{GraphConnect NYC}

Hadoop and Graph Databases (Neo4j): Winning Combination for Bioinformatics

Jonathan Freeman
@freethejazz

Jonathan Freeman
@freethejazz

Open Software Integrators

- Founded January 2008 by Andrew C. Oliver
  - Durham, NC

Revenue and staff has at least doubled every year since 2009.

- New office (2012) in Chicago, IL
  - We're hiring associate to senior level as well as UI Developers (JQuery, Javascript, HTML, CSS)
  - Up to 50% travel (probably less), salary + bonus, 401k, health, etc etc
  - Preferred: Java, Tomcat, JBoss, Hibernate, Spring, RDBMS, JQuery
  - Nice to have: Hadoop, Neo4j, MongoDB, Ruby a/o at least one Cloud platform

Questions to answer

Jonathan Freeman
@freethejazz

- uhh, bioinformatics?
- What is Hadoop? Why is it a good fit?
- And Neo4j? Why the combination?
- I want this now! How do I do it?!?!

{Hadoop + Neo4j = Bioinformatics Win}

Bioinformatics

Jonathan Freeman
@freethejazz

“
dynamic

information processing

system
”

Jonathan Freeman
@freethejazz

Life

http://www.labtimes.org/labtimes/issues/lt2011/lt07/lt_2011_07_26_29.pdf

Jonathan Freeman
@freethejazz

- Storing/Retrieving Biological Data
- Organizing Biological Data
- Analyzing Biological Data

Jonathan Freeman
@freethejazz

Biological Data

- amino acid sequences
- nucleotide sequences
- protein structures

Jonathan Freeman
@freethejazz

- Genetic sequence analysis
- Tracing biological evolution
- Analysis of gene expression
- Studying mutations in cancer
- Predicting protein structure and function
- Molecular Interaction

Jonathan Freeman
@freethejazz

- Genetic sequence analysis
- Tracing biological evolution
- Analysis of gene expression
- Studying mutations in cancer
- Predicting protein structure and function
- Molecular Interaction

