Hierarchical clustering

Procheta Sen

UNIVERSITY OF LIVERPOOL

Why Hierarchical Clustering?

- No need to specify the number of clusters k
- Specifies clusterings at all granularities, simultaneously
- Can be used to organise the data into a hierarchy of subsuming concepts for visualisation (abstraction) purposes

Representation of Hierarchical Clustering

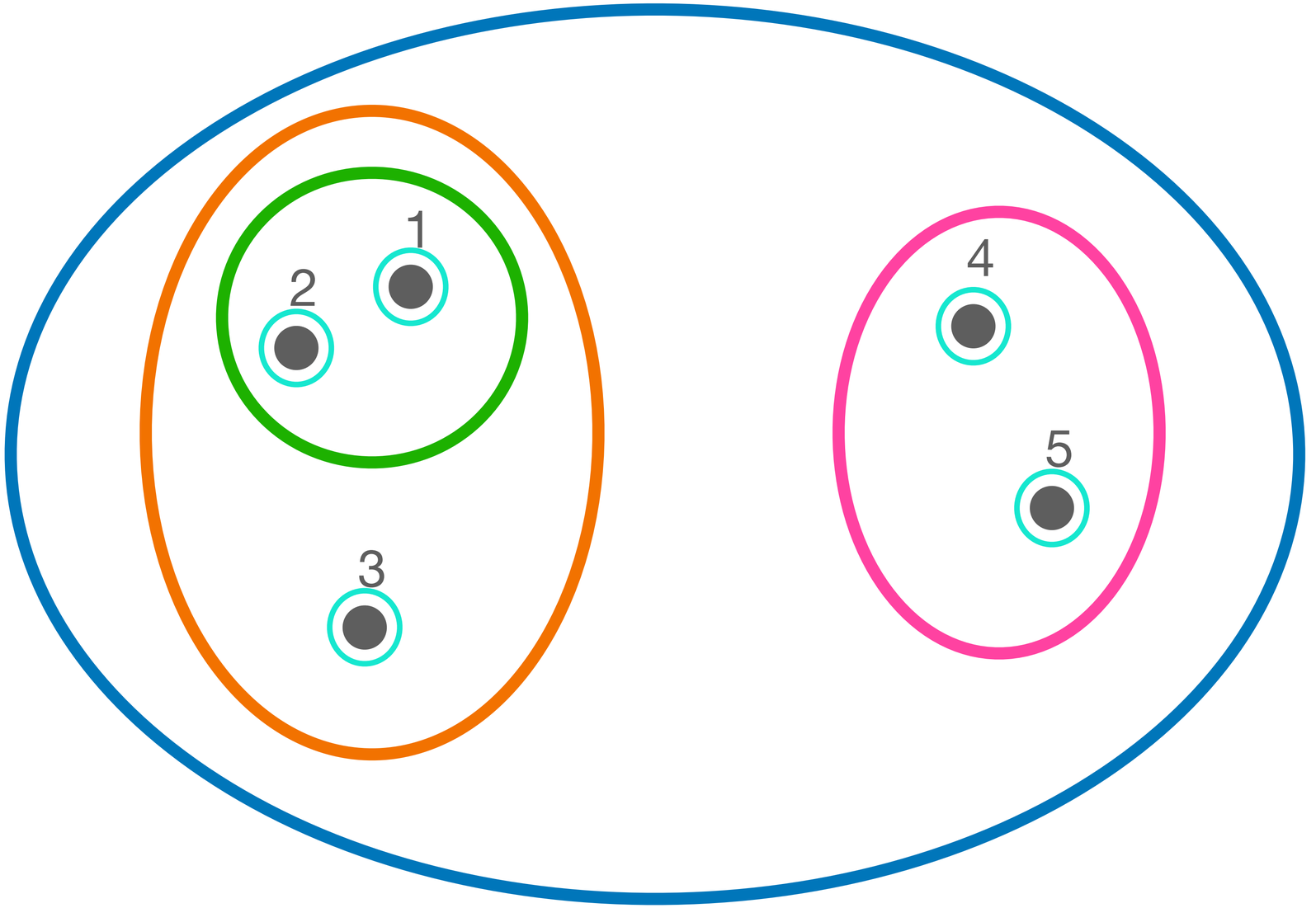
Hierarchy of clusters:

