CODATA 18th International Conference: Frontiers of Scientific and Technical Data (29 September - 3 October, 2002)

Prototype of TRC Integrated Information System for Physicochemical Properties of Organic Compounds: Evaluated Data, Models and Knowledge

Xinjian Yan,

Qian Dong, Xiangrong Hong, Robert D. Chirico, Michael Frenkel

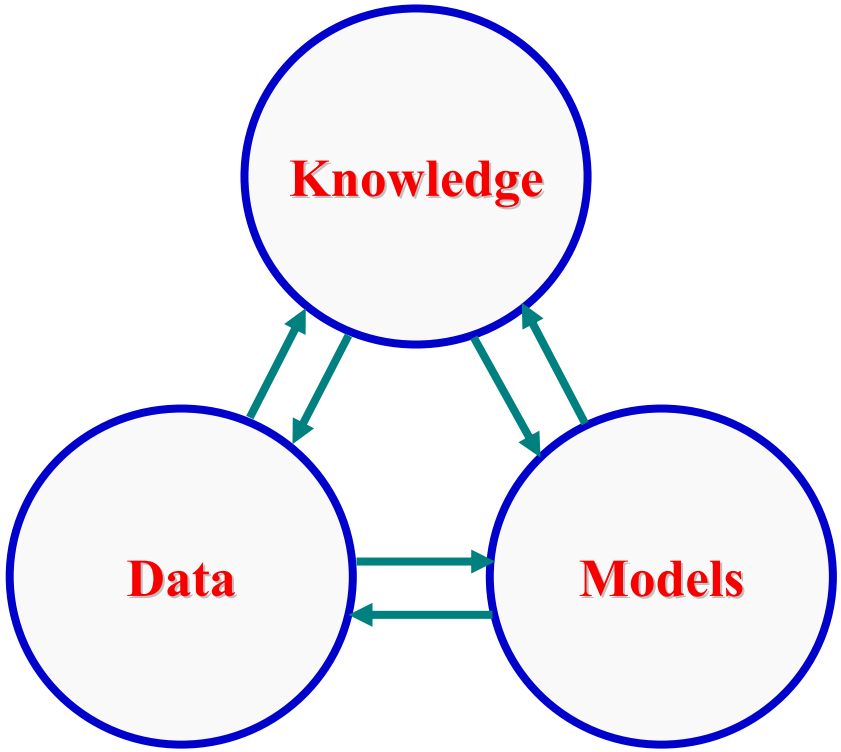
Thermodynamics Research Center (TRC)

