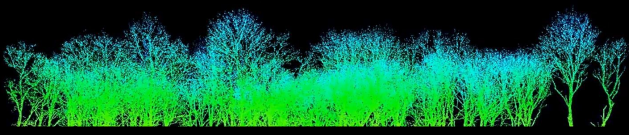


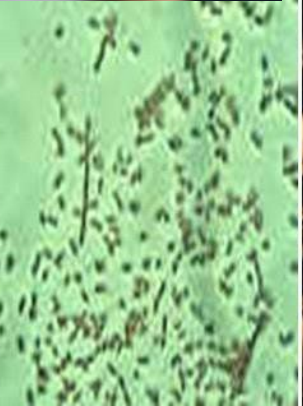
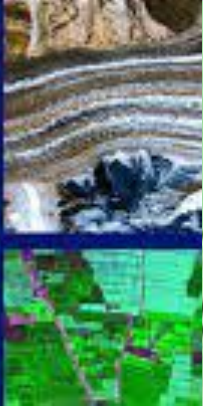
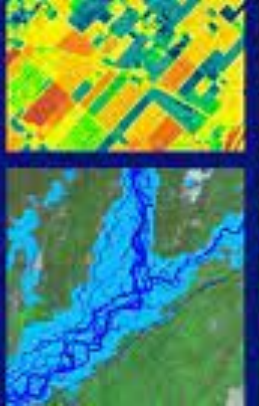
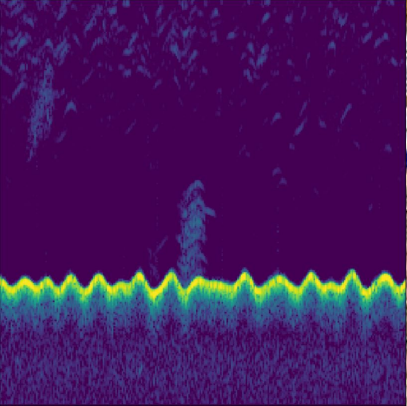
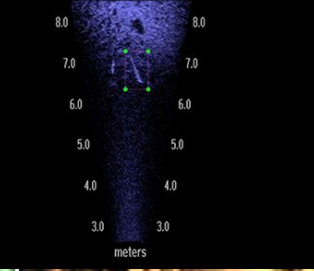
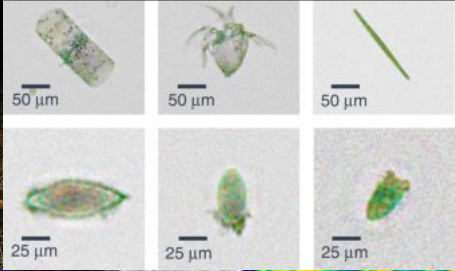


Split Lecture 4: Multimodality

Sara Beery | 3/16/26



Can we build AI systems that share information across data modalities?



Topics today

- Options for combining information
- Language as a “universal” modality
- Remote sensing as a “universal” modality
- Knowledge and structure as a modality
- Key differences in ecology from current multimodal AI



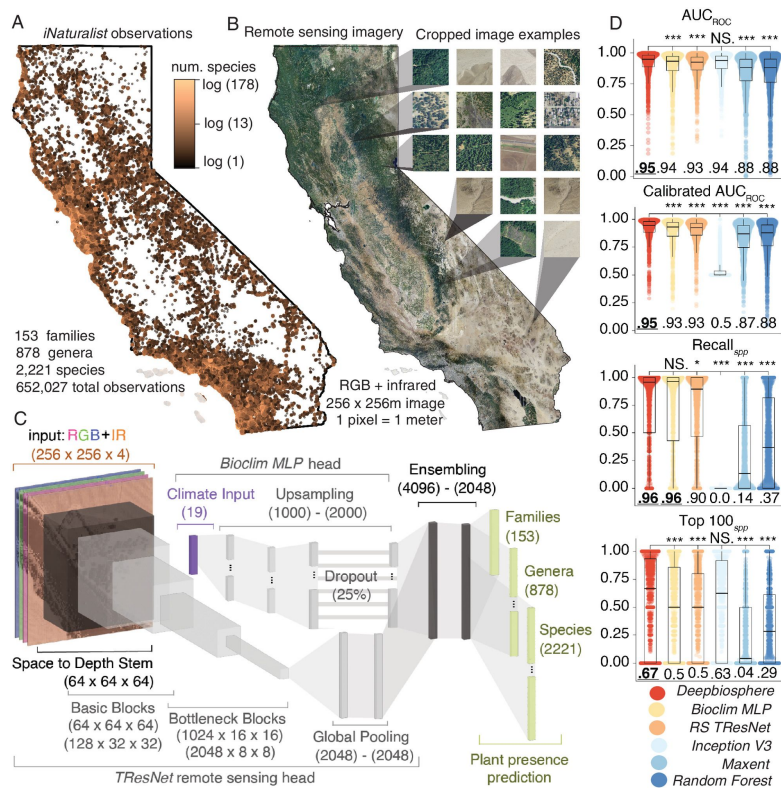
Options for how we can share information

Option 1: Use one modality to train a model for another

E.g. identify species or crop types in aerial data using ground-level data for training

Option 1: Use one modality to train a model for another

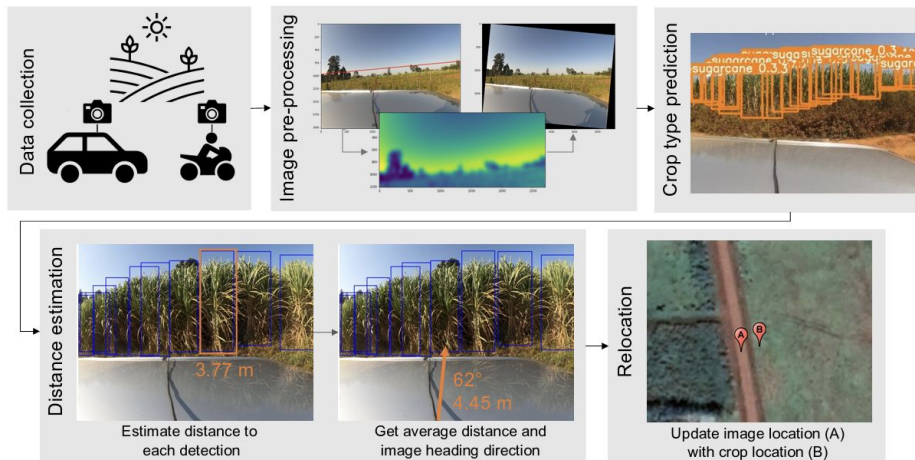
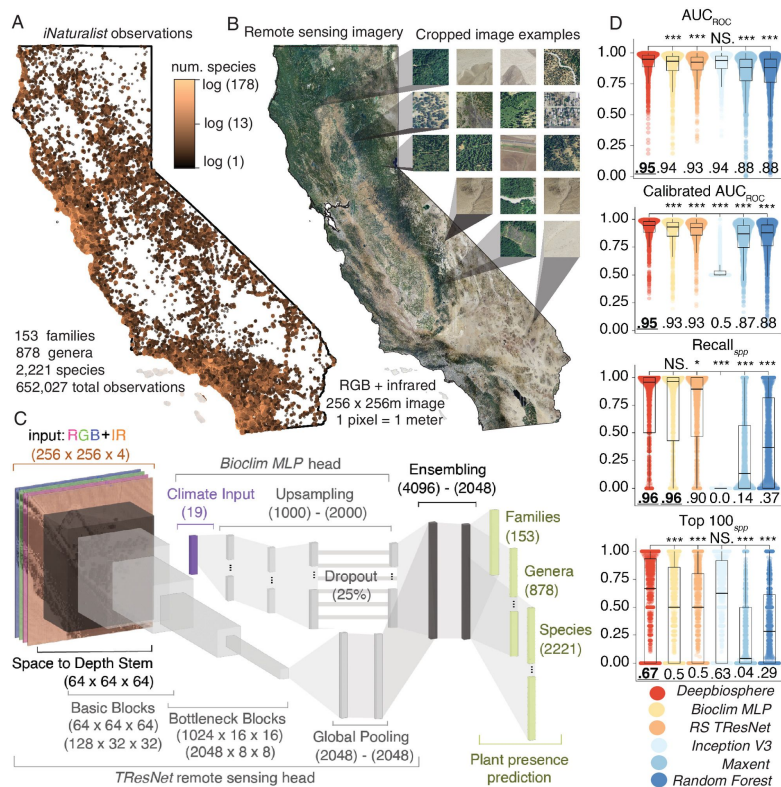
E.g. identify species or crop types in aerial data using ground-level data for training



Deep learning models map rapid plant species changes from citizen science and remote sensing data, Gillespie et al., PNAS 2024

Option 1: Use one modality to train a model for another

E.g. identify species or crop types in aerial data using ground-level data for training



Deep learning models map rapid plant species changes from citizen science and remote sensing data, Gillespie et al., PNAS 2024

Street2Sat: A Machine Learning Pipeline for Generating Ground-truth Geo-referenced Labeled Datasets from Street-Level Images, Paliyam et al., CCAI@ICML 2021

Option 2: Use agreement across modalities

E.g. increase confidence in a prediction when modalities agree

Option 2: Use agreement across modalities

E.g. increase confidence in a prediction when modalities agree



Option 2: Use agreement across modalities

E.g. increase confidence in a prediction when modalities agree

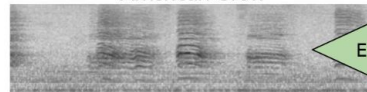


American Crow



Common Raven

Hard



Easy

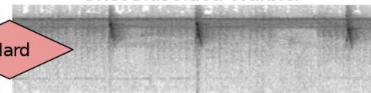


Yellow Warbler



Chestnut-sided Warbler

Easy



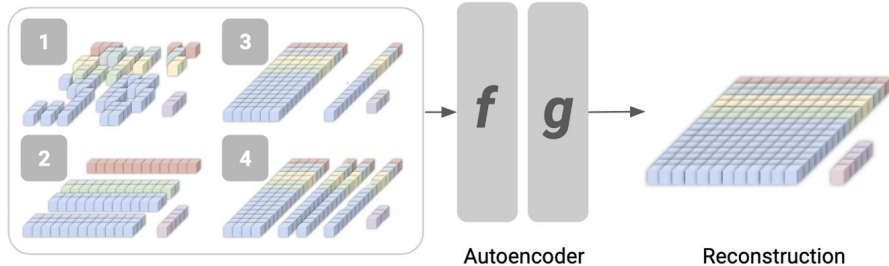
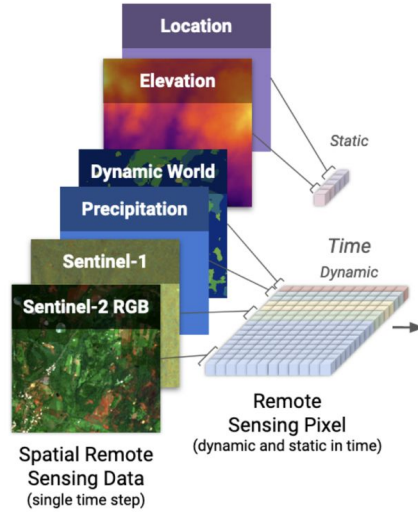
Hard

