Clustering Nature at Its Joints

Natural Kind Realism in the Context of Unsupervised ML Algorithms

Kevin Li

Advised by Wesley Holliday

April 18, 2022

Abstract

In this paper I attempt to draw a conceptual connection between the debate over natural kinds in the Philosophy of Science and the study of clustering algorithms in unsupervised machine learning. They are concerned with the practice of grouping objects together in two different senses of the term "natural". I argue that computational approaches to classification are deserving of some philosophical reconciliation, and that they have implication for whether or not we believe there is some sense in which natural kinds exist. I elucidate a broad view of natural kind realism, which I argue holds a core epistemological disagreement with conventionalism. I then discuss how several theoretical results from the clustering literature should be interpreted under this disagreement. I consider three proposed methods for evaluating clustering algorithms as well as Kleinberg's Impossibility Theorem, and ask whether they offer reasons to believe that our classifications are natural in the realist sense or merely conventional.

Contents

1 Introduction

A fundamental way in which we understand the world is by dividing it up. Organisms are classified as plants, animals, or fungi. Ski routes are more or less difficult if they are blue, black, or double black. People hold different political beliefs if they are Republicans, Democrats, or Libertarians. Classifications help us understand the world, but is there anything to be said about whether some are better than others? Here are a couple examples of what I mean.

1.1 Music Genres

I find musical taste to be one of the most frustrating things to describe. When I get asked what I like to listen to, I'll usually list off a couple of genres: indie, jazz, and hip hop. These give a rough impression of the music I enjoy but aren't entirely correct. Indie, as a genre, is defined as music from independent artists, but I really just use the category to refer to the "chill vibes" music I enjoy from a couple of small bands. I enjoy lots of jazz tracks, but I also love songs that just have elements of jazz in them (which might otherwise be classified in different genres). I also wouldn't identify myself as a full-on hip hop enthusiast, but I'm a fan of at least a couple of artists who fall under the category. Those three genres don't tell the whole story, but they give a fuzzy picture.

Another way of feeling out someone's taste in music is to look at the playlists they listen to. My friend has a playlist called "Lyrical Lemonade," which is full of lyrical rap songs by artists like Kendrick Lamar and J. Cole. The playlist picks out a pretty great subset of rap music with memorable lyrics, interesting metaphors, and poetic messages. I've also seen several versions of the "throwbacks" playlist, a collection of nostalgic 2000s and 2010s songs like "I Gotta Feeling" by The Black Eyed Peas or Jason Mraz's "I'm Yours." The songs in this category don't share much in common other than their popularity and their place in time, but they collectively form a category of music that is quite easily identifiable, cohesive, and frequently listened to together.

"Lyrical Lemonade" and "Throwbacks" aren't proper genres the way that hip hop and jazz are, but they seem to describe a kind of music just the same. In fact, one can be just as clear about the music being referred to by these inventive playlists as with generic genre labels. The following are a couple more examples of playlists, all of which span different traditional genres and are listened to by at least 5 million users on Spotify:

> **Songs to Sing in the Car**: 100 sing-along songs including Don McLean's "American Pie," a folk rock hit from 1971, and Doja Cat's "Kiss Me More," a 2021 Grammy-nominated pop song.
>
> **Chill Hits**: a playlist described simply as "the best new and recent chill tunes."
>
> **Mood Booster**: a collection of "today's dose of feel-good songs."

Each of these playlists group together songs in a pretty reasonable way. The playlists have a theme, the songs follow that theme, and the themes are compelling enough that millions of users listen to the playlist in virtue of this theme. In fact, "Songs to Sing in the Car" is, by name, designed to be played while driving in the same way that one might put on radio stations in the car (which are commonly given by genres like country or R&B). These oddly specific playlists accomplish what widely accepted genres do—they identify a common feature between a set of songs (that is generally accepted by listeners). Somehow, we might still be more intuitively inclined to accept R&B as a legitimate genre over "Mood Boosters". To claim that "Mood Boosters" is a *kind* of music feels awkward and strangely incorrect.

Here are a couple other proposals for how songs could be grouped together, if "Mood Boosters" doesn't feel strange enough:

> **Musical Key**: Classify songs by which of the 12 major or 12 minor keys it is written in.
>
> **Beats per Minute**: Classify songs by their tempo, or in other words, how fast they are played.

Both of these candidate groupings of music, as convoluted as they seem, still have genuine reasons for their way of dividing up music. Young pianists, for example, might care to find songs by key if they can only play easier key signatures. Runners have reason to care about songs in particular BPM ranges—there is evidence that the tempo of a song being played has an effect on how fast a person runs. Neither of these classifications are arbitrary, yet they somehow feel like the wrong way to describe *kinds* of music.

Are we justified in believing that certain classifications of music are better or more correct than others? Is there a reason that genres like country and pop are accepted as proper categories of music, while "Throwbacks" and "125 BPM songs" are not? Finding an explanation for what makes certain classifications "right" is difficult. There are plenty of legitimate reasons for all sorts of groupings (as we've considered above). Some seem to make more sense than others, but it's not immediately clear that one *must* reflect the proper divisions of music. Each classification has its merit, and one could just argue that there isn't really a right answer—there are no true "genres of music".

1.2 The Periodic Table of Elements

Take chemical elements as another example. A common classification system is the Periodic Table, which groups atoms by the number of protons they have in their nucleus (also known as their atomic number). It turns out that grouping atoms by their atomic number is an extremely useful classification for scientists. The elements participate in important chemical properties that have stood the test of time, and it has even been shown that these properties have a periodic dependence on their atomic numbers. In fact, when Dimitri Mendeleev first created the Periodic Table, there were several gaps where elements hadn't been discovered yet, but he was able to correctly predict some of their properties using this system of classification.

In the same way that traditional genres of music divide songs into groups, the Periodic Table is used to classify chemical elements. To group elements by their atomic number is evidently useful, but could we say that it is the *right* way to do so? Just as we might think there is not one proper way to divide music into genres, is there not one proper way to classify elements? We could just as easily group elements by the country they were discovered in, whether or not they can be found in doorknobs, or the number of letters in their names. If classifications are arbitrary, the Periodic Table of Elements is no better than "The Periodic Table of Doorknobs".

In defense of centuries of scientific progress, *something* should be said as to why the former is better than the latter. The difficulty lies in what exactly can be said. It could be argued that the Periodic Table of Elements is more useful, but I could easily come up with a use for any of my other proposed classifications. Knowing what elements are or are not contained in doorknobs is plenty useful for door manufacturers. If instead we argued that the Periodic Table is better because it is endorsed by expert scientists, I would insist that they don't get to decide—I have just as much a right to classify them how I want. We could then push back and say that scientists have a much better reason to endorse the Periodic Table than we do for doorknob manufacturing, since it has been used to predict lots of important chemical phenomena, contributing to our understanding of the world. In some sense, the Periodic Table almost groups elements the way the world meant for them to be grouped—it seems hardly possible that our grouping by doorknobs reflects any deep truth about the universe. But as we travel deeper into this debate, we've found ourselves in a sea of difficult philosophical problems. What does it mean to say that the world has a *true* grouping? If it does, and the chemical elements are one such grouping, could the traditional music genres we've used for decades be one as well?

1.3 Outline of the Following Sections

Classifications play a central role in how we think about the world. We can't help but carve up the different objects we see into kinds, yet we encounter a lot of trouble when we begin to ask why or how we do so. Why do we differentiate between dwarf and non-dwarf planets, but not between the red planets and the planets with stripes? Is green any more of a color than Crayola's Electric Lime? How is it that we have four seasons to a year, and not seven or twelve? The issue of how we arrive at classifications, as well as what they say about the world, is one with philosophical importance. This issue has also become increasingly relevant in the field of computer science, as there is growing popularity in utilizing machine-based approaches to classification. In each field, the problem has come up in different ways. There is, on one hand, the debate over natural kinds in the philosophy of science, which asks if the way that we group particular objects reveals some underlying structure of the natural world. There is, on the other hand, a debate concerning whether computer algorithms can arrive at these underlying classifications of the world, unsupervised. Computer scientists and philosophers have separately tackled the problem of classification, but I argue that there are some senses in which they are talking about the same thing. To the extent that they are, insights in one camp might have implications in the other. This paper explores whether the theoretical results from classification algorithms have implications in the philosophical debate about natural kinds.

The debate over natural kinds is concerned with what kind of knowledge is constituted by the classifications that we make in all sorts of human thinking, from chemical elements to musical genres. I will later spell out what the term "natural kinds" means more formally, but to roughly introduce it: to say that a classification is *natural* is to say that it organizes objects in a way that reflects the genuine way our world is organized. A natural kind is a grouping of objects independent of human interests—it divides the world in the way the world is actually divided, as opposed to how we might, as observers of world, associate objects haphazardly. As I will expand on in Section 2, these debates are generally about (1) whether these kinds exist and (2) whether we are able to obtain knowledge of them, if they do. This section will introduce two positions on this debate in order to contextualize where classification algorithms may be relevant..

If natural kinds are how the world is structured and all we do as human thinkers is reveal them through our practice of classification, it seems plausible that machines, armed with useful data, could do the same thing. Section 3 provides an introduction to clustering algorithms, which are a subset of unsupervised machine learning algorithms that are used to find clusters in large sets of data points. I will explain how these algorithms approach the task of classification, and argue that they can be thought analogously to human classification, at least to the extent in which the disagreements over natural kinds are concerned.

Section 4 considers how clustering algorithms and the classifications that they produce are generally evaluated. If natural kinds exist, and it is possible for us to discover them in our own classifications, we must have a way to pick out the ones which are natural from the ones that aren't. This is, I argue, the role of evaluation procedures for clustering algorithms. I consider three senses in which computationally-generated classifications can be evaluated. Some of the conditions on which classifications are evaluated are human-interest dependent, but the ones which are not can potentially be indicators of natural kinds. In this section, I explain how some ways to evaluate clustering algorithms could serve as natural kind indicators and therefore imply that there are natural kinds we could have knowledge of. Section 5 tests this argument against three different proposals for universal clustering evaluations, to see if the way in which they endorse classifications, in a computational sense, is also an endorsement of them as natural kinds, in a philosophical sense.

Section 6 discusses another way that theoretical results in clustering literature have consequences in the debate over natural kinds. An interesting theorem has shown that there are three basic properties of clustering algorithms that are incompatible in combination. I consider what this implies about the limitations of classification and how the properties could be relaxed to defend against this seeming impossibility.

2 A Closer Look at Natural Kinds

There are various philosophical accounts of what it means to regard a classification as "natural". As I have drawn out in the introduction, we make all sorts of classifications of the objects in the world, and we do so for many different reasons. We not only hold different reasons for making classifications, but we also hold different reasons for claiming that one classification is better than another. The debate over natural kinds in the Philosophy of Science is specifically concerned with how we privilege certain classifications over others for reasons of naturalness. The complexity of this issue lies in what exactly we mean by the term “natural”.

