Goal ⇒ What kind of datasets are needed for the long tail disambiguation task?

- **Questions:**
  - To what extent do we want to represent the long tail? Dromedaris shape, camel shape, or stegosaurus shape?
  - What should a dataset contain to be representative for the long tail?
  - What would be the properties of a dataset that makes it challenging for systems to understand the long-tail context?
  - What defines a long-tail context: time, place, topic, community?
  - Do we need multi-task datasets?
  - Isn't it better to tackle this by focusing on evaluation?
  - How to manipulate the temporal/spatial/topical/community context (Dutch and British football fans talking about the EC-France on a terrace in Paris on the same day) in datasets?
  - How can one create such dataset(s) (crowdsource, expert annotation, automatic acquisition)? The case of ECB+.
- **Hands-on data:**
  - We will prepare a set of datasets we analyzed for 5 disambiguation tasks in common format (TSV) and a metrics file with functions. People can run out-of-the-box evaluations and visualizations, or build their own analysis functions according to their will.
- **Concluding the session+recommendations:**
  - What are the properties of a representative dataset for the long tail?
  - Which methods would be optimal to create such a dataset?
  - Would we create many datasets or one big one?
  - Do we actually need these datasets? Why/why not?