# Aggregation and Structuring of Materials and Chemicals Data from Diverse Sources

**FWP 100250**
**SLAC and Citrine Informatics**
**July 1, 2017-Sep 30, 2019**

Christopher J. Tassone, SLAC National Accelerator Laboratory

U.S. DOE Advanced Manufacturing Office Program Review Meeting
Washington, D.C.
June 11, 2019

*This presentation does not contain any proprietary, confidential, or otherwise restricted information.*

# Overview

## Timeline

- Award issued July 1 2017
- End date September 30, 2019
- Project 80% complete

## Budget

| | FY 17 Costs | FY 18 Costs | FY 19 Costs | Total Planned Funding (FY 19-Project End Date) |
|---|---|---|---|---|
| **DOE Funded** | 142K | 1.05M | 1.1M | $2.3M |
| **Project Cost Share** | 47K | 394K | 309K | 750K |

## Barriers

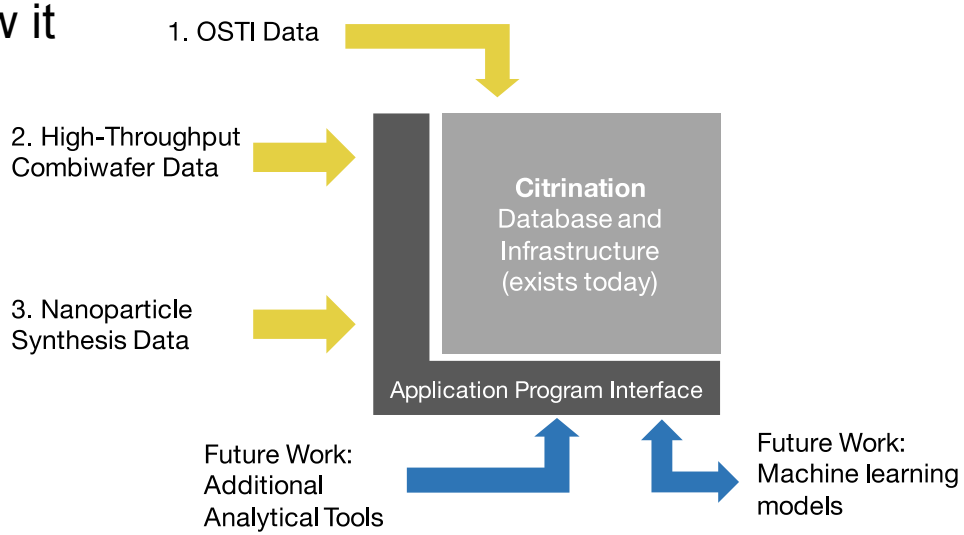
- Key barrier to on-purpose design of materials with targeted properties are data and algorithms which predict the processing pathway for a given property target
  - Machine readable database containing processing-structure-property data
  - Active learning algorithms to predict processing conditions for targeted properties
  - Active learning algorithms guide database building
  - Automated hardware for accelerated characterization

## Partners

- Citrine Informatics provides data aggregation, development of machine learning models, and platform development to support data pipelines
- NREL is responsible for the synthesis of combinatorial libraries
- Colorado School of mines is developing HiTp mechanical property characterization

# Project Objective

**Goal:** Develop a centralized, granular, structured set of materials and chemicals data from diverse sources, and demonstrate how it can be leveraged to accelerate application driven materials R&D




**Problem Statement:** Scientific data is often heavily siloed, requires intensive analyses, with information lost at each step of manipulation resulting in loss of fidelity and slow dissemination of information which could expedite materials discovery and deployment.

**Challenges**

- Automated aggregation of data from centralized data sources
- Automating route data analysis steps
- Building a broadly applicable data handling tools across data types
- Developing data driven predictive frameworks for which no first principle models exist

