NSF CSSI PI Meeting, Seattle, WA, Feb. 13-14, 2020

Award #: OAC1640867

CSSI Element: MADE@UB: Materials Data Engineering at UB

PI: Venu Govindaraju, Co-PIs: Krishna Rajan, Tom Furlani, Srirangaraj Setlur
Institution: University at Buffalo, State University of New York

Material Science Design Challenges

This project includes the leaders in data driven design through Materials Informatics. The project leverages this expertise in the material science domain in conjunction with the computer science expertise to achieve the following objectives :

1) We are aiming to provide a data infrastructure that permits one to link diverse types of data to help users solve complex materials problems. In these cases, one type of data (such as only crystallographic, thermodynamic, kinetic, and other types of data) is typically not sufficient.
2) The bulk of knowledge lies not in well organized databases but is spread over technical papers, books, notes, etc. Further, much of the "data" lies in the empirical knowledge and heuristic interpretations of graphs and figures. The latter is almost untapped in terms of Machine Learning. This program focuses on this challenge.

Computer Science Challenges

The PI and Co-Is from Computer Science are pioneers in developing AI/machine learning technologies for document recognition and data analysis. This project brings computer science expertise into the material science domain for the following objectives :

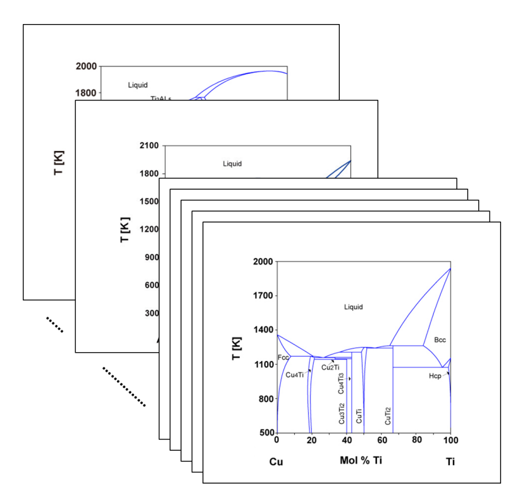
1) We develop state-of-the-art machine learning techniques for extracting data and insights from scientific charts and figures, a field which is still in a nascent stage. We take a two-pronged approach to tap this hitherto unexplored rich source of data: A generalized approach to extract data from common types of charts such as line, bar and pie charts, by parsing commonly occurring elements; and a targeted approach to process specialized diagrams (e.g. phase diagrams) by leveraging domain knowledge.
2) We develop innovative tools that allow practitioners to easily experiment with and understand a variety of machine learning models using their own data, across multiple pipeline stages so that time is spent on analysis rather than programming.

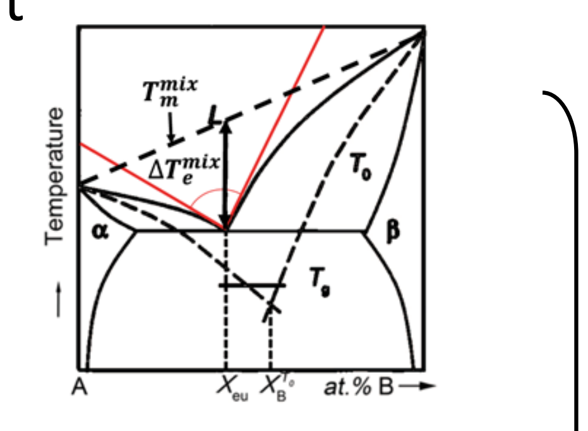
Convergence of Materials Science and Computer Science for Accelerated Discovery

Graphics recognitions to uncover overlooked data for the discovery of new materials

- Infographic tools extracted contours and boundaries from phase diagrams. Phase diagrams are widely available, but are generally used only to visually assess stable microstructures.
- By integrating this data with informatics based analysis of thermodynamic and crystallographic data, coupled with uncertainty quantification techniques, nearly 40 binary alloys that have been overlooked as metallic glass formers were discovered.

