

# Operation: Debunker

DSI Capstone Presentation

Charles Rice

*“As a wizard, I must remind you that words have power,” said Ridcully  
‘As a politician, I must tell you I never forget it,’ replied Vetinari.  
‘Yet you choose to use **these** words at **this** time?’”*

*“Ah, so it is not words, but context that has power?”*

*- Terry Pratchett, **Unseen Academicals***

Whether we take the signified or the signifier, language has neither ideas nor sounds that existed before the linguistic system, but only conceptual and phonic differences that have issued from the system.

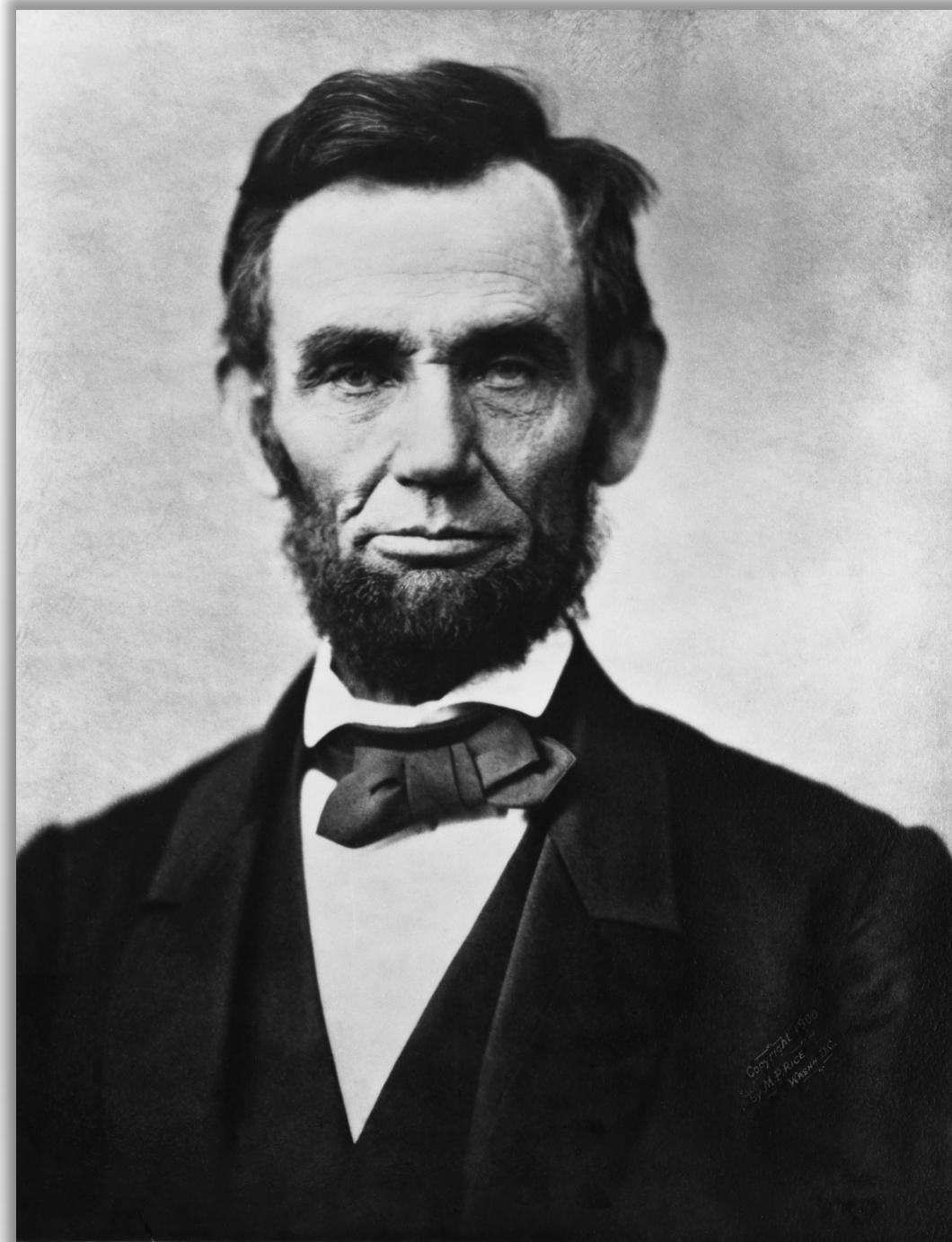
-- Ferdinand de Saussure, "Course in General Linguistics," 1916

The best way to predict the future  
is to create it.

