Summary: Offers robust Self-Tuning ANomaly Detection service (STAND) that enables domain experts to detect anomalies from diverse data sets without requiring data science expertise nor tedious manual tuning.

Intellectual Merit:
A novel anomaly detection cyber-infrastructure, with several advantages over existing capabilities:
• It automatically detects anomalies with higher accuracy than what is achievable by extensive manual tuning;
• It can be applied across a wide range of data types and domains;
• It allows developers to plug in new anomaly algorithms with ease;
• It allows domain experts to steer the anomaly detection process using their domain expertise through feedback.

Data sets considered could come from diverse communities:
• Medicine and biology data for early and automatic detection of diseases. This includes the NSF Biological Sciences (BIO) and Smart and Connect Health (SCH) communities;
• Cybersecurity data for identifying signatures of cyberattacks. This includes the NSF Computer and Information Science and Engineering (CISE) and Electrical, Communications and Cyber Systems (ECCS) communities; and
• Manufacturing data for detecting device failures or predictive maintenance. This includes NSF Engineering (ENG) and Civil, Mechanical and Manufacturing Innovation (CMMI).

Figure: STAND System Architecture
STAND is built on a distributed infrastructure such as Spark to handle big datasets stored in a distributed file system like HDFS or S3.

Given an input data set, the self-tuning framework of STAND identifies a set of detection algorithms from the anomaly detection & feature engineering library, runs these different algorithms concurrently (instantiated with parameters), extracts a set of objects clearly identified as either anomalies or as normal objects from the combined detection results, and then uses this new type knowledge as pseudo-labels to train a binary classification model as an “anomaly classifier”.

The anomaly summarization and explanation component summarizes the results to explain them and sends them to the user.

The human-in-the-loop component accepts domain feedback from users to improve the self-tuning framework or allows users to plug in domain-specific anomaly detection techniques that fit their data well.