Materials Informatics for the ICME CyberInfrastructure

Kim F. Ferris
Pacific Northwest National Laboratory, Richland, WA

Dumont M. Jones
Proximate Technologies, Columbus, OH

May 12, 2011

Project ID: LM038

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

Pacific Northwest
NATIONAL LABORATORY

Project Overview

Project Timeline

- Start: Mid-Aug 2009
- Finish: October 2011
- ~2/3 complete (Feb 2011)

Budget

- Total project funding
  - DOE – $350K
- FY09 Funding - $60K
- FY10 Funding - $170K
- FY11 Funding - ~$120K

Technology Gaps/Barriers

- Greater materials diversity / joining dissimiliar materials
- Information Gaps in ICME Cyberinfrastructure
  - Data sufficiency
  - Incomplete structure among partners
  - Completeness and availability of repositories
- Mg Processing Model Development
  - First principles to Processing - multiscale
  - Mg-sheet forming

Partners

- Ford Motor Company (John Allison)
- Mississippi State University (Mark Horstemeyer, Tomasz Haupt)
- Canada Center for Mineral & Energy Technology (CANMET) (Kevin Boyle)
- University of Virginia (Sean Agnew)
- Metals Bank – Korea (YM Rhyim)
- ESPEI Workgroup (Liu, Wolverton, Haupt)
- Joining Materials Knowledge (Beckham)

Project Overview: Project's place in the cyberinfrastructure

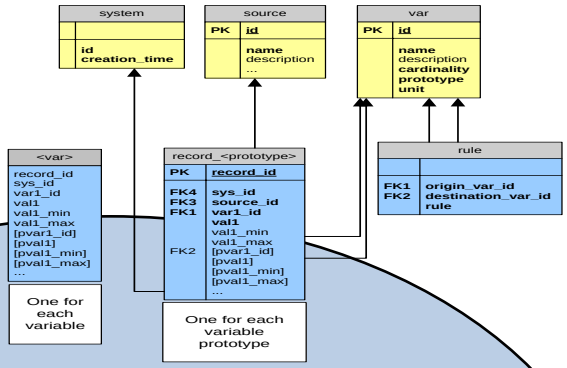
Repository

- Federated/consolidated
- Data exchange support
- ISO 10303 STEP support

1

Analysis DB

- Fills similar role to data warehouse
- Informatics storage/retrieval focused
- Analytic platform, not a re-formed repository



Current Programs

3

2

This Informatics Effort

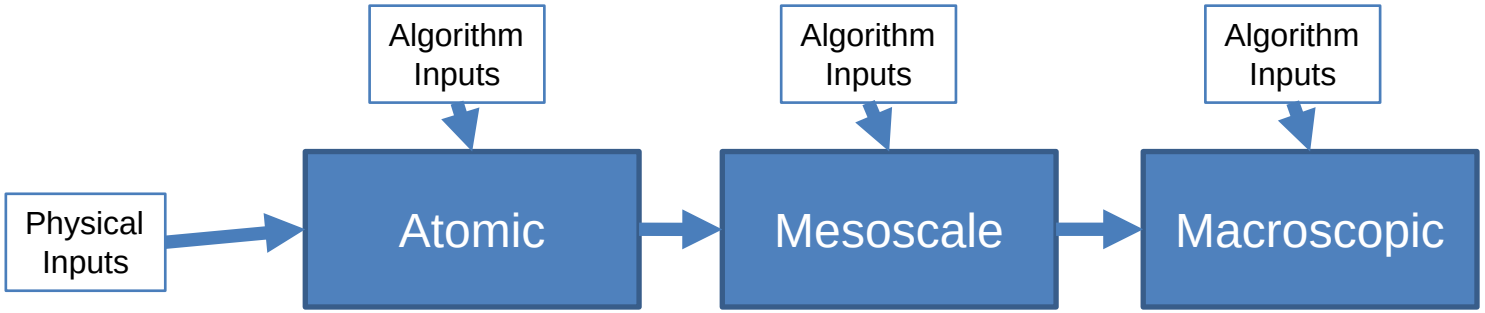
Informatics repository augmentation

- Property correlations
- ID feature space for complex data such as processing sequences
- When can data from distinct systems be safely fused
- Coverage and lineage anomalies
- Insufficient information for prediction (predictor diversity)
- Property definition inconsistencies
- Material definition inconsistencies

Pacific Northwest NATIONAL LABORATORY

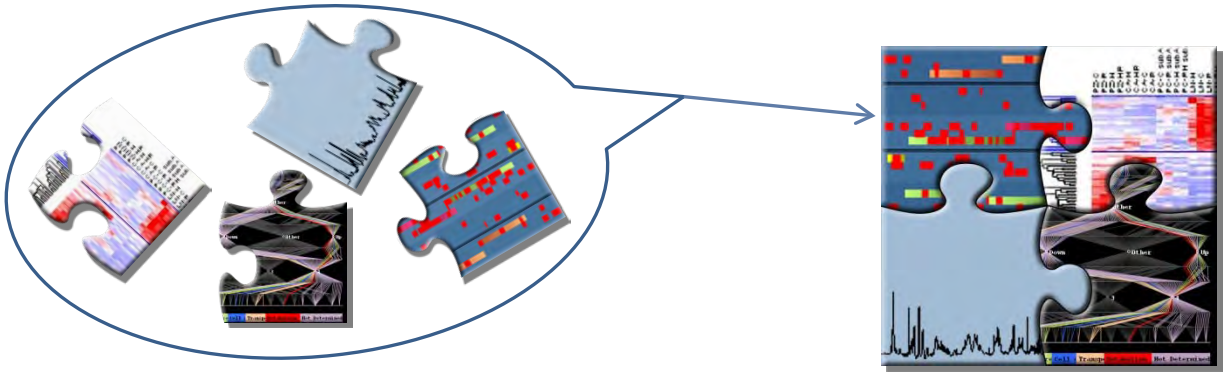
Overview: Basic Tasks & Motivation

1) As parameters are passed from scale to scale, do models adequately represent each scale and what information is lost?



2) How can the 'model' as a whole be verified; reliable data for verification

3) Data from different sources needed for greater variety of materials and joining



4) Does data support design?

