5-2020

# The Two Types of Society: Computationally Revealing Recurrent Social Formations and Their Evolutionary Trajectories

Lux Miranda
*Utah State University*

# THE TWO TYPES OF SOCIETY: COMPUTATIONALLY REVEALING RECURRENT SOCIAL FORMATIONS AND THEIR EVOLUTIONARY TRAJECTORIES

*by*

**LUX MIRANDA**

*Capstone submitted in partial fulfillment of
the requirements for graduation with*

**UNIVERSITY HONORS**

*with a double-major in*

**COMPUTATIONAL MATHEMATICS & COMPUTER SCIENCE**
*in the Departments of Mathematics & Statistics and Computer Science*

Approved:

_____          _____
Mentor & Departmental Honors Advisor          Committee Member
*Dr. Jacob Freeman*          *Dr. John Edwards*

_____
University Honors Program Director
*Dr. Kristine Miller*

**UTAH STATE UNIVERSITY**
Logan, Utah, USA

Spring 2020

*The two types of society: computationally revealing recurrent social formations and their evolutionary trajectories* by Lux Miranda and Jacob Freeman was published in *PLOS One* in May 2020.

# ABSTRACT

Comparative social science has a long history of attempts to classify societies and cultures in terms of shared characteristics. However, only recently has it become feasible to conduct quantitative analysis of large historical datasets to mathematically approach the study of social complexity and classify shared societal characteristics. Such methods have the potential to identify recurrent social formations in human societies and contribute to social evolutionary theory. However, in order to achieve this potential, repeated studies are needed to assess the robustness of results to changing methods and data sets. Using an improved derivative of the *Seshat: Global History Databank*, we perform a clustering analysis of 271 past societies from sampling points across the globe to study plausible categorizations inherent in the data. Analysis indicates that the best fit to Seshat data is five subclusters existing as part of two clearly delineated superclusters (that is, two broad "types" of society in terms of social-ecological configuration). Our results add weight to the idea that human societies form recurrent social formations by replicating previous studies with different methods and data. Our results also contribute nuance to previously established measures of social complexity, illustrate diverse trajectories of change, and shed further light on the finite bounds of human social diversity.

# ACKNOWLEDGMENTS

I'd like to sincerely thank the following people:

- My mentor, Dr. Jacob Freeman, for his enduring and passionate commitment to the success of his students and to the advancement of science. Without him, none of this would be anything, and the world would be worse off. I feel fantastically fortunate to have access to his wisdom, and I consider him a role model.

- My wife, the amazing and brilliant Whitney DeSpain. I wouldn't be half as productive as I am without her continual supply of encouragement, love, and sanity checks.

- Doctors Terry and David Peak who, through the Peak Summer Research Fellowship, provided the generous funding that allowed me to begin work on this project.

- Dr. Peter Turchin, for inspring me to pursue this realm of research, for his excellent work on the *Seshat* Global History Databank, and for his peer-review of this article.

- Dr. John Edwards, for teaching the fabulously fun *Data Science Incubator* course–which taught me the skills to be able to even do this analysis–and for serving as my second comittee member.

- The anonymous reviewer whose comments allowed me to significantly tighten up the Data & methods section.

- Last but not least, Dr. Kristine Miller and all of the wonderful folks running the Honors program. Y'all have helped make my time at USU an absolute blast. *Sapere aude!*

# CONTENTS

# INTRODUCTION

The emerging model of "recurrent social formations" postulates that only a small number of stable social-ecological configurations exist for human societies [1, 2]. The basic idea is this: one can observe a small set of the same empirical regularities in societies cross-culturally and independent of geography or time. These regularities reflect social and environmental conditions and how these conditions have interacted to settle into an overall configuration.

The model of recurrent social transformations combines qualitative insights from modern bio-economic theory with quantitative insights gained using data reduction and algorithmic techniques from computational science. This empirical approach attempts to avoid the pitfalls of the widely-critiqued idea that societies evolve through a linear series of progressively more complex forms toward an ethnocentrically defined endpoint [1–4]. For example, such computational analyses on datasets encoding information on social formations have proven fruitful in the study of social complexity [1, 5, 6]. As these kinds of datasets and analyses continue to emerge, the robustness of previous results to changes in data and method should be explored. Recurrent social formations identifiable by only one method may be nothing more than a mirage—a kind of confirmation bias similar to the phenomenon of $p$-hacking (wherein a single analytical method is misused to artificially create results that are almost certainly false-positive but construed via metrics such as $p$-values to be "significant"). This article contributes to assessing the robustness of recurrent social formations to changes in computational methods and data sets. That is, we use a multi-dimensional clustering algorithm to explore "clumps" in data on human societies indicative of recurrent social formations, and we then compare our results with those identified by researchers using alternative methods and datasets.

In the remainder of this paper, we provide background on the model of recurrent social formations, use clustering to reveal and explore statistically significant typologies of past societies, and we discuss our findings in the context of the conceptual model of recurrent social formations. Our results indicate that the Seshat dataset is robust in that it produces similar results using different methods of analysis. However, our analysis also illustrates the potential to build upon previously-developed methods. In particular, Turchin and colleagues [5] found that the first principal component of the Seshat dataset (PC1) can serve as a useful time-resolved approximation for a society's overall "social

complexity." Our results demonstrate that the PC1 metric does not necessarily capture nuance in the diversity of how societies are "socially complex." Indeed, when comparing across regions, there is significant overlap in PC1 between societies of different clusters. In some extreme cases, societies in entirely different superclusters can have similar social complexity factor scores. This nuance provided by a combination of PC1 and cluster trajectories may be of importance for certain research questions. Further, we find that plotting societies along axes of social "scale" and "non-scale" is a robust process that reproduces the same results as a 2018 study by Peregrine [1], who similarly conducted a social complexity cluster analysis on a sample of past societies but using different methods and source data. In the end, multiple methods can provide an important check against confirmation bias and open-up a broader range of research questions for comparative social scientists.

# BACKGROUND

Comparative social science has a long history of creating typologies of human societies [7]. In anthropology, 19th century theorists proposed that the diversity of human societies resulted from a non-Darwinian evolutionary process in which societies evolve more "Culture" over time at different rates, and Anglo-American societies were viewed as the apex of this universal shared Culture (e.g., [8]). The clearly ethnocentric typologies that resulted from this theory discredited evolutionary anthropology for nearly a half century. However, in the second third of the 20th century, anthropologists again began to study cultural evolution and created typologies of human societies based on empirical characteristics such as family size, levels of political decision making, and subsistence technology (e.g., [7, 9–12]). Although many of these efforts have been criticised as functionalist and overly simplistic, many argue that empirical "social complexity typologies" have a place in modern scholarship [13, 14]. The idea is that comparative social sciences must transition from typologies that are abstract and simple conceptual aids (e.g., band, tribe, chiefdom) to a rigorous, empirical phylogeny of cultures and the kinds of forms they tend to take—similar to biology's transition from pre-Linnean attempts at organism classification to what is now modern biological taxonomy [13, 15].

The emerging conceptual model of "recurrent social formations" attempts to build upon past attempts to categorize human societies in two ways. First, the model reconsiders the theoretical foundation of empirical typologies by using complex systems theory as a conceptual foundation. That is, the model sees typologies as a consequence of social-ecological interactions rather than discrete "stages" of evolution. For example, complex phenomena (such as predator-prey ecologies, gene regulatory networks, weather systems, etc.) are often mathematically modeled as dynamical systems. These systems are usually discussed with emphasis on the systems' attractors—numerical values towards which the systems tend to evolve. Using the concepts of attractor and of repellor as metaphors, Ullah et. al. [2] use this framework to develop a cluster analysis of the *Standard Cross-Cultural Sample* dataset. They observe four distinct clusters of societies based on features of subsistence, mobility, and demographic variables which, by analogy, may form attractors in the underlying dynamical system governing subsistence behavior.

Similarly, Peregrine conducts an exploratory study that conceptualizes variation in human societies as reflective of adaptive landscapes [1]. Adaptive landscapes describe peaks where the fitness of some combination of traits is high and valleys where fitness is low (e.g., due to the interaction of organisms and their environment). In an analysis of the *Atlas of*

*Cultural Evolution*, Peregrine plots morphological traits of human societies in terms of a "Technology Factor" and a "Scale Factor." The Technology Factor is a composite of variables concerning writing, land transport, social stratification, political integration, technological specialization, and money; the Scale Factor is a composite of variables concerning fixity of residence, agriculture, population density, and urbanization. This study, too, finds two superclusters and several smaller, more refined clusters of societies in this space that, by analogy, may reflect peaks and valleys in the adaptive landscape which further correspond with attractors and repellers in the underlying dynamical system.

Second, researchers have begun to use computational methods to analyse human societies and identify not only clusters of societies with related attributes but also quantify trajectories of change in the social complexity of human societies. Rather than assuming evolutionary stages, this approach attempts to quantify measures of social complexity using an explicit methodology. For instance, Turchin et. al. [5] conduct a principle components analysis of social attributes often considered indicators of social complexity. Consistent with earlier work (e.g., a 1962 scale analysis by Carneiro [16]), they find that these attributes all correlate and that one dimension accounts for a significant amount of variation in social complexity traits (the first principle component of a principle components analysis–PC1 or the social complexity factor). In essence, this PC1 metric creates a reasonable way to quantitatively measure and compare the overall social complexity of societies in different world regions. Although their study only performs a cursory cluster identification of societies that share similar attributes, the study does quantify trajectories of change in social complexity and also suggests that such trajectories share many recurring features cross-culturally.

