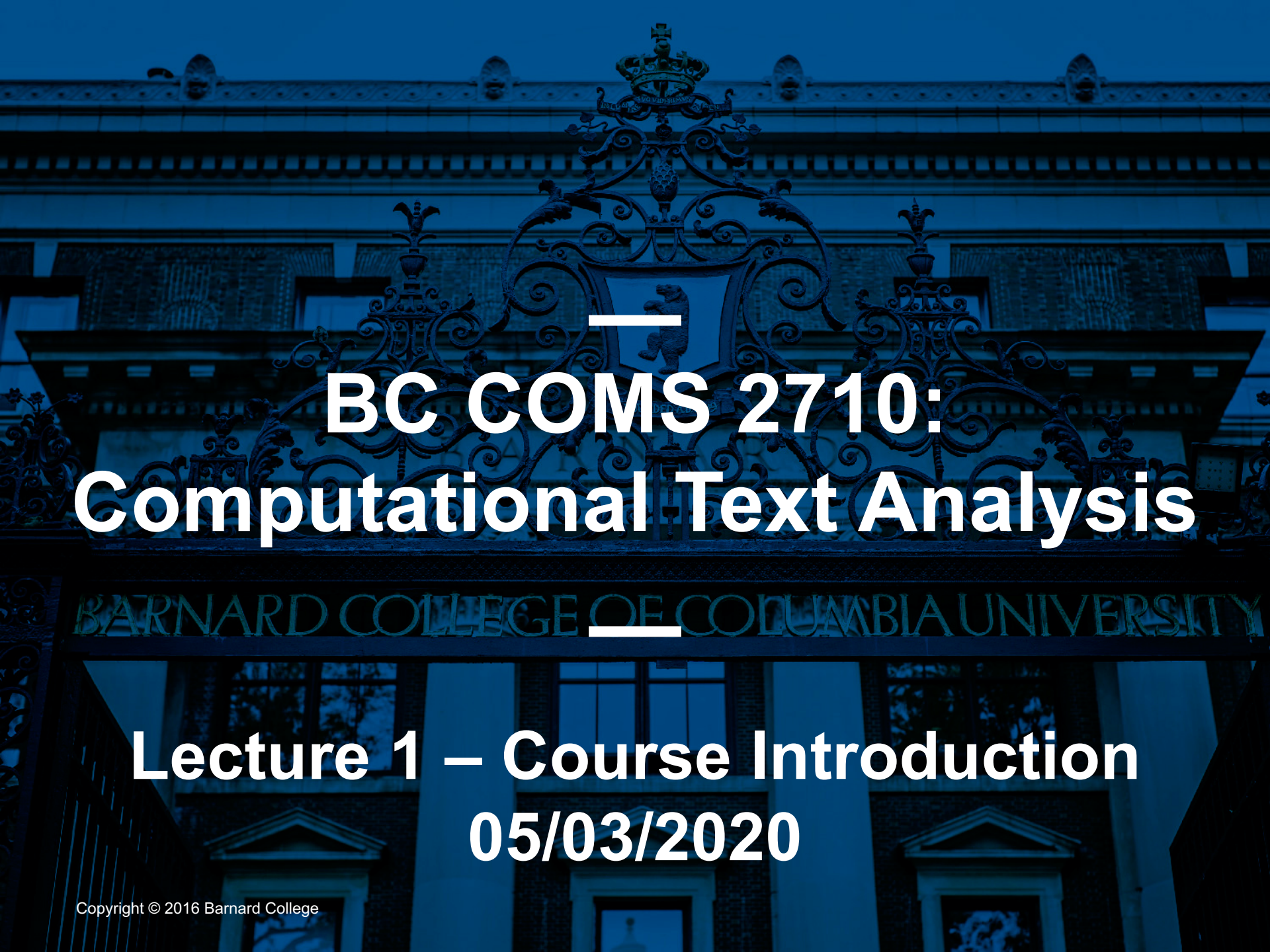


BC COMS 2710: Computational Text Analysis

Lecture 1 – Course Introduction
05/03/2020

What is Computational Text Analysis?

Computational Text Analysis for Social Science: Model Assumptions and Complexity

Brendan O'Connor* David Bamman† Noah A. Smith†*
*Machine Learning Department

Commentary

Adapting computational text analysis to social science (and vice versa)

Paul DiMaggio

Abstract
Social scientists and computer scientist are divided by small differences in perspective and disciplinary divide. In the field of text analysis, several such differences are noted: social scientists models to explore corpora, whereas many computer scientists employ supervised models to tra hold to more conventional causal notions than do most computer scientists, and often favor existing algorithms, whereas computer scientists focus more on developing new models; and com trust human judgment more than social scientists do. These differences have implications that pot practice of social science.

Keywords
Topic models, text analysis, unsupervised models, interpretation, sentiment analysis, supervised

ECR Forum

Computational text analysis: Thoughts on the contingencies of an evolving method

Daniel Marciniak

Abstract
Mapping a public discourse with the tools of computational text analysis comes with many contingenci

Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts

Justin Grimmer
Department of Political Science, Stanford University, Encina Hall West 616 Serra Street, Stanford, CA 94305
e-mail: jgrimmer@stanford.edu (corresponding author)

Brandon M. Stewart
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: bstewart@fas.harvard.edu

Edited by R. Michael Alvarez

Politics and political conflict often occur in the written and spoken word. Scholars have long recognized this, but the massive costs of analyzing even moderately sized collections of texts have hindered their use in political science research. Here lies the promise of automated text analysis: it substantially reduces the costs of analyzing large collections of text. We provide a guide to this exciting new area of research and show how, in many instances, the methods have already obtained part of their promise. But there are pitfalls to using automated methods—they are no substitute for careful thought and close reading and require extensive and problem-specific validation. We survey a wide range of new methods, provide guidance on how to validate the output of the models, and clarify misconceptions and errors in the literature. To conclude, we argue that for automated text methods to become a standard tool for political

What is Data Science?

- *"Data science is the study of extracting value from data" – Jeannette Wing*

What is Data Science?

