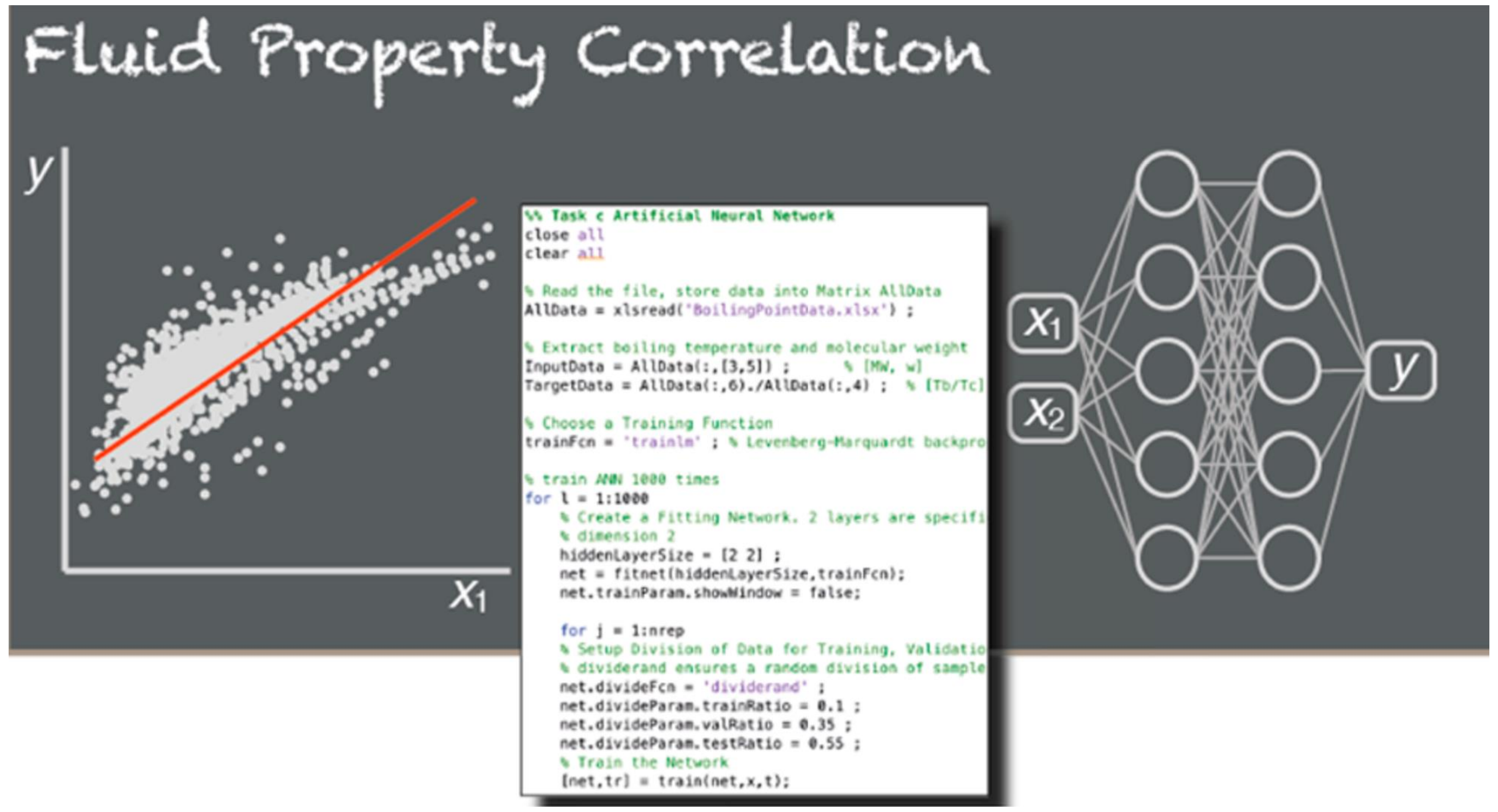
Imperial College London

Employing Machine Learning to Correlate Fluid Properties

Classroom Examples with MATLAB

Erich A. Müller

Department of Chemical Engineering
Imperial College London U.K.



Motivation & Background

- Target audience are 1st year Chemical Engineering undergraduates at Imperial College London. Students have a very mixed-ability background: some have had some basic programming in high school, but many are computer-illiterate.

