Apache Spark & Citizen Science

Using eBird Data to Predict Bird Abundance at Scale

Tom Auer (mta45@cornell.edu), Daniel Fink, and Steve Kelling
Cornell Lab of Ornithology

SPARK SUMMIT 2017

Background

TheCornellLab of Ornithology

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

Why Birds?

Why Birds?

They are living dinosaurs


Why Birds?
There are > 10,000 Species


They are found in all environments
Why
Birds?


Why
Birds?
They can adapt to novel environments


Why
Birds?
They are sensitive environmental indicators


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

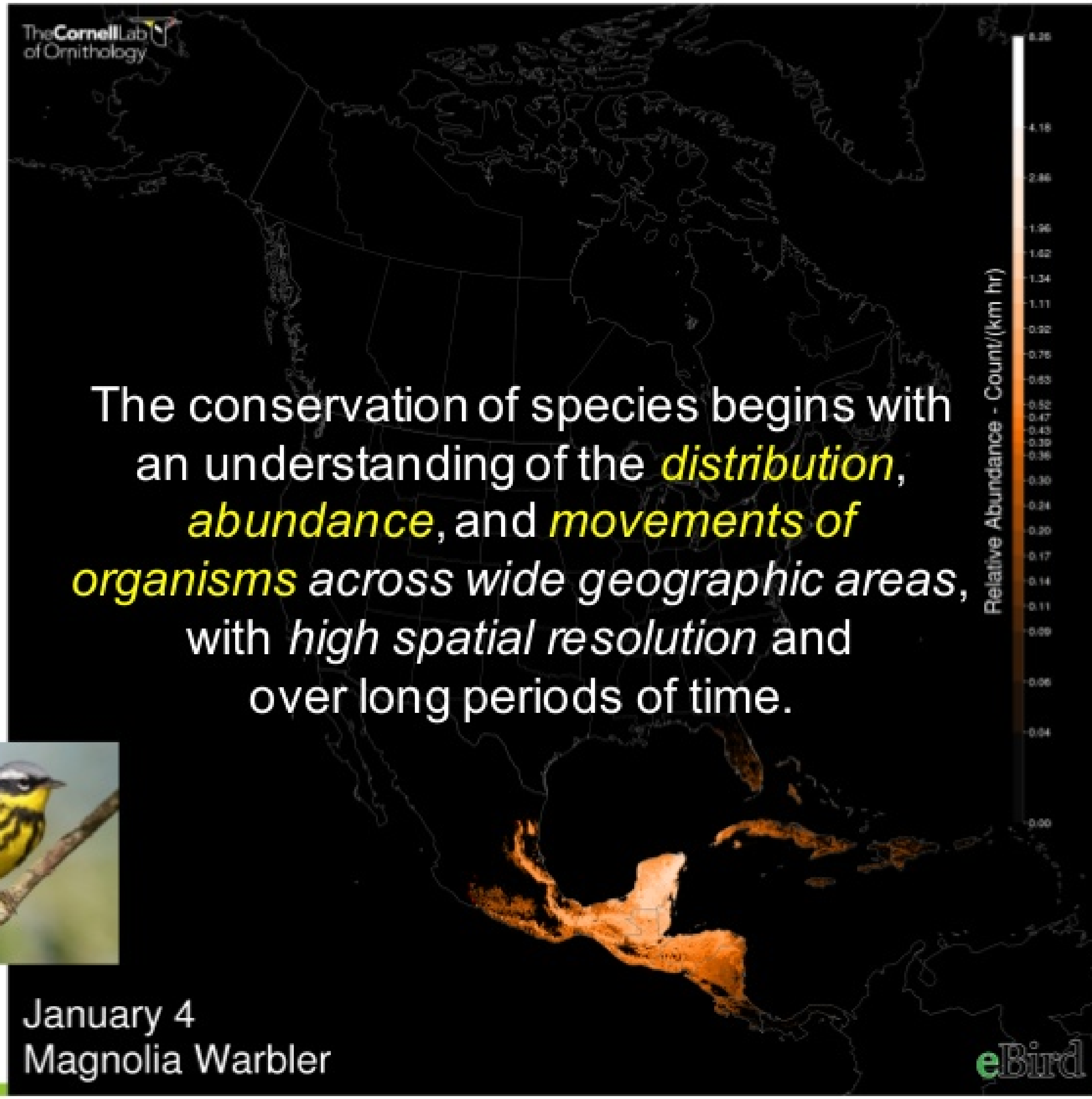

Start
NEW CHECKLIST
My Checklists

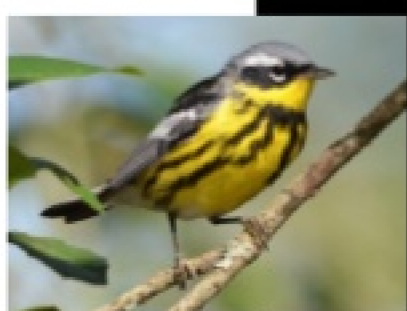

eBird
Welcome to eBird
eBird News and Features
Support eBird
Donate Now
Audubon

eBird

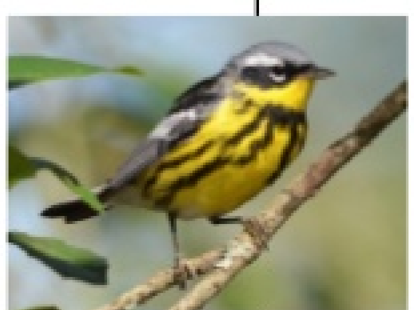
30 million hours collecting bird observations

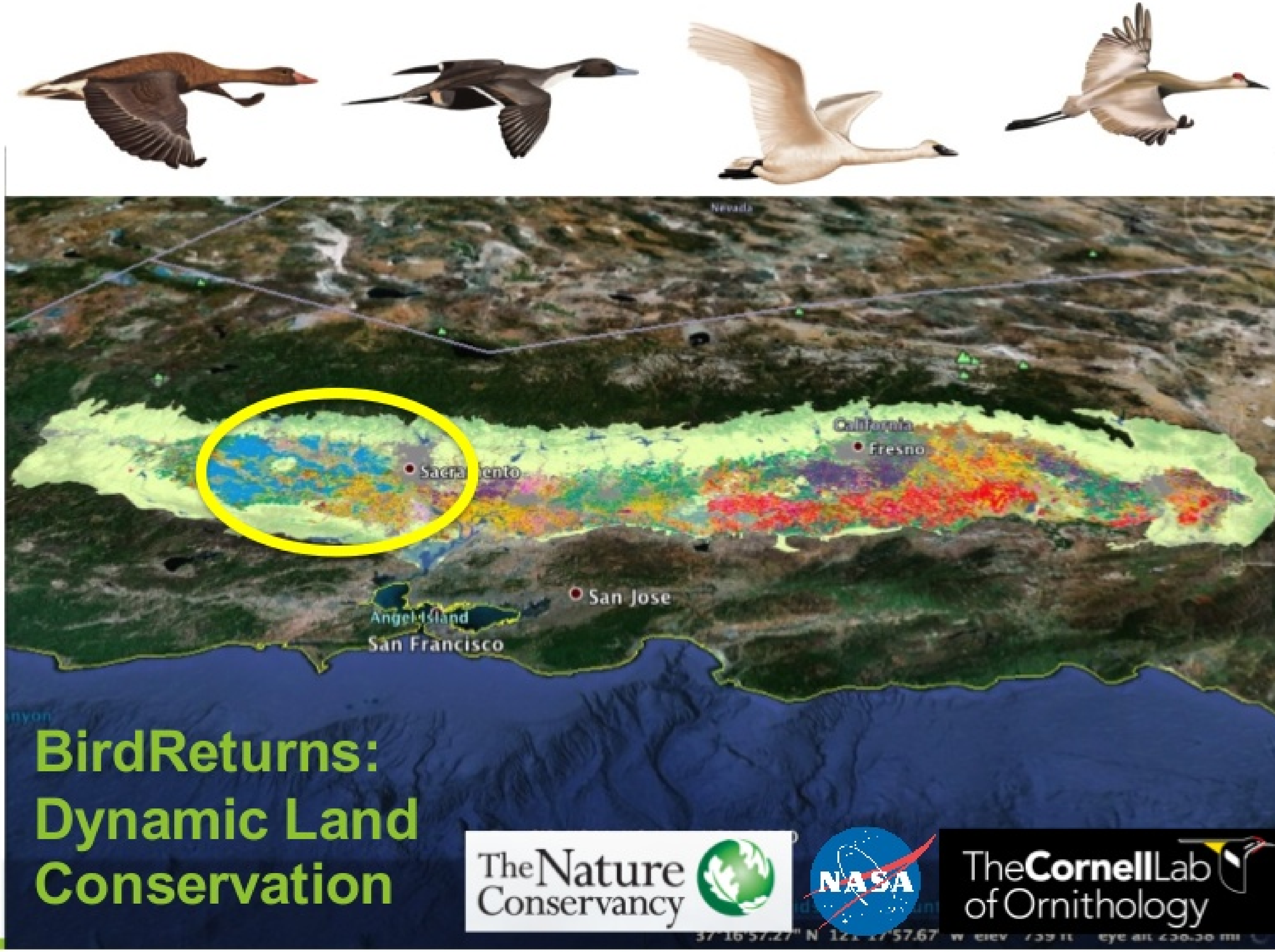
The conservation of species begins with an understanding of the *distribution*, *abundance*, and *movements of organisms across wide geographic areas*, with *high spatial resolution* and over long periods of time.





