

## Task 2: Shaking the Foundations — What Did You Actually Model?

**Weeks:** 4-5

**Title:** *“Is Aqueous Solubility Even a Thing?”*

### Objectives:

- Understand the complexity of the experimental procedure of a given dataset
- Understand how this should guide your thinking about building your model

### Deliverables:

- A 2-page report (font 12) written in a scientific way with proper citations from the literature explaining the meaning and experimental procedures of measuring aqueous solubility
  - **Push your report to the GitHub repo by Monday, 24<sup>th</sup> at 10 PM.**
- Update your Jupyter notebook with a detailed, but concise, introduction to Aqueous solubility, and how this will guide you when building your ML pipeline
  - **Do not forget to explain why the dataset was poorly curated**
  - **This should be unified by the whole group**
  - **Push your update to the GitHub repo by 6 PM the Wednesday before next session**

### Description

Your team and the CEO were quite surprised by the performance of your model on the challenge dataset. This was the only step between you and glory and between the CEO reducing costs. But how did the model fail this tragically??

After a while of dismay, one of you asks, but what exactly is aqueous solubility?

Are we sure that we gave the model what it needs to learn about this property?

You look at each other and the CEO, and you realize your inability to answer this question, for neither of you is a chemist, and neither of you have run such an experiment themselves!

You understand that you cannot move forward with your modelling, unless you talk to (or acquire the knowledge of) a chemist who knows the chemistry behind this property.

Since the next step is **exploration**, you understand that it works better when you can collect different opinions on the same problem. So, you decide to split yourself in pairs; each pair wanders on their own, looks at the literature, asks their chemist friends, and collects knowledge independently. Then, you come back to discuss what each pair understood and make your final understanding. Therefore,

- **Each pair of students will write a 2-page scientific report about their findings to understand what aqueous solubility is.**

## 2<sup>nd</sup> internal meeting

You meet after collecting your thoughts and words and start discussing your findings. You see that all of you understood the main things, but each one of you managed to find something that caught only their attention (because we are unique human beings who will always be different from each other in the way we think).

After discussion, you understand what is essential about this aqueous solubility property, and hopefully you find out that some of the analysis and preprocessing you used when training your models was not that suitable. You also managed to understand why your training data was badly curated. So,

- **You summarize your findings in one short report, and you suggest possible modifications to the ML pipeline based on your new understanding of chemistry.**
- **You add this report to the introduction section of your notebook to be a guideline for you and your CEO**

## Important notes for your deliverables

- You must site every statement in your report from a highly reputable journal/publisher
- You are not allowed to let language models write your report for you
  - If we suspect this, we will ask you to explain your work in details in person