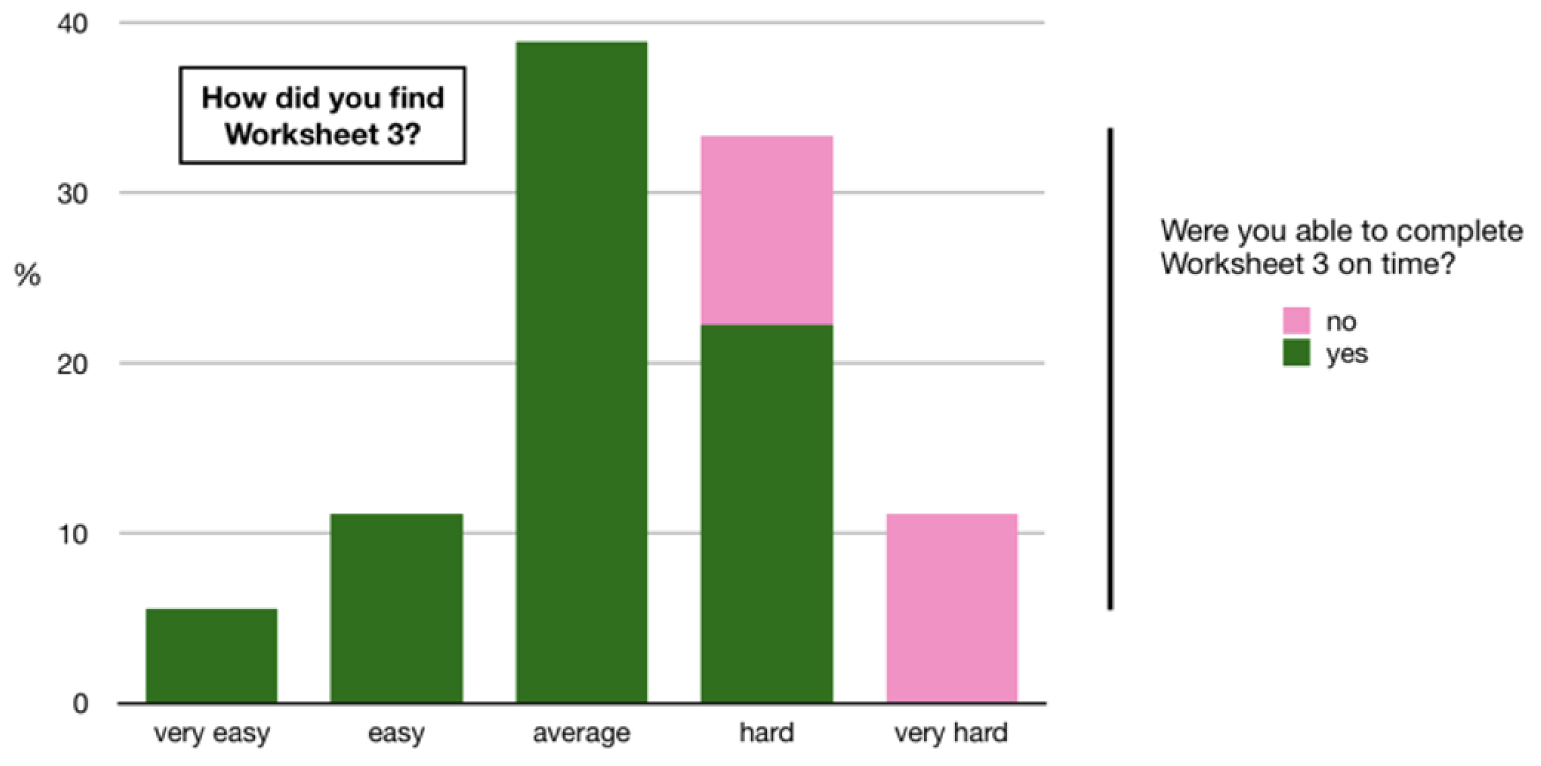


Figure 1 : Poll from a student group of the MATLAB cohort 20-21 performed mid-course.
Distribution suggests a mixed ability course with a large variation in the perceived difficulty.

- Students are enrolled in a mandatory 6-week hands-on “Introduction to programming and MATLAB” course. The course is practical pass/fail module.

- The course covers the basics of programming in the first three weeks and then starts focusing on skillsets needed for the rest of the Chemical Engineering curriculum including plotting, solution of linear sets of equations, ordinary differential equations, etc.

Motivation & Background

- There was a pressing need (and request from students) to be introduced to some (basic) notions of Machine Learning.
- This presentation showcases an example application, provided in the last week of the course aimed at exposing the students to the Machine learning tools that might be useful in further years.

Normal Boiling point

The normal boiling point is defined as the saturation (boiling) temperature of a liquid at 1 atm of pressure.

A related quantity, the standard boiling point, is defined by IUPAC as the saturation temperature of a fluid at a pressure of 1 bar.

It is a key quantity in the design of chemical processes (e.g. distillation towers, solvent extraction processes, etc.)

Most information is collated through empirical correlations based on the mathematical fitting of experimental data

The challenge:

Given a (very large) table of physical properties for many organic substances, can you produce an engineering-quality correlation for the boiling point?

