Genomic Data Processing and Machine Learning Workflows using Spark

SPARK SUMMIT 2017

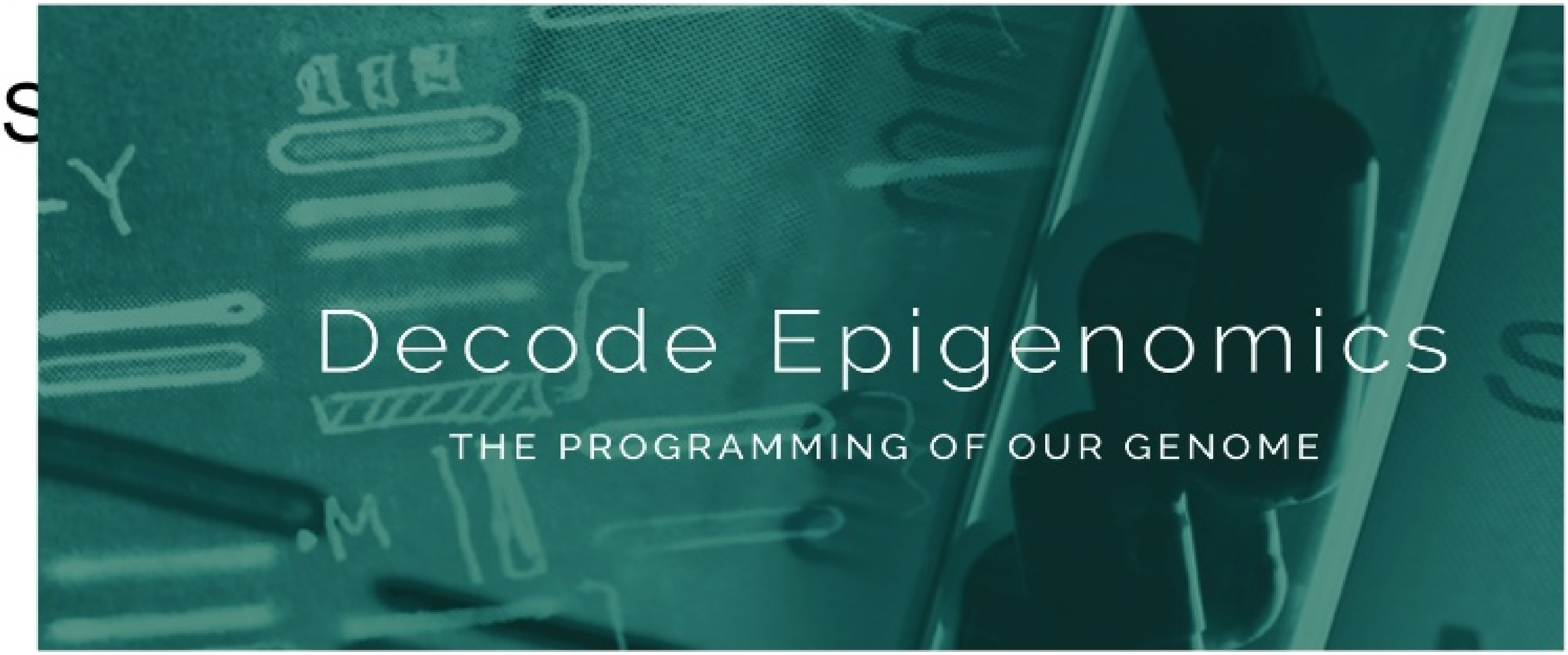
Epinomics

Anupama Joshi

Matei Negulescu

Epinomics

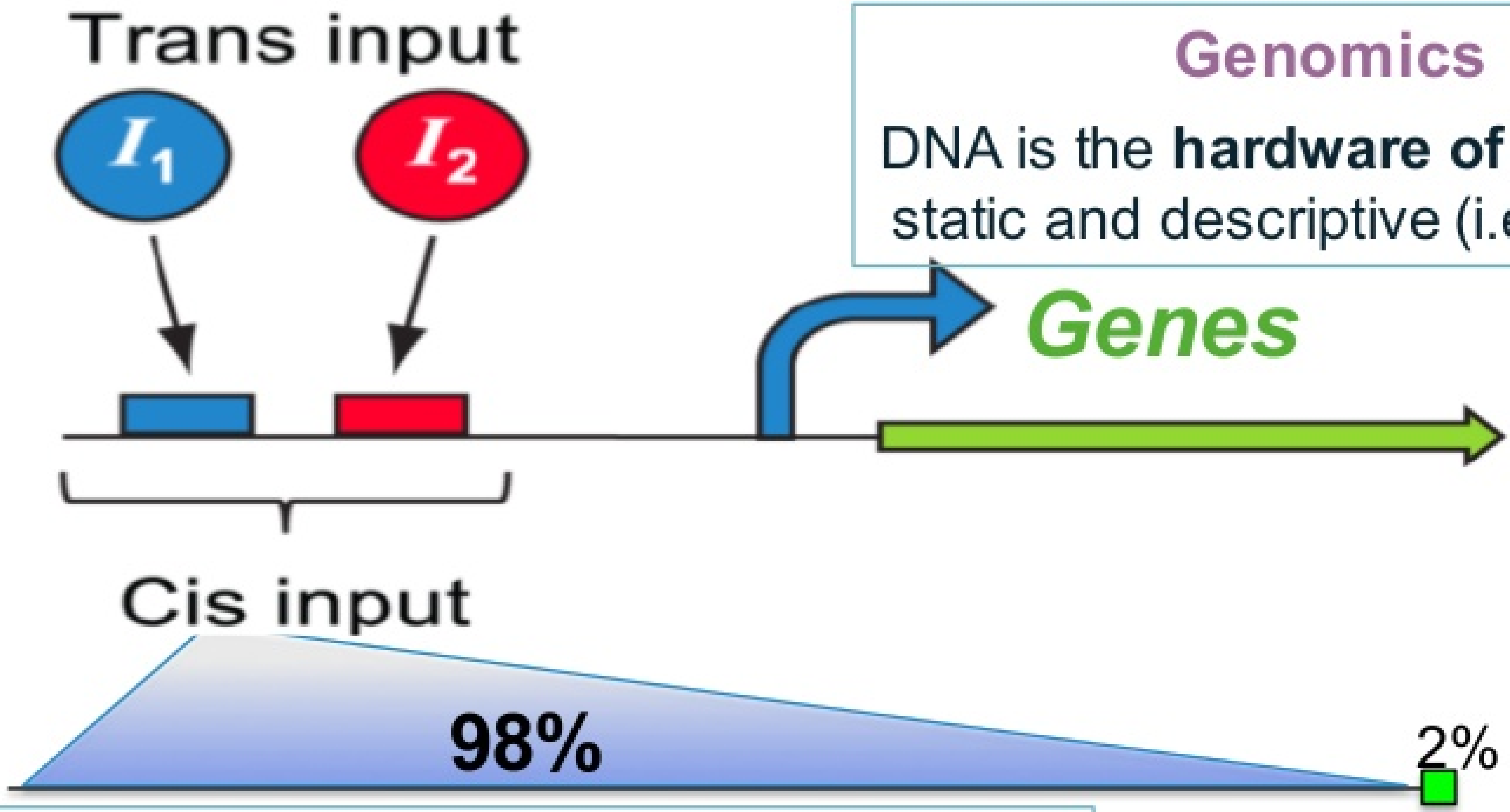
- S



A platform that drives **personalized medicine** by leveraging big data analytics and proprietary **epigenomic** technology.

What is Epigenomics?

