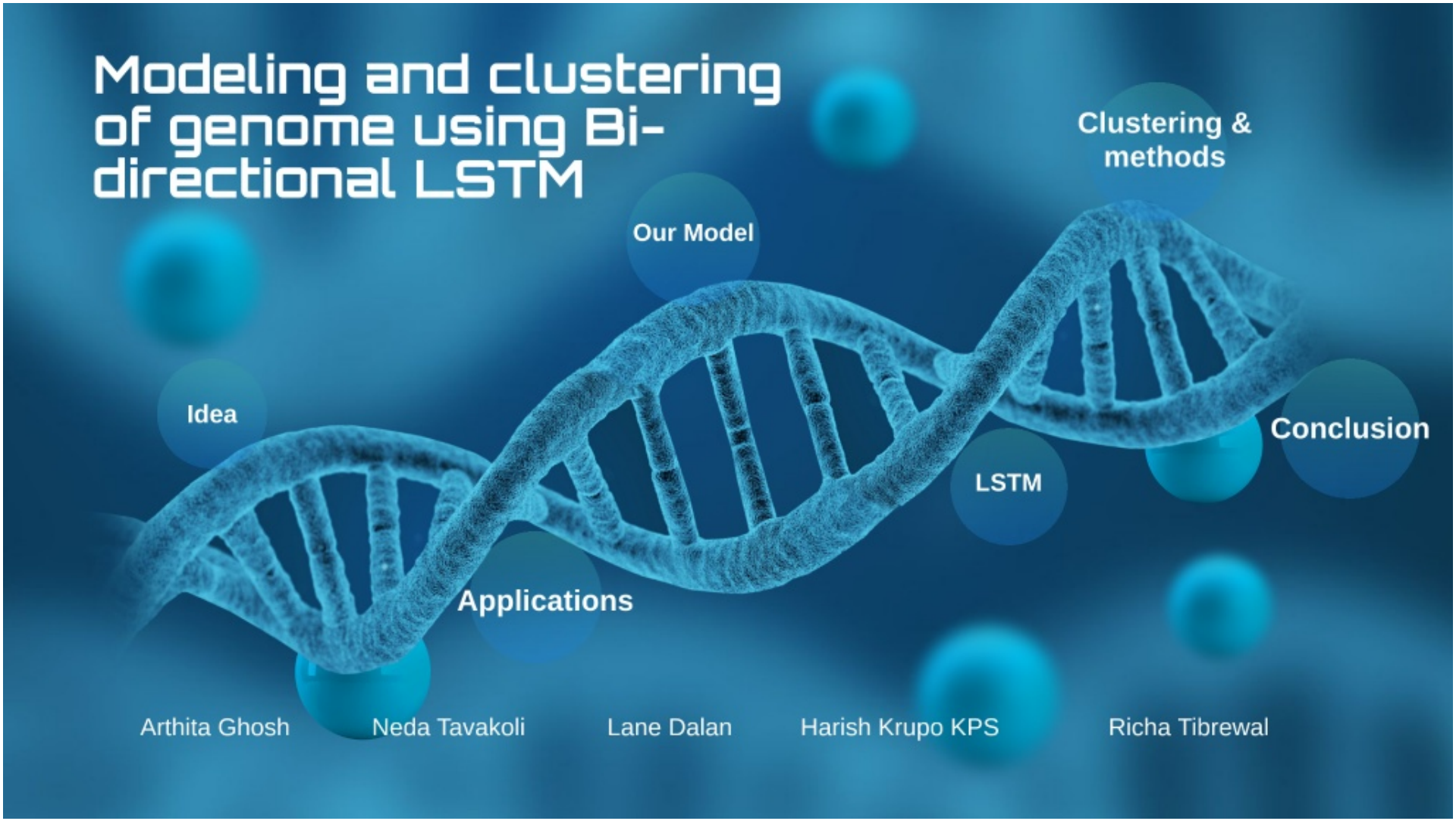

Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

What is a human genome?