# Technical Innovation

## Existing Practices

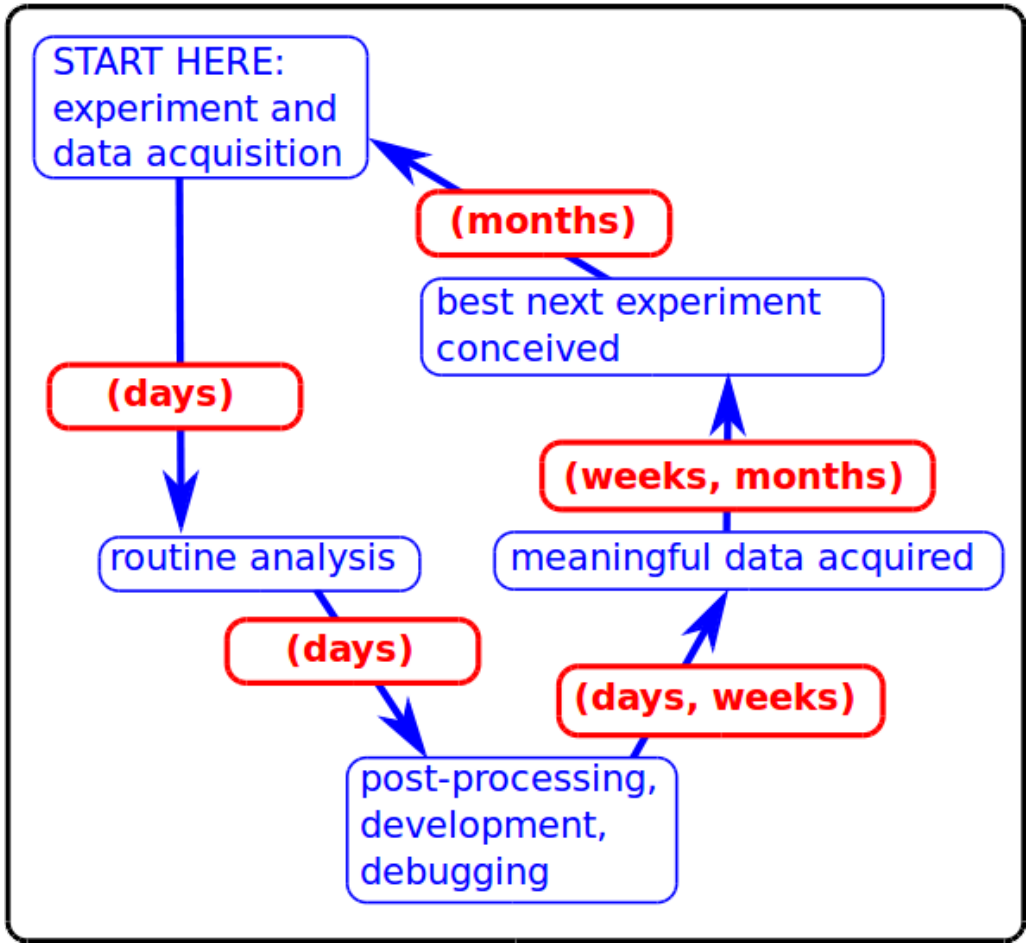
### 1) Data Aggregation

- Aggregation of reports/literature which are not machine readable
- Compilation into structured databases requires time intensive human document extraction

### 2) Knowledge generation

- A user collects a batch of measurements
- Manual manipulation of data to transform data into a processable form
- One or more specialized software packages, or user-developed scripts, provide further processing
- When meaningful results are found (**weeks or months**), the scientist plans the next experiment

