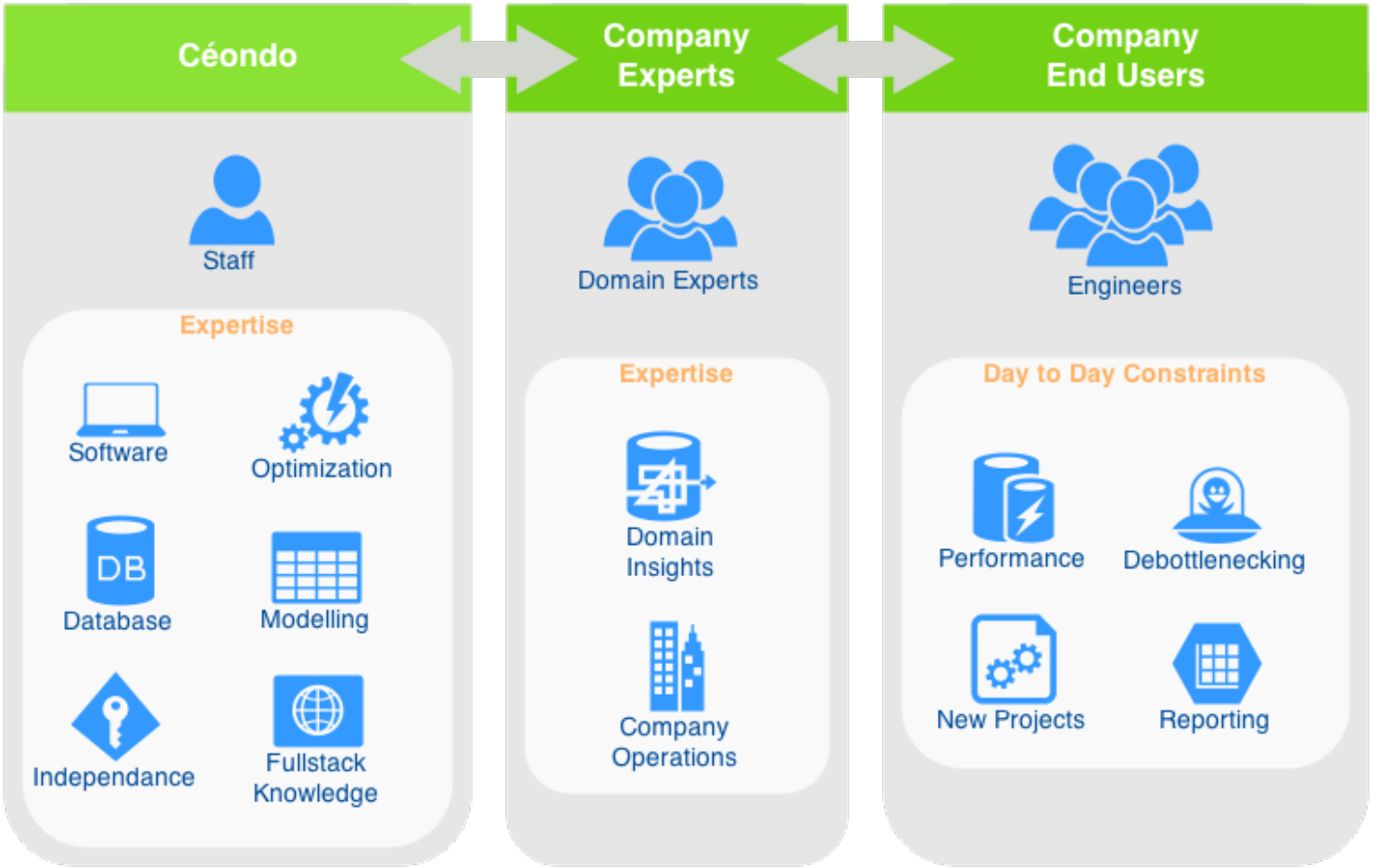
Property Modelling Issues

When Property Prediction Goes ¨Boink¨

Loïc d'Anterroches, PhD
Céondo GmbH

Céondo

Céondo, since 2007



- Modelling, Fluid Phase Equilibria, Chemical Properties & Databases

From Other Presentations

- Bayer: “Model validation needs skills in **statistics** often neglected in process engineering curricula.”
- Alfa Laval: It starts with data, **property prediction** at the start of design.

