UNIVERSITÀ DEGLI STUDI DI GENOVA

Dibris Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi

PALLADIO: A PARALLEL FRAMEWORK FOR ROBUST VARIABLE SELECTION IN HIGH-DIMENSIONAL DATA

UNIVERSITÀ DEGLI STUDI DI GENOVA

Dibris Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi

MATTEO BARBIERI

SAMUELE FIORINI

FEDERICO TOMASI

ANNALISA BARLA

http://slipguru.unige.it/

SUMMARY

- Background: supervised learning and variable selection
- Framework description
- Validation on synthetic datasets
- Conclusions and future works

BACKGROUND

LEARNING FROM EXAMPLES

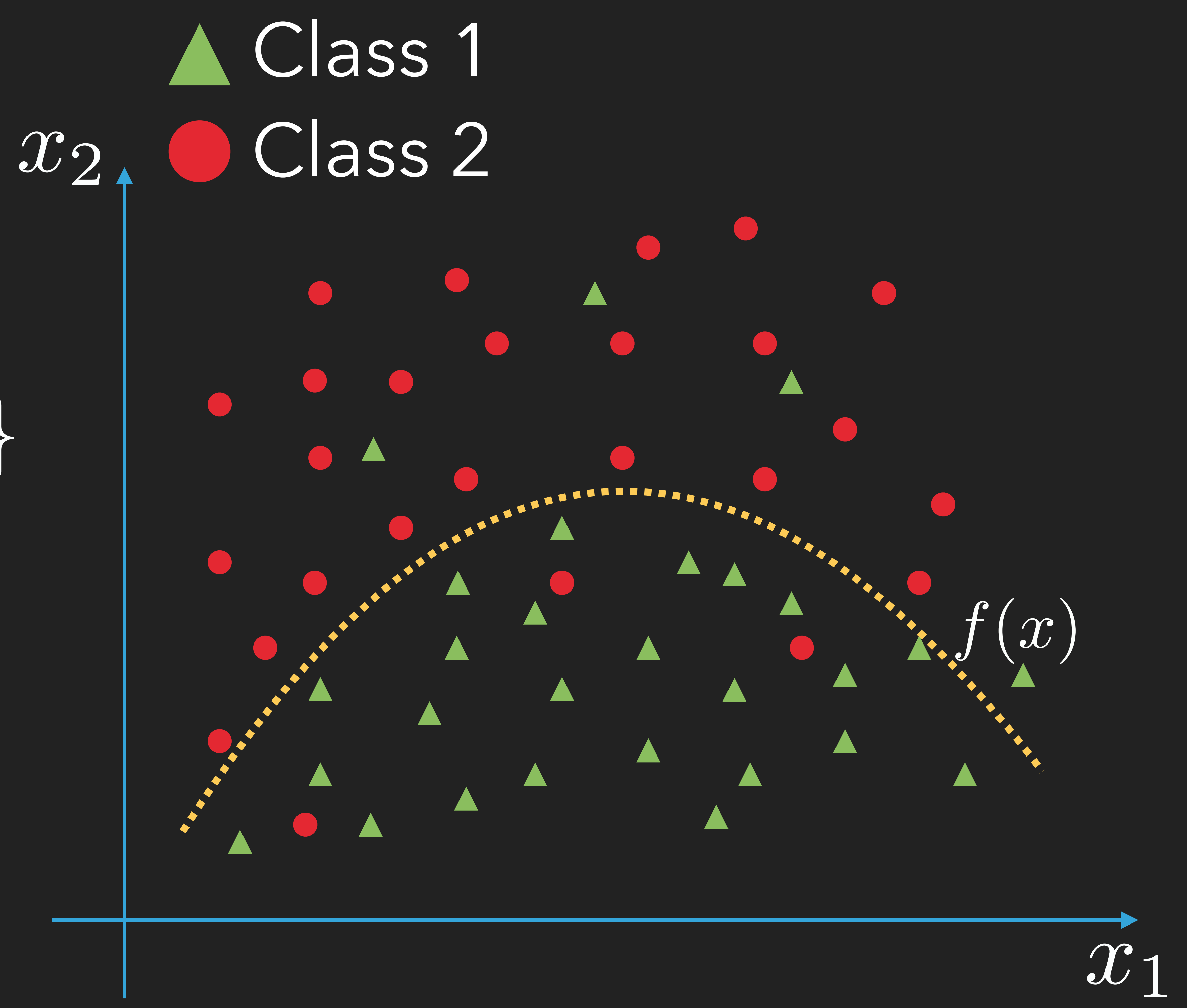
Examples: pairs of the form $(\mathbf{x}, y)$

