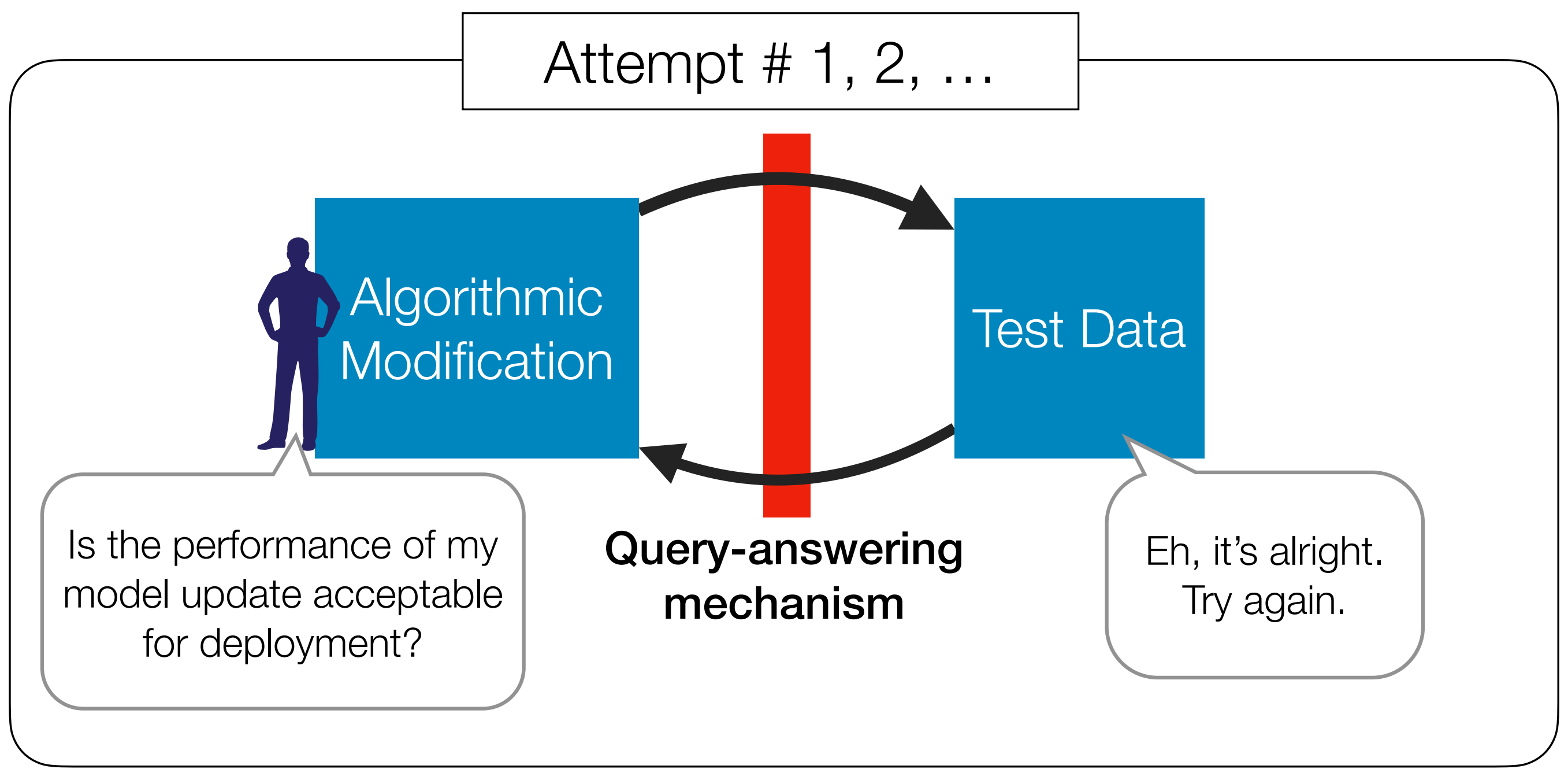
Sequential Algorithmic Modification with Test Data Reuse

Jean Feng[1], Gene Pennello[2], Nicholas Petrick[2], Berkman Sahiner[2], Romain Pirracchio[1], Alexej Gossmann[2]

