

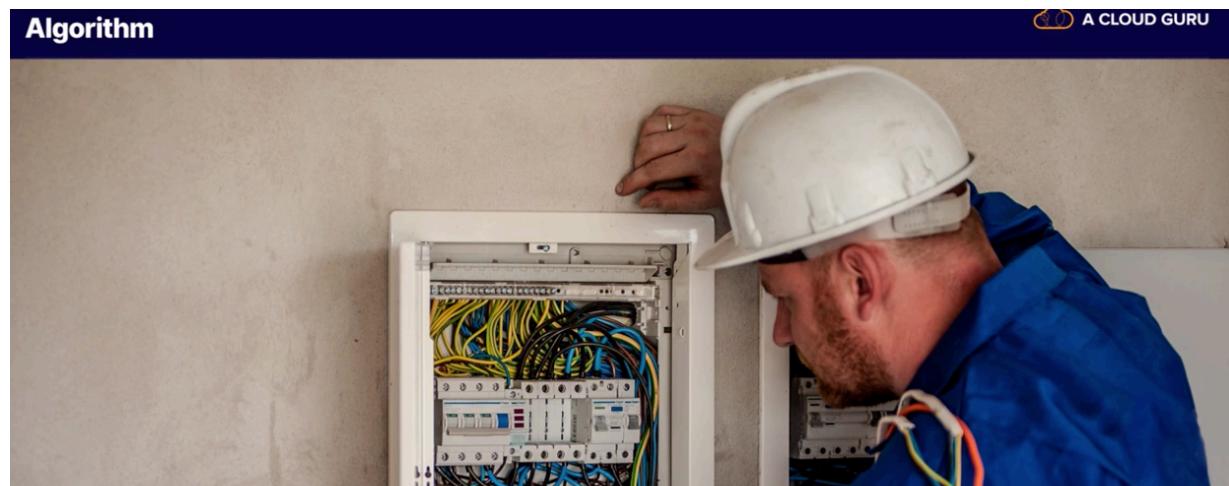
Cloud Guru - 6 - Algorithms

<https://acloud.guru/course/aws-certified-machine-learning-specialty/learn/1efec92d-271f-f097-0c26-236b41ab42b1/f4cd49ae-be82-06a0-7672-7edc64126b7f/watch?backUrl=~2Fcourses&backUrl=~2Fcourses&backUrl=~2Fcourses,~2Fcourses>

Algorithm vs Heuristic

Algorithm	Heuristic
Set of steps to follow to solve a specific problem intended to be repeatable with the same outcome.	A mental short-cut or “rule of thumb” that provides guidance on doing a task but does not guarantee a consistent outcome.

Algorithm = troubleshooting steps to narrow down possibilities to the correct one
Ex: for electrical pb - go through a set of steps to see where the fault is



Heuristic = educating guess

Ex: We try a new pizza place. It could be a good or bad pizza, but we don't know. We are not going to try to see how they make the tough, what is their recipe, match against what we know,... We are willing to take a chance here
We use heuristic a lot in our life



When to use algorithm - trying to **drive out Bias**

Trying to remove assumptions and past experience.

In contrast, computers can't make these assumptions. They don't have that intuition required by a heuristic.

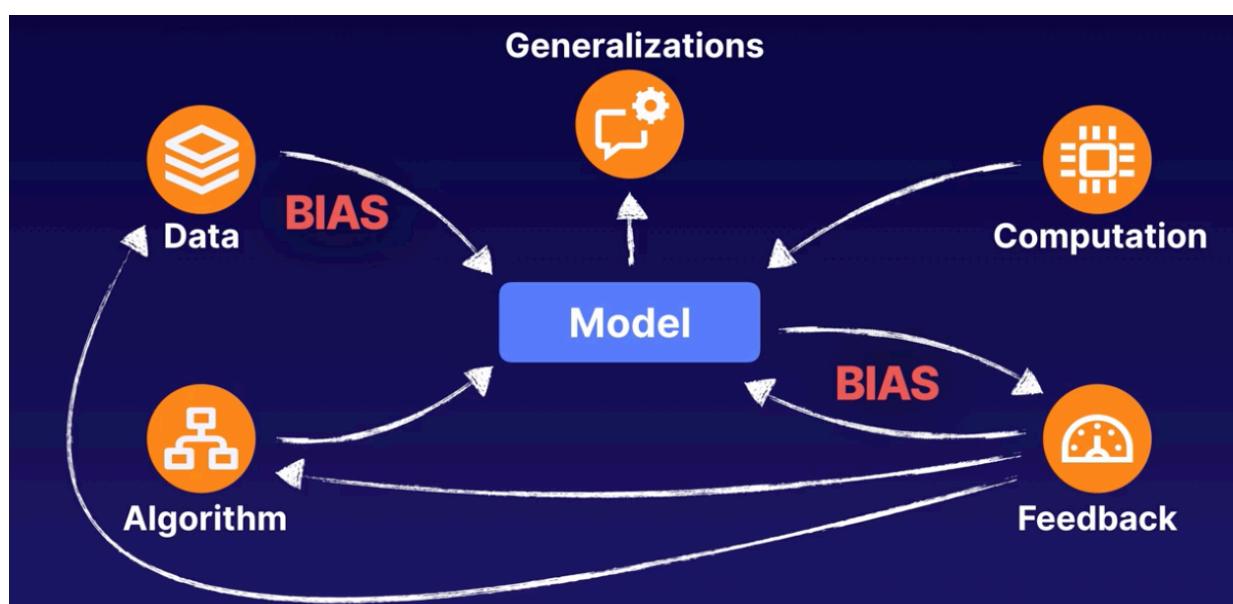
=> **Algorithm** is a **finite process with known input and expected outputs**

We can use this in ML as we want repeatable generalization.

