On the Category of "Religion": A Taxonomic Analysis of a Large-Scale Database

M. Willis Monroe, Rachel Spicer, Gino Canlas, Travis Chilcott, Stephen Christopher, Megan Daniels, Andrew J. Danielson, Matthew Hamm, Caroline Arbuckle MacLeod, William Noseworthy, Ian Randall, Robyn Faith Walsh, Michael Muthukrishna, Edward Slingerland

Abstract

The Database of Religious History (DRH; religiondatabase.org) is a large-scale digital humanities project dedicated to capturing scholarly perspectives on the history of religious groups across the globe. Analysis of the current state of the data shows a remarkable consistency between a taxonomic tree generated from the entries submitted by our experts and larger assumptions within Religious Studies as they pertain to the similarities and differences between religious groups. Additionally, there is broad agreement between how experts answer questions and the tags they use to categorize their own entries, demonstrating a consistency between top-down and bottom-up approaches to describing religious groups. We see both of these results as affirming a commensurable understanding of the category of "religion" while at the same time demonstrating the value of these types of large-scale quantitative analyses for answering larger questions within the field.

Introduction

The Database of Religious History (DRH; religiondatabase.org), created in 2012, is an online quantitative and qualitative encyclopedia of religious cultural history. It consists of entries organized around particular units of analysis, which currently include "Religious Group," "Religious Place," or "Religious Text,"[1] tagged with a particular date range and map. Each entry consists of expert-entered answers (with added comments and sources) to a long questionnaire consisting of a fixed set of questions, specific to the chosen poll. Each entry is also categorized within the database using tags assigned by the expert who created the entry. The combination of these two forms of data in aggregate then allows us to interrogate both the history of religion globally, as well as the category of religion within the field of religious studies. The present study discusses the result of an exercise that we undertook to explore how entries in the DRH might be classified and related to one another through differentiations in their patterns of answers to specific questions. These early results offer an intriguing proof-of-concept analysis of how such a bottom-up approach—one that takes advantage of the unique quantitative nature and

1 A list of entries and experts can be found here: https://religiondatabase.org/browse/. In the interest of transparency, it should be noted that all experts who publish their entries in the DRH receive an honorarium recognizing their academic input. The project recognizes the ubiquity of uncompensated labor in academia and has been able to offer honoraria to all experts thus far for the completion of entries in the DRH.

digital affordances of the DRH—might provide new perspectives on the categorization of religions in the cultural historical record.

The DRH began with the "Religious Group" poll, and from the very beginning confronted scholarly concerns about both aspects of this term. To begin with, scholars in the field of religious studies have, of course, long wrestled with how to define "religion," or if the term even picks out a stable category of human experience.[2] While the directors and editors at the DRH believe that "religion" and "religious" are useful, radial categories for identifying clusters of beliefs and practices that occur across cultures and time,[3] and serve similar functions for the social networks in which they are found, at the end of the day the DRH adopts the approach that experts should not predetermine the parameters of their object of study. In other words, experts should not be concerned about whether or not the group or place or text for which they are interested in preparing an entry counts as "religious." Instead, they are encouraged to look at the relevant poll questions and decide whether or not answering them, or some subset of them, would allow them to provide a coherent account of their group, place or text. It should be noted that no singular questions in the poll determines inclusion within the database. This approach has the added benefit of allowing us to reflect on the applicability of our ontology as new entries come in that might challenge the category of "Religious Group".

Perhaps the more difficult task is how one might define a group. Conscious of concerns about scholars creating artificial groups by assuming anti-historical cohesion, or exaggerating the degree of homogeneity within a group, by imposing etic labels on collections of practitioners, the project created a loose definition to guide our editors and experts in deciding what constitutes a group: "A community or network of people (locatable in space and time) who share common practices, beliefs, and/or institutions, but who are not necessarily conscious members of an explicitly recognized group. The group can be an emic (indigenous) name or category or an etic (scholarly attributed) one."[4] Experts are encouraged to be geographically and temporarily narrow: i.e. to keep the focus on the specific context. The ability to "tag" with labels (e.g.,

2 The literature on the application of the term to the wide range of phenomena that make up religious studies is immense (Stausberg 2010). Suffice to say, the term "religion" is clearly not a universal category, and the degree to which it is entrenched in and encodes particular world views is particularly relevant to this study (McCutcheon 1997, 148–49; Nongbri 2013). For instance, its etymology from Latin shows the caution needed when applying the term to pre-modern contexts (Saler 1987). Likewise, the way the term can coöpt emic terminology (for instance *śāsanā*) through colonial interfaces in non-western contexts is of considerable interest (Hansen 2017).

3 For more on this definition and its application see (Tappenden 2017). While we invoke the language of community in the database, we are not unaware of the challenges and critique that attend its usage. For more on the parameters of such terms within sociological circles, see Brubaker (2004). For the problematic history of the term "community" politically and in the field of religious studies (and the study of Mediterranean antiquity and early Christianity, in particular) see Stowers (2011) and Walsh (2021).

4 Smith even used the terminology of clustering and statistical analysis in his writing about taxonomic projects, i.e. "We must conceive of a variety of early Judaisms, clustered in varying configurations." (Smith 1982, 14). Lehrich writing on Smith notes in his epilogue that he seemed dubious of mathematical techniques while still using the language of clustering and statistics in his discussion of building taxonomic trees (Lehrich 2021, 152).

Christianity, Daoism), however, also allows one to track ties with other groups and larger identities, whether self-affirmed or imposed etically by scholars.[5]

A good illustration of this tagging practice is an entry in the database on the Meo in North-West India (Kukreja 2020). The expert in this case added the following tags to their entry: "Religious Group", "Indic Religious Traditions", "Islamic Traditions", "Tablighi Jamaat", "Meo Muslim". The last two tags were created by the expert within the tagging tree and further identify this particular entry, but they also allow future experts to create and link closely related entries. This tagging system allows the expert a large degree of agency in how their entry is categorized within the database.

Despite the issues surrounding categorization and terminology, whatever these groups are that some scholars have given the moniker of "religious," they are presumed to share certain similarities and differences that can be tracked by the wider categories reflected in tagging labels such as "Judaism" or "Buddhism." In *Imagining Religions*, J.Z. Smith proposed that a "religion"—that is, a "Religious Group"— could be defined or categorized by its "differential quality" (Smith 1982, 1–18). Borrowing from the biological sciences, he saw this project of classification as consisting of a polythetic taxonomy: a scaffolding of differentiating questions until a unique combination was reached at which point the religious group under analysis was differentiated from its neighbors. Determining the nature of these questions, or differentiating qualities, is key to successful comparison. Smith illustrates how taxonomic classifications that ask a series of binary questions arrive at a final determination of to which categories the item under investigation *must* belong, a monothetic or Linnaean taxonomy. This technique is contrasted with a polythetic system of classification where items share a "set of properties" that define a class. This method of classification resembles Wittgenstein's family resemblances in that aspects of the definition are found across the entire group but no one member has all of the criteria.

Proceeding under this theoretical understanding of the "category" of religion and the demarcation of groups, the DRH has been collecting entries from experts in various fields associated with the study of the history of religion. These entries now make up a continually growing corpus of data on discrete religious groups in the historical record, one that allows for large-scale inspection and comparison across the differential qualities that make up a taxonomic structure.[6] Through a large-scale statistical analysis of "Religious Group" entries we have constructed a taxonomic tree that, by comparing similarities and differences across the entire dataset, organizes the entries in the database into precisely the sort of polythetic taxonomy

5 For more on the idea of "religion" as a radial category and the role of ideas from the cognitive sciences within religious studies see Saler (2010) and Slingerland and Bulbulia (2011).

6 These tags generally follow institutional or disciplinary ideas of classification, roughly summarized by the work of the Harper-Collins Dictionary of Religion (1995); for a thorough explanation of the issues of classification and definition see Smith (1996).

described by Smith.[7] To be clear, the analysis performed remains, in J.Z. Smith's often quoted words, within the "scholar's study," but in doing so reflects the perspectives and traditions of a wide-range of scholars all working with a loose, radial conception of "religion."[8] In essence, this exercise allows us to employ an inductive approach to classifying religious groups throughout time and space, by relying on the work of historians in their own fields to provide the raw points of data from which the classifying algorithm draws its comparisons.

In this analysis we are interested in two questions. The first is how religious groups will be classified and related to one another when a simple algorithm sorts them based upon their pattern of responses to specific questions within the "Religious Group" poll. This taxonomic analysis of "Religious Group" entries[9] currently in the DRH, which ignores the actual tags applied to these groups by experts, gives us an opportunity to test established views concerning the similarity or differences between religious groups in the historical record from the ground up.

The second analysis follows from the first: given the bottom-up tree structure thus constructed, how closely does it match the top-down classifications imposed by the experts who entered the data through the use of "Religious Group" tags? Here we are looking to interrogate the difference (or more accurately "distance") between traditional terms used by our experts to categorize entries and the relative position of those entries on the taxonomic tree. This second analysis speaks to how and if systems of classification for religious groups typically employed in the field of religious studies actually track patterns of responses to specific questions about those groups in the DRH.

Methods

The DRH "Religious Group" poll that serves as the data for this analysis consists of questionnaires constituted by a few hundred questions. Each entry is given a time-range and geographical scope by the expert before answering the questionnaire. The questions were designed to be as neutral as possible, to avoid field specific terminology, and offer detailed definitions of specific terms. Much work went into the creation of the questions, including consultation with experts from across the globe over several years.[10] The questions most

7 The process of filtering the data excludes entries which do not contain a sufficient number of "Yes" or "No" answers to allow for comparison; more detail can be found in the Methods section below.

8 A fourth poll type, "Religious Object," will be implemented soon. The plethora of ontologies allows experts to answer questions about a unit of analysis with which they feel comfortable. For instance, an expert on a given place or text might not wish to postulate an organized "group" behind its creation, and places and texts might be used by multiple distinct groups. The back-end mapping of related questions between these different poll types allows a coherent, but complex, picture of the religious landscape of a particular place and time to emerge organically.

9 See van Ravenzwaaij, Cassey, and Brown (2018) for a more thorough introduction to MCMC method.

10 For an introduction to PCA see McCall (2018, 143-147). For stepwise regression see Roalkvam (2020 pt. 6.1). Limitations of both these confirmation methods is that they require 'complete data', and cannot be used on data with missing values, therefore these analyses were run on robustness test data, which is fully imputed, rather than the conservative analysis which did not impute any data.

commonly allow the categorical answers "Yes/No/Field Doesn't Know/I Don't Know," although some provide different categorical options or require a continuous data answer, such as population numbers or size of largest monument. In any case, the result is a standardized, quantitative data point, which means that the answers are standardized across the entire dataset. Additionally, experts are encouraged to add qualitative comments and citations to each answer to allow for narrative description of the trickier clarifying points necessary to approach an answer. Likewise, multiple answers are allowed for any given question, each of which might offer a slightly different time range or geographical area for the answer, enabling a considerable degree of freedom (or complexity) in how the expert answers a particular question.

