Placing Sequences Into Large Trees

RECOMB Satellites: Scientific Communications

Gillian Chu

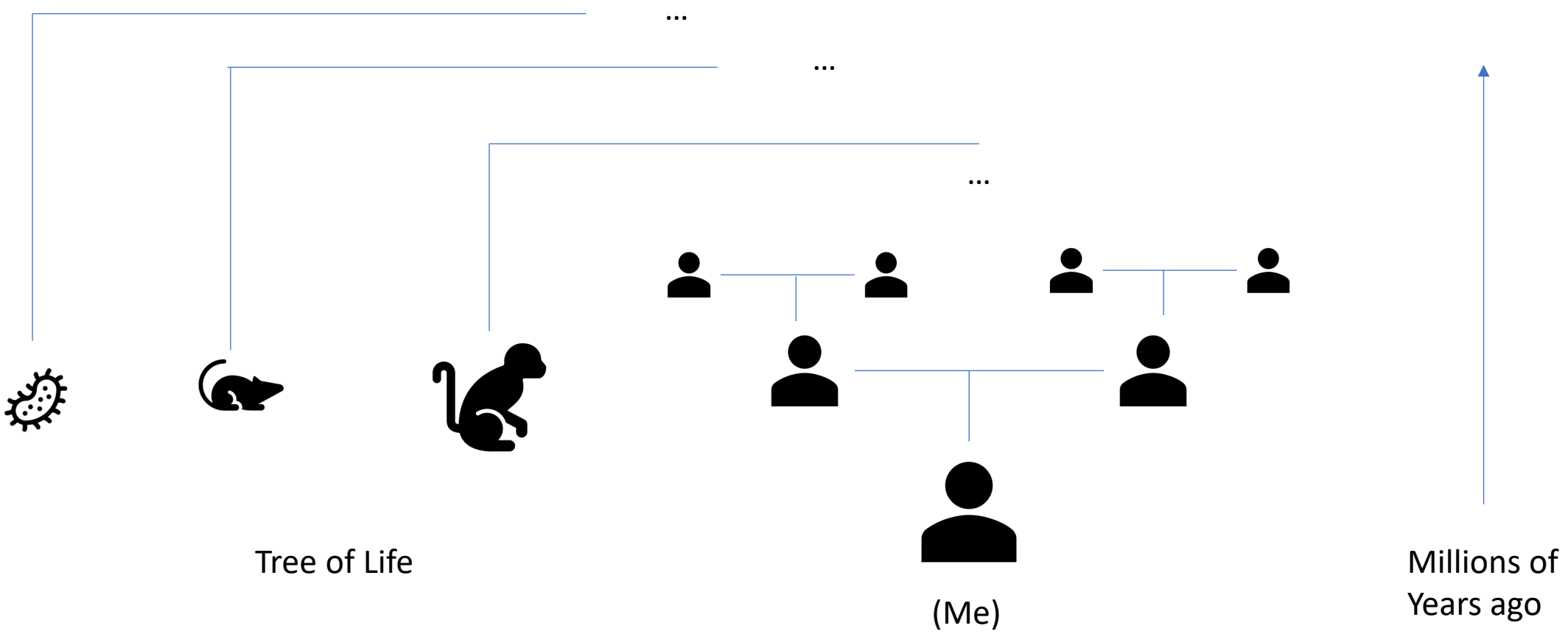
05-21-22

Evolution

“Nothing in biology makes sense except in the light of evolution”

– Theodosius Dobzhansky, 1973 essay in the American Biology Teacher, vol. 35, pp 125-129

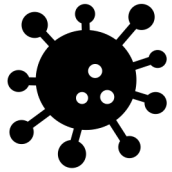
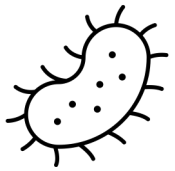
Evolutionary Relationships Summarized in a Tree



Problem: Relationships Between Organisms

Go to a pond, scoop out a bucket of water and then try to identify what new organisms are in your bucket of water.