Jonathan Freeman
@freethejazz

Full Human Genome Sequencing Then

13 Years

$2,700,000,000

Jonathan Freeman
@freethejazz

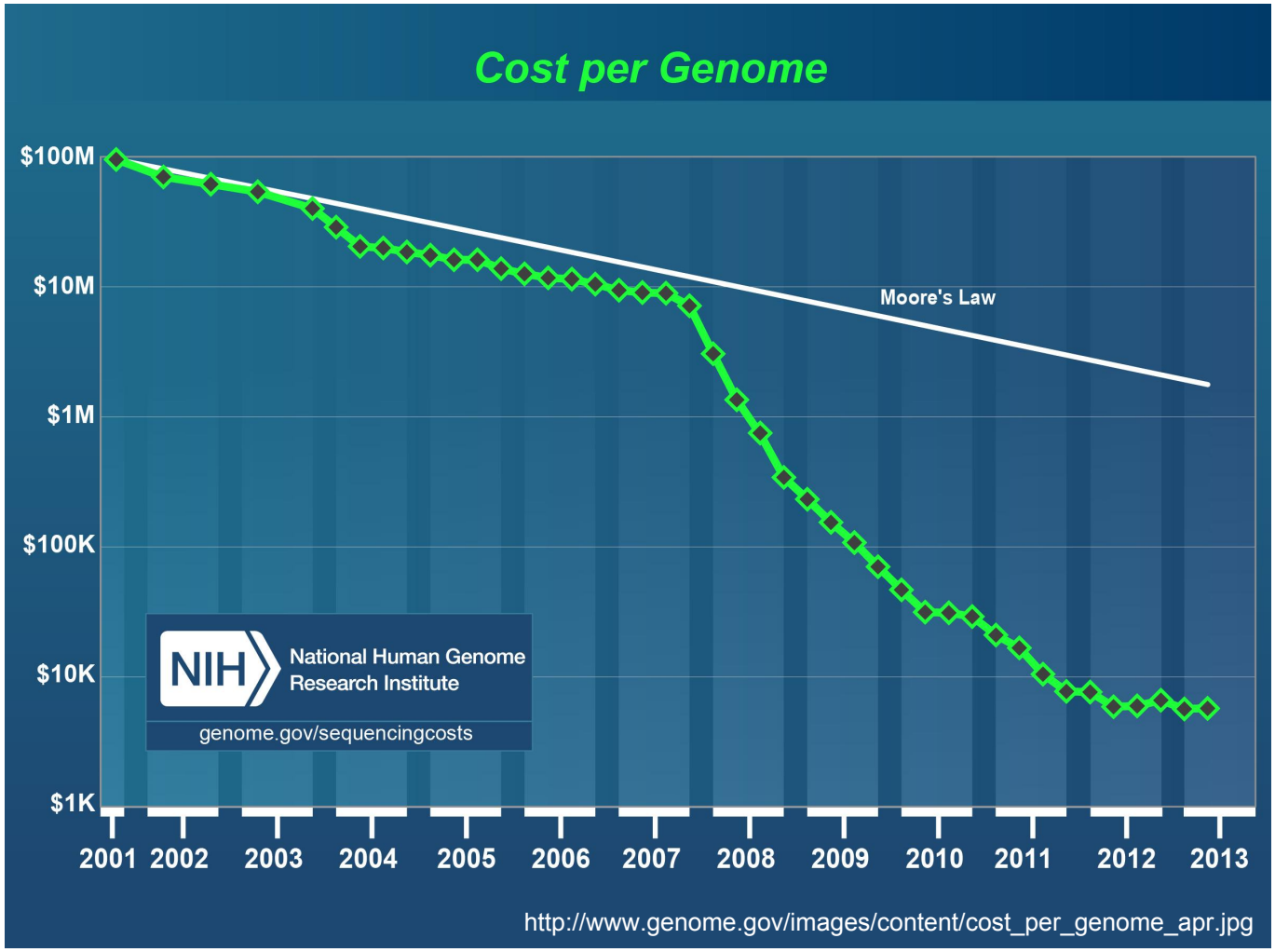
Full Human Genome Sequencing Then

