What did you learn, in your individual session, about static analysis for ML and the pynblint tool?

During the individual session with Luigi, I learned how to use the tool Pynblint, which is a static analysis tool for Jupyter notebooks. I filled out a form with my programming habits, specially related to computational notebooks and I tried Pynblint by analyzing some of my Jupyter notebooks.

Will pynblint be useful to you in your WASP PhD project? Why or why not?

I think Pynblint will be useful for my WASP PhD project because I use Jupyter notebooks in my research and it provides with a nice toolbox to make sure our notebooks are clean, organized and reproducible. Specially when sharing minimal examples of our machine learning (ML) methods and for interactive model evaluation. Particularly, I work on machine learning for molecular science and this tool will allow me to increase the quality of my notebooks before I share, for example, my interactive notebook for visualization of molecules generated with our ML methods.

Ideas for how the tool could be improved?

I think the tool is very useful already and Luigi mentioned a new update will be available in the coming weeks. However, I think it would be useful that this tool could be integrated in development environments such as VS code PyCharm or the actual Jupyter notebook.

What do you see as the limits for static analysis tools in ML? For code, models, and for data?

I think the main limitations of static analysis for ML are related to the fact that is very difficult to predict (or even approximate) the actual behavior of a ML model without running its implementation. Indeed, we often we get surprised by the performance of these methods. Moreover, the performance is heavily dependent on the data that is used, complicating things even more. However, I think we can still benefit from statistical analysis for ML. Some options that come to mind are good practices and suggestions when designing and training models (for example suggesting using the state-of-the-art alternative when one option is clearly superior to the rest) or suggesting standard pre-processing procedures when available for the data we are using. Wrapping up, it is challenging to use statistic analysis for ML, but I still think there a lot potential in it, specially with integrated tools in development environments.