New organism (sequence of letters)


Which is it closest to?
How related is it to everything we've seen so far?
We call this "placing" into the tree of life.

Problems in Placing

When the tree gets large, placing new organisms into our tree becomes difficult.

Two methods:

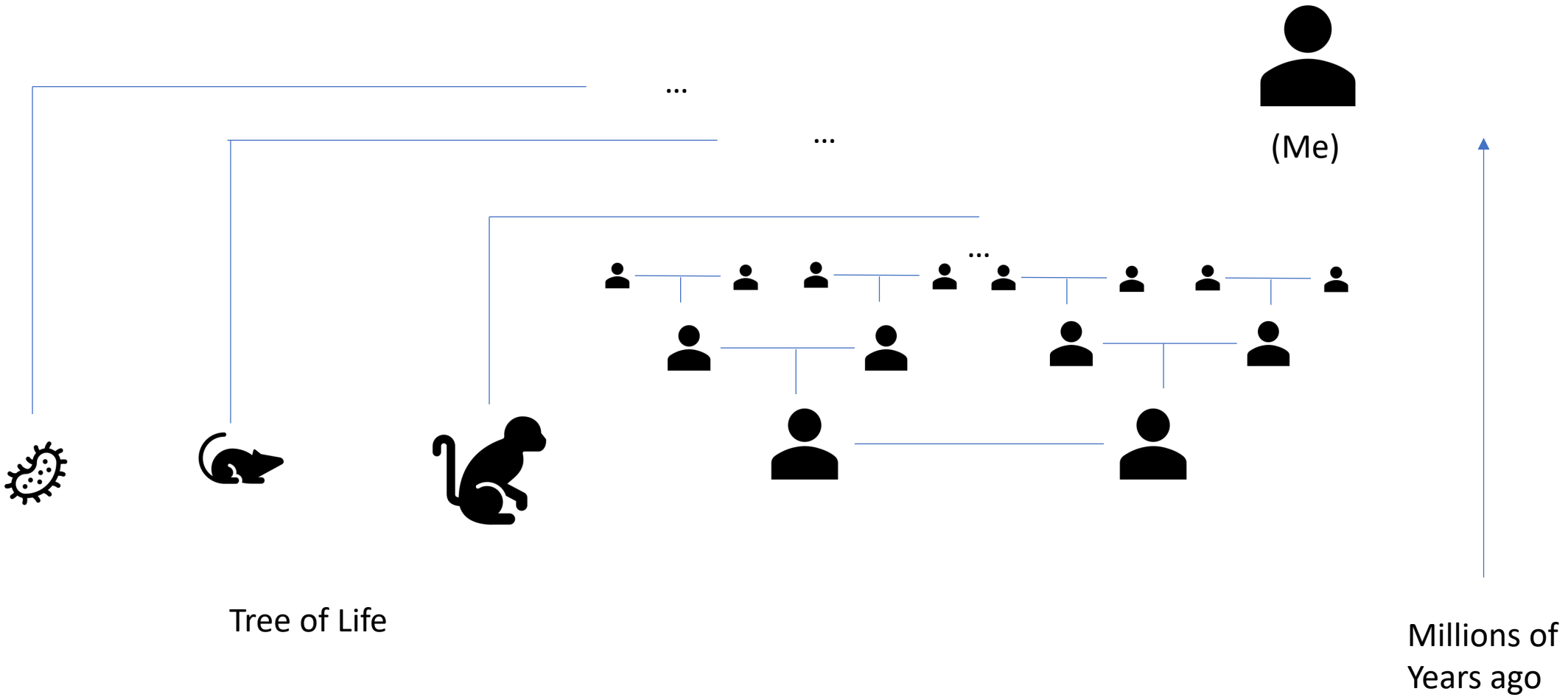
Method A (most accurate)

- Fails on trees with > 5,000 due to some estimation problems

Method B (fast)

