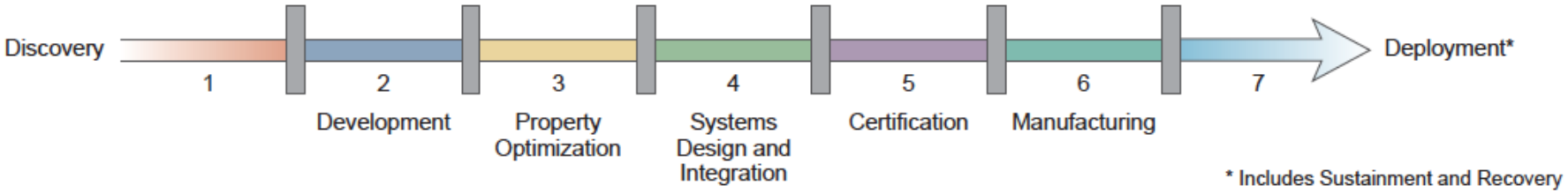




# The Role of HPC on Materials Genomics

# Materials Genome Initiative

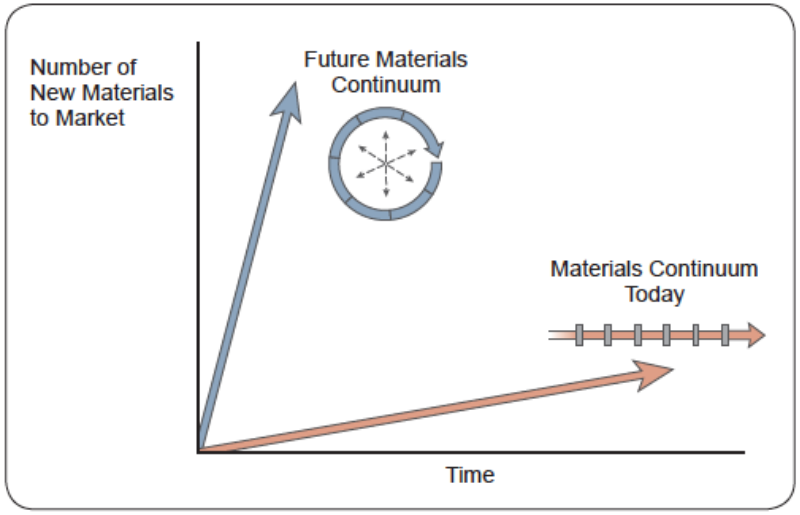
# Materials Genome Initiative

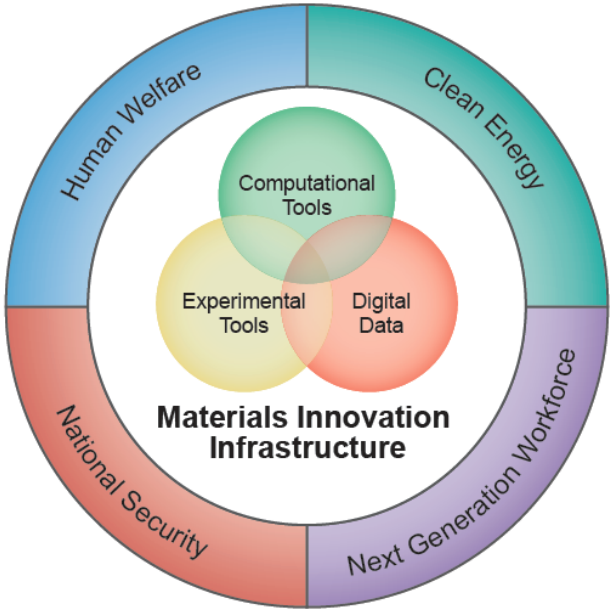



Typical Time Frame ~ 15- 20 years!!!

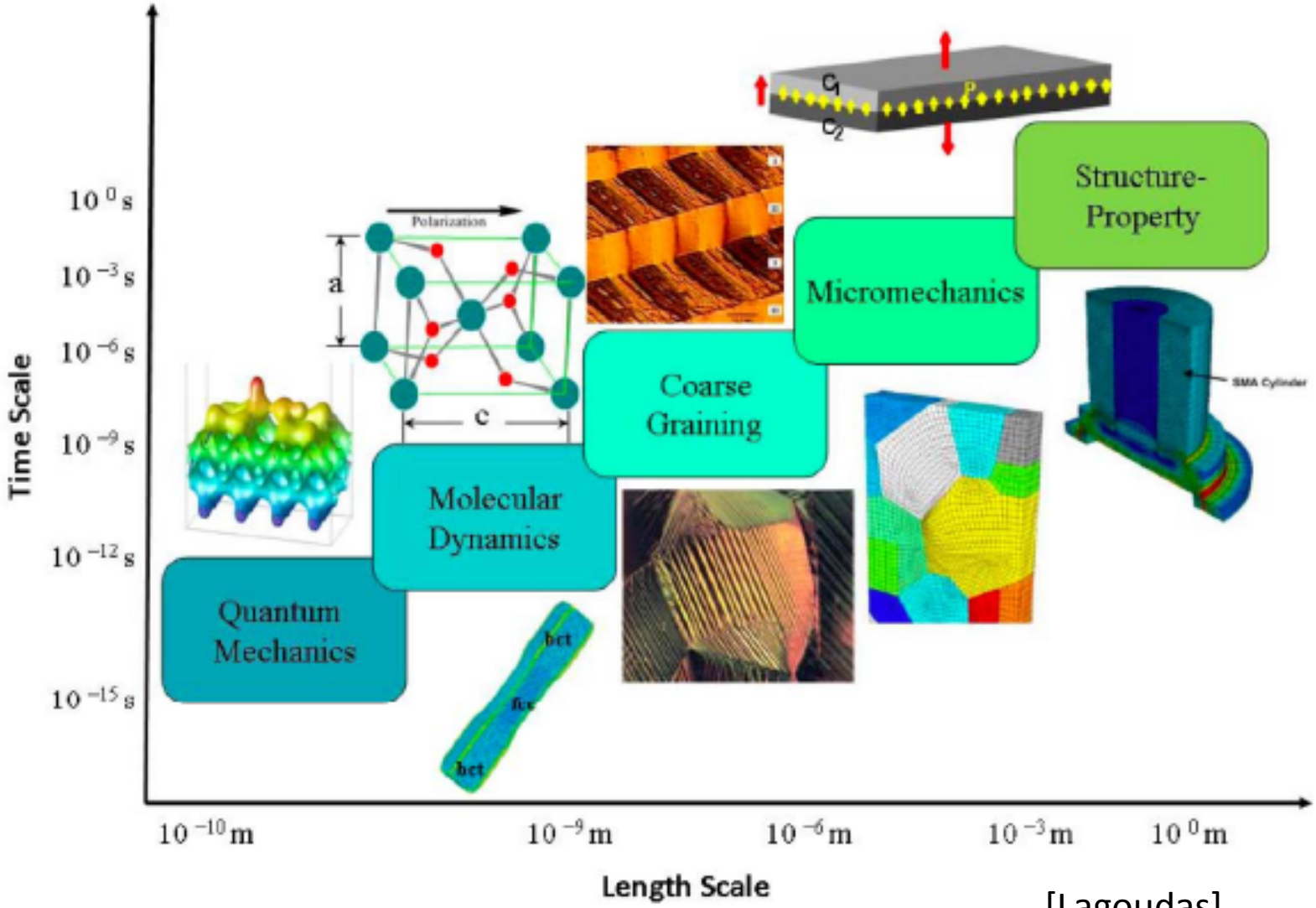
**Current activities:** The Materials Genome Initiative:

