console.cloud.google.com

Machine Learning and Natural Language Processing on Social Media Data at Scale

Agenda

Apache Spark and BigQuery

Natural Language Processing with Spark

Using Pre-Trained and Custom Models

Motivation

What is Apache Spark?

Apache Spark

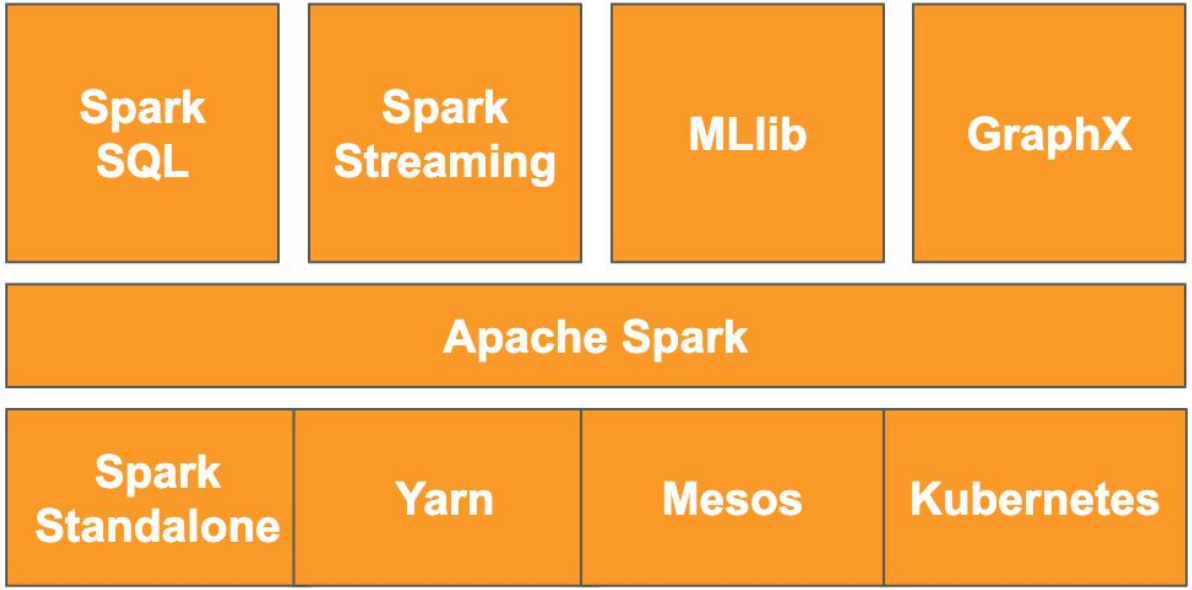
"Unified analytics engine for large-scale data processing"

Load data into memory in a distributed manner

Execute jobs across multiple machines

Available for Python, Scala, Java, R

Vast ecosystem of available tools / runtime environments


Spark SQL
Spark Streaming
MLlib
GraphX
Apache Spark
Spark Standalone
Yarn
Mesos
Kubernetes

What is Cloud DataProc?

Cloud Dataproc

Managed clusters for Apache Spark and Apache Hadoop

Deploy clusters in 90 seconds or less on average

Configurable with OSS

Resizable clusters

Tightly integrated with Google Cloud Platform ecosystem

What is PySpark?

What is BigQuery?

BigQuery

Fully managed, serverless Data Warehouse

Standard SQL

Automatically encrypted, durable, and highly available

Scales behind the scenes to accommodate reads and writes

What is Cloud Storage?

Cloud Storage

Unified object storage

Buckets

Huge storage capacity

High speed and availability

Connectors for various Google Cloud Platform Tools

bit.ly/pyspark-bigquery

git clone
https://github.com/bradmiro/cloud-dataproc

Spark vs BigQuery for Data Processing

Use Spark When...

You want to process your data with Python, Java, Scala, R

You want access to custom ML algorithms

You are working with unstructured data

You have a Spark cluster already set up

Your data is not in BigQuery

Use BigQuery When...

Your data is in BigQuery

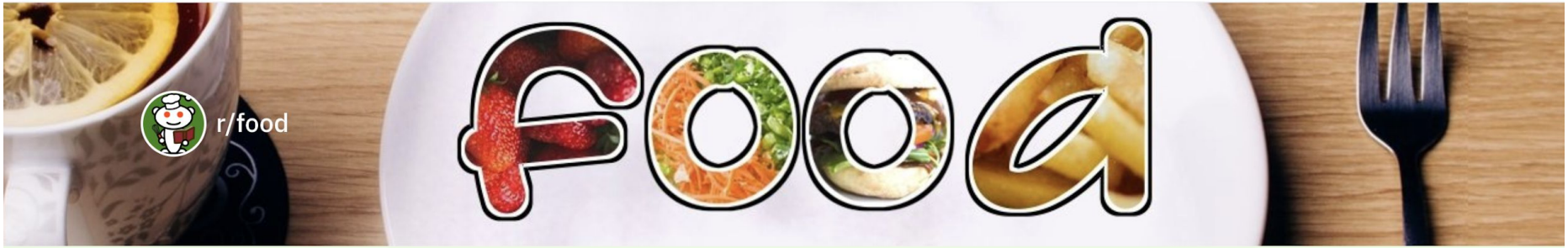
Your data is structured

Your needs can be fulfilled with SQL

You don't need to migrate your data

Intro to Analyzing Reddit

reddit r/food Search r/food pinke1993 132 karma



Posts Quick Search

VIEW SORT HOT

151

Posted by u/AutoModerator 16 days ago

r/Food Bi-weekly Discussion and Requests - June 10, 2019 Discussion

Please use this thread for discussions and requests.

