Wreckignition

Akshay Tiwari, David Kes, Kishan Panchal, Stephen Hsu

Project Objective and Motivation

To predict the age of an individual from an image of their face

Possible Use Cases :

- Consumer profiling at retail stores
- Targeted marketing as per the age group
- Personalized experiences based on age on web applications

Our Motivation

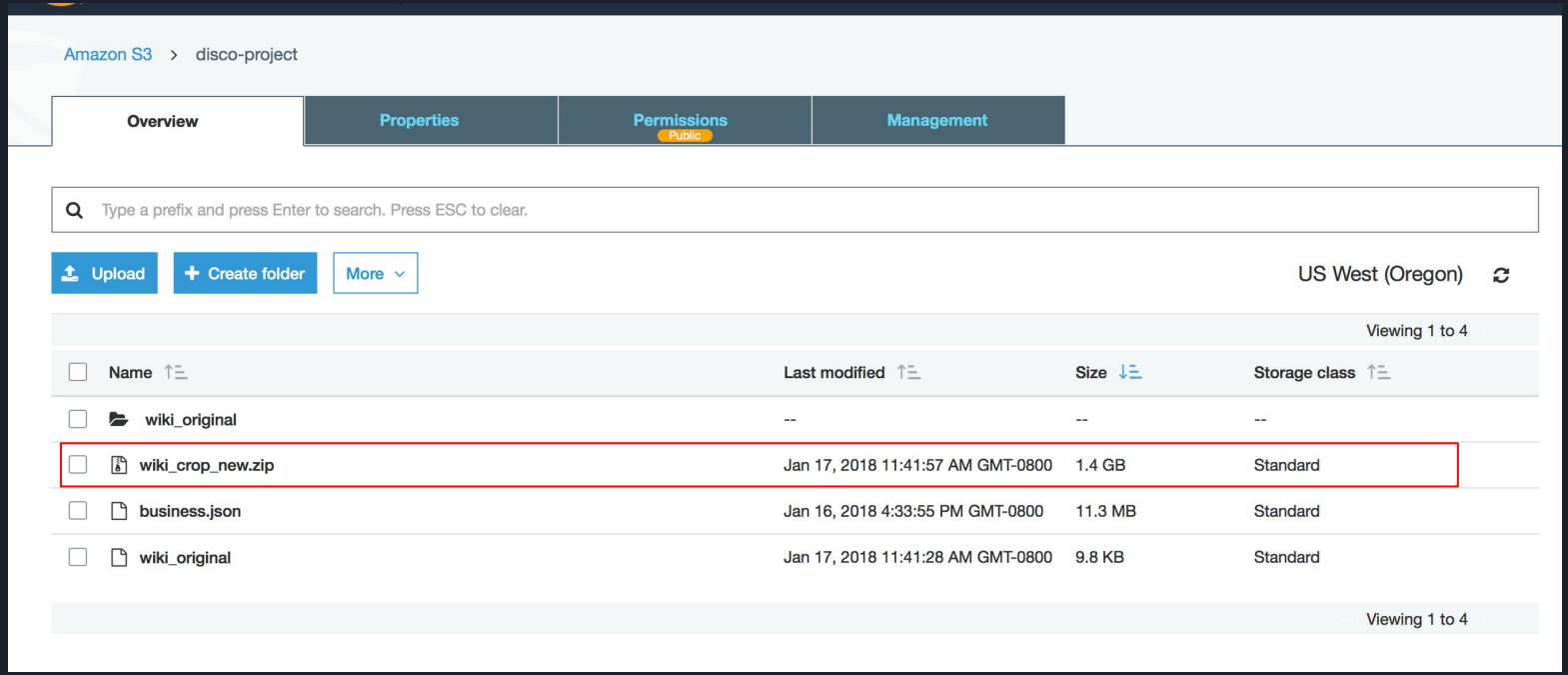
- See how much older a beard makes you look