Option 3: Train representations across modalities

E.g. learn to capture information across modalities directly within your AI model

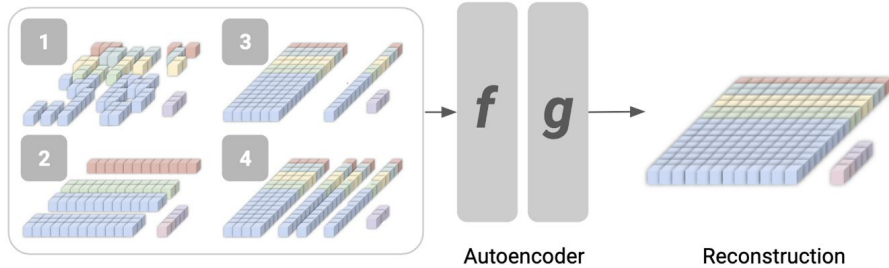
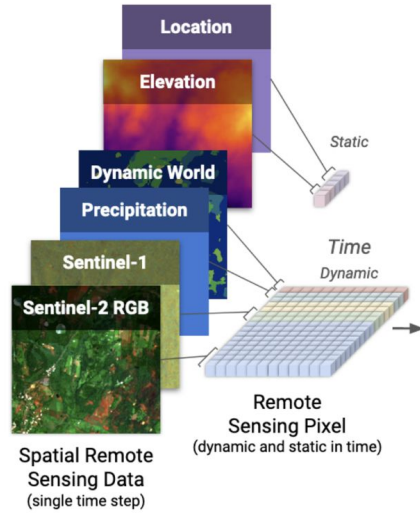
Option 3: Train representations across modalities

E.g. learn to capture information across modalities directly within your AI model

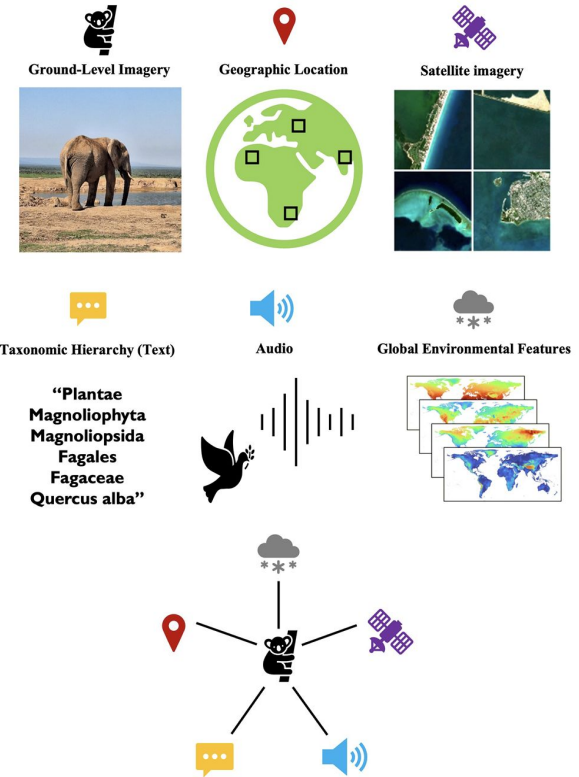


Option 3: Train representations across modalities

E.g. learn to capture information across modalities directly within your AI model



PRESTO: Lightweight, Pre-trained Transformers for Remote Sensing Timeseries, Tseng et al., CCAI@NeurIPS 2024



TaxaBind: A Unified Embedding Space for Ecological Applications, Sastry et al., WACV 2025



Language + X



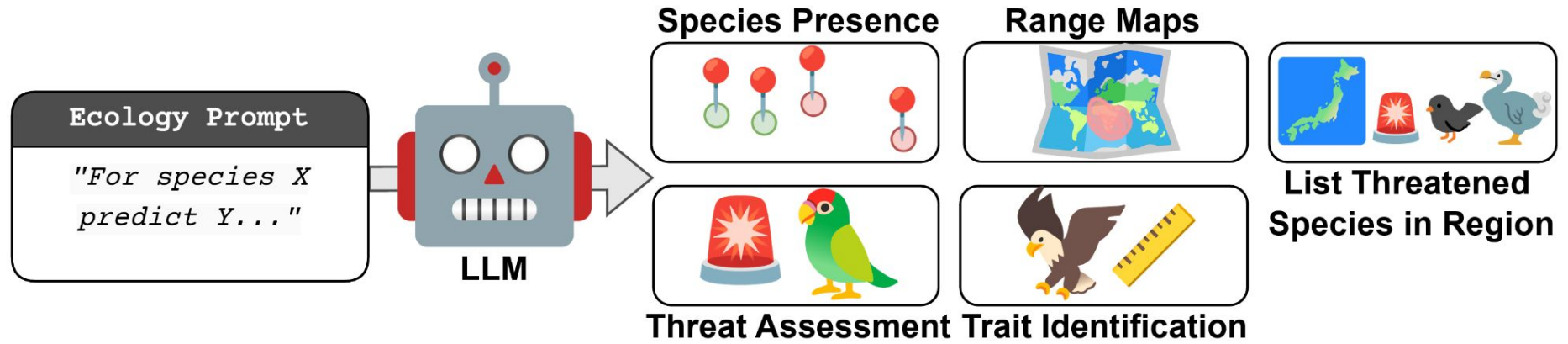
Large Language Models and Generative AI

Large Language Models (LLMs) are AI algorithms trained on extensive text data to comprehend and generate human-like language.

- ChatGPT - OpenAI
- Gemini - Google
- Copilot - Microsoft
- Mistral - MistralAI
- Claude - Anthropic
- DeepSeek - High-Flyer

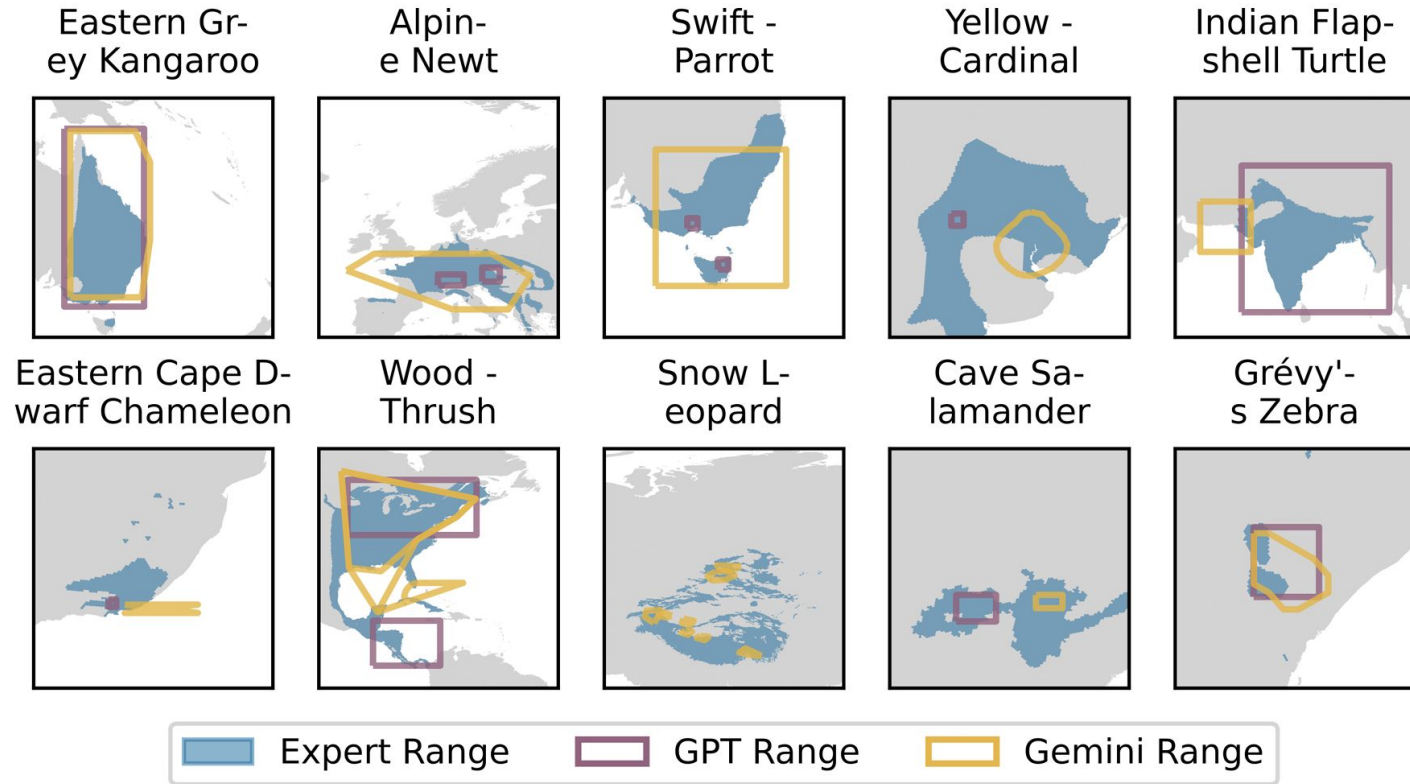


What do LLMs know about ecology?



"While these models could assist in accelerating certain conservation-related tasks and already make ecological information more accessible, their limitations suggest that they are not yet ready for direct application in biodiversity decision making."

What do LLMs know about ecology?



Prompt engineering

- Prompts are brittle. “A photo of a dog” vs “A picture of a dog” will give very different results
- In-context or “chain of thought” prompting
- It can be hard to track/understand errors (reasoning is wrong, answer is right?)

