Chapter 1

Introduction

> The great certainty of the natural sciences in comparison with the study of psychology or consciousness comes exactly from the fact that they choose for their object what is strange, while it is almost contradictory and even absurd to try to choose for one's object what is not-strange.
>
> The Gay Science, Friedrich Nietzsche

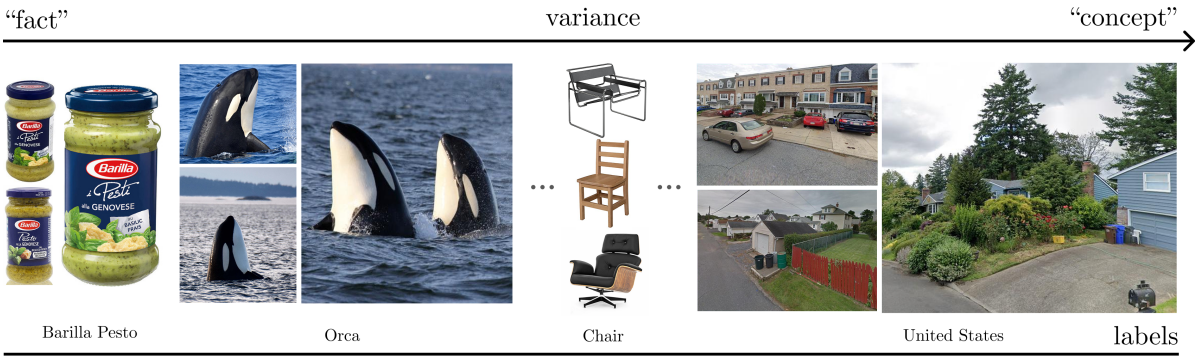
1.1 Philosophical Introduction

The divide between the sciences and the humanities appears foundational in the way institutional knowledge is produced. The one is often asserted as objective or quantifiable while the latter is often cast as subjective or perceptual (phenomenal). Yet for both the goal is to capture what is real, or at least to contour what exists, offering different methodologies with their respective limits. In fact, both can be seen as moments in the production of human knowledge. Human knowledge often starts as a journey from subjective observations, where in an effort of gradually trying to ground them, a form of quantifiable abstraction emerges tasked to provide a guarantee of generality. In the human sciences, one such tradition tries to explain such observables (e.g., social behavior) by trying to abstract their elements into terms of mathematical equations and to draw a parallel between their interactions to those of physical systems [Macy et al., 2024]. A typical example is the German sociologist Nicolas Luhman, who tried to conceptualize social dynamics as dynamics of abstract (mathematical) systems [Luhmann, 2013]. While these approaches have their merit, their migration to such abstractions limits their conclusions from ever bridging back the gap to reality, that these abstractions required in order to function. Their foundational problem is that