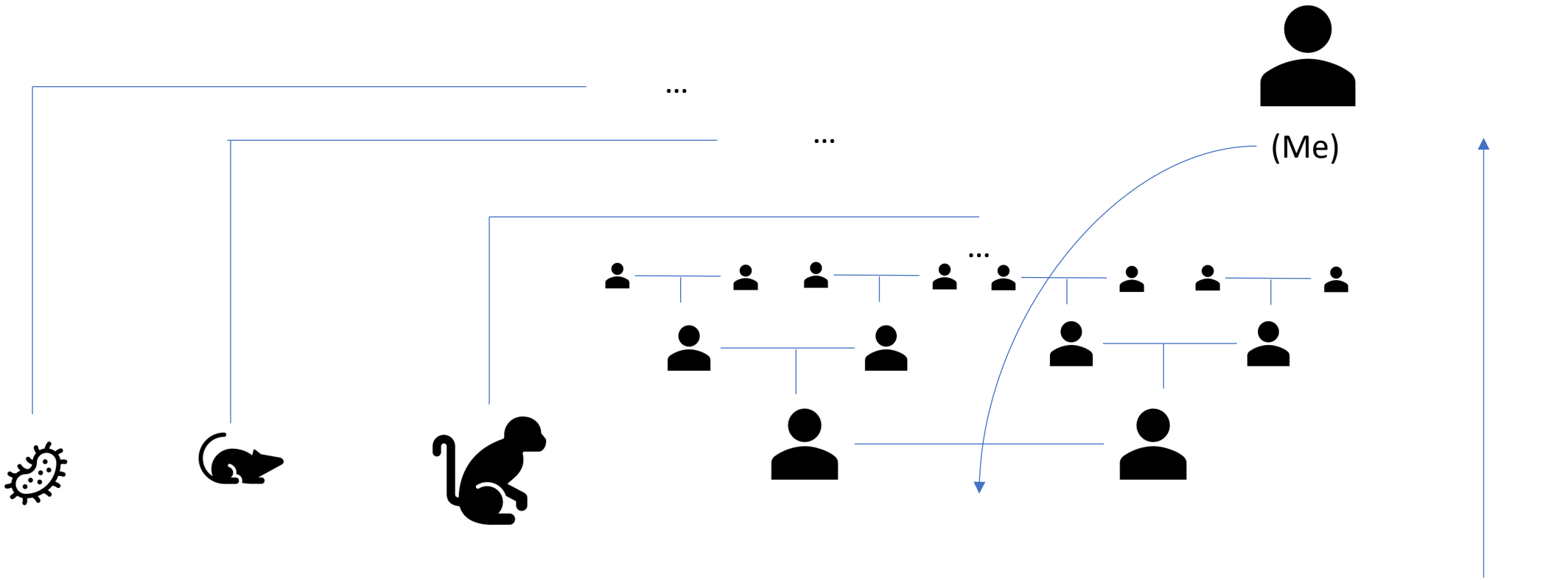
- Does not have as high accuracy as Method A
- Achieves speed by finding a smaller tree to place into
- If we place into a smaller tree:
  - Faster
  - Lower accuracy

Speed Up Placements with Shortcuts



Instead of placing me into a tree describing all organisms, we can use shortcuts:
1. I'm a human, 2. we can use my last name. Use this to narrow down which family tree, then place me.

Speed Up Placements with Shortcuts



Instead of placing me into a tree describing all organisms, we can use shortcuts:
1. I'm a human, 2. we can use my last name. Use this to narrow down which family tree, then place me.

Problems in Placing

When the tree gets really large placing new genetic information into our tree becomes difficult.

Method A (most accurate)

- Fails on trees with > 5,000 due to some estimation problems

Method A' (most accurate)