- *"Data science is the study of extracting **value** from data" –*
  Jeannette Wing

- Value
  - Requires domain expertise to determine what value is
  - *Value from data* is different based on the domain and the needs

What is Data Science?

- *"Data science is the study of **extracting** value from data" –*
  Jeannette Wing

- Extracting
  - emphasizes action on data
  - mining information

What is Computational Text Analysis?

Computational Text Analysis

- "~~Data science~~ is the ~~study~~ practice of extracting value from data" – large ^scale textual

~~Jeannette Wing~~

Adam Poliak

Computational Text Analysis

- *Computational text analysis is not a replacement for but rather an addition to the approaches one can take to analyze social and cultural phenomena using textual data. By moving back and forth between large-scale computational analyses and small-scale qualitative analyses, we can combine their strengths so that we can identify large-scale and long-term trends, but also tell individual stories*

http://coms2710.barnard.edu/readings/Nguyen-et-al-how-we-do-things-with-words.pdf

Computational Text Analysis

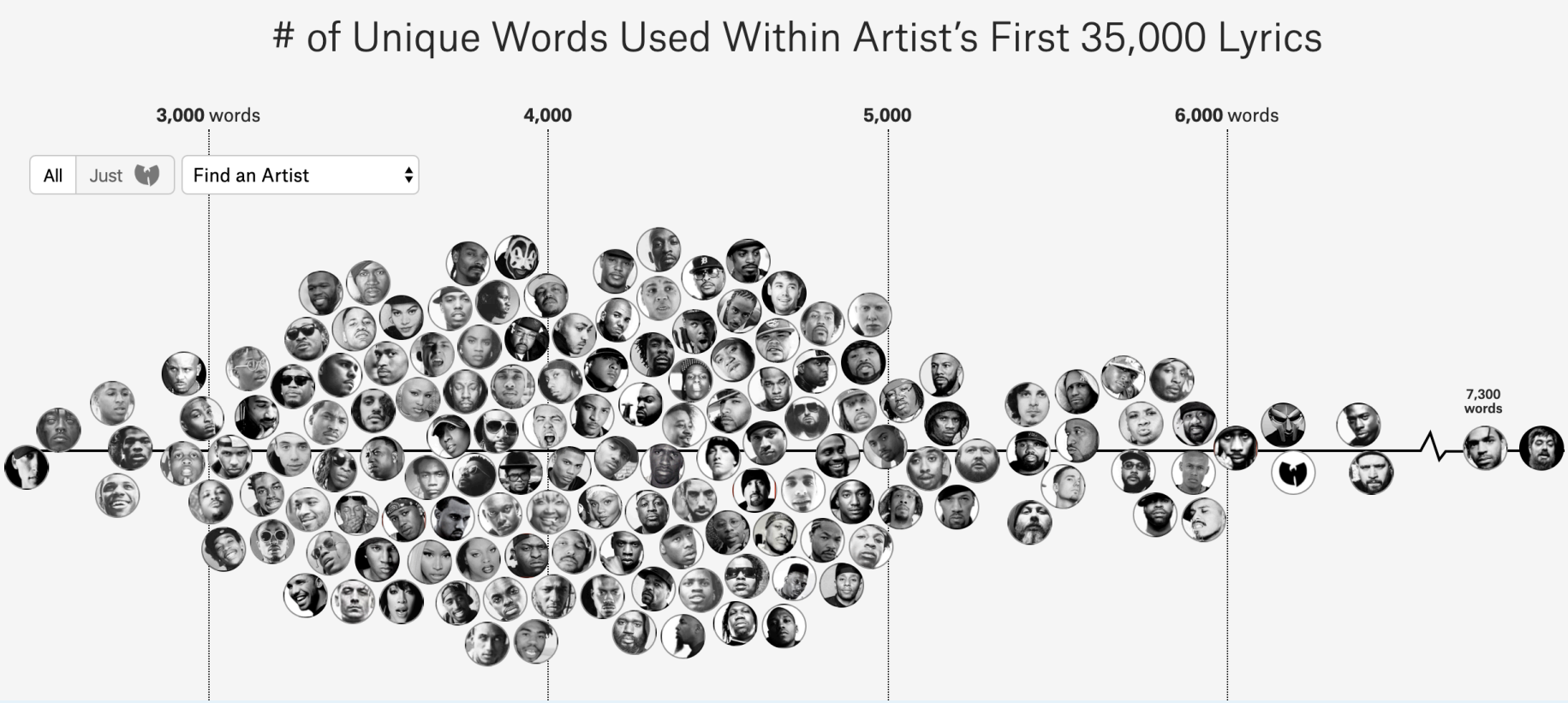
- *Computational text analysis is not a replacement for but rather an addition to the approaches one can take to analyze social and cultural phenomena using textual data. By moving back and forth between **large-scale computational analyses** and small-scale qualitative analyses, we can combine their strengths so that we can identify large-scale and long-term trends, but also tell individual stories*

http://coms2710.barnard.edu/readings/Nguyen-et-al-how-we-do-things-with-words.pdf

What can we do with computational text analysis?

What can we do with large scale textual analysis?

