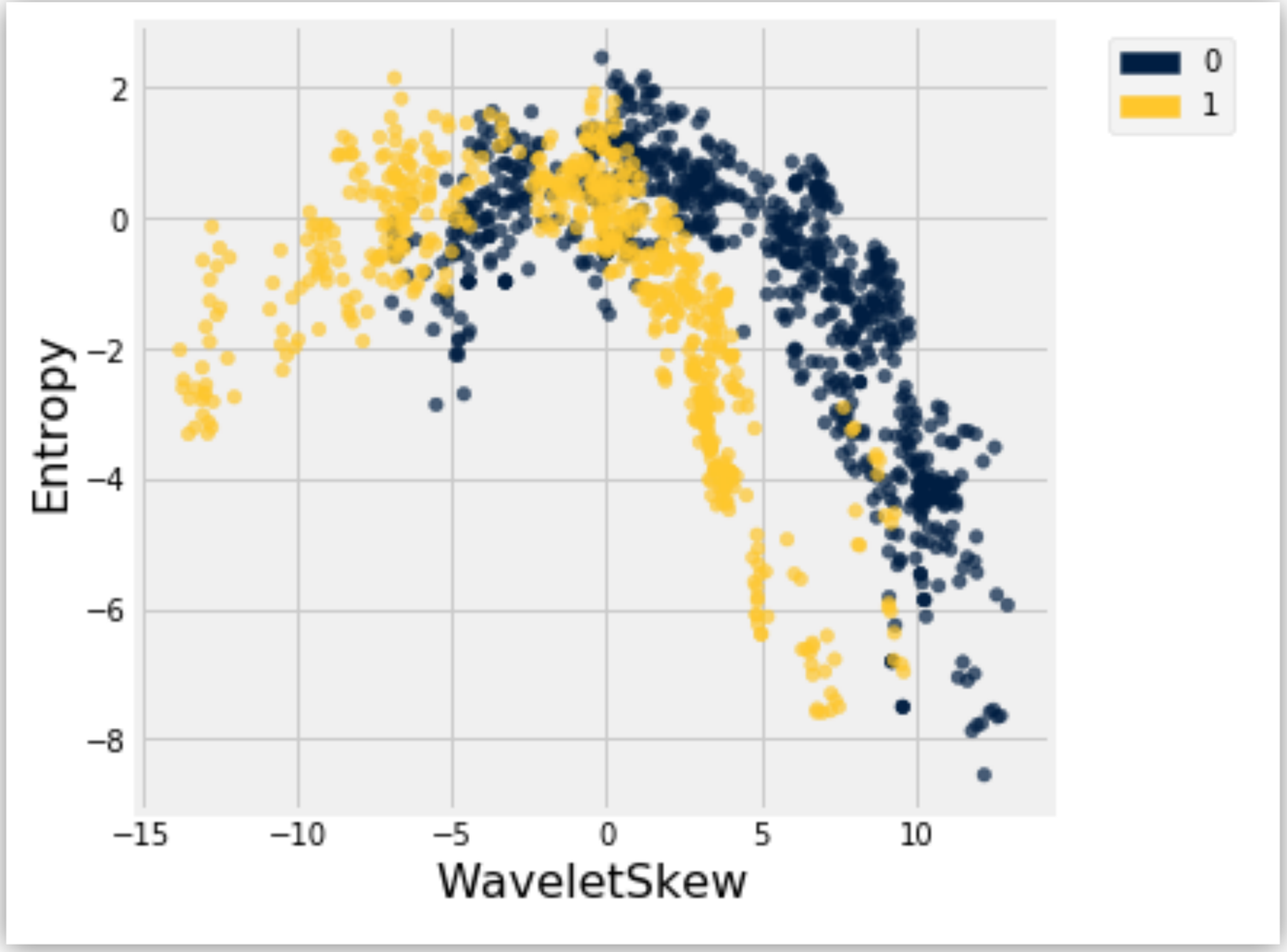
Detecting Counterfeit Banknotes

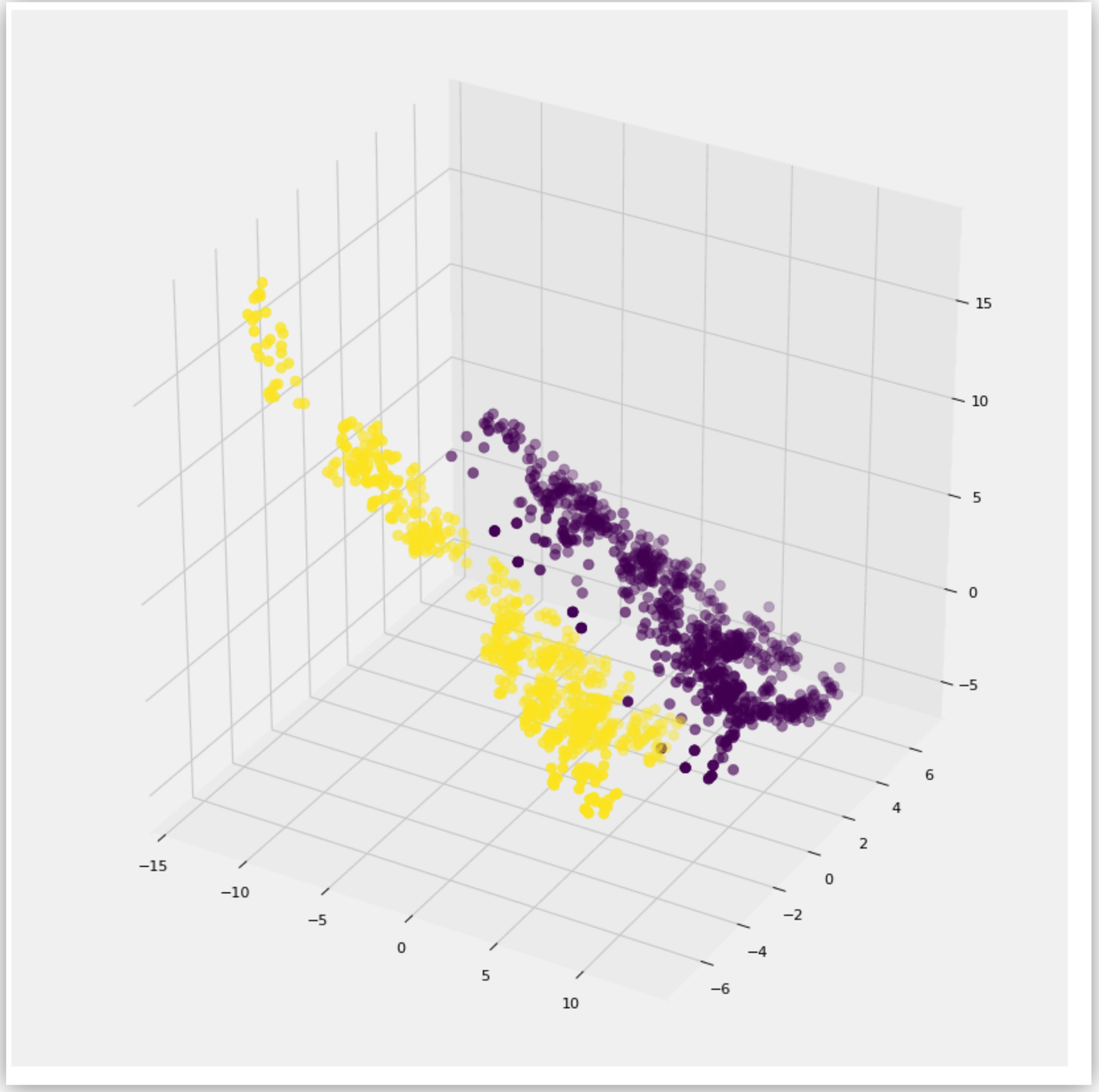
WaveletVar	WaveletSkew	WaveletCurt	Entropy	Class
3.6216	8.6661	-2.8073	-0.44699	0
4.5459	8.1674	-2.4586	-1.4621	0
3.866	-2.6383	1.9242	0.10645	0
3.4566	9.5228	-4.0112	-3.5944	0
0.32924	-4.4552	4.5718	-0.9888	0
4.3684	9.6718	-3.9606	-3.1625	0
3.5912	3.0129	0.72888	0.56421	0
2.0922	-6.81	8.4636	-0.60216	0
3.2032	5.7588	-0.75345	-0.61251	0
1.5356	9.1772	-2.2718	-0.73535	0