The above studies suggest that human societies tend to evolve toward a finite set of recurrent social formations; however, this possibility needs further exploration. Recurrent social formations identifiable by only one method or in one dataset may be nothing more than confirmation bias. In the remainder of this article, we ask: Do we find recurrent social formations in the Seshat database that also replicate the trajectories of change in social complexity identified by Turchin and colleagues? Specifically, our study directly builds on the studies above by using a novel clustering algorithm to evaluate how robust the observation of super-clusters and recurrent changes in social complexity are to a change in method and dataset.

First, we use the clustering algorithm in an attempt to replicate Turchin and Colleagues' results. This assesses whether their results are robust to a change in method of data reduction. Second, we attempt to replicate Peregrine's results of two superclusters and

4

several smaller, minor clusters of societies using the Seshat Database. This assesses whether the observation of two superclusters is robust to changing datasets. In the end, our analysis largely replicates previous findings and adds weight to the emerging model that human societies organize into a finite number of social-ecological configurations constrained by ecology and social evolutionary processes.

# DATA & METHODS

We begin with the *Seshat: Global History Databank* [17]. This dataset encompasses information on over 400 polities from 30 sampling locations across the globe. Seshat encodes social complexity features pertaining to social structures, technologies, information systems, economies, subsistence strategies, and other variables for each polity. This database is designed to measure different aspects of societies and evaluate theories of cultural evolution and the evolution of social complexity [5, 6, 18]. Central to evaluating theories for the evolution of social complexity is a baseline description of differences in social complexity across cases and over time [5].

In accordance with prior analyses, our cluster analysis is conducted on a subset of 51 variables from Seshat that the original authors of the database have deemed reliably identifiable from the archaeological and historical records [5, 6]. These variables encode information such as overall population, largest-settlement population, territory, hierarchy, and boolean variables indicating the presence of various aspects of writing systems, texts in circulation, monetary systems, public infrastructure, and government extent.

Our analysis is conducted on a derivative version of Seshat we have constructed and named *Shiny Seshat*. This iteration upon the original database improves upon the imputation methods used in previous analyses, primes the data to be more appropriately suited for temporally-resolved, polity-wise analysis, and patches a number of human-error typographic mistakes in the original dataset.

## IMPUTATION OF MISSING VALUES

Incomplete entries in the dataset are filled in using statistical imputation—a robust method for performing analysis on incomplete data [19, 20] that has been previously used and explicated in the context of Seshat data [5]. Data entries containing missing information may be subject to systemic bias that has led to their incompleteness; thus, statistical imputation can help alleviate bias in data as opposed to simple list-wise deletion of incomplete entries [20]. Particularly in the archaeological and historical sciences, certain societies and cultures can tend to receive more scholarly attention than other societies and cultures, and this can manifest in Seshat in the form of incomplete data. Therefore, imputation is an important process to represent the greatest amount of social variation in our analysis. However, we were unable to completely replicate the imputation method used by Turchin and colleagues [5].

Fortunately, we managed to improve upon Turchin and colleagues' results using a

new open-source imputation tool for Python known as *datawig*, [21] version 0.1.10. This tool utilizes a deep neural network (DNN) that is especially suited for imputing both numeric and non-numeric data. The imputer's parameters (such as number of hidden layers for a feature, hyperparameter optimization options, or feature encoding options) are customizable, but we found datawig's default, largely automatic determination of these parameters to be quite sufficient for our purposes. We specify only to enable hyperparameter optimization and set the number of training epochs to 1,000.

Central to imputation efforts are the replication of nine "Complexity Characteristics" (CCs) encoding information on polity population (*PolPop*), territory (*PolTerr*), largest-settlement population (*CapPop*), hierarchy (*Hier*), government (*Gov*), infrastructure (*Infra*), writing (*Writing*), written texts (*Texts*), and forms of money (*Money*), respectively [5, 22]. These CCs are useful in that they can serve as broad measures of complexity within these domains even in the absence of completely encoded data. For example, a *Writing* score is assigned based on the values of the "Mnemonic devices," "Non-written records," "Script," and "Written records" features, but only one of these features need be encoded for a given polity to be assigned a *Writing* score.

For each complexity characteristic, we create "regression terms" to input into the imputer in order to provide additional prediction-improving information during the imputation training process. These terms are nearly identical to those indicated in Turchin's piece on fitting regression models to Seshat ( [22] pg. 46). In practice, these terms are simply added as additional feature columns. They are:

$$x_{0,i,t} = \sum_{\tau < t} e^{-(t-\tau-100)/100} Y_{i,t-\tau} \tag{1}$$

$$x_{1,i,t} = \sum_{i \neq j} e^{-\delta_{i,j}/1100} Y_{j,t-1} \tag{2}$$

$$x_{2,i,t} = \sum_{i \neq j} w_{i,j} Y_{j,t-1} \tag{3}$$

where $x_{n,i,t}$ is term $n$ for polity $i$ at time $t$ and $Y_{j,t}$ is the value of complexity characteristic $Y$ for polity $j$ at time $t$. Here, $x_{0,i,t}$ helps encode the history of $Y$ by summing all temporally previous values with an exponential discount that grows greater the older the value is relative to $t$. We use an exponential discount of $e^{-(t-\tau-100)/100}$ as series in Shiny Seshat are sampled at the scale of centuries. This produces a factor of $e^0$ for the most recent previous value, $e^{-1}$ for the second most recent value, $e^{-2}$ for the third, etc.

$x_1, i, t$ helps encode spatial diffusion of $Y$ between polities and includes a similar exponential discount factor; $\delta_{i,j}$ is simply the distance between polities $i$ and $j$ in kilometers, and the $1100$ kilometer constant originates from optimization conducted in Turchin's work [22]. Finally, $x_{2,i,t}$ helps encode linguistic distance; we let $w_{i,j} = 1$ if polities $i$ and $j$ share a common language, $w_{i,j} = 0.25$ if they share a common language family, and $w_{i,j} = 0$ otherwise (a value between 1 and 0.25 for common linguistic genus was not included as this is not readily coded in the current public version of Seshat).

In previous analyses [5, 22], the fidelity of imputation prediction has been quantified using the $\rho^2$ metric [23]:

$$\rho^2 = 1 - \frac{\sum_i (Y_i^* - Y_i)^2}{\sum_i (\overline{Y} - Y_i)^2} \tag{4}$$

Where each $Y_i$ are the actual observations in the test set, $\overline{Y}$ is the mean of all $Y_i$, and $Y_i^*$ are the predicted values. Using this function, $\rho^2 = 1$ is a perfect prediction, $\rho^2 = 0$ is a prediction just as good as simply replacing missing values with the mean of known values, and anything less than zero is a worse prediction than simply predicting all values to be the mean of all $Y_i$. However, when working with the Seshat data, we find on some occasions that we encounter the edge-case of $\overline{Y} = Y_i \,\forall\, i$ (when the perfect prediction *is* the mean of the data), leading to a division by zero.

Specifically, this edge-case arises during the imputer's training step when the equation is used to optimize predictive fit on segments of data automated via $k$-fold cross validation. Every so often, the algorithm happens to sample a subset of data from an integer-valued column (e.g., "Administrative levels") where every data point happens to be the same (e.g., a subsample of polity-centuries that all happen to have three administrative levels). Thus, the perfect prediction (three administrative levels) *is* the mean of the data, and a division by zero occurs and crashes the imputation program.

Thus, we modify the equation and create a new function, $\rho_\mu^2$, to include this additional case:

$$\rho_\mu^2 = \begin{cases} \dfrac{\sum_i -|Y_i^* - \overline{Y}|}{\overline{Y}} + 1 & \text{when } \overline{Y} = Y_i \,\forall\, i \\ 1 - \dfrac{\sum_i (Y_i^* - Y_i)^2}{\sum_i (\overline{Y} - Y_i)^2} & \text{otherwise} \end{cases} \tag{5}$$

Should the $\overline{Y} = Y_i \,\forall\, i$ edge-case occur, this places $\rho_\mu^2 = 1$ as a perfect prediction of all values being the mean with $\rho_\mu^2$ increasingly less than one as predicted values diverge

from the mean. In this manner, the usefulness of the metric is maintained in all cases, despite $\rho_\mu^2 = 0$ being semantically meaningless in the edge-case.

Once each regression variable is coded for, we begin training and testing the imputer on data with known values. We estimate the fidelity of each CC prediction using 5-fold cross-validation. Table 1 indicates these results. In practice, we find that, for this case, $\rho_\mu^2 \approx \rho^2 \approx R^2$.

