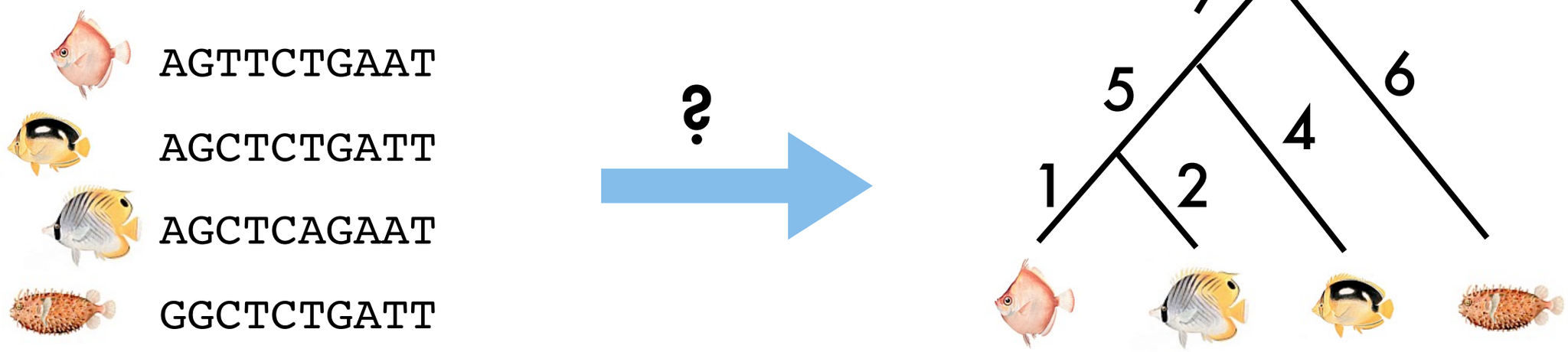
Tree Space: Algorithms & Applications

Part I

Megan Owen
University of Waterloo

Phylogenetic Trees

- a phylogenetic tree:



- questions:
  - how do we infer a tree from data?
  - how do we compare two trees?
  - how do we compute meaningful statistics?

Statistics on Trees

- to estimate a phylogenetic tree, the parameters are:
  - tree topology
  - edge lengths
- this is not a standard statistical problem! (usual parameter space is $\mathbb{R}^d$)
- need to develop new statistical techniques

(My) Ultimate Goal

- develop statistical theory for a space of phylogenetic trees analogous to statistical theory for Euclidean space
- challenges:
  - which space should we use?
  - how to verify theories and algorithms?
  - non-Euclidean behaviour (i.e. sticky means)
  - algorithms need to be practical or biologists won't use them

Outline

1. Tree spaces, including description of BHV tree space
2. Polynomial time algorithm for computing distance in BHV tree space
3. Open problems
4. Mean and variance in BHV tree space
5. Applications of BHV tree space
6. More open problems

Tree Space

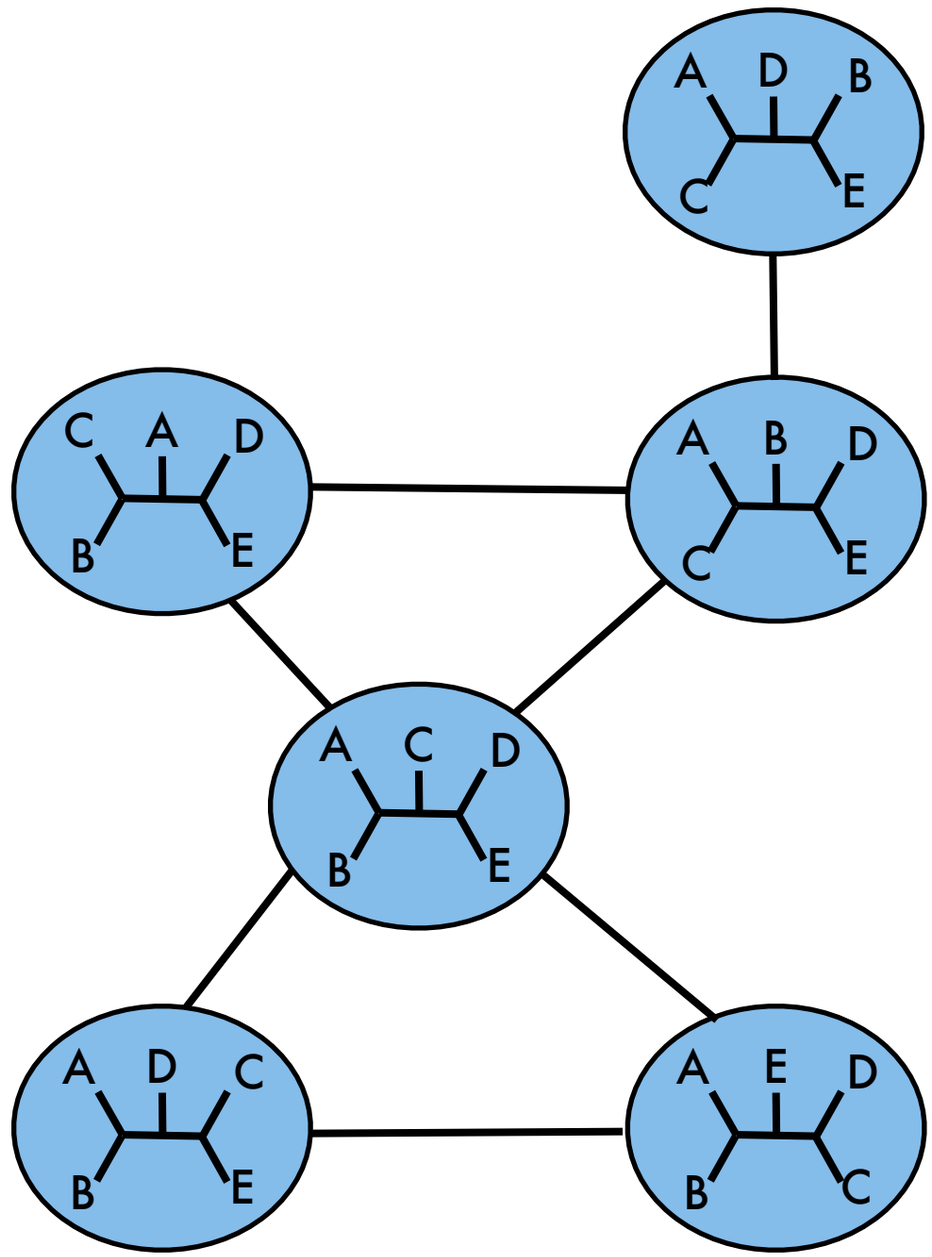
- a tree space is a metric space such that the points of the space are in bijection with some well-defined set of trees
- the metric of the tree space induces a distance between trees
  i.e. BHV tree space, tropical tree space

OR

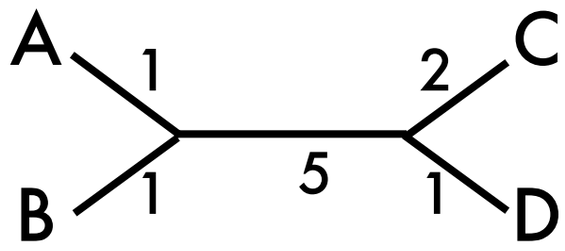
- given a distance measure between trees that is a metric, it induces a tree space i.e. NNI distance, Robinson-Foulds distance

Examples of Tree Spaces

- Example 1: a discrete tree space
  - vertices = tree topologies
  - edge between vertices iff topologies differ by NNI move
  - shortest paths are not unique
  - NP-hard to compute shortest path



Example 2



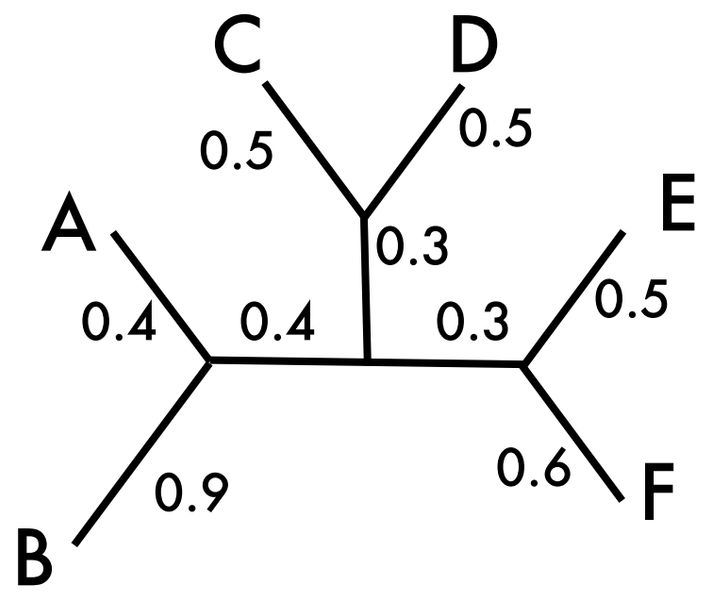
dissimilarity map

$$\begin{array}{c} \\ \mathsf{A} \\ \mathsf{B} \\ \mathsf{C} \\ \mathsf{D} \end{array} \begin{array}{c} \begin{array}{cccc} \mathsf{A} & \mathsf{B} & \mathsf{C} & \mathsf{D} \end{array} \\ \begin{bmatrix} - & 2 & 8 & 7 \\ 2 & - & 8 & 7 \\ 8 & 8 & - & 3 \\ 7 & 7 & 3 & - \end{bmatrix} \end{array}$$

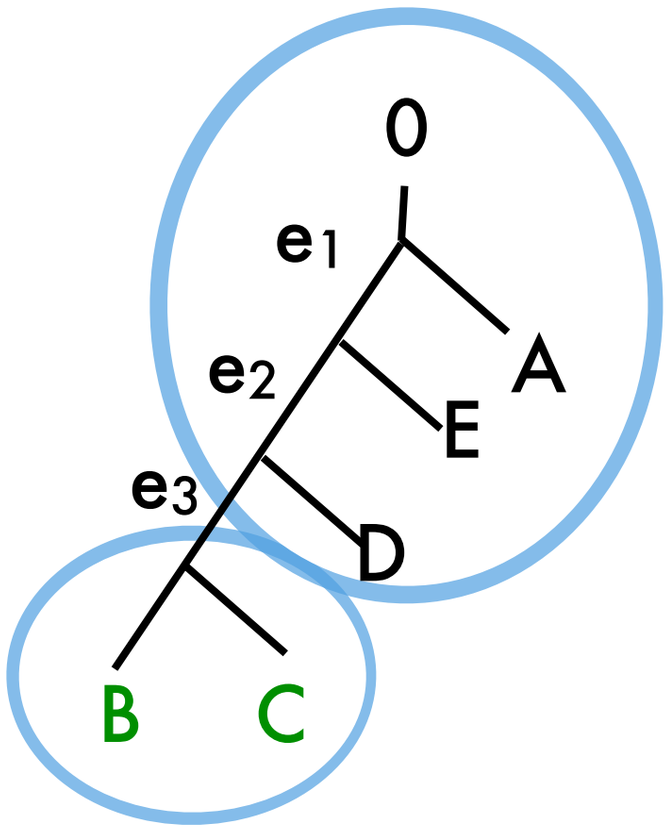
- tropical tree space = set of dissimilarity maps in $\mathbb{R}^{\binom{n}{2}}$ that are realizable as trees
- geodesics are not unique
- no algorithm for computing geodesics

BHV Tree Space

- constructed by Billera, Holmes, Vogtmann, 2001
  - $\mathbb{T}_n$ parametrizes all trees with n leaves and edge lengths
  - includes degenerate trees
  - all interior edges must have length/weight ≥ 0



Splits



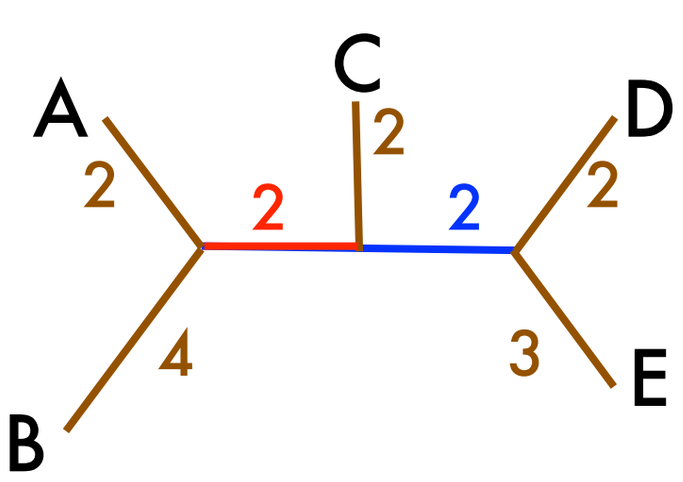
- each tree edge induces a *split*
- a *split* is a partition of the set of leaves:

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

or e_3 = BC | 0AED

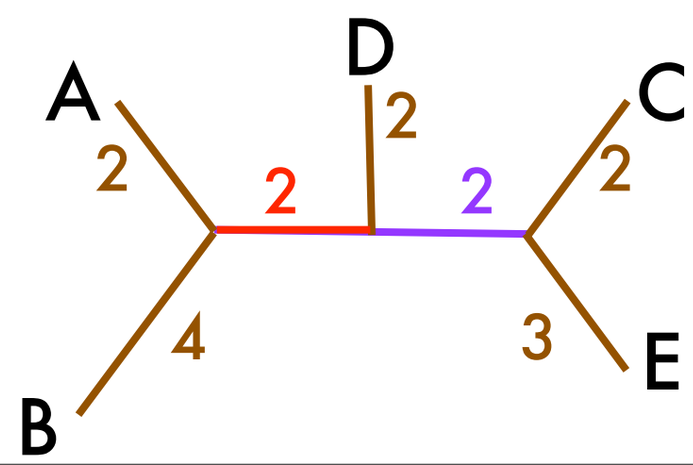
Tree Space

- represent each tree as a vector
- coordinates = splits



A|BCDE B|ACDE … E|ABCD AB|CDE AC|BDE … AE|BCD ABC|DE ABD|CE

(2, 4, 2, 2, 3, 2, 0, 0, 0, 2, 0, ….)

