Teaching the Big Scientific Data Analysis

Boris Sedunov

Russian New University, Russia, https://orcid.org/0000-0002-0300-773X

Abstract: The contemporary Human activity utilizes huge volumes of digital data to solve efficiently multiple socio-economic, scientific and technical problems. Now the big data analysis is mainly oriented to the socio-economic sphere with a goal to lift the profit. The science and technology to penetrate deeper in the nature of objects and systems under investigation prefer to limit the analysis area, concentrating, for example, only on properties of extra pure materials or isolated systems. In science the analysis should be convergent and the initial data may be and should be regularized to diminish the input data errors. The clusters now are considered as a new or still unknown state of matter. The big thermophysical data analysis appears as the most informative way to discover the properties of clusters in pure real gases, because the continuous spectrum of bound states in clusters prevents from the spectroscopic way for clusters' properties evaluation. The goal of the paper is to teach the main principles of the scientific regular data convergent analysis basing on the author's experience to extract clusters' properties in pure real gases from regularized experimental thermophysical big data.

Keywords: Big data analysis, Regularized data, Convergent analysis, Cluster, Molecular interaction

To understand the actual World as it is,
not as we should wish it to be,
is the beginning of wisdom.

Bertrand Russell

Introduction

Various kinds of the Human activity now utilize huge volumes of digital data (Mayer-Schonberger 2013) to rise efficiency in social, economic, political, healthcare, environmental, scientific and technical problems resolution. It launched the Big Data technology, which includes the collection, analysis, and visualization of vast amounts of digital information (Smolan 2012). To extract from the big data useful knowledge and insights a large attention is paid to the computerized big data analysis development (Maheshwari 2014). Now the big data analysis is in a large extent oriented on profitability of the socio-economic sphere. In this direction large perspectives are waited from the Artificial Intellect and Machine Learning technologies (Cielen 2016), which

Sedunov, B. (2021). Teaching the big scientific data analysis. In M. Shelley, W. Admiraal, & H. Akcay (Eds.), *Proceedings of ICEMST 2021--International Conference on Education in Mathematics, Science and Technology* (pp. 134-146), Antalya, TURKEY. ISTES Organization.

are supposed to support the decision making in multivariate situations.

The Scientific Big Data Analysis

In contrast to the socio-economic sphere, in the science and technology researchers to penetrate deeper in the nature of objects and systems under investigation used to limit the area of the analysis, concentrating, for example, only on properties of extra pure materials or isolated systems. For this task the interactive computer aided analysis (Sedunov 2012) seems to be more effective than the direct computer programming for the final result. In this approach the computer is not the main solver, but an effective adviser, providing intermediate and final results in an informative visual form. In the scientific analysis the initial data may be and should be regularized, thus diminishing the input data errors. The raw experimental data regularization means their critical evaluation to remove the data poorly convergent with the bulk of data and the data interpolation to remove large occasional deviations from the general trend (Frenkel 2012). A good example of a huge collection of regularized scientific experimental data present the US National Institute of Standardization and Technology (NIST) electronic databases of pure materials thermophysical properties, such as the Webbook (NIST 2021) or the Database (NIST 2013). The Database provides the second and the third virial coefficients for selected pure gases used in the Semiconductor Industry. The Webbook contains data for 75 pure substances of various types, the convergent analysis of which permits to penetrate in the nature of clusters and molecular interactions (Sedunov 2012). Unlike the paper Handbooks, the Webbook provides an optimal steps selection and up to 12 decimal digits in presentation of data.

The convergent analysis means:

- a mutual correspondence of raw scientific experimental data collected from all World;
- a correspondence of the regularized experimental data to universal polynomials;
- a correspondence of the processing mathematics to the physical nature of values;
- a correspondence of the individual models to the general physical picture of the Cluster World;
- the data processing with account of the thermodynamics correlations between different values.

