PROFESSOR BENJAMIN NOBLE (UCSD)

TEXT IS DATA?

A MOTIVATING QUESTION

WSJ

Marjorie Taylor Greene Easily Wins Re-Election in Georgia

Marjorie Taylor Greene, via NPR

Marjorie Taylor Greene 🇺🇸
@mtgreenee

We need President Trump back in office to unleash American energy dominance and end Joe Biden and the Democrats quest to plunge America into darkness.

Alexandria Ocasio-Cortez
@AOC

Fewer things are more predictable than Republicans having a meltdown when I'm clearing them in debate. 💅

Joe Biden
@JoeBiden

We've come a long way, but I won't stop fighting for hardworking families.

I'll continue to stand against extreme MAGA Republicans' efforts to cut Social Security, Medicare, and Medicaid and to enact massive tax giveaways for the wealthy and big corporations.

WHY TEXT AS DATA?

- Much of politics is communication!
  - Laws, speeches, interviews, tweets, donation emails, advertising, campaign websites, etc.
  - We can only go so far without actually looking at the text.

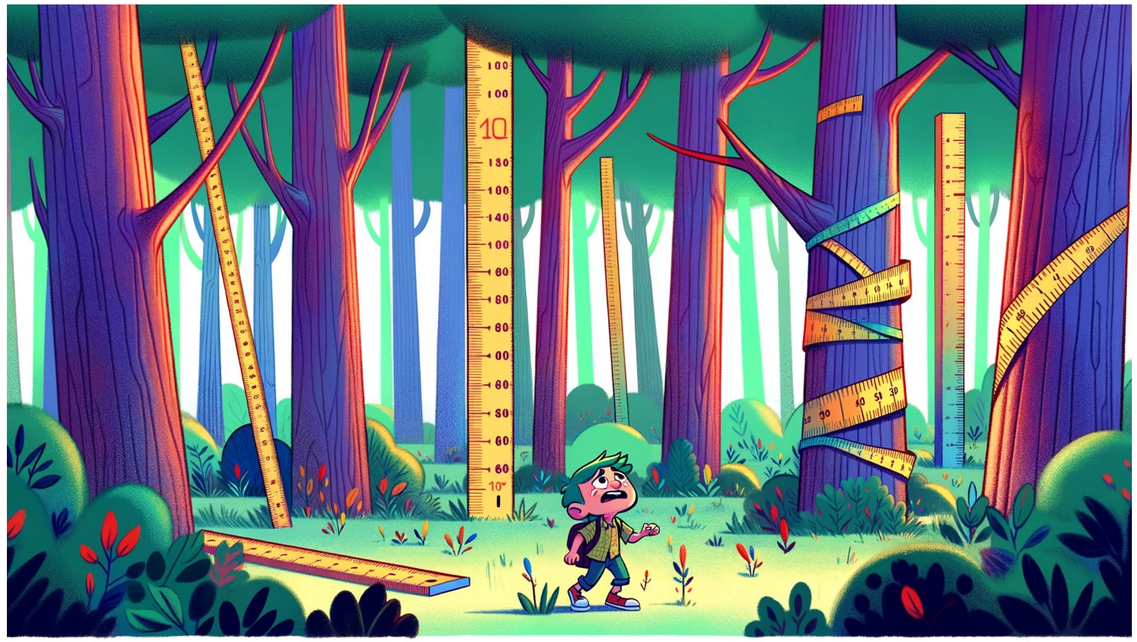
Via DALL-E

WHY NOW?

- Analyzing text at scale...
- Previously:
  - Difficult to acquire documents.
  - Time consuming (or expensive) to read/ analyze.
  - Difficult to categorize text or measure key concepts.

via Library of Congress

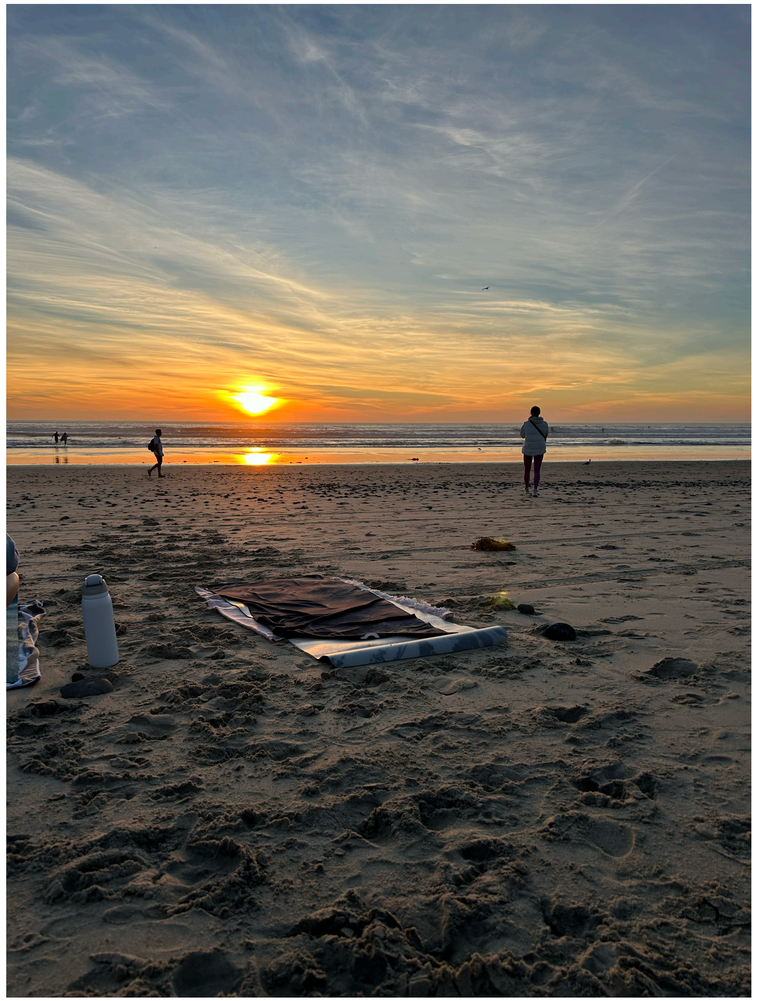
TEXT ANALYSIS AUGMENTS, NOT REPLACES, HUMANS