Now **Bias** can still creep in into ML process:

- by data we select when training or testing, or when we exclude a chunk of data
- feedback loop

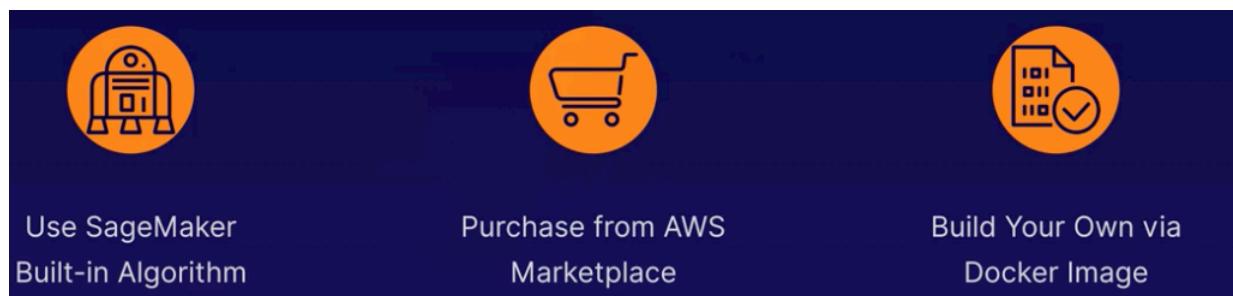
We might skew the model to fulfill our assumptions in the first place



3 types of Learning:

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Training	Training Data and Testing Data	No Training	How to Maximize Reward
Discrete	Classification	Clustering	Simulation-based Optimization
Continuous	Regression	Reduction of Dimensionality	Autonomous Devices

Algorithms:



Regression

Example: linear equation: $y = 2x + 3$

Given x , we can figure out y and vice-versa

We can plot a line too.

Seeing values on that line, we can estimate/predict what is the y value further down on that line



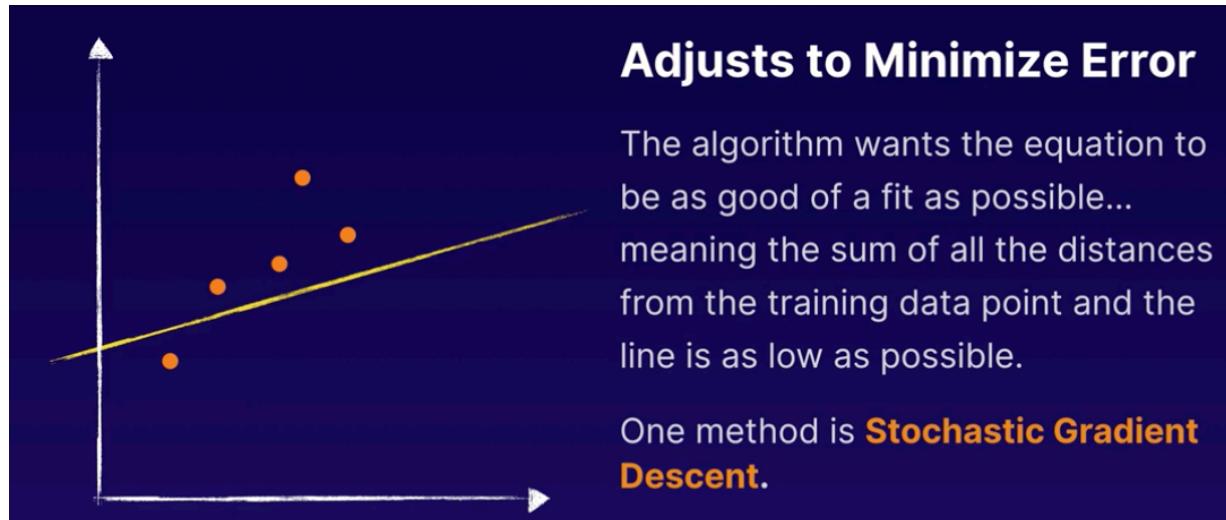
Linear Learner Algorithm



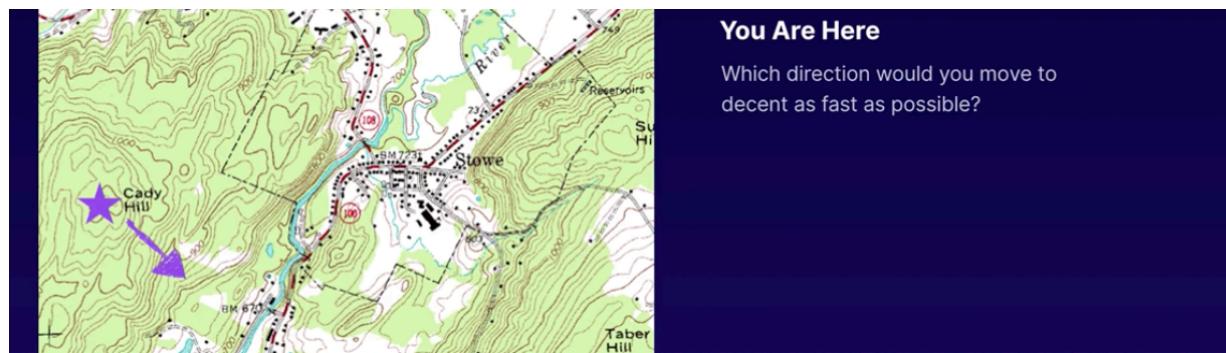
Linear models are supervised learning algorithms for regression, binary classification or multiclass classification problems. You give the model labels (x,y) with x being high-dimensional vector and y is a numeric label. The algorithm learns a linear function, or, for classification problems, a linear threshold function, and maps a vector x to an approximation of label y .



To use this algorithm you need a number or list of numbers which yields some other number...the answer you're after. You can use it to predict a specific value or a threshold for grouping purposes.



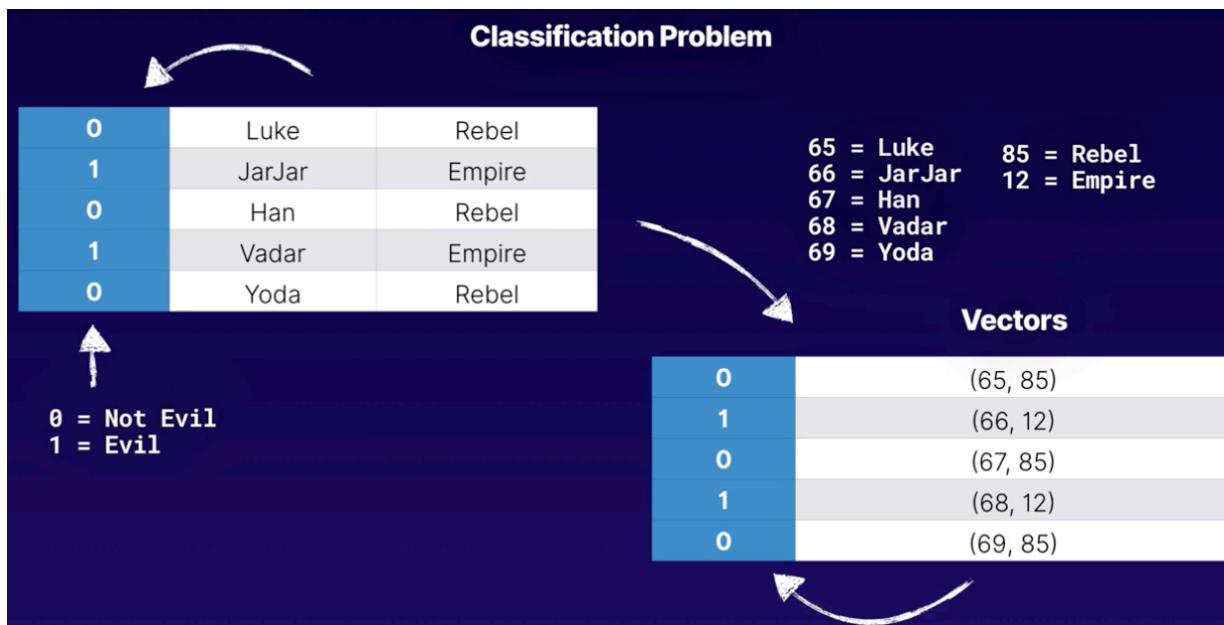
Ex of topographic map to explain Stochastic Gradient Descent
Map is the error.