Goal: Reduce Cost and Time by Half



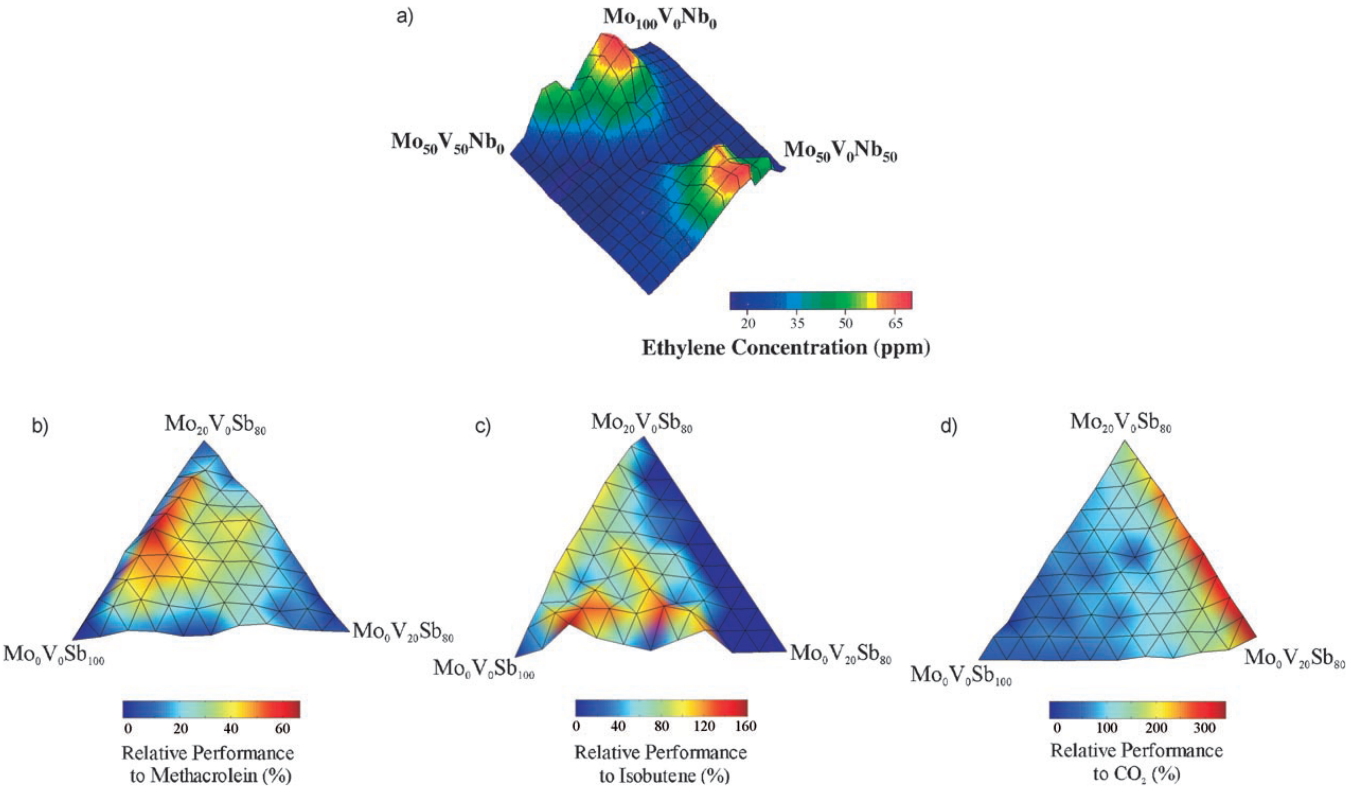



# MGI is (not only about) Multi-Scale Modeling



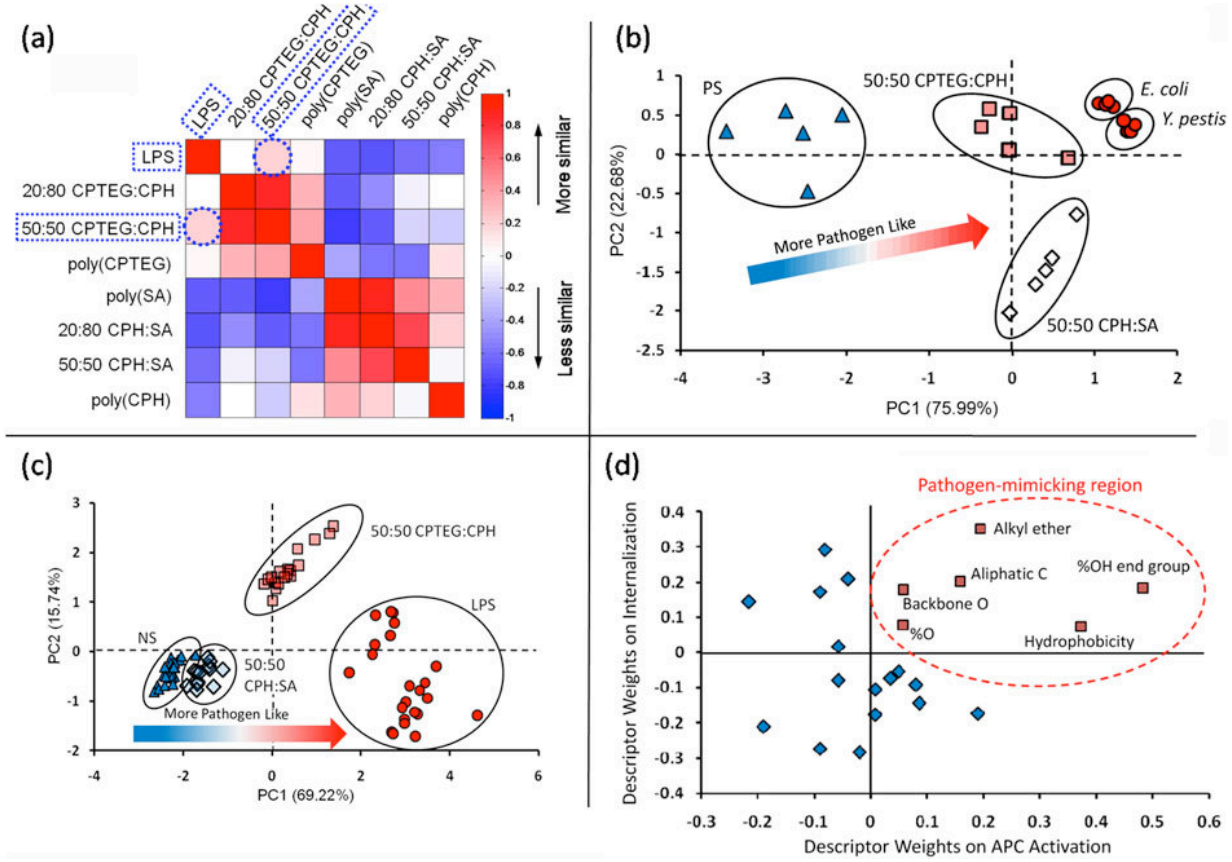

