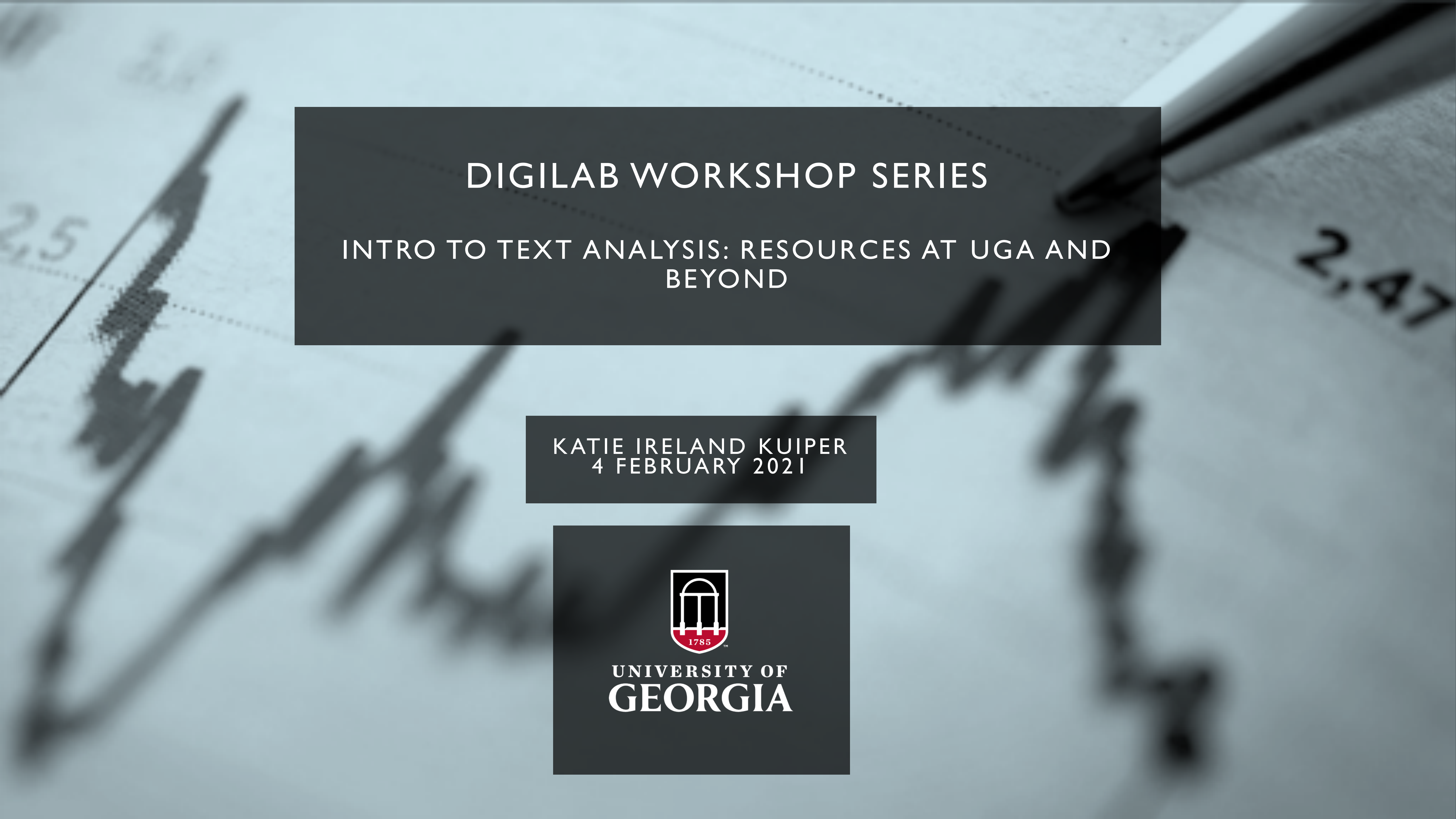


DIGILAB WORKSHOP SERIES

INTRO TO TEXT ANALYSIS: RESOURCES AT UGA AND BEYOND

KATIE IRELAND KUIPER
4 FEBRUARY 2021

1785

UNIVERSITY OF
GEORGIA

WHAT IS A TEXT?

