AbcRanger: a fast and scalable random forest library for ABC model choice and parameter estimation

F.-D. Collin [2] A. Estoup [1] J.-M. Marin [2] L. Raynal [2]

[1]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ. Montpellier [2]Université de Montpellier, CNRS, IMAG UMR 5149

First building block : ABC simulations

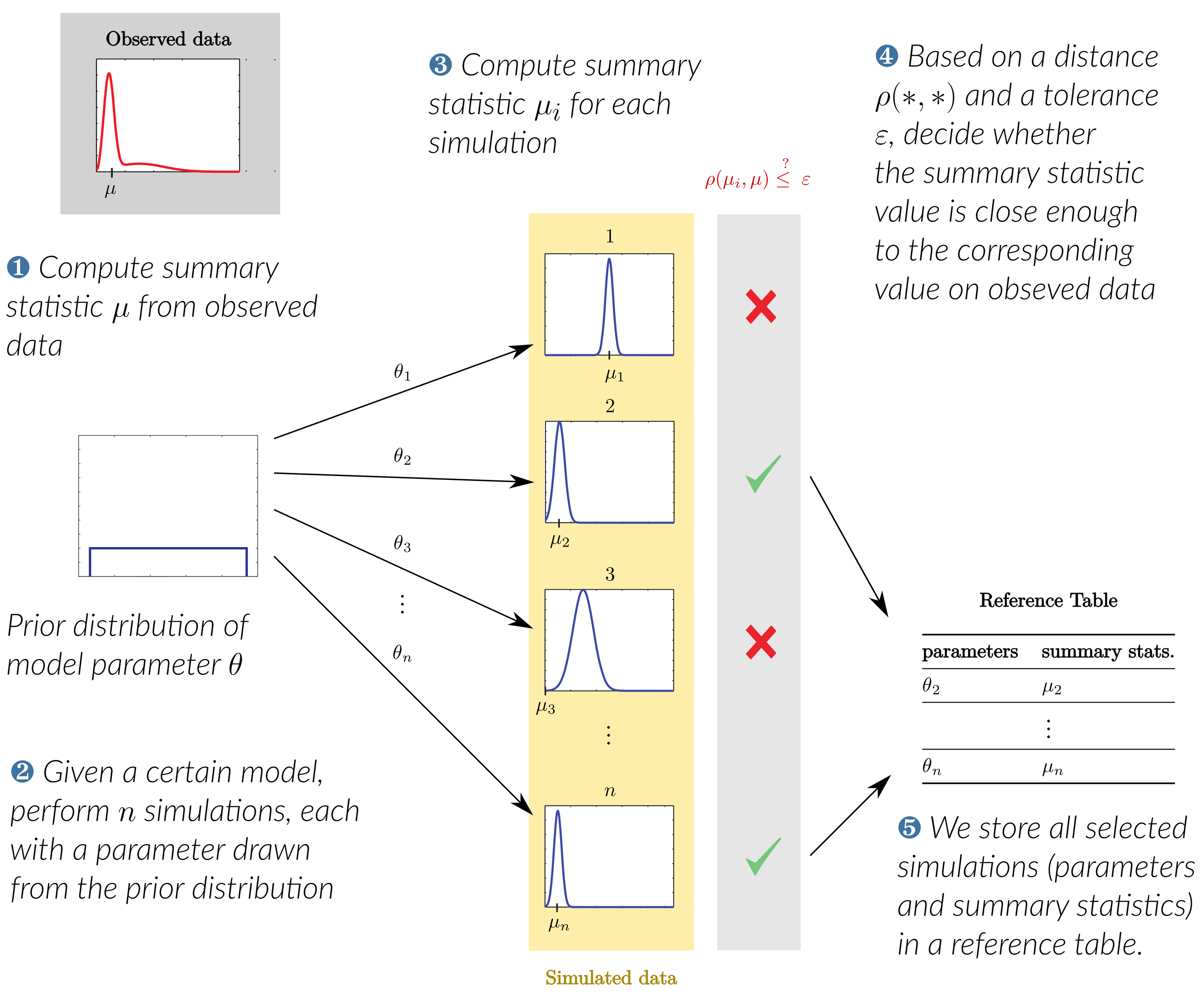


Figure 1. ABC details

Given an observed data, the basic idea of ABC (*Approximate Bayesian Computations*) is to approximate the likelihood of a parametrized model with selected simulations, by comparing the observed data and simulated ones via computed *summary statistics*. The table of summary statistics for simulated data is called the reference table.