boiling point data — Saved to my Mac

	A	B	C	D	E	F
1		name	molweight	critical temperature (K)	acentric factor	boiling point (K)
2	1	(+)-a-pinene	136.23704	647	0.341	428.65
3	2	(+)-camphene	136.23704	638	0.296	432.65
4	3	(-)-a-pinene	136.23704	647	0.341	429.35
5	4	(-)-b-citronellol	156.2682	656.59	0.612	498.65
6	5	(-)-camphene	136.23704	638	0.296	439.95
7	6	(1,1-dimethylbutyl)benzene	162.27492	697.15	0.437	478.65
8	7	(1-butylhexadecyl)benzene	358.65124	851.65	0.759	693.15
9	8	(1-ethyl-1-methylpropyl)benzene	162.27492	697.15	0.437	478.15
10	9	(1-ethylbutyl)benzene	162.27492	697.15	0.437	482.15
11	10	(1-ethyloctadecyl)benzene	358.65124	851.65	0.759	693.15
12	11	(1-hexylheptyl)benzene	260.46308	783	0.844	614.43
13	12	(1-methylenepropyl)benzene	132.20528	666	0.354	455.15
14	13	(1-methylheptyl)benzene	190.32868	712.89	0.565	523.94
15	14	(1-methylnonadecyl)benzene	358.65124	851.65	0.759	693.15
16	15	(1-methylnonyl)benzene	218.38244	738.32	0.659	562.65
17	16	(1-methylpentyl)benzene	162.27492	697.15	0.437	481.15
18	17	(1-octyldodecyl)benzene	358.65124	851.65	0.759	693.15
19	18	(1-propylheptadecyl)benzene	358.65124	851.65	0.759	693.15
20	19	(1-thiapropyl)-benzene	138.2334	712.96	0.372	478.15
21	20	(1R)-(-)-menthyl chloride	174.71356	682.53	0.516	496.15
22	21	(1R,2S,5R)-(-)-menthol	156.2682	658	0.78	489.55
23	22	(1S)-(-)-b-pinene	136.23704	647	0.341	439.15

- **Excel sheet with over 5000 entries**
- Each row corresponds to an individual molecule.
- For each component there are a wealth of data points, including name, CAS number, boiling temperature, molecular weight, etc.

Live Editor – /Users/erichmuller/Dropbox/My Mac (Erichs-MacBook-Retina.local)/Desktop/Mathworks -webinar/files/Machine_Learning_Fluid_Prop...

Q2.m | Temperature.m | Machine_Learning_Fluid_Properties.mlx | +

Importing data from Excel spreadsheet using an autogenerated function by Import Tool:

```
boilingpointdata = importfile("boiling point data.xlsx", "Sheet1", [2, 6032])
```

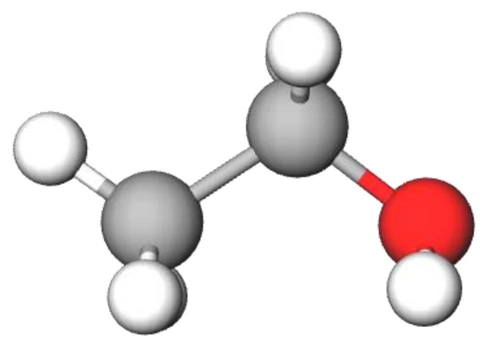
boilingpointdata = 6031×6 table

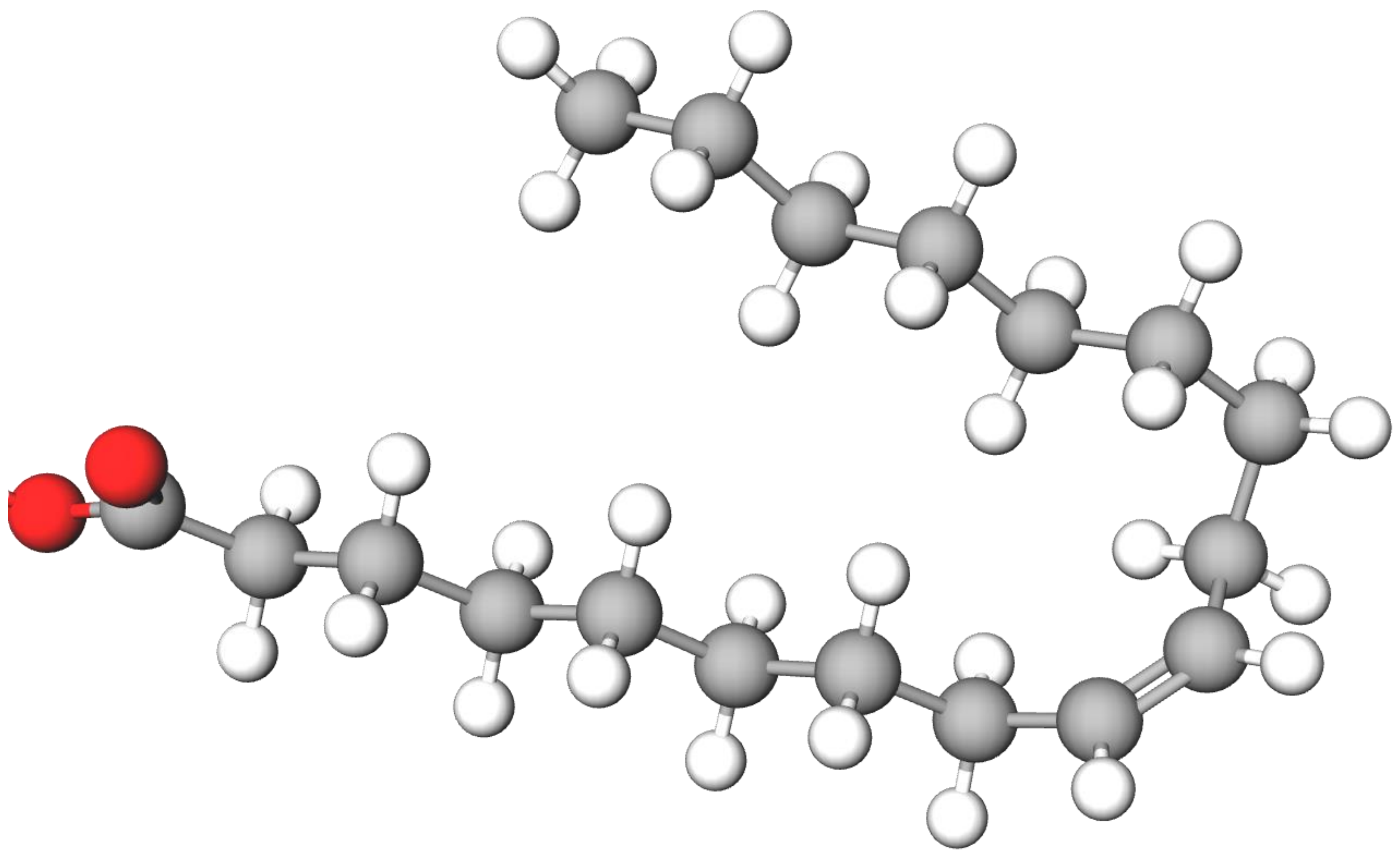
	VarName1	name	molweight	criticaltemperatureK	acentricfactor
1	1	"(+)-a-pinene"	136.2370	647.0000	0.3410
2	2	"(+)-camph…	136.2370	638.0000	0.2960
3	3	"(-)-a-pinene"	136.2370	647.0000	0.3410
4	4	"(-)-b-citron…	156.2682	656.5900	0.6120
5	5	"(-)-camphe…	136.2370	638.0000	0.2960
6	6	"(1,1-dimet…	162.2749	697.1500	0.4370
7	7	"(1-butylhe…	358.6512	851.6500	0.7590
8	8	"(1-ethyl-1-…	162.2749	697.1500	0.4370
9	9	"(1-ethylbut…	162.2749	697.1500	0.4370

```
Mw = boilingpointdata.molweight ;
Tc = boilingpointdata.criticaltemperatureK ;
w = boilingpointdata.acentricfactor ;
Tb = boilingpointdata.boilingpointK ;
```

