

**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”

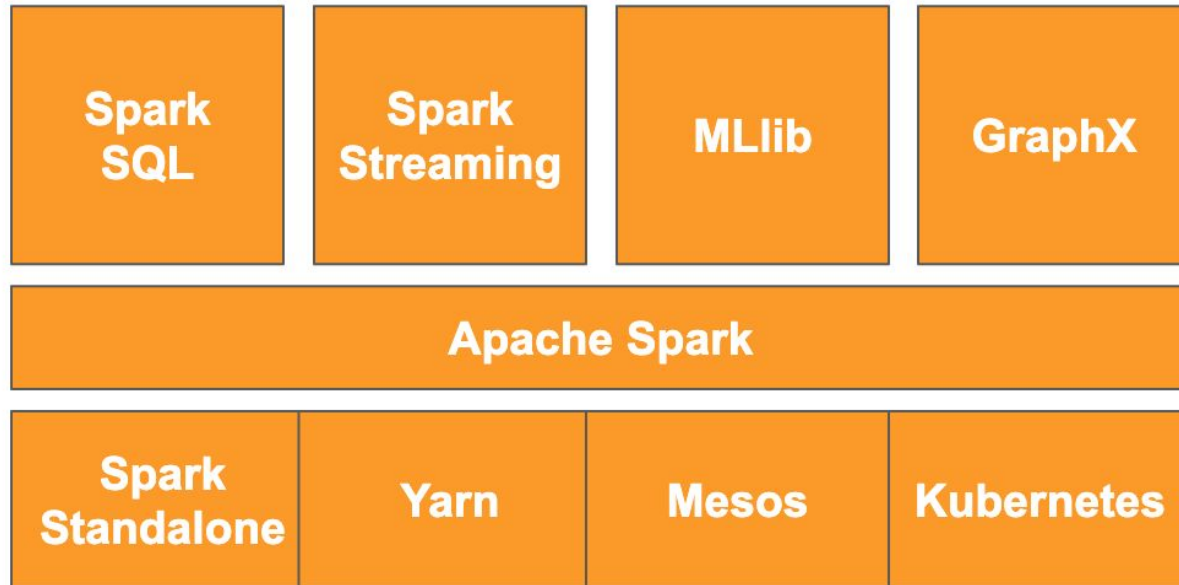
Load data into memory in a distributed manner

Execute jobs across multiple machines

Available for Python, Scala, Java, R

Vast ecosystem of available tools / runtime environments





# 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](https://bit.ly/pyspark-bigquery)**

**git clone**

**<https://github.com/bradmimo/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



Search r/food



pinke1993  
132 karma



r/food

FOOD

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)

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

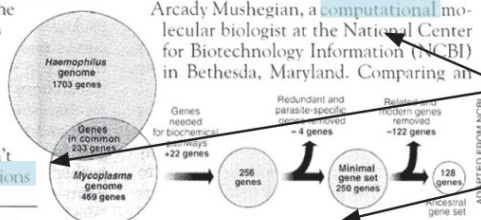
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

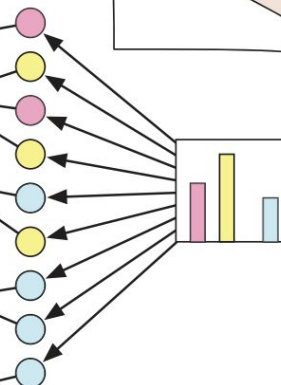


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments



# 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](https://bit.ly/spark-nlp)**

# NLP in Colab

[bit.ly/social-media-nlp](https://bit.ly/social-media-nlp)

# Top Life Pro Tips



97.9k



Posted by u/ProfessorLiftoff 1 year ago 📄 🏆

**LPT: Pay Attention to the smell of your home when you come back from a trip - that's what it smells like to guests all the time, you just get used to it.** [Home & Garden](#)

Whoa! Front page! And all because I stumbled back in my house, half-asleep and jet



99k



Posted by u/jade\_monkey07 1 year ago 📄

**LPT: If you like one song by an artist, but don't dig the rest of their stuff. find out who the producer is and see what other work they've done. The producer can play a big role in how the final song turns out.** [Arts & Culture](#)

Edit 1:Woo, front page! All from a stoner moment while listening to black keys - weight of

