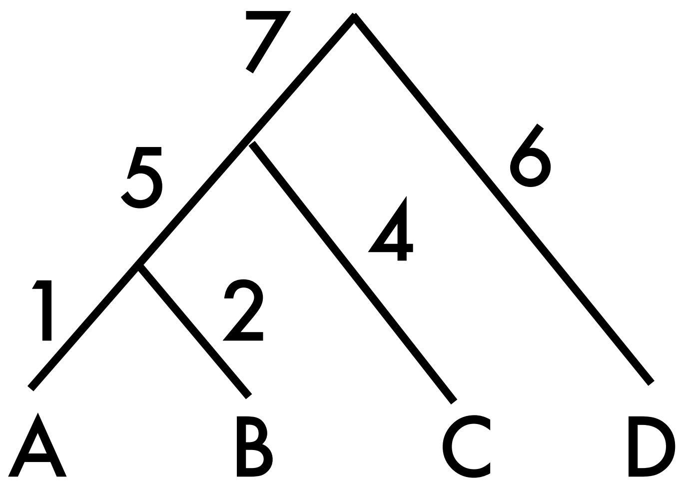
Averaging Metric Trees

Ezra Miller	Duke
Megan Owen	NCSU/SAMSI
Scott Provan	UNC

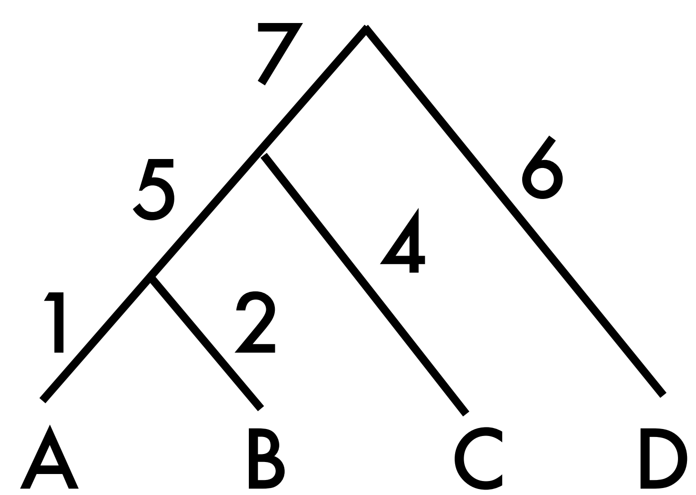
Phylogenetic Trees

- a metric tree:

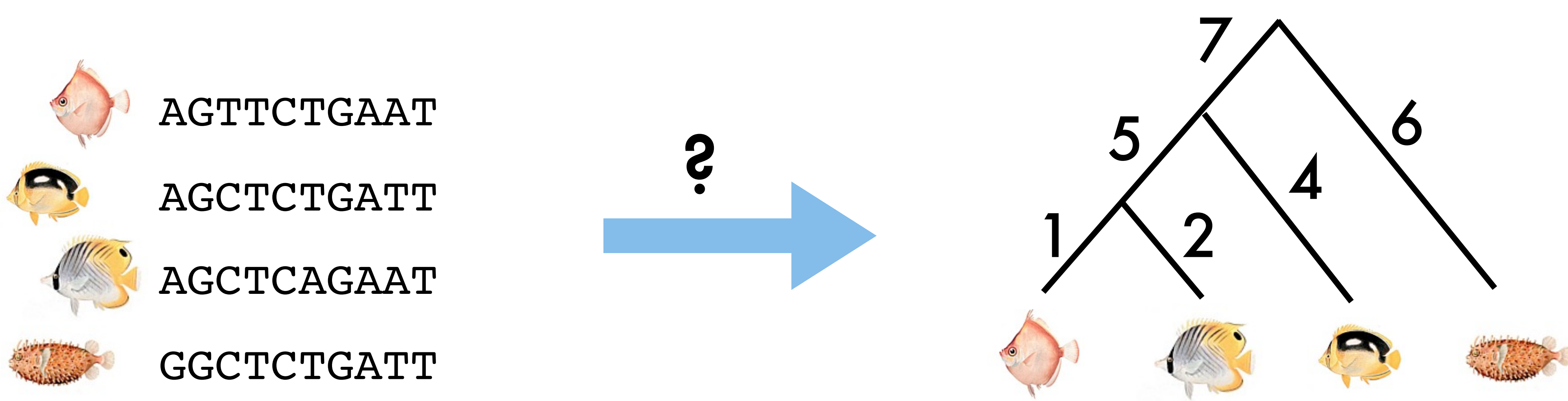


Phylogenetic Trees

- a metric tree:

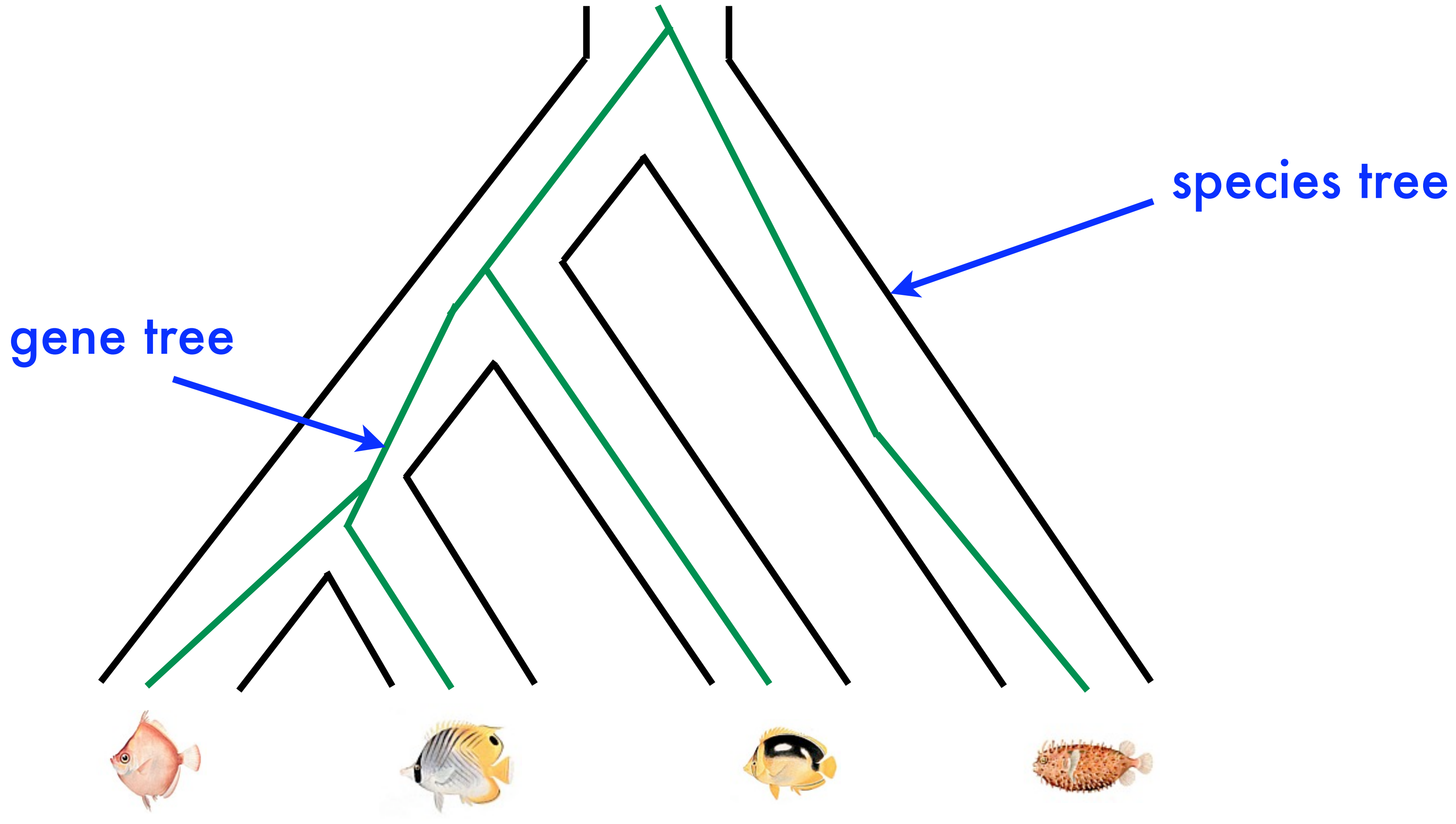


- a phylogenetic tree:



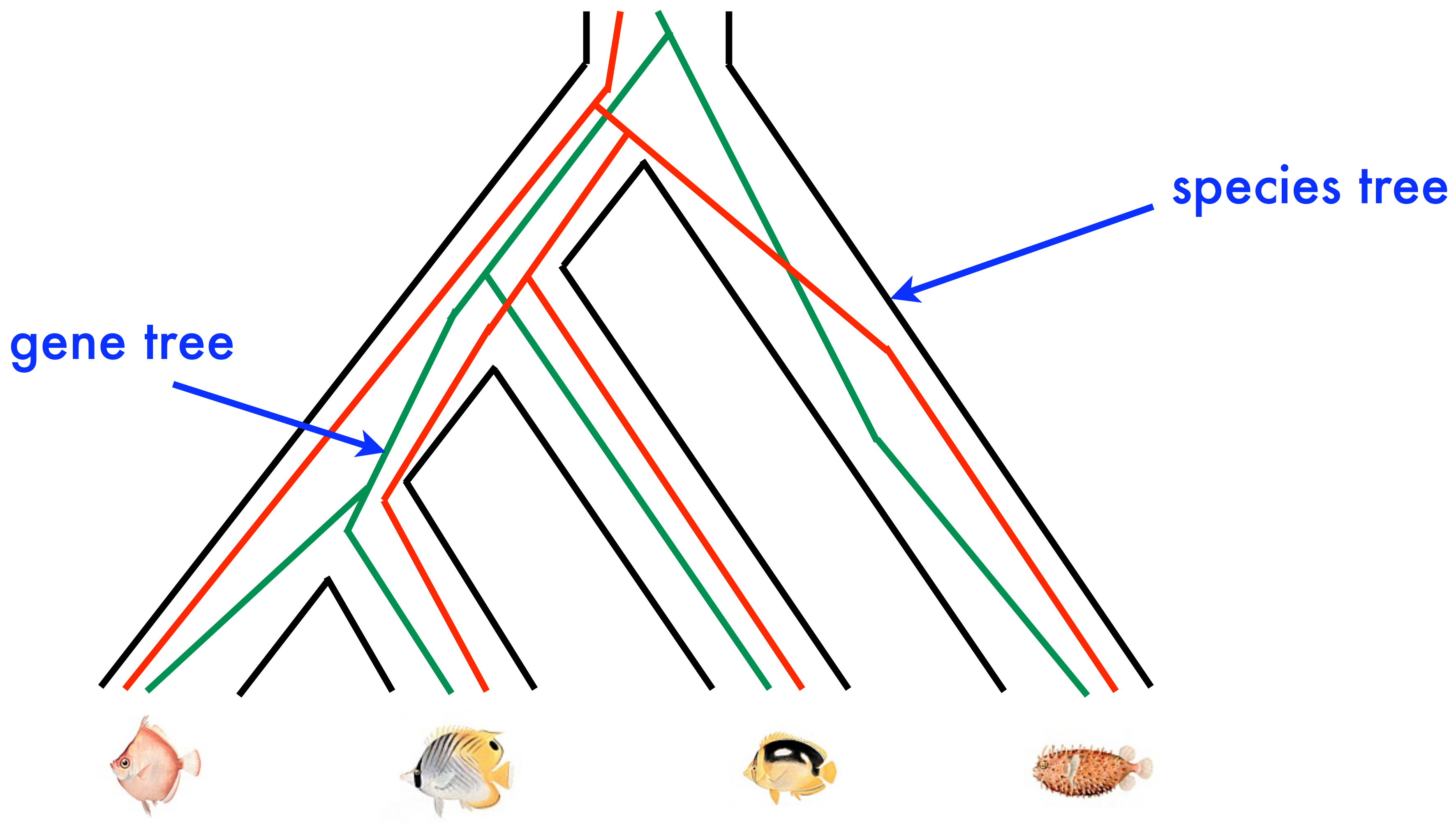
Gene and Species Trees

- want species trees, but DNA gives us gene trees



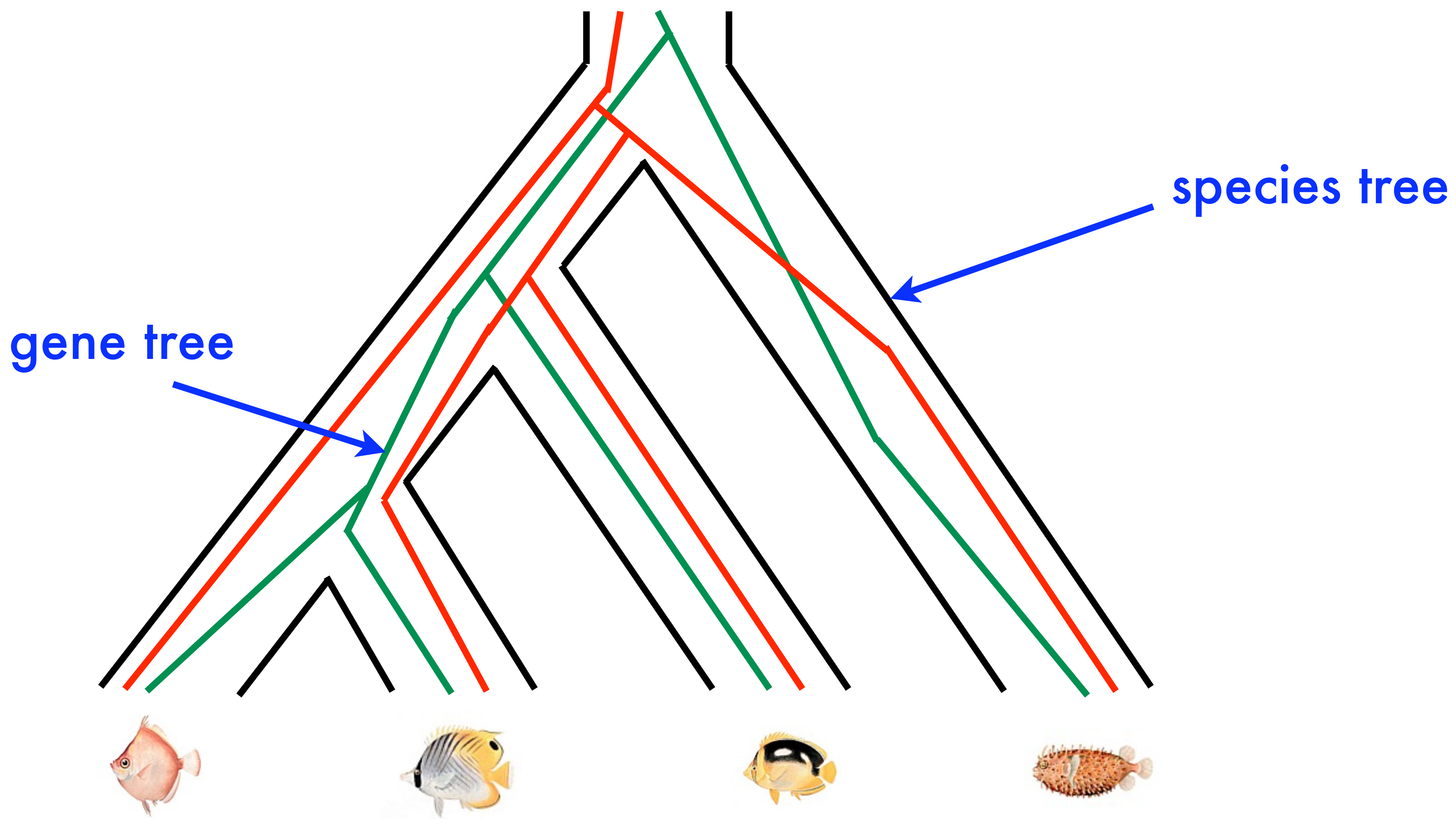
Gene and Species Trees

- want species trees, but DNA gives us gene trees



Gene and Species Trees

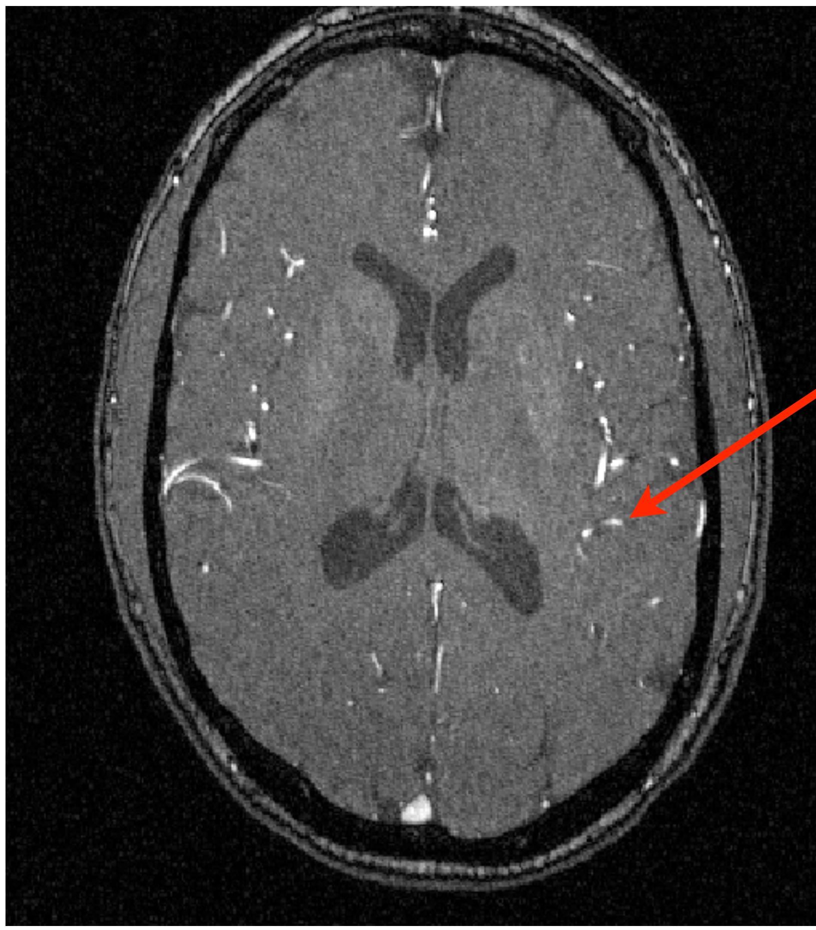
- want species trees, but DNA gives us gene trees



- average of gene trees = species tree ?

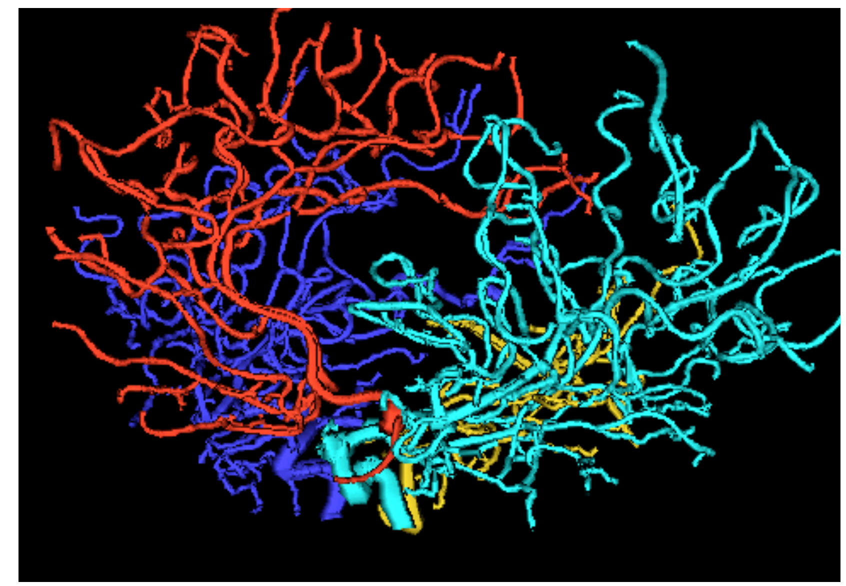
Comparing brains

With Steve Marron, Ipek Oguz , Scott Provan, Martin Styner (all UNC)



blood vessel

MRI scans ⟶



tree representing arteries in a brain

- how do we compare trees to determine changes in brain due to aging or disease?
  - moving average

Figures from Steve Marron

Goal

- goal:
  - compute a meaningful average of a set of metric trees
- metric tree parameters:
  - tree topology
  - edge lengths
- so not a standard statistical problem!

