

# Intro to Data Science

What **R** we doing?

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/08/23

Slides Updated: 2023-08-22

# Agenda

1. Meet the instructor
2. Course Motivation
  - What is data science (DS) & why should we care?
3. Course Objectives
  - **Content:** Critical thinking, analysis, presentation
  - **Skills:** Computing and analysis in R
4. ChatGPT and data science
5. Course Expectations & Syllabus review

# Meet the instructor

- Education
  - PhD from NYU Politics in 2019
  - Postdocs at Princeton Niehaus & NYU CSMaP
- Published some things
  - Methods-ey: external validity [1](#), [2](#); measurement [3](#), [4](#)
  - Substantive: economics & populism [1](#); Covid-19 & U.S. politics [2](#), [3](#); IPE [4](#); academic naval-gazing [5](#)
  - Popular press: [1](#), [2](#), [Podcasts](#)
- Work
  - World Bank / IFC
  - MarketCast

# Meet the instructor

- Current research
  - YouTube + polarization
  - Twitter + misinformation
  - Telegram + white supremacists
- (Throughout the semester, I colorcode data science)

# Why are you here?



Suggested fights

20 last fights



## *DATA SCIENCE vs STEM*

200



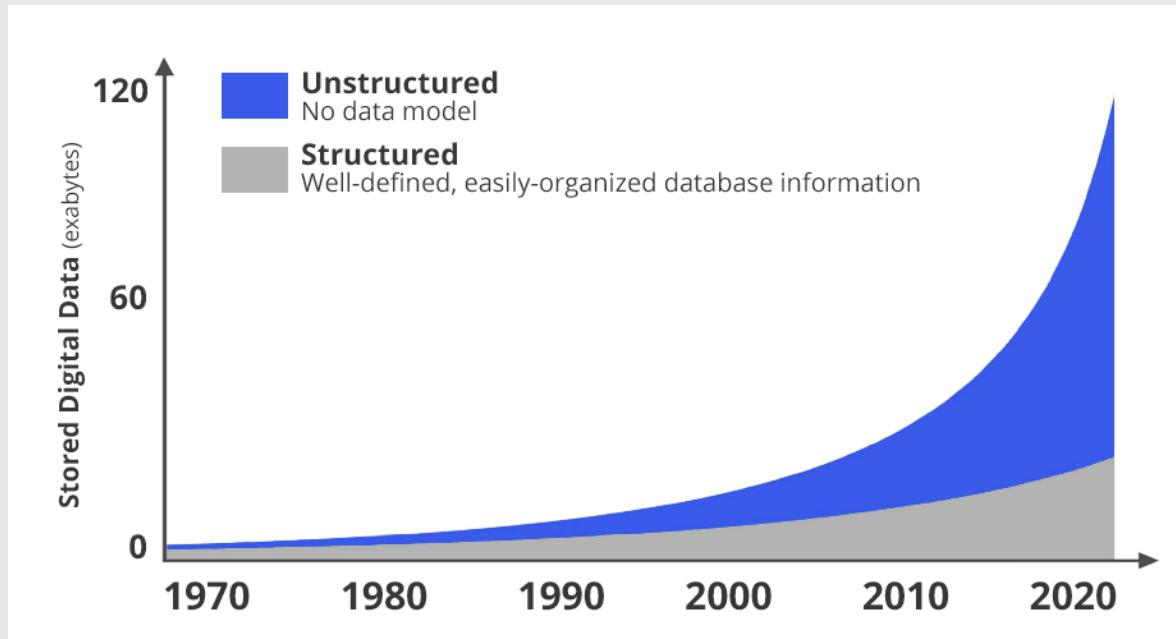
DATA SCIENCE

101

STEM

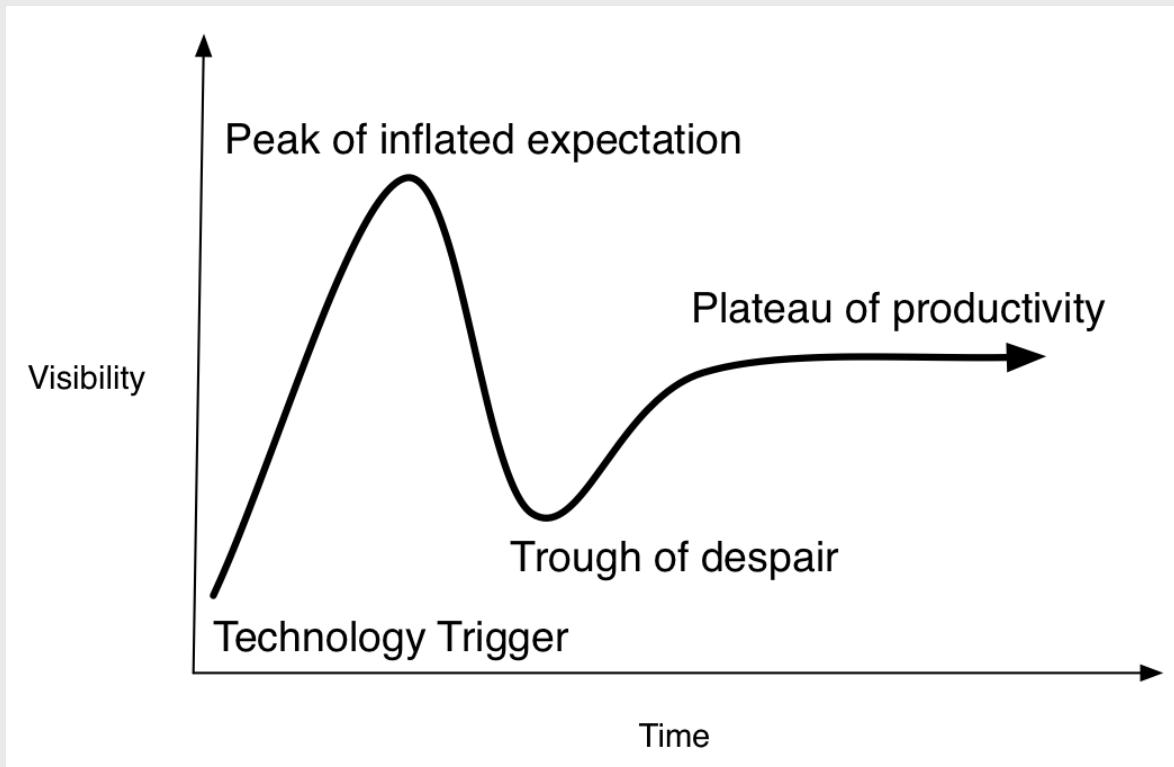
# Is this all just a fad?

- No



# Is this all just a fad?

- But there are faddish qualities



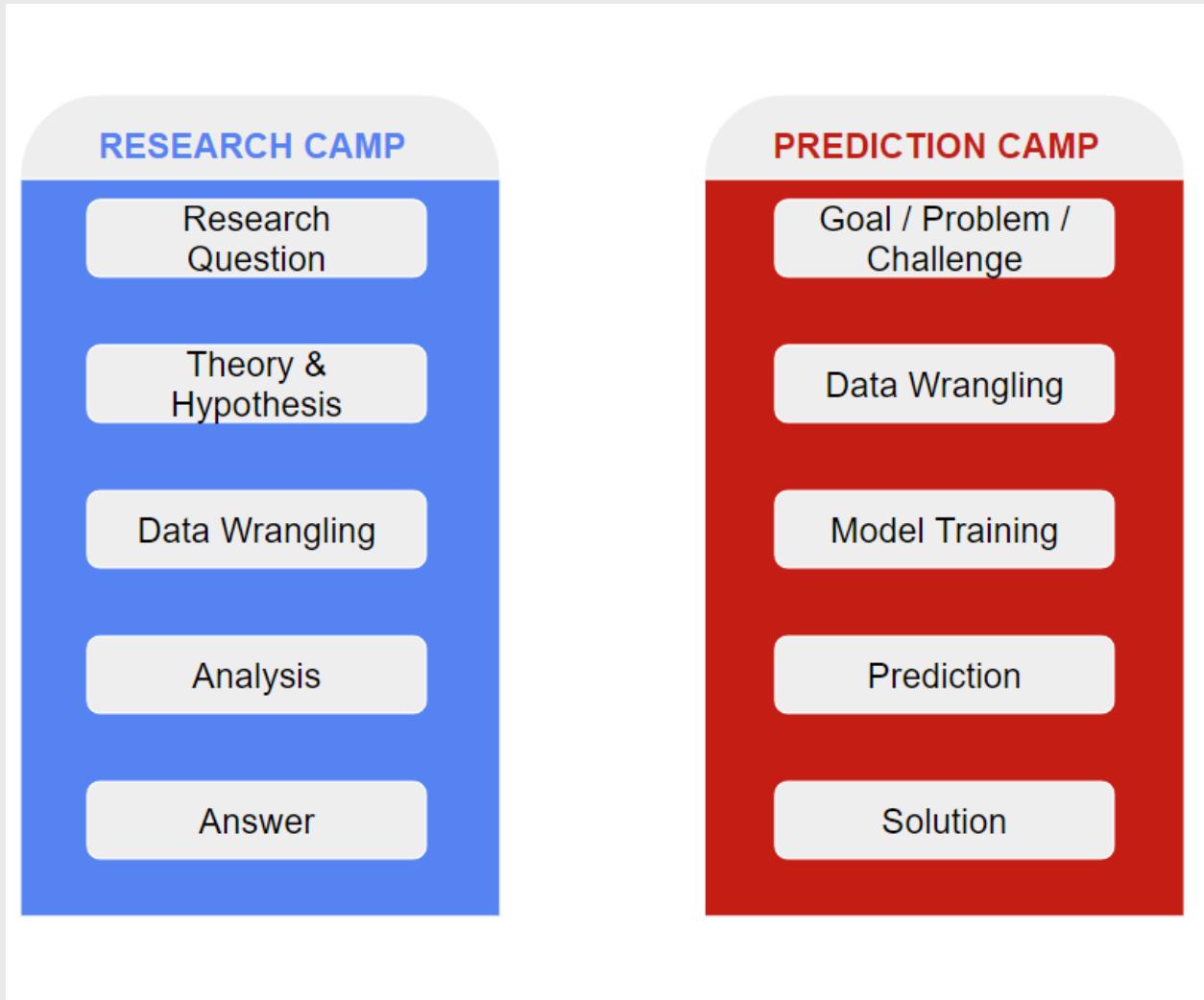
# So what IS data science?

- Split into two camps
  - 1. **Research** camp
    - Focused on **answering a research question**
    - Follows the "scientific method"
    - Goal: contribute to knowledge
    - Domain: academia
  - 2. **Prediction** camp
    - Focused on **making a prediction**
    - Typically unconcerned with theory or *why* a model works
    - Goal: inform a decision / policy
    - Domain: private sector