Before I go on to explain some of these philosophical views, it will be useful to give some strategic vision for what I am setting out to do. There are plenty of open questions about what a natural kind is and how we come to know of them. What I am working towards is not a resolution to these questions, but rather an identification of some points of contention where computer classification algorithms might become relevant. Because I believe there is some sense in which computers and humans engage in the same practice of classification, and the natural kinds debate has thus far only been concerned with human-generated classifications, I argue that some of these philosophical problems will benefit from a discussion of computer-generated classifications. I want to ask: what contributions can be made to our understanding of natural kinds by thinking about classifications computationally? It would be ambitious to argue that computational results will completely overturn and answer all questions concerning natural kinds, but at the very least, they should complicate some of them. Towards this end, the following section will explore a variety of views about natural kinds for the sake of contextualization, but the main consideration will be in finding just *some* of their philosophical disagreements in order to make way for the discussion of classification algorithms.

At the outset, we can frame the philosophical discussion as taking two dimensions—the first concerns what it means for natural kinds to exist, or in other words, what we are claiming when we claim a kind to be “natural” (this I will call the metaphysical debate). The second concerns how we could come to have knowledge of natural kinds, which is epistemological. I think it makes sense to first discuss what metaphysical views are available, and when trouble and ambiguity arises, to turn over and ask how we could obtain knowledge of natural kinds, in order to assess what our metaphysical beliefs about their existence can amount to.

2.1 The Metaphysical Divide

There are various accounts of what it means to say that natural kinds exist. There are generally two views—the natural kinds realist believes that there is at least some sense in which natural kinds exist independently of human interests, and the conventionalist believes otherwise, or that their existence is purely a matter of our interests and actions. I emphasize that the realist believes there is *at least* some sense in which natural kinds exist, because the exact sense in which different realists believe that natural kinds exist is where there is ambiguity and disagreement. Some realists believe that natural kinds exist as a fact about the structure of the external world, while others believe that they exist only as a fact about the structure of human minds. Among the many versions of natural kind realism, they also disagree over what sorts of objects have natural groupings. For example, it might seem plausible for many realists that chemical elements could have a natural division, but it seems less so that human created objects, like furniture or currency, could have a natural division just the same. My point here is just to show that natural kinds realism is not so much a singular metaphysical view, but rather a class of views that each commit to something slightly different. This sub-section will introduce some of these views and bring out a couple points of ambiguity in the term "natural" to demonstrate why this debate is so complicated.

As I explained earlier, my motivating goal is to frame the discussion of machine learning literature around one core disagreement between realists and conventionalists (the disagreement being epistemic success, which I will expound in Section 2.2). Since my focus will be on this one disagreement, is it less important whether one realist view is more plausible than another, so long as they all disagree with conventionalists around this one core issue. The implication I hope to draw is not that machine learning literature suggests a particular version of realism, but just that *some* version of realism is plausible in contrast to conventionalism. What I care about is realism in a broader sense—what are all of these views of realism disagreeing with conventionalists over? With this in mind, the rest of this section is meant to help bring forward this broader view of realism before drawing out their contrast with conventionalism in the following section on epistemology. I will first introduce a common criterion held about natural kind realism in the philosophical literature, and as trouble arises, challenge this criterion in order to expand to a broader position on natural kind realism.

Anjan Chakravartty (2007) argues that a kind is natural if it is mind-independent. A natural kind exists if it does not depend on us having thought it to exist. The mind-independent realist believes that classifications we take as natural are those that don't depend on us to exist. The conventionalist, on the other hand, believes that our classifications are all mind-dependent, meaning that they strictly depend on us, how we think about them, or how we use them—none of them are natural.

In what sense does a mind-independent natural kind exist? If it does not depend on our minds, it follows that natural kinds must exist external to us somewhere in the world itself. For example, the chemical elements of the periodic table are potential candidates for natural kinds under this realist view, since their consistency with chemical phenomena have no dependence on what our minds think of them. This criterion generally works well for the sorts of classifications found in the natural sciences, because the focus of domains like physics, biology, and the earth sciences are on phenomena which we take to already exist mind-independently, and therefore the classifications that explain them can also be taken to exist mind-independently. [1]

The trouble arises when we consider objects which we do not take to exist mind-independently, or objects which we take to exist mind-independently, but are classified in a manner that appears to be mind-dependent. The question, in these cases, is whether there is still something real about these classifications. I will first address the latter—is there a sense in which we generate mind-dependent classifications of mind-independent objects, which we nevertheless consider informed by reality? Realists take natural kinds as reflecting the natural carving

[1] It is worth noting here that even in these domains, there is debate over which classifications are mind-independent. For example, species like "tiger" and "blue whale" are arguably delineated by biologists' interest in evolutionary study.

of the world, but there is also a sense in which natural kinds may not exist in the world external to us, but rather as a reality of the human mind or of human nature. In this case, the term "mind-independence" is complicated by the fact that certain kinds might only be facts about our minds but are nevertheless real. Take, for instance, the classification of color. It is generally taken that the world itself does not "carve up" color. Light exists in the physical world as a spectrum of frequencies, and without the way that humans think about color, there would be no such thing as red, green, or blue in the world itself—there is only red, green, or blue *to us*. Our perception of color is a function of human photoreceptors, which means that what we carve out to be colors is a fact about our minds.

But even so, there seems to be something real about the way we experience color. Take Figure 1. I have proposed two classifications of the visible light spectrum. The first is a generally held consensus on what the colors are (although they differ slightly between cultures and languages). The second is what I consider an arbitrary division of the spectrum. If colors are not natural kinds, at least by the mind-independent criterion that we have discussed thus far, there is no saying whether one of these classification systems is more metaphysically real than the other. If the realist is truly committed to the criterion of mind-independence, it might be that there is simply not much more to say—we may *prefer* or take advantage of the first system more than the second, but our preference does not amount to any metaphysical reality. I would want to argue that, even if color does not reflect a fact about the world itself, it might still reflect a necessary fact about our minds. There seems to be something going on that is stronger than just a preference for one color grouping than another, whether it be years of evolution that have trained our eyes to distinguish between reds and blues, or maybe the coincidental fact that, somehow, languages far and wide have all arrived at words for blue but not for the color of my pillowcase.

This is all to say that there is a sense in which this classification might be natural, even if it does not exist in the world external to us. The study of human minds is considered a science by many who observe it, and, if not color, there are other classifications about

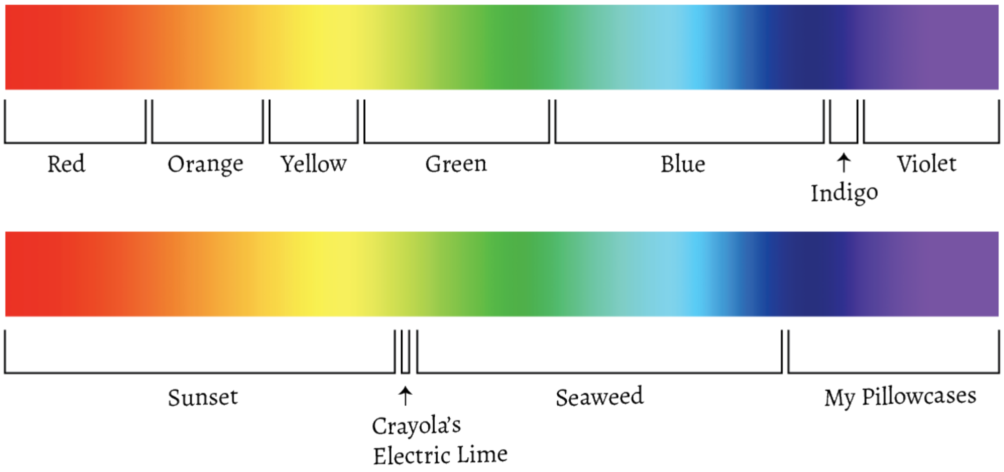

Figure 1: Two ways of classifying the visible light spectrum

our minds, such as mental illnesses or human beliefs, which seem to complicate our use of the criterion of mind-independence. They are mind-dependent in their content, but mind-independent in the fact that there seems to be something *necessary* that does not depend on our attitudes toward them.

Natural kind realism is further complicated by the consideration of classifying objects which only exist mind-dependently. Examples of this form include objects like "dollar bill" and "hundred dollar bill", or kinds in the social sciences like "conservatives" and "liberals". Unlike color, which classifies the visible light out in the external world, these examples stray further into the territory of mind dependence, and it seems aimless to argue that they reflect some natural structure to the world independently of our interests in forming those classifications in the first place.

It is unclear where the realist should draw the line, or in other words, what is a good criterion for capturing what are and are not natural kinds. This dilemma only gets more muddled as we consider other examples that blur the line of mind-independence. Muhammad Ali Khalidi (2016) gives several examples of objects where the mind-independence criterion for natural kinds has ambiguity:

> **Roentgenium**: A chemical element, atomic number 111, whose most stable isotope has mass number 281 and a half-life of 26 seconds. It was first discovered in 1994 when a single atom of roentgenium-272 was produced in the lab by bombarding a target of bismuth-209 with nuclei of nickel-64. Even though the discovery was not certified at that time, in 2002 the experiment was repeated and three more atoms were produced, and this discovery was later certified. It may be that the only atoms of this element that have ever been produced in the universe have been made in the lab, here on earth.
>
> **Triticale**: This plant, which is a hybrid of wheat and rye, was bred to combine the grain quality of wheat with the tolerance of rye, and is mostly used as fodder. Like many hybrids, it is sterile, so it must be chemically treated to double the number of chromosomes and enable it to reproduce itself, which it could not do without intervention. This is done by applying colchicine, a chromosome doubling agent, to a growth point of the plant.
>
> **Dog (Canis familiaris or Canis lupus familiaris)**: As is widely known, dogs have been artificially selected by humans over many generations and originally domesticated from wolves (Canis lupus). The origin of the process is still shrouded in some mystery and estimates for date of domestication vary widely, from roughly 9000 to 34,000 years ago. Moreover, current evidence suggests that the domestication of dogs may have occurred more than once in human history and that some of these lineages did not survive. There is also considerable debate over whether the process originated at human initiative or whether it was largely the fortuitous result of certain members of the wolf species lingering near human settlements. Either way, artificial selection was eventually carried out very deliberately, resulting in many distinct varieties, with distinct characteristics.

These examples demonstrate the difficulty in defining the term “natural kinds” by mind-independence, because there are different ways that humans intervene in the objects of classification, and therefore complicate whether the classifications they fit into can be taken as natural or not. At this point, it is unclear where the criterial line should be drawn between what can be a candidate natural kind and what cannot. As I explained at the beginning, this debate is difficult, but what is important is just that there should, at the very least, be a line *somewhere*—realists and conventionalists could not be disagreeing over nothing, because conventionalists believe that *no* natural kinds exist, and realists believe that at least some do.

I feel that there is still a good deal of ambiguity here, so I will try to do a little more work to frame it. Some realists take natural kinds to be the classifications which exist mind-independently, but this requirement is complicated by the fact that there are other classifications, like color, or objects of classification, like triticale, which we might think of as real but don't fit well into the commitments of mind-independence. The trouble lies in what role we think we play in the game of classification—the conventionalist believes that we are the sole players of this game, but the realists want to argue that some aspect of it is independent of us. What aspect, exactly, is independent of us, is what different views of realism disagree over. The mind-independent realist, the colors realist, and the triticale realist might each be committed to a slightly different criterion for picking natural kinds.