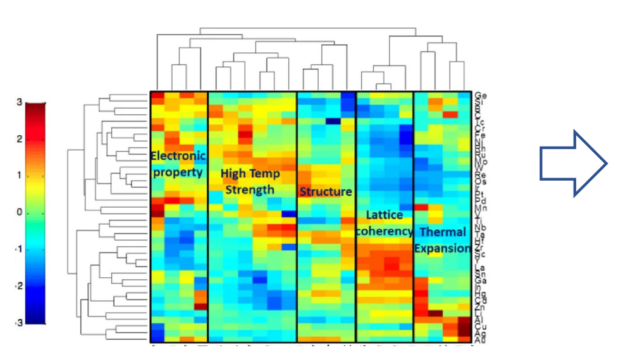
Thermochemical Design Input

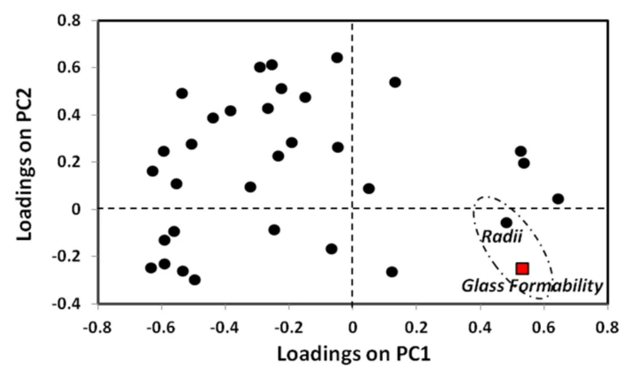




Graphics Recognition to extract key feature

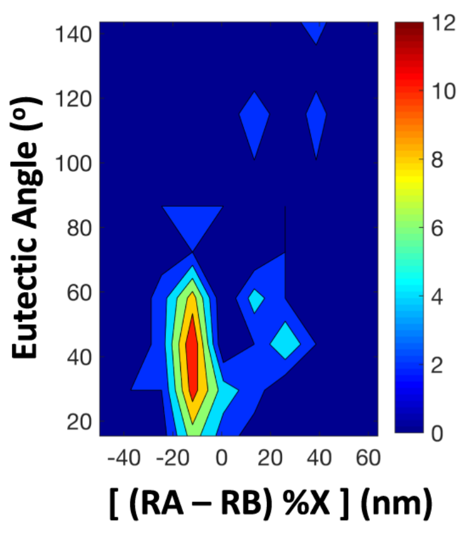
Structural Design Input

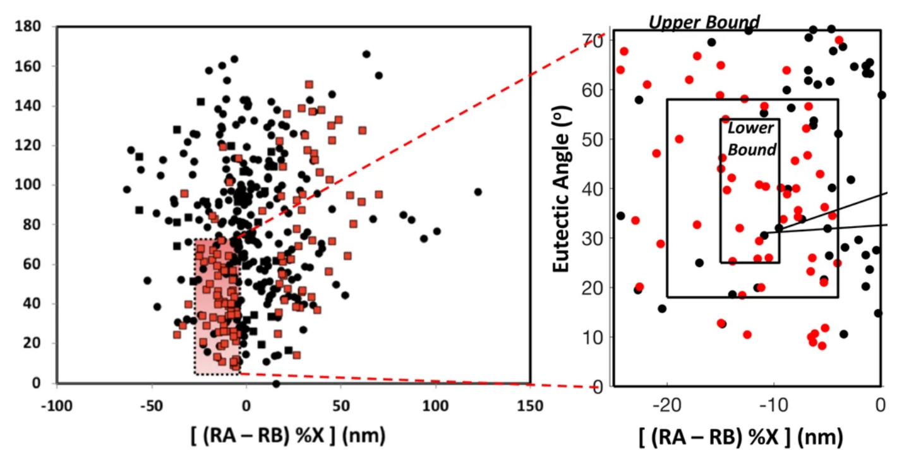




Genetic Programming and Manifold Learning

Uncertainty through **Rough Set Analysis** : Discovery of 'Sleeper' Alloys

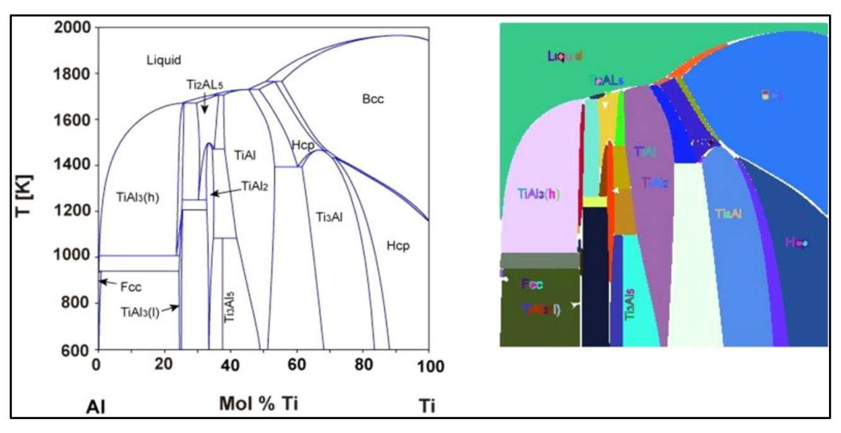




Led to joint MDI / CSE paper : over 1,000 downloads in last year.

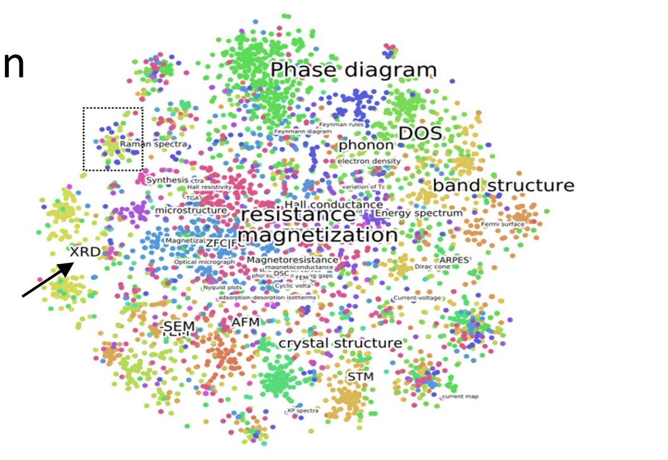
Application of infographic tools for outreach and education activities in materials

- Tools were developed for extracting and more efficiently using data, not just in terms of designing materials but also in how to represent issues.



Representing diagrams in other forms to capture overlooked information

Classifying information and relationships from figures and text



- While these tools were developed for research purposes, they have also been brought into the classroom, for K-12 students.

- Used in our Data Advocacy program, which collaborates with high school teachers to teach students on how to apply data science tools.
- At Western New York Youth Climate Action Summit, infographic tools were used by students to better represent the need for renewable energy, such as solar

Enabled outreach efforts to extend to a network of over 100 teachers and ~ 300 students

MaDE@UB Web portal (madeatub.buffalo.edu)

