



BIO
DIVERSITY
NEXT



CONABIO

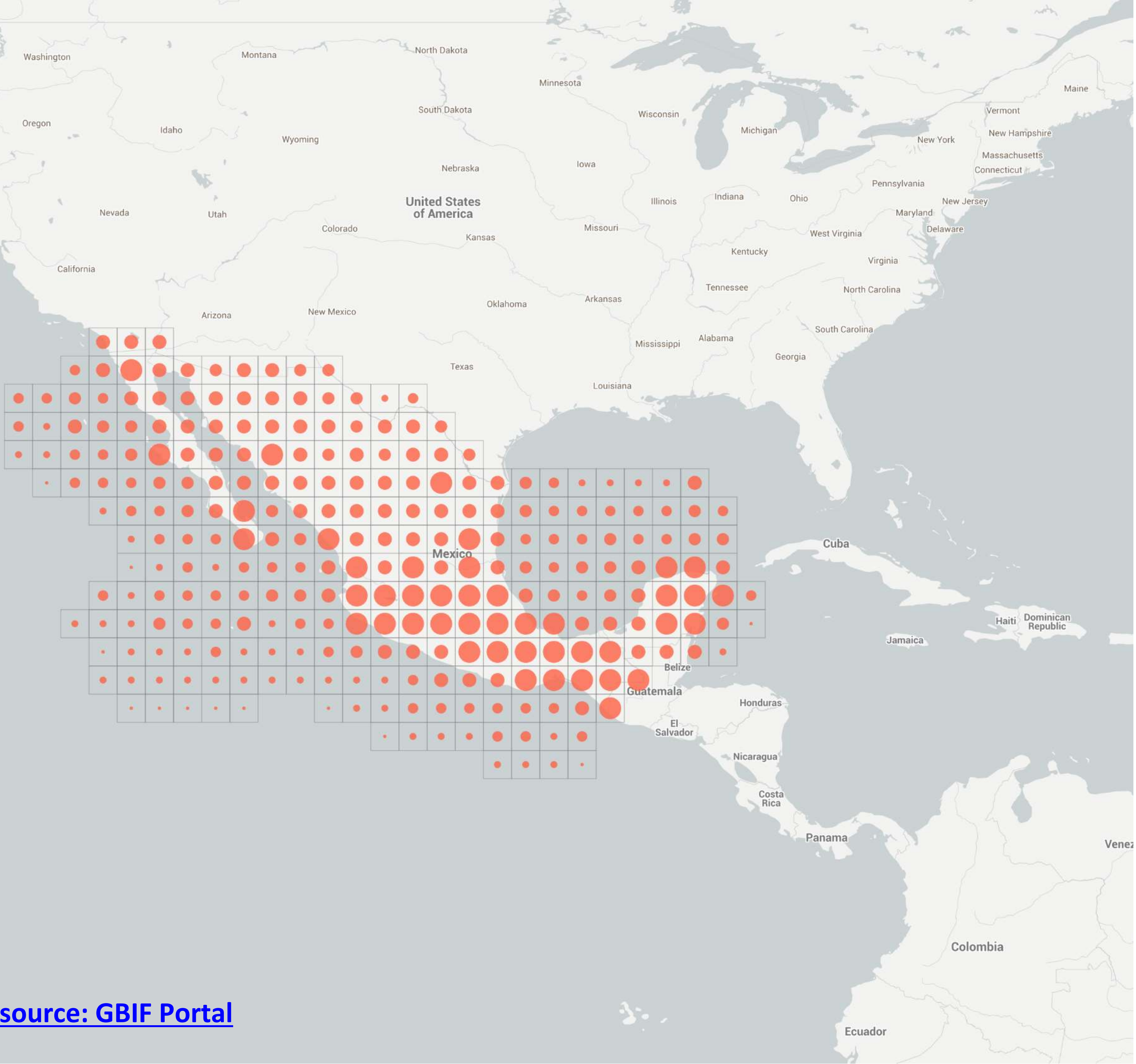
COMISIÓN NACIONAL PARA
EL CONOCIMIENTO Y USO
DE LA BIODIVERSIDAD

Automatizing the detection of erroneous species occurrence records

Presents: Raúl Sierra-Alcocer | National Commission for the
Knowledge and use of Biodiversity, Mexico

