

# **Find similarity in Galaxy tools and predict next tools in workflows**

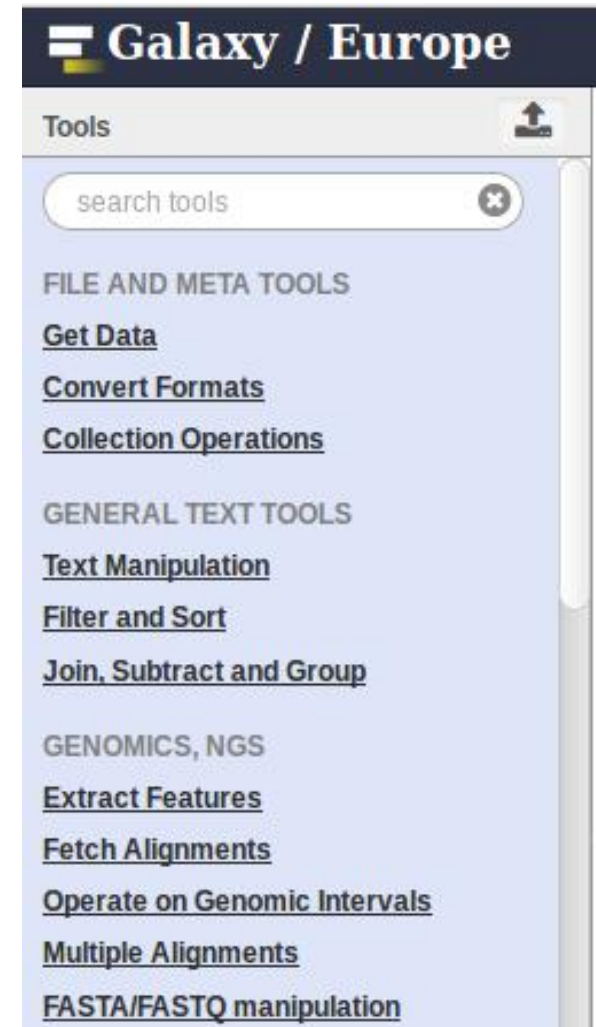
(Master's thesis)