**State of the art:**



# Technical Innovation

## Turning data into information

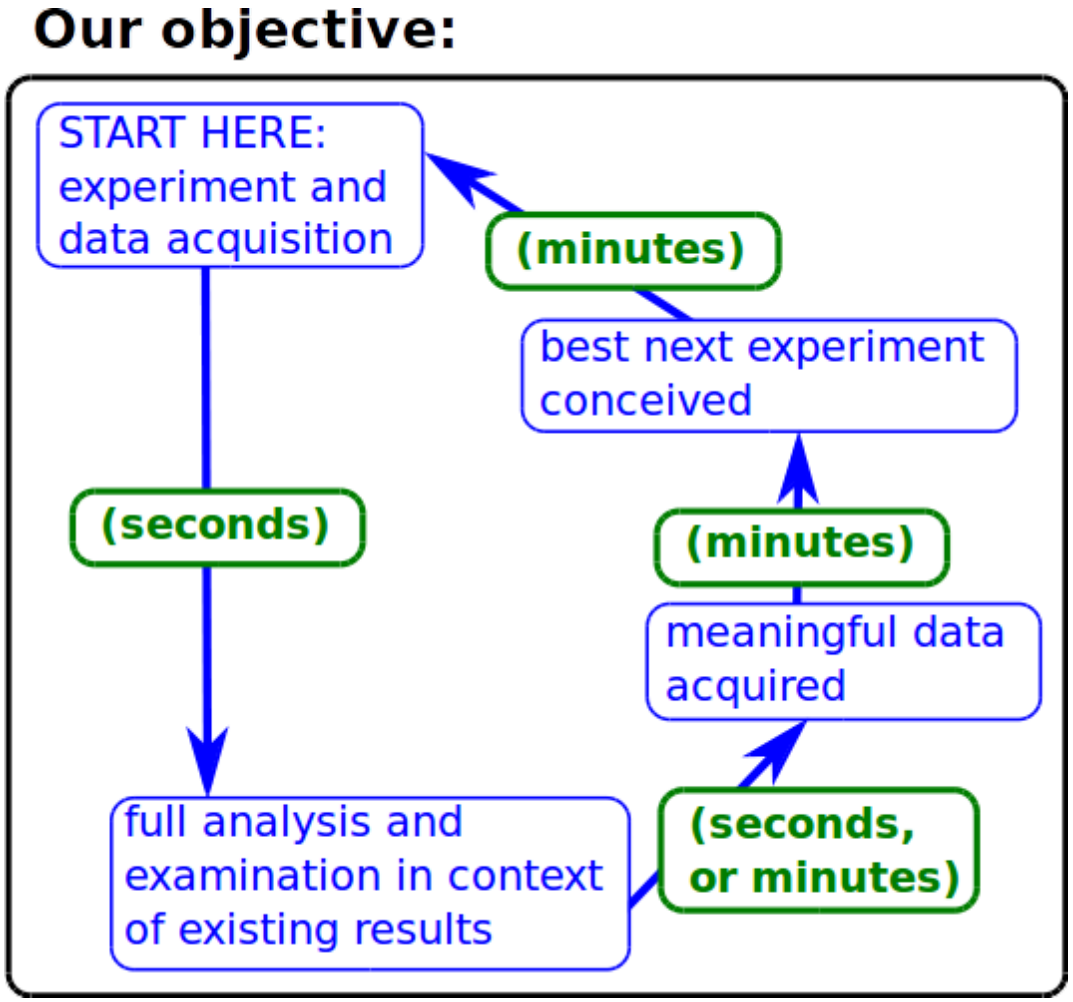
- Existing data is extracted into machine readable form
- As data is captured it is manipulated to:
  - Provide researchers actionable information in real time
  - Structured into a machine readable database

## Experimental Planning

- Automated data handling software extract metrics of interest in real time
- AI makes informed decisions about subsequent experiments in real time

## Cost Reductions

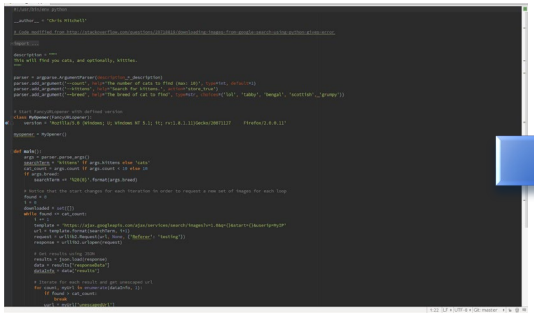
- Experiment efficiency increased
  - Bad data immediately flagged
  - Most valuable experiments performed
- Duplication of effort eliminated



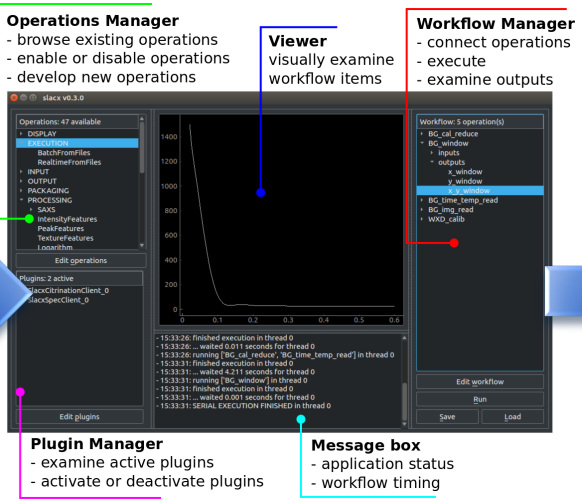

- Basic research informs industry in real time
- Efficiency of analysis intensive experiments increased by up to 3 orders of magnitude

