Data as Culture

Dr. Melanie Walsh // Assistant Teaching Professor // melwalsh@uw.edu

Women in Data Science Conference

April 1, 2022

Information School
UNIVERSITY of WASHINGTON

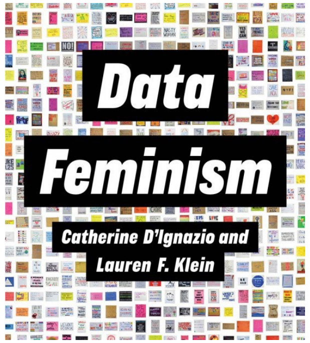
What is data science?
Who counts as a data scientist?

What is data science?
Who counts as a data scientist?

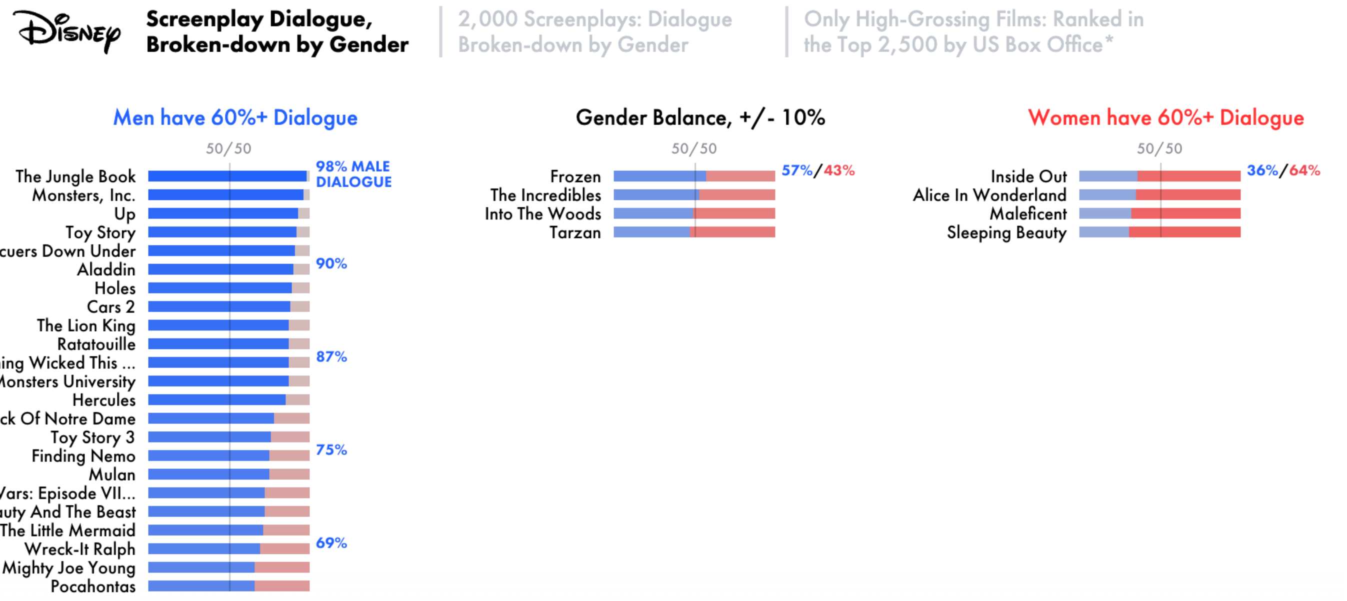
“Many people think of data as numbers alone, but **data can also consist of words or stories, colors or sounds, or any type of information that is systematically collected, organized, and analyzed**.

The *science* in data science simply implies a commitment to systematic methods of observation and experiment.”

-Catherine D’Ignazio and Lauren Klein, *Data Feminism*



What is data science?
Who counts as a data scientist?



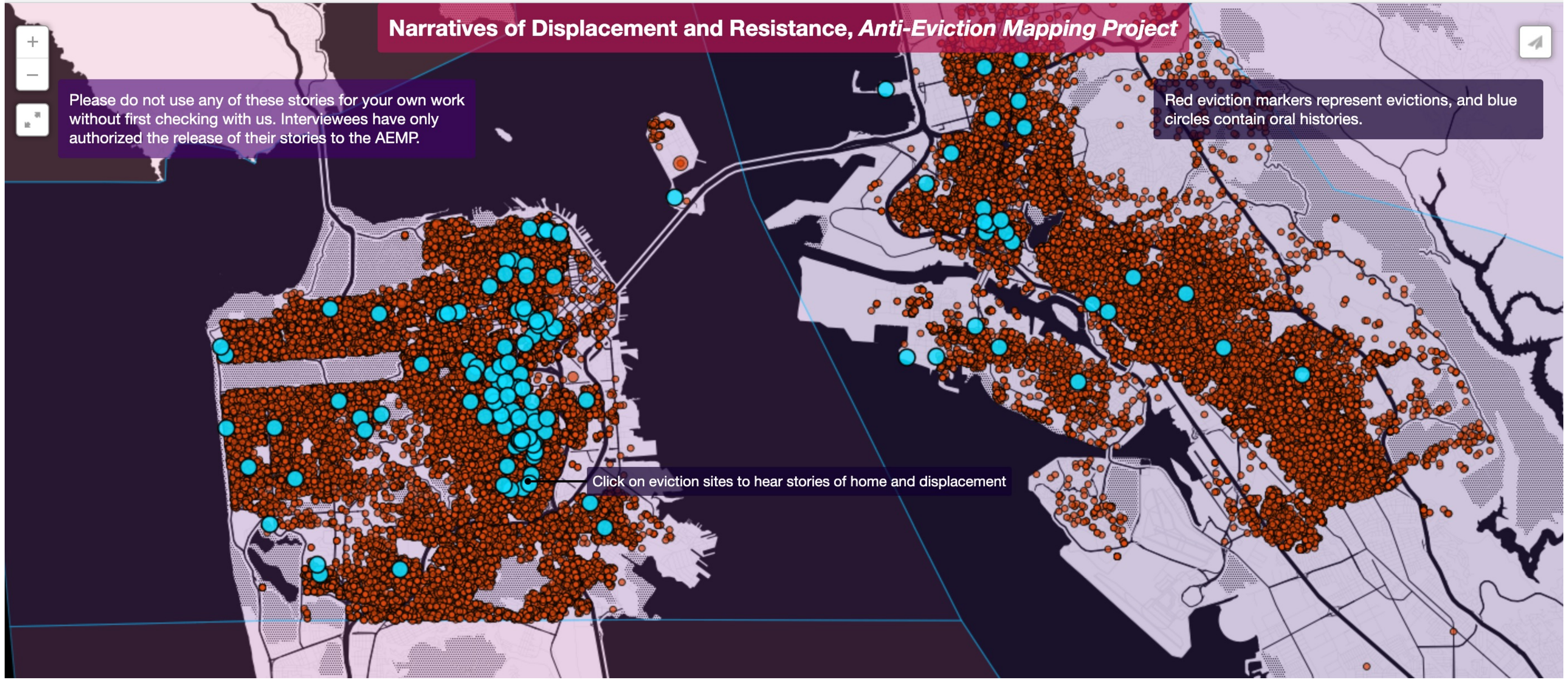
"Film Dialogue from 2,000 screenplays, Broken Down by Gender and Age," Hannah Anderson and Matt Daniels

What is data science?
Who counts as a data scientist?

THE NATIONAL COLORED CONVENTION IN SESSION AT WASHINGTON, D. C.—SKETCHED BY THEO. R. DAVIS.--[SEE FIRST PAGE.]

Colored Conventions Project
https://coloredconventions.org/

What is data science?
Who counts as a data scientist?



Narratives of Displacement and Resistance, *Anti-Eviction Mapping Project*

Please do not use any of these stories for your own work without first checking with us. Interviewees have only authorized the release of their stories to the AEMP.

Red eviction markers represent evictions, and blue circles contain oral histories.

Click on eviction sites to hear stories of home and displacement

The Anti-Eviction Mapping Project
http://www.antievictionmappingproject.net/

What is data science?
Who counts as a data scientist?

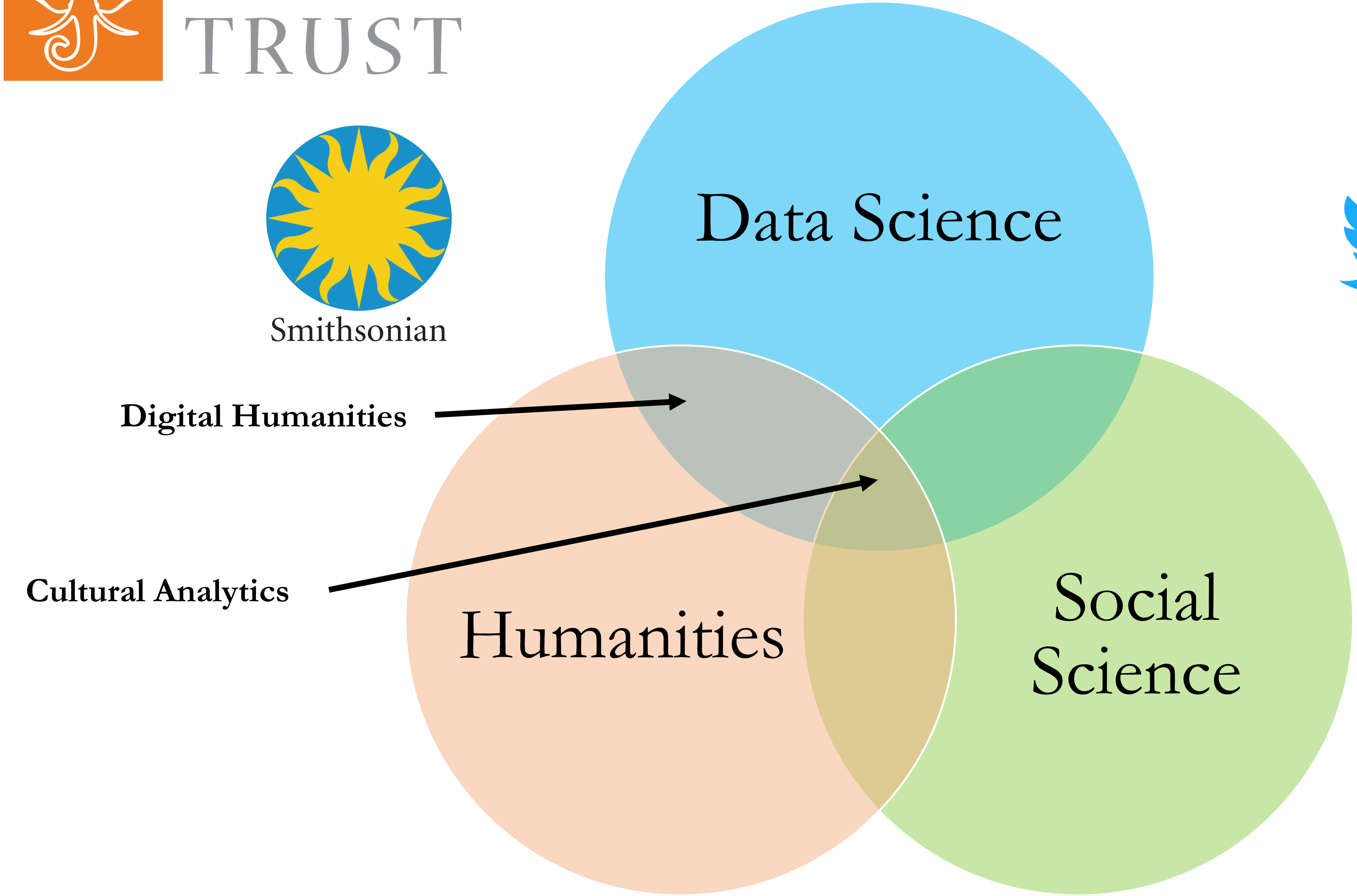
“Throughout this book, we deliberately place diverse data science examples alongside each other. They come from individuals and small groups, and from across academic, artistic, nonprofit, journalistic, community-based, and for-profit organizations. This is due to our belief in a capacious definition of data science, one that seeks to **include rather than exclude and does not erect barriers based on formal credentials, professional affiliation, size of data, complexity of technical methods, or other external markers of expertise.**