Anup Kumar

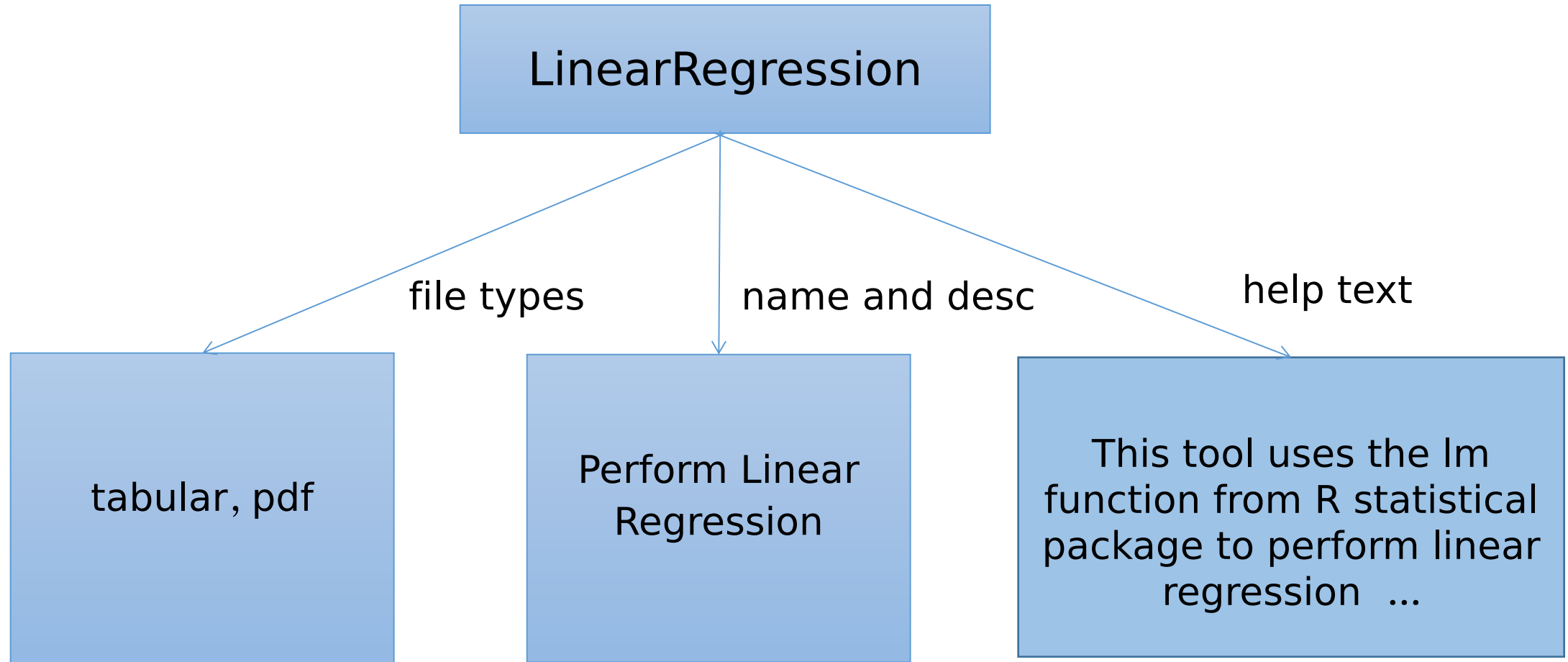
# Find similarity in Galaxy tools

Machine learning (ML) and natural language processing (NLP) approaches

- Paragraph Vectors
- Gradient Descent



# Tool's attributes



# Approach

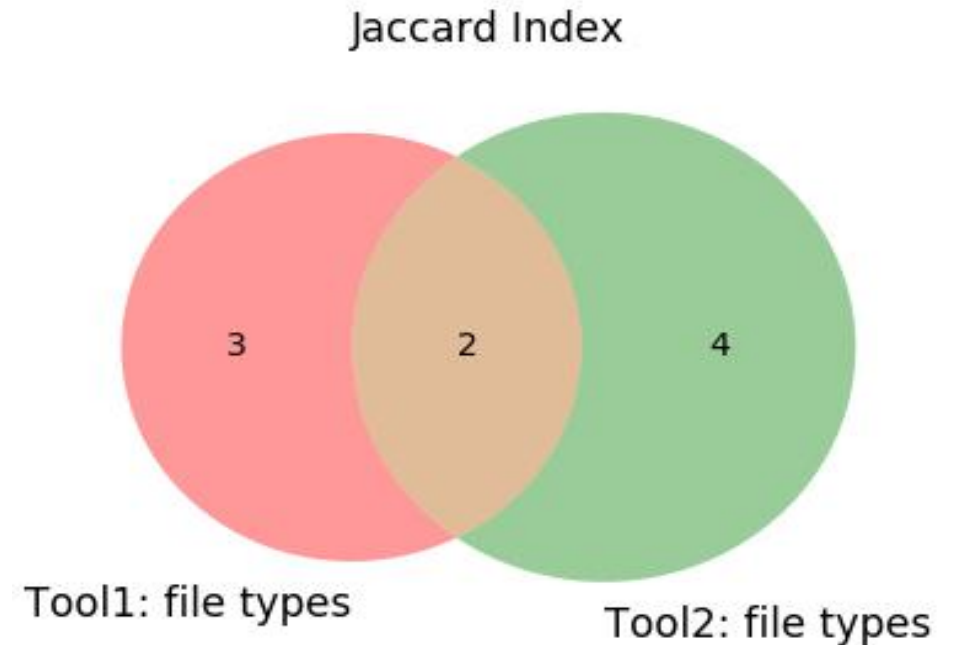
- Extract tool's attributes
- Clean text
- Create sets of tokens
- Learn similarities
- Combine optimally
- Visualize

# Tokens

Attributes/ Tools	LinearRegression	LogisticRegression	Similarity
Input, output	<b>'pdf'</b> , <b>'tabular'</b>	<b>'tabular'</b>	?
Name, description	<b>'regress'</b> , <b>'linear'</b> , <b>'perform'</b>	<b>'logist'</b> , <b>'regress'</b> , <b>'perform'</b>	?
Help text	<b>'regress'</b> , <b>'assumpt'</b> , <b>'lm'</b> , <b>'statist'</b> , <b>'linear'</b> ...	<b>'vif'</b> , <b>'regress'</b> , <b>'glm'</b> , <b>'car'</b> , <b>'inflat'</b> , <b>'function'</b> , <b>'statist'</b> , <b>'logist'</b> ...	?

# Compute similarity

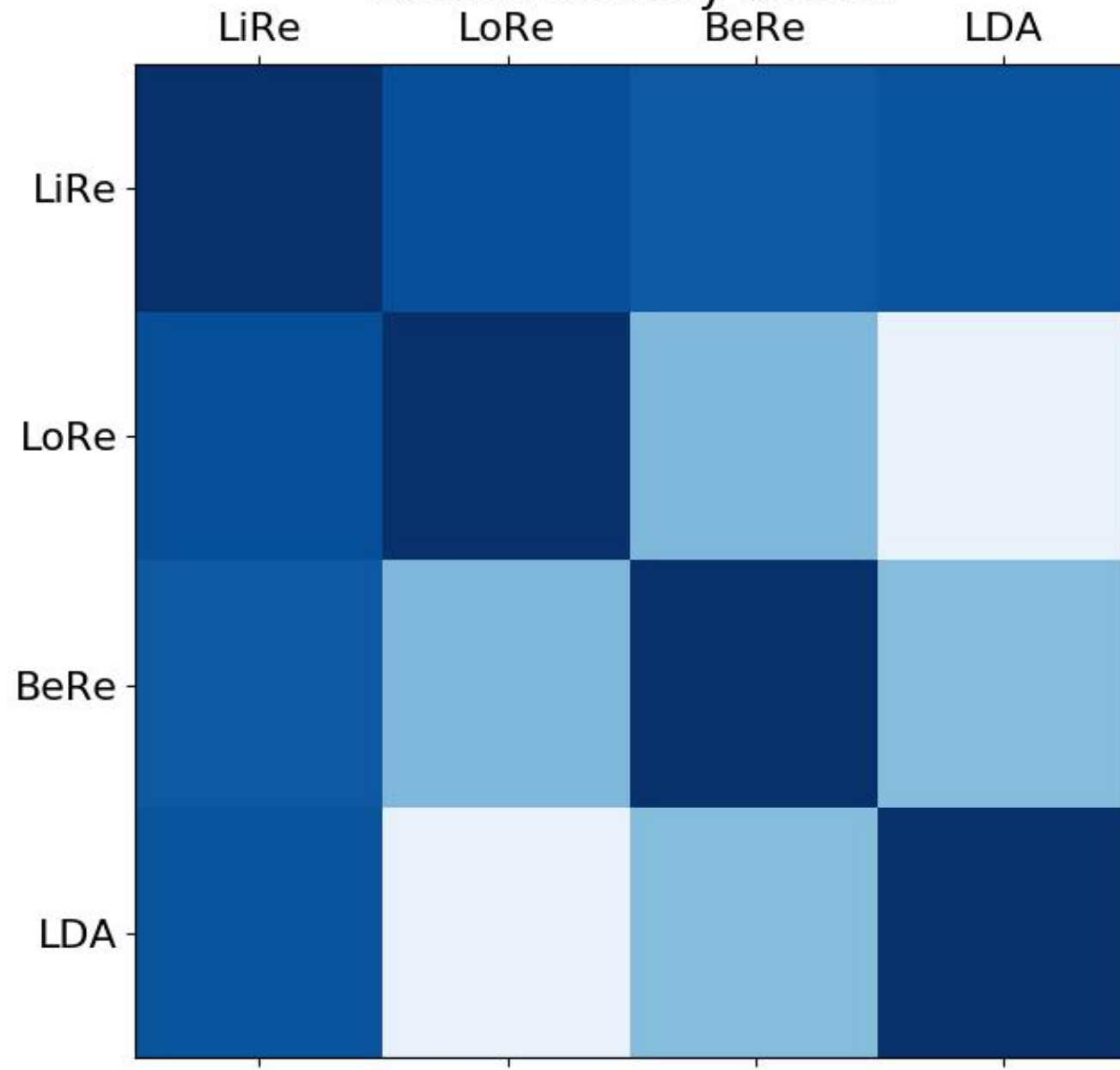
- Compute Jaccard Index for input/output