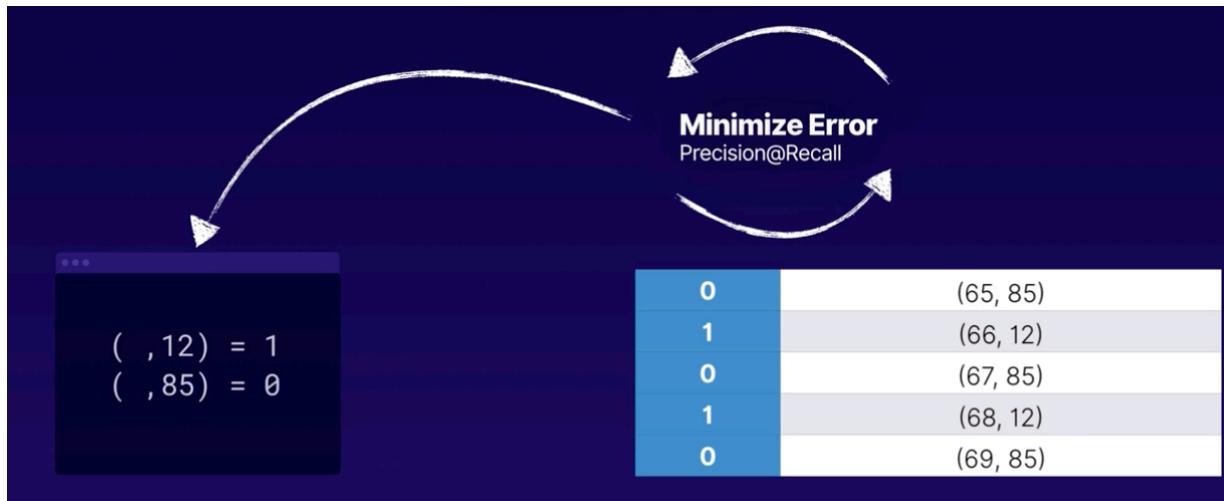
Another idea - idea coming down
Some may be stuck in pools, some may make their way all the way to the ocean



For linear regression, we need numbers -> convert the data and end up with vectors



Linear learning will learn through trial/error process, is that the 2nd value in the vector is the most accurate to predict outcome, whether somebody is a 0 or 1



1	Very Flexible Linear Learner can be used to explore different training objectives and choose the best one. Well suited for discrete or continuous inferences.	2 Built-in Tuning Linear Learner algorithm has an internal mechanism for tuning hyperparameters separate from the automatic model tuning feature.	3 Good First Choice If your data and objective meets the requirements, Linear Learner is a good first choice to try for your model.
---	---	--	--

Examples when to use Linear Learner Algorithm - Use Cases:

- **Predict quantitative value based on given numeric input**

Example: Based on last five years ROI from marketing spend, what can we expect to be this year's ROI?

- **Discrete Binary Classification Problems**

Example: Based on past customer response, should I mail this particular customer? Yes or No?

- **Discrete Multiclass Classification Problems**

Example: Based on past customer response, how should I reach this customer? Email, Direct Mail, or Phone Call?

It works well when we have good continuous data.

With sparse dataset, need to adjust this with Factorization Machine

(34, 4, 56, 34, 6, 73, 3, 4, 12, 65, 63...)

Linear Learner



(34, 4, , 34, 6, , 3, , 12, 65, , ...)

Factorization Machines

▲ ▲ ▲ ▲

a sparse dataset



Factorization Machines Algorithm

General purpose supervised learning algorithm for both binary classification and regression. Captures interaction between features with high dimensional sparse datasets.

SUPERVISED



To use this algorithm you need a number or list of numbers which yields some other number...the answer you're after. You can use it to predict a specific value or a threshold for placing into one of two groups. It is a good choice when you have "holes" in your data.

Things to know about Factorization Machines Algorithm

1	Considers only pair-wise features. Amazon SageMaker's implementation of factorization machines will only analyze relationships of two pairs of features at a time.	4	Really needs LOTS of data. To make up for the sparseness, needs lots of data. Recommended dimension of the input feature space is between 10,000 and 10,000,000.
2	CSV is not supported. CSV is not a good choice for sparse dimensions. File and Pipe mode training are supported using recordIO-protobuf format with Float32 tensors.	5	CPUs rock sparse data better. AWS recommends CPUs with factorization machines for the most efficient experience.
3	Doesn't work for Multiclass Problems. Factorization Machines algorithm can be run in either binary classification mode or regression mode.	6	Don't perform well on dense data. Other algorithms are much more performant when you have a full set of data.

Example: Movie recommendation engine
 4 movies watched by 4 people.
 Nobody has watched all movies —> perfect situation for a recommender engine

Objective: We want to find a movie that Dante is going to like

Dante	(1, , , 0)	$0 = \text{Did not like the movie}$ $1 = \text{Liked the movie}$
Bob	(, 1, , 0)	"Clerks"
Jay	(, , 1,)	"Mallrats"
Veronica	(1, 1, 0,)	"Dogma"
Caitlin	(1, 0, , 1)	"Clerks 2"

=> do one hot encoding that creates a matrix for us to evaluate with movies and people that like them

p_Bob	p_Jay	p_Veronica	p_Caitlin	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2
1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1
0	0	1	0	1	0	0	0
0	0	1	0	0	1	0	0
0	0	0	1	1	0	0	0
0	0	0	1	0	0	0	1

1,0,0,0,0,1,0,0
 0,1,0,0,0,0,0,1
 0,0,1,0,1,0,0,0 X
 0,0,1,0,0,1,0,0
 0,0,0,1,1,0,0,0
 0,0,0,1,0,0,0,1

To note if we had 100,000 viewers and 10,000 films, we would end up with a really huge matrix, filled for the most with 0

