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

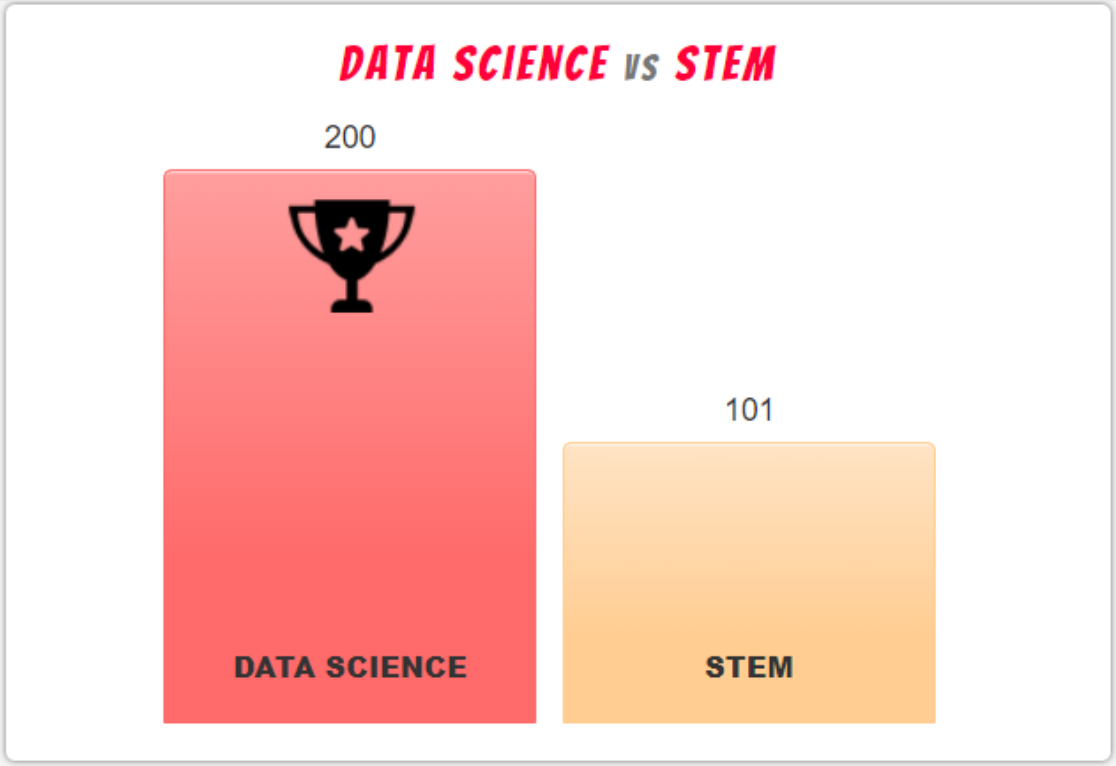
Meet the instructor

- 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

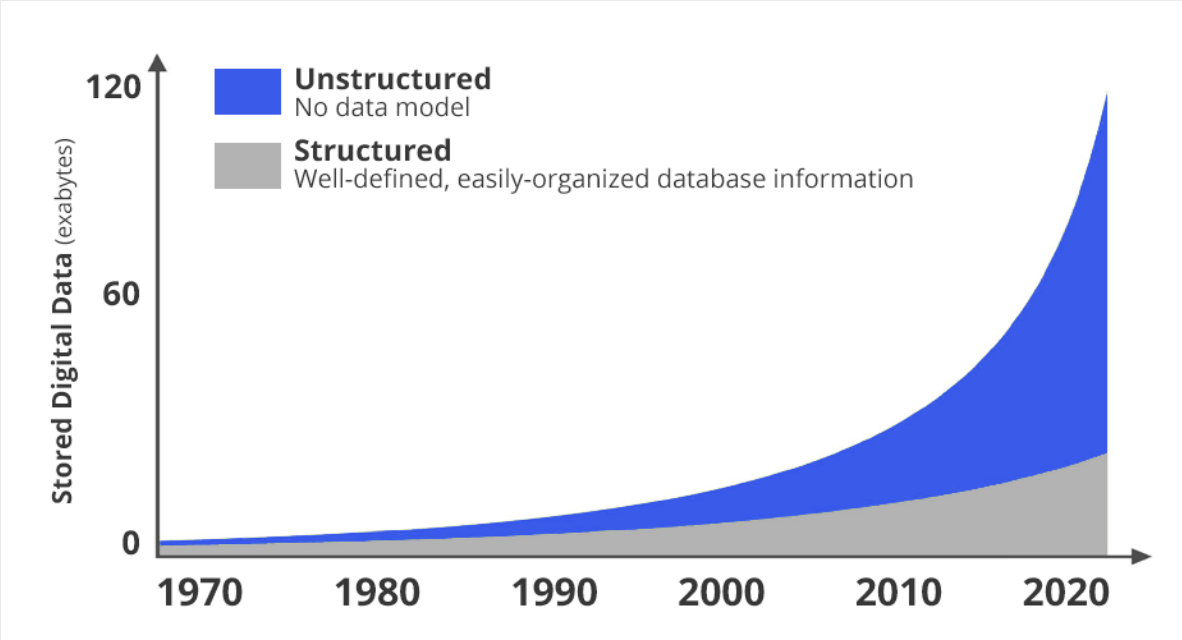
- Current research
  - YouTube + polarization
  - Twitter + misinformation
  - Telegram + white supremacists
- (Throughout the semester, I colorcode data science)

Why are you here?



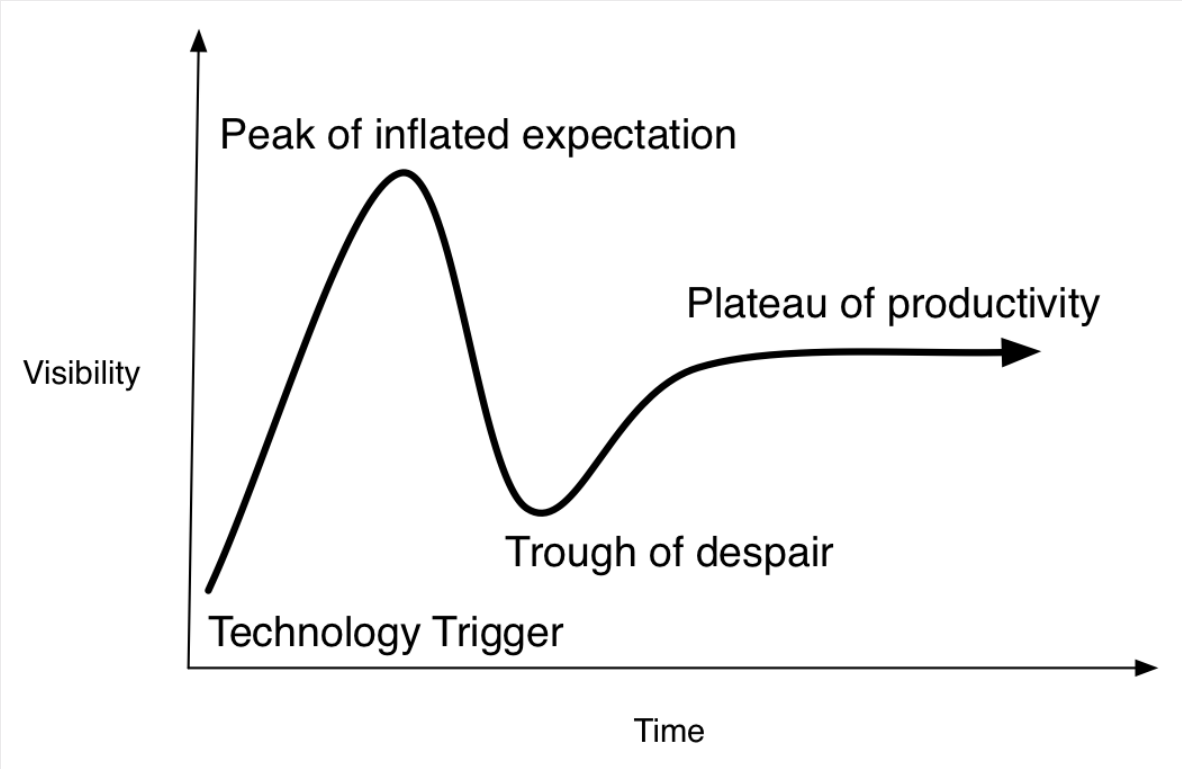
Is this all just a fad?

- No



Is this all just a fad?

- But there are faddish qualities

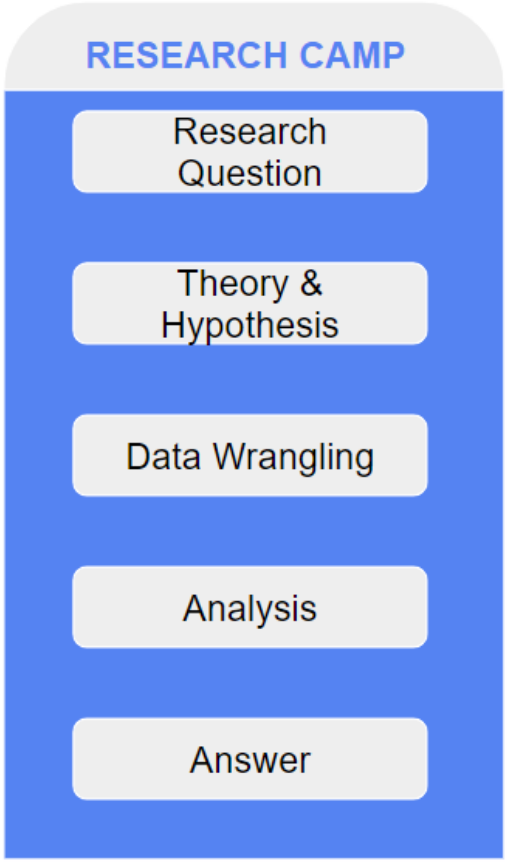


So what IS data science?

