RANDOM FOREST & BAGS OF LITTLE BOOTSTRAP

Project_name <- randomForestBLB

Presenter <- Zihao_Xu

Agenda

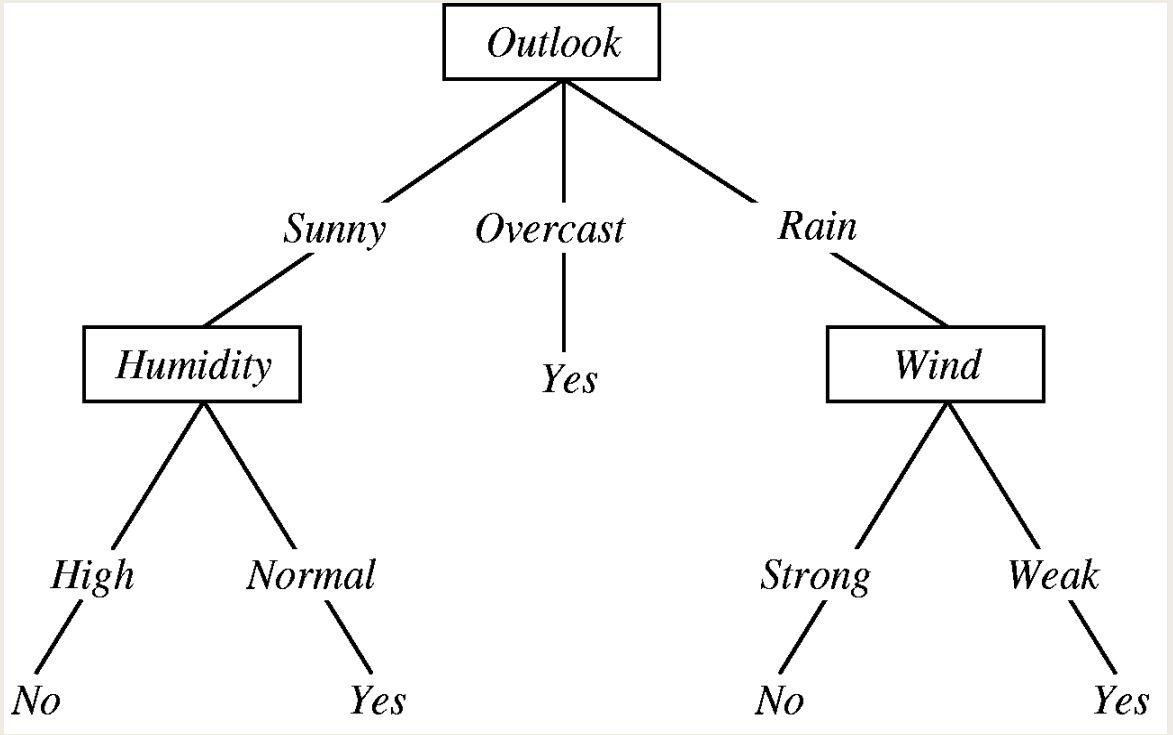
- Intro to Random Forest (RF) through CART
- Intro to Bags of little bootstrap (BLB)
- Integration: how can RF and BLB work together

What is Classification And Regression Tree (CART)?

- **What** Classification or Regression predictive model
- **How** Uses greedy binary splits
- **Adv.** Easy to understand, visualize, implement
- **Limitations** Variance & Overfitting

Illustration: Classification Tree

Obs.	Outlook	Humidity	Wind	Play?
1	Sunny	High	Normal	Yes
2	Overcast	High	Weak	Yes
3	Sunny	High	Strong	No
…	…	…	…	…
New Comer	Rain	Normal	Weak	???