- Learn dense vectors for name, description and helptext\*
- ['regress', 'linear', 'perform'] = [ 0.98, 0.07, ... , 0.12 ]
- Compute cosine distance between dense vectors

\*[[https://cs.stanford.edu/~quocle/paragraph\\_vector.pdf](https://cs.stanford.edu/~quocle/paragraph_vector.pdf)]

Tools similarity matrix



- LiRe - Linear Regression
- LoRe - Logistic Regression
- BeRe - BestSubsetsRegression
- LDA - LDA Analysis




# How to combine ?

- 3 similarity matrices, one for each attribute
- How to combine them ? Take average ?
- Optimal combination, learn weights for each tool
- Similarity:

$$\arg \max_{(w_1, \dots, w_n)} \sum_{i=1}^N w_i \cdot s_i$$



# Optimization

s1	Input, output				s2	Name, desc.				s3	Helptext			
	1	0.34	0.65	0.9		1	0.56	0.6	0.9		1	0.6	0.90	0.7
	0.34	1	.66	...		0.56	1	...	...		0.6	1	...	...
	0.65	0.66	1	...		...	...	1	...		0.9	...	1	...
	0.9	...	...	1		...	...	...	1		0.7	...	...	1

$$\text{Minimize } ([1.0, 1.0, 1.0, \dots, 1.0] - [w1 \cdot s1 + w2 \cdot s2 + w3 \cdot s3])$$

$$\text{where } w1 + w2 + w3 = 1$$

# Example

- Tool: LinearRegression
- Similarity for input/output:  $\mathbf{s1} = [ 0.8, 0.5, 1.0, \dots ]$
- Similarity for name, desc:  $\mathbf{s2} = [ 0.09, 0.09, 0.005, \dots ]$
- Similarity for helptext:  $\mathbf{s3} = [ 0.45, 0.36, 0.001 \dots ]$
- Optimal weights:  $w1 = 0.55, w2 = 0.10, w3 = 0.35$
- Similarity:

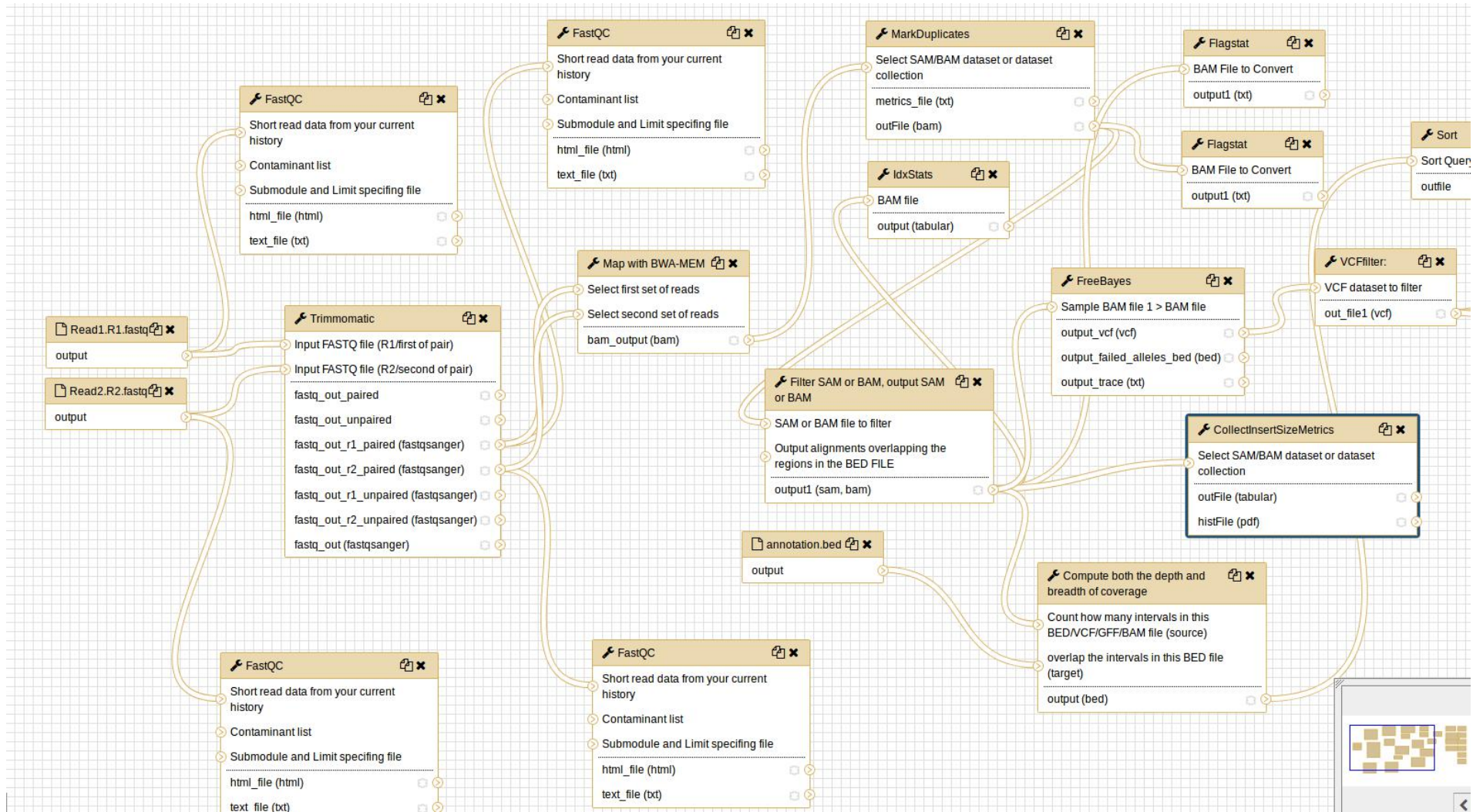
$$[w1 \cdot s1 + w2 \cdot s2 + w3 \cdot s3]$$