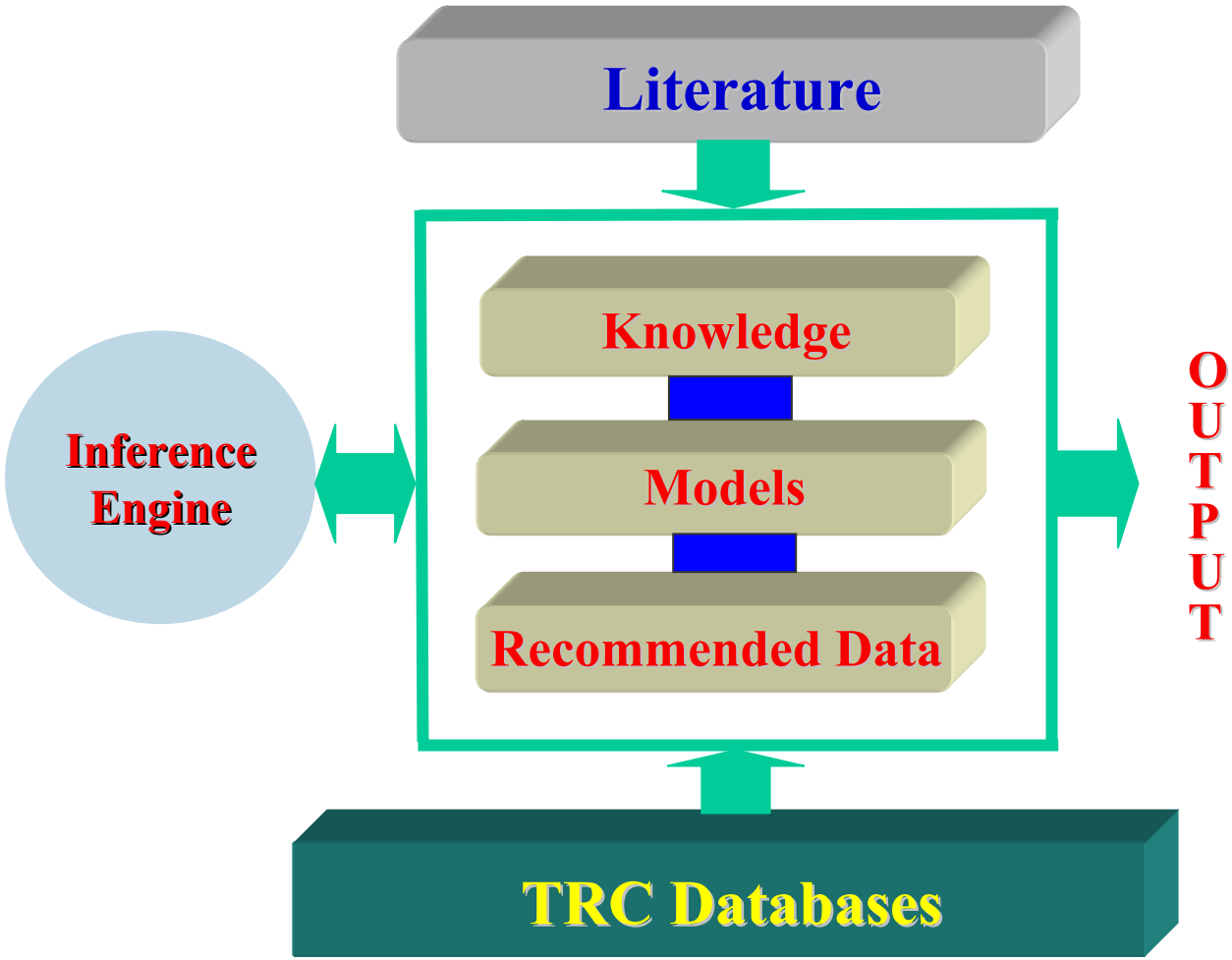
National Institute of Standards and Technology

Introduction

- **Requirement:** Industrial and scientific developments require high quality data and models
- **Key Point:** High quality data system needs strong support from comprehensive knowledge base
- **Aim:** Develop a system with high quality data and models fully supported by domain knowledge

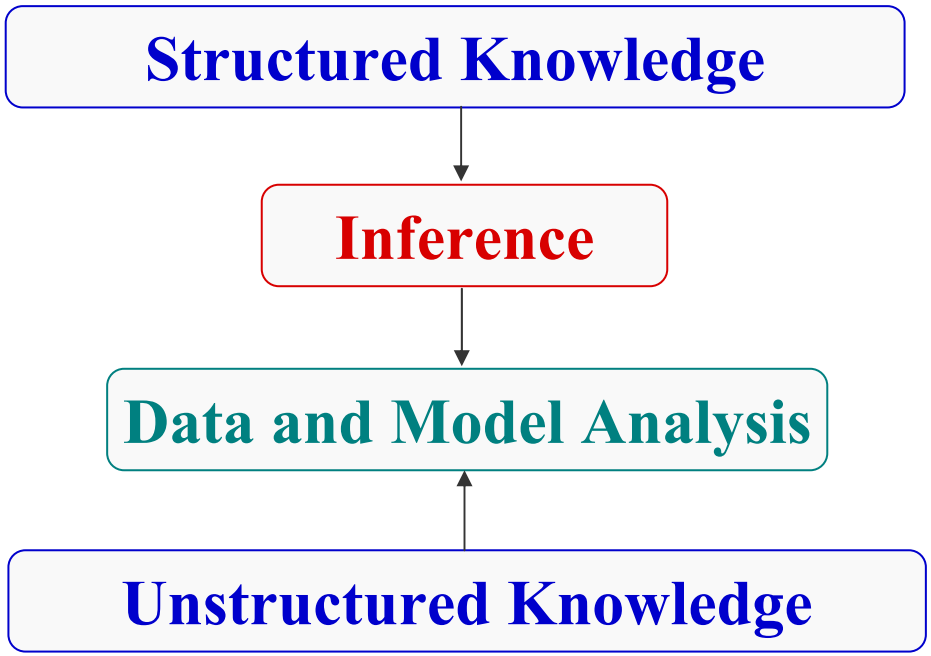
The Relationship between Data, Model and Knowledge





TRC Integrated Information System (TIIS) Structure

The Support of Knowledge to Data and Model Analysis



Data Background - TRC Databases

- **Databases:**

  Source Database, Table Database, Density Database, Vapor Pressure Database, Ideal Gas Database, etc.