Outline

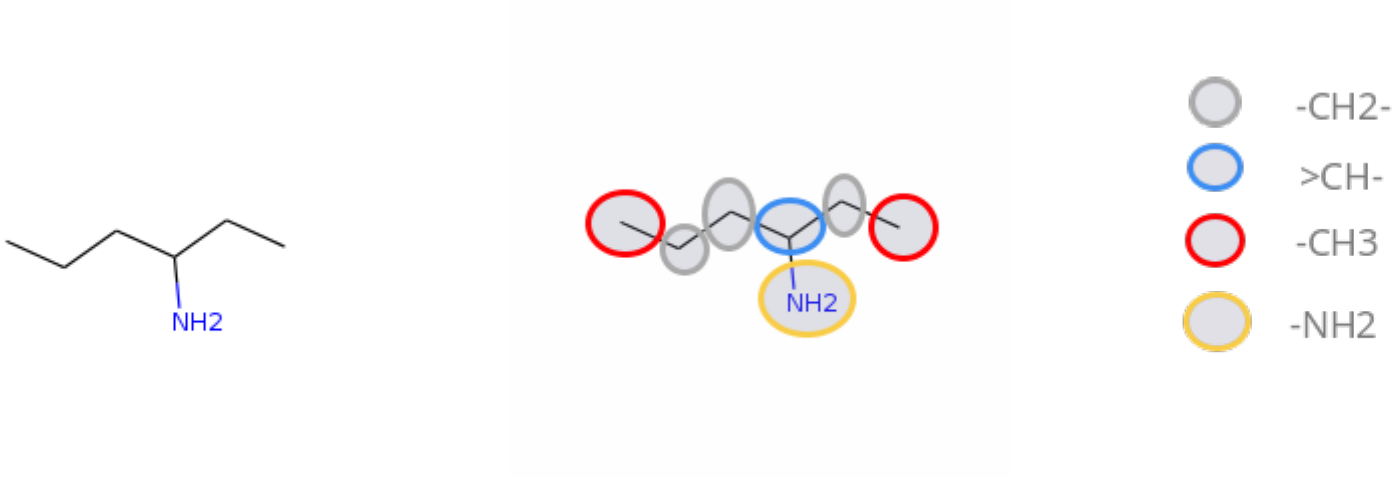
A PhD thesis in 20 minutes

- Create a group contribution method.
- Use it.
- Discover some issues.
- Go to see what the best are doing.
- The future.

A Group Contribution Method

Is Lego for Chemical Engineers

Take a molecule, cut it into pieces, you get the groups:



Then you postulate: $p = A_p + \sum_i n_i \times G_{p_i}$

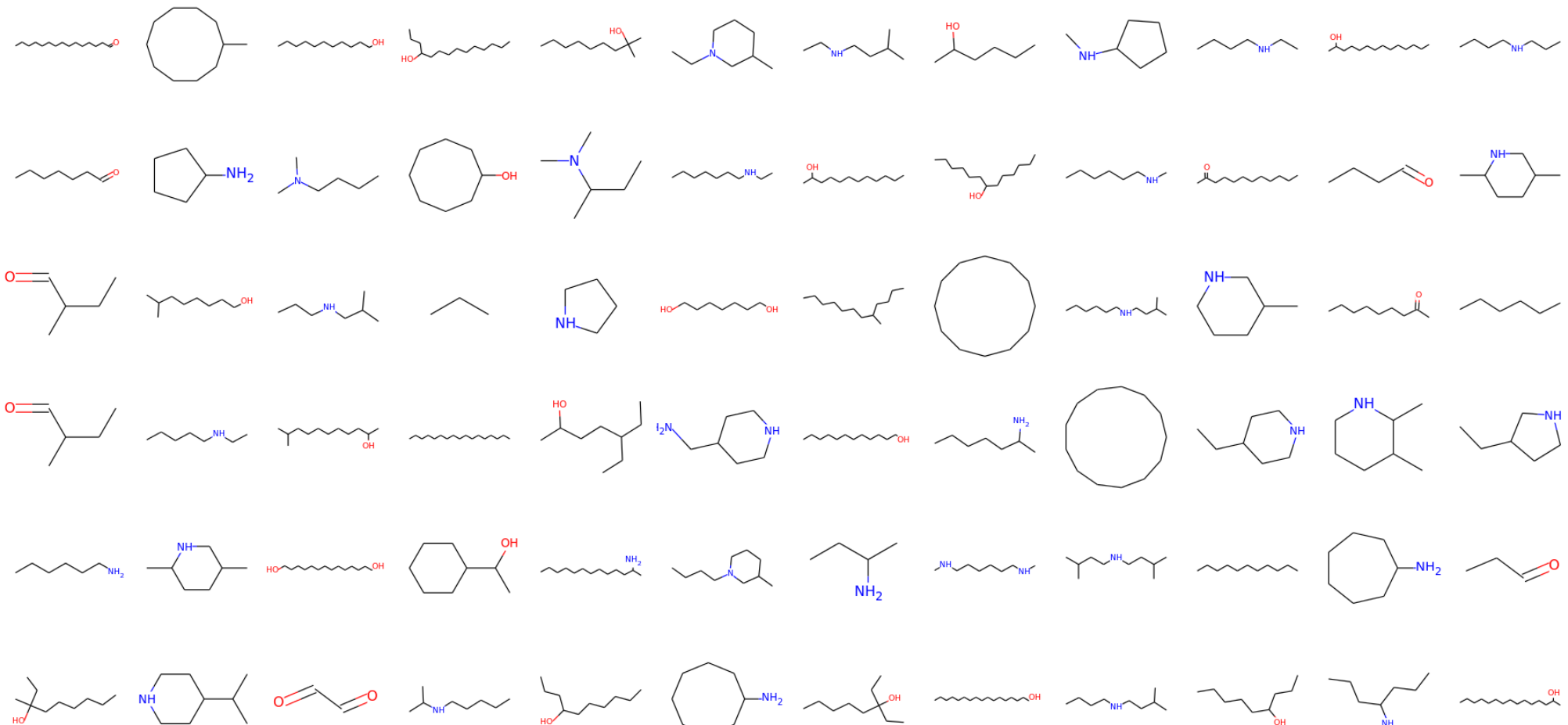
More complex equations and descriptors (2nd order, group/group interactions) are possible.