- any coherent stretch of language (R. Nordquist)
- a piece of written or spoken material in its primary form
- a text is any object that can be ’read’ ; a coherent set of signs that transmits a message (Wikipedia)
- the main body of a piece of language
- the written words in a book, on the internet, etc (Cambridge Dictionary)

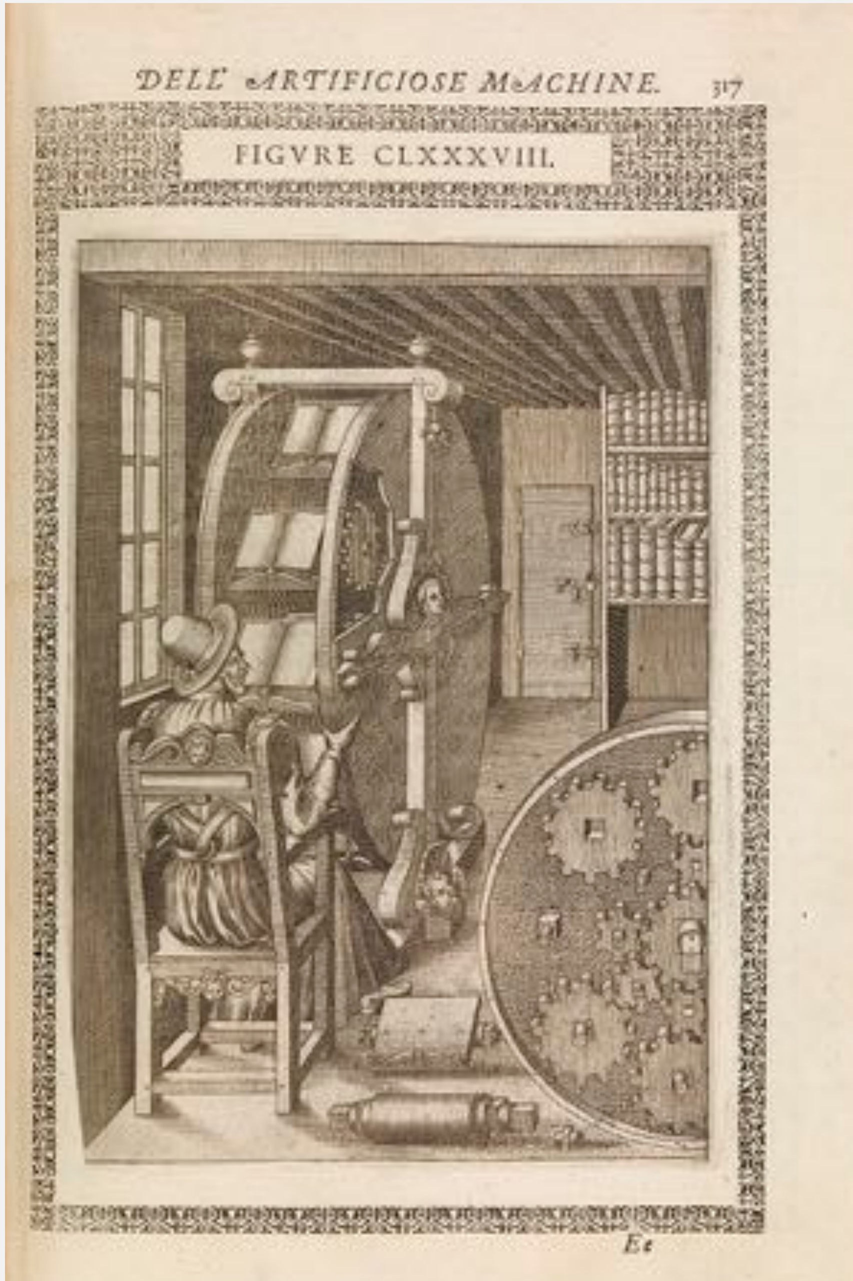
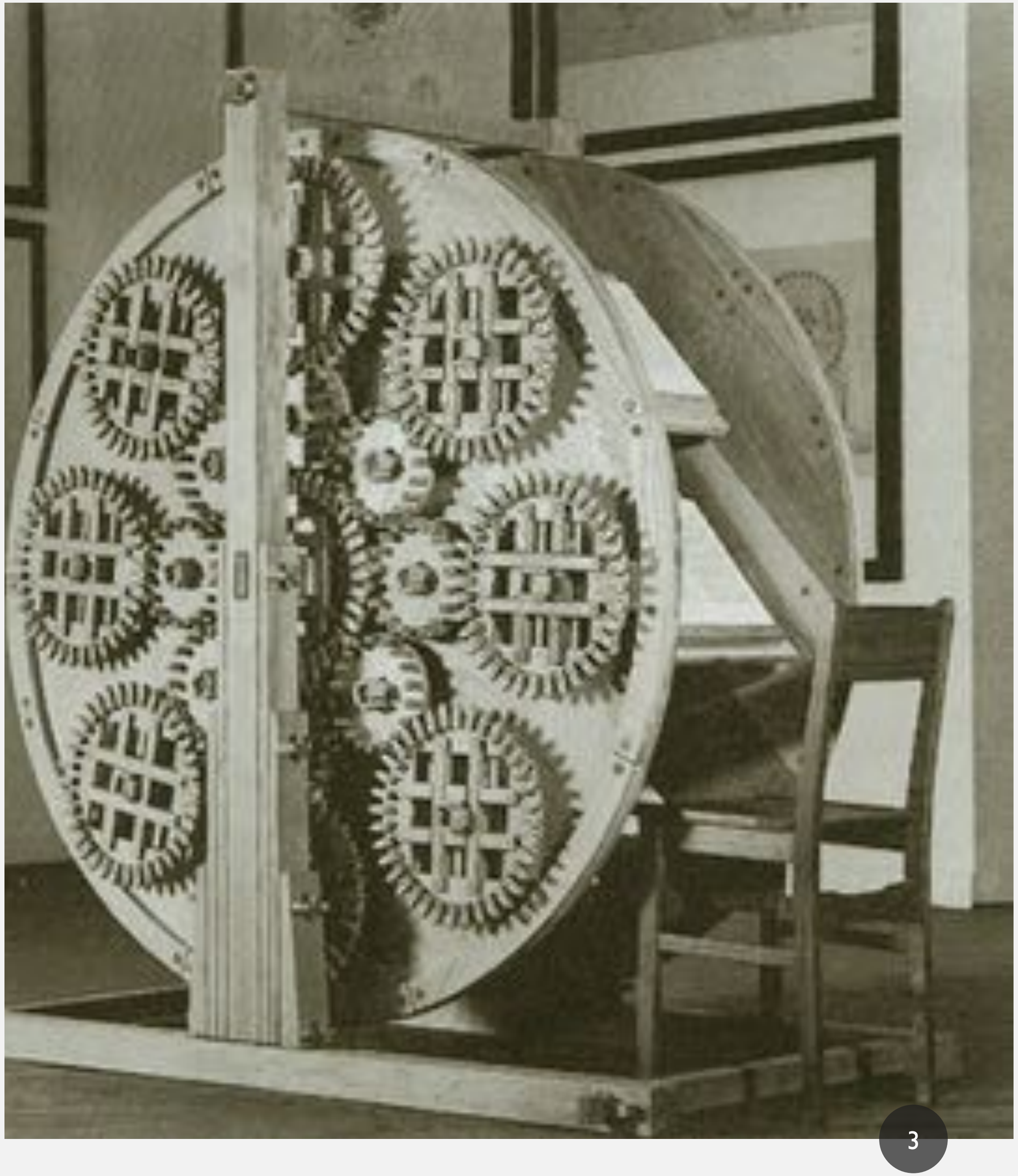


