Employing Mathematics to Identify the Real Shakespeare

by Paul Chambers, PhD

Alternative authorship theories for the identity of William Shakespeare are dismissed by all but a few professors and Shakespeare scholars, who accept the traditional attribution to William Shakspere of Stratford (Niederkorn). This view is epitomized by William Hunt, a Harvard Scholar who wrote his dissertation on Elizabethan England: "No, absolutely no competent student of the period, historical or literary, has ever taken this theory seriously. First of all, the founding premise is false—there is nothing especially mysterious about William Shakespeare, who is as well documented as one could expect of a man of his time. None of his contemporaries or associates expressed any doubt about the authorship of his poems and plays" (Blakemore).

The contentious debate has continued unabated since the 19th Century. In the 21st Century, however, extraordinary new tools have emerged to resolve complex issues across a wide range of disciplines. With the advent of fast and powerful computers, Artificial Intelligence and Machine Learning have revolutionized many fields and are currently actively employed in areas as diverse as the financial sphere to determine fraud and investment strategies, in the business world to evaluate product potential and marketing, and in the health care sector to predict the progress of diseases and the probabilities of patient hospitalization. These powerful technologies have been brought to bear to resolve the Shakespeare authorship question.

The Dartmouth Study of 2007

One area where Machine Learning has proven useful is the field of text analytics. With the advent of social media, efforts to categorize and analyze textual material using artificial intelligence have become an active area of Data Science. Text analytics is highly effective as a means of supplementing and extending human abilities, adding speed and accuracy for a quantitative, as opposed to a qualitative, assessment of text data (Sabo).

An attempt to apply text analytics to resolve the issue of the Shakespeare authorship using modern computer science was conducted by three students at Dartmouth College. In 2007, they wrote a paper addressing the authorship question using analytics (Seletsky et al). They chose three candidates for evaluation: Sir Francis Bacon, Christopher Marlowe, and Edward de Vere, 17th Earl of Oxford. They employed a series of analytical language metrics to distinguish among the authors, including character usage, word lengths, and the ratio of unique words. What they found surprised them.

For a comparison to Shakespeare's work, they used the following plays of Christopher Marlowe: *Dido, Queen of Carthage; Tamburlaine part 1; Tamburlaine part 2; The Jew of Malta*; and *Edward II*. For Francis Bacon, they used the prose works *The Great Instauration*, *Preparative toward a Natural and Experimental History*, and *New Atlantis*. Since no known plays are attributable to Edward de Vere, they compared his poetry to the poetry of William Shakespeare.

The first analytical test they conducted was a comparison of character distributions. This meant evaluating the frequency of appearance of individual letters in the works. For Marlowe they found a significant difference between the usage of his letters: Marlowe tended to use the vowel "e" far more often than Shakespeare. The overall differences were so large in this case that the authors concluded with high statistical certainty that the works of Shakespeare and Marlowe originated from different sources.

Sir Francis Bacon fared no better in this test. Bacon seemed to use longer words than Shakespeare and had significantly more usage of the letters "t," "i," and "e." They also found a significant difference between Bacon's work and Shakespeare's so as to make it unlikely that they came from the same

Paul Chambers *holds a doctorate in engineering from the University of Maryland at College Park. He has performed data analytics as a contractor for the EEOC in Washington DC where he developed algorithms to detect statistical demographic pay disparities and for the Center for Medicare Services in Baltimore where he performed statistical modeling for the care and treatment of dialysis patients. He has also served as a Senior Data Scientist in the private sector for Hitachi Consulting.*

source—although the smaller number of characters made it more difficult to reach a conclusion with confidence. While Oxford also seemed to have a different frequency of letter usage from Shakespeare, the small corpus of his work caused this result to be the least reliable. However, his match, based on statistical tests, was far closer than the other two candidates for this metric.

The second test employed was word length analysis, which compared the distributions of words and their lengths used in the corpus of papers. The first thing the scholars noticed was that Shakespeare used significantly more four-letter words than three-letter words while Marlowe used more three-letter words than any other size. Although there were clearly differences, this metric was unable to definitively distinguish between the works of Shakespeare and Marlowe with confidence.

Francis Bacon was another matter. The word length distributions of Bacon and Shakespeare were so statistically different that the authors concluded it was extremely unlikely that Bacon ever wrote under the name of Shakespeare. Applying this same metric to Edward de Vere, however, the authors were surprised to find that based on word length analysis, the works of Shakespeare and Oxford were virtually indistinguishable with high statistical confidence, based on a p-value, a measure of statistical confidence, of p = 0.4 (with a maximum possible value of 1.0, p = 0.05 is usually considered the cut off point for hypothesis testing). Stated statistically, based on this metric, the hypothesis that the plays of Bacon and the plays of Shakespeare were written by the same author was rejected, while the hypothesis that the poems of Shakespeare and the poems of Oxford were written by the same author was not rejected. While this doesn't mean necessarily that Oxford wrote the works of Shakespeare, the statistical match for this metric was so close that the authors concluded that "the two may have written under the same name" from this test alone.

The last metric employed was the proportion of unique words. This is a novel analytic that calculates the proportion of words that appear just once compared to total words in a corpus. The five plays of Marlowe showed an average ratio of 0.207 with a very small variance (a statistical measure of the overall degree of disparity between each ratio and the average) of 0.0005. Francis Bacon showed a similar result. His average ratio was 0.204, very similar to Marlowe's, again with a small variance of 0.0012. Shakespeare's corpus showed an average ratio of 0.16 with a variance of only 0.0002. Because the margins of error were so small and the ratios were so consistent and precise for each author, it was clear that both Marlowe and Bacon exhibited statistically significant differences from Shakespeare. Based on this metric, the hypothesis that the plays of Marlowe and Shakespeare were written by the same author was rejected, and the hypothesis that the plays of Bacon and Shakespeare was written by the same author was also rejected, a compelling indication that neither man wrote under the pseudonym "William Shakespeare."

