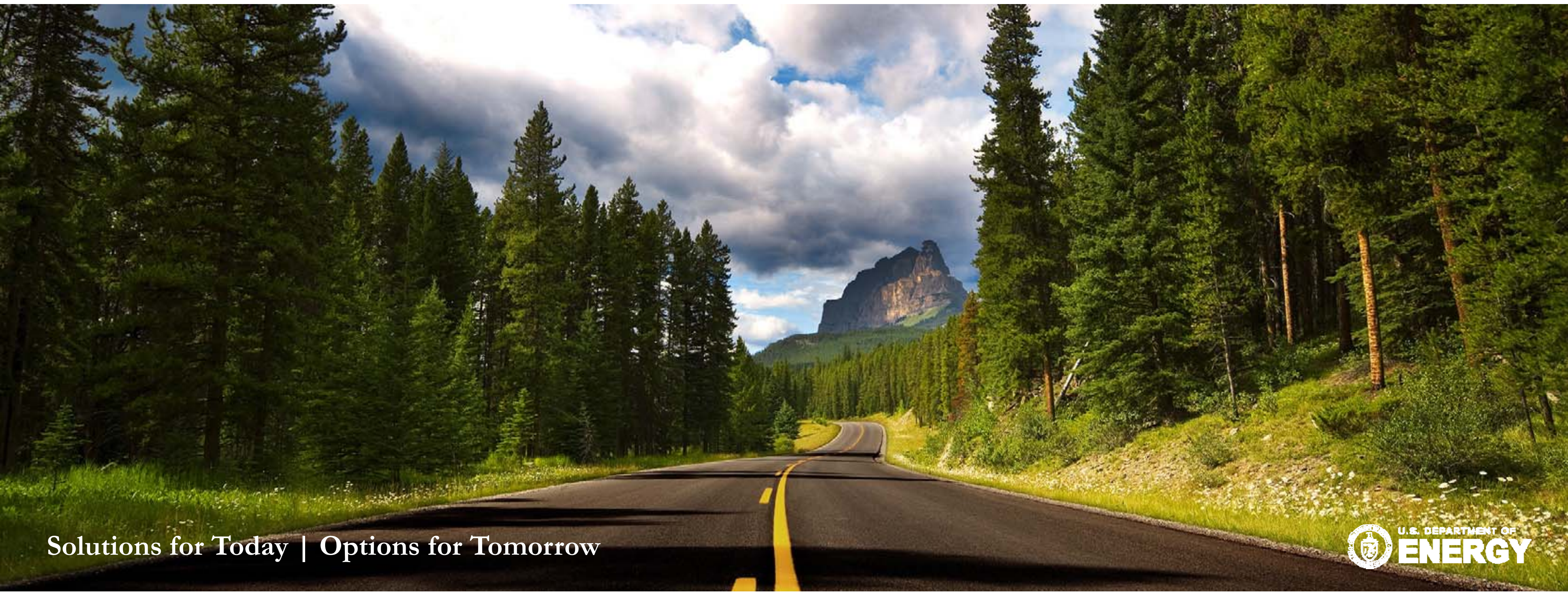
Use of Data Analytics in Advanced Alloy Development: Trends and Modeling

PI: Slava Romanov, NETL (romanov@netl.doe.gov)

Crosscutting Research: Computational Materials

March 21, 2017



PROJECT Goals and Objectives

FY 2017

Overarching Goal:

- **MDA Tools to Reduce Time and Cost of Alloy Development, Certification, and Qualification**

Objectives:

- **Establish a reference methodology for non-linear model development**

  Milestone: Complete the linear exploratory data analysis using Materials-CRADLE methodology / FE0028685 Contract Final Report (09/30/2017)

- **Calibrated and validated tools for selective model application within non-random data space**

  Milestone: Design algorithms for mining the associative data patterns and cluster analysis to partition the multi-dimensional data space / NETL Technical Report on data-driven modeling approaches (09/30/2017)

Presentation Outline

NETL and CWRU collaborative research

- NETL and CWRU Team
- Project Scope
- NETL Data Management
- CWRU Energy CRADLE
- Exploratory Data Analysis
- Clustering, Classification and Visualization
- Data-Driven Non-Linear Modeling
- SUMMARY

Project Team

NETL and CWRU (FE0028685 Contract)

- Dr. Vyacheslav (Slava) Romanov, NETL Project PI
- Dr. Jefferey Hawk, NETL TPL
- Dr. Siddharth Maddali, ORISE Fellow (currently ANL)
- Narayanan Krishnamurthy, ORISE Fellow, Univ. Pittsburgh
- Prof. Jennifer Carter, CWRU Contract PI
- Prof. Roger French, CWRU Contract Co-PI
- Prof. Laura Bruckman, CWRU Contract Co-PI
- Dr. Mohamed Elsaeiti, CWRU (appointment ended)
- Amit Kumar Verma, CWRU

Alloys Pilot

Data Curation and Mining

Identify and collect data (with material and model pedigree information) on 9Cr steel family of alloys

- mechanical properties (tensile, creep, low cycle fatigue, and creep-fatigue)
- microstructures (austenite grain size, lath size, carbide size, carbide volume fraction)
- results of computational modeling and
- design data

Store, share, and pre-process data

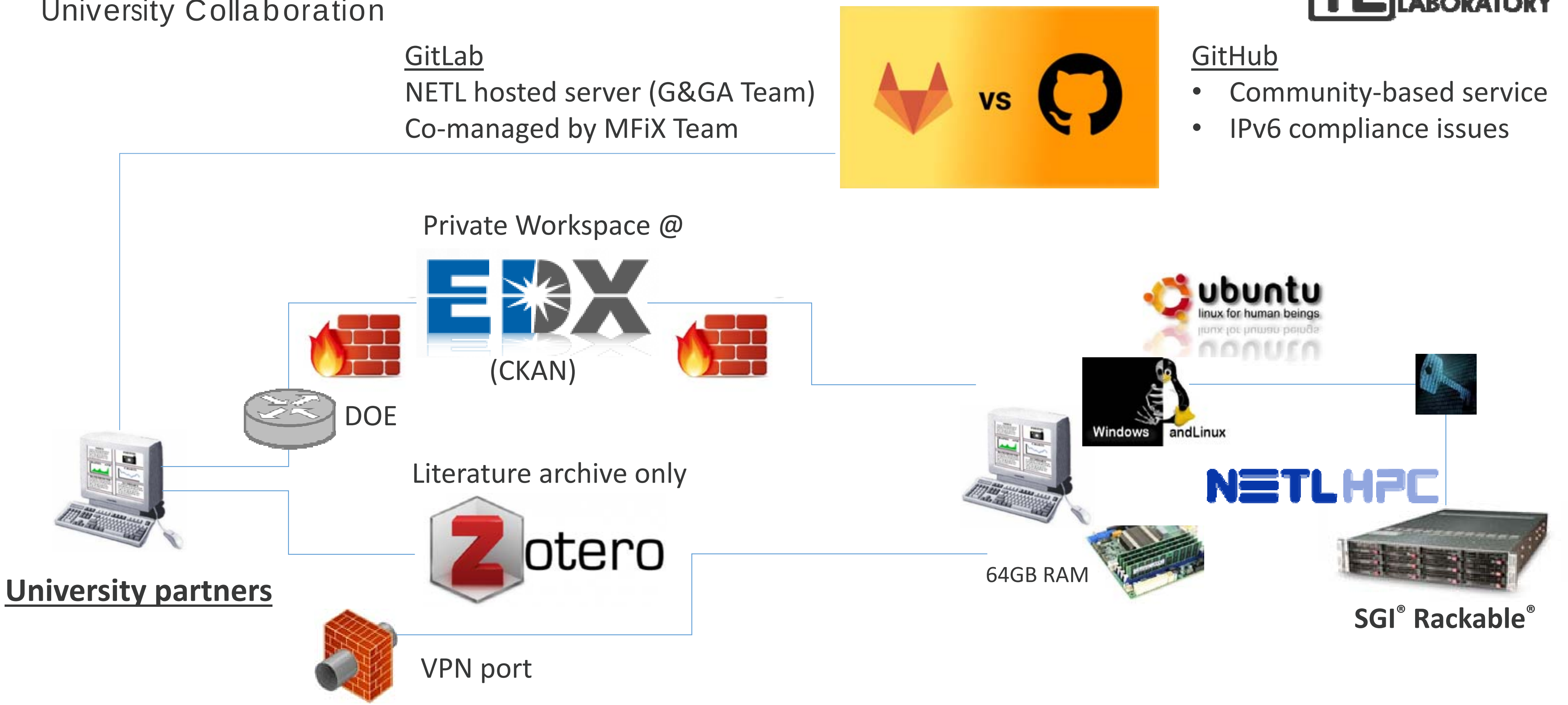
- cleaning, parsing, and validating
- building, managing, and maintaining tidy data sets
- preserving data and model provenance

Analyze the data

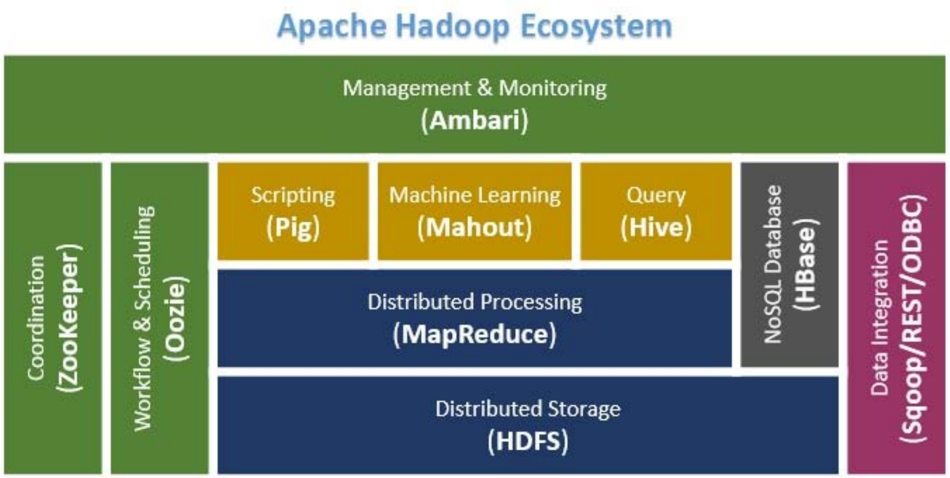
- data analytics
- uncertainty quantification
- identifying data gaps and outliers

