# The Community Labeling and Sharing of Security and Networking Test datasets (CLASSNET)

Jelena Mirkovic (USC/ISI), John Heidemann (USC/ISI) ,
Wes Hardaker (USC/ISI), Michalis Kallitsis (Merit Networks, Inc)

# Goals of the Project

- Share interesting datasets for networking and cybersecurity data science
  - Real traffic from real networks
  - Both ongoing and curated datasets
  - Labeled, thus suitable for ML workflows
- Handle privacy issues via mix of technical and human approaches
  - Anonymization, restricted access, DUAs, user vetting
- Enable community discussion around datasets and labels
  - Collaborative labeling

# Privacy Handling and Data Access

- Each data provider hosts their own datasets, and gives us metadata and DUAs (similar to IMPACT)
  - Researchers find datasets on our COMUNDA portal and request them by signing a DUA
  - For non-anonymized datasets containing legitimate traffic we also require that the researcher has IRB approval
- We vet researchers in agreement with data provider
  - Fast track: US/EU researchers, published research
  - Slower track: non-US/EU researchers, no publications
- Researchers receive an email detailing how to access datasets:
  - Download, onsite, in cloud

# Labeling Approach

- Datasets initially labeled by provider
  - Labels can be per record (e.g., list of scanners) or per event (e.g., start and stop time of an attack, attack target, attack type)
  - Labels can be program-generated
    - Run this program with this input to produce labels for each record in the dataset
- Researchers that requested the given dataset can
  - Suggest label corrections
  - Produce and submit their own set of labels
- This addresses elusive ground truth in labeling
  - E.g., “our system has 99% accuracy on Mao-Franz labels”

# Participate and Contribute

- COMUNDA portal https://comunda.isi.edu
  - Search for datasets using keywords or just browse
  - Request a dataset
  - Rate and comment on a dataset (later there will be threaded conversation feature)
- Contribute your labels or correct existing labels
  - Documentation is at COMUNDA site
- Consider contributing your datasets
- Send us feedback about the portal
  - It's fairly new so there may be glitches
- Direct feedback: sunshine@isi.edu