Scaling Data Products Under Startup Constraints

A Case Study of ML Bias Testing

TinyData

Scaling Data Products Under ~~Startup~~ Constraints

...

A Case Study of ML Bias Testing

TinyData

Edwin Ong @edwin

Co-Founder, TinyData

Founded CastTV (acquired by Tribune)

Founded FileFish (acquired by Oracle)

Stanford Symbolic Systems

TinyData

TinyData

- Help other companies make data products

- Make our own data products

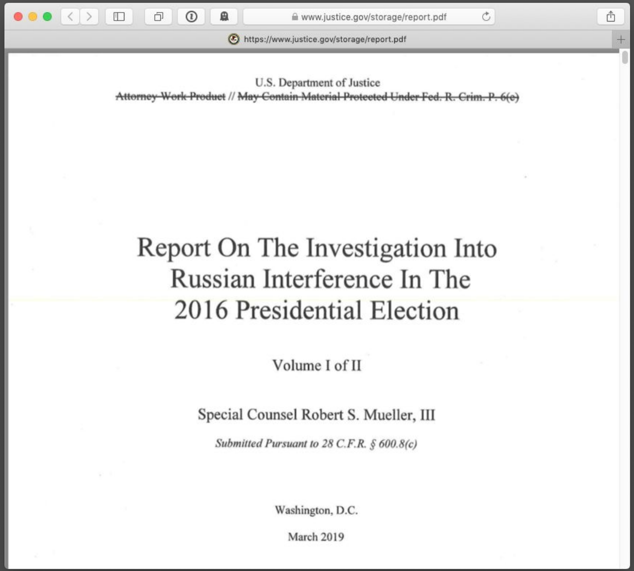
BLOCKSPUR Dashjoy

Problem: Testing Machine Learning in Production

- Tools for machine learning testing *in training*
- Not as many tools for machine learning testing in production
- Different tools needed because ML testing is different from traditional software testing

TinyData

Traditional Software Has Deterministic Outcomes



U.S. Department of Justice
~~Attorney Work Product~~ // ~~May Contain Material Protected Under Fed. R. Crim. P. 6(e)~~

Report On The Investigation Into
Russian Interference In The
2016 Presidential Election

Volume I of II

Special Counsel Robert S. Mueller, III

Submitted Pursuant to 28 C.F.R. § 600.8(c)

Washington, D.C.

March 2019

	Request URL	Status codes	↓ Redirects
>	https://www.justice.gov/storage/report.pdf	200	0

Traditional Software Has Deterministic Outcomes

[pods/probe/exec-liveness.yaml]

```yaml
apiVersion: v1
kind: Pod
metadata:
  labels:
    test: liveness
  name: liveness-exec
spec:
  containers:
  - name: liveness
    image: k8s.gcr.io/busybox
    args:
    - /bin/sh
    - -c
    - touch /tmp/healthy; sleep 30; rm -rf /tmp/healthy; sleep 600
    livenessProbe:
      exec:
        command:
        - cat
        - /tmp/healthy
      initialDelaySeconds: 5
      periodSeconds: 5
```

```
kubectl describe pod liveness-exec
```

TinyData

ML Has Probabilistic Outcomes

Dog vs Muffin given
new user input

TinyData

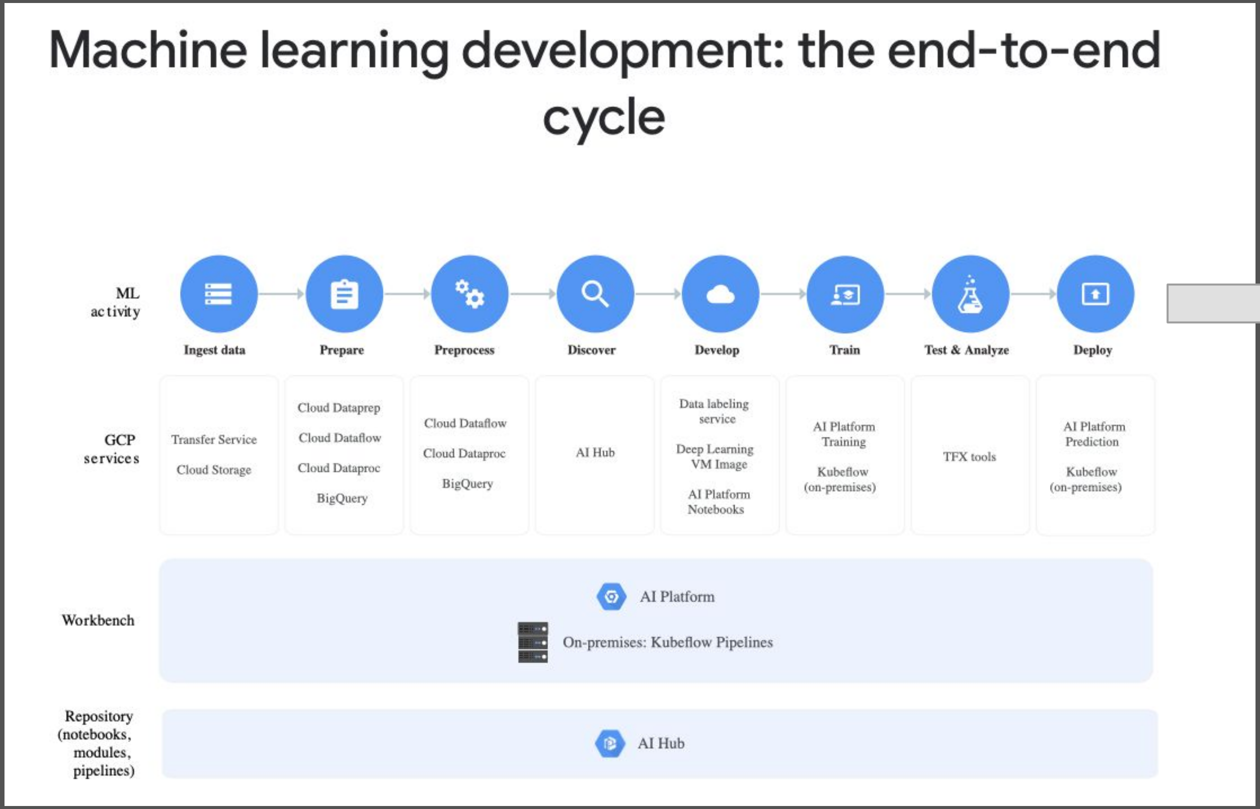
ML Has Probabilistic Outcomes That Change Over Time

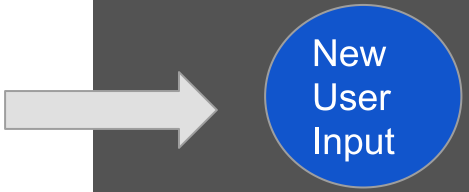
Version 1: Muffin (59%)

Version 2: Muffin (66%)

TinyData

ML Platforms Often End at Deploy





Production Testing

ML Chaos Engineering

Requirements for Production ML Testing Tool

1. "Entropy": Generation of new inputs against model servers

2. Recording of outputs from model servers

3. Feedback loop for additional training

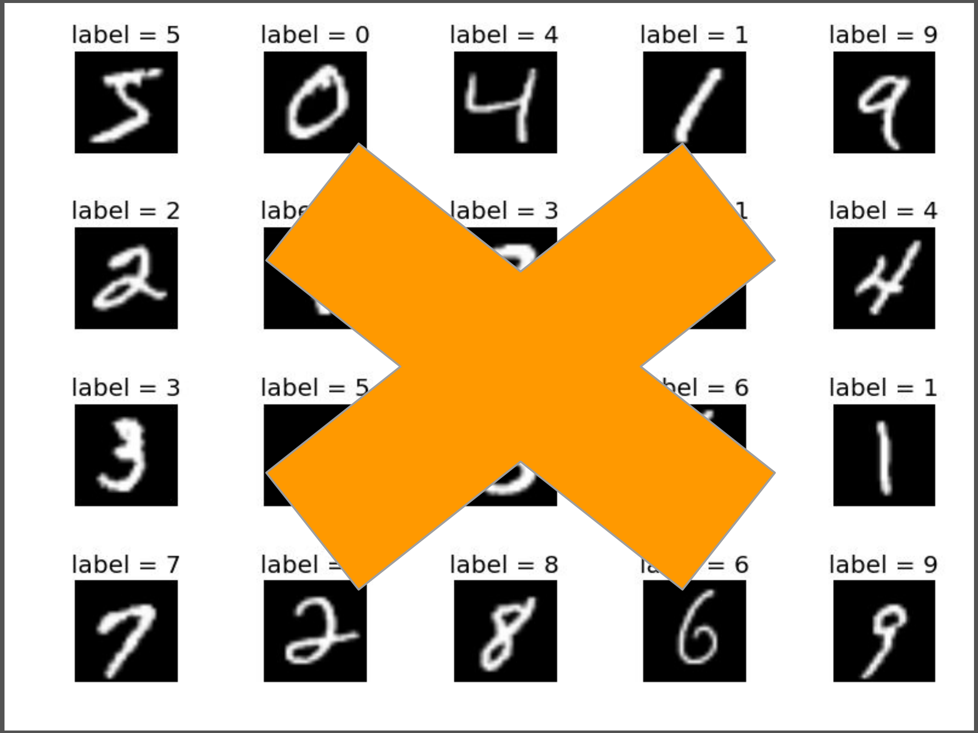
TinyData

Challenges for Building as a Startup

1. Need access to non-toy model servers

2. Need access to generated data for testing model servers

Access to Non-Toy Model Servers



TinyData

Non-Toy Model Servers: Commercial Cloud Services

aws

Azure

Google Cloud Platform

clarifai

IBM Watson™

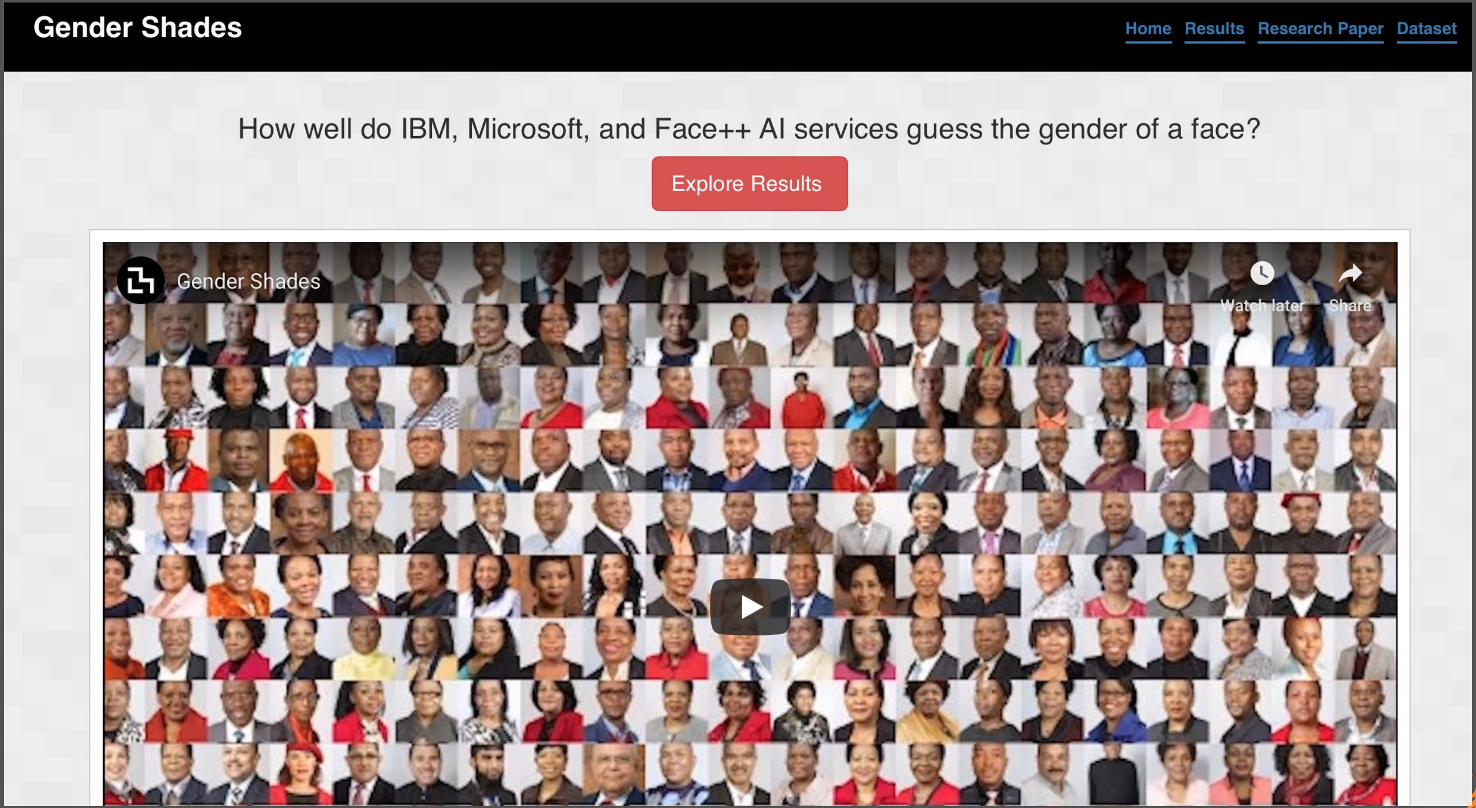
TinyData

Commercial Image Recognition Services

- Opaque systems
- Object and scene detection, facial recognition, facial analysis, NSFW detection, text detection
- Facial analysis includes gender detection