The hidden parameters extraction from big experimental data

The main task of the big data analysis in science is to understand and evaluate the hidden mechanisms, governing the behavior of complex systems. The blind spot in thermodynamics are clusters, which are considered as a new state of matter (Yarris 1991), or to be more exact, still unknown state of mater. Their detailed investigation is important both for science and technology. For clusters cognition it is essential to build realistic and comprehensive models of molecular interactions. The main task of the big data analysis in thermal physics of real gases is to extract from experimental thermophysical data the parameters of clusters and molecular interactions. This task becomes much more realizable for pure gases; therefore, we start our analysis from extra pure gases. The hidden parameters extraction from experimental data is known as an inverse problem

(Aster 2012), which results strongly depend on the initial data and processing accuracy: small errors in initial data turn out to be huge errors at the finish. And the data processing methods should correspond to the physics of systems under investigation. So, the convergent processing of raw experimental data is a very important operation for the scientific analysis to be successful.

The goal of this paper is to describe the main principles and the possible traps of the convergent scientific big data analysis basing on the author's experience in the clusters' properties and molecular interactions extraction from regularized experimental thermophysical big data for pure real gases. The possible traps result from misunderstandings of the cluster physics. So, the deeper is penetration in the cluster physics, the more correct are the data processing methods. Some of these principles and the recommendations how to escape from possible traps may be extended to other scientific and technical fields.

Method

A rich structure of pure real gases filled with multiple cluster fractions opens ways for statements of new relations and introduction of still unknown variables. Among these statements a very important is: the n-particle cluster fraction, including the monomer fraction, in a pure real gas behaves as an ideal gas:

$$P_n = RT\, D_n, \qquad (1)$$

where P_n and D_n are partial pressure and molar density of the n-particle cluster fraction, with n = 1 for monomers. The second statement is: in a pure real gas the basic particles' molar Gibbs energy is the same for all clusters and monomers! The basic particles of a fluid are particles corresponding to its chemical nature. Their molar density D is shown in the thermophysical databases, as the fluid total density. In a neat fluid we have only one type of basic particles and one chemical potential G for its basic particles. The third statement is: the total pressure P and the integral molar density of all free moving particles (clusters and monomers) $D_p = \Sigma\, D_n$ correspond to the ideal gas law: $P = RT\, D_p$.

The most effective and informative variables selection

For the computer aided analysis of pure fluids' precise thermophysical properties it was important to select the adequate method and the most informative variables. For equilibrium clusters concentrations, as for chemical compounds, the fundamental Mass action law **(**Koudryavtsev 2001) is valid, and its proper utilization may bring valuable results. To use this law, we recommend the series expansions of thermophysical values by the new variable - the monomer fraction density (MFD), D_m (Sedunov 2008):

$$D_n = C_n\, D_m^{\,n}, \qquad (2)$$

where Cn is the apparent equilibrium constant for n-particle complexes, including real and virtual clusters. The virtual clusters are not bound by the attraction forces, but appear in series expansions due to repulsions. Only D_m, as an effective argument for series expansions, provides the correspondence of the n-th expansion term to properties of the n-particle complexes: including bound by attraction forces clusters and instant colliding complexes. For this reason, the series expansion (2) may be named canonical.

The monomer fraction density

In pure real gases the variable D_m means an average molar density of basic particles, temporarily not bound in clusters. This definition is rather vague, but the D_m can be defined by the phenomenological way via the molar Gibbs energy G named also as the chemical potential of basic particles. In the chemical equilibrium the chemical potential for all basic particles G is equal to the chemical potential G_m of monomers (Sedunov 2008):

$$G = G_m = G_{int} + RT \ln (D_m V_q). \quad (3)$$

The G_{int} is the part of G connected with internal movements of basic particles: molecular rotations and vibrations, $V_q = h^3 N_A^4 / (2\pi MRT)^{3/2}$ is the molar quantum volume (Kittel 1969) proportional to the cube of the thermal de Broglie wavelength; h is the Plank's constant, M is the basic particles' molecular mass in kg/mol, N_A is the Avogadro number, R is the universal gas constant. From the Equation (3) we come to the differential Equation for D_m (Sedunov 2008):