# Unpopular Life Pro Tips



0



Posted by u/Tachypsychias 11 hours ago

**LPT: Befriend neighbors along your usual walking routes to get their Wi-Fi passwords to save on Data!**



8 Comments



Share



Save

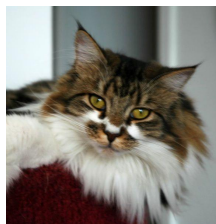


# Supervised Learning

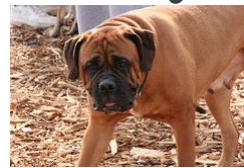
(no labels)

# Labeled Training Dataset

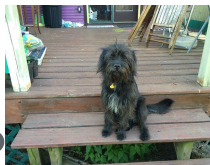
*Examples of cats*



*Examples of dogs*



# Training a Model

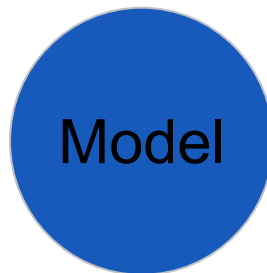


Machine  
Learning  
Algorithm



Model

# Making Predictions



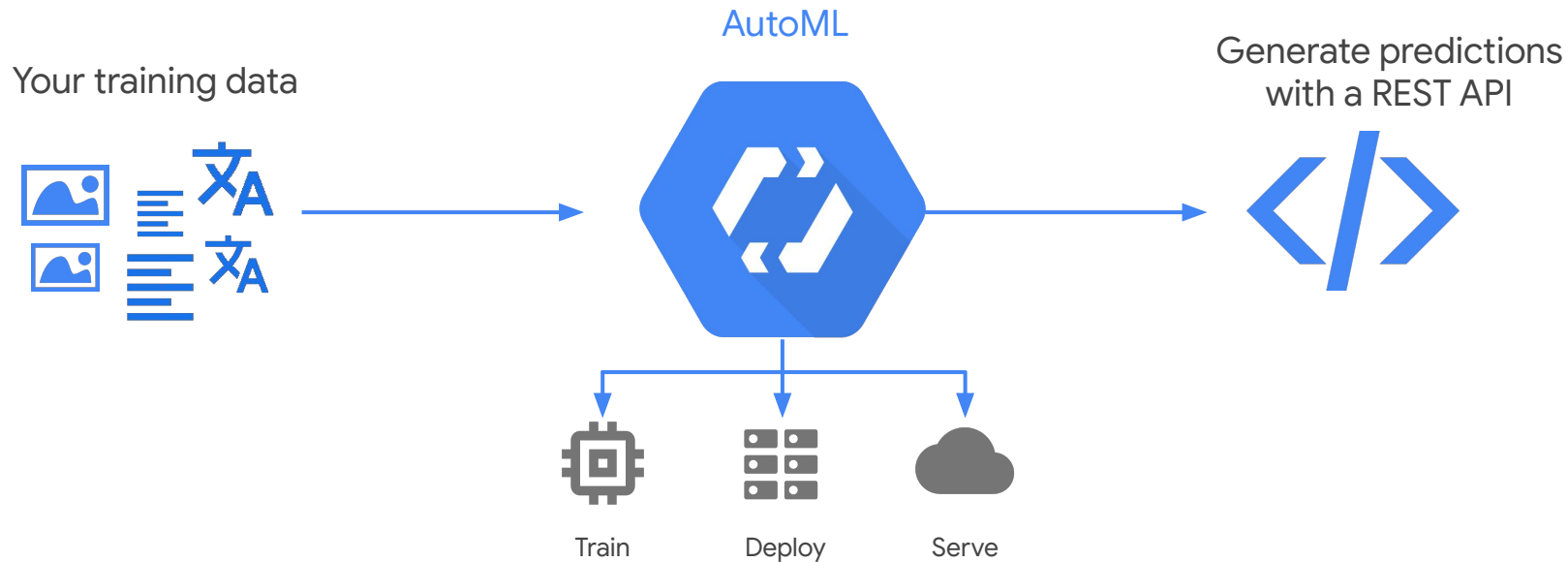
“Dog”