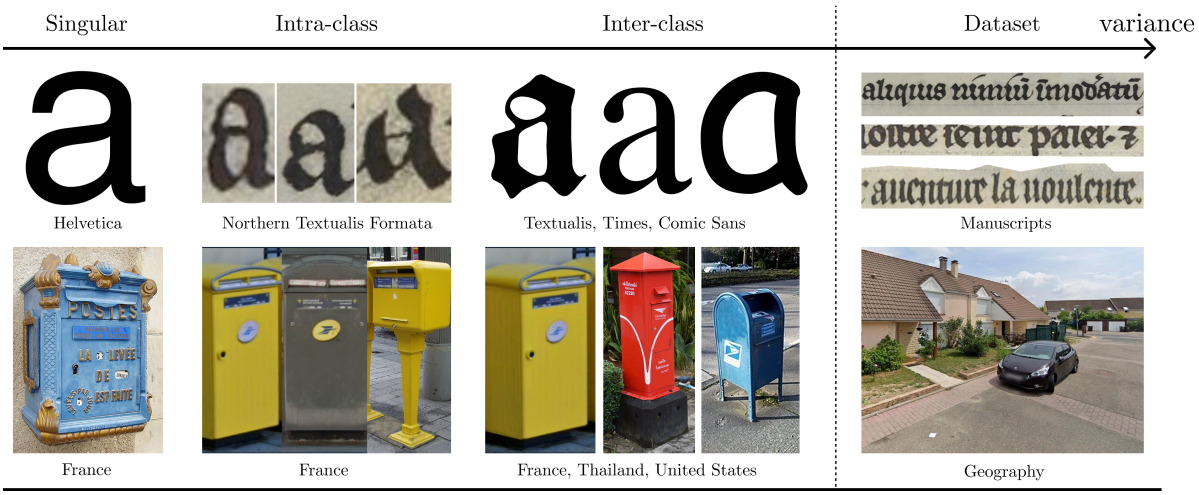
instead of focusing in understanding the *anthropologic complexity* of their input data, i.e., the way their data is both perceived and constructed by humans, their goal is more to capture their *Kolmogorov complexity*, trying to form a compressed representation of them as a physical world that can claim to exist outside humanity. But, to paraphrase Prof. Dawkins, in order to understand that part of the physical reality that is human, models need to be as "drunk" on symbols as they are on "facts" [Dawkins, 2000]. For example, in order for a visual system to understand the "nature" of a church, it would need to incorporate the factual existence of god inside culture irrelevant of whether it is factual in respect to physics. Thus, instead of trying to reduce the humanities into physics, what if we could elevate physics into the humanities? One way would be to create a physical system that has the power of understanding and manipulating two very central modalities of human culture, images and text, just like humans do. Performing such a migration of human perception into machinic perception, can in fact offer a form of *objective subjectivity*. While models are subjective as they depend on the data, architecture and task they were trained on, they remain objective as all their inference is operated via weights and computational processes that can be replicated and analyzed through digital computers. By making technology aware of both the processes of design and emergence that govern human societies, such a technology can allow for the first time in history a proper automation of the production of human knowledge. However, why would someone want to involve digital computation in the production of human knowledge in the first place?

1.2 Motivation

A great part of human knowledge, involves the recurrent synthesis of external information into archives. Yet, as these archives get large, it becomes increasingly hard for humans to manually produce knowledge and ground it at scale. Exploratory Data Analysis [Tukey, 1977], has been a common technique that researchers in the digital humanities have been using for almost 50 years to understand data of a non-human scale. However, their observations could only be as good as their models. This endeavor has been historically hard, as researchers really failed to properly formalize the way humans perceive the world. Yet, after tremendous progress in statistical modeling, machine learning, and optimization, self-supervised, generative, deep learning models are slowly turning the question of aligning a model's perception to human perception into a practical one. Modern approaches are graduating from performing retrieval tasks to meet the challenge of performing reliable mining and summarization of unseen



(a) **Signifier's Paradox.** While a category correctly describes a part of the visual world what it describes may vary significantly.



(b) **Hidden Structure in Visual Variation.** The visual appearance of the signified images may hide distinct structure of visual variation, either behind or across classes.

Figure 1.1: **Motivation.** While labels may be consistent the visual appearance of the signified images, may vary significantly (**top**). The variance of images can be attributed to inter-class variation of a given sub-category or intra-class variation across categories (**bottom**).

information, a fundamental skill required for producing new knowledge. When this involves the high-level abstract representation of text, automating search and compilation of information have been two of the biggest computational breakthroughs of the last 20 years. However, when it comes to low-level forms of information such as images (and sound), this task comes with increased degrees of freedom. For a system to automatically extract knowledge from such data, it is important that it is able to navigate both how humans perceive images in terms of sensory input (e.g., [Wertheimer, 1938; Blake and Zisserman, 1987; Biederman, 1987]), and in terms of labels (e.g., [Barthes, 1990; Rosch and Lloyd, 2024; Barriuso and Torralba, 2012]).