As an added piece of data each expert is asked to "tag" their entry with a hierarchical list of terms from a tree of religions not unlike what one might find in a standard encyclopedia or reference guide. The expert can add as many tags as they want, and even suggest their own tags for insertion at different levels within the tree. This system is meant to capture the expert's knowledge about the location of their entry within a traditional form of classification. The current study is based upon a snapshot of DRH data from 6th November 2020, when the DRH encompassed 458 "Religious Group" entries from 234 experts.[11]

In order to perform this analysis a significant amount of processing had to take place beforehand to clean up the data so that the algorithms we employed would be able to cluster entries efficiently. Because the data from each entry generally takes the form of standardized answers to shared questions, they can be visualized as data points representing presence or absence. Categorical DRH questions with "Yes" or "No" (as well as "Field Doesn't Know" and "I Don't Know" answers) can be presented in a matrix (rectangular table of data, Figure 1). Each "Yes" answer is coded as 1 and each "No" is coded as 0. The columns of the matrix represent DRH questions and the rows represent one of the tripartite social divisions of each DRH entry—religious specialists, elites and non-elites (general populace). Depending on the analysis condition, answers of "Field Doesn't Know", "I Don't Know" and unanswered questions are treated as missing values or imputed to yes or no (for full details of analysis conditions see the supplementary material). Questions not employing these standard four answers as options were not included in the analysis. As polls in the DRH are structured hierarchically, more-specific follow-up questions are only asked if the answer to the overarching question is yes. Consequently, it can be inferred that a "No" answer to an overarching question also applies to subsequent follow up questions. For example the question "A supreme high god is present?" is only asked after a yes response to the overarching question "Are supernatural beings present?", as there can only be a supreme high god if supernatural beings are also present. Therefore, in this

11 Because of the promise of this technique, we decided that the data currently available in the DRH was adequate to provide a proof of concept. We intend to continue running the same taxonomic analysis at regular intervals as our coverage increases, and will post these updated taxonomies both on our website and online version of the taxonomy tree linked to below in the Results section: https://rachel-spicer.shinyapps.io/drh_tree/

analysis, if the answer to an overarching parent question is “No”, all follow-up questions are assumed (imputed) to have the same answer.

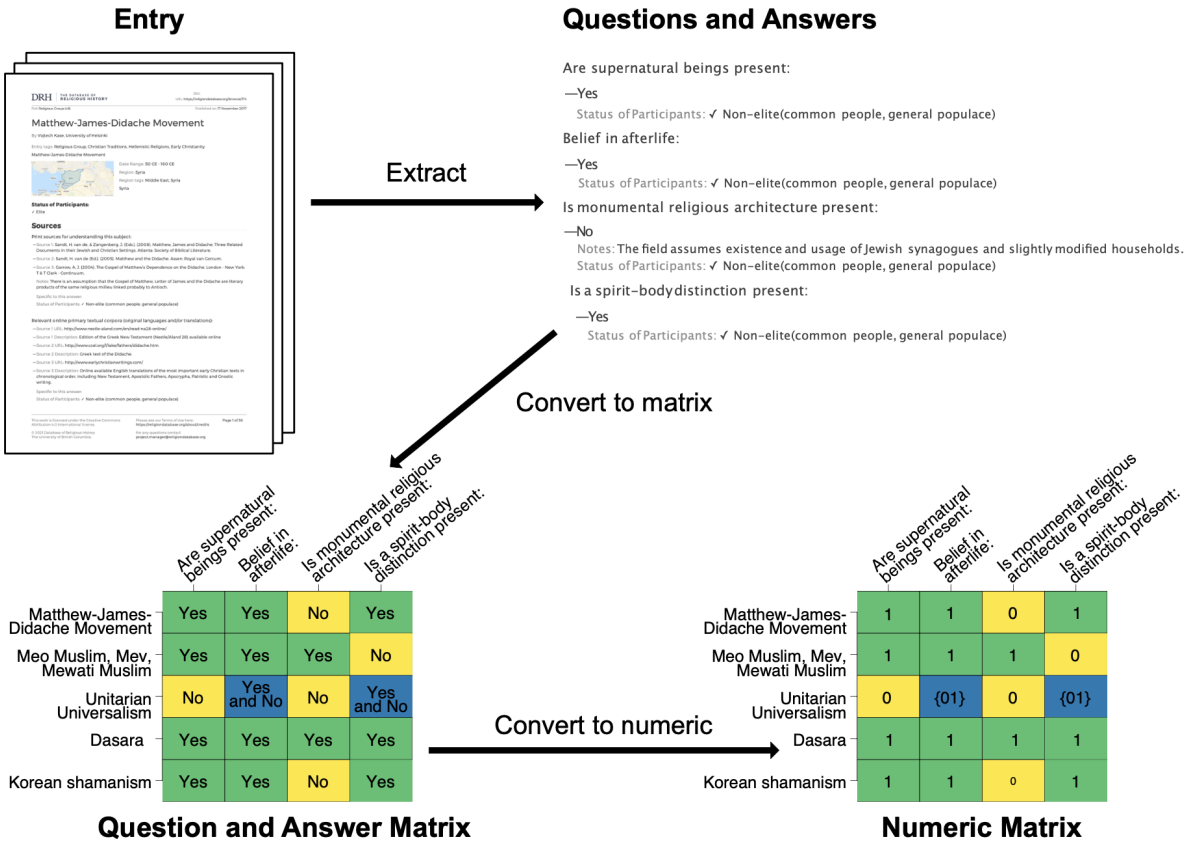

Figure 1. How DRH entries are represented in a matrix for analysis. Question and answer pairs are first extracted for each entry and group of people. These question and answer pairs are then transformed into a matrix, where each row represents an entry and group of people and each column represents a question.

Once the data is processed into this matrix, an analysis program called BEAST2 (Remco Bouckaert, Timothy G Vaughan, and Barido-Sottani, Joëlle 2019) is used to produce taxonomies. These are calculated based on how similar DRH entries are to one another depending upon the pattern of their answers, using a method called Markov Chain Monte Carlo (MCMC).[12] Each row represents an entry (or one social division of an entry), and this row is compared with every

12 These included feedback on polls from participants at the “Workshop on Ritual and the Evolution of Religion and Morality,” Organized by the Cultural Evolution of Religion Research Consortium (CERC, UBC) and the research project “Ritual and the Emergence of Early Christian Religion” (REECR, University of Helsinki), Vancouver, BC, November, 2014; “Religion in the Text and on the Ground: the Convergence of Historiography and Ethnography in Religious Studies” (with Fred Tappenden, McGill), CERC 2nd Plenary Meeting, McGill University, Montreal, QB, May 2015; and “Religion, Ritual, Conflict, and Cooperation: Archaeological and Historical Approaches,” Center for Advanced Studies in the Behavioral Sciences (CASBS), Stanford University, April 29-30, 2016.

other row to figure out which rows are most similar based on the absence or presence of answers in the columns of questions. Thousands of possible trees are generated from this data, each of which differ in how the entries are grouped. The trees differ because each tree is a “probability-based” guess on how best to fit all the entries into a structure that puts similar entries as close together as possible while accounting for uncertainty. The MCMC algorithm samples from these trees, producing a probability distribution of the most likely trees based on the input data and parameters in the model (Figure 2). These trees are then averaged to produce a single consensus tree. This consensus tree can then be examined for groups of similar entries, which we have called clusters.

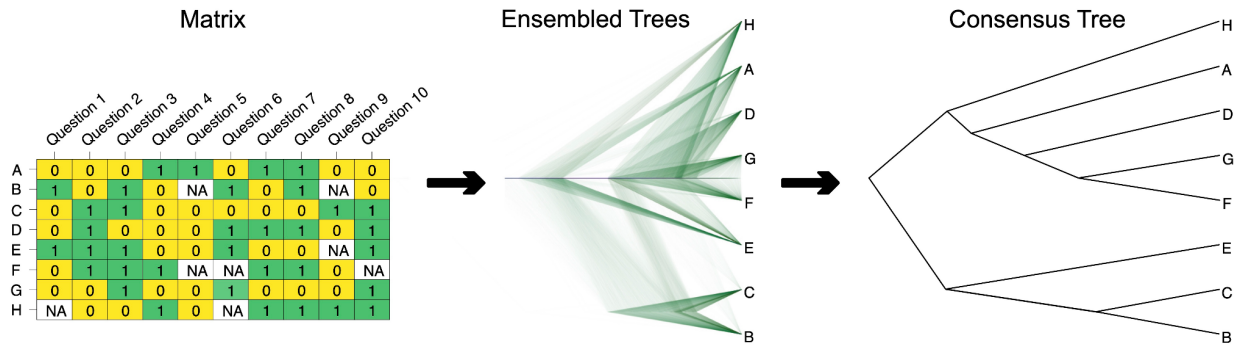

Figure 2. An overview of the creation of a taxonomy using the BEAST2 software. From a matrix of questions and answers, where 1 indicates yes, 0 indicates no and NA indicates missing answers, BEAST2 generates many of the most likely possible taxonomies (thousands to millions). These taxonomies are referred to as "ensembled" trees. A single consensus tree is then produced by averaging these ensembles.

The main advantage of using BEAST2 to produce trees, rather than alternative methods such as hierarchical clustering, is that BEAST2 allows for the inclusion of uncertainty. This means that where there are exceptions to general answers, or conflicting evidence, this can be included in the model. For example the entry Mesopotamian city-state cults of the Early Dynastic periods (Kelley 2019) answers the question “Does membership in this religious group require sacrifice of adults?” as “No.” However, it also notes in a comment that a specific instance of adult sacrifice was identified: “An exception is the discovery of sacrificed attendants (musicians and others) in the Royal Burials at Ur, dating to the Early Dynastic IIIA period. In this case a particular palace or city-state cultic tradition (possibly influenced by unique political events at a moment in history) is apparently behind the practice of human sacrifice associated with royal burial.” BEAST2 allows for the encoding of uncertainty between multiple possible values, in this case between Yes and No, which more accurately reflects the messiness behind coding the historical data for data analysis. The ability to incorporate (accurate) historical uncertainty allows the algorithm to better understand and predict the end result. In this analysis, if a question was answered as both Yes and No for the same entry it was encoded as uncertainty between Yes and

No. This is a crude approach as accuracy of these multiple, disparate answers was not checked prior to analysis.