Image courtesy of Wikimedia commons



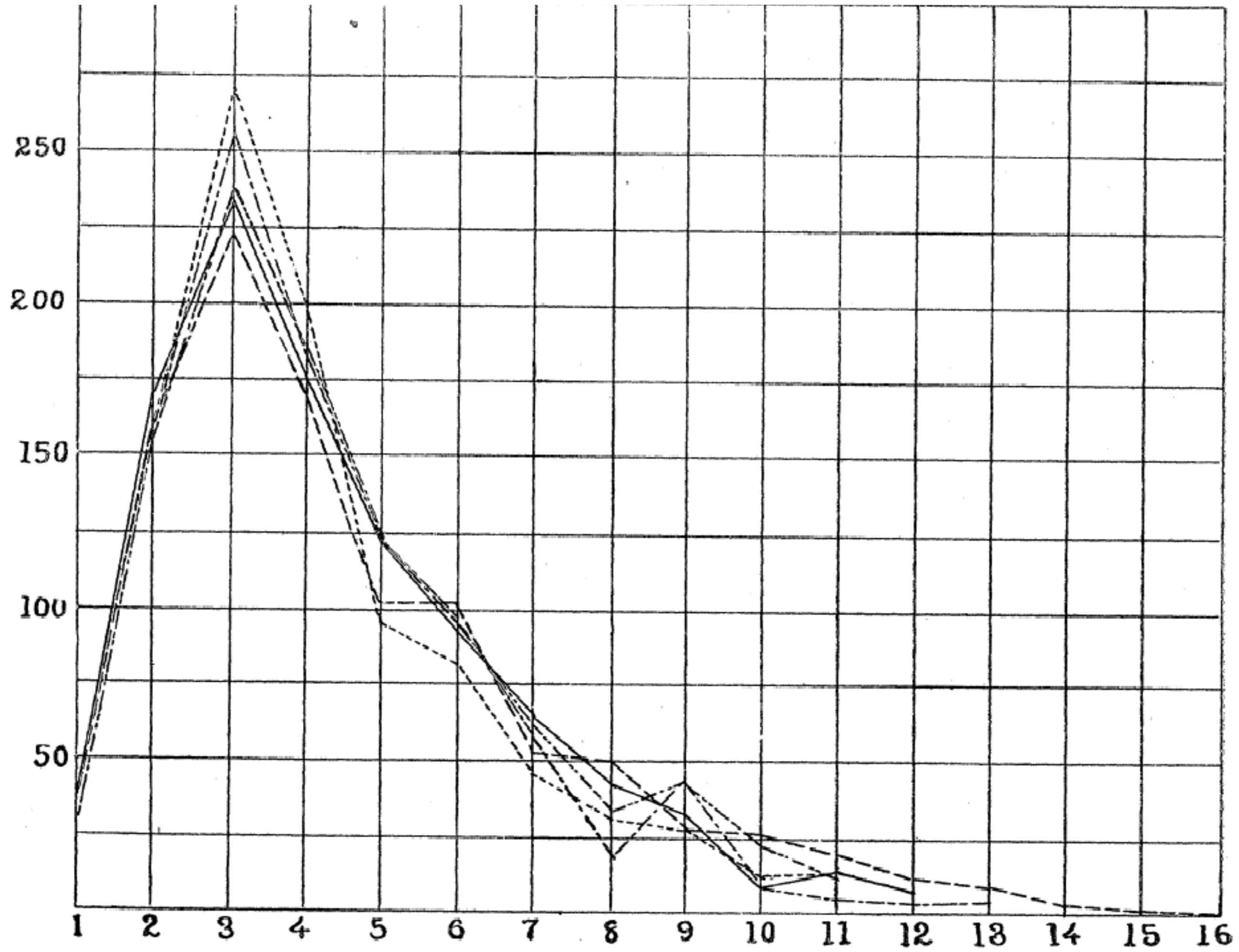
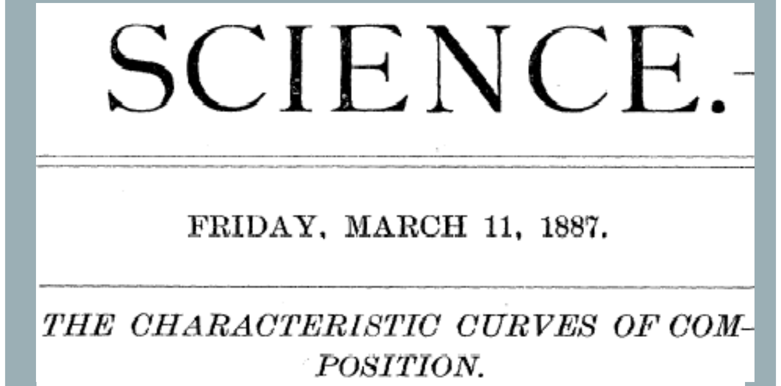