Our rules will apply in this thread, if you see a rule breaking comment please report it or message the mods.

We have two threads per-week, Monday to Thursday & Friday to Sunday. Since this thread is likely to fill up quickly comment sorting is set to "new" (instead of "best" or "top").

89 Comments Give Award Share Save

COMMUNITY DETAILS

r/food

16.2m Members

2.5k Online

Cooking, restaurants, recipes, food network, foodies, talk about it here!

JOINED

CREATE POST

COMMUNITY OPTIONS

Resolving host...

What types of foods are people talking about?

Unsupervised Learning

(no labels)

Topic Modeling

[“hamburger”, “hotdog”, “ribs”, “baked beans”]

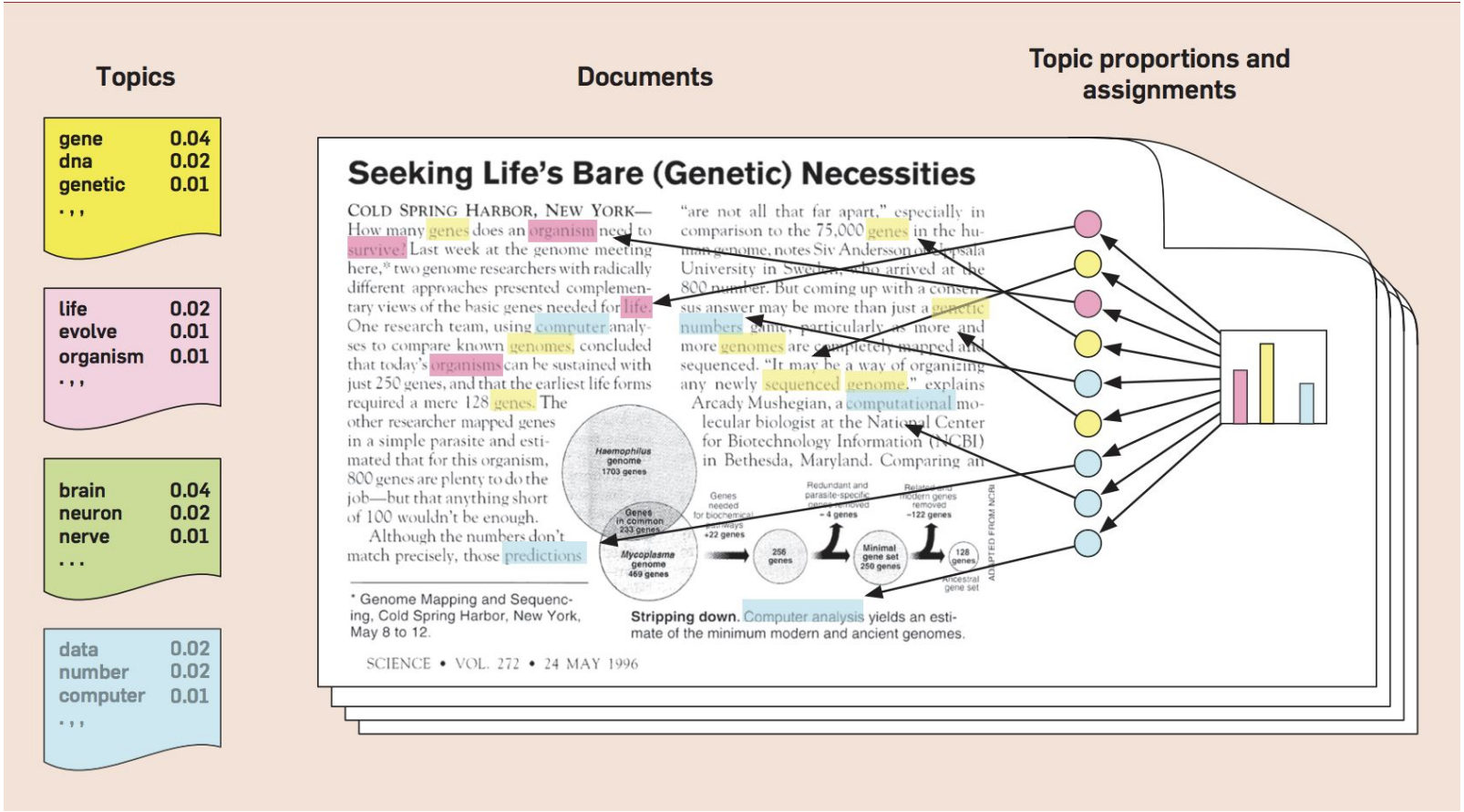
Topic Modeling

["hamburger", "hotdog", "ribs", "baked beans"]

Topic Modeling

["cupcake", "flour", "sugar", "muffin"]

Latent Dirichlet Allocation (LDA)



David Blei, Probabilistic Topic Models, Communications of the ACM | April 2012 | Vol. 55 | No. 4

Text Pre-Processing

Tokenization

“What time is it?” →
[“what”, “time”, “is”, “it”]

