# Replicability and Reproducibility in Crowdsourcing & Social Media

Panos Ipeirotis

Stern School of Business
New York University

# Two Main Types of Research

- Get data sets from a company
  - Usually contain user behavior information
  - NDA's, IP agreements, and other strings attached
  - Example: Travelocity browsing and purchase data
- Generate data sets through crowdsourcing
  - Get humans by paying (e.g., Mechanical Turk)
  - Get humans through social media, email, etc.
  - Example: Hire humans to participate in quizzes

# Replicability with Corporate Data: Hard

- Impossible to share corporate data publicly
  - Even if possible, privacy snafus prevent most companies from even allowing public disclosure of their user's data

- Negatives:
  - Unclear how to deal with this issue
  - If replicability impossible, is it even research?
- Positive:
  - Should stop taking too seriously N=1 experiments
  - ***Reproducibility*** only way forward

# Replicability with Collected Behavioral Data: Easier

- When hiring/engaging humans, possible to share data
  - Need to pass IRB (often easy, sometimes a pain)
  - Anonymization not that hard
- Can simply post data on a web page
  - But this allows only for statistical double-checking
- Can also share code for replicating experimental setting
  - **Costly** to run again the experiment
    - Thousands of dollars needed to pay participants, or to run advertising campaigns on the web
  - In reality, we want to reproduce in **similar**, not identical settings
    - If result not robust to small perturbations, not a result
    - It is a resource waste to try to **replicate**

# The End

- Experiments carried our in my field
- What is reproducible in them?
- Which tools are used?
- Why (or why not) reproducibility would be desirable?
- What are the existing barriers (if any) to reproducibility?
- Existing reproducibility efforts in her/his field, if any