Database Infrastructure to Support Knowledge Management in Physicochemical Data

- Application in NIST/TRC SOURCE Data System

Qian Dong, Xinjian Yan, Robert D. Chirico, Randolph C. Wilhoit, Michael Frenkel

Thermodynamics Research Center (TRC)
National Institute of Standards and Technology (NIST)
Boulder, CO, U.S.A.

Evolution of Database Technology

Data file	**Hierarchical and Network Models**	**1960s**
Database	**Relational Data Model DBMS implementation**	**1970s**
R_Database **O_Database** **R_O_Database**	**Ubiquitous SQL RDBMS Advanced data models**	**1980s**
Warehouse **Web Database**	**Data Mining and Knowledge Management**	**1990s**

Hot Topic – Data Mining (DM), Knowledge Discovery in Databases (KDD) and Knowledge Management (KM)

- As of 2002 fall, a quick Google search gives 700,000 web pages with the exact phrase match of DM or KDD and 900,000 pages of KM
- Similar search through IEEE Explore and Web of Science discloses thousands of scientific papers published in research and application areas.
- Application Areas:

banking and credit	bioinformatics
customer relationship	Internet advertising
healthy care	e-commerce
Insurance	manufacturing
Marketing and retails	communications….

What is KDD, DM and KM?

Knowledge Discovery in databases (KDD) –

A process of non-trivial extraction of implicit, previously unknown and potentially useful information from large collections of data