1 Day

$5,000

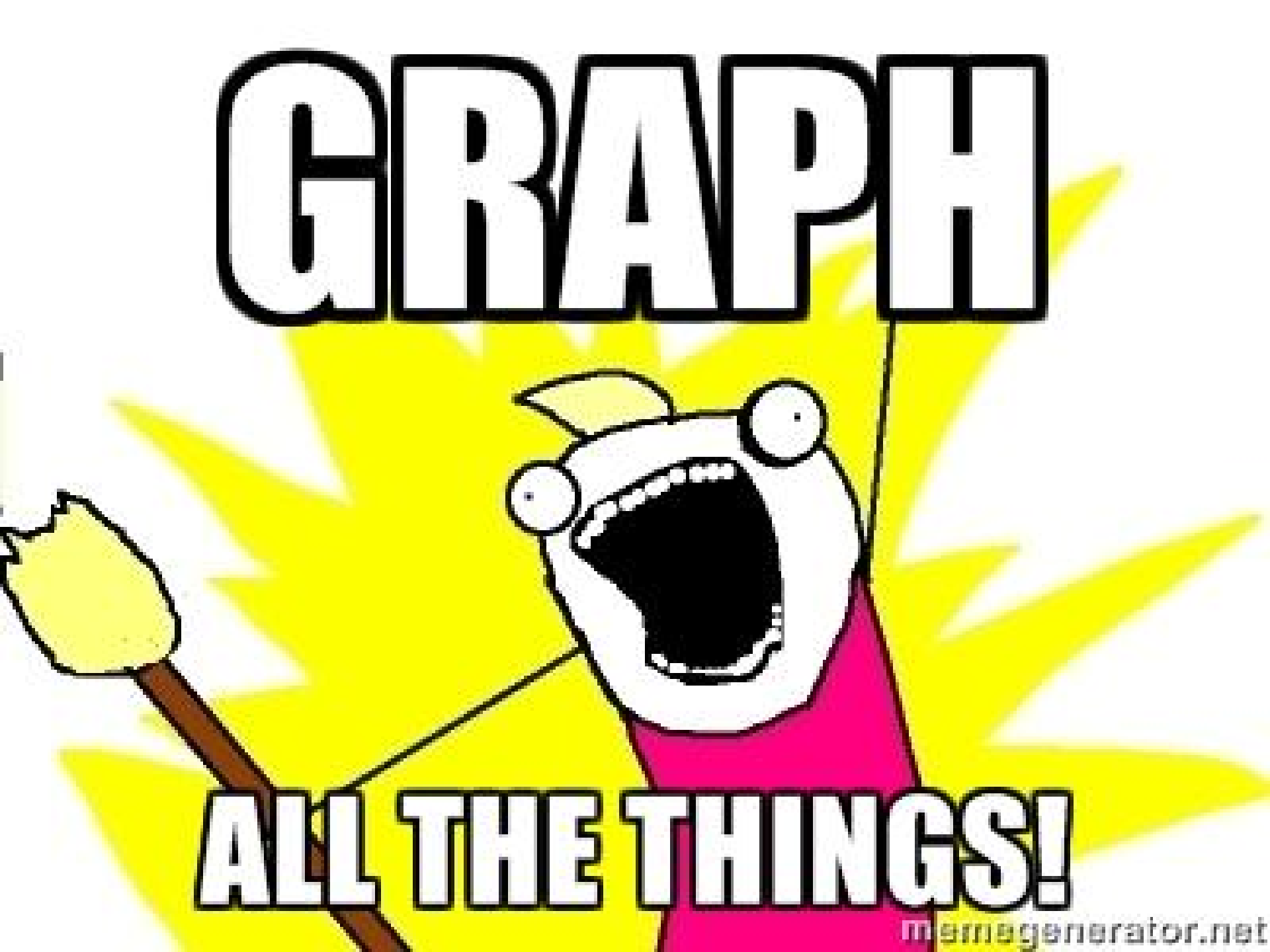
Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz



Jonathan Freeman
@freethejazz

So what are we waiting for?


