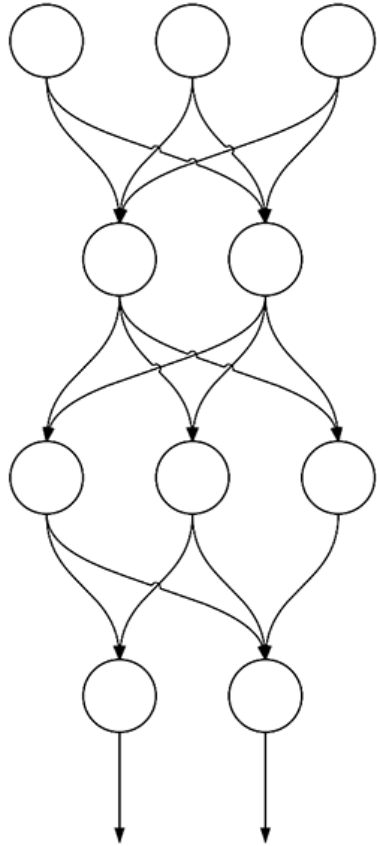




How to Feed Your Robot

Building Machine Learning
Datasets for Social Good



Evan Tachovsky, Rockefeller Foundation

@evantachovsky

September 28, 2018

Meet Charlie



Photo credit to [Eric Fleming](#) on Flickr

Computer Vision, Simplified

1. Train an algorithm with labeled data

2. Use the algorithm to **classify** new pictures of dogs and cats

Show **a picture** Show **a label**



"This is a **dog**"



"This is a **cat**"



"This is a **dog**"



"This is a **cat**"



"This is a **dog**"

...lots of other examples...



"This is a **cat**"

New **picture**

Algorithm **classifies**



"I'm 0.79 sure
that is a **cat**."

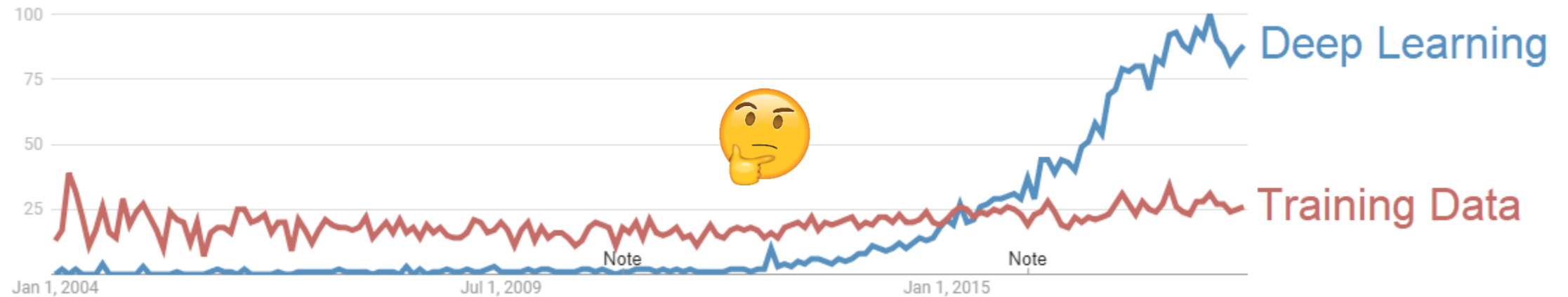


"I'm 0.87 sure
that is a **dog**."



Training Data: The Secret Superhero

What happened here?



