Big Data and Cloud Computing, 19-20 (2nd Module)

Inês Dutra
DCC-FCUP
ines@dcc.fc.up.pt

April 15th

Data Mining and Machine Learning: recap

- Learning?
  - "An agent **learns** if it improves its performance in future tasks after making observations about the past or current world." (Mitchel)

Data Mining and Machine Learning: recap

- Learning?
  - Given observations O, described by features $f_1, f_2, \ldots, f_n$, the task of a machine learning algorithm is:
    - to find patterns based on features $f_1, f_2, \ldots, f_n$ (all or some of them), that distinguish among different groups of observations OR
    - to find a function that will **predict** new observations

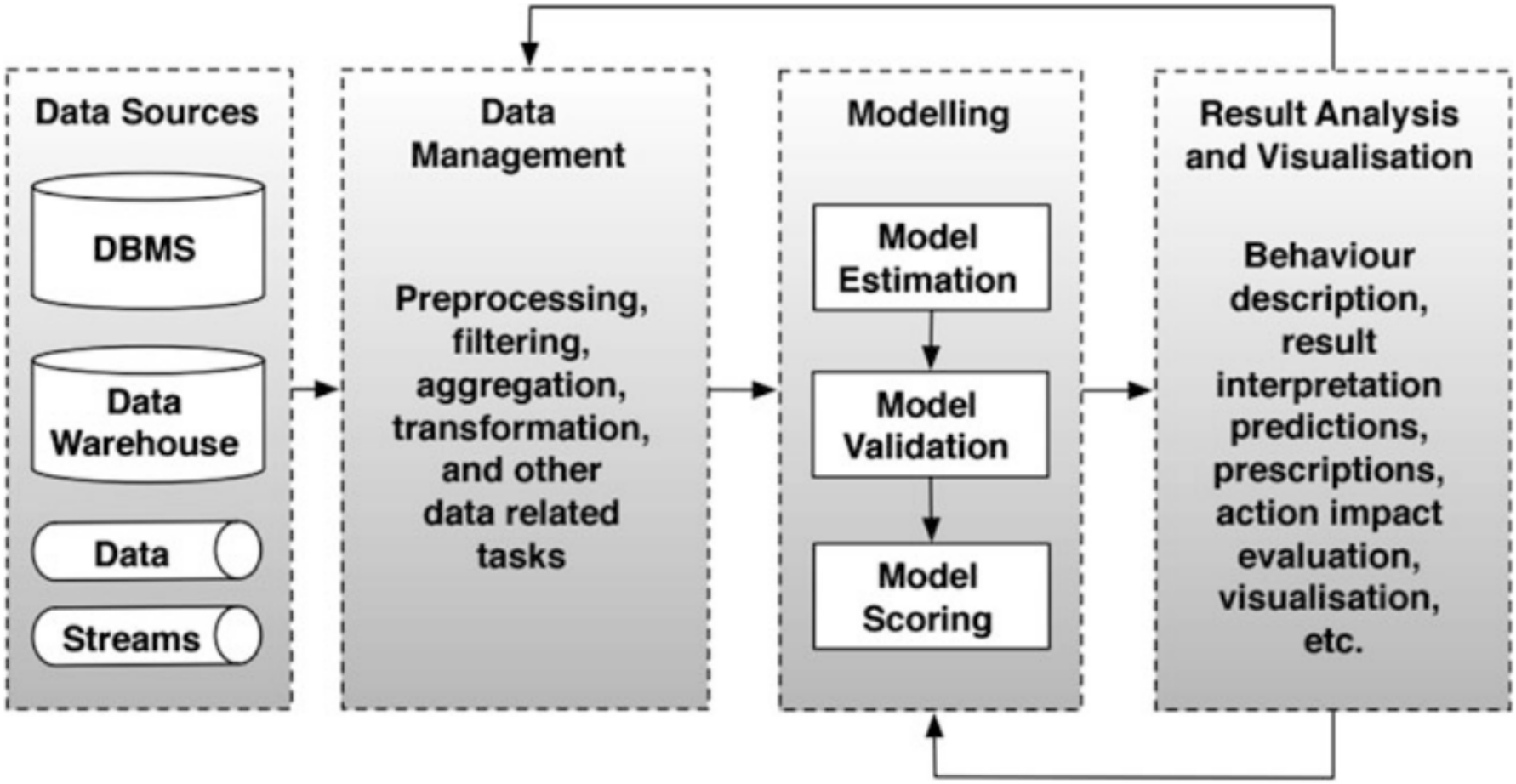
Data Mining and Machine Learning: recap

- Learning?
  - Can be **supervised**:
    - Given features $f_1, f_2, \ldots, f_n$, and a special feature, the **target** variable (ground truth), find a model that can predict the target variable for new observations that are described by features $f_1, f_2, \ldots, f_n$
    - The supervised learning task can be **classification** or **regression**
  - Can be **unsupervised**: find subgroups of patterns, no target variable is known or provided
    - clustering
    - association rules
  - Other learning methods: reinforcement learning, matrix factorization for recommender systems

Data Mining and Machine Learning: recap

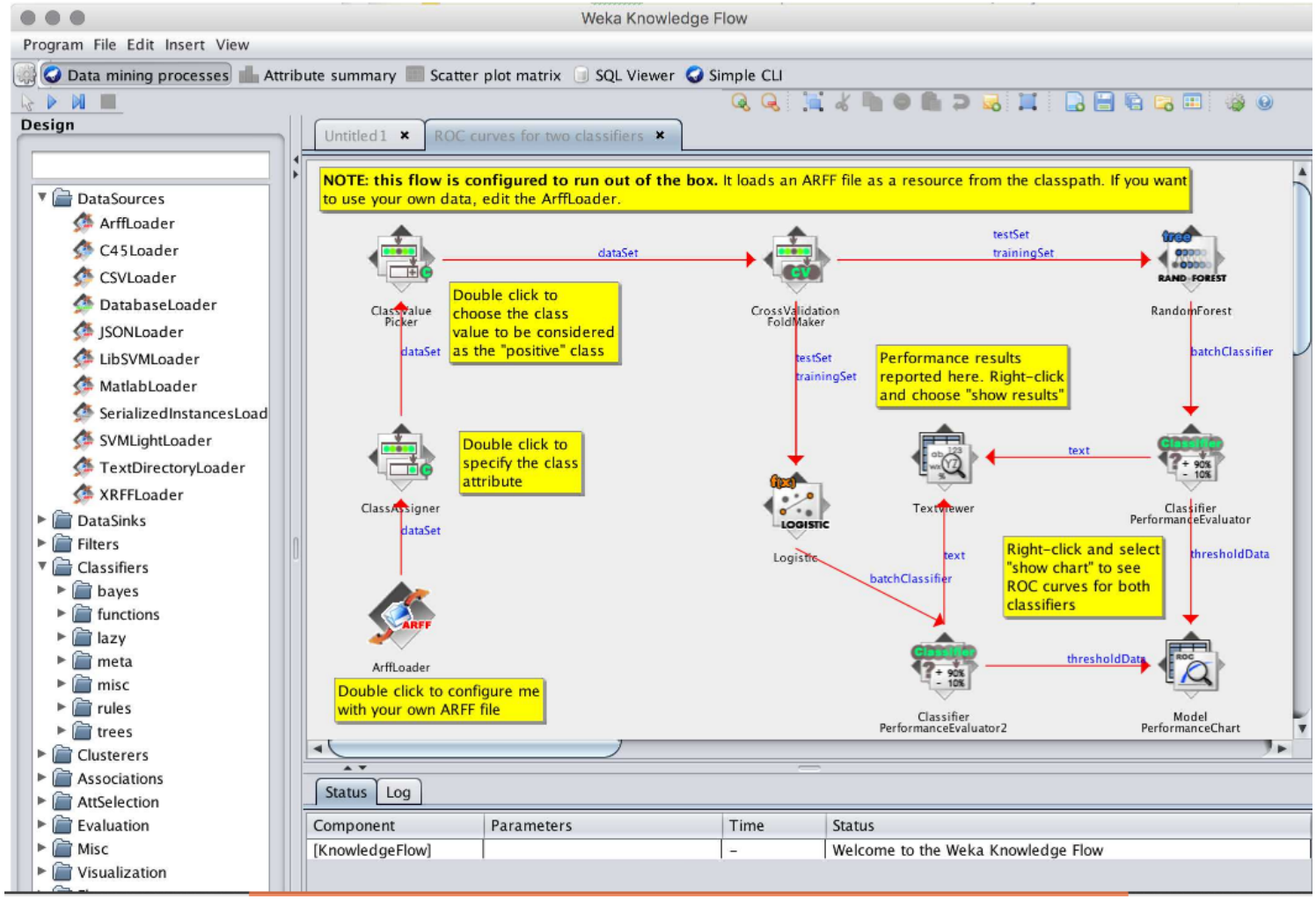
- Workflow (Dataflow - Knowledgeflow):
  - Data preprocessing
    - transformation: normalization, standardization, averaging, median, denoising, filtering
    - preparation: depends on the task, algorithm, package or library being used
  - Machine learning task, algorithm
  - Validation: cross-validation, bootstrapping
- Workflow tools: WEKA KnowledgeFlow, RapidMiner, Orange3, Taverna, Condor DAGMan, Pegasus, Google Dataflow, Google Composer (Apache Airflow)

Data Mining and Machine Learning: workflow



Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S., Buyya, R.: Big data computing and clouds: trends and future directions. J. Parallel Distrib. Comput. 79–80, 3–15 (2015)

Example of workflow in WEKA



`java -jar weka.jar` ⇒ KnowledgeFlow

Example of workflow with Orange3

(installation needed, go to https://orange.biolab.si/)
orange.canvas

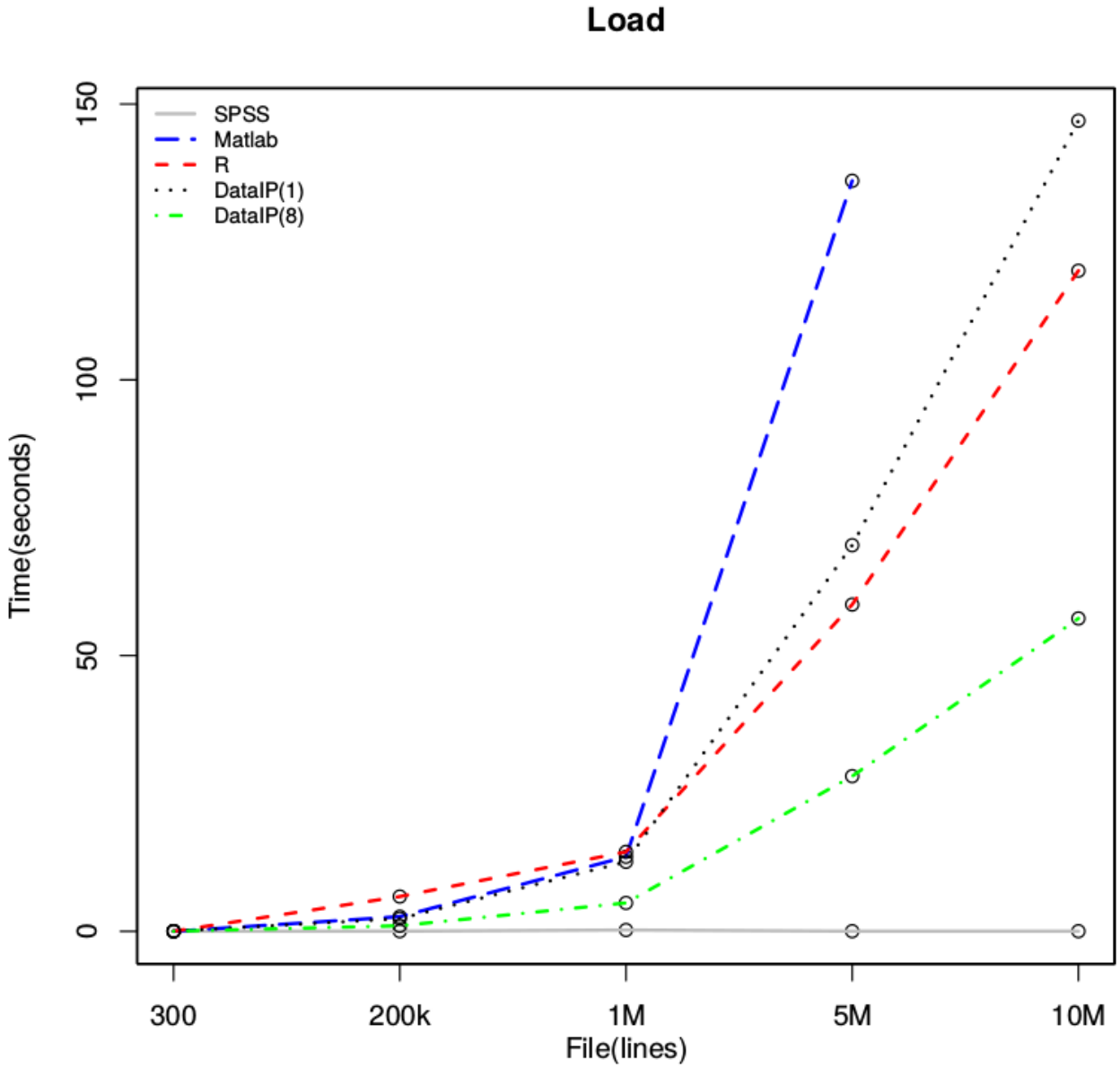
Limitations

- Most systems and tools for data analysis are not scalable
- I/O, memory, computing power