Authors: Raúl Jiménez Rosenberg, Raúl Sierra-Alcocer

**Species occurrence records,
National Biodiversity Information System,
CONABIO**



Species in Mexico	
Region	Mexico
Records	~ 14,000,000
Species	~ 70,000

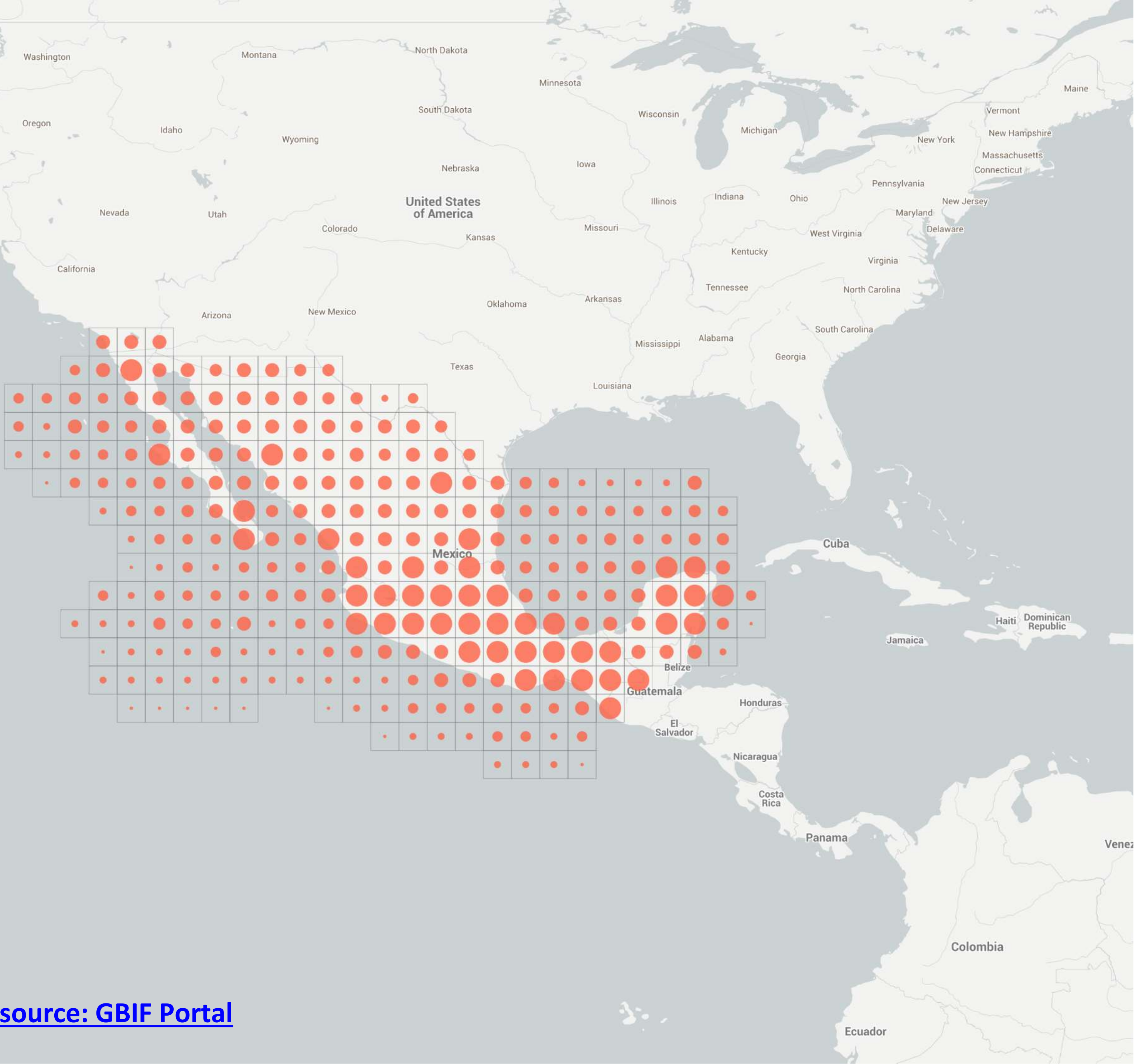
[Image source: GBIF Portal](#)

Validation at CONABIO

- Taxonomy : 70% validated at 100%, representing 90% of the records
- Geography: 99.84% at country level, 90.47% at state, 22.63% at county level
- Organism environment: Not processed (35.11%); **Not valid (0.48%)**; Valid (64.40%)

