Microsoft ML for Apache Spark

Unifying Machine Learning Ecosystems at Massive Scales

Mark Hamilton

Microsoft, MIT

marhamil@microsoft.com

Overview

- Background
  - Spark + SparkML
  - MMLSpark
- Unifying ML Ecosystems
  - LightGBM, CNTK, Vowpal Wabbit
  - Multilingual Bindings
- Microservice Orchestration
  - Cognitive Services on Spark
- Model Deployment with Spark Serving
- Use Cases
  - The Snow Leopard Trust

APACHE Spark™

Vowpal Wabbit LightGBM CNTK

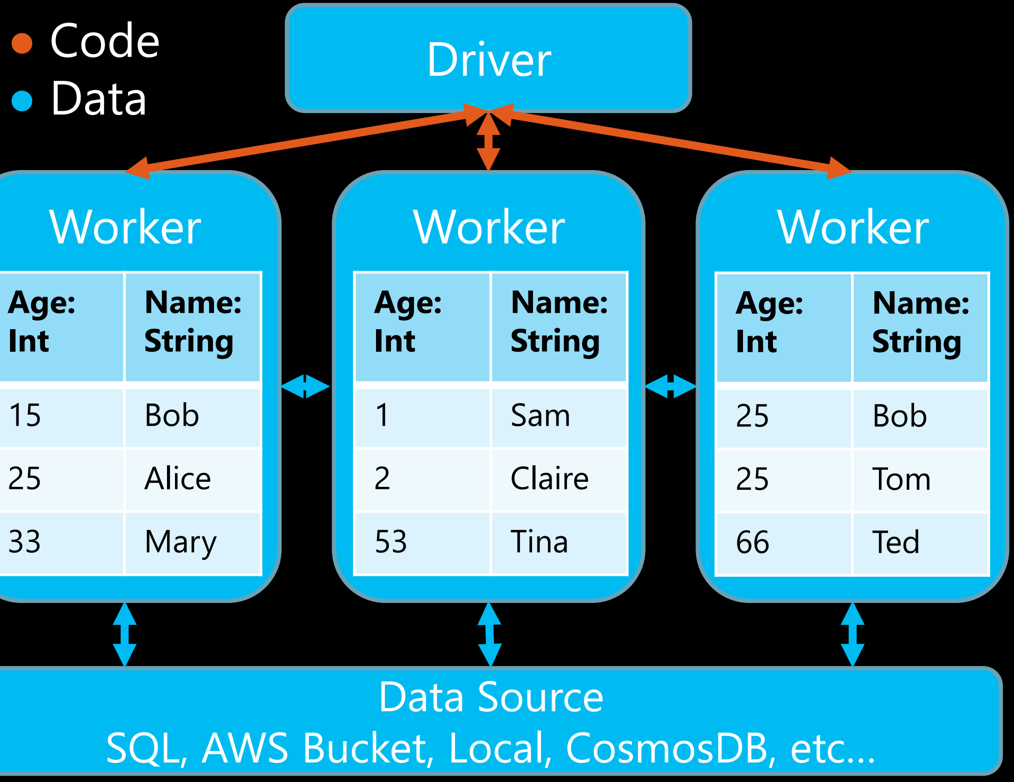
Cognitive Services Kubernetes

Snow Leopard Trust

THE MET

APACHE Spark™

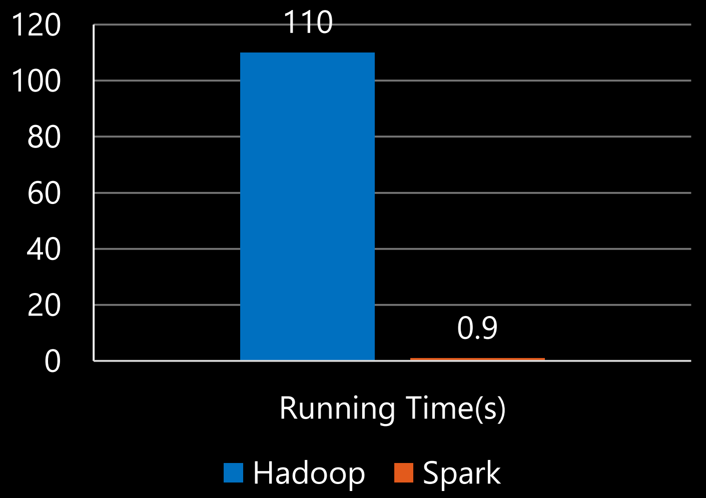
- A **fault-tolerant distributed** computing framework
- Map Reduce + SQL
- Whole program optimization + query pushdown
- Elastic
- Scala, Python, R, Java, Julia
- ML, Graph Processing, Streaming



Age: Int	Name: String
15	Bob
25	Alice
33	Mary

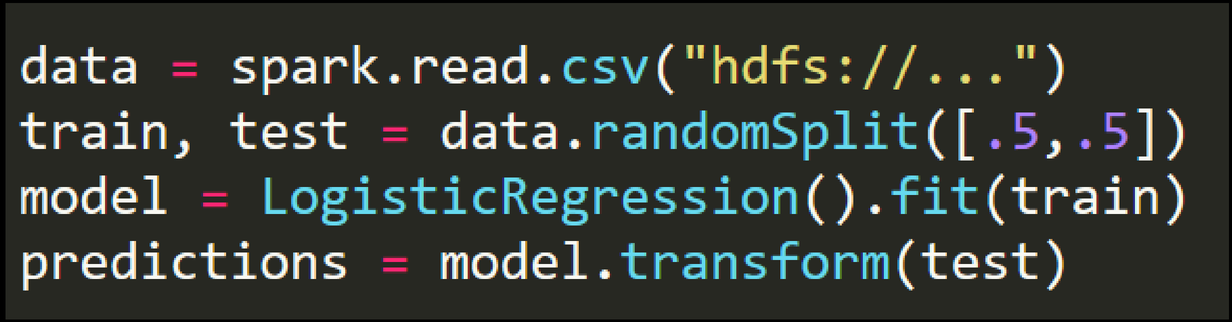
Age: Int	Name: String
1	Sam
2	Claire
53	Tina

Age: Int	Name: String
25	Bob
25	Tom
66	Ted

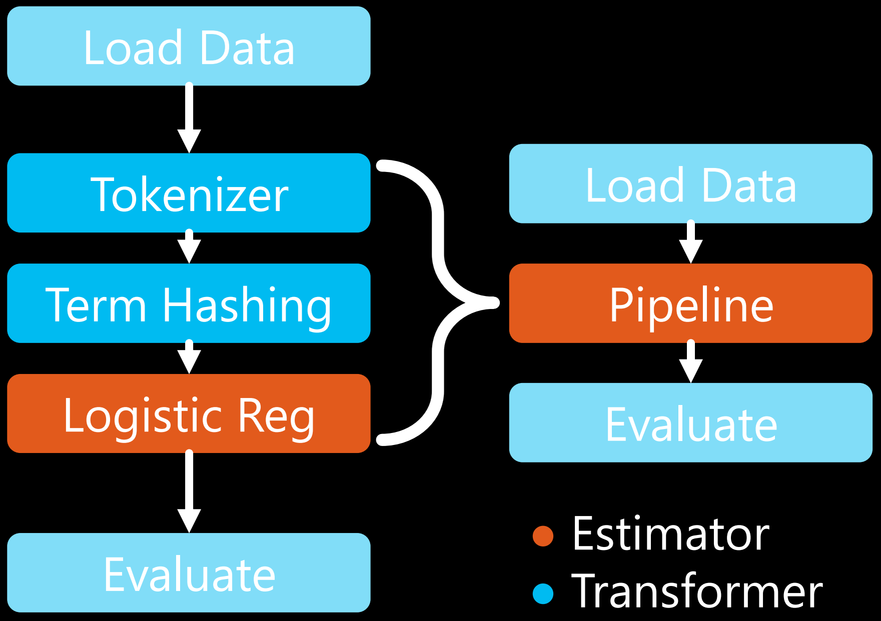


APACHE Spark™ ML

- High level library for distributed machine learning
- More general than SciKit-Learn
- All models have a uniform interface
  - Can compose models into complex pipelines
  - Can save, load, and transport models



```
data = spark.read.csv("hdfs://...")
train, test = data.randomSplit([.5,.5])
model = LogisticRegression().fit(train)
predictions = model.transform(test)
```



Microsoft Machine Learning for Apache Spark v0.18

Microsoft's Open Source Contributions to Apache Spark

Distributed Machine Learning

Fast Model Deployment

Microservice Orchestration

Multilingual Binding Generation

www.aka.ms/spark

Azure/mmlspark

Unifying Machine Learning Ecosystems

- Goals
  - Same API
  - Composable
  - Batch, Streaming, Serving
  - Elastically Distributed
  - Fault Tolerant
  - Multi-Language
  - Data Source Agnostic

Image Processing with Open CV

Gradient Boosting with LightGBM

Deep Learning Pipelines (Databricks)

Text Analytics with Vowpal Wabbit

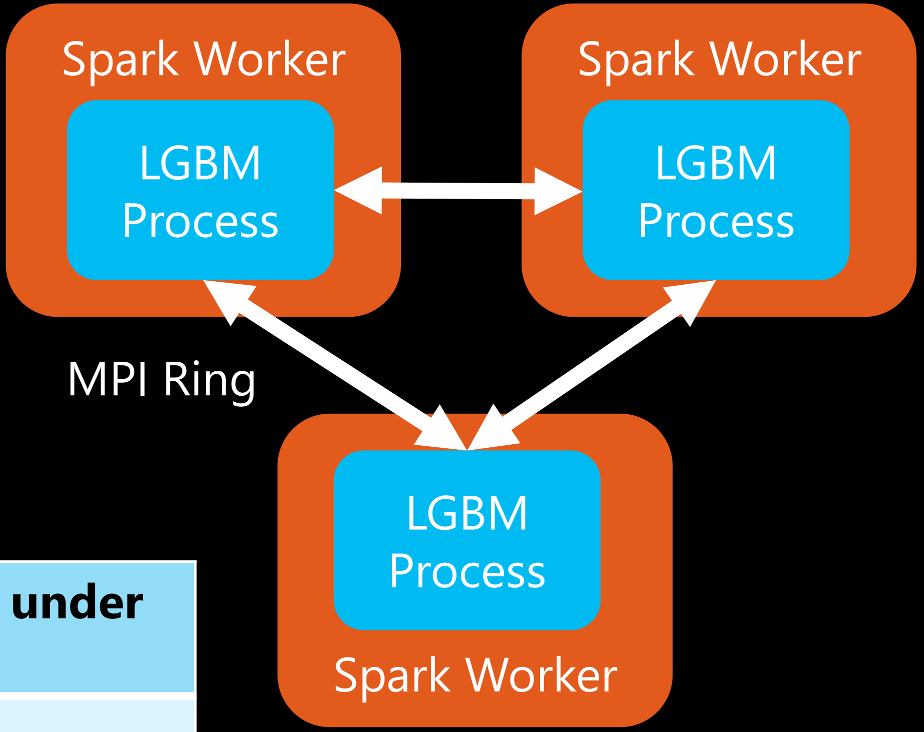
Distributed Model Interpretability with LIME

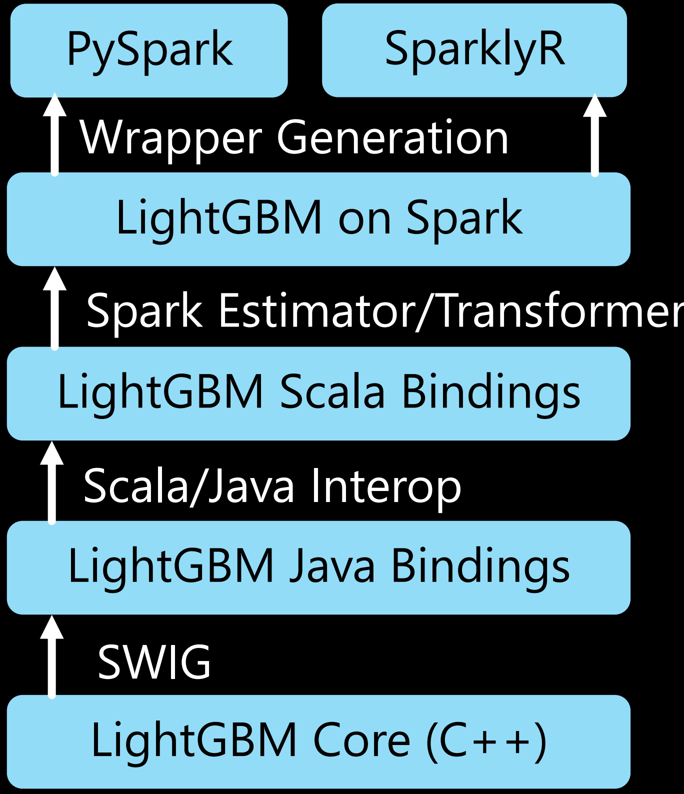
Deep Learning with CNTK

Markus Cozowicz
marcozo@microsoft.com
Data Scientist

Example Backend: LightGBM on Spark

- Barrier Execution for Synchronizing Workers
- Fast Socket/MPI communication
- `mapPartitions` for Transformer



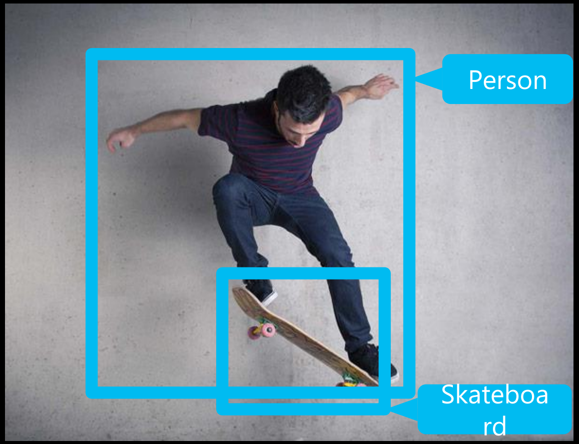


Framework	**Time(s)**	**Area under ROC**
XGBoost	52.60	.808
SparkML GBT	82.78	.788
LightGBM	**45.39**	**.812**

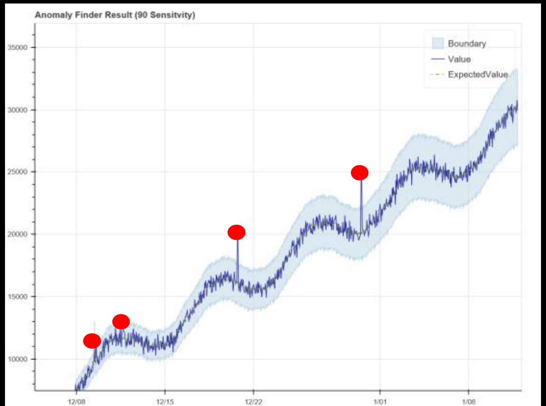
Ilya Matiach, ilmat@microsoft.com
Developer, Azure ML

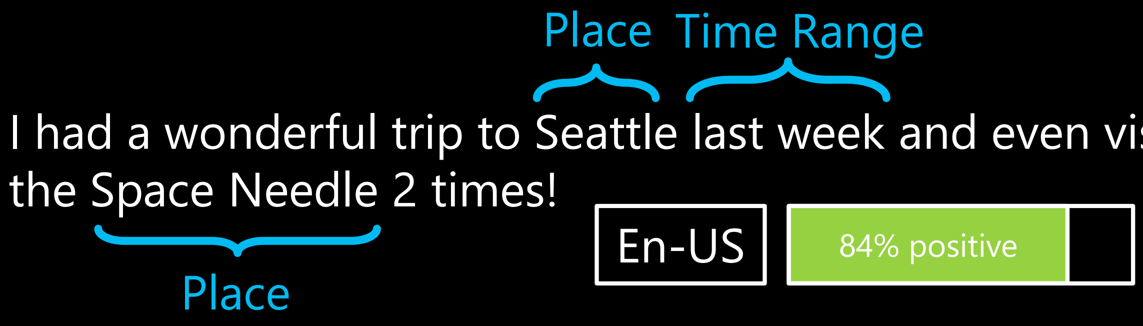
Cognitive Services

- High quality pre-built intelligent services
- No time intensive model training or deployment
- Leverage Microsoft Research and Azure ML
- **Available as Docker Containers**



→ “Hello World”





Place Time Range

I had a wonderful trip to Seattle last week and even visited the Space Needle 2 times!