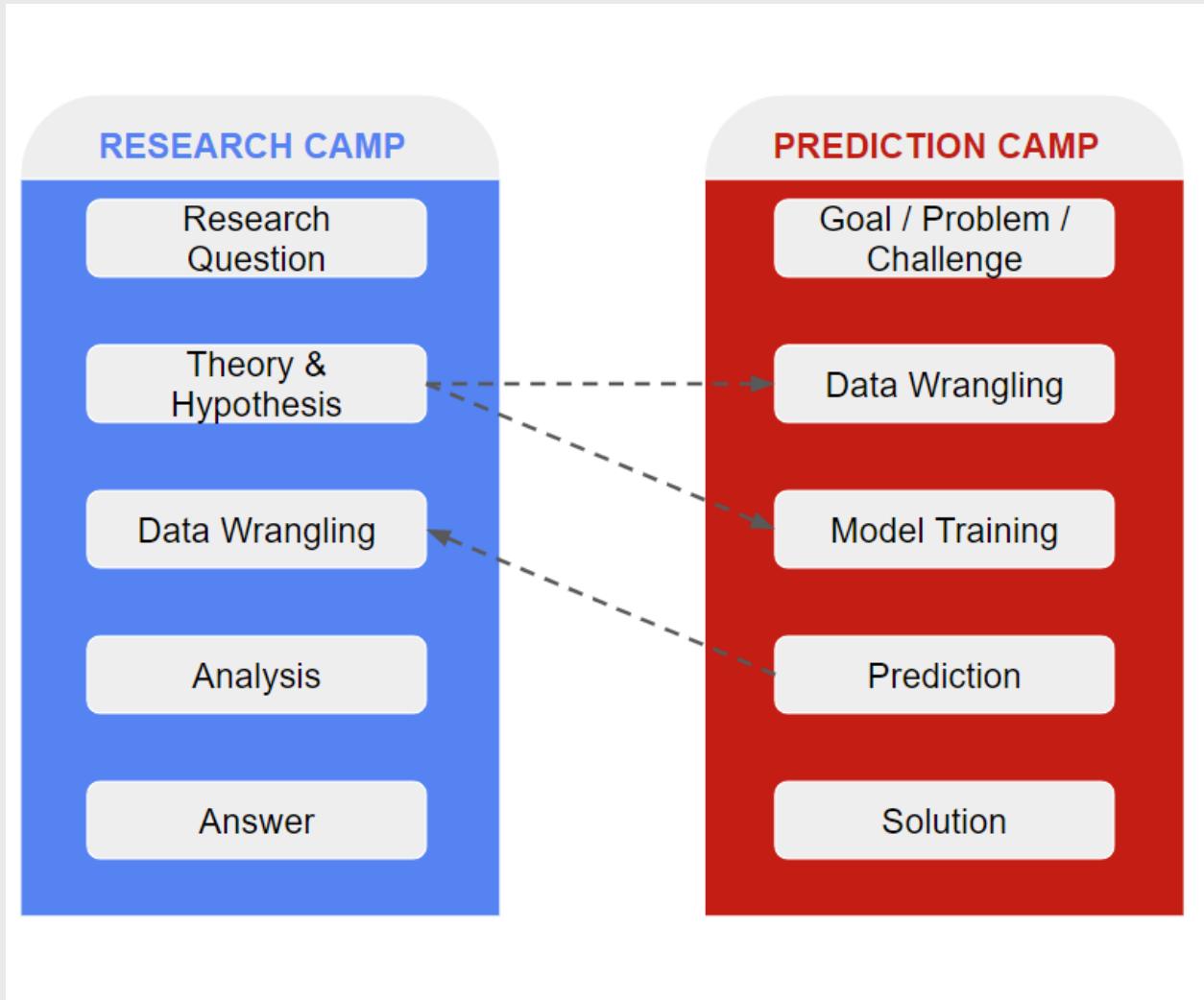
# The Two Camps



# The Two Camps

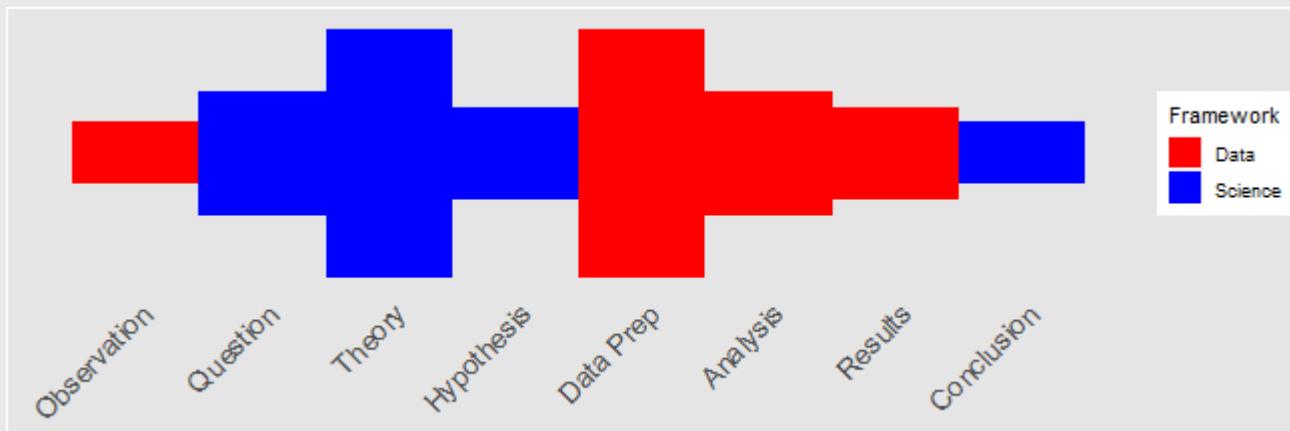


# The Two Camps



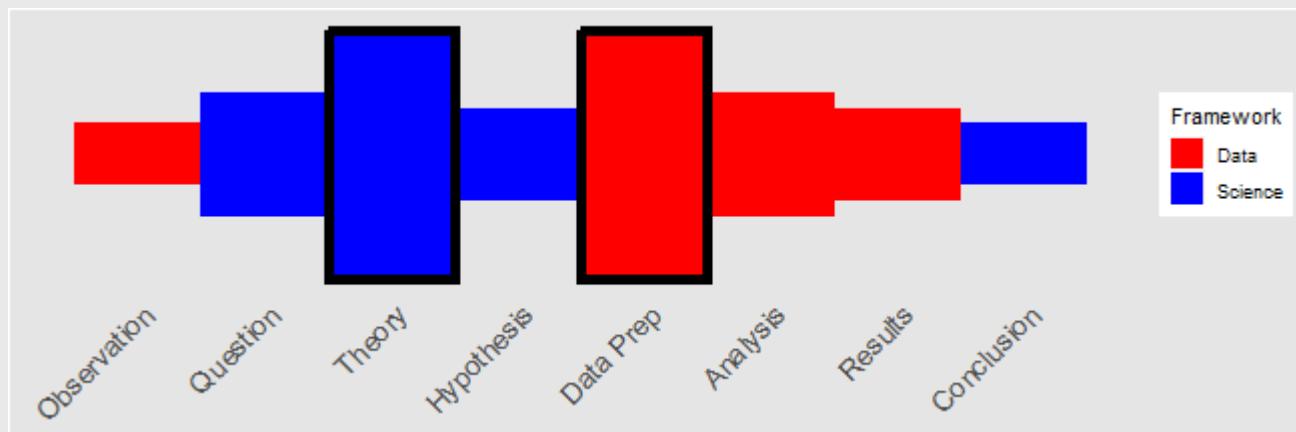
# Research Camp

- The scientific method
  1. Observation → Question
  2. Theory → Hypothesis
  3. Data Collection / Wrangling → Analysis
  4. Results → Conclusion



# Research Camp

- The scientific method
  1. Observation → Question
  2. Theory → Hypothesis
  3. Data Collection / Wrangling → Analysis
  4. Results → Conclusion



# Research Camp

## Echo Chambers, Rabbit Holes, and Algorithmic Bias: How YouTube Recommends Content to Real Users

Megan A. Brown,<sup>1†</sup> James Bisbee,<sup>1</sup> Angela Lai,<sup>1,4</sup>  
Richard Bonneau,<sup>1,3,4</sup> Jonathan Nagler,<sup>1,2,4</sup> Joshua A. Tucker<sup>1,2,4</sup>

<sup>1</sup>Center for Social Media and Politics, New York University

<sup>2</sup>Politics Department, New York University

<sup>3</sup>Biology Department, New York University

<sup>4</sup>Center for Data Science, New York University

<sup>†</sup>To whom correspondence should be addressed: [meganbrown@nyu.edu](mailto:meganbrown@nyu.edu)

August 23, 2023

Abstract

# Research Camp

## 1. Observation → Question

- Observation is facilitated by **data** (Descriptive analysis)



# Research Camp

## 1. Observation → Question

- Observation is facilitated by **data** (Descriptive analysis)

The image shows a screenshot of a CBS News video player. The main content area features a split-screen interview. On the left, a Black male anchor in a blue patterned shirt and dark tie looks directly at the camera. On the right, a white female anchor in a bright pink V-neck top also looks at the camera. Below the anchors is a red horizontal bar with the text 'PRES. USES SOCIAL MEDIA TO DENOUNCE "RIGGED" ELECTION'. The CBS News logo is visible in the bottom right corner of the video frame. At the bottom of the screen, there is a navigation bar with icons for play, volume, and a timestamp '0:06 / 11:40'. Below the video, there is a section for 'U.S. elections' with a note from AP about Joe Biden's victory. A 'SHOW ME' button is also present. To the right of the video, there is a sidebar titled 'Up next' which lists several other news clips from various networks like FOX, CBS, NBC, and CNN, each with a thumbnail, title, and view count.

PRES. USES SOCIAL MEDIA TO DENOUNCE "RIGGED" ELECTION

LIVE CBSN

SENATE HEARING ON RUSSIAN INTERFERENCE IN 2016 ELECTION  
cbsnews.com/hearing

U.S. elections

Robust safeguards help ensure the integrity of elections and results. Learn more

Trump continues to push false claims of election fraud in Facebook video

12,798 views • Dec 3, 2020

CBS News 3.29M subscribers

President Trump posted a long Facebook video where he repeatedly denounced the November election as "rigged," even though Attorney General William Barr said the Justice Department has seen no evidence of election fraud. CBS News White House correspondent Paula Reid joins CBSN's

942 182 SHARE SAVE

SUBSCRIBE

Up next

HOW IT STARTED: Senate Hearing On FBI Investigation...  
NewsNOW from FOX 47K views • 3 hours ago

Mary Trump Says Trump's Legal Battles Could Prevent a 2024...  
The View 5.2K views • 1 hour ago

Trump WH, State Dept. Push Ahead With Holiday Parties...  
MSNBC 9.9K views • 2 hours ago

Black Home Ownership - If You Don't Know, Now You Know...  
The Daily Show with Trevor Noah 119K views • 3 hours ago

President Risks Handing Democrats The Senate By...  
The Late Show with Stephen Colbert 2.1M views • 1 day ago

'MOST IMPORTANT SPEECH'  
Trump gives 'most important speech he's made, calls for Tu...  
MSNBC 340K views • 16 hours ago

Wisconsin Supreme Court Rejects Trump Lawsuit | MTP...  
MSNBC 15K views • 56 minutes ago

Attorney General William Barr's job in jeopardy  
ABC News 57K views • 5 hours ago

Mary Trump Says It's 'Impossible' for Trump 'to...  
The View 7.5K views • 1 hour ago

A Fool: MAGA Fans Turn On Barr After Debunking Trump's...  
MSNBC 875K views • 19 hours ago

16 / 75

# Research Camp

## 1. Observation → Question

- Observation is facilitated by **data** (Descriptive analysis)

The image shows a screenshot of a CBS News video player. The main content area features a split-screen interview. On the left, a Black male anchor in a white shirt and dark tie looks directly at the camera. On the right, a female anchor with long dark hair, wearing a pink top, also looks at the camera. Below the anchors is a red horizontal bar with the text 'PRES. USES SOCIAL MEDIA TO DENOUNCE "RIGGED" ELECTION'. The CBS News logo is visible in the bottom right corner of the video frame. At the bottom of the screen, there is a navigation bar with icons for play, volume, and time (0:06 / 11:40). A small note below the play bar says 'U.S. elections'. The video has 12,798 views and was posted on Dec 3, 2020. A 'SUBSCRIBE' button is located in the bottom right corner of the video frame.

Up next

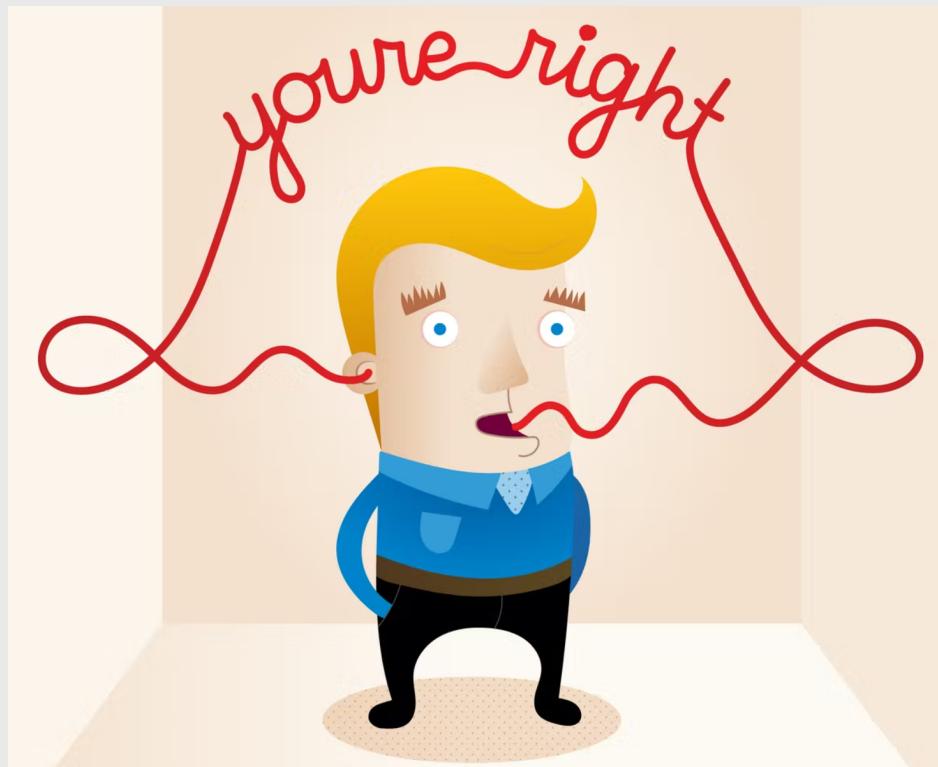
- HOW IT STARTED: Senate Hearing On FBI Investigation I...  
ABC News  
47K views • 3 hours ago
- Attorney General William Barr's job in jeopardy  
ABC News  
57K views • 5 hours ago
- The Full Story of Trump and COVID-19 | NowThis  
NowThis News  
1.8M views • 1 month ago
- Live: New York Gov. Andrew Cuomo Holds Briefing On Cov...  
NBC News  
9.2K watching
- See Bernie Sanders' reaction to Trump floating 2024...  
CNN  
963K views • 18 hours ago
- Mary Trump Says Trump's Legal Battles Could Prevent a 2024...  
The View  
5.5K views • 1 hour ago
- Trump releases Facebook video full of false claims about...  
CBS News  
14K views • 4 hours ago
- Election Lawsuits Meltdown... With Prejudice!  
LegalEagle  
996K views • 4 days ago
- Second Georgia Senate election hearing  
11Alive  
9K watching
- 'A Fool': MAGA Fans Turn On Barr After Debunking Trump's...  
MSNBC  
875K views • 19 hours ago

17 / 75

# Research Camp

## 1. **Observation** → **Question**

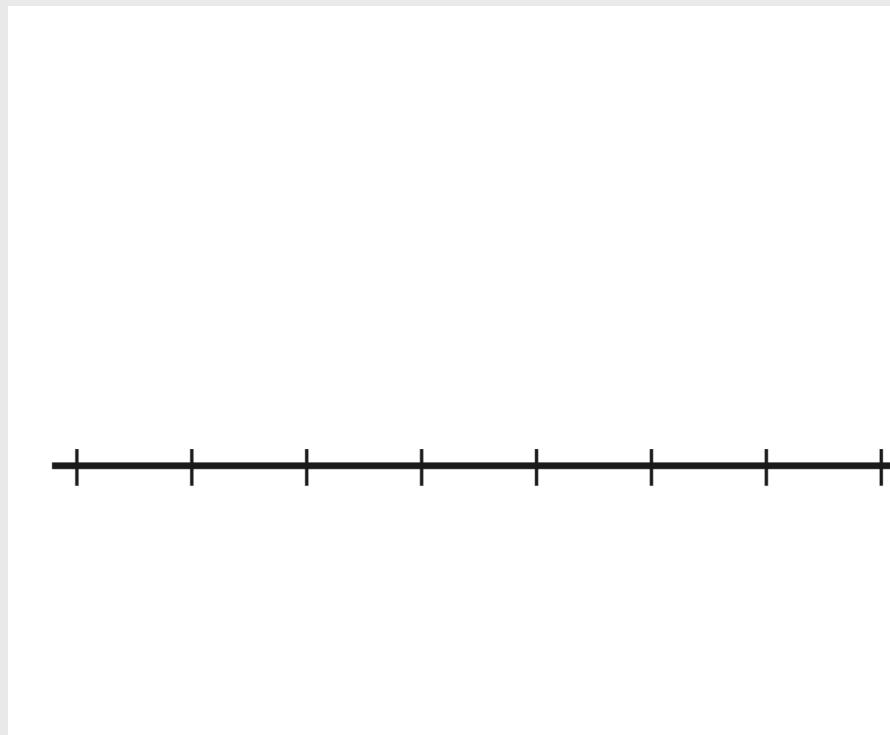
- The question pertains to science
- I.e., does YouTube's algorithm put users into "echo chambers"?