Pacific Northwest
NATIONAL LABORATORY

Project Objectives

Challenges:

- Greater diversity of materials in design problems
- Reliable information needed for new materials and model development, and verification
- Mg-alloy performance properties can conflict. Use data-driven design/informatics to address this challenge?

Objectives:

- Develop materials informatics techniques for robust model development
- Identify gaps in the fundamental information contained in the ICME knowledge infrastructure
- Augment knowledge –address gaps/barriers to developing Mg alloys
- Collaborate with current repositories on knowledge content

Pacific Northwest
NATIONAL LABORATORY

Milestones / Deliverables

Month/Year	Milestone / Deliverables
Aug 2009	Program start. ✓
Feb 2010	Milestone: Assess data support needs within context of workgroup research. Deliverable: Preliminary toolkit and data-storage design, assessment of data support. ✓
Sep 2010	Milestone: Toolkit for informatics data assessment Preliminary Mg-alloy data assessment * Deliverables: Create ✓ and deliver toolkit for lightweight alloys. Validated tools against prototype alloy design data ✓.
Jun 2011	Milestone: Extend work to multiple repositories Deliverable: Apply toolkit to address global data sufficiency.

Pacific Northwest
NATIONAL LABORATORY

Technical Approach

Task 1

Informatics Toolkit Design

- Contextual assessment of data support for workgroup models development
- Identify algorithms and storage approach for materials knowledge retention and generation

Task 2

Informatics Toolkit Development

- informatics toolkit for data sufficiency/completeness, support for model development, and knowledge augmentation
- Analytic support for cyberInfrastructure

Task 3

Assessment of data support for model development

- Contextual assessment of data support compatible with heterogeneous sources
- Collaborative contact with data sources (MSSU, CanMet, MetalsBank, DataSoc)

Pacific Northwest
NATIONAL LABORATORY

Accomplishments & Impacts

Assessed data support for first principles development and materials repository

→ Design of prototype data schema
Knowledge retention model for workgroups

→ Guidance to First Principles and Cyberinfrastructure Workgroups for data completeness

Data sufficiency – informatics tools to assess modeling power of finite data sets

→ Data analytics for first principles and sheet forming groups for *robust* models

Tools to assess repository support for information-driven design

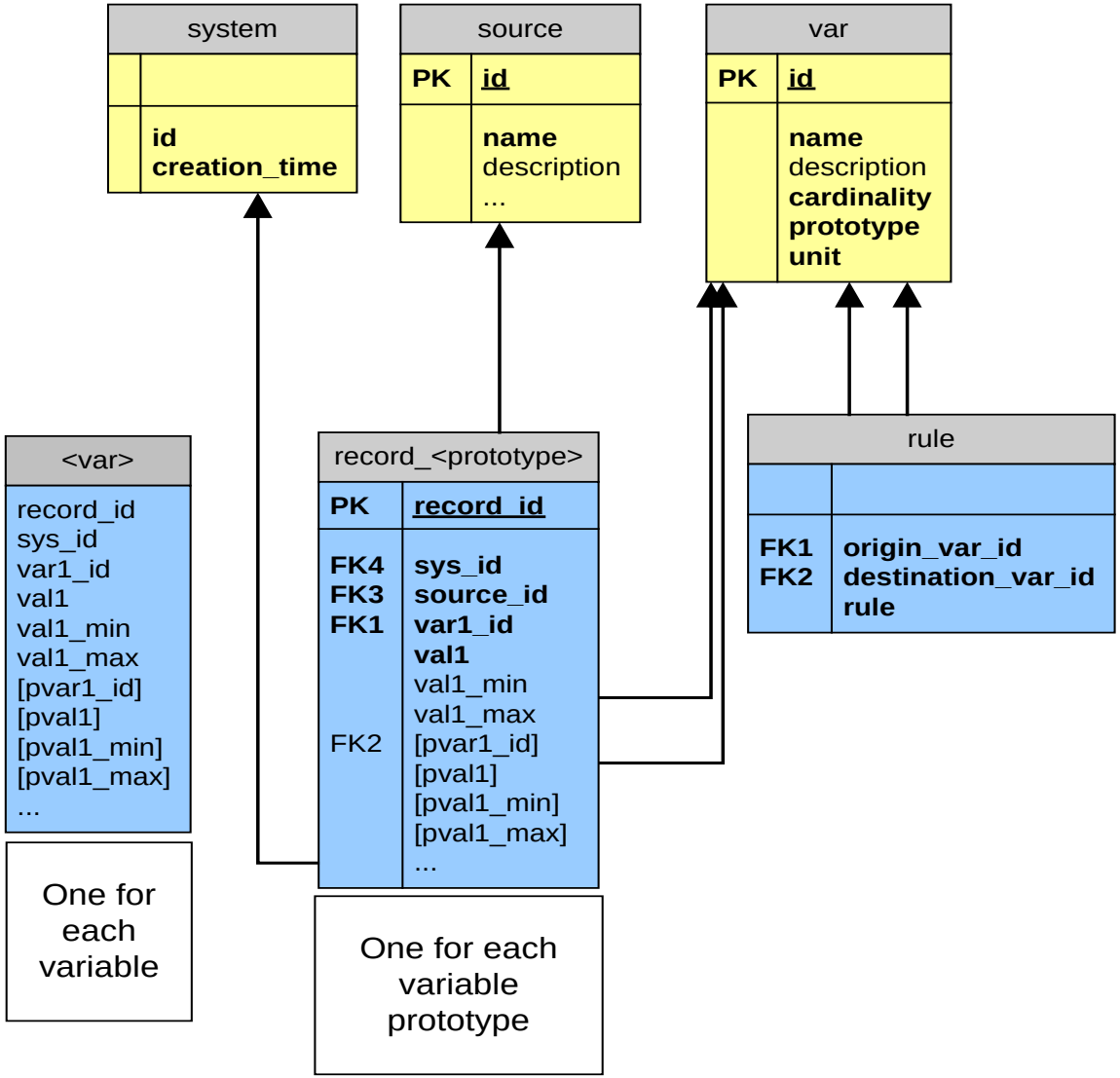
Knowledge content of multiple information sources

- Repository/repository
- PI/repository
- PI/PI

→ Guidance on integrating ICME applications with repositories and cyberinfrastructure

