Dept. of Computer Science, University of Copenhagen

The geometry and statistics of geometric trees

Current Topic Workshop on Statistics, Geometry, and Combinatorics on Stratified Spaces arising from Biological Problems

Mathematical Biosciences Institute, Ohio, May 25 2012

Aasa Feragen

aasa@diku.dk

In collaboration with!

Marleen de Bruijne

Jens Petersen

Pechin Lo

CPH Lung imaging

Megan Owen

Can compute...

Francois Lauze

Mads Nielsen

Math and imaging

Asger Dirksen

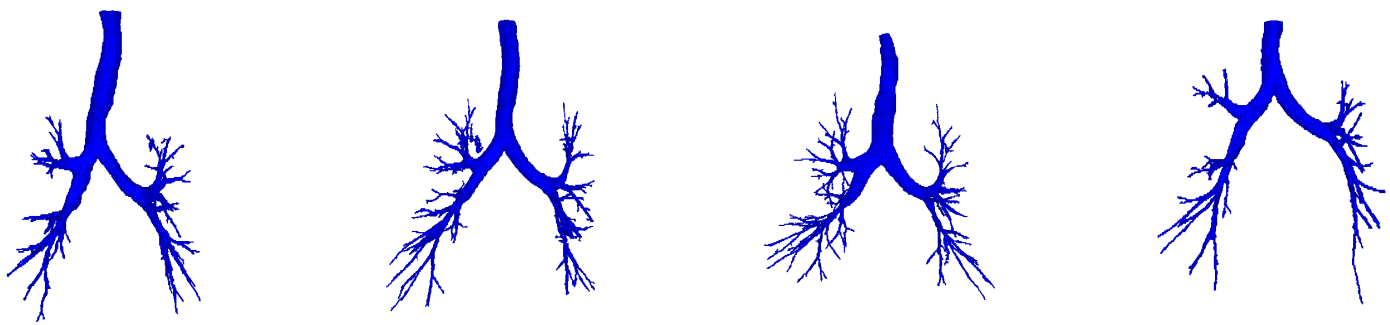
Laura Hohwü Thomsen

Mathilde Marie Winkler

The MDs!

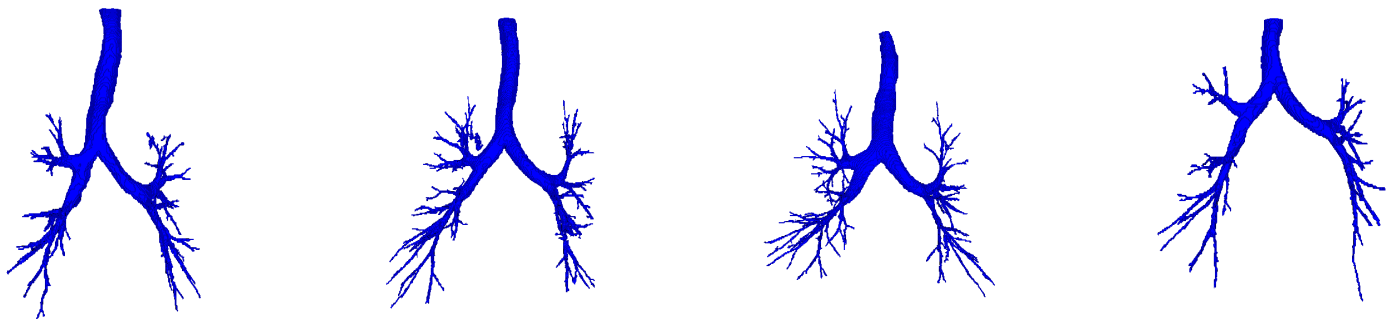
Airway shape modeling

Starting point: What does the average human airway tree look like?