Instructions encoded within non-coding sequence



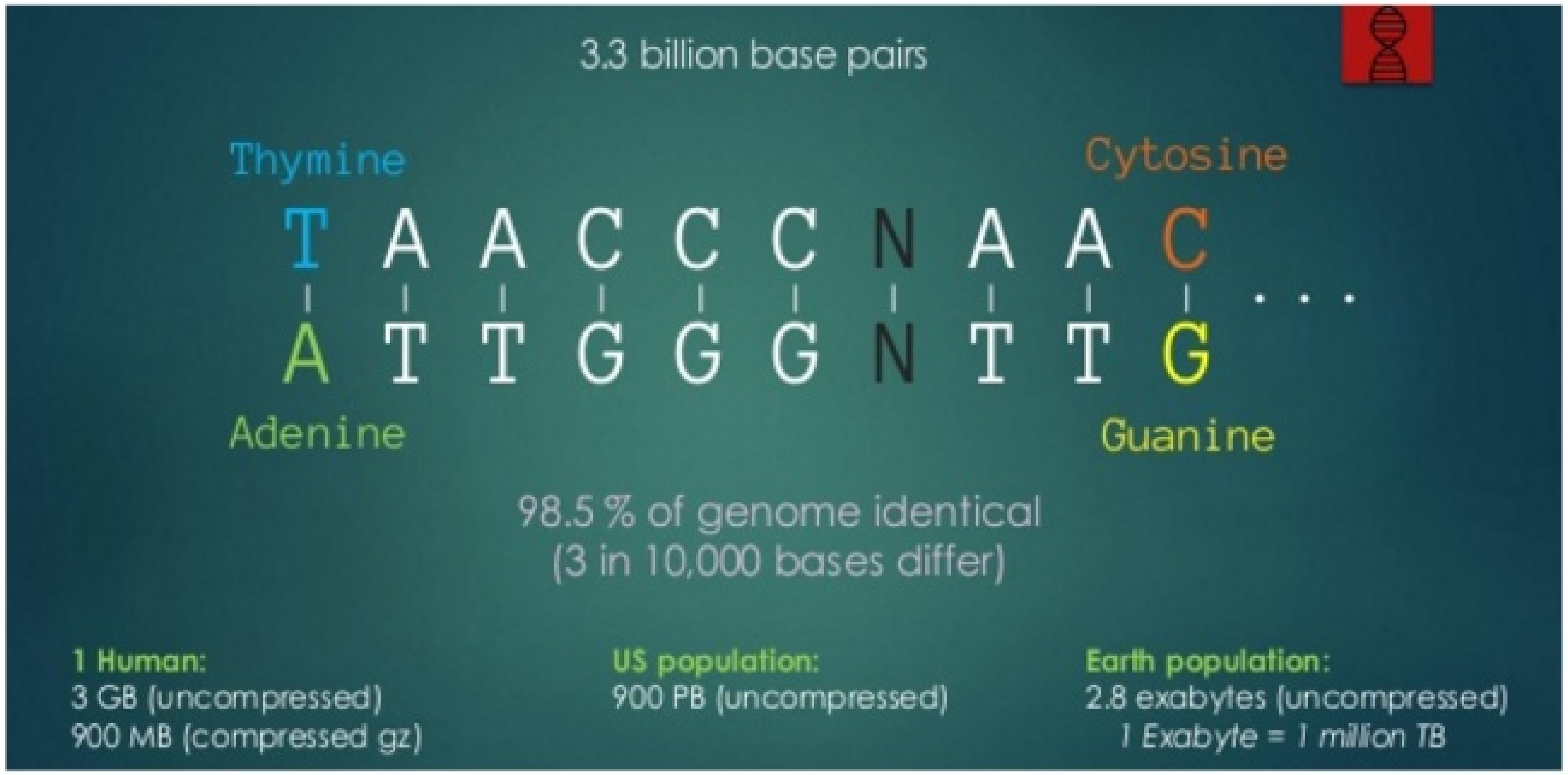
Genomics

DNA is the **hardware of the body**: static and descriptive (i.e. nature).

Epigenomics

Software layer: dynamically turns genes on or off (i.e. nature and nurture).

Typical Genomic data



3.3 billion base pairs

Thymine

Cytosine

T A A C C C N A A C

A T T G G G N T T G . . .

Adenine

Guanine

98.5 % of genome identical
(3 in 10,000 bases differ)

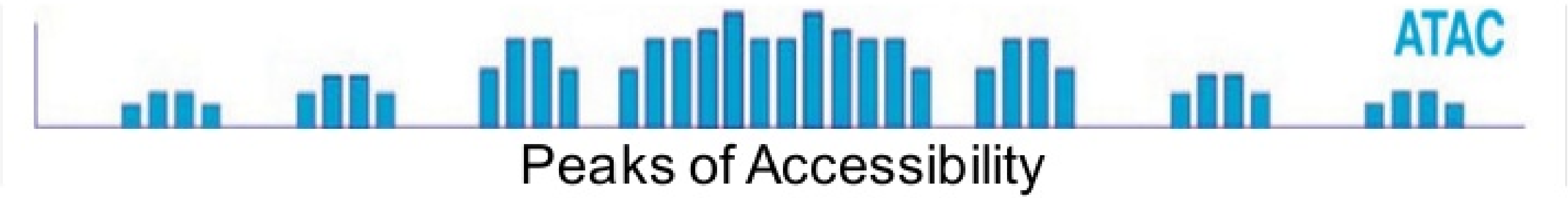
1 Human:
3 GB (uncompressed)
900 MB (compressed gz)

US population:
900 PB (uncompressed)

Earth population:
2.8 exabytes (uncompressed)
1 Exabyte = 1 million TB

- Typical genomic sequencing data contains the protein letters **ATCG** .
- Most research work focuses on **variation** from standard genome sequences.

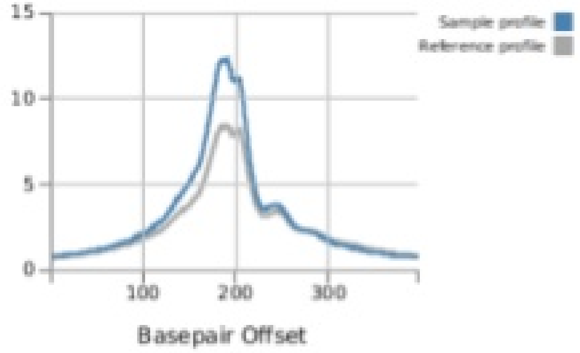
Epigenomic Data



Fragment Data

Single fragment where DNA was accessible during the experiment.

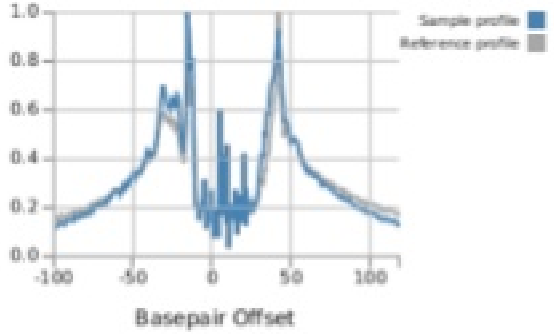
```
chr1 713701    714600    +
chr1 804976    805650    +
```



Peaks Data

Aggregated regions of the genome where DNA was accessible during the experiment.

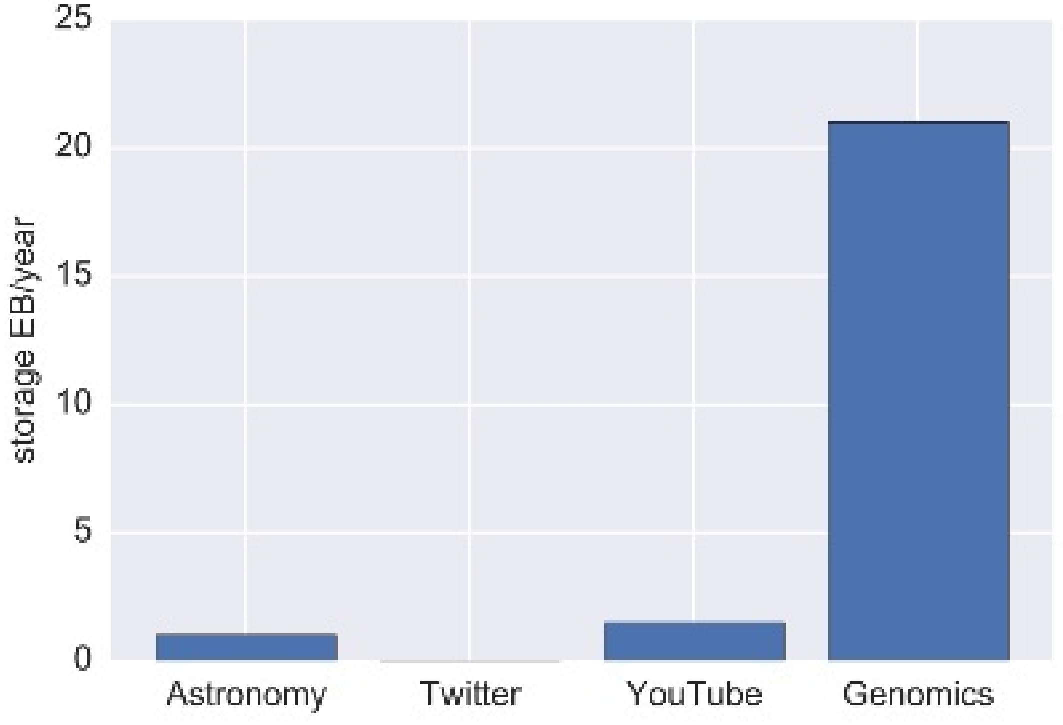
```
chr1 713701    714600    peak.1    899  +
chr1 804976    805650    peak.2    674  +
```

