Prototype-Based Methods in Explainable-AI and Emerging Opportunities in Geosciences

Anushka Narayanan[1,2] and Karianne J. Bergen[1,2,3]

[1]Data Science Institute, [2]Dept. of Earth, Environmental and Planetary Sciences, [3]Dept. of Computer Science

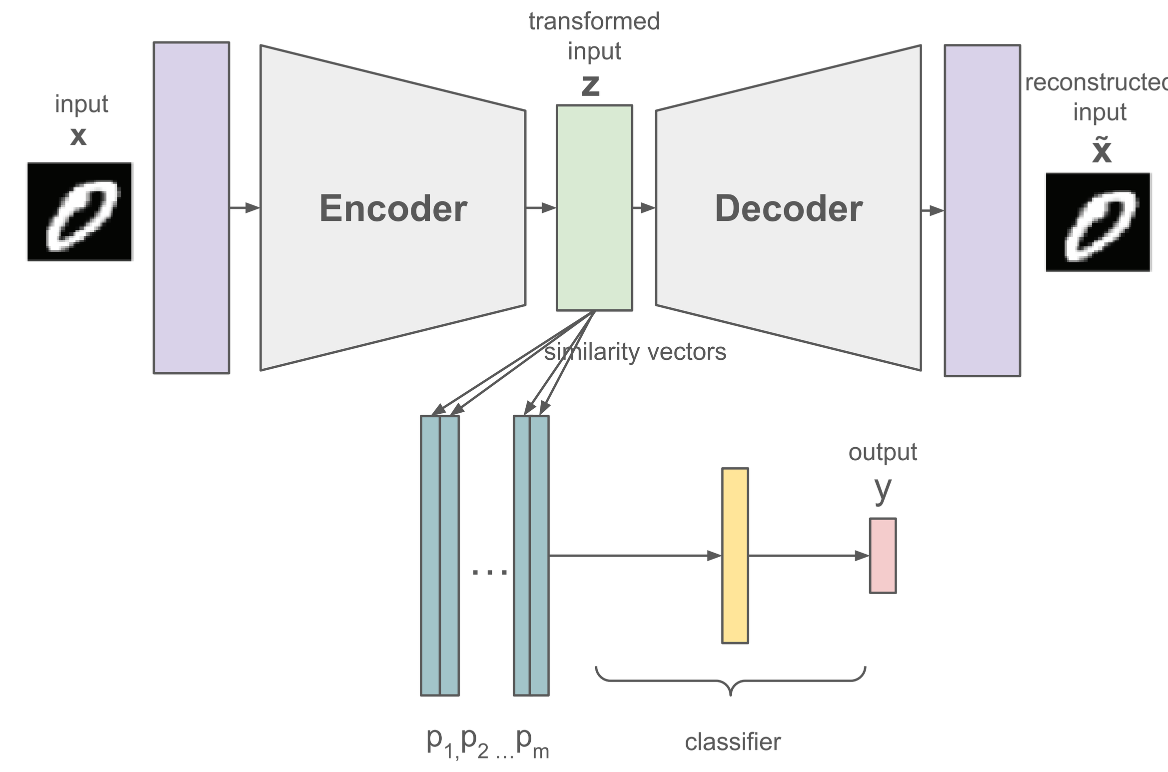
Brown University, Providence, RI, USA

Contact: anushka_narayanan@brown.edu

Motivation: Why Prototype-XAI for Geosciences?

- Prototype-based methods are (1) **intrinsically interpretable**, (2) produce predictions & explanations by comparing data with "prototypical" instances (learned examples representative of the training data).
- **Position:**
  - Prototype-XAI offers an under-utilized alternative to *post-hoc* methods
  - Prototypes offer reasoning via inspection of **similarity to prototypes** (typical features / patterns in the data) – mimics the human reasoning process
  - Prototype methods **show potential for geoscientific learning tasks**
- We highlight **differences between geoscientific datasets and the standard benchmarks** used to develop XAI methods, and discuss how specific geoscientific applications may benefit from modifying existing XAI methods

General Prototype-XAI Architecture



Case Studies: Prototype methods and their relevance to Geosciences

Development and Visualization

Image-Sized Prototypes [1]

