Olha Horlova

PhD candidate, 3rd year

Advisor: **Stefano Ceri**

Co-advisor: **Abdulrahman Kaitoua**

Thesis submission: May 2020

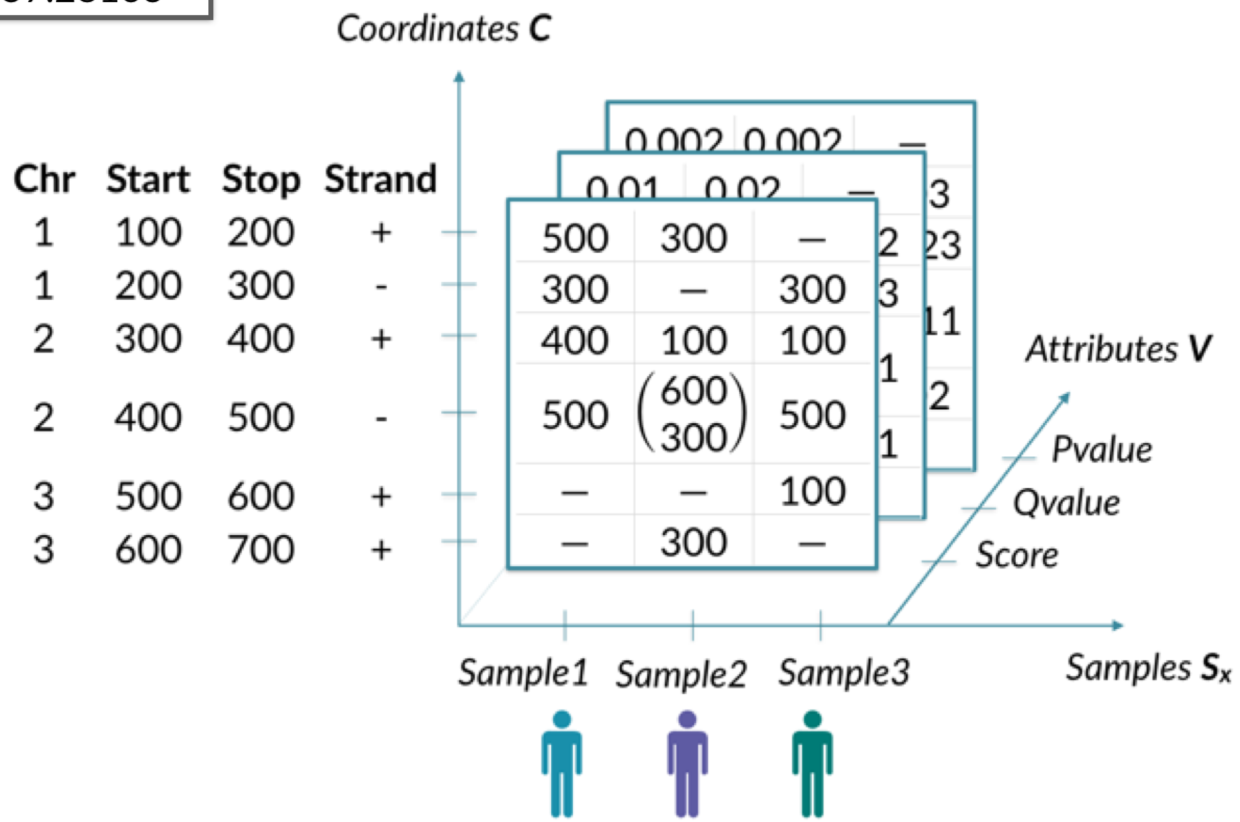
THE EVOLUTION OF GENOMIC DATA MODEL FOR THE CLOUD

Publications:

- **Olha Horlova**, Abdulrahman Kaitoua, Volker Markl, Stefano Ceri. Multi-Dimensional Genomic Data Management for Region-Preserving Operations. In Proceedings of the 35th IEEE International Conference on Data Engineering (ICDE 2019). Macau SAR, China. 8-11 April, 2019. DOI: https://doi.org/10.1109/ICDE.2019.00107
- **Olha Horlova**, Abdulrahman Kaitoua, Stefano Ceri. Array-based Data Management for Genomics. Accepted for the 36th IEEE International Conference on Data Engineering (ICDE 2020). Dallas, TX, USA. 20-24 April, 2020.
- Masseroli M, Canakoglu A, Pinoli P, Kaitoua A, Gulino A, Horlova O, Nanni L, Bernasconi A, Perna S, Stamoulakatou E, Ceri S. Processing of big heterogeneous genomic datasets for tertiary analysis of Next Generation Sequencing data. Bioinformatics (Oxford, England), 2018. DOI: https://doi.org/10.1093/bioinformatics/bty688

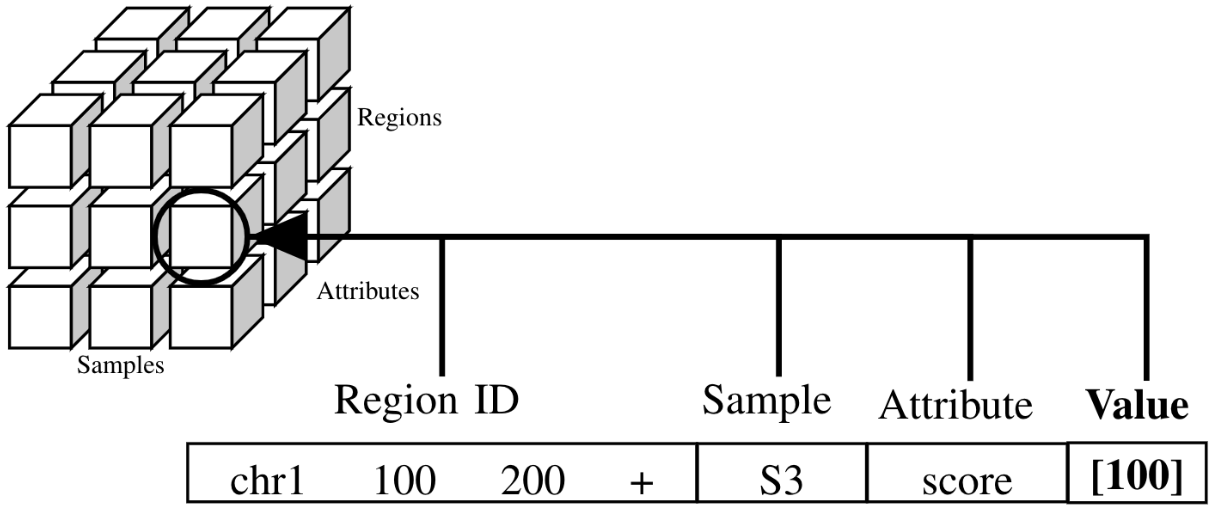
Genomic Data Model: Row to Array

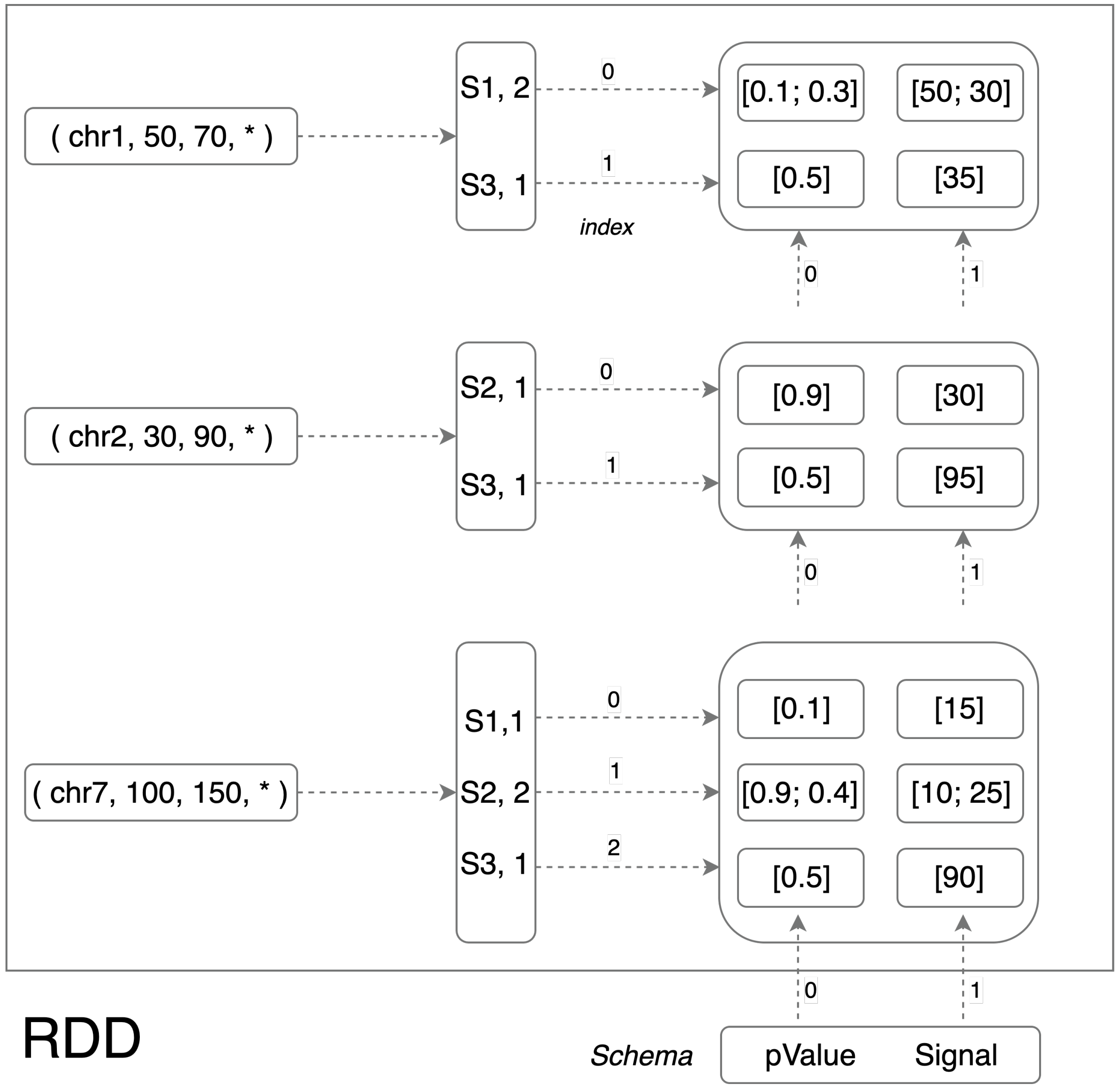
ID	CHR	START	STOP	STRAND	[Pvalue]
1	chr1	46149951	46155132	+	56.89183
1	chr1	49960544	49963154	+	61.99932
1	chr2	61242871	61247159	*	59.90095
2	chr1	161367656	161370907	*	56.08152
2	chr6	32934039	32944374	—	56.16523
3	chr4	175203948	175207120	*	56.97809
3	chr8	100023190	100026831	—	62.72371
3	chr20	34327086	34332343	*	57.28163



Chr	Start	Stop	Strand	Sample1	Sample2	Sample3
1	100	200	+	500	300	—
1	200	300	-	300	—	300
2	300	400	+	400	100	100
2	400	500	-	500	$\begin{pmatrix}600\\300\end{pmatrix}$	500
3	500	600	+	—	—	100
3	600	700	+	—	300	—

