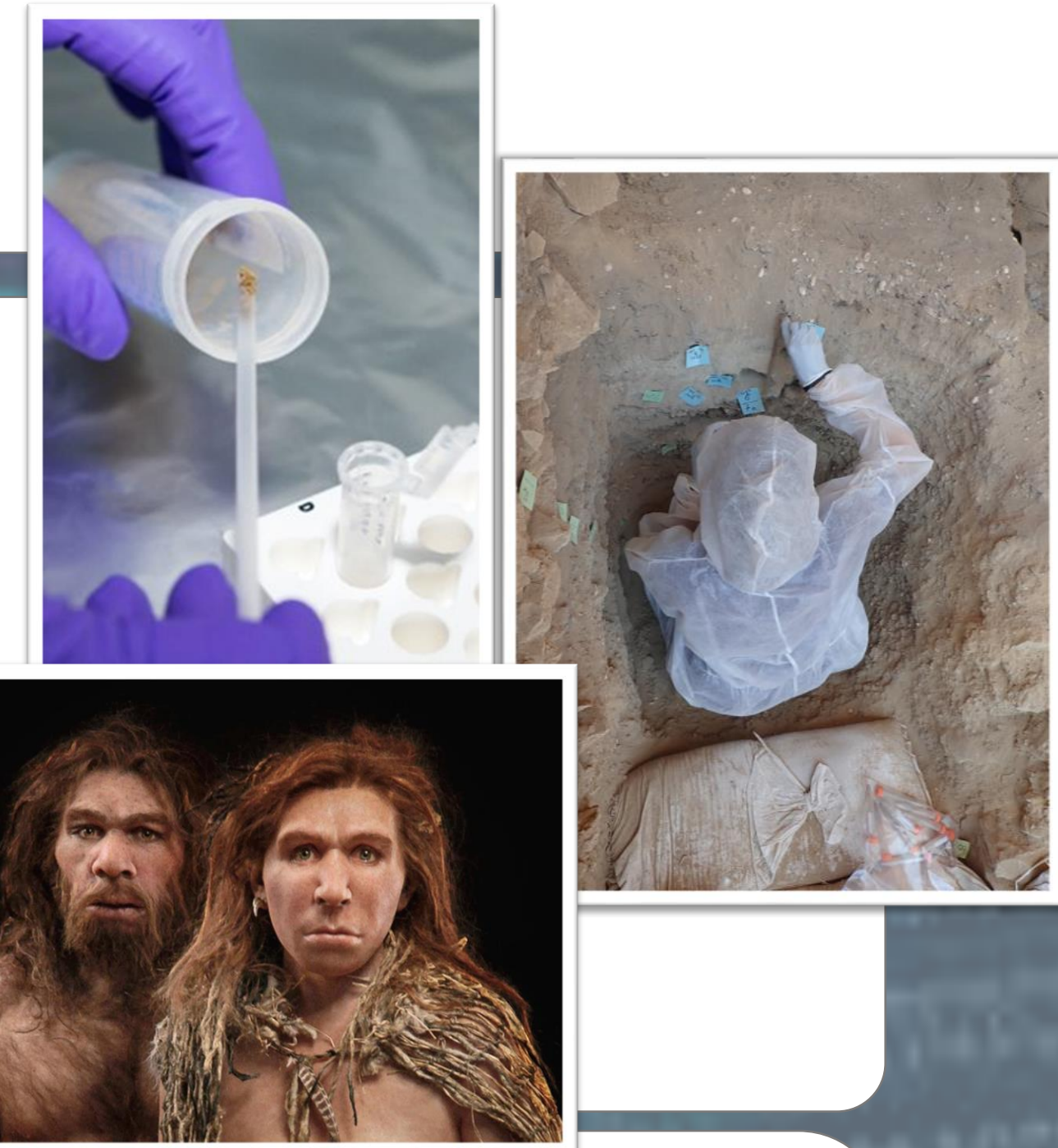


exML – An Explainable Maximum Likelihood Tool for Proportion Estimation in DNA Data