{1,2,3,4,5}

{1,2,3} {4,5}

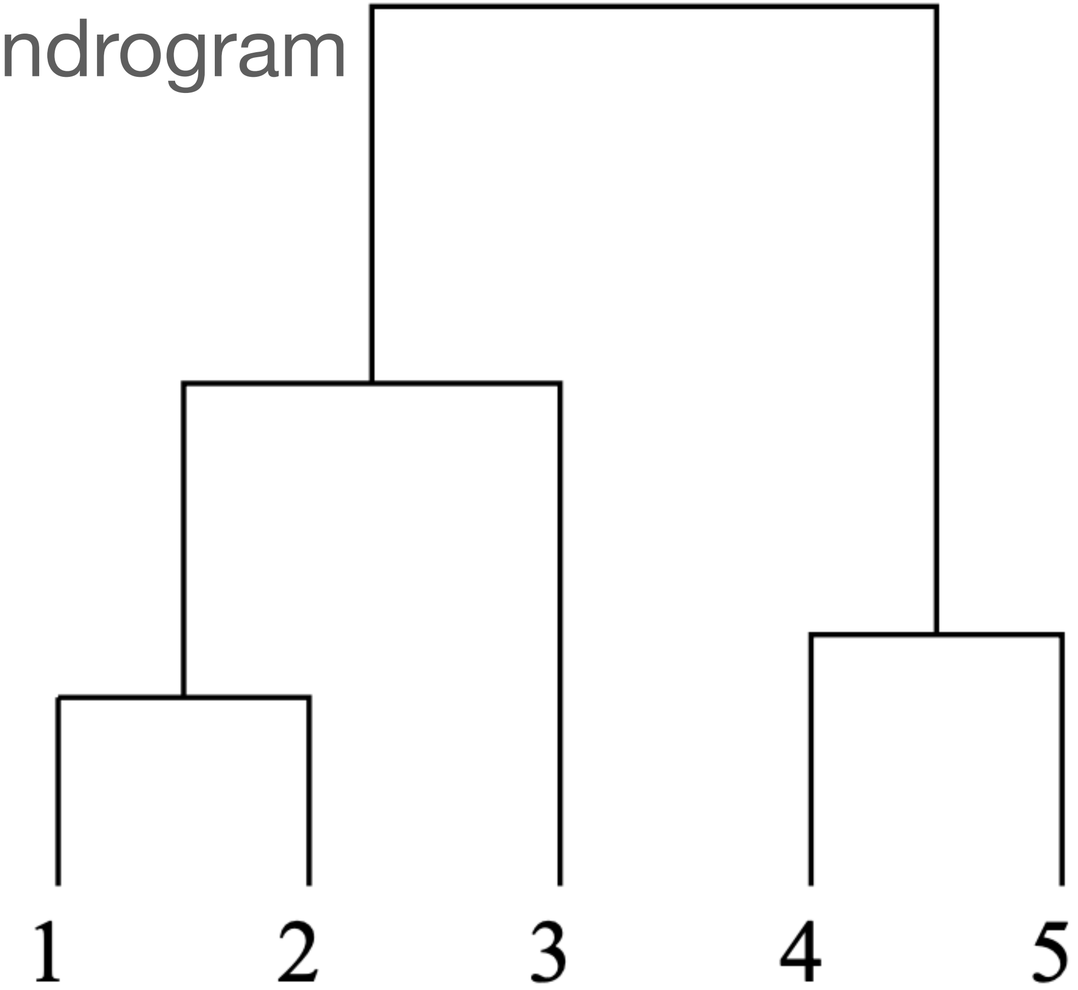
{1,2} {3} {4,5}

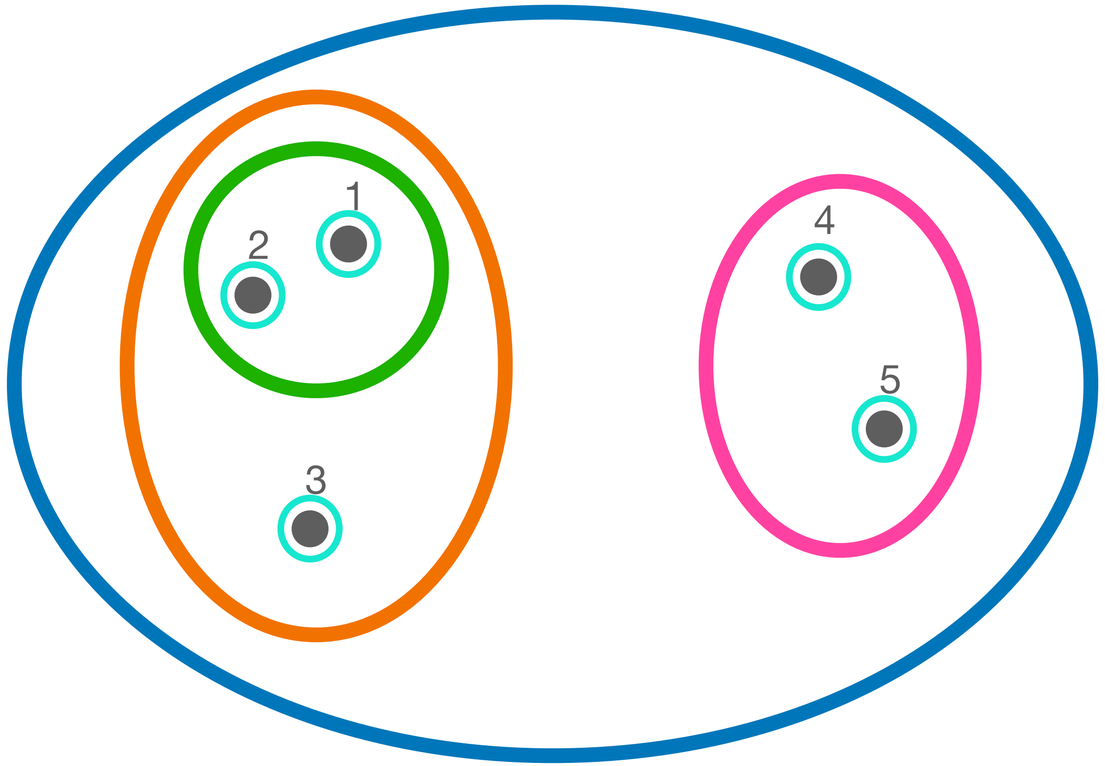
{1}{2}{3}{4}{5}



Representation of Hierarchical Clustering

Dendrogram





Hierarchy of clusters:

{1,2,3,4,5}

{1,2,3} {4,5}

{1,2} {3} {4,5}

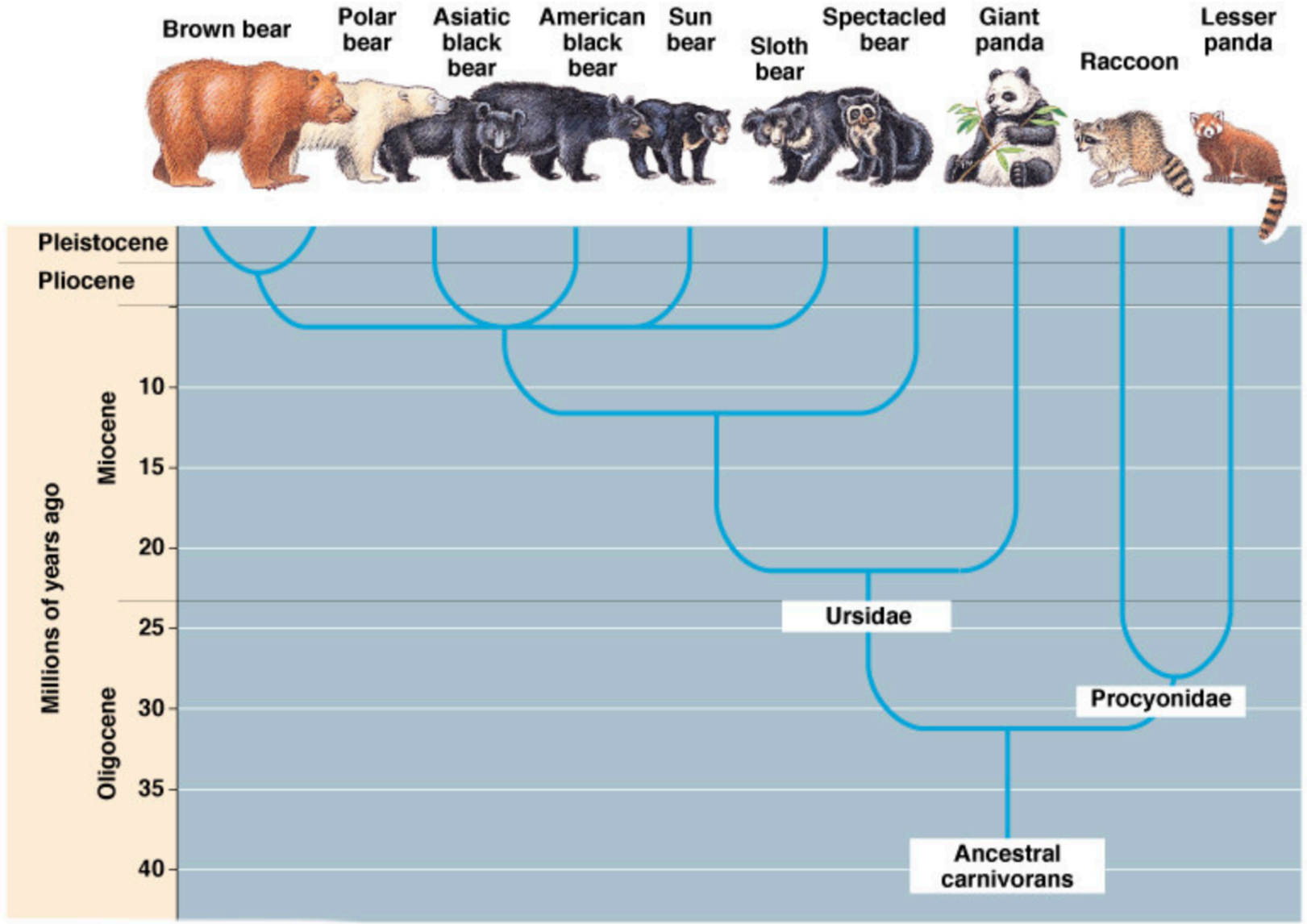
{1}{2}{3}{4}{5}

Example: Charting Evolution through Phylogenetic Trees

1. Generate the DNA sequences
2. Calculate the edit distance between all sequences.
3. Calculate the DNA similarities based on the edit distances.
4. Construct the phylogenetic tree.

