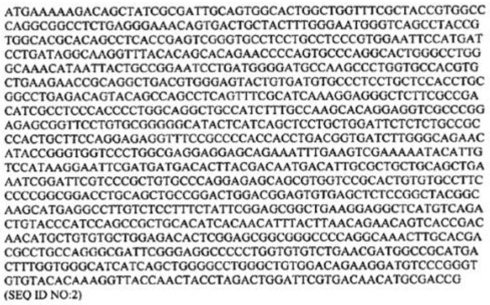
Lecture 3 – Data & Python & Sequence

Agenda

- Different data types
- Introduction to Python programming
- Sequence data
- Sequence comparison and alignment score

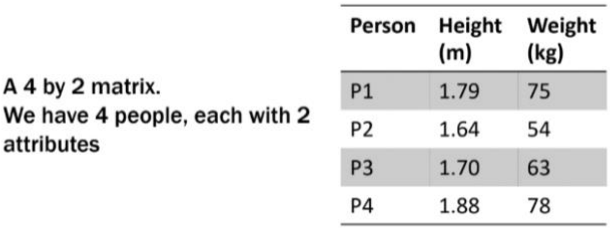
Data Types We Will Encounter

- Sequential Data
  - Ex) DNA sequence



```
ATGAAAAAGACAGCTATCGCGATTGCAGTGGCACTGGCTGGTTTCGCTACCGTGGCC
CAGGCGGCCTCTGAGGGAAACAGTGACTGCTACTTTGGGAATGGGTCAGCCTACCG
TGGCACGCACAGCCTCACCGAGTCGGGTGCCTCCTGCCTCCCGTGGAATTCCATGAT
CCTGATAGGCAAGGTTTACACAGCACAGAACCCCAGTGCCCAGGCACTGGGCCTGG
GCAAACATAATTACTGCCGGAATCCTGATGGGGATGCCAAGCCCTGGTGCCACGTG
CTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGATGTGCCCTCCTGCTCCACCTGC
GGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAAGGAGGGCTCTTCGCCGA
CATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCACAGGAGGTCGCCCGG
AGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATTCTCTCTGCCGC
CCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTGGGCAGAAC
ATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATACATTG
TCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTGA
AATCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTC
CCCCGGCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGC
AAGCATGAGGCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGA
CTGTACCCATCCAGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGAC
AACATGCTGTGTGCTGGAGACACTCGGAGCGGCGGGCCCCAGGCAAACTTGCACGA
CGCCTGCCAGGGCGATTCGGGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGA
CTTTGGTGGGCATCATCAGCTGGGGCCTGGGCTGTGGACAGAAGGATGTCCCGGGT
GTGTACACAAAGGTTACCAACTACCTAGACTGGATTCGTGACAACATGCGACCG
(SEQ ID NO:2)
```

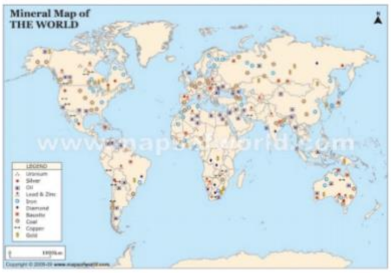
- Data Matrix
  - Consist of collection of records, each collection made up of a fixed set of attributes
  - Can be represented by [nxm] matrix; n rows for n object, m columns for m attributes
  - If you shuffle the entire column or the entire row at once, you would not change the data



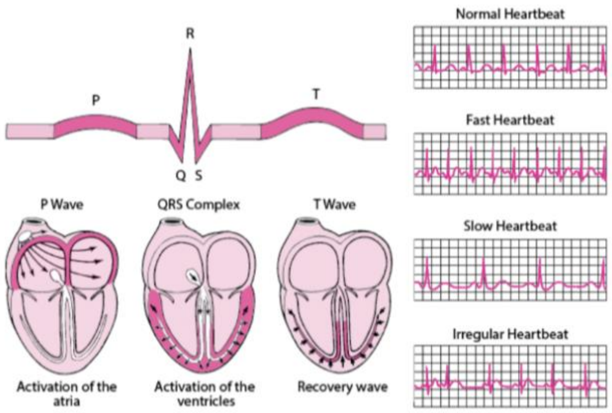
A 4 by 2 matrix.
We have 4 people, each with 2 attributes