Simply the sum total of an organism's DNA

Human genome

↓

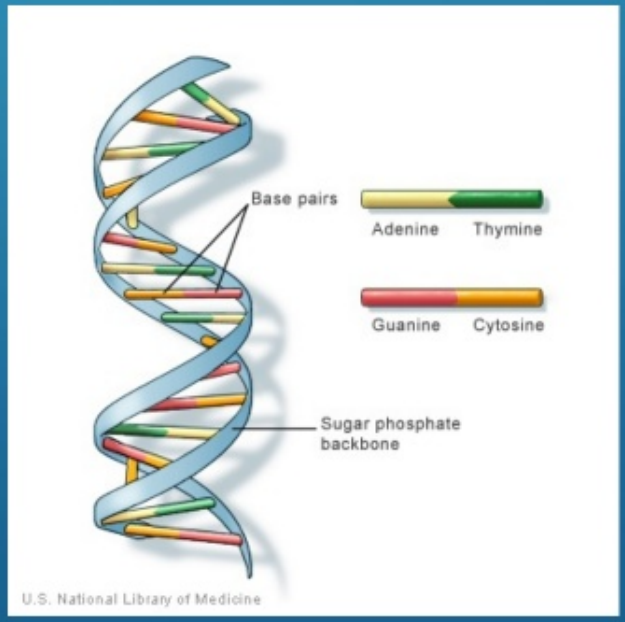
23 pairs of chromosomes

↓

DNA

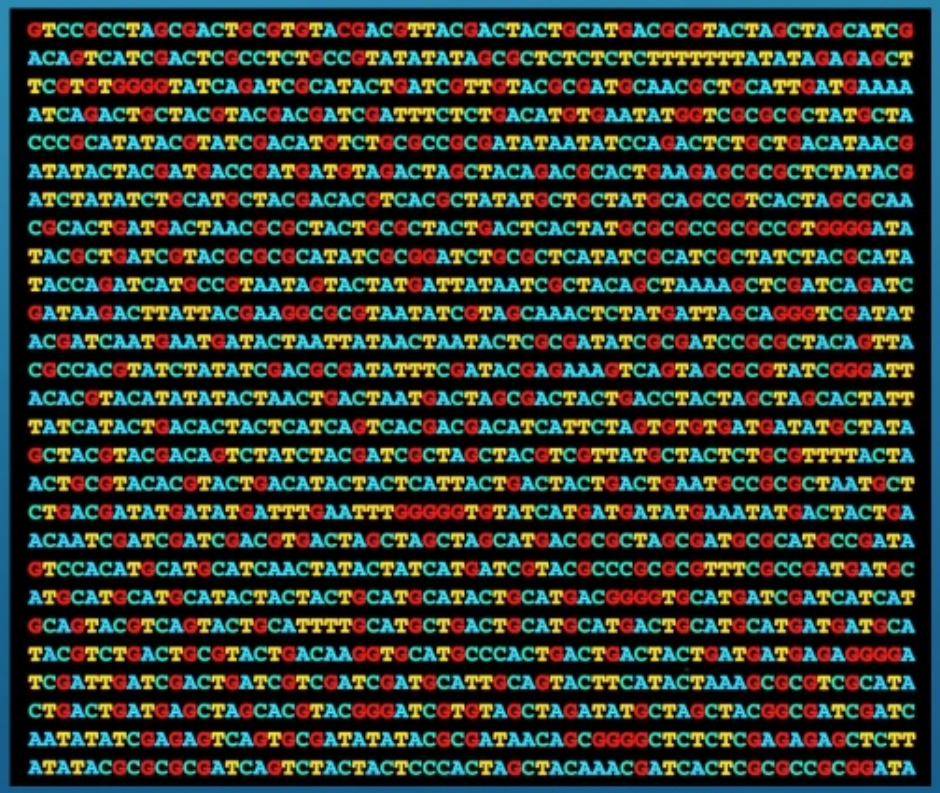
↓

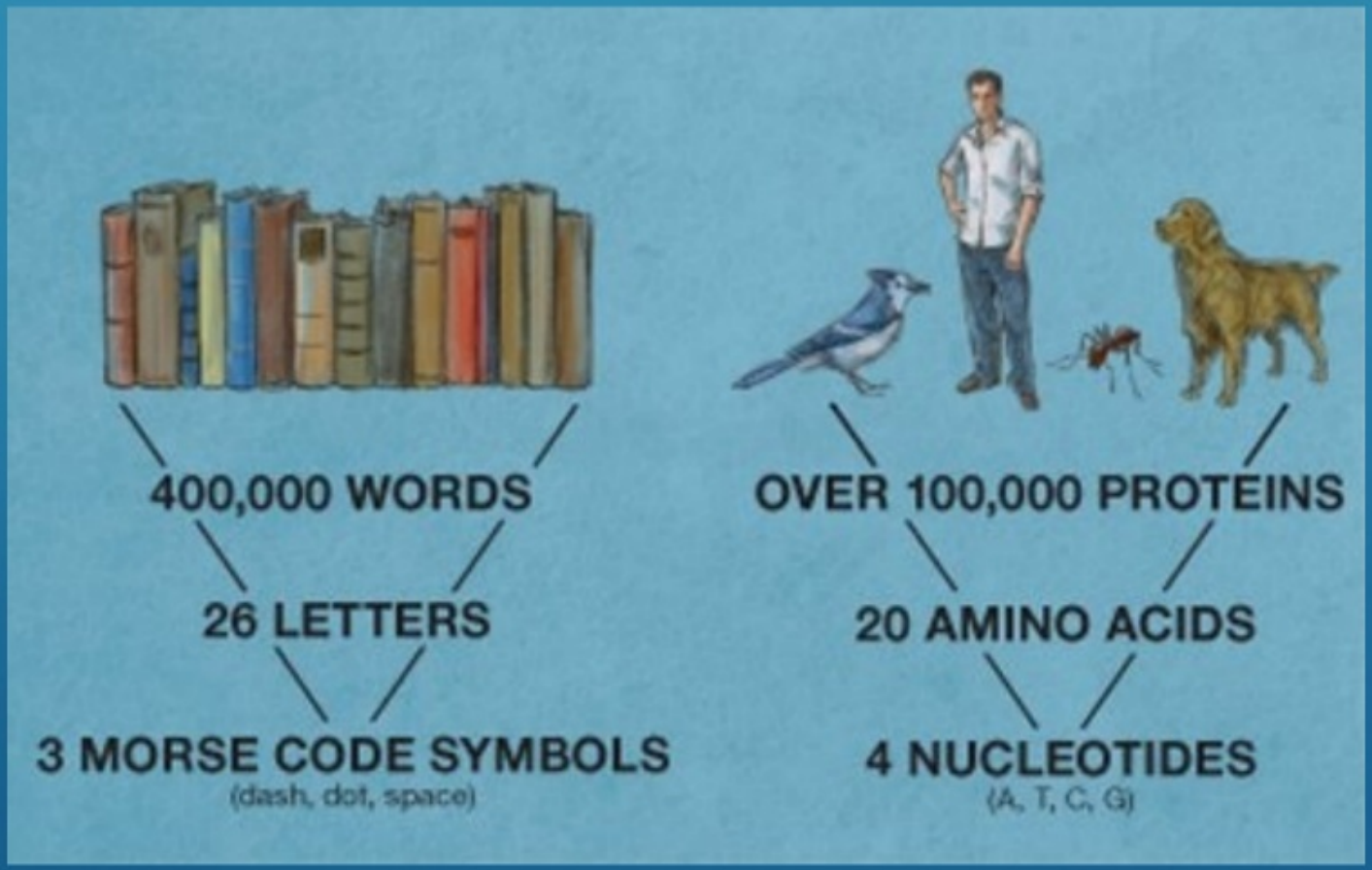
A,T,G,C sequences



~ 6.4 billion characters (A,T,G,C) in a human genome

Repeating sequences in the genome

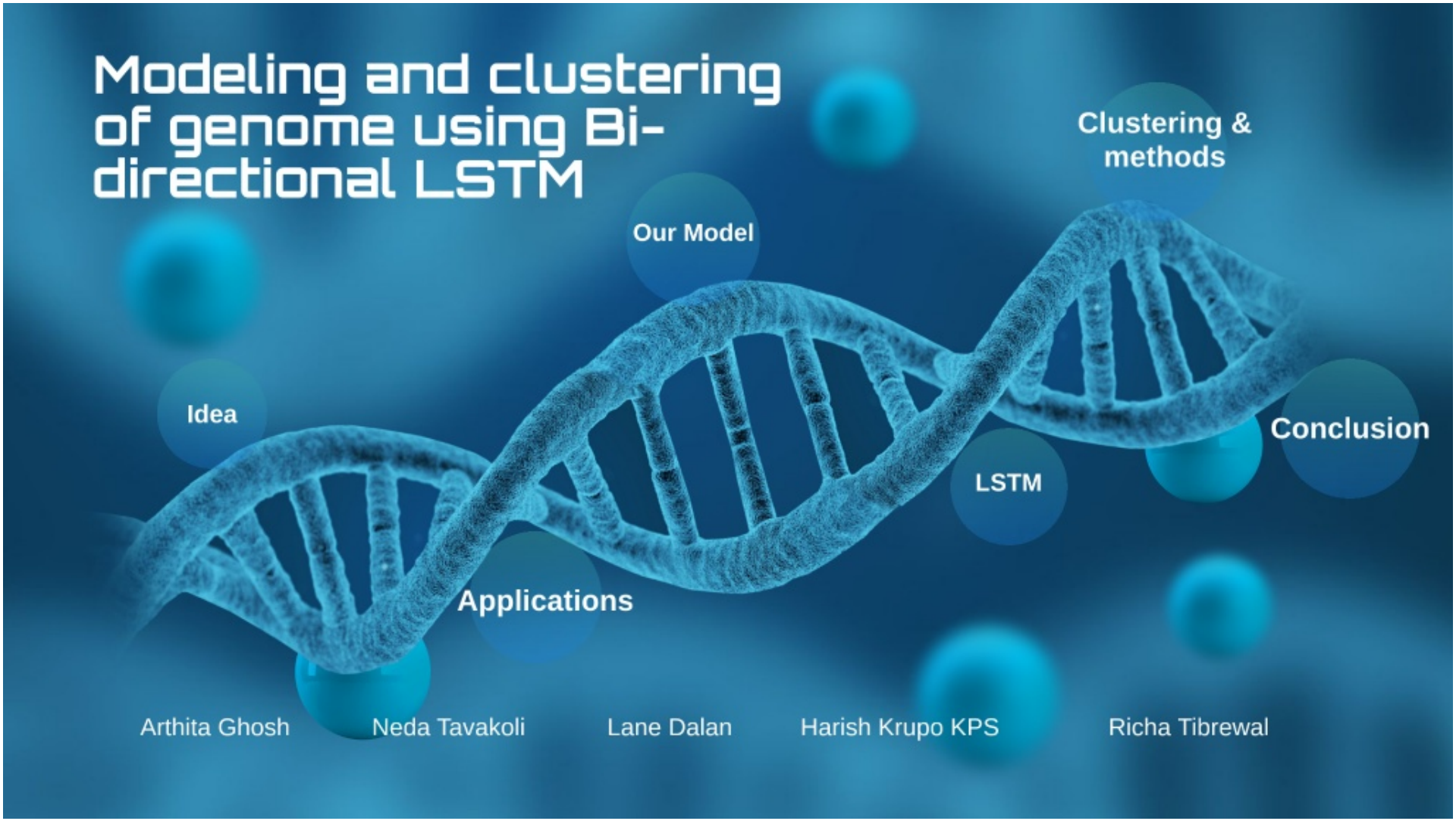




Language of Life: Francis Collins

Our objective

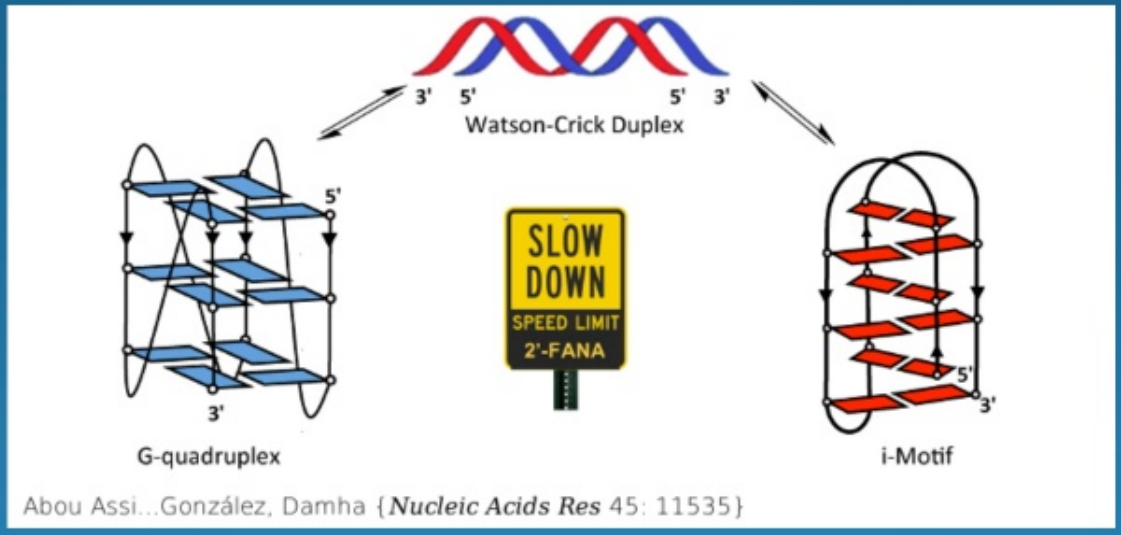
- Learning repeating sequences in a genomic sequences.
- Clustering similar genome sequences.


Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

Applications: Why Deep learning?