Place

En-US

84% positive

Vision

Object, scene, and activity detection

Face recognition and identification

Celebrity and landmark recognition

Emotion recognition

Text and handwriting recognition (OCR)

Customizable image recognition

Video metadata, audio, and keyframe extraction and analysis

Explicit or offensive content moderation

Speech

Speech transcription (speech-to-text)

Custom speech models for unique vocabularies or complex environment

Text-to-speech

Custom Voice

Real-time speech translation

Customizable speech transcription and translation

Speaker identification and verification

Language

Language detection

Named entity recognition

Key phrase extraction

Text sentiment analysis

Multilingual and contextual spell checking

Explicit or offensive text content moderation

PII detection for text moderation

Text translation

Customizable text translation

Contextual language understanding

Decision

Q&A extraction from unstructured text

Knowledge base creation from collections of Q&As

Semantic matching for knowledge bases

Customizable content personalization learning

Search

Ad-free web, news, image, and video search results

Trends for video, news

Image identification, classification and knowledge extraction

Identification of similar images and products

Named entity recognition and classification

Knowledge acquisition for named entities

Search query autosuggest

Ad-free custom search engine creation

Azure Cognitive Services on Spark

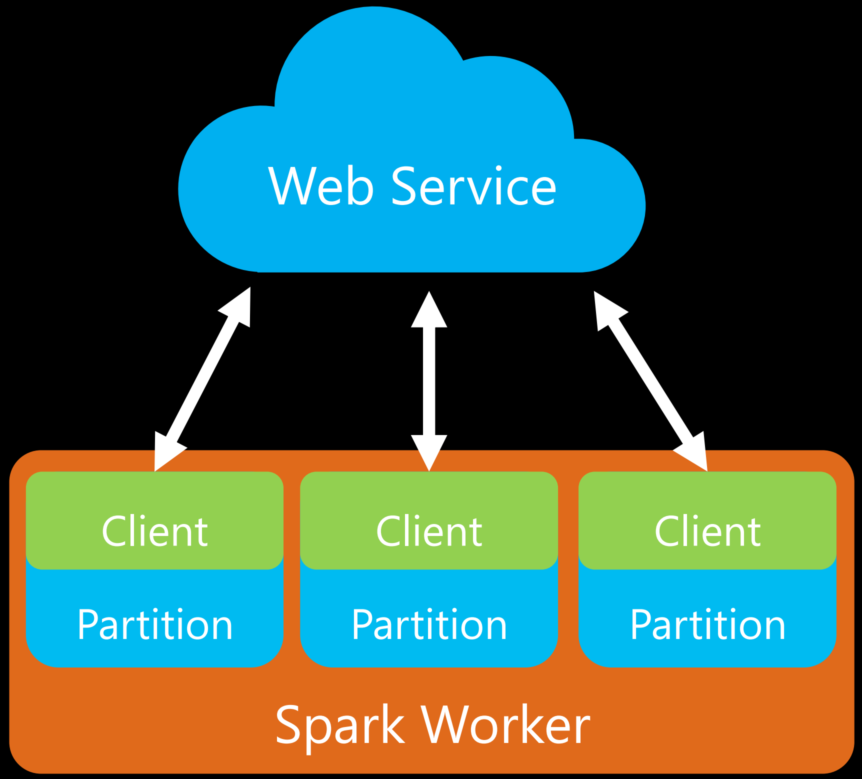
- Easy to use integration between Spark and the Azure Cognitive Services
- Composable and pipelinable with all other SparkML models!
- Exponential Backoffs, Backpressure, Batching, Async Parallelism
- Fully Fluent API

```
val df = new TextSentiment()
  .setTextCol(“text”)
  .setOutputCol(“sentiment”)
  .transform(inputs)
```

Features	**Time (s)**	**Errors #**
None	30.8	18993
EBO+BP	1163.0	0
EBO+BP+B	57.1	0
EBO+BP+B+P	**49.7**	**0**

HTTP on Spark

- Full Integration between HTTP Protocol and Spark SQL
- Spark as a Microservice Orchestrator
- Spark + X
- Support for all Spark Languages



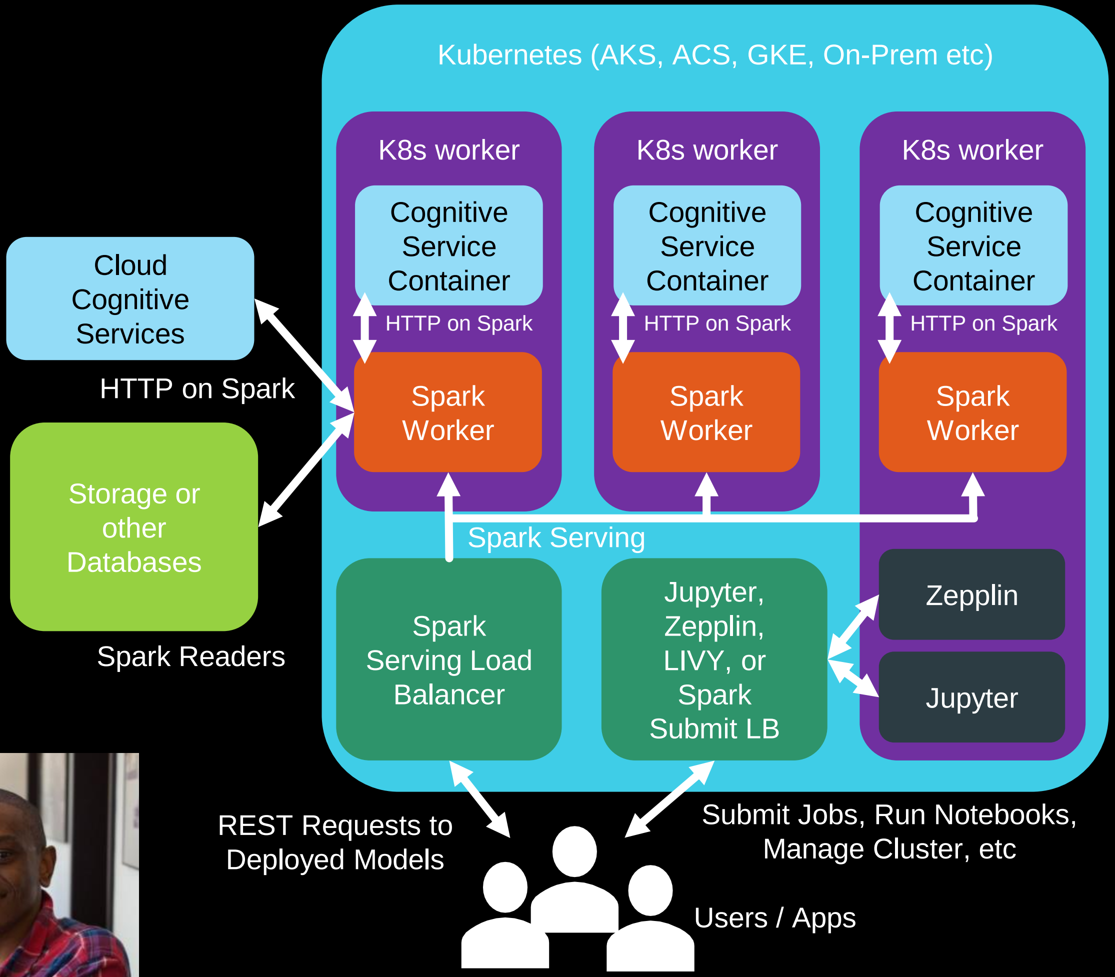
```
df = SimpleHTTPTransformer()
  .setInputParser(JSONInputParser())
  .setOutputParser(JSONOutputParser()
    .setDataType(schema))
  .setOutputCol("results")
  .setUrl(…)
```

Deploying on Kubernetes

- Works on any k8s cluster
- Helm: Package Manager for Kubernetes

```
helm repo add microsoft \
    https://microsoft.github.io/charts/repo
helm update

helm install microsoft/spark --version 1.0.0
```


Kubernetes (AKS, ACS, GKE, On-Prem etc)
K8s worker
Cognitive Service Container
HTTP on Spark
Spark Worker
K8s worker
Cognitive Service Container
HTTP on Spark
Spark Worker
K8s worker
Cognitive Service Container
HTTP on Spark
Spark Worker
Cloud Cognitive Services
HTTP on Spark
Storage or other Databases
Spark Readers
Spark Serving
Spark Serving Load Balancer
Jupyter, Zepplin, LIVY, or Spark Submit LB
Zepplin
Jupyter
REST Requests to Deployed Models
Submit Jobs, Run Notebooks, Manage Cluster, etc
Users / Apps

Dalitso Banda, dbanda@microsoft.com
Microsoft AI Development Acceleration Program

Model Deployment with Spark Serving

- Sub-millisecond RESTful Model Deployment on Spark Clusters

Batch API:

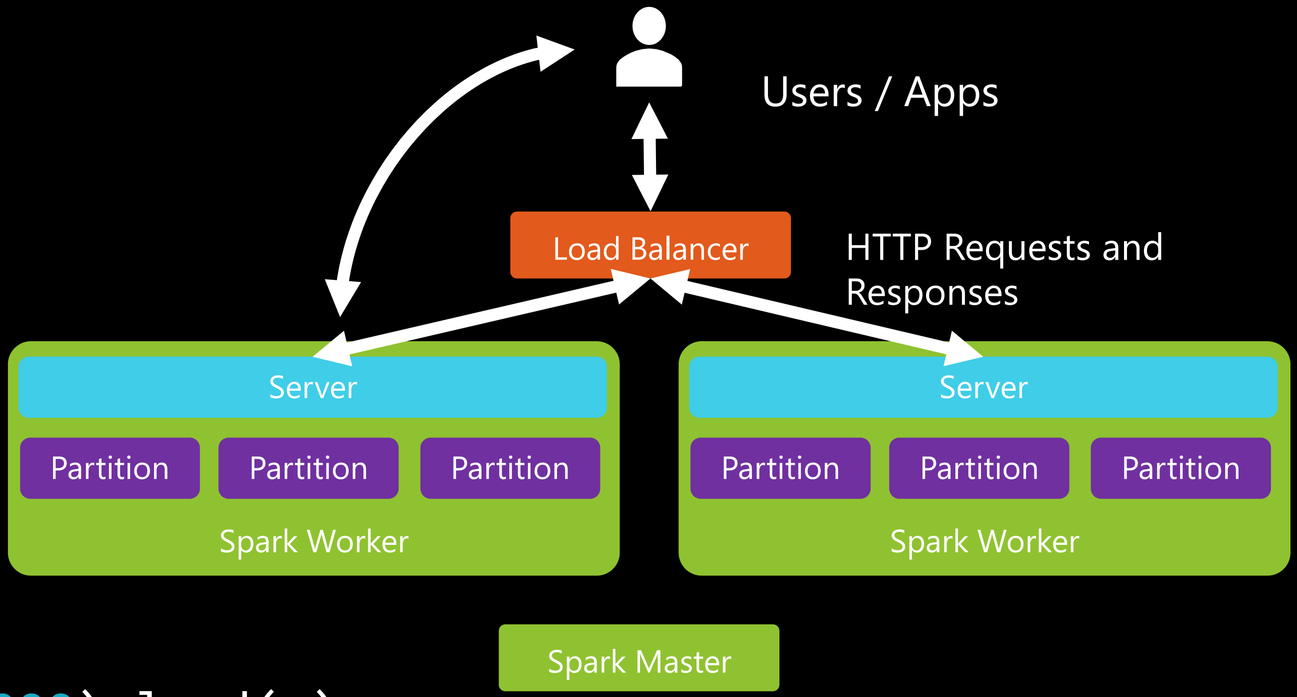
```
spark.read.parquet.load(…)
  .select(…)
```

Streaming API:

```
spark.readStream.kafka.load(…)
  .select(…)
```

Serving API:

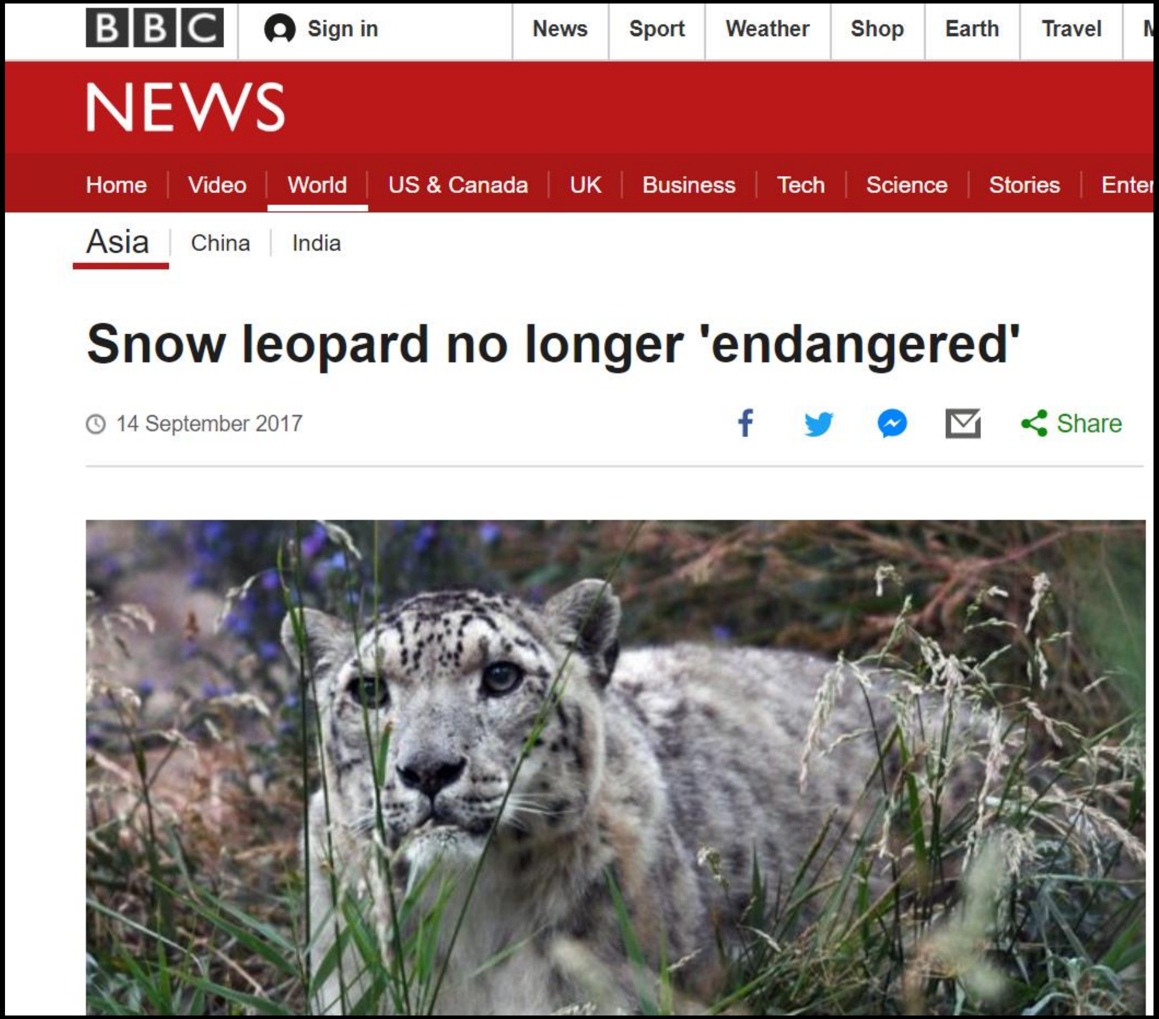
```
spark.readStream.server(“0.0.0.0”, 5000).load(…)
  .select(…)
```



AI for Earth

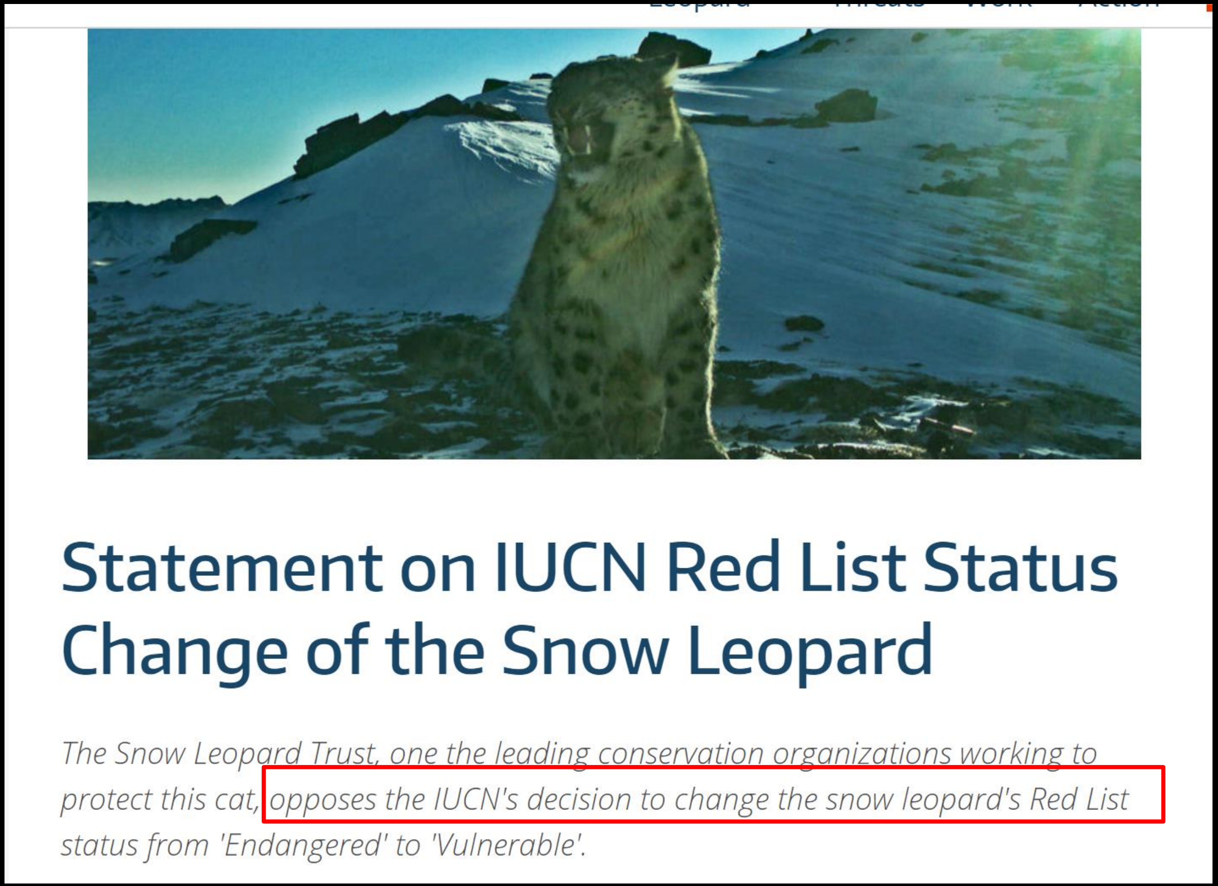
Snow Leopard Trust

Endangered Status Matters



Snow leopard no longer 'endangered'

14 September 2017

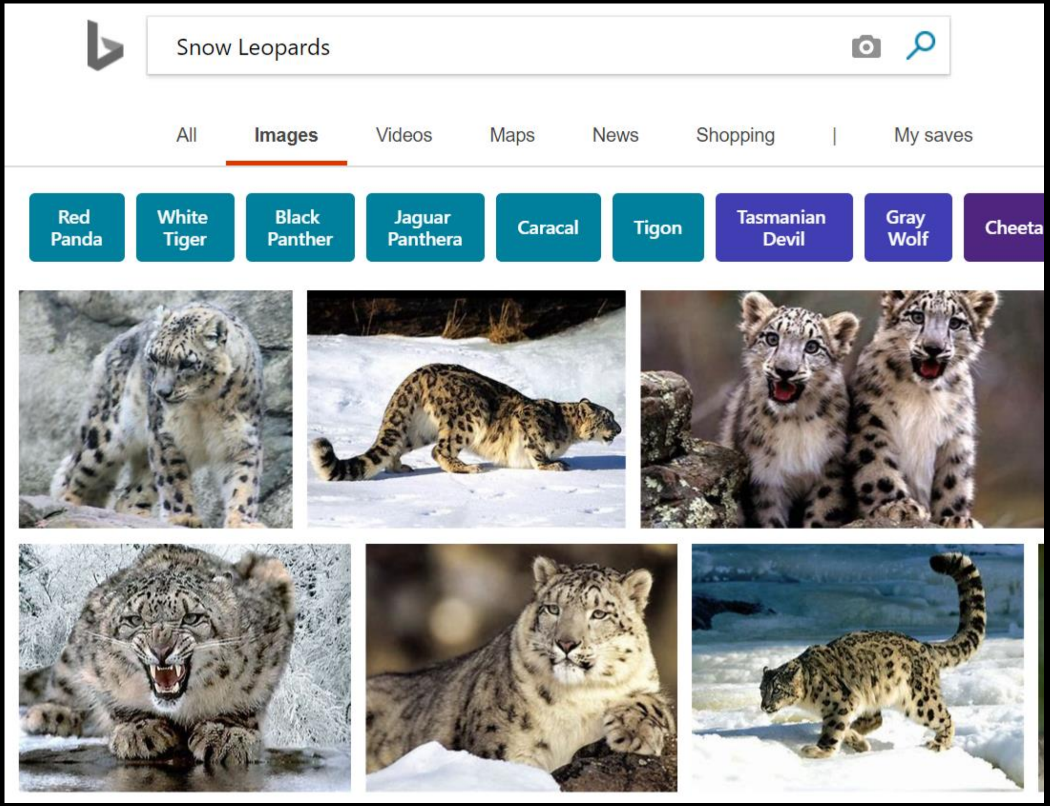


Statement on IUCN Red List Status Change of the Snow Leopard

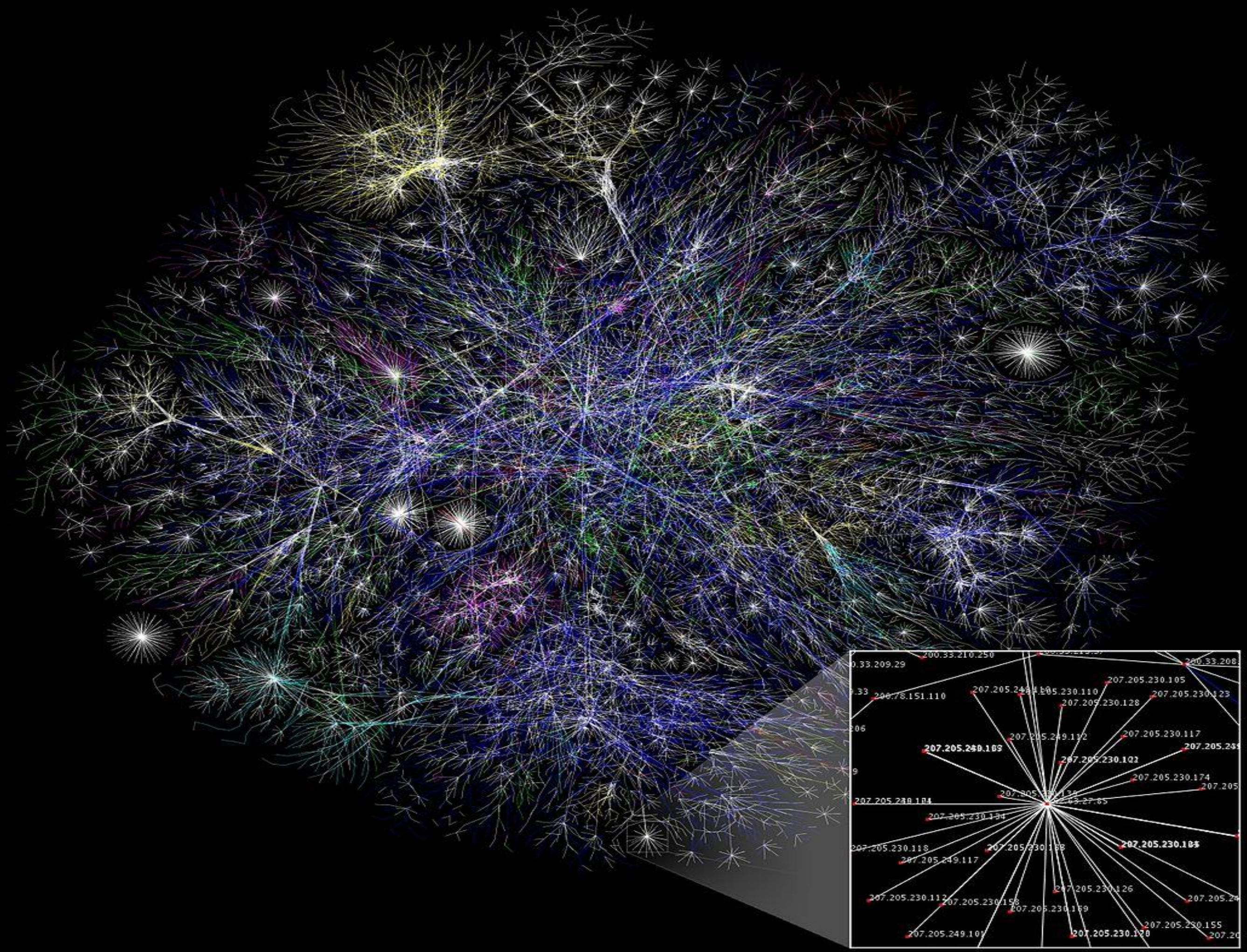
The Snow Leopard Trust, one the leading conservation organizations working to protect this cat, opposes the IUCN's decision to change the snow leopard's Red List status from 'Endangered' to 'Vulnerable'.