... (1362 rows omitted)

Counterfeit Banknotes




0
1
2
0
−2
−4
−6
−8
Entropy
−15
−10
−5
0
5
10
WaveletSkew


15
10
5
0
−5
6
4
2
0
−2
−4
−6
−15
−10
−5
0
5
10

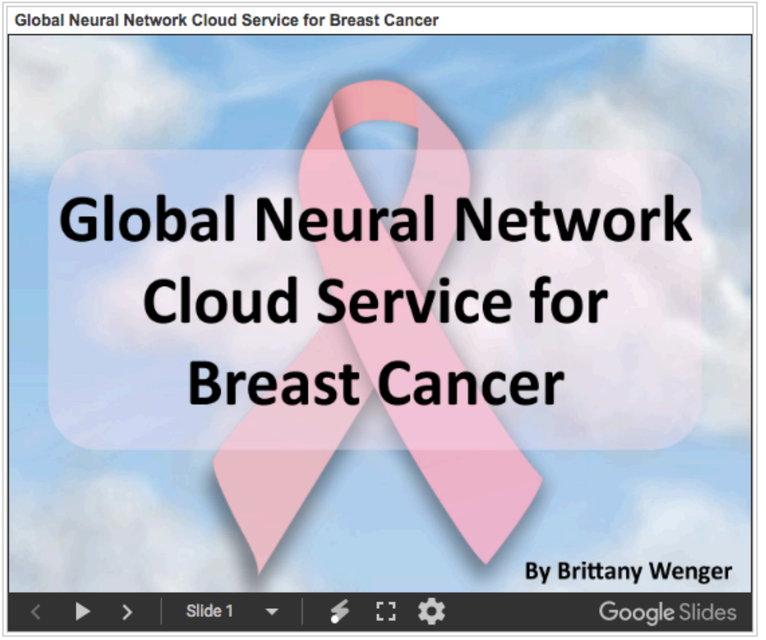
Finding distance between two points.

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}.$$

Finding distance between three points.

$$D = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2}$$

17-year-old Brittany Wenger won Google Science Fair for building a computer program doctors use for Breast Cancer Detection

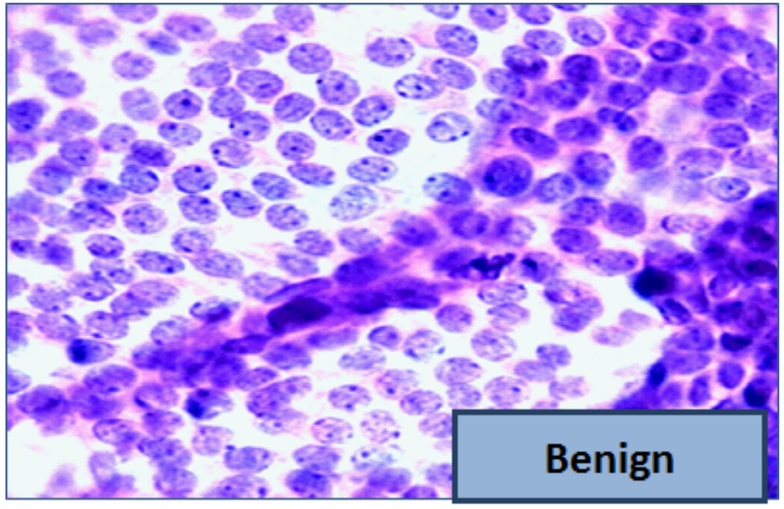


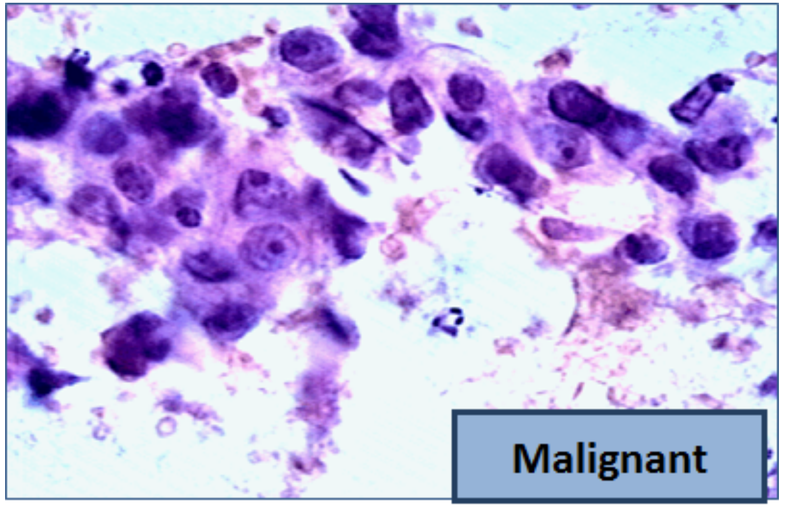
Global Neural Network Cloud Service for Breast Cancer

http://cloud4cancer.appspot.com

Leveraging the computing power of the cloud to assist with medical diagnosis can become an effective tool for doctors to provide more consistent and reliable care. Artificial neural networks detect patterns too complex to be recognized by humans and can be applied to breast mass malignancy classification when evaluating Fine Needle Aspirates (FNAs). This project teaches the cloud how to diagnose breast cancer by implementing a custom-crafted neural network that consumes FNA data collected by the University of Wisconsin to answer the question – is a mass malignant or benign?

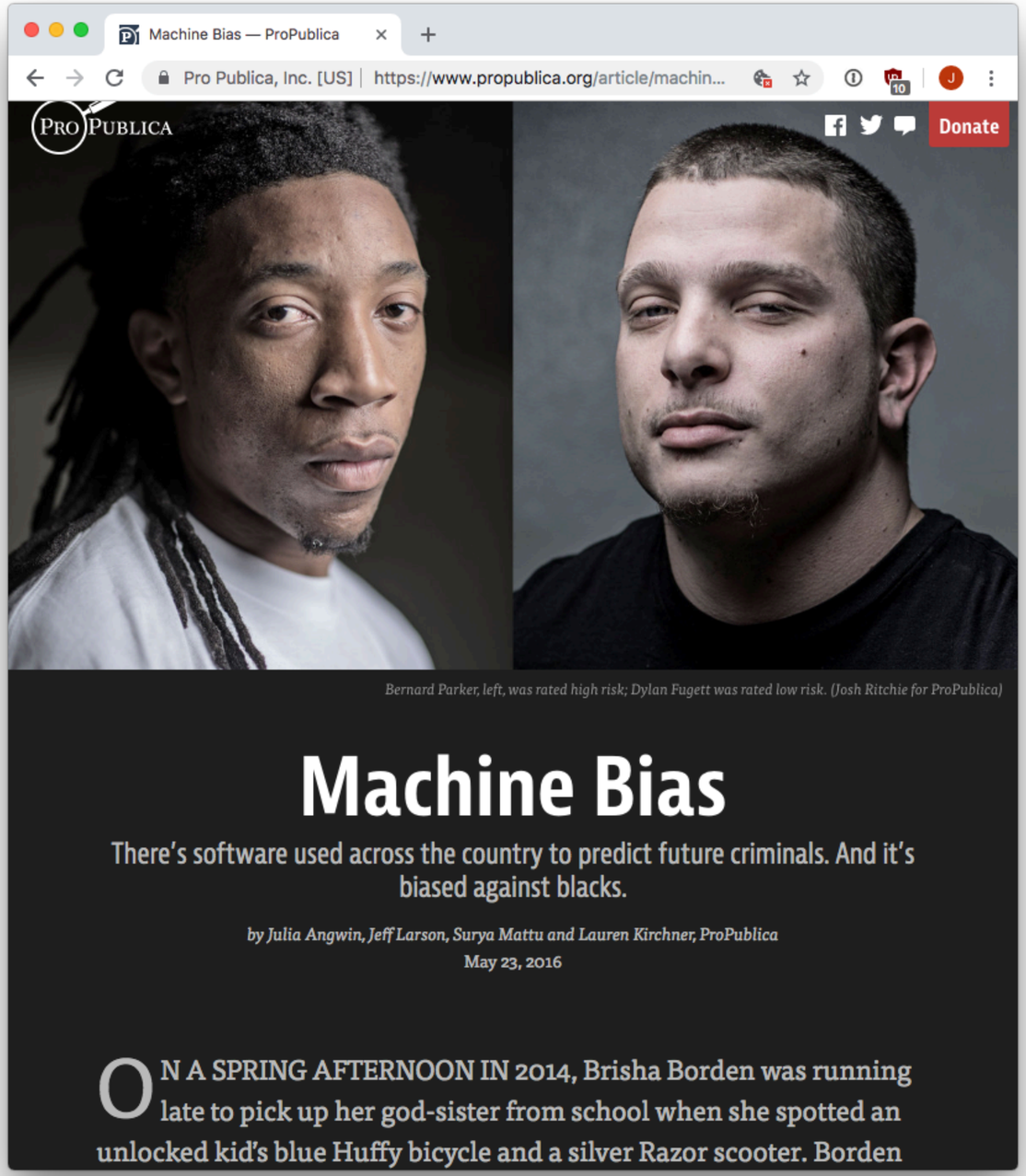
Medical usage demands neural networks achieve accuracy with their diagnosis and reduce malignant false negatives. Building on data collected by the University of Wisconsin in the early 1990s, this project first evaluates three modern commercial neural network implementations. Information regarding potential indicators of breast


