



# Apache Spark & Citizen Science

Using eBird Data to Predict Bird Abundance at Scale

**Tom Auer** ([mta45@cornell.edu](mailto:mta45@cornell.edu)), Daniel Fink, and Steve Kelling  
Cornell Lab of Ornithology

# Background

# The **Cornell** Lab of Ornithology



Our mission: To interpret and conserve the earth's biological diversity through research, education, and citizen science *focused on birds.*

# Why Birds?



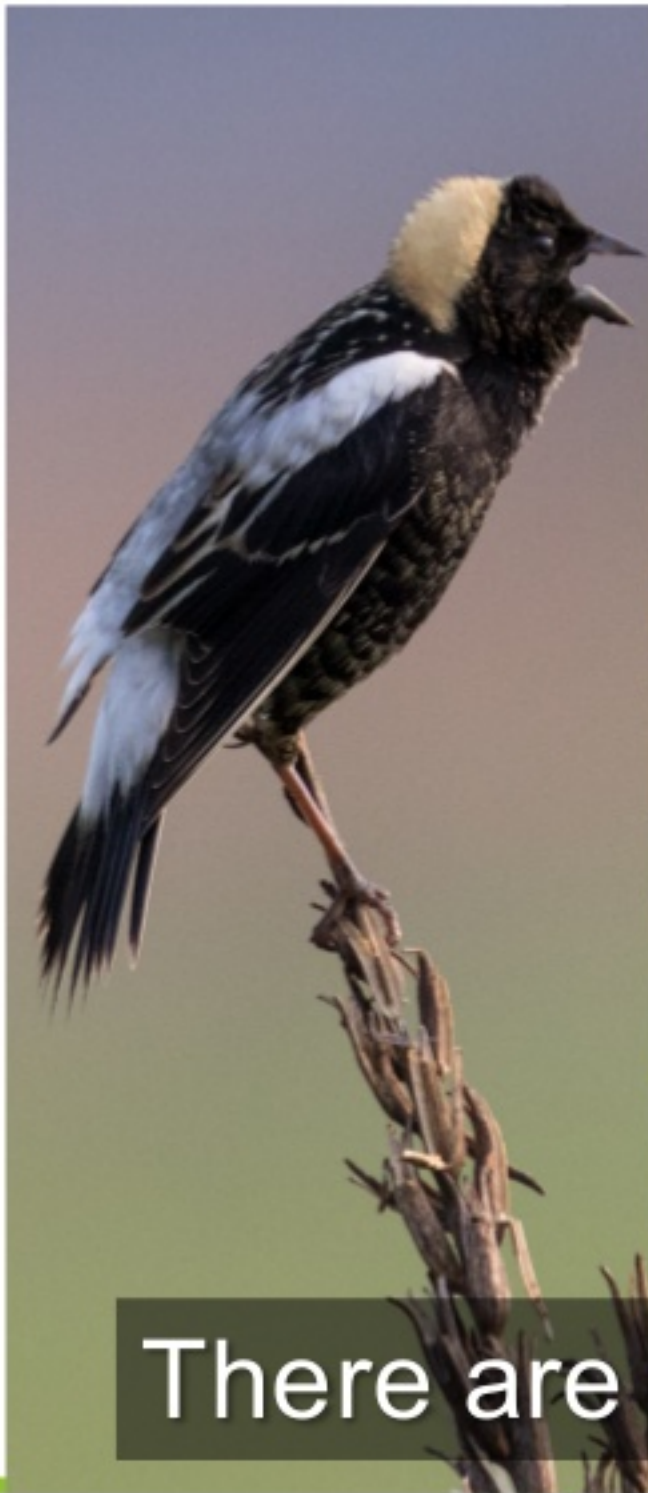


A photograph of a bird of prey, possibly a hawk or eagle, standing on a dead animal (likely a rabbit or small deer) in a snowy field. The bird has brown and white mottled feathers and a sharp beak. The background is a snowy landscape with some bare branches.