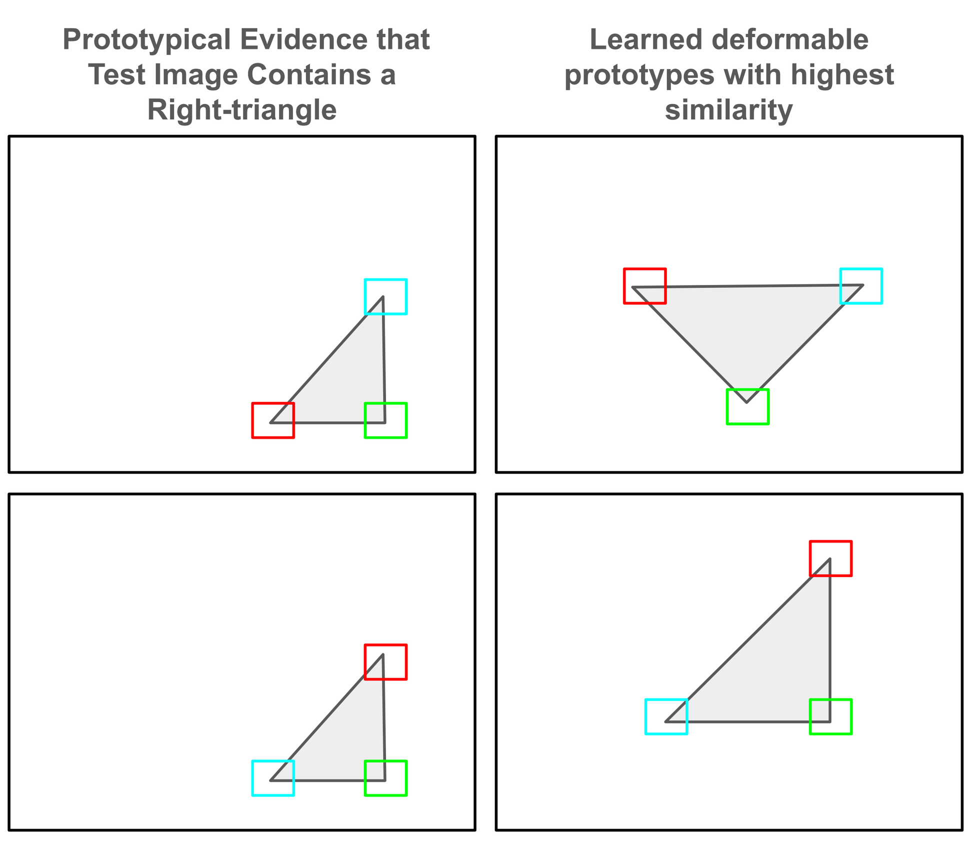
- **Explanation:** This input image resembles a learned prototypical image of the target class.
- **Geoscientific use-cases: climate phase prediction, dimensionality reduction**

Patch-based Prototypes [2]

- This specific patch in input image resembles a learned prototypical local patch / pattern in a training image
- **generating local spatial feature patterns**

Spatially Deformable Prototypes [3]

- This organized cluster of prototypical patches resembles a set of prototypical patches within an image of the target class. (see Fig below)
- **organization of feature patterns, detecting multi-scale feature patterns**



Types of Prototypes

Multi-Variable Prototypes [4]

- This multi-variable input shares similarities with single variables prototypes and their relationships with each other.
- **multi-forcing classification, multi-variable feature attribution, multi-spectral imagery classification**

Sequential Prototypes [5]

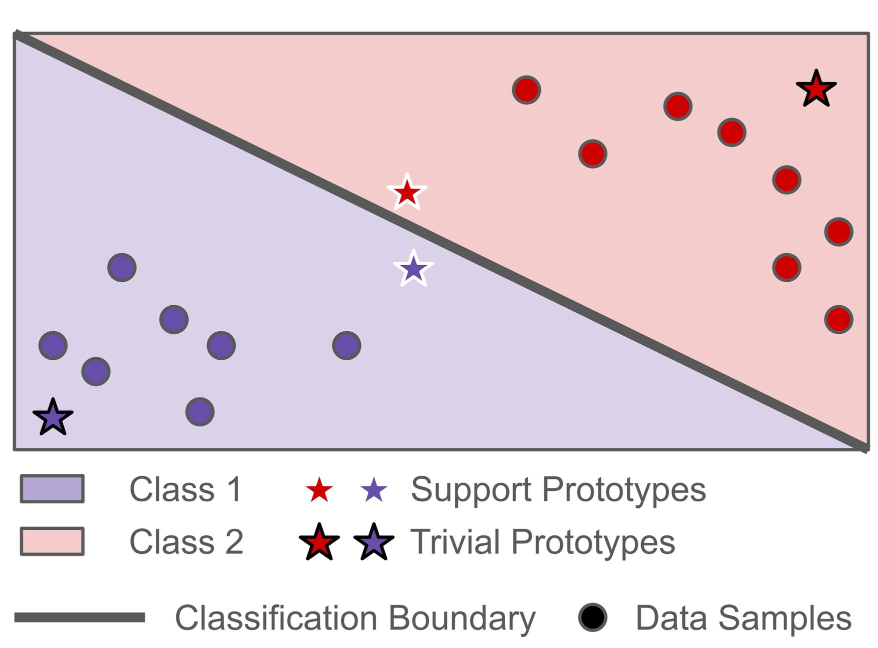
- A window of this input sequence resembles a prototypical window sequence in the target class.
- **temporal forecasting, climate prediction**

