New Ways of Analyzing Variability: A Digital Ceramic Attribute Analysis

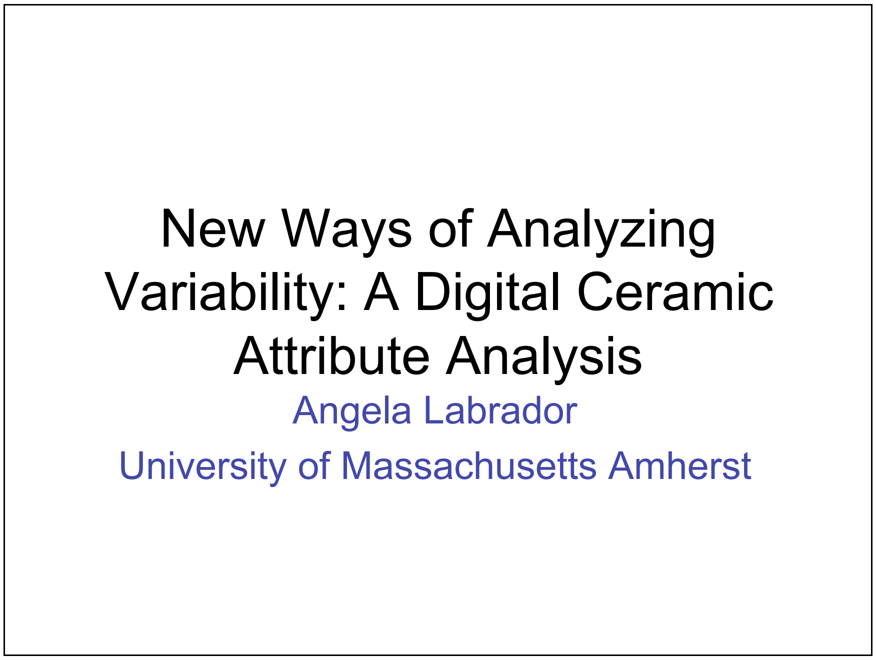
Angela Labrador

presented at the Computer Applications and Quantitative Methods in Archaeology Annual Meeting, "Digital Discovery: Exploring New Frontiers in Human Heritage"
Fargo, North Dakota, April 21, 2006

Abstract:

Our understanding of Algonquin cultures has been limited by the application of typologies to Northeast pre-Contact assemblages. Because artifact types by definition create homogeneity of a complex of attributes, the resulting model obscures variability within each class and reduces culture-historical "stages" to static snapshots of cultural development. In order to record variability in new ways, I utilize an "attribute analysis of technical choice," an alternative method in which a range of attribute states are measured and a minimal vessel count is determined. By recording a range of attribute states, I am able to discern patterns of interdependent behavioral "modes" of manufacturing techniques and formal concepts while remaining sensitive to variance among vessels. In this paper I discuss my attempt to computerize this alternative method using Knowledge Discovery tools. I recognize the manual task of grouping sherds into sets of vessel lots as a clustering exercise, which can be both tedious and subjective. Automating this task with the computer not only saves time, but also addresses the subjective limitations to manual sorting. The end result is an open source, digital toolkit made available to regional archaeologists to enable a more efficient application of attribute analyses to a wider range of ceramic datasets. Moreover, such a toolkit has the potential to allow us to look at variability, not as error in our data, but as the expression of the range of human behavior.

do not cite without permission from the author



New Ways of Analyzing Variability: A Digital Ceramic Attribute Analysis

Angela Labrador

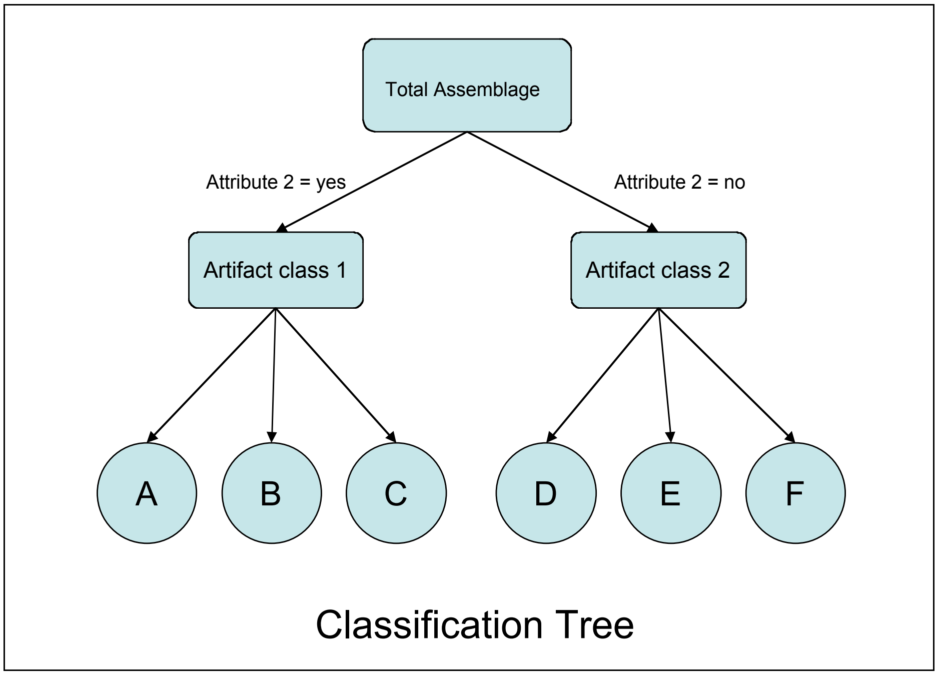
University of Massachusetts Amherst

Slide 1

Typology has been fundamental to the development of archaeological theory and remains the formal language with which we pursue much archaeological research (Adams and Adams, 1991:313). Artifact types, allow us to isolate and organize units of data for research and to convey information about the archaeological record to other professionals (Handsman, 1977:37). However, artifact types create a seeming homogeneity of a complex set of attributes. Thus, typological models also run the risk of obscuring variability within each artifact class and reducing culture-historical "stages" to static snapshots of cultural development (Chilton, 1999:44; Smith, 1979:823).

This paper explores an "attribute analysis of technical choice" (Chilton, 1996) , an alternative method for observing and interpreting ceramic assemblages introduced to New England by Dincauze (1975) and Kenyon (1979) and further developed by Elizabeth Chilton (1996, 1999, 2000). With the support of a grant from the Robert E. Funk Memorial Archaeology Foundation, I have been developing a digital toolkit to further refine this method and adapt it for practicable use in a computer-equipped lab. This paper details the process of my attempt to computerize this alternative method using Knowledge Discovery tools. First, I discuss the problems of traditional ceramic typologies in New England. I next outline how I gathered and structured my data, and then describe the data mining methods that I used. Finally, I detail my important lessons learned and look toward the future of my project.

In New England, archaeologists often rely upon stylistic typologies of ceramics to interpret Woodland chronologies and cultural sequences (Chilton, 1996:6). This focus on attributes of style has limited our ability to observe variability in other attributes of ceramic vessels, which offer insight into production and usage of ceramics (Chilton, 1996:7; Chilton, 1999:45). Another stumbling block is that ceramic type names used in New England have been imported from other regions to the South and West, with the oft-untested presumption that cultural diffusion from those regions is responsible for cultural developments within New England (Chilton, 1996:8; Chilton, 1999: 45).