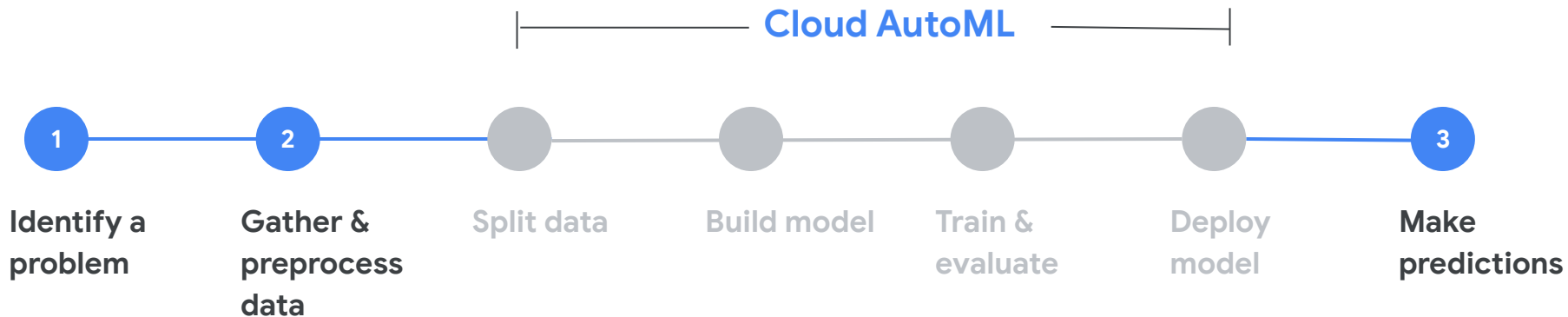


Cloud AutoML

# What is Cloud AutoML?



# How can Cloud AutoML help?



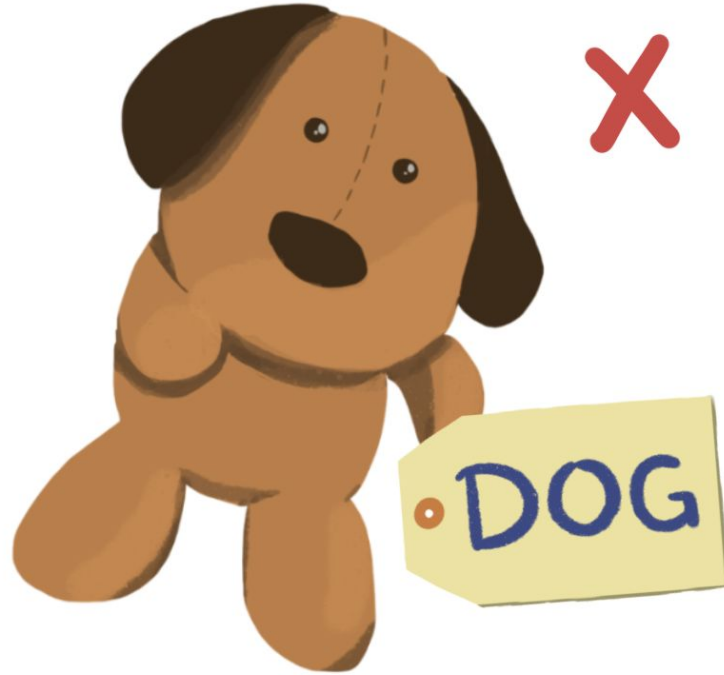
# Evaluating Your Model

**Dog classifier: dog or not dog**

**Accuracy: % correct predictions**

# All errors are not equal

# FALSE POSITIVE





# FALSE NEGATIVE

X



TRUE NEGATIVE



TRUE POSITIVE

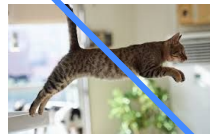
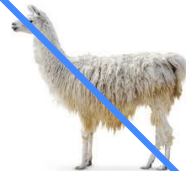


FALSE NEGATIVE



FALSE POSITIVE



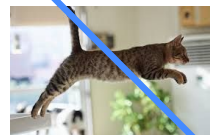
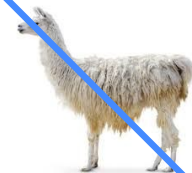
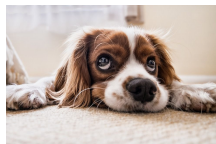


