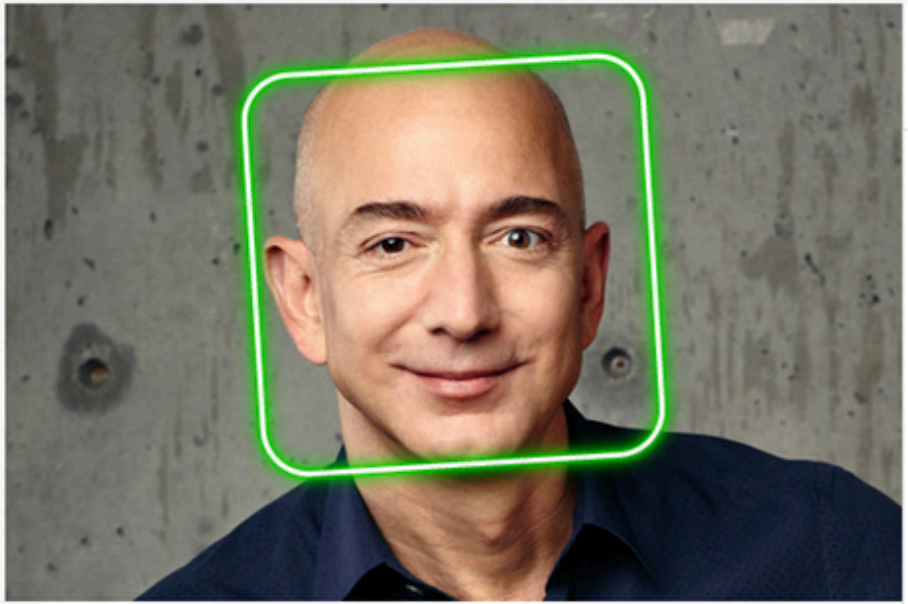
Adversarial data & Invariants

Celebrity recognition

Rekognition automatically recognizes celebrities in images and provides confidence scores (Your images aren't stored.)