Thus far, it has been difficult to draw the line in the sand, so I think it makes sense to turn to epistemology and ask *why* we suspect that some of our classifications are more than just arbitrary, and *how* we come to have knowledge of this. Whatever it is that natural kind realists and conventionalists are disagreeing about on the metaphysical level, the disagreement should also surface at the level of epistemology. In other words, if natural kind realists think that some classifications are more than arbitrary or human-interested, what is it about our epistemic beliefs that indicates so? If conventionalists believe otherwise, what is it about our epistemic beliefs that fail to really demonstrate the existence of natural kinds? Rather than taking the debate as a matter of what "natural" is and what it is not, it may be easier to simply ask how the two views disagree in their account of epistemological evidence, namely, what persuades us to even entertain the fact that natural kinds might exist.

2.2 Epistemological Considerations: Epistemic Success

What is it about our classifications that makes realists think that there is something natural to begin with? Regardless of whether natural kinds exist or not, why might we *suspect* there is something real about our classifications, independent of human interests?

I turn back to the earlier discussion of the Periodic Table. When Dmitri Mendeleev formulated the first iteration of the Periodic Table in 1869, there were plenty of missing elements which had yet to be observed. The system of classification, nevertheless, correctly predicted how newly-discovered elements would behave. For centuries, chemists have continued to use (and revise) this system, because it organizes atoms in a way that predicts chemical phenomena consistently. The same way that scientists are more convinced of hypotheses that receive evidential support over time, some classifications continue to be used and endorsed because they group objects in a way that predicts how future objects, based on those groups, will behave. It is this miracle of induction that suggests that some classifications are stronger and possibly reflective of the way the world is really carved.

A possible motivation for thinking of some kinds as natural (as opposed to arbitrary or human-interested) is the fact of epistemic success. Classifications which are able to help us predict future events are the ones which are suspiciously non-arbitrary. If classifications were simply a matter of preference, it is a wonder that they have helped us categorize diseases to create vaccines, categorize human populations to win elections, or categorize behavior to design widely-successful marketing campaigns. Classifications which display epistemic success are, in other words, predictive. We can hypothesize about phenomena using these classifications, and they tend to predict and explain how events unfold. Disease categories predict what treatments will work, demographic categories predict how citizens will vote, and customer segmentation categories predict what advertisements people will be receptive to.

Epistemic success is a motivation for thinking that some classifications reflect natural kinds, but it is not yet clear how exactly it implies that they do. Before discussing whether epistemic success aligns to what a realist might take as "natural," I will first explain how it implies that certain classifications are ruled out. If epistemic success is an indicator for natural kinds, then the classifications which cannot display epistemic success must also not be natural kinds.[2] In other words, there are certain classifications which are not predictive at all, and if epistemic success is the reason we take certain classifications as natural, then

[2] This can be challenged by views of realism which do not necessarily take epistemic success to be an epistemological indicator of naturalness.

non-predictive classifications are certainly ruled out. There are core differences between classifications which are predictive and those which are not. Some classifications are used to group objects by certain properties, but there are no additional properties of those objects that can be predicted, and therefore no possibility for the classifications to demonstrate their predictive power. For example, consider two classifications: climbing grades, which are used to classify the difficulty of climbing routes, and psychiatric taxonomy, which is used to classify mental disorders. Psychiatric taxonomy is a predictive classification because the illnesses in each category tend to be receptive to certain treatments, so the matter of how effective psychiatric taxonomy is at dividing up mental disorders can be tested against the success of mental health professionals in finding treatments. Climbing grades could also be considered a predictive classification if every new climbing route can be classified well into one of the existing difficulty categories. I argue that this is a different sort of epistemic success. The success of climbing grades is different from the success of psychiatric taxonomy because the categories in the latter case are used to predict phenomena which are not essential properties to the categories themselves, while the properties of objects in the former categories are so in virtue of belonging to that category. For instance, mood and personality disorders tend to have particular treatments that are effective to each, but the disorders themselves are not categorized on the basis of what treatments will be effective to them. This is not to say that the psychiatric taxonomic system, as it stands today, is certainly a natural kind, but that there are ways to confirm or deny its "correctness" as a kind. Its structure is not purely dependent on how we choose to see it—we play a role in revising the classification, but it is the success of its predictions that determines how it will evolve. In contrast, a V5 climbing route may require certain techniques and footholds that a V4 does not, but it is on these bases that they are classified as V4 or V5 to begin with. The category of V5 climbing routes does not predict that something is the case about the climbing routes which belong to that category, rather, the things that are the case about the climbing routes are what make them a part of that category.[3] The difference between the predictive possibility of some classifications over others is a matter of what they are attempting to predict and whether those facts are contingent or necessary. If they are contingent, then there can be success or failure. If they are necessary, there can only be success. A treatment may or may not work for a mental disorder. but a V2 climbing route is necessarily climbable without advanced footholds.

At this point, I have explained how there are some classifications which are not predictive and therefore are not natural (under the assumption that predictive ability is what indicates naturalness). What is left is to ask whether the ones which *can* be predictive are indeed natural. There are classifications which seem to *discover* something about reality, because if they were simply *inventions* of the mind, it is a wonder that they have served us so well in our predictions about the world. The concern, however is whether a classification's predictive ability is enough to qualify it as a natural kind. Consider the following case of prediction. A snack company comes up with two ways of categorizing their customers into user segments for the purpose of creating targeted advertisement campaigns. They might design different advertisements for users based on the country they live in or based on whether they are over or under the age of 26. If the country-based advertisements in turn result in more people buying snacks and the age-based advertisements do not, we could argue that the former classification has correctly predicted customers' buying habits or food tastes and is therefore the better classification on the basis of this predictive power. The question then, is whether this in turn implies that the country-based classification is also the more *natural* classification. Does this experiment really give us reason to believe that countries is a natural kind, and the age is not? The conventionalist could argue that neither are—the former classification is only more predictive than the latter because we have arbitrarily chosen

[3] I would like to qualify this point—although climbing grades do not really predict whether a route uses certain footholds or techniques, it may be the case that some additional properties common to each difficulty level are discovered (for example, that V4 routes tend to stress a particular set of a climber's muscle groups that routes of other difficulties do not). In this case, it *is* possible for climbing grades to be predictive. Whether or not a classification can be predictive is not a matter of what it classifies or how it does so, rather it is a matter of whether it has found something to predict.

what we are predicting. The realist, in defense, could argue that it is not that prediction does not indicate natural kinds, but just that this example has not been predictive *enough* to do so. We just need to test these user segments in more ways. If we continually revise our classification over many years and with many experiments, as we have with the Periodic Table, we might eventually see more of its predictive potential and be better convinced of its naturalness.

Whether or not epistemic success amounts to an indication of natural kinds is a matter of how we *interpret* the nature of epistemic success. This is where the metaphysical disagreements between realists and conventionalists come apart at the level of epistemology. As Anjan Chakravartty (2007) puts it, our epistemological endorsements for certain classifications become a kind of data for metaphysical views. He writes, “The fact that kinds are posited to account for epistemic success ultimately places constraints on what kinds are taken to be, because the epistemic success that some categories afford and others do not amounts to a repository of empirical data for thinking about the nature of kinds” (4). Realists and conventionalists disagree over how we are to interpret this repository of empirical data. We have in front of us the fact that our “best” classifications are the ones that have the highest epistemic success, and it is left up to metaphysics to then interpret and account for what the epistemic success of those classifications mean. Do they suggest that we have landed upon natural kinds, or is there a way to account for epistemic success as some kind of anthropocentric trick of the mind?

Chakravartty offers two ways to account for the epistemic success of certain classifications, one of which implies that there is something natural about them, the other of which attempts to explain success as mere human interest. On one hand, human thinking can be thought of as, metaphorically, a “filter” (13). Our process of picking out classifications that make successful inductive inferences is how we come to discover natural kinds. This view takes epistemic success as our means of triangulating on the kinds that are natural, filtering out the ones that are not. The other characterization is that of a “lathe”—our inductive practices

Figure 2: This is a lathe– a block of wood can be carved on it by spinning it and placing tools against it to pick away at small pieces.

cause us to shape our classifications a certain way, and thus the classifications only exist, or are legitimated by, the interests they serve. In other words, it is our inductive interests that determine our classifications, not the shape of the world. The conventionalist takes epistemic success as a reflection of what our interests are, like creating advertisement campaigns or doing chemical experiments, rather than as a reflection of how the world is actually divided. Realists take advertising or chemical experiments as ways of filtering down, from the broad range of possible classifications, the ones that tell us how the world is actually divided. Because each view disagrees over what kinds exist metaphysically, there is a difference of interpretation over what it is we think we have knowledge of when we talk of epistemic success.

To summarize, I have argued that out of the many ambiguities about what the realist believes about natural kinds, a core disagreement with the conventionalist lies in how they interpret epistemic success. The realist and conventionalist should both have an explanation for why it is that some of the classifications we use are good at predicting things about the world. The conventionalist argues that when our classifications get better at predicting things, they are only better because we decide what we wish to predict. The realist adopts the view that when our classifications get better at predicting things, they are getting closer to catching onto a natural kind. It is not necessarily that every realist believes that epistemic success is *enough* to suggest that natural kinds exist, or that epistemic success is a sufficient condition for a classification to be endorsed as natural. It is that, under a broad view of realism versus conventionalism, this is at least one point of disagreement—whether epistemic success acts as a filter or as a lathe. The following sections consider how we should think about classification algorithms along this disagreement. What metrics are available, in the computational sense, for demonstrating the epistemic success of a classification? Are they better interpreted by the realist as a filter, or by the conventionalist as a lathe?

3 An Introduction to Clustering

The primary goal of this paper is to draw out similarities between the way that humans and machines engage in the practice of classification—if our judgements about the naturalness of human classifications should be compatible with our metaphysical beliefs about natural kinds, then our judgements about the naturalness of machine-produced classifications should be as well. Section 3.1 will give a brief introduction to how unsupervised clustering algorithms classify objects, and Section 3.2 offers a defense of why they are worthy of philosophical investigation. Section 3.2 broadly addresses how computer and machine classification can be compared, but there are two important footnotes in Section 3.1 that address specific issues that initially arise in my introduction to clustering algorithms.

3.1 Machine Learning

Machine learning (ML), broadly, is the study of computer algorithms that learn patterns from data. ML techniques have varying levels of human supervision, or in other words, they are given more or less human input in making decisions. Supervised ML algorithms are given a pattern and asked to apply that pattern to new things—they take in data that is already categorized, look for patterns to understand how the data is categorized, and then use that understanding to categorize new data they haven't seen before. Consider the task of teaching a computer to recognize handwritten digits. We want to ask a machine to determine, from

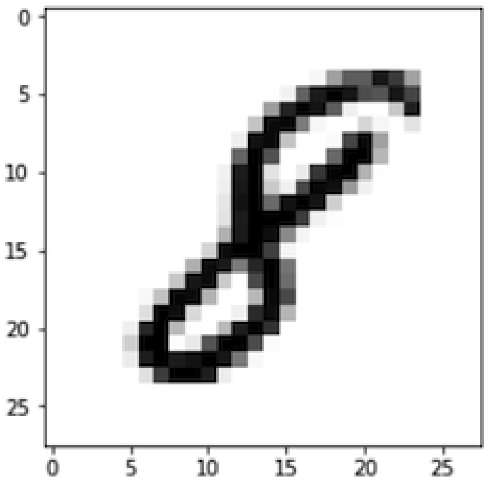

Figure 3: Sample image from the MNIST dataset at index 7777

an image of a handwritten digit, which number (0-9) it is. The MNIST dataset contains 60,000 images of these handwritten digits, which all look something like the one in Figure 3.