Wanted: Parametric statistical model for trees, allowing variations in branch count, tree-topological structure and branch geometry

Airway shape modeling

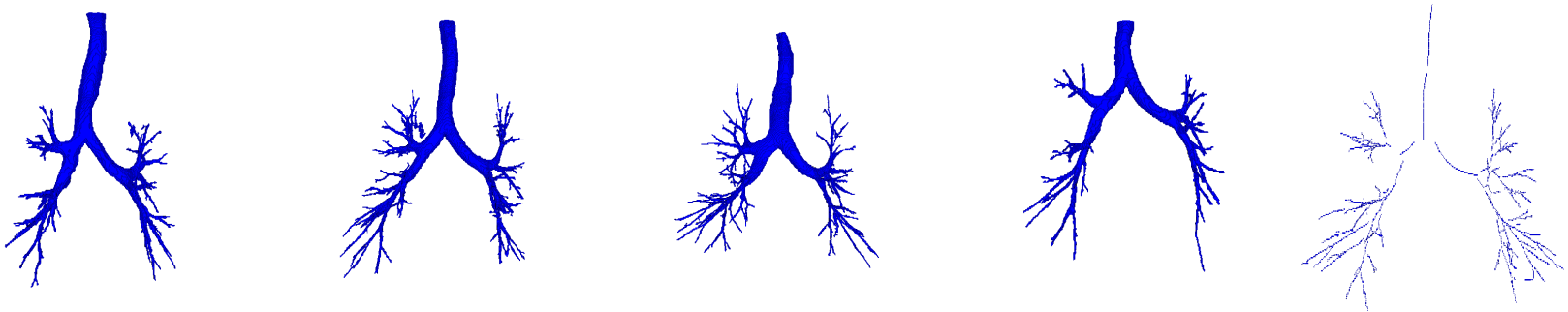


- Smoker's lung (COPD) is caused by inhaling damaging particles.
- Likely that damage made depends on airway geometry
- Reversely: COPD changes the airway geometry, e.g. airway wall thickness.
- $\rightsquigarrow$ Geometry can help diagnosis/prediction.

Airway shape modeling

Properties of airway trees:

- Topology, branch shape, branch radius
- Somewhat variable topology (combinatorics) in *anatomical* tree
- Substantial amount of noise in *segmented* trees (missing or spurious branches), especially in COPD patients,
  i.e. inherently incomplete data



Airway shape modeling

Wanted properties:

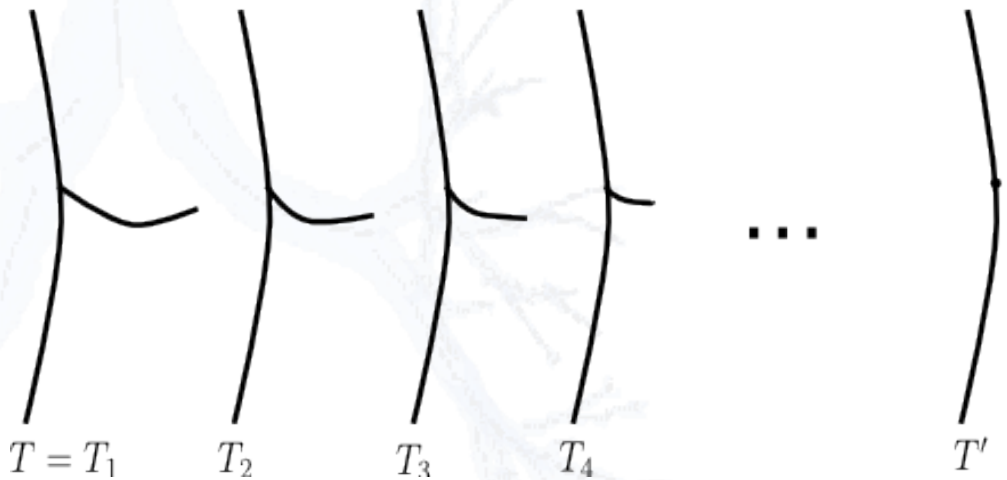


Figure: Tolerance of structural noise.

Airway shape modeling

Wanted properties:

