The title of this talk isn’t actually an error, or just a placeholder that I forgot to get rid of when I was making my slides. We’ll come back to the reason for it near the end of this talk.

Hi, I’m Paige, and I work in Richter Library, in an area called Digital Humanities. That area often involves working with various sorts of data. As a librarian, sometimes I do my own research, but a lot of the time, I help other people with their research; helping them figure out what kind of data they have, or can create, and what to do with it. Several months ago, someone asked me if I was a data scientist, and I was completely caught off guard. And they said "well, you work with data, right?" And I replied "yeah, but..." I really didn’t know what to say. And that, too, is something that we’ll come back to later.

I say that I work with data, but in some ways, it feels more accurate to say that I work with various types of mess. Sometimes I’m trying to clean up the mess, or help other people figure out how to clean it up. And sometimes I want them to appreciate the messiness, and understand that it’s not going to go away, and that they shouldn’t try to make it go away.

Today I’m going to talk about mess in the context of textual data. Textual data can mean data that’s made up of words. It can also mean data that’s contained within texts. A text isn’t just something you send on your phone. It could be a newspaper, or a book, or letters, emails, tweets – and textual data sometimes contains other sorts of info, like data about places, or numbers. There are a lot of possibilities, even though I’m only going to have time to talk about a couple today.

In data science, you may encounter the terms "unstructured" and "structured" data. I’m going to show you a few examples to help explain what these are – but the point I want to make right now is that depending how much you structure your data, you can ask different questions of it. Although you might hear people talk about unstructured, or raw, data, it’s almost always structured at least a little in order to make it compatible with whatever computer programs you’re using to explore it.

Let me start by showing you some minimally structured literary data. Working with literary data might involve just taking the text of a novel, or several novels, or poems, or song lyrics, and putting it into an analysis tool. One
of my favorites is called AntConc, and it’s one of my preferred tools because it’s free, works on multiple platforms, and has plenty of tutorials.²

A while ago I bought ebooks for the Harry Potter series and converted them to plain text – one text file per book, so that when I analyze it, I can see which book certain words came from. This is an example of minimally structured data, since I didn’t break the novels into separate files for each chapter or break it down even further than that.

When you’re analyzing (sometimes you’ll hear this called text mining or distant reading) a book, one place to start is seeing which words are the most frequently used. It can also be interesting to see which words appear most frequently right near each other. These words are called collocates, which just means that they’re located together. With AntConc, I can see collocates for any word that I search for, and see how frequently they’re used in the text, and see how statistically unique the pairing is. I’m about to show you the collocates for ‘he’ and ‘she’ in the Harry Potter series; and to give you an example of the statistical uniqueness, one of the most common words is ‘is’ – like ‘he is’ and ‘she is.’ But ‘is’ isn’t statistically unique, because it’s used with both genders. So you’re about to see the words that are paired with “she”, but are less likely to be paired with “he”, and vice-versa.

‘She’

herself, added, says, smiled, snapped, screamed, shrieked, thinks, crying, liked, pretty, waved, impatiently, trembling, coldly, replied, upset, marched, shocked, peered, hi, forgot, baby, wore

‘He’

realized, remembered, wished, named, imagining, doubted, daring, revenge, experience, sworn, trusts

When I look at these groups of words, the differences between them stand out right away. The ‘he’ words are words that I associate with adventure and excitement – they’re active words, words that sound like they ought to be part of a story that I’d want to read, or to be part of. The ‘she’ collocates, on the other hand, are words that I associate with being on the sidelines, or with being kind of a stereotypical girl – words that either put girls on the sidelines, or make them into something for boys to look at. It says everything to me that ‘she’ is associated with forgot, while ‘he’ is associated with remembered. I don’t think that this was intentional on J.K. Rowling’s part when she wrote the series – but this analysis supports arguments that people have made about how several of our stereotypes (about gender, among other things), are so deeply ingrained that even well-intentioned people unintentionally reinforce them.