The branches in the tree represent some underlying differentiation between the entries on each side of the divide. In order to understand what is driving that division we can examine which questions differ between the two groups. These *discriminating questions* are calculated by comparing the percentage of questions with Yes and No answers within each cluster, and identifying the questions which have the highest percentage difference in answers between clusters. This is then confirmed using principal components analysis (PCA) and stepwise regression.[13] PCA is a method of summarizing data based on patterns in the data and is used to identify variables (in this case questions) that are important in creating those patterns. Stepwise regression automatically tries models with different combinations of variables in order to find which variables best explain the differences between clusters.

The end result is a taxonomic tree of "Religious Group" entries showing the most likely relationships between entries across the entire database.[14] This tree structure takes into account uncertainty in the data. The overall patterns present in the tree represent two major levels of analysis. At the macro-scale (i.e. the early branches in the tree), the structure represents major groupings of entries across the dataset, which may or may not reflect current categorical understandings of religious "families." At the most minute level (i.e. the leaves of the tree), the structure represents the algorithm's best guess as to which entries are most like other entries. However, because our data is somewhat sparse,[15] two "Religious Group" entries might be located quite close to each other despite having significant differences. This is because the algorithm is forced to assign a branch to every entry, which means that sometimes dissimilar entries are paired with one another because we lack any otherwise more similar entries that would have been placed between them. Paired entries whose first common branch is far to the left in the tree are not significantly related and are only paired through the necessity of the algorithm.

13 This is a sub-question that only appears if the expert answers "Yes" to the parent question, "Reincarnation in this world," so "reincarnation in this world" is the understood subject (indicated by square brackets).

14 The term Hinduism, as we now use it in religious studies, was undoubtedly shaped by the history of usages by colonialists in the British Raj. However, as Lorenzen (1999, 655) argues, "Hinduism wasn't invented by anyone, Indian or European," although previous scholars have argued that the term Hinduism, while also derived from a Persian geographical descriptor referring to the Indus River Valley, was a colonial construction. Pennington (2005) argues that the agency of Indian authors who argued with, against, and responded to British colonial authors needs to be taken into account. Bayly (2004) however, points toward the ways that both French and Indian authors misinterpreted the history of "Brahmanism" or "Hinduism." Recent scholarship on Hinduism in Bali has shown that there is also intentional construction of Hinduism in Indonesia, resulting from a confluence of Balinese, Dutch colonial, Indian, Japanese, and Indonesian State influences (McDaniel 2010; Picard 2011).

15 These categories include the two-fold division of western/non-western or world religions/primal religions (Tsonis 2013), as well as Bellah's (2011) threefold division into tribal/archaic/axial. Recent attempts to overturn the "world religions" paradigm include the concept of indigenous religion advanced by Cox (2007; 2013; 2017), but recently challenged by Tsonis (2017).

After creating the taxonomic tree of entries, we turned to comparing this tree with a similar structure generated instead from the tags each expert applied to their entries. These tags represent the expert's own intuition concerning what categorical relationships might exist between their entries and other entries in the database. In order to compare the expert-sourced tagging tree classification system of religions to the taxonomy derived from the quantitative answers to DRH polls, the distance between each pair of tags is first calculated. From these distances, the shortest and longest distance between each entry and every other entry is then calculated (Figure 3).

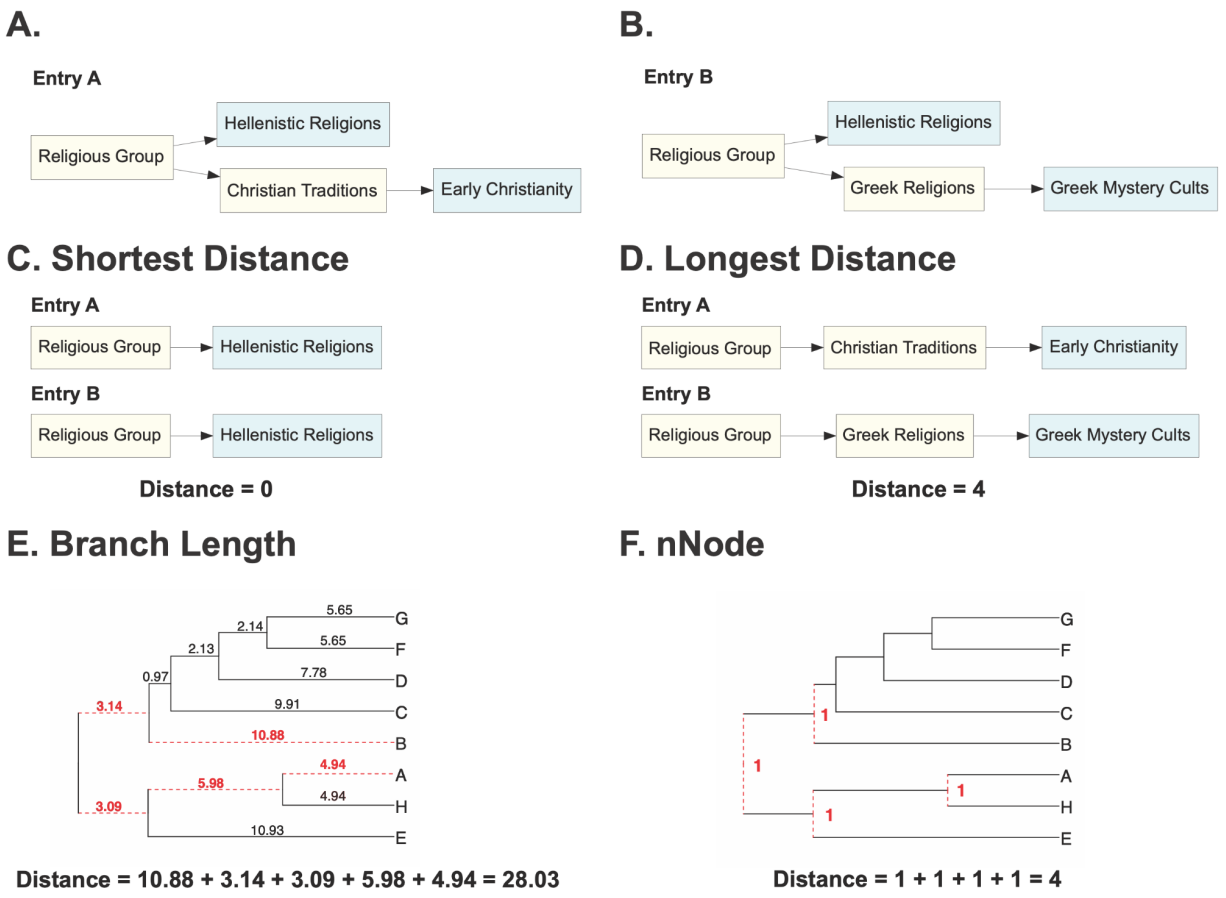

Figure 3. Methods used for calculating the distance between entries using the tagging tree and the taxonomies. Sections A and B show example "Religious Group" tags for two example entries. C and D demonstrate the two methods used for calculating distance between entries using the "Religious Group" tagging tree. C) The shortest distance between entries is calculated using the entries' tags, by finding the pair of tags with the shortest distance between them. In this case the distance is 0 as both entries share the same "Religious Group" tag, Hellenistic Religions. D) The longest distance between entries is calculated by finding the tags that are most disparate in the tagging tree for each pair of entries. E and F show how distance between entries is calculated using the taxonomies. E) The distance between two entries is calculated by summing the total length of branches between them in the tree. F) The distance between two entries is calculated by counting the total number of nodes (nNode) between the entries in the tree.

Two different methods are used for calculating the distance between entries in the taxonomy: branch length and the shortest number of nodes (nNode) between each pair of entries in the taxonomy (Figure 3). All methods (shortest tagging distance, longest tagging distance, branch length and nNode) used for calculating the difference between entries produce distance matrices that reflect the distance between each pair of entries. Kendall correlations—that is, for each method, the similarity/divergence of each pair of entries is ranked, and these rankings are compared between methods, with similar rankings between methods leading to a higher score—are then calculated to find the correlation between each method of calculating distance between entries (branch length, nNode, longest and shortest distance between tags in the tagging tree).

Results

Study 1: Bottom-Up Classification of Religious Groups Based on Pattern of Poll Question Answers

In study one we employed the BEAST2 algorithm to generate a dendrogram of the relationship between entries in the DRH solely through analysis of experts' answers to questions about these religious groups, ignoring both the tags applied by experts to their own entries and the entry names or other classifying information (geography or time range). The overall results are presented in Figure 1.

