The field of phylogenetics explores evolutionary relationships between species. By analyzing heritable genetic information for species, the biologist constructs a genetic tree having species as leaves and the branch points indicating when previous ancestors diverged. The length of the edges captures a sense of “evolutionary time,” indicating how genetically close the species on one side of the edge are to those on the other.

Mathematics is perfectly poised to help build these trees, and the tools of algebra, probability, statistics, combinatorics, geometry, and analysis have provided great advances. Indeed, the past decade has seen an explosion of activity in phylogenetics, involving nearly every field of mathematics. Some wonderful books in this area include the classics by Joseph Felsenstein (Inferring Phylogenies, Sinauer Associates, 2004) and Charles Semple and Mike Steel (Phylogenetics, Oxford University Press, 2003).

Today, the prevalence of data has allowed phylogenetics to have a great impact on old frameworks. After about 130 years of status quo, a recent paper in *Nature* provided a challenge to traditional dinosaur classification. Building a phylogenetic tree from new data, the authors argued an alternative relationship between dinosaurs, one in which the stegosaurus and the tyrannosaurus are closer evolutionary cousins than previously assumed (Matthew Baron, David Norman, Paul Barrett. “A new hypothesis of dinosaur relationships and early dinosaur evolution.” *Nature* 543 (2017): 501–506).

Although the original intent of phylogenetics was to build evolutionary trees, the technique can be applied more generally to organize and classify items by similarity. It has become an increasingly useful tool for data analysts who want to classify and relate various items given certain measured traits even if there was no evolutionary process that created them.

This article is far less ambitious than a re-examination of long-dead dinosaurs. But it is possibly more practical. We focus not on creatures, but rather on craft beer. Beer is more accessible to the masses than is wine, but as Charles Bamforth reasoned, it is arguably as subtle and complex in its nature (Grape vs. Grain: A Historical, Technological, and Social Comparison of Wine and Beer. Cambridge University Press, 2008).

Although beer is quite ancient, with recorded history dating back as early as 4000 BC, the modern classification of beer styles began in 1977 with the publication of Michael Jackson's seminal book *The World Guide to Beer* (Prentice Hall). This classification includes two main types of beer, ales and lagers, along with numerous substyles including pilsners, stouts, porters, and wheat beer.

In this study, we compare and classify beer not based on classical labels, but rather using actual data and phylogenetic techniques.
Beer Genes
What types of data should be used to categorize and compare beer? To answer this question, we turned to some experts in the rich history of beer and the complicated brewing process: brewers at the San Diego Brewing Company.

The San Diego Brewing Company was founded by Alonzo Horton in 1896 as the first commercial brewery in the area, and it quickly became one of the largest manufacturing enterprises in the country. The company closed in 1942 (its location was needed for a naval base) but was reopened in 1993 by Scott Stamp. Today San Diego has more than 130 breweries and is internationally recognized as America’s “craft beer capital.”

With the aid of Stamp and head brewmaster Jarrod Davis, we decided on 11 important, quantifiable, and collectable measurements that cover the full beer spectrum, from brewing to finish. The data for each beer includes gravities (density of liquid before and after fermentation), IBU (international bitterness units), SRM (the standard reference method for color), weights (grain, yeast, wet hops, and dry hops), fermentation (time and temperature), and cost per keg. This last data point crudely captures the notion of complexity and risk involved in creating that particular beer. To keep the evolution analogy, we will refer to these categories as beer genes.

We collected data for 21 types of beer (both traditional and seasonal) from the San Diego Brewing Company. These included drafts with names such as American Stout, Berliner Weisse, Infinitude Double IPA, Lakshmi Imperial Red, Old Town Nut Brown, and San Diego Amber. The raw data can be found at maa.org/mathhorizons/supplemental.htm.

The phylogenetic problem is then to reconstruct a tree that best represents this data. There are a variety of methods for doing this. One type of methodology is the exhaustive search method. This builds all possible trees from a given data set and evaluates them based on some optimality criterion. Unfortunately, given $n$ species (or types of beer), there are $(2n - 5)!! = (2n - 5) \cdot (2n - 7) \cdots 1$ distinct binary trees. For our 21 beers, there are over $8.2 \times 10^{21}$ distinct trees to choose from and optimize.

One way to circumvent this exponential search is metric based: construct a “distance” between every pair of species that records the difference between them based on the genetic characteristics. The raw data for $n$ species is converted to an $n \times n$ matrix, where $ij$-entry is the distance between beers $i$ and $j$. If each beer is viewed as an 11-dimensional vector (from its 11 genes), the distance between two types of beer is the distance of their respective points in $\mathbb{R}^{11}$ (using some metric in this space).

In our work, we normalized the values for each trait, so that they fell in the unit interval. Then we scaled each dimension using weights that were determined based on the brewmaster’s opinions on the importance of the different qualities. We then calculated the distances using the classical Euclidean metric.

Beer Trees
Even with a distance-based approach, there is usually not a unique tree that best fits the data. Indeed, only when a matrix satisfies a quite stringent condition (called four point) is the tree unique (see Semple and Steel).