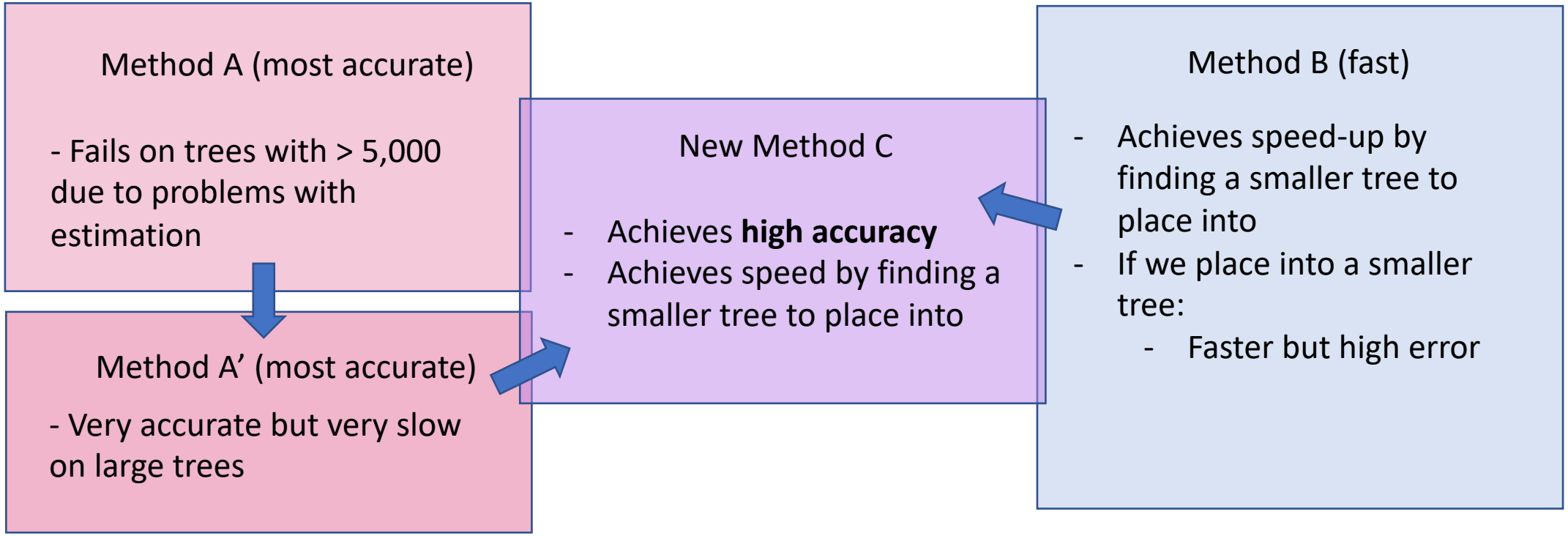
- Runs on trees with > 5,000 but is very slow

Method B (fast)

- Achieves speed-up by finding a smaller tree to place into
- If we place into a smaller tree:
  - Faster but high error

New Method Combining Accuracy and Scalability

When the tree gets really large placing new genetic information into our tree becomes difficult.



Method A’ (most accurate)

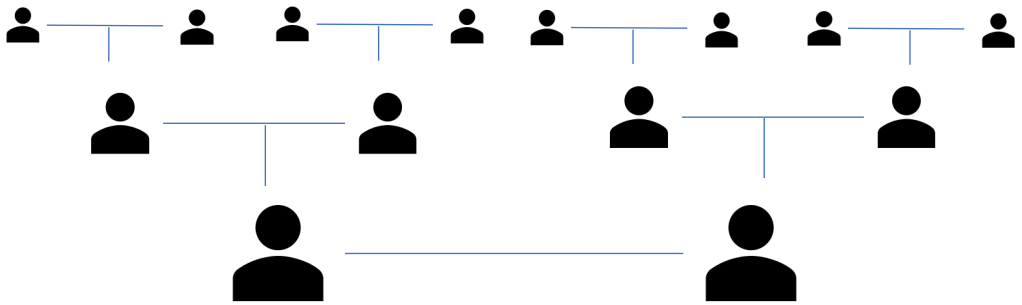
- Runs with **low error** on trees with > 5,000 but is **very slow**

New Method C

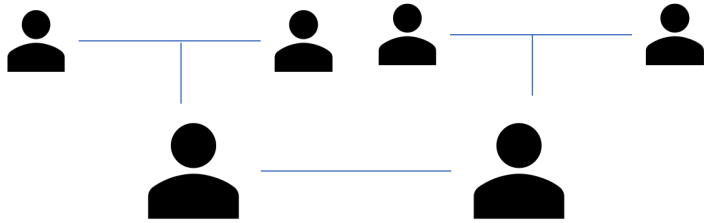
- **Low error and very fast**
  - Find a smaller tree to place into