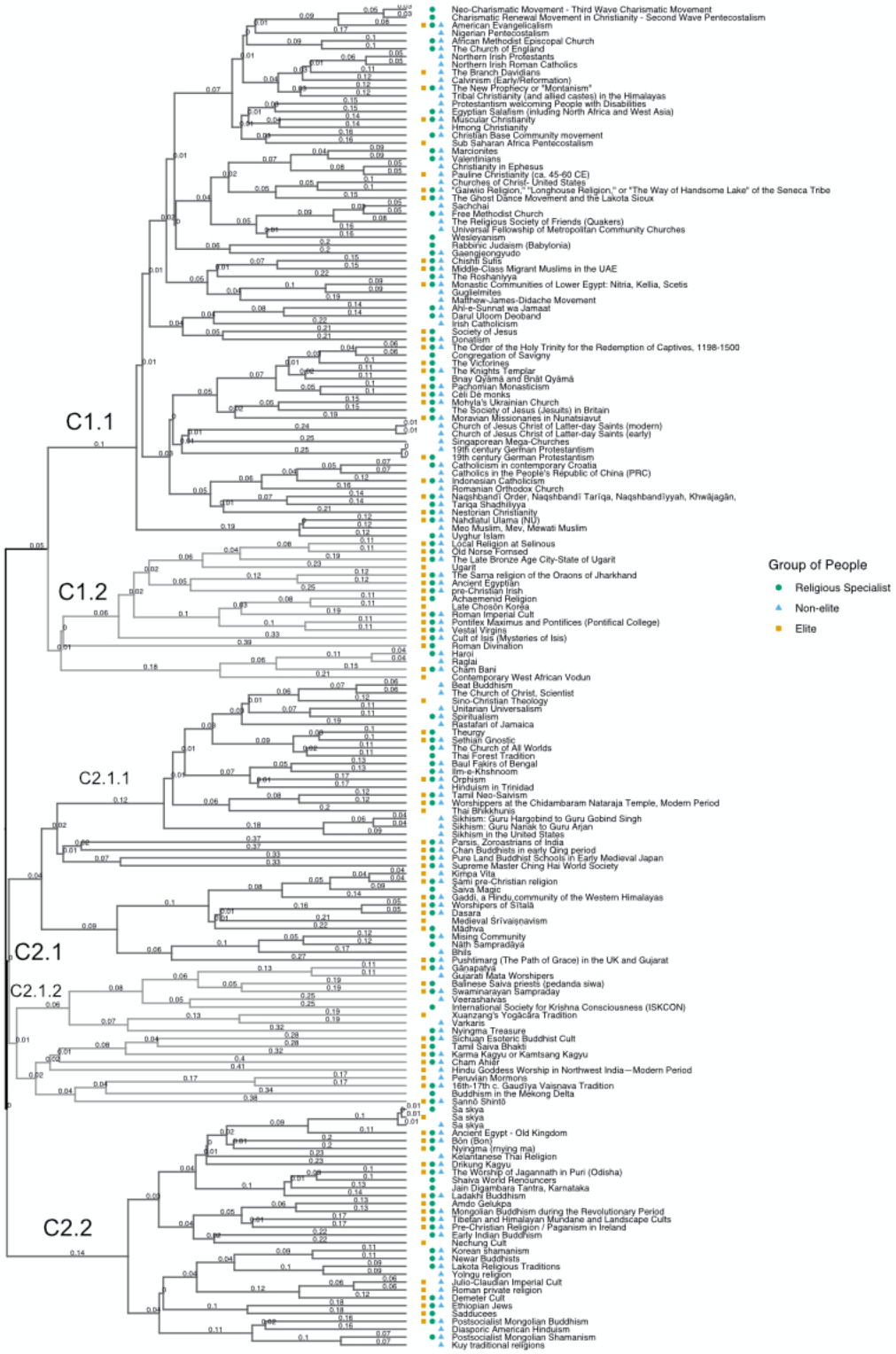

Figure 4. Dendrogram of sample of 171 “Religious Group” entries in the DRH

Figure 4 depicts the overall dendrogram of 171 “Religious Group” entries present in the DRH as of November 6th 2020, including only entries where at least 50% of questions were answered

(questions were simultaneously filtered to include only questions answered by at least 50% of entries). Given its complexity, this figure has also been made available on-line (https://rachel-spicer.shinyapps.io/drh_tree/) to facilitate browsing specific portions of the tree and zooming in on details. In the sections below we will focus on specific sub-sets of the overall tree.

We find that, on the whole, the entries divide into two distinct clusters (termed Cluster 1 and Cluster 2). Each of these is further divided twice in C1.1 and C1.2, and C2.1 and C2.2. Finally, C2.1 is further divided into two smaller clusters (C2.1.1 and C2.1.2). This section initially treats the overall split in C1 and C2, and then the subdivisions of each cluster in turn. This is followed by an in depth discussion of pairs of entries in the tree which either confirm existing scholarly opinions or offer interesting and unexpected results.

Cluster 1 vs. Cluster 2:

The first major division identified by the algorithm splits the dendrogram into two large branches, or "clusters": what we will refer to as Cluster 1 (C1.1 + C1.2) vs. Cluster 2 (C2.1 + C2.2).

Although there are some minor exceptions, the division between C1 and C2 seems to map quite well onto what were traditionally given broad-brush labels: "Western" versus "Eastern" religions. However, our results can more accurately categorize these groups as those that originated geographically in the Mediterranean basin and West Asia (C1) and South, Southeast, and East Asia (C2). Despite the fact that the model does not factor geographical information into its clustering algorithm we can clearly see patterns of colonial and missionary activity. C1 (especially entries in C1.1) trace their origins back to the Mediterranean basin yet can be found around the globe. C1 contains clear examples of Christian missionary impact, such as Nigerian Pentecostalism, Hmong Christianity, and Indonesian Catholicism. Other groups in C1, such as the Chishti Sufis (origins in Afghanistan), the Darul Uloom Deoband (origins in India), Uyghur Islam (Central/East Asia), and Nahdlatul Ulama (Indonesia), all "originate" as South/East/Southeast Asian groups, but the prominence of Meccan-centered discourse among them might help to explain why they are showing up as linked to the Mediterranean/West Asian cluster in our analysis.

Similarly, the placing of certain groups in C2 demonstrates the dispersal of Buddhism along land and sea trade routes. For instance, when it comes to the East Asia entries in Cluster 2, their placement there can be attributed to the influence of South Asian Buddhism, the mechanism of which was the dispersal of Buddhism along land and maritime trade routes (Zürcher 1959, Ch'en 1964, Ch'en 1973). Buddhism brought with it belief in reincarnation—which is, as we see below, a distinguishing question for Cluster 2, and one not found in pre-Buddhist East Asian religions.

East Asian entries that have ended up in Cluster 1 are the product of Christian missionary activity, primarily after the 16th century as a result of maritime trade (Wills 2010).

While we can understand this split as reflecting certain views of broad-brush categorizations, the model can provide more information in the form of which questions drive this split. The primary drivers are questions related to reincarnation, with another contributor being the nature of a high god as unquestionably good (Table 1).

Table 1: Discriminating questions driving the C1 vs. C2 split

	C1		C2	
Question:	Yes	No	Yes	No
Reincarnation in this world:	4.65%	94.19%	81.18%	16.47%
[Reincarnation][16] in a human form:	2.33%	96.51%	69.41%	20.00%
[Reincarnation] in animal/plant form:	1.16%	97.67%	57.65%	28.24%
Reincarnation linked to notion of life-transcending causality (e.g. karma):	2.33%	96.51%	60.00%	27.06%
The supreme high god is unquestionably good:	91.86%	6.98%	36.47%	55.29%

The single most powerful discriminating question between these clusters is "[Reincarnation] in a human form", which 96.51% of the C1 groups answer in the negative and 69.41% of the C2 groups answer in the affirmative. The question "Is there reincarnation in this world?" is a close second, with 94.19% of C1 groups answering in the negative, and 81.18% of C2 groups answering in the affirmative.

Outside of questions covering reincarnation, the nature of a high god as unquestionably good is also powerfully discriminative. While this question is not as starkly one-sided for C2, the signal from C1 (91.86% answered in the affirmative) is very strong. This result is perhaps not terribly surprising given that both Jewish and later Christian scripture and discourse emphasize concepts like "mercy" and the redemptive nature of God's covenant(s) with those who follow the Abrahamic traditions.

Divisions Within C1: Cluster 1.1 vs. Cluster 1.2:

16 For additional discussion and critique of Smith's claim, see: Smith (1982, xi); Smith (2004, 5); Schilbrack (2017, 161-178).

The first major division within C1 is a bifurcation between two clusterings, C1.1 and C1.2 (Figure 5)

The immediate impression given by this split is a distinction between Abrahamic groups versus other "Mediterranean/West Asian" groups. Jewish, Christian, and Muslim groups are all well represented in C1.1 in a variety of historical forms. C1.2, on the other hand, represents a wide range of comparatively more ancient traditions from the Mediterranean basin, with a few outliers (e.g., "Late Chosoň Korea" and "The Sarna religion of the Oraons of Jharkhand" stand out in particular). The presence of Late Chosoň Korea (Shababo, 2019) in C1.2 is likely due to the influence of Confucianism, which has features—such as a belief is a supreme high god and absence of belief in reincarnation—that position it similarly to other ancient Mediterranean religions. Although this entry currently appears to be an outlier, it is indicative of what will likely be a much greater East Asian presence in C1 and likely C1.2 (depending on the scholar) once we have more non-Buddhist East Asian entries available in the database.

It is worth noting the small sample size of ancient Mediterranean/West Asian traditions reflected in the dendrogram. On the basis of many of the discriminating questions that drive the C1.1 and C1.2 split — messianism, proselytization, and the (non)exclusive worship of a high-deity — we may anticipate that this division will become more accentuated as additional ancient Mediterranean exemplars are added to the DRH.

Figure 5. Zoom in on split between C1.1 and 1.2 within C1

The discriminating questions driving the C1.1-C1.2 split are found in Table 2:

Table 2: Discriminating questions driving the C1.1 vs. C1.2 split

	C1.1		C1.2	
Question:	Yes	No	Yes	No
Are messianic beliefs present:	95.59%	2.94%	5.56%	94.44%
Is the messiah's purpose known:	77.94%	8.82%	0.00%	100.00%
Are grave goods present:	10.29%	75.00%	88.89%	0%
Does the religious group actively proselytize and recruit new members:	77.94%	17.65%	5.56%	94.44%
[Are grave goods present] Personal effects:	7.35%	77.94%	77.78%	5.56%
Is it permissible to worship supernatural beings other than the high god:	10.29%	88.24%	83.33%	16.67%
Does the religious group in question possess its own distinct written language:	13.24%	85.29%	72.22%	22.22%
The monarch is seen as a manifestation or emanation of the high god:	2.94%	95.59%	61.11%	33.33%
Are the group's adherents subject to institutionalized punishment enforced by an institution(s) other than the religious group in question:	91.18%	5.88%	33.33%	50.00%
The supreme high god is a sky deity:	32.35%	67.65%	83.33%	16.67%

Unlike the split between C1 and C2, narrowing in on the split between sub-clusters in C1 we find a larger diversity of differentiating questions. The presence of messianic beliefs is the strongest differentiator followed closely by a couple of questions that address burial practices. A few others deal with the way in which the group might bring in new members or interact with authorities. The questions addressing worship of other supernatural beings and the existence of a sky deity address the nature and construction of the supernatural objects of worship, and while their results in this table perhaps align with expectations it is important to note that they are not the strongest signal differentiating entries between C1.1 and C1.2. One more take away would be the way in which these questions might designate religious groups that are often assumed to be de-facto aligned with (or categorized as) state religions versus religions that are potentially in competition with a state-sanctioned belief structure.

We recognize that scholars will have various ideas about the presence of monotheism as a distinguishing factor for religious groups. Some readers may be familiar with the Hindu reformist movements of the Arya Samaj and the Brahmo Samaj, which are Hindu - a religion

commonly described at the introductory level as polytheistic - although these reformist streams of Hinduism are intentionally monotheist. Scholars of Mediterranean religions might also not refer to Abrahamic religions as monotheist, arguing that the term reproduces colonialist and Romantic assumptions.[17] It is worth noting that the significant, but weakly discriminating question, "Is it permissible to worship supernatural beings other than the high god(s)?" is arguably not strictly concerned with monotheism, but rather helps to place a group on a spectrum of henotheism to polytheism.