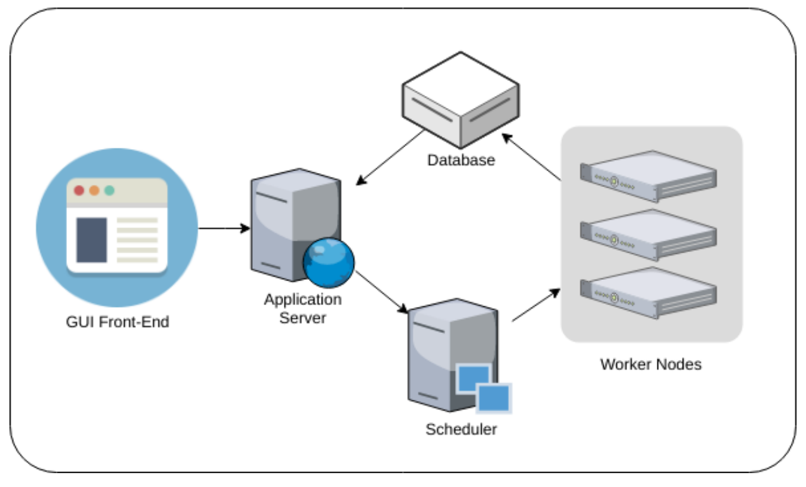
Machine Learning Toolkit

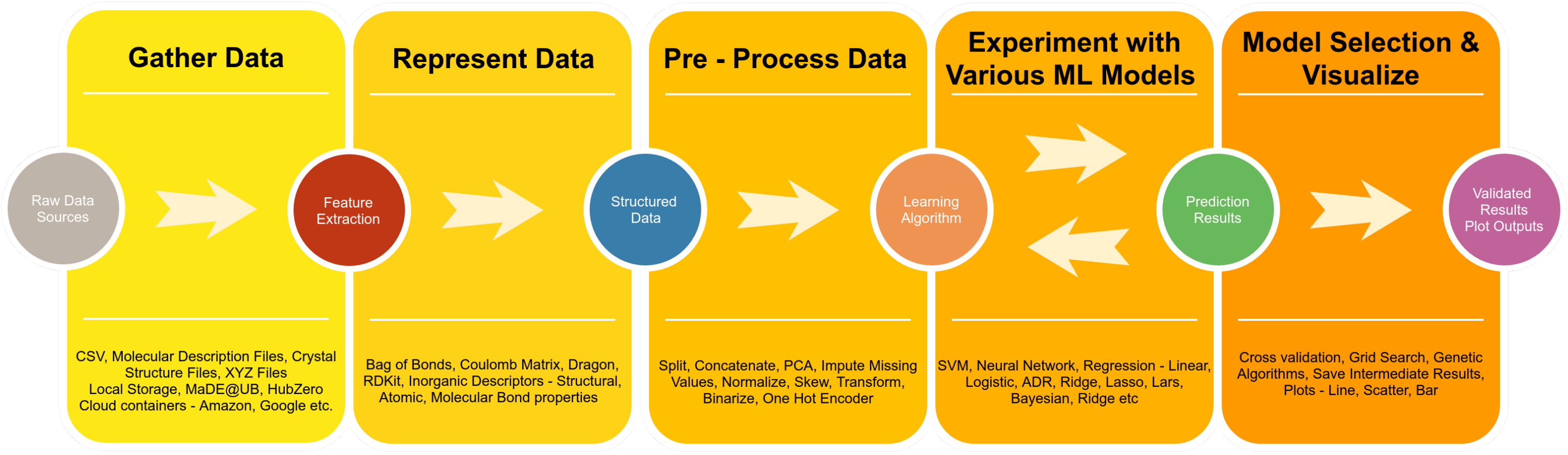
- ❑ Contributed vital modules to ChemML– an open-source package to setup, execute and explore data processing pipelines.
- ❑ Inorganic descriptor modules, neural networks added to cater to wide variety of use cases.
- ❑ ChemML was integrated with a web framework and GUI to facilitate easy exploration of data science pipelines. It can also be used as a pedagogical tool to teach data science concepts.

ChemML Repository

MADE@UB ML Toolkit

Web Framework Architecture





CSV, Molecular Description Files, Crystal Structure Files, XYZ Files
Local Storage, MaDE@UB, HubZero
Cloud containers - Amazon, Google etc.

Bag of Bonds, Coulomb Matrix, Dragon, RDKit, Inorganic Descriptors - Structural, Atomic, Molecular Bond properties

Split, Concatenate, PCA, Impute Missing Values, Normalize, Skew, Transform, Binarize, One Hot Encoder

SVM, Neural Network, Regression - Linear, Logistic, ADR, Ridge, Lasso, Lars, Bayesian, Ridge etc

Cross validation, Grid Search, Genetic Algorithms, Save Intermediate Results, Plots - Line, Scatter, Bar