Pacific Northwest
NATIONAL LABORATORY

Accomplishment #1: Analysis Database Schema



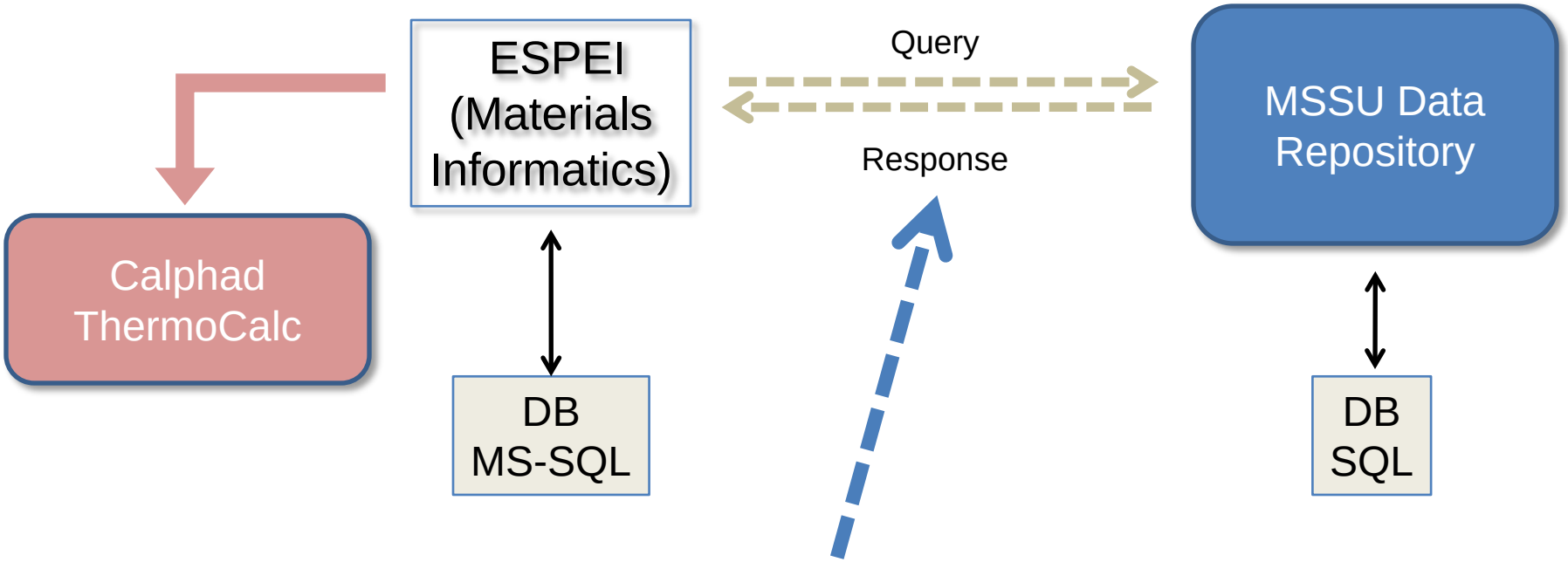
- Developed prototype schema as example for knowledge retention
- Identify key weaknesses in conversion from conventional storage
- Strongly-typed records
- Formed design constraints for information transport

Ferris, KF, and DM Jones. 2010. "**Assessing Data in Support of Structure-Processing-Property Relationships in Mg-alloy Design**" TMS 2010.

Jones, DM and KF Ferris. 2010. "**Generic Materials Property Data Storage and Retrieval for Alloy Material Applications,"** TMS 2010

Technical Accomplishment 2: Enabling the Phase Diagram Infrastructure to MSSU Repository

First principles workgroup: Bridging first principles calculation to phase diagrams

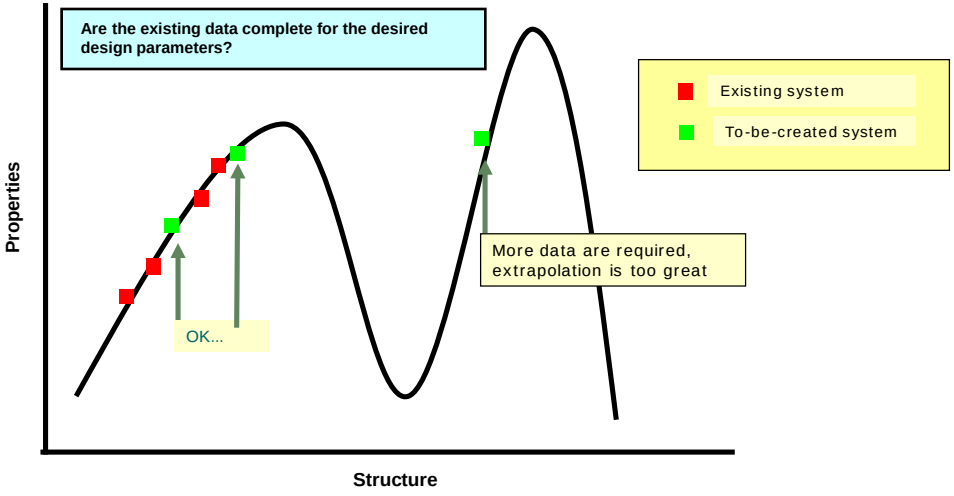


PNNL Contribution

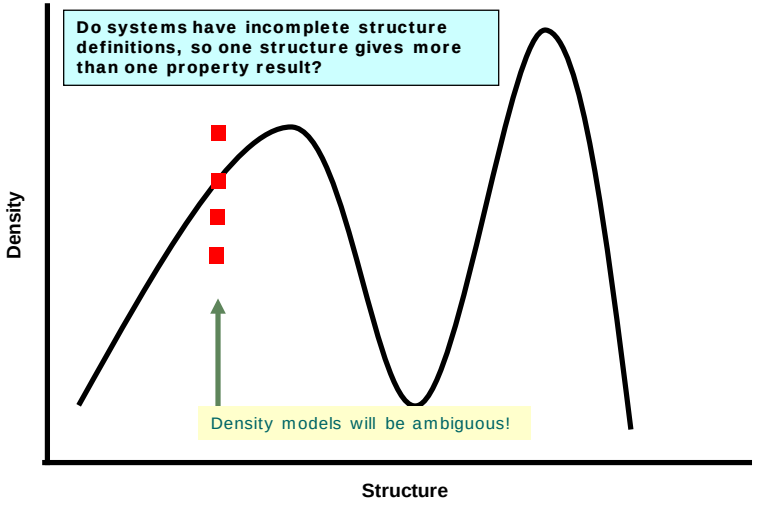
- Informatic effort provided Query/Response 'glue' for ESPEI, MSSU Cyberinfrastructure, first principles, Calphad.
- Automated procedure, retaining critical information for data
- 'Glue' formalism applicable to other information joining situations