Understanding all functions and interactions of genome is challenging

Genome data → **DL (Bi-LSTM)** → **Extract highly complex patterns (Motifs)**



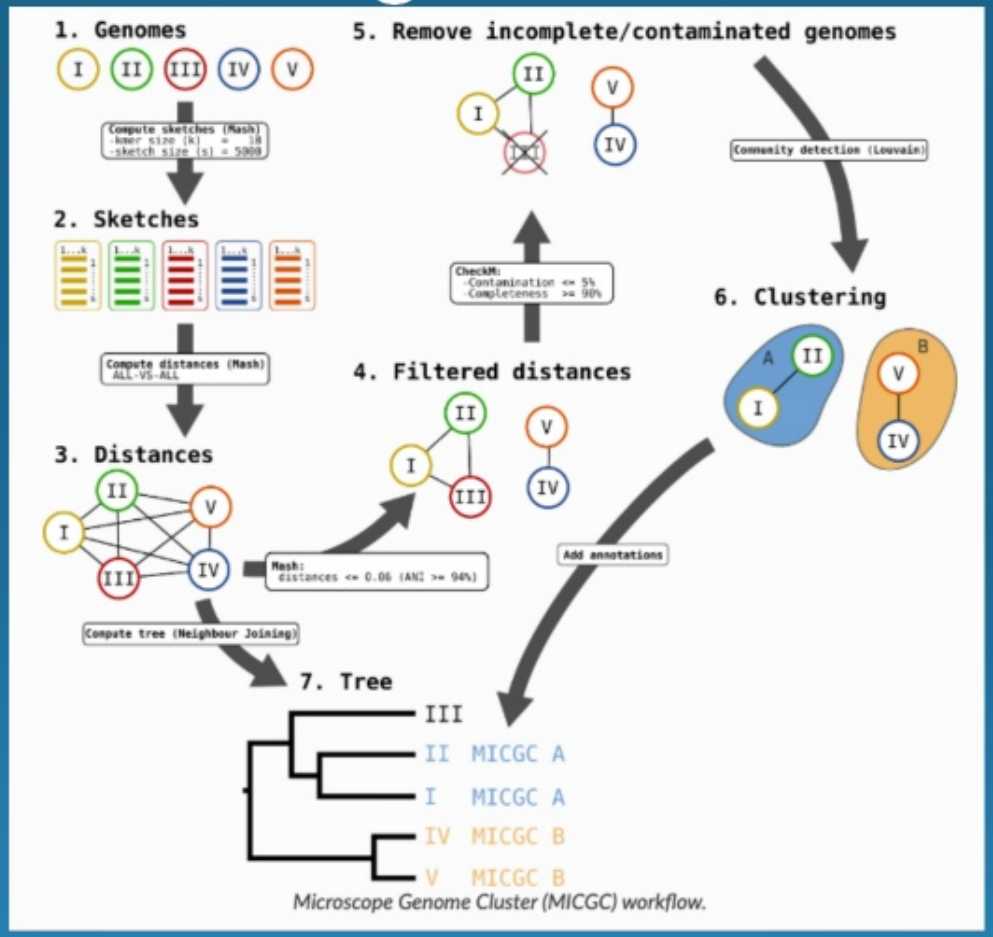
Abou Assi...González, Damha {*Nucleic Acids Res* 45: 11535}

Applications: Clustering Genome

Traditional Genome clustering:

- Computationally intensive
- Impractical for large genome

Deep learning: Train once , use multiple times



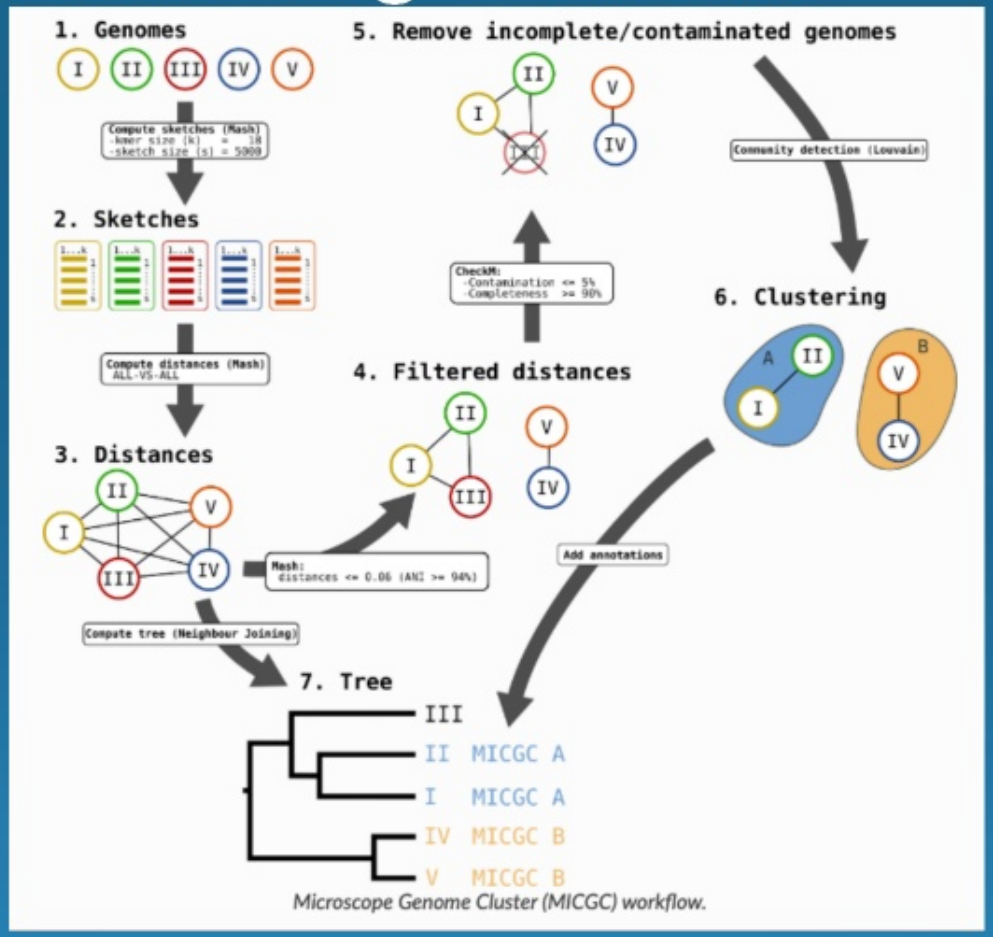
Microscope Genome Cluster (MICGC) workflow.

Applications: Clustering Genome

Traditional Genome clustering:

- Computationally intensive
- Impractical for large genome

Deep learning: Train once , use multiple times



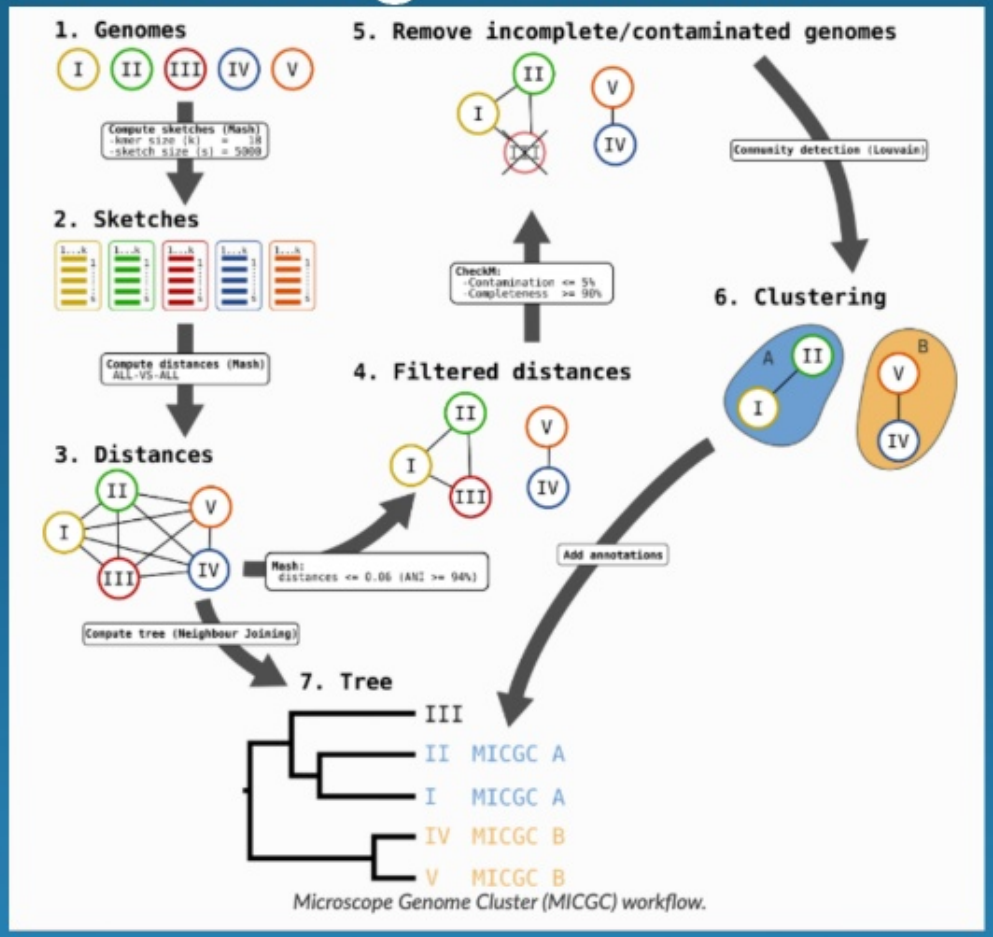
Microscope Genome Cluster (MICGC) workflow.