- **A Comprehensive Physicochemical Data System:**

  Source Database contains more than 100 physical and chemical properties, over 2 million experimental records for 32,000 chemical systems (pure compounds, mixtures, and reaction systems)

Information for Recommended Data (RD)

Detailed information is crucial for a good understanding of data. The following information has been prepared for recommended data (also for experimental data).

- The uncertainty values of RD
- The number of data points used for obtaining RD
- The discreteness of the data used to process RD
- The description about the selection of RD
- The grade of RD

Data Processing for RD

- For compounds having multiple values, a weighted average method is used to obtain recommended data
- For compounds having only one or two values, the data are inspected by:
  - A. Theories and thermodynamics relationships
  - B. Comparison with the values from models
  - C. Comparison with other well characterized sources
  - D. Similar compounds
- For doubtful data, original articles are reviewed

Criteria and Methods of Evaluating Models

The major problem in model evaluations is that very little attention is paid on the prediction abilities of models. The following factors have been considered in our evaluating and selecting of models for TIIS.

- **Prediction ability**
- **Complexity of compounds used in developing and testing models**
- **Diversity of compounds used in developing and testing models**
- **Reliability of each parameter (how many and how well data were used in obtaining each parameter)**
- **Similarity analysis**

Example of Prediction Ability of Models

WJ is a simple model with about 20 parameters, while MP is a model using 167 parameters. MP is better than WJ in correlating data, but not in predicting for new compounds.

Ranges of deviations (Dev, K) of the Tc predicted by WJ and MP group contribution models for the compounds having experimental data reported between 1996 and 2001, and number of compounds in each range.

	WJ	MP
Correlated result		
Total number	**467**	**434**
Dev, K	**6.3**	**5.7**
Predicted result		
Total number	**48**	**42**
Dev, K	**10.2**	**13.0**

Examination on the Reliability of Parameters

167 MP groups and estimated occurring frequency in compounds that have Tc data published before 1987 (N1) and up to now (N2).

	N1	Groups	N1	Groups	N1	Groups
Summary	0	40	1	36	2	19

			1987	Now
			361	600
Part1	Part2	Bond	N1	N2
C=	rC=	1	0	1
CH_2	C*	1	0	0
CH_2	I	1	0	0
rC	rrC=	1	0	0
rC	rC=	1	0	0
S	S	1	0	1
rCH	rrCH	1	0	2
CH	CH=	1	0	0
CH	C=	1	0	0
CH	CHO	1	0	0
C	C=	1	0	1
CH_2	C=	1	0	0
=C=	=O	2	0	0
rC	rrC	1	0	0
C=	=C=	2	0	0
CH=	CN	1	0	1
CH=	COOo	1	0	0
CH=	COOH	1	0	0
CH=	CHO	1	0	1
CH=	F	1	0	0
CH=	rC=	1	0	1
CH=	C*	1	0	0
CH=	C=	1	0	0
CH=	=C=	2	0	0
C*	C*	3	0	0
rO	rN=	1	0	1
CH_3	Br	1	0	0
CH_3	I	1	0	0
CH_3	rC	1	0	4
rC=	I	1	0	0
rC=	NH	1	0	0
rC=	COOH	1	0	0
Cl	CO	1	0	0
rC=	CO	1	0	1
O	CO	1	0	0
rC=	Br	1	0	6
CO	CO	1	0	0
rC=	rrC=	1	0	0
H	CHO	1	0	0
NH_2	NH	1	0	0
H	COOH	1	1	1
CH_3	F	1	1	1
C	O	1	1	23
CH*	CH*	3	1	1
CH*	C*	3	1	1
rC=	COO	1	1	3
C=	Cl	1	1	4
C	COOH	1	1	1
C	rC=	1	1	7
CH=	COO	1	1	2
CH=	O	1	1	2
CH=	Cl	1	1	1
CH_3	C*	1	1	1
CH=	CH=	1	1	1
CH_3	CH_3	1	1	1
rC=	CN	1	1	1
CH_2	CHO	1	1	8
CH_3	CN	1	1	1
CH_3	NO_2	1	1	1