[Lagoudas]

# MGI is (not only about) High Throughput Synthesis/ Characterization



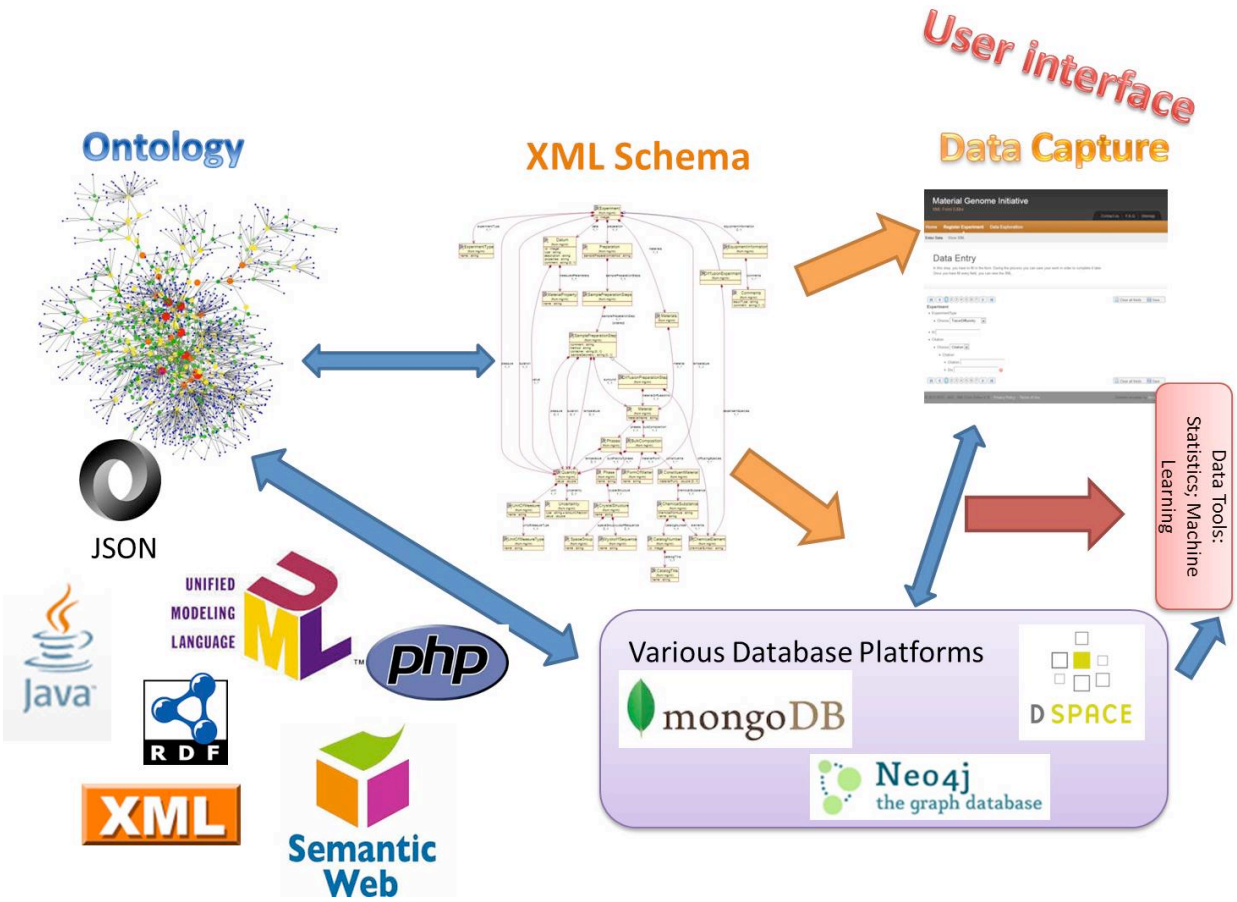

[Maier 2012]