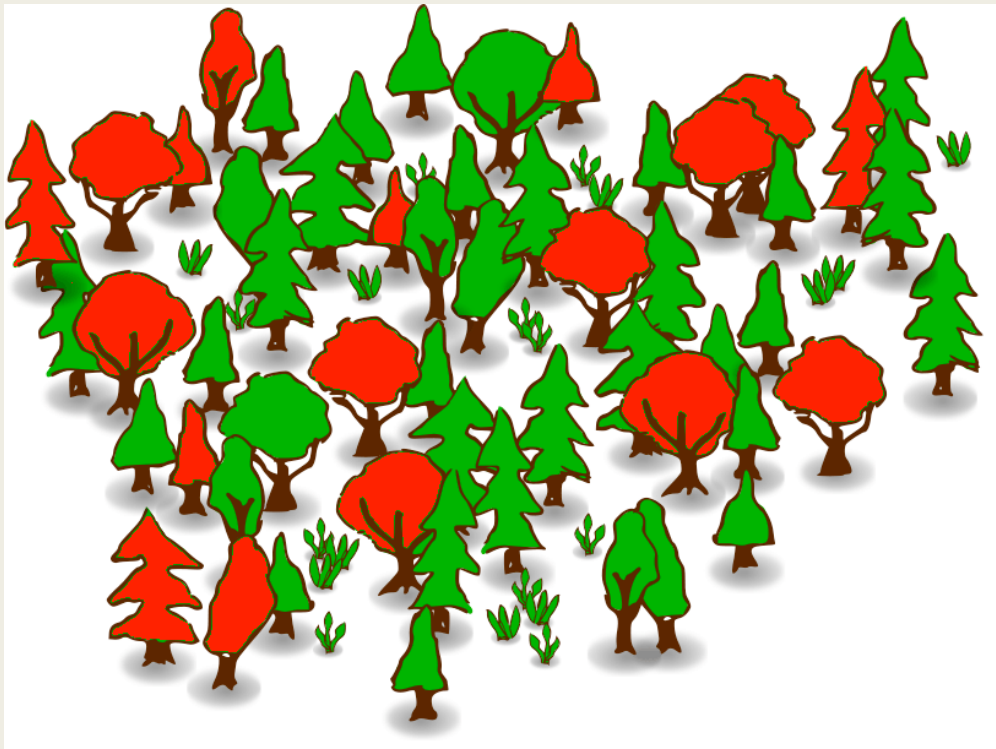