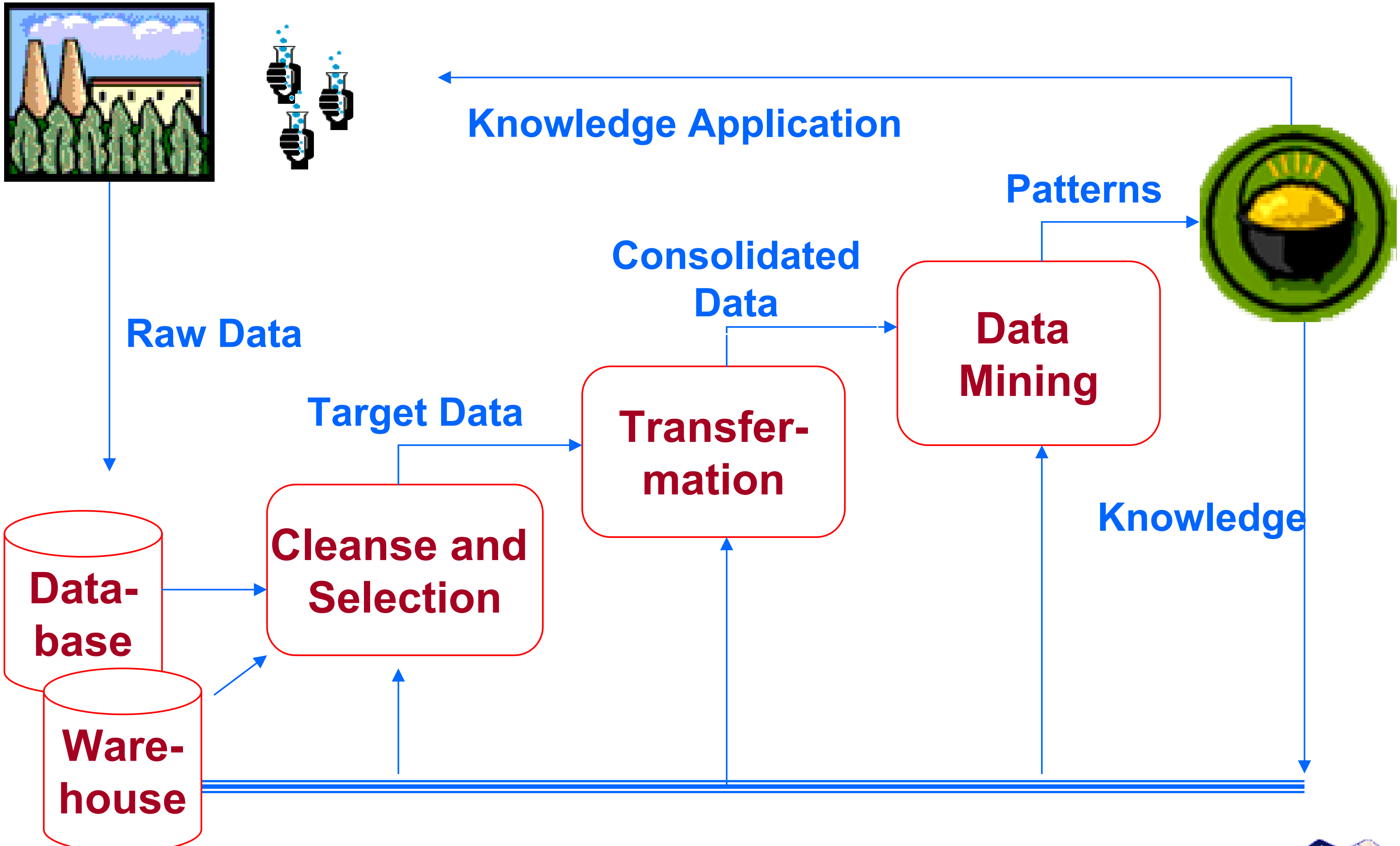
Data Mining (DM) -

A step in the knowledge discovery process – application of specific algorithms for extracting patters (models) from a large set of data.

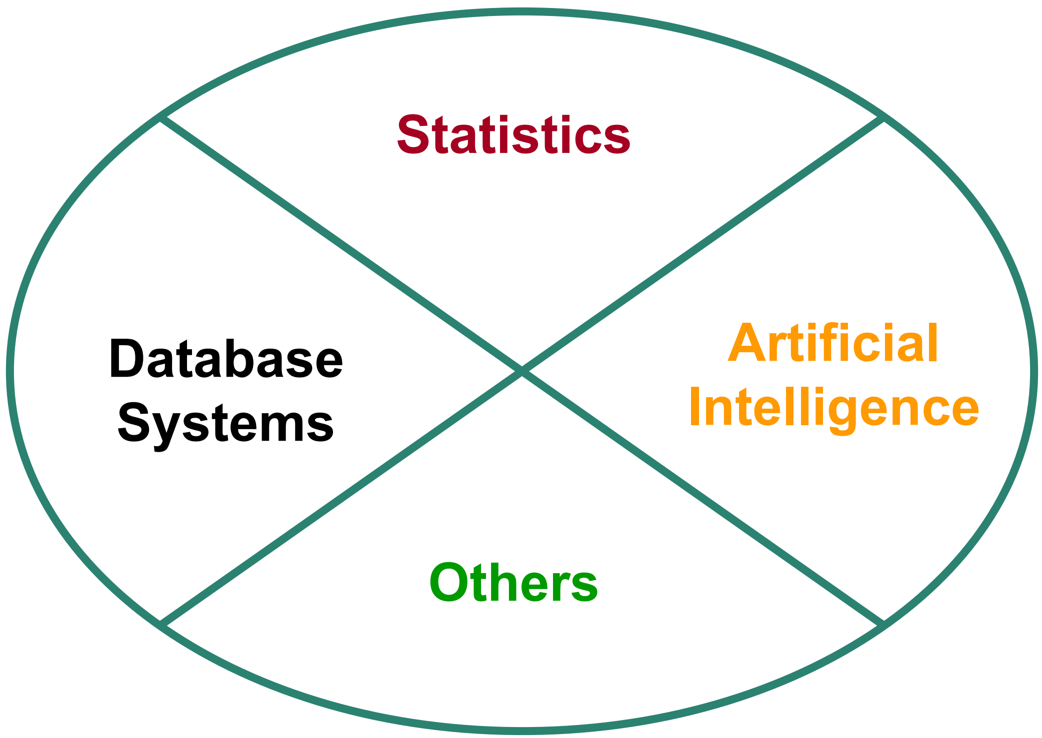
Knowledge Management (KM) –

Efforts to capture, store, and deploy knowledge

Knowledge Discovery Process



Confluence of Statistics, AI and DBMS



NIST/TRC SOURCE Data System

- ❖An extensive and historical repository system: experimental thermophysical and thermochemical properties and relevant measurement information
- ❖Comprehensive coverage: up to 100 properties of over 2 million records for 32,000 chemical systems including pure compounds, mixtures, and reaction systems
- ❖Information Complexity: bibliographic information, chemical system identifiers, sample purity, property values and the relevant state variables, estimates of uncertainty, measurement methods, and a variety of other metadata
- ❖Sophisticated data organization and structures: a complete measurement is described by six types of records involving 37 tables, within each data table up to near 80 columns

Challenges in Massive-Scale Data Collection and Data Entry

How to automate the process of data collection and extraction with a goal of covering essentially all experimental data available from the world's literature in this field?

How to ensure efficient data entry with error prevention mechanism?