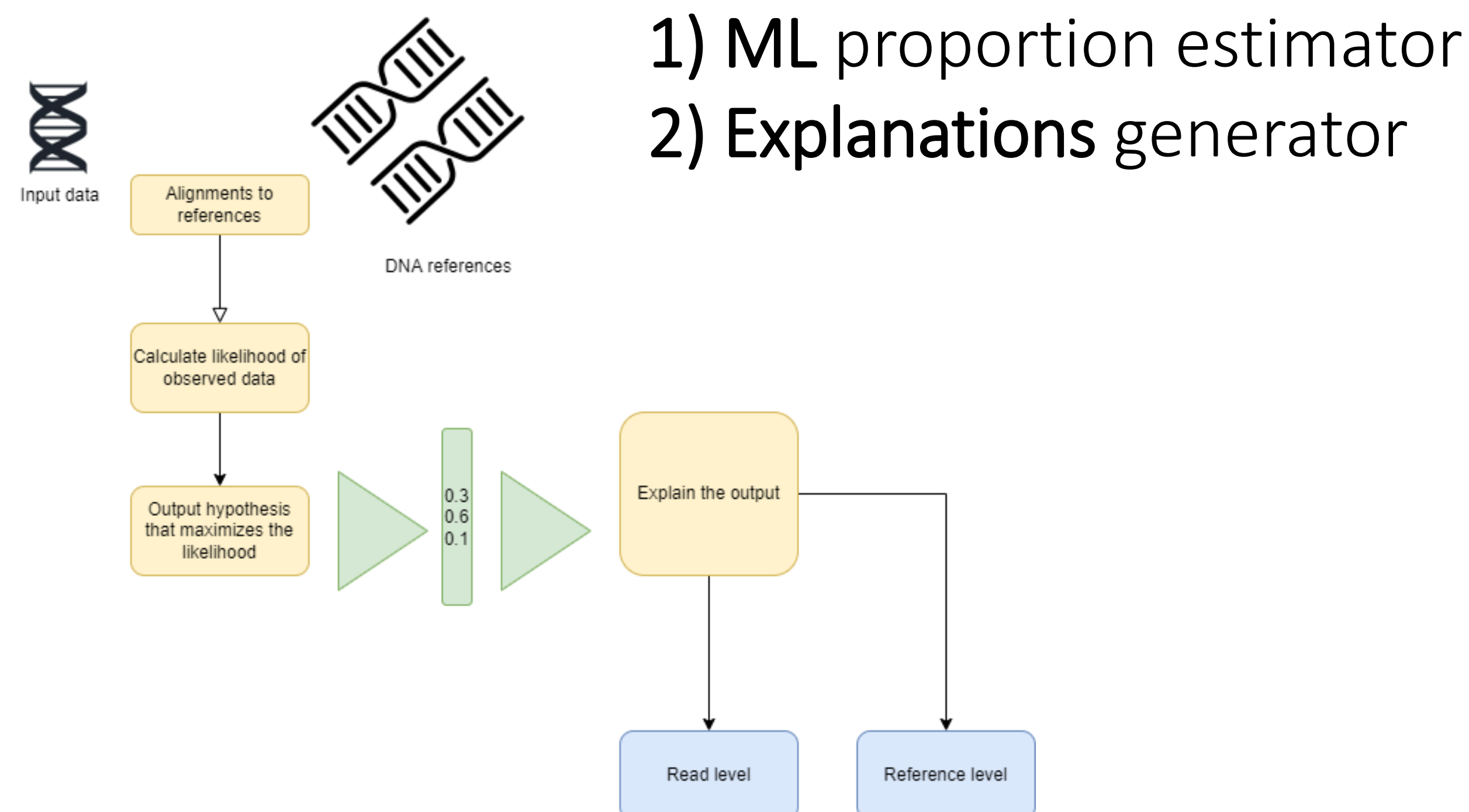
Amit Bergman, Viviane Slon, Daniel Deutch

Abstract

- We use maximum likelihood algorithm to **estimate proportions of species** in genomic datasets
- We attach **explanations of multiple flavors** to the output to assist the data analysis
- The scenario on which we demonstrate our work is datasets of **ancient DNA of hominids**