Method B (fast)

- **Faster but high error**
  - Find a smaller tree to place into



Large Trees
- Higher accuracy
- Takes longer



Smaller Trees
- Lower accuracy
- Faster

Method A’ (most accurate)

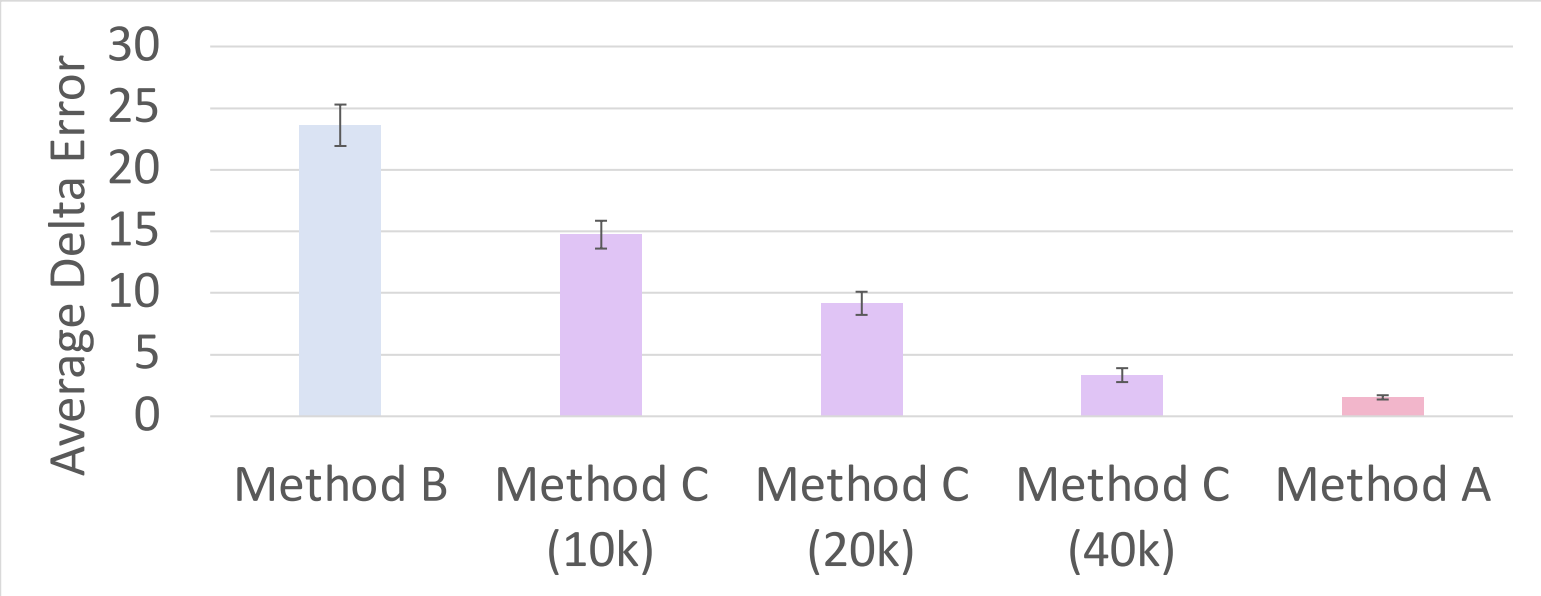
- Runs with **low error** on trees with > 5,000 but is **very slow**

New Method C

- **Low error and very fast**
  - Find a smaller tree to place into

Method B (fast)

- **Faster but high error**
  - Find a smaller tree to place into



Conclusions

We are able to quickly and accurately place new organisms into the tree of life!

This helps us:

- Identify new organisms
- Understand different organism communities (e.g. pond)
- Build and update large trees – like the tree of life