Such markers, after all, have long been used to prevent women from fully engaging in any number of professional fields, even as those fields—which include data science and computer science, among many others—were largely built on the knowledge that women were required to teach themselves.”

-Catherine D’Ignazio and Lauren Klein, *Data Feminism*.”

HATHI TRUST

Smithsonian



Data Science

Digital Humanities

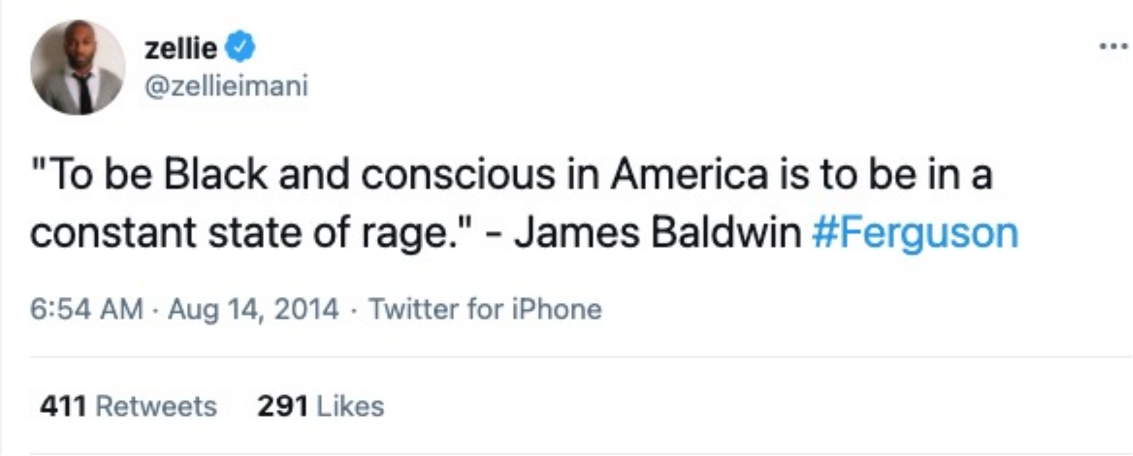
Cultural Analytics

Humanities

Social Science

goodreads

amazon

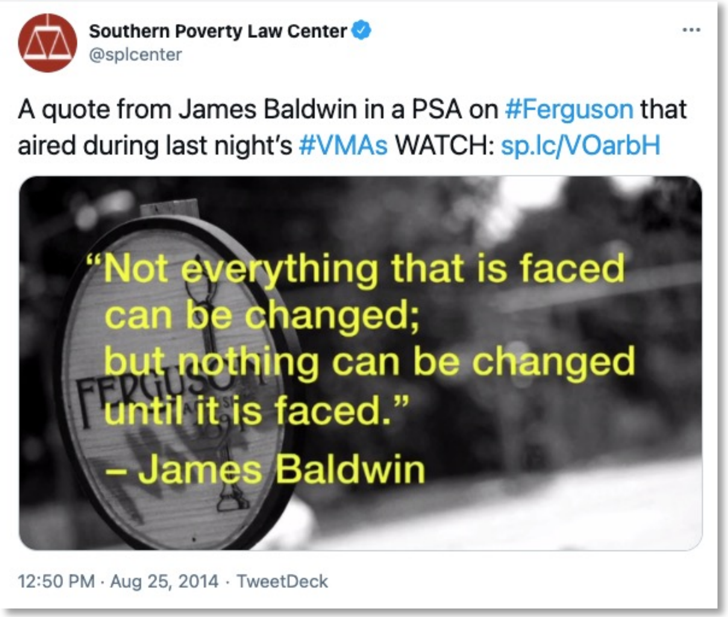


zellie @zellieimani

"To be Black and conscious in America is to be in a constant state of rage." - James Baldwin #Ferguson

6:54 AM · Aug 14, 2014 · Twitter for iPhone

411 Retweets 291 Likes



Southern Poverty Law Center @splcenter

A quote from James Baldwin in a PSA on #Ferguson that aired during last night's #VMAs WATCH: sp.lc/VOarbH

"Not everything that is faced can be changed; but nothing can be changed until it is faced."
– James Baldwin

12:50 PM · Aug 25, 2014 · TweetDeck

My Training

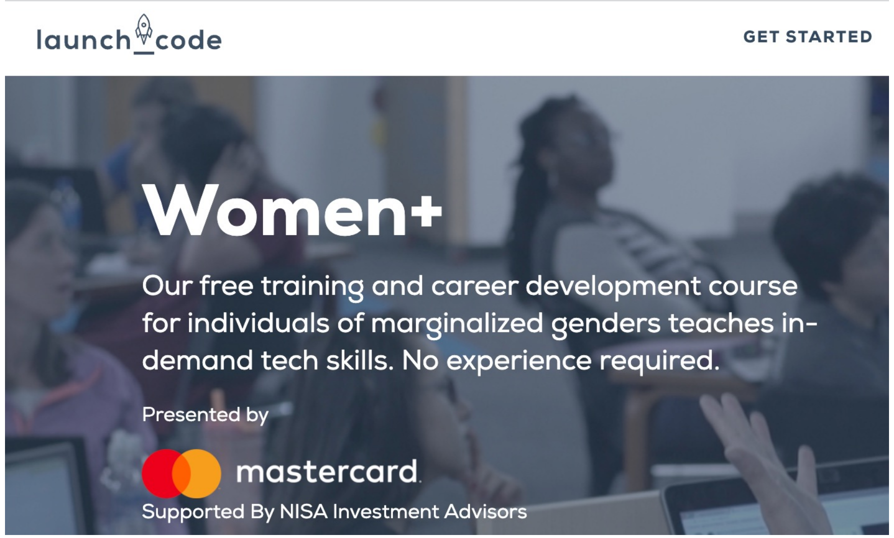
- Graduate fellowships and research assistantships that included programming training
- Online resources like CodeAcademy
- Data science courses and bootcamps for women+ hosted by local non-profits

Washington University in St. Louis
Arts & Sciences

Humanities Digital Workshop

Graduate and Undergraduate Fellowships

codecademy



launchcode

GET STARTED

Women+

Our free training and career development course for individuals of marginalized genders teaches in-demand tech skills. No experience required.

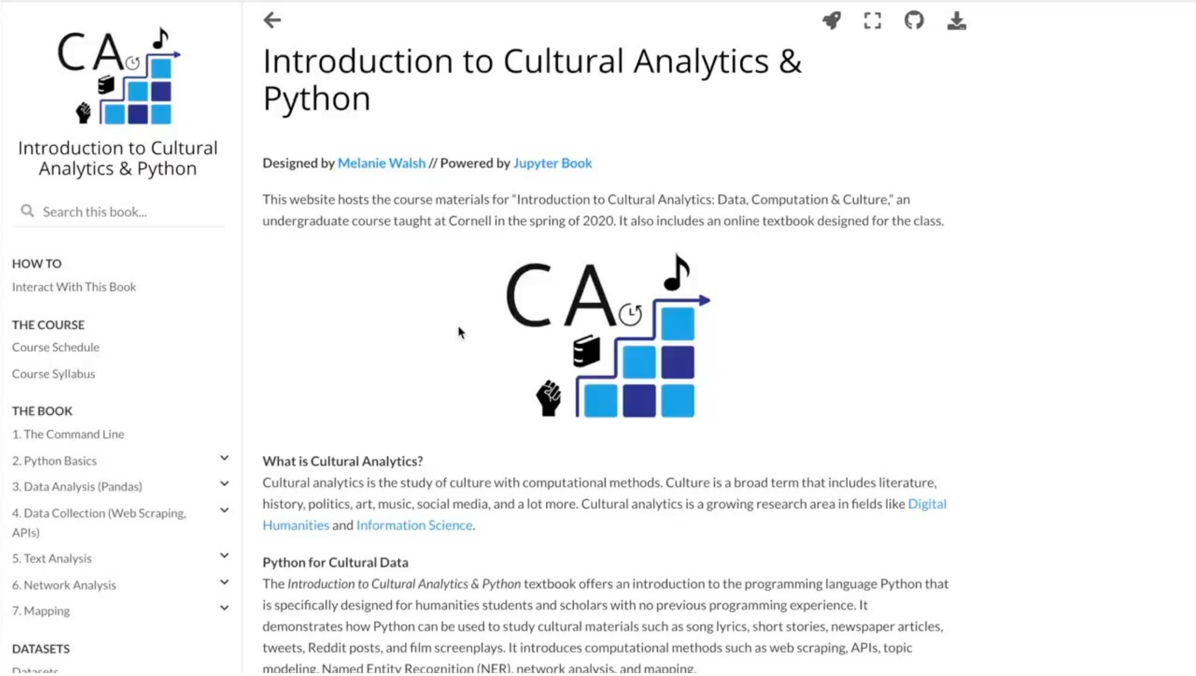
Presented by

mastercard

Supported By NISA Investment Advisors

Introduction to Cultural Analytics & Python

- Free, open-source, interactive textbook:
  - Python
  - Data analysis (Pandas)
  - Text analysis
  - Network Analysis
  - Mapping



Introduction to Cultural Analytics & Python

Designed by Melanie Walsh // Powered by Jupyter Book

This website hosts the course materials for "Introduction to Cultural Analytics: Data, Computation & Culture," an undergraduate course taught at Cornell in the spring of 2020. It also includes an online textbook designed for the class.