FIG. 2.—SHOWING FIVE GROUPS, OF ONE THOUSAND WORDS EACH, FROM 'OLIVER TWIST.'



SCIENCE.

FRIDAY, MARCH 11, 1887.

THE CHARACTERISTIC CURVES OF COMPOSITION.

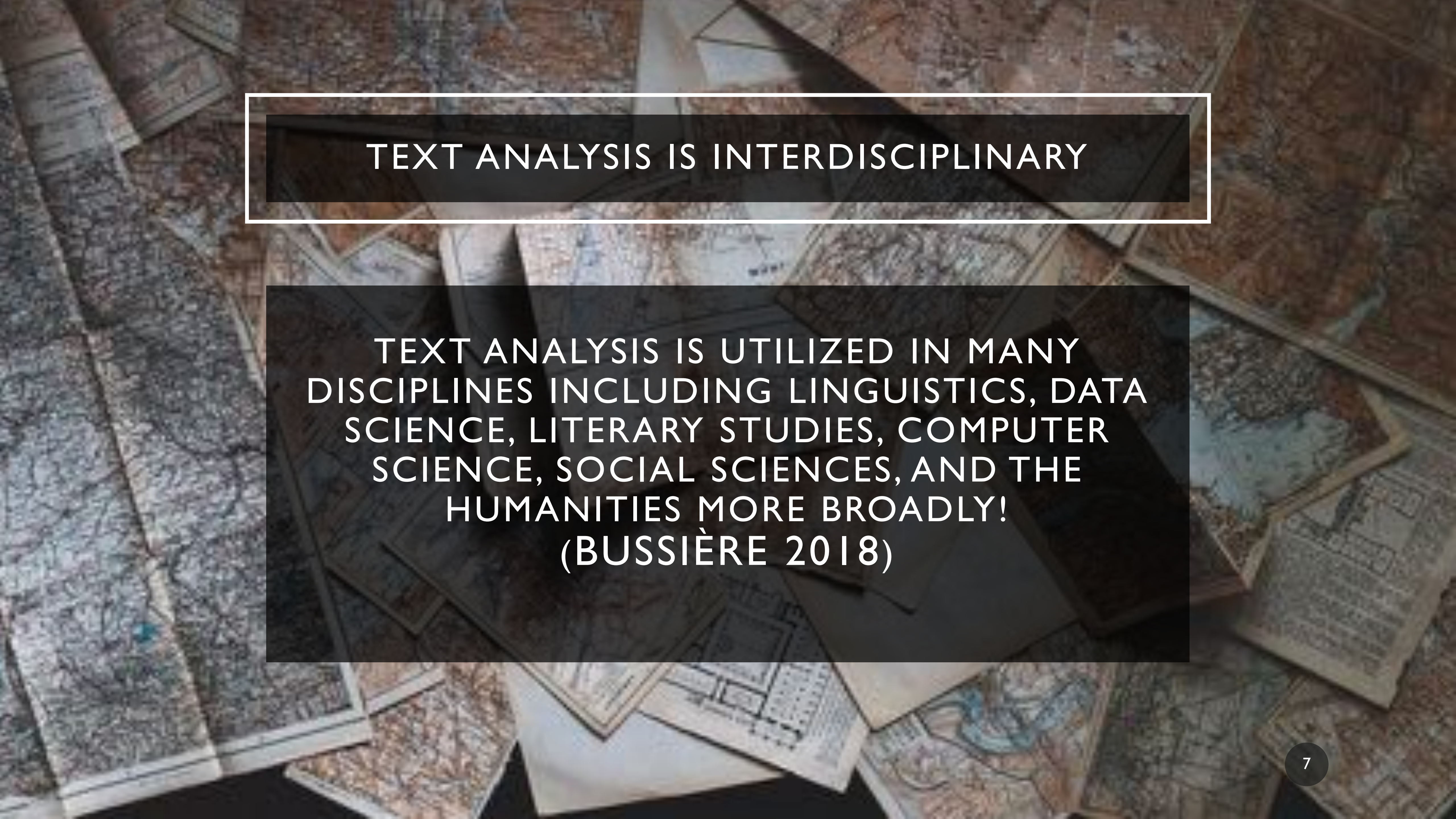
Thomas C Mendenhall (1887) wrote one of the first statistical text analyses (Norman; Madigan & Lewis)