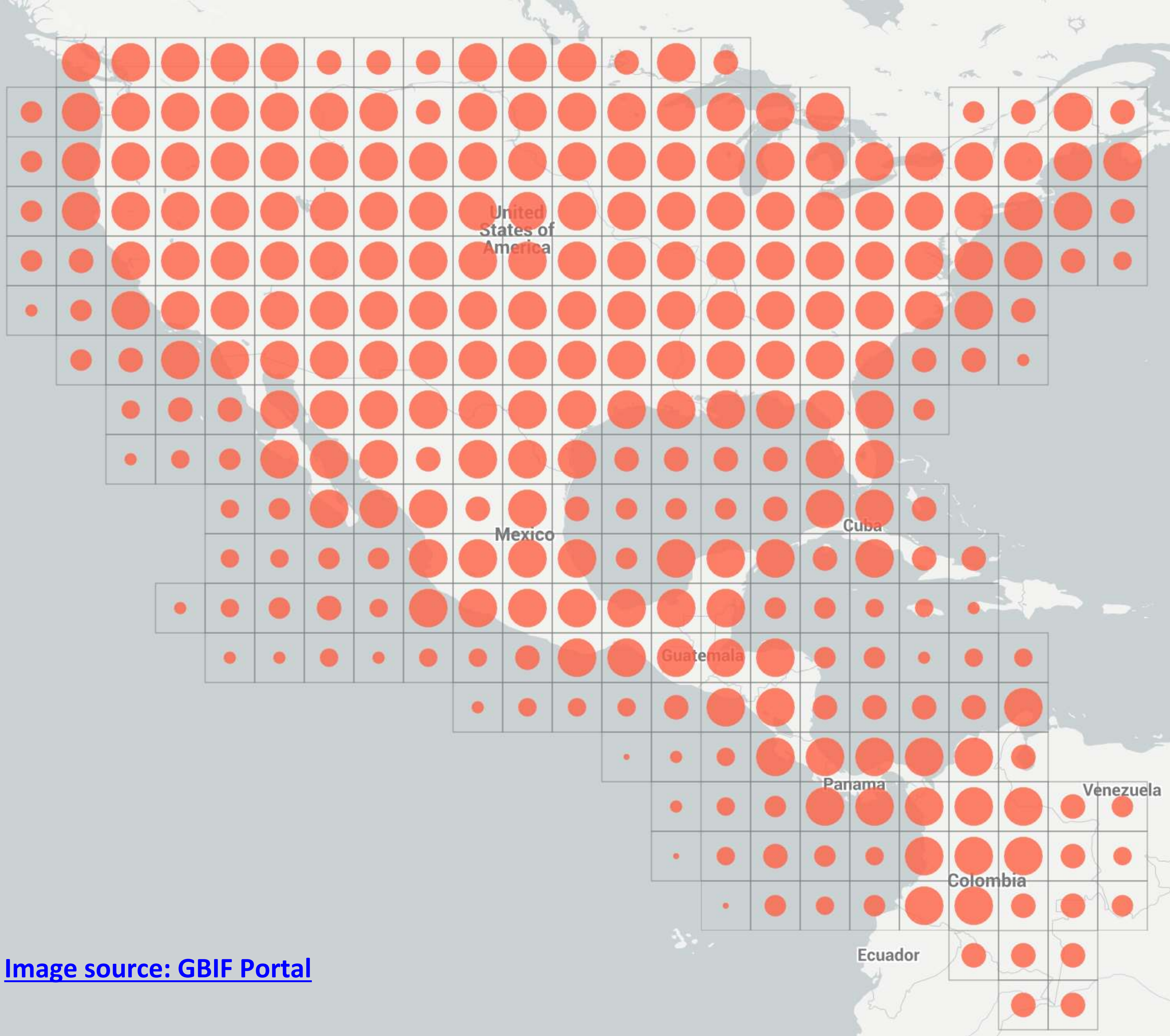
**This project's
focus**

**Species occurrence records,
National Biodiversity Information System,
CONABIO**



Species in Mexico	
Region	Mexico
Records	~ 14,000,000
Species	~ 70,000

[Image source: GBIF Portal](#)



National Biodiversity Information System without borders

	Species that share genus with species in Mexico
Region	From Colombia to the USA
Records	~ 300,000,000 (Main source: GBIF)
Species	?