To teach a machine how to recognize these digits, we can give it a set of *labeled* data, which is a set of images whose digits have already determined. Using this information, different ML algorithms can then be leveraged to make educated guesses about unlabeled data, that is, images whose digits haven't been determined yet. By taking in examples of images that have already been categorized, it can "learn" how to classify new ones.

Digit recognition is an example of an ML problem where we already know how our data should be classified—it should fall in one of ten different buckets, one for each digit. When this is the case, *supervised* ML is useful. We can give the algorithm a sense of how this system of classification works by feeding it examples. There are other cases, however, where we don't know how our data should be classified, and we instead want to first figure out how it should. For example, if we were interested in researching different types of cancer by grouping cancer patients, we might not know exactly how many kinds of cancer there are in the dataset, or what those kinds are.

This is where clustering algorithms come in. Clustering techniques (or at least the ones relevant to this paper,) fall under the umbrella of unsupervised ML algorithms. In such cases, we don't give the algorithm examples of how to classify the data. Instead, we ask the algorithm to help us *decide* how to classify it. In other words, there is no predetermined structure to the data that we wish to teach the machine—we want it to find structure on its own. A clustering algorithm is a computational procedure that accomplishes this task. It takes in some information about a set of objects as input, and outputs what it considers to be the best way to put them into groups (or clusters). The following is an example of what a clustering algorithm might involve.

Consider the task of dividing up the customers of a credit card company. We wish to find a way of grouping customers into credit groups which may then help the company decide what credit plans to offer. This is a task where the "labels," or the credit groups, are not decided ahead of time—we simply want to know what clusters form, and design credit plans accordingly.

A clustering algorithm is initially given a set of objects and information about their similarity to one another. These objects are represented as a set of data points, each of which represents a particular customer. Each customer is differentiated by some information we know about them (also known as their features).[4] For example, take the objects labeled A through I in Figure 4. They have the following features:

Gender: A number, either 0 for male, 1 for female, or 2 otherwise.

Children: The number of children they have.

Age: Their current age.

[4] If the goal in mind is to eventually argue that clustering algorithms can produce natural classifications, it might seem suspect that the algorithm takes as input a set of features. These features seem to be carefully chosen, and therefore interest-dependent. In fact, input features such as gender are themselves controversially conventional kinds. I'd argue that the choice of input features may be interest-dependent, but the classification which results is distinctly new—just because the input features are interest-dependent does not mean that those same interests will decide the classification which is outputted. It is also a fact of *human* classification that the kinds we theorize about are interdependent, and many of our classifications cannot escape use of other systems of classification whether we take them as natural or not. For example, to say that elements are naturally divided by their atomic numbers supposes another natural division between the protons and neutrons in an atom's nucleus. If human-generated classifications rely on the use of other kinds, and we are still willing to argue that they might amount to knowledge of natural kinds, computer-generated classifications should be granted this reliance on other kinds (or input features) as well.

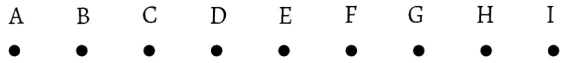

Figure 4: Initially, a clustering algorithm is given a set of unrelated data points.

Tax Bracket: A number from the range 1-7, each of which corresponds to one of the seven federal income tax brackets.

Distance to Nearest City: The number of miles that lie between a customer's residential address and the nearest city.

An example of a data point, for a 25 year old male from San Jose with two children, with a yearly salary of 50 thousand dollars, might look like this:

Person A: (0, 2, 25, 3, 42)

Aside from the set of data points, a clustering algorithm is also given information about how similar or dissimilar the objects are.[5] This is represented by a distance function, which tells us how "far" any two data points are from each other (in some sense, a measure of their dissimilarity). Here is an example possible distance function, taking as input two data points (customers) x and y:

$d(x, y)= 5 * |x_{gender} - y_{gender}| + |x_{child} - y_{child}| + 0.5 * |x_{age} - y_{age}| + |x_{taxbracket} - y_{taxbracket}| + |x_{distcity} - y_{distcity}|$

Given two data points, the distance function will compute a value to represent their dissimilarity based on their features, which I have depicted in Figure 5. All clustering

[5] There is an interdependence in human notions of kind and similarity that are not captured by clustering algorithms. It is often that we judge the strength (or naturalness) of classifications by whether or not they group together similar objects, but it also the case that we judge the similarity of two objects in terms of whether they are of the same kind or of a different kind. In other words, kinds are a function of similarity, but similarity is also a function of kinds. For example, we might be inclined to argue that humans are more similar to gorillas than to elephants because both are apes and elephants are not. We could also argue that what justifies the category of apes is the fact that it groups together similar animals, such as humans and gorillas. The classifications we take to be natural depend on what objects we think of as similar, but what objects we think of as similar also depend on what classifications we put them into. Clustering algorithms fail to capture this interdependence because they are given as input a distance function, which defines the similarity (or dissimilarity) between objects, and clusters them after the fact. The point here is just that more work needs to be done. The philosopher must resolve this circularity between similarity and kinds to understand how it affects our notion of naturalness. If what the computer scientist is after is a notion of naturalness that aligns to human thinking, they must reconcile this interdependence. It may be a reason to find plausible clustering algorithms that don't require predetermined distance functions.

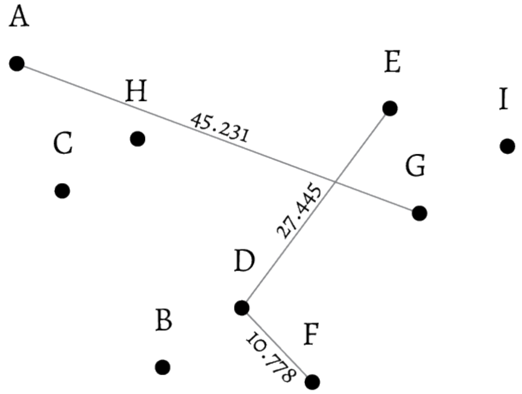

Figure 5: After a distance function has been determined on these data points, they will bear a spatial relationship to one another. As an example, three distance calculations are illustrated—each point is some distance away from every other point, which reflects how similar or dissimilar the two objects they represent are based on the input features.

algorithms start with these two basic inputs: the data points, and a distance function. From here, each algorithm implements a different procedure to determine the best way to group data points. Some procedures prefer clusters where the points within a single cluster are closer together, and some procedures emphasize clusters where the points in different clusters are farther apart. Some procedures group the data points pair by pair, and others will guess a couple of "central data points" and check whether the rest of the points fall around them reasonably. Regardless what procedure is used, all clustering algorithms eventually output one arrangement of clusters, which puts the original data points into separate groups, as illustrated in Figure 6. The output is a classification of the customers into distinct groups based on their background, which the company can then use to design credit plans to target to each corresponding group.

Clustering algorithms can be used in a wide range of cases, usually where the structure of a dataset is not obvious and needs to be revealed. For example, clustering can be used in market research to help companies find general "profiles" that a consumer base can be grouped into based on various features like gender, age, and buying habits. These algorithms can also be used in medical research to help doctors determine the right way to categorize

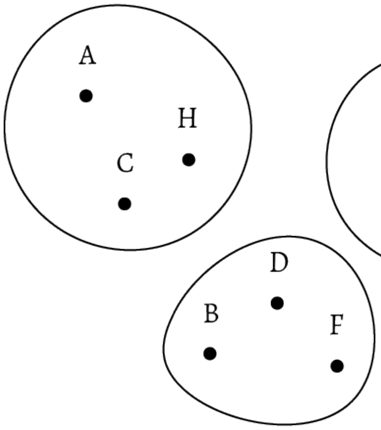

Figure 6: A clustering algorithm's goal, given a set of data points and a distance function, is to decide on the best way to group the points into clusters.

illnesses based on patient data—patients who have similar data will naturally fall into similar groups, which can then serve as a basis for diagnosis and treatment.

Clustering seems to solve a very familiar classification problem. Given a set of individual objects, these algorithms determine how they should be grouped, unaided by pre-existing knowledge about groupings. One might even say that the resulting classifications, given their lack of human supervision, arise *naturally*. It is in the remainder of this paper where I investigate if there is a philosophically plausible sense of "natural" that computationally-generated classifications really amount to.

3.2 A Defense for Comparing Human and Machine Classification

The comparison I am attempting to draw is extremely broad. Philosophers and computer scientists are both tinkering with the idea of classification, but with possibly different motivations in mind. I am not committed to the view that these two problems are identical or equivalent—rather, my goal here is to draw a couple of connections between them in the hope of finding some modest insights.

Before I move forward, I'd like to put some pressure on what this paper is generally trying to attempt, which is a comparison between two different takes on classification. The debate over natural kinds concerns the way that humans generate kinds, but clustering algorithms are specifically concerned with how machines generate kinds. What I am trying to do is take the insights drawn from clustering literature and apply them to our understanding of natural kinds, but a concern I have somewhat downplayed thus far is how comfortable we should be in treating these two types of classification processes (namely the human ones and the machine ones) as equivalent. Whether or not they are equivalent seems to be a consequential assumption, because if humans and machines have drastically different ways of generating classifications, it would restrict the extent to which I can draw this comparison. Before I dive too deep, I would like to put up a defense for this practice, at least in the way I intend to take it up.

The difference between machine and human classification, in the sense that matters for the case at hand, is what classifications they are capable of generating. In one sense, we can think of computers as being a human-operated tool, just as a hammer or a drill, which helps us accomplish goals that we could otherwise achieve ourselves if only at a slower and more inefficient pace. Taken this way, there is nothing really unique to what machines are doing—they just hold more computational power to draw out kinds that we could otherwise find ourselves. Taken another way, machines could be thought of as epistemically superior, where their algorithms bear a unique way of thinking that humans are not capable of even with all the time in the world. Clustering algorithms could then be generating classifications that humans could not comprehend, which might lead us to think that it is unfair to hold them to the same standards, or to take what computers do and argue that we work under the same conditions. What kind of knowledge, exactly, do the results of machine learning algorithms constitute? What relationship does that knowledge share with the limited, rational, human mind?

This question, even just as it relates to the project of classification, could be another paper unto itself. I would like to diminish its importance by arguing that, regardless of how we come to answer it, we could still make use of clustering algorithms to the extent that I implicate them in this paper. From a metaphysical perspective, we are concerned with whether natural kinds exist. This matter of existence seems to not depend at all on whether humans *or* machines come to know of these kinds, if they presumably hold independently.[6] In other words, regardless of what kinds we, or machines, come up with, there is (in the realist's eyes) a right answer to whether natural kinds exist or not. The concern I've highlighted above, about how machines generate kinds, is an epistemological one. However clustering algorithms come up with these classifications, we can admit that at the very least, the humans who use these algorithms in practice recognize what they generate and incorporate them into our knowledge of systems of classification. Clustering results, even if we were to regard them as distinct from the kinds that humans generate, are still "epistemological data" nonetheless—they are processes that we recognize as generating classifications, and therefore realists and conventionalists should both have accounts of what they amount to, metaphysically. Clustering is just another form of "epistemological data". If we use computational reasoning as one of our strategies to arrive at classifications, the theoretical results from clustering warrant an explanation by realists and conventionalists. We want to know if our clustering practices are better explained as a filter on natural kinds, or if they result in groupings just because we want them to (in the manner of a lathe).

