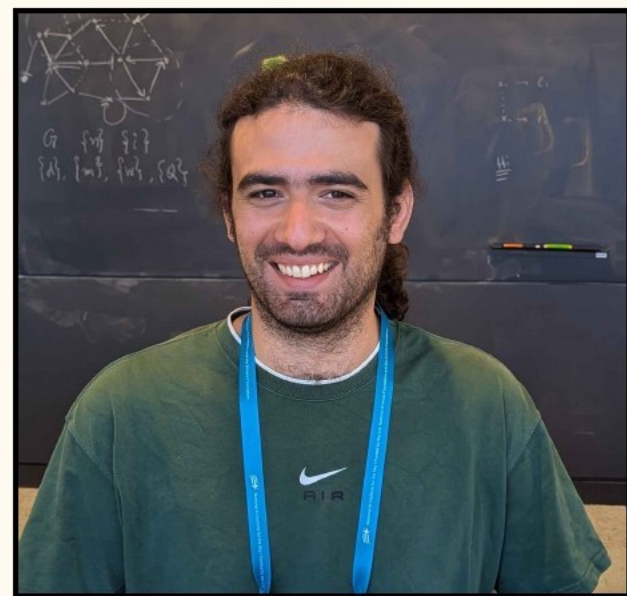


Optimal Phylogenetic Reconstruction from Sampled Quartets

STOC 2026

Dionysis Arvanitakis

Northwestern University



Vaggos Chatziafratis

UC Santa Cruz



Yiyuan Luo

UC Santa Cruz



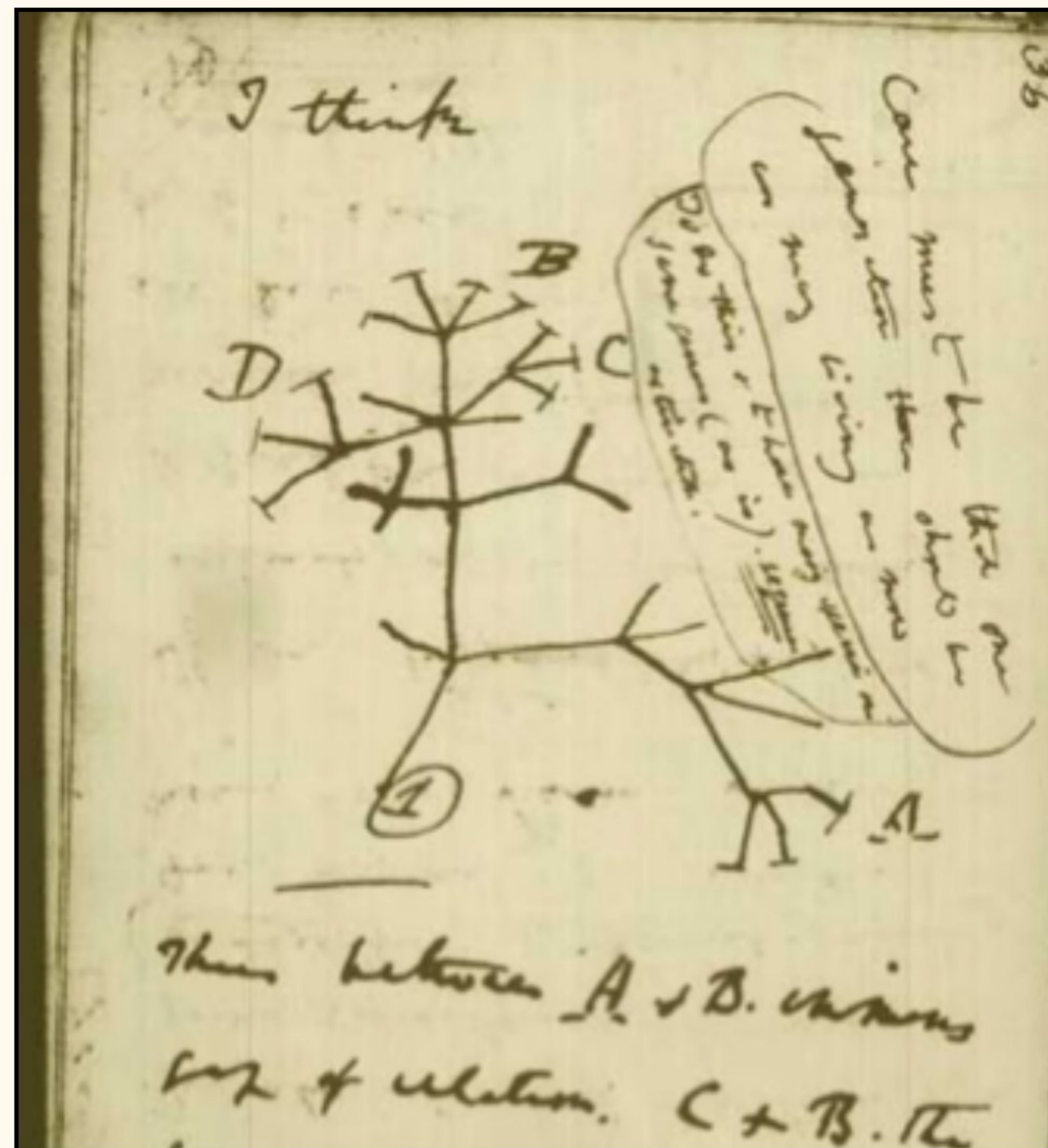
Konstantin Makarychev

Northwestern University

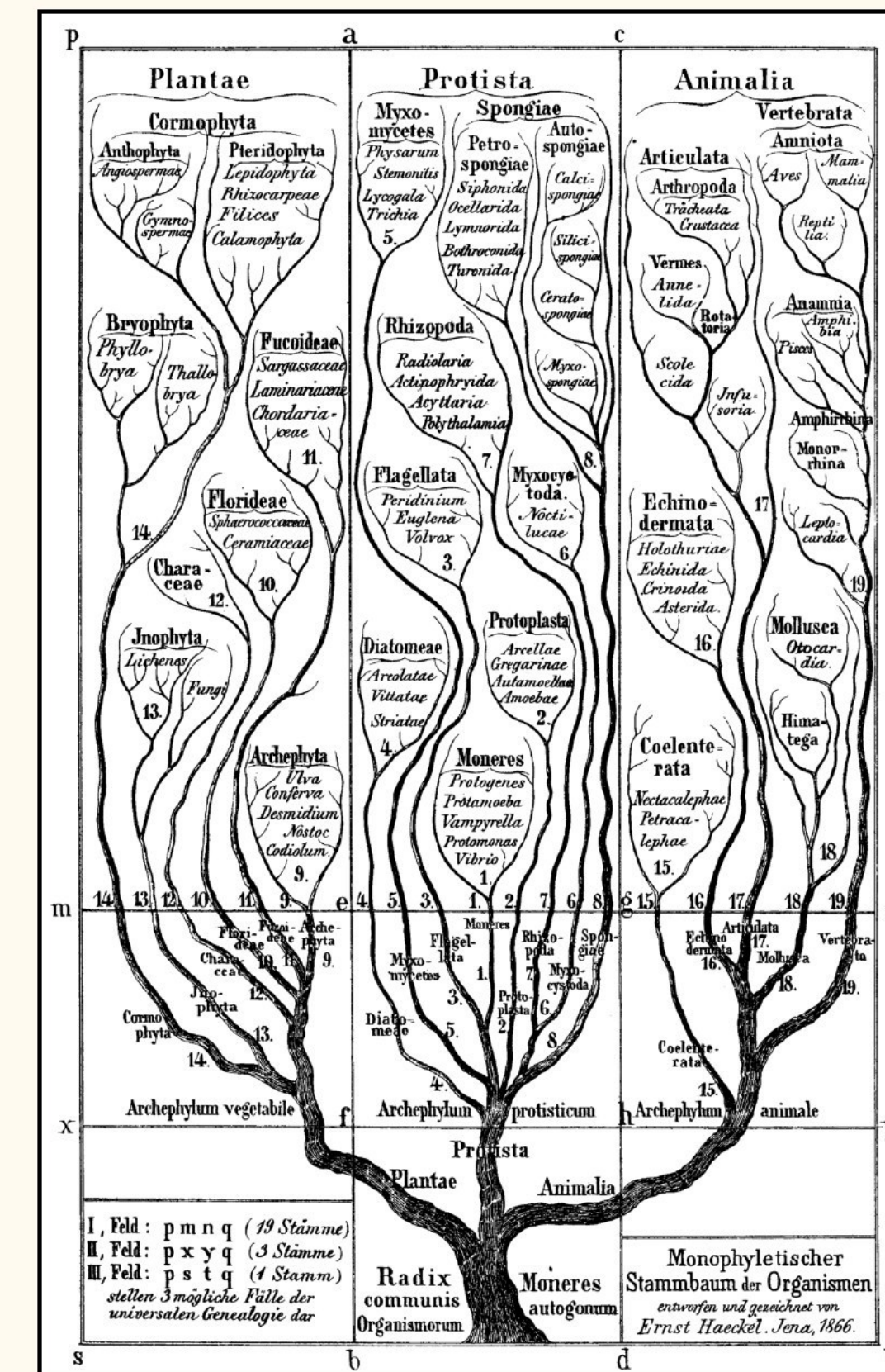


Phylogenetic Reconstruction

Starting in the 19th century:

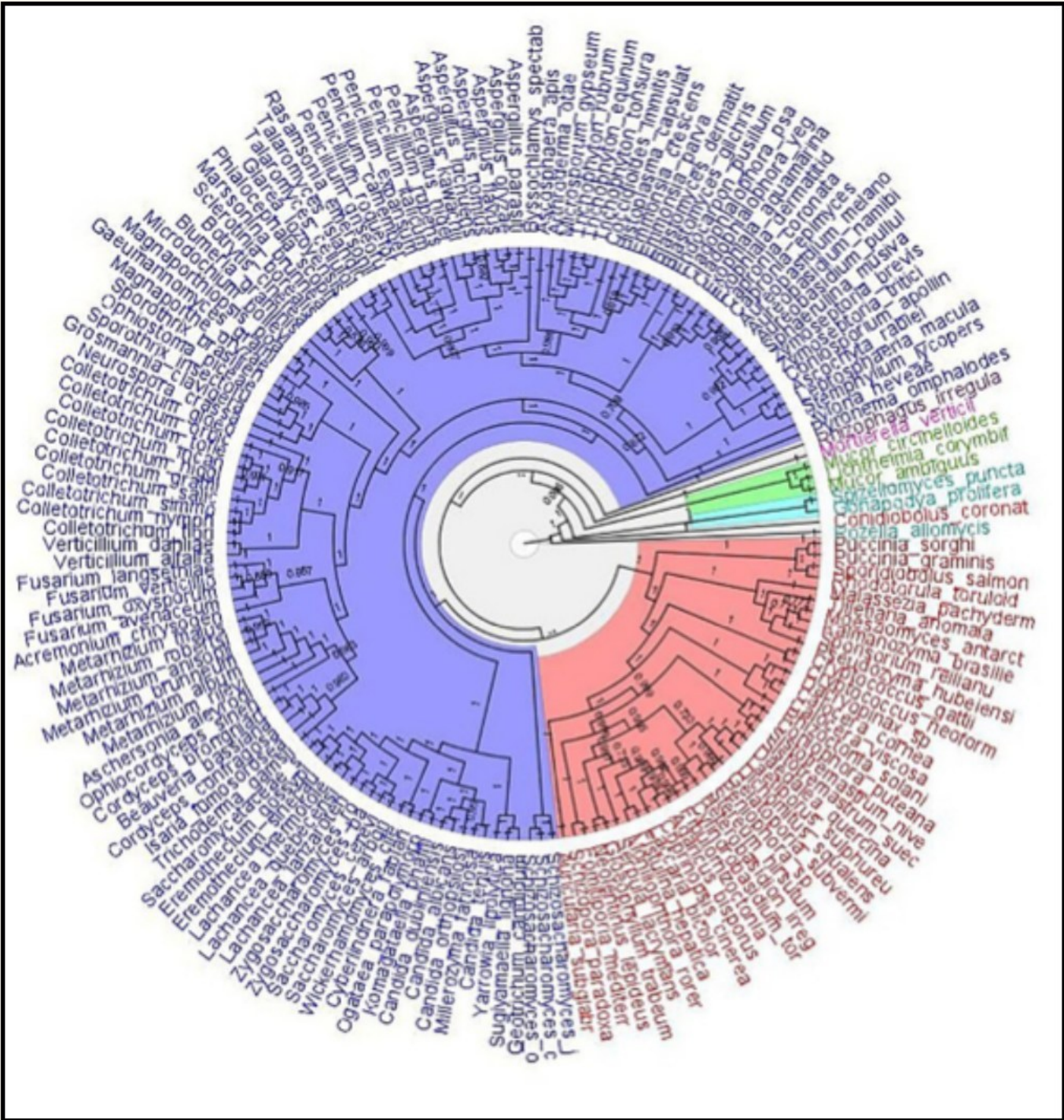


A page from Darwin's notebook (1837)



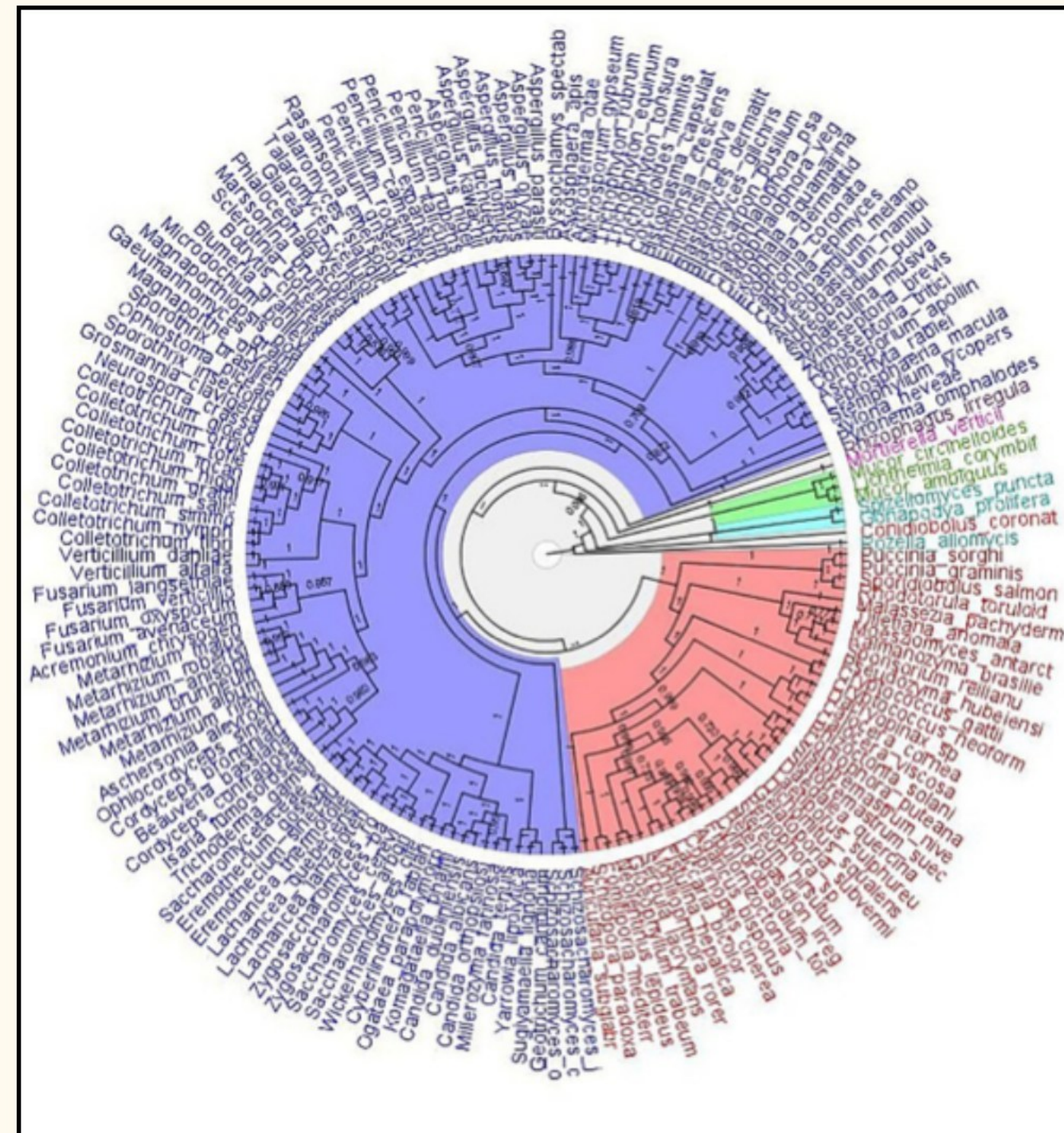
Haeckel's tree of life (1866)

Phylogenetic Reconstruction



Phylogenetic tree on fungi

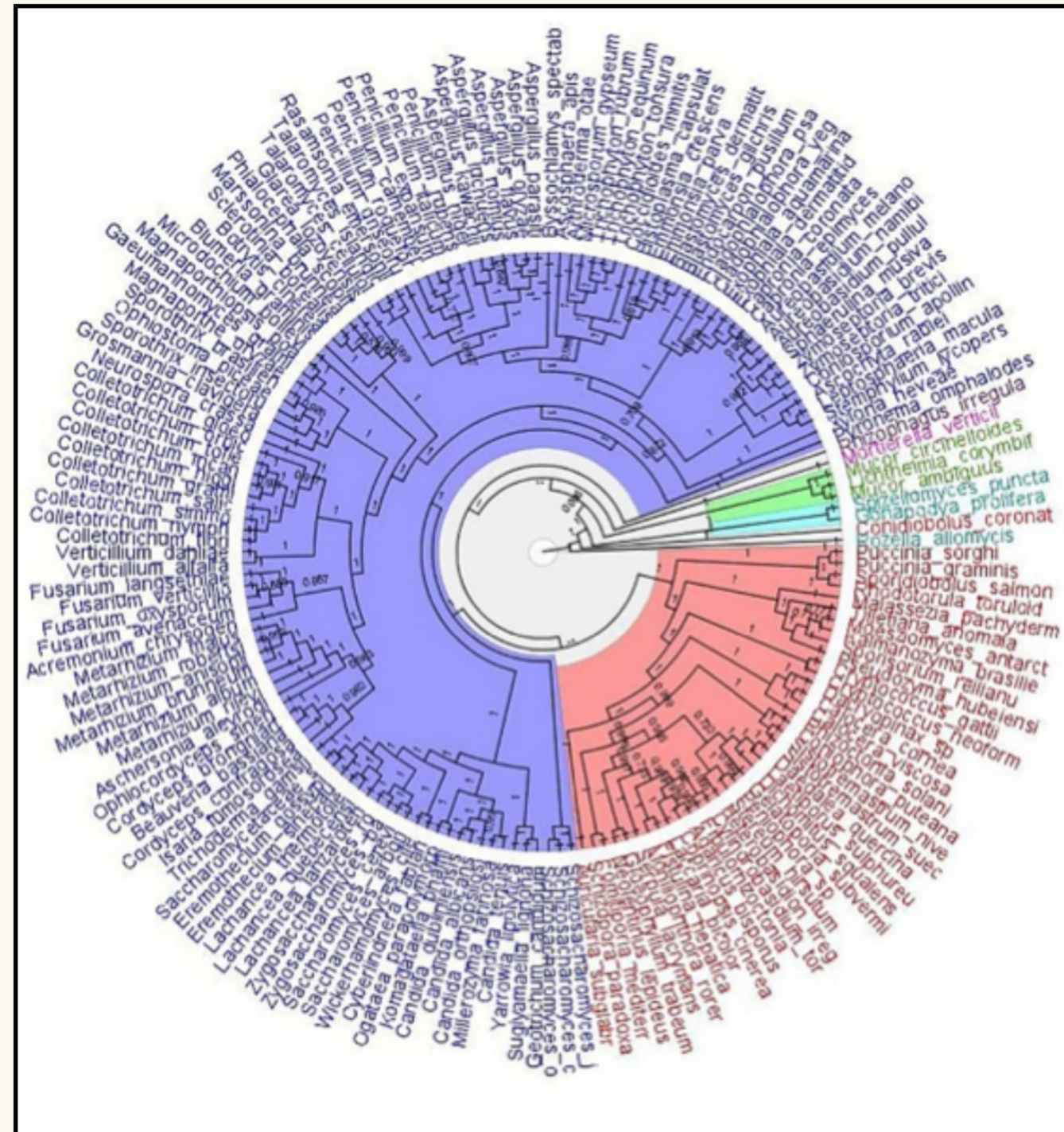
Phylogenetic Reconstruction



Phylogenetic tree on fungi

Q1: How to reconstruct the tree?

Phylogenetic Reconstruction



Phylogenetic tree on fungi

Q1: How to reconstruct the tree?

Q2: How much information is needed?

Phylogenetic Reconstruction

Set of n species {  ,  ,  ,  , ...,  ,  }

Phylogenetic Reconstruction

Set of n species {  ,  ,  ,  , ...,  ,  }

Goal: Recover the phylogenetic tree on these species

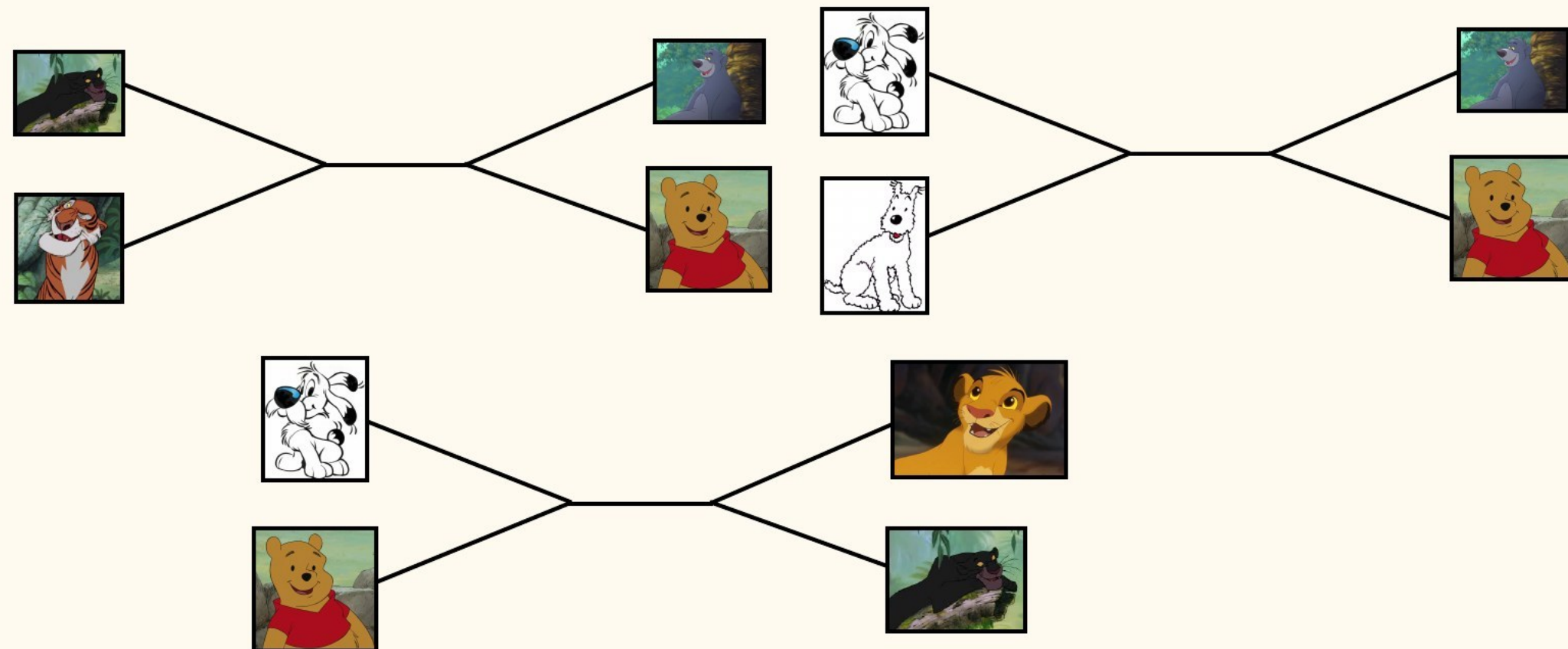
Phylogenetic Reconstruction

Set of n species {  ,  ,  ,  , ...,  ,  }

Goal: Recover the phylogenetic tree on these species

Approach (subtree amalgamation)

