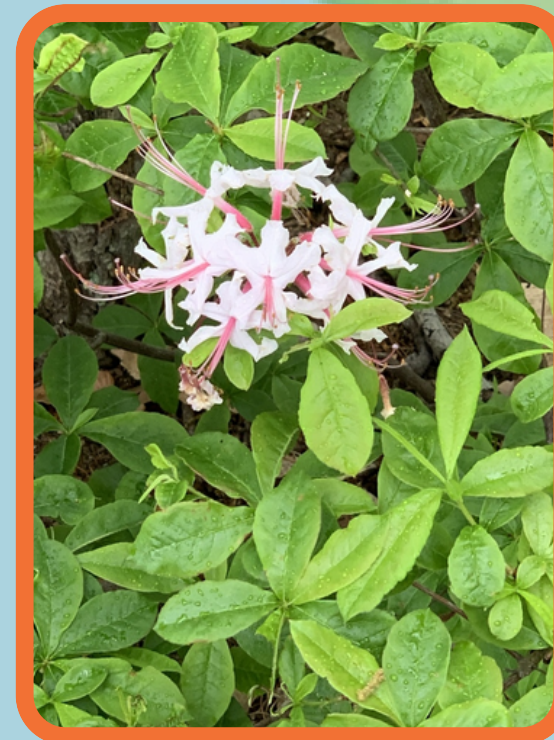


# HIKING WITH INATURALIST

*Determining User Behavior  
with Crowd-Sourced Data  
for Environmental Science*

LISA MCCOOL-GRIME





# AGENDA

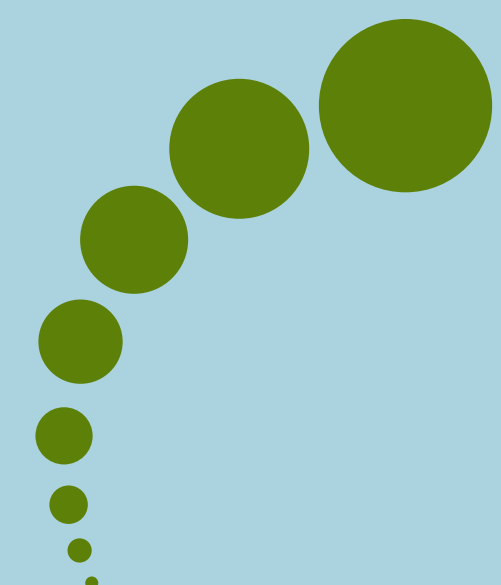
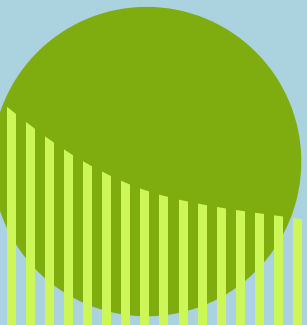
**DATA**

**PROBLEM**

**IMPACT**

**SOLUTION**

**NEXT STEPS**



# INATURALIST DATASET

SOCIAL NETWORK SHARING  
BIODIVERSITY INFORMATION

OVER 100 MILLION NATURAL  
OBSERVATIONS SINCE 2015

MY SUBSET: 1,050,151 U.S. OBSERVATIONS  
FROM 300 USERS

Crested Coralroot Orchid (*Bletia spicata*) VU Research Grade Follow

Observed: July 2022 Submitted: July 2022

Map Satellite

Virginia, US (Obscured)

Details

abelkinser and annegp faved this observation

Notes

Group of 5 flower spikes (tallest is about 2 feet))  
Flora of Virginia lists this as Hexalectris spicata  
In King William County, VA

Activity

annegp suggested an ID Improving Jul '22

Crested Coralroot Orchid Compare Agree

catullus suggested an ID Jul '22

Crested Coralroot Orchid

Community Taxon

Crested Coralroot Orchid (*Bletia spicata*) VU

Cumulative IDs: 4 of 4

0 2/3rds 4

Agree Compare About

Annotations

Attribute	Value	Agree	Disagree
Plant Phenology	Select		
Sex	Select		

Encompassing Places

Standard:

- North America (Continent)
- United States (Country)
- Virginia, US (State)
- King William County, US, VA (County)

Community Curated:

- Rolling Coastal Plain (Region)
- Atlantic Southern Loam Hills, US (Region)
- Mid-Atlantic States, US (Region)
- Chesapeake Bay watershed (Drainage)
- BIOL272 project area (Unknown)
- United States Atlantic Coastal Plain (Land feature)
- US Eastern States, US (Colloquial)
- Eastern United States (Region)
- Newcomb's Wildflower Guide Range (Region)
- Southeast US conifer savannas, US (Zone)

[More](#)

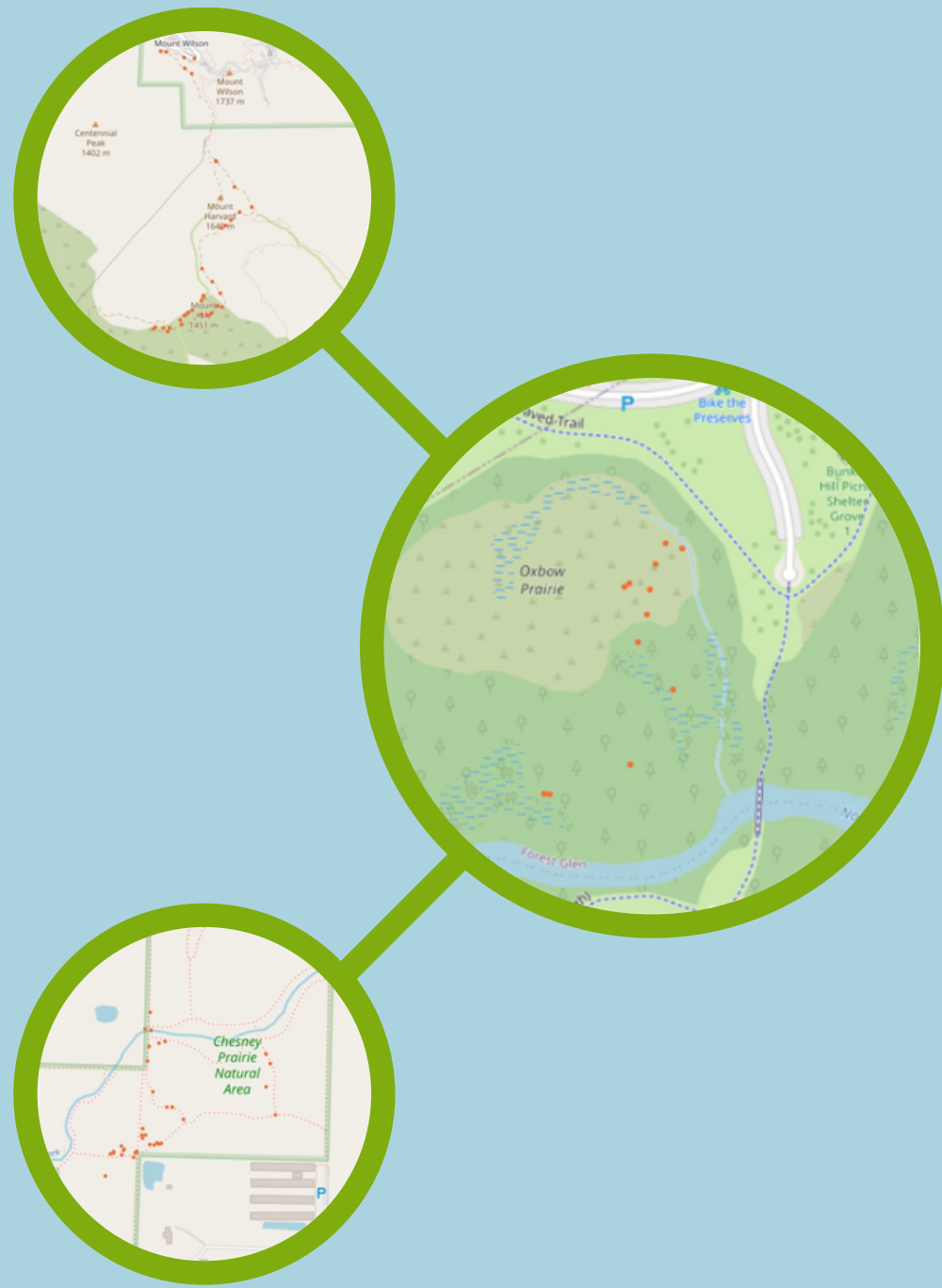
Why the Coordinates Are Obscured

- Geoprivacy is obscured: Observer has chosen to obscure the coordinates.
- Taxon is threatened, coordinates obscured by default: One of the taxa suggested in the identifications, or one of the taxa that contain any of these taxa, is known to be rare and/or threatened, so the location of this observation has been obscured.