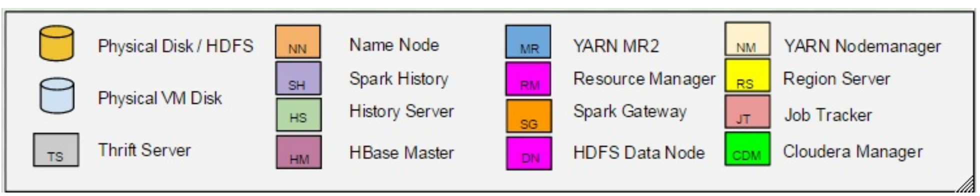
Data Management

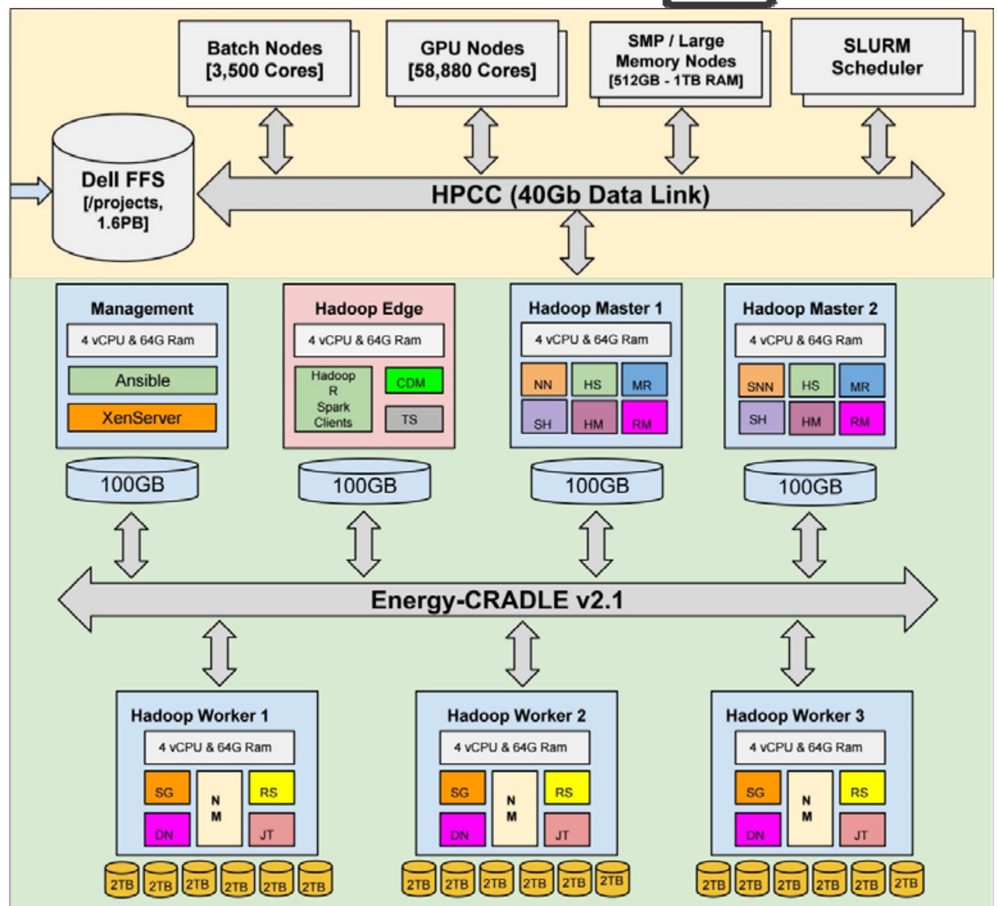
University Collaboration



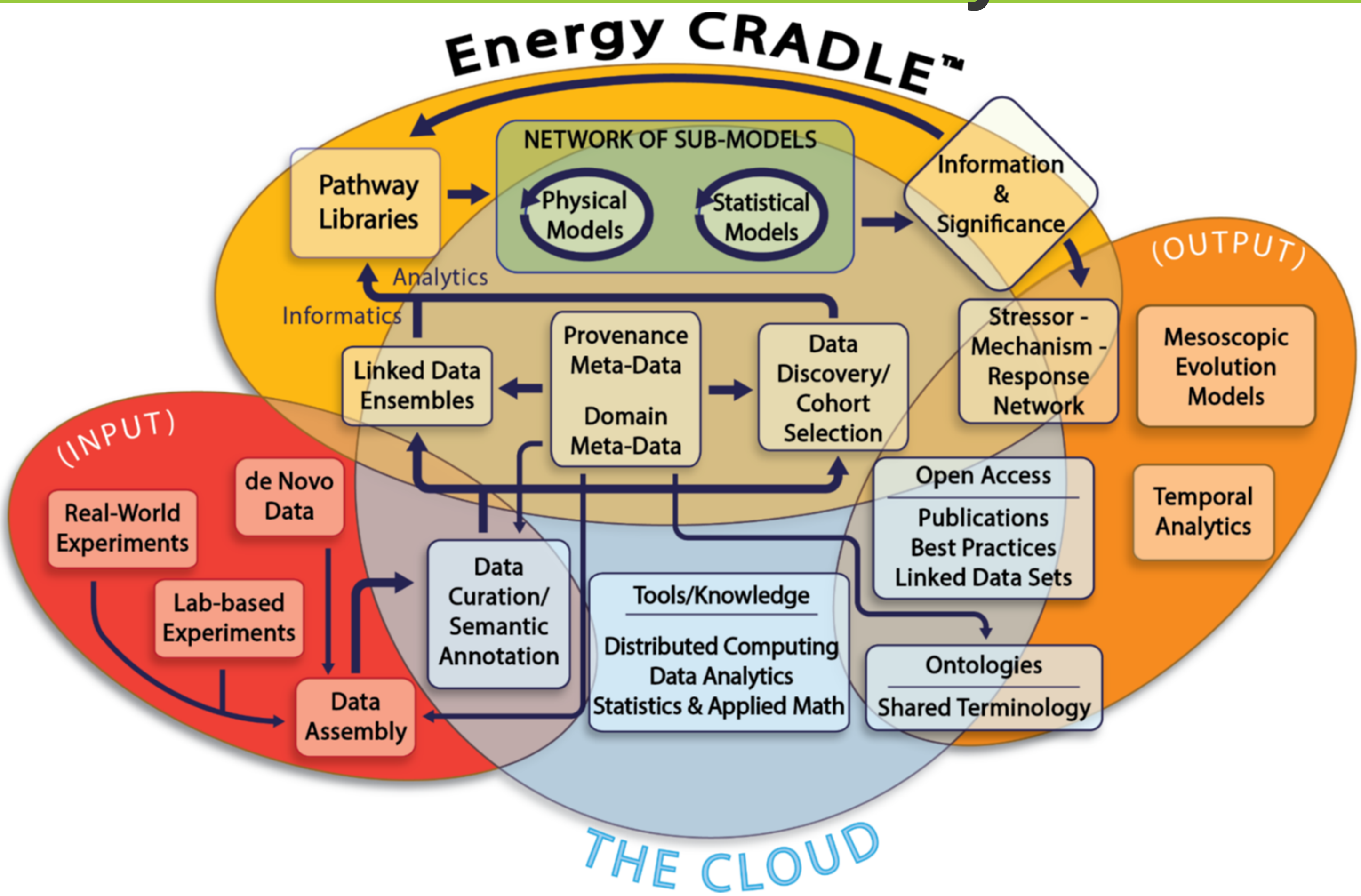
Case Western Reserve University







Case Western Reserve University

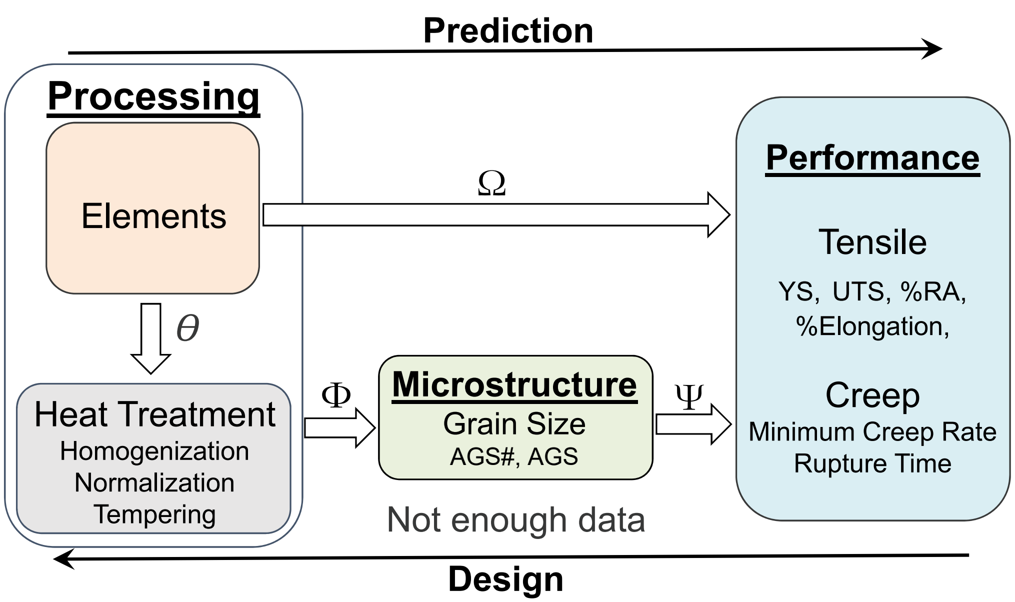


Support Contract Technical Approach

Case Western Reserve University

Utilize data analytics to guide the design and development of the next generation of 9Cr-steel alloys:

- Optimize chemistry for performance
- Optimize heat treatment for performance of a particular chemistry



Quantitative mapping functions (Φ,Ψ) are needed for efficient design of a material to achieve a performance metric for an application. These mapping functions provide statistically-derived models of the relationships between processing (stressors) and microstructure metrics that capture the physics of damage.

Data Analytics Workflow

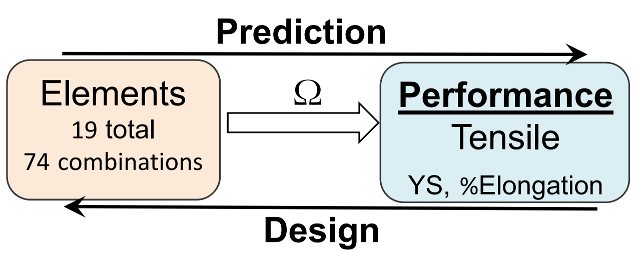
Case Western Reserve University

Data → Ingestion & Cleaning → Exploratory Data Analysis → Linear Model → Validation

Exploratory Data Analysis →
- Pair-wise Correlation Plots
- Stepwise Linear Regression
- Domain Study

Acquired from **NETL Albany, OR**

- 30 years of development work
- 47 total variables / metrics
    - 22 Variables
    - 826 Observations

