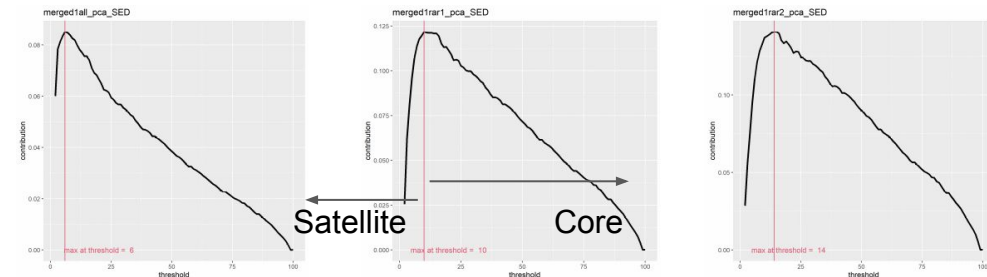# Topic 1: February 2022

Work done by Kadir Bice & Michaela de Melo & Masumi Stadler (discussed bi-weekly with team)

1 **Finalizing cleaning/filtering data:**
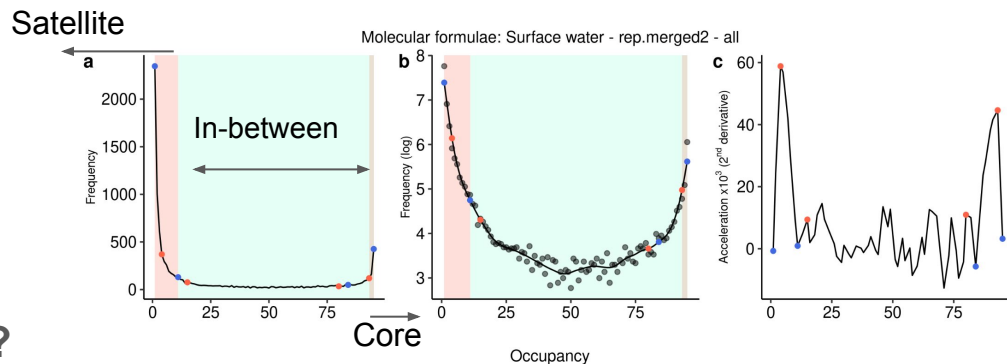
- Replicate merging
  - Replicates:
    - SW = 3 analytical replicates
    - SED = Up/Mid/Down
  - Merging thresholds:
    - MF/Peak in at least 1 sample
    - MF/Peak in at least 2 samples
- Rarity cut-off
  - None
  - At least in more than 1 sample
  - At least in more than 2 samples
- Summary table on Github

2 **How can we classify core and satellite?**

Literature threshold (LT) → PCA driven approach:
Threshold changes with processing method.

Emergent threshold (ET) → 2nd derivative of frequency-occupancy curve:

Molecular formulae: Surface water - rep.merged2 - all

Very likely we keep MF/MZ and replicate merging 1 and replicate merging 2 datasets, but drop rarity cutoffs.

Next steps:

- Explore how smoothing methods affect ET
- Explore 'in-between' group of ET and decide whether to keep it as a third category
- Finalize machine learning threshold (Random forest)