Stemming

[“cars”, “car’s”, “car”] →
[“car”, “car”, “car”]

Removing Stop Words

[“the”, “dog”, “is”, “big”] → [“dog”, “big”]

Term Frequency

number of times
a term occurs in a document

Inverse Document Frequency

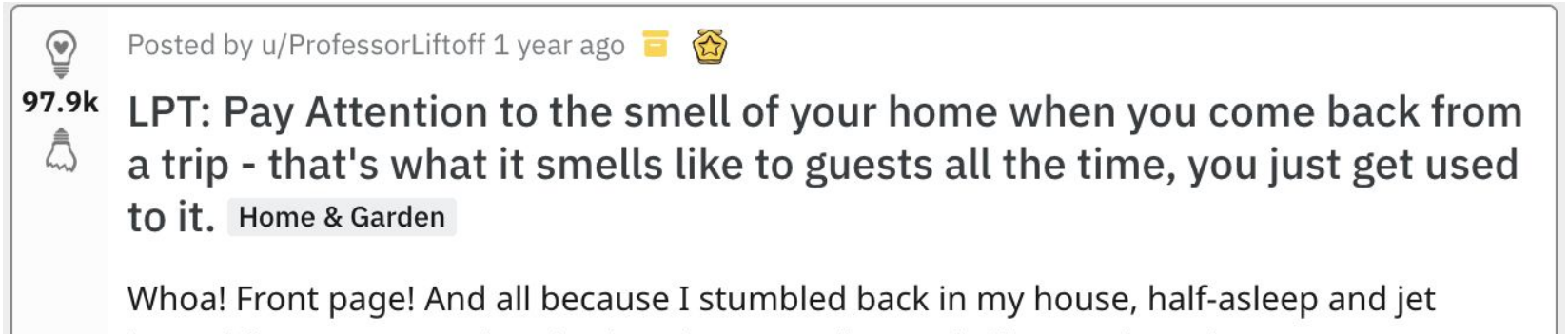
Inverse of % of document term appears in

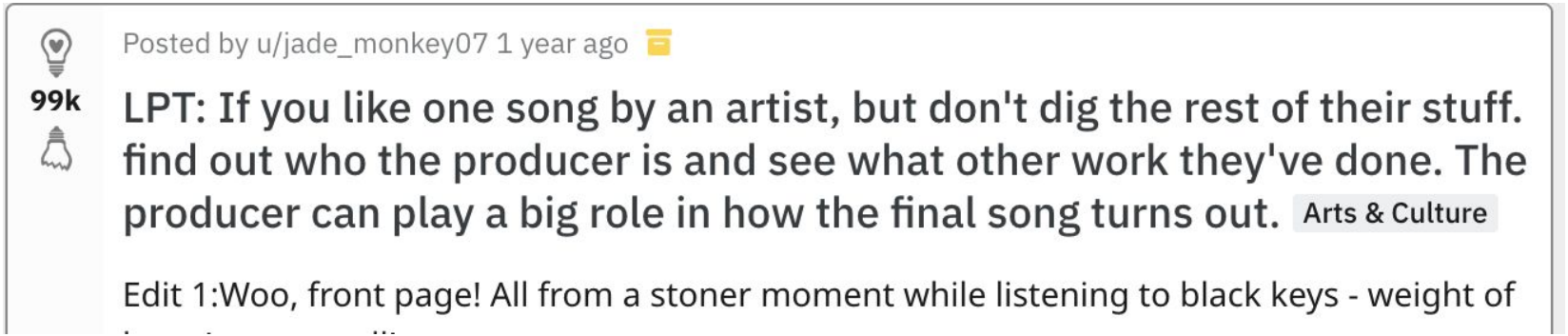
bit.ly/spark-nlp

NLP in Colab

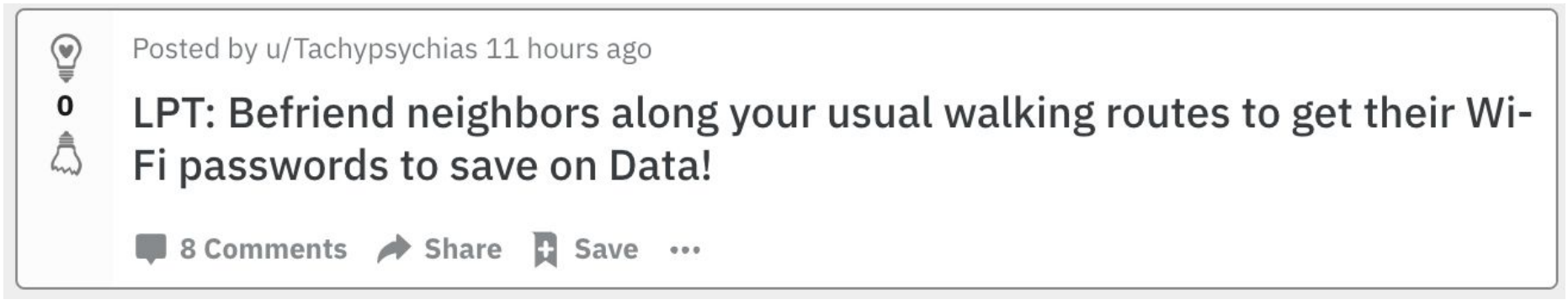
bit.ly/social-media-nlp

Top Life Pro Tips





Unpopular Life Pro Tips



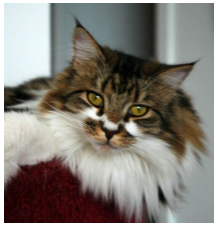
Supervised Learning

(no labels)