The most significant discovery came when they compared the unique word ratios of Edward de Vere's poetry to Shakespeare's poems. De Vere's works had a ratio of 0.31 while Shakespeare had a ratio of 0.30. This was an almost exact match. It lent further confirmation to the results from the word length ratios, essentially that the works of the Earl of Oxford were analytically indistinguishable from Shakespeare's, leading them to suggest that "perhaps de Vere was Shakespeare" and that the "Oxfordian camp may have some veracity" (Seletsky 4). After considering the personal connections and autobiographical elements of de Vere's life in the works, these authors ultimately concluded that they were "very doubtful that Shakespeare did in fact write his plays."

An Independent Study Using Text Analytics

This result was intriguing enough to warrant further analysis. Toward this end, I recently applied modern text mining analysis to the problem. My approach differs from the Dartmouth group because I employ an unsupervised learning methodology. In this analysis I only seek similarities among the works using a technique called text mining. Text mining analytics is currently used in such diverse applications as spam filtering, business intelligence, and fraud detection (Williams).

For comparison, I chose nine contemporary authors for affinity to Shakespeare together with two authors from the 19th and 20th Centuries as a sanity check. For analysis from Shakespeare's era, in addition to Oxford, I chose the poets and playwrights John Donne, Edmund Spenser, Christopher Marlowe, John Webster, John Fletcher, Thomas Dekker, Ben Jonson, and Francis Beaumont; I also included the modern American poets Walt Whitman and Ogden Nash for contrast. While some authors, such as Marlowe and Spenser, were writing during the same time as Oxford, other authors such as Donne, Beaumont, Fletcher and Webster, were writing later than early Shakespeare and Oxford. Because Shakespeare's early poems were very popular and printed numerous times, these later authors may have been influenced by his work. This therefore provides an acid test for de Vere as it compares his work to authors who would have had access to Shakespeare's poetry and may, in turn, have been influenced by it.

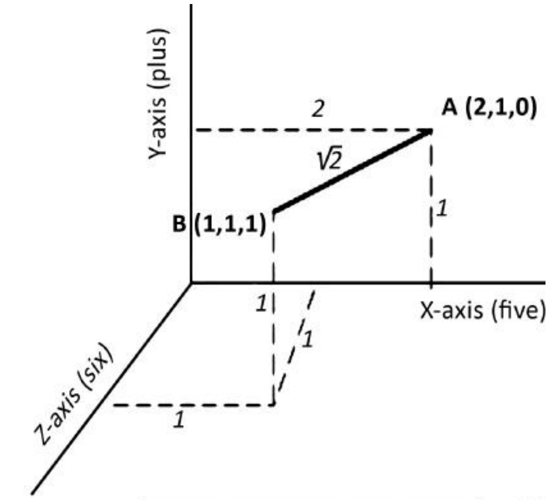
Text mining involves creating a term document matrix from a corpus of works. The works of each author were assembled into a single document for each. Stop words like "the" and "and" together with punctuation were removed from each corpus. Modern spellings were used where possible. The terms used by each author were then automatically counted and placed in a table that shows word occurrence together with the number of appearances of each word by author. An example is shown in figure 1. This term document matrix was created with the R package TM (text mining) and has frequencies for more than 10,000 different words.

Because the total body of de Vere's extant publications is so small—fewer than 4,000 words—to get a meaningful comparison I broke up longer works into fragments to get comparable document sizes. I separated the *Sonnets* and the *Rape of Lucrece* into three pieces and *Venus and Adonis* into two parts. There are other ways of comparing documents of disparate sizes but breaking the longer works into smaller parts has advantages.

	abate	abide	abjure	able	abound	about	abridg-ment	absence	abusd	aby	accent	accidents
1	1	2	1	2	1	1	1	1	1	2	1	1
2	0	0	0	1	0	0	0	1	1	0	1	0
3	1	0	0	1	0	3	1	0	1	0	2	1
4	0	2	0	0	0	2	0	1	0	0	0	0
5	0	2	1	0	1	0	0	2	0	0	0	0
6	0	0	0	0	0	0	0	3	7	0	1	2
7	1	0	0	2	0	0	0	0	1	0	0	1
8	0	2	1	0	0	0	0	0	0	0	0	2
9	0	1	0	2	1	2	0	5	1	0	0	0
10	0	0	0	0	0	1	0	1	1	0	1	0

Figure 1: First portion of term document matrix showing word frequencies for a series of contemporary authors to Shakespeare.

The next step is determining the degree of similarity between the works of each author. This is accomplished by means of a distance metric. Because a number is assigned to each word in the corpus, it represents a point in a multidimensional space. Figure 2 illustrates this for two documents A and B. Only three words appear in each document, say "five," "plus," and "six," but their frequencies vary. In A, the first word appears twice, the second once, and the third not at all, yielding the point (2,1,0). In B, each word appears exactly once for (1,1,1). The distance between the two points is calculated from the simple formula for distance metric provided in figure 2.



$$D(A, B) = \sqrt{(2-1)^2 + (1-1)^2 + (0-1)^2} = \sqrt{2}$$

Figure 2: Example of two points in a term document matrix using only three words together with the distance metric formula.

