RAJASTHAN HACKATHON 3.0, TRACK: **BIO-INFORMATICS**

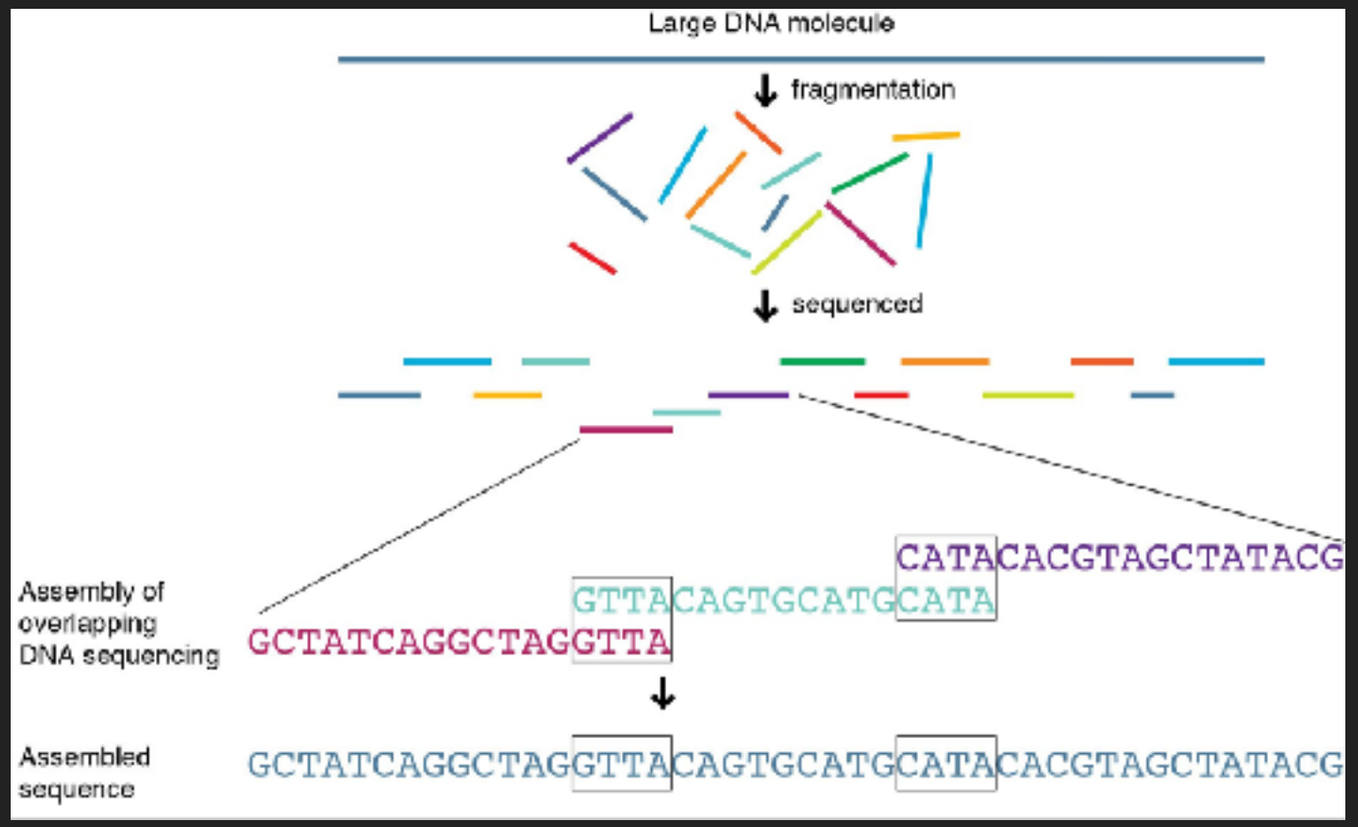
A DISTRIBUTED COMPUTING BASED GENOME SEQUENCING SOLUTION

HUMANS-FOR-HUMANS

WHAT IS GENOME SEQUENCING (PROBLEM STATEMENT)

WHAT?

- Genome sequencing is a computationally intensive and expensive technique of sequencing the genotype of a human.
- A single sequence costs over 1000$ for sequencing and takes days for processing.