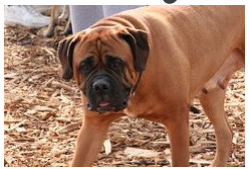
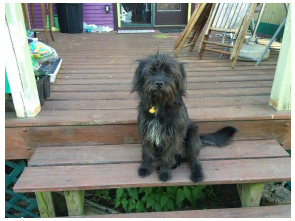
Labeled Training Dataset

Examples of cats





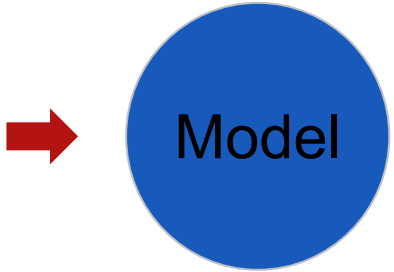

Examples of dogs


Google Cloud

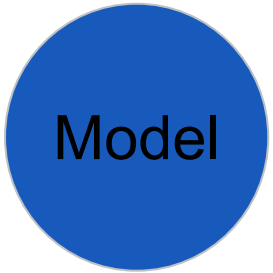
Training a Model

Machine Learning Algorithm

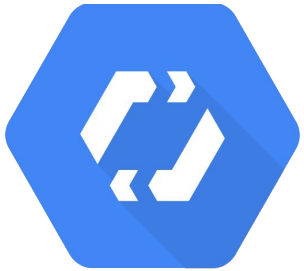


Model

Making Predictions

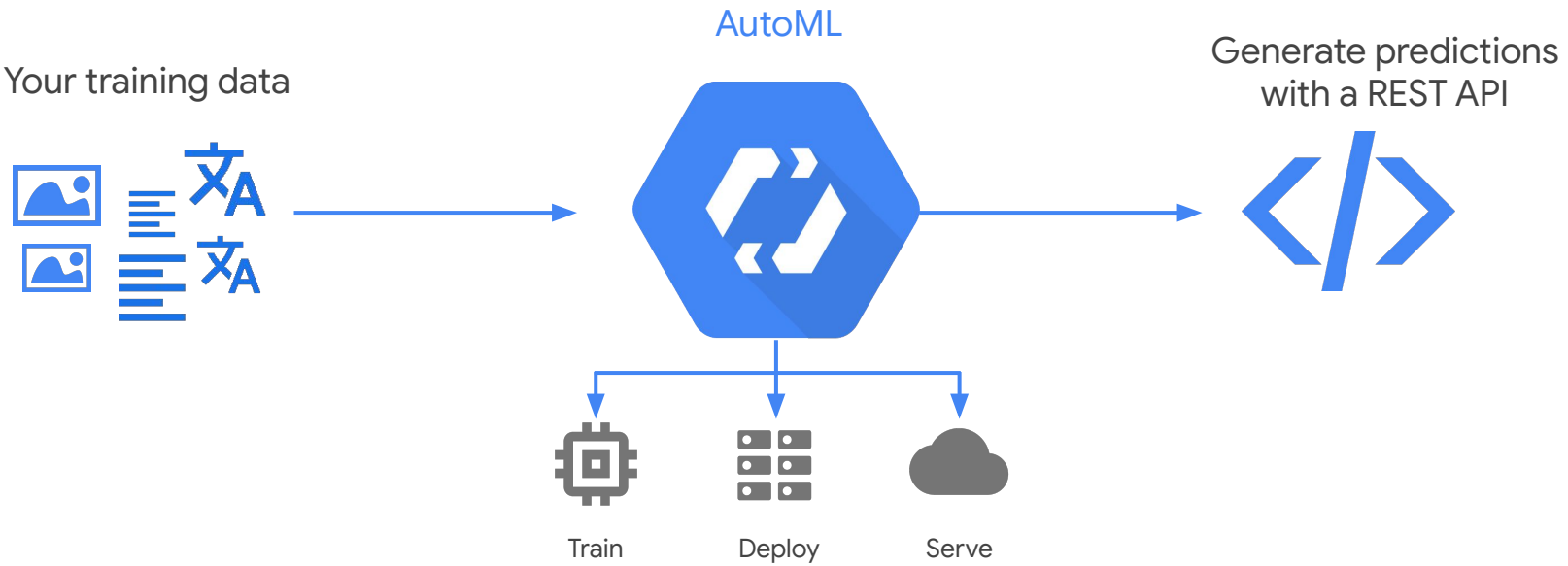


“Dog”



