cloudera

Petascale Analytics in Genomics with Hadoop

2 June 2016, Strata+Hadoop World, London

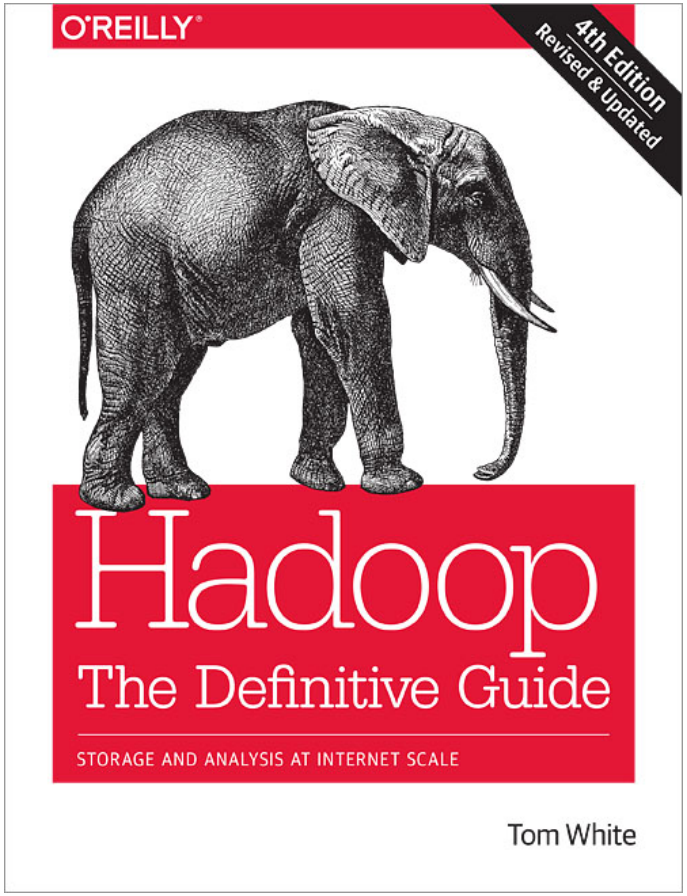
Tom White | @tom_e_white

About Me

- Data Science Team at Cloudera

- Apache Hadoop Committer, PMC Member, Apache Member

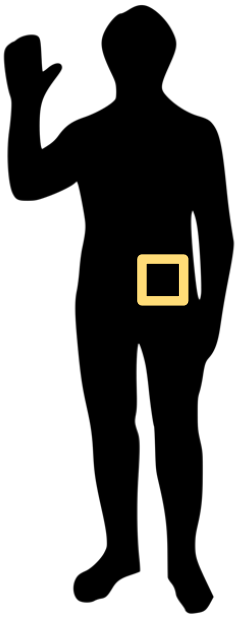
- Author of “Hadoop: The Definitive Guide”

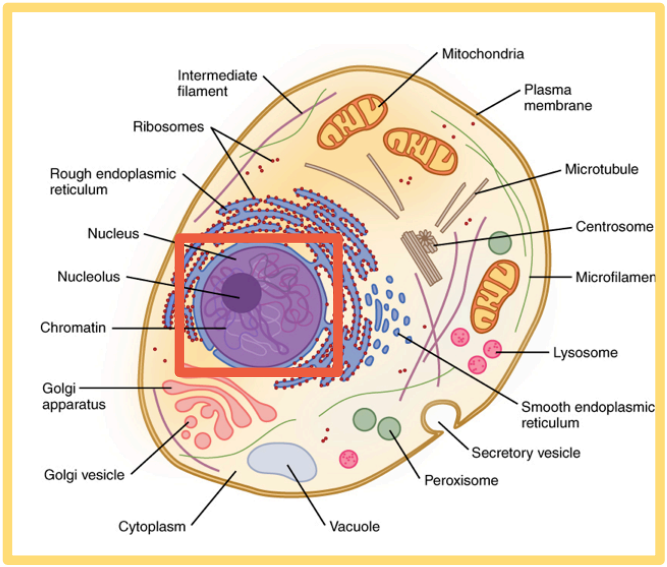


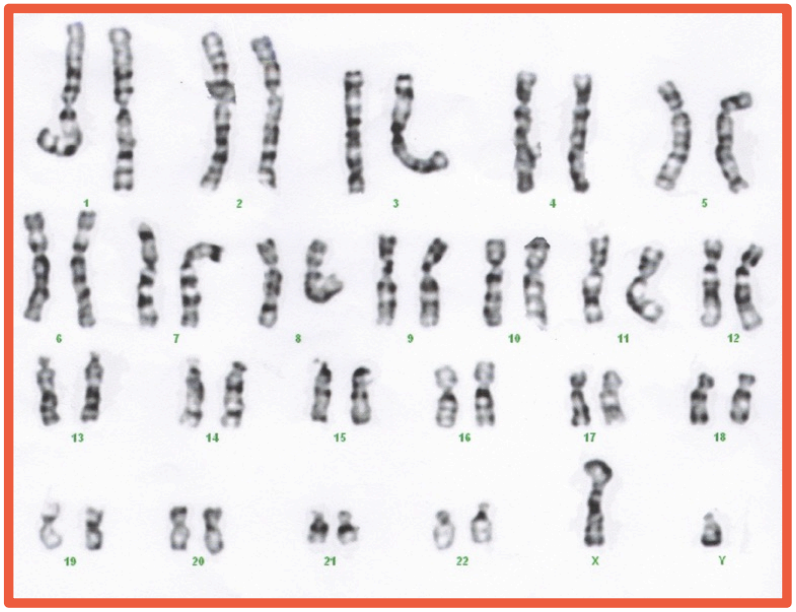
cloudera

What is **genomics**?