# Visualizer and References

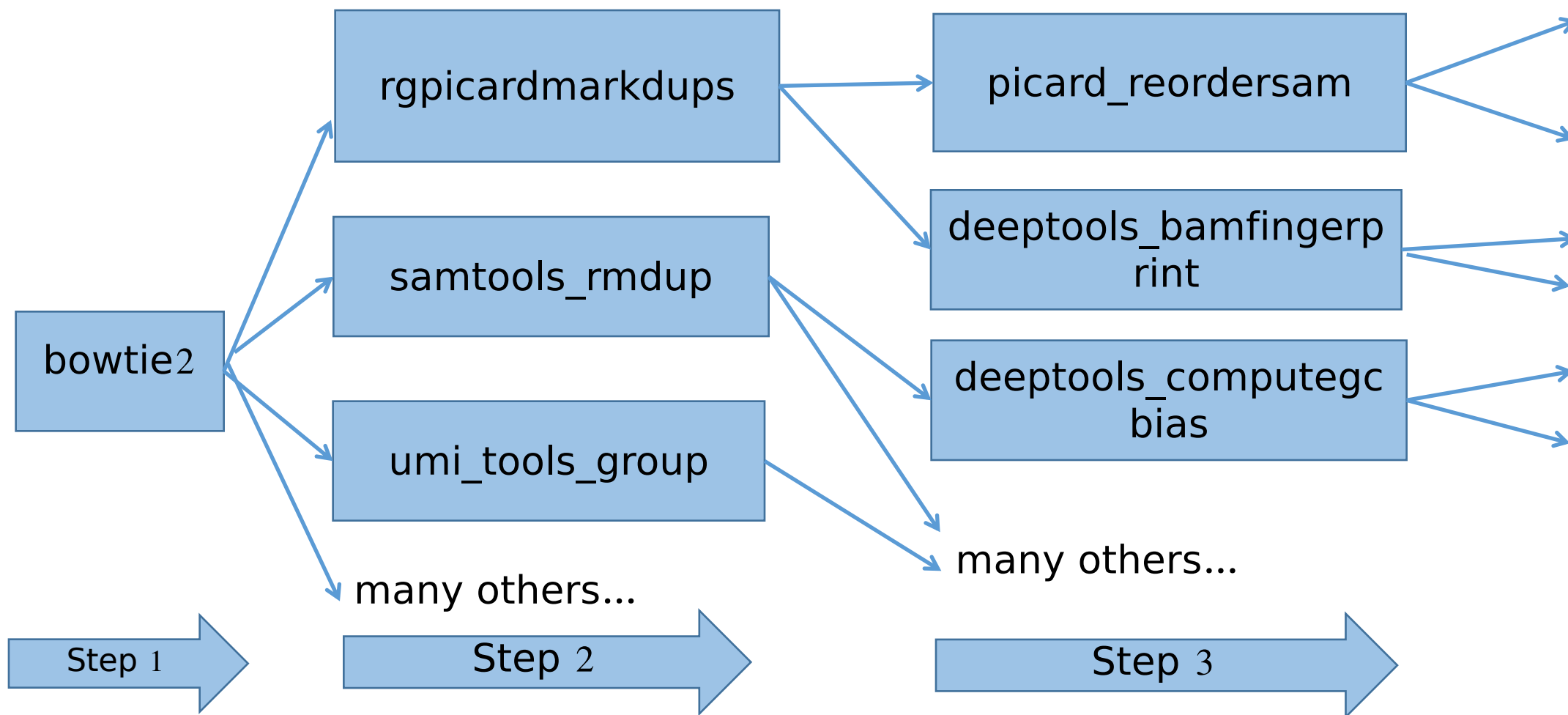
- Static website: results for  $\sim 1000$  tools
- [https://rawgit.com/anuprulez/similar\\_galaxy\\_tools/master/viz/similarity\\_viz.html](https://rawgit.com/anuprulez/similar_galaxy_tools/master/viz/similarity_viz.html)
- [https://github.com/anuprulez/similar\\_galaxy\\_tools](https://github.com/anuprulez/similar_galaxy_tools)
- [https://cs.stanford.edu/%7Equocle/paragraph\\_vector.pdf](https://cs.stanford.edu/%7Equocle/paragraph_vector.pdf)
- <https://arxiv.org/pdf/1607.05368.pdf>

# **Predict next tools in Galaxy workflows**

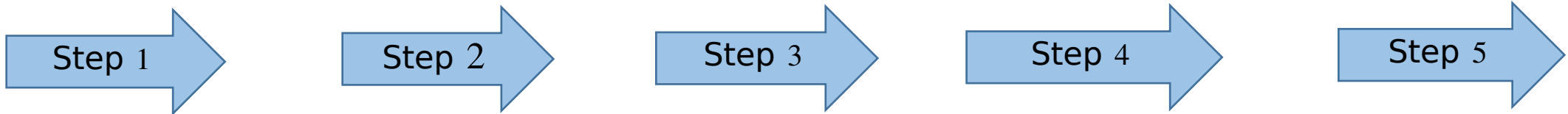
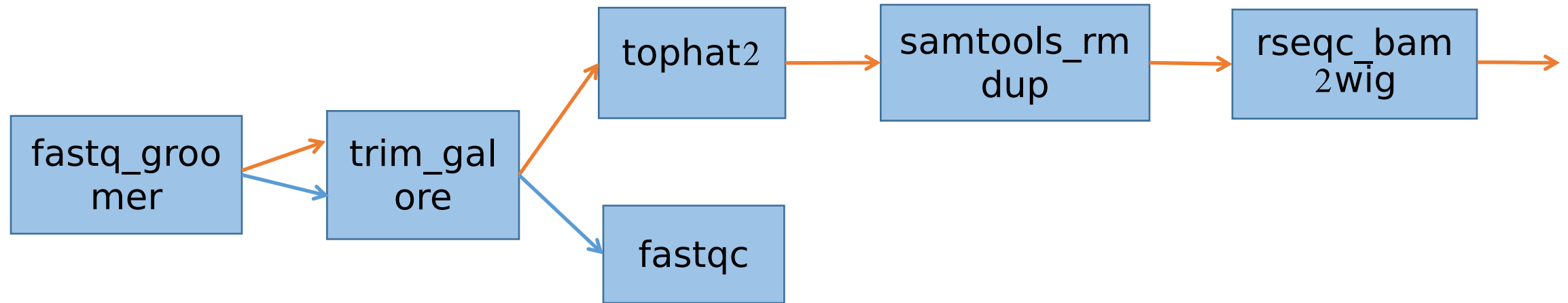
# Galaxy workflow