The important takeaway is just that clustering algorithms generate classifications that humans both accept and incorporate into knowledge. Whether they do it in any manner that is comparable to humans is less important for my purposes here, because so long as we

[6] It is possible, however, that machines may not come to know of all the kinds that humans could. This depends on the sense in which we take "exist independently". If a kind exists independently in the sense that it exists in the world external to us, it can be argued that machines have access to the same world and can therefore discover the same natural kinds. If it is the case that the term "exists independently" is also used to refer to existence of mental states or conditions, there is a possibility that the natural kinds concerning humans are not perceivable by machines. Take color as an example. If we believe that our classification of color is a natural carving of human eyesight (but not of the world itself,) it is possible for humans to arrive at knowledge of this carving, because we perceive through the "carved lenses" of our photoreceptors. It seems, on the other hand, implausible that a machine could be given the visible light spectrum and instantly know how to carve out the categories of "red," "orange," "yellow," etc. This isn't to say it is impossible—it may be that a robot who observes humans for a long time could catch on to how we carve color, or that it can discover the right clusters by observing which light frequencies are also crayon colors and book covers. My point here is only that the fact of whether machines can catch on to the same classifications as humans is complicated by our exact metaphysical commitments on natural kinds.

take them to produce legitimate classifications, they are a part of how we find candidates for natural kinds. We should therefore still expect whatever metaphysical picture we choose to defend to be able to explain what it is we think we are arriving at.

4 Evaluation of Clustering Algorithms

There are a variety of uses for clustering techniques. One such use is to classify a set of data points and draw meaning out of it.[7] A credit card company can use a clustering algorithm to group their customers, and the output of the algorithm can then be interpreted as natural customer segments to design credit card offerings around. A doctor might cluster patients using data about their symptoms and preconditions in order to learn more about the kind of disease that is associated with a specific cluster.

Since clustering algorithms are used by a wide variety of people with different interests, there is usually little consideration for how "correct" their results are. They are correct if they serve the interests of the people who use them, whether that be to create useful genres for song recommendation engines or to effectively cater education programs to particular groups of children. In this way, the process of clustering is often treated like a lathe because the classifications that form are created for and only exist by way of what their intended use is, for humans. However, there have been several attempts to find a general procedure to evaluate the effectiveness of clustering algorithms, which tackles an overarching problem concerning all clustering applications. Ulrike Von Luxburg (2012) poses the question of whether there is a universal way to evaluate clustering algorithms. Is there something to be said about how good a clustering algorithm is, or if it gives us a better classification than another algorithm? An evaluation of clustering algorithms is a way to endorse certain classifications over others. The question I ask is, do these endorsements coincide with the

[7] There are other uses of clustering algorithms that are not as relevant for the comparison being made in this paper. For example, clustering can be used to reduce the dimensionality of data, which means that data points with many distinguishing features can be made simpler by figuring out which features are the most differentiating. These uses of clustering do not take the outputs to mean anything—they serve the sole function of making other computational tasks less intensive.

way that realists might endorse certain classifications as natural kinds? To reiterate some of the work in Section 2, we might suspect that certain classifications arrive at natural kinds because they demonstrate epistemic success, and this serves as “epistemological data” we must account for in our metaphysical views on natural kinds (the filter versus lathe metaphor). What follows is an introduction to the various forms of clustering evaluation available to us. I consider whether they endorse classifications in a similar manner to how we would endorse human-generated classifications, or in other words, whether their forms of reasoning coincide with our notion of epistemic success. If they can be taken as indicators of epistemic success, I argue that they become another form of epistemological data on natural kinds—it is then possible to implicate them in the debate expounded in Section 2 and investigate whether these accounts of epistemic success are better interpreted by the realist or by the conventionalist.

4.1 Three Levels of Evaluation

It is possible to evaluate a classification at three different levels of generality—the first is within a particular algorithm, the second is between different algorithms in a particular domain, and the third is domain-independently. I argue that the first level does not pick out classifications in any sense of the term “natural”. I argue that the second level picks out classifications that have epistemic success but, due to their domain-specific context, are better interpreted as conventional. I argue that the third level avoids this problem, but needs to meet several important conditions in order to be consistent with a realist view of natural kinds.

The narrowest form of evaluation is the first (Level 1). Given that we input a set of data points to an algorithm, it returns one possible arrangement of clusters. The algorithm must have a way to reason why it will return one set of clusters over another. This is one sense in which a classification is evaluated—the algorithm itself is a form of evaluation because it decides on what clustering it thinks is best for a given dataset. This evaluation is just a matter of the algorithm's mathematical clustering procedure. For example, an algorithm may prefer one clustering to another because it minimizes the total distance of points in the same cluster and maximizes the distance of the points in separate clusters. At this level, one classification is chosen over another without any consideration for the actual problem to which it is being applied. This kind of evaluation doesn't help us on the matter of judging whether classifications should be taken as natural, mainly because it doesn't concern how the classification is going to be used which means it has not yet been tested inductively. An algorithm can (and will) choose a classification, but it is not at this level that we have really committed to saying that the classification is meaningful or good in any sense, especially in the sense of epistemic success.

The next level of generality (Level 2) is evaluating different algorithms against each other in the context of a domain. At this level, we have an interest in mind when using clustering algorithms—they should give us classifications that will be helpful for a particular task. Suppose a data engineer at Spotify is asked to use clustering algorithms to improve their music recommendations. The engineer might try several different clustering algorithms on the music that a user has listened to a lot, generate genres around them, and then recommend new music based on those genres. The algorithms will each return some set of clusters, which is an example of each algorithm making its own Level 1 evaluation. At the second level, the engineer will have to evaluate which of these algorithms provided the most successful genres for recommending new music to users. Here, inductive success plays a factor. The engineer will be able to see, as time goes on, which newly recommended songs were listened to more and which genres were used to recommend those songs. The clustering algorithm which produced those genres might then be endorsed for choosing classifications which are more predictive of a user's music taste than others. Through this process, there is a way to evaluate different clustering algorithms in terms of how they succeed in a given domain.

This kind of evaluation and endorsement of classifications is more relevant to our comparison to natural kinds, because it is at this level that we begin to favor classifications with epistemic success. We endorse clustering algorithms, or (to draw out the terminological connection) endorse inductive practices, on the basis of their track record of consistency with phenomena, which varies from application to application. In one case, the phenomena might be the attraction of new listeners to playlists, in another it may be the advances in the study of new disease categories. Clustering results can have epistemic success, which might lead the realist to wonder: do these algorithms arrive at candidate natural kinds? We could be led to believe, at this point, that clustering algorithms serve as a kind of filtering mechanism. They pick out natural kinds through a process of inductive testing. We see which classifications are predictive and which are not, and through many iterations, we slowly triangulate onto the kinds which reflect the way the world is structured.

What is troubling, though, is that our methods of evaluation are still contextualized to a particular practice or domain of usage. How we judge the epistemic success of clustering algorithms differs from field to field. It is true that we are guided by something stronger than simply a mathematical preference for one classification over another, since they must stand the test of induction, but it is still humans who set these inductive interests to begin with. The conventionalist has reason to push back and argue that the clustering results we take to be predictive are still just legitimated by our interests and could still be otherwise—the classifications did not “exist” independent of us, and it is our human interests which determine the shape of the wood on the lathe and therefore what it is our algorithms are carving around.

Rather than pushing back on this point about whether domain-interested (Level 2) evaluation procedures can be made consistent with natural kind realism, I think there is still another path yet unexplored. This brings me to the third and most general way that people have attempted to evaluate clustering algorithms, which is in a domain-independent manner. These evaluation mechanisms attempt to distinguish good classifications from bad ones without regard for what the results will be used for. They will argue for a way to judge the quality of clustering algorithms and their results in a universal way that applies to any and all algorithms. I hope to build a clearer picture of what this entails in the next section, where I attend to some concrete examples of these evaluation mechanisms. For now, I want to start a broader discussion of, if there are universal clustering standards, what the implications would be in our debate about whether classification algorithms work as filters or lathes. I hope this will give some initial sense of what we could conclude about these evaluations, or what they would need to have in order to be interpreted as either a filter or a lathe. It will become apparent that each of the specific attempts at universal evaluation will lead us to slightly different ways of thinking about realism and conventionalism.

4.2 Domain-Independent Clustering Evaluation

A domain-independent (Level 3) clustering evaluation procedure will ideally give us a way to distinguish good clustering algorithms from bad ones, without having to appeal to the domain in which they are used. I argue that if this is successful, the realism interpretation is more plausible—the existence of these evaluation standards seems better accounted for by the view that we are discovering natural kinds rather than the view that we are inventing them through anthropocentric interests. I will first try to explain why a Level 3 evaluation would be incompatible with the "lathe" theory, which may help motivate a second explanation that attempts to account for this in "filter-realist" terms.

To reiterate, we take the lathe account of natural kinds to be saying that the classifications we endorse by epistemic success are not natural kinds, because all classifications are conventional. In other words, they take the shape that we are interested in seeing them take rather than the shape that exists in the world. They come about because of the purposes they serve us. From this position, it seems implausible that we could come up with a *general* justification for the classifications we endorse—the conventionalist argues that the only reason we could make such an endorsement to begin with is because they align to specific interests. If clustering algorithms produce different candidates for natural kinds, and a Level 3 evaluation procedure judges these algorithms regardless of their application to any particular domain of human interest, it must be judging these classifications in an interest-independent sense. Our evaluation gives us a set of classifications that we can endorse without a domain-related justification. How, then, could these classifications have come about in a conventional way? A domain-independent evaluation metric gives us a kind of epistemological data that seems incompatible with conventionalism. It seems that we have stumbled upon a way to pick out classifications that sidesteps the worry of how they align to our domain-related interests.

Domain-independent clustering evaluations seem to put pressure on the conventionalist, but I would like to frame this argument more positively in terms of compatibility with a realist account of natural kinds. As posed earlier in Section 2, the realist interprets epistemic success as a way of indicating to us which of our classifications are better fit to understand the world, and are therefore triangulating onto natural kinds. But as drawn out from the snack food company example, inductive success is not enough to call our classifications natural—some classifications are only predictive about certain things. In addition, some classifications have been more inductively tested than others. The snack food company used a country-based classification to predict consumer buying habits, but a classification like the Periodic Table has arguably *more* epistemic success because it has been used to predict countless chemical phenomena for hundreds of years. The realist is in search of an indicator of naturalness that is over and above epistemic success—the classifications cannot just be predictive for specific inductive interests, and we must also be able to measure how some classifications are more or less predictive than others. Level 3 evaluations satisfy this first requirement. Because they are domain-independent, they can judge a classification without regard for predictive interests we have in mind. Level 3 evaluations should also satisfy the second requirement. They must be able to measure, across all classifications, the *degree* of epistemic success. Because these evaluations are universal, they can judge any and all clustering algorithms and outputs, which gives us a universal measure. The question is whether that universal measure is tracking epistemic success.

The problem is this. Level 2 evaluations capture the important requirement that we have set out thus far, which is that our best candidates for natural kinds should be predictive. By applying algorithms in a particular domain, we can tell which of the classifications we generate carry more or less epistemic success. The worry is that epistemic success, at Level 2, is contextualized to particular domain interests, which means that they are only good at predicting the things we want it to predict. Level 3 evaluations seem to solve this problem because they endorse classifications domain-independently, but how do they capture the notion of epistemic success without regard for a particular domain? In other words, we cannot endorse classifications that only serve particular inductive interests (because they would be taken as conventional), but we also cannot endorse classifications that do not have inductive success at all, because such is the motivation we have for thinking classifications might be natural in the first place. There are a couple ways for us to deal with this inductive consideration.