$$\partial D_m / \partial P |_T = D_m / (RTD). \quad (4)$$

The Figure 1 shows the D_m in comparison with D and $D_p = P/RT$. The D_p means the integral molar density of all free moving particles: clusters and monomers. For the differential equation (4) numerical solution we have found an original expression (5) containing the differential ΔD_P both in the numerator and the denominator:

$$D_{m\,i} = D_{m\,(i-1)} (1 + \Delta D_P / (2D_{(i-1)})) / (1 - \Delta D_P / (2D_i)). \quad (5)$$

As an initial condition, we suggest: $D_{m\,1} = 2 D_{P\,1} - D_1$. For this condition to be precise the initial pressure P_1 should be in the ideal gas zone.

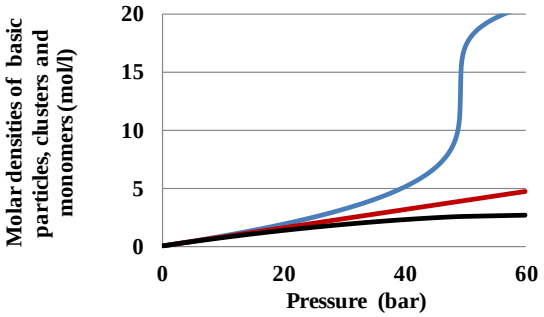

Figure 1. Molar densities of all basic particles D (blue line), free moving particles D_p (red line) and the monomer fraction D_m (black line) in Argon at a supercritical temperature, T = 151 K.

The Figure 1 shows that at pressures over 10 bar the three lines diverge and near critical pressure the total density D quickly grows due to the growth of the clusters number and the number of particles in them. If we remove the MFD from D and D_p, we find the total density of basic particles contained in the cluster fractions (D - D_m) and the molar density of clusters, as free moving particles (D_p - D_m), Figure 2.

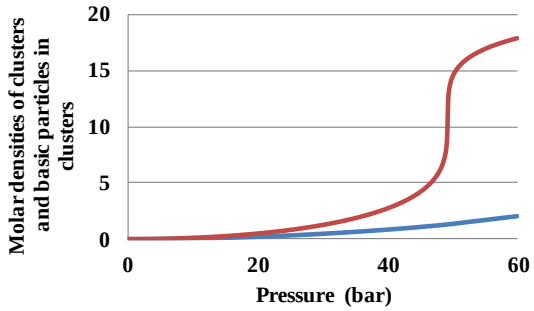

Figure 2. The total density of basic particles contained in clusters (D - D_m) (red line) and the molar density of clusters (D_p - D_m) (blue line) in Argon at a supercritical temperature, T = 151 K.

The difference between two lines quickly grows near critical pressure, demonstrating growth of the particle numbers in clusters. The ratio (D - D_m) to (D_p - D_m), shown at the Figure 3, demonstrates the average number of basic particles in a cluster.

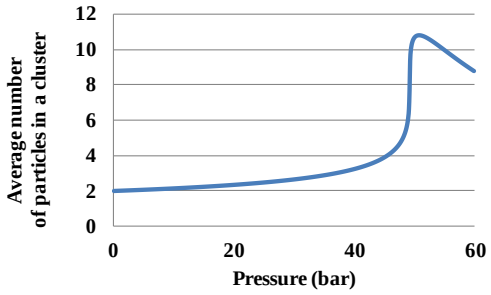

Figure 3**.** The basic particles average number in an Argon cluster at a supercritical temperature, T = 151 K.