Tree Space Framework

continuous, polyhedral space of phylogenetic trees

- *Geometry of the space of phylogenetic trees*, Billera, Holmes, and Vogtmann, 2001.

= tree complex

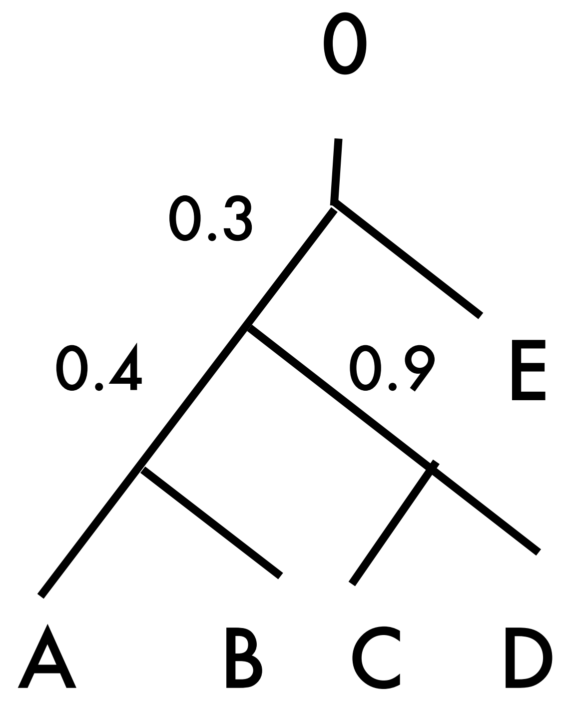
- *Shellability of complexes of trees*, Trappmann and Ziegler, 1998.
- *The tree representation of σ_{n+1}*, Robinson and Whitehouse, 1996.

+ metric (geodesic distance)

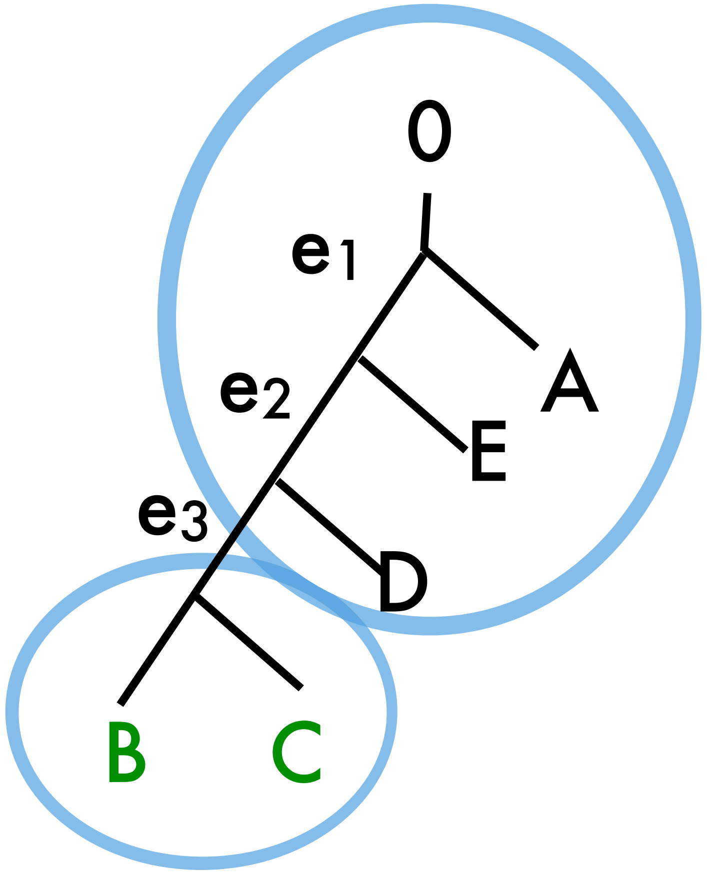
- computable in polynomial time (Owen and Provan, 2009)

Tree Space $\mathbb{T}_n$

- trees in $\mathbb{T}_n$ have:
  - n leaves
  - interior edges with lengths ≥0



Splits



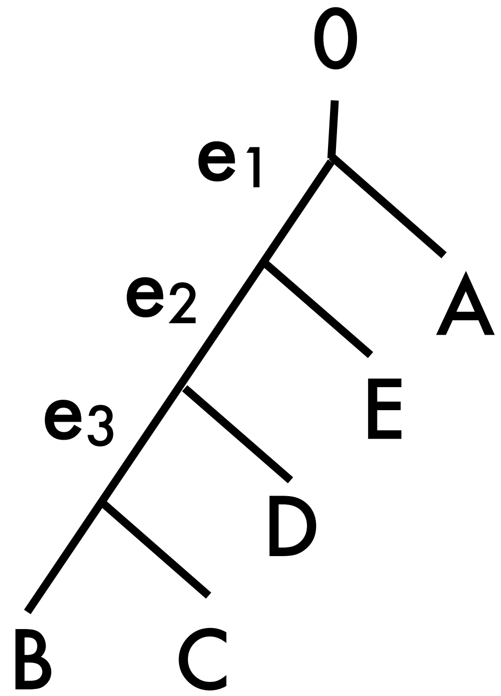
- each interior edge induces a *split*
- a *split* is a partition of the set of leaves plus the root 0:

$$e_3 = \{ \{B,C\}, \{0,A,E,D\} \}$$

$$\text{or} \quad e_3 = BC \mid 0AED$$

Split Compatibility

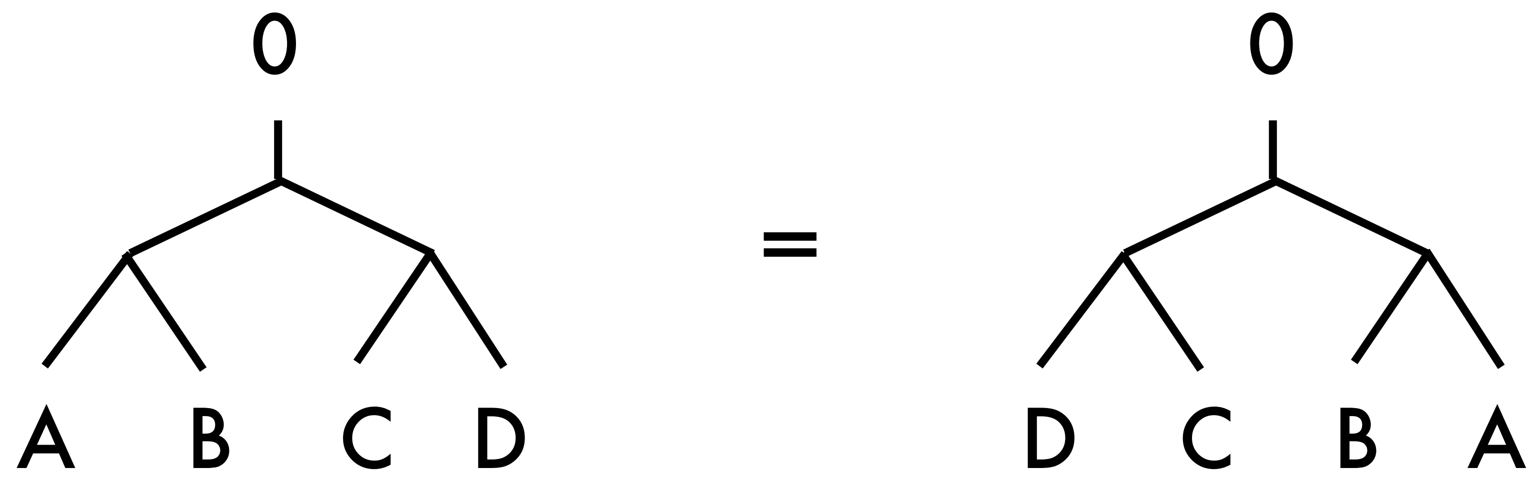
- $e_x = X|X'$ is *compatible* with $e_y = Y|Y'$ if there exists a tree containing both splits



ex. e_3 = BC | 0AED is compatible
with e_2 = BCD | 0AE
but not with f = AB | 0CDE

