Lecture 7: Efficiency in Training, Eval, and Deployment

Sara Beery | 4/1/25

Challenge: AI models require expensive hardware
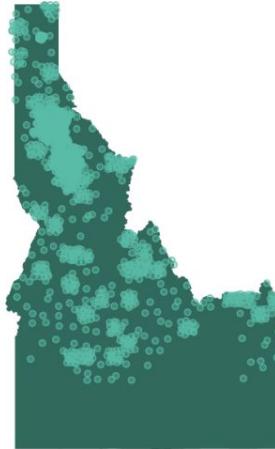
These are inaccessible to many potential AI users in biodiversity

Increased efficiency reduces cost

MegaDetector is used to process data for NGOs and conservation organizations globally
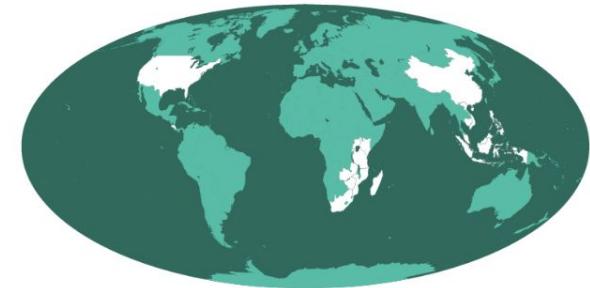
Seeing biodiversity: perspectives in machine learning for wildlife conservation, Tuia*, Kellenberger*, Beery*, Costelloe*, et al., Nature Communications (to appear)
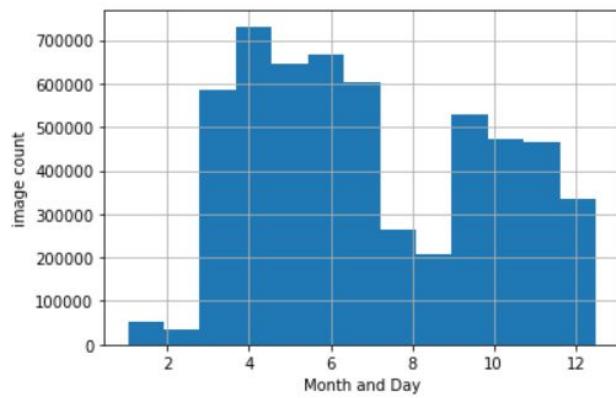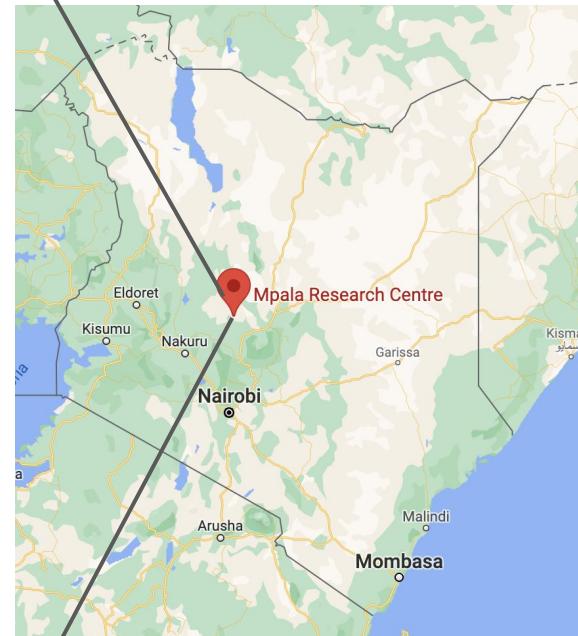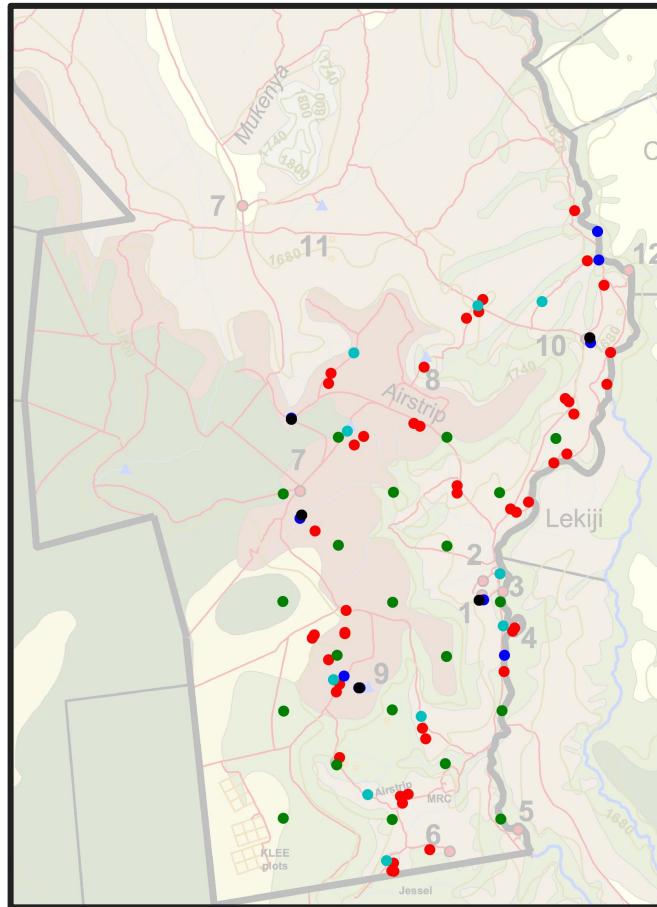
Bandwidth is limited in the field

