

# Intro to Data Science

What **R** we doing?

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/01/09

Slides Updated: 2023-01-03



# Agenda

## 1. Meet the instructor

- Prof. Bisbee: *james.h.bisbee@vanderbilt.edu*

## 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. Course Expectations & Syllabus review

# Meet the instructor

- 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
  - Monkey Cage articles [1](#), [2](#)
  - [Podcast](#) / Radio interviews

# Meet the instructor

- Current research
  - YouTube + polarization
  - Twitter + misinformation
  - Telegram + white supremacists
- Is my current research agenda data science?

# What is "data science"?

- What is **data**?
- What is **science**?

# What is **data**?

- "It is a capital mistake to theorize before one has data." Sherlock Holmes
  - Data **informs**
- "Torture the data, and it will confess to anything." Ronald Coase, Nobel Prize Laureate in Economics
  - Data **lies**
- "Here's an open secret of the big data world: all data is dirty. All of it." Meredith Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*
  - Data is **invalid**

# What is science?

- Simplification, codification, abstraction
  - Science identifies patterns in data...
  - ...to make predictions about the future
- As such, it is inherently:
  - Causal
  - Empirical
  - Theoretical

# What is data science?

- Data: informs / lies / invalid
- Science: simplification / codification / abstraction
- Data + science = ?



# Why are you here?



Suggested fights

20 last fights



## DATA SCIENCE vs STEM

200



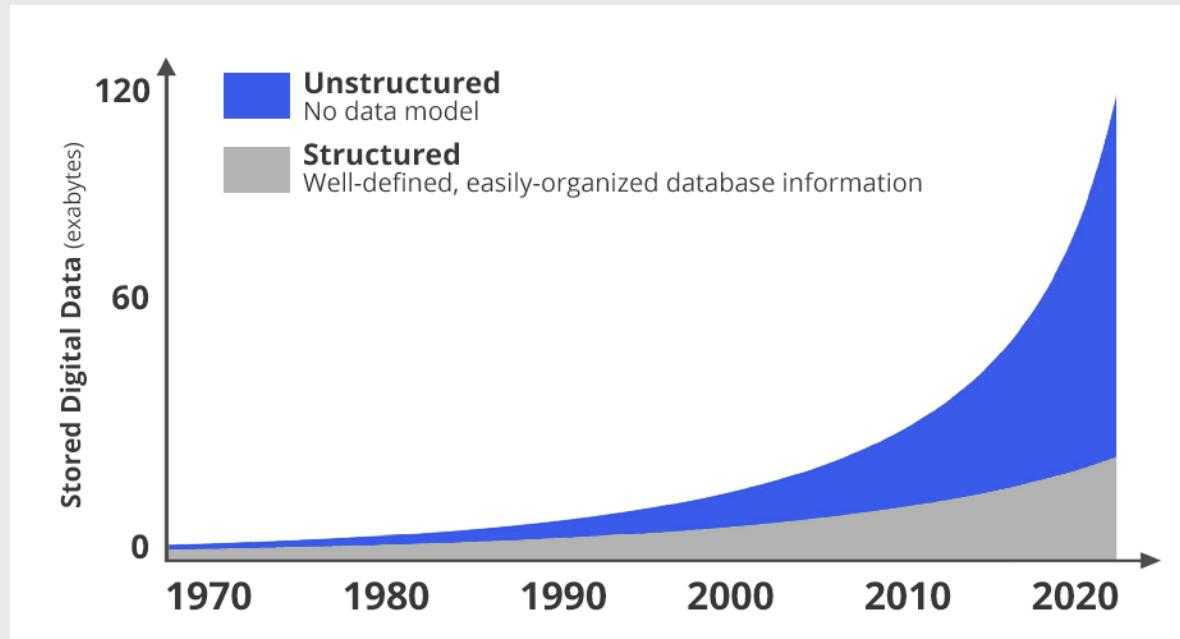
DATA SCIENCE

101

STEM

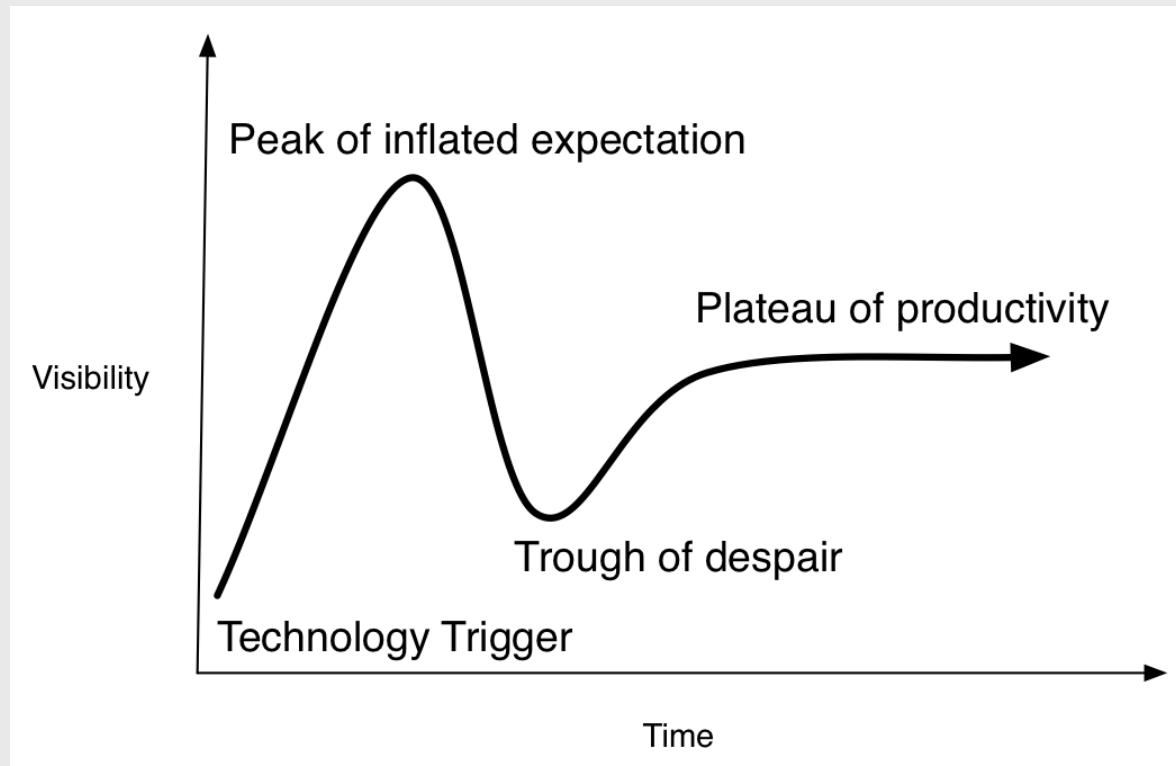
# Is this all just a fad?

- No



# Is this all just a fad?

- But there are faddish qualities



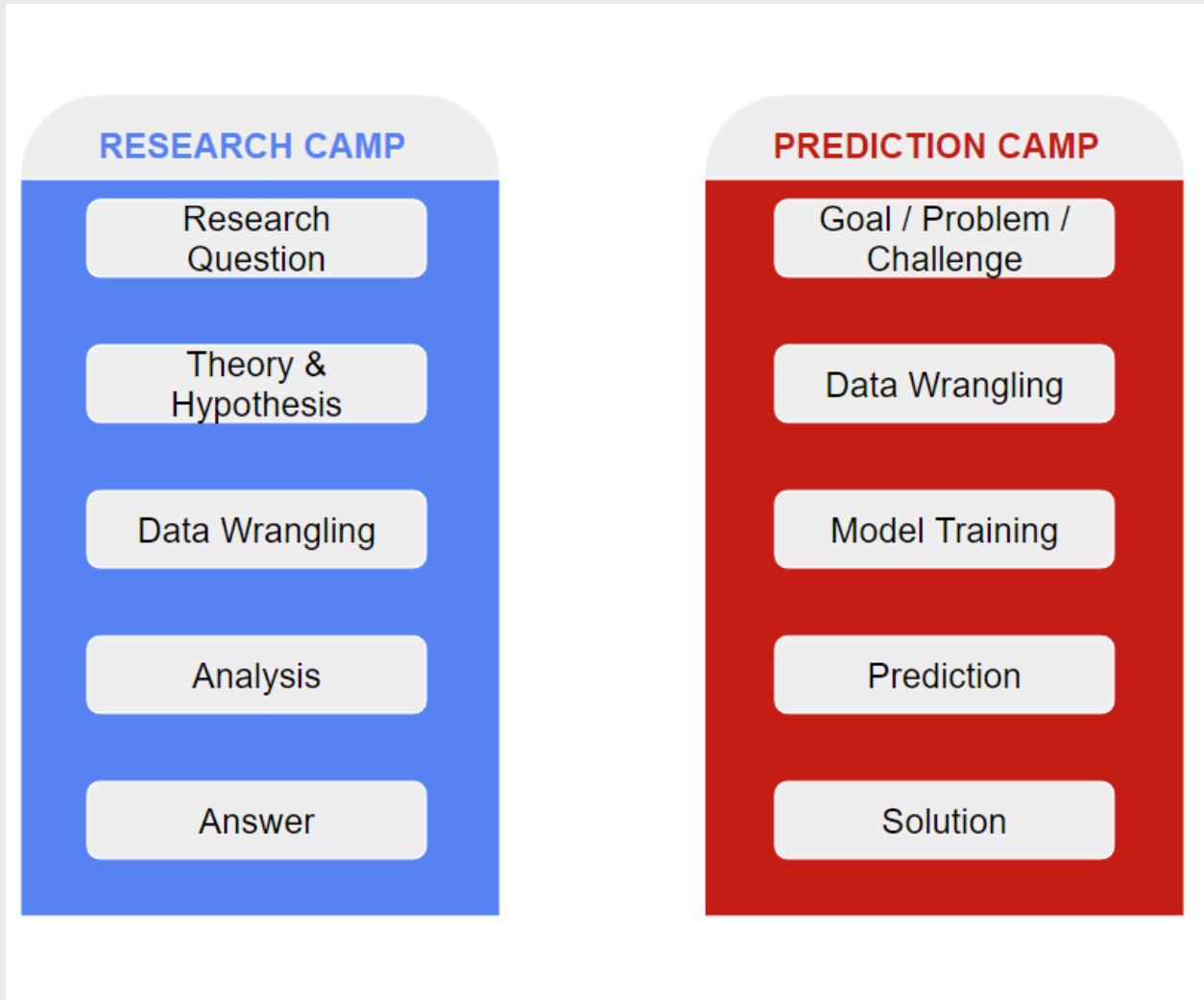
# So what IS data science?