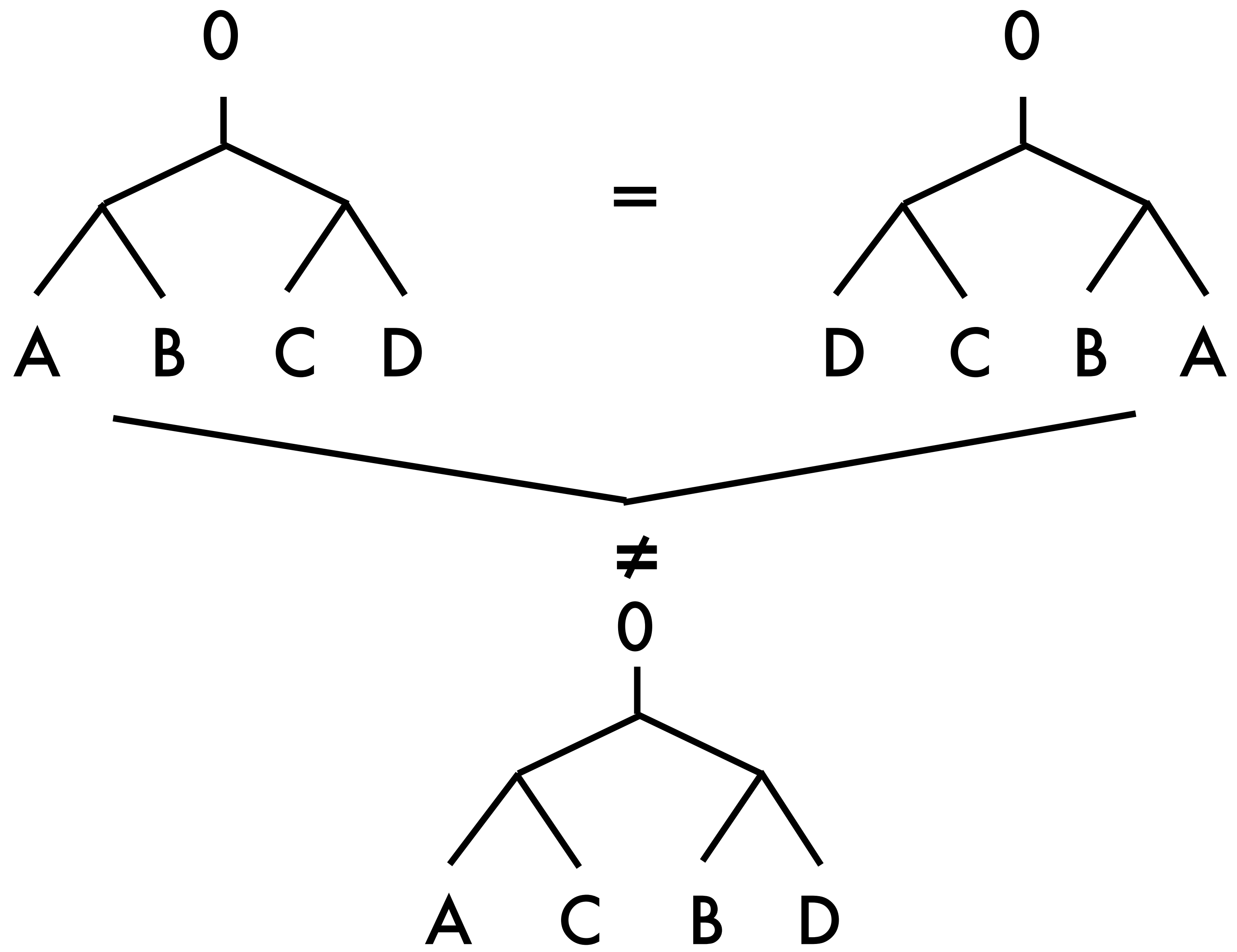
The trees

- embedding in plane irrelevant

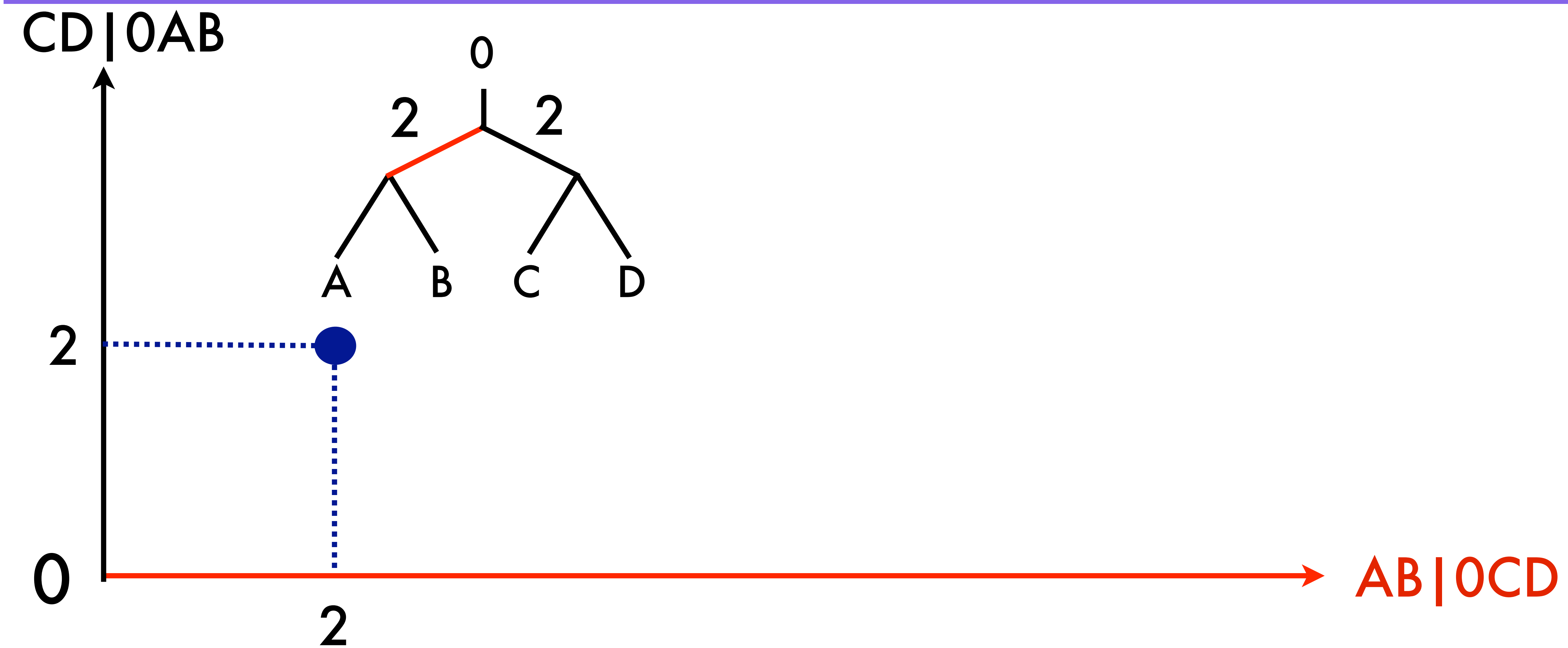


The trees

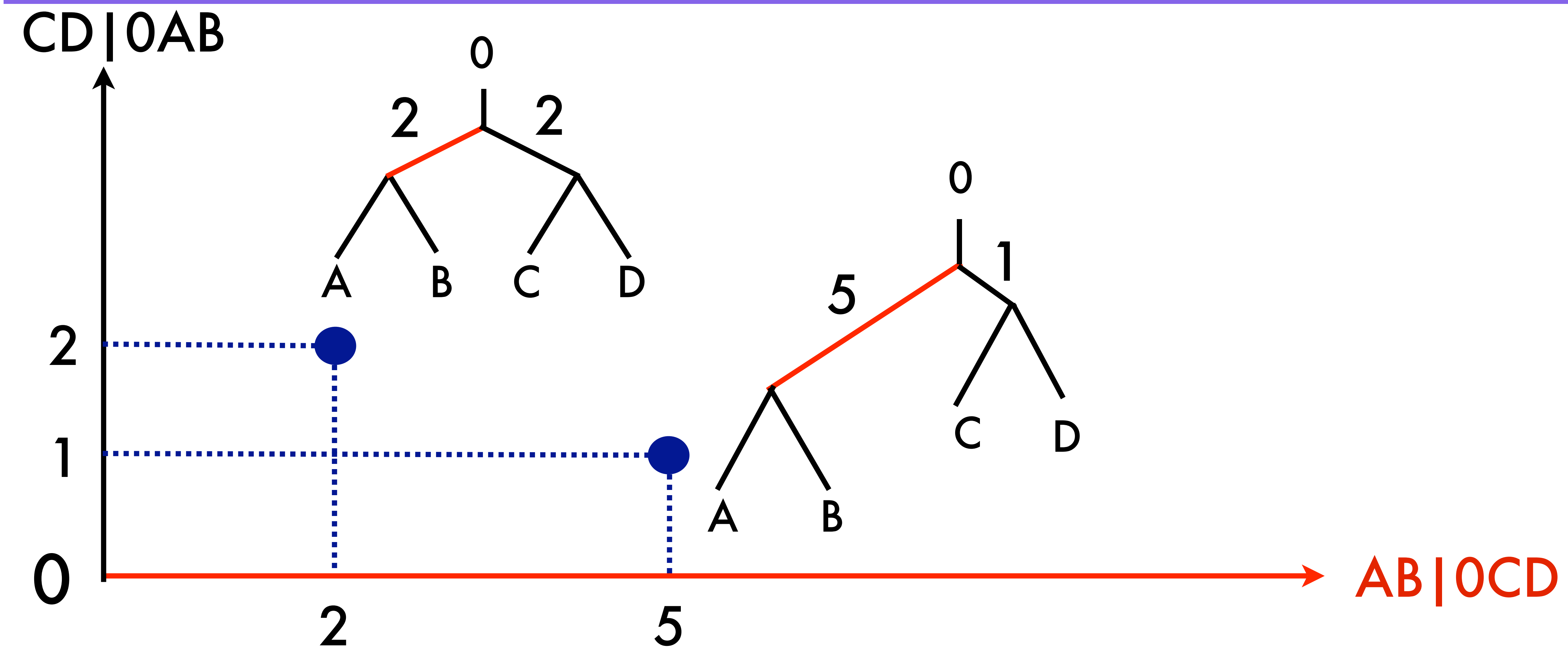
- embedding in plane irrelevant



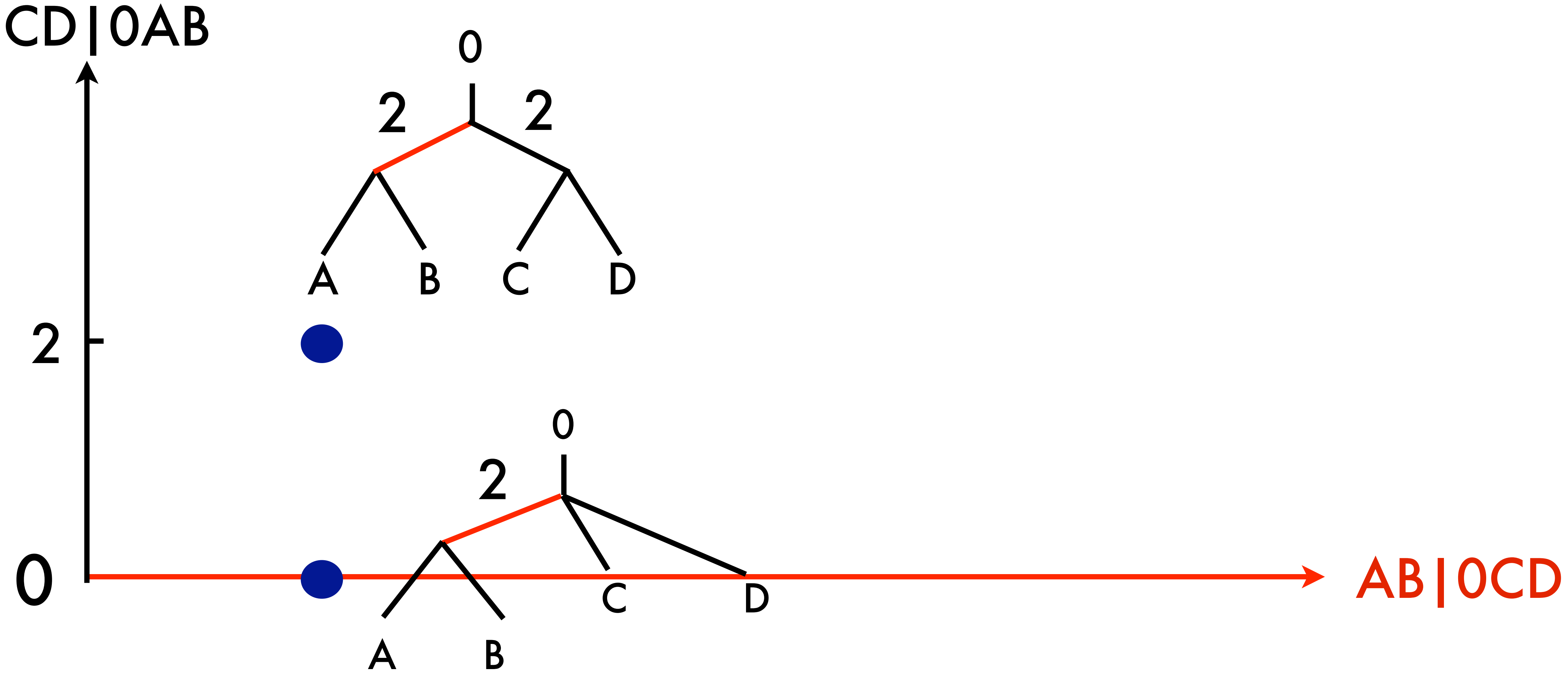
Orthants



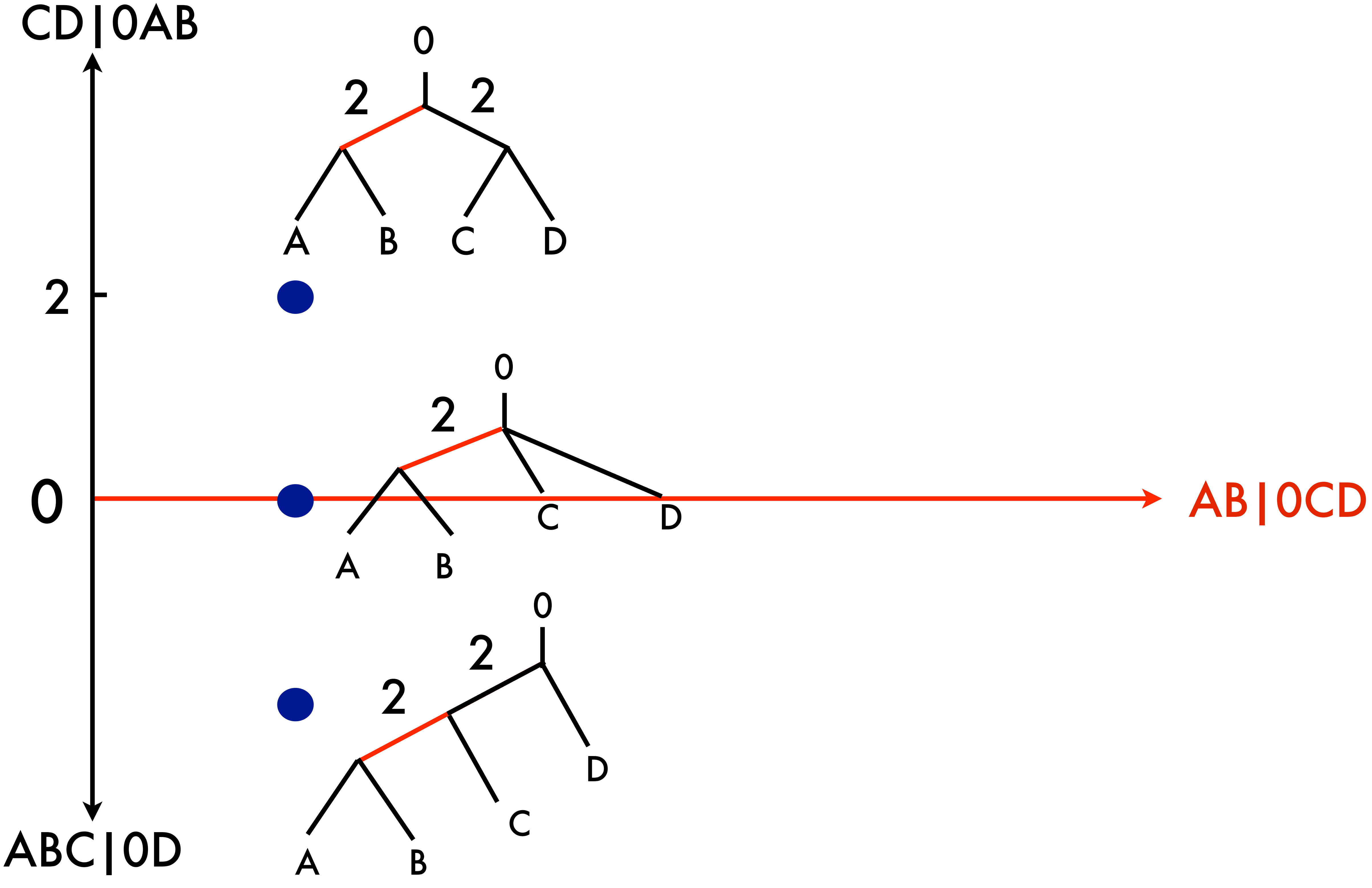
Orthants



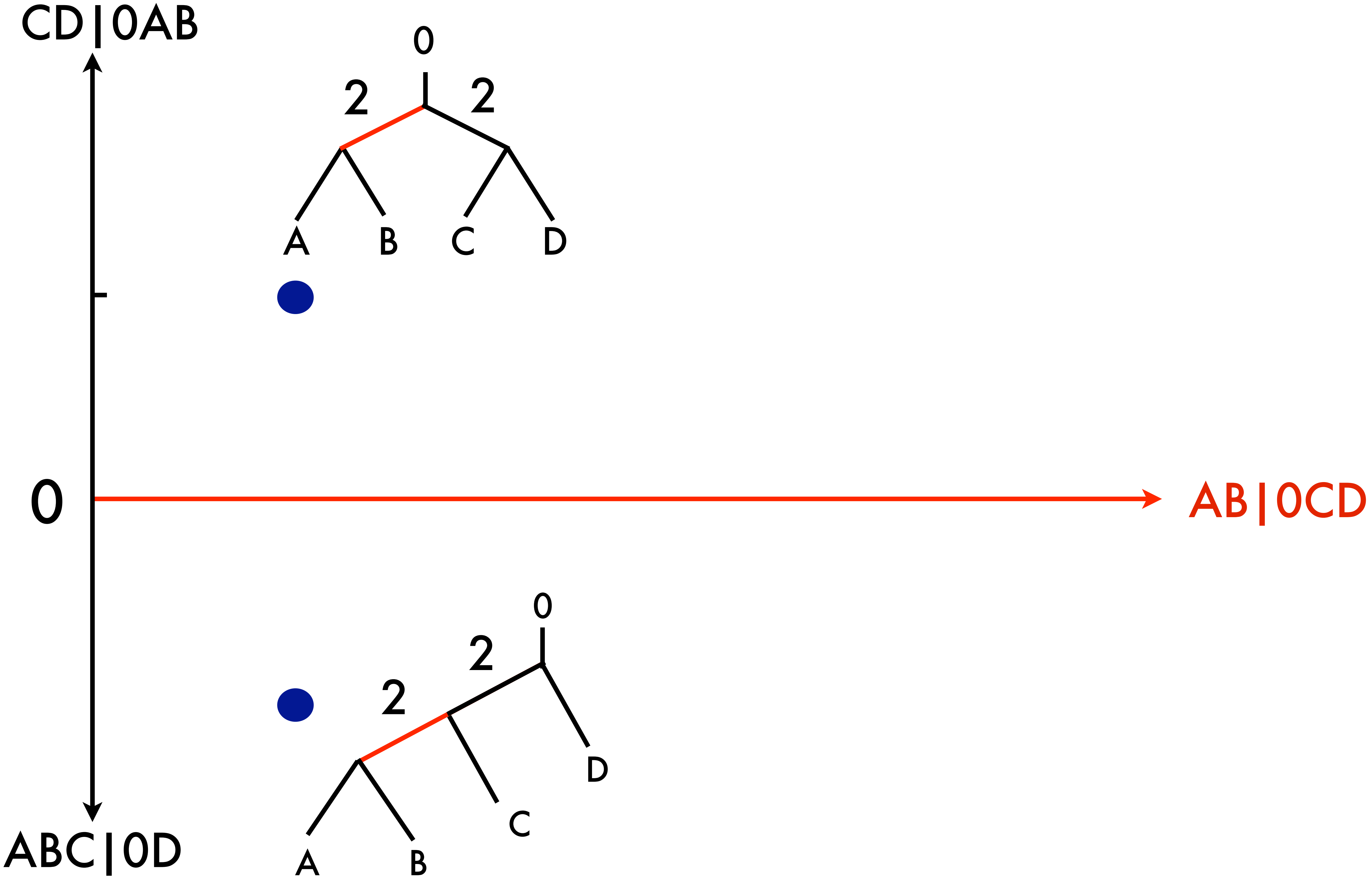
Orthants



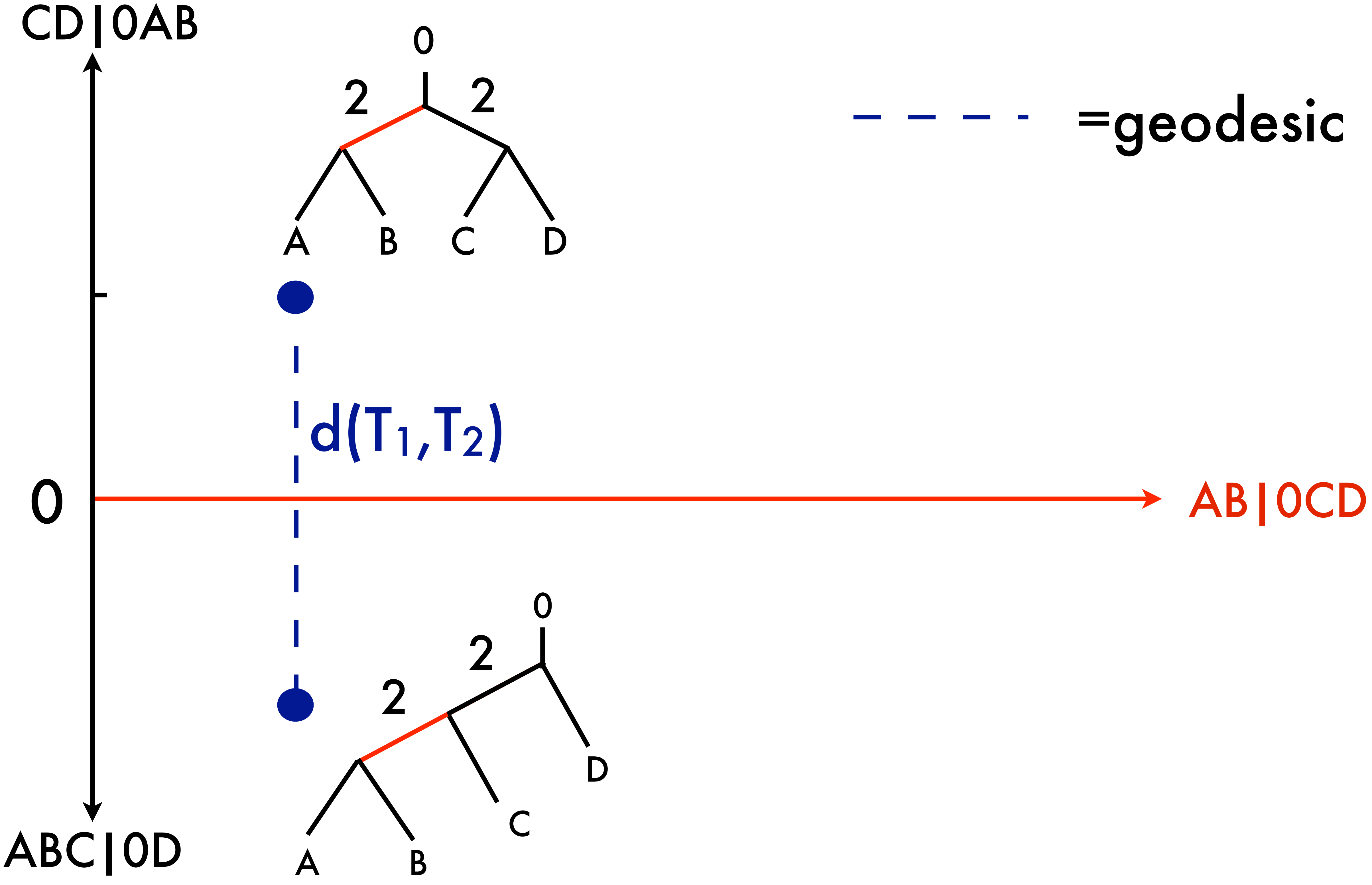
Orthants



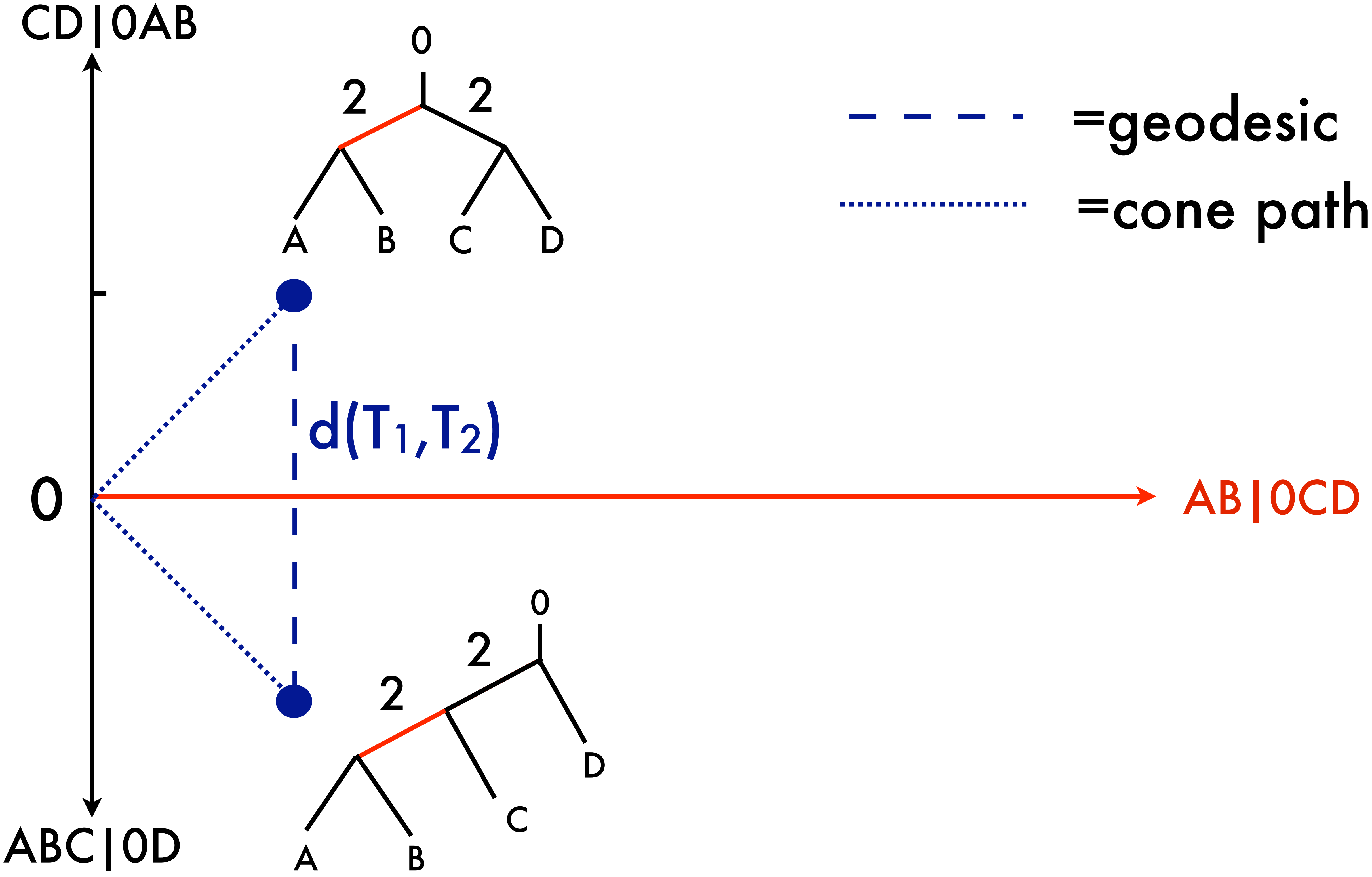
Orthants



Orthants

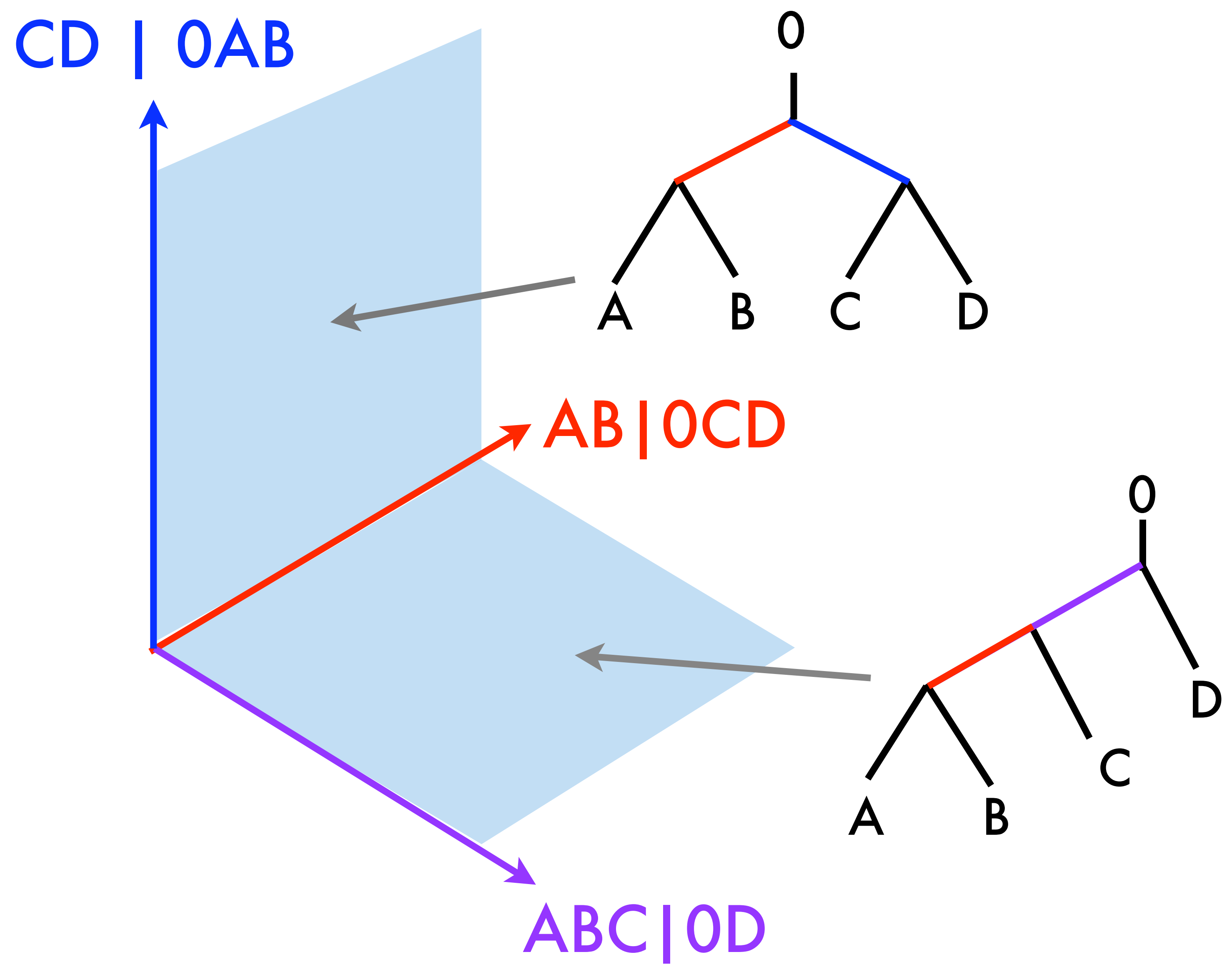


Orthants

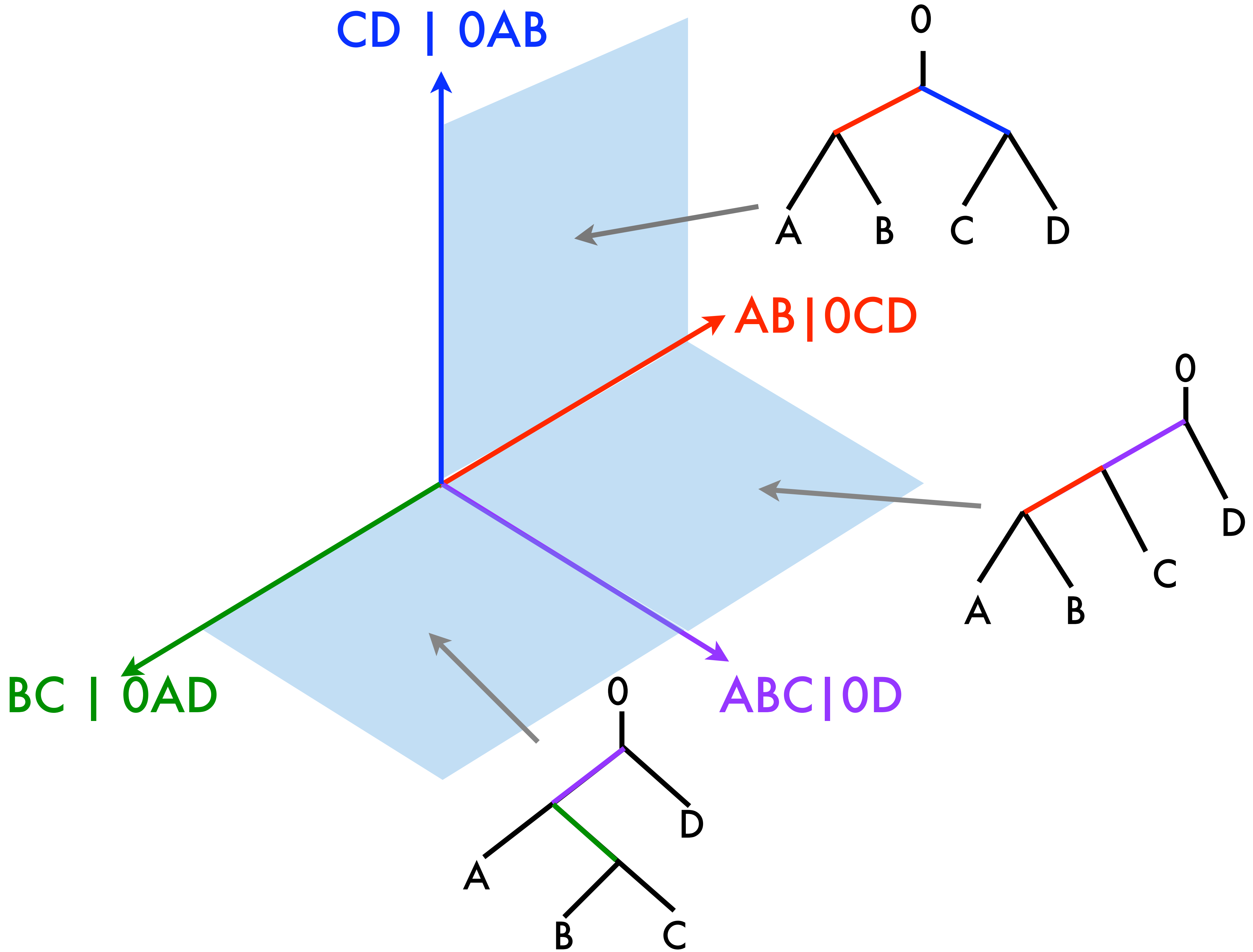


CD|0AB
0
2
2
A
B
C
D
=geodesic
=cone path
d(T1,T2)
0
AB|0CD
0
2
2
A
B
C
D
ABC|0D

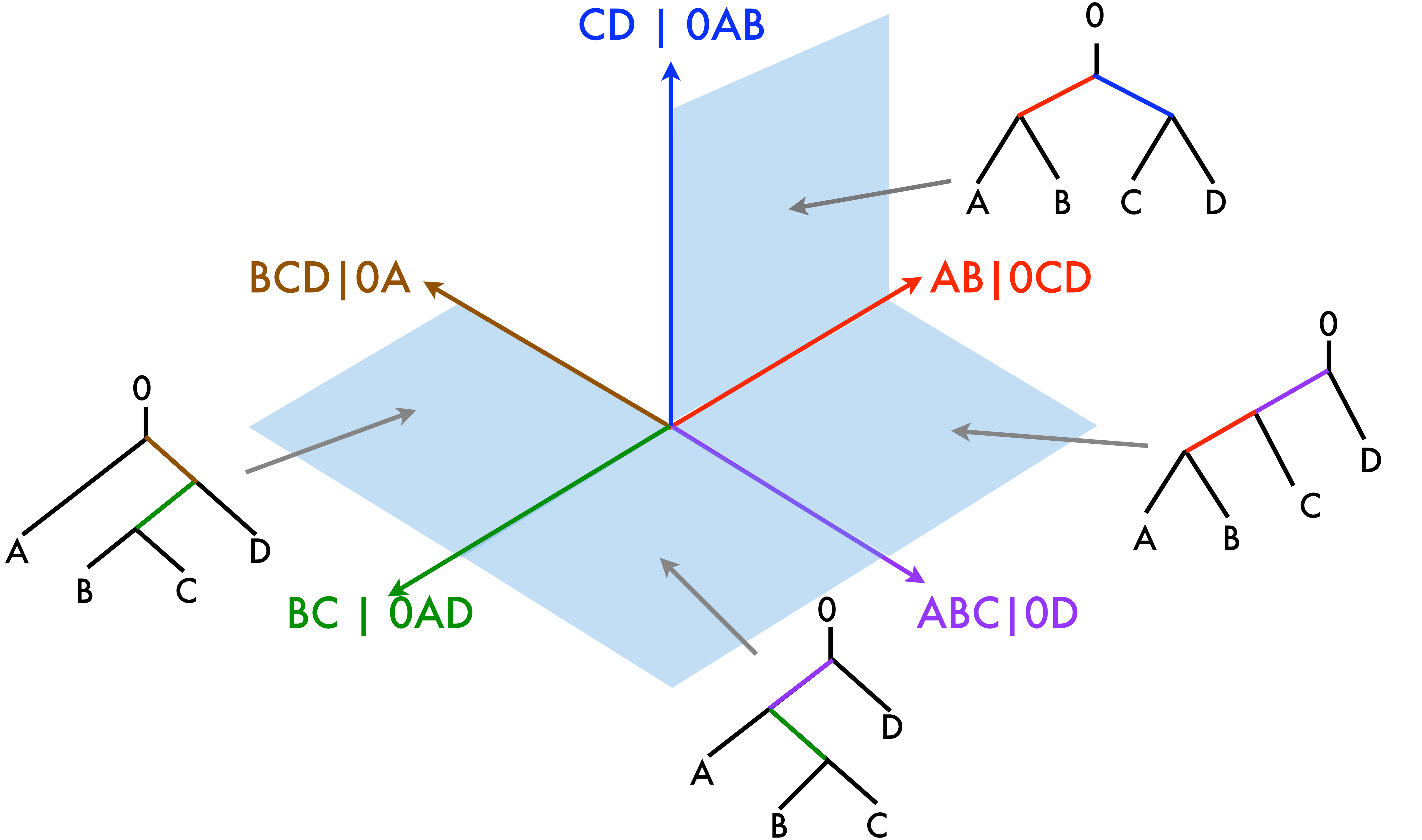
Structure of $\mathbb{T}_4$



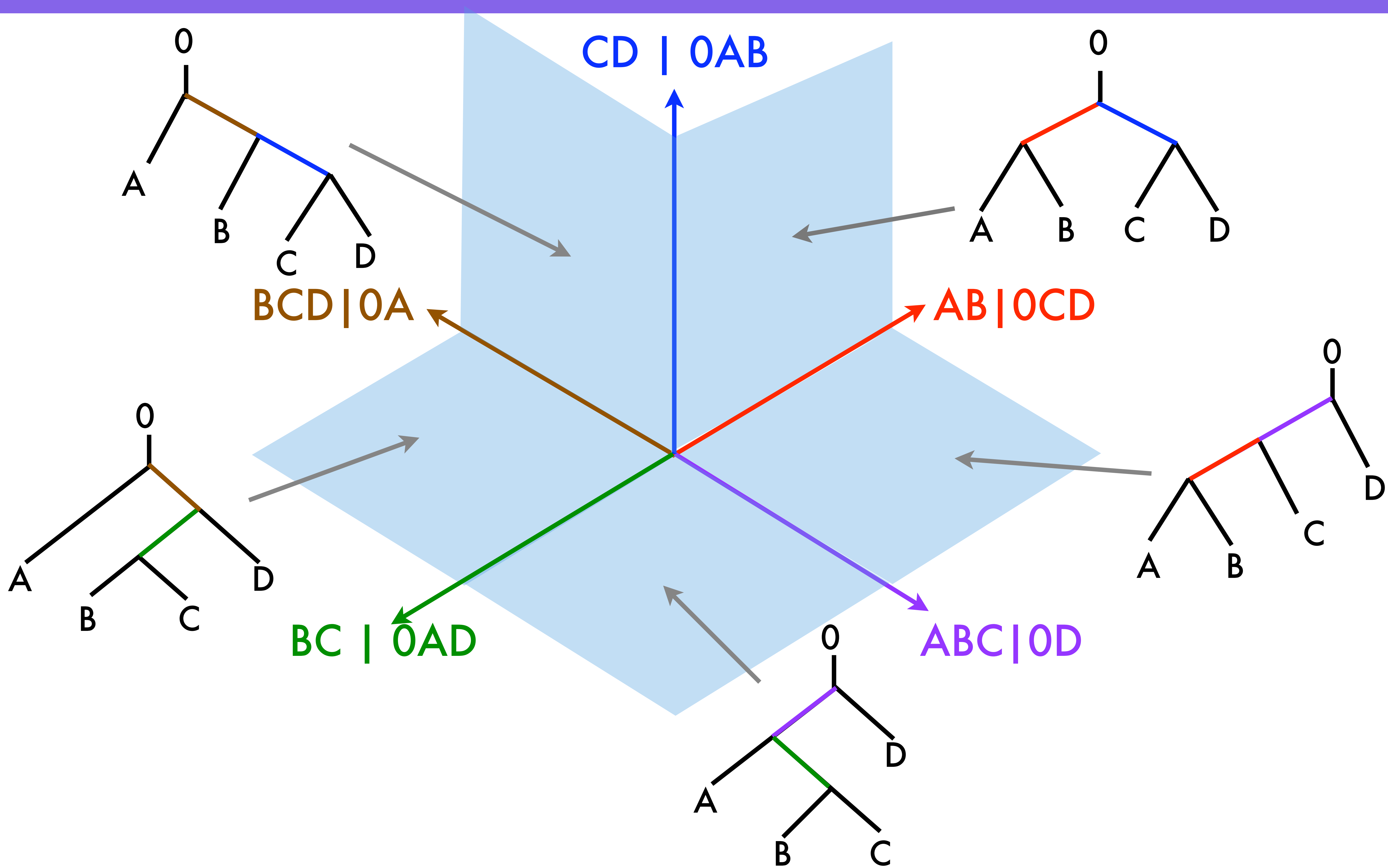
Structure of $\mathbb{T}_4$