Technical Accomplishment 3: Tools to Assess Data Completeness

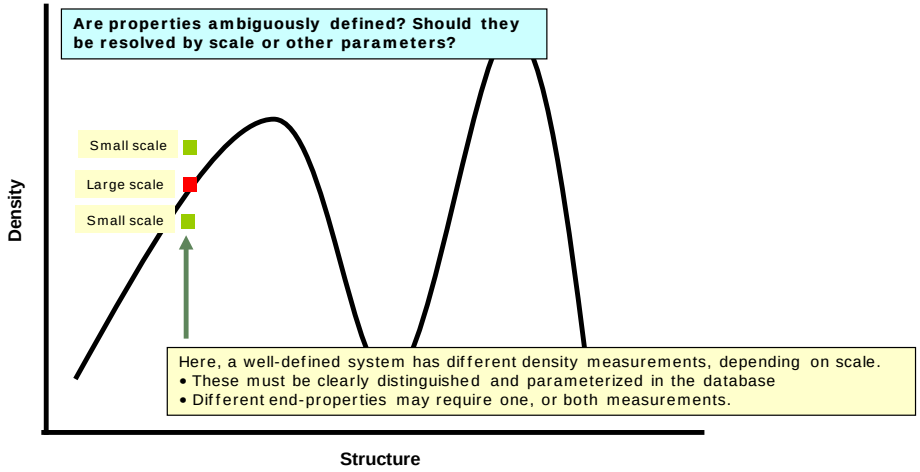
Data Completeness



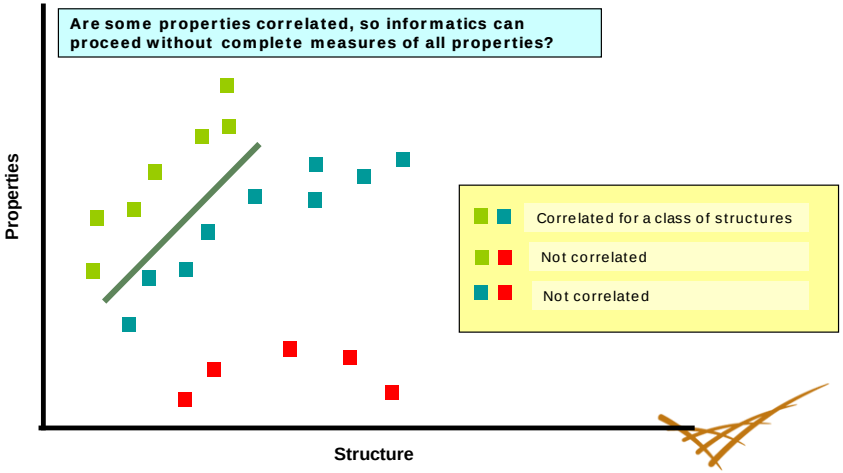
System Definition



Property Definition & Scale



Property Correlation



Pacific Northwest NATIONAL LABORATORY

Data alone do not guarantee successful information-driven design.

Incomplete System Definition Can Lead to Incorrect Conclusions

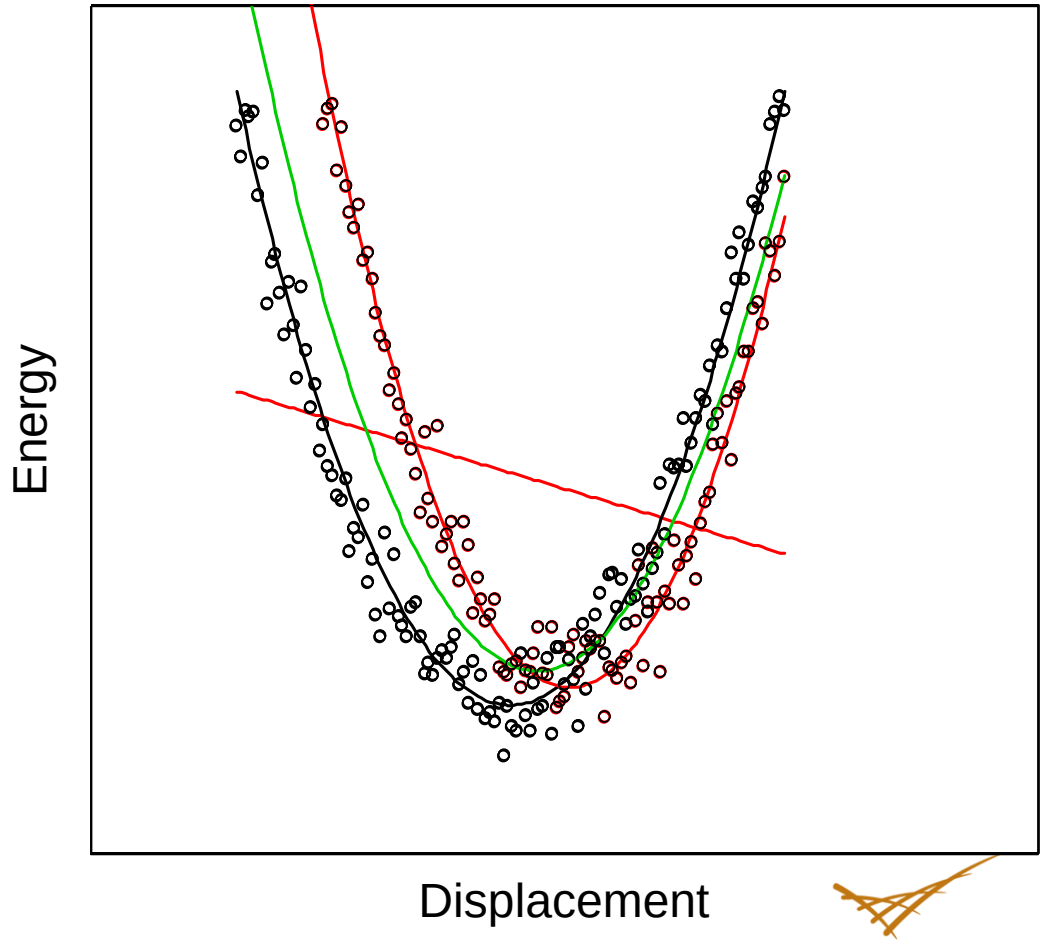
Example: Predicting mechanical constants from individual data points