Complexity of Organic Compounds - Definition

Group/ complexity	=1	>1		
CH	1	1	CH3-CH(CH3)-CH3	= 2
C	2	2		
C=C (double bond)	2	2		
=C=	2	2		
C*C (triple bond)	2	2		
F, Cl, Br, I	3	5	2 (when groups >4)	
CN	3	4		
N	3	4		
NC	3	4		
S	3	4		
SH	3	4		
CHO	4	10		
CO	4	10		
COO	4	10		
COOH	4	10		
N=	4	10		
NH	4	10		
NH=	4	10		
NH2	4	10		
NO2	4	10		
O	4	10		
OH	4	10	OH-CH2-CH2-OH	= 18
SO	4	10		
SO2	4	10		

Ring / complexity	3	5	Including fused ring
Terminals / complexity	6 (C=1)	3 (C=2)	1 (C=3)

C atoms / complexity

1- 10	1	11- 20	2	21- 30	3
31- 40	4	41- 50	5	> 50	6

Example of Complexity for Compounds Having Critical Temperature (Tc) Data

	CN	AC
Tc before 1996*	500	14
Tc after 1995**	100	21

CN - Compound Number; AC - Average Complexity

* 500 compounds having critical temperature reported before 1996.

** 100 new compounds reported between 1996 and 2001.

Example of Using the Information from Similar Compounds to Judge Uncertainty of the Value Estimated by Models

Property Estimation by Thermodynamic Models

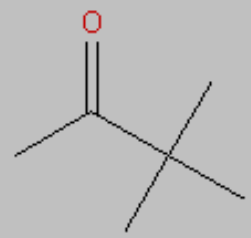
CASRN 75978 | Calculate | Info | Print | Exit

Select a Property

- normal freezing point
- (•) normal boiling point
- critical temperature
- critical pressure
- critical density/volume
- acentric factor
- standard state Gibbs energy of formation
- enthalpy change of atmospheric boiling
- enthalpy change of atmospheric melting
- ideal gas enthalpy of formation at 298 K
- ideal gas Gibbs energy of formation at 298 K
- vaporization heat at normal boiling point
- Gibbs energy at 1 bar
- Enthalpy of ideal gas
- Entropy of ideal gas
- vapor pressure
- Surface tension
- density of liquid
- heat capacity
- Thermal conductivity
- Viscosity
- Thermal diffusivity

Select a Model

- (•) Joback
- Constantinou/Gani
- Marrero/Pardillo



Normal boiling point by Joback method = 387.3 (K)

The Information of Similar Molecules

No	CASRN	Similarity	PValue	MValue	PVError	Formula
1	67641	4	329.3	321.9	7.4	C3H6O
2	563804	5	367.1	367.2	0.1	C5H10O
3	78933	6	352.7	344.8	7.9	C4H8O
4	108101	7	389.1	390.1	1	C6H12O
5	96220	8	375	367.7	7.3	C5H10O
6	107879	8	375.2	367.7	7.5	C5H10O
7	141797	9	402.8	394.6	8.2	C6H10O
8	463821	10	282.6	310.6	28	C5H12
9	591786	10	400.3	390.5	9.8	C6H12O
10	589388	10	396.5	390.5	6	C6H12O
11	106354	12	420.4	413.4	7	C7H14O
12	123193	12	416.8	413.4	3.4	C7H14O
13	110430	12	423.8	413.4	10.4	C7H14O