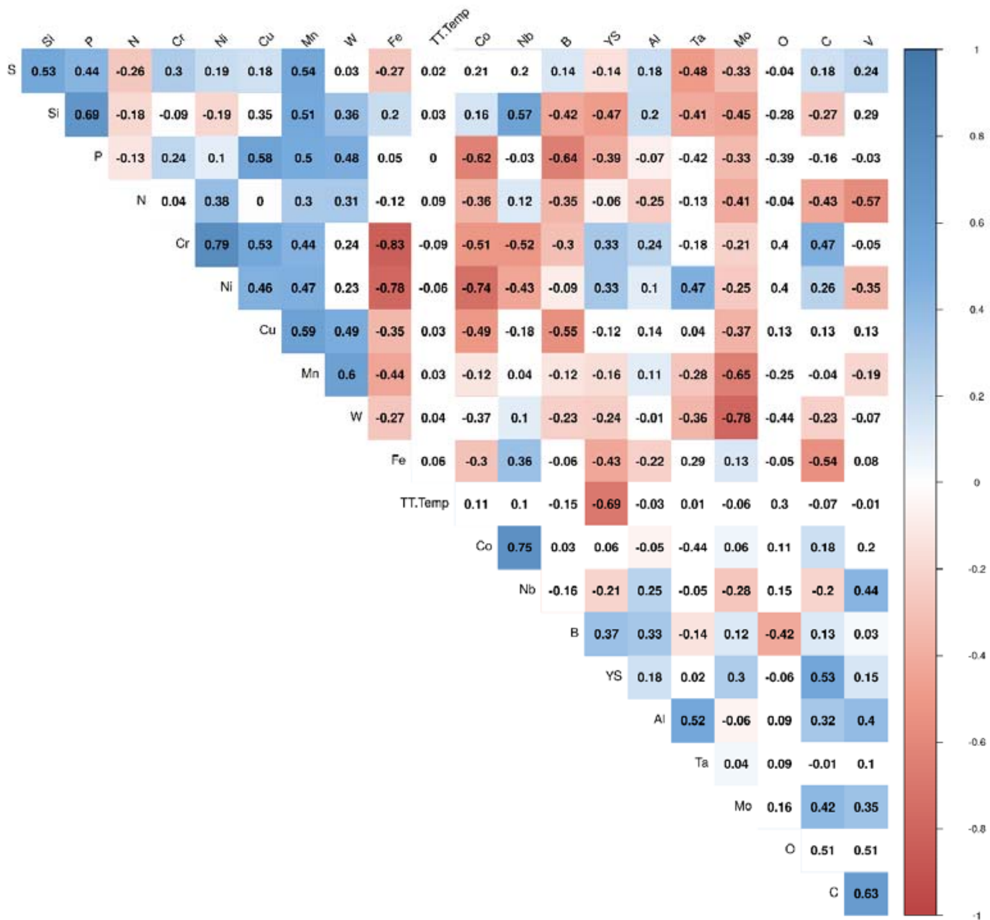


Exploratory Data Analysis

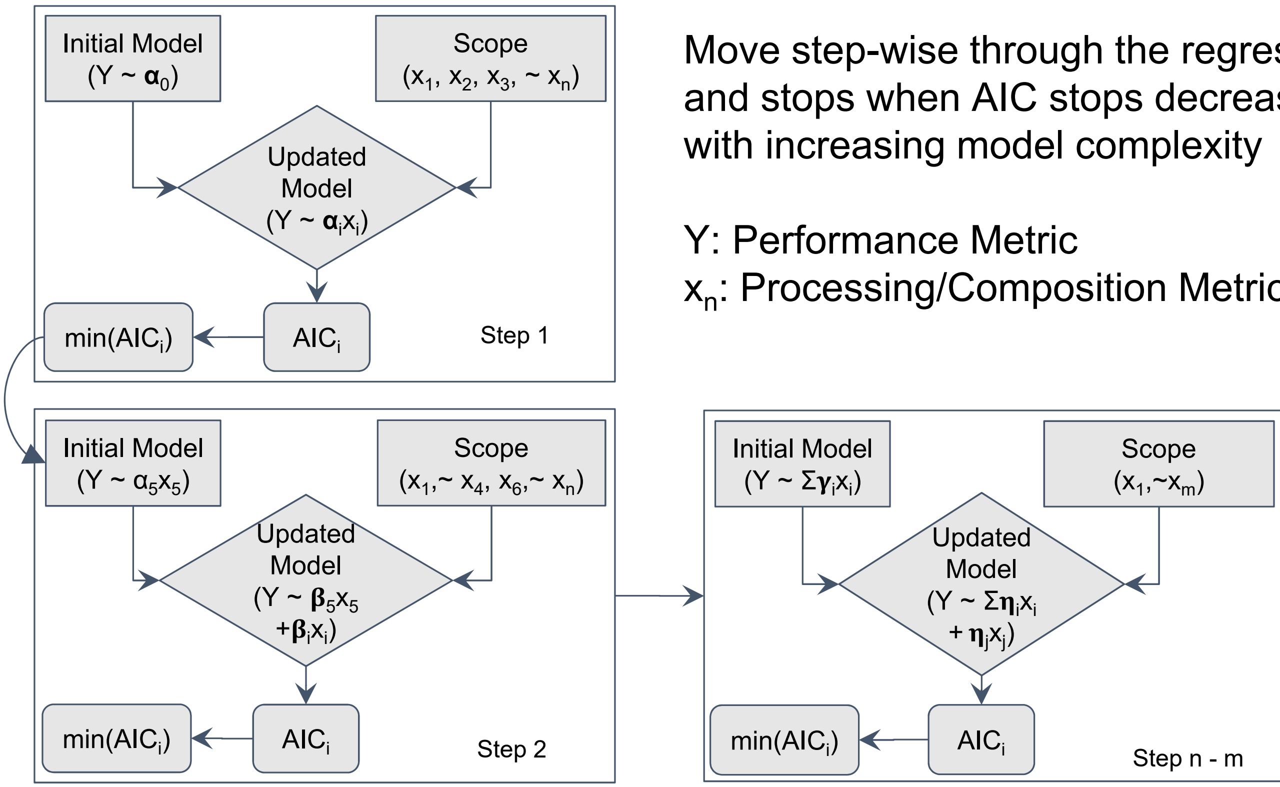
Pair-wise correlation plots

- Captures uni-variate relationships
- Separates linear and non-linear relationships
- Shows interdependency between variables

Upper diagonal: linear correlation coefficients



Stepwise Linear Regression



Move step-wise through the regression and stops when AIC stops decreasing with increasing model complexity

Y: Performance Metric
x_n: Processing/Composition Metrics

Basics of linear regression:

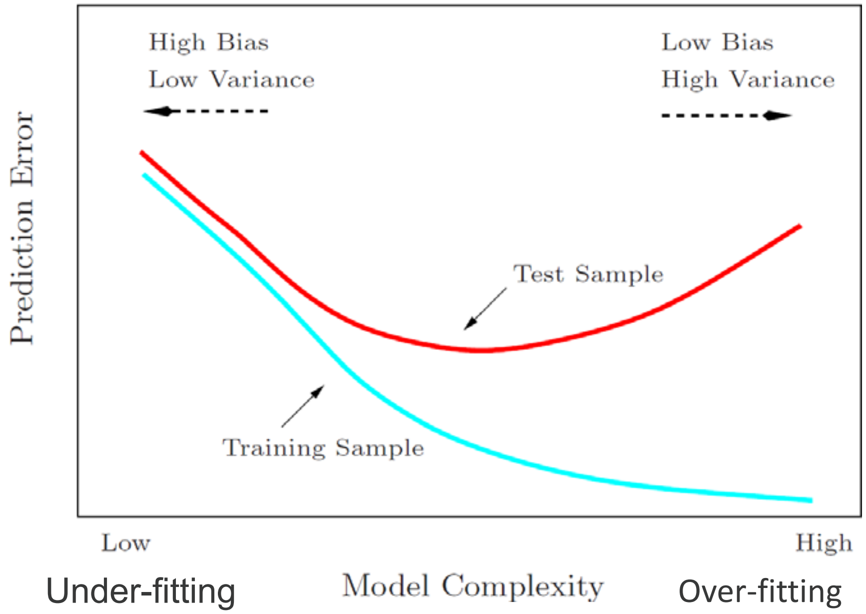
- Multi-variable correlations
- Assumes linear relationships
- Can be generalized to nonlinear correlations

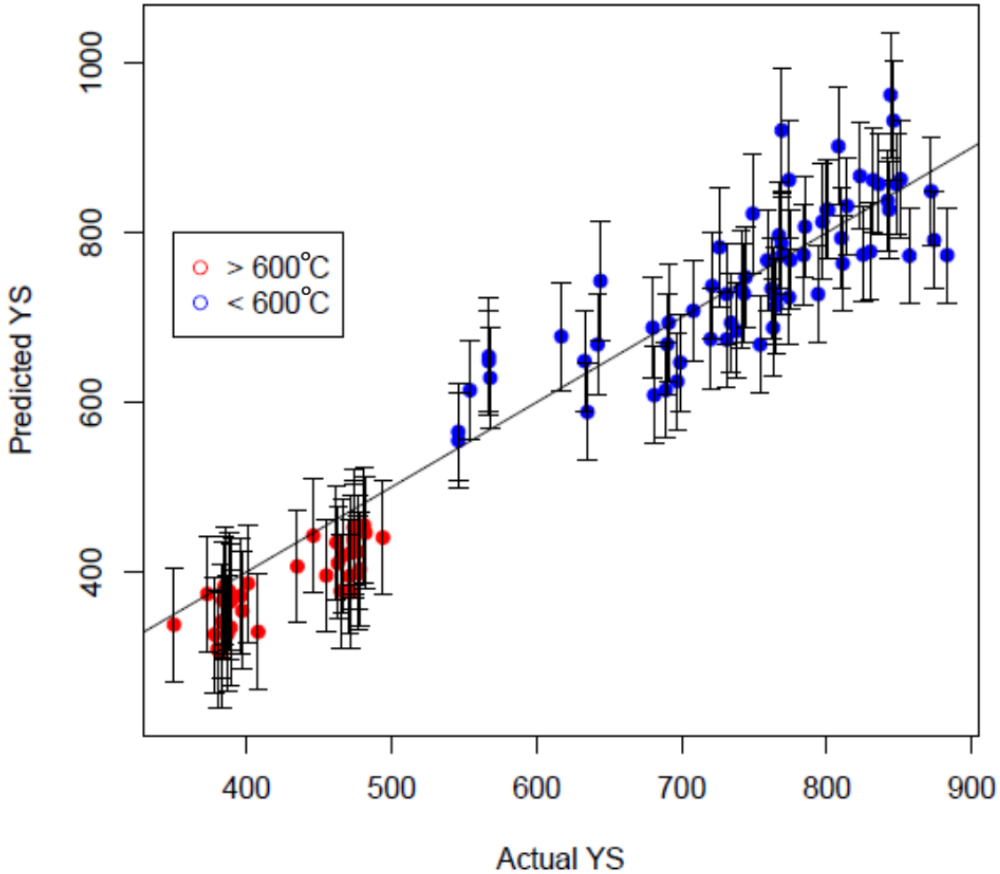
Semi-gSEM

Linear Model Validation

Bias-Variance Trade-offs

Homogenized (Test Dataset) vs Non-homogenized (Training Dataset)

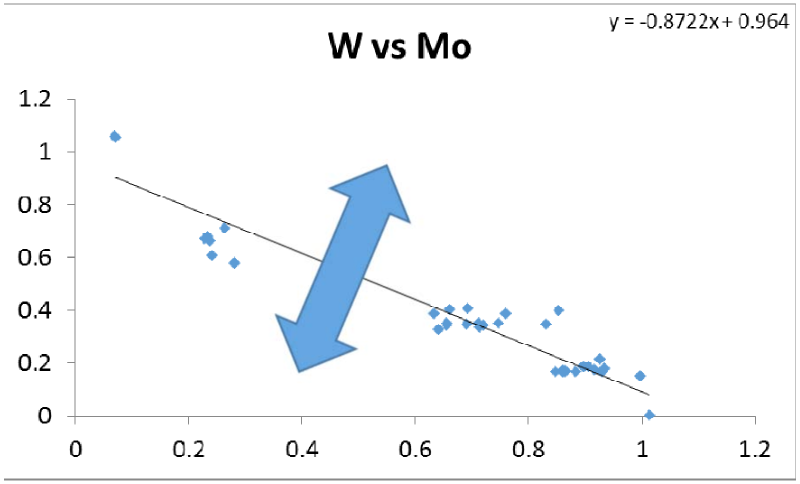




Exploratory Data Analysis

Multicollinearity

Garbage In, Garbage Out:
Data Quality is the Foundation for Good Analytics



Multicollinearity increases standard errors of the regression coefficients. It makes some variables statistically insignificant when they should be significant.