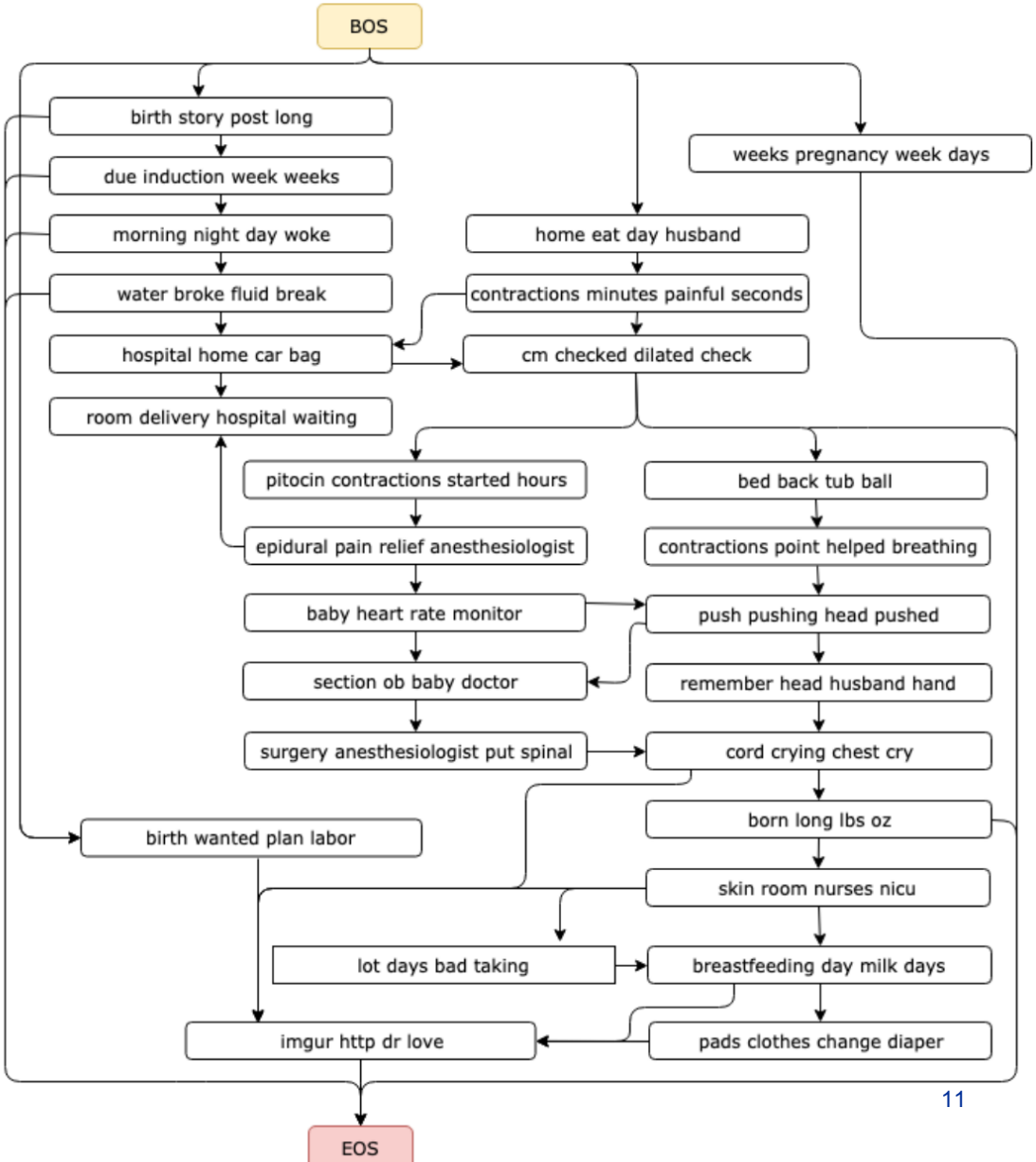
- Sort artists by their vocabulary



https://pudding.cool/projects/vocabulary/index.html

What can we do with large scale textual analysis?

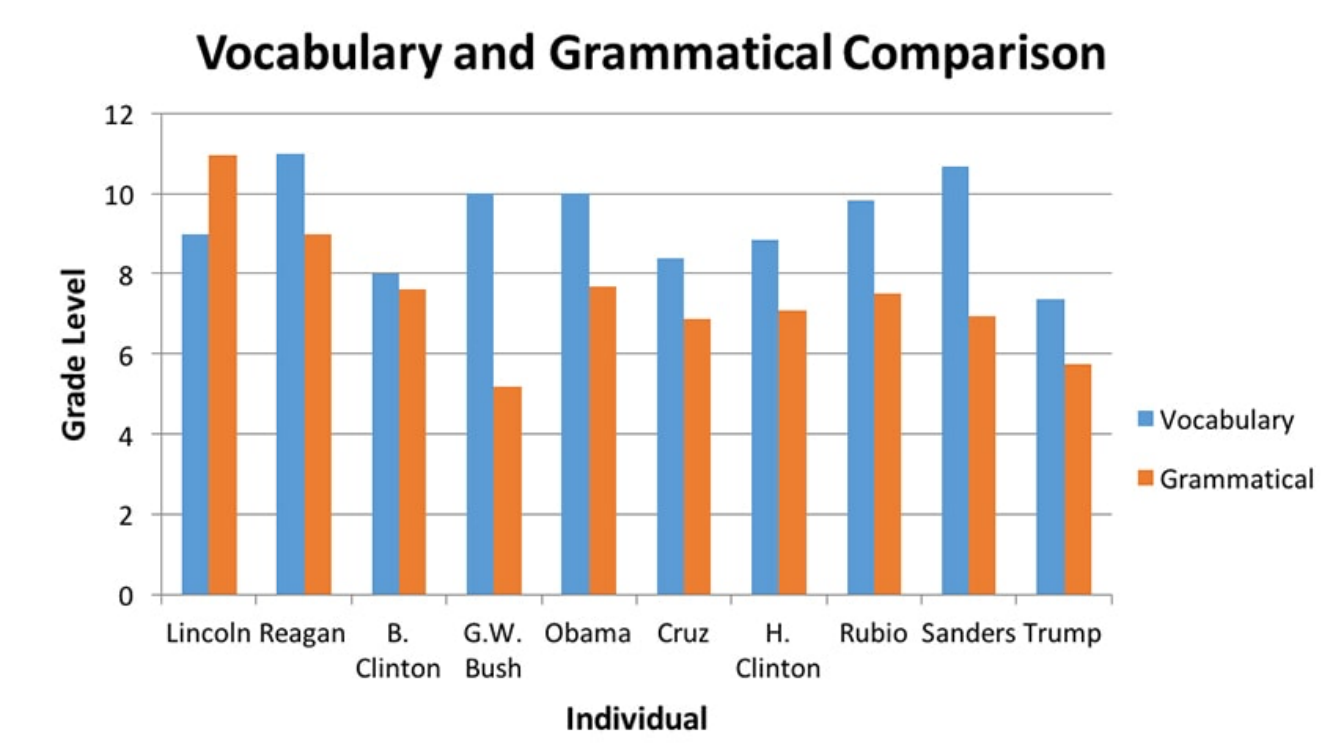
- Identify flow of topics in birthing narratives



https://maria-antoniak.github.io/resources/2019_cscw_birth_stories.pdf

What can we do with large scale textual analysis?