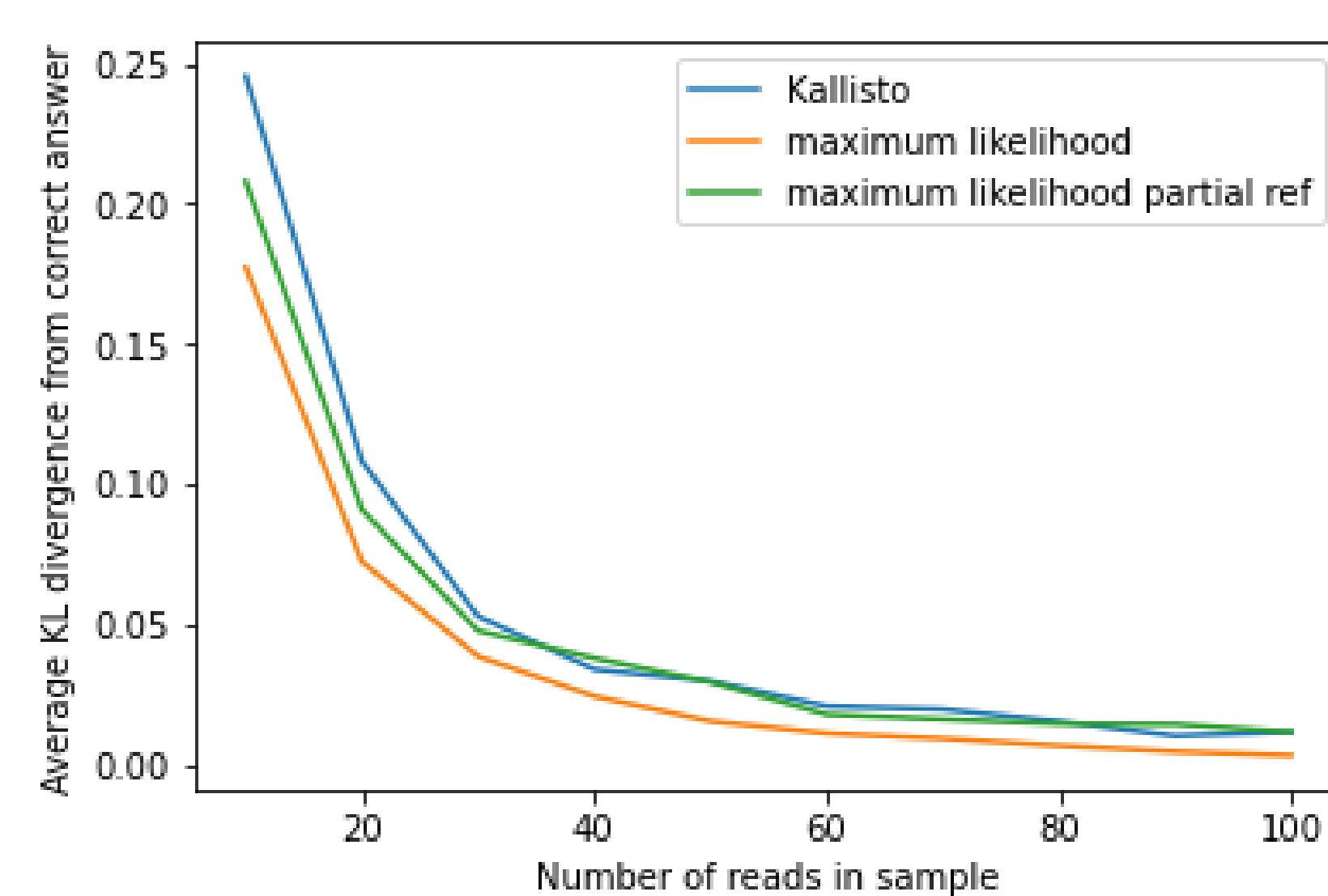


System design

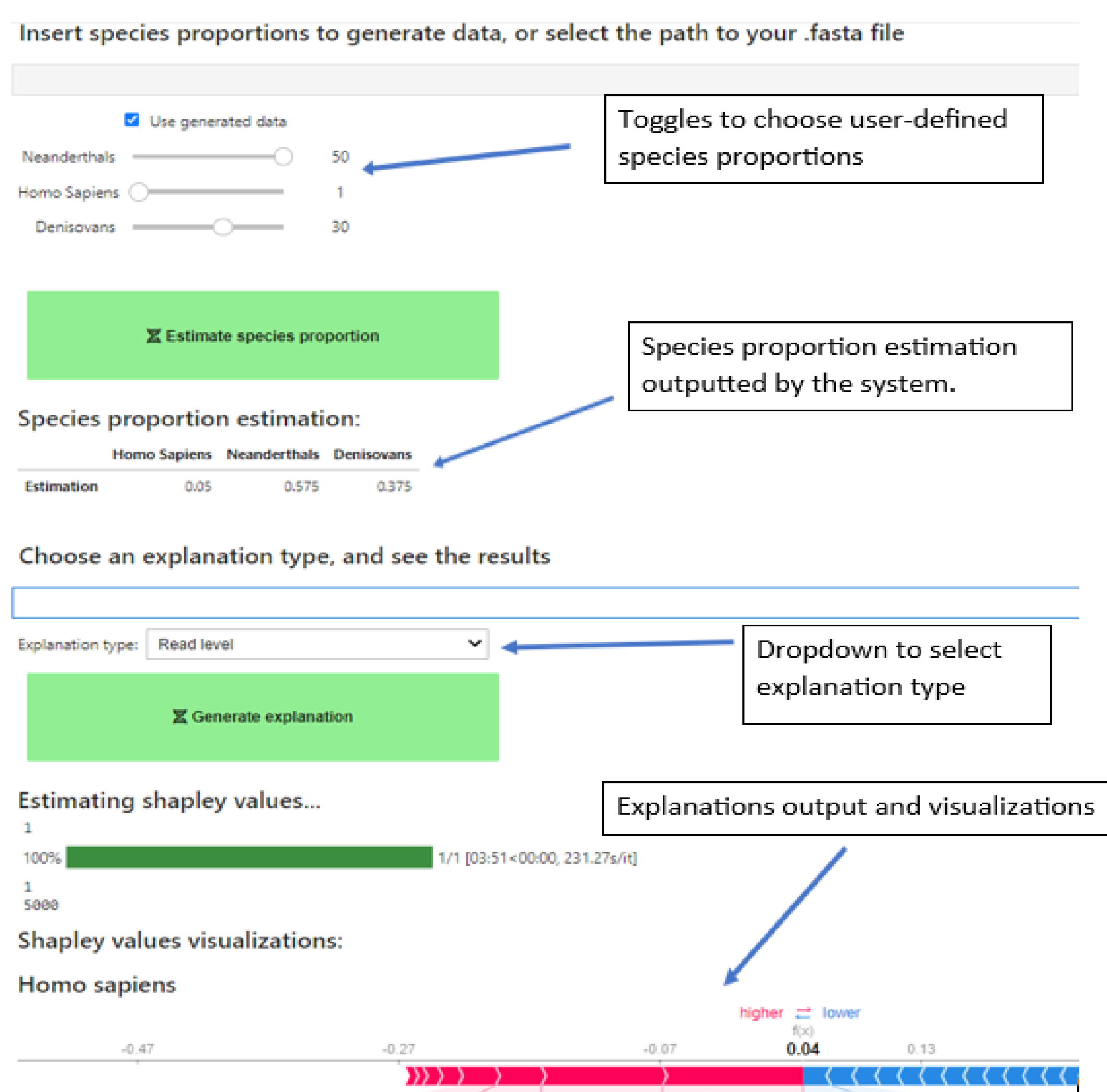


Accuracy

- We compared our work to an existing method² to **classify ancient DNA**
- Attained **better accuracy** on simulated data