MISSION:  
to connect people to nature and  
advance biodiversity science and conservation

# PROBLEM: HOW DOES USER BEHAVIOR OVERLAP WITH HIKING BEHAVIOR?



Can iNaturalist data be used to find common nature trails?

What metrics can describe users who commonly use nature trails?  
What metrics describe other ways of engaging with iNaturalist?

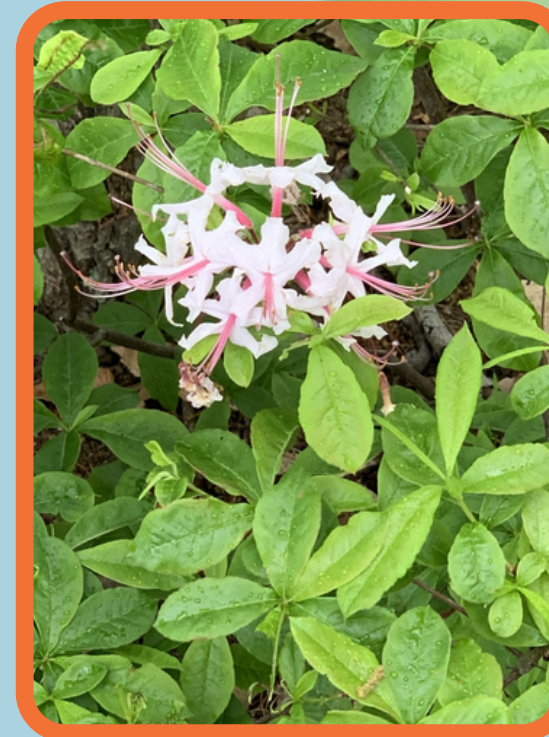


# IMPACT

## USER ENGAGEMENT/EXPANSION

Add slideshow of individual user hikes  
to the personalized Year-in-Review

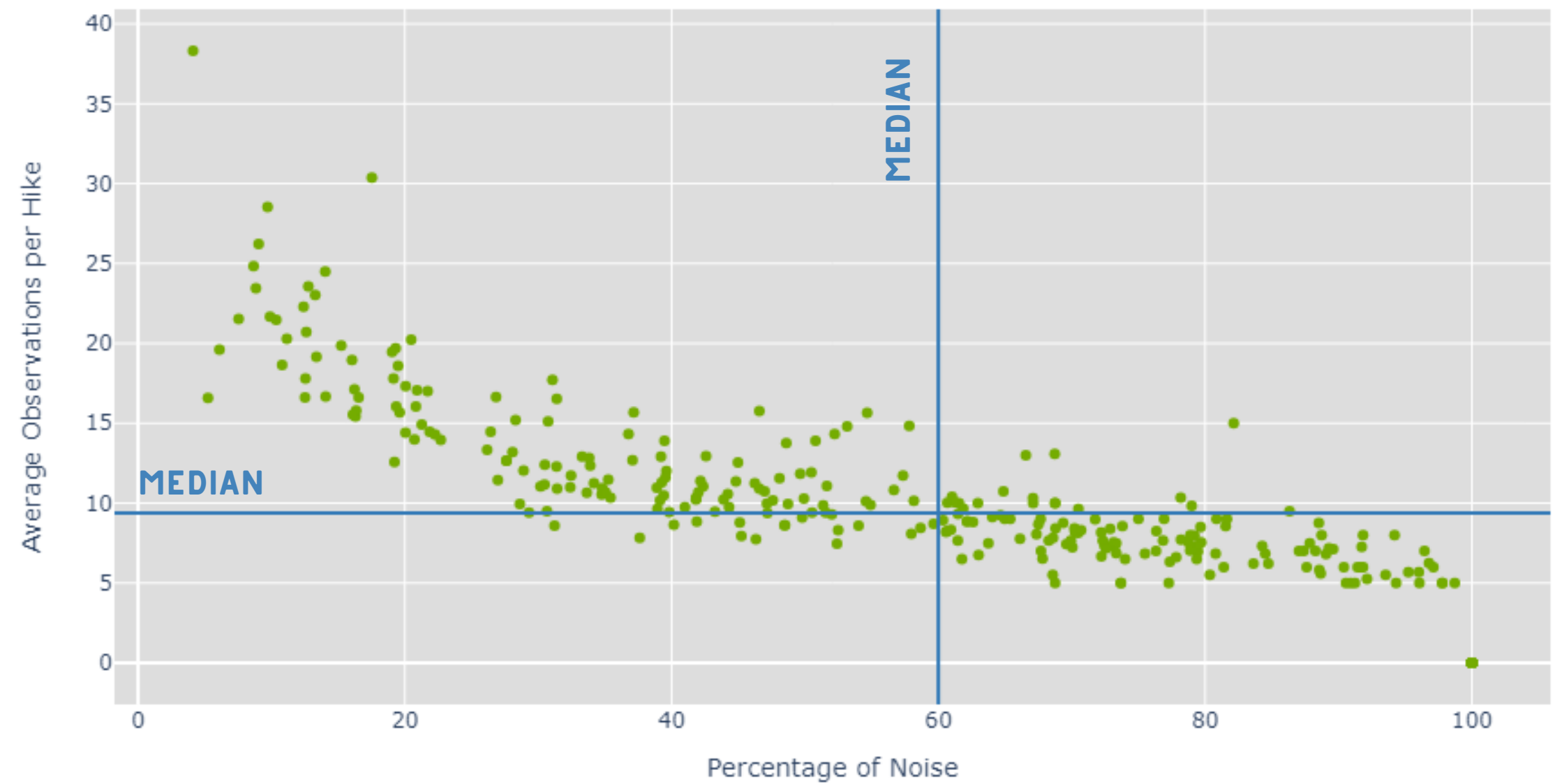
Create interactive “scavenger hunt” in partnership  
with State or Local Parks with trails



