[Title Slide]

Imagine for a moment that you are in the midst of a project in which you need to compare two large, complex digitized texts. You are unfamiliar with these texts, having never read them before. Furthermore, the metadata about these documents, which might have given you some clue as to their contents is not terribly revealing. Short of reading both of these texts in their entirety (a choice that may or may not be feasible) what will you do?

Such a scenario may seem uncommon in a group such as this one, but it often confronts first time users of digital repositories. While the collections of digitized texts these repositories contain are a god-send for researchers who would otherwise have to travel long distances in order to physically access them, they can also provide rather daunting challenges for the layperson researcher.

Why are they daunted? Consider things for a moment from their perspective. They are confronted with massive collections of texts, many of which will be unfamiliar or unknown to them. Furthermore, despite the diligent, if not heroic efforts of those who process and maintain these collections to ensure that each piece has been indexed with unique and painstaking metadata, such information may yet be insufficient for the layperson or first time user to make a judgment on the content, appropriateness or usefulness of a text.

With these problems in mind, what tools are available to these users in order to assist them in taking these massive amounts of textual information and rendering them into something more easily digestible?

[Slide 2]

Edward Tufte, a noted statistician and graphic designer, provides one potential solution to this problem. In his 1983 work *The Visual Display of Quantitative Information* Tufte asserted that, "often the most effective way to describe, explore and summarize a set of numbers-even a very large set-is to look at pictures of those numbers."

Now Tufte of course was talking about statistical and numerical data, the backbone of most data visualization. Such data is easily quantified and visualized. How can such principles be applied to digitized text? How do we "look at pictures" of words?

The answer to this question requires a specific type of data visualization: The Word Cloud.

Word Clouds are a relatively popular form of data visualization. They are commonly shared in social media, and frequently pop up in the news after a the delivery of a major speech, or the release of a large document as a way of providing a brief, but visually striking summary of its contents. They've also been used for a variety of tasks within the Digital Humanities.

[Slide 3]

For example, Google has utilized one "novel" (pun intended) approach to word clouds in the digital humanities in the bibliographic records on Google Books. Many of these records include a word cloud containing common terms and phrases in the book. Clicking on any one of these terms takes you to every instance of that term or phrase in the digitized text.

[Slide 4]

Word Clouds have also already seen some use in textual comparison and analysis. Brad Borevitz, on his website, *State of the Union*, has taken the text of every Presidential address to Congress, from President Washington's 1790 speech to the one President Obama gave a month ago, and broken them down into complex, interactive word clouds in order to compare and contrast them. His research has come up with some fascinating data about the ways our leaders have used certain key terms over the past two centuries, such as the marked increase of the use of the term "freedom" in contrast to the relative decline of the term "justice."

[Slide 5]

While all of these clouds were created using different software, their creation was based on the same principles: a machine readable text was inputted into software that broke it down to its basic lexical components, (or words). The individual words in the text were sorted, and assigned values according to the frequency of times they were used within the text. Commonly utilized words, such as the, a, an, and etc… were filtered out of the equation and discarded. Finally the remaining words were randomly arranged into a "cloud" in which the relative sizes of the words depend on their frequency of use. With results like this:

[Slide 6 (the Bill of Rights), Slide 7 (The Gettysburg Address]

As you can see, the resulting clouds make the original text generally unrecognizable, thus one cannot claim to have read the Bill of Rights or Gettysburg Address by having read its Word Cloud. However, by viewing the pattern of word size, one can gain an idea of the major concepts covered by a text.

Furthermore, as I briefly alluded to in mentioning Brad Borevitz's work, if more than one text is involved, multiple clouds can be utilized in order to compare the differences between the texts.

[Slide 8]

For example: between 1854 and 1861, Kansas Territory generated four different proposed State Constitutions. Using word clouds, we can compare and contrast the ideals argued for in these documents. In particular I would like to draw your attention to some key differences between the Topeka Constitution on the one hand, and the Lecompton Constitution on the other:

The Topeka Constitution repeatedly uses the term "Assembly" or "General Assembly" to refer to the legislative body in the state. This contrasts sharply with the Lecompton document, which heavily emphasizes the term "Legislature". This difference in terms suggests a potential difference in philosophy regarding the mission and makeup of these bodies. While the term "law" appears in equal size in both documents, the term "Governor" appears slightly larger in the Lecompton cloud, meaning that it is mentioned with greater frequency, and hinting at a potentially larger role for the Governor under this particular constitution.

[Slide 9]

Of course, the primary thing that sets the Lecompton Constitution apart from the other three Constitutions generated by the territorial conventions, is it's protection of the institution of slavery, with the ultimately unsuccessful goal of bringing Kansas into the Union as a Slave State. Hence, it is unsurprising that the Lecompton cloud is the only one to have the terms "Slave" "Slaves" or "Slavery" appearing within. However, what is striking is their relative size within the cloud, none of these words are particularly large, meaning none of them occur with great frequency within this document. Yet their very presence within this document is what sets it apart from the proposed constitutions that both preceded and succeeded it. Such a distinction could catch the attention of a sharp-eyed researcher, even if they were unfamiliar with Kansas' troubled antebellum history.

So, how does this work?

[Slide 10, link to wordle.net for demonstration]
(Poe text: www.bartleby.com/195/10.html)

[Slide 11]

The benefits of creating word clouds are straightforward: Wordle is free and it is relatively easy to use. Furthermore, one of the byproducts of word cloud production is the creation of a collection of beautiful textual portraits of some of the key texts in your collection that can then be used for outreach or to highlight your holdings and generate patron interest.

On the other hand, the use of word clouds also has a number of drawbacks. First off, in order for a text to be useable, it must be machine readable. Hence a scanned copy of a census record written in longhand won't work, unless it has been transcribed into machine readable format beforehand. Furthermore, the system doesn't remove synonyms, so an overarching theme can be lost as instead of one gigantic word depicting the whole idea, it becomes lost in a few words of roughly equal size. Finally, the filter for common words, doesn't work well with archaic texts, such as biblical language from the King James Version, meaning that you have to manually remove the "thee" and "thou" "thy" and "ye".

With these limitations in mind, I believe it's fair to say that the use of word clouds on digitized text is probably not going to replace close reading anytime soon. However, that really isn't the intention. Going back to that hypothetical first time user I mentioned at the beginning of this presentation, suppose you suggest that they take the digitized text they're struggling with and feed it through Wordle? Then have them check out the results: Does the textual portrait that was generated contain terms or ideas they're interested in? If they're comparing two texts, are there noteworthy differences between the major terms in both clouds? Perhaps a portrait of the text will make it seem a little less daunting?

[Slide 12]

In the end, despite their limitations, word clouds provide us with one of a variety of tools for analyzing unfamiliar digitized texts by breaking them down to their basic lexical components and then illustrating which of those components are the most prominent. While word clouds are by no means a replacement for the close

reading and analysis necessary to truly become acquainted with a text, these textual portraits can provide users with a useful first-glance tool when searching or comparing unfamiliar digital texts with minimal metadata.

[Closing Slide]