# Research Camp

## 2. Theory → Hypothesis

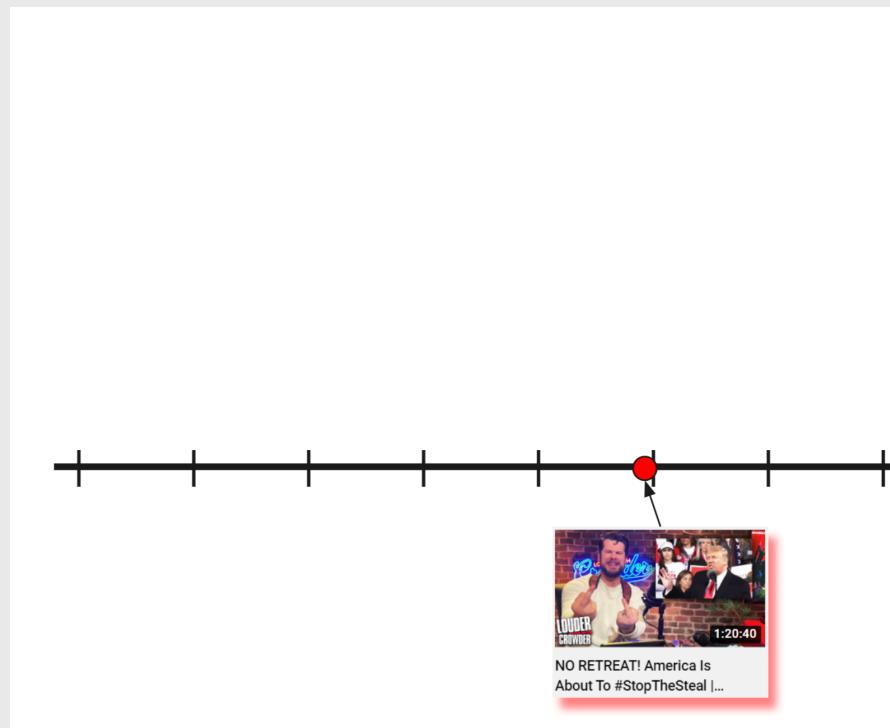
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

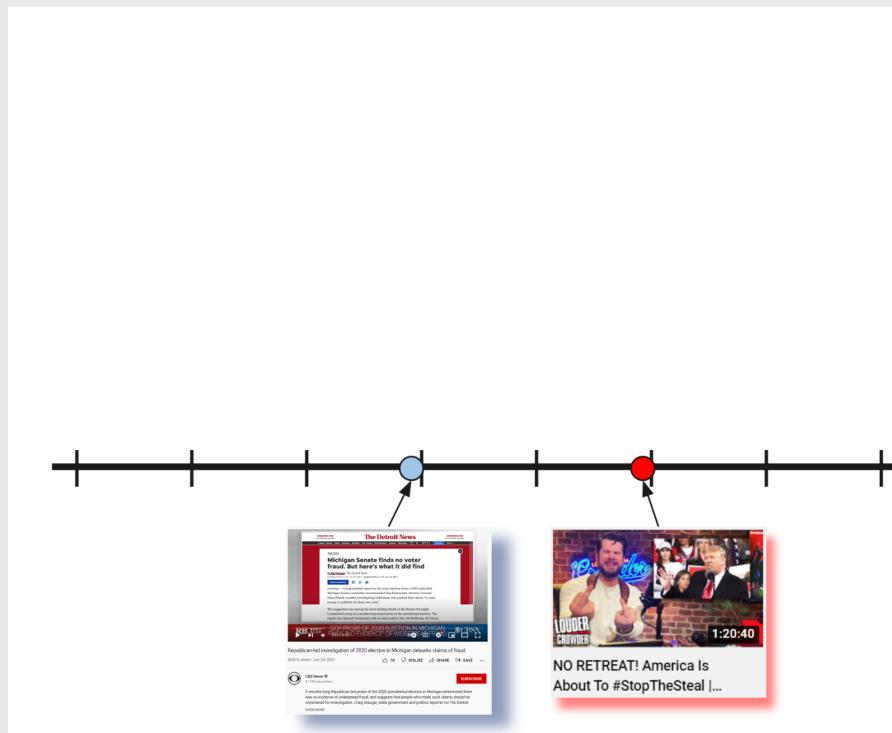
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

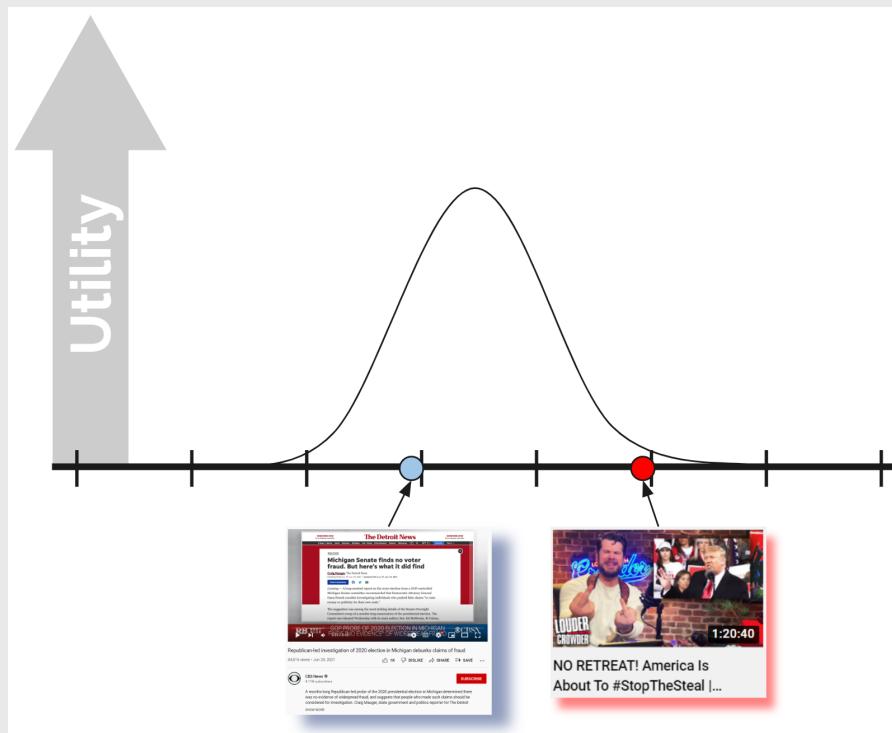
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

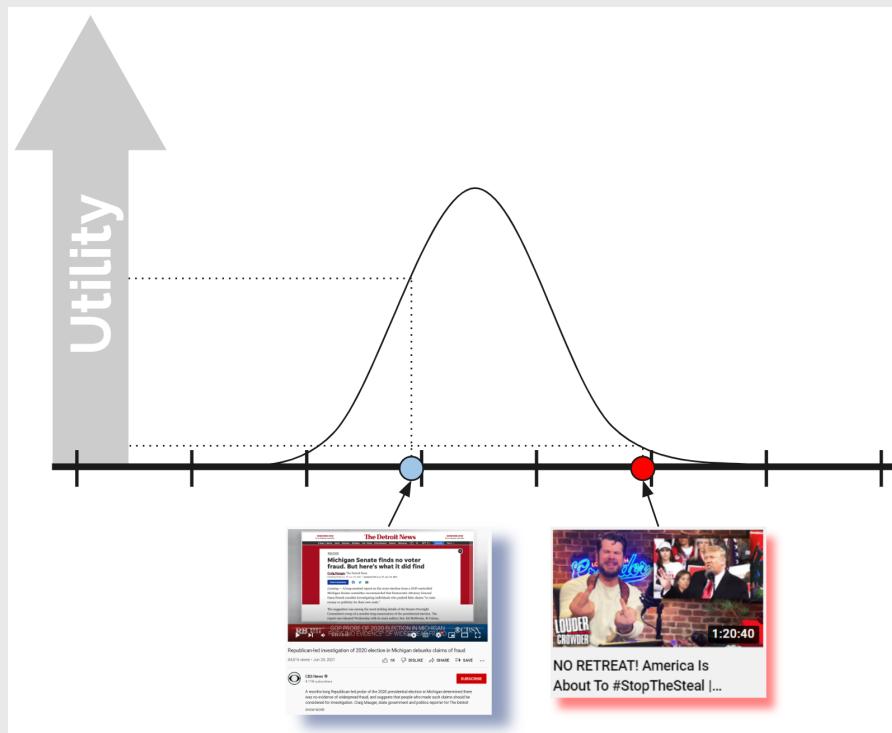
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict
- YouTube wants users to watch more videos

**Deep Neural Networks for YouTube Recommendations**

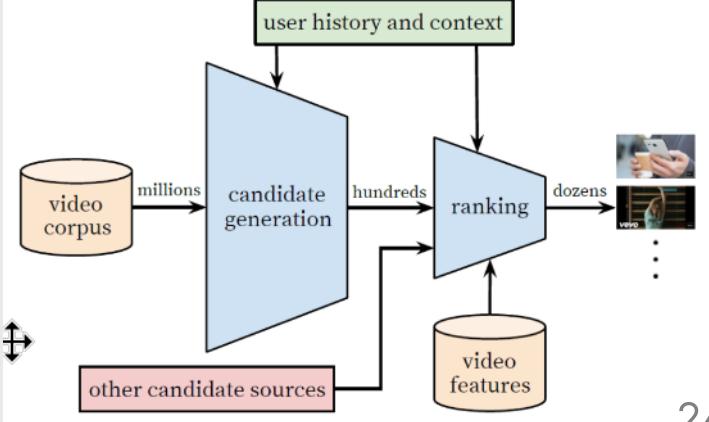
Paul Covington, Jay Adams, Emre Sargin  
Google  
Mountain View, CA  
[{pcovington,jka,msargin}@google.com](mailto:{pcovington,jka,msargin}@google.com)

**ABSTRACT**  
YouTube represents one of the largest scale and most sophisticated industrial recommendation systems in existence. In this paper, we describe the system at a high level and focus on the dramatic performance improvements brought by deep learning. The paper is split according to the classic two-stage information retrieval dichotomy: first, we detail a deep candidate generation model and then describe a separate deep ranking model. We also provide practical lessons and insights derived from designing, iterating and maintaining a massive recommendation system with enormous user-facing impact.

**Keywords**  
recommender system; deep learning; scalability

**1. INTRODUCTION**  
YouTube is the world's largest platform for creating, sharing and discovering video content. YouTube recommendations are responsible for helping more than a billion users



$$P(w_t = i | U, C) = \frac{e^{v_i, u}}{\sum_{j \in V} e^{v_j, u}}$$


The diagram illustrates the YouTube recommendation system architecture. It starts with a large "video corpus" (millions of videos) which feeds into a "candidate generation" module. This module outputs "hundreds" of candidates to a "ranking" module. The "ranking" module outputs "dozens" of recommended videos. The "ranking" module receives input from "user history and context" and "video features". There is also a feedback loop from the "ranking" module back to "candidate generation". Additionally, "other candidate sources" are integrated into the system.

# Research Camp

## 2. Theory → Hypothesis

- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict
- YouTube wants users to watch more videos
- Hypotheses fall out naturally from well-done theory
- **H1:** *YouTube's recommendation algorithm should suggest liberal content to liberals and conservative content to conservatives.*

# Research Camp

## 3. Data Collection / Wrangling → Analysis

- Data collection separates "Data Science"...
- ...from "Science, with data"

- Recruit YouTube users to install [extension](#)



YouTube Recommendation Downloader

Offered by: csmappplugin

# Research Camp

## 3. Data Collection / Wrangling → Analysis

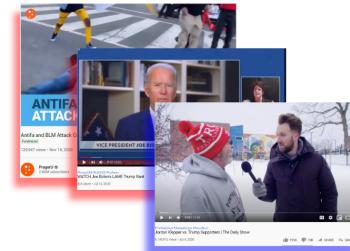
- Data collection separates "Data Science"...
- ...from "Science, with data"

- Recruit YouTube users to install **extension**
- Start on randomly assigned **seed video**



YouTube Recommendation Downloader

Offered by: csmappplugin

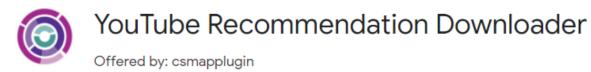


# Research Camp

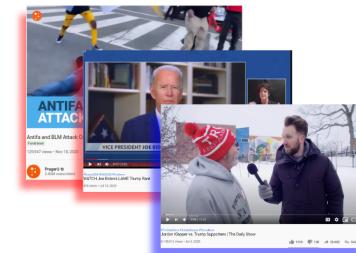
## 3. Data Collection / Wrangling → Analysis

- Data collection separates "Data Science"...
- ...from "Science, with data"

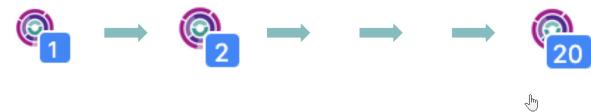
- Recruit YouTube users to install **extension**



- Start on randomly assigned **seed video**



- Follow **traversal rule** to select recommended video



# Research Camp

## 3. Data Collection / Wrangling → Analysis

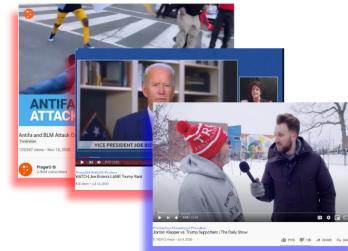
- Data collection separates "Data Science"...
- ...from "Science, with data"

- Recruit YouTube users to install **extension**



YouTube Recommendation Downloader

Offered by: csmappplugin



- Start on randomly assigned **seed video**

- Follow **traversal rule** to select recommended video

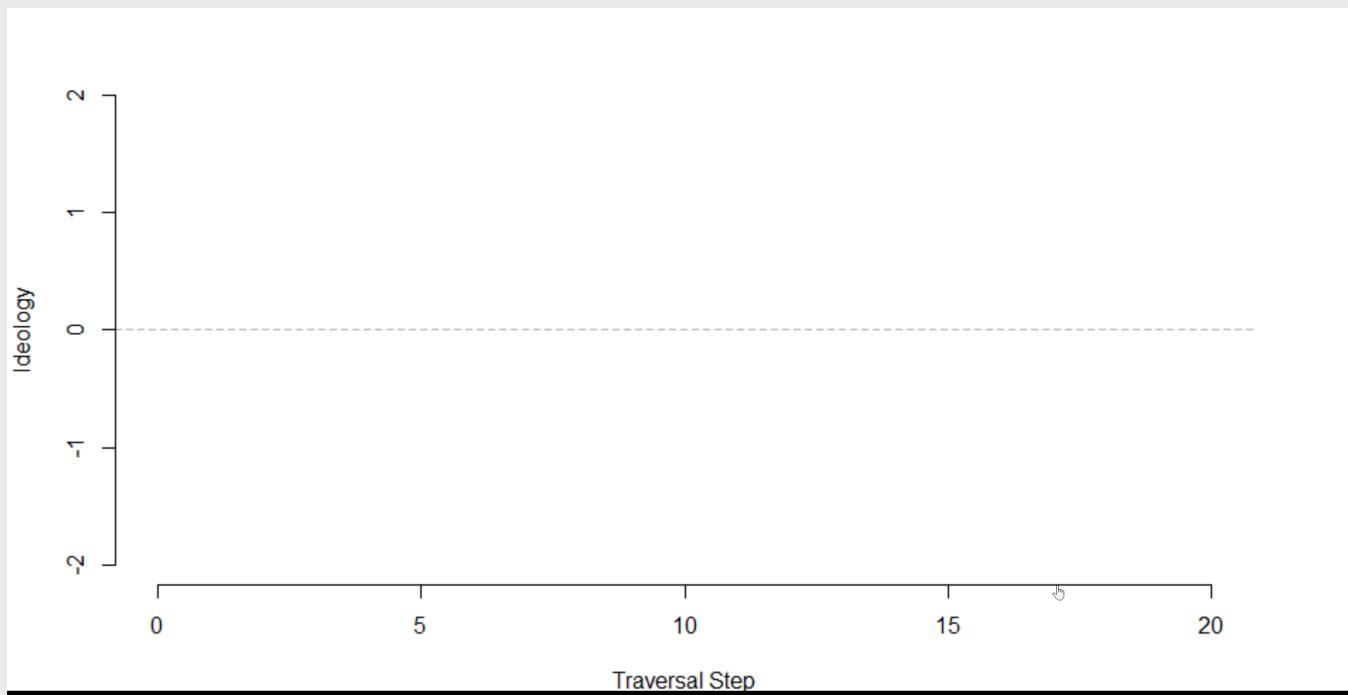


- Short **survey** on demographics, politics, and **BELIEFS ABOUT THE 2020 ELECTION**