**Table 1.** Fidelity metrics for the prediction of each Complexity Characteristic (CC)

| Complexity Characteristic | $\rho^2 \approx \rho_\mu^2 \approx R^2$ |
|---|---|
| PolPop | 0.86 |
| PolTerr | 0.58 |
| CapPop | 0.80 |
| Hier | 0.89 |
| Gov | 0.92 |
| Infra | 0.88 |
| Writing | 0.88 |
| Texts | 0.95 |
| Money | 0.79 |

During exploratory analysis, we discovered that *not* including spatial and linguistic distance led to better predictions for every CC except for *Texts* and *Money*. We hypothesize this is due to the imputer's DNN being somewhat sensitive to including too much irrelevant information during the training phase. This hints at the possible theoretical consequence that cultural diffusion may, then, be a largely irrelevant factor in the development of many societal characteristics. However, exploring this implication is beyond the scope of this study. Thus, we leave it at that and only include $x_{1,i,t}$ and $x_{2,i,t}$ for the *Money* and *Texts* variables to improve the overall prediction accuracy.

After each CC is imputed, we further impute every missing value in all other columns in Seshat, allowing the imputer to use the already-imputed CCs as input. Additionally, the imputer allows for the possibility of imputing non-numeric values. We utilize this to impute categorical features such as "bureaucracy source of support," "degree of centralization," "linguistic family," etc. We exclude from imputation only features indicating proper names of cultures, places, and rituals. Predictive power for the individual variables is comparable to that of the CCs themselves.

## DATA CLEANING & REORGANIZATION

Beyond imputation of missing values, the most immediately recognizable difference between Seshat and Shiny Seshat is that Shiny Seshat reorganizes information from

individual listings of data points into a matrix-like format where rows are polities during a specific century and columns are features. For brevity, we dub each "polity during a specific century" as a "polity-century," indicating that the row is not only sociopolitically distinguished but temporally distinguished as well, e.g. "The Ottoman Empire 1500CE-1600CE" or "Woodland Cahokia 300BCE-200BCE." This avoids previous ambiguity using the term "polity" without regards to the polity's internal chronology, as a single polity's features will usually take on multiple values throughout its tenure. Thus, time-resolved analyses are done at the scale of *polity-centuries* rather than at the scale of *polities*.

The following changes are also made:

- Seshat encodes binary features on a scale of "present," "inferred present," "inferred absent," and "absent." Mirroring previous work [6, 22], these features are converted to the numeric forms of 1.0, 0.9, 0.1, and 0.0, respectively

- Values encoded as ranges in Seshat are stored as medians in Shiny Seshat. For example, if Seshat indicates that a particular polity has between 6 and 7 administrative levels (ranges such as this typically indicate uncertainty and/or organizational complexity), we encode this as "6.5" administrative levels. For analytic purposes, this simplifies the encoding while still representing the full information for nearly all ranges.

- Polity-centuries spanning multiple Natural Geographic Areas (NGAs) are also more clearly indicated as such in Shiny Seshat in the form of a simple list of NGAs for each polity-century. If one wishes to compare NGAs instead of individual polity-century (such as we do for cluster trajectories in the following sections), it requires only a few simple data transformations to wrangle the dataset into an appropriate form.

- We perform a principal component analysis in the same manner as Turchin et. al. [5] and include the first principal component (PC1) for each polity-century, though this component differs slightly from the one from Turchin et. al.; the PC1 of Shiny Seshat only accounts for 68% of the variance (our code which performs this is included in SI File 2 ). PC2 through PC6 account for 13%, 7%, 6%, 4%, and 2% of the remaining variance, respectively, while PC7 through PC9 all account for less than one percent of variance. Another difference from the original dataset is that the eigenvalue for our PC2 exceeds the standard significance threshold of 1.0; see the Appendix for details on the PCA.

## ALGORITHM

Sparse Subspace Clustering (SSC) is a clustering algorithm capable of efficiently dealing with sparse, highly-dimensional data [24]. The algorithm is resilient to missing, erroneous, and noisy data, and the algorithm is not overly sensitive to data points near subspace intersections. The number of clusters $k$ need not be known prior to clustering; however, a handful of hyperparameters are still required for the algorithm to function. The algorithm is summarized from the work of Elhamifar and Vidal [24] in Algorithm 1.

---

**Algorithm 1** Sparse Subspace Clustering (adapted from Elhamifar and Vidal [24])

---

**Input:** A matrix of data points $Y$

1. Solve the sparse optimization program (equation 8)

2. Normalize the columns of $C$ as $\boldsymbol{c}_i \leftarrow \frac{\boldsymbol{c}_i}{||\boldsymbol{c}_i||_\infty}$

3. Form an adjacency matrix $W = |C| + |C|^T$

4. Perform spectral clustering on $W$

**Output:** Cluster labels for the data points in $Y$

---

We choose to cluster using SSC as Seshat is, indeed, quite sparse in some categories prior to imputation, highly dimensional, and contains lower-dimensional "subspaces" with meaningful interpretations (our typologies in question).

The algorithm operates on the principle of "self-expressiveness" [24]. That is, we start by assuming that every data point $\boldsymbol{y}_i$ can be expressed as a linear combination of every other point (with a total of $n$ points):

$$\boldsymbol{y}_i = c_{i0}\boldsymbol{y}_0 + c_{i1}\boldsymbol{y}_1 + \cdots + 0 \cdot \boldsymbol{y}_i + \cdots + c_{in}\boldsymbol{y}_n \tag{6}$$

In essence, a greater weight $c_{ij}$ indicates that data point $\boldsymbol{y}_j$ belongs in the same cluster as $\boldsymbol{y}_i$, and a weight $c_{ij}$ approaching 0 indicates that $\boldsymbol{y}_j$ is in a different cluster from $\boldsymbol{y}_i$.

In the semantics of polities as data points, this means the algorithm operates on the assumption that no human society has a single feature of social complexity that is entirely unique. That is, a polity's quantitative particularities can always be expressed as some weighted mixture of the aspects of other polities.

Now, we define a matrix $Y = [\boldsymbol{y}_1 \cdots \boldsymbol{y}_n]$ and formulate the equation

$$Y = YC + E + Z, \ \text{diag}(C) = 0 \tag{7}$$

where $C$ is a matrix of weights, $E$ is a matrix to account for error in the dataset, and $Z$ is a matrix to account for noise in the dataset [24]. We wish to find a $C$, $E$, and $Z$ that solve equation 7, thus we frame this as a sparse optimization program

$$\min \ ||C||_1 + \lambda_e ||E||_1 + \frac{\lambda_z}{2} ||Z||_F^2 \tag{8}$$
$$\text{such that } Y = YC + E + Z, \ \text{diag}(C) = 0$$

where $\lambda_e = \alpha_e / \sqrt{n}$, $\lambda_z = \alpha_z / \sqrt{n}$, and $F$ indicates the Frobenius norm [24]. We utilize an Alternating Direction Method of Multipliers (ADMM) optimizer (algorithm also provided by Elhamifar and Vidal [24]) to solve this program.

Lastly, we formulate a matrix $W = |C| + |C|^T$. This matrix serves as an adjacency matrix for a graph—effectively turning linear combination weights between data points into edge weights between nodes. Using a hyperparameterized threshold $\rho$ to determine when a weight is too small to indicate a connection, we are left with a graph containing a discrete number of connected components. These components encode the clustering [24]. In practice, we simply count the number of connected components and feed the graph into a spectral clustering function [25] to create labels for the data points in Y.

Our implementation of this algorithm is available as Python code (SI File 1).

## CLUSTER OPTIMIZATION

We begin by sampling from Shiny Seshat the 51 variables of analysis used in prior works (see Whitehouse et. al. [6] Extended Data Table 5 for a full list of these variables); this is the subset of data that we will perform clustering on. We then normalize each of these features using min/max normalization. SSC involves a high number of matrix multiplications, so this prevents floating-point overflow while still maintaining sufficient information to perform clustering. We also collapse polity-centuries into data points representing the entire base polity by simply taking the mean across all time periods. Further, we found that including too many highly-imputed data points diminished the algorithm's ability to converge on a good clustering. Thus, for the analysis, we have paired down the dataset to only include data points with at least 75% encoding for the fifty-one features, leaving us

with 271 polities.

Our primary means of gauging how good a clustering we have is silhouette analysis [26]. A "silhouette coefficient" between -1 and 1 is calculated for each data point. This coefficient is a measure of how close a data point is to the center of the cluster it has been assigned. A silhouette coefficient close to 1 indicates that a data point is very near the center of its assigned cluster. Conversely, a silhouette coefficient close to -1 indicates that a data point is much closer to a *different* cluster's center than it is its own cluster's center. A silhouette coefficient close to 0 indicates that a data point is near a boundary between clusters roughly equidistant between the centers of its assigned cluster and another cluster.

SSC requires a number of hyperparameters. We use hyperopt, a hyperparameter optimization library [27], to select hyperparameters and find a clustering that minimizes the number of data points with a negative silhouette coefficient. We found, however, that this process does not converge upon the most optimal labelling but rather a labelling that is "fairly close" to optimal. Thus, after performing automatic clustering, we manually optimize the labeling by iterating over data points with a negative silhouette to relocate them to the cluster where they have the highest silhouette. Specially, we simply iterate through each negative-silhouette data point, re-compute is hypothetical silhouette coefficients were it belonging each of the clusters, and re-label it to whichever cluster in which it has the highest silhouette coefficient. Alternatively, using other optimization criteria that do not involve the silhouette coefficient such as MDL or information entropy would perhaps provide better, more streamlined automatic clustering should this process need to be carried out again in future analyses.