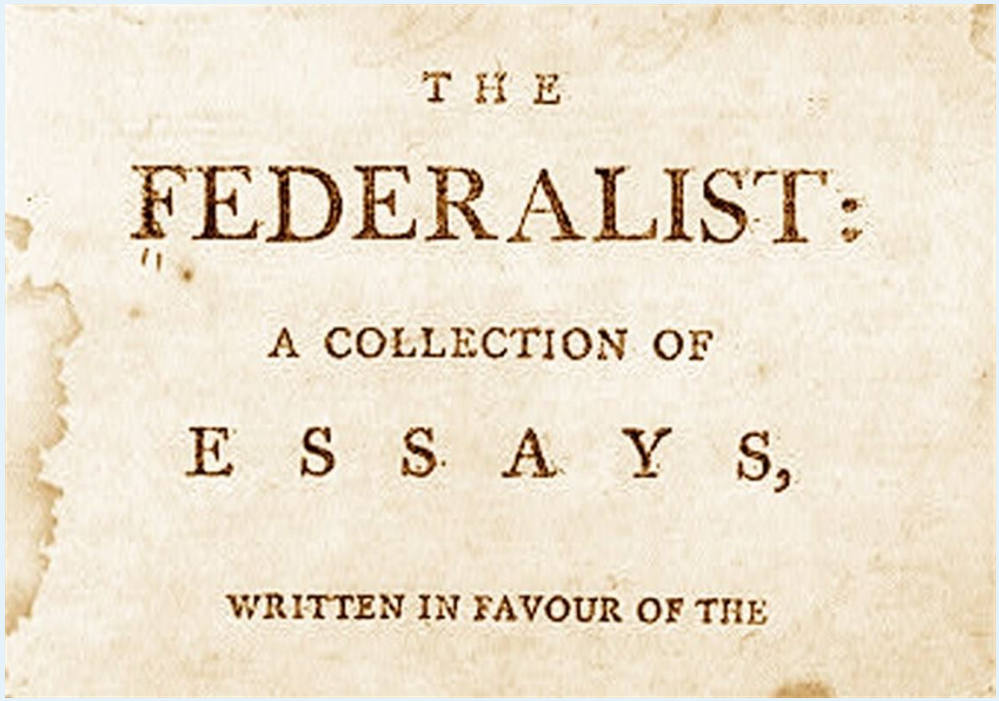
- Categorize the level of presidential candidates' speeches



https://arxiv.org/pdf/1603.05739.pdf

What can we do with large scale textual analysis?

- Who wrote the anonymous Federalist Papers?


THE
FEDERALIST:
A COLLECTION OF
ESSAYS,
WRITTEN IN FAVOUR OF THE

https://www.jstor.org/stable/2283270

What can we do with large scale textual analysis?

Computational Text Analysis in this course

- Aggregate large scale textual data
- Text Processing
- Discovering patterns in data

Course Objectives

Learn the tools and gain the confidence to independently:

1. Aggregate large scale textual data
2. Text processing
3. Discovering patterns in data

Course Outline

- Python Overview — **Week 1**
  - Introduction to Python
  - Pandas
- Lexical based analysis methods — **Week 2 - 3**
  - Text Processing
  - Word & Document Representation
  - Topic Modeling
- Data Collection — **Week 4**
  - Web Scraping
  - APIs

Course Outline

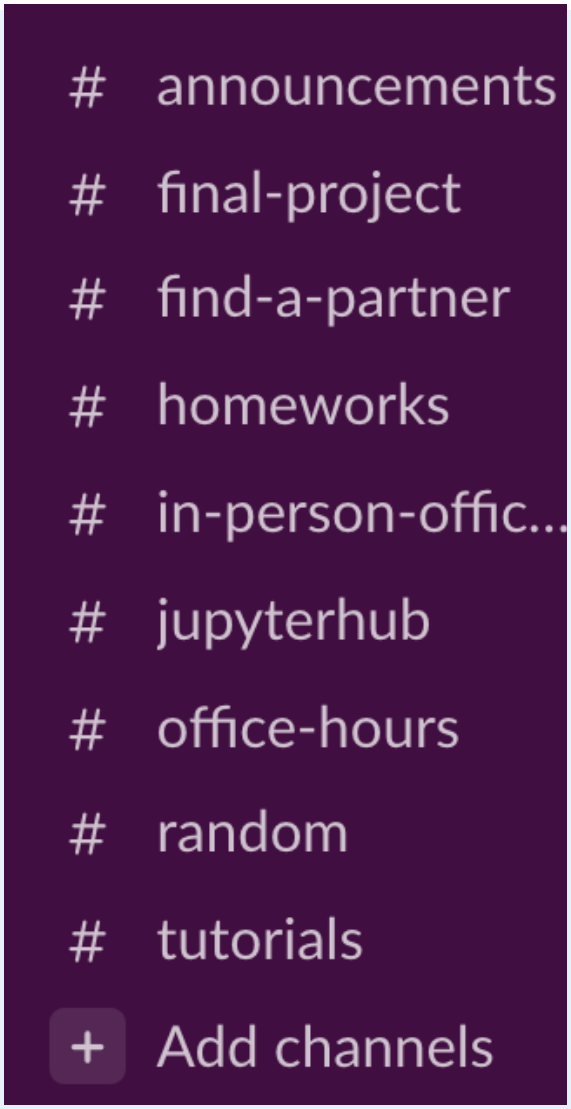
- Machine Learning **Week 5**
  - Regression & Classification
  - Clustering
- Advanced Topics & Final Projects **Week 6**