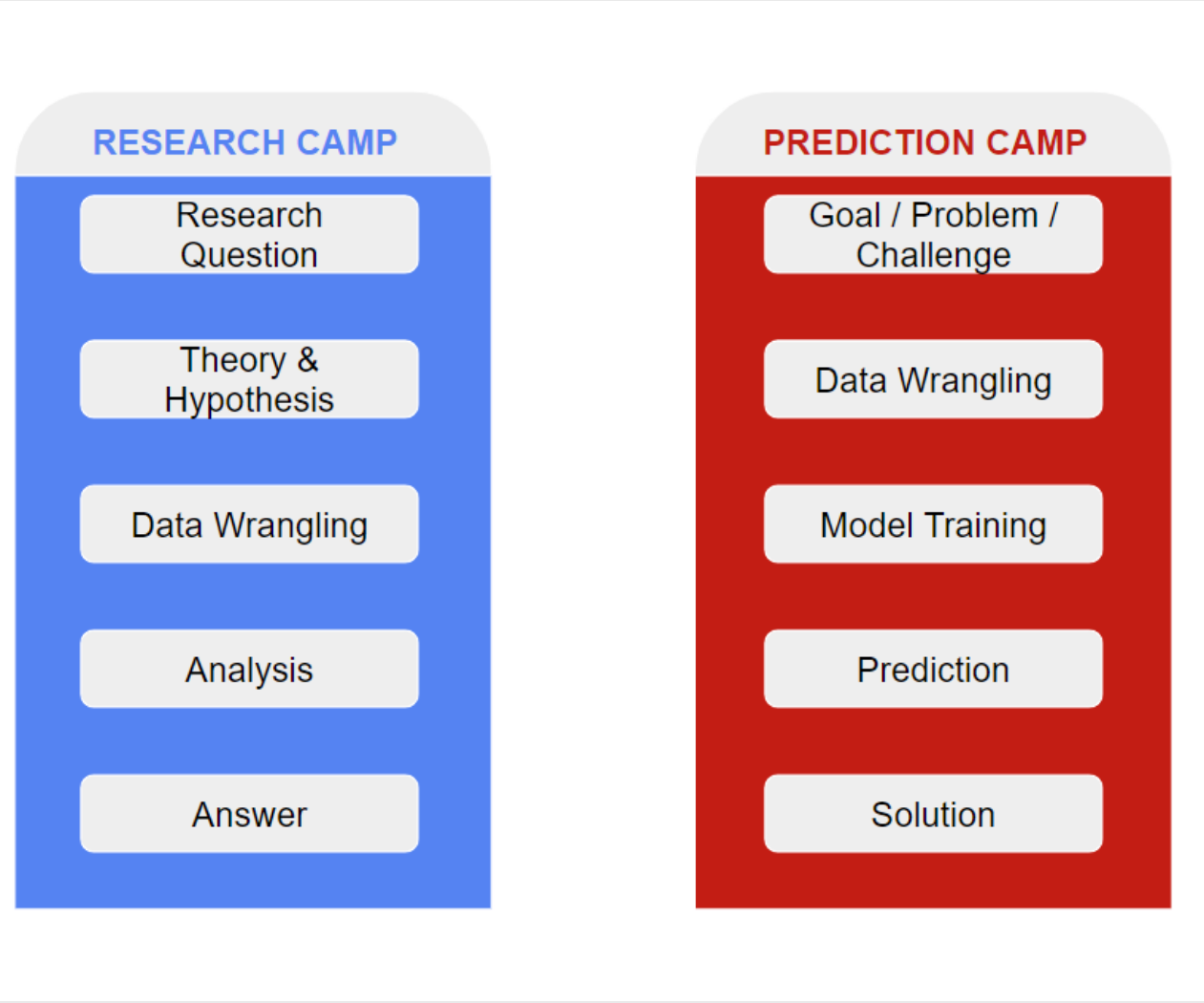
- Split into two camps

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

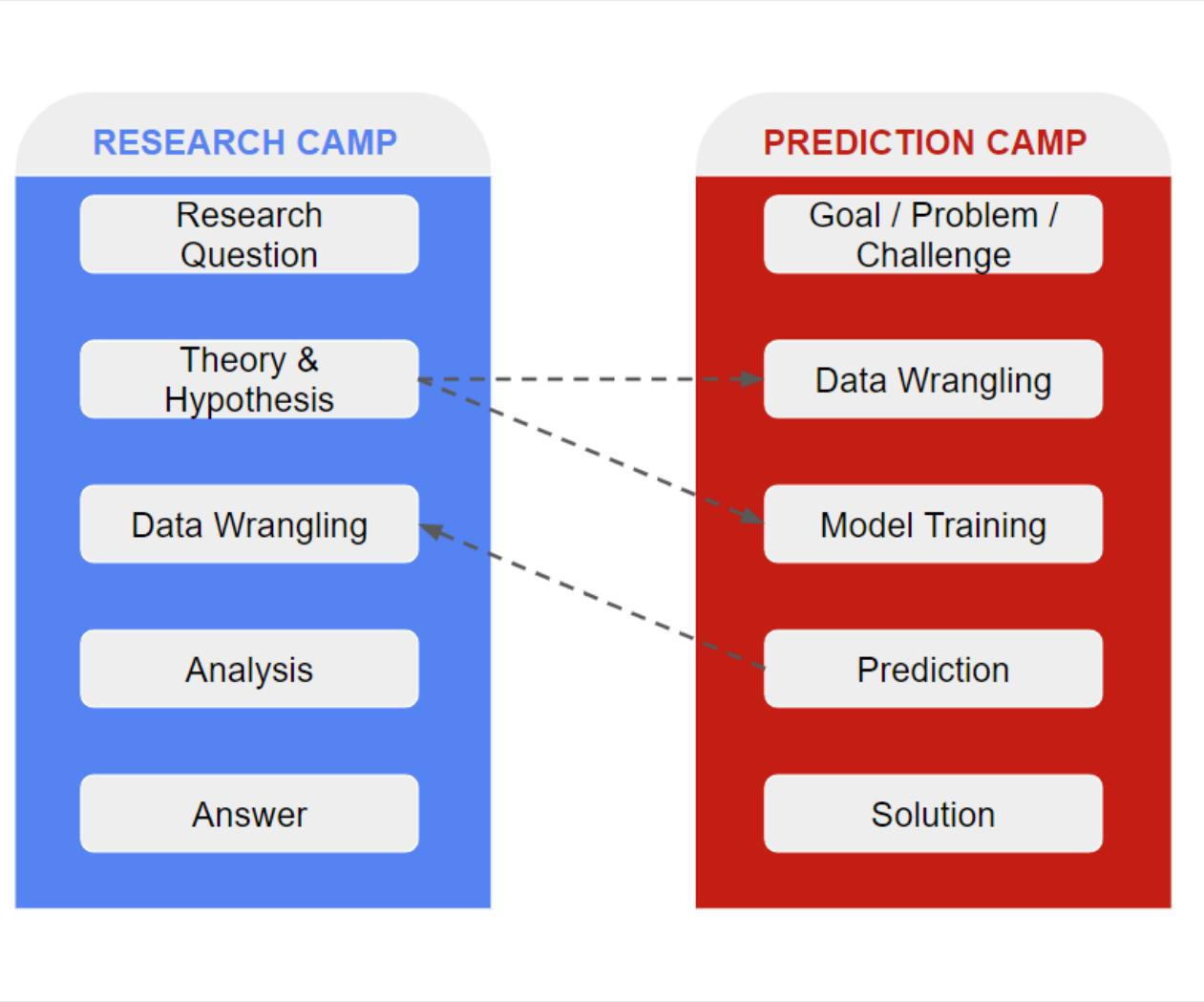
The Two Camps



The Two Camps

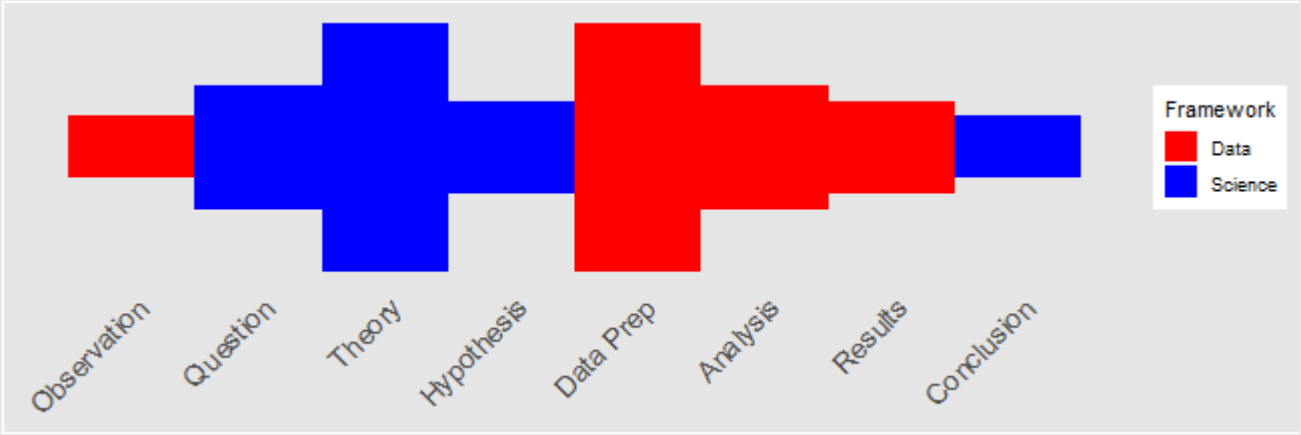


The Two Camps



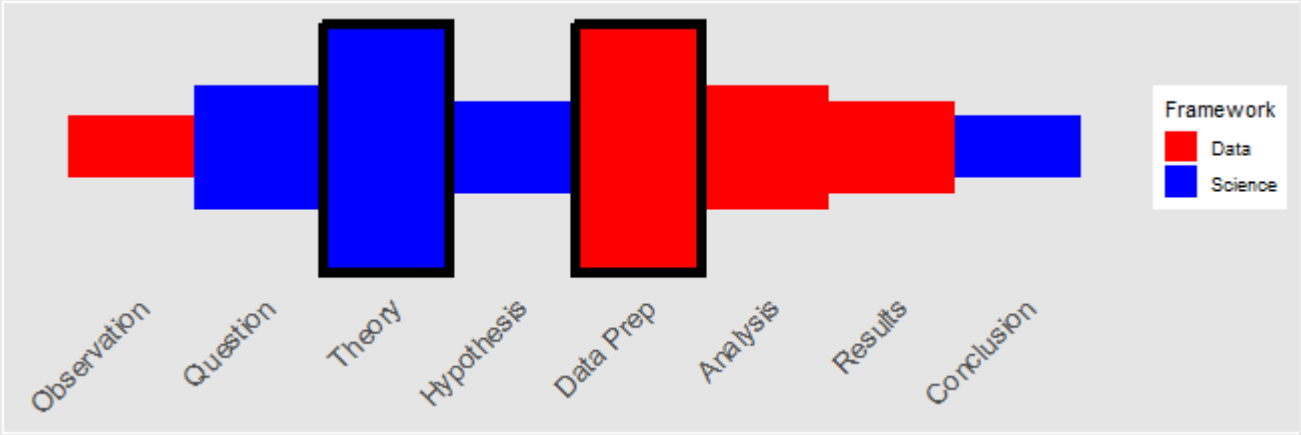
Research Camp

- 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,[1‡] James Bisbee,[1] Angela Lai,[1,4]
Richard Bonneau,[1,3,4] Jonathan Nagler,[1,2,4] Joshua A. Tucker[1,2,4]

[1]Center for Social Media and Politics, New York University

[2]Politics Department, New York University

[3]Biology Department, New York University

[4]Center for Data Science, New York University

[‡]To whom correspondence should be addressed: meganbrown@nyu.edu

August 23, 2023

Research Camp

1. **Observation** → **Question**
    - Observation is facilitated by data (Descriptive analysis)

Research Camp

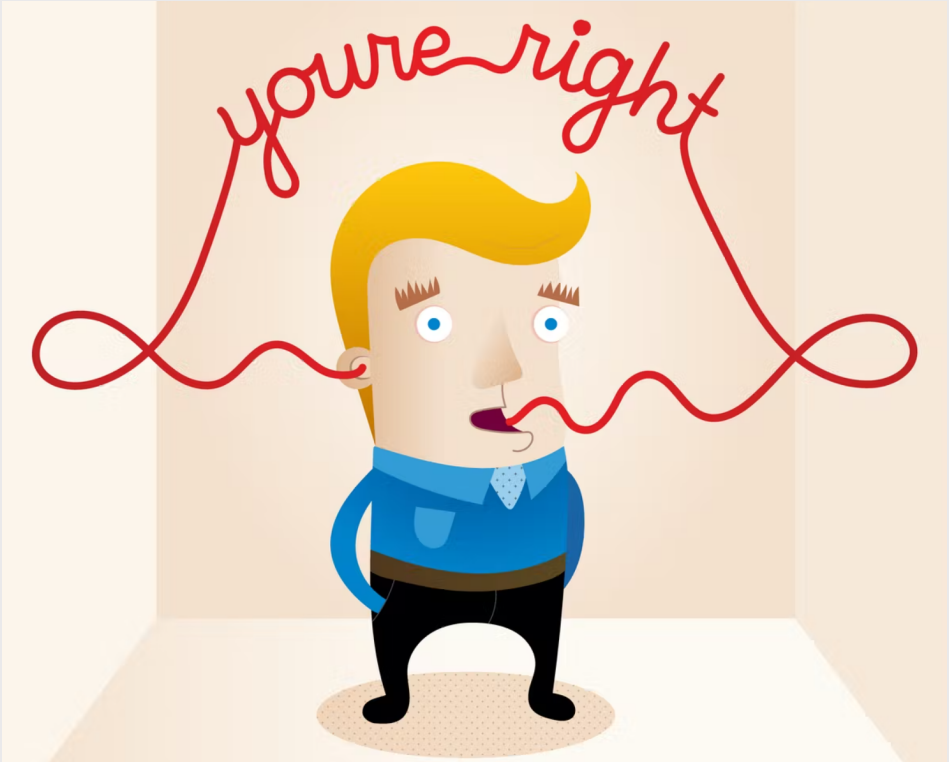
1. **Observation** → **Question**
    - Observation is facilitated by data (Descriptive analysis)

Research Camp

1. **Observation** → **Question**
    - Observation is facilitated by data (Descriptive analysis)