# MGI is (not only about) Materials Informatics



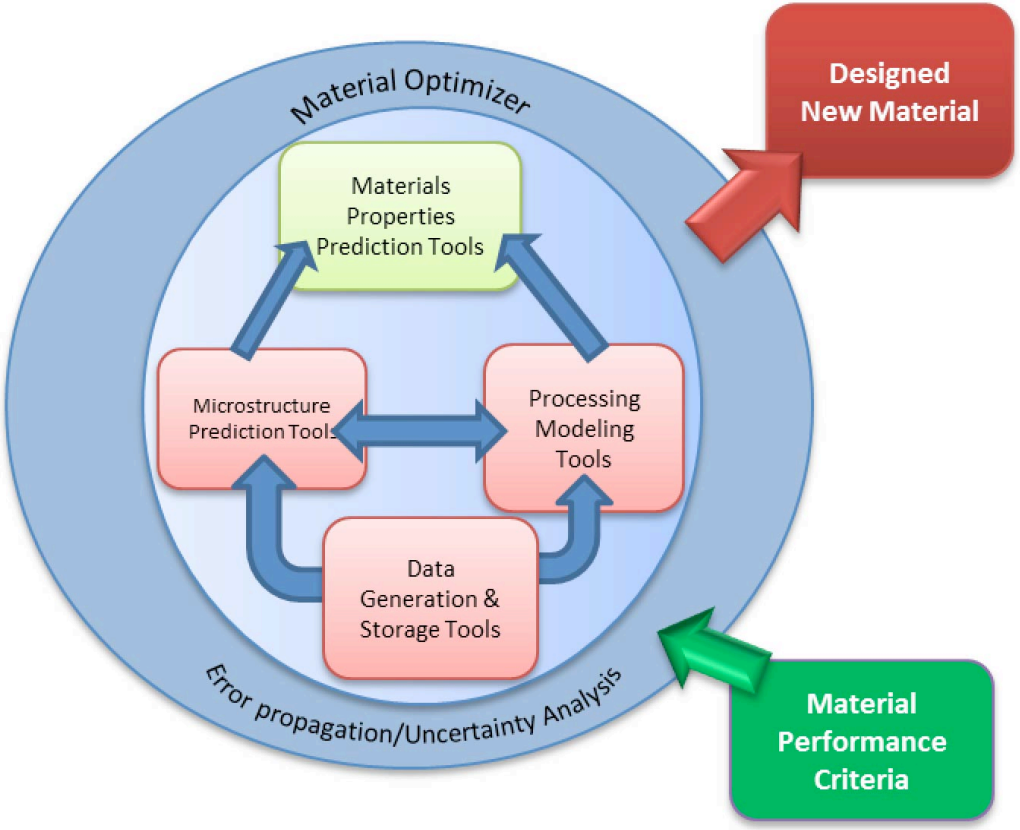

[Ulery et a, Nature 2011]

# MGI is (not only about) Information Infrastructure



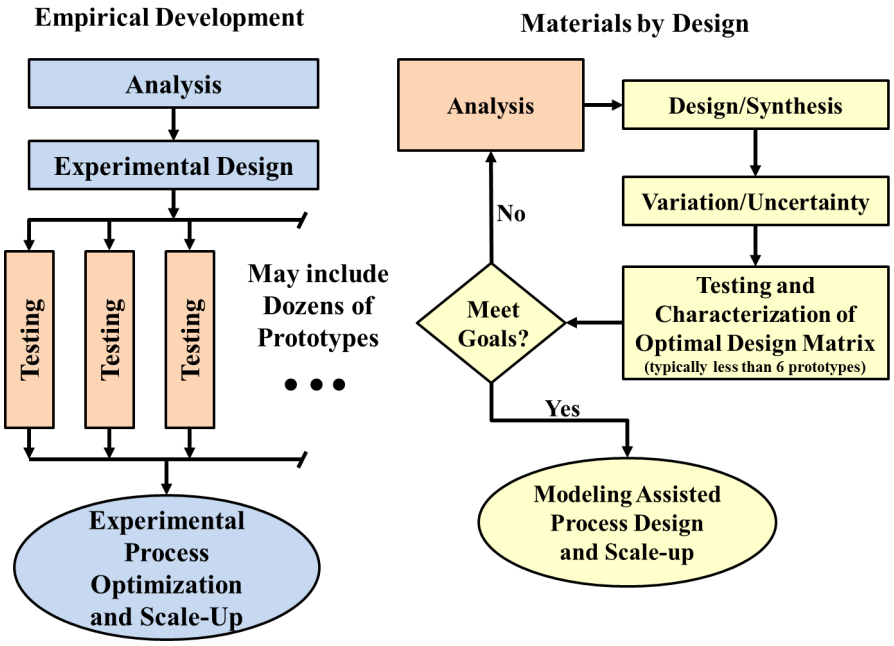

[NIST, Materials Genomics Group]

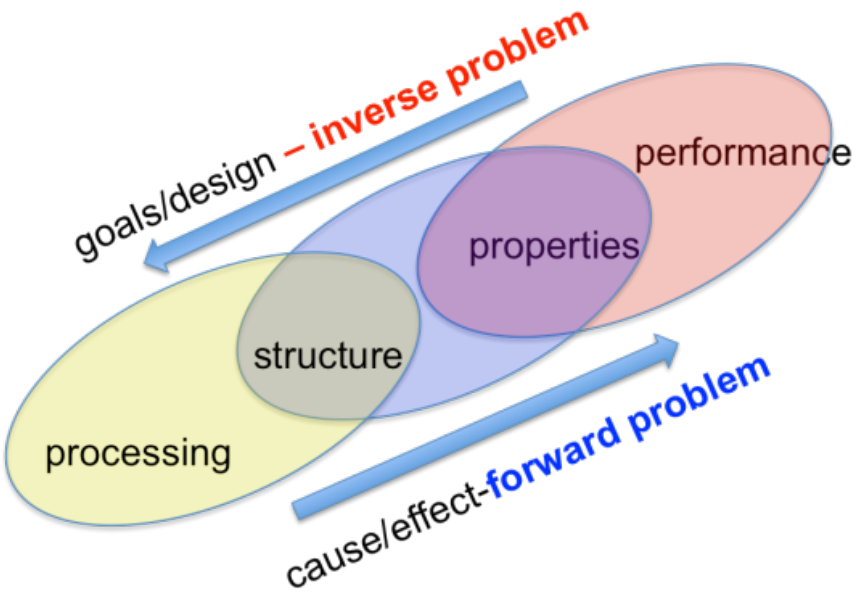
# MGI is about integration of all the above (but this is not all)




[NIST, Materials Genomics Group]

# MGI is Ultimately About Design



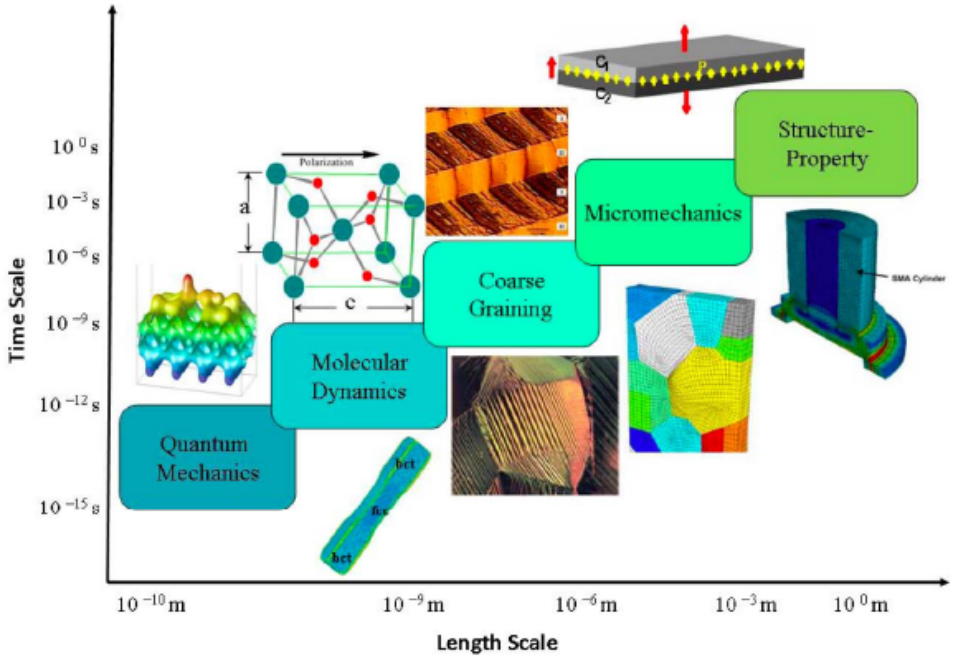

[Olson 1997,2012]