Via DALL-E

ABOUT ME

- Professor Ben Noble.
  - From St. Louis, MO. Live in San Diego, CA.
  - My research: congressional and presidential rhetoric.
  - Hobbies include: yoga, board games, and cooking.



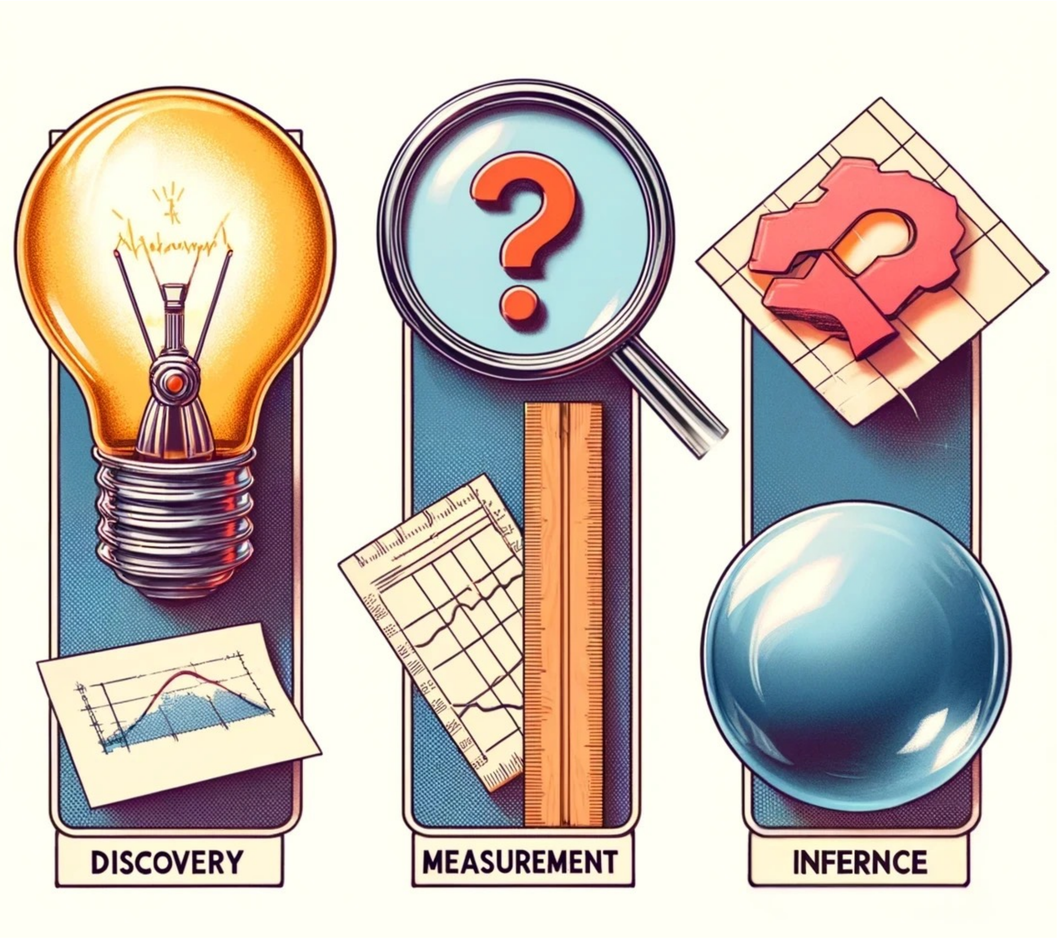
INTRODUCE YOURSELVES!

A LITTLE MORE ABOUT THESE MODULES

- Two modules: computer assisted text analysis I and II.
  - Module I: text is data (this lecture), dictionaries, topic models.
  - Module II: supervised learning, word embeddings, hack-a-thon.
- Each session: half lecture, half lab.