- 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

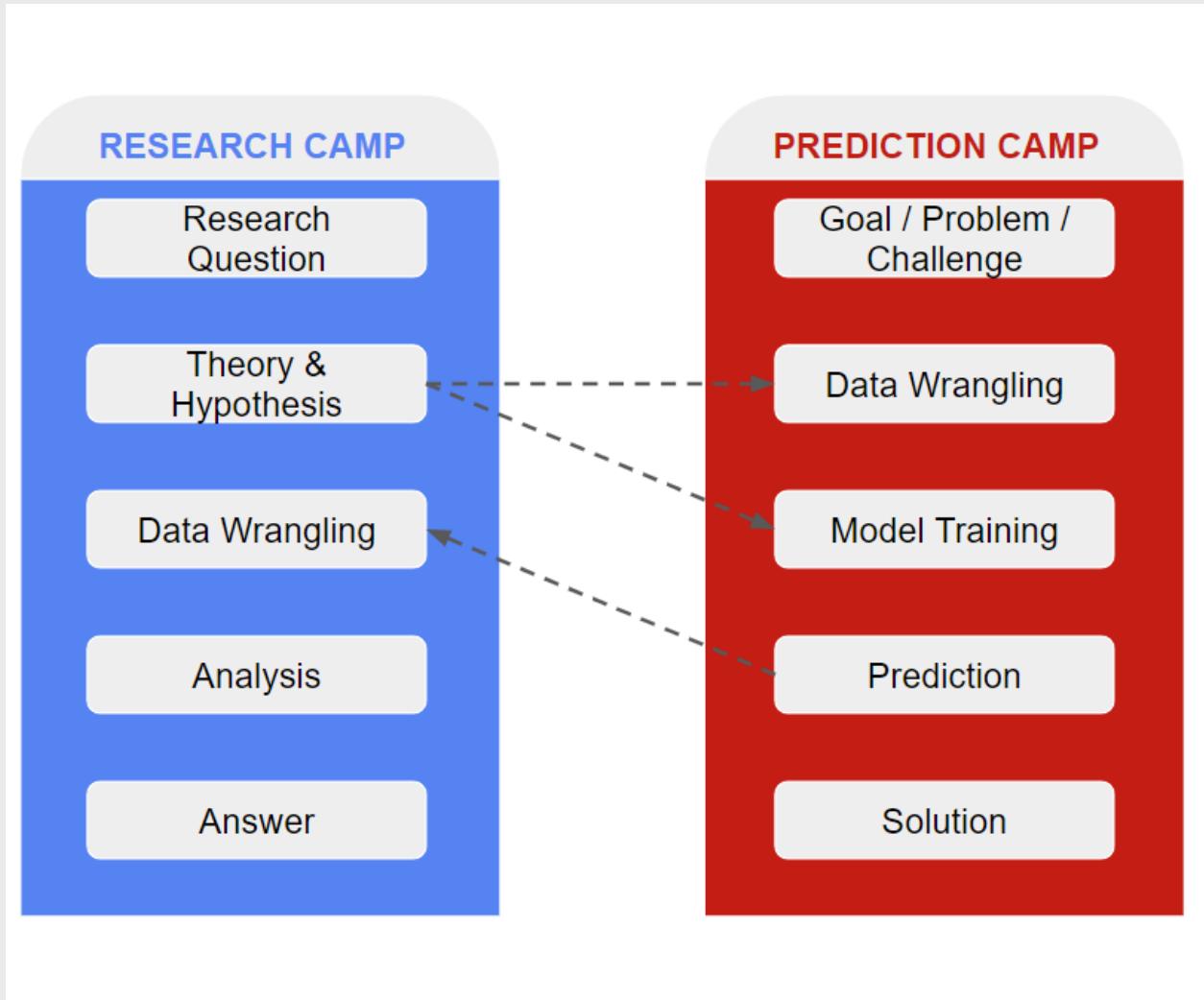
# The Two Camps



# The Two Camps

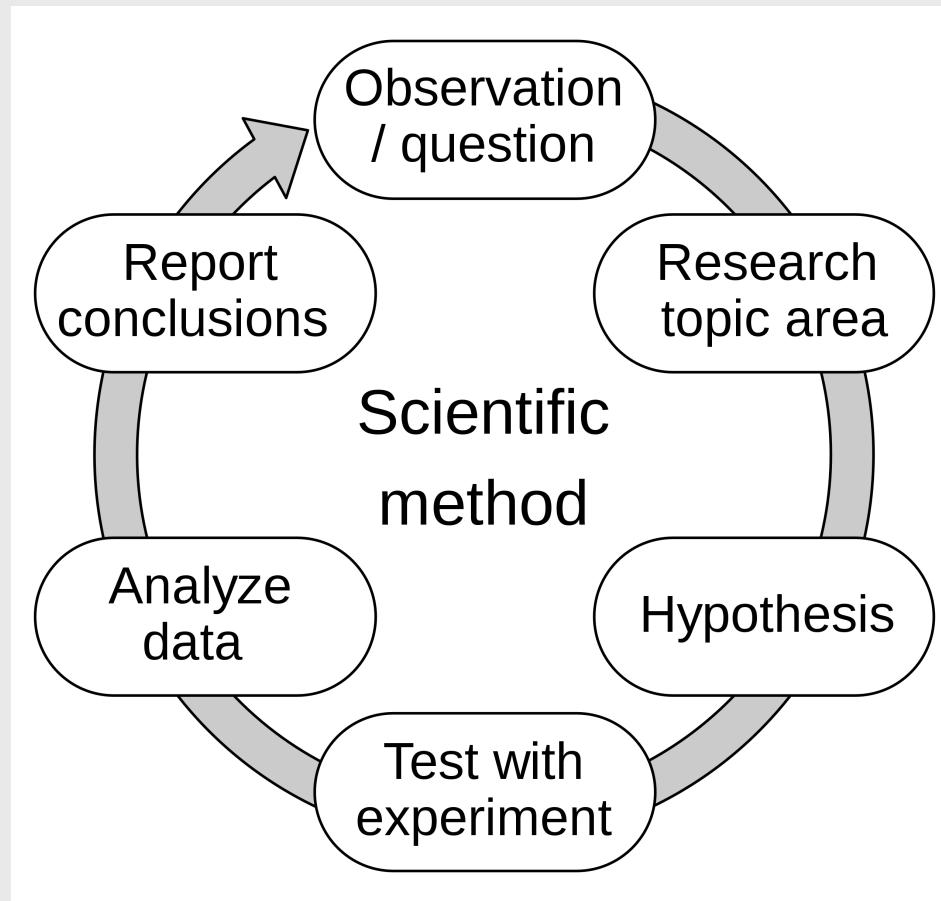


# The Two Camps



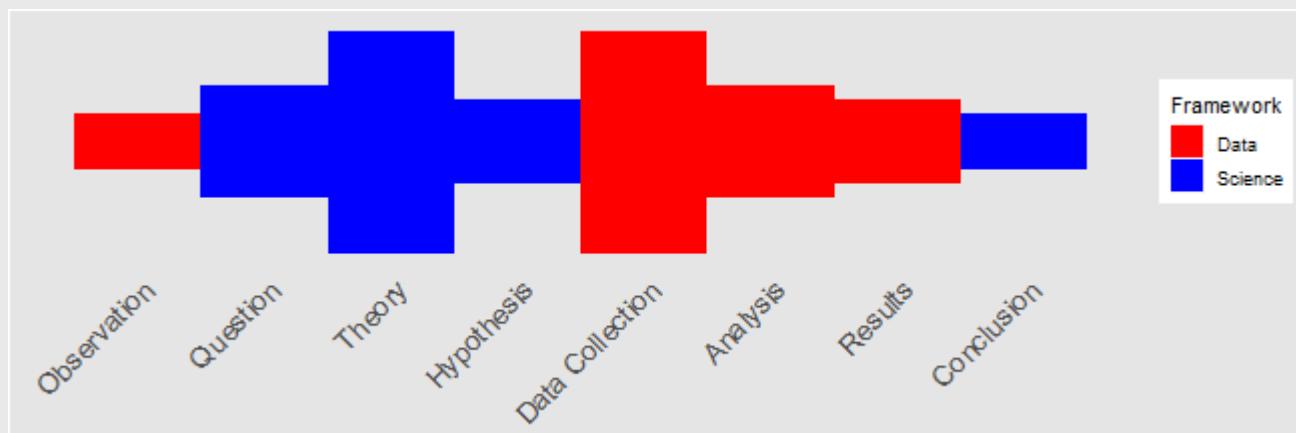
# Research Camp

- The scientific method



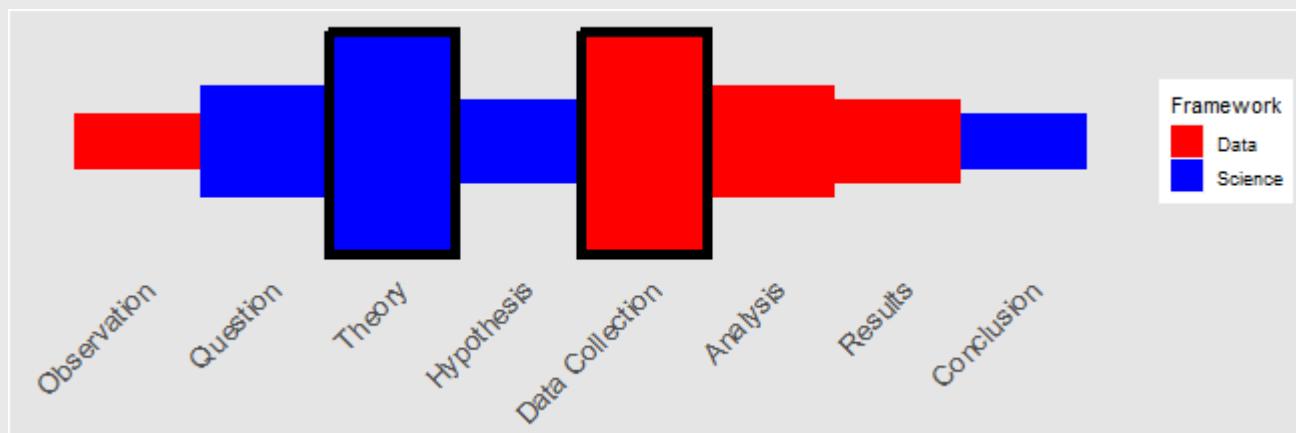
# Research Camp

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



# Research Camp

- The scientific method
  1. Observation → Question
  2. Theory → Hypothesis
  3. Data Collection → 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

August 24, 2022

### Abstract

To what extent does the YouTube recommendation algorithm push users into echo chambers, ideologically biased content, or rabbit holes? Despite growing popular concern, recent work suggests that the recommendation algorithm is not pushing users into these echo chambers. However, existing research relies heavily on the use of anonymous data collection that does not account for the personalized nature of the recommendation algorithm. We asked a sample of real users to install a browser extension that downloaded the list of videos they were recommended. We instructed these users to start on an assigned video and then click through

# 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 displays 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'. A small note below the bar says 'U.S. elections' and 'The AP has called the Presidential race for Joe Biden. See more on Google.' 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 MSNBC, 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

AUTOPLAY

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

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

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

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

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

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

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

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

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

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

22 / 74

# Research Camp

## 1. Observation → Question

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

The screenshot shows a CBS News video player interface. At the top, there are two video frames: the left one shows a man in a suit and tie, and the right one shows a woman in a pink top. Below these is a red banner with white text that reads "PRES. USES SOCIAL MEDIA TO DENOUNCE 'RIGGED' ELECTION". To the right of the banner is a "LIVE" CBSN logo. The main video frame has a play bar at the bottom with a progress indicator showing 0:06 / 11:40. Below the play bar, there's a "U.S. elections" section with a note from AP about Biden winning. A "SHOW ME" button is also present. On the right side of the screen, there's a sidebar titled "Up next" with several news items listed:

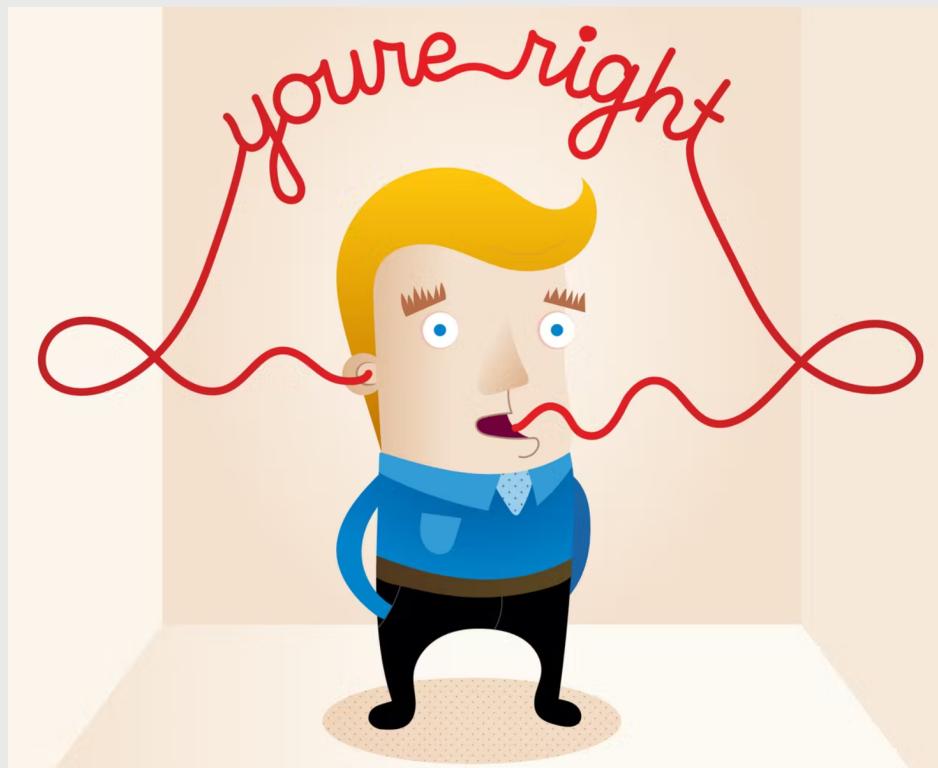
- HOW IT STARTED: Senate Hearing On FBI Investigation (47K views • 3 hours ago)
- Attorney General William Barr's job in jeopardy (57K views • 5 hours ago)
- The Full Story of Trump and COVID-19 (1.8M views • 1 month ago)
- Live: New York Gov. Andrew Cuomo Holds Briefing On Covid-19 (9.2K watching)
-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

At the bottom of the screen, there's a caption about Trump's Facebook video, a like/dislike count (942 upvotes, 182 downvotes), a share button, a save button, a subscribe button, and a page number "23 / 74".

# Research Camp

## 1. **Observation** → **Question**

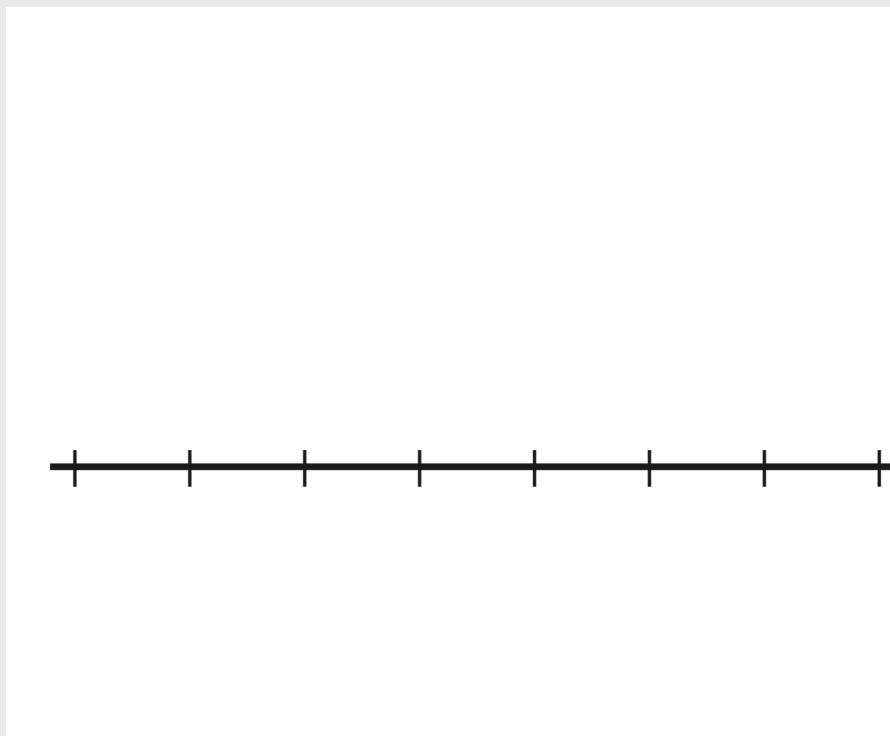
- The question pertains to science
- I.e., does YouTube's algorithm put users into "echo chambers"?



