



BC COMS 2710: **Computational Text Analysis**

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

Lecture 1 – Course Introduction 05/03/2020



What is Computational Text Analysis?

BIG
DATA
& SOCIETY

Big Data & Society
July–December
© The Author
Reprints and
sagepub.com/
DOI: 10.1177/
bds.sagepub.c
SAGE

ECR Forum

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)

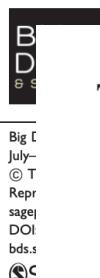
Paul DiMaggio

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 train 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 computers trust human judgment more than social scientists do. These differences have implications that potentially practice of social science.

Keywords

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



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 contingencies: corpus curation, data processing and analysis, and visualization. However, the complexity of algorithms

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 scientists, methodologists must contribute new methods and new methods of validation.

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

practice

- “~~Data science is the study of extracting value from data~~” –

large ^ scale textual

~~Jeannette Wing~~

Adam Poliak



- *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>



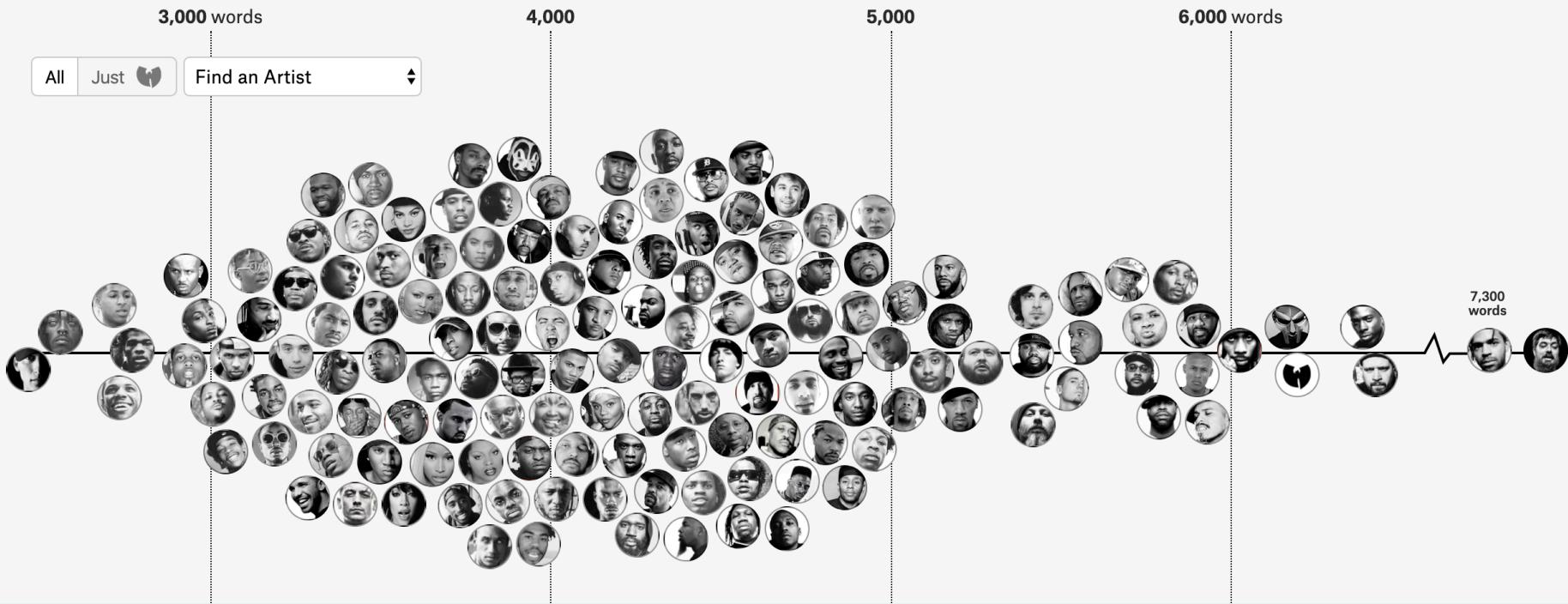
What can we do with computational text analysis?

What can we do with large scale textual analysis?