Benign


Malignant

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
5	1	1	1	2	1	3	1	1	0
5	4	4	5	7	10	3	2	1	0
3	1	1	1	2	2	3	1	1	0
6	8	8	1	3	4	3	7	1	0
4	1	1	3	2	1	3	1	1	0
8	10	10	8	7	10	9	7	1	1
1	1	1	1	2	10	3	1	1	0
2	1	2	1	2	1	3	1	1	0
2	1	1	1	2	1	1	1	5	0
4	2	1	1	2	1	2	1	1	0

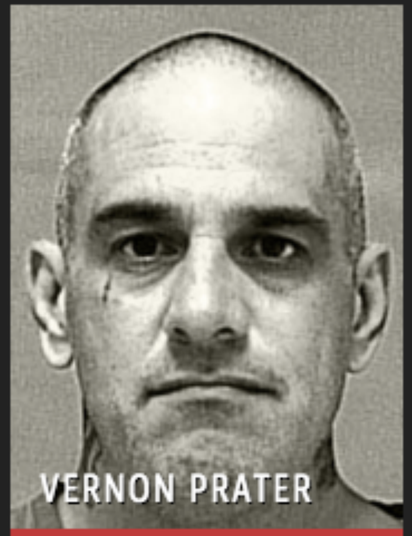
Problems with Machine Learning



https://www.propublica.org/series/machine-bias

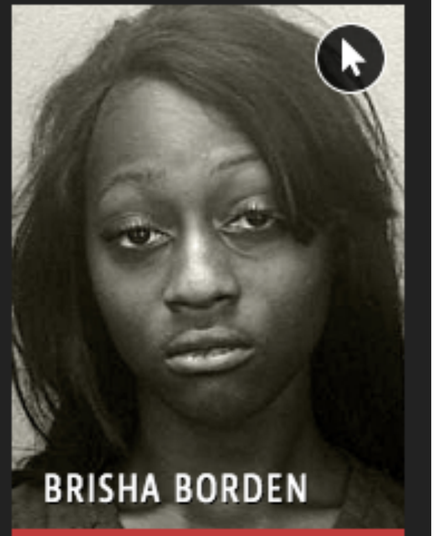
Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: **"Was one of your parents ever sent to jail or prison?" "How many of your friends/ acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?"** The questionnaire also asks people to agree or disagree with statements such as **"A hungry person has a right to steal"** and **"If people make me angry or lose my temper, I can be dangerous."**

Two Petty Theft Arrests



VERNON PRATER

LOW RISK 3

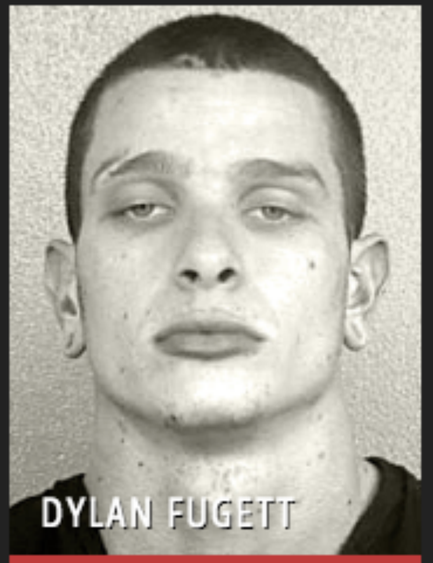


BRISHA BORDEN

HIGH RISK 8

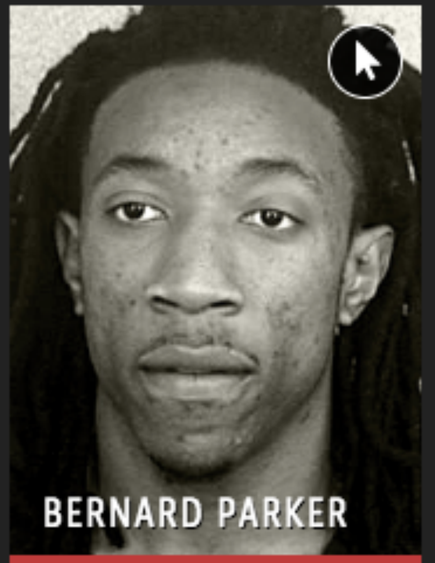
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests



DYLAN FUGETT

LOW RISK 3

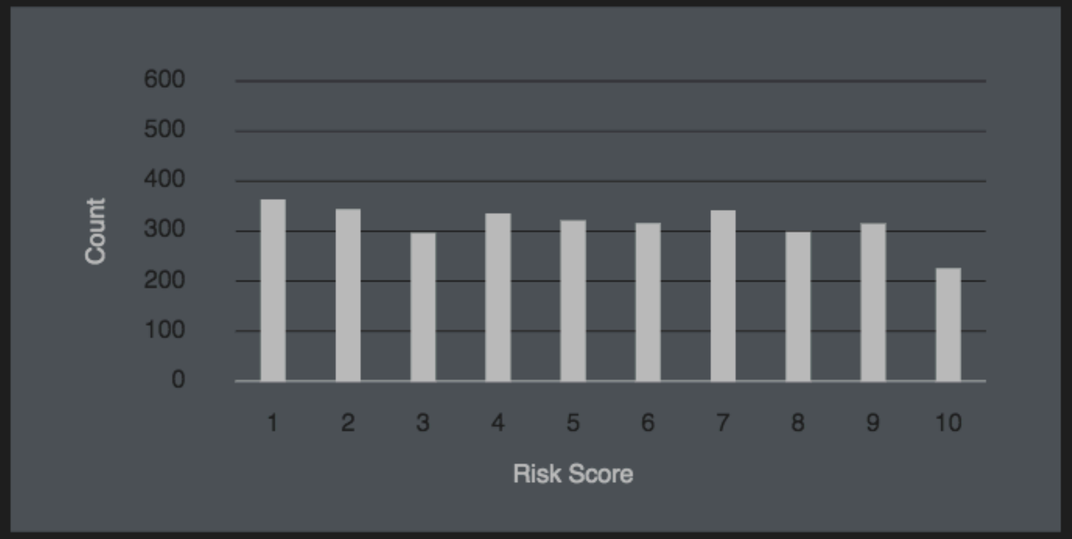


BERNARD PARKER

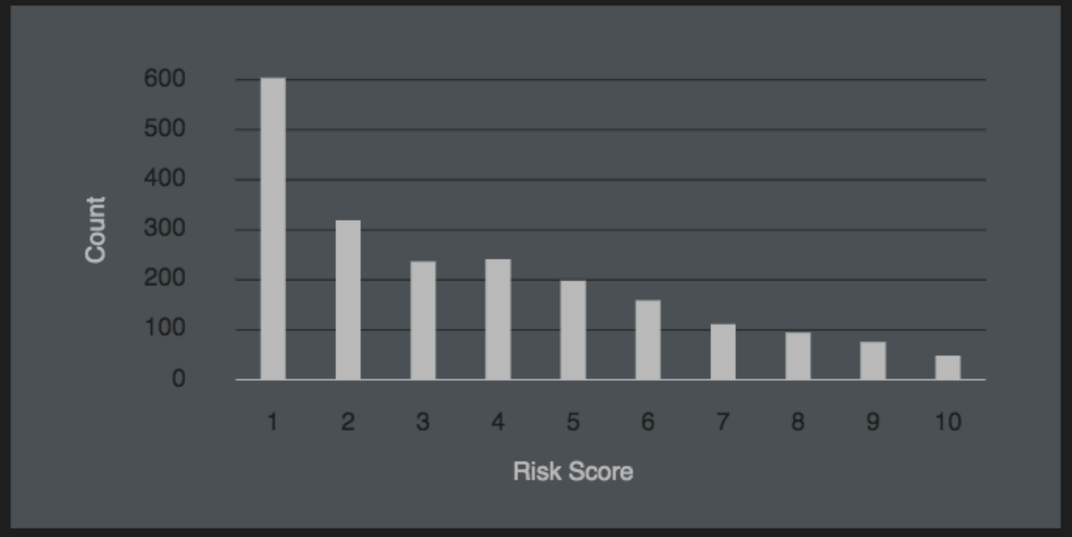
HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Black Defendants' Risk Scores



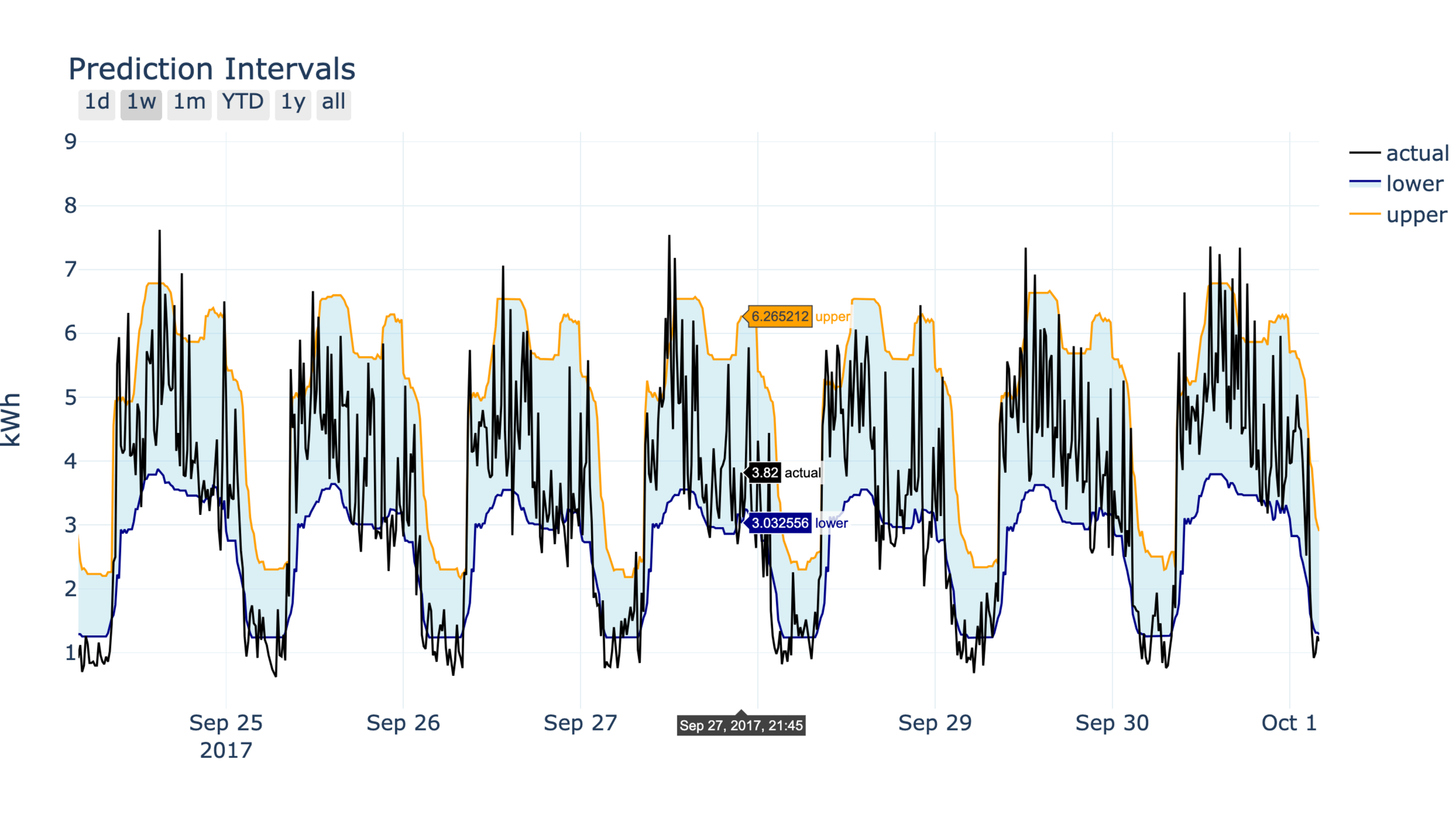
White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

“All models are wrong but some are useful”

— George Box, famed statistician in 1978

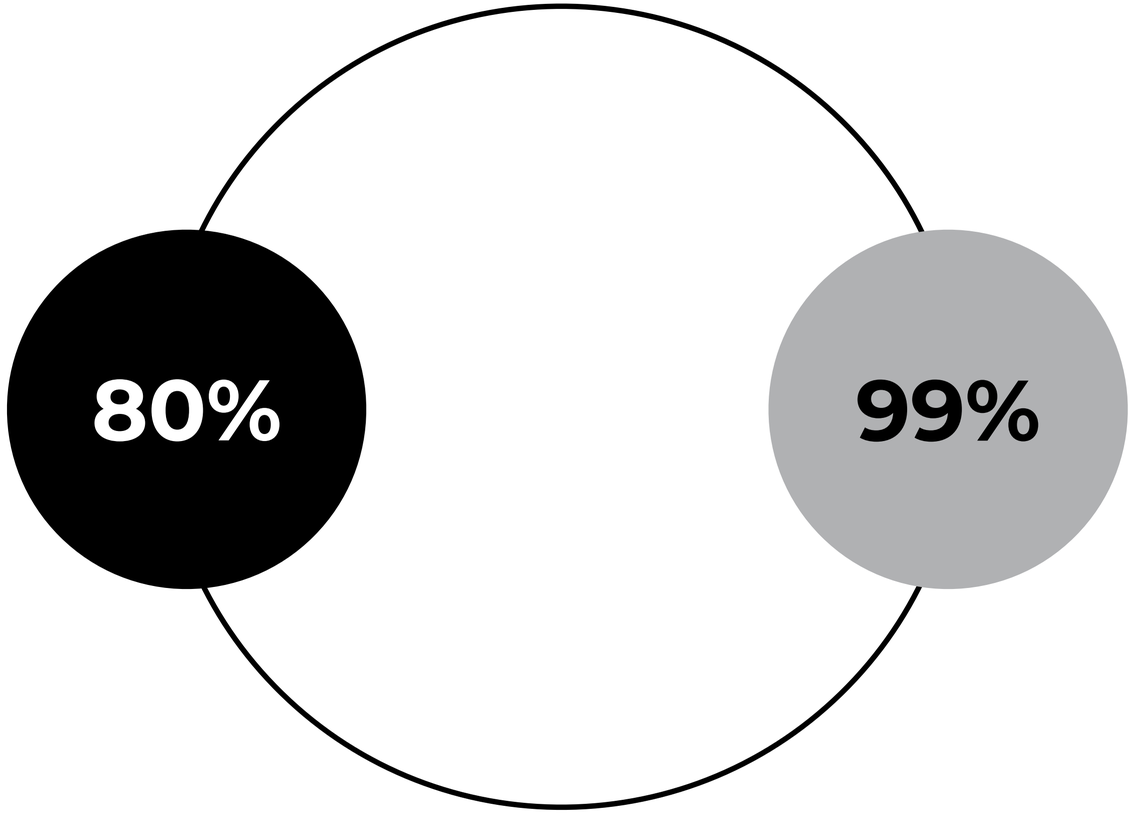

Prediction Intervals
1d
1w
1m
YTD
1y
all
actual
lower
upper
kWh
9
8
7
6
5
4
3
2
1
6.265212 upper
3.82 actual
3.032556 lower
Sep 25
2017
Sep 26
Sep 27
Sep 27, 2017, 21:45
Sep 29
Sep 30
Oct 1

Machine learning is almost never 100% accurate.

