



# **IBSS: Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks**

## **Research questions and challenges at University of Arkansas (UARK)**

Xuan Shi  
GRA: Bowei Xue  
University of Arkansas



**Spatiotemporal Modeling of Human Dynamics Across Social Media and Social Networks  
Interdisciplinary Behavioral and Social Science Research, National Science Foundation**

## **UARK's goals in this project**

- **Develop high performance computing (HPC) solutions and tools to support:**
  - **Social media data pre-processing, transformation, and other tasks**
  - **Data mining and machine learning, e.g. spatial cluster and correlation analytics**
  - **Agent based modeling (ABM)**
  - **Visualization**

# Data preprocessing and transformation

- **Text retrieving**
  - Based on key words of interest to this project
- **Text cleaning**
  - Uncertain patterns or terminologies in social media [e.g. abbreviated terms, jargons, etc.]
- **String matching**
  - Lots of algorithms can be applied
- **Transformation and localization**
  - Key words replacement and standardization
  - Vocabulary unique in California

# Data mining and machine learning

- **High frequency word construction**
  - Time series of wildfire events in social media and network
  - Identification of duplicated/retweeted tweets
- **Supervised classification**
  - Training data development and validation
  - Similarity and difference
  - Exemplar identification through Affinity Propagation (AP)
  - Support Vector Machine (SVM) for classification



# HPC solutions

- **When large volume of data has to be processed, parallel and distributed computing solutions have to be developed, such as GPUs and MICs**
  - **In the case of AP, for example, 10K points need 4GB memory, 20K points need 16GB memory, and 40K points need 64GB memory**
  - **The time used for training SVM classifiers grows when the size of the training data increases.**
    - **The theoretical computation complexity of building a SVM classifier lies between  $O(n^2)$  to  $O(n^3)$**
  - **In the case of ABM, data communication may have to be executed multiple times in order to complete the computational processes thus increase the difficulty and challenge in development.**