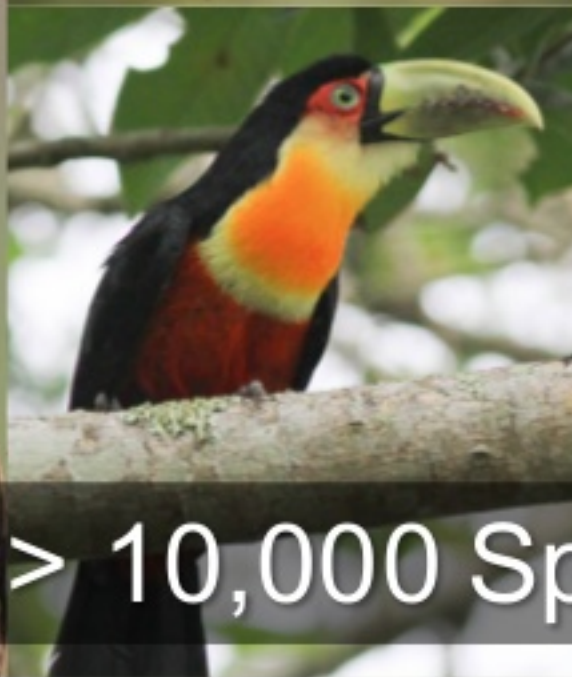
# Why Birds?

They are living dinosaurs





Why Birds?



There are > 10,000 Species





They are found in all environments



Why  
Birds?



A small brown bird with streaked plumage is perched on a white plate. On the plate is a sandwich with a pinkish filling and a yellow filling, possibly cheese or fruit. The bird is looking towards the left. In the background, there is a dark, curved object with the word "Smithology" written on it. The scene is set on a dark table.

# Why Birds?

They can adapt to novel environments



# Why Birds?

They are sensitive environmental indicators

A large Indian Vulture is shown in flight, its wings spread wide, against a background of a dense, green forest. The bird is brown with a lighter-colored head and neck. The text is overlaid on the left side of the image.

# Indian Vulture

India & Pakistan  
30,000,000 in 1990  
functionally extinct today



# Why Birds?

They are the most easily observed,  
counted, and studied of all widespread  
animal groups





# eBird

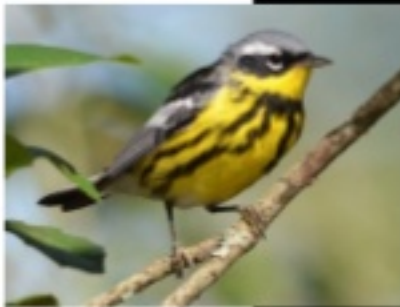


A world map with a dark background, where landmasses are highlighted in a glowing yellow-gold color. The map is covered with numerous small, bright yellow-gold dots, representing bird observation points. These dots are most densely clustered in North America, Europe, and parts of Asia and Africa, with more sparse distributions in South America and Australia. The overall effect is a global network of bird sightings.

# eBird

**30 million hours** collecting bird observations

The conservation of species begins with an understanding of the *distribution*, *abundance*, and *movements of organisms* across wide geographic areas, with *high spatial resolution* and over long periods of time.