# MGI AND HARDWARE/SOFTWARE/ INFORMATICS INFRASTRUCTURE

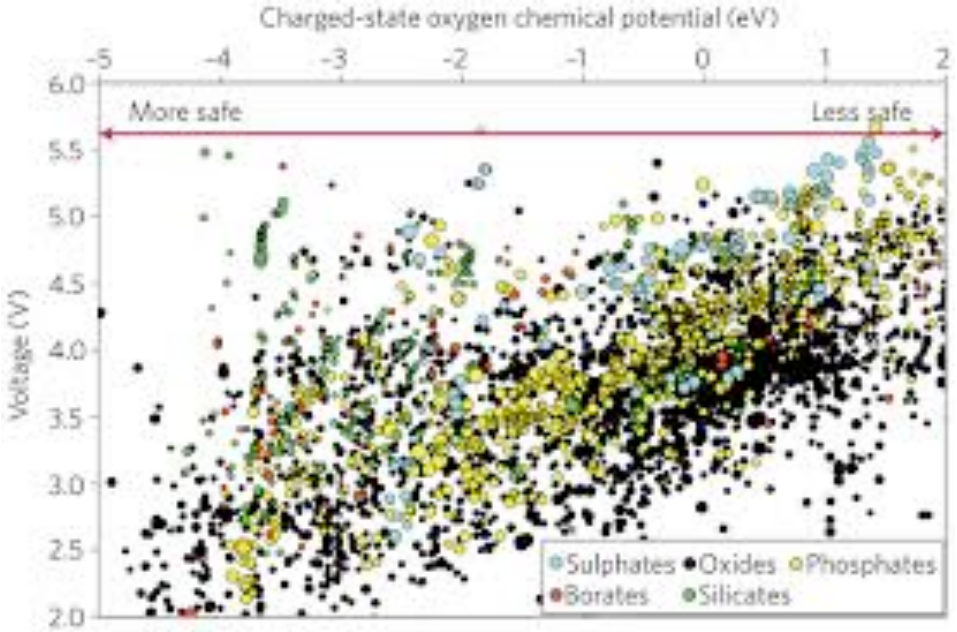
# Large Materials Simulation Problems

- Large problem sizes
  - # atoms
  - # DOFs
  - # grid points
- Multi-scale model integration
  - DFT+MD+FEA



# High Throughput Materials Computations

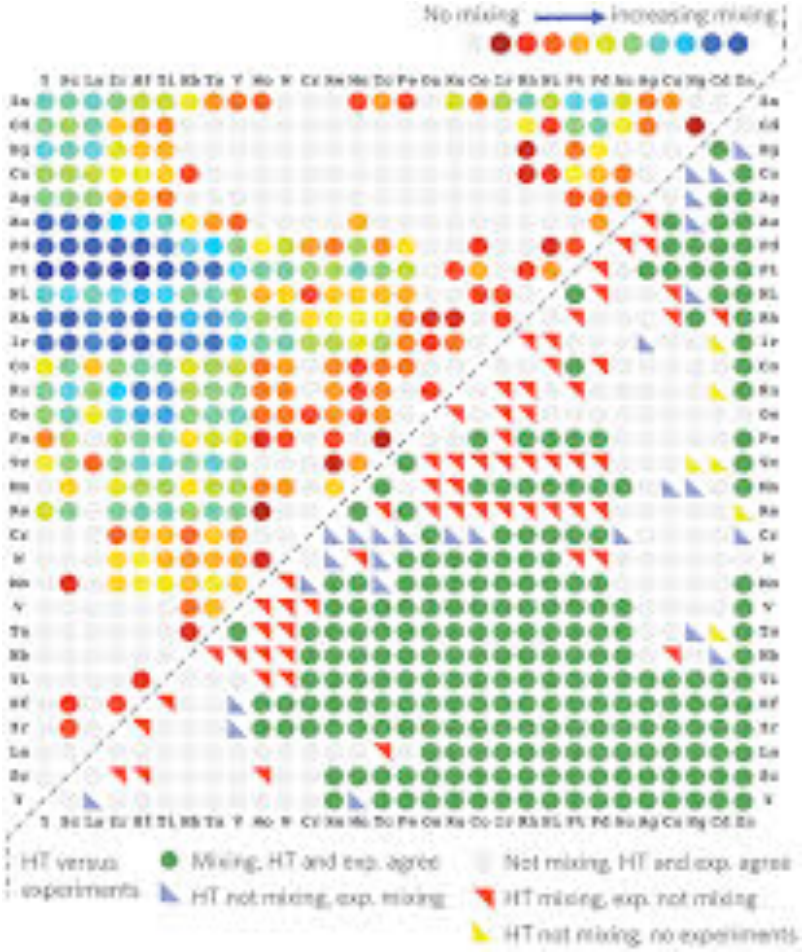
- Many (relatively small) problems
- Massively parallel computing tasks (e.g. high-throughput ab initio)
- Embarrassingly parallel simulations (e.g. Monte Carlo)




[Ceder, 2013]

# Materials Informatics

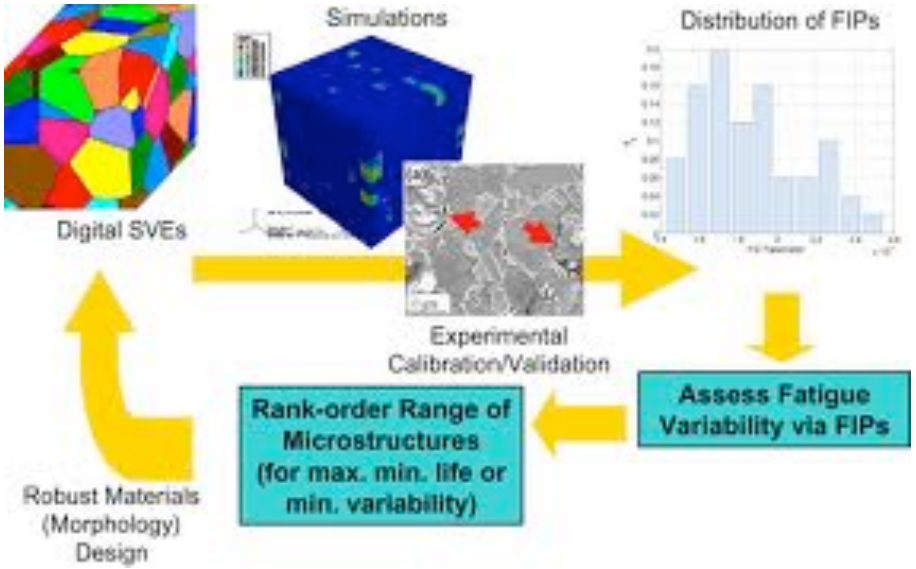
- Identify correlations between materials descriptors and performance indicators
- Correlations between multi-dimensional data points
- Use sophisticated informatics approaches (i.e. classification/ regression)