[1]University of California, San Francisco, [2]U.S. Food and Drug Administration

Introduction

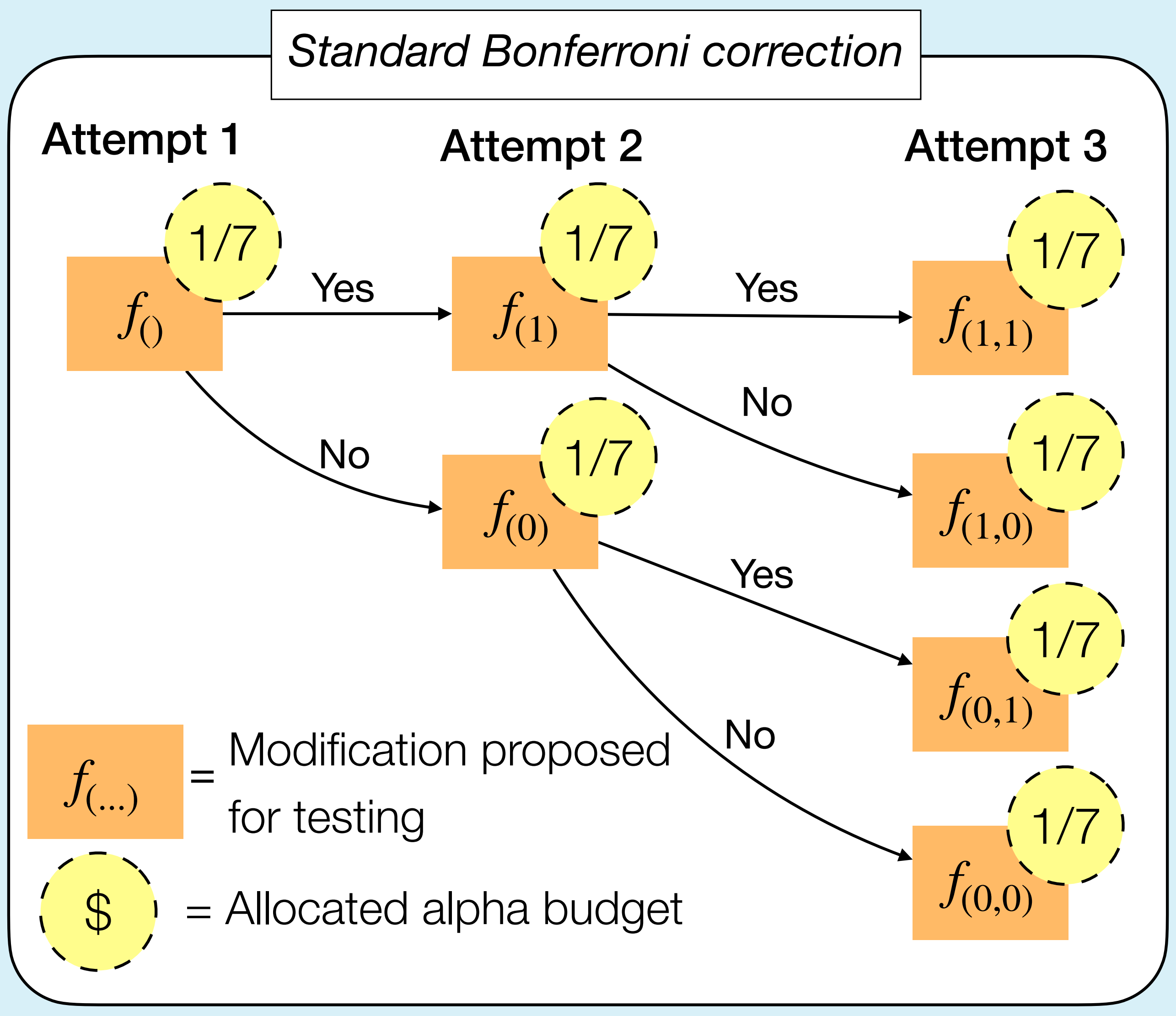
- After initial release of a ML algorithm, a model developer may choose to update the algorithm, e.g. retrain it on subsequently collected data or add newly discovered features.
- Because each modification introduces a risk of deteriorating performance, it must be validated on a test dataset.
- In cases where it is not practical to assemble a new dataset to test each modification, ***how can we reuse an existing test dataset to validate sequential algorithmic modifications?***


Attempt # 1, 2, …
Algorithmic Modification
Test Data
Query-answering mechanism
Is the performance of my model update acceptable for deployment?
Eh, it's alright. Try again.

Background

- Under uncontrolled test data reuse, the model developer can learn too much information about the test data with each query. This can lead to the approval of models that are overfit to the test data.
- Methods that allow for valid test data reuse restrict the amount of information leaked with each query, either by perturbing the query result with random noise (i.e. differential privacy) or by restricting the number of bits of information returned.
- Existing methods require *large* datasets. Our interest is in methods suitable for ***small*** datasets.