Up to 700K hi-resolution images per month

Commercial edge-based AI camera traps in development

Focused on anti-poaching
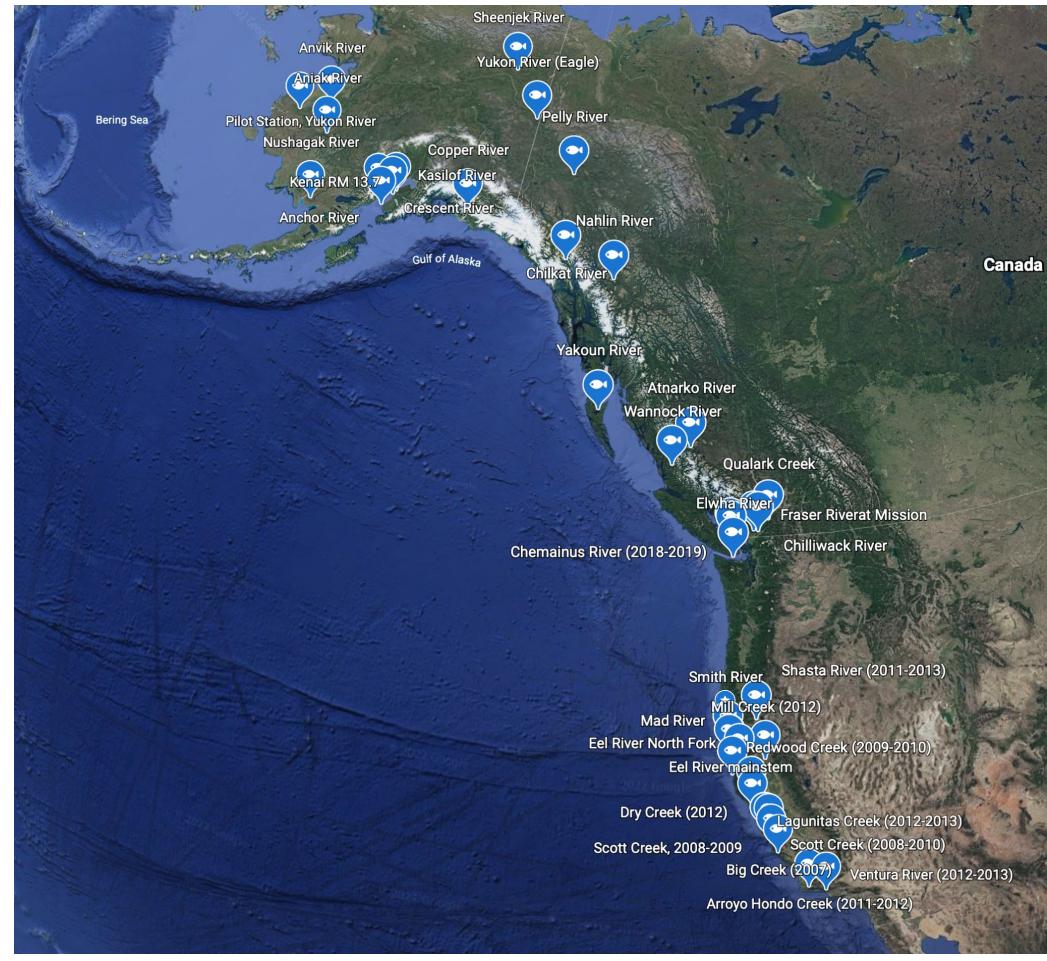
Focused on anti-poaching & human-wildlife conflict