TinyData

GenderShades.org



Testing Commercial Systems for Gender Bias

- Testing = Finding cases where trained systems fail
- Hypothesis: Gender labels are trained on traditional images
- What if we generate “non-traditional” images?

TinyData

Training Data vs Test Data

Training Data

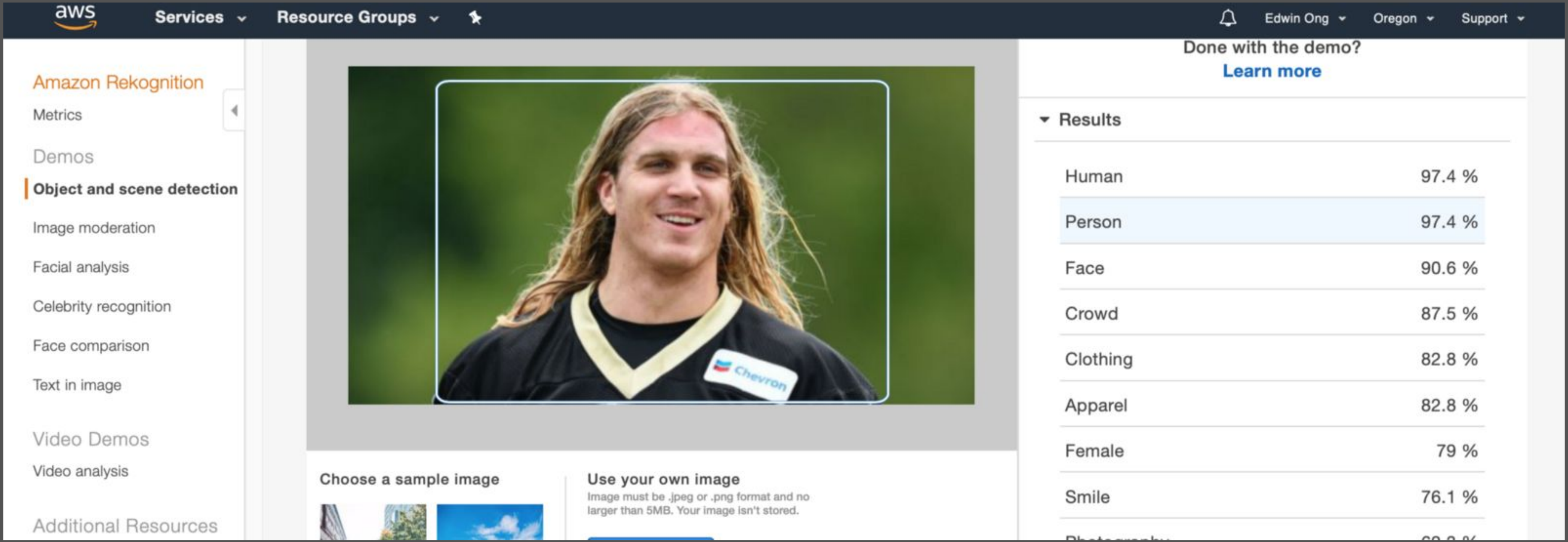
Test Data

TinyData