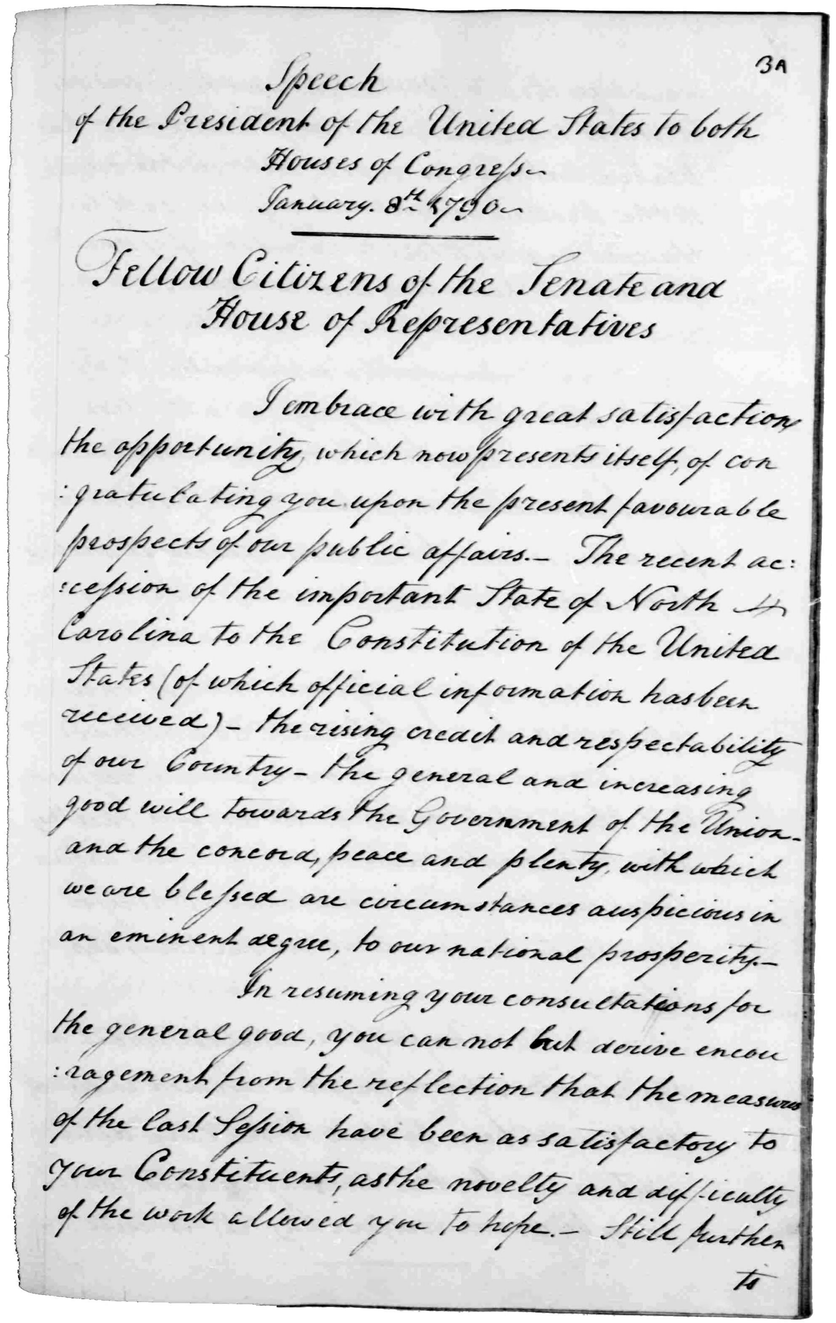
THREE STAGES OF RESEARCH

- Discovery: defining the research question, conceptualizing.



Via DALL-E

TEXT IS DATA?



3A

Speech
of the President of the United States to both
Houses of Congress
January. 8th 1790.

Fellow Citizens of the Senate and
House of Representatives

I embrace with great satisfaction the opportunity which now presents itself, of con: gratulating you upon the present favourable prospects of our public affairs. The recent ac: cession of the important State of North Carolina to the Constitution of the United States (of which official information has been received) the rising credit and respectability of our Country the general and increasing good will towards the Government of the Union and the concord, peace and plenty with which we are blessed are circumstances auspicious in an eminent degree, to our national prosperity.

In resuming your consultations for the general good, you can not but derive encou: ragement from the reflection that the measures of the last Session have been as satisfactory to your Constituents, as the novelty and difficulty of the work allowed you to hope. Still further

to

MARCH 07, 2024

Remarks of President Joe Biden — State of the Union Address As Prepared for Delivery

BRIEFING ROOM › SPEECHES AND REMARKS

The United States Capitol

Good evening.

Mr. Speaker. Madam Vice President. Members of Congress. My Fellow Americans.