# Research Camp

## 3. Data Collection / Wrangling → Analysis

- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

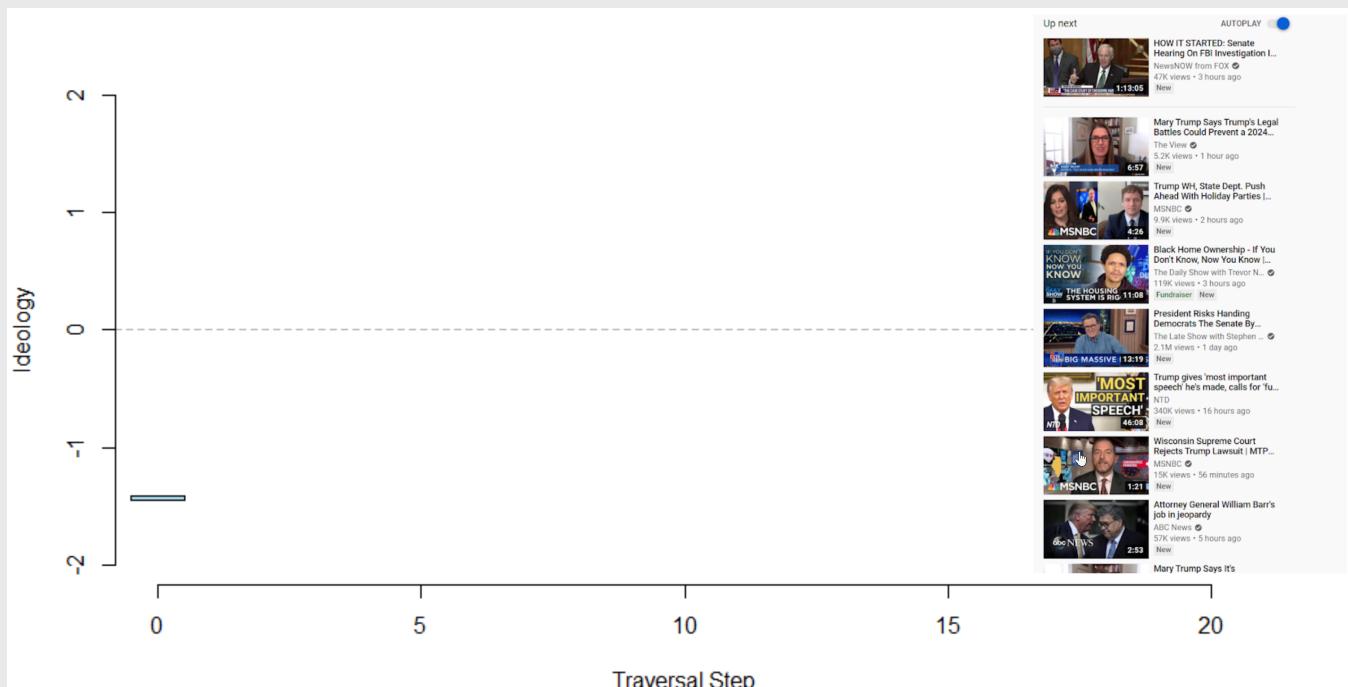
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

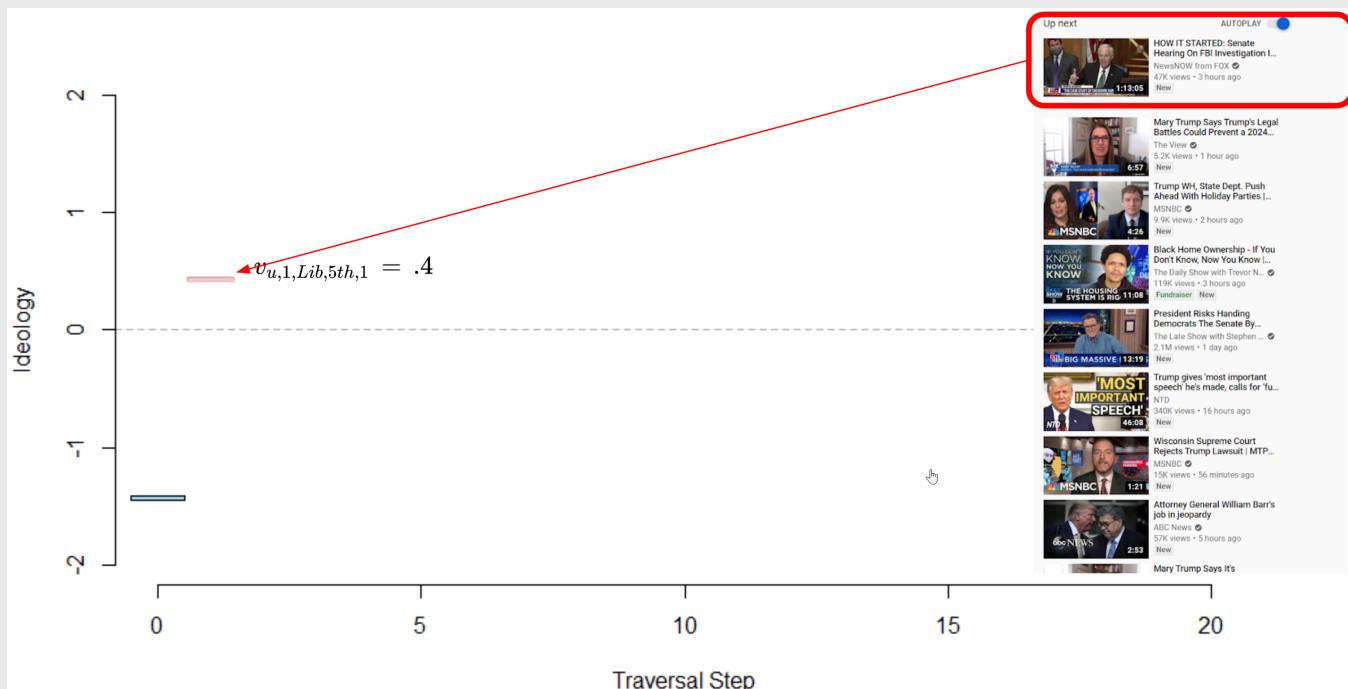
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

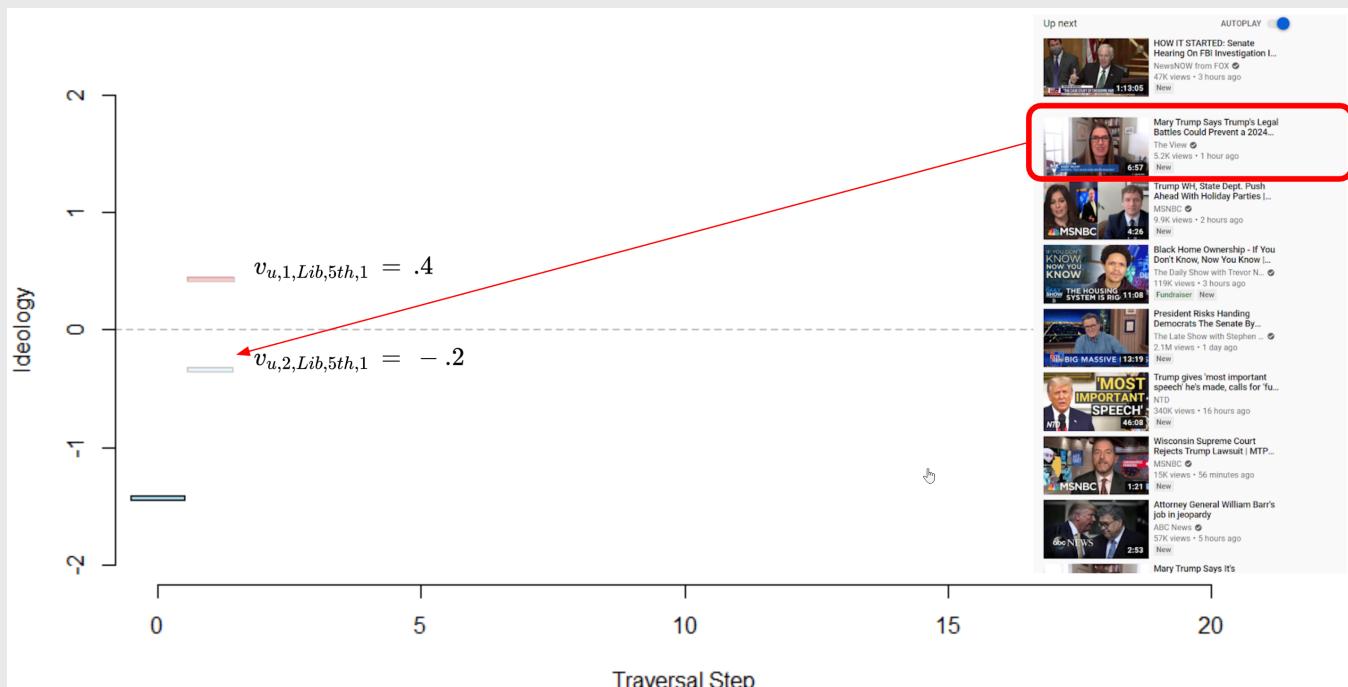
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

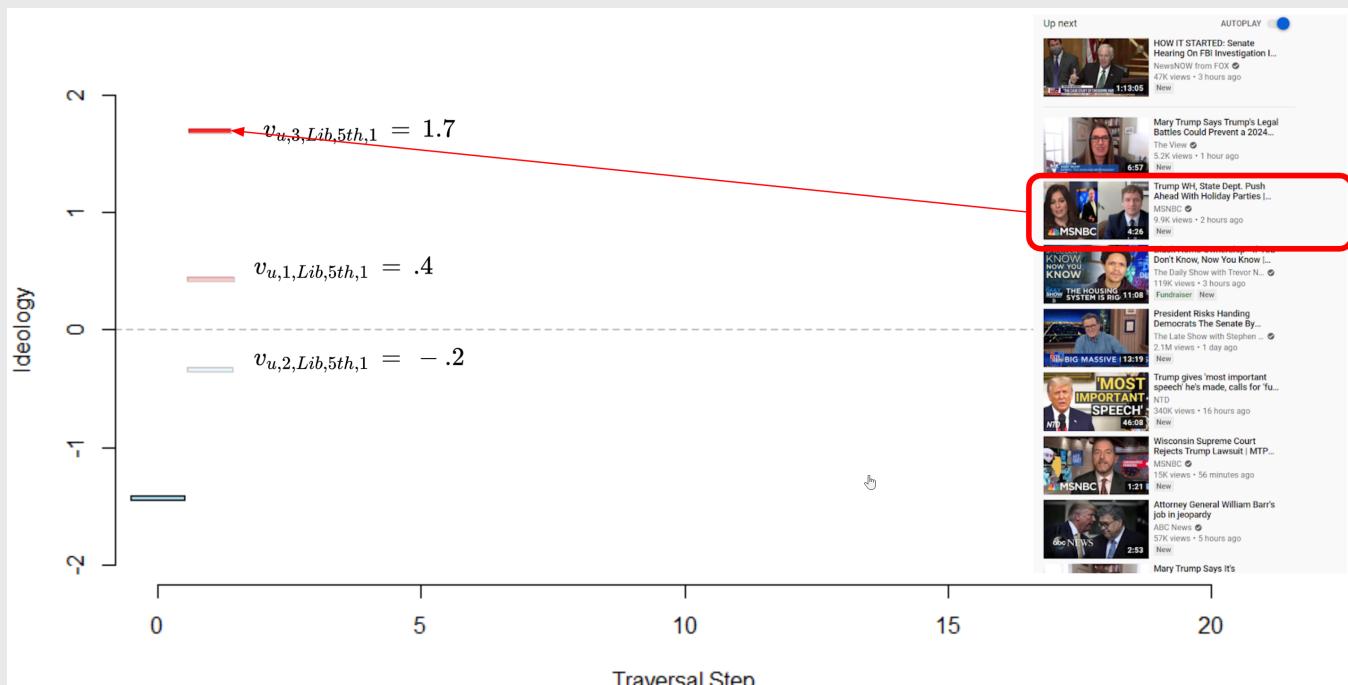
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

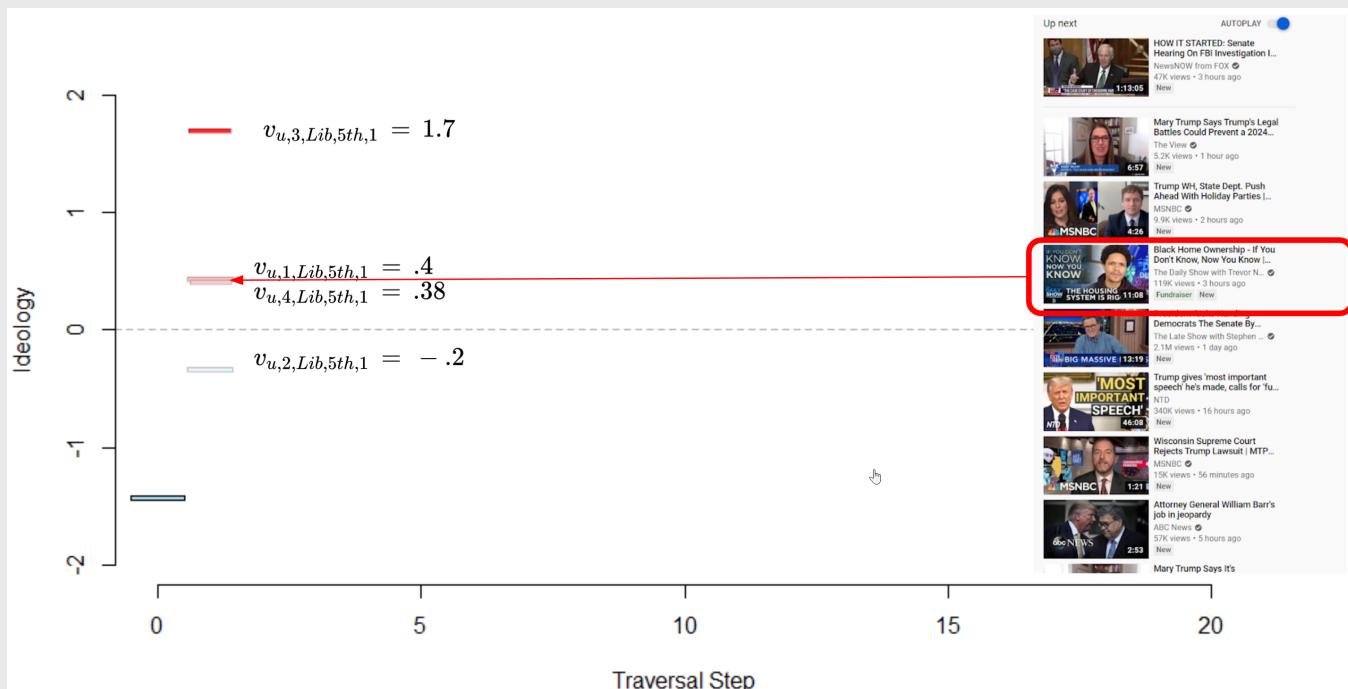
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