nnot currently be displayed.



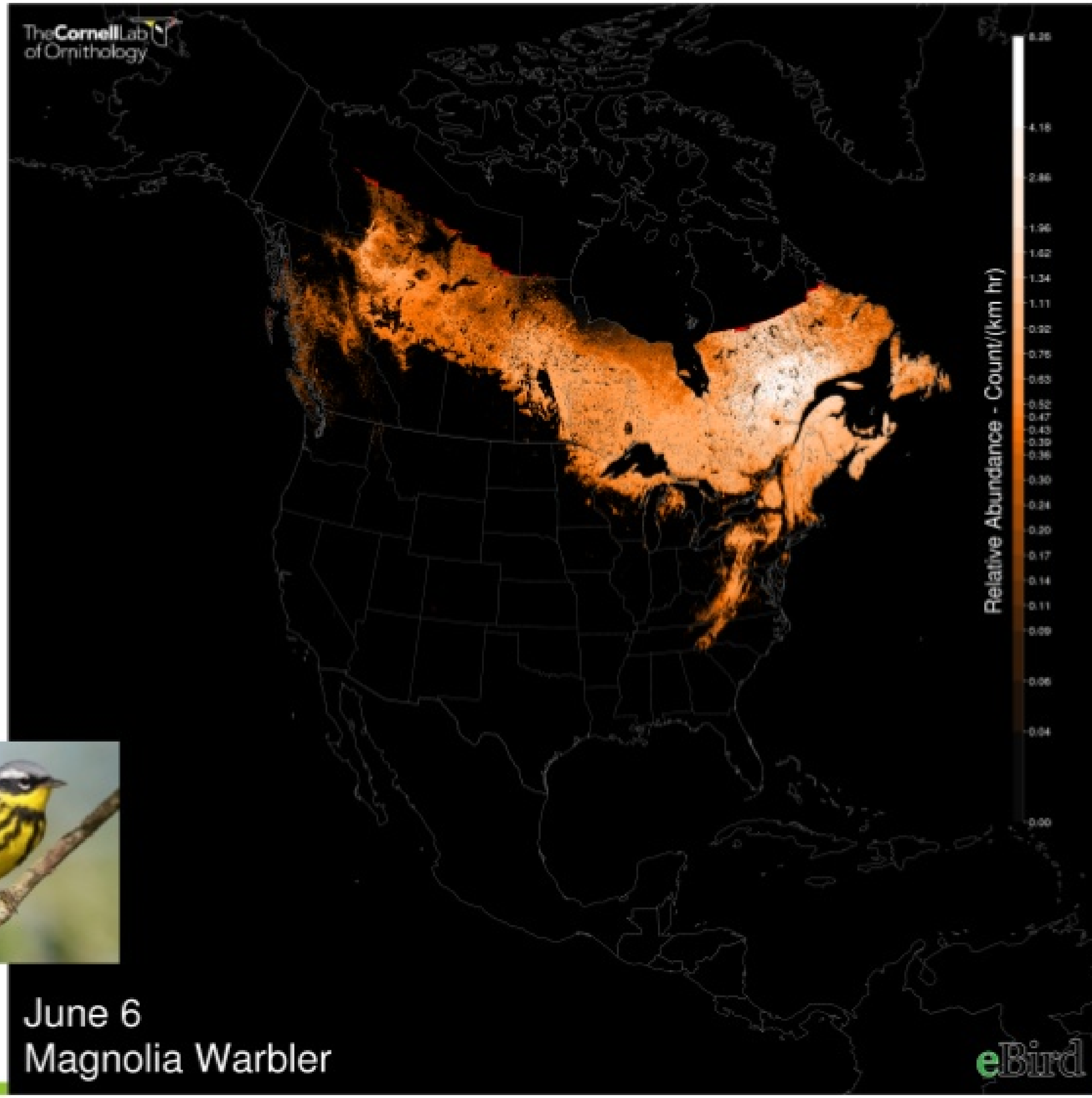


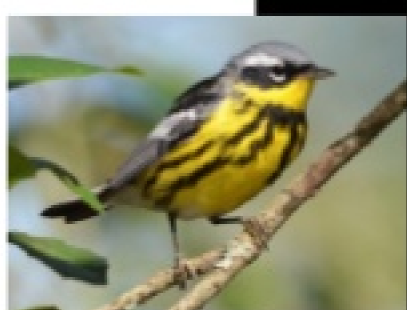
BirdReturns: Dynamic Land Conservation

ebird.org/birdvis