January 4  
Magnolia Warbler

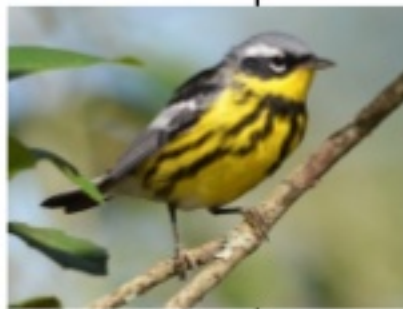
eBird

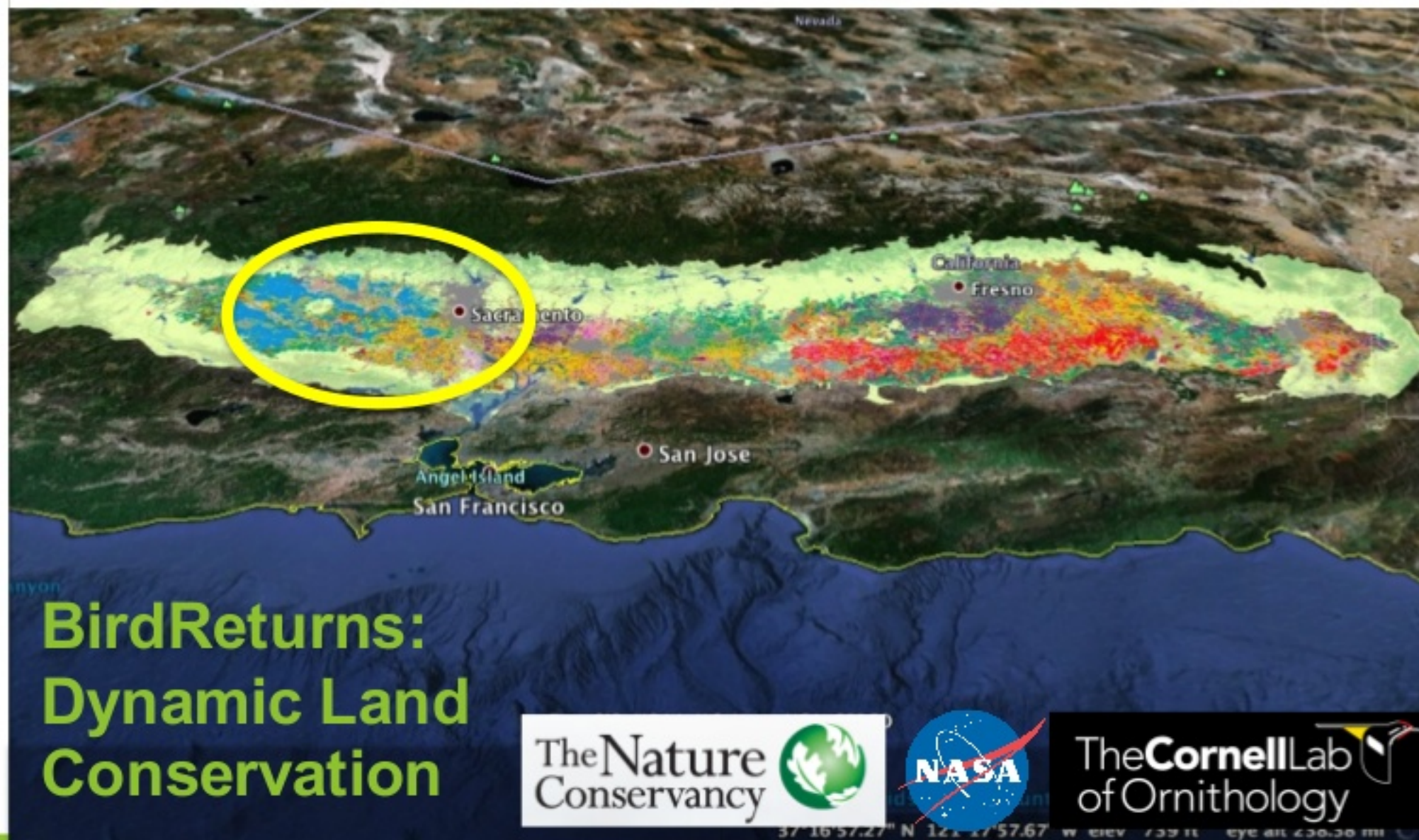




Microsoft  
Azure

not currently be displayed.





## BirdReturns: Dynamic Land Conservation

The Nature  
Conservancy



The Cornell Lab  
of Ornithology



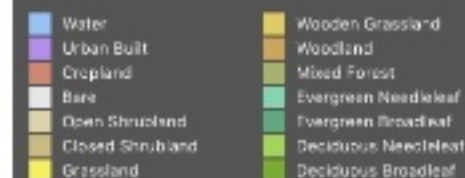




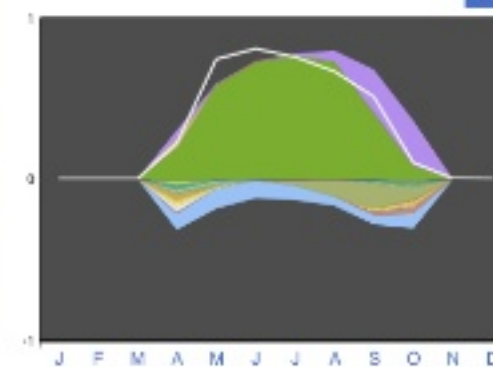
## Abundance

## Habitat

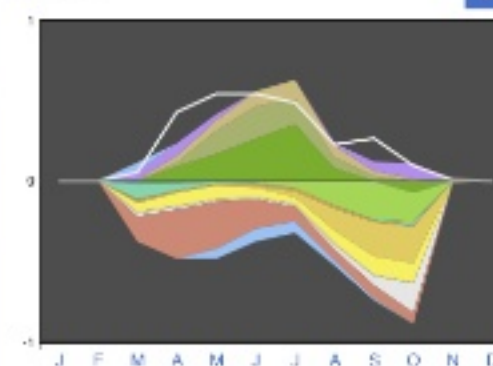
## Habitat Legend

[show/hide](#)