What is Cultural Analytics?
Cultural analytics is the study of culture with computational methods. Culture is a broad term that includes literature, history, politics, art, music, social media, and a lot more. Cultural analytics is a growing research area in fields like Digital Humanities and Information Science.

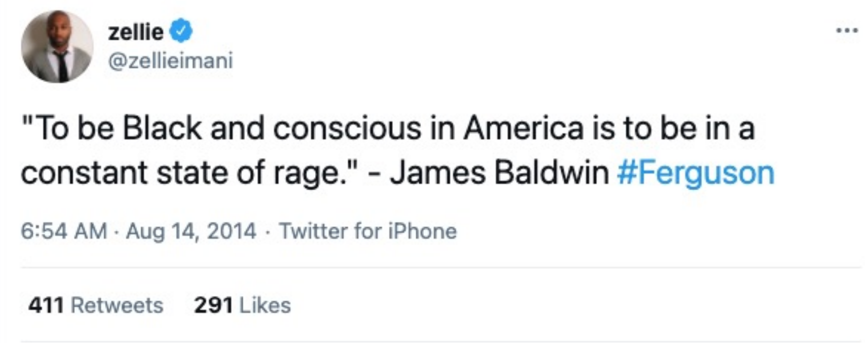
Python for Cultural Data
The *Introduction to Cultural Analytics & Python* textbook offers an introduction to the programming language Python that is specifically designed for humanities students and scholars with no previous programming experience. It demonstrates how Python can be used to study cultural materials such as song lyrics, short stories, newspaper articles, tweets, Reddit posts, and film screenplays. It introduces computational methods such as web scraping, APIs, topic modeling, Named Entity Recognition (NER), network analysis, and mapping.

Voted "Best Digital Humanities Training Material" 2021

Tweets of a Native Son: The Quotation and Recirculation of James Baldwin from Black Power to #BlackLivesMatter

American Quarterly, 2018

www.TweetsofaNativeSon.com

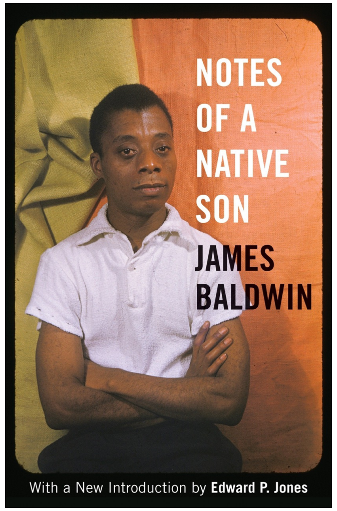


zellie
@zellieimani

"To be Black and conscious in America is to be in a constant state of rage." - James Baldwin #Ferguson

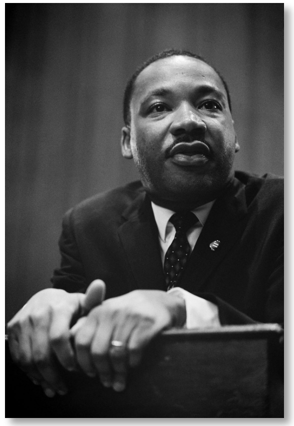
6:54 AM · Aug 14, 2014 · Twitter for iPhone

411 Retweets 291 Likes



Motivating Questions

- Why were people quoting James Baldwin in #BlackLivesMatter tweets so often?
- How does Baldwin's quotation compare to other African American writers?



Martin Luther King
(1929-1968)

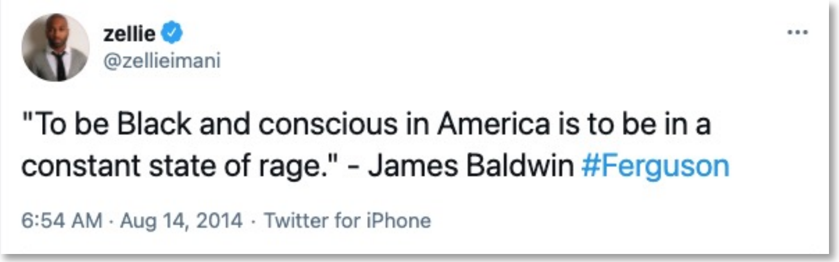
James Baldwin
(1924-1987)

Toni Morrison
(1931-2019)

Langston Hughes
(1901-1967)

Motivating Questions

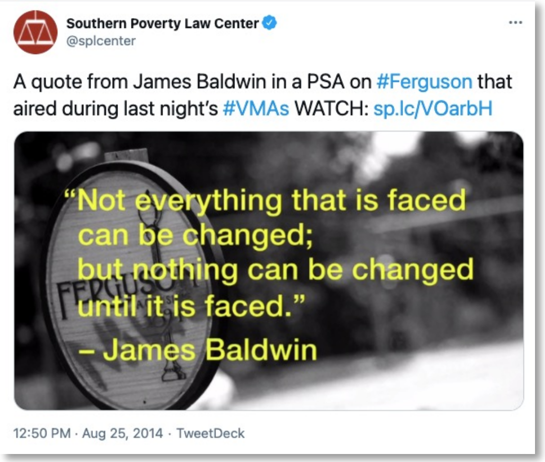
- Which Baldwin quotations were most popular? Why?
- What makes a quotation go viral?



zellie
@zellieimani

"To be Black and conscious in America is to be in a constant state of rage." - James Baldwin #Ferguson

6:54 AM · Aug 14, 2014 · Twitter for iPhone



Southern Poverty Law Center
@splcenter

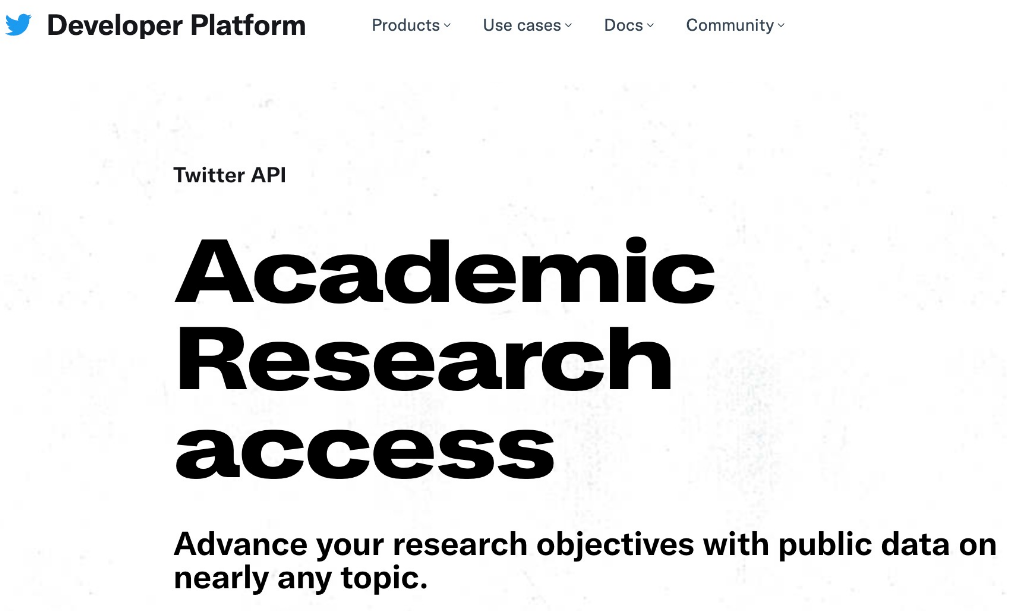
A quote from James Baldwin in a PSA on #Ferguson that aired during last night's #VMAs WATCH: sp.lc/VOarbH

"Not everything that is faced can be changed; but nothing can be changed until it is faced."
– James Baldwin

12:50 PM · Aug 25, 2014 · TweetDeck

BEYOND THE HASHTAGS
#Ferguson, #Blacklivesmatter, and the online struggle for offline justice
DEEN FREELON
CHARLTON D. MCILWAIN
MEREDITH D. CLARK
CMSi
CENTER FOR MEDIA & SOCIAL IMPACT
SCHOOL of COMMUNICATION
AMERICAN UNIVERSITY • WASHINGTON, DC
43
Tweet

Freelon, McIlwain, and Clark shared ~40 million tweet IDs related to the #BlackLivesMatter movement between 2014-2015



Twitter API

Academic Research access

Advance your research objectives with public data on nearly any topic.

Free access to all historical tweets for researchers, faculty, and graduate students who apply — 10 million tweets per month

BEYOND THE HASHTAGS
#Ferguson, #Blacklivesmatter, and the online struggle for offline justice
DEEN FREELON
CHARLTON D. MCILWAIN
MEREDITH D. CLARK
CMSi
CENTER FOR MEDIA & SOCIAL IMPACT
SCHOOL of COMMUNICATION
AMERICAN UNIVERSITY • WASHINGTON, DC
43
Tweet

Freelon, McIlwain, and Clark shared ~40 million tweet IDs related to the #BlackLivesMatter movement between 2014-2015