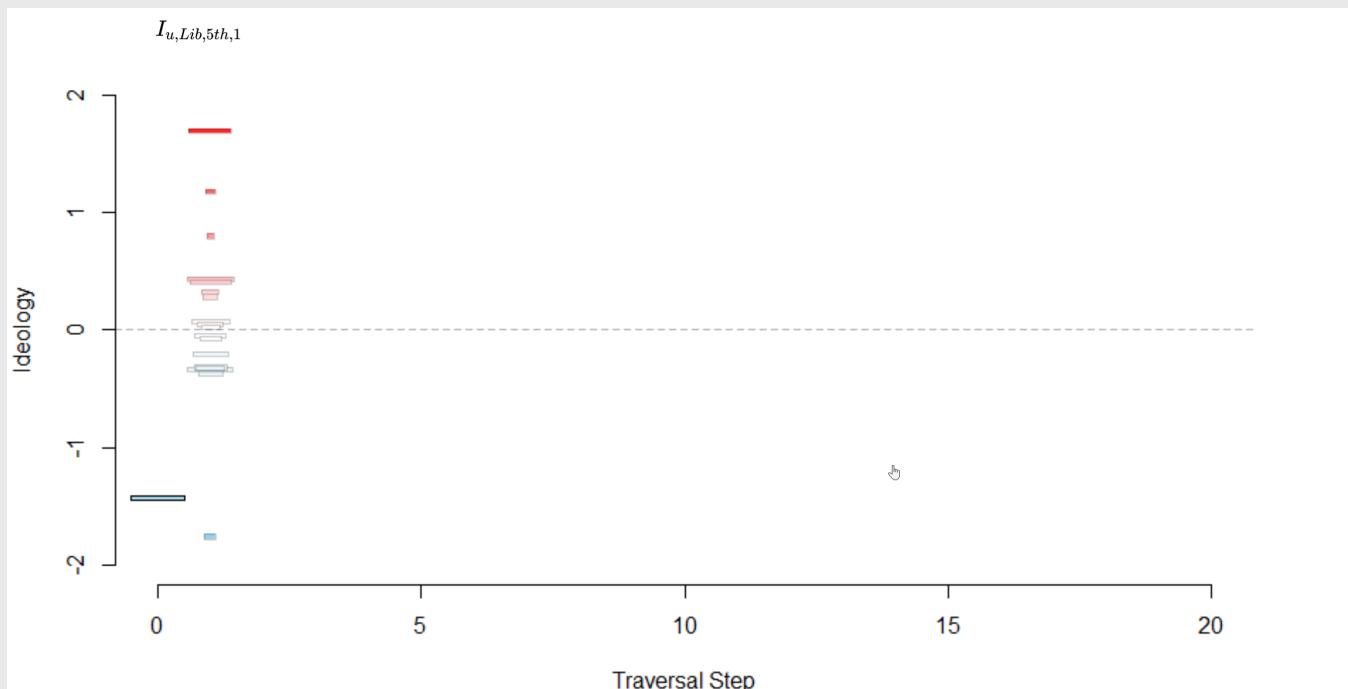
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

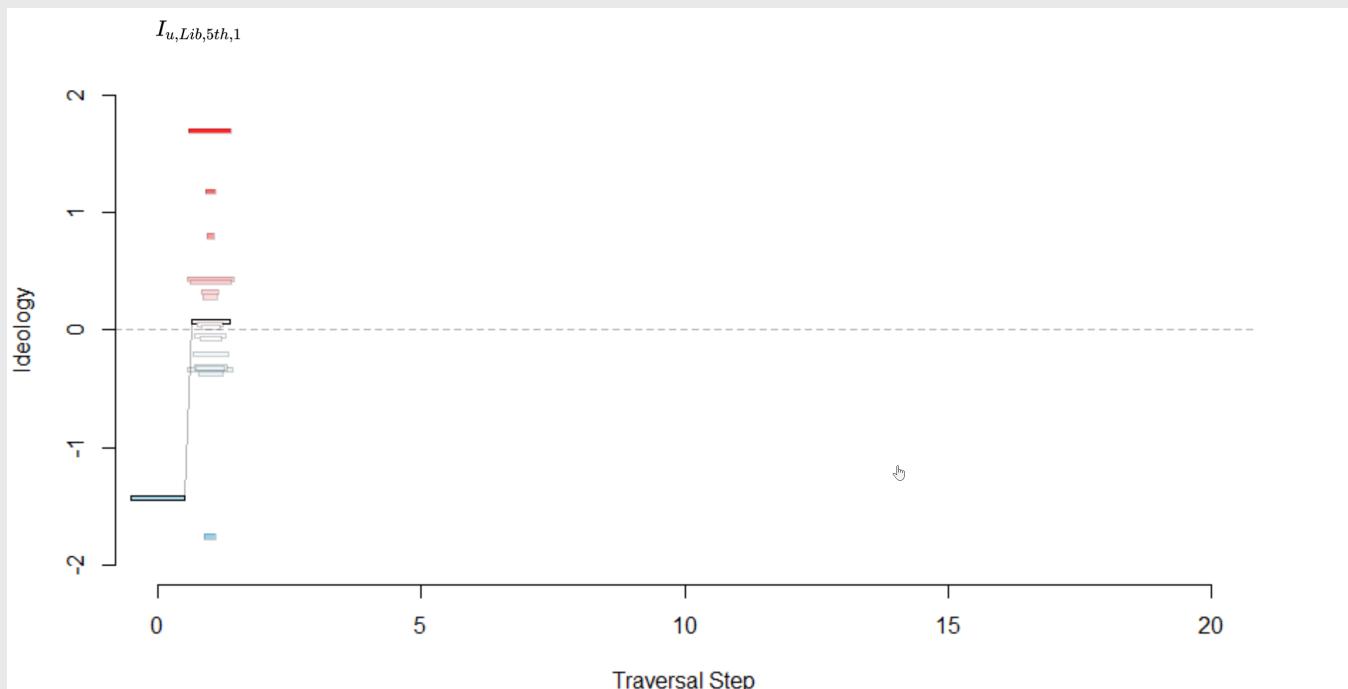
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

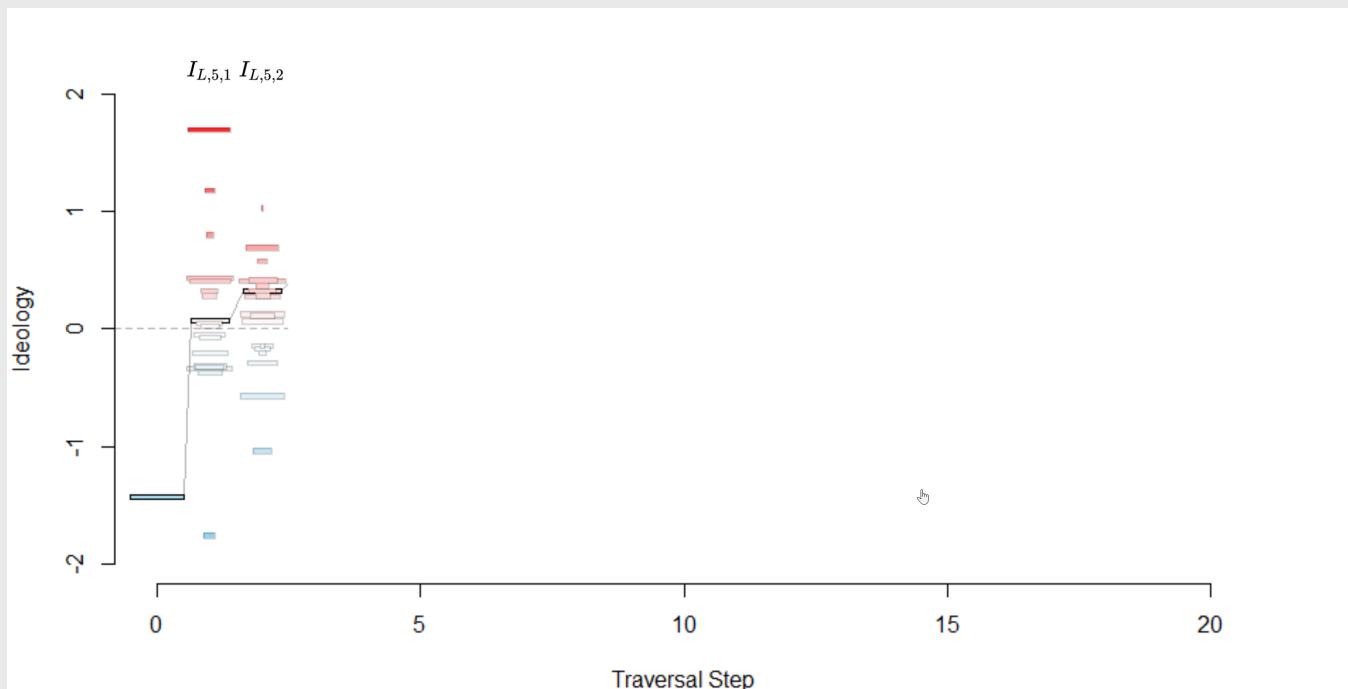
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

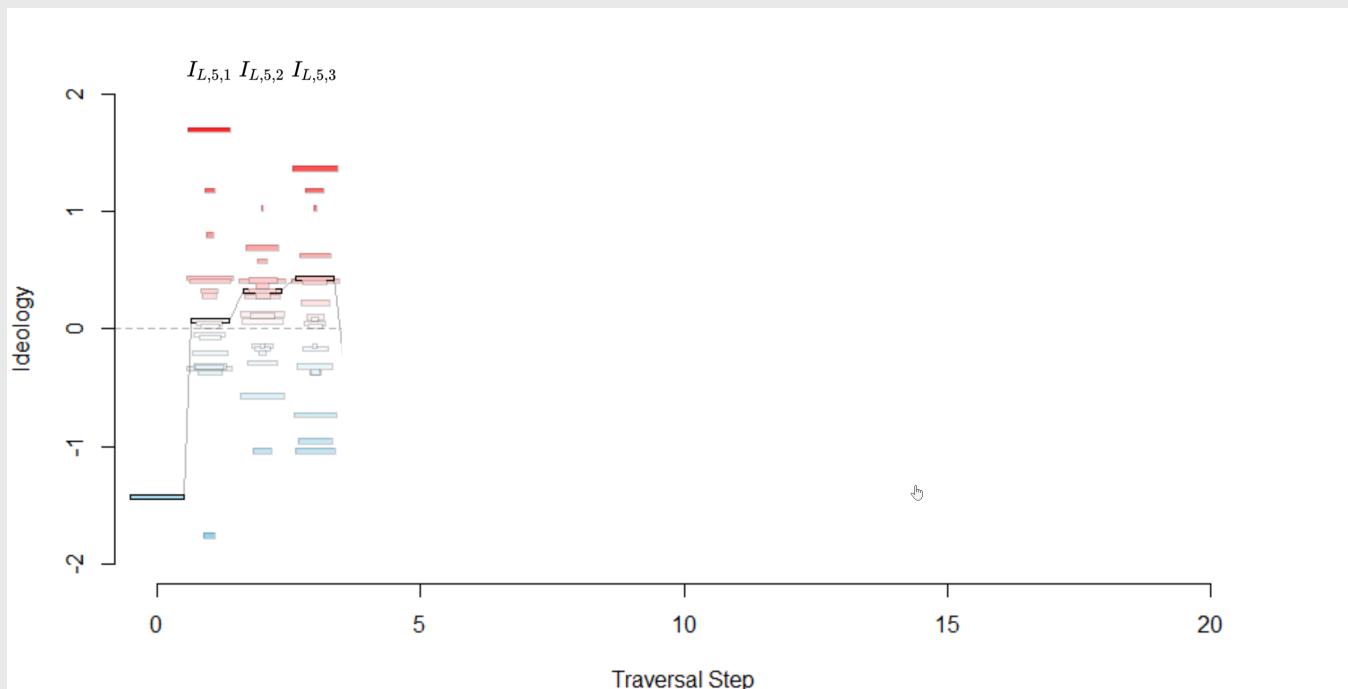
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

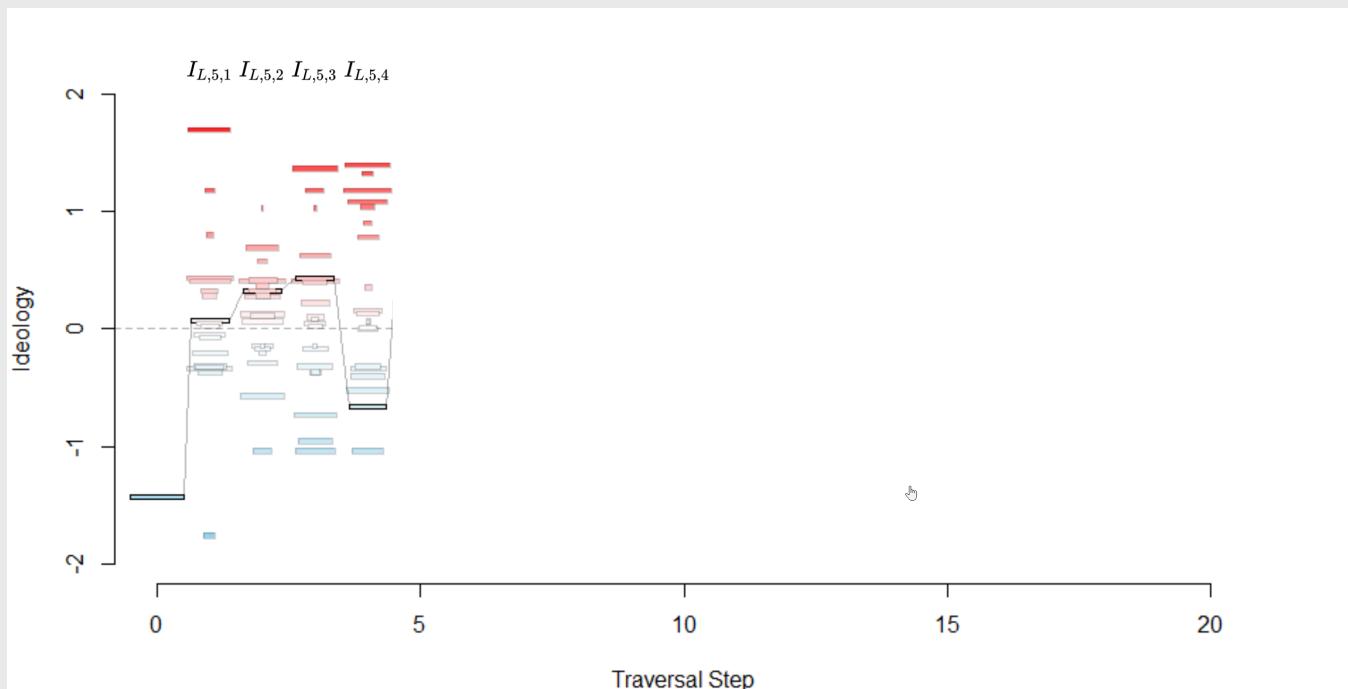
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