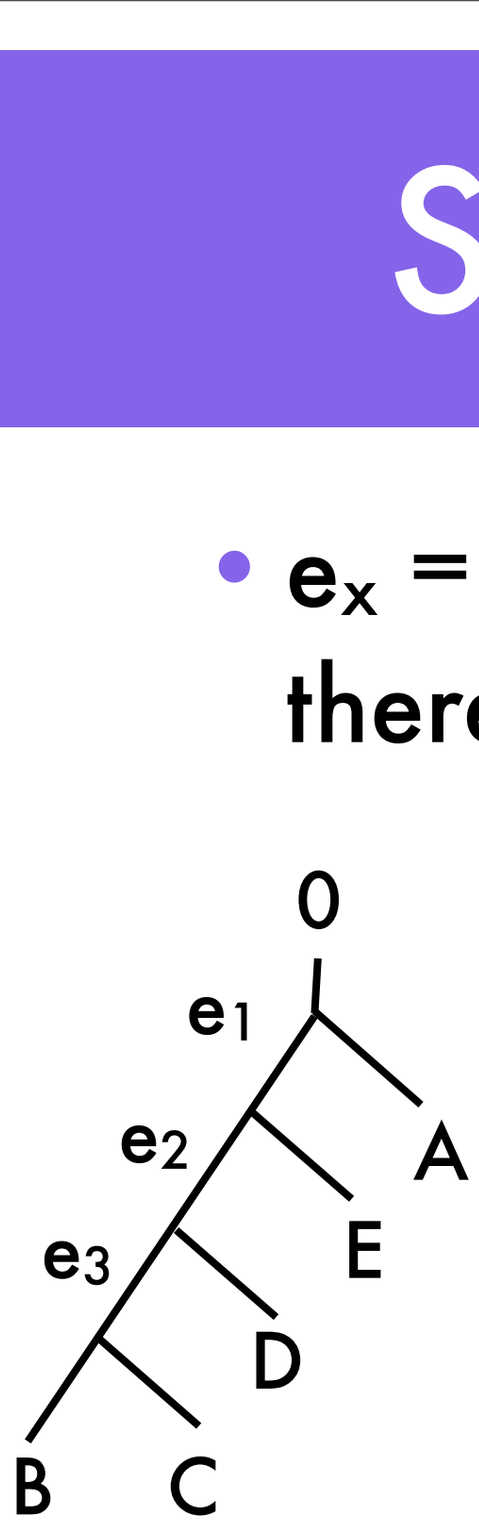


A|BCDE B|ACDE … E|ABCD AB|CDE AC|BDE … AE|BCD ABC|DE ABD|CE

(2, 4, 2, 2, 3, 2, 0, 0, 0, 0, 2, ….)

Split Compatibility

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



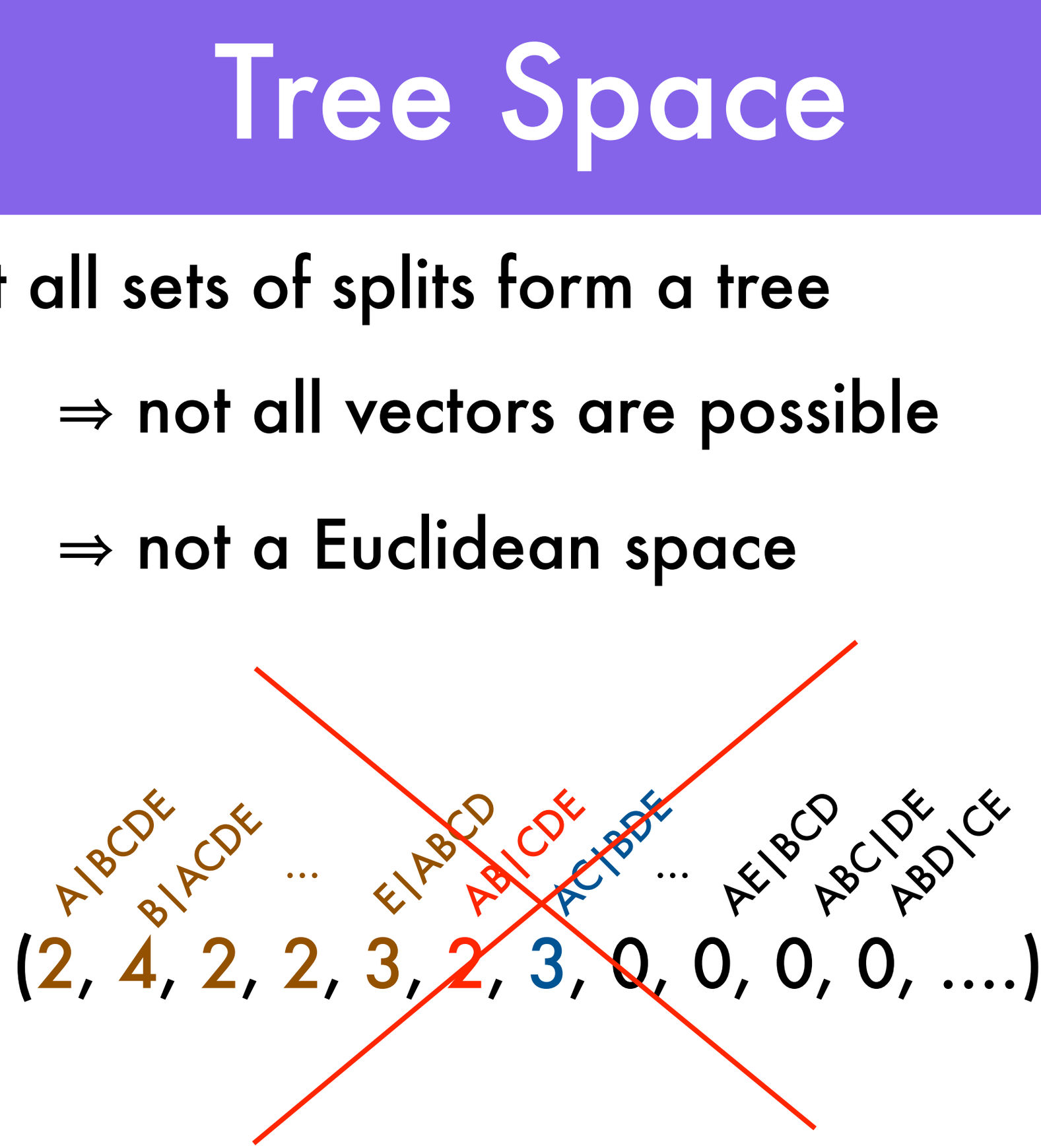
ex. e_3 = BC | 0AED is compatible with e_2 = BCD | 0AE but not with f = AB | 0CDE

