To Measure Meaning
(in Big Data)

don’t give me a map,
give me transparency and reproducibility

Laura K. Nelson
Northeastern University

VocalVault
Bill N. & Tyne
1-800-868-3404

Text as Data (also works with images)

- Classification
- Clustering
- Feature Selection

Text as Data (also works with images)

- Classification
  - Words
  - Texts
- Clustering
  - Words
  - Texts
- Feature Selection
  - Words / n-grams
  - A bunch of other stuff

Approaches to Text Analysis

- Deductive / Theory Testing
  - Lexical-based
  - Supervised Machine Learning / Deep Learning / Neural Networks
  - Word Embeddings
- Inductive / Exploratory
  - Lexical-based
  - Unsupervised Machine Learning
    - topic modeling, clustering, word embeddings
  - Deep Learning / Neural Networks (maybe?)

Approaches to Text Analysis

- Deductive / Theory Testing
  - Lexical-based
  - Supervised Machine Learning / Deep Learning / Neural Networks
  - Word Embeddings
- Inductive / Exploratory
  - Lexical-based
  - Unsupervised Machine Learning
    - topic modeling, clustering, word embeddings
  - Deep Learning / Neural Networks (maybe?)

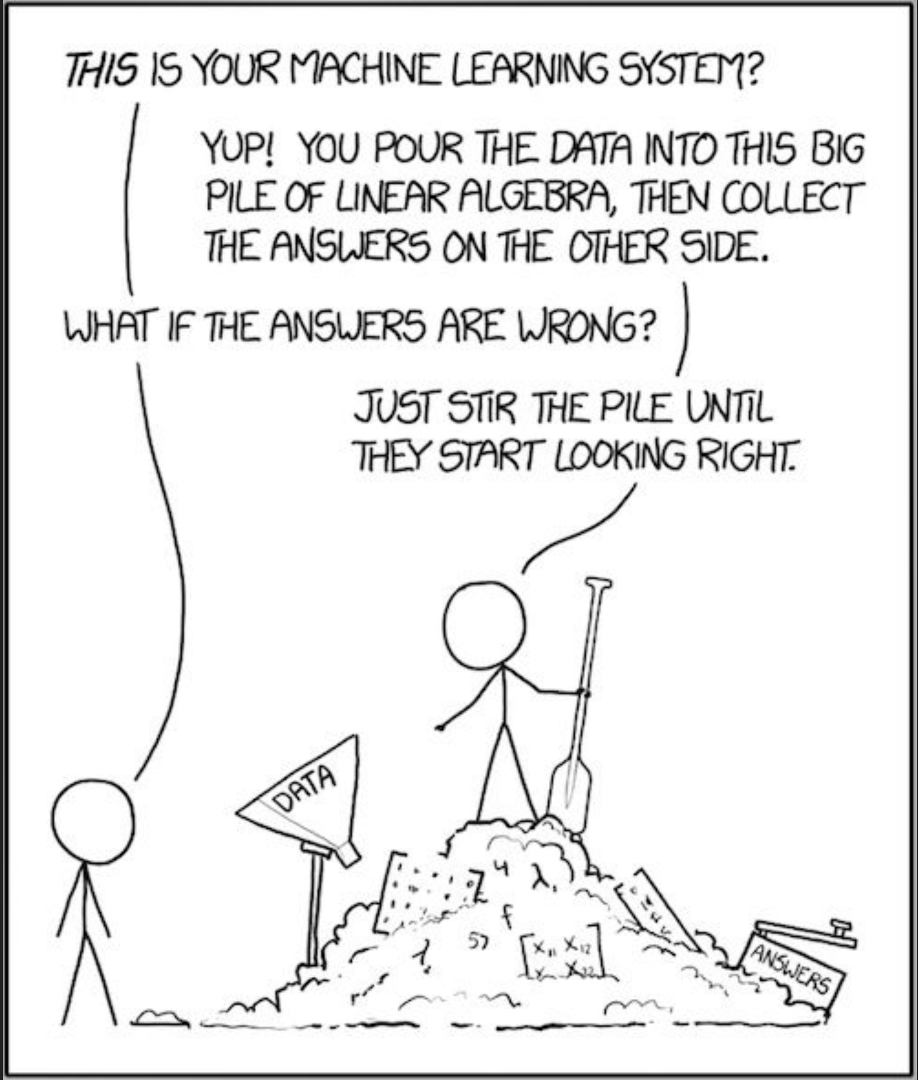
Inductive / Exploratory

What is the discourse* related to X?

How has the discourse around X changed?

How and why is the discourse around X different for groups Y and Z?

*discourse, frames, etc.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

computation (statistics) == GOOD

human judgement == BAD

When it comes to formal analyses, we might say that bad sociologists code, and good sociologists count. The reason is that the former disguises the interpretation and moves it backstage, while the latter delays the interpretation, and then presents the reader with the same data on which to make an interpretation that the researcher herself uses.

Lee, Monica and John Levi Martin. 2015. “Coding, Counting and Cultural Cartography.” *American Journal of Cultural Sociology* 3 (1): 1-33.
https://doi.org/10.1057/ajcs.2014.13

