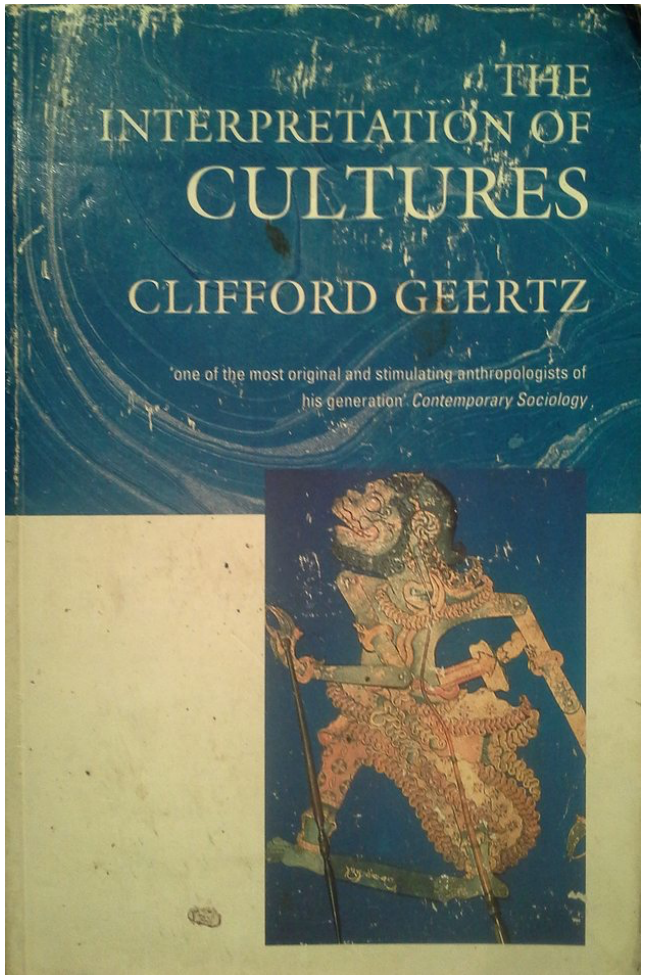
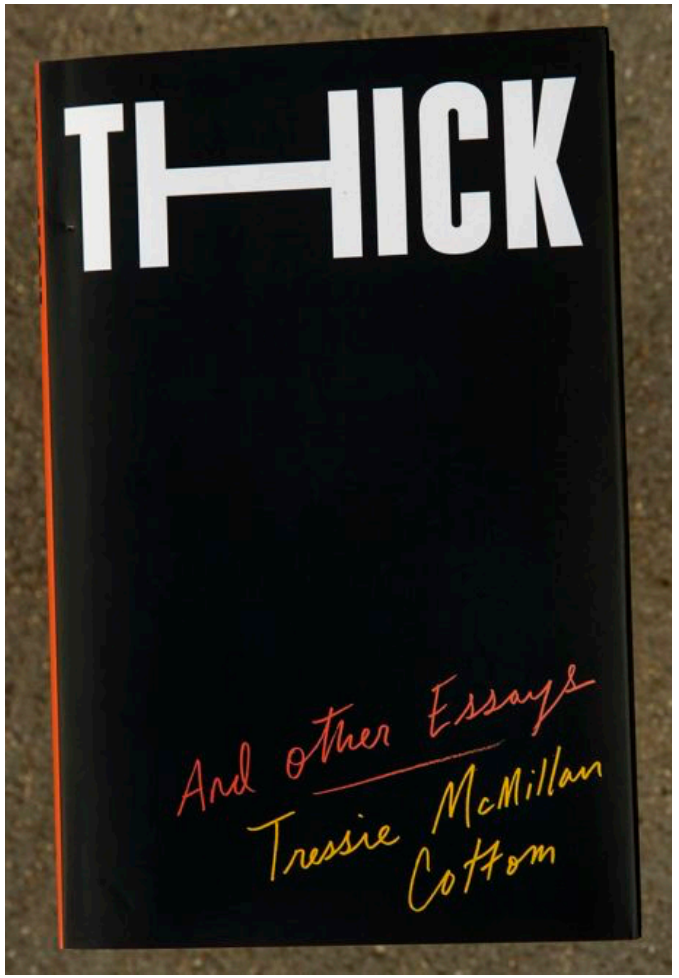
Arguing with Textual Data

(and *arguing* with textual data)

Today's intertext: Nguyen et al., "How we do things with words: Analyzing text as social and cultural data" (2019)

Textual data lets us answer “thick” questions





Thick questions + computation…

- Humans can…
  - Make logical inferences
  - Resolve ambiguities
  - Apply cultural knowledge

- Computers can…
  - Detect large-scale patterns
  - What else?