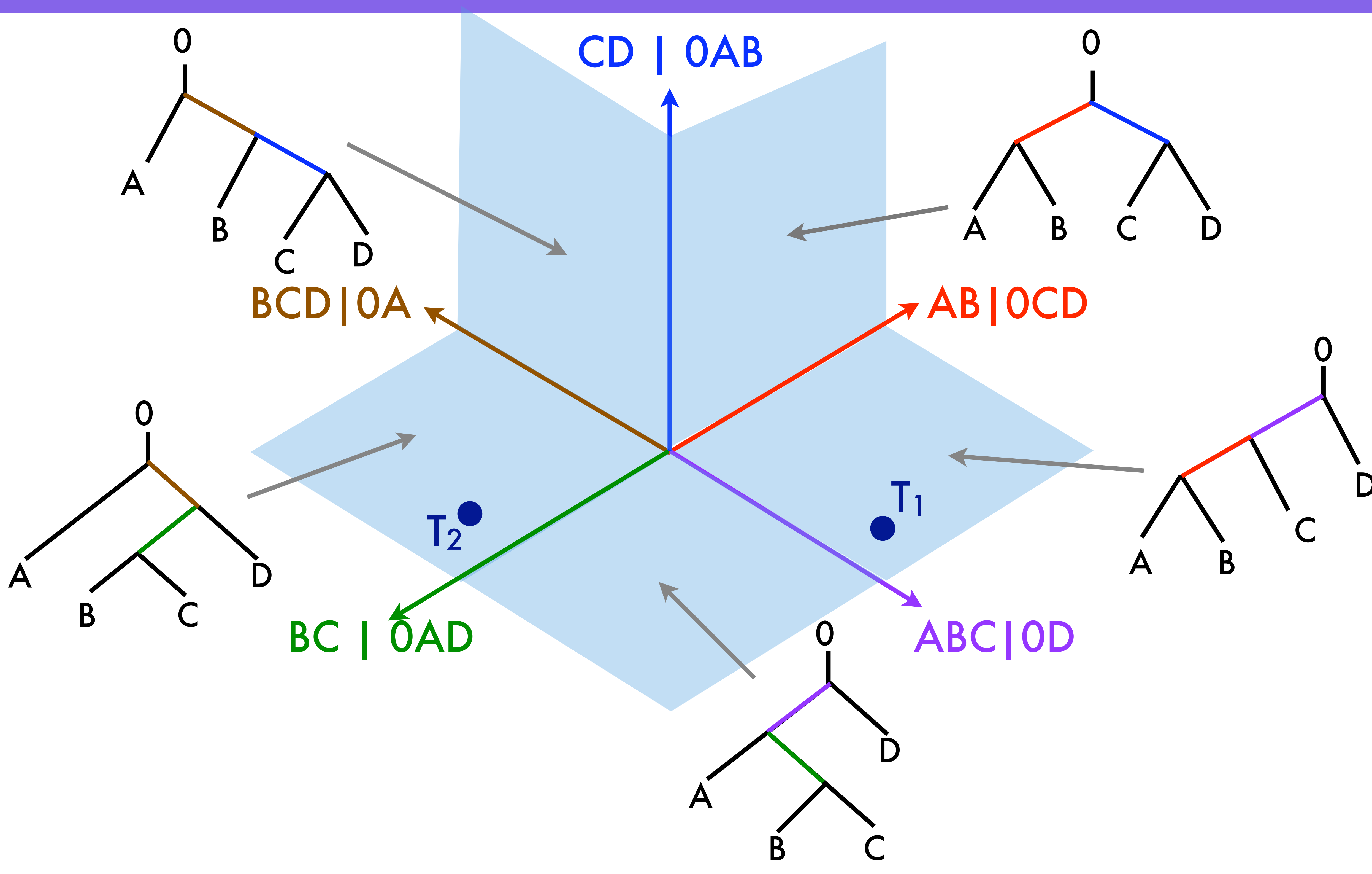
Structure of $\mathbb{T}_4$



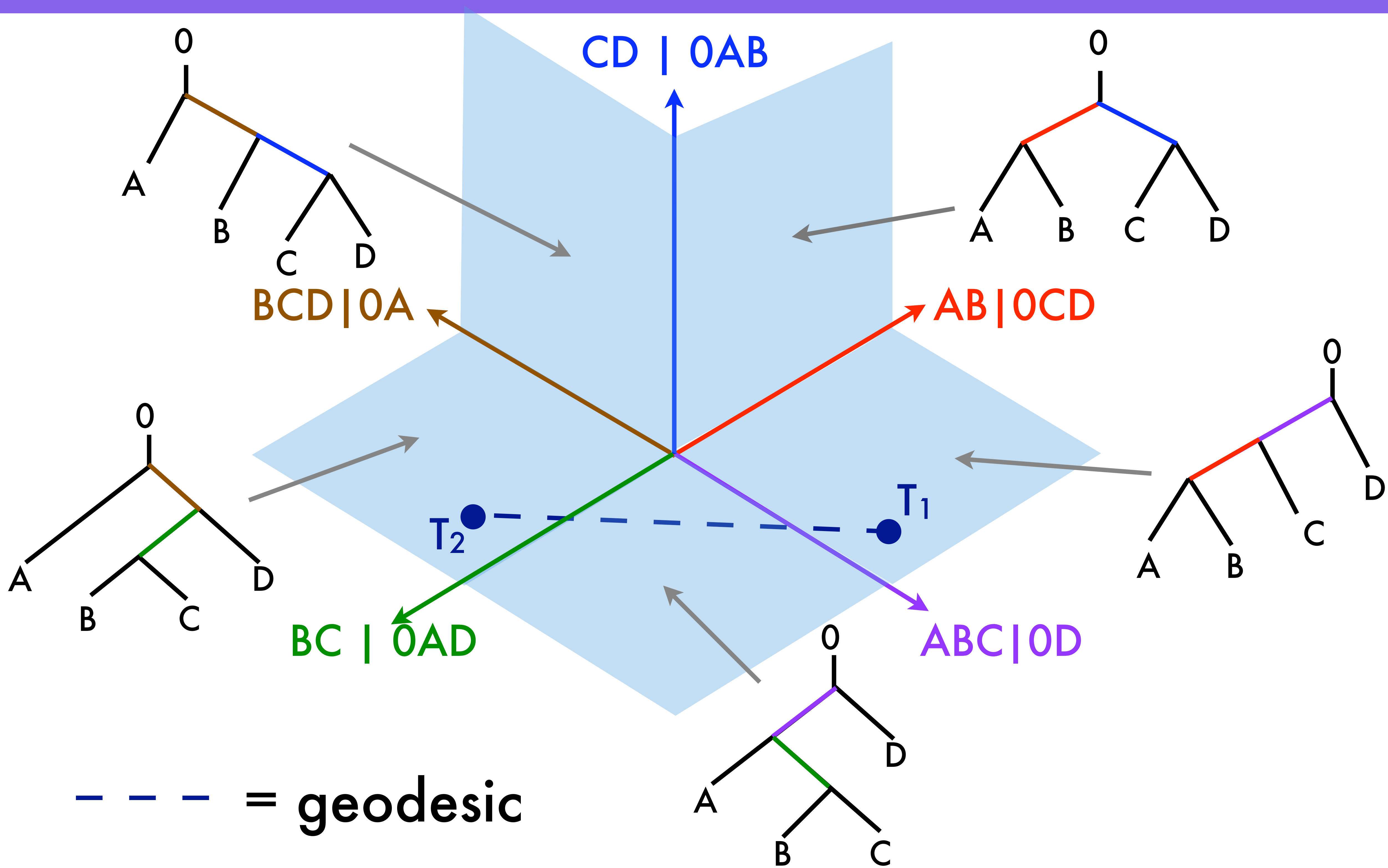
Structure of $\mathbb{T}_4$



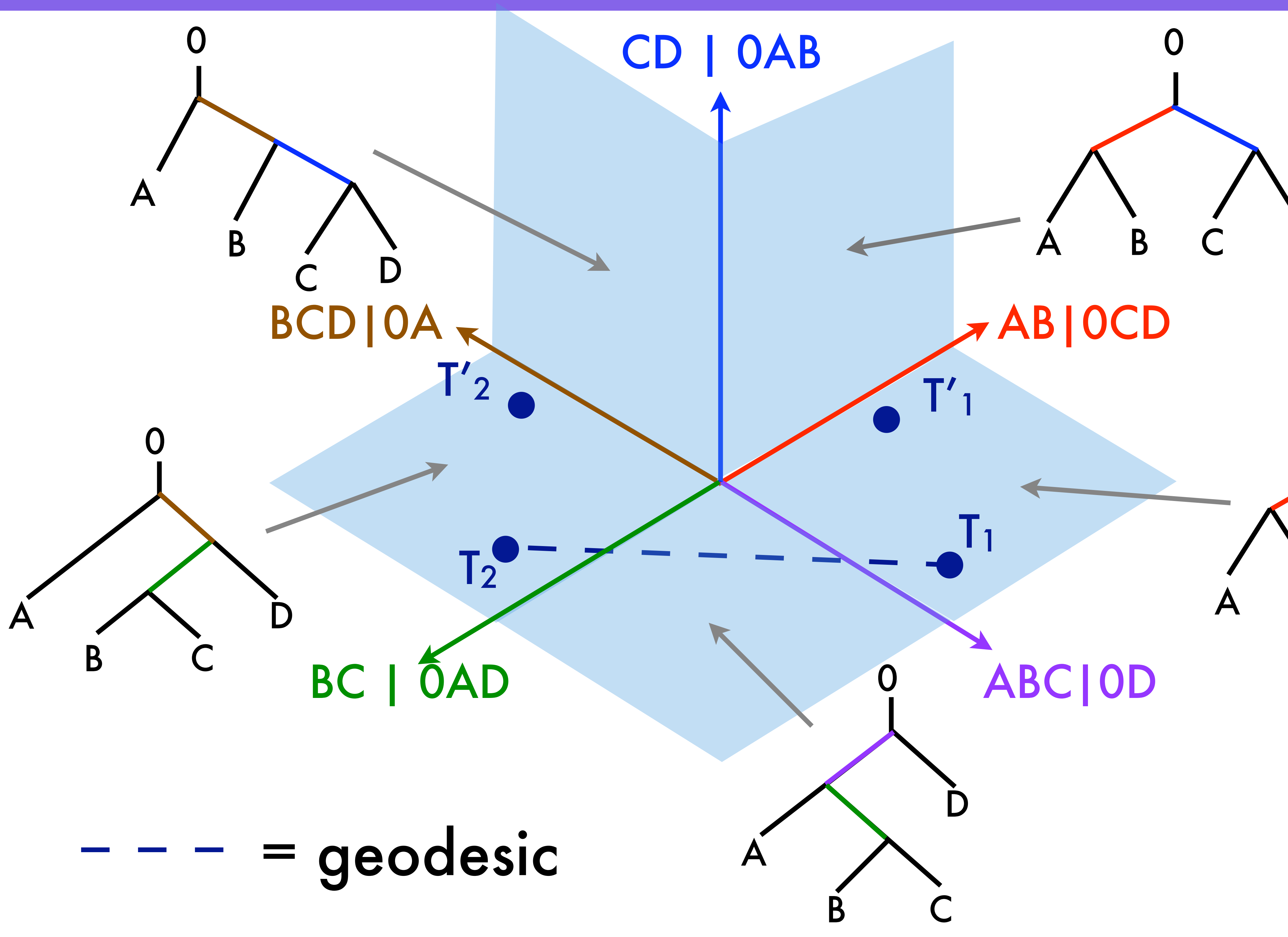
Structure of $\mathbb{T}_4$



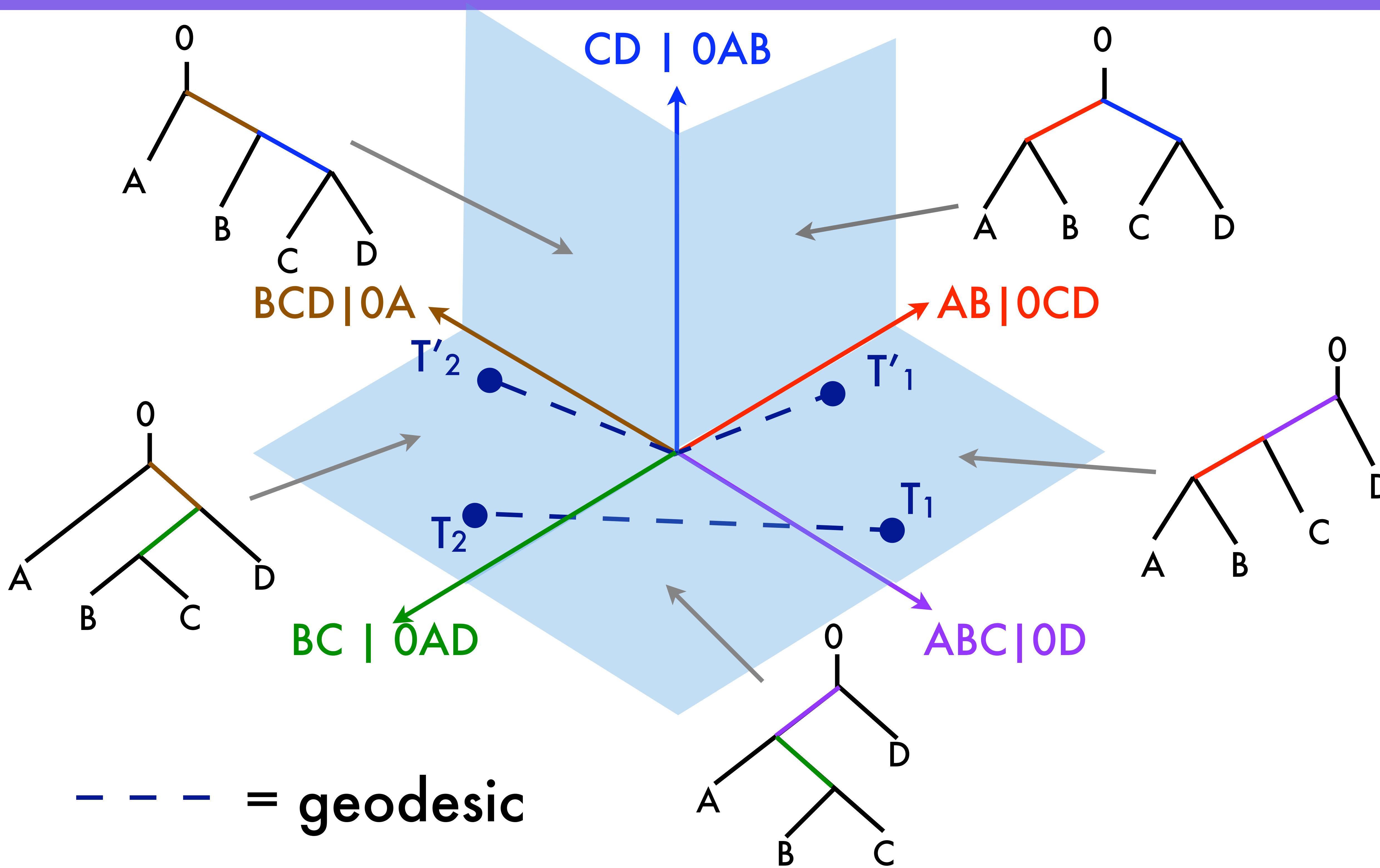
Structure of $\mathbb{T}_4$



Structure of $\mathbb{T}_4$



Structure of $\mathbb{T}_4$



$\mathbb{T}_n$ is CAT(0)

- CAT(0) space (non-positively curved)

  ⇒ unique geodesic (shortest path between two points)

  ⇒ well-defined mid-point tree

- geodesic distance = length of geodesic between two trees T_1 and T_2, in
  - computable in polynomial time $O(n^4)$ (Owen and Provan, 2010)

Average or Mean Trees

- mean tree

  = center of mass of given set of trees

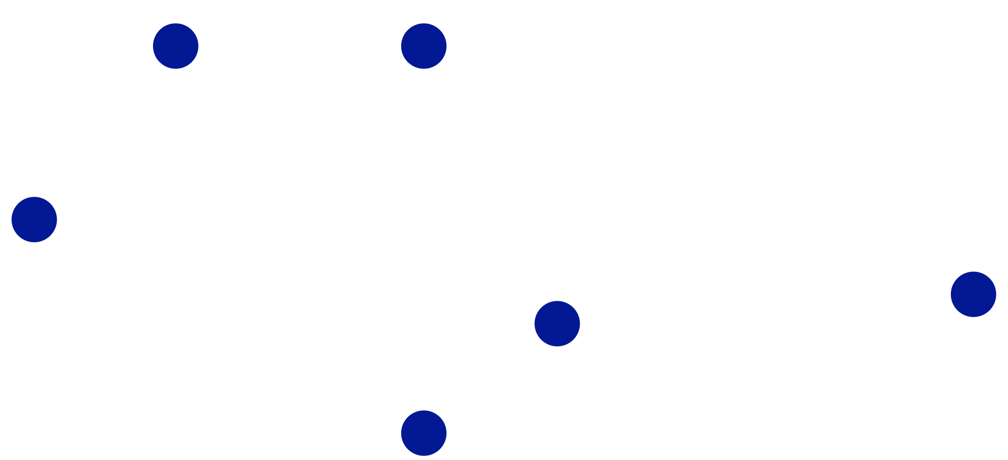