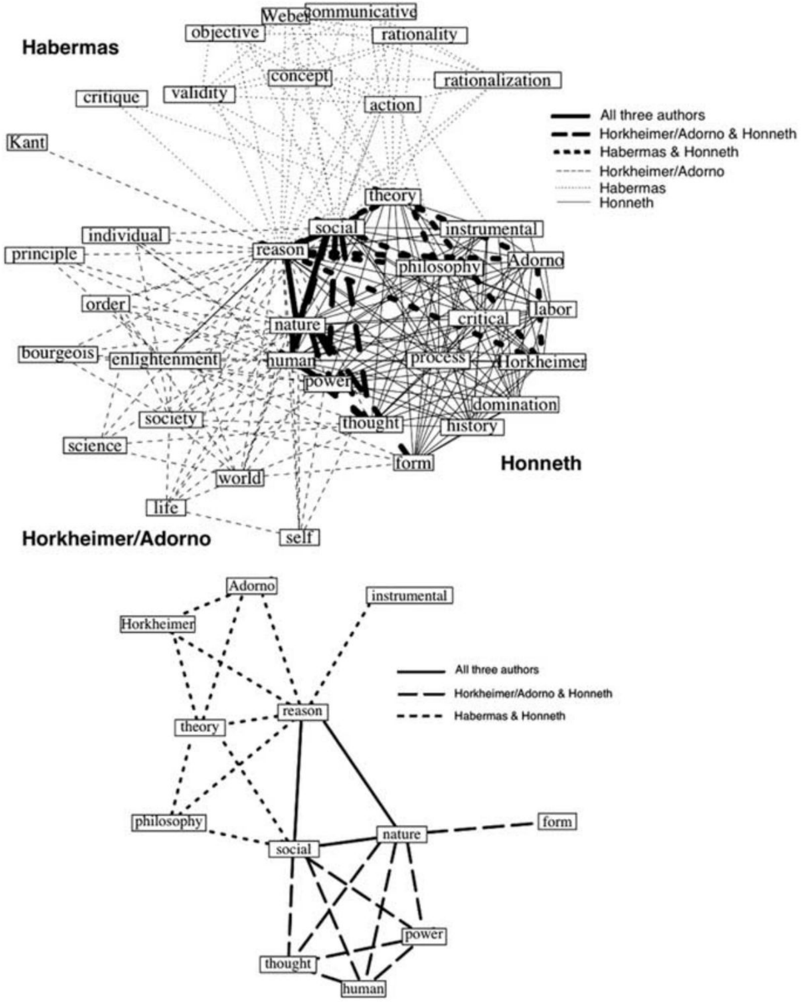
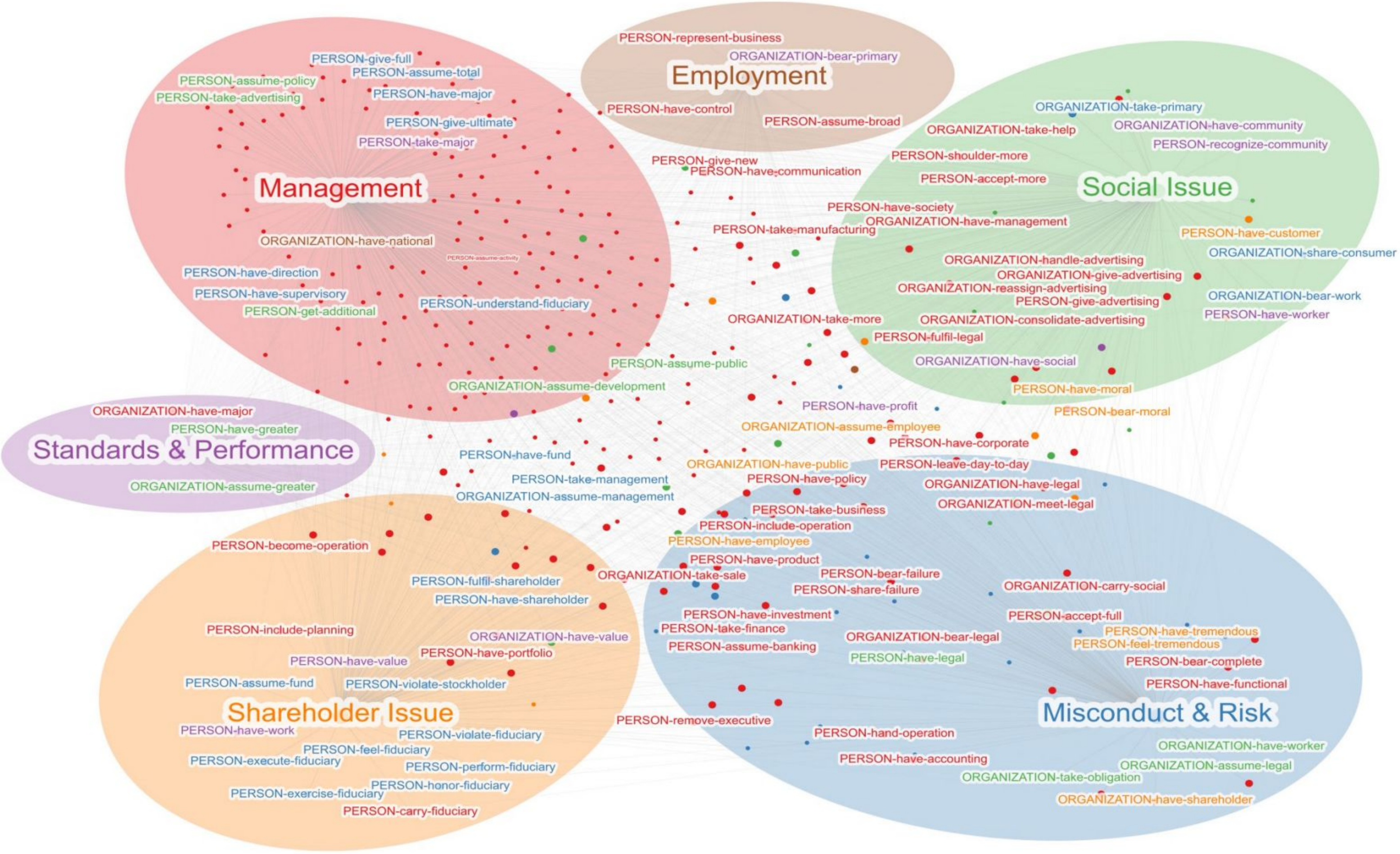


Figure 3: Overlaps of concept maps of three generations.