e.g. $\mathbf{x} = [x_1, x_2]$ and $y \in \{▲, ●\}$

Goal: infer function f such that

$$f(\mathbf{x}) \sim y$$

More generally: $\mathbf{x} \in \mathbb{R}^d$



VARIABLE SELECTION

The process of identifying the subset of relevant variables

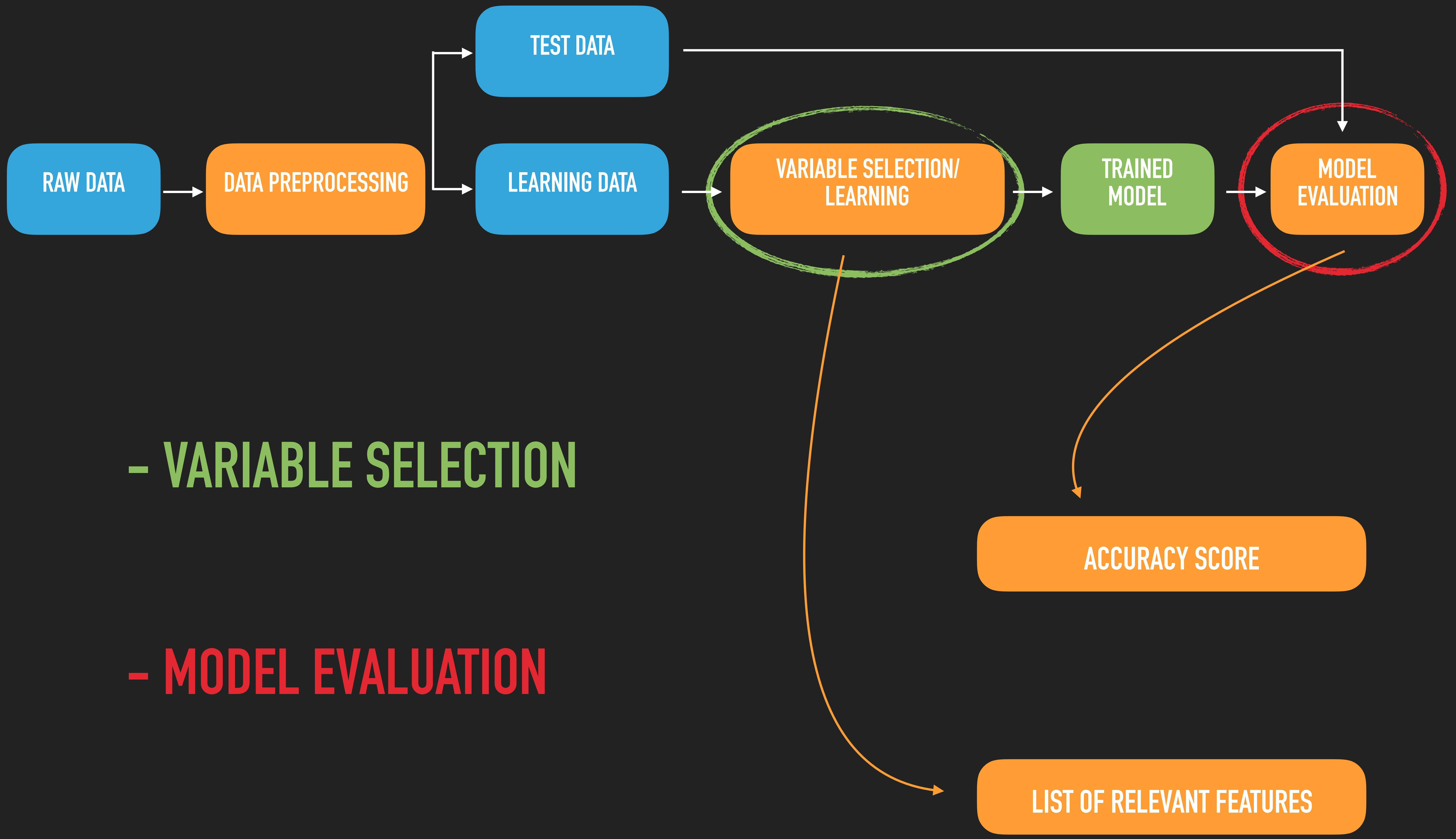
ASSUMPTION

Not all variables are relevant for the problem

PURPOSE

- Reduce computational time
- Enhance interpretability

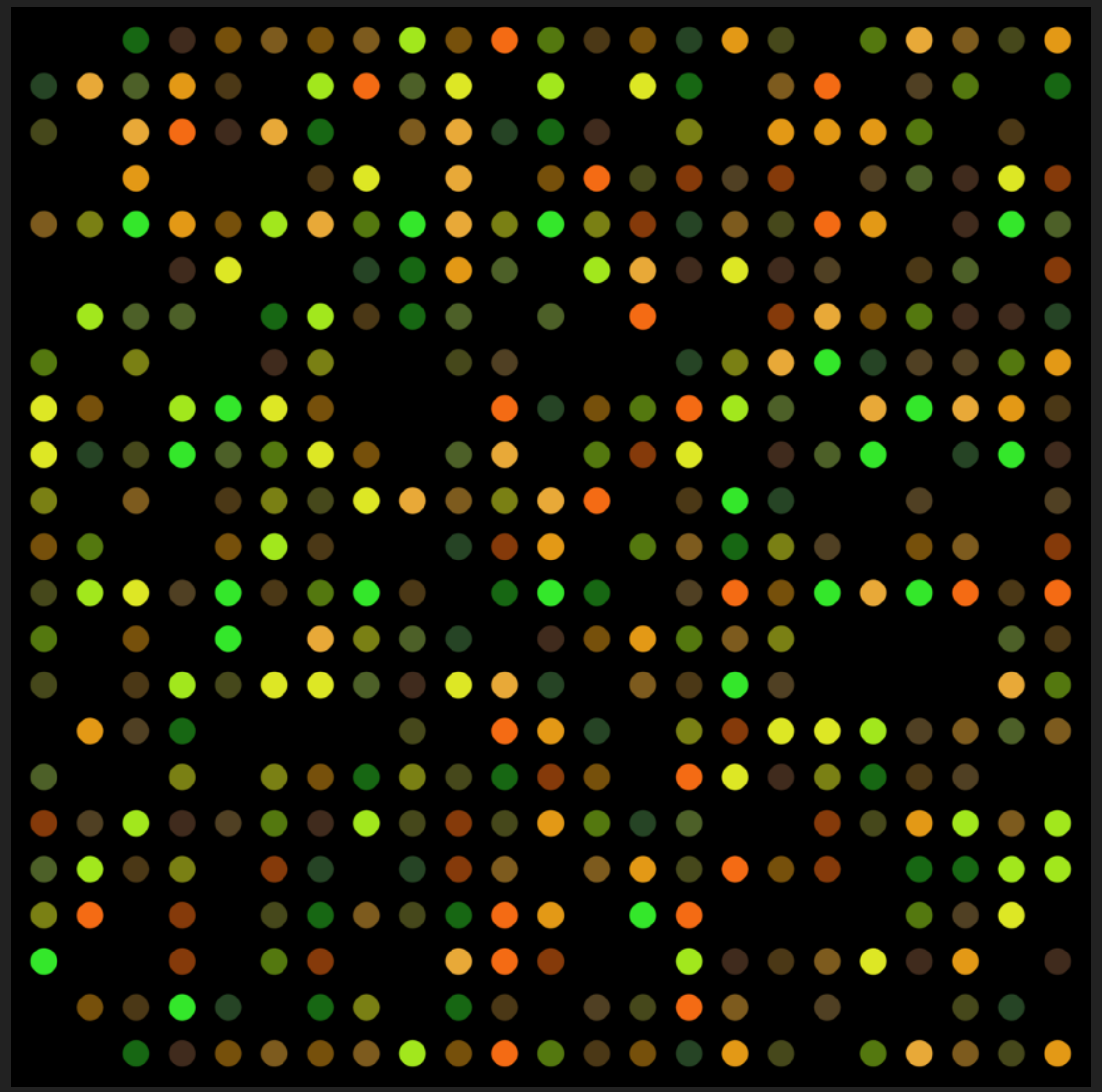
THE LEARNING PIPELINE

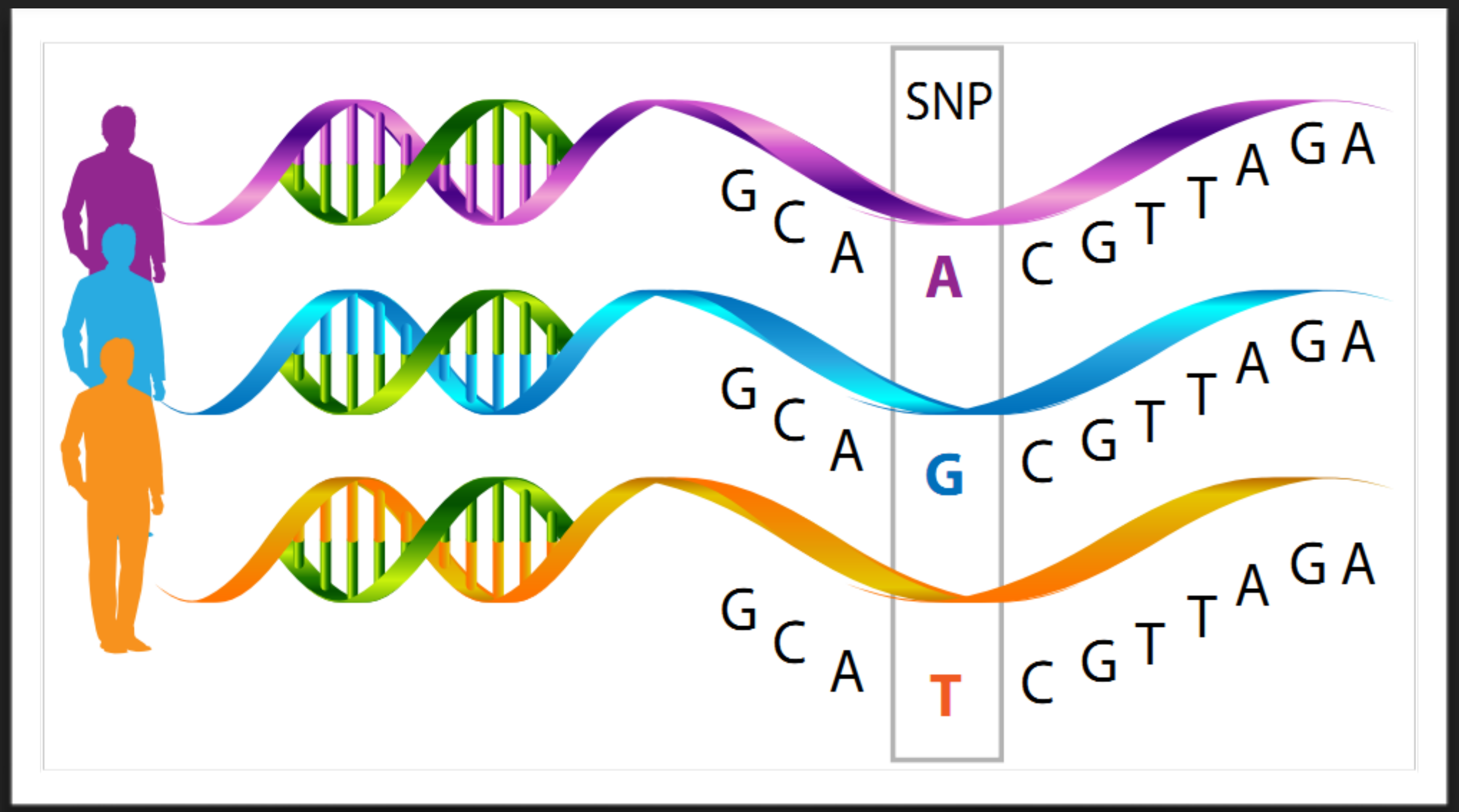


- VARIABLE SELECTION

- MODEL EVALUATION

EXTREME CASES: n << d, WEAK CORRELATION





Few examples (10^2), high dimensionality (10^4 - 10^6)