# Research Camp

## 2. Theory → Hypothesis

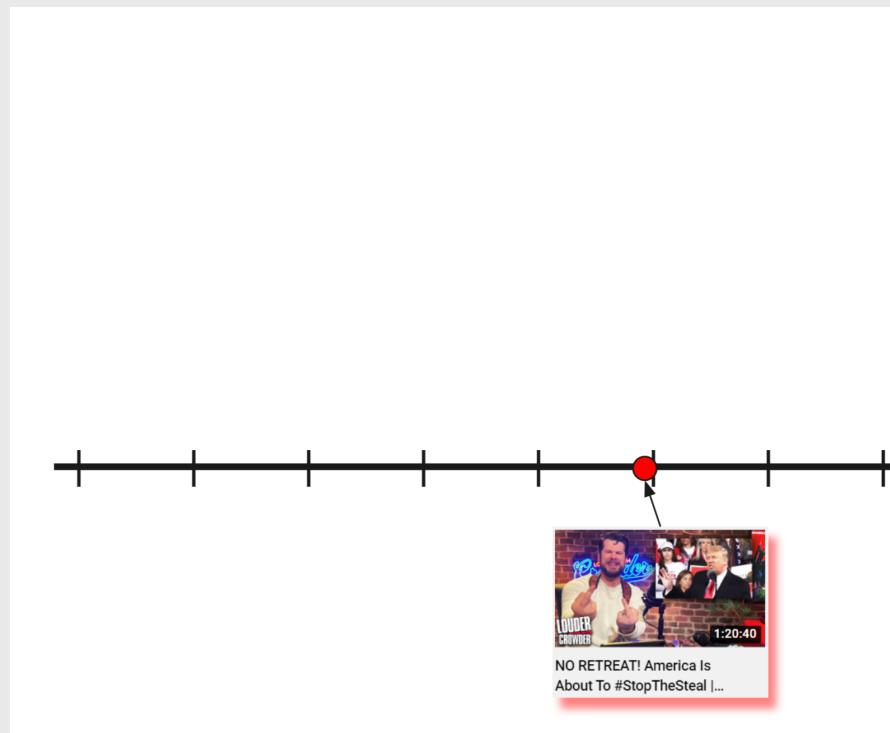
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

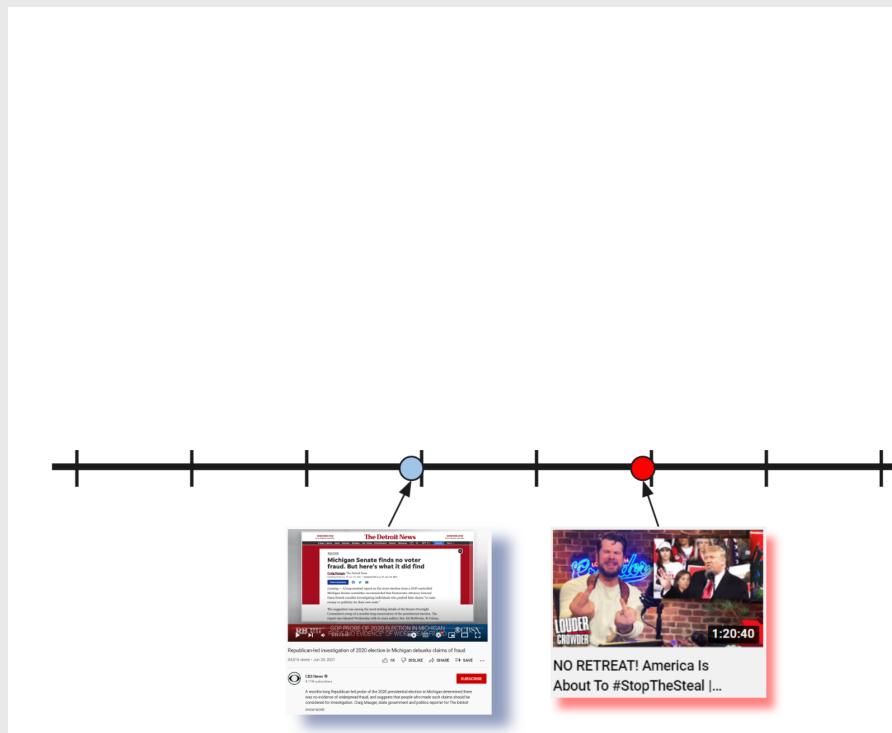
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

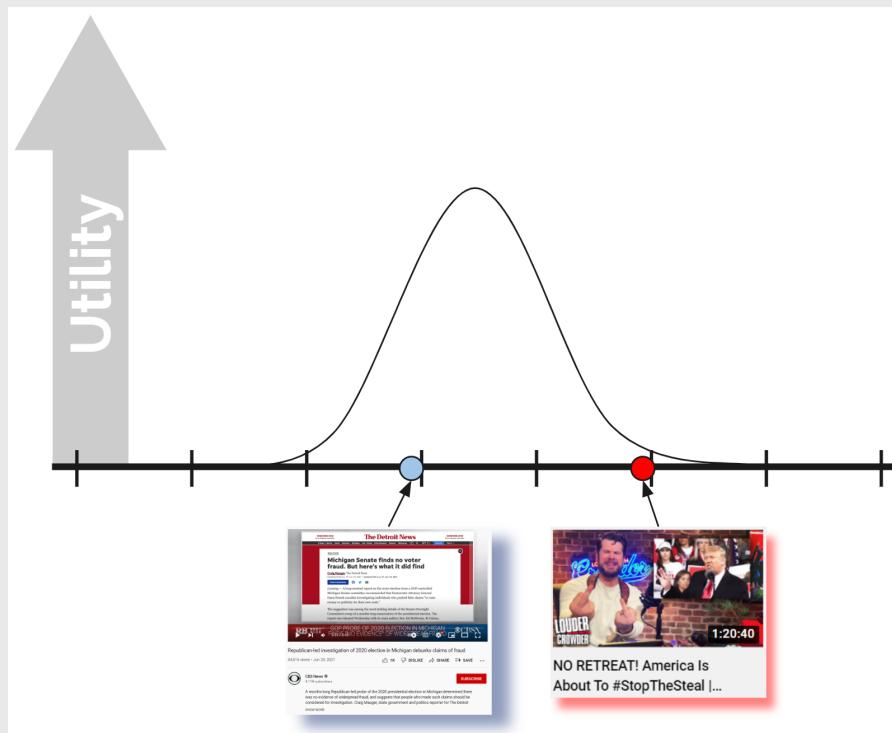
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# Research Camp

## 2. Theory → Hypothesis

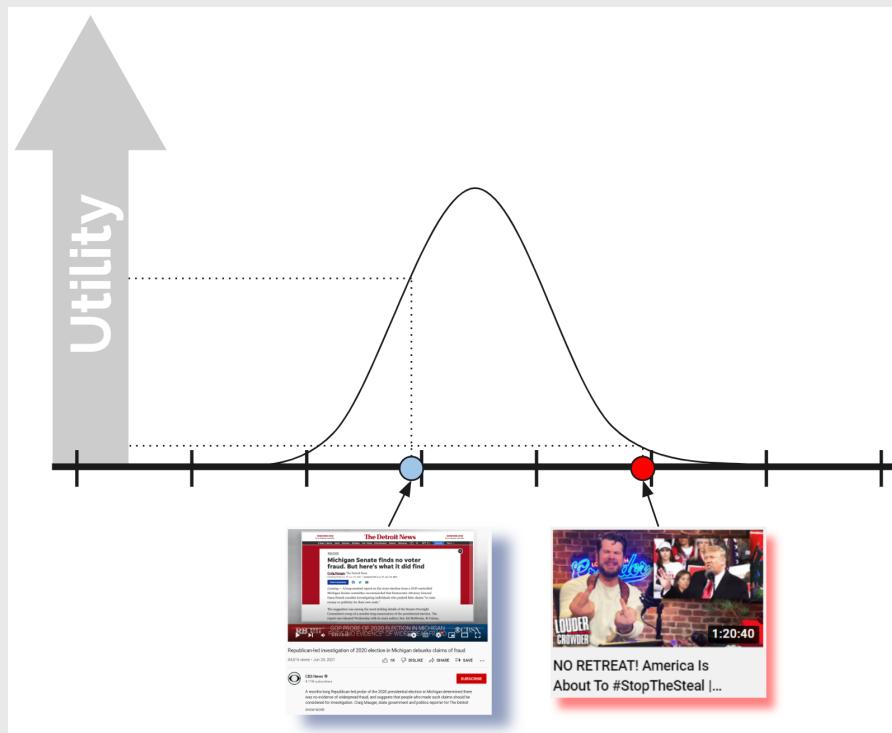
- Theorizing requires abstraction & simplification
- I.e., people (in general) avoid conflict



# 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**

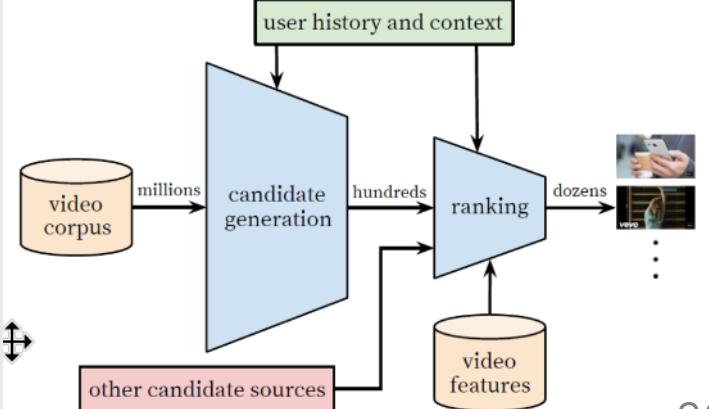
Paul Covington, Jay Adams, Emre Sargin  
Google  
Mountain View, CA  
[{pcovington,jka,msargin}@google.com](mailto:{pcovington,jka,msargin}@google.com)

**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" stage. This stage outputs "hundreds" of candidates. These candidates then pass through a "ranking" stage, which outputs "dozens" of results. The ranking process takes into account "user history and context", "video features", and "other candidate sources". The final output is a list of video thumbnails, represented by small images of hands holding phones.

# 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 → Analysis

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

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



YouTube Recommendation Downloader

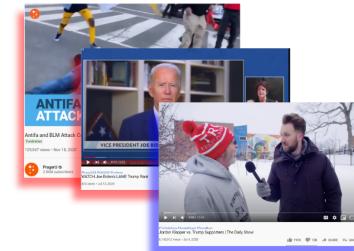
Offered by: csmappplugin

# Research Camp

## 3. Data Collection → Analysis

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

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



# Research Camp

## 3. Data Collection → Analysis

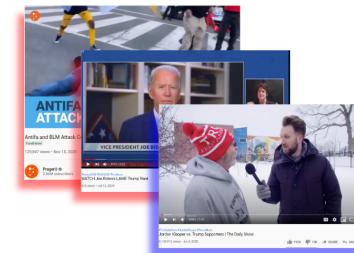
- Data collection separates "Data Science"...
- ...from "Science, with data"

- Recruit YouTube users to install **extension**



YouTube Recommendation Downloader

Offered by: csmapp plugin



- Start on randomly assigned **seed video**



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

# Research Camp

## 3. Data Collection → Analysis

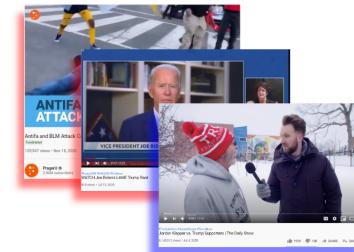
- Data collection separates "Data Science"...
- ...from "Science, with data"

- Recruit YouTube users to install **extension**



YouTube Recommendation Downloader

Offered by: csmapp plugin



- Start on randomly assigned **seed video**

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

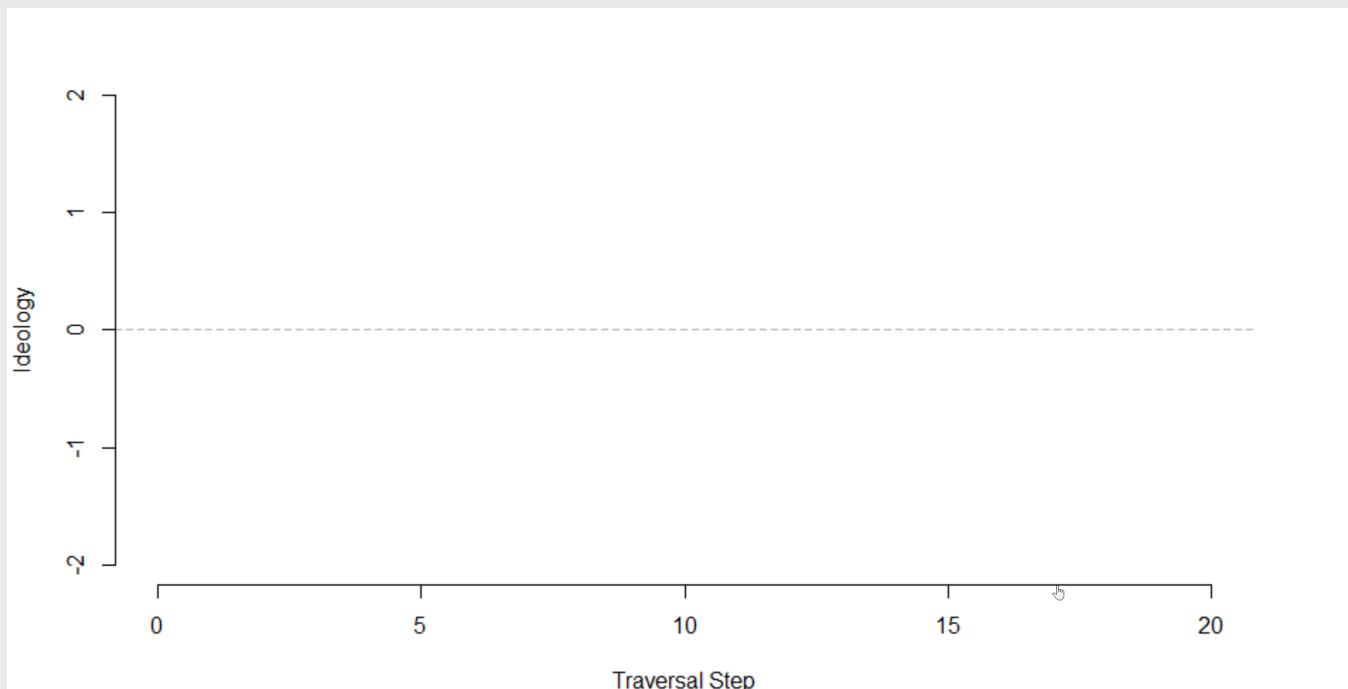


- Short **survey** on demographics, politics, and **BELIEFS ABOUT THE 2020 ELECTION**