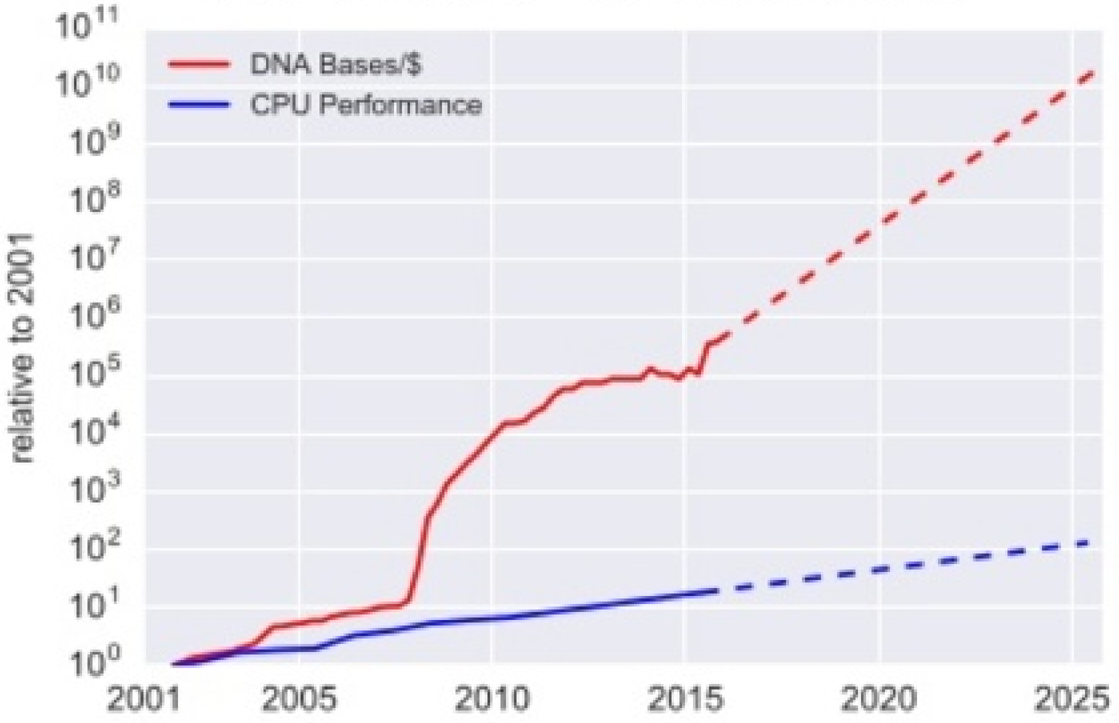


SPARK SUMMIT 2017

EPINOMICS

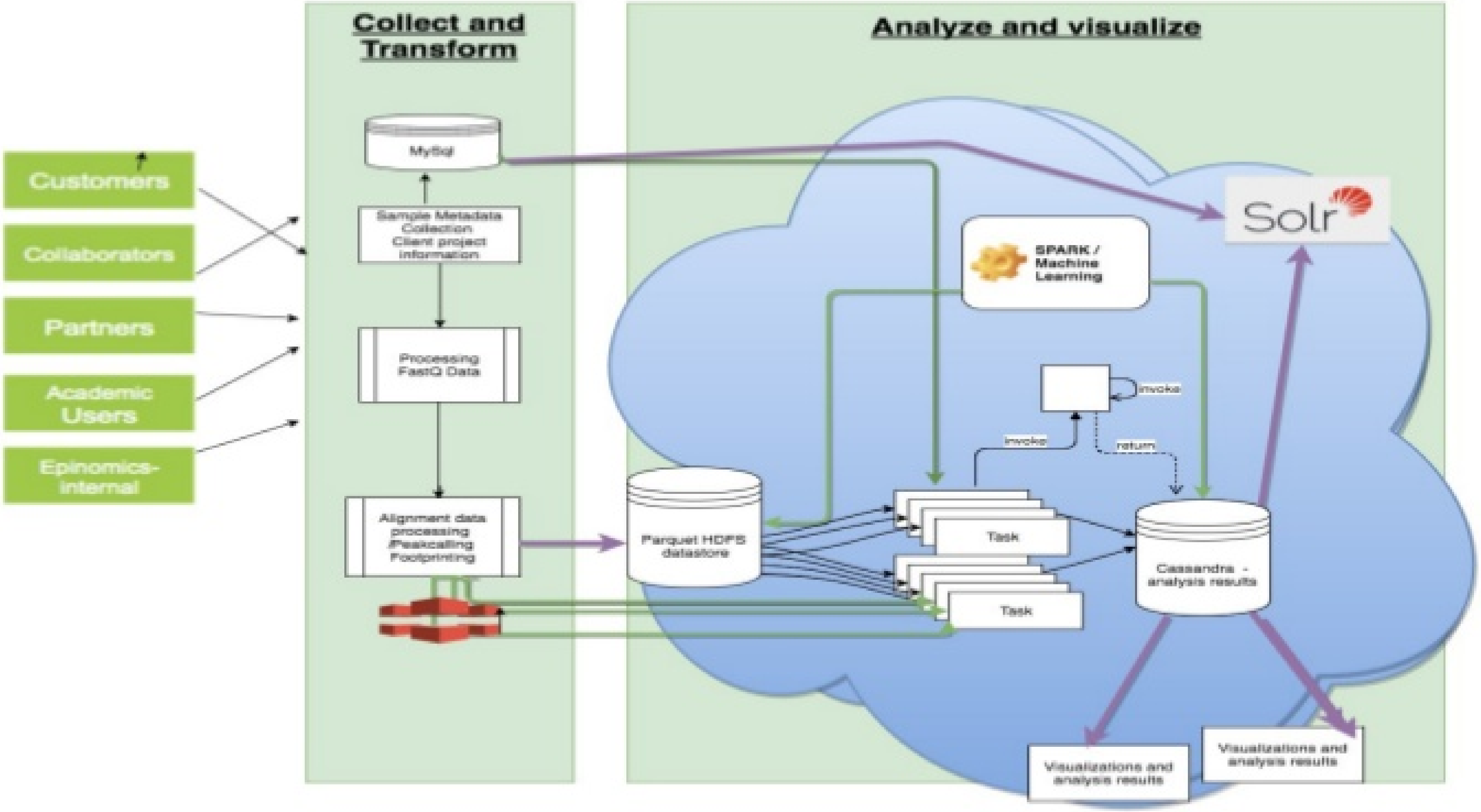
Genomic Data Growth


Data Acquisition (2015)


Moore’s law vs Genomics

Stephens, et al., Big Data: Astronomical or Genomical? (2015)

Data @ Epinomics



Goal: A Map of Human Health

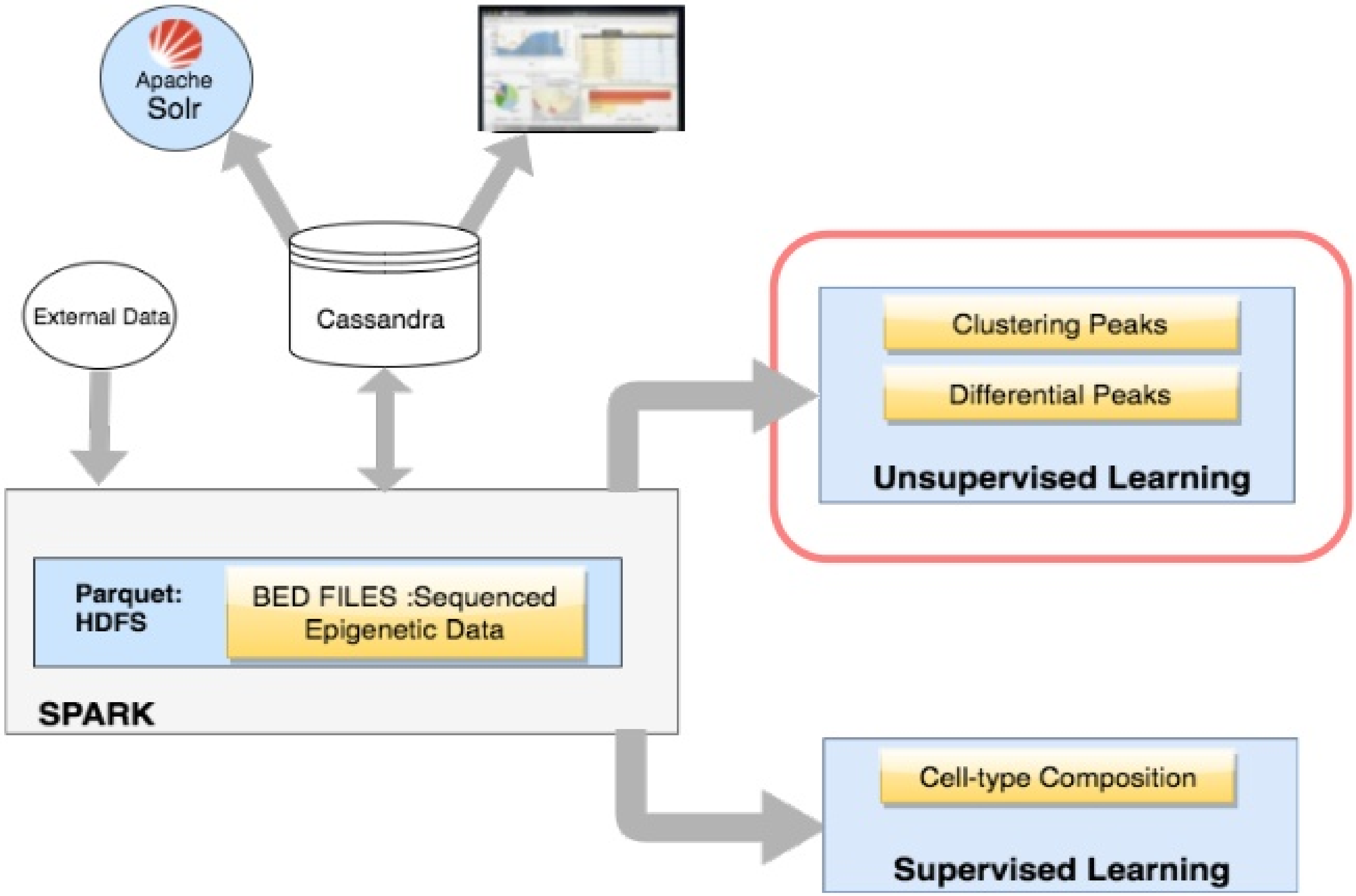
Assessing data quality

Finding patterns in the data

- **Clusters of similar data**
- **Significant differences between groups**
- **Finding unique fingerprints**