- Recover the relationship between smaller subsets of species



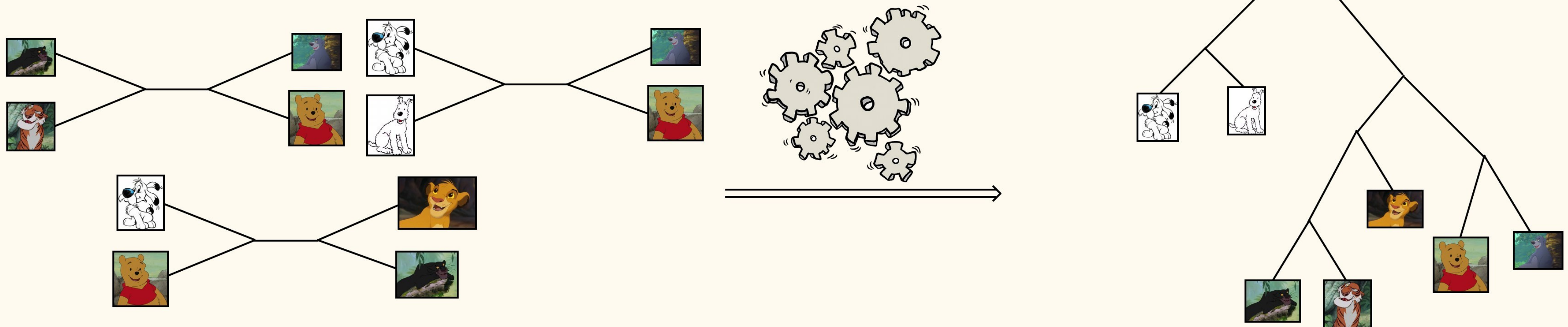
Phylogenetic Reconstruction

Set of n species {  ,  ,  ,  , ...,  ,  }

Goal: Recover the phylogenetic tree on these species

Approach (subtree amalgamation)

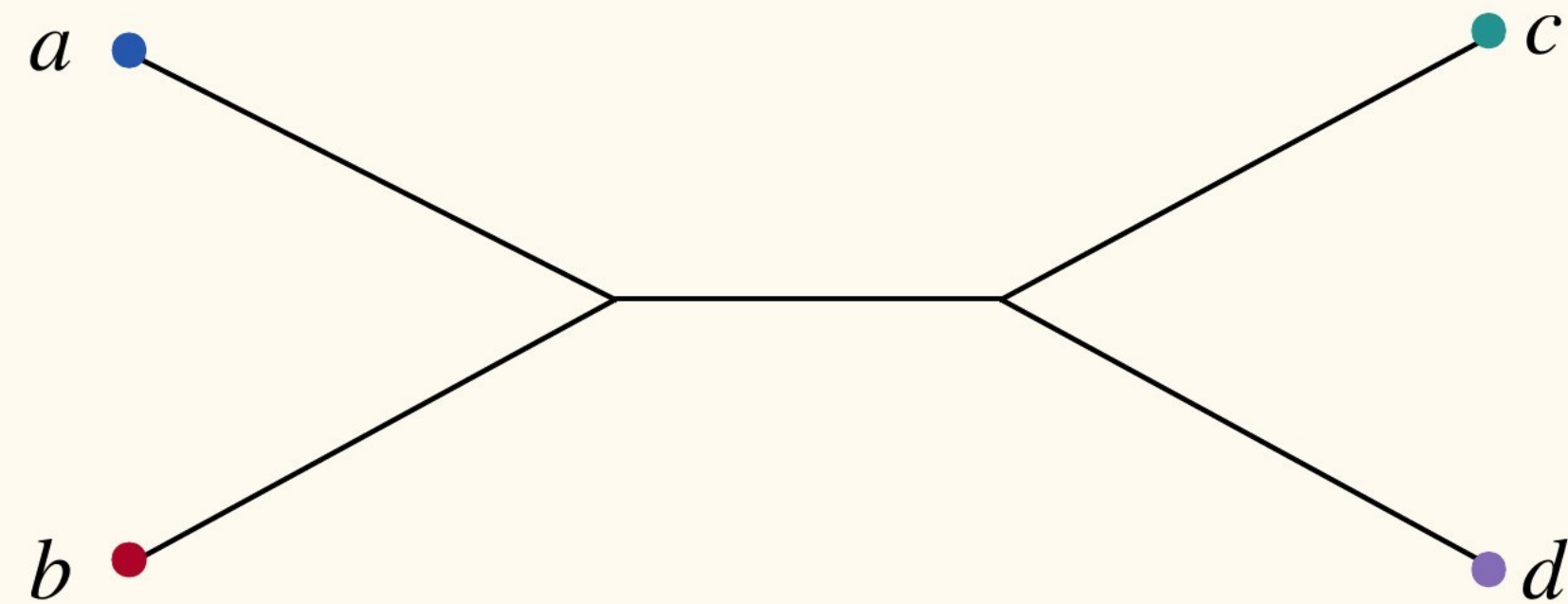
- Recover the relationship between smaller subsets of species
- Combine this information to recover the entire tree



Quartet trees

The smallest piece of phylogenetic information

- (Unrooted) phylogeny on 4 species



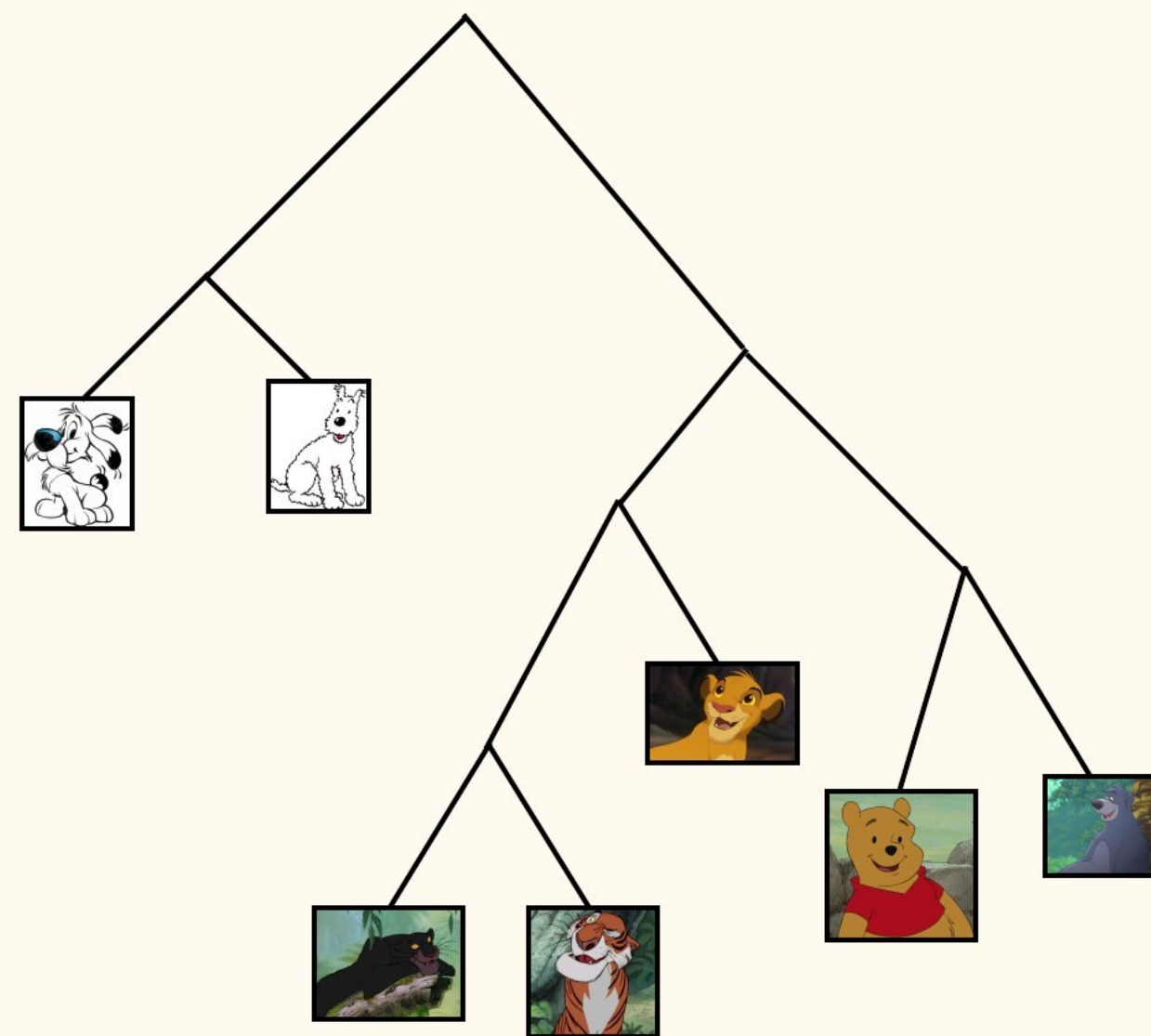
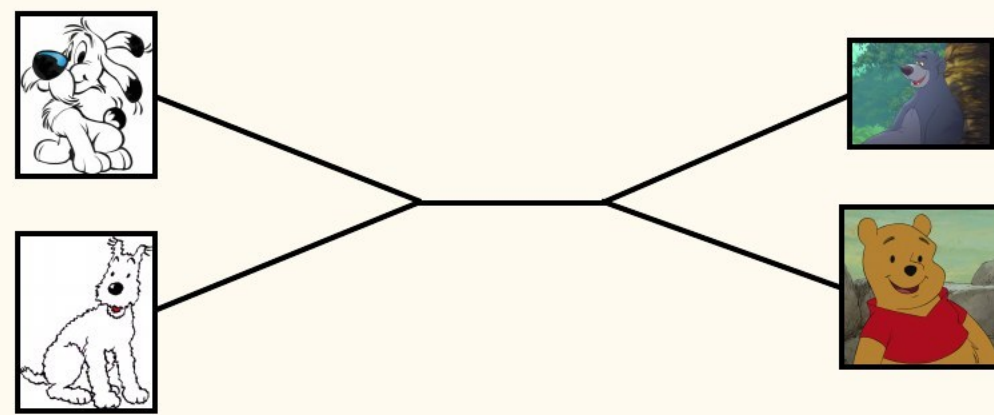
$ab|cd$: In the phylogenetic tree there exists an edge **e** such that if removed:

- *a*, *b* are in **one connected component**
- *c*, *d* in the **other connected component**

Quartet trees

ab|cd: In the phylogenetic tree there exists an edge **e** such that if removed:

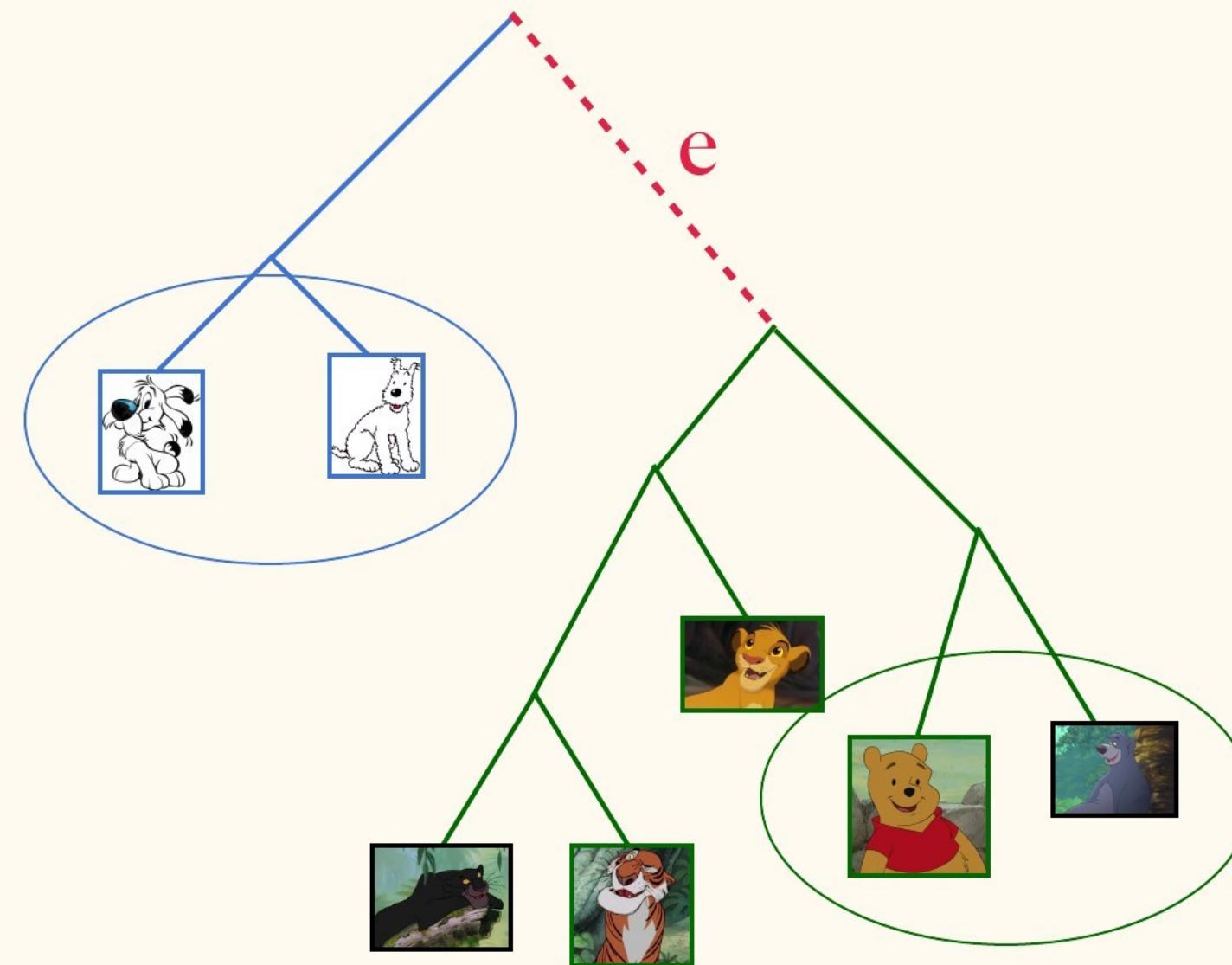
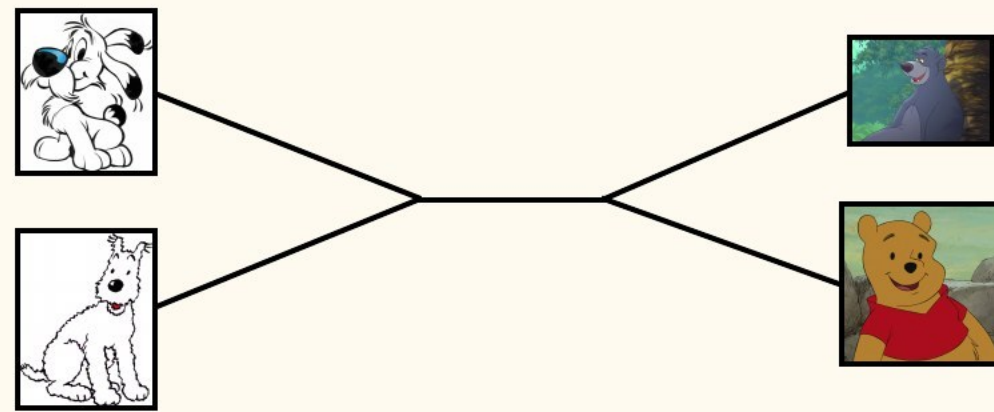
- a, b are in **one connected component**
- c, d in the **other connected component**



Quartet trees

ab|cd: In the phylogenetic tree there exists an edge **e** such that if removed:

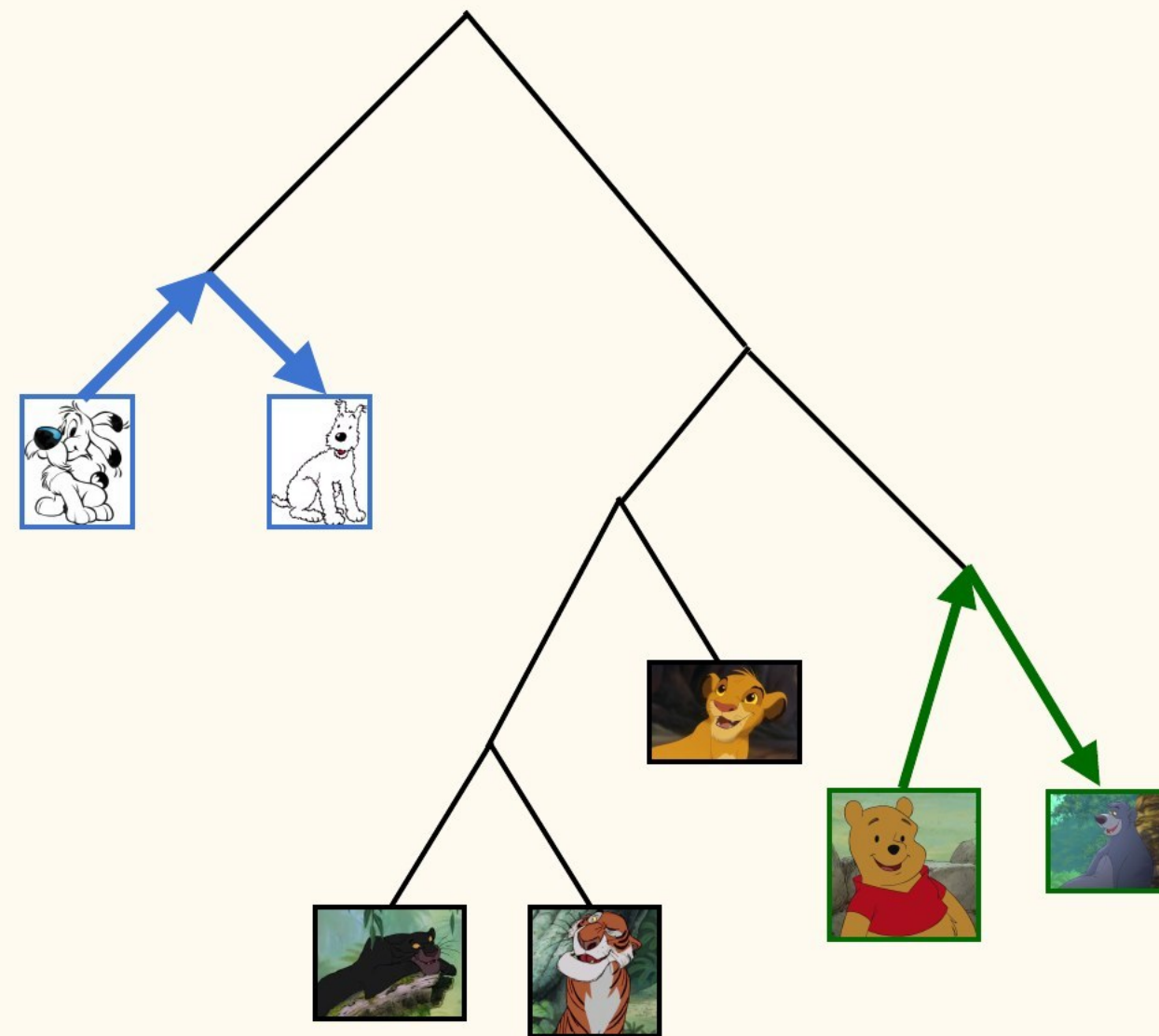
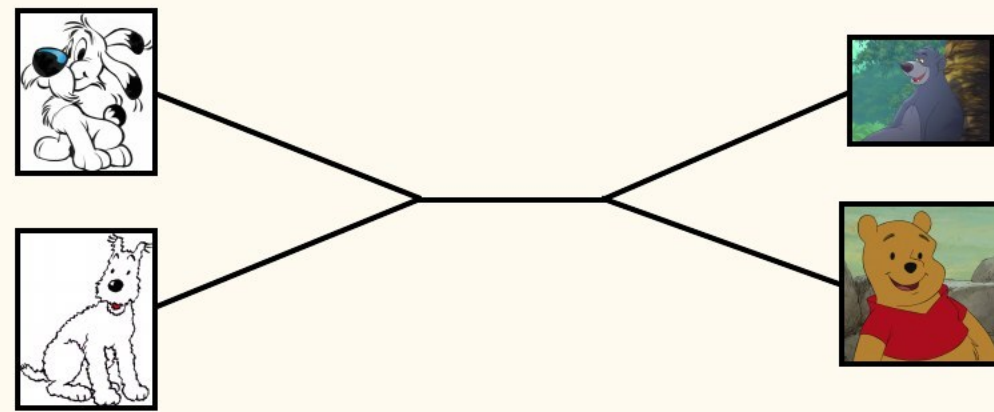
- a, b are in **one connected component**
- c, d in the **other connected component**



Quartet trees

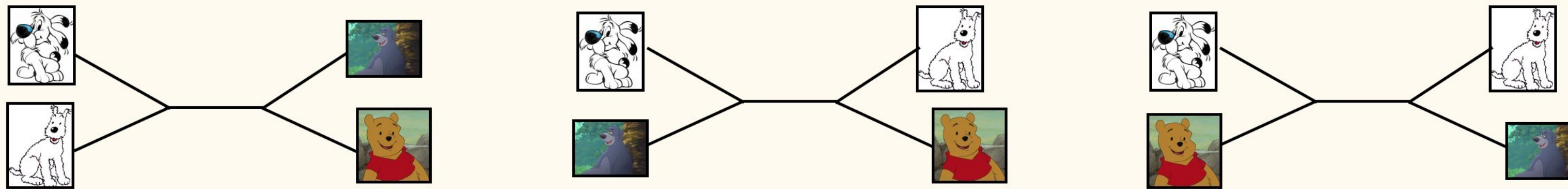
Equivalently:

$ab|cd$: In the phylogenetic tree the path between a and b is disjoint from the path between c and d



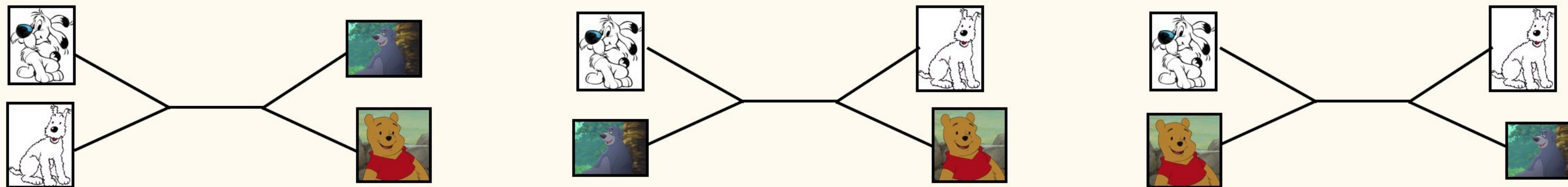
Quartet trees

- For a set $\{a, b, c, d\}$: 3 different possible quartet trees



Quartet trees

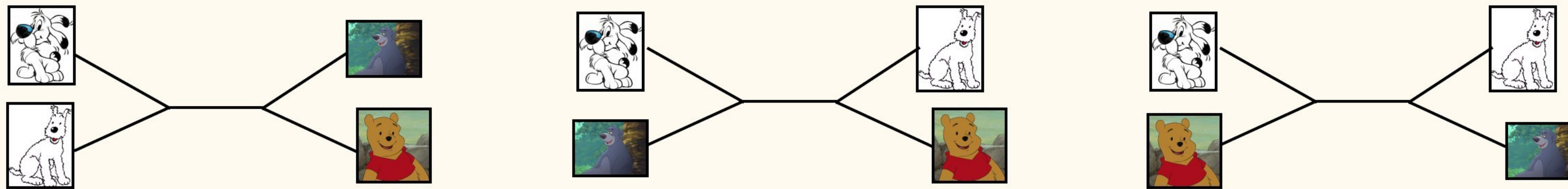
- For a set $\{a, b, c, d\}$: 3 different possible quartet trees



- For a tree T and set $\{a, b, c, d\}$, exactly 1 out of 3 agrees with T : $T(a, b, c, d)$

Quartet trees

- For a set $\{a, b, c, d\}$: 3 different possible quartet trees



- For a tree T and set $\{a, b, c, d\}$, exactly 1 out of 3 agrees with T : $T(a, b, c, d)$
- The set of all $\binom{n}{4}$ quartet trees determines the phylogeny

Maximum Quartet Consistency (MQC)

Definition (Maximum Quartet Consistency)

Given a set $\mathcal{Q}$ of quartet constraints $q_i = a_i b_i | c_i d_i$ find the tree T that **maximizes** the satisfied constraints

Maximum Quartet Consistency (MQC)

Definition (Maximum Quartet Consistency)

Given a set $\mathcal{Q}$ of quartet constraints $q_i = a_i b_i | c_i d_i$ find the tree T that **maximizes** the satisfied constraints

- NP-hard even when the instance is satisfiable [*Steel Journal of classification 1992*]

Maximum Quartet Consistency (MQC)

Definition (Maximum Quartet Consistency)

Given a set $\mathcal{Q}$ of quartet constraints $q_i = a_i b_i | c_i d_i$ find the tree T that **maximizes** the satisfied constraints

- NP-hard even when the instance is satisfiable [*Steel Journal of classification 1992*]
- Approximation resistant under UGC [*Chatziafratis, Makarychev FOCS 2023*]

Maximum Quartet Consistency (MQC)

Definition (Maximum Quartet Consistency)

Given a set $\mathcal{Q}$ of quartet constraints $q_i = a_i b_i | c_i d_i$ find the tree T that **maximizes** the satisfied constraints