Ex-ante interpretations are problematic because they involve the necessarily subjectively driven exclusion of linguistic units or the grouping of particularities into labeled categories beyond the observer’s sight.

Goldenstein, Jan, and Philipp Poschmann. 2019. “Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling.” *Sociological Methodology*. Online First. doi:10.1177/0081175019852762.

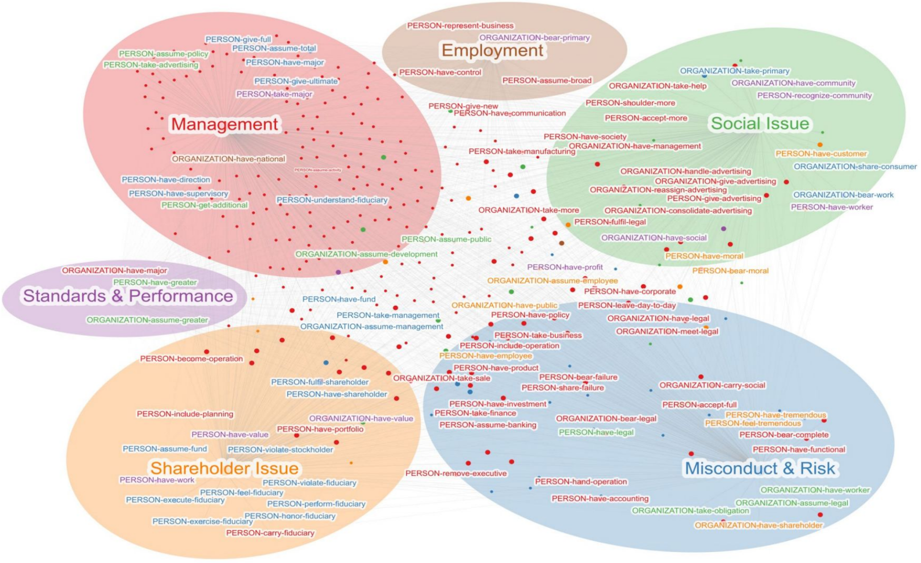
Figure 6. Map 2: Semantic associations of triplets in era 3 (large nodes and nodes with labels changed their semantic association over time).



Ex-ante interpretations are problematic because they involve the necessarily subjectively driven exclusion of linguistic units or the grouping of particularities into labeled categories **beyond the observer’s sight**.

Goldenstein, Jan, and Philipp Poschmann. 2019. “Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling.” *Sociological Methodology*. Online First. doi:10.1177/0081175019852762.

Figure 6. Map 2: Semantic associations of triplets in era 3 (large nodes and nodes with labels changed their semantic association over time).



