MATERIALS INFORMATICS: An Introduction

Krishna Rajan
Director:
NSF International Materials Institute :
Combinatorial Sciences and Materials Informatics Collaboratory
(CoSMIC-IMI)
Iowa State University

OVERVIEW

1. What is "materials informatics" ?

2. Why do we need informatics for materials science and engineering?

3. What experimental and computational resources and tools are needed to enable materials informatics?
    - Data generation / combinatorial experiments / high throughput experimentation / reference libraries and databases
    - Data warehousing
    - Dimensionality reduction
    - Clustering analysis
    - Predictive modeling techniques
    - Visualization techniques
    - Cyber infrastructure

DATA DRIVEN MATERIALS SCIENCE

Data + Correlations + Theory = Knowledge Discovery

- Combinatorial experimentation
- Digital libraries & data bases

+

- **Data mining**
- **Dimensionality reduction**

+

- Atomistic based calculations
- Continuum based theories

=

- Materials discovery
- Structure-property-processing relationships
- Hidden data trends

Information is multivariate, diverse , can be very large and access / expertise is globally distributed

WHY MATERIALS INFORMATICS?

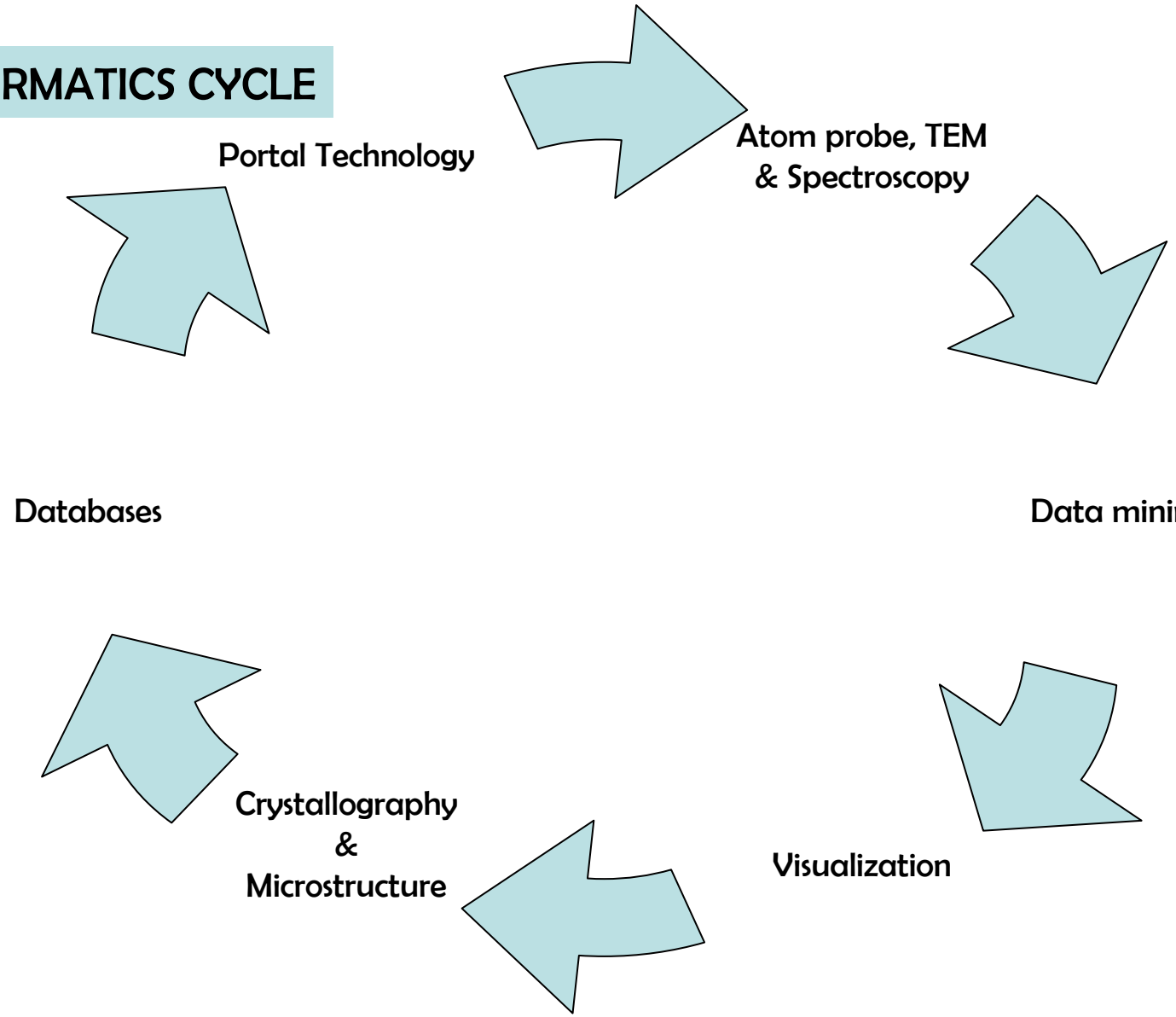
Potential of informatics:

