How to Feed Your Robot

Building and Maintaining Open Machine Learning Datasets 🤖

Evan Tachovsky, Rockefeller Foundation
@evantachovsky

2019-05-08

Supervised ML, grossly simplified

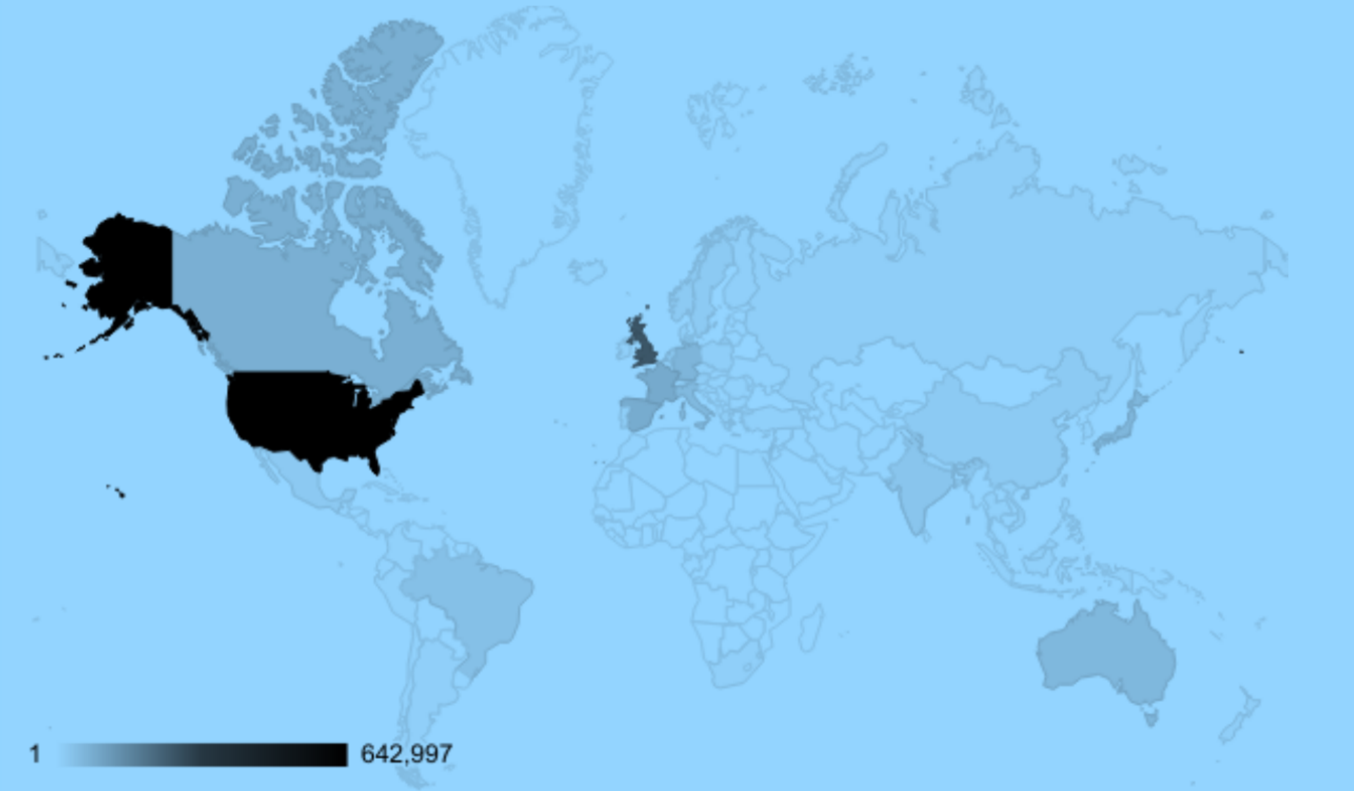
1. **Train** an algorithm with labeled data

Picture	Show **a label**
🐶	“This is a **dog**”
🐱	“This is a **cat**”
🐶	“This is a **dog**”
🐱	“This is a **cat**”
…lots of other examples…	
🐶	“This is a **dog**”

