











## LinkIt: Making Sense of Disparate Data

Members of the team:

 <b>Ethan Garnica</b>	 <b>Andrew Liddle</b>	 <b>Alejandro Ramirez</b>	 <b>Prerana Kharat</b>
 <b>Sayali Badole</b>	 <b>Nyra Usi</b>	 <b>Alexander Verdugo</b>	 <b>Rishabh Shetty</b>

### Project Mission / Motivation

The goal of this project is to create a tool that helps data scientists create more precise data analytics faster and more easily. Data is a valuable resource that is not always easily accessible due to poor database organization. Data housed in different systems such as, sales, support, product, and marketing have incompatible data

categorization, even if they were developed by the same organization!

A company could mandate that all the different systems converge on a single data model, but repeated failures show that such projects take years, cost millions, and are likely to start rotting almost as soon as the project is complete. Integrating data is hard which is why there is a multi-billion dollar industry trying to address it.



## Project Objective

The basic objective and the workflow of LinkIt involves the following phases:

1. **Loading** a small sample of data from each of their data sets
2. **Classifying** the data and metadata to determine the semantic meaning of each column of data (e.g. name, address, phone number, social security number, etc.) and cataloging the results.
3. **Finding linkages** of columns across datasets (e.g. there is a customer name in both of these tables)
4. **Viewing** the Catalog to allow a user to understand all of their data.

When the process is complete, users will be able to refer to the catalog to generate rich queries over the composite data set.

## Project Approach

To implement this project, two teams of people will work on the classifier framework, plugins and the catalog part of the project.

Team 1: Classifier Framework (*Andrew Liddle, Alexander Verdugo, Alejandro Ramirez*)

- Create the classifier framework
- Develop and test a dynamic loading framework for the plugins
- Provide an ability to run the classifier and update the catalog
- Integration and test of all components (*shared task*)

Team 2: Plugins (*Nyra Usi, Sayali Badole, Ethan Garnica, Prerana Kharat, Rishabh Shetty*)

- Develop and test plugins for generic content types
- Develop and test type-specific plugins
- Ability to designate a field or group of fields as Personally Identifiable Information (PII)
- Integration and test of all components (*shared task*)

All team members will have input on the format of the Catalog. Since the Catalog is an extension of the Framework, the Framework team will address any issues that arise with the Catalog.

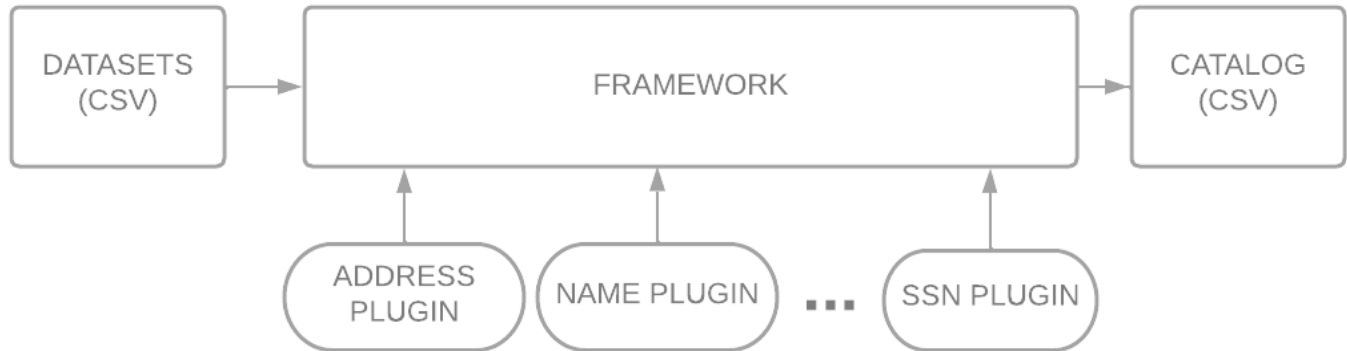
Tools used: GitHub (source code management) and Python.

Python is chosen for its extensive libraries and is compatible with data cleaning.



GitHub is chosen because it is the most commonly used version control tool.

Flowchart is as follows:



## Dependencies / Risks

Dependencies:

- Availability of data sets for testing and training the classifier framework and plugins.
- Access to necessary APIs or data sources for integrating external data.
- Adequate funding to purchase any necessary hardware or software.
- Compliance with data privacy and security regulations.

Risks:

- Inaccurate classification of data due to errors in the plugins or the classifier framework.
- Lack of compatibility with different data sources and systems, leading to integration issues.
- Inadequate or incorrect data labeling, leading to incorrect catalog entries.
- Potential data privacy or security breaches during data integration.
- Technical difficulties in integrating different data sources and systems.

## Project Deliverables

There are a total of 3 features: the Plugins, the Framework, the Catalog.

The Plugins are used to identify and categorize data. This means all plugins will take in the data from the columns of the table and provide a score with how well the inputted data matches the categories. Generally



## CST499 - CS Capstone Project Proposal

one plugin will be created for each category. The categories are: name, phone number, address, social security number, email address, URL, social media handle, currency, credit card information, generic text, generic number, generic date, generic ID.

The Framework is used to find, load, and execute the plugins. From the plugin results, the framework will populate the catalog.

The Catalog is a CSV file that categorizes the data found in the table columns and presents a confidence score on that chosen category.