Total Assemblage

Attribute 2 = yes

Attribute 2 = no

Artifact class 1

Artifact class 2

A

B

C

D

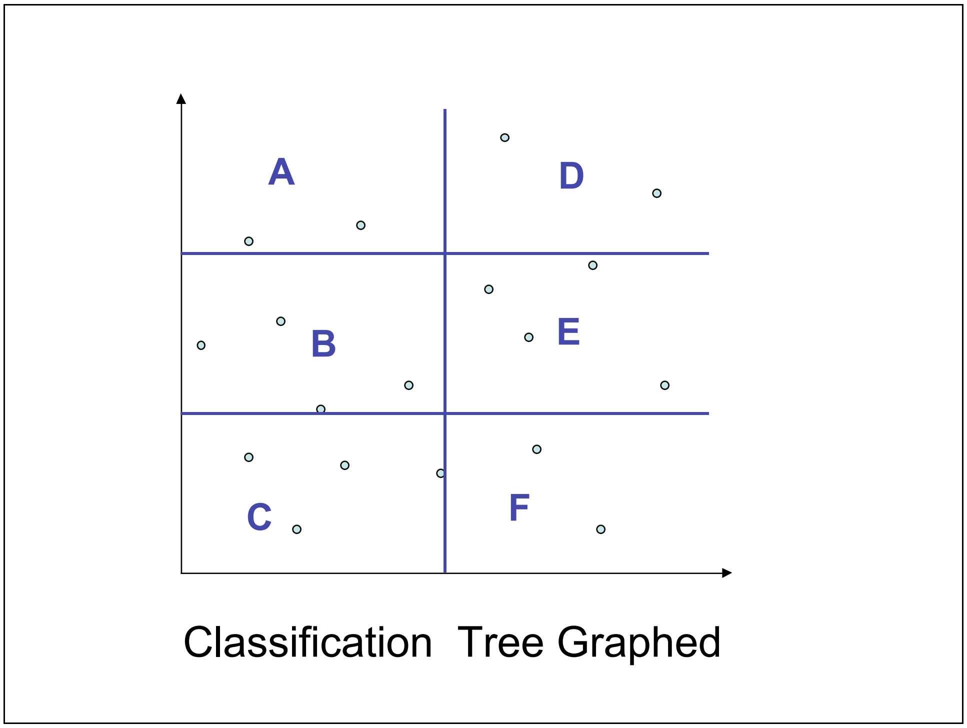
E

F

Classification Tree

Slide 2

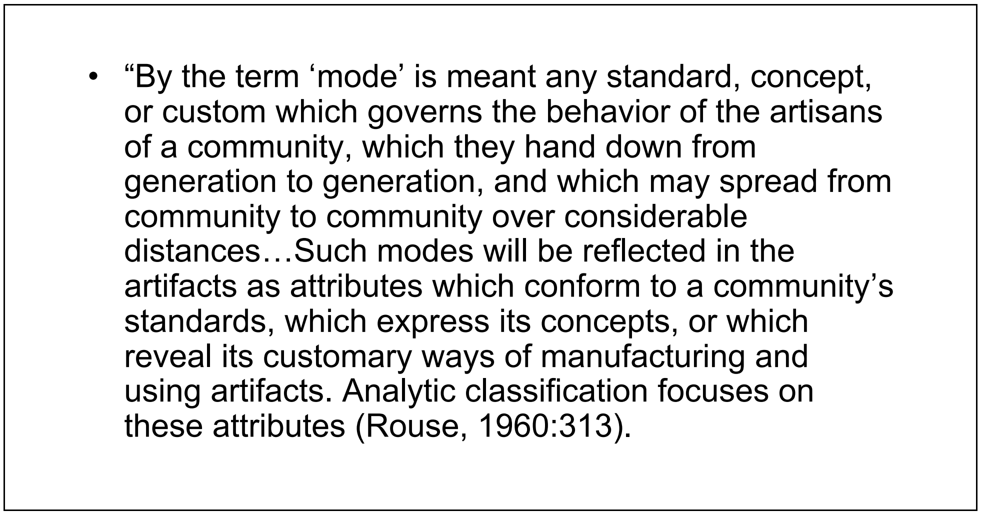
Finally, the taxonomic sorting of New England ceramics into traditions, types, and subtypes presumes a hierarchical structure of attribute complexes . Such structure is inherent in all classificatory schemas, which divide space into mutually exclusive entities and model the boundaries between the classes (Hand, et al., 2001:180).



Classification Tree Graphed

Slide 3

An attribute analysis is an alternative, descriptive approach to identifying "variation *and* co-variation across objects" (Chilton, 1999:46). Irving Rouse defined this method as "analytic classification," which is the attempt to identify behavioral "modes" inherent in artifacts (Dincauze, 1968: 7; Rouse, 1960:313).

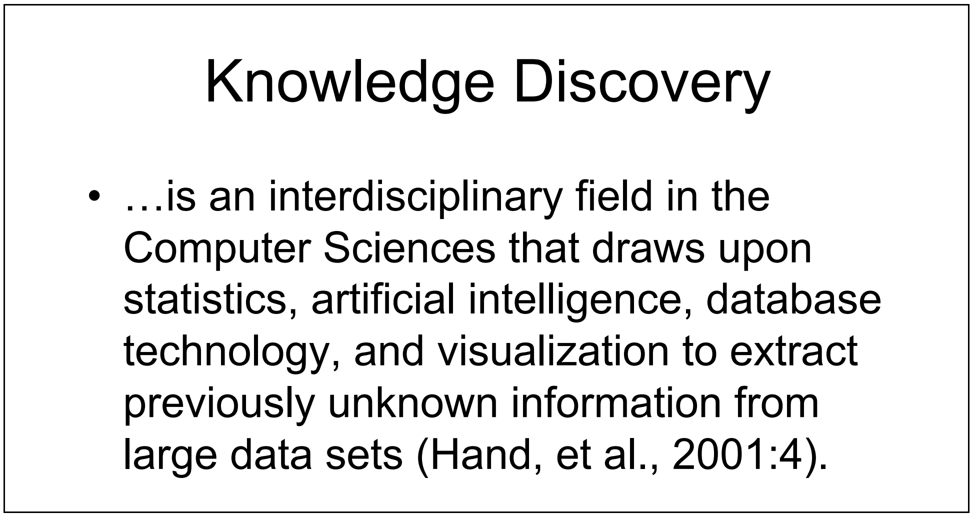


- "By the term 'mode' is meant any standard, concept, or custom which governs the behavior of the artisans of a community, which they hand down from generation to generation, and which may spread from community to community over considerable distances…Such modes will be reflected in the artifacts as attributes which conform to a community's standards, which express its concepts, or which reveal its customary ways of manufacturing and using artifacts. Analytic classification focuses on these attributes (Rouse, 1960:313).