Applications: Clustering Genome

Traditional Genome clustering:

- Computationally intensive
- Impractical for large genome

Deep learning: Train once , use multiple times



Microscope Genome Cluster (MICGC) workflow.

Applications: Clustering Genome using Deep learning

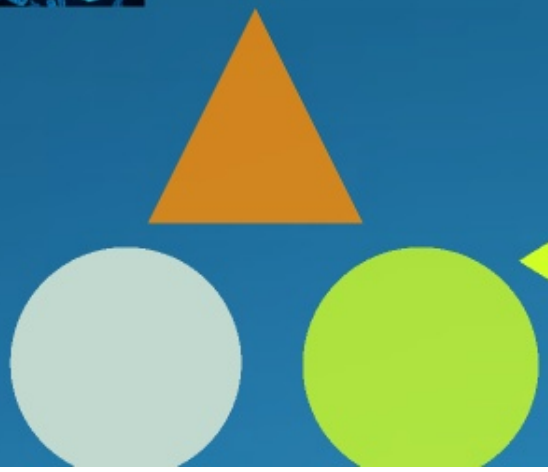
Bi-LSTM

Genome data → Build a model

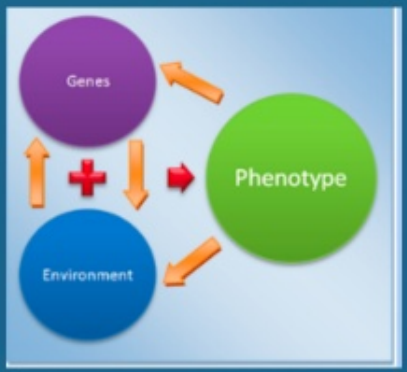
Train

Build a model → Vector representation of Genome

Vector representation of Genome → ML (DBSCAN, Kmeans, GMM)

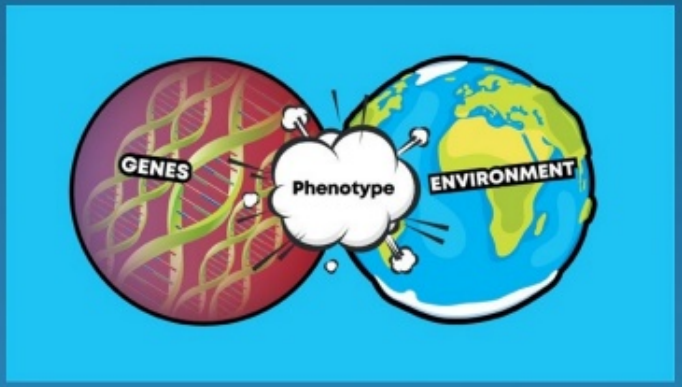


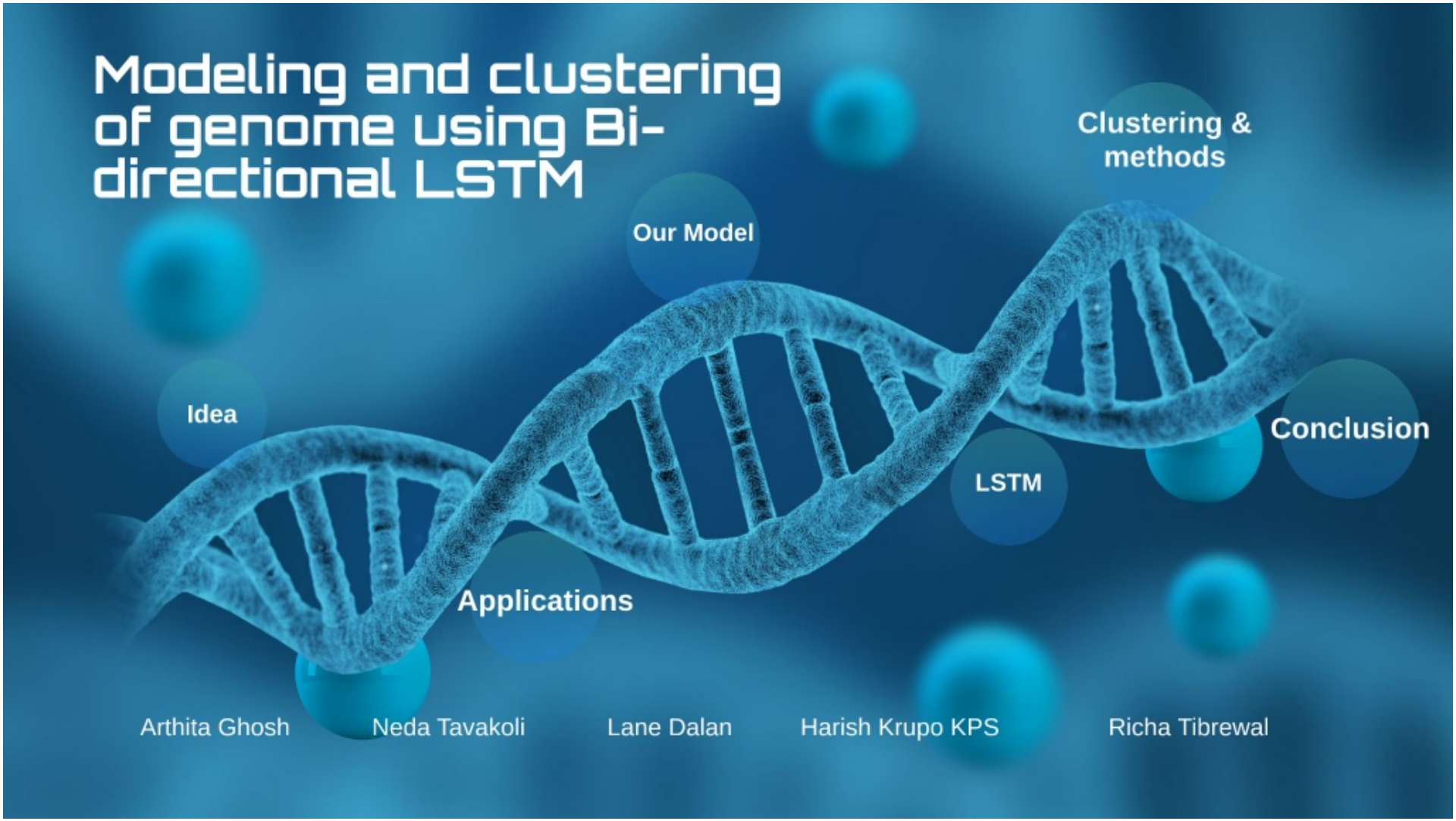
Other Applications:



- Modeling genome using LSTM:
  - Automatically extract novel features from input data to:
    - Connecting genotype to phenotype
    - Predicting regulatory function
    - Classifying mutation types

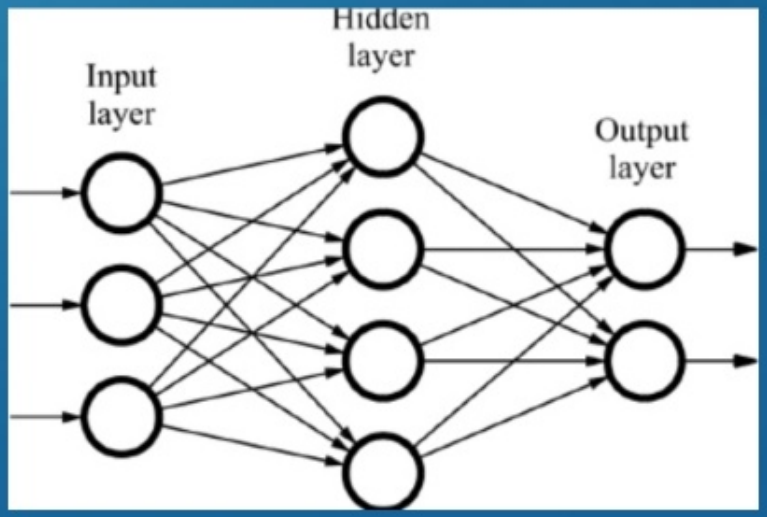
Phenotype: The physical characteristics
Genotype: The genetic composition