Accuracy in Machine Learning

Face detection on Facebook

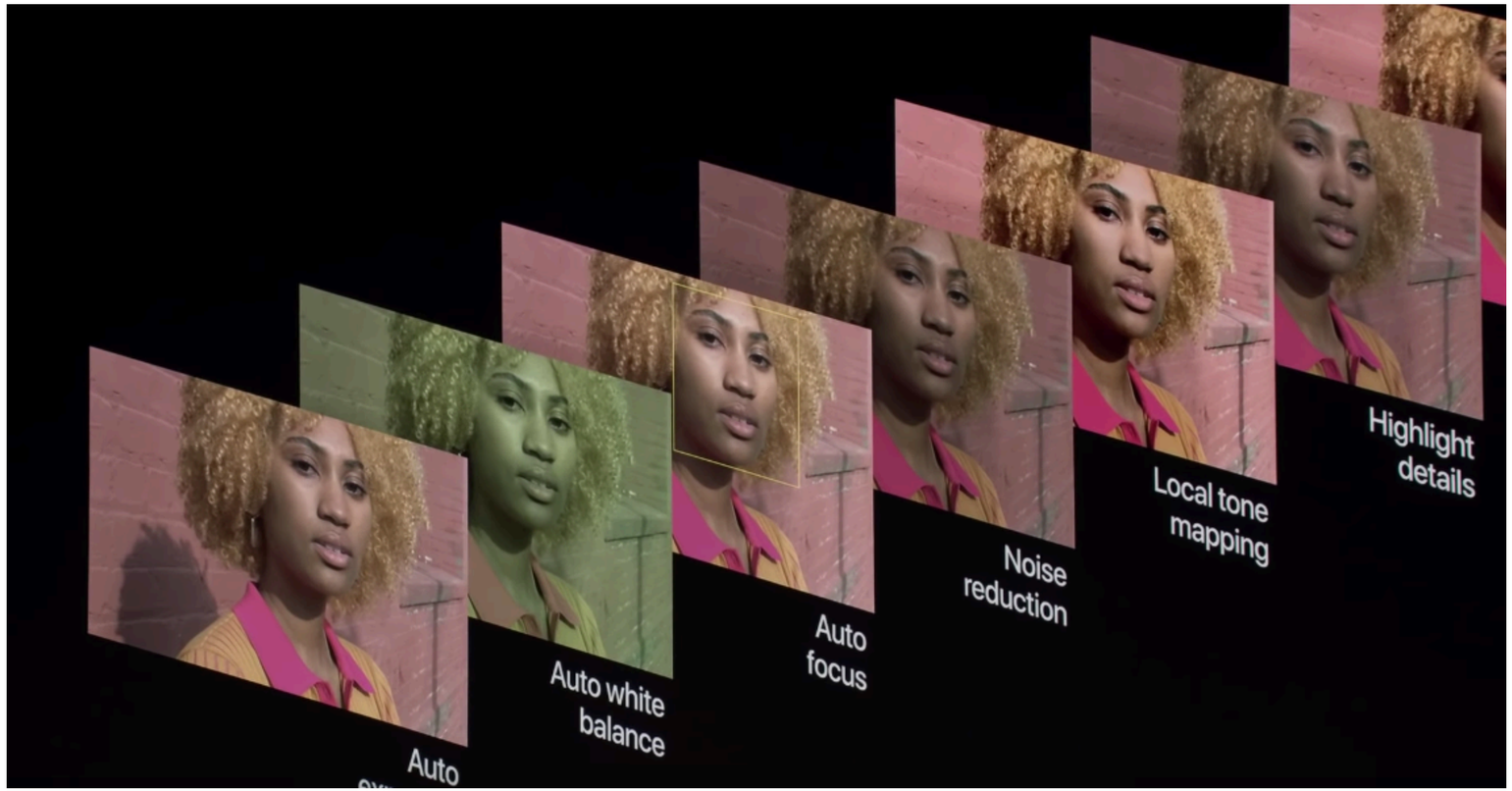
Facebook tries to determine if you're in the photo. Not so much penalty if it's wrong.



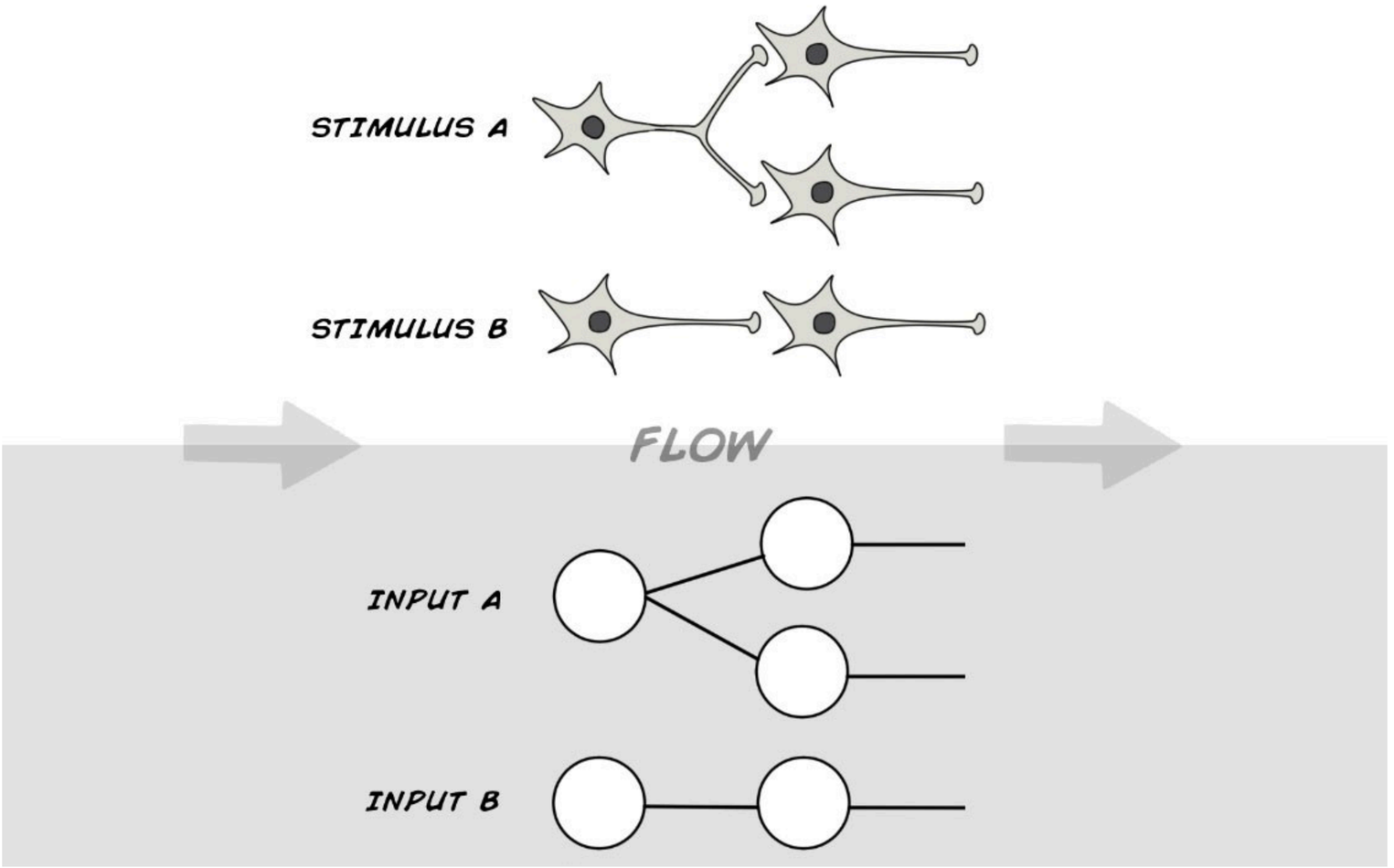
Self-driving cars

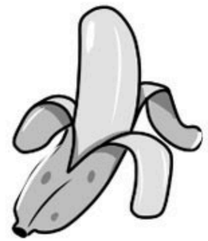
A misinterpretation by the computer could have disastrous consequences.