Actionable Insight

- Diagnostics, new drugs, dosage, safety



External Data

Cassandra

Parquet: HDFS

BED FILES :Sequenced Epigenetic Data

SPARK

Clustering Peaks

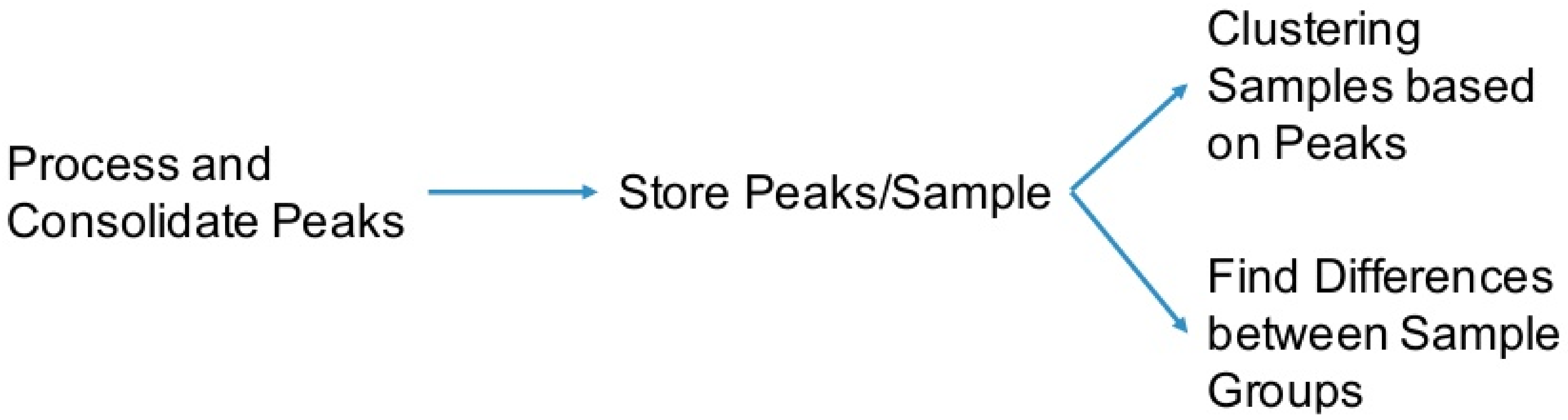
Differential Peaks

Unsupervised Learning

Cell-type Composition

Supervised Learning

Unsupervised Patterns of Accessibility



Process and Consolidate Peaks → Store Peaks/Sample

Store Peaks/Sample → Clustering Samples based on Peaks

Store Peaks/Sample → Find Differences between Sample Groups

GraphX

cassandra

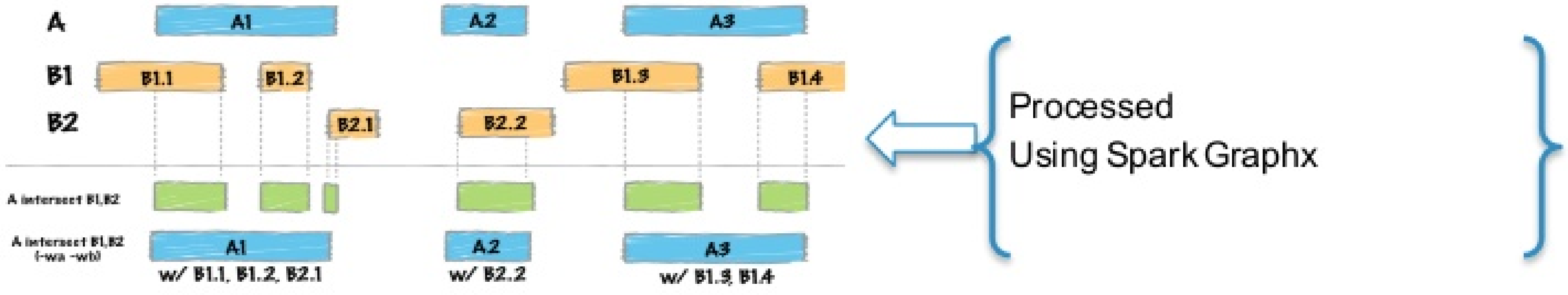
Spark MLlib

Peaks Processing

Each sample will have between 150K to 200K peaks

A typical biological experiment can have between 10 to 200 samples.

Consolidate and process overlapping peaks



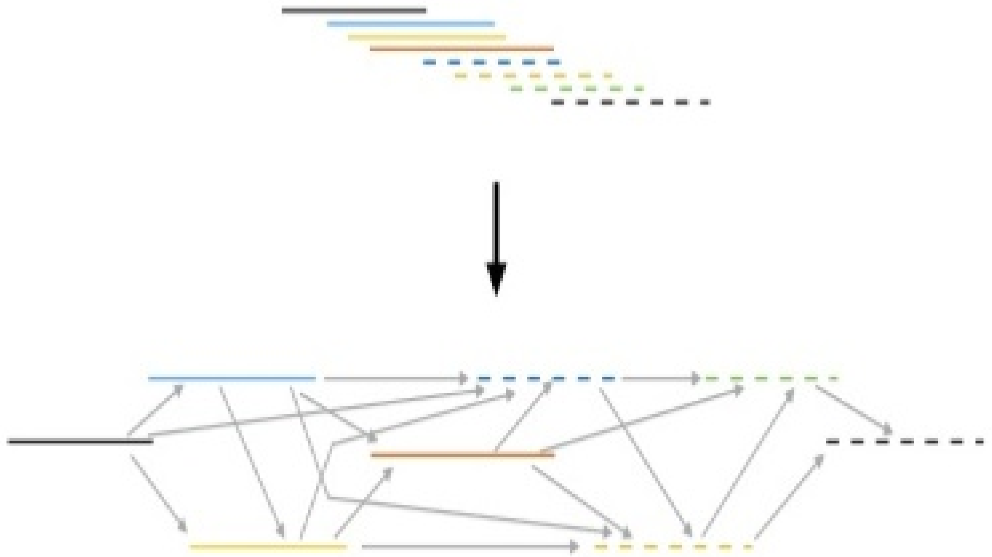
Processed
Using Spark Graphx

Source -:http://bedtools.readthedocs.io/

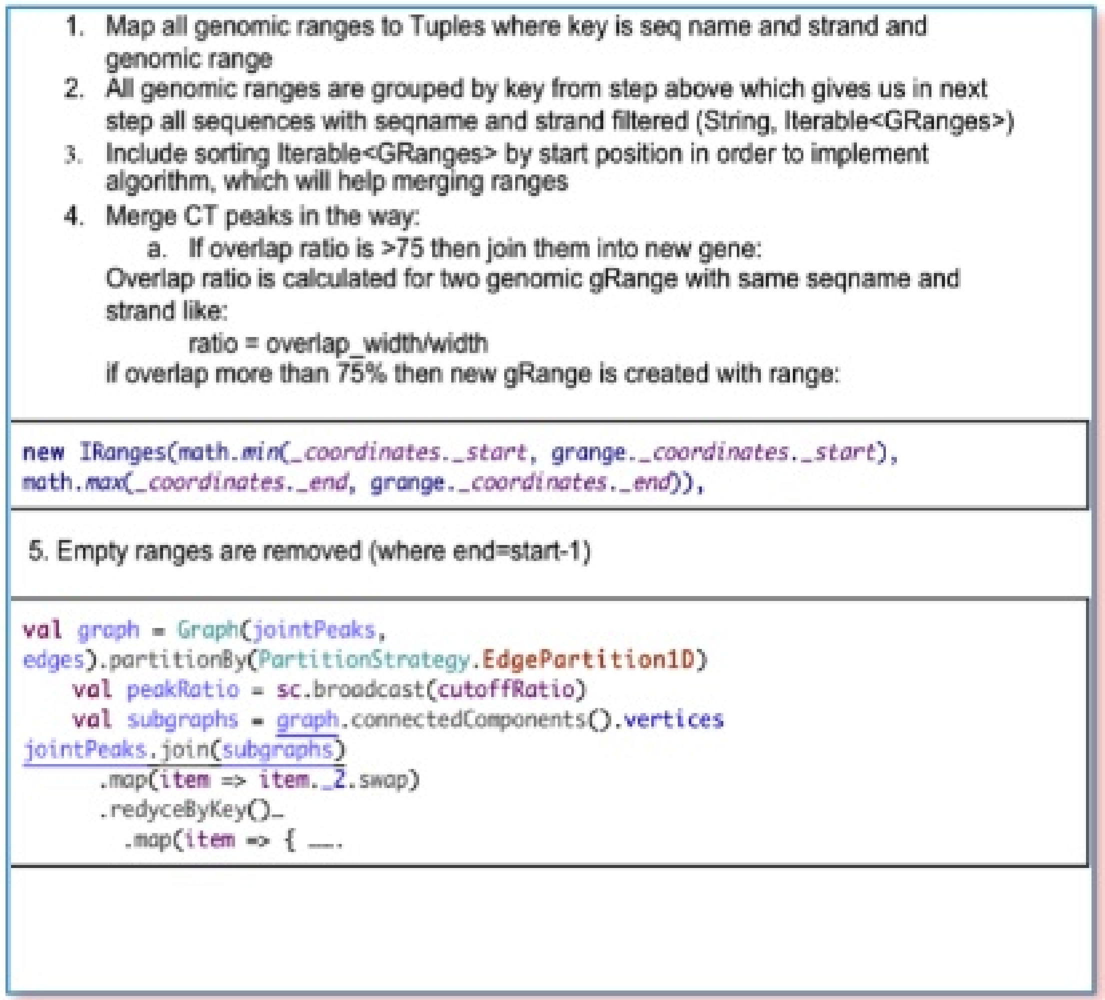
A typical experiment will have between 300K to 600k overlapping peaks. (depending on dataset and sequencing depth)