Fig 1 provides a silhouette plot for each data point. We indicate the most "archetypal" polities for each cluster in table 2. These are the polities with the highest silhouette coefficient in their given cluster. The average silhouette score across all clusters is 0.18. Although no standard exists for what constitutes a "significant" average silhouette (especially in the social sciences), we can compare this score against a score distribution obtained from performing the same clustering process on similar data sets constructed randomly.

We construct a dataset in the same shape and general form as our clustering input dataset, but instead filled with uniformly random values. It contains the same number of rows and columns corresponding to each entry of actual data. For each column, we note the minimum and maximum values taken on by the actual data, and generate a new uniformly random number within these bounds for each entry. We then perform optimized clustering on this dataset and calculate its silhouette score.
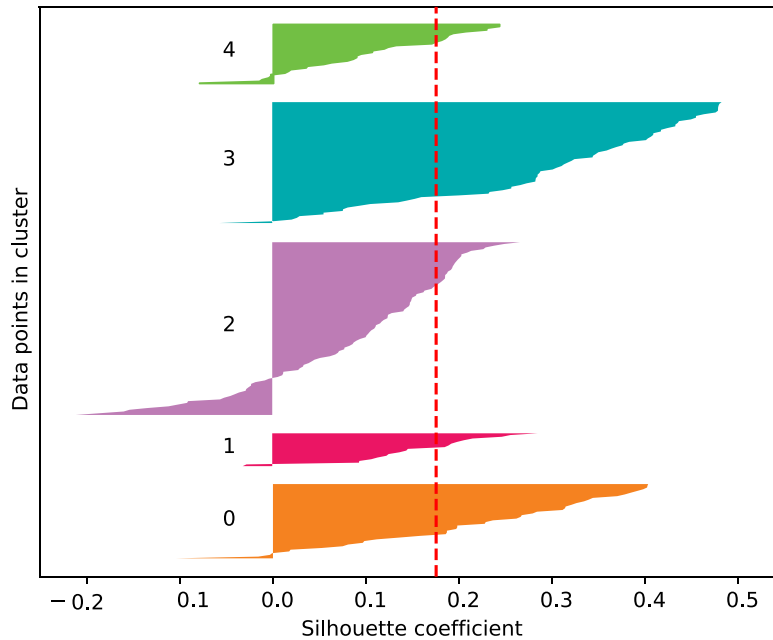
**Fig 1.** Silhouette plot for each data point categorized into clusters. The dashed line indicates the average silhouette of 0.18.

We repeat the above process 540 times and collect the results. We find that the mean silhouette score from clustering each of 540 random data sets is $-0.15$ and that the silhouette score exceeds 0.18 in only 3.9% of cases. In other words, there is a low probability that we are detecting a "false positive" cluster signal in the Shiny Seshat data. In terms of qualitative significance of the clustering, we explicate the uniqueness and significance of individual feature distributions in the following section.

**Table 2.** Archetypal polities. These are the polities which have some of the highest silhouette scores in their respective clusters.

| Cluster # | Archetypal Polity | Approx. Era | Mean PC1 |
|-----------|-------------------|-------------|----------|
| 0 | Woodland Cahokia | 600 BCE - 700 CE | -2.6 |
| 1 | Cahokia Proper | 1100 CE - 1300 CE | -1.5 |
| 2 | Roman Kingdom | 700 BCE - 500 BCE CE | 0.8 |
| 3 | Papal States | 1500 CE - 1600 CE | 2.6 |
| 4 | Ottoman Empire | 1600 CE - 1900 CE | 3.0 |

# RESULTS

To sum up our results, we conceptually replicate previous studies. We find two equally viable clusterings with insignificantly different average silhouette coefficients: A five-cluster solution, and a two-supercluster solution. In this supercluster solution, Clusters 0 and 1 from the five-cluster solution are grouped into a *Supercluster A* and Clusters 3 and 4 form *Supercluster B*, with Cluster 2 from the five-cluster solution split arbitrarily between the two superclusters. The five-cluster solution clearly covaries with Turchin and colleagues' PC1 social complexity factor. Similarly, we replicate Peregrine's two supercluster morphospaces. The observation of recurrent social formations is robust to our new method and change in dataset. We also document that "cluster trajectories" tell regional histories of how societies evolve and move between clusters in the longue durée. These patterns provide a foundation for understanding the causal forces that drive changes between forms of society.

## CLUSTER ANALYSIS VS. THE SOCIAL COMPLEXITY FACTOR

Using the first principal component (PC1) of the Seshat dataset (which both our analysis and that of Turchin et. al. [5] found to encode a majority of the data's variance), we may quantify each polity's complexity in terms of the variables encoded in Seshat. The PC1 metric exemplifies the organization of the five clusters into two superclusters, with Clusters 0 and 1 having generally low PC1 values, Clusters 3 and 4 having generally high PC1, and Cluster 2 having a large variance in PC1 "bridging the gap" between the two superclusters (Fig 2). This trend extends to many of the distributions of the clusters' complexity characteristics; we generally see a "clumping" of superclusters with Cluster 2 spanning a large variance in the middle, or a near-linear scaling of distribution means (SI Fig 1 ).
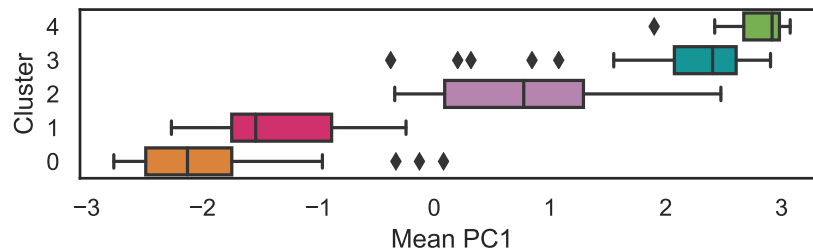


**Fig 2.** Distribution of Social Complexity (PC1) values for clustered polities.

We also analyze the distributions of Complexity Characteristics (CCs) between clusters (SI Fig 1). We see that, for example, social hierarchies within clusters are roughly

normally distributed and that more socially complex clusters tend to have taller hierarchies (Cluster 0 has shorter hierarchies than Cluster 1, Cluster 1 has shorter hierarchies than Cluster 2, etc). Mann-Whitney U tests indicate that nearly all distributions of Complexity Characteristics (CCs) between clusters are unique ($p < 0.05$ in all cases and $p < 0.001$ in nearly all cases). What is more interesting for revealing the nature of these clusters is seeing which distributions are *not* likely to be unique; namely, *CapPop* for Clusters 1 and 2 ($p > 0.26$), *Hier* for Clusters 1 and 2 ($p > 0.11$), *Money* for Clusters 0 and 1 ($p > 0.12$), *Money* for Clusters 2 and 3 ($p > 0.21$), *Writing* for Clusters 0 and 1 ($p > 0.35$), and *Writing* for Clusters 3 and 4 ($p > 0.26$).

These results indicate two things. First, there is clear reason to differentiate subclusters within their superclusters, but the *Money* variable may be measuring something that unifies clusters into superclusters in the first place. Second, it seems the development of a written script is almost synonymous with a society existing in any of Clusters 2, 3, or 4. Whether this relationship is causal or simply highly correlated is yet to be explored.

Table 2 lists some of the most archetypal polities in each of our four clusters. These polities are considered among the best fit data points in their respective clusters, and are thus especially representative of the typical quantitative characteristics of the polities in each cluster. This allows us to discuss cluster characteristics in terms of actual historical examples. The Seshat Knowledge Graph [17] provides qualitative info on these polities:

Exemplary of Cluster 0, Woodland Cahokia is the period prior the rise of the urban city of Cahokia proper. Populations were small and foraging was important for subsistence in this period. Cultures in this period practiced mound-building and pottery, and there is evidence for some high-status burials and crop cultivation in the latter half of the period. Cahokia proper is exemplary of Cluster 1, with the sudden emergence of Cahokia as a immense and densely populated center with a population capable of great feats of cooperation such as mound-building and constructing large wooden palisades.

The Roman Kingdom period stands out as a Cluster 2 society as the small, disjoint villages of the Copper, Bronze, and Iron age (Cluster 0 societies) give way to the beginnings of Rome as a city-state and, following the Kingdom period's conclusion, to the rise of the Roman Republic (a Cluster 3 society). Despite Rome's classification as a Cluster 4 society during the Principate and the Dominate, we see a return to Cluster 3 following the Empire's collapse and a thorough settling-in to this cluster as the Papal States become quantitatively exemplary of Cluster 3 (see Fig 3 to follow this journey).

The Ottoman Empire stands out with the highest silhouette score in Cluster 4. The vast territory of the empire stands in contrast to the relatively small region of the Italian

peninsula encompassed by the Papal States. Although the Ottoman empire had a shorter religious hierarchy in comparison to the Papal States, political hierarchies are matched or greater. Further, the Ottoman Empire was consistently a unitary state throughout its tenure, whereas the Papal States' degree of centralization fluctuated over the centuries from strong singular bureaucracies to loose associations of cities. Yet, both of these Supercluster B societies represent a much greater state of social complexity than societies of Supercluster A.