Data Description

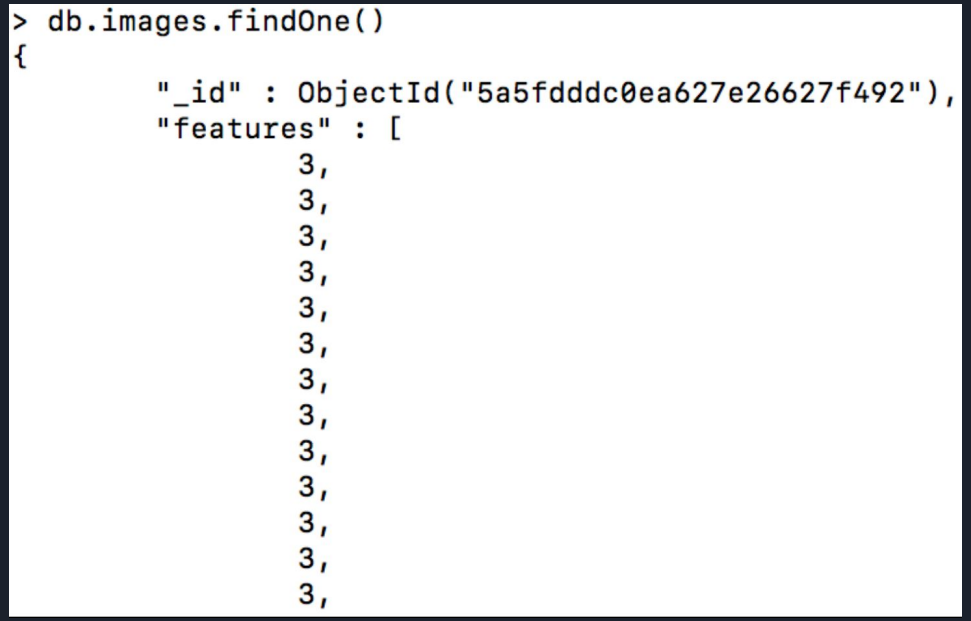
- 1.4 gb
- 124,368 photos
- .jpg format
- Images from imdb and wikipedia

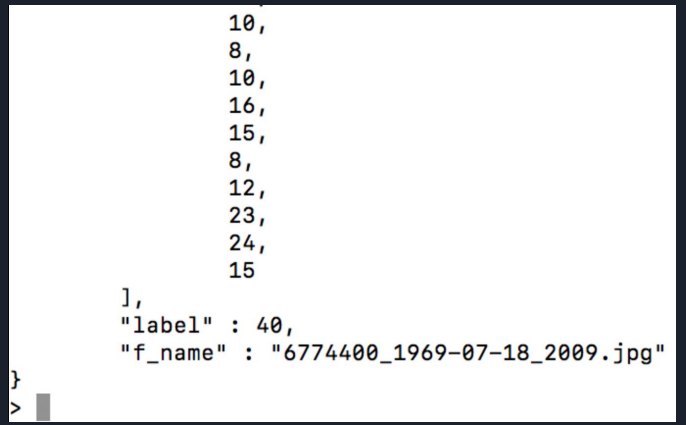
Source: https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

Data on S3


Amazon S3 > disco-project
Overview
Properties
Permissions
Public
Management
Type a prefix and press Enter to search. Press ESC to clear.
Upload
Create folder
More
US West (Oregon)
Viewing 1 to 4
Name
Last modified
Size
Storage class
wiki_original
--
--
--
wiki_crop_new.zip
Jan 17, 2018 11:41:57 AM GMT-0800
1.4 GB
Standard
business.json
Jan 16, 2018 4:33:55 PM GMT-0800
11.3 MB
Standard
wiki_original
Jan 17, 2018 11:41:28 AM GMT-0800
9.8 KB
Standard
Viewing 1 to 4