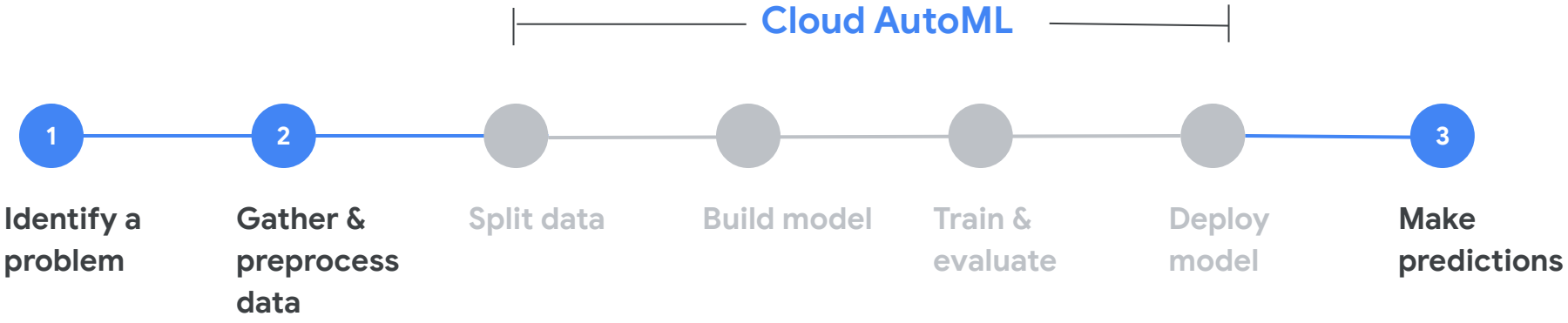
Cloud AutoML

What is Cloud AutoML?



Google Cloud

How can Cloud AutoML help?