Further, the clustering allows us to create a kind of analog for the social complexity trajectories of natural geographic areas (NGAs) provided by the PC1 metric (Figs 3,4,5,6). These cluster trajectories illustrate a similar journey through time of cluster membership for the polities occupying an NGA. Temporally, NGAs almost always begin in Cluster 0 and eventually move on to the other clusters. Long-term shifts in cluster membership are usually accompanied by large shifts in PC1 (such as with Fig 3), whereas more rapid fluctuations between clusters usually are accompanied by a relatively stable, if noisy, PC1 (such as with Fig 5). Notably, societies *never* remain in Cluster 2 for the amount of time recorded for the other clusters, suggesting that societies in Cluster 2 are perhaps in an unstable or transitional state. In all cases, time spent in Cluster 2 is typically limited to 200-500 years, whereas time spent in all other clusters can stretch on for millennia (Fig 7a). Further, when accounting for all trajectories, we observe cluster shift frequencies that indicate the vast majority of societies leave Supercluster A without returning, societies tend to pass through Cluster 2 on to Supercluster B, and a majority of societies do not leave Supercluster B once they have entered, and those that do are likely to return (Fig 7b).

Not all geographic areas are very complete in their cluster trajectories due to their data sparseness. The trajectories presented here have been chosen as they are among the most complete trajectories and display dynamics exemplary of their siblings (see the SI File 3 for all generated cluster trajectories).

## THE SOCIAL COMPLEXITY MORPHOSPACE

Building from our cluster analysis, the data replicate the observation of an empirical morphospace of societal scale and technology as observed in a 2018 study by Peregrine [1]. Peregrine's study analyzes a different dataset, the *Atlas of Cultural evolution*, and uses a different cluster-revealing methodology involving Guttman scaling and morphospace analysis. The *Atlas* encodes similar information to Seshat; to help reduce the data's dimensionality, Peregrine utilizes scale and technology factors derived from the Murdock-Provost scale of cultural complexity [28] by Chick [29]. From Peregrine's data,
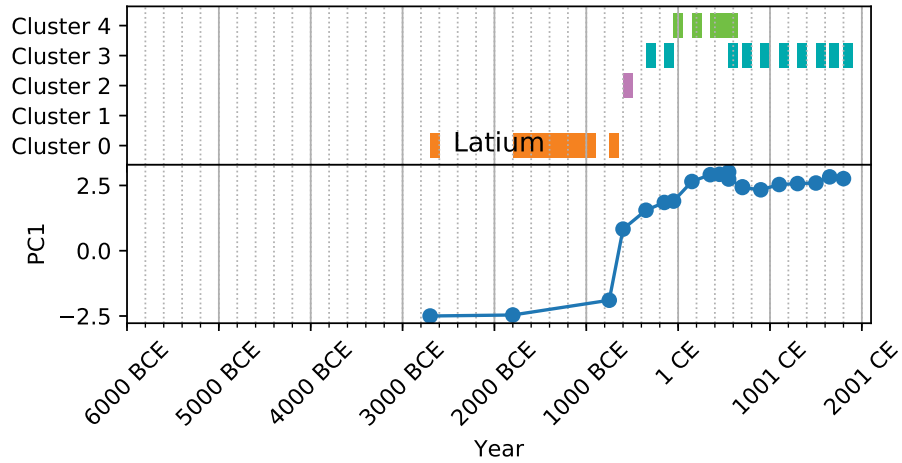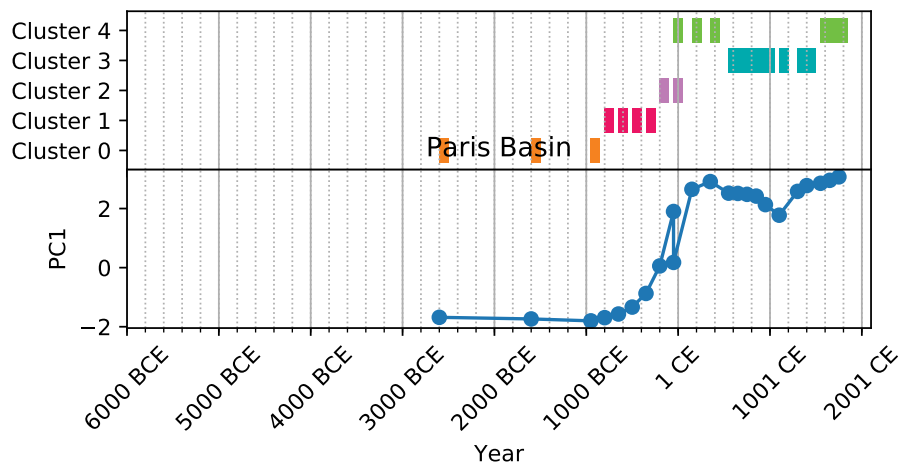
**Fig 3.** Cluster trajectory for Latium



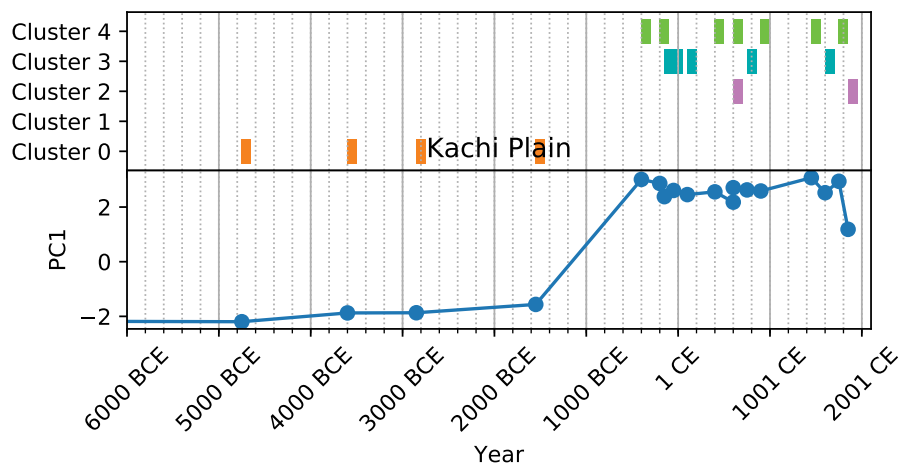**Fig 4.** Cluster trajectory for the Paris Basin



**Fig 5.** Cluster trajectory for Kachi Plain
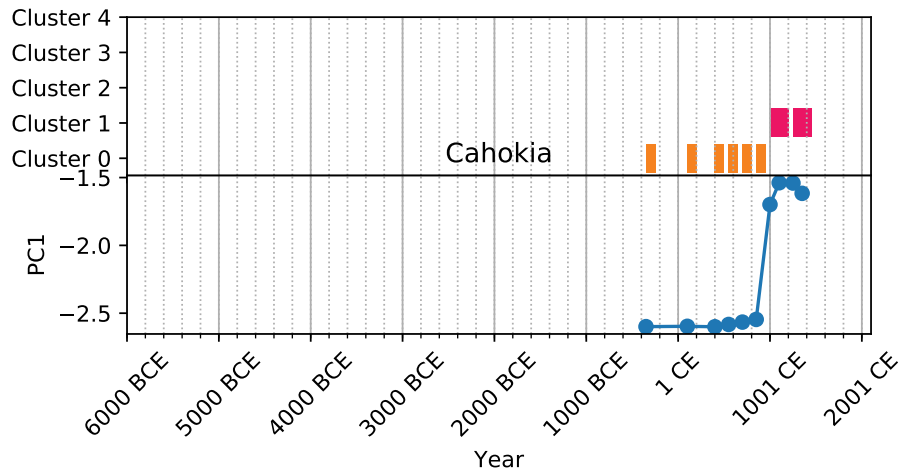
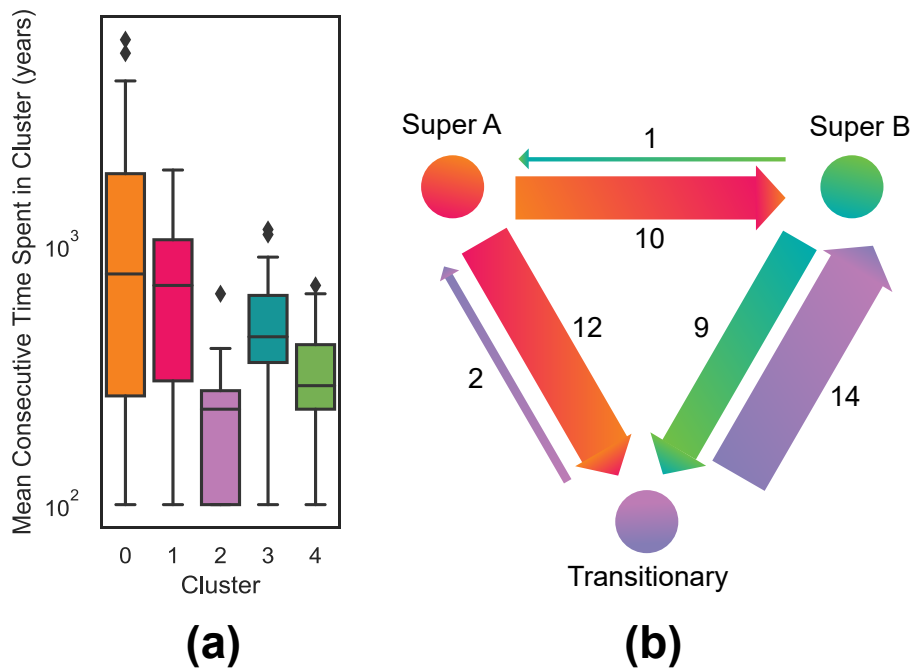**Fig 6.** Cluster trajectory for Cahokia



**(a)**



**(b)**

**Fig 7. (a)** The mean consecutive time societies spend in each cluster. **(b)** Diagram of the total number of observed societal transitions between Supercluster A, the transitionary Cluster 2, and Supercluster B. Arrow width is proportional to each count.

the Technology Factor is a composite of variables concerning writing, land transport, social stratification, political integration, technological specialization, and money; and the Scale Factor is a composite of variables concerning fixity of residence, agriculture, population density, and urbanization.