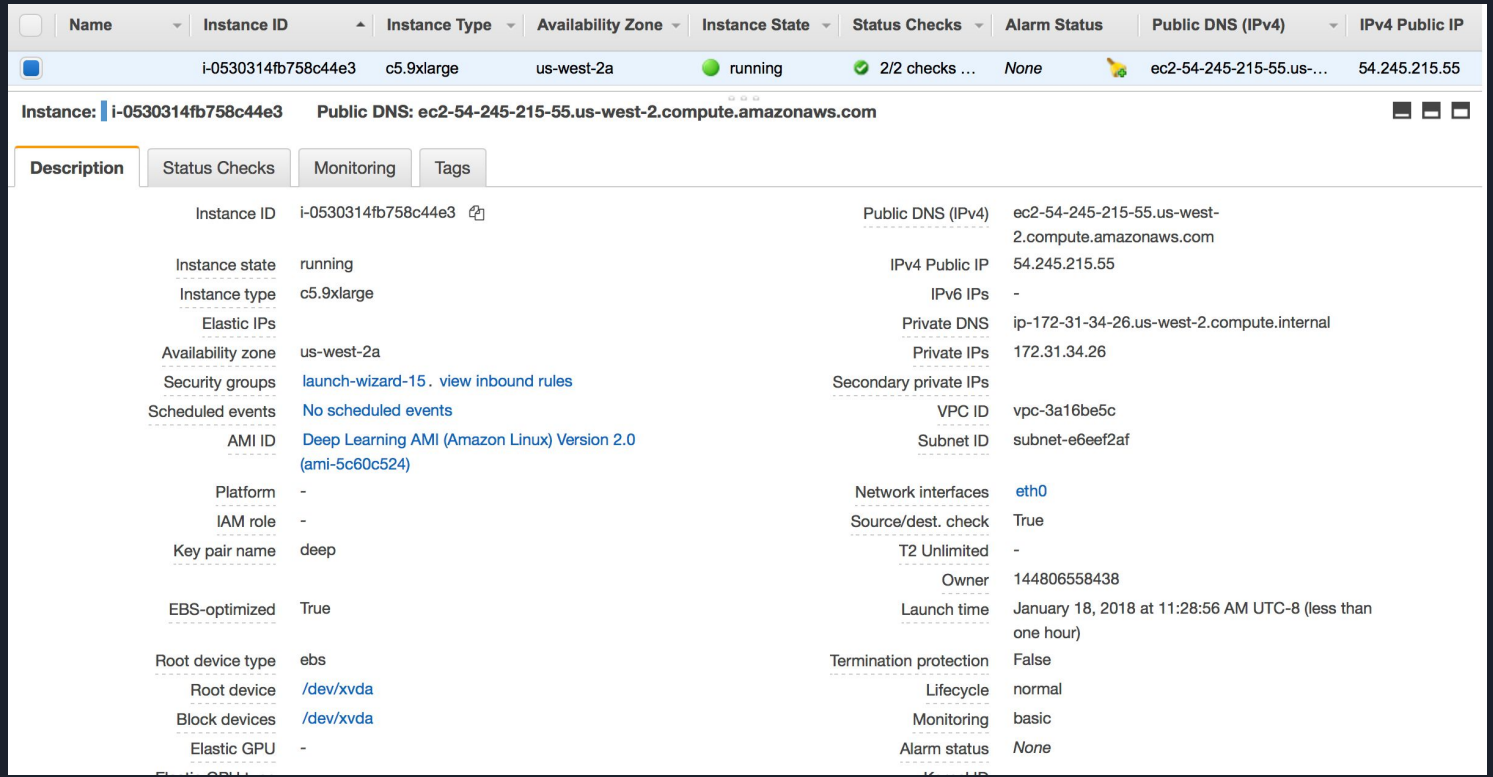
MongoDB





Screenshots of the top and bottom parts of output since there are 16,384 pixels

EC2 Specs



Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP
	i-0530314fb758c44e3	c5.9xlarge	us-west-2a	running	2/2 checks ...	None	ec2-54-245-215-55.us-...	54.245.215.55

Instance: i-0530314fb758c44e3 **Public DNS: ec2-54-245-215-55.us-west-2.compute.amazonaws.com**