- Sort artists by their vocabulary

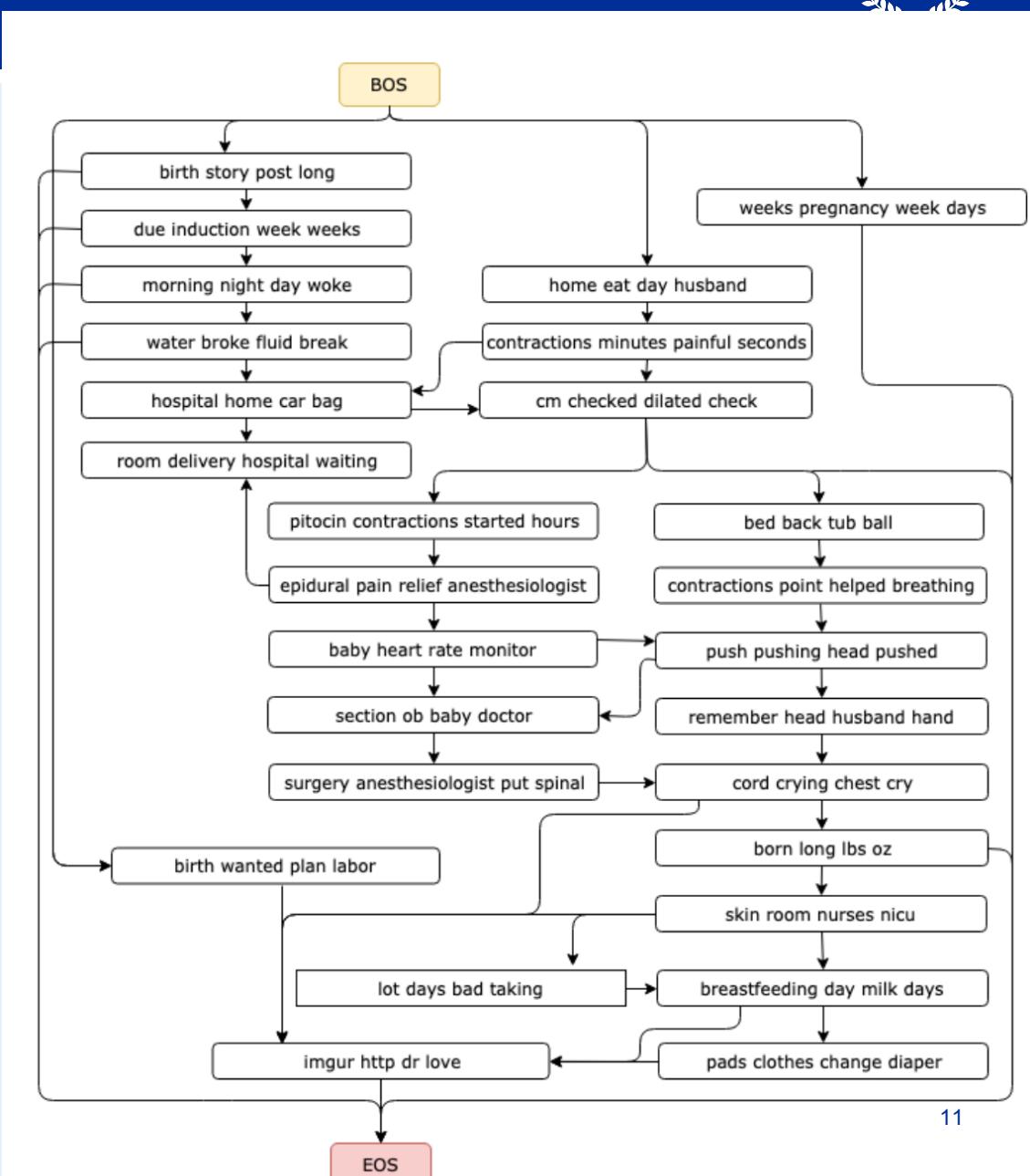
of Unique Words Used Within Artist's First 35,000 Lyrics



What can we do with large scale textual analysis?