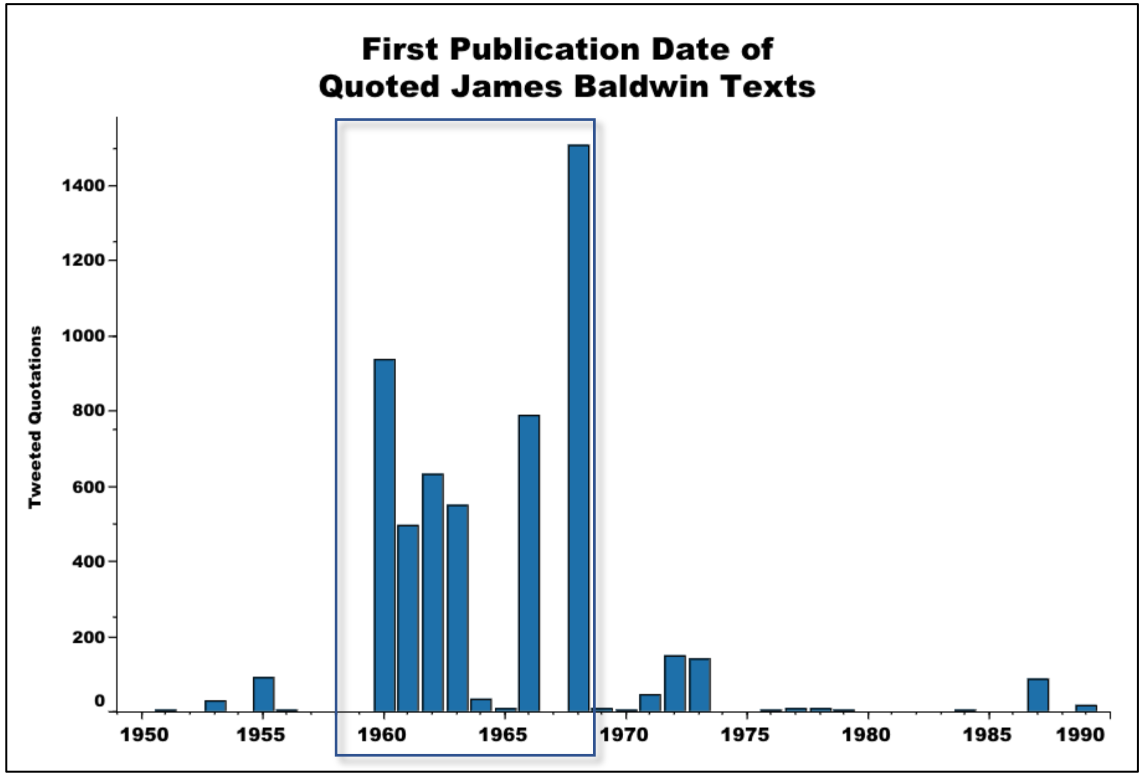
The Most Cited Canonical African American Authors in #BlackLivesMatter Tweets

THE NORTON ANTHOLOGY OF AFRICAN AMERICAN LITERATURE THIRD EDITION

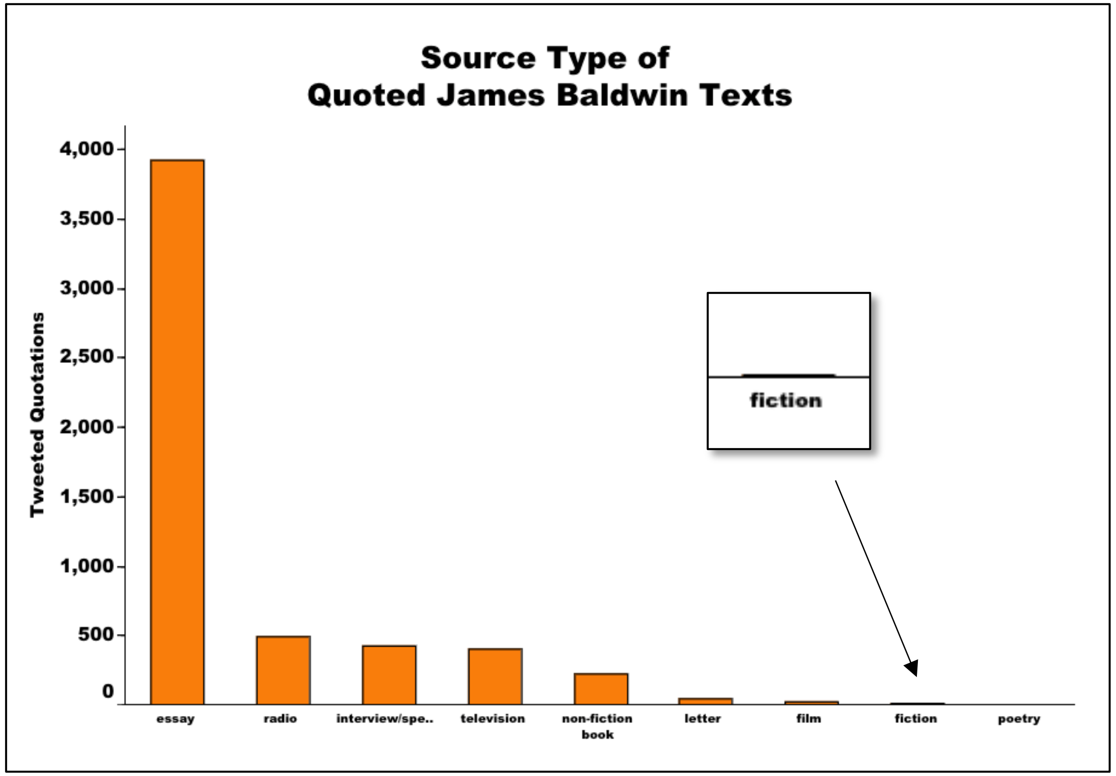
BEYOND THE HASHTAGS

2014 – 2015

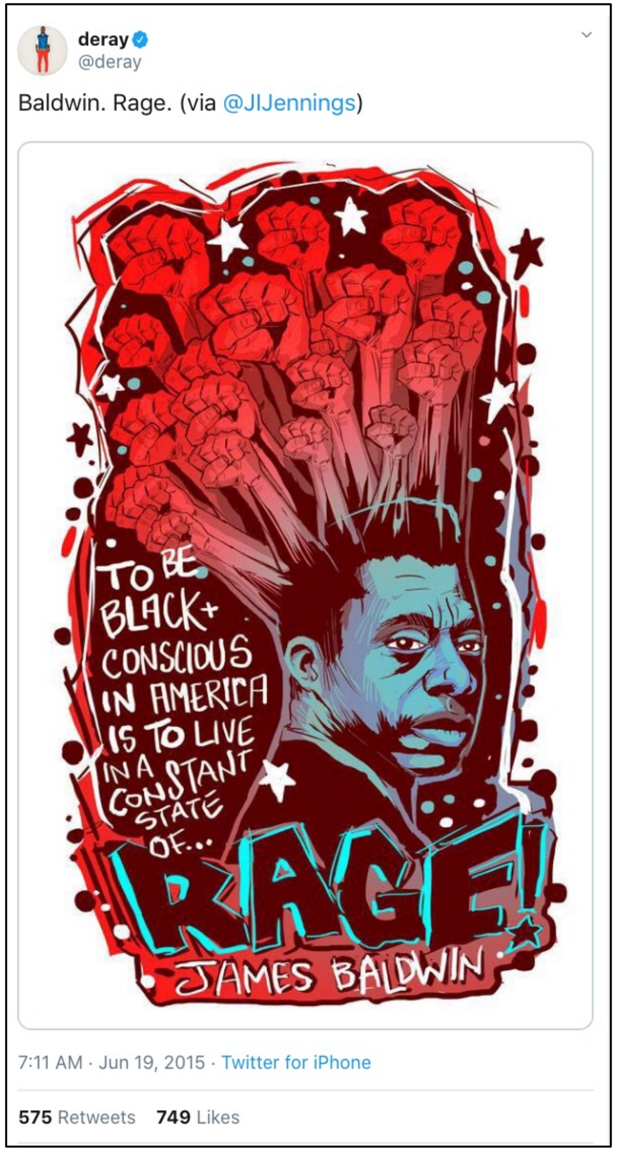
Hover for more information ?



First Publication Date of
Quoted James Baldwin Texts



Source Type of
Quoted James Baldwin Texts



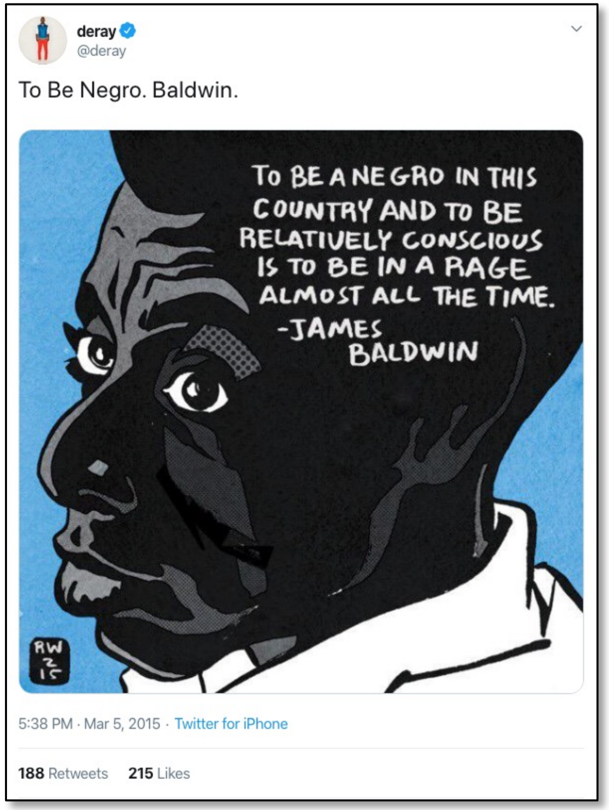
deray
@deray

Baldwin. Rage. (via @JlJennings)

TO BE BLACK+ CONSCIOUS IN AMERICA IS TO LIVE IN A CONSTANT STATE OF...
RAGE!
JAMES BALDWIN

7:11 AM · Jun 19, 2015 · Twitter for iPhone

575 Retweets 749 Likes



deray
@deray

To Be Negro. Baldwin.

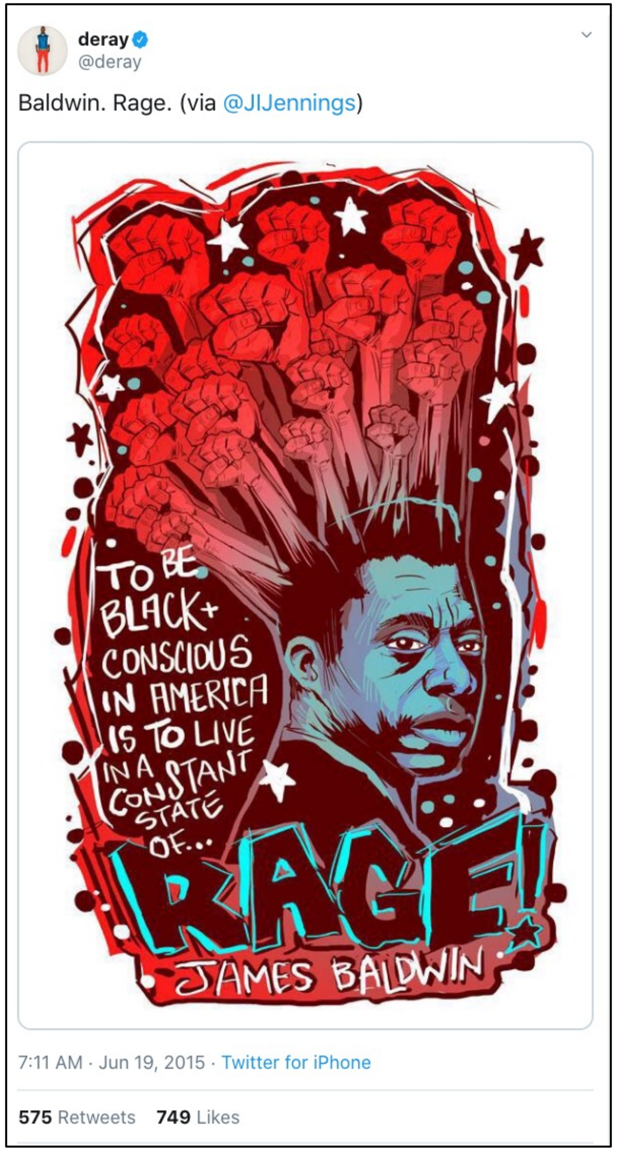
TO BE A NEGRO IN THIS COUNTRY AND TO BE RELATIVELY CONSCIOUS IS TO BE IN A RAGE ALMOST ALL THE TIME.
-JAMES BALDWIN
RW 2 15

5:38 PM · Mar 5, 2015 · Twitter for iPhone

188 Retweets 215 Likes

"To be Black and conscious in America is to be in a constant state of rage."