General strategy of the project

A first empirical observation is that the boiling point is proportional to the molecular weight





<u>Ethanol</u>

MW = 46.07

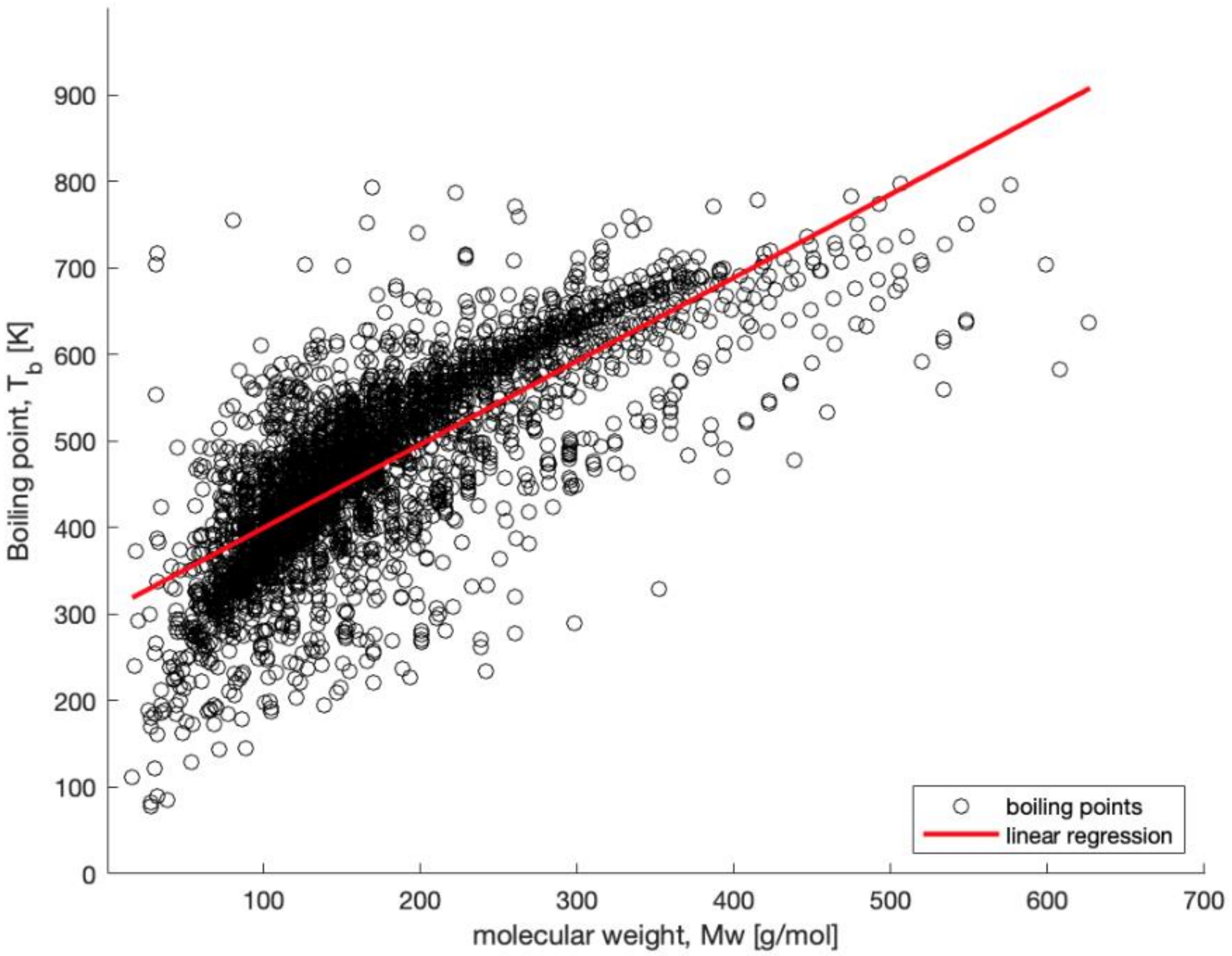
T_b = 78.4 ºC

<u>Oleic acid</u>

MW = 282.47

T_b = 360 ºC

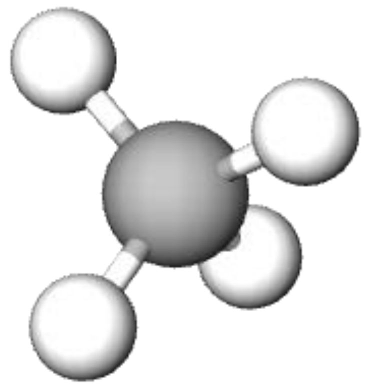
TASK 1 : Linear fit of the boiling point to the molecular weight

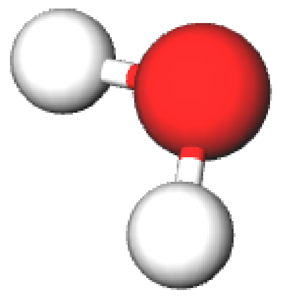


- The correlation is rather poor (R^2 = 0.76)
- There is *some* trend, but obviously there are other parameters which are also of importance.