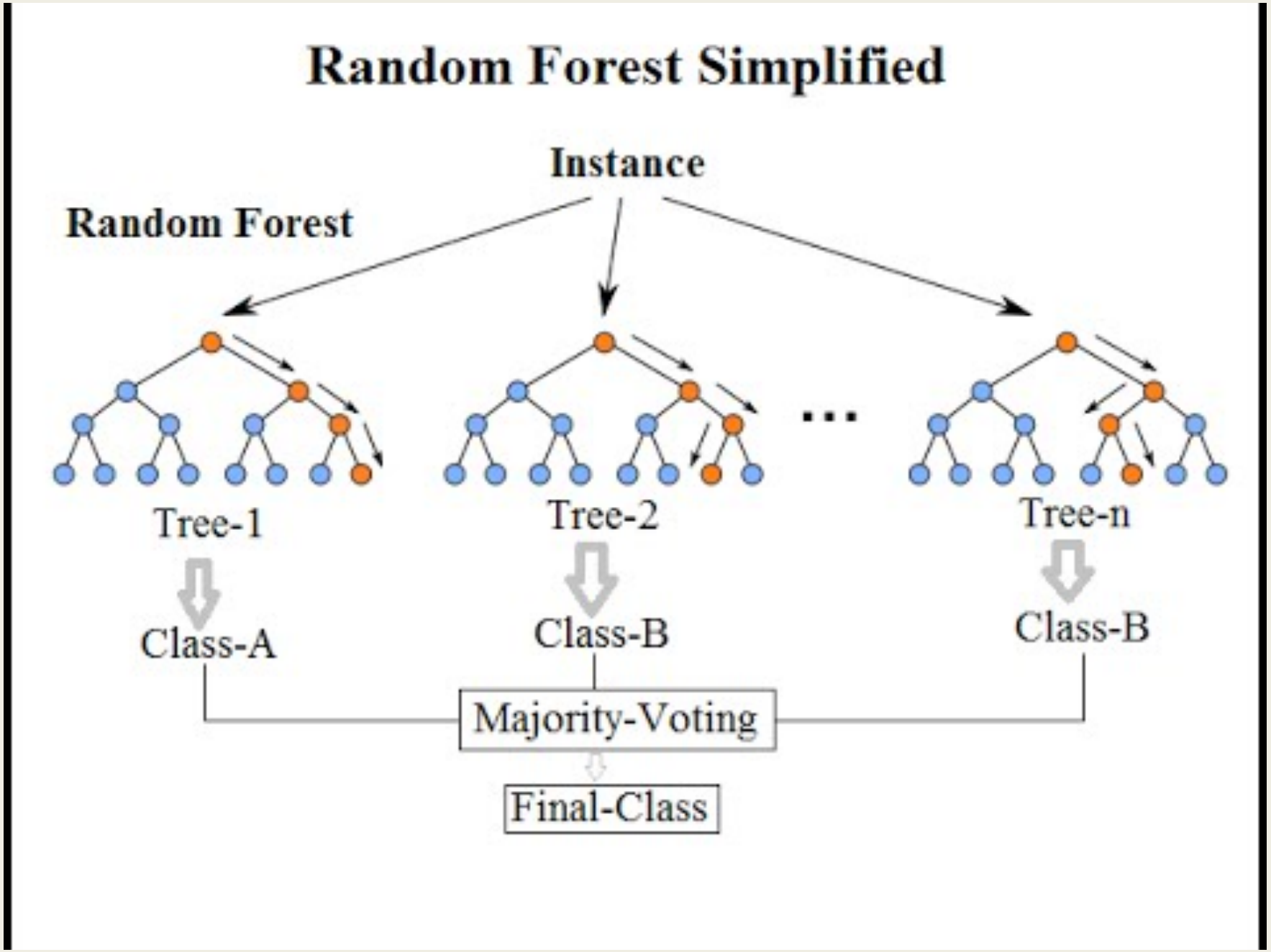
Yes!!!!!!

