## Data overview

For the course will we use the data published from the Amboseli Baboon project. The project task is to predict microbiome composition (the abundence of each taxa) in future time points. We will evaluate the model performance of using the Bray-Curtis distance between the measured and predicted microbiome composition.

**Data for milestone 1 +2**

Before milestone 1 you will receive two data files, train data and metadata. These files include a full-time series of 60 animals. You will also receive information on 20 animals with long-sliced time series that you will need to extrapolate on during milestone 3 (n = 5 latest time points for each animal).

1. Train data – a feature table of microbial relative abundance across samples. note that columns are microbes and rows are samples. Because the data is in relative abundance the sum of each row is equal to 1.
2. Train metadata: per sample metadata includes the following information:
   1. sample – a unique ID for each sample
   2. baboon\_id – a unique ID for each animal
   3. collection data
   4. sex
   5. age – age at sample collection
   6. social group – which social group the animal belongs to at sample collection
   7. group\_size – the size of the social group at sample collection
   8. rain\_mounth\_mm – amount of rain in mm
   9. season – dry or wet
   10. hydro\_year – the hydrological year in the reserve
   11. month
   12. diet\_PC – PCA on the food frequency questionnaire (13 PCs)

**Data for Milestone 3**

Before milestone 3 you will receive extra data that includes:

1. Sliced long time series - a metadata file for the sliced time series you received before.
2. Short time series - an additional 20 baboons in the same format. You will receive a limited number of samples from each baboon (n=10) and will need to extrapolate a full-time series. You will get a feature table for limited number of first timepoints and metadata for the complete time series.