Description | Status Checks | Monitoring | Tags

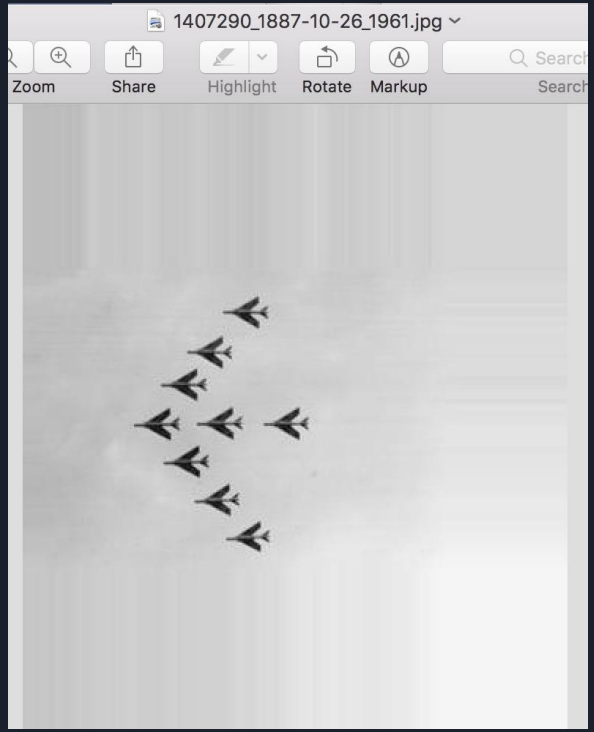
Instance ID	i-0530314fb758c44e3	Public DNS (IPv4)	ec2-54-245-215-55.us-west-2.compute.amazonaws.com
Instance state	running	IPv4 Public IP	54.245.215.55
Instance type	c5.9xlarge	IPv6 IPs	-
Elastic IPs		Private DNS	ip-172-31-34-26.us-west-2.compute.internal
Availability zone	us-west-2a	Private IPs	172.31.34.26
Security groups	launch-wizard-15. view inbound rules	Secondary private IPs	
Scheduled events	No scheduled events	VPC ID	vpc-3a16be5c
AMI ID	Deep Learning AMI (Amazon Linux) Version 2.0 (ami-5c60c524)	Subnet ID	subnet-e6eef2af
Platform	-	Network interfaces	eth0
IAM role	-	Source/dest. check	True
Key pair name	deep	T2 Unlimited	-
		Owner	144806558438
EBS-optimized	True	Launch time	January 18, 2018 at 11:28:56 AM UTC-8 (less than one hour)
Root device type	ebs	Termination protection	False
Root device	/dev/xvda	Lifecycle	normal
Block devices	/dev/xvda	Monitoring	basic
Elastic GPU	-	Alarm status	None

Data Processing

- Convert images to their pixel representations
- Calculate ages by subtracting date of birth from the date the photo was taken

Data Processing

- Filtered
  - Outliers
  - Corrupted images
- Resized photos
  - 128 x 128 pixels = 16,384 pixels
- Changed from RGB to grayscale