"To be a Negro in this country and to be relatively conscious, is to be in a rage almost all the time."



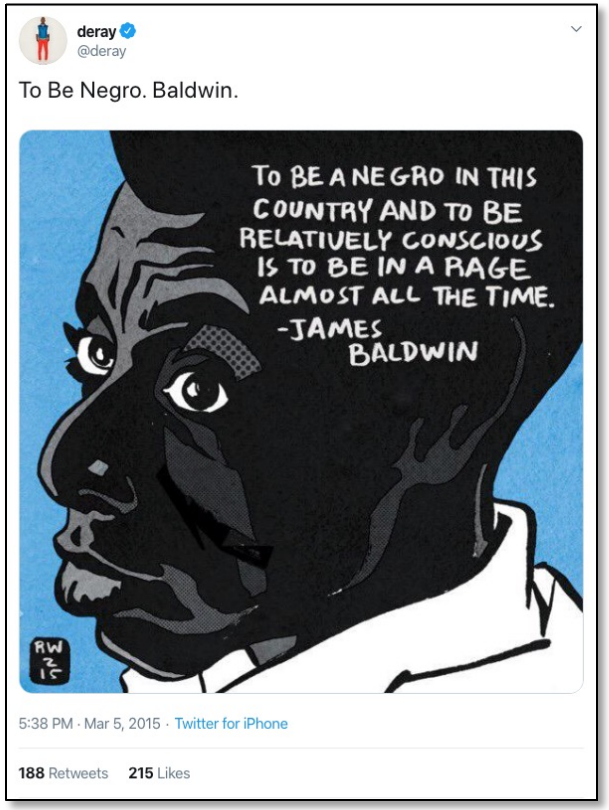
deray @deray

Baldwin. Rage. (via @JIJennings)

TO BE BLACK+ CONSCIOUS IN AMERICA IS TO LIVE IN A STANT CONSTANT STATE OF...
RAGE!
JAMES BALDWIN

7:11 AM · Jun 19, 2015 · Twitter for iPhone

575 Retweets 749 Likes



deray @deray

To Be Negro. Baldwin.

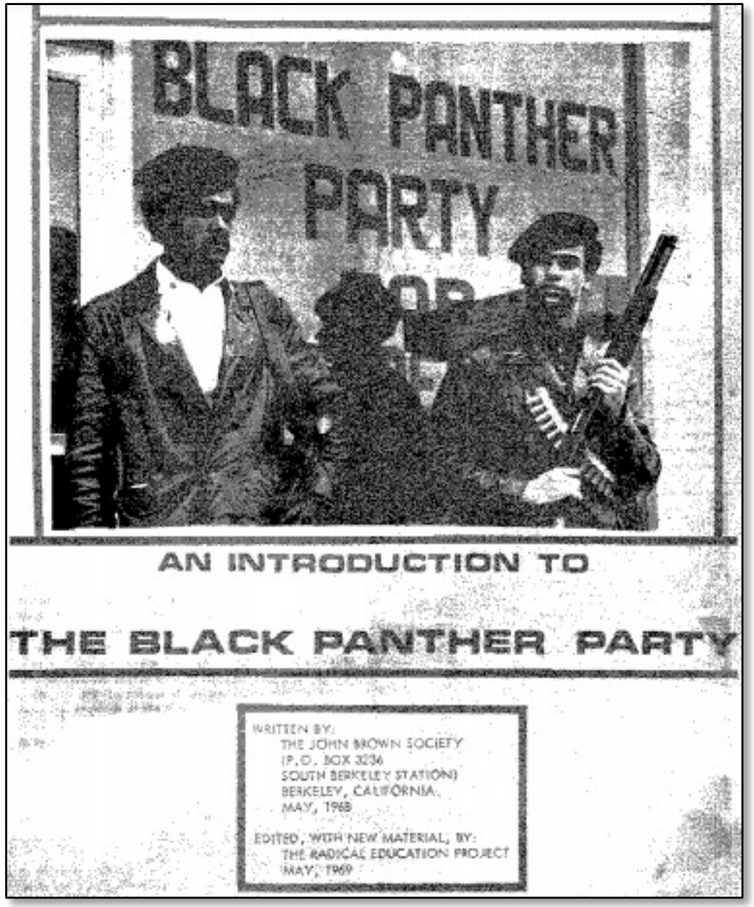
TO BE A NEGRO IN THIS COUNTRY AND TO BE RELATIVELY CONSCIOUS IS TO BE IN A RAGE ALMOST ALL THE TIME.
-JAMES BALDWIN
RW 3 15

5:38 PM · Mar 5, 2015 · Twitter for iPhone

188 Retweets 215 Likes

"To be Black and conscious in America
is to be in a constant state of rage."

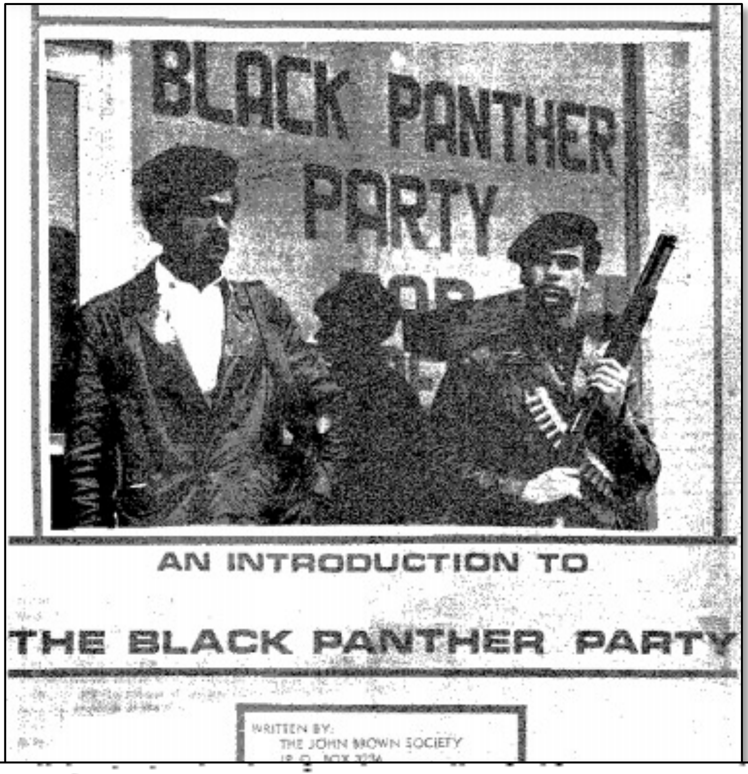
"To be a Negro in this country and to be relatively conscious,
is to be in a rage almost all the time."



AN INTRODUCTION TO THE BLACK PANTHER PARTY

WRITTEN BY:
THE JOHN BROWN SOCIETY
(P.O. BOX 3236
SOUTH BERKELEY STATION)
BERKELEY, CALIFORNIA
MAY, 1968

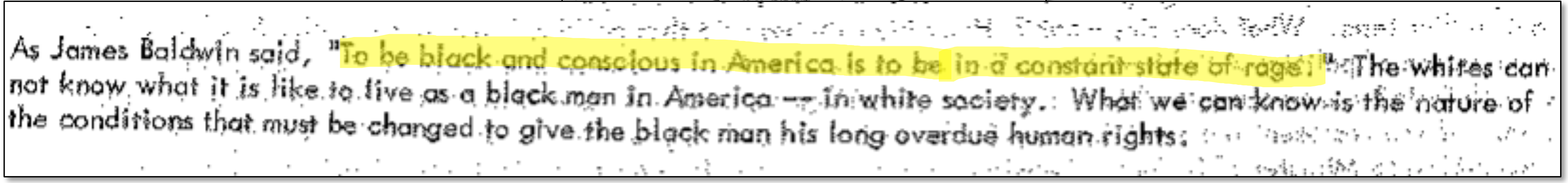
EDITED, WITH NEW MATERIAL, BY:
THE RADICAL EDUCATION PROJECT
MAY, 1969



AN INTRODUCTION TO

THE BLACK PANTHER PARTY

WRITTEN BY:
THE JOHN BROWN SOCIETY
P.O. BOX [illegible]



As James Baldwin said, "To be black and conscious in America is to be in a constant state of rage." The whites can not know what it is like to live as a black man in America -- in white society. What we can know is the nature of the conditions that must be changed to give the black man his long overdue human rights.

Humanities ⟷ Data Science

- Data and computational methods offer new window into how culture circulates among “everyday” people

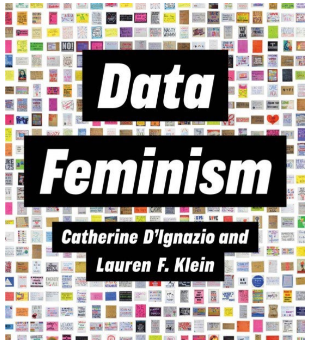
- Fully understanding this data requires historical context and analog research

“Big Dick Data”

“*Big Dick Data* is a formal, academic term that we, the authors, have coined to denote big data projects that are characterized by **patriarchal, cis-masculinist, totalizing fantasies of world domination** as enacted through data capture and analysis.