Source: [Google Trends](#)

Building a Better Dataset: ImageNet



[Learn more](#) about **Dr. Fei-Fei Li** and the history of ImageNet

Building a Better Dataset: ImageNet

2007



Fei-Fei Li
starts work on
ImageNet

2009



**ImageNet
competition**
launches

2012



AlexNet wins,
igniting the deep
learning craze

2017



ImageNet is
essentially a solved
task; competition
ends

ImageNet helped us **benchmark**, quickly **test** new techniques, and **pre-train** models

SQuAD2.0

The Stanford Question Answering Dataset

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

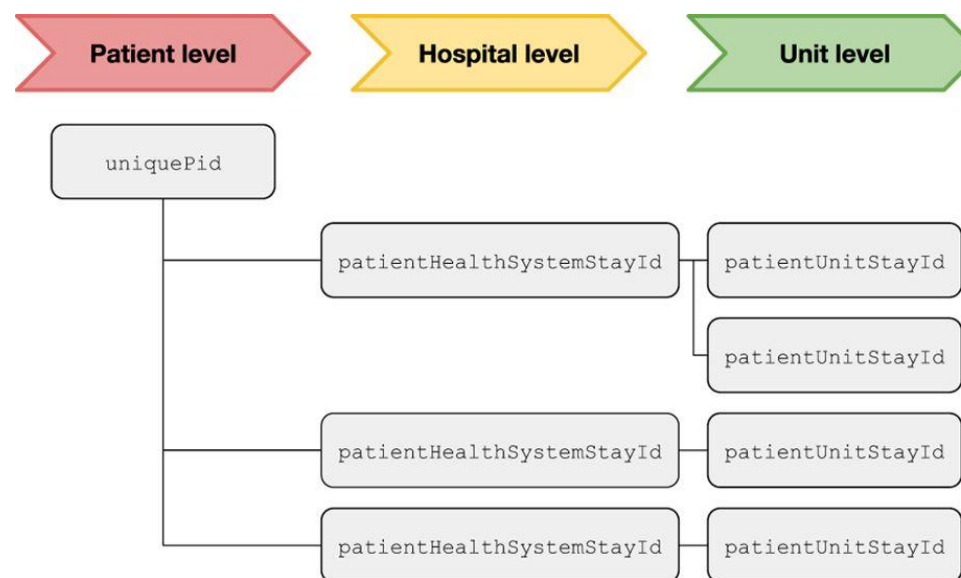
graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

The eICU Collaborative Research Database, a freely available multi-center database for critical care research

Tom J. Pollard, Alistair E. W. Johnson , Jesse D. Raffa, Leo A. Celi, Roger G. Mark & Omar Badawi

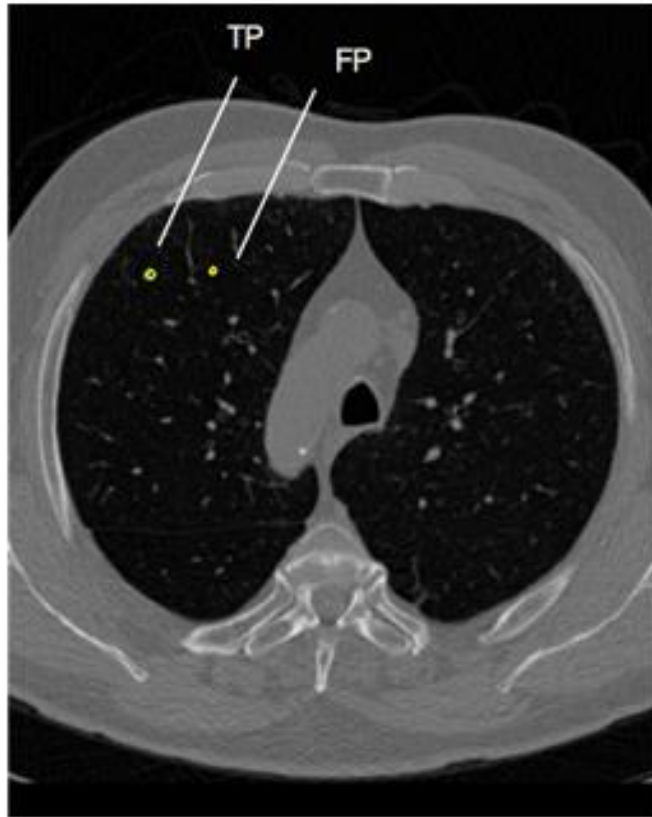


What Makes a Good Training Dataset?