Slide 4

In an attribute analysis of technical choice, vessels, and not sherds are the minimal units of analysis, allowing us to see the ceramic vessel as an aggregative record of technical choices made in sequence by a Woodland potter (Chilton, 2000:102). The initial step of an attribute analysis is to identify the range of attributes that will be recorded for each ceramic sherd and record these attribute states in a database. The next step is to manually sort the sherds into vessel lots using clues gathered from the attribute states, a process akin to sorting out several jumbled jigsaw puzzles. The final step is to interpret what these vessel lots mean.

Thus, the goal of an attribute analysis is not to divide the assemblage up into culture-historic types in order to reconstruct a temporal and cultural sequence. Rather, an attribute analysis helps us get at more fundamental questions about behavior and technology during the Woodland Period. For instance, Elizabeth Chilton's 1996 comparative attribute analysis of Connecticut River Valley and Mohawk Valley ceramics allowed her to identify important cultural differences in intended vessel function between the different groups.

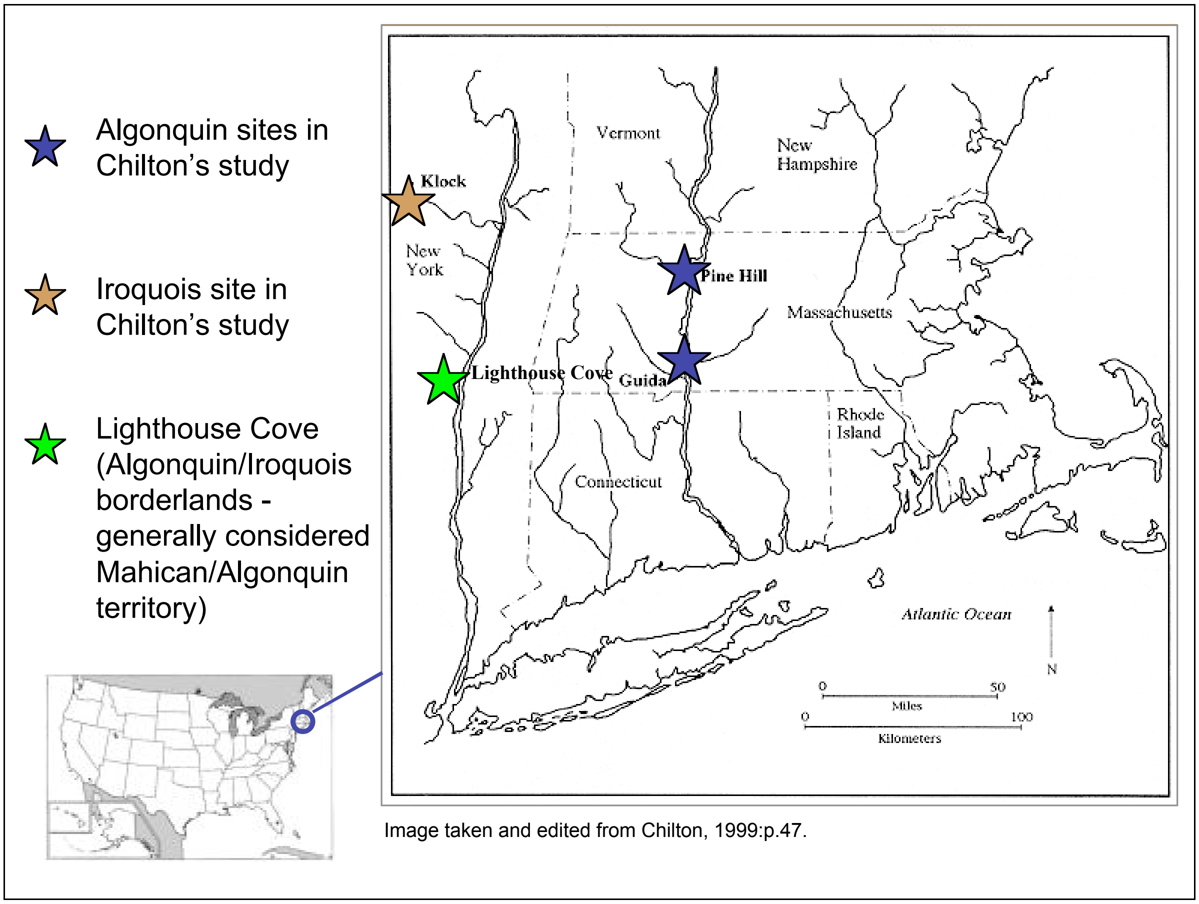


Knowledge Discovery

- …is an interdisciplinary field in the Computer Sciences that draws upon statistics, artificial intelligence, database technology, and visualization to extract previously unknown information from large data sets (Hand, et al., 2001:4).

Slide 5

Inspired by Chilton's work as well as the Knowledge Discovery and Data Mining class I was taking in the Computer Science department during Spring 2005, I decided to try and computerize her methodology using techniques I was learning about in my course. While I would still manually complete the first step of recording the attribute states of each ceramic sherd, I recognized the manual task of grouping sherds into sets of vessel lots as a clustering exercise, which can be both tedious and subjective. Automating this task with the computer not only saves time, but also addresses the subjective limitations to manual sorting. With the combination of these two methods drawn from different disciplines, I could bring to light more evidence of variability in the technological process of ceramics production and usage within the regional type-sequences while also making the process more efficient and practicable for a lab tech to accomplish.



★ Algonquin sites in Chilton's study

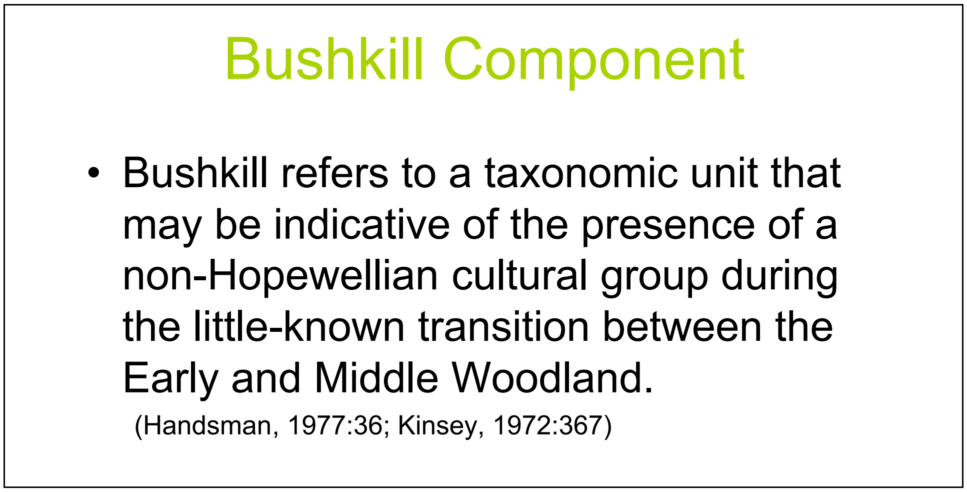
★ Iroquois site in Chilton's study

★ Lighthouse Cove (Algonquin/Iroquois borderlands - generally considered Mahican/Algonquin territory)

Image taken and edited from Chilton, 1999:p.47.