-Abraham Lincoln

-Michael J. Astrue, Commissioner  
Social Security Trust Fund

Attributed to Lincoln in 2008

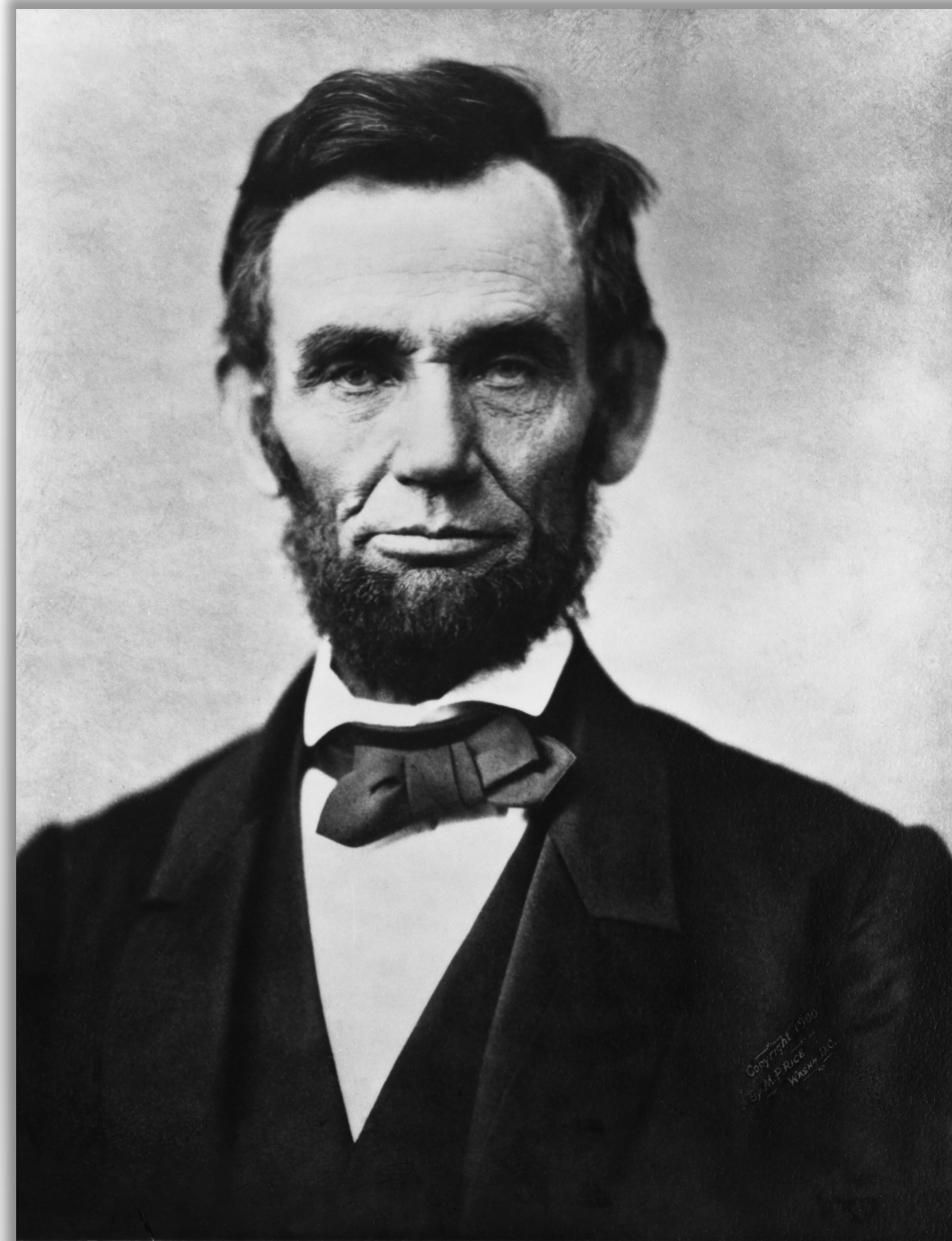


The best way to predict the future  
is to create it.