Dataframe Creation

```
df = spark.read.format("com.mongodb.spark.sql.DefaultSource").option("uri","mongodb://127.0.0.1/
df.show()
```

```
+--------------------+-----+--------------------+
|            features|label|              f_name|
+--------------------+-----+--------------------+
|[210, 211, 205, 2...|   67|10049200_1891-09-...|
|[59, 54, 46, 40, ...|   27|10110600_1985-09-...|
|[28, 28, 28, 28, ...|   46|10126400_1964-07-...|
|[251, 245, 154, 5...|   43|1013900_1917-10-1...|
|[148, 148, 148, 1...|   48|10166400_1960-03-...|
|[88, 88, 88, 88, ...|   38|102100_1970-10-09...|
|[172, 155, 164, 1...|   29|1024100_1982-06-0...|
|[248, 253, 253, 2...|   25|10292500_1984-03-...|
|[101, 118, 137, 1...|   68|1035700_1945-11-2...|
|[143, 143, 143, 1...|   60|10416800_1907-01-...|
|[1, 0, 0, 0, 0, 0...|   35|10525500_1916-02-...|
|[74, 74, 74, 74, ...|   64|1054800_1947-09-1...|
|[244, 242, 240, 2...|   25|10623500_1931-09-...|
|[96, 96, 96, 96, ...|   19|10726900_1991-02-...|
|[23, 23, 32, 33, ...|   42|10870400_1971-06-...|
|[255, 255, 255, 2...|    1|10898800_1951-06-...|
|[193, 194, 199, 2...|   51|10967900_1956-03-...|
|[105, 116, 123, 1...|   20|10996600_1988-06-...|
|[48, 39, 39, 45, ...|   25|11035100_1984-08-...|
|[158, 158, 158, 1...|   31|1121500_1976-07-3...|
+--------------------+-----+--------------------+
only showing top 20 rows
```

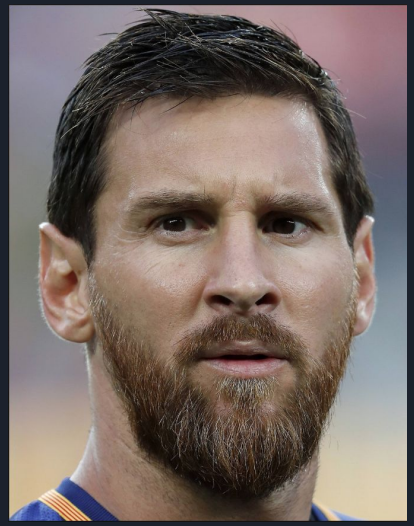
Machine Learning with Spark

```
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression(maxIter=10, fitIntercept=True)
lrmodel = lr.fit(train)
```

```
validpredict = lrmodel.transform(test)
validpredict.show()
```