Input and output are weakly correlated

Classification accuracy close to chance

A PARALLEL FRAMEWORK FOR ROBUST VARIABLE SELECTION IN HIGH-DIMENSIONAL DATA

PALLADIO

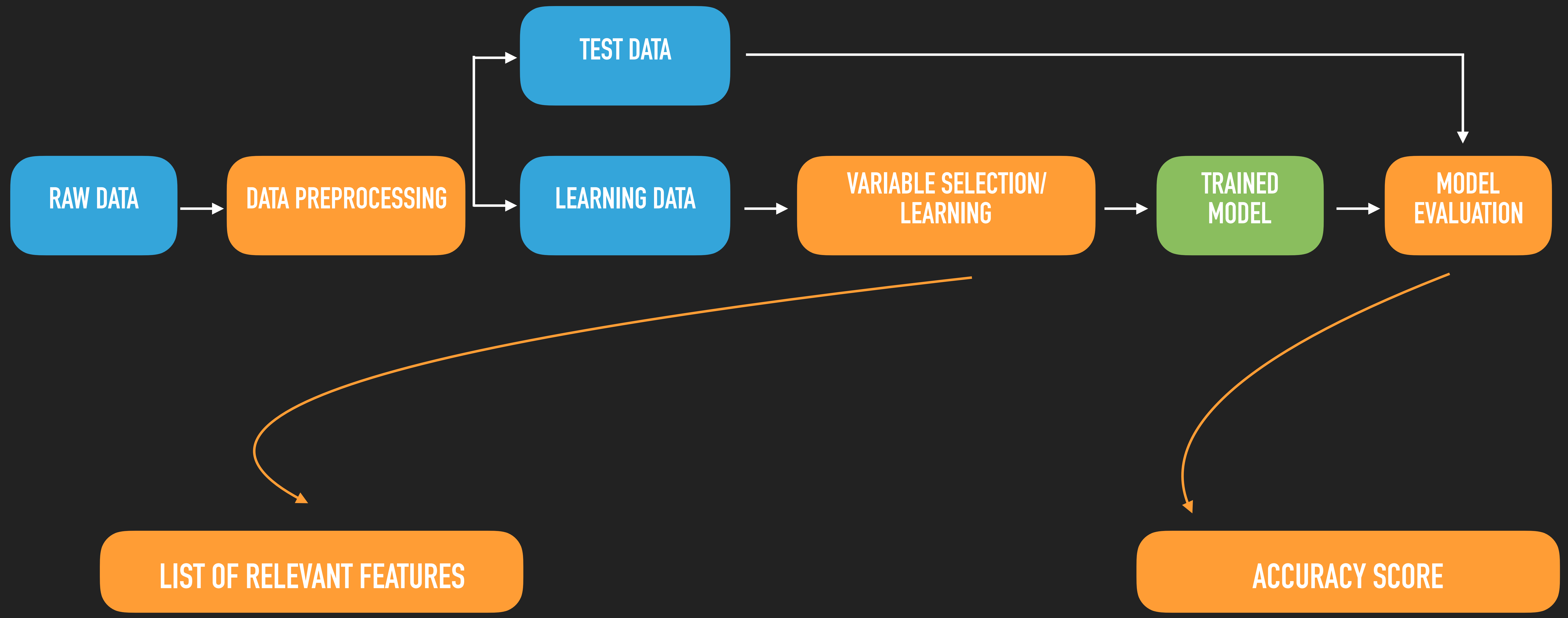
MPI

GOALS

- Select a subset of relevant variables

- Provide a measure of the reliability of the results

SCHEMA



One accuracy score, one list of variables

Accuracy = 54%

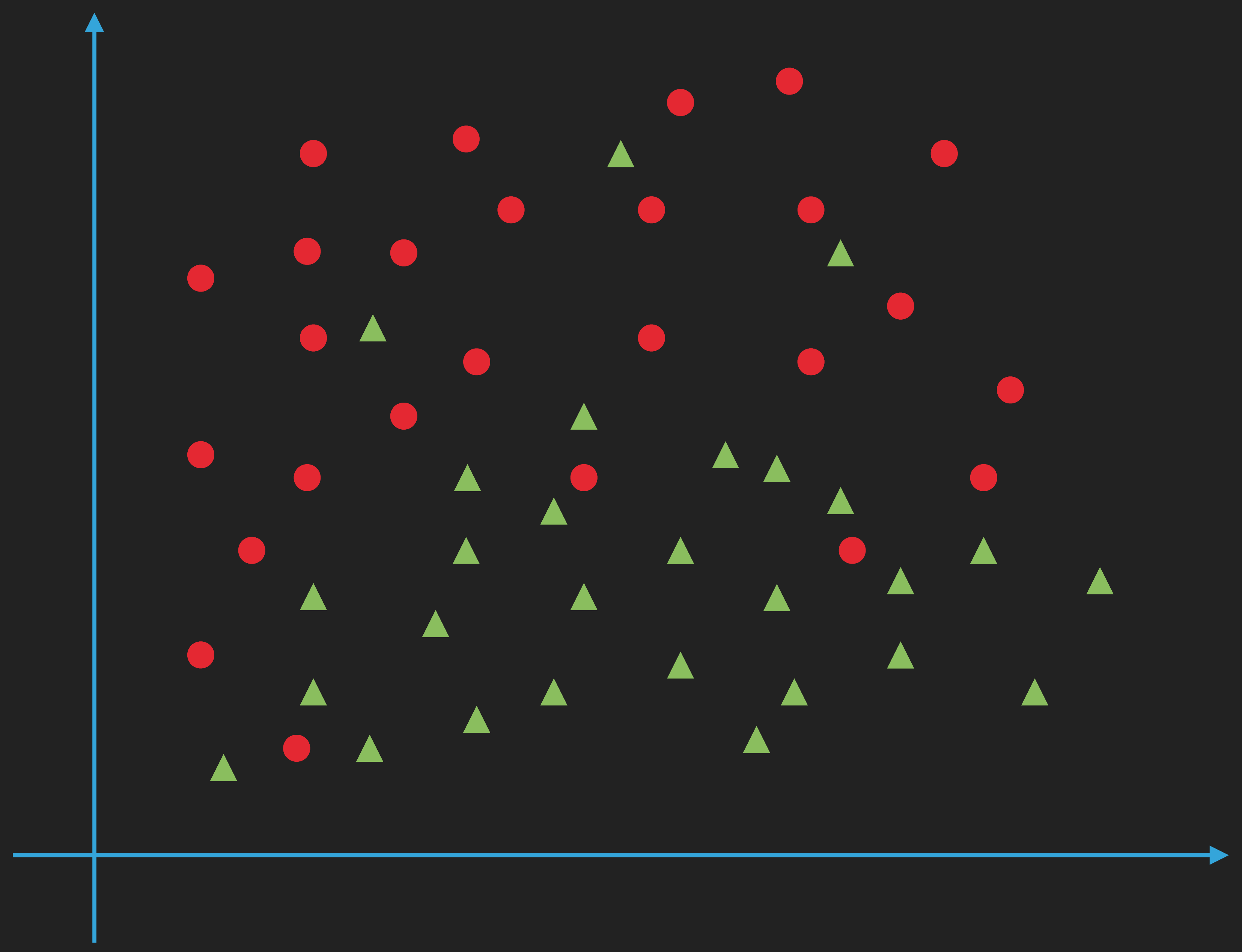
Is it good? (chance is 50%)

Does it depend on the split?