Peaks Processing

Merges overlapping peaks of two genomic ranges vectors using GraphX library



Nodes are peaks and edges are overlaps



1. Map all genomic ranges to Tuples where key is seq name and strand and genomic range
2. All genomic ranges are grouped by key from step above which gives us in next step all sequences with seqname and strand filtered (String, Iterable<GRanges>)
3. Include sorting Iterable<GRanges> by start position in order to implement algorithm, which will help merging ranges
4. Merge CT peaks in the way:
   a. If overlap ratio is >75 then join them into new gene:

   Overlap ratio is calculated for two genomic gRange with same seqname and strand like:

   ratio = overlap_width/width

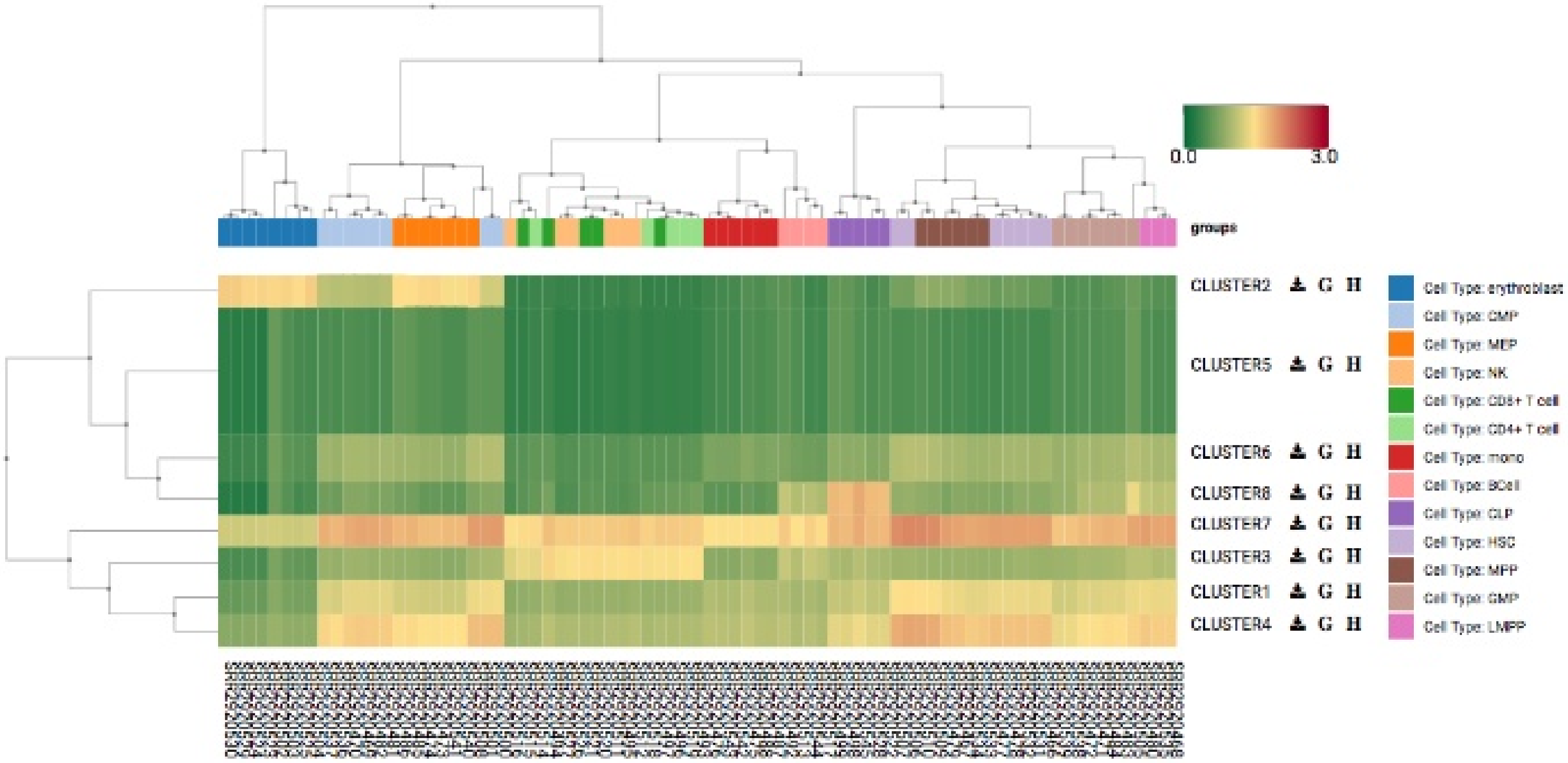
   if overlap more than 75% then new gRange is created with range:

```
new IRanges(math.min(_coordinates._start, grange._coordinates._start),
math.max(_coordinates._end, grange._coordinates._end)),
```

5. Empty ranges are removed (where end=start-1)

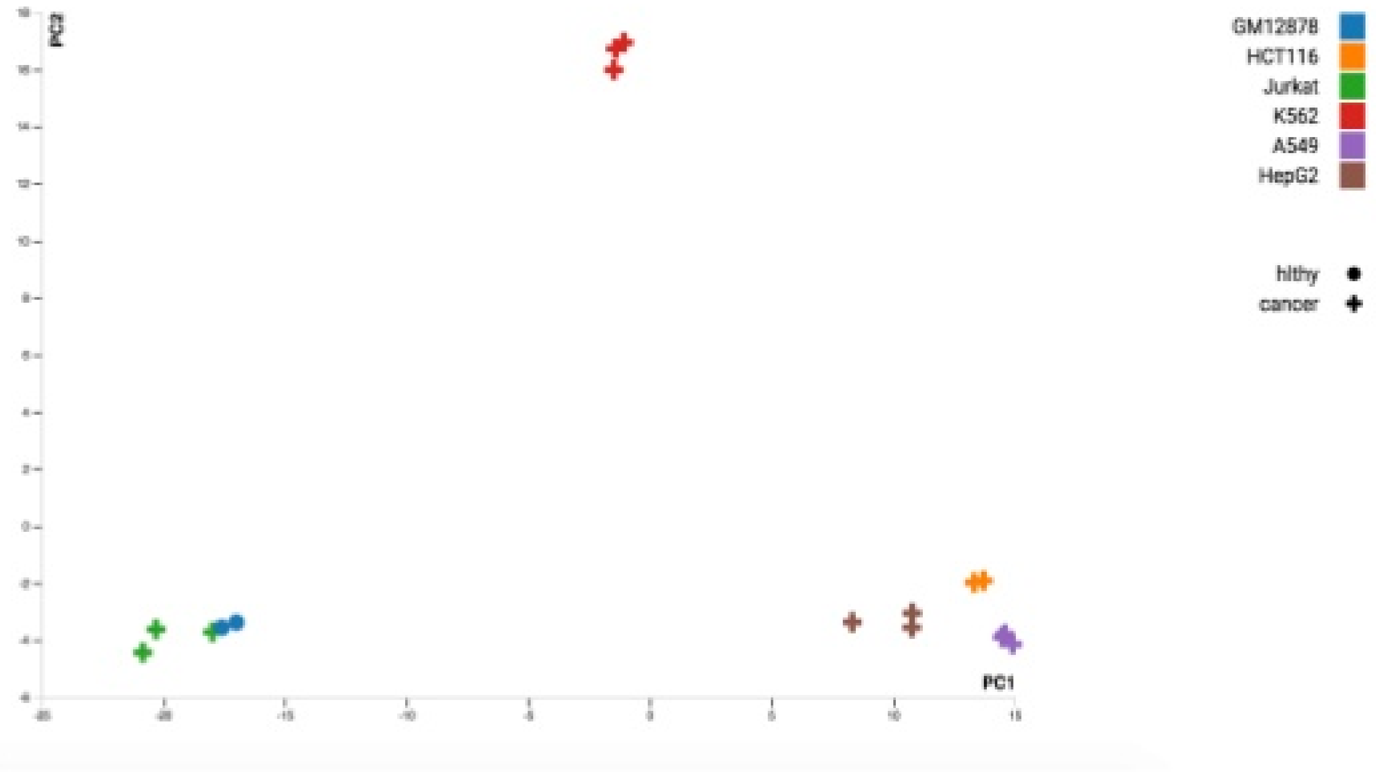
```
val graph = Graph(jointPeaks,
edges).partitionBy(PartitionStrategy.EdgePartition1D)
    val peakRatio = sc.broadcast(cutoffRatio)
    val subgraphs = graph.connectedComponents().vertices
jointPeaks.join(subgraphs)
      .map(item => item._2.swap)
      .redyceByKey()_
        .map(item => { ___.
```