We focused on two classical approaches to phylogenetic tree construction, both of which are fast in execution (suitable for large data sets) and involve clustering, comparing objects based on their pairwise similarities. We outline the two methods and encourage the reader to explore them in greater detail. Numerous software packages are available to construct these trees; we chose PHYLIP (Phylogeny Inference Package from the Department of Genome Sciences at the University of Washington), which does not have strict requirements for classical genetic data as the primary inputs.

The first method, neighbor joining, which is illustrated in figure 1, begins with a tree in the shape of a star with all species (types of beer in

Figure 1. An illustration of the neighbor joining method.
our case) evenly spaced. The algorithm uses the distance matrix to find the two species separated by the minimal distance. These species are then connected to a new vertex which is connected to the central vertex. It then recalculates the distance matrix, replacing the pair of vertices with the newly added vertex. The algorithm repeats this procedure until it yields a tree in which the species separated by the least distances are paired closest together. (This method was introduced by Naruya Saitou and Masatoshi Nei in “The neighbor-joining method: A new method for reconstructing phylogenetic trees,” *Mol. Biol. Evol.* 4 [1987]: 406–425.)

Figure 2 shows the output of the neighbor joining method for our 21 beers. The lengths of the edges joining two types of beer indicate how close they are to each other.

The second method is UPGMA (unweighted pair group method with arithmetic mean), which again looks to the distance matrix to find the closest types of beer. Now, however, the tree is constructed from the bottom up. We assume the final tree is *rooted*; that is, there is a common ancestor for all objects of comparison. Moreover, we assume that there is a constant rate of evolutionary change along each edge, resulting in all leaves being equidistant to the root. After clustering two groups, the algorithm recalculates the distance matrix, where the distance between clusters A and B is the average of all distances between pairs of species between A and B. This continues until all species have been considered and we are left with a rooted tree. Figure 2 shows the output for our 21 beers. (UPGMA was introduced by Robert Sokal and Charles Michener in “A statistical method for evaluating systematic relationships,” *Univ. Kans. Sci. Bull.* 38 [1958] 1409–1438.)

**A New Classification of Beer**

Notice the similarities and differences in the trees in figures 2 and 3: other than the obvious distinction of one being a rooted tree, both have three large subtrees (shaded in orange, purple, green) that contain the same sets of beer. In fact, the two orange clusters and two purple clusters have identical branchings. The similarities are to be expected since both use the same underlying matrix.

It was quite fascinating to see that our trees classified beers not into the classical groups of ales and lagers, but rather into three distinct categories. The orange group focused on beers malty in taste, with a high alcohol content. The purple group contained high bitterness as well as high hop content. The green group was wide ranging, both in taste (from spicy to sour) and color (light to dark); it’s not too surprising to see the vast differences in branchings of this green group between the two tree types.

In choosing our 21 beers, we purposefully added one into the mix (Dark and Stark Weizenbock) that is a very small, obscure, hybrid beer style. We were curious to see where this

---

**Figure 2.** The San Diego Brewing Company beers clustered using the neighbor joining method.
Moreover, phylogenetic trees can be generalized into structures called *phylogenetic networks*, which open the door to further geometric and combinatorial mysteries and complexities. (See Satyan Devadoss and Samantha Petti, “A space of phylogenetic networks,” *SIAM J. Appl. Algebra Geom.* 1 [2017]: 683–705.)

Mathematics is a powerful tool for analyzing the complex and beautiful real world. But as this project showed, the *art* of using mathematics to help make sense of data is part of the challenge and part of the fun.

**Growing a Better Tree**

We produced two phylogenetic trees that aim to describe the San Diego Brewing Company’s products. But there are many more ways to use phylogenetics to investigate them. Choosing different weights for the beer genes, obtaining different beer genes, and employing different reconstruction algorithms are but a few choices that might result in different phylogenies. This is why it is vital to understand the area of study (beer processing, fermentation time, quality of hops, and so on) and to gain wisdom from professionals in the field to produce meaningful analyses.

Beyond tree reconstruction, mathematics offers a far larger world to explore. For example, instead of creating one phylogenetic tree from the given species, there is a geometric *space of trees* in which each point in the space is a different tree. For our 21 types of beer, the resulting tree space would be 18-dimensional complex and would have beautiful combinatorial and topological properties. (See Louis Billera, Susan Holmes, Karen Vogtmann, “Geometry of the space of phylogenetic trees,” *Adv. Appl. Math.* 27 [2001]: 733–767.) Moreover, phylogenetic trees can be generalized into structures called *phylogenetic networks*, which open the door to further geometric and combinatorial mysteries and complexities. (See Satyan Devadoss and Samantha Petti, “A space of phylogenetic networks,” *SIAM J. Appl. Algebra Geom.* 1 [2017]: 683–705.)

Mathematics is a powerful tool for analyzing the complex and beautiful real world. But as this project showed, the *art* of using mathematics to help make sense of data is part of the challenge and part of the fun.

**Satyan L. Devadoss** is the Fletcher Jones professor of applied mathematics at the University of San Diego (USD). He never turns down an invitation for a pint.

**Lia Hebert** was raised in the snow globe town of Park City, Utah. After graduating from USD, she returned to the mountains to work in disaster recovery and business continuity.

**Sophia Raglione** was raised in a small town outside of Portland, Oregon. She followed the sun to San Diego, and after graduating from USD, began her adventure in data analysis.