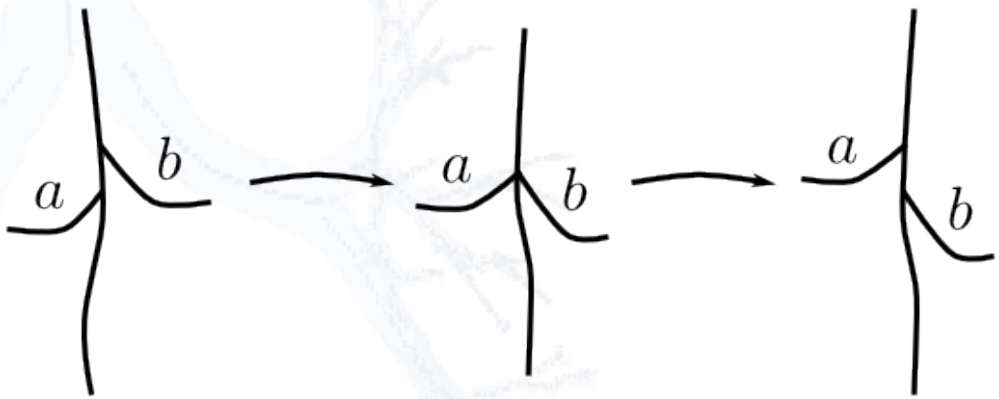
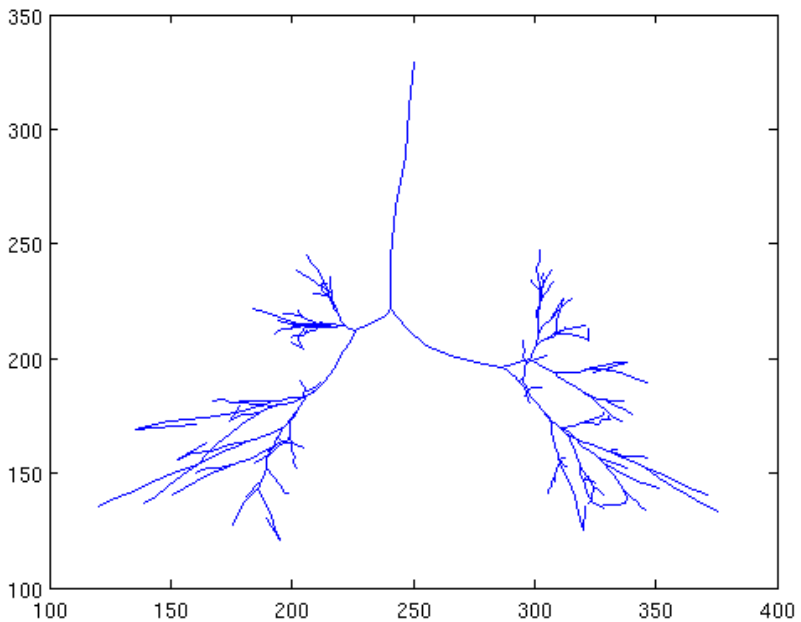


Figure: Handling of internal structural differences.

Airway shape modeling

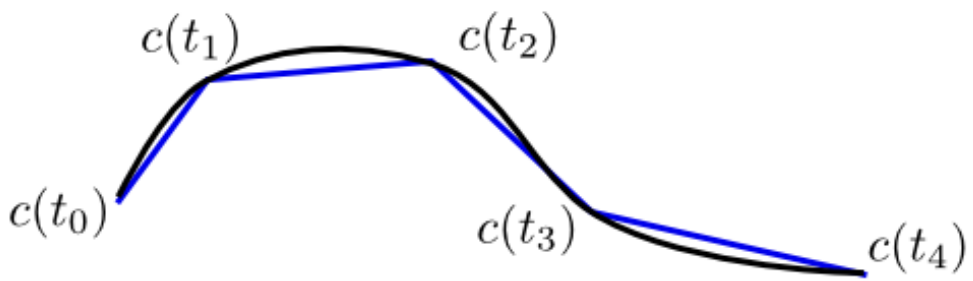
We shall consider airway centerline trees embedded in $\mathbb{R}^3$.



A little metric geometry – geodesics

- Let (X, d) be a metric space. The length of a curve $c : [a, b] \to X$ is

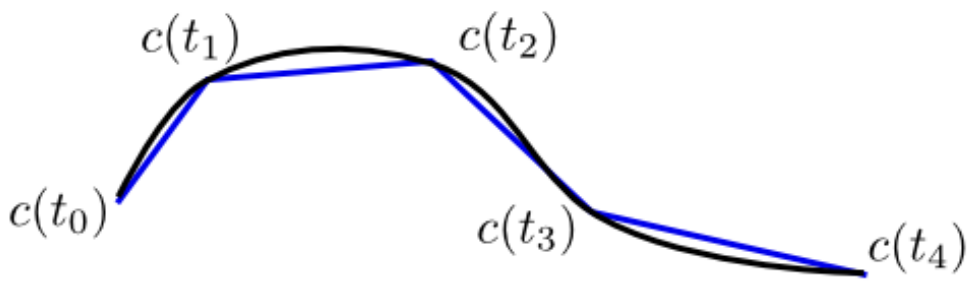
$$l(c) = sup_{a=t_0 \leq t_1 \leq \ldots \leq t_n = b} \sum_{i=0}^{n-1} d(c(t_i, t_{i+1})).$$