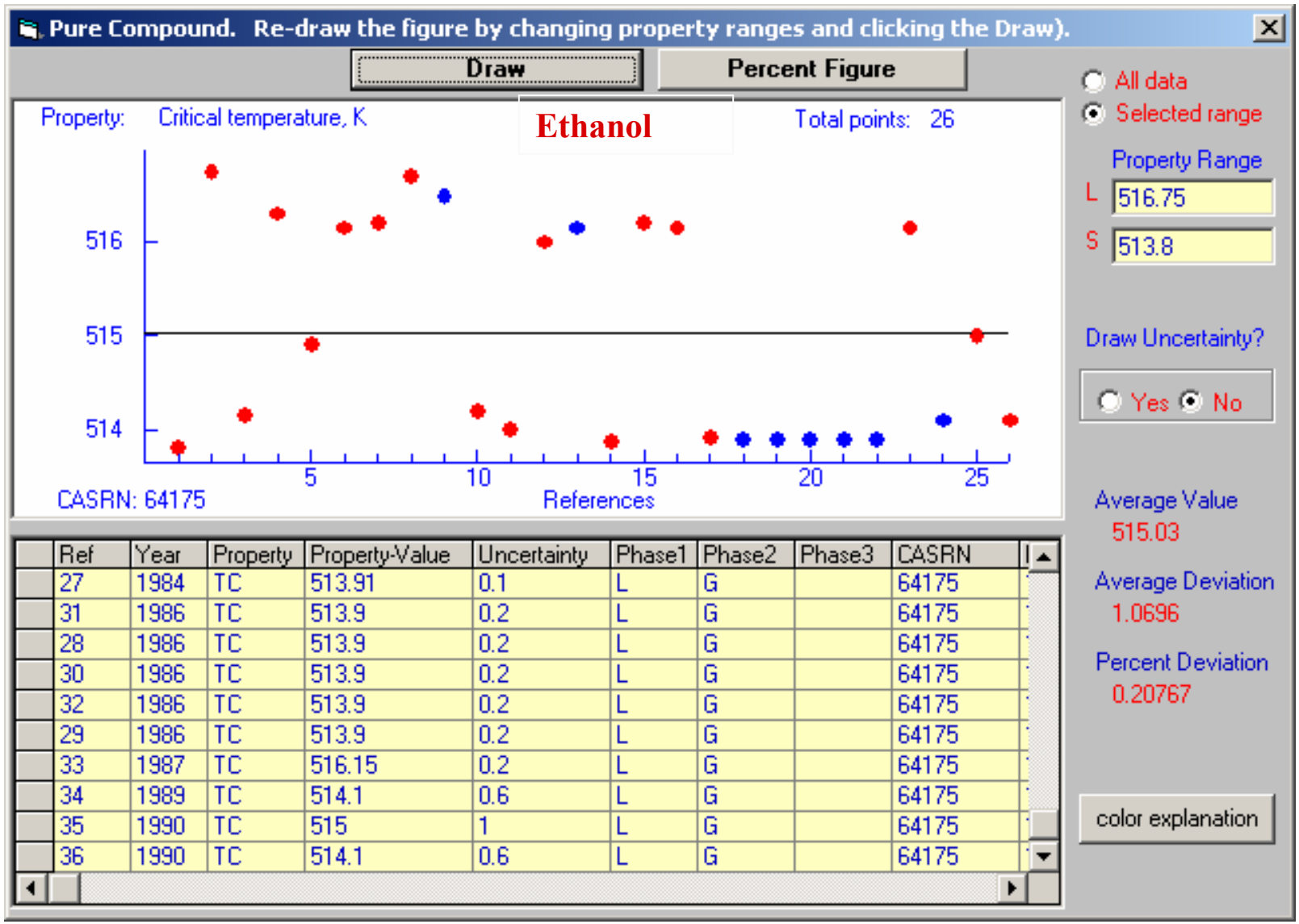
Knowledge is the key to evaluate and understand scientific data as well as models

- 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 their absolute correctness
- The true value of physicochemical property needs repeated experimental examination
- The above problems are also true for models

