# AL/ML TECHNOLOGIST CASE STUDY

# 

# Overview

We really appreciate your time and effort working on this home task as part of our interview process. It is meant for you to demonstrate your skills as well as get a taste of what we are after.

The task includes four parts and we hope you enjoy working on it. Please send over your **code** as well as your findings in a **2-­3 page document**, and also be ready to present your ideas and lead the discussion at our next meeting. With that in mind we encourage you to visualise the results and bring them to life with colour and graphs. You can use the software or statistical package of your choice for this analysis.

The file, available [HERE](https://drive.google.com/file/d/1j63oc3O6ABoX0axgrVvgeYeDhyvKkA1S/view?usp=sharing) (~50MB zipped), contains technology news articles from the Guardian newspaper, including their tags. The core of this challenge is to use the article body text to predict the document tags. While generally our work is not focused on NLP, we find this a fun dataset to experiment with.

We are more interested in the approach you take than the accuracy score. Given the time constraints we do not expect a perfect solution but we would like to see you address the challenge and demonstrate your ability. Please describe additional analysis you would suggest for further exploration.

*Please do not share this task with anyone or make it public in any way. Please also delete all the task materials, data and derivatives after your presentation. Your co­operation with this request is much appreciated.*

# Part 1: warm up

1. What are the most common topics in the dataset?
2. Which of these do you expect to be easiest to predict from the text?

# Part 2: full on

1. Which tags can be predicted best from the text?
2. Is this according to your hypothesis? If not, can you speculate why?

# Part 3: mental stretches

In this section we would like to understand your approach to the challenge, but it is not necessary to provide any code.

1. We would like to look into your understanding of deep neural networks. Consider a remote camera trap that has collected hundreds of hours of video. You’ve been asked to produce a list of timestamps and their associated animals.
   1. How would you approach this problem?
   2. Which tools, libraries and cloud services might you use?
   3. Where would you look for training data?
   4. How could you reduce your need for training data?
2. In this section we focus on using different types of hardware for ML models. What are the fundamental differences in design between a CPU and GPU? What is bandwidth and latency? What are the differences between model parallelism and data parallelism? In your view, what is the most promising hardware for the future of AI?

# Part 4: wizard

7. Changing chairs for a moment, which question would you add to this task? Answer your question!

We hope you enjoyed working on this task! Feel free to contact us at tiffiny.cazin@digicatapult.org.uk with any questions.

Your effort is much appreciated!