- NP-hard even when the instance is satisfiable [*Steel Journal of classification 1992*]
- Approximation resistant under UGC [*Chatziafratis, Makarychev FOCS 2023*]
- PTAS for dense instances ($|\mathcal{Q}| = \Theta(n^4)$) and planted instances with $\Theta_\epsilon(n^2 \log(n))$ quartets
[*Jiang, Kearney, Li. FOCS 1998*], [*Snir, Yuster. Journal on Discrete Mathematics 2011*]

Maximum Quartet Consistency (MQC)

Definition (Maximum Quartet Consistency)

Given a set $\mathcal{Q}$ of quartet constraints $q_i = a_i b_i | c_i d_i$ find the tree T that **maximizes** the satisfied constraints

- NP-hard even when the instance is satisfiable [Steel *Journal of classification* 1992]
- Approximation resistant under UGC [Chatziafratis, Makarychev *FOCS* 2023]
- PTAS for dense instances ($|\mathcal{Q}| = \Theta(n^4)$) and planted instances with $\Theta_\epsilon(n^2 \log(n))$ quartets [Jiang, Kearney, Li. *FOCS* 1998], [Snir, Yuster. *Journal on Discrete Mathematics* 2011]
- Practical algorithms and local to global compatibility phenomena [Snir, Rao *Molecular phylogenetics and evolution* 2012], [Alon, Snir, Yuster *SODA* 2014]

Maximum Quartet Consistency (MQC)

Maximum Quartet Consistency is just a proxy, the goal is to **recover** the phylogeny

Maximum Quartet Consistency (MQC)

Maximum Quartet Consistency is just a proxy, the goal is to **recover** the phylogeny

In this talk:

- (-) Stronger assumptions: Impose probabilistic model on the input quartets

Maximum Quartet Consistency (MQC)

Maximum Quartet Consistency is just a proxy, the goal is to **recover** the phylogeny

In this talk:

(-) Stronger assumptions: Impose probabilistic model on the input quartets

(+) More ambitious goal: Approximately **recover** the phylogeny

Planted model with random noise

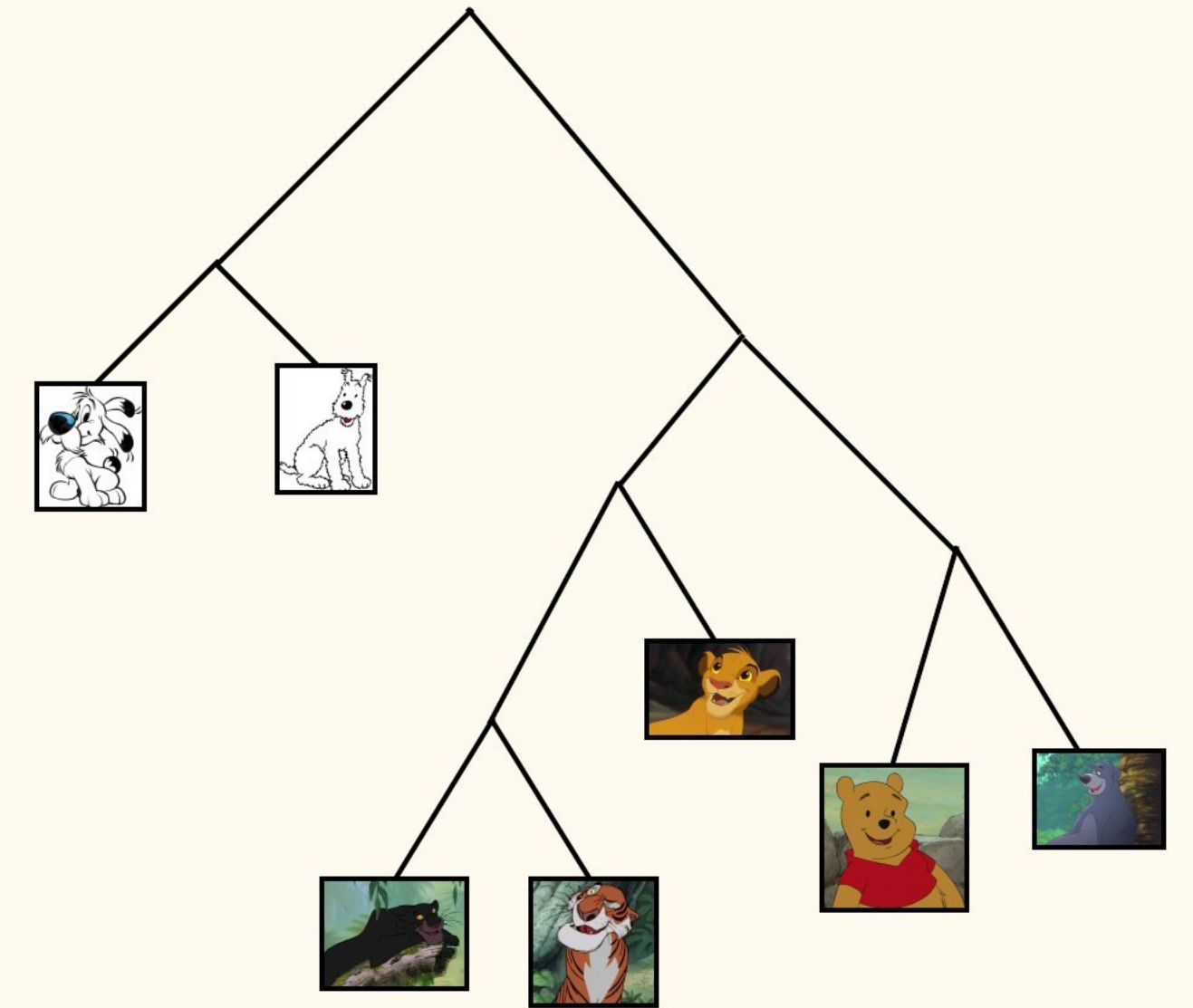
We assume that there is a ground truth tree T

Access to quartets $\mathcal{Q}$ from the following procedure RCN(η):

- Sample 4 distinct species $\{a, b, c, d\}$, uniformly at random

- W.p. η , get the correct quartet $T(a, b, c, d)$

- W.p. $1 - \eta$, get a random incorrect quartet

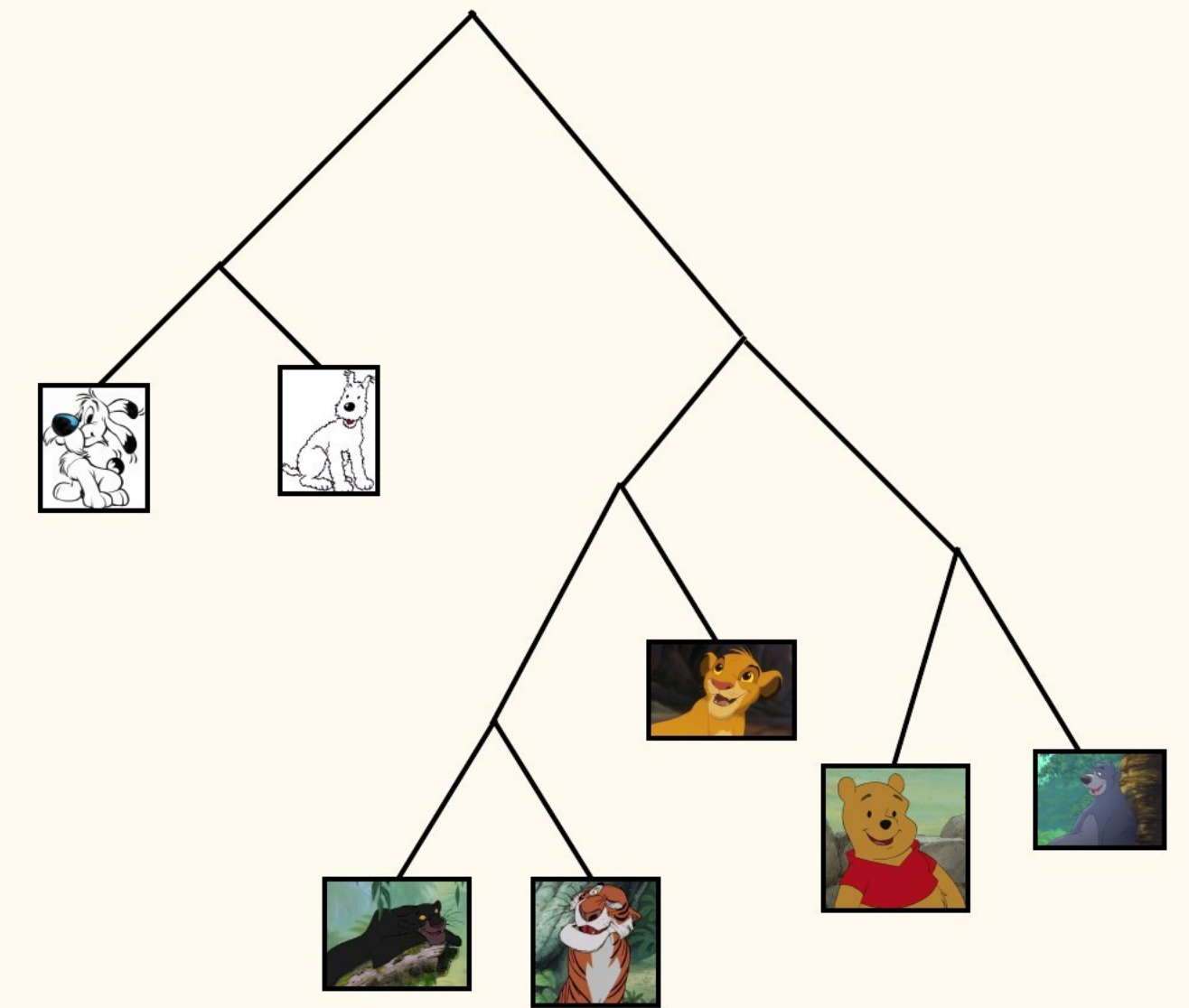


Planted model with random noise

We assume that there is a ground truth tree T

Access to quartets $\mathcal{Q}$ from the following procedure RCN(η):

-Sample 4 distinct species $\{a, b, c, d\}$, uniformly at random



Planted model with random noise

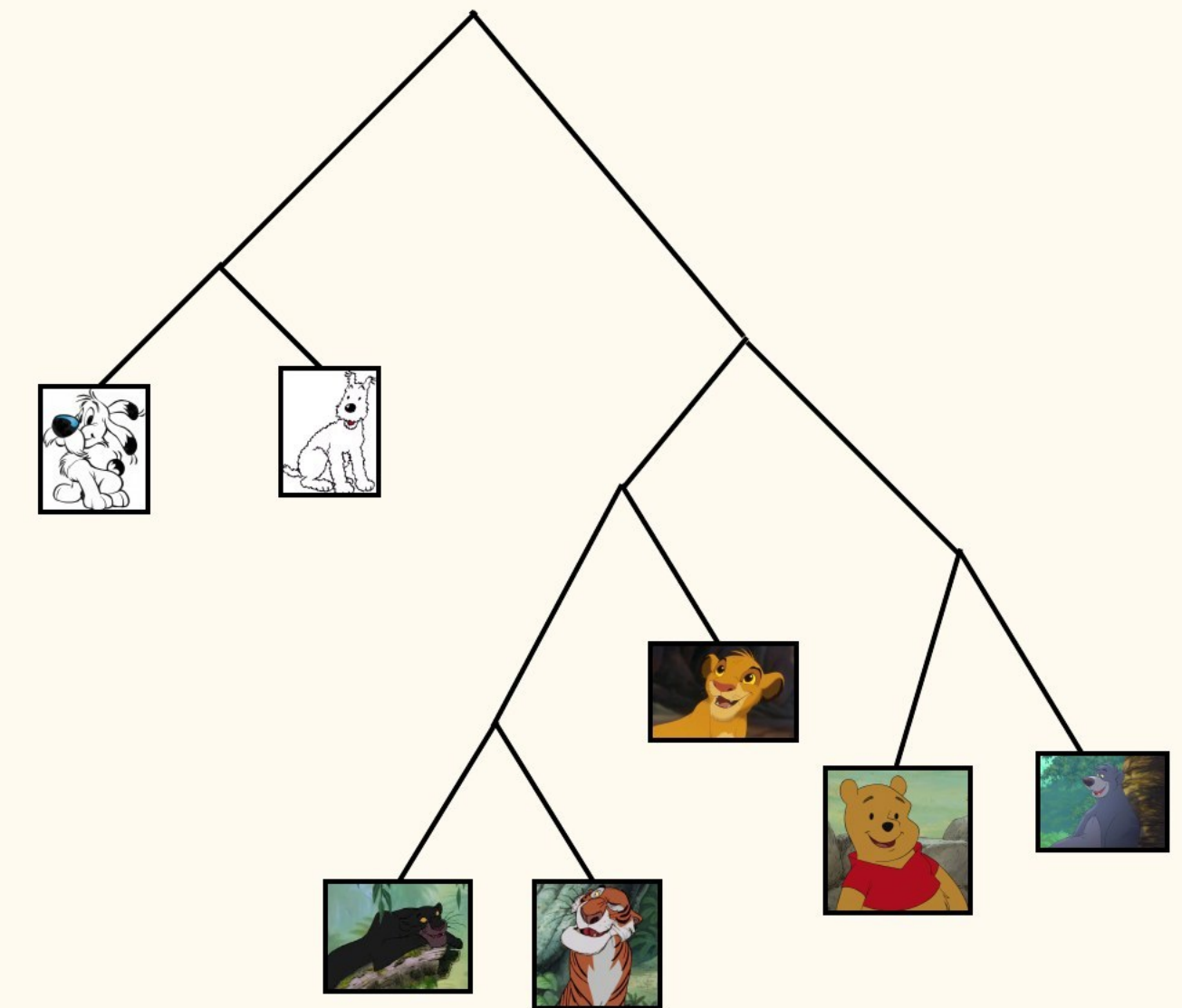
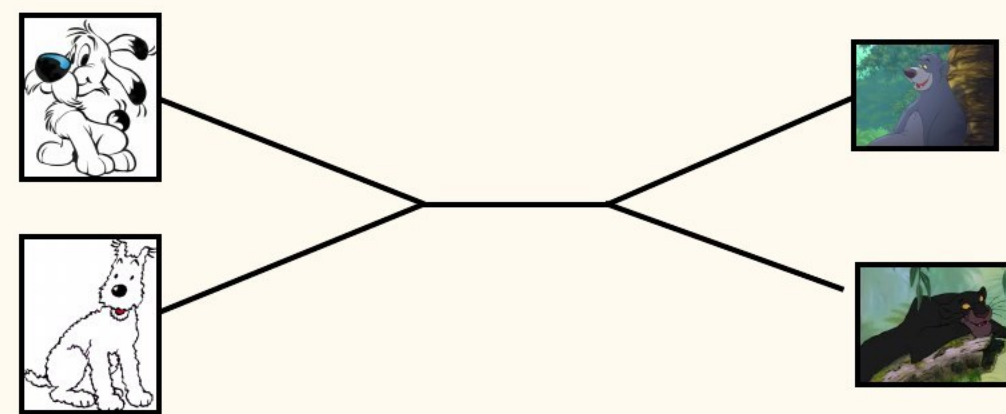
We assume that there is a ground truth tree T

Access to quartets $\mathcal{Q}$ from the following procedure RCN(η):

-Sample 4 distinct species $\{a, b, c, d\}$, uniformly at random



-W.p. η , get the correct quartet $T(a, b, c, d)$



Planted model with random noise

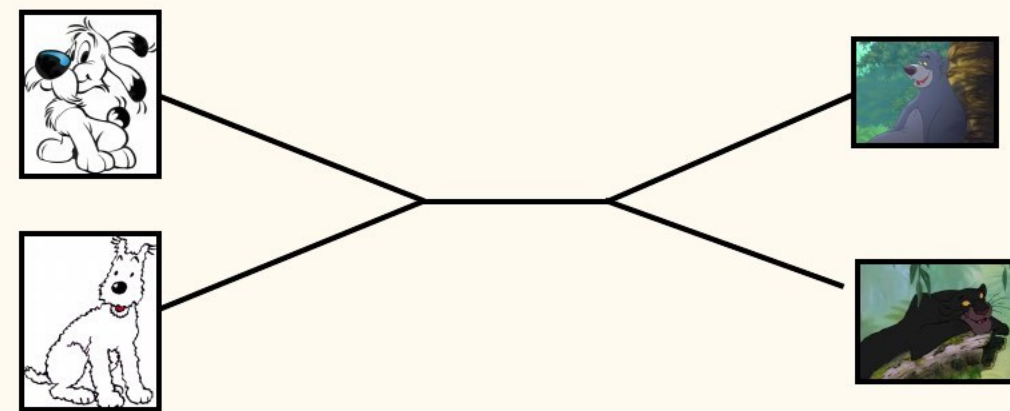
We assume that there is a ground truth tree T

Access to quartets $\mathcal{Q}$ from the following procedure RCN(η):

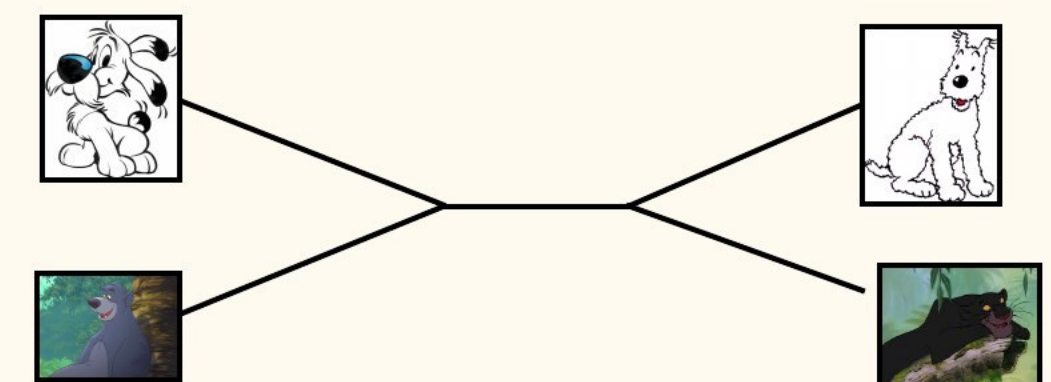
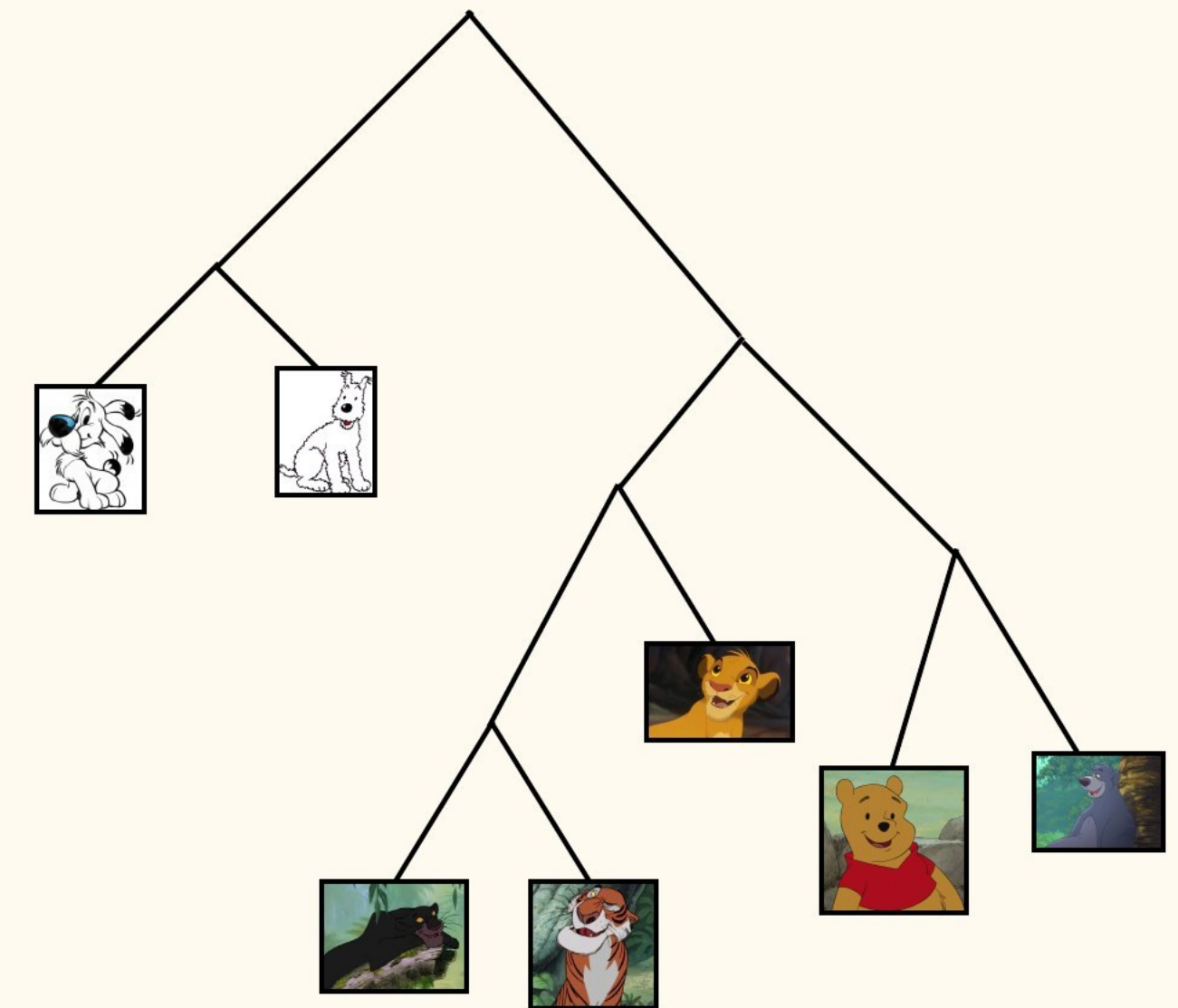
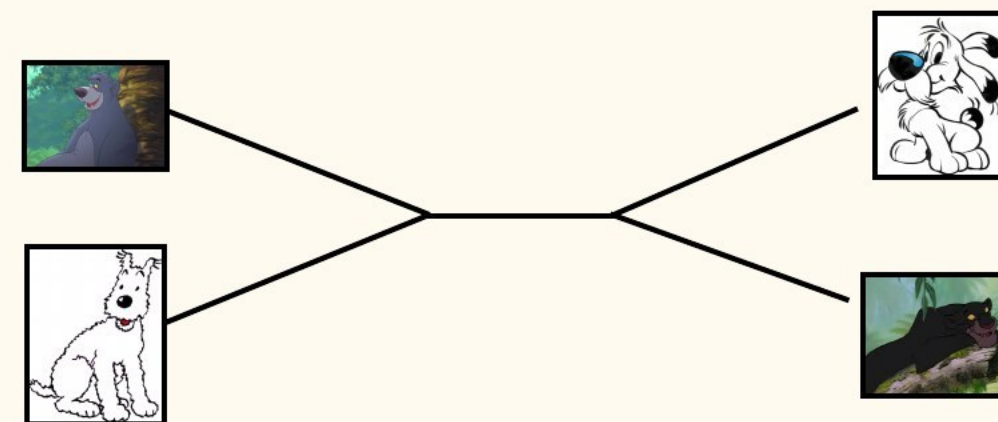
-Sample 4 distinct species $\{a, b, c, d\}$, uniformly at random



-W.p. η , get the correct quartet $T(a, b, c, d)$



-W.p. $1 - \eta$, get a random incorrect quartet



Our Results

We provide a polynomial time algorithm that **recovers** a tree **close** to T using **linearly many samples**.

Our Results

We provide a polynomial time algorithm that **recovers** a tree **close** to T using **linearly many samples**.

Q: How to measure distance between trees?