Céondo

A New Group Contribution Method

For Amines, Alcohol, Acids (Groups)
Boiling Point (Property)

Based on Joback's groups and a linear model for T_{boil}



A New Group Contribution Method

The groups for Amines, Alcohol, Acids

-CH3 -CH2- >CH- >C< =CH =C<

-CH2- (ring) >CH- (ring) >C< (ring) -OH (alcohol)

-O- (nonring) >C=O (nonring) O=CH- (aldehyde)

-NH2 >NH >NH (ring) >N-

We have 17 groups (A subset of Joback's 41 groups).

A New Group Contribution Method

The equation for the boiling point

We **postulate** that: $T_{boil} = A + \sum_i n_i \times G_i$

The group contribution method is defined.

We need to find the 17 G_i and the constant A.

I told you, this is simple and empirical.

Working extremely well for a large number of properties.

Céondo

Regression of the Parameters

First we need Experimental Data

Collected around **2000** data points from the literature.

Cleaned them because *0K* is not a good boiling point.

Discarded data points based on statistical analysis (standard deviation).

Associated for each experimental data point the group decomposition of the molecule.

Regression of the Parameters

Build the regression problem

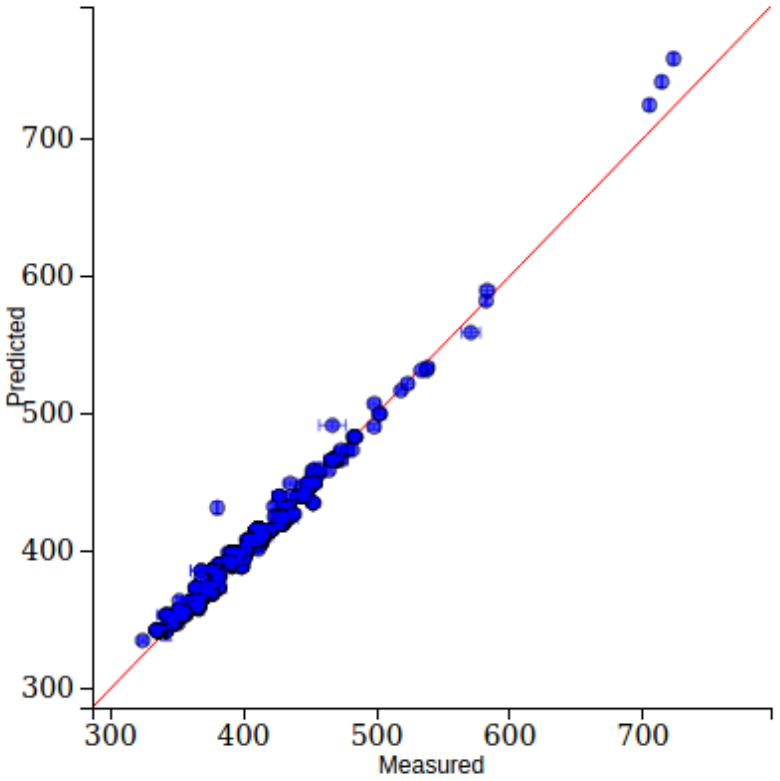
Constant	-CH2-	-CH3-	>CH-	...	Tboil (K)
1	1	2	**0**	...	365.12
1	5	2	**0**	...	289.63
1	4	1	2	...	325.89
1	10	3	2	...	402.19
1	1	2	**0**	...	367.12
...	...	...	...	...	...

A table with 1315 lines, because **we removed *bad* data**.

Céondo

Regression of the Parameters

This is extremely fast



Group	Contribution	Error	Significance	Data Points	Original
Constant	237.85	5.9648	★★★	1315	198.0
-CH3	24.14	2.9841	★★★	1168	23.58
-CH2-	16.85	0.0821	★★★	1079	22.88
>CH-	0.06	3.0002		301	21.74
>C<	-19.06	6.0916		15	18.25
=CH2	446705.39	0.0000		2	18.18
=CH-	24.96	0.0000		0	24.96
=C<	-446681.73	3.1549		2	24.14
-CH2- (ring)	19.30	0.9998	★★★	156	27.15
>CH- (ring)	1.73	2.0009		16	21.78
>C< (ring)	2.16	5.9893		3	21.32
-OH (alcohol)	78.85	2.9997	★★★	496	92.88
-O- (nonring)	18.81	3.1339	★★★	2	22.42

Céondo

Testing the New Method

Look! It Works!