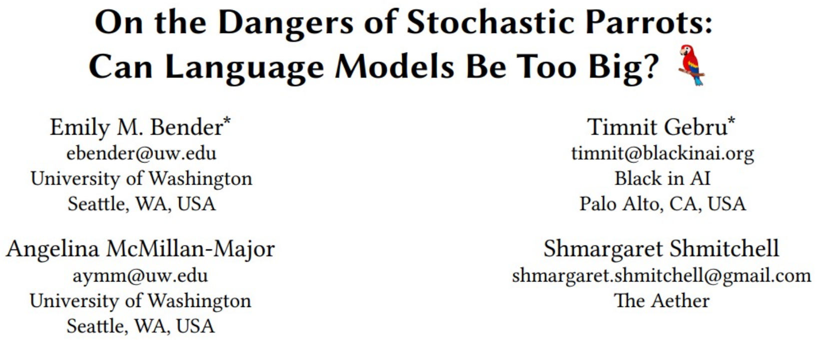
Big Dick Data projects **ignore context, fetishize size, and inflate their technical and scientific capabilities.**”

- Catherine D’Ignazio and Lauren Klein



On The Dangers of Stochastic Parrots: Can Language Models Be Too Big?

- Research paper that points out problems with large language models — natural language processing models that are trained on very large text datasets often collected from the web, such as GPT-3, GPT-2, or BERT
- These models can generate text, answer questions, summarize documents, etc.
- BERT is currently used in Google's search algorithm



On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

On The Dangers of Stochastic Parrots: Can Language Models Be Too Big?

- Risks and criticisms of large-language models:
  - These models are very expensive to train
  - These models use a lot of resources and may be harmful to the environment
  - The datasets used encode various kinds of bias (sexism, racism, toxicity, etc.)
  - **The datasets are so big they can’t be properly documented**

On The Dangers of Stochastic Parrots: Can Language Models Be Too Big?

“When we rely on ever larger datasets we risk incurring **documentation debt**, i.e. putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc. **While documentation allows for potential accountability, undocumented training data perpetuates harm without recourse.** Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones.

The solution, we propose, is to budget for documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget.”

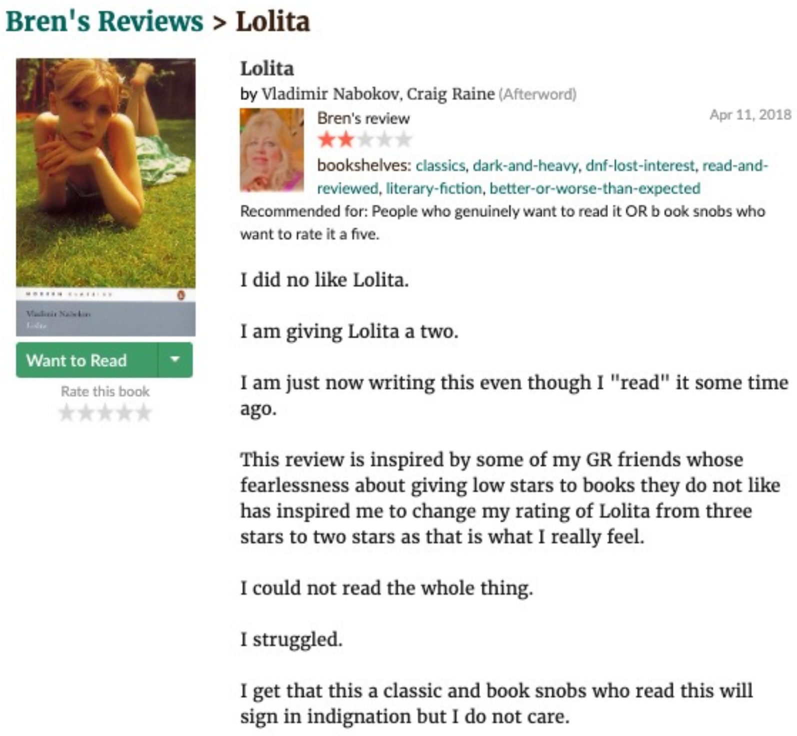
-Bender, at al, “On The Dangers of Stochastic Parrots”

The Goodreads "Classics": A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism

Melanie Walsh and Maria Antoniak

Post45 and *the Journal of Cultural Analytics*, 2021

goodreads



Bren's Reviews > Lolita

Lolita

by Vladimir Nabokov, Craig Raine (Afterword)

Bren's review ★★☆☆☆ Apr 11, 2018

bookshelves: classics, dark-and-heavy, dnf-lost-interest, read-and-reviewed, literary-fiction, better-or-worse-than-expected

Recommended for: People who genuinely want to read it OR b ook snobs who want to rate it a five.

I did no like Lolita.

I am giving Lolita a two.

I am just now writing this even though I "read" it some time ago.

This review is inspired by some of my GR friends whose fearlessness about giving low stars to books they do not like has inspired me to change my rating of Lolita from three stars to two stars as that is what I really feel.

I could not read the whole thing.

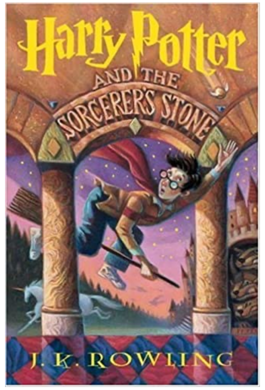
I struggled.

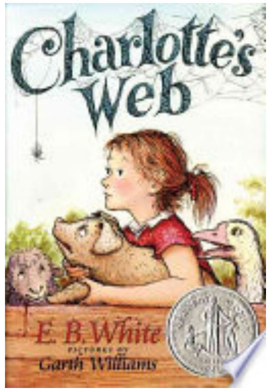
I get that this a classic and book snobs who read this will sign in indignation but I do not care.

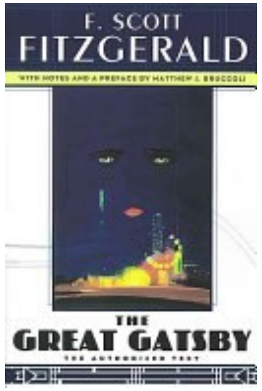
Motivation

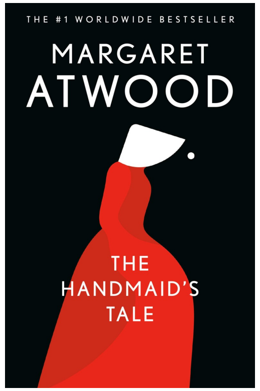
- Goodreads is the largest social networking site for readers on the internet (120 million users) and a subsidiary of Amazon
- “Classics” is one of the most active Goodreads categories, with some of the most rated and reviewed books across the entire site

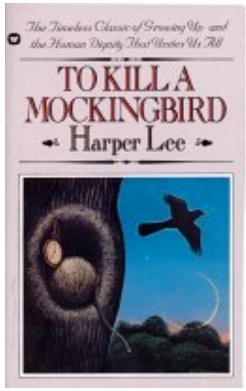
Which of these books are "classics"?

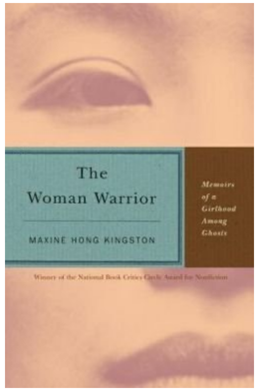


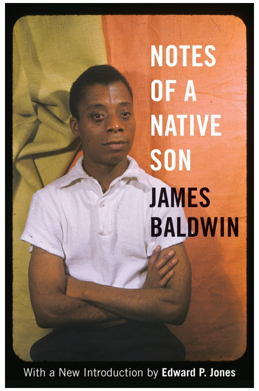


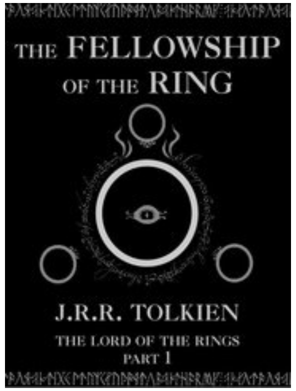




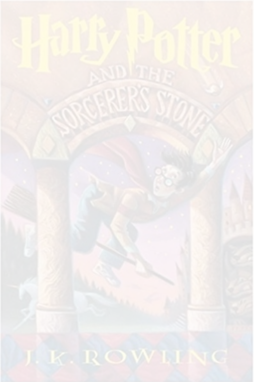


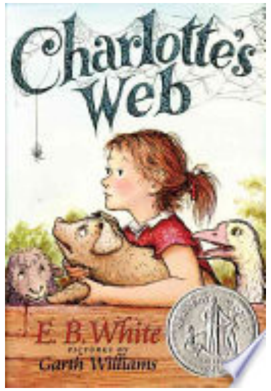


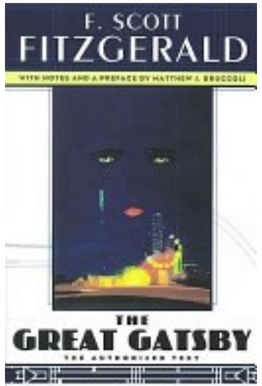


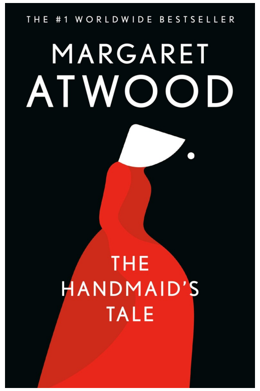


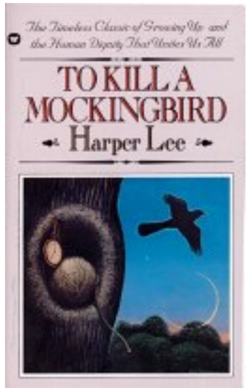
Which of these books are “classics”?

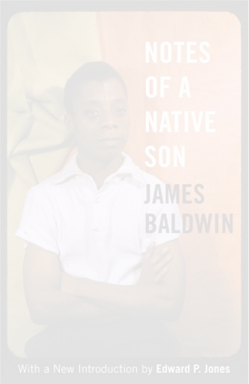


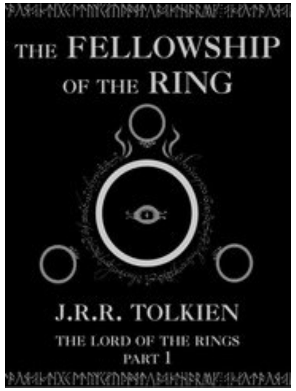







The Timeless Classic of Growing Up and the Human Dignity That Unites Us All
TO KILL A
MOCKINGBIRD
Harper Lee





Motivating Questions

- Why are the classics so popular on Goodreads? Which books have readers "shelved" as classics most often?

- What do people talk about when they talk about the classics? Why do they love them and hate them?