cloudera







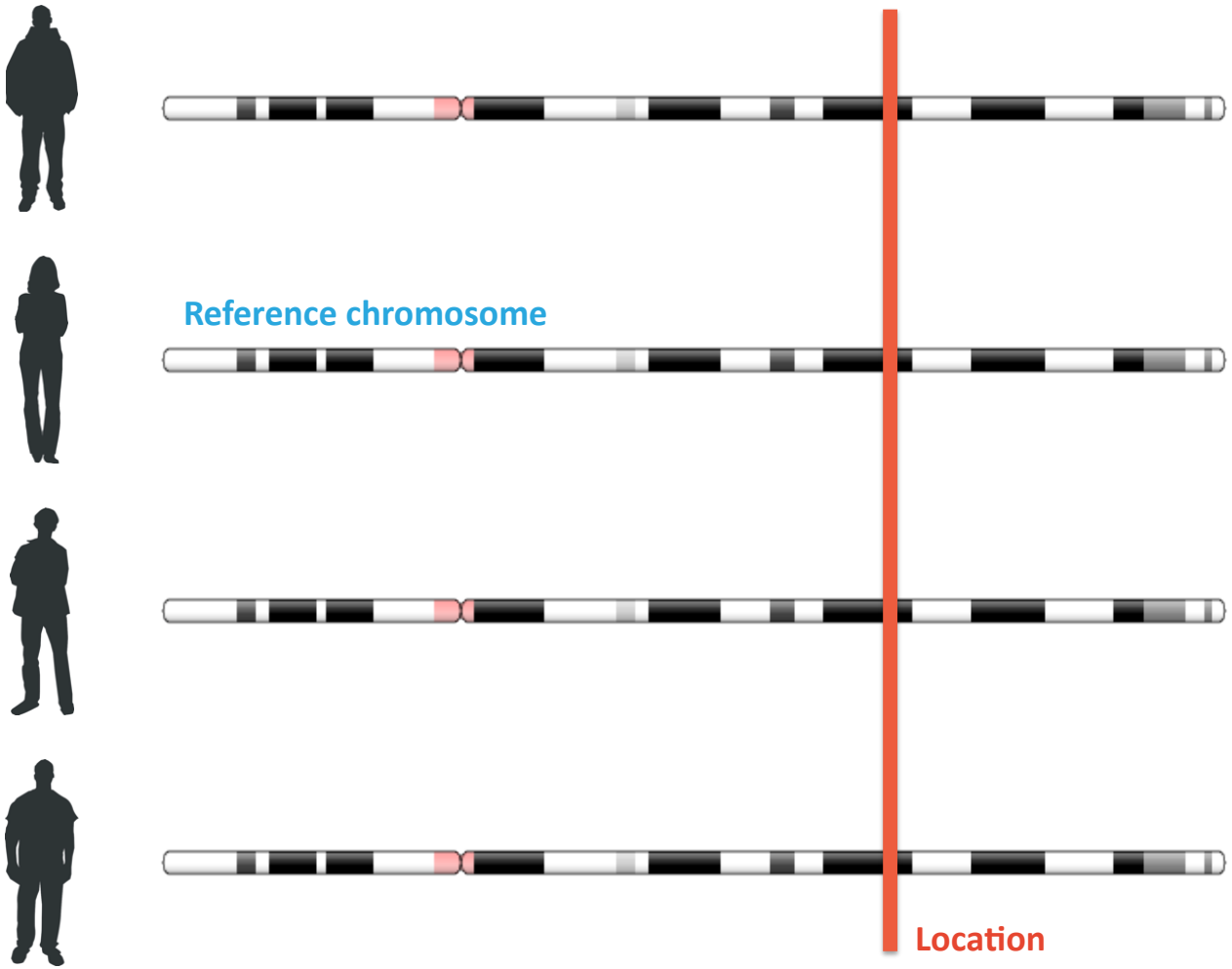
Organism

Cell

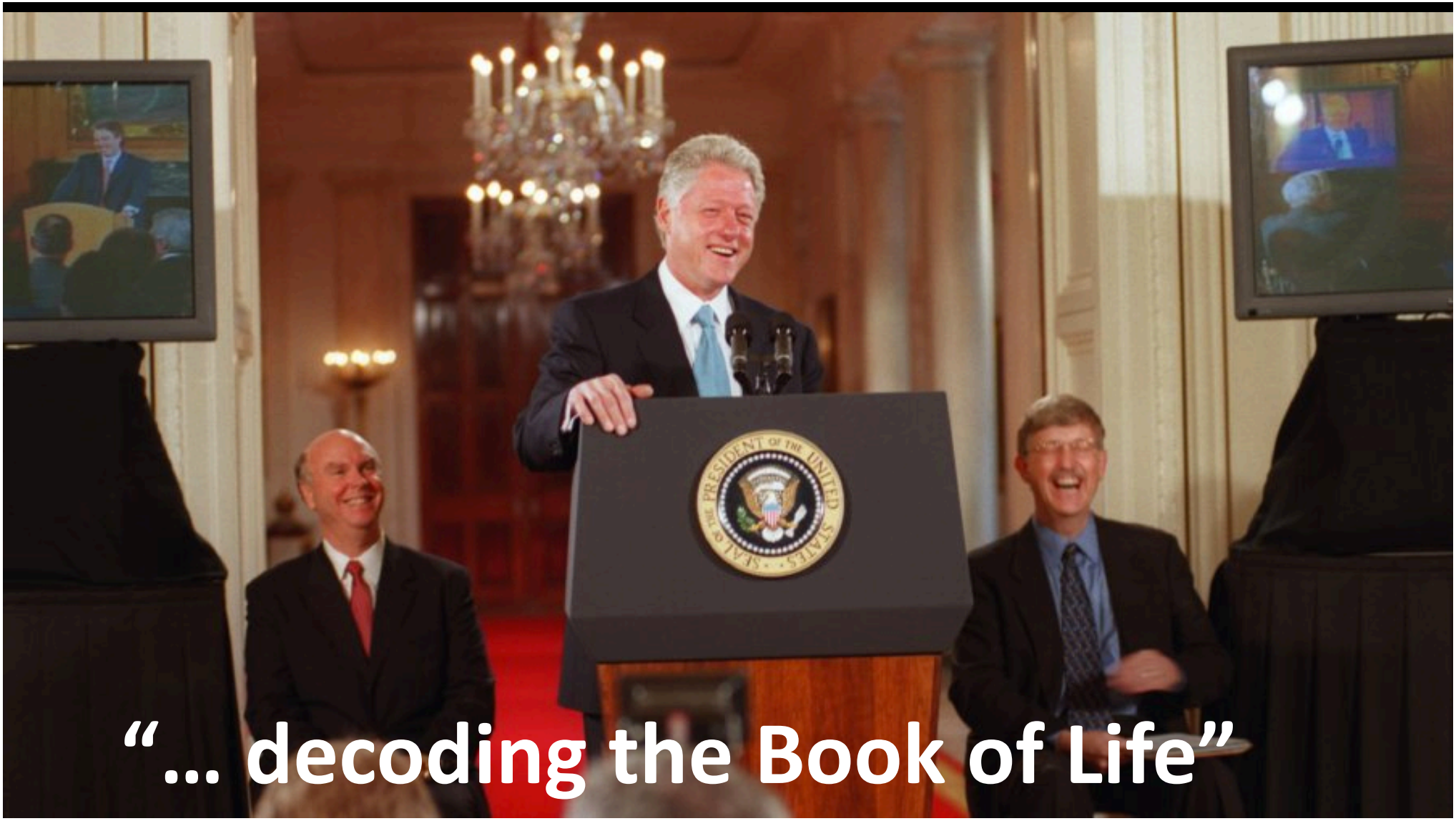
Genome

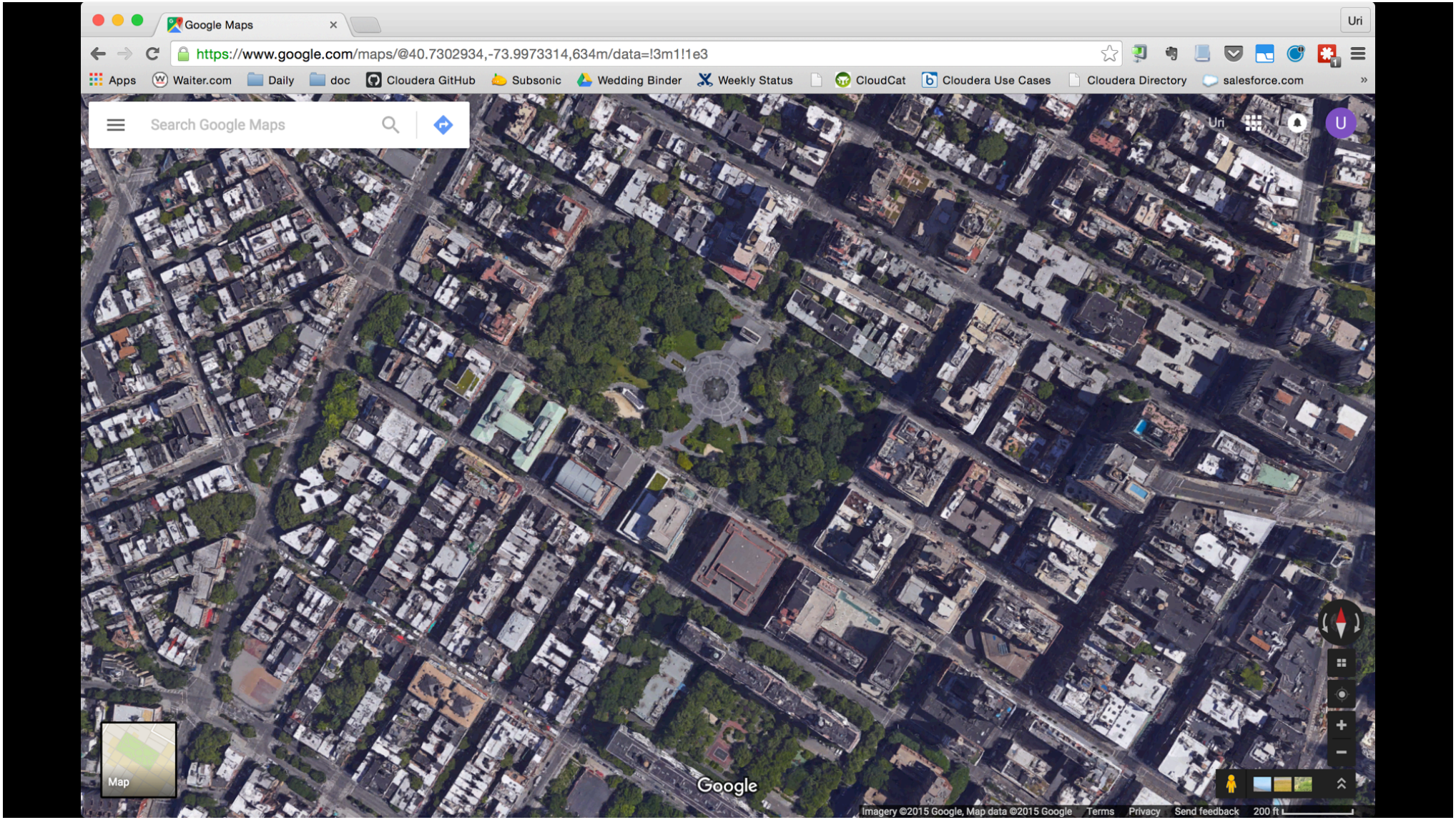
cloudera

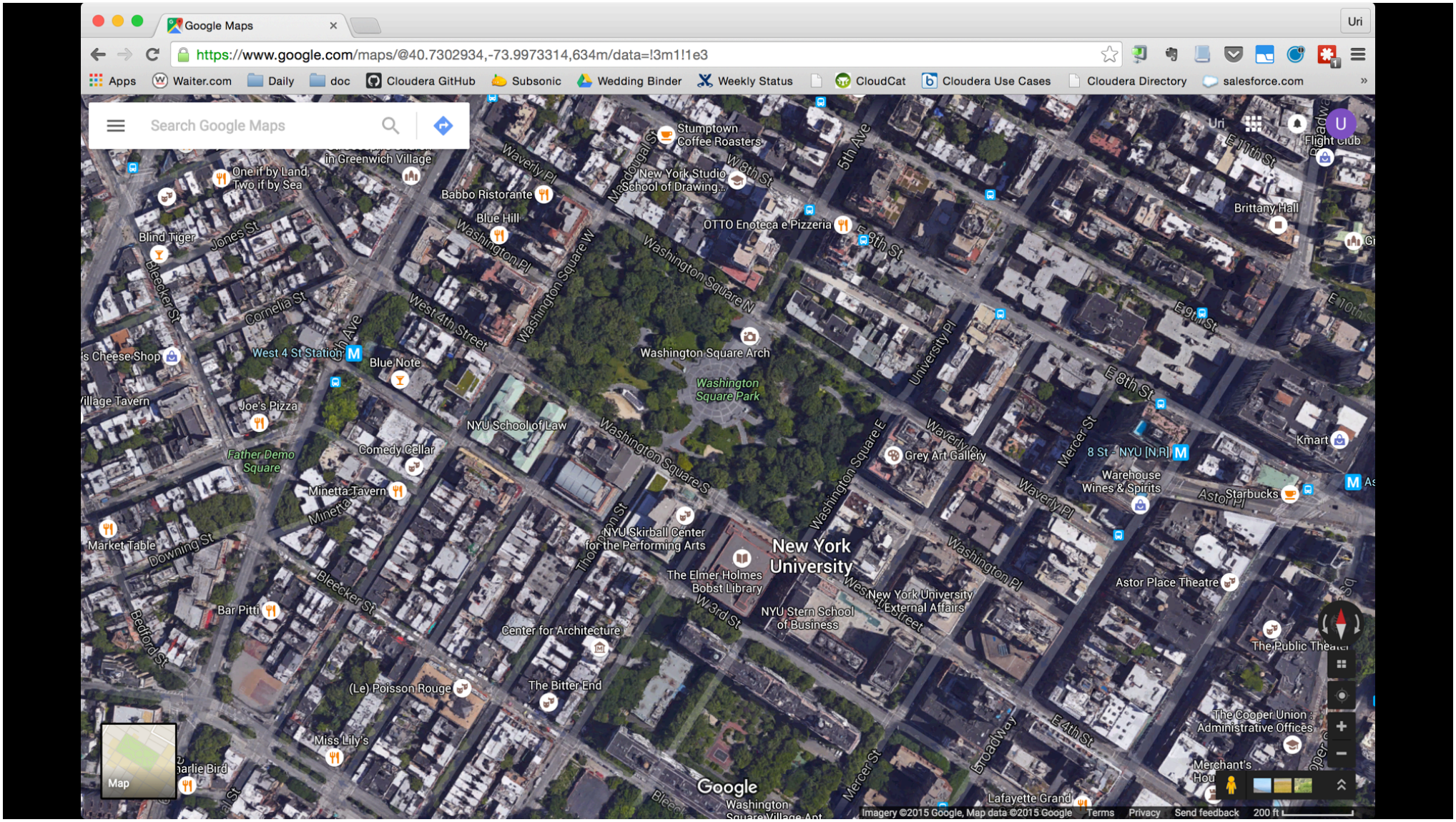