Logistics

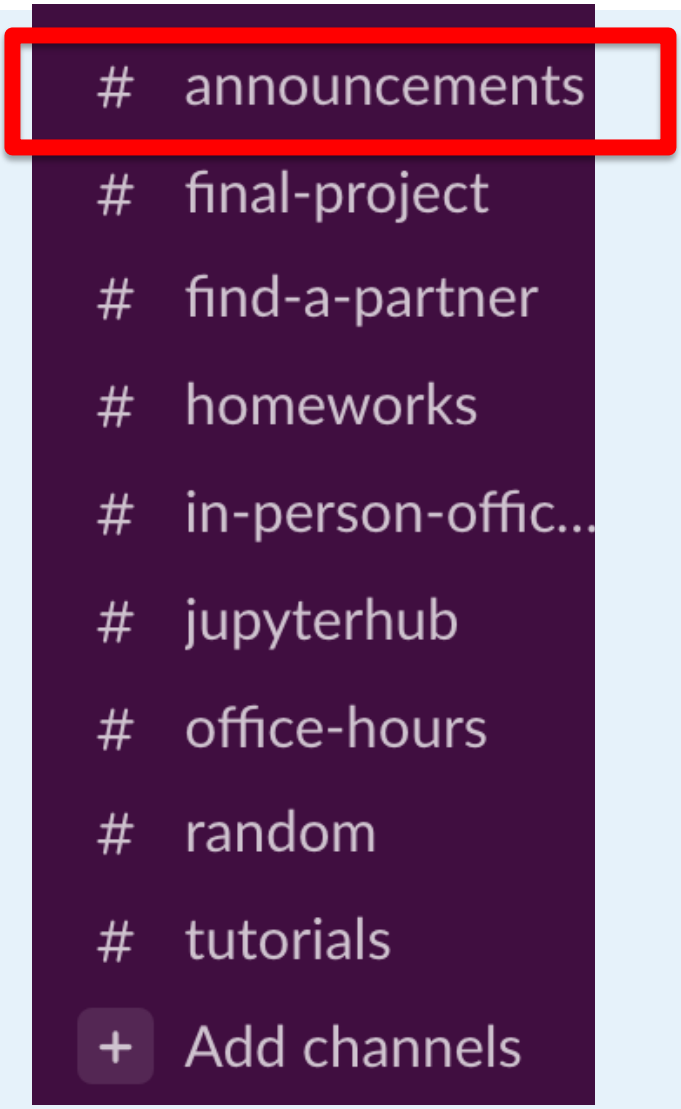
Communication

- Course webpage:
  - https://coms2710.barnard.edu/
- Slack:
  - https://bc-coms-2710-summera.slack.com/
- Zoom link:
  - Same for lectures and office hours
- Gradescope:
  - Submitting assignments

