HIKING WITH INATURALIST

Determining User Behavior with Crowd-Sourced Data for Environmental Science

LISA MCCOOL-GRIME

AGENDA

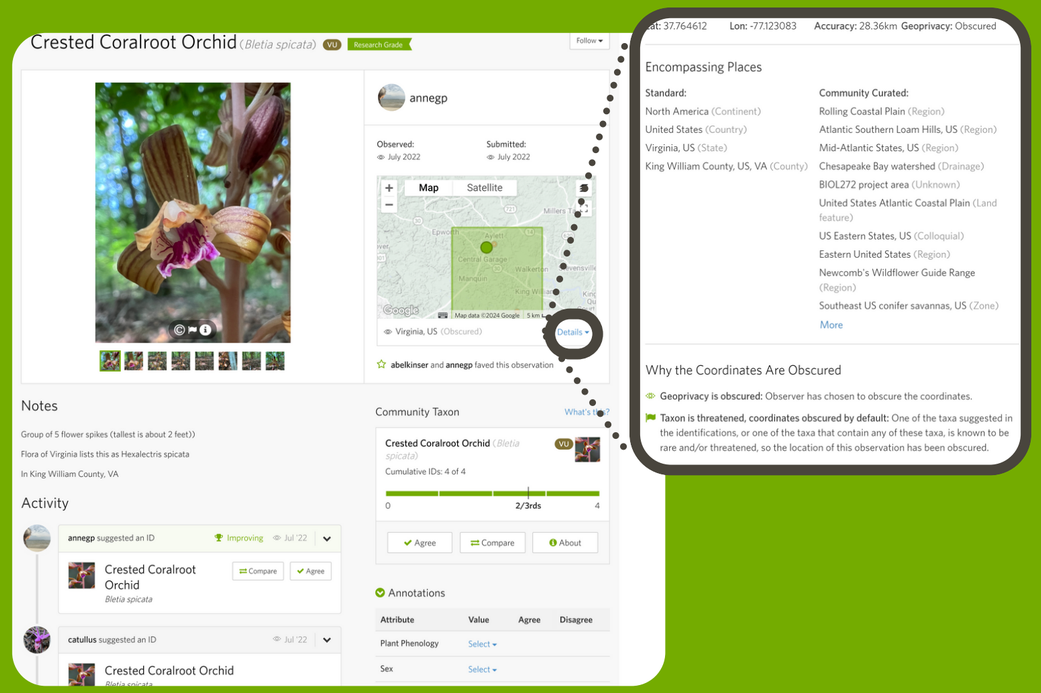
- DATA
- PROBLEM
- IMPACT
- SOLUTION
- NEXT STEPS

INATURALIST DATASET

SOCIAL NETWORK SHARING BIODIVERSITY INFORMATION

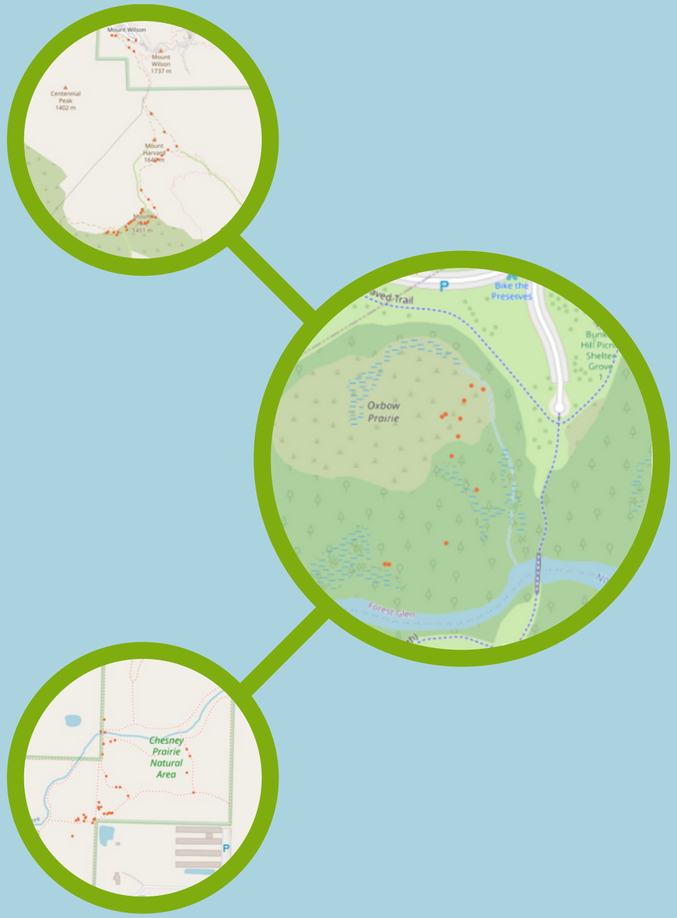
OVER 100 MILLION NATURAL OBSERVATIONS SINCE 2015