Multiple testing framework


Standard Bonferroni correction

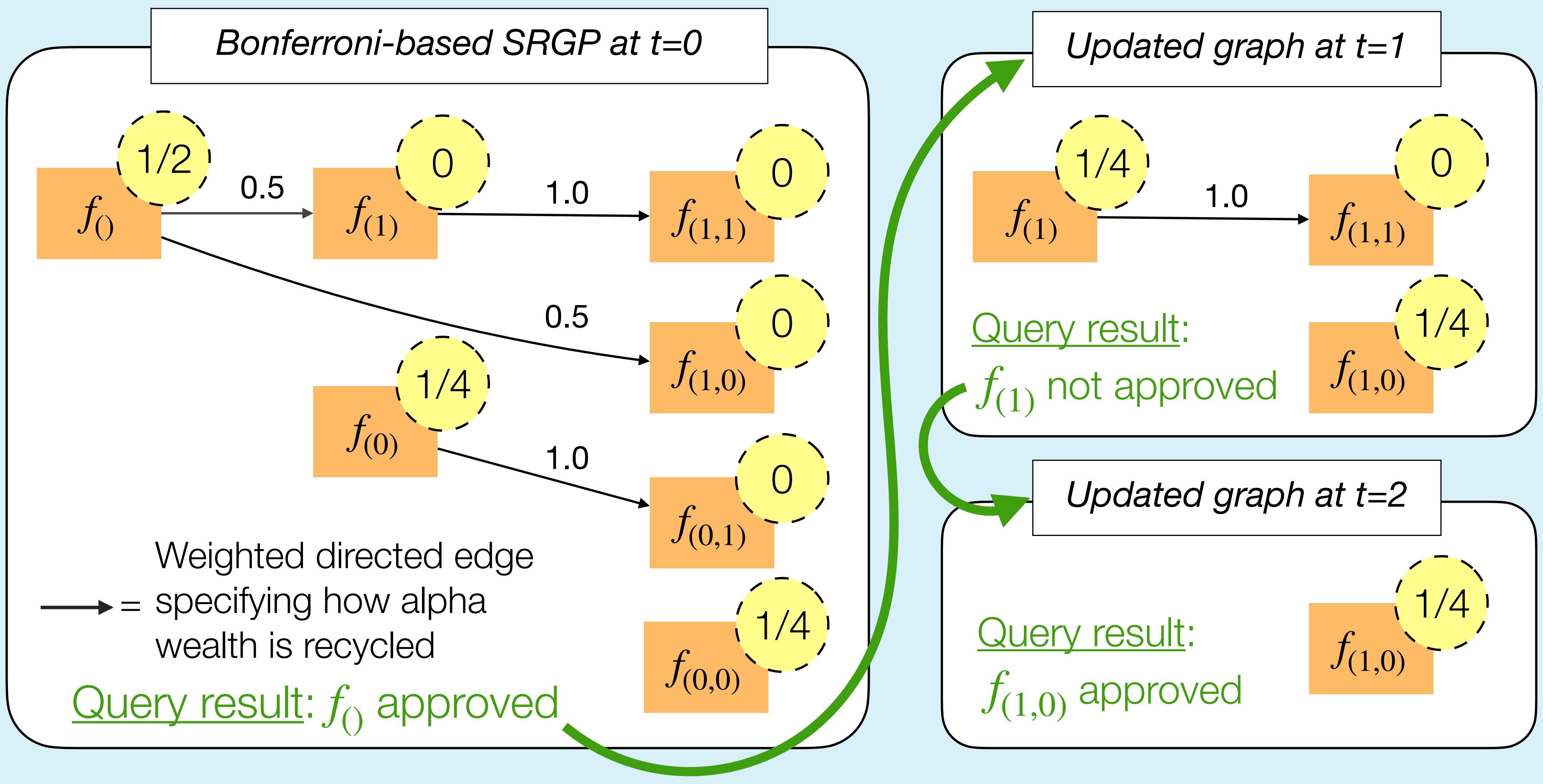
$f_{(...)}$ = Modification proposed for testing

$ = Allocated alpha budget

- The adaptive test strategy can be viewed as a tree of hypothesis tests. Control of the Family-Wise Error Rate (FWER) implies control over the probability of inappropriately approving at least one algorithmic modification.
- Recent work has proposed to control the FWER using a standard Bonferroni correction. However, Bonferroni corrections (standard or weighted) are known to be overly conservative.