1. **Sufficiently large**, generally millions of labeled examples
2. **Accurate labels** and **standardized objects**
3. **Representative** of the problem we're working on

How Training Data are Made

Expert Labeled



Crowd Labeled



How Training Data are Made

☐ I'm not a robot  reCAPTCHA

Select all images below that match this one:



   [Verify](#)

☐ I'm not a robot  reCAPTCHA

Select all images below that match this one:



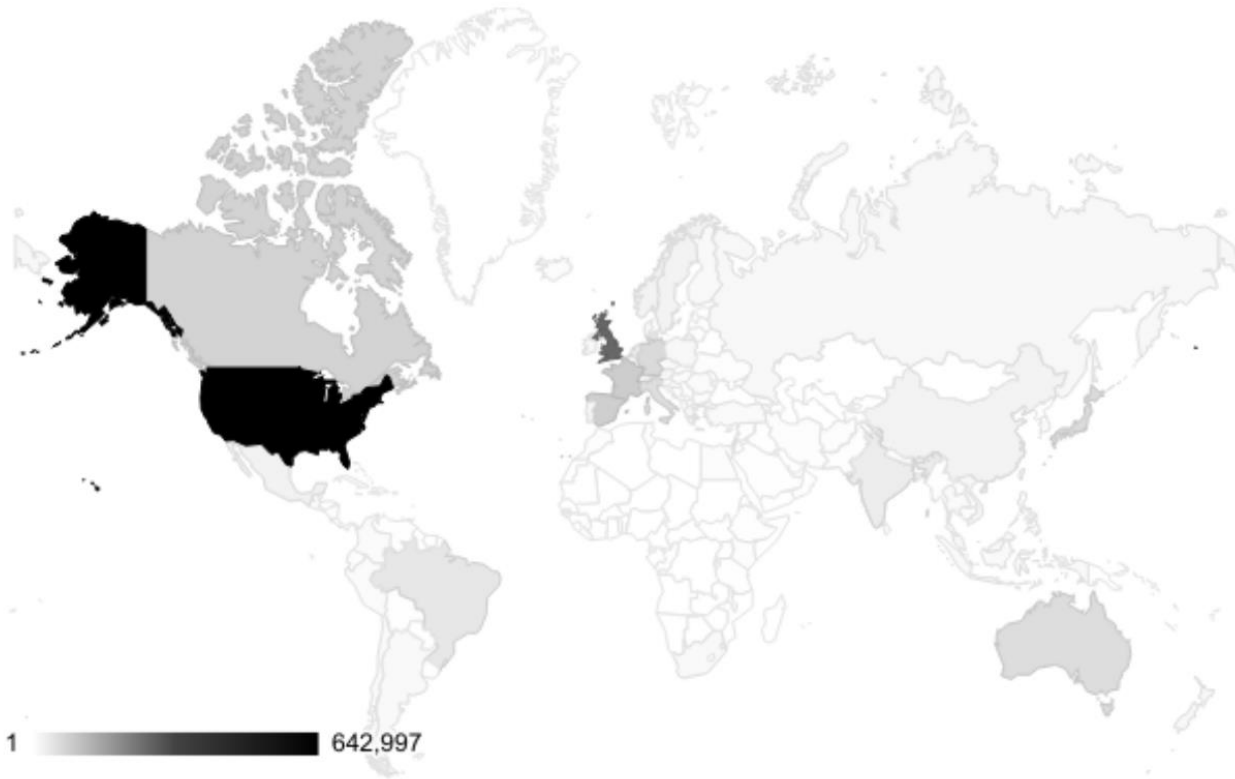
   [Verify](#)

1. For many important problems, we simply **don't have the datasets we need**
2. Many datasets we rely on are **systematically biased**

Bias and Error: What's Missing?