The difference for the term document matrix in this case is that the corpus of words is much larger. The same operations to calculate the distance between each point in space, each corpus, are used except that it is in a much higher dimensional space, in this case over 10,000 dimensions. Once the distances between each document are calculated, the results are grouped together using a process called agglomerative clustering. In this method, each document is assigned first to its own cluster. Then the algorithm finds pairs of clusters that are closest to each other and merges them. The pair of documents in each new cluster can be represented by a tree-like structure called a dendrogram. Then the distances are computed between the new clusters and the closest clusters are linked together in a higher level tree-like structure. This process is continued until a complete tree structure is produced showing the documents, their groupings, and their nearest interrelationships based on the distance metric.

The result is a hierarchical clustering Dendrogram that shows the closest connections and inter-relatedness between the documents based on their word frequencies. This Dendrogram was calculated from the term document matrix using the *hclust()* command ("average" method) from the "stats" package in R and is shown in figure 3.

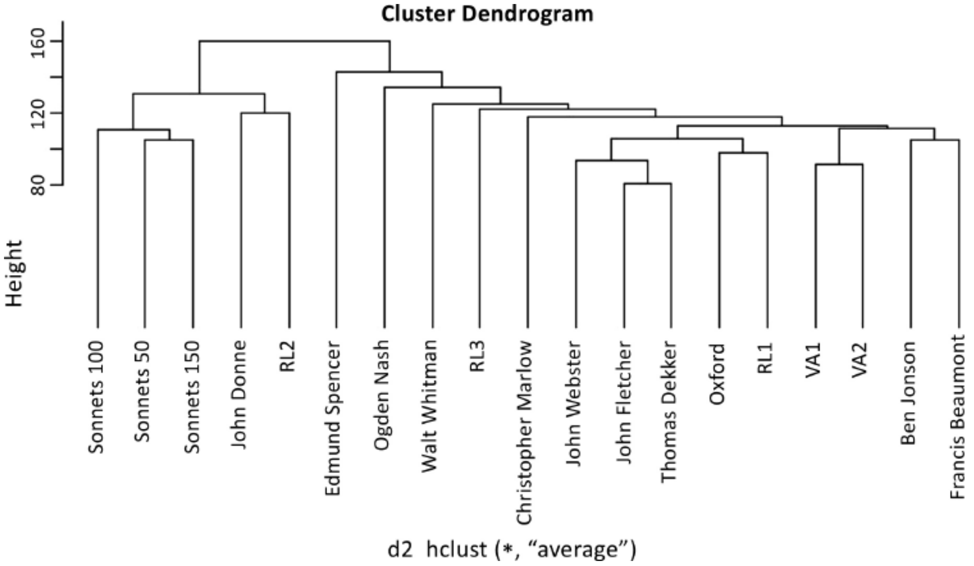


Figure 3: Cluster Dendrogram comparing a series of authors to Shakespeare's works.

Although this technique may deny poetry its rhyming scheme, flow, sounds, even its human and dramatical elements, it does allow for strictly mathematical and analytical comparisons. What emerges from this Dendrogram is that the *Sonnets* all cluster together even though they were broken into smaller pieces and assigned their own document in the matrix. The two parts of *Venus and Adonis* also cluster next to each other. The American poets Ogden

Nash and Walt Whitman cluster adjacently, a reasonable result since they are both from later centuries than the other poems that were compared. Of further interest is that Francis Beaumont clusters next to Ben Jonson. This is significant in that Beaumont was Jonson's student. Even though it transforms poetry into mathematics, this analysis yields remarkably consistent results and thereby shows merit.

The most striking aspect of the Dendrogram is that Oxford clusters directly with the first part of *Rape of Lucrece* and adjacent to both parts of *Venus and Adonis*. Of the contemporary authors considered, his work is clearly closest to the earliest poems of Shakespeare. What makes this exciting is that Oxford's works stop appearing in print just when Shakespeare's poems begin to appear in the same year, 1593 (Leubering). The fact that De Vere's poems cluster next to Shakespeare's early work seems to be too much of a coincidence. This has all the earmarks of an author writing under a new name and is consistent with the Oxfordian theory of authorship. That Oxford's early poems do not cluster near the *Sonnets* is not surprising, as his later work likely matured and exhibits disparate word frequency usage. John Donne died in 1631, and his poems were published posthumously in 1633. It should not be surprising that Donne's poetry may have been strongly influenced by Shakespeare's work and potentially explains why his material clusters near the *Sonnets*, published in 1609.

While this result is thought-provoking, it is not by itself definitive. It doesn't prove that Edward de Vere was the true author of Shakespeare's works. However, it does constitute an important piece to the authorship puzzle. As it turns out, there is a systematic mathematical way to assemble the pieces: Bayesian Analysis.

Bayesian analytics is based on Thomas Bayes' theorem and considers the probability of an event happening given that a prior event has already occurred. It is given by a simple formula which relates the probability of hypothesis H before getting the evidence, to the probability of the hypothesis after obtaining the evidence. Used analytically, it provides a systematic framework for ascertaining the likelihood of belief in a hypothesis based on probabilities and probability distributions. While Bayesian analysis has been criticized by classical statisticians as being subjective, it does provide a way of making the subjectivity explicit (Britannica "Bayesian analysis"). Bayesian analysis has found application in statistical decision theory to make better decisions as well in bioinformatics to calculate the probability of an individual having a specific genotype. For instance, to determine the chances of being affected by a genetic disease or the likelihood of being a carrier for a recessive gene of interest, Bayesian analysis is performed using family history or genetic testing to predict whether an individual will develop a disease or pass one on to their children (Kraft 790–97).

