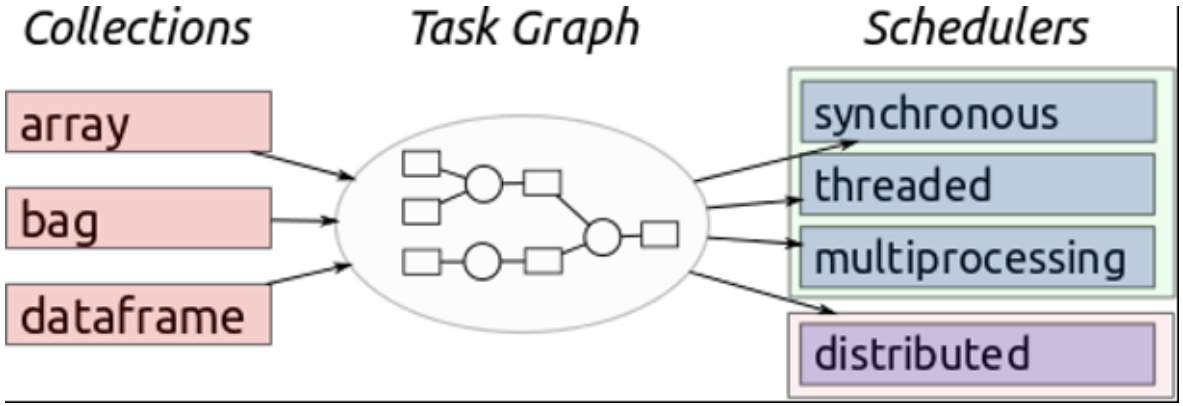
zühlke
empowering ideas

Distributed machine learning with dask and scikit-learn

Dask

A flexible parallel computing library for analytic computing

- Dynamic Task Scheduling
- "Big Data" Collections: parallel numpy and pandas objects
- Distributes on cluster with 100s of machines
- Task Graphs: more complex than map/reduce