- **What is the bibliographic source?**
- **What chemical compounds were studied?**
- **What was the nature of the particular chemical samples ?**
- **What mixtures or reactions involving the samples were studied?**
- **What properties were measured?**
- **How were the properties measured?**
- **What were the numerical values obtained?**

Challenges in Assurance of Data Integrity

❖A large-scale numeric database without critical evaluation may have an error rate of 2-5%

❖Common occurred errors – (a) typographical; (b) unit-conversion; (c) report interpretation; (d) metadata compilation; (e) errors in original report …

How to detect anomalous values and how to enforce the scientific data integrity?

Challenges in Data and Models Evaluation

- Scientific experiment is a complicated process
- Experimental data tend to have uncertainty or error
- Evaluation of scientific data is extremely difficult, no way to guarantee its absolute correctness
- The true value of physicochemical property needs repeat examinations
- The above problems are also true for models

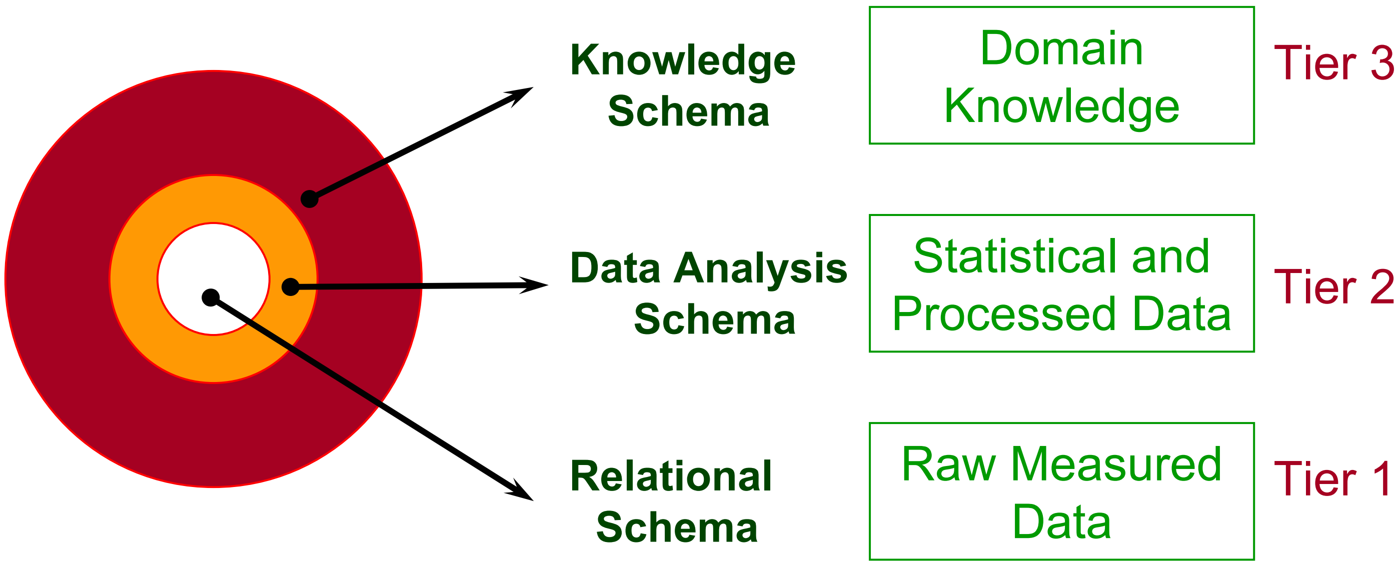
Domain Knowledge

1. Relational database principle, definition and structure of physicochemical data, database schema and relations
2. More than one hundred thermophysical and thermochemical properties for pure compounds, binary and ternary mixtures, and chemical reactions
3. Measurement techniques and sample purity
4. Chemical characteristics of substances, mixture and reaction systems
5. Relations of chemical structure and property, …
6. Knowledge gained from statistical data analysis…
7. …