## Area 1

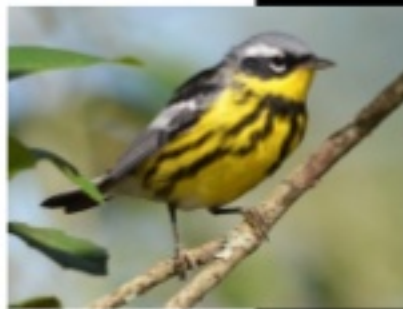


## Area 2

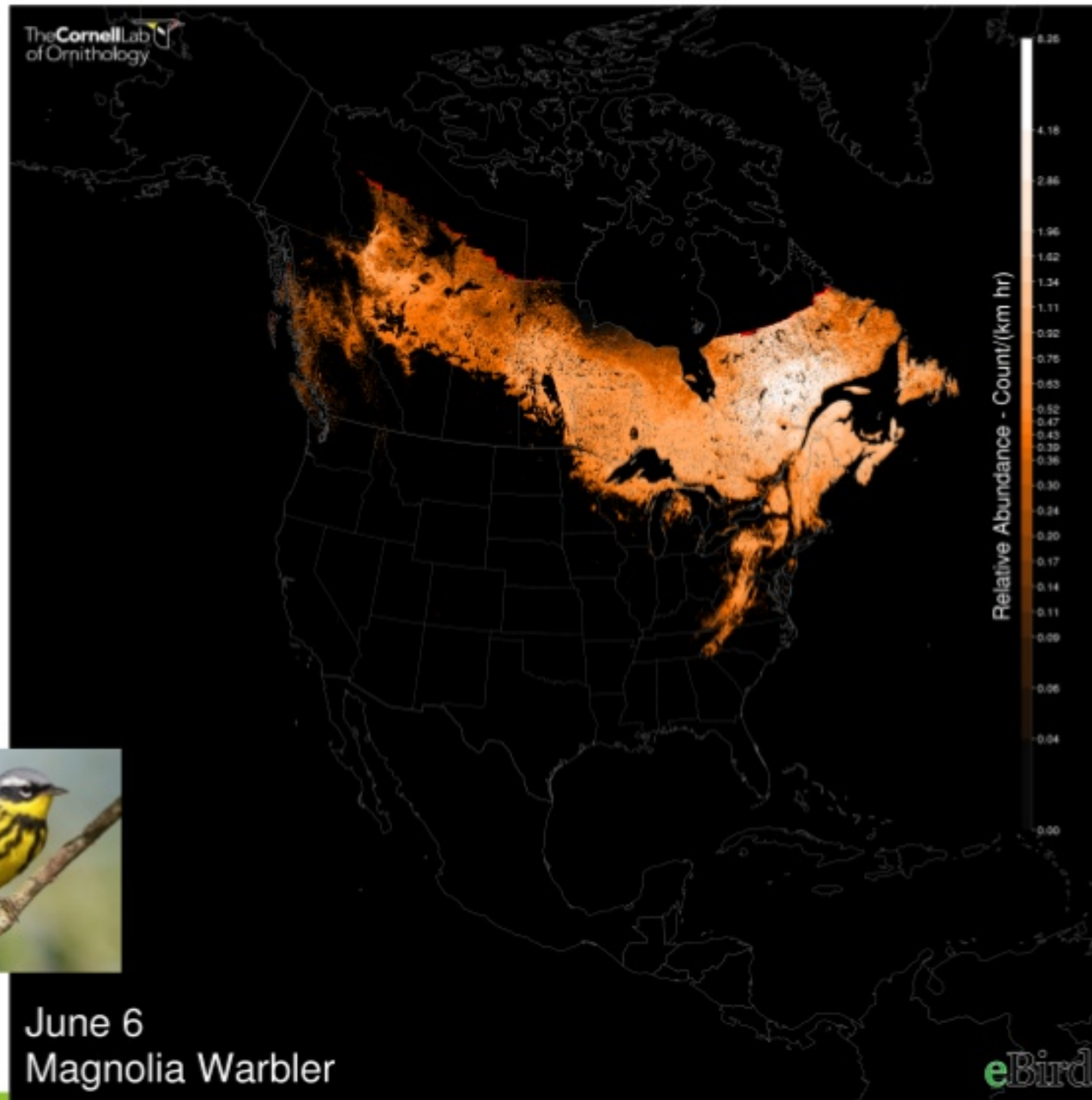


# Technical Experience





June 6  
Magnolia Warbler



# Linking Populations and Environment



$$\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2}$$

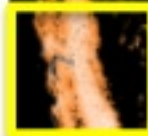


Machine Learning





# Scaling: Ecological Challenges



**Tree Swallow, March**



# SpatioTemporal Exploratory Model (STEM)

## 1. Divide

- Partition extent into regions
- Train and predict models within regions



100  
randomized  
replicates

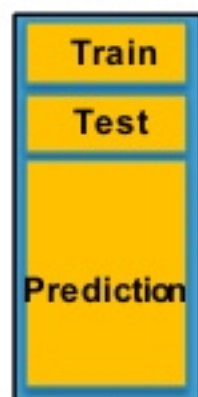
## 2. Recombine

- Average predictions across all models for each location within regions



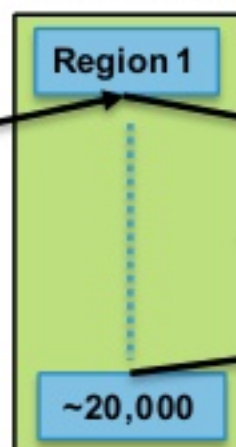


**Step 1:**  
Divide into Regions

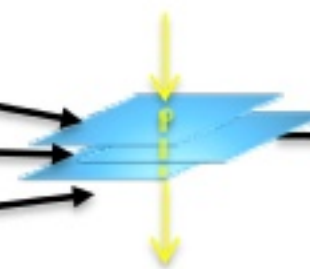


Randomized  
Partitions

**Step 2:**  
Fit & Predict in each Region



**Step 3:**  
Summarize by Location



**Step 4:**  
Write Out

**.gz**

**12m rows  
2gb**

**3.1b rows  
750gb**

**1b rows  
100gb**

**52m rows  
12gb**

# Code





# Current Modeling

- Sampling to address class imbalance
- First stage: binary response GBM
  - Calibrate with GAM
- Second stage: Poisson response GBM

*Models use weights*

# Future Modeling

- “Occupancy” Models
- Semi-parametric learning: GamboostLSS
- Statistical/Machine Learning models: suRFing



# What have we tried?

- HPC Parallelization
- Hadoop MapReduce
- ~~SparkR~~
- Spark 2.x



Cornell University  
Center for Advanced Computing

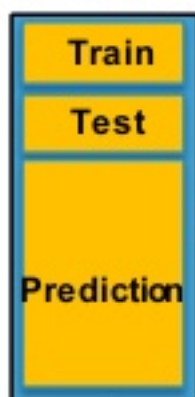