One analysis of ‘she’ and ‘he’ is only the beginning of developing research and an argument. If I wanted to develop this further, I might want to also look at the collocates for ‘his’ and ‘her’ as well.³ Or I might want to structure this data more by breaking down my text files into one per chapter. That would make it possible for me to see whether each of these words occur throughout the novel – do they spike in frequency at certain dramatic points or are they consistently distributed? There are other ways I could structure the data further, too: I could encode it in order to show how each word and sentence is associated with character, gender, Muggle or witch/wizard, and so on. Structuring the data that carefully would

³I haven’t delved down further into this analysis yet, but if and when I do, it will be interesting to see how the ‘she’ collocates get linked to Hermione vs. other female characters. I expect to see that they cluster around Hermione more earlier, and less later. I also expect that if I investigated, I would find that ‘baby’ is strongly associated with Lily Potter – so a related question would be whether any of these other words are strongly linked to one female character in particular.
also let me track which words belong to the narrator, who doesn’t have a name themself, but who certainly has opinions about various people and things!

I haven’t done that, and I’m not sure I’m interested enough to put in the time that structuring the data that intensely would take. But I’m aware of another data researcher who has put a little more effort into working with the Harry Potter books, so let me show you what he’s done.

Skyler Johnson was interested in how many spells were cast throughout the whole Harry Potter series, and being able to show where in the series they were cast. To do this, he needed: a snippet of text from every time in the books when someone casts a spell; information about which book that text snippet came from; and information about what position in the series each text snippet held (so that the visualization would show which spell was cast first, and which was cast last).

I suppose that Johnson could have opened up a spreadsheet and labeled it "text snippet," "book number," and "position", and started reading and taking notes. But that would have taken a long time; and you can actually get most of this information quite easily with a tool like AntConc. What you need, though, to do that, is a list of all the spells that someone casts through the whole series. If you have that list, then you can use it in order to retrieve all those text snippets, and the info about which book they come from. The point that I want to make right now is how powerful that list is for being able to answer questions about spell casting in the whole series.

If you can define something by making a list of all its features, you can use that list to detect its presence, and see where it exists – or where it doesn’t exist. In mathematics and data work, we’d call this a set. Sometimes I work with sets; and sometimes the sets I work with are part of what’s called a controlled vocabulary. Both sets and controlled vocabularies can be used to define things in various contexts, and we’ll come back to why that’s important later.

You can use text mining with literary data, but you can do this sort of analysis on all sorts of texts. Let me give you another example – it’s older, but I’m using it today because it makes a good parallel to what we’ve been looking at with Harry Potter text data; and because it’ll help me show you how that data work can go from clean to messy very quickly.

Last fall I was collecting data from the presidential campaign about the speeches made by Hillary Clinton and Donald Trump – the text of what they said when they appeared. I was also able to get hold of the text of each of the three presidential debates, structured so that it was broken down by speaker. I used another tool, called Voyant, to compare what Clinton and Trump said during each of those debates, and which words were comparatively distinctive, i.e., more common from one speaker than the other. The distinctiveness in this case is similar to what I was looking at for when I examined the collocates for 'he' and 'she' in Harry Potter.

**First presidential debate**

Clinton (distinctive words): information, working, sure, proposed, justice

Trump (distinctive words): clinton, leaving, agree, she’s, wrong

**Second presidential debate**

Clinton: court, supreme, worked, try, start

Trump: she’s, hillary, disaster, bad, inner

**Third presidential debate**

Clinton: donald, clear, security, stand, undocumented

Trump: she’s, hillary, bad, iran, russia
You can see a definite pattern here: Clinton was talking about various issues, and Trump was mostly talking about Clinton. That only starts to shift in the third debate.\footnote{This analysis was performed with Geoffrey Rockwell and Stefan Sinclair’s \texttt{Voyant}; and I am also grateful to Stefan for the script that facilitated splitting the debate text according to the speaker.}

I started putting this dataset together because a professor came to me and asked about using text mining to answer the question “What kind of understanding of history does Donald Trump have?” This professor was used to making arguments by reading texts the old-fashioned way, with his own eyes, and thinking about them, and making arguments.\footnote{(A method which I’m quite fond of myself, lest anyone think I’m suggesting that text mining and distant reading are the sole future of analysis.)} He wanted to know whether it was possible to use data analysis to answer it differently. We talked about the fact that using text or data mining would allow him to think about this question on a larger number of texts than if he were just reading speeches one at a time – and he might be able to do comparisons by looking at not just Donald Trump’s speeches, but Barack Obama’s, and George W. Bush’s, and...well, as many presidents as we could find speeches for. But there was a more difficult challenge that we had to deal with. How do we define when Trump is talking or thinking about American history?\footnote{Though indeed, the question of Donald Trump’s sense of international history is certainly relevant.}

Now, you could start by taking the dataset and searching for "history," and you could go a step further and see what the collocates are.