Slack

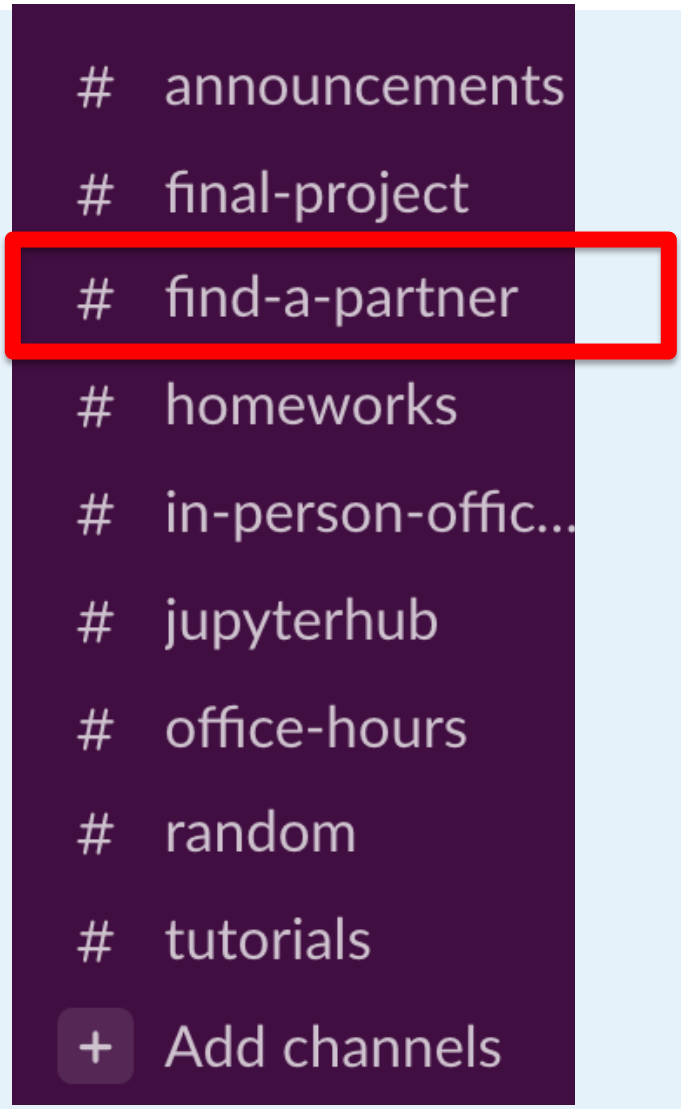


Slack - Announcements



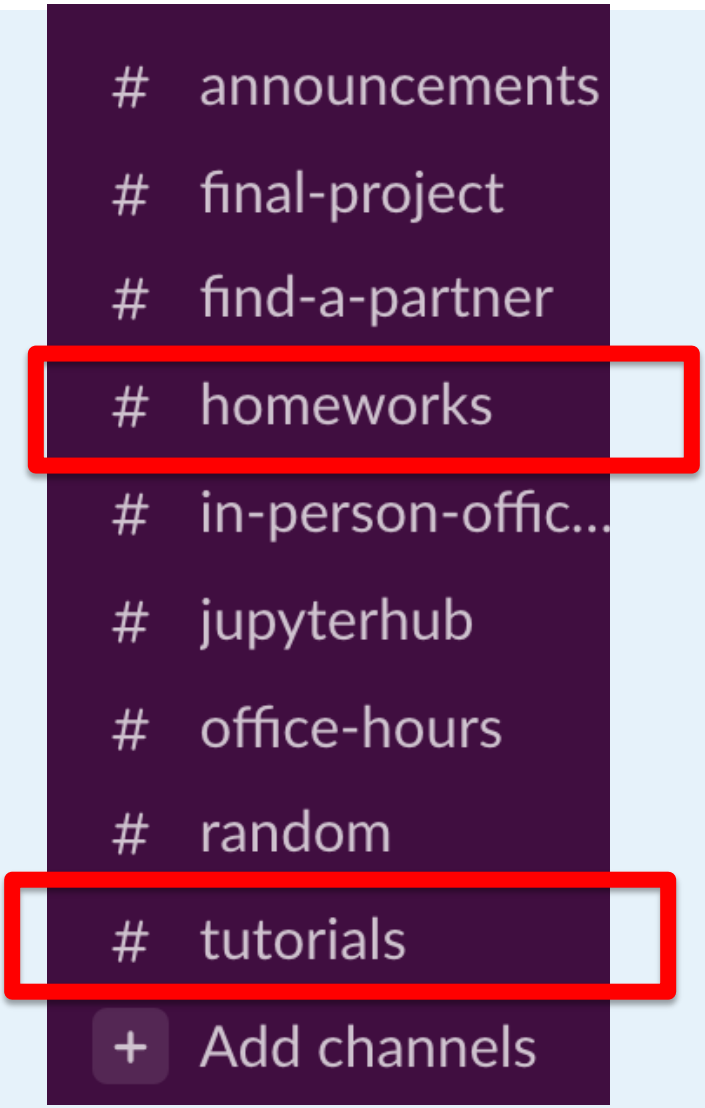
- course staff post course wide announcements
- Do not post here
- Encouraged to reply to posts that we create there

Slack – Find-a-Partner



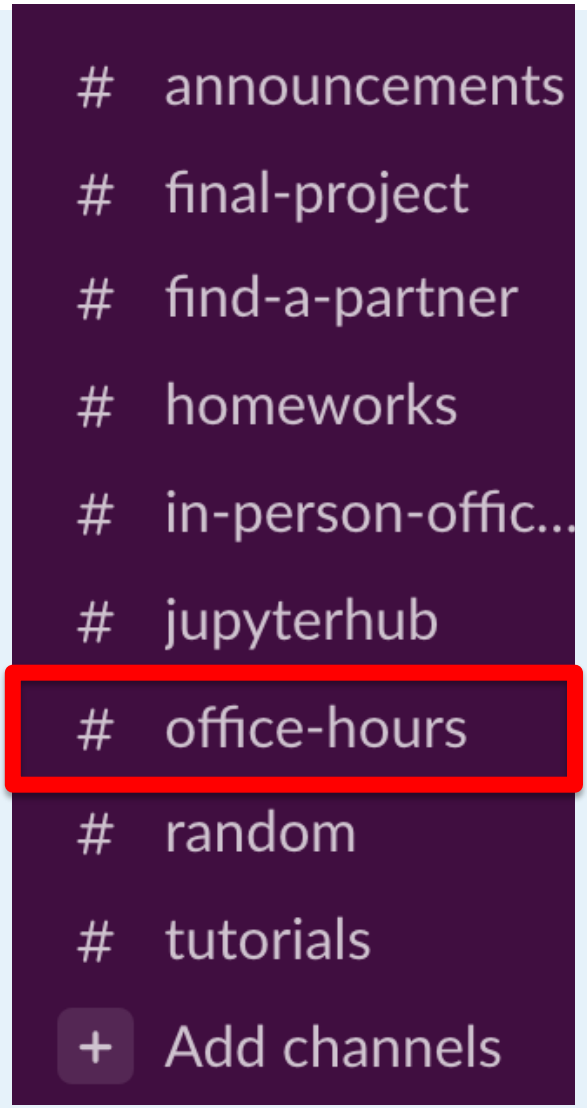
- Use this channel to find partners
- Different parts of course can be completed in pairs

Slack – Homeworks/Tutorials



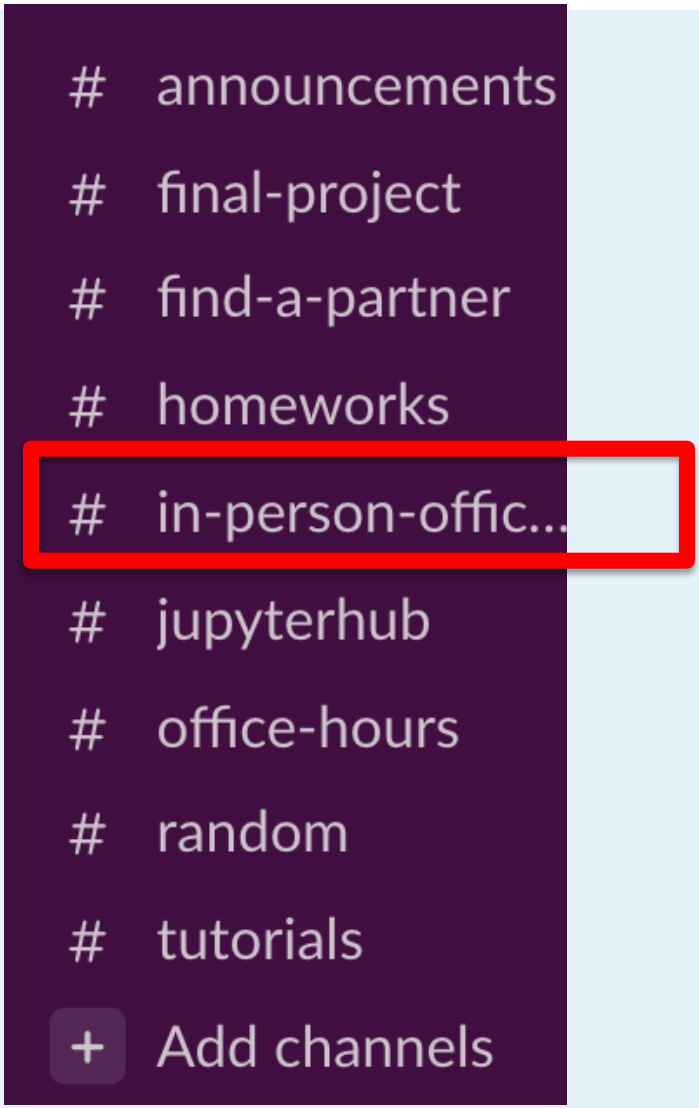
- Ask questions when working on homework, labs, and projects
- **Do not** post solutions