Molecule	Measured T_{boil} (K)	Predicted T_{boil} (K)
2-Hexanol	410.65 ± 4.00	415.60 ± 2.02
1-Butanamine, N,N-dimethyl-	366.65 ± 3.00	367.33 ± 4.04
2-Dodecanol	518.15 ± 3.00	516.69 ± 2.20
Butanal	347.94 ± 0.30	351.92 ± 2.15
Isobutyl-propyl-amine	397.15 ± 2.00	397.74 ± 2.45

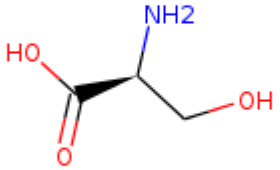
10 years ago, the same without uncertainty. Now, we have it!

But using the regression dataset to validate the model...

Céondo

The Method in Real Life

Serine, Alanine, the Amino Acids



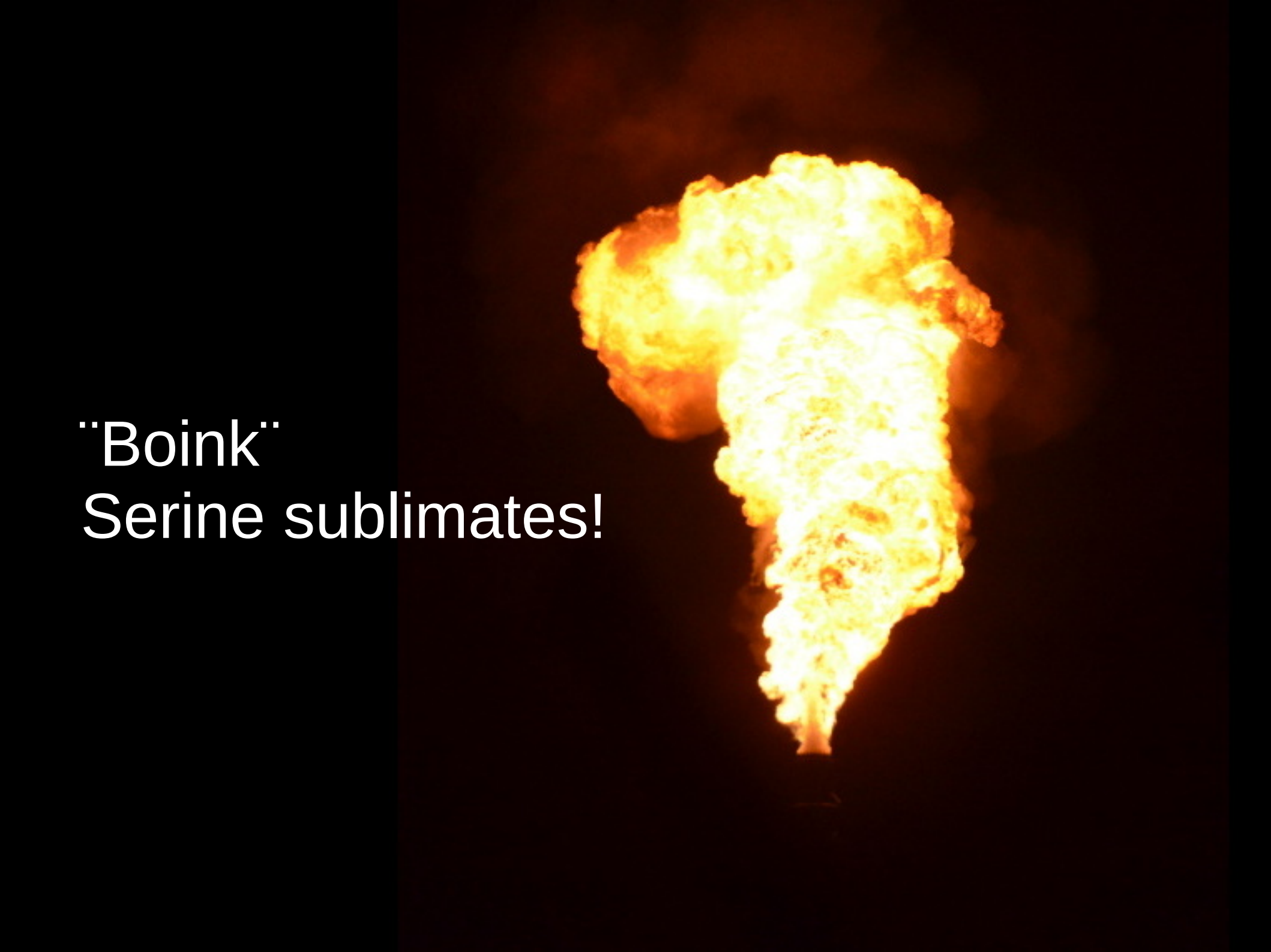
$$T_{boil} = 523.01 \pm 2.77K \qquad T_{boil} = 451.46 \pm 2.50K$$

We can predict for new molecules and the it is good!