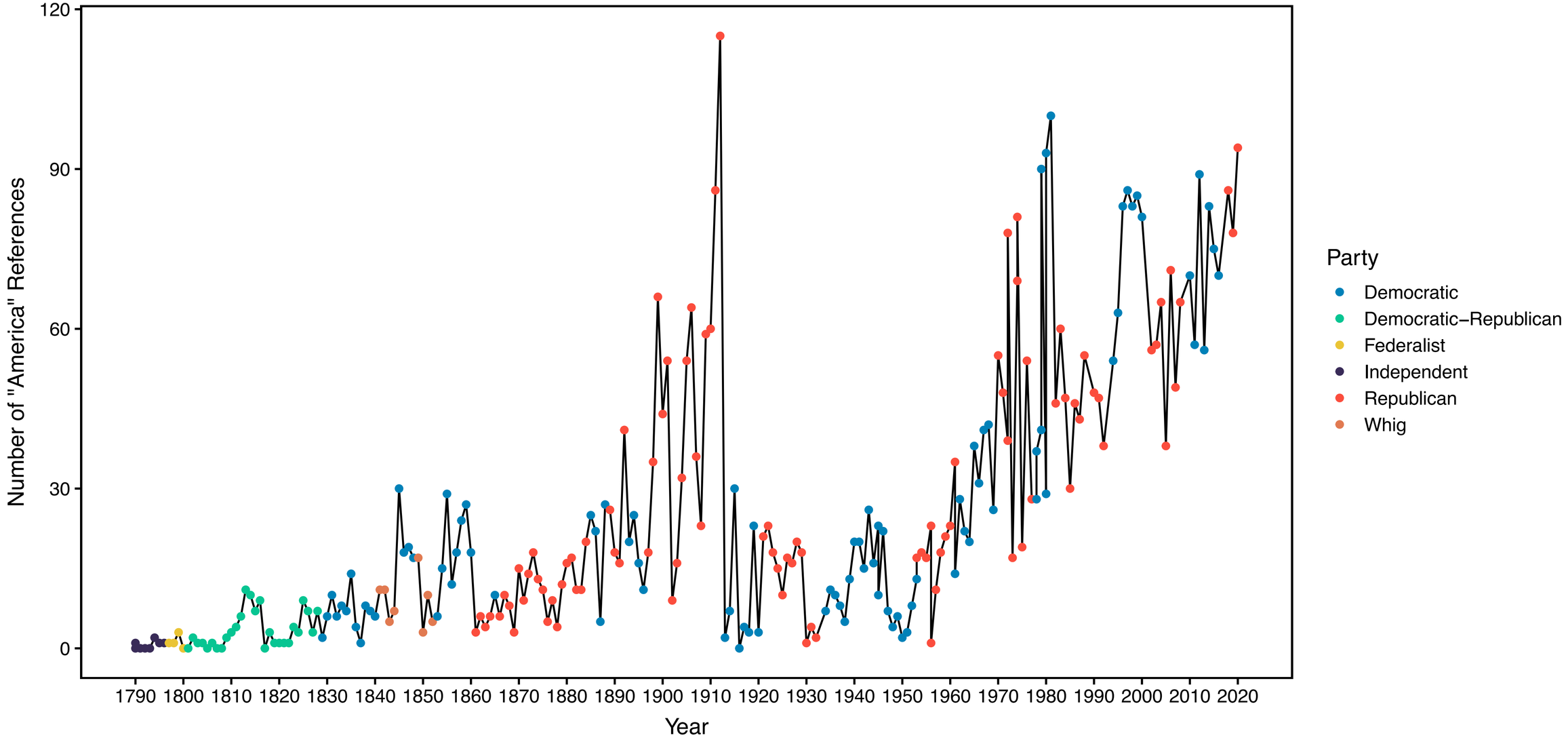
In January 1941, President Franklin Roosevelt came to this chamber to speak to the nation.

He said, “I address you at a moment unprecedented in the history of the Union.”

Hitler was on the march. War was raging in Europe.

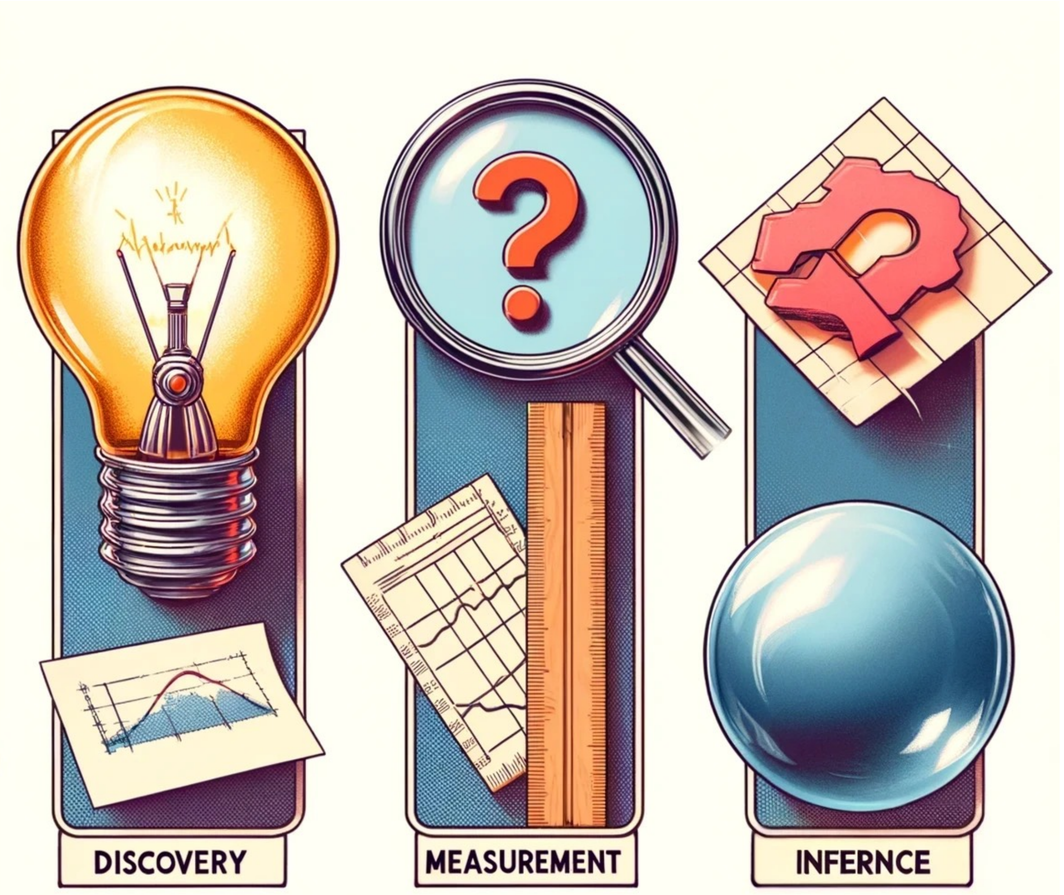
President Roosevelt’s purpose was to wake up the Congress and alert the American people that this was no ordinary moment.