Possible solutions:

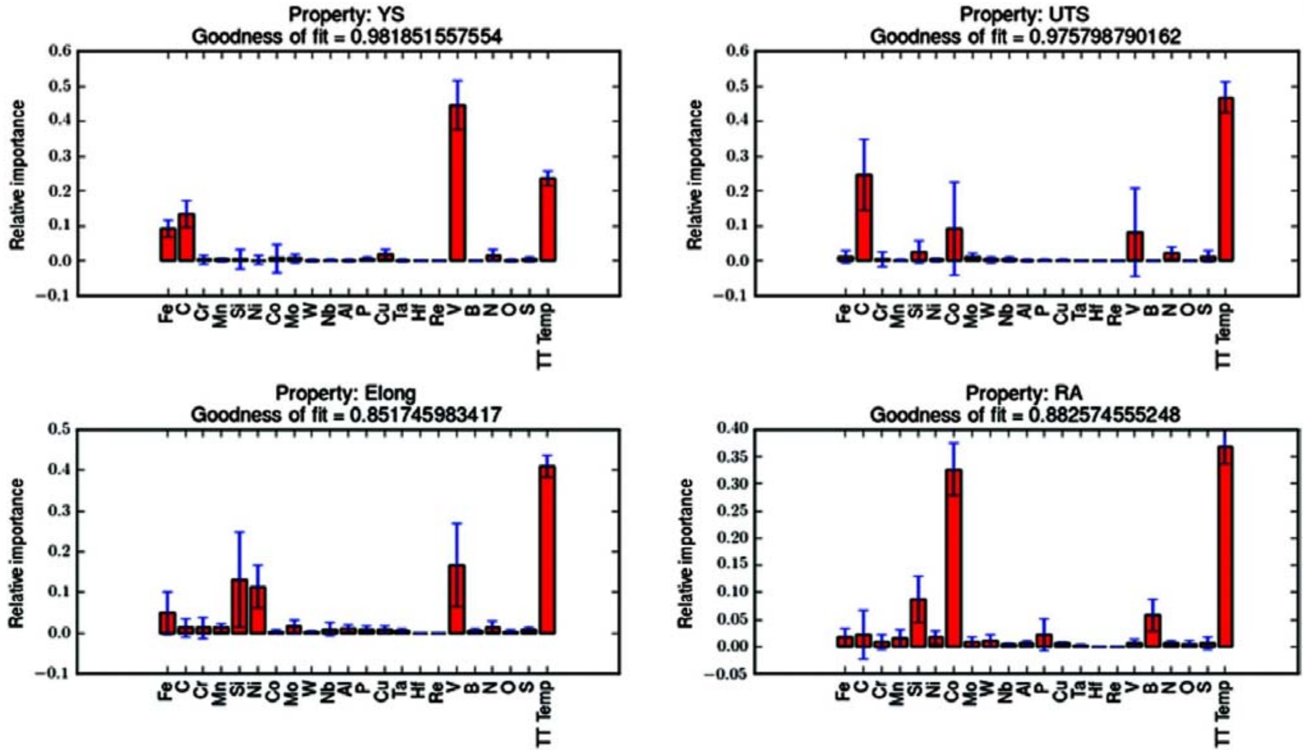
- Stepwise (AIC, p-value) regression or best subsets (adjusted R^2) regression
- Partial Least Squares Regression (PLS) or Principal Components Analysis (PCA) or Nonnegative Matrix Factorization (NMF)

Non-linear, ensemble learning methods

randomForest (Python 3)

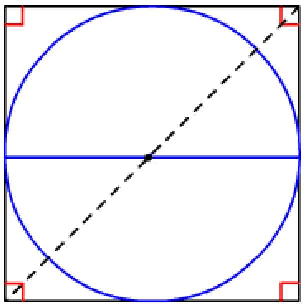
Non-parametric models train decision trees based on large number of data points and specialized sampling techniques (sampling-with-replacement) and do well in regression and classification tasks.

Ensemble learning methods operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees, to correct for decision trees' habit of overfitting to their training set.



Curse of Dimensionality

High-dimensional data



- Naively, each additional dimension doubles the effort needed to try all binary combinations.
- There is little difference in the distances between different pairs of samples (Euclidian equidistance + noise).

Hypersphere with radius r and dimension d, volume:

$$\frac{2r^d \pi^{d/2}}{d\,\Gamma(d/2)}$$

Hypercube with edges of length 2r, volume: $(2r)^d$

Volume ratio: $$\frac{\pi^{d/2}}{d 2^{d-1} \Gamma(d/2)} \to 0$$

CLUSTERING SOLUTIONS:

PLAN "A"

Do dimensionality reduction first

PLAN "B"

$$d(i,j) = \lim_{p\to\infty} \sqrt[p]{|x_{i1}-x_{j1}|^p + |x_{i2}-x_{j2}|^p + \cdots + |x_{il}-x_{jl}|^p} = \max_{f=1}^{l} |x_{if}-x_{jf}|$$

Dimensionality Reduction & Visualization

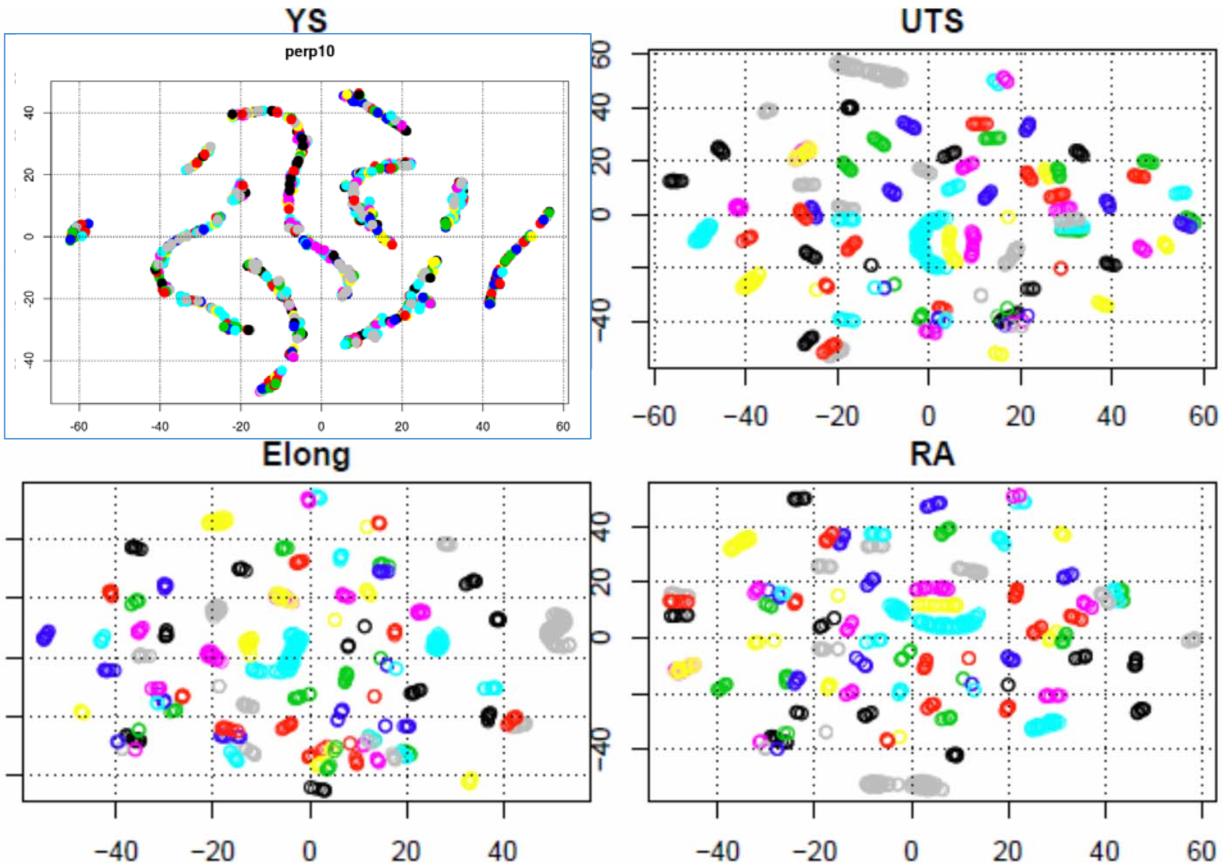
Heavy-tailed symmetric distribution to alleviate problems w/crowding & cost function optimization

- Gradient descent algorithm

t-Distributed Stochastic Neighbor Embedding

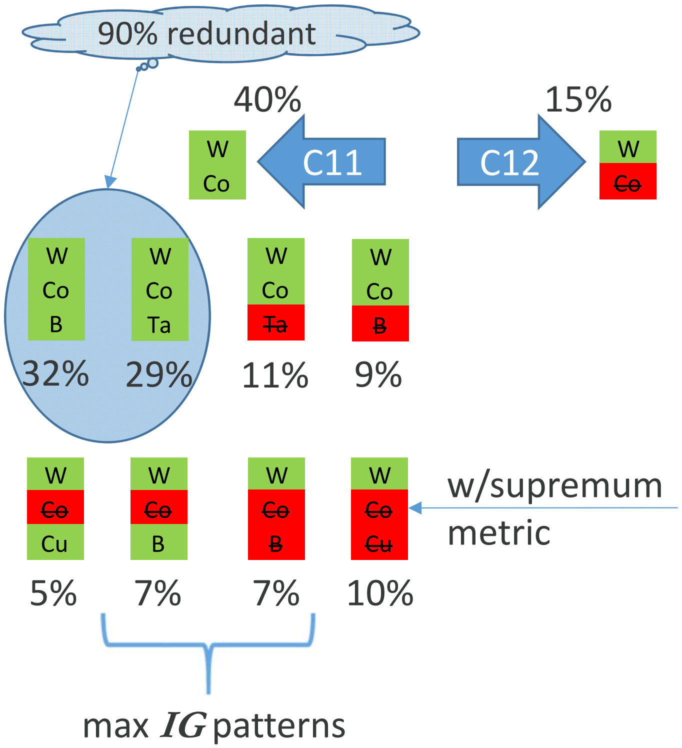
- Induced placement probability
- Preserve neighborhood identity
- Multiple low-D images of object

Perplexity defined as 2 to the power of Shannon entropy (2^H)



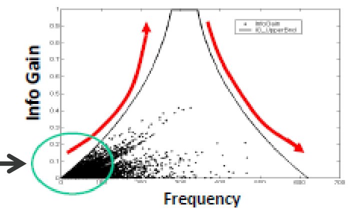
Pattern Discovery

Pattern-Based Classification



Information Gain vs. Pattern Frequency

- Computation on real datasets shows: Pattern frequency (if not too frequent) is strongly tied in with the discriminative power (information gain)



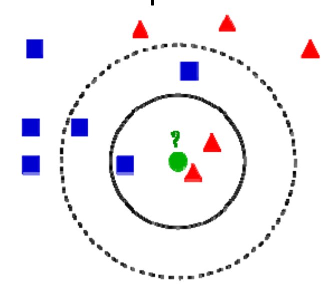
- Information Gain upper bound monotonically increases with pattern frequency
  - **Information Gain Formula:** $IG(C \mid X) = H(C) - H(C \mid X)$

Entropy of given data: $H(C) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

Conditional entropy of study focus: $H(C \mid X) = \sum_j P(X = x_j) H(C \mid X = x_j)$

Extended k-NN (ENN) Algorithms

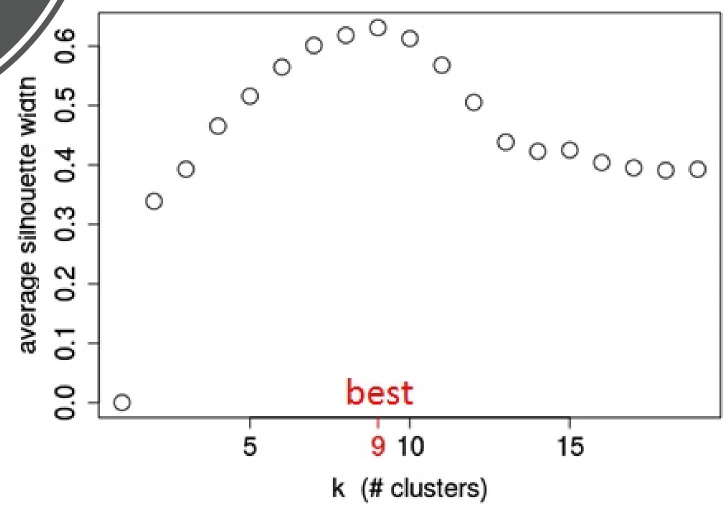
- k-NN (k-nearest neighbors) is a type of instance-based learning, or lazy learning, where the function is only approximated locally (by training instances) and all computation is deferred, until classification or regression in response to query.