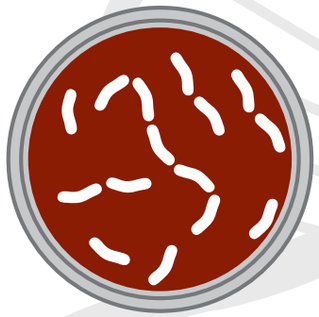
WHY?

- **Genome Sequencing can detect early on the genetic disorders in cardiovascular and tumour diseases.** Hence it is important to make it computationally faster and cheaper for widespread use. 'Humans for Humans' does that exactly leveraging the power of distributed volunteer computing.

The Whole Genome Sequencing (WGS) Process

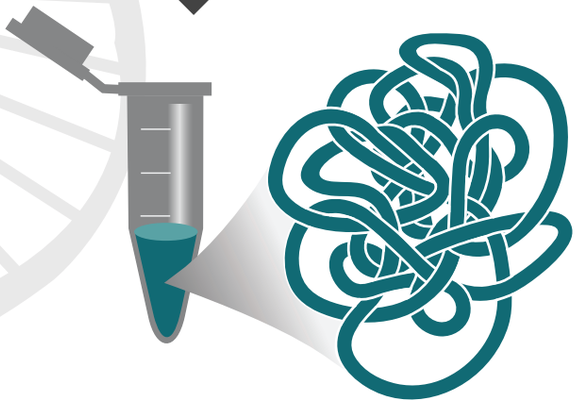
WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

Bacterial Culture



1. DNA Extraction

1 Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.



2. DNA Shearing

2 DNA is cut into short fragments of known length, either by using enzymes "molecular scissors" or mechanical disruption.

3. DNA Library Preparation

3 Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a "DNA library."

4. DNA Library Sequencing

4 The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a "DNA read."

5. DNA Sequence Analysis

5 The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.

WHY IS THE PROCESS COMPUTATIONALLY EXPENSIVE

PHRED QUALITY SCORE

Calculate the probability of the occurrence of a sequence of genomes. There exists millions of probable sequence, out of which the sequence having the maximum accuracy (Avg. PHRED SCORE of 20 i.e. 99%) is the chosen sequence.