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

New **picture**	Algorithm **classifies**	
😸	“I’m 0.79 sure that is a **cat**.”	🤖
🐶	“I’m 0.87 sure that is a **dog**.”	🤖

Massive Bias Problems



60%

images in ImageNet and Open Images comes from **six countries**

Source: *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World by Shreya Shankar et al.*

Current Project

Interview **~20 data set builders, maintainers, and funders** to understand how we can build just and effective datasets for development and humanitarian use

- Semi-structured conversations
- Spread across industry, academic, research institutions
- No attribution unless the info was published
- Most interviewees were working on text or image datasets

Alex Ratner	*Alistair Johnson*	*Andrew McCallum*	*Brandeis Marshall*	*Celina Lee*
Courosh Mehanian	*Desmond Patton*	*Donghui Li*	*George Azzari*	*Jayme Garcia Arnal Barbedo*
Leo Anthony Celi	*Mutale Nkonde*	*Mutembesa Daniel*	*Nick Adams*	*Rashida Richardson*
Rediet Abebe	*Sebastian Ruder*	*Sue Márquez*	*Tariq Khokhar*	*Tony Hey*

Thank you to everyone who took the time to talk! Any errors are on me.

Five lessons

1. Motivations shape dataset
2. Transactional labels are always worse than you think
3. Find your place on the labeling spectrum
4. Don’t ignore shelf life
5. Think infrastructure, not research

1. Motivations shape datasets

Commercial

- I want to build a new product
- This dataset will give us a moat
- This dataset will make our algorithm more robust