[Curtarolo, 2012]

# Simulation-driven Materials Optimization

- Requires combining:
  - Large simulations (for detailed understanding)
  - High-throughput calculations (of simplified models)
  - Materials informatics
  - Advanced Optimization Schemes



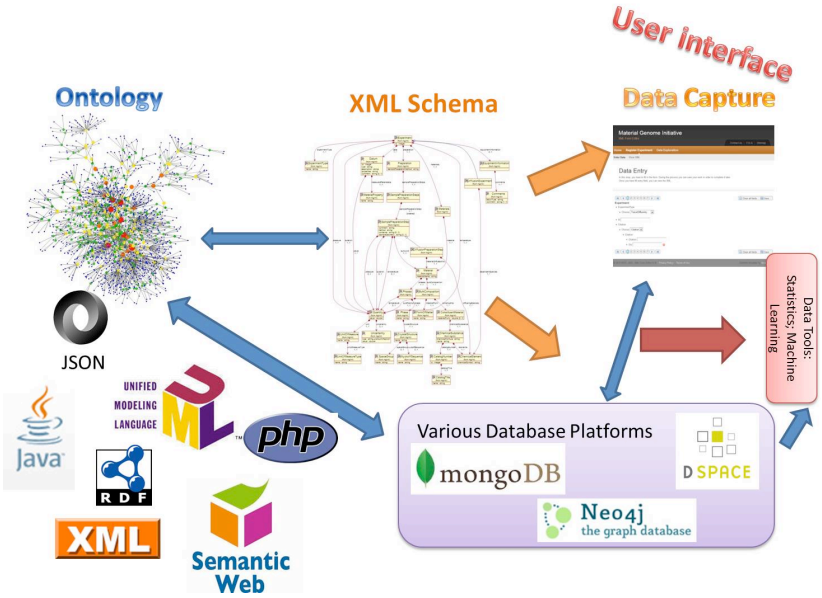

[Kalidindi and McDowell]

# Materials Data Mining

- How do we exploit previous work?
  - How do we extract information from highly non-structured sources (i.e. text + plots + micrographs)
  - How do we encode this data into useful information?

# Information Infrastructure

- Achilles Heel of MGI:
  - No one (except for very few groups) is thinking about how to encapsulate information
  - Without II, MGI cannot be realized
  - II is about enabling information exploitation




[NIST, Materials Genomics Group]

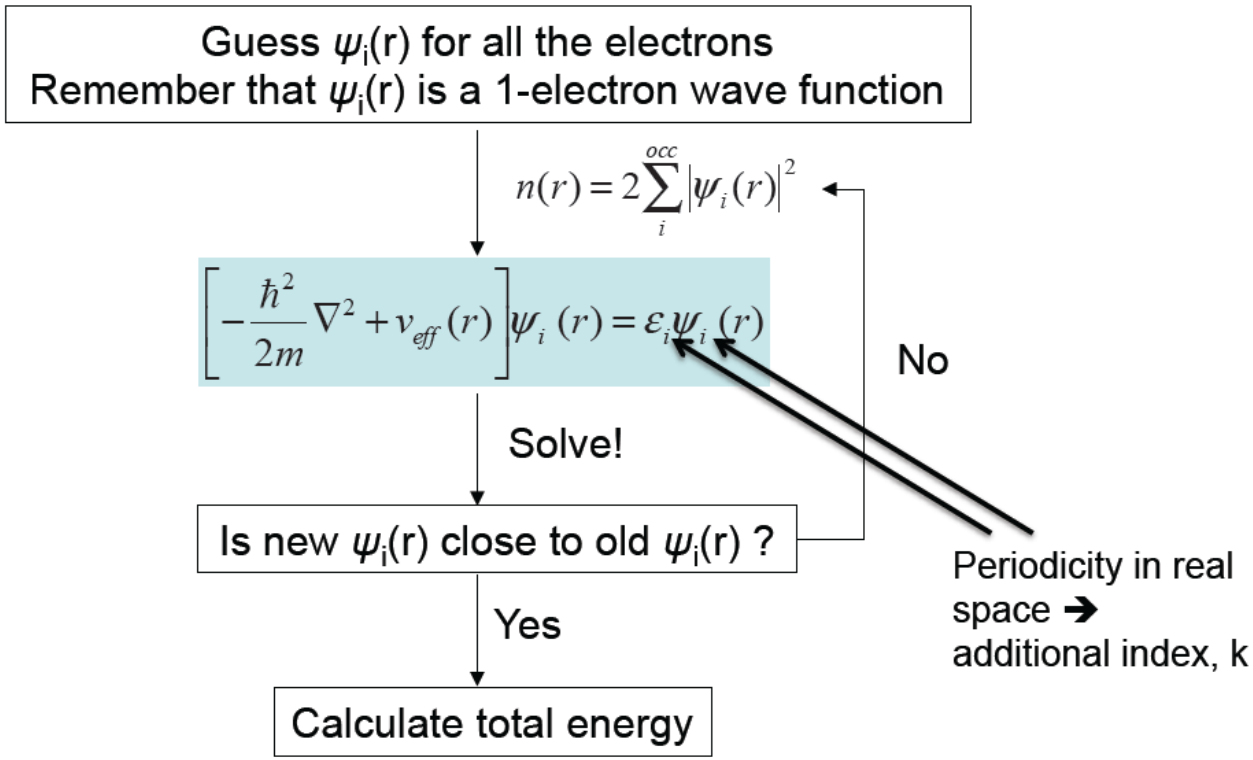
# Opportunities

- We have the expertise
  - Methods: advanced synthesis, characterization, computational materials science, informatics, design, etc
  - Materials: multi-functional materials, advanced structural materials, etc
- We have the resources
  - New SC resources at TAMU provide us with unique competitive advantages
  - Only a few other groups would have similar access to resources
  - We should aspire to try and do the (so far) impossible/ impractical