A little metric geometry – geodesics

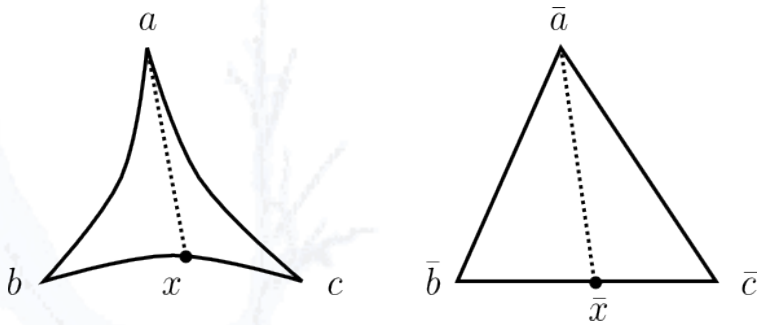
- Let (X, d) be a metric space. The length of a curve $c\colon [a, b] \to X$ is

$$l(c) = sup_{a=t_0 \leq t_1 \leq \ldots \leq t_n = b} \sum_{i=0}^{n-1} d(c(t_i, t_{i+1})).$$



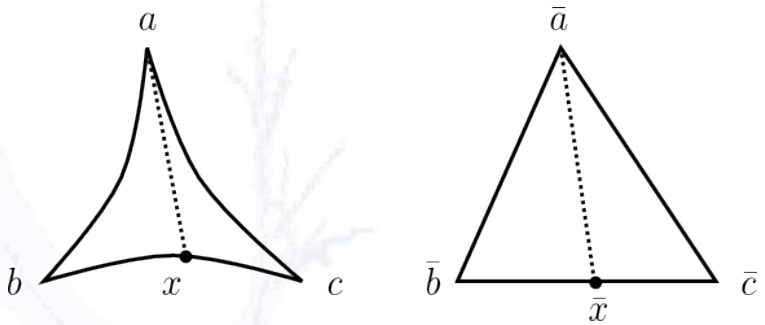
- A *geodesic* from x to y in X is a path $c\colon [a, b] \to X$ such that $c(a) = x, c(b) = y$ and $l(c) = d(x, y)$.
- (X, d) is a *geodesic space* if all pairs x, y can be joined by a geodesic.

Curvature in metric spaces



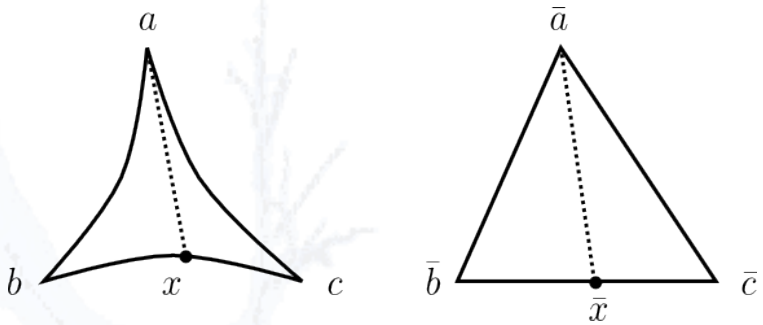
- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.

Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.
- A space has non-positive curvature if it is locally $CAT(0)$.