Done with the demo?
Download SDKs

▼ Results

Jeff Bezos
Learn More

Match confidence 100%

▶ Request

▶ Response

Choose a sample Image

Use your own image

Upload or drag and drop

Use image URL Go

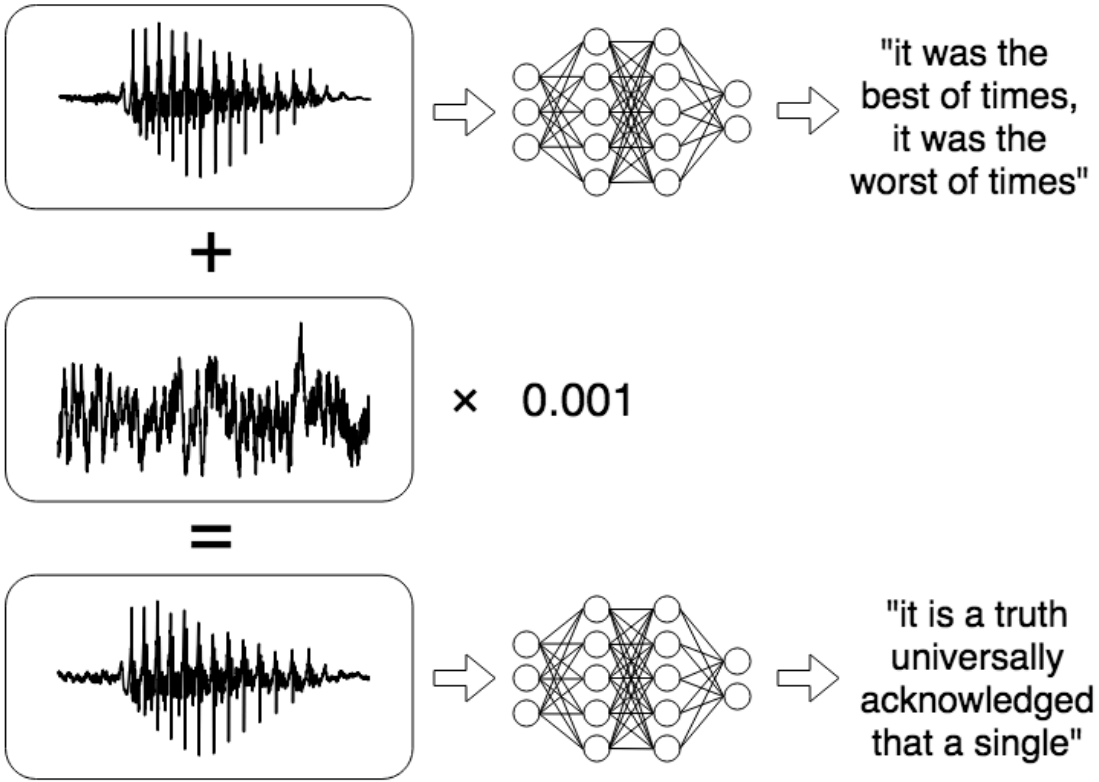
AWS Rekognition

Adversarial Image Generation (e.g. Sharif et al. 2017)

Digital manipulation
to dodge recognition

In real life - via 3D
printed glasses

Adversarial Audio Generation (e.g. Carlini & Wagner, 2018)



- Modify data slightly such as to obtain wrong class

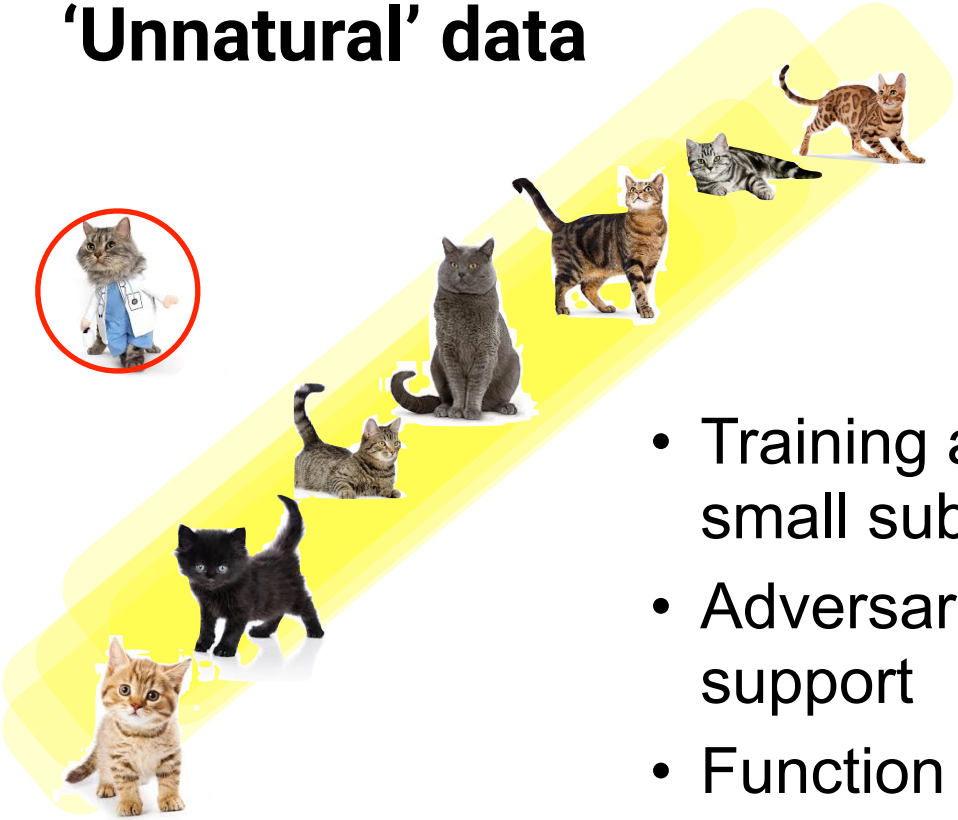
$$\underset{\delta}{\text{maximize}} \; l(f(x+\delta), y)$$

$$\text{subject to } \|\delta\| \leq \epsilon$$

Different norms
Different datasets
Different papers …

Why does this work?