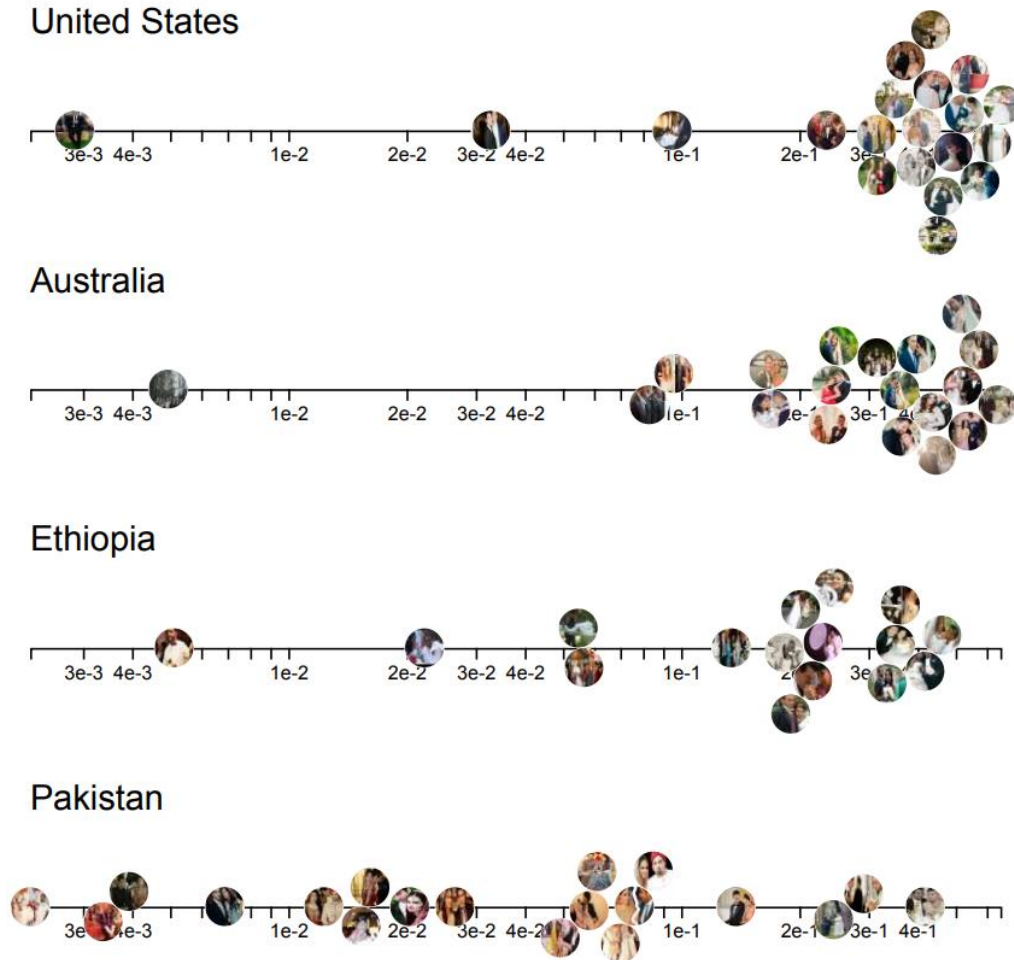
60%

of location-based data in ImageNet and Open Images comes from **six countries** in **North America and Europe**



Source [*No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World*](#) by Shankar et al.

Bias and Error: The Result



Images from Ethiopia and Pakistan are classified **less consistently** than images from the U.S. and Australia

Source [No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World](#) by Shankar et al.

Step 5: Behold the monstrosity that we have created

Not every sentence is going to contain obvious sentiment words. Let's see what it does with a few variations on a neutral sentence:

```
text_to_sentiment("Let's go get Italian food")
```

```
2.0429166109408983
```

```
text_to_sentiment("Let's go get Chinese food")
```

```
1.4094033658140972
```

