BDCC 19/20
Worksheet #3
May 7th, 2020

Questions about paper:

- **Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence**

1. Figure 1 shows the standar Python ecosystem for machine learning, data science, and scientific computing. Search the internet and write a paragraph about each of the tools appearing in the blue boxes. What is the tool in the orange box? What is it for and how does it work?

2. What are the differences between Spark and Dask?

3. "Pandas' data frame format uses columns to separate the different fields in a dataset and allowseach column to have a different data type…Laying out the data contiguously by column enables SIMD by allowing the processor to group, or coalesce, memory accesses for row-level processing,making efficient use of caching while lowering the number of accesses to main memory."
Given the sentence above, what is the consequence of this memory arrangement in an algorithm that will cluster instances using, for example, a k-means algorithm, where distances between pairs of rows need to be calculated?

4. Figure 2 shows a python code example. What does the `make_pipeline` function do?

5. Why is bagging easier to parallelize than boosting?

6. What is the purpose of Neural Architecture Search? (Section 3.3)

7. Explain the sentence: "The CoCoA framework preserves locality across compute resources on each physical machine to reduce the amount of network communication needed across machines in the cluster"

8. What is the relation between Dask, RAPIDS and cuML?

Questions about paper:

- From Persistent Identifiers to Digital Objects to Make Data Science More Efficient

1. Given the sentence below, what is necessary to make data of better quality and better organized?

“The biggest factors for these inefficiencies seem to be a bad quality of data/metadata and a bad data organization, i.e., if a rough description of data is available, it is difficult to find out how to access them, where the corresponding metadata is to enable processing, etc.”

2. Do you think that the points mentioned in the paper are the only ones to make data of better quality? (you may need to complement your reading with this report).

3. What is a Digital Object and how can it help to improve data access?