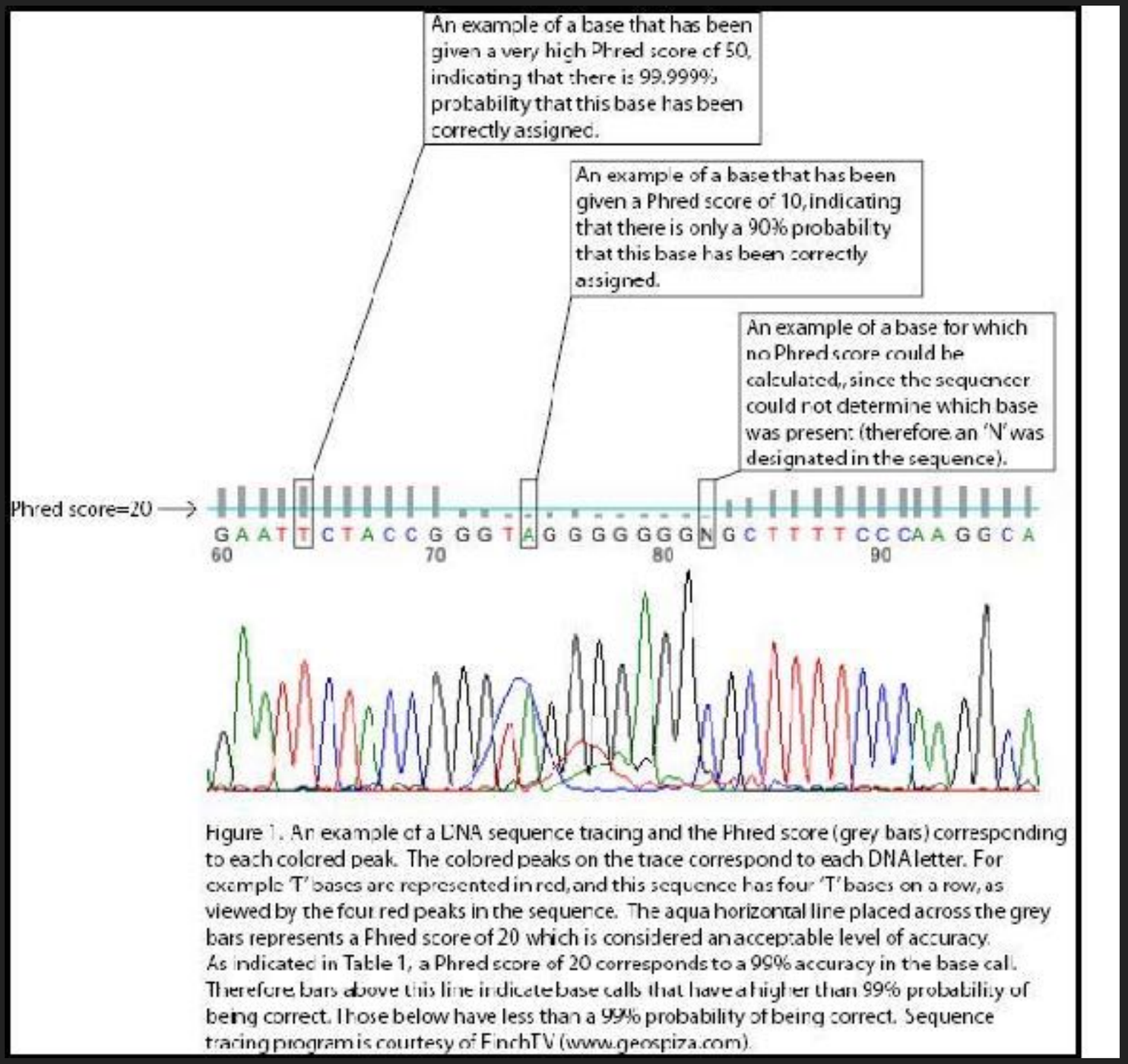
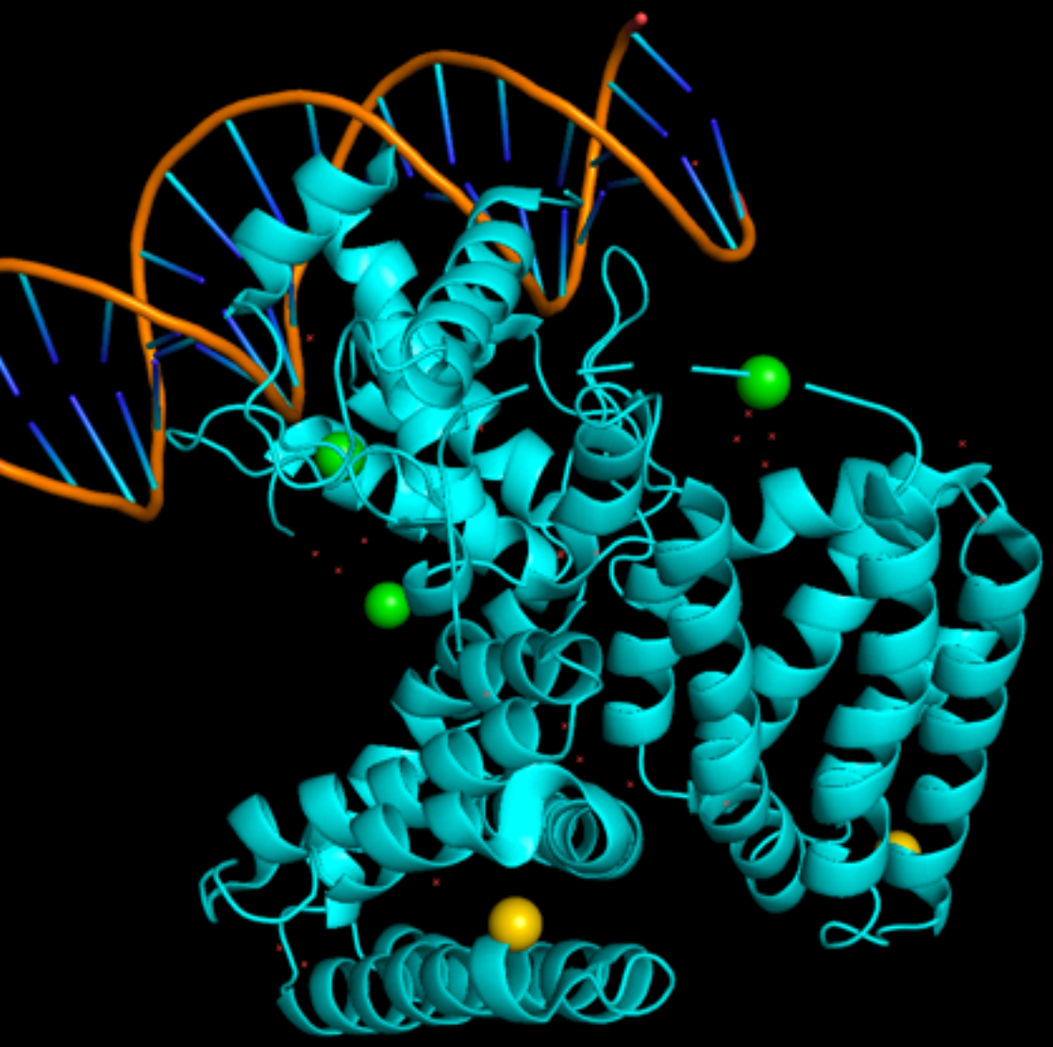