Photo by Thomas Millot on Unsplash

- Early studies can be also traced back to automatic translation projects in the 1940s and 50s (Witten).
- Kucera and Nelson Francis's work on the Brown corpus, a 1 million word database of American English (1967).
- First transcribed corpus of spoken language was created in 1971 by the Montreal French Project, 1 million words (Sankoff & Sankoff 1973).
- Another highly influential study is Kretzschmar et al. 2004's work on the US Tobacco Industry Documents Corpus.

WHAT IS TEXT ANALYSIS?

- Encompasses the processes involved in analyzing computerized natural language databases (ie. Corpora) in order to:
  - Organize/reformat
  - Describe
  - Understand
  - Investigate linguistic and rhetorical devices
  - Study specific genres, contexts, and/or discourses



TEXT ANALYSIS IS INTERDISCIPLINARY

TEXT ANALYSIS IS UTILIZED IN MANY DISCIPLINES INCLUDING LINGUISTICS, DATA SCIENCE, LITERARY STUDIES, COMPUTER SCIENCE, SOCIAL SCIENCES, AND THE HUMANITIES MORE BROADLY!
(BUSSIÈRE 2018)

METHODS

- distant reading
- natural language processing
- machine learning
- corpus linguistics, corpus-based analysis

you're always trying to talk about collocations what is it with you?
collocation
I just think they're neat!

YOU SHALL KNOW A WORD BY THE COMPANY IT KEEPS. JR FIRTH (1957)

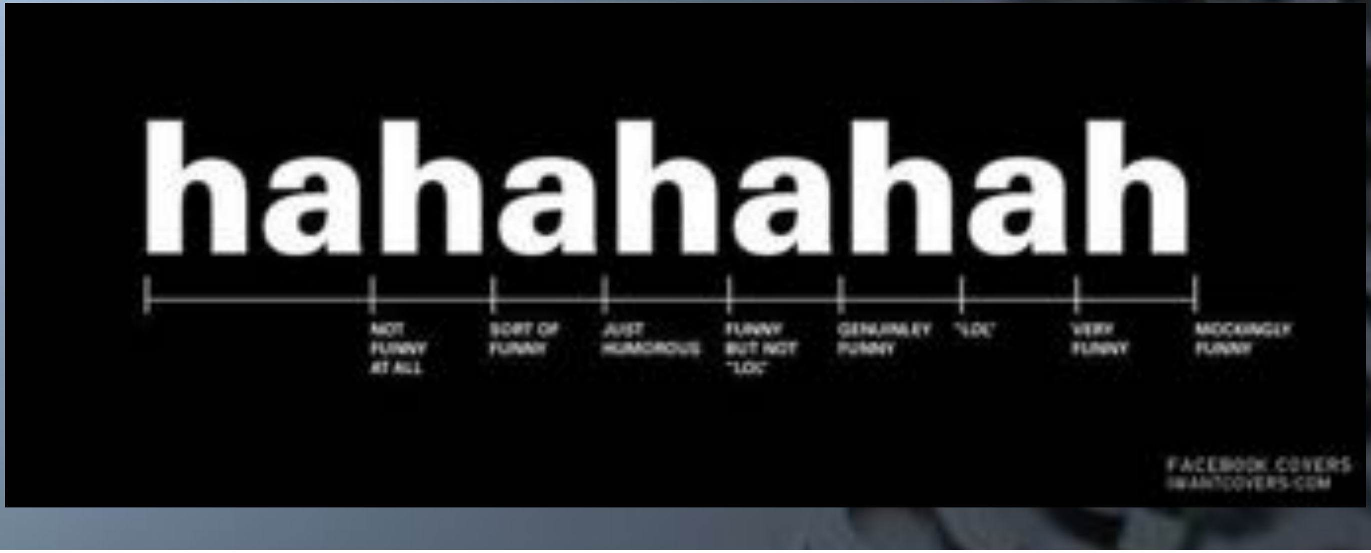
METHODS

document classification

topic modeling

sentiment analysis

frequency analysis,
collocational analysis



TEXT ANALYSIS IS SCALABLE.