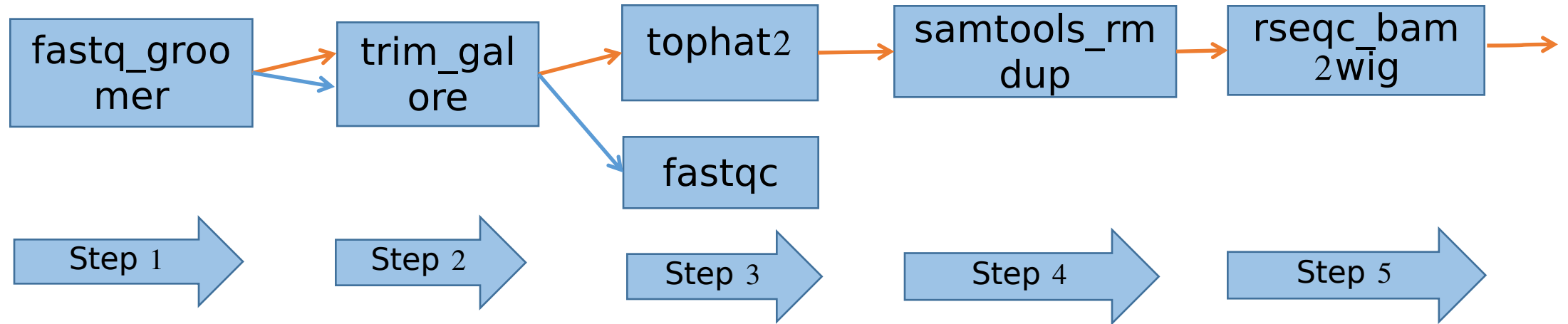
# Next tools ?



# Workflow as a sequence



# Data preprocessing



- fastq\_groomer, trim\_galore (Step 1)
- fastq\_groomer, trim\_galore, tophat2, fastqc (Step 2)
- fastq\_groomer, trim\_galore, tophat2, samtools\_rmdup (Step 3)
- fastq\_groomer, trim\_galore, tophat2, samtools\_rmdup, rseqc\_bam2wig (Step 4)