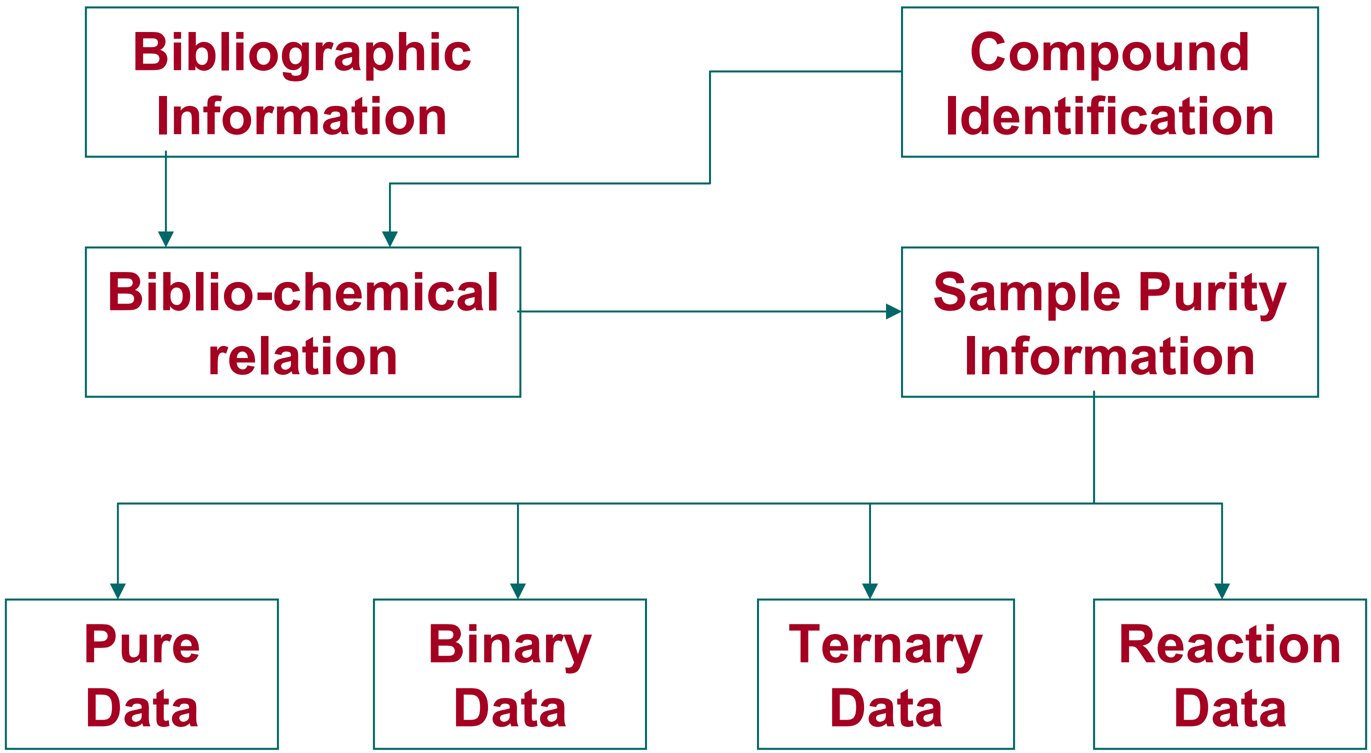
General Process of knowledge Management

- ✓ Requirements analysis – Identify the scope of the knowledge-based system, typically in terms of its expected competency
- ✓ Conceptual Modeling – Based on the defined scope, create a glossary of terminology (concepts) for the application domain and define interrelationships between terms
- ✓ Knowledge capture – Real-time and automated capture from Web and full-text as well as knowledge discovery in databases
- ✓ Knowledge base construction – Creation of knowledge schema based on the conceptual modeling (in the form of rules, facts, cases, or constraints)
- ✓ Reasoning and Validation – Testing the competence of knowledge base against requirements