The uncertainty information **creates trust**.

Céondo

¨Boink¨
Serine sublimates!



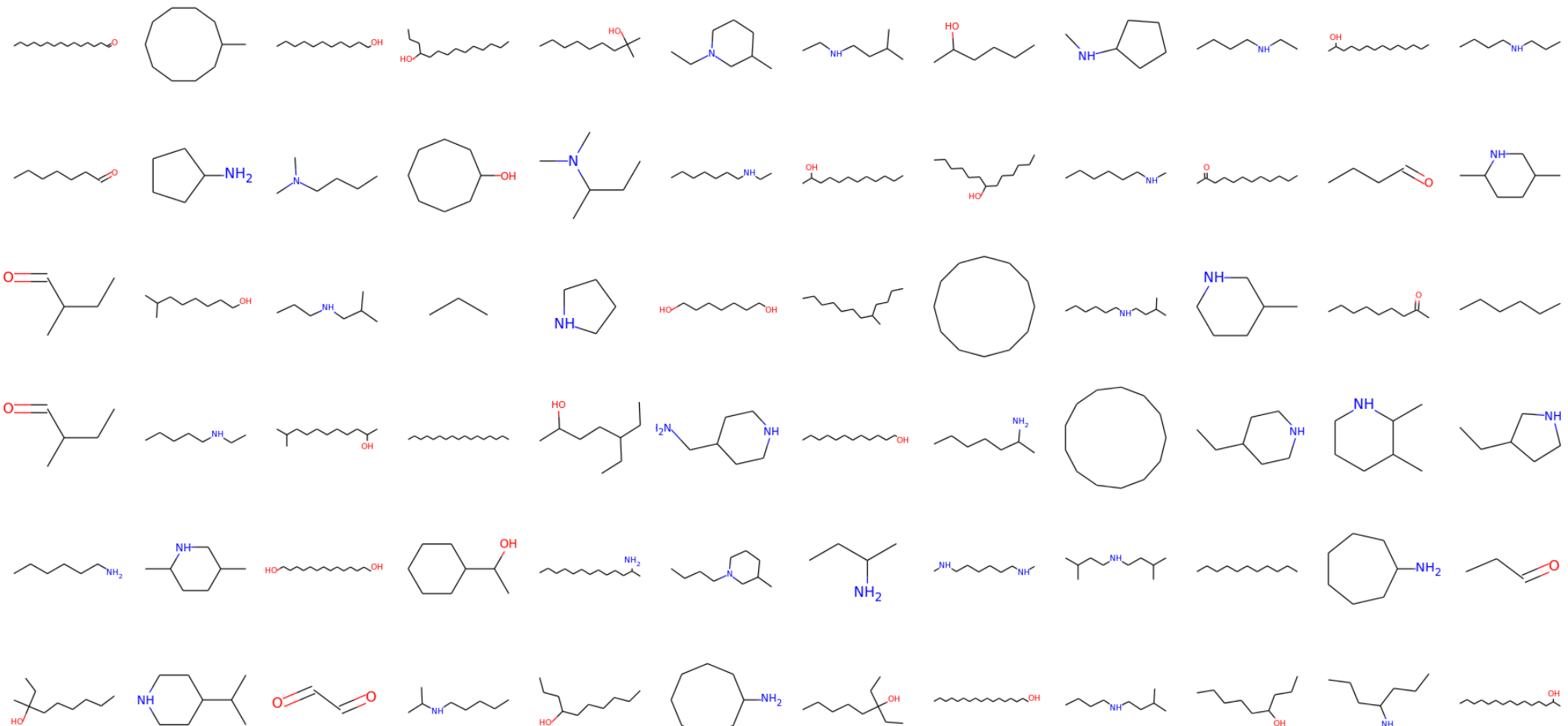
“There are three kind of lies:
Lies, damned lies and statistics”

A politician
Reported by Mark Twain

Céondo

It went ¨Boink¨, Why?

Even so 1315 data points for 18 parameters. We never had the O=CH- HO- groups together with the -NH2 group.



Where It Goes ¨Boink¨

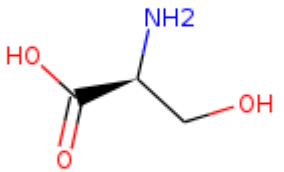
Group/Group Interactions

For this small model: (17 x (17-1))/2 + 17 = 153

153 group/group interactions to validate.

End result is that it is very easy to use your model outside of its domain of application, even if the stats are good!

And it is not a really complex molecule!



Céondo

Where It Goes ¨Boink¨

But the Prediction Interval is Good!

T_{boil}=451.46 ±2.50K

Confidence interval is based on the **degree of freedom**.