- Harmonic behavior
- Two different phases w/ different responses

Knowing there are two different groupings is critical to correct predictions/models/behaviors

Not having enough information about this data may lead to –

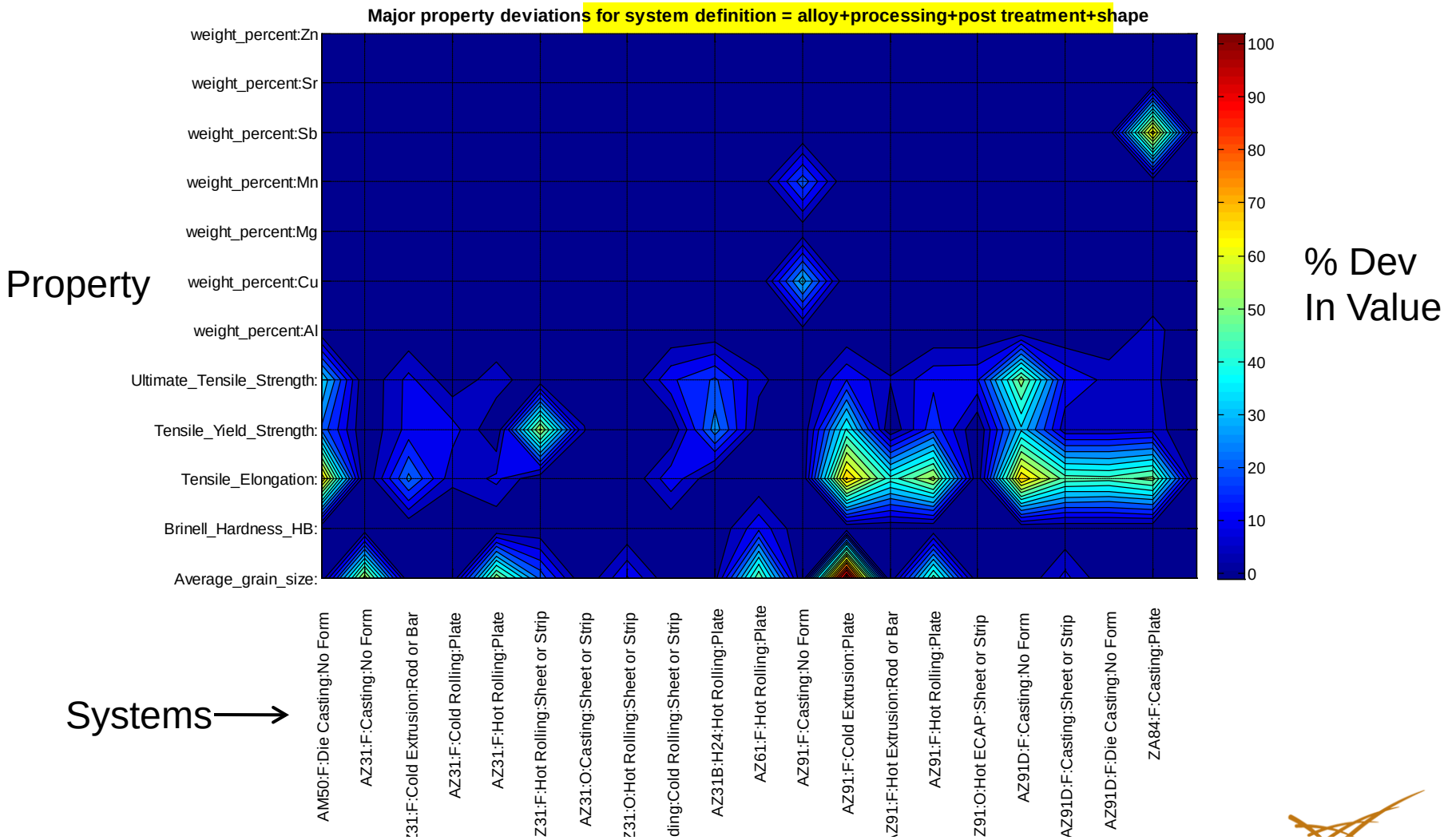
- Linear fit
- Quadratic model for average behavior
- Averaging non-equivalent system (two phases)
- Greater property value uncertainties; validation/verification difficulties



Step-through animation sequence illustrates challenge

Pacific Northwest
NATIONAL LABORATORY

Property Value Deviations Resulting From Incomplete System Definition (Metals Bank)



- **Illustration: % Deviation in Property Values for 'Nominally' Equivalent Systems**
- **Problem: Improperly joined information not useful for verification and validation**

Tools for data assessment (Metals Bank):
Contribution of system definition components to property error

Principle: the importance of a system definition component relates to the observed increase in error when that component is removed.

Example: for density, observed error using alloy+shape as a system definition is nearly the same as that for alloy alone.

Shape is not a crucial system definition component *for density.*

Create reduced system definitions and their associated error for each measured property by aggregation

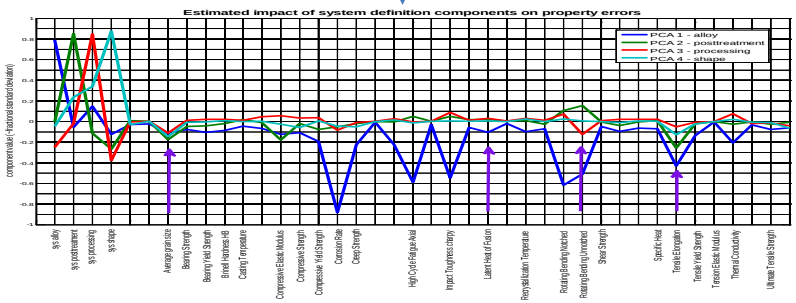
↓

Principal components analysis, using indicator variables to describe the reduced System definition components