Bayes' theorem relates the probability of belief in a hypothesis to a prior belief based on the acquisition of new evidence (Equation 1).

$$P(H|E) = \frac{P(E|H)}{P(E)} P(H)$$

Here P(H|E) is the probability of hypothesis H occurring given that event E has already occurred. This is directly proportional to the probability of event E given hypothesis H times the initial probability of H divided by the probability of event E given by P(E).

As an example, consider a deck of playing cards. If a face card is drawn from the deck, what is the probability that the card is a King? Since there are 13 possible cards in the deck, one for each of the four suits, the probability that any card drawn is a King would be P(King) = 1/13. Since every King is a face card, the probability of a King being a face card is 100%, P(Face|King) = 1. Each suit has three face cards (Jack, Queen, King), so the probability of drawing a face card is P(Face) = 3/13. Using Bayes' theorem to determine the probability that the card is a King given that a face card is drawn would be P(King|Face) = (1/13) / (3/13) = 1/3. This result is accurate: since there are only three possible face cards the probability that a drawn face card is a King must be exactly 1/3.

In this case, it seemed that Bayes' theorem was a complicated way to get a simple result, but there are situations where the theorem has advantages. Consider a legal case where the guilt or innocence of a criminal defendant is at issue. This can be determined from a modified version of Bayes' theorem based on the law of total probability (Fienberg 771–88) (Equation 2):

$$P(G|(E_n \text{ and } H)) = \frac{P(G|H)P(E_n\,|(G \text{ and } H))}{P(G|H)\;P\,(E_n\,|(G \text{ and } H)) + P(NG|H)\;P(E_n|(NG \text{ and } H))}$$

Where P(G|H) = probability of Guilt given events H (a summation of prior events), P(NG|H) = probability of Not Guilty given events H and E_n represents the current event under analysis. The denominator reflects the total probability of the event E_n, given the two possibilities, in this case guilt or innocence, under consideration.

While this equation looks formidable, it basically allows for the calculation of the probability of a defendant's guilt based on a summation of prior evidentiary events. The equation is used in an iterative fashion to incorporate new knowledge as it becomes available. The probability in the belief of guilt is derived from prior belief in guilt G or innocence NG considering the new evidence E_n for each iteration.

As an example, consider the hypothetical criminal defendant. Since a defendant is entitled to a presumption of innocence in the American system of justice, he can reasonably be assigned an initial probability of guilt of say just 10%. Bayesian analysis requires a non-zero starting point and there must be some finite probability of guilt for a defendant to be accused or arrested.

Belief in the defendant's innocence, NG, is therefore P(NG) = .90, while belief in his guilt is just P(G) = 0.10 starting out. A blood sample is found at the scene with Type O, a match for the defendant's blood type. But, type O blood is found in 45% of the population, so if the defendant is innocent there is still a 45% chance that his blood would match by chance, $P(E_1|NG) = .45$. If the suspect is guilty, however, then there is a 100% chance that his blood will be a match to the crime scene serum, $P(E_1|G) =1$. Applying these proportions into the Bayesian formula for analysis, the probability of innocence now dips slightly to ~80% while belief in the guilt of the defendant increases to ~20%. For the first iteration (Equation 3):

$$P(G|(E_1)) = \frac{P(G)P(E_1|(G\ ))}{P(G)\ P\ (E_1|(G)) + P(NG)\ P(E_1|(NG))} = \frac{(0.1)(1)}{(0.1)(1) + (0.9)(0.45)} = 0.198$$

A partial fingerprint is found at the scene. It matches the defendant's reasonably well, but there is a 21% chance that the match could occur randomly. The next iteration of analysis incorporates this probability building on the results from the prior evidence, where now $P(E_2|NG) = 0.21$ and $P(E_2|G) = 1$. The second iteration uses the values for guilt and innocence calculated in the first iteration, P(G) = 0.198 and P(NG) =.802. This new evidence raises belief in the suspect's guilt to 54% (Equation 4):

$$P(G|(E_1 + E_2)) = \frac{P(G)P(E2|(G\ ))}{P(G)P(E_2|(G)) + P(NG)\ P(E_2|(NG))} = \frac{(0.198)(1)}{(0.198)(1) + (0.802)(0.21)} = 0.54$$

Finally, DNA is extracted from the dried blood sample. The DNA is a match for the defendant's with only a 1.7% chance of error. The low likelihood of this evidentiary result occurring randomly if the suspect is innocent seriously lowers his odds of being not guilty, $P(E_3|NG) = 0.017$, thereby increasing the probability of his overall guilt significantly. Incorporating this final evidence into the analysis raises belief in the defendant's guilt to an overwhelming P(G|H) = 98.6% while belief in his innocence falls to an abysmal P(NG/H) = 1.4% (Equation 5):

$$P(G|(E_3{+}H)) = \frac{P(G)P(E3|(G\ ))}{P(G)P(E3|(G)) + P(NG)\ P(E3|(NG))} = \frac{(0.54)(1)}{(0.54)(1) + (0.46)(0.017)} = 0.986$$

Stated another way, the defendant is almost 70 times more likely to be guilty than innocent. This result would easily meet a "beyond reasonable doubt" standard and would be enough to convict the suspect of the crime.