TEXT IS DATA?


Number of "America" References
120
90
60
30
0
1790 1800 1810 1820 1830 1840 1850 1860 1870 1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020
Year
Party
Democratic
Democratic–Republican
Federalist
Independent
Republican
Whig

THREE STAGES OF RESEARCH

- Discovery: defining the research question, conceptualizing.
- Measurement: actually measuring the quantity we care about.



Via DALL-E

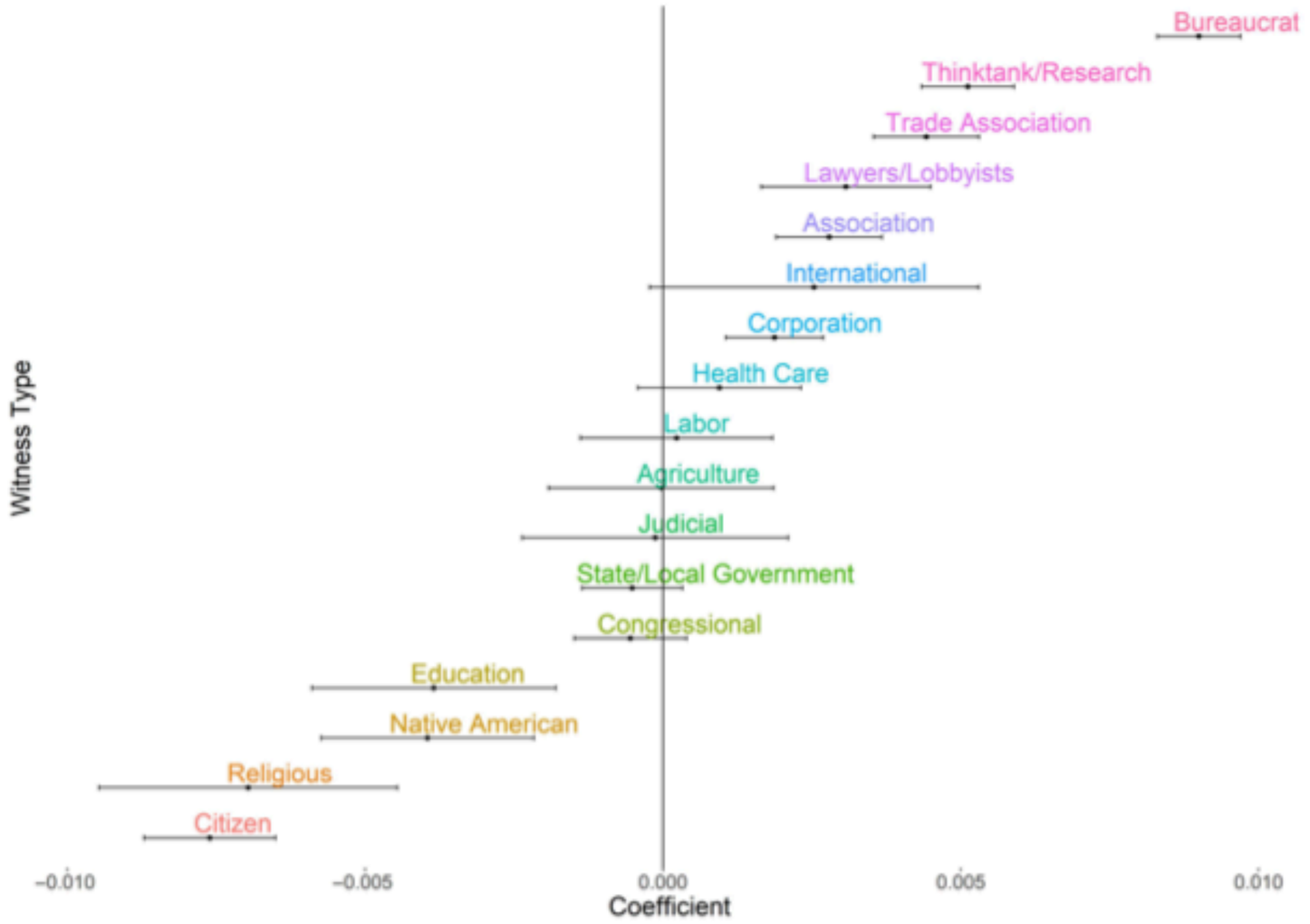
THE CHALLENGE OF MEASUREMENT: AN EXAMPLE

- Ban, Park, and You (2023): How much information do witnesses convey in congressional hearings?
  - What is “information” and how do we measure it in text?

WIRED, Pablo Martinez Monsivais

HOW WOULD YOU MEASURE “INFORMATION” IN TEXT?

FIGURE 3. Proportion of Keywords by Witness Type



Note: Vertical lines indicate 95% confidence interval. The reference group is the witnesses from nonprofit organizations.