# Research Camp

## 3. Data Collection → Analysis

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



# Research Camp

## 3. Data Collection → Analysis

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



# Research Camp

## 3. Data Collection → Analysis

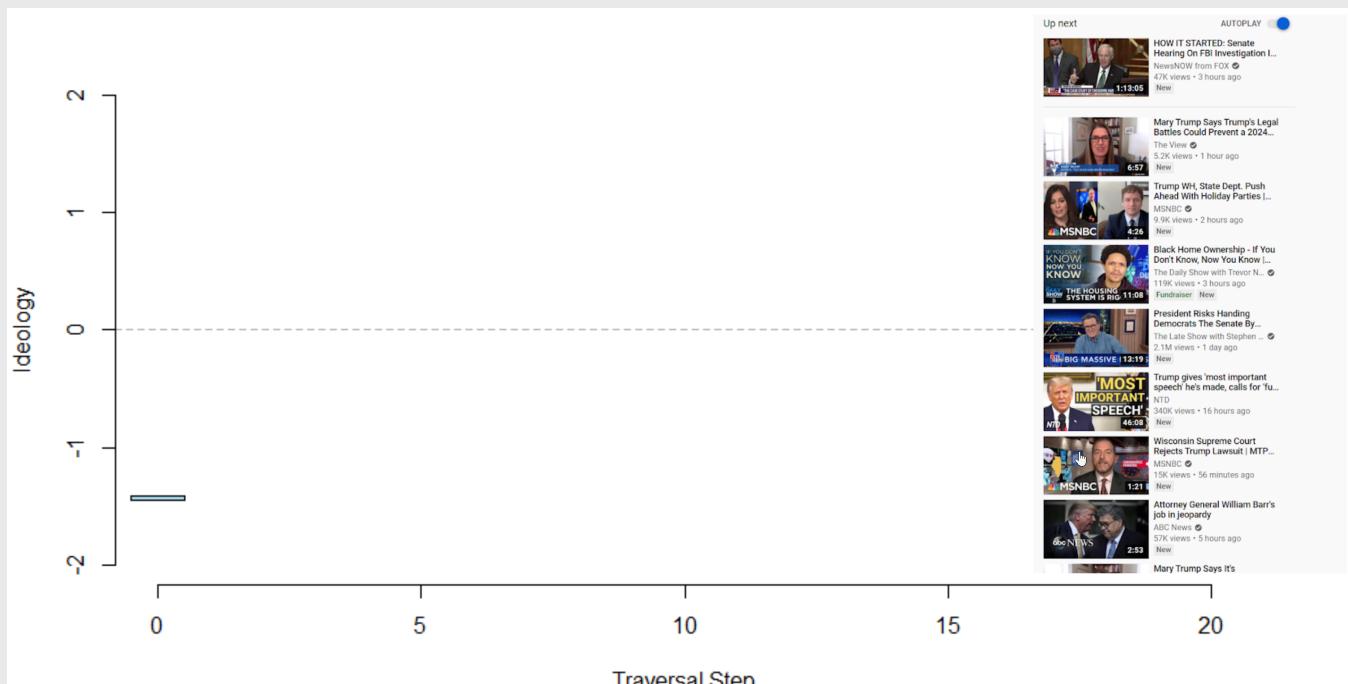
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

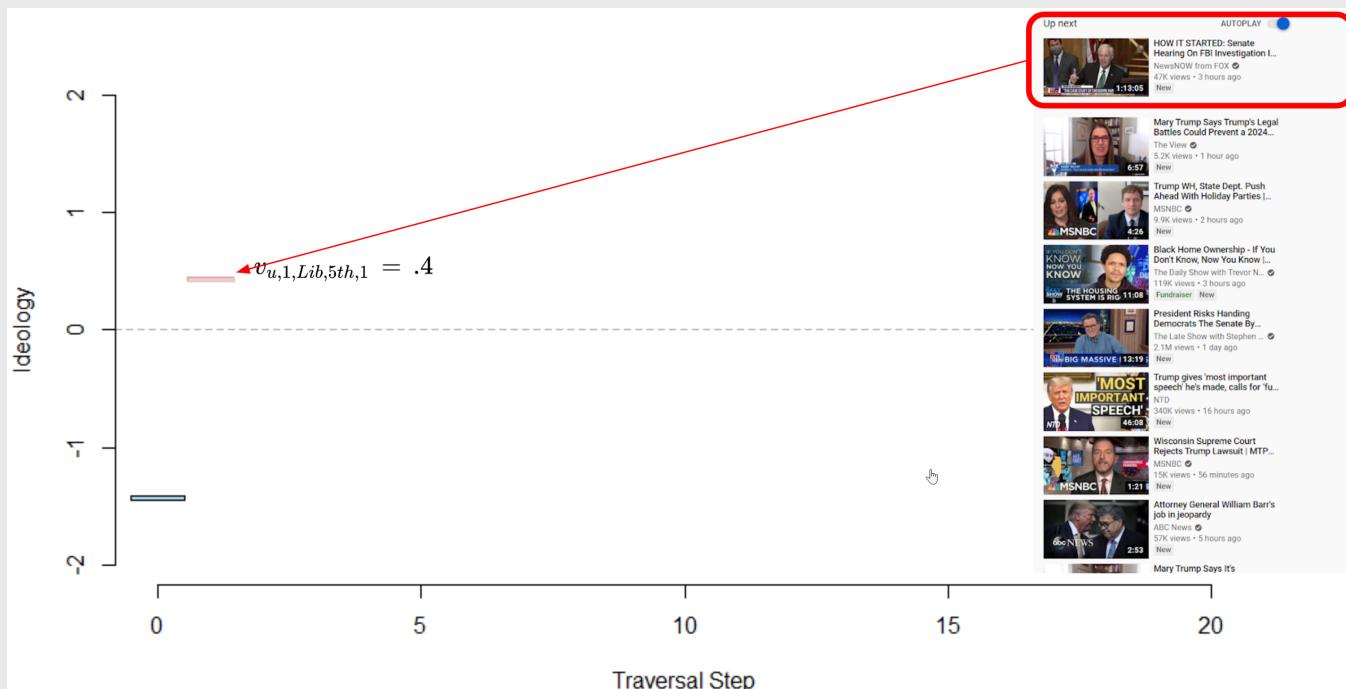
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

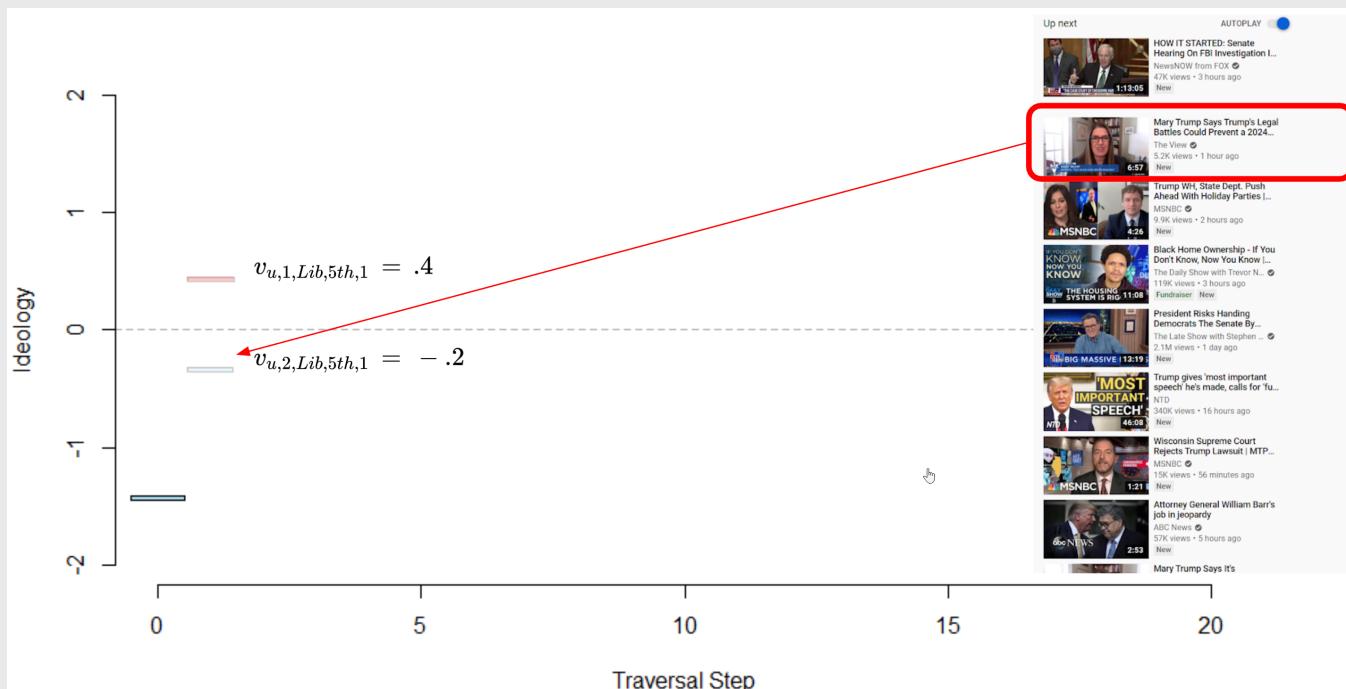
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

### 3. Data Collection → Analysis

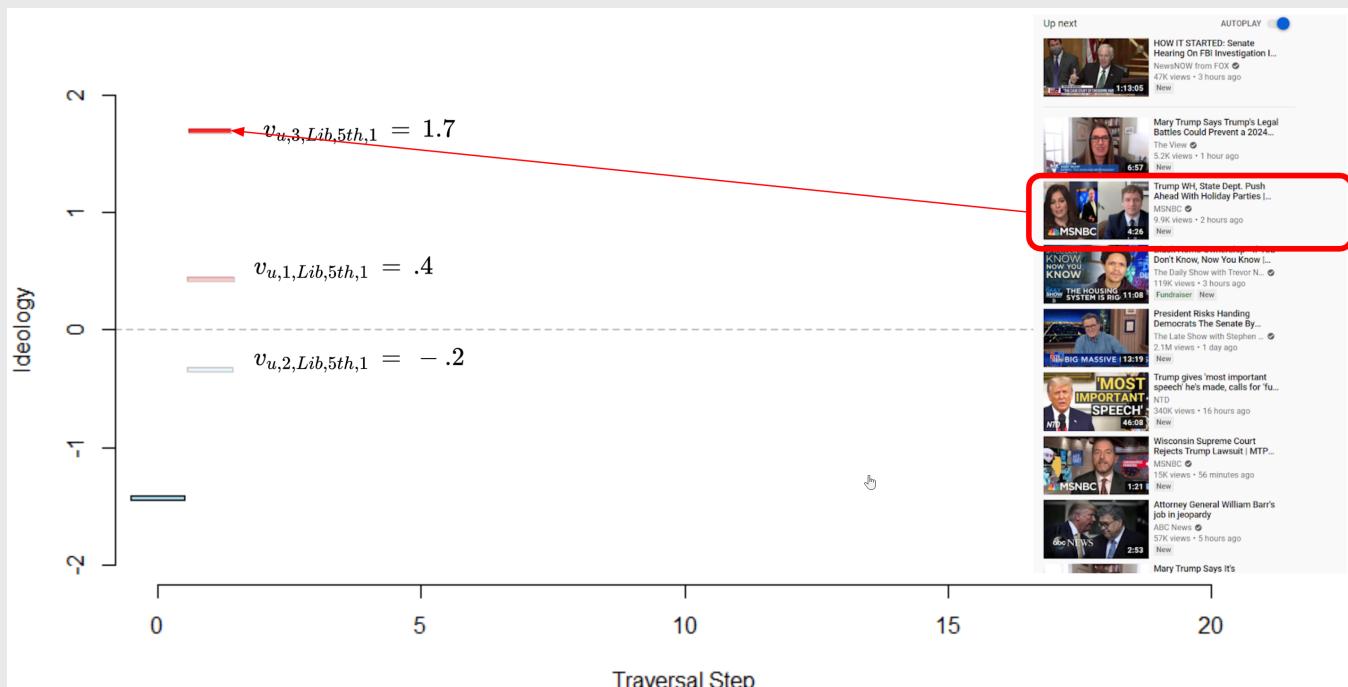
- Analysis is informed by the **data** you have collected...
  - ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