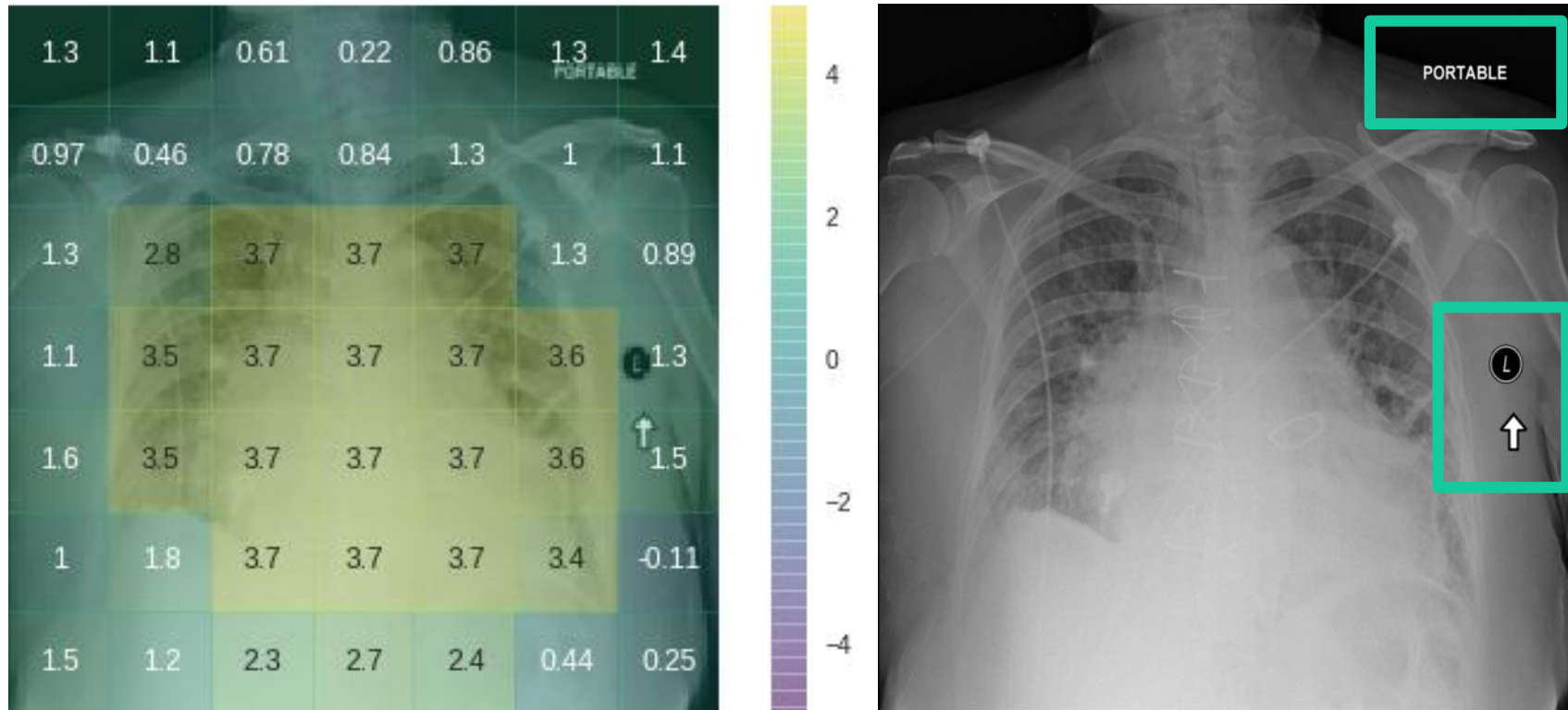
```
text_to_sentiment("Let's go get Mexican food")
```

```
0.38801985560121732
```

Source [How to make a racist AI without really trying](#), by Rob Speer

Quirks in Collection Procedures

$P(\text{Cardiomegaly})=0.752$



Source [What are radiological deep learning models actually learning?](#) by John Zech

How Can Philanthropy Help?