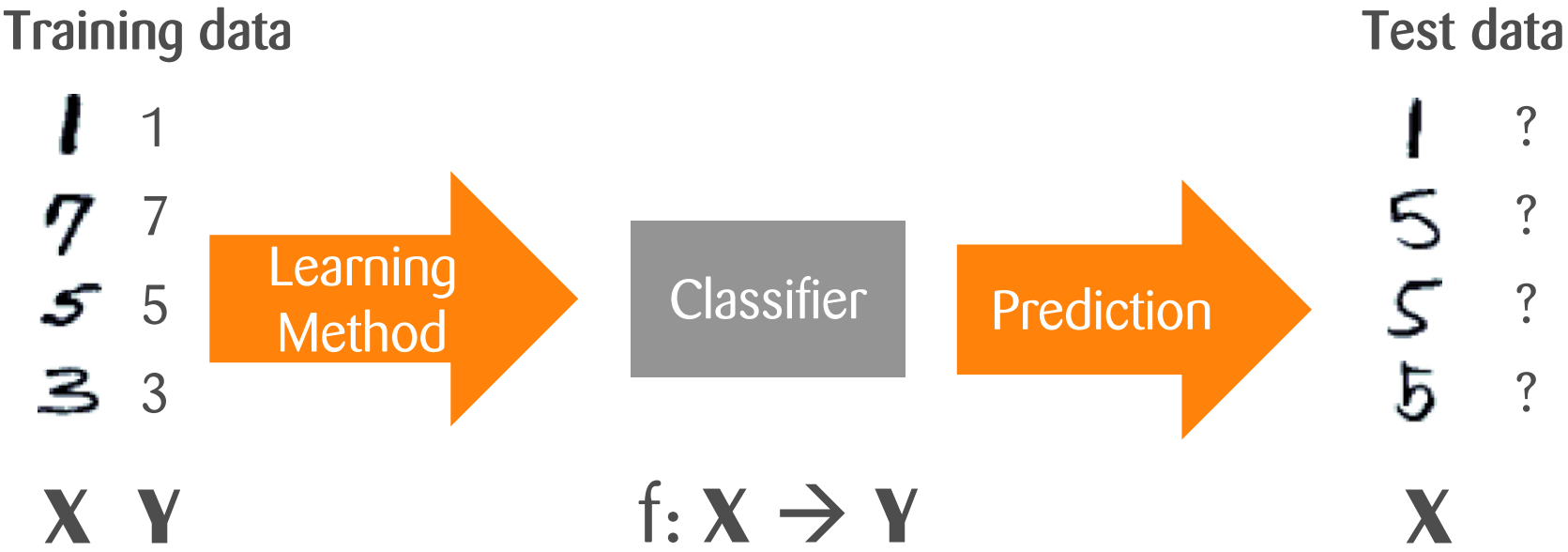


Jupyter Notebook Demo

Parallel Numpy and Distributed Computing

Classification Example

Back to the 90s: Handwritten digit recognition

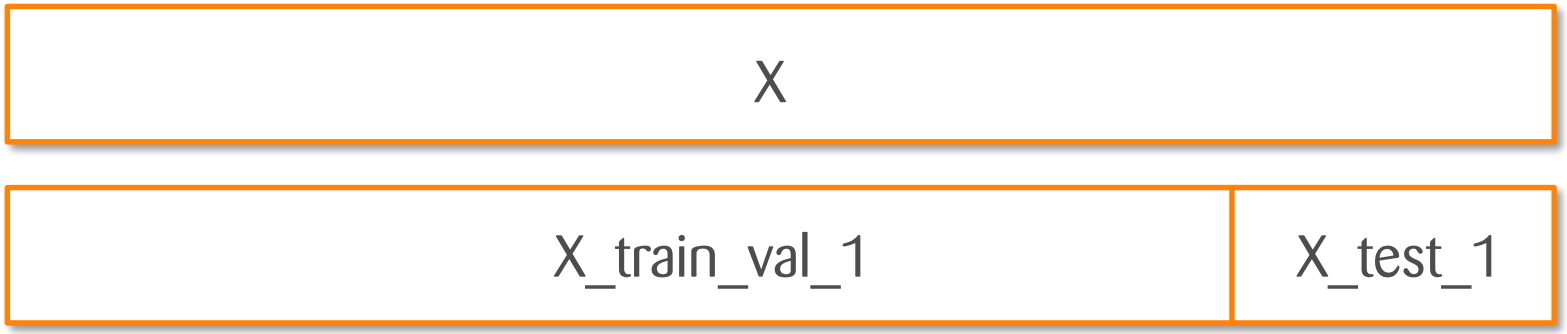


Fit Model

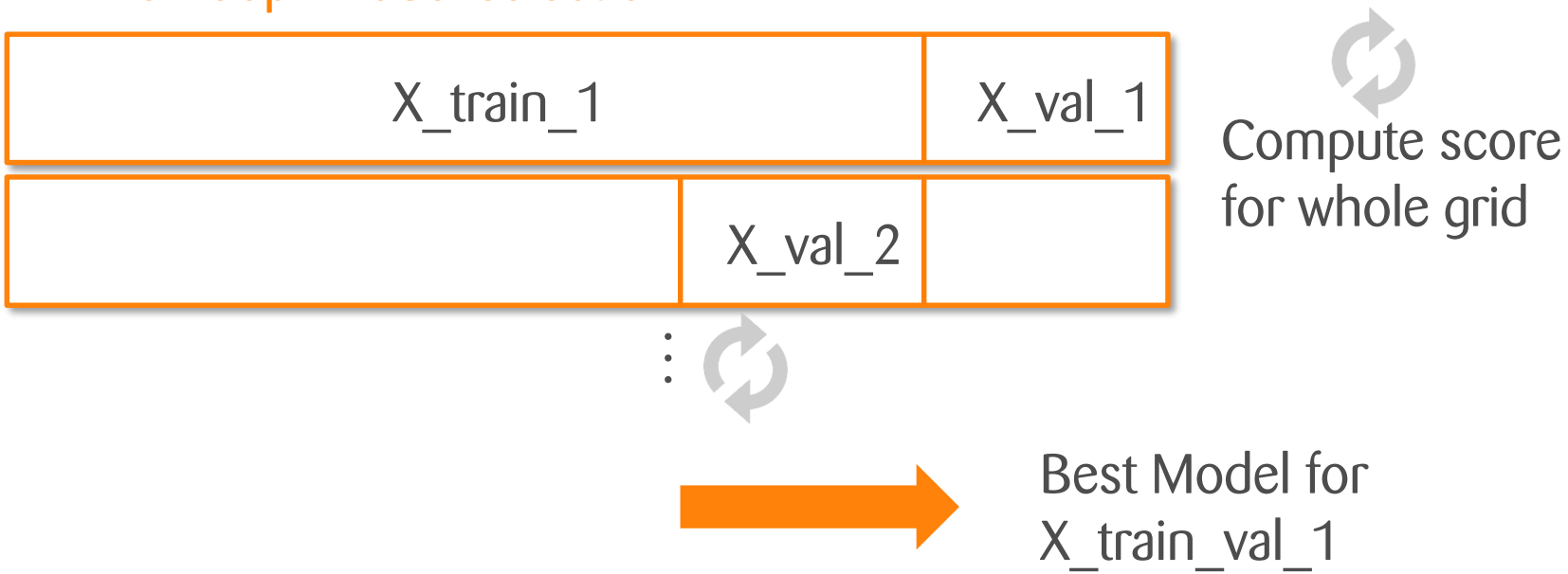
Generalization

Nested cross validation

A quick refresher on model selection and expected error



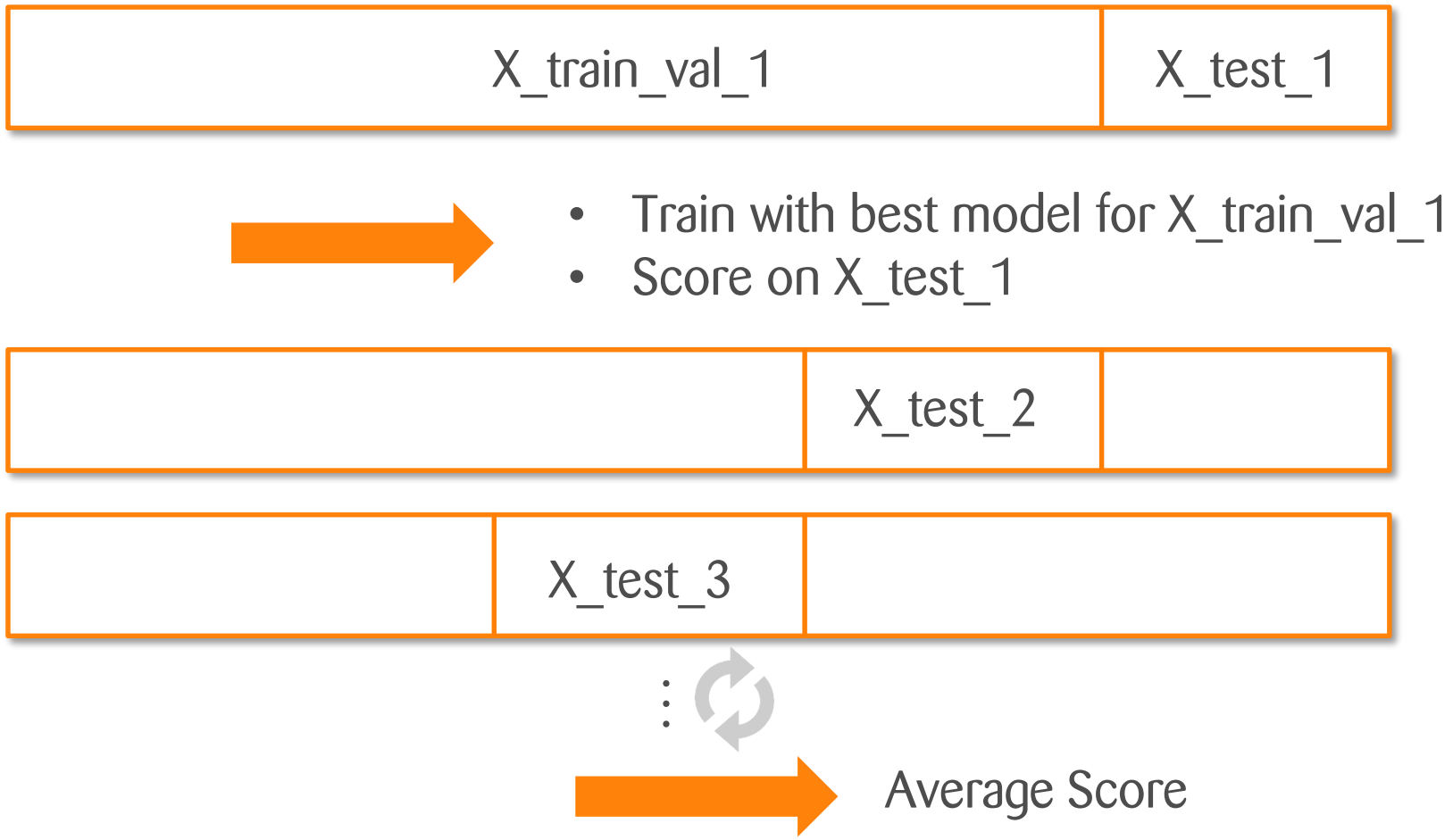
1. Inner loop: model selection



Nested cross validation

A quick refresher on model selection and expected error

2. Outer loop: expected error



Jupyter Notebook Demo 2

Train an SVM to recognize handwritten digits

There is more!

What we haven't shown

- GridSearchCV: github.com/dask/dask-learn

- Run Dask on Amazon Cloud
- Complex Task Graphs

- http://matthewrocklin.com/blog/
- http://dask.pydata.org