Learning Latent Data Structure from Observations

Imagine obtaining a copy of a mysterious document, such as the Voynich manuscript ([1]; Figure 1). Its contents form lines of repetitive symbols and resemble text in a foreign, maybe fictional language. After a closer inspection, you form hypotheses about the letters, and identify sequences which we call words, as a basis for a further analysis. But could you be certain about them? Lets assume many handwriting styles present in the document. Would you be able to map them to a common alphabet? This is clearly a task for modern machine learning, which should take advantage of the volume of data, and identify recurrent patterns automatically.

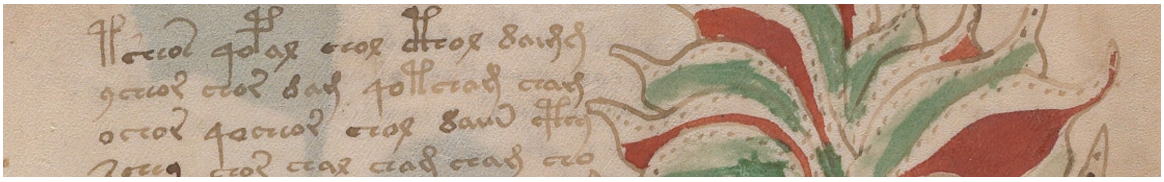

Figure 1: Excerpt from the Voynich manuscript. It resembles text, but the alphabet, text, and meaning remain a mystery.

In order to do so, you decide to scan the documents. However, you soon run into a problem: modern machine learning systems, including optical character recognition systems (OCR), are built by presenting to a computer vast amounts of *labeled* data. In order to use artificial intelligence, you are forced to laboriously type in the transcriptions for numerous pages of text [2]. Clearly, not being able to reliably read the text in the first place, you cannot label the data. Even though some patterns are visible to the naked eye, we lack the theoretical understanding of the problem, and specific algorithms that should follow. There has to be a better way!

This project is here to help. We are working on algorithms that, based on large collections of data, discover the latent structure and bring forth hypotheses about it. In our example, the algorithms can analyze scanned lines of text, in an attempt to derive structure and form hypotheses about characters, words, or maybe even grammar. We will develop ways to steer the algorithms with little domain knowledge, which comes natural to a person, but not to the algorithm. This includes specifying the expected number of characters, or enforcing a heavy-tailed zipfian distribution of discovered word units, known to be prevalent across most natural languages.

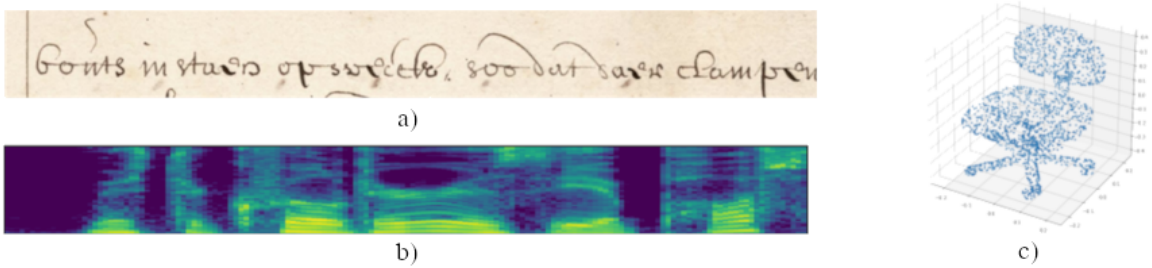

Figure 2: Examples of data modalities present in our work: a) handwriting [3], b) speech [4], c) point clouds [5]

Apart from handwriting, our algorithms will be applicable to other data types which we are currently working on ([3–5]; Figure 2), popular in the domain of machine learning. We will be able to analyze audio signals and discover phonemes - the basic "units of speech", analyze the contents and actions present in video clips, or even segment objects represented as point clouds. Such point clouds are collected in the real world with depth cameras, and our algorithms would discover sub-objects like furniture parts.

Our research effort will bring both theoretical and practical contributions. On the theoretical side, we want to understand what is needed for machines to automatically understand the structure of the data: discover meaningful entities, model their dynamics and the relations between them. On the practical side, we will ensure that the emerged data representations allow applying existing, proven machine learning models, with little amounts of manually labeled data.

1 References

[1] "Voynich manuscript," Dec. 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[3] J. Chorowski, N. Chen, R. Marxer, H. J. Dolfing, A. Łancucki, G. Sanchez, S. Khurana, T. Alumäe, and A. Laurent, "Unsupervised Neural Segmentation and Clustering for Unit Discovery in Sequential Data," in *Workshop on Perception as Generative Reasoning, NeurIPS 2019*, Vancouver, Canada, Dec. 2019.

[4] J. Chorowski, R. J. Weiss, S. Bengio, and A. v. d. Oord, "Unsupervised Speech Representation Learning Using WaveNet Autoencoders," *IEEE/ACM TASLP*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[5] M. Stypułkowski, M. Zamorski, M. Zięba, and J. Chorowski, "Conditional Invertible Flow for Point Cloud Generation," in *Published in Sets & Partitions Workshop at NeurIPS 2019*, Vancouver, Canada, Dec. 2019.