Unsupervised Learning

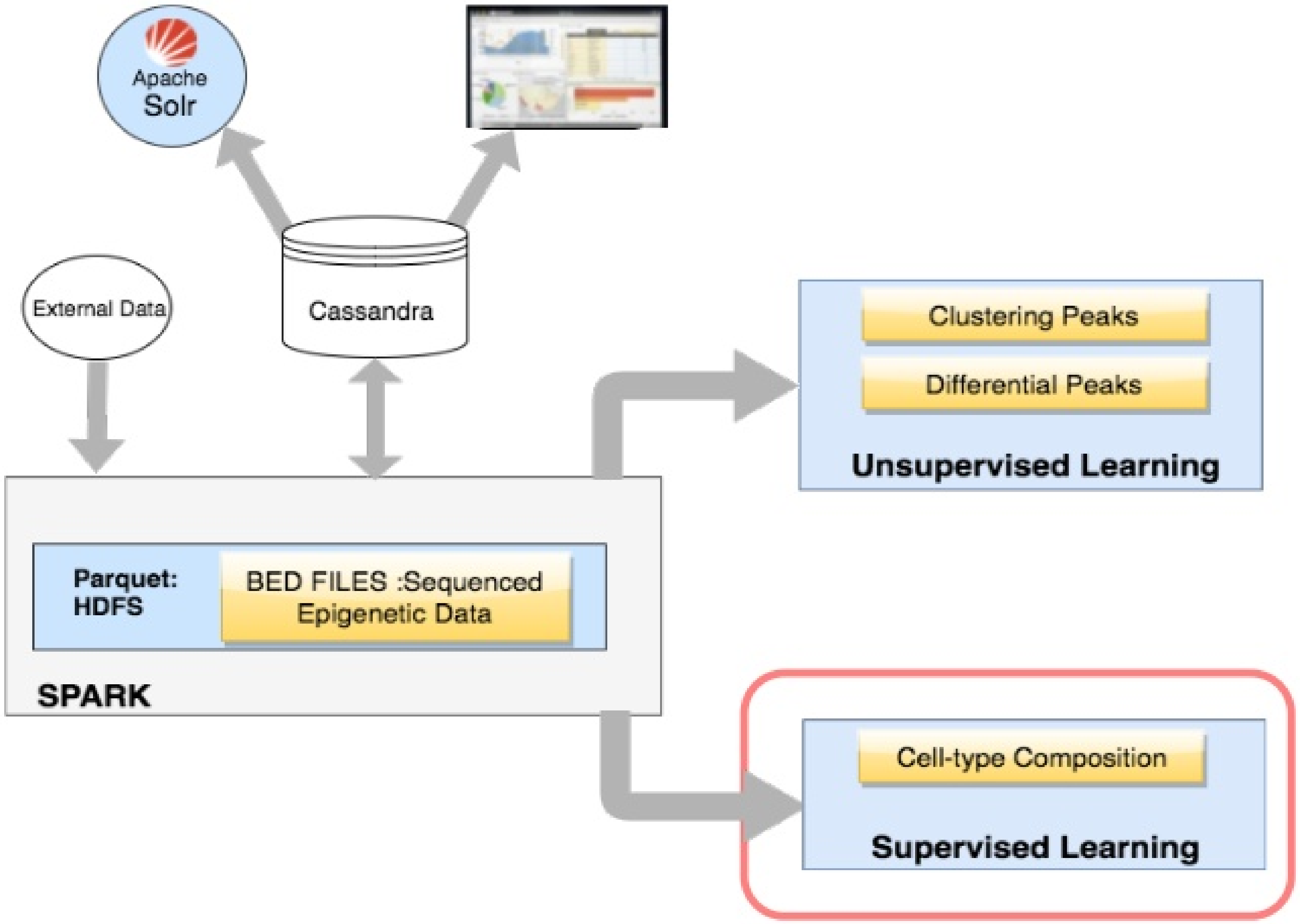


K-means and hierarchical clustering

Unsupervised Learning



Clustering similar datasets with PCA



Apache Solr

External Data

Cassandra

Clustering Peaks

Differential Peaks

Unsupervised Learning

Parquet: HDFS

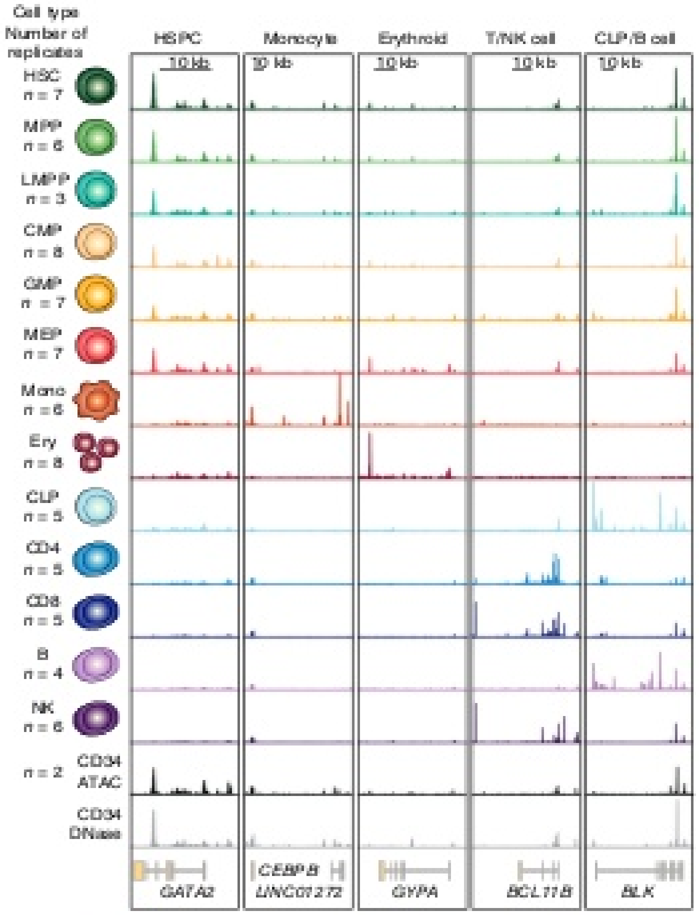
BED FILES :Sequenced Epigenetic Data

SPARK

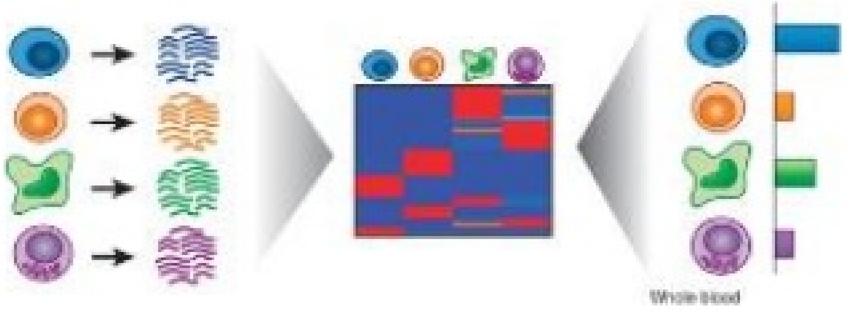
Cell-type Composition

Supervised Learning

Supervised Learning – Cell composition



Epigenome of each cell-type is unique fingerprint