Tree Space

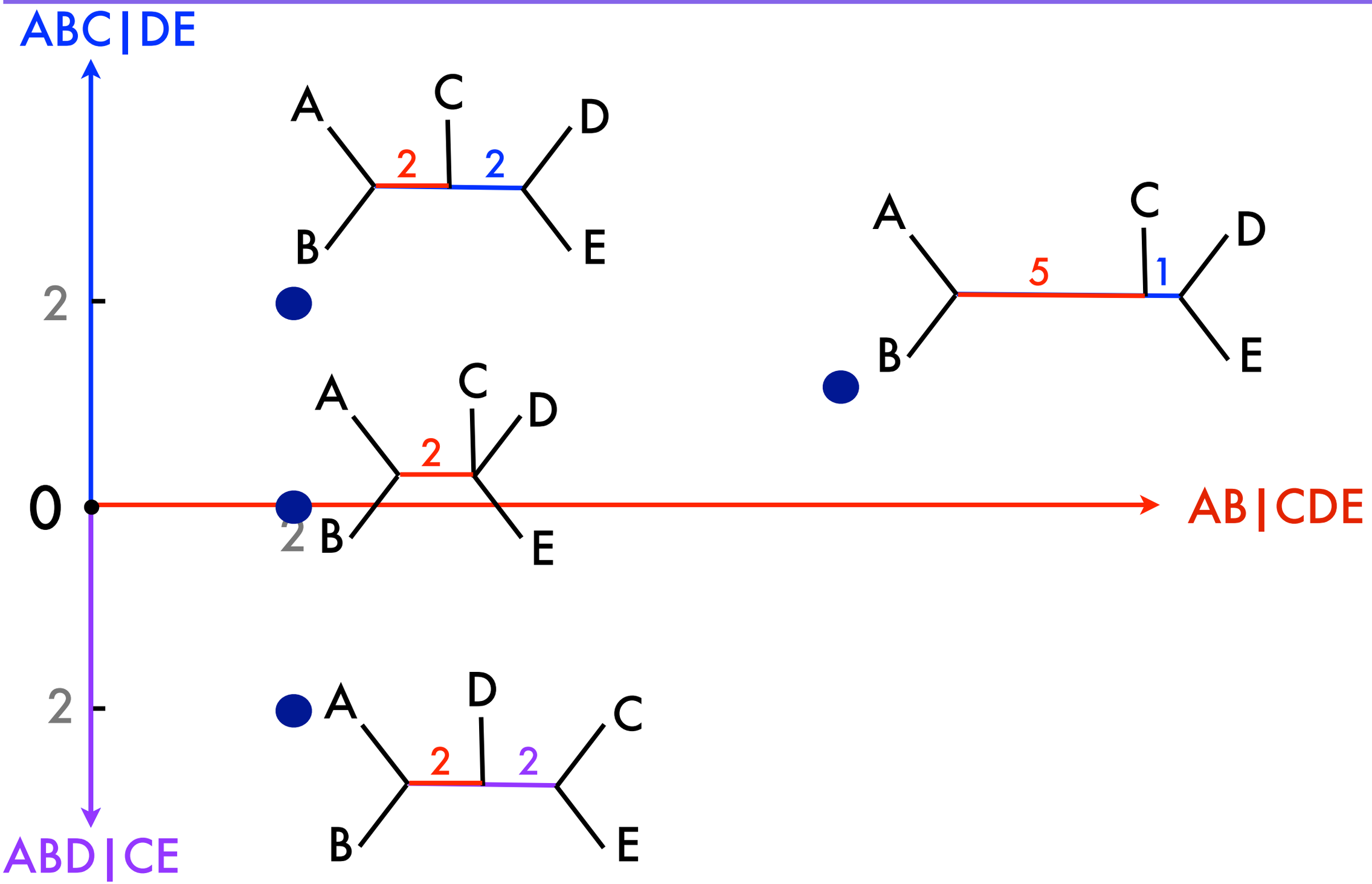
- not all sets of splits form a tree

  ⇒ not all vectors are possible

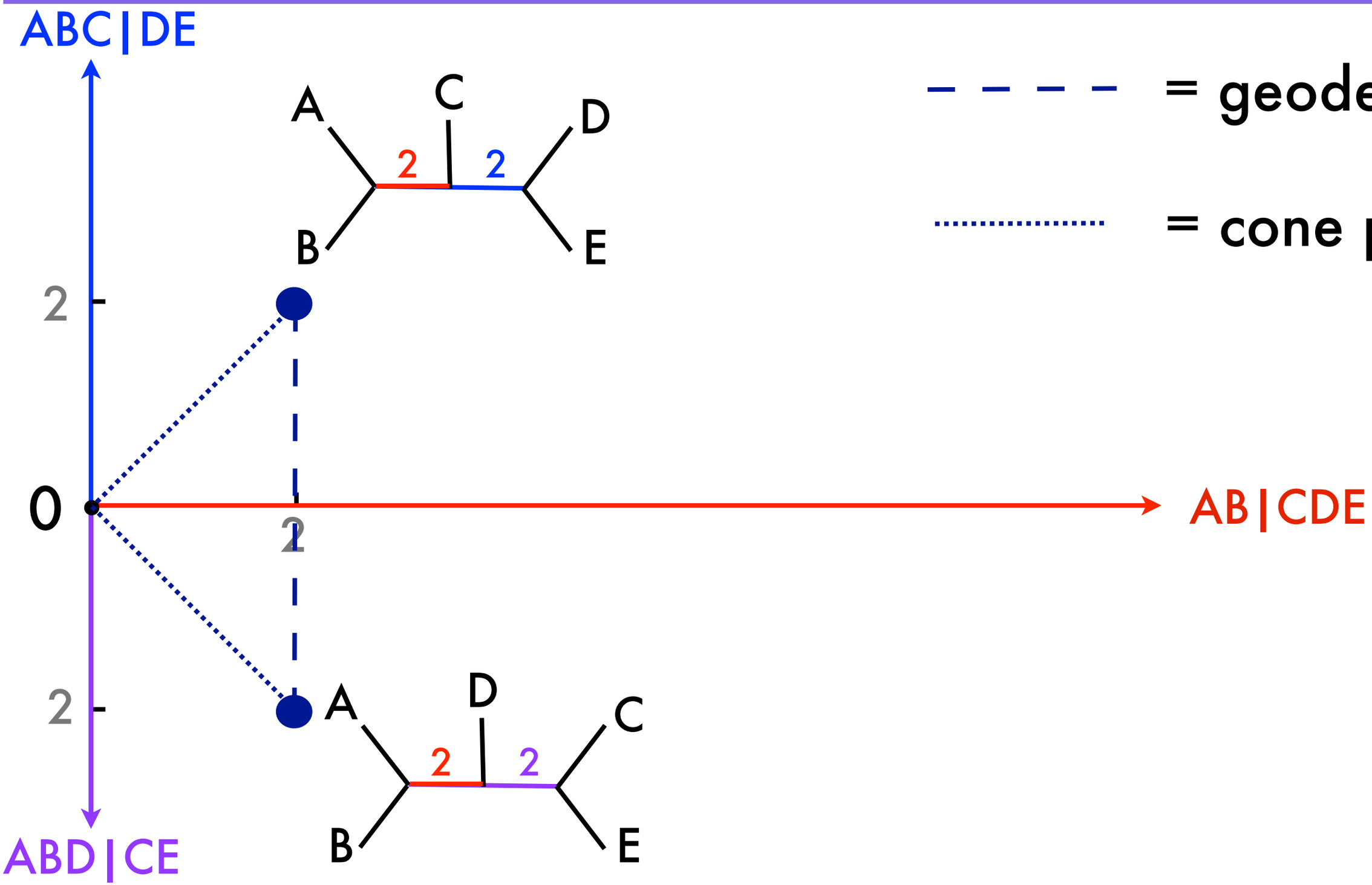
  ⇒ not a Euclidean space



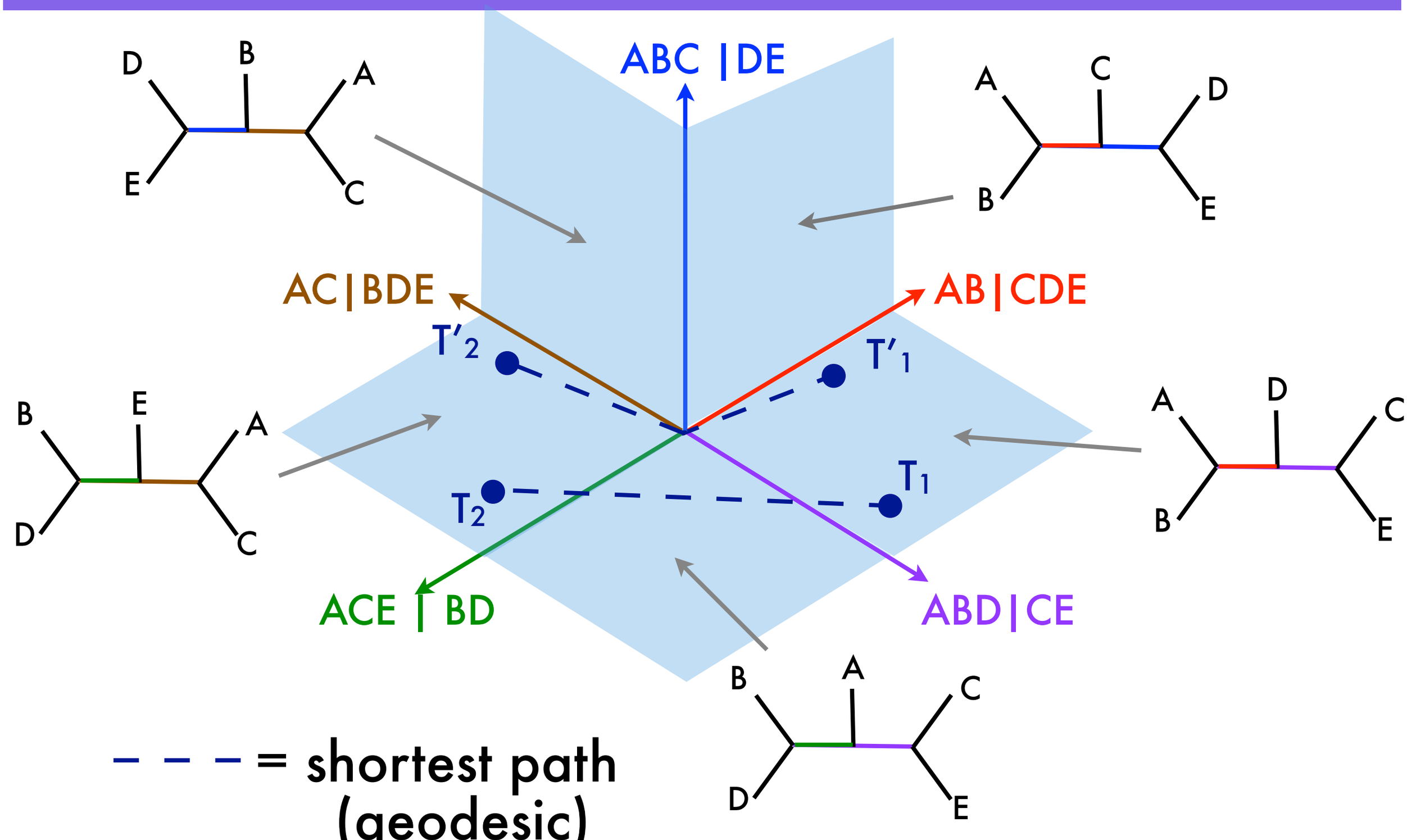
Tree Space



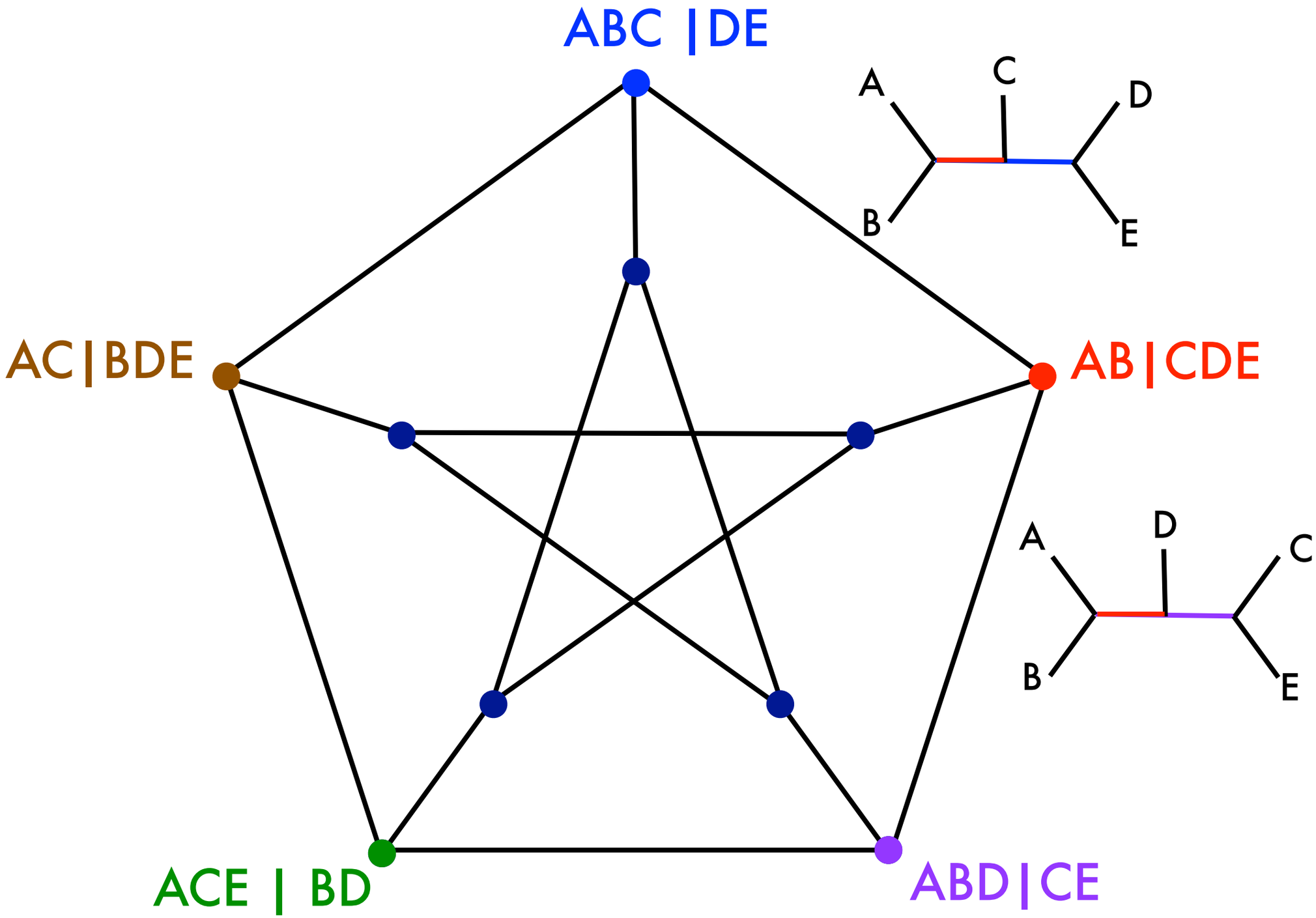
Tree Space



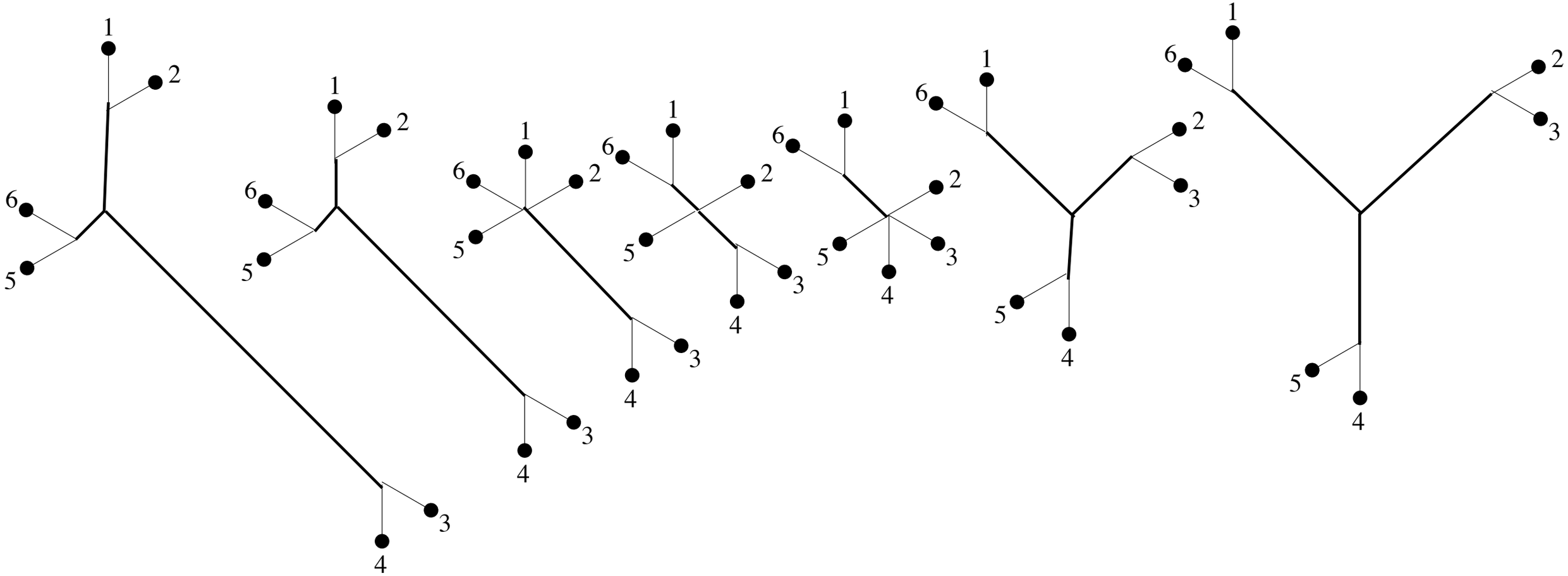
Tree Space


ABC | DE
AC | BDE
AB | CDE
ACE | BD
ABD | CE
T′2
T′1
T2
T1
D
B
A
E
C
A
C
D
B
E
B
E
A
D
C
A
D
C
B
E
B
A
C
D
E
= shortest path
(geodesic)

Structure of $\mathbb{T}_4$



Geodesics



Tree Space Properties

Theorem (Billera, Holmes,Vogtmann, 2001):
Tree space has global non-positive curvature (CAT(0)).

$\Rightarrow$ unique geodesics (shortest paths)

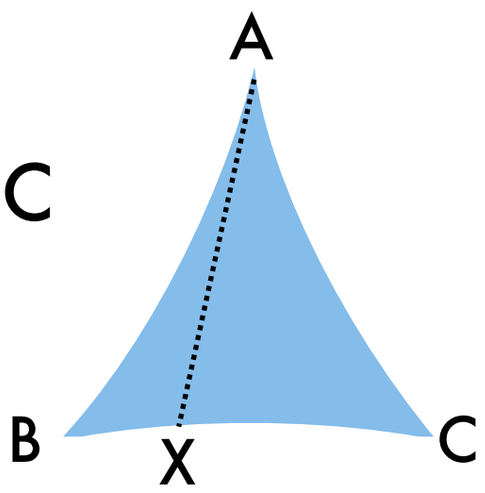
$\Rightarrow$ well-defined mid-point tree

- geodesic distance = length of shortest path (geodesic) between two trees T_1 and T_2
- computable in polynomial time via GTP algorithm (O., Provan, 2011)

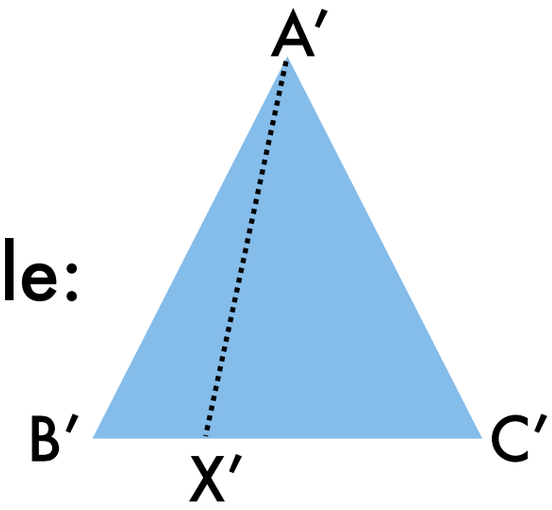
Non-positive Curvature

- *non-positive curvature (NPC)* = triangles are at least as thin as in Euclidean space
- *global non-positive curvature* = all triangles are at least as thin as in Euclidean space = CAT(0)

triangle in a NPC space:



Euclidean comparison triangle:



$d(A,X) \leq d'(A',X')$

- CAT(0) ⇒ unique shortest paths (*geodesics*)

CAT(0) Cubical Complexes

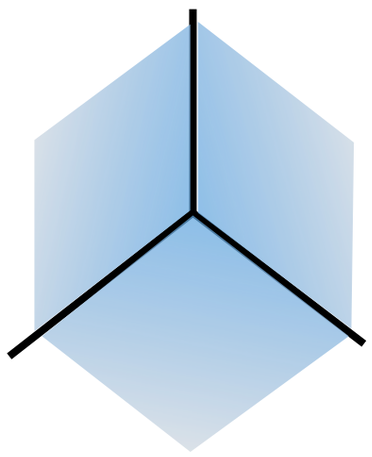
Theorem (Gromov, 1987):