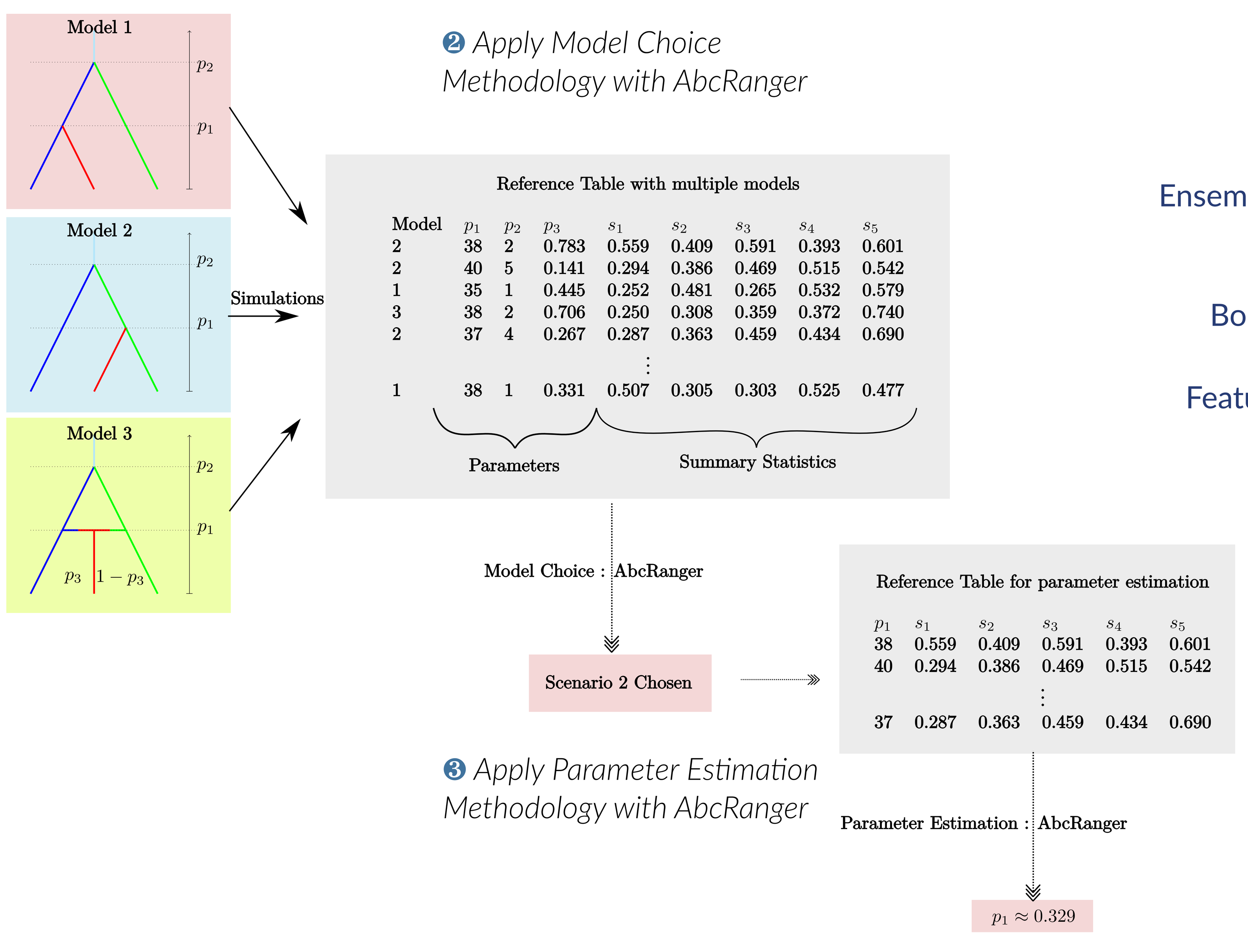
ABC posterior methodologies

Model choice: Simulate data for several models and *choose the best model to fit our data*

Parameter estimation: Simulate data for one model and *infer one or several parameters for this model given the observed data*

A sensible workflow is to first choose a model and then infer its parameters.