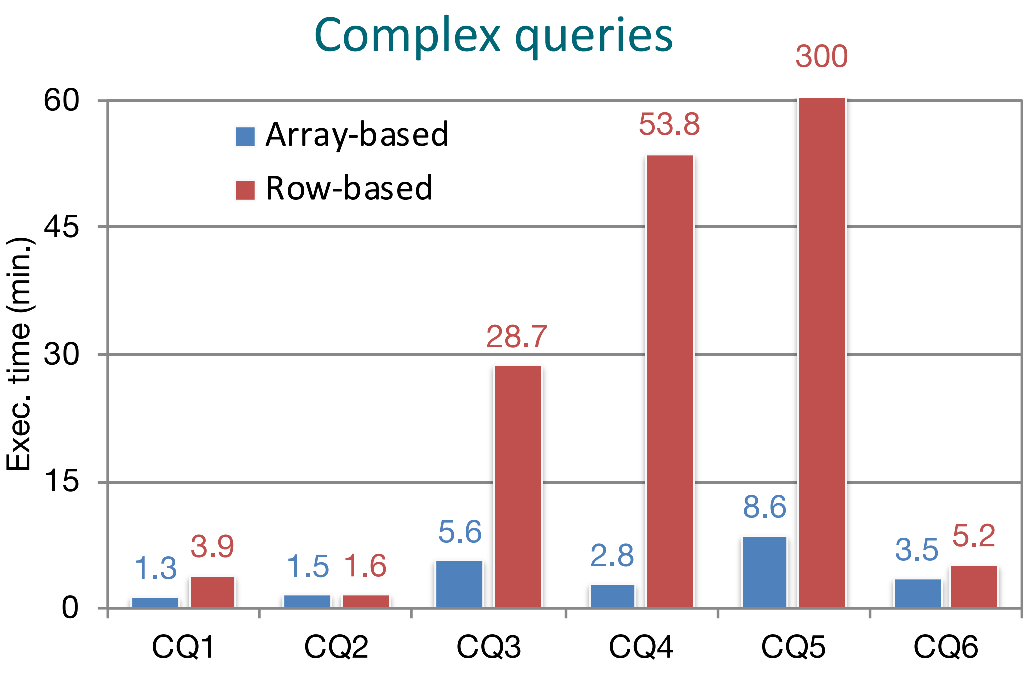
Spark RDDs organization to support array model