cloudera

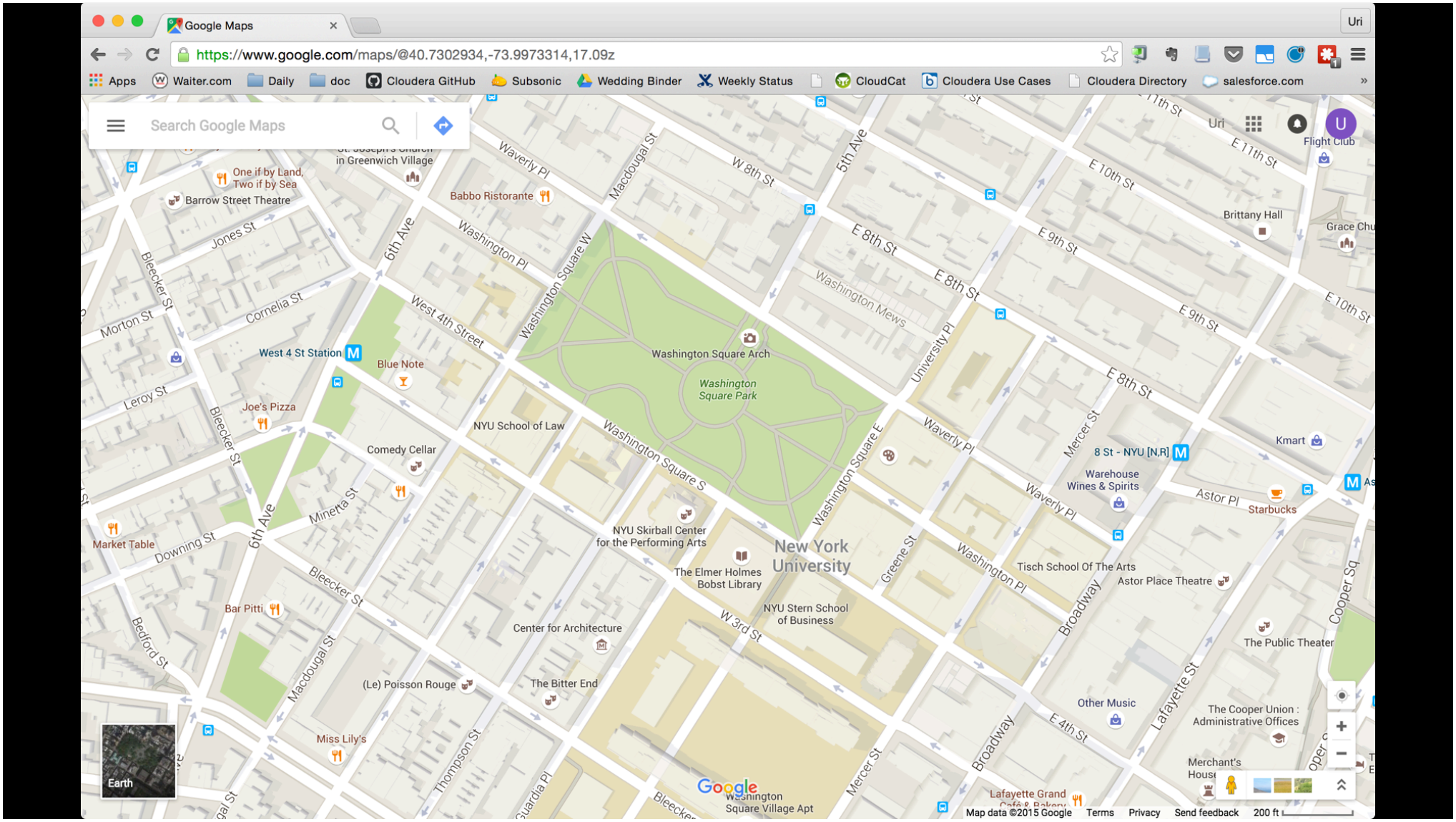


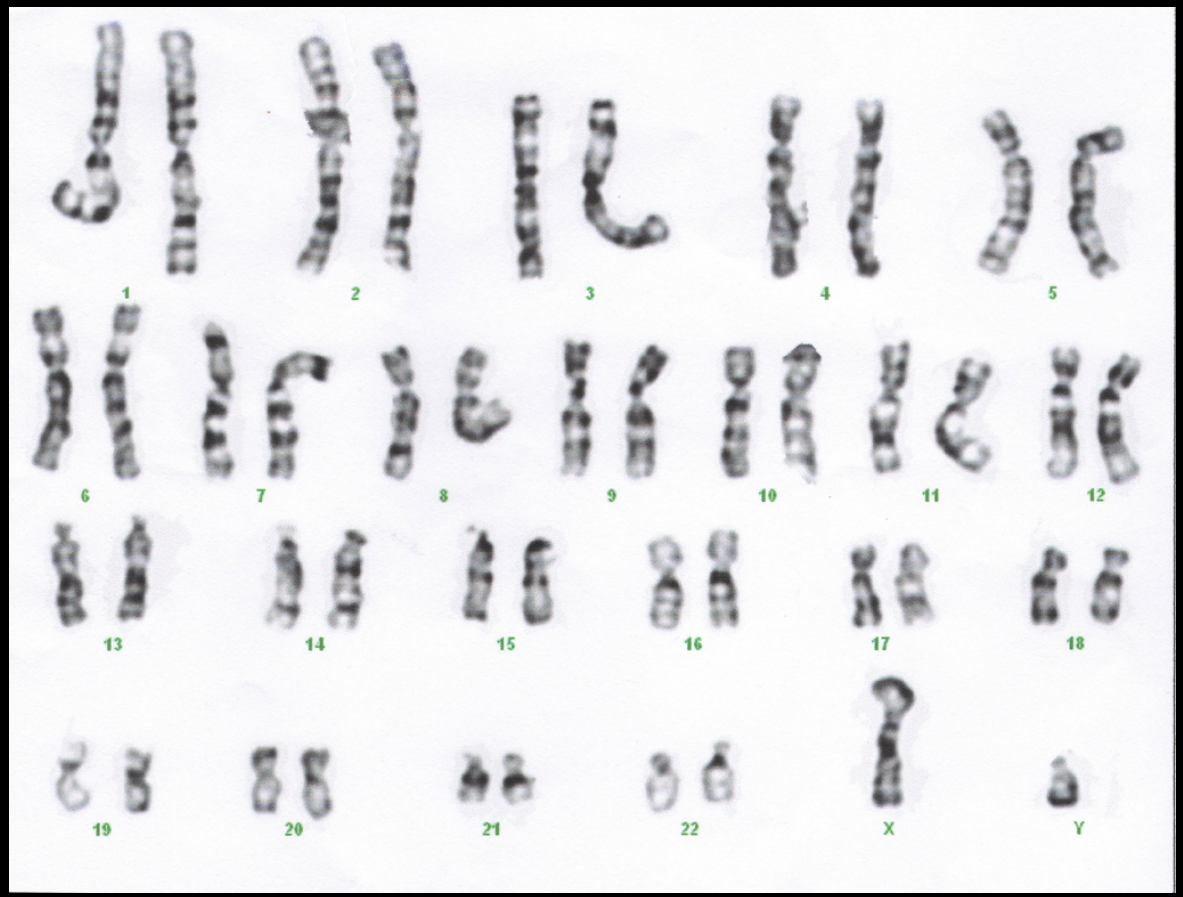
cloudera

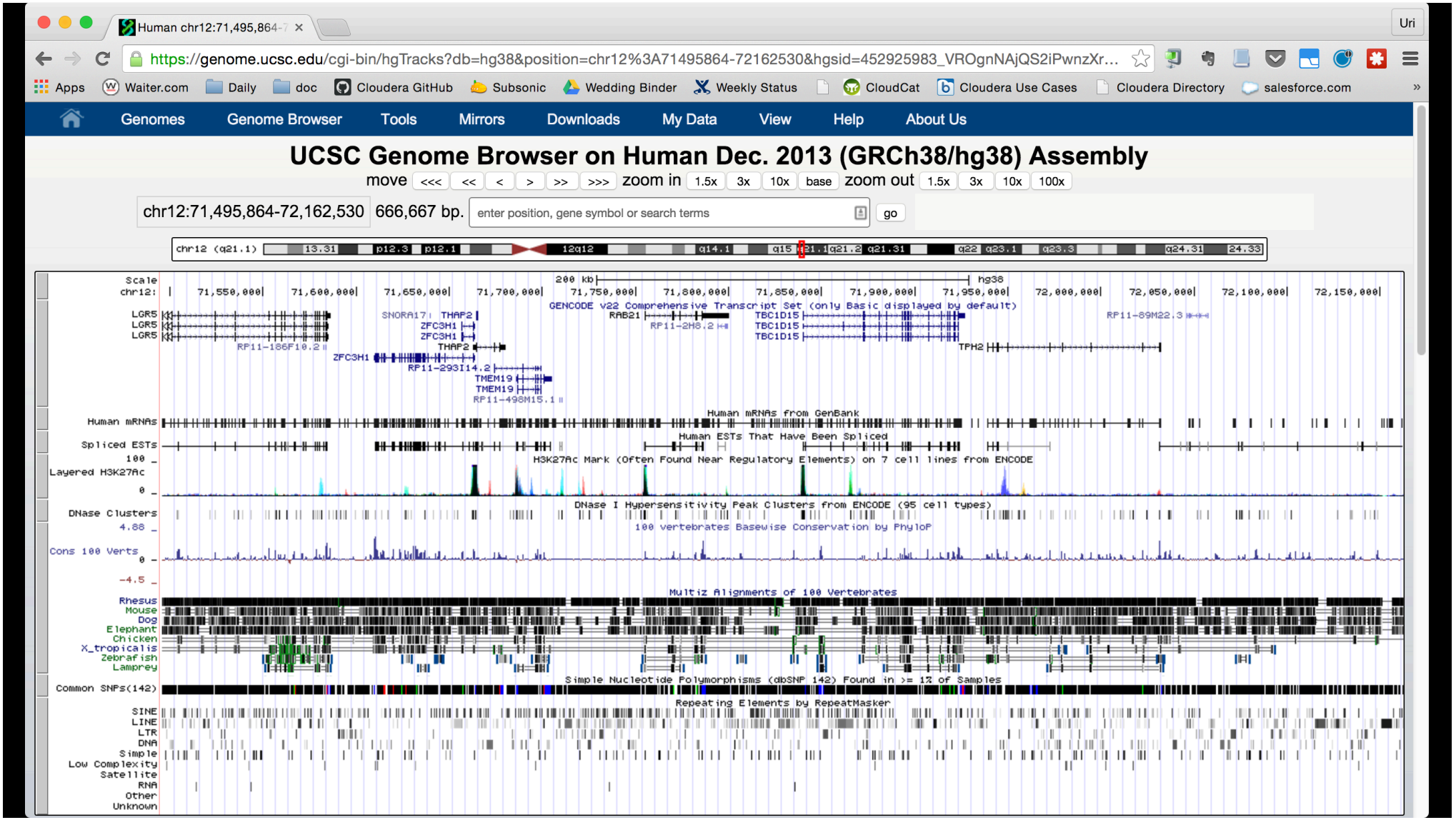

“… decoding the Book of Life”