- Other fits (logarithmic, quadratic, etc.) will clearly not be successful.

```
Clin = polyfit(Mw, Tb, 1) ;
Mwplot = linspace(floor(min(Mw/10)*10), ceil(max(Mw/10)*10), 50) ;
% plot
figure(1)
hold on
plot(Mw, Tb, 'ok')
plot(Mwplot, polyval(Clin, Mwplot), '-r', 'linewidth', 2)
```

Second ansatz : the boiling point has some relation with the acentric factor





Methane

MW = 16.04

T_b = -161.5 ºC

ω = 0

Water

MW = 18.01

T_b = 100 ºC

ω = 0.344

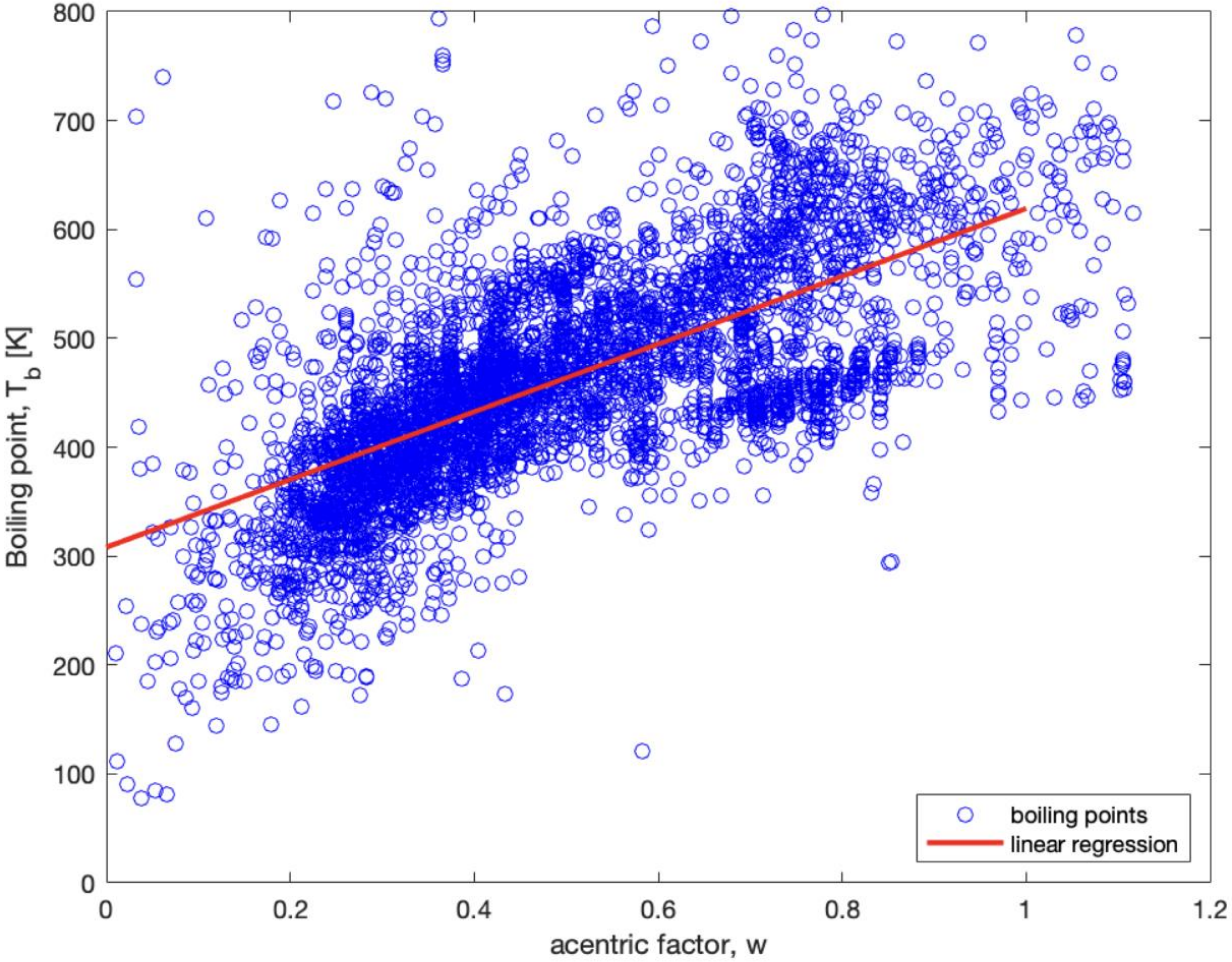
- The acentric factor is an empirical number