1. Encouraging commercial and academic organizations to build datasets that are representative and just
2. Funding new datasets for geographies and problems the market won't
3. Coordinating demand in the social sector for encourage private sector collaboration

Identifying Damage

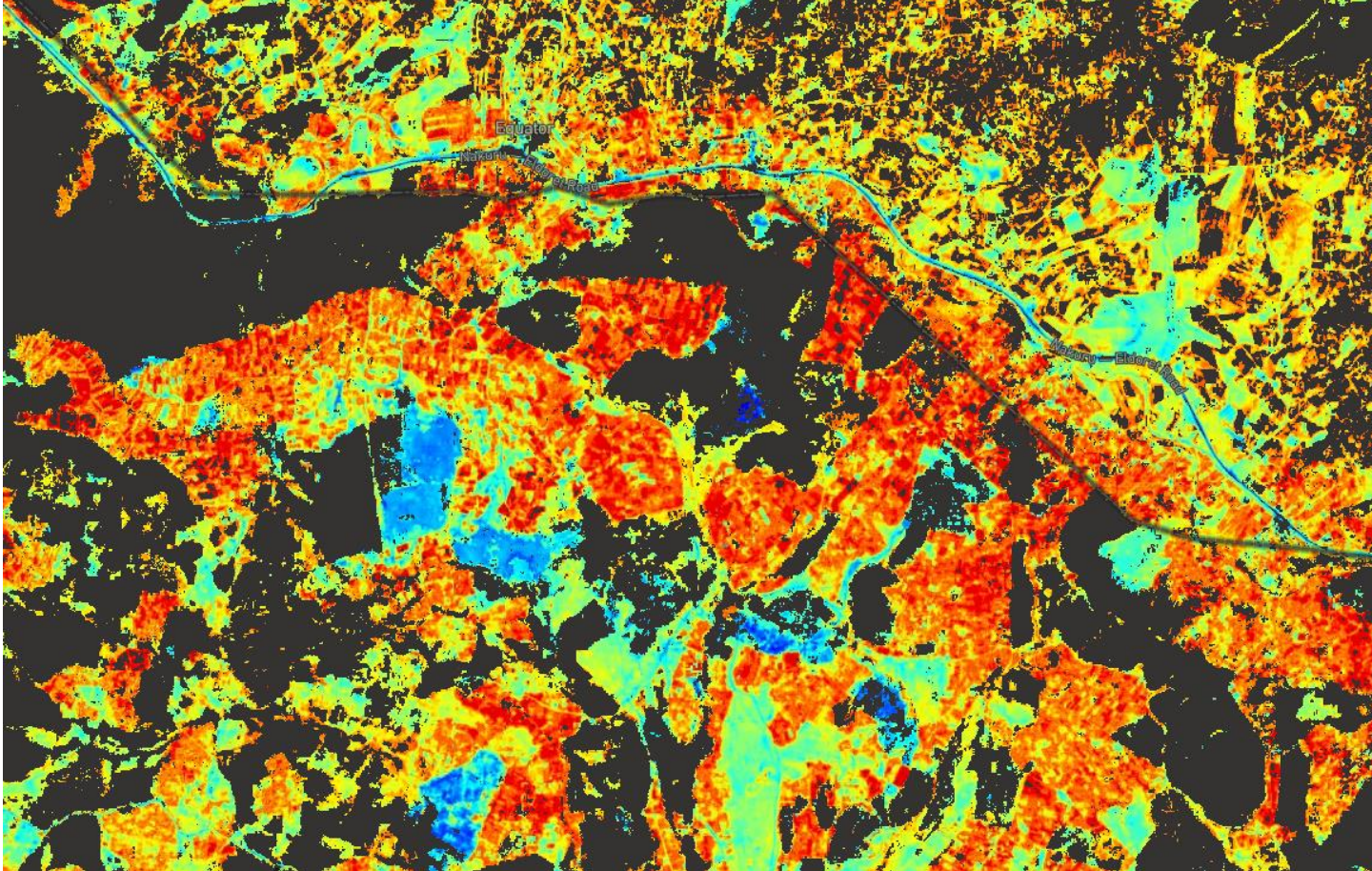


Key Issue:

Requires domain knowledge and ground access.