We see that there is a wide range of pressures, where dimers dominate and a wider range, where the average

number of basic particles in a cluster is under three. But then the average number of particles in a cluster quickly grows. We should not take the average numbers for maximal reachable ones. It may be shown that the maximal numbers can overcome 1000. At supercritical pressures the average number of particles in a complex starts falling. But this complex is not a cluster, but a large pore in the infinite cluster of the liquid-like fluid (Sedunov 2013, March). So, the MFD may be found from known isothermal data for total pressure P and total density D of basic particles and used for the clusters' properties analysis. And the MFD based cluster expansion corresponds to the Mass action law! Due to this remarkable feature the neat equilibrium fluids stay as an adequate platform for clusters and molecular interactions investigation and as an advanced platform for thermodynamics and molecular physics education (Sedunov 2020).

The apparent PDT equilibrium constants

The figure 4 shows the Pressure-Density-Temperature (PDT) interaction function $C_{2+} = (D - D_p)/D_m^2$ to be expanded in a series by D_m.

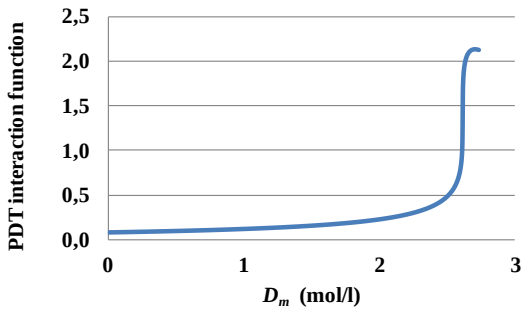

Figure 4. The Pressure-Density-Temperature (PDT) interaction function $C_{2+} = (D - D_p)/D_m^2$.

To compare our analysis method with the theory of virial expansion (Mayer 1977). we consider the PDT cluster expansion. Its theory is reflected in equations (1-5). Only the second coefficient C_2 reflecting dimers' PDT relations is equal to - B - the second virial coefficient. The figure 1 helps to understand why C_2 has an opposite sign to B: C_2 results from the expansion of D_p by D_m, which is smaller than D_p, but B follows from the expansion of D_p by D, which is larger than D_p. The figure 5 shows the $C_2(T)$ changing sign at the Boyle point. It shows that at this point the real dimer fraction becomes weaker than the virtual one.

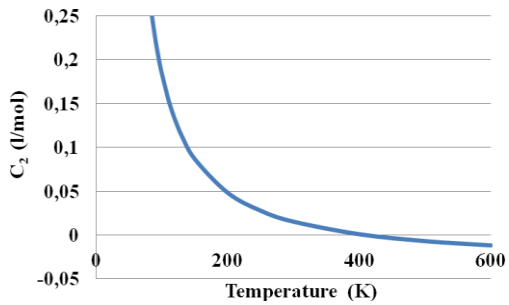

Figure 5. The temperature dependence of the apparent equilibrium constant for dimers in Argon.

The Figure 6 demonstrates the found C_3 coefficient coincidence with its model expression $C_{3mod} = 2C_2^2$. The physical sense of the model expression is: the trimer is formed by forming one of two new dimers around an existing dimer. It gives the factor 2. The same pair interaction apparent equilibrium constants C_2 for old and new dimers tell about an open linear structure of the trimer isomer in this case. So, the big data analysis gives not only quatitative measure of the $C_3(T)$, but also the structure of the formed trimer.

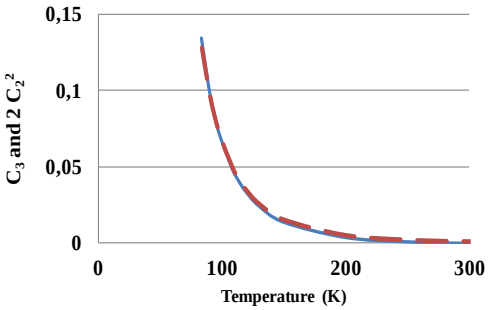

Figure 6. The coincidence of the C_3 (T) (blue line) with its model $2\ C_2^2$ (red dashed line) in Argon.