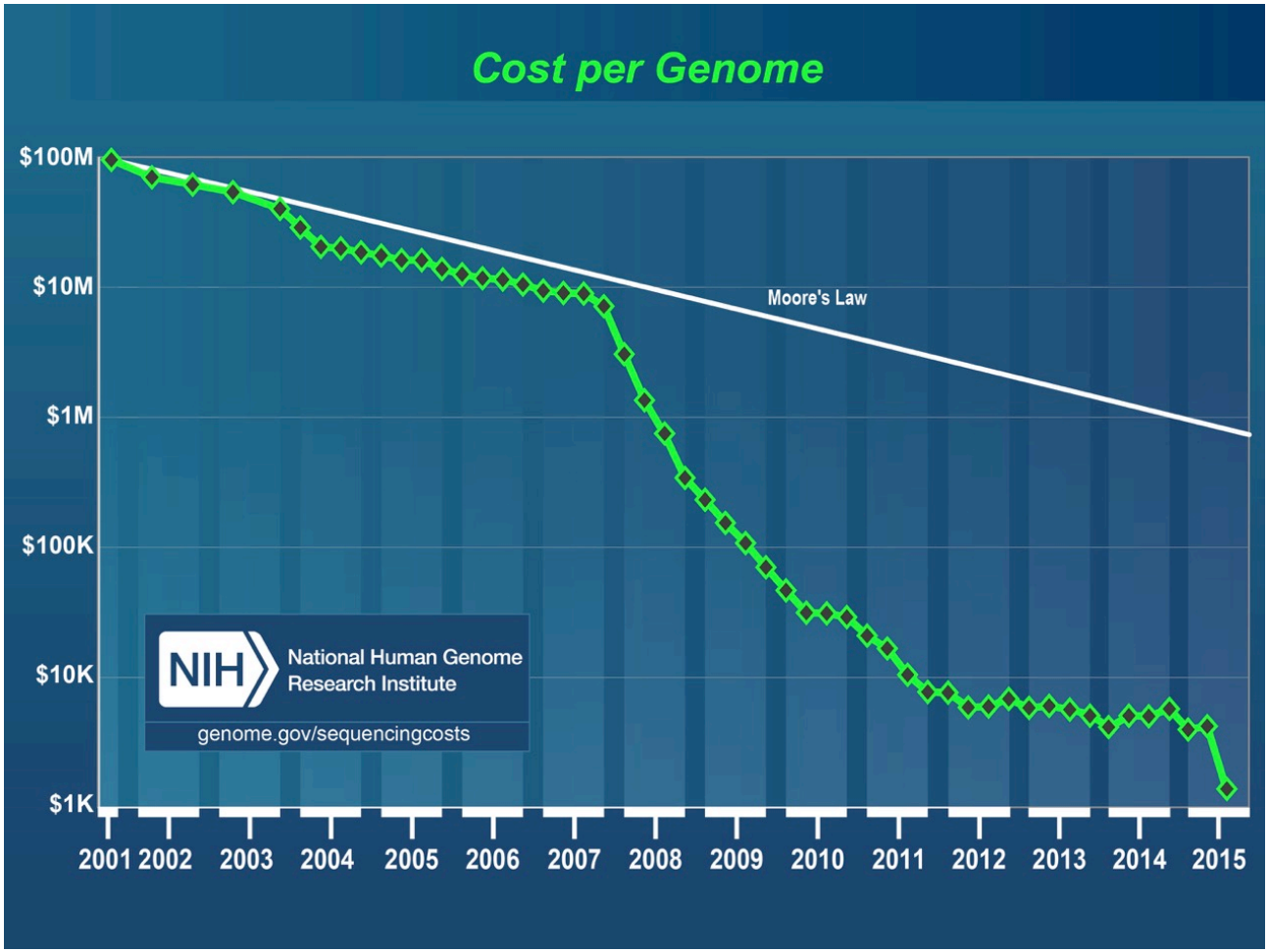

Google Maps
Uri
https://www.google.com/maps/@40.7302934,-73.9973314,634m/data=!3m1!1e3
Apps
Waiter.com
Daily
doc
Cloudera GitHub
Subsonic
Wedding Binder
Weekly Status
CloudCat
Cloudera Use Cases
Cloudera Directory
salesforce.com
Search Google Maps
Uri
U
Map
Google
Imagery ©2015 Google, Map data ©2015 Google
Terms
Privacy
Send feedback
200 ft


Google Maps
https://www.google.com/maps/@40.7302934,-73.9973314,634m/data=!3m1!1e3
Apps
Waiter.com
Daily
doc
Cloudera GitHub
Subsonic
Wedding Binder
Weekly Status
CloudCat
Cloudera Use Cases
Cloudera Directory
salesforce.com
Search Google Maps
Stumptown Coffee Roasters
New York Studio School of Drawing...
Babbo Ristorante
Blue Hill
OTTO Enoteca e Pizzeria
Brittany Hall
Flight Club
One if by Land, Two if by Sea
Blind Tiger
Jones St
Bleecker St
Cornelia St
West 4th Street
Washington Square W
Washington Square N
Washington Square Arch
Washington Square Park
University Pl
E 9th St
E 8th St
E 10th
West 4 St Station
Blue Note
Joe's Pizza
NYU School of Law
Comedy Cellar
Father Demo Square
Minetta Tavern
Market Table
Downing St
Bar Pitti
Bedford St
Center for Architecture
(Le) Poisson Rouge
The Bitter End
Miss Lily's
Washington Square E
Washington Square S
Waverly Pl
Grey Art Gallery
8 St - NYU [N,R]
Mercer St
Kmart
Warehouse Wines & Spirits
Starbucks
Astor Pl
NYU Skirball Center for the Performing Arts
Thompson St
New York University
The Elmer Holmes Bobst Library
W 3rd St
NYU Stern School of Business
New York University External Affairs
Washington Pl
Astor Place Theatre
The Public Theater
The Cooper Union : Administrative Offices
E 4th St
Broadway
Merchant's
Lafayette Grand
Map
Google
Imagery ©2015 Google, Map data ©2015 Google
Terms
Privacy
Send feedback
200 ft


Google Maps
https://www.google.com/maps/@40.7302934,-73.9973314,17.09z
Apps
Waiter.com
Daily
doc
Cloudera GitHub
Subsonic
Wedding Binder
Weekly Status
CloudCat
Cloudera Use Cases
Cloudera Directory
salesforce.com
Search Google Maps
Uri
Flight Club
Babbo Ristorante
Barrow Street Theatre
One if by Land, Two if by Sea
Brittany Hall
Washington Square Arch
Washington Square Park
West 4 St Station
Blue Note
Joe's Pizza
Comedy Cellar
NYU School of Law
Washington Mews
Kmart
8 St - NYU [N,R]
Warehouse Wines & Spirits
Starbucks
Market Table
NYU Skirball Center for the Performing Arts
New York University
The Elmer Holmes Bobst Library
Tisch School Of The Arts
Astor Place Theatre
Bar Pitti
NYU Stern School of Business
Center for Architecture
The Public Theater
(Le) Poisson Rouge
The Bitter End
Other Music
The Cooper Union : Administrative Offices
Miss Lily's
Earth
Google
Washington Square Village Apt
Lafayette Grand
Map data ©2015 Google
Terms
Privacy
Send feedback
200 ft