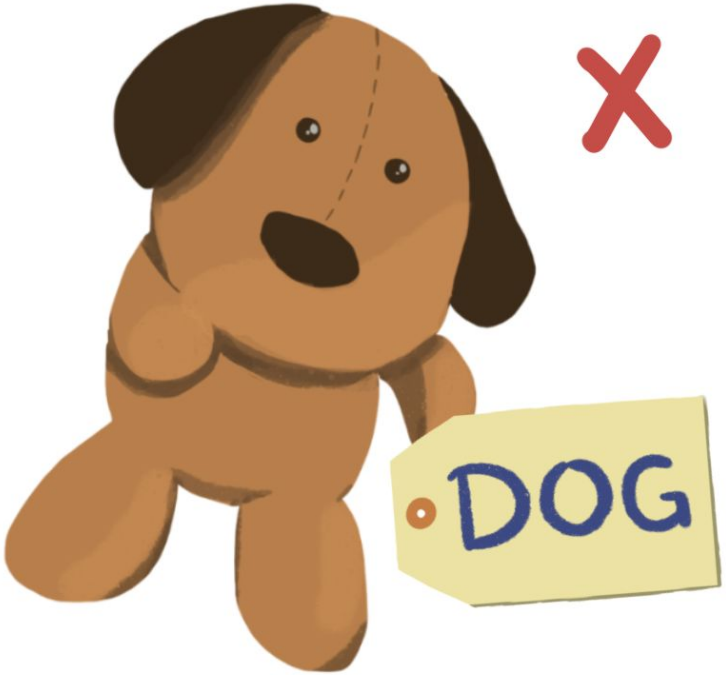
Evaluating Your Model

Dog classifier: dog or not dog

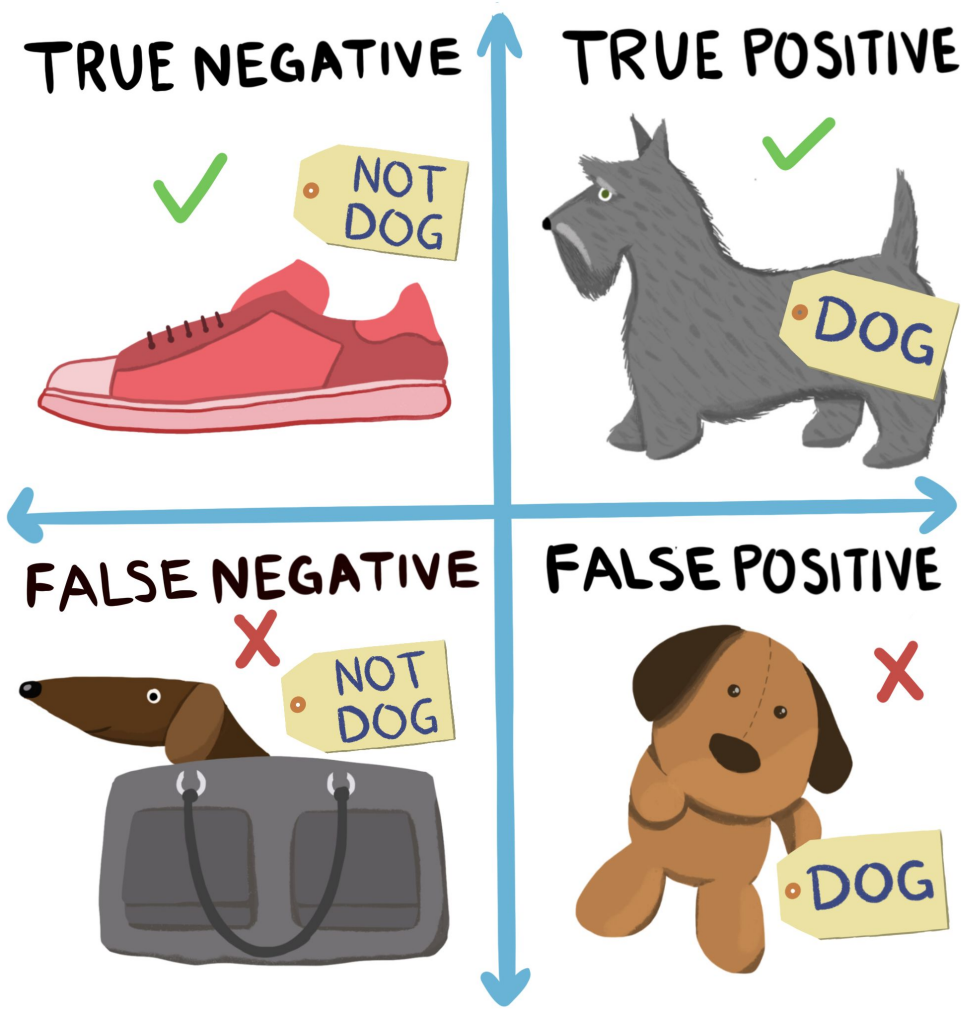
Accuracy: % correct predictions

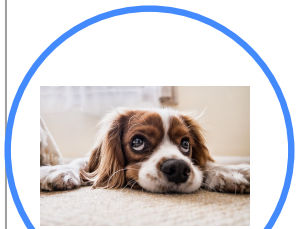
All errors are not equal

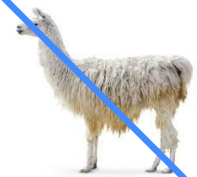
FALSE POSITIVE



FALSE NEGATIVE


TRUE NEGATIVE
NOT DOG
TRUE POSITIVE
DOG
FALSE NEGATIVE
NOT DOG
FALSE POSITIVE
DOG



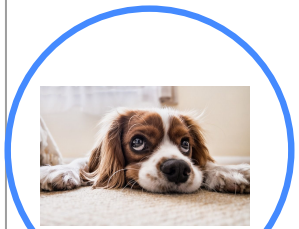



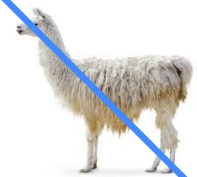

Precision

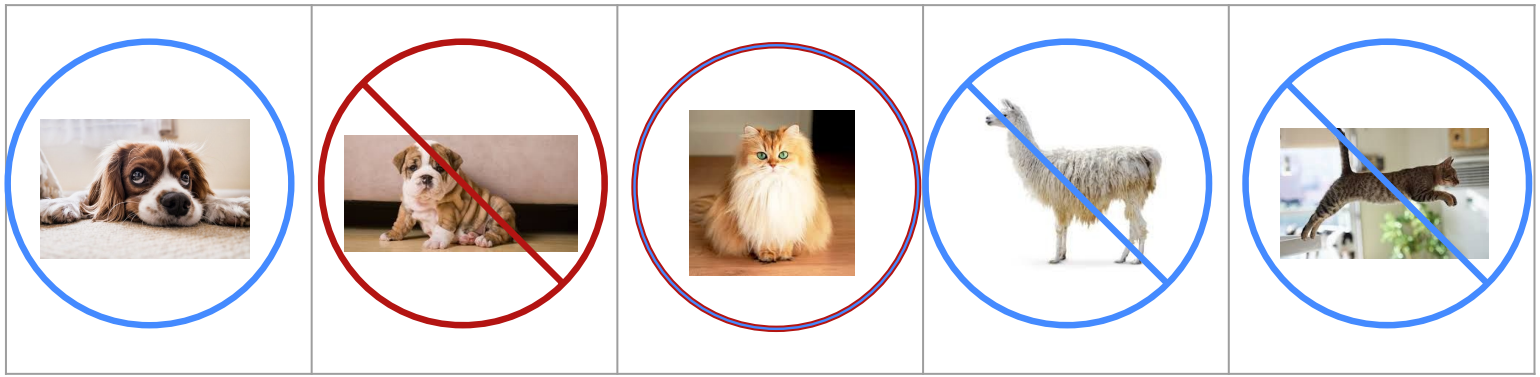
Recall

Precision

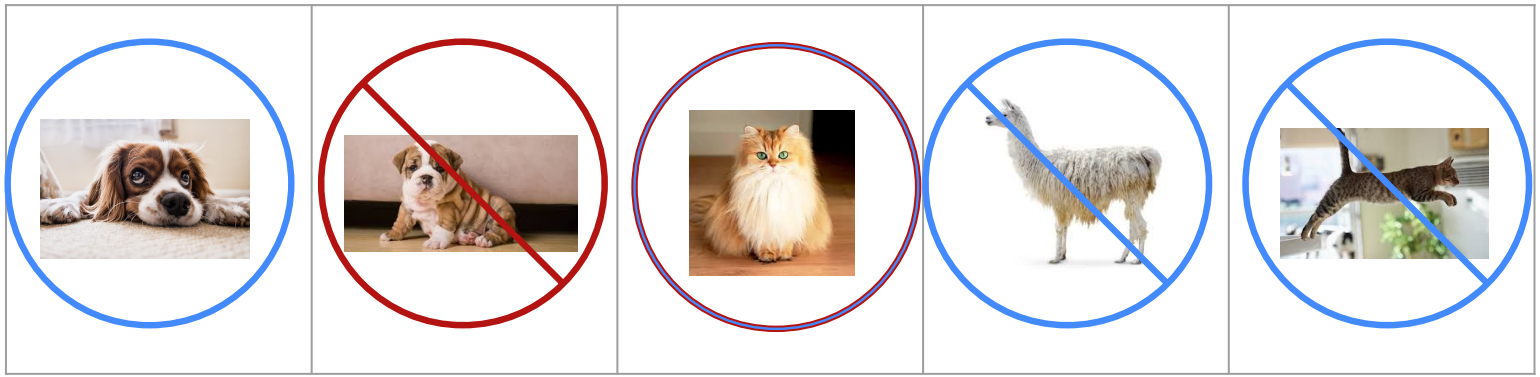
How often was the model right when it labeled something a dog?