The $C_3(T)$ deviation from the linear model would say about the soft structural transition to the closed triangular trimer structure. The possibility to express the C_3 via C_2 and the perfect coincidence of the physically clear model $2\ C_2^2$ and real function C_3 (T) in a wide temperature range witness in favor of the canonical cluster expansion, while the virial expansion does not provide the B_3 correlation with B_2. It results from a wrong argument D for virial expansion, being the sum of the clusters' basic particles partial densities. The equation (2) tells about the Mass action law, if D_m serves as an expansion argument, but with D as an argument we never come to the Mass action law. So, a wrong argument D in virial expansions resulted in the fundamental Mass action law ignoring.

The fluid potential energy

To find the clusters' bond parameters we introduce another informative variable - the fluid potential energy (Sedunov 2012):

$$U = E(T, P) - E(T,0), \qquad (6)$$

where E(T, P) and E(T,0) are molar internal energies at the pressure P and zero pressure, correspondingly. We expand in series by D_m the positive potential energy density: W = - UD. The n-th terms W_n of this expansion give the corresponding contributions of n-particle complexes to the total potential energy density W (Sedunov 2013).

The fluid potential energy-based cluster analysis

The $ln\ (W_n(T))$ provide estimations of clusters' bond energies via the tangents of lines slopes, figure 7. For

dimers the estimation of E_2 in K coincides with its estimation through the constant volume heat capacity C_v and molar internal energy $E(T, P)$: $E_2 = -\ limit\ \partial C_v / \partial E|_T$. A correlation of two different big data analysis methods confirms both methods.

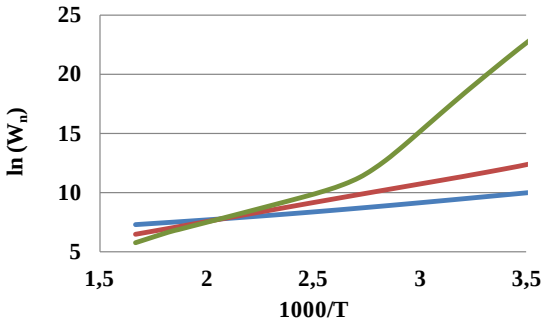

Figure 7. The graph for the Water vapor clusters averaged bond energies estimation: dimers (blue line), trimers (red), tetramers (green).

Results

The soft structural transitions discovery

The bond energies E_n at the figure 7 are determined by tangents of the corresponding lines slope. For Water clusters bond energies E_n are shown at the Table 1.

Table 1. Water vapor n-particle clusters bond energies En averaged for T: T1 < T < T2

n	2	3	4	5
T_1 (K)	274	274	274	390
T_2 (K)	378	600	350	600
E_n (K)	1647	3299	14742	4926

The results at the Table 1 show that $E_n \sim (n - 1)\ E_2$ for trimers in the whole investigated temperature range but for tetramers only at $T > 390$ K. The bend of tetramers' graph in the temperature range 350 -390 K shows the soft structural transition from tightly bonded tetramers at $T < 350$ K to loosely bonded at $T > 390$ K. A similar transition was noticed also in the tetramers of Methanol. But in Neon the coexistence of normal trimers with giant bonds trimers was discovered.

Linear and 3D clusters in noble gases

The bond energies E_n in the noble gas Helium E_n grow with the numbers n of particles in small clusters linearly: $E_n = E_d\ (n - 1)$, where E_d is the dimers' bond energy, figure 8. The teacher should explain the difference between open linear and closed structures of clusters.