Data Extraction Tools and Outcomes

Sub-Task Chart
Upstream tasks feed their output as the input of downstream tasks

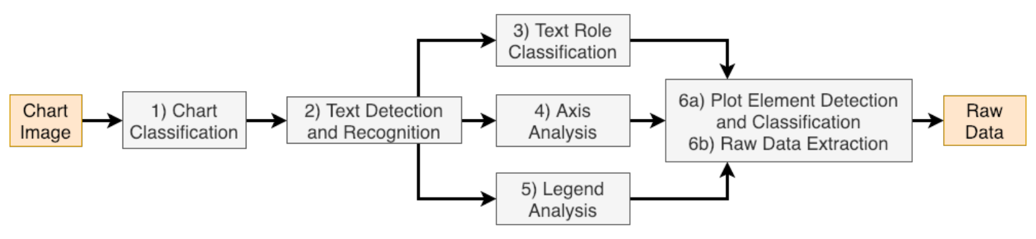
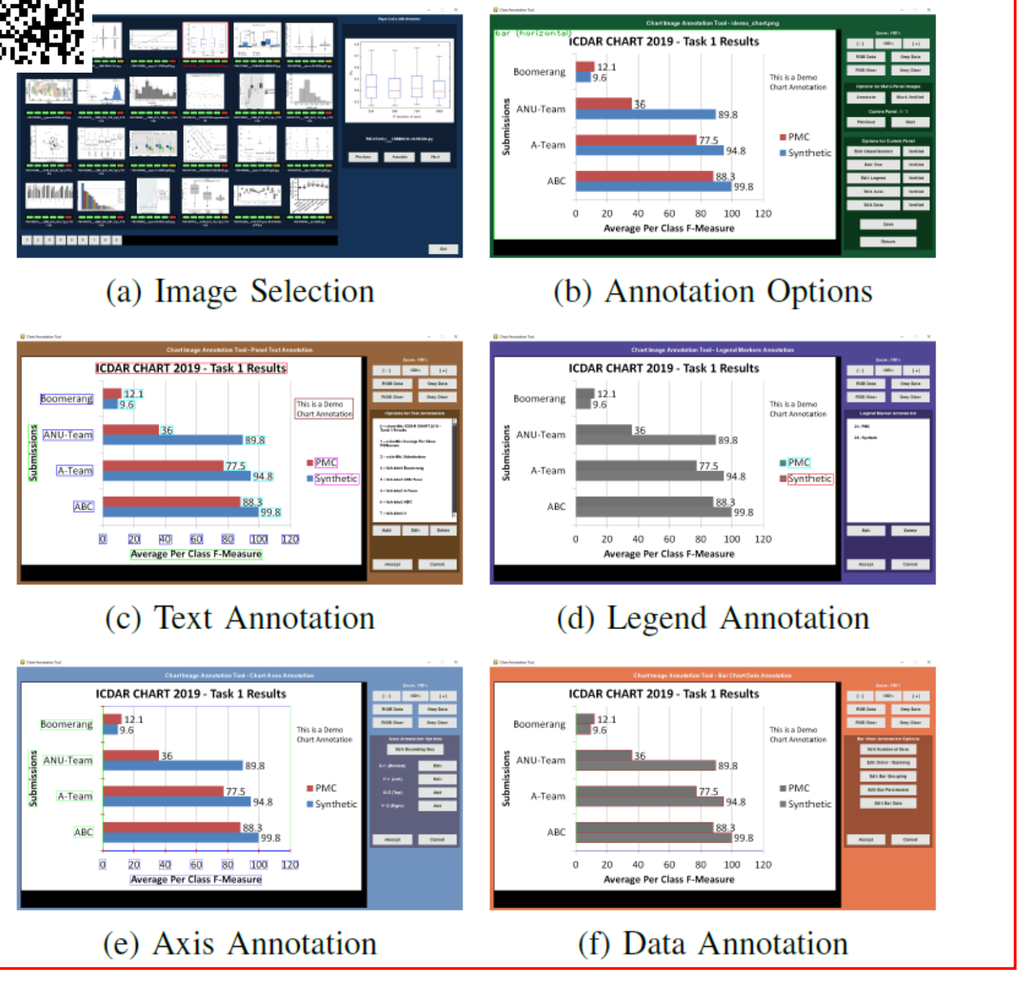


Chart Image
1) Chart Classification
2) Text Detection and Recognition
3) Text Role Classification
4) Axis Analysis
5) Legend Analysis
6a) Plot Element Detection and Classification
6b) Raw Data Extraction
Raw Data