# Technical Approach

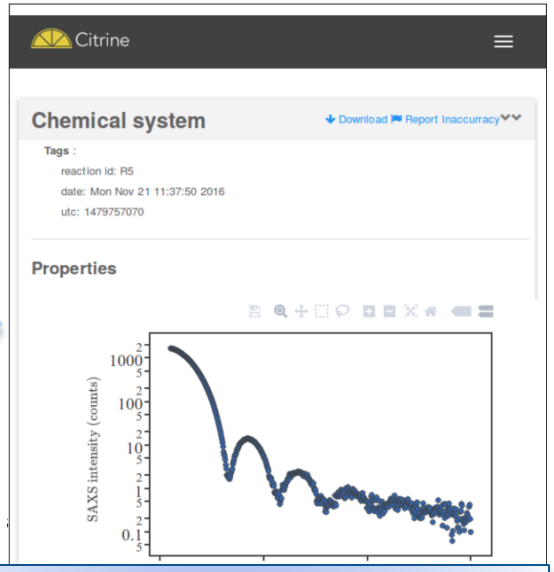
## Computational Tools



Prototype Scripts

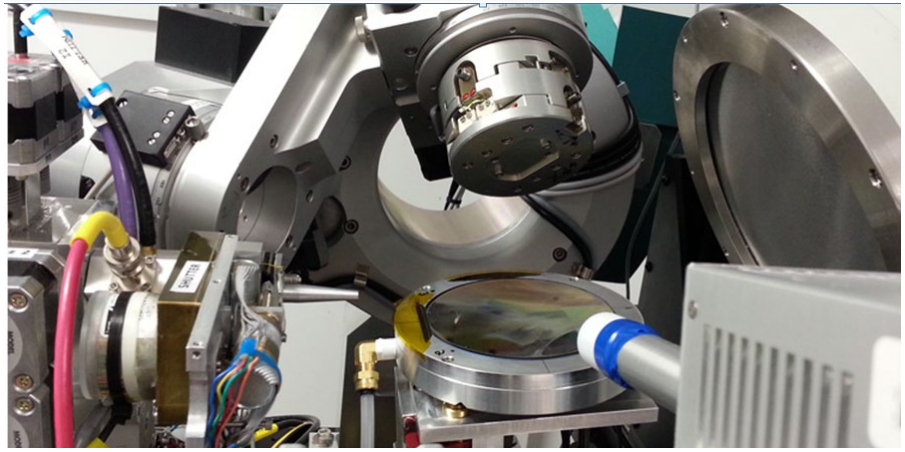


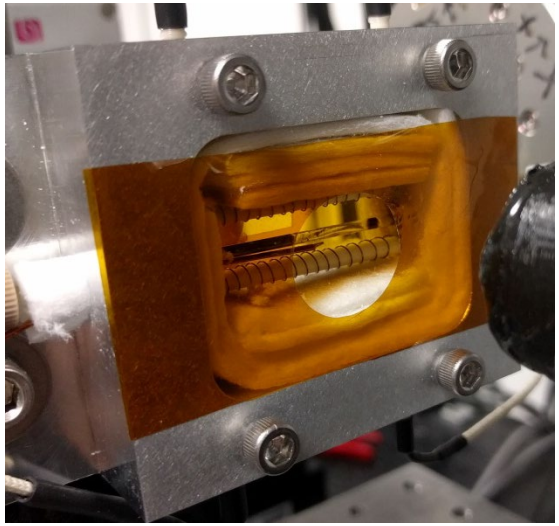

Production level software




Structured data integrated into Citrination

## Experimental Hardware





# Technical Approach

## Project Participants

**SLAC:** Position as a national user facility enables broad dissemination of best practices, accessible tools for industry and academia.

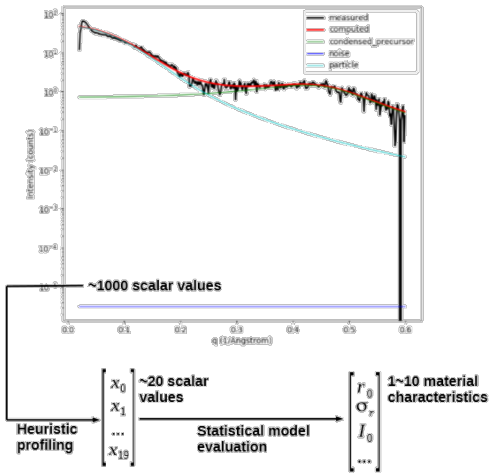
- Develop automated data handling software
- Automate the experimental planning and execution process
- Discover high performing metallic glasses and associated processing
  - AI driven prediction of ternary glass forming ability
  - AI driven prediction of alloy toughness and ductility
- Discover high performing propane dehydrogenation catalysts and associated synthesis pathways
  - AI driven synthesis of novel nanoparticle catalysts
  - AI driven catalyst target selection

