# CADRE Project - The Dilemma

- Libraries cannot provide researchers with **sustainable, standardized access** to licensed datasets for text & data mining
- It is cost-prohibitive for most individual libraries to **develop and implement infrastructure** to provide access to licensed big data sets and large or unwieldy open data sets
- Many researchers who could benefit from text and data mining library-acquired resources, lack programming skills and would only be able to do so via a **graphical user interface**

# CADRE Project - The Solution

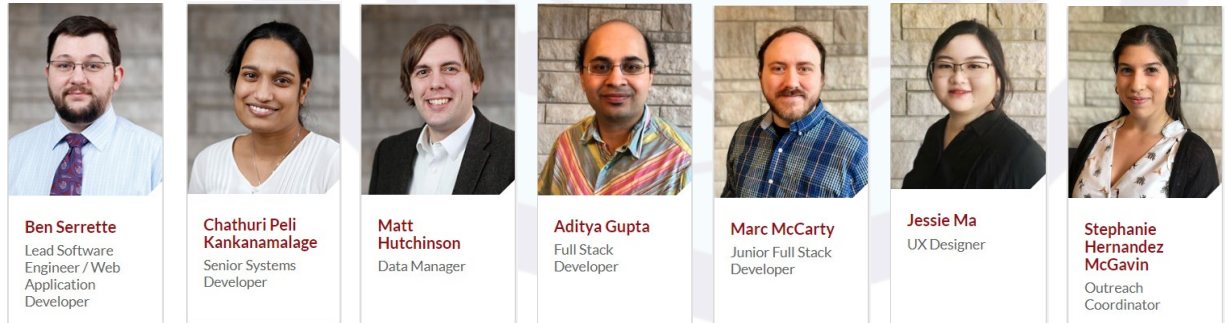
- CADRE is a cloud-based platform that will provide **secure access to library-licensed datasets** and open, non-consumptive datasets

- By sharing the cost of this solution across a large number of academic libraries, we are able to provide a **superior solution at a lower cost to members**, as well as a free service tier for non-members

- CADRE will feature a graphical **user interface**; standardized , multiple data formats; shared and custom **computational resources**; and a **space to share and store** queries, algorithms, derived data, results of analyses, workflows, and visualizations.

# **CADRE Project** - Indiana University Network Science Institute (IUNI)

- **IUNI** – http://iuni.iu.edu **a unique startup** in an established academic institution
  - A **cross-campus, transdisciplinary institute** that brings together faculty who engage in **network research** from various scientific fields
- **IUNI's mission**
  - To **strengthen the theories**, methods, analytic tools, and practice **of network science**, and to **foster collaborative, interdisciplinary network science approaches** to understanding and improving the complex challenges of our world
- **IUNI's Teams**
  - A team of IT professional
  - A team of research scientists



# CADRE Project – Goals
## Identify Constituents' Needs

- Understanding **users' needs** and expectations
  - User stories, Product Owner Council meetings, **Communication**
- **Informatics/computer science researchers** and labs
  - **APIs, Notebooks,** Access to **Raw Data** and **Cloud Compute** Resources
- **Science of Science** community
  - **Interface Access to Databases** and Cloud Native Technologies
- **Library and research** community outside of computer science
  - **Web Interface** guiding Query Building and **suggesting the most appropriate backend technology** on a case by case basis

# CADRE Project – Goals
## Research Asset Commons

- **Federated Login**
  - Access from **any affiliated institution** using single sign on ( CILogon, inCommon, Shibboleth etc.)
  - **Restricted access to proprietary resources**, based on login credentials
- **Collaboration**
  - Ability to **save and share** with specific users, community or the public **metadata, queries, results, annotations, visualizations, algorithms, code, containers and virtual machines**
  - Community building and collaboration based around **same data access privileges** and goals
- **Reproducibility, Replicability, Provenance and Transparency**
  - Use of the same, **well documented original datasets**
  - **DOIs** identifying any and every data change or permutation
  - **Saved and shared** workflows a pipelines trough **Packages and Containers**
  - **Ability to publish using unique identifiers** leading back to Research Asset Commons

# CADRE Project – Goals
## Identify the Proper Technology

- **Raw data access**
  - Access to **XML, JSON, CSV** etc. files in their native form. **Containerized** tools and packages
  - Access to data using **cloud native technologies** like **U-SQL and Athena/Glue**
  - Access to cloud **distributed computing** using **Databricks** and **SPARK** on **HDInsight** and **EMR**
- **Database access**
  - Researching on currently available **cloud and serverless Relational Database** implementations for each dataset and query type
  - Researching on currently available **Graph Database** implementations for each dataset and query type. Currently comparing **Neo4j, Tiger Graph, AgensGraph, cloud native and in-memory** alternatives
- **Web interface**
  - **Guided Query Building**
  - Ability to **suggest the most appropriate technology** on a case by case basis
  - **User control** over execution and use of resources

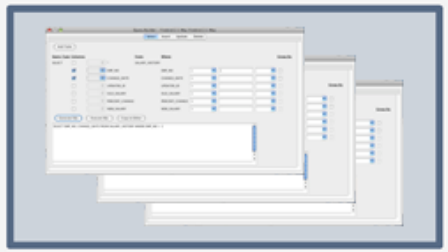
# SHARED BIGDATA-GATEWAY FOR RESEARCH LIBRARIES (SBD-GATEWAY)

## AUTHENTICATION

FEDERATED SINGLE SIGN ON

- Custom
- InCommon
- Shibboleth
- Globus

## WEB QUERY INTERFACE



## RESEARCH ASSET COMMONS

Saved and Shared Results, Tools and Datasets

- Visualizations
- Annotations
- Data Derivatives
- Metadata

## ACCESS BY PROGRAMMING AND VISUALIZATION TOOLS


## APPLICATION PROGRAMMING INTERFACE (API)

## CLOUD COMPUTE GATEWAY

### Machine Learning

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

### Text Analytics

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

### Database Analytics

- Analytics Tool
- Analytics Tool
- Analytics Tool
- New Analytics Tool

### Data Processing and Querying


Future processes

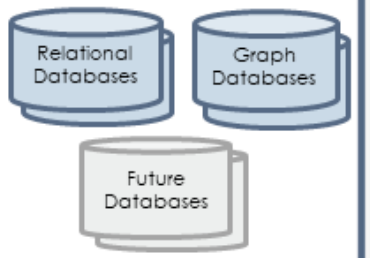
## CLOUD AND LOCAL COMPUTE RESOURCES


## AUTHORIZATION

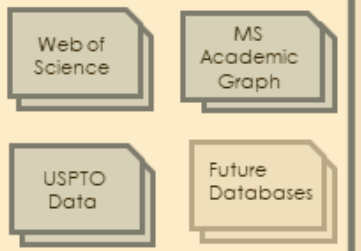
GRANULAR DATASET PERMISSIONS

## CLOUD STORAGE

### Indexed Storage




### Raw Format Data Lake Storage




## LOCAL AND INSTITUTIONAL STORAGE