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

SOLUTION



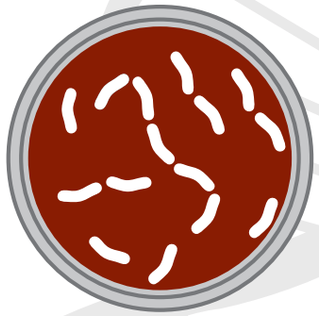
DISTRIBUTED VOLUNTEER BASED CLOUD COMPUTING SOLUTION FOR REDUCING COMPUTATION TIME AND EXPENSES.

HUMANS-FOR-HUMANS

The Whole Genome Sequencing (WGS) Process

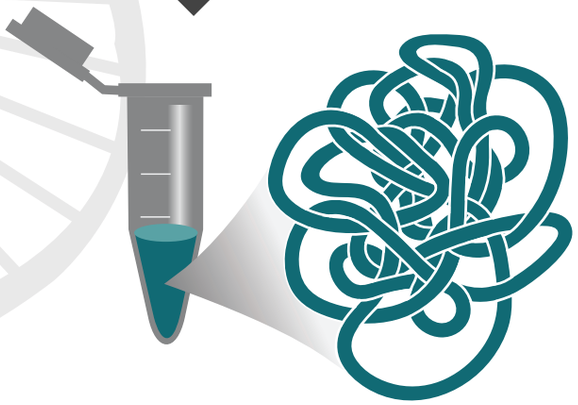
WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

Bacterial Culture



1. DNA Extraction

1 Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.



2. DNA Shearing

2 DNA is cut into short fragments of known length, either by using enzymes “molecular scissors” or mechanical disruption.