Some challenges with the ”thick questions + computation” nexus:

- All texts are socially and culturally situated
- Most social/cultural concepts are highly contested
- Choices in how we operationalize/analyze those concepts are therefore contested too

What do to?

- Be intentional and reflexive
- Be aware of the choices we make
- Be open to iteration, both within each step of the research process and between steps
- Be aware of the limits of our conclusions

The Research Process

- Research questions
- Data (acquisition and compilation)
- Conceptualization
- Operationalization
- Analysis

Research Questions

- Can text analysis provide a new perspective on a “big question” that has been attracting interest for years?
- Can we raise new questions that have only recently emerged, or that new forms of textual data help prompt?
- How can we explain what we observe?

Research Questions

- In general, *insight-driven:*
- We aim to describe a phenomenon or explain how it came about

Research Questions

- Success isn’t just measured by performance or accuracy
- Success is measured by how well we answer our research questions

Data

- After the research question (or sometimes before), we need to decide on our data sources, collect and compile them, and inspect (or create) any associated metadata

Data

- Some concerns:
  - Potential harms of data collection on the people/events it documents
  - Issues w/ data quality (comprehensiveness, representativeness, etc)
  - Big != better:
    - E.g. https://www.youtube.com/watch?v=e9CHVEheqlo

Data

- Some concerns:
  - Potential harms of data collection on the people/events it documents
  - Issues w/ data quality (comprehensiveness, representativeness, etc)
  - Big != better:
    - E.g. https://www.youtube.com/watch?v=e9CHVEheqlo

Data

- Additional considerations:
  - Do we need a comparison set?
  - Should we remove any data that’s diluting the signal?
  - Do we have metadata that can help us assess the data’s comprehensiveness and/or representativeness? Can we create some?

Conceptualization

- How do we translate “thick” cultural or social concepts into measurable quantities?
  - First, define the concepts! Ask: who are the domain experts and how have they approached the topic?
  - Need definitions of concepts that are flexible enough to apply to our data, yet formal enough for computational research.
  - Background concept vs. systematized concept
  - In the humanities/social sciences, rarely a single ground truth!

Operationalization

- Essentially, developing measures for the concept.
- Key questions:
  - Are we measuring what we intended to measure?
  - Does our operationalization match our conceptual definition?
- Mind the gaps!

Operationalization

- Modeling considerations:
  - Variables
    - Interesting ones: time, space, social network
  - Categorization scheme
    - What choices are we making? What is implicit and what is explicit?
  - Supervised vs. unsupervised
    - Supervised:
      - When we have a clear definition of the concept we want to explore
      - When we have good labels
    - Unsupervised:
      - Exploration
      - Want theory to emerge from data
      - When there is a clear way of measuring a concept, based on strong assumptions
  - Units of interest
    - Novel, chapter, document, sentence, word, etc.
    - Can be a unit we impose, too (e.g. that birth story project)

Operationalization

- Data pre-processing
  - Time consuming
  - Involves many many choices
- Do you clean the data? If so, how?
- Do you tokenize the data? If so, how?
- Do you lowercase, remove punctuation, lemmatize, stem, normalize?
  - E.g. “U.S.A. vs. USA”, “apple” vs. ”Apple”, “No” vs. “Nooooo”
- Issues of inflection?
  - What is not contained within text?
- Remove what’s not relevant?
  - E.g. stoplist.
- Mark up or tag text
  - E.g. POS tagging, NER

Operationalization

- Types of approaches:
  - Dictionary approaches
    - E.g. word lists, word ranks
  - Supervised models
    - Classifiers
    - Sentiment analysis (is supervised!)
  - Unsupervised models
    - Clustering
    - Topic modeling
    - Word embeddings

Operationalization

- Can iterate within approaches and among them
  - E.g. more stopwords in topic model, or topic model -> word count
  - E.g.: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Validation

- Goal of validity is to determine the extent to which a given measurement tool measures what it is supposed to measure
  - In NLP and machine learning, usually compare machine-generated labels against annotated sample. Human as "gold standard."
    - These are what scores like "accuracy," "precision," and "recall" measure.
    - But maybe they're just indicators of *reliability*
  - In humanities disciplines, computational analyses validated by close reading
    - Does the model tell us anything new?
  - Content validation through stratified random sample
    - Select observations from the full range of scores and ask:
      - Do these make sense in light of the systematized concept?
      - If not, what seem to be missing?
      - Or is something extraneous being captured?
  - Compare to other approaches that aim to capture same concept
    - E.g. sentiment analysis vs. dictionary
  - Can always iterate! Return to the model or try something new.

Analysis

- How do our models help us answer our research question(s)?
- Look at errors and failures as well as successes
- What are the implications for the research question(s)?
- What are the limits of those explanations?
- Move back and forth b/t large scale and small scale
- Consider what new questions the analysis has prompted