Focused on real-time animal behavior monitoring

Sonar deployment to monitor salmon returns
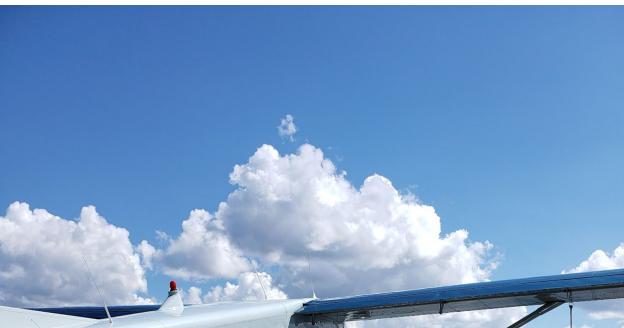
Photos courtesy of ADFG

We need near-real-time counts from remote field sites

We need near-real-time counts from remote field sites

This requires edge-based models that are robust and reliable even as environmental conditions change

www.inaturalist.org

How It Works

1. Record your observations

2. Share with fellow naturalists

3. Discuss your findings

Observations
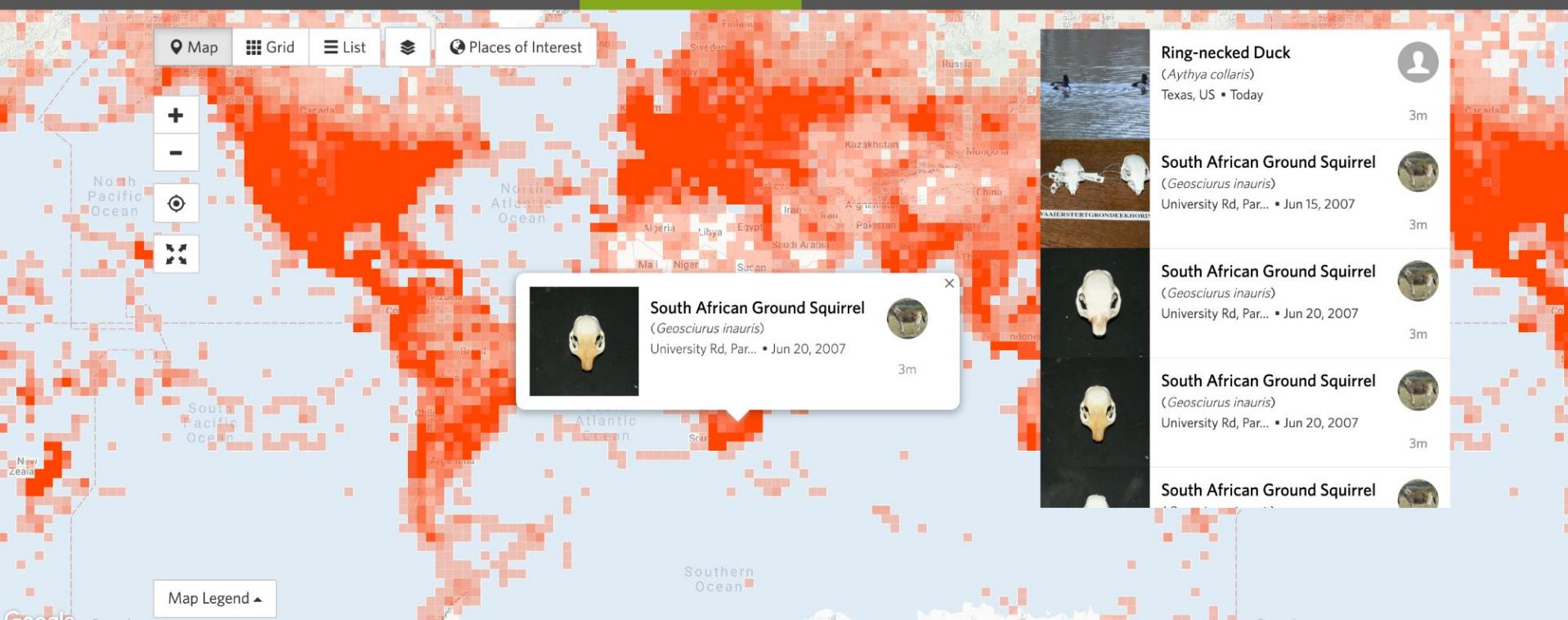
| Species | | Location | | Go | Filters |