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

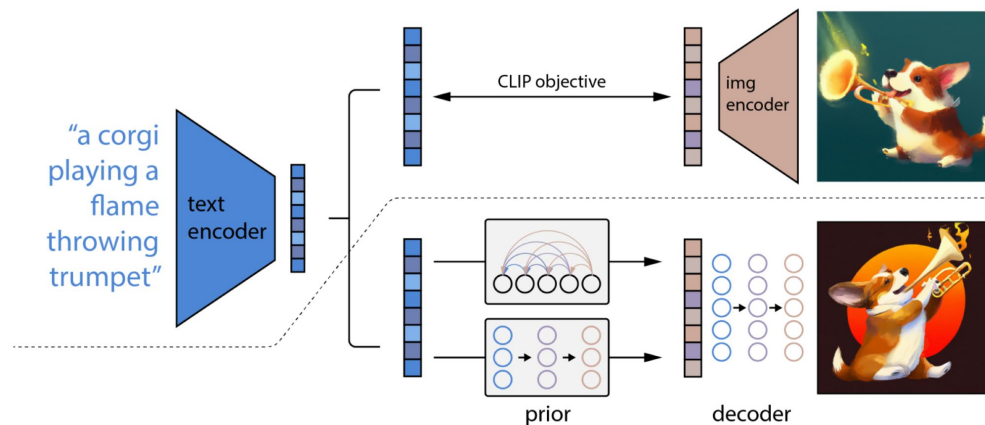
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

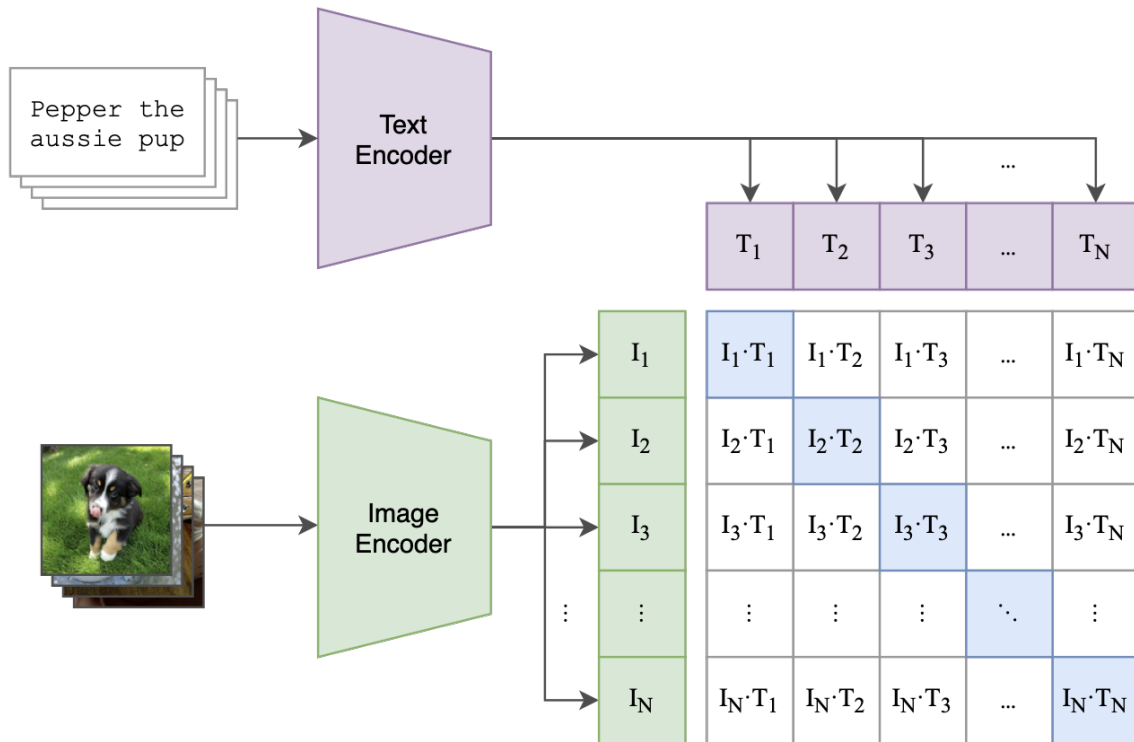
Generative vision-language models

- Text-conditional image/video generation
 - Stable diffusion, DALL-E, Sora
- Image-conditioning in generative language models
 - Ie “Does this image contain an American Robin?”
 - Use image encoders that are accessible/interpretable to the language models, but more tokens per image than CLIP (ie 170 tokens for an image in GPT-4O)



Joint embeddings

- CLIP - contrastive image/language training
- SigLIP introduces sigmoid loss to handle normalization issues with softmax



Joint embeddings with language



🔍 "A California condor tagged with a green 26"

Joint embeddings with language

🔍 "A California condor tagged with a green 26"



Multimodal AI model
(e.g. CLIP)



Joint embeddings with language



🔍 "A California condor tagged with a green 26"

Multimodal AI model
(e.g. CLIP)



Joint embeddings with language

Most Similar



0.95



0.91



0.62



0.17



0.06

Least Similar

 *"A California condor tagged with a green 26"*

Language + X embeddings enable open-ended interactive search

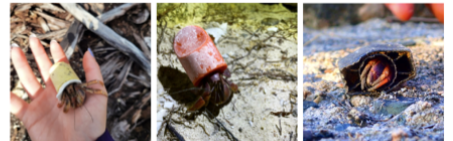
A tamandua pup being carried on its mother's back



Everted osmeterium



A hermit crab using plastic waste as its shell



Kahili ginger plant with open fruit capsule showing seeds



Melanistic leopard



INQUIRE

A humpback whale fluke with identifiable markings



A satin bowerbird's ornamented nest



A reef with manmade structures and debris



Strawberry poison-dart frog with the "la gruta" morph



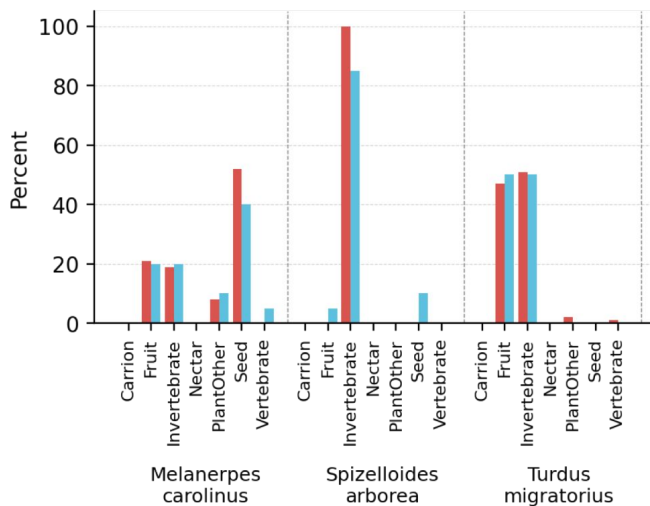
California condor tagged with a green "26"



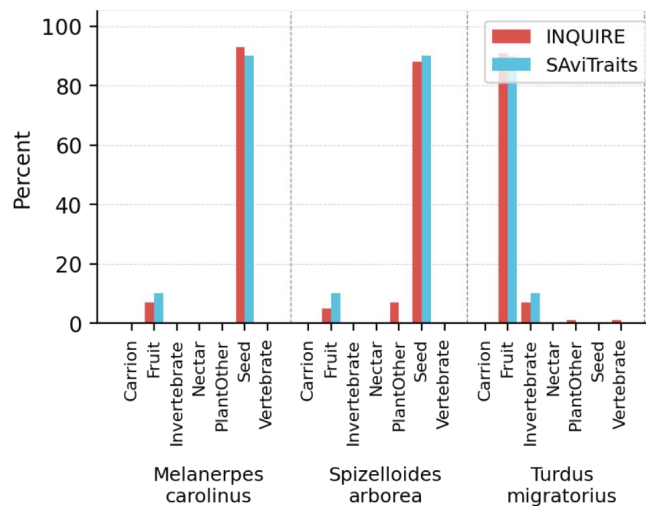
Using interactive search to discover ecological trends



Summer diet composition



Winter diet composition



**work with
Marta
Jarzyna*

Code assistants

- Can make AI more accessible
- Doesn't require foundational knowledge
- What are the pros and cons?



GitHub Copilot: AI that builds with you

 GitHub Copilot

Ask a question or type '/' for topics



interested.

can you provide imagery for the first 2 weeks of august

I'll retrieve the satellite imagery for Odemira for the first two weeks of August 2023.

Spotify

Added Recent Imagery to Data Pane

I've retrieved satellite imagery for Odemira from the first two weeks of August 2023. The results show six satellite images from two different Sentinel-2 satellite tiles (29SNB and 29SNC) on the following dates:

- August 2, 2023
- August 7, 2023
- August 12, 2023

These images cover the area of Odemira during the specified time period. The images are from the Sentinel-2 mission, which provides high-resolution optical imagery useful for monitoring land use, vegetation, and potential disturbances.

Would you like me to help you interpret these images or provide more context about the disturbance alerts we found earlier?

Ask a question



Layers

- Location Layer
- Disturbances
- Sentinel-2



3 km

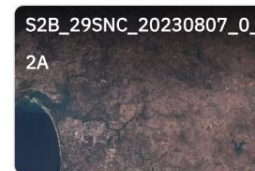
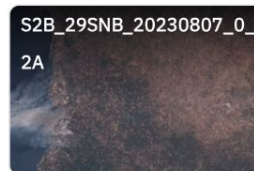
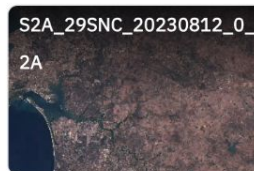
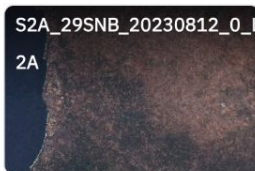
Background tiles: © OpenStreetMap contributors

lat, lon: 37.463, -8.761

The planet's most powerful AI monitoring system.

Data Visualization Recent Imagery Export data

Recent Imagery



*slide from Craig Mills, Land and Carbon Lab



Agentic AI systems

Query: How many muffins can each kid have for it to be fair?