GUI



GUI of the system

The GUI allows to:

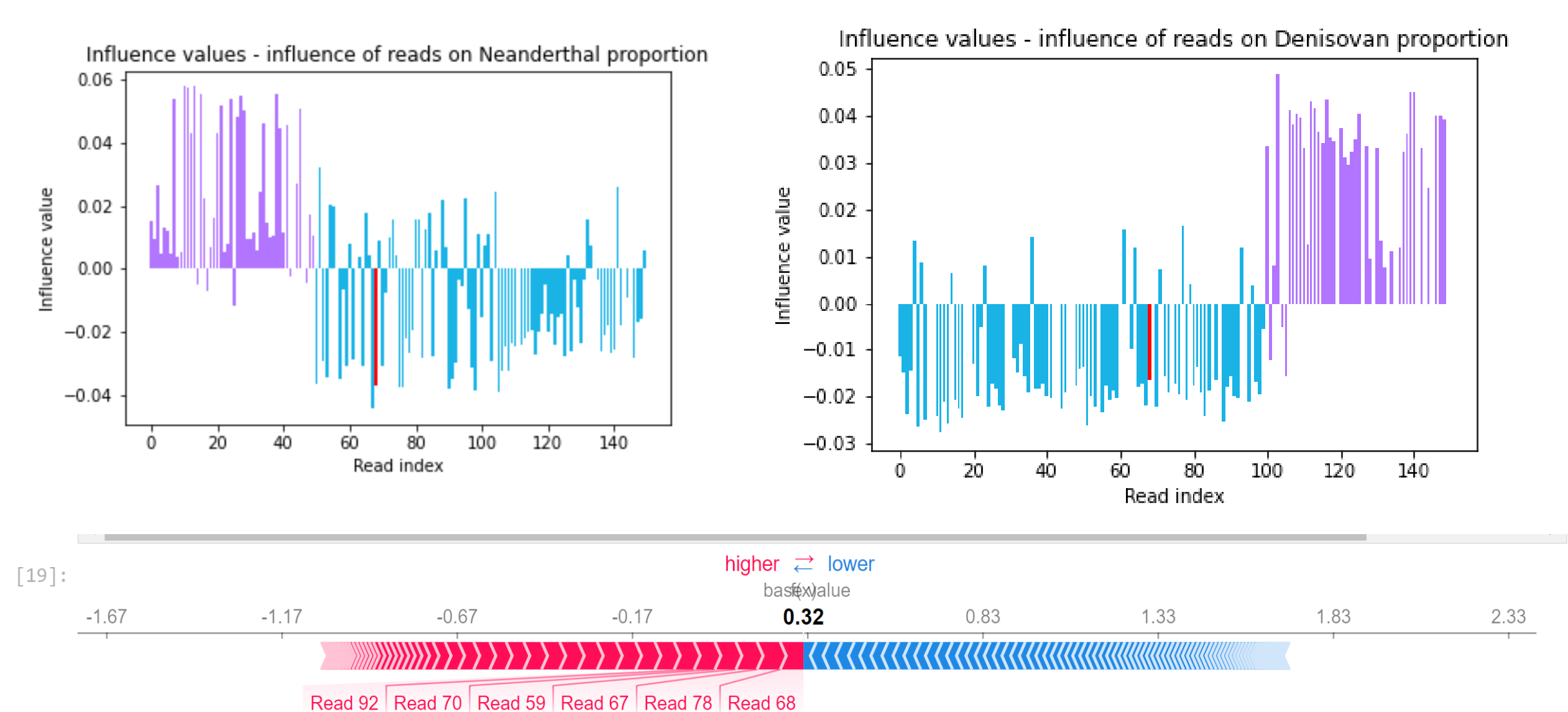
- Provide or simulate genomic dataset
- View proportion estimation on that data
- Choose explanation method and visualize it