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Y

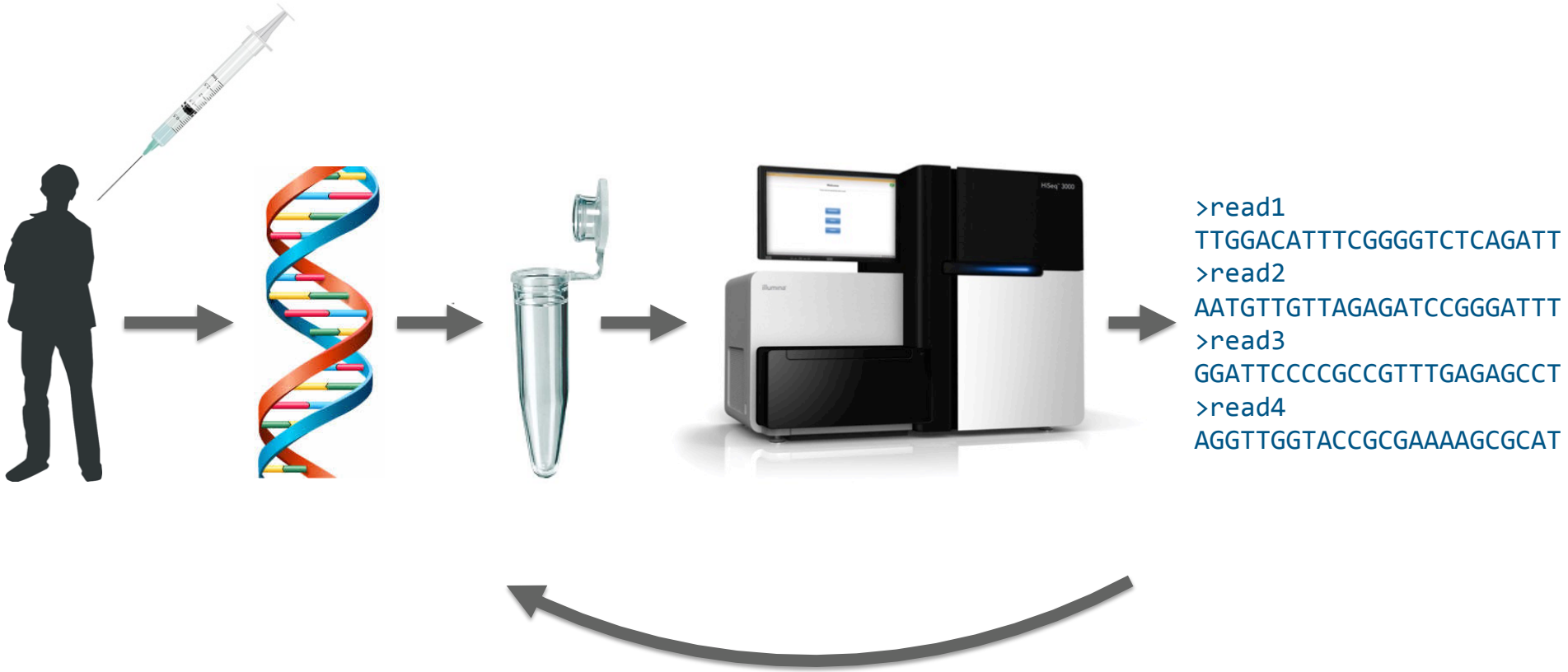

Human chr12:71,495,864-7
https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr12%3A71495864-72162530&hgsid=452925983_VROgnNAjQS2iPwnzXr...
Apps
Waiter.com
Daily
doc
Cloudera GitHub
Subsonic
Wedding Binder
Weekly Status
CloudCat
Cloudera Use Cases
Cloudera Directory
salesforce.com
Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr12:71,495,864-72,162,530
666,667 bp.
enter position, gene symbol or search terms
go
chr12 (q21.1)
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE
DNase I Hypersensitivity Peak Clusters from ENCODE (95 cell types)
100 vertebrates Basewise Conservation by PhyloP
Multiz Alignments of 100 Vertebrates
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Repeating Elements by RepeatMasker



cloudera

What is **bioinformatics**?

cloudera

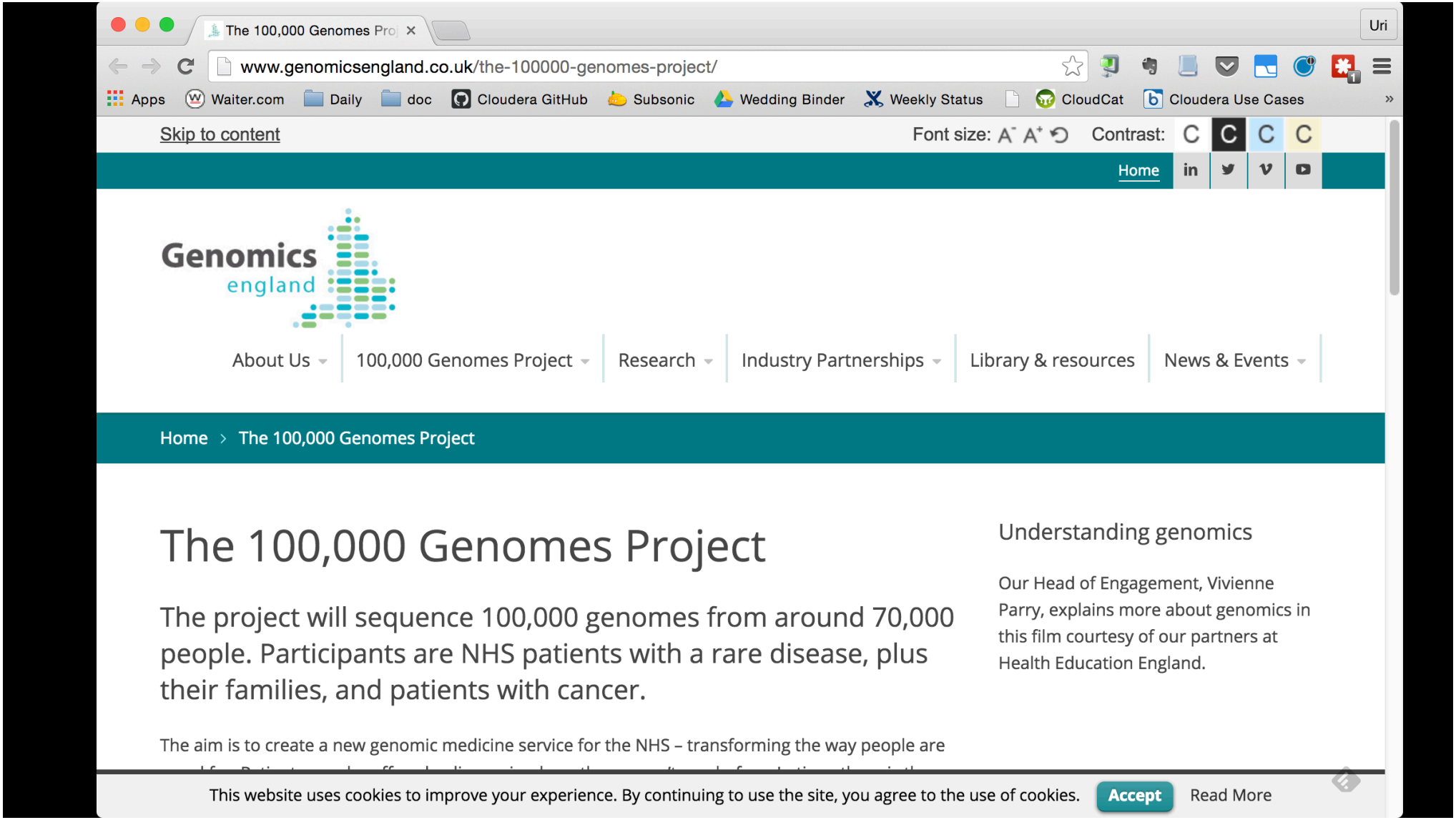


Bioinformatics!

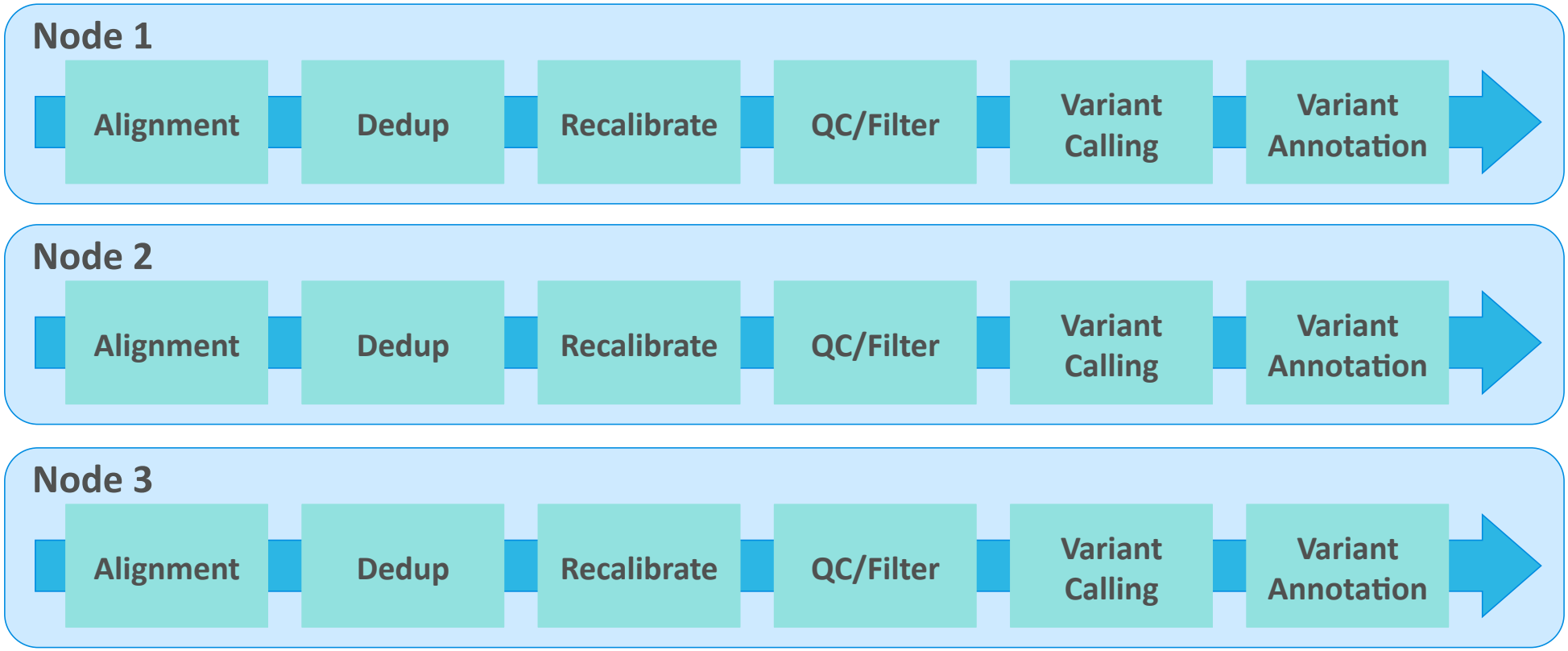
cloudera

Pipelines!