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



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?

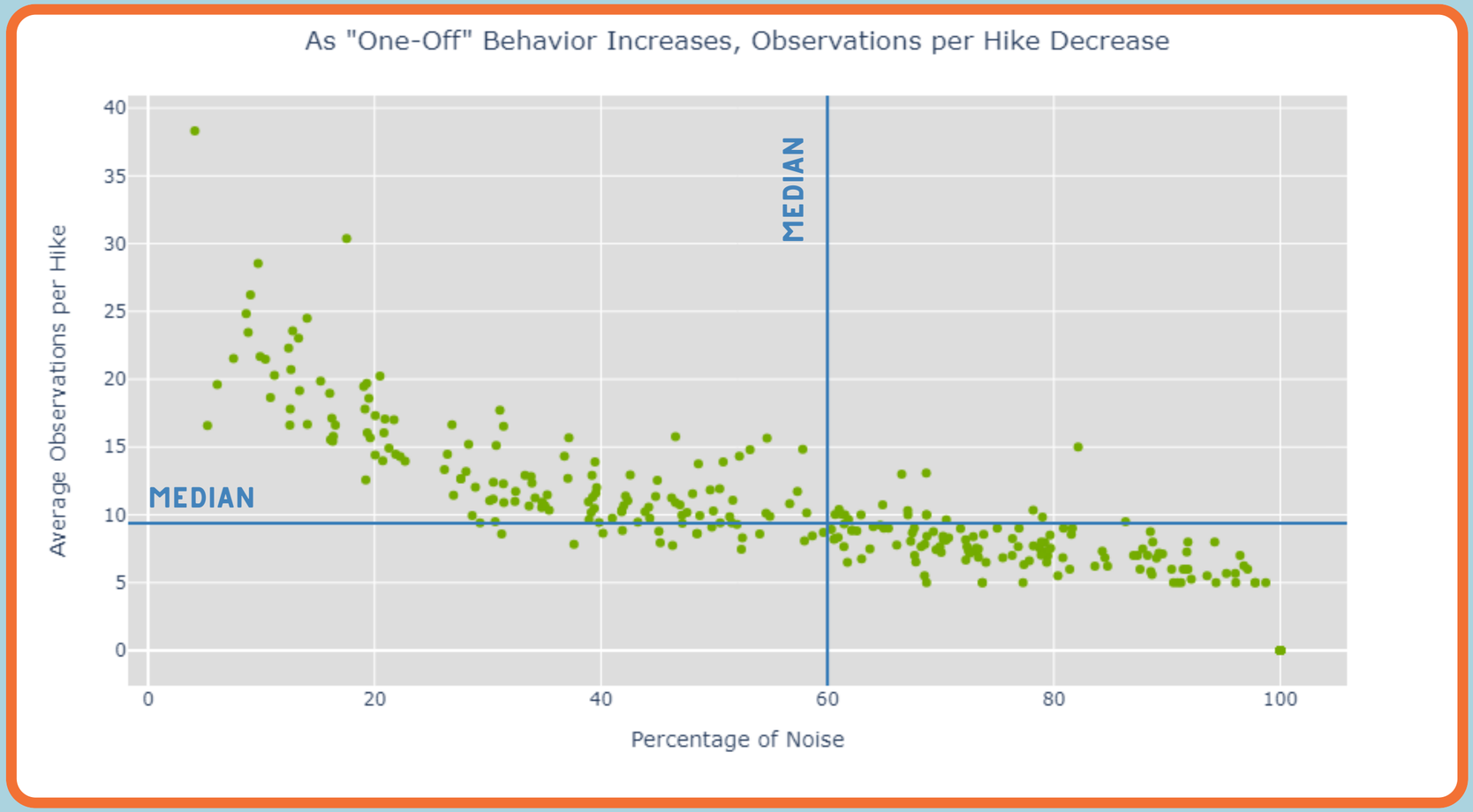
IMPACT
USER ENGAGEMENT/EXPANSION

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

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

IMPACT

SCIENTIFIC STUDY:
METRICS THAT DESCRIBE USER BEHAVIOR



SOLUTION:
CLUSTERING WITH DBSCAN
(SLIDE ONLY FOR FRIDAY FOR MY DATA SCIENCE PEEPS)