We can take Level 3 evaluations as a way to filter our Level 2-endorsed classifications. In this way, Level 3 evaluations are still endorsing classifications domain-independently, but the classifications must first pass the test of a Level 2 classification, namely that they have proven some level of epistemic success within a particular domain. By adopting a two-tier evaluation, we can guarantee that we only endorse classifications that are well suited to our inductive interests, but in addition, we do not necessarily endorse them solely by these inductive interests. We pick the ones that not only succeed in their domain, but also satisfy a more universal evaluative standard. Under this view, Level 3 classifications do not need to consider epistemic success, because it is Level 2 classifications that first filter classifications that meet this requirement. The problem, though, is that we are still missing a way to measure the *degree* of epistemic success if we only consider epistemic success at Level 2. For example, two different domain-specific evaluation procedures might endorse our snack-food company's country-based classification and the Periodic Table, but if a Level 3 evaluation procedure has no way of comparing one classification's inductive success to another's, it might arbitrarily choose the former classification over the latter, even though it is part of the realist's view that classifications with more epistemic success should be more indicative of naturalness. Our Level 3 evaluation procedure must not only be universal in the domain-independent sense, but should also be universal in the sense that it gives us a universal metric of epistemic success. Otherwise, there is no way to judge the predictive strength of one classification over another.

In order for us to plausibly interpret Level 3 evaluation procedures as picking out classifications in a realist manner, they must offer a way to judge the degree of a classification's predictive strength, and therefore the degree to which we believe it to be natural. The realist is in search of a Level 3 evaluation that does not only endorse classifications domain-independently, but also tracks and measures some notion of epistemic success. In the next section, I consider three different proposals for Level 3 evaluations. For each, I will ask whether it satisfies the requirements necessary for us to take it as epistemological evidence that is more plausibly interpreted as filtering natural kinds than creating merely conventional classifications. Namely, the evaluation should be interest-dependent, and it should be able to measure and discriminate between different degrees of epistemic success.

5 Proposals for Domain-Independent Clustering Evaluation

At the most general level, there have been several attempts to propose a method of evaluating, roughly, the "good-ness" of classifications outputted by clustering algorithms. What we are looking to do is consider whether these evaluation standards are better interpreted as picking out conventional classifications or natural kinds. The following are three proposals for evaluation that Luxburg (2012) expounds.

5.1 Benchmark Datasets

The quality of a clustering algorithm can be tested by running them on datasets where we believe we already have the "right answer" to. The idea of evaluating on benchmark datasets is to pick out a couple instances of classification where we are fairly certain about how things should be grouped to see if the clustering algorithm in question will arrive at the same grouping. For example, if we have a group of images that are either pictures of cars or of other random objects, we are fairly sure that the pictures of cars should be one of the outputted clusters. We can then take this dataset to be a benchmark and run all of our candidate clustering algorithms against it—the ones which correctly group the cars will be endorsed, and the ones that group them in some other way will not. These benchmark datasets can be decided by field consensus, but will generally be cases where there seems to be a "correct" grouping. In other words, we test how well our algorithms find our best natural kinds candidates, and if they do well, we endorse the other classifications that those algorithms generate.

This appears to be a reasonable way to evaluate clustering algorithms. If they perform well in instances where we know how things should be classified, we will expect that they should similarly perform well in instances where we are not as certain. The status of benchmark datasets is similar to, in scientific realism, the idea of our "best scientific theories." These are the theories which have accurately explained evidence and predicted new observations time after time, so often that we hold them to high epistemic regard as being candidates for yielding genuine knowledge of the world. These theories are often the ones we use to reason about other theories that we are less certain of—scientists are more inclined to explore hypotheses that are consistent with the best theories out there so far, since the hypotheses that contradict them are less likely to be confirmed. Similarly, if we choose benchmark datasets that reflect the classifications we take, so far, to be our best candidates for natural kinds, the clustering algorithms we endorse will be ones that consistently produce groupings compatible with them. For instance, chemical elements are taken to be one of our best candidates for natural kinds, because so many of our theories make good use of them and no other alternative for grouping atoms has been nearly as predictive. Consider a benchmark dataset comprised of various data points which each represent atoms, with their various properties contained as features. If a clustering algorithm were able to correctly group these data points by their number of protons, we would think the algorithm has arrived at the "correct" grouping. If the same algorithm performed well on other paradigm examples of natural kinds, we could be convinced that the algorithm has generally found the right way to reason about kinds and can be used to bring more clarity to datasets we don't yet know how to classify.

An initial problem with this procedure is that it assumes that algorithms that perform well on some datasets will also perform well on others, which may not be the case. It might happen to "answer correctly" on our benchmark datasets, but when applying the same lines of reasoning to others, it may answer very poorly. A more pressing problem, though, is that it seems to select classifications in an interest-dependent way. In some sense, this procedure evaluates algorithms domain-independently, since it grades them on a general set of benchmark datasets, but these datasets themselves will lie in particular domains, which means that the algorithms we endorse will tend to work well for the domains included in our benchmark. As mentioned in the last section, it is also important that a Level 3 evaluation measures whether some classifications are more or less inductively successful, but an algorithm's ability to classify benchmark datasets properly does not necessarily mean that the classifications it produces are predictive in any way.

This first proposal falls short in a couple of places, partly because it depends too heavily on how we regard classifications we already take to be "natural" and not enough on the actual algorithms themselves, which gives no guarantee on how well they perform on new datasets.

5.2 Convergence

David Pollard (1981) proves that the k-means clustering algorithm converges almost surely as sample sizes increase. The k-means algorithm refers to a clustering technique where the number of clusters is given beforehand, as a parameter k. For example, a k-means clustering where $k = 5$ will find a way to cluster data points into 5 different clusters, where each cluster

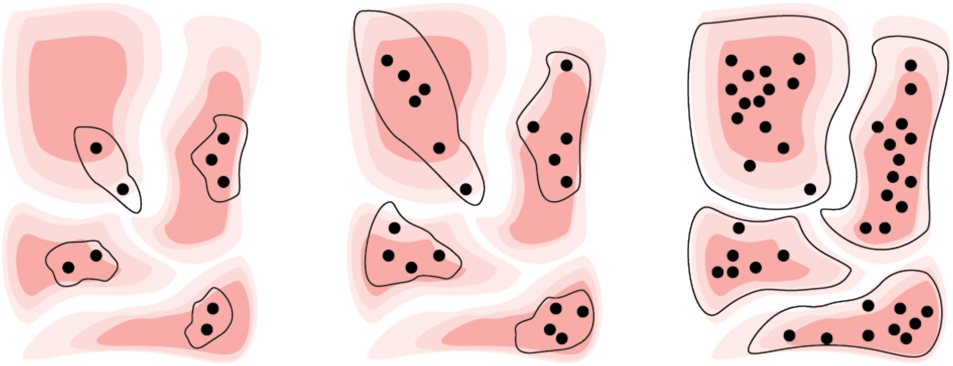

Figure 7: As more and more data points are sampled, an algorithm that converges will eventually find clusters that align to the underlying distribution of the dataset.

is defined by some “central” point inside of it. The result from Pollard shows that as you continue to draw new sample points from the same distribution, the algorithm will converge upon a clustering almost surely. “Almost surely” is a probabilistic term, which roughly refers to a form of convergence where the result of the clustering algorithm will start to look like to the underlying distribution of the data. The k-means algorithm is shown to eventually settle on a set of clusters that, as the number of data points gets larger, matches the underlying distribution of the dataset. I have illustrated an example in Figure 7.

This procedure of evaluation is, then, just to ask whether a clustering algorithm converges or not. If it can be proven to converge, we endorse the classification. Convergence avoids some of the pitfalls of benchmark datasets. It seems to be truly domain-independent, because it is not up to us whether or not the data converges—it will uncover whatever is true to the nature of the dataset. If a clustering algorithm converges on a classification, we cannot simply alter it to our interests. It can only ever converge on whatever distribution is true to the data points that are drawn.

The question that remains is if convergence captures some notion of epistemic success, and whether it does so in a way that is more compatible with the realist or conventionalist interpretation. If a classification converges, it tells us that as more data points are added, they will start to consistently fall into the same groupings. This suggests that the the classification is predictive—if we can prove that it converges, it means our classification will slowly align to the underlying distribution of the dataset, and we can therefore become increasingly certain about where the data points will fall. The classifications that converge seem to also be the ones with epistemic success, but where the problem lies is that convergence cannot judge *degrees* of epistemic success.

Convergence is an absolute measure—an algorithm either converges, or it does not. This evaluation procedure is a fact about a mathematical limit, which means that classifications are either proven to converge or proven to not converge. Because it is absolute, there is no sense in which, under this evaluation, one classification is more or less predictive. In fact, convergence implies that the algorithms which converge are *certainly* predictive, meaning that as we take in more data points, we can be almost sure that we have landed on the underlying structure of the dataset. If the realist is just looking for a way to judge whether one classification is more predictive than another, in a relative manner, how did we land on something stronger? It seems implausible that we could even say something stronger about our classifications. Given that we generate classifications within the limited inventory of human experience, it seems that we could not say absolutely that our best classifications are indeed natural kinds, just that they are more likely to coincide with natural kinds than others. Our process of generating kinds is only so much as a game of induction, the same way that science is—we can be optimistic about our best scientific theories, but there is always the possibility that they are wrong, because we can only theorize with the inventory of limited human experience. Convergence seems to endorse classifications in a way that has transcended epistemic limits. What then, is the problem with how we have taken the fact of an algorithm's convergence?

The main underlying problem is that Pollard's proof of k-means convergence assumes as a given fact that our data points are drawn from an underlying distribution. What we are taking as fact is that such a distribution does exist and showing that a k-means clustering algorithm converges upon it. What we don't have, but what we want, is the implication in the other direction—that an algorithm whose clustering output converges implies that an underlying distribution exists. The notion of convergence does not cohere without a given distribution that points are drawn from, but what we are looking to use convergence for is to ask if such a distribution exists at all. We are somewhat misusing the idea of convergence to make realist implications that were already assumed to be there.

Another limitation of convergence is that it has been shown only for k-means clustering algorithms, a subset of clustering algorithms where we decide beforehand how many clusters we wish to group the data points into. This assumption does not seem natural by any means: how could we be sure that the natural world is split into any particular number of kinds? The number of clusters, k, is human input that is taken as information for k-means clustering algorithms to produce classifications. The results, even if they were to converge, assume a structure to the data that we cannot be certain actually holds. A given algorithm may only converge if we decide it should output 5 clusters, but it would be wrong to imply that we have therefore discovered that there are naturally 5 kinds. It may be that the natural world is carved into 11 kinds, but if k-means algorithms do not converge at $k = 11$, convergence would not be a good indication of what is natural.

To summarize, convergence is an evaluation procedure that endorses classifications that provably arrive at a dataset's underlying distribution. It is better than benchmark datasets because it reflects a domain-independent fact about any clustering algorithm, but it fails to plausibly suggest that the classifications it endorses are natural, because it assumes that there is a natural division to the objects it classifies in the first place. Lastly, I will ask whether the metric of stability solves these problems—whether or not it can capture the notion of epistemic success domain-independently (in the manner of a filter,) and measure relative degrees of predictive strength.