↓

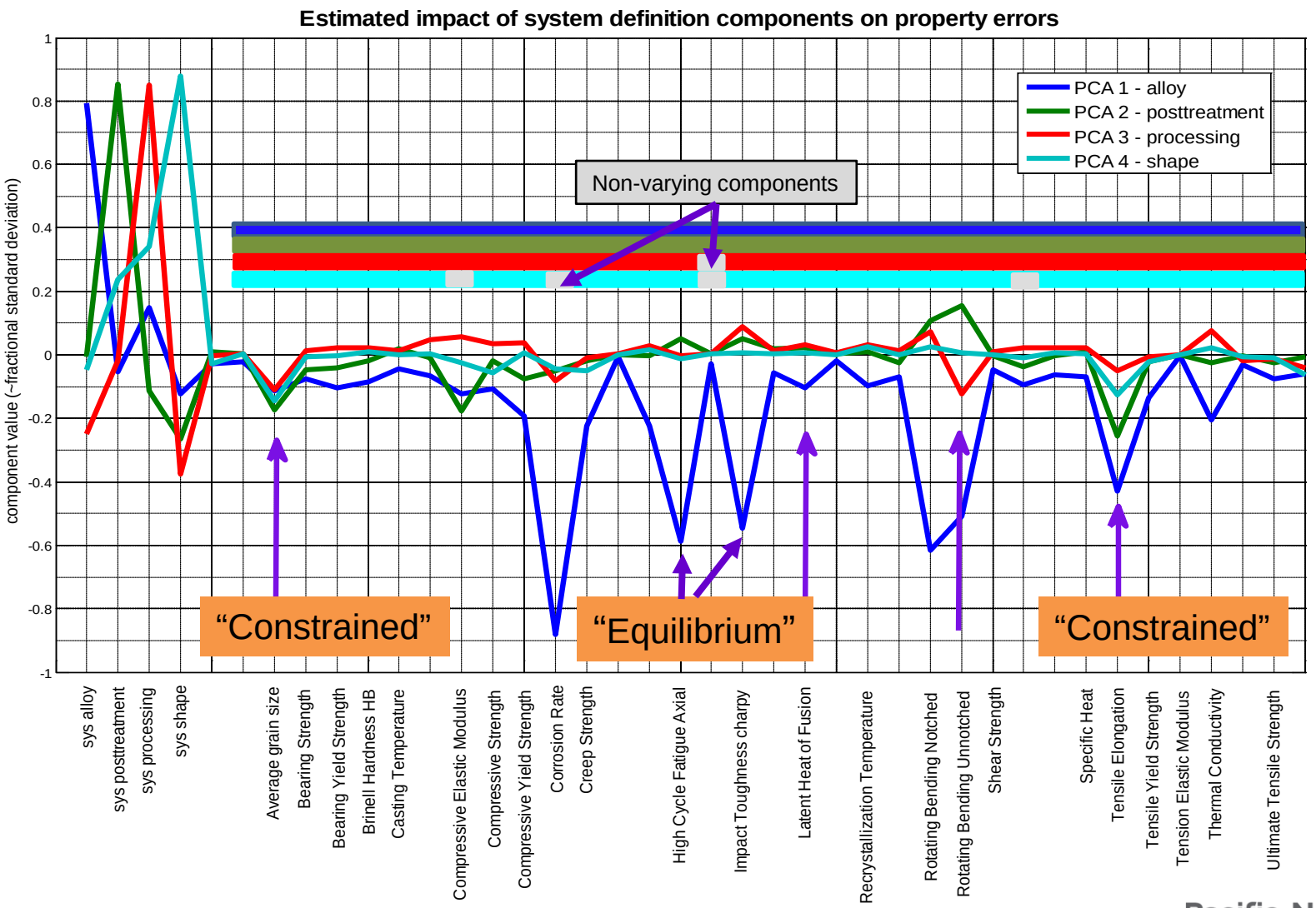
Normalize components to unit indicator variable height for comparison; flag non-varying components

↓



Pacific Northwest
NATIONAL LABORATORY

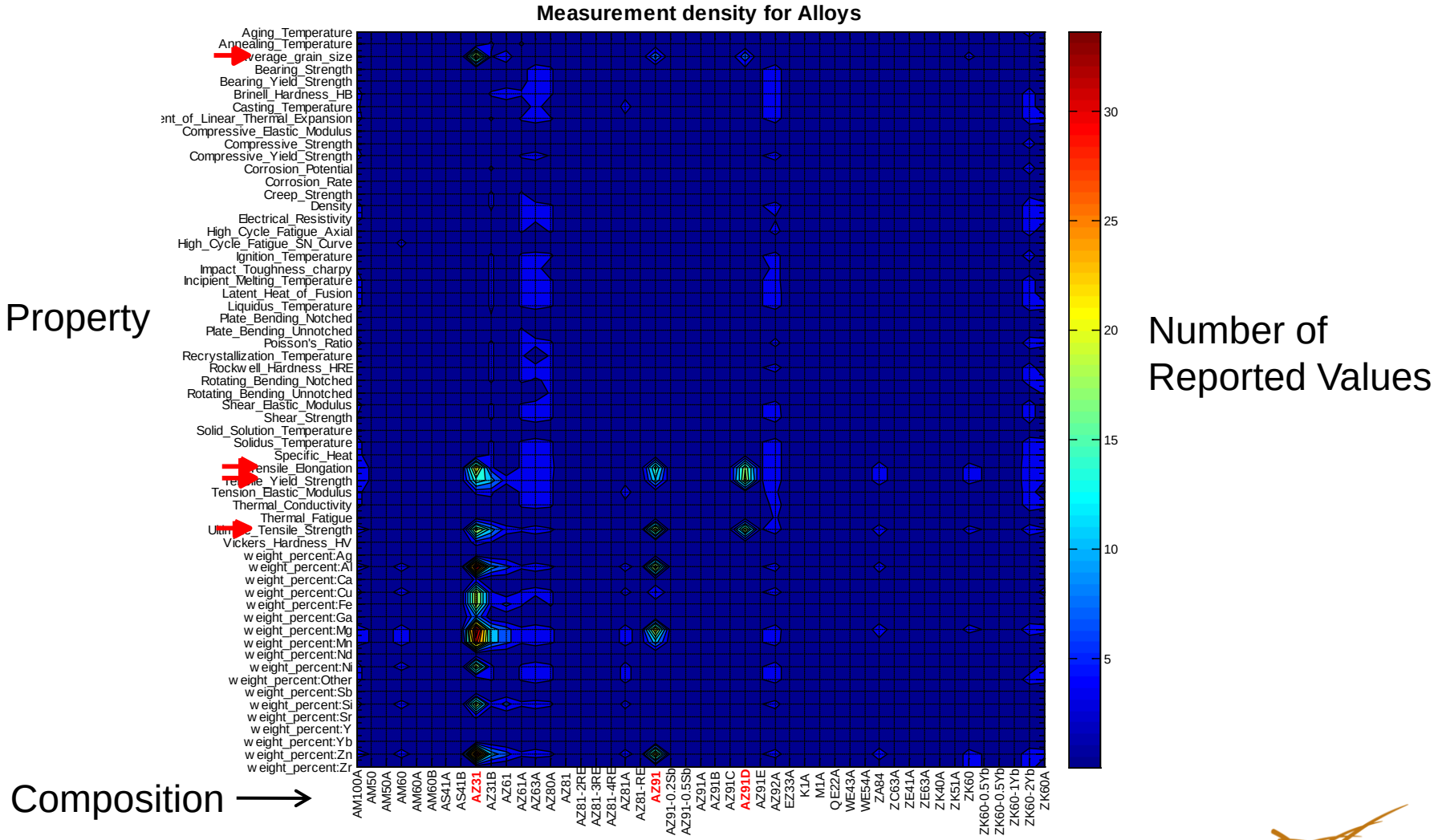
Tools to identify system definition components controlling property value deviations