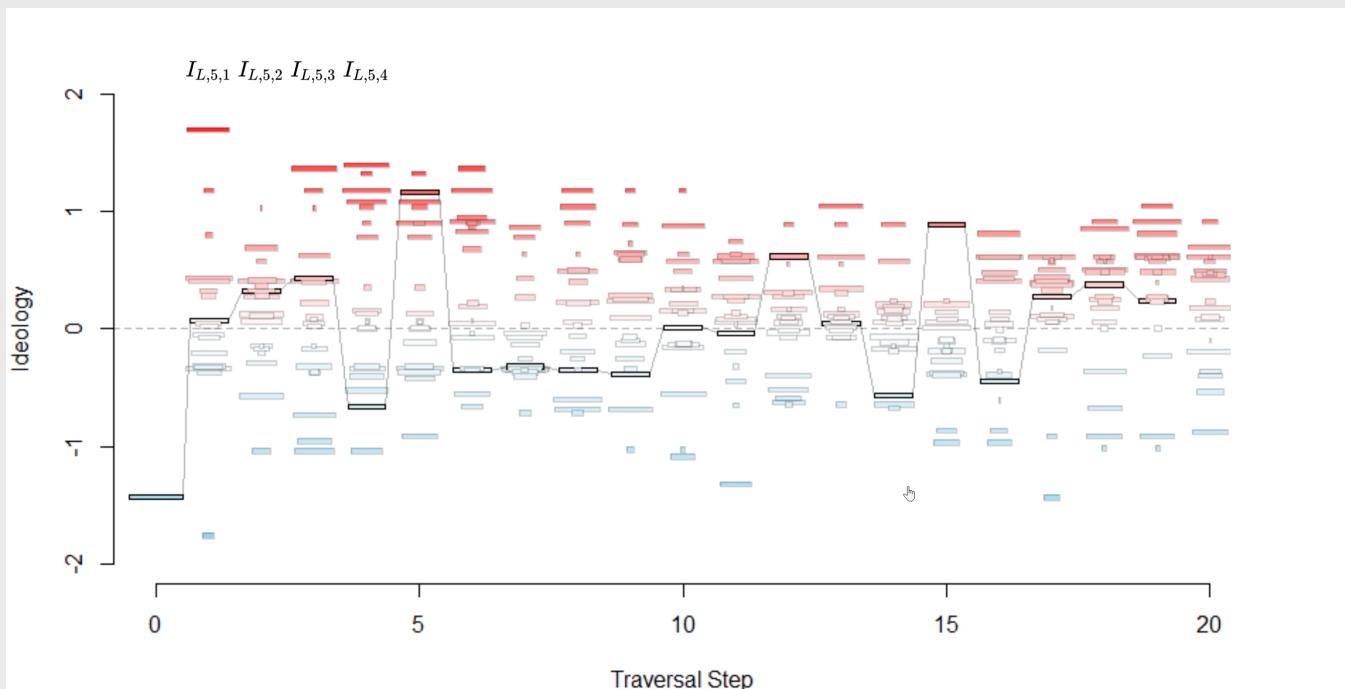
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection / Wrangling → Analysis

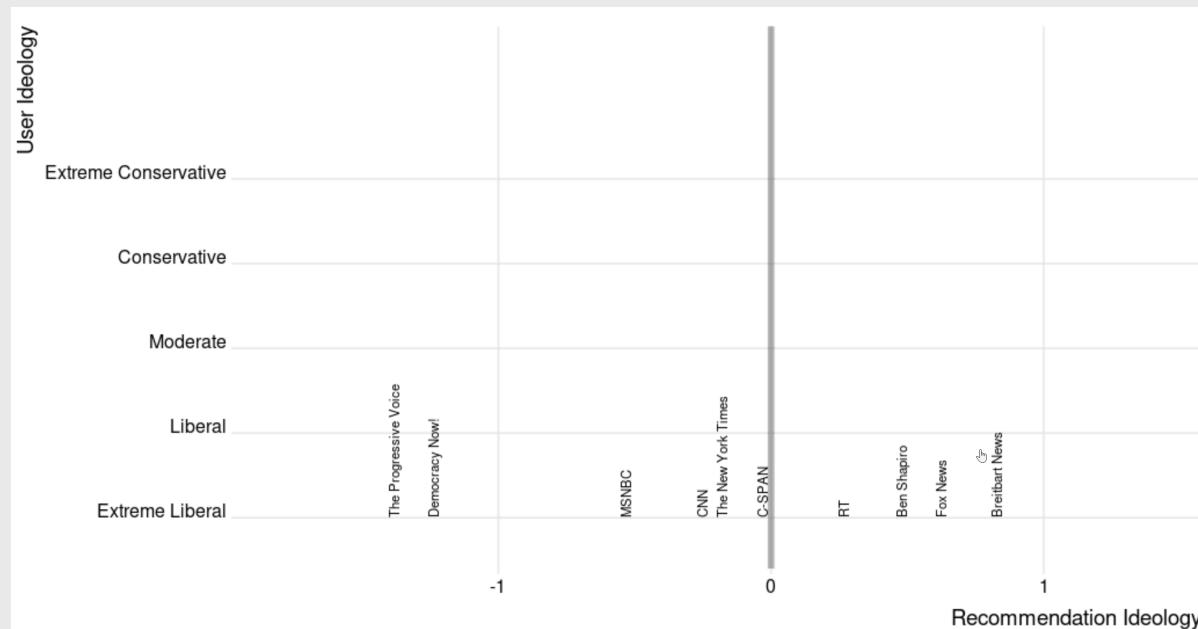
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 4. Results → Conclusion

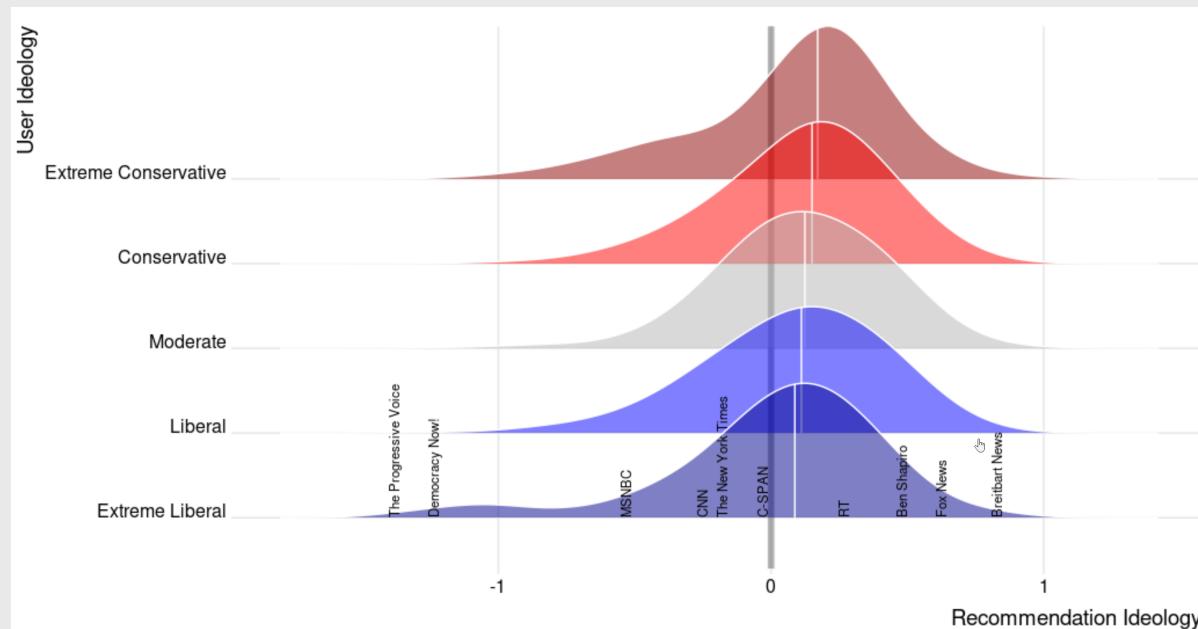
- Results fall out naturally from the analysis...
- ...and must be interpreted in terms of the theory and hypotheses...
- ...to draw conclusions



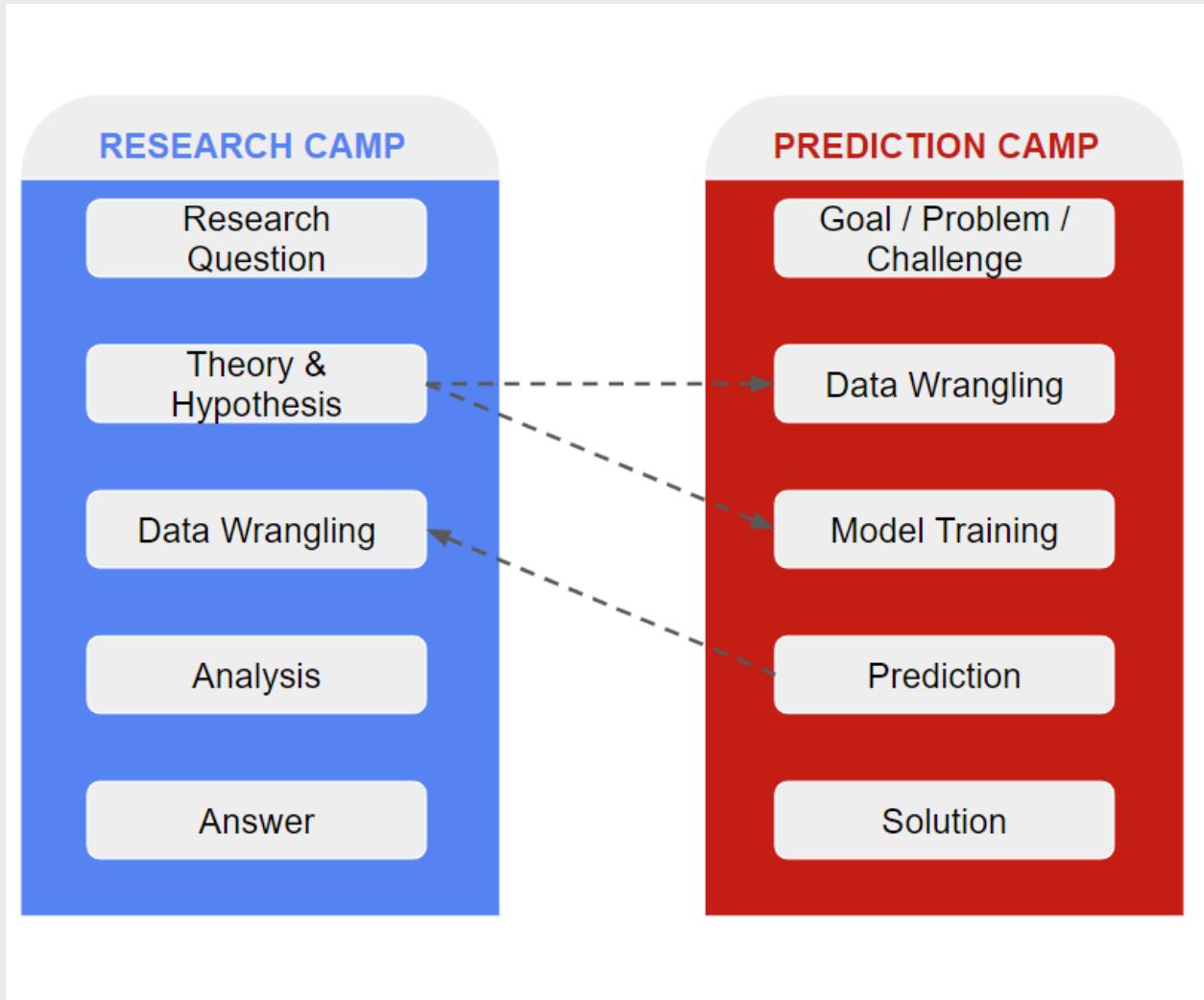
# Research Camp

## 4. Results → Conclusion

- Results fall out naturally from the analysis...
- ...and must be interpreted in terms of the theory and hypotheses...
- ...to draw conclusions



# The Two Camps



# Prediction Camp

- **Goal/Problem/Challenge:** Measure the ideology of a YouTube video

# Prediction Camp

- **Data Wrangling:** Get matrix of links shared on political subreddits

## The Ideology of a Video in 3 Steps: Step 1

Behavior:  
Sharing URLs

Posted by u/santanzchild Constitutional Conservative  
6 hours ago 2

AOC, a Sitting Member of Congress,  
Weaponized Her Followers in an  
Attempt to Silence a Free Press  
[redstate.com/jenav...](http://redstate.com/jenav...) 2

1.2k 323 Comments Share ...

Posted by u/oz4ut Conservative 3 hours ago

Joe Biden's Abortion Policies Are  
Grounds For Excommunication  
[thefederalist.com/2021/0...](http://thefederalist.com/2021/0...) 2

308 131 Comments Share ...

Posted by u/guanaco55 Conservative 3 hours ago

The QAnon Takeover Of The GOP Is  
A Fantasy Of Dems And The Media --  
The GOP Civil War Is Between  
Populists and the Establishment  
[thefederalist.com/2021/0...](http://thefederalist.com/2021/0...) 2

303 43 Comments Share ...

# Prediction Camp

- **Data Wrangling**: Get matrix of links shared on political subreddits

# The Ideology of a Video in 3 Steps: Step 1

## Behavior: **Sharing URLs**

+

Domain:  
**Subreddits**

Posted by u/santanzchild **Constitutional Conservative**  
6 hours ago   2

AOC, a Sitting Member of Congress, Weaponized Her Followers in an Attempt to Silence a Free Press

[redstate.com/jenvan...](http://redstate.com/jenvan...) ↗



r/Conservative

Posted by u/oz4ut · Conservative · 3 hours ago

## **Joe Biden's Abortion Policies Are Grounds For Excommunication**

[thefederalist.com/2021/0... ↗](https://thefederalist.com/2021/0...)



r/neutralnews

↑ 308 ↓ 131 Comments ↗ Share ...

Posted by u/guanaco55 [Conservative](#) 3 hours ago

**The QAnon Takeover Of The GOP Is A Fantasy Of Dems And The Media -- The GOP Civil War Is Between Populists and the Establishment**

[theredlist.com/2021/0...](https://theredlist.com/2021/0...) ↗

thefederalist.com/2021/01/



r/SandersForPresident

# Prediction Camp

- **Data Wrangling:** Get matrix of links shared on political subreddits

## The Ideology of a Video in 3 Steps: Step 1



Posted by u/santanzchild Constitutional Conservative  
6 hours ago 2

AOC, a Sitting Member of Congress,  
Weaponized Her Followers in an  
Attempt to Silence a Free Press  
[redstate.com/jenav...](http://redstate.com/jenav...) 2

↑ 1.2k ↓ 323 Comments Share ...

Posted by u/oz4ut Conservative 3 hours ago

Joe Biden's Abortion Policies Are  
Grounds For Excommunication  
[thefederalist.com/2021/0...](http://thefederalist.com/2021/0...) 2

↑ 308 ↓ 131 Comments Share ...

Posted by u/guanaco55 Conservative 3 hours ago

The QAnon Takeover Of The GOP Is  
A Fantasy Of Dems And The Media --  
The GOP Civil War Is Between  
Populists and the Establishment  
[thefederalist.com/2021/0...](http://thefederalist.com/2021/0...) 2

↑ 303 ↓ 43 Comments Share ...