Quartet Distance

Definition (quartet distance)

The quartet distance between two phylogenetic trees T and T' is defined as the fraction of quartets on which the two trees disagree, i.e.,

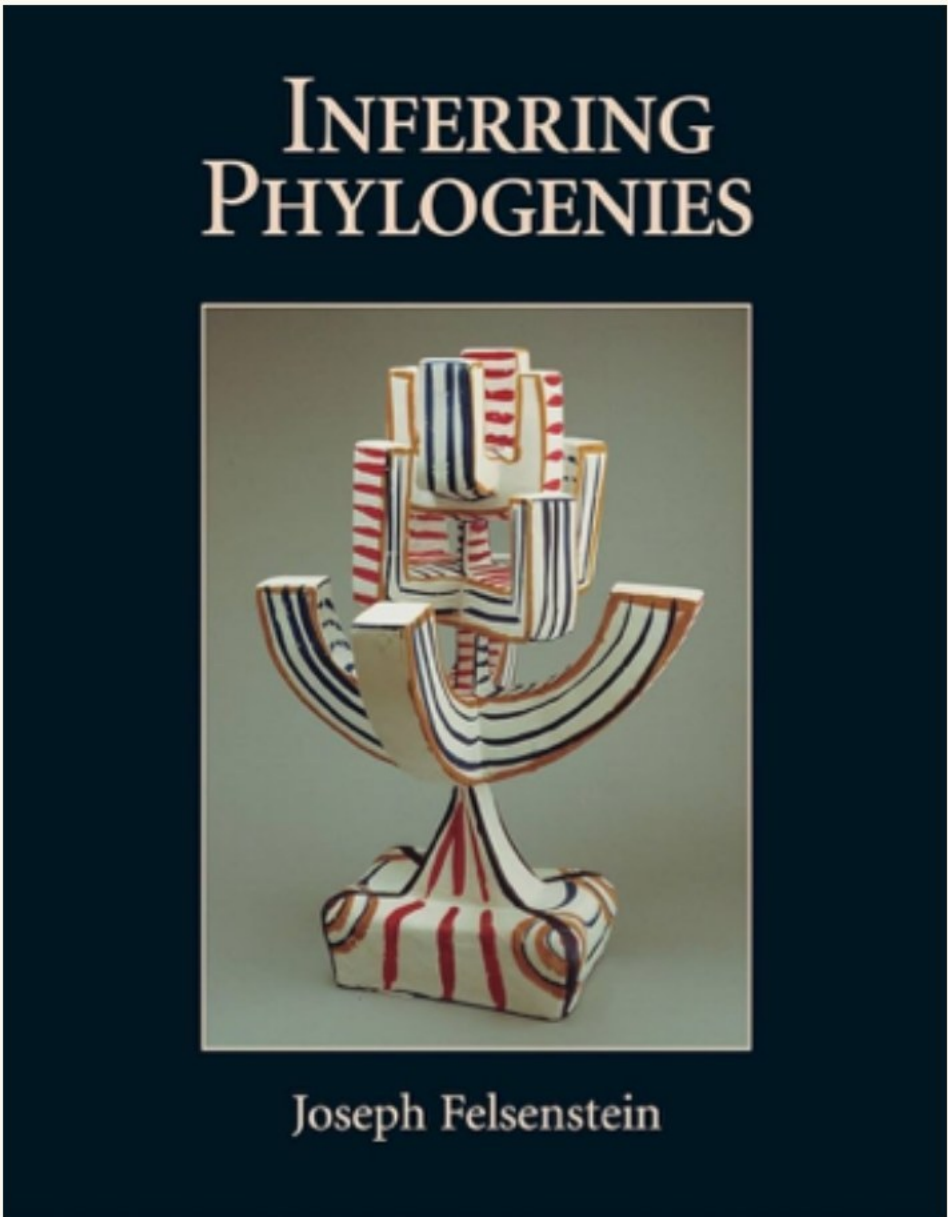
$$d(T, T') = \frac{|\{a, b, c, d\} : T(a, b, c, d) \neq T'(a, b, c, d)|}{\binom{n}{4}}$$

Quartet Distance

Definition (quartet distance)

The quartet distance between two phylogenetic trees T and T' is defined as the fraction of quartets on which the two trees disagree, i.e.,

$$d(T, T') = \frac{|\{a, b, c, d\} : T(a, b, c, d) \neq T'(a, b, c, d)|}{\binom{n}{4}}$$



30 Consensus trees and distances between trees	521
Consensus trees	521
Strict consensus	522
Majority-rule consensus	523
Adams consensus tree	524
A dismaying result	525
Consensus using branch lengths	526
Other consensus tree methods	526
Consensus subtrees	528
Distances between trees	528
The symmetric difference	529
The quartets distance	530
The nearest-neighbor interchange distance	530
The path-length-difference metric	531
Distances using branch lengths	531
Are these distances truly distances?	533
Consensus trees and distances	534
Trees significantly the same? different?	534
What do consensus trees and tree distances tell us?	535
The total evidence debate	536
A modest proposal	537

Our Results

Theorem (Optimal Reconstruction)

There exists an algorithm with runtime $O_\epsilon(\text{poly}(n))$ that given $O_\epsilon(n)$ samples from the RCN model recovers a tree $\hat{T}$ that is within distance ϵ to the ground truth tree T :

$$d(T, \hat{T}) \leq \epsilon$$

Our Results

Theorem (Optimal Reconstruction)

There exists an algorithm with runtime $O_\epsilon(\text{poly}(n))$ that given $O_\epsilon(n)$ samples from the RCN model recovers a tree $\hat{T}$ that is within distance ϵ to the ground truth tree T :

$$d(T, \hat{T}) \leq \epsilon$$

Remarks:

- Lower bound of $\Omega(n)$ for any meaningful reconstruction (optimal samples)

Our Results

Theorem (Optimal Reconstruction)

There exists an algorithm with runtime $O_\epsilon(\text{poly}(n))$ that given $O_\epsilon(n)$ samples from the RCN model recovers a tree $\hat{T}$ that is within distance ϵ to the ground truth tree T :

$$d(T, \hat{T}) \leq \epsilon$$

Remarks:

- Lower bound of $\Omega(n)$ for any meaningful reconstruction (optimal samples)
- We achieve recovery without even seeing all the species (requires $\Omega(n \log(n))$)

Our Results

Theorem (Optimal Reconstruction)

There exists an algorithm with runtime $O_\epsilon(\text{poly}(n))$ that given $O_\epsilon(n)$ samples from the RCN model recovers a tree $\hat{T}$ that is within distance ϵ to the ground truth tree T :

$$d(T, \hat{T}) \leq \epsilon$$

Remarks:

- Lower bound of $\Omega(n)$ for any meaningful reconstruction (optimal samples)
- We achieve recovery without even seeing all the species (requires $\Omega(n \log(n))$)
- Exact reconstruction requires $\Theta(n^3)$ samples

Our Results

Theorem (Optimal Reconstruction)

There exists an algorithm with runtime $O_\epsilon(\text{poly}(n))$ that given $O_\epsilon(n)$ samples from the RCN model recovers a tree $\hat{T}$ that is within distance ϵ to the ground truth tree T :

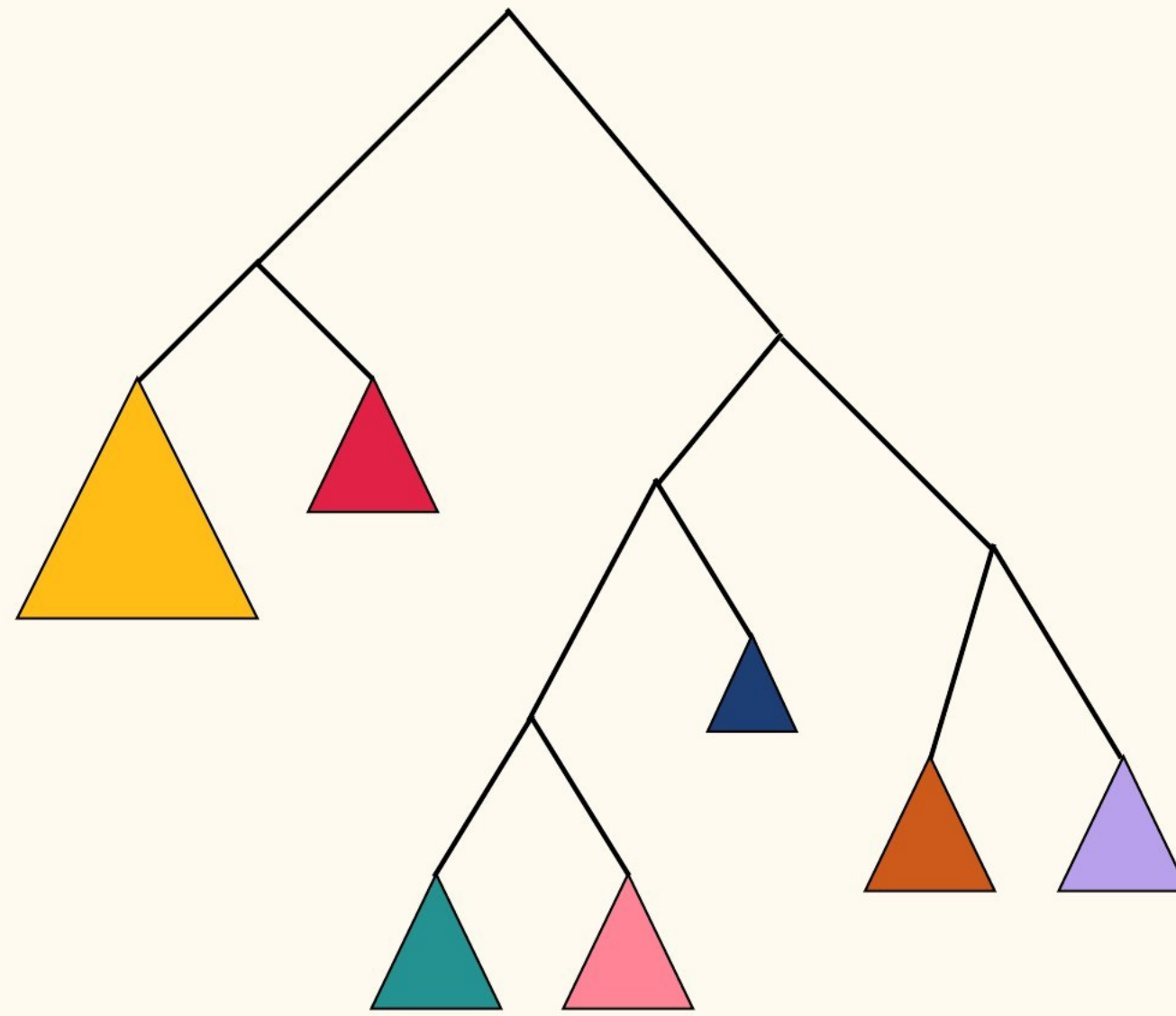
$$d(T, \hat{T}) \leq \epsilon$$

Remarks:

- Lower bound of $\Omega(n)$ for any meaningful reconstruction (optimal samples)
- We achieve recovery without even seeing all the species (requires $\Omega(n \log(n))$)
- Exact reconstruction requires $\Theta(n^3)$ samples
- Improves upon [Snir, Yuster 2011]

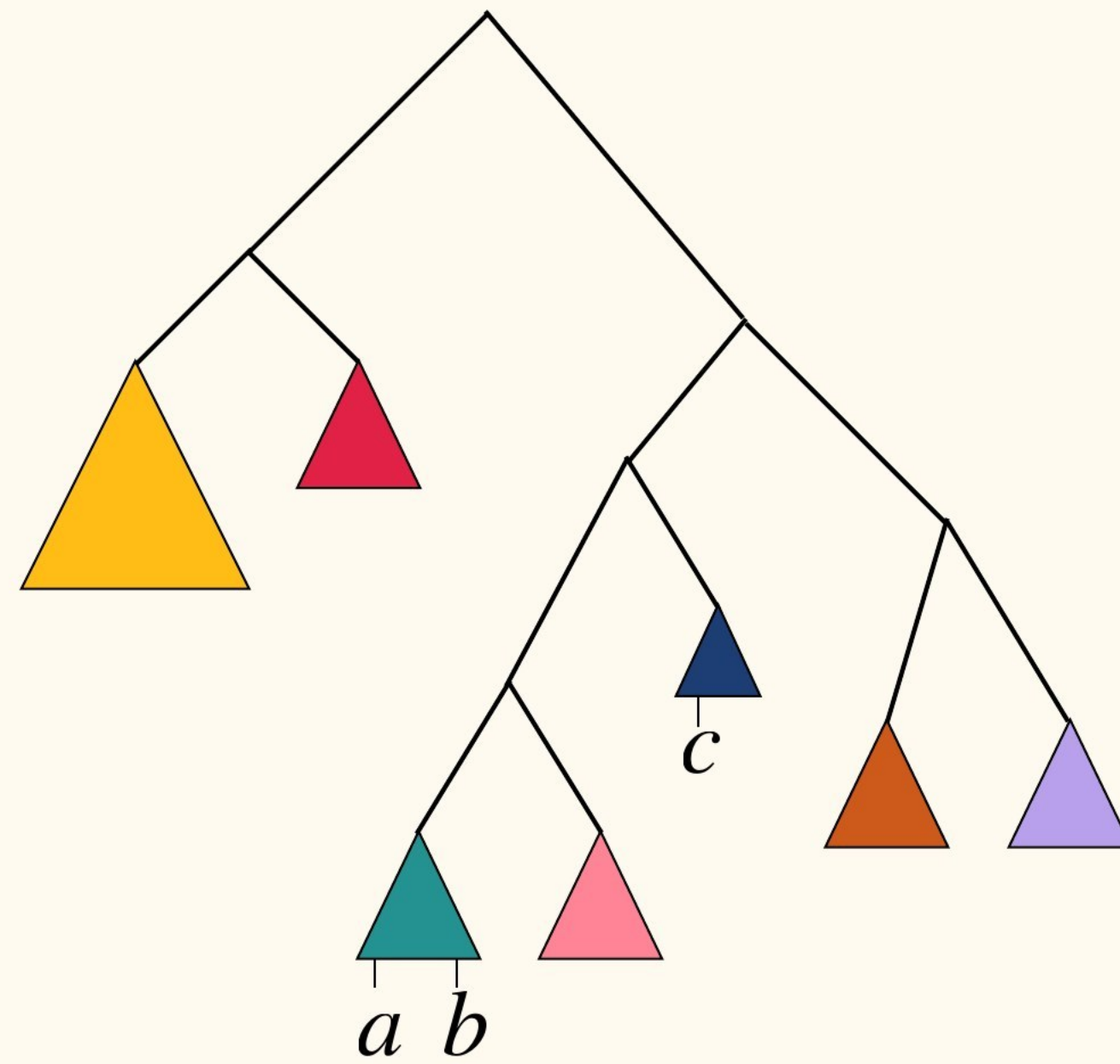
Proof sketch

Consider the ground truth tree T (assume for now that it is more or less balanced)



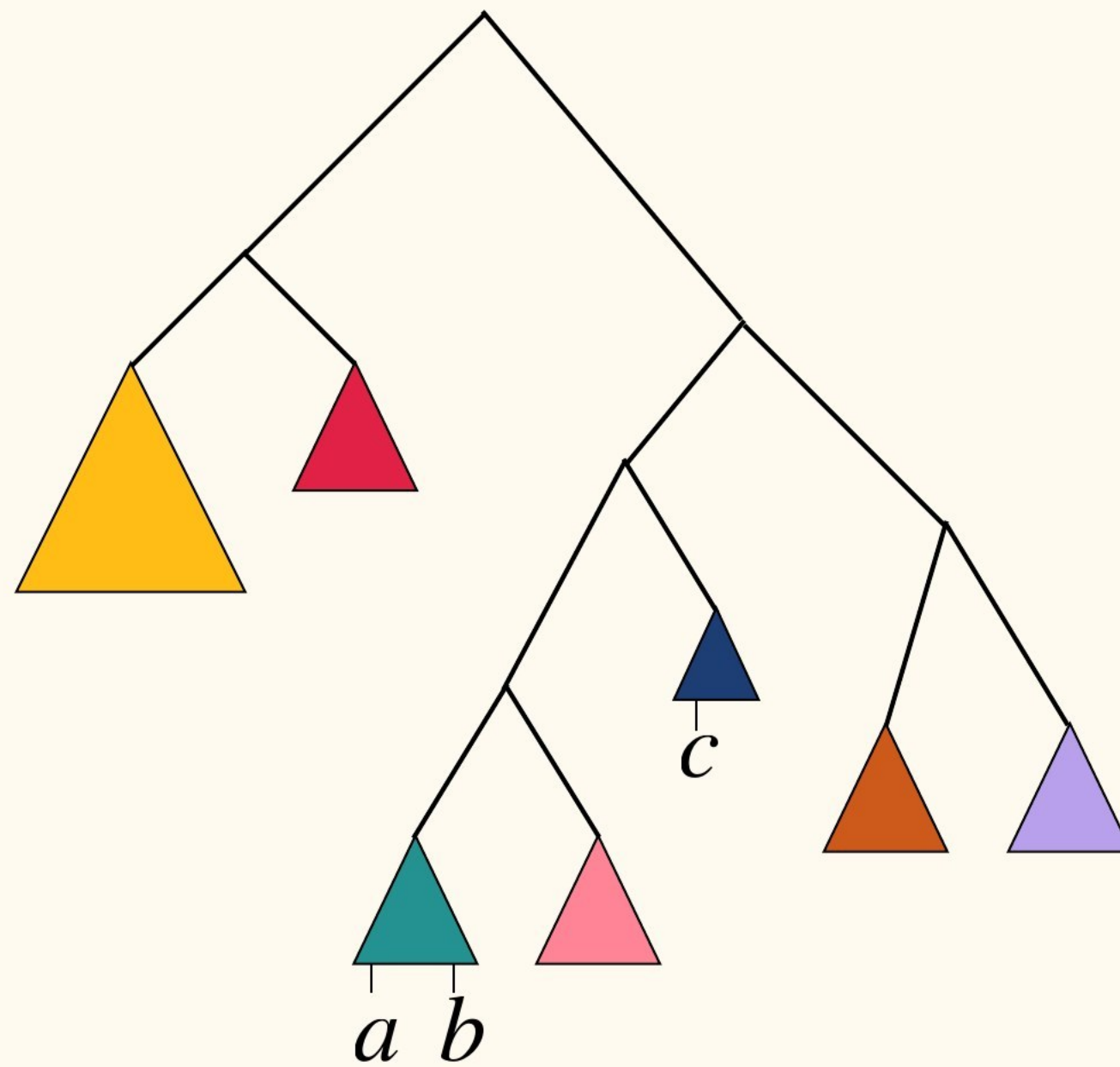
Proof sketch

Consider the ground truth tree T (assume for now that it is more or less balanced)



Proof sketch

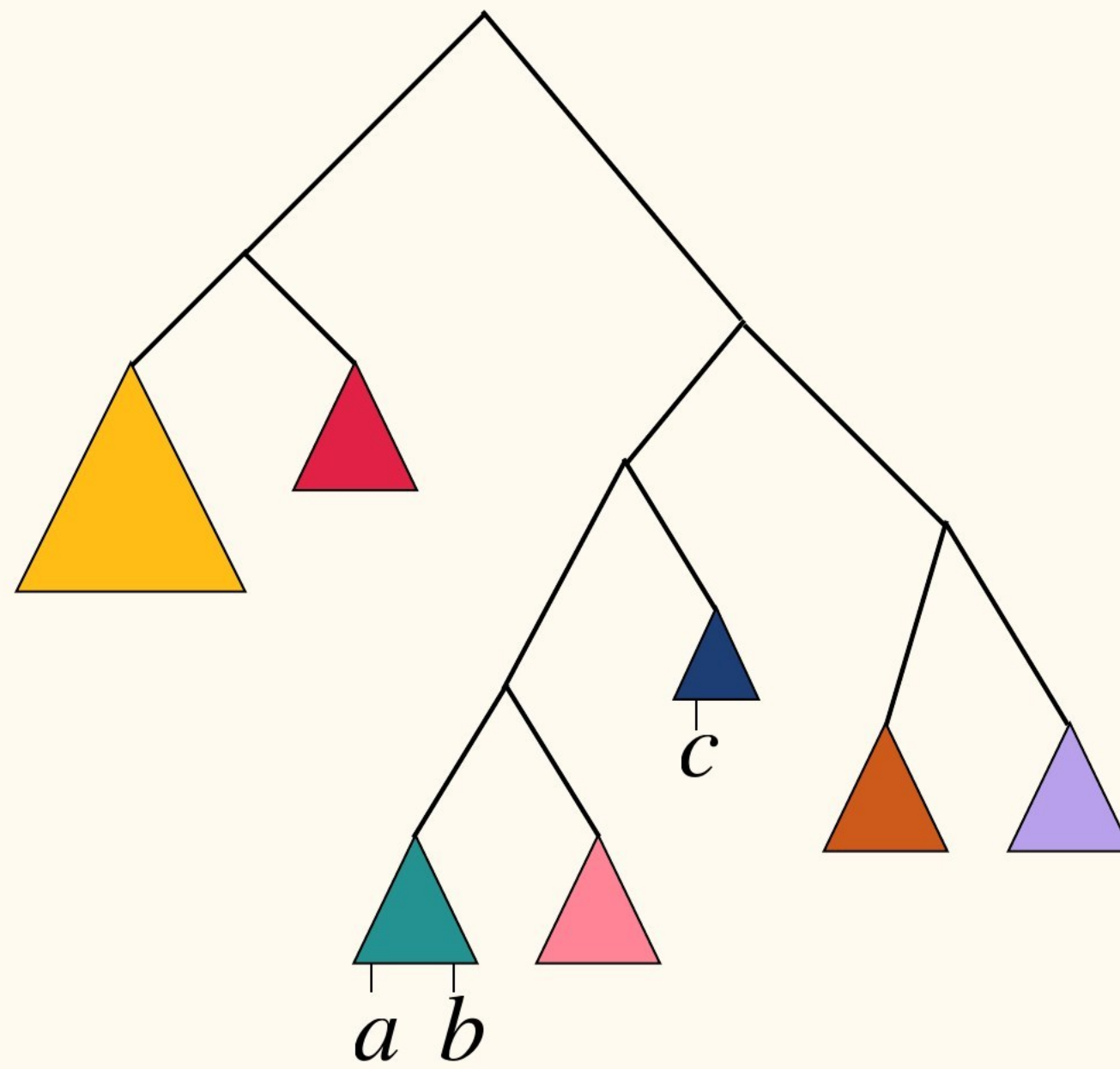
Consider the ground truth tree T (assume for now that it is more or less balanced)



 Observation: The probability that a quartet $ab \mid ______$ appears is **slightly higher** than the probability of $ac \mid ______$

Proof sketch

Consider the ground truth tree T (assume for now that it is more or less balanced)



 Observation: The probability that a quartet $ab \mid ______$ appears is **slightly higher** than the probability of $ac \mid ______$

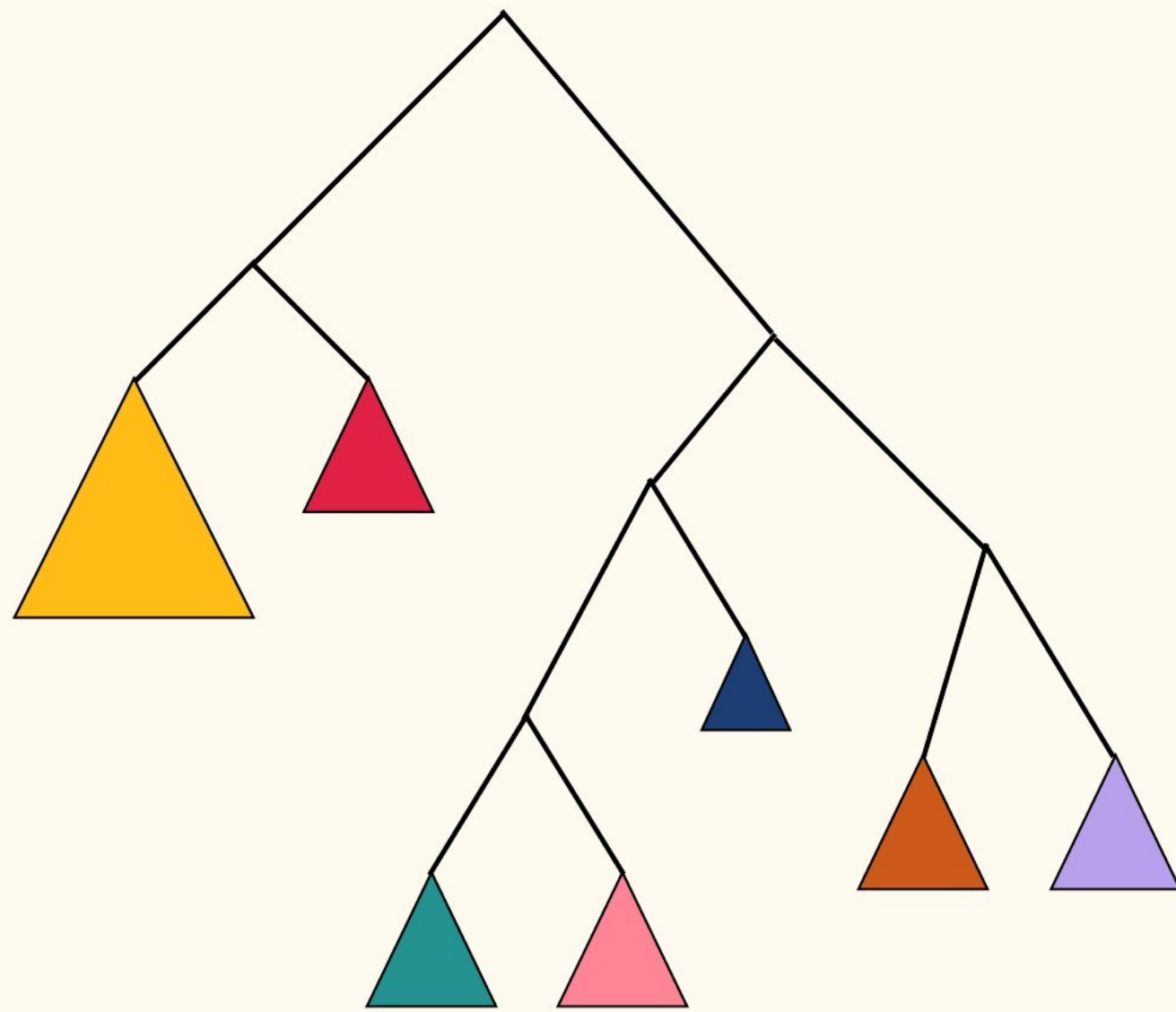
$$\mathbb{P}_{q \sim \text{RCN}}(q = ab \mid ______) \propto 1 - \frac{2|T_{ab}|}{n} + O\left((|T_{ab}|/n)^2\right)$$

Proof sketch



Observation: The probability that a quartet $ab \mid _ _$ appears is **slightly higher** than the probability of $ac \mid _ _$

Construct a graph on the leaves of the tree!