That is a huge waste and why Factorization Machine does not support CSV => rather they use recordIO protobuf with Float32 Tensor

p_Bob	p_Jay	p_Veronica	p_Caitlin	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2
1	0	0	0	0	1	0	0
0	1	0	0	0	0	0	1
0	0	1	0	1	0	0	0
0	0	1	0	0	1	0	0
0	0	0	1	1	0	0	0
0	0	0	1	0	0	0	1

p_Dante	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2
1	1	0	0	0

(1, , , 0)

Start building a profile of Dante's movies preferences vs other's preference, and summing that

p_Bob	p_Jay	p_Veronica	p_Caitlin	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2		
1	0	0	0	0	1	0	0		
0	1	0	0	0	0	0	1		
0	0	1	0	1	0	0	0		
0	0	1	0	0	1	0	0		
0	0	0	1	1	0	0	0		
0	0	0	1	0	0	0	1		
p_Dante	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2	Sum	m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2
1	1	0	0	0	-4	-1	1	0	0
						-1	0	0	1
						0	0	0	0
						-1	1	0	0
						0	0	0	0
						-1	0	0	1
						-4	2	0	2

We want anything positive

m_Clerks	m_Mallrats	m_Dogma	m_Clerks_2
-1	1	0	0
-1	0	0	1
0	0	0	0
-1	1	0	0
0	0	0	0
-1	0	0	1
Sum	-4	2	Nope
Dante	(1, , , 0)		
I recommend “Mallrats”.			

Use case for Factorization Machine Algorithm:

- **High Dimensional Sparse Data Sets**

Example: Click-stream data on which ads on a webpage tend to be clicked given known information about the person viewing the page.

- **Recommendations**

Example: What sort of movies should we recommend to a person who has watched and rated some other movies?

Clustering

Unsupervised - no need for training data

Unsupervised algorithms that aim to group things such that they are with other things more similar than different.

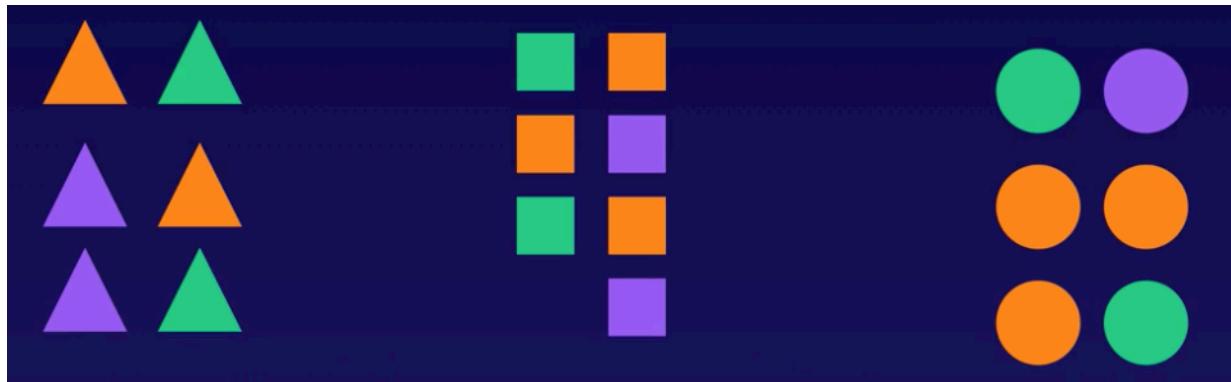


Specify which column a clustering algorithm need to pay attention to to group the data together

Use the number of sides to cluster!



Clustering
Algorithm



We could also cluster based on color



Popular clustering algorithm: K-Means

K-Means



Unsupervised algorithm that attempts to find discrete groupings within data, where members of a group are as similar as possible to one another and as different as possible from members of other groups. The Euclidean distance between these points represents the similarity of the corresponding observations.

UNSUPERVISED



K-Means will take in a list of things with attributes. You specify which attributes indicate similarity and the algorithm will attempt to group them together such that they are with other similar things. "Similarity" is calculated based on the distance between the identifying attributes.

SageMaker K-means - things to know:

1

Expects Tabular Data

Rows represent the observations that you want to cluster and the columns represent attributes of the observations.

4

CPU Instances Recommended

GPU instances can be used but SageMaker's K-Means can only use one GPU.

2

Define the Identifying Attributes.

You must know enough about the data set to propose attributes that will define similarity. If you have no idea, there are ways around this too!

5

Training is Still a Thing

You want to be sure your model is still accurate and using the best identifying attributes. Your data just doesn't have labels.

3

SageMaker uses a Modified K-Means

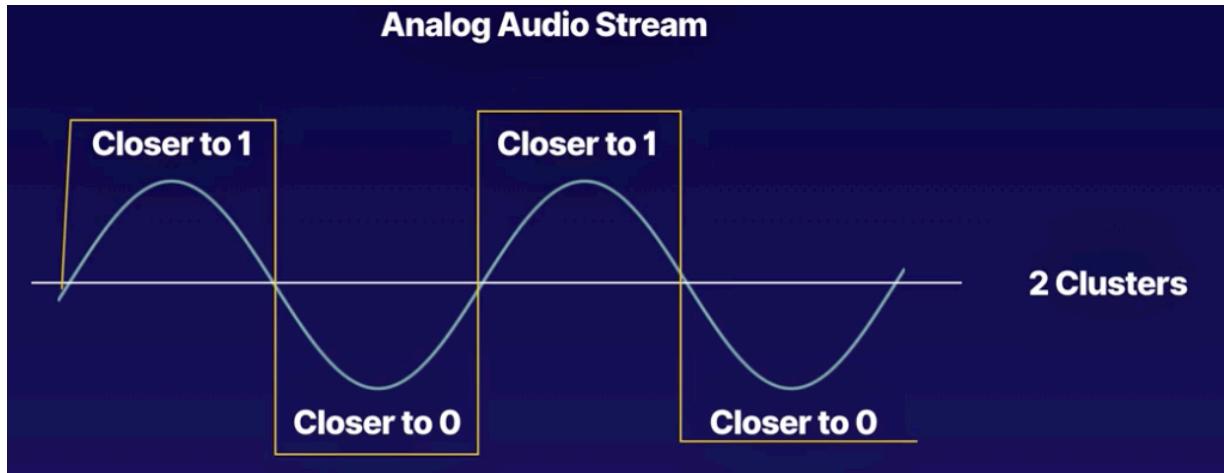
AWS uses a modified version of the web-scale K-Means algorithm, which it claims to be more accurate.

