

The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)

Determination of Accuracy and Probability in the Analysis of Large-Scale Biomedical Data

Stella Vetova

Primary funding for the presented work was provided by the National Science Fund, Ministry of Education and Science, Republic of Bulgaria under contract **KP-06-N37/24**, research project “Innovative Platform for Intelligent Management and Analysis of Big Data Streams Supporting Biomedical Scientific Research”.

**The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)**

Large Scale Data

1.Generated heterogeneous amount of data used for further processing and analysis.

2.Features of Big Data :

- complexity;
- variety;
- volume ;
- application opportunities.

**The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)**

Areas of Large-Scale Data Application

-science, business, medicine, biomedicine,
bioinformatics, agriculture, biology;

Gathering Large-Scale Data Application

-laboratory results, clinical test, patients' exams,
symptoms data captured by the means of
telemedicine,
etc.

Workflow samples

Definition for workflow - a sequence of functions defined to perform a single task;

Advantages:

1. computes complex tasks;
2. includes data visualization and analysis based on the principles of segmentation, diagnosis and therapy;

Workflow Samples

3. Workflows are applied for complex automatic analysis to improve the interpretation and reporting, reducing time and providing ease of decision making in the process of disease diagnosis and treatment, classification and detection;

Methods, Algorithms and Tools for Large-Scale Data Workflow Analysis in Biomedicine

Data-Classification	Clustering	k-medoids algorithm;
		k-Means-partitioning algorithm;
		k-Nearest-Neighbor (KNN);
		Unweighted-Pair-Group-Method-with-Arithmetic-Mean (UPGMA);
		Neighbor-Joining (NJ) method;
		Fitch-and-Kitsch method;
	Distance-computation	Euclidean distance;
		Hamming distance;
		Manhattan distance;
		Minkowski distance;
Data-Learning	Deep-Learning	Convolutional-Neuron-Networks (CNN);
		Deep-Boltzmann-Machine;
		Self-Organizing-Map (SOM);

Issue	Approach	Methods
Dimensionality-reduction	Mapping methods	Principal-Component-Analysis (PCA);
		Singular-value-decomposition;
	Non-Linear Mapping methods	Kernel-Principal-Component-Analysis;
		Laplacian-eigenmaps;
		Sammon's mapping;

Methods, Algorithms and Tools for Large-Scale Data Workflow Analysis in Biomedicine

Workflow platforms and design tools	Galaxy
	myExperiment
	Taverna
	MapReduce
	Hadoop
	Spark
	GraphLab
	Pregel
	Closha
	Preglix
	Pegasus
	Kepler
	Chipster
	Mike
	GraphFlow
	Gromacs

Storage Environment	Cloud storage
	Local storage
Programming Languages	Bash
	R
	Perl
	Python

**The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)**

Large-Scale data visualization tools

BIFLOWS

Open-source web tool for bioimage analysis workflows

•Main Features:

- 1. Import of image databases containing annotation and their organization as bioimage analysis tasks;
- 2. Bioimage analysis workflow encapsulation;
- 3. Image processing and visualization in combination with the results;
- 4. Automatic evaluation of the workflows performance.

STU-tool

R-based STU tool for bioinformatics

•Main features:

- 1. Identifies spatial patterns alignment of tissue images and performs visualization;
- 2. Works with RNA count and images;
- 3. Constructs a 3D model of the tissue on the base of cell segmentation.

**Java Script web-
based API**

Web solution for large-scale data processing and visualization

•Main tasks:

- 1. Input data processing and change;
- 2. Storing the computations result as an output variable.