Remote Camera Trapping

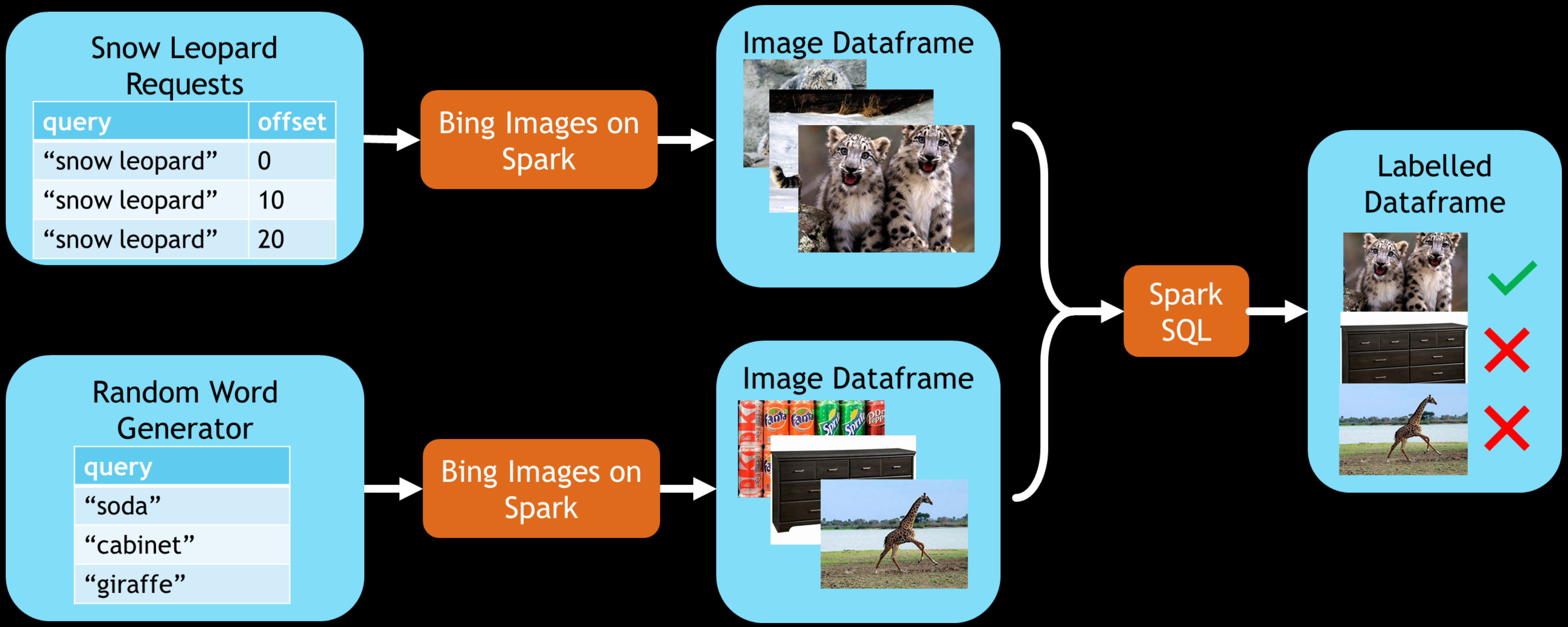
Creating a labelled Training Dataset



Snow Leopards



Creating a labelled Training Dataset



Snow Leopard Requests

query	offset
“snow leopard”	0
“snow leopard”	10
“snow leopard”	20

Bing Images on Spark

Image Dataframe

Random Word Generator

query
“soda”
“cabinet”
“giraffe”

Bing Images on Spark

Image Dataframe

Spark SQL

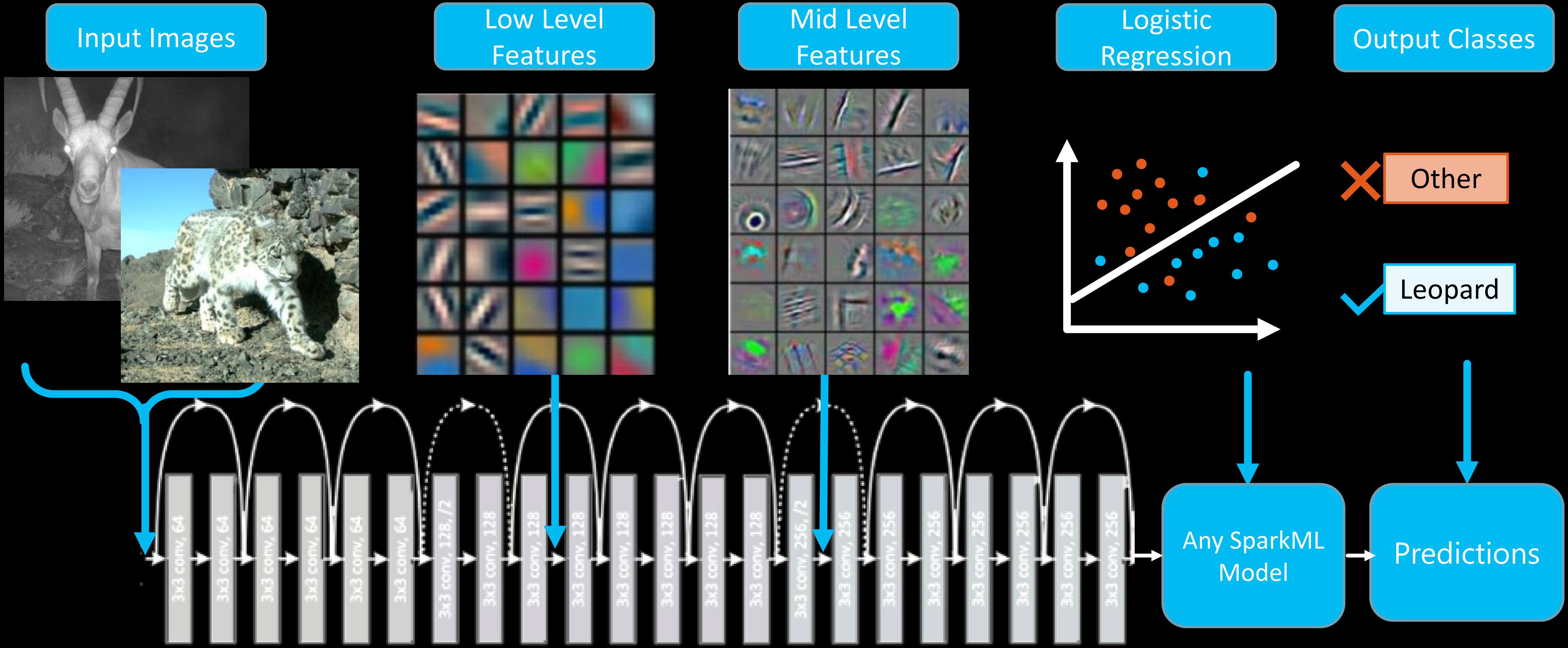
Labelled Dataframe

✓

✗

✗

Transfer Learning with ResNet 50



Input Images

Low Level Features

Mid Level Features

Logistic Regression

Output Classes

Other

Leopard

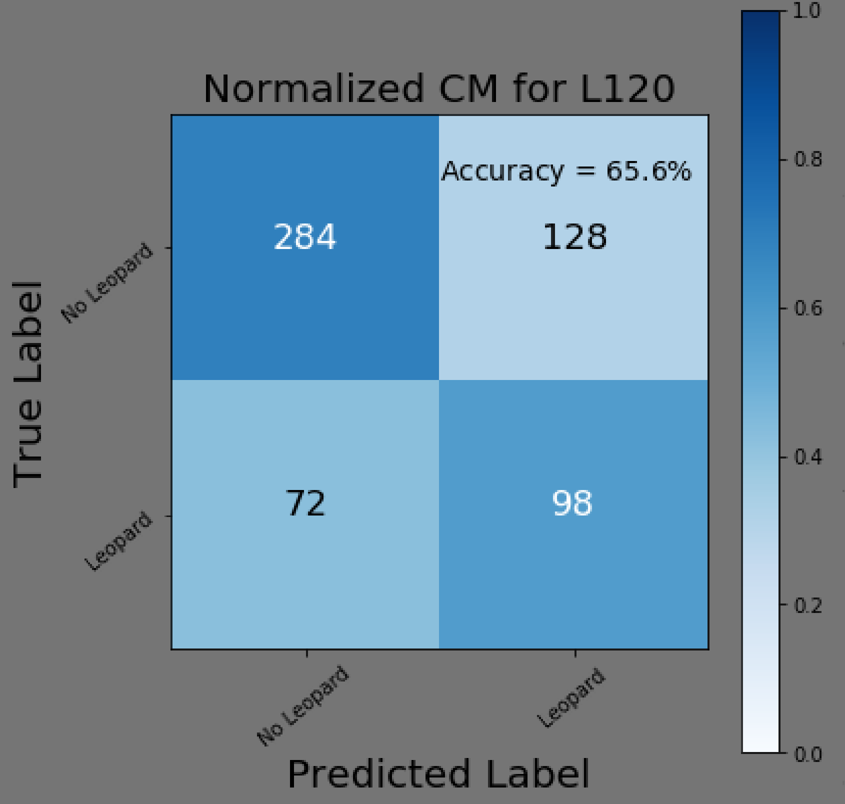
Any SparkML Model

Predictions

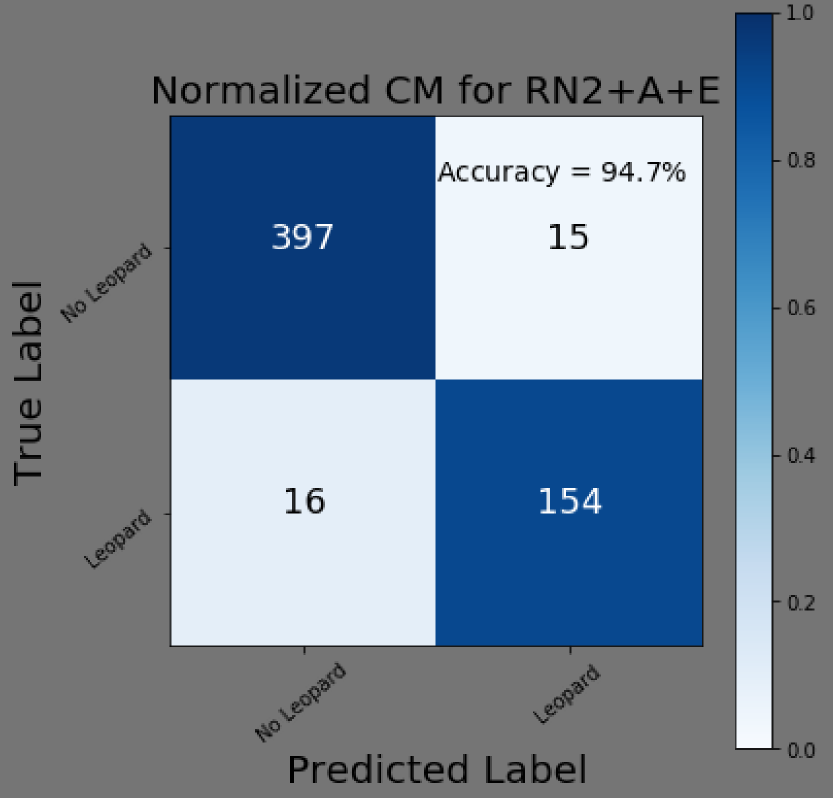
Filters from Zeiler + Fergus 2013

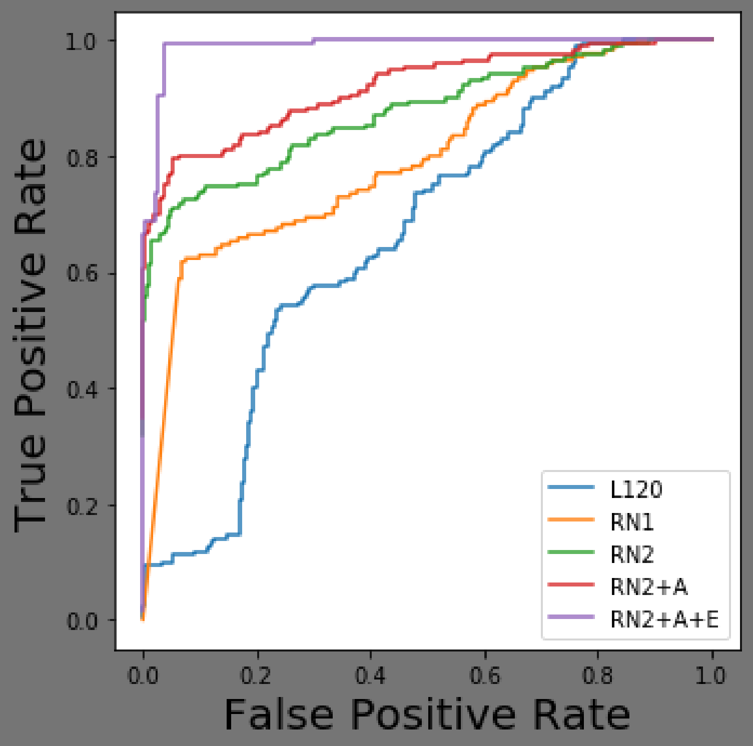
Performance

Without Deep Featurization



With Deep Featurization, Augmentation, and Temporal Ensembling

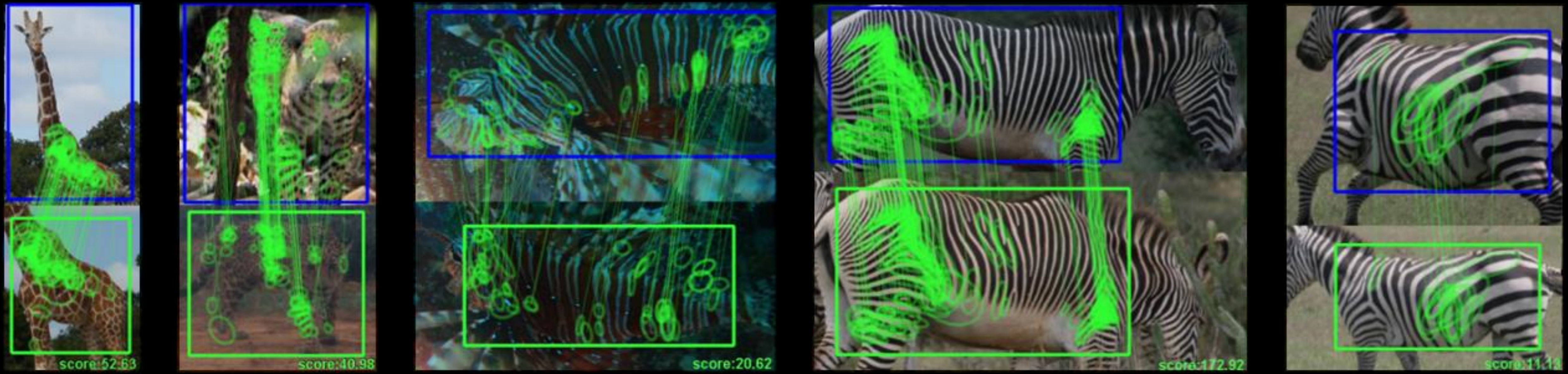




Accuracy 65.6%

Accuracy 94.7%

Goal: Identify Individual Leopards

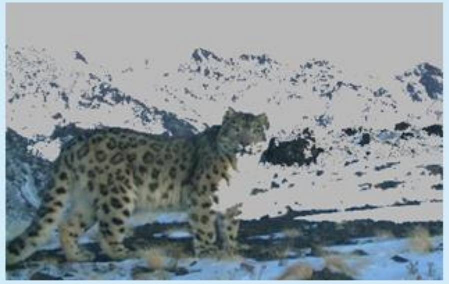
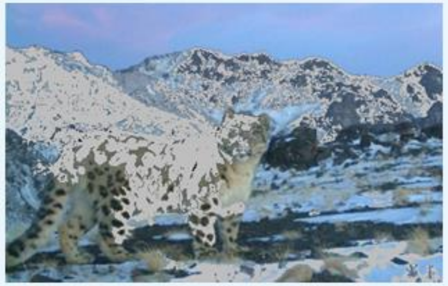
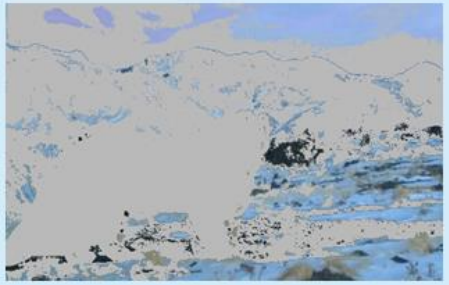


