**Harry Langford**
hjel2@cam.ac.uk

# Machine Learning Supervision 1:

The ticks do not align with the exam questions – there is a lot more information in the exam questions and you have to get the knowledge from the supervisiosn. The course is quite disjoint.

Anwering the examquestion is a different story.

Read the recommended reading.

Validation and testing data – get the difference between it.

You are allowed to retest on the validation data – but on the testing data you are only ever allowed to use it once. You must never reuse test data again. Both are used for evaulation – but you are allowed to reuse the validation data but not test data. Test data must only ever be tested on once.

"What types of words have their sentiments flipped": Verbs and ajectives mainly.

Think about "how is the setup similar to what you have done before" and how is it different?

When choosing a metric to use you should use a metric which does not allow the system to cheat and get a high score without solving the problem properly. That is how you should analyse metrics.

Baseline approach to a problem: The simplest approach to a problem which you should compare the model to. A baseline approach is a "simple and tractible" approach. An approach that you can cod eup in 5 minutes. If by doing something simple you get a very high accuracy then maybe you shoul dnot bother doing anything more complicated. It is a simple and tractible approach you should compare your models with.

When doing (a small number of) binary features we should model the absence of features ie:

$$P(A \cap B \cap C') = P(A) \times P(B) \times P'(C)$$

With some features – like when the features are words – you should not model the absence since the probabilities are so low that the absence of a feature is meaningless.

Burstiness:
Words come very infrequently but when they do come they come a lot.

"stop words" are words which are grammatically necessary.