From Fact to Concept. The most established way that humans can abstract, share and communicate the structure of the visual world is to associate it with labels. A label is a signifier, a symbol or a text, that is inferred by instances of the signified, which it is meant to signify. For the process of signification to be useful for a specific task, labels need to effectively abstract or group visual variation. For example, in a text manuscript, the label of the character 'a' should be associated with all the different ways 'a' is inscribed. If this letter comes from a specific typeset, the association is trivial. However, in all other cases, deriving this association necessitates learning to properly organize its variation and discriminate it from other forms. As demonstrated in Fig. 1.1a, this brings us close to a paradox. While labels may correctly abstract the visual world, the visual variation that they capture may diverge significantly. To clarify this further let's pick the example of the *Barilla pesto*. As an industrial object it should carry almost no variation.[1] Of course the illumination, the sensor, the environment, or the background where it is captured may vary, but as an actual, signified object it is almost identical. In other words this *signifier's paradox* is especially relevant even if we had a perfect way of isolating and grouping all the salient parts that associate an image with its label. This creates a spectrum between labels that are more "factual" and labels that are more "conceptual". In the words of the media scholar McKenzie Wark, "a good fact is true for something in particular, while a good concept is slightly true about a lot of things" [Wark, 2023].

Functional Ontology of Categories. This is in fact a very old problem in the philosophy of categorization [Rosch and Lloyd, 2024; Murphy, 2004; Malisiewicz and Efros, 2009]. Arguing against the unchangeable form of Platonic metaphysics [Reeve et al., 2004] and its criticism through the rule-based definitions of categories of Aristotle's [Studtmann, 2024], Wittgenstein famously discussed in his philosophical investigations [Wittgenstein, 2009] that categories such as the word "game", one which he himself would use to describe the way language evolves, group instances of things that may share common properties, for example competitiveness or fun, but which are not necessarily consistent or discriminative across all members of its class.[2] Inspired by Wittgenstein, the cognitive psychologist Eleanor Rosch tried to show that human visual categorization is performed through what she called a prototype theory [Rosch,

[1] In fact this corresponds to a shape of Barilla pesto discontinued in 2022 for reasons of sustainability, which became so recognizable that the company added a warning label so that one could learn its new appearance after the packaging changed.

[2] Don't say: "There must be something common, or they would not be called 'games'"—but look and see whether there is anything common to all. [Wittgenstein, 2009]

1973]. For Rosch a set of empirical data points can all be associated to a centroid that effectively clusters them together, during our perception. Instead of categories being defined by their decision boundary (i.e., "where someone subject draws the line") they maybe associated to a central object that is *typical* of that class to which visual samples are grounded in order to be classified. For a single category these centroids may be multiple in order to properly accommodate all the things a category may represent. In a follow-up work, Medin and Schaffer tried to prove that visual inference is instead performed by associating the input instance to a set of samples stored in memory that are associated to a category [Medin and Schaffer, 1978].

What is an 'a'? While this may seem like an abstract problem it is an integral part of the very nature of categorization, which extends to something as trivial as our example of a signifier: the letter 'a'.[3] In the Middle Ages, professional handwritters, formally called scribes would copy notable texts like the Bible in order to pass knowledge from generation to generation and assert authority [Coulson and Babcock, 2020]. The script they would adopt for writing a text would try to inscribe consistent aesthetic qualities that would reflect its origin and purpose. To date these texts and agree on a common basis of categorization, palaeographers try to visually abstract a certain family of letters into textual descriptions. These Aristotelian definitions of categories, called typologies [Derolez, 2003] suffer from two main problems: **(a)** they are vague in their transmission leaving space for ambiguity and **(b)** as they are defined using text can't grasp the minute but significant variance of the visual data they aim to represent. This way, an alternative approach like the prototypes of Eleanor Rosch [Rosch, 1973] can allow to better organize existing categorization and even locate minute visual structure in its variation.