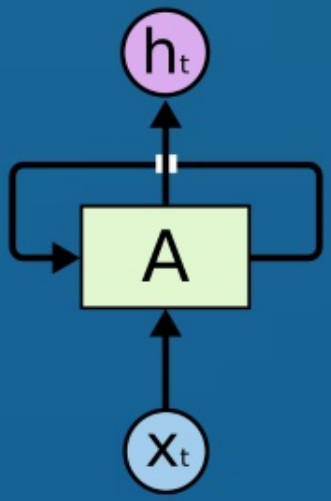

Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

Recurrent Neural Networks

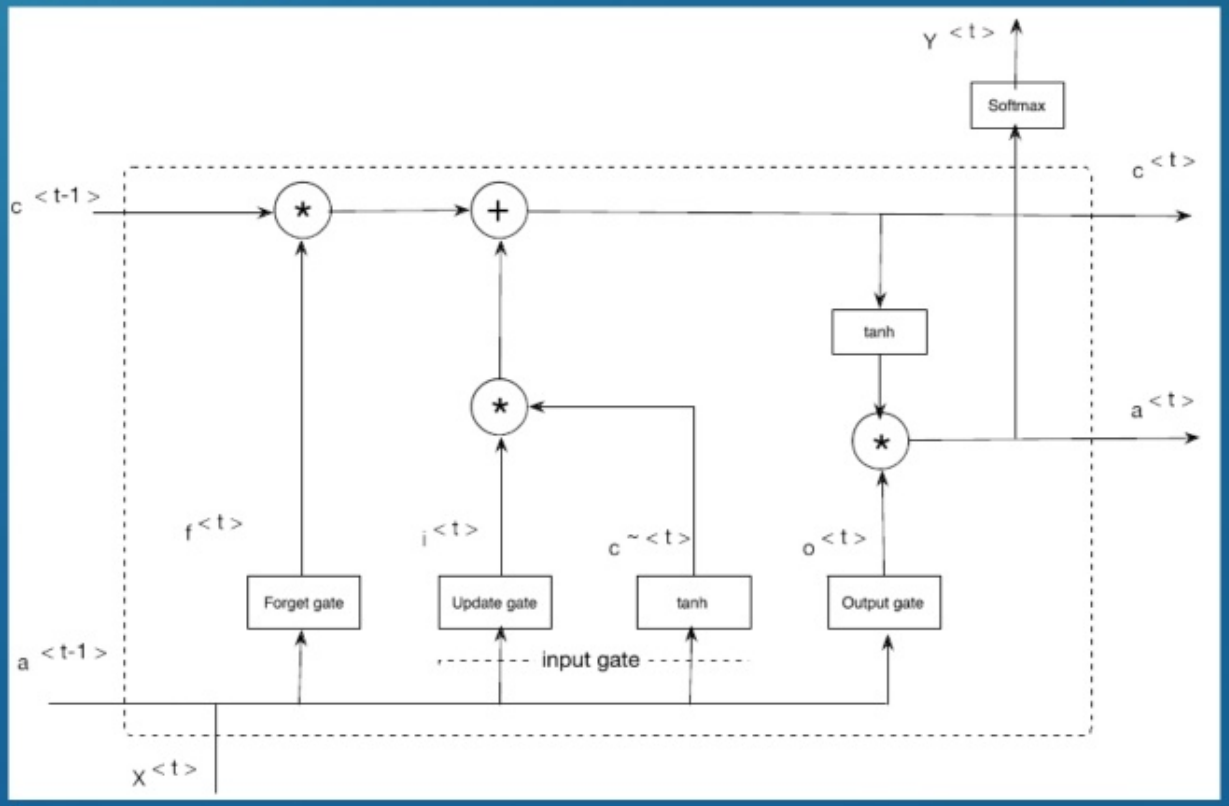
Traditional Network



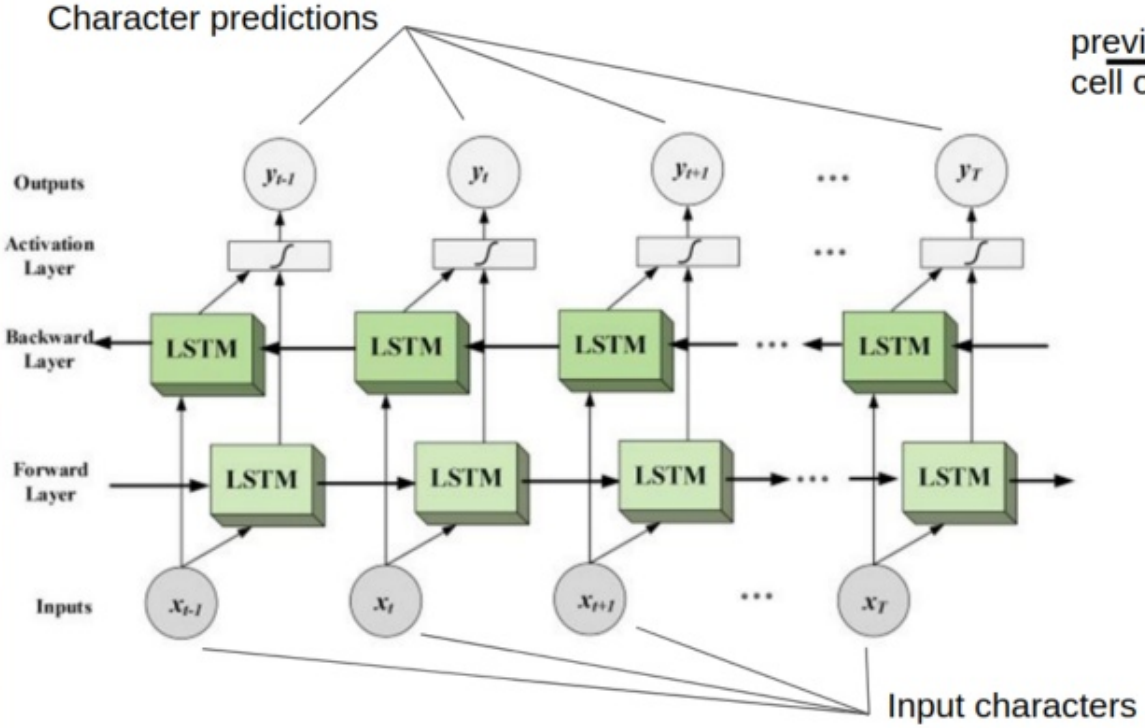
Recurrent Network

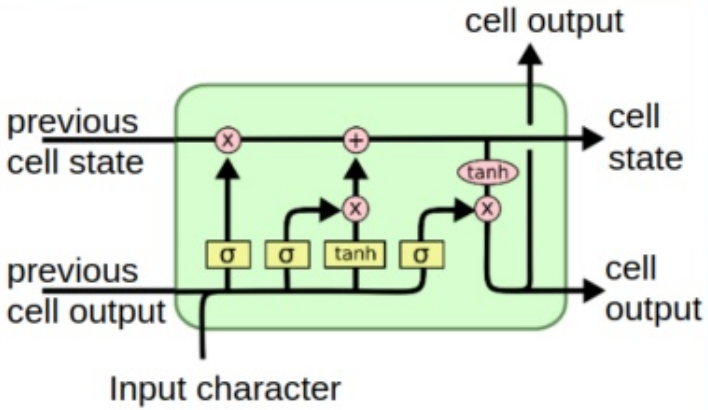


LSTM, Long Short Term Memory



Our Model

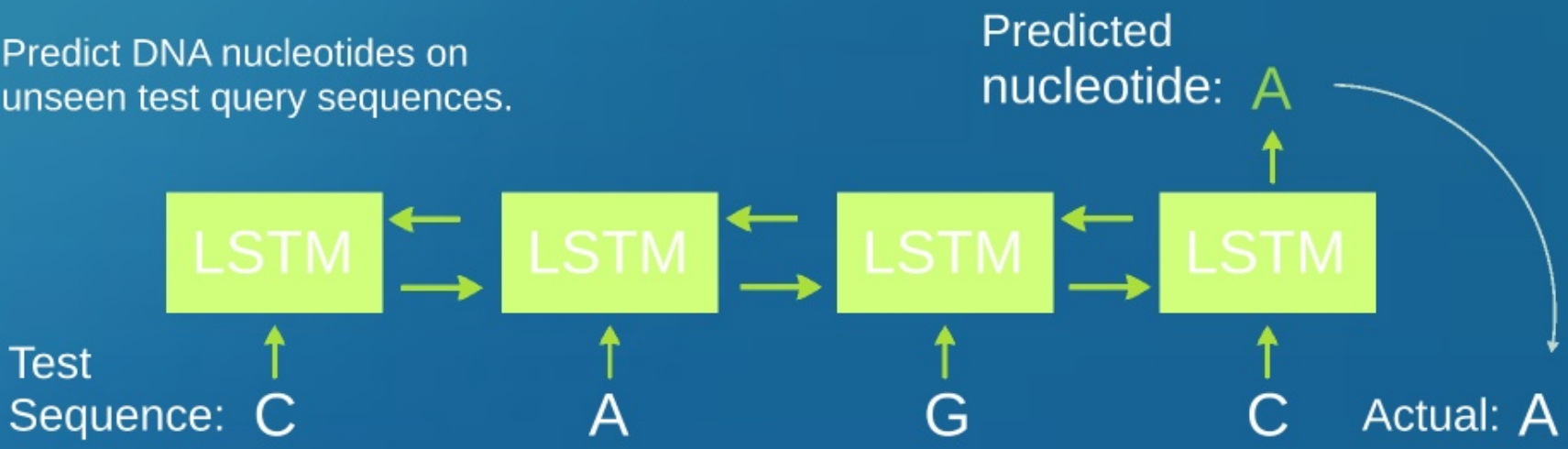




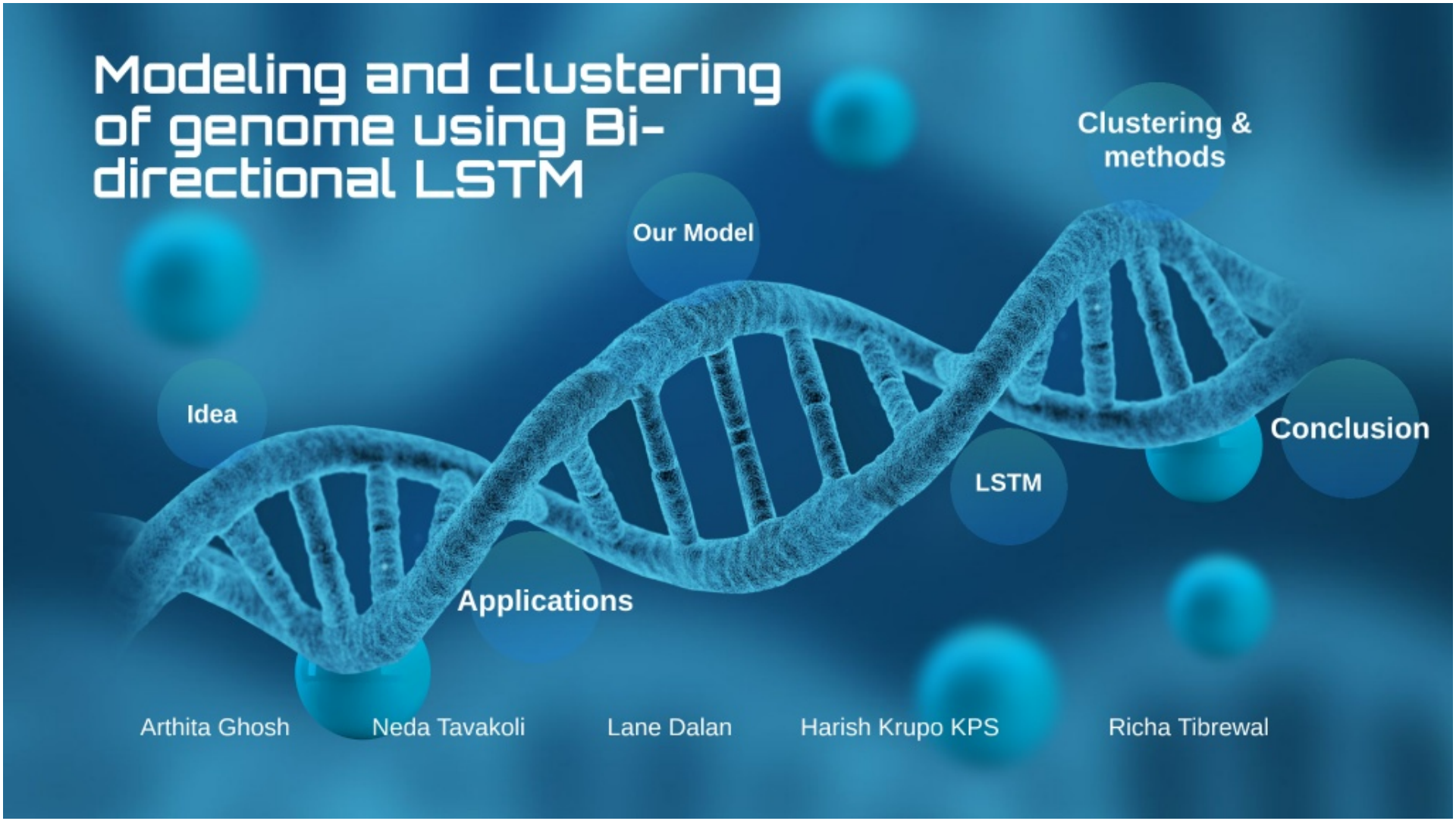
Vectorization of the input sequence is taken from the hidden layer of the last cell.

Validating DNA Model

- Predict DNA nucleotides on unseen test query sequences.



- Cluster with alternative method of DNA embedding and compare results
  - fastDNA looks promising


Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

LSTM

- LSTMs excel for sequential data
- Following slides explain text generation using LSTM
- Can be easily extended to DNA just by changing input to DNA sequences

Text generation using LSTM

Steps:

- Dictionary Building