Slide 6

I wanted to choose a site that lies geographically in between the sites that Chilton looked at in her 1996 study. In 2002, I worked as a field student for Dr. Christopher Lindner at a multi-component site known as Lighthouse Cove, which overlooks the confluence of the Esopus Creek and Hudson River in the Tivoli Bay region of the Hudson Valley (Lindner, 2003:1). Not only is Lighthouse Cove in the right geographical location for my study, but also unique to the site is Lindner's research design, which explicitly focuses on the intriguing Bushkill component (Chilton, personal communication, 2/9/2005).



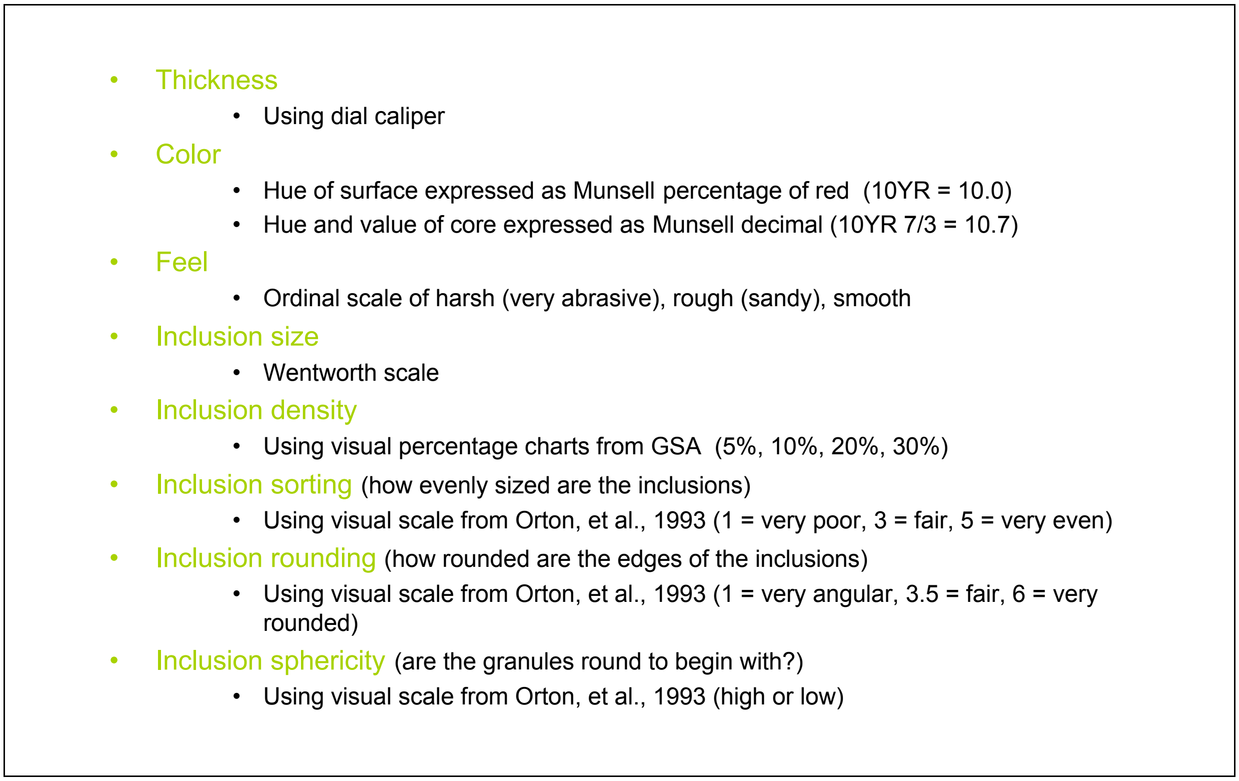
Bushkill Component

- Bushkill refers to a taxonomic unit that may be indicative of the presence of a non-Hopewellian cultural group during the little-known transition between the Early and Middle Woodland.

(Handsman, 1977:36; Kinsey, 1972:367)

Slide 7

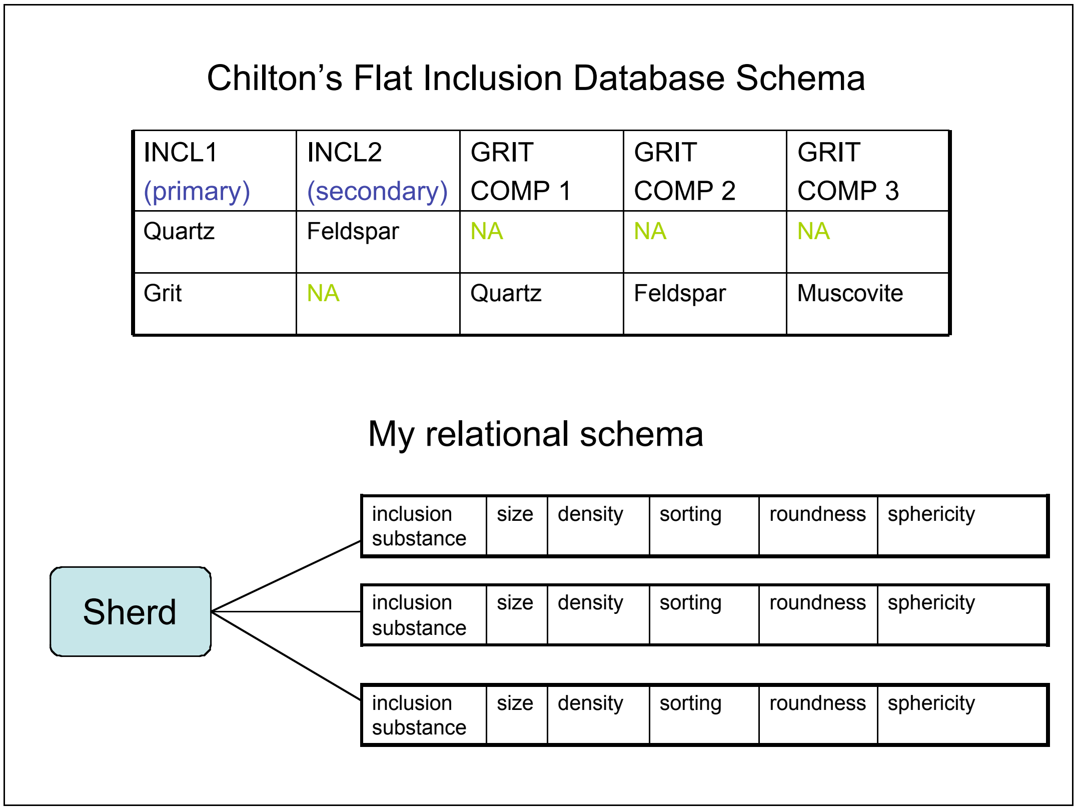
Artifacts identified as belonging to the “Bushkill component” may also provide insight into the debate surrounding the “Iroquoian” and “Algonquin” dichotomy that some archaeologists have read into Northeast assemblages. A ceramic attribute analysis of the Lighthouse Cove assemblage stands to contribute to our knowledge about the lifeways of Early to Middle Woodland people in the Mid Hudson Valley who resided within the borderlands between Iroquois and Algonquin people.