Classification is (often) not final. The practical motivation of this thesis is that behind concepts like a country's name hides salient visual variation that can be extracted and grouped. For example, behind the name of a country one could discover visual structure that can be grouped as a set of architectural (windows, roofs), infrastructural (road marks, electricity poles), or regulatory (license plates) visual elements. In Fig. 1.1b we locate two types of hidden structure of visual variation. The first type, concerns intra-class variation, or the minute differences between elements that are assigned on the same parent class. For example, given a script type of written charac-

[3] Douglas Hoffstader, famously wrote in 1985 that "the center problem of AI is the question: What is the letter 'a'?" [Douglas, 1985].

ters or equivalently a species of birds, one would like to understand whether there is minor variation that lies behind their underlying morphology, be it ascenders or beaks, which could characterize them as subtypes or subspecies. Even if this classification is arbitrary, this analysis can both allows us to establish its validity in light of a certain context [Stutzmann, 2016], or prove that it is arbitrary (i.e., statistically irrelevant) as in the case of the famous "Salamander's tale" [Dawkins, 2005]. The second type of hidden structure concerns inter-class variation, that lies behind the images of a certain domain. For example, given images that are conceptually identified by their geographical location, there may be elements that well characterize them, e.g., post-boxes [geohints, 2023]. While in the first case each label can be better structured in regard to participating into one of further categories, in this case each label can be decomposed and grouped into a ("mid-level") visual vocabulary [Singh et al., 2012]. Choosing which type of hidden structure of visual variation is more relevant to pursue, is sensitive to the input domain. For example, finding subfamilies of post-boxes for a specific country may be ill-posed due to lack or data or too trivial. Inversely, performing discovery of all the elements that make a manuscript typical to its family may be straight-forward as we already know that characters are expected to be the "visual vocabulary" of a manuscript. In both cases the potential of discovering further visual structure starting from predefined classification, is the fundamental motivation behind *image data mining*.

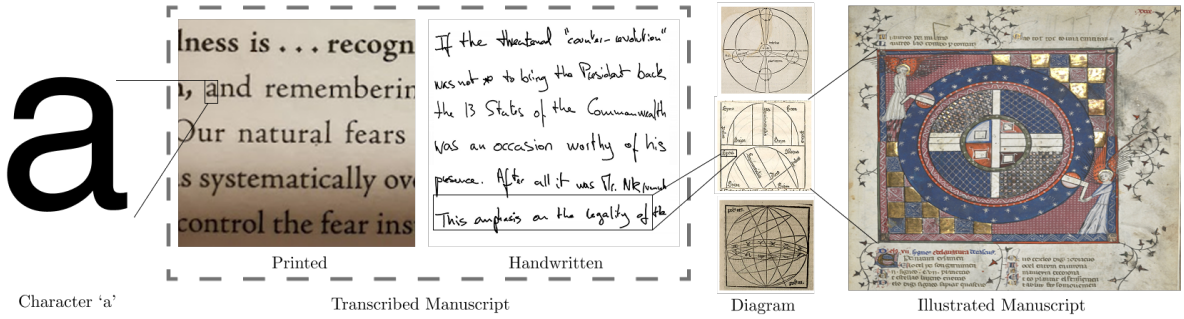
1.3 Goal

The goal of this thesis is to perform Image Data Mining (see Sec. 2). It aims to discover informative visual structure that lies behind the variation of established human categories and present it in a human interpretable way. Visual structure is defined as groups of individual elements (image patches, segments, etc.) that can be clustered together. Humans both perceive the visual world as a combination of such structure and in turn construct it as such, because of how they learn to perceive it. A written manuscript is composed of printed characters that readers learn to recognize (Fig. 1.2a), and a street is composed of road tracks that drivers learn to identify. As the datasets studied in this thesis, concern relevant domains (Fig. 1.2b), our goal is to build on systems that are able to identify this type of structure, as are for example discriminative classifiers, or segmentation networks. However, even if a system has the capacity to detect written characters it may not provide an adequate representation to differentiate and discover new visual structure.