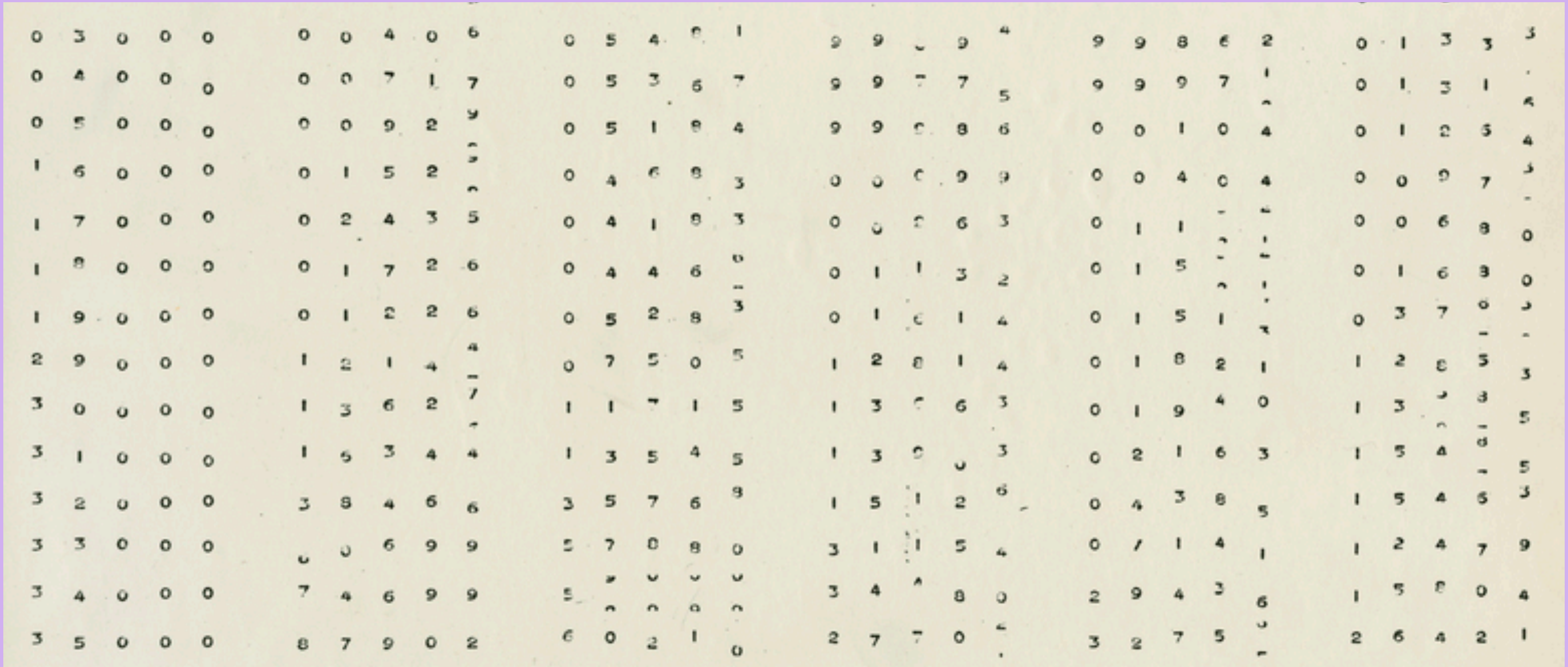
Methodological

- We need a benchmark for the field
- If we have this dataset, we can figure out this method
- I’m curious about __________.

Applied

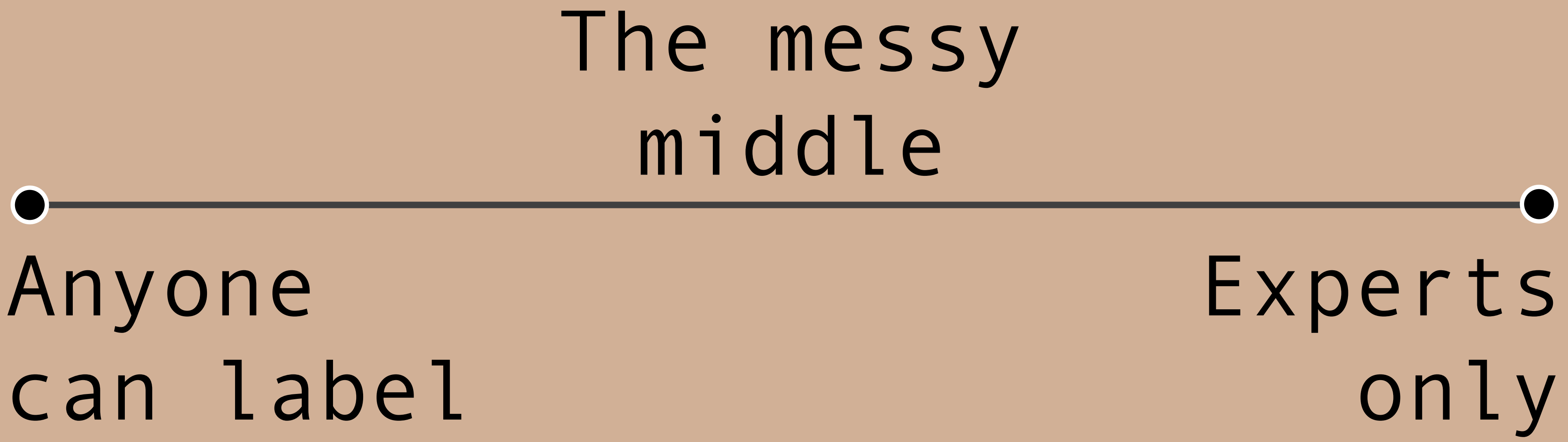
- How do I understand these data?
- How to I automate this annoying task?

2. Transactional labels are worse than you think



03000	00406	054[illegible]1	99[illegible]94	99862	01333
04000	00717	05367	99775	9997[illegible]	0131[illegible]
05000	0092[illegible]	05184	99[illegible]86	00104	01254
16000	0152[illegible]	04683	00[illegible]99	00404	0097[illegible]
17000	02435	04183	00[illegible]63	011[illegible][illegible]	00680
18000	01726	0446[illegible]	01132	015[illegible][illegible]	01630
19000	01226	05283	01[illegible]14	0151[illegible]	037[illegible][illegible]
29000	1214[illegible]	07505	12814	01821	12853
30000	1362[illegible]	11715	13[illegible]63	01940	13[illegible]35
31000	15344	13545	13[illegible][illegible]3	02163	154[illegible]5
32000	38466	35769	15126	04385	15453
33000	[illegible]0699	57080	31154	07141	12479
34000	74699	5[illegible][illegible][illegible][illegible]	34[illegible]80	29436	15[illegible]04
35000	87902	60210	2770[illegible]	3275[illegible]	26421