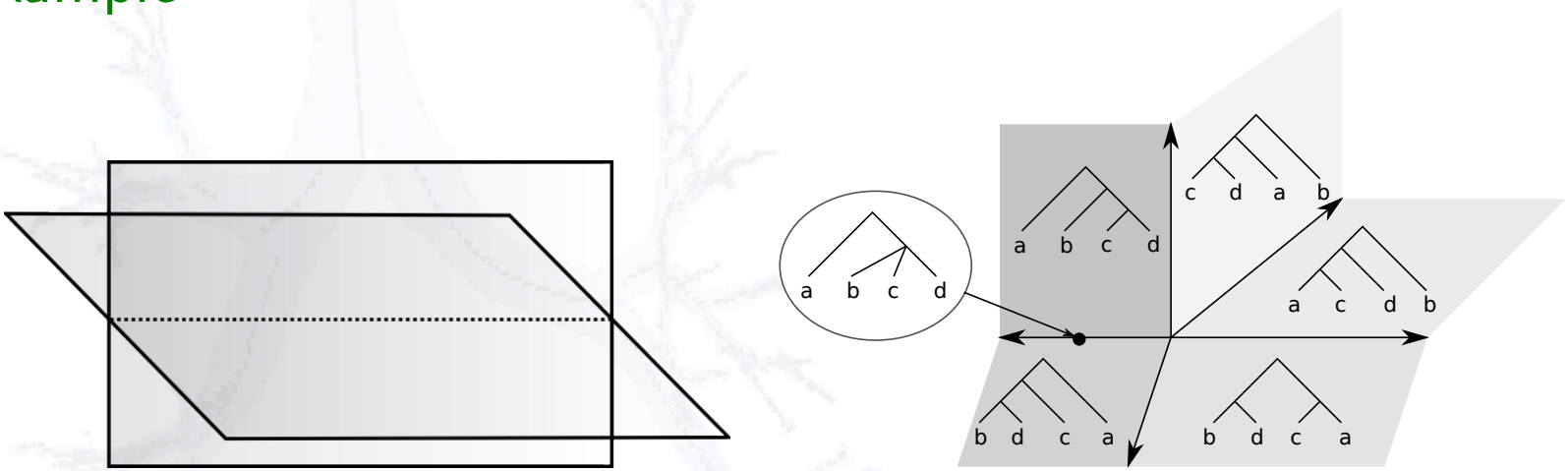
Curvature in metric spaces



- A $CAT(0)$ space is a metric space in which geodesic triangles are "thinner" than for their comparison triangles in the plane; that is, $d(x, a) \leq d(\bar{x}, \bar{a})$.
- A space has non-positive curvature if it is locally $CAT(0)$.
- (Similarly define curvature bounded by κ by using comparison triangles in hyperbolic space or spheres of curvature κ.)

Curvature in metric spaces

Example



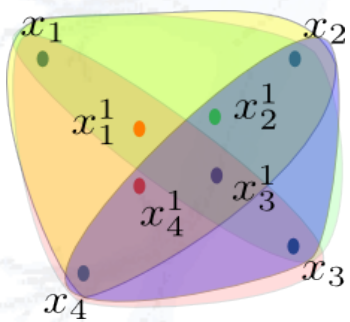
Theorem (see e.g. Bridson-Haefliger)

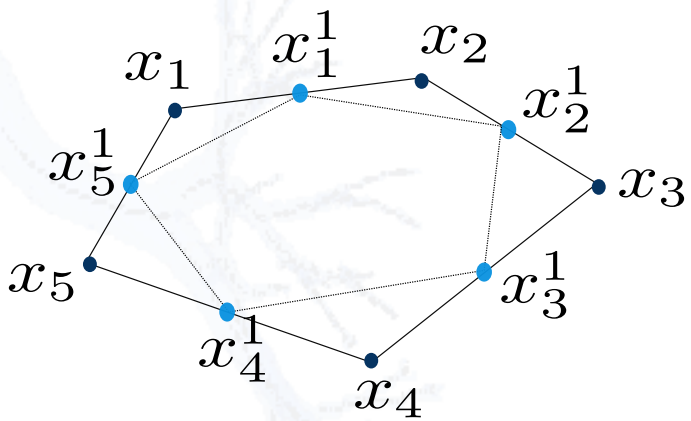
Let (X, d) be a $CAT(0)$ space; then all pairs of points have a unique geodesic joining them. The same holds locally in $CAT(\kappa)$ spaces, $\kappa \neq 0$.

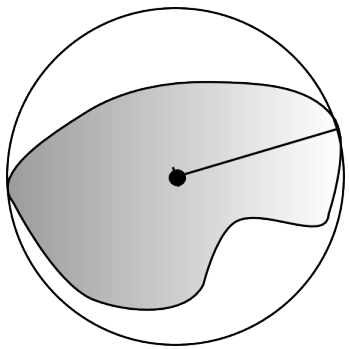
Statistics in metric spaces?

Theorem (Sturm)

Frechet means are unique in $CAT(0)$ spaces.