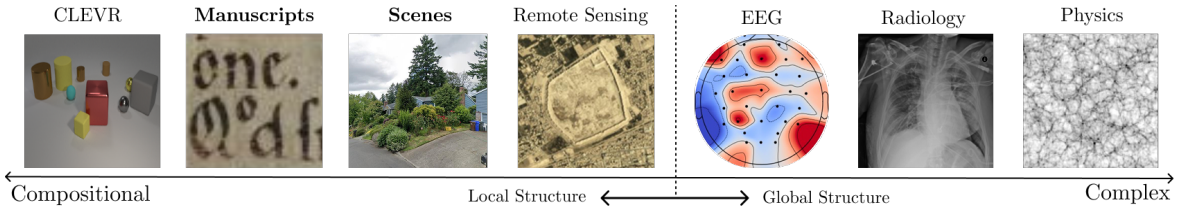
The importance of this problem becomes very apparent in geographical data. Consider for a moment the *"all text in nyc"* artwork that lifts all text appearing in the Google street view scenes of New York in a searchable interface [Zhao, 2024]. Having this annotation one could detect the statistics of certain phrases or find strange words that appear inside the city's billboards[4]. However, unlike text where these compositional attributes are often predefined and few, properly naming elements of an input scene is often ambiguous [Blake and Zisserman, 1987]. Similarly, with searching for text, one could try instead to search for predefined attributes that a system can learn to identify using supervised learning, by relying on standard semantic-segmentation methods [He et al., 2017] or using more open-ended text-based approaches, using multimodal models such as CLIP [Radford et al., 2021]. For example, one could search for images that contain a "car", "traffic lights", or "sky rise buildings that repeat on the side of a large boulevard". Yet even if we had the resources to list and count occurrences across all this combinatory of attributes, they may still be too generic to capture the minute important details under analysis. In other words even if our data are compositional, we may not already know what they are composed of. For example, in the case of UK windows, most buildings in the world have windows, yet the ones found in the UK have a very typical and consistent appearance. While one may describe them as "UK, white, UPVC windows", simply using a description isn't informative either: if a certain type window is only associated with a certain place in the world, then the discovery has already taken place. "As soon as something is named it is recoverable" [Baudrillard, 1999].

Focus of this thesis. More Concretely, this thesis aims at designing methods, that can mine visual structure in two practical cases of visual variation (Fig. 1.3): **(a)** transcribed documents, in order to analyze minute variation across characters (Chap. 3), and **(b)** labeled large-scale image datasets, in order to extract typical visual structure that summarizes a certain label, be it location, time, or type of scene (Chap. 4). We first focus on the micro-variation of written characters (Fig. 1.3a). For this domain, we aim to extract an aggregate representation of the font (or script) of a given set of documents. Later, by spatially aligning these representations for each character across different collections of fonts our goal is to be able to perform visual comparison that is interpretable while being quantifiable. This can provide us with clues about the evolution of the written character morphology that is hard to summarize otherwise. Our second goal, is to discover macro-variation that lies behind labeled datasets

[4] As the author's first name: https://www.alltext.nyc/search?q=Yannis appearing almost 100 times!



(a) **Example of Compositionality.** A letter can be part of a line of a written manuscript, which can be contained in an illustration, that can itself be part of an illustrated manuscript.



(b) **Compositional vs. Complex datasets.** Compositional datasets, as the ones studied on this thesis (**bold**) can be reduced to local structure which is informative for their global analysis. However, this is not the case for complex datasets, where analysis needs to be aggregated into a more global structure for it to be informative.

Figure 1.2: **Nature of the Studied Datasets.** This thesis focuses on compositional datasets, where local visual structure can be extracted and is informative of their parent class.

