Here Be Dragons (authors)

1. Problem
How can we make self-selected social media populations be representative of the population in a region?

Solution: Bin users into demographics and then poststratify (re-weight by bin) to get a representative sample.

Q1: How do we assign users to demographic groups (bins)?
Q2: How to effectively poststratify?

2. M3: Multimodal, multilingual, multi-attribute demographic inference
Given standard information from a user’s bio, like their name and profile image, the M3 system predicts gender, age, and whether the account belongs to an organization.

Data Augmentation: The M3 system needs also data augmentation, which is done by adding co-ranking on synthetic text and by adding new groups of features. The system is designed to operate on profiles written in any of the languages spoken in Europe and neighboring countries—32 in total.

3. Demonstration: Estimating population counts in Europe
How many people live in a particular region of Europe? Here, we test whether we can estimate population from the non-representative sample by using M3 to stratify by demographics and then using census information to learn inclusion probabilities for each population stratum, which lets us poststratify counts to get a more precise population count estimate.

Using three months of 10% samples of Twitter data, we train all users using M3 and estimate population counts for NUTS3 regions in Europe. We measure performance using the MAPE and leave-one-out cross-validation. Best models are compared to commercial systems for estimation, each using all the demographic information the models differ in how much demographic information they use and their structure. We find that inclusion probabilities help to reduce errors by 30%. Our best model substantially improves on MAPE. Further, each model learns the estimates, shown here as a lower state of the art performance.

MAPE. Further, each model learns the estimates, shown here as a lower state of the art performance.

4. Try it yourself!
Demo: http://www.m3inference.org/demos
Paper: https://doi.org/10.1145/3308558.3313684
M3 library: pip install m3inference
Inclusion Probos: http://www.m3inference.org/inclusionprobos
Auto-Poststratifier: coming soon!