PYTHON LIBRARIES

- **spaCy**: pos tagging, tokenization, dependency parsing, etc. Check out this tutorial for more about NLP with spaCy
- **CoreNLP**: lemmatization, pos tagging, tokenization, named entity recognition
- **NLTK**: Natural Language ToolKit; contains over 100 corpora, includes options for tokenization, tagging, parsing, document classification
- **Gensim:** useful for various types of topic modeling
- **PyNLPI:** open-source NLP library; great for of tasks ranging from building simplistic models and extraction of n-grams and frequency lists, with support for complex data types and algorithms
- **Pattern**: useful for web-crawling (webscraping) for creating your own corpora; includes options for tokenizing, pos tagging, etc
- **Polyglot**: very useful library for other languages than English
- **TextBlob:** includes options for pos-tagging, noun phrase extraction, classification, translation and sentiment analysis

- **Tidytext**: helpful for data formatting and visualization; works well with other packages in the Tidyverse (Silge & Robinson 2016)
- **Textmining/tm:** includes options for data processing, metadata management, and creation of term-document matrices (Feinerer 2020; Feinerer et al. 2008)
- **Syuzhet:** package created specifically for sentiment analysis by Jockers
- **Text2vec**: dtm, vectorizing data, supports topic modeling and collocational analysis, too
- **StringR**: supports regex, pattern matching, useful for string manipulation
- **spacyR**: NLP package originally created for Python; useful for tokenization and works well with quanteda and tidytext
- **Quanteda**: incredibly useful package; includes preprocessing abilities, dtm function, as well as statistical analyses options like document classification and topic modeling
- **Ggplot2**: great way to visualize your data

RESOURCES AT UGA

- Corpus Server
- Upcoming Courses
- Digilab Resources
- Data Office Hours

THE UGA CORPUS SERVER

- The corpus server utilizes CQP. (Corpus Query Processor).
- For access to the corpus server, please email linglab@uga.edu.
- On the corpus server, we have access to many awesome databases of language, including:
- **EuroParl**
- **SpokenBNC 2014**
- **COHA: Corpus Of Historical American English**
- **The Brown Corpus**
- The Digital Archive of Southern Speech (DASS)
- Ancora

COURSES AT UGA

- This Fall 2021:
- Natural Language Processing: LING 4570/6570
- Style: ENGL/LING 4826/6826
- American English: ENGL/LING 4010/6010
- Note: These all count toward the Digital Humanities Undergraduate certificate!