- Identify flow of topics in birthing narratives

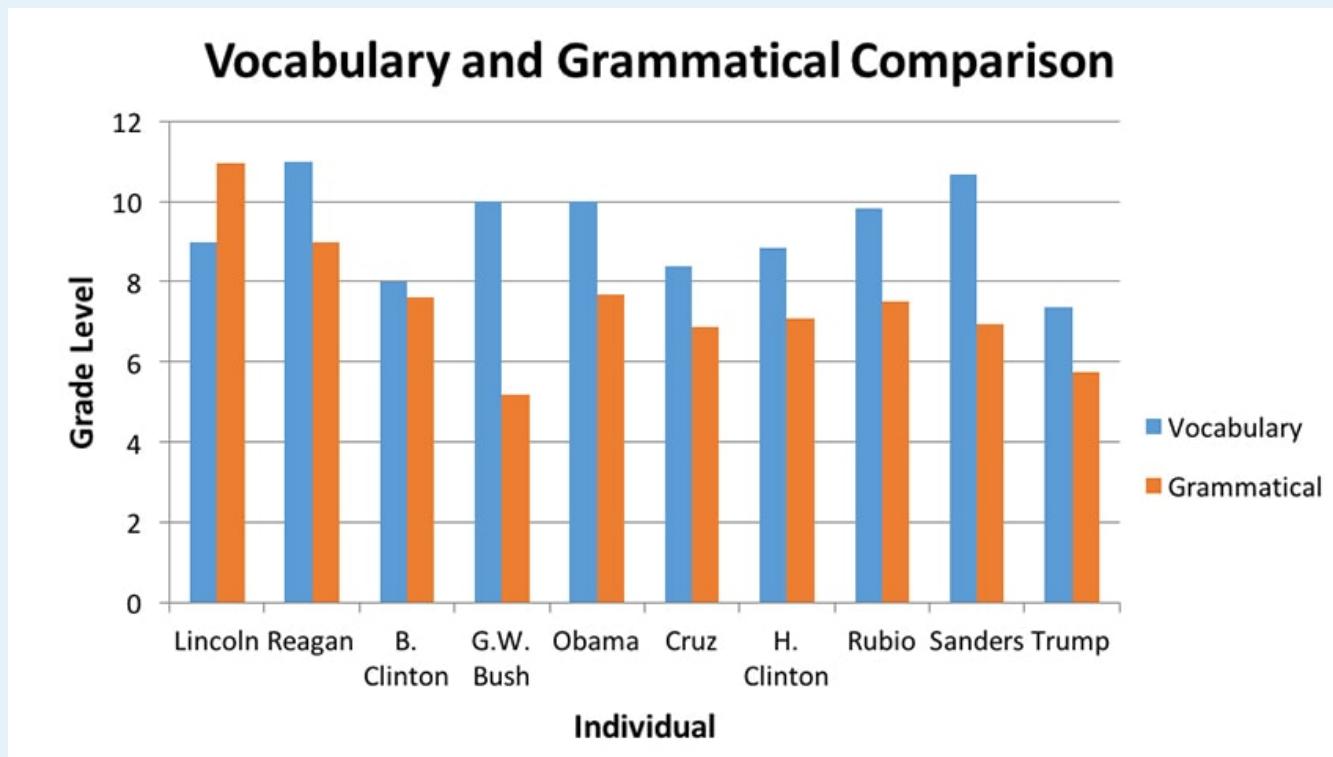


https://maria-antoniak.github.io/resources/2019_cscw_birth_stories.pdf

What can we do with large scale textual analysis?