Divisions Within C2: Cluster 2.1 vs. Cluster 2.2:

17 Critique of the term "monotheism" has a long history in the field. See, for example, Hayman (1991).

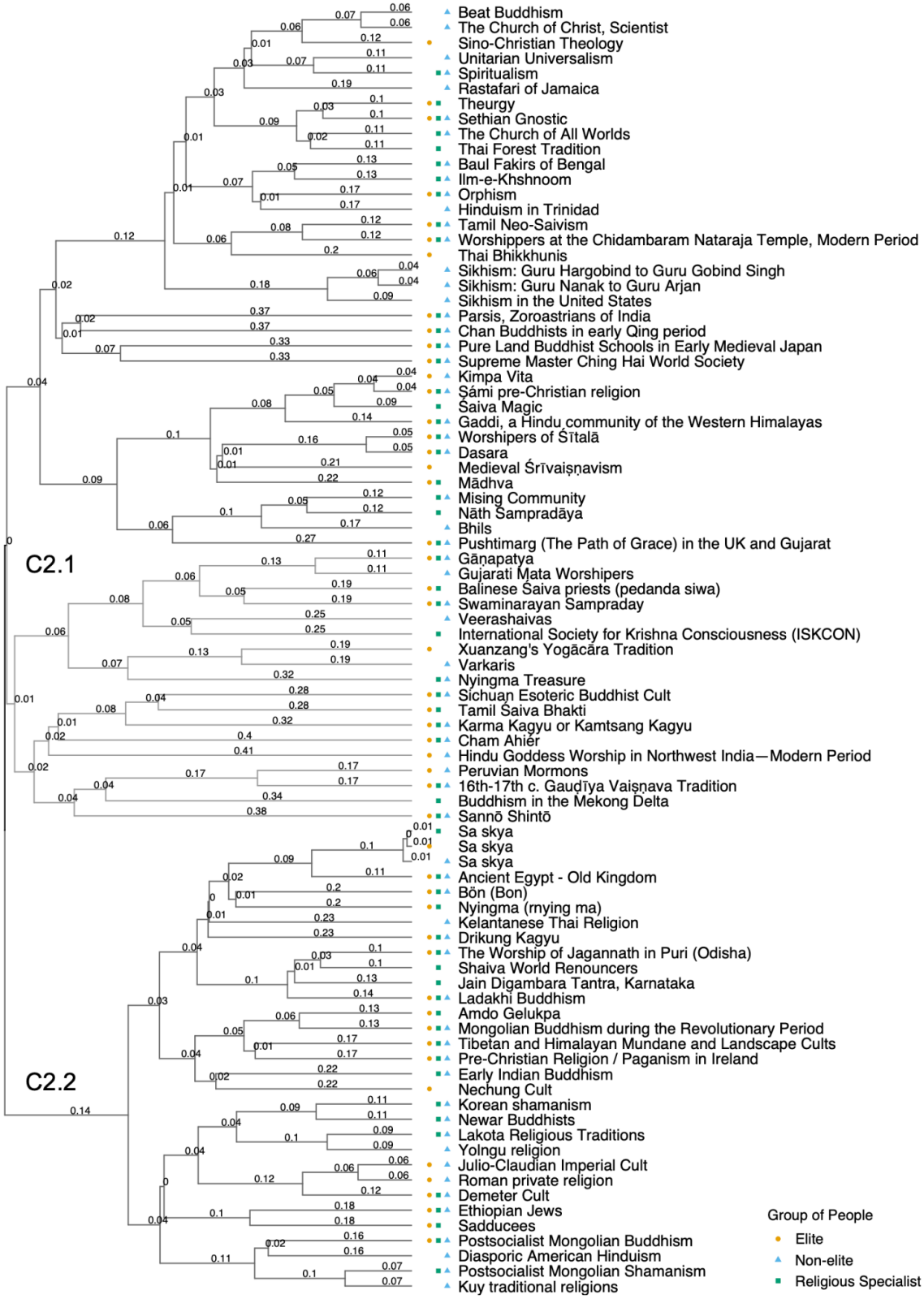

Figure 6. Zoom in on split between C2.1 and 2.2 within C2

Like C1, C2 can also be split into two further clusters (See Figure 6; C2.1 has a further subdivision discussed below). The discriminating questions driving the first split (C2.1 and C2.2) are shown in Table 3.

Table 3: Discriminating questions driving the C2.1 vs. C2.2 split

	C2.1		C2.2	
Question:	Yes	No	Yes	No
The supreme high god has knowledge of this world:	81.48%	7.41%	0.00%	96.77%
The supreme high god communicates with the living:	70.37%	12.96%	0.00%	96.77%
A supreme high god is present:	87.04%	11.11%	6.45%	93.55%
The supreme high god has deliberate causal efficacy in the world:	64.81%	14.81%	3.23%	96.77%
The supreme high god has indirect causal efficacy in the world:	59.26%	14.81%	3.23%	96.77%
Is it permissible to worship supernatural beings other than the high god:	74.07%	14.81%	3.23%	96.77%
The supreme high god exhibits positive emotion:	70.37%	16.67%	3.23%	93.55%
The supreme high god is anthropomorphic:	64.81%	25.93%	0.00%	100.00%
The supreme high god is unquestionably good:	57.41%	29.63%	0.00%	100.00%

Unlike the split between C1.1 and C1.2, the two clusters here are almost entirely differentiated here by conceptions of high gods. There is a strong emphasis in C2.2 on the non-existence of a supreme high god and therefore negative answers to a range of questions about the high god's nature. The fact that 93.55% of entries in C2.2 answer the question of the existence of a supreme high god in the negative suggests a very strong differentiator. However, this sharp division should not be read as attesting to entirely positive answers in C2.1, as we can see in Table 2 many of the affirmative answers in C2.1 are below 80%, suggesting that the negative signal coming from C2.2 plays a stronger role in differentiating than the positive answers from C2.1. This is an important point as it stands in contrast with the split between C1.1 and C1.2 which was driven by more bi-modal distributions of answers between the two clusters.

Further analysis indicates that the majority of groups in C2.1 are South Asian theistic traditions (Hindu, Sikh, Zoroastrian diaspora) that subscribe to variously conceived high gods. Of these,

most are identified as Hindu *bhakti* traditions.[18] The high percentage of positive answers about the existence of a supreme high god among these religious groups is consistent with their respective worldviews. The fact that many of the affirmative answers in C2.1 are below 80% might be explained as a result of the different ways in which these groups conceive of their supreme high god. However, it is also likely that the clustering of traditions that are not primarily traced to South Asia, which make up around a quarter of the groups in C2.1 (clustering mostly towards the top third of C2.1.1), may also help explain this variability. Further analysis is needed to test these hypotheses and tease these variables apart.

Analysis of C2.2 indicates that the majority of entries clustered here, on the other hand, are generally identified as Buddhist or Buddhist-influenced traditions, primarily in South and Southeast Asia. While these traditions often hold theistic beliefs and engage in theistic practices, it is generally a feature of Buddhist thought that there is no supreme deity and that everything, even gods, are subject to impermanence (*anitya/anicca*). The high percentage of negative answers about the existence of a supreme high god among entries identified as Buddhist or being influenced by Buddhism is consistent with this central doctrine.

Taken together, these clusterings suggest that the categories of "Buddhist" and "Hindu" capture meaningful differences between and similarities among the general worldviews of large and diverse groups that are typically identified as either Hindu or Buddhist, the largest differentiator being belief in a supreme high god. Further analyses are needed however. These include analyses aimed at accounting for the variation of groups that are included in these clusters but typically fall outside of the label "Hindu" and "Buddhist;" why some "Buddhist" groups are clustering with a greater majority of "Hindu" traditions in C2.1; and why some "Hindu" groups are clustering with a greater majority of "Buddhist" groups in C2.2. Analyses also need to be done to shed light on why many of the affirmative answers in C2.1 found in Table 2 are below 80%.

Division within C2.1: Division between Cluster 2.1.1 and 2.1.2

Finally, as suggested earlier, there is a smaller division within C2.1 that might suggest a division of this cluster into two further clusters (C2.1.1 and C2.1.2). Here the split between these smaller clusters suggests a heavy emphasis on monumental architecture as defining inclusion in C2.1.2. The differentiating questions can be found in Table 4.

Table 4: Discriminating questions driving the C2.1.1 vs. C2.1.2 split

	C2.1.1		C2.1.2	
Questions:	Yes	No	Yes	No
[Monumental architecture] mass gathering point	13.89%	86.11%	88.89%	0%
[Monumental architecture] temples:	16.67%	83.33%	100%	0%
Are there different types of religious monumental architecture:	80.56%	19.44%	100%	0%

18 For a critique of the term "world religions" see Masuzawa (2005).

[Monumental architecture] altars:	16.67%	83.33%	88.89%	5.56%
[Monumental architecture] tombs:	8.33%	91.67%	66.67%	33.33%
Is monumental religious architecture present:	44.44%	55.56%	100%	0%
Does membership in this religious group require sacrifice of time (e.g., attendance at meetings or services, regular prayer, etc.):	27.78%	69.44%	83.33%	16.67%

The most defining feature of this division is a series of questions that address the existence of monumental architecture. Following from the discussion above, the majority of groups in C2.1 are South Asian theistic traditions and, of these, most of them are identified as Hindu *bhakti* traditions (this is especially true within C2.1.2). Monumental architecture has generally played a central role in South Asian religions over the last 1500 years, especially among those with a devotional orientation, whether Hindu, Buddhist, Sikh, or otherwise. For this reason, it is unclear without further analysis what is driving the divisions between C2.1.1 and C2.1.2. Moreover, why do we find that 80.56% of the groups of C2.1.1 recognize "different types of religious monumental architecture," yet only 44.44% of these same groups recognize the presence of "monumental religious architecture?" They are nearly identical questions, so one would think that the answers to them would match up more closely. To make sense of what is driving these divisions, further analyses of expert answers and the questions posed to them in the poll, which extends beyond the scope of this article, need to be performed. As with the analysis in the previous section, which discusses the division between C2.1 and C2.2, the clustering of traditions that are not primarily traced to South Asia may also help account for the unexpected variability that is driving the division here.

Study 2: Aligning the generated taxonomy with expert tags

For the second study, we compared the generated taxonomy discussed above with a tree structure derived from the group-based tags that experts assigned to their entries. The results are displayed in Figure 7 below.

These two structures show a degree of similarity (Figure 7). This suggests that the top-down conceptual categories employed by our experts match the bottom-up categories derived from a tag-blind analysis of patterns of answers to specific questions.