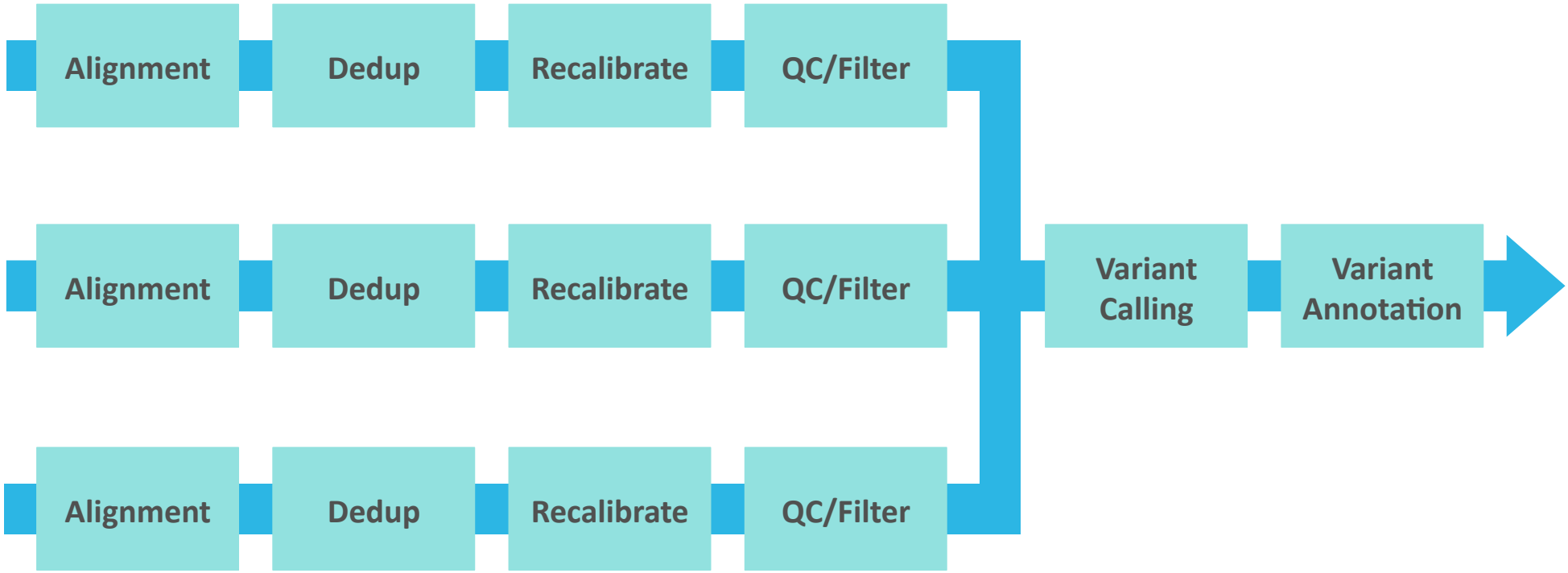
cloudera


The 100,000 Genomes Pro
www.genomicsengland.co.uk/the-100000-genomes-project/
Apps
Waiter.com
Daily
doc
Cloudera GitHub
Subsonic
Wedding Binder
Weekly Status
CloudCat
Cloudera Use Cases
Skip to content
Font size:
Contrast:
Home
Genomics england
About Us
100,000 Genomes Project
Research
Industry Partnerships
Library & resources
News & Events
Home
The 100,000 Genomes Project
The 100,000 Genomes Project
The project will sequence 100,000 genomes from around 70,000 people. Participants are NHS patients with a rare disease, plus their families, and patients with cancer.
The aim is to create a new genomic medicine service for the NHS – transforming the way people are
Understanding genomics
Our Head of Engagement, Vivienne Parry, explains more about genomics in this film courtesy of our partners at Health Education England.
This website uses cookies to improve your experience. By continuing to use the site, you agree to the use of cookies.
Accept
Read More

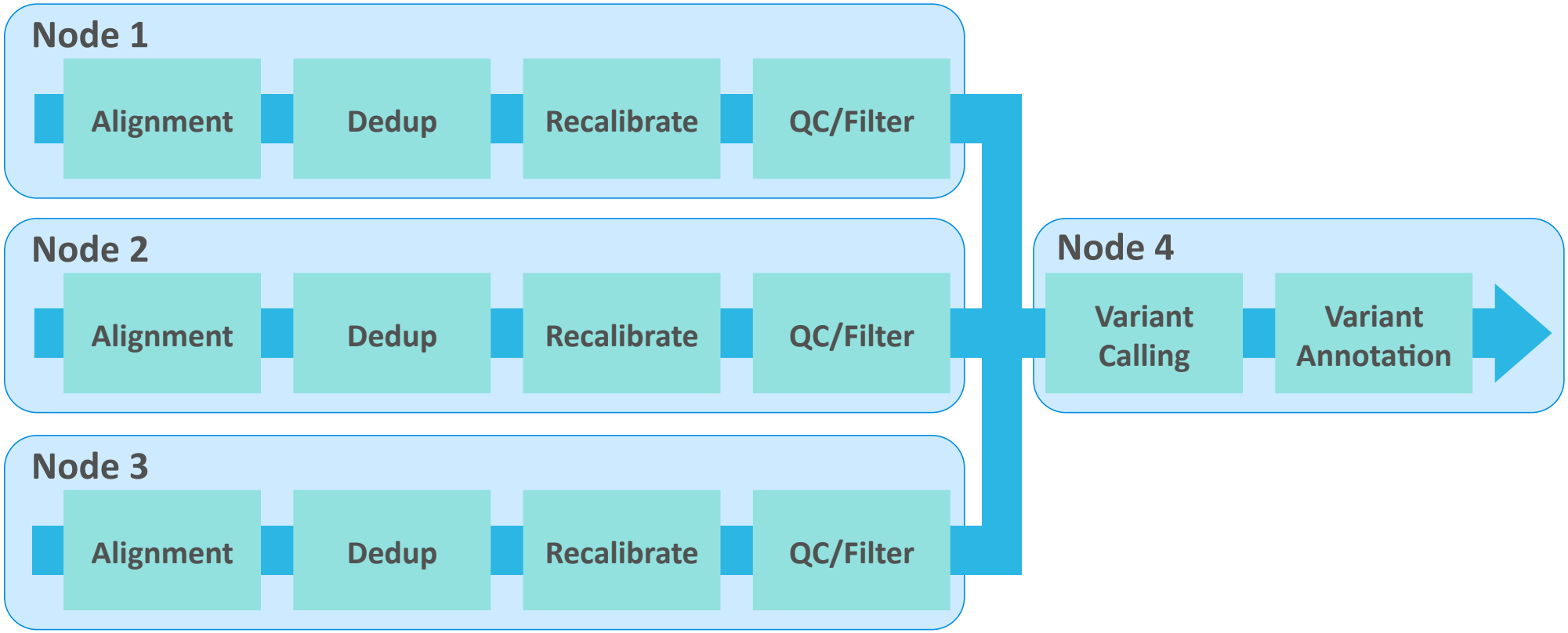
It’s pipelines all the way down!



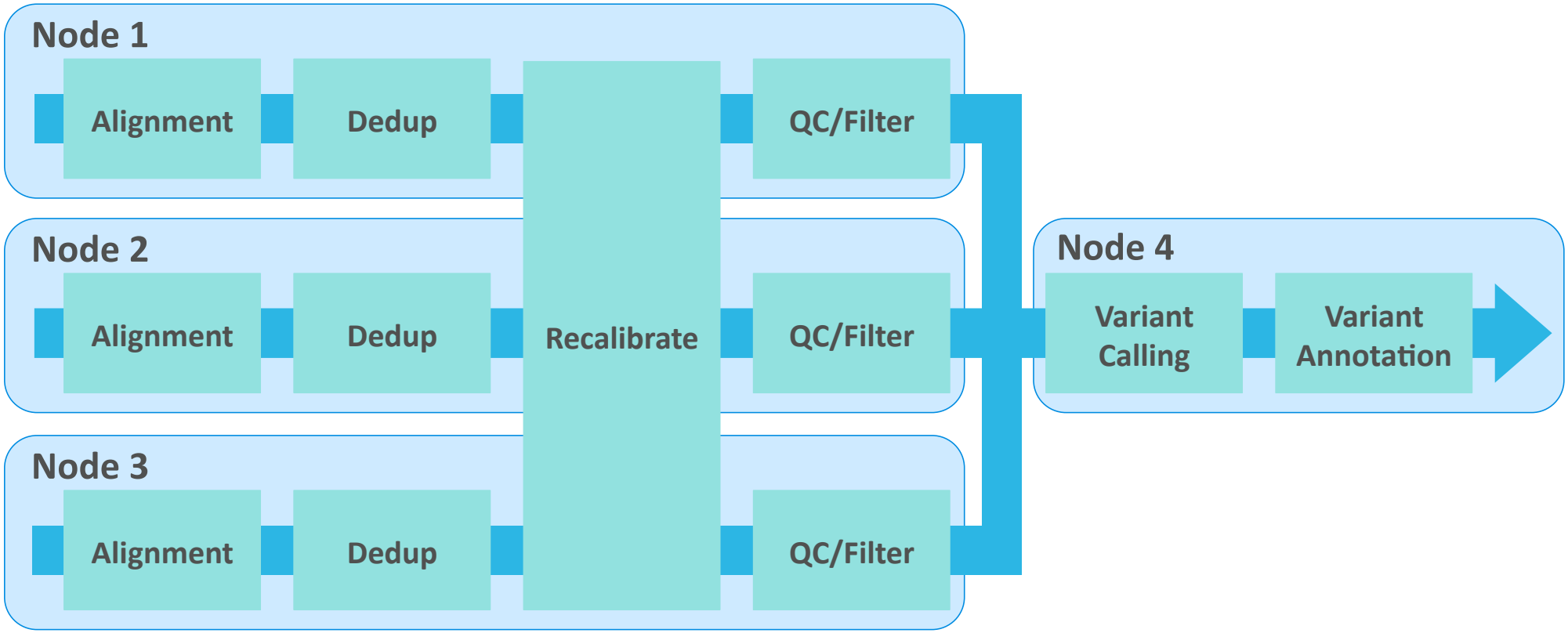
cloudera



cloudera



cloudera



cloudera

How can **Hadoop** be used in bioinformatics?

