# ECPR Methods Summer School: Big Data Analysis in the Social Sciences

**Pablo Barberá**

School of International Relations
University of Southern California
pablobarbera.com

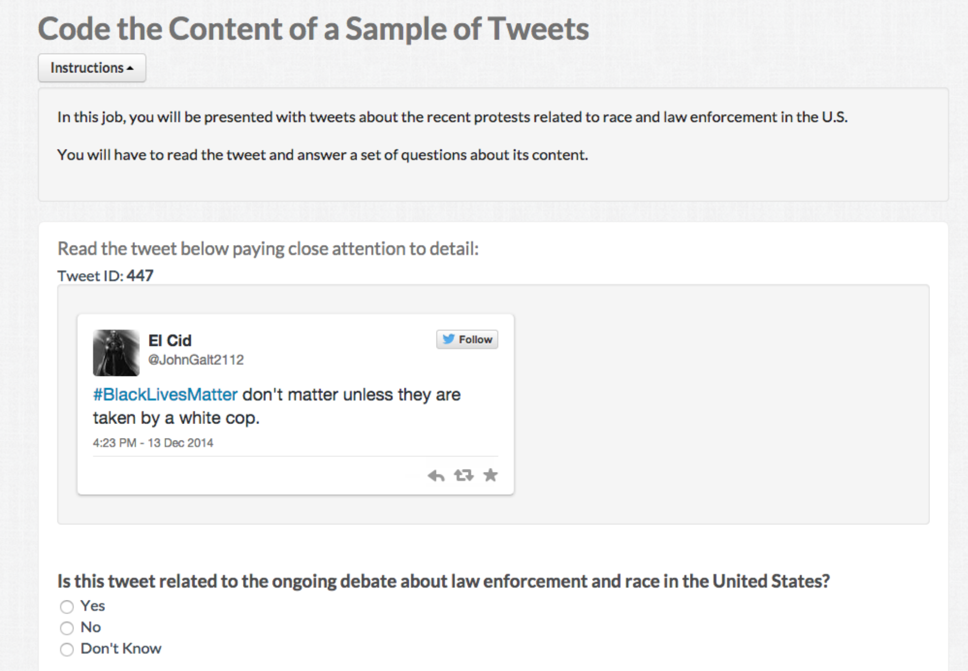
Networked Democracy Lab
www.netdem.org

Course website:

github.com/pablobarbera/ECPR-SC104

How do I create my own training dataset for supervised learning?

CrowdFlower



# Code the Content of a Sample of Tweets

In this job, you will be presented with tweets about the recent protests related to race and law enforcement in the U.S.

You will have to read the tweet and answer a set of questions about its content.

Read the tweet below paying close attention to detail:

Tweet ID: **447**

**El Cid**
@JohnGalt2112

#BlackLivesMatter don't matter unless they are taken by a white cop.

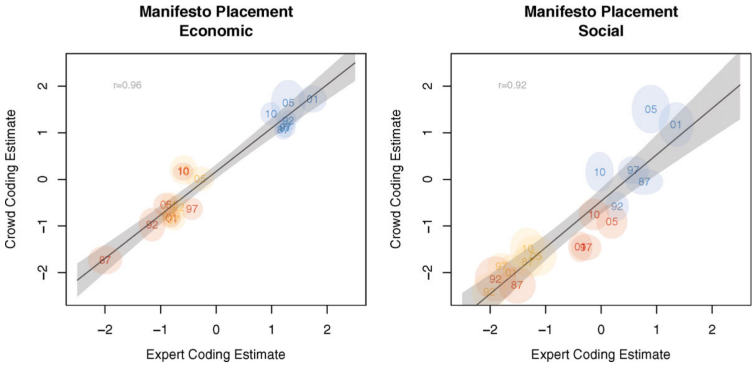
4:23 PM - 13 Dec 2014

**Is this tweet related to the ongoing debate about law enforcement and race in the United States?**

○ Yes
○ No
○ Don't Know

# Crowd-sourced text analysis (Benoit et al, 2016 APSR)

**FIGURE 3. Expert and Crowd-sourced Estimates of Economic and Social Policy Positions**

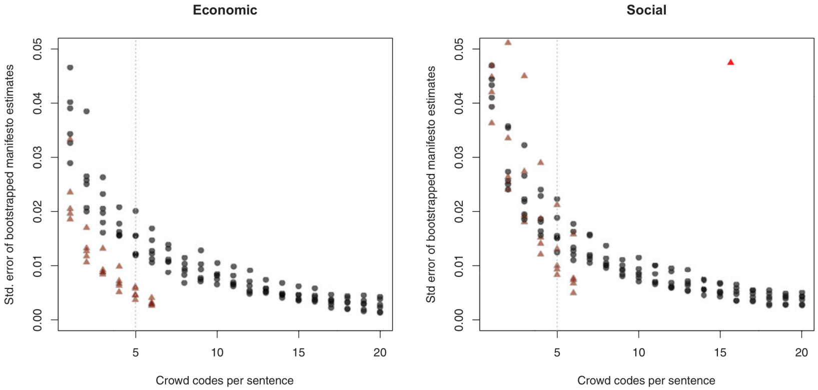


Manifesto Placement Economic

Manifesto Placement Social

# Crowd-sourced text analysis (Benoit et al, 2016 APSR)

**FIGURE 5. Standard Errors of Manifesto-level Policy Estimates as a Function of the Number of Workers, for the Oversampled 1987 and 1997 Manifestos**



*Note:* Each point is the bootstrapped standard deviation of the mean of means aggregate manifesto scores, computed from sentence-level random n subsamples from the codes.

# Demo outline

Steps:

1. Clean and save dataset of tweets
2. Import data to Crowdflower and prepare task
3. Launch task to coders (internal or external)
4. Export coded data and open in R

# Preparing data for Crowdflower

Script: `day2/04-crowdsourcing.Rmd`

1. Read a sample of tweets in JSON format
2. Take random sample of 100 tweets
3. Prepare tweet to be embedded on html format
4. Export dataset as .csv file for crowdflower

# Crowdflower instructions

Creating a new task:

1. Create a “Data for Everyone” account on CrowdFlower
2. Log in as “customer”
3. “Create job” and choose “start from scratch”

   (but many useful templates to get you started)
4. Upload file with data
5. “Build Job”
   - Graphical editor enough for most scenarios
   - Here we use CML editor to add embedded tweets
   - CML code: `crowdflower-task-code.txt`
6. Create test questions
   - Coders will need to get 80% right
7. Modify “Settings”
   - Number of coders (“Judgments per Row”)
   - Tweets per page (“Rows per Page”)

# Task settings

CrowdFlower offers options to:

- Use internal and/or external coders
- Select coders from specific countries or that specific certain languages
- Different tiers (levels of quality) for coders
- Dynamic judgments mode (advanced)

# Exporting data from Crowdflower

Five types of data:

1. Full: individual codings, one per line
2. **Aggregated:** “best” codings for each tweet
3. Source: original data (and test questions)
4. Contributor: statistics for each coder
5. Json: full data in json format

Importing file:

- `R`: `read.csv` function