- Data set Generation
- Training
- Prediction

Text generation using LSTM

Dictionary building and encoding

char to id

```
{
"a" : 1,
"b" : 2,
"c" : 3,
⋮
}
```

id to char

```
{
1 : "a",
2 : "b",
3 : "c",
⋮
}
```

DNA sequence to id

```
{
"AAA" : 1,
"ATG" : 2,
"GTC" : 3,
⋮
}
```

id to DNA sequence

```
{
1 : "AAA",
2 : "ATG",
3 : "GTC",
⋮
}
```

Text generation using LSTM

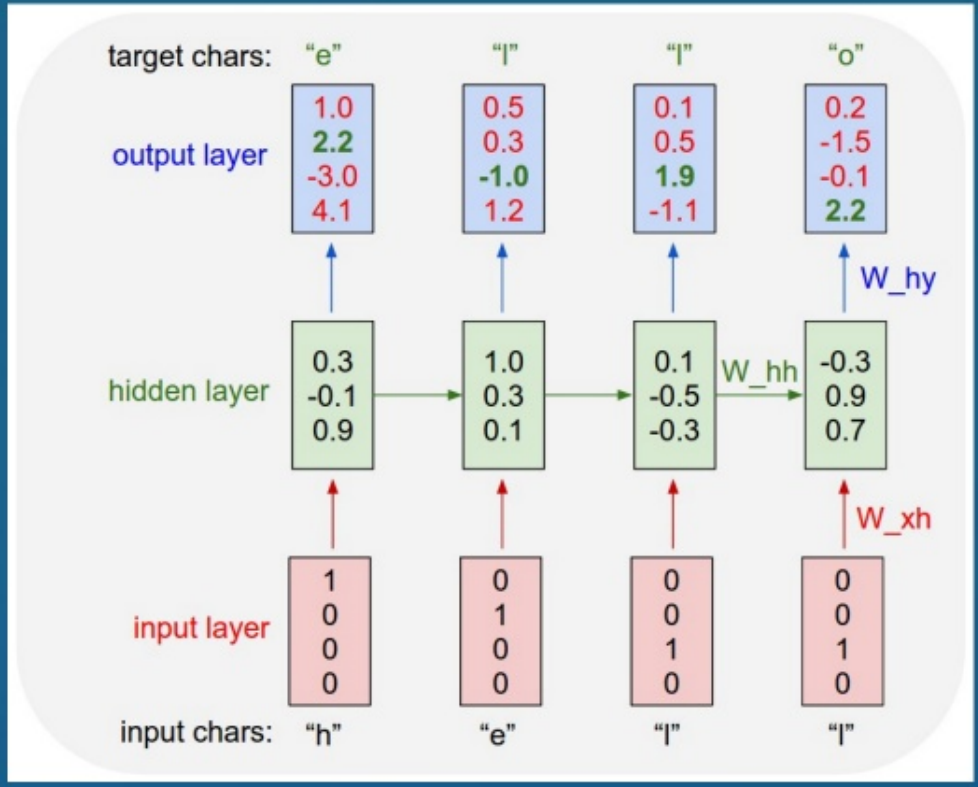
Data Set generation:

- The encoded sequence is split into chunks of size L
- L is the size of the unrolled LSTM

```
Dataset =
{
    {
        input = "31 21 40 40"      (encoding of the characters: "hell")
        target = "21 40 40 56"     (encoding of the characters: "ello")
    }
      .
      .
      .
}
```

Text generation using LSTM

Training and prediction



Text generation using LSTM

Training:

Hidden state update:

h(t) = tanh(dot(W_hh, h(t-1)) + dot(W_xh, x(t)))

Output generation:

y(t) = dot(W_hy, h(t))