\textbf{Frequent and statistically significant collocates for ‘history’ in a selection of Trump and Clinton speeches}

\begin{itemize}
  \item Clinton: our, us, shown, racial, must
  \item Trump: in, our, country, seen, worst
\end{itemize}

It’s interesting, but not terribly illuminating. It doesn’t feel like even the start of a conclusive answer to the question. But part of working with data is figuring out what questions you have to ask to move towards a larger answer, and in this case, there are a couple of ways to move forward. One would be to get a larger set of texts, including speeches from other presidents, or presidential candidates, so that we can try to establish some sort of basis for comparison – that is, what’s normal for a president’s sense of history? The collocates might be illuminating. But the problem is really that people don’t always use the word "history" when they’re talking or thinking about history.

A second option for us would be similar to what Skyler Johnson did with the Harry Potter spells question. We could make a list of all the words that might count as an indicator that the president was talking or thinking about history; and then we could use that set as a comparison against the various speeches. And that’s a much messier question, because I think that depending on who you asked, you would get a really different list, because the question is “what counts as American history?”

On some level, everything counts, because historians are interested in all sorts of things – even how people shopped for groceries. But to answer this history professor’s question about what Donald Trump’s understanding of American history is, we would probably want to start by making a list of important moments in American history. I think that a lot of people’s lists would involve various wars: the Revolutionary War, the Civil War, the two World Wars, the Vietnam War, and so on. But my list, and plenty of other people’s lists, might involve various constitutional amendments, such as the 19th amendment granting women the right to vote; Supreme Court decisions like Loving v. Virginia, invalidating the laws that prohibited interracial marriage; or Brown vs. the Board of Education, desegregating schools; or Roe v. Wade, legalizing abortion; or the HIV and AIDS crisis that first peaked in the 1990s. Or legalizing marriage equality just a couple years
This talk doesn’t have a name • July 20, 2017

ago. That’s just a tiny start.

To answer this question, you would need to create that list. You might make it a set; you might structure it as a controlled vocabulary. You might want to structure it so that you could designate types of historical moments: wars, court verdicts, natural disasters, and so on. While a text file of a novel is mostly unstructured data, a set or a controlled vocabulary is highly structured data. Everything in it has been carefully chosen.

As soon as you started working to create that set/vocabulary, your data work would get incredibly messy. You would have some people arguing that certain moments don’t count as sufficiently important to be considered capital-A, capital-H American History. Another problem, no less important, is that some things are harder to identify with a single name or phrase. Some things have different names, depending on who the community is who is talking about them. And some moments, I would say, don’t have names quite yet, for all sorts of reasons.

Working with data can get messy really quickly. And that mess isn’t just something to be cleaned up – it’s people’s lives. People’s lives can be genuinely complicated in so many different ways, and figuring out how to handle that mess is a vital part of working with data. When people don’t consider the mess; or when they try to shove it out of the way so that it doesn’t complicate their analysis, then there’s a good chance that they’re not representing the reality that people are living.

Here’s one really common and classic example of not accurately representing people’s realities that I’m sure you’ve all seen before, in one form or another:

<table>
<thead>
<tr>
<th>Race/Ethnicity:</th>
</tr>
</thead>
<tbody>
<tr>
<td>African American/African/Black/Caribbean</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
</tr>
<tr>
<td>Caucasian</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
</tr>
<tr>
<td>Native American</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td>Prefer not to answer</td>
</tr>
</tbody>
</table>

This question asks people about their ethnic background, but leaves out some groups (such as near or middle eastern heritage) entirely; and doesn’t allow people to identify themselves as having multiple ethnicities, unless they want to lump themselves into that vague and mysterious "Other" category. While the Harry Potter spells list was long and inclusive, this list is really short and incredibly exclusive. I’m showing it to you because I want to make the point that textual data, in various forms, is all over the place, in areas where it’s not always very obvious, but where it’s still having an impact. There are so many examples of this, and I’m sorry I only have time to give you a couple, but I wanted to plant this seed in your heads.