Our Goodreads "Classics" Dataset

- ~900 Goodreads reviews for each of 144 classic texts
    - 300 oldest, newest, and default reviews per book
    - Filtered to English-language reviews
- **127,855 total** Goodreads reviews (2007-2019)

Topic Modeling

school, high, time, class, first,
remember, years, year, english, still,
think, college, grade, since

School

school, high, time, class, first, remember, years, year, english, still, think, college, grade, since

"This was the first Toni Morrison I read for 10th grade English while I was in high school. I couldn't get into at the time… I was more prepared for the novel this time around."

-Goodreads User

Topic Modeling

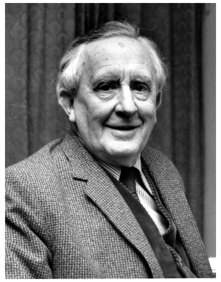
love austen jane mr emma character novels characters romance time

poirot mystery christie books murder story series tolkien agatha end

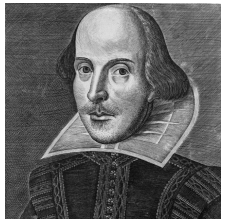
play love shakespeare hamlet plays good macbeth othello king romeo

Jane Austen

Agatha Christie



J.R.R. Tolkien

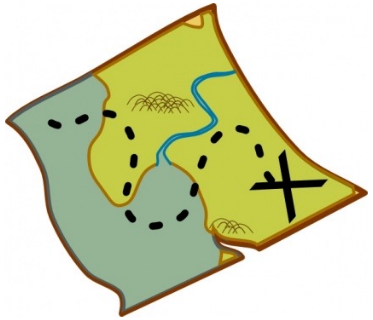


William Shakespeare

“Authorless” Topic Modeling

novel young first mrs woman love miss two family friend old marriage sister years

war world man back great time king men journey first adventure



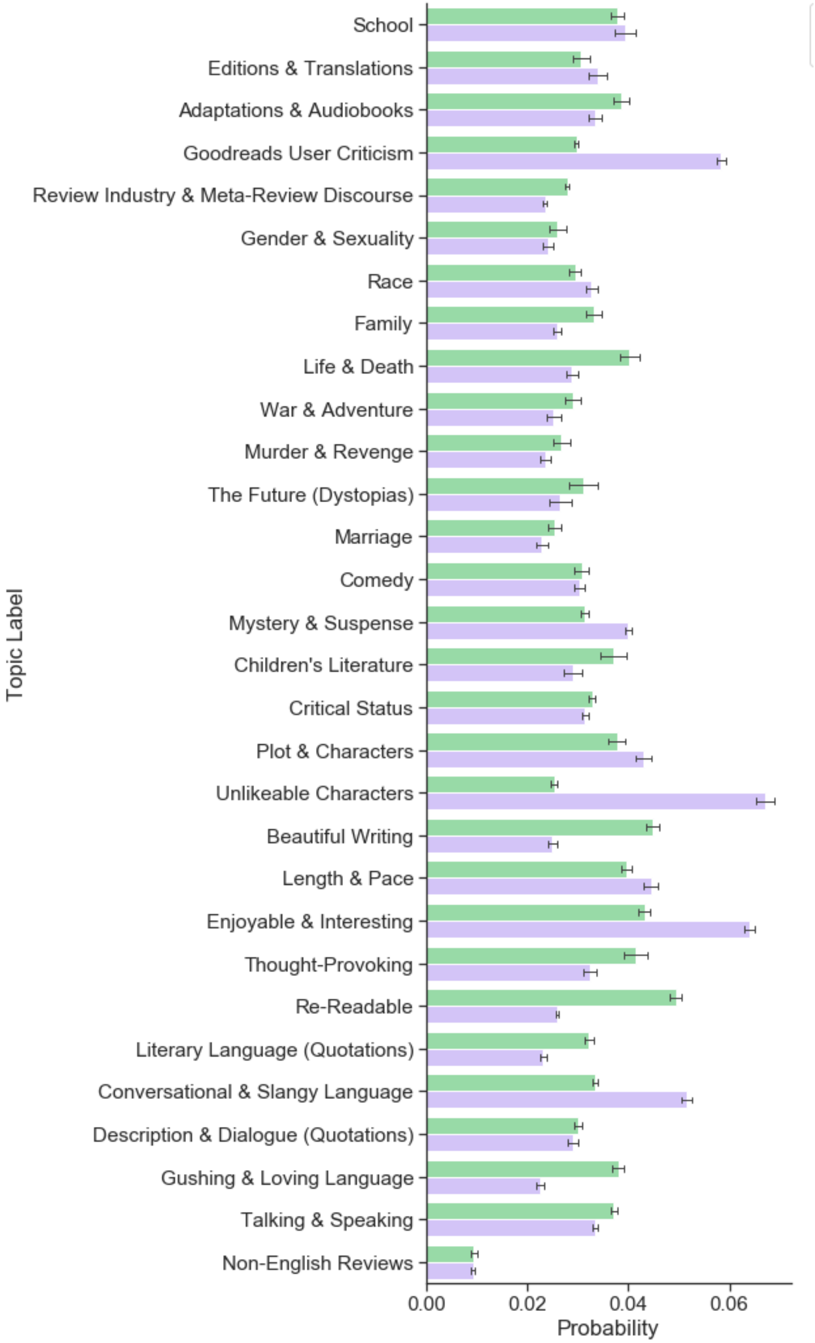
man evil death good murder love father play king two wife end crime revenge

Laure Thompson and David Mimno, “Authorless Topic Models: Biasing Models Away from Known Structure”

Topics in Goodreads Reviews of Classics

- School
- Editions & Translations
- Adaptations & Audiobooks
- Goodreads User Criticism
- Review Industry & Meta-Review Discourse
- Gender & Sexuality
- Race
- Family
- Life & Death
- War & Adventure
- Murder & Revenge
- The Future (Dystopias)
- Marriage
- Comedy
- Mystery & Suspense
- Children's Literature

- Critical Status
- Plot & Characters
- Unlikeable Characters
- Beautiful Writing
- Length & Pace
- Enjoyable & Interesting
- Re-Readable
- Literary Language (Quotations)
- Conversational & Slangy Language
- Description & Dialogue (Quotations)
- Gushing & Loving Language
- Talking & Speaking
- Non-English Reviews

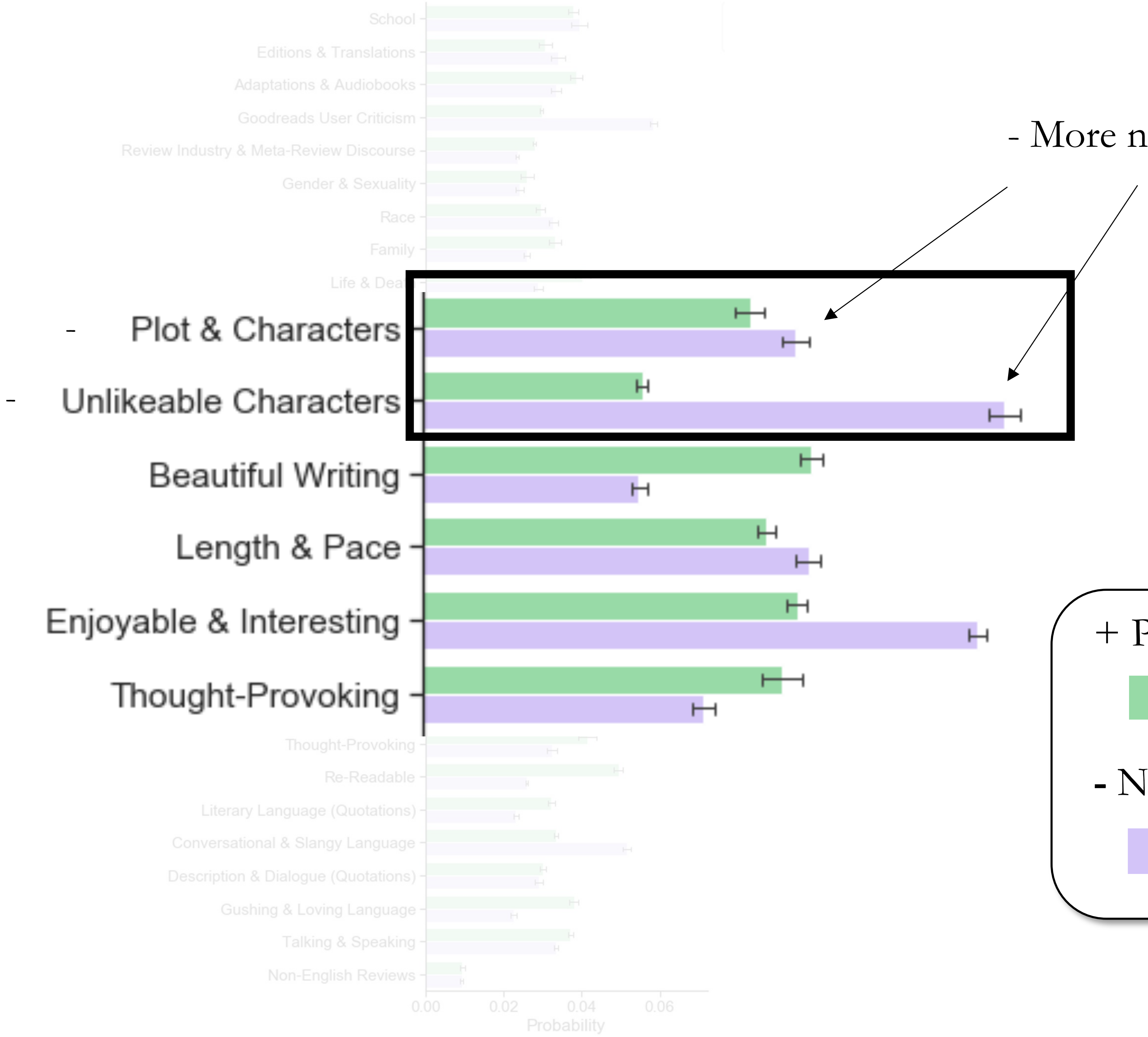


+ Positive Ratings (4-5 stars)

🟩

- Negative Ratings (1-3 stars)