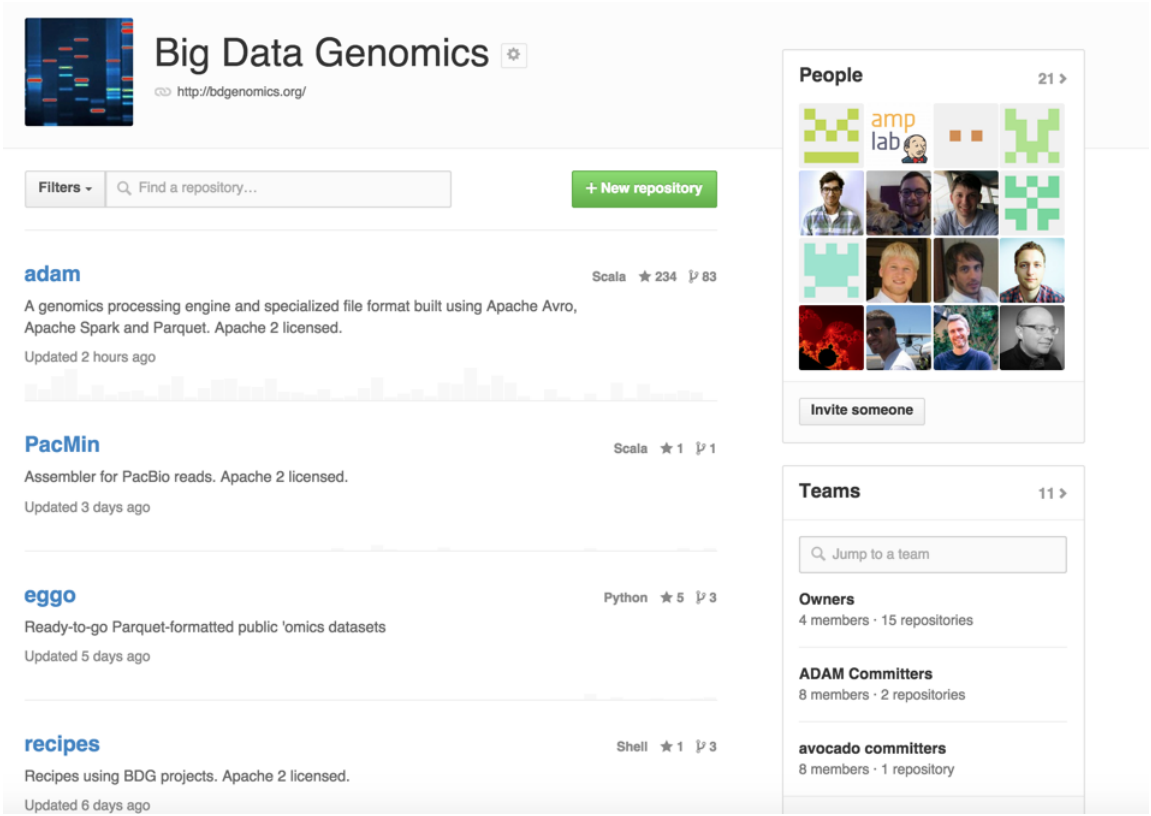
cloudera

Genomics on Hadoop – A Potted History

- 2010 - Hadoop-BAM - MR input/output formats for bio (BAM, VCF, etc)
- 2011 - Seal - MR tools for reads
- 2012 - SeqPig - Pig interface for Hadoop-BAM
- 2013 - ADAM - a genomics analysis platform on Spark, Avro, and Parquet
- 2013 - OpenCGA - a variant store built on HBase
- 2014 - Halvade - a tool to run the GATK best practices pipeline using MR
- 2014 - Guacamole - Spark variant caller for ADAM
- 2015 - GATK4 - a toolkit for running genomics pipelines on Spark
- 2016 - Hail - PLINK-like tool for whole genome association analysis

Spark + Genomics = ADAM

- Hosted at Berkeley and the AMPLab
- Apache 2 License
- Contributors from both research and commercial organizations
- Core spatial primitives, variant calling
- Avro and Parquet for data models and file formats



cloudera

Genome Analysis Toolkit (GATK)

- Developed by the Broad Institute
- Core is MIT license, some proprietary tools on top
- Version 4 has been re-written to use Spark, now competitive with ADAM for speed
- Uses existing bio file formats for input and output, but Spark RDDs for intermediate data

Bioinformatics File Formats

- Hand crafted
- Poorly specified
- Text based
- Unsplittable (in the Hadoop sense)