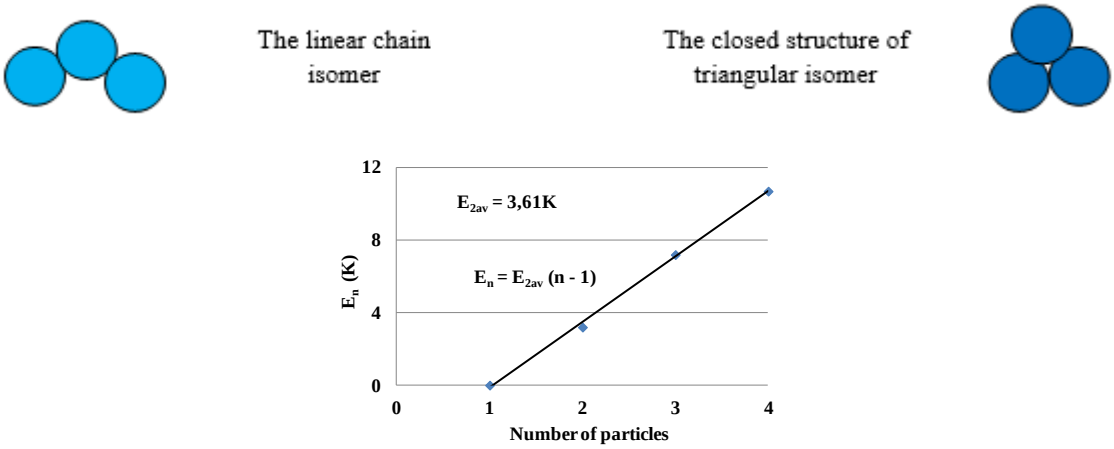

Figure 8. The linear Helium small clusters' bond energy E_n dependence on the number n of particles in a cluster.

But in Argon for clusters with numbers n of particles from 2 to 5 we see a quasi-parabolic dependence of E_n / E_d on $(n - 1)$, figure 9. The found E_n / E_d ratios are close to 1, 3, 6, 9 values. The corresponding numbers of bonds may be prescribed to dimer, 2D closed triangular trimer, and 3D clusters: tetramer and double tetramer. So, the bond energies estimations show the clusters' structures.

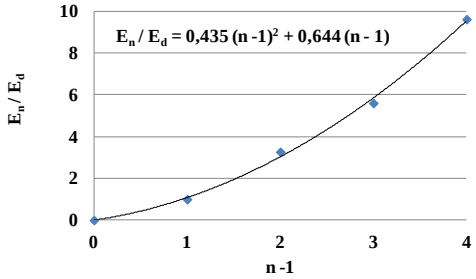

Figure 9. The quasi parabolic bond energy E_n dependence on the number n of particles in an Argon cluster.

The linear E_n / E_d dependence on the n for Helium origins from the spherical symmetry of the atoms' electron shells, resulting in a linear chain structure of clusters. But the parabolic dependence for Argon tells about directional bonding of atoms, resulting in the tightly bonded clusters. The E_n / E_d estimation for Argon shows the role of the electron shells structure in the directional bonds formation. So, the computer aided big data analysis in the physics of clusters may act as the computer tomography, showing the nanosized clusters' structures (Sedunov 2013, March) instead of the short wavelength spectroscopy, which could destroy clusters with small bond energy.

The traps in the thermophysical data analysis

The raw experimental data contain two sorts of errors: random and systematic. The random errors can be reduced with the interpolation by high order polynomials. But at the polynomial type selection the teacher

should take into account the natural laws, which rule the system under investigation. Otherwise the generalization may lead to wrong data, as it happened with viscosity data for Xenon, figure 10.

The erroneous viscosity data for Xenon

To study the viscosity η density dependence we use the value $V_{vis} = \partial(\eta/\partial D)|_T /\eta_0$, which may be named as the characteristic viscosity volume. Here η_0 is the zero-pressure viscosity. The V_{vis} dependence on pressure is shown at the figure 10 for Xenon and Krypton. The viscosity data have been taken from the Webbook (NIST 2021). We see that V_{vis} for Krypton behaves quite naturally, but V_{vis} for Xenon shows unnatural growth in the ideal gas zone. This error contradicts with the theory, which states that the ideal gas viscosity does not depend on density.