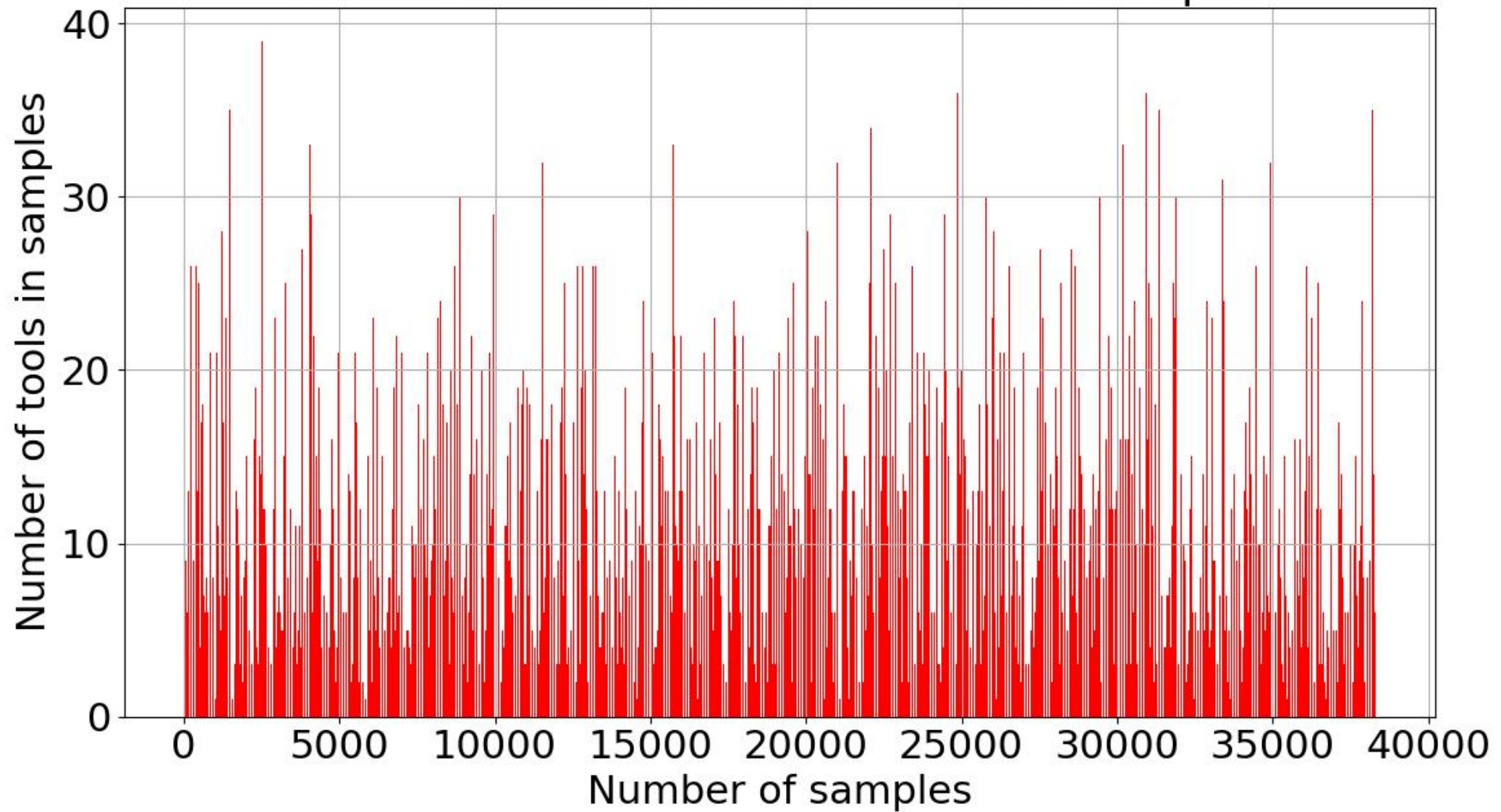


# Data preprocessing

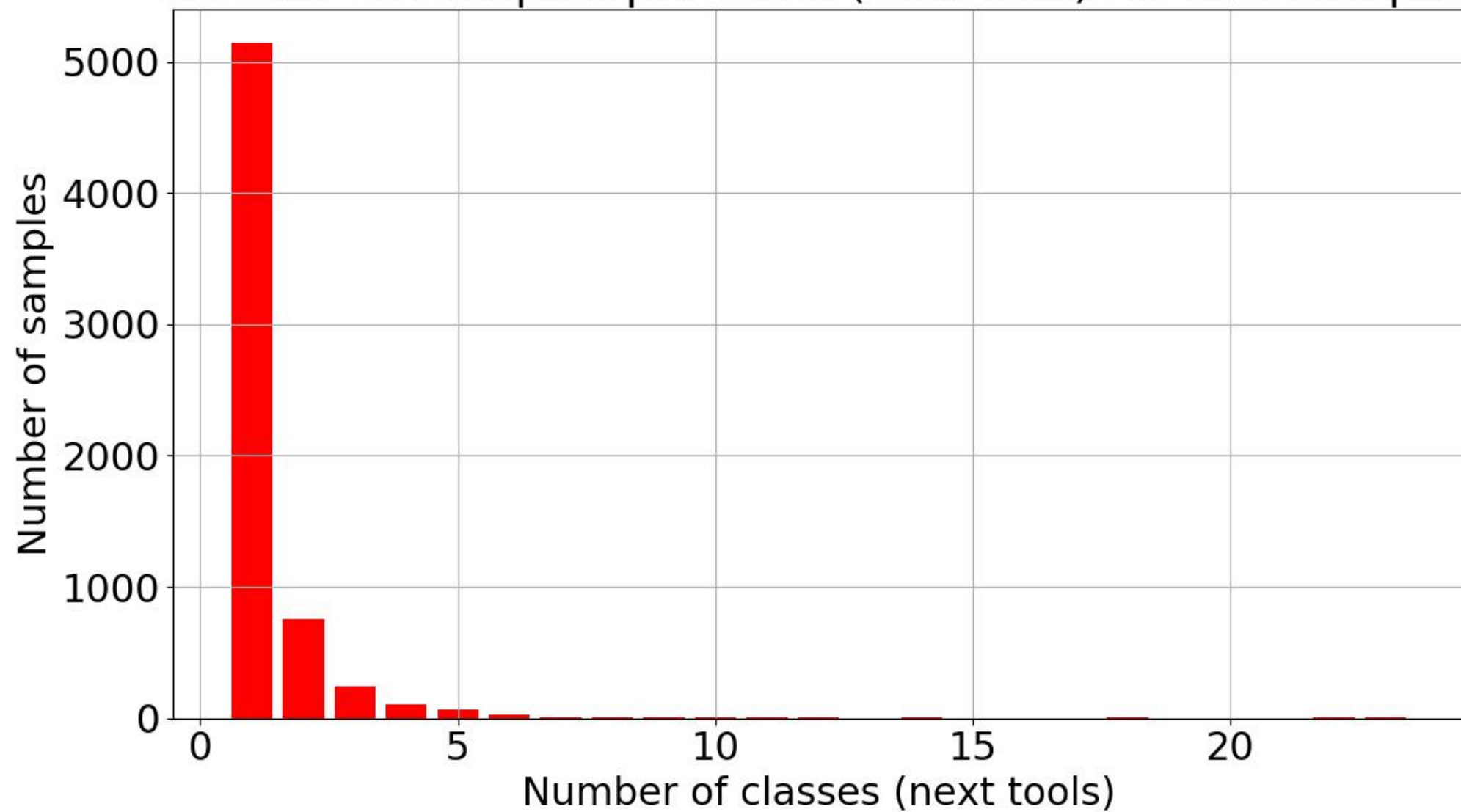
‘fastq\_groomer’: 1, ‘trim\_galore’: 2, ‘tophat2’: 3,  
‘samtools\_rmdup’: 4, ‘rseqc\_bam2wig’: 5, ‘fastqc’: 6