➊ Compute simulations with several models, and the reference table with model-indexed lines using a simulator (DIYAC, PyABC etc.)



Reference Table with multiple models

Model	p_1	p_2	p_3	s_1	s_2	s_3	s_4	s_5
2	38	2	0.783	0.559	0.409	0.591	0.393	0.601
2	40	5	0.141	0.294	0.386	0.469	0.515	0.542
1	35	1	0.445	0.252	0.481	0.265	0.532	0.579
3	38	2	0.706	0.250	0.308	0.359	0.372	0.740
2	37	4	0.267	0.287	0.363	0.459	0.434	0.690
				⋮				
1	38	1	0.331	0.507	0.305	0.303	0.525	0.477

Parameters / Summary Statistics

Reference Table for parameter estimation

p_1	s_1	s_2	s_3	s_4	s_5
38	0.559	0.409	0.591	0.393	0.601
40	0.294	0.386	0.469	0.515	0.542
			⋮		
37	0.287	0.363	0.459	0.434	0.690

Figure 2. ABC workflow with AbcRanger

Challenges of ABC

in the context of population genetics recent advances

Number of simulated data : could be > 100 000

Number of summary statistics : could range from several hundred to tens of thousands (scenario with several populations and combinatory "explosion") : how to select the *meaningful* ones?

Classical Methods for ABC (k-nn and local methods) doesn't cope very well with this situation.

Our solution

[3] and [4] proposed a novel approach, relying on *Random Forests* to provide both model choice and parameter estimation methodologies

Second building block : Random Forests

CART

Random Forests are based on a CART (*Classification and Regression Trees*) algorithm.