- ❑ ML based methods to extract data from general charts have been developed as well as a specialized tool for annotating charts.
- ❑ A competition was organized in ICDAR 2019, the flagship conference for Document Recognition in collaboration with Adobe Research to encourage the community to engage in chart data extraction research.
- ❑ Datasets including synthetic and those drawn from scientific papers were assembled for the competition.
- ❑ The overall task of data extraction was divided into sub-tasks with metrics designed or adapted for each stage.
- ❑ This competition for comprehensive end-to-end data extraction is the first of its kind..
- ❑ Innovative techniques developed to extract and summarize content from instructional videos won best paper awards at ICDAR 2019 and CBDAR 2019.

Annotation Tools for Different Chart Elements


ICDAR CHART 2019 - Task 1 Results
Boomerang
ANU-Team
A-Team
ABC
Submissions
Average Per Class F-Measure
PMC
Synthetic

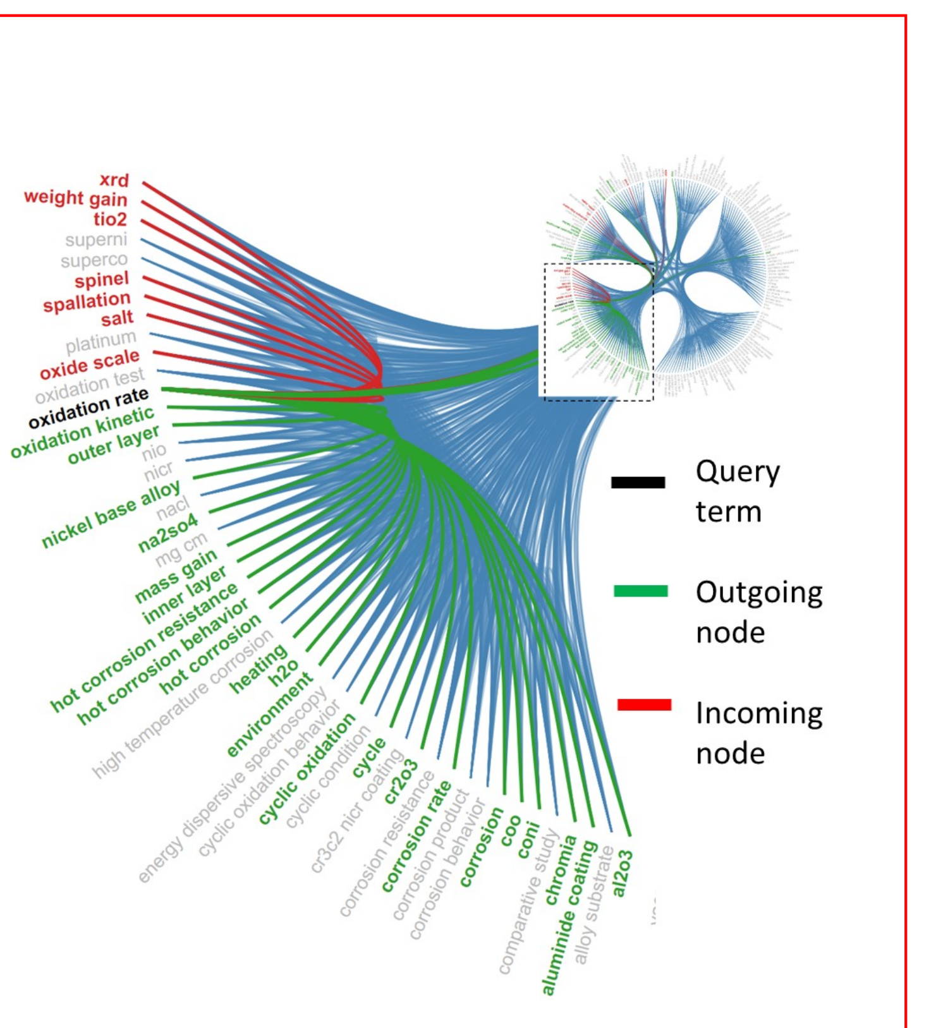
(a) Image Selection (b) Annotation Options