‘Unnatural’ data



- Training and ‘natural’ test data live in small subset
- Adversarial data is slightly off that support
- Function behavior undefined away from where data occurs

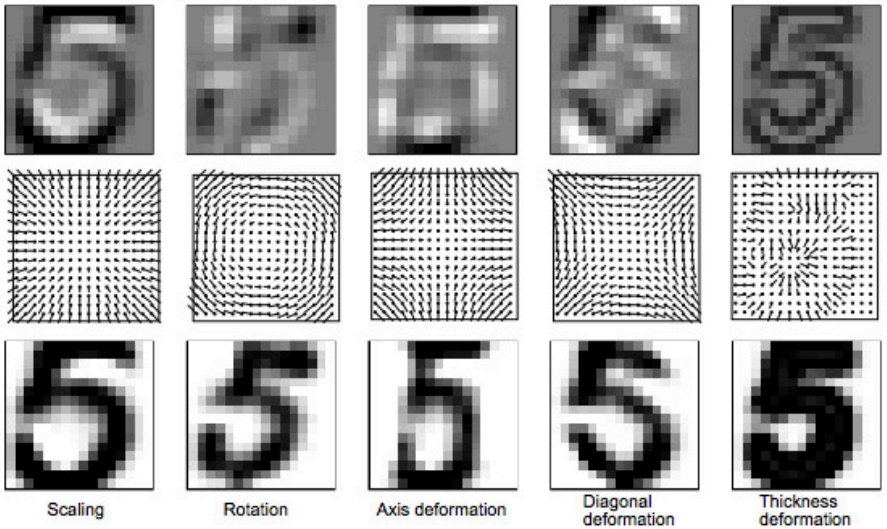
Wow. Breathtaking. Is this new?

Spam defenses

- While TRUE
  - Mail host extends dataset and trains new classifier
  - Spammer's e-mails are rejected
  - Spammer finds a modification that succeeds
- Examples
  - Add highly scoring words (or sentences) to email
  - Add highly scoring sentences (and vary them)
  - Change or forge header ('Dear Alex, …')

Invariances

- Tangent Distance (Simard et al., 1995)
  - Invariance transforms don't change the label
  - Explore data and their neighborhood



