- Introduction
  - Due to federal regulations, cannot collect race and ethnicity information when taking applications for credit and other financial products outside of the mortgage, yet government regulators are charged with enforcing fair lending laws.
  - Assessment of potential fair lending violations requires a statistical analysis of loan level micro-data.
  - This has two general consequences for when researchers want to analyze fair lending outcomes
    - 1.) How do we accurately measure race and ethnicity?
    - 2.) How do we use these estimates to measure fair lending outcomes?
- Background
  - Measuring race and ethnicity
    - Might want to mention something about what race/ethnicity means conceptually, maybe reporting in Census or something. I think demographers there have papers out about this.
    - Might want to discuss reporting of race/ethnicity in HMDA, emphasizing self-reporting, reporting by loan officers, non-reporting.
    - Bertrand and Mullainathan (first names, Massachusetts Birth records)
    - Might want to mention DMV records...
    - Audit studies (Yinger; Ayres, etc.)   **Comment [BES1]:** Is this really relevant here?
    - Elliott et. al.
      - Identify the need/desire to compare health outcomes on the basis of race/ethnicity using administrative healthcare data that often do not record patient race/ethnicity (this establishes the commonality)
      - Recognize common use of proxies for race/ethnicity based on the distributions of individuals (in population) across race/ethnicity based on surname and geography.
      - Propose using Bayesian updating methodology on commonly available datasets, which leads to an overall improvement in accuracy/performance of proxy relative to the use of a surname only or geography only proxy
      - Compare estimates to reported race/ethnicity info using dataset of medical insurance holders
    - Population matching   **Comment [BES2]:** Might want to move this up and conclude this section with the Elliott et al. paper...which naturally leads to the next section.
      - Using data on distribution of users of product to update estimates
      - Somewhat more complicated, not sure how much this improves estimates
      - Discuss paper using Myspace/Facebook data
  - Outcome estimation
    - McCaffery and Elliott discusses differences in usage of race/ethnicity measures when testing outcomes
      - Direct use   **Formatted**
      - Assignment (classification)
      - PIML
    - Discuss use of population average as a consistent estimator
    - Geronimus, Bound, and Neidert discuss role of omitted variable bias

- Discuss how Disparate Impact analysis specifically concerns <u>differential</u> OVB between reported value and proxy
- Data
  - Proxy Construction
    - US Census Records
    - Social Security Administration
  - Validation <u>exercise performed on data where race/ethnicity/sex is reported</u>
    - HMDA+
- Estimation of Race
  - Proxy Construction
    - Methodology
      - Could be expanded to examine additional info, such as multiple last names, intermarriage, income
  - Accuracy of Estimates
    - Show table of correlations
    - What to think about how much this measure is able to provide signal of race relative to noise. We can examine using ROCs
    - Provide AUC of ROCs to test for ability to serve as significant predictor of race
      - Do for Joint proxy v. geo or name
      - Discuss how this relates to use of threshold rule
- Estimates of Outcomes using Proxy Measures of Race/Ethnicity
  - Potential Concerns
    - Omitted Variables Bias
      - Show simulated examples w/ & w/o OVB
    - <u>Measurement error? (If not, remind me why)</u>
    - <u>Timeliness of data? (Not sure…but it is a potential concern going forward)</u>
    - Validity of out-of-sample proxies
      - Discuss potential impacts of non-alignment of population and sample distributions on estimates
      - Could using income correct for this?
  - Actual Outcomes Using Outside Data – Same HMDA + dataset
    - Omitted Variable Bias – Show results for house price regression
    - Unbiased estimates – which GSE do GSE-quality loans get sent to?
      - Overall results are biased, subset of loans with tighter underwriting (hence, fewer "unobservables") closely match reported results.
- Conclusion