Applying Bayesian Mathematics

Since the above equation relates to a binary outcome, in this case guilt or innocence, it is equally applicable to the Shakespeare authorship question. It can be used to provide a likelihood of Oxfordian authorship of the works of William Shakespeare compared to William Shakspere from Stratford upon Avon, designated as Shakspere for this analysis. Application of Bayesian analyses to the authorship question has been realized previously in a book-length work that considered a wide range of factors (Sturrock). The example analysis that follows considers only a small number of select factors, specifically temporal correlations and the text-based analytical results presented above. The focus here is on relevant substantive events and results, and not carefully chosen trivia or arcana.

To apply the full Bayesian analytical framework to the case for Oxford's authorship first requires a starting point. This involves comparing what is known of the two most popular candidates: Edward de Vere and William Shakspere from Stratford. The choice of a starting point is subjective; the best that can be hoped for is a reasonable estimate that can be fairly justified based on the known historical and literary evidence.

The evidence in favor of Oxford's authorship candidacy is compelling by any sensible standard (Bethell 45–61). Oxford wrote some of his poetry in iambic pentameter, a style invented by his uncle, Henry Howard, Earl of Surrey, and used by Shakespeare. Many of Shakespeare's plays such as *Hamlet*, *Romeo and Juliet*, *Othello*, and *All's Well That Ends Well* inexplicably feature numerous events from Oxford's personal life. Indeed, Shakespeare's masterpiece *Hamlet* seems to be a virtual biography of Edward de Vere (Londre). Multiple contemporary authors list Oxford as the best playwright of the Elizabethan Court, especially for comedy (Francis Meres in 1598), yet, surprisingly, none of his plays have survived—even though numerous letters and correspondence are extant including 23 early poems.

Contemporary author George Puttenham wrote in the *Art of English Poetrie* (1589), "And in her Majesties time that now is are sprong up an other crew of Courtly makers Noble men and Gentlemen of her Majesties owne servauntes, who have written excellently well as it would appeare if their doings could be found out and made publicke with the rest, of which number is first that noble Gentleman Edward Earle of Oxford" (Nelson 386) In 1622, Henry Peacham published *The Compleat Gentleman*, which not only ranked Oxford as the best poet of the Elizabethan age but failed to even mention William Shakespeare, a telling omission (Anderson, Epilogue).

> In the time of our late Queene Elizabeth, which was truly a golden Age (for such a world of refined wits, and excellent spirits it produced, whose like are hardly to be hoped for, in any succeeding Age)

> above all others, who honoured Poesie with their pennes and practise (to omit her Majestie, who had a singular gift herein) were Edward Earle of Oxford, the Lord Buckhurst, Henry Lord Paget; our Phoenix, the noble Sir Philip Sidney, M. Edward Dyer, M. Edmund Spencer, M. Samuel Daniel, with sundry others;

Oxford had access to the best education the age could provide, first with private tutors, then at St. John's College at Cambridge University. He studied law at Gray's Inn. He had direct access to the vast libraries of Sir Thomas Smith and Sir William Cecil while growing up in their homes, an important point given that public libraries did not exist in Elizabethan England. He became fluent in French, Italian and Latin. He traveled to France, Germany and Italy for 16 months and spent time in the courts of France and Italy, where 10 of Shakespeare's plays are set. He could write convincingly about the nobility because he was the senior Earl of the Elizabethan Court. Equally important, de Vere had a theatrical background, serving as patron of two theatrical troupes, Oxford's Men and Oxford's Boys. *Shakespeare's Sonnets* complain of the maladies of old age, lameness, and the loss of his good name, all which Oxford had to endure. Although subjective, the weight of literary and historical evidence in Oxford's favor suggests a substantive starting point for his candidacy.

Shakspere, by contrast, had no known formal education and never travelled outside of England. According to the archival records, he was a successful businessman, real estate investor and actor in the Lord Chamberlain's Men. He was relatively young when the *Sonnets* were written and was not known to have been lame or to have suffered a tarnished reputation. Given the disparities in education, access to libraries, the literary world and worldliness, it would not be inappropriate to start both men out at an equal 50/50 probability of Shakespearean authorship despite the near consensus of scholarly opinion to the contrary.

However, bowing to the weight of academic scholarship and giving Shakspere the benefit of the doubt, for the sample analysis that follows, Oxford will start at a 5% probability of authorship with Shakspere at 95%, P(Oxford/Author) = P(OA) = 0.05, P (Oxford/Not Author) = P(ONA) = 0.95. A higher starting point could be justified, and a lower one could be taken as well. However, 5% is not unreasonable based on the literary, historical, and biographical evidence in Oxford's favor and his popularity among non-Stratfordians as the leading alternative authorship candidate. The choice of a starting probability, however, is subjective and the reader is invited to choose a starting point that seems most reasonable and appropriate for the analysis.

Once a starting point is selected, the first key piece of evidence available to input into the Bayesian analysis is the start of Shakespearean publication.

The first publication ascribed to William Shakespeare appeared in 1593 as the long poem *Venus and Adonis*, the same year that Oxford ceases publication. "Oxford's 23 acknowledged poems were written in youth, and, because he was born in 1550, Looney proposed that they were the prelude to his mature work and that this began in 1593 with *Venus and Adonis*. This theory is supported by the coincidence that Oxford's poems apparently ceased just before Shakespeare's work began to appear" (Leubering). This timing is crucial to the Oxfordian theory of authorship. For the Bayesian analysis, the first event E_1 is the concatenation of two actions: the first is that Shakespeare's works begin to appear in print and that Oxford's works cease appearing in print in the very same year, 1593. The probability of both instances occurring in that same year is the probability of E_1 occurring, $P(E_1)$, given that Shakespeare first starts publishing in 1593.