Anomaly Detection Prototypes [6]

- This input sequence significantly differs from prototypical sequences in the training data.
- **anomaly detection, extreme event detection**

Trivial and Support Prototypes [7]

- This input may resemble prototypes representative of a target class / boundary. (see Fig below)
- **outlier detection, extreme event detection**



Prototypes for Various Learning Tasks

Decision Trees for Learning Tasks [8]

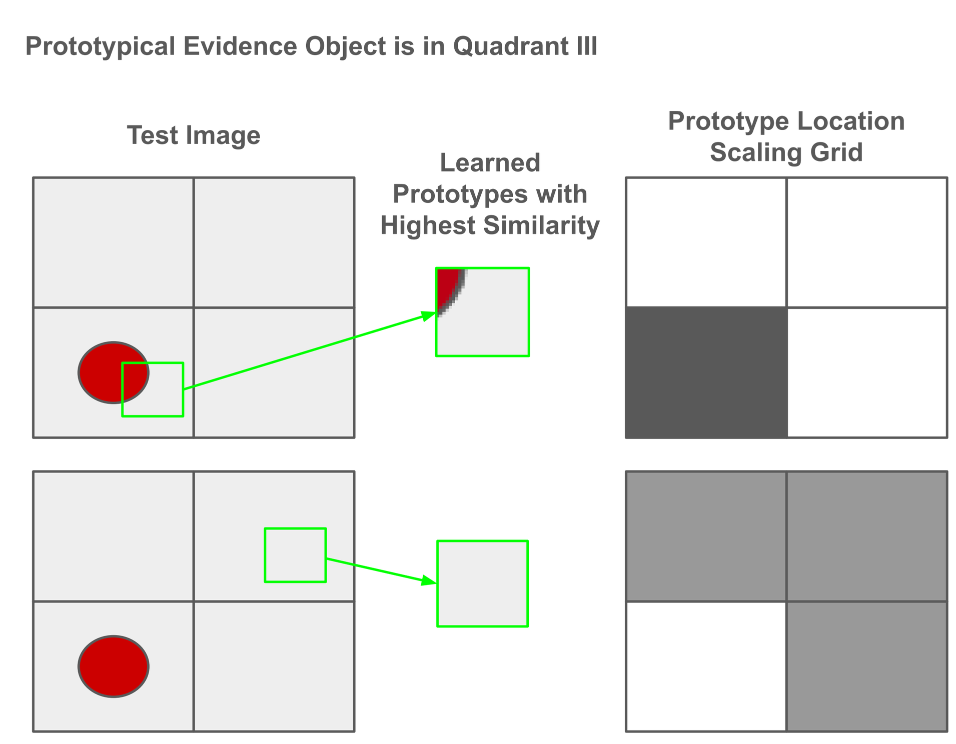
- The prediction is derived from a tree-like reasoning process on whether the input contains similarities to certain prototypes.
- **multi-variable, extreme event prediction**

Learning Tasks with Negative Reasoning [9]

- This input image contains similar prototypical parts of the target class and does not contain prototypical parts from another class.
- **climate prediction and classification, extreme event prediction**

Learning Tasks with Location Scaling [10]

- This input image contains similar prototypical parts of the target class only in specific regions of the image. (see Fig below)
- **generating spatially relevant feature patterns**