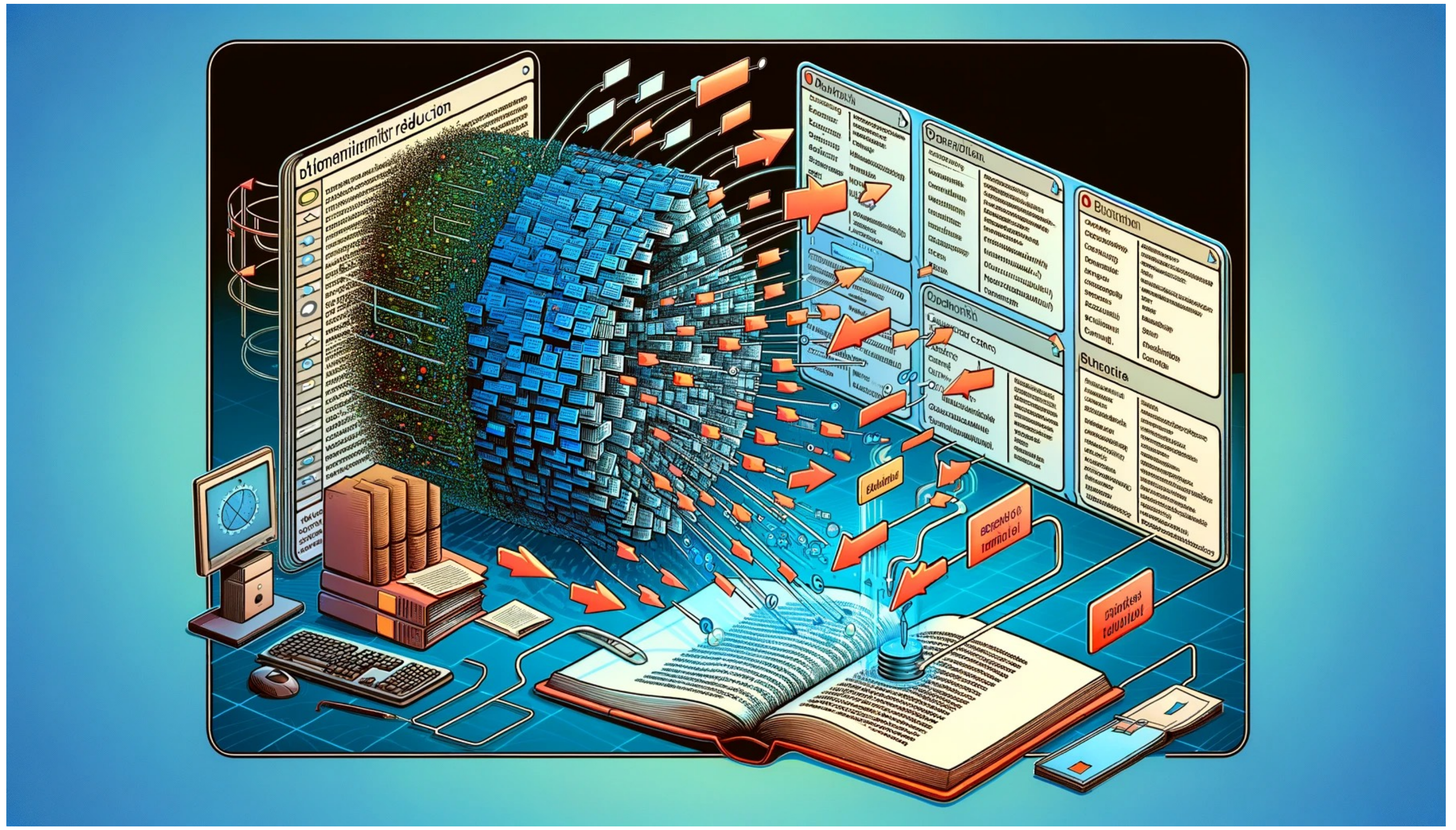
Ban, Park and You (2023)

PRINCIPLES OF MEASUREMENT

- “The goal in measurement is to instantiate some concept within our hypothesis or theory in order to facilitate quantification” (Grimmer, Roberts and Stewart 2022).
  - Measures should have clear goals (no “best” measurement).
  - Construction of the measure should be explainable and reproducible.
  - Validate!

WHY IS TEXT “DIFFERENT”

- Text is **high-dimensional**, not low-dimensional.
- Text is **unstructured**, not “rectangular.”
- Text requires **validation.**