$$\omega = -\log_{10}(p_r^{sat}) - 1 , \quad \text{at} \quad T_r = 0.7$$

- Its value is close to zero for noble gases and increases as the molecule becomes non-spherical and/or polar.

- It is commonly tabulated (along critical properties)

Just out of curiosity:
Linear fit of the boiling point to the acentric factor



TASK 2 : Multivariate correlation

- Assume that the boiling point is a linear function of **both** the molecular weight and the acentric factor

$$T_b = \theta_0 + \theta_1 \omega + \theta_2 MW$$

- Scale the boiling temperature with the appropriate critical temperature. This scales the T_b values from 0.7 to 1

$$y = T_b / T_c$$

- Solve the problem by matrix manipulation

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{100} \end{pmatrix} = \begin{pmatrix} 1 & \omega_1 & MW_1 \\ 1 & \omega_2 & MW_2 \\ \vdots & \vdots & \vdots \\ 1 & \omega_{100} & MW_{100} \end{pmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix} = \mathbf{X}\theta$$

$$\theta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

Classical procedure

Determine a priori the mathematical (plausible) relationship

Employ physical insights

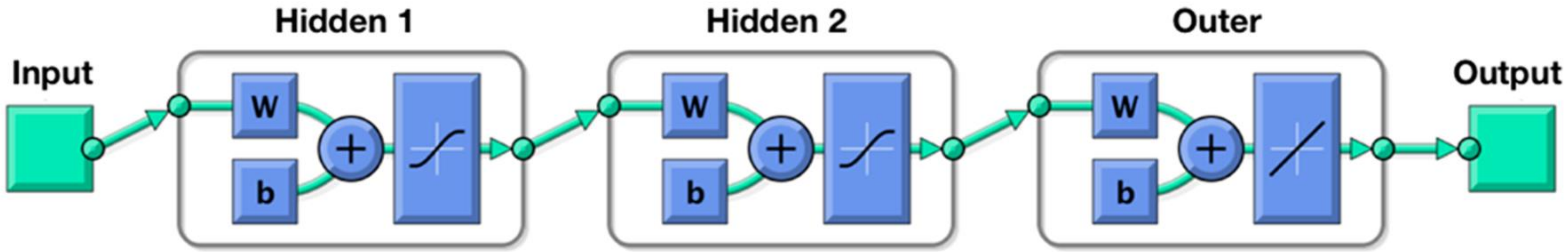
Solve the minimum likelihood problem

```
% sample 100 observations
[~, idx] = datasample(Tb,100, 'replace', false) ;
Tbr_train = Tb(idx)./Tc(idx) ;
Mw_train = Mw(idx) ;
w_train = w(idx) ;
% create matrixes for linear regression
X = [ones(size(Tbr_train)), Mw_train, w_train] ;
y = Tbr_train ;
A = X'*X ;
b = X'*y ;
theta = A\b ;  % estimated parameters (solution)

TrainedModel = @(x) [ones(size(x,1),1), x]*theta ;
Tbr_pred = TrainedModel([Mw, w]) ;

% calculate correlation coefficient
[R,P] = corrcoef(X*theta,Tbr_train) ;
fprintf('The correlation coefficient is %.2f \n', R(2))
```

```
The correlation coefficient is 0.81
```

TASK 3 : Employ an Artificial Neural Network (ANN)