A Man with Long Hair

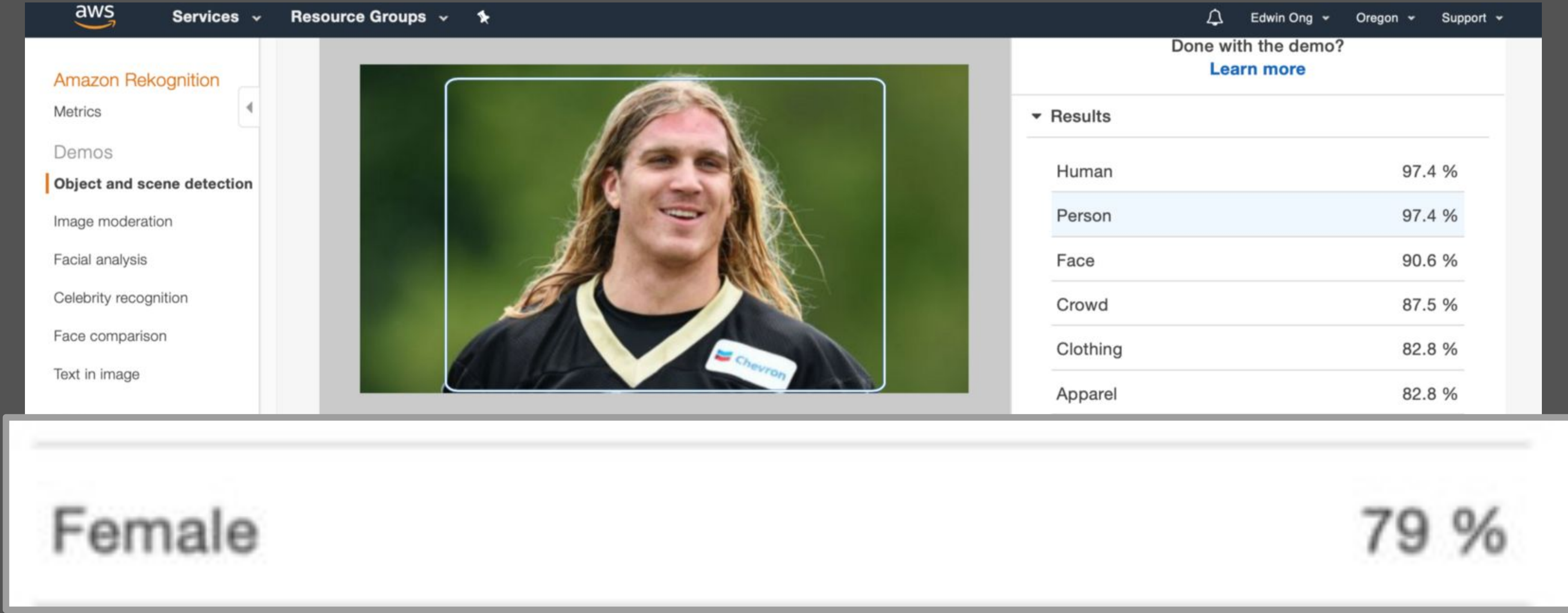
TinyData

A Man with Long Hair



TinyData

A Man with Long Hair

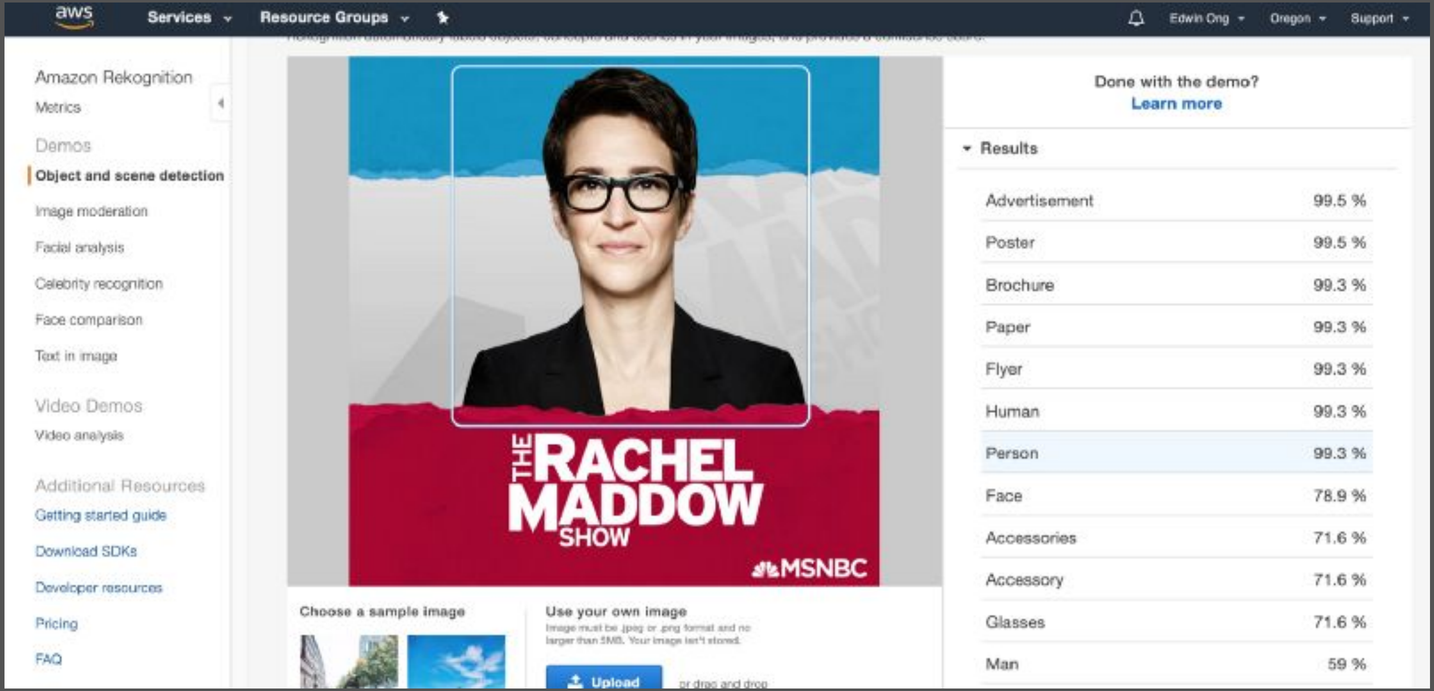


TinyData

A Woman with Short Hair

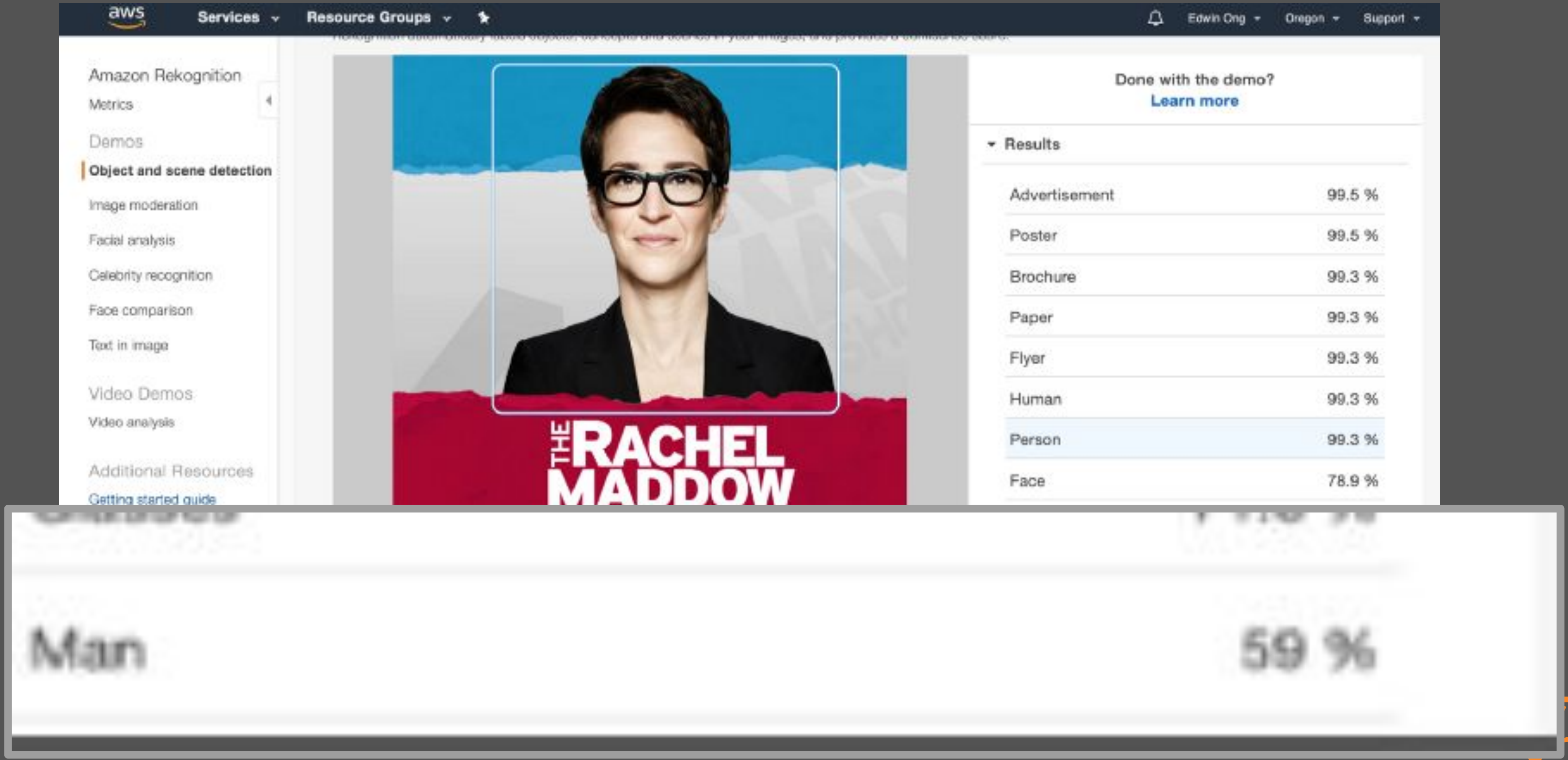
TinyData

A Woman with Short Hair

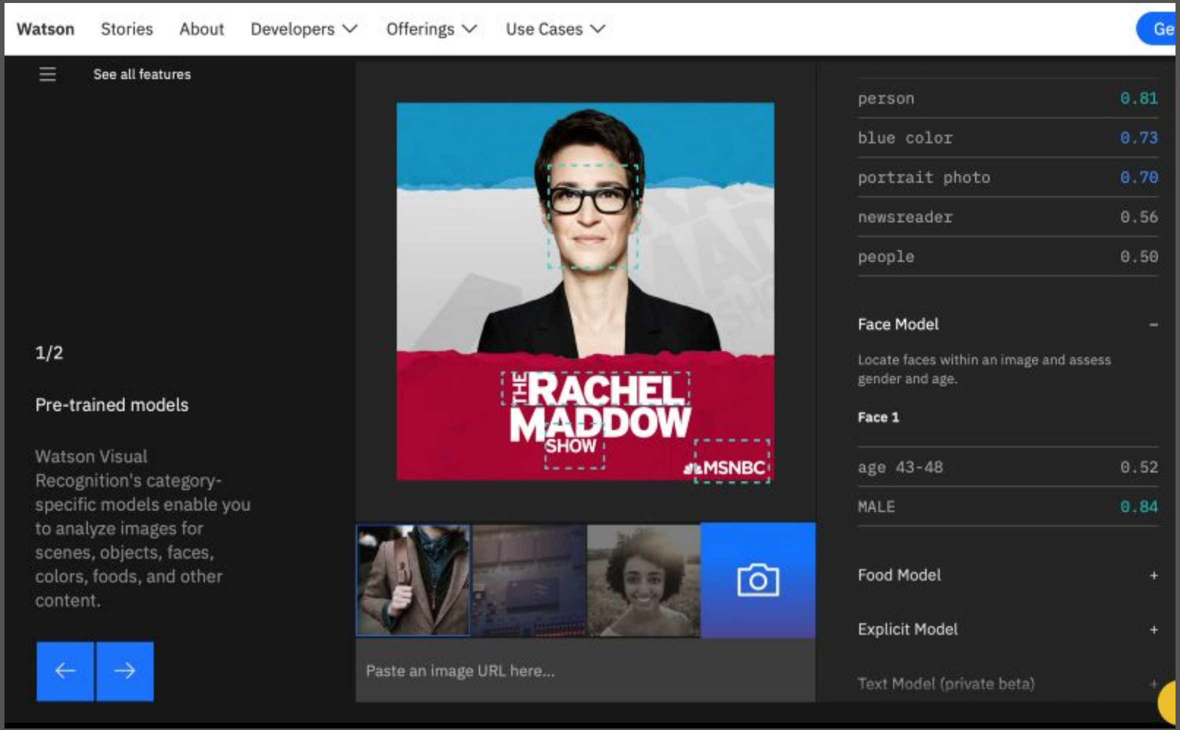


TinyData

A Woman with Short Hair