Generated Code

```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    muffin_patches = image_patch.find("muffin")  
    kid_patches = image_patch.find("kid")  
    return str(len(muffin_patches) // len(kid_patches))
```

Execution

```
muffin_patches =  
image_patch.find("muffin")
```



```
kid_patches =  
image_patch.find("kid")
```



```
► len(muffin_patches)=8
```

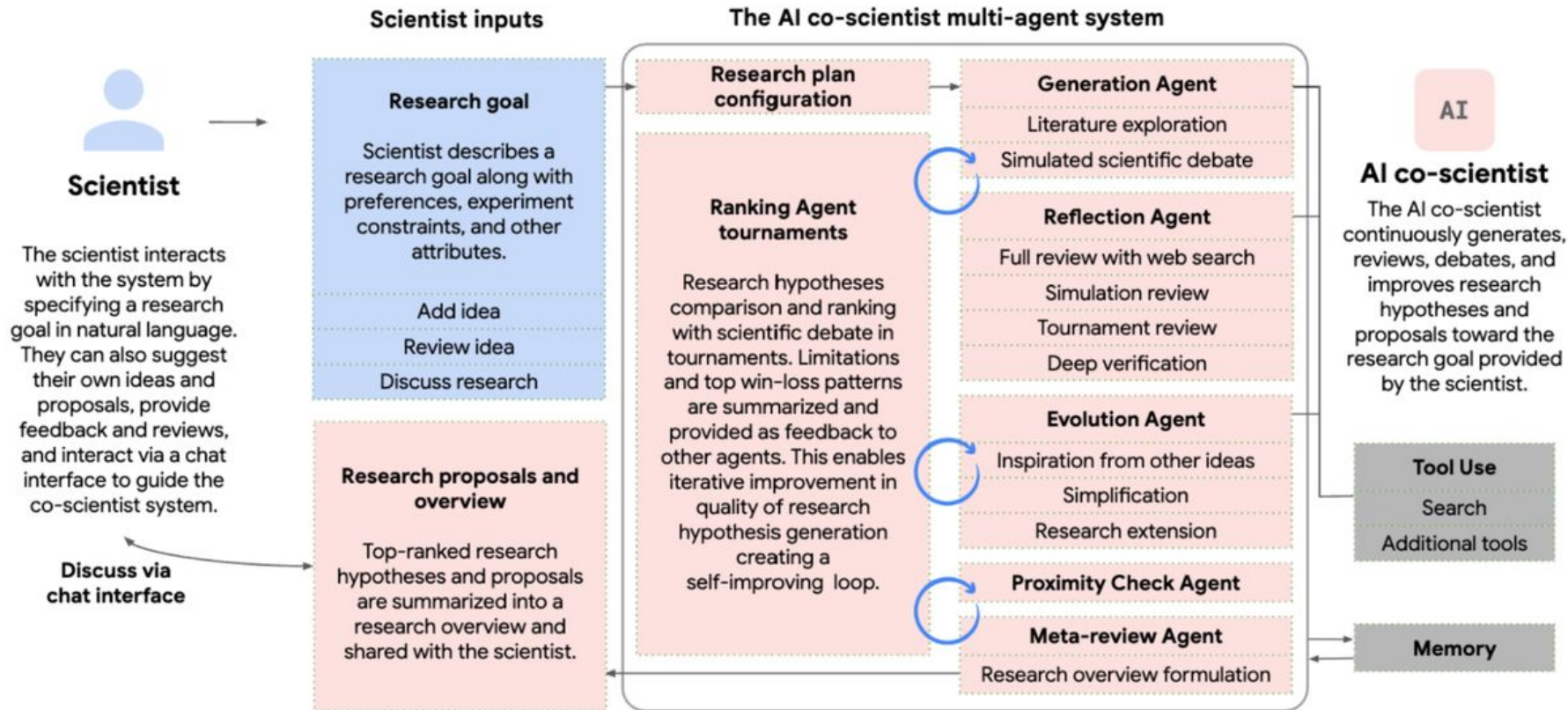
```
► len(kid_patches)=2
```

```
► 8//2 = 4
```

```
Result: 4
```

- AI that writes and executes code (ie Viper GPT, many more)
- Enables more complex reasoning
- Provides structure to force reliance on real data and functions, can help reduce hallucinations