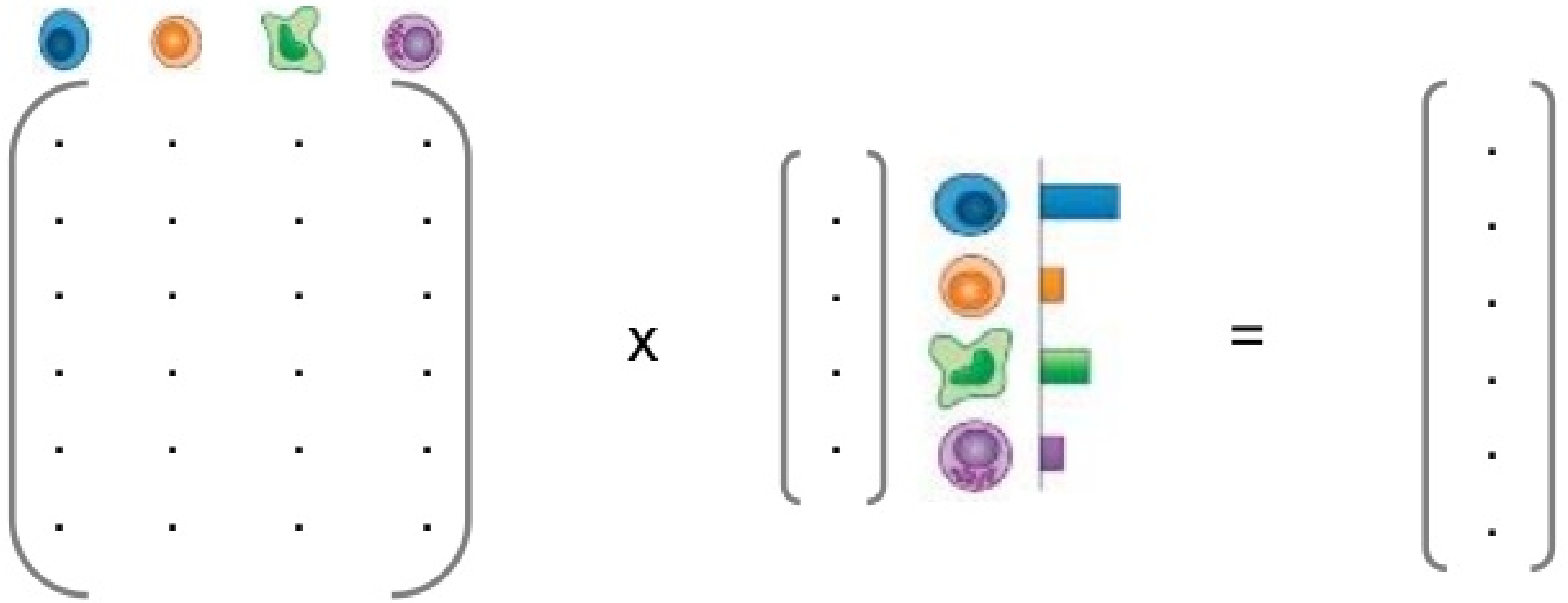
Mixed sample's signature can be deconvolved into pure cell type signals

Corces et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution

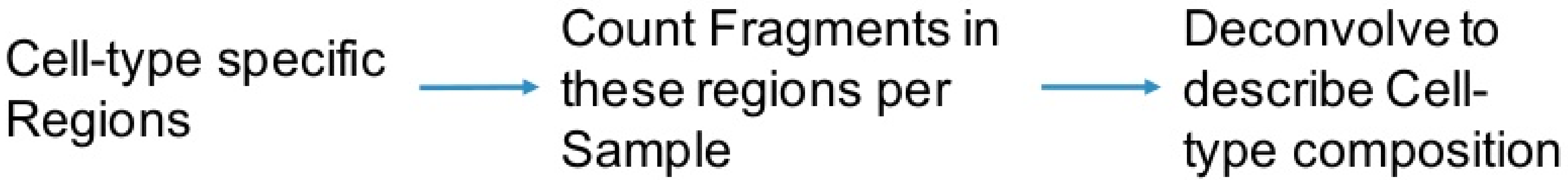
Supervised Learning – Cell composition

Cell type signature
Number of reads at
specific sites

Sample signature
Number of reads
at specific sites



Supervised Learning – Cell composition



Cell-type specific Regions

Count Fragments in these regions per Sample

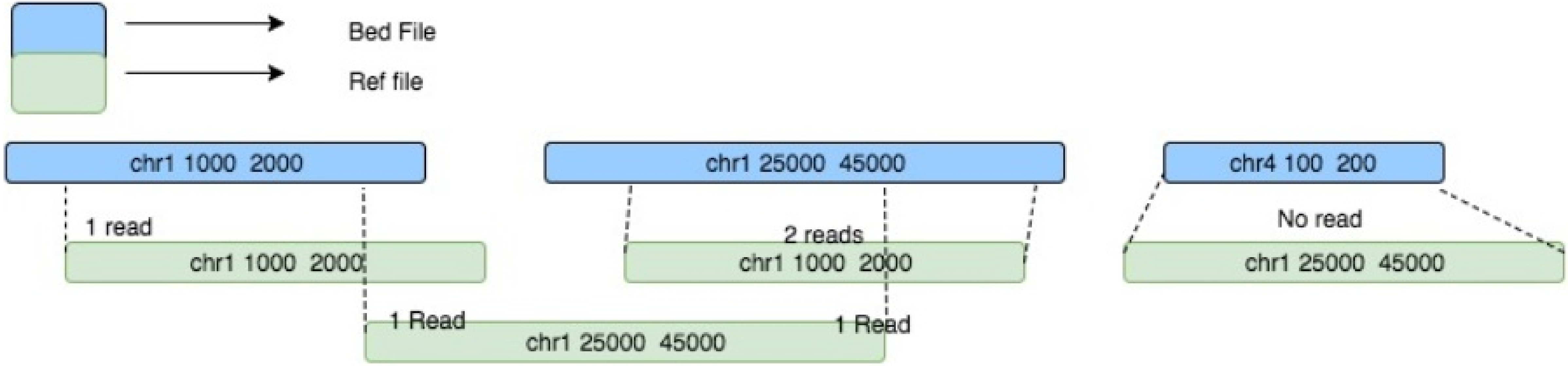
Deconvolve to describe Cell-type composition