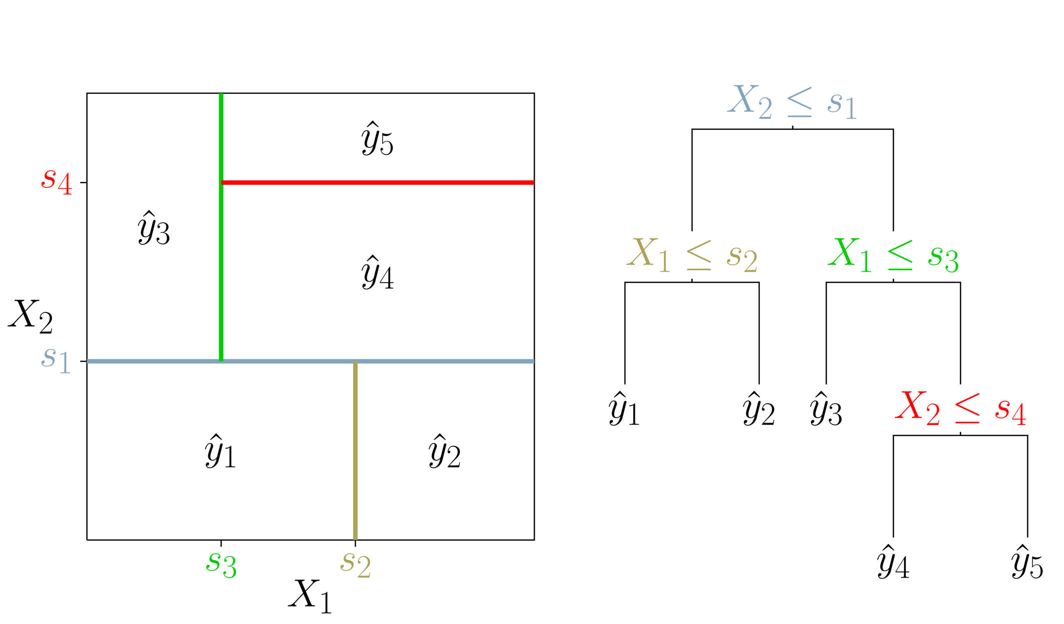
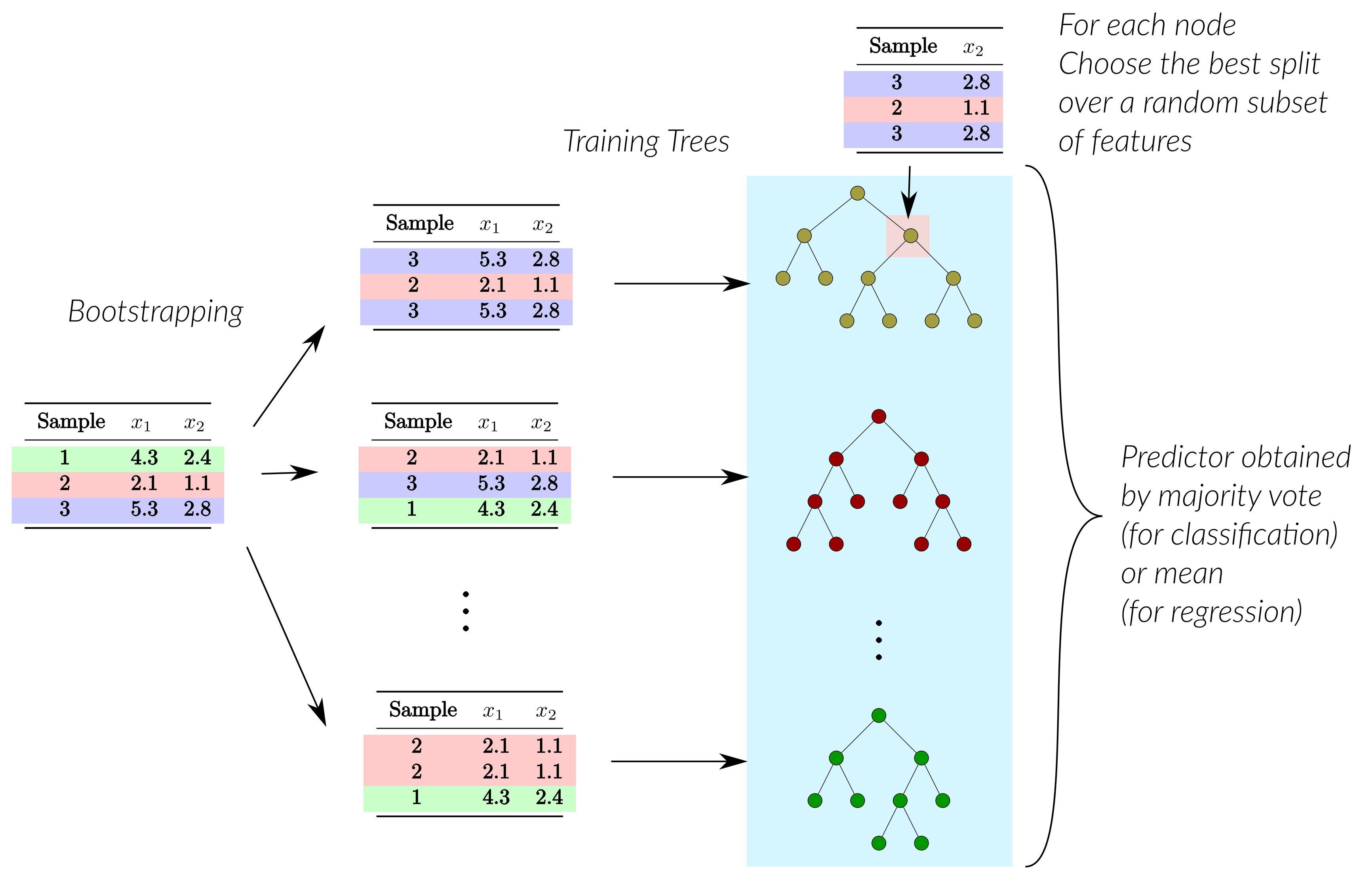


Figure 3. An example of CART and the associated partition of the two dimensional predictor space. Each splitting condition takes the form $X_j \leq s$ and the prediction at a leaf is denoted $\hat{y}_\ell$.

A CART is a *machine learning algorithm* whose principle is to partition the predictor space into disjoint subspaces, in an iterative manner, and each one is assigned a prediction value which will be used for test data falling in this subspace.
Once the partitioning is done, we have a binary tree structure which could predict outcomes from an input data, either classes or continuous values.