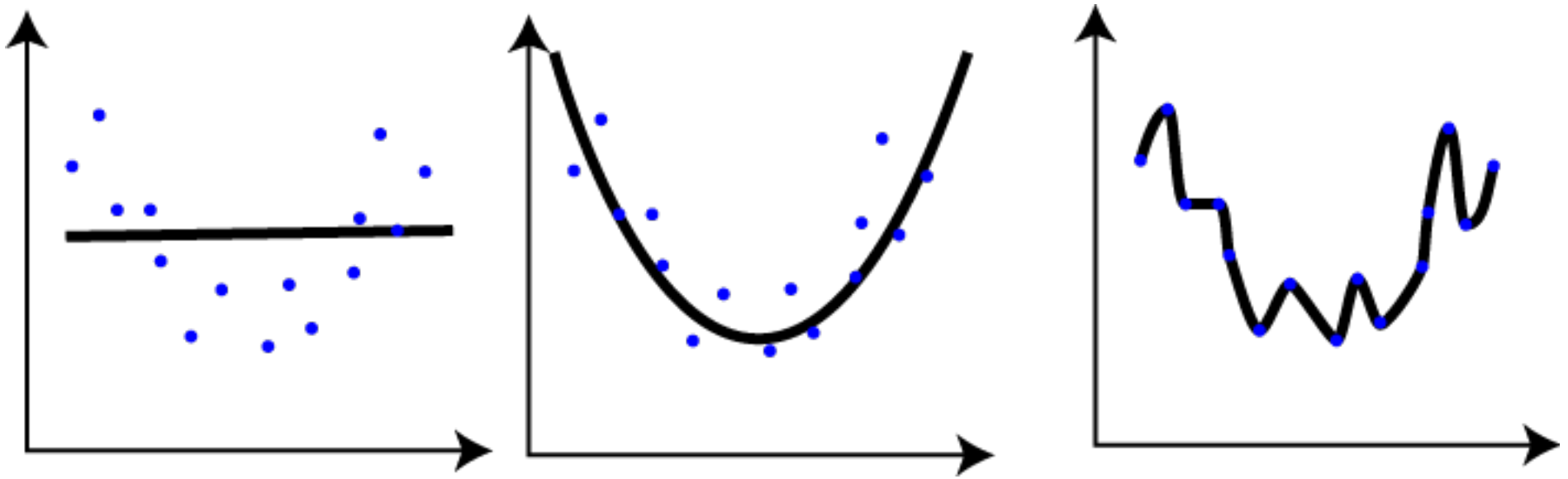
Group	Contribution	Error	Significance	Data Points	Original
Constant	237.85	5.9648	★★★	1315	198.0
-CH3	24.14	2.9841	★★★	1168	23.58
-CH2-	16.85	0.0821	★★★	1079	22.88
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
-OH (alcohol)	78.85	2.9997	★★★	496	92.88
-O- (nonring)	18.81	3.1339	★★★	2	22.42
>C=O (nonring)	54.07	0.4509	★★★	294	76.75

Céondo

Where It Goes ¨Boink¨

Over fitting

Over fitting because of the sparse problem, it is not caught by the statistics.



shapeofdata.wordpress.com

Céondo

Some Points to Look At

- How the regression dataset is built.
- Over fitting.
- Group/Group interactions.
- Validation/Statistics

The Work of the Best

- Picking two fairly new publications (not from the 70's), known to be good publications (2009, 2013).
- DISCLAIMER: In the following slides a lot will be about the non provided information. I am not judging the quality of the work! Just the trust factor.
- Please take it with humour... especially if you recognize your work!

The Dataset

The Unknown Man but we Trust Them!

Liquid viscosity data were taken from the Dortmund Data Bank (DDB) [6]. The DDB contains approximately 103,000 viscosity data points from 2630 references and approximately 2400 components,

of the adjusted value to estimate the liquid viscosity using the method proposed in this work, a relative absolute deviation of 15.3% in viscosity is obtained for the 813 components or 12,139 data points

From 100,000 points to 12,000 points. Data removed because not complete, questionable, not fitting nicely.

This is for viscosity, this requires a point at 2 different temperatures.

$$\ln\left(\frac{\eta}{1.3cP}\right) = -dBv\left(\frac{T - T_v}{T - (T_v/16)}\right)$$

Céondo

Over Fitting

We can nearly only trust them...

Group ID	Group contribution, dBv_i ($\times 10^3$)	Number of components	Absolute mean deviation
1	13.9133	520	0.2
2	11.7002	70	0.2
3	−11.0660	46	0.2
4	2.1727	344	0.2
5	4.5878	90	0.2
6	37.0296	22	0.3
7	21.3473	331	0.2
8	5.9452	28	0.4

Group ID	Group contribution, dBv_i ($\times 10^3$)	Number of components	Absolute mean deviation
76	16.3525	4	0.1
77	−2.6553	6	0.0
78	−61.2368	6	0.1
79	−7.5067	2	0.0
80	4.1408	8	0.1
81	−46.5613	2	0.0
82	122.6902	3	0.0
86	−70.9713	1	
88	29.0985	1	

Céondo

Over Fitting

We trust them

CHO–O	−231.63	5.00	13.92[b]	0.72	−231.63	5.00
CO–C,O	−238.01	1.26	10.39	0.97	−238.01	1.00
CO–2O	−299.61	0.29	8.31	0.54	−294.10	0.33

[a]Cd is a double-bonded C-atom; Cb is a C-atom in the aromatic ring; Ct is a triple-bonded C-atom; *ortho*(alkyl-alkyl), *meta*(alkyl-alkyl), and *para*(alkyl-alkyl) are interactions of small alkyl substituents (methyl and ethyl) in di-substituted benzenes.