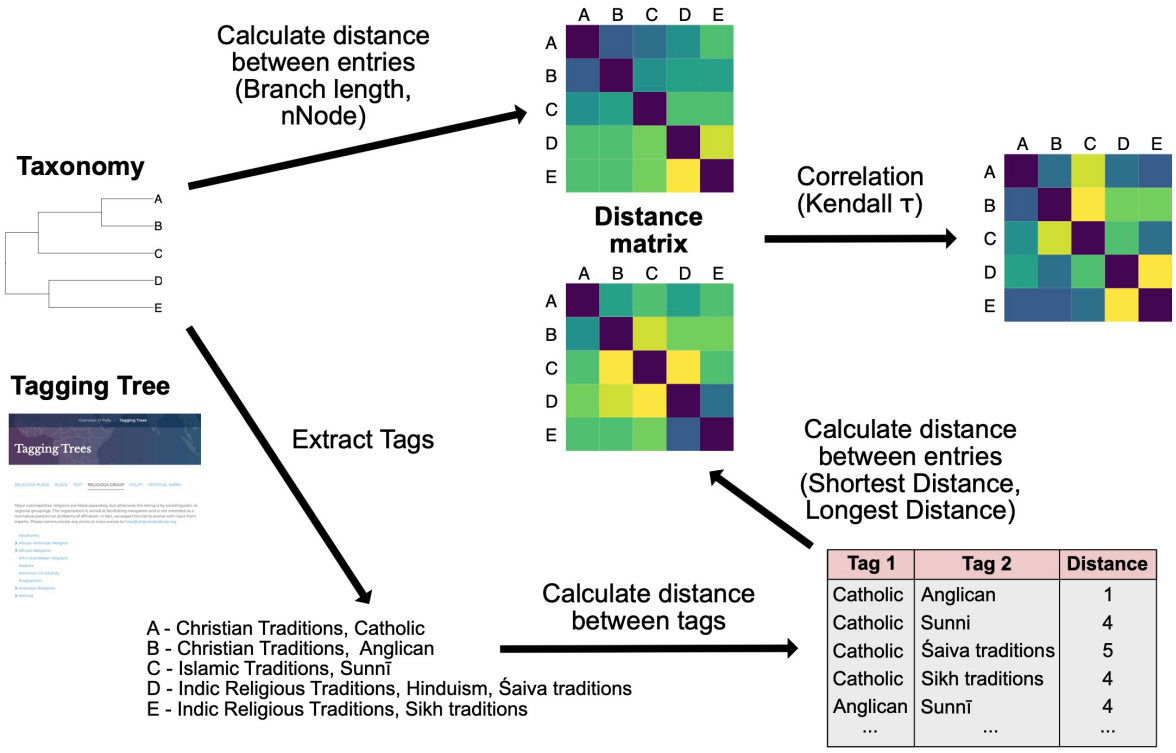

Figure 7. The methods used for comparing the data generated taxonomy with the expert tagging tree. To find the distance between entries using the tagging tree, the tags for each entry are first extracted. The distance between each pair of tags in the tagging tree is then calculated. From these paired distances between tags, the distance between entries is calculated based on their tags. The distance between entries in the taxonomy is calculated using branch length and nNode. All four of these methods produce a pairwise distance matrix which shows the distance between an entry and any other entry. These distance matrices are then compared to areas of agreement and disagreement between methods for calculating distance.

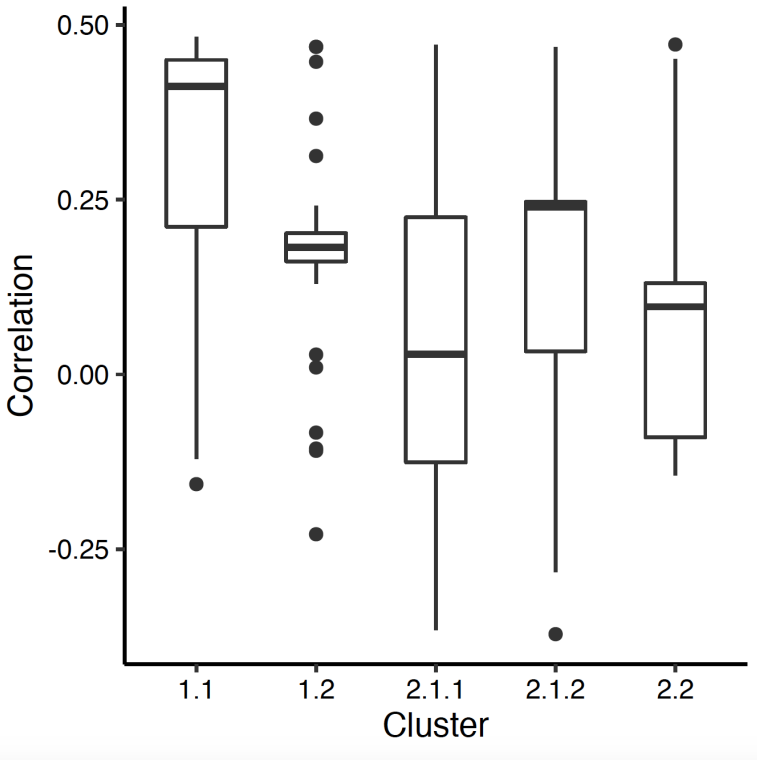

Figure 8: Degree of within-cluster correlation between the generative tree and the tree derived from group-based tags.

Broadly, each cluster finds internal agreement between the structure derived from expert-sourced answers and the tags these same experts applied to their entries (Figure 8). This indicates that across all entries (in all geographical areas and temporal periods) the tags used by experts reflect a certain degree of the underlying organization of the entries as generated by the individual answers themselves. However, Cluster 1 clearly performs the best out of all the clusters. We find this unsurprising as the types of tags applied to entries in Cluster 1 have perhaps the longest history and therefore most stable representations within the field of religious studies (in Western academia). We suggest that this is due to the legacy of church history and theology as precursors to many existing concepts within religious studies scholarship.

For instance, in Cluster 1.1 the entry “Christianity in Ephesus” (Proctor 2020) has the tags: Religious Group, Asia Minor, New Religious Movement, Christian Traditions, Roman Religious Traditions, Anatolian Religions, Early Christianity, and Ancient Christianity in Rome (eight tags). In contrast, in Cluster 1.2, the entry “Ancient Egyptian” (Simpson 2020) has the tags: Religious Group, Egyptian Religions, African Religions (three tags). The capacity for the smaller minute branches within C1.1 to match the underlying tree derived from the answers is driven by the specificity of the tags that are present on entries in that cluster which may be caused by existing disciplinary biases which have created a potent vocabulary of terms and names to describe the diversity of religious groups. On the contrary, Cluster 2 and its components have the

least internal agreement (although still average positive agreement). This, likewise, is not surprising as the terminology used by experts to describe and categorize entries within these clusters is still an active area of scholarship, at least within Western academia.

Discussion

With regard to Study 1, we find that the large splits and major clusters within the generated tree roughly match existing scholarly judgments about the geographical and temporal spread of religious traditions. Study 2 adds nuance to this conclusion by showing that the coherence of these clusters is most pronounced in the branches that have a longer and more entrenched position within the history of the field itself.

General Observations

Our tree seems to replicate standard surveys of "world religions" that offer various high-level taxonomies of religious groups or traditions.[19] However, the cluster division and entry positions offer much more information for fine-grained analysis. Looking in more detail at specific pairings, we find a mix of predictable, puzzling and interesting results when we look to the far right of the dendrogram, and note which religious groups have the closest relationships. As mentioned in the methods section, the way the tree is generated requires that all entries must end up with a physically-paired entry, even if some of these pairs are in fact more distant than we would consider statistically significant for the purposes of direct comparison. To illustrate this point, in Figure 1 above we have overlaid a vertical line on the tree marking entries which fall within a subjective measure of similarity (branches to the right of the line) and those which are merely clustered out of necessity (branches to the left of the line).

Confining our discussion to paired entries with close relationships, we can note that, in certain cases, the similarity between paired entries is obviously the result of geographical, temporal, or doctrinal connections. For instance the similarity between "Neo−Charismatic Movement −Third Wave Charismatic Movement" (Womack 2020a) and "Charismatic Renewal Movement in Christianity −Second Wave Pentecostalism" (Womack 2020b) is attributable to their geographical area (North America), time period (latter half of the 20th. c.), and shared connection to Pentecostal traditions. It is worth noting that these two entries were written by the same expert. Likewise two entries representing groups dating between the 11th and 15th centuries in North-Western Europe, "The Order of the Holy Trinity for the Redemption of

19 To be clear, Nongbri warns against this approach being used to determine whether or not "Capitalism" is a "religion", but the an entry in the database would in fact open up exactly the type of analysis present in our discussion section: what sorts of questions drive a hypothetical "Capitalism" entry towards or away from other entries in the database, and how might those particular vectors speak to underlying questions of how these groups are formed?

Captives, 1198−1500" (Blair 2019) and "Congregation of Savigny" (Doss 2019) show close similarity. Unlike the previous example these two entries were written by different experts, but reflect underlying similarities between these Christian orders.

When a religious group has multiple entries across different time periods we find that their entries still cluster closely, as is the case for "Church of Jesus Christ of Latter−day Saints (modern)" (Pepper 2019a) and "Church of Jesus Christ of Latter−day Saints (early)" (Pepper 2019b). Similarly, we find that when an expert has differentiated the answers in their entry by category of group member "Elite", "Non-elite", and "Religious Specialist," all three sets of answers cluster very closely, as is the case for the entry "Sa skya" (Wojahn 2020). A particularly interesting observation in this case is that the "Elite" and "Religious Specialist" sets of answers cluster more closely than those of the "Non-Elite," which probably reflects a more generic distinction between elite and "folk" understandings and practices.

There are other cases where the results are not entirely surprising but point to an intriguing area for further study or reflection. For instance, "Northern Irish Protestants" (Ward 2019a) and "Northern Irish Catholics" (Ward 2019b) cluster more closely together than the latter does with another entry for a Catholic tradition elsewhere in Europe. While this is a small data point, it suggests that a geographical or local cultural signal is having an impact on how these two entries are paired in the tree rather than an overwhelming doctrinal signal. The close pairing of "Sachchai" (BK 2020) and "Free Methodist Church" (Lane 2020) also offers an intriguing look into how the expression of charismatic traditions within the questions of the database show close similarity.

Finally, we find cases where the placement of entries within clusters and pairing of entries is not entirely expected. This may reflect surprising and important connections or differences of scholarly opinion, but in some cases prompts reconsiderations of both the overall model as well as how our questions might be influencing the placement of entries with the tree. Notably the entry on "Peruvian Mormons" (Palmer 2020) appears in Cluster 2.1 rather than with the other entries on the Church of Latter Day Saints in Cluster 1.1. Here the major differences come down to whether "missions" are considered a form of "pilgrimage" and how the two groups might answer the question of "Reincarnation in this world". The expert for Peruvian Mormons wrote: "Inasmuch as Mormons believe they will be resurrected and go to the Celestial Kingdom and that this world will become the Celestial Kingdom, they believe in reincarnation on this world," whereas the LDS entry states that there is no reincarnation. Digging into the details of the answers and comments here provides us insight into how experts wrestle with these exact questions and reflects divergent scholarly opinions about the beliefs of these groups.