Previous QA workflow does not scale



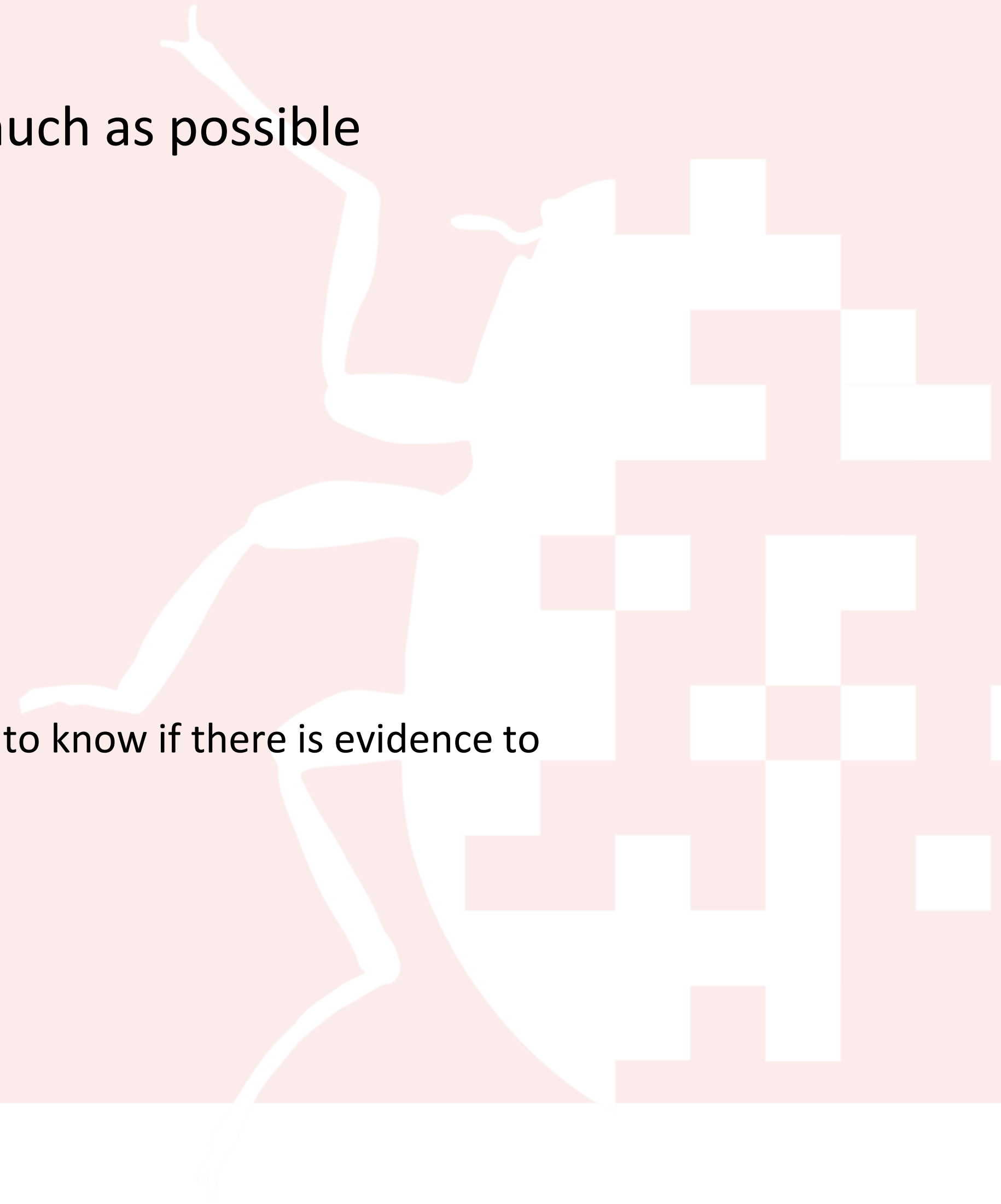
Line of work

- Our databases have lots of information, use as much as possible
- Keep it simple

Elements

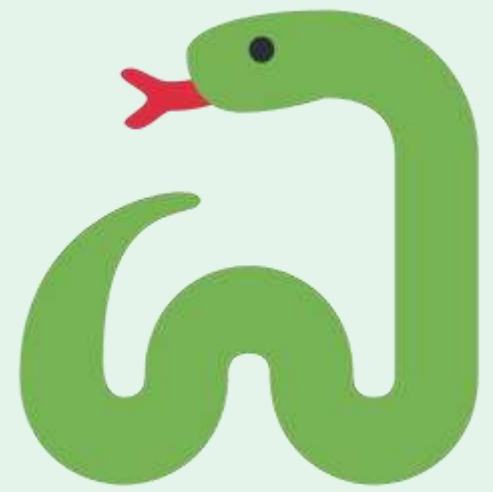
- Climatic variables
- Other species records

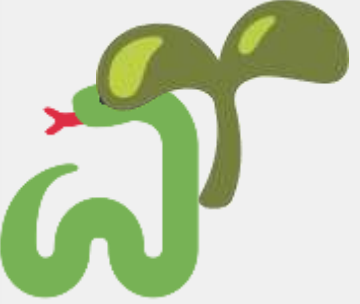
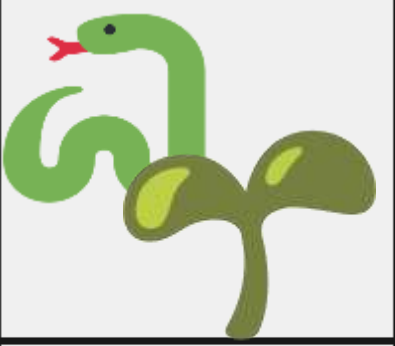