We present simple, roughly analogous alternatives to Peregrine's Scale and Technology factors derived from Seshat data alternatively named factors of "Scale" and "Non-scale" for greater clarity and rigor. We categorize Seshat's population, territory, and hierarchy features as features of scale and all other features (such as variables concerning infrastructure, writing, and economy) as features of non-scale. In the same manner as we did prior to clustering, we then min/max normalize all features of scale to be between 0 and 1. We then create a scale factor for each polity by simply summing together each polity's normalized scale features. Since all non-scale features are binary in nature (Shiny Seshat encodes them as present with a 1 and absent with a 0, with varying degrees of uncertainty assigned intermediary values), we finally assign polities a non-scale factor that is analogous to the total number of non-scale features that are listed present for each polity. This method is further replicated by creating axes of normalized Scale CCs (*PolPop*, *PolTerr*, *CapPop*, and *Hier*) and Non-scale CCs (*Govt*, *Infra*, *Writing*, *Texts*, and *Money*).

In Fig 8, we plot polities along the axes of scale and non-scale factors. This plot shows almost precisely the same morphospace curve as Peregrine's study [1], and the plot also shows that our own clusters are quite clearly clumped together in this space. Density analysis of the space exemplifies the two superclusters that polities tend to exist in, as well as an additional, smaller smattering between the two clusters representing the centroid of Cluster 2. These results are consistent with the proposition that human societies are well described by recurrent social formations driven by underlying social-ecological interactions.
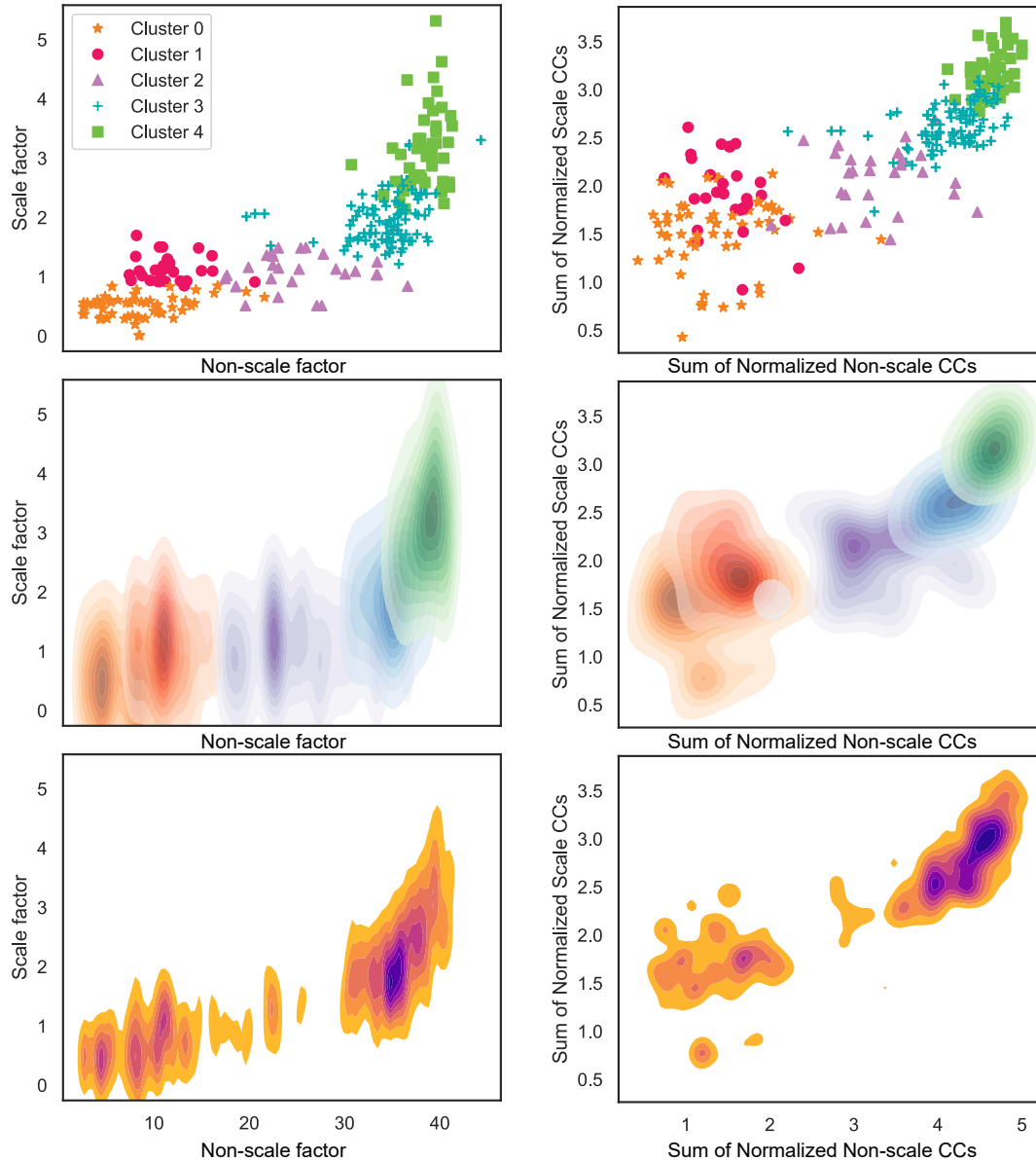
**Fig 8.** The empirical morphospace of cultural complexity: all datapoints by cluster (top plots), per-cluster probability density (middle plots), and overall probability density (bottom plots).

# DISCUSSION & CONCLUSION

In this analysis, we have algorithmically uncovered discrete clusters of societies based on features of government, economy, technology, religion, military, information systems, and population variables provided by the *Seshat: Global History Databank*. Analysis indicates that solutions of two and five clusters are the best fit to Seshat's data, with

lower-complexity Clusters 0 and 1 in the first supercluster, higher-complexity Clusters 3 and 4 in the second supercluster. Cluster 2 is a kind of intermediary, "transient" cluster of societies transferring between the two superclusters. Results hint at the possibility of the development of a written script playing a role in the shift from the first supercluster to the second, although further exploration is needed to determine if this relationship is causal or simply highly correlated.

Our cluster trajectories indicate that, while Seshat and the corresponding Social Complexity (PC1) metric are resilient to differing methodologies, the PC1 metric does not always capture much of the diversity between societies. Indeed, two societies with different technologies, social organizations, and cluster membership may be calculated to have a near-identical PC1. Tying in cluster analysis to study societies in terms of both PC1 and typology may be of use to scholars seeking to utilize a more comprehensive approach to quantification. For example, large changes in PC1 were shown to cross-culturally precede the development of judgemental deities [6], thus contradicting the Moralizing Gods hypothesis. We hypothesize that these large changes in PC1 may also temporally coincide with shifts in cluster membership.

We offer the interpretation of our results within the framework of dynamical systems theory. We hypothesize that there exists an underlying model with attractors in the space of scale and non-scale that manifests in the form of the clusters that we have found. With this interpretation, our analysis is ultimately more exploratory than explanatory. Though, our results offer robustness to the theory behind recurrent social formations. Our methods and dataset differ wholly from those of the study by Peregrine [1], yet we seem to observe the very same phenomena of social complexity morphospaces. In the future, predictive mathematical models should be constructed to describe the attractors that lead to clumping in the morphospaces and shed light onto the dynamics that create these apparent "social steady-states."