I’ve been talking about messy data, and messy situations. I know that mess is often used as a word that has bad associations – people may tell you that something that’s messy has to be cleaned up before it’s worth anything. And I want to push back hard against that assumption. Sometimes there are things about data that are easy to clean up – typos in a scan of a novel, for instance. A lot of the time, however, it’s not that simple. Let me give you a few examples of that:

One thing we need to consider is where data comes from, and how accessible it is. When I gathered the Clinton and Trump speeches, I was able to get a few of them from the UCSB Presidency Project Archive, and

---

7I should say, ostensibly, everything in a controlled vocabulary has been carefully chosen. In fact, I suspect that the care that is used in the selection process does not extend to consideration of the impact of the choices made. Greater care in that regard would almost certainly take more time and more effort; and currently, my sense is that in situations where controlled vocabularies are being created, expediency is the greater priority.

8This is just one instance out of many problematic survey questions about racial and ethnic heritage.
others, I needed to collect directly from the candidates’ websites, or from news sites. I needed to be careful because sometimes the candidates would release their planned texts to the news sites in advance – but then go off script. Trump, in particular, was known to do this, and I wanted to make sure that my data captured what he had said. That meant looking for transcripts of the talks as actually given, and sometimes those were a little tricky to find. Still, I think it took me three or four hours – so it was relatively straightforward.

However, Trump and Clinton are major public figures, so it would be surprising if I couldn’t track down their words. And outside of people who are in major political roles, there are several factors that can shape the question of whether material will be available to turn into data that’s machine-readable; data that computers can work with.

Sometimes those factors involve race, so that white people’s data is more likely to be available.

Sometimes those factors involve gender, so that men’s data is more likely to be available.

Those factors can involve sexuality, so that straight people’s data is more likely to be available.

They can also involve wealth, so that wealthier people’s data is more likely to be available.

Or those factors can involve status, and affect the availability of data related to people who are incarcerated or who are undocumented immigrants.

I want to say very clearly that this is wrong, and that various people are working to improve the situation; but I want to be realistic about acknowledging that this happens, and is a fundamental aspect of working with data.

Each of these factors can intersect – people from less wealthy communities are more likely to have materials in formats that don’t survive as well over time, or isn’t printed on sturdy paper, or is printed with ink that runs, or on cheap equipment, so that the paper doesn’t scan well. Or, if the data is made, and is made machine-readable structured data, as in a controlled vocabulary, it’s more likely to have been created not by the community itself, but by an outside group (in many cases, white people.) Unfortunately, there have been good examples of these types of data problems in libraries as well as in other contexts.

Those examples involve library catalog subject headings – data which altogether, make up a really huge controlled vocabulary that shapes people’s abilities to find books and resources. Those subject headings, in some ways, try to describe the whole world, in the sense that there are books about almost every subject in the whole world. But these subject headings – these pieces of data, were created by individuals, and small groups of individuals – almost always very privileged (white) people. And the data that they created to organize libraries reflected their biases, and the biases of their times.

One example of this involves people who identify as LGBTQ. To the degree that books about LGBTQ people were available at all, they were classified in a subject heading that labeled them as abnormal, that treated their sexuality as a disease. In 1946, content related to queer sexuality was given the subject heading “Homosexual” – an acknowledgement that this sexuality existed; but it was still classified by the Library of Congress Subject Headings as an abnormality until 1972. It took more time, and more advocacy, for the subject heading “homo-

---

9My collection of speeches is available at my [GitHub repository](https://github.com), though at this point, I really need to go back and update it to reflect more recent speeches.

10One excellent discussion of the history of queer-related subject headings is Emily Drabinski’s Queering the Catalog: Queer Theory and the Politics of Correction, The Library Quarterly: Information, Community, Policy, Vol. 83, No. 2 (April 2013), pp. 94-111. And for a quick discussion of Library of Congress data intended for a lay audience, see Melissa Levine’s recent post at The Conversation.
sexual” to be changed to names that LGBTQ people themselves prefer, like “lesbian”, “gay”, “bisexual,” “transgender,” etc. There’s still more room for improvement.

To give you another example: for a long time, the Haitian religion Vodoun was classified in library catalogs under the term "voodooism." Voodoo is a word that became prevalent via the US military occupation of Haiti in the early 20th century. It was a term used by white Marines who were stationed there; it is associated with devil worship; and it has nothing to do with Haitian Vodoun. One of my colleagues here at the University of Miami, Kate Ramsey, who teaches in the History Department, was involved in advocating for the Library of Congress to change this, and start classifying materials related to Vodoun under the subject heading "Vodou," because "voodoo" is a pejorative and misleading term. I’m happy to tell you that the Library of Congress listened to the people who were advocating for this change, and that it was made in 2013.