The first poetry attributable to Oxford was published in 1573. He died in 1604. If Oxford is the author writing under a pen name for a specific reason, such as a desire to distance himself from the political and satirical nature of the works, it is not surprising that his published poetry would cease as Shakespeare's begins to appear. Assuming this is the case, the probability of E_1 given that Oxford is the author could be as high as 1, $P(E_1|OA) = 1$.

But if Oxford is not the true author writing under a pseudonym, there is only about a 1 in 30 chance of these two instances being coincident by pure chance in the same year, 1593, $P(E_1|ONA) = 0.0333$, treating E_1 as a random event. In the absence of a specific and relevant historical reason why Oxford should permanently cease publishing precisely in 1593 at the age of 43 (i.e., he departs England never to return or the Queen issues an edict banning his poetry, etc.), the mathematical probability must account for the unlikely event that he should decide to completely stop publishing in the same year that Shakespeare begins to do so, given the 31-year period from the start of Oxford's publishing career to his death (1573–1604). To assume these occurrences are uncorrelated presents a steep probabilistic hurdle for event E_1.

However, Oxford's departure from publishing at the age of 43 could be treated as a form of early retirement. In a recent meta-analysis, the factors considered affecting early retirement were family obligations, organizational pressures, workplace time for retirement, job stress, job satisfaction, income, financial security, physical health, and mental health (Topa et al). Even if these factors could be appropriately evaluated in the case of Edward de Vere, it is not clear that modern statistical retirement models would apply to him. Alternatively, aggregating the ages of final publication for a series of poets who were Oxford's contemporaries yields an average age of last publication of 47.3, with a standard deviation of 10.6. Based on these statistics and assuming an underlying Gaussian distribution the probability of retirement from publishing at age 43 for the Earl of Oxford would be 6.9%, $P(E_1|ONA) = 0.069$ (Probability Calculator).

Incorporating the above probabilities of this timing evidence into the Bayesian inference calculation, the first iteration raises belief in Oxfordian authorship to a 43% probability, while Shakspere falls from 95% probability of authorship to 1− 0.433 = .567 or ~57%. (Equation 6)

$$P(OA|(E_1)) = \frac{P(OA)P(E_1|(OA))}{P(OA)\,P(E_1|(OA)) + P(ONA)\,P(E_1|(ONA))} = \frac{(0.05)(1)}{(0.05)(1) + (0.95)(0.069)} = 0.433$$

The first timing event is significant and substantially affects belief in these two authorship candidates, placing them on nearly equal footing. However, significant assumptions have gone into calculation of the E_1 probabilities. The reader is invited to make his or her own assumptions and calculate probabilities for this event that seems most logical and reasonable.

The second piece to the Bayesian framework involves the multi-year gap in the publication record of the plays occurring in 1604, contemporaneously with the death of Oxford, an event designated as E_2 for this analysis, consisting of the concatenation of the break in publication *and* the death of Oxford the same year. Oxford's recent biographer, Mark Anderson (2005), notes that from 1593 through 1603 the publication of new plays appeared at the rate of two per year and whenever an inferior or pirated text was published it was typically followed by a genuine text described on the title page as "newly augmented" or "corrected." After the publication of the Q1 and Q2 *Hamlet* in 1603 and 1604, no new plays were published until 1608. Anderson observes that, "After 1604, the 'newly correct[ing]' and 'augment[ing]' stops. Once again, the Shakespeare enterprise appears to have shut down" (Bethell 45–61).

To incorporate this result mathematically in the alternative that Oxford is not the author, note that Shakespeare's poems and plays were published over a 17-year period from 1593 to 1609. The odds that an extended multi-year gap, assuming one occurs at all, should randomly start in the record would be only about 1/15 to occur in any specific year. In the absence of some significant event in Shakspere's life that should shut down his production, such as chronic debilitating illness or leaving England altogether, the probability for this event must be assigned randomly over the period of Shakespeare's publishing career. If Oxford is not the true author of the works, then his death should be irrelevant to the gap in Shakespeare's publication record and there should be no correlation between the two events. The odds of this happening by chance in the same year of his death would be about P(E_2|ONA) = 0.06666. I know of no other way to reasonably calculate this probability, although the reader is invited to assign a probability to this event that seems most rational.

If Oxford is the real author, however, a disruption in publication would be expected upon his death and P(E_2|OA) =1. Integrating this result into the analysis and iterating yields only an 8% probability that Shakspere is the true author of the works, while belief in Oxford's authorship rises to 92% (Equation 7):

$$P(OA|(E_2+E_1)) = \frac{P(OA)P(E_2|(OA\ ))}{P(OA)P(E_2|(OA))+P(ONA)P(E_2|(ONA))} = \frac{(0.433)(1)}{(0.433)(1)+(0.567)(0.0666)} = 0.92$$

The probabilities of belief in the traditional authorship have now been reversed from the initial starting point.

Even if we assume that contemporary references to the author "Shakespeare" are intended to refer to Shakspere, a dubious assumption (see Chiljan, Wildenthal et al), these could still be allusions to Shakspere as a front man rather than as the true author—and is therefore not a relevant factor. Moreover, this issue is already considered in the initial probability assignment weighting the scholarly consensus in Shakspere's favor. It bears mentioning again that Henry Peacham was a contemporary of Shakespeare's who very clearly alludes to Oxfordian authorship in his book, *The Complete Gentleman*.