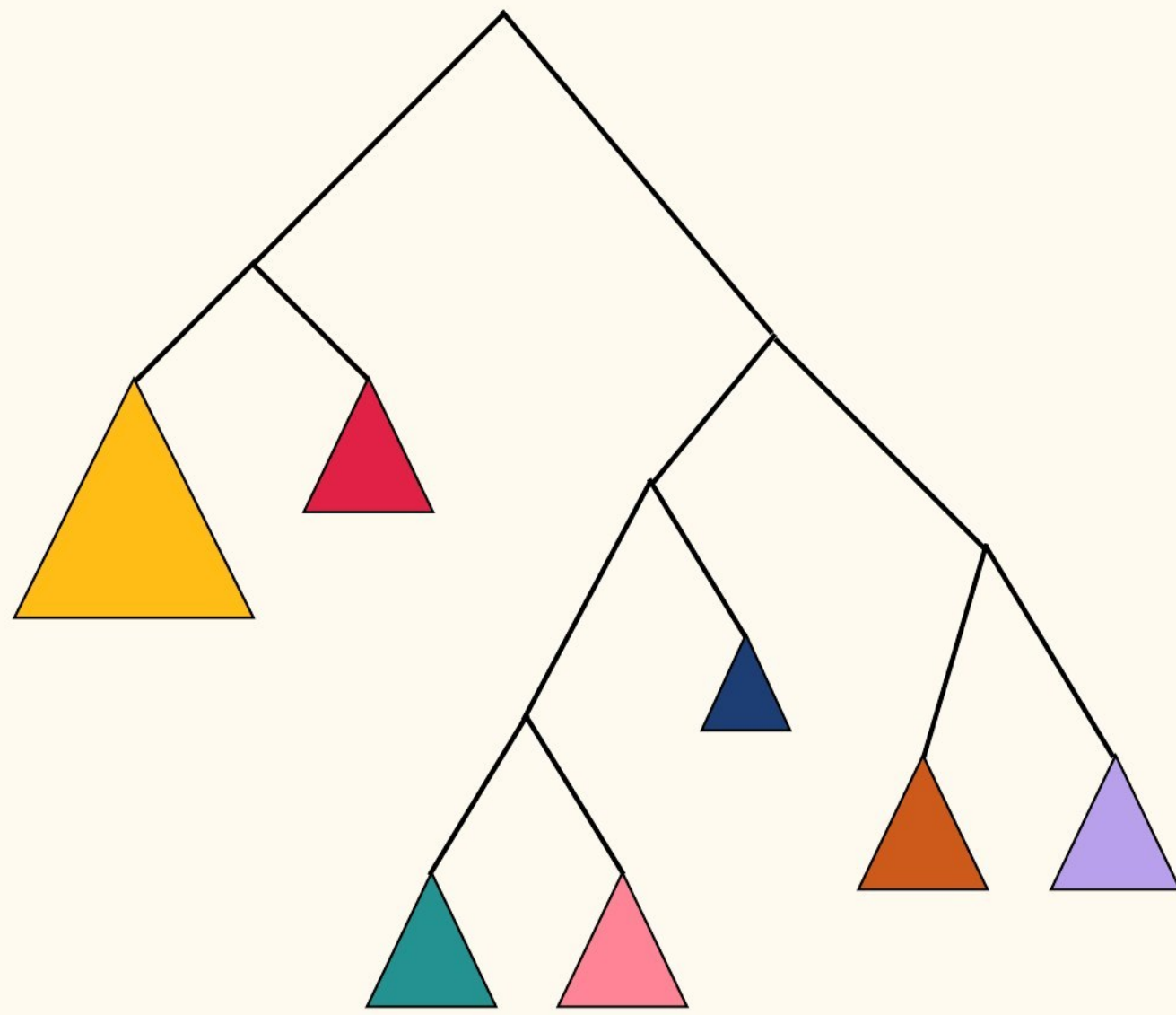
Proof sketch



Observation: The probability that a quartet $ab \mid __\$ appears is **slightly higher** than the probability of $ac \mid __\$

Construct a graph on the leaves of the tree!

-Add edge (a, b) if a quartet $ab \mid __\$ appears in the samples



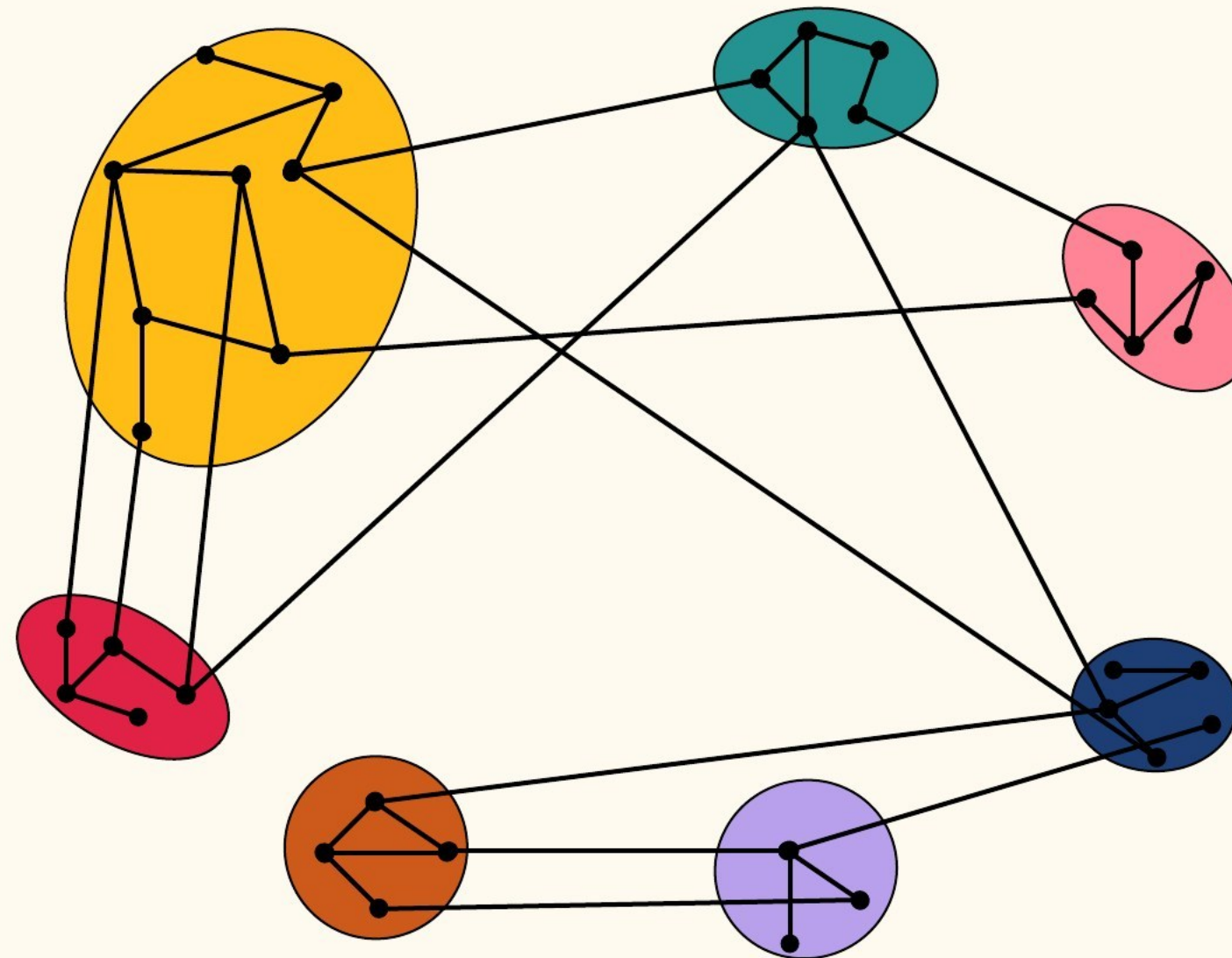
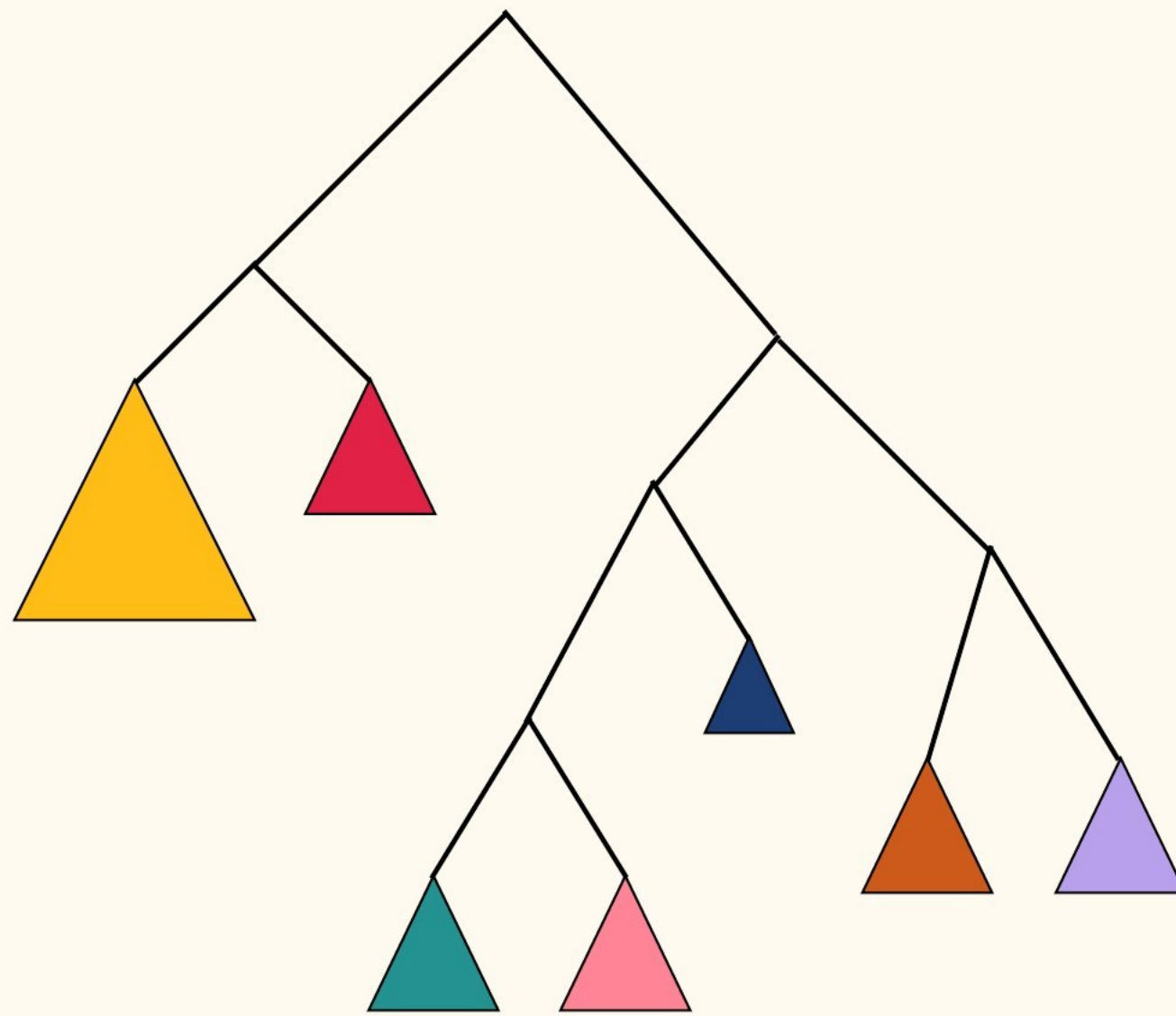
Proof sketch



Observation: The probability that a quartet $ab \mid _ _$ appears is **slightly higher** than the probability of $ac \mid _ _$

Construct a graph on the leaves of the tree!

- Add edge (a, b) if a quartet $ab \mid _ _$ appears in the samples
- Structure like a **stochastic block model**!



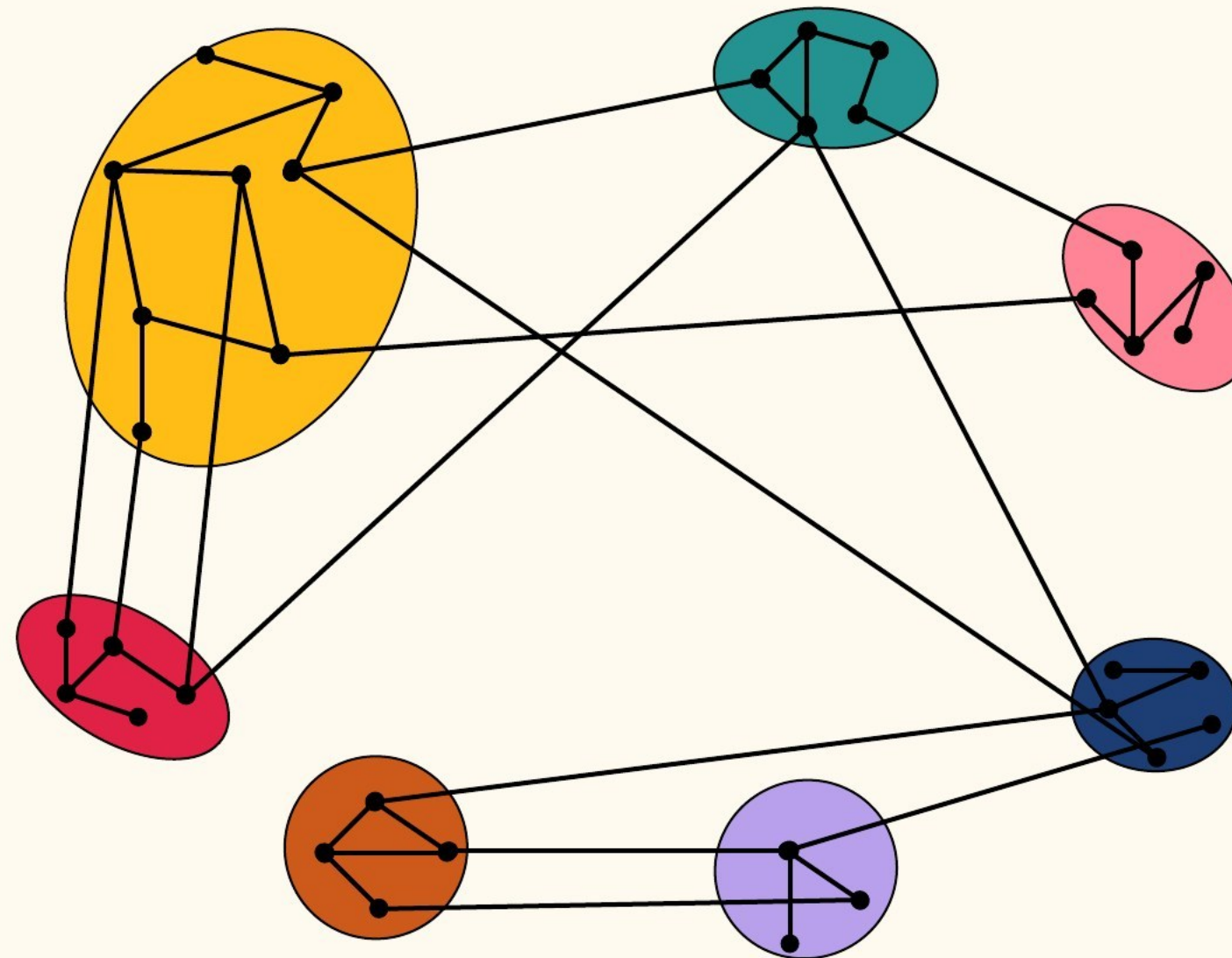
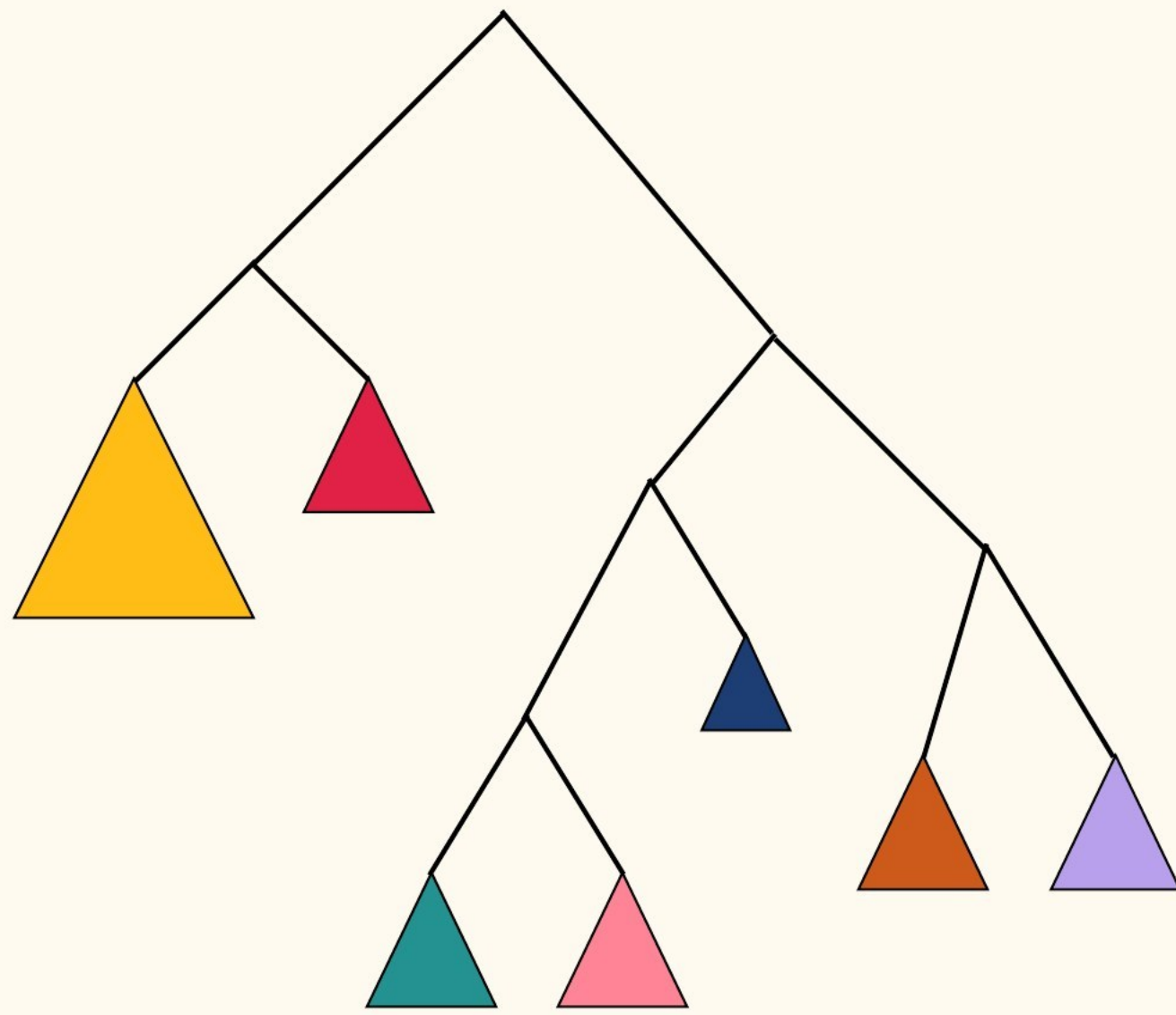
Proof sketch



Observation: The probability that a quartet $ab \mid _ _$ appears is **slightly higher** than the probability of $ac \mid _ _$

Construct a graph on the leaves of the tree!

- Add edge (a, b) if a quartet $ab \mid _ _$ appears in the samples
- Structure like a **stochastic block model**!



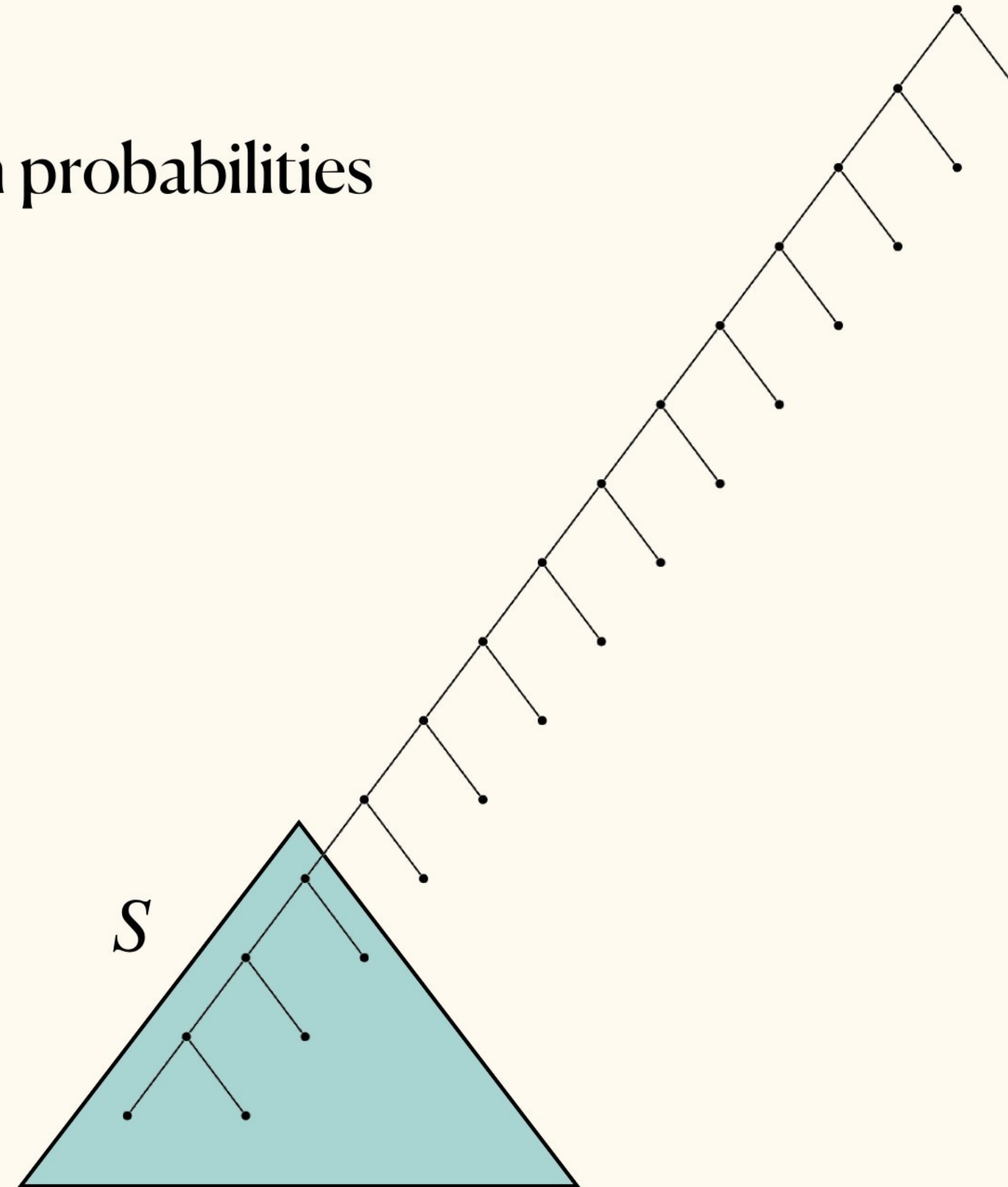
- Recover the communities
- Use them to reconstruct the tree

Proof sketch

The situation is more subtle:

Edge probabilities **change gradually**.