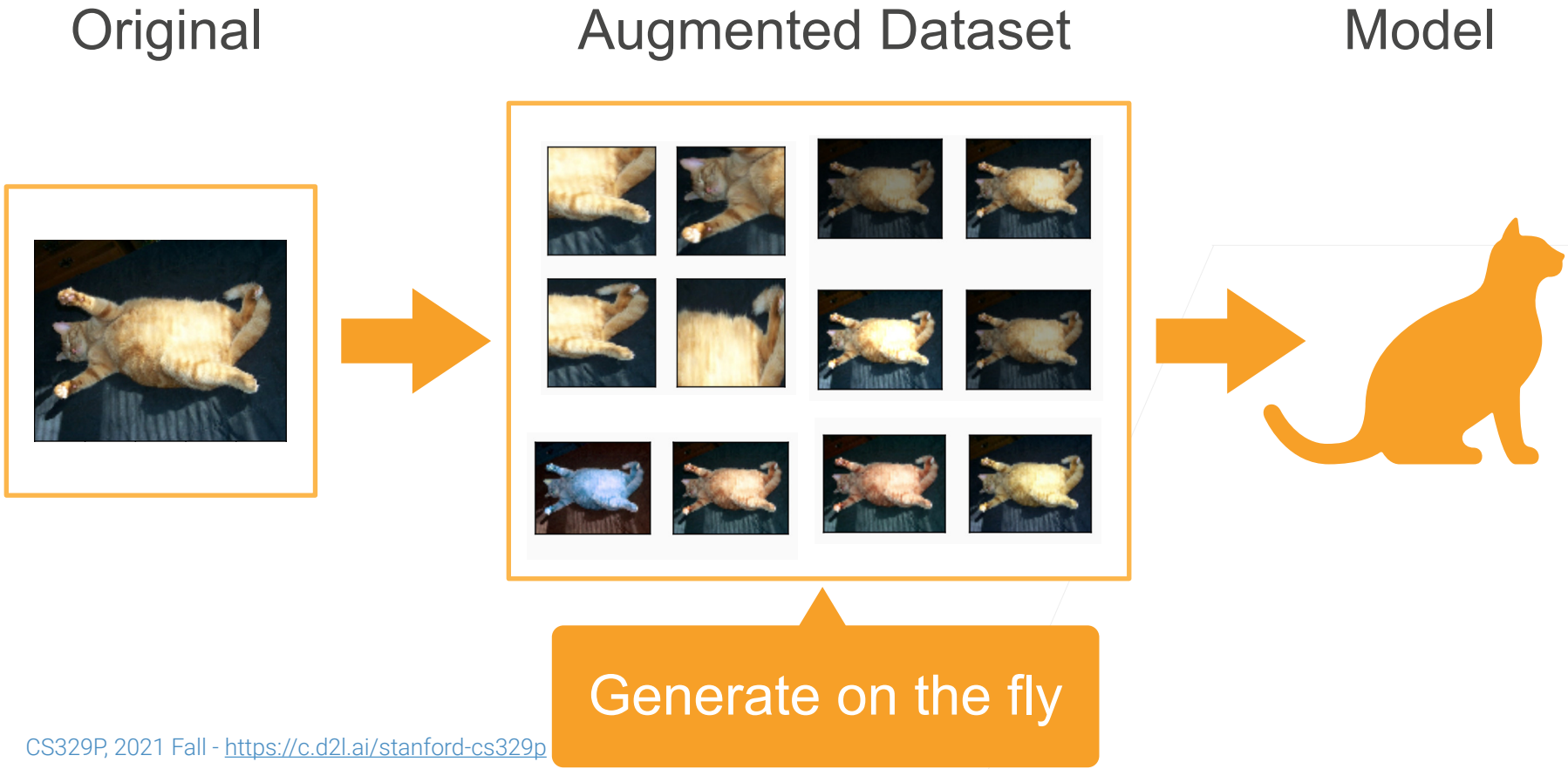
Invariances

- **Virtual Support Vectors** (Schoelkopf, 1997)
  Only change the data at the boundary (not enough RAM)
- **Data augmentation for training**
  - **Imagenet** (pretty much every paper)
    Cropping, scaling, change mean, per channel, …
  - **Speech Recognition**
    Background noise, scenes, …
  - **Document Analysis**
    Random substrings, word removal, insertion

Data Augmentation

- Use prior knowledge about invariances to augment data
  - Add background noise to speech
  - Transform / augment image by altering colors, noise, cropping, distortions

Training with Augmented Data



Flip

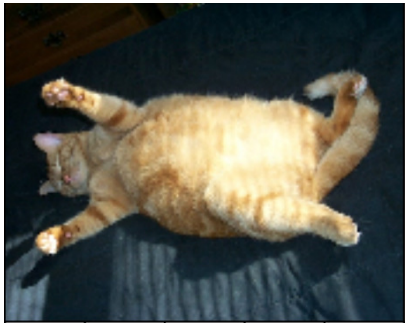
- Left-right flip

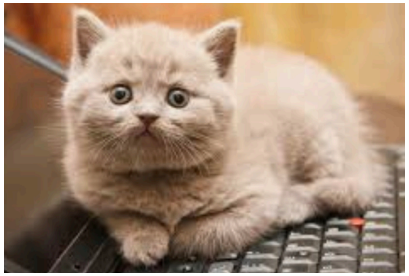
  vertical

- Top-bottom flip

  horizontal

Doesn't always makes sense



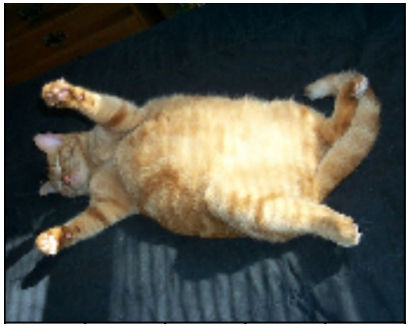


Crop

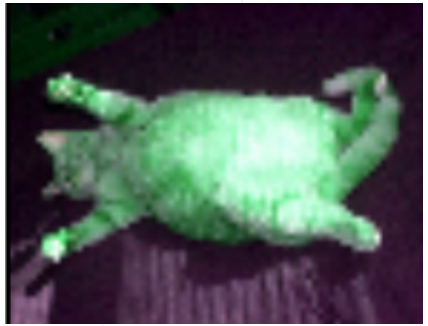
- Crop an area from the image and resize it
  - Random aspect ratio (e.g. [3:4, 4:3])
  - Random area size (e.g. [8%, 100%])
  - Random position

Color

Scale hue, saturation, and brightness (e.g. [0.5, 1.5])