Domain Knowledge

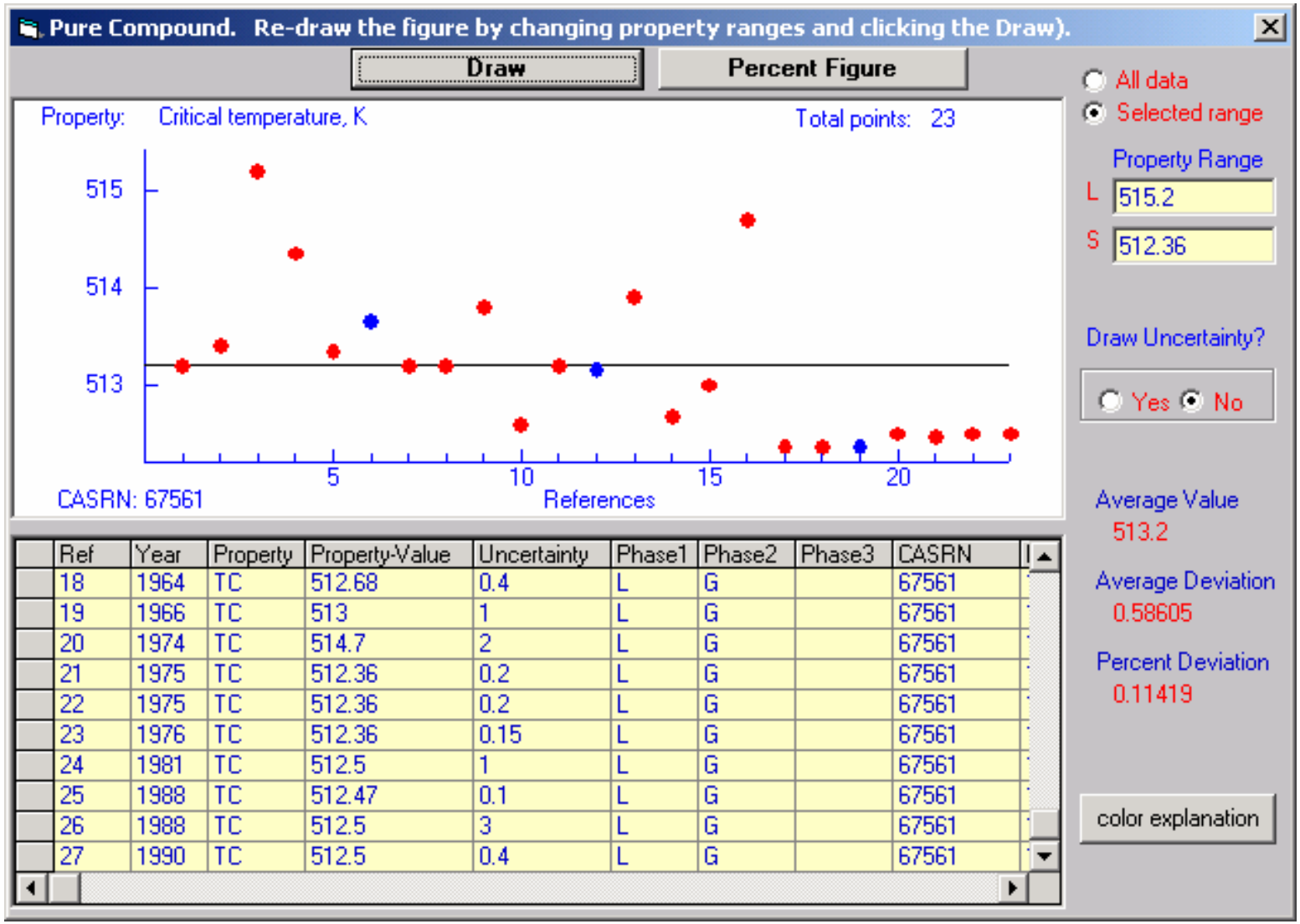
- Thermophysics theory and concept
- Experimental and theoretical research methods
- Evaluation and comment on experimental data
- Compound physical and chemical characteristics
- Models (introduction, evaluation and comment)
- Molecular structure and interaction information
- Terminology
- Unit
- ……

Example about Knowledge and the Selection of Ethanol's Recommended Critical Temperature