(Fig. 1.3b). For example, starting from images that come from a specific country, one would like to discover consistent visual structure that is typical of that location. These can be for example road-tracks, utility poles, post-boxes, windows and more. Such an approach allows human users to arrive at a high level understanding of how a certain label is represented inside a visual collection, potentially providing them with visual cues that can help them understand this collection on a higher level.

1.4 Challenges

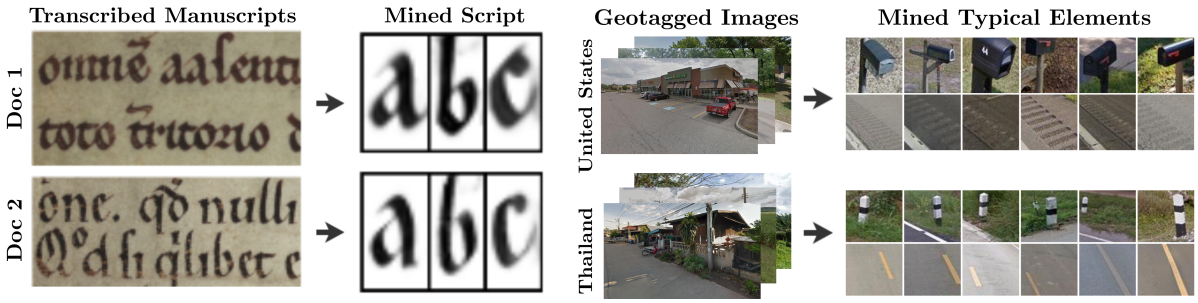
Image Data Mining, is challenging as it requires discovering and summarizing visual structure that is not already predefined by the input labels. Both because it is meant to be applied on datasets that concern the digital humanities, and because it is meant to be used by researchers to aid their analysis, the way we design these processes needs to be informed by the way humans already perceive and navigate visual data,

both in terms of labels and in terms of visual perception. This entails three different challenges.

Reconciling Supervised and Unsupervised Discovery. In order to identify visual structure that is common, one needs to also decide how to group it. In fact these two procedures, counting and grouping, are inseparable. This raises a problem of discovery. A rich unsupervised object discovery literature [Villa-Vásquez and Pedersoli, 2024], tried to show how useful or intuitive visual structure can emerge using certain architectures and training pipelines. However, the default use case of Image Data Mining, isn't simply the unsupervised discovery of visual structure as is for example the case with object discovery, but the identification of informative elements that best correspond to the labels of a weakly-labelled dataset [Singh et al., 2012]. While there exists a rich literature of categorical supervision, it mainly focuses on reliably recognizing predefined human categories [Wang et al., 2022]. This raises the challenge of reconciling both supervised and unsupervised approaches to discover non-annotated elements in our input dataset, while bringing this discovery closer to human perception.

Lack of Ground Truth A more foundational challenge, is that unlike object discovery, where the aim is to discover commonly occurring visual structure, in image data mining the visual structure that is expected to be discovered is in the form of clusters of *frequent outliers*, that are neither pre-assumed nor trivial. This locates Image Data Mining, in a regime of lack of ground truth that relies on subjective evaluation. It turns it into a task for which it is hard to measure progress in high granularity, and where its evaluation benefits from interdisciplinary expertise. Not only that but answers to the question of what an informative cluster is for a certain label may be multiple and sensitive to the input dataset. For example, while for purposes of classification two different measures of similarity, e.g., CLIP [Radford et al., 2021] or DINO [Caron et al., 2021] may have a similar performance, their qualitative performance, i.e., their retrieved nearest neighbors for the same dataset, may be noticeably different.

Interpretable and Faithful Summarization. The discovered output visual structures need to be summarized to a human level, in order to be compiled towards a form of conclusive evidence. Such evidence needs to be the most representative for the posed question and context. Not only that, but it also needs to be intimately connected to the way the model actually represents the input data. This connection



(a) **Inter**: *script* of historical Manuscripts. (b) **Intra**: *typical elements* of different countries.