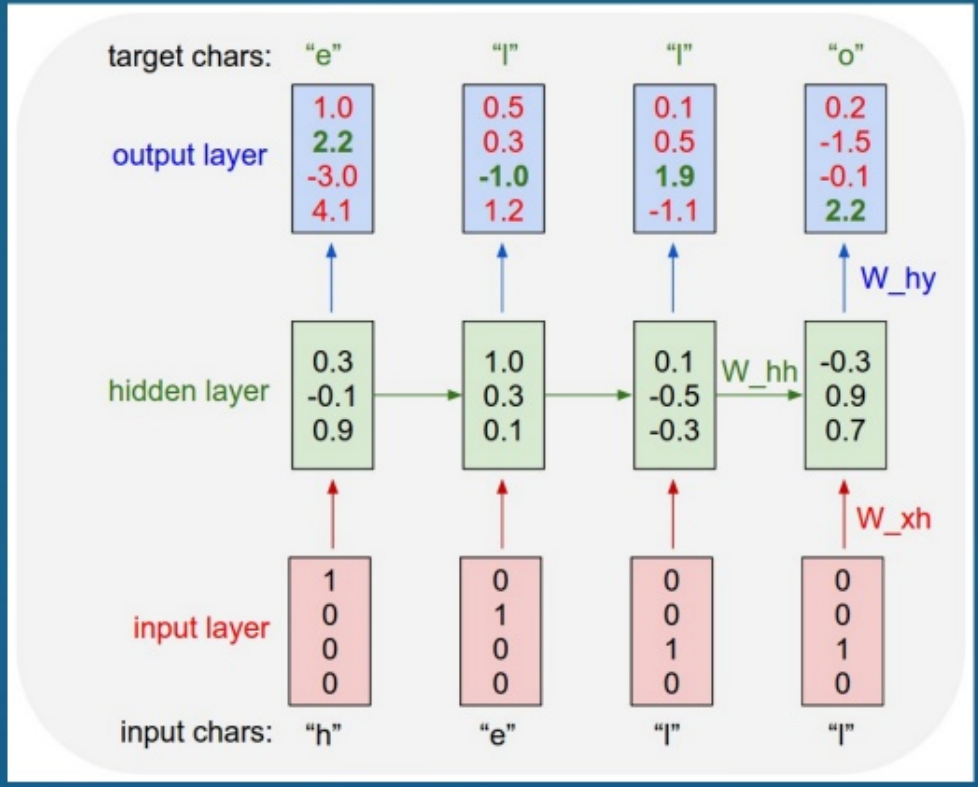
Text generation using LSTM

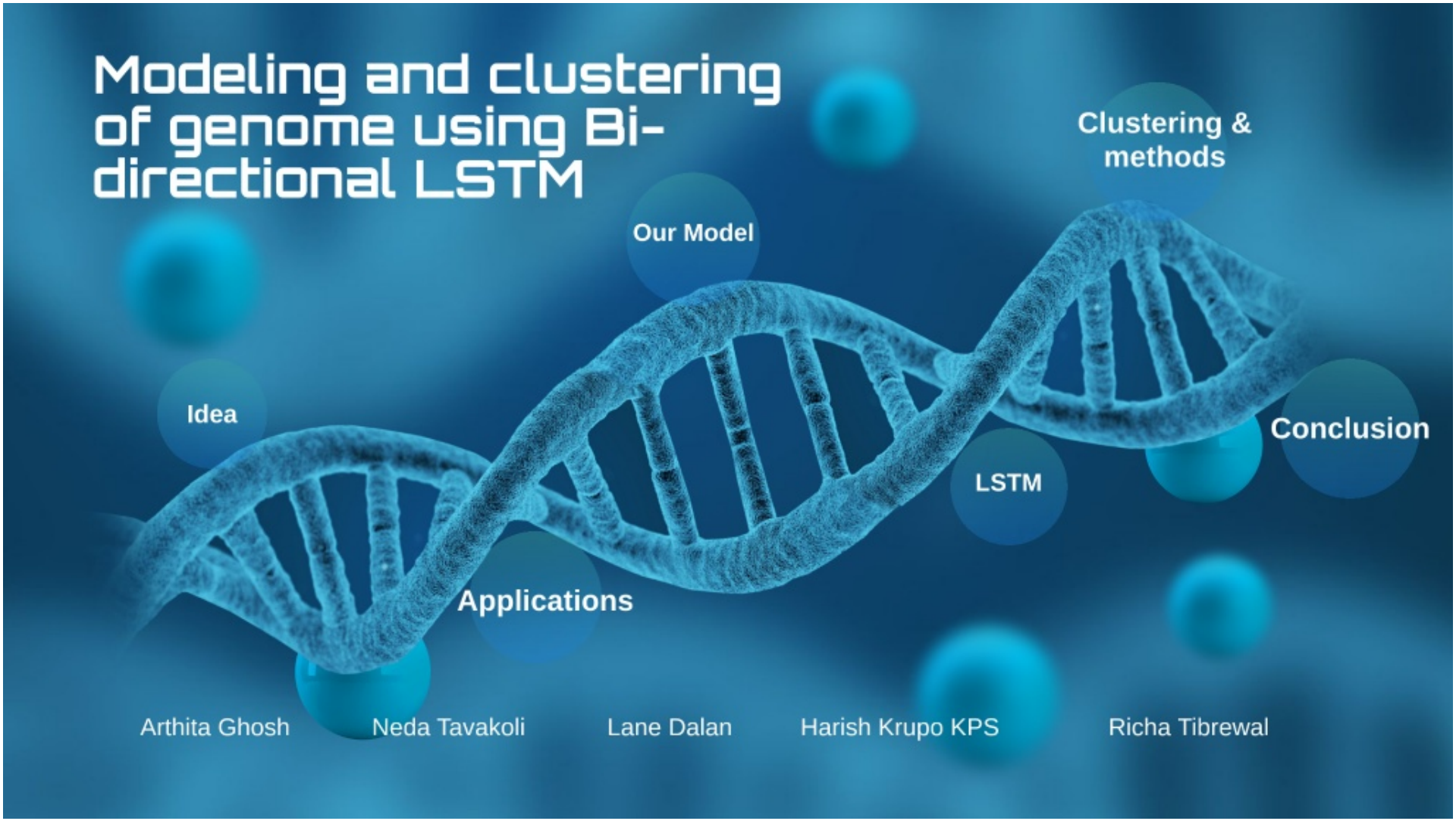
Prediction:

- Given an input string "Hel"
- Convert each character to its corresponding id
- Run each id through the model
- Once, completed start picking next character from predicted values

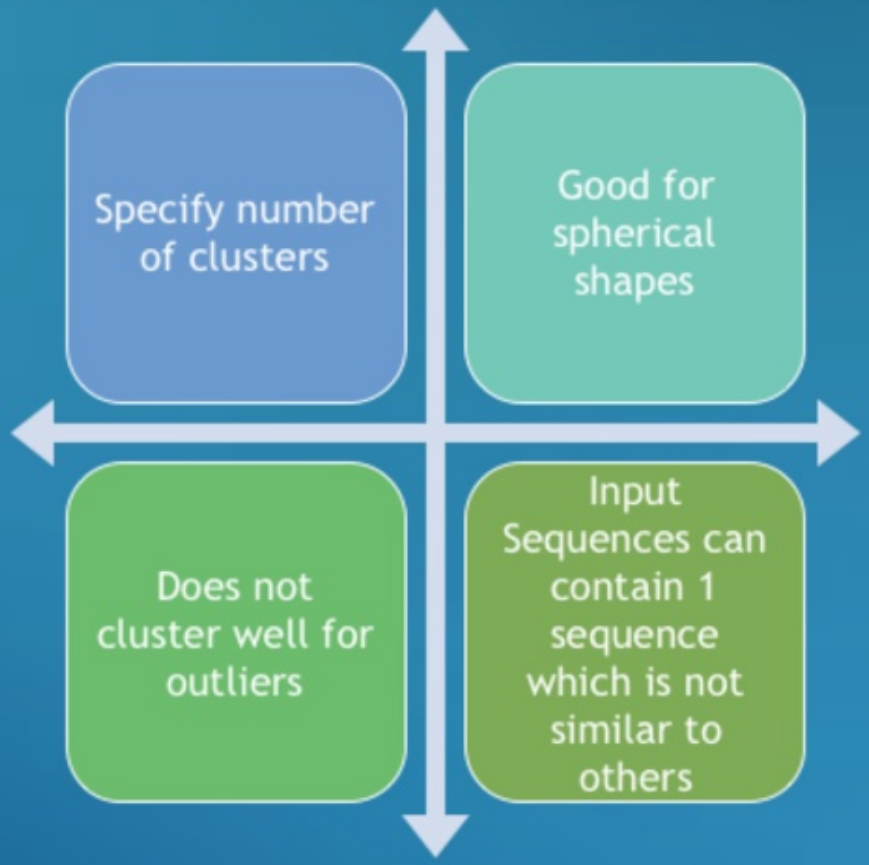
Text generation using LSTM

Training and prediction



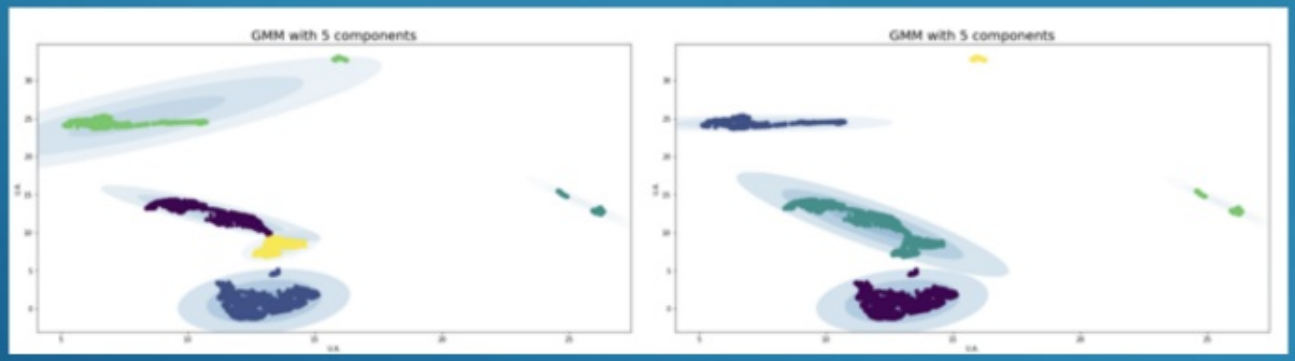

Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

K Means


Specify number of clusters
Good for spherical shapes
Does not cluster well for outliers
Input Sequences can contain 1 sequence which is not similar to others