# IMPACT

## SCIENTIFIC STUDY: METRICS THAT DESCRIBE USER BEHAVIOR

As "One-Off" Behavior Increases, Observations per Hike Decrease



# SOLUTION:

## CLUSTERING WITH DBSCAN

(SLIDE ONLY FOR FRIDAY FOR MY DATA SCIENCE PEEPS)



**Irregularly shaped clusters (hikes)  
with noise (one-off observations)**



**Ran DBSCAN user by user**



**Feature engineered distance and time  
columns in several ways**



**Optimization: Feature selection,  
epsilon (radius of neighborhood around a  
point) and minimum number of samples**



**MinMax Scaler  
(helps give limited range for epsilon)**



**Optimization: Performed EDA on meta  
dataset of silhouette scores for various  
features and hyperparameters**

# VALIDATION TIPS FOR UNSUPERVISED LEARNING (AGAIN, ONLY FRIDAY)



## INTERNAL VALIDATION

Silhouette Scores

Silhouette scores are **RELATIVE** to features.

To compare between feature selection, I computed silhouette scores relative to their own features **AND** relative to the larger set of all numerical features.



## EXTERNAL VALIDATION

Non Machine Learning Model

Questions to ask to help with external validation:

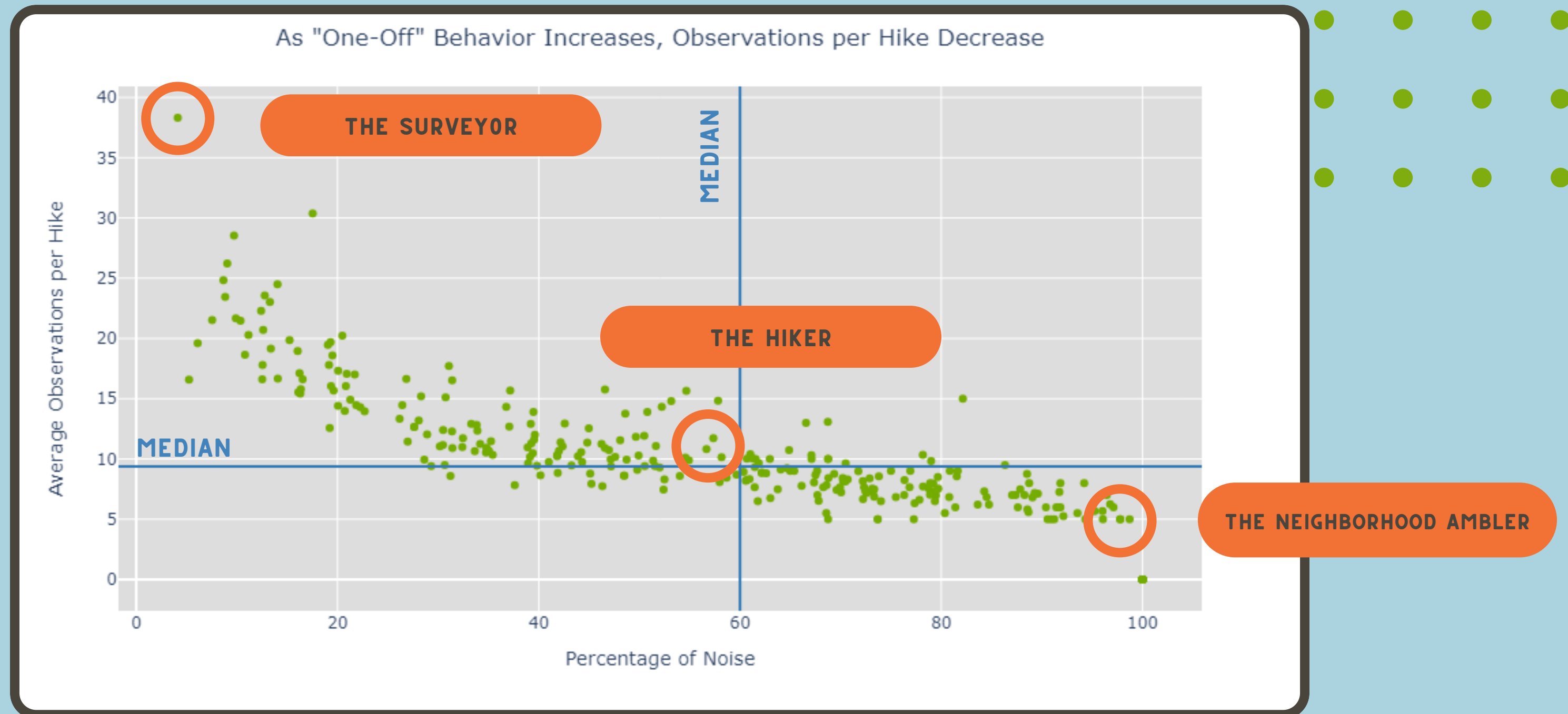
- Is there someone with subject expertise to validate the emerging clusters?
- How would humans have approached this problem **BEFORE** machine learning? Does your model improve on this approach?

My external validation solution: group by dates, choose a minimum sample size--call those “hikes”.  
This provided a **BASELINE MODEL** to compare to my DBSCAN models, allowing me to find a **GOLDILOCKS** version--  
not too strict, not too loose.

