


# Twitter De-Identification

Jonathon Storrick

Jon.Storrick@gmail.com



Center for Computational Analysis of  
Social and Organizational Systems  
<http://www.casos.cs.cmu.edu/>



## Why It's Necessary

Facebook says Cambridge Analytica may have had data on 87 million people

by Heather Kelly @heatherkelly  
April 4, 2018, 8:57 PM ET



June 2020

2

Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## Why It's Necessary

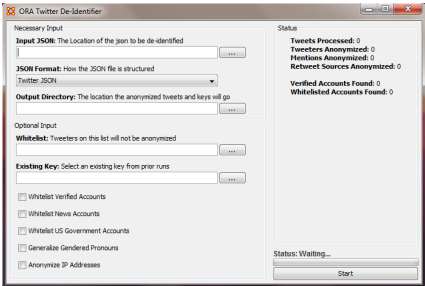
- After the Cambridge Analytica scandal, there is MASSIVE concern for how data is stored.
- EU passes General Data Protection Regulation.
- Personally Identifiable Information.
- The more information we gather about a given individual, the more likely it is we'll be able to reverse engineer their real identity.
- That can cause issue with grants, data transfer, and may limit the amount of data you can collect for a given subject.
- Because Twitter said it is

CASOS  
June 2020

Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## The Solution

- We developed the Twitter De-Identifier, a standalone tool for processing Twitter data.
- Reduces PII, handles large datasets, and removes only superfluous information
- For information on how to access the De-Identifier, please email Dr. Carley



CASOS  
June 2020



Carnegie Mellon  
IST Institute for Software Research

## De-Identifier: the Challenges

- While a typical tweet is limited to 280 characters (mostly), an individual tweet has 10-20x as much info associated with it. Each tweet would need to be carefully handled such that no user could take a De-ID tweet and find its source.
- A record of the anonymization needs to be kept, in case project heads absolutely need it, and to keep consistent anonymizations across multiple datasets.
- Speed. A twitter dataset can contain millions of tweets.
- Not removing data that is of analytic use

CASOS  
June 2020

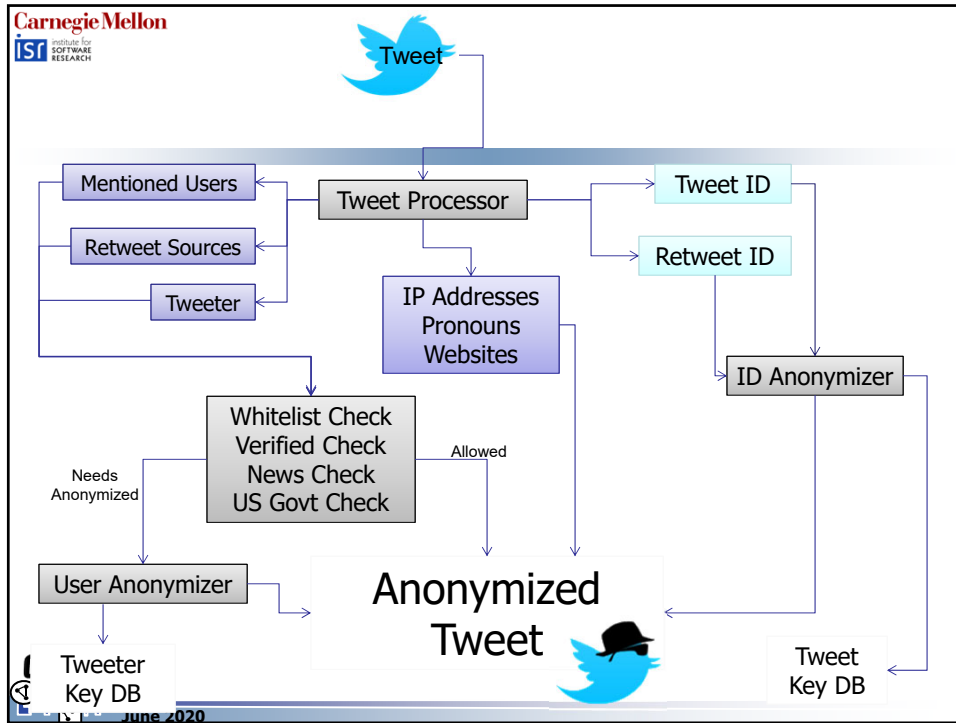
Carnegie Mellon  
IST Institute for Software Research

## The Approach

- Direct Identifiers: Tweet ID's, Tweet Usernames, Mentions
- Indirect Identifiers: User Profiles, Locations, Dates
- Masking: Should something need to be anonymized, its relevant portion is replaced by pseudo-random text
- Recognizing data that doesn't need to be anonymized.
  - News reports, verified individuals, etc.

CASOS  
June 2020





**Carnegie Mellon**  
IST Institute for SOFTWARE RESEARCH

## Operation Speed

- The primary bottleneck – read/write speed. Twitter data is far too large to fit entirely in Memory.
- Even then, it can process 20k per minute with a typical non-SSD hard drive.

**CASOS**  
June 2020



Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## Demo

CASOS  
June 2020

Carnegie Mellon  
IST Institute for SOFTWARE RESEARCH

## Summary of Features

- It must be capable of importing a tweet in Json format, and exporting a de-identified tweet in the exact same format.
- It must remove as much personally-identifiable information as possible, without removing information important to analysis.
- Users must have options in what gets anonymized. If they want to leave certain users or agencies un-anonymized, they should be able to.
- De-Identified ID's should be carried throughout the process. If a tweeter is "00001" in one place, he should be "00001" in every other place.
- A lookup table for tweets and users should be output to allow for looking into specific agents or to keep De-Identified ID's the same across multiple runs.

CASOS  
June 2020