**Evaluation**

As previously stated, ETL took a considerable length of time. The majority of the time spent working on this dissertation, in fact.

Modern ETL systems have a great many convenience features – automated rollbacks, in particular – that would have simplified the task somewhat.

A great deal of time was spent running and re-running the same processes with small bugfixes. Even working across different machines was difficult with terabytes of data.

It was also a brand-new idea for me. Machine learning problems I’d tackled in the past had datasets fully prepared already, albeit with some pre-processing required. But ETL is much deeper. I found it very hard, in particular, to reach a stage when I could view my dataset as immutable.

As this dissertation progressed, a new repository, archiving all feeds, was established by OpenTrainTimes[[1]](#footnote-1). However, switching to the new mirror would have incurred too much work, both from an extraction, but particularly a transformation, perspective.

There is scope for comprised data integrity. With no archived CORPUS information, it is possible, though unlikely, that TIPLOCs and STANOXs in 2018 referred to different locations as they do in 2020.

1. https://networkrail.opendata.opentraintimes.com/ [↑](#footnote-ref-1)