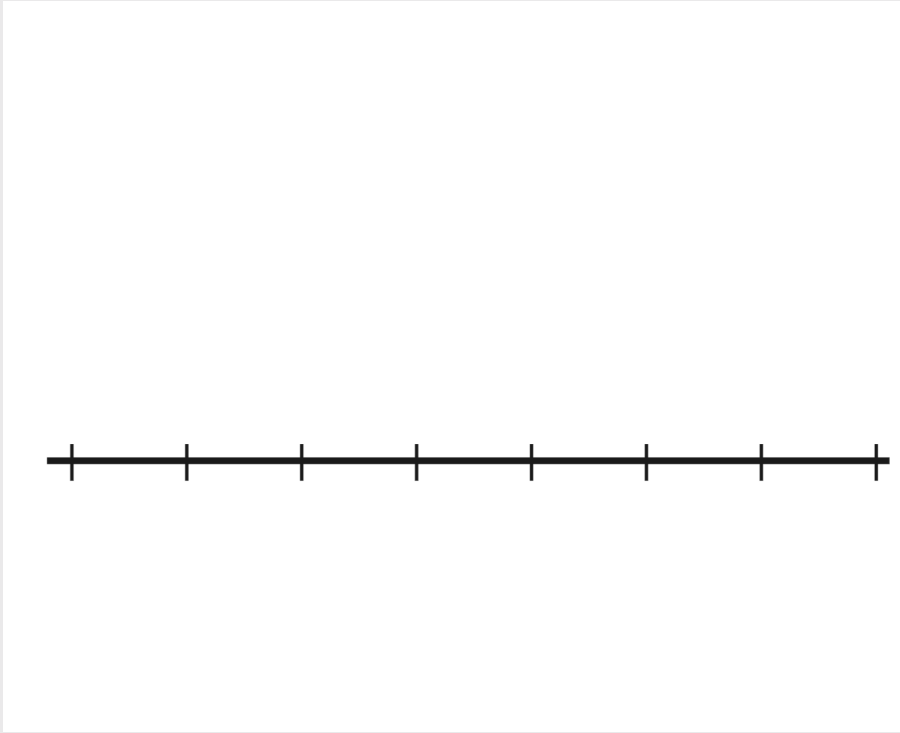
Research Camp

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



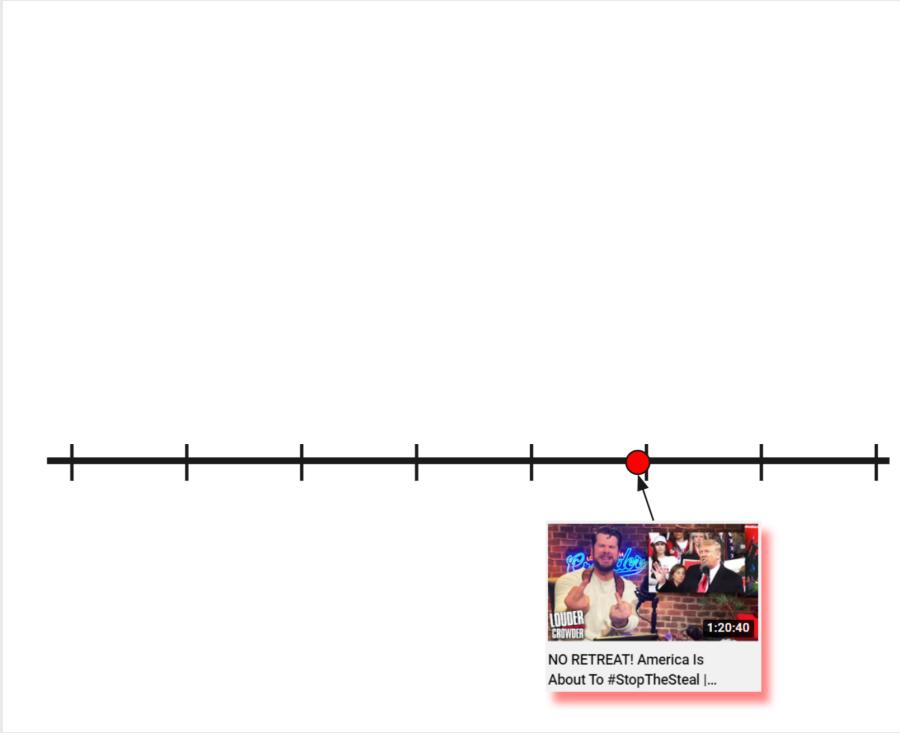
Research Camp

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

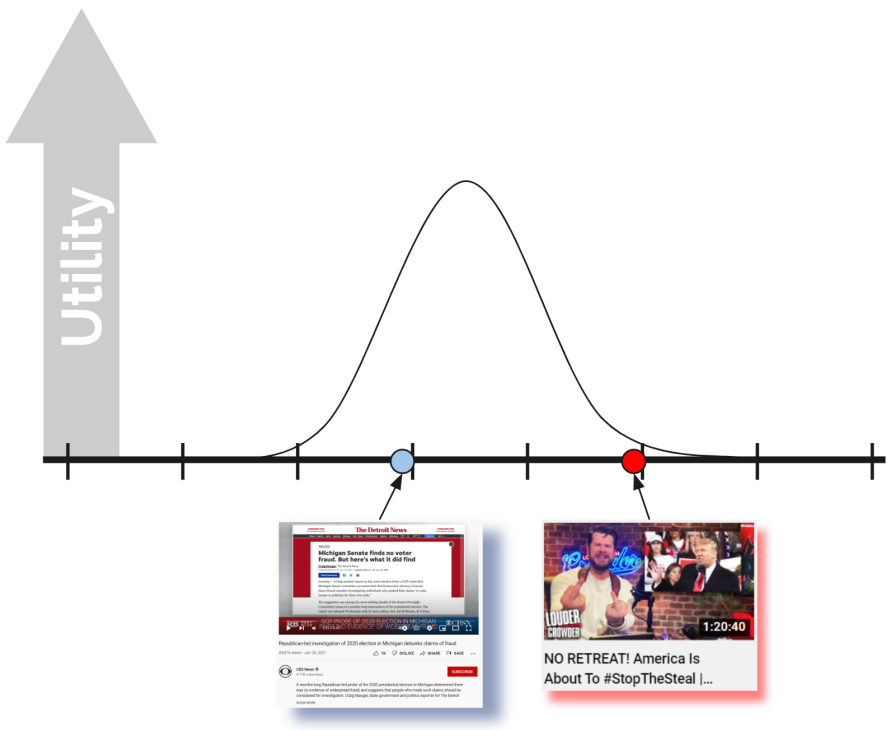


Research Camp

2. **Theory** → **Hypothesis**
    - Theorizing requires abstraction & simplification
    - I.e., people (in general) avoid conflict

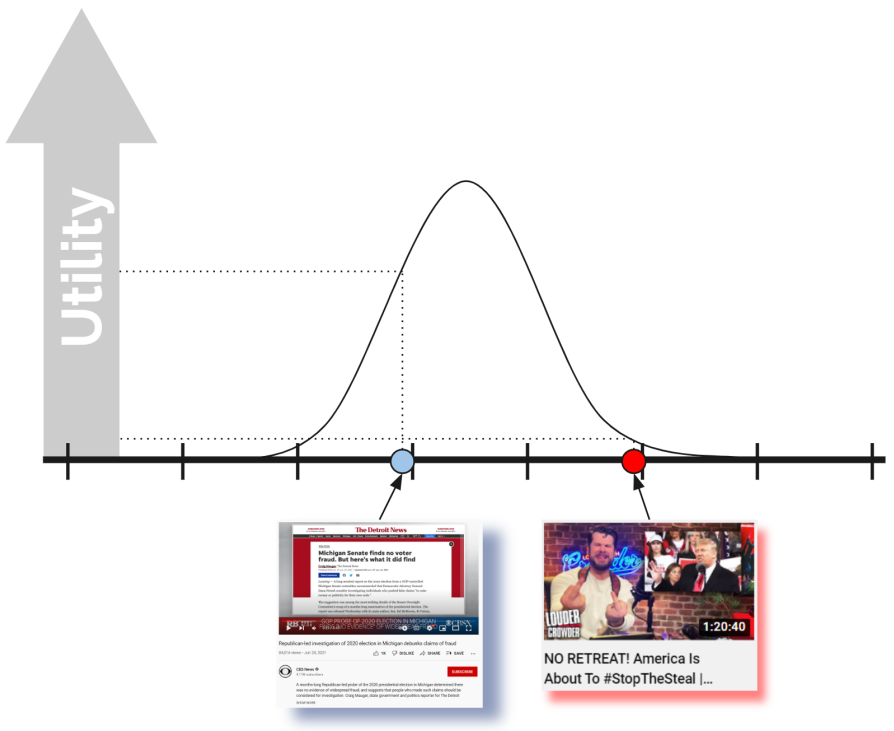
Research Camp

2. **Theory** → **Hypothesis**
    - Theorizing requires abstraction & simplification
    - I.e., people (in general) avoid conflict



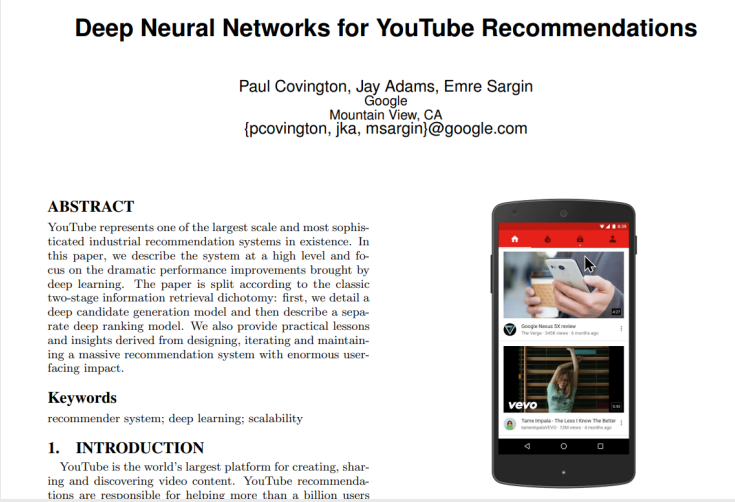
Research Camp

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

ABSTRACT

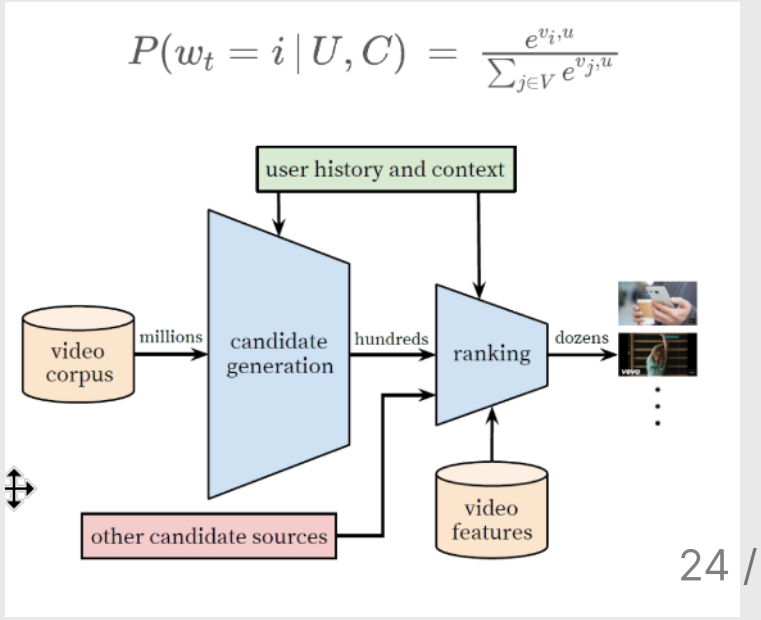
YouTube represents one of the largest scale and most sophisticated industrial recommendation systems in existence. In this paper, we describe the system at a high level and focus on the dramatic performance improvements brought by deep learning. The paper is split according to the classic two-stage information 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

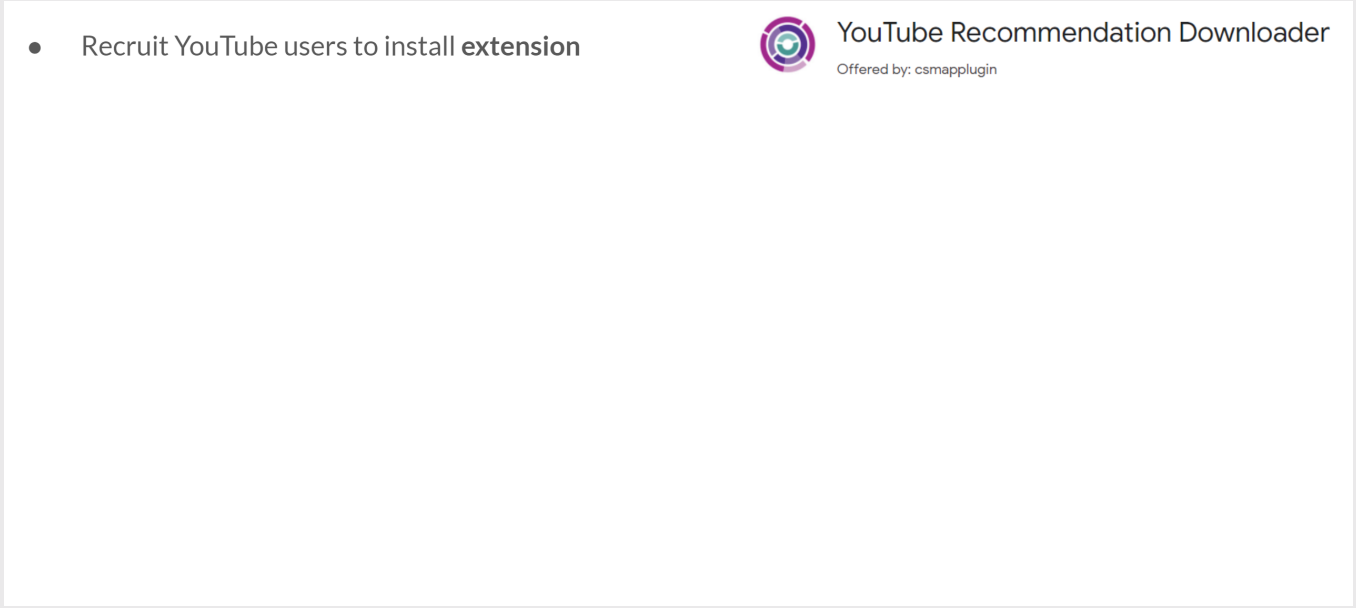


Research Camp

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 suggested liberal content to liberals and conservative content to conservatives.*

Research Camp

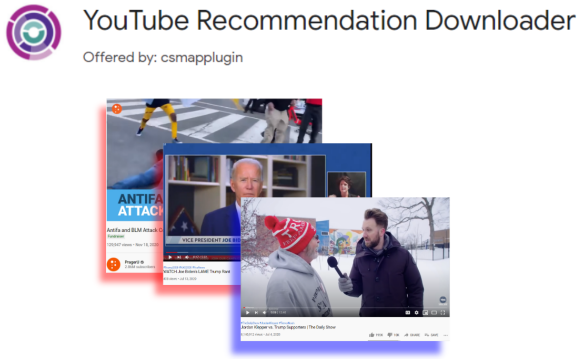
3. **Data Collection / Wrangling → Analysis**
    - Data collection separates "Data Science"...
    - ...from "Science, with data"



Research Camp

3. **Data Collection / Wrangling → Analysis**
    - Data collection separates "Data Science"...
    - ...from "Science, with data"

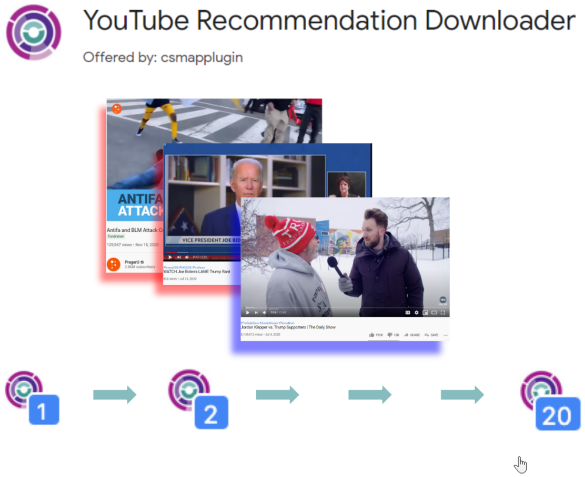
- Recruit YouTube users to install **extension**
- Start on randomly assigned **seed video**



Research Camp

3. **Data Collection / Wrangling → Analysis**
    - Data collection separates "Data Science"...
    - ...from "Science, with data"

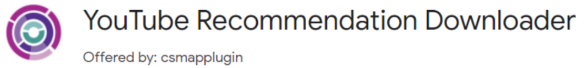
- Recruit YouTube users to install **extension**
- Start on randomly assigned **seed video**
- Follow **traversal rule** to select recommended video



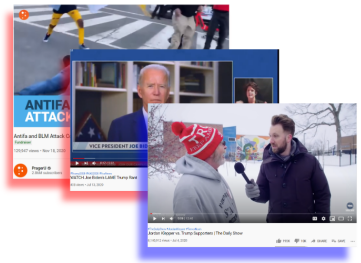
Research Camp

3. **Data Collection / Wrangling → Analysis**
    - Data collection separates "Data Science"...
    - ...from "Science, with data"