- No assumption is made with respect to the mathematical structure of the “correlation”
- The *features* are ω and MW (by simple inference)
- Solve the problem *training* with 100 randomly selected data points
- MATLAB has a built-in ANN encoder

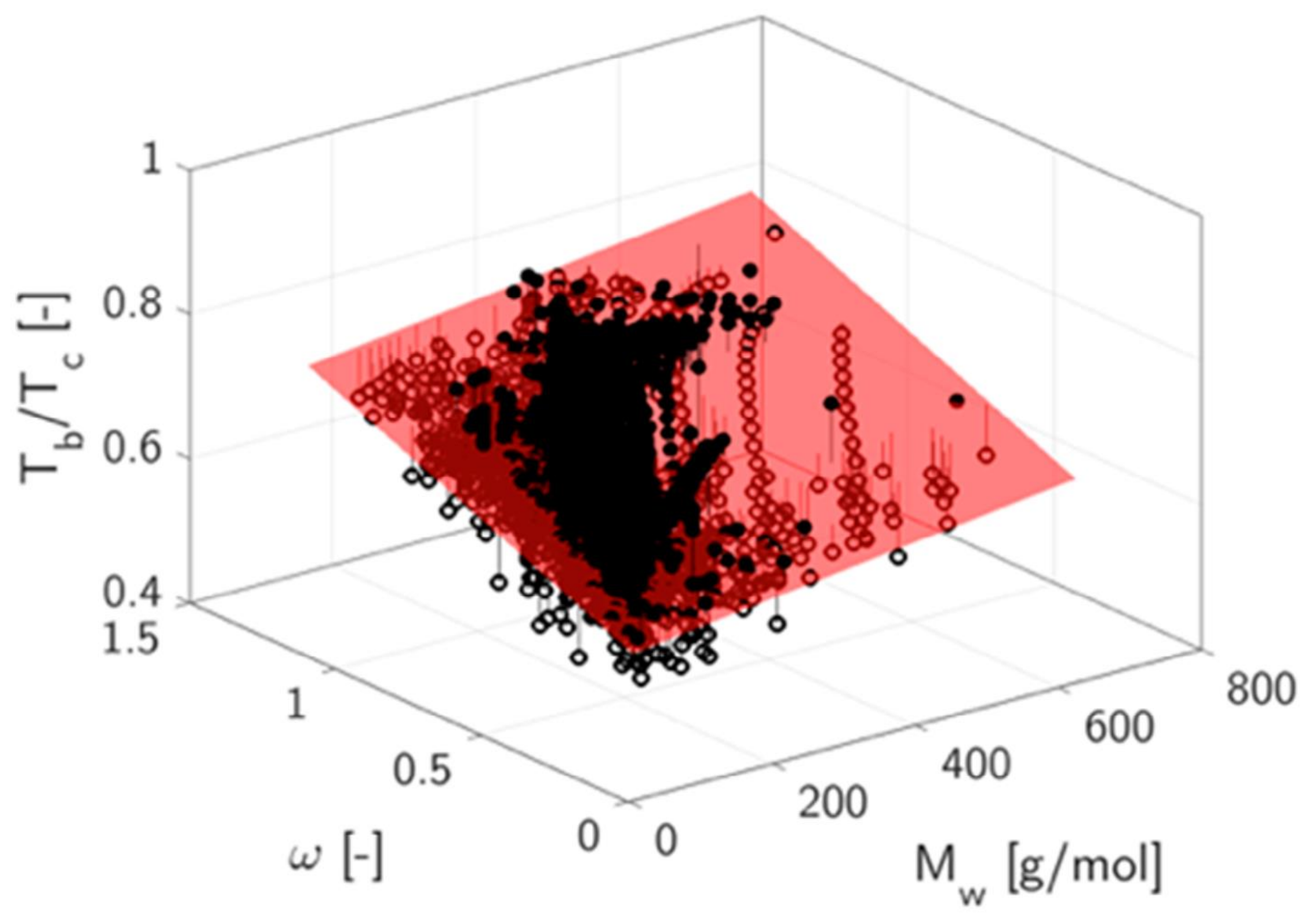


- The ANN is composed of two hidden layers with a tan−sigmoid transfer function and an outer layer with a linear transfer function (gray boxes).
- The weights (W) and biases (b) are optimized using the Levenberg−Marquardt algorithm.
- Green boxes represent the algorithm input (left) and output (right)

```
% Solve an Input-Output Fitting problem with a Neural Network
% Input and target data.
x = InputData';
t = TargetData';

% Choose a Training Function
trainFcn = 'trainlm' ; % Levenberg-Marquardt backpropagation.

% Create a Fitting Network. 2 layers are specified here, each of dimension
hiddenLayerSize = [2 2] ;
net = fitnet(hiddenLayerSize,trainFcn);

% Setup Division of Data for Training, Validation, Testing
% dividerand ensures a random division of samples
net.divideFcn = 'dividerand' ;
net.divideParam.trainRatio =  0.2  ;
net.divideParam.valRatio =  0.4
net.divideParam.testRatio =  0.3  ;

% Train the Network
[net,tr] = train(net,x,t);

% Estimate the output y based on the trained network (net) and the inputs
y = net(x) ;

% calculate correlation coefficient
[R,P] = corrcoef(t,y) ;
fprintf('The correlation coefficient is %.2f \n', R(2))
```