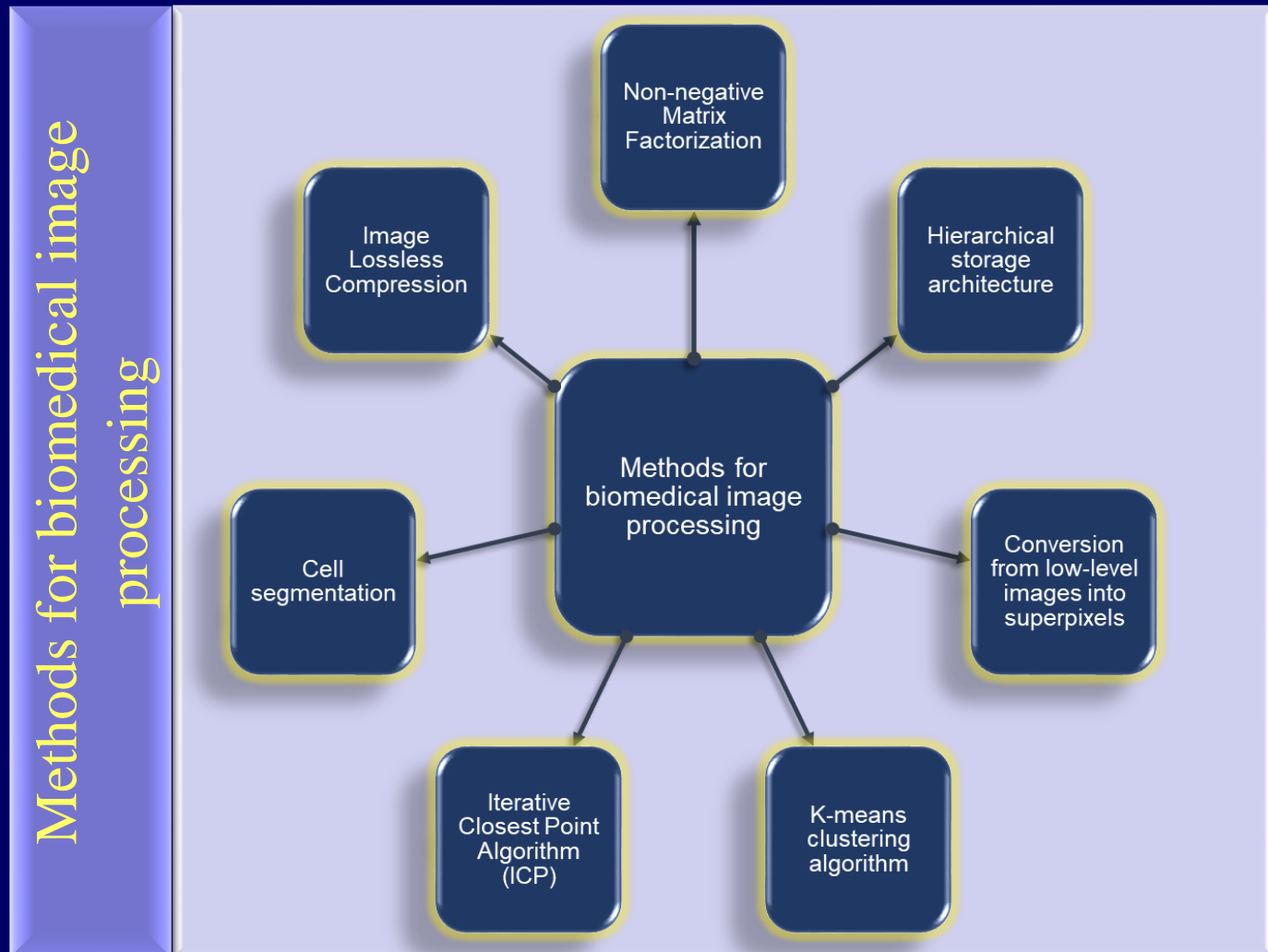
FAN-C

A framework for matrix generation, analysis and visualization in the field of bioinformatics

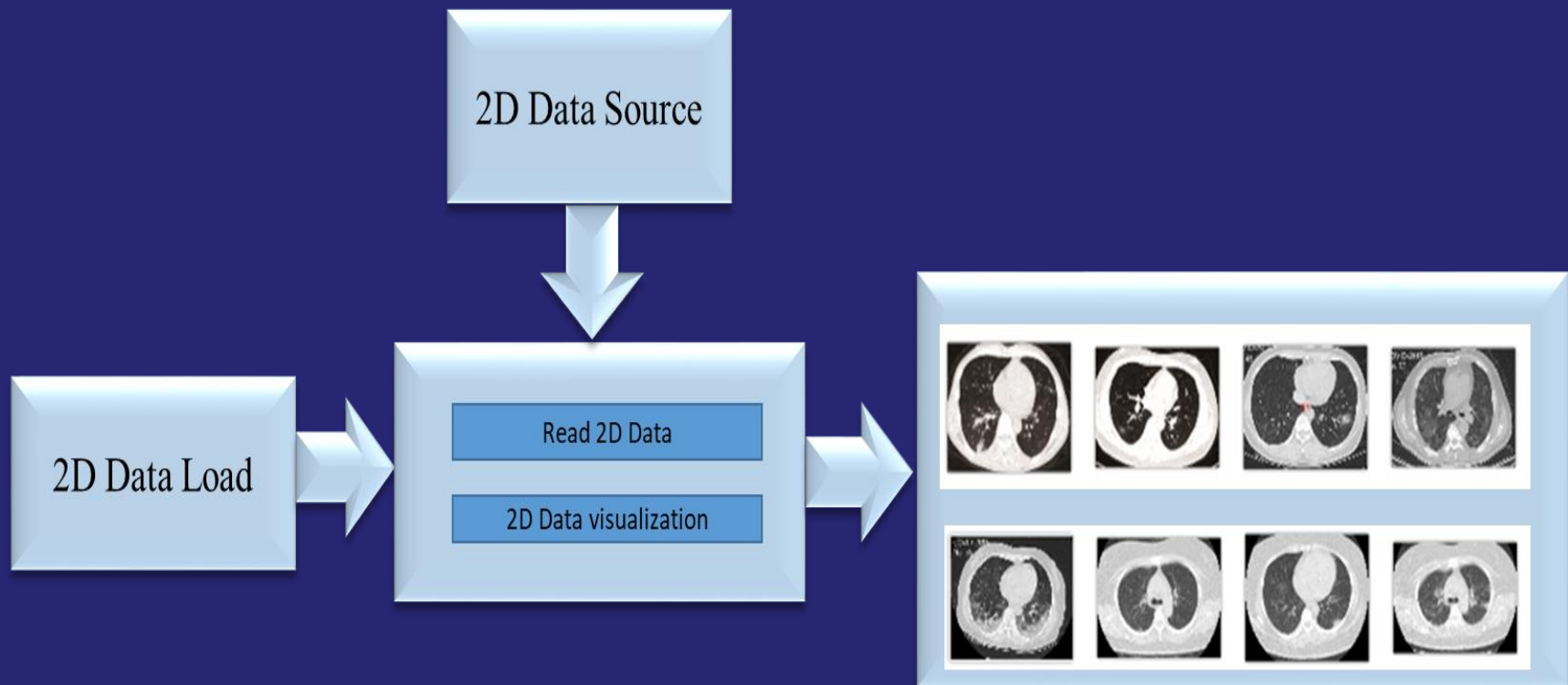
•Main features:

- 1. Hierarchical storage architecture;
- 2. Enables the import of variety of text-based matrix inputs;
- 3. Option for adaptation of the automated FASTQ-to-matrix pipeline to the requirements of the scientific experiments and Hi-C analysis performed;
- 4. Enables the running of the pipeline functions separately and enables individual setting for each of them;
- 5. Users can choose filters through the Python API which is a component of the framework;
- 6. Automatically generated diagnostic plots with filtering statistics which task is informing the user of issues.