- Recruit YouTube users to install **extension**



- Start on randomly assigned **seed video**

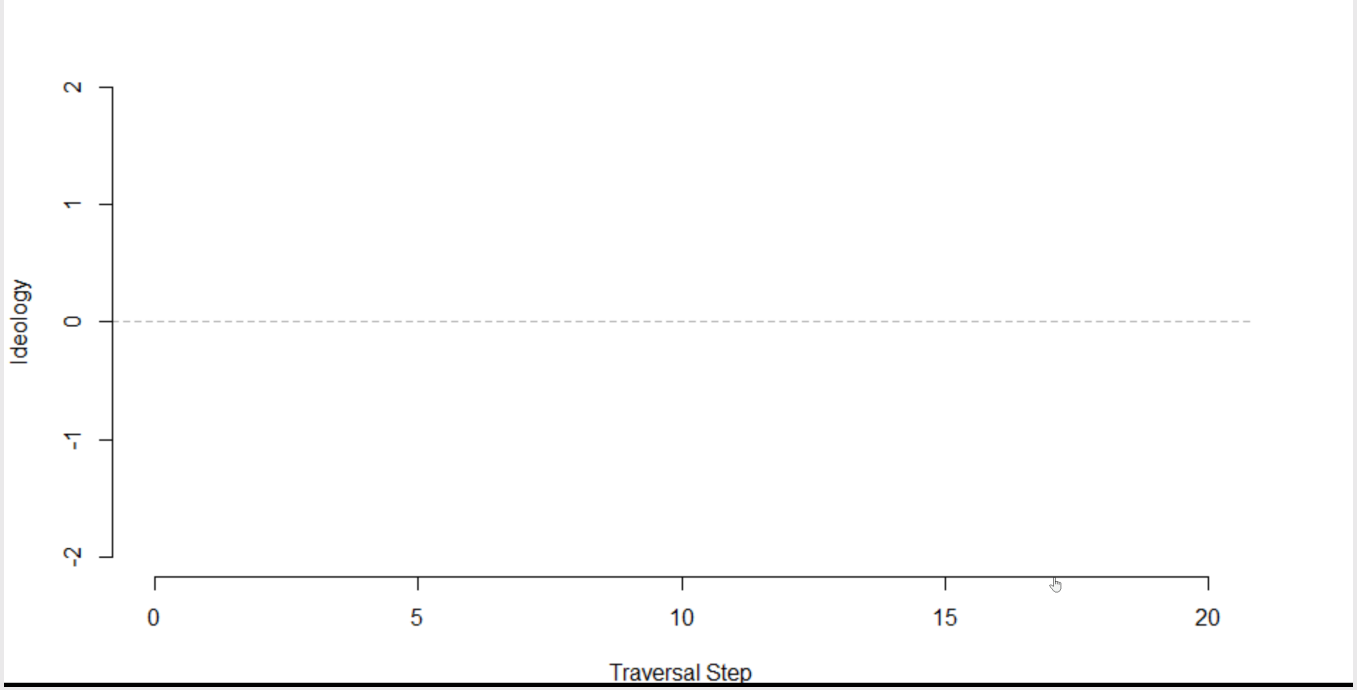


- Follow **traversal rule** to select recommended video

- 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

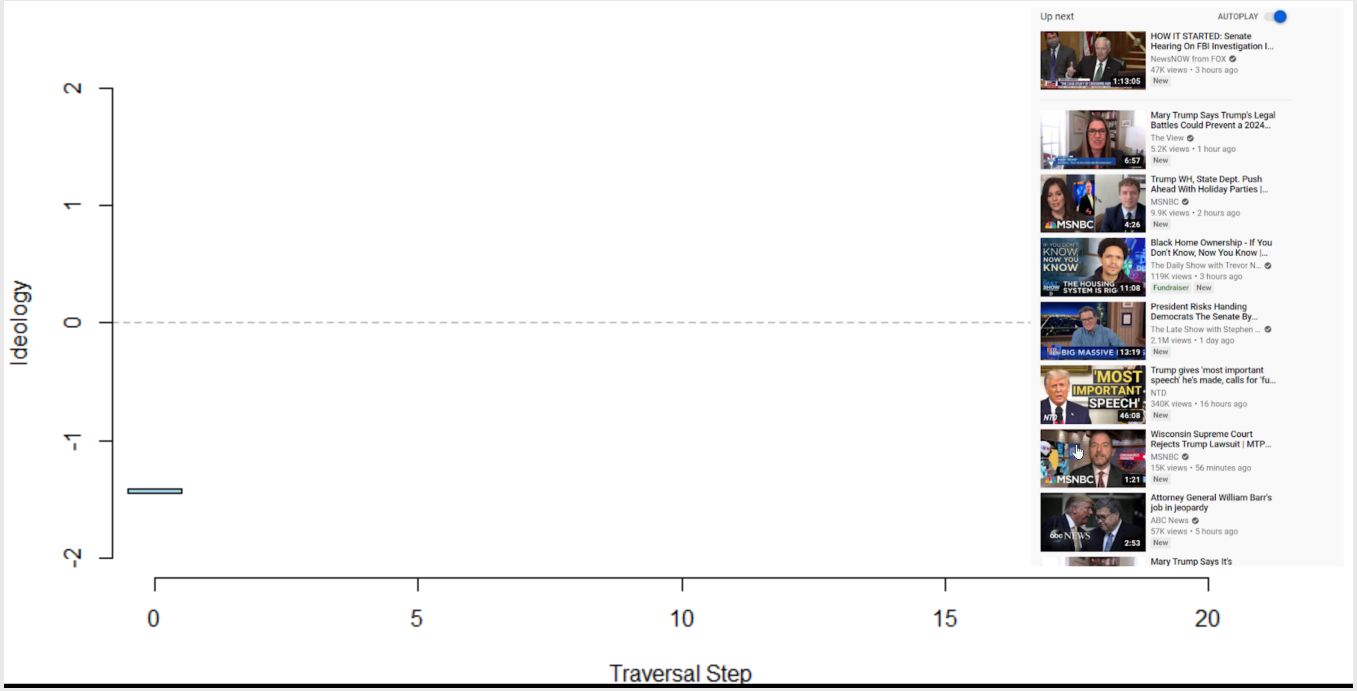
3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated

Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated

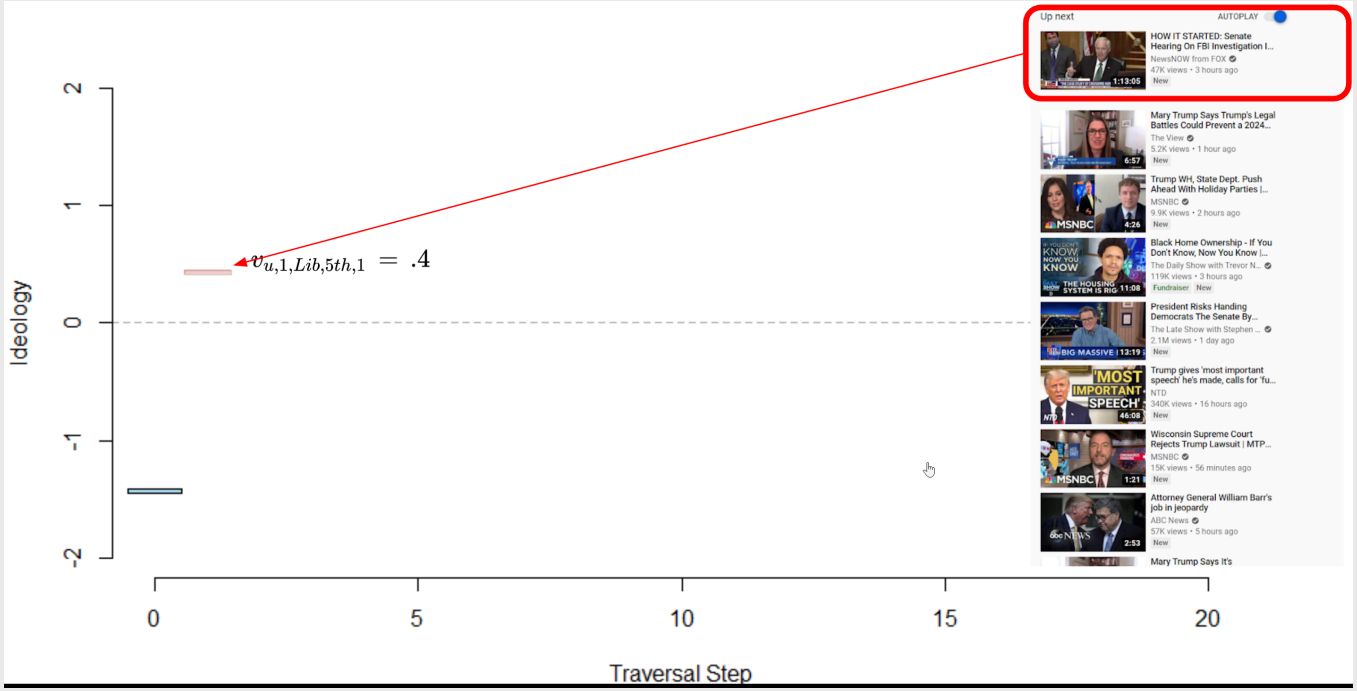
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



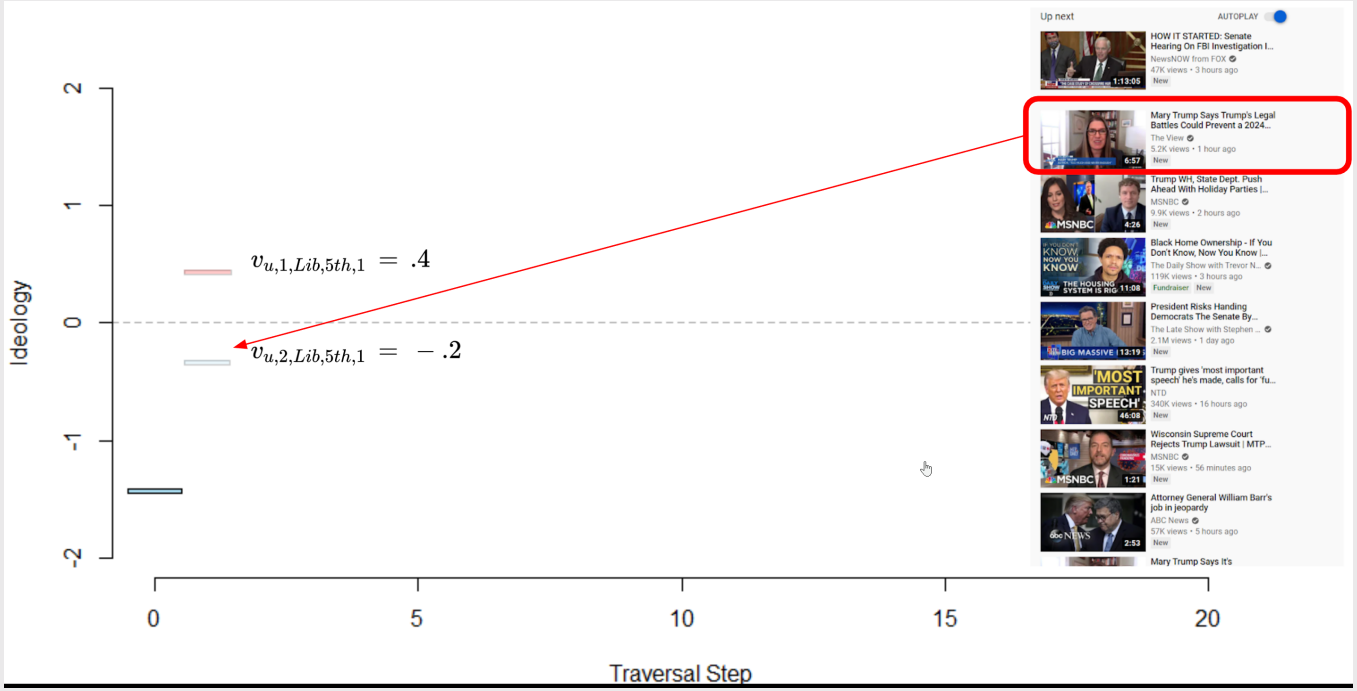
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated



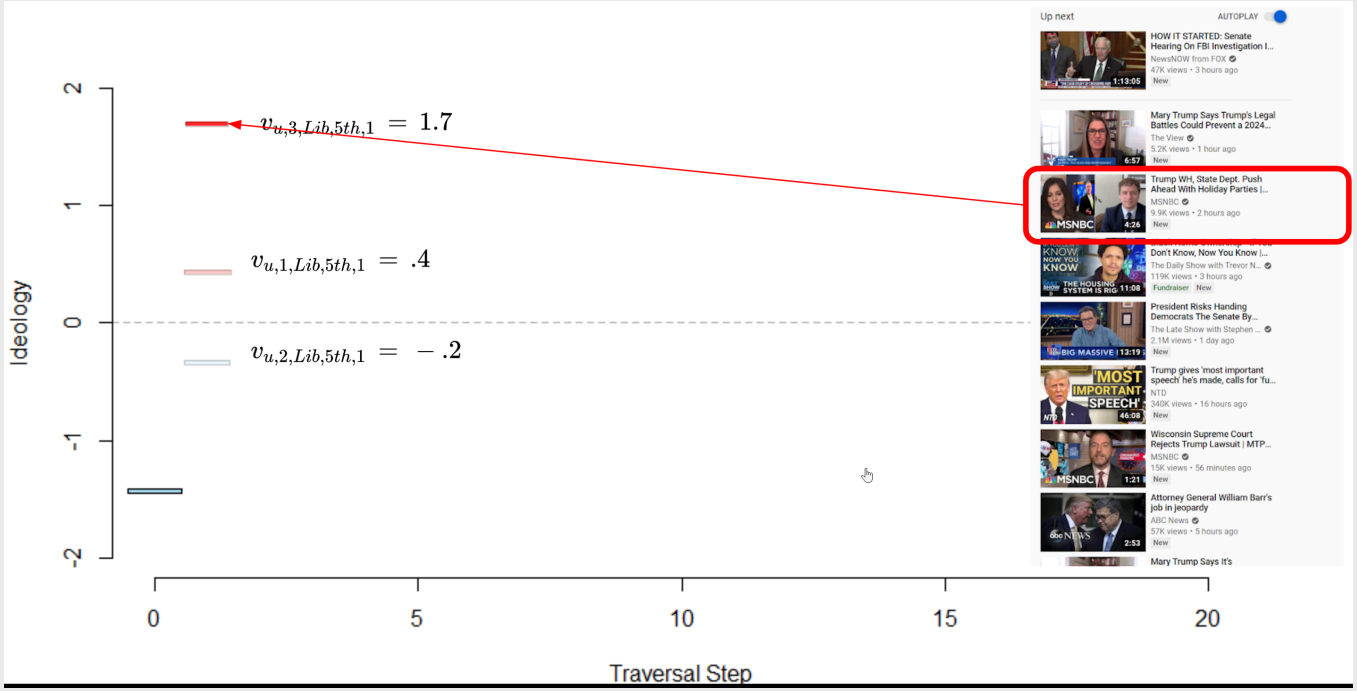
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