6

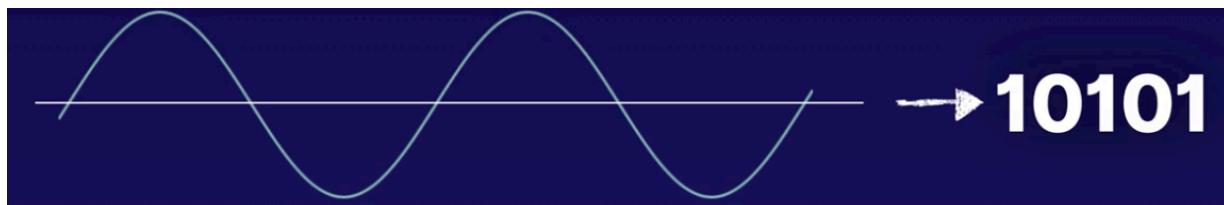
You Define Number of Features and Clusters

You must define the number features for the algorithm to analyze and the number of clusters you want.

Example of application: audio signal processing



Translate audio signal into digital signal with 0 and 1



The digitization process clips a bit of the fidelity from the analog



Other example: cluster the MNIST dataset
Store coordinates of white/black pixels and build vectors

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

Classification



Popular classification algorithm in SM: K-Nearest Neighbor

K-Nearest Neighbor



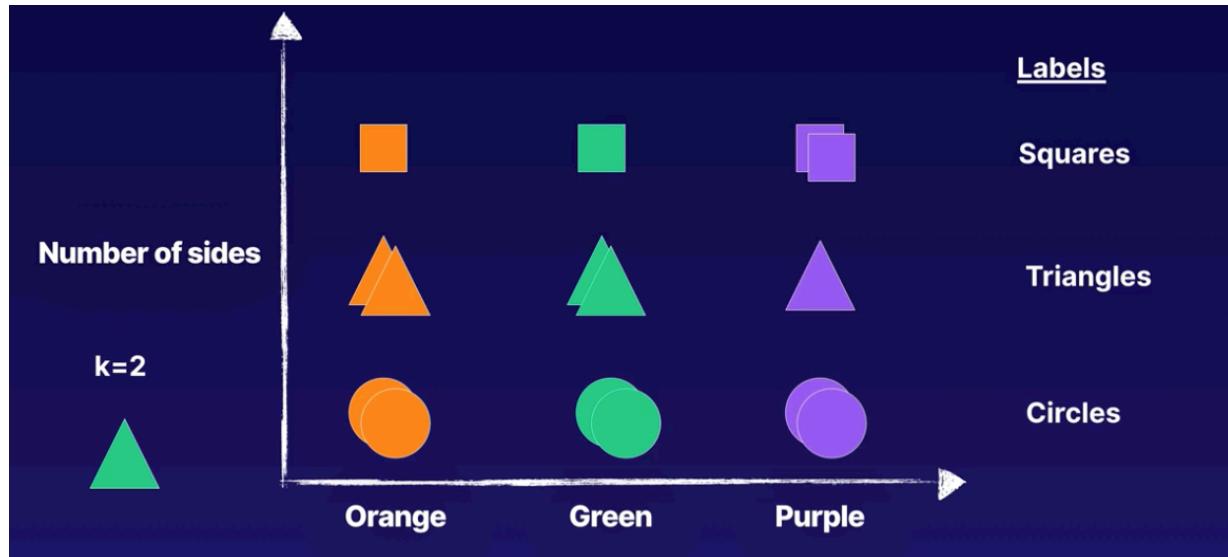
SUPERVISED

An index-based, non-parametric method for classification or regression. For classification, the algorithm queries the k points that are closest to the sample point and returns the most frequently used label as the predicted label. For regression, the algorithm queries the k closest points to the sample point and returns the average of the feature values as predicted value.



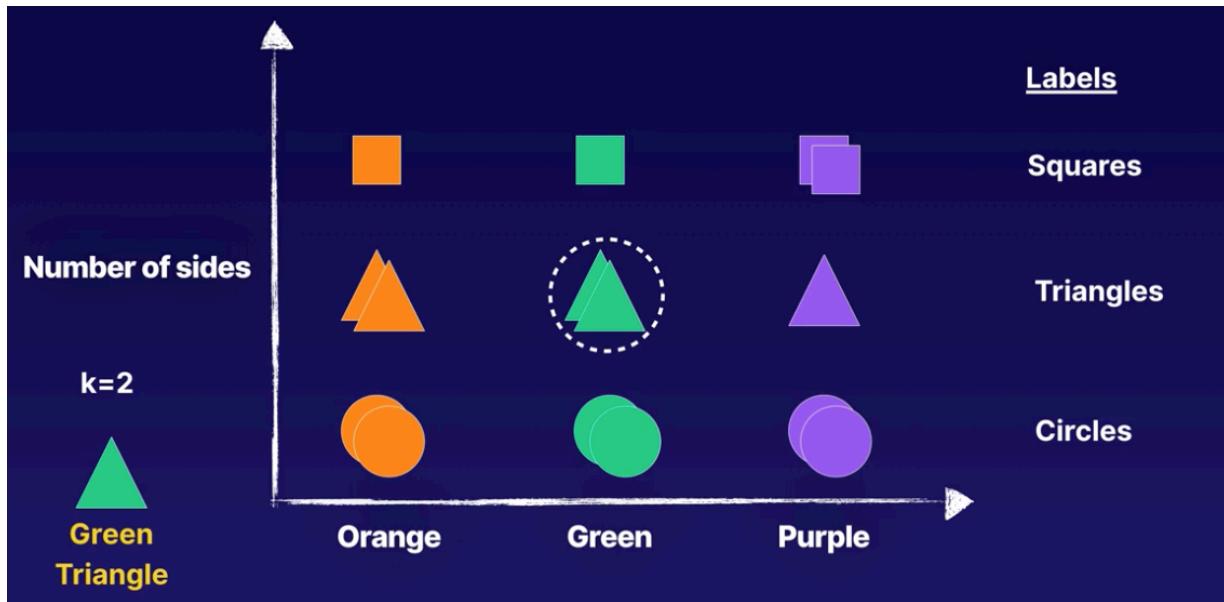
Predicts the value or classification based on that which you are closest. It can be used to classify or to predict a value (average value of nearest neighbors).