A cubical complex is CAT(0)

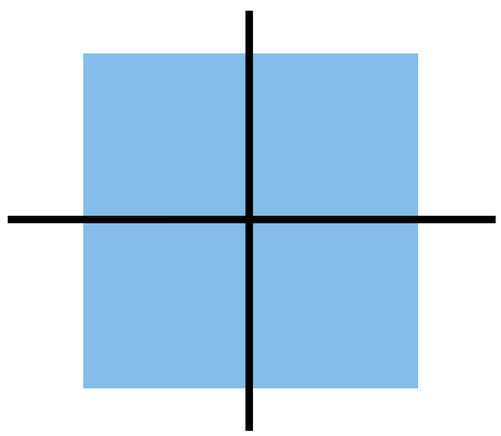
⇔ it is simply connected and the link of any vertex is a flag simplicial complex

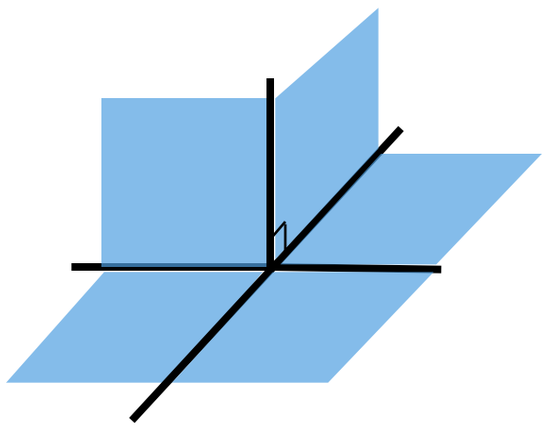
⇔ it is simply connected and if a vertex is incident to K edges, any pair of which specify a square, then these K edges also specify a K-dimensional cube.

not CAT(0):

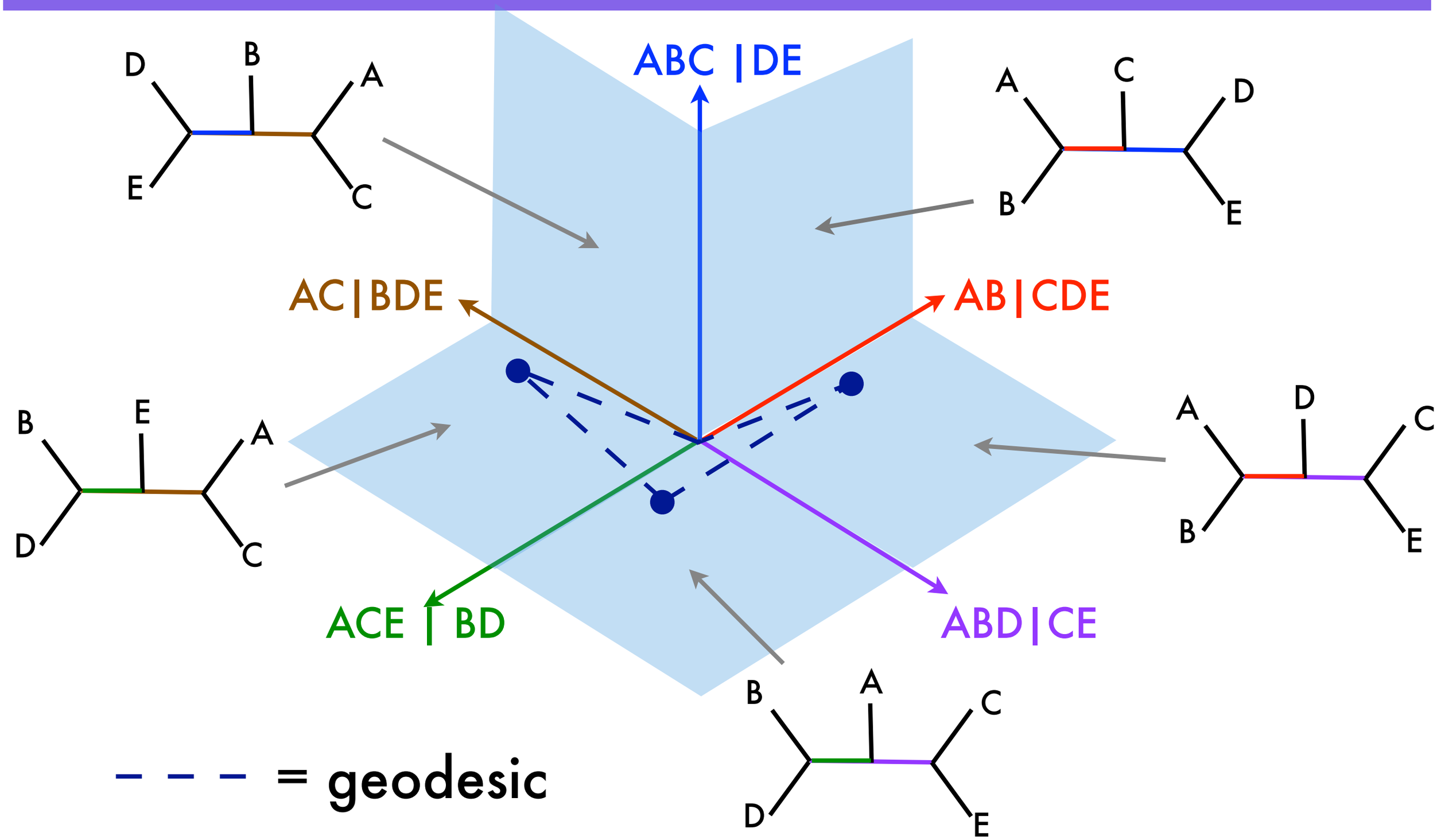


CAT(0):





Thin Triangle



Outline

1. Tree spaces, including description of BHV tree space
2. Polynomial time algorithm for computing distance in BHV tree space
3. Open problems
4. Mean and variance in BHV tree space
5. Applications of BHV tree space
6. More open problems

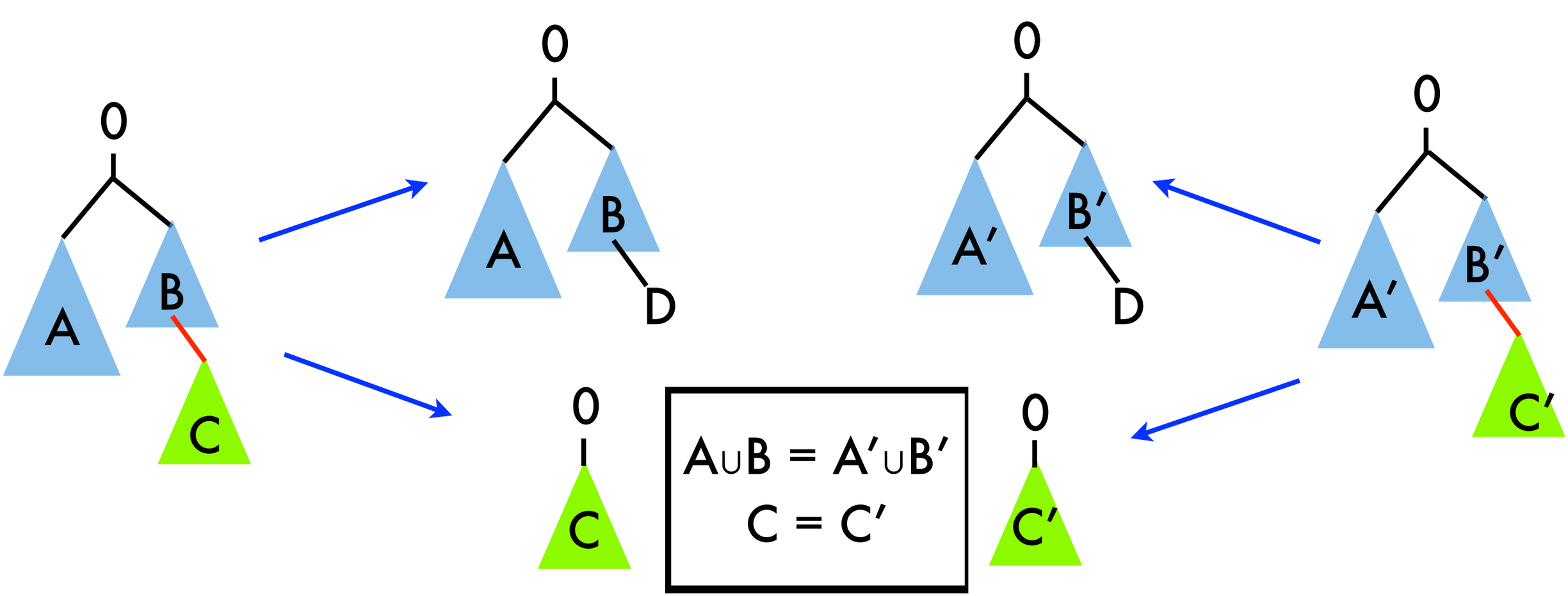
Size of $\mathbb{T}_n$

- if trees have n leaves, then:
  - orthants have (n-2)-dimensions
  - (2n - 3)!! tree topologies
    = (2n - 3)!! orthants

- orthant = non-negative part of R^n

Common Edges

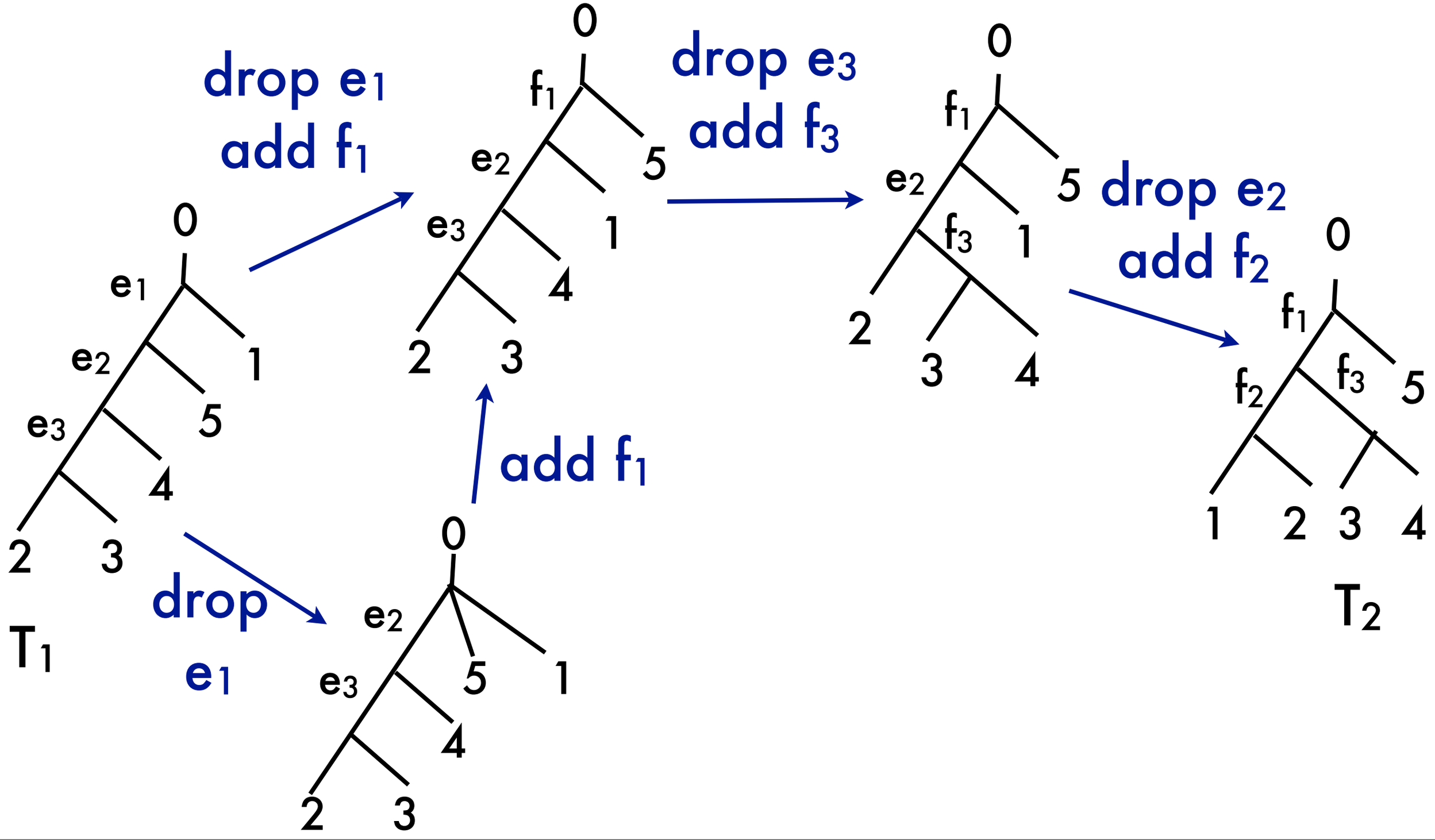
Lemma (Billera, Holmes, Vogtmann, 2001):
If e is a common edge, then every tree on the geodesic also contains e