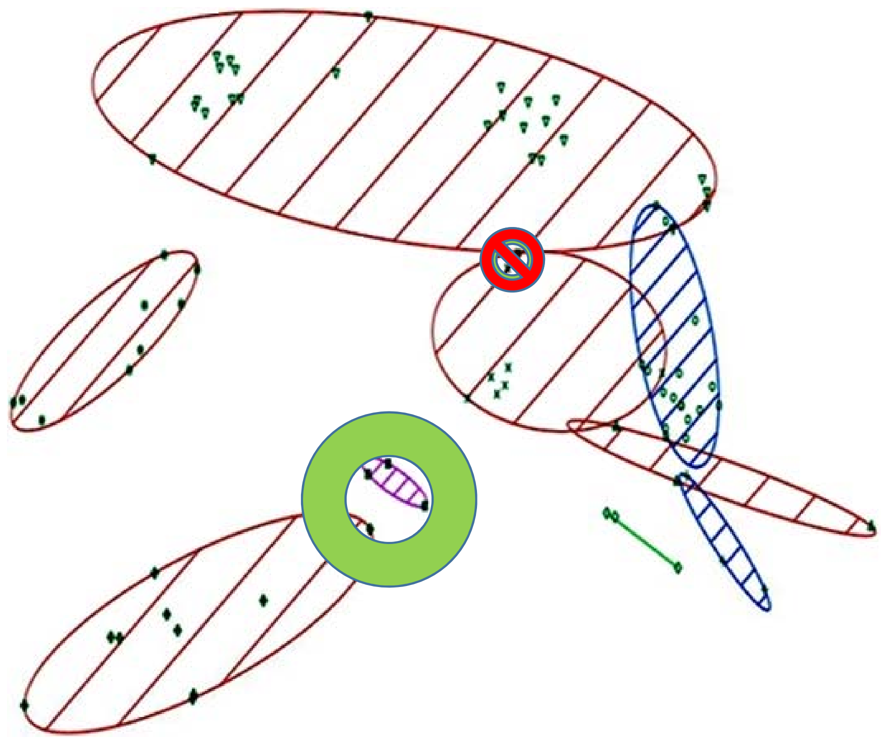
Other Data Mining Algorithms

Partitioning Around Medoids

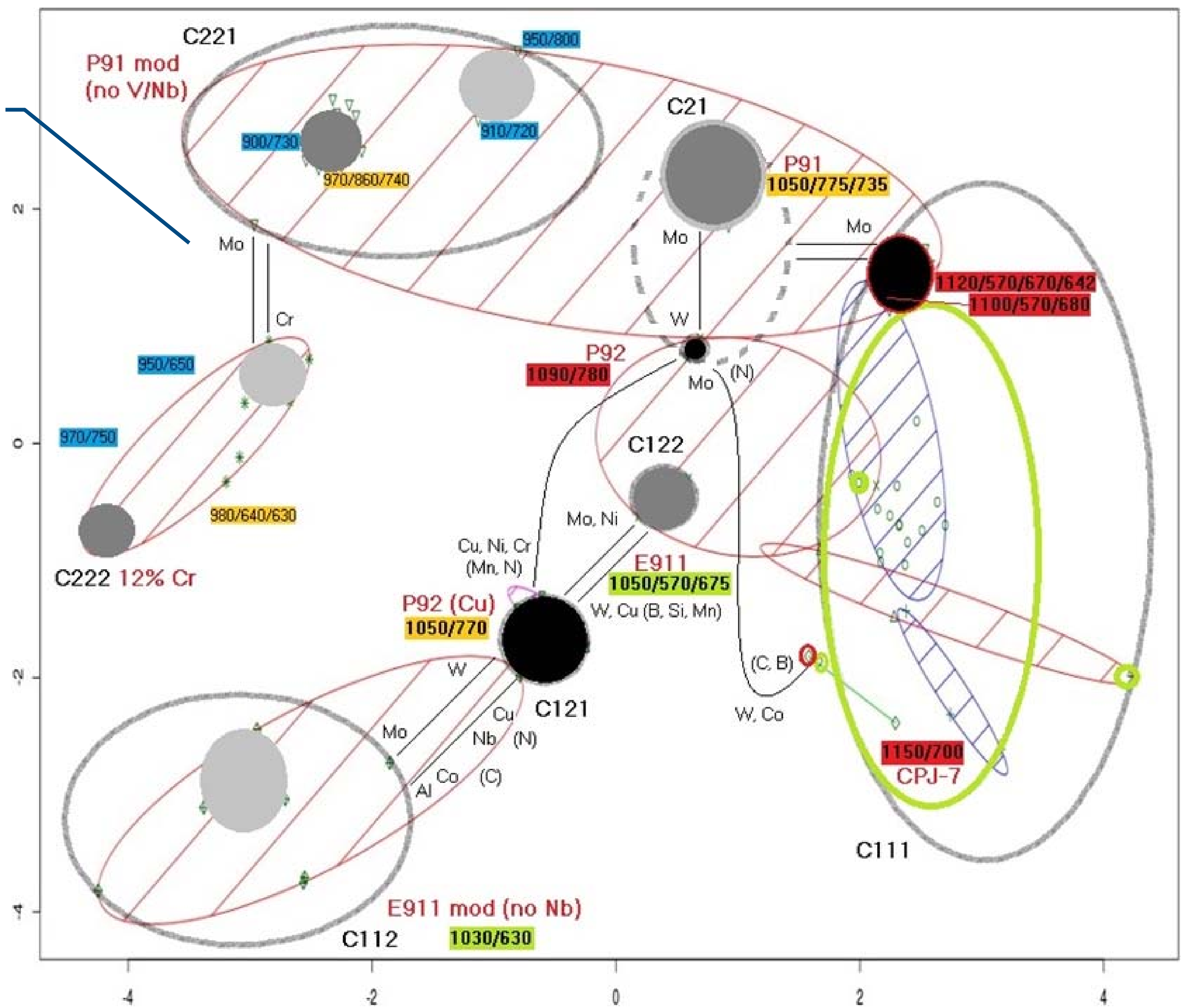
pam() clustering assessment



(ranges from -1 to 1)



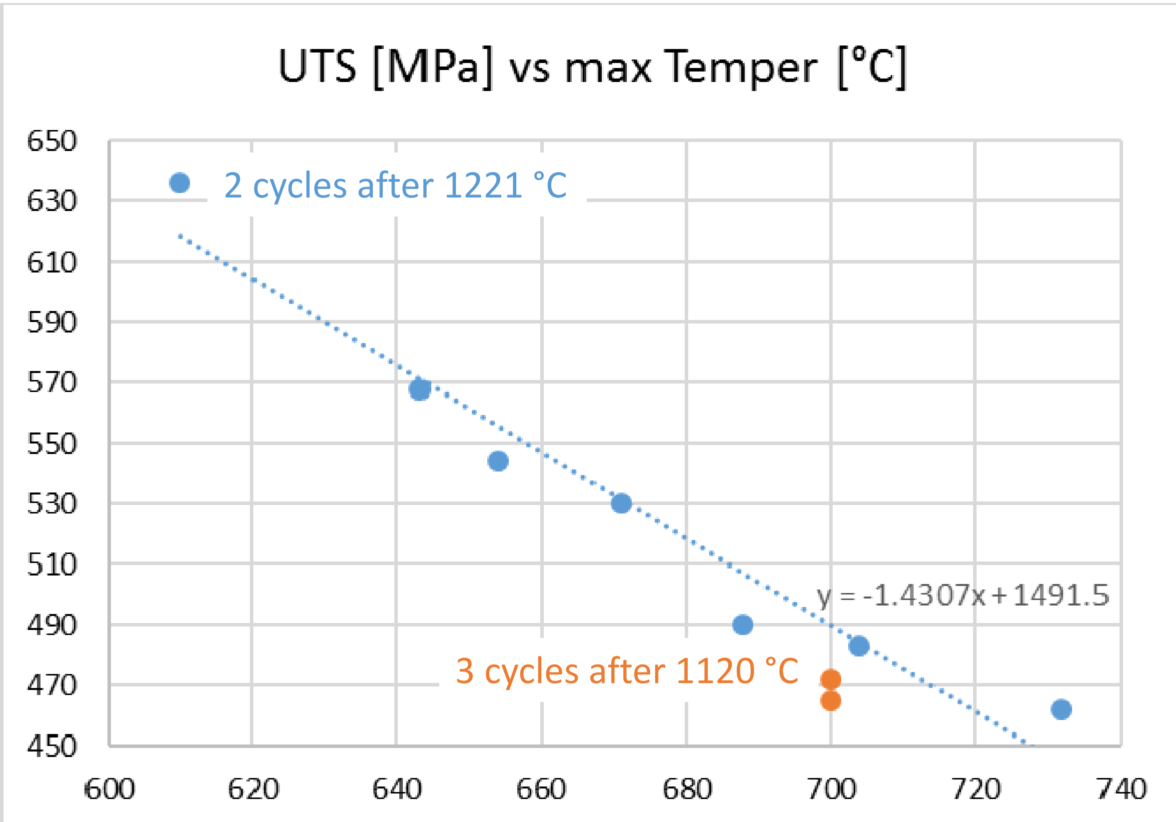
Limited process variability within {no V} cluster



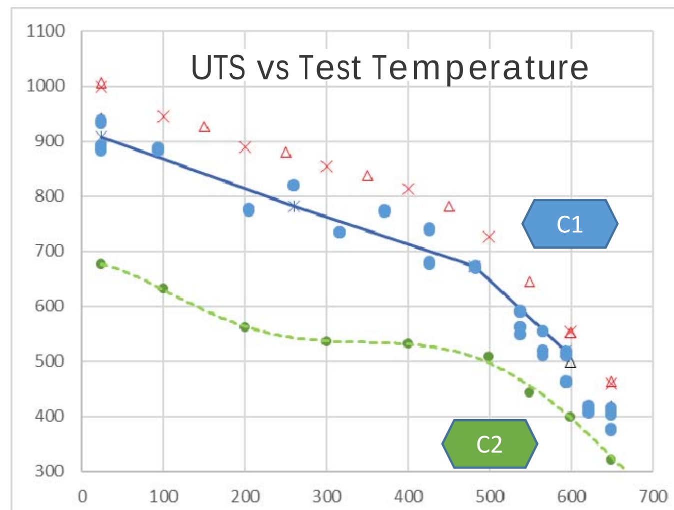
Process is pre-defined by composition

Utilizing “scientific constants”

Composition #37 @593-593.3 °C – Cluster C1(Gate)