What is Random Forest?

Two Sources of Randomness

- **Bootstrap Aggregation (Bagging) :** randomized sample of rows in training set with replacement
- **Neglect some features:** uncorrelate the trees by randomly select some, but not all, variable in the construction of each tree




Random Forest Simplified
Instance
Random Forest
Tree-1
Tree-2
...
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class

Major Limitation of RF: Performance

- Physical memory requirement: need to store n distinct observations
- Greediness of CART: only wants the BEST split
- Takes forever to run RF on Large Datasets

Bags of Little Bootstraps (BLB)

- Introduced by *Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, Michael I. Jordan*
- To replace Bootstrapping!
- Steps: **Subsample** ------ > **Resample** ---------- > **Combine** results of resamples --------- > **Combine** results of subsamples

Illustration : Subsample *(Just for intuition!!!)*

Each letter is an observation

Sample size, n = 9:

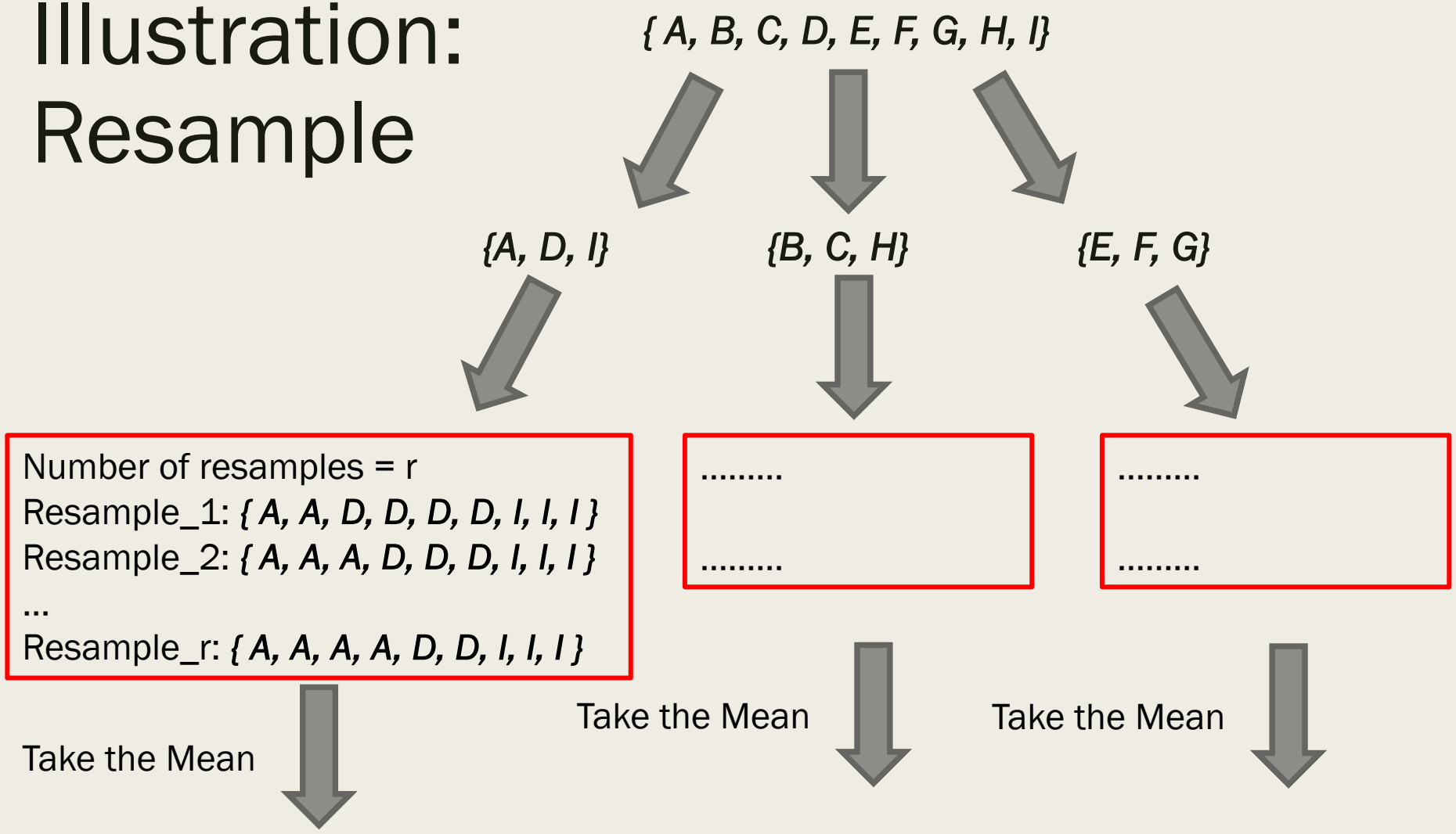
{ A, B, C, D, E, F, G, H, I }

Number of subsamples, s = 3

Size of each subsample, b = 3

{A, D, I} ***{B, C, H}*** ***{E, F, G}***

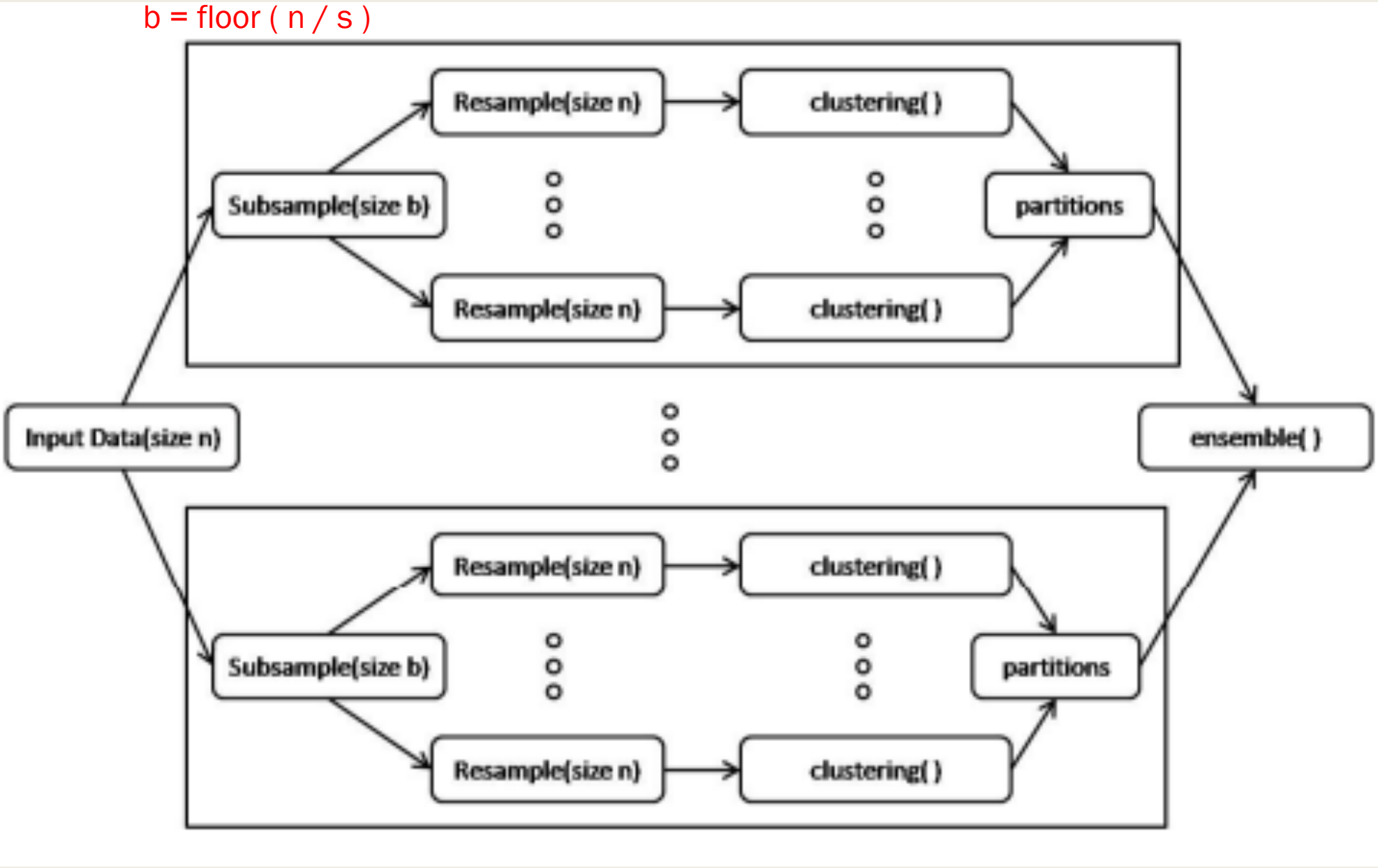
Illustration: Resample



(Result_1 + Result_2 + Result_3) / 3 = **RESULT!!**

Subsample

b = floor (n / s)



Why use BLB?