Figure 1.3: **Goal.** Given labeled datasets, our goal is to summarize and compare their most informative elements in a human interpretable way. **(a)** given a collection of annotated manuscripts, our goal is to extract and summarize their script allowing experts to perform comparative quantitative morphological analysis (Chap. 3). **(b)** given a collection of annotated datasets of street-view images, our goal is to extract and summarize their typical visual structure, so that non-expert users can understand their labeling in relation to further visual elements (Chap. 4).

should be clear to the target user otherwise a future analyst that uses this tool can easily manipulate its results and project their biases in order to arrive to their desired conclusions [Sculley and Pasanek, 2008]. Unfortunately, most deep-learning architectures output high-dimensional representations of their input which can then be aggregated in multiple ways and whose features are often polysemantic [Olah et al., 2020]. This raises the challenge of how we can design approaches, which while being able to properly represent the data can provide interpretable summaries faithful to how the model actually perceives them.

1.5 Contributions

The approach of this thesis is to mine informative visual structure, using implicit or explicit, synthesis based compositional methods. These approaches align more with human perception [Biederman, 1987] and their predictions can be directly interpreted. We use two methods of the recent literature that can best capture the properties of the underlying visual structure. Sprite based methods [Smirnov et al., 2021; Monnier et al., 2021], are able to capture minute details of the input images by effective modeling the input data as a combination of a fixed subset of components. Diffusion models, through denoising learn to compose patches of the input scene into a more complex layering that can reproduce more abstract visual structure [Ozguroglu et al., 2024; Kamb and Ganguli, 2024]. Escaping the purism between unsupervised and supervised learning, our work adapts these methods to guide the discovery of new visual structure starting

from existing weak-annotation. For example, in the case of letters we don't want a system to have to both learn from scratch what a "d" is (which can be ambiguous as we discuss in Chapter 3) when our goal is to capture its shape. Instead, we would like to construct a system that learns the shape of "d", in a way that allows us to compare it with other shapes. Similarly, we would like to prime our system with the ability to recognize common visual structure or discriminate countries, so we can later find what visual structure of each country make it typical. Starting from established categories this enables mining visual structure that is not already annotated in the original training set, yet which corresponds to meaningful information.

Analyzing morphological variation of characters. To study images characterized by intra-class variation we use **sprite**-based methods [Monnier et al., 2020; Smirnov et al., 2021], that learn a decomposition of input images through a set of differential transformations of smaller input images, which in computer graphics are called sprites. Given an input text line, a neural network is trained to select and learn their appearance in order to effectively reconstruct all the text lines of an input dataset, on average. What sprites get to capture are interpretable summaries of the appearance of input characters. As they are very close in reconstructing the target images, they can be used to reliably capture the average written morphology of a given type. We show that this method can operate in a variety of printed and handwritten datasets [Siglidis et al., 2024a] and that using the learned sprites produced by this method can enable quantitative morphological palaeographical analysis [Vlachou-Efstathiou et al., 2024].

Mining typical summaries of labeled datasets. When studying inter-class variation of complex data that can hardly be characterized by a repetition of the same input image we rely on another synthesis method that is known as **Diffusion Models** [Ho et al., 2020] which learn to synthesize images by removing Gaussian noise across multiple resolutions. This allows them to learn hierarchical and robust representations of the input scene [Li et al., 2023a]. In fact, paired with text conditioning these models can learn to compose arbitrary combinations of visual concepts prompted through words, into an output image. By using them as a strong prior, we can identify how informative parts of the input image are for a target class, by averaging the difference in denoising performance using the information of a target class. Those that can be denoised significantly better in presence of the target class are clustered using features of the diffusion model and are then ranked to produce the final visual vocabulary. This type of approach allows us to summarize an input label in a way that is both

interpretable and that is directly connected to what the model perceives and groups as the most typical visual structure of a set of input labels [Siglidis et al., 2024b].