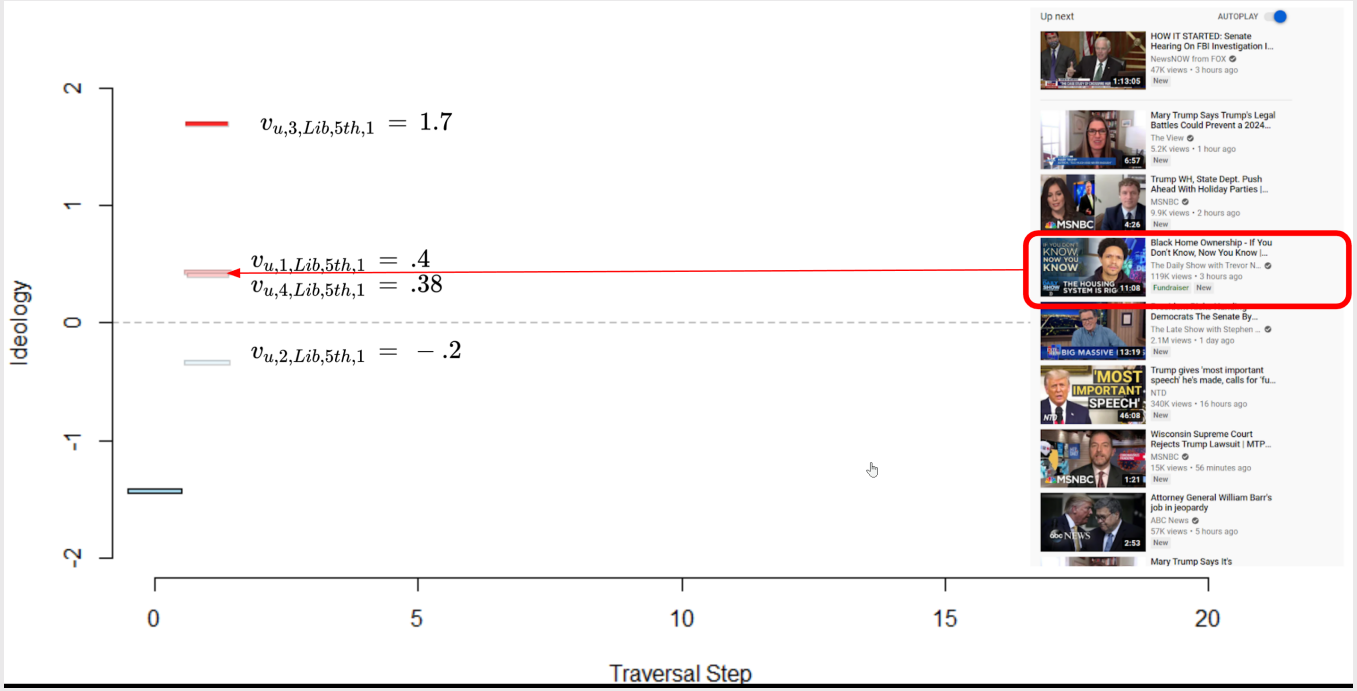
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



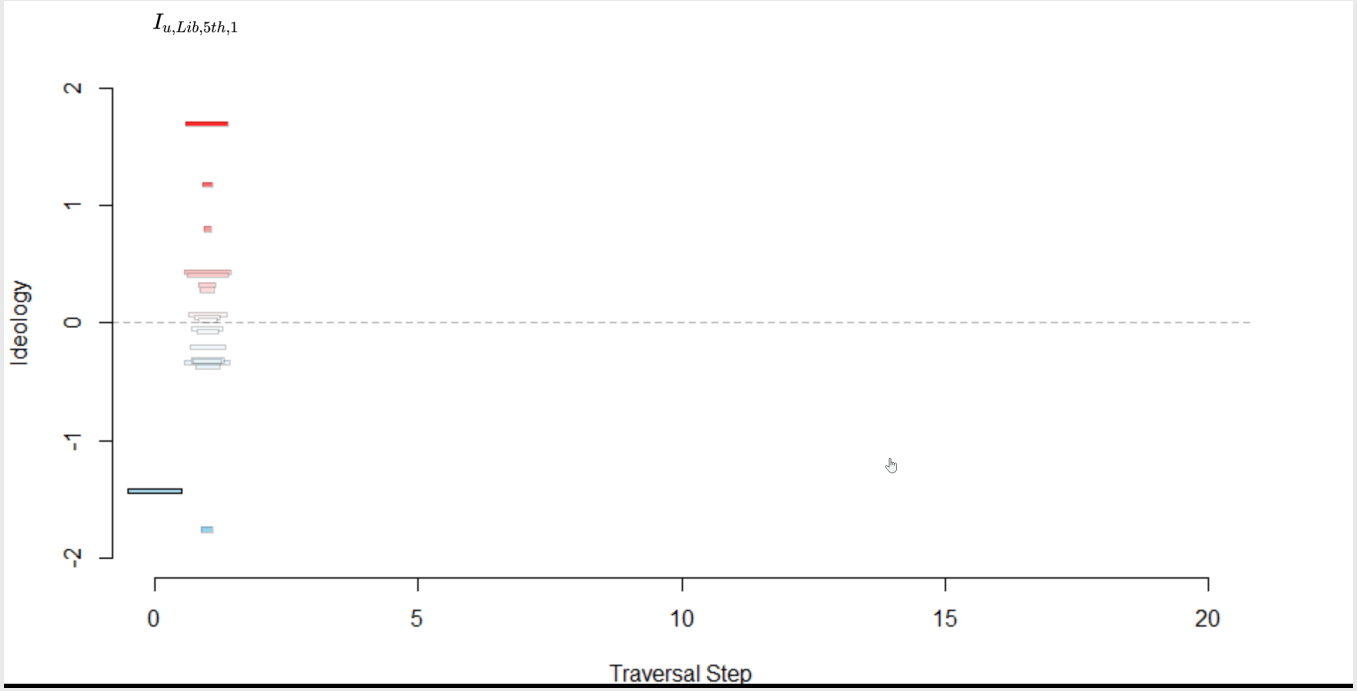
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



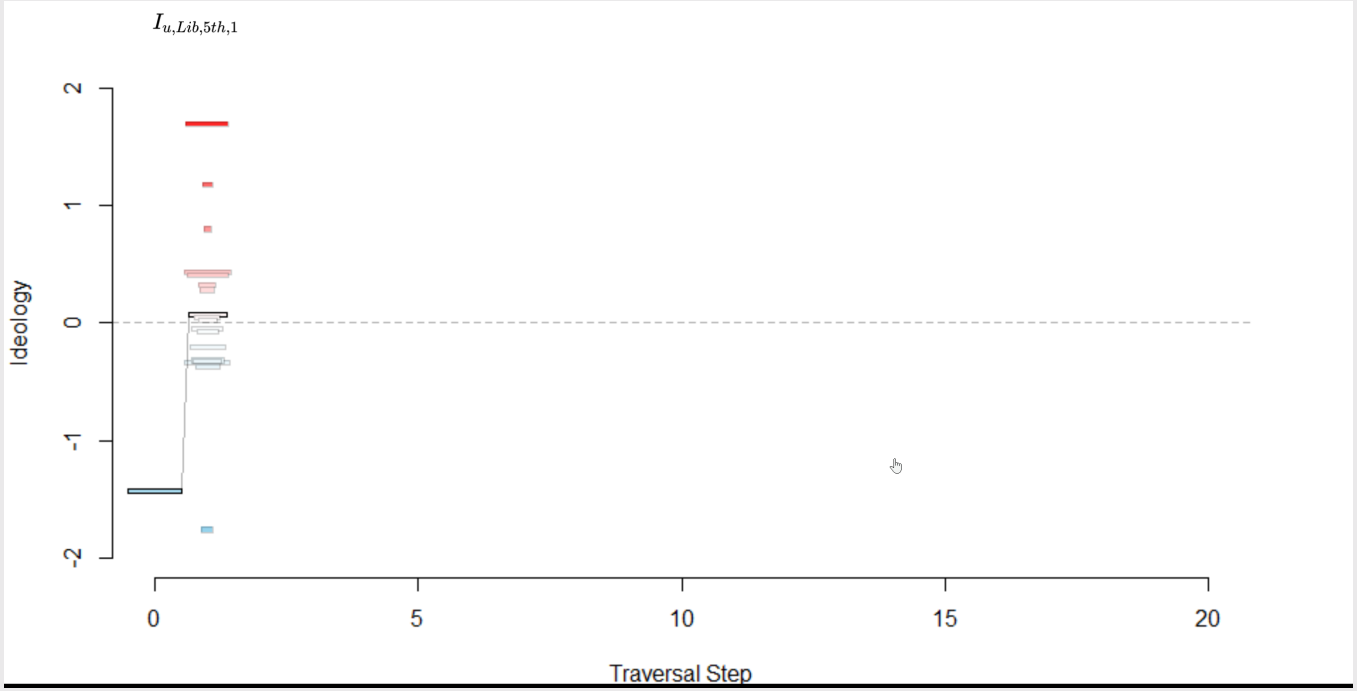
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



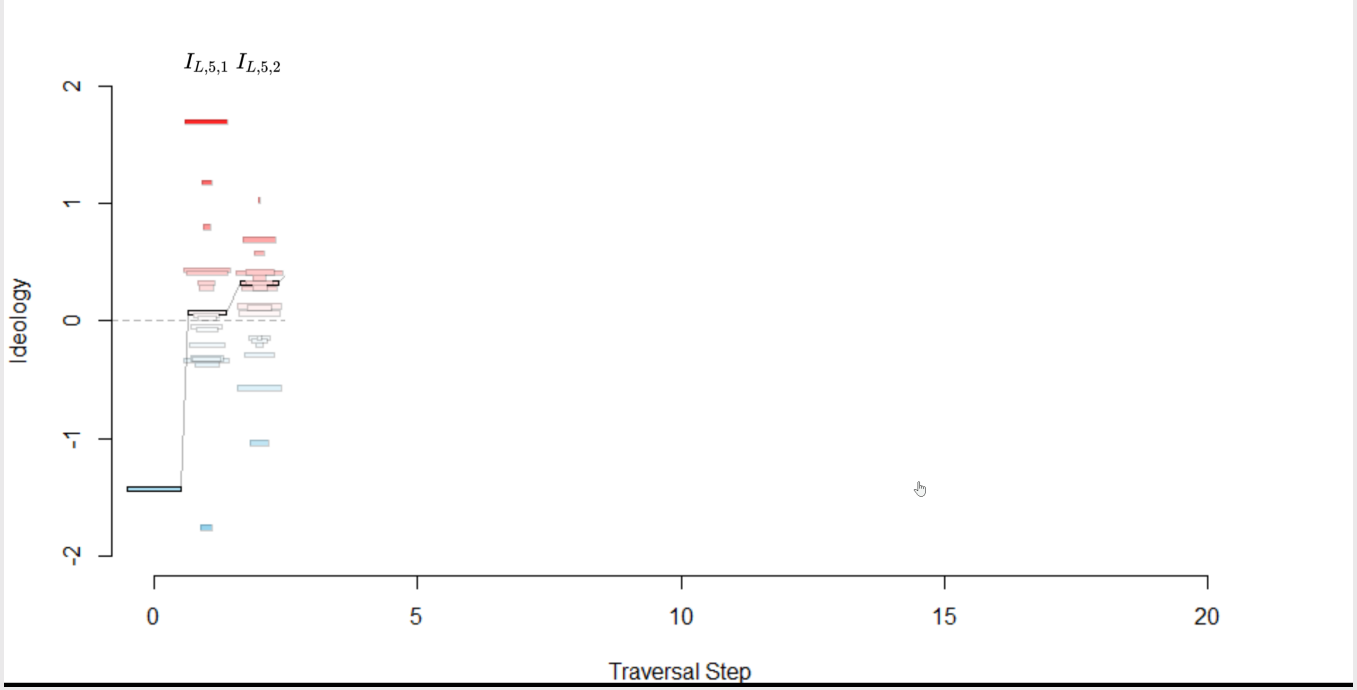
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected…
   - …and the hypotheses you have generated



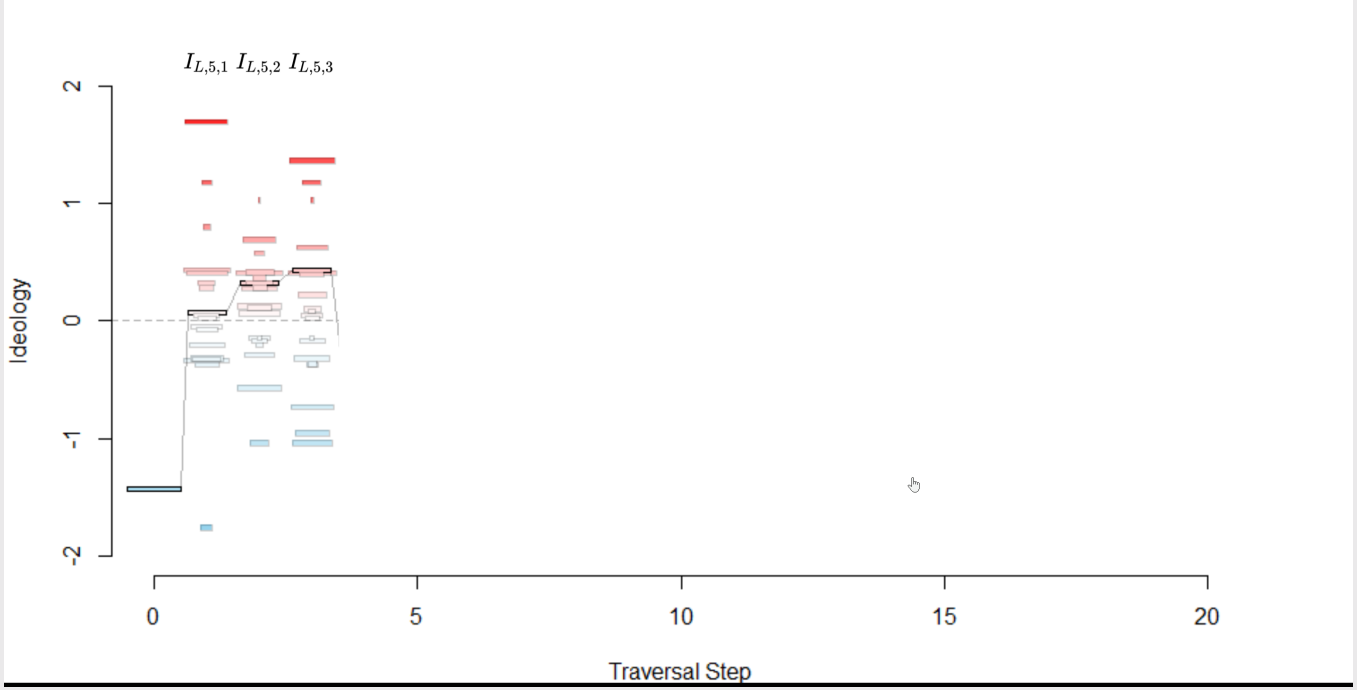
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated



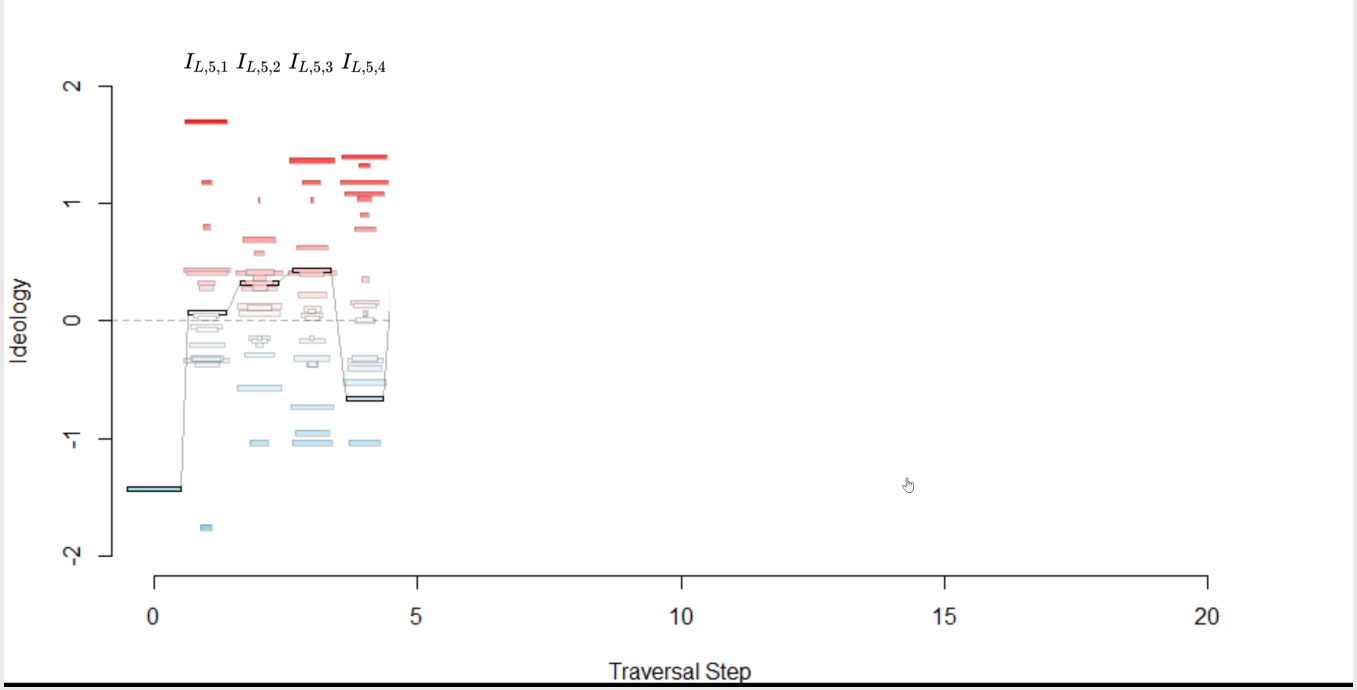
Research Camp

3. **Data Collection / Wrangling → Analysis**
    - Analysis is informed by the data you have collected...
    - ...and the hypotheses you have generated



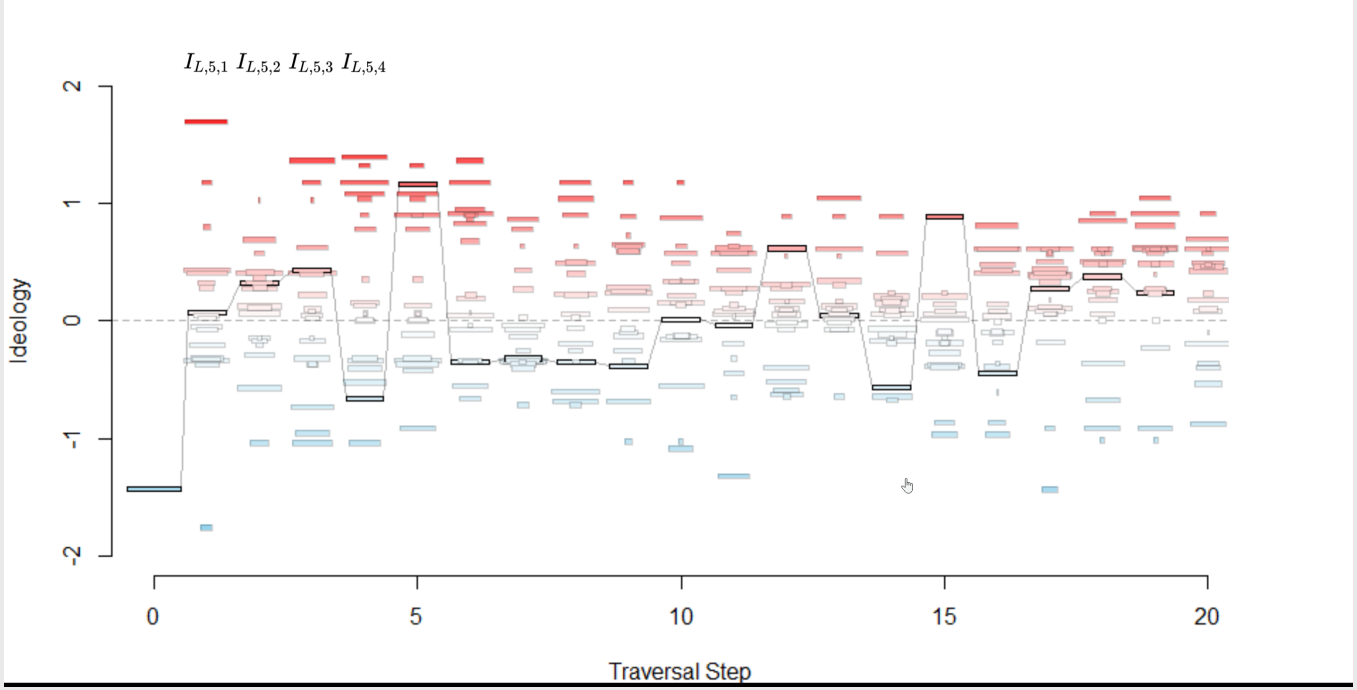
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated



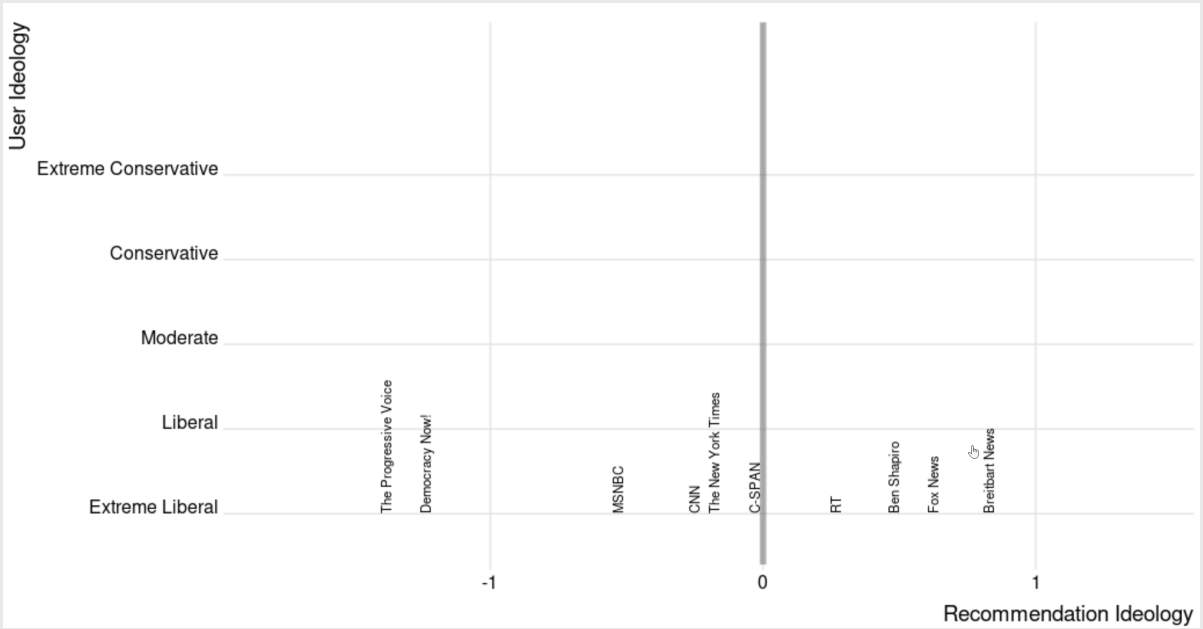
Research Camp

3. **Data Collection / Wrangling → Analysis**
   - Analysis is informed by the data you have collected...
   - ...and the hypotheses you have generated