Discussion

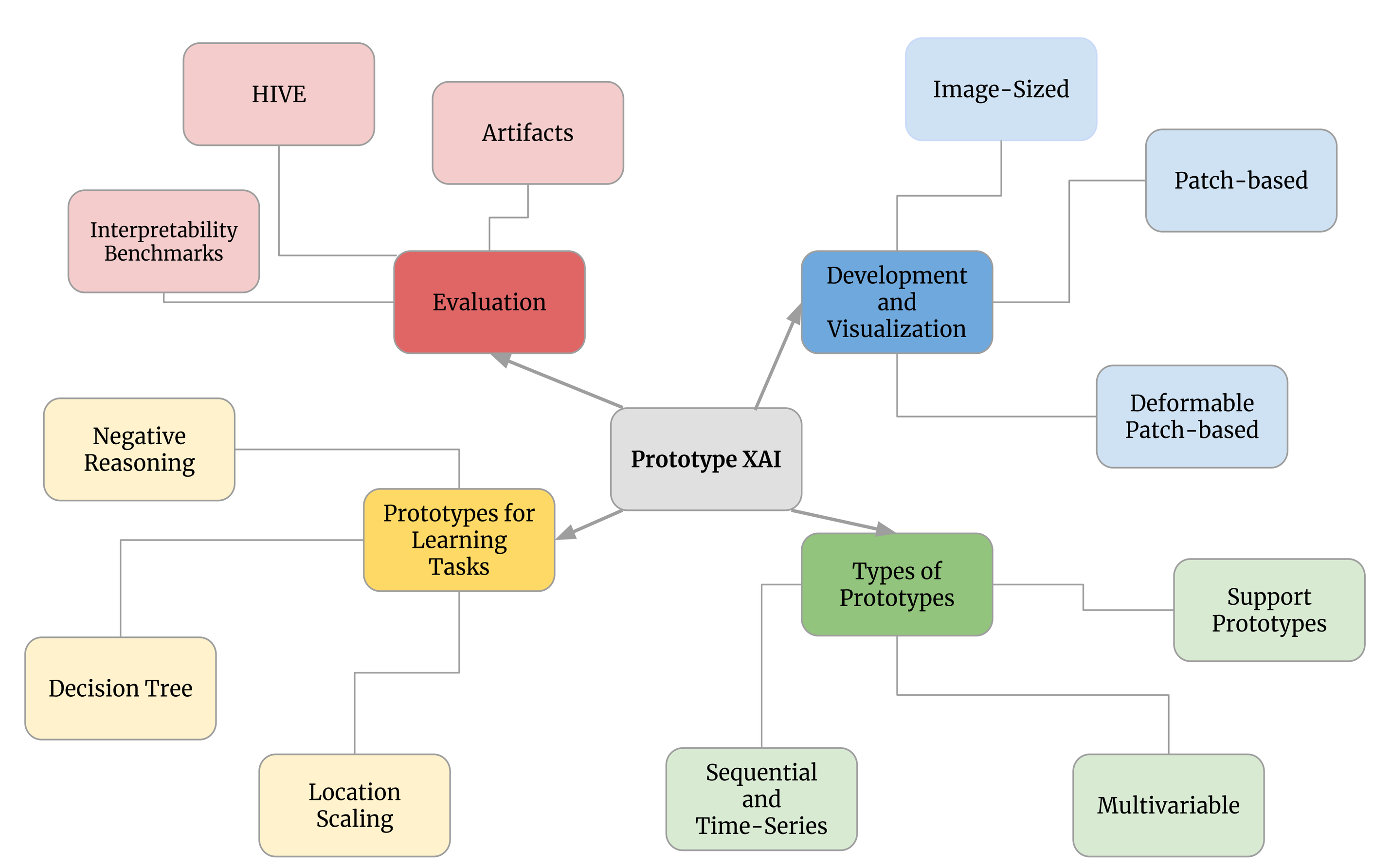
Geoscientific data **has unique characteristics** vs. natural images & text data typically used in prototype-based XAI research that require **particular attention** when using / developing prototype-based techniques.

Opportunities in Geosciences

- Image sized prototypes alternative to PCA-derived global climate modes vs patch prototypes for localized feature patterns
- Location- and channel-specific prototypes with sequential prototypes for interpretable spatiotemporal forecasting (ongoing!)
- Prototype anomaly scores, support prototypes for identifying subtle patterns, domain shift, extreme events
- Multi-variable prototypes for identifying salient individual and combined spectral channels for remote sensing

Limitations and Pitfalls of XAI

- Need **simple, non-duplicative** prototypes, e.g. scientist-in-the-loop pruning
- **Evaluation of robustness and reliability** to avoid artifact-induced explanation variability



Acknowledgements: This work was supported by the SciAI Center, and funded by the Office of Naval Research (ONR), under Grant Number N00014-23-1-2729.

References: [1] Li et al (2018). [2] Chen et al (2019). [3] Donnelly et al (2022). [4] Ghosal et al (2021). [5] Ming et al (2019). [6] Li et al (2023). [7] Wang et al (2023a). [8] Nauta et al (2021). [9] Singh et al (2021). [10] Barnes et al (2022).