Article 1:

In the article, "Hidden Markov Model for Stock Trading", Nguyen outlines a four-state HMM forecasting monthly closing S&P500 stock prices. To determine this optimal number of states, the researchers evaluated their model using the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the Hannan-Quinn information criterion (HQC) and the Bozdogan Consistent Akaike Information Criterion (CAIC). After their analysis, they found that a model with 4 states was best for S&P500 stock. They then detailed a three-step prediction process: first, they calibrated their model using training data, then found a previous data set with a similar likelihood as the training, and lastly, generated a prediction using the difference of stock prices of the previous month with a consecutive month. Their procedure rendered impressive results. Even though at times the traditional historical average model (HAR) outperformed their HMM, cumulative squared predictive error analysis proved that in out-of-sample predictions, their model was better. After determining the "validity" of the model, Nguyen put the model to work, trading S&P 500 stock using monthly data in accordance with economic theory (i.e – when the stock return is positive the next month, buy the stock; if it is negative, sell, else, do nothing), and under the assumption that the stock prices would not change month to month. In comparison to the other models (Hand Buy-and-Hold), their HMM provided both higher costs (meaning more trades) and profit percentages, which Nguyen in part attributed to the model's sensitivity to local changes in data.

Article 2:

In the article, "Gene finding and the Hidden Markov models", the authors explain how HMM can be used to locate genes. They first provide a biological background on amino acids and a way to represent proteins viable for statistical and computational manipulation/analysis. In simple prokaryotic cells, gene finding is relatively straightforward, because it is easy to scan the genetic code for start and stop codons. This allows researchers the ability to identify open reading frames (ORFs) and then apply probabilistic weights to these ORFs to siphon out real genes. However, such methods are more complicated in eukaryotes, as the genetic code is littered with noncoding genetic material (introns). Without flexibility and versatility that can take into account these introns, researchers often turn to HMM to both estimate the emission and transition matrices for a particular observed sequence and a hidden sequence as well as estimate the most likely hidden sequence based on an observed sequence, emission and transition matrices. In the latter application, the most likely hidden sequence can be found using the Viterbi algorithm. This algorithm uses dynamic programming to find the most likely hidden sequence associated with a particular DNA sequence through recursion, tabular computations (akin to memoization), and traceback. Because there is often a problem with numerical stability, the last columns of the maximum likelihood matrix generated through a traceback may consist of only zeros. To circumvent this issue, they recommend using log-likelihoods instead (similar to what we do in Bio52).

While the article did not go explicitly in depth on how one can use HMM to find genes, it does detail how one could use it in identifying certain segments of amino acid sequences. In a broad sense, this process takes a set of hidden states where each segment in a particular biological sequence corresponds to one hidden state, and then for each hidden

state, particular segments of the biological sequence is outputted according to the maximum likelihood hidden sequence. Unfortunately, the initial probabilities, transition matrix and emission matrix are often unknown, nor can they often be accurately estimated given a particular, known sequence. However, iteratively applying the process often leads to reliable estimates for the parameters. The identification of hydrophobic and hydrophilic segments in proteins illustrated the use of HMM with both supervised learning to get initial estimates for parameters and unsupervised learning to refine these parameters based on a particular known sequence. The outputted data visualizations demonstrated the correct detection of hydrophobic and hydrophilic regions in the proteins respectively.

Individual Project Source:

In the CityLab article, "The Geography of Coronavirus", it attributes other factors beyond population density to the rapid spread of the contagion. While bustling tourist hubs, industrial centers and densely populated areas certainly play a role in the spread, the article reminds its readers that demographics, social dynamics, cultural inclinations, and economic footprint also determine a location's vulnerability to the most destructive effects of the novel coronavirus. In fact, at the time of publishing, COVID-19 was spreading through urban America at relatively the same rate as in rural America. Furthermore, the article repeatedly cites Jeff Kolko's analysis of the COVID-19, pointing to correlations between COVID-19 per capita death rates and counties with older populations, larger percentages of minorities and colder, wetter climates (although this might be due to the fact that ski resort towns were hit fairly hard). As the article beautifully puts, "it is not density in and of itself that seems to make cities susceptible, but the kind of density and the way it impacts daily work and living." Clearly, there are other factors lurking beneath COVID-19 statistics—factors that belong to human geography *and* physical geography. In other words, race, socioeconomic status, climate, and workforce composition have a stake in the toll the novel coronavirus has on communities too; it's not just population density alone.