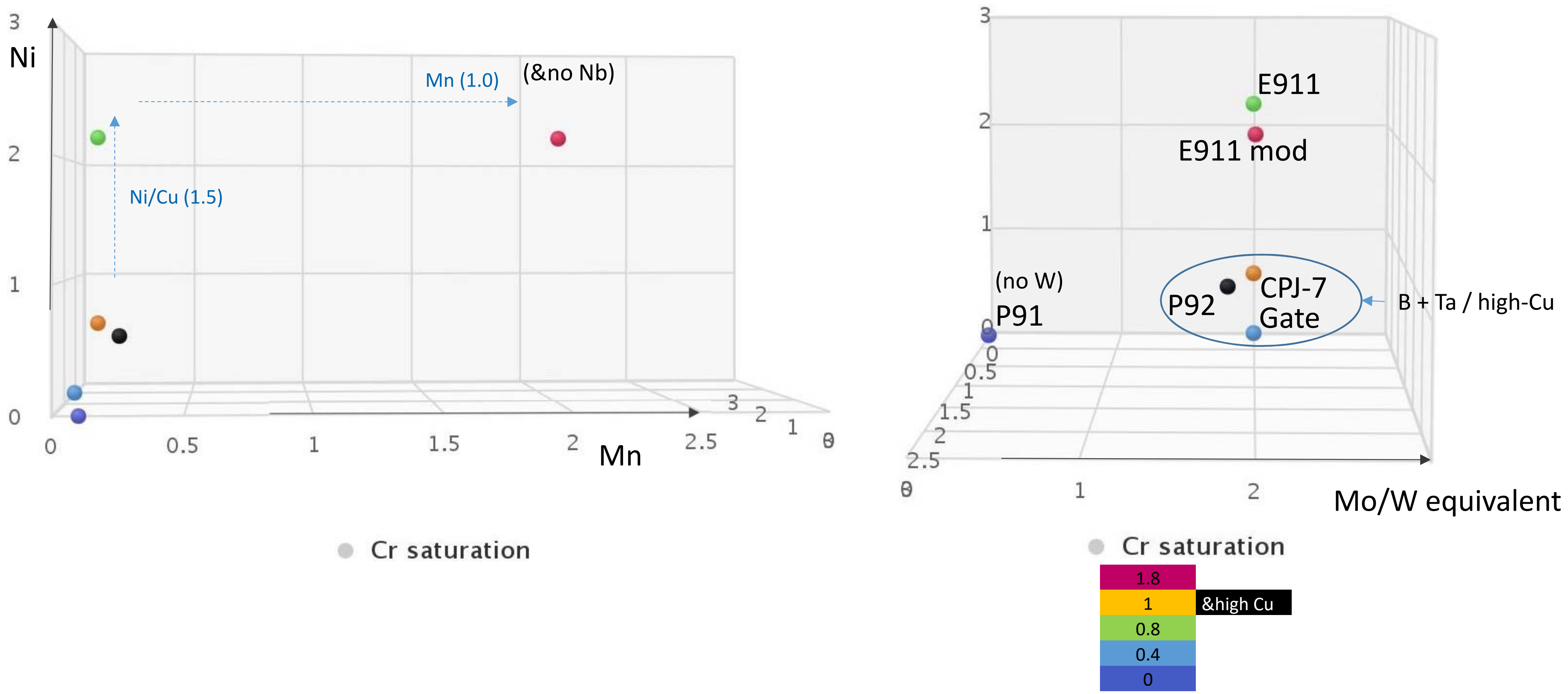


Limited data with control variables can be used to extract ***priors*** for the other subsets



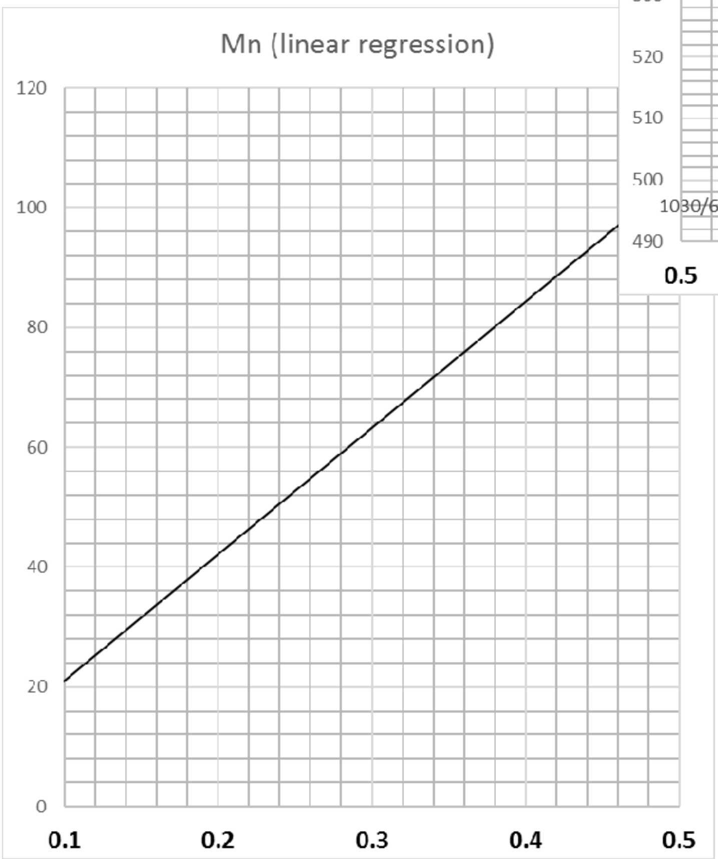
Cluster Separation (Java)

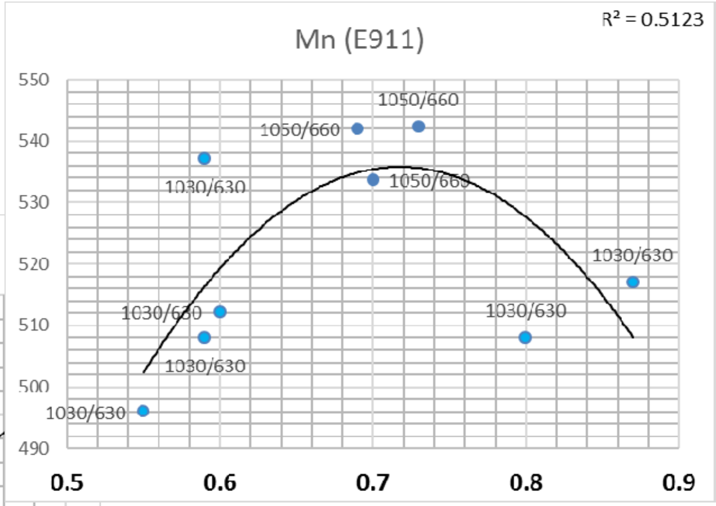
Cr, Mo/W/Co, V/Nb, Ni, Mn



Gate, P92 vs E911

UTS [MPa] vs Mn [%wt.]

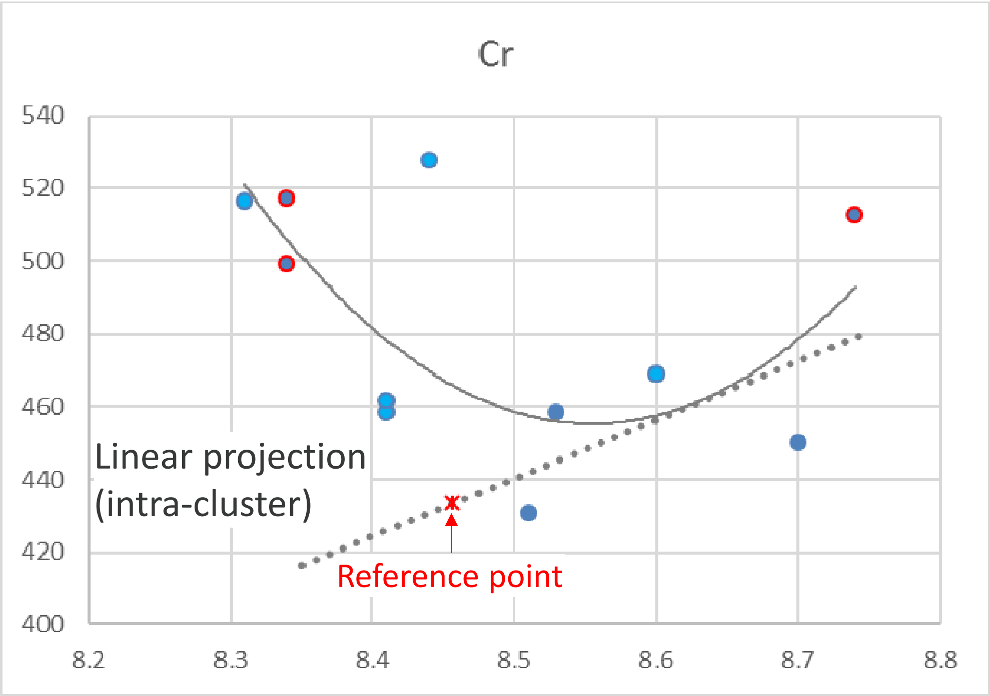


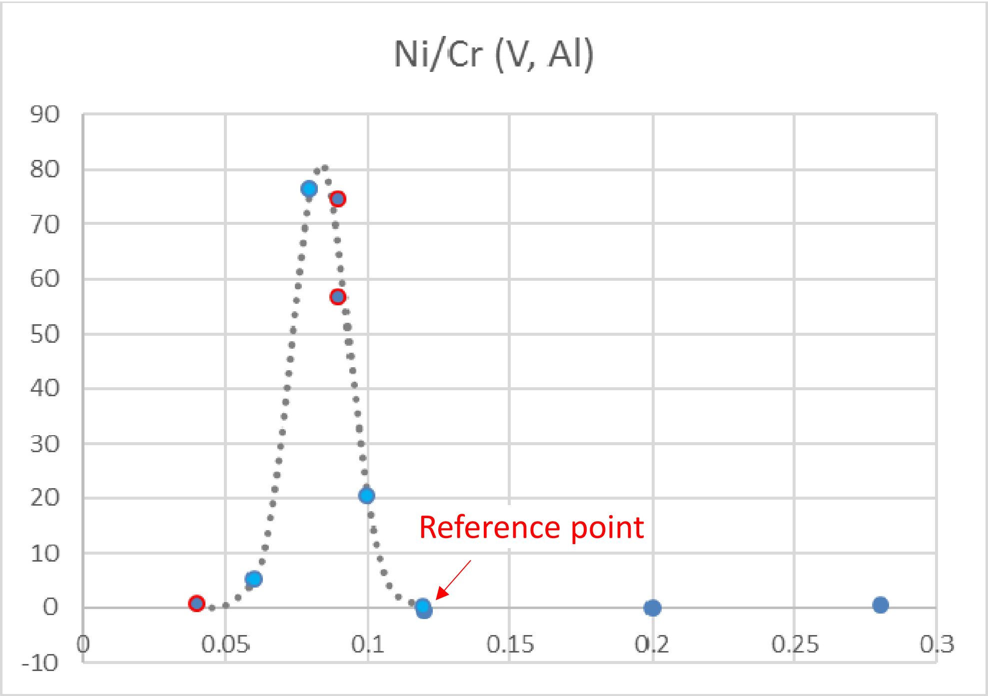


Cluster-based piece-wise polynomial fit reveals the range of applicability for linear regression modeling

Cluster C21 (P91)

UTS [MPa] for low Ni & Cr [%wt.]