The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)



Distance-Based Workflow Sample and results



[illegible]

The 47th International Conference
APPLICATIONS OF MATHEMATICS IN ENGINEERING AND ECONOMICS
(AMEE 2021)

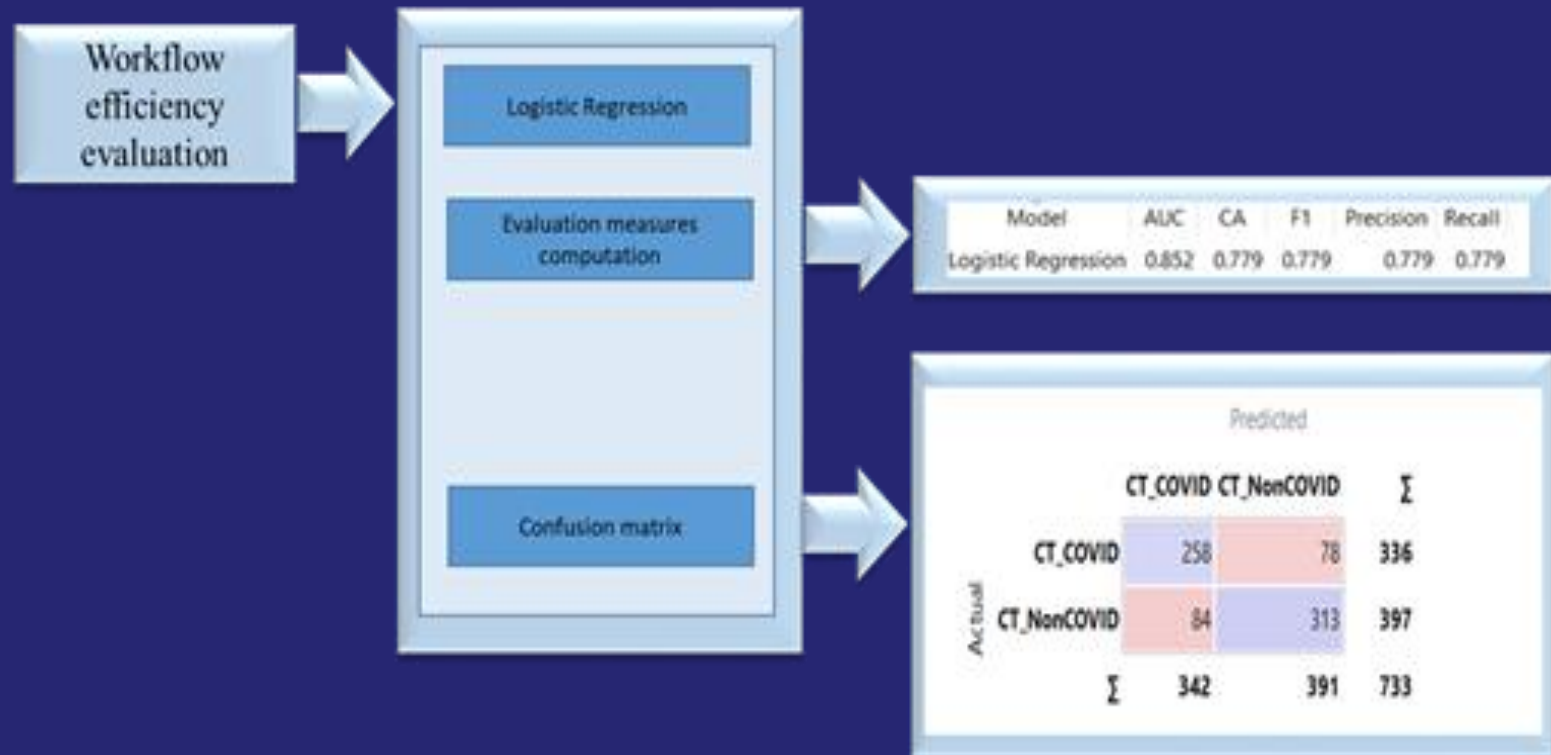


Image meta data and feature vectors

Thank You!

Stella Vetova

vetova.bas@gmail.com