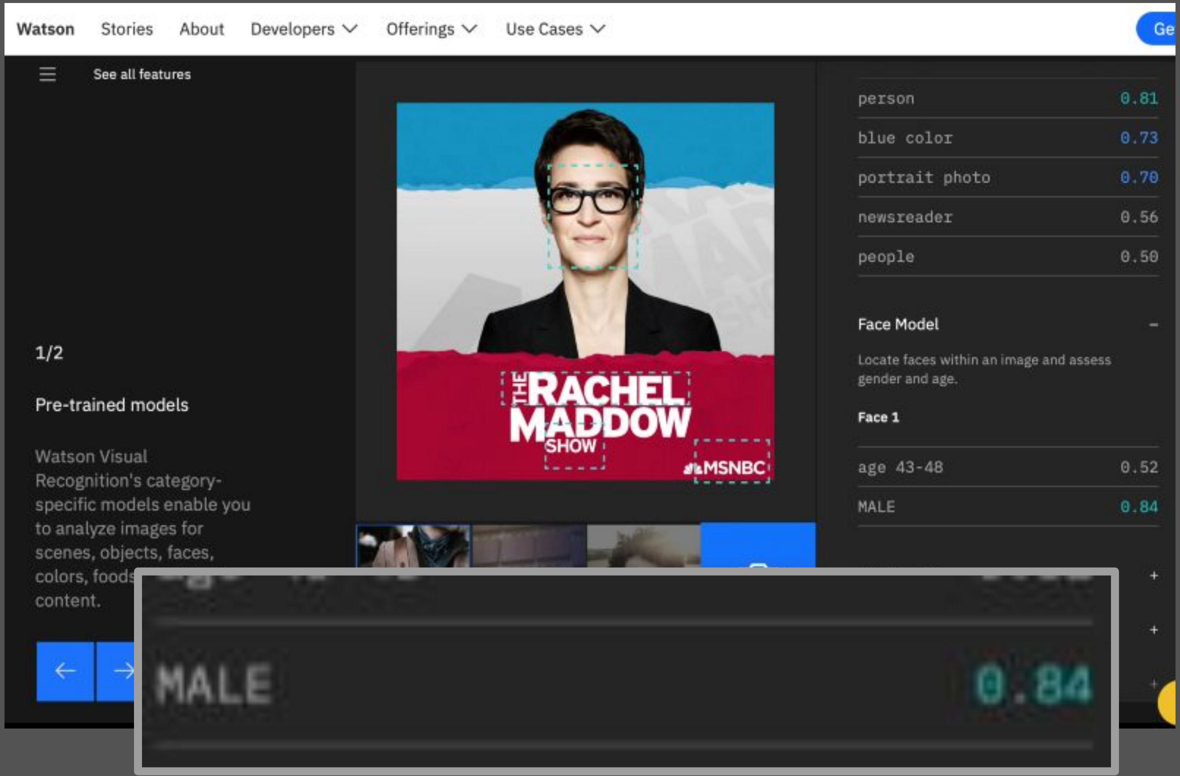


A Woman with Short Hair

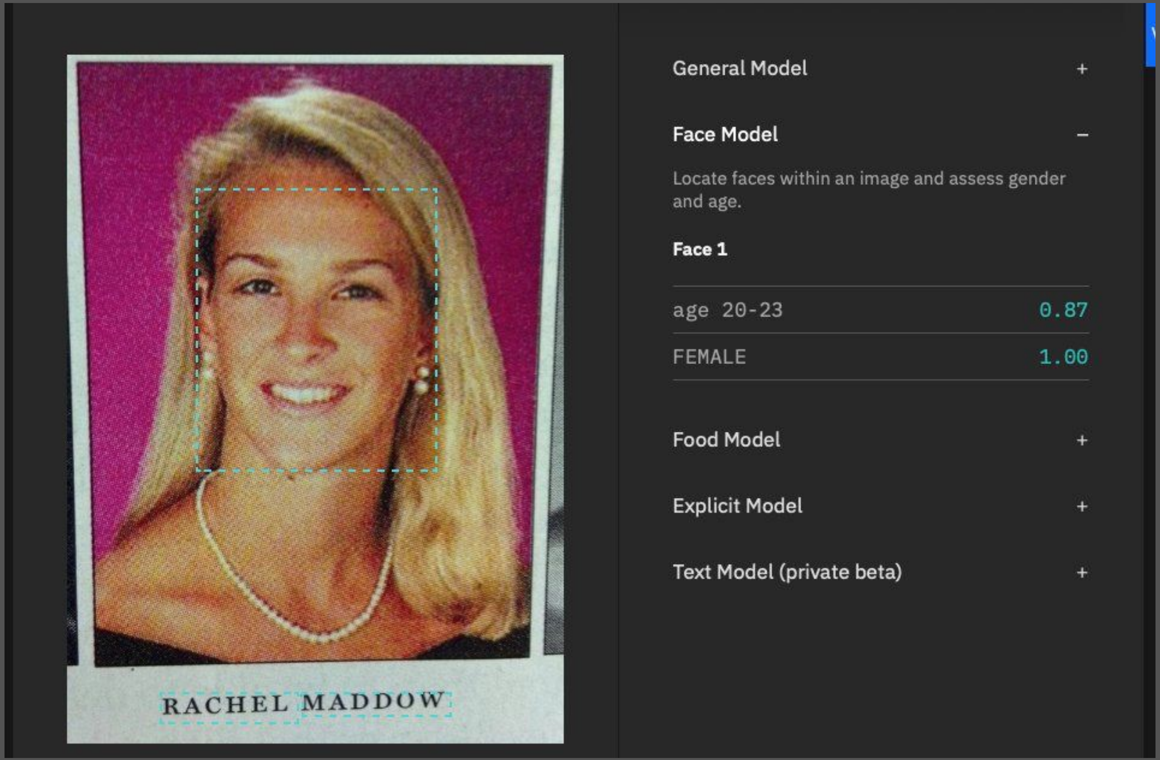


TinyData

A Woman with Short Hair

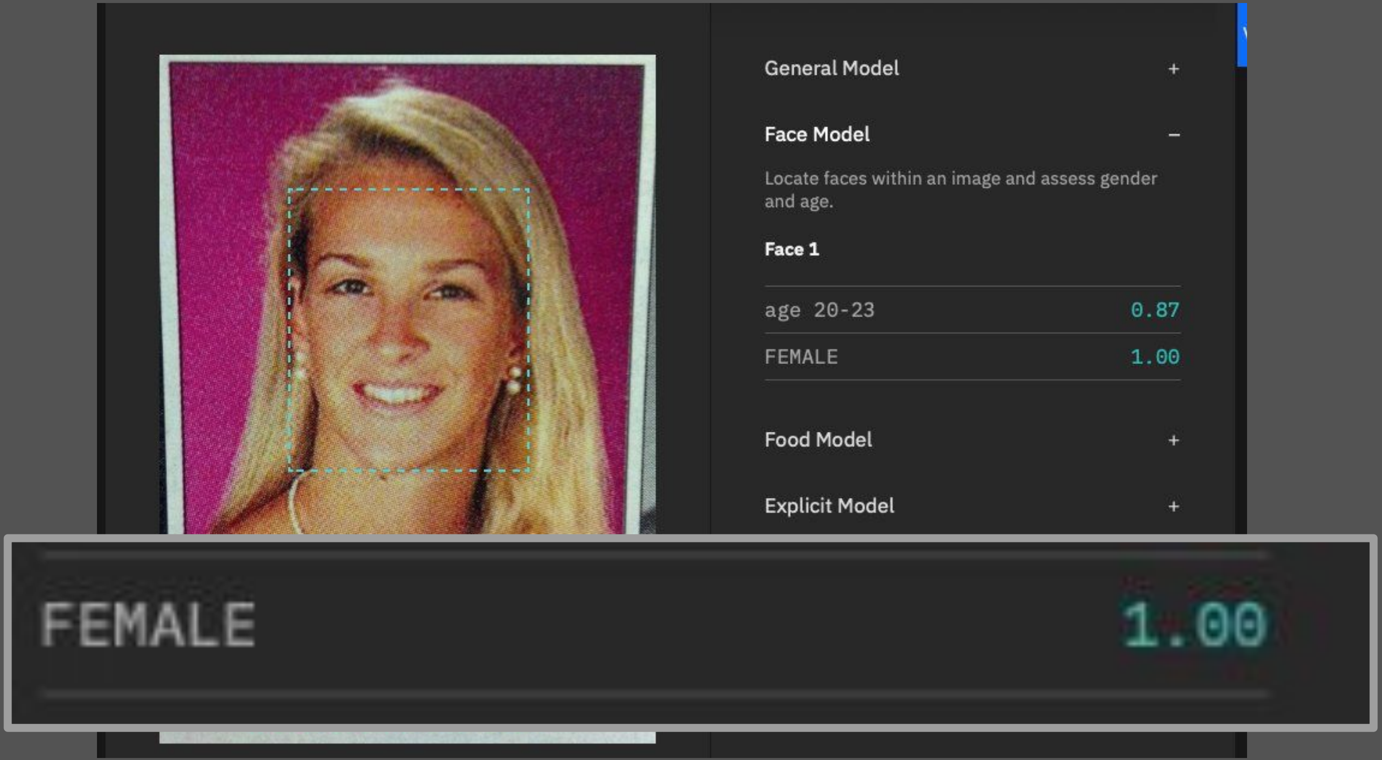


Woman with Long Hair

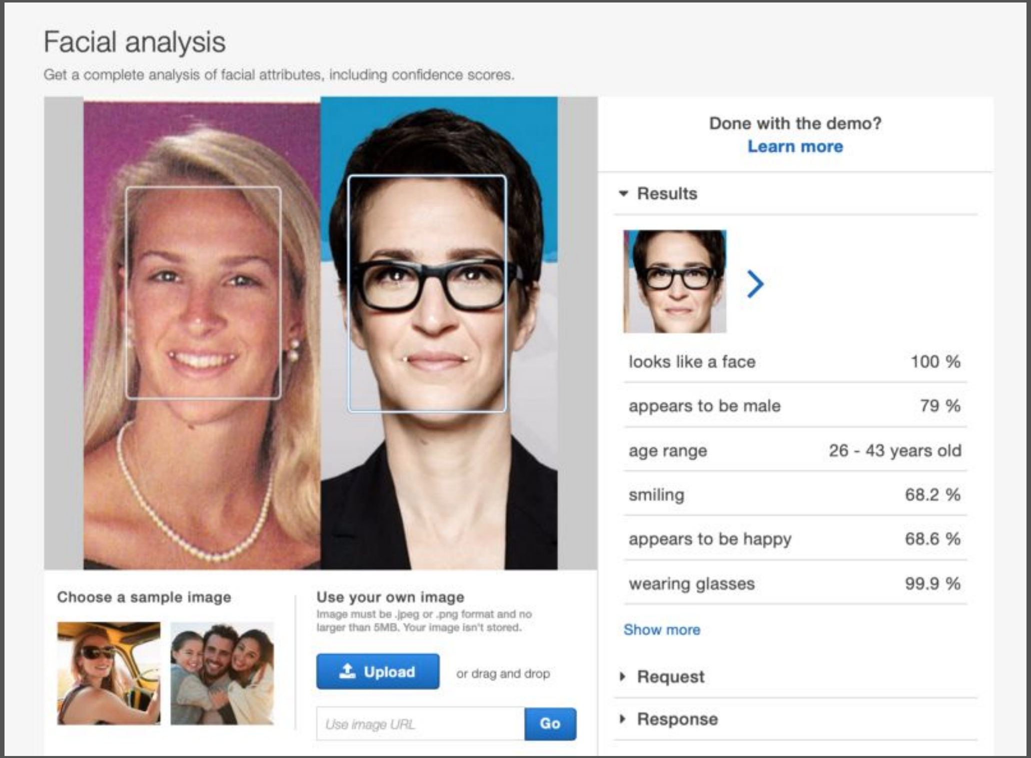


TinyData

Woman with Long Hair

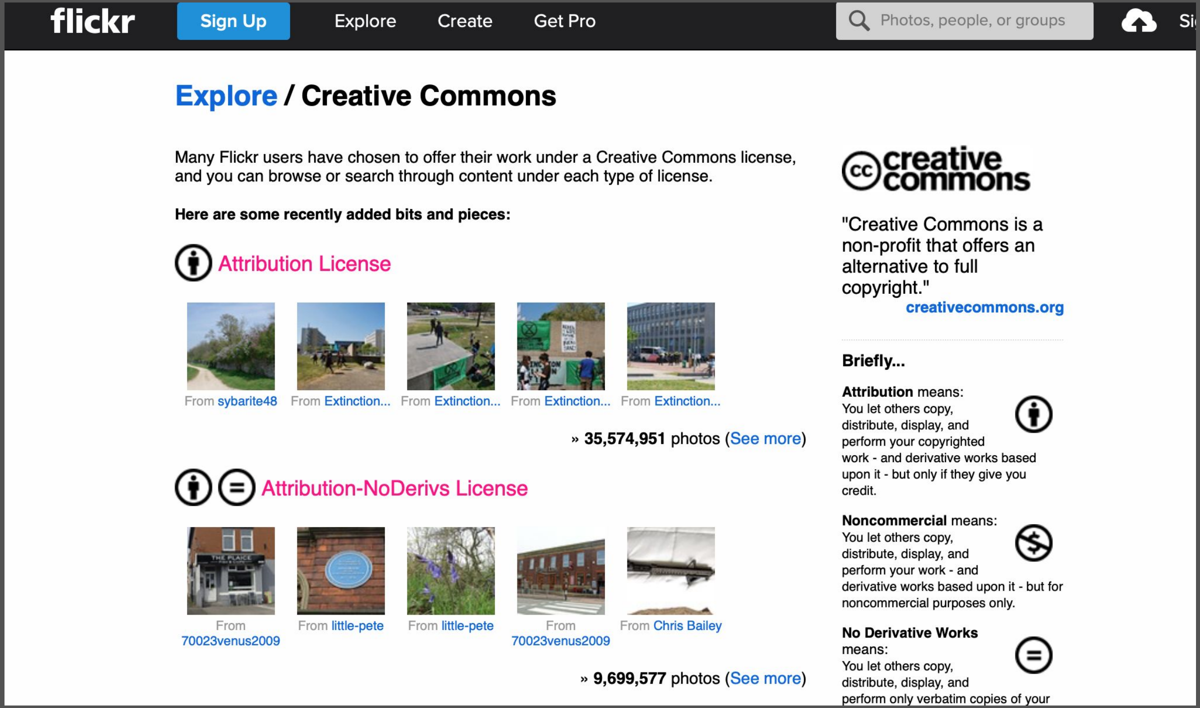


“Facial Analysis”?