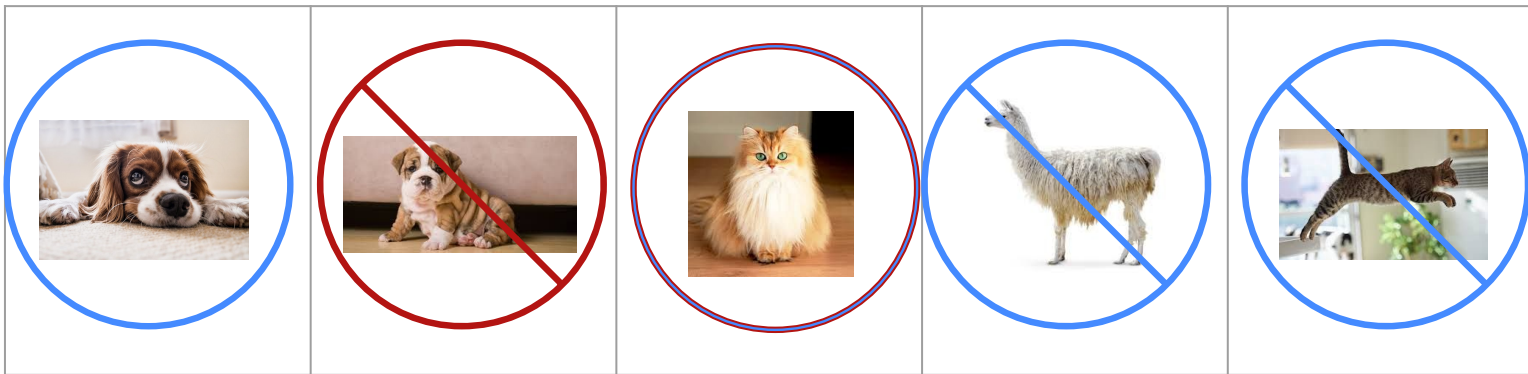
**Precision**

**Recall**

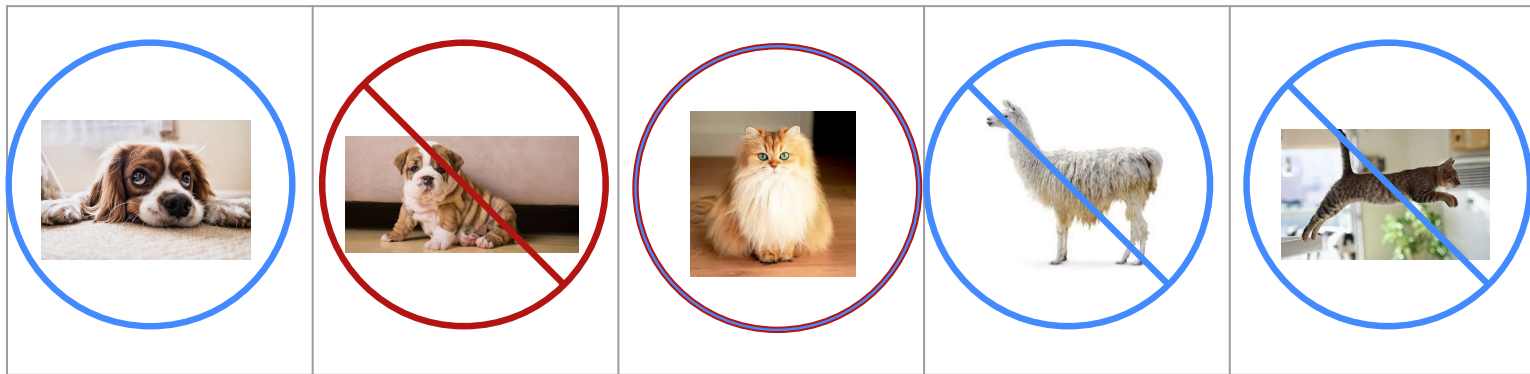
# Precision

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





2 labeled “dog”



2 labeled “dog”

Precision: 0.5



# Precision

# **dogs** labeled “dog”

# examples labeled “dog”

# Precision

$$\frac{\text{\# dogs labeled "dog"}}{\text{\# examples labeled "dog"}} \rightarrow \frac{\text{\# true positives}}{\text{\# false positives} + \text{\# true positives}}$$

# Precision

$$\frac{\text{\# dogs labeled "dog"}}{\text{\# examples labeled "dog"}} \rightarrow \frac{\text{\# true positives}}{\text{\# false positives} + \text{\# true positives}}$$