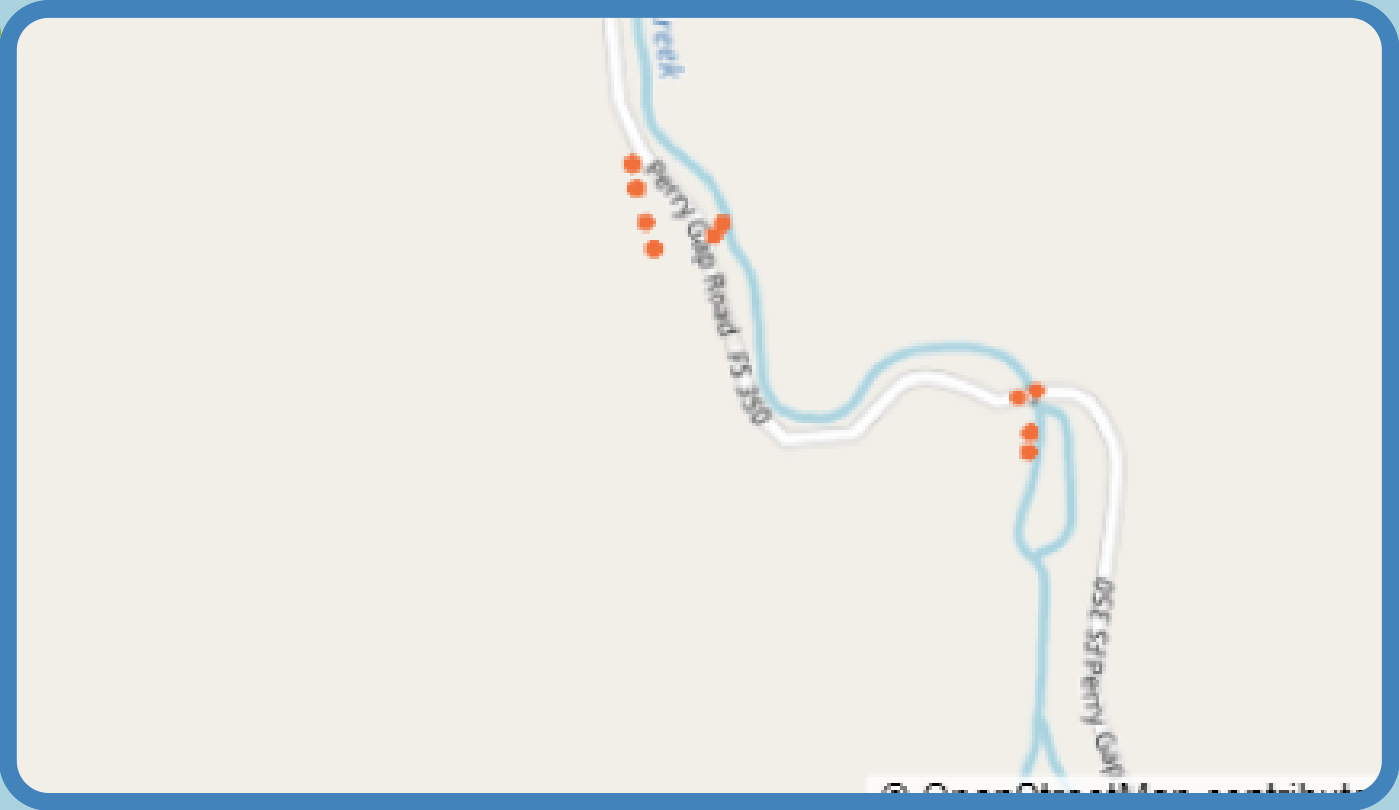