Three-Tier Infrastructure for Supporting Knowledge Management



Relational Schema



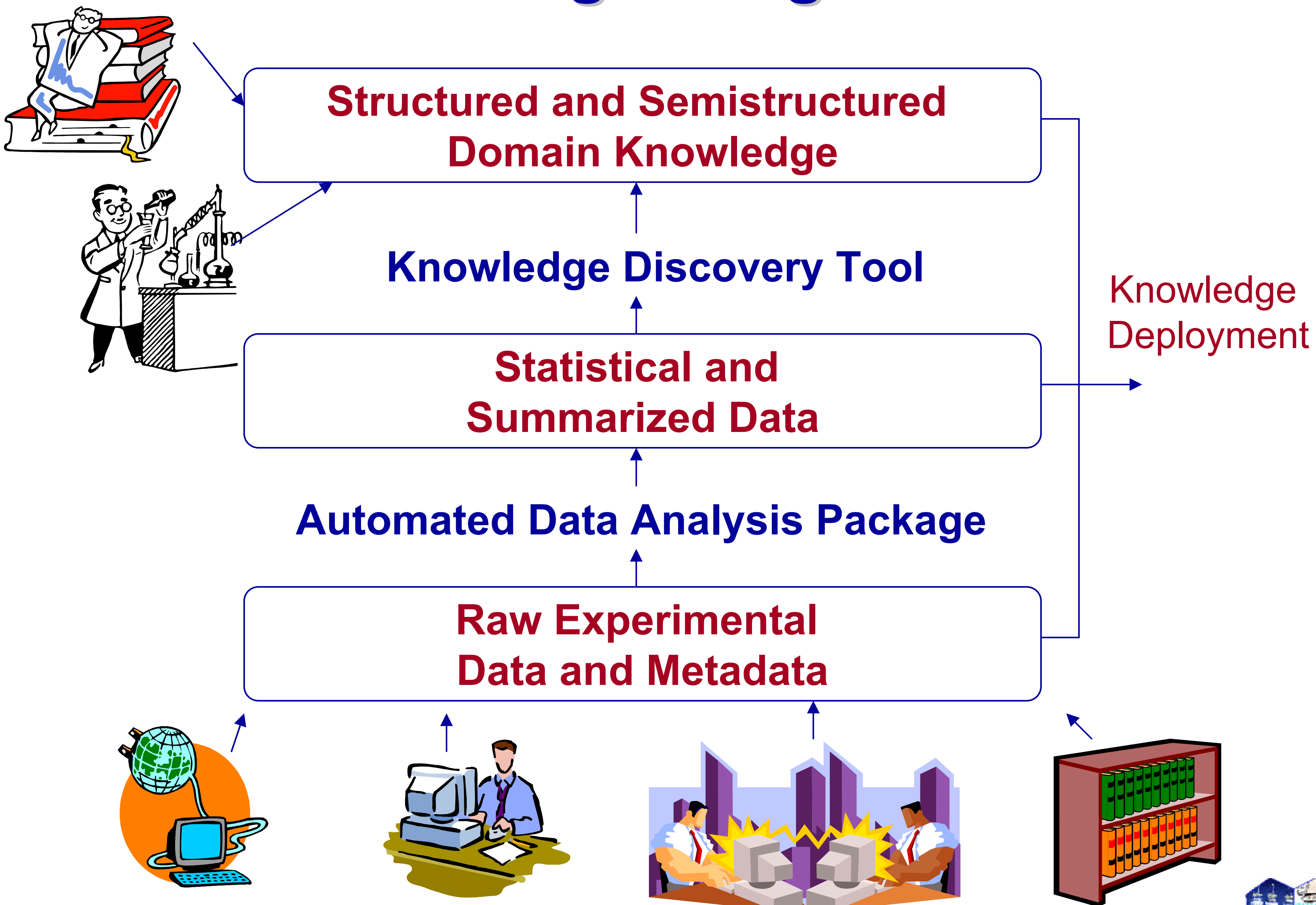
Data Analysis Schema

- ***Content:*** weighted-average-value, deviation, average-deviation, suggested-uncertainty, total-data-points, selected-data-points and others
- ***Nature of Content:*** time-stamped statistical data for each compound/property in the database
- ***Method of Generation:*** periodically via a process of data cleaning, data transformation, and data integration to provide information from historical perspective and are typically summarized
- ***Functionality:*** to enable a combination of statistical data with experimental data as a basis to discover and capture relevant knowledge from the database

Knowledge Schema

- Structured Domain Knowledge
  - thermophysical and thermochemical principles glossary
  - over 100 predefined property dictionary
  - relational data structure and relation map
  - recommended-data index
  - predict-model class
  - property-compound/system summary
- Semistructured Domain Knowledge
  - characteristics of pure substances and chemical systems
  - experimental case study
  - abstract and summary of original reports
  - All factual text domain knowledge