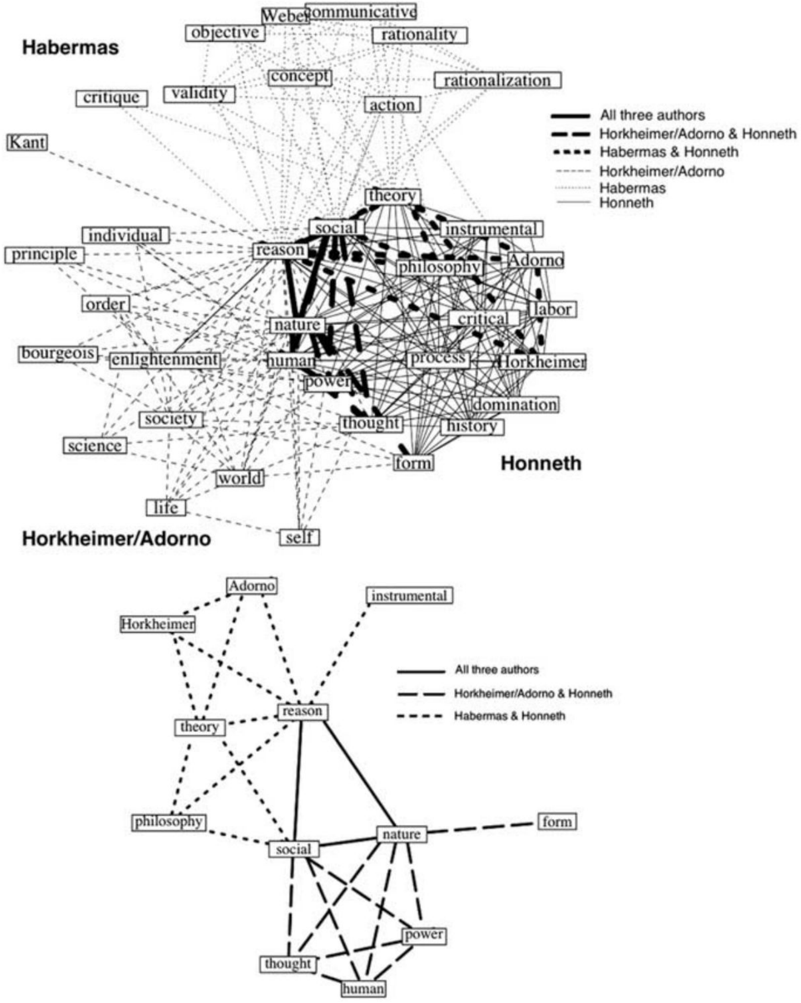


Figure 3: Overlaps of concept maps of three generations.

- Constrain the “semantic surrounding” to the paragraph in which their chosen keywords occurred
- Include only adjectives and nouns (and excluding proper nouns) in the text used to construct their topic model
- Exclude a full 38 of the 70 semantic patterns they estimated and pool the resulting 32 topics into six semantic groups
- Label the six semantic groups with their own subjectively chosen phrases
- Use the number of unique semantic triplets (rather than frequency) per main era (era defined through yet another ex-ante choice of clustering cutoff) as the relevant textual characteristic of their data

Belies an ontological truth about text as data

Belies an ontological truth about text as data

Text conveys a vast amount of information, much of it ambiguous and only some of which is relevant for a research question or purpose.

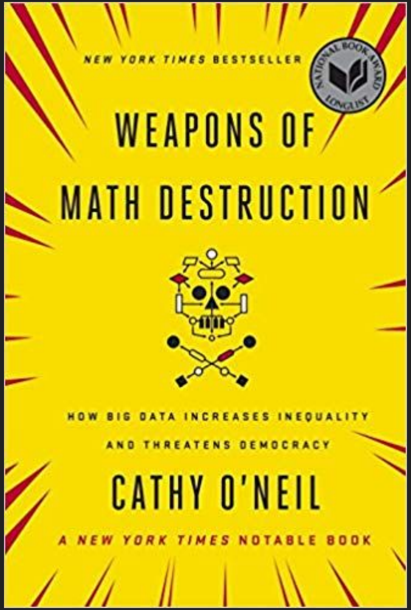
Belies an ontological truth about text as data

Text conveys a vast amount of information, much of it ambiguous and only some of which is relevant for a research question or purpose.

It is simply impossible to represent text in a way truly absent any **subjective** decisions and have those representations be **analytically useful** or meaningful.