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

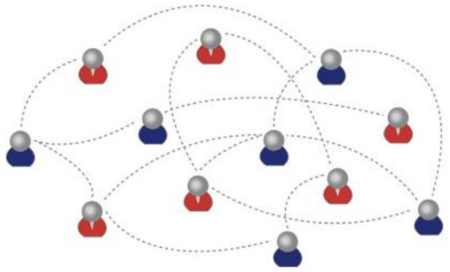
- Spatial Data
  - Geographic locations and spatial information



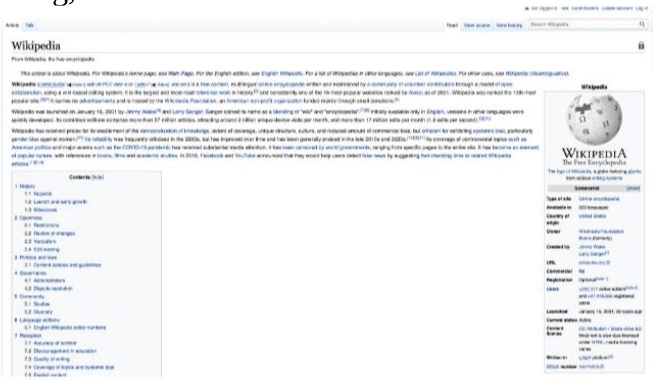
- Temporal Data
  - Data involving time



- Graph or Networks
  - Objects and connections (the link)
  - Social network and PPI network



- Text
  - Long, short sentences or documents



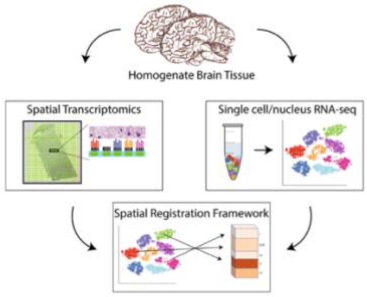
Wikipedia

From Wikipedia, the free encyclopedia

Contents

WIKIPEDIA
The Free Encyclopedia

- Multi-Modality Data
  - Video: temporal images, audio, transcript
  - Electronic health records: data matrix, images, text
  - Spatial transcriptomics: spatial data, sequence, data matrix



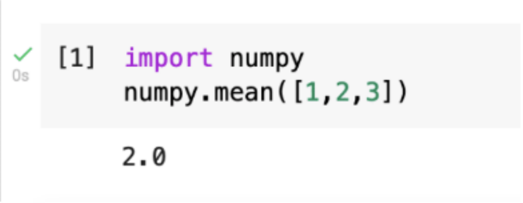
- Unknown Data Types
  - Data not shown

Python Programming

- Programming
  - communicating with the computer, asking to do something (task)
- Code
  - message we send to the computer
- Python
  - the software used to send the message + it translates human language to computer language