Another unexpected group of divisions occurred between Cluster C1.2 and C2.2, in which entries related to the same ancient Mediterranean and North African cultures were divided. For instance,

a general entry on “Ancient Egyptian” (Simpson 2019) was placed in C1.2, while a more specific entry on “Ancient Egypt - Old Kingdom” (Arbuckle 2020) was placed in C2.2. While we would expect that nuances in the more limited time period that is represented within the Old Kingdom entry should cause it to differ slightly from the more general entry on the whole of Egyptian history, we did not expect the entries to show up in entirely separate clusters. In addition, in this case, the discriminating question does not seem to relate to reincarnation, as both entries agree that there was no “reincarnation in this world”. They disagree, however, regarding the presence of a “supreme high god”, which was another discriminating factor between clusters C1 and C2. The more general entry notes that there was a high god, but provides no additional commentary, whereas the Old Kingdom entry states that there was none, but goes on to explain that there were several significant supernatural beings present during the period in question. While ancient Egyptian religion was generally polytheistic, there were certain times when specific gods were more popular, such as the rise of the god Amun in the New Kingdom or Aten during the Amarna Period, which may account for the discrepancy between the two answers. This helps to show the importance of including entries based on more discrete time periods within the larger religious history of a region in order to arrive at more nuanced and precise conclusions. Nevertheless, as more entries are added to the database, it seems likely that these two entries will end up closer together, given that many of the other answers are similar.

An additional surprising case is the placement of the “Sadducees” (Matson 2020), within Cluster 2.2. With their geographic location in the eastern Mediterranean, and their chronological position at the nexus of the ancient Mediterranean and Abrahamic groups, it might be assumed that they would fall within Cluster 1. Indeed, on the basis of the discriminating questions that drove the C1 and C2 split, the Sadducees measure quite closely with the C1 groups in their lack of belief in reincarnation (resurrection). Likewise, within the breakdown between C1.1 and C1.2, the Sadducees predominantly overlap with C1.1 in terms of messianism, apparent proselytization, and in not allowing the worship of other supernatural beings. Among these discriminating questions they only really differ from C1.1 in their presence of grave goods. Collectively, these positions might assume eventual placement within C1.1 as perhaps befits their historical situation and relation to Judaism.

The Sadducees’ placement within C2, however, is perhaps more the result of particular understandings of their supreme high god, where they stood in contrast to C1 that largely tends to see this deity as “unquestionably good.” Furthermore, within C2, the Sadducee entry’s negative response to nearly all questions related to aspects of the supreme high god places them firmly in line with C2.2 groups, except for one major caveat in that they do in fact possess a supreme high god, whereas the C2.2 groups overwhelmingly do not. Such an apparent similarity between the Sadducees and C2.2 may then be something of a mirage, as so many of the C2.2 discriminating questions are hierarchically predicated on the lack of a supreme high god. This factor might also account for the placement of other Mediterranean entries such as Julio-Claudian Imperial Cult (Bell 2021) in C2.2. Similar to the situation in relation to Ancient Egypt discussed above, it will be interesting to see the effect that additional entries have on this clustering.

At first glance, we found it surprising that Haroi (Quang 2020b), Raglai (Quang 2020a), and Cham Bani (Noseworthy 2020) were together with the Mediterranean/West Asian groupings in Cluster 2, and so closely associated with Contemporary West African Vodun (Atte-oudeyi 2020). Looking more closely at these entries compared to Cluster 2 (where we might assume they would belong) we find the following discriminating questions:

- Assigned at a specific age
- Assigned at birth (membership is default for this society)
- Does the membership in this religious group require sacrifice of property/valuable items
- Is there violent conflict (with groups outside the sample region)
- Does the religious group in question possess its own distinct written language
- The monarch is seen as a manifestation or emanation of the high god
- The supreme high god is a sky deity
- Supernatural punishments are meted out in the afterlife
- Are messianic beliefs present
- Is the messiah's purpose known
- [Are there special treatments for adherents' corpses:] Internment
- Are grave goods present
- Does the religious group in question provide public food storage

The evidence of this cluster points toward a Robert Orsi-like emphasis of understanding "religion as relationships" among humans, their families, their societies, the realm of the supernatural and the human realm(s) (Orsi 2005). For example, we see membership is defined at a specific age, often assigned at birth, and requires the sacrifice of property/valuable items. Members also tend to possess their own distinct written language, while we see that there are examples of violent conflict with groups outside the sample region. There tend to be clear mechanisms for divinities (high god or no) to relate to human society, inclusive of messianic beliefs in at least some of the cases. These could be viewed as Orsi's (2005) "relationships between heaven and earth" although we can rephrase Orsi's language as "realm of the supernatural and human realm(s) to fit the cases in our evidence. Special treatments for adherents' corpses, including internment and grave goods also tend to be present, as does public food storage, suggesting that in these cases relations to earth itself are also present. While our evidence reaffirms some ideas scholars in Religious Studies have about how religions are conceived of, broadly speaking, it also refines the ways that we think about specific cases. Typically, we might regard such groups as Southeast Asian Haroi and Raglai Religion or West African Vodun as "traditionalist." The Cham Bani religious group - nearby Raglai and Haroi communities in Vietnam - has also been described as variously syncretic and traditionalist, although Cham Bani are often interpreted as a form of Islam by outsiders and scholars alike. In both West Africa and Southeast Asia, we find layers of influences including traditional religions, Islam, and colonial era-Christian missions. In both regions, local indigenous groups develop systems that incorporate spirits and saints, have understandings of the cosmos that

include cosmological dualism (emphasizing balance of competing elements), along with traditional healers. Such similarities in the cultural contact zones where we find Vodun—and perhaps also traditional Yoruba religion, although we do not have an entry for it yet in the database—and the Bani, Haroi, and Raglai communities of Southeast Asia could have emerged as a result of similar historical experiences. That traditional religions where membership is assigned at birth, at a specific age, in the context of rituals, and require sacrifices yet do not necessarily retain official political support, as we see in this cluster, speaks to the relational elements of the religious communities in question. This particular result demonstrates that scholars of religious studies have been addressing traditional religions that grow out of comparable historical contexts using similar language, while also demonstrating the ways in which the DRH can lead to thought-provoking comparisons.

Conclusion

One obvious limitation to the current study is that we are not carving up reality at its joints, or even directly analyzing the lived world of religious experience. The analysis above is constructed from experts' answers to questions about units of analysis, i.e. the "Religious Group" poll, that they themselves constructed. However, this expert knowledge is quite fine-grained: experts are answering very specific questions about daily practice, ritual infrastructure, and beliefs. Therefore the standardized poll structure with multiple levels of questions removes the expert from existing narratives and allows them to focus on single answers, producing snippets of atomic data in the form of answers which are more "objective" and therefore comparable, while still remaining highly informed by existing scholarship.

The tree and its component clusters thus constructed from the data is then even more remarkable for its consistency with prior intuitions. The fact that a completely unintelligent algorithm, churning through the answers to very specific questions, produced a dendrogram that roughly mirrors traditional models in religious studies of the major divisions between large religious groups, etc., is significant. It shows that rather than being a completely artificial scholarly category, our entries represent units of analysis that capture real variance in the world, as understood by historians. This is true even though our dataset is incomplete and there are groups and traditions elsewhere in the world with no relation to those on the current tree. Furthermore, the groupings produced by the DRH have significant potential to intervene in recent conversations regarding long-standing (and often highly problematic) categories used to define and categorize religious traditions throughout history.[20]

For instance, if the sorts of questions the DRH asks about religious groups were solely the product of older historical assumptions about religion, we would expect that the major division

20 Note that we are not proposing a phylogenetic tree here as that would imply inheritance and change over time, two forms of analysis that are not part of the current project. The emphasis on taxonomy is to look at similarities and differences between entries across the existing corpus.

(C1 vs C2) in the tree would replicate those perspectives by, for instance, being driven by divisions between groups that have a supreme high god and those that do not, or between those who do or do not worship supernatural beings. This seems not to be the case, because discriminating questions that address these issues that were fundamental to earlier theories of religion are not present at the root of our tree, but rather are scattered through the branches. This suggests that recent skepticism concerning the usefulness of religious categories, such as the assertion that "*the entire set of categories* used to divide human groups must be reconceived" in the study of religion (Tsonis 2017: 59, emphasis original), are exaggerated.

Anomalous results produced by our method show us that certain questions have a highly discriminating power within the model but also allow further analysis to try to fill in entries around them, with the goal of reconstructing a missing cluster or refining our questions to better model the beliefs and practices of the group in question. This is in stark contrast to previous top-down methods where particular properties of groups selected by a theorist were used as a black-box criteria to cluster groups. Our methods are open and inspectable. For instance, Church of Christ Scientist (Prince 2020) and Unitarian Universalism (Applewhite 2020) both appear in cluster 2.1.1. This might, at first, seem strange, but their clustering is likely driven by a negative answer to the question: "Are supernatural beings present?" This suggests that they have a fundamental difference from the entries in C1, but that their similarity to other groups in C2.1.1 is perhaps not as strong, and is then overshadowed by the algorithm's instance of distancing them from C1. More entries would almost certainly produce a context in which these two entries would find a tighter fit, via a hypothetical cluster of entries which fit between C1 and C2.1.1.

The DRH and the taxonomic method introduced in this paper can also be used to explore other interesting questions in religious studies. For instance, the method outlined above could be used to address one of the concluding exercises suggested by Nongbri (2013: 155-156) by creating an entry for, as suggested by Nongbri, "Capitalism." Once such a "Religious Group" entry was created and the questionnaire for it completed, we could interrogate the ways in which the membership practices and beliefs of "Capitalism" were similar or different to other groups that historically have been considered to be "religions".[21]

The biggest limitation of our current study is the preliminary and limited scope of the dataset. While our coverage does span most of the globe, the lack of entries for certain traditions almost certainly drives some of our ambiguous results, as noted above. However, despite this relative paucity of data, this early analysis is still remarkable for its outcome. As our coverage expands and deepens, these sorts of analyses may produce more surprises. We intend to rerun this analysis on a regular basis and keep a current version of the output live on our project website so that the shape of the resulting tree will change as new entries are added.