Example:



Hyper parameter: $k = 2$

Find the 2 nearest neighbors to this input shape (a triangle)



- | | | |
|--|--|--|
| 1
You choose the number of "neighbors".
You include a value for k , or in other words, the number of closest neighbors to use for classifying. | 2
KNN is a <i>lazy</i> algorithm.
Does not use training data points to generalize but rather uses them to figure out who's nearby. | 3
The training data stays in memory.
KNN doesn't "learn" but rather uses the training dataset to decide on similar samples. |
|--|--|--|

Use cases for K-nearest neighbors:

- **Credit Ratings**

Example: Group people together for credit risk based on attributes they share with others of known credit usage.

- **Product Recommendations**

Example: Based on what someone likes, recommend similar items they might also like.

K-nearest neighbor is a bit a matter of stereotyping
So be aware of Bias

Example when classification has gone wrong:

Redlining

The practice of literally drawing lines around neighborhoods and classifying those as:

- A - Best
- B - Still Desirable
- C - Definitely Declining
- D - Hazardous

Private and public entities would then use those redline maps to deny home loans, insurance, and other services for those less desirable areas – regardless of the qualifications of the applicant.

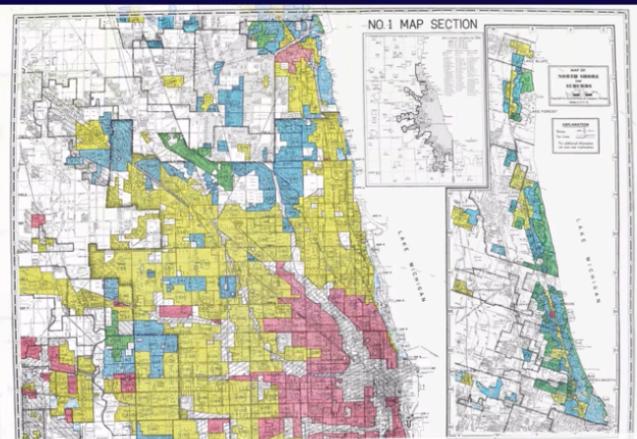
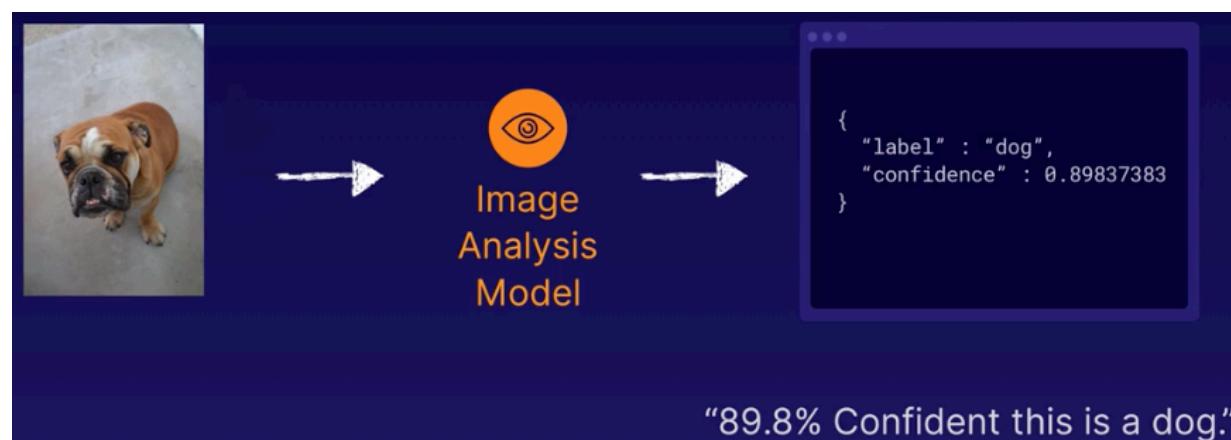
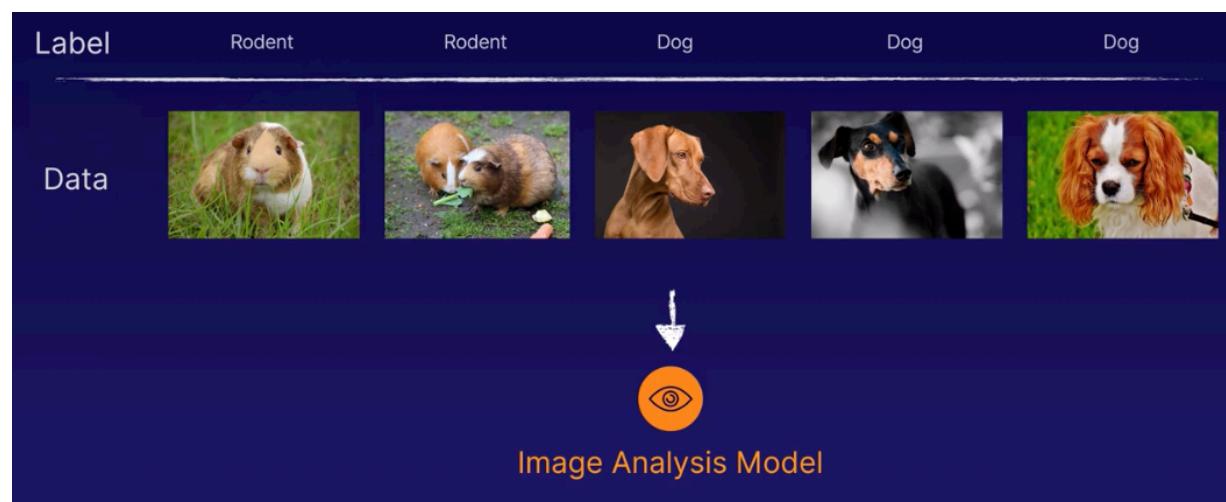
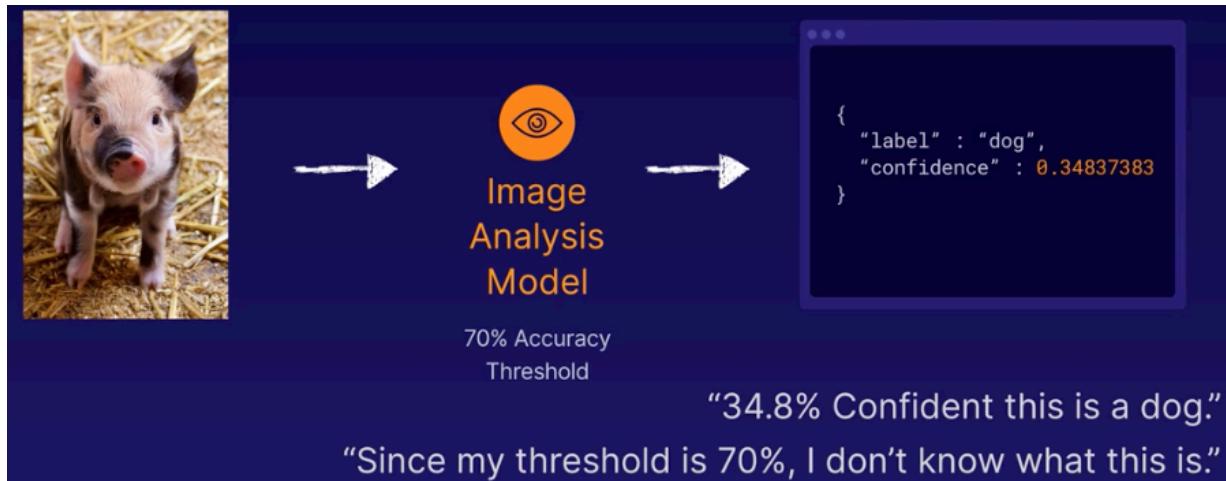