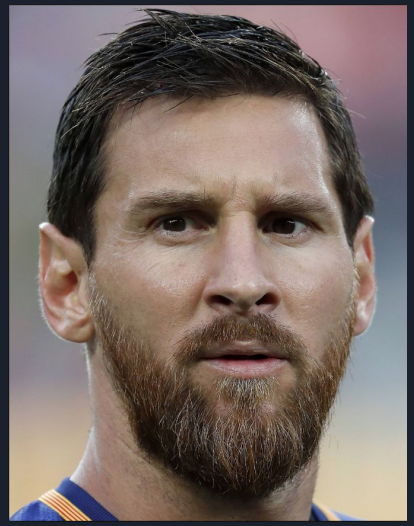
```
+--------------------+-----+--------------------+--------------------+--------------------+----------+
|            features|label|              f_name|       rawPrediction|         probability|prediction|
+--------------------+-----+--------------------+--------------------+--------------------+----------+
|[35.0,35.0,35.0,3...|   61|10000029563647_19...|[-6.8463414207276...|[1.73151962189526...|      38.0|
|[67.0,67.0,67.0,6...|   21|31843216_1990-06-...|[-8.4015258969212...|[1.57755623720897...|      20.0|
|[70.0,69.0,70.0,7...|   30|680054_1980-04-22...|[-6.8102111892276...|[2.69520661417808...|      24.0|
|[103.0,103.0,102....|   28|1000001462396_198...|[-6.8023374140069...|[4.59565307231250...|      26.0|
|[107.0,110.0,110....|   22|10000034806862_19...|[-5.9757387137851...|[1.68526814958488...|      26.0|
|[138.0,137.0,142....|   23|19187533_1960-09-...|[-9.1310373750813...|[1.13957032763565...|      23.0|
|[0.0,0.0,0.0,0.0,...|   22|10000038381611_19...|[-1.6583188926933...|[0.00114766716239...|      24.0|
|[0.0,0.0,0.0,0.0,...|   27|26472686_1945-06-...|[-1.6583188926933...|[0.00114766716239...|      24.0|
|[9.0,11.0,15.0,16...|   45|11989440_1963-05-...|[-5.4610321967234...|[3.67441441861105...|      24.0|
|[64.0,60.0,56.0,7...|   23|1000003786440_194...|[-6.8831384133220...|[1.02834016861535...|      29.0|
|[148.0,160.0,164....|   22|10000010188227_19...|[-7.0117629701875...|[1.54456932767558...|      24.0|
|[158.0,158.0,158....|   38|3073788_1963-07-2...|[-6.9809366168129...|[2.13630314581950...|      39.0|
|[162.0,161.0,161....|   63|1000003648861_194...|[-7.2180631630133...|[2.77020444094900...|      24.0|
|[184.0,183.0,183....|   28|1000004111781_198...|[-8.9626908732952...|[6.03601725909524...|      34.0|
|[206.0,206.0,206....|   20|10000036749635_19...|[-7.3814005863340...|[5.89397355545400...|      23.0|
|[22.0,26.0,32.0,3...|   19|100000679542_1974...|[-4.0980093097604...|[2.41935949336375...|      43.0|
|[169.0,158.0,149....|   22|10000039658977_19...|[-6.3305491154704...|[1.02846832448655...|      24.0|
|[174.0,173.0,172....|   19|10000034785683_19...|[-6.7499471833244...|[3.04088607067367...|      21.0|
|[210.0,204.0,205....|   30|10000035072071_19...|[-10.506759506541...|[2.00713806130717...|      47.0|
|[55.0,53.0,51.0,5...|   35|1000008584471_197...|[-6.7005840200514...|[2.82929443090967...|      43.0|
+--------------------+-----+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

Test Inputs



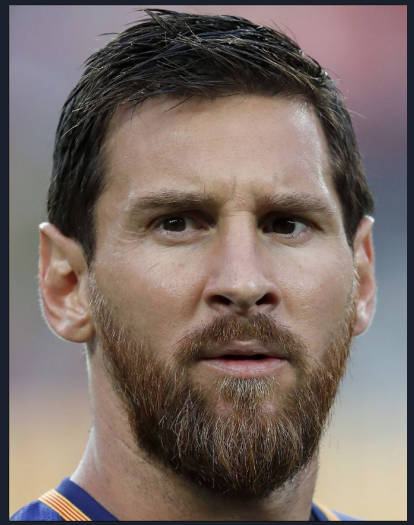
	Predicted Age	Actual Age
No Beard		
Beard		

Test Inputs



	Predicted Age	Actual Age
No Beard		29
Beard		30

Test Input Predictions



	Predicted Age	Actual Age
No Beard	29	29
Beard	48	30

Test Input

	Predicted Age	Actual Age
No Beard		29
Beard		31

Test Input Predictions

	Predicted Age	Actual Age
No Beard	21	29
Beard	45	31

Results

```
validpredict = lrmodel.transform(df)
validpredict.show()
```

```
+--------------------+-----+--------------------+--------------------+----------+
|            features|label|       rawPrediction|         probability|prediction|
+--------------------+-----+--------------------+--------------------+----------+
|[255.0,255.0,255....|   29|[15.0964586767863...|[2.48927573228792...|      29.0|
|[79.0,108.0,105.0...|   31|[7.89848711842482...|[0.00576605419693...|      21.0|
|[118.0,135.0,133....|   30|[8.52674352613275...|[0.00369337149909...|      48.0|
|[22.0,21.0,21.0,2...|   32|[5.32746676125807...|[0.00202276754133...|      45.0|
+--------------------+-----+--------------------+--------------------+----------+
```

BONUS ROUND

	Predicted Age	Actual Age
No Beard		???
Beard		???

	Predicted Age	Actual Age
No Beard	27	???
Beard	37	???

Lessons Learned

- Images are difficult to deal with in pre-processing but doable
- You should zip your data folder prior to uploading it to S3
- Setting up everything on EC2 instances is very time consuming
- Unstructured data can still be interpreted via PySpark even if not all deep learning algorithms are supported in Spark ML