Subtree S does not exhibit a gap in probabilities



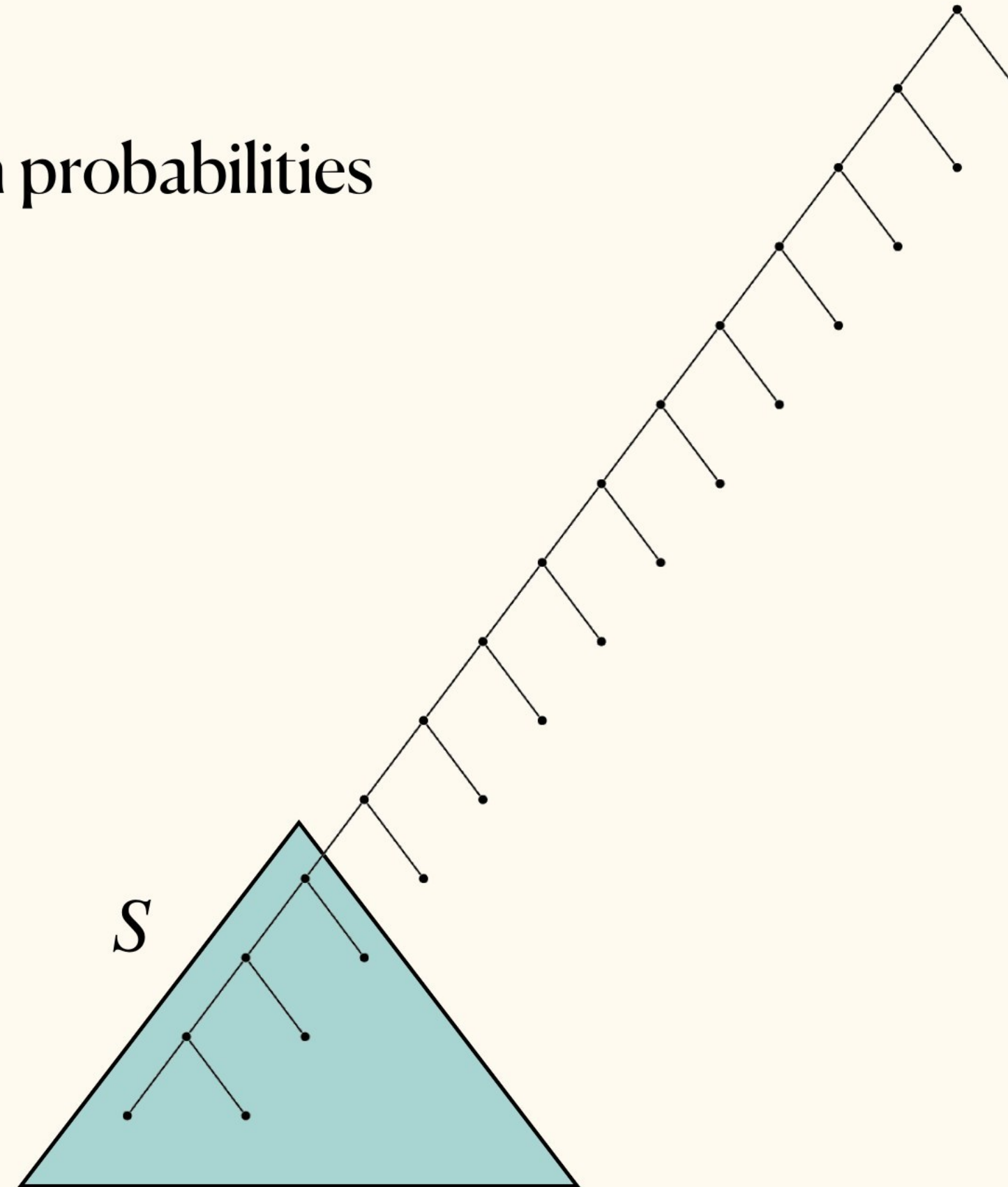
Proof sketch

The situation is more subtle:

Edge probabilities **change gradually**.

Subtree S does not exhibit a gap in probabilities

Introduce a buffer set B



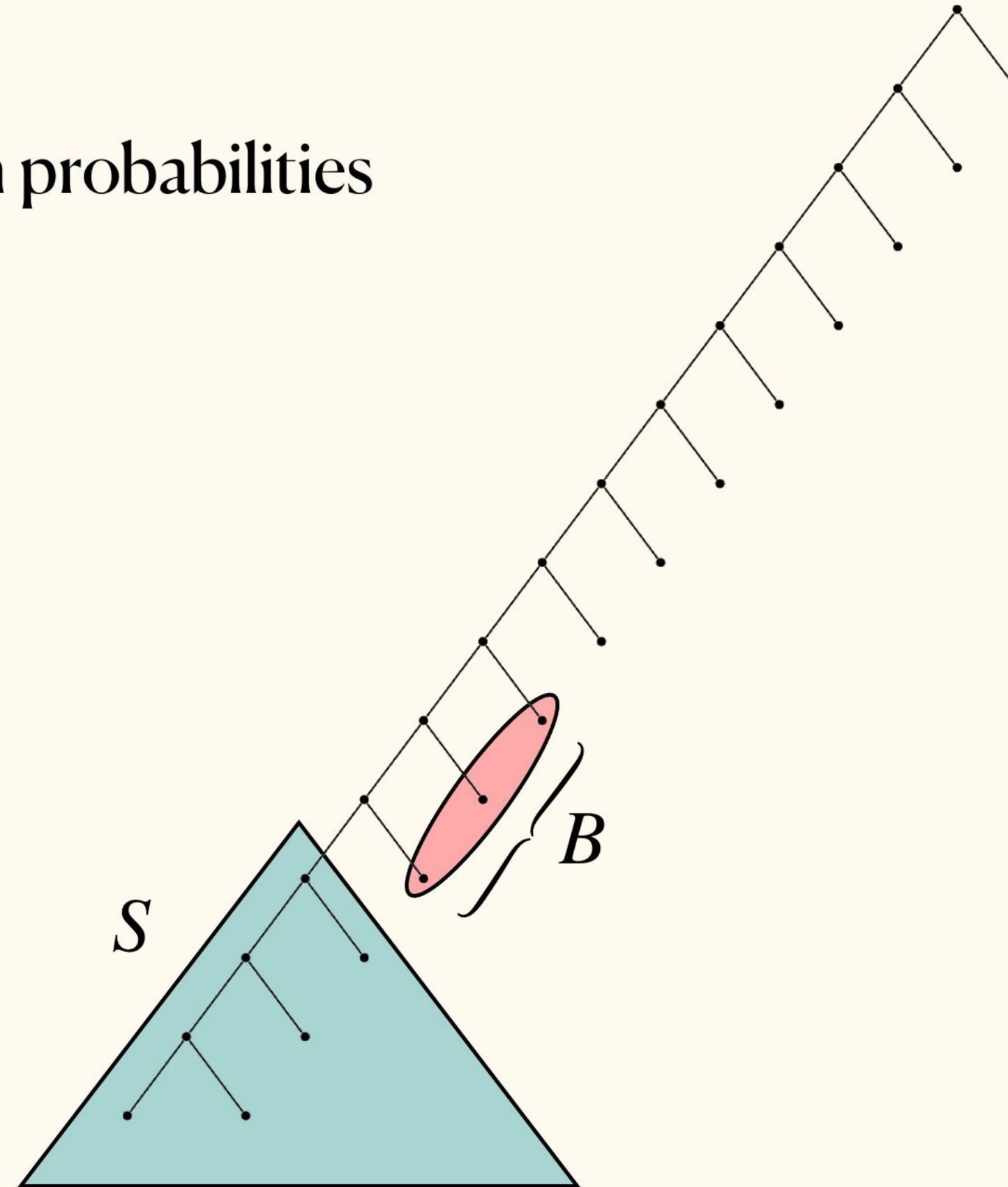
Proof sketch

The situation is more subtle:

Edge probabilities **change gradually**.

Subtree S does not exhibit a gap in probabilities

Introduce a buffer set B

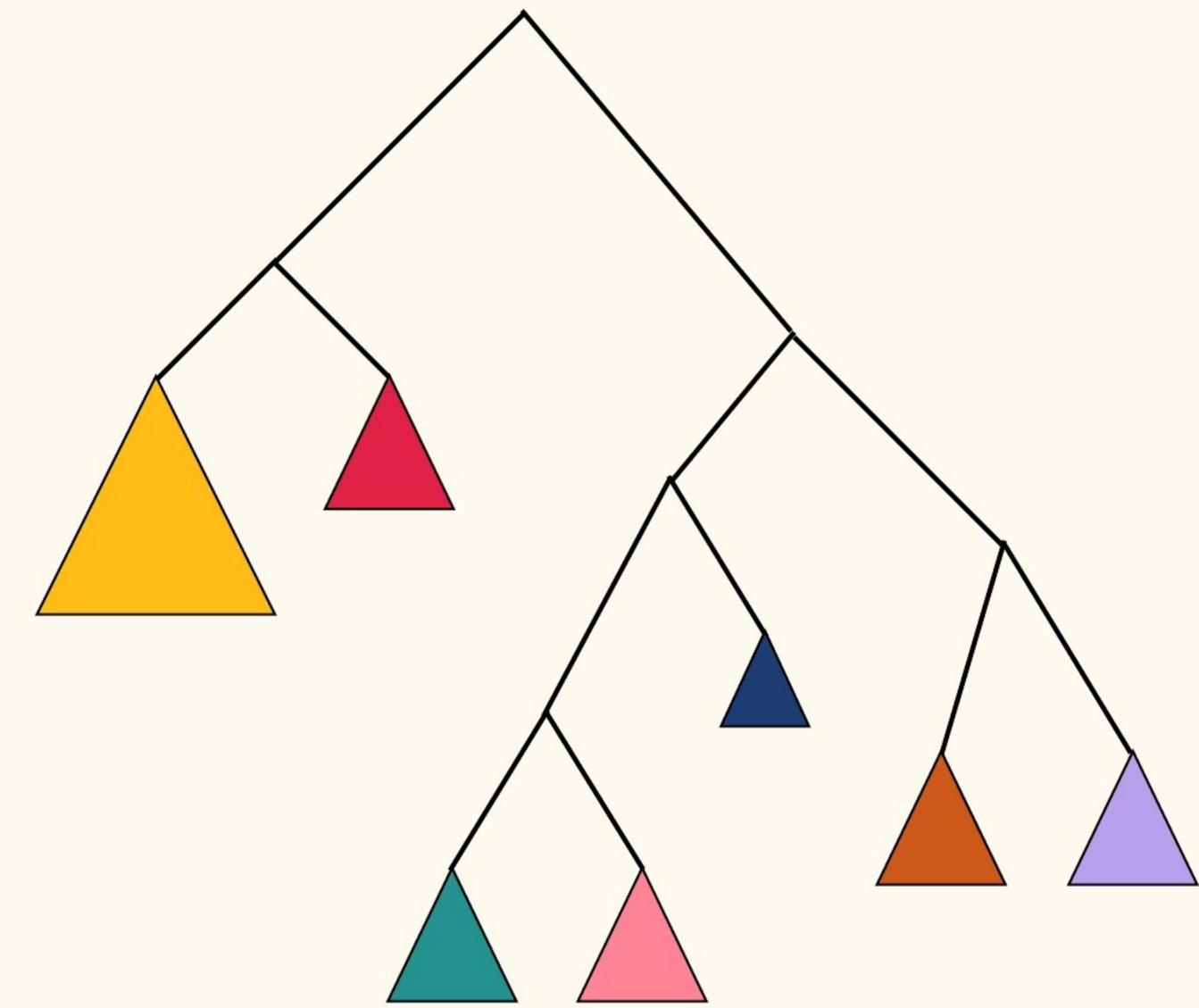
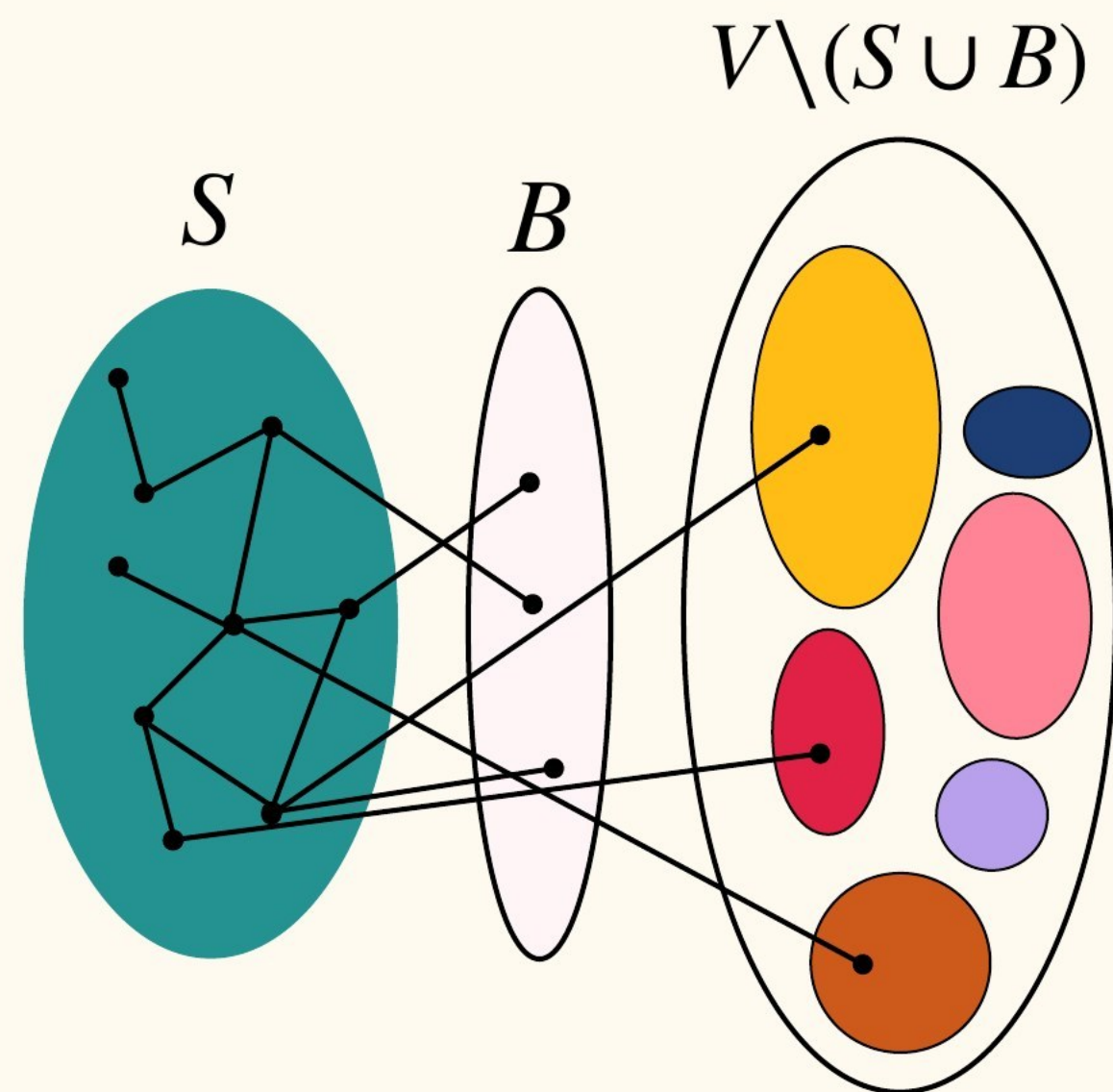


Proof sketch

Add edge (a, b) if a quartet $ab | _ _$ appears in the samples

Consider a subtree with leaves S , $|S| \approx \varepsilon n$ 

Buffer B , $|B| \approx \varepsilon^2 n$

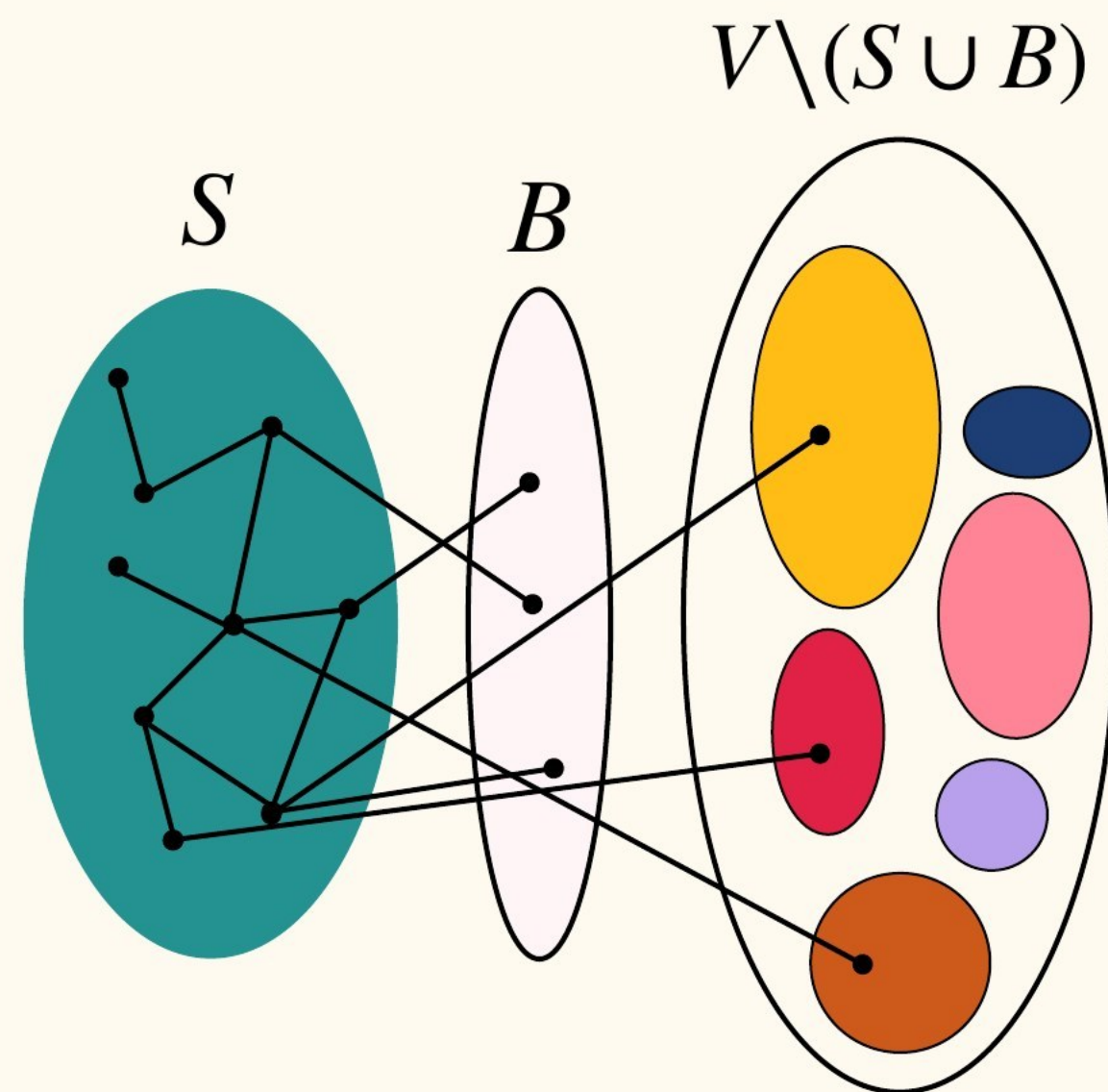


Proof sketch

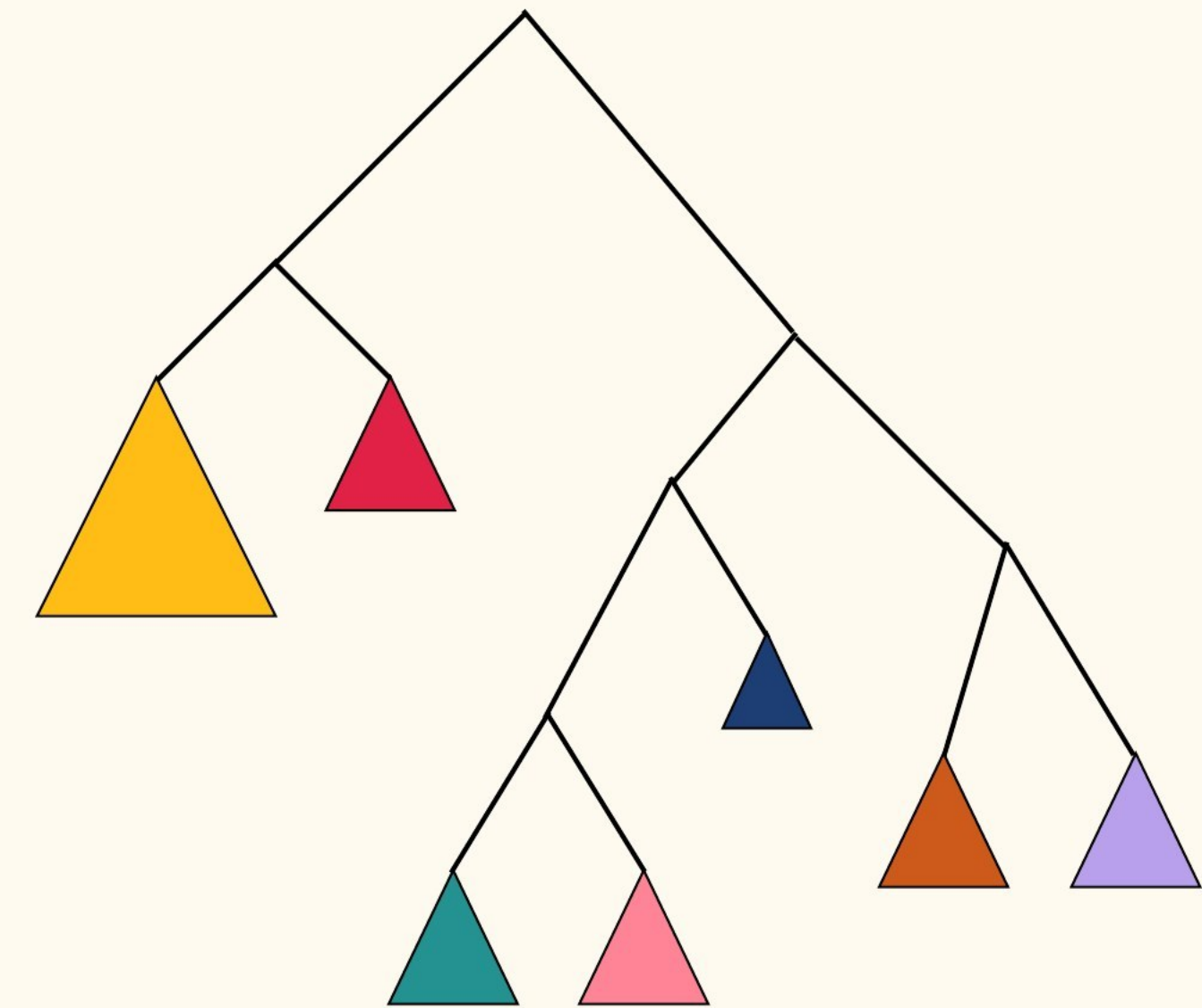
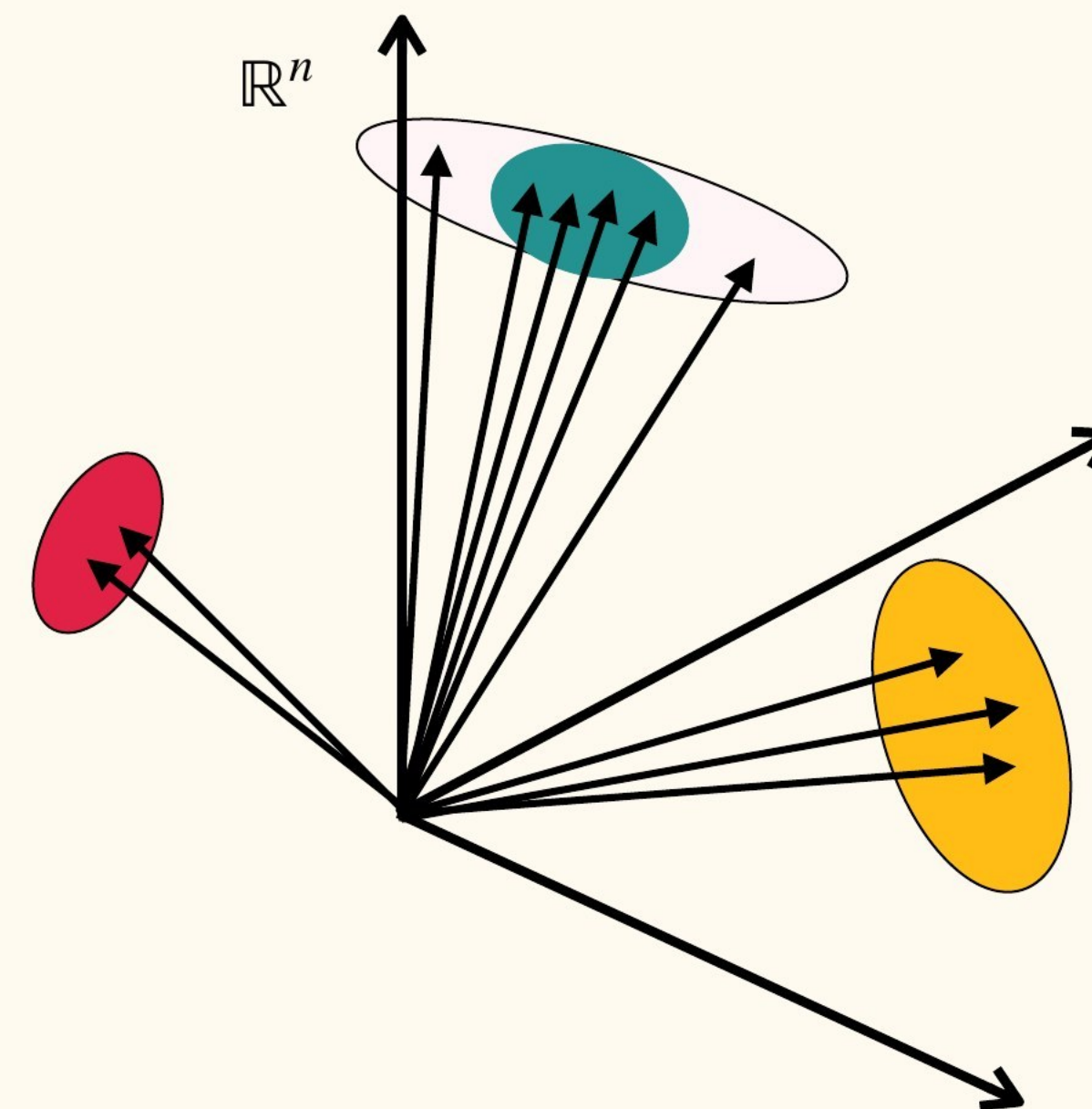
Add edge (a, b) if a quartet $ab | _ _$ appears in the samples