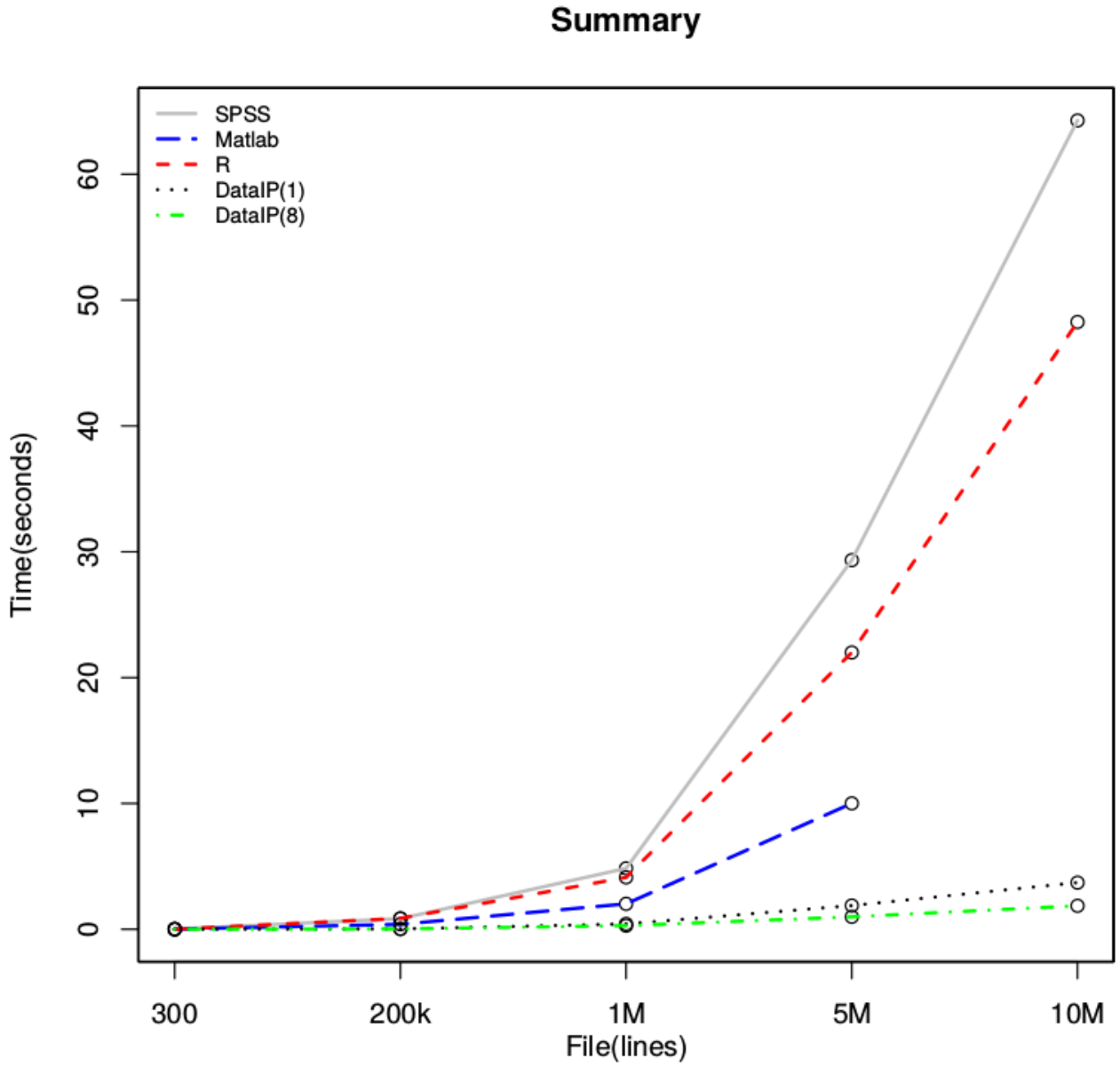
Scalability

- Computational resources: memory, CPU, I/O, storage
- I/O:
  - Experiment 1: SPSS, MatLab, R and DataIP (in-house implementation)
    - dataset of patients, originally 200K entries, 6 numeric variables without nulls
    - varying sizes: 300, 200k, 1M, 5M, 10M
  - Experiment 2: job that needs to fetch data files from a remote site

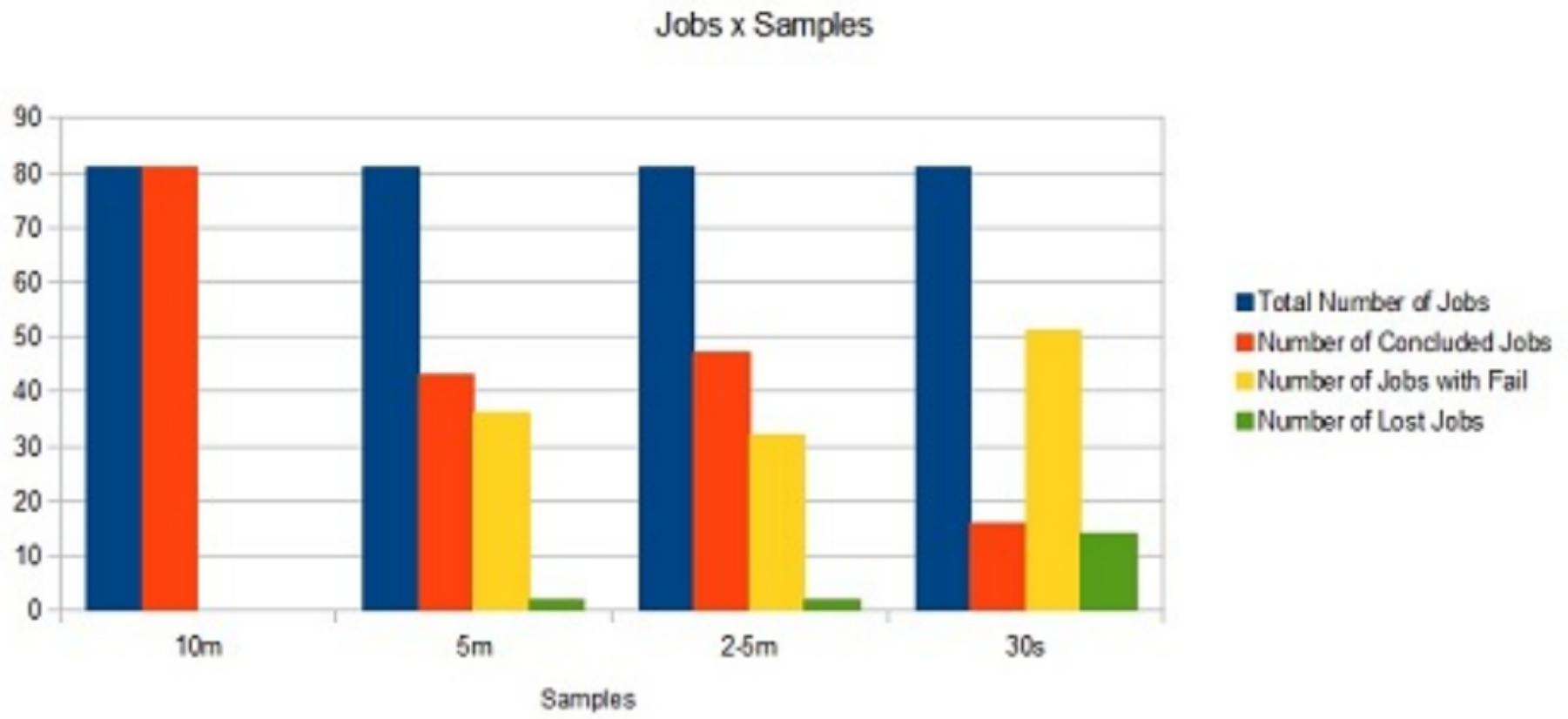
Scalability: Experiment 1, I/O



Scalability: Experiment 1, simple computing: summary



Scalability: Experiment 2, file transfer



Scalability

- Alternatives
  - break file in multiple smaller files that can be read in parallel: useful if the reading can be done in parallel
  - undersampling: need to be careful about data distribution
  - use of specific hardware and software: distributed disks, distributed file systems, distributed databases, in-memory databases, parallel and distributed software
  - work with compressed files: zip, parquet, CSR, CSR5 (for sparse matrices) etc