Explanations

The system provides the user explanations that shed light on what led it to the output

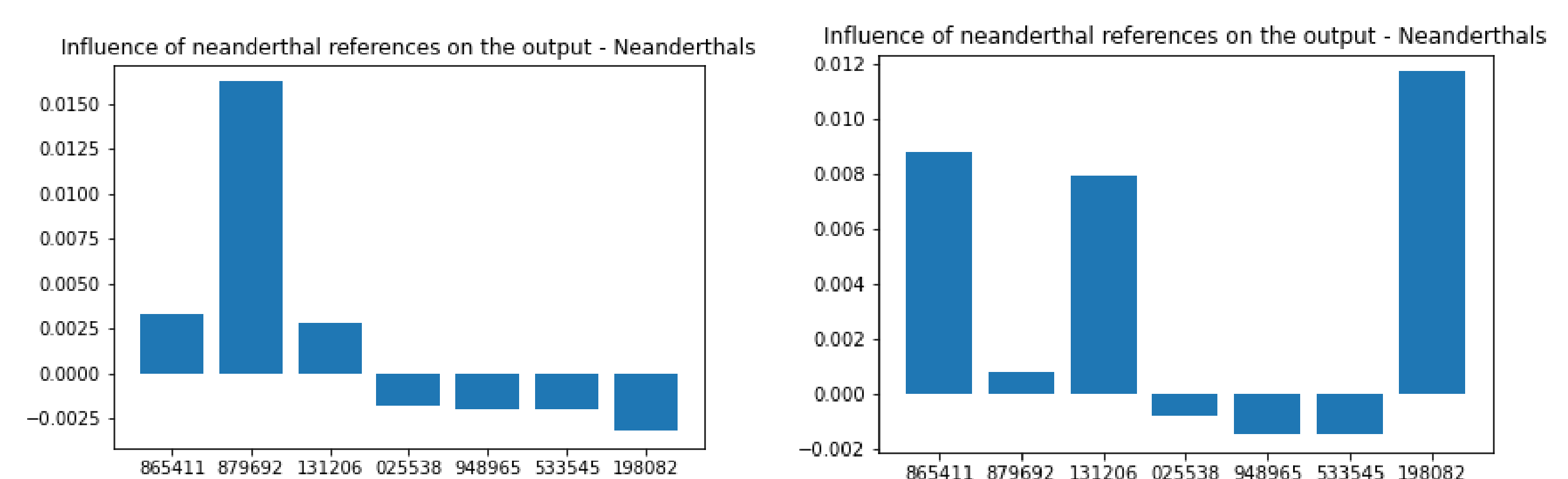
Data-level explanations

Quantify how each datum from the dataset influences the output + counter factual



Reference-level explanations

Quantify how each reference from the reference dataset influences the output



Code availability

- The system, GUI, sample datasets and use case notebooks are available under [this github repository](https://github.com/bergmanam/exML).

Future plans

- Apply the tool on datasets from other domains
- Improve theoretical results that show the quality of the explanations

References

- Slon et al., *Science* 2017
- Vernot et al., *Science* 2021