Consider a subtree with leaves S , $|S| \approx \varepsilon n$ ▲

Buffer B , $|B| \approx \varepsilon^2 n$

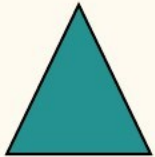


Using an SDP algorithm **recover** S up to $\approx \varepsilon^2 n$ **error** [MMV16]

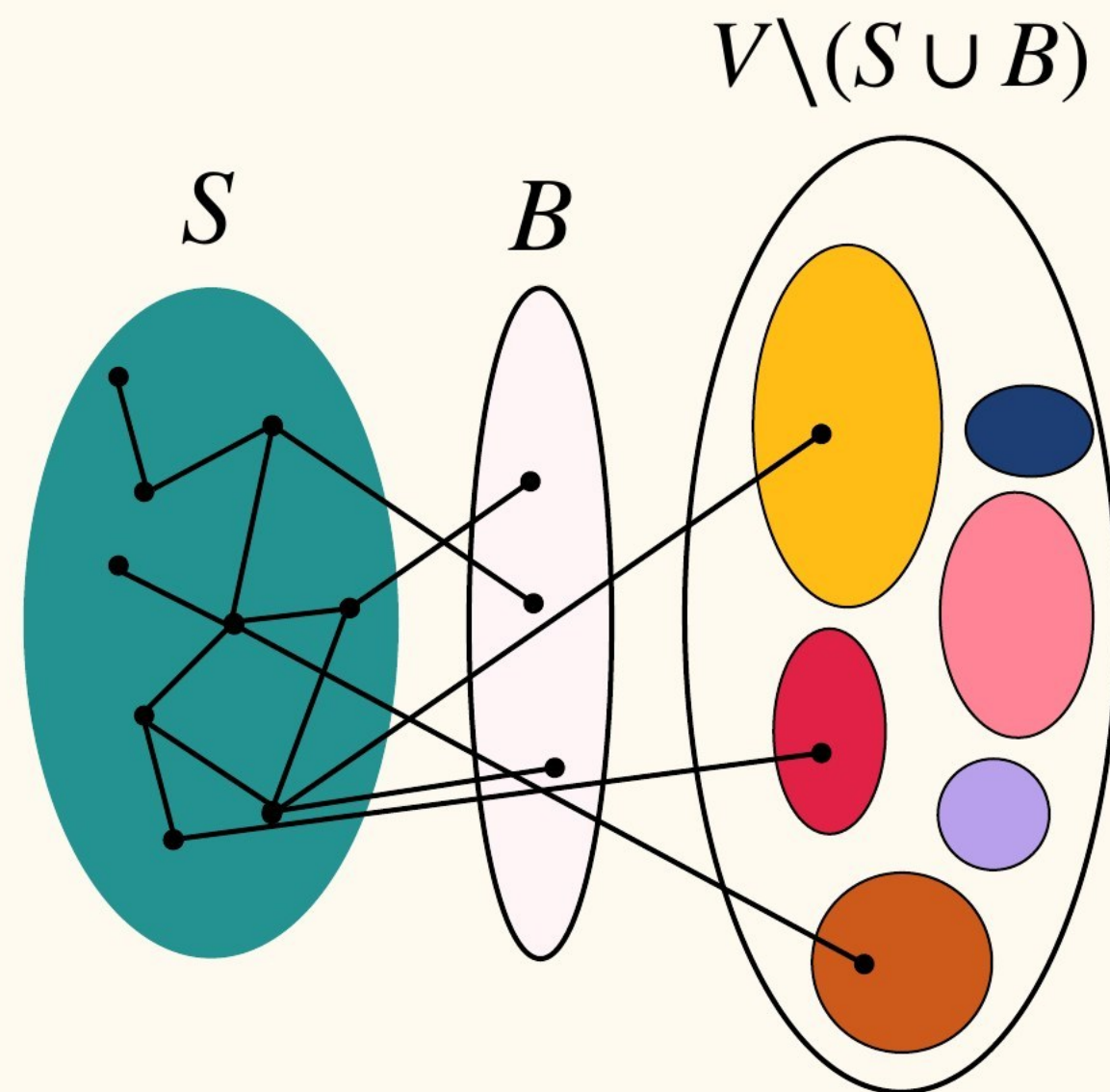


Proof sketch

Add edge (a, b) if a quartet $ab | _ _$ appears in the samples

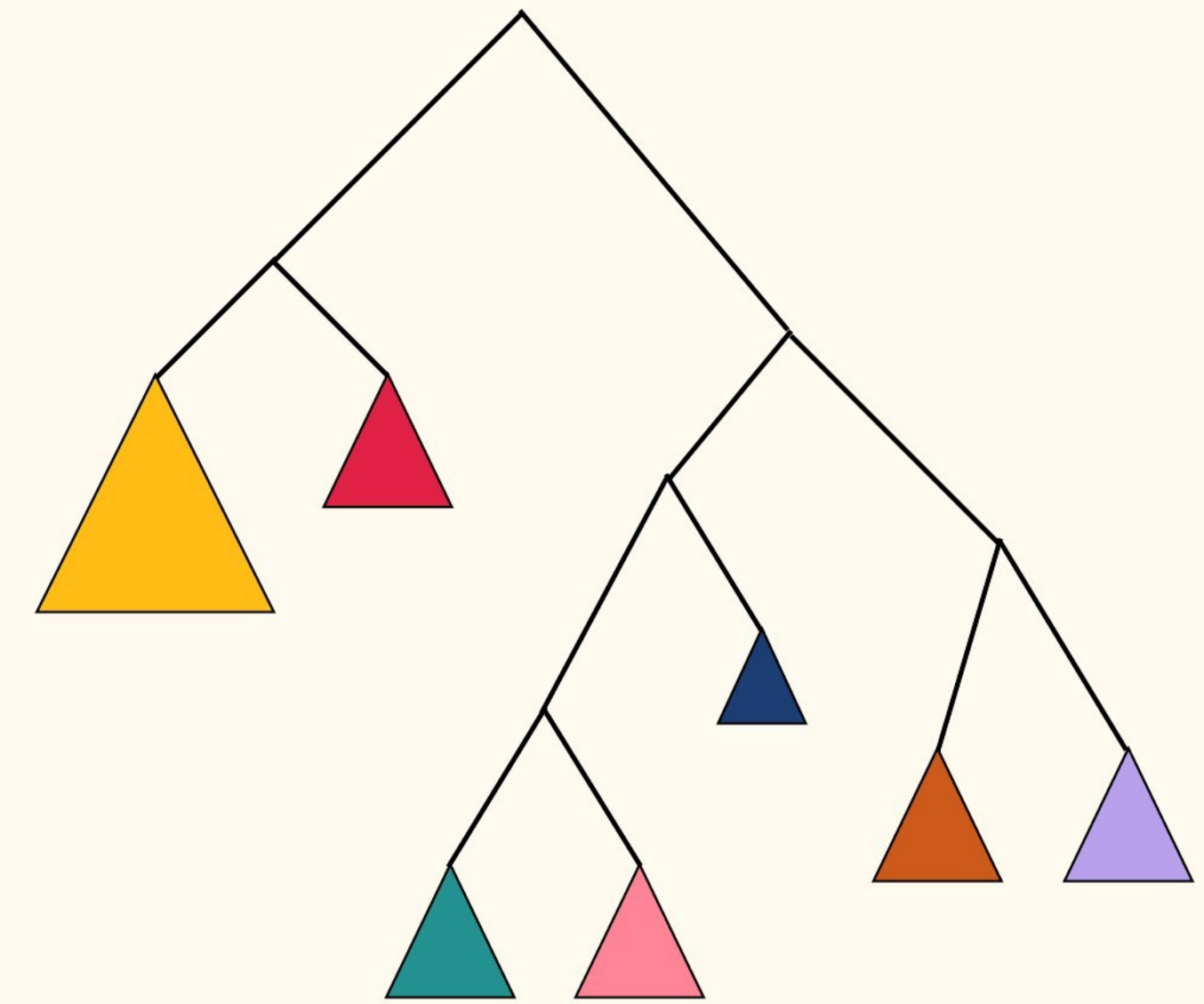
Consider a subtree with leaves S , $|S| \approx \varepsilon n$ 

Buffer B , $|B| \approx \varepsilon^2 n$



Using an SDP algorithm **recover** S up to $\approx \varepsilon^2 n$ **error** [MMV16]

Repeat the procedure $1/\varepsilon$ **times**

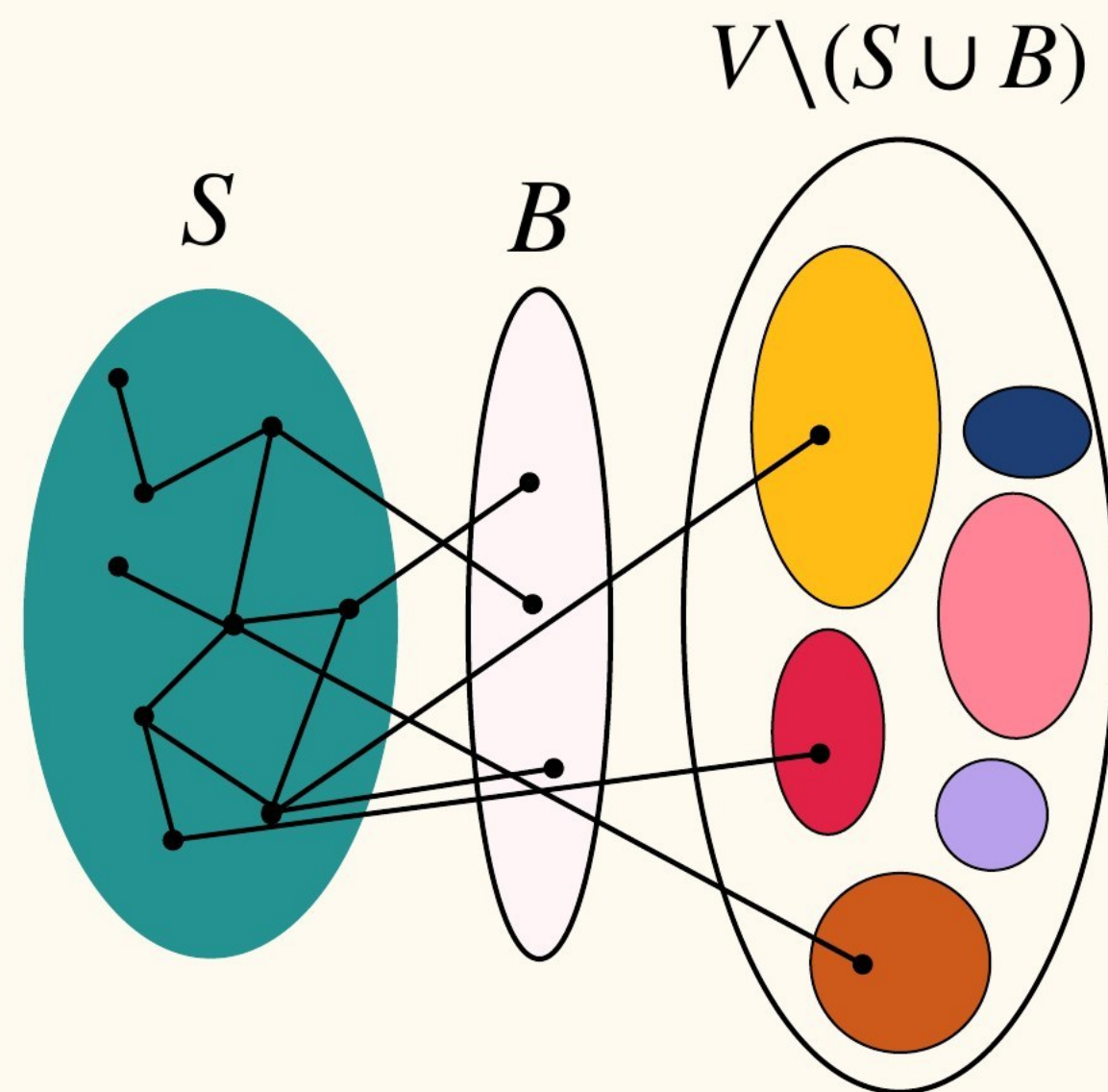


Proof sketch

Add edge (a, b) if a quartet $ab | _ _$ appears in the samples

Consider a subtree with leaves S , $|S| \approx \varepsilon n$ 

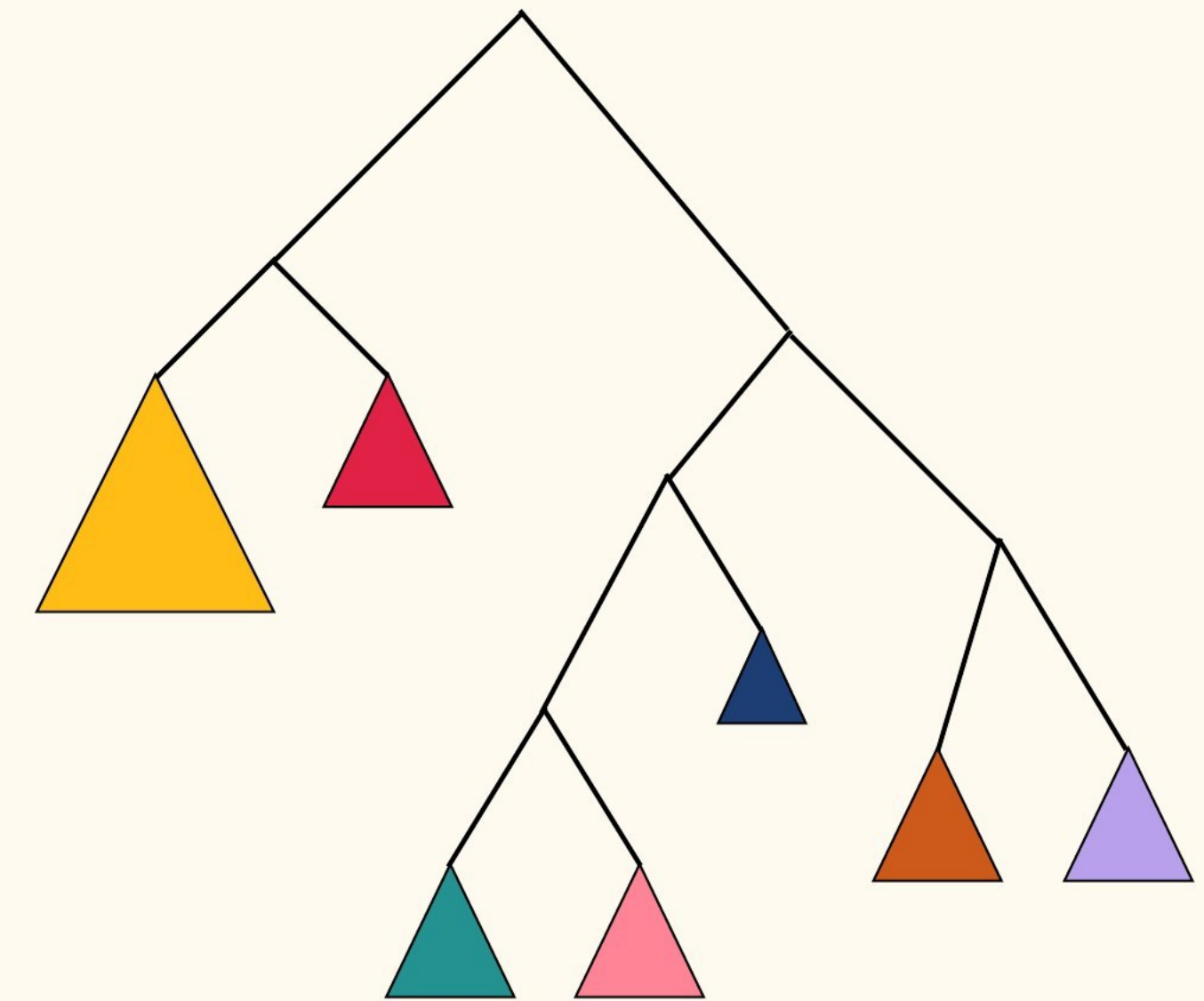
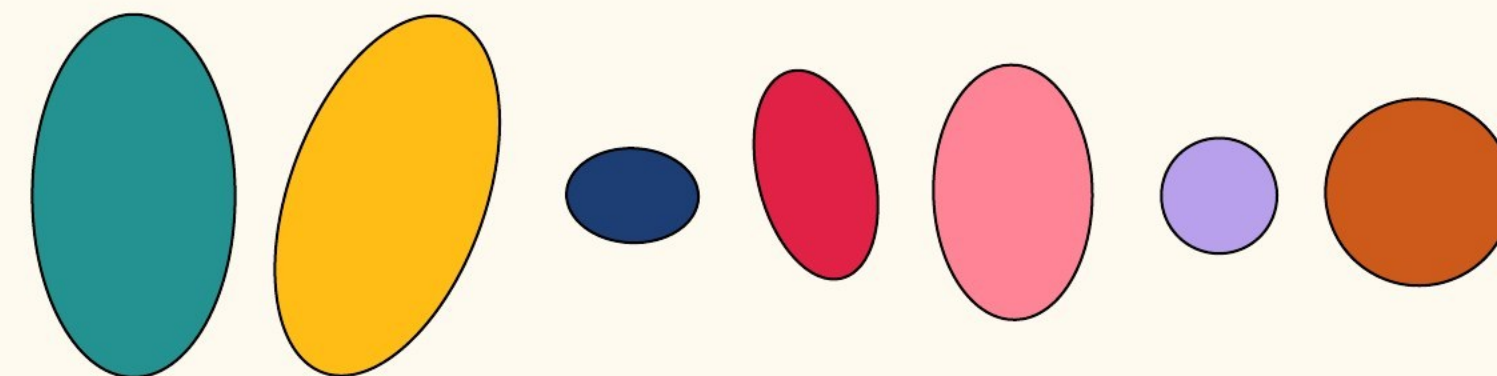
Buffer B , $|B| \approx \varepsilon^2 n$



Using an SDP algorithm **recover** S up to $\approx \varepsilon^2 n$ **error** [MMV16]

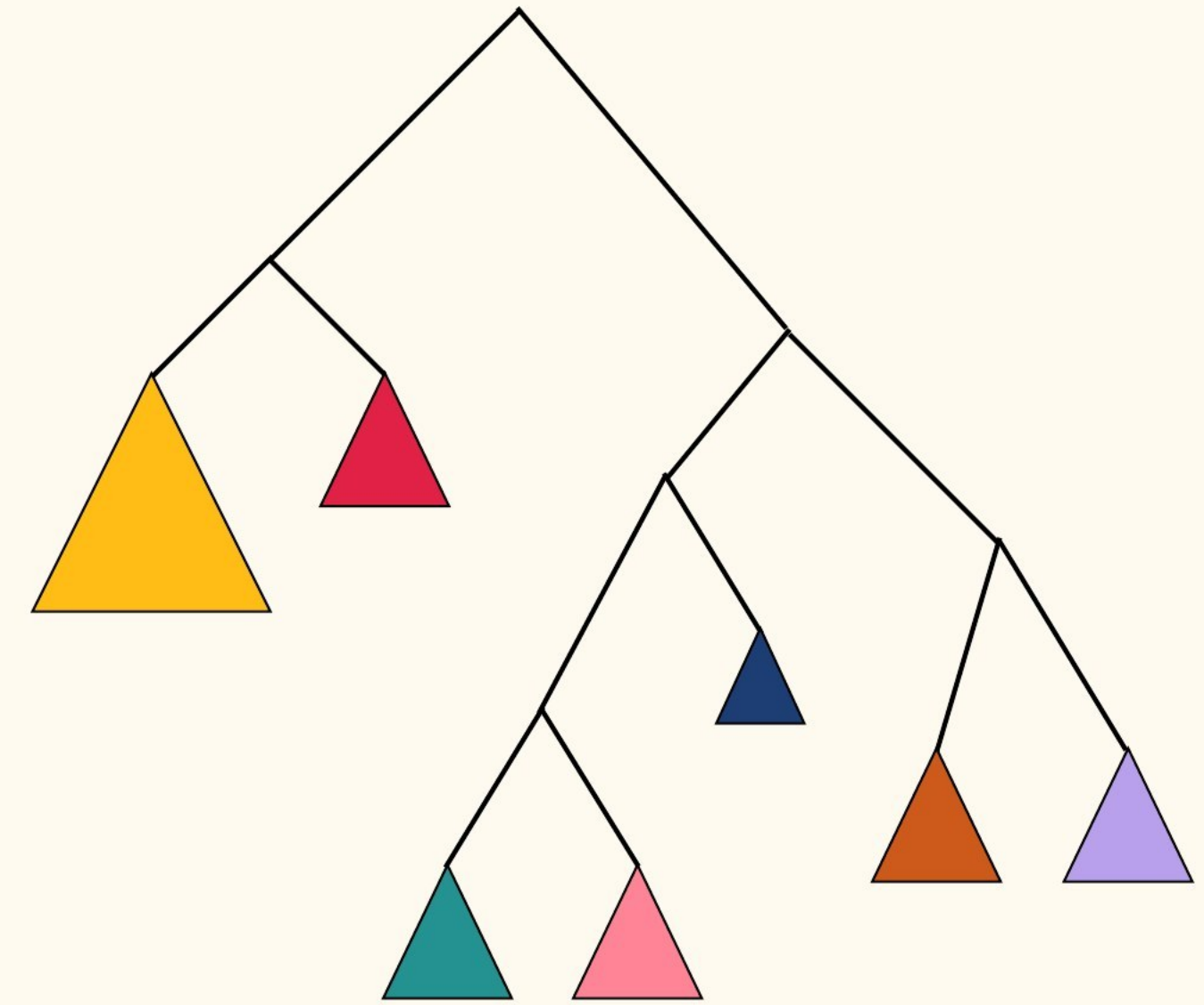
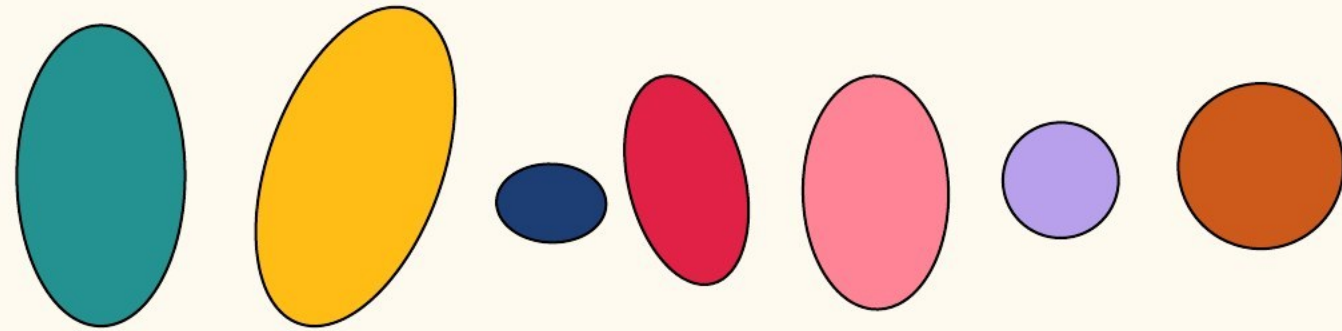
Repeat the procedure $1/\varepsilon$ **times**