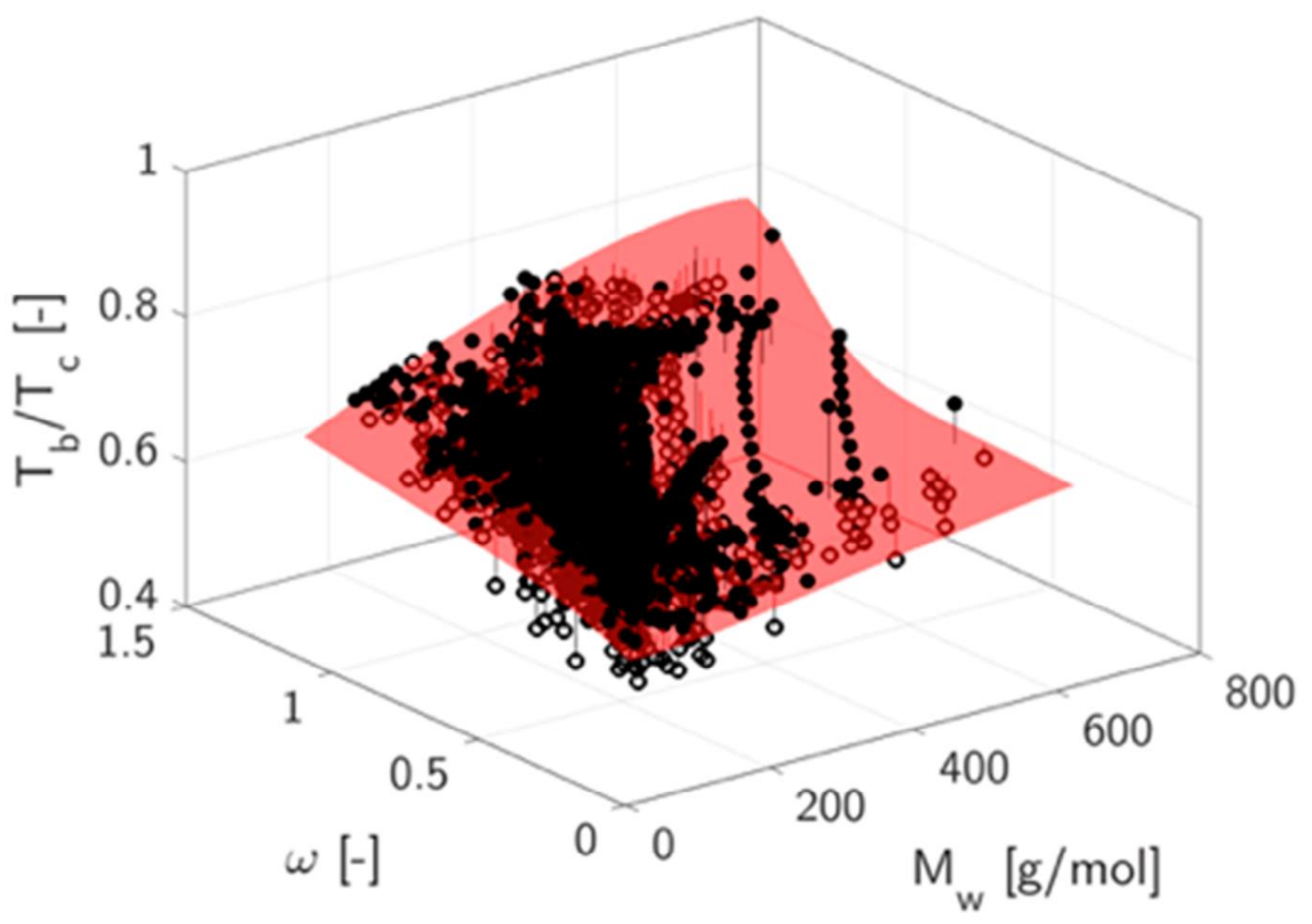
```
The correlation coefficient is 0.88
```

Multivariate



R^2 = 0.84
AAD = 2.7 %

ANN



R^2 = 0.89
AAD = 2.2 %

Quality of fit is comparable to available engineering correlations in the open literature. {

Conclusions

- The exercise has been extremely well received by the students, who come back asking for more material to expand their understanding of the topic.

- The example can be expanded and improved easily, although the ML correlation is already quite good. Other examples in physical property prediction/correlation come to mind.

- Machine learning has crept up in a large number of the final year research projects and has proven to be an extremely popular topic (and a skill requested by employers).

- A final year elective on Machine learning in Process Engineering is now being developed.

Please direct the questions to

- Lisa Joss lisa.lj.joss@gmail.com
- Erich A. Müller e.muller@imperial.ac.uk

More details

L. Joss and E. A. Müller, “Machine Learning for Fluid Property Correlations: Classroom Examples with MATLAB,” *J. Chem. Educ.*, **96**(4), 697–703, 2019.

http://www.molecularsystemsengineering.org

MOLECULAR SYSTEMS ENGINEERING