GMM

- Takes into account elliptical shapes
- Central Limit Theorem
- Different initialization points may lead to different configurations

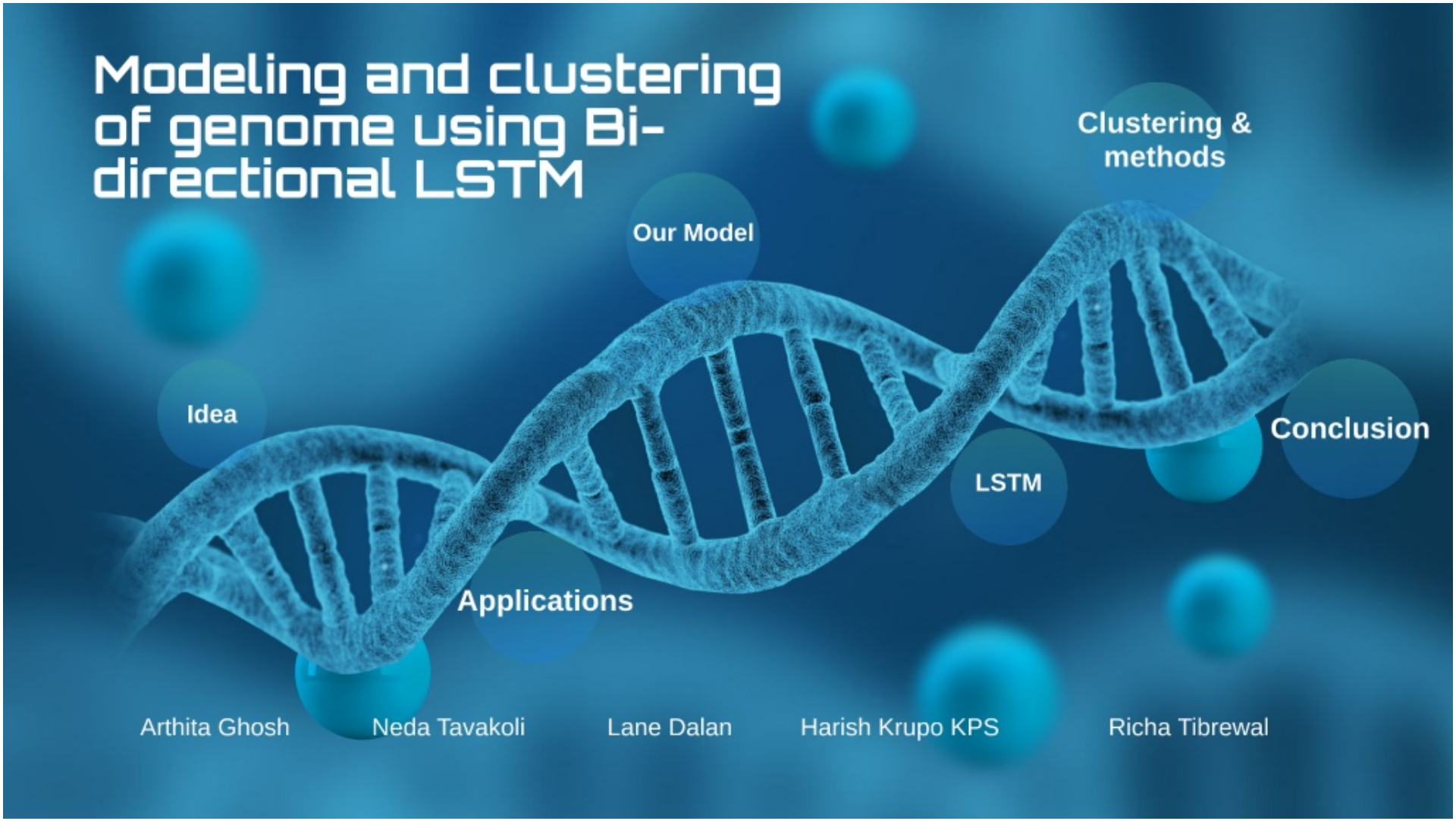



DBSCAN
Finds association and structures in data that are hard to find manually
Find patterns
Predict trends
Works well with outliers and noise

Preliminary Results

Clustering Method	Number of clusters	Silhouette Score
K Means	50	0.198
	2	0.932
	10	0.210
	5	0.579
GMM	2	0.935
DBSCAN	2	0.935

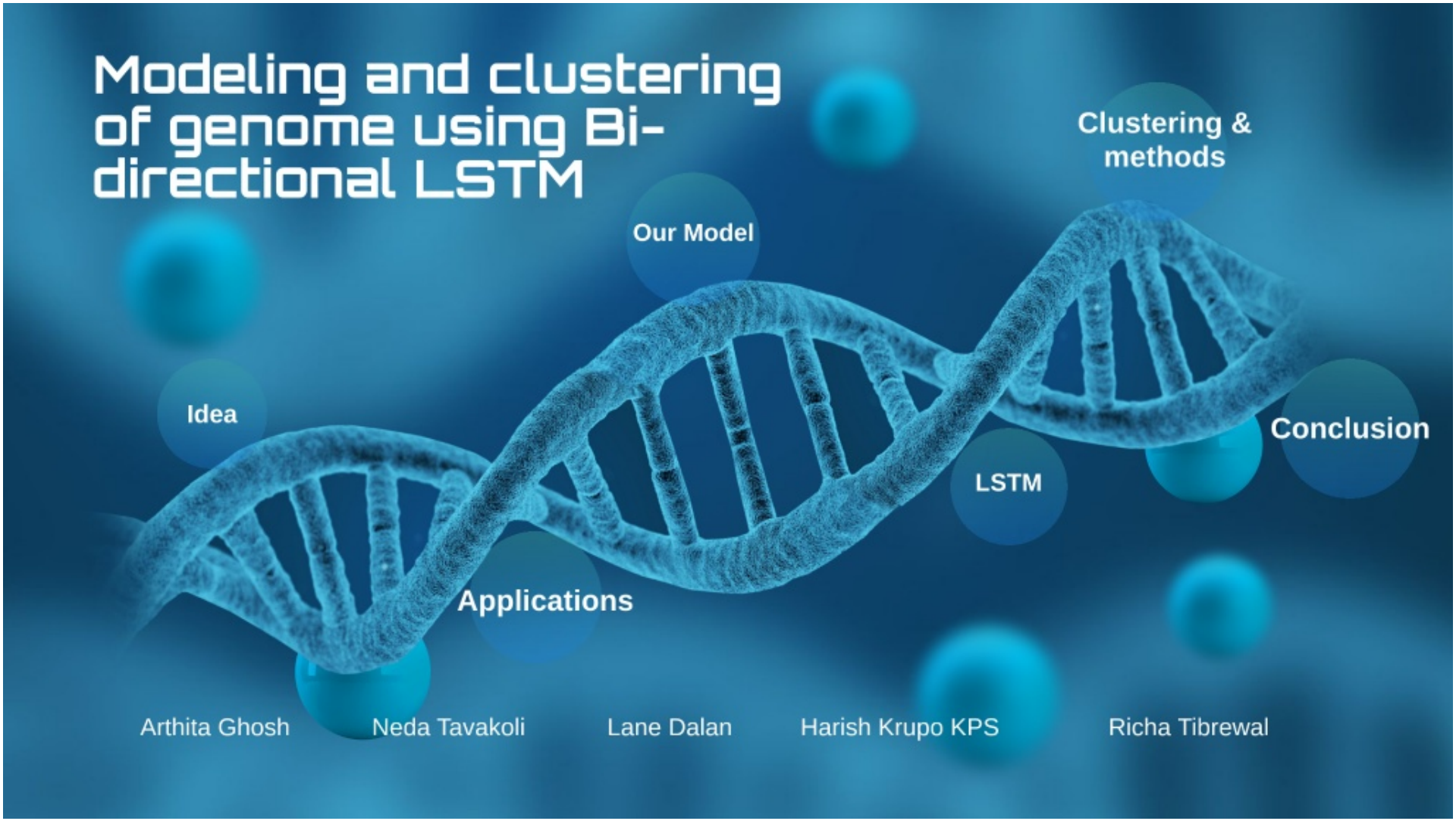
- 207 Query Sequences
- Each sequence has 80 characters


Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal

Conclusion

- Traditional methods for genome clustering are computationaly expensive.
- Utilize bi-directional LSTM to model and cluster genome data more efficiently

Thank You !!


Modeling and clustering of genome using Bi-directional LSTM
Clustering & methods
Our Model
Idea
Conclusion
LSTM
Applications
Arthita Ghosh
Neda Tavakoli
Lane Dalan
Harish Krupo KPS
Richa Tibrewal