[b]The calculated variance for this parameter was overestimated due to the existence of few or only low quality experimental data involving the group of interest.

Céondo

Group/Group Interactions

The new methods take it into account

Group ID	Interacting groups[a]	Group contribution, $T_{v,i}$
135	OH–OH	−1313.5690
136	OH–NH_2	−41.9608
137	OH–NH	−1868.6060
140	OH–EtherO	−643.4378
145	OH–CN	−345.7844
146	OH–AO	50.2582
[illegible]	[illegible]	[illegible]
194	Ketone–Ketone	1985.8270[b]
204	Alde–AO	161.7447
206	Nitro–Nitro	1839.2630
209	CN–AN6	718.1262
218	COOH–NH_2	(do not estimate)[c]

25 group/group interaction parameters. But with more than 100 groups, we have thousands of possible interactions.

Céondo

Validation/Statistics

Not enough data

No splitting into regression dataset and test dataset.

One can do repeated sampling by excluding compounds and looking at the regressed parameters and predicted properties.

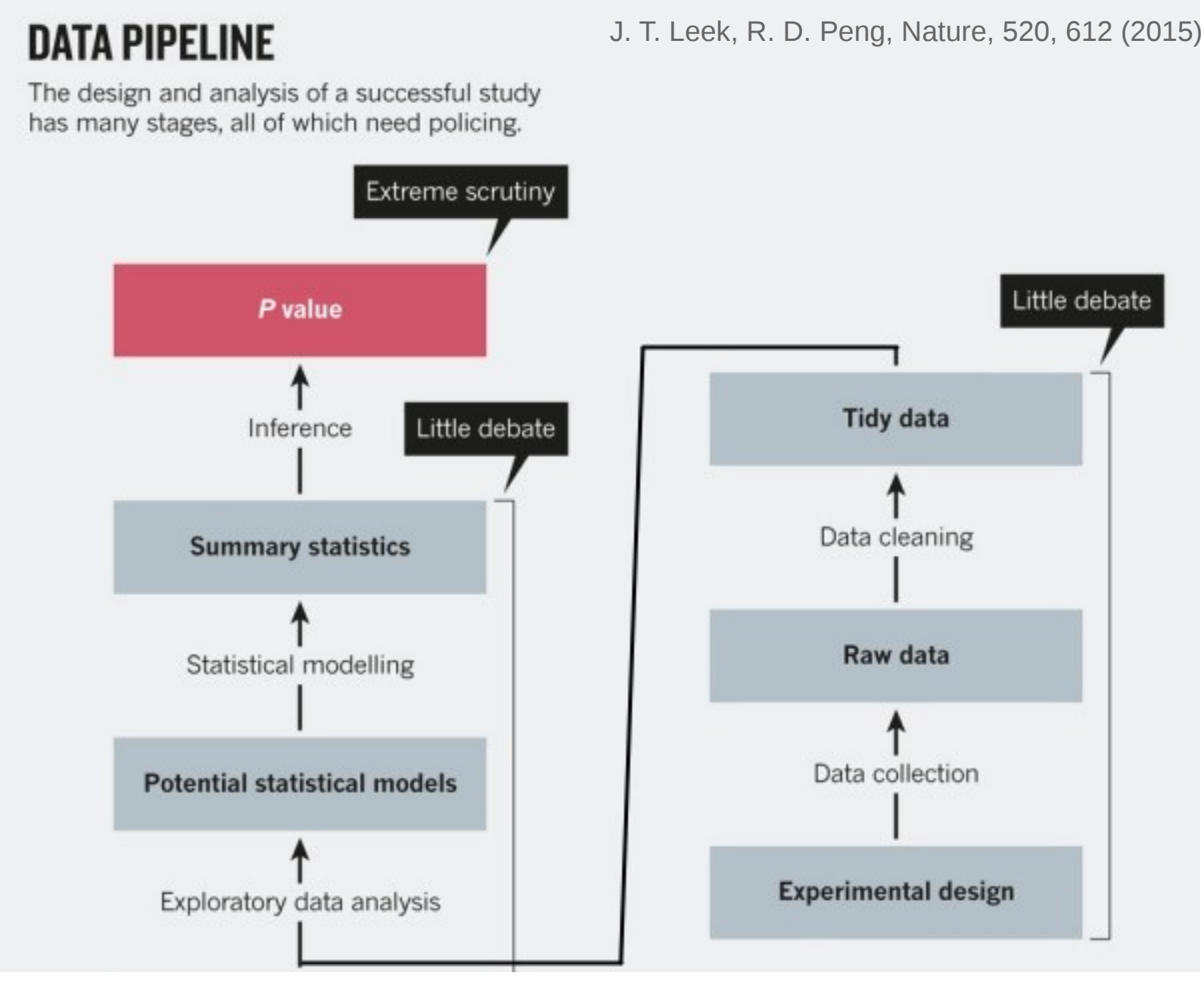
If you over fitted a significant number of group contributions:

Does the relative mean deviation has a sense?