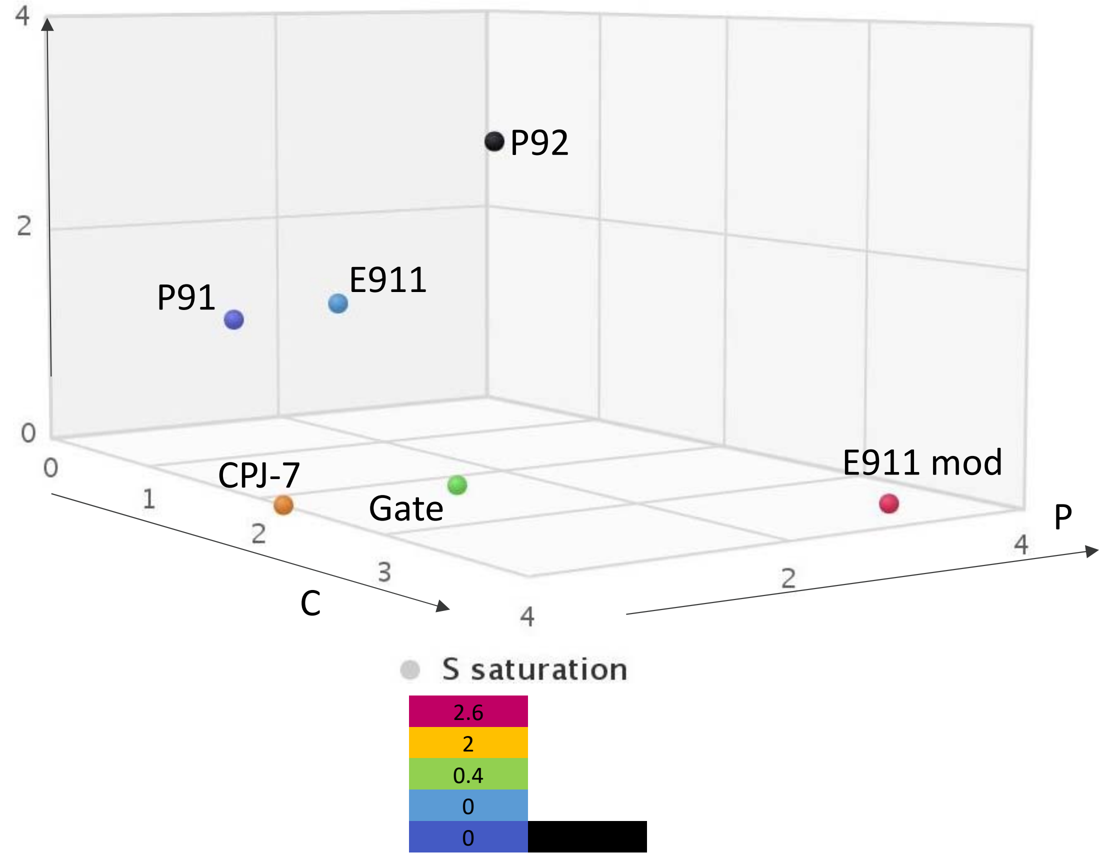


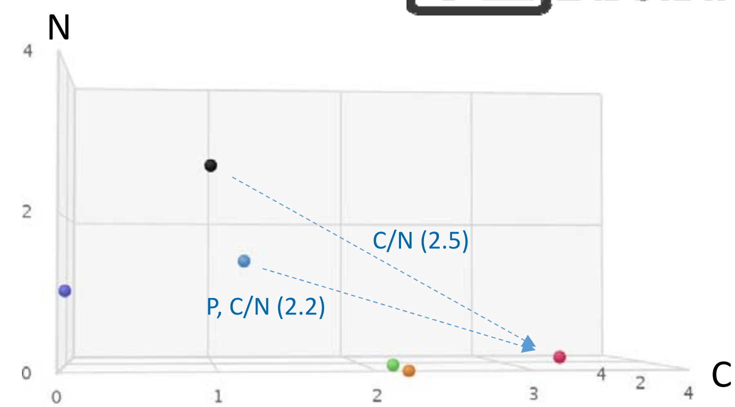


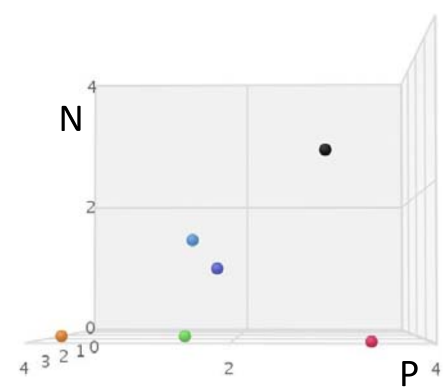
Intra-cluster reference points for locally-linear contributions are adjusted to fit the linear trends for global contributors.

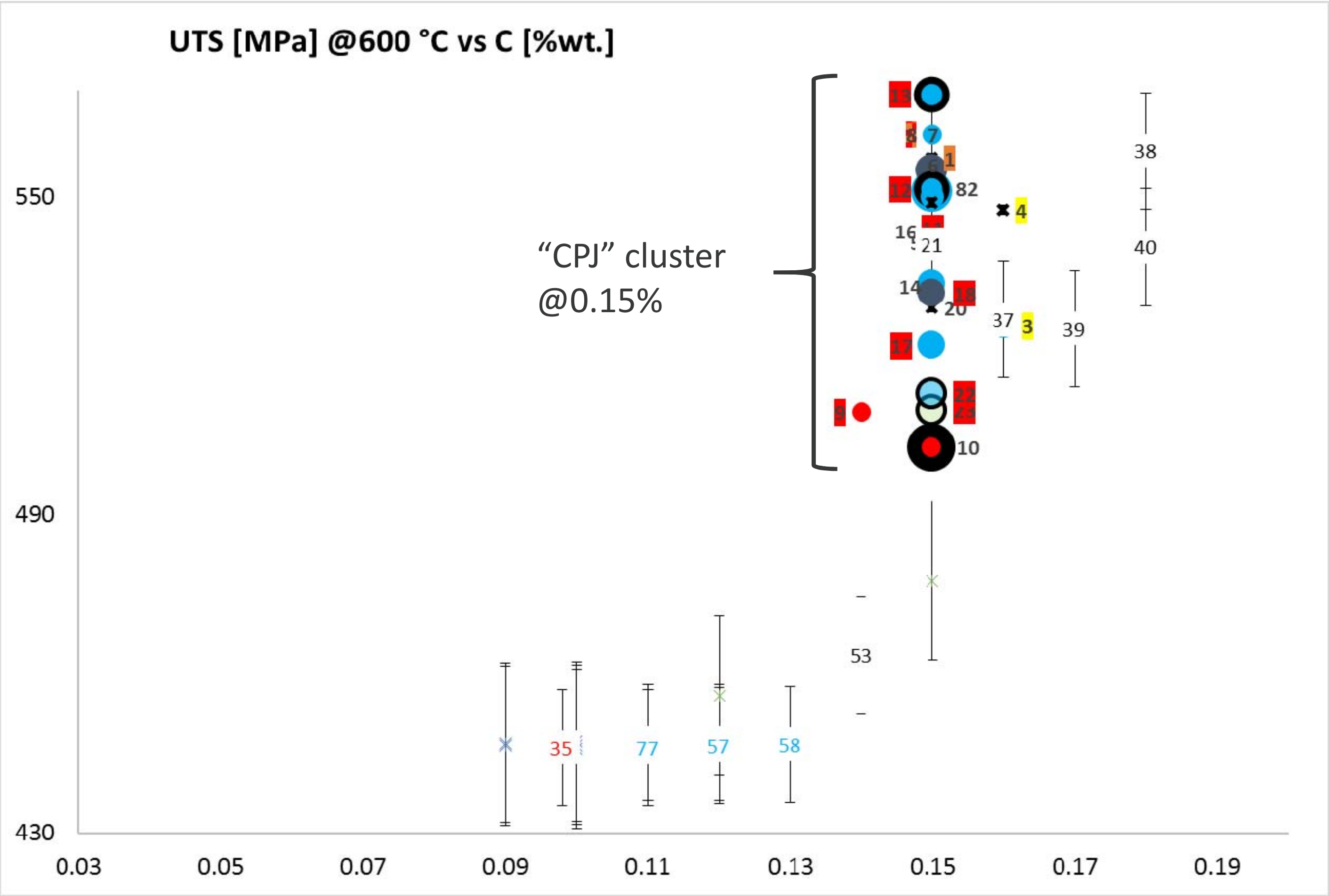
Cluster Separation (Java)