- Two important signatures: "thermodynamic," "constrained"
- *Error for core mechanical properties depend on all components*

Pacific Northwest NATIONAL LABORATORY

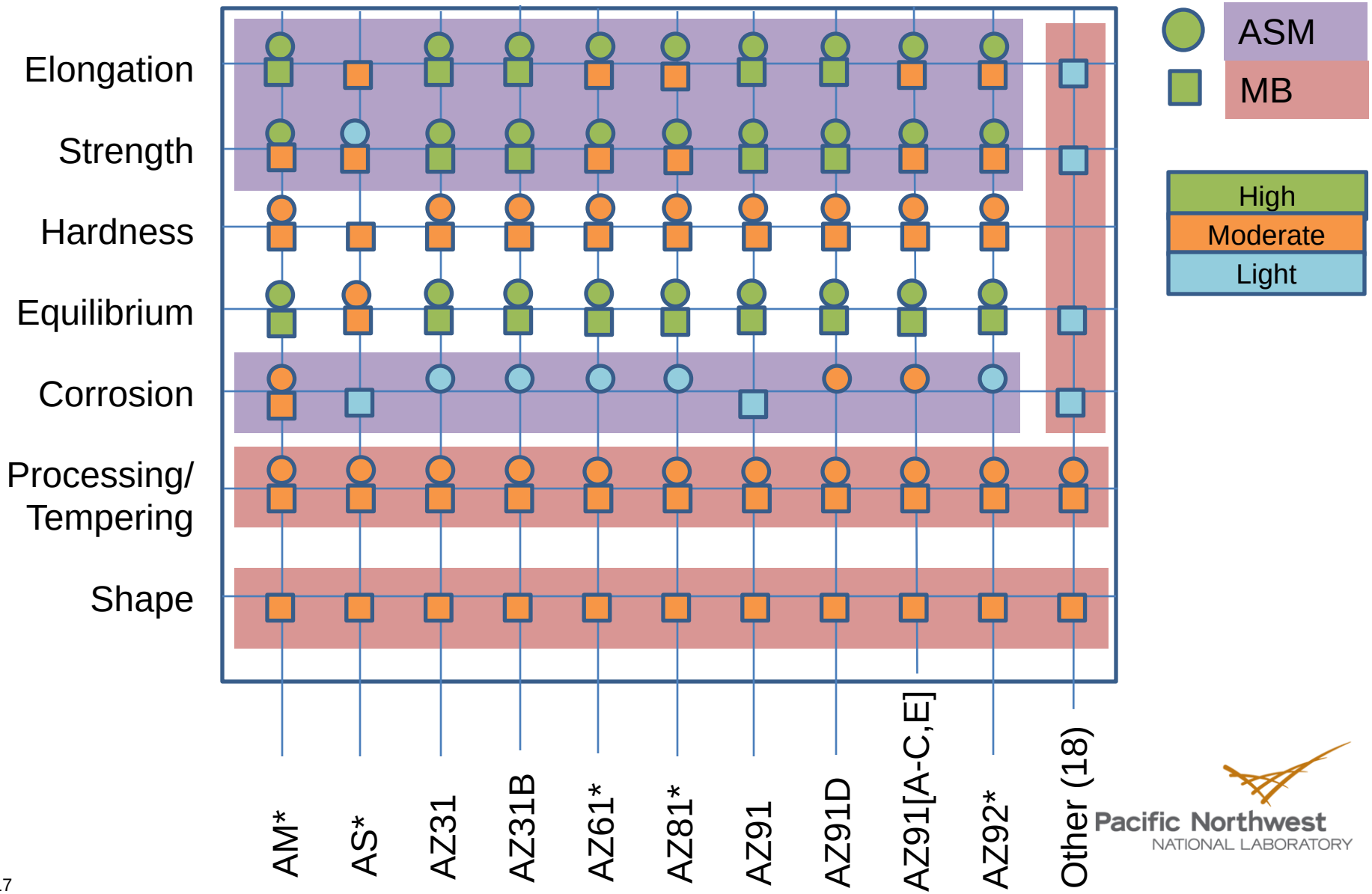
Basic data support: Tools for data density... what properties, alloys measured, what aren't.