My advice: look at the dataset used to regress when you predict. Many tools show the experimental values.

Can You Trust Your Models?

J. T. Leek, R. D. Peng, Nature, 520, 612 (2015)



Céondo

The Work of the Best
The New Friends are

- "New Group-Contribution Approach to Thermochemical Properties of Organic Compounds: Hydrocarbons and Oxygen-Containing Compounds", Univ. Rostock, **NIST**, J. Phys. Chem. Ref. Data, Vol. 42, No. 3, 2013
- "Estimation of pure component properties. Part 4: Estimation of the saturated liquid viscosity of non-electrolyte organic compounds via group contributions and group interactions", **University Oldenburg**, SASOL, Y. Nannoolal et al. / Fluid Phase Equilibria 281 (2009) 97–119

New Statistical Tools

But not yet general

- Sequential evaluation of the covariance matrix.
  - Uncertainty in sequential (residual) regression.
- Introduction of the concept of “effective number of data points”.

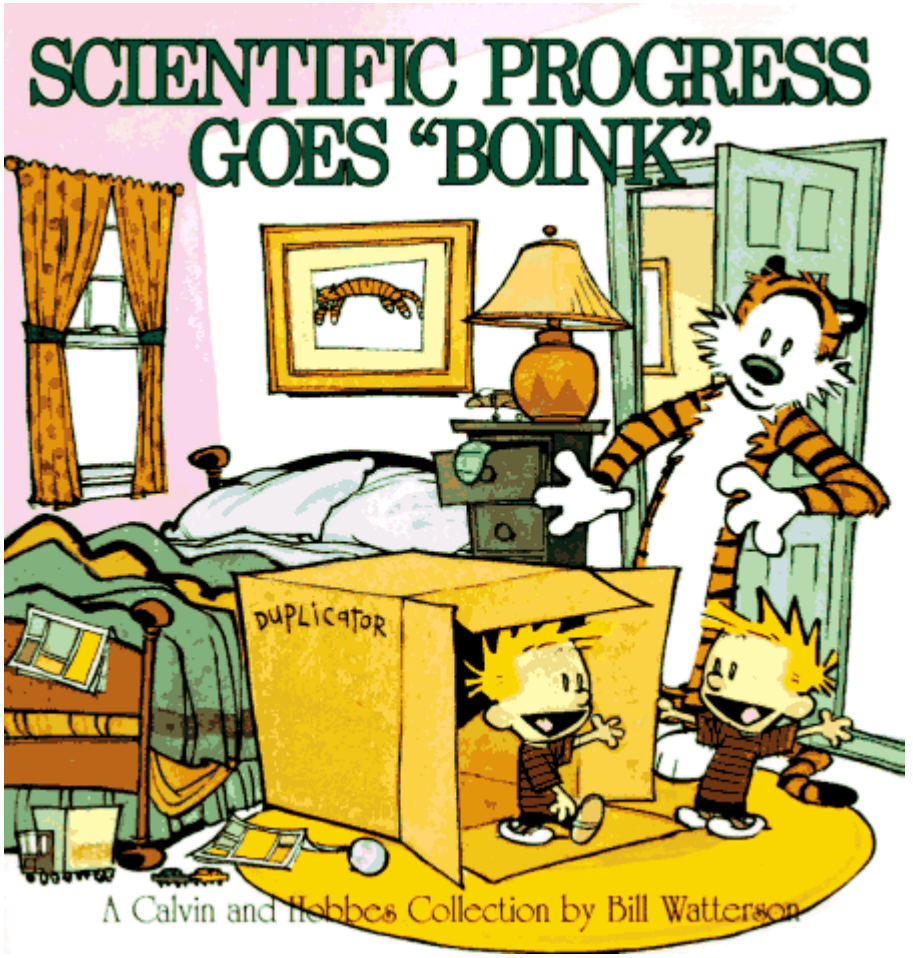
S. P. Verevkin et al. J. Phys. Chem. Ref. Data, Vol. 42, No. 3, 2013

Céondo

Conclusion
And future

- The group contribution methods do work and are proved.
- New statistical tools are coming to bring us better understanding of the uncertainties to alleviate the current limitations. Take the current stats with caution.
- We are not going to know all the group/group interactions, so we need at least which molecules were in the regression dataset.
  - Note that this problem also affects UNIFAC.

Céondo



Thank you!

Céondo