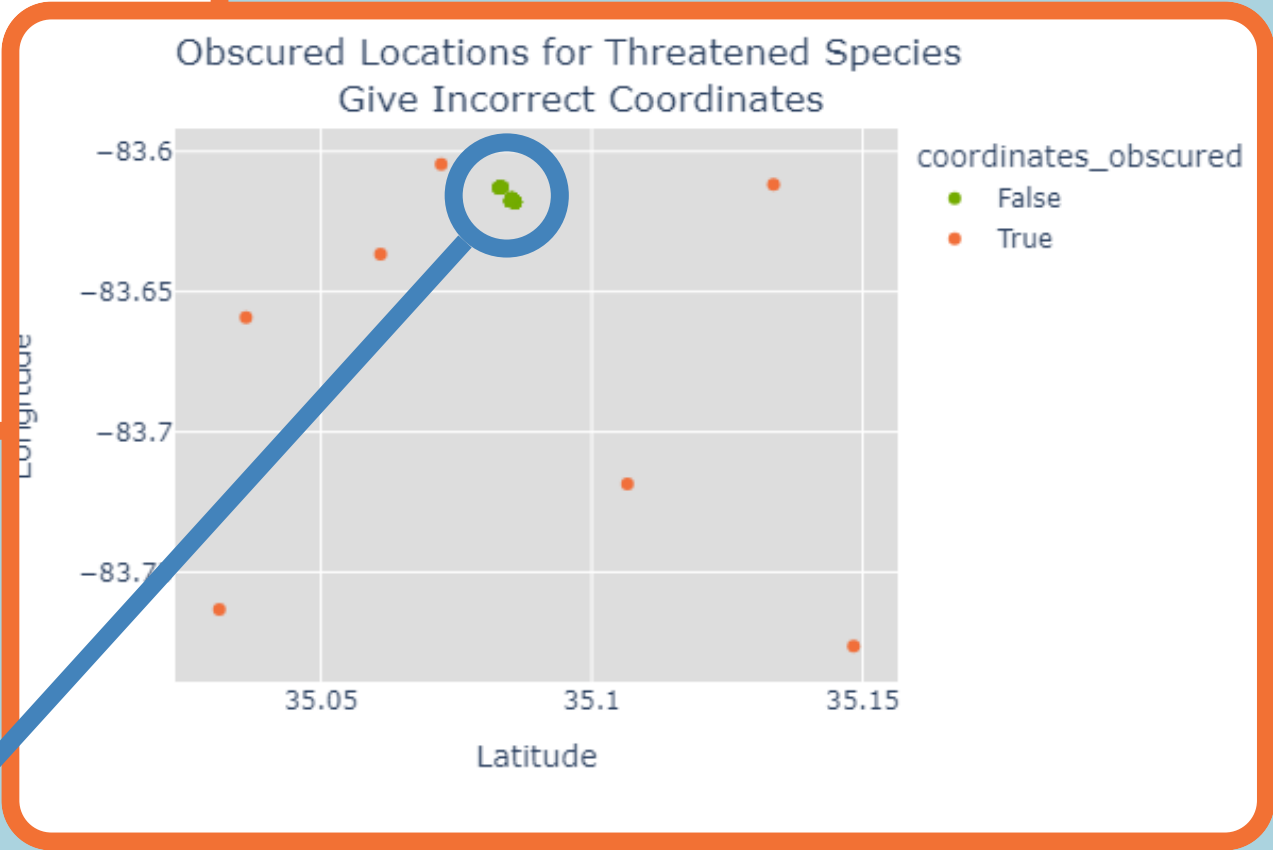
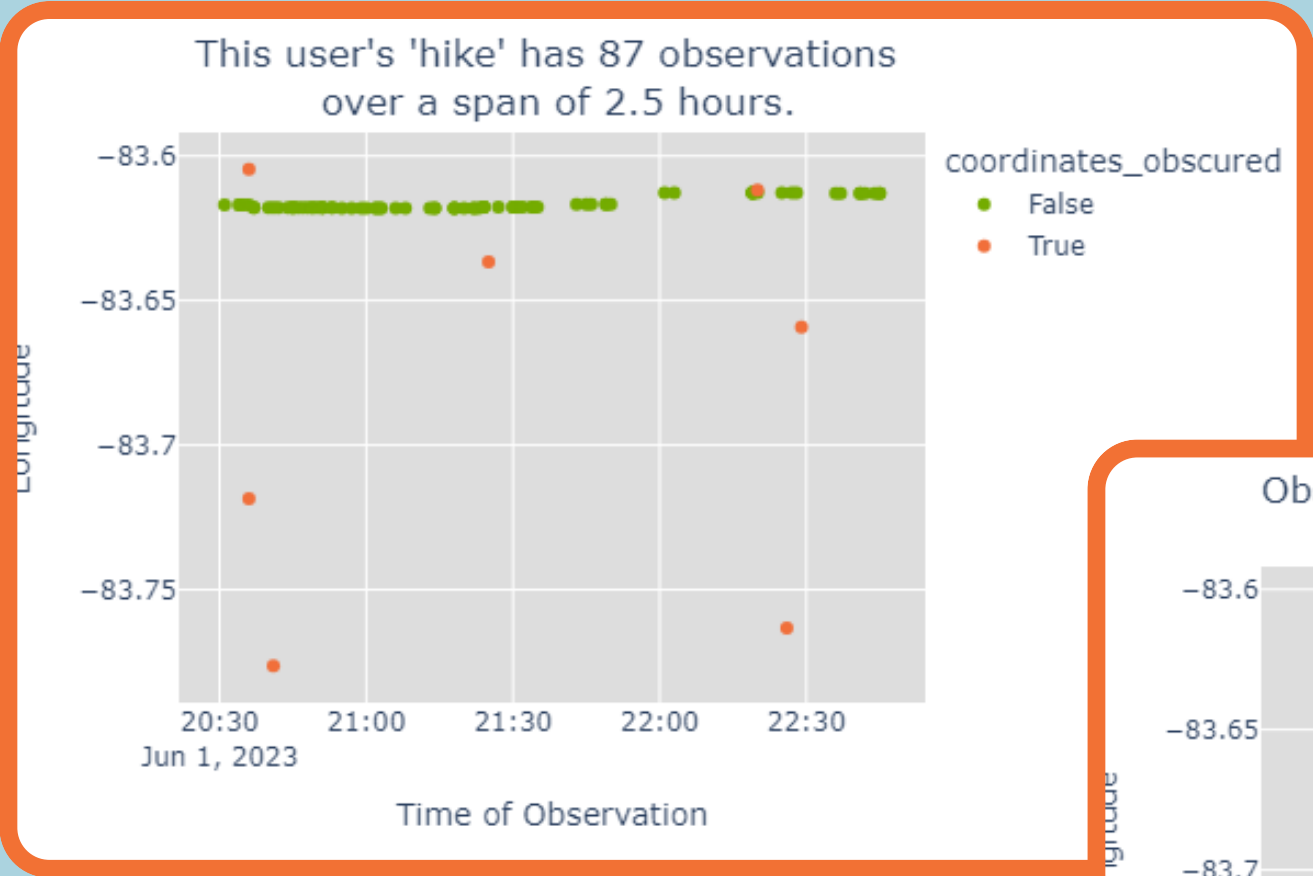