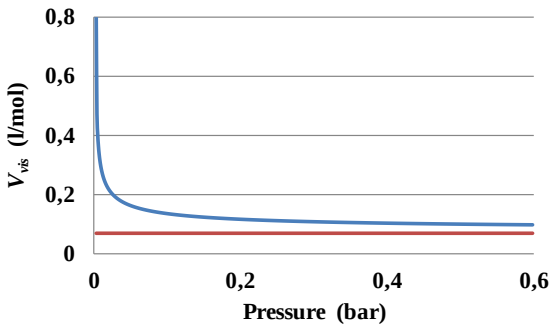

Figure 10. The characteristic viscosity volume V_{vis} dependence on pressure for Xenon at 180 K (blue line) and Krypton at 150 K (red line).

A wrong utilization of universal constants

The universal constants, such as the Universal gas constant R, are regularly updated. But the data in databases reflect their value for the moment of data generation. So, to analyze the scientific data the teacher should find the apparent constant value, for example, R_a as the zero-pressure limit of P/TD, figure 11. For different gases in the Webbook (NIST 2020) R_a may differ. For Helium R_a = 8,3148 J/(molK), for Argon R_a = 8,31451 J/(molK). So, to escape from this error we should for every gas find its own R_a value.

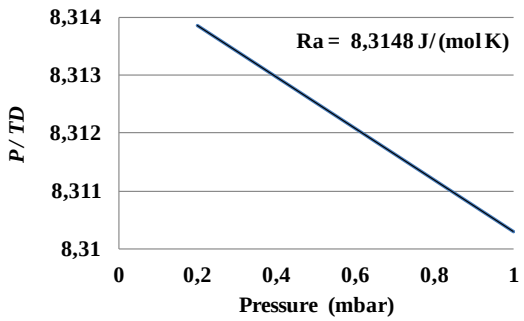

Figure 11. Estimation of the apparent gas constant R_a for Helium data from the NIST Webbook (NIST 2021).

Discussion

Wrong utilization of the virial cluster expansions

The virial expansions perfectly help to regularize raw experimental data. But their utilization to investigate clusters' properties (Mayer 1977) does not correspond to the Mass action law (Koudryavtsev 2001) and gives erroneous results. Feynman wrote (Feynman 1972) that the virial approach silently supposes that the virial expansion n-th term corresponds to the n-particle cluster properties. But it is not so! The expansion terms with $n > 2$ collect contributions from different cluster fractions. And Feynman concludes that the virial cluster expansions program was stopped (Feynman 1972). The next shortcoming of the virial cluster expansion is in its concentration only on the PDT relations, which do not give the clusters' bond parameters.

Wrong interpretation of virtual contributions in the expansion coefficients

To explain the change of sign in the second virial coefficient a new factor: $(exp(-E/RT) - 1)$ was introduced (Mayer 1977). This factor excludes the Boltzmann law application to clusters and builds a boundary between clusters and chemical compounds. But the (-1) addition to the Boltzmann factor results from virtual contributions, being inevitable at series expansions of experimental data. So, the Boltzmann law is valid for clusters, but we should understand the virtual nature of the apparent equilibrium constants falling in the negative zone. The teacher should be ready to meet the revolutionary concepts, such as the virtual cluster.

The Lennard-Jones model limits

The spherical symmetry of atomic interactions in the widely used Lennard-Jones model (Lennard-Jones 1924) results from a simplified vision of the electron shells. For noble gases the spherical symmetry was confirmed only for Helium and explained by the spherical symmetry of its s-type electron shells. All other noble gases demonstrate the directional bonding near the triple point. And trimers in Neon possessing giant bonds run away from the Lennard-Jones model. It shows a significant role of quantum effects in atomic and molecular interactions. The teacher should see and explain students the obsolete theories and their ranges of application. So, the big scientific data analysis opens new and promising directions for research!