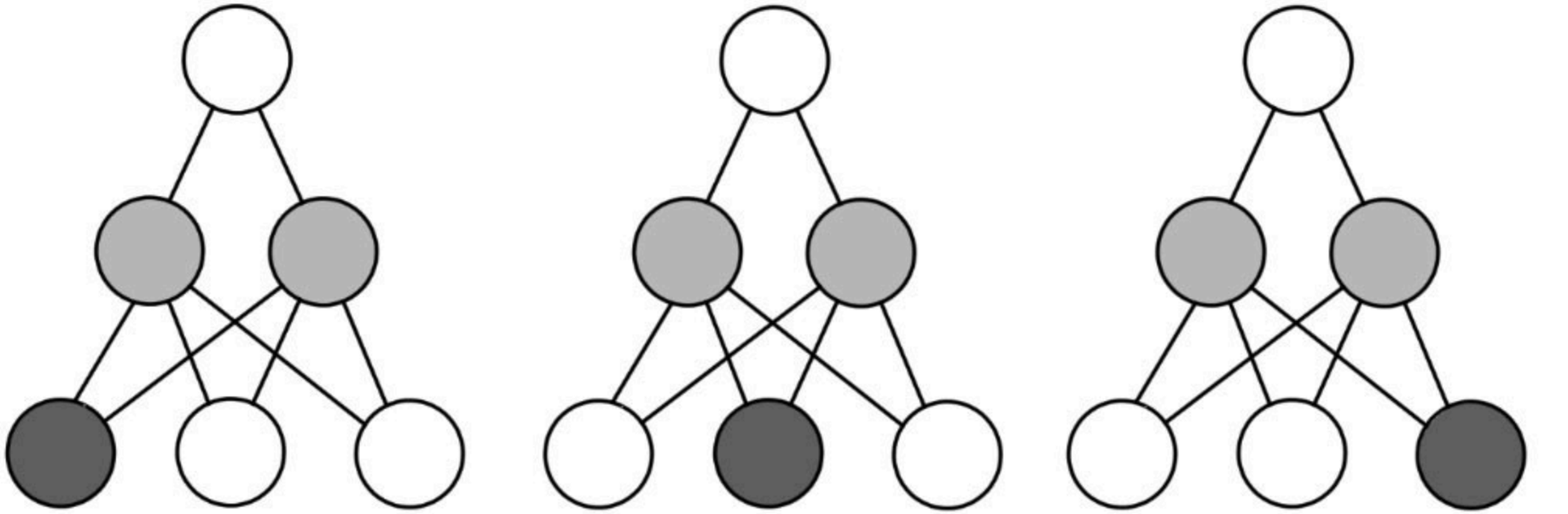
Neural Network Machine Learning

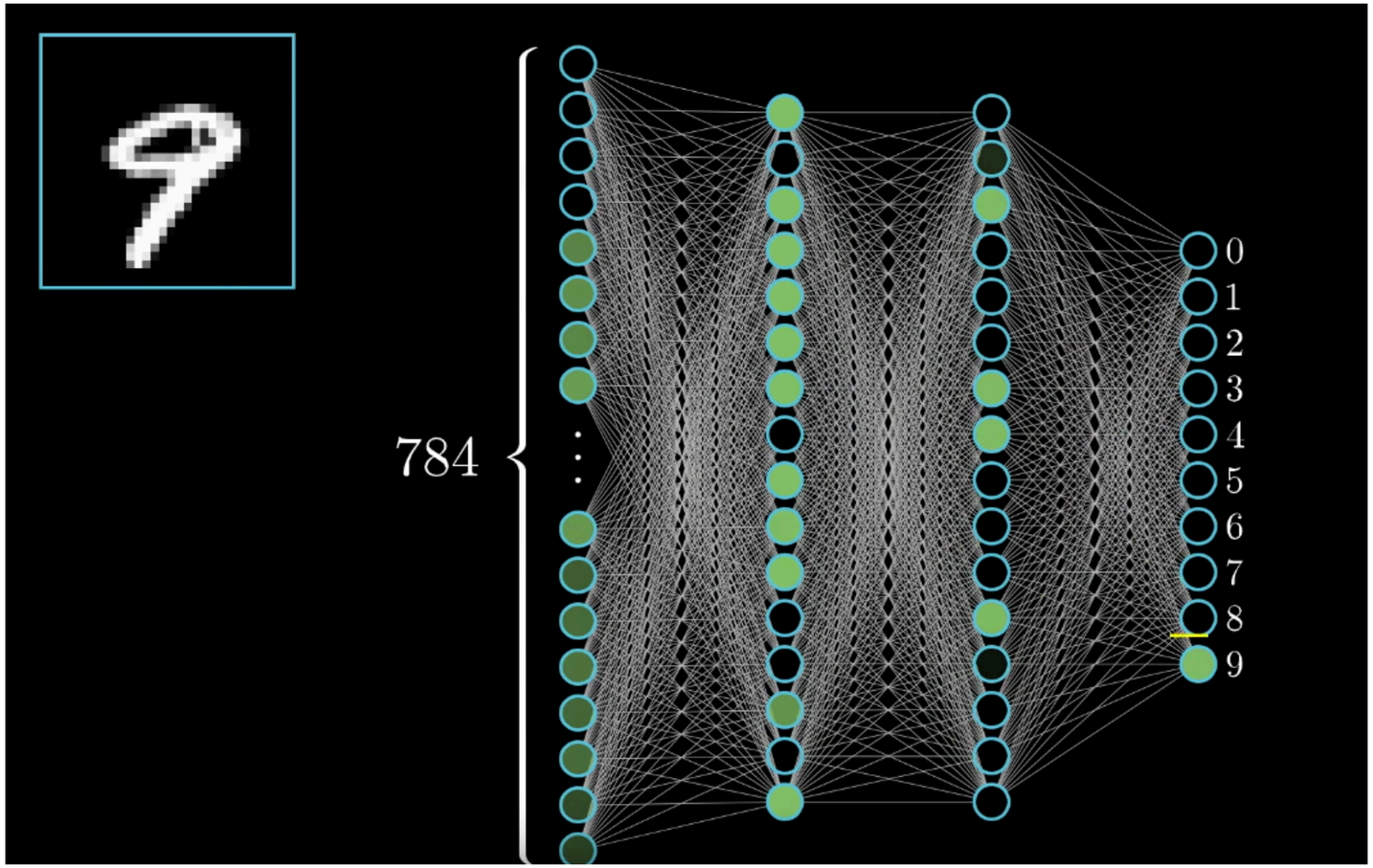


https://youtu.be/wFTmQ27S7OQ?t=4092

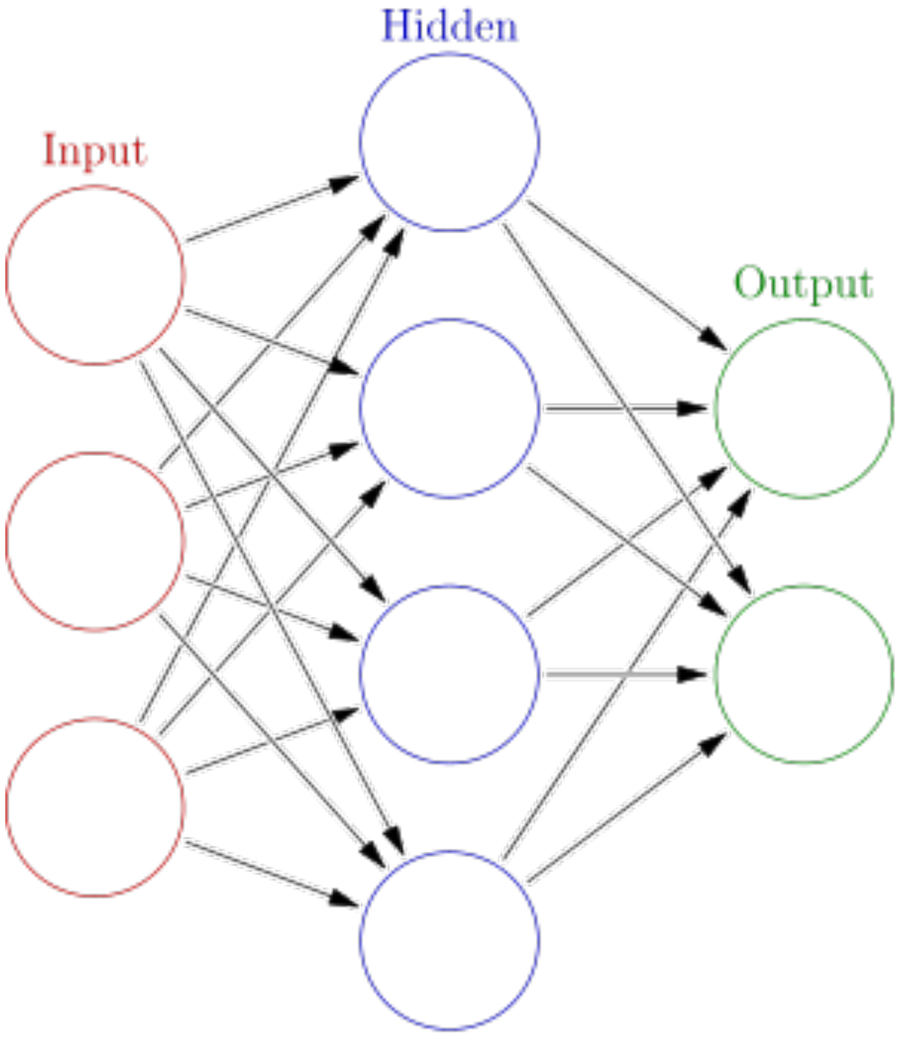

STIMULUS A
STIMULUS B
FLOW
INPUT A
INPUT B

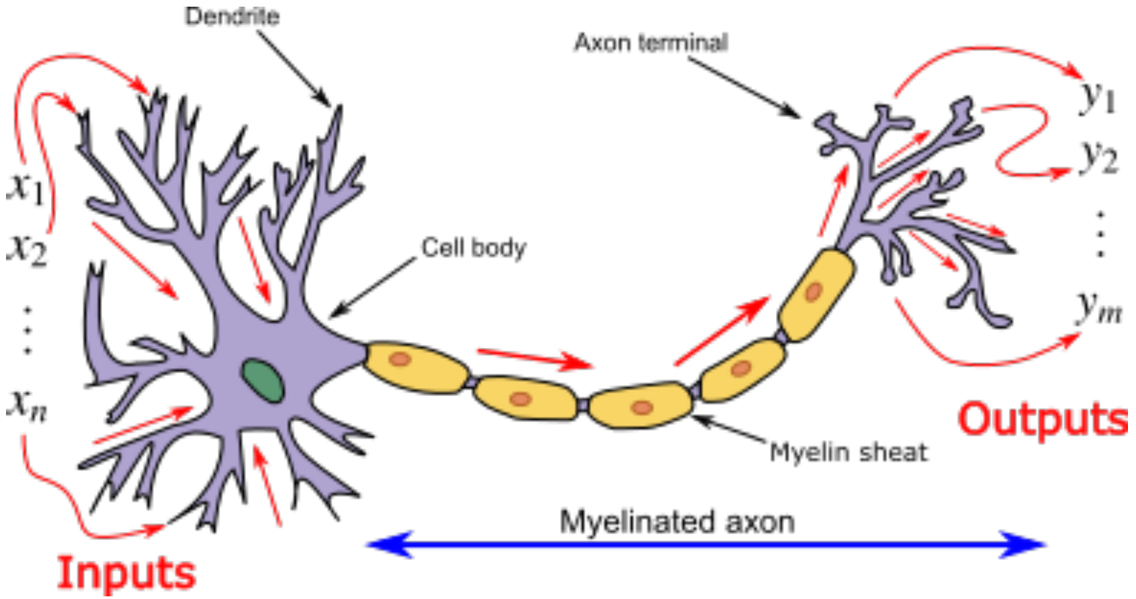



APPLES
BANANAS
ORANGES

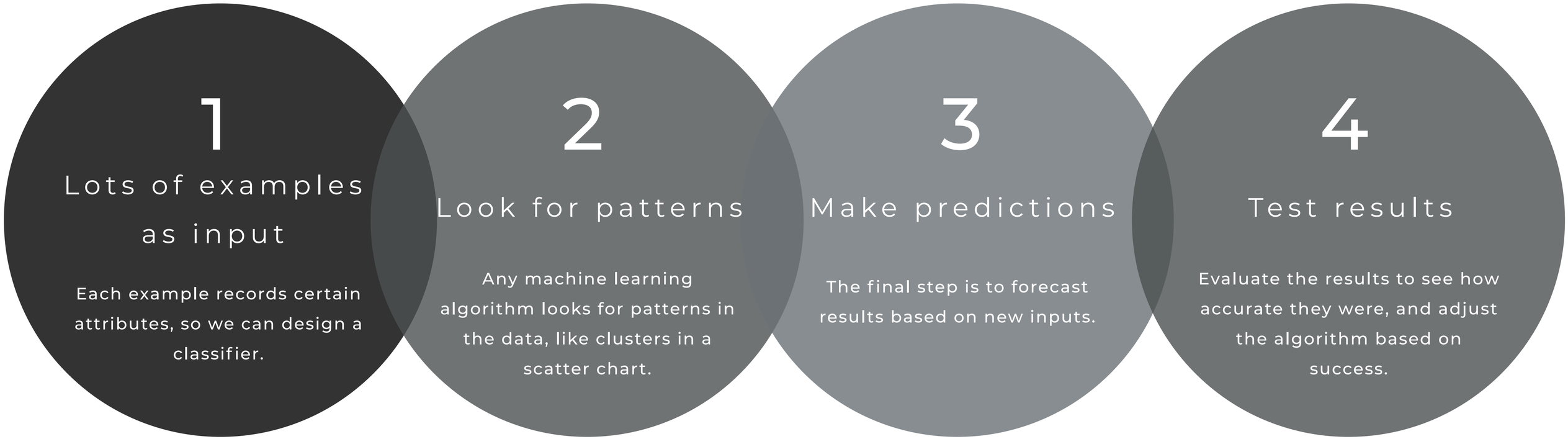


https://youtu.be/aircAruvnKk?t=303


Hidden
Input
Output


Dendrite
Axon terminal
x_1
x_2
x_n
Cell body
y_1
y_2
y_m
Outputs
Myelin sheat
Myelinated axon
Inputs

Machine Learning Process



How does a machine "learn"?

1

GATHERING lots of data, recording various attributes

The more data you have, the more accurate your system will be.

2

TRAINING data to look for specific groupings, correlations in the data.

Look for how data clusters around certain areas, and what other pieces of data correlate to that.