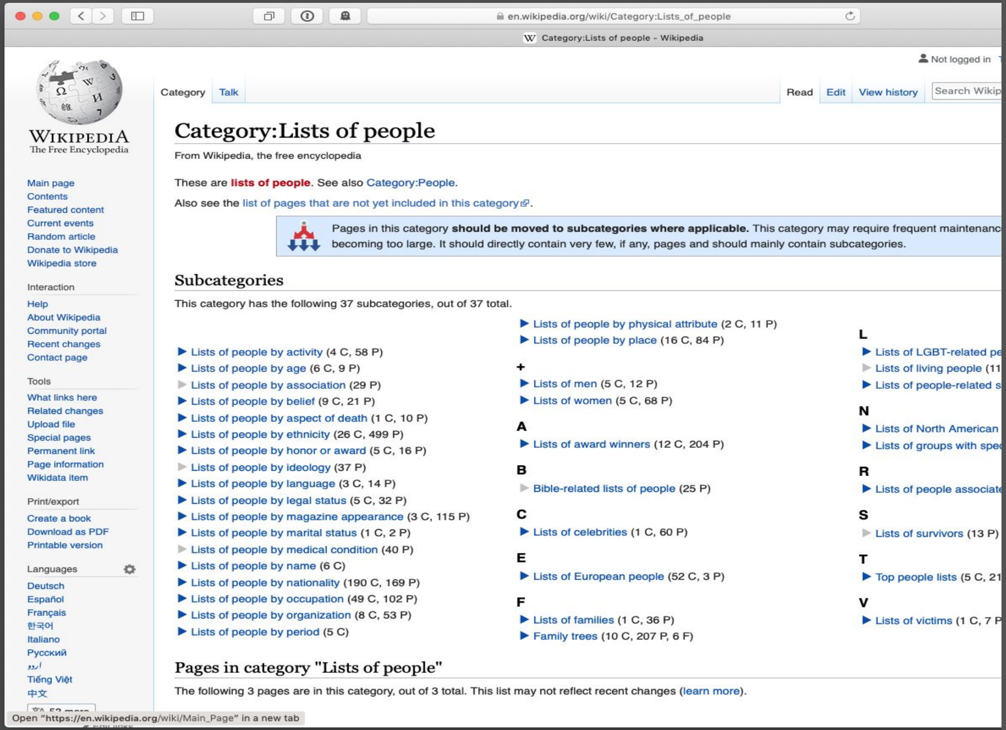
TinyData

Data Generation



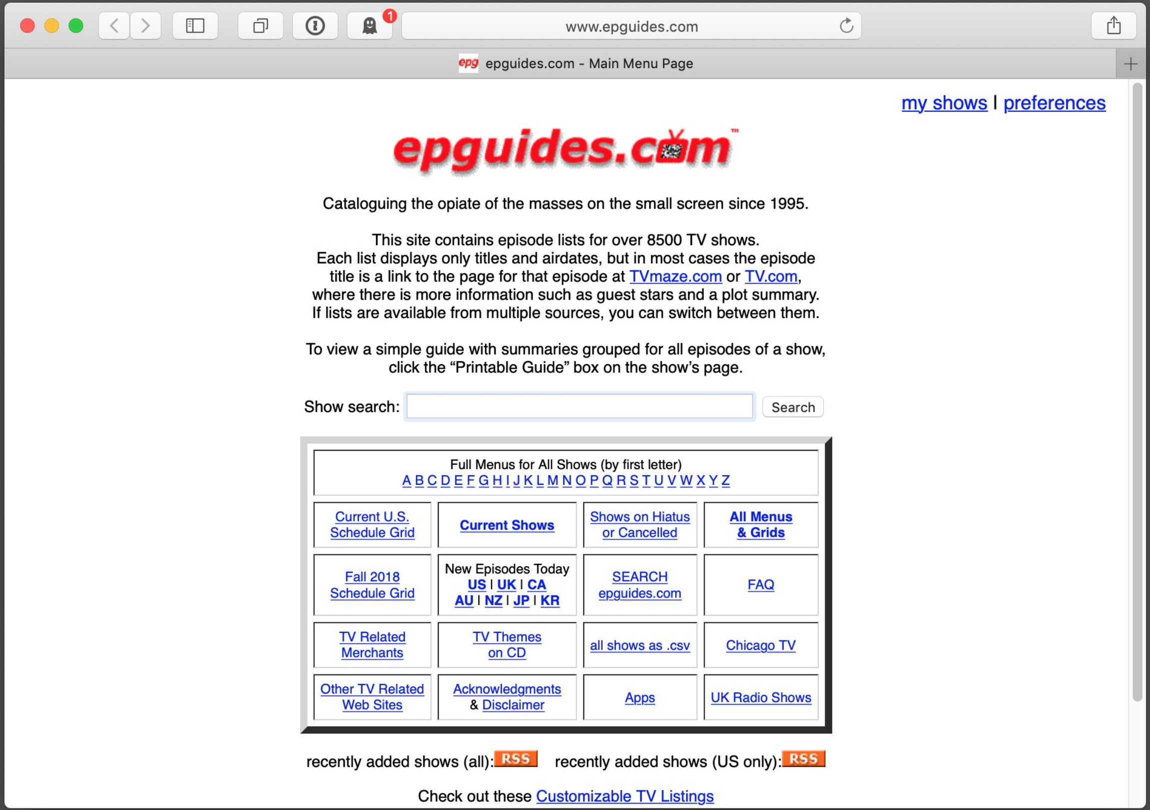
TinyData

Data Generation



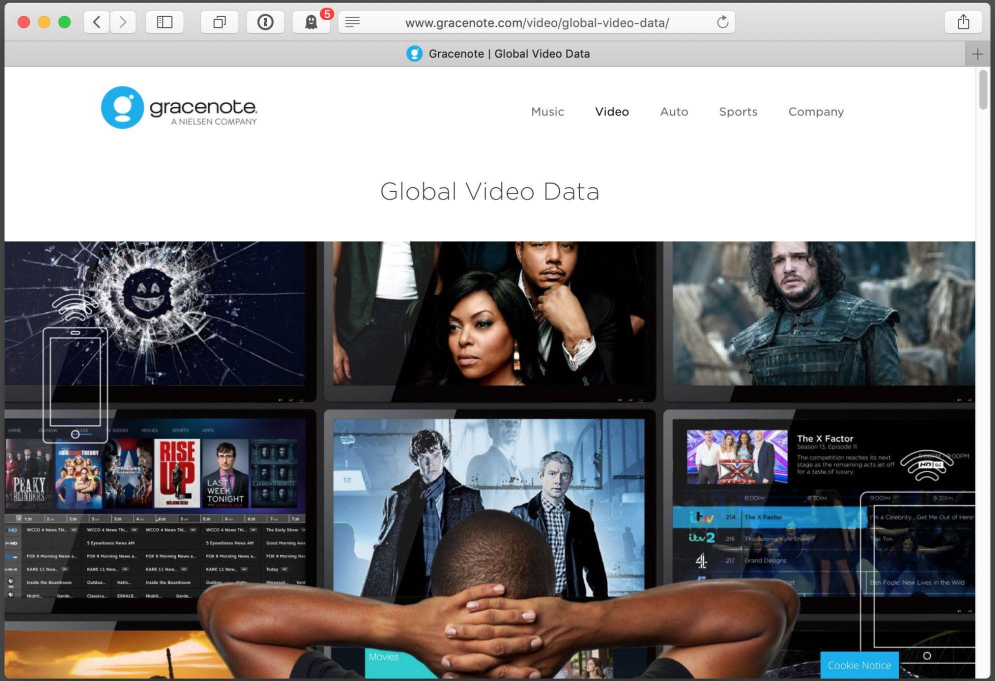
TinyData

Prototype Data



TinyData

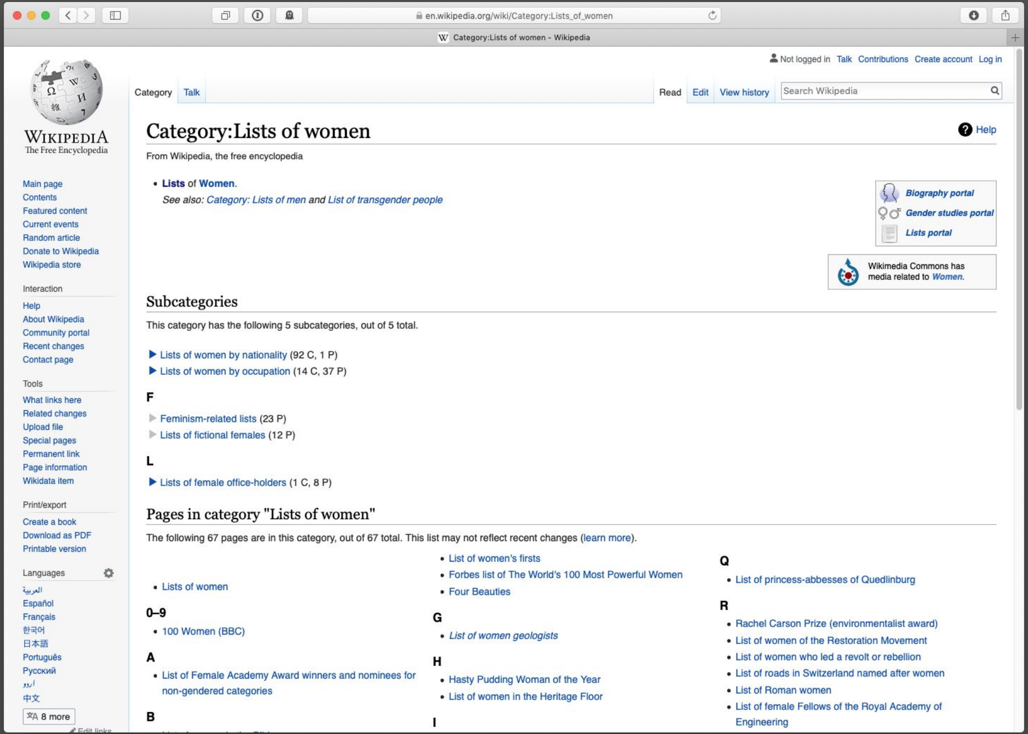
Global Standard



Data Generation

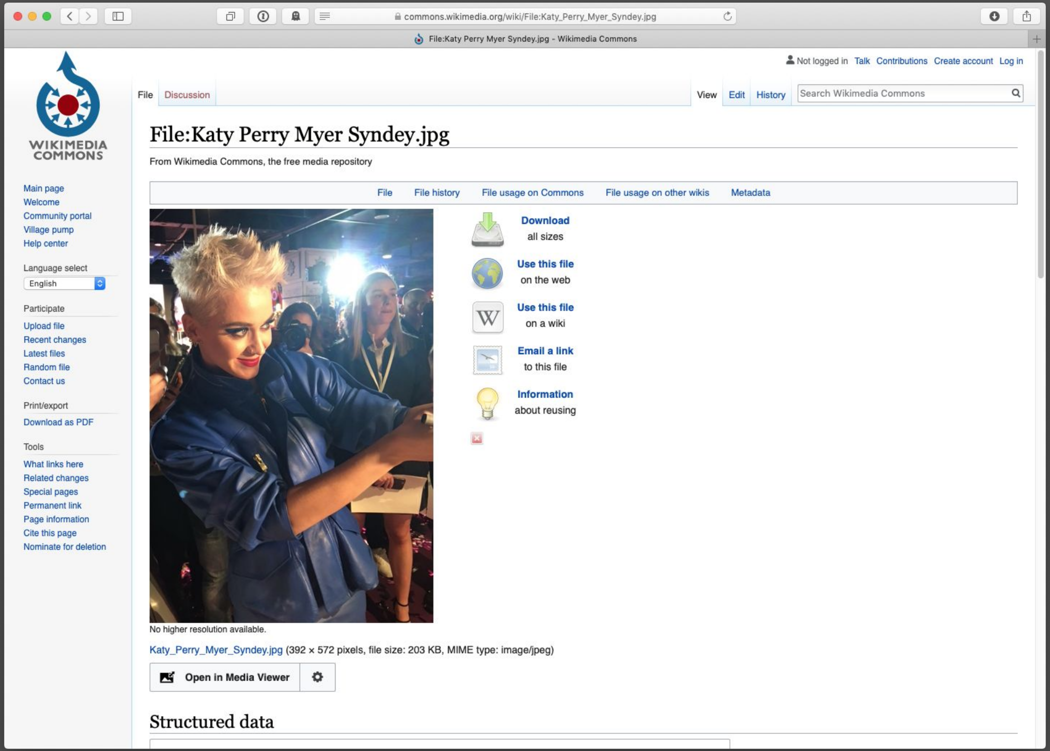
TinyData

Data Generation

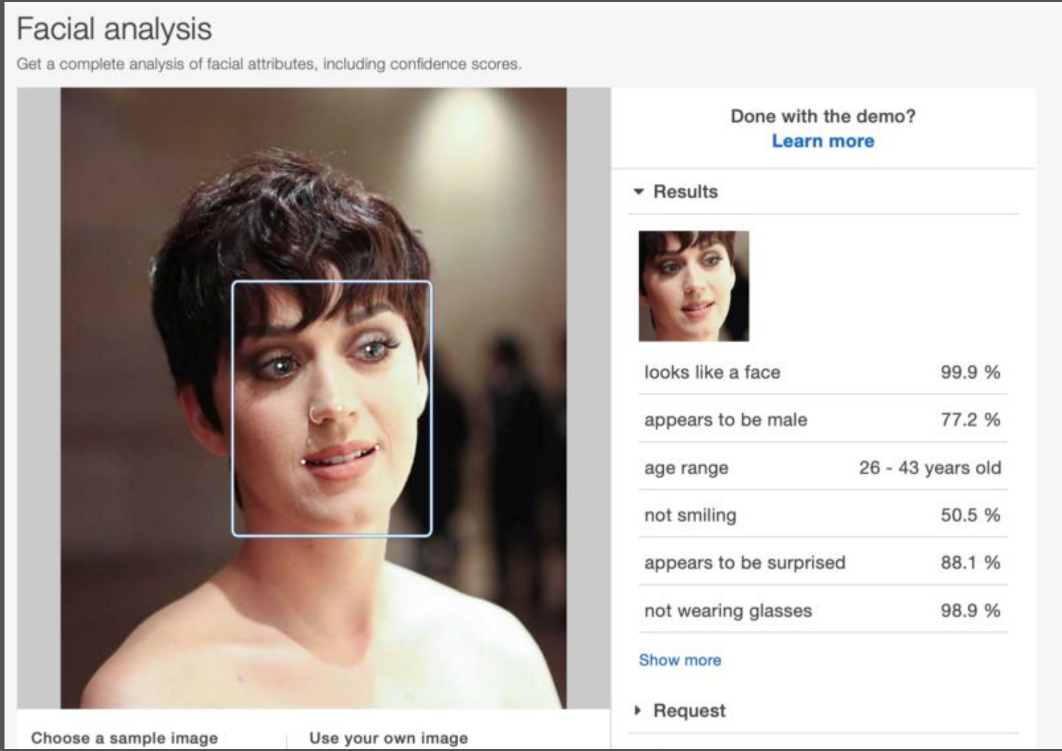


TinyData

Data Generation

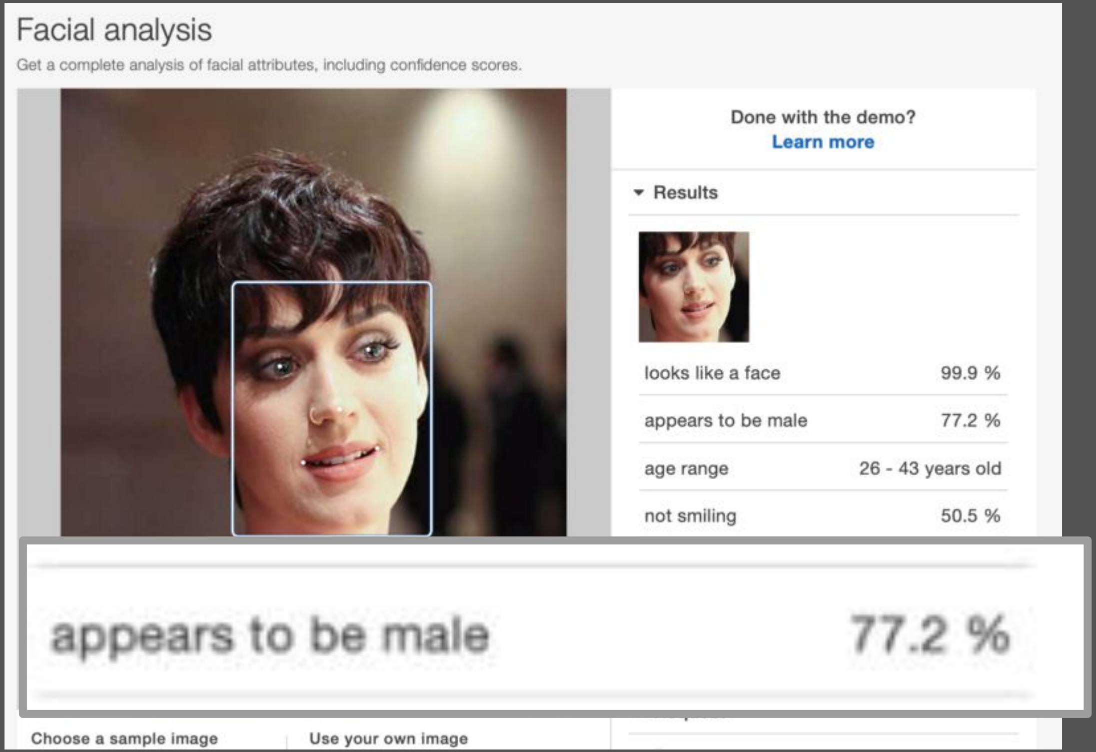


Woman with Short Hair



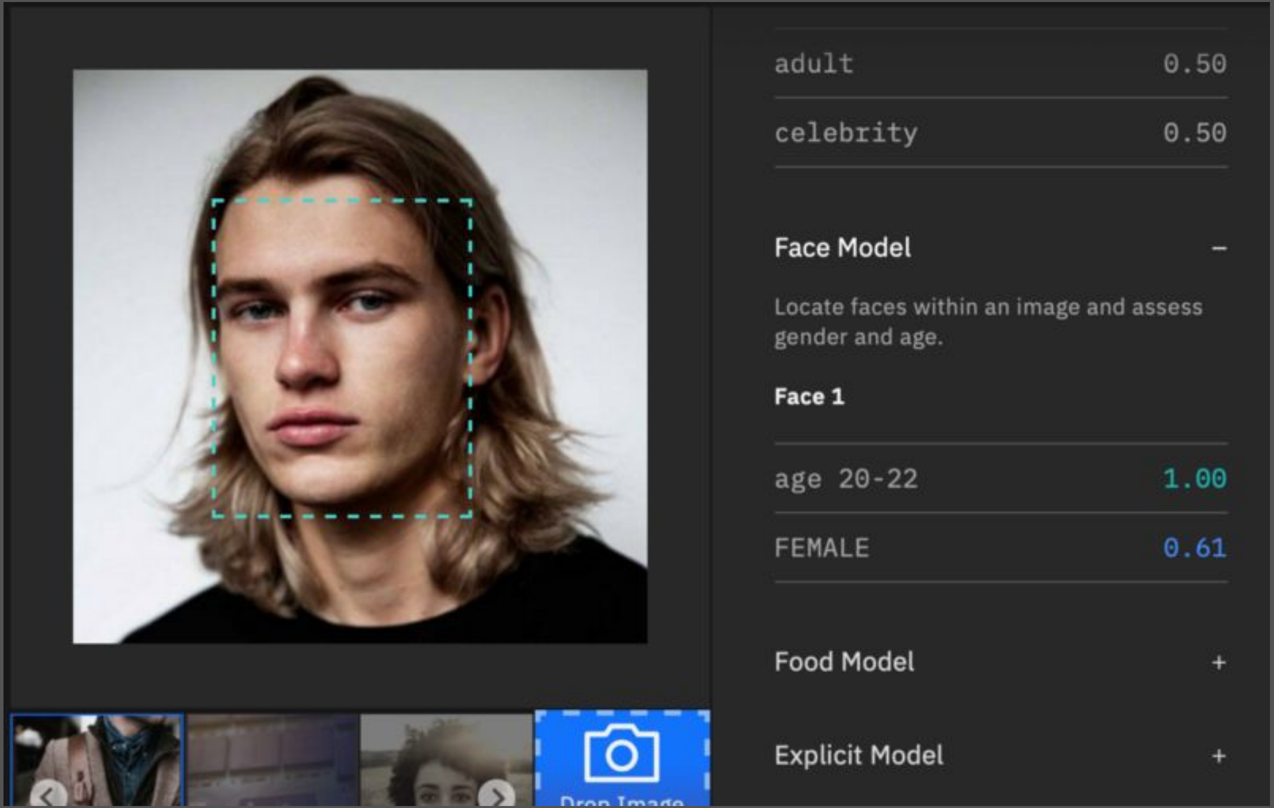
TinyData

Woman with Short Hair



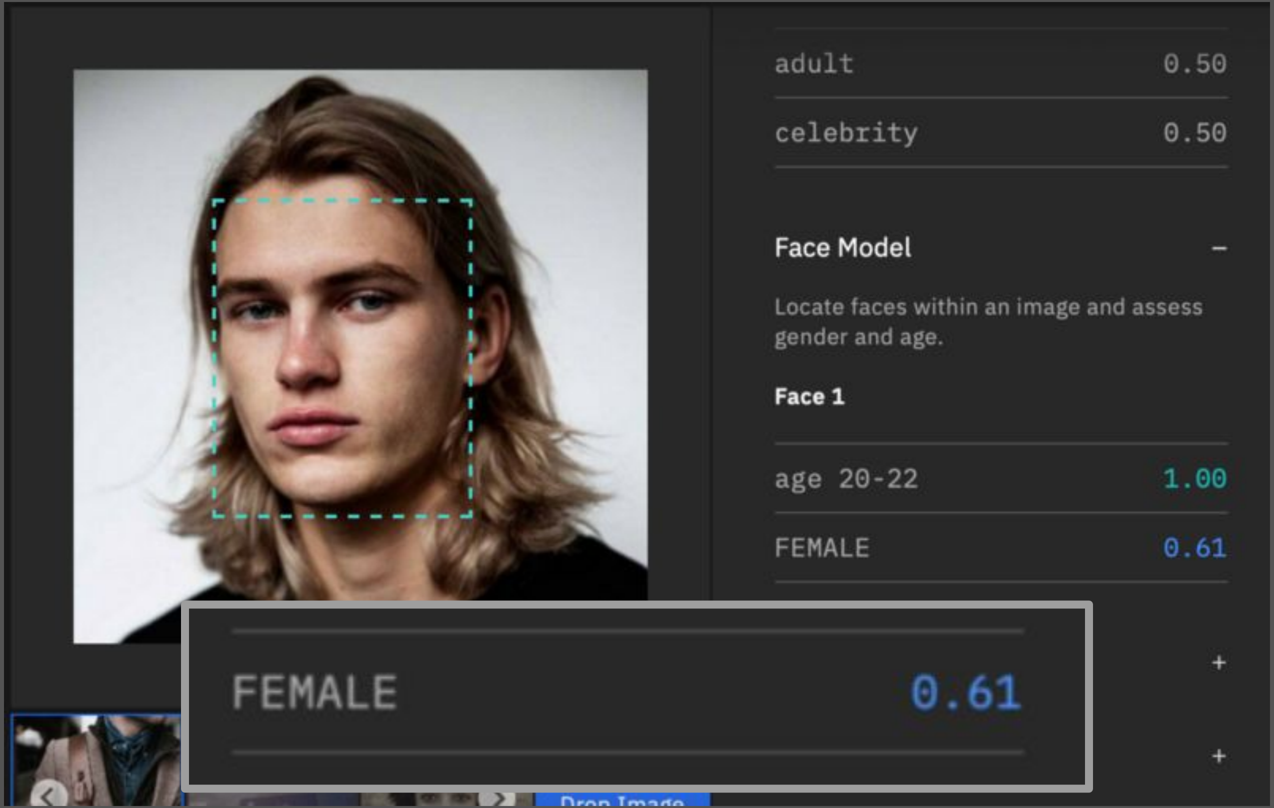
TinyData

Man with Long Hair

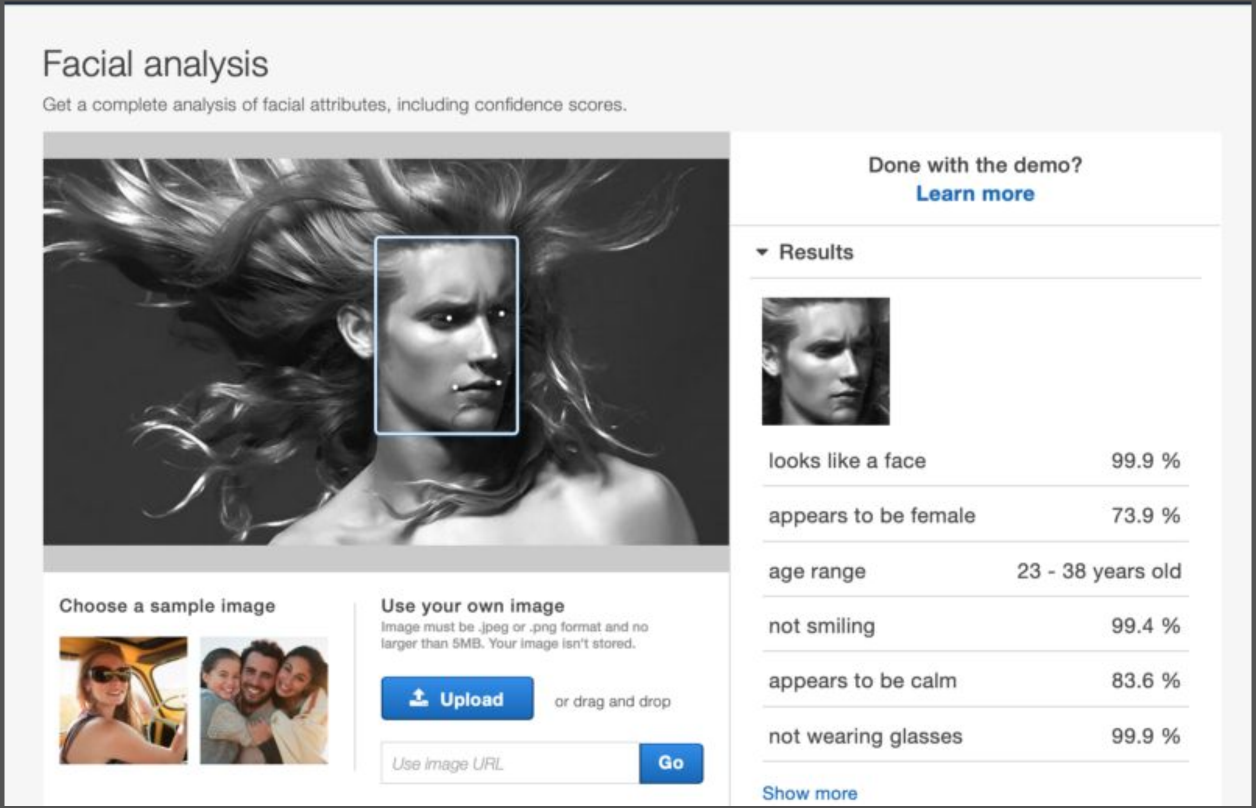


TinyData

Man with Long Hair