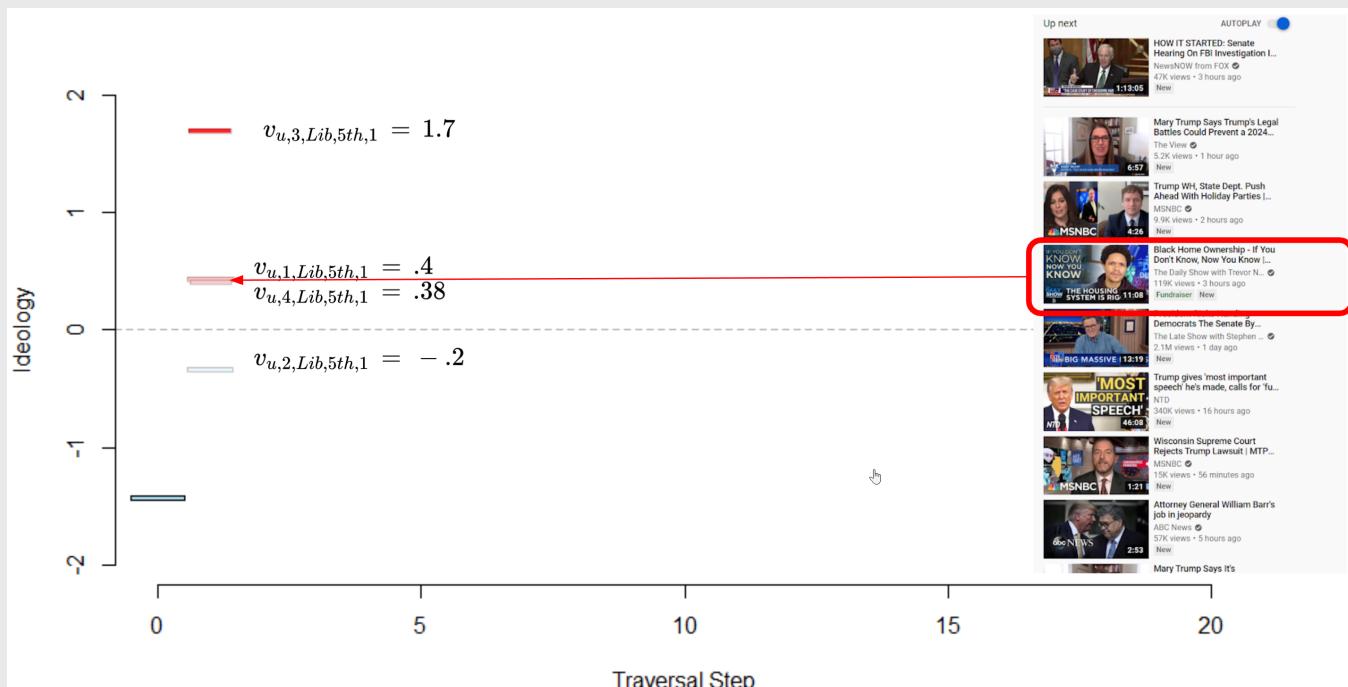
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

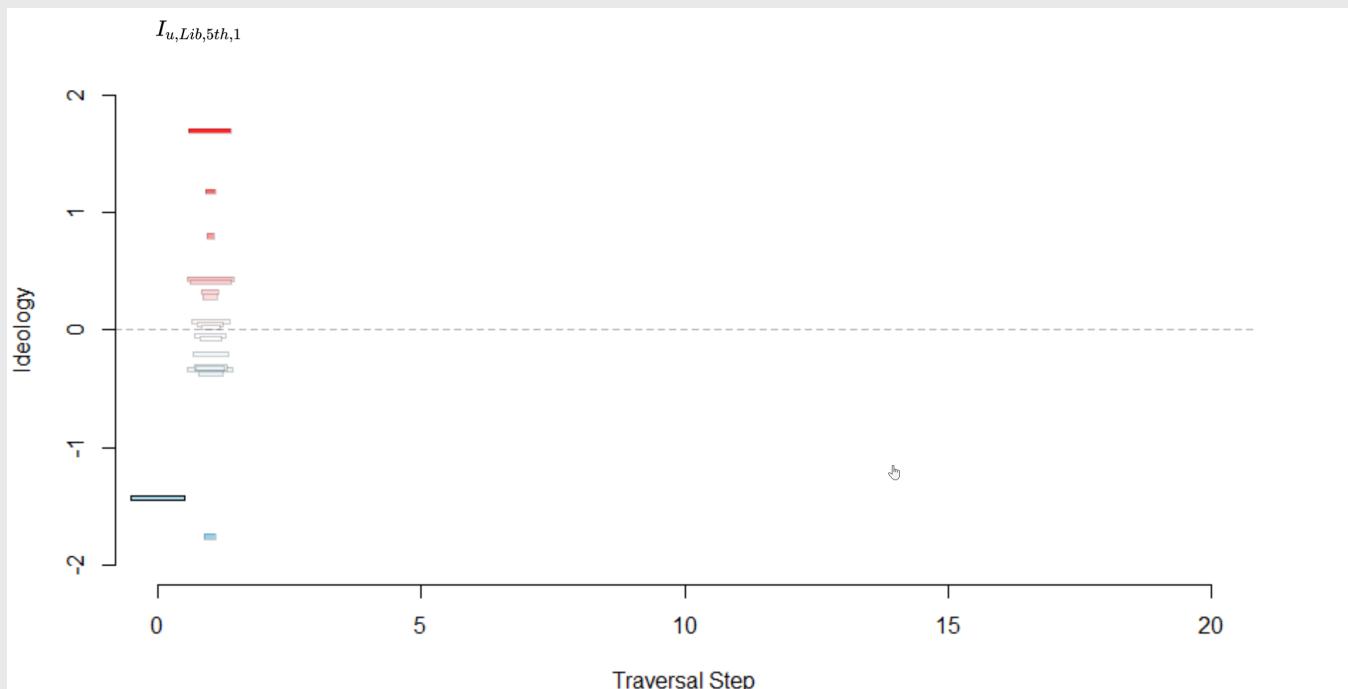
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

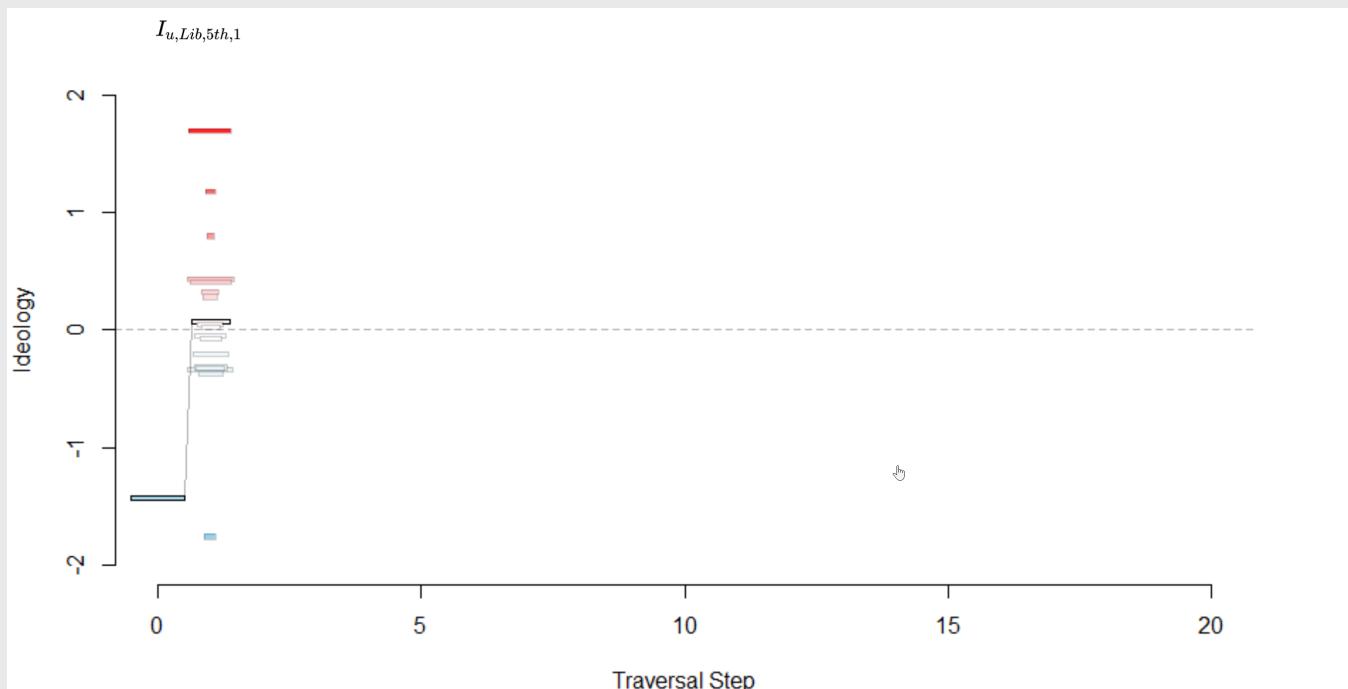
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

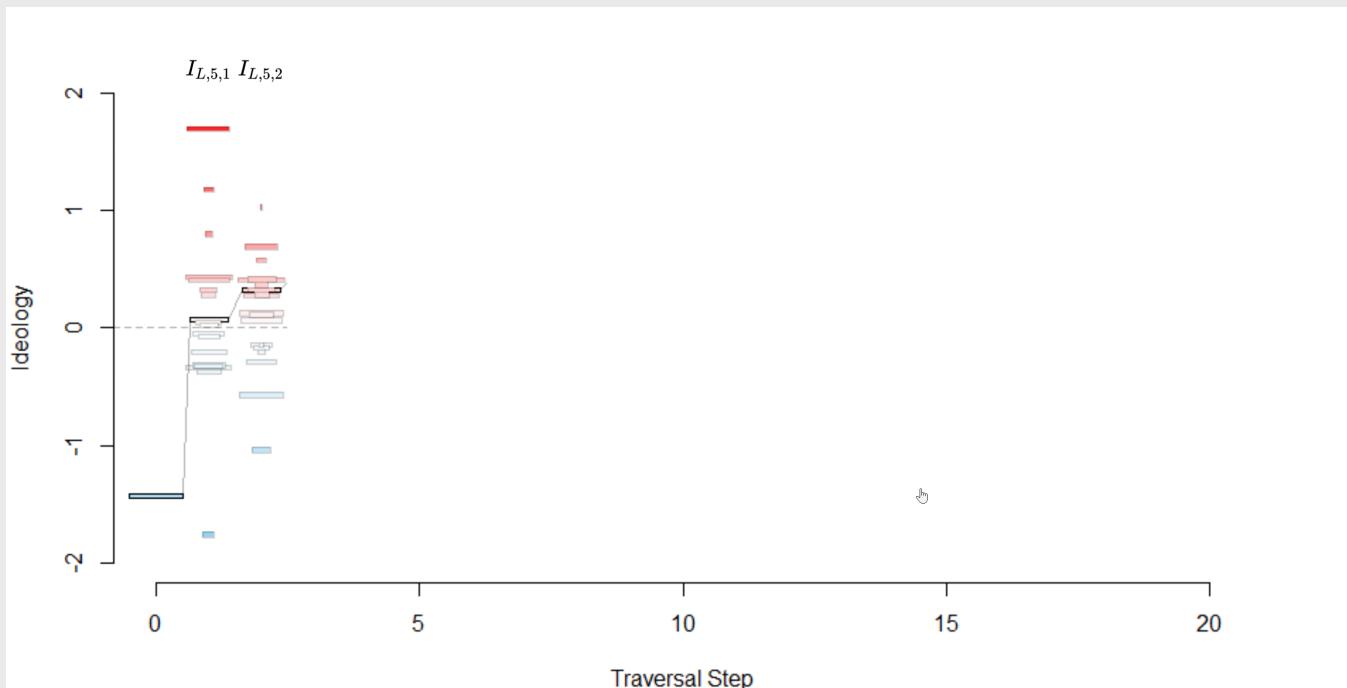
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

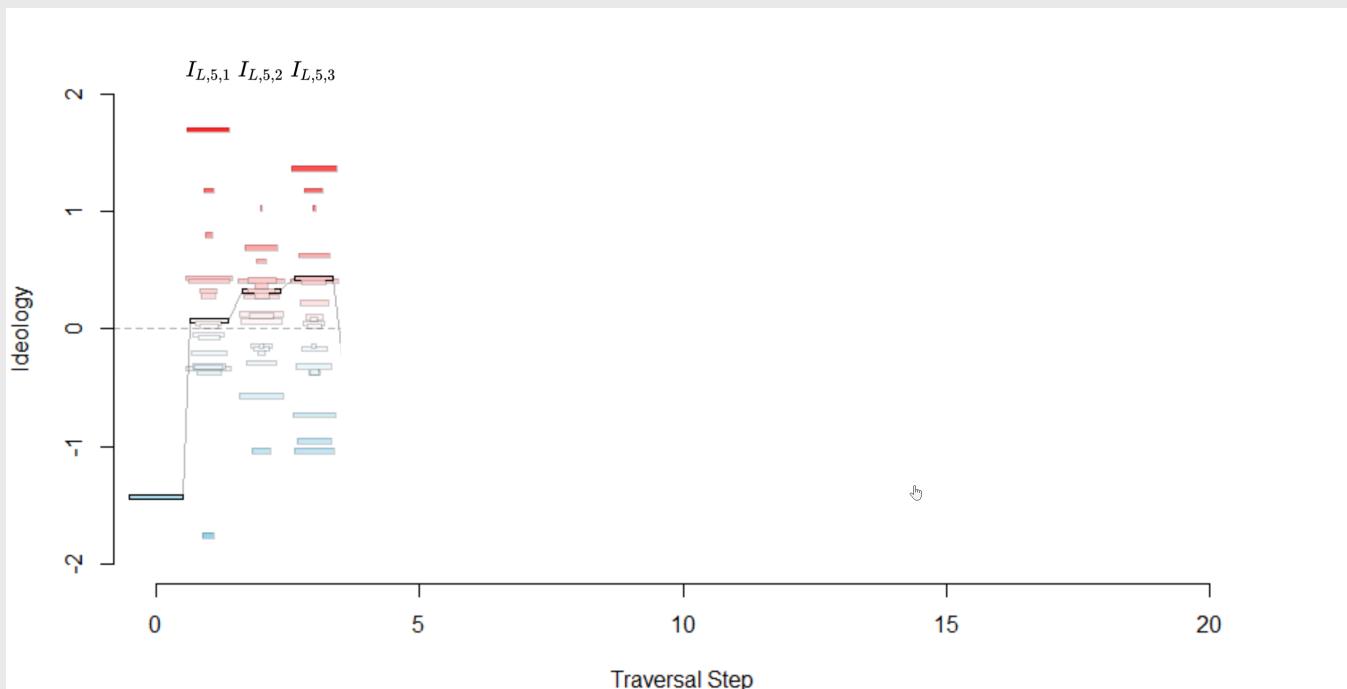
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