🟪



- More negative ratings

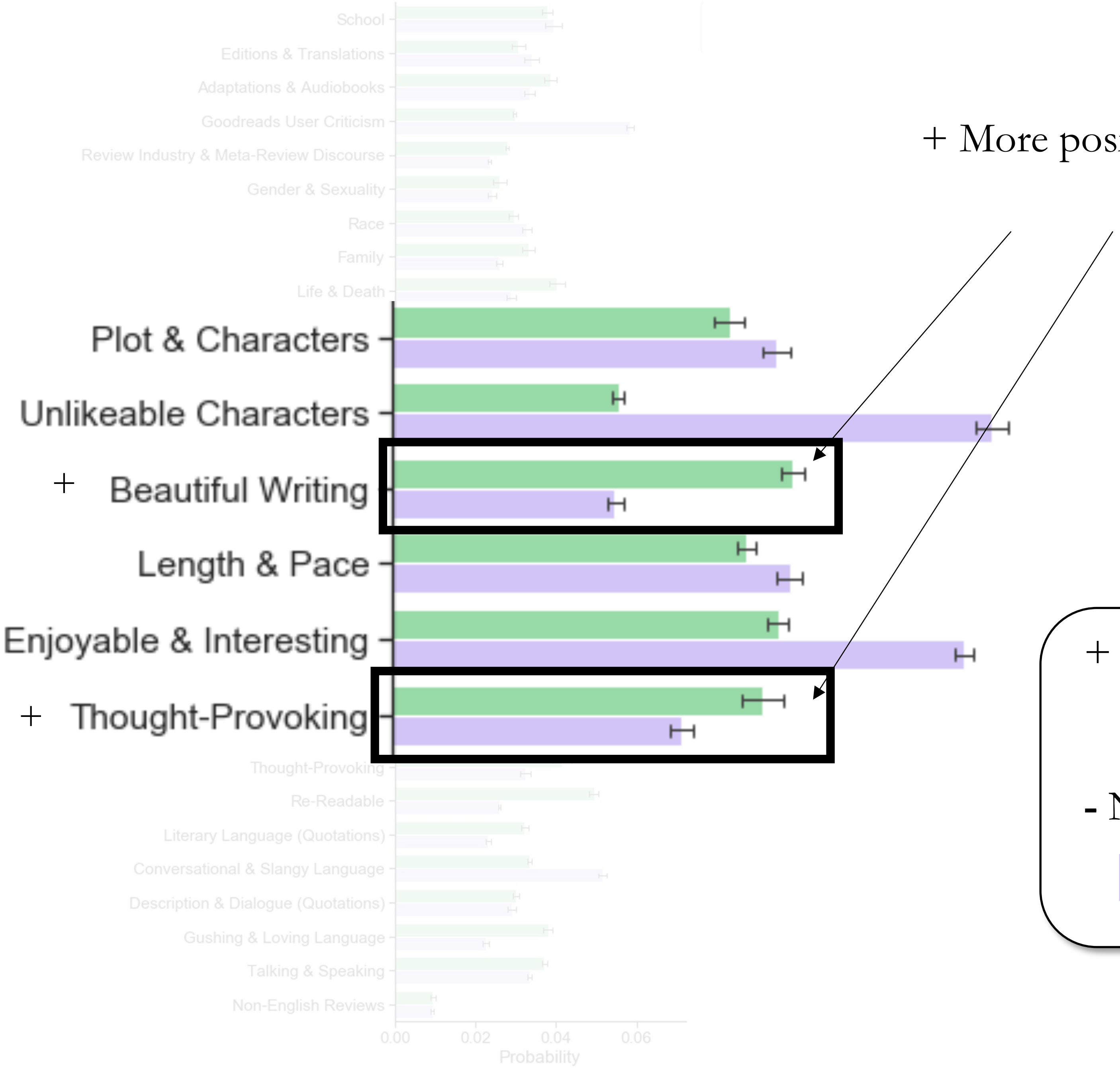
+ Positive Ratings (4-5 stars)

🟩

- Negative Ratings (1-3 stars)

🟪

+ More positive ratings



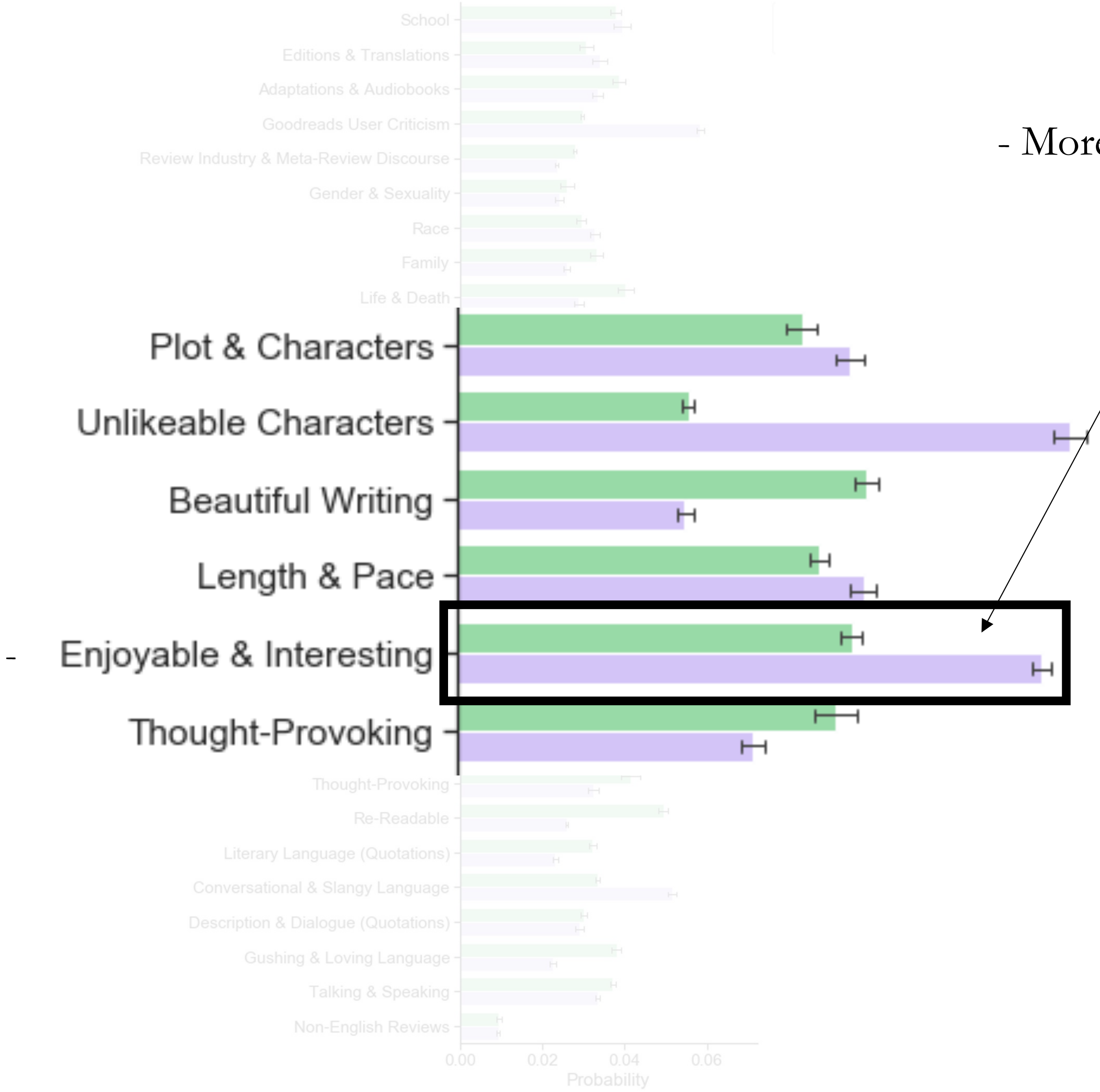
+ Positive Ratings (4-5 stars)

🟩

- Negative Ratings (1-3 stars)

🟪

- More negative ratings?



+ Positive Ratings (4-5 stars) 🟩

- Negative Ratings (1-3 stars) 🟪

Unlikeable Characters

characters, character, didn, felt, could, found, writing, plot, good, boring, feel, get, main, nothing

"A flat, disappointing story populated with one-dimensional characters that I grew to dislike. The narrator is a whiny young man prone to fainting and with very little backbones… it's also, frankly, boring"

-Rachel (Kalanadi)

Humanities ⟷ Data Science

- Data and computational methods offer new window into how culture circulates among “everyday” people

- Fully understanding this data requires historical context and analog research

- **Focus on individual examples can raise questions about ethics — user privacy, consent, surveillance, etc.**

User Ethics & Privacy

- Though Twitter and Goodreads data is public, many users have an expectation of privacy

- Exposing users' posts to an unexpected audience can have harmful consequences

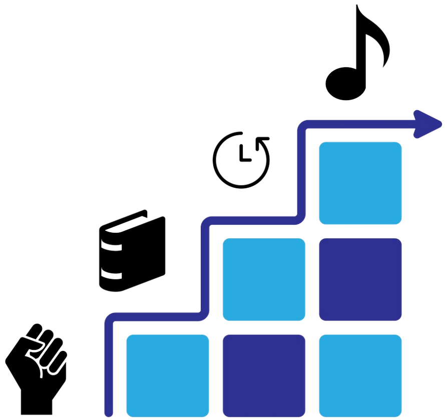
- We sought permission to quote and attribute posts in published research

To analyze complex data, students need…

- Technical skills
  - Maybe Python, R, JavaScript
  - Maybe pandas or dplyr, packages for working with data
- Critical reading, thinking, and research skills
  - Biography of the data
    - Where did this data come from? Who collected it? How was it collected? Why was it collected?
  - Ethical considerations
    - Is it ethical to collect and/or analyze this data?
    - What are the consequences of this data in the world?

Humanities data can help students develop critical data science skills because humanistic data is often:

- Messy, ambiguous, hard to reduce
- Requires historical, social, and human context to understand
- Fun and engaging for students



Lean into your liberal arts educations!

- Technical skills
  - Maybe Python, R, JavaScript
  - Maybe pandas or dplyr, packages for working with data
- Critical reading, thinking, and research skills
  - Biography of the data
    - Where did this data come from? Who collected it? How was it collected? Why was it collected?
  - Ethical considerations
    - Is it ethical to collect and/or analyze this data?
    - What are the consequences of this data in the world?

What is data science?
Who counts as a data scientist?

“This is due to our capacious definition of data science, one that seeks to **include rather than exclude and does not erect barriers based on formal credentials, professional affiliation, size of data, complexity of technical methods, or other external markers of expertise.**

Such markers, after all, have long been used to prevent women from fully engaging in any number of professional fields, even as those fields—which include data science and computer science, among many others—were largely built on the knowledge that women were required to teach themselves.”

-Catherine D’Ignazio and Lauren Klein, *Data Feminism*

Thank you!

Dr. Melanie Walsh // Assistant Teaching Professor //melwalsh@uw.edu

Women in Data Science Conference

April 1, 2022

Information School
UNIVERSITY of WASHINGTON