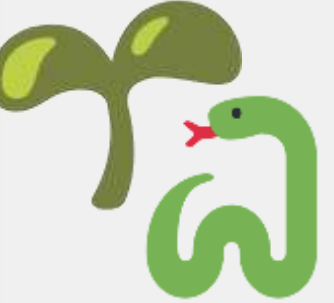
Note: We are not trying to model the species distribution, we want to know if there is evidence to believe the record



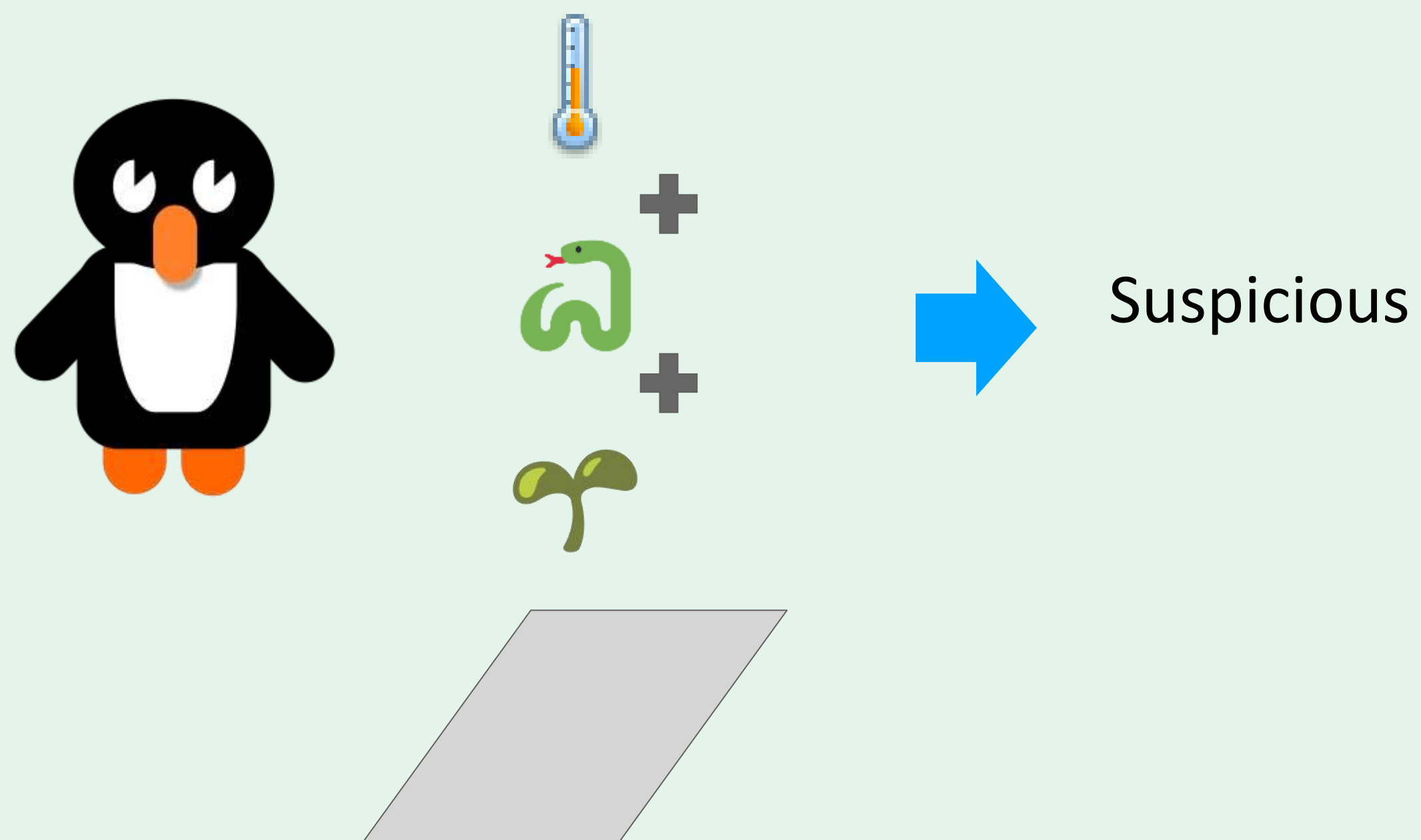
Method Basic Idea

For each species S , calculate score based on co-occurrences:



Method Basic Idea



Fauna score $\log \frac{P(\text{penguin} = 1 | \text{snake}, \dots, Y_k)}{P(\text{penguin} = 0 | \text{snake}, \dots, Y_k)}$

Climate score $\log \frac{P(\text{penguin} = 1 | \text{thermometer}, \dots, Y_k)}{P(\text{penguin} = 0 | \text{thermometer}, \dots, Y_k)}$

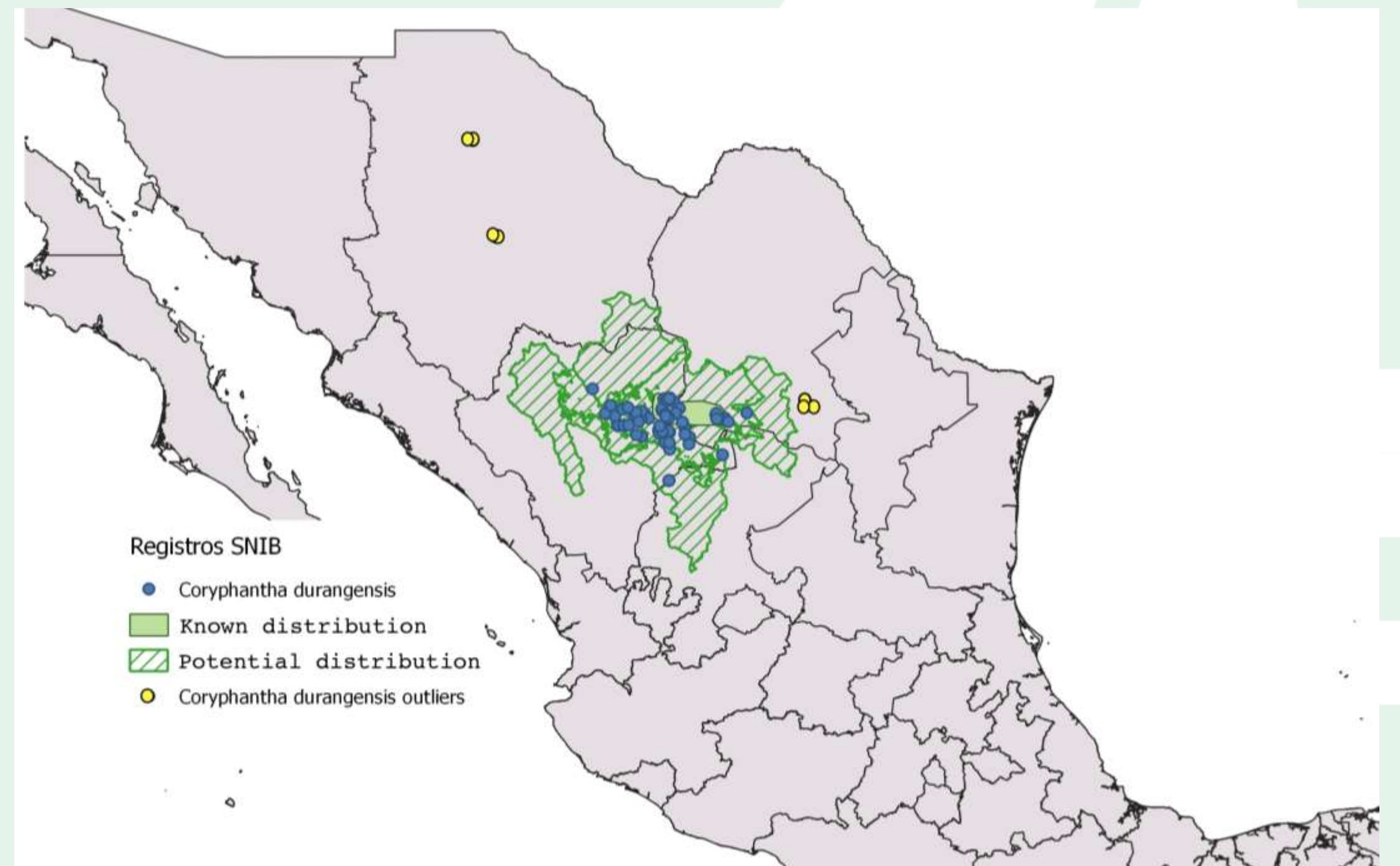
Plants score $\log \frac{P(\text{penguin} = 1 | \text{plant}, \dots, Y_k)}{P(\text{penguin} = 0 | \text{plant}, \dots, Y_k)}$