Image Analysis

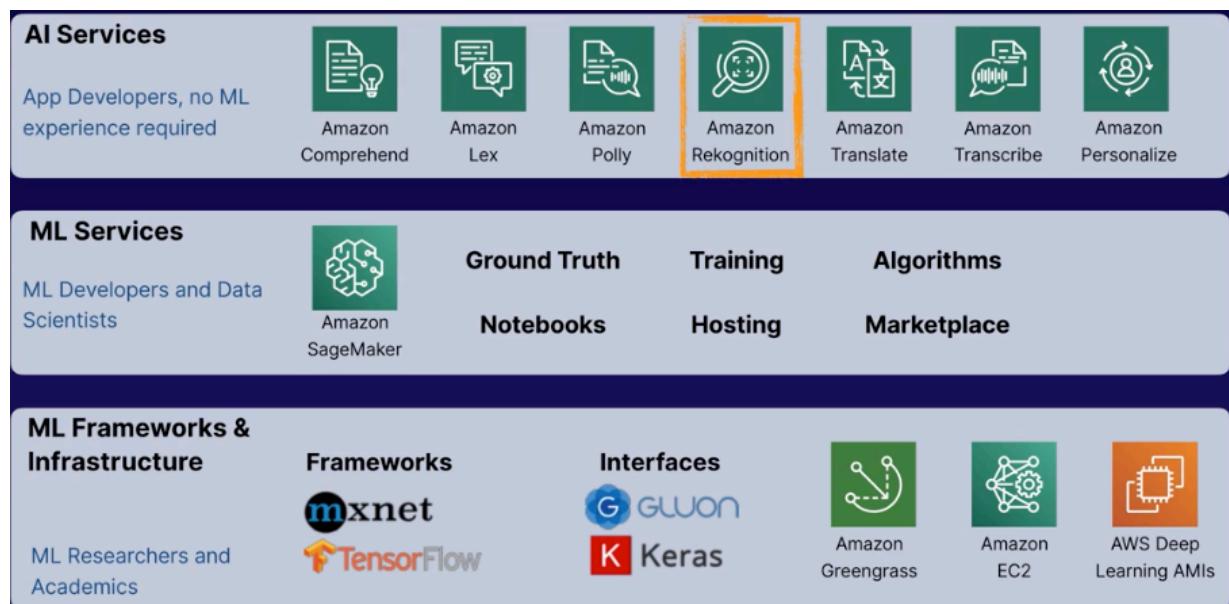
It's another form of classification, looking at pictures or images.



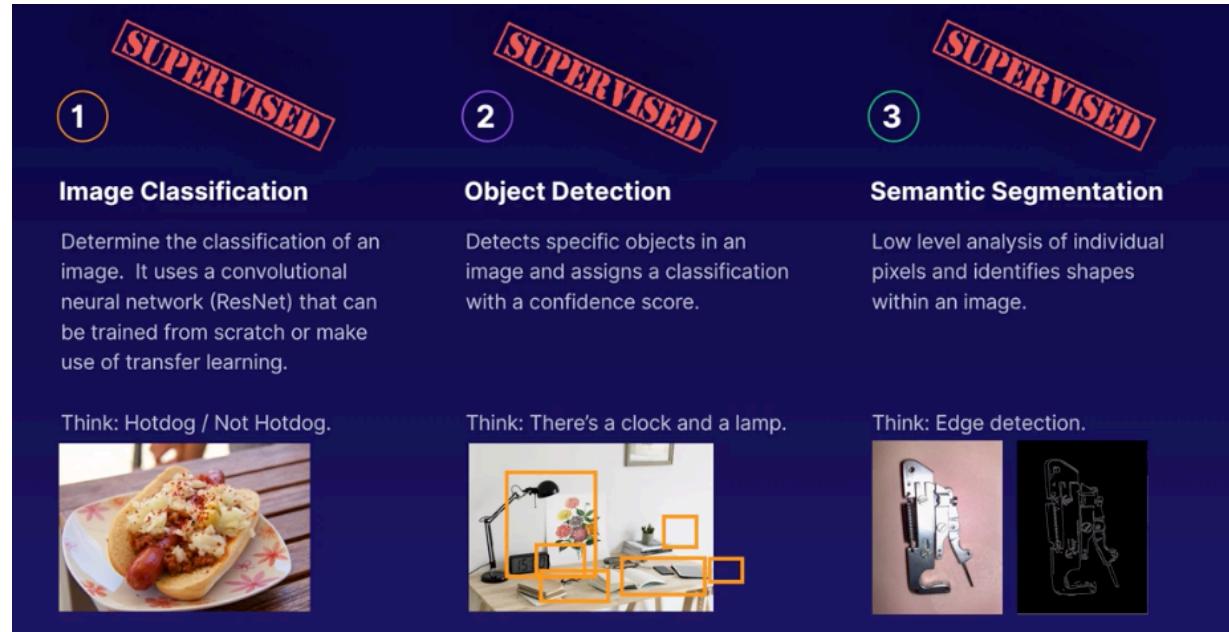
If we pass an image of a pig that we haven't trained the model on, confidence will be lower



AWS has a service doing a lot of that Rekognition



3 Image Analysis algorithms from Amazon



This is an example of Transfer Learning
 Leveraging ImageNet: huge DB of images

IM+GENET

dog
14,197,122 images, 21841 synsets indexed

SEARCH

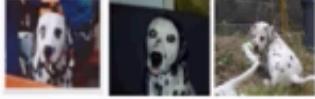
Home | About | Explore | Download

Not logged in. Login | Signup

ImageNet is under maintenance. Only ILSVRC synsets are included in the search results. Rank by Popularity Depth