https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

Example: Charting Evolution through Phylogenetic Trees



https://www.bearsinmind.org/Page/The-evolution-of-bear-species

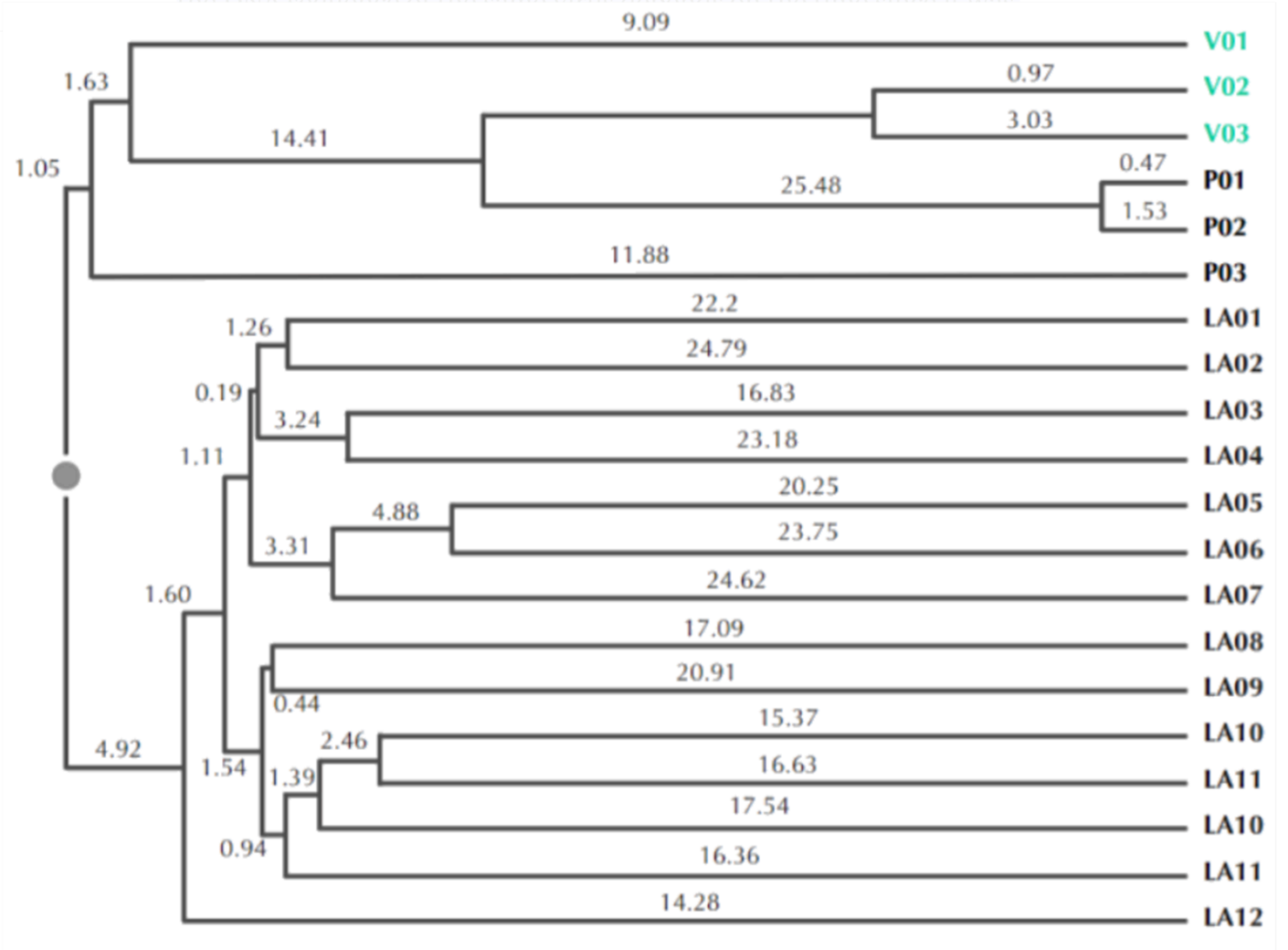
Example: Tracking Viruses through Phylogenetic Trees

- Viruses such as HIV have high mutation rates, which means the similarity of the DNA sequence of the same virus depends on the time since it was transmitted. This can be used to trace paths of transmission.

- Such method was used as evidence in a court case, wherein the victim's strand of HIV was found to be more similar to the accused patient's strand, compared to a control group.