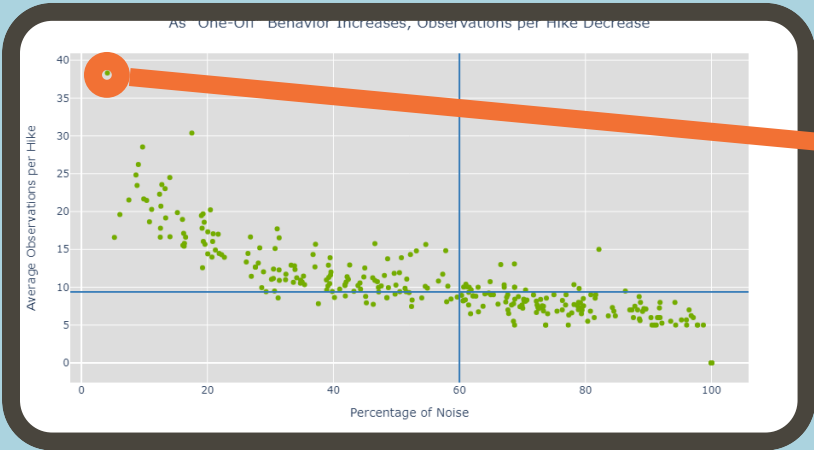