# High-throughput Ab Initio Materials Data Mining

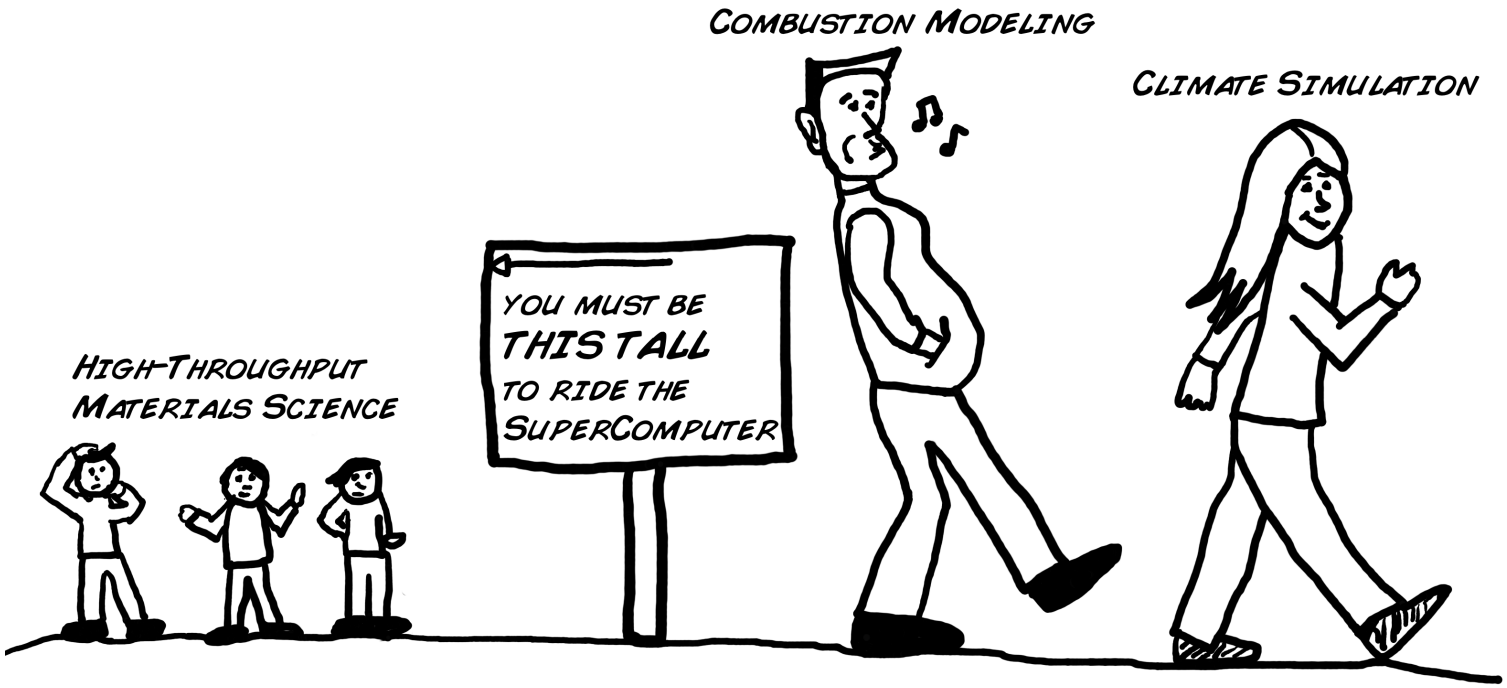
# DFT in a nutshell




$$E_{elec} = \sum_{i=1}^{N}\varepsilon_i - \frac{1}{2}e^2\iint\frac{n(r)n(r')}{|r-r'|}d^3r\,d^3r' + \int n(r)\{\varepsilon_{xc}[n(r)] - v_{xc}[n(r)]\}d^3r$$

$$E = E_{elec} + nuclear - repulsion - energy$$

# How does HT Ab Initio Stack with other HPC Applications




Note: this only applies to 'trivial' DFT calculations

Cost increases as $N^{4-5}$ where N = size of system

In DFT, we reach the limit of what is possible to do quite fast

[Jain 2014]

# Challenges for HT Ab Initio

- Failures are common:
  - Problematic structures
  - Wrong set of parameters
  - Poor convergence
  - Issues on hardware side (connections to SC resources, hardware malfunction, etc)
- Workflows are quite dynamic
  - If a structure is a metal, you do not want to calculate band gap, for example
- Automation is critical

# TAMMAL

- **T**exas
- **A**&M
- **M**aterials
- **M**odeling
- **A**utomation
- **L**ibrary

# TAMMAL

- Object Oriented:
  - Job Object
  - Calculation Sequence Object
  - Workflow Object
  - Post-processing Objects
- Highly customizable
  - Users can create their own workflows
- Minimal User Interaction
  - Type one command, forget about it

# METADATA

- For Data Mining to work, we need data about data (meta-data)
- TAMMAL stores all information necessary to duplicate calculation (code, coder version, input parameters, etc) as well as calculation results in JSON files
  - JSON format: Human readable, dictionary-based serialization of data

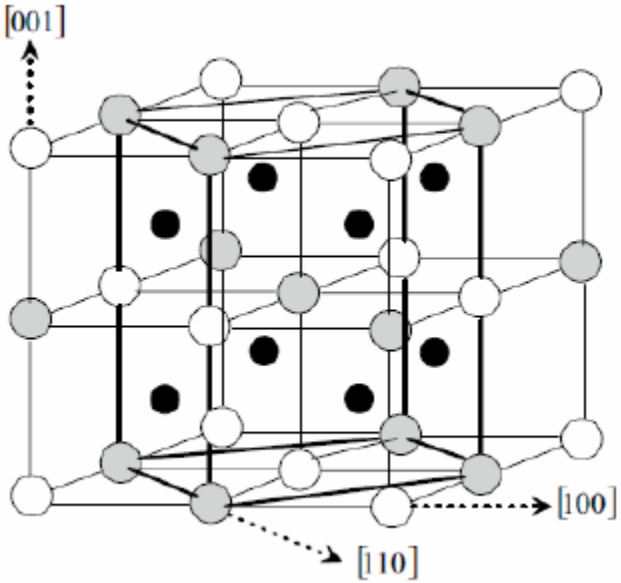
# DATABASE

- All JSON files created by TAMMAL will be stored in MONGO DB
- MONGO DB:
  - Non-SQL Database System
  - Highly Efficient (map/reduce)
  - Highly Scalable
- Using Python implementation (pymongo)