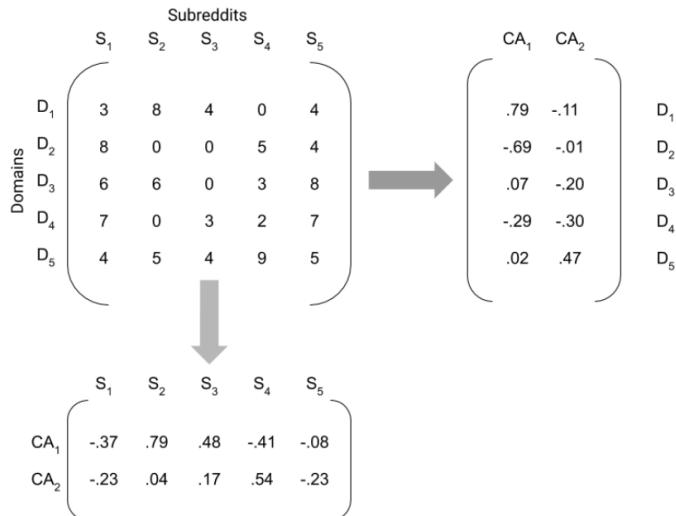
r/Conservative



r/neutralnews



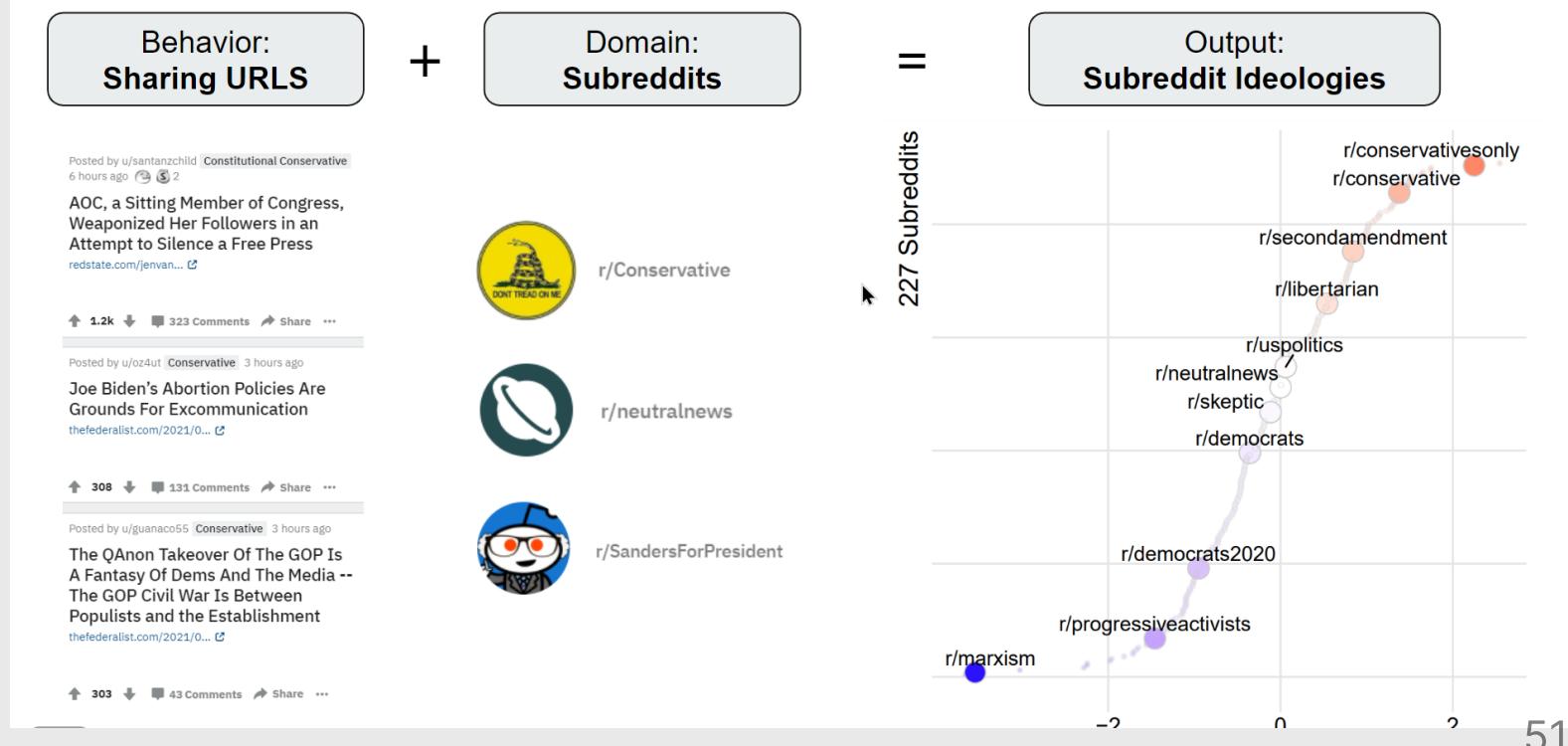
r/SandersForPresident



# Prediction Camp

- **Data Wrangling:** Correspondence Analysis to estimate ideology scores for subreddits

## The Ideology of a Video in 3 Steps: Step 1



# Prediction Camp

- **Data Wrangling:** Get matrix of YouTube videos shared on scored subreddits

## The Ideology of a Video in 3 Steps: Step 2

# Prediction Camp

- **Data Wrangling:** Get matrix of YouTube videos shared on scored subreddits

## The Ideology of a Video in 3 Steps: Step 2

Behavior:  
Sharing Videos

Interview with Thomas Biryani by a reporter from an abc local texas affiliate's live feed:  
<https://www.youtube.com/watch?v=X3WYY0fsF-I>  
r/PublicFreakout Posted by u/elseman 20 days ago

52 37 Comments Share ...

A wand with a twist! I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnhkNLUAvw> Credit:tpowen!\_   
r/Wandsmith Posted by u/tphilly3 16 days ago

44 15 Comments Share ...

Made a video about the G14 and my setup! Check it out if you're interested! It would be greatly appreciated! <https://www.youtube.com/watch?v=crcTp9vYE&feature=youtu.be> r/Zephyrus14 Posted by u/alexszurkus 1 month ago

11 30 Comments Share ...

why is jimin like this full video:  
<https://www.youtube.com/watch?v=iIhaZl1436M&t=173s> r/heungtan Posted by u/yangtiglighthere 14 days ago

74 6 Comments Share ...

# Prediction Camp

- **Data Wrangling:** Get matrix of YouTube videos shared on scored subreddits

## The Ideology of a Video in 3 Steps: Step 2

Behavior:  
Sharing Videos

+

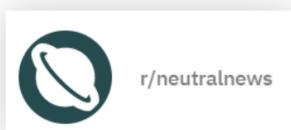
Domain:  
Ideological Reddit

Interview with Thomas Biryani by a reporter from an abc local texas affiliate's live feed:  
<https://www.youtube.com/watch?v=X3WYY0fsF-I>  
r/PublicFreakout Posted by u/elseman 20 days ago  
52 comments Share ...

A wand with a twist. I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnkhNUAvew> Credit:tpowen!\_   
r/Wandsmith Posted by u/tphilly3 16 days ago  
44 comments Share ...

Made a video about the G14 and my setup! Check it out if you're interested! It would be greatly appreciated! <https://www.youtube.com/watch?v=crcTp9vYEY&feature=youtu.be> r/Zephyrus14 Posted by valeruszurkus 1 month ago  
111 comments Share ...

why is jimin like this full video:  
<https://www.youtube.com/watch?v=vIhaZtI436M&t=173s>   
r/heungtan Posted by u/yangtiglighthere 14 days ago  
74 comments Share ...



# Prediction Camp

- **Data Wrangling:** Get matrix of 60k YouTube videos shared on scored subreddits

## The Ideology of a Video in 3 Steps: Step 2

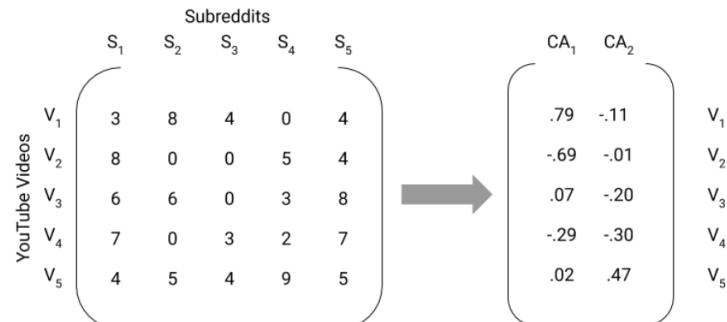
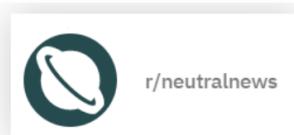


Interview with Thomas Biryani by a reporter from an abc local Texas affiliate's live feed:  
<https://www.youtube.com/watch?v=X3WYY0fsF-I>  
r/PublicFreakout Posted by u/eliseann 20 days ago  
52 comments Share ...

A wond with a twist. I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnhkNUlAew> Credit:tpowen...  
r/Wandsmith Posted by u/timothy3 56 days ago  
44 comments Share ...

Made a video about the G14 and my setup! Check it out if you're interested! It would be greatly appreciated! <https://www.youtube.com/watch?v=crcTp9vYEY&feature=youtu.be>  
r/Zephyrus14 Posted by valeriszurkus 1 month ago  
11 comments Share ...

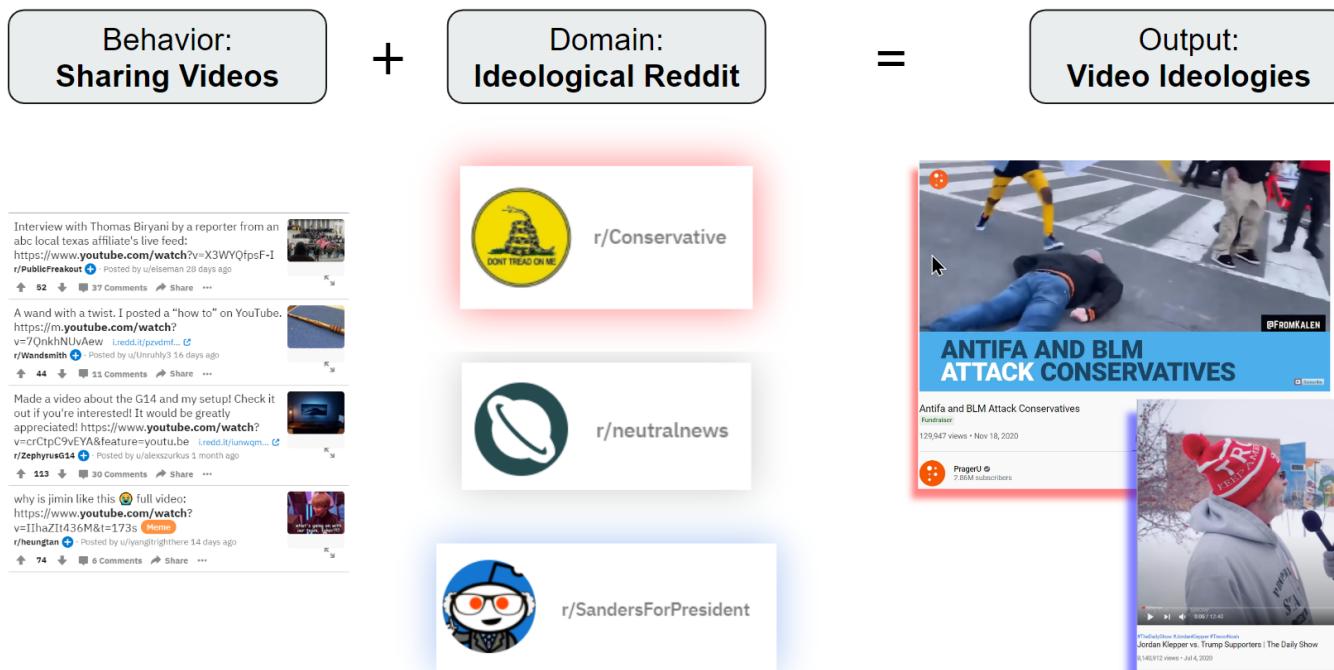
why is jimin like this full video:  
<https://www.youtube.com/watch?v=1haZtI436M&t=173s> Home  
r/heungtan Posted by u/yangtighere 14 days ago  
74 comments Share ...



# Prediction Camp

- **Data Wrangling:** Calculate video ideology as weighted mean of subreddits

## The Ideology of a Video in 3 Steps: Step 2



# Prediction Camp

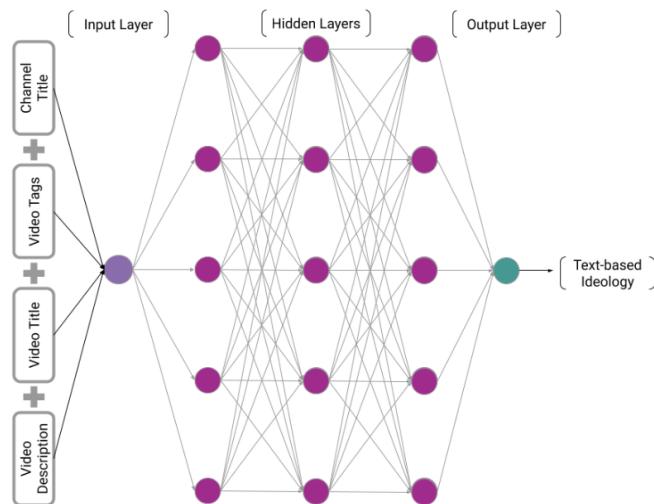
- **Model Training:** BERT transformer trained on 60k videos

## The Ideology of a Video in 3 Steps: Step 3

Training Data:  
67k Coded Videos

+

Classifier:  
BERT Transformer



# Prediction Camp

- **Prediction:** Measure the ideology of a YouTube video

## The Ideology of a Video in 3 Steps: Step 3

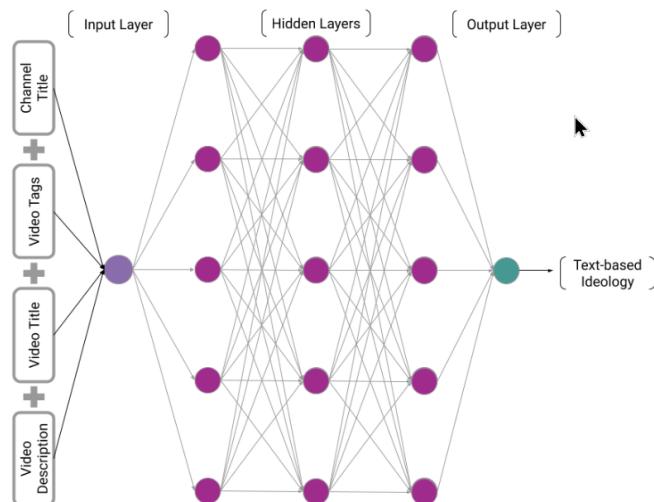
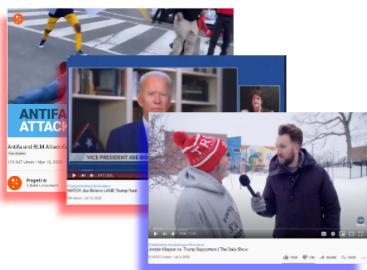
Training Data:  
67k Coded Videos

+

Classifier:  
BERT Transformer

=

Output:  
Any Video's Ideology



# Course Objectives

- This course is the menu, not the food
  - Look over many different fields, methods, and tools
  - You pick those you like, and take more advanced classes to dig into them
- But we are very **hands on**
  - You must download `R` and `RStudio` prior to next class (Problem Set 0)
  - You must work through first HW using an `.Rmd` file

# Learning goals

1. Generate a sophisticated research question based on clearly described assumptions and a narrowly defined hypothesis.
2. Describe the data used to investigate this research question, including univariate and multivariate visualizations and summary statistics.
3. Apply the appropriate methods to answer the research question and evaluate the hypothesis.
4. Acknowledge limitations of method and results, and describe a superior empirical setting that would overcome these limitations.