We
Robotics

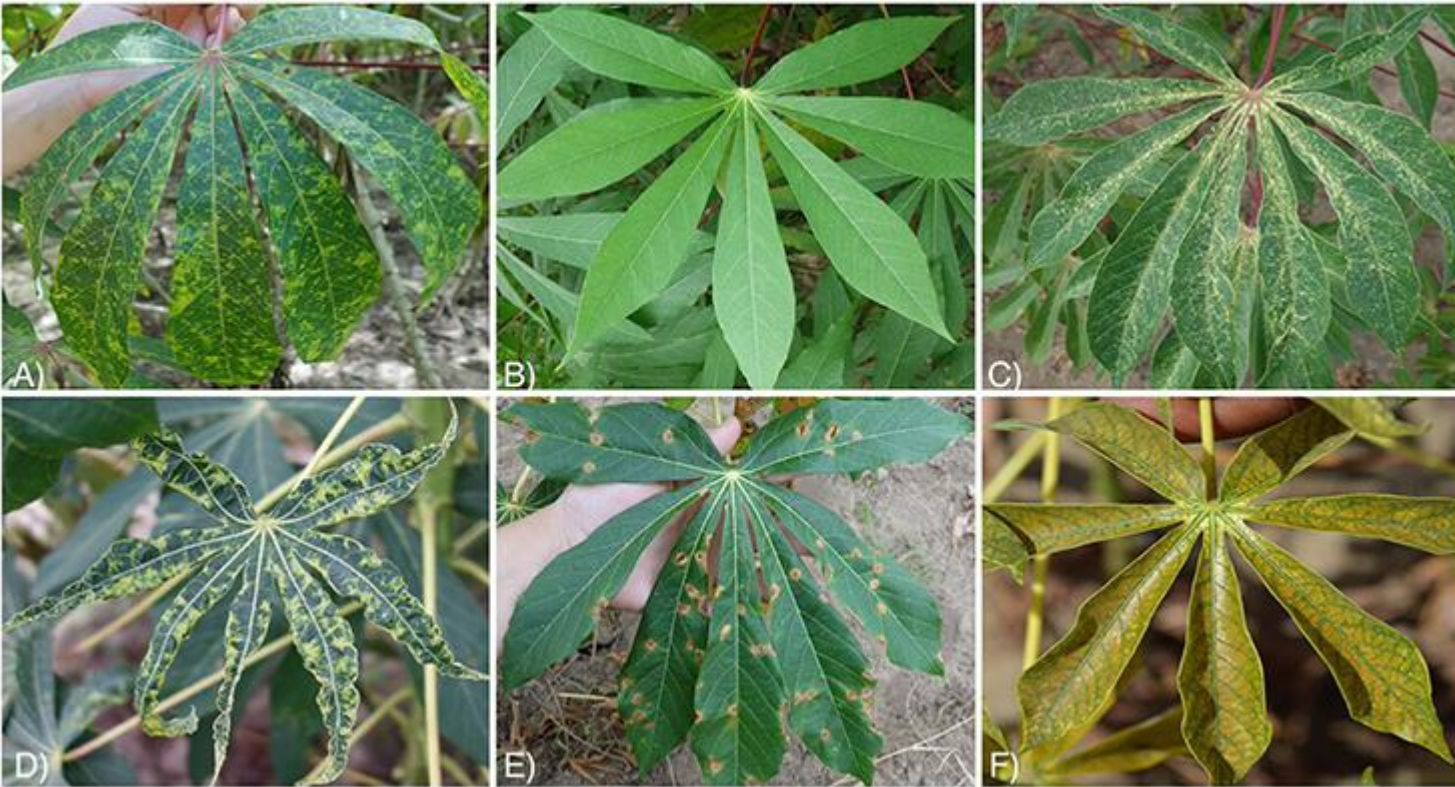
Estimating Crop Yields



Key Issue:

Limited records of land ownership. Boundaries are hard to tag remotely.

Classifying Plant Diseases

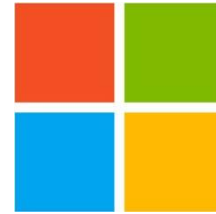


Key Issue:

Academic-collected datasets are too small and commercial datasets are expensive. How do we **incentivize collaboration**?

Source [Deep Learning for Image-Based Cassava Disease Detection](#) by Ramcharan et al.

Thank you and
thanks to



Read More

<https://www.are.na/evan-tachovsky/in-training>

Get in
touch

@evantachovsky