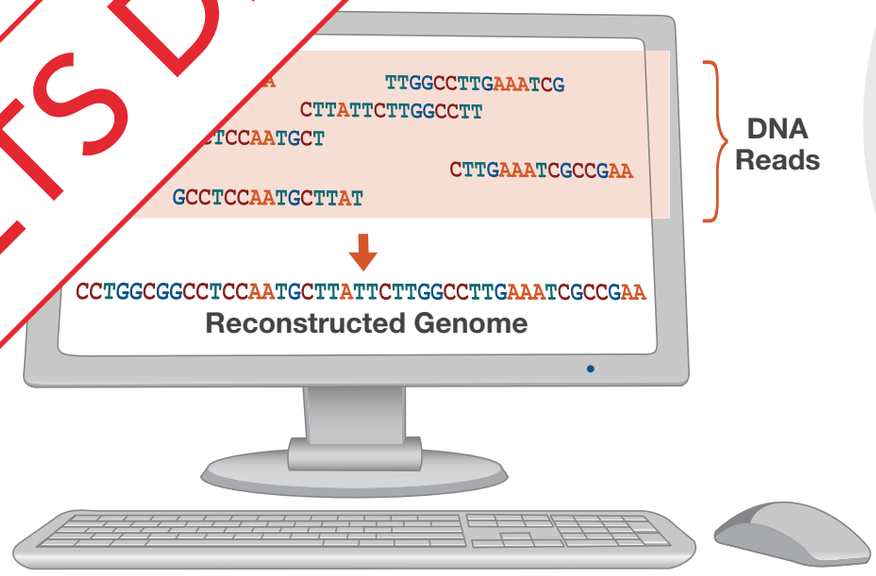
3. DNA Library Preparation

3 Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a “DNA library.”

4. DNA Library Sequencing

4 The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a “DNA read.”

5. D…



5 The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.

GETS DISTRIBUTED

HOW?