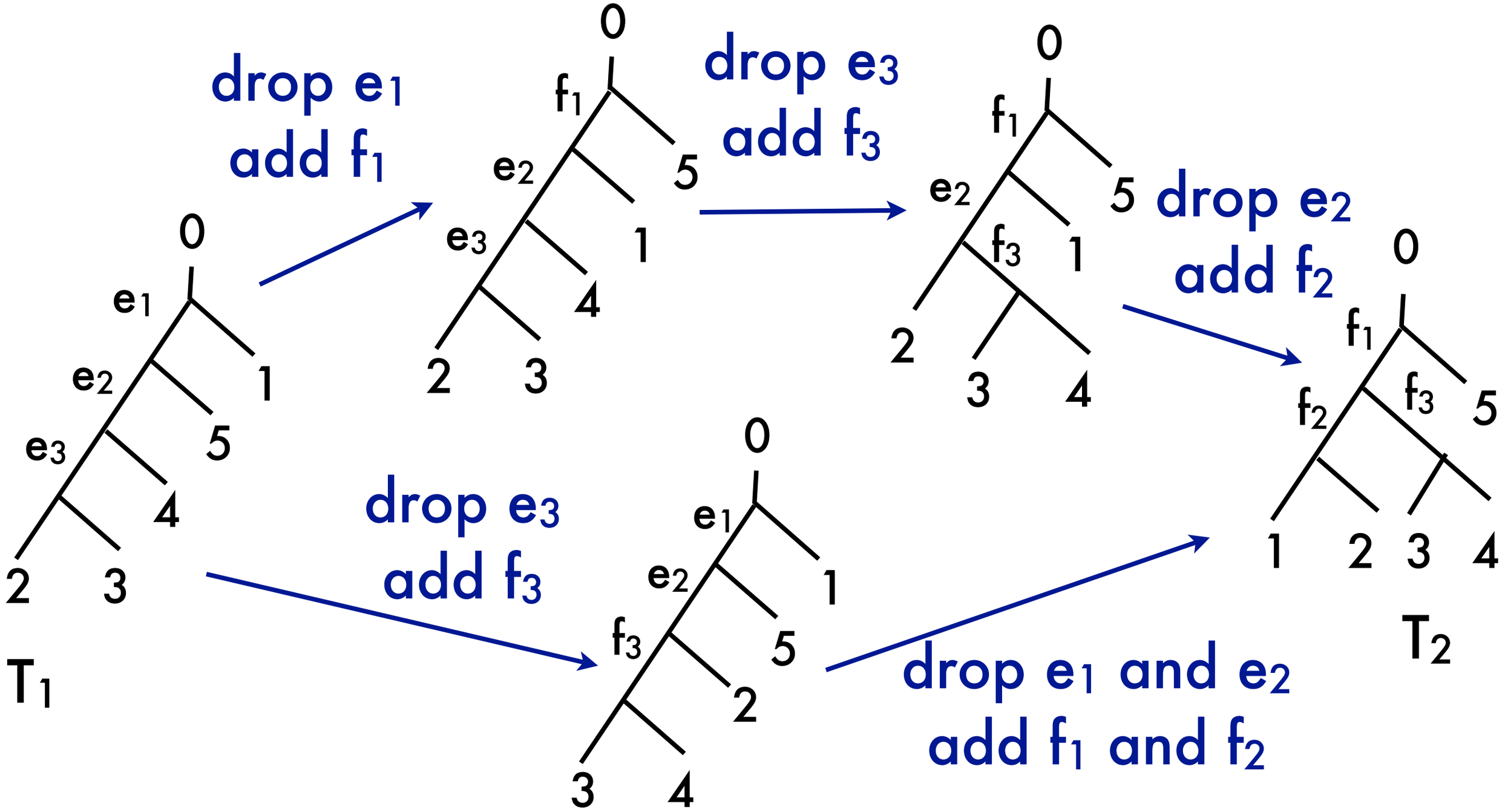
Geodesic Distance

- so can restrict problem to computing geodesic distance between two trees T_1 and T_2 with no common edges
- two previous exponential algorithms:
  - GeoMeTree (Kupczok et al., 2008)
  - GeodeMaps (O., 2011)
- $\sqrt{2}$- approx. algorithm (Amenta et al, 2007)

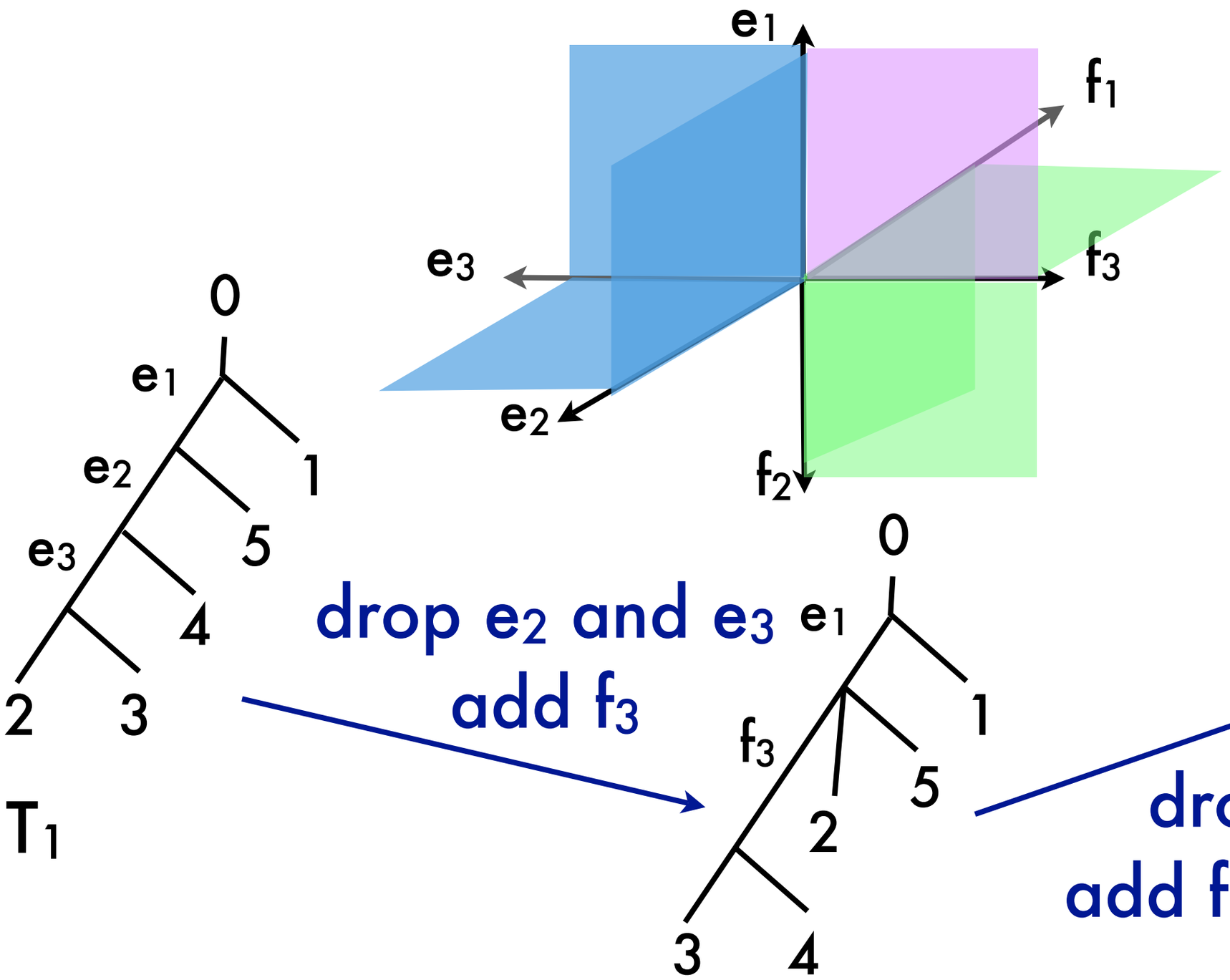
Geodesic Combinatorics



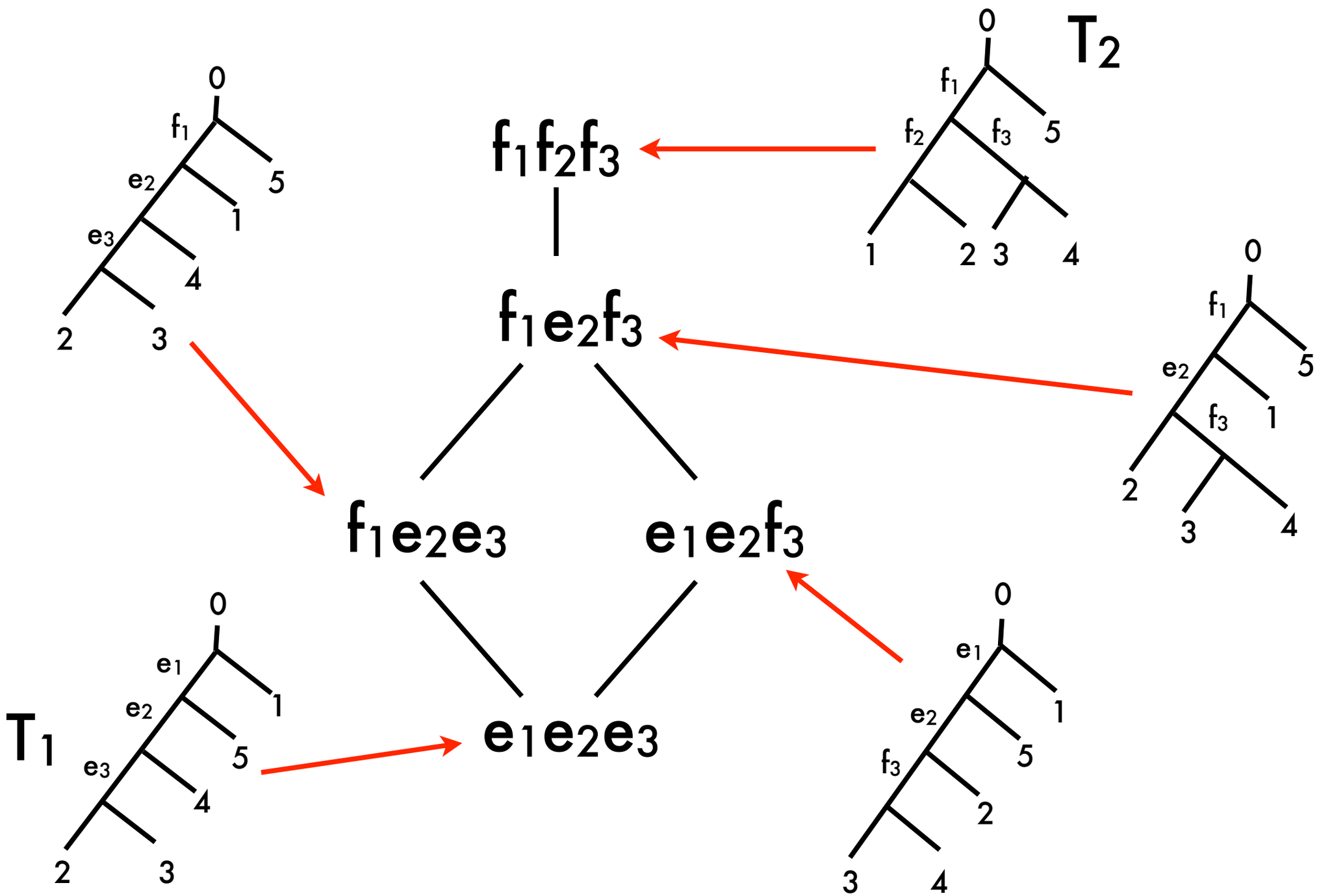
Geodesic Combinatorics



Path Spaces


e1
f1
e3
f3
e2
f2
0
e1
e2
1
5
e3
4
2
3
T1
drop e2 and e3
add f3
0
e1
1
f3
5
2
3
4
drop e1
add f1 and f2
0
f1
f2
f3
5
1
2
3
4
T2

Geodesic Combinatorics



Characterizing Geodesics

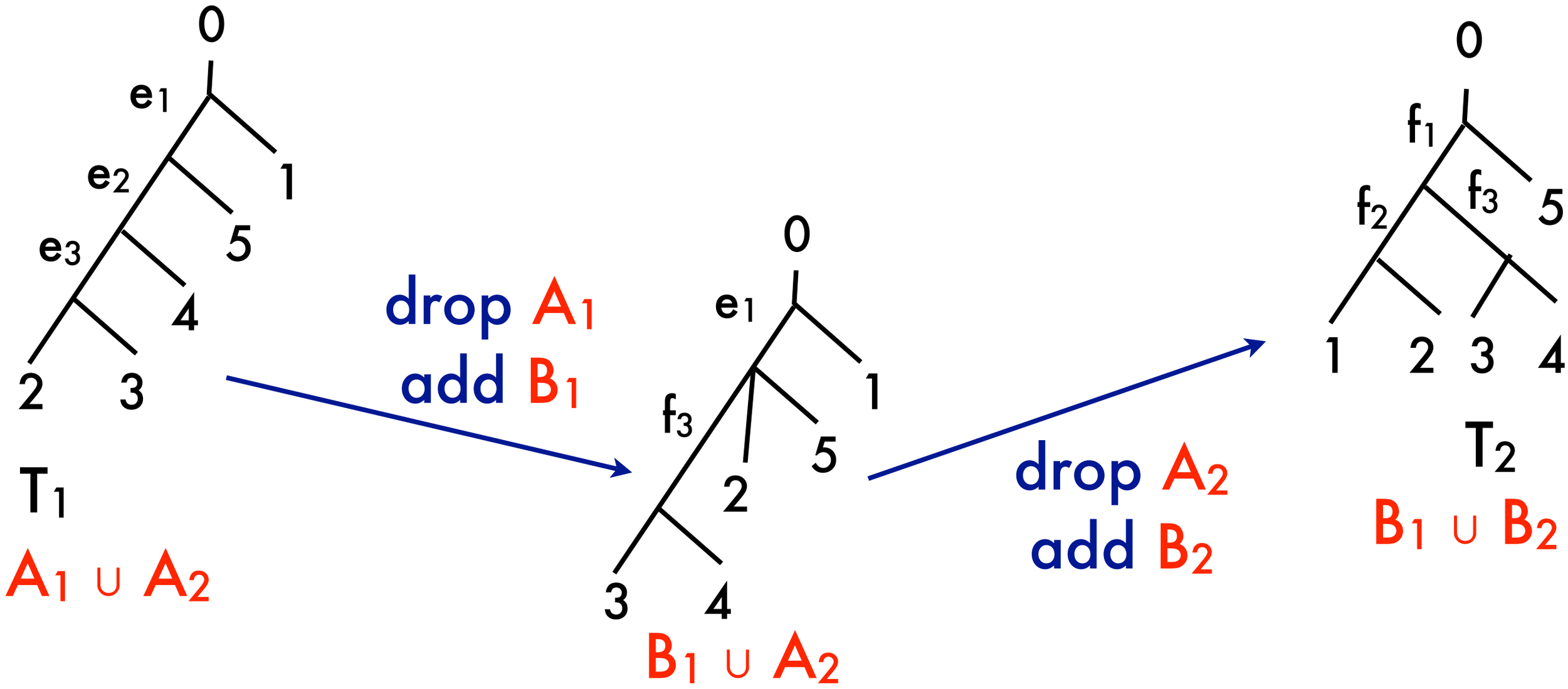
- at i^{th} transition between orthants:
  - edges A_i are dropped
  - edges B_i are added
- $(A_1, \ldots, A_k)$ partitions $E(T_1)$ and $(B_1, \ldots, B_k)$ partitions $E(T_2)$
- geodesic characterized by 3 properties
- Property 1:

  A_i and B_j compatible for all $i > j$

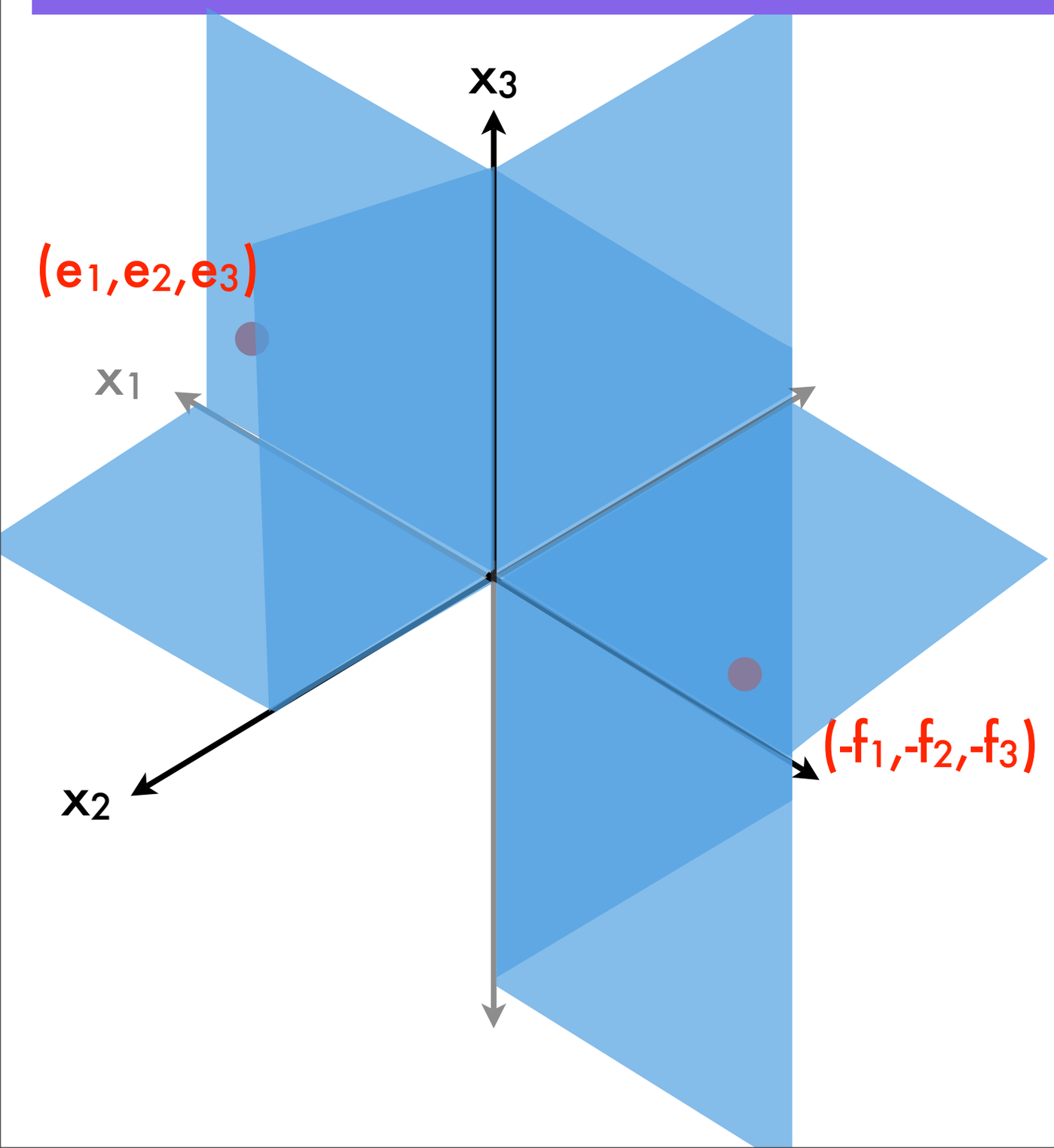
Property 1

- Property 1:

  A_i and B_j compatible for all i > j



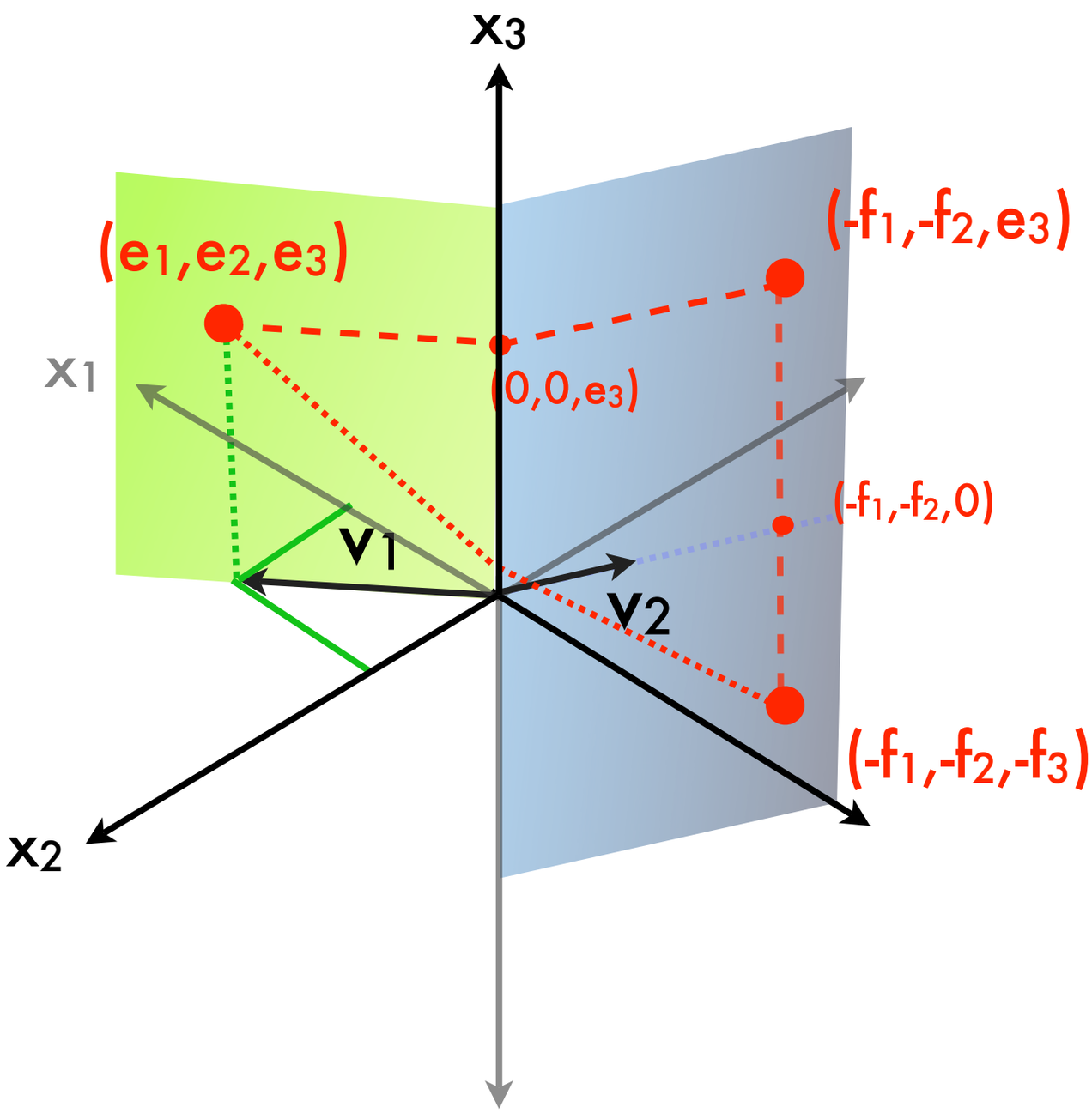
Property 2

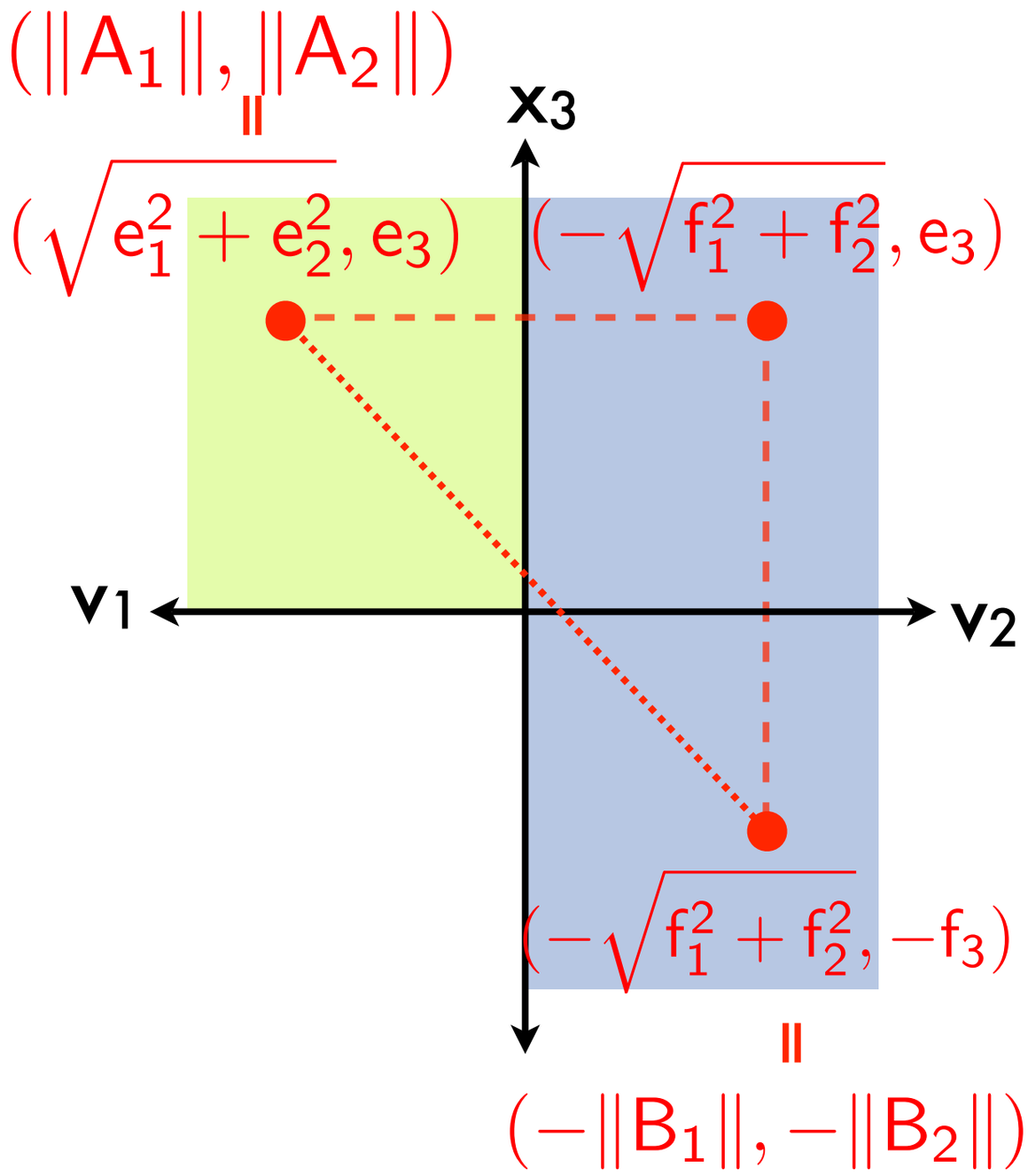


$A_1 = \{ e_1, e_2 \}$
$B_1 = \{ f_1, f_2 \}$

$A_2 = \{ e_3 \}$
$B_2 = \{ f_3 \}$

Isometric to part of $\mathbb{R}^k$





..... = geodesic

Property 2

- line from $(||A_1||,\ldots,||A_k||)$ to $(-||B_1||, \ldots, -||B_k||)$ is the geodesic in our region of $\mathbb{R}^k$ iff

$$\frac{||A_1||}{||B_1||} \leq \frac{||A_2||}{||B_2||} \leq \ldots \leq \frac{||A_k||}{||B_k||}$$