📍 Map | ▦ Grid | ☰ List | ⬒ | 🌐 Places of Interest

Ring-necked Duck
(*Aythya collaris*)
Texas, US • Today
3m

South African Ground Squirrel
(*Geosciurus inauris*)
University Rd, Par... • Jun 15, 2007
3m

South African Ground Squirrel
(*Geosciurus inauris*)
University Rd, Par... • Jun 20, 2007
3m

South African Ground Squirrel
(*Geosciurus inauris*)
University Rd, Par... • Jun 20, 2007
3m

South African Ground Squirrel

South African Ground Squirrel
(*Geosciurus inauris*)
University Rd, Par... • Jun 20, 2007
3m

Map Legend ▲

Real-time, on-device fine grained categorization

Training efficiency

- Fine tuning often decreases training time, and thus costs
- Reducing sample overlap or removing noisy training data can reduce training costs without impacting performance, or sometimes improving performance (Coresets, DataComp)
- Smaller models train faster

Training on the edge

- Power
- Hardware
- Bandwidth
- Verification
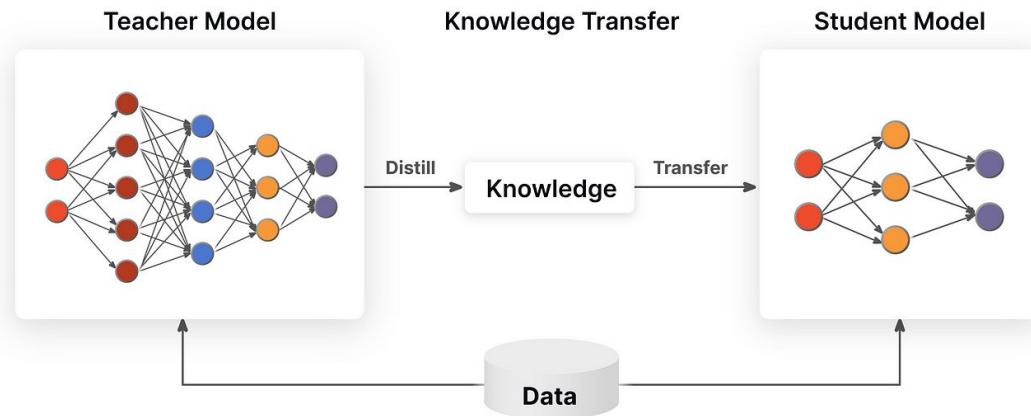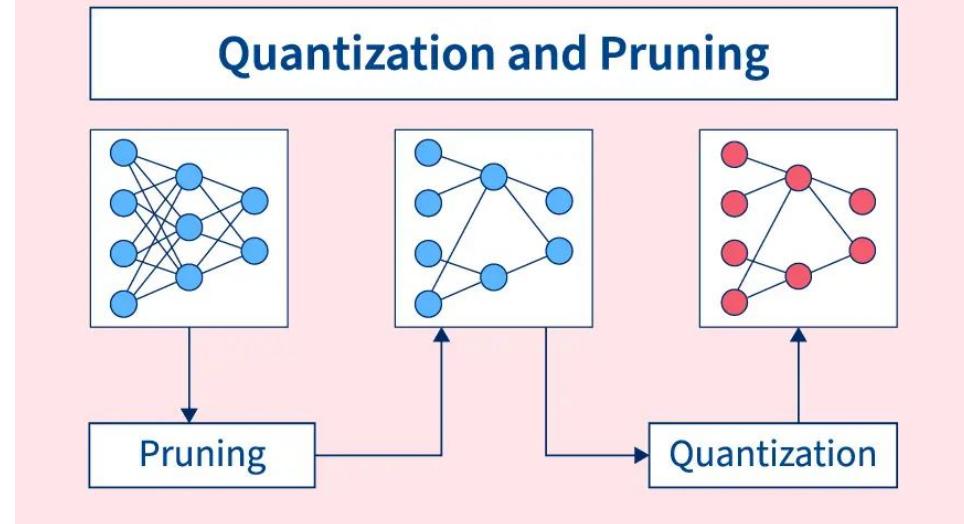
Evaluation efficiency
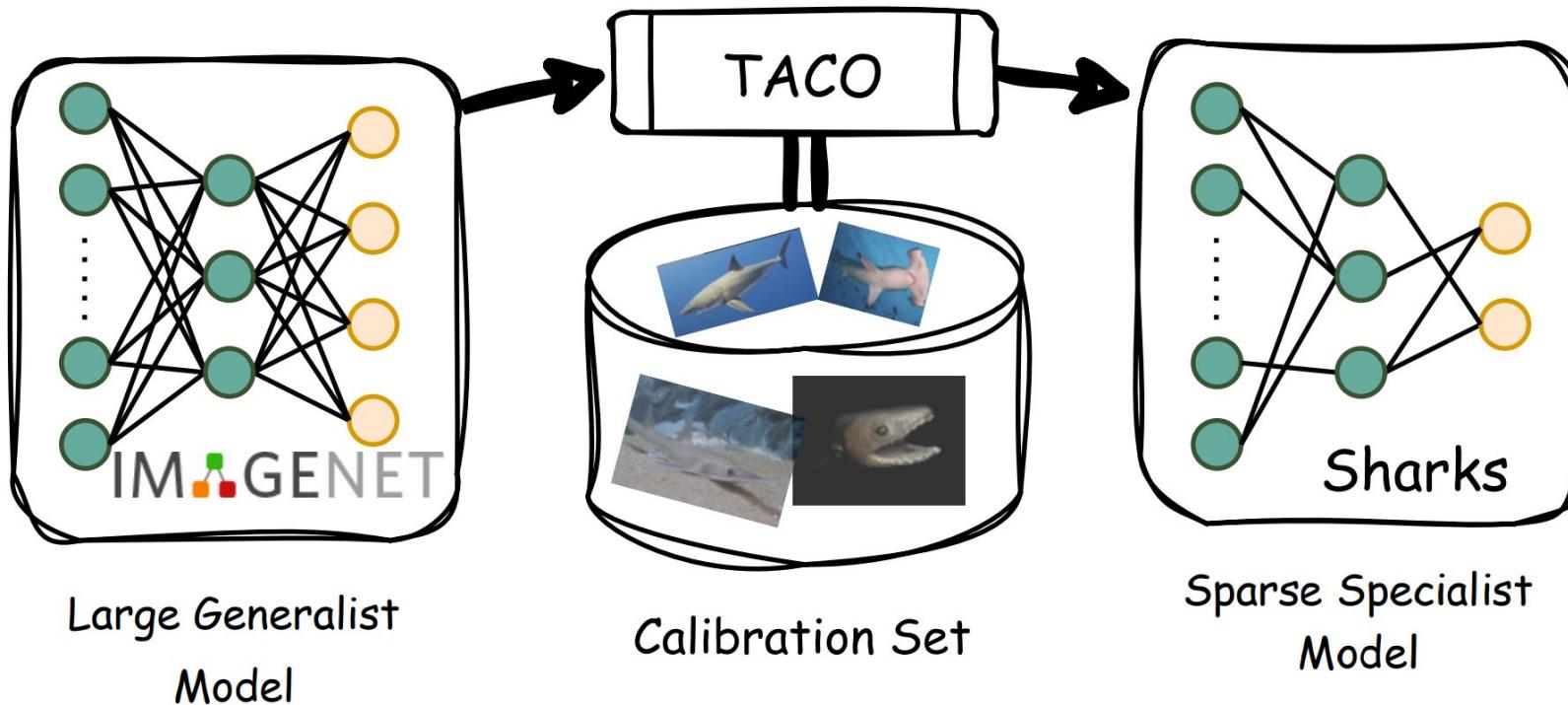
- Inexpensive proxy tests (ie brightness)
- Small (but representative) test sets
- Striation and multiple metrics
 - Make the most of your inference calls
- Active testing (soon!)

Inference efficiency

- Quantization
- Pruning
- Distillation
- Routing

We can quickly compress large generalist models into accurate and efficient specialists

Large Generalist Model

Calibration Set

Sparse Specialist Model

Vision Models Can Be Efficiently Specialized via Few-Shot Task-Aware Compression, Kuznedelev,…, Beery, Kurtic, Alistarh et al. 2023

Federated learning

- Maintains data privacy
- Can be efficient at the edge
- Requires bandwidth and synchronization