Agentic AI



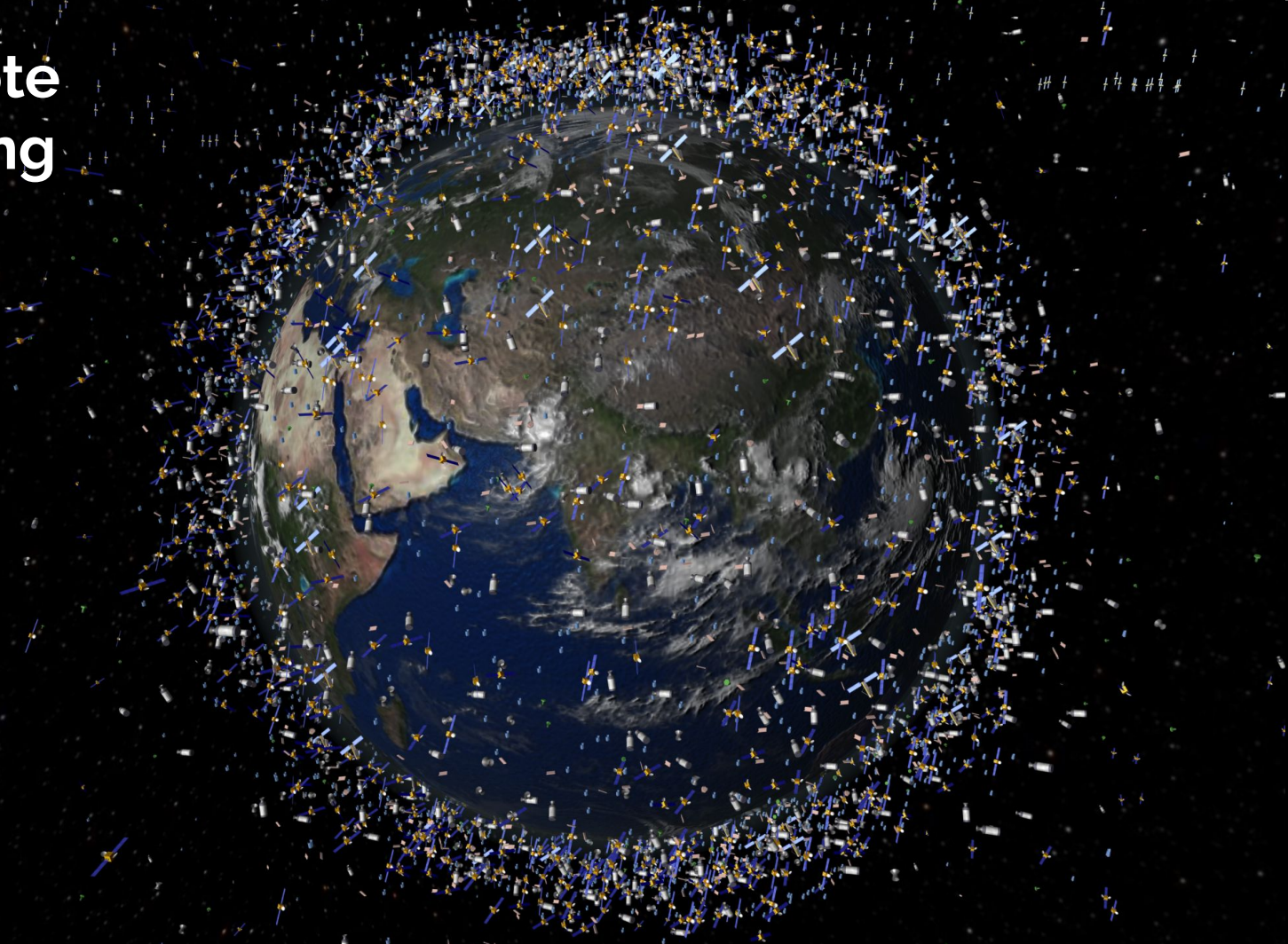
The AI co-scientist system design



Remote sensing + X

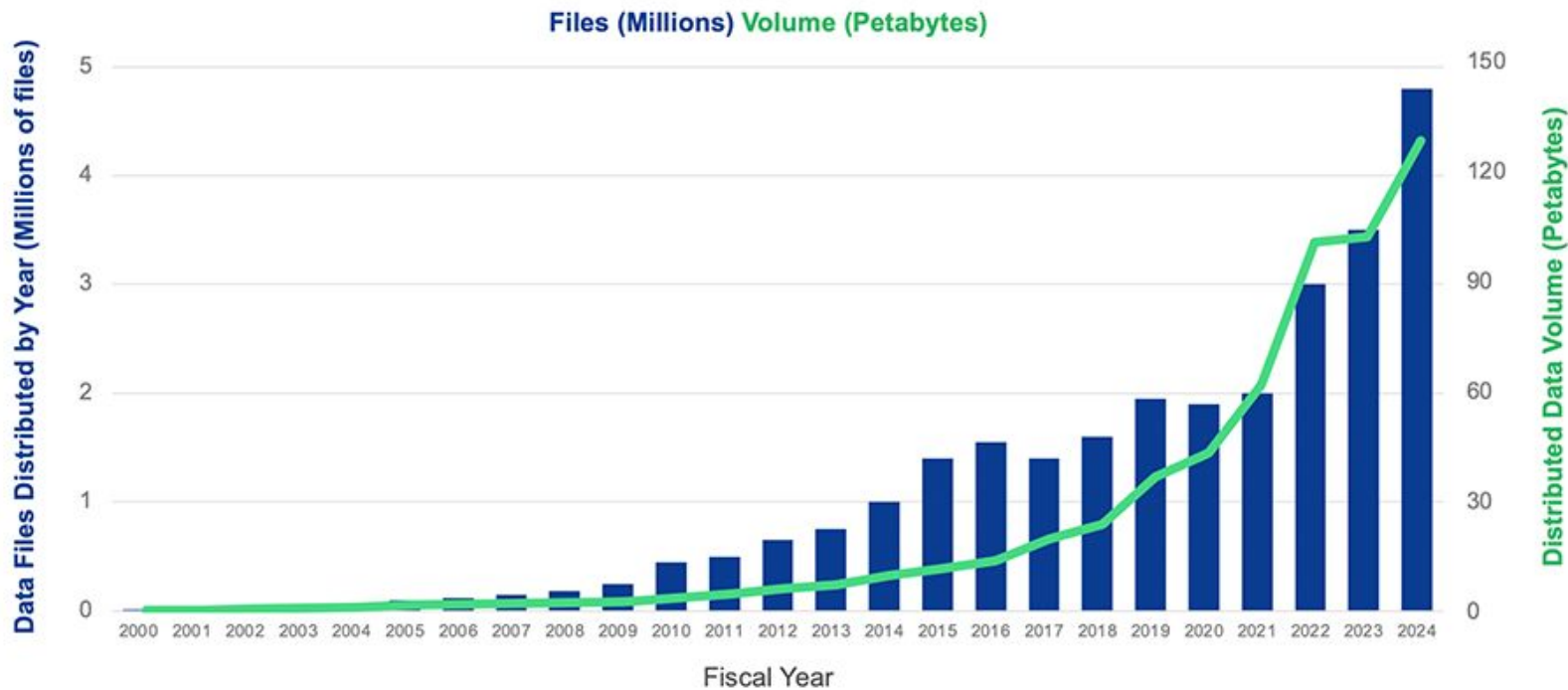


Remote sensing

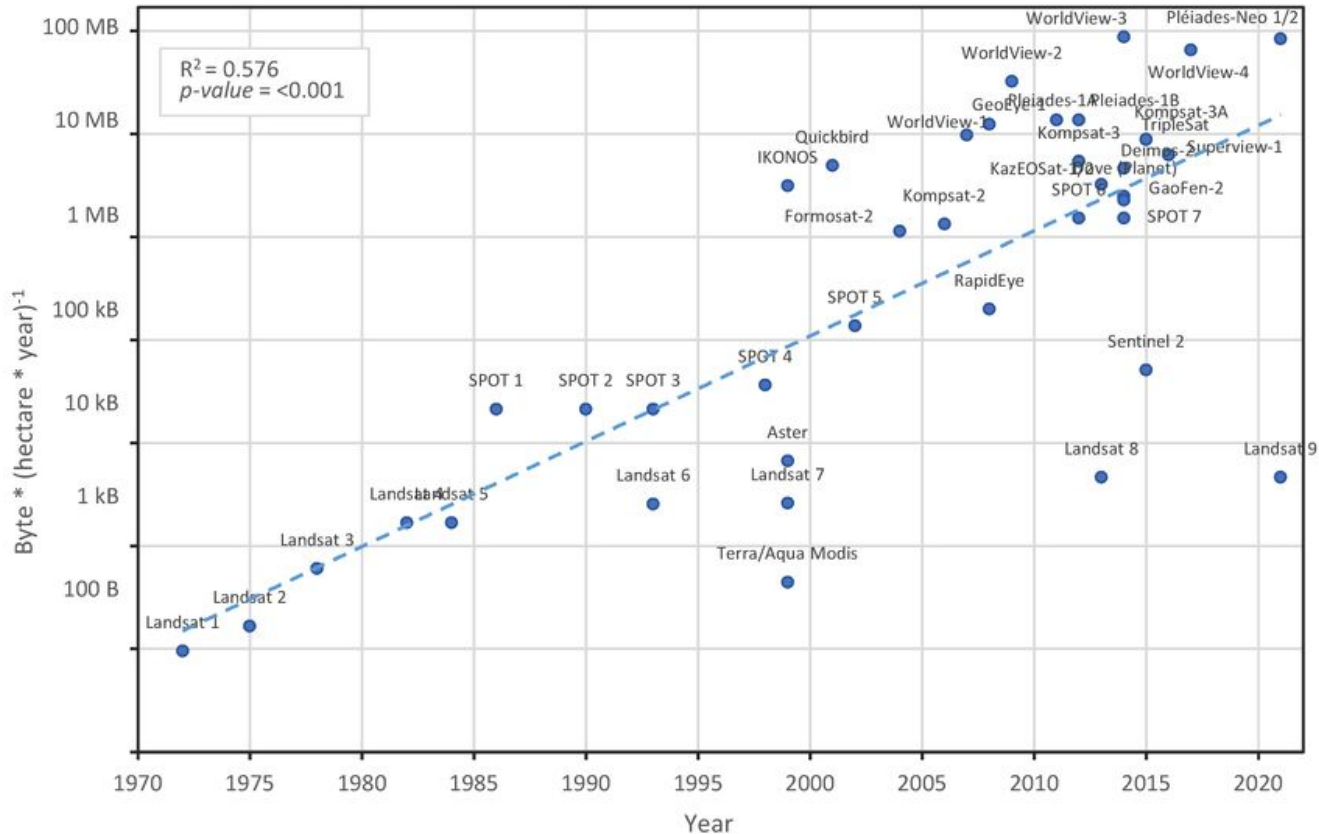




Total Data Volume and Data Files Distributed by Year



Evolution of data volume per hectare across platforms

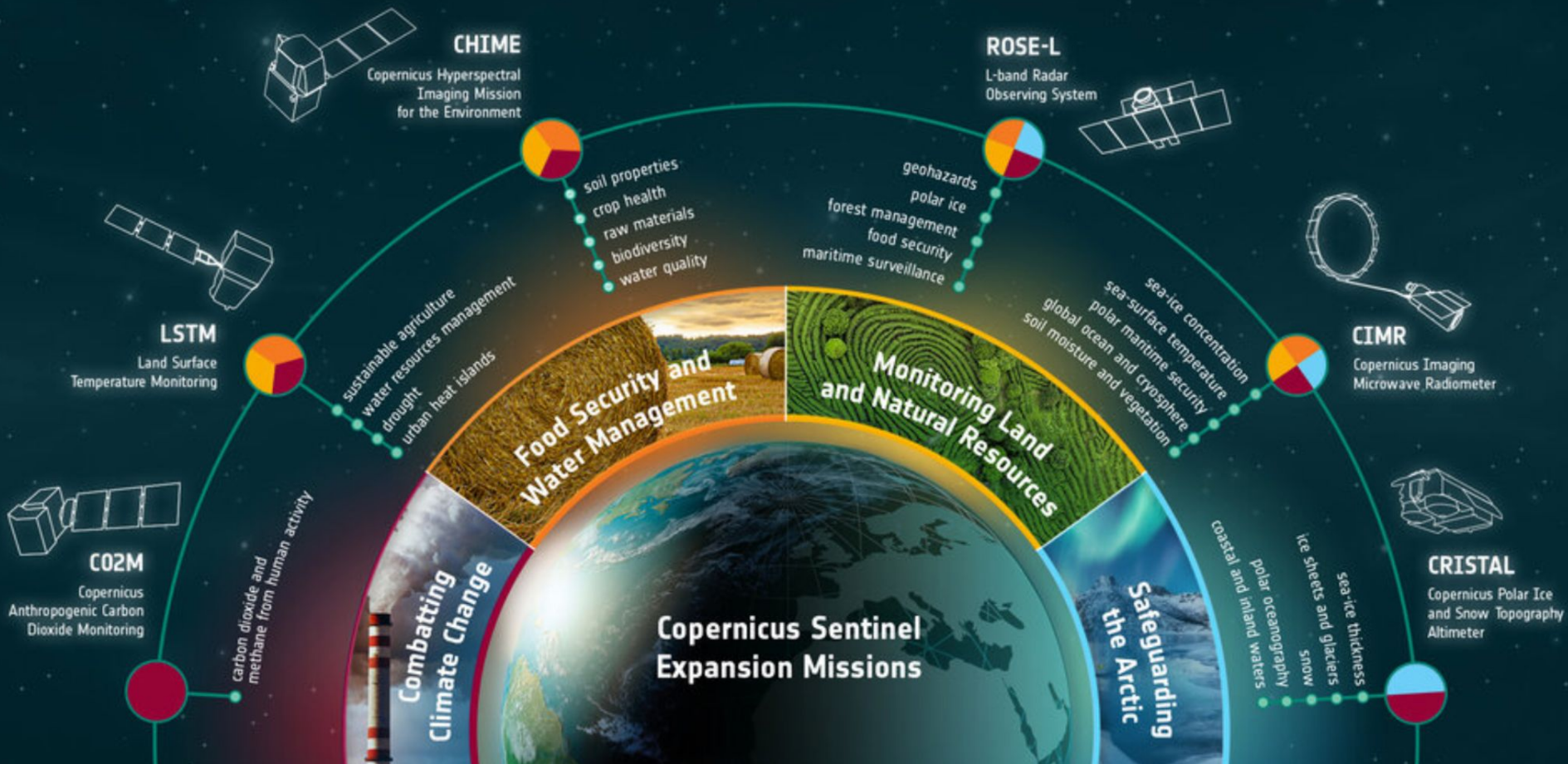




PROGRAMME OF THE
EUROPEAN UNION

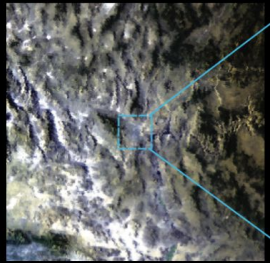


co-funded with

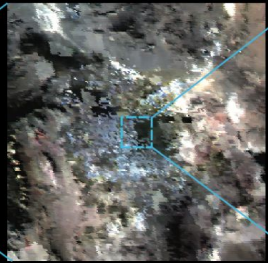


Spatial resolutions

Las Vegas, Nevada, USA



GOES-18 at 2000m/px



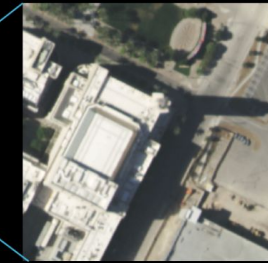
MODIS at 250m/px



Landsat 9 at 30m/px



Sentinel 2 at 10m/px



NAIP at 0.6m/px

Time steps

Las Vegas, Nevada, USA



Dec. 25, 1973



Dec. 3, 1982



Dec. 9, 1993



Dec. 24, 2001



Dec. 25, 2013



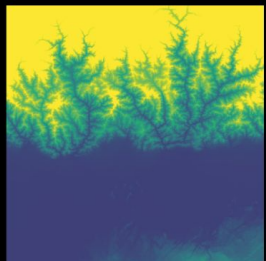
Dec. 28, 2023

Modalities

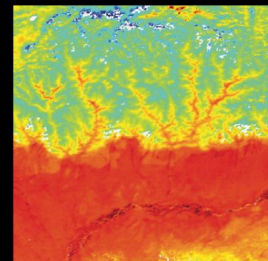
Eastern Himalayas

Washington, USA

English Channel



ALOS DEM



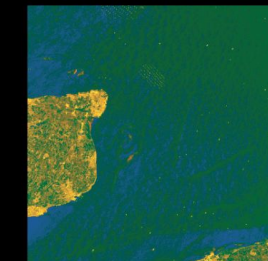
MODIS day temp.



16 day NDVI (Nov.)