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

Feature engineered distance and time columns in several ways

MinMax Scaler
(helps give limited range for epsilon)

Ran DBSCAN user by user

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

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

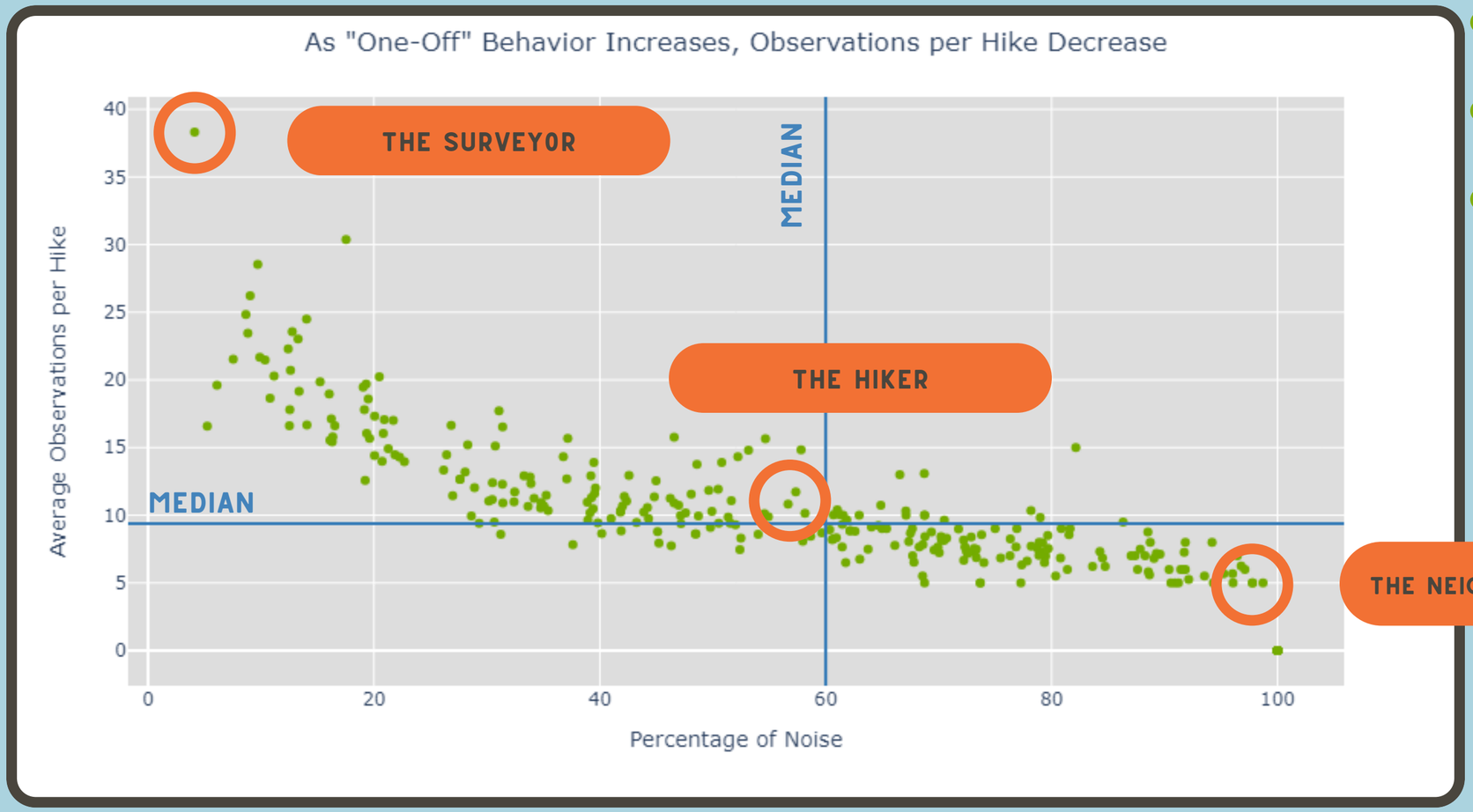
Questions to ask to help with external validation:

- 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?

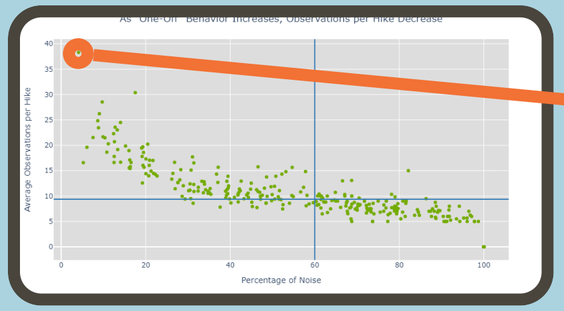
My external validation solution: group by dates, choose a minimum sample size--call those “hikes”.
This provided a BASELINE MODEL to compare to my DBSCAN models, allowing me to find a GOLDILOCKS version--
not too strict, not too loose.

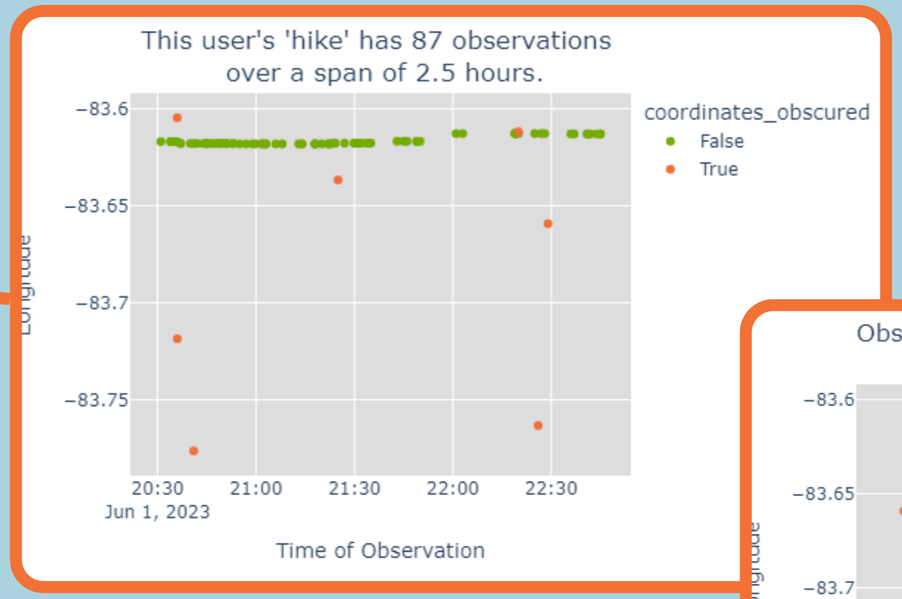
NEXT STEPS

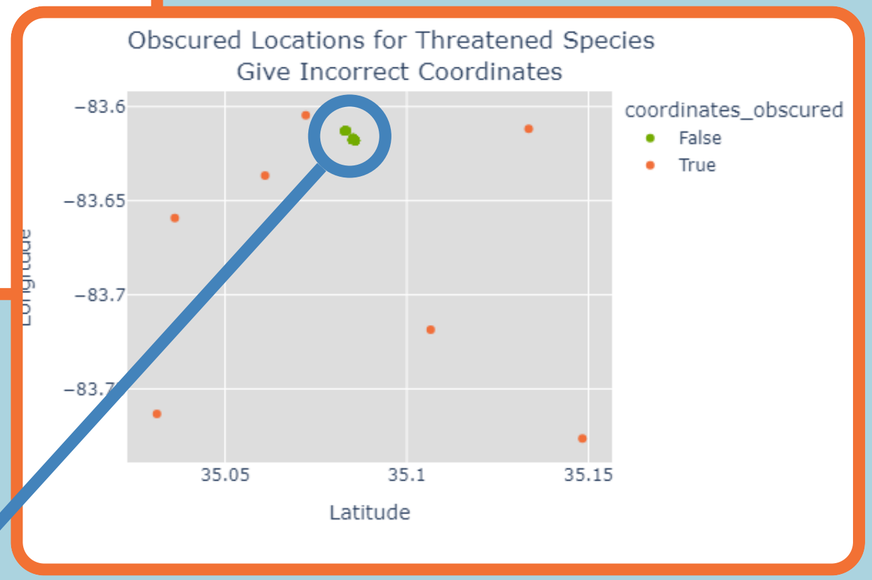
TARGETED USER ENGAGEMENT

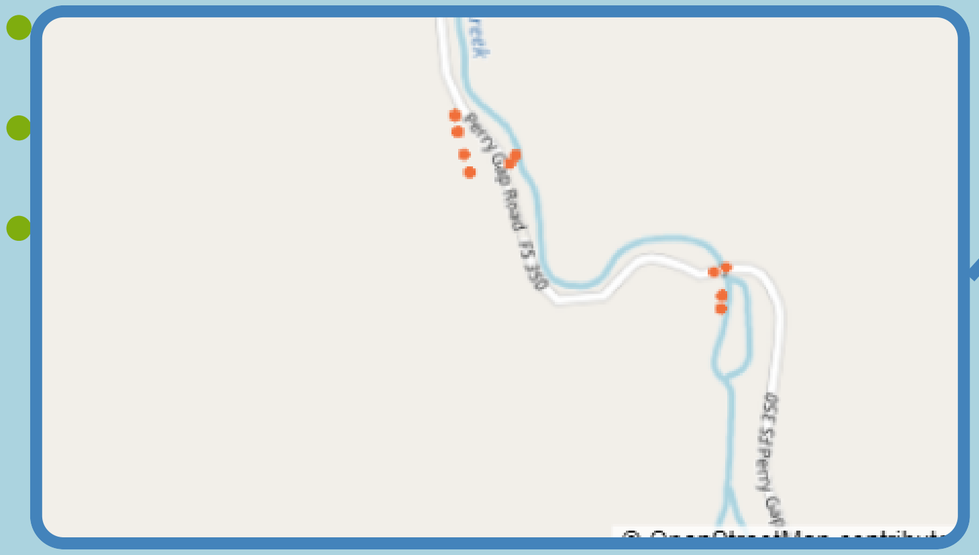


THE SURVEYOR









THANK YOU