**Goal ⇒** How should we evaluate the long tail disambiguation task(s)?

- **Questions for discussion (30 mins):**
  - How do we identify the long tail in each disambiguation task?
  - What are baselines for the long tail?
  - Should the long tail problem be tackled by evaluating on datasets with controlled properties, or by evaluation metrics?
  - Why can not all disambiguation tasks have the same evaluation metric?
  - Can we balance the long tail skewness by using the right evaluation metric?
  - Should there be weighted evaluation and what would this look like? (micro-average, annotation disagreement)?
  - Should we score the interpretation of the text or the knowledge derived from the text through reasoning? Accuracy versus Q&A.
  - Should evaluation addresses different ground truths (expert, crowd, etc..)
  - Should evaluation be done at different levels of granularity?

- **Hands-on data (90 mins):**
  - We will prepare a Google spreadsheet with an overview of existing evaluation metrics across NLP disambiguation tasks
  - For a fixed evaluation dataset and system output, measure the degree to which the long tail phenomena are prominent in different evaluation metrics.

- **Concluding the session+recommendations (30 mins):**
  - Which metric would you recommend for the SemEval 2018 Long Tail disambiguation task(s)?
  - Are the results from the hands-on generalizable over all disambiguation tasks?
  - How to incentivise systems that can deal with various contexts?

- **Preparing presentation (15 mins)**