16 day NDVI (May)

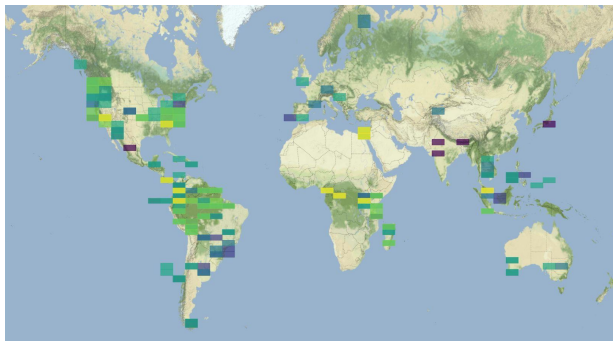
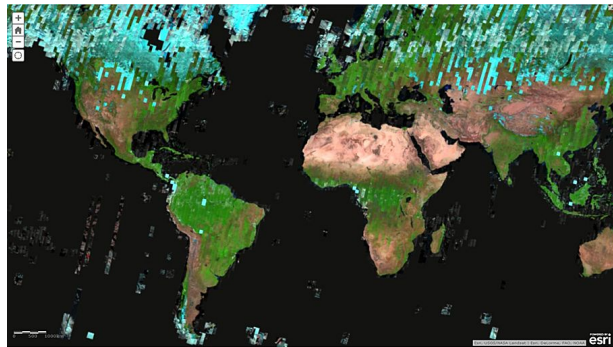
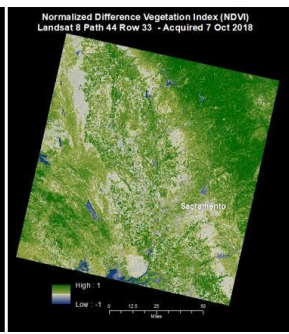
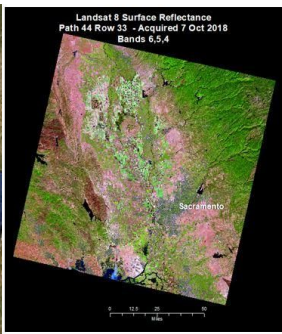


SAR (Sentinel 1)



**Where does ground
level data add value?**

Heterogeneous Sampling



Think about the relative tradeoffs for each modality

Mobile Sensors

Satellite (optical, SAR, LiDAR)



UAV (RGB, thermal, LiDAR)



On-Animal Sensors

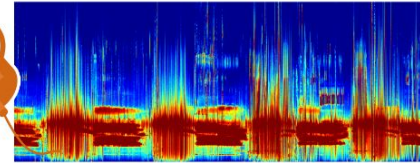


Stationary Sensors

Camera Traps

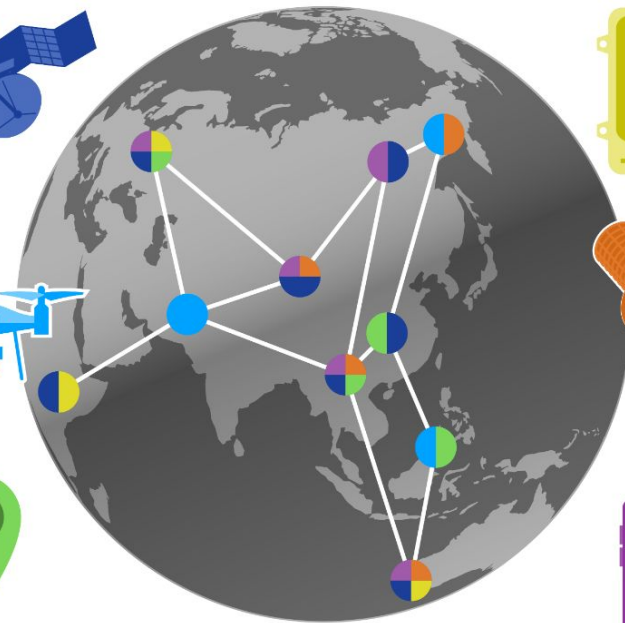
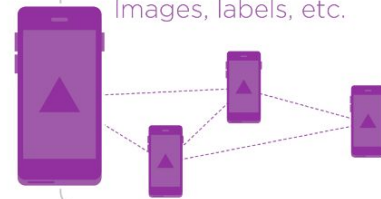


Bioacoustic Sensors



Community Science

Images, labels, etc.



Multiview aerial and street level imagery for the same tree instance

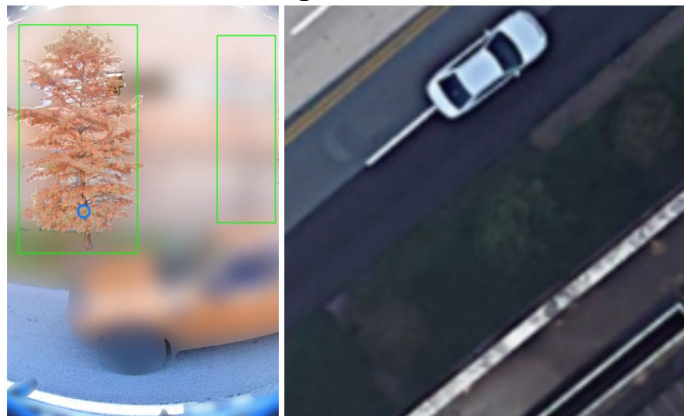
Sioux City, Fraxinus (Ash)