- Categorize the level of presidential candidates' speeches

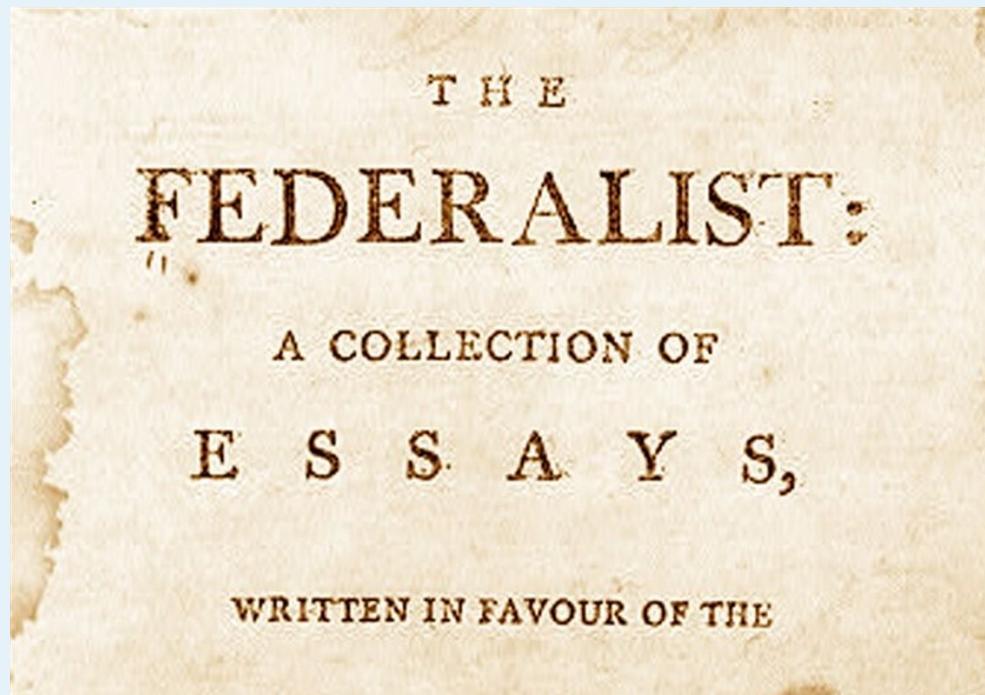


<https://arxiv.org/pdf/1603.05739.pdf>

What can we do with large scale textual analysis?



- Who wrote the anonymous Federalist Papers?



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

What can we do with large scale textual analysis?



A
l
o
l
c
.

Computational Text Analysis in this course



- Aggregate large scale textual data
- Discover patterns in data

Course Objectives



Learn the tools and gain the confidence to independently:

1. Aggregate large scale textual data
2. Discover patterns in data



Course Outline

- Python overview

Course Outline



- Python Overview
 - Introduction to Python
 - Pandas

- Lexical based analysis methods
 - Text Processing
 - Word & Document Representation
 - Topic Modeling

- Data Collection
 - Web Scraping
 - APIs

Week 1

Week 2 - 3

Week 4



- Machine Learning Week 5
 - Regression & Classification
 - Clustering

- Advanced Topics & Final Projects Week 6

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





— Logistics —



- 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

- # announcements
 - # final-project
 - # find-a-partner
 - # homeworks
 - # in-person-offic...
 - # jupyterhub
 - # office-hours
 - # random
 - # tutorials
- + Add channels



Slack - Announcements

announcements

final-project

find-a-partner

homeworks

in-person-offic...

jupyterhub

office-hours

random

tutorials

+ Add channels

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

Slack – Find-a-Partner



announcements
final-project
find-a-partner
homeworks
in-person-offic...
jupyterhub
office-hours
random
tutorials

+ Add channels

A list of Slack channels. The channel '# find-a-partner' is highlighted with a red rectangular box around its text and the entire list item.

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

Slack – Homeworks/Tutorials



```
# announcements  
# final-project  
# find-a-partner  
# homeworks  
# in-person-offic...  
# jupyterhub  
# office-hours  
# random  
# tutorials
```

+ Add channels

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

Slack – Office-Hours



```
# announcements  
# final-project  
# find-a-partner  
# homeworks  
# in-person-offic...  
# jupyterhub  
# office-hours  
# random  
# tutorials  
+ Add channels
```

- 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

```
# announcements  
# final-project  
# find-a-partner  
# homeworks  
# in-person-offic...  
# jupyterhub  
# office-hours  
# random  
# tutorials
```

+ Add channels

- Potential in-person office hours



- 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

Assignments

Learn By Doing



Assignments

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

Daily-ish Exercises



Saturday night @ midnight

- Due ~~M/T/W/R~~ *Saturday night @ midnight*
- Complete individually
- ~1.5 hours long

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 (optional)
 - 1 question for future work
- Goal: Examples of computational text analysis
 - Preparation for final projects