Sample	Label (next tool(s)/classes)
1,2 (fastq_groomer, trim_galore)	3, 6 (tophat2, fastqc)
1,2,3	4
....	....
....	....

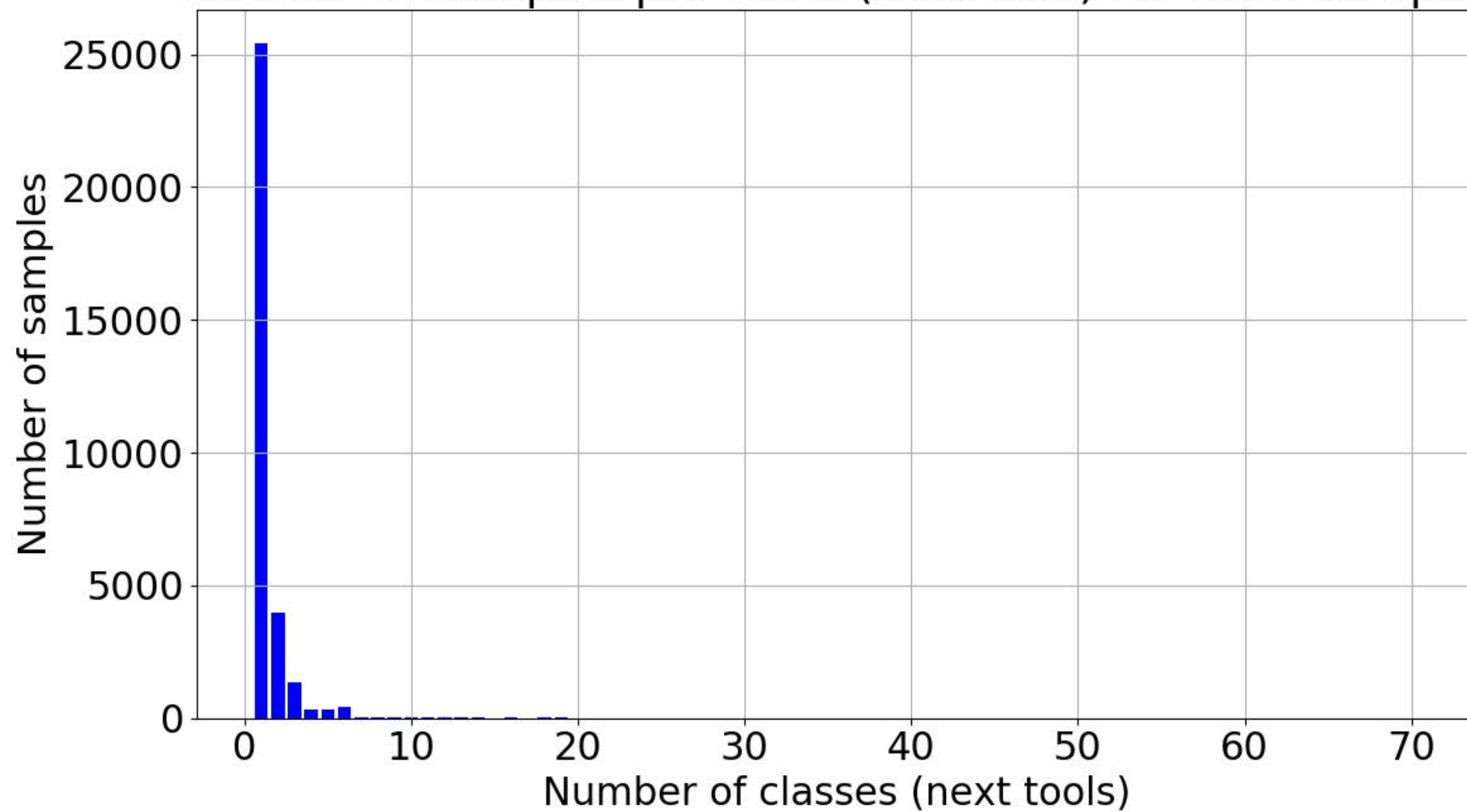
Distribution of number of tools in samples