HOW IT WORKS

- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV

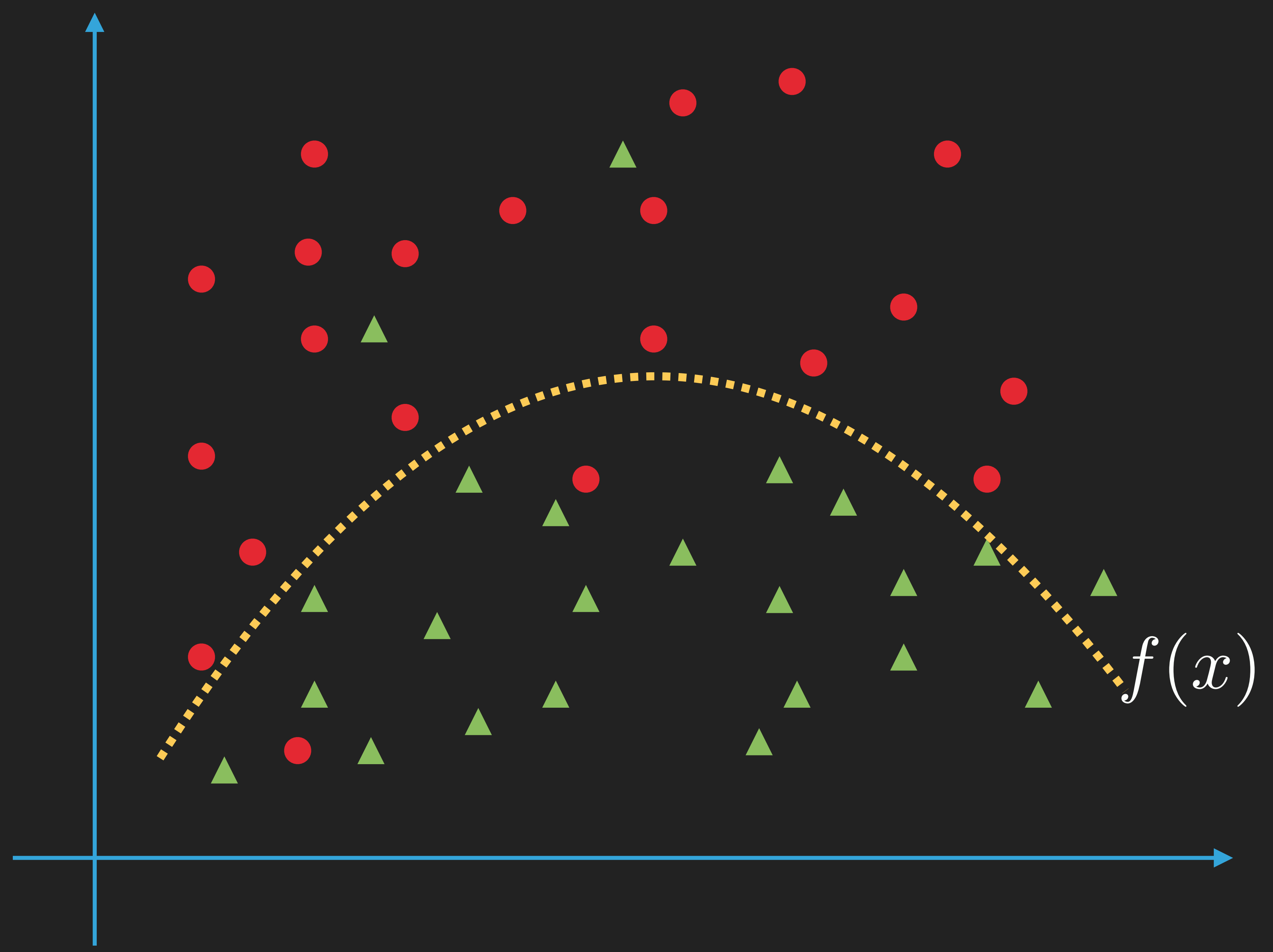
EXPERIMENT 1



HOW IT WORKS

- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV

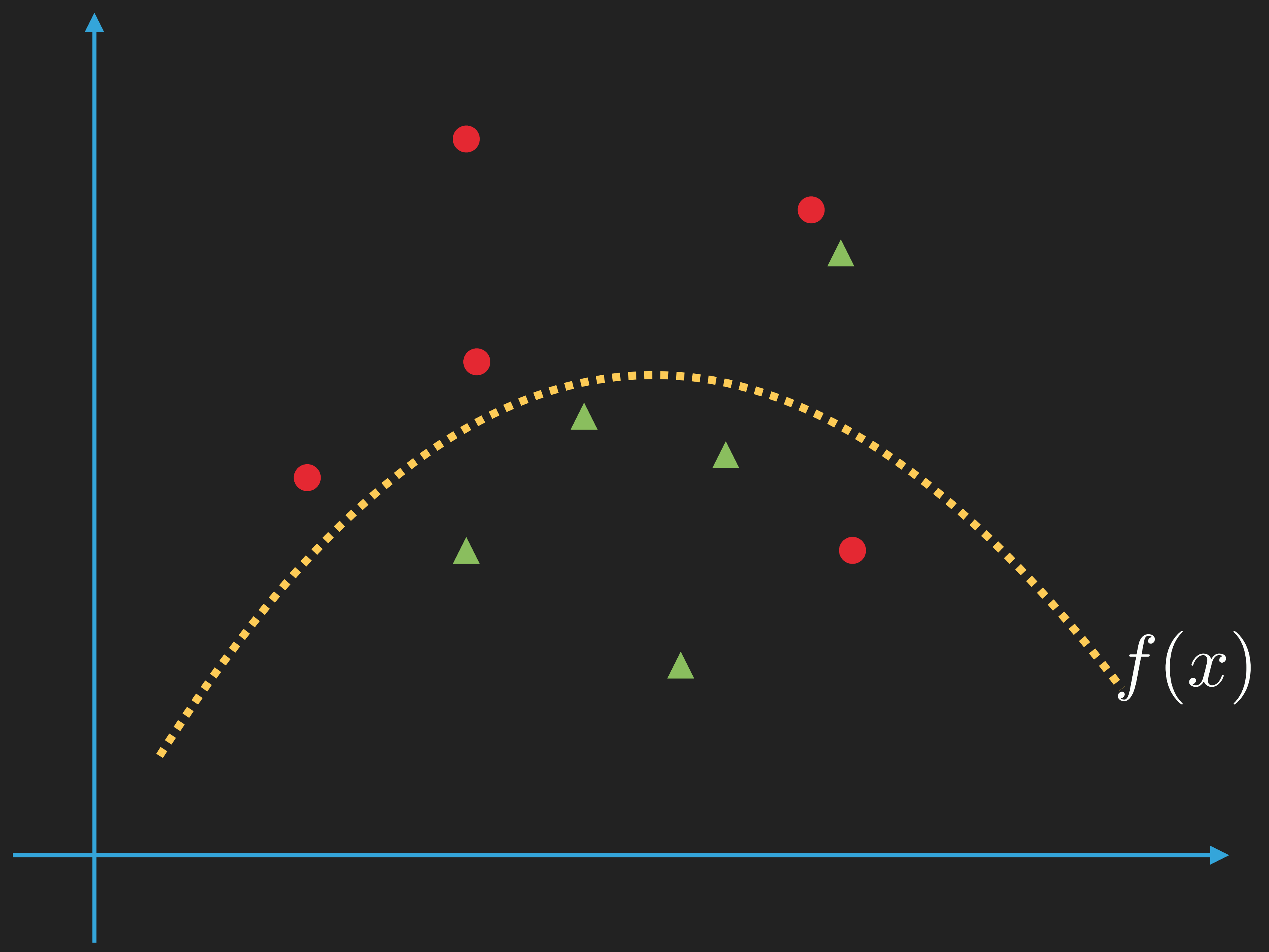
EXPERIMENT 1: LEARNING SET



HOW IT WORKS

- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV

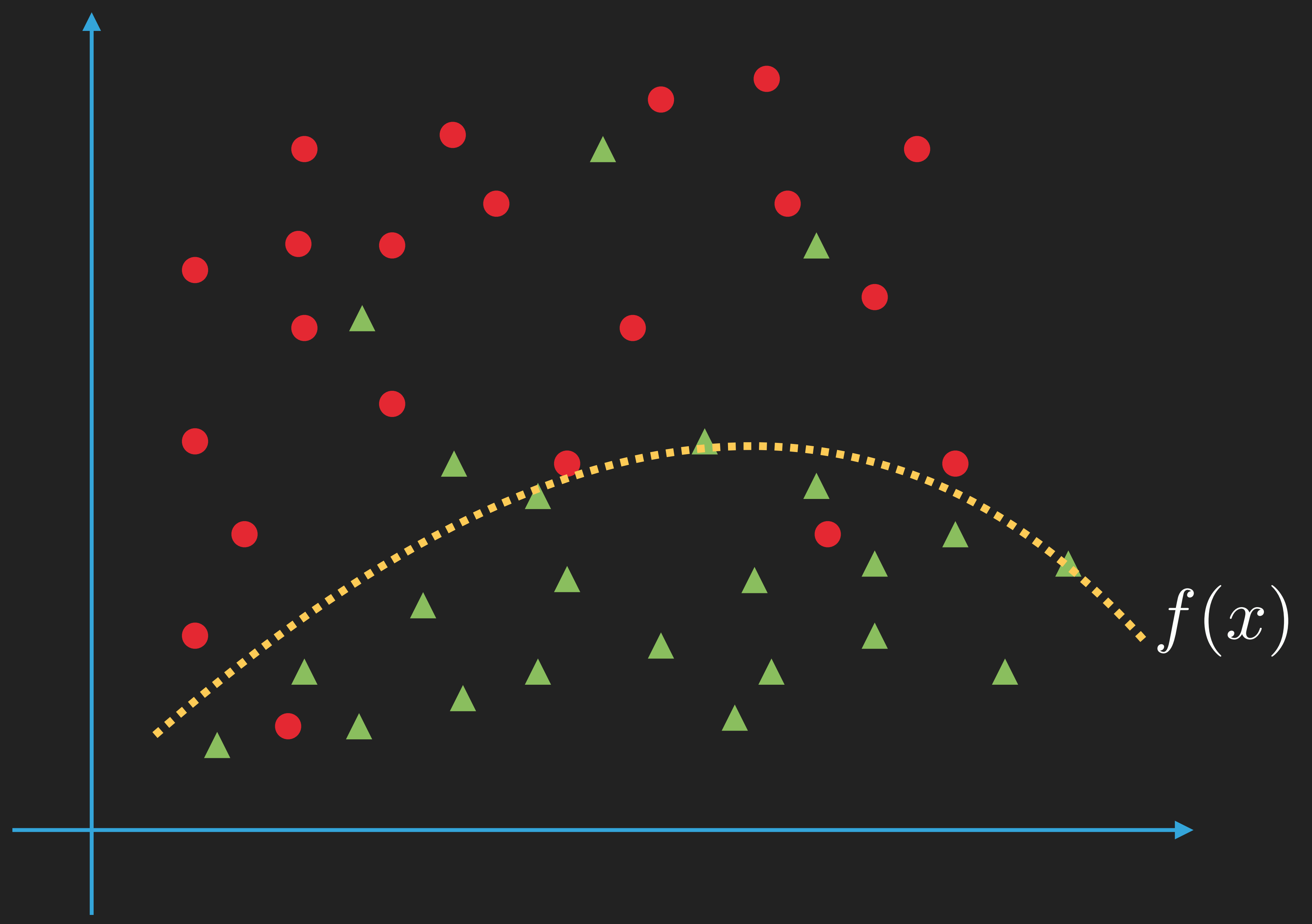
EXPERIMENT 1: TEST SET



HOW IT WORKS

- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV

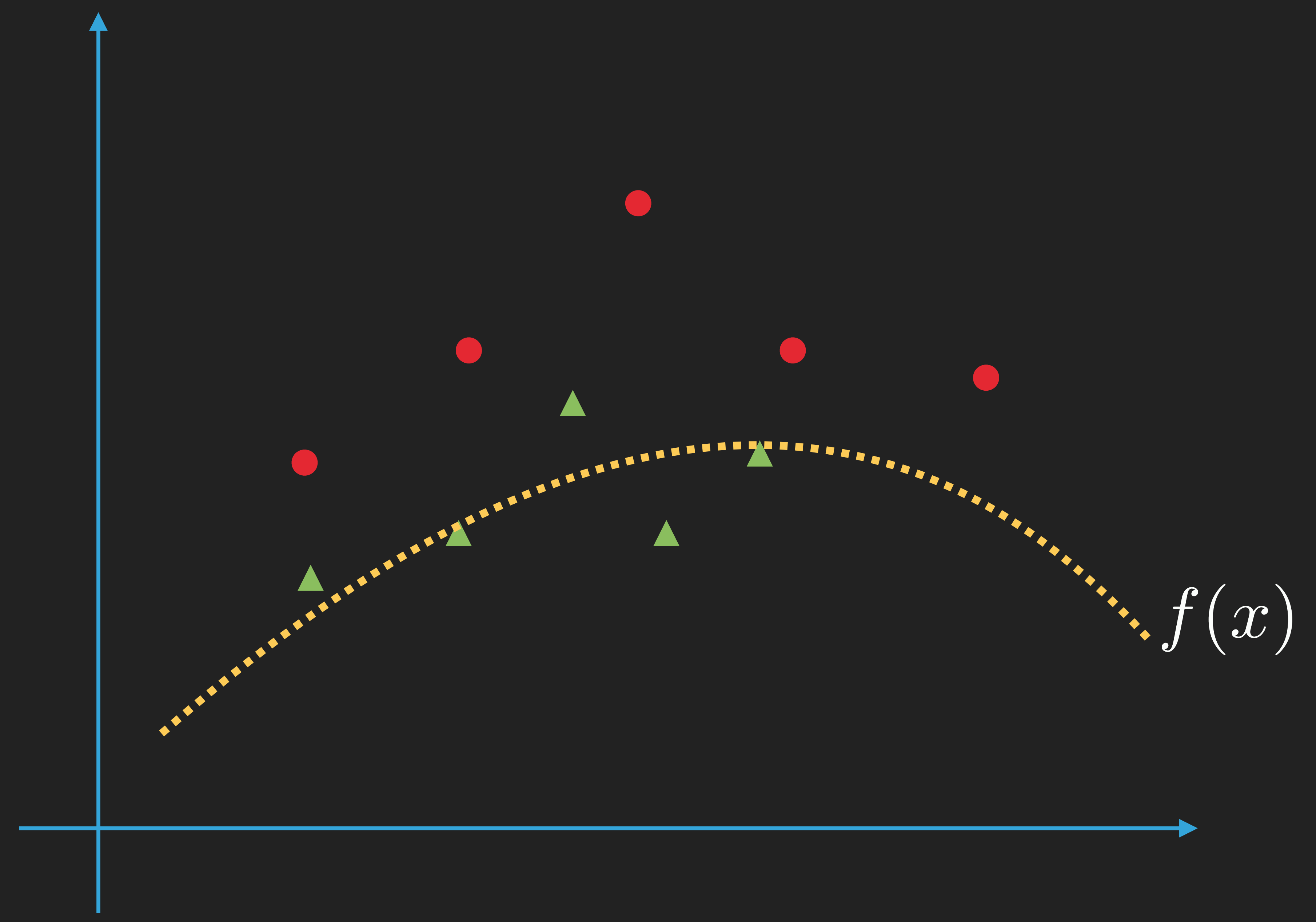
EXPERIMENT 2: LEARNING SET



HOW IT WORKS

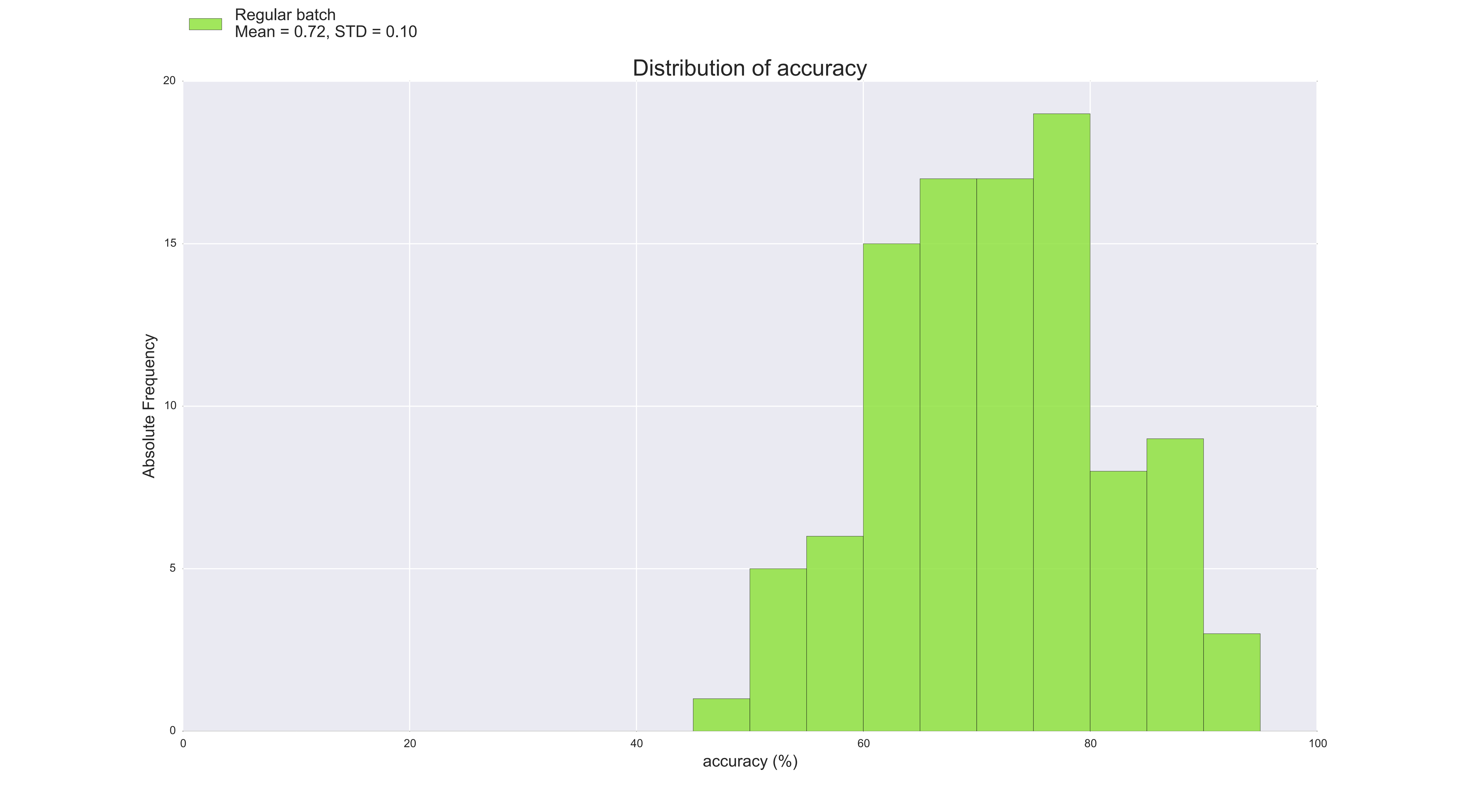
- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV

EXPERIMENT 2: TEST SET



HOW IT WORKS

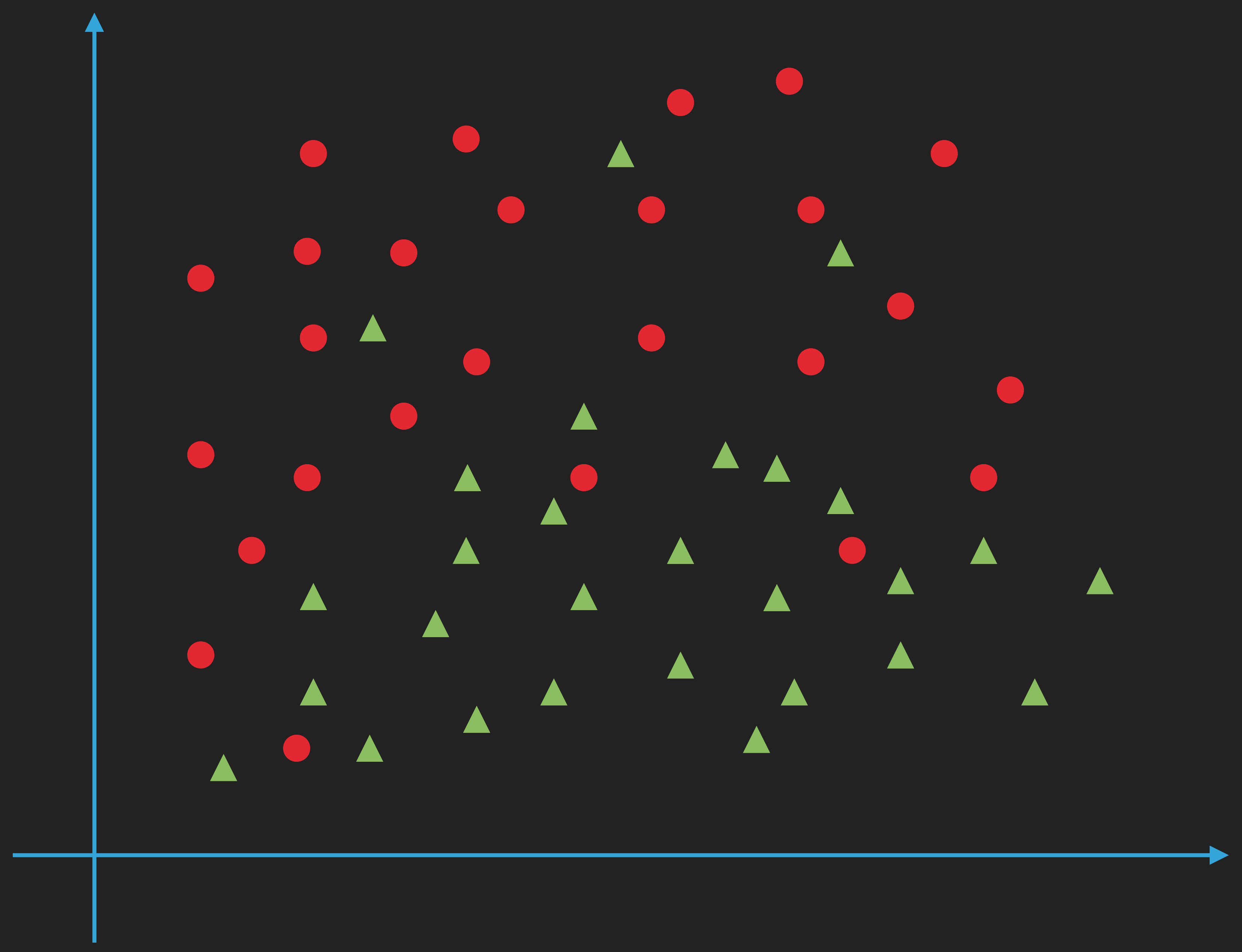
- Repeat the experiments many times (100)
  - Resample learning and test set with MCCV