- Thickness
  - Using dial caliper
- Color
  - Hue of surface expressed as Munsell percentage of red (10YR = 10.0)
  - Hue and value of core expressed as Munsell decimal (10YR 7/3 = 10.7)
- Feel
  - Ordinal scale of harsh (very abrasive), rough (sandy), smooth
- Inclusion size
  - Wentworth scale
- Inclusion density
  - Using visual percentage charts from GSA (5%, 10%, 20%, 30%)
- Inclusion sorting (how evenly sized are the inclusions)
  - Using visual scale from Orton, et al., 1993 (1 = very poor, 3 = fair, 5 = very even)
- Inclusion rounding (how rounded are the edges of the inclusions)
  - Using visual scale from Orton, et al., 1993 (1 = very angular, 3.5 = fair, 6 = very rounded)
- Inclusion sphericity (are the granules round to begin with?)
  - Using visual scale from Orton, et al., 1993 (high or low)

Slide 8

Once I had completed the logistical steps of securing permission, funding, and lab space, I then had to actually get in the lab and start identifying the range of attributes I would look at, develop standard metrics, and design the data structure that would organize the data. Although I had Chilton's study and her complete dataset as a starting point, the fact that I would be using a computer to perform the vessel lot cluster step, meant that all of the attributes had to be understood by whatever algorithm I would use. I thus drew from and adapted several ordinal and ratio scales for measuring the attribute states I would record.



Chilton’s Flat Inclusion Database Schema

INCL1 (primary)	INCL2 (secondary)	GRIT COMP 1	GRIT COMP 2	GRIT COMP 3
Quartz	Feldspar	NA	NA	NA
Grit	NA	Quartz	Feldspar	Muscovite

My relational schema

Sherd

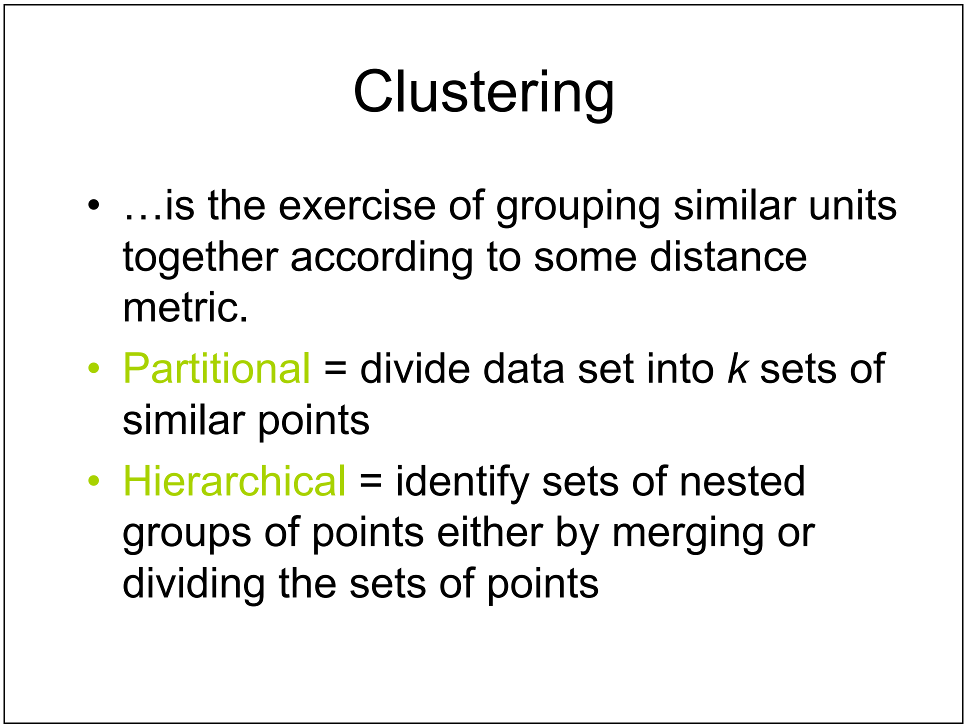
inclusion substance	size	density	sorting	roundness	sphericity

inclusion substance	size	density	sorting	roundness	sphericity

inclusion substance	size	density	sorting	roundness	sphericity

Slide 9

Next, I recognized that the flat, spreadsheet like structure of Chilton’s data could be problematic when it came to identifying temper material. In her study, she identified “INCL1” as a primary inclusion type and “INCL2” as secondary. For temper with more than 2 substances, she coded “INCL1” as “GRIT” and recorded 3 GRITCOMP variables with substance type. I didn’t want to presume such arbitrary hierarchy in the Lighthouse Cove data, especially since after cataloging all 1700 sherds, I had already observed quite the range in inclusion material and number of inclusion types. Instead, I created a relational schema in which any ceramic sherd could have any number of inclusion types. Inclusion type-specific attributes such as maximum size, density, etc. were recorded within the related table on a per-inclusion type basis. Such a schema has no hierarchy of inclusion types built into it and can better manage a wider range of variability in the dataset. However, simultaneous to my recognition of the need for a relational schema was my knowledge that very few clustering algorithms handle relational data well. So, I knew that flattening my data in the “right” way would be tricky. I thus decided to write a flattening method based on inclusion type so that each ceramic sherd would have a field for each inclusion type found in the entire dataset. To a relational data model proponent as myself, such a method results in a rather ugly table with lots of zeroes that reminds me of several flat, mainframe databases I’ve had to upgrade over the past years. However, I figured that as long as my clustering method could handle multiple dimensionality well, a standard computer processor wouldn’t be taxed when running the actual algorithm, and the 0’s (symbolic of NA’s) would literally sort themselves out.



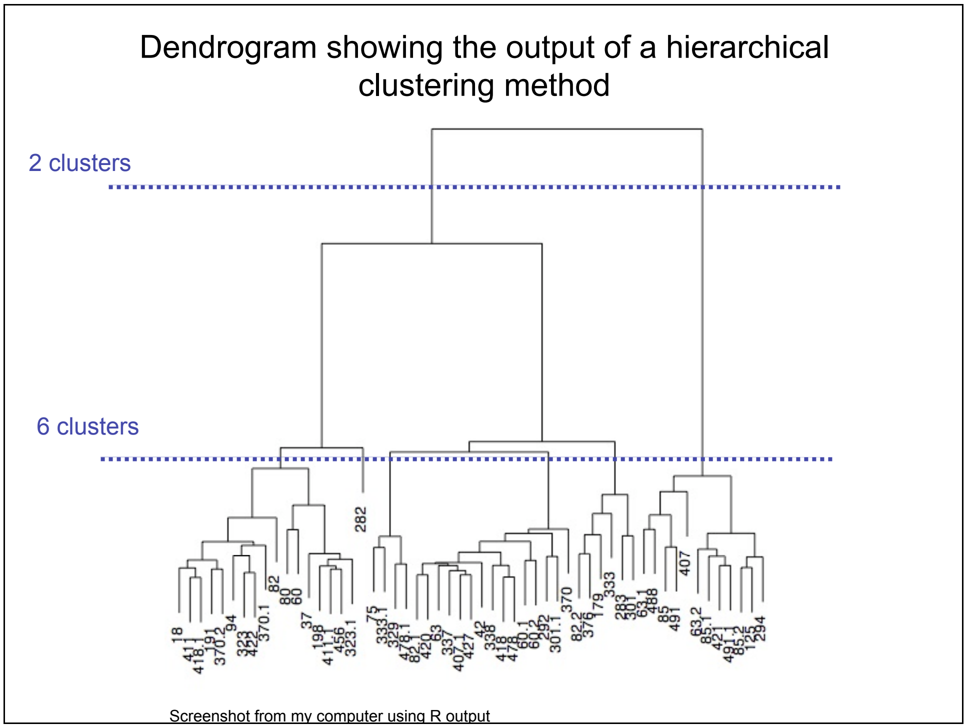
Clustering

- …is the exercise of grouping similar units together according to some distance metric.
- Partitional = divide data set into *k* sets of similar points
- Hierarchical = identify sets of nested groups of points either by merging or dividing the sets of points