3. Find your place on the labeling spectrum



The messy middle

Anyone can label

Experts only

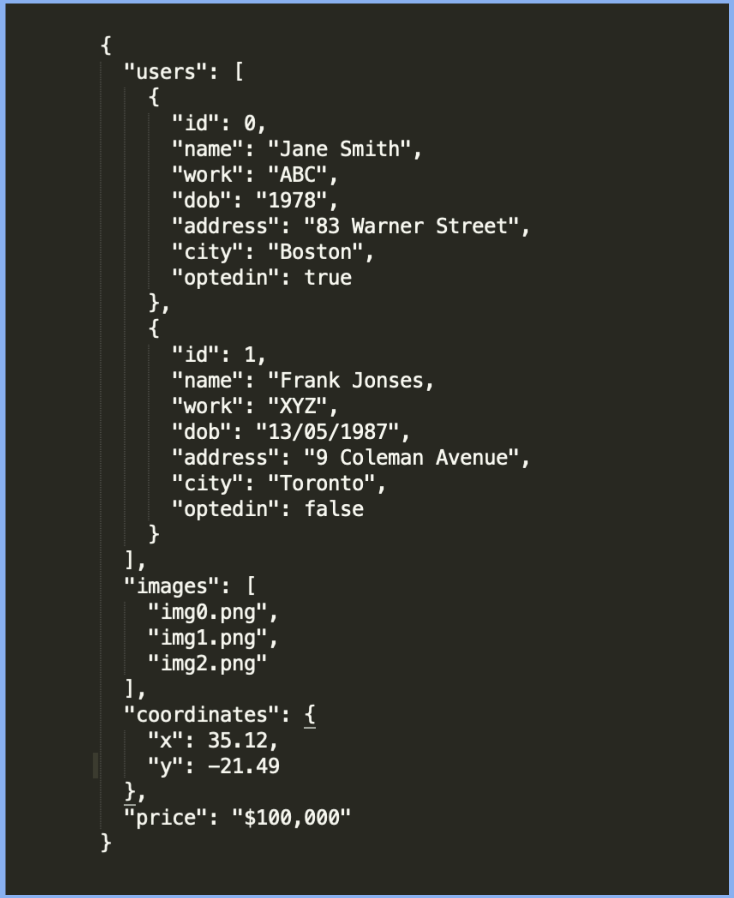
3. Find your place on the labeling spectrum

Strategies for the **messy middle**

- Incentives, not just financial
- New tools to make tasks easier
- Change the classification problem

4. Don’t ignore shelf life

Use Buy:
2020-02-28



```
{
  "users": [
    {
      "id": 0,
      "name": "Jane Smith",
      "work": "ABC",
      "dob": "1978",
      "address": "83 Warner Street",
      "city": "Boston",
      "optedin": true
    },
    {
      "id": 1,
      "name": "Frank Jonses,
      "work": "XYZ",
      "dob": "13/05/1987",
      "address": "9 Coleman Avenue",
      "city": "Toronto",
      "optedin": false
    }
  ],
  "images": [
    "img0.png",
    "img1.png",
    "img2.png"
  ],
  "coordinates": {
    "x": 35.12,
    "y": -21.49
  },
  "price": "$100,000"
}
```

5. Think infrastructure, not research

- Ubiquitous, not bespoke
- For a community, not a team
- A separate budget, not a budget line
- Measure and reward contributions

Background Reading:
https://www.are.na/evan-tachovsky/in-training

Thank you! Get in touch

Evan Tachovsky,
Rockefeller Foundation
@evantachovsky
2019-05-08