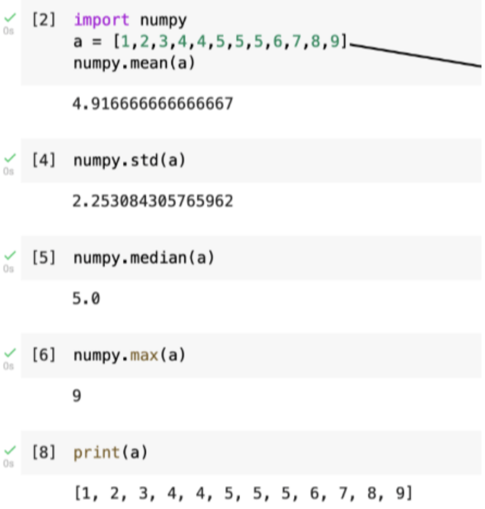
Writing Python Code

Ex) Calculating mean of some values



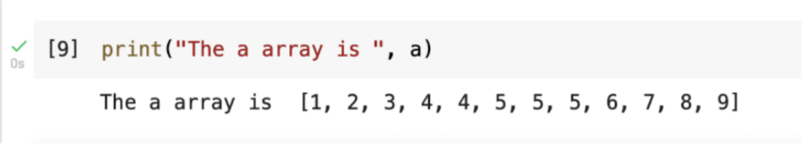
```
[1]  import numpy
     numpy.mean([1,2,3])

     2.0
```

- What is NumPy?
  - Additional plug-in to make Python more powerful
  - Load them by *import.numpy* when using
  - Other plug-in: SciPy, Pandas
- What if we want to calculate other things?
  - For example: mean, variance, median, max, min, etc
  - Store the array in a variable to avoid entering them every time



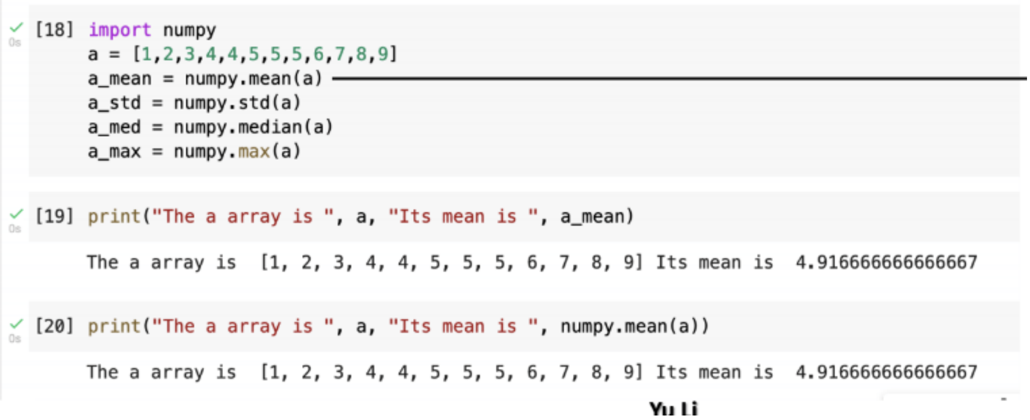
```
[2]  import numpy
     a = [1,2,3,4,4,5,5,5,6,7,8,9]
     numpy.mean(a)

     4.916666666666667

[4]  numpy.std(a)

     2.253084305765962

[5]  numpy.median(a)

     5.0

[6]  numpy.max(a)

     9

[8]  print(a)

     [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9]
```

- What is "print"?
  - Asking the computer to print something.
  - If the variable is storing some values, everything in "" will we printed.



```
[9]  print("The a array is ", a)

     The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9]
```

- Other Example:



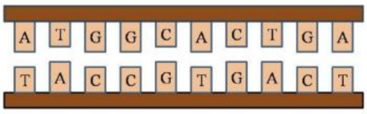
```
[18] import numpy
     a = [1,2,3,4,4,5,5,5,6,7,8,9]
     a_mean = numpy.mean(a)
     a_std = numpy.std(a)
     a_med = numpy.median(a)
     a_max = numpy.max(a)

[19] print("The a array is ", a, "Its mean is ", a_mean)

     The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9] Its mean is  4.916666666666667

[20] print("The a array is ", a, "Its mean is ", numpy.mean(a))

     The a array is  [1, 2, 3, 4, 4, 5, 5, 5, 6, 7, 8, 9] Its mean is  4.916666666666667
```

  - a_mean = numpy.mean(a): calculate the mean of array a then store the calculated mean into the variable a_mean.

Sequence Data

DNA sequences are where the genetic information is hidden and to understand central dogma thorough study of DNA is needed. Human genotypes are determined by sequences.