https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

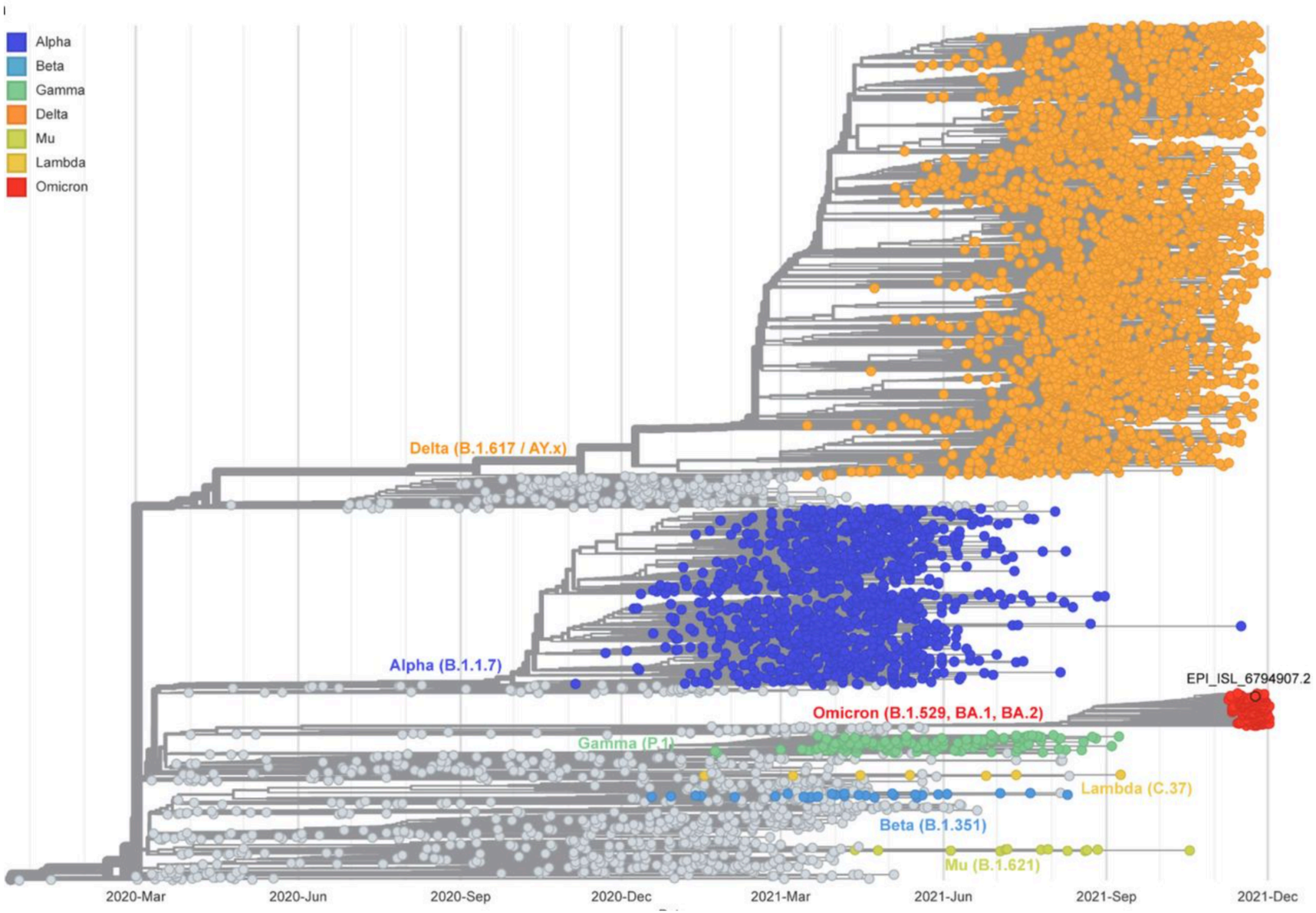
Example: Tracking Viruses through Phylogenetic Trees



V1–3 are victim’s strands, P1–3 are accused patient’s, and LA1–12 are the control group

https://towardsdatascience.com/hierarchical-clustering-and-its-applications-41c1ad4441a6

Example: Global phylogeny of SARS-CoV-2 highlighting the Omicron lineage.



Planas, D., et al. (2022). Considerable escape of SARS-CoV-2 Omicron to antibody neutralization. *Nature*, *602*(7898), 671-675.

Hierarchical Clustering

Two approaches

- **Agglomerative** clustering
  - Start with singletons (clusters with exactly one instance) and iteratively merge the most *similar* two clusters
  - Bottom-up approach
- **Divisive** (conglomerative) clustering
  - Start from one big cluster with all data instances and repeatedly partition it
  - Top-down approach