Random Forests



Sample	x_1	x_2
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Sample	x_1	x_2
3	5.3	2.8
2	2.1	1.1
3	5.3	2.8

Sample	x_1	x_2
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Sample	x_1	x_2
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

Sample	x_2
3	2.8
2	1.1
3	2.8

Figure 4. Random Forest

Random Forests [1] are a three pronged extension of CART:

Ensemble method Training a *set* of CART (not just one), and getting the majority vote (resp. mean) for classification (resp. regression)

Bootstrapping Training data is random sampled (with replacement) for *each* tree

Feature bagging At each node of a growing tree, find the best split on a random subset of the features

Advantages in an ABC setting :

- robust to noise
- (almost) free variable importance
- free (out-of-bag) cross-validation procedure
- easy parallelization
- good scaling properties on number of samples and on number of features (summary statistics)
- classifier and regressor

Computational challenges with ABC/Random Forests

With 100 000 lines and more than 10 000 summary statistics, each tree could reach over 1 gigabyte of memory size. Typically we need 500 or 1000 trees for good prediction performance, so, even with state of the art RF packages like [5], memory constraints are preventing completion of the training.

A new implementation of Random Forest for ABC

Since ABC procedures only use trained Random Forests on a known set of observations, we have altered the random forest training computation by using only a subset of in-memory trees at a time and accumulating the required outcomes (predictions and statistics)

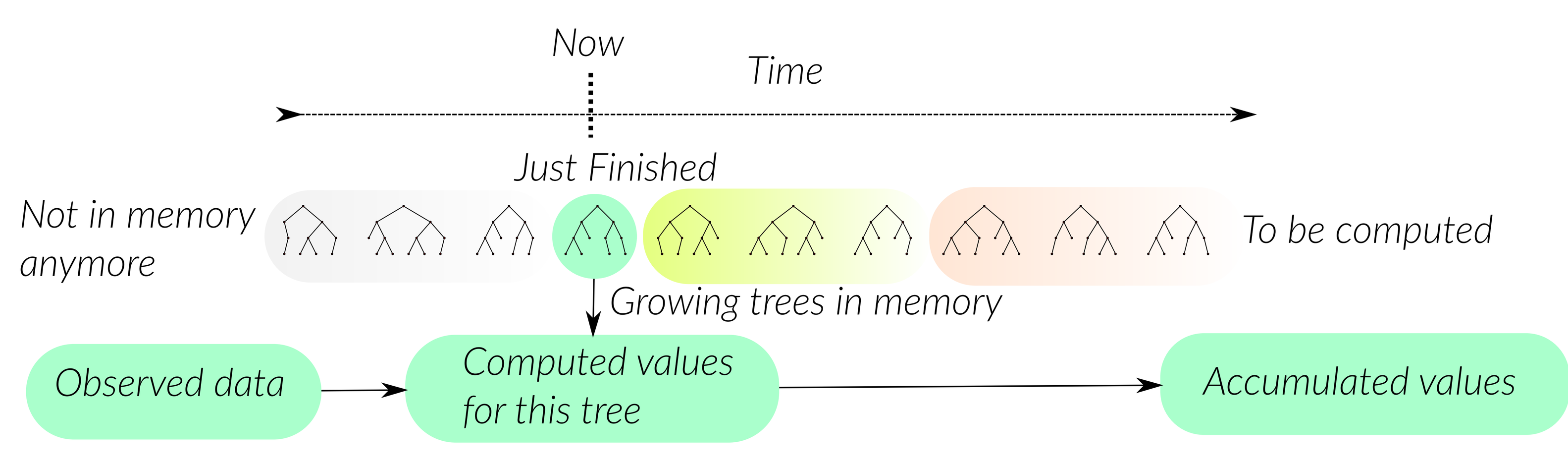


Figure 5. Window of growing trees

References

[1] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis, 1984.

[3] Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable abc model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

[4] Louis Raynal, Jean-Michel Marin, Pierre Pudlo, Mathieu Ribatet, Christian P Robert, and Arnaud Estoup. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728, 10 2018.

[5] Marvin N Wright and Andreas Ziegler. Ranger: a fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.