# SUPPORTING INFORMATION
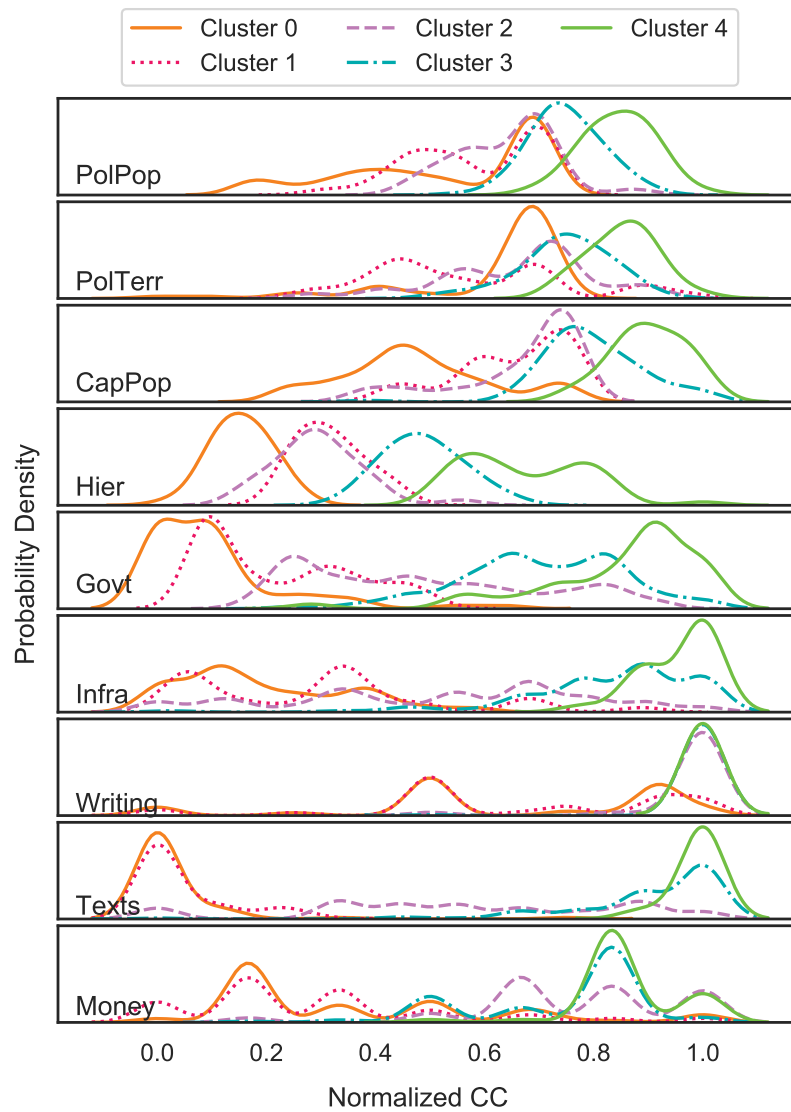


**S1 Fig. 1    Distributions of Complexity Characteristics (CCs) across clusters.** Using a kernel-density estimation.

**S1 File 1.    Clustering and plot generation code.** We include our Python 3 implementation of the clustering algorithm and all analysis and plot-generation code.

**S1 File 2.    Shiny Seshat scrubbing code.** We include our Python 3 program that begins

with the original, untampered Seshat database and performs the entire process of turning it into Shiny Seshat (including all error correction, Complexity Characteristic creation, imputation of missing values, etc.).

**S1 File 3.    A complete collection of cluster trajectories.** We include trajectories for all Natural Geographic Areas (NGAs) for which there is sufficient data (all polities with at least 75% complete encoding for the 51 features of analysis; see the Data and methods section for details).

# REFLECTION

It has been a thrilling adventure to pull together the expertise from people at three entirely different university departments in order to complete this project, for the project has truly lain at the point of intersection between the my three fields of study: computer science, mathematics, and anthropology. In this way, in combination with the research necessitating the application of concepts from so many of my high-level courses (some of which I even completed while concurrently working on the project), I genuinely feel that there isn't a more apt word to describe it than "capstone."

My career goals lie in academic research, so this experience has unsurprisingly been invaluable in preparing me for such a line of work. It has exposed me to the entire research process—from background research, to exploration and discovery, to analysis, to writing, editing, presentation, journal submission, peer review revisions, and publication. It has taught me how to function as a researcher working with other researchers—how to collaborate, communicate, and rely on different peoples' differing skillsets to accomplish a complex task.

Dr. Freeman has been an incredible mentor. He has shown me so many of the ropes and guided me along at every step of the process. I have to particularly thank him for helping "de-engineer" my research mindset. When I began to jump into all this, I did so at the (mostly unreluctant) abandonment of my previous career plans of becoming an engineer. The first time I tried to present results at one of our weekly lab meetings, I dived straight into the mathematical details of the clustering algorithm. Dr. Freeman had to step in and stop me—lest I bore a room of anthropologists to sleep. Since then, my scientific presentation skills have, I hope, greatly improved. He's really helped me convert my brain from thinking only of the "hows" of engineering to also thinking of the "whys" of science, and for this I am very grateful.

The project has, by necessity, given me a good taste of research from each of my fields. It was quite a cool experience to, for the first time, have to parse through an academic paper to take note of the mathematics driving the algorithm I'm trying to use, throw some linear algebra libraries at it to program it all up, and have the darn thing *work*. A part of me didn't actually *believe* that it would work. At first it all just seemed like a bunch of arcane incantations. Yet, there it was, right in front of me, the fruit of my efforts: *clusters!*

To then also use the very methods that I learned in my data science classes, and to *then* even cite some of the very papers that we had to read for Dr. Freeman's *History and Theories of Anthropology* course in order to make some of the points we wanted to make... It has been so satisfying to fit together the puzzle pieces from so many different corners of my studies.

And I've, of course, not only had to fit the pieces together but also critically think about broader topics in order to understand the larger picture. I've had to learn how to think about and interpret both the data and the results of my analysis to understand what it entails for human societies at large. I've had to develop the crucial skill of being able to realize when the data is saying something semantically significant—and when it isn't.

Working so intimately with imputation has also required me to think a lot about how useful it is to fill in missing data with what are essentially fake values. A significant portion of the analysis was dedicated to trying to fill out those values with better estimates, understand how to evaluate those estimates and, once its all said and done, understand the types of things that these estimates can actually tell us. This being the first time I'd done such a thing, I found it to be quite an exercise in data-scientific thinking.

But one of my favorite aspects of the project is how it has brought me a closer to communities I wasn't a part of before. From meeting regularly with our local USU lab, to presenting preliminary results at the Peak fellowship symposium, to ultimately getting involved with the PEOPLE 3,000 research group of archaeologists studying paleoclimate and the "peopling of the Earth," I've met so many brilliant scientists that I've been able to converse with, bounce ideas off of, and garner new understandings from.

Although not directly related to this project, my involvement with the PEOPLE 3,000 group is a direct result of me seeking out Honors capstone mentorship from Dr. Freeman (though some members of the group also ended up providing me with some very good feedback on some early drafts of the manuscript). I first met the bulk of this group through attending a workshop in Vernal, Utah, where I evolved into the token code monkey in the room helping archaeologists construct a new database of global radiocarbon records. Sitting in that conference room, collaborating with so many different researchers from around the world hailing from countries like Norway, Chile, and the UK, was the first time I honestly felt like I was a part of a kind of global community. It was an incredible experience, and I've very much continued to enjoy remotely working with the group since.

But, ultimately, the *raison d'être* of any research, I think, is the impact that it is meant to have on people. While this particular project is largely theoretical without any immediately obvious applications, its true purpose is that it is but one more step into an

emerging field of exciting research that marries quantitative methods with the study of large-scale human behavior. The dream is for this field to one day inform policy—for governments to be able to make decisions rooted in a comprehensive science of social change and to be able to fully predict the long-term impacts that those decisions will have on the social systems they are attempting to affect.

In essence, the hope is that this kind of science will be able to promote human well-being, reduce violent conflict among nations, and allow for ecological and social sustainability across generations. My own research is maybe a long way from being able to do any of that, but I think it's important to not lose sight of that goal. I hope that each hour we work will bring us a little closer to it.

# REFERENCES

1. Peregrine P. Toward a theory of recurrent social formations. In: Sabloff JA, Sabloff PLW, editors. The Emergence of Premodern States: New Perspectives on the Development of Complex Societies. vol. 2. SFI Press; 2018.

2. Ullah IIT, Kuijt I, Freeman J. Toward a theory of punctuated subsistence change. Proceedings of the National Academy of Sciences. 2015;112(31):9579–9584. doi:10.1073/pnas.1503628112.

3. Spencer CS. Testing the Morphogenesist Model of Primary State Formation. In: Prentiss AM, Kuijt I, Chatters JC, editors. Macroevolution in Human Prehistory: Evolutionary Theory and Processual Arcaheology; 2009.

4. Atkinson QD, Whitehouse H. The cultural morphospace of ritual form. Evolution and Human Behavior. 2011;32(1):50–62. doi:10.1016/j.evolhumbehav.2010.09.002.

5. Turchin P, Currie TE, Whitehouse H, François P, Feeney K, Mullins D, et al. Quantitative historical analysis uncovers a single dimension of complexity that structures global variation in human social organization. Proceedings of the National Academy of Sciences. 2018;115(2):E144–E151. doi:10.1073/pnas.1708800115.

6. Whitehouse H, François P, Savage PE, Currie TE, Feeney KC, Cioni E, et al. Complex societies precede moralizing gods throughout world history. Nature. 2019;568(7751):226–229. doi:10.1038/s41586-019-1043-4.