Slack – Office-Hours



- Changes to Office Hours will be posted here
- Ask questions about Office Hours posted here
- Fill out poll for times

Slack – In-person-office-hours



- Potential in-person office hours

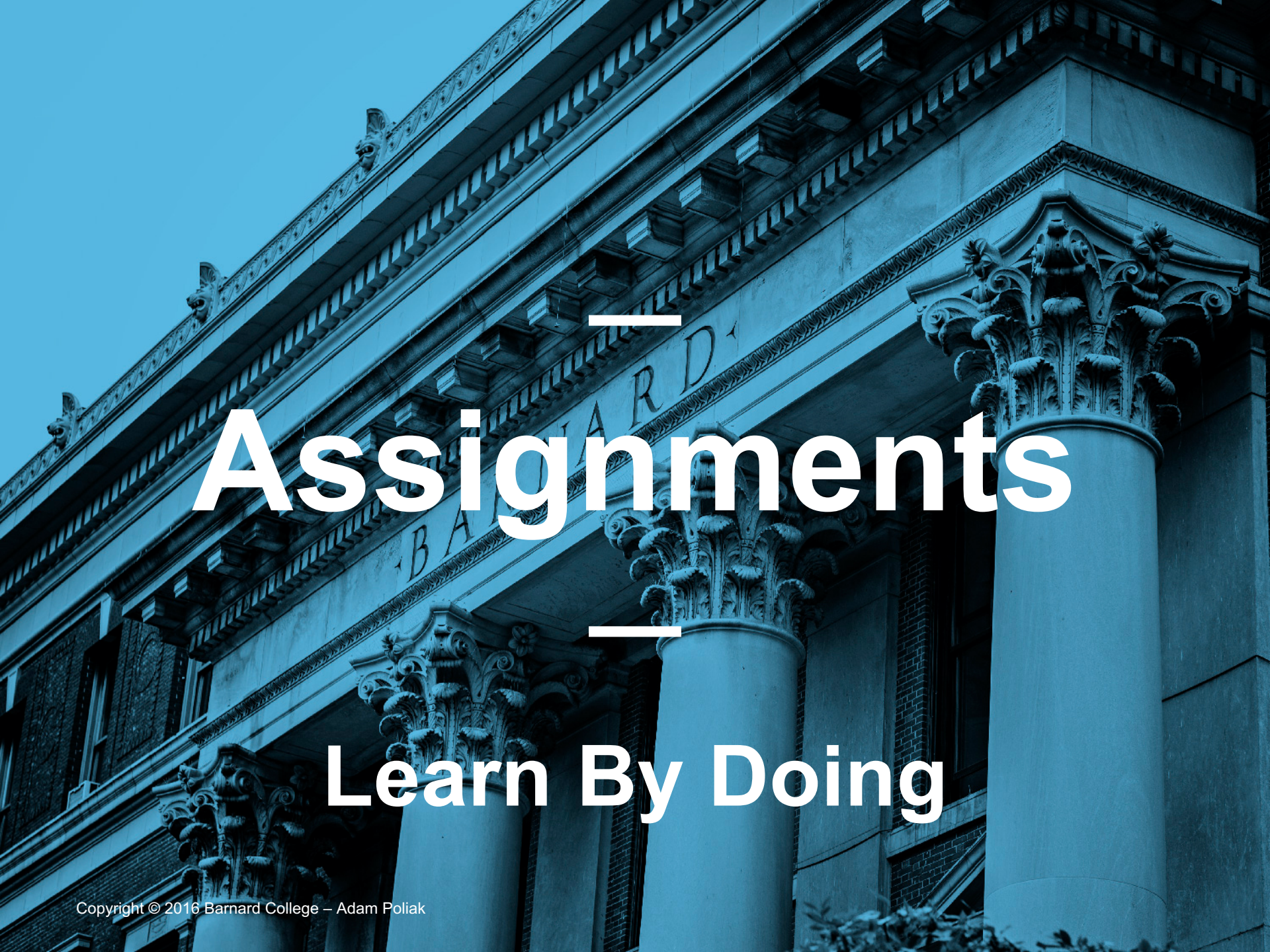
Course Meetings: 10:45am – 12:25pm (EST) MTWR

- Live classes
  - Primarily lectures
  - Q/A
  - Recorded
  - Discussions and exercises about course material
- Readings:
  - Readings associated with the lecture's material
  - Distributed on course schedule

Special dates

- No lectures: May 17th, 18th, May 31st

- Guest Speakers:
  - Maria Antoniak:
    - PhD student @ Cornell – June 1st
  - Lucy Li
    - PhD student @ Berkeley – June 9th



Assignments

Learn By Doing

Assignments

- Daily-ish exercises/tutorials
- Reading reflections
- 4 ~week long homeworks
- Final Project

Daily-ish Tutorials

- Due ~~M/T/W/R midnight~~ Saturday night @ midnight
- Complete individually
- ~1.5 hours long
- 2 or 3 a week

Reading reflections

- Due Sunday midnight
- For each reading:
  - 3-4 sentence summary
  - 1 sentence about something in particular that you like
  - 1 sentence about something you didn't like or something you found confusing and you'd like me to explain
  - 1 question for future work
- Goal: Examples of computational text analysis
  - Preparation for final projects
- Complete individualy

