Imagine you're a digital artist. Over the years, you've developed a distinctive visual style - your own palette, composition, and brushwork. One day, you discover that a popular AI image generator is producing visuals eerily similar to your work. They're not exact copies, but something about them feels undeniably familiar. You begin to suspect that your art was used to train the model. But how can you know for sure?

Today, no reliable tools exist that allow creators, publishers, or institutions to verify whether their content was used to train large generative models such as ChatGPT, DALL·E, or Stable Diffusion. With high-profile lawsuits emerging (e.g., Getty Images vs. Stability AI, New York Times vs. OpenAI), it's clear that the ability to audit the origins of training data - even post hoc - is urgently needed.

This project aims to develop practical and scalable methods for determining whether - and to what extent - a specific dataset was used to train a large generative model. Our central hypothesis is that training data leaves subtle statistical traces in the model's behavior. If we can amplify and aggregate these traces, we can detect the use of our data in model training even when it is not memorized verbatim.

In Task I, we will develop techniques that go beyond binary yes/no answers and instead estimate how much of a dataset was used during training - crucial for legal contexts that consider the quantity and importance of reused material. In Task II, we will strengthen these weak signals, making it possible to detect small datasets - such as a single artist's portfolio - with high reliability. Task III will address the challenge of lacking known "non-member" data by generating synthetic reference sets or statistically correcting for distribution mismatch. Finally, in Task IV, we will extend our techniques to multimodal generative models that process both text and images - systems that are increasingly prevalent in today's AI landscape.

This project is both scientifically and socially relevant. It supports data owners, enables independent audits, and provides tools for regulators and legal professionals. In the long run, it contributes to a more transparent and responsible AI ecosystem. All developed methods will be released as open-source tools to ensure accessibility and real-world impact - empowering creators, journalists, and data rights advocates alike.