  = tree T′ minimizing sum of square geodesic distances from T′ to each tree in a given set $\mathcal{T}$

$$\text{mean tree} = \underset{T'}{\operatorname{argmin}} \sum_{T \in \mathcal{T}} d(T, T')^2$$

Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)

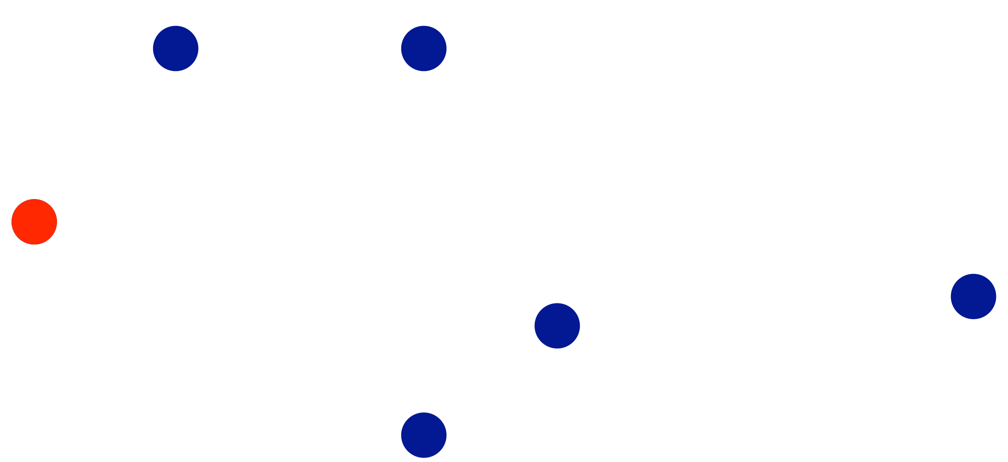


Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)

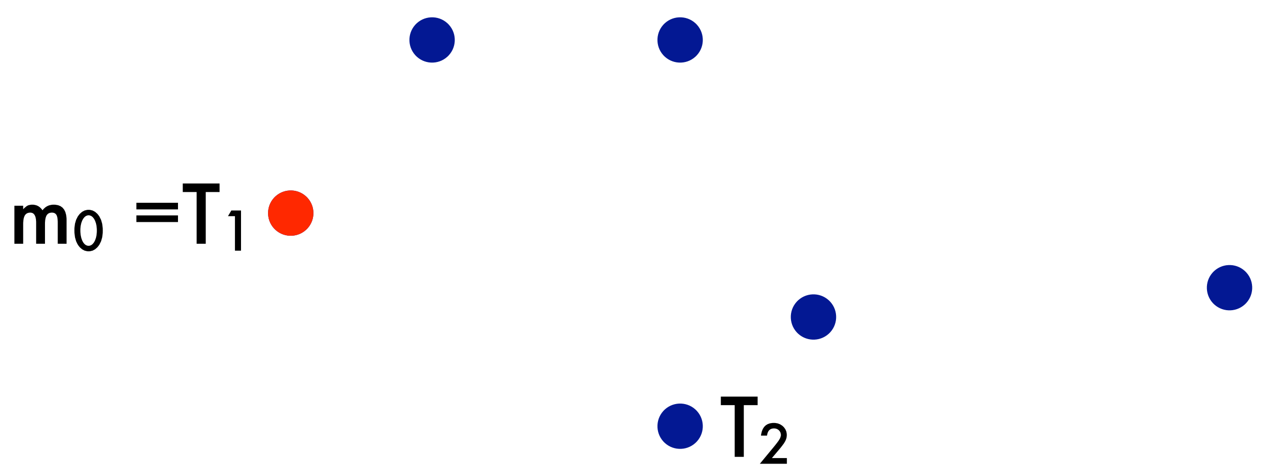
$m_0 = T_1$



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

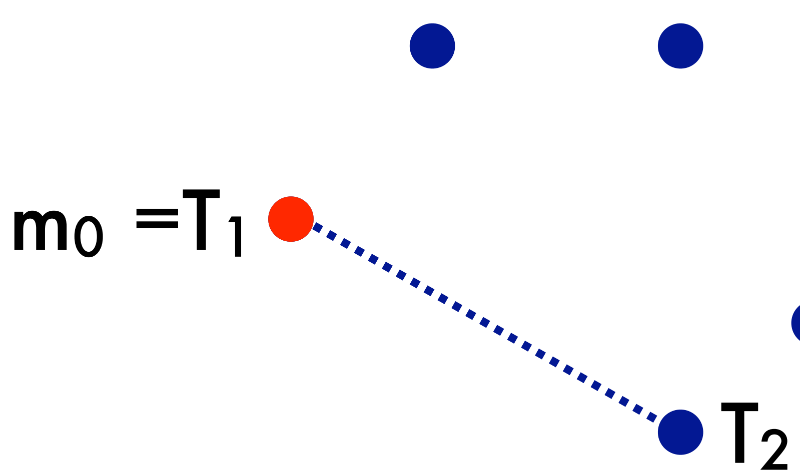
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

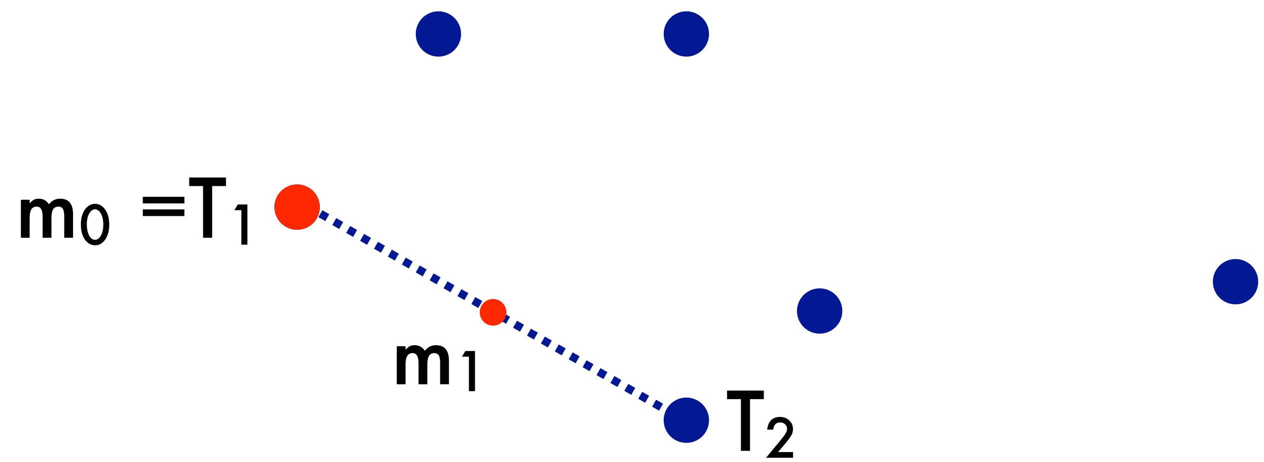
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

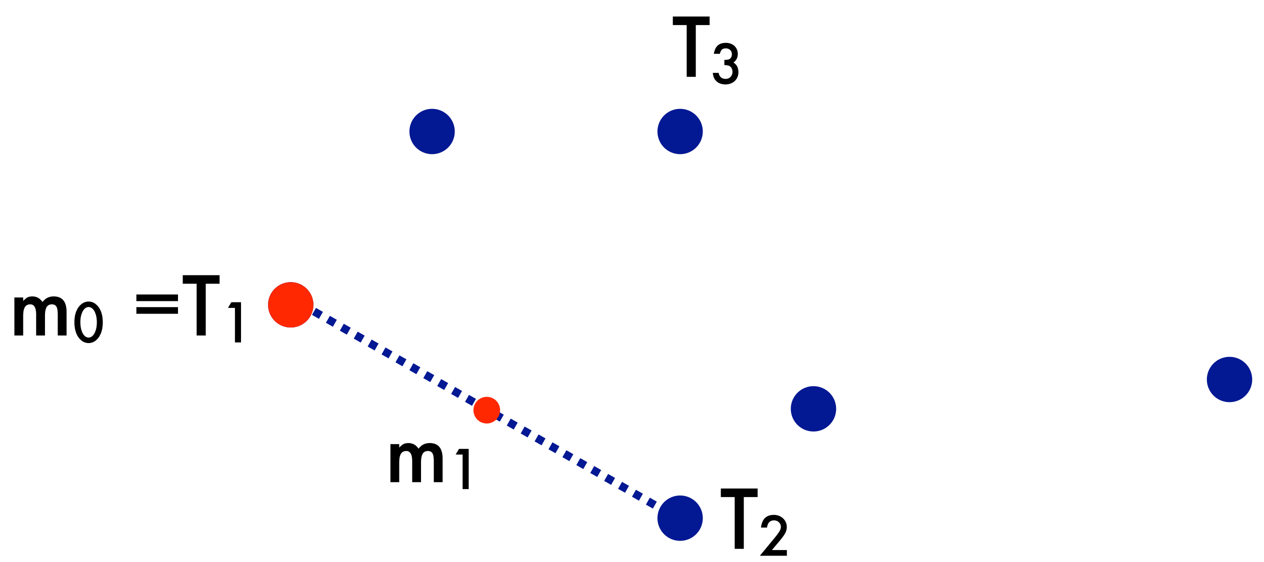
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

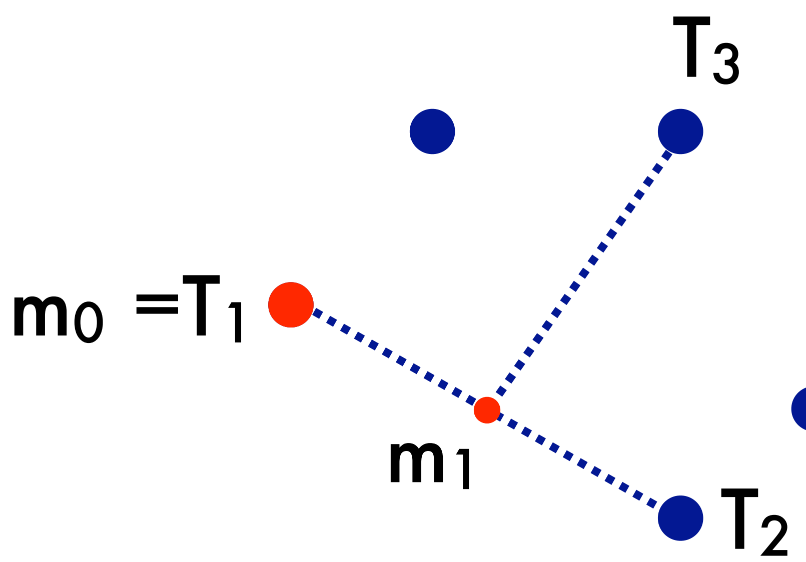
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

