Text Analytics and Sentiment Analysis

World Changers Shaped Here

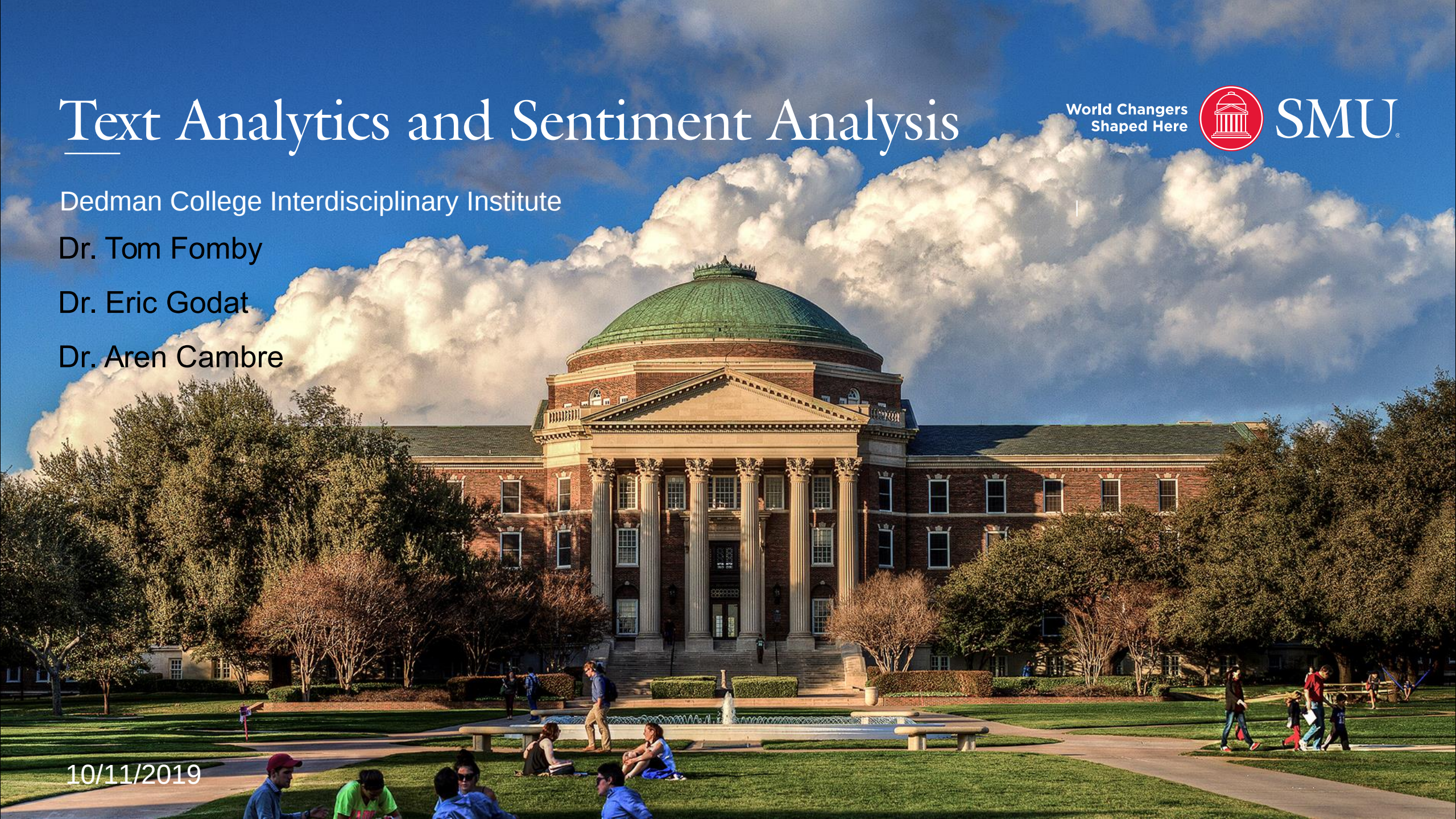
SMU

Dedman College Interdisciplinary Institute

Dr. Tom Fomby

Dr. Eric Godat

Dr. Aren Cambre



10/11/2019

- Ph.D. in Theoretical Particle Physics from SMU
- Data Science Research Applications Developer for the Office of Information Technology
- Interim Academic Technology Service Director for Dedman College