1.6 Thesis outline

This thesis consists of five chapters, which are organized as follows:

Chapter 2: Related Work. This chapter discusses related work in the area of Image Data Mining. It explores image data mining around three axes: discovery, mining and summarization. It outlines an evolution of methods and problems in each axis and clarifies how they are related to our work.

Chapter 3: The Learnable Typewriter, A Generative Approach to Text Analysis. Then, this thesis focuses in how to capture the inter-variation of characters from printed or real manuscripts of input text lines using a sprite-based approach. Through an architecture that is explicitly designed for reconstructing text lines using weak supervision it allows to capture the morphology of written characters [Siglidis et al., 2024a] for various datasets of printed font [Vincent, 2007], ciphers [Knight et al., 2011] and historical fonts [Seuret et al., 2023]. The power of this approach, is further demonstrated through an application to palaeographic analysis [Vlachou-Efstathiou et al., 2024] that allows qualitative and quantitative comparison in both rare [Camps et al., 2022] and established typologies [Derolez, 2003].

Chapter 4: Diffusion Models as Data Mining Tools. In the next chapter the focus turns to detecting visual structure of macro-variation across a variety of annotated image datasets of varying sizes. Starting from a pre-trained diffusion model [Rombach et al., 2022], a diffusion-based score is introduced that allows to mine informative elements from the dataset given an input conditioning. This allows to extract visual summaries for the labels of input datasets, scaling across a variety of data, such as historical cars [Lee et al., 2013] (10K), portraits [Chen et al., 2023] (25K), geographical data [Luo et al., 2022] (350K), and places [Zhou et al., 2017a] (1.8M), enabling an interpretable understanding of the meaning of a label assigned inside the context of an input dataset [Siglidis et al., 2024b].

Chapter 5: Conclusion. This thesis concludes by summarizing our work and discussing future directions, open problems and implications.

1.7 Publications

The following three publications are presented in the manuscript [Siglidis et al., 2024a; Vlachou-Efstathiou et al., 2024; Siglidis et al., 2024b]:

- Ioannis Siglidis, Nicolas Gonthier, Julien Gaubil, Tom Monnier, and Mathieu Aubry [2024]. The Learnable Typewriter: A generative approach to text analysis. ICDAR.
- Malamatenia Vlachou-Efstathiou, Ioannis Siglidis, Dominique Stutzmann, and Mathieu Aubry [2024]. An interpretable deep learning approach for morphological script type analysis. IWCP.
- Ioannis Siglidis, Aleksander Holynski, Alexei A. Efros, Mathieu Aubry, and Shiry Ginosar [2024]. Diffusion models as data mining tools. ECCV.

Our code was open-sourced[5] and presented using specialized webpages[6] which contain additional visualizations and results. Our work on The Learnable Typewriter received **the best paper award** at ICDAR 2024, and our work on Diffusion Models as Data Mining Tools aside from being published as a Poster on ECCV, was invited for a spotlight talk at the *Workshop for Visual Concepts*[7].

Not presented in this thesis. During my PhD, I was a joint first author in the following publication [Astruc et al., 2024], which started as a group hackathon idea from Nicolas Dufour and was performed under the supervision of Loic Landrieu:

- *Guillame Astruc, *Nicolas Dufour, *Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. [2024]. Openstreetview-5m: The many roads to global visual geolocation. CVPR.

The goal of this work was to release the largest open-source dataset for global scale visual geolocation, that had significantly more well-defined label-image associations, and a more balanced coverage than existing open datasets, in order to facilitate turning geolocation into a standard benchmark for evaluating image models as opposed to well established datasets like ImageNet [Deng et al., 2009].

[5] https://github.com/ysig
[6] https://ysig.github.io/phd/#papers
[7] https://sites.google.com/stanford.edu/visual-concepts-workshops/eccv24