cloudera

BAM files

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

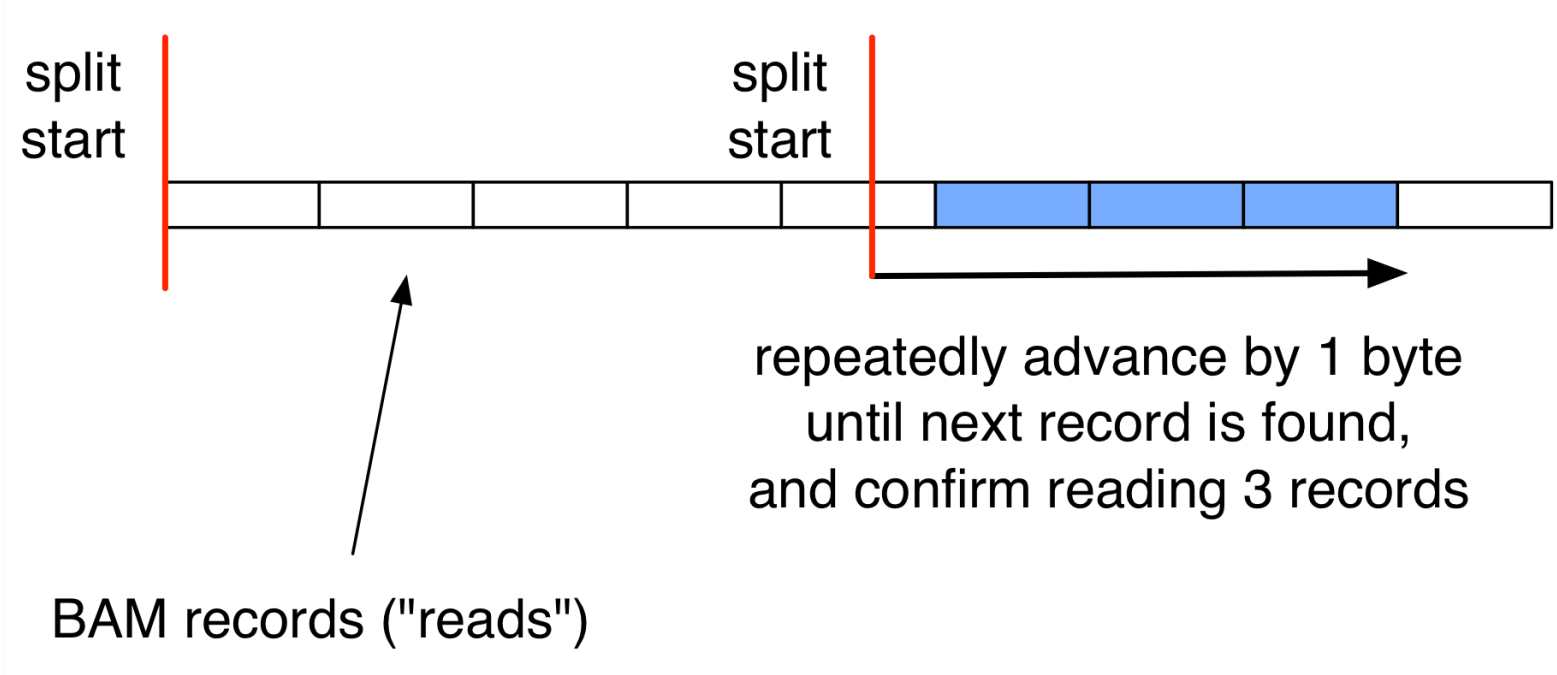
Reads

Position

Sequence

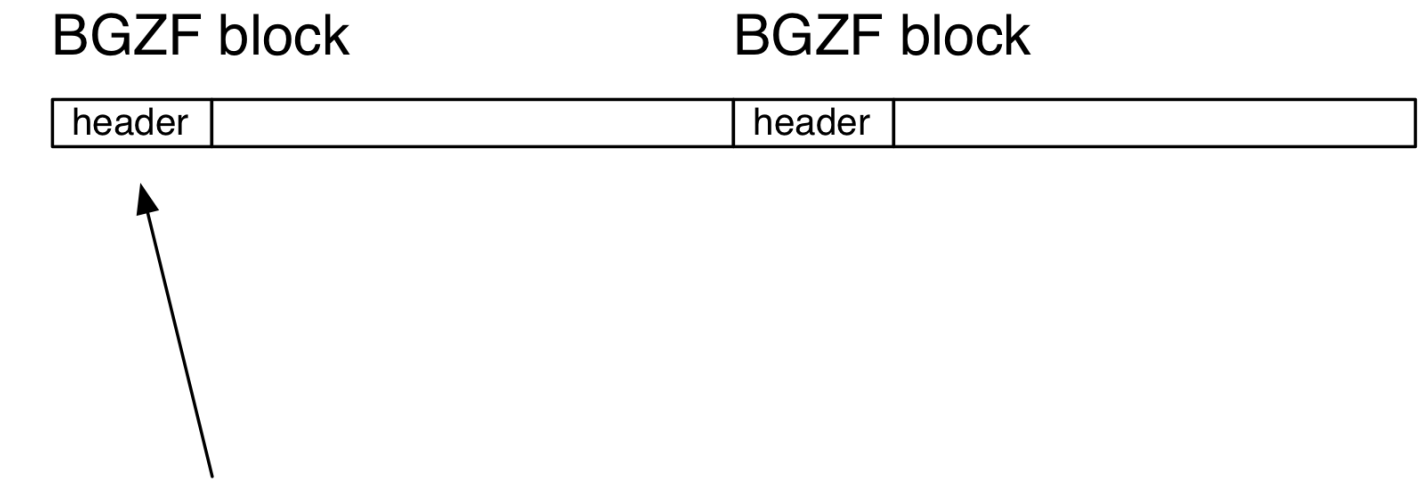
cloudera

Example: splitting BAM files



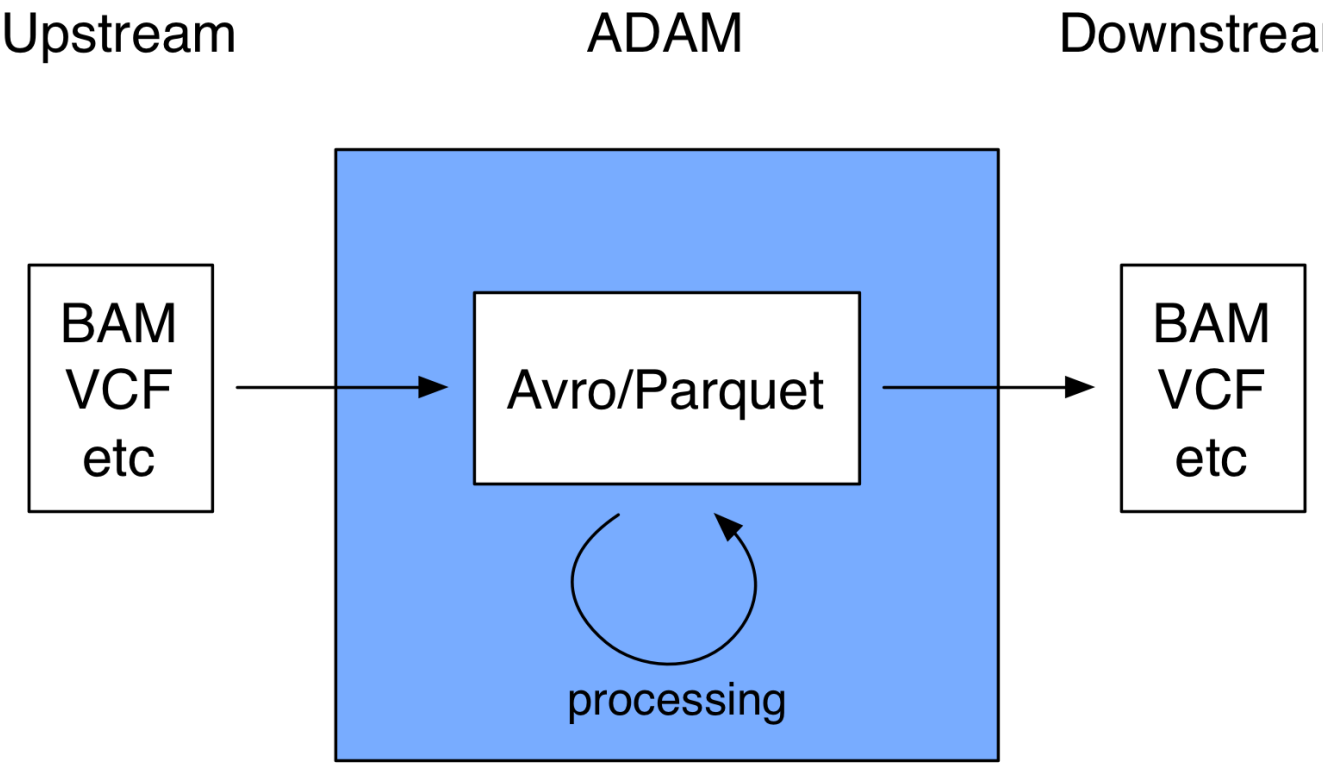
cloudera

Example: splitting BAM files (BGZF compression)



header includes block length (as gzip extension),
so easy to index and split

Why not use Hadoop formats?



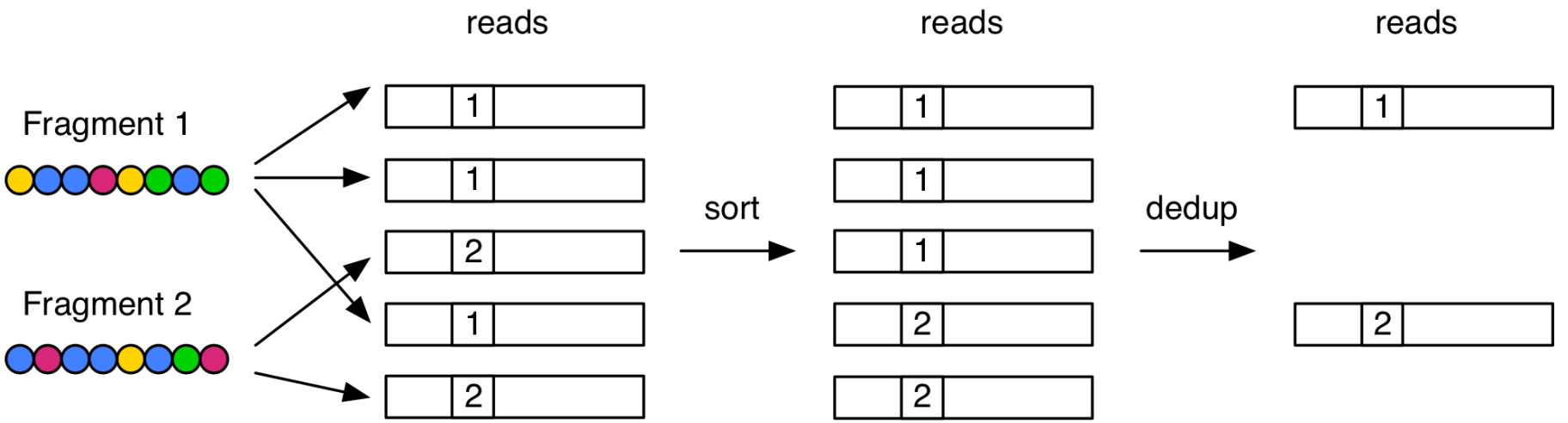
cloudera

cloudera

Dedup

cloudera

Mark Duplicates



Method

Code

```
/**
 * Main work method.  Reads the BAM file once and collects sorted information about
 * the 5' ends of both ends of each read (or just one end in the case of pairs).
 * Then makes a pass through those determining duplicates before re-reading the
 * input file and writing it out with duplication flags set correctly.
 */
protected int doWork() {
    // build some data structures
    buildSortedReadEndLists(useBarcodes);
    generateDuplicateIndexes(useBarcodes);

    final SAMFileWriter out =
        new SAMFileWriterFactory().makeSAMOrBAMWriter(outputHeader, true, OUTPUT);
    final CloseableIterator<SAMRecord> iterator = headerAndIterator.iterator;
    while (iterator.hasNext()) {
        final SAMRecord rec = iterator.next();
        if (!rec.isSecondaryOrSupplementary()) {
            if (recordInFileIndex == nextDuplicateIndex) {
                rec.setDuplicateReadFlag(true);
                // Now try and figure out the next duplicate index
                if (this.duplicateIndexes.hasNext()) {
                    nextDuplicateIndex = this.duplicateIndexes.next();
                } else {
                    // Only happens once we've marked all the duplicates
                    nextDuplicateIndex = -1;
                }
            } else {
                rec.setDuplicateReadFlag(false);
            }
        }
        recordInFileIndex++;
        if (!this.REMOVE_DUPLICATES || !rec.getDuplicateReadFlag()) {
            out.addAlignment(rec);
```

cloudera


```
@Option(shortName = "MAX_FILE_HANDLES",
        doc = "Maximum number of file handles to keep open when spilling " +
              "read ends to disk. Set this number a little lower than the " +
              "per-process maximum number of file that may be open. This " +
              "number can be found by executing the 'ulimit -n' command on " +
              "a Unix system.")
public int MAX_FILE_HANDLES_FOR_READ_ENDS_MAP = 8000;
```

Spark Implementation

```
JavaPairRDD<String, Iterable<GATKRead>> keyedReads = ...;

JavaPairRDD<String, PairedEnds> keyPairs =
   keyedReads.flatMapToPair(keyedRead -> { ... };

JavaPairRDD<String, Iterable<PairedEnds>> keyedPairs =
   keyPairs.groupByKey(numReducers);

JavaRDD<GATKRead> markedDups = markPairedEnds(keyedPairs,
   scoringStrategy, finder, header);
```

cloudera

Lessons Learned

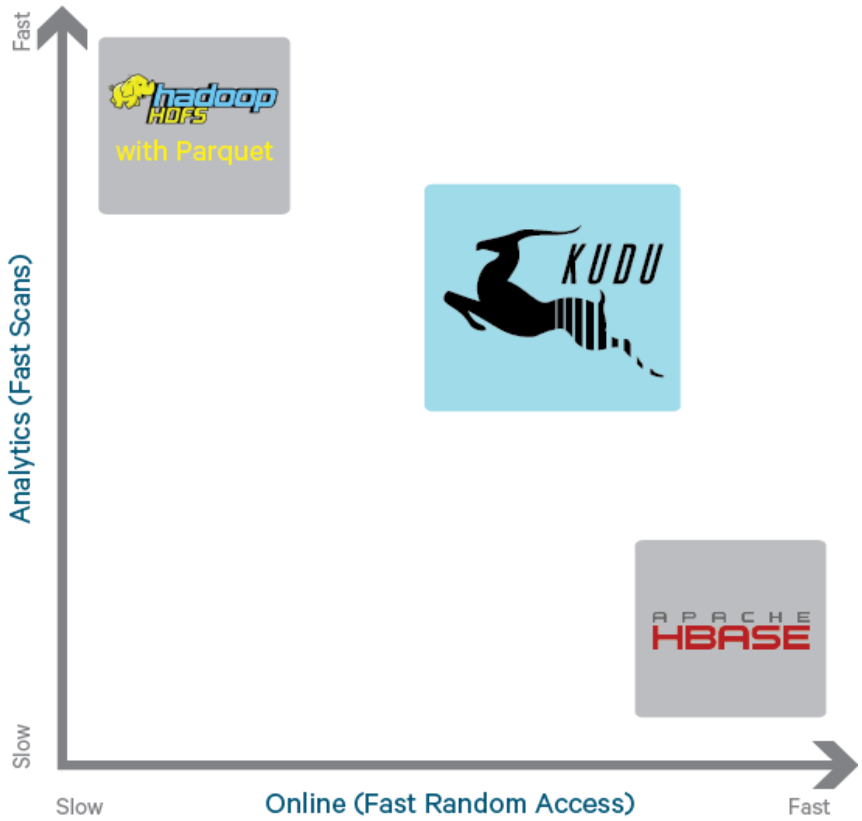
1. Figure out how to read and write existing formats efficiently
2. Spark is a great API, but developers need to understand consequences of e.g. the shuffle, serialization cost
3. Work with domain experts, on existing projects, if possible

Future developments

cloudera

Kudu for Variant Stores

- Kudu fills gap between HDFS and HBase
- Fast scans and updateable
- Add new annotations to genomics data (variants) without rewriting whole dataset
- Key = genome position
- Range partitioning



cloudera

Hail

- Scalable variant analytics in Spark
- Command-line tools like PLINK
- Parquet-based storage by default, other storage possibilities like Kudu

cloudera

Links

- ADAM
  - https://github.com/bigdatagenomics/adam
- GATK4
  - https://github.com/broadinstitute/gatk

cloudera

Acknowledgements

UCBerkeley
Matt Massie
Frank Nothaft
Michael Heuer

Tamr
Timothy Danford

MSSM
Jeff Hammerbacher
Ryan Williams

Cloudera
Uri Laserson
Sandy Ryza
Sean Owen

cloudera

cloudera

Thank you

@tom_e_white

tom@cloudera.com