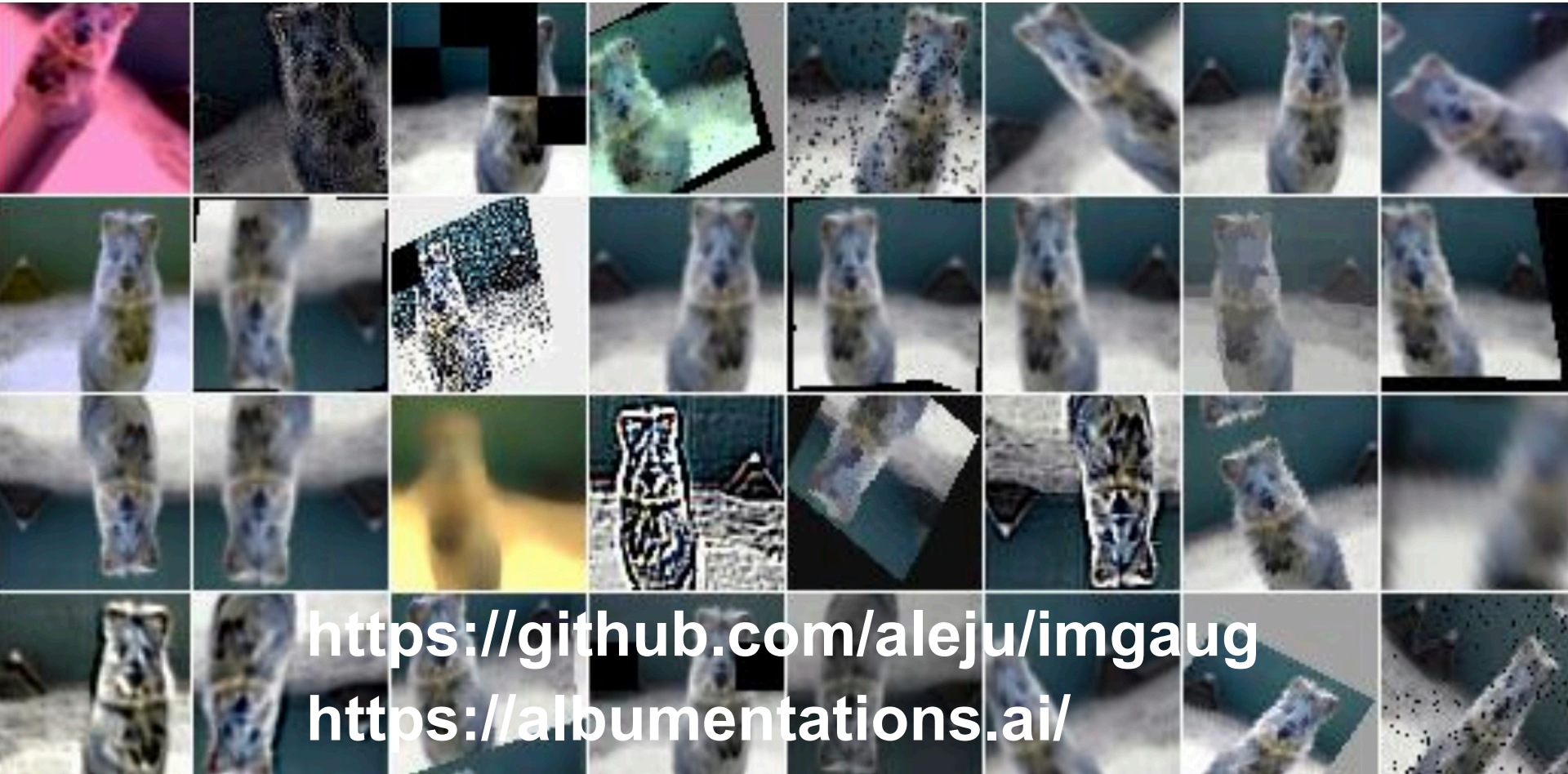


Brightness



Hue

Many Other Augmentations



Invariant and robust loss

- **Convex loss** (Teo et al, 2005)
  - Family of transformations $\delta \in \Delta$
  - Penalty for extreme transformations $1 \geq \eta(\delta) \geq 0$
  - Find the 'worst' possible example at each step

Adversarially Robust Networks

$$L(x, y, f) = \sup_{\delta \in \Delta} \eta(\delta) l(f(x + \delta), y)$$

e.g. adversarial example generator Finds worst possible

Reduced penalty for extreme distortions

Key Takeaways

- Invariances
  - We **know** that the transformation keeps outcome unchanged.
  - Add it to dataset to get more robust estimate.
- Adversarial Data
  - We **don't know** that the transformation **should** keep outcome unchanged.
  - Using it changes outcome.
  - Defense by treating it as invariance.