GRAPH
ALL THE THINGS!
memegenerator.net

Jonathan Freeman
@freethejazz

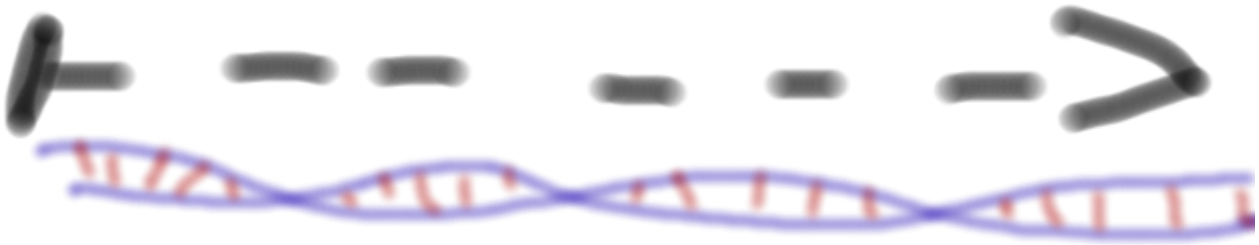
well, the thing about that…

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

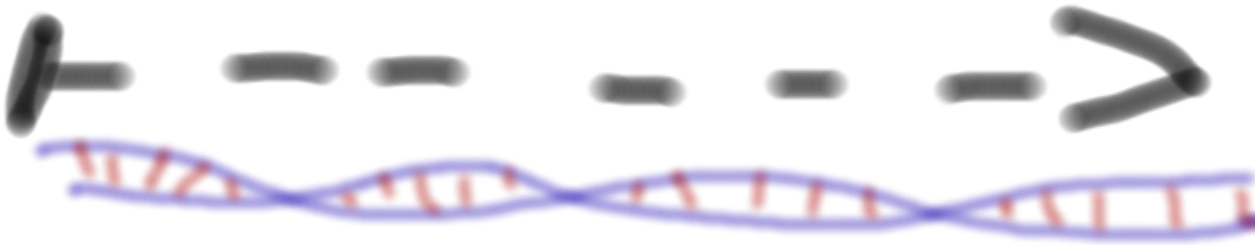


Jonathan Freeman
@freethejazz

...
ATTCCAGGAGTATTGACACCAT...

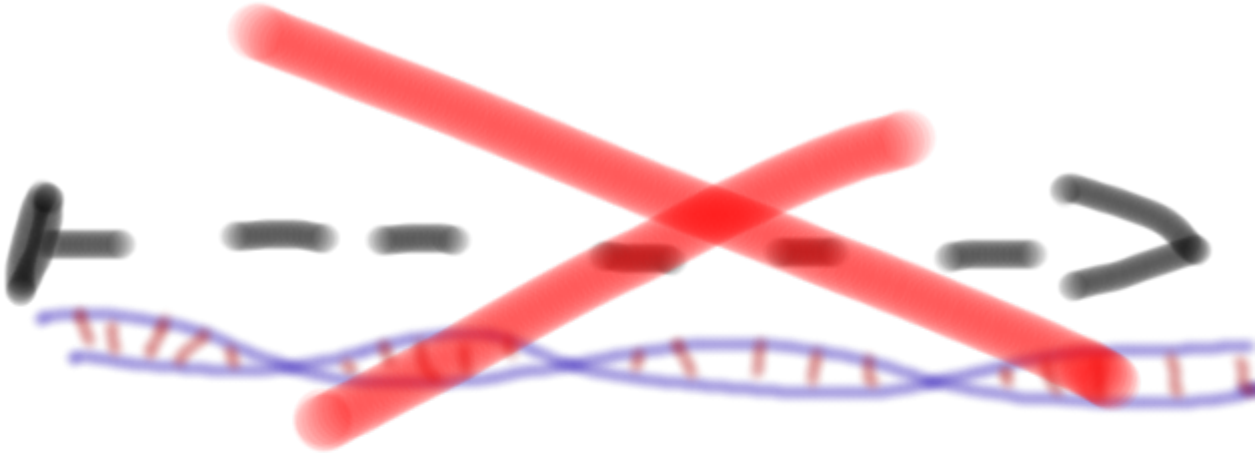
Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz



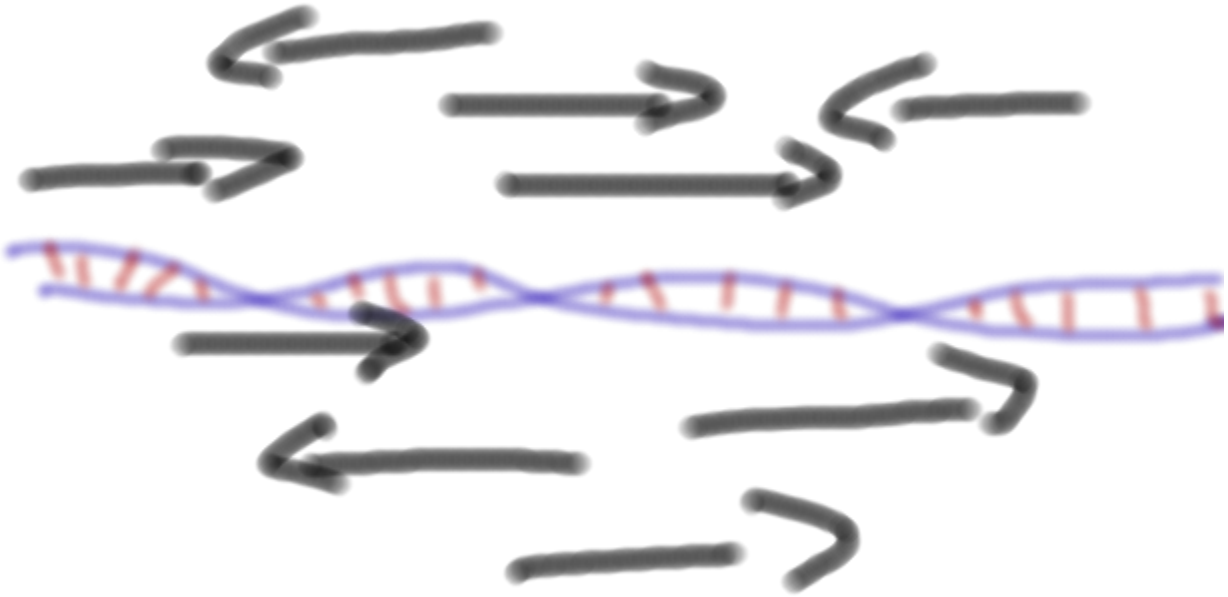
Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz



Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz



Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

AGGATTACCAGGA
CAAAGGATT
TTACCAGGATACCAG
TGACAA
AAGGATTAC
GATACCAGTA
CAAGGATT
GTGACAA

{Hadoop + Neo4j = Bioinformatics Win}

Hadoop

Jonathan Freeman
@freethejazz

Infrastructure for distributed computing

HDFS

A distributed file system.

MapReduce

An implementation of a programming model for processing very large data sets.

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

APACHE HBASE

Jonathan Freeman
@freethejazz

Jonathan Freeman
@freethejazz

Infrastructure for distributed computing

HDFS

A distributed file system.

MapReduce

An implementation of a programming model for processing very large data sets.

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz

AGGATTACCAGGA
CAAAGGATT
TTACCAGGATACCAG
TGACAA
AAGGATTAC
GATACCAGTA
CAAGGATT
GTGACAA

Jonathan Freeman
@freethejazz

...
ATTCCAGGAGTATTGACACCAT...

Jonathan Freeman
@freethejazz

1000 CPU hours

Jonathan Freeman
@freethejazz

3 hours
$85
OSS

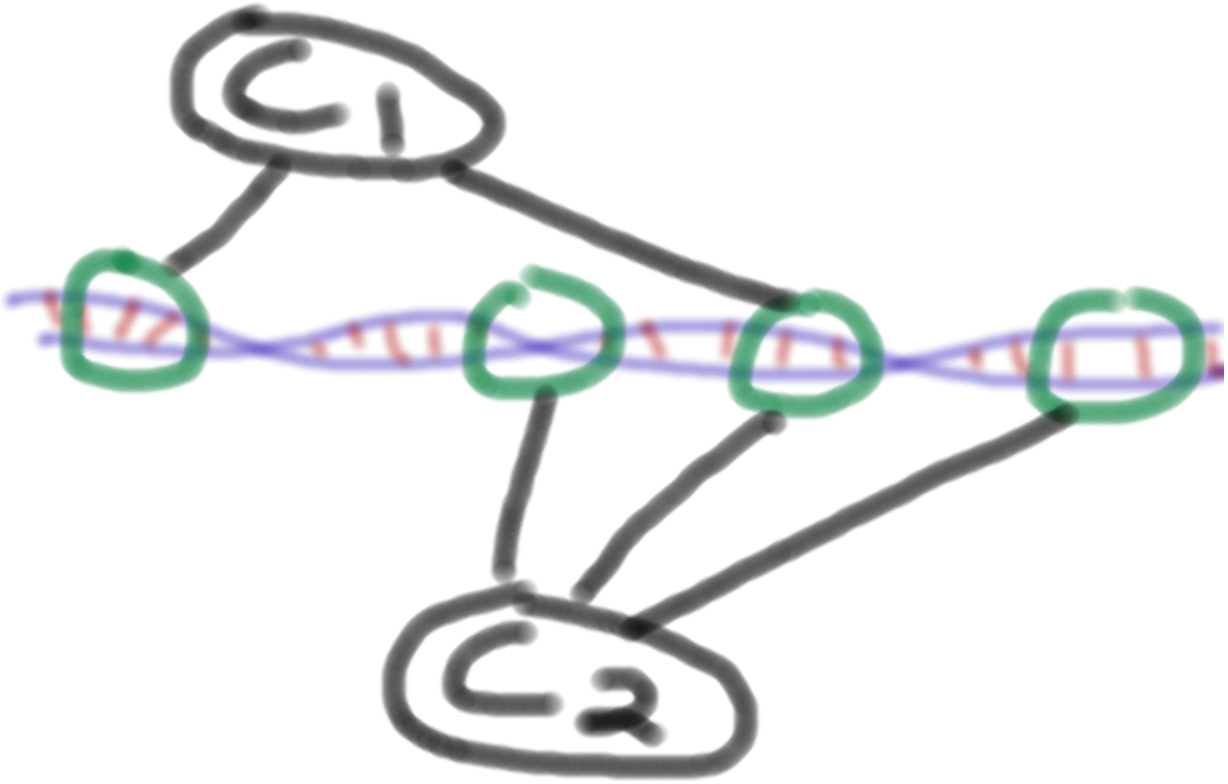
http://bowtie-bio.sourceforge.net/crossbow/index.shtml