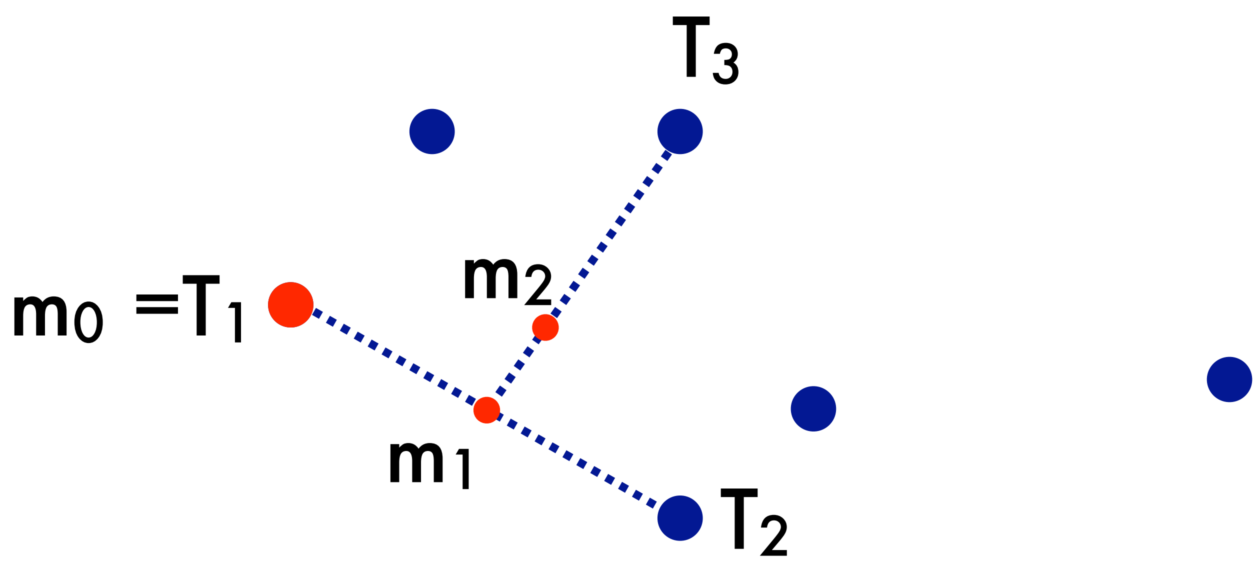
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

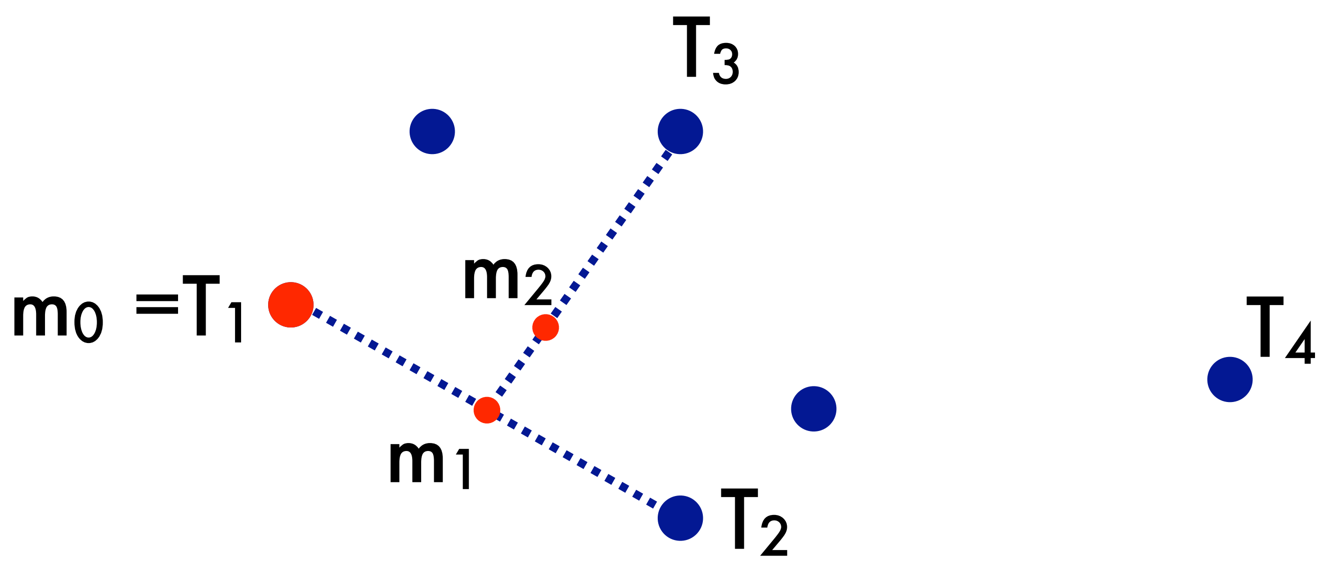
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

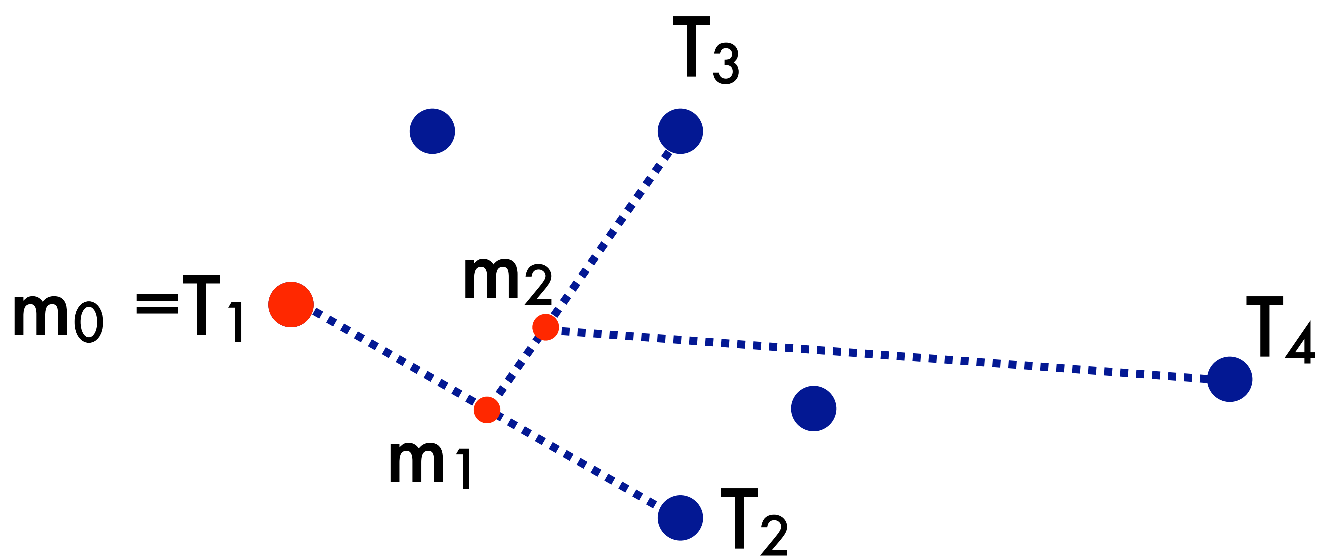
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

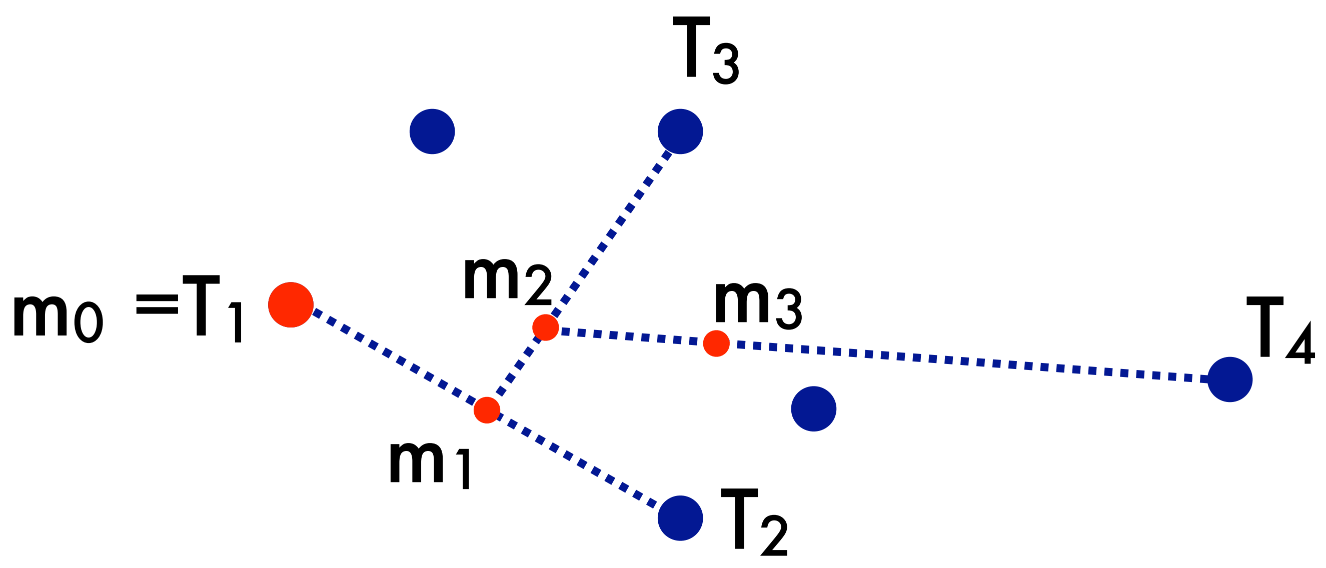
- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



Mean Trees

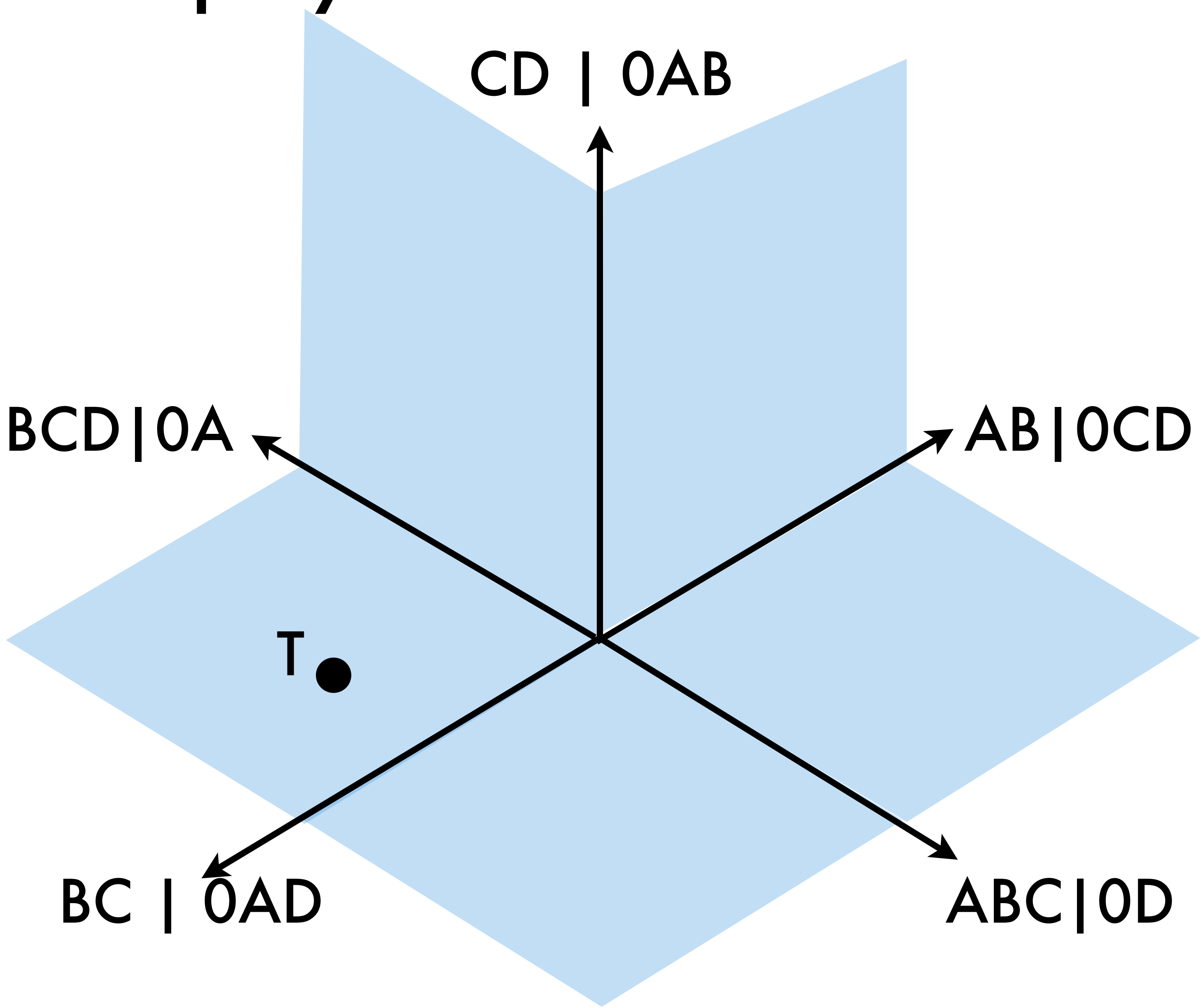
Theorem (Sturm, 2003): the following algorithm converges to the mean tree:

- $m_0 = T_1$
- i^{th} iteration:
  - randomly choose tree T_i from given set
  - $m_i = \frac{1}{i+1}$ (geodesic from m_{i-1} to T_i)



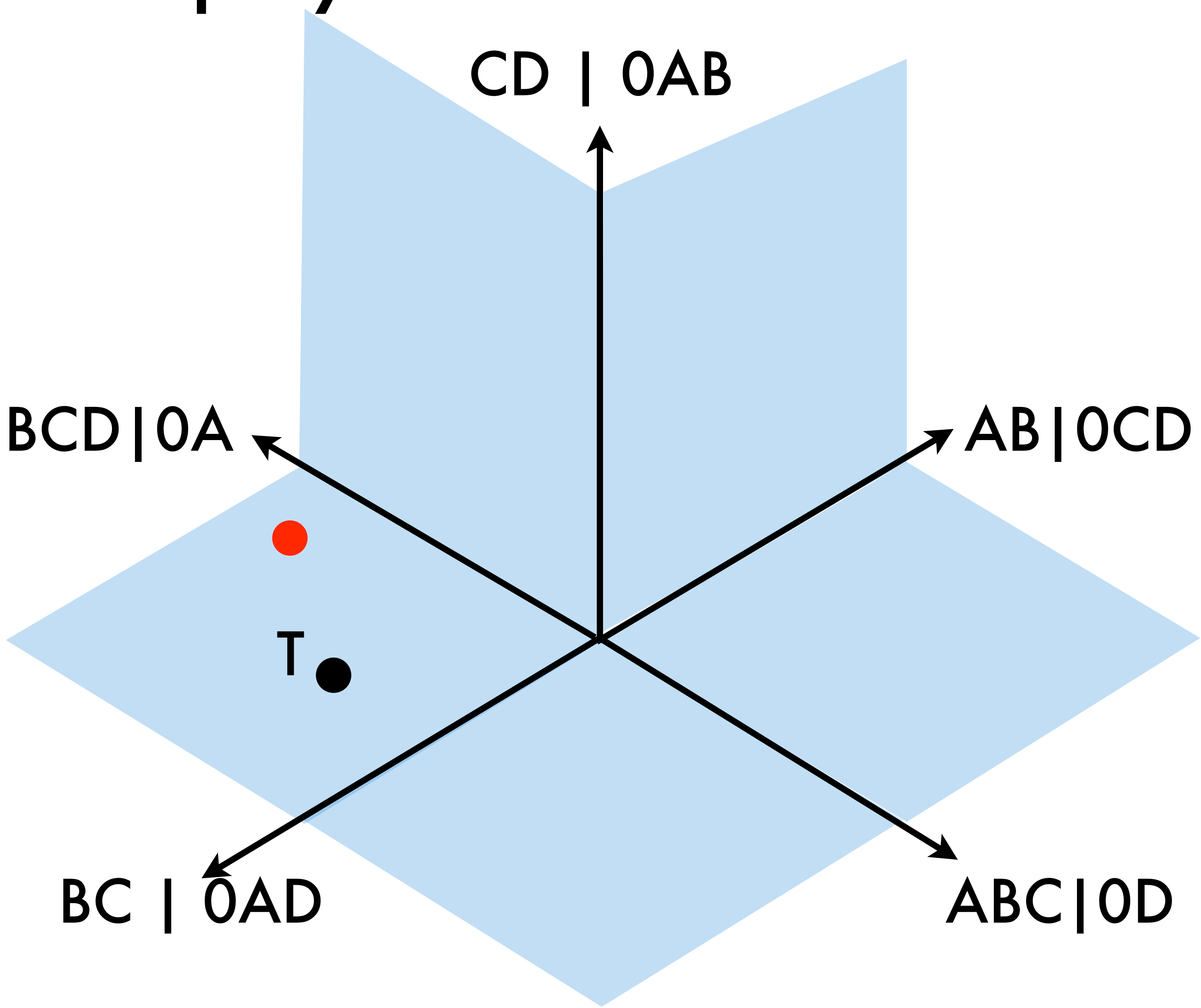
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



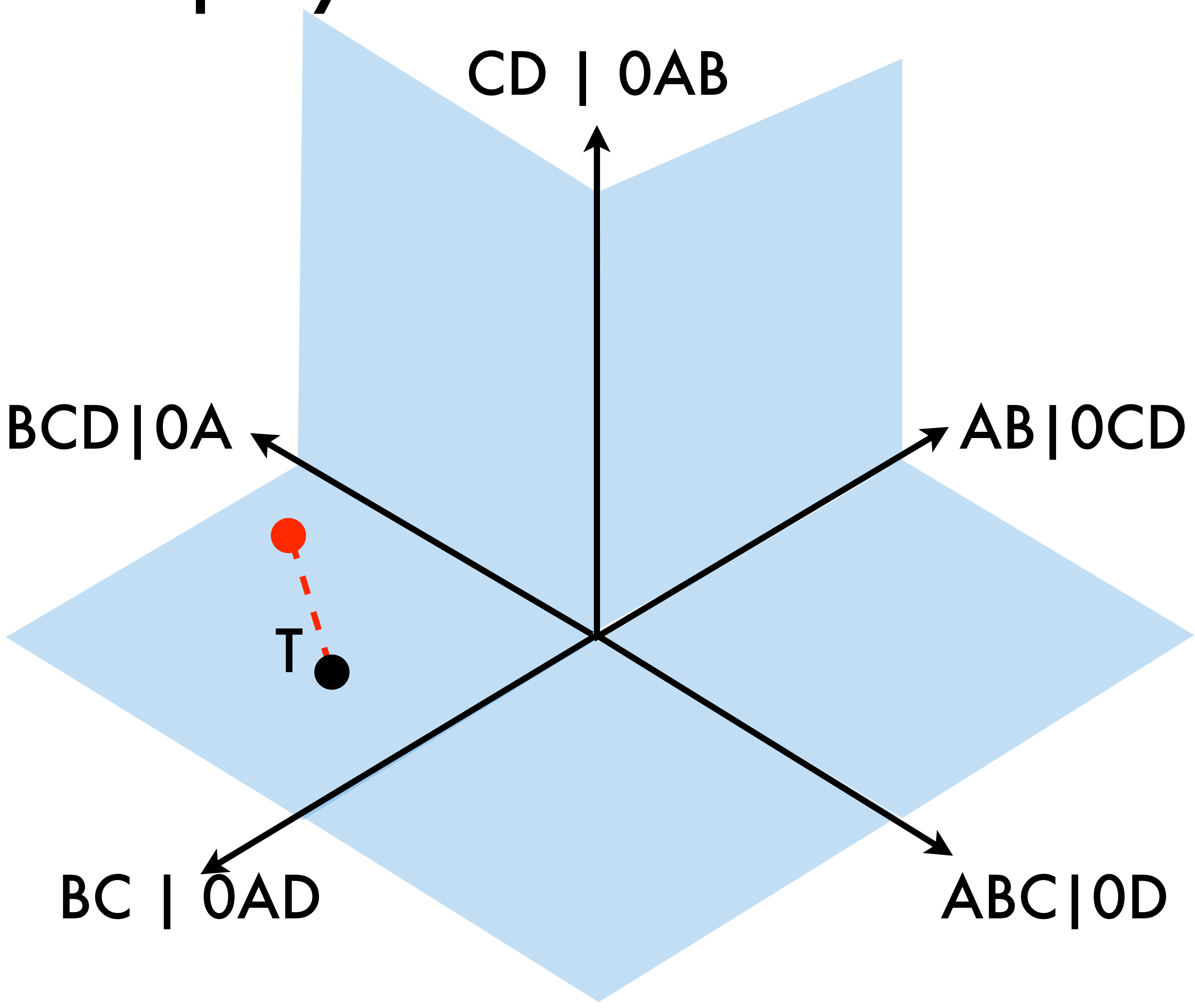
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