$\Rightarrow$ geodesic distance = Euclidean distance

$$= \sqrt{\sum_{i=1}^{k} ||A_i|| + ||B_i||}$$

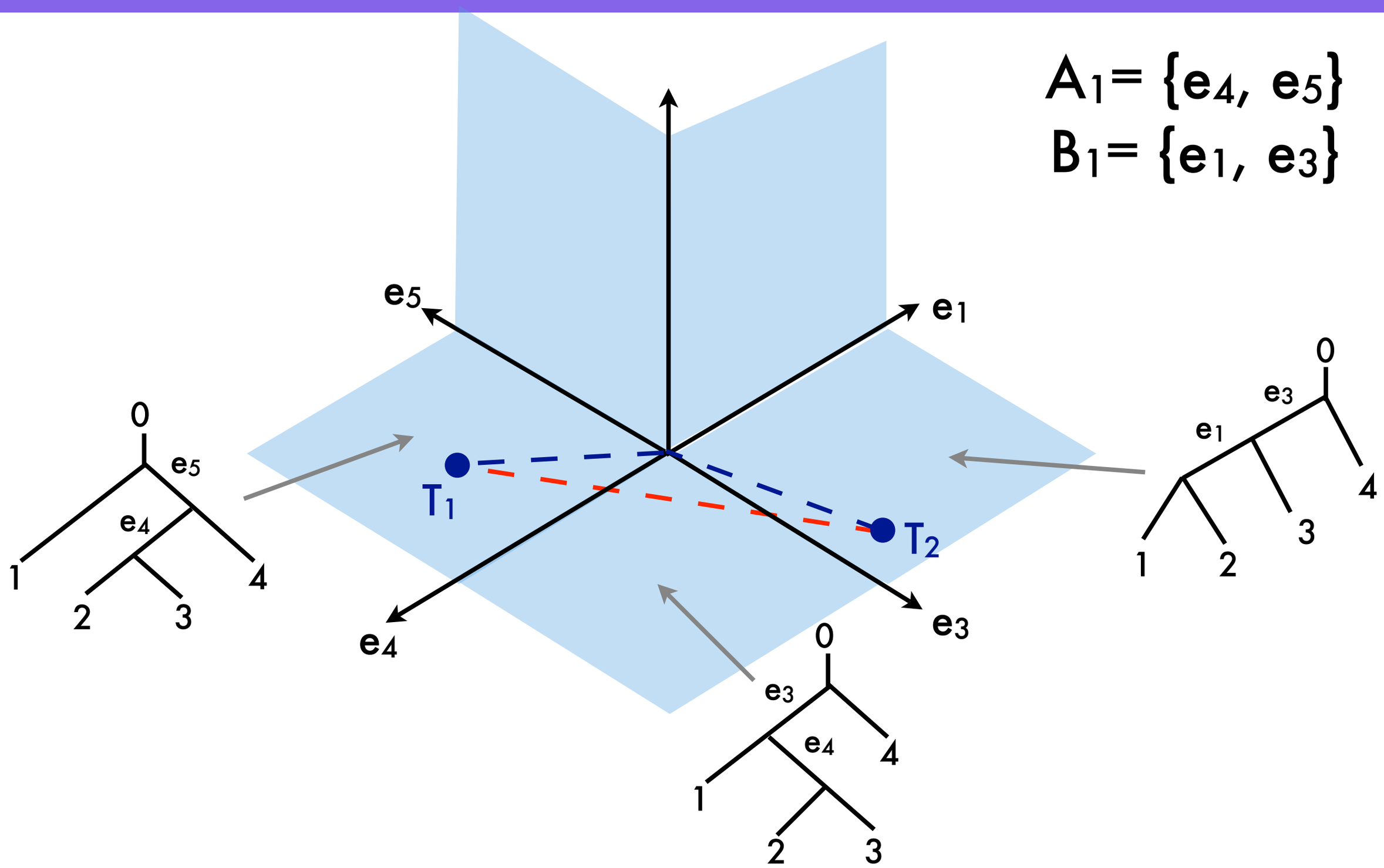
- Property 2:

$$\frac{||A_1||}{||B_1||} \leq \frac{||A_2||}{||B_2||} \leq \ldots \leq \frac{||A_k||}{||B_k||}$$

Property 3

$A_1 = \{e_4, e_5\}$

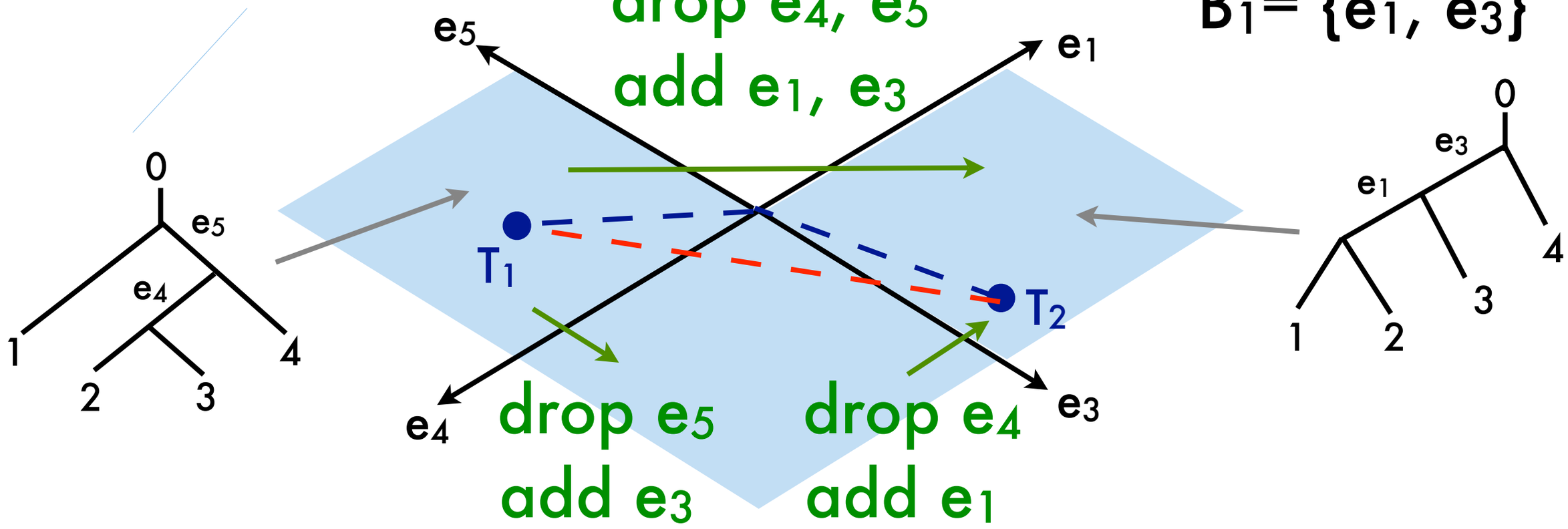
$B_1 = \{e_1, e_3\}$



Property 3

$A_1 = \{e_4, e_5\}$

$B_1 = \{e_1, e_3\}$

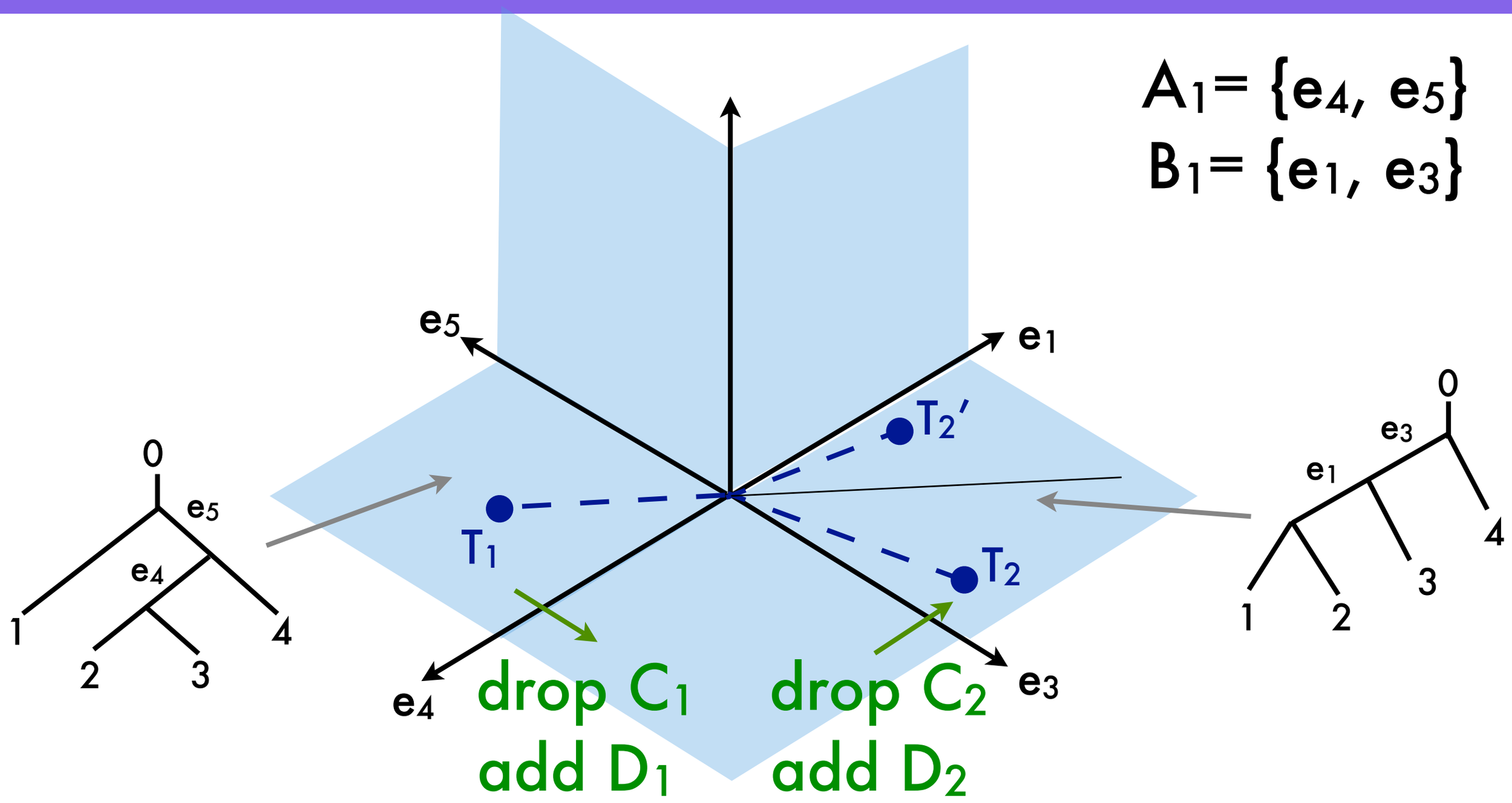


For (A_i, B_i), $\exists$ partition $C_1 \cup C_2$ of A_i, and partition $D_1 \cup D_2$ of B_i, such that C_2 is compatible with D_1.

Property 3

$A_1 = \{e_4, e_5\}$

$B_1 = \{e_1, e_3\}$



Want $\frac{\|C_1\|}{\|D_1\|} < \frac{\|C_2\|}{\|D_2\|}$

Property 3

- Property 3:

  For each pair (A_i, B_i), ∄ partition $C_1 \cup C_2$ of A_i, and partition $D_1 \cup D_2$ of B_i, such that C_2 is compatible with D_1 and $\frac{\|C_1\|}{\|D_1\|} < \frac{\|C_2\|}{\|D_2\|}$.

Theorem:

Partitions $(A_1, \ldots, A_k)$ and $(B_1, \ldots, B_k)$ represent the geodesic iff Properties 1, 2, and 3 hold.

Geodesic Algorithm

Initialize: $A_1 = E(T_1)$, $B_1 = E(T_2)$ (cone path)

P1 and P2 hold.

Iterative Step: P1 and P2 hold for $(A_1, \ldots, A_r)$ and $(B_1, \ldots, B_r)$.

Does (A_i, B_i) satisfies Property 3 for every i?

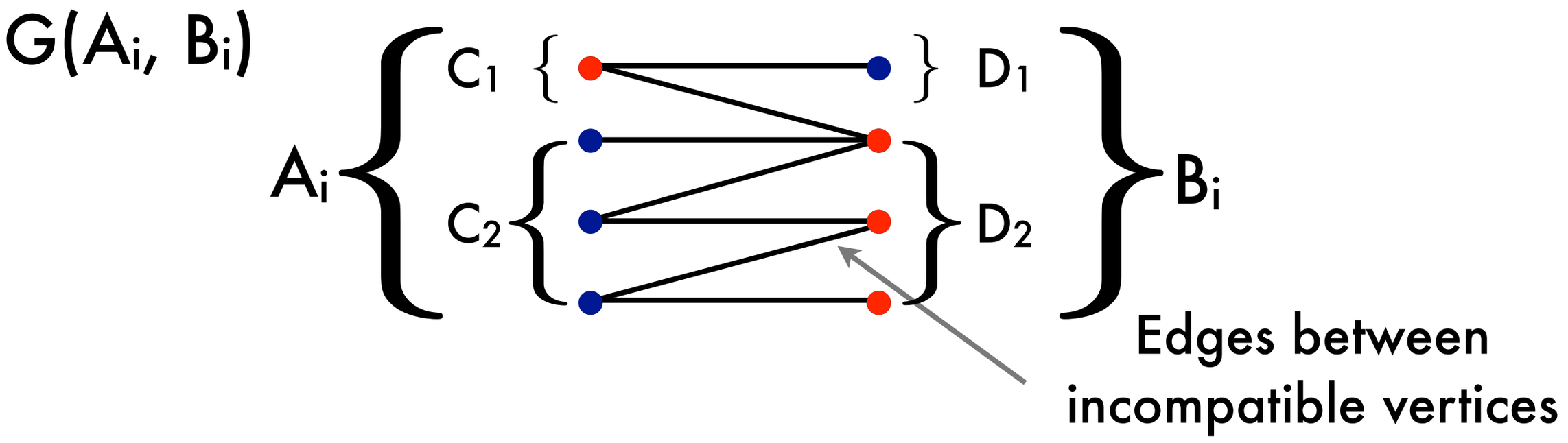
No: split blocks A_i and B_i, and re-index the new partition to get $(A_1, \ldots, A_{r+1})$ and $(B_1, \ldots, B_{r+1})$.

Yes: we are done.