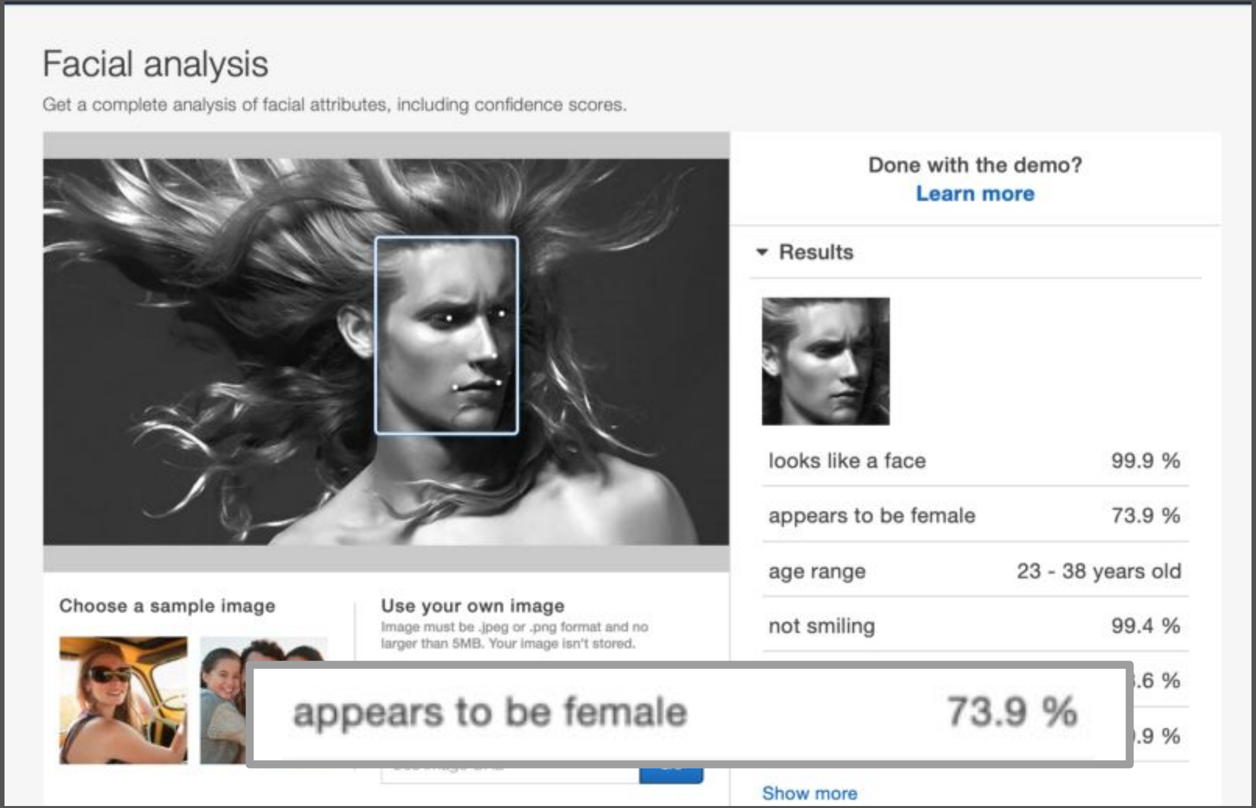


Man with Long Hair



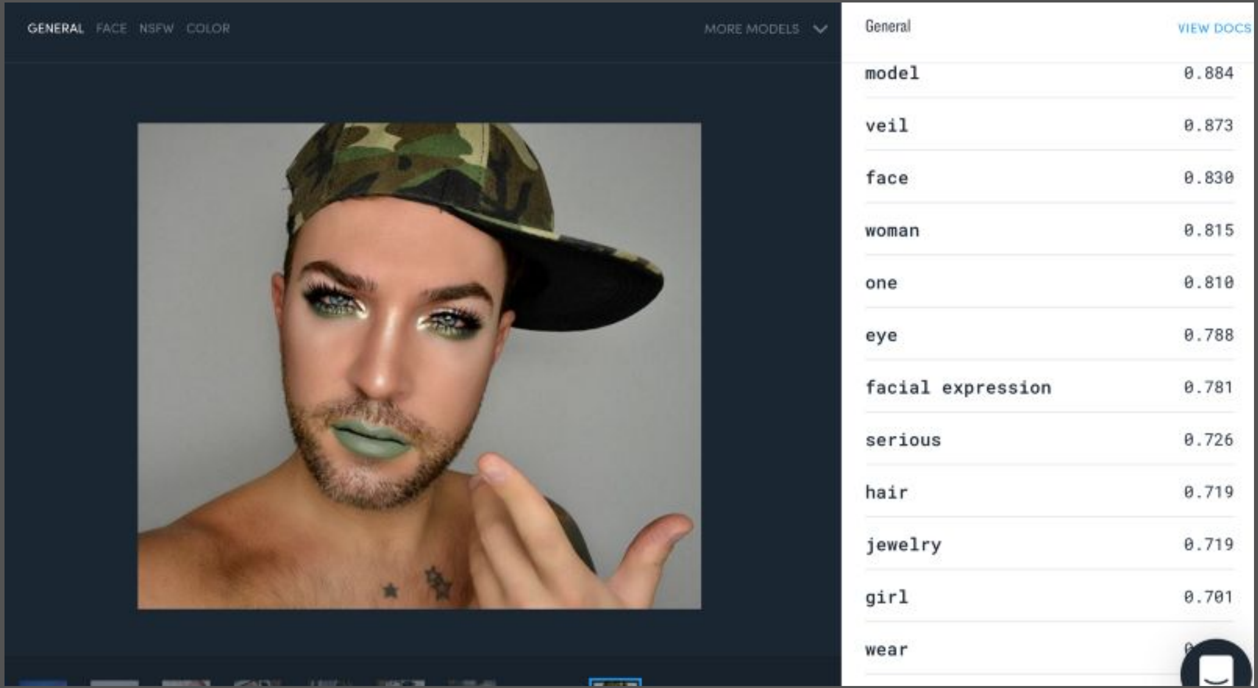
TinyData

Man with Long Hair



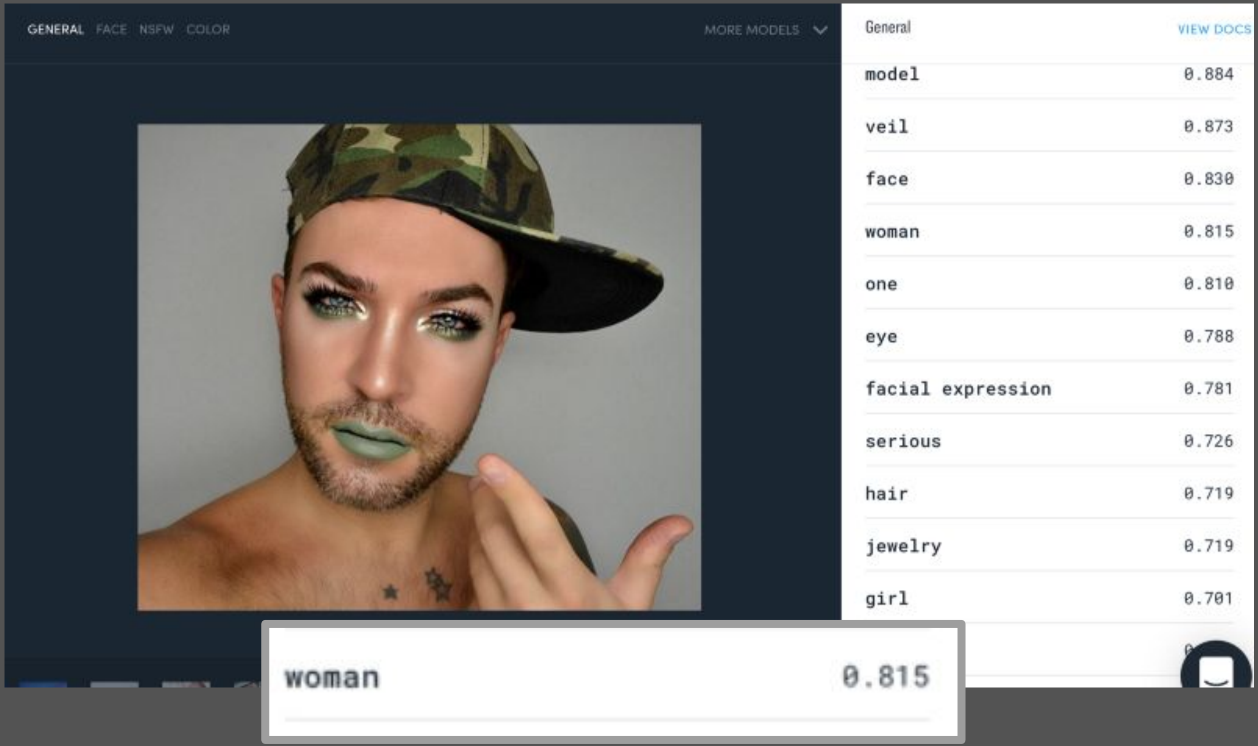
TinyData

Man with Makeup



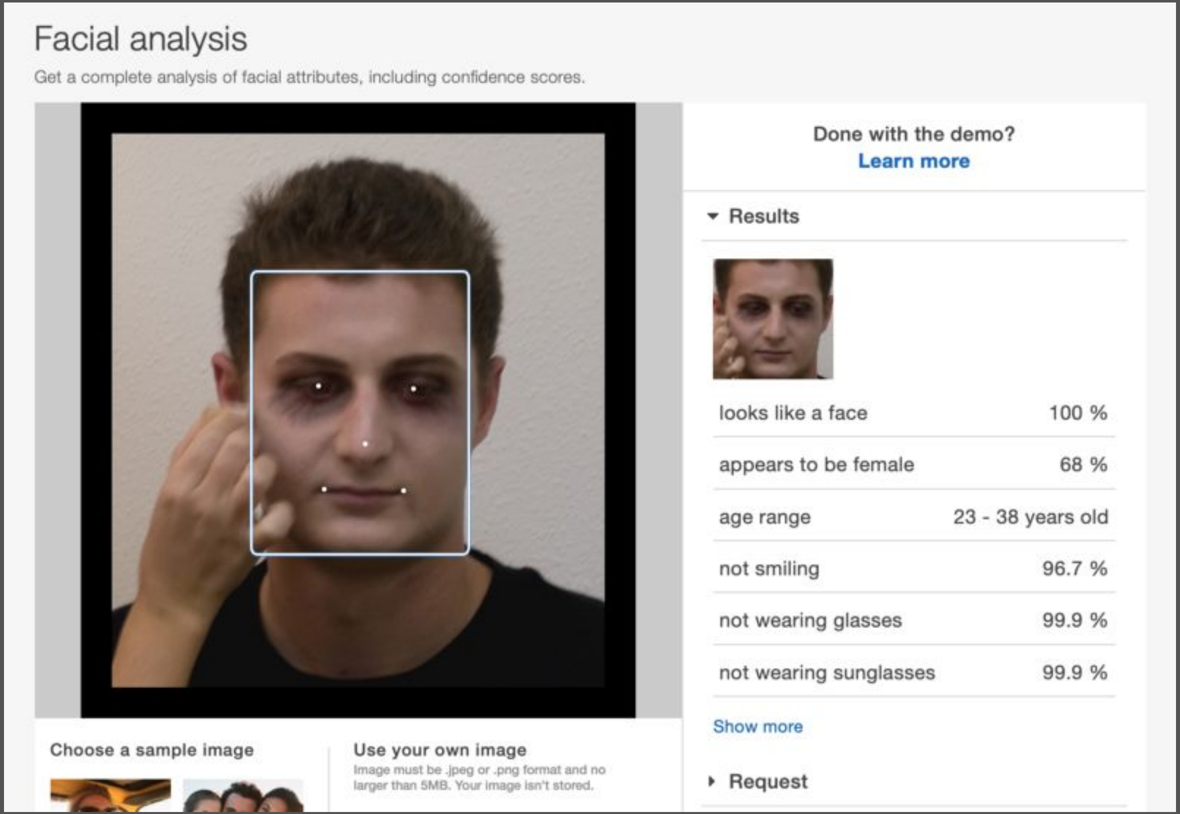
TinyData

Man with Makeup



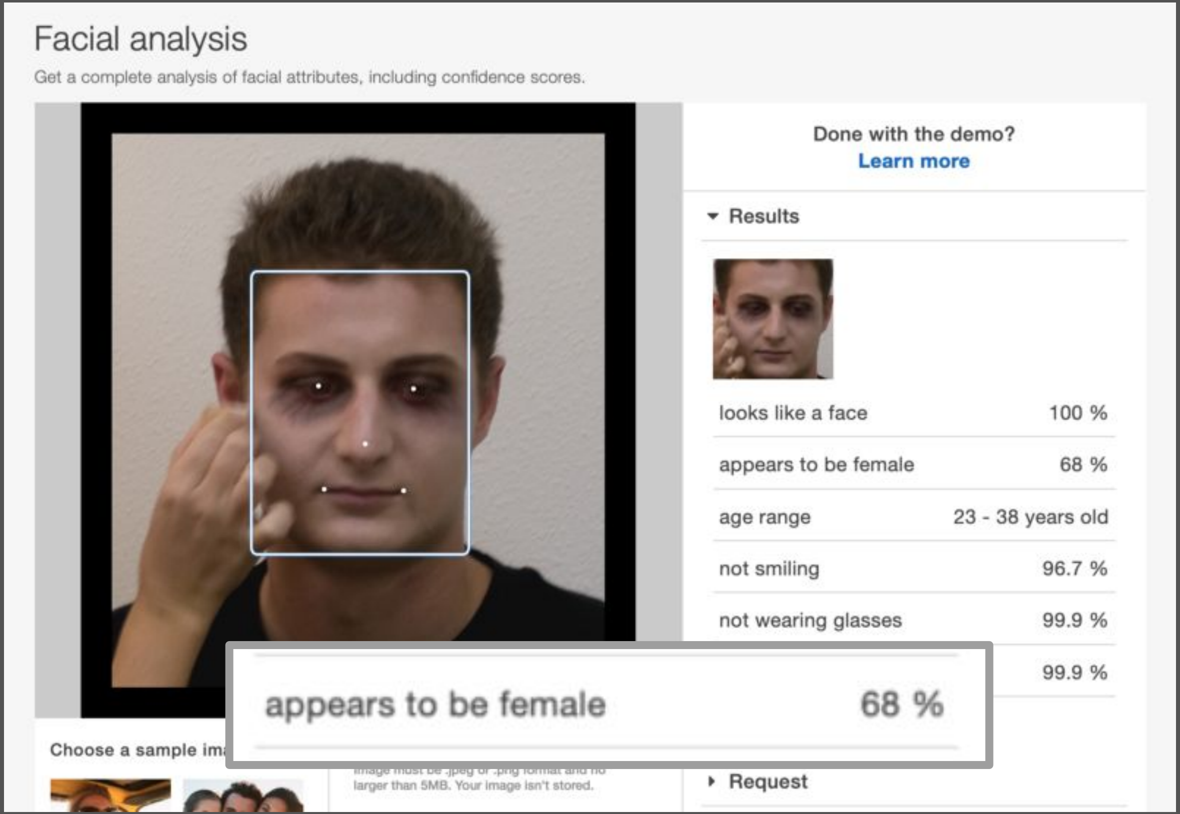
TinyData

Man with Makeup



TinyData

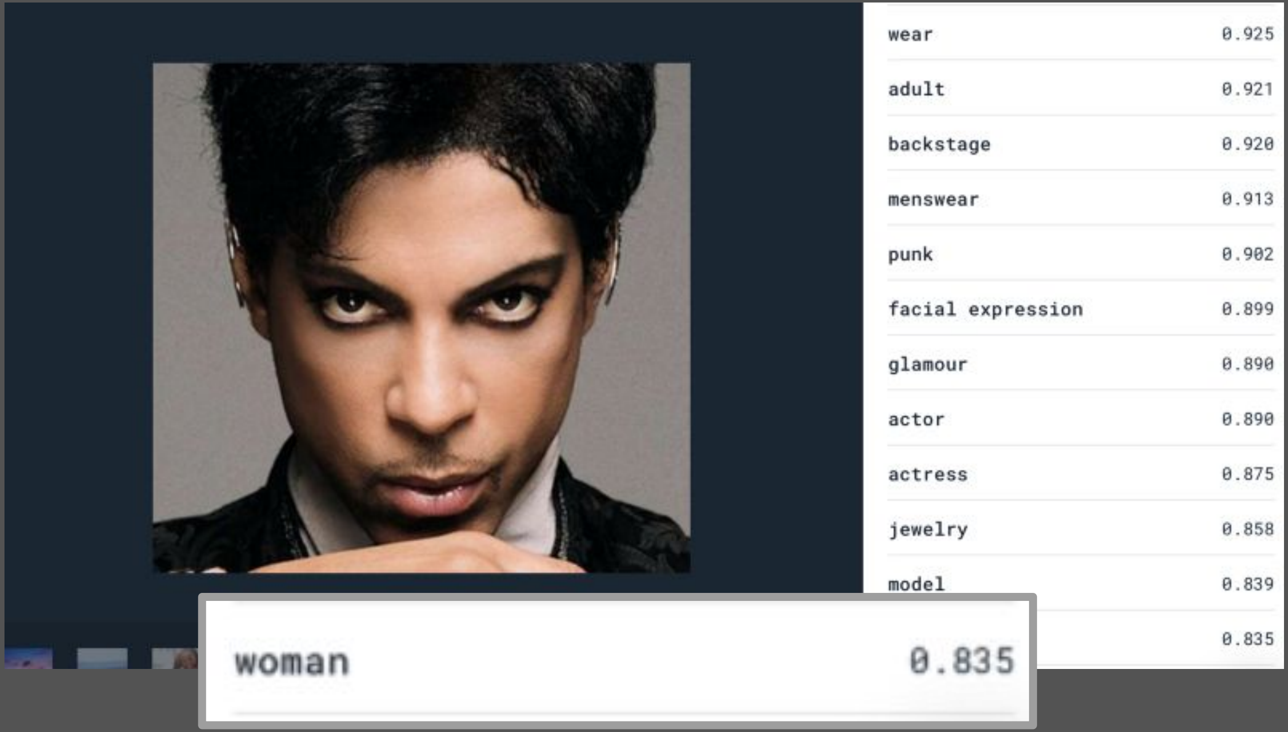
Man with Makeup



Man with Makeup

wear	0.925
adult	0.921
backstage	0.920
menswear	0.913
punk	0.902
facial expression	0.899
glamour	0.890
actor	0.890
actress	0.875
jewelry	0.858
model	0.839
woman	0.835

Man with Makeup



TinyData

Automating Data Generation + Testing

TinyData Home Projects Datasets APIs Search Search