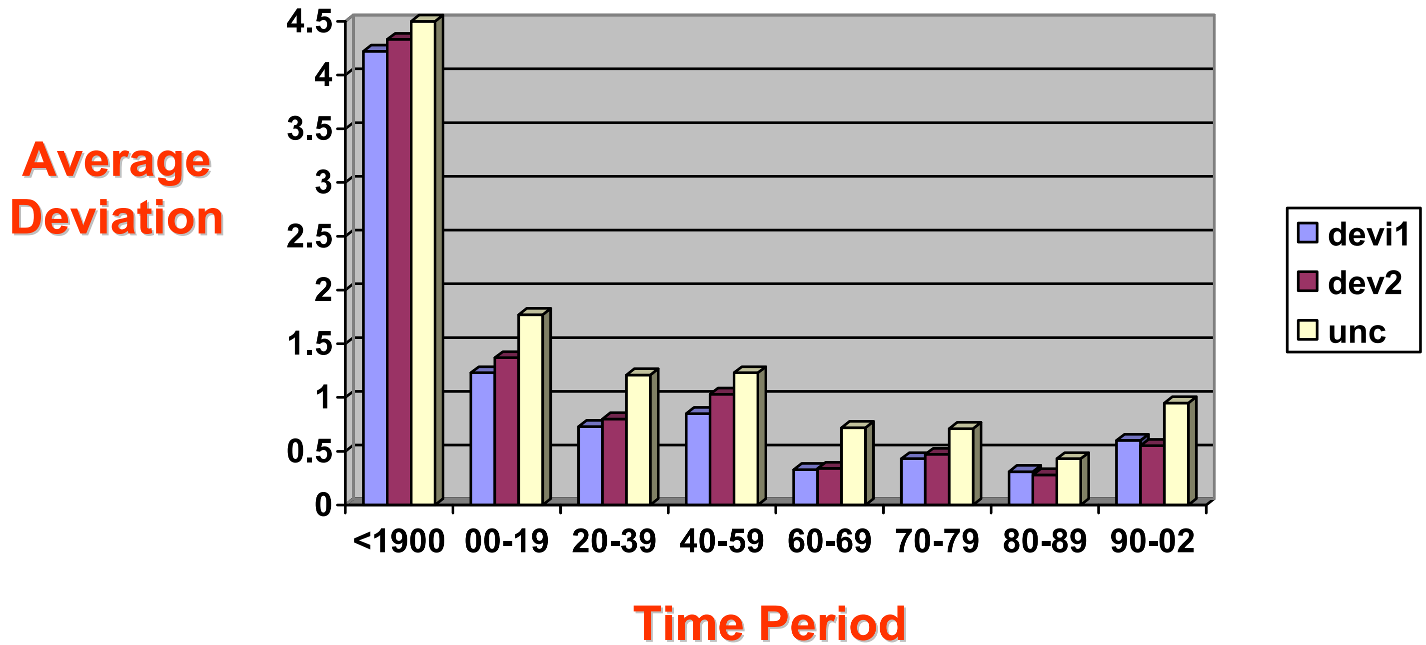
Knowledge Integration



Data Mining and Knowledge Discovery (1)
- Time factor on the average deviation of critical data

<u>Methods:</u> Analyzing deviations of the Tc data published through 1822-2001. Select the compounds for which there are multiple experimental data or their property values have been measured with high accuracy.

<u>Average Deviation (K) of Tc Values through 1822 to 2001:</u>



Data Mining and Knowledge Discovery (2)

- Author factor on the average deviation of critical data

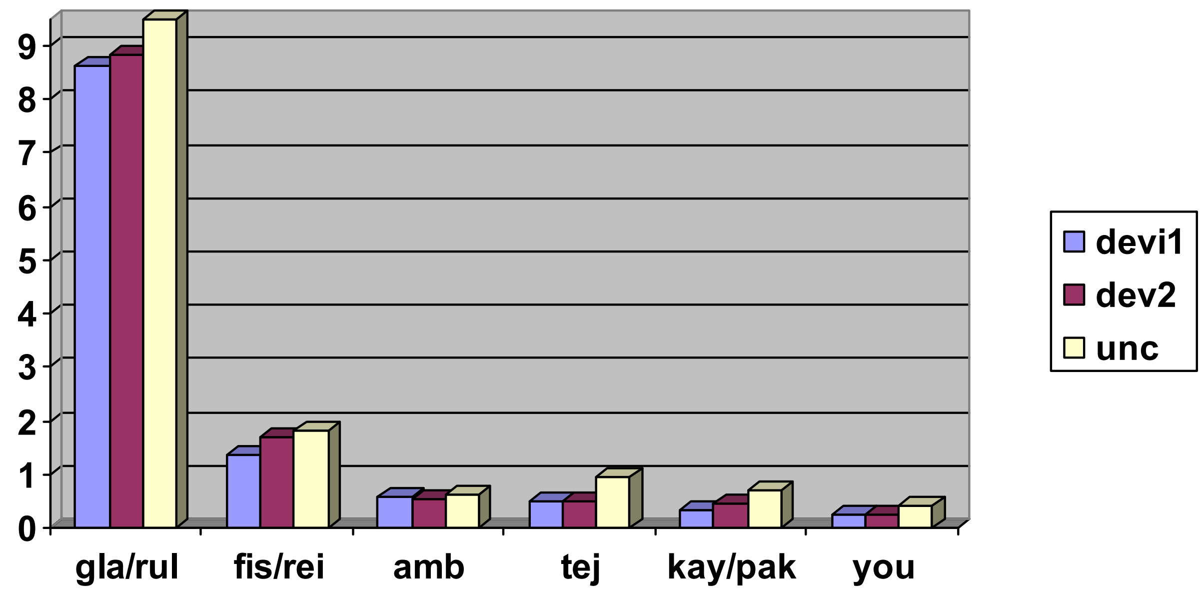
Methods:

Analyzing deviations of the data published by different authors. Select the compounds for which there are multiple experimental data or their property values have been measured with high accuracy

Average Deviation (K) of Tc Values by Different Authors:

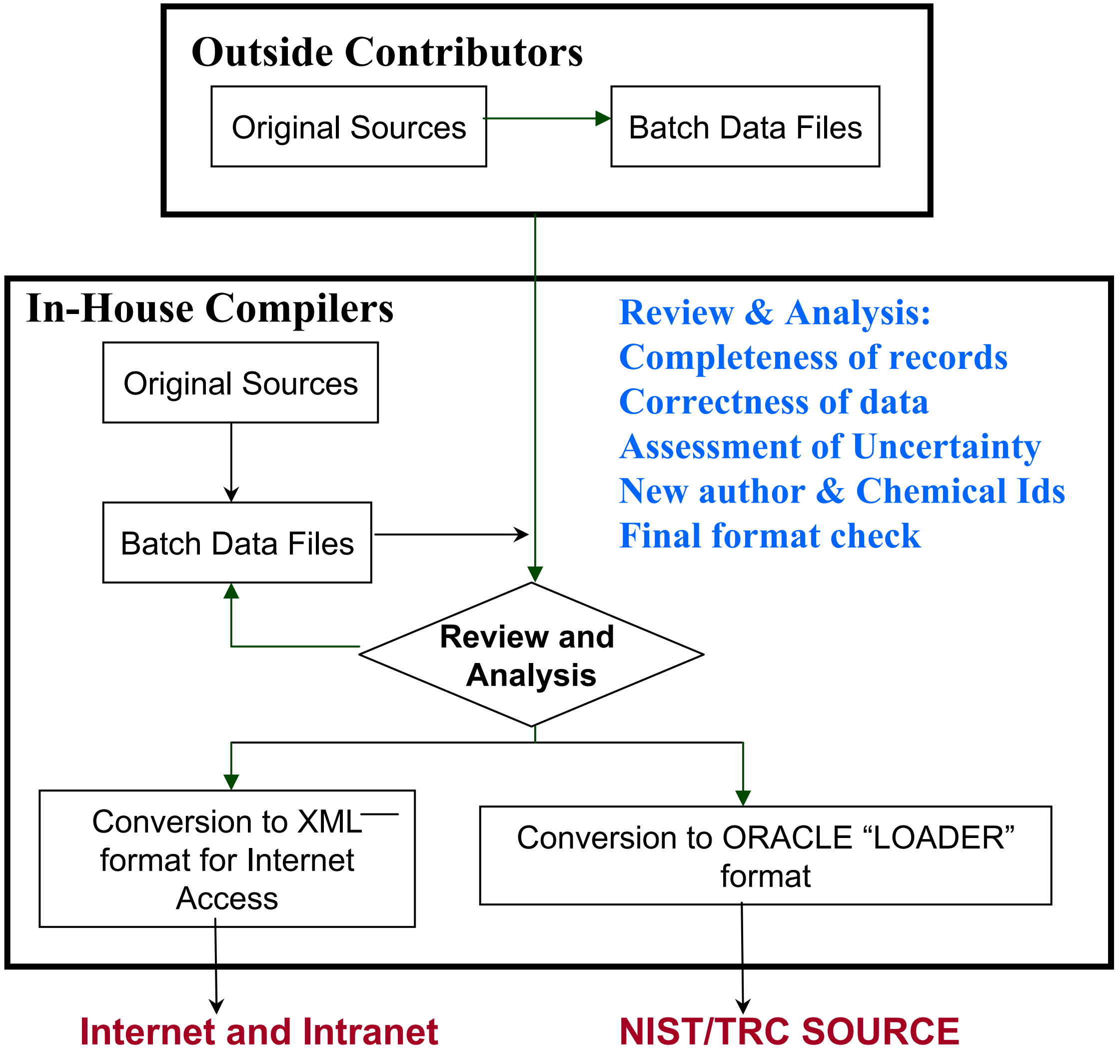
Average Deviation

Glaser, F./Ruland, H.
Fischer/Reinkober
Ambrose, D.
Teja, A. S.
Kay, W.D./Pak, S.C.
Young, C. L.