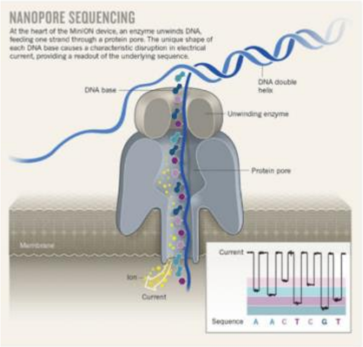
- DNA sequence
  - Composed of A,T,C,G
  - Complementary double strand
  - Approximately 3 billion base pairs



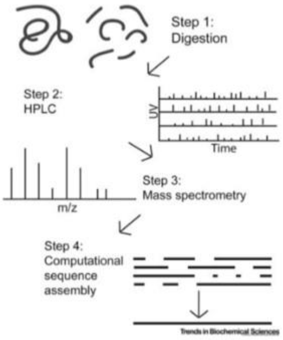
- RNA sequence
  - Composed of A,U,C,G
  - Single stranded
- Protein sequence
  - Usually composed of 20 amino acids
  - Multiple sequence alignment

Obtaining Sequences

- DNA/RNA sequencing
  - Under active development
  - From short reads to long reads
- Nanopore Sequencing
  - DNA goes through a chemical pore
  - Different bases → electrical current changes
  - Sequence by detecting the current change
  - Can sequence very long samples (up to 3Mb)
  - High error rate (5%)
  - Under active development


NANOPORE SEQUENCING
At the heart of the MinION device, an enzyme unwinds DNA, feeding one strand through a protein pore. The unique shape of each DNA base causes a characteristic disruption in electrical current, providing a readout of the underlying sequence.
DNA base
DNA double helix
Unwinding enzyme
Protein pore
Membrane
Ion
Current
Current
Sequence A A C T C G T

- Protein Sequencing (mostly based on mass spectrometry)
  - Break the long sequence into short pieces
  - Each piece can be determined by mass spectrometry
  - It is determined by the weight of each pieces
  - Assemble the short pieces into the raw sequence



Raw Data – What do we do to them?

- DNA sequences
  - Quality control
  - Map reads to reference genome
  - Variant calling
  - Phenotype associated variants
- Protein sequences
  - Sequence comparison
  - Multiple sequence alignment
  - As similar sequence my indicate similar structure which results in similar function
  - Similar sequence may also indicate common ancestor

So how do we compare two sequences?

⇨ Sequence Comparison and Alignment Score

Sequence Alignment

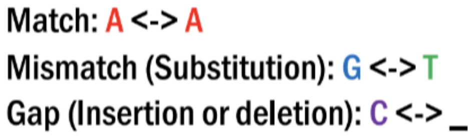
To determine the similarity between sequences and identify regions of similarity

- Why is it important?:
  - For biomolecular function and property prediction
  - For evolution, identifying conservative region, investing mechanism

Sequence Alignment and Sequence Similarity

- Pairwise sequence alignment
  - Arrange two sequence to maximize the similarity between them

  - Gap can be inserted in the sequences
- Defining sequence similarity

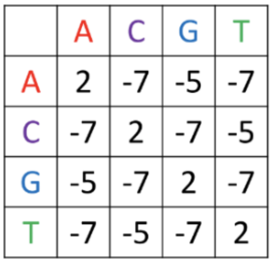


Match: A <-> A

Mismatch (Substitution): G <-> T

Gap (Insertion or deletion): C <-> _

- Sequence alignment score
  - Scoring matrix:



	A	C	G	T
A	2	-7	-5	-7
C	-7	2	-7	-5
G	-5	-7	2	-7
T	-7	-5	-7	2

  - Gap penalty: -10

<u>Finding the Best Pairwise Alignment</u>

Known: two sequences, scoring matrix

- Straight Forward solution: Enumeration
  - Enumerate all possible alignments
  - Calculate the scores for all the alignment
  - Find the one with the highest score
- Problem:
  - Too many possible combination
  - Need Dynamic Programming