Sequentially Rejective Graphical Procedures (SRGPs)

- SRGPs perform alpha recycling, in which the alpha allocated to a rejected null hypothesis can be "recycled" to test a subsequent hypothesis. The resulting procedures are more powerful for approving algorithmic modifications.


Bonferroni-based SRGP at t=0

→ = Weighted directed edge specifying how alpha wealth is recycled

Query result: $f_{()}$ approved

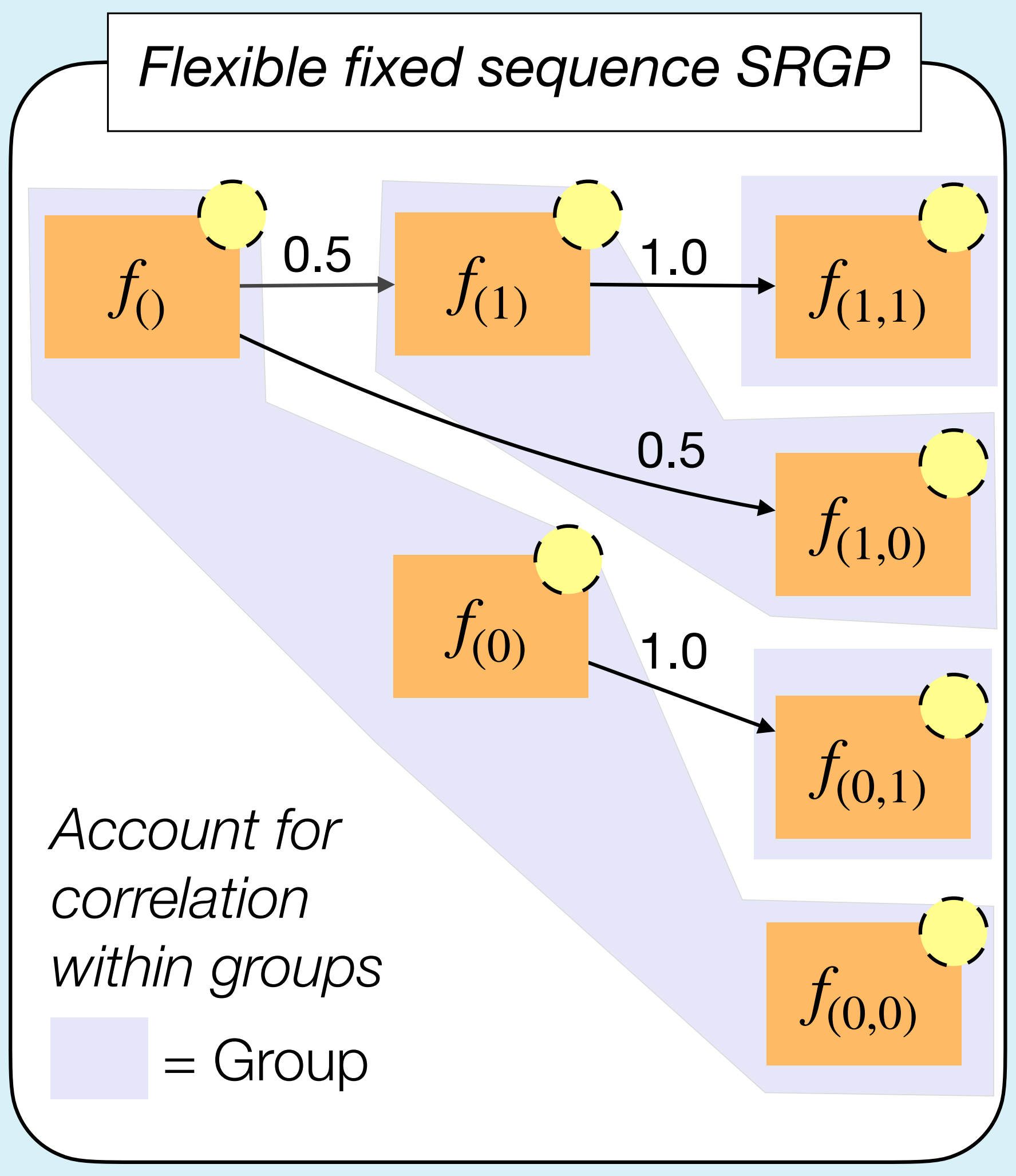
Updated graph at t=1

Query result: $f_{(1)}$ not approved

Updated graph at t=2

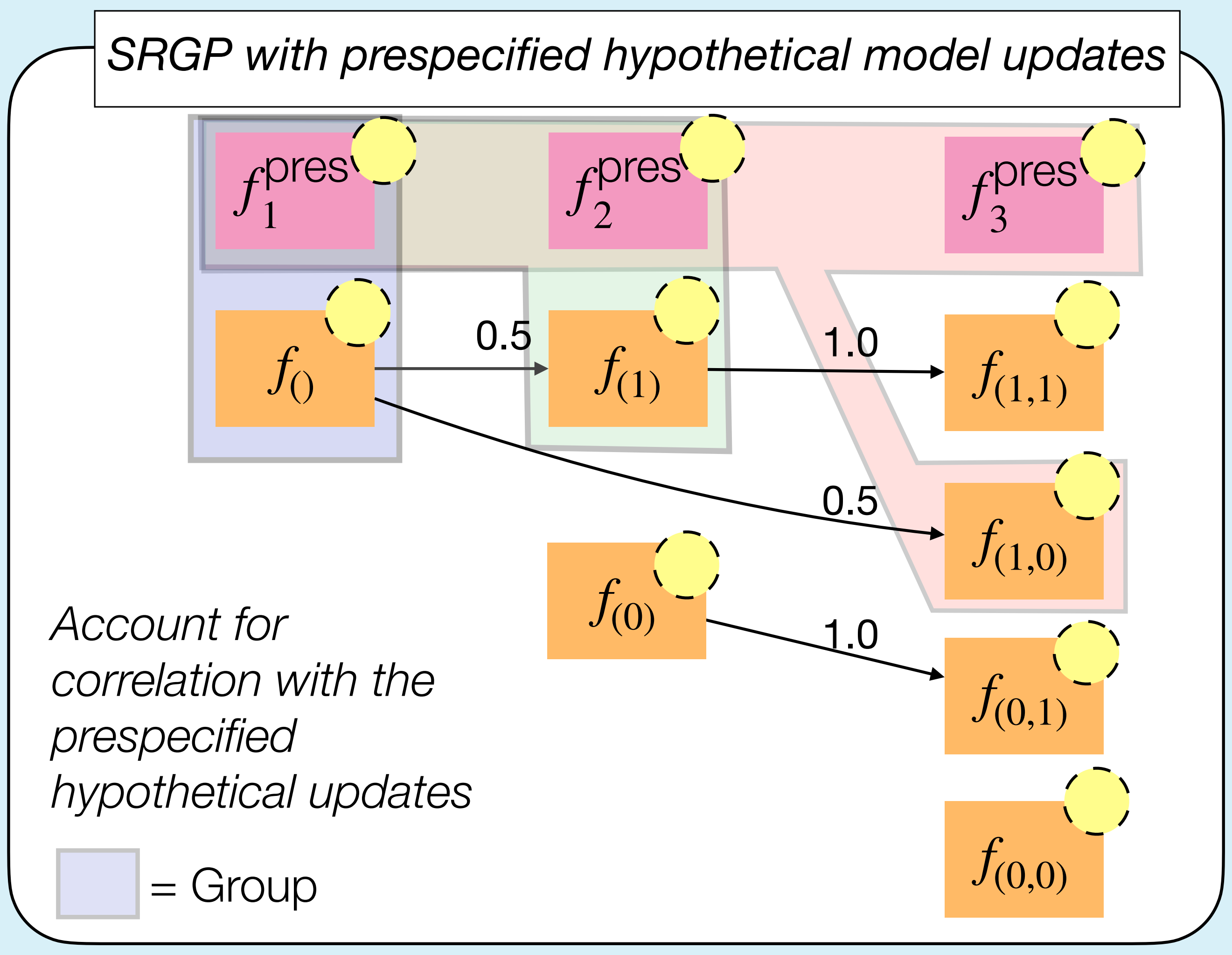
Query result: $f_{(1,0)}$ approved

- In practice, we expect many modifications to be highly correlated. We can design more powerful SRGPs that account for correlation between model updates. Care must be taken since we do not observe the "counterfactual" modifications.