Authors

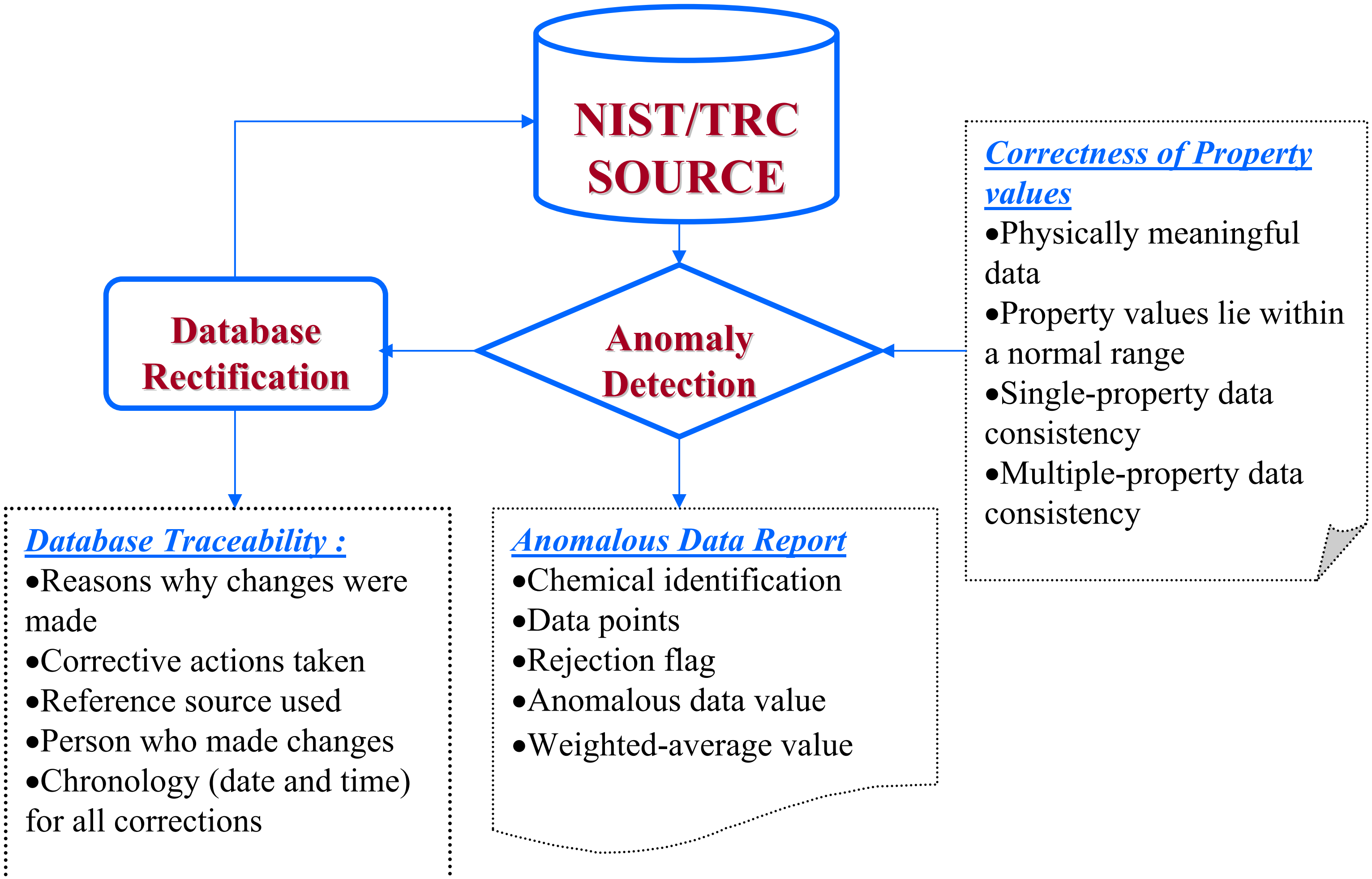
Intelligent-Guided Data Entry



Functionalities

- distributed batch files
- preparation, review, and editing
- hierarchical system
- guiding extraction
- assuring completeness
- validating prepared data
- Automate uncertainty assessment

Automation of Anomaly Detection



Systematic and Integrated Process