- PhD. in Economics from University of Missouri at Columbia
- Professor of Economics at SMU
- Focus in predictive analytics and applied econometrics

- Doctorate in Engineering Management from SMU
- Director of Web Development for the Office of Information Technology
- Adjunct Lecturer in the Department of Economics

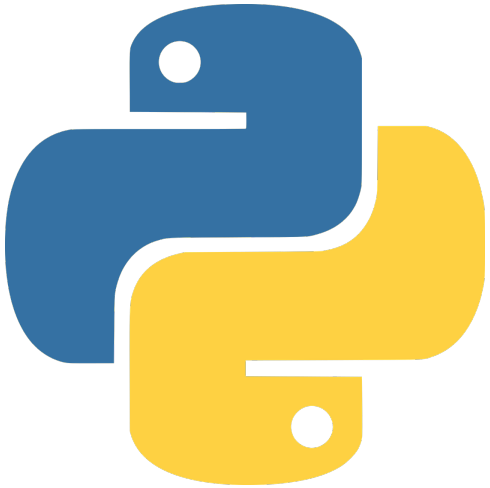
What is Text Analytics?

- **Text Analytics** or **Text Mining** is a method for extracting meaning from text beyond simple reading
  - Leverages programming and software tools
  - Relies on statistical techniques and Natural Language Processing (NLP)
  - Can be simple or complex:
    - Word counts to sentiment analysis
  - Text manipulation and preprocessing (cleaning)
- Used in a wide variety of disciplines
  - Just about anything that people write about
    - Economics: quarterly reports, federal reserve reports
    - Marketing: product reviews
    - Biomedical: patient records
    - Politics: speeches, court rulings
    - Humanities: literature, historical records, etc.
    - Scientific papers: literature reviews

Text Mining Tools

- R vs Python
  - Matter of preference
    - Both have text analysis libraries
      - e.g. NLTK in Python, tm in R
    - Open source
  - We will use R with:
    - Tidyverse
    - RStudio

VS



- Other tools:
  - SAS: Text Miner and Teragram
  - SPSS: Modeler Premium
  - Mathematica: Wolfram Language built-in functions
  - MatLab: Text Analytics Toolbox

Coding Basics

- Objects
  - `1`
  - `"test"`
  - `x <- 1`
  - `y <- "test"`

Coding Basics

- Vector
  - `scouts <-  c("Sam", "Ethan", "Davian", "Hayes", "Max")`

- Function

```
addnumbers <- function(m,n) {
    z = m + n
    print(z)
}
```

- Function that returns a value

```
addnumbers <- function(m,n) {
    z = m + n
    return(z)
}
```

Coding Basics

- Data Frame
  - `scouts <-  c("Sam", "Ethan", "Davian", "Hayes", "Max")`
  - `ranks <- c("Webelos", "Bear", "Lion", "Bear", "Wolf")`