Source: HotSpotter - Patterned Species Instance Recognition

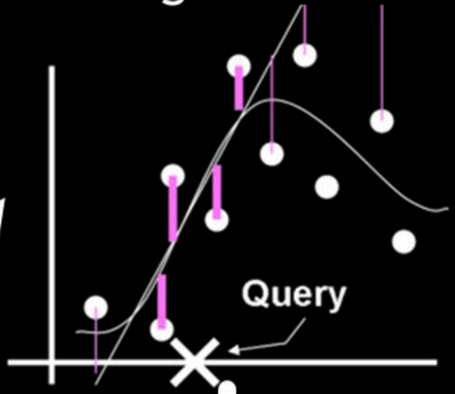
Automating Detection with LIME on Spark

Input Image

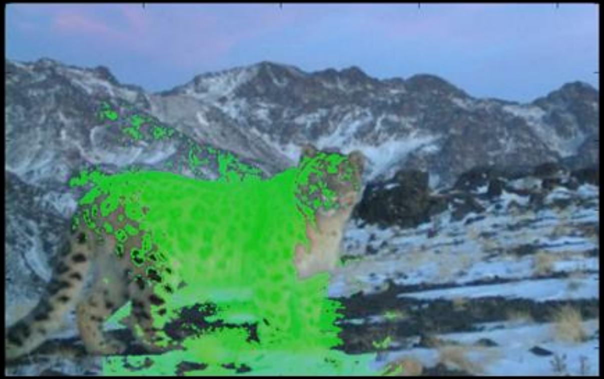
P(leopard) = .88

Perturbed Photos	P(leopard)
	.92
	.56
	.003

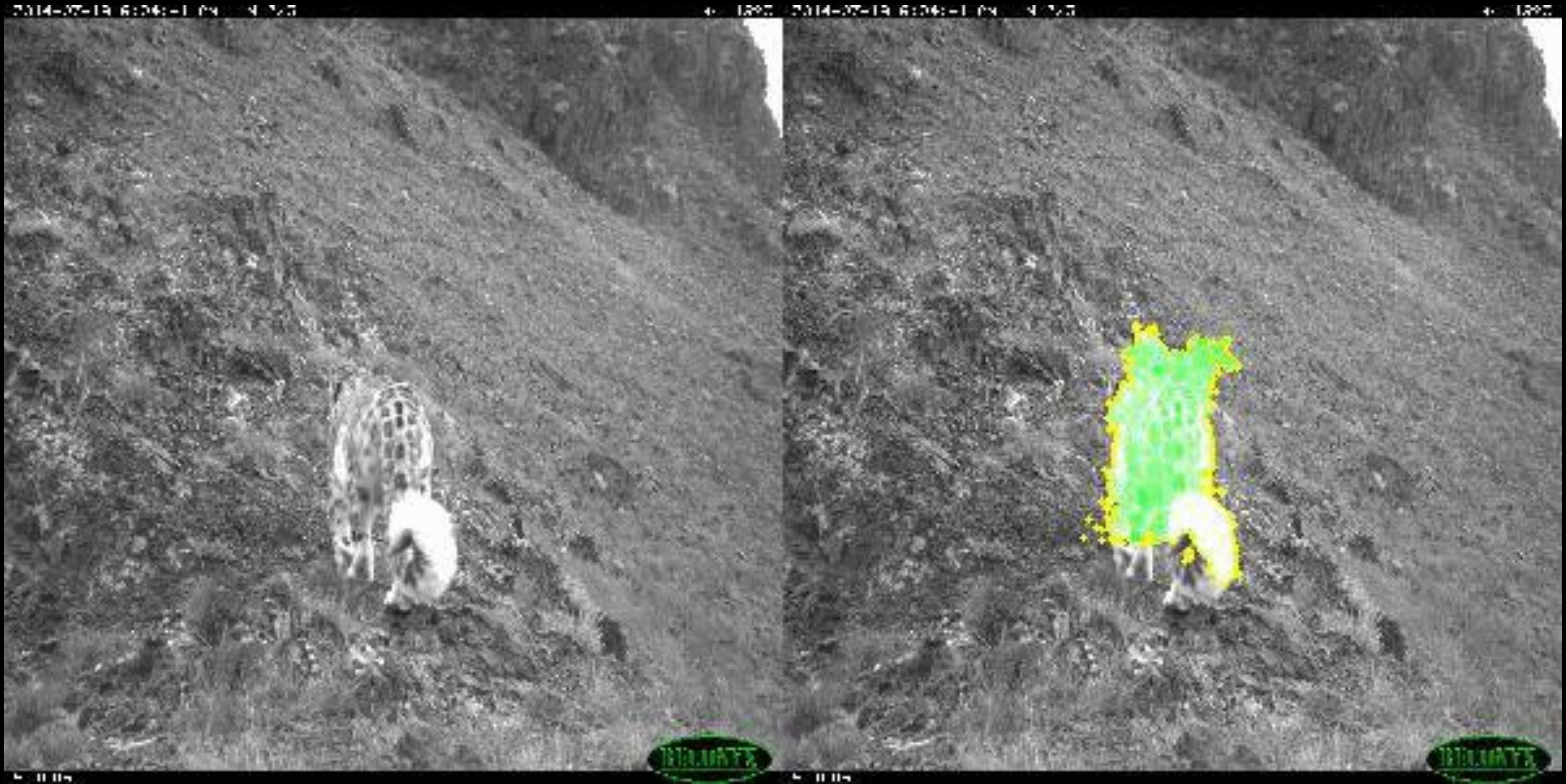
Locally Weighted Regression



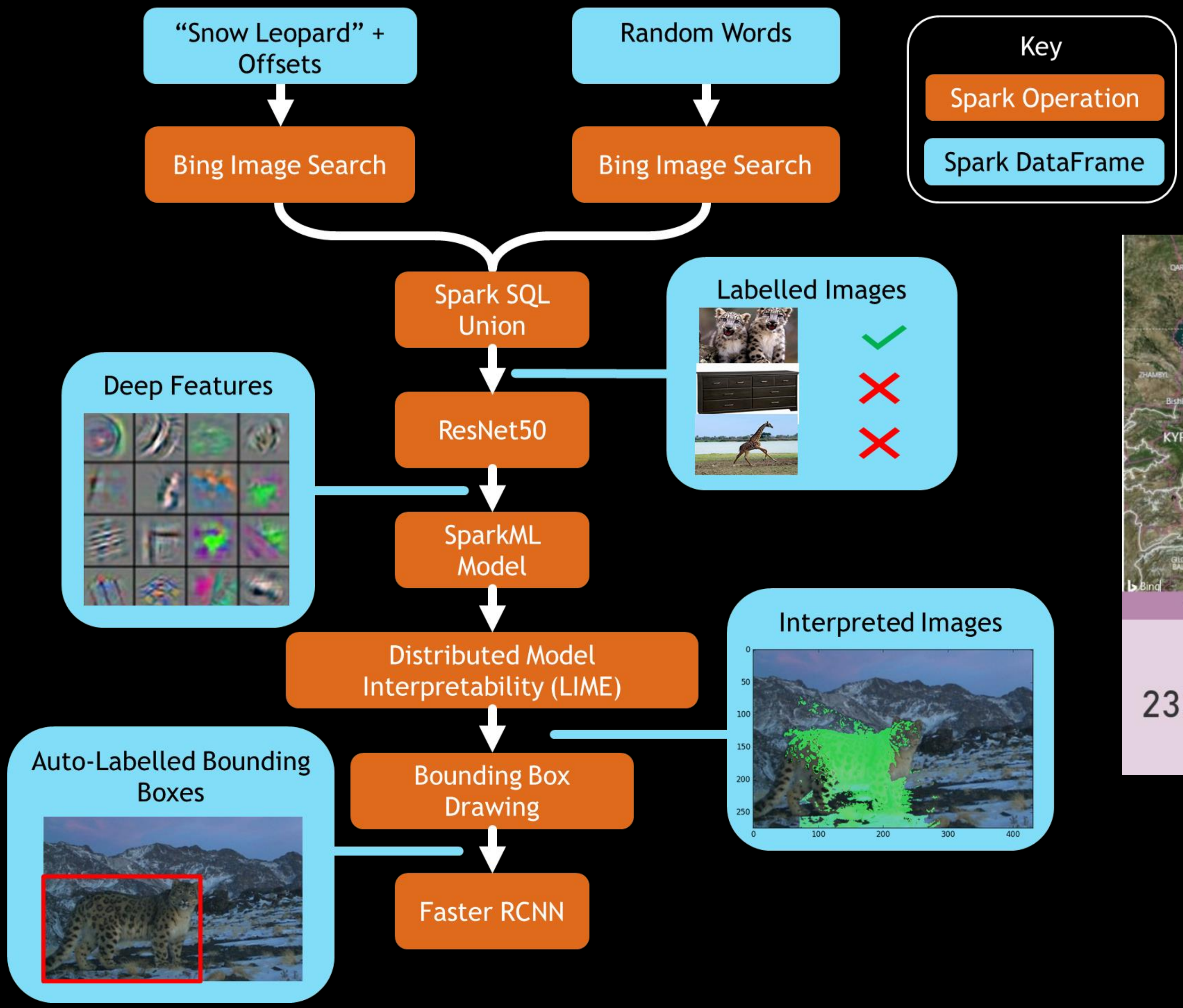
Explanation



LIME on Spark



End to End Architecture



“Snow Leopard” + Offsets

Random Words

Key

Spark Operation

Spark DataFrame

Bing Image Search

Bing Image Search

Spark SQL Union

Labelled Images

✓

✗

✗

Deep Features

ResNet50

SparkML Model

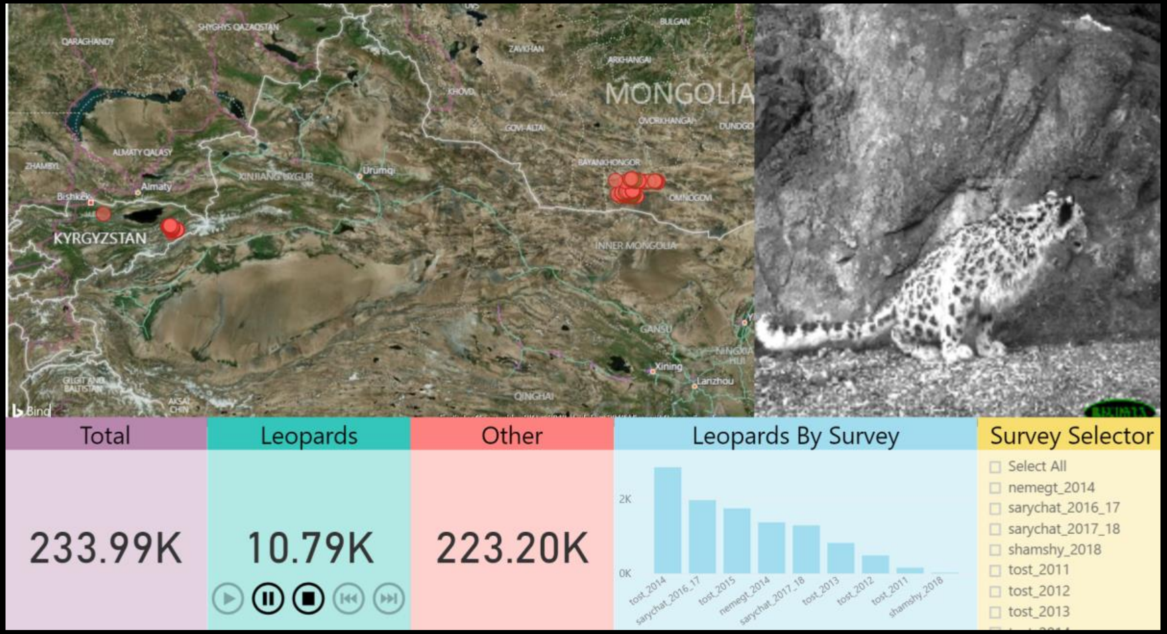
Distributed Model Interpretability (LIME)