Synset: Newfoundland, Newfoundland dog

 Definition: a breed of very large heavy dogs with a thick coarse usually black coat; highly intelligent dogs and vigorous swimmers; developed in Newfoundland.
Popularity percentile: 70%
Depth in WordNet: 8

Synset: dalmatian, coach dog, carriage dog

 Definition: a large breed having a smooth white coat with black or brown spots; originated in Dalmatia.
Popularity percentile: 65%
Depth in WordNet: 8

Synset: Eskimo dog, husky

 Definition: breed of heavy-coated Arctic sled dog.
Popularity percentile: 62%
Depth in WordNet: 9

Synset: German shepherd, German shepherd dog, German police dog, alsatian

 Definition: breed of large shepherd dogs used in police work and as a guide for the blind.
Popularity percentile: 61%
Depth in WordNet: 10

Synset: African hunting dog, hyena dog, Cape hunting dog, Lycaon pictus

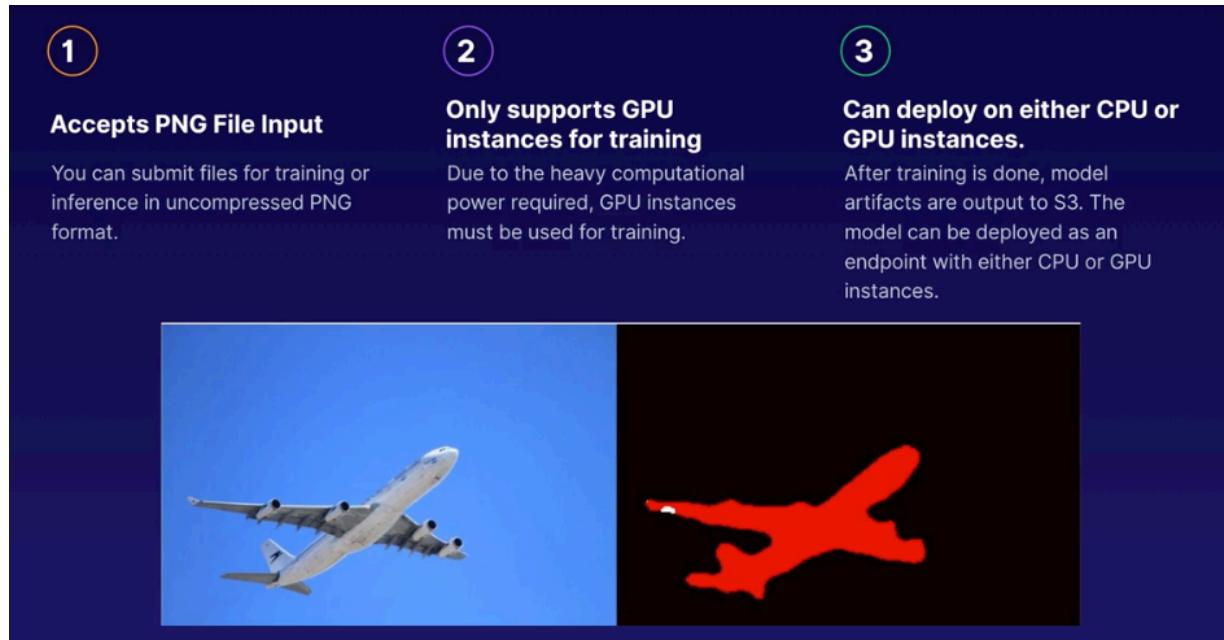
 Definition: a powerful doglike mammal of southern and eastern Africa that hunts in large packs; now rare in settled areas.
Popularity percentile: 58%
Depth in WordNet: 13

Synset: dog-sled, dog sled, dog sleigh

 Definition: a sled pulled by dogs.
Popularity percentile: 50%
Depth in WordNet: 8

© 2016 Stanford Vision Lab, Stanford University, Princeton University support@image-net.org Copyright infringement

Semantic segmentation:



CityScapes - classification available to help train/test software riding models

Group	Classes
flat	road · sidewalk · parking* · rail track*
human	person* · rider*
vehicle	car* · truck* · bus* · on rails* · motorcycle* · bicycle* · caravan** · trailer**
construction	building · wall · fence · guard rail* · bridge* · tunnel*
object	pole · pole group* · traffic sign · traffic light
nature	vegetation · terrain
sky	sky
void	ground* · dynamic* · static*

Use Cases for Image Analysis

- **Image Metadata Extraction**

Example: Extract scene metadata from an image provided and store it in a machine-readable format.

- **Computer Vision Systems**

Example: Recognize orientation of a part on an assembly line and, if required, issue a command to a robotic arm to re-orient the part.

Anomaly Detection

Using Random Cut Forest



Random Cut Forest

Detect anomalous data points within a set, like spikes in time series data, breaks in periodicity or unclassifiable points. Useful way to find outliers when it's not feasible to plot graphically. RCF is designed to work with n-dimensional input.



Find occurrences in the data that are significantly beyond "normal" (usually more than 3 standard deviations) that could mess up your model training.

UNSUPERVISED

1

Gives an Anomaly Score to Data Points

Low scores indicate that a data point is considered "normal" while high scores indicate the presence of an anomaly.

2

Scales Well

RCF scales very well with respect to number of features, data set size and number of instances.

3

Does not benefit from GPU

AWS recommends using normal compute instances (ml.m4, ml.c5)

Logic behind Random Cut Forest



- **Quality Control**

Example: Analyze an audio test pattern played by a high-end speaker system for any unusual frequencies.

- **Fraud Detection**

Example: If a financial transaction occurs for an unusual amount, unusual time or from an unusual place, flag the transaction for a closer look.

Another anomaly Detection algorithm form SM: IP Insights

IP Insights



Learns usage patterns for IPv4 addresses by capturing associations between IP addresses and various entities such as user IDs or account numbers.

UNSUPERVISED



Can potentially flag odd online behavior that might require closer review.

1 Ingests Entity/IP address Pairs

Historic data can be used to learn baseline patterns

4 GPUs recommended for Training

Generally, GPUs are recommended but if the dataset is large, distributed CPU instances might be more cost effective.

2 Returns Inferences via a Score

When queried, the model will return a score that indicates how anomalous the entity/IP combination is, based on the baseline.

5 CPUs recommended for Inference

Does not require costly GPUs for normal inference activities.

3 Uses a Neural Network

Uses a neural network to learn latent vector representations for entities and IP addresses. (learns patterns...)

• Tiered Authentication Models

Example: If a user tries to log into a website from an anomalous IP address, you might dynamically trigger an additional two-factor authentication routine.

• Fraud Detection

Example: On a banking website, only permit certain activities if the IP address is unusual for a given user login.