# ChatGPT in the classroom

- Are we at the precipice of a new era in human-computer relations?
  - ChatGPT can achieve these learning goals!
  - But it needs to be used wisely...it is still a tool
- It can make coding (the hardest part of this class) easier
- But it can also prevent you from learning

# AI in the labor market

- McKinsey told AT&T in 1980 that, by 2000, cell phones would be a niche market of 900,000 subscribers
- Is AI-assisted work is the future?
  - Profound gains in productivity already
- Will this be like automation and globalization for US manufacturing?
  - What skills will be valuable in 5 years? 10 years?

# AI in the labor market

- My answer: prepare you for both possibilities
  - If AI is a "fad", make sure you can do this work unassisted
  - If AI is the new normal, make sure you can work with it productively
- The one thing you **shouldn't** do
  - Take shortcuts / cheat
- You will still have an interview in which you are asked something like the following: "How is overfitting different from underfitting, and why should we care?"
  - **You** need to know this answer

# Grades

Item	Percent	Points
pset 1	5%	10
pset 2	5%	10
pset 3	5%	10
pset 4	5%	10
pset 5	5%	10
pset 6	5%	10
pset 7	5%	10
pset 8	5%	10
Midterm	20%	40
Final Exam	20%	40
Quizzes	20%	40
Totals	100%	200

# Grades: PSets

- 9 in total, only 8 are graded
  - Pset 0 doesn't count
- Posted to **Brightspace** on Mondays at noon
- Due **Friday by midnight**
  - Each day late is -1 point
  - After 3 days, scored zero
- Restrictions:
  - Open book / open note / open Campuswire
  - Can collaborate but submissions must be your own
- **Must submit a record of ChatGPT work with the problem set**

# Grades: Exams

- 2 in total: midterm on October 25th, final on December 11th
- 20% of final grade
- Restrictions:
  - **In-class with pen and paper**

# Grades: Quizzes

- Taken at end of each lecture
- Password protected
  - Only students in class can take them
  - 50% of quiz grade is just taking it (sign affidavit)
  - 50% of quiz grade is four questions related to lecture

# Not Graded: HW

- You should work through the homeworks prior to each lecture
- Open the `.Rmd` file and Knit it
- Read the output and try and answer the prompts
- **Not graded**, but enormously helpful in preparing you to keep up with lectures

# The Syllabus

Date	Lecture	DOW	Learning Goal	Assignments	Quizzes
23-Aug-23	Intro to Data Science Part 1	Wed	The scientific method, the camps of analysis	Pset 0 assigned	
28-Aug-23	Intro to Data Science Part 2	Mon	ChatGPT and the New Frontier of Data Science		Quiz 1
<b>30-Aug-23 BREAK</b>		<b>Wed</b>			
4-Sep-23	Intro to R Part 1	Mon	Objects, functions, %>%, and <-		Quiz 2
6-Sep-23	Intro to R Part 2	Wed	Visualization in R		Quiz 3
11-Sep-23	Intro to R Part 3	Mon	More visualization	Pset 1 assigned	Quiz 4
<b>13-Sep-23 Review of R</b>		<b>Wed</b>			
18-Sep-23	Data Wrangling	Mon	Replicability, R, and tabular data	Pset 2 assigned	Quiz 5
20-Sep-23	Univariate Analysis	Wed	Summaries of a single variable		Quiz 6
25-Sep-23	Multivariate Analysis Part 1	Mon	Summaries of multiple variables	Pset 3 assigned	Quiz 7
27-Sep-23	Multivariate Analysis Part 2	Wed	Visualizations of multiple variables		Quiz 8
2-Oct-23	Multivariate Analysis Part 3	Mon	Uncertainty and bootstrapping	Pset 4 assigned	Quiz 9
<b>4-Oct-23 Multivariate Review</b>		<b>Wed</b>			
9-Oct-23	Regression Part 1	Mon	The concept of a linear regression	Pset 5 assigned	Quiz 10
11-Oct-23	Regression Part 2	Wed	Interpreting a linear regression output and evaluating model		Quiz 11
16-Oct-23	Regression Part 3	Mon	Multiple regression and categorical predictors		Quiz 12
18-Oct-23	Regression Review	Wed			
23-Oct-23	Midterm Review	Mon			
<b>25-Oct-23 Midterm Exam</b>		<b>Wed</b>			
30-Oct-23	Classification Part 1	Mon	The concept of a logistic regression	Pset 6 assigned	Quiz 13
1-Nov-23	Classification Part 2	Wed	Interpreting a logistic regression output and evaluating model		Quiz 14
6-Nov-23	Classification Part 3	Mon	Using models for prediction	Pset 7 assigned	Quiz 15
<b>8-Nov-23 Classification Review</b>		<b>Wed</b>			
13-Nov-23	Clustering & NLP Part 1	Mon	k-means clustering	Pset 8 assigned	Quiz 16
15-Nov-23	Clustering & NLP Part 2	Wed	k-means clustering on text		Quiz 17
<b>20-Nov-23 BREAK</b>		<b>Mon</b>			
<b>22-Nov-23 BREAK</b>		<b>Wed</b>			
27-Nov-23	Clustering & NLP Part 3	Mon	Sentiment analysis	Pset 9 assigned	Quiz 19
<b>29-Nov-23 Clustering &amp; NLP Review</b>		<b>Wed</b>			
4-Dec-23	Ethics	Mon	The risks of rapid technological change		Quiz 20
<b>6-Dec-23 Final Review</b>		<b>Wed</b>			
<b>11-Dec-23 Final Exam</b>		<b>Mon</b>			

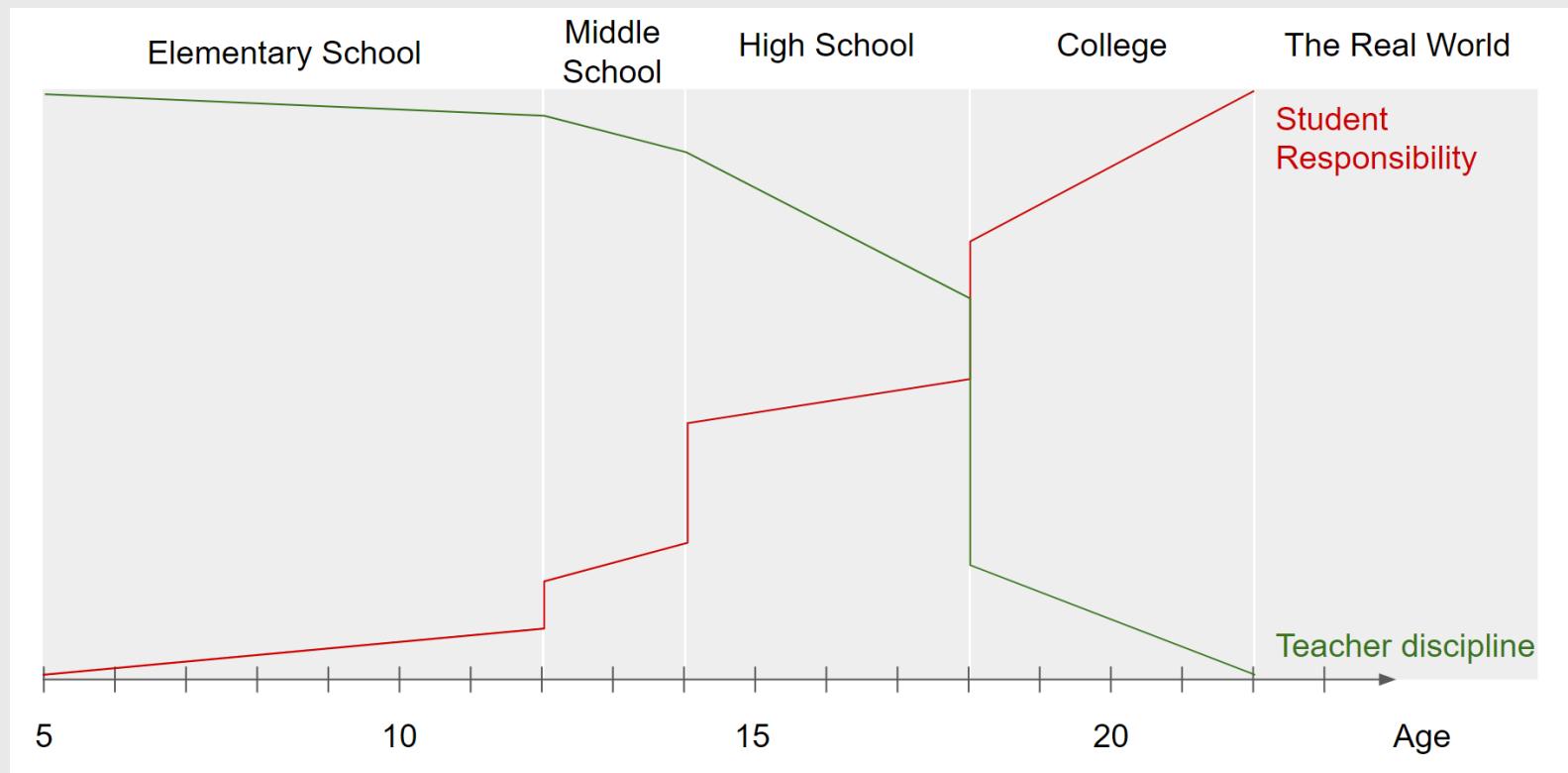
# Honor Code

- Students are assumed to have read and agreed with the [Vanderbilt University Academic Honesty policy](#)
- Violations of this policy may result in:
  - An F for the semester (at minimum)
  - Suspension for a semester
  - Expulsion
- However, except where **explicitly noted**, this course is collaborative
  - Open book, open note, open internet
  - Can rely on Campuswire for help
  - Can work together on problem sets (but must submit own work)
- **Can't collaborate on exams**

# Resources

- Campuswire (place for **questions**)
  - Post questions on the class feed
- Brightspace (place for **submissions**)
  - Submit problem sets, quizzes, and exams
- GitHub (place for **materials**)
  - Find all in-class materials
- TA recitations / labs (place for **hands-on help**)
- Office hours (place for **hands-on help**)

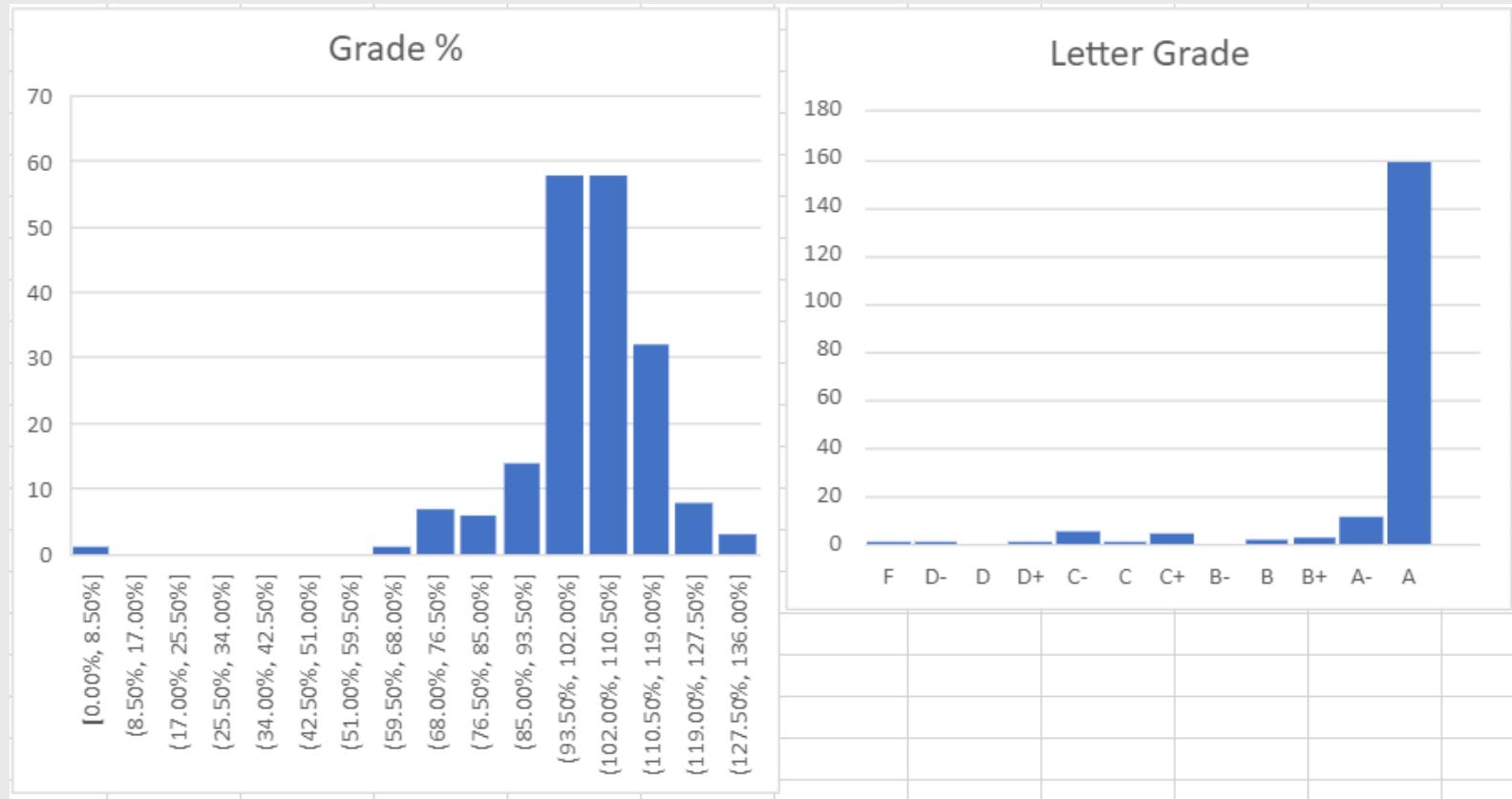
# Teaching Philosophy



# Teaching Philosophy

- This course is **inherently** hard
  - Learning **R** is challenging
- But the goal is to **encourage** you to pursue data science
- As such, the **nature** of the material is at odds with the **goal** of the class
- My solution: grade leniently
  - + lots of extra credit

# Previous Semester



# Conclusion

- Let's have a great semester!
- Homework:
  1. Work through Intro\_Data\_Science\_hw.Rmd
  2. Complete Problem Set 0 (on Brightspace)
  3. Create an OpenAI account (<https://platform.openai.com/signup>)