  - `pack <- data.frame(scouts, ranks)`

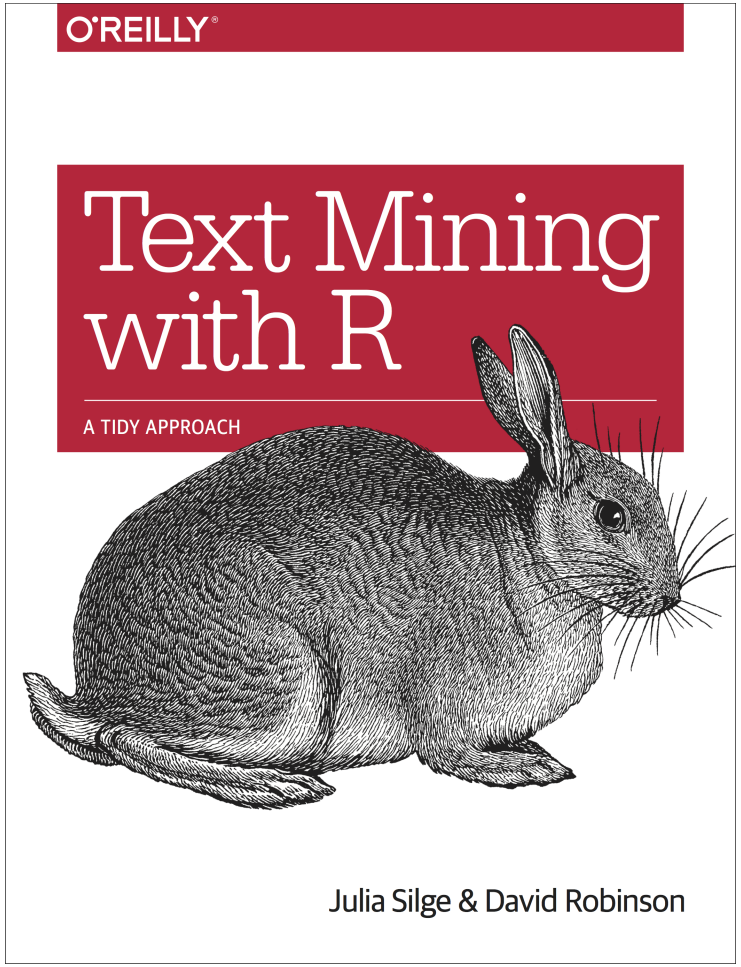
- Some notes on R:
  - Libraries add functions, data, etc. to R
  - The arrow, `<-` , is an assignment operator
    - The equal sign, =, is used in function parameters in R
  - You will see this combination of characters %>%
    - Called a “pipe”
    - “take this data” %>% “do this to it”
    - Can chain multiple pipes together

Usenet

- In general
  - A distributed bulletin board posting system
    - Users could post messages
  - Established in 1980, still active today
  - Think “Reddit before the internet really took off”
  - Origin of terms like “FAQ” and “spam”
- Usenet20
  - 20,000 messages from 20 newsgroups in 1993
    - Lots of text by lots of writers on a variety of different topics
  - Commonly used in text mining and machine learning experiments
    - Somewhat clean and structured
  - Publicly available here: http://qwone.com/~jason/20Newsgroups/
- If you are interested in Usenet data for research, talk to us afterwards

Ready for Launch

- We will walk through an example text analysis in R
  - Goal: Motivate the “why” more than describe the “how”
  - Demonstrate capabilities
  - Detailed lesson:
    - Chapter 9 from *Text Mining with R* by Julia Silge and David Robertson
    - Free textbook available here: https://www.tidytextmining.com/
- This is for you
  - Please stop us to ask questions
  - We will take a break halfway through
- Any questions before we get started?



RStudio

Wrap Up

- Want to learn more?
  - *Text Mining with R* by Julia Silge and David Robertson
    - Free textbook available here: https://www.tidytextmining.com/
  - Keep an eye out for:
    - Aren's Text Mining Practicum
    - Data Science Workshops from OIT and SMU Libraries
- Code available on GitHub here:
  - https://github.com/SouthernMethodistUniversity/DCII-Text_Analytics_2019

Closing

- Next Session:

  November 15th 2:00 – 4:00 Fondren Science 133

  Dr. Jo Guldi

  What was Gladstone's favorite word? Ophthalmia. What was the most important year in the nineteenth century? 1832. How does Dr. Guldi know these things? Humanists like Dr. Guldi are at the forefront of a revolution in using computers to understand text, with applications in contemporary politics, information retrieval, and app design.

- Thank you for coming!