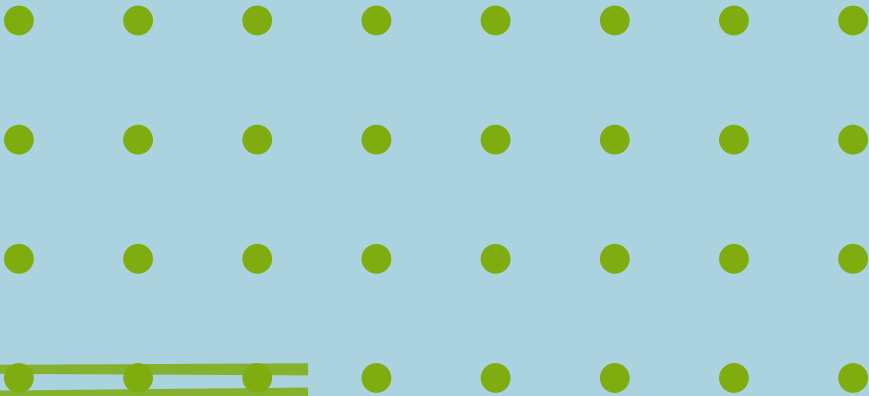
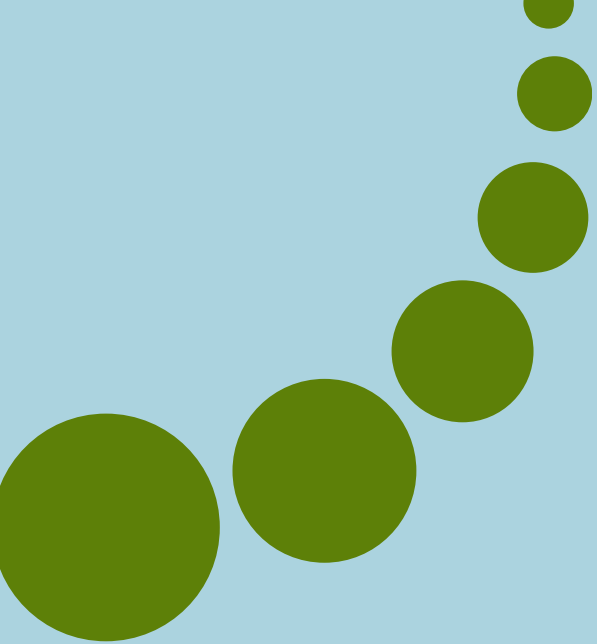
# NEXT STEPS

## TARGETED USER ENGAGEMENT



# THE SURVEYOR





# THANK YOU