Al, C, N, P, S





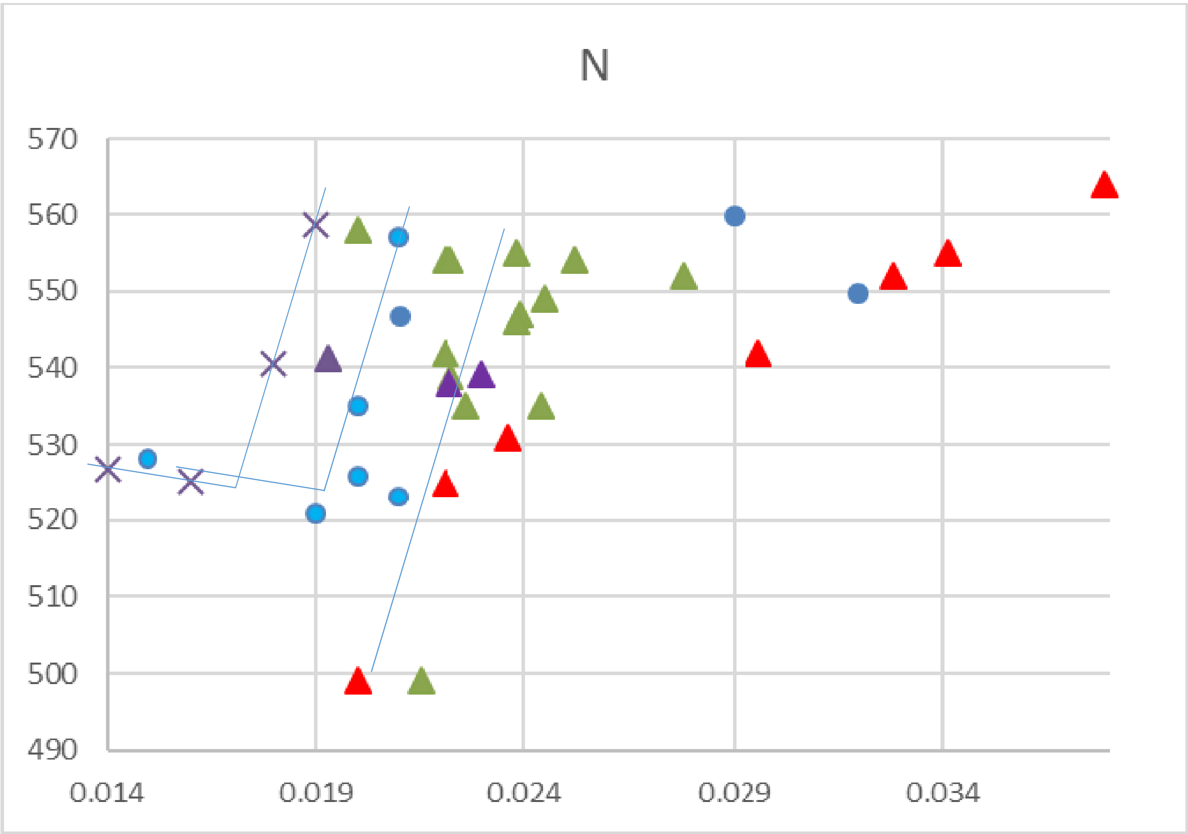




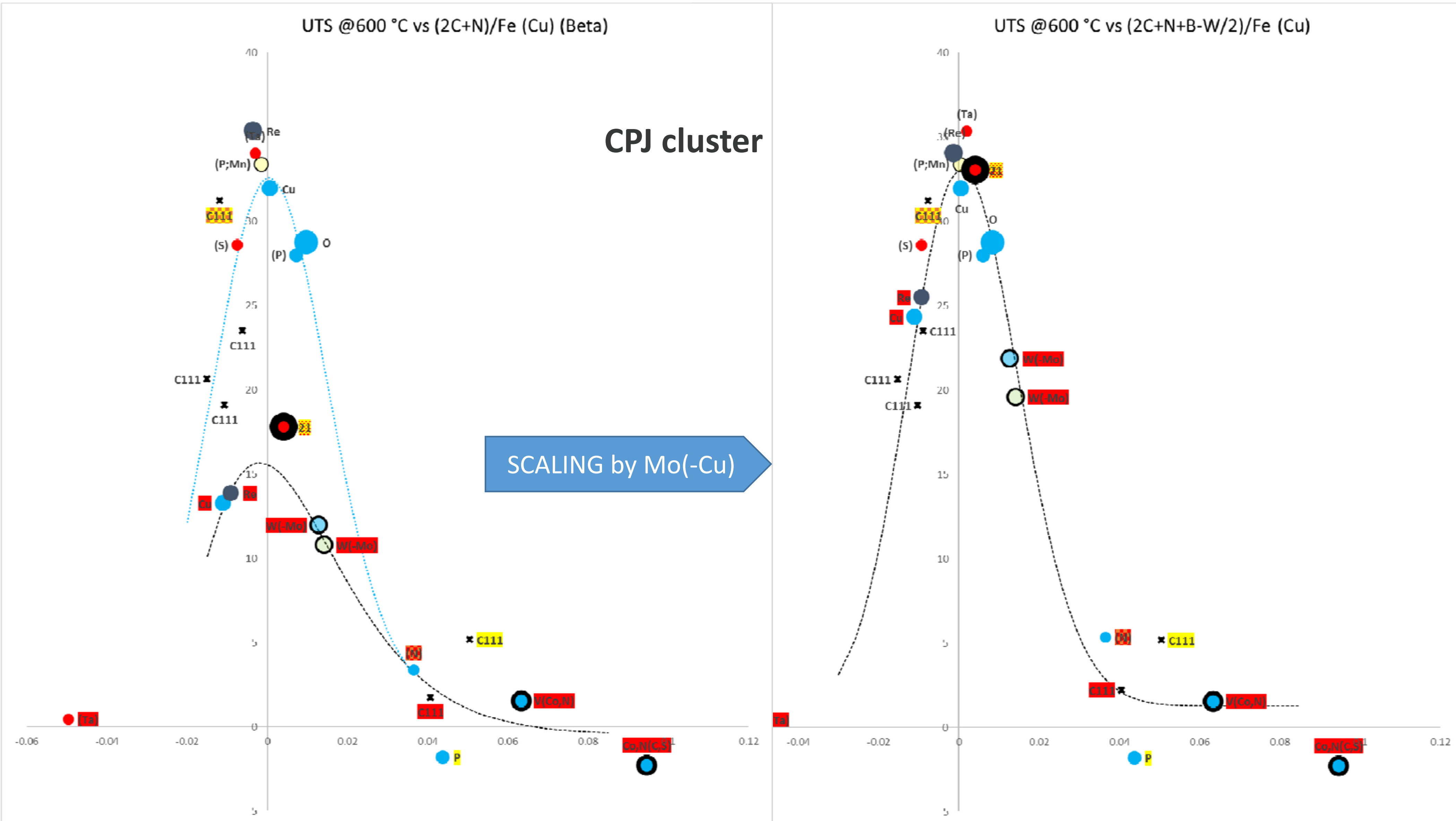
C1 & C21 clusters projected on to “CPJ” cluster median using global regression by Mn and maximum temper T (subsequent to intra-cluster alignments)

Cluster C11

UTS [MPa] vs N [%wt.]

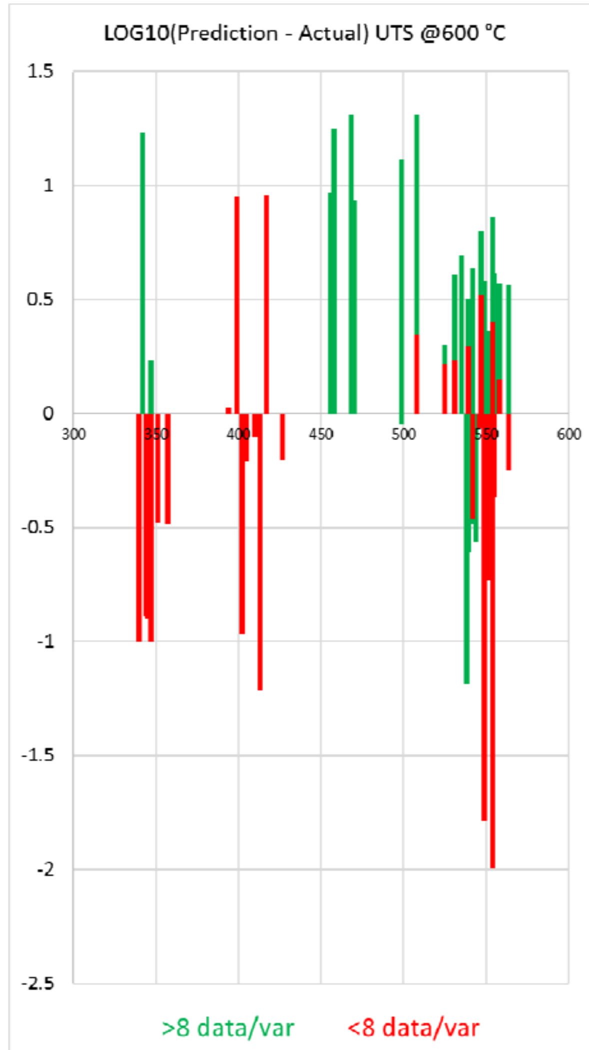


High-fidelity discrimination between the non-linear model adjustments due to secondary contributors is problematics with limited and multi-collinear data.



UTS @600 °C vs (2C+N)/Fe (Cu) (Beta)
CPJ cluster
SCALING by Mo(-Cu)
UTS @600 °C vs (2C+N+B-W/2)/Fe (Cu)

Cluster-based models' cross-validation: Error vs Actual



Experimental error
(by compositions #1, 80)

Precision of data reporting

Overfitting

SUMMARY

Results & Conclusions

- **Heritage data compiled and ingested into "machine-readable" format**
- **Exploratory data analysis (EDA) for models development:**
  - Pair-wise correlation plots indicated strong univariate relationships between compositional variables;
  - Linear regression models explained 85% of the variance in yield strength; subsequent model cross-validation indicated a change in strengthening mechanism at ~500 °C.
- **Ensemble learning methods provided rank-ordering of contributors and illustrated that the models are composition-group dependent.**
- **Cluster analysis confirmed that dependencies between the contributors were a result of a bias around human-derived design parameters.**
- **Non-linear models provide an order of magnitude better fit than global linear models, for the similar number of data points per model variable.**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference therein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed therein do not necessarily state or reflect those of the United States Government or any agency thereof.


NATIONAL ENERGY TECHNOLOGY LABORATORY
Thank You!
Questions?

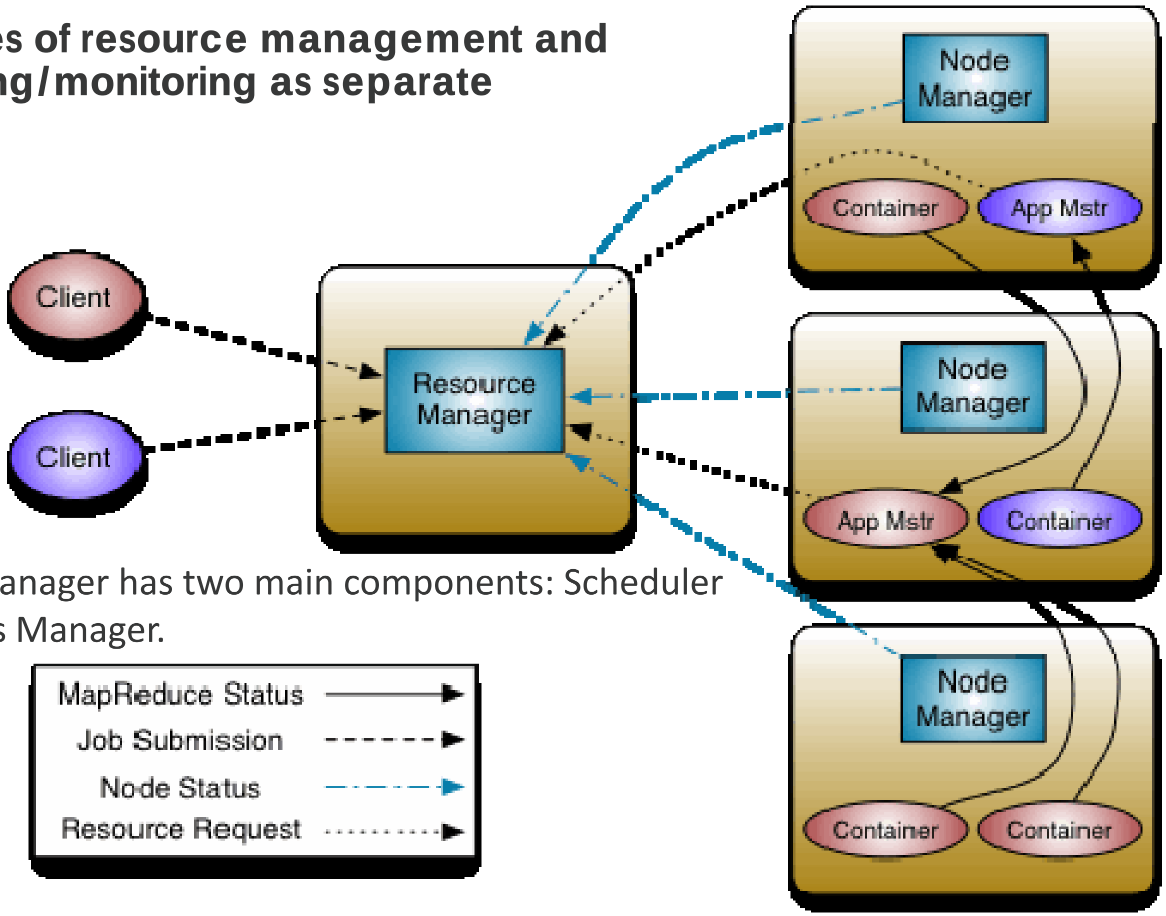
YS correlation with concentration

Test Temperature (°C)	Rank Contributors	adj-R^2 value
RT - 800	V, -Fe, -Si, -Cu, -N, -Co, -P, Nb	0.88
RT - 599	V, -Fe, -Si, -Cu, Cr, -P, Ni, Nb, -N	0.91
600 - 800	B, V, -Cu, Ni, Mo, Nb	0.94

Domain knowledge	Coefficients
Below 600 °C • Cr enhances sub-boundary strengthening • V, Nb enhances precipitation strengthening • Ni aid in solid solution strengthening	Positive
Above and equal to 600 °C • Mo, Ni aid in solid solution strengthening • B reduces coarsening rate • V, Nb enhances precipitation strengthening	Positive

Yet-Another-Resource-Negotiator

Functionalities of resource management and job scheduling/monitoring as separate daemons



The Resource Manager has two main components: Scheduler and Applications Manager.