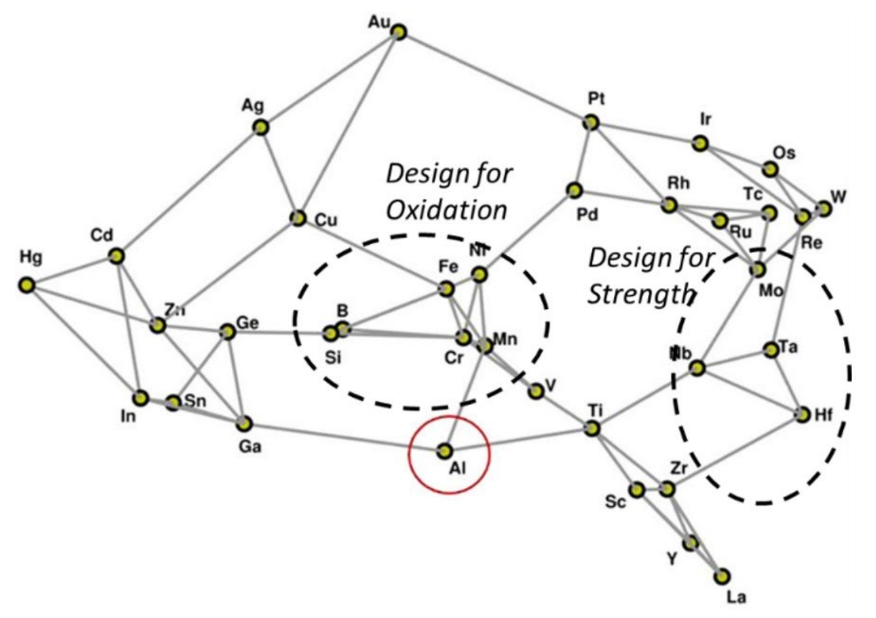
(c) Text Annotation (d) Legend Annotation

(e) Axis Annotation (f) Data Annotation

Alloy Design Foundry Tools and Capabilities

•Natural Language Processing : Capture chemistry-processing-performance relationships hidden in the large amounts of legacy information : Included is visualization methods, such as inference through Sankey plots



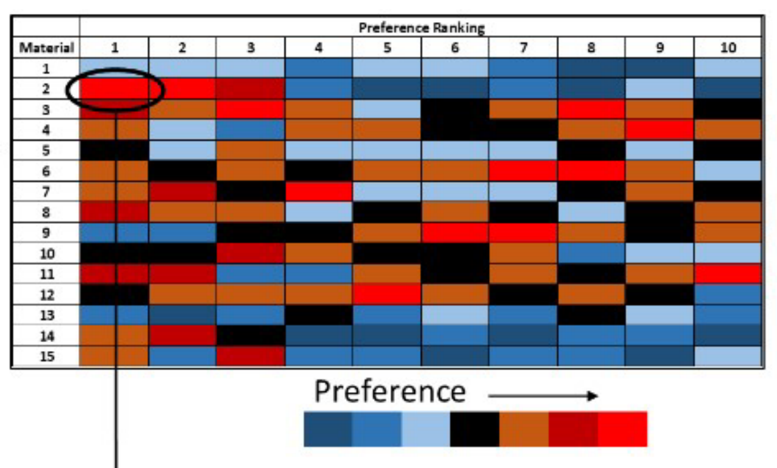


•Manifold Learning: Identify new unexplored systems through exploration of chemistry-performance relationships in new systems (linear and non-linear dimensionality reduction tools)

•Decision Theory : Select most promising materials through consideration of chemistry-structure-performance trade-offs in sparse and uncertain data.

•Data Connectivity : Unsupervised data tools for exploration of connectivity of data (Topological Data Analysis)

Output



For preference criteria:
Yield Strength > Formability = Machinability = Joinability > Ductility = Cohesive Energy

Preferred Material :
51.6 Ni – 21.5 Cr – 5.5 Fe – 2.5 Co – 0.01 C – 1.0 Mn – 0.1 Si – 13.5 Mo – 4.0 W