Conclusion

The computer aided analysis of big regular thermophysical data results in fundamental discoveries:

- the monomer fraction density based canonical type of the cluster expansion, as an opposition to virial expansions;
- unknown before bond energy values for different gases, temperatures and cluster isomers;
- a new concept of the soft structural transitions in the cluster fractions;

- apparent equilibrium constants for combinations of real and virtual clusters;
- directional bonding in noble gases, rejecting the Lennard-Jones model.

No trivial results of the big scientific data analysis in the real gases thermal physics confirm the validity of its main principles. The found principles of the big data analysis may be transferred to other scientific spheres.

Recommendations

- In the big data analysis teaching we should see the difference between socio-economic and scientific spheres.
- The initial data for the big data scientific analysis should be regularized: they should be carefully selected to remove the outstanding data; they should be interpolated with polynomials, reflecting scientific relations.
- The teacher should select the data processing mathematics, which reflects the scientific relations.
- The teacher should be able to introduce and teach new variables and new concepts.
- The teacher should be able to revise traditional theories and to escape from traps in the analysis.

References

Aster, R.C., Borchers, B., & Thurber, C. (2012). *Parameter Estimation and Inverse Problems*, 2nd edition. Elsevier.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science: Big Data, Machine Learning, and more, using Python tools*, Manning publications.

Feynman, R. (1972). *Statistical mechanics; A set of lectures*, *W.A.* Benjamin, Inc., Massachusetts.

Frenkel, M. (2012). *Industrial Use of Molecular Thermodynamics Workshop*, Lyon, France.

Kittel, Ch. (1969). *Thermal physics*, New York: Wiley & Sons, Inc.

Koudryavtsev, A.B., Jameson, R.F. & Linert, W. (2001). *The Law of Mass Action*, Berlin: Springer-Verlag.

Lennard-Jones, J. E. (1924). On the Determination of Molecular Fields, *Proc. R. Soc. Lond. A 106 (738)*, 463 477, doi:10.1098/rspa.1924.0082.

Maheshwari, A. (2014). *Data Analytics Made Accessible*, Kindle Edition.

Mayer, J.E. & Goeppert-Mayer, M. (1977). *Statistical Mechanics*, 2nd Ed. New York: Wiley & Sons.

Mayer-Schonberger, V., & HCukier, K. (2013). *Big Data. A Revolution That Will Transform How We Live, Work and Think*, An Hachette UK Company, Great Britain.

NIST. (2013). *Database Thermophysical Properties of Gases Used in the Semiconductor Industry*.

NIST. (2021). *Webbook Thermophysical Properties of Fluid Systems*. General format. Retrieved from http://webbook.nist.gov/chemistry/fluid.

Sedunov, B. (2008). Monomer fraction in real gases. *Int. J. of Thermodynamics, 11(1)*, 1-9.

Sedunov, B. (2012). Equilibrium molecular interactions in pure gases. *J. of Thermodynamics, 2012*, Article ID 859047, 13 pages.

Sedunov, B. (2013). Proceedings of the JEEP-'13, *Joint European days on Equilibrium between Phases*, Nancy, France, MATEC Web of Conferences, 3, 01002, DOI:10.1051/matecconf/20130301002 .

Sedunov, B. (2013, March). Proceedings of the JEEP-'13, *Joint European days on Equilibrium between Phases*, Nancy, France: MATEC Web of Conferences, 3, 01062, DOI:10.1051/matecconf/20130301062

Sedunov, B. (2020). An advanced platform for thermodynamics education. Part one: Small density pure real gases, *Int. J. of Thermodynamics, 23(3),* 224–233.

Smolan, R., & Erwitt, J. (2012). *The Human Face of Big Data*, Against All Odds Productions.

Yarris, L. (1991). *Clusters: A New State of Matter*, Berkeley LAB Publication Archive. General format. Retrieved from https://www2.lbl.gov/Science-Articles/Archive/clusters.