- Management of informational complexity
- Accelerated discovery
- Identifying new pathways
- Building new learning communities through cyber-infrastructure

Realizing the potential:

- Data mining and statistical learning
- Cyber infrastructure
- Research platforms
- Impact on education – new paradigm for materials education

THE INFORMATICS CYCLE



SOURCES OF DATA : diversity of databases

Reference libraries

- crystallographic
- thermodynamic
- properties

handbooks

Literature data

- dispersed
- books/ reviews

Experiments

- systematic data collection –slow
- combinatorial experiments- high throughput
- in-situ / dynamic experiments- time series data

REFERENCE LIBRARIES

- Crystallographic –
  - hierarchical database- group theory driven

- Thermodynamic –
  - primary database- ie. Heat capacities – thermochemical data
  - derivative database- free energy data…computational phase diagrams

- Property databases …
  - meta database…building on primary and derived data but organized phenomenologically…eg. strength ..UTS / % RA / .2% off set yield ….foundations for “handbooks”

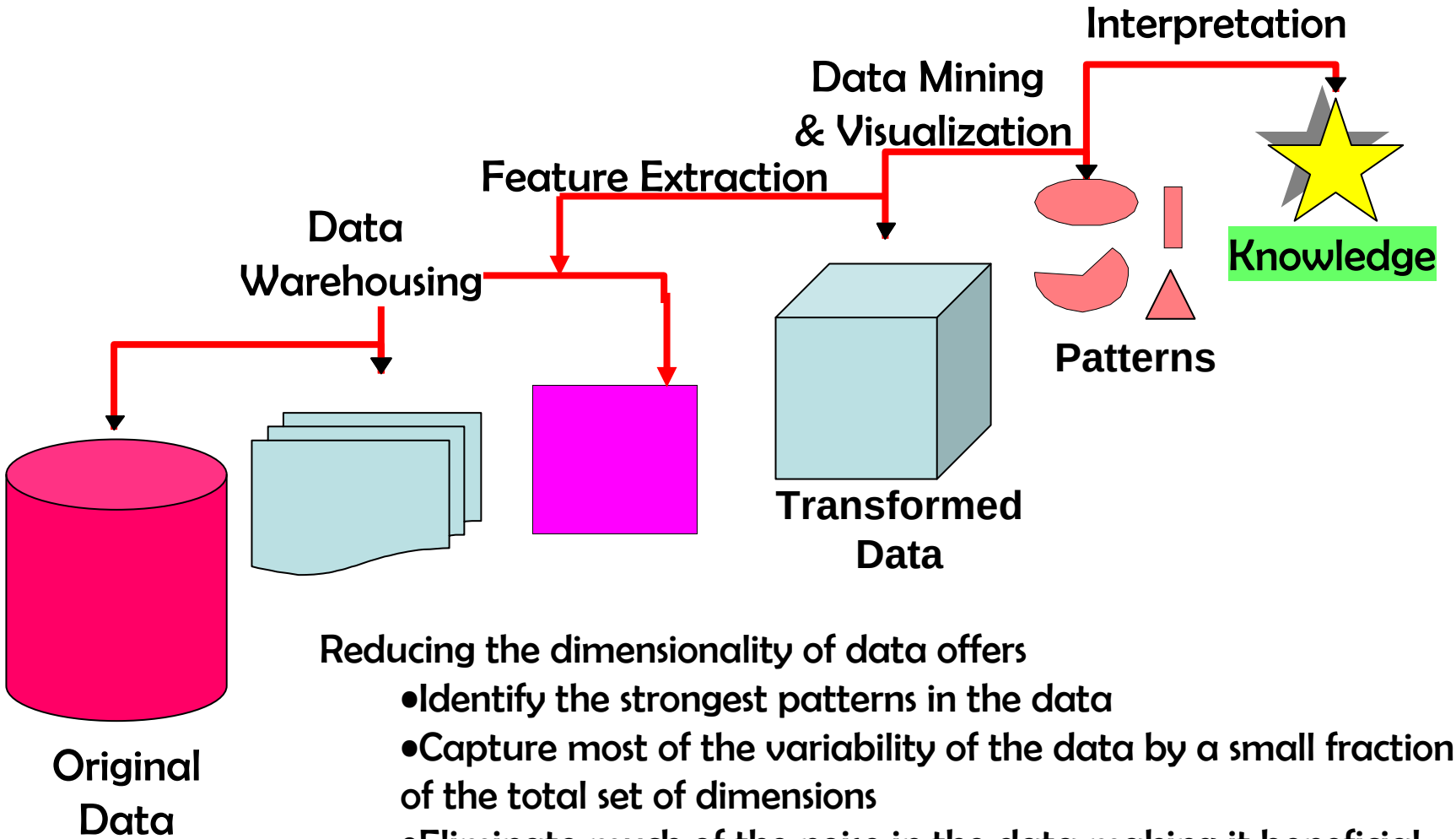
DATA STORAGE ⟷ **Database administration & management**
- thermodynamic, crystallographic & property data bases
- combinatorial experimental data