{Hadoop + Neo4j = Bioinformatics Win}

And Neo4j?

Hadoop + Neo4j = Bioinformatics Win

Jonathan Freeman
@freethejazz



Jonathan Freeman
@freethejazz

```
MATCH (snp)<-[:INFLUENCED_BY]-(conditions)
WHERE snp.id = “rs1234”
RETURN conditions;
```

Jonathan Freeman
@freethejazz

```
MATCH (p)-[:GENOME_CONTAINS]->(snp)
      (snp)<-[:INFLUENCED_BY]-(conditions)
WHERE p.name = “Jonathan Freeman”
RETURN conditions;
```

Jonathan Freeman
@freethejazz

```
MATCH (p)-[:GENOME_CONTAINS]->(snp)
      (snp)<-[:INFLUENCED_BY]-(conditions)
WHERE c.name = “Parkinsons”
RETURN p;
```

{Hadoop + Neo4j = Bioinformatics Win}

How can I haz?!?!?!1

Jonathan Freeman
@freethejazz

Step 1: Get local copies

- Hadoop: http://www.neo4j.org/download
- Neo4j: http://hadoop.apache.org/releases.html#Download
- Batch Importer: https://github.com/jexp/batch-import

Jonathan Freeman
@freethejazz

Step 2: Familiarize yourself with the languages

- MapReduce: http://hadoop.apache.org/docs/r0.18.3/mapred_tutorial.html
- Pig: http://pig.apache.org/docs/r0.12.0/start.html
- Hive: https://cwiki.apache.org/confluence/display/Hive/GettingStarted

Jonathan Freeman
@freethejazz

Step 3: Find a dataset

- Typical starter data: http://www.gutenberg.org/
- Amazon's public data sets: http://aws.amazon.com/publicdatasets/

Jonathan Freeman
@freethejazz

Step 4: Start Playing!!!

Jonathan Freeman
@freethejazz

Step 5: Take Hadoop to the cloud

- http://aws.amazon.com/elasticmapreduce/

Jonathan Freeman
@freethejazz

Doing this in production?

http://blog.xebia.com/2012/11/13/combining-neo4j-and-hadoop-part-i/
http://blog.xebia.com/2013/01/17/combining-neo4j-and-hadoop-part-ii/

{Hadoop + Neo4j = Bioinformatics Win}

Thank You

@freethejazz

Jonathan Freeman
@freethejazz

Image Attribution:

Sand Timer: *http://bit.ly/HyCAgy*

Money: *http://bit.ly/1e4lhS6*

Scraggly DNA drawings: Jonathan Freeman :)