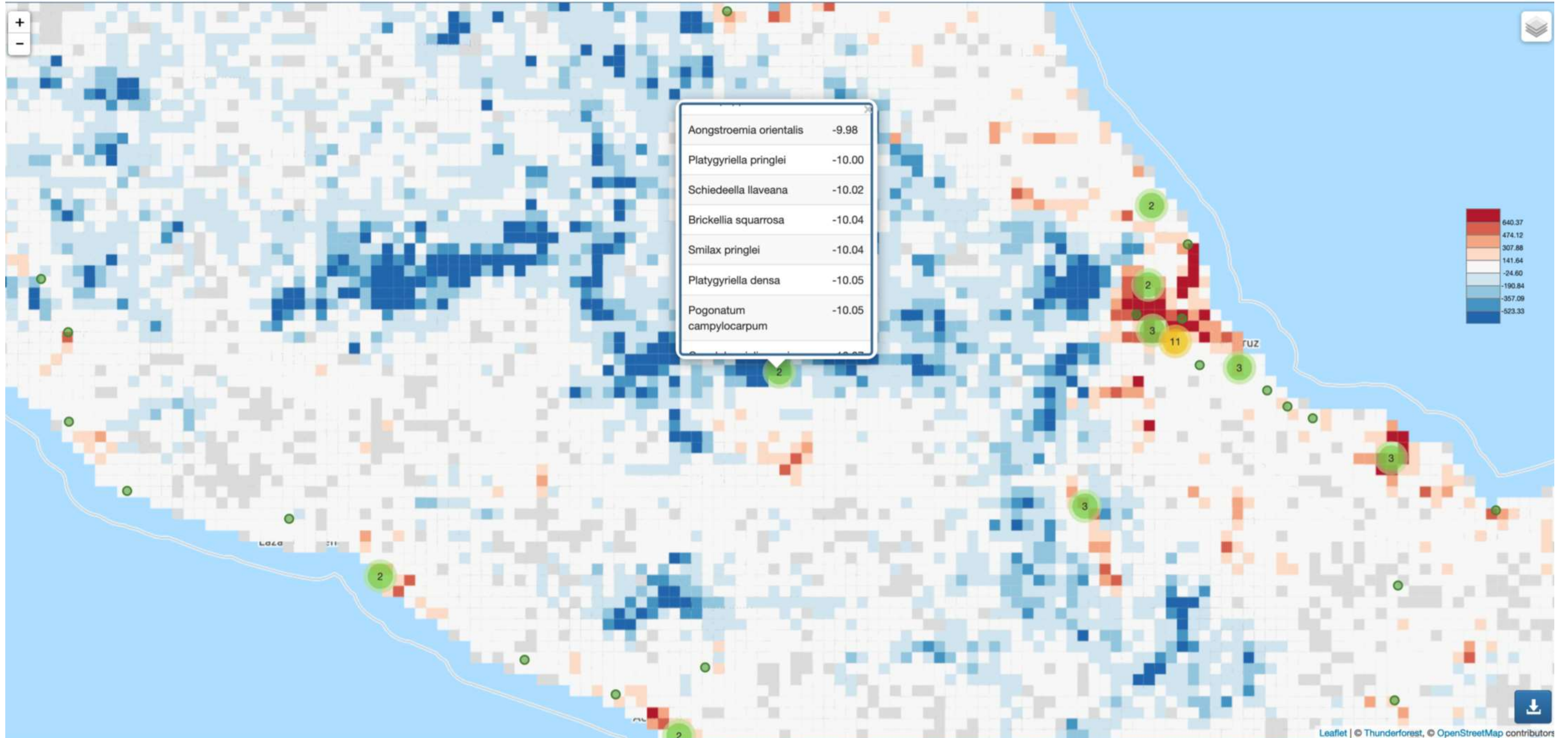
Proof of concept evaluation

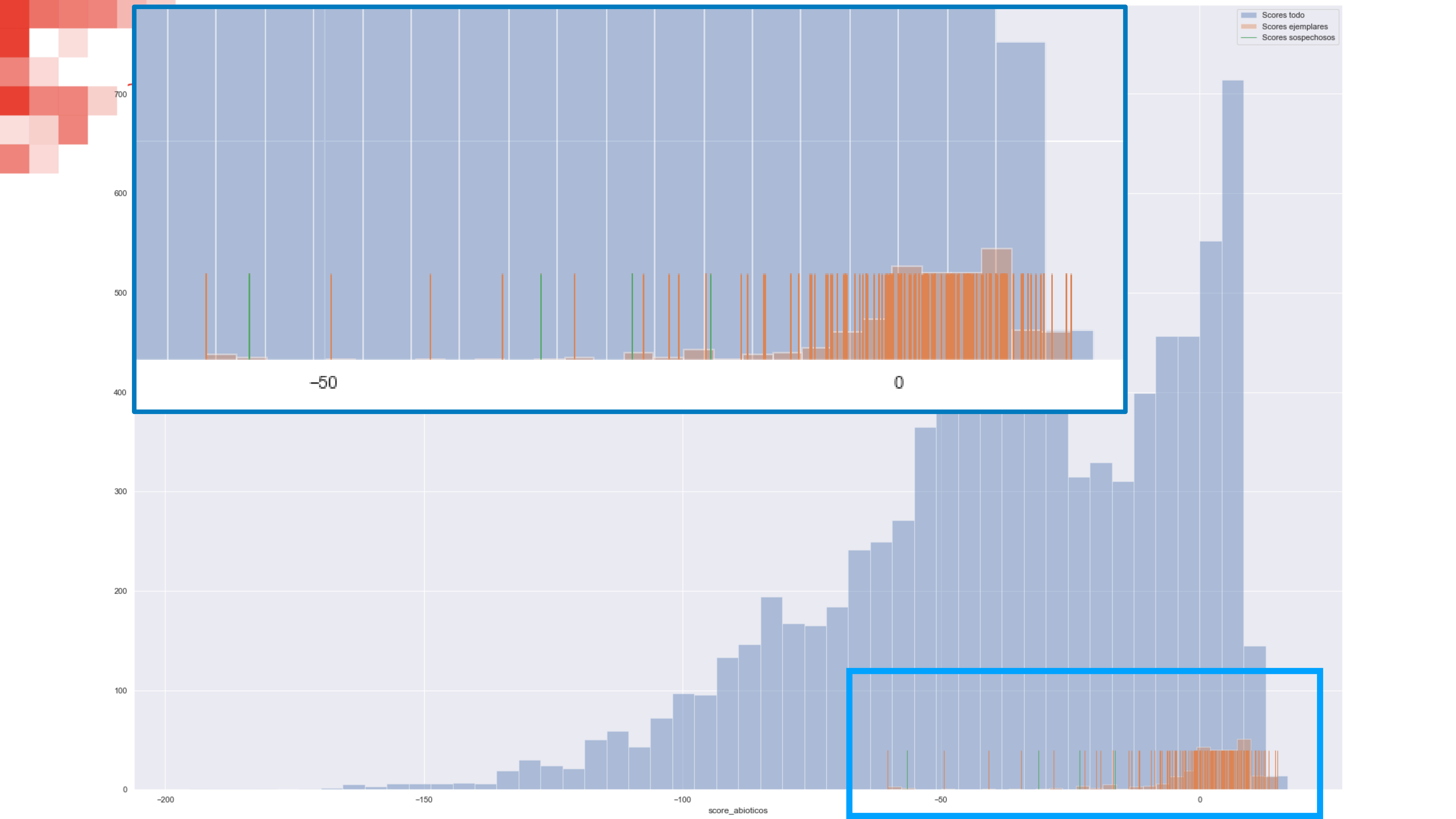
1 person, 13 days (104 hours), 60 cactus species, ~8000 points

QA quick workflow:

1. If inside known distribution then valid, else
2. If inside potential distribution then valid, else
3. If on border or no potential distribution map and records not too scattered, then check land cover maps and literature







Test results with climate data

		Flagged	
		False	True
Actual	False	6223	787
	True	411	271

Precision	Recall
0.25	0.46

Proportion of suspicious records in data	Proportion of suspicious records in flagged data
0.08	0.25

Summary (First impressions, future work, questions)

Strengths:

- Intuitive
- Use all the data available

Weaknesses:

- Large number of variables
- Sensitive to outliers

Questions:

- What is a helpful output?
- How to extrapolate to new regions?

Work:

- Improve recall
- Test pines, mammals, **birds**
- Outlier detection strategy (~cross-validation)
- Communication with QA analysts



BIO
DIVERSITY
NEXT



CONABIO

COMISIÓN NACIONAL PARA
EL CONOCIMIENTO Y USO
DE LA BIODIVERSIDAD

Automatizing the detection of strange species occurrence records

Presents: Raúl Sierra-Alcocer | National Commission for the
Knowledge and use of Biodiversity, Mexico

Authors: Raúl Jiménez Rosenberg, Raúl Sierra-Alcocer

Summary (First impressions, future work, questions)

Strengths:

- Intuitive
- Use all the data available

Weaknesses:

- Large number of variables
- Sensitive to outliers

Questions:

- What is a helpful output?
- How to extrapolate to new regions?

Work:

- Improve recall
- Test pines, mammals, **birds**
- Outlier detection strategy (~cross-validation)
- Work with QA analysts

Thank you, raul.sierra@conabio.gob.mx