- Data density map for lightweight alloy section of Metals Bank
- Illustrates dense and sparse data regions; information gaps
- Caveat: Data by itself does not necessarily construe knowledge

Pacific Northwest NATIONAL LABORATORY

Tools for data assessment: Data coverage summary (Metals Bank/ASM)



Summary of Technical Accomplishments

- Assessed data support in current data resources and current state of data capabilities (Milestone Feb 2010)
- Analysis repository schema and tested implementation. (Milestone Feb 2010)
  - Design helps resolve incomplete system or property definitions
  - Allows direct entry of "as is" data in its original form
  - Accommodates differing system and property definitions
- Translated knowledge retention schema into information transfer method for first principles workgroup (ThermoCalc-ESPEI-MSSU-First Principles)
- Repository assessment tools; assess support for information-driven validation, verification and design (Milestones Sep 2010).
  - Diagram database content; compare data population (single & combined)
  - Diagram of property value deviations as a function of system; identify system/property combinations with ambiguous property values
  - Map and diagram impact of system definition components against property deviations
  - Systematically look for associations between properties and processing variables

Pacific Northwest
NATIONAL LABORATORY

Collaborations

- Partners/Collaborators
  - Mississippi State University (materials repository)
  - Pennsylvania State University (ESPEI integration with repository)
  - ThermoCalc (ESPEI integration, Calphad model)
  - Materials Atlas (Iowa State)
  - CANMET (Kevin Boyle)
  - Sean Agnew (University of Virginia)
  - MetalsBank (materials repository)
  - Discussion in progress for collaborations with other repository developers (ASM)
- Technology Transfer
  - Techniques/design recommendations for software/repository integration (Pennsylvania State University collaboration)
  - Informatics tools / database design available to ICME community
  - Sharepoint site: https://spteams1.pnl.gov/sites/mat_informatics/default.aspx

Pacific Northwest
NATIONAL LABORATORY

Future Work

FY2011

- Informatics support for multiple data resources
  - Repository+Repository, Individual+Repository, Individual+Individual
  - Display ‘global’ information gaps – multi-resource
  - Physics-based data merging
  - ‘Multi-repository’ additions to system definitions
- Materials informatics Toolkit
  - Data completeness and support
  - Multiphysics deconvolution for knowledge/system definition augmentation
  - Modeling power visualization for workgroup validation and verification tasks

Follow-on Proposal

- Feature development to identify new or missing multiphysics
- Bridge new compositions to enable more robust design
- Granularity and multiscale hierarchy in data resources
- Robust strategies to identify/address global information gaps

Technical Back-Up Slides (limit of 5)

Technical Approach – Data Completeness

- Identify and fill gaps in ICME knowledge infrastructure required for Mg/Alloy design
  - Identify data ambiguities -- “same” measurement, different value (missing context in repository).
  - Identify data discrepancies between current repositories
  - Assure data are sufficiently complete to engage Phase II modeling – structure-processing/property relations for alloys
- Method: materials informatics techniques
  - Look for structure or processing data correlated with final properties
    - Examine impact of associating data across multiple length scales
    - Examine whether data forms must be updated to enable modeling
  - Anomaly detection
    - Property values that are not defined by specified structure-processing information
    - Ambiguities across repositories
- Collaborate with current repositories on inputs and outcomes

Pacific Northwest
NATIONAL LABORATORY

Technical Approach - Knowledge Augmentation

- What: after identifying data completeness problems, suggest knowledge augmentation in current alloy/processing databases
  - Make outcomes available for updates to repositories as needed
  - Informatics tools
- Why: Current data may be incomplete or ambiguous for modeling purposes
  - Assess and enable overall data support for information-driven design
  - Examine issues with data ambiguities
  - Knowledge assessments are measured against a problem (what is to be learned)
- How:
  - Look for structure or processing data correlated with final properties
    - Examine impact of associating data across multiple length scales
    - Examine whether data forms must be updated to enable modeling
  - Anomaly detection
    - Property values that are not defined by specified structure-processing information
    - Ambiguities across repositories

Pacific Northwest
NATIONAL LABORATORY

Example: Data completeness and knowledge augmentation in MSSU repository

- *Goal: Identify compositions and processing conditions for Mg alloys consistent with improved properties*
- *Data resource: MSSU/CAVS Materials Properties repository*

- Within the repository
  - Available: Approximately 30 Mg alloys
  - Available: Stress-strain curves for each alloy
  - *Not available*: Detailed composition or processing information
  - *Not available*: Yield strength numbers deduced from the stress-strain curve

- In advance of detailed analysis (knowledge augmentation)
  - Processing sequences should be added to the repository
  - (Bulk) composition should be added to the repository
  - A yield stress feature should be added to the stress-strain curves
  - Other features, based on experience and theory
  - *Try for a small set first, assessing potential association with yield stress*

More generally, systematically identify missing information, anomalies in characterization, and associations between structure and properties…

Pacific Northwest
NATIONAL LABORATORY

Example: Knowledge augmentation; where, how

- It is easy to intermingle properties and systems…

Table 1. An example of alloy property data [CRC Handbook 64th edition]

System	Name	Composition	Density (g/cm^3)	Tensile Strength Kg/m.m	Form
1	Alnico I	Fe 37-39; Al 12; Ni 20-22; Co 5	6.9	2.9	Cast, isotropic
2	Alnico II	Fe 52.5; Al 12; Ni 24-26; Cu 6; Co 12.5	7.1	2.1	Cast, isotropic
2	Alnico II	Fe 52.5; Al 12; Ni 24-26; Cu 6; Co 12.5	7.1	45.7	Sintered, isotropic
3	Alnico III	Fe 59-61; Al 12; Ni 24-26; Cu 3	6.9	8.5	Cast, isotropic
4	Alnico IV	Fe 55-56 ; Al 12; Ni 27-28; Co 5	7.0	6.3	Cast, isotropic
5	Hypothetical Calculation	Fe 40, Ni 24, Co 5	9.0	60	Repeated fcc unit cell

- Inconsistent or ambiguous properties (name, form)
  - Difficult to compare across systems
- Merged systems (e.g. Alnico II)
- Merged properties (e.g. Form)
- *Augmentation in this case: can rectify ambiguities*

Pacific Northwest
NATIONAL LABORATORY

More Detailed Summary -Toolkit Accomplishments

- **Repository assessment tools; assess support for information-driven validation, verification and design**
  - **Diagram of database content**
  - **Ability to survey data population and compare repository content**
  - **Locate information gaps; identify/prioritize maximum leverage points**
  - **Identify data conditions conducive to robust validation of model predictions/experimental measurements**
  - **Diagram of property value deviations as a function of system**
  - **Identify repository system/property combinations with ambiguous property values (anamolies/system and measurement uncertainties)**
  - **Indicate which systems/property combinations are most susceptible to large property variations**
  - **Map and diagram impact of system definition components against property deviations**
  - **Systematically look for associations between properties and processing variables**

Pacific Northwest
NATIONAL LABORATORY