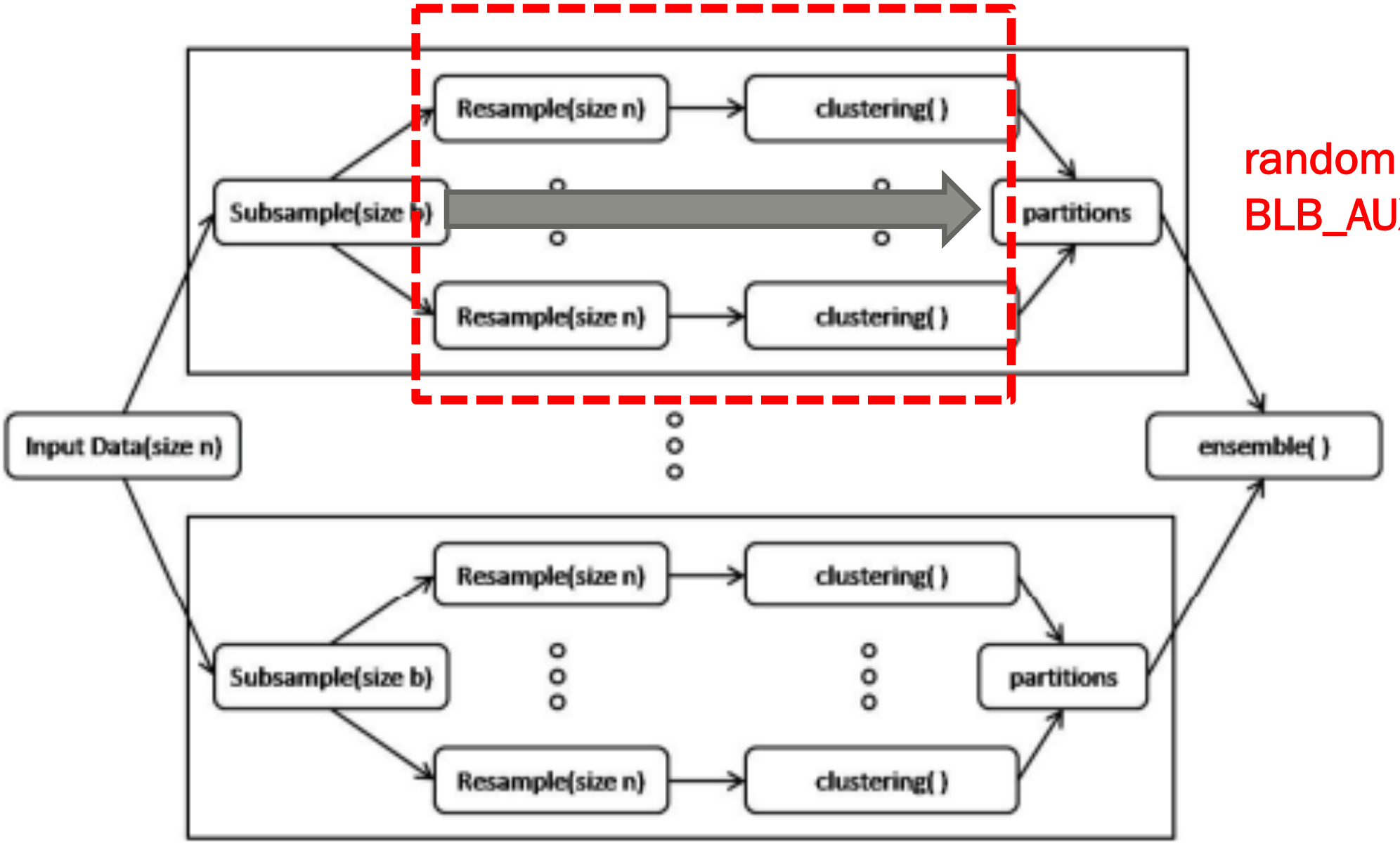
- Replace *Bootstrapping* with *BLB*!!!
- How can BLB-RF make a difference?
  - *Advantages:*
    - Each subsample only has b unique values!
    - Scales in **b,** O(b), instead of **n,** O(n)
    - Easily parallelizable to speed up execution

How to implement BLB-RF?

	randomForest	randomForestBLB_AUX	randomForestBLB
Input	N observations	B observations (subsample)	N observations
Resampling methods	Bootstrap	Multinomial (n, b, c(1/b, 1/b .... 1/b))	BLB
Output	Prediction using entire sample	Prediction using the subsample	Prediction using the entire sample w/ Parallelism

How to implement BLB-RF?