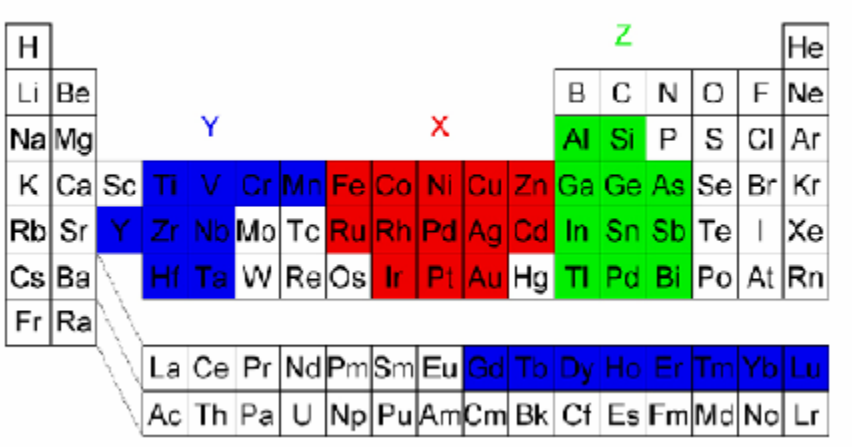
# DATA Mining

- Using Python-based Scikit to mine data
- Scikit is a series of libraries for machine learning:
  - Classification
    - Trees, Neural Networks, etc
  - Regression
    - GA-based
    - Multi-variable
  - Dimensional Reduction
    - Principal Component Analysis

# TAMMAL Example: Heusler Alloys



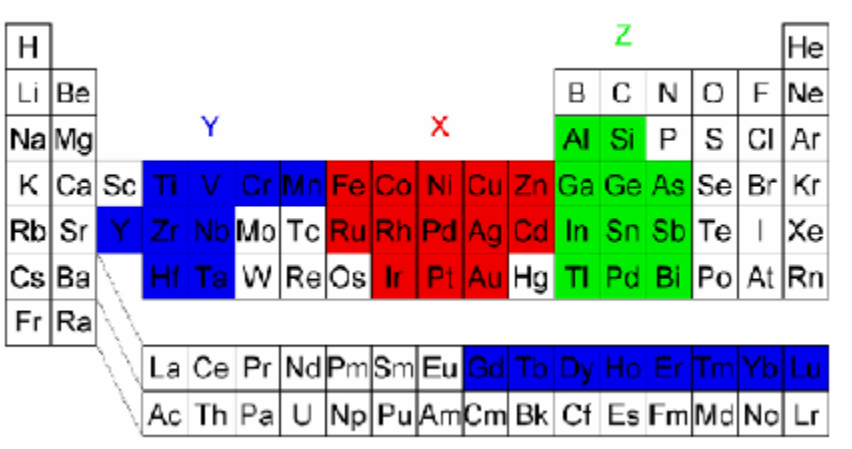

Schematic of the $L2_1$ Conventional Cell in
Heusler Alloys [Williams, 2009]




Typical compositions of $X_2YZ$ Heusler Alloys [Williams, 2009]

# TAMMAL Example: Heusler Alloys

- A week ago:
  - We started calculating ground state of all (simple) Heusler alloys in periodic table
  - 1300 compounds
  - Completed in 5 days
- Next week:
  - Start calculation of other properties
  - Do what other groups have been unable to do (due to lack of computational resources)

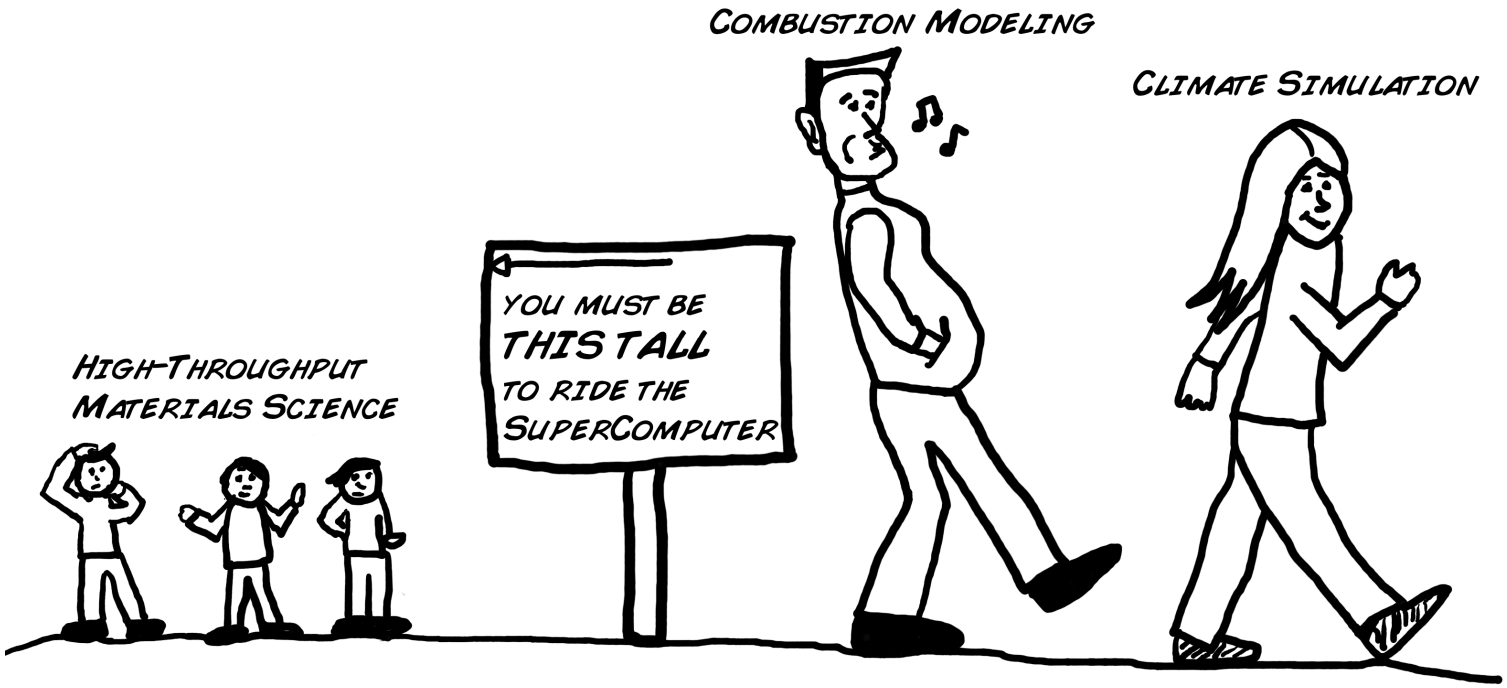



Typical compositions of $X_2YZ$ Heusler Alloys [Williams, 2009]

# TAMMAL in TAMU SC

- We can start calculating non-ground state properties
  - No other group is doing this at this moment
- We can expand TAMMAL to automate other Materials-based simulation tasks

# High throughput is a valid use of SC Resources!




[Jain 2014]