Slide 10

After conducting the attribute analysis on a small sample of my dataset, and with the nearing deadline for this paper, it was time to start clustering. There are two basic classes of clustering models: partitional and hierarchical. Partition-based algorithms require one to set *k* at the beginning. This would of course be a problem for my dataset where each data point = a ceramic sherd, because *k* was the entity that I wanted to solve for. That is, I wanted to know *how many* vessels were represented in my dataset of sherds. I knew that a hierarchical method would be the route to take, but I wasn't sure if I wanted to pursue an agglomerative or divisive algorithm.

After listening to Dr. Chilton describe her method of manual sorting as a "lumping" process, I knew to concentrate on the agglomerative hierarchical methods which start the clustering exercise with each point in its own cluster and successively merge points into clusters until you're left with a single supercluster (Hand, et al., 1998:302-308). However, the trick here is to know when to cut the tree off to choose the optimal number of clusters, what specific clustering method to use, and how to handle outliers in the data.

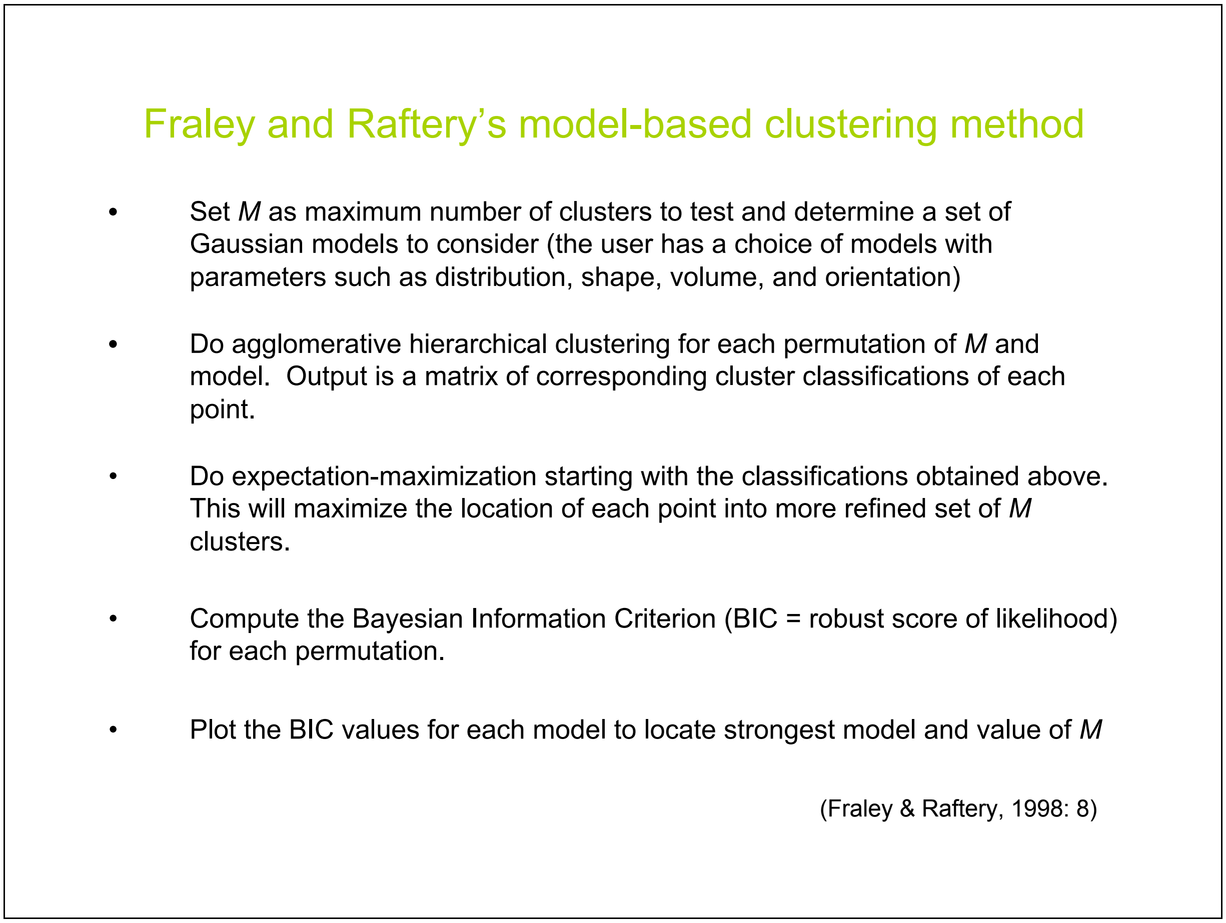


Dendrogram showing the output of a hierarchical clustering method

Screenshot from my computer using R output

Slide 11

I found a promising solution in Fraley and Raftery's model-based clustering method,



Fraley and Raftery’s model-based clustering method

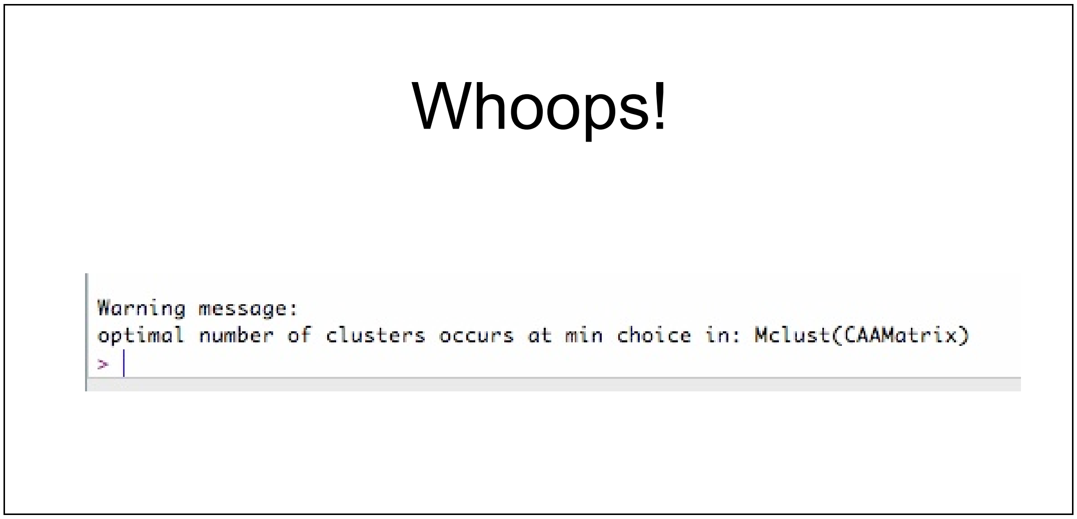
- Set *M* as maximum number of clusters to test and determine a set of Gaussian models to consider (the user has a choice of models with parameters such as distribution, shape, volume, and orientation)
- Do agglomerative hierarchical clustering for each permutation of *M* and model. Output is a matrix of corresponding cluster classifications of each point.
- Do expectation-maximization starting with the classifications obtained above. This will maximize the location of each point into more refined set of *M* clusters.
- Compute the Bayesian Information Criterion (BIC = robust score of likelihood) for each permutation.
- Plot the BIC values for each model to locate strongest model and value of *M*