Flexible fixed sequence SRGP

Account for correlation within groups

▢ = Group

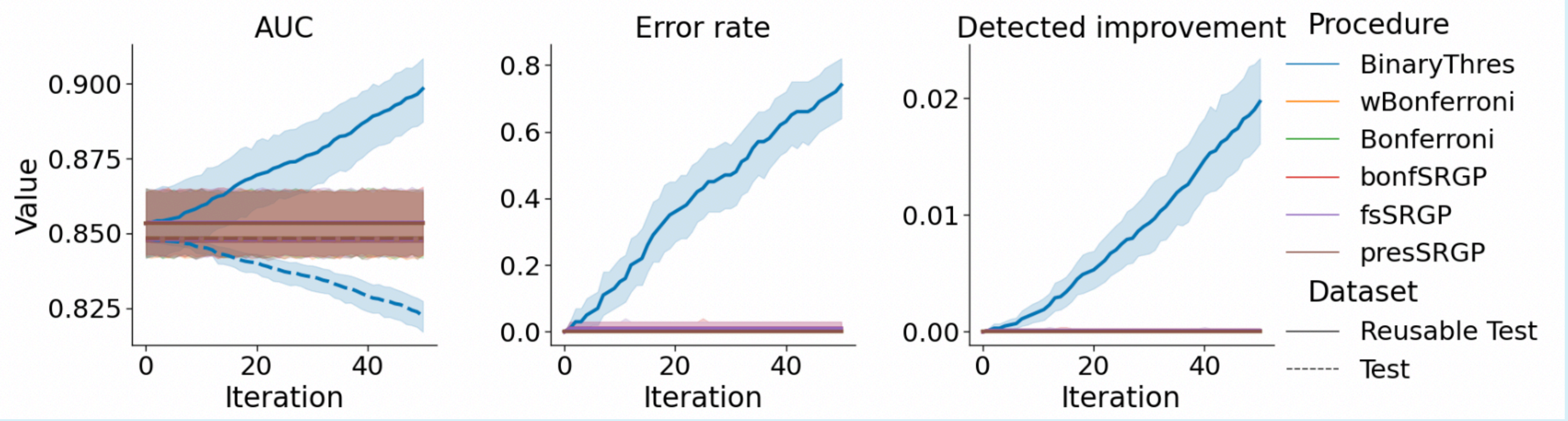

SRGP with prespecified hypothetical model updates

Account for correlation with the prespecified hypothetical updates

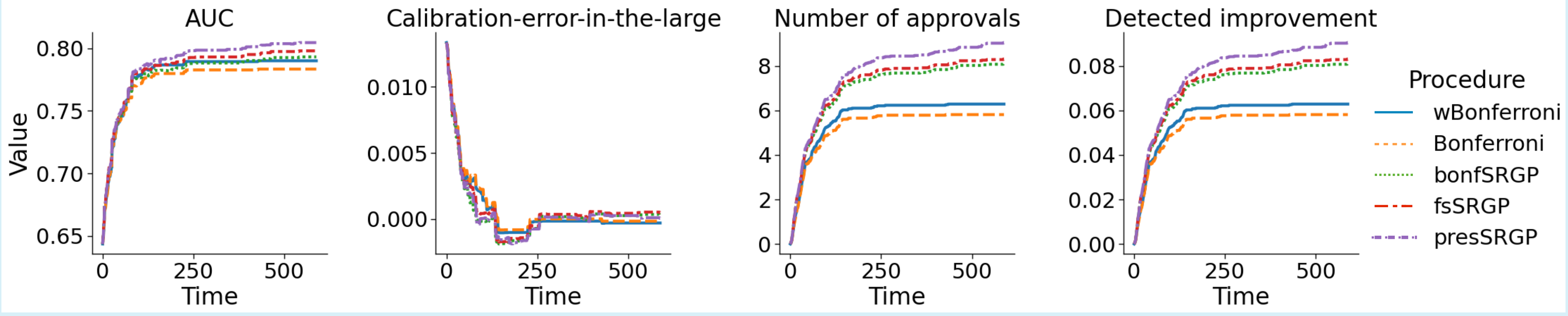
▢ = Group

Experiments

Experiment 1: Approval of algorithmic modifications increasingly overfit to the test data, as generated by an adversarial model developer. BinaryThres does not control FWER so it incorrectly approves many model updates. The SRGPs and Bonferroni-based procedures appropriately control the probability of approving bad updates.



Experiment 2: Approval of gradient boosted trees (GBTs) for predicting acute hypotension episodes in the eICU dataset. GBTs are continually retrained on a stream of patient data. Performance is evaluated on independent test data. SRGP with prespecified hypothetical updates (presSRPG) approves the most model updates and achieves the best performance. (n=500)



Funding: This work was supported by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award Center of Excellence in Regulatory Science and Innovation grant to University of California, San Francisco (UCSF) and Stanford University, U01FD005978. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.