4 Homeworks

- Based on the previous week's material
- JupyterNotebook containing a mix of programming and written analysis
- Goal: gain comfort and confidence in textual analysis
- Can work in pairs

4 Homeworks

- Readability of Inaugural Addresses
  - Due Monday 05/10 – available online
- Exploring NYTimes Obituaries
- Scraping and finding biases in CULPA reviews
- Machine Learning

Final Project

- Develop Research Question
- Collect Textual Data to Answer Question
- Data Exploration & Analysis
- Machine Learning
  - Prediction or clustering

Final Project – Deliverables

- Project ideation – Friday May 21st
- Project proposal – Friday June 4th
- Project presentations – Monday June 14th
- Project submissions – Friday June 18th
- http://coms2710.barnard.edu/final_project

Grading

Participation	5%
4 Homeworks	30%
Reading reflections	10%
Daily Tutorials	15%
Final Project	35%

Participation Grade

- During class meetings:
  - Topic discussion
  - Asking questions
- Asynchronous
  - Active on Slack (questions & answering)
  - Watching lectures

Assignment Logistics

- Distribution:
  - Instructions:
    - https://coms2710.barnard.edu/schedule.html
  - Materials:
    - Columbia JupyterServer
- Gradescope (for submission)

Course Staff

Course staff

Adam Poliak (apoliak@barnard.edu)

- PhD in Computer Science from Johns Hopkins University
- First year at Barnard
- Research:
  - Natural Language Processing
  - Data Science applied to text data

Course staff - TA

Gauri Narayan
gn2271@barnard.edu

- BA Computer Science, Barnard ‘20
- Master’s Computer Science, Columbia
- TA-ed 2 previous NLP classes
- 2 hours of office hours a week

Course staff - Preceptor

Susu Rawwagah

Barnard Political Science ’21

Our job is to help you succeed!

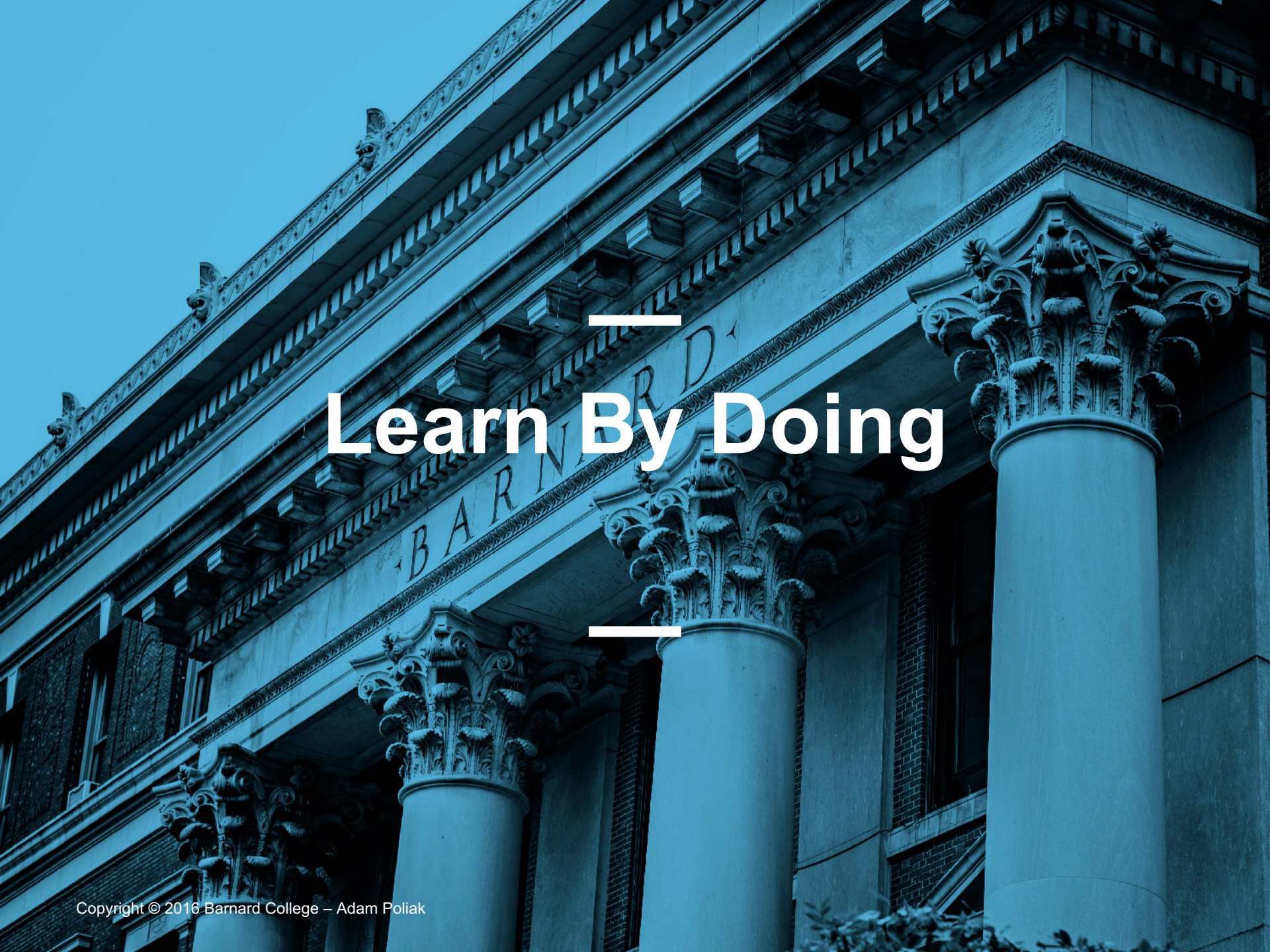
Office Hours

- Roughly 6 hours a week
- Times based on your interests
  - Complete poll found in Slack
- Possibly additional by appointment

Course Policies

Collaboration

- Encouraged to discuss problems
- Do not share solutions



Learn By Doing

Jupyter Lab Demo