ADDITIONAL TOOLS

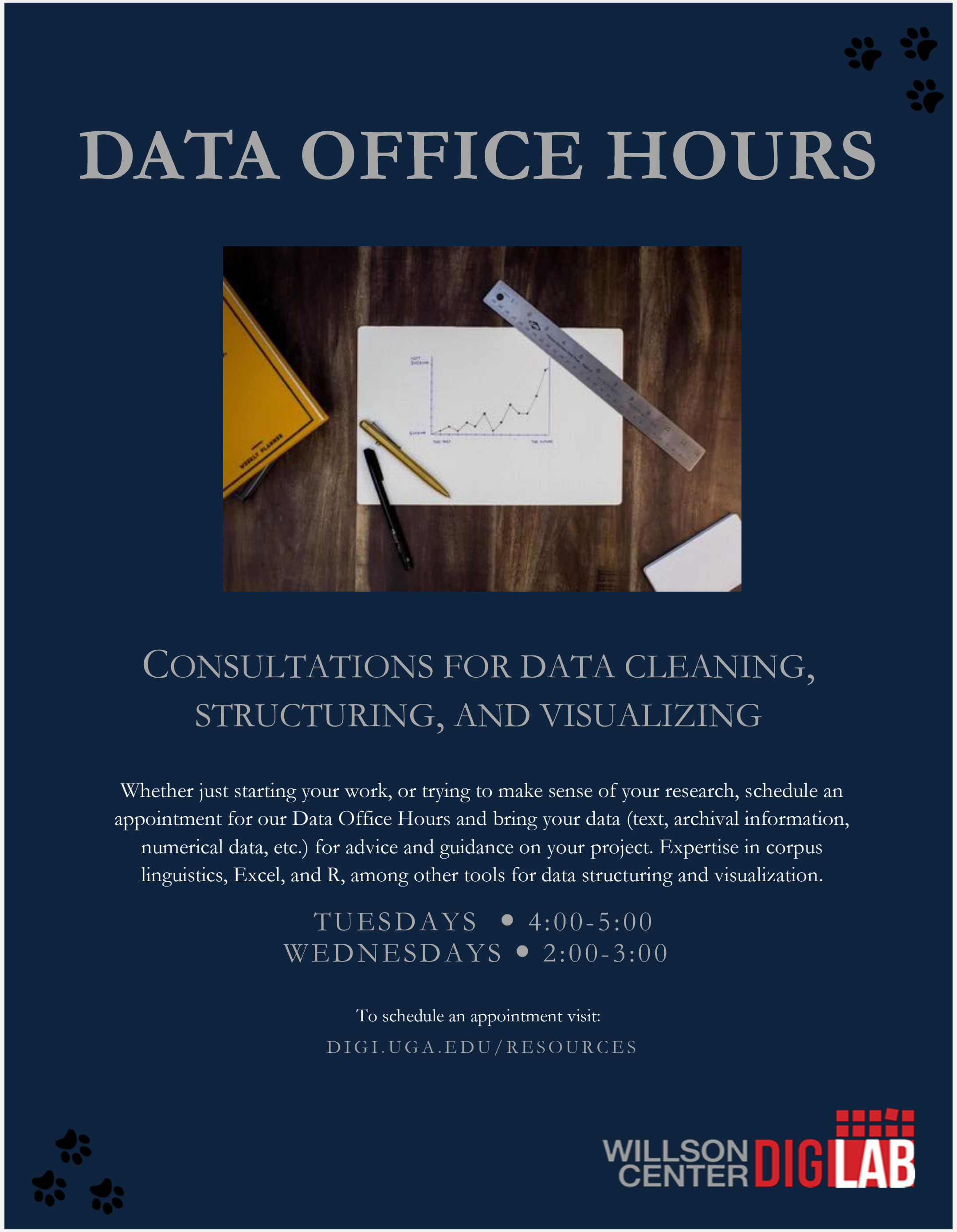
- AntConc: A free corpus analysis toolkit for concordancing and corpus-based methods
- Voyant Tools: web-based text reading and analysis environment
- Google Books Ngram Viewer: online search engine that charts the frequencies of any set of comma-delimited search strings
- Wordseer: text analysis environment that combines visualization, information retrieval, and nlp methods
- Tapor: web-based set of text analysis tools

ADDITIONAL TOOLS

- TextArc: A visual representation of a text.
- MALLET: Maps patterns across texts with various tools.
- Perl: was originally created to be a general purpose programming language to help with reports; includes many excellent text-specific functions; supports powerful regular expressions, string processing, and parsing
- SketchEngine: text mining app based out of the EU; includes options for your own corpora and includes 500+ other corpora
- http://corsis.sourceforge.net/: open source corpus software written in C
- ICECUP 3.1, Fuzzy Tree Fragments: based at UCL, set of corpus tools for parsed corpora like **ICE-GB** and **DCPSE**

ADDITIONAL DATA SOURCES

- Linguistic Data Consortium
- The World Wide Web
- UGA Library Databases
- The Linguistic Atlas Project
- The Hathi Trust Digital Library
- Project Gutenberg
- MONK: Metadata Offer New Knowledge: text analysis suite and public domain TEI texts
- TAPoR: Text Analysis Portal for Research at McMaster Uni
- Martin Weisser's list of historical corpora
- CLARIN historical corpora



DATA OFFICE HOURS

CONSULTATIONS FOR DATA CLEANING, STRUCTURING, AND VISUALIZING

Whether just starting your work, or trying to make sense of your research, schedule an appointment for our Data Office Hours and bring your data (text, archival information, numerical data, etc.) for advice and guidance on your project. Expertise in corpus linguistics, Excel, and R, among other tools for data structuring and visualization.

TUESDAYS • 4:00-5:00
WEDNESDAYS • 2:00-3:00

To schedule an appointment visit:

DIGI.UGA.EDU/RESOURCES

WILLSON CENTER DIGILAB

RECOMMENDED RESOURCES

- Brezina's *Statistics in Corpus Linguistics*
- Evert's work on collocations and corpus methods
- University of Lancaster Corpus for Schools
- Natural Language Processing with Python by Bird et al.; Na-Rae Han's python tutorials
- Silge and Robinson's Text Mining with R
- University of Birmingham, UK Centre for Corpus Research
- HELSINKI's VARIENG Center for Research