(Fraley & Raftery, 1998: 8)

Slide 12

which compares models of differing number of clusters and clustering methods, and uses Bayesian statistics to choose the optimum combination. Each component of their method is a probability distribution for each model. The first step in their strategy is to determine the maximum number of clusters to consider and the range of Gaussian models to test. Next, an agglomerative hierarchical clustering for the chosen method is completed and outputs the classification for each of the clusters. Next, iterative relocation of the data points is completed using the classifications from the hierarchical method as starting points. Then a robust score of the likelihood of each component is computed for each parameterization of the EM over the full range of number of clusters, resulting in a matrix of BIC’s. Finally, these scores are plotted for each model to determine what the optimal number of clusters would be for each (Fraley & Raftery, 2002).

Luckily, Fraley and Raftery's model-based clustering method, named Mclust, is available for free as a package for the R language. In the kind of frenzy only worked up when procrastination, spring weather, and approaching deadlines arise, I learned how to read and write in R and carefully studied the entire Mclust documentation and associated packages. Finally, it was time to feed Mclust my data. I wrote my program, ran it, and held my breath as the algorithm churned away. What model would it choose? How many vessels did I have? Success seemed a moment away! Finally, the moment passed and the console returned my answer.



Whoops!

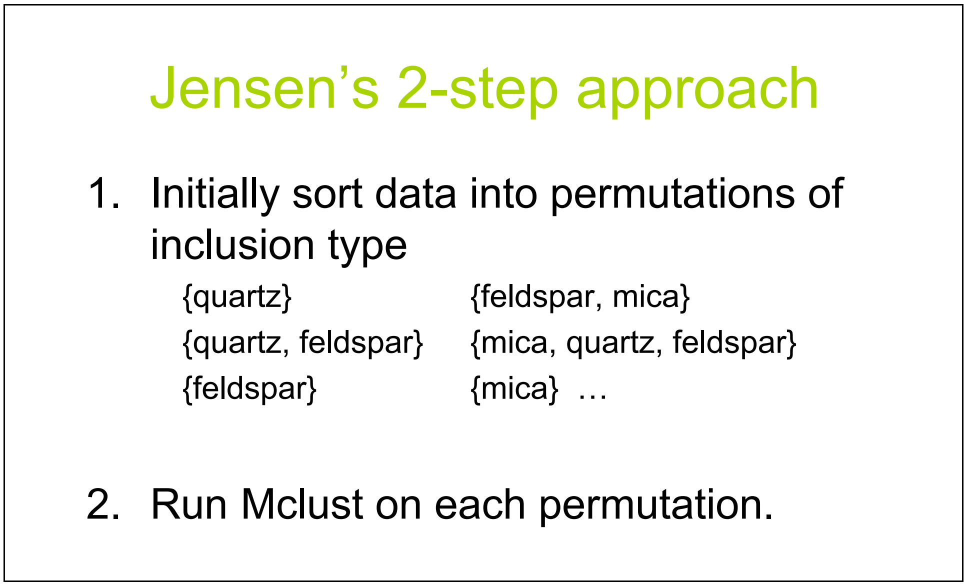
```
Warning message:
optimal number of clusters occurs at min choice in: Mclust(CAAMatrix)
> |
```

Slide 13

That's right, Mclust was telling me I had one cluster – 1 vessel.

The answer I had received was perhaps the worst answer that a clustering algorithm can return. After going through the requisite panic phase, I decided it was time to reevaluate and pay a visit to my Knowledge Discovery professor, Dr. David Jensen. Together, we were able to identify two major problems with my approach. First, my data structure was seriously contributing to error. Second, and more importantly, I didn't have enough data points yet to be getting any meaningful answers from such a complex model-based solution.

Jensen pointed out the structural problems in my dataset, which were due to my need to flatten a relational structure.



Jensen's 2-step approach

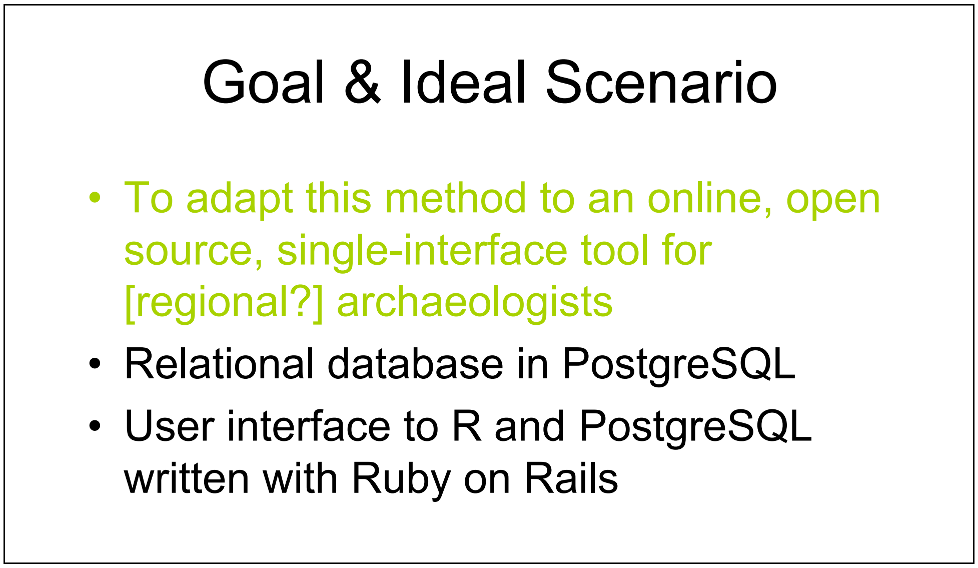
1. Initially sort data into permutations of inclusion type

   {quartz}
   {quartz, feldspar}
   {feldspar}
   {feldspar, mica}
   {mica, quartz, feldspar}
   {mica} …

2. Run Mclust on each permutation.

Slide 14

He advised me to establish a new 2 step method: to first initially sort my data into permutations of my inclusion type combinations and then to run Mclust on each permutation. This makes sense because I already know that those ceramics with mutually differing inclusion types probably can't belong to the same vessel. Next, he reminded me of the bias/variance tradeoff that we often struggle with and said that until I got more data points for each permutation, my model would be skewed by the variance in the few data points I had.   In the meantime, he advised me to test my permutations using simpler agglomerative techniques. It followed that once I conducted the initial sort and used a simpler algorithm, I started seeing the glorious dendrograms and clusters I had been dreaming of. With my confidence reinstilled, I have been able to return to the lab and continue the actual attribute analysis.