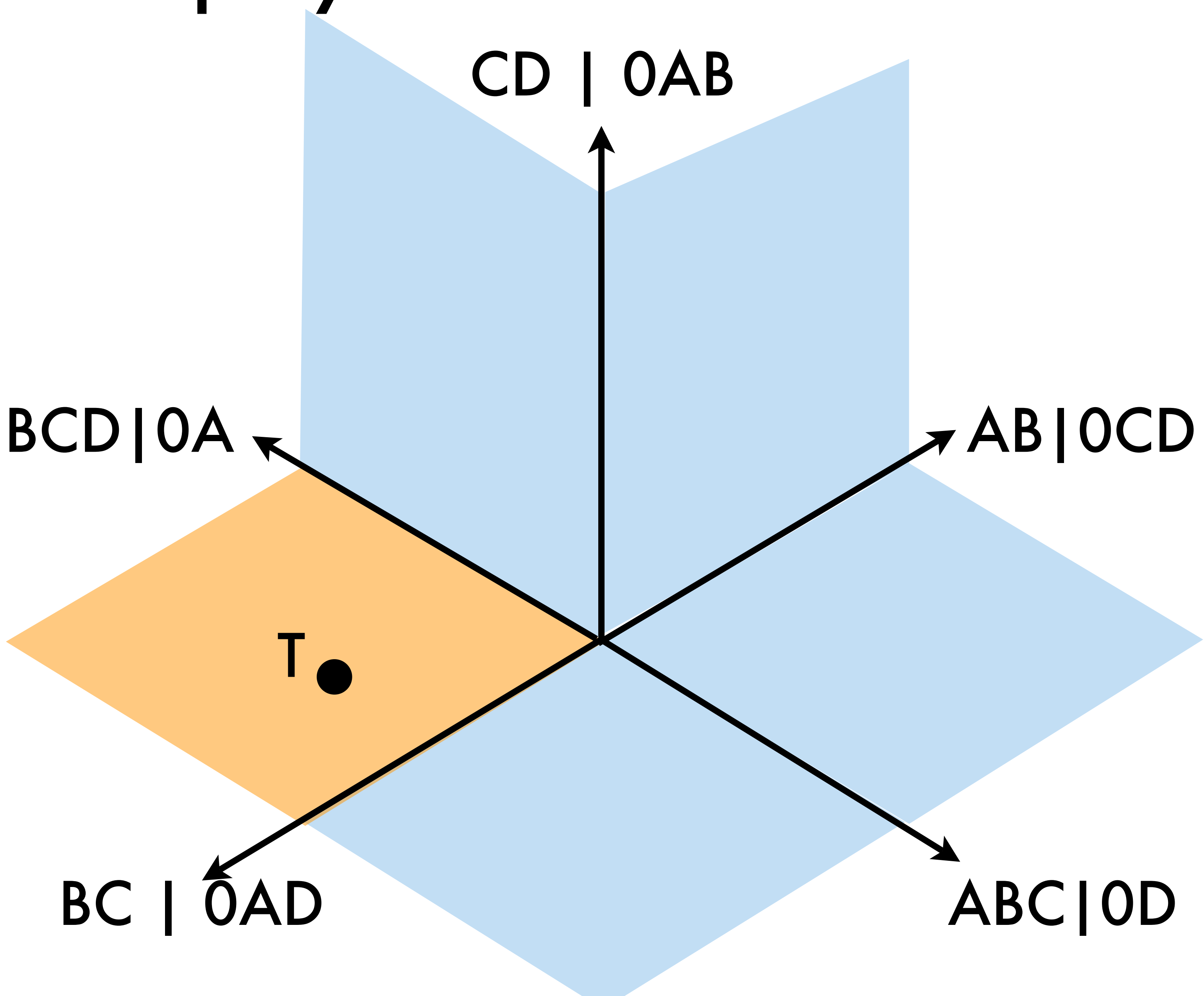
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



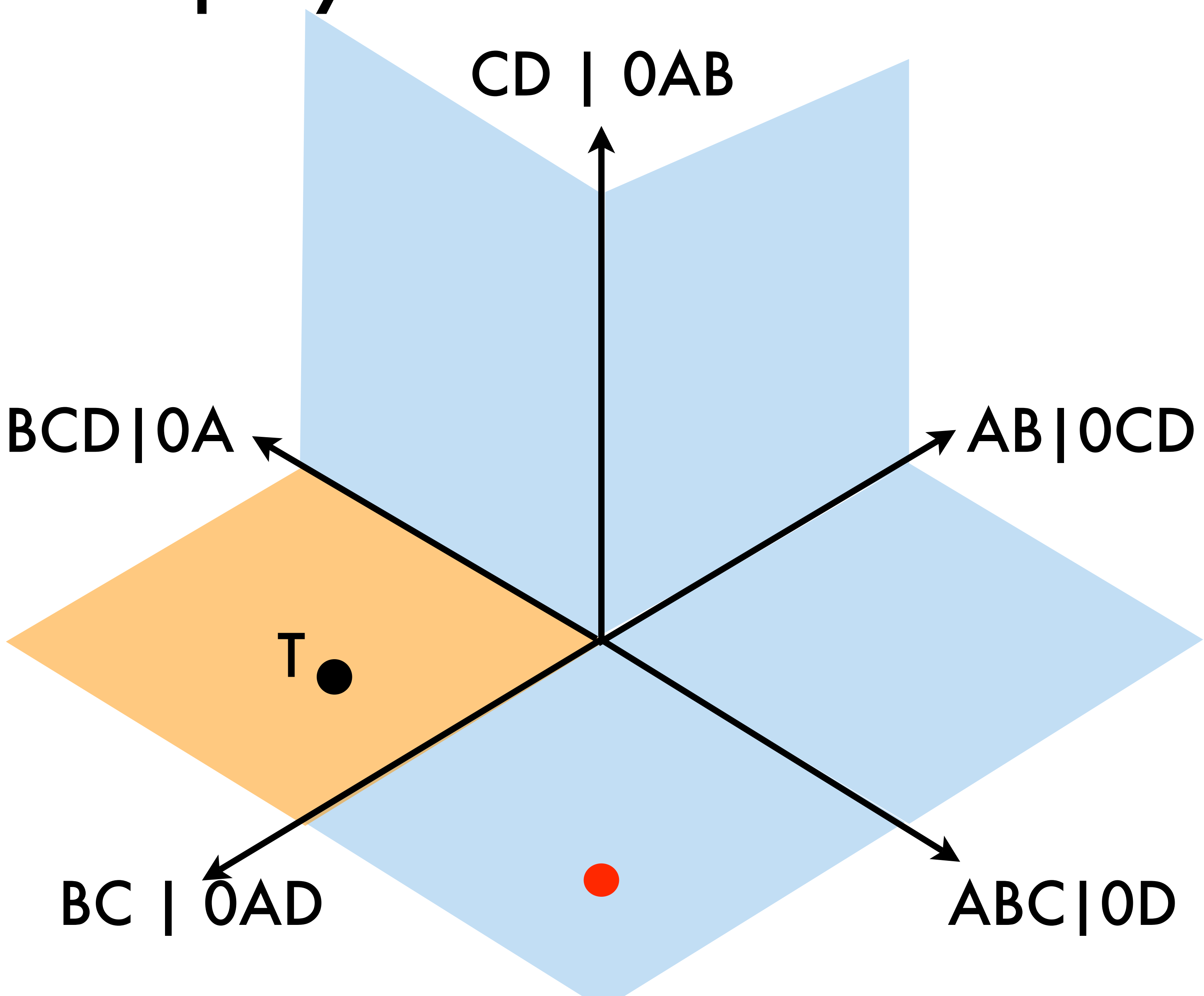
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



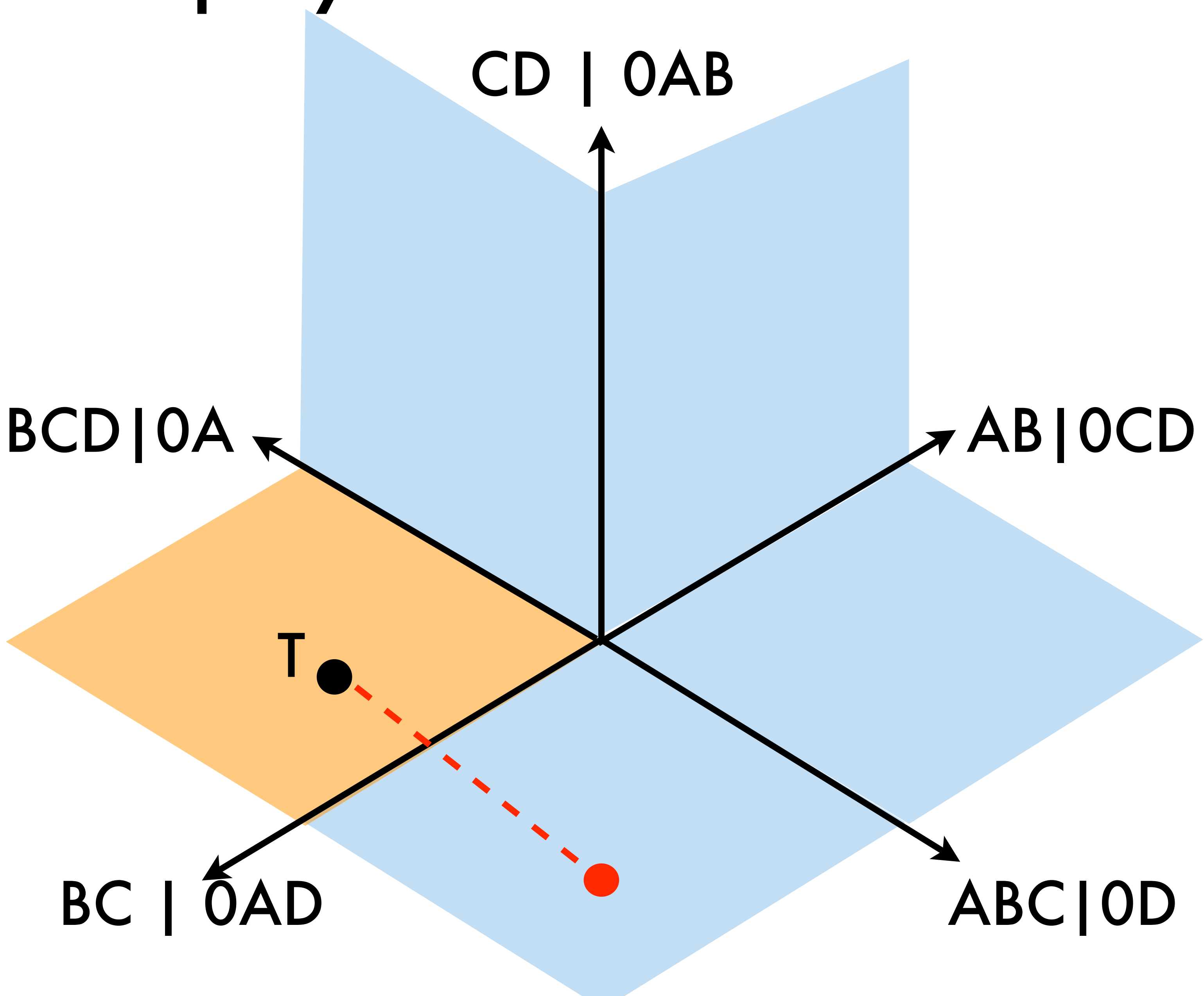
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



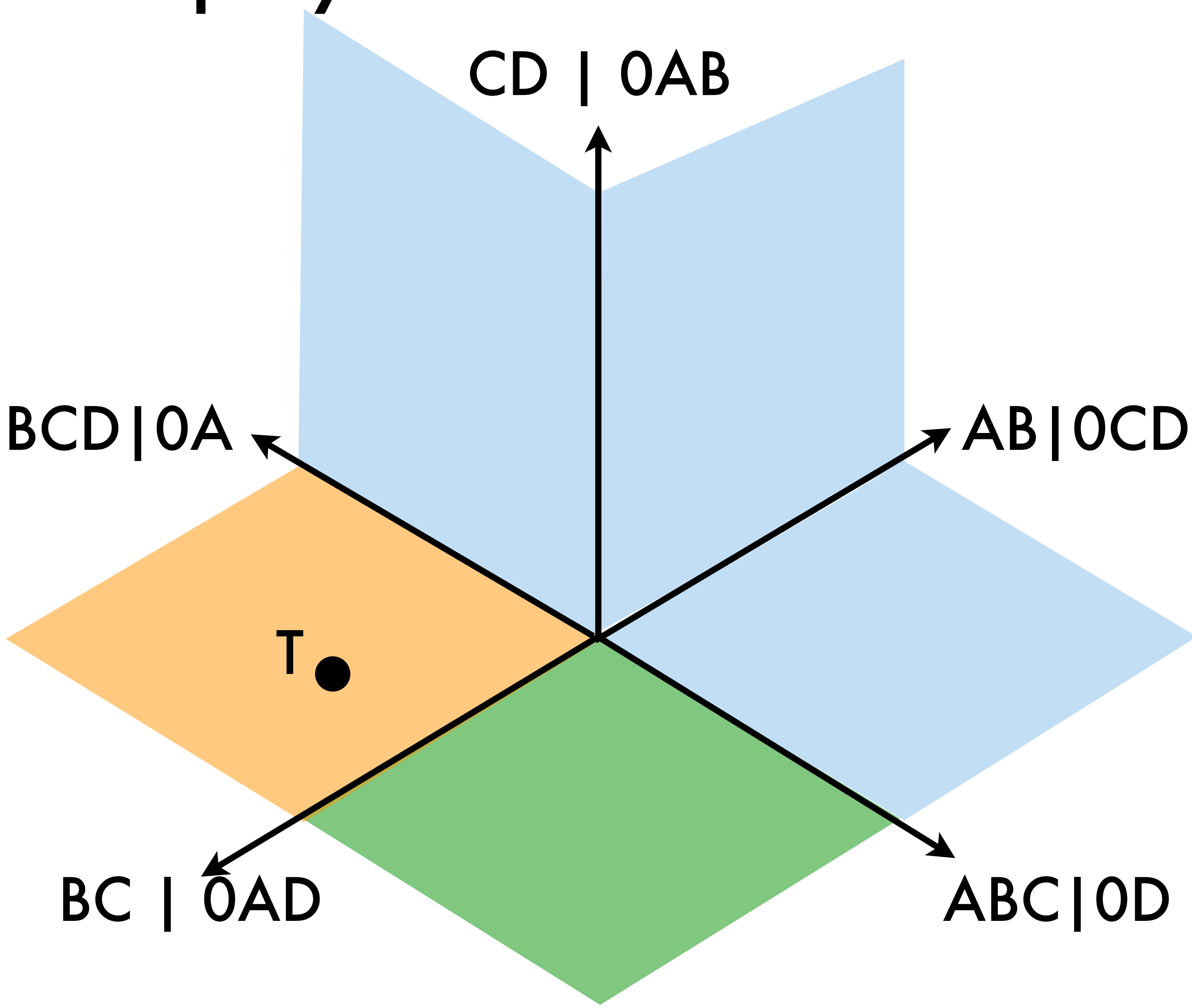
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



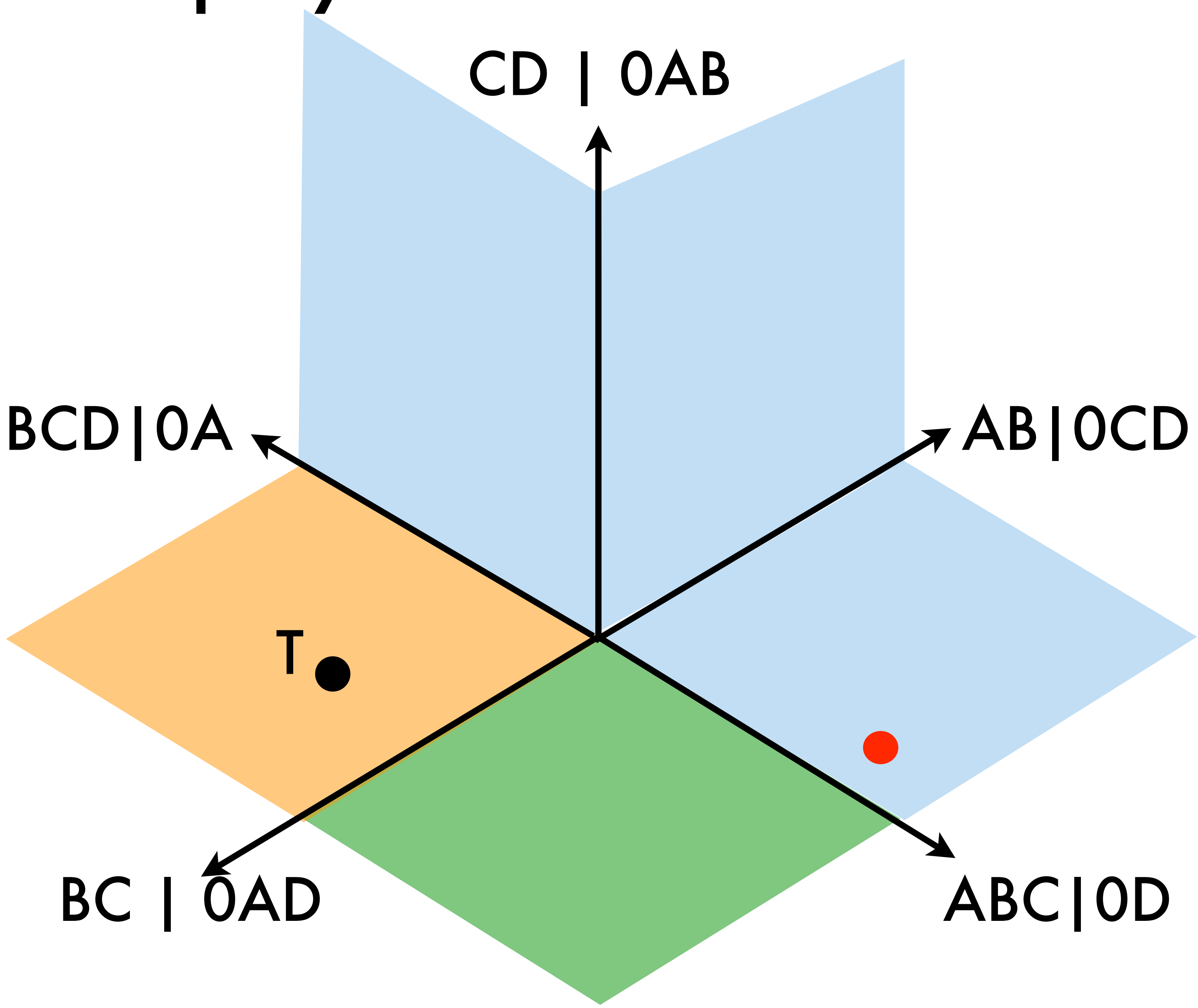
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