-Peter Drucker, Management Guru, 1986

-Alan Kay, Chief Scientist, Atari, 1982

-Denis Gabor, "Inventing the Future," 1963



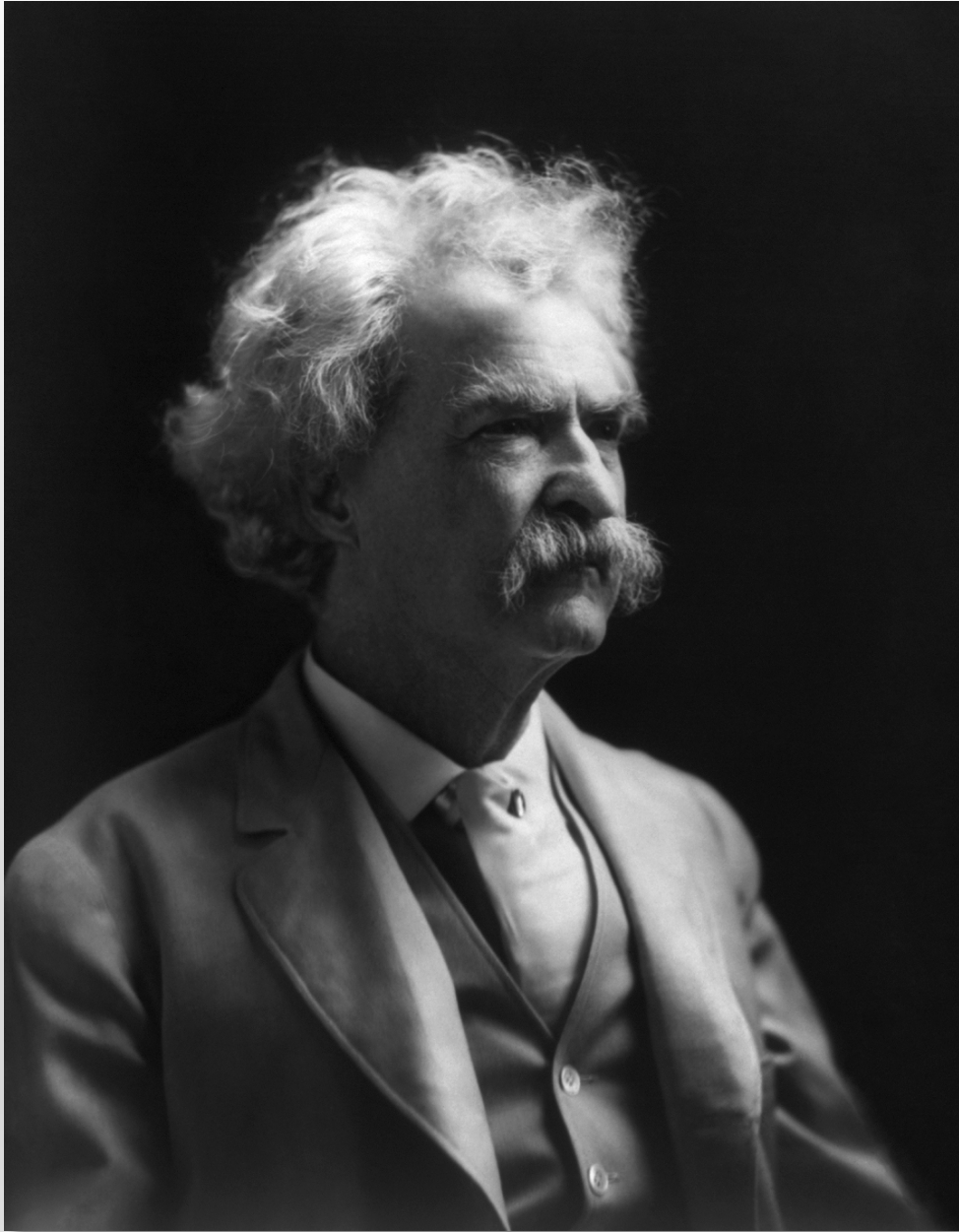
# Writers and Their Sayings

- We know that writers develop unique styles
- Can those styles be at all predictive of what they say, or are reported to have said?
- Mark Twain (1835 – 1910)
- Oscar Wilde (1854 – 1900)
- Abraham Lincoln (1809 – 1865)

# You Will, Oscar...

- Lincoln was a politician, not a writer
- Twain was a writer – 3m+ words
- Wilde was a wit
- All three present problems for this project
- Furthermore, styles evolve over time
- Words into numbers





“Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: "There are three kinds of lies: lies, damned lies and statistics.”

*- Mark Twain's Own Autobiography: The Chapters from the North American Review*



# Cleaning and Preprocessing

- Corpora were collected from Project Gutenberg
- Both Twain and Lincoln have “Collected Works” files
- Wilde had to be assembled
- All three were cleaned by hand
  - Approximately 100,000 words removed
  - Editor’s marks, formal language, dates, times, salutations, valedictions

# Preprocessing, Continued

- Initially planned to use TFIDF
- So I split the data into pseudo-documents at the character, word, and sentence level
- Equal numbers of observations (1,000, 500, and 100) for all three

# Still Preprocessing ...

- Started with a basic CountVectorizer
- Manual pseudo-gridsearch tweaking hyper parameters
- Wound up using character level n-grams, 1-4 inclusive
  - This conforms to research indicated character-level ngrams have more predictive power
- Naïve Bayes for the modeling

## ...Some are useful

- Initial precision scores were extremely high – between .8 and .95
- Ran a train-test split for Naïve Bayes and jerry-rigged a cross-validation
- Mean precision scores remained high, although there were some interesting patterns
- Scores fell off using TFIDF vectorizer, so stayed with CVEC
- Naïve Bayes performed so well that I did not see the need to test others.

So, after all that, did the model do what it was supposed to?

No.

No, it did not.

Of the 91 quotations with known attributions, only 8 were correctly guessed by the model.

Including *that* quote

# Next Steps

- Correct the class imbalance
- Identify a source of 'modern' text that more closely matches what I'm trying to debunk
- Add authors over time
- Identify a way to streamline the pre-processing



Words are the currency of thought, the circulating medium of ideas. And we pay as steep a price for muddled thinking as for muddled speaking.