HOW IT WORKS

- Permutation test
  - Labels in the learning set are shuffled

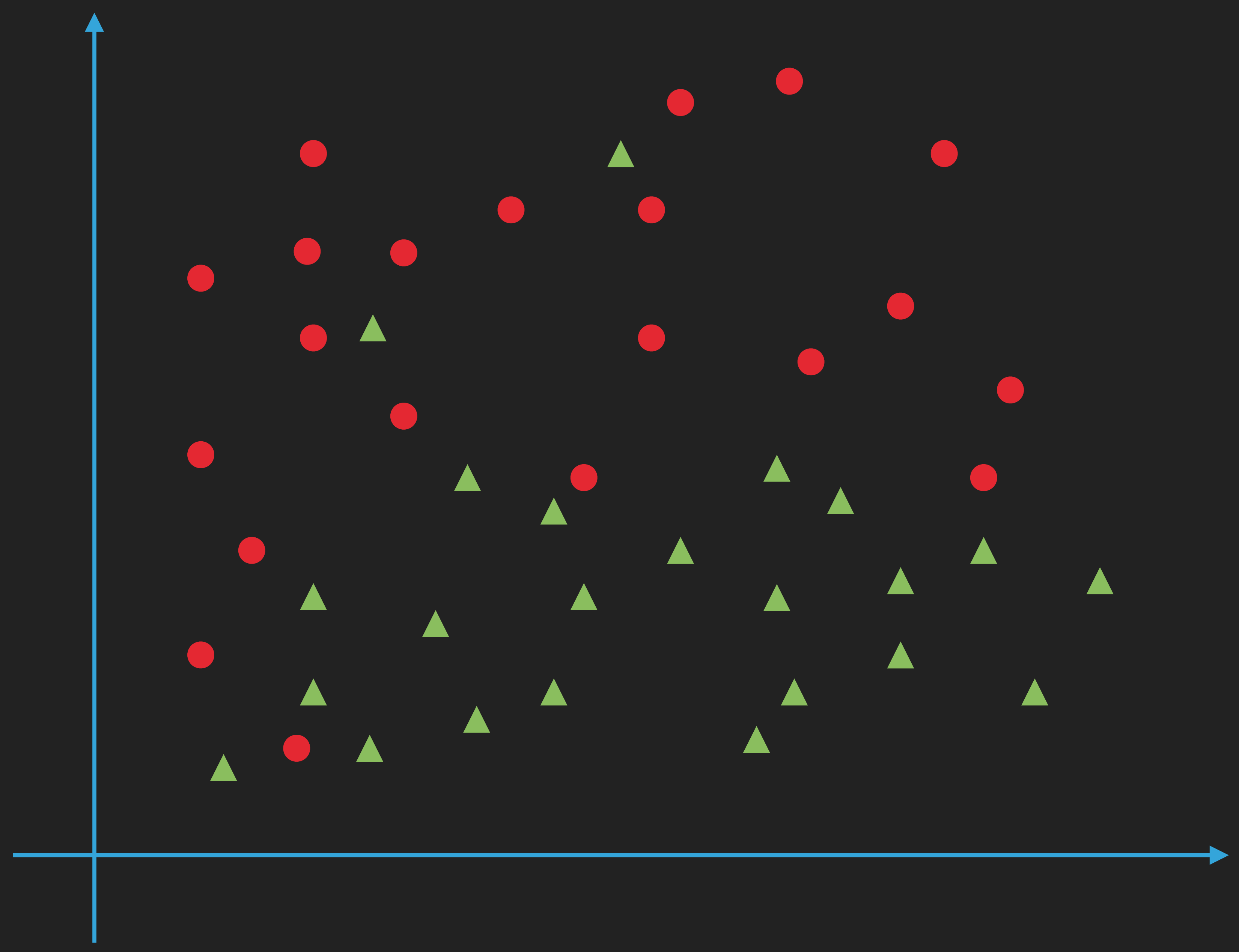
EXPERIMENT 1



HOW IT WORKS

- Permutation test
  - Labels in the learning set are shuffled

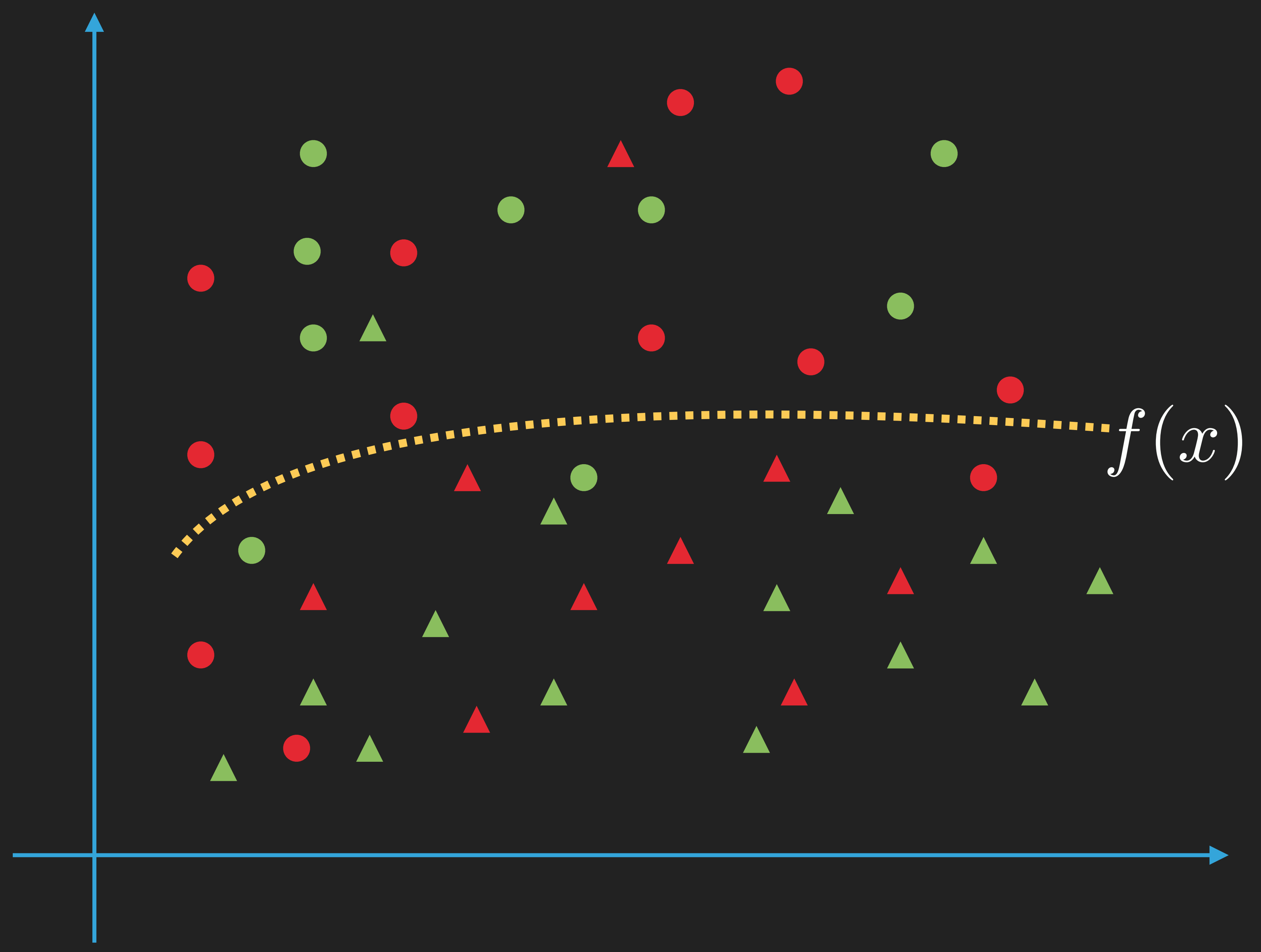
EXPERIMENT 1: LEARNING SET



HOW IT WORKS

- Permutation test
  - Labels in the learning set are shuffled

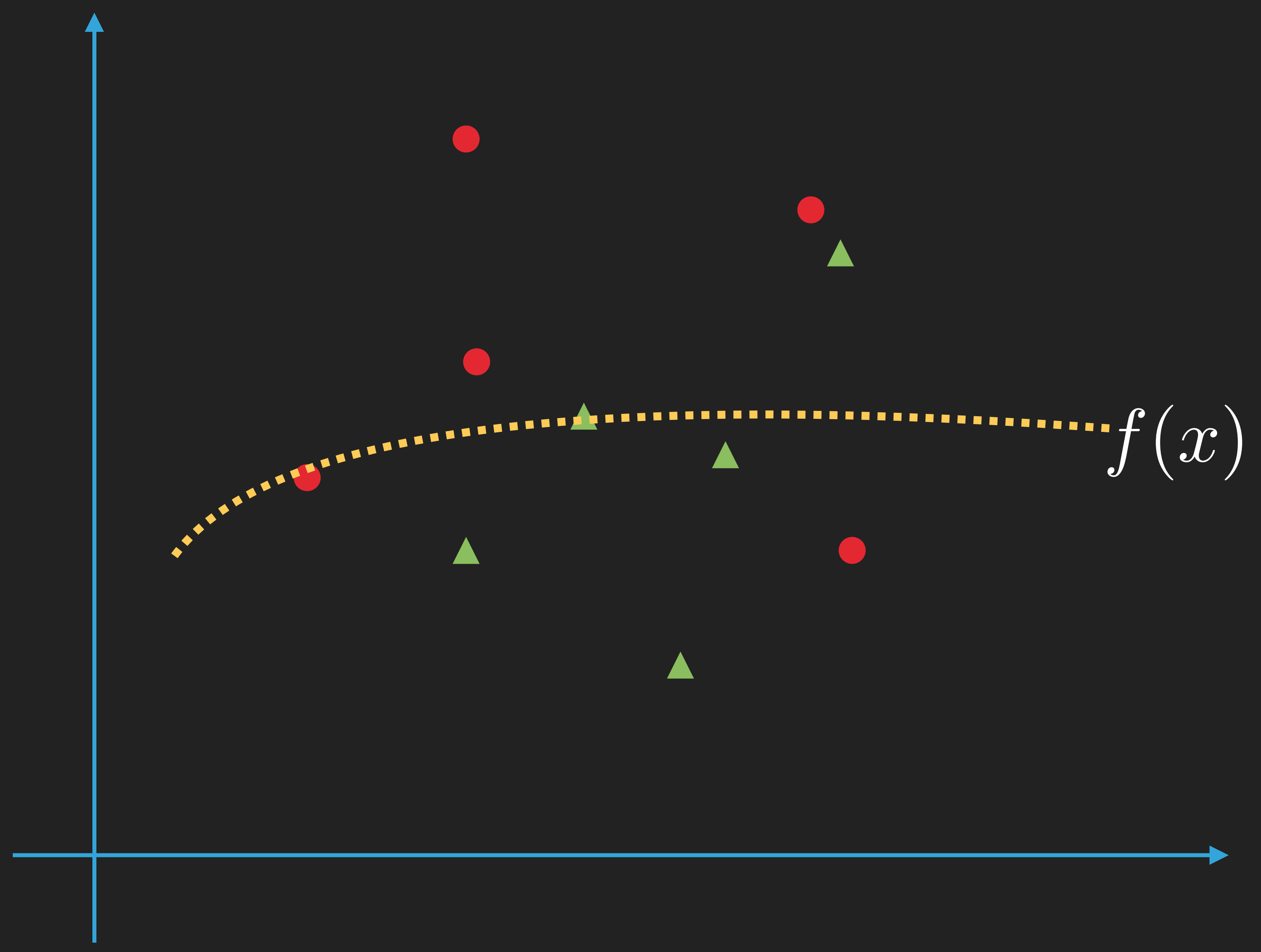
EXPERIMENT 1: LEARNING SET



HOW IT WORKS

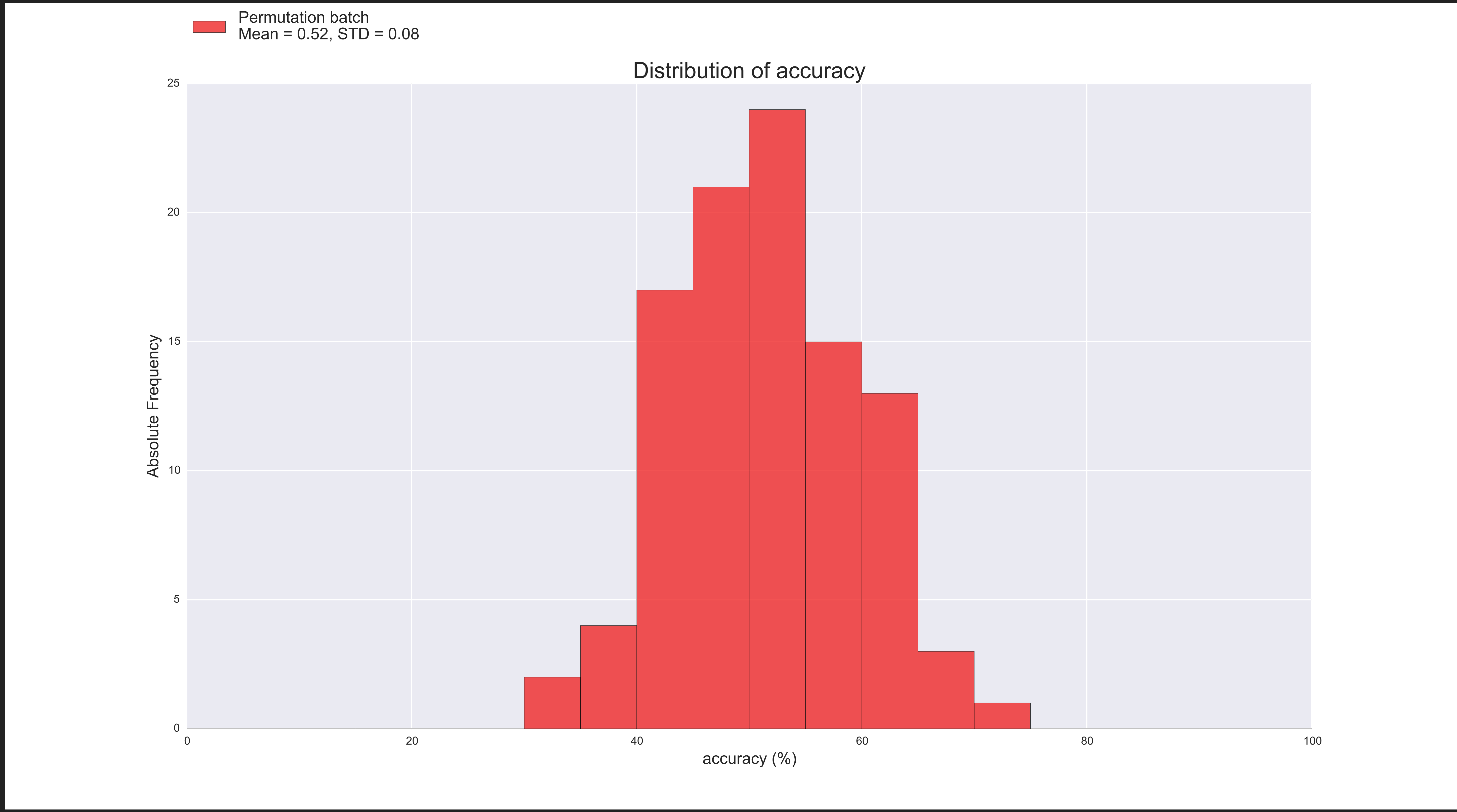
- Permutation test
  - Labels in the learning set are shuffled