Now consider the analytical results. From the Dartmouth study, Oxford's poems match most closely to the work of Shakespeare among the three contemporary candidates considered. If Oxford is not the author, he would have only a 1/3 chance of matching closest statistically to the works of Shakespeare in any given metric, yet he clearly comes closest over the most popular alternative candidates Christopher Marlowe and Sir Francis Bacon in all three. The odds of this happening by chance are only 1/27 or 3.7%.

This analysis is similar to that of the controversial Monty Hall problem, which generated much debate among statisticians (Monty Hall was host of the popular TV game show Let's Make a Deal, on which contestants often participated in games of chance such as the one described here). Hall offers a contestant a choice of three doors. Behind one door is a very good prize, while the other two doors hide less desirable ones. After the contestant selects a door (but before it's opened), Hall opens one of the other two doors to reveal a lesser prize. He then asks if the contestant would like to switch his or her selection to the other unopened door. Contestants almost never choose to switch; most seem to believe that their odds of having selected the right door have now increased from 1 in 3 to 1 in 2.

However, the mathematics of probability indicates that they should always switch. The odds of selecting the best door in the first round are only 1 in 3. Put another way, the odds against choosing the best door are 2/3 so they are twice as likely to have chosen the wrong door in the first round. Once one of the lesser prizes is eliminated, switching doubles their chances of winning since they were originally 2/3 likely to have been wrong but become 2/3 likely to be right by switching to another door. Although there was considerable furor over the right strategy, the controversy was settled when Monte Carlo simulations conducted at Los Alamos confirmed that contestants who switched won 66.7 percent of the time while contestants who stuck with their first choice won only 33.3 percent of the time (Wikipedia, Monty Hall).

This is why Oxford is so unlikely to have been the closest statistical match in all three analytical categories in the event that Shakspere is the true author. As in the Monte Hall problem, his work is twice as likely not to be closest as to be closest statistically to the works of Shakespeare in any given text analytic, if none of the three candidates is the true author of Shakespeare's works. He therefore must beat long odds to match most closely in all three analytics if none of them are the author and the comparison is just random, $P(E_3|ONA) = P(E_4|ONA) = P(E_5|ONA) = 0.333$. If Oxford is the true author, then he would be expected to be closest statistically in every analytic and $P(E_3|OA) = P(E_4|OA) = P(E_5|OA) = 1$. Incorporating these probabilities into the Bayesian formula iteratively and performing the analysis as shown above increases belief in Oxford as the author to $P(OA) = .9968$ or 99.7%, while belief in Shakspere as the true author drops to 0.3%.

Lastly, Oxford was compared to eight of his contemporary authors in the text mining analysis and clustered closest to the earliest poems attributed to Shakespeare. There is only a 1/9 chance or a probability ~11% of this happening randomly if Oxford is not the true author of the works, $P(E_6|ONA) = 0.111$. In this case, Oxford would be no more likely to cluster next to *Venus and Adonis* and the *Rape of Lucrece* than any other author from the period. If Oxford is the real author writing under a new pen name, then his published poetry would be expected to cluster closest to the earliest works of Shakespeare and $P(E_6|OA) = 1$. Adding these factors into the analysis and performing the last iteration of the calculation, as demonstrated above, raises the final probability in our belief in Oxford as the true author to 99.96% while our belief in Shakspere's authorship declines to 0.04%.

Stated another way, based on the above sample analysis, the Earl of Oxford is over 2,790 times more likely to have authored the works attributed to William Shakespeare than William Shakspere of Stratford. Adding the analytical findings to the Bayesian inference calculation validates the result, as the probabilities derived from the analytical outcomes approximate those derived from the temporal correlations. A similar analysis applied to Marlowe would fail quickly as the text analytical results rule him out even in the unlikely circumstance that he somehow faked his death in 1593 in order to publish anonymously. Bacon is even more strongly ruled out by the statistical text-based analytic results.

Likewise, there are no additional factors that can be incorporated into the analysis in Shakspere's favor. No known letters, manuscripts, or publications under a different name are ascribed to him and therefore nothing to statistically compare. Only six signatures are known, all spelled differently, and none spelled "Shakespeare" or "Shake-speare". So little is known about his life that it is difficult to determine any valid temporal correlations with the publication of the works. For instance, he seems to have had no connection whatsoever with publication of the *Sonnets* in 1609, even though he was still alive at the time. His passing in 1616 appears to have gone by unnoticed by the literary community and his will contains no reference to plays, poetry, or manuscripts.

While the choice of a starting point for the Bayesian analysis is highly subjective, it bears mentioning again that the autobiographical nature of the plays and other literary and historical factors entitle Oxford to a non-zero starting point. Most particularly, as a contemporary, Henry Peacham's allusion in 1622 to the Earl of Oxford as the best poet of the Elizabethan Age over Edmund Spenser with no mention of Shakespeare justifies a ~5% starting point for Oxfordian authorship. The only body of work conceivably surpassing Spenser's *The Faerie Queene* is the portfolio of William Shakespeare. The limited corpus of Oxford, totaling 23 poems, would hardly qualify.

The Shakespearean actor and director Orson Welles once said, "I think Oxford wrote Shakespeare. If you don't agree, there are some awful funny coincidences you have to explain away" (Tynan). Bayesian analysis provides a systematic framework to evaluate these coincidences and other factors mathematically. The prime advantage of the framework is its flexibility. Factors can be added, subtracted or modified as desired or as new information becomes available. For instance, if a 1% starting point for Oxford's candidacy is used in the above example analysis, the end probability for belief in Oxfordian authorship would be 99.8% or 546 times more likely that Oxford is the true author of the works of Shakespeare than Shakspere. While even statisticians may sometimes disagree on the calculation of probabilities, the reader is encouraged to choose his or her own starting point, factors, and probabilities based on facts and assumptions that seem most reasonable to apply to the analysis of the authorship question.