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

Grading



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

Participation Grade

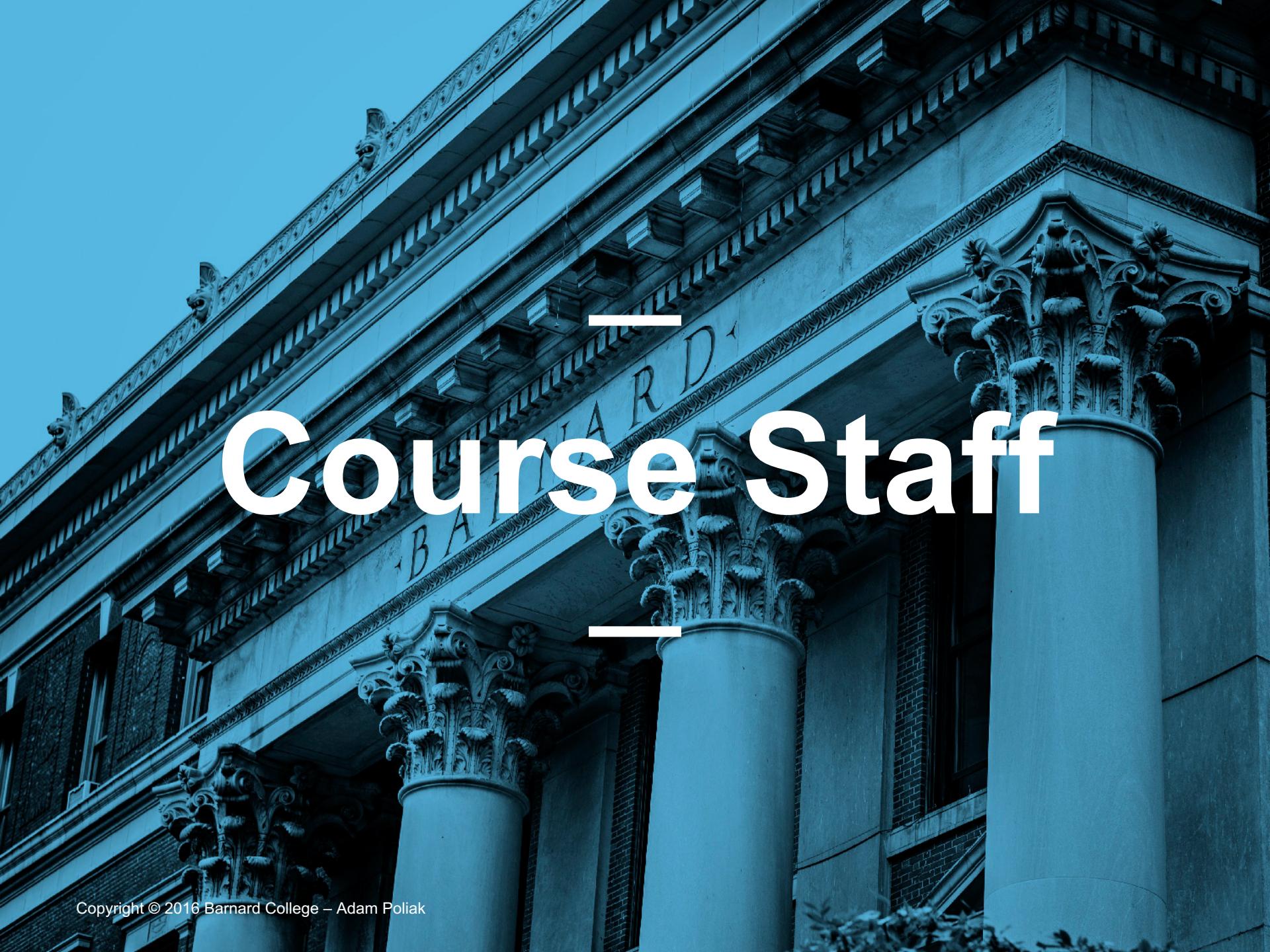


- During class meetings:
 - Topic discussion
 - Asking questions
- Asynchronous
 - Active on Slack (questions & answering)
 - Watching lectures

Assignment Logistics



- Distribution:
 - Instructions:
 - <https://coms2710.barnard.edu/schedule.html>
 - Materials:
 - Columbia Jupyter Server
- Gradescope (for submission)