# RDD pipe()



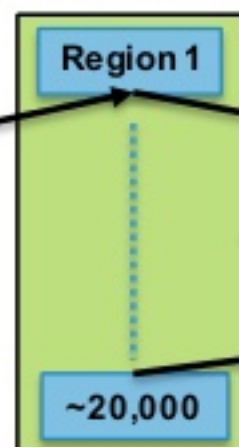


**Stage 0:**  
Divide into Regions

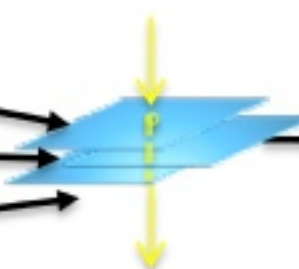


Randomized  
Partitions

**Stage 1:**  
Fit & Predict in each Region



**Stage 2:**  
Summarize by Location



**Stage 3:**  
Write Out

.gz

12m rows  
2gb

3.1b rows  
750gb

1b rows  
100gb

52m rows  
12gb

```

pipe(
  FileFormat(sparkContext.hadoopConfiguration, "text", "org.apache.hadoop.mapreduce.lib.input.FileInputFormat") \
    .load(sparkContext.hadoopConfiguration, "text", "org.apache.hadoop.mapreduce.lib.input.FileInputFormat") \
    .mapPartitionsWithIndex(
      lambda p, index, _ : (p.map(lambda line: line.split("\t")), index) \
    ) \
    .groupByKey() \
    .map(lambda p: (p[0].split("\t")[0], p[1].join(p[0].split("\t")[1:len(p[0].split("\t"))] + "EOL"))) \
    .groupByKey() \
    .map(lambda p: (p[0].split("\t")[0], p[1].join(p[0].split("\t")[1:len(p[0].split("\t"))] + "EOL"))) \
    .groupByKey().mapValues(list) \

```

# Spark!

- Fast: ~25% faster than MapReduce
- Portable: HPC, Azure
- Scalable: data volume doubled



Cornell University  
Center for Advanced Computing



Microsoft Azure



# What's next?

- RDD pipe()?
- Spark DataFrames
- More Spark!



# Summary

RDD pipe() allows us to keep our code base within our community language and use new R modeling libraries, while leveraging the speed of Spark for parallelizing our modeling workflow to address ecological challenges.





# Thank You.

Tom Auer (mta45@cornell.edu)