Number of samples per class (next tool) for test samples

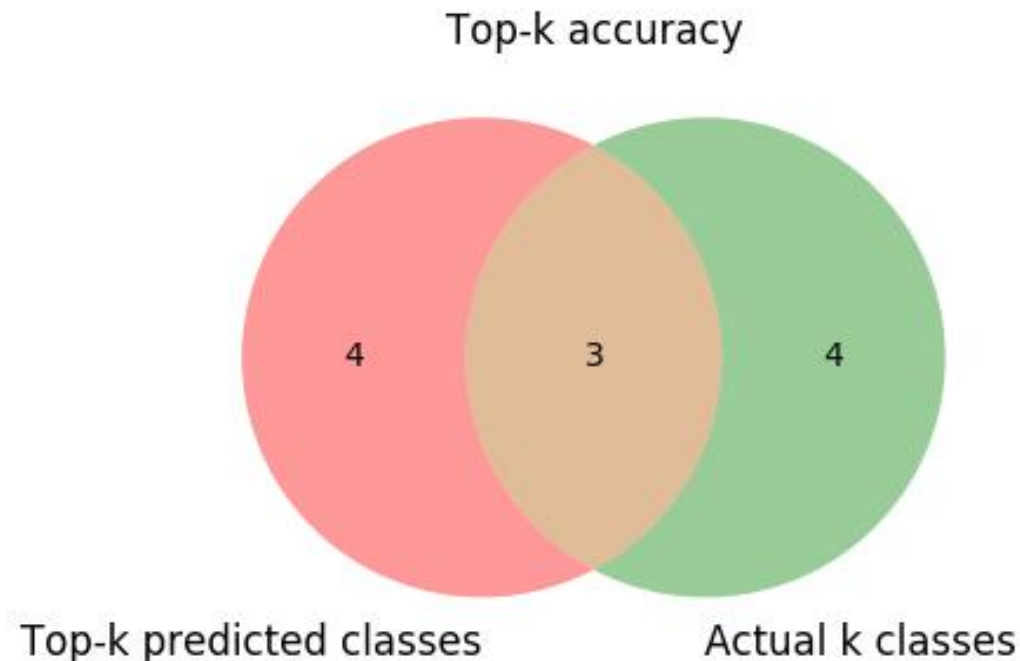


Number of samples per class (next tool) for train samples

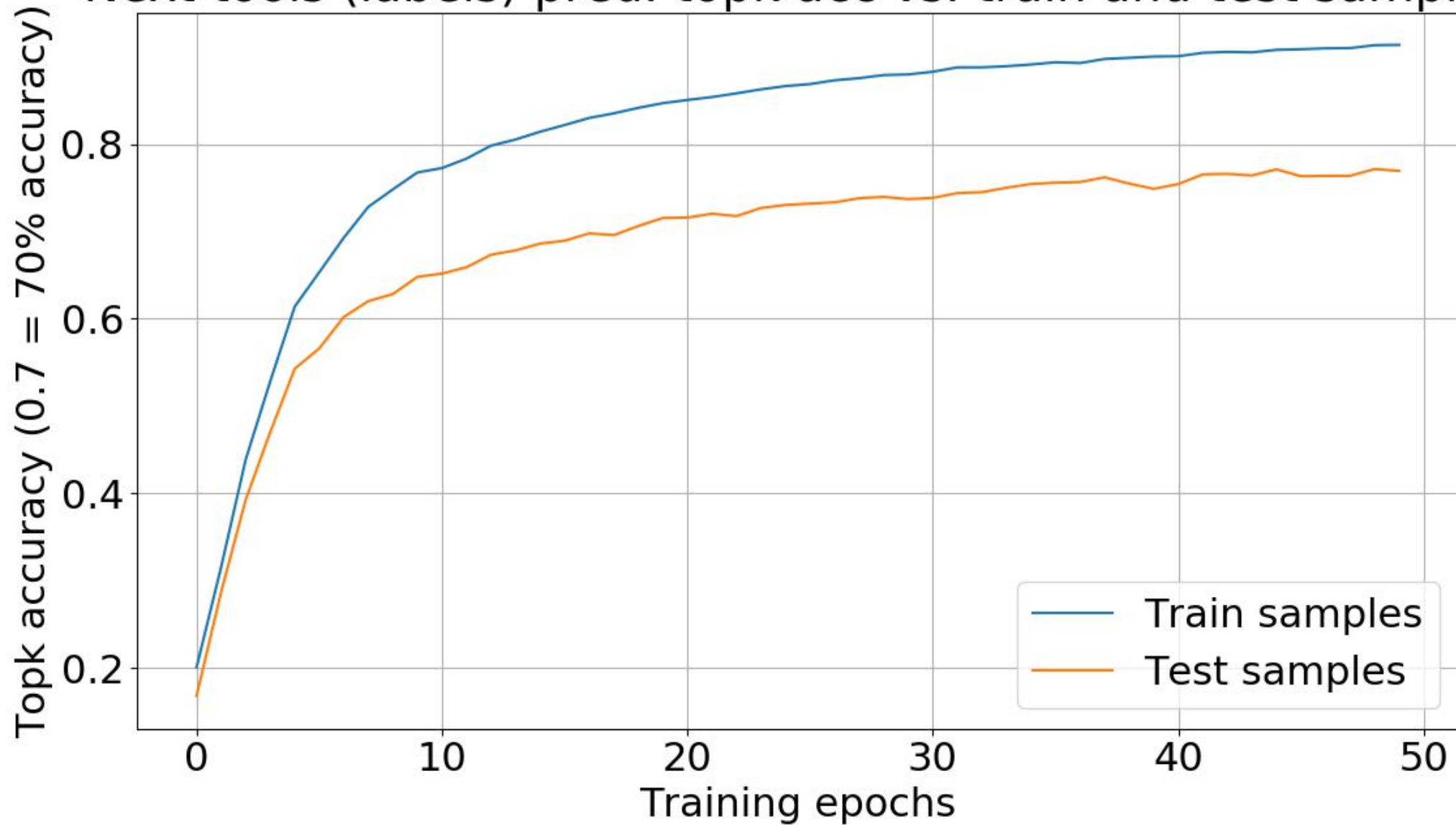


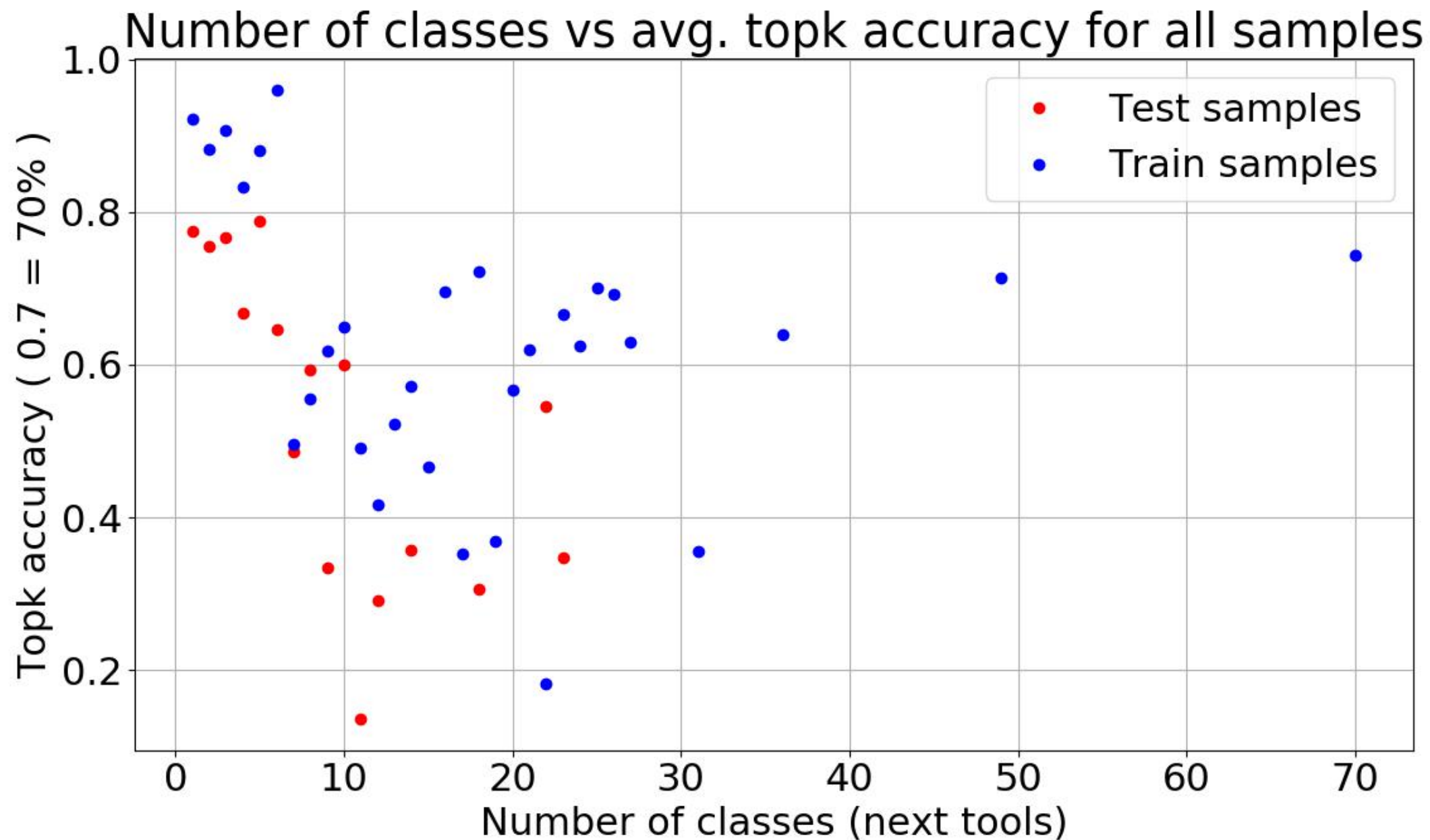
# Classification

- Multi label, multi class classification
- Long short term memory (LSTM) networks
- Topk accuracy



Next tools (labels) pred. topk acc vs. train and test samples





# References

- [https://github.com/anuprulez/similar\\_galaxy\\_workflow](https://github.com/anuprulez/similar_galaxy_workflow)
- <https://arxiv.org/pdf/1511.03677.pdf>
- <https://arxiv.org/pdf/1604.04573.pdf>
- <https://arxiv.org/pdf/1506.00019.pdf>



**Thank you for your attention**

**Questions ?**