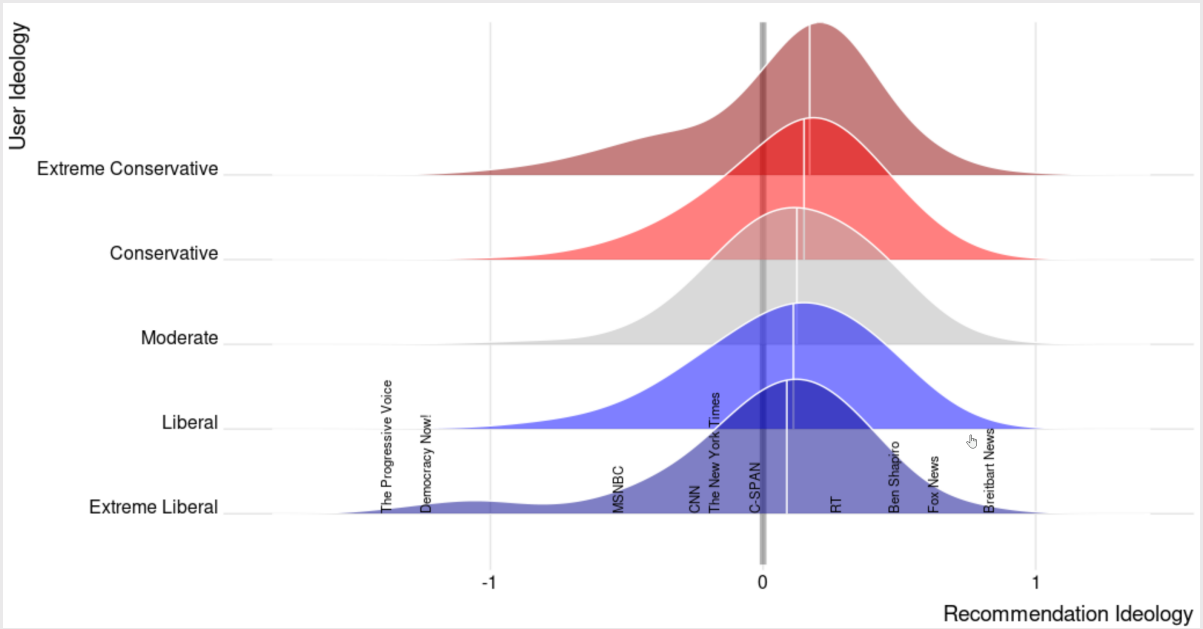
Research Camp

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

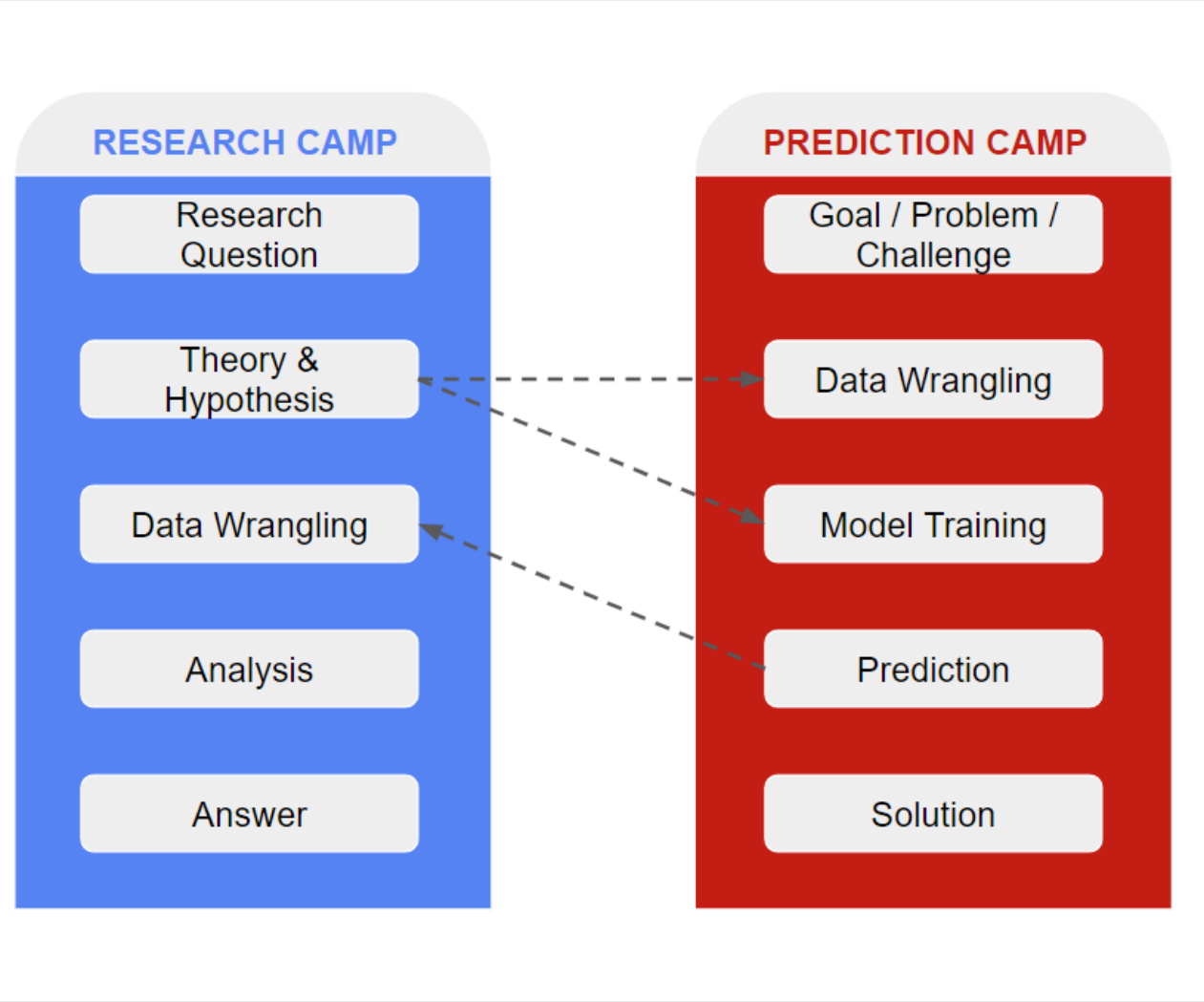


Research Camp

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

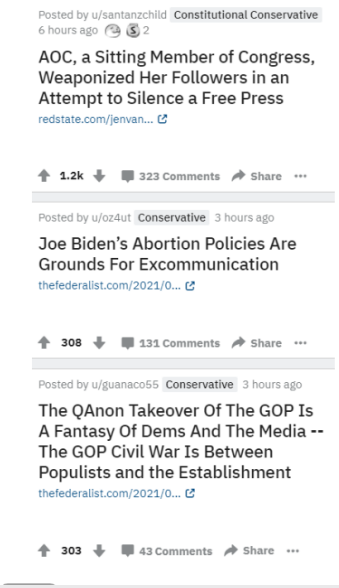
- **Goal/Problem/Challenge**: Measure the ideology of a YouTube video

Prediction Camp

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

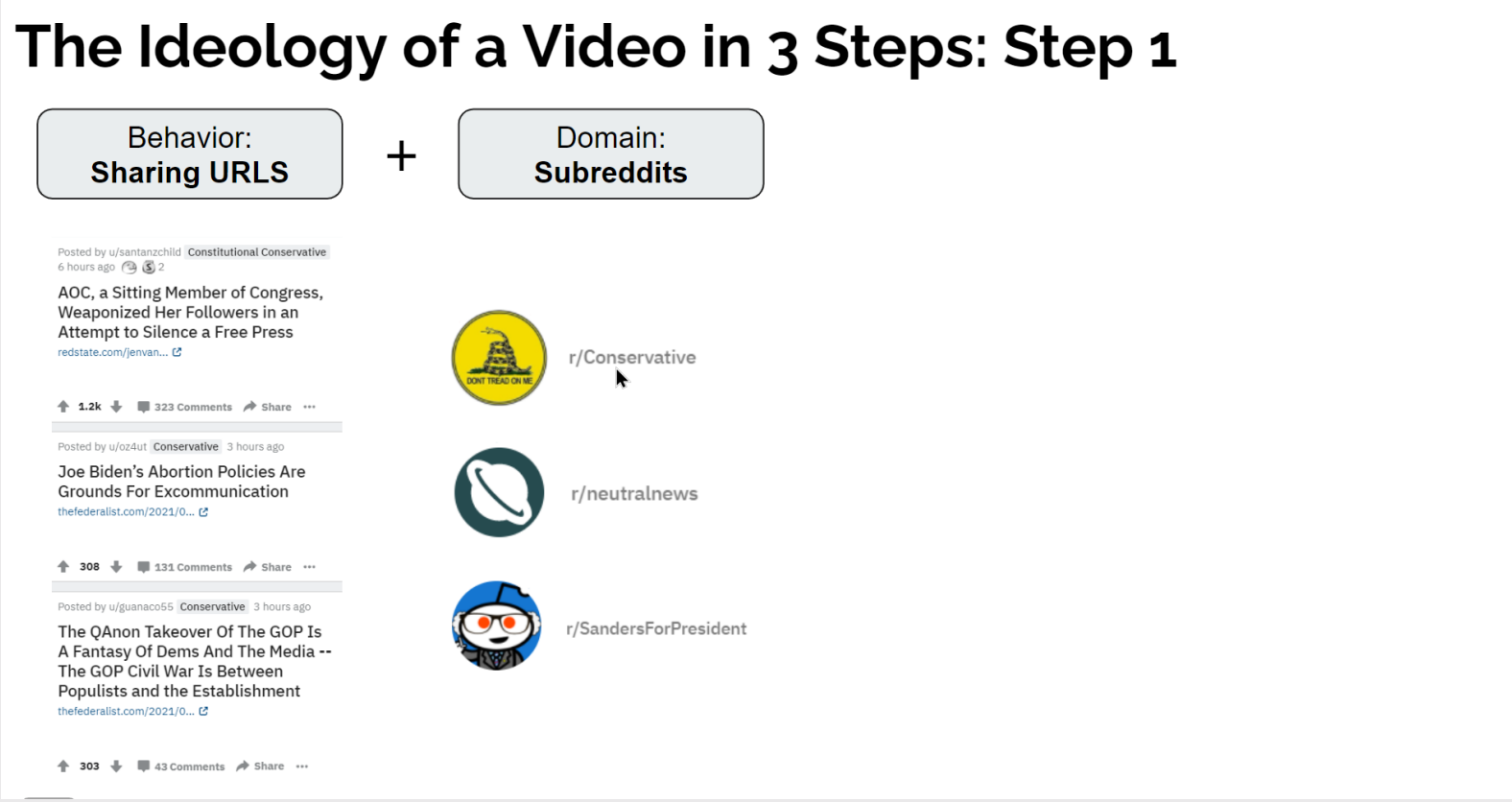
The Ideology of a Video in 3 Steps: Step 1

Behavior:
Sharing URLS



Prediction Camp

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

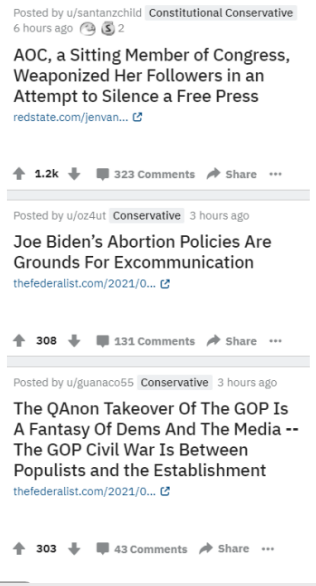


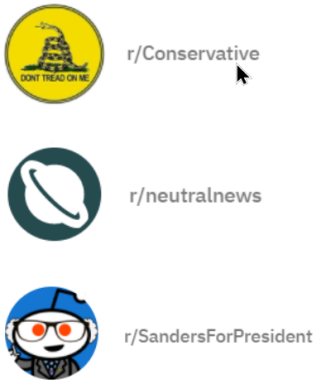
Prediction Camp

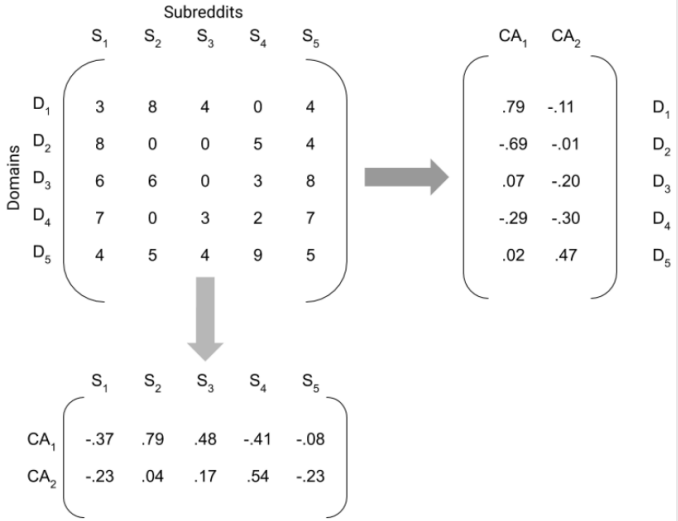
- **Data Wrangling**: Get matrix of links shared on political subreddits

The Ideology of a Video in 3 Steps: Step 1

Behavior: **Sharing URLS** + Domain: **Subreddits**

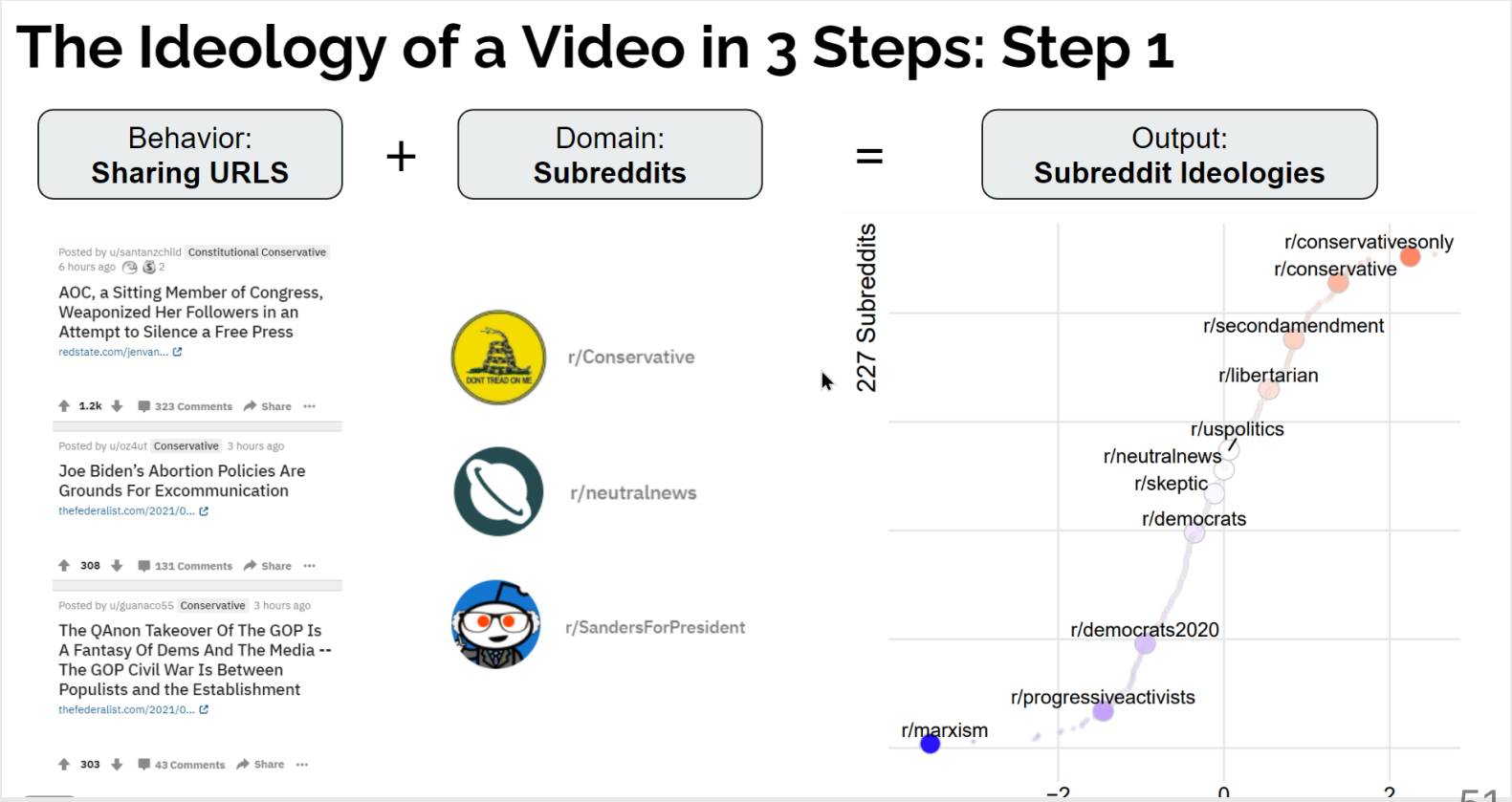






Prediction Camp

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



Prediction Camp

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

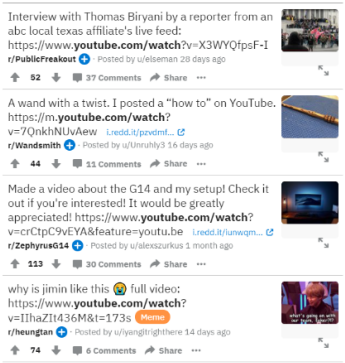
The Ideology of a Video in 3 Steps: Step 2

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

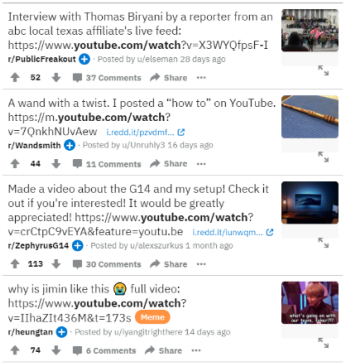


Prediction Camp

- **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**

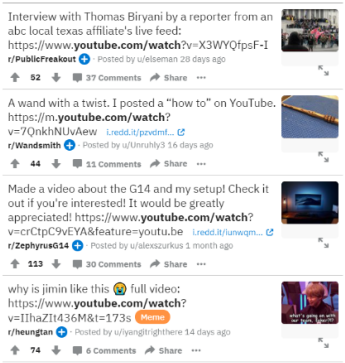


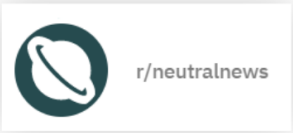
Prediction Camp

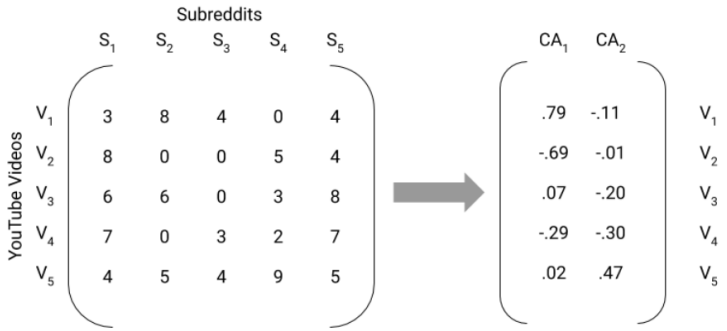
- **Data Wrangling**: Get matrix of 60k YouTube videos shared on scored subreddits