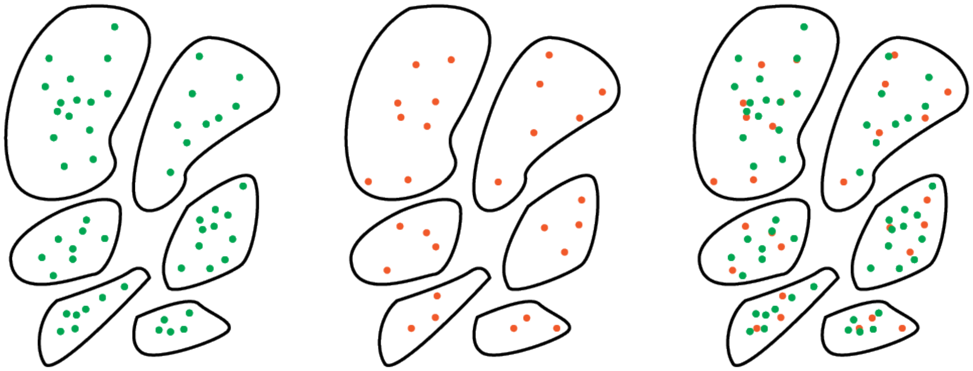

Figure 8: A stable clustering is one that maintains its clusters even as different data points are taken into account. The example above illustrates a cluster that is stable over two halves of the dataset.

5.3 Stability

Stability is a measure of how much a cluster changes based on what data points are used in generating it. For a given algorithm, we can perform clustering multiple times, each on a different set of data points. A stable result is one in which the resulting groupings do not change much from one set of data points to another. In other words, the algorithm should maintain the same clusters even as we fluctuate what data points we feed into it. This is illustrated in Figure 8.

There are several ways that clustering stability can be tested (Luxburg 2010). We might choose to run a clustering algorithm on a random half of our dataset, run it again on the other half, and then see how closely the two clusters resemble each other. We might also choose to initially cluster on all of the available points and see how much the results change if we only provide the algorithm smaller subsets of the data. In all cases, the general principle remains the same: a stable clustering will commit to more or less the same groups even as we alter, remove, or add data points into consideration.

Stability is strong on the issue of domain-independence. It describes a generic property of the algorithm's output, as opposed to how benchmark dataset evaluations use a particular choice of classifications to grade algorithms against. The property of stability does not depend on how we think of it in any particular domain, since it refers to simply how well a grouping can accommodate data points, a fact that applies to any and all applications of clustering.

Stability also captures the requirement of epistemic success. What we want is for the classifications that we endorse to explain past events and to hold consistent for future events. If we take past events to be the data points we already have, a clustering algorithm that produces a stable result will explain them well because it is on that basis that it arrives at its clustering in the first place. Although it cannot be said what new data points might be generated, the fact that a stable clustering can withstand varying data points seems to suggest that it will be consistent as new points come along. Otherwise, we would simply say that it is not a stable classification.

Stability is also strong where convergence is not—it is a relative metric, because certain clustering results can be more or less stable based on how many data points it remains stable across. This gives us an evaluation procedure which can judge, between two classifications that both display epistemic success, which of them is stronger. Because stability is a relative metric, it suggests a version of natural kind realism that admits to "degrees" of naturalness. We can take this two ways—it either means that there are varying degrees to how certain we can be about our classifications, or that natural kinds themselves only exist in degrees of reality. This is a distinction between degrees in an epistemological sense, and degrees in a metaphysical sense. The latter sense is more difficult to defend and to make sense of, but it seems compatible to at least argue for the former, that our epistemic beliefs about what the natural kinds are admits to degrees, just as our best scientific theories are only the best because we take them with higher degrees of certainty than other theories.

Is stability, as a metric of epistemic success, better interpreted as a filter or a lathe? The key strength of stability is that it endorses classifications which do not vary based on the data points it is created from, or in other words, the observations that it takes into account. A stable clustering should maintain its structure even if we were to give it a different subset of objects in the dataset—the resulting classification is only loyal to the underlying structure of the dataset (if one exists). If natural divisions exist, then the data points that are generated will also have some natural tendency to fall along those divisions, and stability will be able to detect when this tends to occur. If different sets of data points all seem to conform to one stable clustering configuration, it would suggest that there is some natural division that lies underneath them. Stability seems to filter out human-independent classifications because if they were merely conventional, the data points would never cluster in a stable way since there is no necessary "shape" which they should stabilize to—it would be possible for one subset of the data points to cluster one way, and another subset of data points to cluster in another way. [8]

Another way to put this is that the conventionalist lathe seems to imply that the groupings we create are self-interested because of the motivations we have for using them. A stable clustering is resistant to these interests, since it withstands the variability of the data points we take into account. It does not "decide" to be stable or unstable based on what shape we want the groupings to take. It will be what it is, solely as a function of all the data points we test against it at any point in time. It operates as a filter by rejecting unstable groupings in favor of what naturally arises from the tendencies of any and all data points that are considered.

Out of the three considered, stability seems to be the universal clustering evaluation metric that most plausibly suggests that natural kinds exist because it picks out arrangements of clusters that, when given one set of data points, are predictive of how other data points will

[8] As I've outlined earlier, my discussion is contextualized by a broader view of natural kinds realism that is only committed to the idea that natural kinds exist, even though different metaphysical views disagree over which of them do. Stability only endorses classifications that remain stable over time. Whether we think it picks out natural kinds is complicated by what types of classification we think are natural. For example, certain classifications of humans may be useful for social scientists in predicting future behavior, but as migratory patterns or other social phenomena affect humans, their classifications will also evolve. A stability evaluation will discriminate against these kinds of classifications because their underlying tendencies will change (and therefore their clustering will be unstable), even though they are nonetheless valuable for certain predictive interests, and under *certain* views of natural kind realism, may still be considered natural kinds.

fit. Stability is a way to capture the idea of prediction without being tied to any particular domain interest, and domain interest is what the conventionalist thinks that prediction can only amount to. I think there is still room to push back and interrogate whether stability really does capture prediction in all the senses that we normally take prediction in science and everyday life. This requires a deeper philosophical investigation into what prediction precisely means—what counts as evidence of an accurate prediction, what counts as prediction or merely self-confirmation, and what we mean when we judge one classification to have *more* epistemic success than another.

This concludes the sections that discuss clustering evaluation procedures. To summarize, evaluations by benchmark datasets fail to be truly domain-independent, and the property of convergence implausibly implies that there are classifications which we certainly know to be natural. Stability offers a way to capture the idea of epistemic success of classifications in a domain-independent way, and it also gives us a way to judge the degrees of our epistemic certainty of natural kinds. Thus, stability is the most compatible with a realist interpretation of epistemic success, although this can be further tested by clearing up what should be meant by epistemic success or prediction. The next section looks at a different theoretical result that doesn't necessarily have to do with evaluation, but nevertheless asks if there is a limit to what clustering algorithms can discover about natural kinds.

6 Kleinberg's Impossibility Theorem

As a consequence of the broad array of use cases for clustering (and therefore the broad array of implementations of clustering algorithms,) there has been a large deal of ambiguity about how we could reason about these algorithms in a general way. Evaluation is one attempt to solve this problem, but yet another way to generally reason about clustering algorithms is to formalize the task of clustering and to ask if there are any basic facts that hold true of all clustering algorithms. Jon Kleinberg (2003) finds an interesting result, which is that, given three basic properties which we intuitively hold to be true across any and all clustering algorithms, it is provably impossible to satisfy all of them at once. This is proven by assuming any two of these properties, and mathematically deducing that satisfying the third is impossible.

This brings out a surprising limitation of clustering, which is that any algorithm will have to sacrifice one of these three basic properties, and that clustering necessitates trade-offs. This section will look at what these three properties are and what they mean in relation to our more general notions of human classification, to ask if the computational impossibility that Kleinberg proves is also a human, epistemic impossibility when it comes to our ability to discover natural kinds (if we take them to exist).

6.1 Scale Invariance

The first of these is a property that concerns the relationship that holds between any two data points. This relationship is encoded in a clustering problem as a distance function. To reiterate, a distance function takes as input the features of each data point and returns a number that represents how "far" they are from each other. A clustering algorithm is considered scale-invariant if it returns the same clustering of a set of points even if we were to scale all distance values between points either up or down (as illustrated in Figure 9). For example, if we have several data points whose closeness measures fall somewhere between 0 and 10, we should be able to multiply each of their distances by 5, and a scale-invariant clustering algorithm run on these points should return the same clusters.

Scale-invariance should hold for clustering algorithms because it holds for the way we normally produce classifications. There is no particular "scale" to which we think about the relationship between objects. In human terms, objects simply can be more or less similar to one another, at least as far as we are concerned with in discussing kinds. We take it that objects which are more similar should be in the same cluster, and objects that are more dissimilar should be in different ones. If distance functions encode this similarity relationship, they should hold for any scale—distance is only a numerical representation of

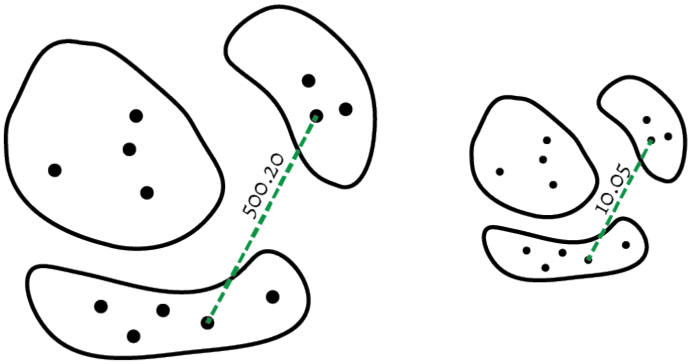

Figure 9: Scale invariance is satisfied if an algorithm returns the same clustering, even as all the distances between points are scaled up or down.

what is otherwise a comparative metric by human terms, so if these numbers were to scale across the board, they should not change how the objects are grouped.

6.2 Consistency

Consistency is the second property we take as desirable for clustering algorithms. The basic idea is this: for a given clustering result, if we shrink the distances between points in the same cluster and expand the points between different clusters, we should get the same clustering if we run the algorithm on these revised distance measures. To put it another way, we should be able to push the points inside of a cluster closer together and the points outside of a cluster farther apart, but end up with the same result (illustrated in Figure 10).

For similar reasons as scale-invariance, this property should necessarily hold for any general process of classification. If distances are being shrunk within clusters and expanded between clusters, we are only "committing harder" to the similarity beliefs we had to begin with. Objects in the same cluster are now considered more similar to each other, and objects in different clusters are considered more dissimilar. It shouldn't be the case that, in human classification processes, such a revision would motivate someone to decide that a object belongs to a different clustering—if anything, it should make us more certain of the

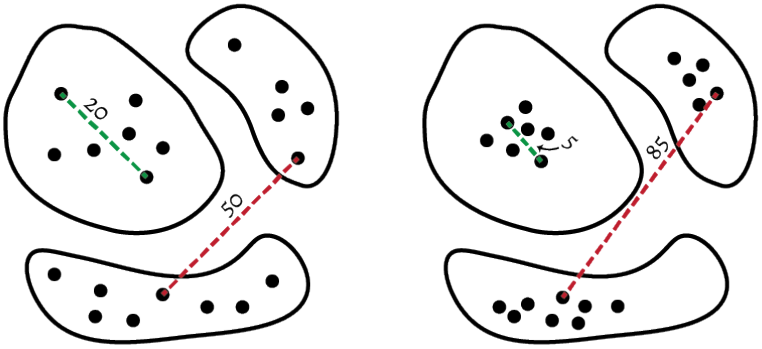

Figure 10: Consistency is satisfied if an algorithm returns the same clustering even if points within a cluster are moved closer and points between clusters are moved farther apart.

categories we initially started with.