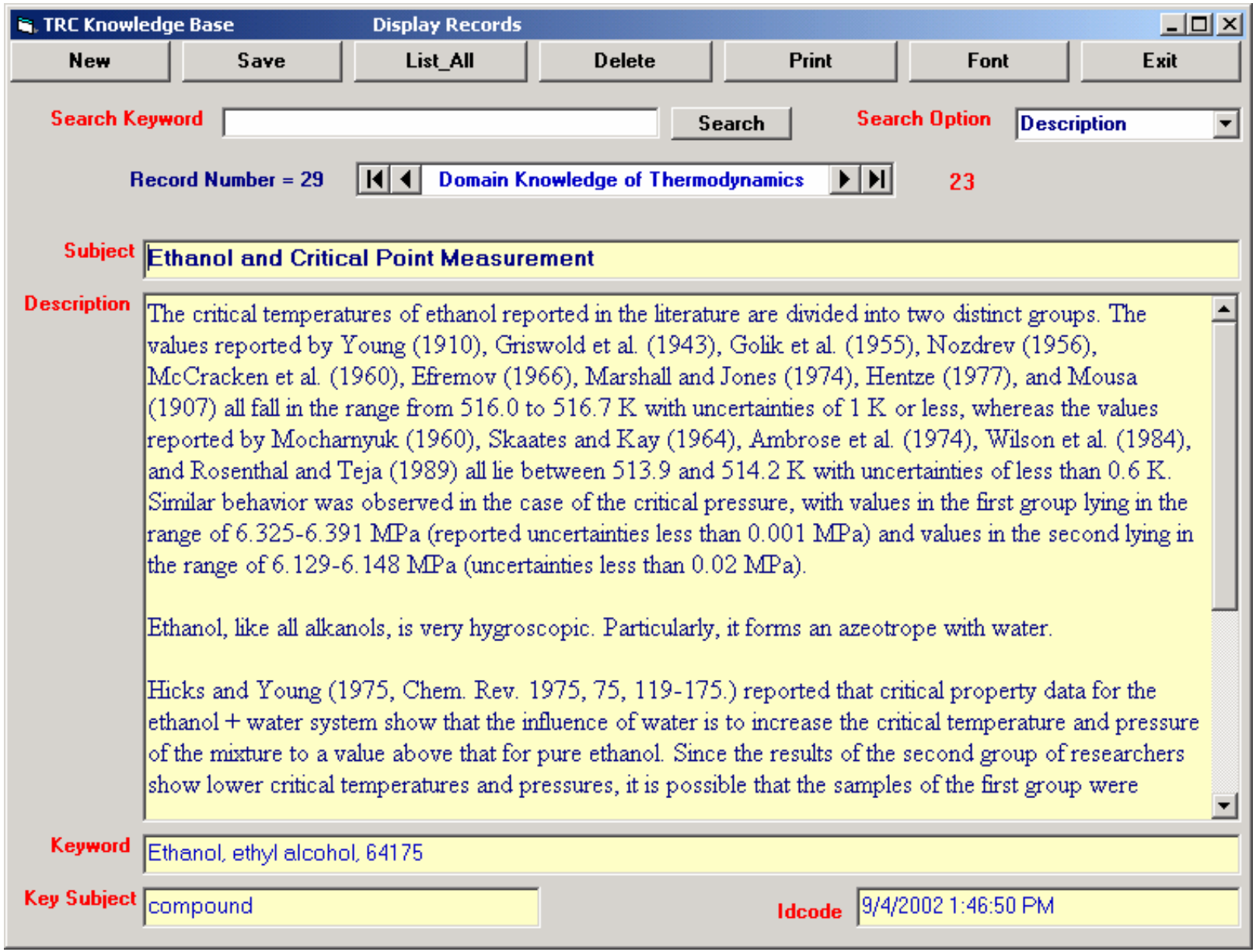
Ref	Year	Property	Property-Value	Uncertainty	Phase1	Phase2	Phase3	CASRN
27	1984	TC	513.91	0.1	L	G		64175
31	1986	TC	513.9	0.2	L	G		64175
28	1986	TC	513.9	0.2	L	G		64175
30	1986	TC	513.9	0.2	L	G		64175
32	1986	TC	513.9	0.2	L	G		64175
29	1986	TC	513.9	0.2	L	G		64175
33	1987	TC	516.15	0.2	L	G		64175
34	1989	TC	514.1	0.6	L	G		64175
35	1990	TC	515	1	L	G		64175
36	1990	TC	514.1	0.6	L	G		64175

Example about Knowledge and the Selection of Ethanol's Recommended Critical Temperature



Ref	Year	Property	Property-Value	Uncertainty	Phase1	Phase2	Phase3	CASRN
18	1964	TC	512.68	0.4	L	G		67561
19	1966	TC	513	1	L	G		67561
20	1974	TC	514.7	2	L	G		67561
21	1975	TC	512.36	0.2	L	G		67561
22	1975	TC	512.36	0.2	L	G		67561
23	1976	TC	512.36	0.15	L	G		67561
24	1981	TC	512.5	1	L	G		67561
25	1988	TC	512.47	0.1	L	G		67561
26	1988	TC	512.5	3	L	G		67561
27	1990	TC	512.5	0.4	L	G		67561

Example of Knowledge Supporting System



Summary

- Uncertainty is everywhere
- Our knowledge on uncertainty is very limited
- Our awareness on uncertainty is low
- Knowledge is crucial to decrease the uncertainty
- For building a high quality information system, it is necessary to develop a strong ability for analyzing the uncertainty of data, models and text information