DATA CURATION ⟷ **Oracle, Unix, Supercomputing, SQL**
- taxonomy and ontology of materials science data
- data sharing , networking / cyber infrastructure

DATA REPRESENTATION ⟷ **JAVA, HTML, Python**
- object oriented programming language
- visualization of high dimensional data

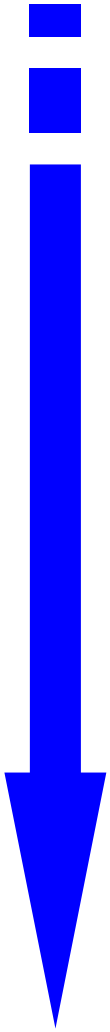
KNOWLEDGE DISCOVERY ⟷ **Data mining algorithms**
- Clustering analysis
- Quantitative Structure-Activity Relationships (QSAR) for materials design

DATA MINING and KNOWLEDGE DISCOVERY



Reducing the dimensionality of data offers

- Identify the strongest patterns in the data
- Capture most of the variability of the data by a small fraction of the total set of dimensions
- Eliminate much of the noise in the data making it beneficial for both data mining and other data analysis algorithms



- Data generation
  - Size and diversity
  - Combinatorial experiments
- Data storage and organization
  - Large data sets and computer memory
- Data query
  - Linking computer language to scientific theory and paradigms
- Data transfer and sharing
  - Cyber infrastructure
- Seeking correlations among diverse data sets
  - Curse of dimensionality
- Mining the data
  - Developing classification and predictions- QSAR
- Interpretation
  - Linking theory to data mining /
- Defining information space
  - Defining criticality and nature of descriptors

INFORMATICS STRATEGY: QSAR...following the biologists

$$\text{Functionality} = \mathcal{F}\,(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 \ldots\ldots)$$

Issues:

- how many variables?
- which variables are important?
- classify behavior among variables
- making quantitative predictions ...relate functionality to variables ...
  - traditionally we describe them by empirical equations:
  - Quantitative Structure Activity Relationships (QSARs) are derived from data mining techniques not assuming a priori which physics is the most important

Need to build database with these variables

COMPUTATIONAL ISSUES

Establish multivariate database:

Seek DIVERSITY in datasets

Data can come across length and time scales

Focus on properties of signal / macroscopic behavior rather than noise/ error. Assume complexity !!!

- Utilize data dimensionality reduction techniques
- Analyze variation and correlation in data
- Establish correlations across diverse data sets (ie. length & time scales
- Identify outliers: explore cause
- Develop predictive models
  - Target requirements of missing data
  - Quantitatively assess data diversity

Model relationships in data to seek heuristic relationships:

Advanced statistical learning tools can deal with:

- skewed data
- missing data
- differentiate between local and global minima
- ultra large scale datasets
- variable uncertainty

- Singular value decomposition
- Cluster analysis
- Partial least squares
- Support vector machines
- Association mining
- Fuzzy clustering

WHY COUPLE COMPUTATIONAL MATERIALS SCIENCE & INFORMATICS ?

- Accelerated insertion of materials into engineering systems
- Rapid multiscale design and optimization of materials properties
- Establishment of new structure –property correlations among large, heterogeneous and distributed data sets
- Discovery of new chemistries and compounds
- Formulation and / or refinement of new theories for materials behavior
- Rapid identification of critical data and theoretical needs for future problems