Reference Regions

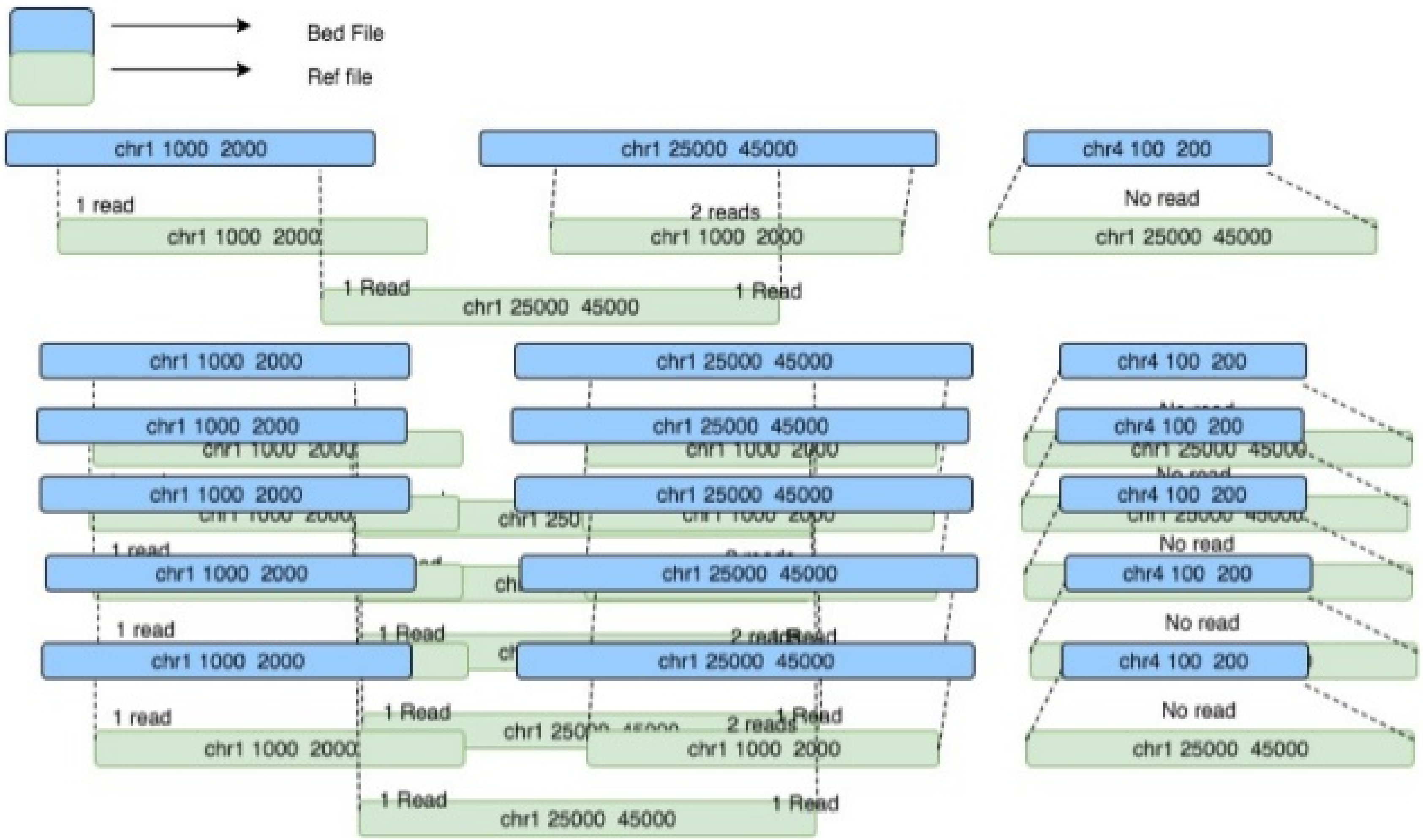
Clinical sample with mixed cells

Composition of Cells in Sample

Counting Reads within Windows



Counting Reads within Windows



🟦 → Bed File

🟩 → Ref file

chr1 1000 2000

1 read

chr1 1000 2000

1 Read

chr1 25000 45000

1 Read

chr1 25000 45000

2 reads

chr1 1000 2000

chr4 100 200

No read

chr1 25000 45000

chr1 1000 2000

1 read

chr1 1000 2000

1 Read

chr1 25000 45000

chr1 25000 45000

2 reads

1 Read

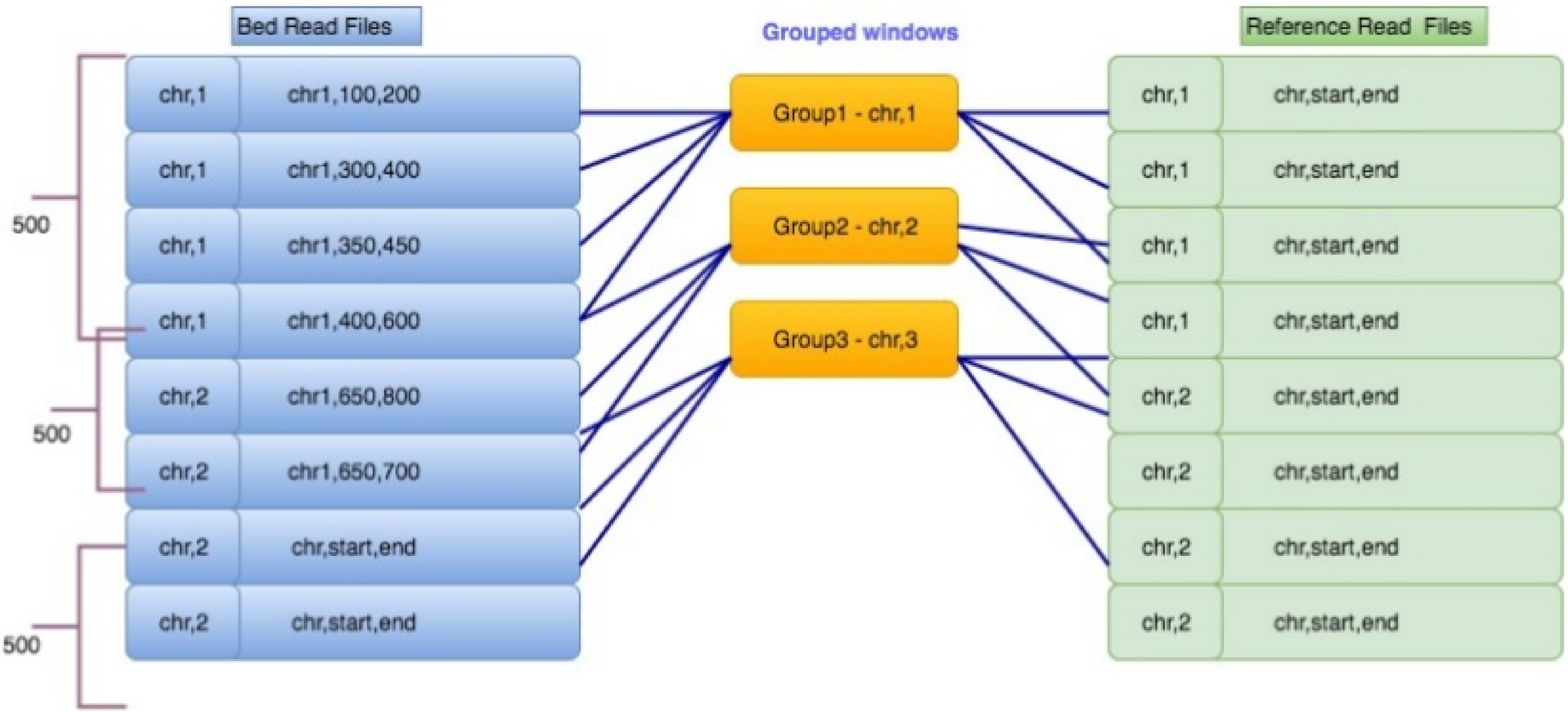
chr1 1000 2000

chr4 100 200

No read

chr1 25000 45000

Counting Reads – Range joins



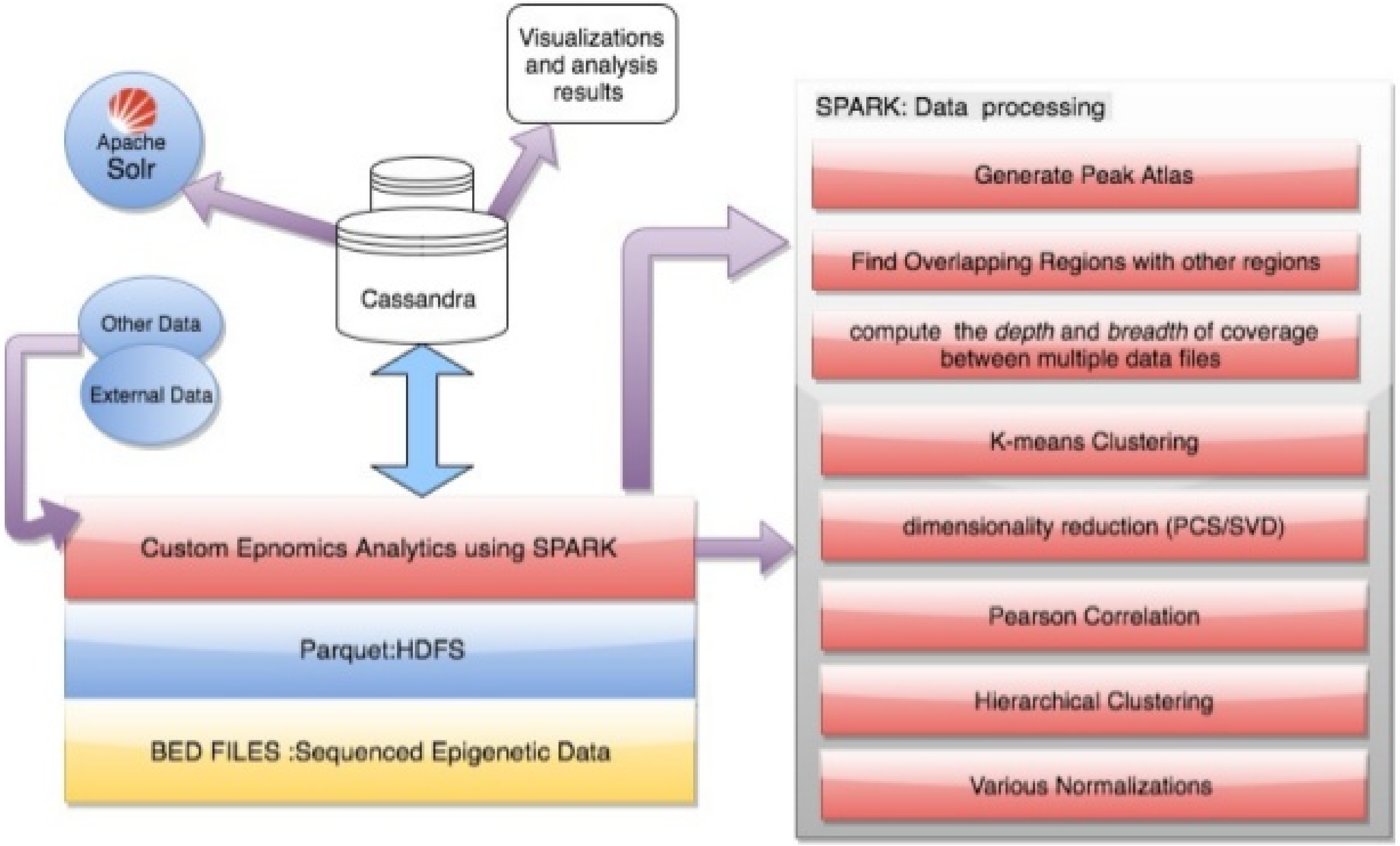
Bed Read Files

chr,1	chr1,100,200
chr,1	chr1,300,400
chr,1	chr1,350,450
chr,1	chr1,400,600
chr,2	chr1,650,800
chr,2	chr1,650,700
chr,2	chr,start,end
chr,2	chr,start,end

Reference Read Files

chr,1	chr,start,end
chr,1	chr,start,end
chr,1	chr,start,end
chr,1	chr,start,end
chr,2	chr,start,end
chr,2	chr,start,end
chr,2	chr,start,end
chr,2	chr,start,end

Building a Personalized Medicine Workflow


Visualizations and analysis results
Apache Solr
Cassandra
Other Data
External Data
Custom Epnomics Analytics using SPARK
Parquet:HDFS
BED FILES :Sequenced Epigenetic Data
SPARK: Data processing
Generate Peak Atlas
Find Overlapping Regions with other regions
compute the *depth* and *breadth* of coverage between multiple data files
K-means Clustering
dimensionality reduction (PCS/SVD)
Pearson Correlation
Hierarchical Clustering
Various Normalizations

Conclusion

Epinomics is building **a map of human health** through epigenomics.

ML pipelines combine Spark processing with traditional computing and algorithms.

Spark helps to process **tens of TB of genomic data** for personalized medicine applications.

SPARK
SUMMIT
2017

Thank You.

Anupama Joshi – anupama.joshi@gmail.com

Matei Negulescu – mnegules@uwaterloo.ca