2 labeled “dog”



2 labeled “dog”

Precision: 0.5

Precision

$$\frac{\text{\# } \textbf{dogs} \text{ labeled “dog”}}{\text{\# examples labeled “dog”}}$$

Precision

$$\frac{\text{\# dogs labeled “dog”}}{\text{\# examples labeled “dog”}} \Rightarrow \frac{\text{\# true positives}}{\text{\# false positives + \# true positives}}$$

Precision

$$\frac{\text{\# dogs labeled “dog”}}{\text{\# examples labeled “dog”}} \Rightarrow \frac{\text{\# true positives}}{\text{\# false positives + \# true positives}}$$

How much can we trust our model when it says “dog”?

When you’d want a high-precision model

- Deciding which book to read next
- Deciding who to hire

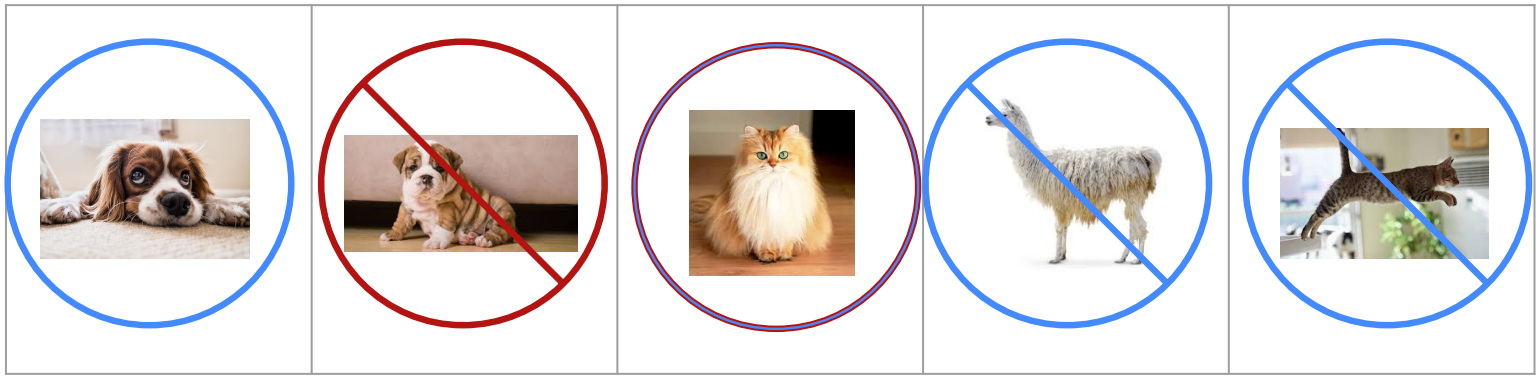
When you'd want a high-precision model

- Deciding which book to read next
- Deciding who to hire

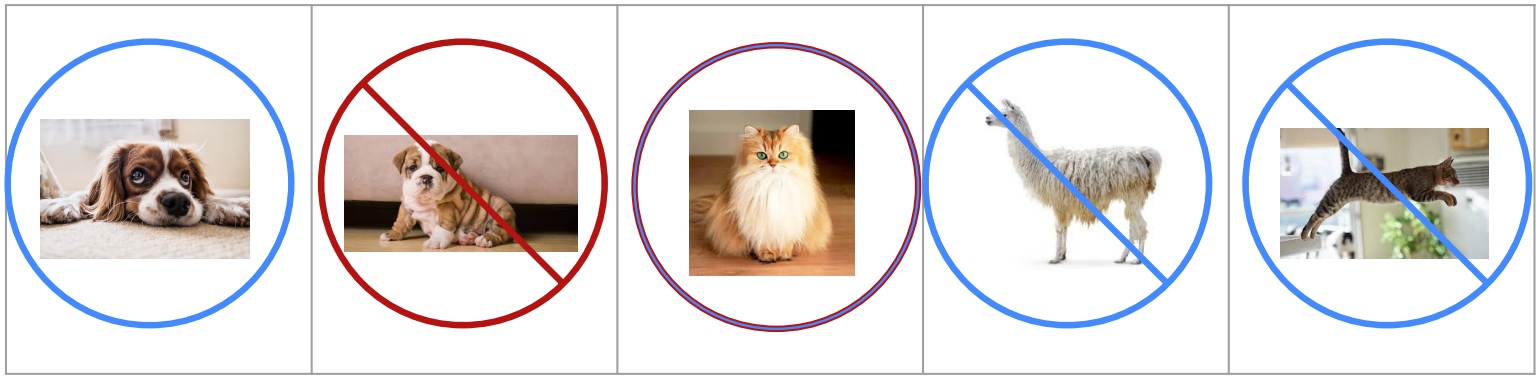
More false negatives

Recall

How well does our model catch all the dogs?