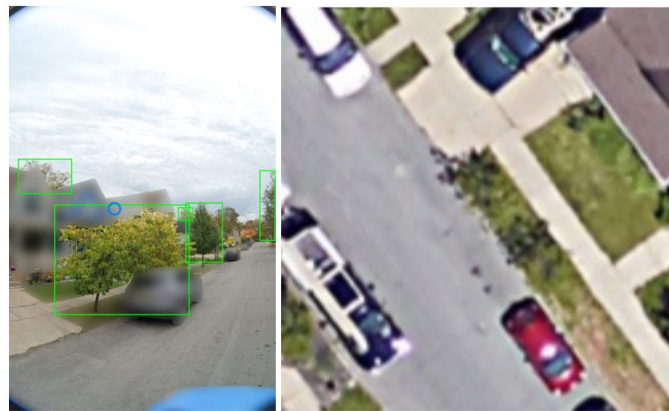
Sioux City, Tilia



Pittsburgh, Taxodium

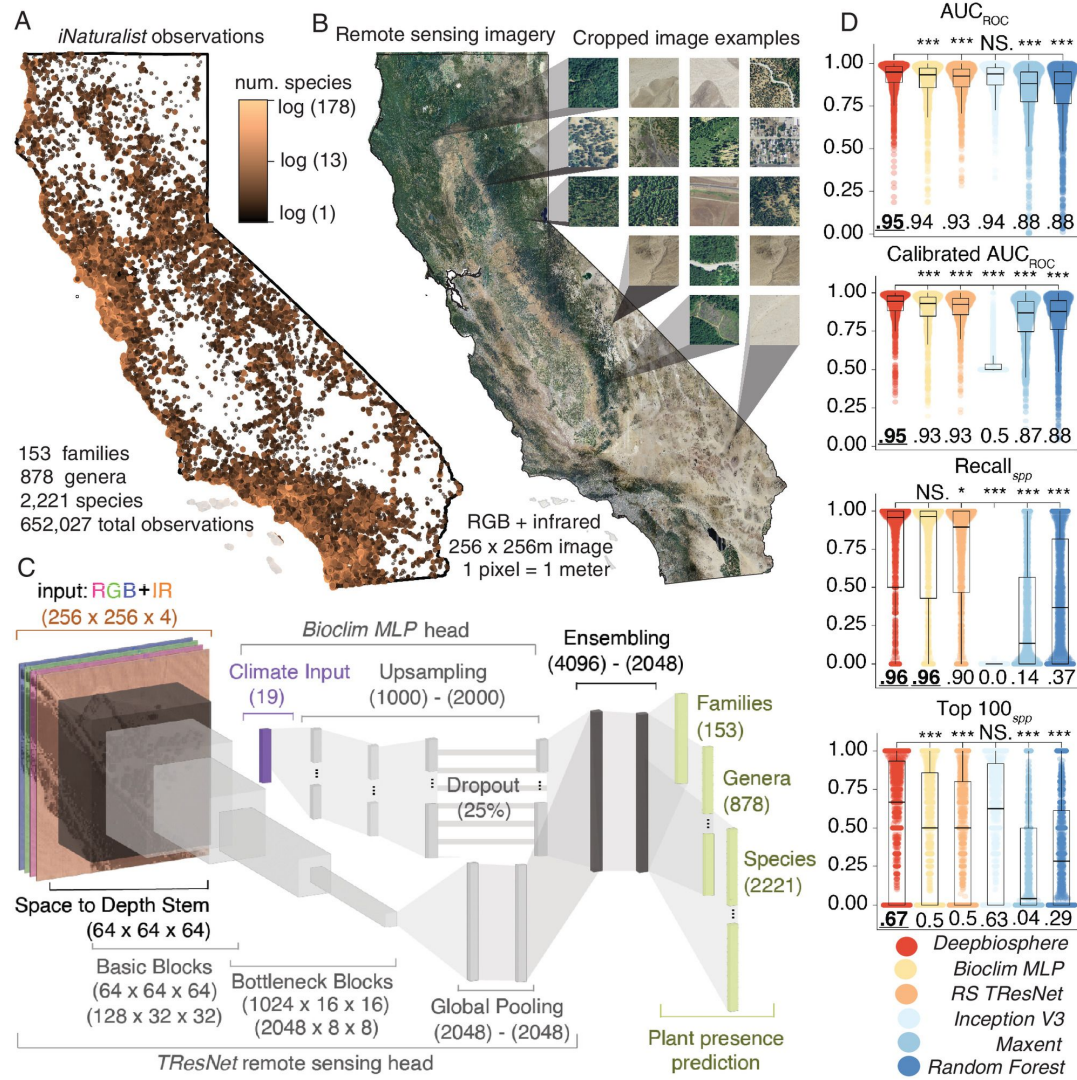


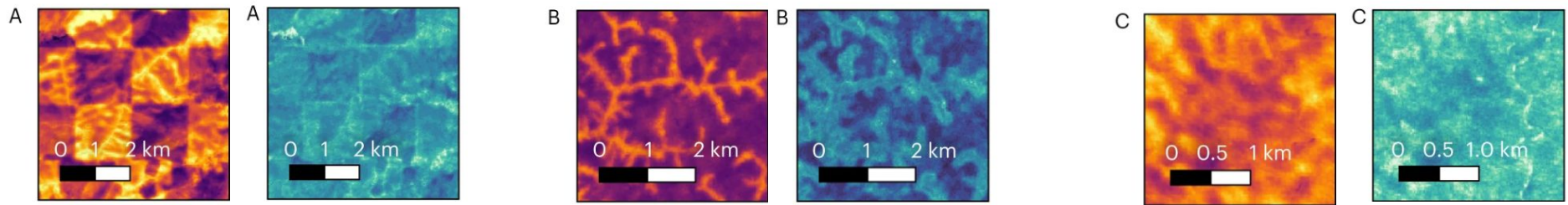
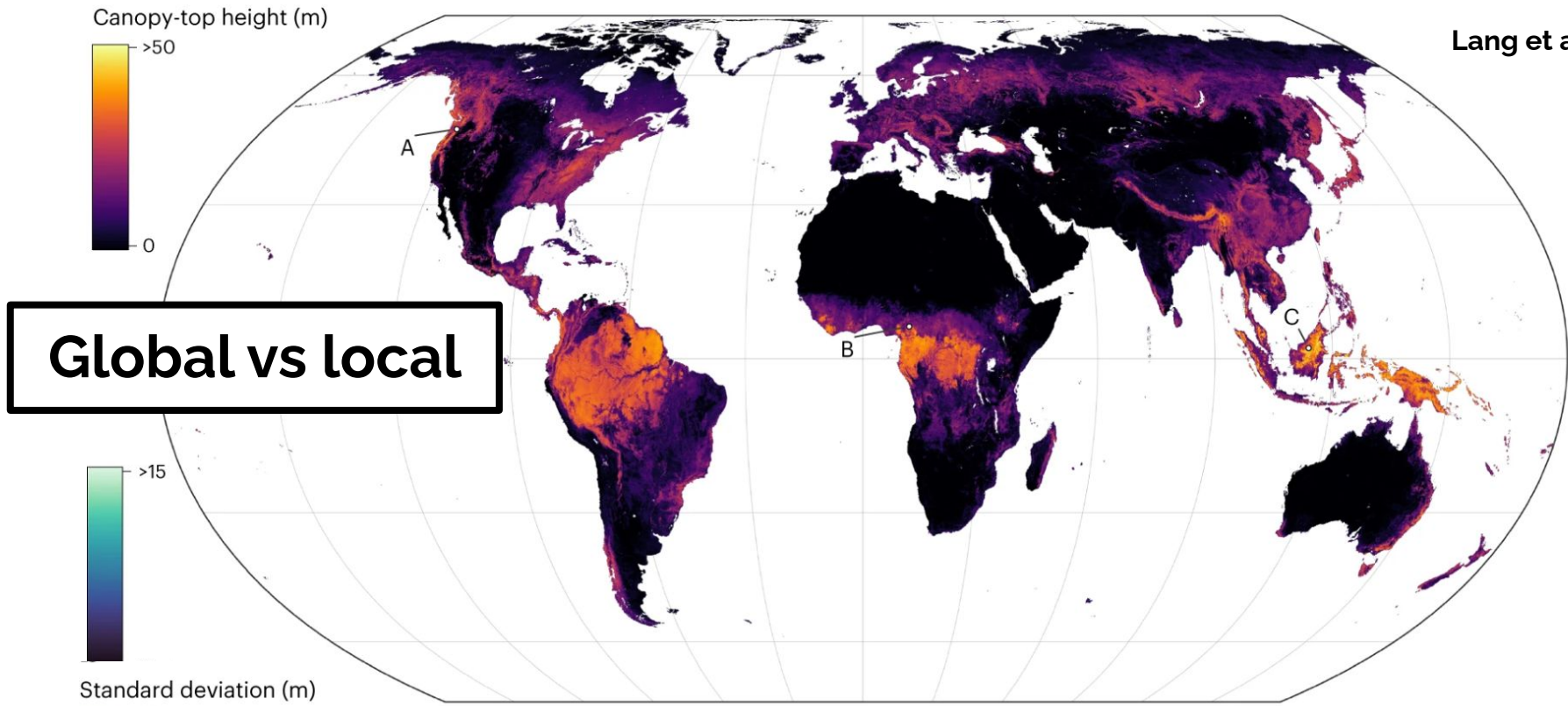
Buffalo, Cercis (Redbud)

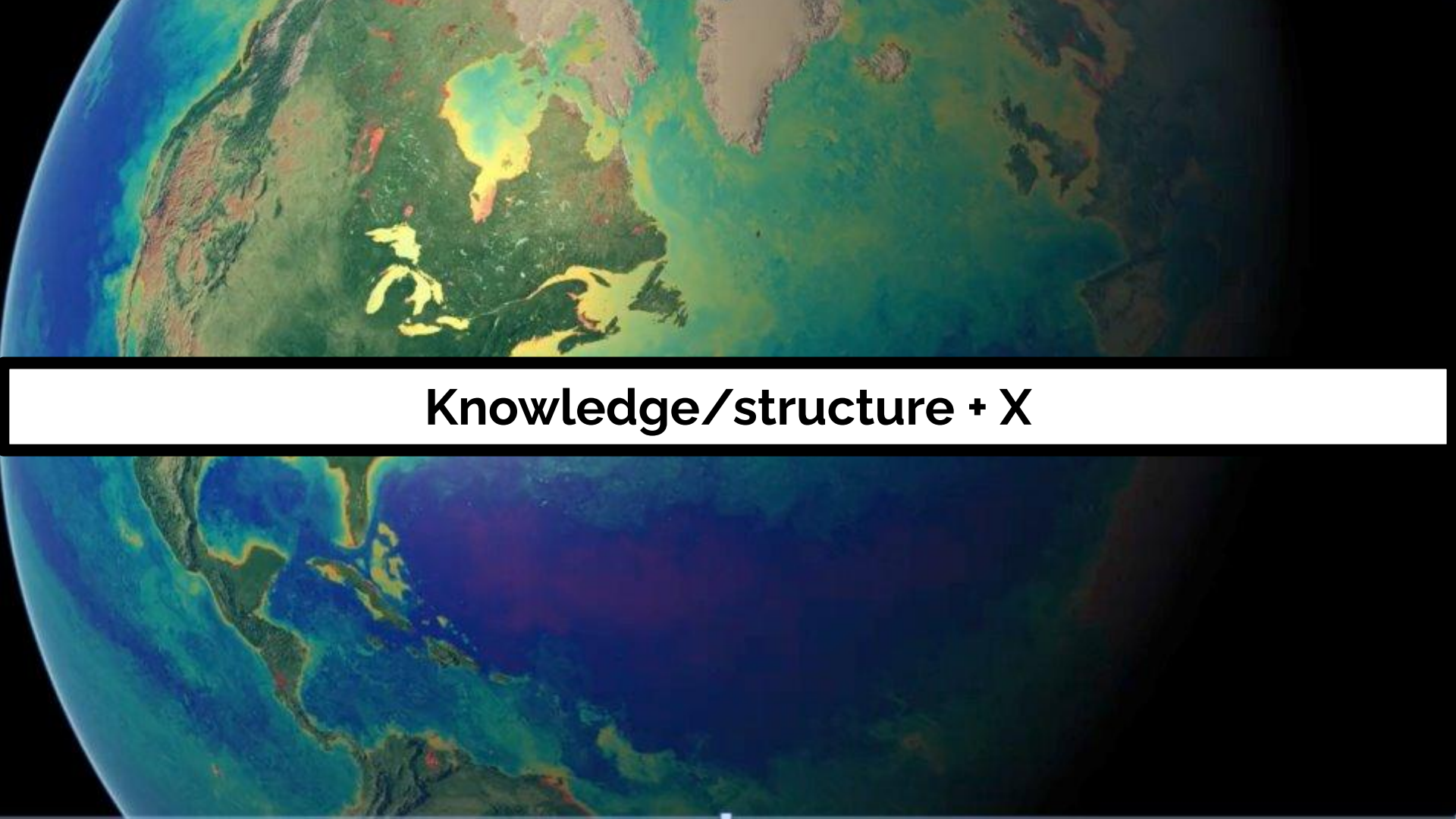


Improved species distributions from RS+iNat

Show finer-scale species mapping for trees and understory plants, and show detection of change in plant communities





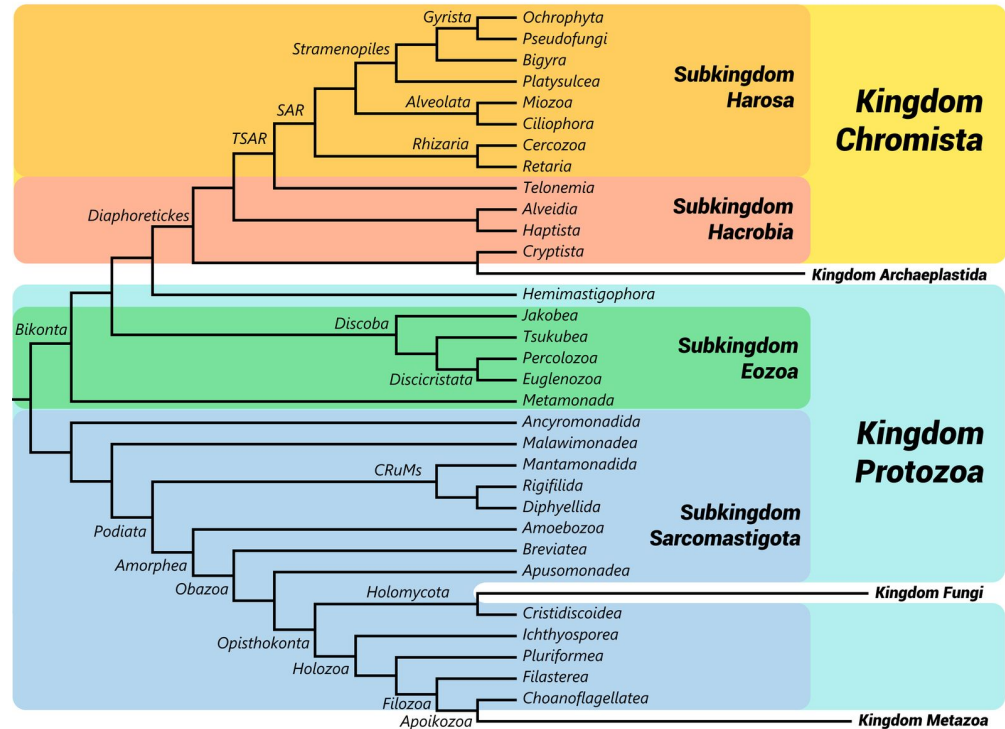


Knowledge/structure + X

Knowledge and structure

i.e. *taxonomy*:

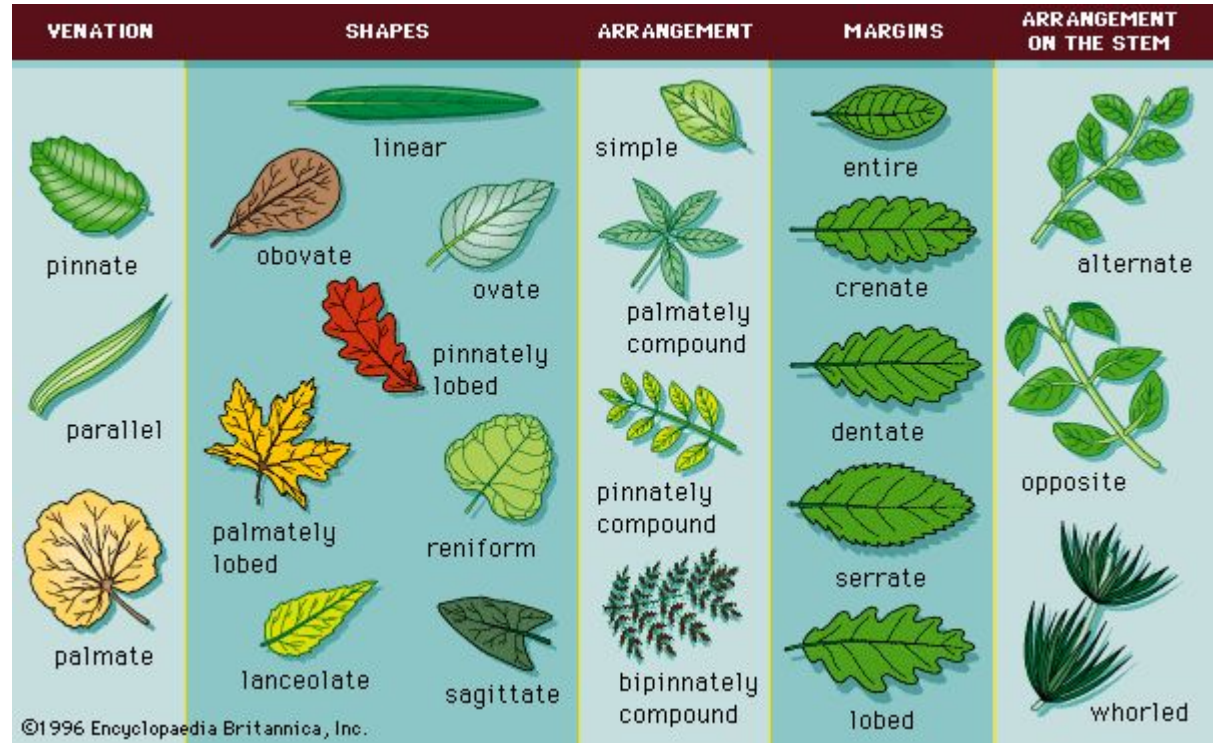
how organisms are
named and
categorized



<https://organismalbio.biosci.gatech.edu/biodiversity/phylogenetic-trees/>
https://evolution.berkeley.edu/evolibrary/article/phylogenetics_02