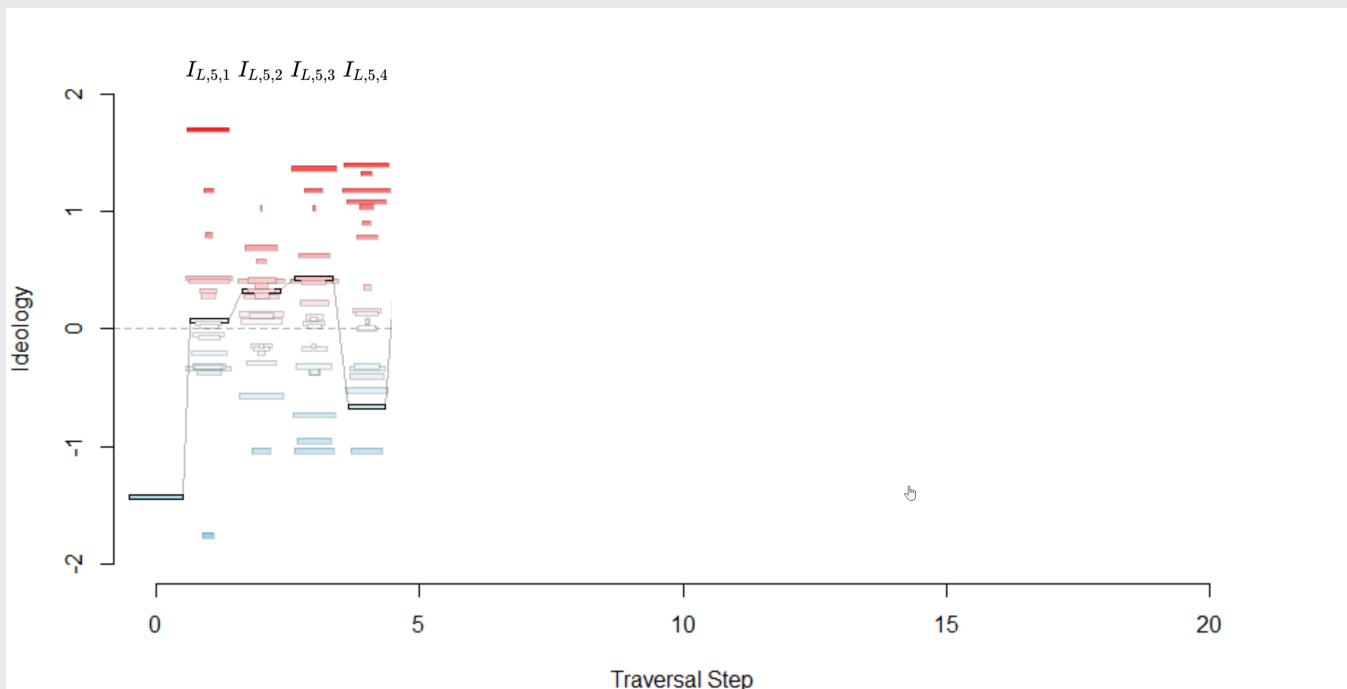
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

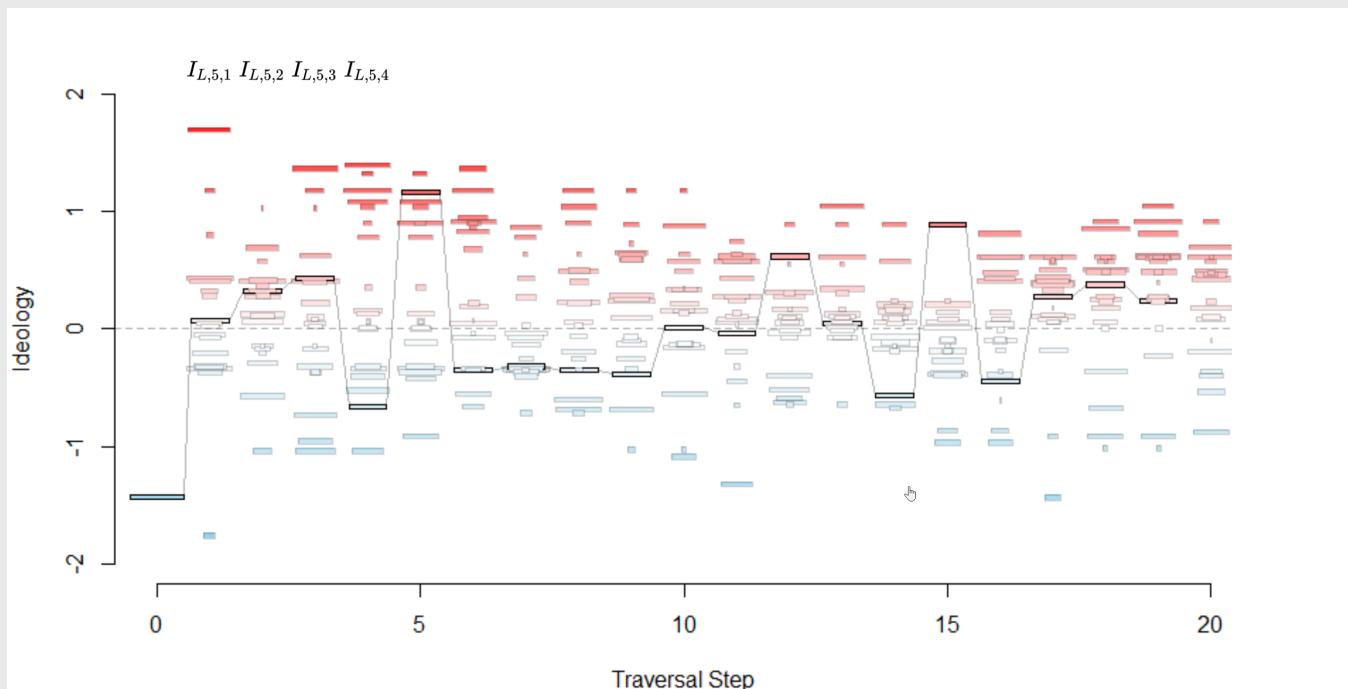
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 3. Data Collection → Analysis

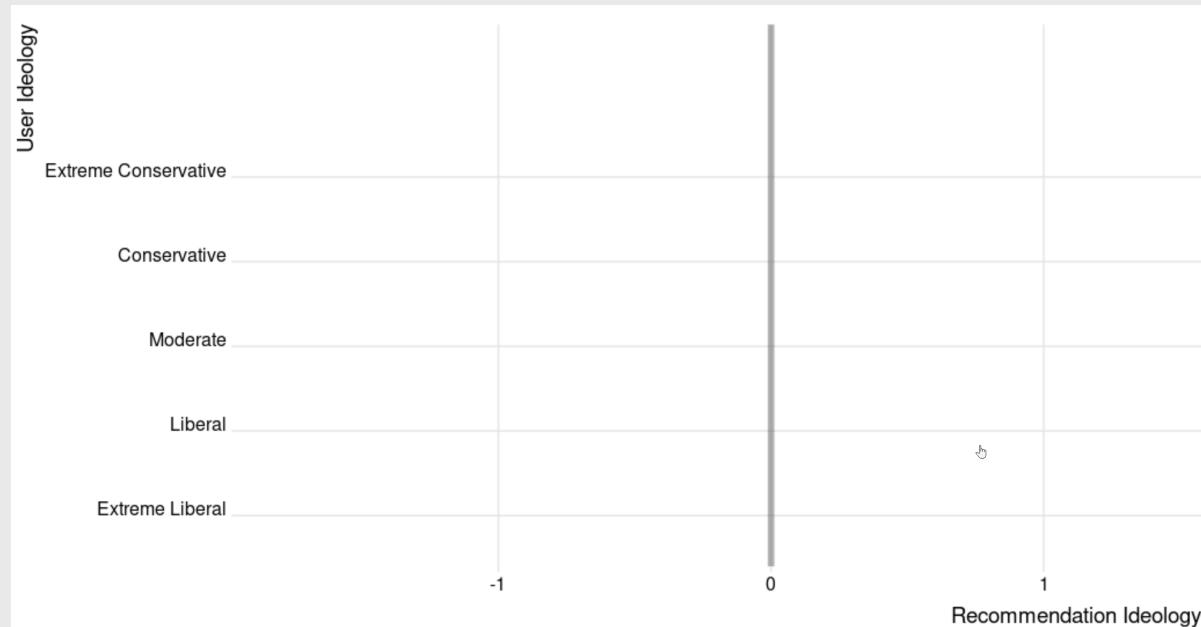
- Analysis is informed by the **data** you have collected...
- ...and the **hypotheses** you have generated



# Research Camp

## 4. Results → Conclusion

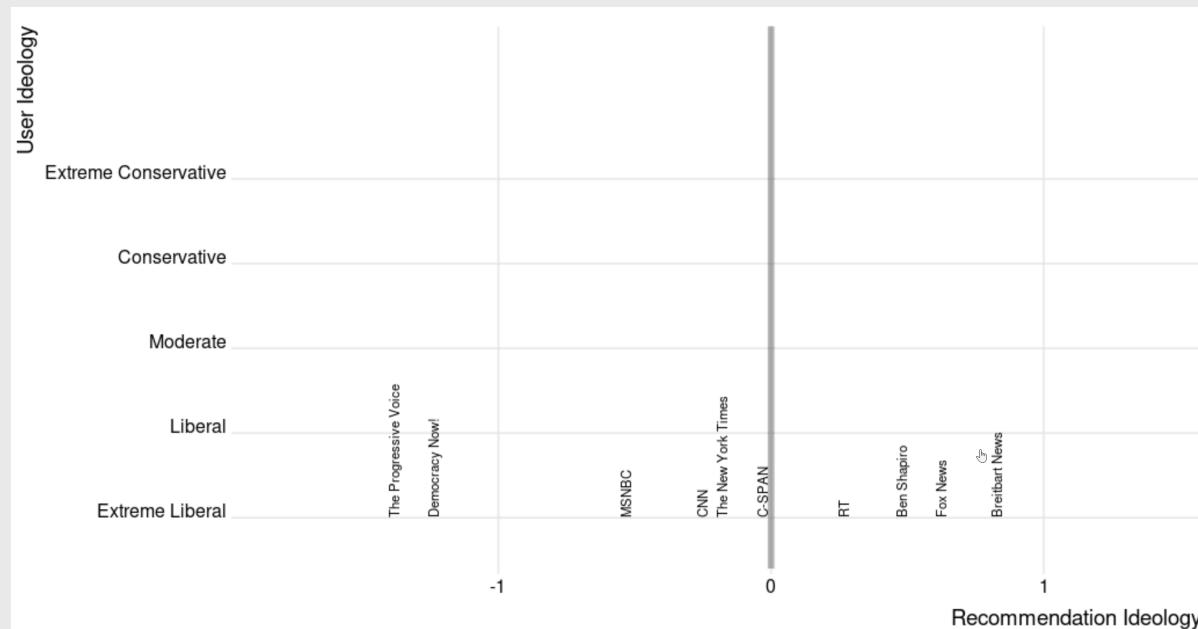
- Results fall out naturally from the analysis...
- ...and must be interpreted in terms of the theory and hypotheses...
- ...to draw conclusions



# Research Camp

## 4. Results → Conclusion

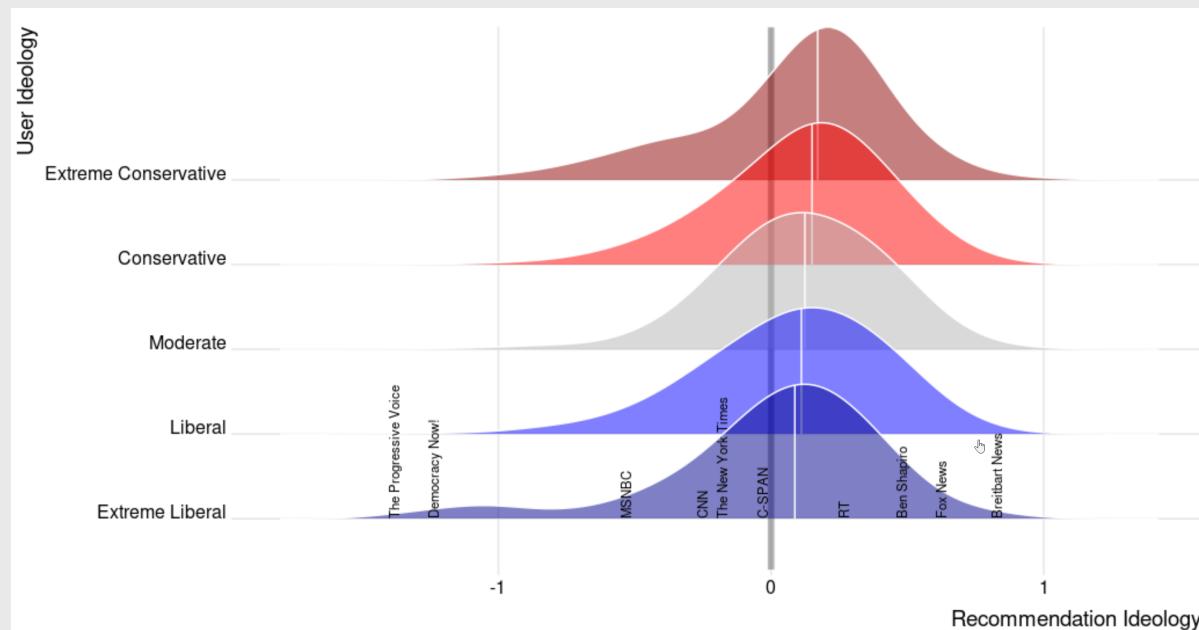
- Results fall out naturally from the analysis...
- ...and must be interpreted in terms of the theory and hypotheses...
- ...to draw conclusions



# 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



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

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

# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

## 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

- **Goal:** Measure the ideology of a YouTube

## 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/jenav...](http://redstate.com/jenav...)

1.2k 323 Comments Share ...



r/Conservative

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...)