EXPERIMENT 1: TEST SET

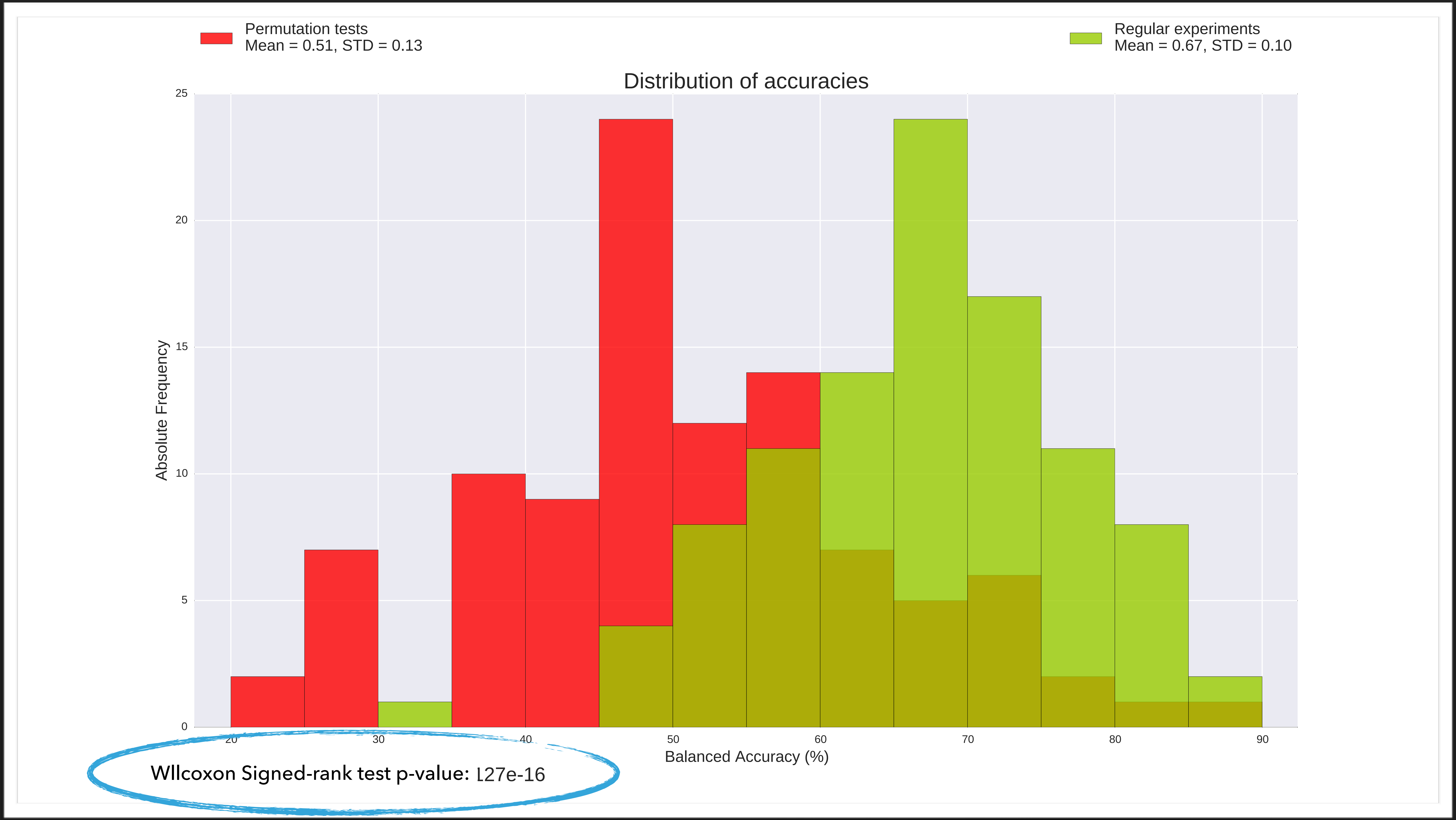


HOW IT WORKS

- Permutation test
  - Labels in the learning set are shuffled



RESULTS - DISTRIBUTION OF ACCURACIES



RESULTS - VARIABLES SELECTION FREQUENCIES

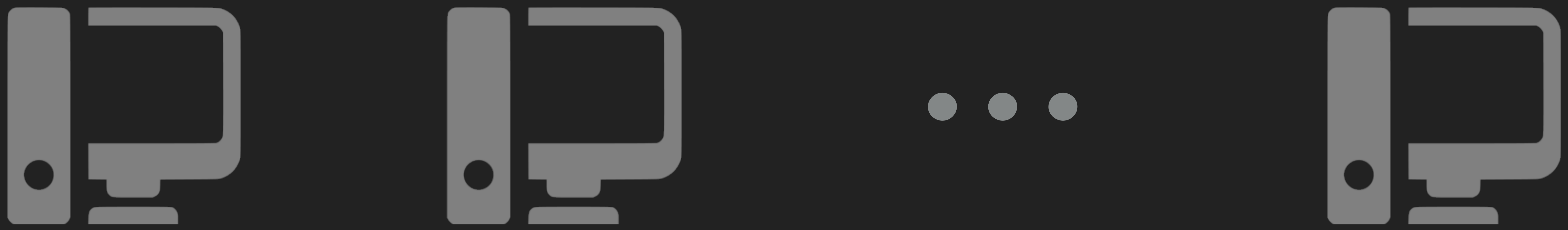
EXPERIMENT 1	VAR #4, VAR #8, VAR #71, …
EXPERIMENT 2	VAR #4, VAR #29, VAR #17, …
…	VAR #3, VAR #78, VAR #2, …
EXPERIMENT 100	VAR #4, VAR #17, VAR #31, …

RESULTS – VARIABLES SELECTION FREQUENCIES

VARIABLE	FREQUENCY
VAR #4	97%
VAR #7	91%
VAR #29	83%
VAR #17	78%
…	…
VAR #78	12%

WORKLOAD DISTRIBUTION

MPI is used to distribute jobs in a cluster



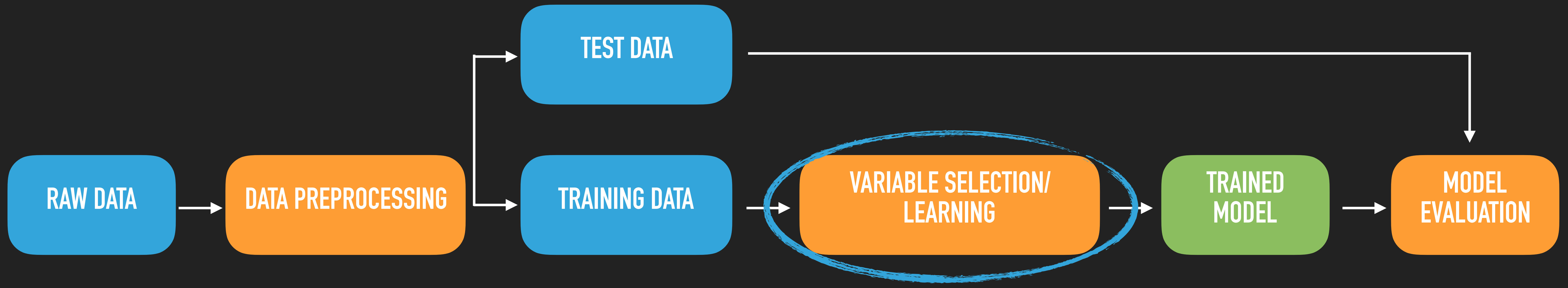
Experiment #1
Experiment #1
Experiment #2
Experiment #2
...
...
Experiment #20
Experiment #20

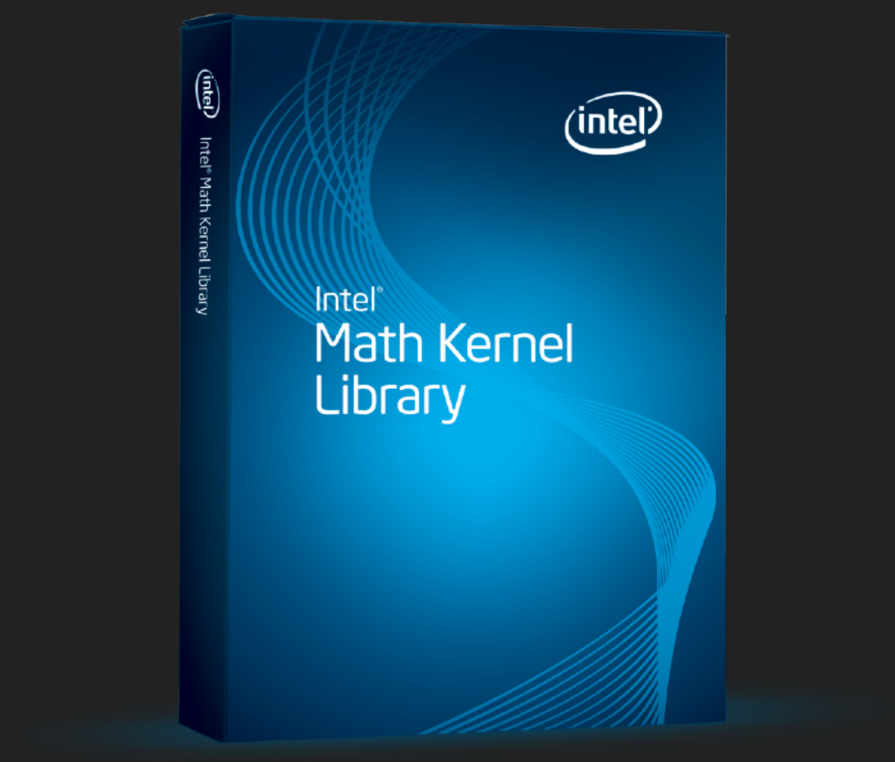
Experiment #21
Experiment #21
Experiment #22
Experiment #22
...
...
Experiment #40
Experiment #40

Experiment #81
Experiment #81
Experiment #82
Experiment #82
...
...
Experiment #100
Experiment #100

MKL/CUBLAS ACCELERATION

Libraries for linear algebra





NVIDIA
CUDA

OUR MACHINES

2 X 

Intel® Xeon® CPU E5-2630 v3
8 cores 2.4 GHz
32 GB of RAM
NVIDIA Quadro K2200

1 X 

Two Intel® Xeon® CPUs E5-2630 v3
8 cores 2.4 GHz (each)
128 GB of RAM
NVIDIA Tesla K40c

VALIDATION ON SYNTHETIC DATASETS

REAL VALUED DATASETS - DATA

- 45 datasets of different sizes:

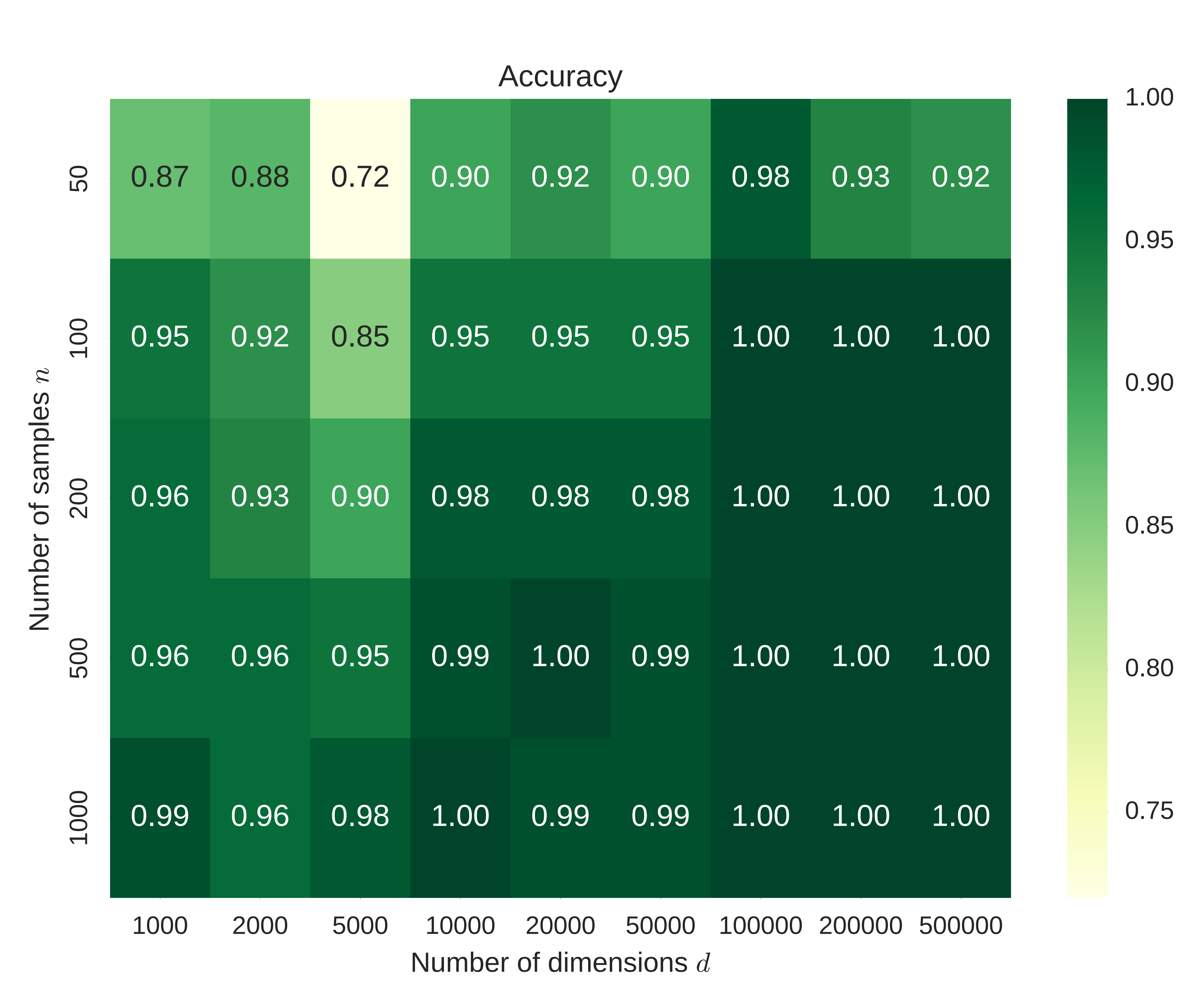
  $50 \leq n \leq 1000$

  $1000 \leq d \leq 500000$

- Binary classification task
- The number of relevant variables was proportional to the data dimensionality (between 25 and 100 relevant variables)

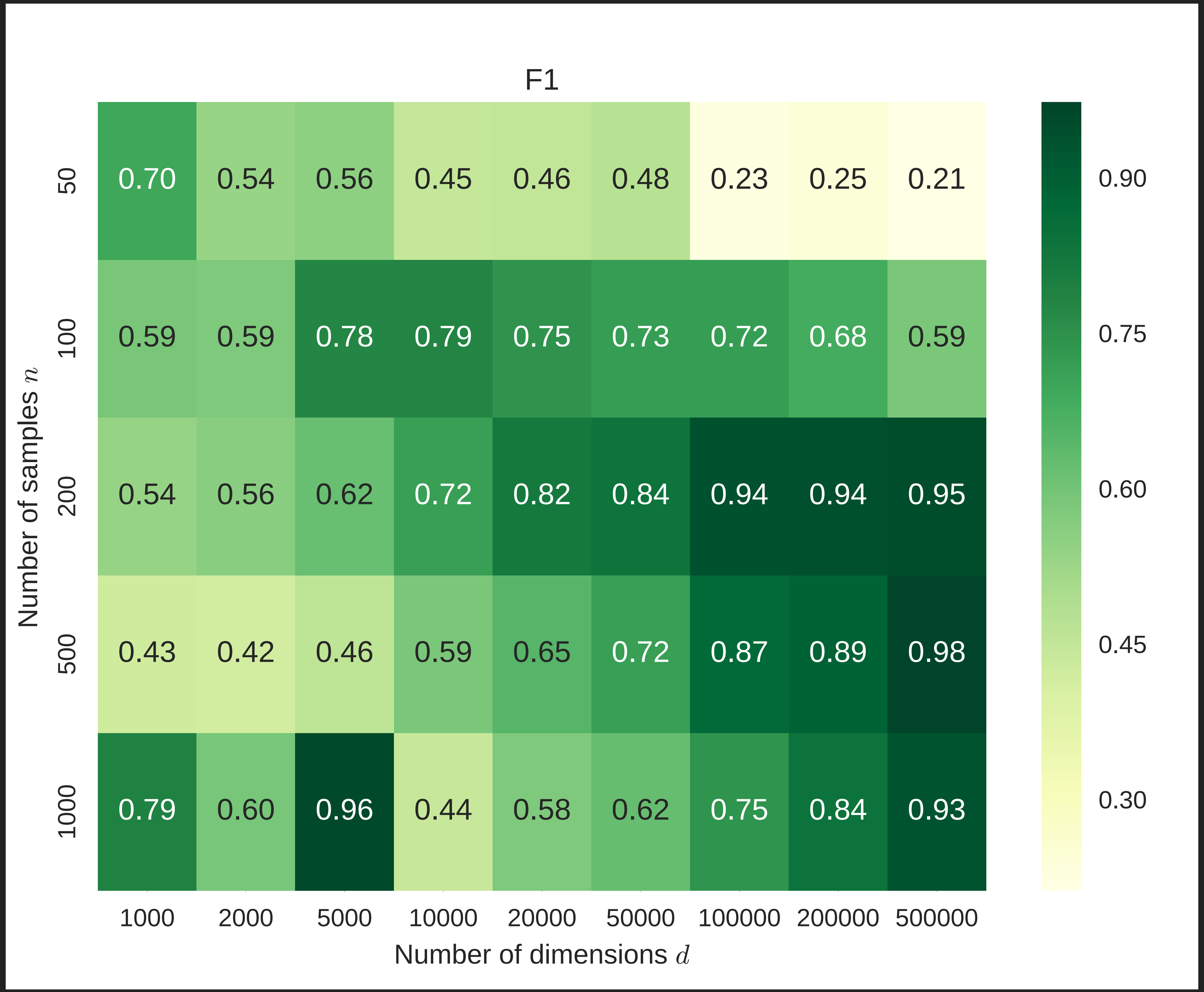
SYNTHETIC DATASETS – ACCURACY RESULTS

Accuracy values for all 45 experiments

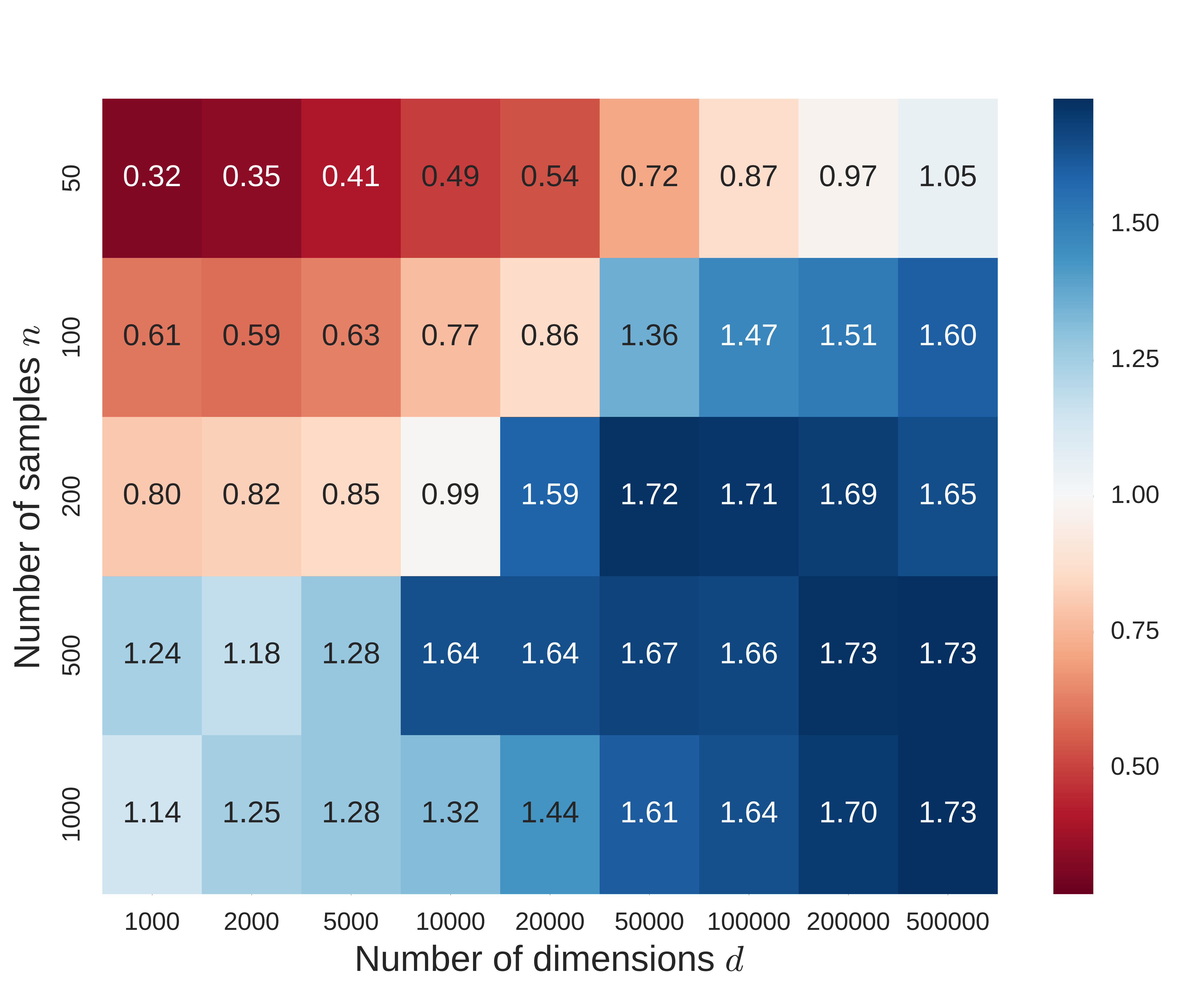


SYNTHETIC DATASETS – VARIABLE SELECTION RESULTS

F1 Scores for all 45 experiments



CPU VS GPU SPEEDUP



CONCLUSIONS AND FUTURE WORK

CONCLUSIONS

- The framework allowed us to obtain clear results (either positive or negative) on both synthetic and real datasets
- The MPI implementation allows for the analysis of large datasets in a reasonable amount of time

FUTURE WORKS

- Extension to the regression problems
- Inclusion of wrappers for more learning algorithms
- A different job distribution strategy (Master/Slave) for heterogeneous clusters
- Obtaining a single model through a final training step

AKNOWLEDGEMENTS

THANKS!

[HTTPS://GITHUB.COM/SLIPGURU/PALLADIO](https://github.com/slipguru/palladio)

Matteo Barbieri, 2016