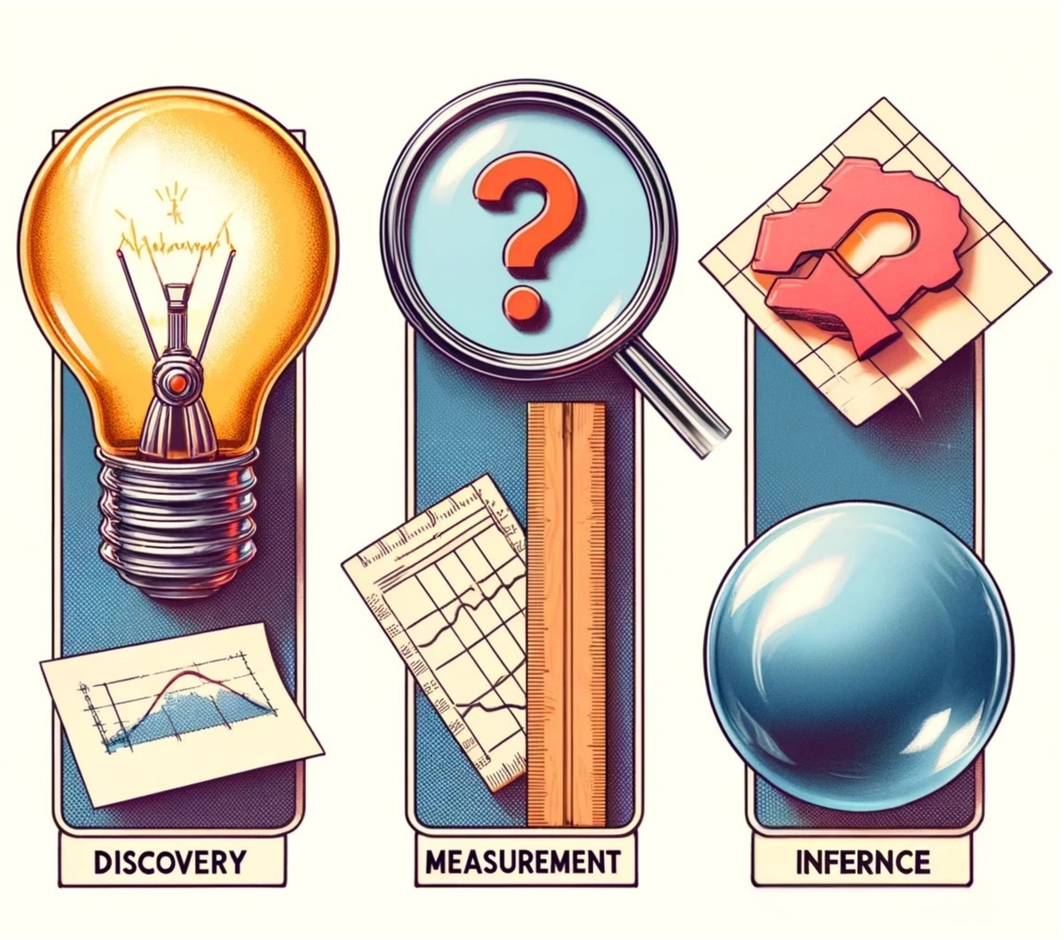
Via DALL-E

TEXT ANALYSIS ISN'T MAGIC (SORRY)

- Outputs are **estimates** and require **validation**.
  - Fidelity (how accurate is your measure).
  - Aggregation error (aggregated estimates are unbiased).
  - Face validity (are you measuring what you think you're measuring).
  - Hypothesis (convergent) validity (do your measures correlate with what they should).

THREE STAGES OF RESEARCH

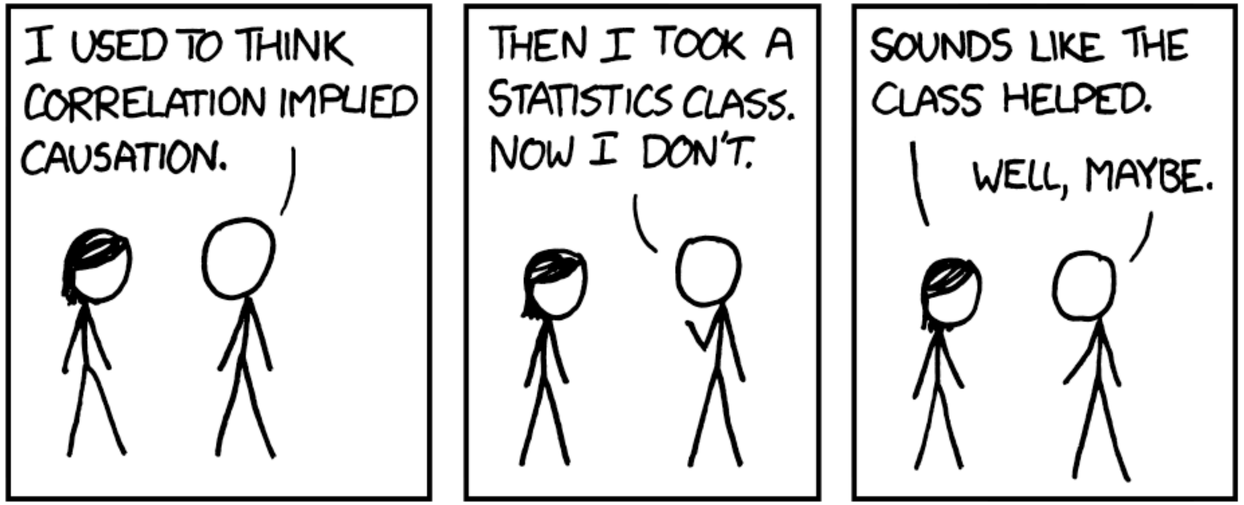
- Discovery: defining the research question, conceptualizing.
- Measurement: actually measuring the quantity we care about.
- Inference: prediction and causal inference from existing data.



Via DALL-E

PREDICTION VERSUS CAUSAL INFERENCE

- Prediction: Focused on relationships between variables **to predict an outcome.**
- Causal inference: Focused on relationships between variables **to tell stories about whether X causes Y.**



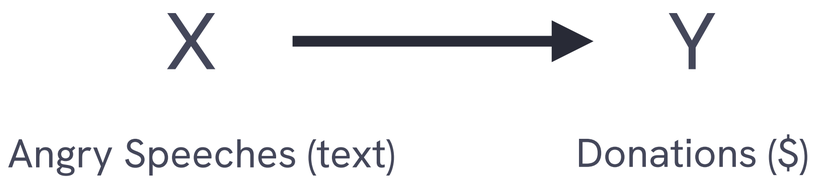
Via XKCD

HOW TO USE TEXT AS DATA

- Text as outcome (e.g., do candidates use more negative rhetoric when it's closer to election day?)