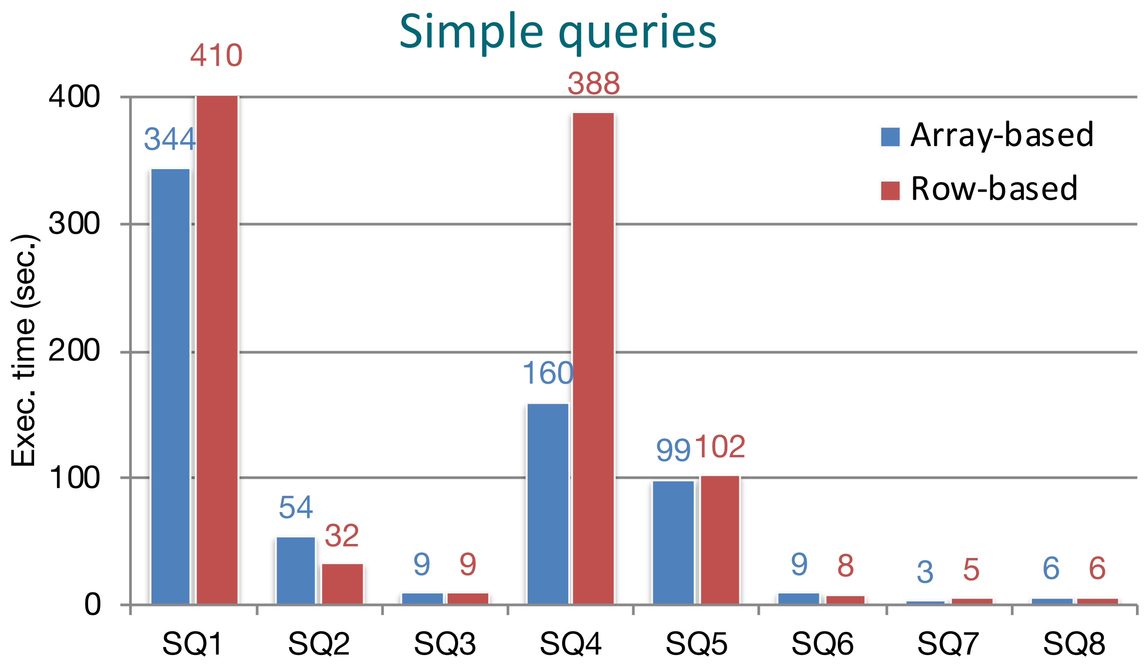


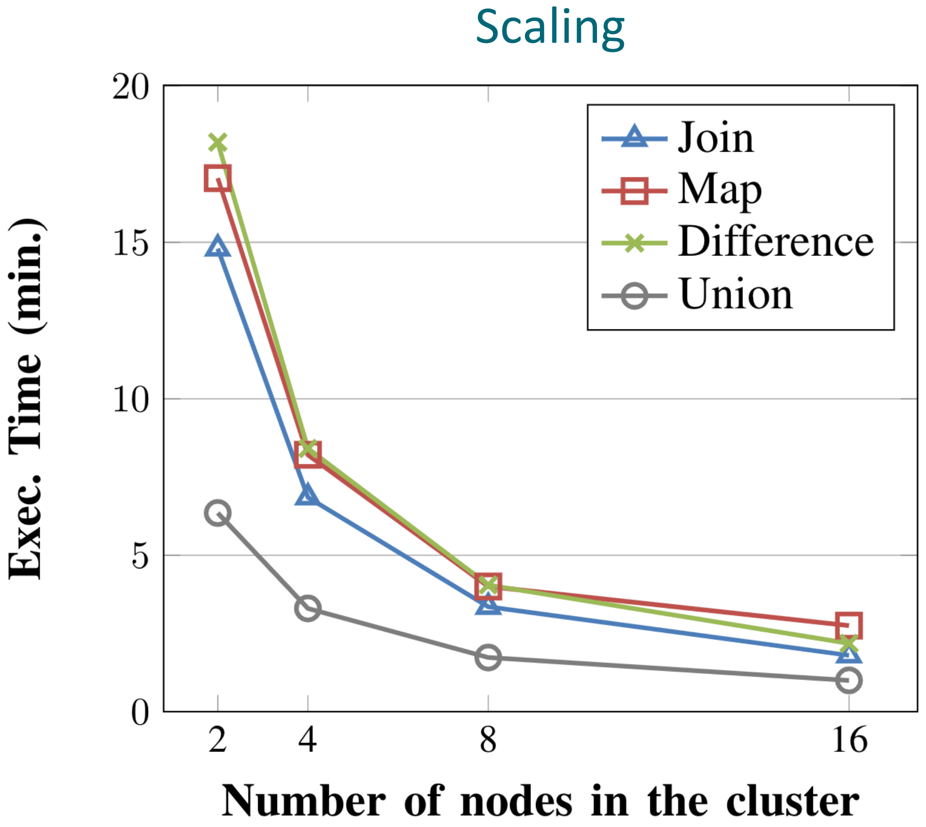


```
ArrayModel(key: RegionKey, value: RegionData)

RegionKey(chrom: String, start: Long, stop: Long, strand: Char)

RegionData(Replication: Array[(Long, Int)], Attribute: Array[Array[Array[GValue]]])
```

Model Evaluation


Complex queries


Simple queries


Scaling

Towards spatial and temporal applications

We can map genomic coordinates to:

- longitude and latitude of locations in spatial data
- time intervals of temporal data

Example #1: Find minimum distance offices of public or private organizations closest to given locations (e.g. for all banks, the closest bank office from home)

Example #2: Find the closest time events in different countries when a certain climate event occurred (e.g., for each nation/state/region, the event closest in time to Xmas 2019 when temperature was higher than 40 degrees Celsius)

THANK YOU!