SPARK SUMMIT 2017

Technical Experience

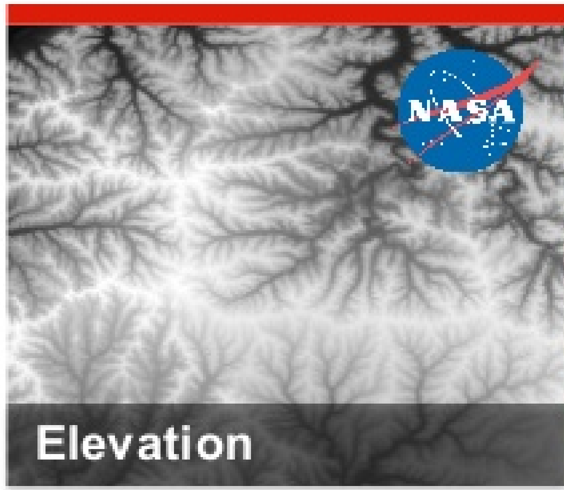


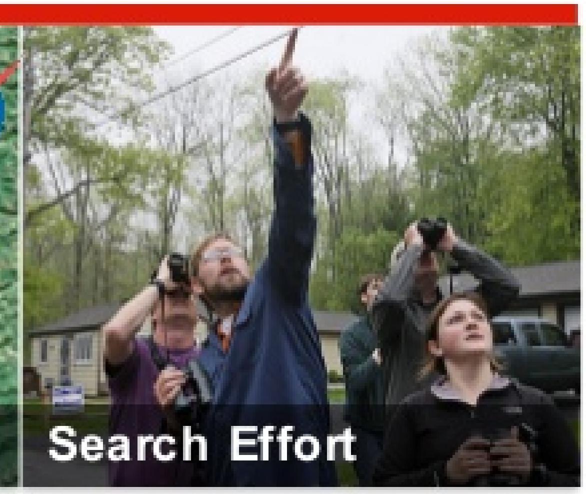


SPARK SUMMIT 2017

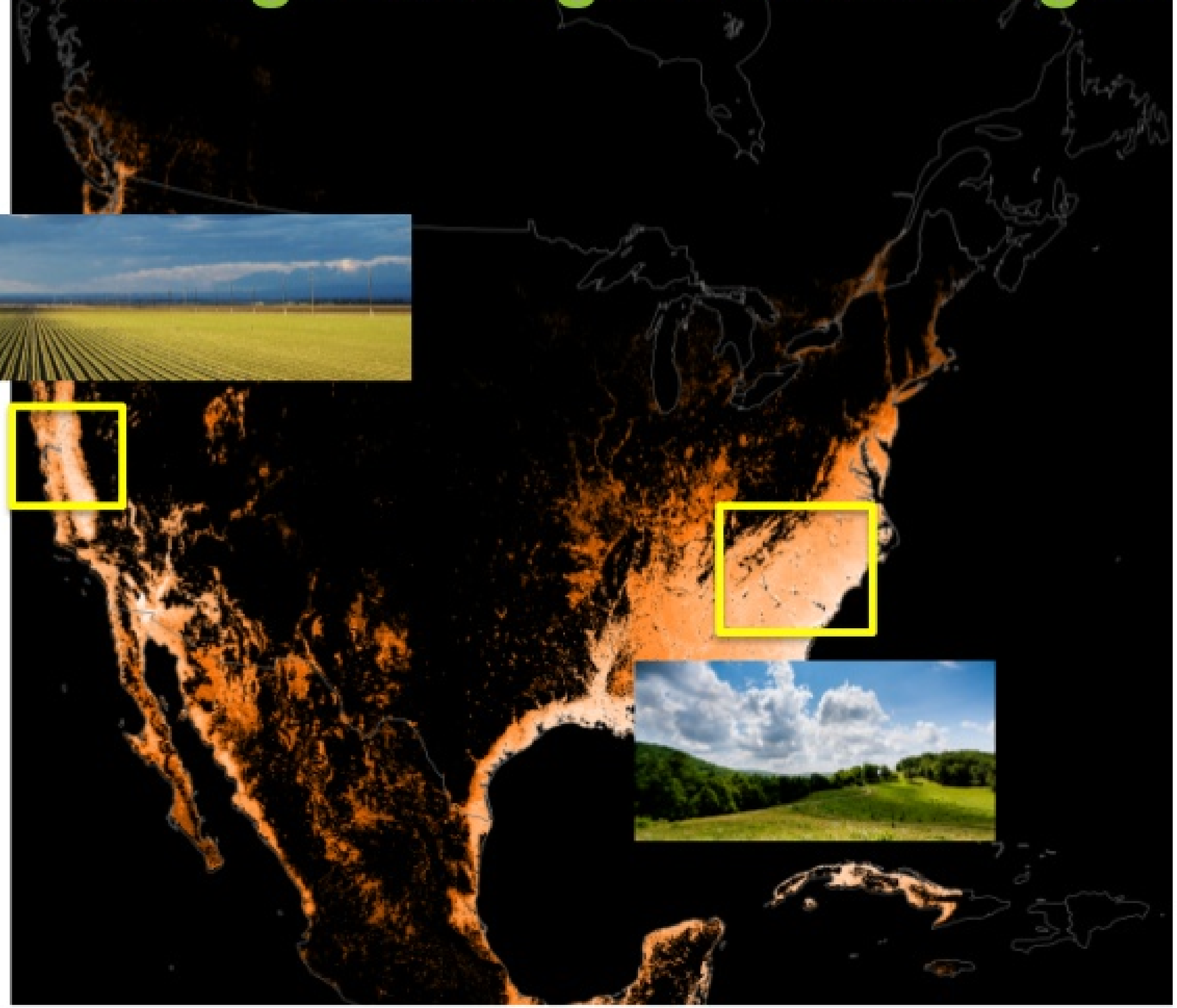
Linking Populations and Environment

Machine Learning





Scaling: Ecological Challenges



Tree Swallow, March

SpatioTemporal Exploratory Model (STEM)