Conclusions

This analysis does not rely on autobiographical parallels in the plays, educational backgrounds, or hypotheses. It depends only on the historical timing of events and text mining analytics. The only historical information is used to assign a starting point of belief in the two leading alternatives of authorship, a starting point weighted heavily in favor of Shakspere as the author bowing to the preponderance of scholarly opinion.

No single event or analytic proves the case for the Earl of Oxford. Rather, it is the combination of low probability events and analytics, taken in total, that leads to a final probability or likelihood for his authorship.

However, this is not all. With stunning clarity, the results of 21st Century Machine Learning and Text Mining Analytics are consistent with the views of the subject matter experts, the doubting authors of the 19th Century. In this case, historians are not subject matter experts; writers are. The opinions of each of the three great writers from the 19th Century who doubted Stratfordian authorship all agree with the analytical results from the 21st Century. Mark Twain thought the true author was a lawyer. Modern analytics are consistent with this view. Walt Whitman thought the true author of Shakespeare's canon was an Earl. The analytics are consistent with this belief. Henry James doubted that either Shakspere or Francis Bacon wrote the plays. Text analytics are consistent with neither man being the author. Even in hindsight, the accuracy of their beliefs is astonishing. This agreement between analytics and subject matter experts represents the ultimate standard of Data Science. When the judgment of experts and the results of modern analytics converge with this level of precision, objective truth is revealed.

In the absence of an authenticated original Shakespearean manuscript it may prove impossible to determine the true creator of Shakespeare's works with a consensus of certainty. However, modern text mining and machine learning techniques can shed light on the authorship question. While the works stand on their own, uncertainty as to the true author denudes the poems and plays of historical context. As the above analysis illustrates, there is a not insignificant probability that the identity of the greatest artist of all time has been lost to our collective consciousness.

Works Cited

Anderson, Mark. *Shakespeare By Another Name: The Biography of Edward de Vere, Earl of Oxford, The Man Who Was Shakespeare*. Kindle Edition, 2011.

Awati, Kailash. "A Gentle Introduction to Cluster Analysis Using R," 2015. https://eight2late.wordpress.com/2015/07/22/a-gentle-introduction-to-cluster-analysis-using-r/.

Bethell, Tom. "The Case for Oxford (and Reply)". *Atlantic Monthly*. October 1991, 268:4, 45–61, 74–78. Retrieved December 16, 2010.

Blakemore, Bill (October 14, 2011). " 'Anonymous': Was Shakespeare A Fraud?". *ABC News*. Retrieved August 26, 2012.

Encyclopedia Britannica. Entry on Bayesian Analysis. https://www.britannica.com/science/Bayesian-analysis?utm_campaign=b-extension&utm_medium=chrome&utm_source=ebinsights&utm_content=Bayesian%20 analysis.

Fienberg, S. E., and M.J. Schervish. "The Relevance of Bayesian Inference for the Presentation of Statistical Evidence and for Legal-Decision-Making," *Boston University Law Review*, 66, 771–88, 1986.

Kraft, Stephanie A, Devan Duenas, Benjamin S. Wilfond, and Katrina Goddard. "The evolving landscape of expanded carrier screening: challenges and opportunities". *Genetics in Medicine*. 21 (4): 790–797. doi:10.1038/s41436-018-0273-4. PMC 6752283. PMID 30245516.

Leubering, J.E. Entry on "Edward DeVere, 17th Earle of Oxford," *Encyclopedia Britannica*, current online edition.

Londré, Felicia. "Hamlet as Autobiography," *Bulletin of the Faculty of Letters*, Hosei University (Tokyo, Japan), No. 39 (1993). Republished on the Shakespeare Oxford Fellowship website.

Nelson, Alan H. *Monstrous Adversary*: *The Life of Edward De Vere, 17th Earl of Oxford*. Liverpool University Press, 2003.

Niederkorn, William S. "The Shakespeare Code, and Other Fanciful Ideas from the Traditional Camp", *The New York Times*, August 30, 2005.

Probability Calculator. https://www.calculator.net/probability-calculator.html?val2mean=47.3&val2deviation=10.578&val2lb=42&val2rb=44&calctype=normal&x=53&y=18#probofnd.

Sabo, Tom. https://blogs.sas.com/content/sascom/2018/05/29/5-remarkable-uses-for-text-analytics/.

Seletsky, Oleg, Tiger Huang, and William Henderson-Frost. "The Shakespeare Authorship Question," Dartmouth College, December 12, 2007, 1–13. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.7146&rep=rep1&type=pdf.

Sturrock, Peter. *Shakespeare, A Scientific Approach to the Authorship Question.* Palo Alto, CA; Exoscience Publishing, 2013. https://shakespeareoxfordfellowship.org/aka-shakespeare-scientific-approach-authorship-question/.

Topa, Gabriela, Marco Depolo, and Carlos-Maria Alcover, "Early Retirement: A Meta-Analysis of Its Antecedent and Subsequent Correlates" *Frontiers in Psychology*, 4 January 2018. doi: 10.3389/fpsyg.2017.02157.

Tynan, Kenneth. *Personna Grata*. London: Putnam, 1954.

Wikipedia. Entry on "Monty Hall Problem" https://en.wikipedia.org/wiki/Monty_Hall_problem.

Williams, Janet. August 6, 2018. https://www.promptcloud.com/blog/9-best-examples-of-text-mining-analysis/.