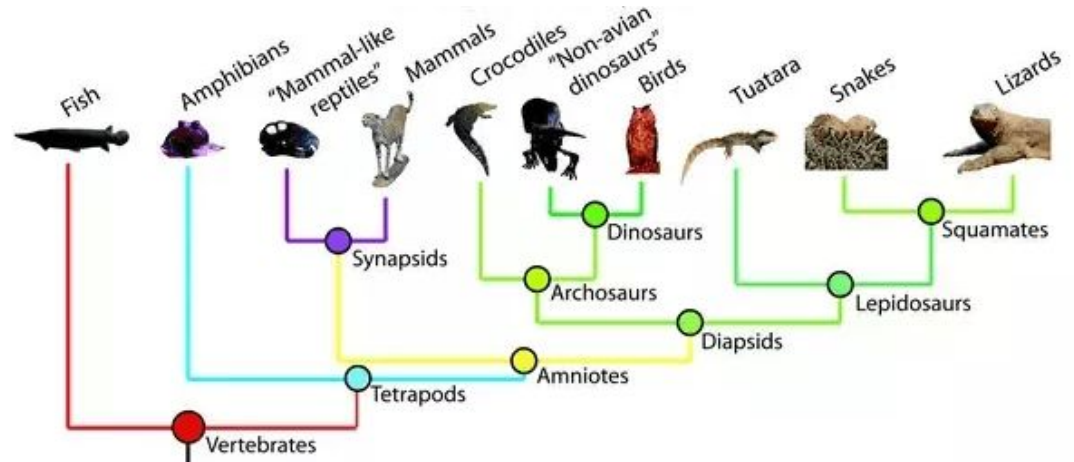
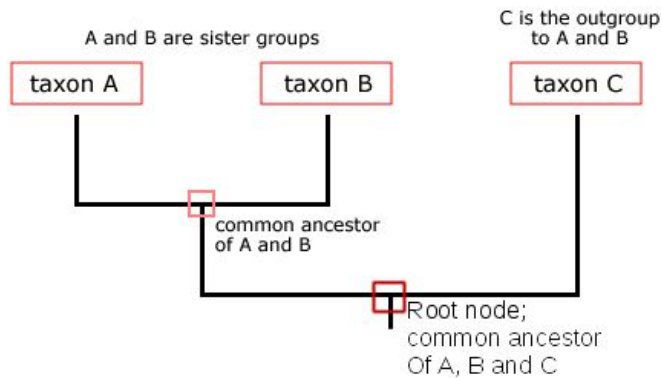
Knowledge and structure

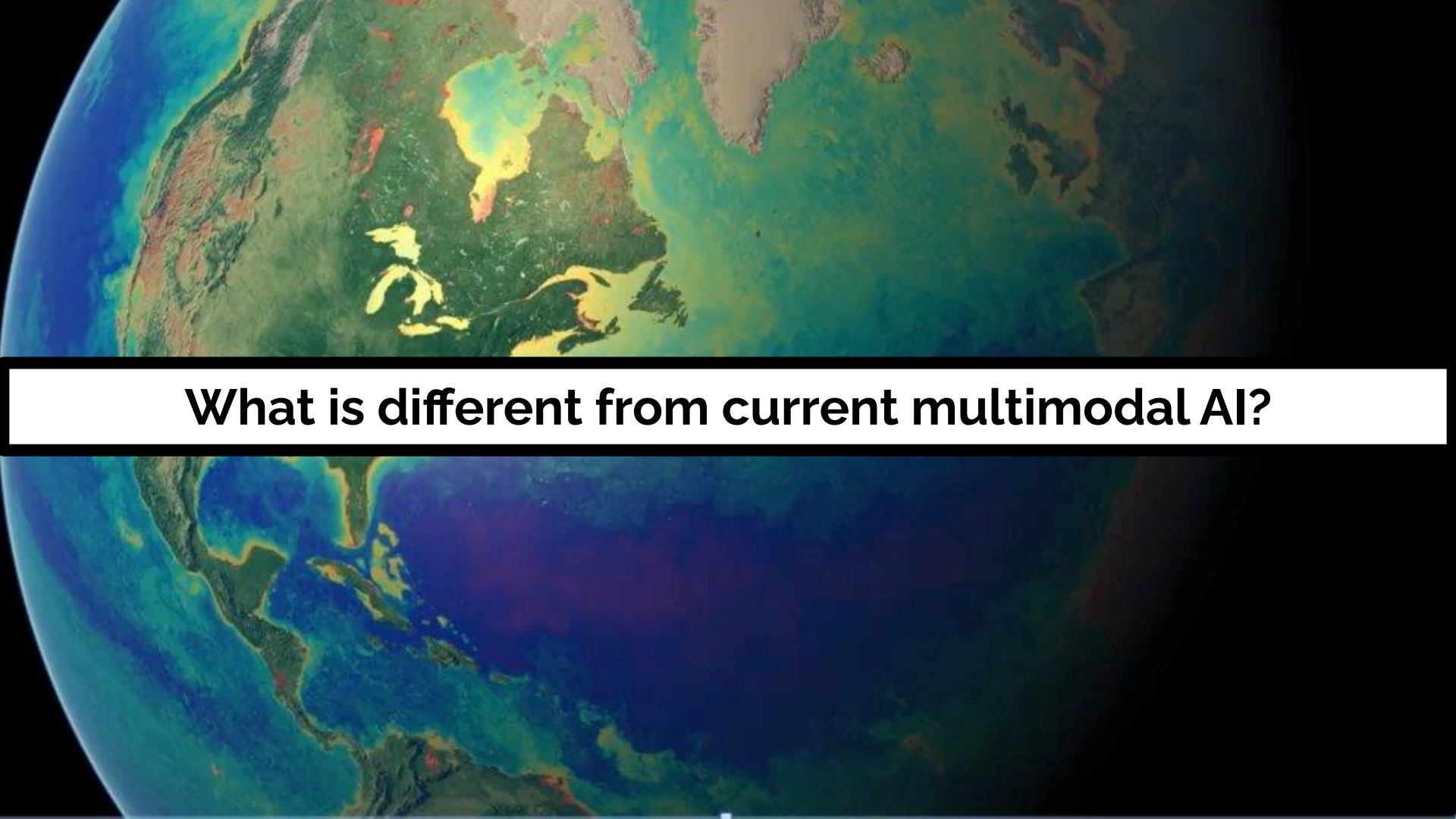
i.e. *morphology*:
the size, shape,
and structure of
organisms



Knowledge and structure

i.e **phylogeny**: the representation of evolutionary history and the relationships between groups of organisms



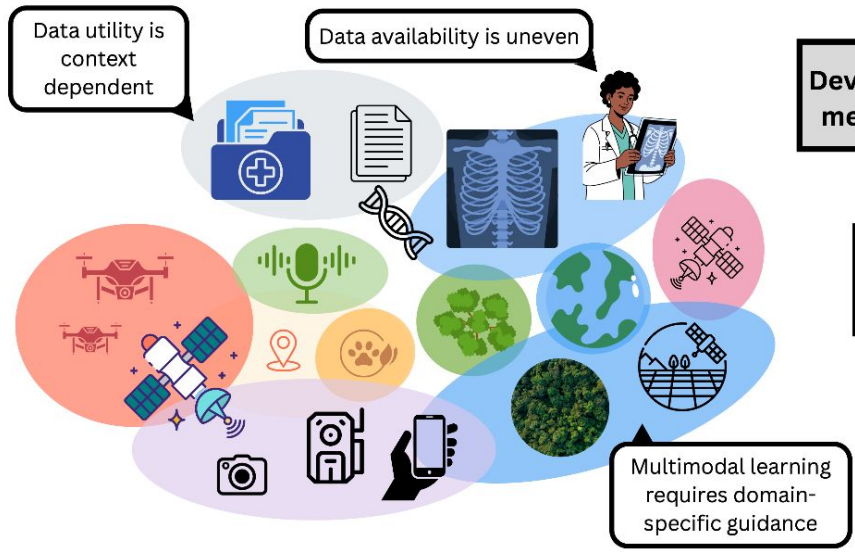
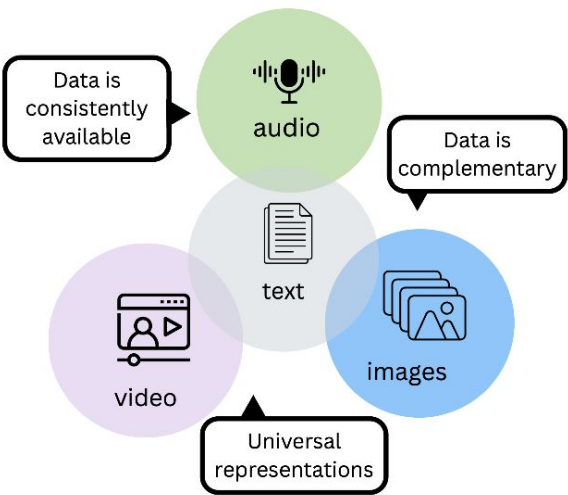


What is different from current multimodal AI?

Multimodal Benchmarks

Real-World Multimodal Datasets

Call to Action



Develop novel methods to meet domain challenges

Center context-aware evaluation

Diversify multimodal benchmarks