The Ideology of a Video in 3 Steps: Step 2

Behavior: **Sharing Videos** + Domain: **Ideological Reddit**





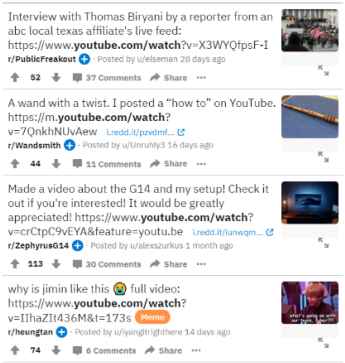


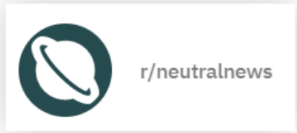
Prediction Camp

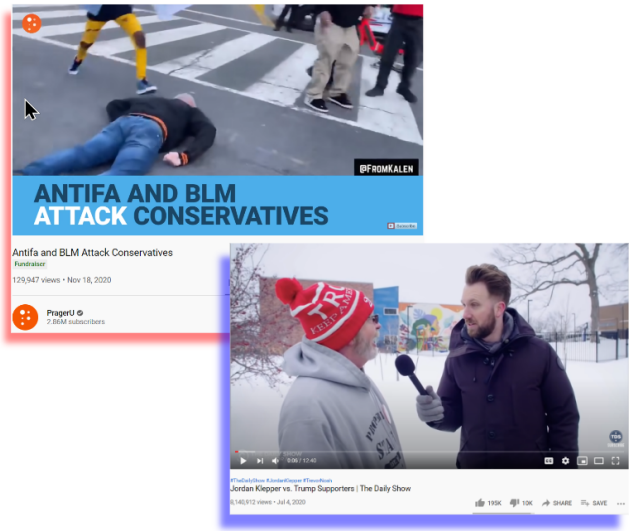
- **Data Wrangling**: Calculate video ideology as weighted mean of subreddits

The Ideology of a Video in 3 Steps: Step 2

Behavior: **Sharing Videos** + Domain: **Ideological Reddit** = Output: **Video Ideologies**







Prediction Camp

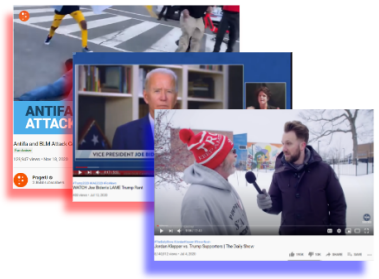
- **Model Training**: BERT transformer trained on 60k videos

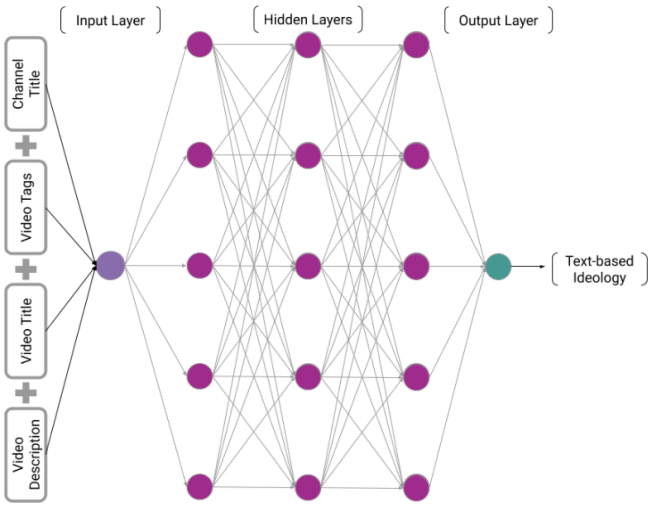
The Ideology of a Video in 3 Steps: Step 3

Training Data:
67k Coded Videos

+

Classifier:
BERT Transformer



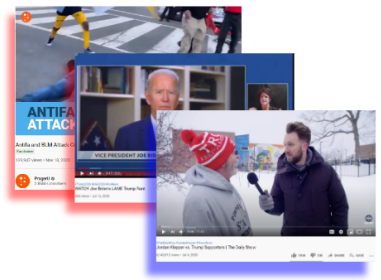


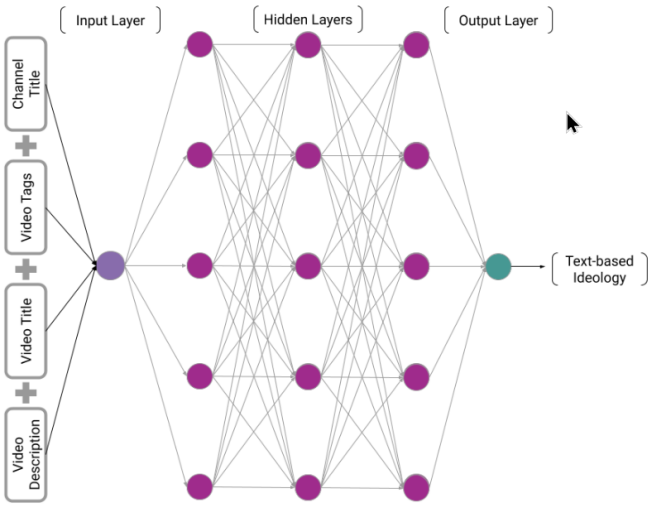
Prediction Camp

- **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
30-Aug-23	BREAK	Wed			
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
13-Sep-23	Review of R	Wed			
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
4-Oct-23	Multivariate Review	Wed			
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			
25-Oct-23	**Midterm Exam**	**Wed**			
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
8-Nov-23	Classification Review	Wed			
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
20-Nov-23	BREAK	Mon			
22-Nov-23	BREAK	Wed			
27-Nov-23	Clustering & NLP Part 3	Mon	Sentiment analysis	Pset 9 assigned	Quiz 19
29-Nov-23	Clustering & NLP Review	Wed			
4-Dec-23	Ethics	Mon	The risks of rapid technological change		Quiz 20
6-Dec-23	Final Review	Wed			
11-Dec-23	**Final Exam**	**Mon**			

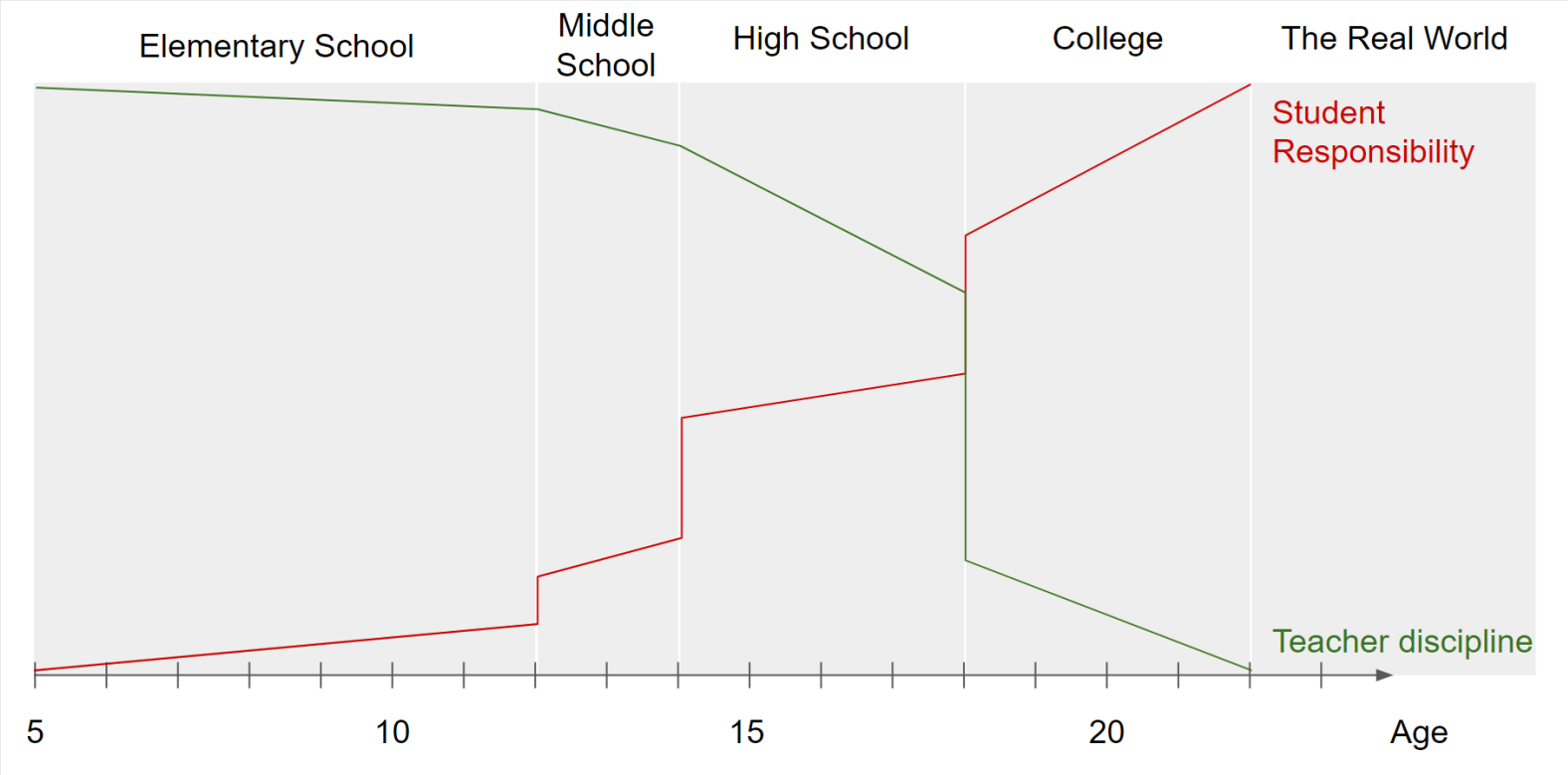
Honor Code

- 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**)

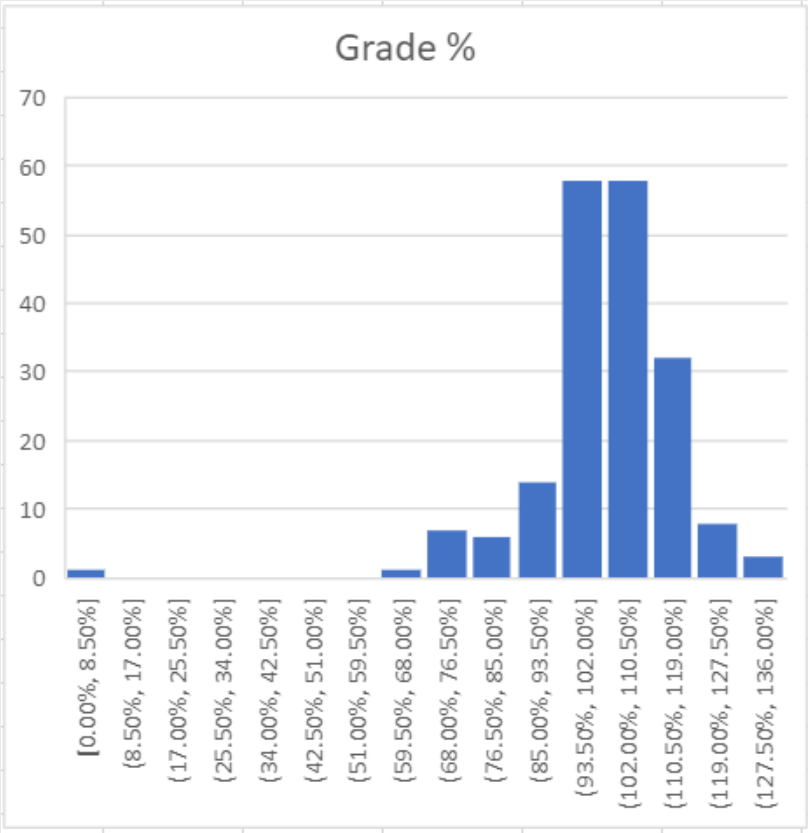
Teaching Philosophy

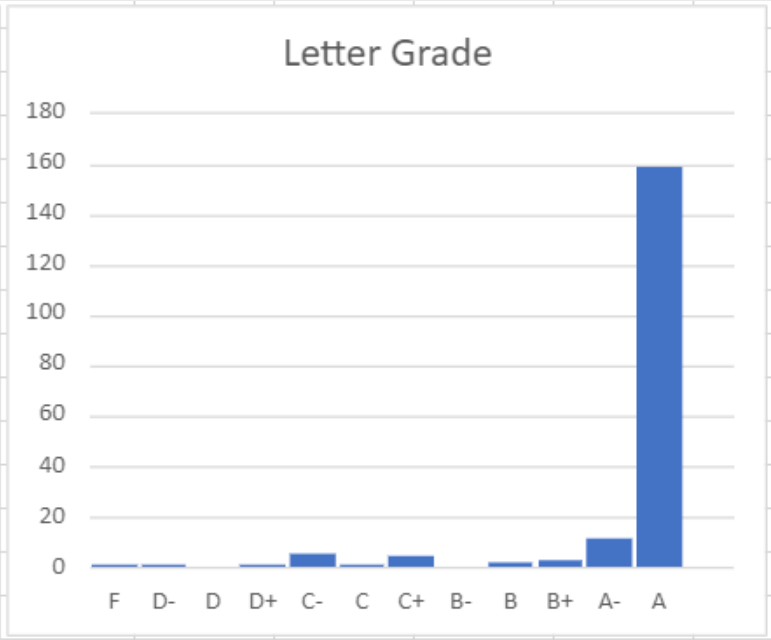


Teaching Philosophy

- 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)