Checking Property 3

- Property 3:

  For each pair (A_i, B_i), ∄ partition $C_1 \cup C_2$ of A_i, and partition $D_1 \cup D_2$ of B_i, such that C_2 is compatible with D_1 and $\frac{\|C_1\|}{\|D_1\|} < \frac{\|C_2\|}{\|D_2\|}$



Checking Property 3

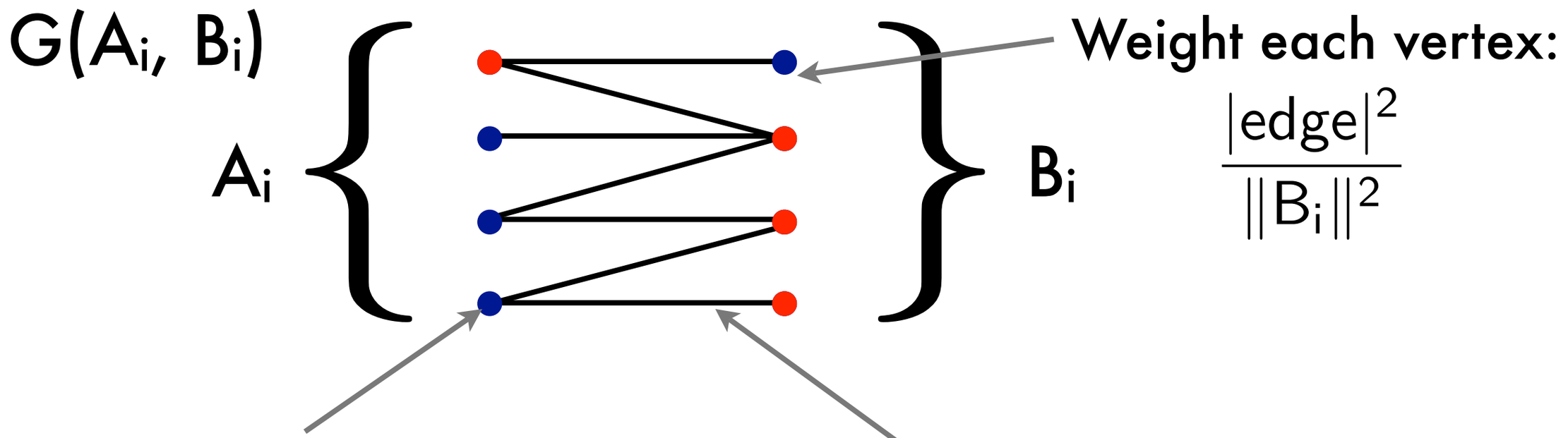
For each pair (A_i, B_i), $\nexists$ partition $C_1 \cup C_2$ of A_i, and partition $D_1 \cup D_2$ of B_i, such that C_2 is compatible with D_1 and $\frac{\|C_1\|}{\|D_1\|} < \frac{\|C_2\|}{\|D_2\|}$.

Can assume $\|A_i\| = \|B_i\| = 1$.

Then $$\frac{1 - \|C_2\|^2}{\|D_1\|^2} < \frac{\|C_2\|^2}{1 - \|D_1\|^2}$$

or $$\|C_2\|^2 + \|D_1\|^2 = \sum_{e \in C_2} |e|^2 + \sum_{f \in D_1} |f|^2 > 1$$

Property 3



We can add an orthant if there is a min. weight vertex cover of $G(A_i, B_i)$ with weight < 1.

Complexity: $O(n^3)$ (Solve as max. flow problem.)

Geodesic Algorithm

Initialize: A_1 = all splits of T_1, B_1 = all splits of T_2 (cone path)

Iterative Step: Current orthant sequence given by $(A_1, \ldots, A_r)$ and $(B_1, \ldots, B_r)$.

Does (A_i, B_i) satisfy the Shortcut Property for any i?

Yes: split blocks A_i and B_i, and re-index the new partition to get $(A_1, \ldots, A_{r+1})$ and $(B_1, \ldots, B_{r+1})$.

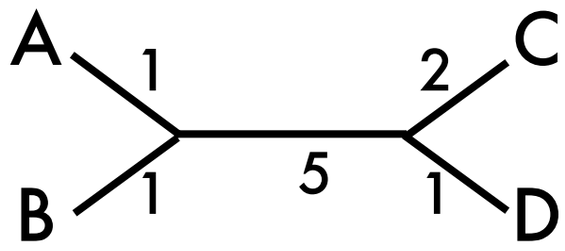
No: we are done.

Total time: $O(n^4)$

Open Problems

- BHV space: put L_1 metric on orthants, instead of L_2 metric
  - geodesic distance is the weighted Robinson-Foulds distance
  - geodesics are not unique so not CAT(0)
- what if put L_∞ metric on orthants?

OP: Tropical Tree Space



dissimilarity map

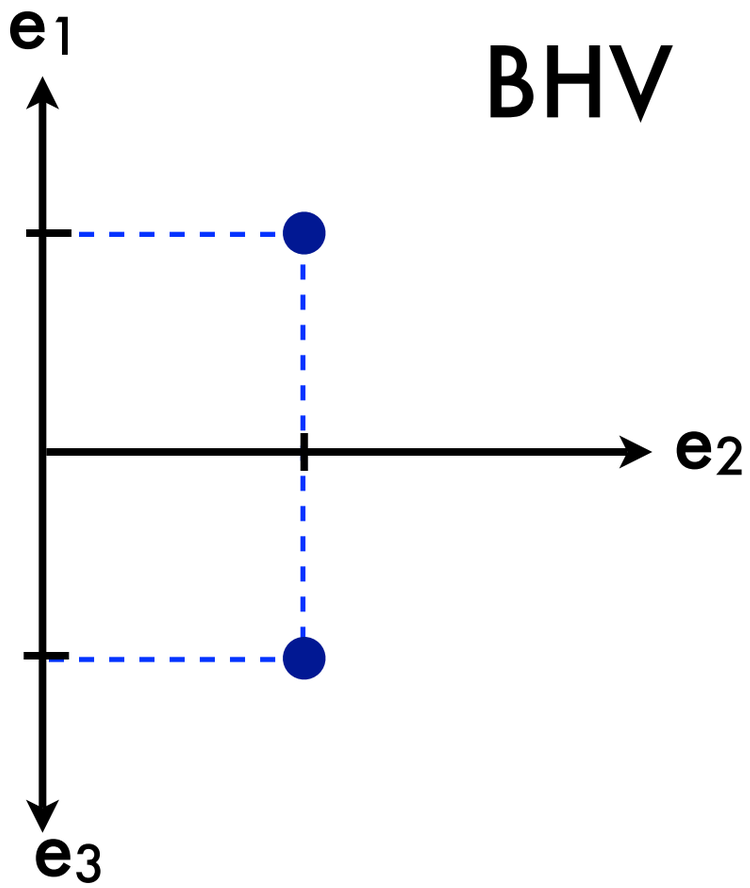
$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} - & 2 & 8 & 7 \\ 2 & - & 8 & 7 \\ 8 & 8 & - & 3 \\ 7 & 7 & 3 & - \end{bmatrix} \end{array}$$

- tropical tree space = set of dissimilarity maps in $\mathbb{R}^{\binom{n}{2}}$ that are realizable as trees
- open problems:
  - algorithm for computing geodesic
  - how often are geodesics not unique?

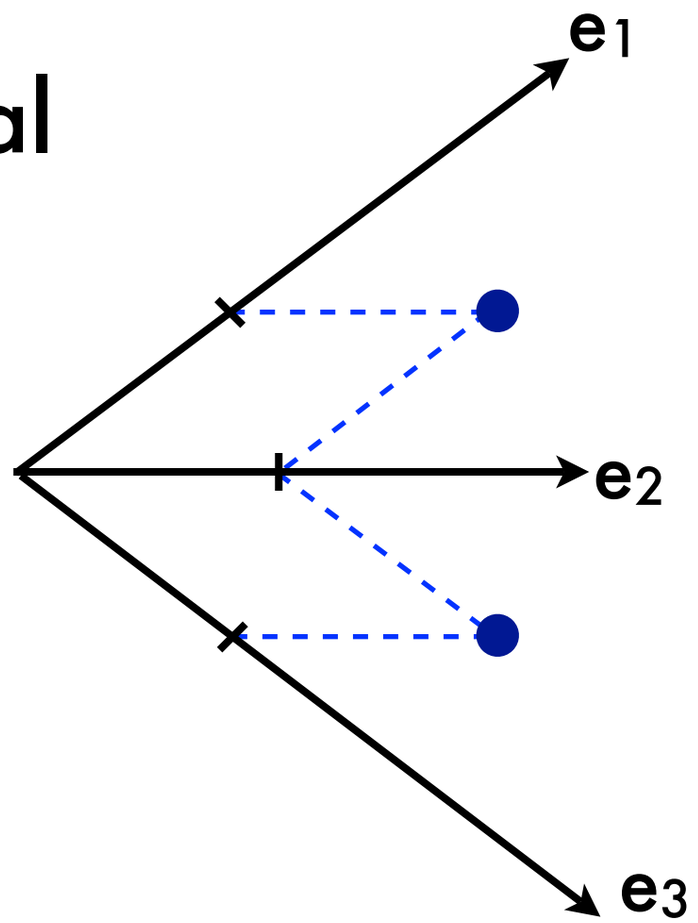
OP: Tropical Tree Space

- tropical and BHV versions of tree space:
  - same combinatorics
  - different geodesics

BHV

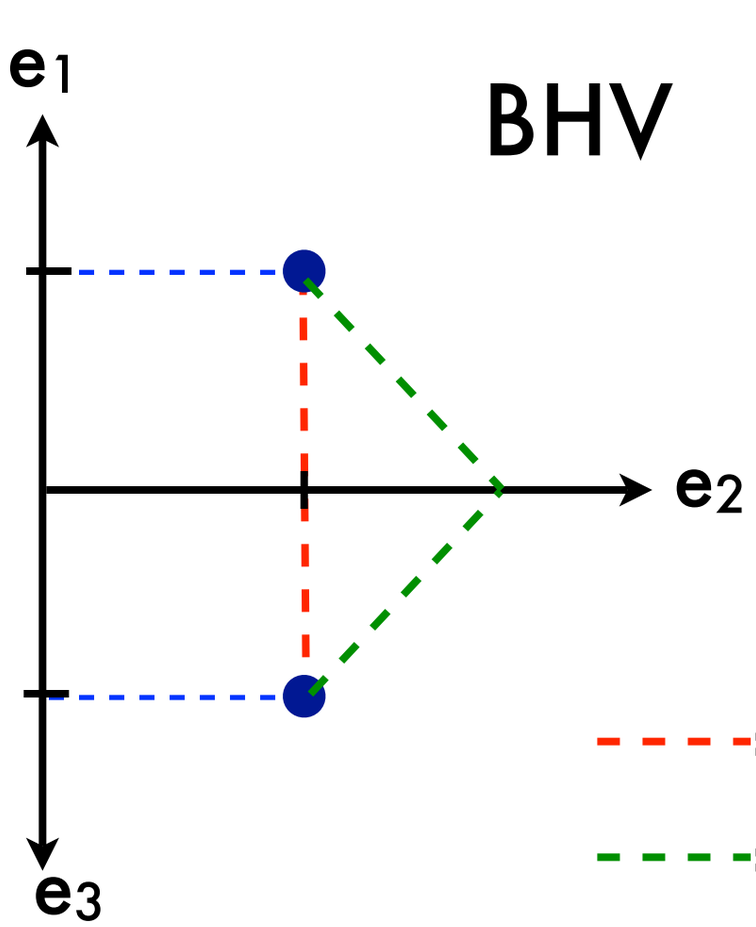


tropical



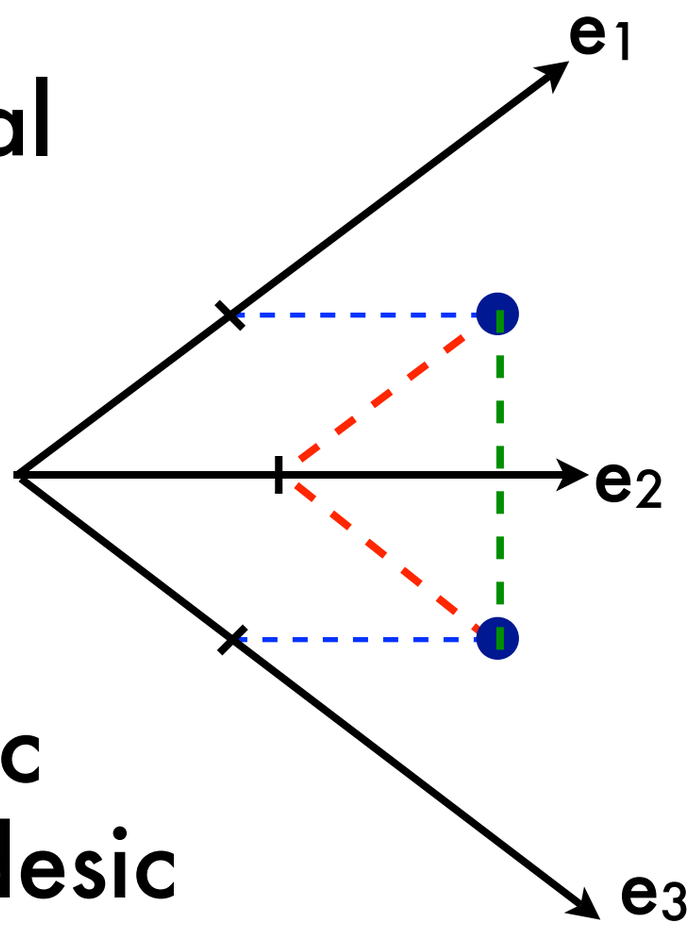
OP: Tropical Tree Space

- tropical and BHV versions of tree space:
  - same combinatorics
  - different geodesics



BHV

tropical



= BHV geodesic

= tropical geodesic

OP: Phylo Orange Space

- phylogenetic orange/edge-product space:
  - "compatification of BHV tree space at ∞"
  - all trees with all edge lengths ∞ are identified
    - space where trees are identified iff they induce the same Markov process on their leaves
- what is a natural metric for this space?
- properties of the space under this metric?
- how to compute distances?

OP: Other Tree Spaces

- space of unlabelled trees with n-leaves? (Feragen et al. 2010, 2011; Hultman 2007)
- space of trees with different, but overlapping taxa sets (i.e. for supertrees)
  - what about just one potentially missing taxon? (i.e. rogue taxon)
- space of phylogenetic networks?

OP: Visualization

- what is the best way to visualize a set of points in some tree space?

- Multi-Dimensional Scaling (Hillis, Heath, St. John, 2005)
- tree of trees (Nye, 2008; Chakerian and Holmes, 2010)

End of Part I

L. Billera, S. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27:733-767, 2001.

M. Owen and S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 8:2-13, 2011.

GTP code: http://www.stat-or.unc.edu/webspace/miscellaneous/provan/treespace