2 labeled “dog”



2 dogs

Recall: 0.5

Recall

$$\frac{\text{\# dogs } \textbf{labeled} \text{ “dog”}}{\text{\# actual dogs}}$$

Recall

$$\frac{\text{\# dogs } \textbf{labeled} \text{ “dog”}}{\text{\# actual dogs}}$$

$$\frac{\text{\# true positives}}{\text{\# false negatives + \# true positives}}$$

Recall

$$\frac{\text{\# dogs } \textbf{labeled} \text{ “dog”}}{\text{\# actual dogs}} \quad \Rightarrow \quad \frac{\text{\# true positives}}{\text{\# false negatives + \# true positives}}$$

Does our model miss any dogs?

When you'd want a high-recall model

- Security Screening
- Music recommendations

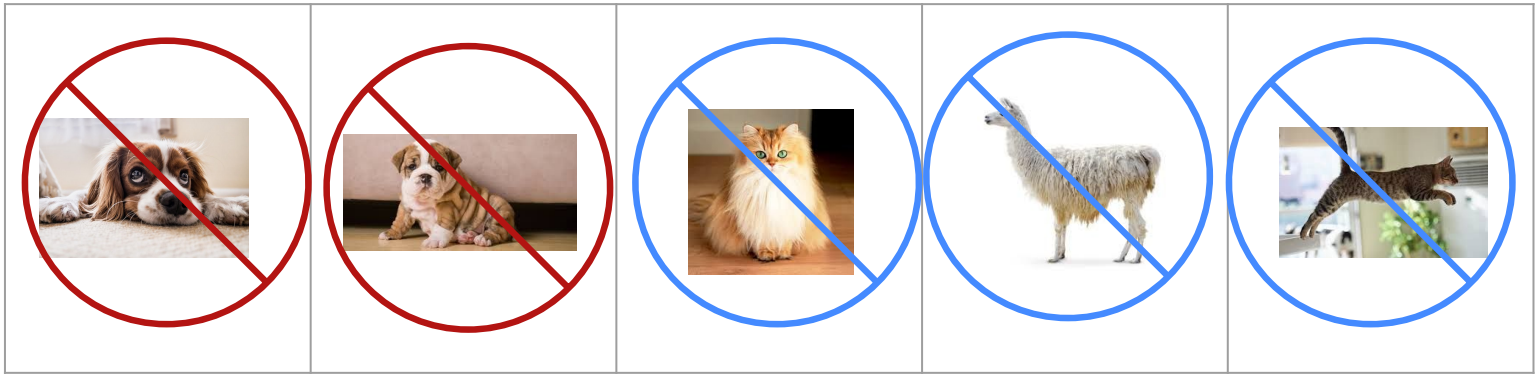
When you'd want a high-recall model

- Security Screening
- Music recommendations

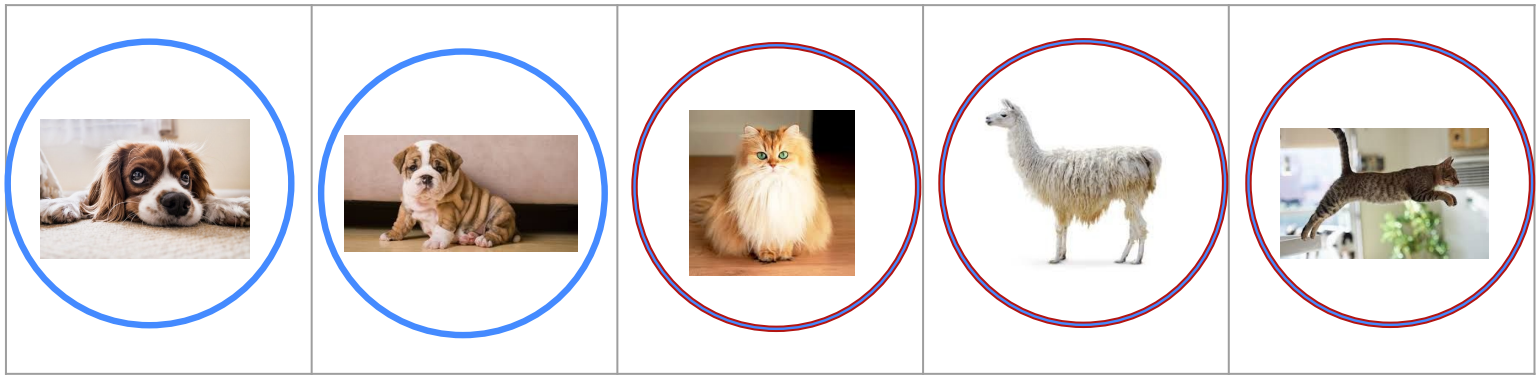
More false positives

There’s a natural tradeoff between precision and recall

100% Precision



100% Recall



Scaling Machine Learning

AutoML Natural Language

Batch predictions coming soon!

AutoML Tables

Deep Learning that integrates with BigQuery

Spark NLP

Learn more about NLP here:

https://towardsdatascience.com/july-edition-natural-language-processing-272f28835af7

bit.ly/social-media-ws