1. Divide

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

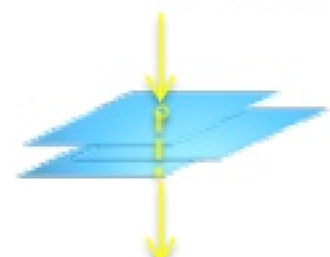
100 randomized replicates

2. Recombine

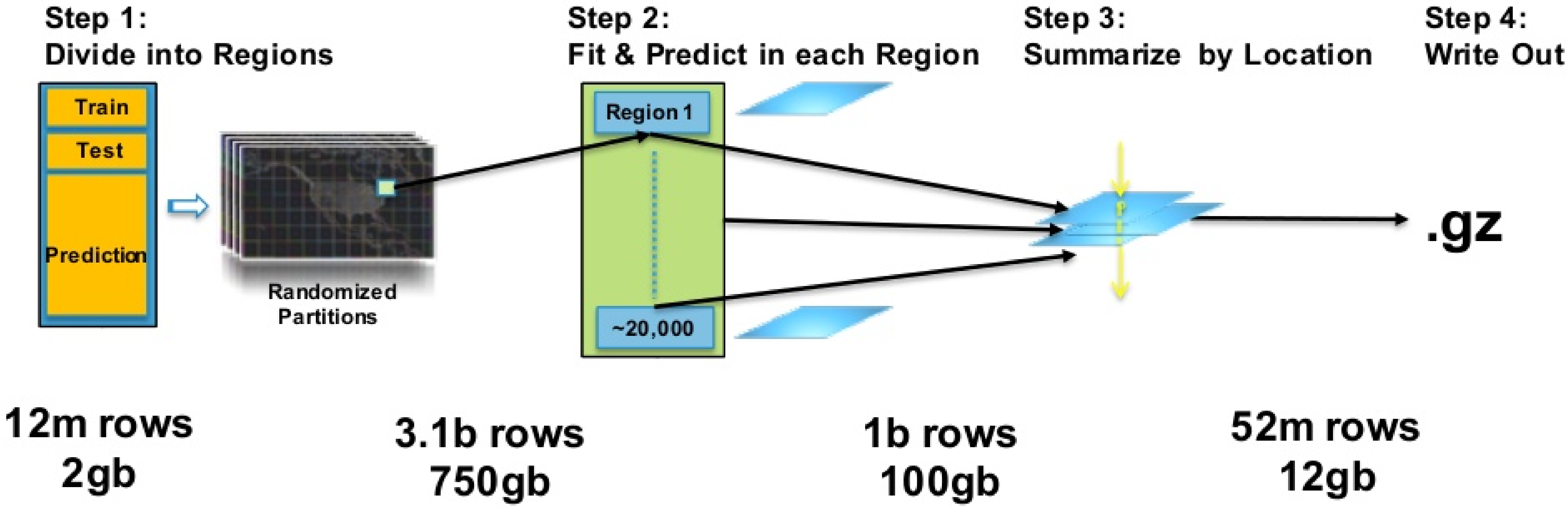
- Average predictions across all models for each location within regions

Region

Location



SPARK SUMMIT 2017


Step 1:
Divide into Regions
Train
Test
Prediction
Randomized
Partitions
Step 2:
Fit & Predict in each Region
Region 1
~20,000
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

SPARK SUMMIT 2017

Future Modeling

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

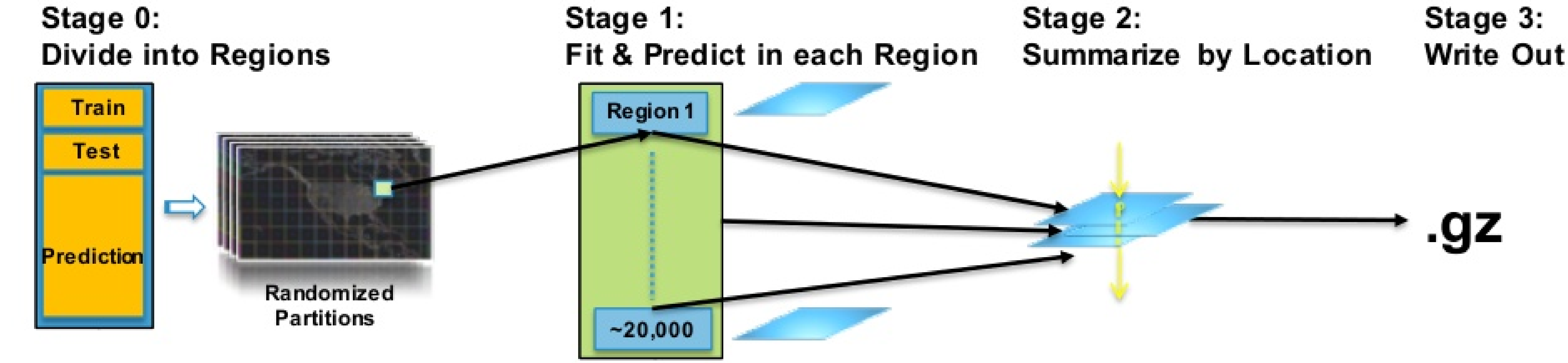
What have we tried?

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

SPARK SUMMIT 2017

RDD pipe()

SPARK SUMMIT 2017



12m rows 2gb	3.1b rows 750gb	1b rows 100gb	52m rows 12gb

```
.map(lambda p: (p[0].split("\t")[0], p[0].split("\t")[1]+"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)

LEON LEVY FOUNDATION

Microsoft Azure

SPARK SUMMIT 2017