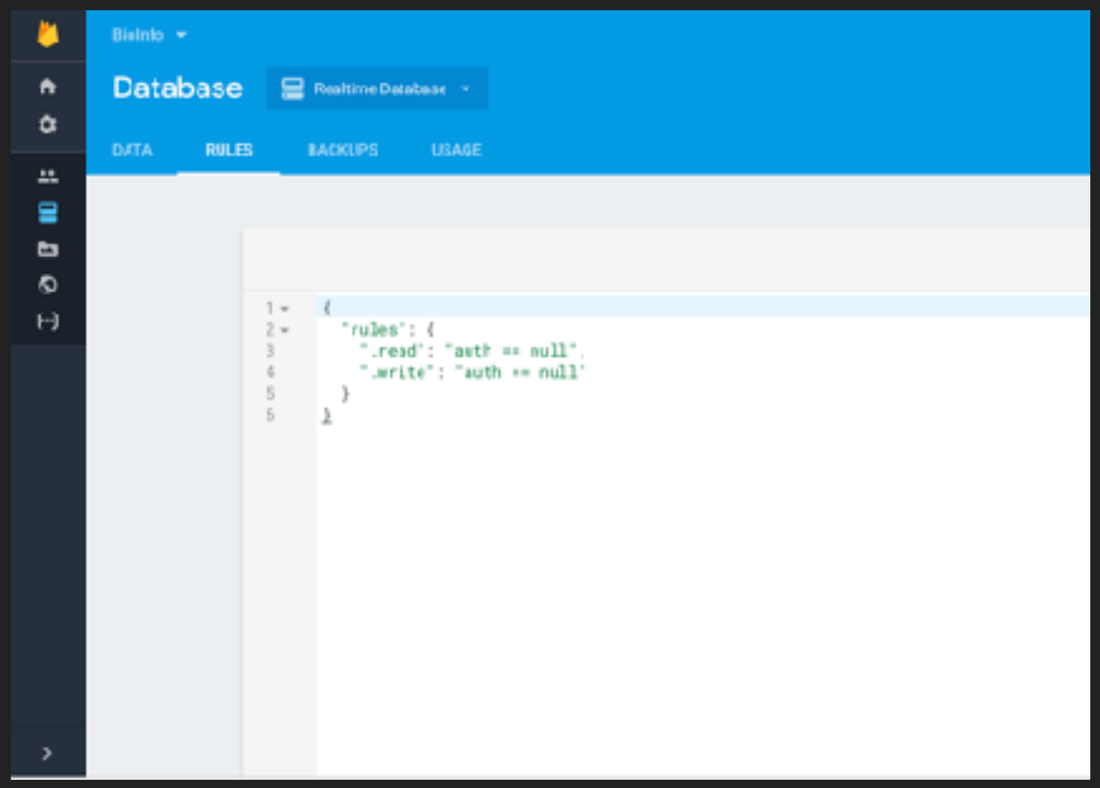
DISTRIBUTED

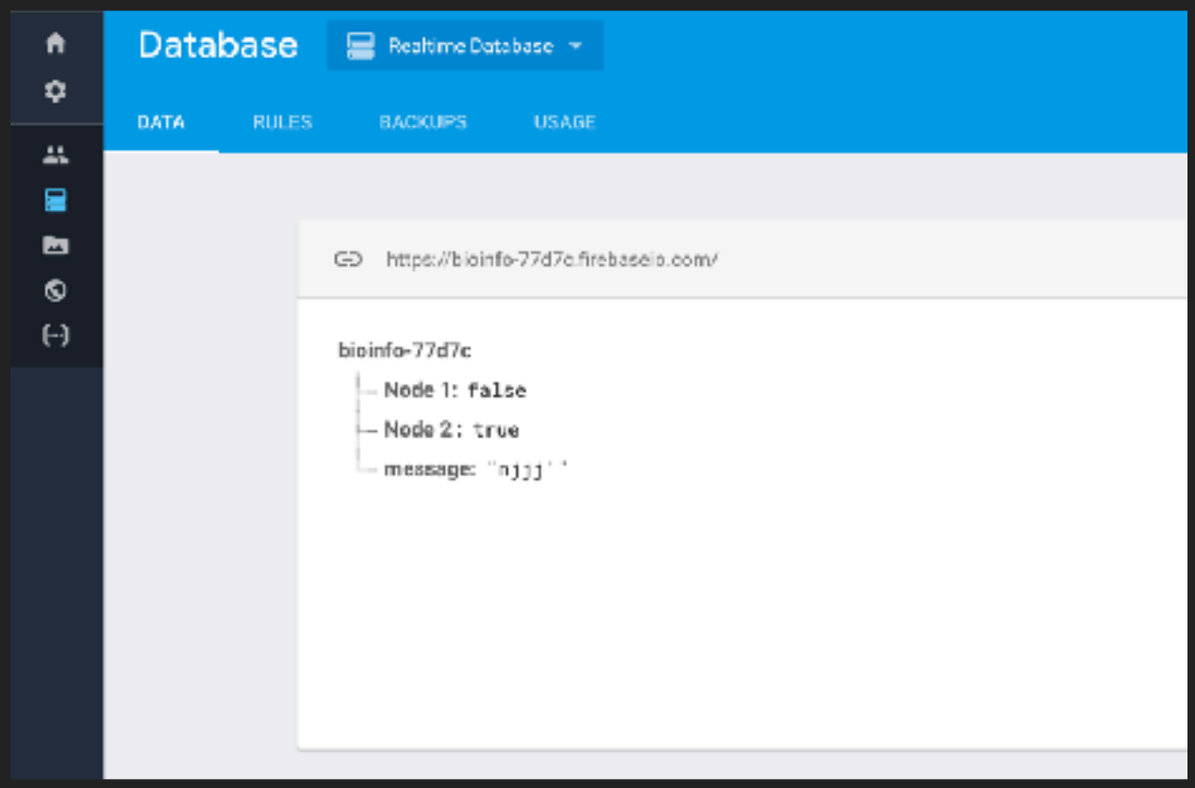
Networks as message flows

- Nodes interact via a *network*
  - Humans interact via spoken words
  - Particles interact via fields
  - Computers (nodes) interact via IP, SCTP
- We model those interactions as **discrete *messages* sent between nodes**
- **Messages take *time* to propagate**
  - This is the **"slow"** part of the distributed system
  - We call this "latency"
- **Messages can often be lost**
  - This is another "unreliable" part of the distributed system
- **Network is rarely homogenous**
  - Some links slower/smaller/more-likely-to-fail than others

HUMANS-FOR-HUMANS, A DISTRIBUTED VOLUNTEER COMPUTING BASED GENOME SEQUENCING SOLUTION

CLOUD COMPUTING





NODES AVAILABLE TO THE MAIN SERVER
(MASTER - SLAVE SYSTEM)

VOLUNTEER BASED

- The system is deployed as volunteer based. All phones which download the app.
- Once they allow the server and become an active node, the data is processed and returned back to the server.

LEARNING NETWORK

- The system eventually learns which node (mobile system) is more reliable.
- Designed as a binary classifier, with number of failures per month, joining time/ day and efficiency (computations / time)

RAJASTHAN HACKATHON 3.0, TRACK: **BIO-INFORMATICS**

A DISTRIBUTED COMPUTING BASED GENOME SEQUENCING SOLUTION

HUMANS-FOR-HUMANS