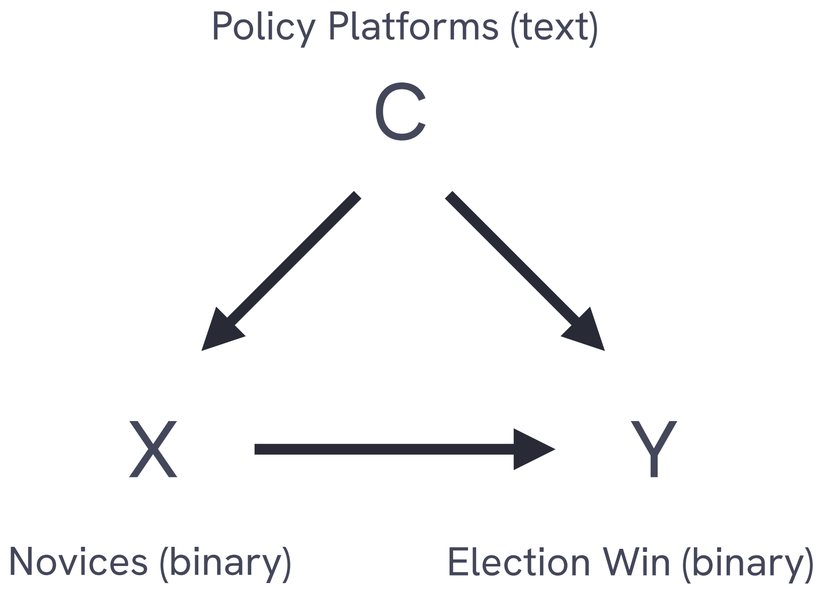
HOW TO USE TEXT AS DATA

- Text as outcome (e.g., do candidates use more negative rhetoric when it's closer to election day?)
- Text as treatment (e.g., do angrier speeches lead to an increase in donations to a politician?)



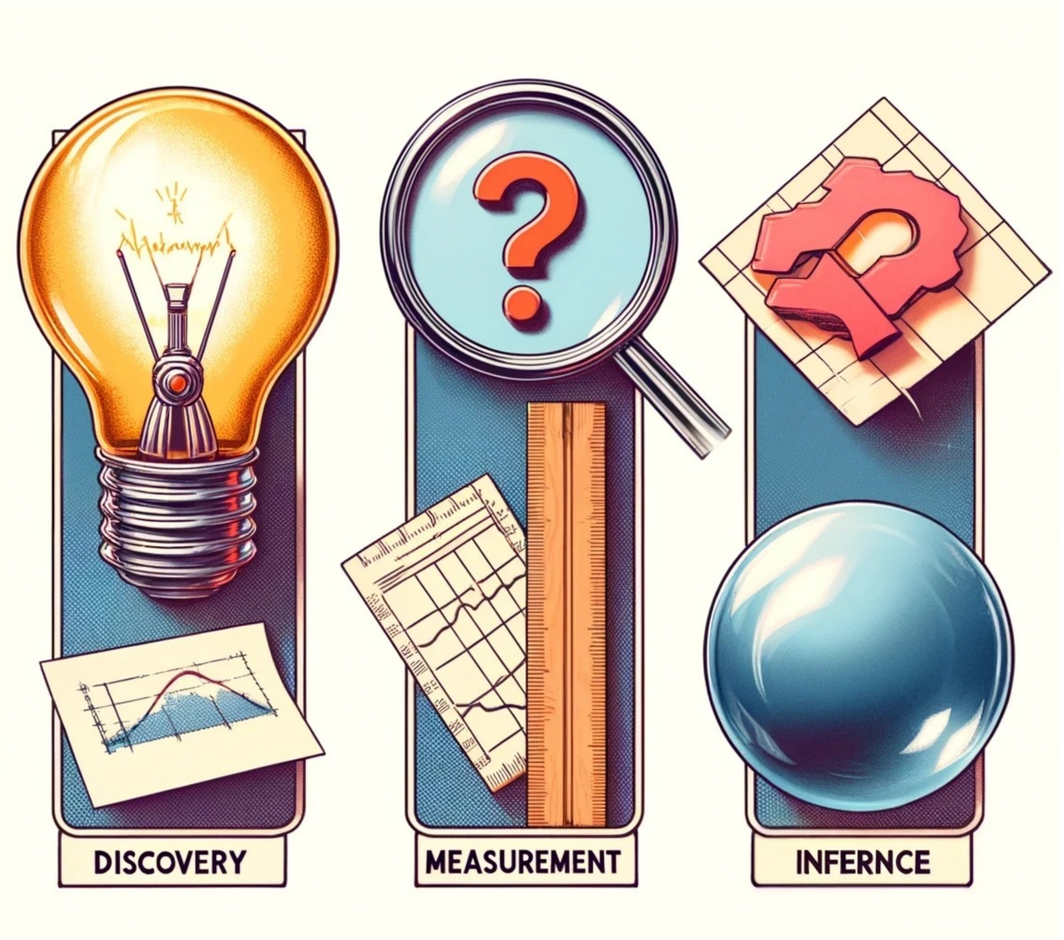
HOW TO USE TEXT AS DATA

- Text as outcome (e.g., do candidates use more negative rhetoric when it's closer to election day?)
- Text as treatment (e.g., do angrier speeches lead to an increase in donations to a politician?)
- Text as confounder (e.g., are political novices more likely to win elections, controlling for the policy platforms).



THREE STAGES OF RESEARCH

- Discovery: defining the research question, conceptualizing.
- Measurement: actually measuring the quantity we care about.
- Inference: prediction and causal inference from existing data.



via DALL-E