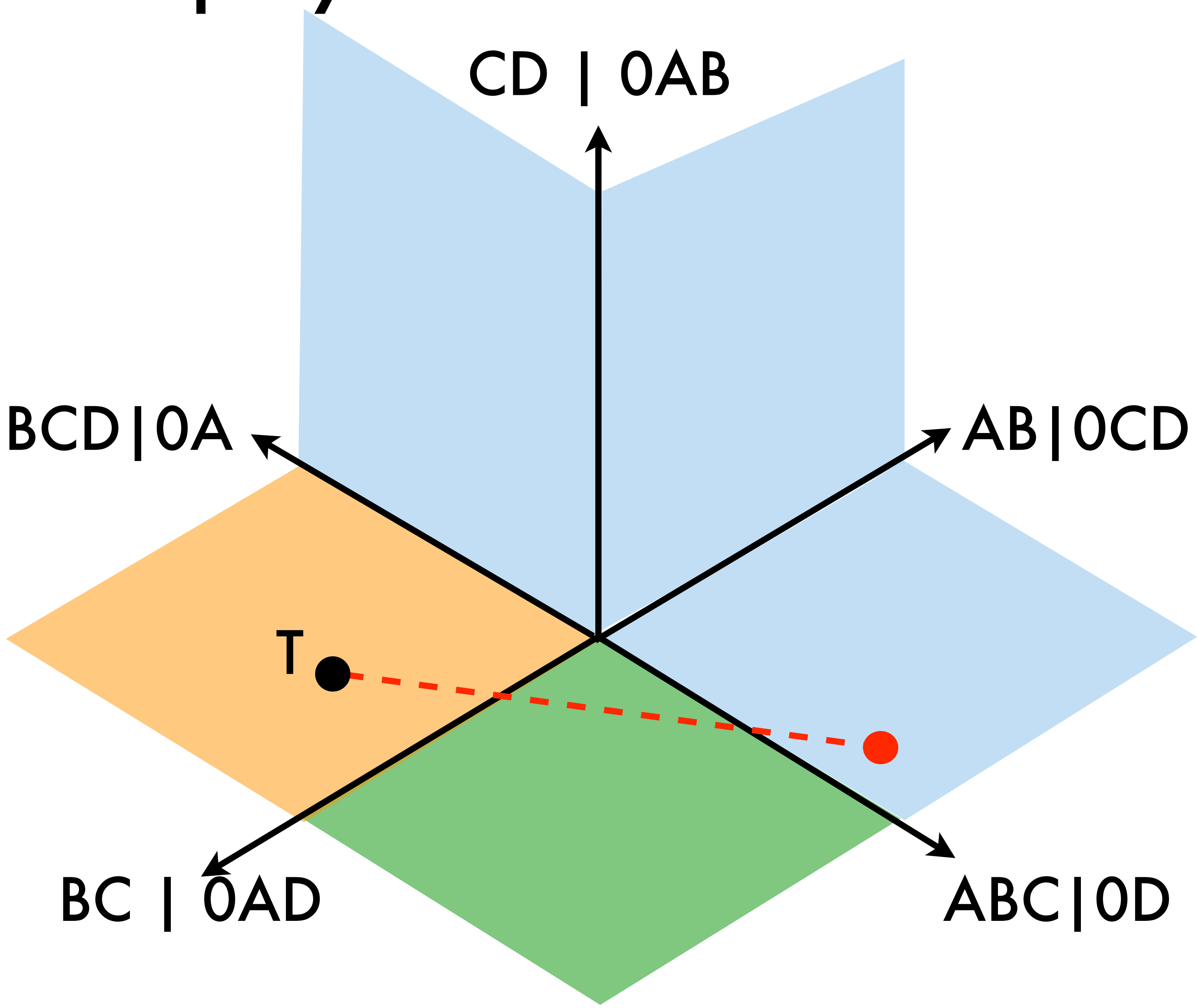
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



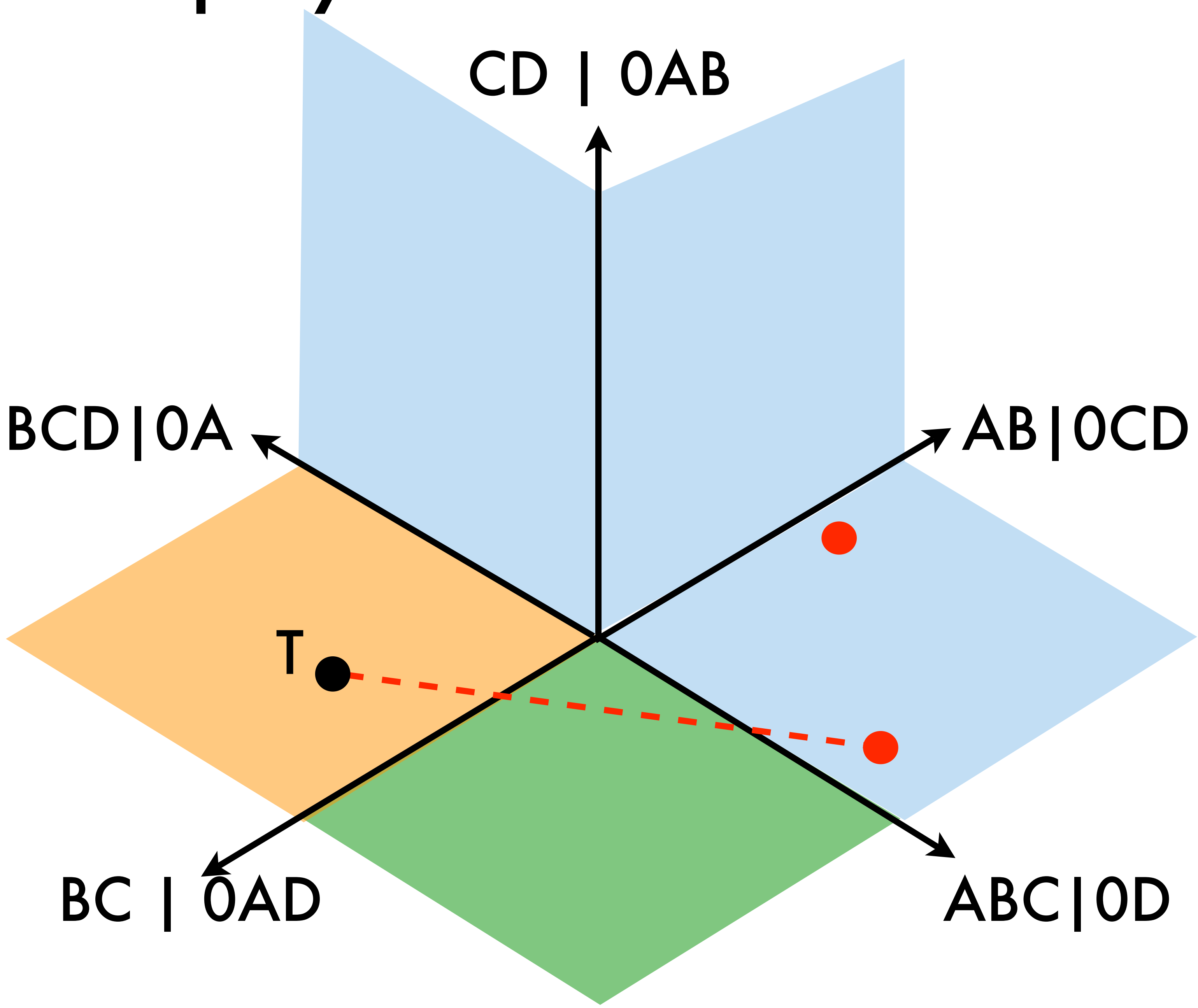
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



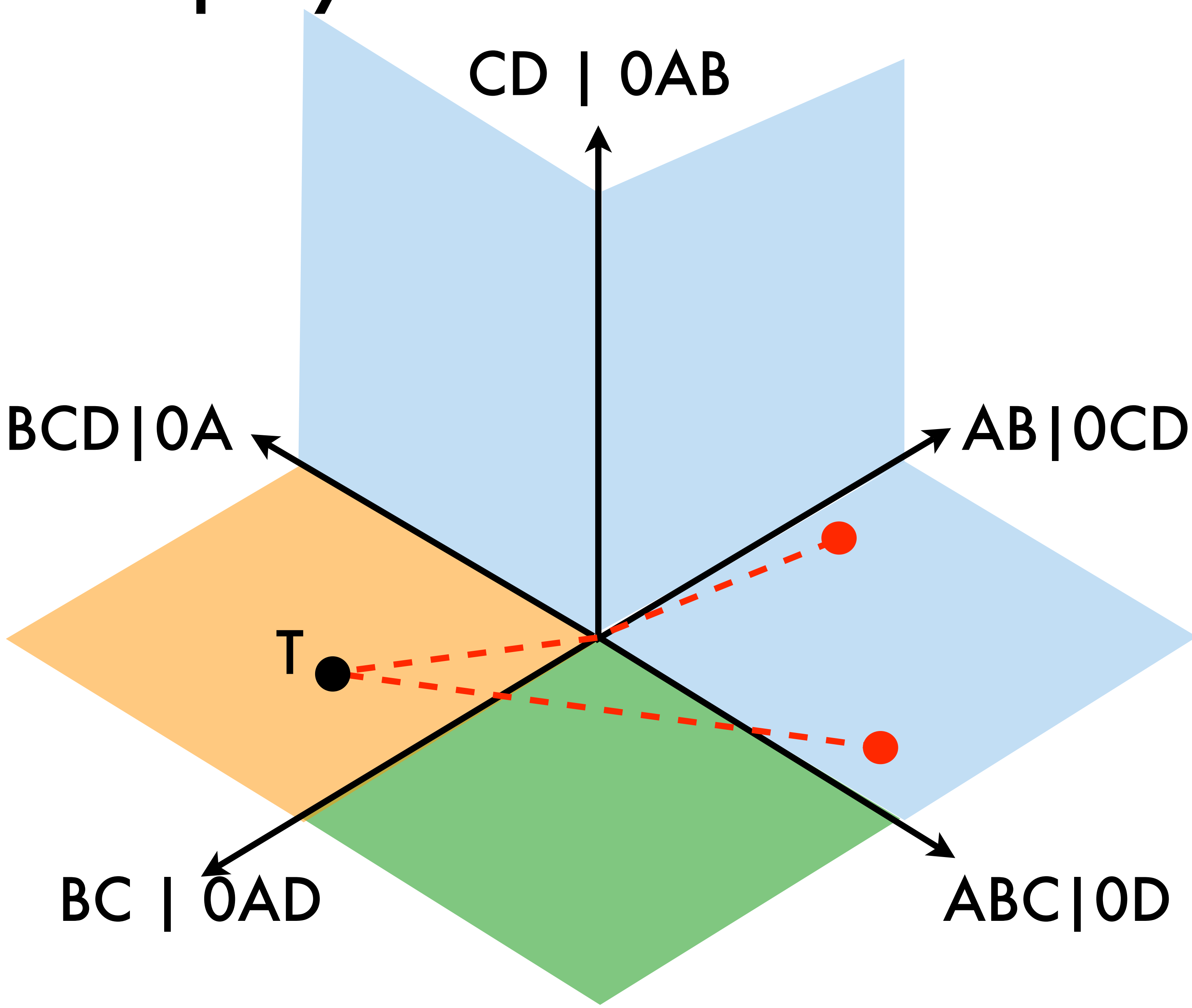
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



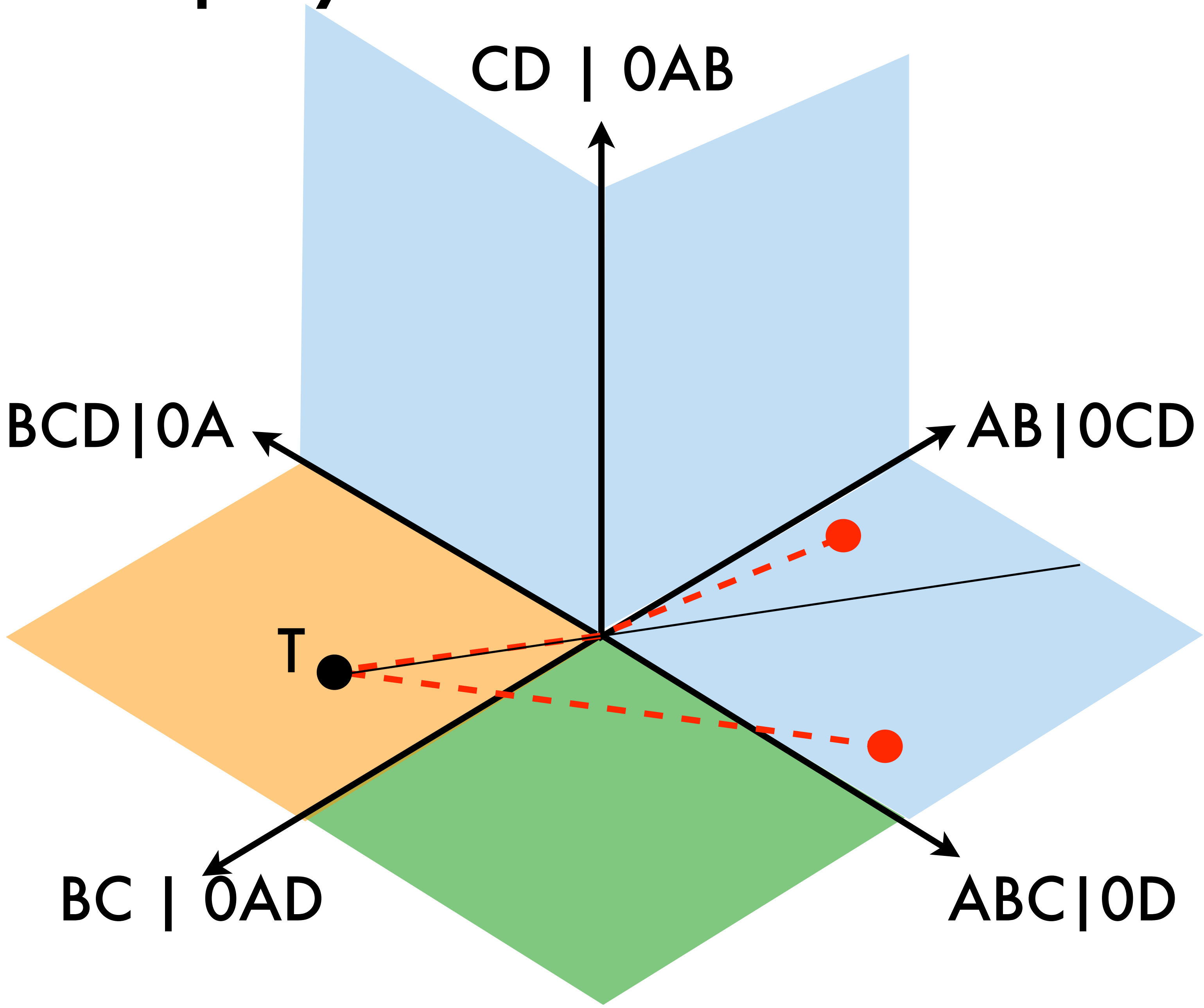
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



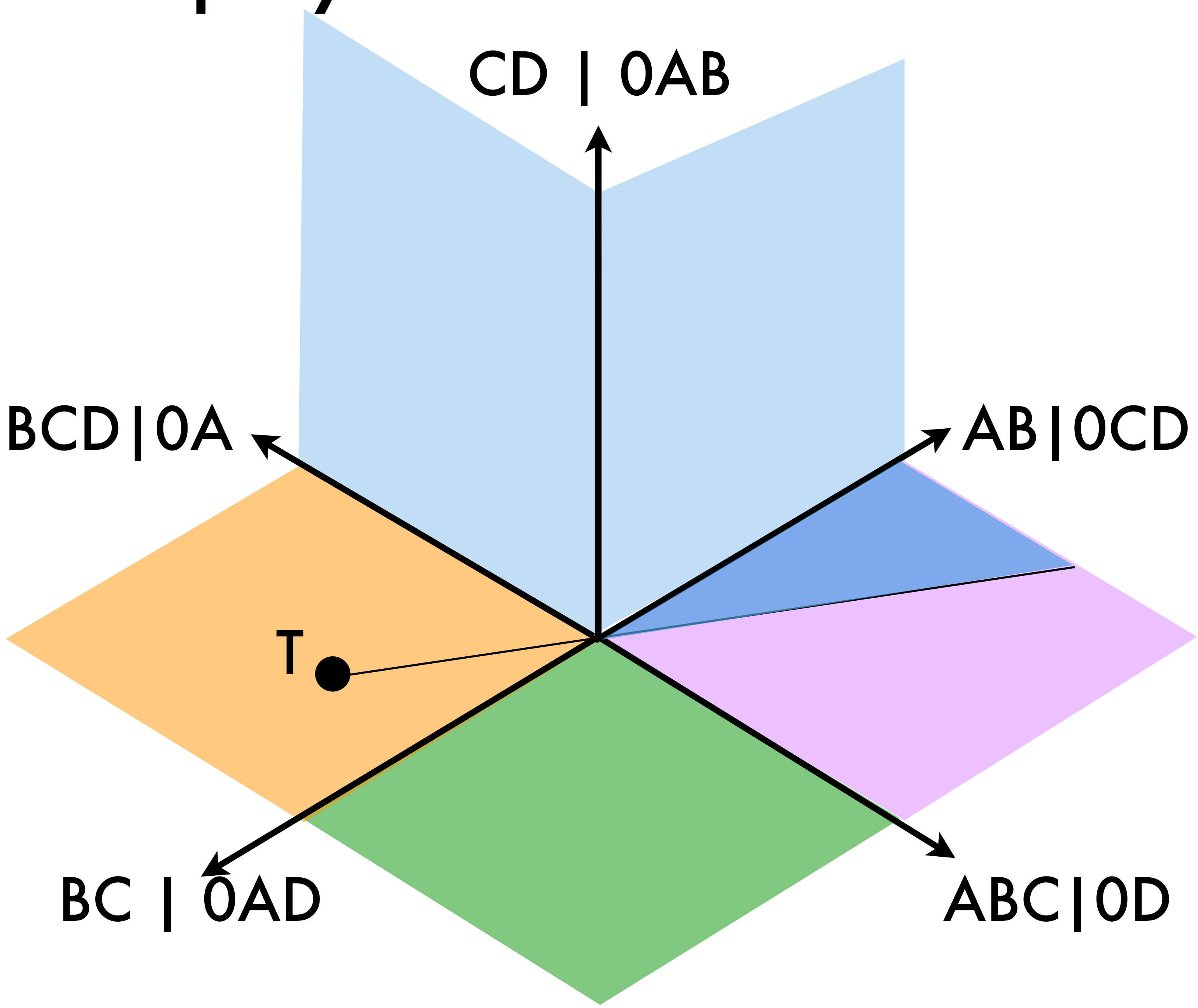
Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



Mean Trees

- combinatorial type of geodesic to a fixed tree T induces a polyhedral subdivision on tree space



Mean Trees

- combinatorial type of the geodesic to a fixed tree T induces a polyhedral subdivision on tree space
- use non-linear optimization to improve Sturm's algorithm:
  - once in correct polyhedral subdivision, gradient descent method will give minimum

Current and Future Work

- determine convergence of algorithm
- grouping similar trees using Principal Component Analysis
- using the geodesic distance and tree space to do statistics on trees

Thank You

- *A fast algorithm for computing geodesic distances in tree space* (Owen and Provan, 2010)

  http://arxiv.org/abs/0907.3942

samsi
NSF • Duke • NCSU • UNC • NISS