COMING UP NEXT…

11 Feb. Intro to text analysis

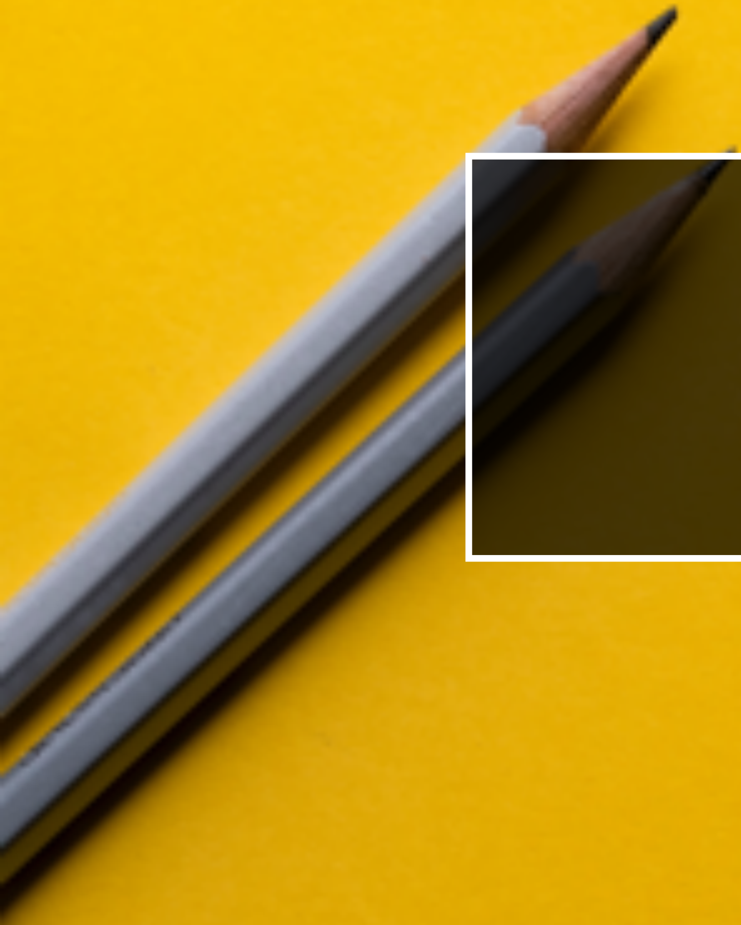
With Jupyter Notebooks and Python

18 Feb. Intro to R for text analysis

25 Feb. Advanced R for text analysis

IN PREPARATION FOR NEXT WEEK

- Download and install:
- Anaconda
- Jupyter notebooks



THANKS FOR LISTENING!

KATHERINE.KUIPER25@UGA.EDU

PLEASE FILL OUT THIS SURVEY.

WORKS CITED

- Bird, Steven, Ewan Klein, and Edward Loper. 2019. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit
- Blaette, Andreas. 2020. Introducing the ‘polmineR’-package. https://cran.r-project.org/web/packages/polmineR/vignettes/vignette.html.
- Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics.*
- Brown, Simon. 2016. Tips for Computational Text Analysis. https://matrix.berkeley.edu/research/tips-computational-text-analysis
- Bussiere, Kirsten. 2018. Digital Humanities - A Primer.
- Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference,* University of Birmingham, UK.
- Evert, Stefan. 2003. The CQP Query Language Tutorial.
- Evert, Stefan. 2007. Corpora and collocations. http://www.stefan-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf
- Feinerer et al. 2008.
- Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
- Firth, JR. 1957. Papers in Linguistics. London: OUP.
- Garber, Megan. 2013. Behond, the Kindle of the 16th Century. *The Atlantic. https://www.theatlantic.com/technology/archive/2013/02/behold-the-kindle-of-the-16th-century/273577/*
- Han, Na-Rae. Python 3 tutorials. http://www.pitt.edu/~naraehan/python3/.
- HathiTrust. https://www.hathitrust.org/about.
- Jockers, Matthew. 2020. Introduction to the Syuzhet Package. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html
- Kretzschmar, William, C. Darwin, C. Brown, D. Rubin, D. Biber. Looking for the Smoking Gun: Principled Sampling in Creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics.* 32:1.
- Laudun, John. Text Analytics 101. https://johnlaudun.org/20130221-text-analytics-101/
- Loria, Steven. 2020. TextBlob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/

- 2020.Modern Perl: Why Perl Rules for Text. https://somedudesays.com/2020/02/modern-perl-why-perl-rules-for-text/
- https://monkeylearn.com/text-analysis/
- Millot, Thomas. Photo. Unsplash
- Nordquist, R. 2019. "Definition and Examples of Text in Language Studies. https://www.thoughtco.com/text-language-studies-1692537
- Norman, Jeremy. Thomas Mendenhall Issues One of the Earliest Attempts at Stylomtery. Historyofinformation.com https://www.historyofinformation.com/detail.php?id=4120
- O'Connor, Brendan, David Bamman, and Noah Smith. 2011. Computational Text Analysis for Scoial Science: Model Assumptions and Complexity.
- Parlante, Nick. 2002. Essential Perl. http://cslibrary.stanford.edu/108/EssentialPerl.html.
- Project Gutenberg. https://www.gutenberg.org
- Sankoff, D. & Sankoff, G. Sample survey methods and computer-assisted analysis in the study of grammatical variation. In Darnell R. (ed.) *Canadian Languages in their Social Context* Edmonton: Linguistic Research Incorporated. 1973. 7–64.
- Witten, Ian. 2004. Text mining. https://www.cms.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf