Data Type

- Sequential Data: In a sequence where order matters
- Data Matrix: A collection of records (n rows) and a fixed set of attributes (m columns)
- Spatial Data: Geographic locations and spatial information involved
- Temporal Data: Data involving time (with built-in support)
- Graph or Networks: Object connections
- Text: Sentences and documents
- Multi-Modality Data: Involving 2 or more kinds of data types
- Unknown Data Type: Data not shown

Python Programming

- Numpy: An additional plug-in to make Python more powerful

Python	**Meaning**
`import numpy`	Imports the entire numpy library into the Python file
`a = [1,2,3,4,5]`	Store the array [ ] in a variable a
`numpy.mean(a)`	Calculate the mean value stored in a
`numpy.std(a)`	Calculate the standard deviation stored in a
`numpy.median(a)`	Calculate the median value stored in a
`numpy.max(a)`	Calculate the maximum value stored in a
`print(a)`	Output the value stored in variable a
`print("a")`	Output a (inside the quotation mark " ")

Sequence Data

- For Central Dogma, Genetic information in DNA sequences

1. DNA sequence
    - A, T, C, G
    - 3 billion pairs
    - Doubled strand
2. RNA sequence
    - A, U, C, G
3. Protein sequence

    - 20 amino acids
    - Multiple sequence alignment

- Get sequence by
    1. Nanopore sequencing
        - DNA goes through chemical pore
        - Sequencing by detecting electrical current change caused by different bases
        - Long (3Mb)
        - High error rate
    2. Protein sequencing
        - Break long chains into short and short will be determined by mass spectrometry (weight)
        - Form raw sequence from short

- Raw Data and Handling
    1. DNA sequence

        Step 1: Quality Control

        Step 2: Mapping
        - Mark duplicates, sort and merge alignments

        Step 3: Variant calling
        - Variations recalibration, scoring, and filtering

        Step 4: Phenotype-associated variant
        - Link genetic variant to phenotype (observable trait)
    2. Protein sequence
        - Compare and multiple sequence alignment
        - Sequence-to-structure-to-function paradigm
            - Similar sequence = similar structure = similar function
        - Homology
            - Similar sequence = common ancestor

Sequence Comparison and Alignment Score

Compare 2 sequences by

1. Sequence Alignment: Determine similarity and its region
    - Biomolecular Functions and property prediction for Sequence-to-structure-to-function paradigm
    - Evolution, identifying conservation regions, investigating mechanisms for Homology
2. Pairwise sequence alignment: maximise 2 sequences' similarity by inserting gaps and score an alignment
    - Match (remain), Mismatch (substitute), Gap (insert or delete)

      I. Enumeration: calculate scores for all possible alignments, highest score = most similar

      II. Dynamic Programming: length n, (2n n) = (2n)!/(n!)^2