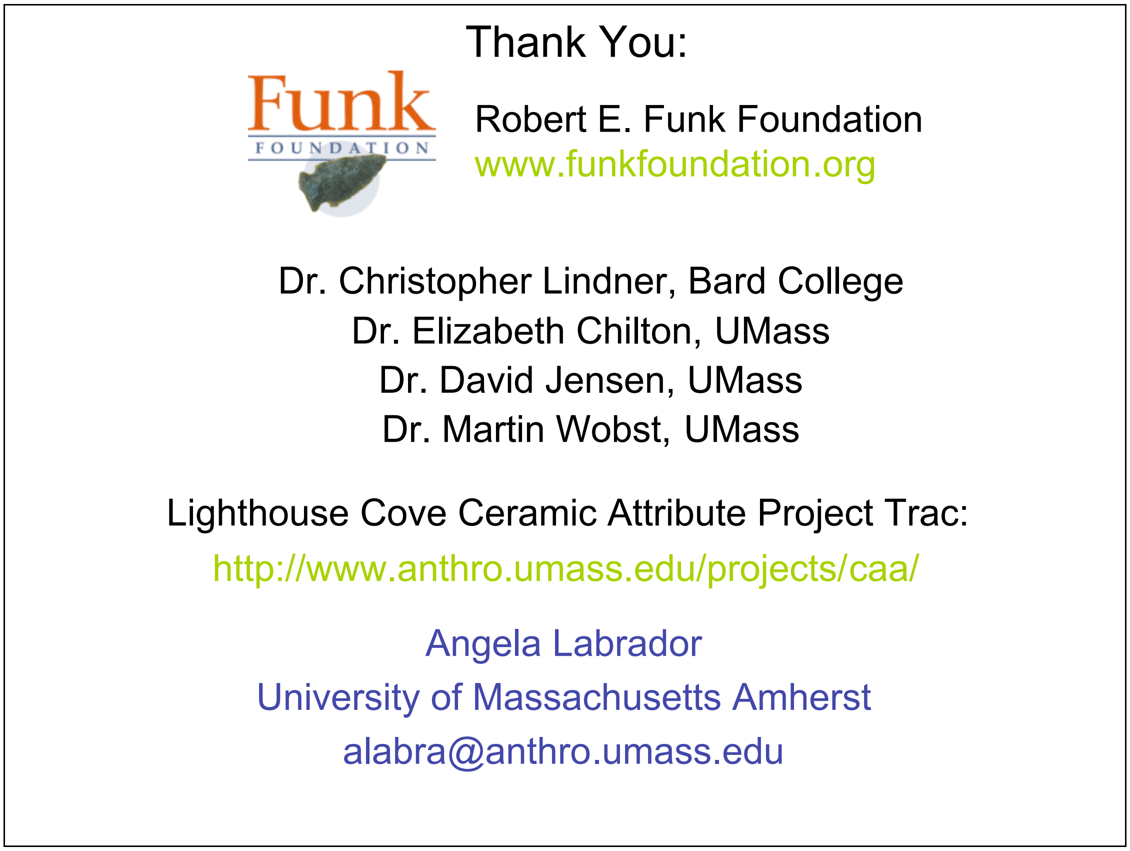


Goal & Ideal Scenario

- To adapt this method to an online, open source, single-interface tool for [regional?] archaeologists
- Relational database in PostgreSQL
- User interface to R and PostgreSQL written with Ruby on Rails

Slide 15

As with most things, this entire process has taken me much longer than I had anticipated. However, my long term goals still remain the same: to adapt this method to an online, open source, single-interface tool for regional archaeologists to conduct their own attribute analyses. Archaeologists would be able to enter their ceramic attribute data into the relational database and with the press of a button be able to retrieve their minimal vessel count for their assemblage at any time. Behind the scenes, the program would flatten the database into a matrix, initially sort it and divide it into inclusion permutations, and use Mclust to identify clusters for each permutation. This method is advantageous not only because it saves the archaeologist time when it comes to the manual sorting of sherds into vessel lots, but also because the algorithm can be run at any time. As more data is added to the database, the vessel count will automatically grow without the archaeologist having to manually return to the original vessel lots for comparison. This is especially beneficial for a field school site such as Lighthouse Cove where each year, the dataset of ceramic sherds grows steadily. CRM firms and museums could also benefit from such a program.



Thank You:

Funk
FOUNDATION

Robert E. Funk Foundation
www.funkfoundation.org

Dr. Christopher Lindner, Bard College
Dr. Elizabeth Chilton, UMass
Dr. David Jensen, UMass
Dr. Martin Wobst, UMass

Lighthouse Cove Ceramic Attribute Project Trac:
http://www.anthro.umass.edu/projects/caa/

Angela Labrador
University of Massachusetts Amherst
alabra@anthro.umass.edu

Slide 16

As I look forward to more lab work, and more programming, I also look forward to more synthesis of the actual results of the attribute analysis. I have funds set aside for petrographic analysis of the ceramics and carbon-14 dating of associated features on site. I plan on future publications to detail not only the progress (and hopefully the completion) of the computer program, but also the archaeological interpretations of these possible Bushkill-component potters. In the meantime, I welcome all those who are interested to visit my project-tracking site at the following url to watch my progress and provide me with any feedback you may have on my project.

Works Cited

Adams, William Y. and Ernest W. Adams

1991 Archaeological Typology and Practical Reality. Cambridge: Cambridge University Press.

Chilton, Elizabeth S.

1996 Embodiments of Choice: Native American Ceramic Diversity in the New England Interior. Ph.D. dissertation, Department of Anthropology, University of Massachusetts Amherst.

1999 One Size Fits All: Typology and Alternatives in Ceramic Research. *In* Material Meanings: Critical Approaches to the Interpretation of Material Culture. Elizabeth S. Chilton, ed. Pp. 44-60. Salt Lake City, UT: University of Utah Press.

2000 Ceramic Research in New England: Breaking the Typological Mold. *In* The Archaeological Northeast. Second Edition. Mary Ann Levine, Kenneth E. Sassaman, and Michael S. Nassaney, eds. Pp. 97-111. Westport, CT: Bergin & Garvey.

Dincauze, Dena F.

1968 Cremation Cemeteries in Eastern Massachusetts. Papers of the Peabody Museum of Archaeology and Ethnology, Harvard University. 59(1).

Fraley, C. and Raftery, A. E.

1998 How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. The Computer Journal 41:578-588.

Hand, David, Heikki Mannila and Padhraic Smyth

2001 Principles of Data Mining. Cambridge, MA: MIT Press.

Handsman, Russell

1977 The Bushkill Complex as an Anomaly: Unmasking the Ideologies of American Archaeology. Ph.D. dissertation, Department of Anthropology, American University.

Kinsey, W.F.

1972 Archaeology in the Upper Delaware Valley. Harrisburg, PA: Pennsylvania Historical and Museum Commission.

Lindner, Christopher

2003 Research Design for the Lighthouse Cove Site, Saugerties, Ulster Co. NY. Submitted to the New York State Museum.

Rouse, Irving
1960 The Classification of Artifacts in Archaeology. American Antiquity. 25(3):313-323.

Smith, Michael E.
1979 A Further Criticism of the Type-Variety System: the Data Can't be Used. American Antiquity. 44(4):822-826.