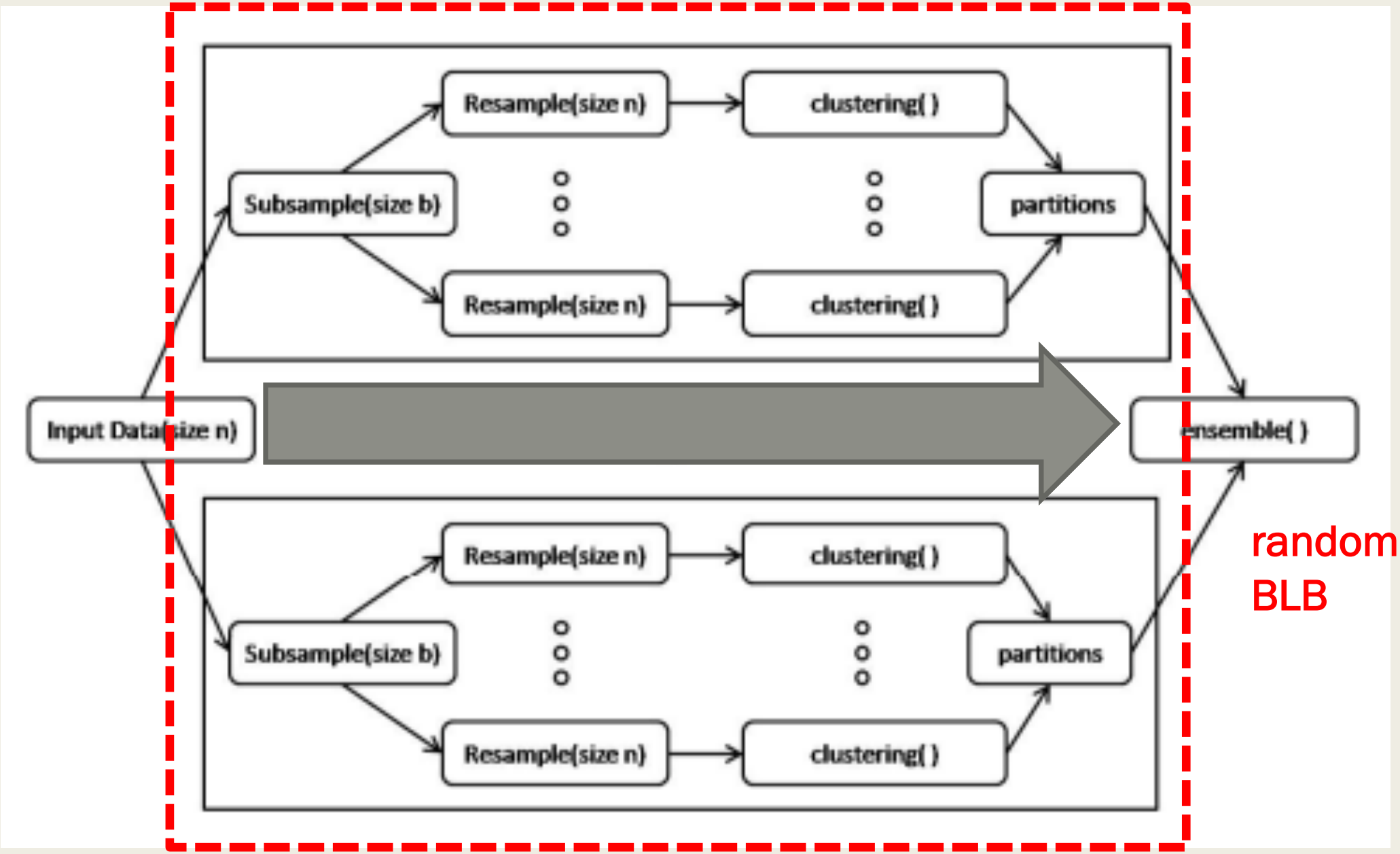
	randomForest	randomForestBLB_AUX	randomForestBLB
Input	N observations	B observations (subsample)	N observations
Resampling methods	Bootstrap	Multinomial (n, b, c(1/b, 1/b .... 1/b))	BLB
Output	Prediction using entire sample	Prediction using the subsample	Prediction using the entire sample w/ Parallelism



randomForest
BLB_AUX

How to implement BLB-RF?

	randomForest	randomForestBLB_AUX	randomForestBLB
Input	N observations	B observations (subsample)	N observations
Resampling methods	Bootstrap	Multinomial (n, b, c(1/b, 1/b .... 1/b))	BLB
Output	Prediction using entire sample	Prediction using the subsample	Prediction using the entire sample w/ Parallelism



randomForest
BLB

Thank you for listening and enjoy your PIZZA!