6.3 Richness

The final property is richness. It roughly says that any possible arrangement of clusters could be outputted for a given distance function between a set of points, as illustrated in Figure 11. This isn't to say that every possible clustering must be favorable or practical (in a loose sense) in some way, but just that no possibility has been structurally excluded from consideration. If we could play around with the distance measures between points, a clustering algorithm should theoretically be able to output any arrangement of clusters—this includes groupings where every point is its own cluster, every point is part of one large cluster, any point can be clustered with any other point, etc.

In general terms, what richness holds is that no structure of the data should be out of the question—that our algorithms should be open to any and all possibilities. Our algorithms should only choose one grouping of objects over another for reasons of the perceived similarity between them. If we are potentially in search of natural kinds, we should maintain richness as a property. We don't have certain knowledge of the structure of the world and how it is divided, so our epistemic practices for discovering kinds should also maintain that any

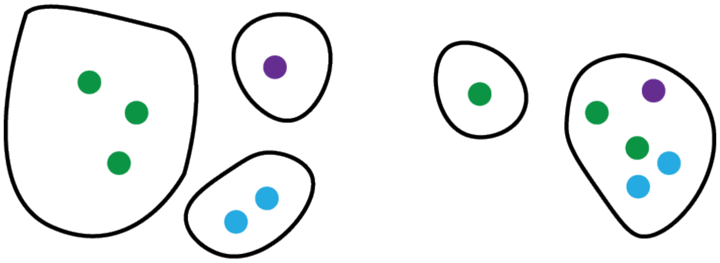

Figure 11: Richness is satisfied if an algorithm is able to return any of the possible arrangements of the same data points, given the right distance function.

structure is epistemically possible. There is always the chance that future observations will lead us to revise and generate new arrangements of clusters, so to foreclose that possibility would be an unmotivated limitation.

6.4 From Impossibility to Possibility

As much as it seems that scale-invariance, consistency, and richness are all fundamental properties that should hold for our practices of classification (and therefore the clustering algorithms we use along the way,) Kleinberg has proven that no single algorithm can satisfy all three of these. How are we to deal with this impossibility?

We could initially just argue that Kleinberg has abstracted the task of machine-based classifications in a way that does not also abstract human-based classification, in which case impossibility serves as only a constraint on the former and not the latter. Kleinberg's result is theoretically true for clustering algorithms, but it might just mean that computers are "weaker" classifiers than humans. It goes back to the overarching comparison problem that I touched on earlier—computers and humans are different epistemic agents when it comes to generating kinds. Kleinberg's Impossibility Theorem might lead us to think that we should just be more pessimistic about whether machines can arrive at natural kinds, but it still remains a possibility that humans may not be subject the same impossibility constraints. Where the problem lies, then, is in the fact that we haven't properly formalized clustering algorithms such that they can fully replicate how we, as humans, generate kinds.

But given the more intuitive explanations I provided in the above subsections, I maintain that each of the constraints seem to not only hold for algorithms, but hold true for human reasoning about classification, at least if want to remain consistent with the realist possibility that we can discover natural kinds. So I pose the question: what if we are indeed subject to the same impossibility constraints? Where can we find wiggle room in these three properties where we would still be able to make classifications, and more importantly, do so in a way that could be thought of as discovering natural kinds?

Richness is where we could push back a bit. First, it seems that we might be able to structurally exclude trivial classifications, namely the output that assigns each data point into its own cluster, or the output that clusters all of them together. Each of these are somewhat meaningless classifications, because they don't create any division between the objects. They might be a valid output for a clustering algorithm, but to the extent of how, and more importantly, why we generate classifications in the first place, they wouldn't be meaningful results. We care to find classifications where multiple objects participate in a single grouping, and at the same time, at least two distinguishing groups are meaningfully categorized apart from one another. This is one way we could reasonably violate the third property.

In a less trivial way, we could also just concede that some of the possible classifications of our objects *should* be excluded. As exemplified in the k-means algorithm, one way they could be excluded is if we specifically require that the data points settle into a certain number of clusters. A question remains as to whether this would prevent us from being able to uncover natural kinds, if we are admitting to, by our own volition, excluding some structures of the data from being possible outputs—the ones we exclude might just be the natural way to classify the objects (if one does exist), in which case we have prevented ourselves from discovering natural kinds. In defense, I argue that this concern can be distinguished from our other conventionalist concerns. Violating richness only admits to influencing the *structure* of our resulting classifications, but if, by some viable evaluation procedure, we still find that the classifications we endorse have epistemic success, it may just be the case that the natural world also happens to not align to the sorts of clustering structures that we have ourselves excluded. This can be accounted for in a revised picture of filtering. The fact that we violate richness implies that the filters we have are imprecise, and that some of the natural kinds which exist cannot be discovered because of assumptions we make about which clusterings are possible. But an imprecise filter, in this sense, only implies that some natural kinds won't make their way through it—it does not admit conventional classifications to "leak through" (because they would not be endorsed by an indicator like stability). We can still be realists about the classifications that we endorse, it might just be that there are some natural kinds that we miss out on because we have violated richness. We are simply forced to admit to our epistemological limitations, namely that we will not be able to uncover all of the different natural classification structures that exist in the world.

7 Conclusion

I want to now summarize the work I have done in this thesis and briefly discuss some lingering questions. The broader intention of this paper has been to draw a conceptual connection between the philosophical debate over natural kinds and the theoretical work concerning unsupervised clustering algorithms. These two problems are generally concerned with the same practice of classification—we take a set of objects, and on some basis, sort them into groups.

The philosophical question that arises from classification is whether there is a natural way to group these objects, and if so, whether it is possible that our practices of classification, whether in the sciences or in every life, might allow us to discover them. The realist believes that there is a natural way to group these objects, and the conventionalist believes that there is not. There are many ways to understand what the term “natural” means, as well as what criterion holds for a kind to be considered natural. As a result, there are many versions of natural kind realism, each of which is committed to a different metaphysical view. I have argued, however, that in a broad sense, they all disagree with conventionalists over at least one thing: epistemic success. It is a wonder how some of the classifications we make fare so well with new information—as we gather more data about objects, they continue to conform well to classifications we have already made with them. The realist, who believes that natural kinds exist (in at least some broad sense) will argue that these classifications reflect the way things must naturally be grouped if it can predict how new information will conform to it. The conventionalist might instead argue that what we have taken to be prediction is really just self-confirmation. In other words, the classifications that succeed only do so because they succeed at what we want them to. I used Chakravartty’s metaphor of a filter and a lathe to elucidate this difference in interpretation.

We take epistemic success to be a virtue of our best classifications, and the realist and the conventionalist disagree over whether we should take this as indication of some notion of naturalness or not. It is here that the computational question of classification becomes relevant. Clustering algorithms are another way that humans generate classifications of objects (as I have argued in Section 3.2). Is there a way to know if the classifications they output have epistemic success? If they do, is it better interpreted from the realist or the conventionalist angle? In Section 4, I introduced three different ways that clustering algorithms are evaluated, and argued that the sort of evaluation procedure that can be universally applied to clustering algorithms domain-independently is the most plausibly consistent with a realism interpretation of natural kinds. In Section 5, I discussed whether three particular proposals for this form of evaluation live up to this standard. Evaluations by benchmark datasets fail to be truly domain-independent, and the property of convergence implausibly implies that there are classifications which we certainly know to be natural. Stability offers a way to capture the idea of epistemic success of classifications in a domain-independent way, and it also gives us a way to judge the degrees of our epistemic certainty of natural kinds.

Lastly, I discussed the limitations to our ability to filter natural kinds by Kleinberg's Impossibility Theorem in Section 6. There are three basic properties of classifications that hold in both clustering algorithms and in what we would require in a realist account of classification, but I argued that richness is a property that can be relaxed and still remain consistent with the belief that it is possible for us to have knowledge of natural kinds.

Aside from simply arguing in this paper that there are similarities between how philosophers and computer scientists think about classification, I have tried to make the stronger point that the comparison is genuinely consequential for both sides—that philosophical work has implications for computer scientists, and that the work of computer scientists has implications for how we philosophize about natural kinds. This paper has mainly focused on the latter direction of influence, but it is interesting to consider what value the philosophical literature on natural kinds offers to computer scientists. Clustering algorithms are commonly applied with the idea that they will reveal a "natural" grouping of data. Baked into this belief about clustering algorithms is an unsure concept of "natural". As Von Luxburg wrote, "It is often presumed that for any situation where clustering may be used there is a single 'right' clustering" (2012, p66). If it is the proclaimed goal of clustering algorithms to discover a natural grouping, it seems important to know whether natural kinds exist. If they do, it would be important for computer scientists to be clear, philosophically, about what "natural" means in order their for clustering algorithms, in both implementation and evaluation, to accurately coincide with such a concept. If they do not, then computer scientists must be forced to revise their intentions and be clearer about what exactly is being produced by their classification algorithms, if not what was mistakenly taken as natural.

I will leave off with a couple of thoughts on how the conversation between philosophers and computer scientists on the issue of classification might continue as algorithms become increasingly capable of conducting human labor. Classifications are an important part of the natural and social sciences. In fact, it could be argued that classifications are fundamental to scientific understanding—we theorize in terms of kinds, and progress in scientific fields is often taken by the revision or evolution of the kinds we adopt. If clustering algorithms engage in classification in the same way humans do, and there is an inevitable future where clustering algorithms improve at what they set out to do, it presents an interesting dilemma where clustering algorithms may at some point be better classifiers than humans. If they become better at classifying, they will also become better at science. Under these circumstances, it seems possible to ask whether science is a kind of human labor that machines could eventually overtake. If clustering algorithms, today, can be said to discover facts about the world, couldn't the clustering algorithms of tomorrow be *in charge* of this process of discovery? I would argue that in order to defend against this possibility, there must be something fundamentally human to either our process of generating classifications or of endorsing them. It presents an interesting philosophical question regarding the limits of computation in terms of what knowledge can be generated and what sort of epistemic authority algorithms can potentially hold. As algorithms begin to play larger roles in the world around us, it feels imperative that our philosophical investigations of them do as well.

References

[1] Chakravartty, Anjan. "Last Chance Saloons for Natural Kind Realism."

[2] Khalidi, Muhammad Ali. "Mind-dependent kinds." Journal of Social Ontology 2.2 (2016): 223-246.

[3] Kleinberg, Jon. "An impossibility theorem for clustering." Advances in neural information processing systems 15 (2002).

[4] Von Luxburg, Ulrike. "Clustering stability: an overview." (2010).

[5] Von Luxburg, Ulrike, Robert C. Williamson, and Isabelle Guyon. "Clustering: Science or art?." Proceedings of ICML workshop on unsupervised and transfer learning. JMLR Workshop and Conference Proceedings, 2012.

[6] Pollard, David. "Strong consistency of k-means clustering." The Annals of Statistics (1981): 135-140.