Projects / Gender Classification Tests

Label		aws	Azure	clarifai	IBM Watson		
female		male	female	male	male	Edit	Delete
female		male	female	female	male	Edit	Delete
female		female	female	female	female	Edit	Delete
female		male	female	male	male	Edit	Delete
female		male	female	female	male	Edit	Delete
female		male	male	male	female	Edit	Delete
female		male	male	male	male	Edit	Delete

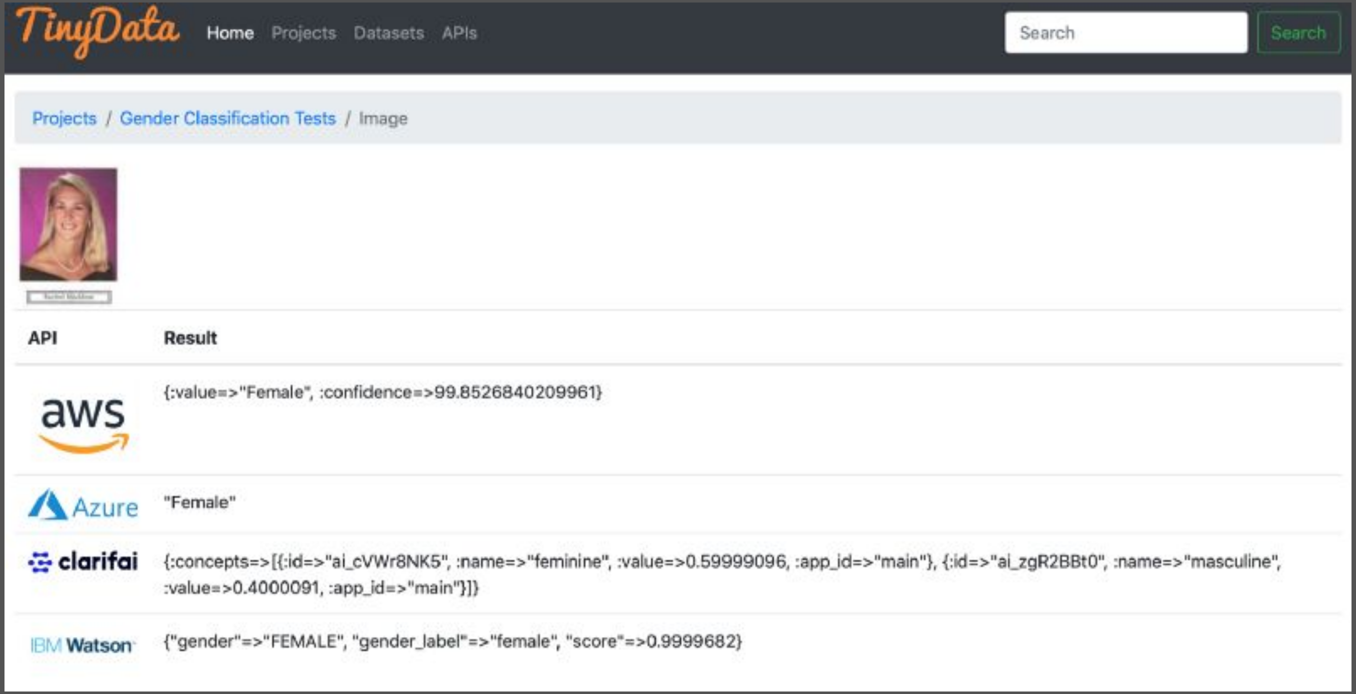
TinyData

Automating Data Generation + Testing

female		female	female	female	female	Edit	Delete
female		male	female	male	female	Edit	Delete
female		female	female	male	female	Edit	Delete
female		male	female	female	male	Edit	Delete
female		male	female	female	female	Edit	Delete
female		female	male	male	male	Edit	Delete
female		male	male	male	female	Edit	Delete
female		female	female	female	female	Edit	Delete
female		female		female	female	Edit	Delete
female		female		female	female	Edit	Delete

TinyData

Tracking Results Over Time



TinyData

Takeaways

- Even the best trained commercial ML systems are far from perfect

- Systems return different results over time as new versions get deployed

- Cumbersome & intractable to test without tools & automation

TinyData

Scaling Data Products as a Startup

- Bootstrap servers with commercial APIs
- Bootstrap data with open web, public & synthetic datasets
- Automation is startups' best friend

TinyData

Questions / Comments

edwin@tinydata.co

Twitter: @edwin

TinyData