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

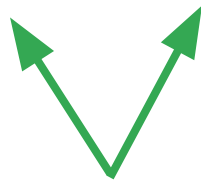
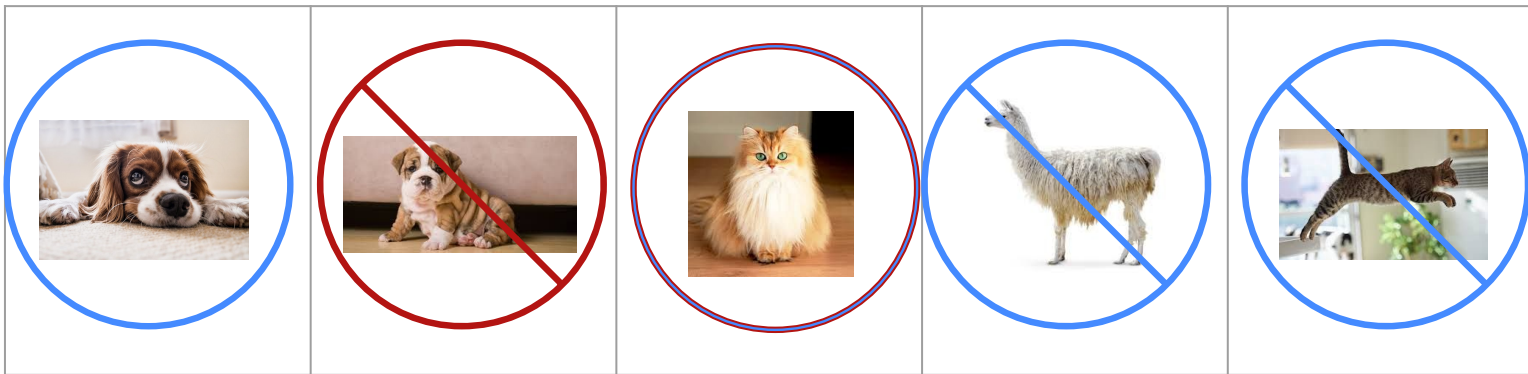
# When you'd want a high-precision model

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

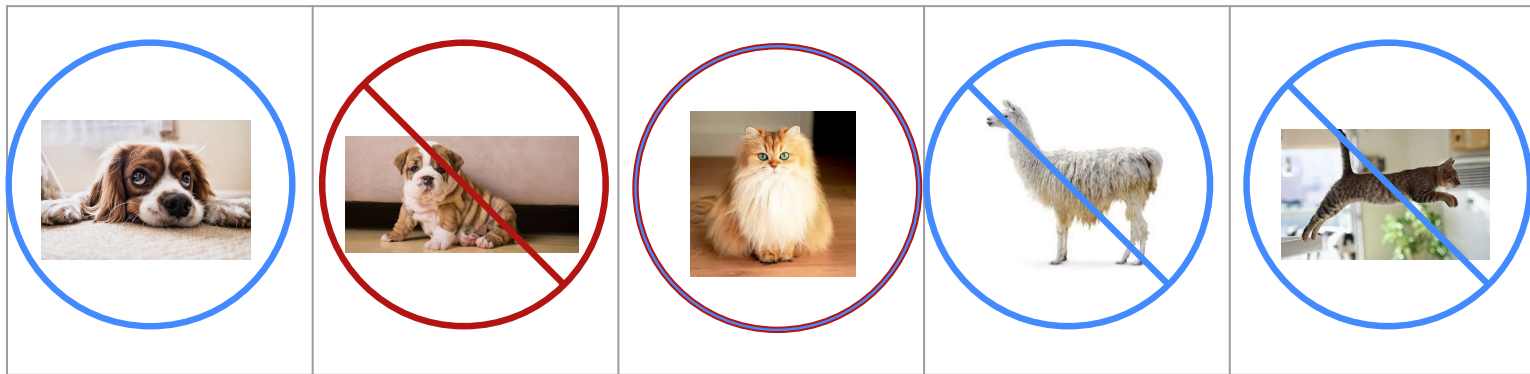
**More false negatives**

# Recall

How well does our model catch all the dogs?



2 labeled "dog"



2 dogs

Recall: 0.5



# Recall

# dogs **labeled** “dog”

---

# actual dogs

# Recall

$$\frac{\text{\# dogs **labeled** "dog"}}{\text{\# actual dogs}}$$



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

# Recall

$$\frac{\text{\# dogs labeled "dog"}}{\text{\# actual dogs}} \quad \rightarrow \quad \frac{\text{\# true positives}}{\text{\# false negatives} + \text{\# true positives}}$$

Does our model miss  
any dogs?