“Pure” Representation

Transparency and Replicability



NEW YORK TIMES BESTSELLER

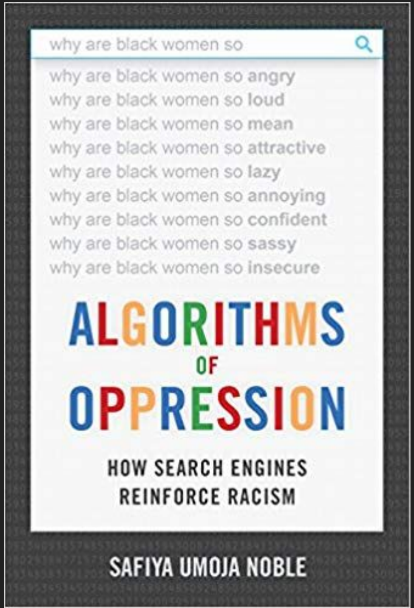
NATIONAL BOOK AWARD LONGLIST

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

A NEW YORK TIMES NOTABLE BOOK



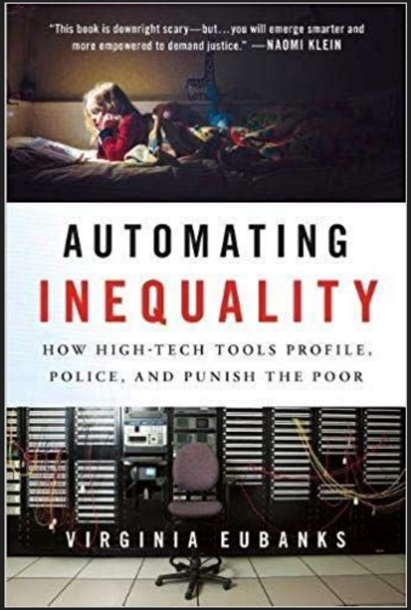
why are black women so

why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

SAFIYA UMOJA NOBLE

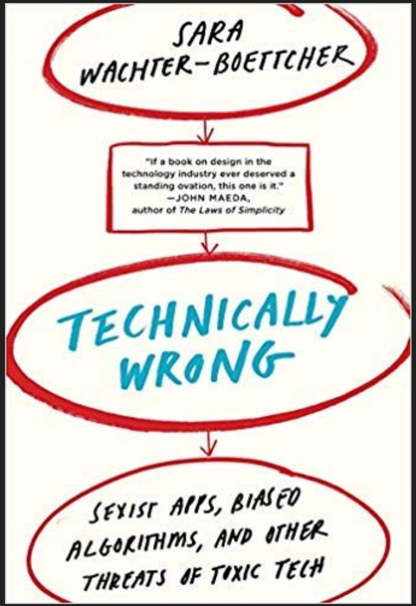


"This book is downright scary—but… you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN

AUTOMATING INEQUALITY

HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR

VIRGINIA EUBANKS



SARA WACHTER-BOETTCHER

"If a book on design in the technology industry ever deserved a standing ovation, this one is it."
—JOHN MAEDA,
author of *The Laws of Simplicity*

TECHNICALLY WRONG

SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH

Information Extraction

Is the information extracted from the text the most relevant information to the social process/concept/question?

Were the techniques (computational or otherwise) used to extract this information the most accurate techniques available?

Is the method used the most transparent and replicable available?

Within reason, if the authors altered linguistic key decision points, would they extract the **same information** from the text?

Is the authors’ interpretation reproducible?

Guidelines

1. Is the author extracting the most relevant information?
2. Are the methods the most accurate available?
3. Are the methods transparent and replicable?
4. Is the conclusion robust to sensitivity checks?
5. Is the interpretation reproducible?

Text as Data for Inductive Analysis

Text as Data for Inductive Analysis

✓ Information extraction tools

Text as Data for Inductive Analysis

✓ Information extraction tools

✓ Embrace and acknowledge researcher degrees of freedom

Text as Data for Inductive Analysis

- ✓ Information extraction tools
- ✓ Embrace and acknowledge researcher degrees of freedom
- ✓ Reproducible interpretation

References

- Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press.
- Goldenstein, Jan, and Philipp Poschmann. 2019. “Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling.” *Sociological Methodology*. Online First. https://doi.org/10.1177/0081175019852762.
- Lee, Monica, and John Levi Martin. 2015. “Coding, Counting and Cultural Cartography.” *American Journal of Cultural Sociology* 3 (1): 1–33. https://doi.org/10.1057/ajcs.2014.13.
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.
- O’Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
- Wachter-Boettcher, Sara. 2018. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*. W. W. Norton & Company.