I wish I had more time to go into more depth with you about either of these examples, or just the way that data is structured throughout so many aspects of our lives. But I thought that it was worth telling you about these examples in order to show you just a small glimpse of some of the other work and history around data. Data is important not just because we can analyze it for research purposes, but because it affects how various aspects of our worlds are structured. This is just one reason why being alert about messiness and data is so important; and it’s why I told you at the beginning of this talk that sometimes my job is to help people appreciate messiness, rather than try to clean it up; or to encourage them that creating data is worthwhile, even if it’s going to be messy data.

One woman whose work I’ve learned from in this regard is artist and data researcher Mimi Onuoha, who writes that:

For every dataset where there’s an impetus for someone not to collect, there’s a group of people who would benefit from its presence. More data doesn’t always mean better answers, but in cases where data is used as the end-all tool of proof or a definitive measure for change, then it’s clear that lacking it can be a serious structural disadvantage.

Data can have a tremendously positive impact. I want to acknowledge that it’s also possible for data to be used against people, and that’s an important consideration as well; and it’s a reason that we need smart people working with data who can think alertly about how we’re collecting it, how we’re shaping it, and its impact on people’s lives.

In closing, I want to leave you with four questions to think about:

• Who does your data impact? (How?)
• What data might be worth trying to create? (Why?)
• How can you develop a plan to work with your data?
• How will you advocate for and about this data?

As Mimi Onuoha has argued, there are a lot of instances where people’s realities are not represented, and that can have serious consequences for how local and federal money is spent, what services are provided; for understanding what people experience. But there isn’t always an obvious way forward with these questions. Creating and working with data can

---

11For a shorter summary of Vodoun and voodoo in relation to LCSH, see this discussion; however, I highly recommend Kate Ramsey’s article From ‘Voodooism’ to ‘Vodou’: Changing a US Library of Congress Subject Heading: Journal of Haitian Studies, Vol. 18, No. 2, Special Issue on Vodou and Créolité (Fall 2012), pp. 14-25 for a more in-depth discussion.

12Mimi Onuoha, Missing Datasets, November 15, 2015. I was reminded of Onuoha’s work by Jer Thorp’s recent essay [You Say Data, I Say System], which, needless to say, I also recommend.
be long, hard, and exhausting work. I want you to know that specifically because I might look nice today, in my "I'm giving a talk" outfit, but sometimes when I work with data, I swear, I let my hair frizz, and my apartment gets really sloppy, because I'm trying to figure out a problem, and it's taking all my time. Working with data doesn't always mean being good at it.

At the start of this talk, I said that I wasn't sure whether I was a data scientist. One of the reasons that I'm not sure about that is that I don't actually spend much time analyzing data – and most of the time, I'm helping other people think about and plan their projects, rather than working on my own. But it's also that I spend less time analyzing data or writing computer programs to do stuff with it than I spend helping people think about data before they even work with it – and help them understand, in depth, what that data could mean – not because it's been analyzed, but just because it exists. I'm not sure that that's being a data scientist, exactly – though I think it's important work, and it's work that any of you could learn to do.

Being a data scientist is a hot career, according to various magazines and newspapers and websites – but there's a lot of other important work that involves data that might be data science – or might not. One last example to drive this home. Earlier I said that some realities are challenging as data because they're difficult to name. One of those realities involves the #blacklivesmatter movement that was founded by Alicia Garza, Patrice Culors, and Opal Tometi. Their work to start that movement and grow it to where it is now was important – in so many ways – but it was important from a data perspective, because the hashtag that they created helped to make the problem of police violence more visible to broader audiences than it was before. And their work wasn't computer programming; it was organizing and activism.

To come back to the title of this talk, I think that some of those roles for people who work with data, work with creating data in powerful and valuable ways don't have names yet. And I want to encourage all of you, if you keep working with data: if you think that there is work that needs to be done, that is worth doing, to trust your instincts and try to do it even when people push back and things are hard. I want to encourage you to find friends and work together with them, because that can make the hard parts more possible to get through, even when they're exhausting. Because your instincts, and your questions, and your arguments are important to the work that will happen with data now, and going forward.

References