Recover sets $S_1, S_2, \dots, S_{1/\varepsilon}$



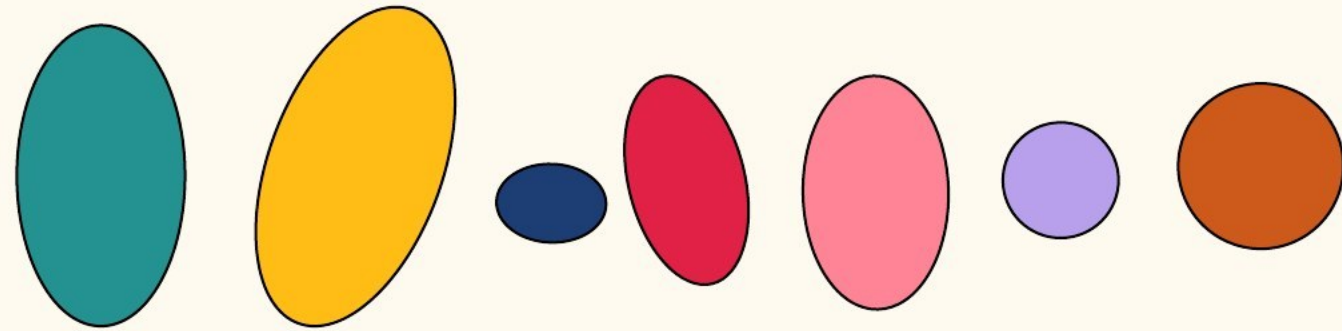
Proof sketch

Recover sets $S_1, S_2, \dots, S_{1/\varepsilon}$

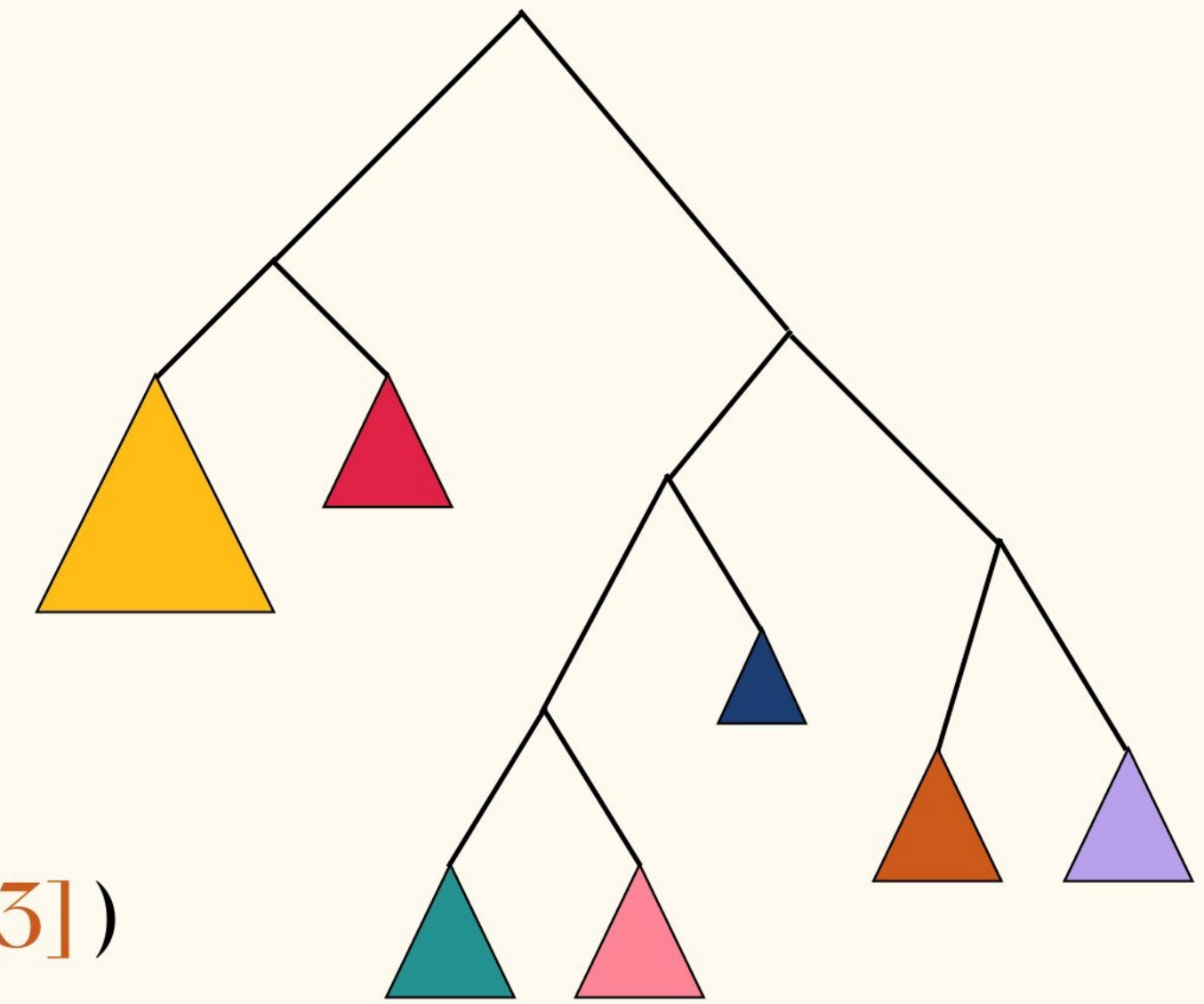
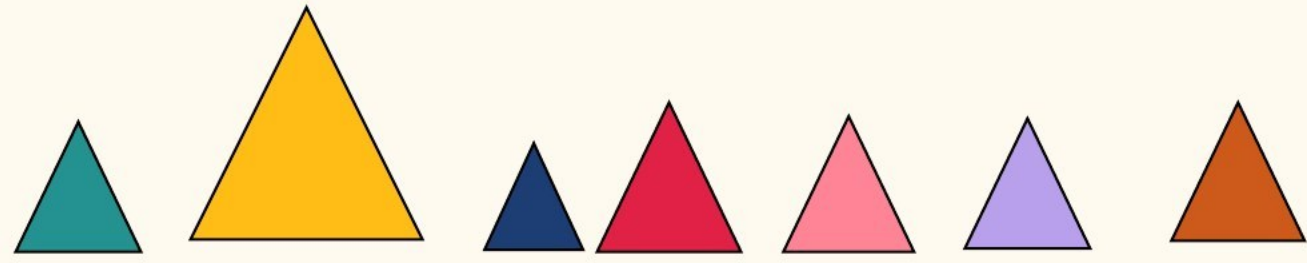


Proof sketch

Recover sets $S_1, S_2, \dots, S_{1/\varepsilon}$

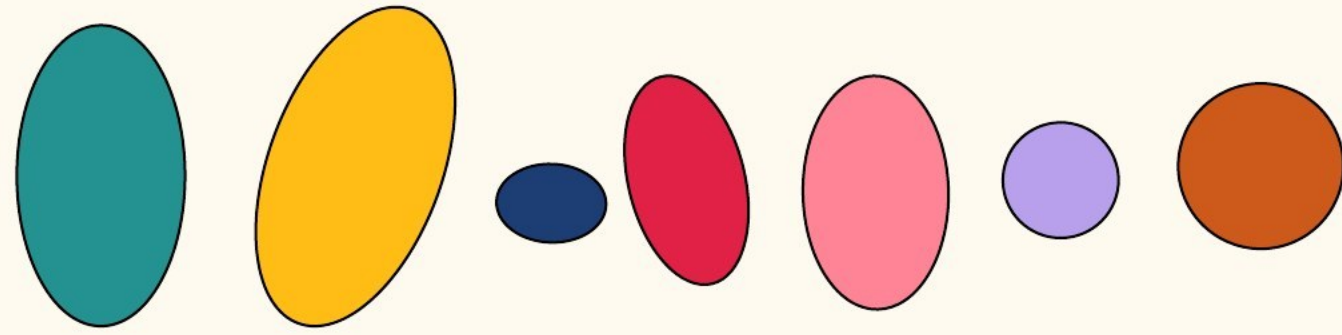


Pick an arbitrary tree structure within sets (model trees, coarse solutions [SY16], [CM23])

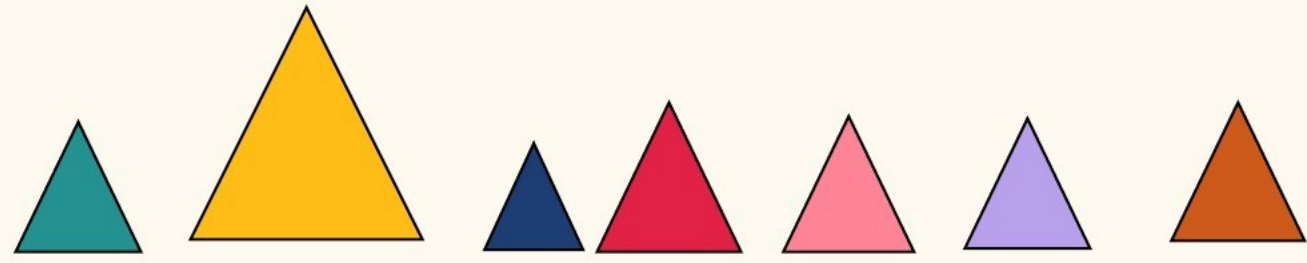


Proof sketch

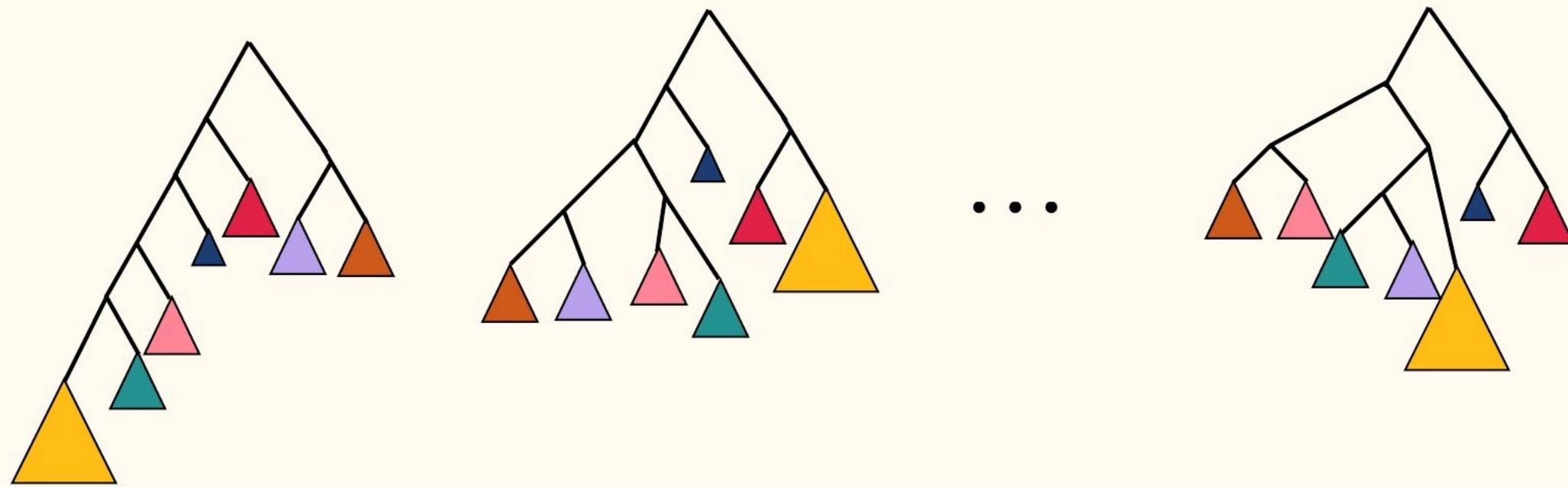
Recover sets $S_1, S_2, \dots, S_{1/\varepsilon}$



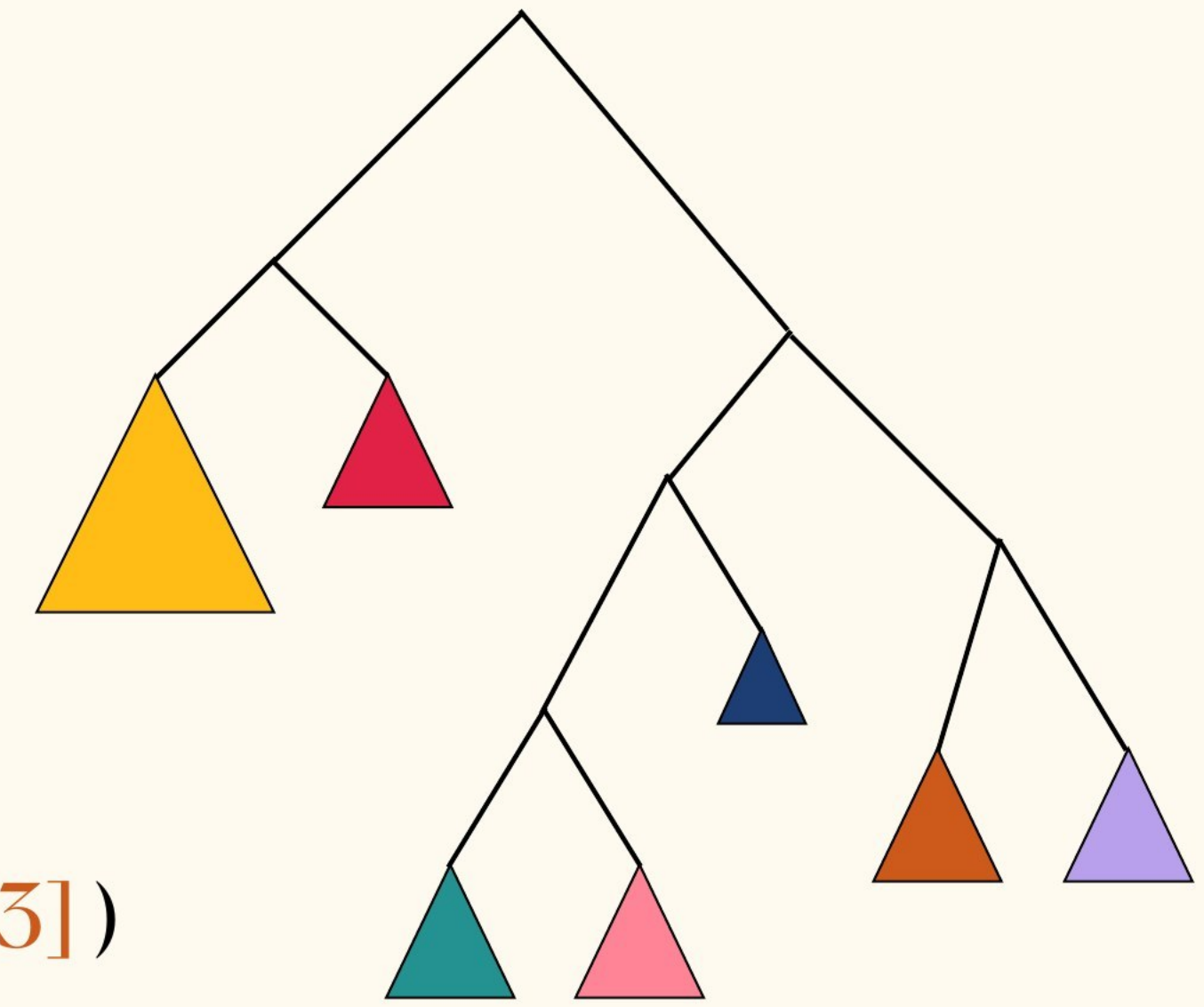
Pick an arbitrary tree structure within sets (model trees, coarse solutions [SY16], [CM23])



Look at all the $(1/\varepsilon)^{1/\varepsilon}$ ways to connect the communities

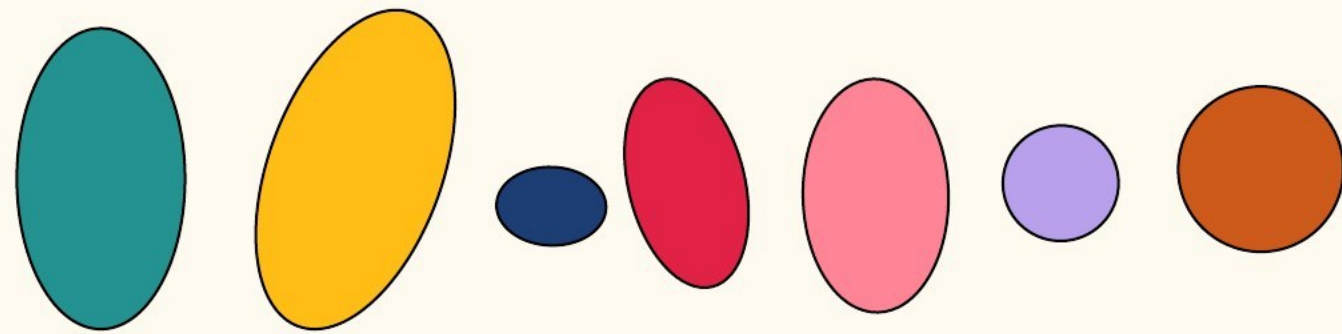


Pick the one that satisfies **most quartets** (ERM)

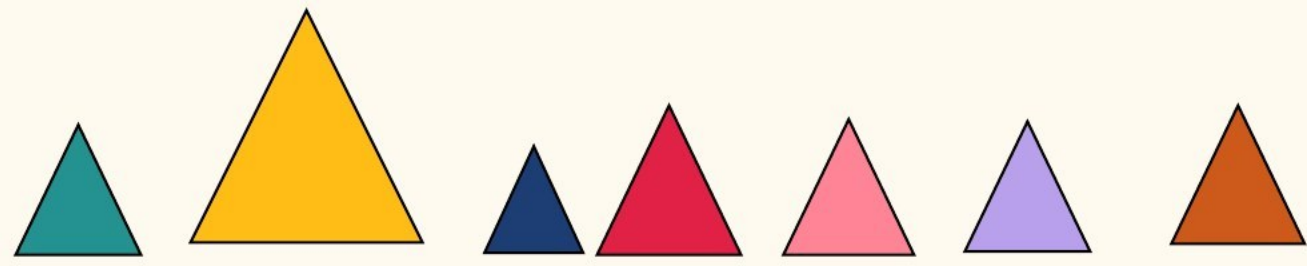


Proof sketch

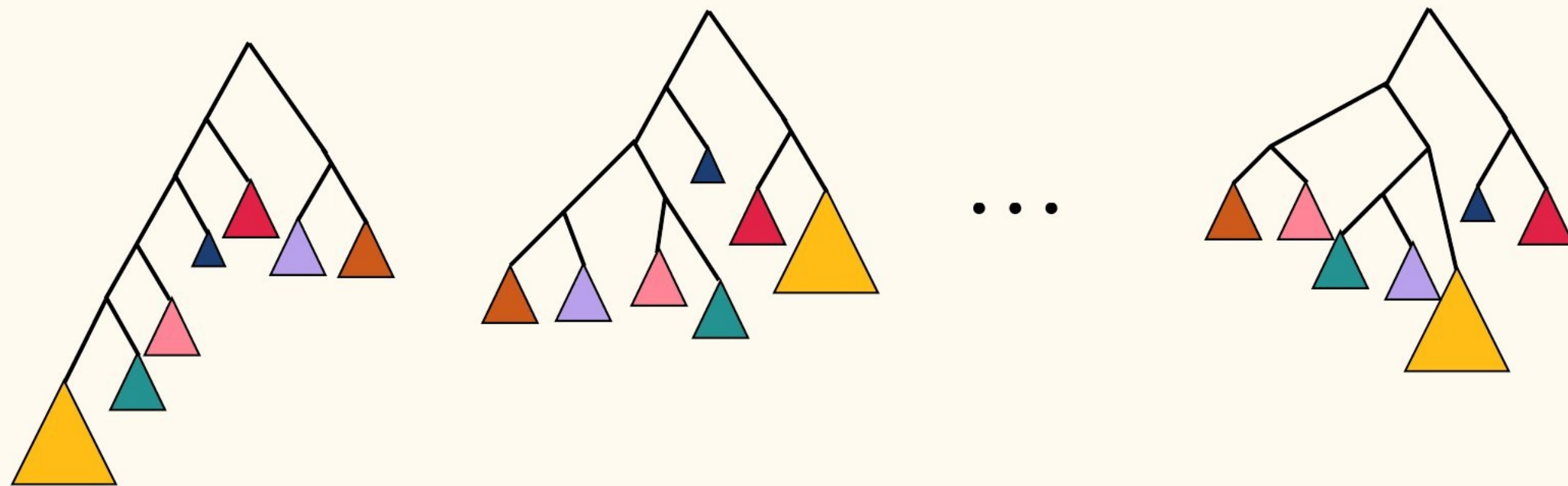
Recover sets $S_1, S_2, \dots, S_{1/\varepsilon}$



Pick an arbitrary tree structure within sets (model trees, coarse solutions [SY16], [CM23])

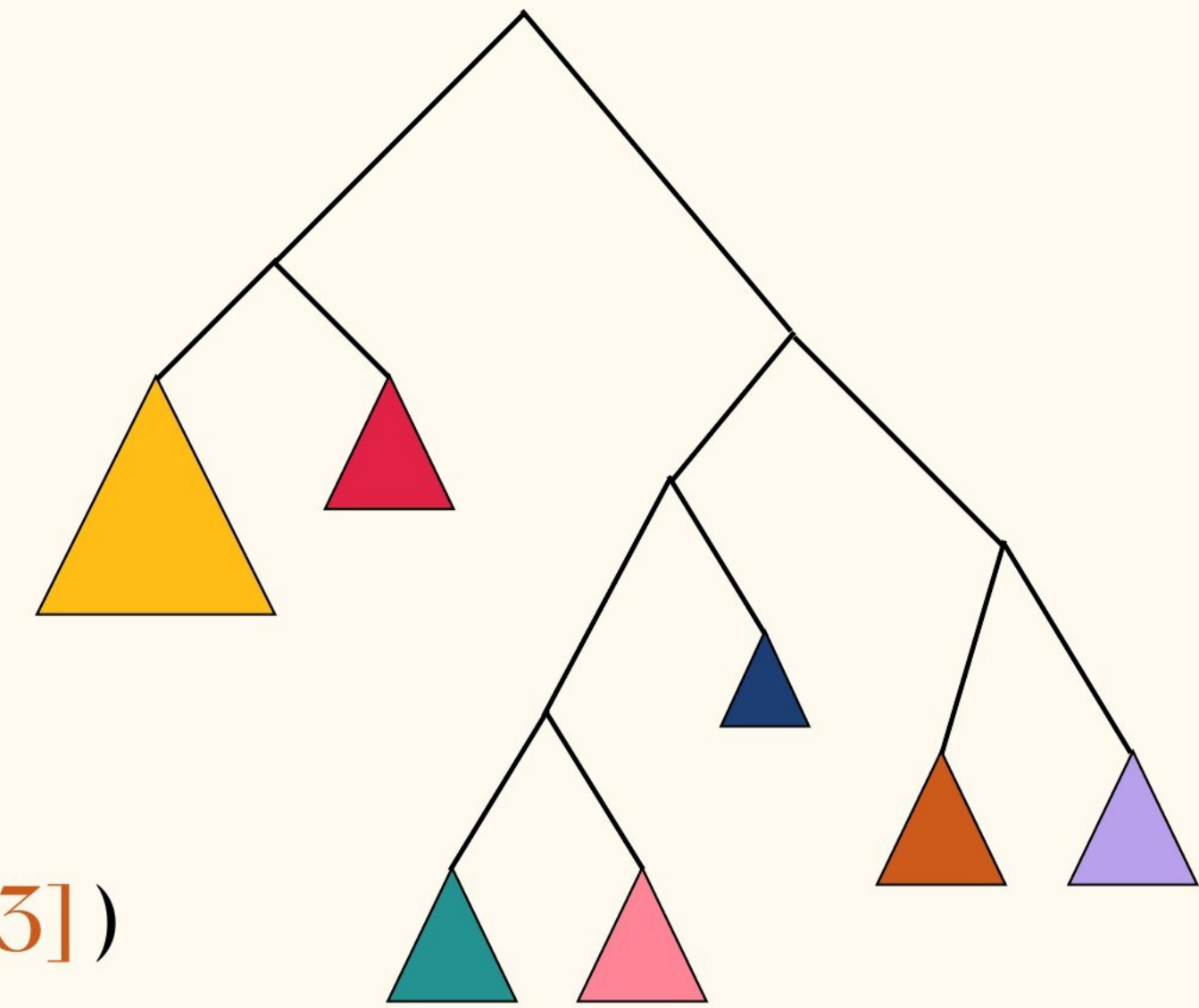


Look at all the $(1/\varepsilon)^{1/\varepsilon}$ ways to connect the communities



Pick the one that satisfies **most quartets** (ERM)

By a *uniform convergence* property it will satisfy $1 - \varepsilon$ of **all quartets**



The sample complexity of tree learning

We have designed an efficient, polynomial time algorithm that given $O_\epsilon(n)$ noisy uniform quartets finds a tree that is close to the ground truth.

Q: What about other distributions?

We show that $O_{\epsilon,\delta}(n)$ samples suffice, information-theoretically, to **learn the tree** in a distribution-free PAC model

- Bound on the Natarajan dimension of quartets (analog of VC for multi-class classification)
- Linearly many samples suffice to learn a complicated structure

The sample complexity of tree learning

We have designed an efficient, polynomial time algorithm that given $O_\epsilon(n)$ noisy uniform quartets finds a tree that is close to the ground truth.

Q: What about other distributions?

We show that $O_{\epsilon,\delta}(n)$ samples suffice, information-theoretically, to **learn the tree** in a distribution-free PAC model

- Bound on the Natarajan dimension of quartets (analog of VC for multi-class classification)
- Linearly many samples suffice to learn a complicated structure

Thank you!