3

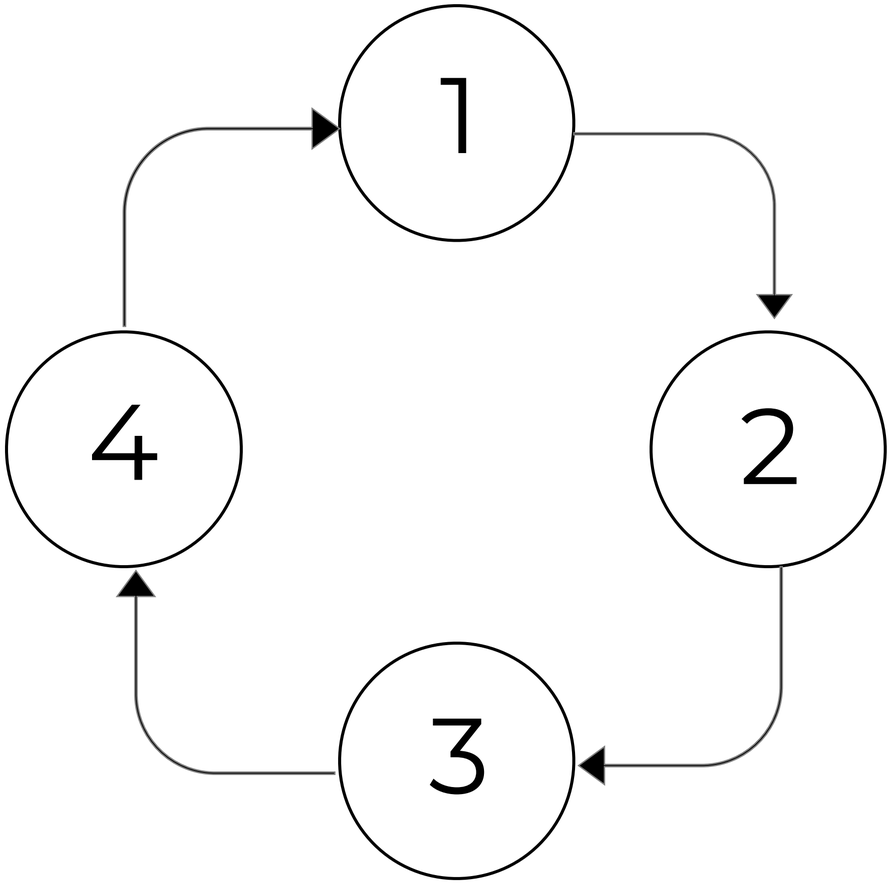
PREDICTING a new piece of datum based on similar characteristics.

Machine learning is about predicting outcomes based on historical data.

4

EVALUATING and testing effectiveness for improvement.

We evaluate how accurate our predictions are, and create a confidence score. And use those predictions to feed back into the system to improve it.



title	calories	protein	fat	sodium	rating
Lentil, Apple, and Turkey Wrap	426	30	7	559	2.500
Boudin Blanc Terrine with Red Onion Confit	403	18	23	1439	4.375
Potato and Fennel Soup Hodge	165	6	7	165	3.750
Spinach Noodle Casserole	547	20	32	452	3.125
Korean Marinated Beef	170	7	10	1272	4.375
Ham Persillade with Mustard Potato Salad and M...	602	23	41	1696	3.750
Yams Braised with Cream, Rosemary and Nutmeg	256	4	5	30	3.750
Banana-Chocolate Chip Cake With Peanut Butter ...	766	12	48	439	4.375
Beef Tenderloin with Garlic and Brandy	174	11	12	176	4.375
Peach Mustard	134	4	3	1394	3.125
Raw Cream of Spinach Soup	382	5	31	977	4.375
Sweet Buttermilk Spoon Breads	146	4	5	160	1.875
Mozzarella-Topped Peppers with Tomatoes and Ga...	107	5	7	344	5.000
Tuna, Asparagus, and New Potato Salad with Chi...	421	10	33	383	5.000
Asian Pear and Watercress Salad with Sesame Dr...	345	11	19	423	4.375

title	calories	protein	fat	sodium	rating
Lentil, Apple, and Turkey Wrap	426	30	7	559	2.500
Boudin Blanc Terrine with Red Onion Confit	403	18	23	1439	4.375
Potato and Fennel Soup Hodge	165	6	7	165	3.750
Spinach Noodle Casserole	547	20	32	452	3.125
Korean Marinated Beef	170	7	10	1272	4.375
Ham Persillade with Mustard Potato Salad and M...	602	23	41	1696	3.750
Yams Braised with Cream, Rosemary and Nutmeg	256	4	5	30	3.750
Banana-Chocolate Chip Cake With Peanut Butter ...	766	12	48	439	4.375
Beef Tenderloin with Garlic and Brandy	174	11	12	176	4.375
Peach Mustard	134	4	3	1394	3.125
Raw Cream of Spinach Soup	382	5	31	977	4.375
Sweet Buttermilk Spoon Breads	146	4	5	160	1.875
Mozzarella-Topped Peppers with Tomatoes and Ga...	107	5	7	344	5.000
Tuna, Asparagus, and New Potato Salad with Chi...	421	10	33	383	5.000
Asian Pear and Watercress Salad with Sesame Dr...	345	11	19	423	4.375

We're going to use these

title	calories	protein	fat	sodium	rating
Lentil, Apple, and Turkey Wrap	426	30	7	559	2.500
Boudin Blanc Terrine with Red Onion Confit	403	18	23	1439	4.375
Potato and Fennel Soup Hodge	165	6	7	165	3.750
Spinach Noodle Casserole	547	20	32	452	3.125
Korean Marinated Beef	170	7	10	1272	4.375
Ham Persillade with Mustard Potato Salad and M...	602	23	41	1696	3.750
Yams Braised with Cream, Rosemary and Nutmeg	256	4	5	30	3.750
Banana-Chocolate Chip Cake With Peanut Butter ...	766	12	48	439	4.375
Beef Tenderloin with Garlic and Brandy	174	11	12	176	4.375
Peach Mustard	134	4	3	1394	3.125
Raw Cream of Spinach Soup	382	5	31	977	4.375
Sweet Buttermilk Spoon Breads	146	4	5	160	1.875
Mozzarella-Topped Peppers with Tomatoes and Ga...	107	5	7	344	5.000
Tuna, Asparagus, and New Potato Salad with Chi...	421	10	33	383	5.000
Asian Pear and Watercress Salad with Sesame Dr...	345	11	19	423	4.375

to predict these

Step 1 - Separate our data

calories	protein	fat	sodium	rating
426	30	7	559	2.500
403	18	23	1439	4.375
165	6	7	165	3.750
547	20	32	452	3.125
170	7	10	1272	4.375
602	23	41	1696	3.750
256	4	5	30	3.750
766	12	48	439	4.375
174	11	12	176	4.375
134	4	3	1394	3.125
382	5	31	977	4.375
146	4	5	160	1.875
107	5	7	344	5.000
421	10	33	383	5.000
345	11	19	423	4.375

Step 1 - Separate our data

calories	protein	fat	sodium	rating
426	30	7	559	2.500
403	18	23	1439	4.375
165	6	7	165	3.750
547	20	32	452	3.125
170	7	10	1272	4.375
602	23	41	1696	3.750
256	4	5	30	3.750
766	12	48	439	4.375
174	11	12	176	4.375
134	4	3	1394	3.125
382	5	31	977	4.375
146	4	5	160	1.875
107	5	7	344	5.000
421	10	33	383	5.000
345	11	19	423	4.375

Training Data

Testing Data

calories	protein	fat	sodium	rating
426	30	7	559	2.500
403	18	23	1439	4.375
165	6	7	165	3.750
547	20	32	452	3.125
170	7	10	1272	4.375
602	23	41	1696	3.750
256	4	5	30	3.750
766	12	48	439	4.375
174	11	12	176	4.375
134	4	3	1394	3.125

Training Data

We only use training data to build our model (classifier)

382	5	31	977	4.375
146	4	5	160	1.875
107	5	7	344	5.000
421	10	33	383	5.000
345	11	19	423	4.375

Testing Data

calories	protein	fat	sodium	rating
426	30	7	559	2.500
403	18	23	1439	4.375
165	6	7	165	3.750
547	20	32	452	3.125
170	7	10	1272	4.375
602	23	41	1696	3.750
256	4	5	30	3.750
766	12	48	439	4.375
174	11	12	176	4.375
134	4	3	1394	3.125

Training Data

We only use training data to build our model (classifier)

382	5	31	977	4.375
146	4	5	160	1.875
107	5	7	344	5.000
421	10	33	383	5.000
345	11	19	423	4.375

Testing Data

Then we input some of our test data to see if we get the correct result.

calories	protein	fat	sodium
426	30	7	559
403	18	23	1439
165	6	7	165
547	20	32	452
170	7	10	1272
602	23	41	1696
256	4	5	30
766	12	48	439
174	11	12	176
134	4	3	1394

X_train

rating
2.500
4.375
3.750
3.125
4.375
3.750
3.750
4.375
4.375
3.125

X_test

Training Data

382	5	31	977
146	4	5	160
107	5	7	344
421	10	33	383
345	11	19	423

y_train

4.375
1.875
5.000
5.000
4.375

y_test

Testing Data