7. Johnson AW, Earle TK. The evolution of human societies: from foraging group to agrarian state. Palo Alto, CA: Stanford University Press; 2000.

8. Tylor EB. Primitive Culture. vol. 1; 1871.

9. Fried MH. The evolution of political society: An essay in political anthropology. New York: Random House; 1967.

10. Sahlins M, Service ER. Evolution and culture. Ann Arbor, MI: University of Michigan Press; 1960.

11. Steward, Julian Haynes. Theory of culture change: The methodology of multilinear evolution. Urbana, IL: University of Illinois Press; 1972

12. White, Leslie A. The evolution of culture: the development of civilization to the fall of Rome. Routledge; 2016

13. Currie TE, Mace R. Mode and tempo in the evolution of socio-political organization: reconciling 'Darwinian' and 'Spencerian' evolutionary approaches in anthropology. Philosophical Transactions of the Royal Society B: Biological Sciences. 2011;366(1567):1108–1117. doi:10.1098/rstb.2010.0318.

14. Peregrine P. Power and legitimation: Political strategies, typology, and cultural evolution. The comparative archaeology of complex societies. 2012; p. 165–191.

15. Shennan S. Genes, Memes, and Human History: Darwinian Archaeology and Cultural Evolution. 1st ed. Thames & Hudson; 2003.

16. Carneiro RL. Scale analysis as an instrument for the study of cultural evolution. Southwestern Journal of Anthropology. 1962;18(2):149–169.

17. Turchin P, Brennan R, Currie TE, Feeney KC, François P, Hoyer D, et al. Seshat: The Global History Databank. 2015; p. 32.

18. Mullins DA, Hoyer D, Collins C, Currie T, Feeney K, François P, et al. A Systematic Assessment of "Axial Age" Proposals Using Global Comparative Historical Evidence. American Sociological Review. 2018; p. 31.

19. Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons; 2004.

20. Yuan YC. Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0). SAS Institute, Rockville, MD. 2000; p. 13.

21. Biessmann F, Salinas D, Schelter S, Schmidt P, Lange D. "Deep" Learning for Missing Value Imputationin Tables with Non-Numerical Data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management - CIKM '18. Torino, Italy: ACM Press; 2018. p. 2017–2025. Available from:
`http://dl.acm.org/citation.cfm?doid=3269206.3272005`.

22. Turchin P. Fitting Dynamic Regression Models to Seshat Data. Cliodynamics: The Journal of Quantitative History and Cultural Evolution. 2018;9(1). doi:10.21237/C7CLIO9137696.

23. Turchin P. Complex Population Dynamics: A Theoretical/Empirical Synthesis. Monographs in Population Biology. Princeton University Press; 2000.

24. Elhamifar E, Vidal R. Sparse Subspace Clustering: Algorithm, Theory, and Applications. arXiv:12031005 [cs, math, stat]. 2012;.

25. von Luxburg U. A Tutorial on Spectral Clustering. arXiv:07110189 [cs]. 2007;.

26. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.

27. Bergstra J, Yamins D, Cox DD. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. 2013; p. 9.

28. Murdock GP, Provost C. Measurement of Cultural Complexity. Ethnology. 1973;12(4):379. doi:10.2307/3773367.

29. Chick G. Cultural Complexity: The Concept and Its Measurement. Cross-Cultural Research. 1997;31(4):275–307. doi:10.1177/106939719703100401.

# APPENDIX

In the original principal component analysis (PCA) of Seshat conducted by Turchin and colleagues [5], the first principal component (PC1) was found to be the only significant PC as it was the only one with an eigenvalue above the standard threshold of 1.0. Our replicated PCA using Shiny Seshat, however, provides us with a significant PC1 but also a PC2 eigenvalue of 1.42 — just above the significance threshold (Table 3).

The loadings across PC1 appears to also be less consistent than the prior analysis (Table 4). This may be an artifact of differences in imputation procedure. Loadings for PC2 indicate that it is primarily encoding the *Money* complexity characteristic (CC) and secondarily encoding the *Writing* CC (Table 5).

There appears to be no significant difference in PC2 between most polity-centuries regardless of supercluster membership (SI Fig. 2). Interestingly, however, the greatest variation in PC2 is seen in Cluster 2, the "transitionary" cluster. Whether this is perhaps a clue about the process that causes shifts in supercluster membership, or whether it bears any semantic meaning at all, is yet to be determined; further research and analysis is required.

**Table 3.** PCA eigenvalues

| PC | Eigenvalue |
|----|------------|
| 1  | 7.5604     |
| 2  | 1.4152     |
| 3  | 0.7268     |
| 4  | 0.6183     |
| 5  | 0.4121     |
| 6  | 0.2549     |
| 7  | 0.0605     |
| 8  | 0.0227     |
| 9  | 0.0148     |

**Table 4.** Loadings for PC1

| Rank # | CC | Loading |
| --- | --- | --- |
| 1 | Hier | 0.494 |
| 2 | Money | 0.459 |
| 3 | PolTerr | 0.400 |
| 4 | PolPop | 0.392 |
| 5 | CapPop | 0.326 |
| 6 | Writing | 0.297 |
| 7 | Texts | 0.137 |
| 8 | Govt | 0.095 |
| 9 | Infra | 0.089 |

**Table 5.** Loadings for PC2

| Rank # | CC | Loading |
| --- | --- | --- |
| 1 | Money | 0.831 |
| 2 | Writing | 0.115 |
| 3 | Texts | 0.034 |
| 4 | Infra | 0.032 |
| 5 | Govt | 0.019 |
| 6 | Hier | -0.231 |
| 7 | PolTerr | -0.259 |
| 8 | PolPop | -0.259 |
| 9 | CapPop | -0.323 |
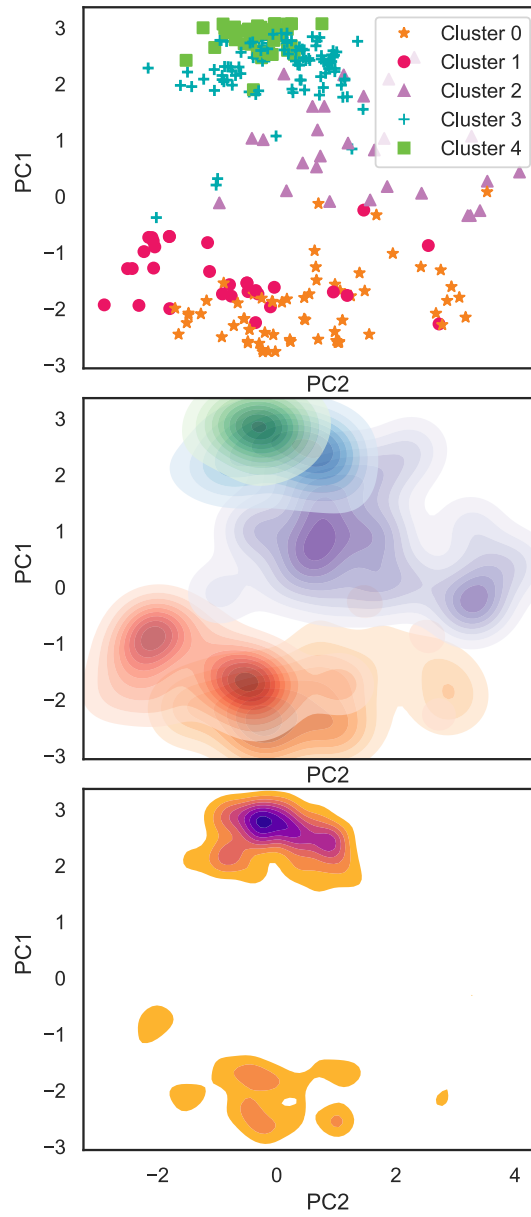
**SI Fig. 2 PC2 plotted against PC1** with datapoints by cluster (top), per-cluster probability density (middle), and overall probability density (bottom).

# ABOUT THE AUTHOR



**Lux Miranda** is graduating from Utah State University with a Bachelor's of Science, double-major in Computational Mathematics and Computer Science, minor in Anthropology. *The two types of society: computationally revealing recurrent social formations and their evolutionary trajectories* is their first scientific publication. Miranda's second publication is up-and-coming with co-author Darcy Bird and will feature a synthetic global dataset of archaeological radiocarbon ages.

Mx. Miranda's academic accomplishments include being part of the inaugural cohort of Peak Summer Research Fellows, receiving a NASA Space Grant Consortium Fellowship to develop flight software for CubeSats, serving as a recitation lecturer for the *Linear Algebra & Differential Equations* course at USU, developing intelligent autonomous systems under an internship at the Space Dynamics Laboratory, and winning a Hackathon by writing a poetry-writing program known as *Will A.I. Shakespeare*.

Mx. Miranda is interested in continuing to use mathematical and computational techniques to help answer big questions about human societies and how they evolve. With their background in space engineering, they are also particularly interested in researching the social structures and practices that will best sustain human settlements in space and on other worlds. They are planning on pursuing graduate studies in computer science beginning Spring 2021.