21 Code and data are available here: https://github.com/religionhistory/religion_taxonomy

Finally, we note that the way in which the DRH collects data diverges from alternative "big data" approaches to history (e.g., Turchin 2018). The flexibility of the question and answer system allows experts to answer multiple times for the same question, change the coverage and scope of their answer, and embed rich media and qualitative comments. This allows the expert to answer questions in a way which best preserves their own intuitions concerning the source material. The ability of the DRH system to encompass multiple units of analysis and allow the periodic updating of polls responds to concerns that, as scholars and scientists studying the phenomenon of religion, we are "we are inextricably stuck with asking just our questions and using just our tools in posing those questions" (McCutcheon 2001: 78). We hope that the analysis presented here shows the potential of the DRH as a tool for scholars of religion, and look forward to this tool being used in the future in ways that we cannot possibly anticipate.

References

Applewhite, C. 2020. "Unitarian Universalism". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/967/#/

Atte-oudeyi, A. 2020. "Contemporary West African Vodun". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/885/#/

Bayly, S. 2004. Imagining 'Greater India': French and Indian Visions of Colonialism in the Indic Mode, *Modern Asian Studies*, 38 (3), 703 - 744.

Bell, T. 2021. "Julio-Claudian Imperial Cult". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/940/#/

Bellah, R. 2011. *Religion in Human Evolution: From the Paleolithic to the Axial Age*. Cambridge: Belknap Press of Harvard University Press.

BK, A. 2020. "Sachchai". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/953/

Blair, H. J. H. 2019. "The Order of the Holy Trinity for the Redemption of Captives, 1198-1500". *Database of Religious History*, Retrieved August 06, 2019, from https://religiondatabase.org/browse/676/#/

Brubaker, R. 2004. *Ethnicity without Groups*. Cambridge, Mass.: Harvard Univ. Press.

Ch'en, K. 1964. *Buddhism in China: A Historical Survey*. Princeton, NJ: Princeton Univ. Press.

———. 1973. *Chinese Transformation of Buddhism*. Princeton, NJ: Princeton Univ. Press.

Cox, J. 2007. *From Primitive to Indigenous: The Academic Study of Indigenous Religions*. Hampshire: Ashgate.

Cox, J. 2013. "Reflecting critically on indigenous religions." In *Critical Reflections on Indigenous Religions*, edited by James Cox, 1–18. Hampshire: Ashgate.

Cox, J. 2017. "Restricted Access Kinship and Location: In Defence of a Narrow Definition of Indigenous Religions." In *Religious Categories and the Construction of the Indigenous*, edited by C. Hartney and D. Tower, 38-57. Leiden: Brill.

Doss, J. 2019. "Congregation of Savigny". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/661/

Hansen, A. 2017. 'Buddhist Communities of Belonging in Early-Twentieth-Century Cambodia'. In *Theravāda Buddhist Encounters with Modernity*, edited by J. Schober and S. Collins, 1st ed., 62–77. New York : Routledge, 2017. | Series: Routledge critical studies: Routledge. https://doi.org/10.4324/9781315637600.

Hayman, P., 1991. "Monotheism-- A Misused Word in Jewish Studies," JJS 42 (1991): 1-15.

Kelley, K. 2019. "Mesopotamian city-state cults of the Early Dynastic periods" *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/479/

Kukreja, R. 2020. 'Meo Muslim, Mev, Mewati Muslim'. *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/645/#/

Lane, S. 2020. "Free Methodist Church". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/879/

Lehrich, C. 2021. *Jonathan Z. Smith on Religion*. Key Thinkers in the Study of Religion. Abingdon, Oxon ; New York, NY: Routledge.

Lorenzen, D. N. 1999. Who Invented Hinduism? *Comparative Studies in Society & History*, 41 (4): 630-659.

Masuzawa, T. 2005. *The Invention of World Religions.* Chicago: The University of Chicago Press.

Matson, J. 2020. "Sadducees". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/848/#/

McCall, G. S. 2018. *Strategies for Quantitative Research: Archaeology by Numbers*. Abingdon, Oxon ; New York, NY: Routledge.

McCutcheon, R. T. 1997. *Manufacturing Religion: The Discourse on Sui Generis Religion and the Politics of Nostalgia*. New York: Oxford University Press.

McCutcheon, R. T. 2001. *Critics Not Caretakers: Redescribing the Public Study of Religion. SUNY Series, Issues in the Study of Religion*. Albany: State University of New York Press.

McCutcheon, R.T. 2010. 'Will Your Cognitive Anchor Hold in the Storms of Culture?' *Journal of the American Academy of Religion* 78 (4): 1182–93.

McDaniel, J. 2010. Agama Hindu Dharma Indonesia as a New Religious Movement: Hinduism Recreated in the Image of Islam, *Nova Religio: The Journal of Alternative and Emergent Religions*, 14(1), 93 - 111.

Nongbri, B. 2013. *Before Religion: A History of a Modern Concept*. New Haven: Yale University Press.

Noseworthy, W. 2020. "Cham Bani". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/476/#/

Orsi, R. 2005. *Between Heaven & Earth: The Religious Worlds People Make and the Scholars Who Study Them*. Princeton, NJ: Princeton University Press.

Palmer, J. 2020? "Peruvian Mormons". Database of Religious History, Retrieved November 20, 2020, from https://religiondatabase.org/browse/891/#/

Pennington, B. K. 2005. *Was Hinduism Invented?: Britns, Indians, and the Colonial Construction of Religion*, New York: Oxford University Press.

Pepper, K. 2019a. "Church of Jesus Christ of Latter-day Saints (early)". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/486/#/

Pepper, K. 2019b. "Church of Jesus Christ of Latter-day Saints (modern)". *Database of Religious History*, Retrieved September 16, 2021, from

https://religiondatabase.org/browse/525/#/
Picard, M. 2011. From Agama Hindu Bali to Agama Hindu and Back: toward a relocalization of Balinese Religion? In M. Picard and R. Madinier (Eds.), *The politics of religion in Indonesia: syncretism, orthodoxy, and religious contention in Java and Bali* (pp. 117 – 141). New York: Routledge.
Prince, A. 2020. "The Church of Christ, Scientist". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/826/
Proctor, T. 2020. "Christianity in Ephesus", *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/948/
Quang, I. 2020a. "Raglai". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/727/#/
Quang, I. 2020b. "Haroi". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/752/#/
Ravenzwaaij, Don van, Pete Cassey, and Scott D. Brown. 2018. 'A Simple Introduction to Markov Chain Monte–Carlo Sampling'. *Psychonomic Bulletin & Review* 25 (1): 143–54. https://doi.org/10.3758/s13423-016-1015-8.
Bouckart R., T. G. Vaughan, and J. Barido-Sottani. 2019. 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis'. *PLOS Computational Biology* 15 (4): e1006650. https://doi.org/10.1371/journal.pcbi.1006650.
Roalkvam, I. 2020. 'Algorithmic Classification and Statistical Modelling of Coastal Settlement Patterns in Mesolithic South-Eastern Norway'. *Journal of Computer Applications in Archaeology* 3 (1): 288–307. https://doi.org/10.5334/jcaa.60.
Saler, B. 1987. 'Religio and the Definition of Religion'. *Cultural Anthropology* 2 (3): 395–99.
———. 2010. 'Theory and Criticism: The Cognitive Science of Religion'. *Method & Theory in the Study of Religion* 22 (4): 330–39. https://doi.org/10.1163/157006810X531111.
Schilbrack, K. "A Realist Social Ontology of Religion," *Religion* 47: 161-78.
Shababo, G. 2019. "Late Chosŏn Korea". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/228/
Simpson, B. 2020. "Ancient Egyptian". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/738/
Slingerland, D., and J. Bulbulia. 2011. 'Introductory Essay: Evolutionary Science and the Study of Religion'. *Religion* 41 (3): 307–28. https://doi.org/10.1080/0048721X.2011.604513.
Smith, J.Z. 1982. *Imagining Religion: From Babylon to Jonestown*. Chicago: University of Chicago.
Smith, J.Z., ed. 1995. *Harper-Collins Dictionary of Religion*. San Francisco: Harper-Collins.
Smith, J.Z.. 1996. 'A Matter of Class: Taxonomies of Religion'. *The Harvard Theological Review* 89 (4): 387–403.
Smith, J. Z. 2004. *Relating Religion: Essays in the Study of Religion*. Chicago: University of Chicago Press.
Stausberg, M. 2010. "Distinctions, Differentiations, Ontology, and Non-humans in Theories of Religion". *Method & Theory in the Study of Religion* 22/4: 354–274. 99
Stowers, S. 2011. 'The Concept of "Community" and the History of Early Christianity'. *Method & Theory in the Study of Religion* 23 (3): 238–56. https://doi.org/10.1163/157006811X608377.
Tappenden, F. S. 2017. 'The Database of Religious History and the Study of Ancient Mediterranean Religiosity'. *Journal of Cognitive Historiography* 3 (1–2): 32–42.

https://doi.org/10.1558/jch.34448.
Tsonis, J. 2013. *Don't Say All Religions are Equal Unless You Really Mean It: John Hick, the Axial Age, and the Academic Study of Religion*. Ph.D. Dissertation. Sydney: Macquarie University.
Tsonis, J. 2017. "Against 'Indigenous Religions': A Problematic Category that Reinforces the World Religions Paradigm" In *Religious Categories and the Construction of the Indigenous*, edited by Christopher Hartney and Daniel Tower, 58-73. Leiden: Brill.
Turchin, P. 2018. "Translating Knowledge about Past Societies into Seshat Data." *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* 9 (1). https://doi.org/10.21237/C7CLIO9139209.
Walsh, R. F. 2021. *The Origins of Early Christian Literature: Contextualizing the New Testament within Greco-Roman Literary Culture*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108883573.
Ward, S. 2019a. "Northern Irish Roman Catholics". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/483/
Ward, S. 2019b. "Northern Irish Protestants". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/484/
Willis, J., ed. 2010. *China and Maritime Europe, 1500-1800: Trade, Settlement, Diplomacy, and Missions*. Cambridge, UK: Cambridge University Press.
Wojahn, D. 2020. "Sa skya". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/841/#/
Womack, N. 2020a. "Charismatic Renewal Movement in Christianity - Second Wave Pentecostalism". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/892/
Womack, N. 2020b. "Neo-Charismatic Movement - Third Wave Charismatic Movement". *Database of Religious History*, Retrieved September 16, 2021, from https://religiondatabase.org/browse/975/#/
Zürcher, E. 1959. *The Buddhist Conquest of China: The Spread and Adaptation of Buddhism in Early Medieval China*. Sinica Leidensia, v. 11. Leiden: Brill.