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

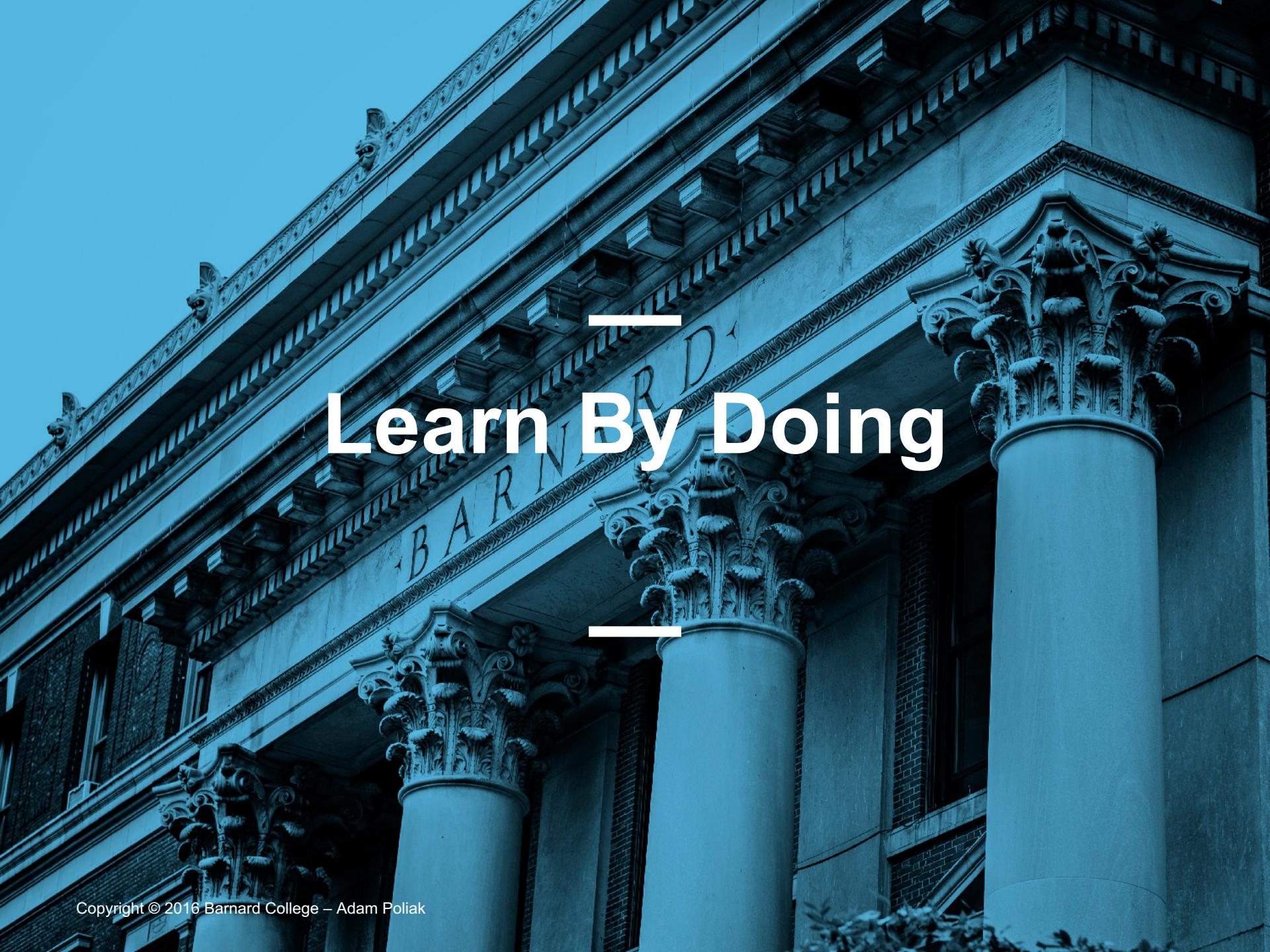


Course Policies

Collaboration



- Encouraged to discuss problems
- Do not share solutions



Learn By Doing

Jupyter Lab Demo