# When you'd want a high-recall model

- Security Screening
- Music recommendations

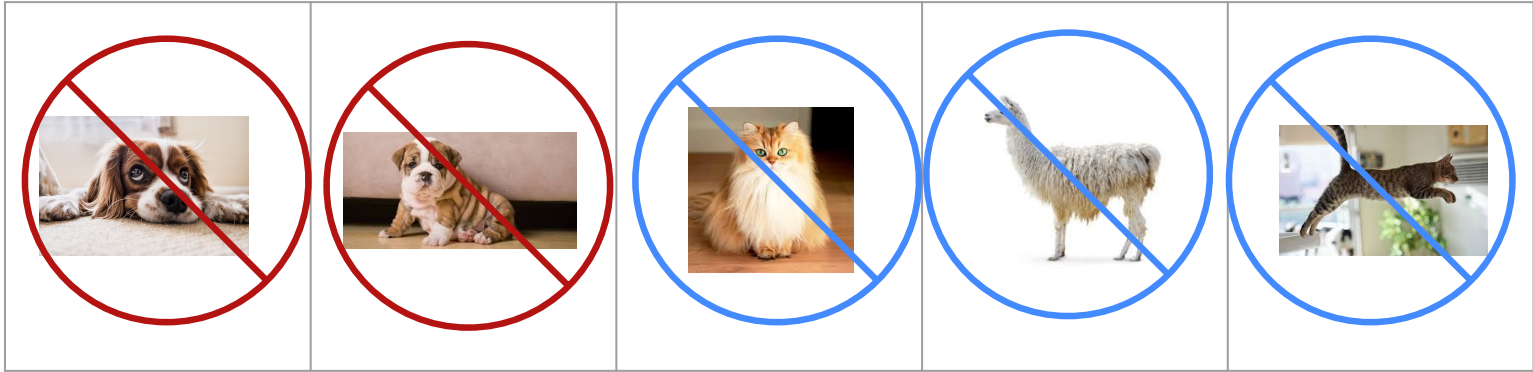
# When you'd want a high-recall model

- Security Screening
- Music recommendations

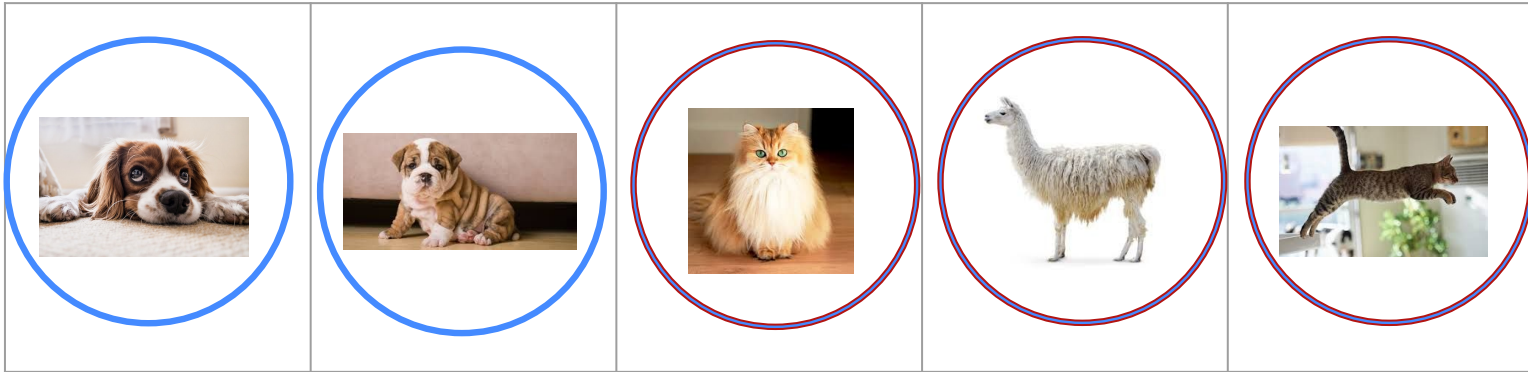
**More false positives**

**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](https://bit.ly/social-media-ws)**