Interpreted Images

Auto-Labelled Bounding Boxes

Bounding Box Drawing

Faster RCNN



Total	Leopards	Other
233.99K	10.79K	223.20K

Leopards By Survey

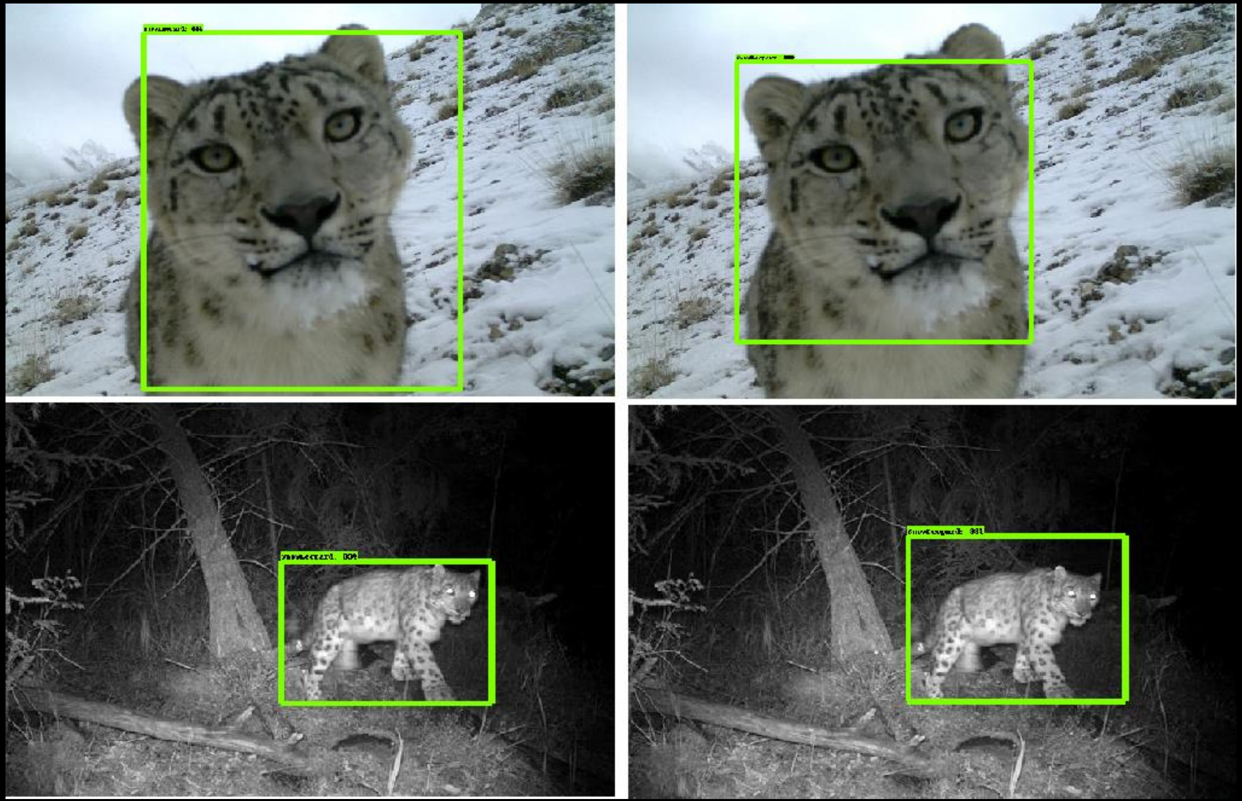
Survey Selector

- □ Select All
- □ nemegt_2014
- □ sarychat_2016_17
- □ sarychat_2017_18
- □ shamshy_2018
- □ tost_2011
- □ tost_2012
- □ tost_2013

Results

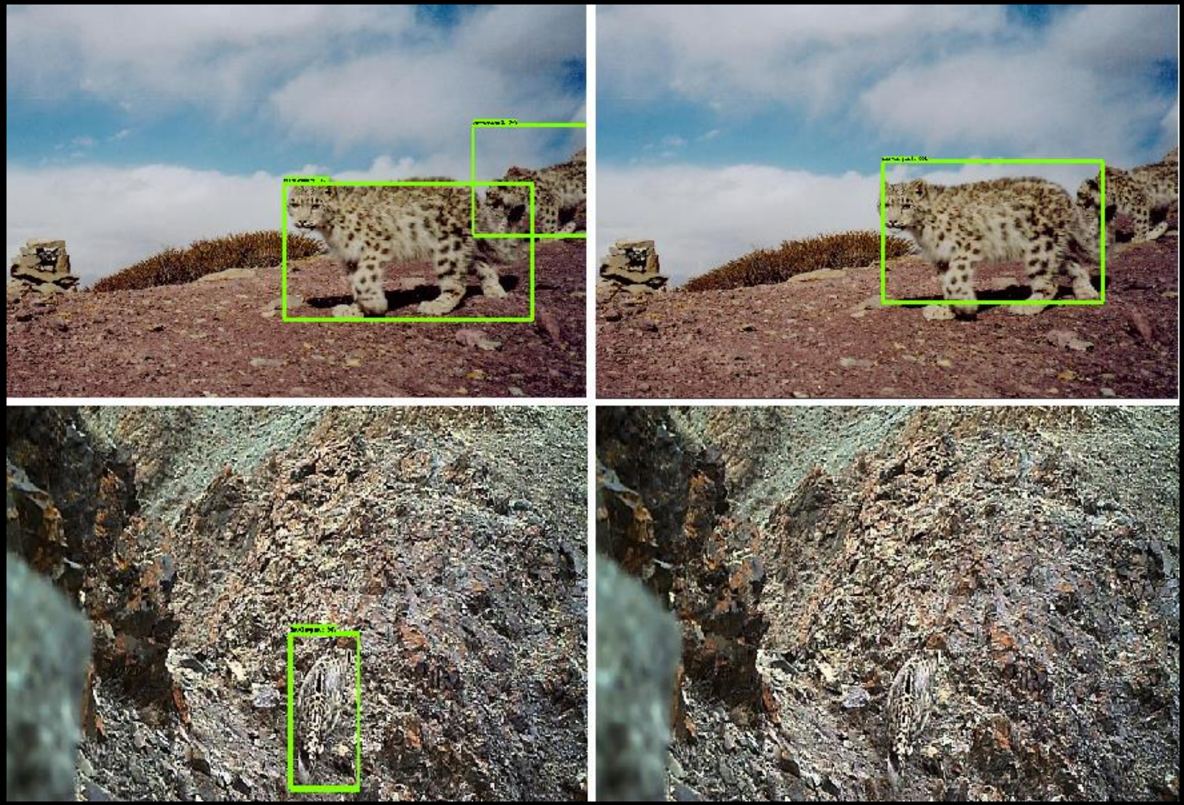
Human Labels

Unsupervised FRCNN Outputs



Human Labels

Unsupervised FRCNN Outputs



Microsoft Machine Learning for Apache Spark v0.18

Microsoft's Open Source Contributions to Apache Spark

Distributed Machine Learning

Fast Model Deployment

Microservice Orchestration

Multilingual Binding Generation

www.aka.ms/spark

Azure/mmlspark

Thanks to

- You all!
- **Ilya Matiach: LightGBM on Spark**
- **Markus Cozowicz: VW on Spark**
- Sudarshan Raghunathan, Christina Lee, Daniel Ciborowski, Eli Barzilay, Tong Wen, Pablo Castro, Chris Hoder, Ryan Gaspar, Henrik Neilsen, Andrew Schonhoffer, Joseph Sirosh
- Microsoft NERD Garage Team + MIT Externship Program
- Snow Leopard Trust: Koustubh Sharma, Rhetick Sengupta, Jeff Brown, Michael Despines
- Microsoft Development Acceleration Team:
  - Dalitso Banda, Casey Hong, Karthik Rajendran, Manon Knoertzer, Tayo Amuneke, Alejandro Buendia
- Azure CAT, AzureML, and Azure Search Teams

Get in Touch

- Support: mmlspark-support@microsoft.com
- Me: marhamil@microsoft.com
- Github : Azure/mmlspark
- Website: www.aka.ms/spark
- Paper: www.aka.ms/spark-paper
- Contributions Welcome!
- Check out our MSR Podcast on Oct 2

Backup Slides

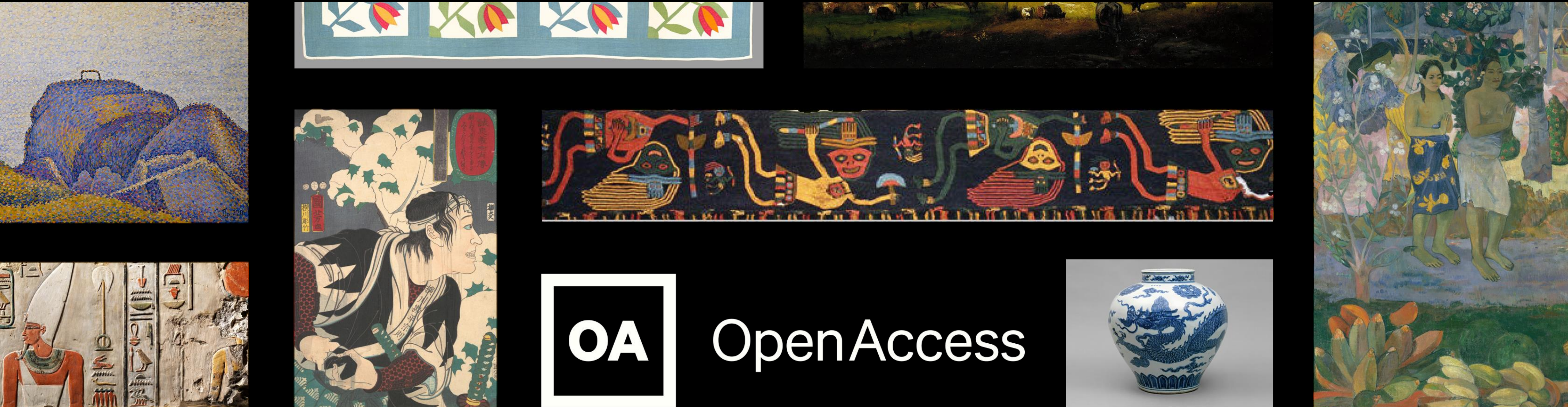
AI for Cultural Institutions

Microsoft

THE MET

Celebrating 2 years of Open Access at The MET

- In 2016 The MET Released 400k images under open access
- This past winter the MET released a new subject-keyword dataset of image annotations
- MIT, The MET, and Microsoft participated in a 3-day hackathon to create intelligent experiences using the new collection



Goals:

Create new works of art

Use new work to explore existing art

Explore further with intelligent search

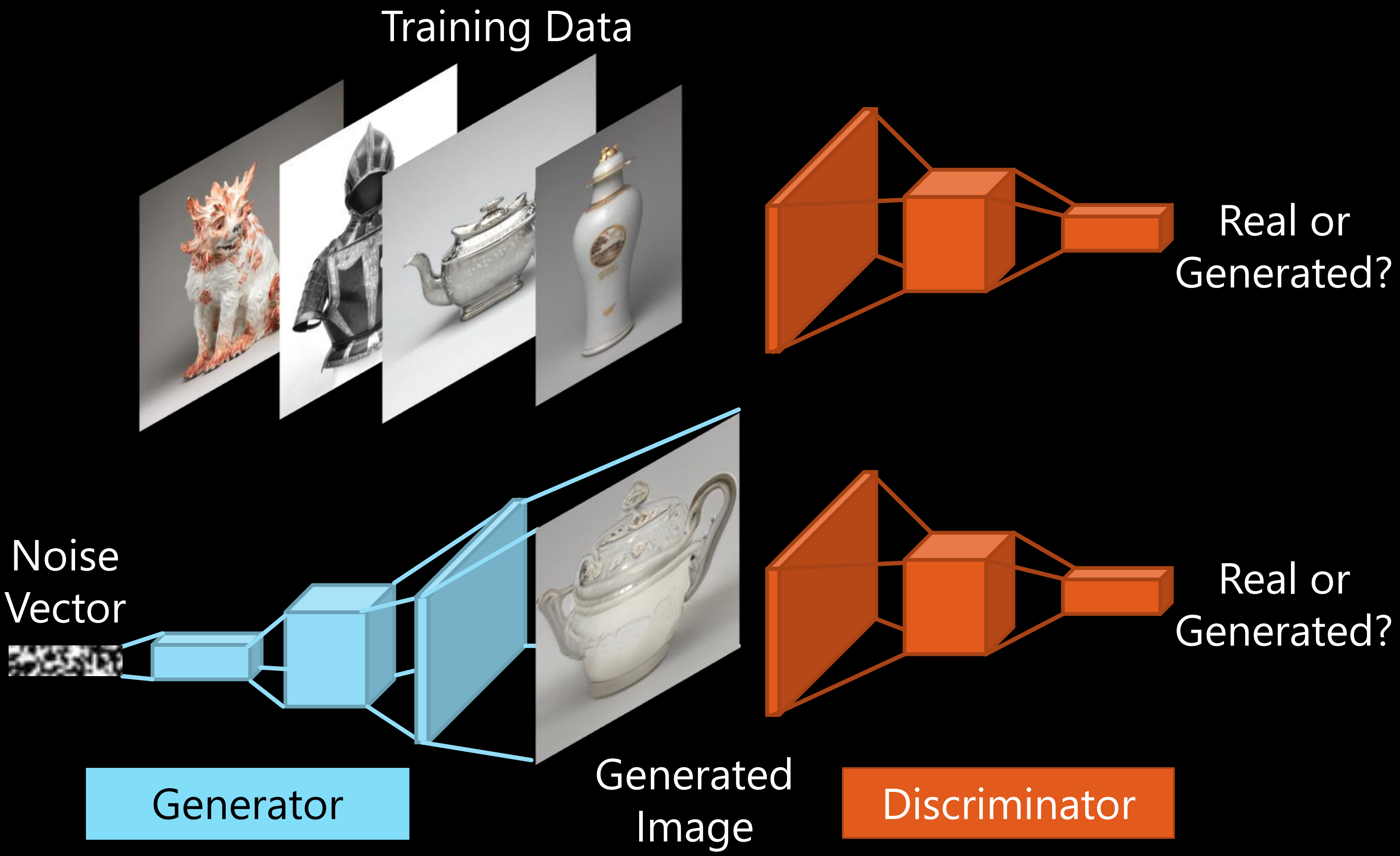
Needed Technologies:

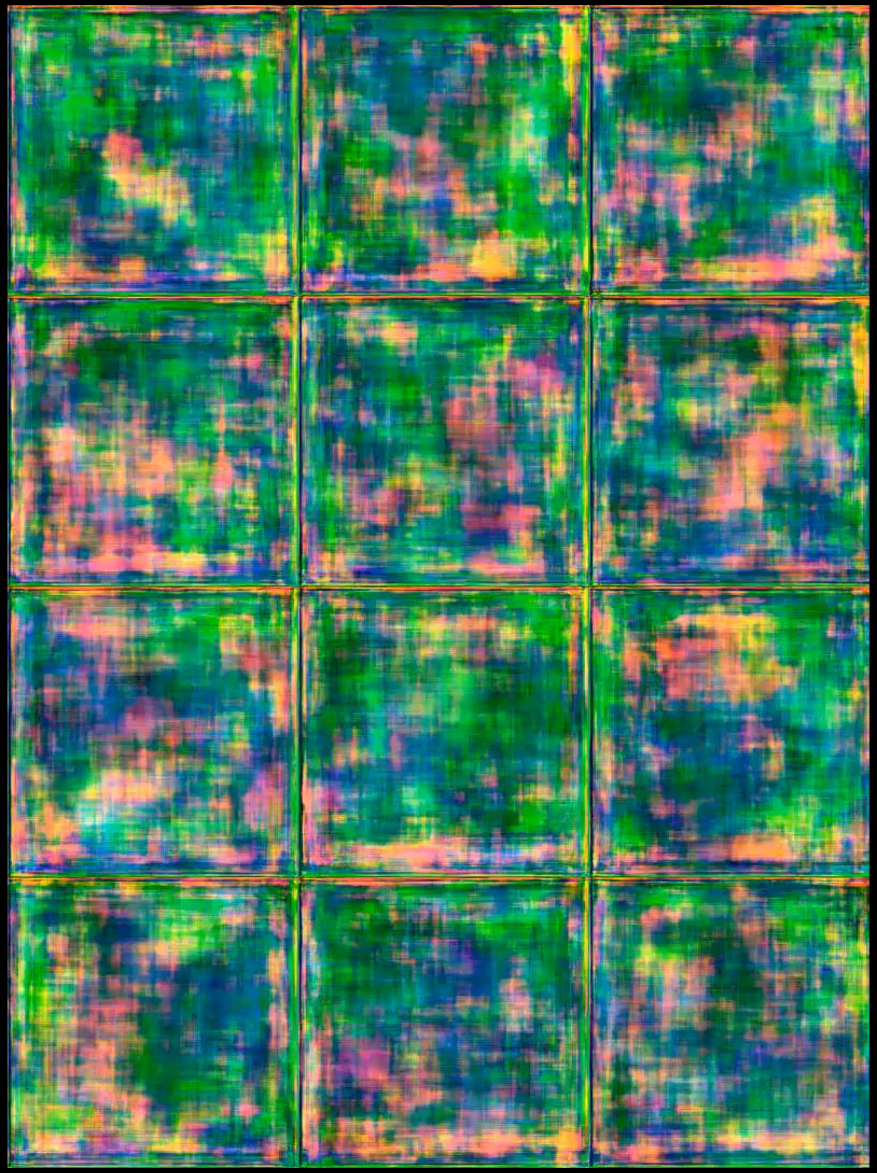
Generative Adversarial Networks

Reverse image search

Elasticsearch with Cognitive Services

Generative Adversarial Art





Custom Reverse Image Search

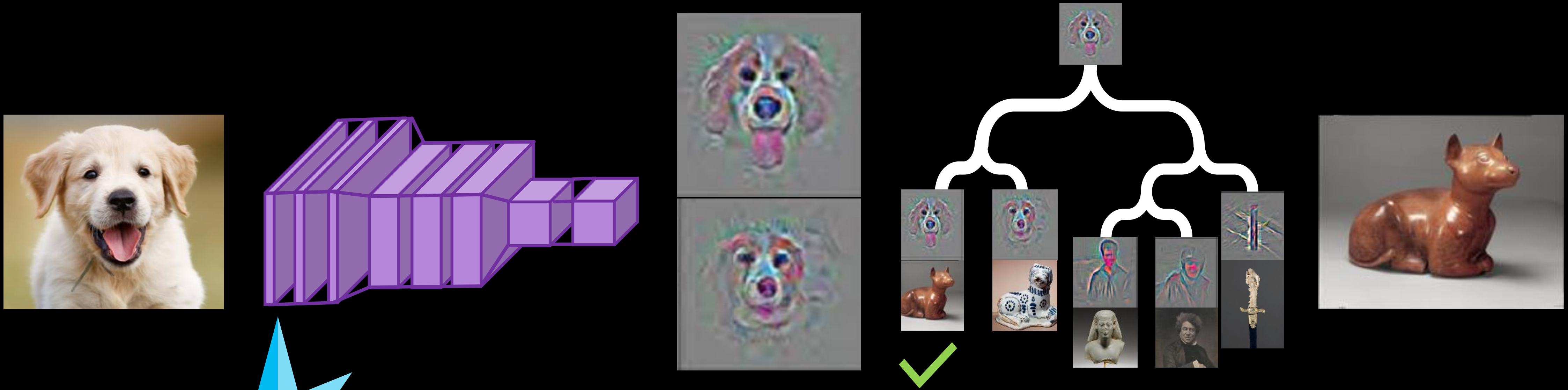
Query Image

ResNet Featurizer

Deep Features

Fast Nearest Neighbor Lookup

Closest Match

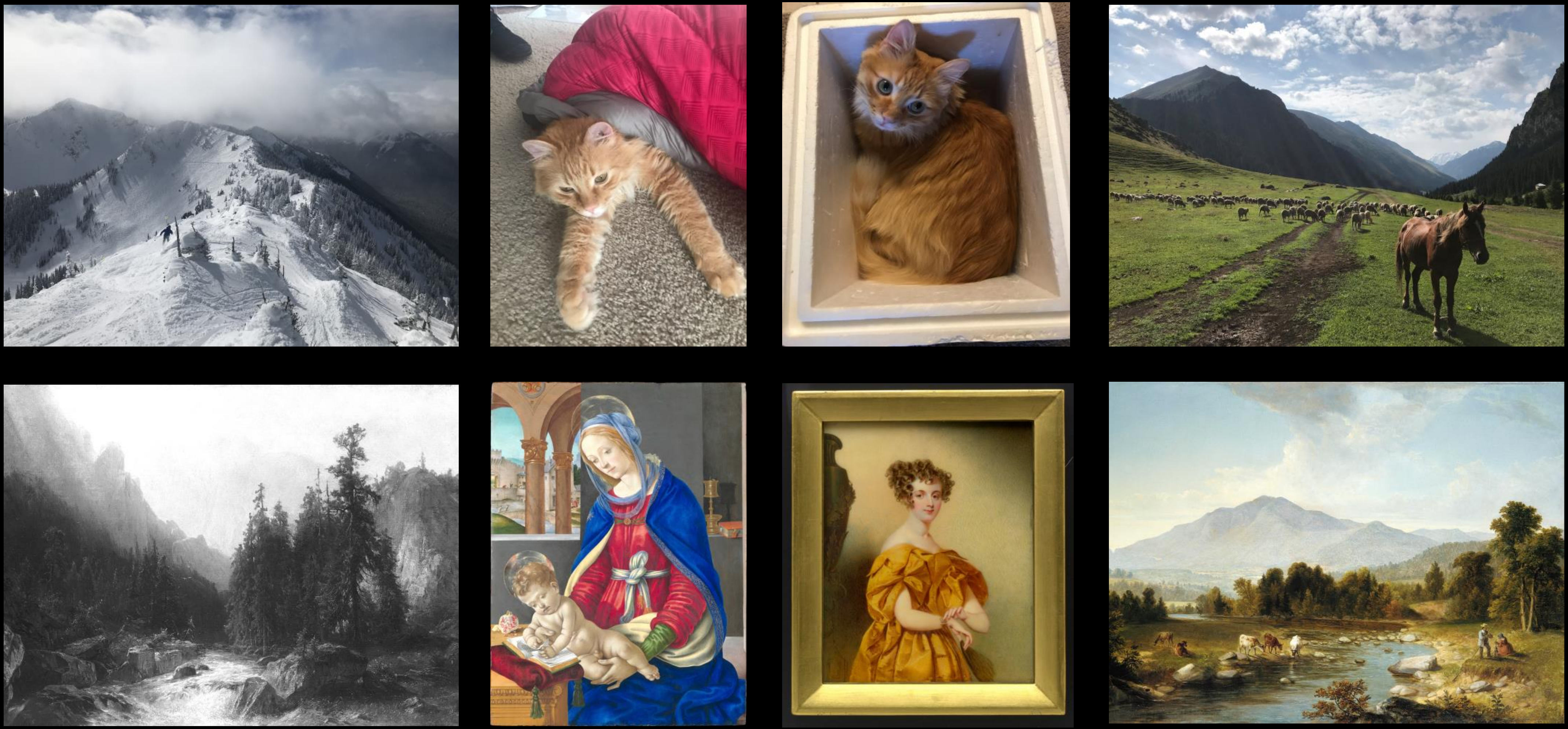


MMLSpark

SparkML LSH or Annoy

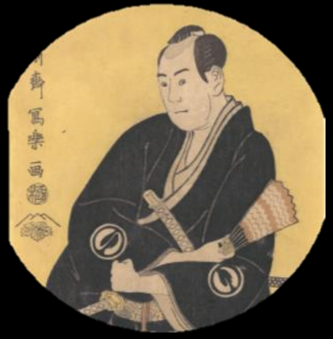
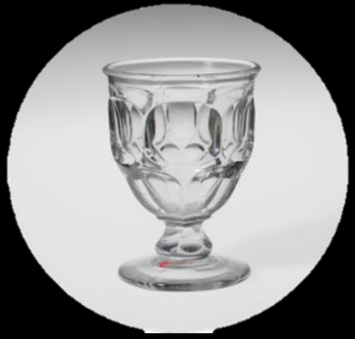
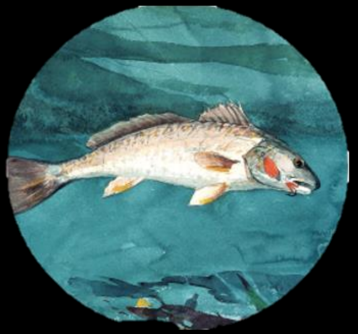
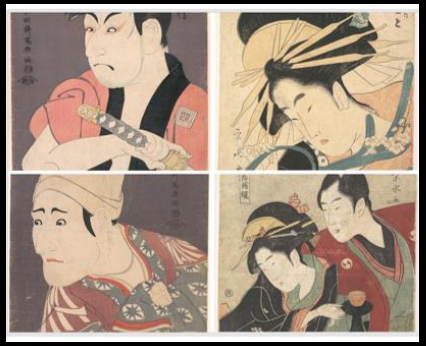
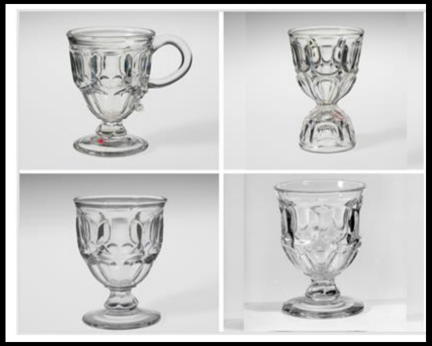
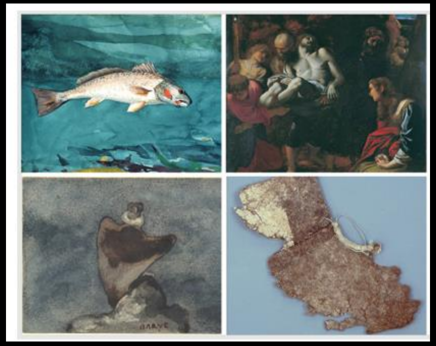
Filters from Zeiler + Fergus 2013