308 131 Comments Share ...



r/neutralnews

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...)

303 43 Comments Share ...



r/SandersForPresident

# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

## 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 ...



r/Conservative

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 ...



r/neutralnews

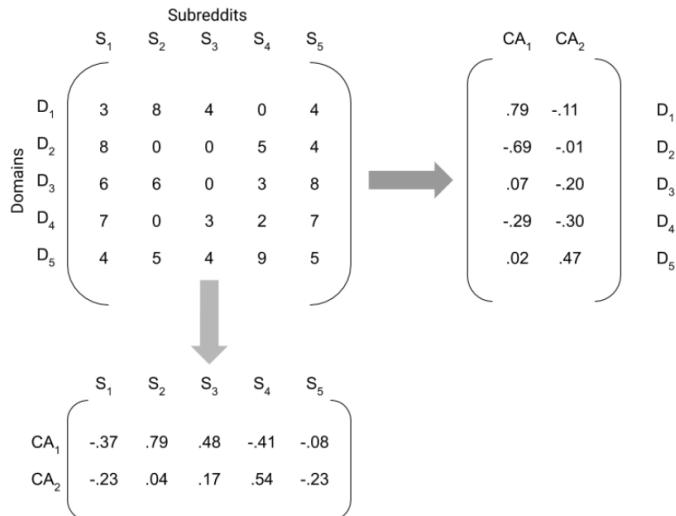
Posted by u/guanaco55 Conservative 3 hours ago

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/SandersForPresident



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

## 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...)

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...)

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...)

303 43 Comments Share ...



r/Conservative

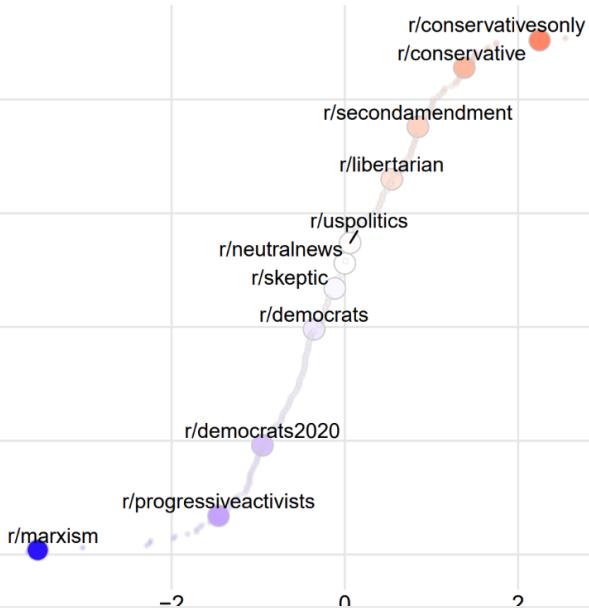


r/neutralnews



r/SandersForPresident

227 Subreddits



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

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

# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

## 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=X3WYQfpsF-I>  
[r/PublicFreakout](#) Posted by u/eseeman 28 days ago

A wand with a twist. I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnhkNLUnew> [Leed.it/pvmdm...](#)  [r/Wandsmith](#) [Posted by u/Unholy3](#) 16 days ago

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=cCtC9vEYA&feature=youtu.be> [Leed.it/junwqm...](#)  [r/ZephyrusG14](#) [Posted by u/alekszukus](#) 1 month ago

why is jimin like this  full video:  
<https://www.youtube.com/watch?v=I1haZ1436M&t=173s> [Meme](#)  [r/heungtan](#) [Posted by u/yengfrifghere](#) 14 days ago

# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

## 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=X3WYQfpsF-I>  
r/PublicFreakout Posted by u/eseeman 28 days ago  
62 Comments Share ...

A wand with a twist. I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnhkNLUnew> Leed.it/powdm! ...  
r/Wandsmith Posted by u/Unholy3 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=cCtPc9vEYA&feature=youtu.be> Leed.it/junwpm...  
r/ZephyrusG14 Posted by u/alekszukus 1 month ago  
123 Comments Share ...

why is jimin like this full video:  
<https://www.youtube.com/watch?v=I1haZl1436M&t=173s>   
r/heungtan Posted by u/yengtrifighthere 14 days ago  
74 Comments Share ...



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

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

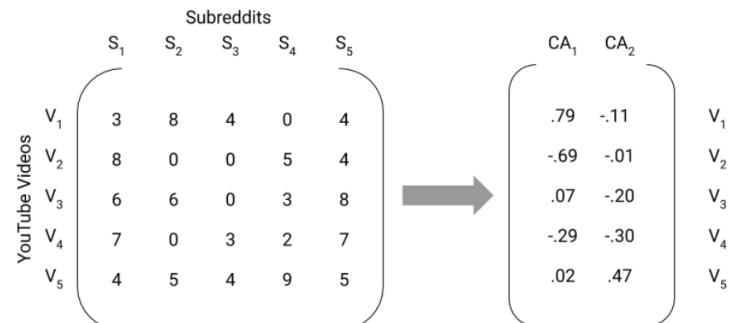
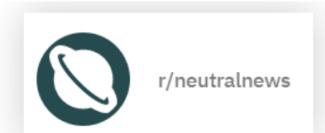
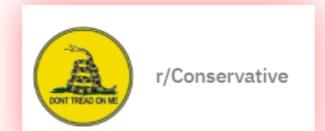
Behavior:  
Sharing Videos + Domain:  
Ideological Reddit

Interview with Thomas Biryani by a reporter from an abc local texas affiliate's live feed:  
<https://www.youtube.com/watch?v=X3WYQfpsF-I>  
r/PublicFreakout Posted by u/eseeman 28 days ago  
62 comments

A wand with a twist. I posted a "how to" on YouTube.  
<https://m.youtube.com/watch?v=7QnhkNUjNew> r/Wandsmith Posted by u/lurhnl3 16 days ago  
44 comments

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=cCtpC9vEYA&feature=youtu.be> r/ZephyrusG14 Posted by u/alekszukus 1 month ago  
123 comments

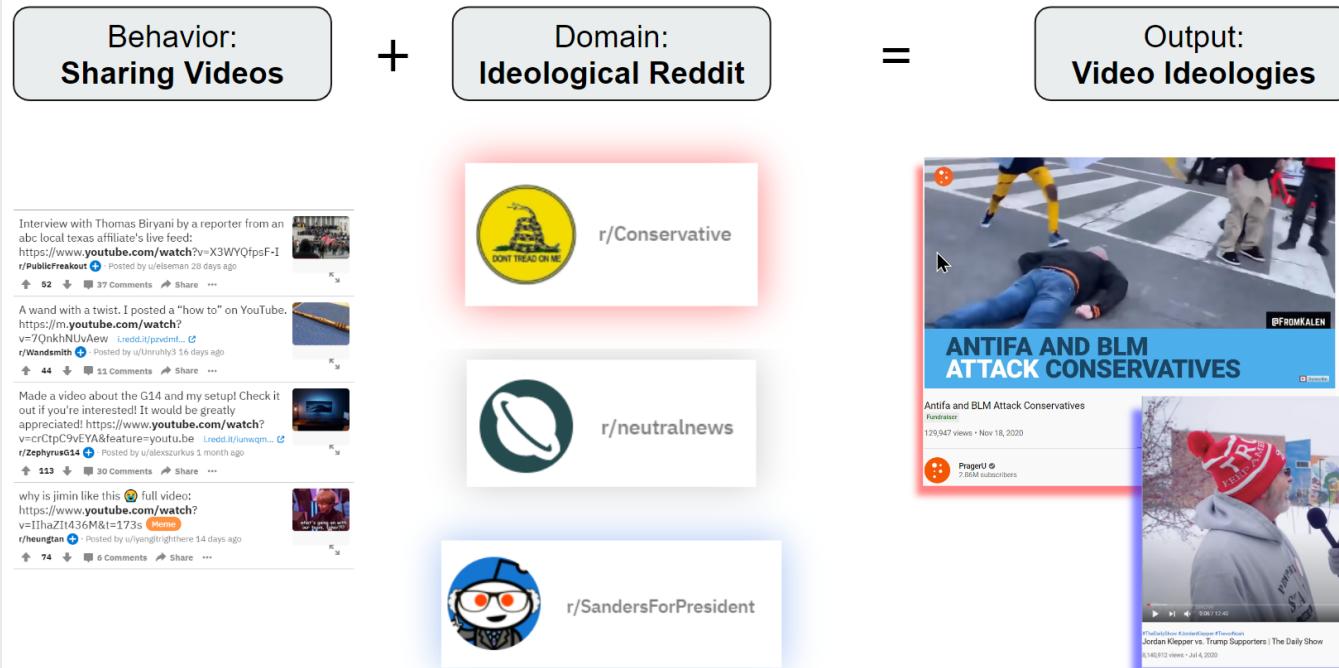
why is jimin like this full video:  
<https://www.youtube.com/watch?v=I1haZ1436M&t=173s> r/heungtan Posted by u/yengtigriffithere 14 days ago  
74 comments



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

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



# Prediction Camp

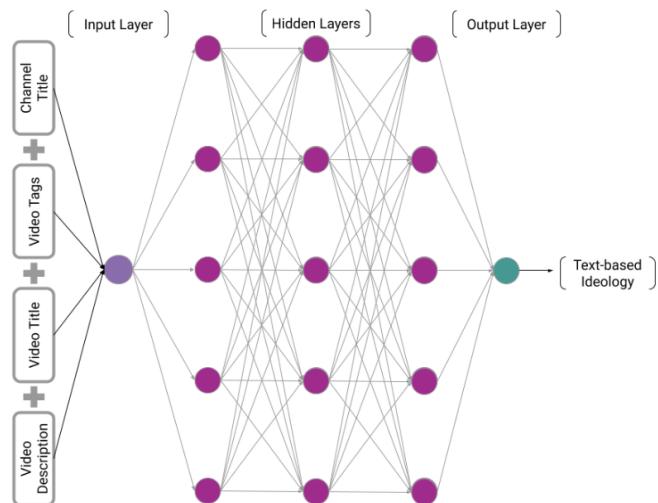
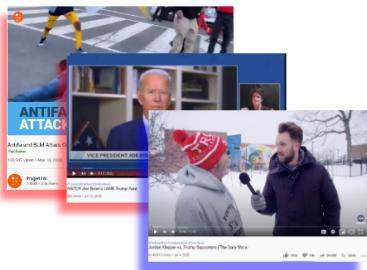
- **Goal:** Measure the ideology of a YouTube

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

Training Data:  
67k Coded Videos

+

Classifier:  
BERT Transformer



# Prediction Camp

- **Goal:** Measure the ideology of a YouTube

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

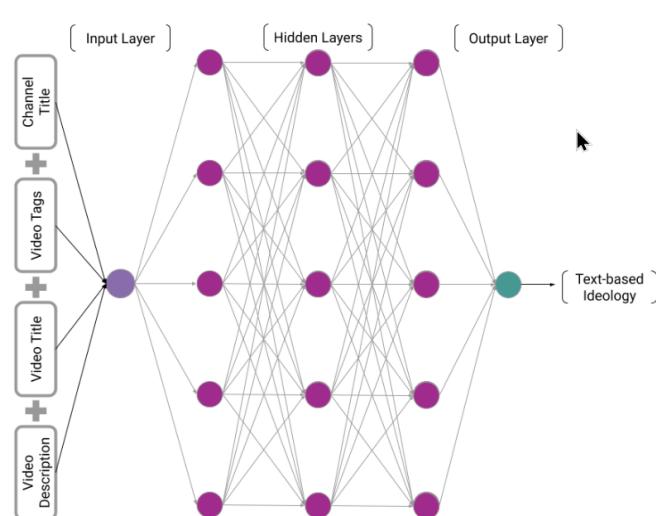
Training Data:  
67k Coded Videos

+

Classifier:  
BERT Transformer

=

Output:  
Any Video's Ideology



# Preview of Semester

- 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
  - You must work through the first homework assignment using an **.Rmd** file

