Hannah Busch, An Artificial Eye for Palaeography. Applying Deep Machine Learning for the Study of Medieval Latin Script. Lightning Talk, 13th Schoenberg Symposium on Manuscript Studies in the Covid-19 Age, Nov 18–20, 2020. DOI: 10.5281/zenodo.4302210

Hi!
I'm Hannah Busch a PhD researcher at the Huygens Institute for the History of the Netherlands in Amsterdam. And this over there is my research project. In this presentation I will talk about my ongoing PhD research where I investigate the application of deep machine learning for the study of medieval Latin paleography.

Imagine you're browsing through the amazing collections of digitized medieval manuscripts available via the world wide web and then a manuscript touches your attention. You wonder where and when it was written didn't you come across something similar a couple of weeks ago. You want to find out more so you check the manuscripts descriptions provided by the collection there is no information about dating or localizing the origin of the manuscript. You try to remember where you saw that other manuscript and maybe you even start browsing through collections you have accessed over the past weeks, month, years, or compare images by using the IIIF manifests. But that takes way too much time and random keyword searches also don't give you any significant results.
Wouldn't it be wonderful to have an image reverse search for medieval Latin manuscripts? With one drag and drop you could search the entire IIIF universe of digitized manuscripts and receive images of manuscripts showing similar script features in return. In addition it would give you information how similar the two script samples are.

But how do we train an artificial paleographic eye? To be able to train an artificial intelligence system that can assist manuscript scholars in doing their research we need to provide the computer with the expert's knowledge: the so called ground truth. The ground truth consists of training data in form of images and labels that carry information about the place and date of the manuscript's origin. With the training data the artificial intelligence is supposed to find patterns in the script that are unique to their origin.

For the training processes pages will be cut into snippets and fed to the machine. The ground truth labels serve two purposes. On the one hand they give feedback to the machine as part of the supervised learning approach, on the other hand they can help the babysitter of the training process—which is me—to get a better understanding of how the system might work. Where it works well and where it fails.

As a result of the training process a script fingerprint for each training sample will be stored. But beside experimenting with the potentials of artificial intelligence my research is concerned with the following questions: Does the available material meet the requirements for the application of artificial intelligence? For example, does the data correspond to the fair principles? But also which patterns does the machine recognize? Here we can think of the good patterns, but also about bed patterns like this barcode finder that we accidentally developed. And finally, can the artificial intelligence system learn features that correspond to detailed studies of script characteristics?

My research is part of a larger research project called “Digital Forensics for Historical Documents. Cracking cold cases with new technology.” at the Huygens Institute and the International Institute of Social History in Amsterdam. The project is funded by the Royal Netherlands Academy of Arts and Sciences. It will run until 2022.

Thank you for watching this presentation, and if you are interested in more projects at our institute check out the talks from my colleagues Evina and Mariken.