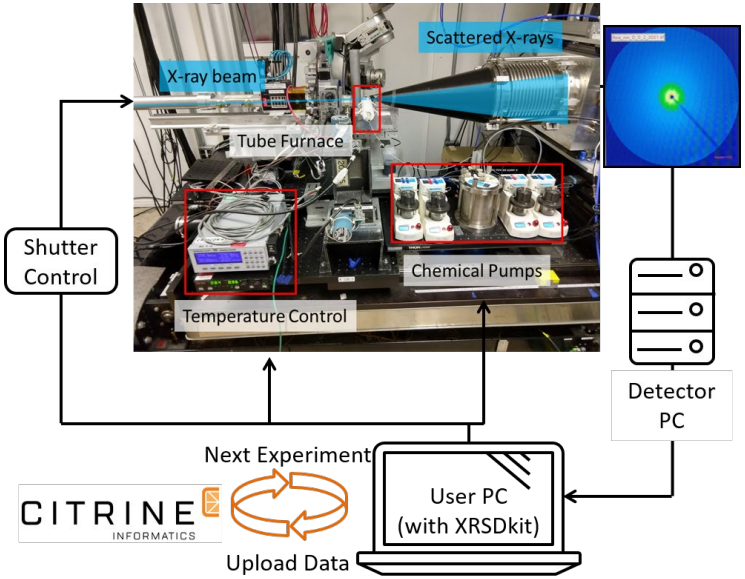
**Citrine:** Industry leader in machine learning on materials data.

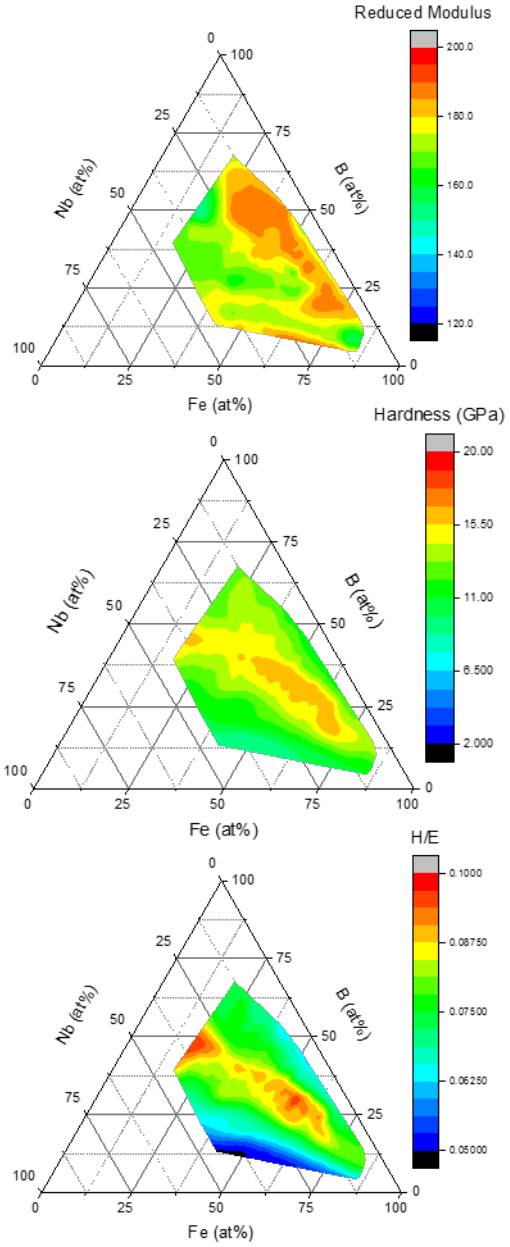
- Automated extraction of information from centralized databases and literature
- Aggregation of data into Citrination database to enable cloud based dissemination
- Develop active learning and machine learning models to guide R&D
- Develop platform API to accommodate project demands

# Results and Accomplishments

- Automated data assessment and analysis
- Data aggregation from Diverse Sources and Web visualization
- Trained ML model for prediction of ternary GFA
  - Discovery of 4308 new glass forming alloys across 5 ternaries
- AI driven optimization of nanoparticle synthesis
  - HiTp reactor capable of 100s of syntheses/day
  - AI guided synthesis optimization validated to reach target NP



# Transition and Deployment

- **Target audience:** National user facilities, individual researchers, and industry.

The effect of aggregating data from diverse sources and capturing data as it is collected will increase efficiency of individual researchers, while providing a robust database of process-structure-property relationships to be leveraged downstream by industry. Expediting the materials discovery to deployment process significantly, and maintaining the US manufacturing competitive advantage.

- **Distribution:**
  - Software developed is freely available, under a DOE-approved BSD-like license (https://github.com/slaclab/paws & https://github.com/slaclab/xrsdkit)
  - Aggregated data is freely available via the Citrination platform (https://citrination.com/)
- **Adoption:**
  - **National Laboratories:** Data handling tools adopted by software developers at Berkeley Labs and under continued development by CAMERA
  - **Industry:** A.I. guided R&D methods and data pipelines developed deployed through CRADAs to develop proprietary materials