# 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	25%	50
Final Exam	25%	50
Quizzes	10%	20
Totals	100%	200

# The Syllabus

Date	Lecture	DOW	Goal	Assignments	Quizzes
9-Jan-23	Intro to Data Science	M	The scientific method, the camps of analysis	Pset 0 assigned	Quiz 1
11-Jan-23	Intro to R	W	Install and open R, packages, tidyverse functions		Quiz 2
16-Jan-23	BREAK	M			
18-Jan-23	Data Wrangling Part 1	W	Missingness and data types		Quiz 3
23-Jan-23	Data Wrangling Part 2	M	mutate(), ifelse(), and spread()	Pset 1 assigned	Quiz 4
25-Jan-23	R & Data Wrangling Review	W			
30-Jan-23	Univariate Part 1	M	Summaries of a single variable	Pset 2 assigned	Quiz 5
1-Feb-23	Univariate Part 2	W	Visualizations of a single variable		Quiz 6
6-Feb-23	Multivariate Part 1	M	Summaries of two variables	Pset 3 assigned	Quiz 7
8-Feb-23	Multivariate Part 2	W	Summaries of more than two variables		Quiz 8
13-Feb-23	Multivariate Part 3	M	Visualizations of multiple variables	Pset 4 assigned	Quiz 9
15-Feb-23	Multivariate Review	W			
20-Feb-23	Regression Part 1	M	The concept of a linear regression	Pset 5 assigned	Quiz 10
22-Feb-23	Regression Part 2	W	Interpreting a linear regression output and evaluating model performance		Quiz 11
27-Feb-23	Regression Part 3	M	Uncertainty and bootstrapping	Pset 6 assigned	Quiz 12
1-Mar-23	Regression Review	W			
6-Mar-23	Midterm Review	M			
<b>8-Mar-23</b>	<b>Midterm Exam</b>	<b>W</b>			
13-Mar-23	BREAK	M			
15-Mar-23	BREAK	W			
20-Mar-23	Classification Part 1	M	The concept of a logistic regression	Pset 7 assigned	Quiz 13
22-Mar-23	Classification Part 2	W	Interpreting a logistic regression output and evaluating model performance		Quiz 14
27-Mar-23	Classification Part 3	M	Using models for prediction	Pset 8 assigned	Quiz 15
29-Mar-23	Classification Review	W			
3-Apr-23	Clustering Part 1	M	k-means clustering	Pset 9 assigned	Quiz 16
5-Apr-23	NLP Part 2	W	k-means clustering on text		Quiz 17
10-Apr-23	NLP Part 3	M	Sentiment analysis	Pset 10 assigned	Quiz 18
12-Apr-23	NLP Review	W			
17-Apr-23	Advanced Topics in DS	M	Random forests, neural networks, image as data		Quiz 19
19-Apr-23	Ethics	W	The risks of rapid technological change		Quiz 20
24-Apr-23	Final Review	M			
<b>26-Apr-23</b>	<b>Final Exam</b>	<b>W</b>			

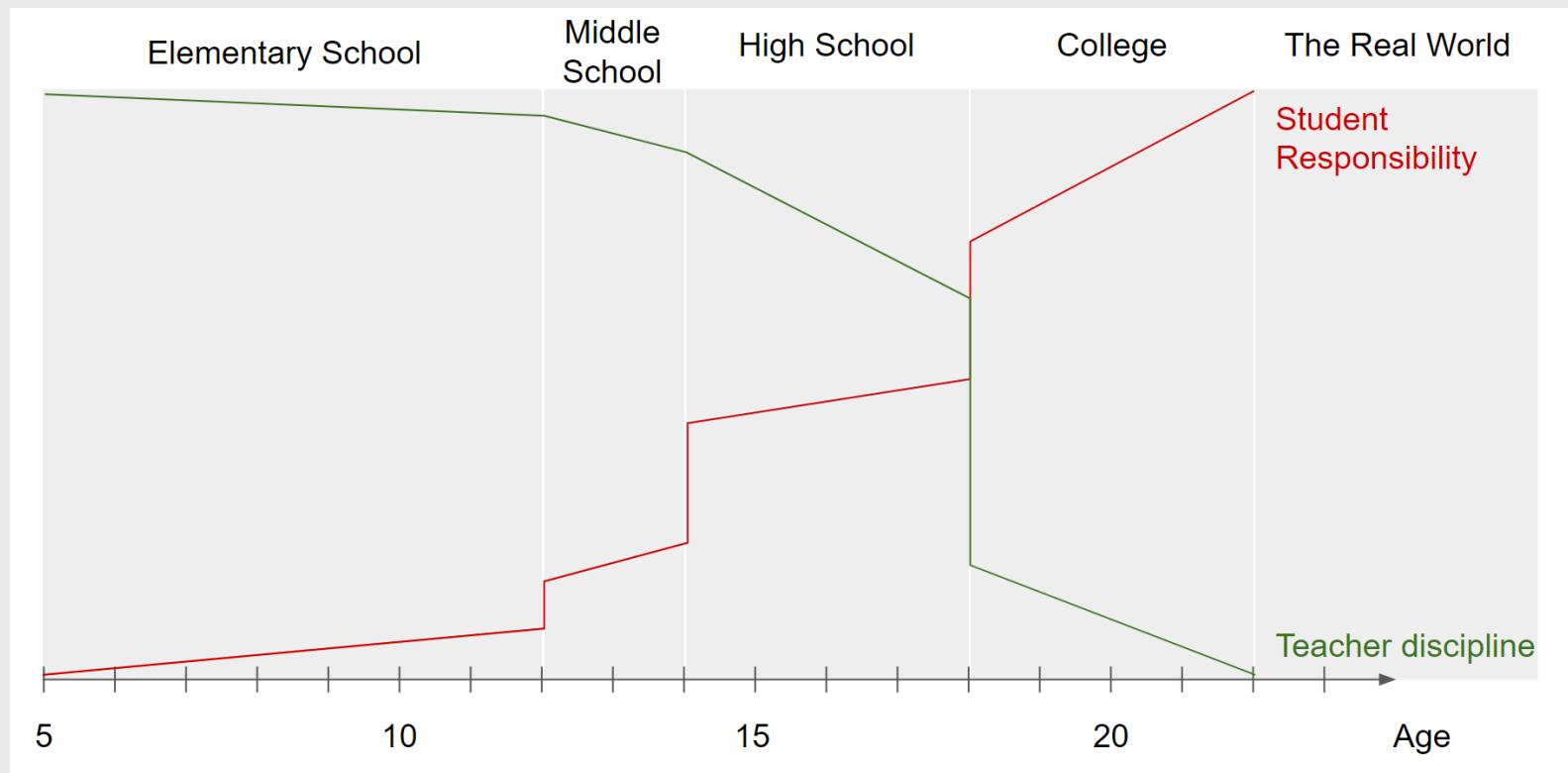
# Honor Code

- 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 (place for **hands-on help**)
- Office hours (place for **hands-on help**)

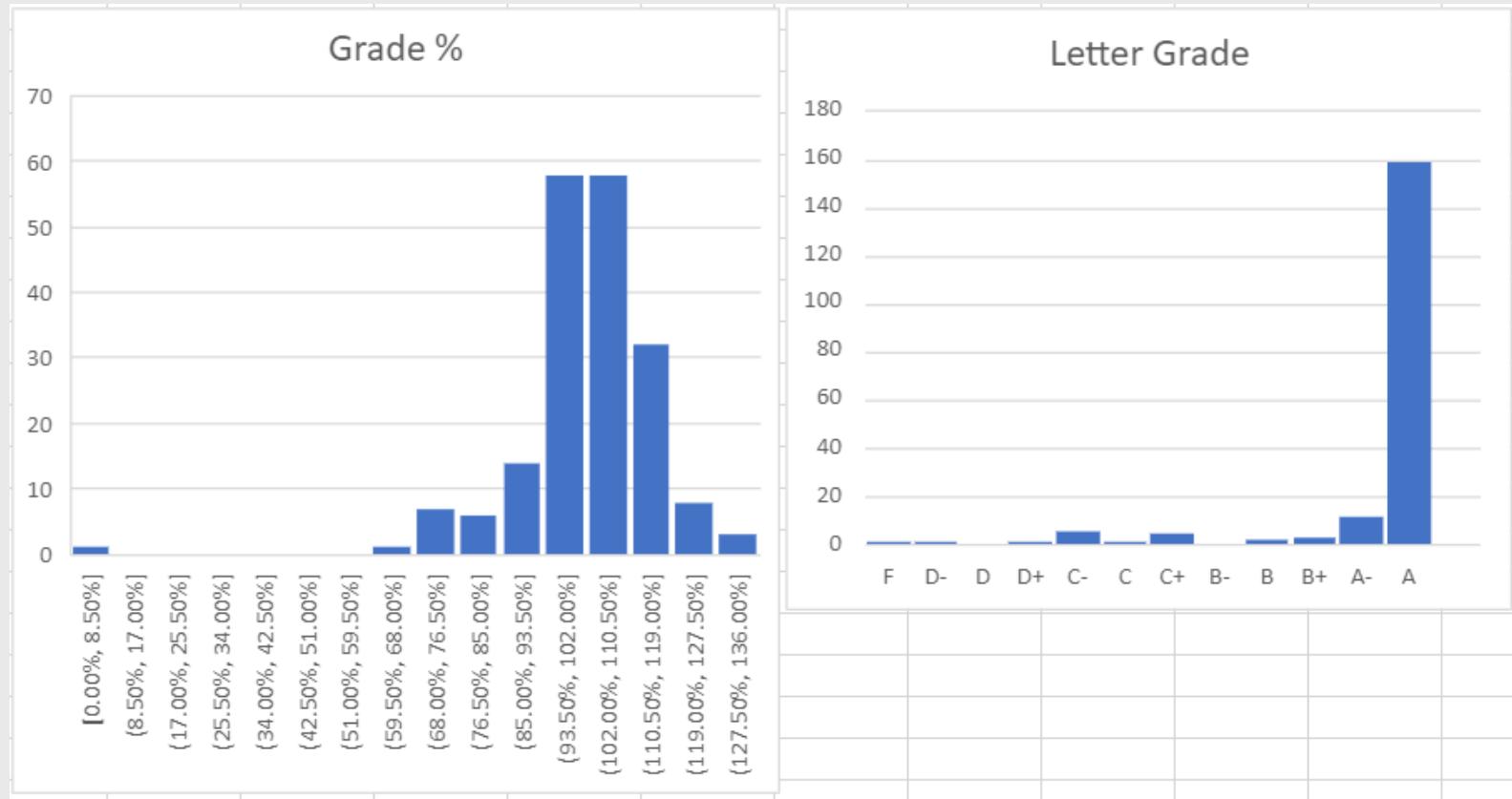
# Teaching Philosophy



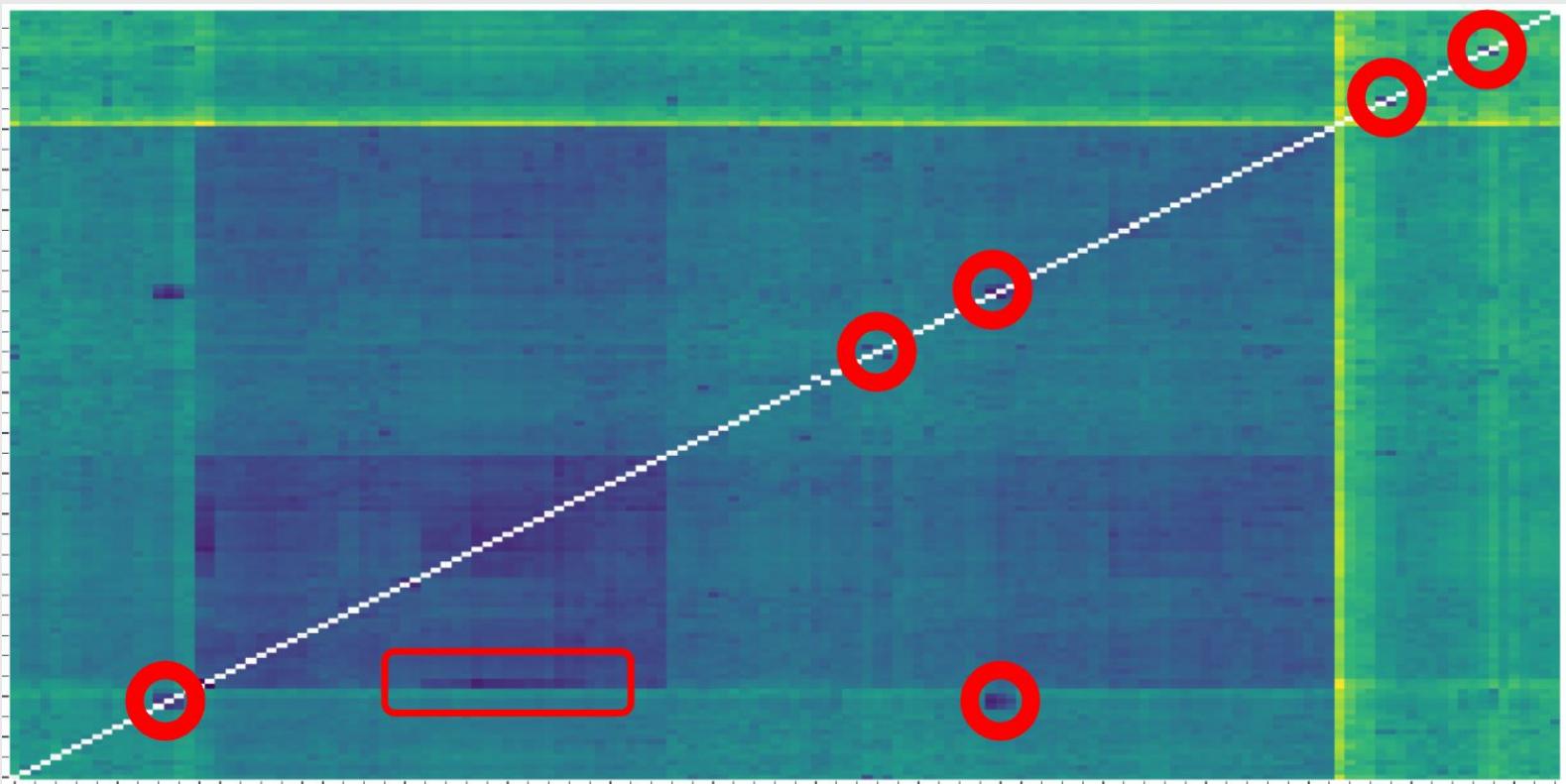
# Teaching Philosophy

- 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



# Previous Semester



# Conclusion

- Go to Brightspace and take the **1st** quiz
  - The password to take the quiz is 3326
- Homework:
  1. Work through Intro\_Data\_Science\_hw.Rmd
  2. Complete Problem Set 0 (on Brightspace)