Example Nearest Neighbors

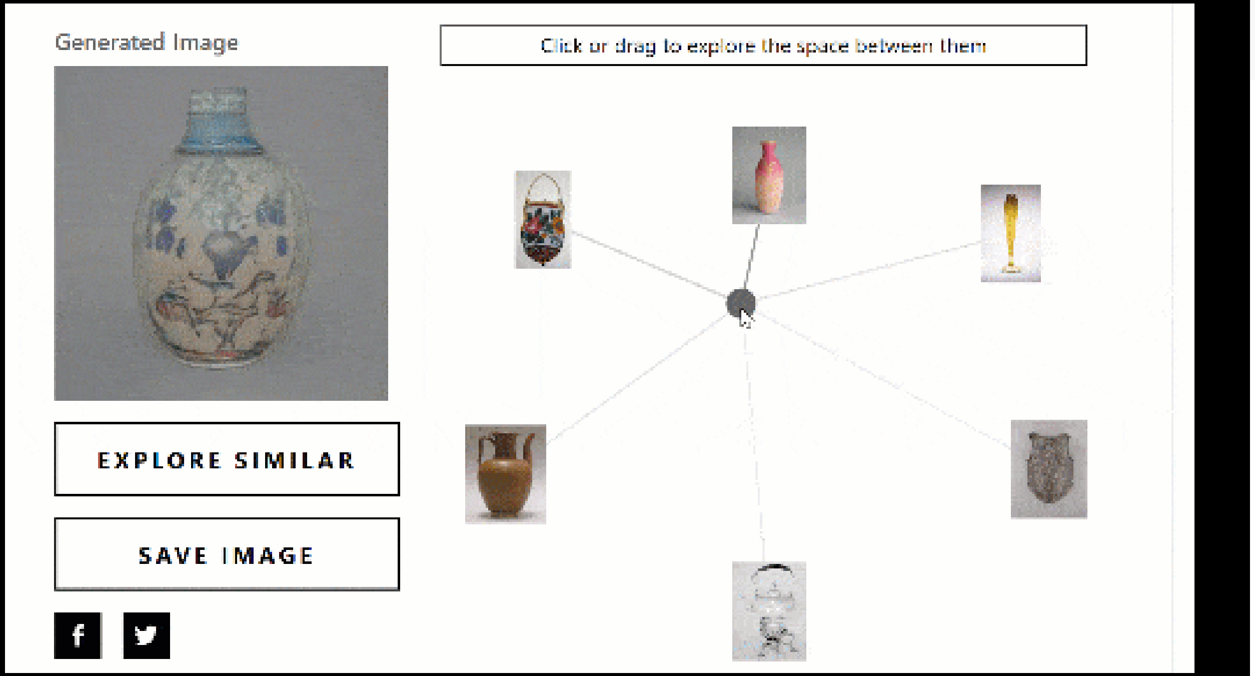


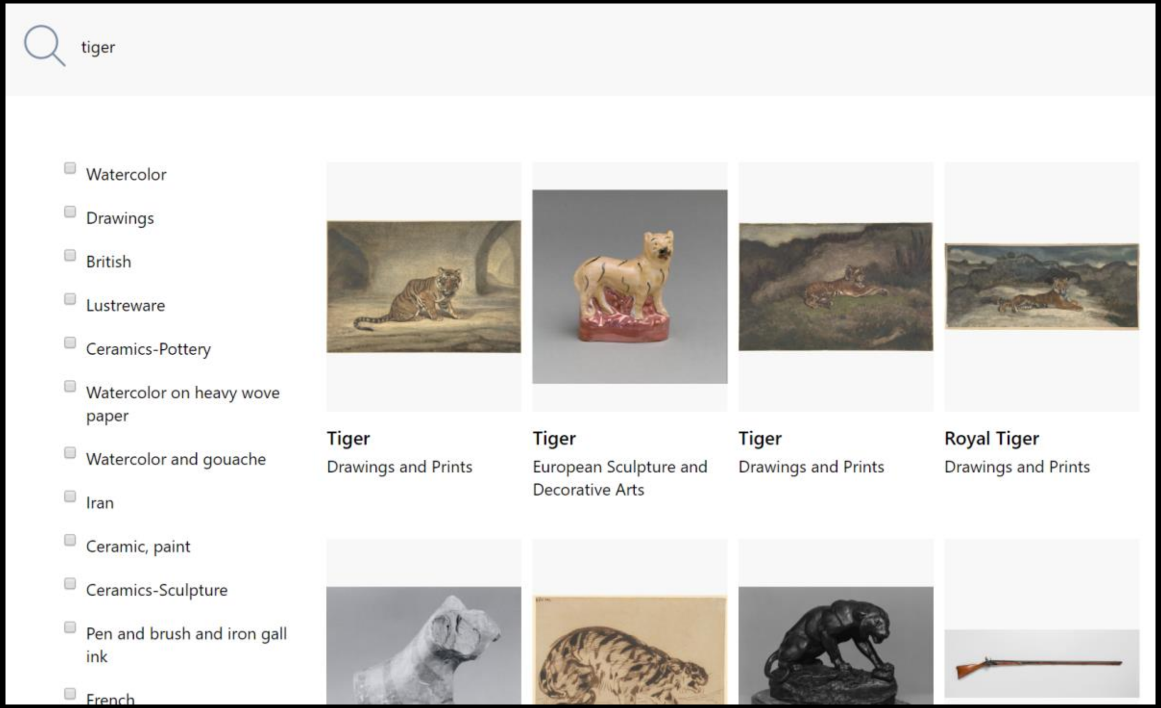
Intelligent Search Index

- Pipe images through Computer Vision API to annotate image for searching
- Stream images and intelligent annotations to Azure Search

Query Image:			
Describe Image Output:	A picture containing a person	A picture containing a glass, cup	A fish swimming underwater
Deep Feature Nearest Neighbors:			

End Application: Gen Studio




tiger

Tiger
Drawings and Prints

Tiger
European Sculpture and Decorative Arts

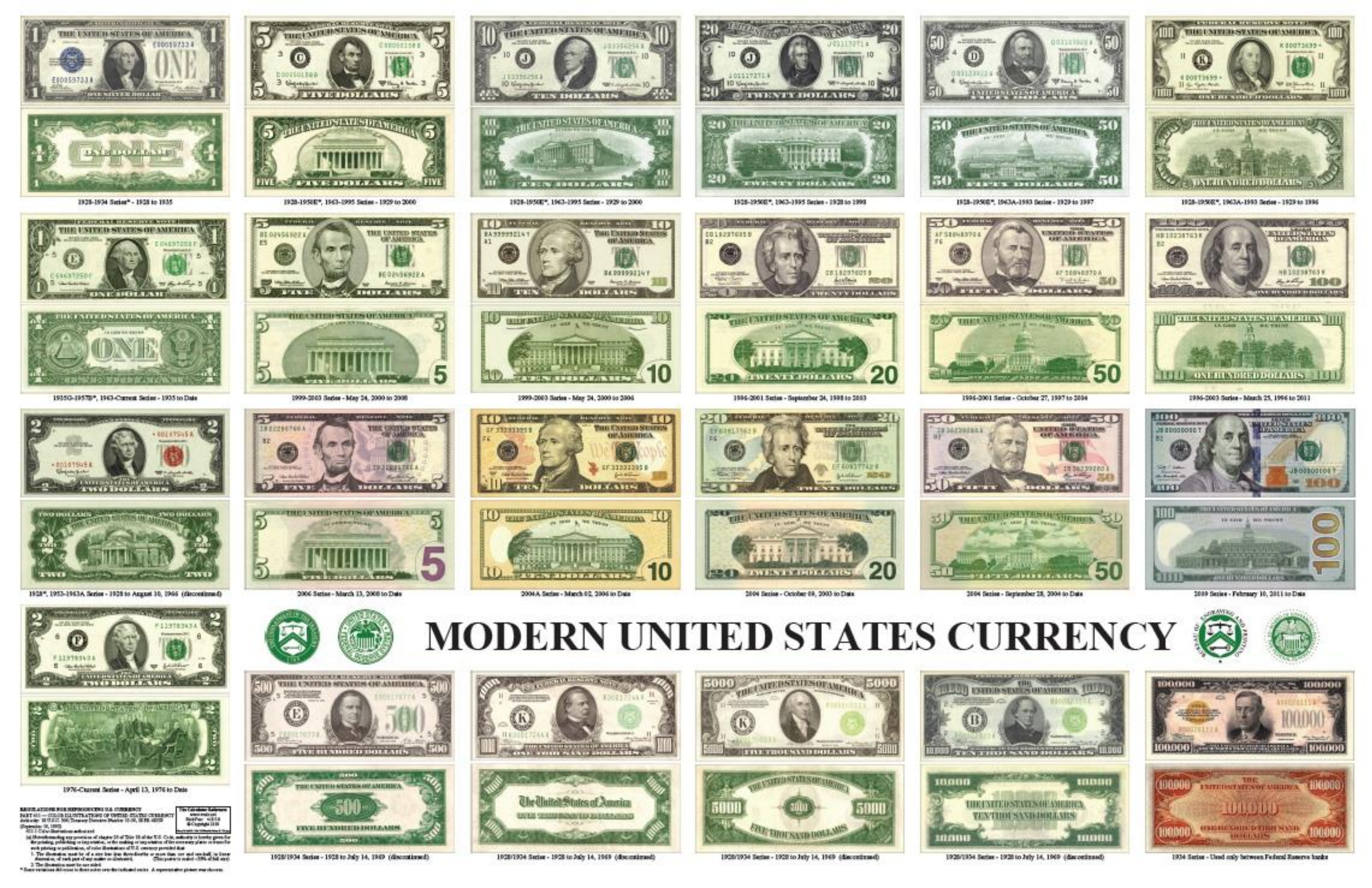
Tiger
Drawings and Prints

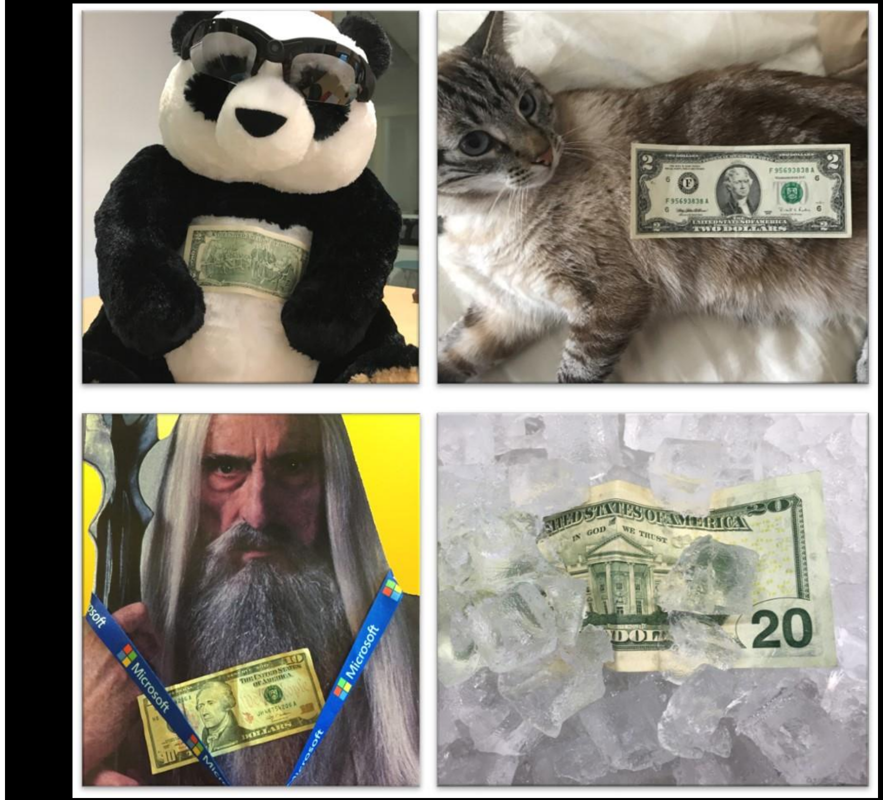
Royal Tiger
Drawings and Prints

AI for Accessibility

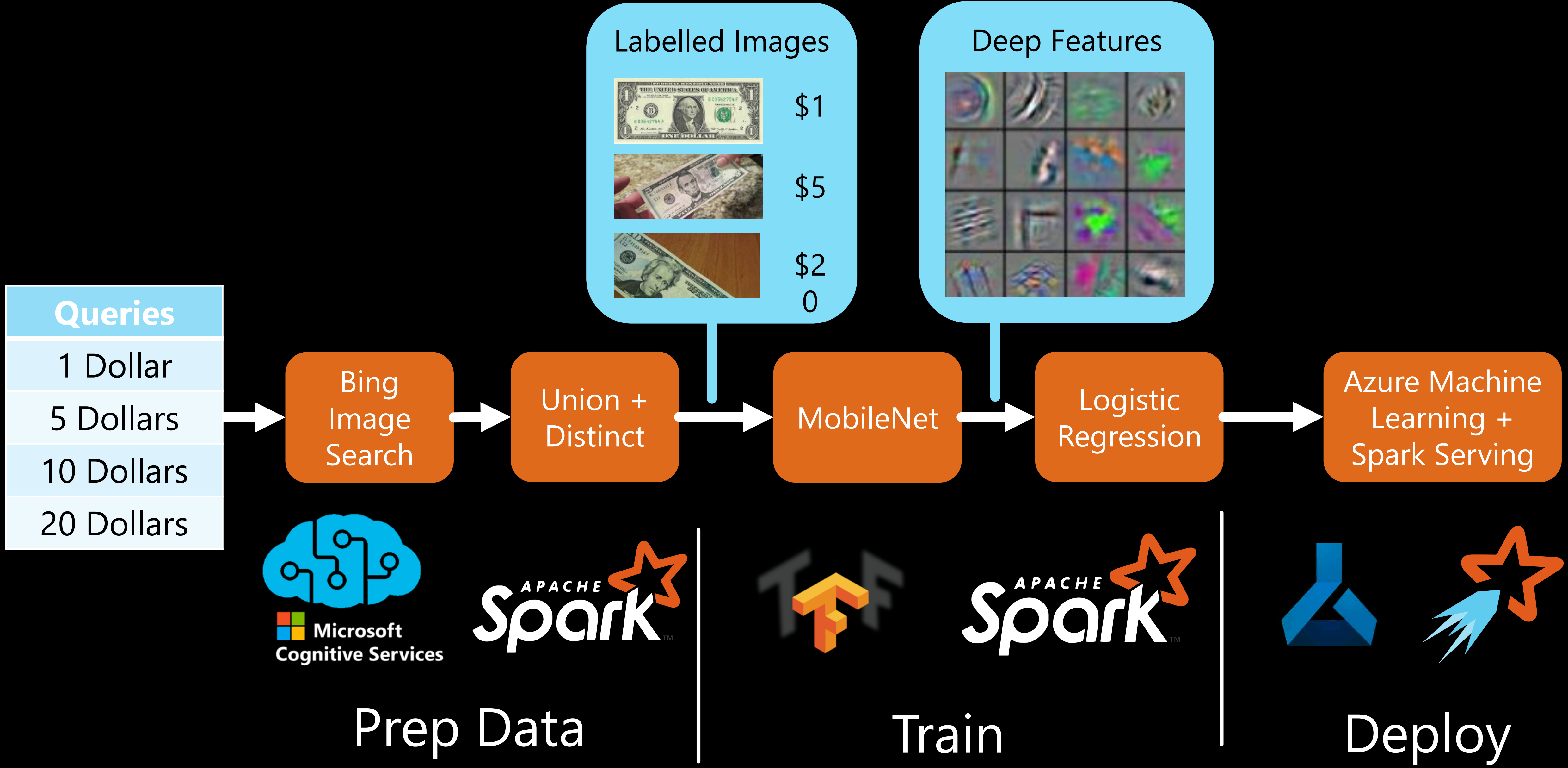
Seeing AI

Currency Identification


1928-1934 Series* - 1928 to 1935
1928-1950E*, 1963-1995 Series - 1929 to 2000
1928-1950E*, 1963-1995 Series - 1929 to 2000
1928-1950E*, 1963-1995 Series - 1928 to 1998
1928-1950E*, 1963A-1993 Series - 1929 to 1997
1928-1950E*, 1963A-1993 Series - 1929 to 1996
1935G-1957B*, 1963-Current Series - 1935 to Date
1999-2003 Series - May 24, 2000 to 2008
1999-2003 Series - May 24, 2000 to 2006
1996-2001 Series - September 24, 1998 to 2003
1996-2001 Series - October 27, 1997 to 2004
1996-2003 Series - March 25, 1996 to 2011
1928*, 1953-1963A Series - 1928 to August 10, 1966 (discontinued)
2006 Series - March 13, 2008 to Date
2004A Series - March 02, 2006 to Date
2004 Series - October 09, 2003 to Date
2004 Series - September 28, 2004 to Date
2009 Series - February 10, 2011 to Date
MODERN UNITED STATES CURRENCY
1976-Current Series - April 13, 1976 to Date
1928/1934 Series - 1928 to July 14, 1969 (discontinued)
1928/1934 Series - 1928 to July 14, 1969 (discontinued)
1928/1934 Series - 1928 to July 14, 1969 (discontinued)
1928/1934 Series - 1928 to July 14, 1969 (discontinued)
1934 Series - Used only between Federal Reserve banks
REGULATIONS FOR REPRODUCING U.S. CURRENCY



A Familiar Architecture...



Queries
1 Dollar
5 Dollars
10 Dollars
20 Dollars

Bing Image Search → Union + Distinct → MobileNet → Logistic Regression → Azure Machine Learning + Spark Serving

Labelled Images

$1

$5

$20

Deep Features

Prep Data

Train

Deploy