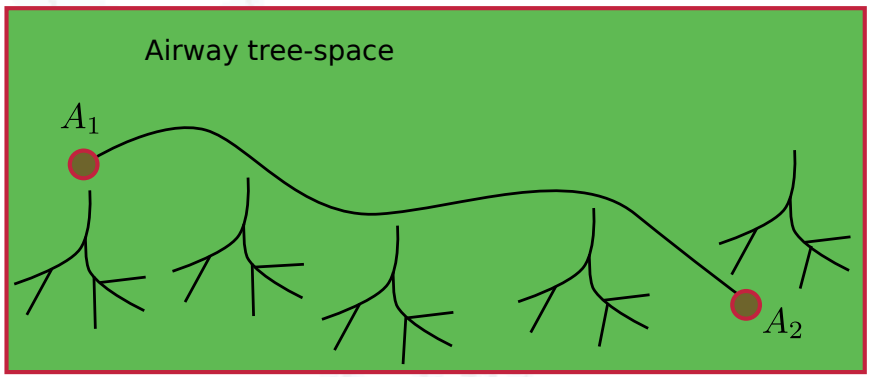


Other midpoint-finding algorithms also converge in CAT(0) spaces:

- centroid (BHV)
- Birkhoff shortening
- circumcenters

Plan of action?

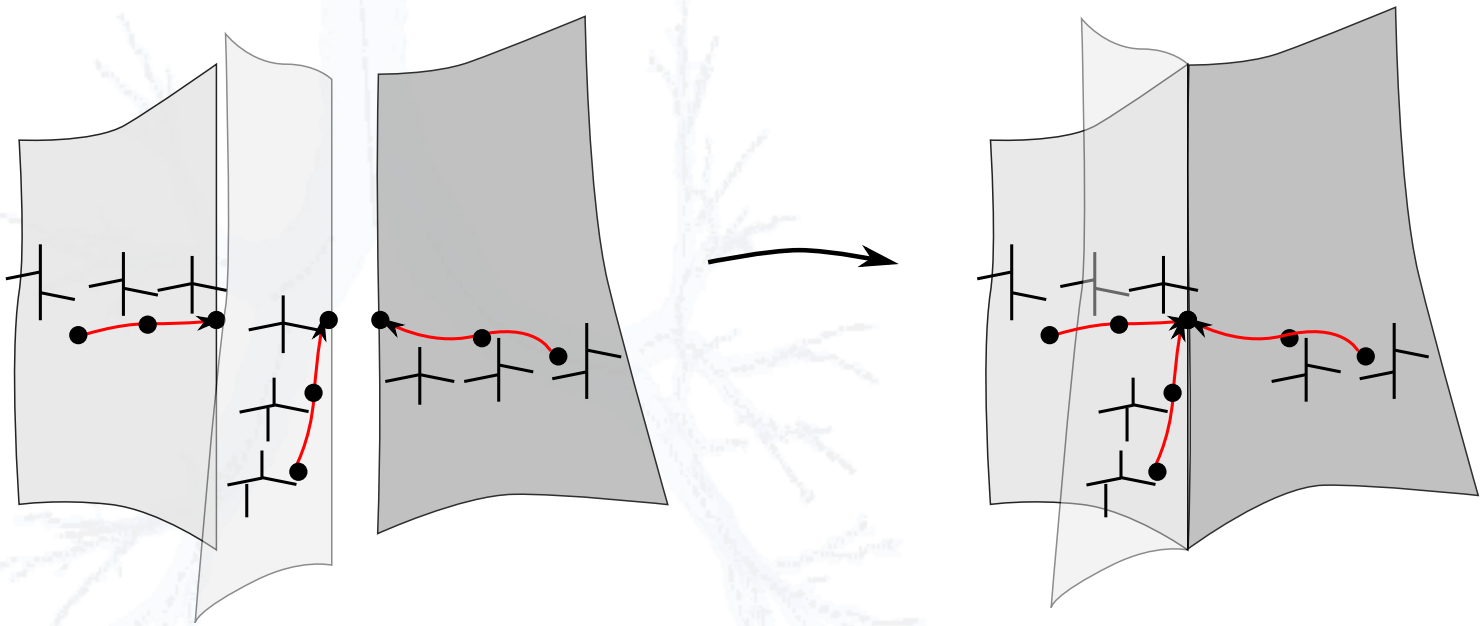
- Tree representation where size, topology and edge geometry can be consistently and simultaneously compared
- **Geodesic** metric tree-space!



- Do statistics in this space ($CAT(0)$-ish space?)

A space of tree-like shapes: Intuition

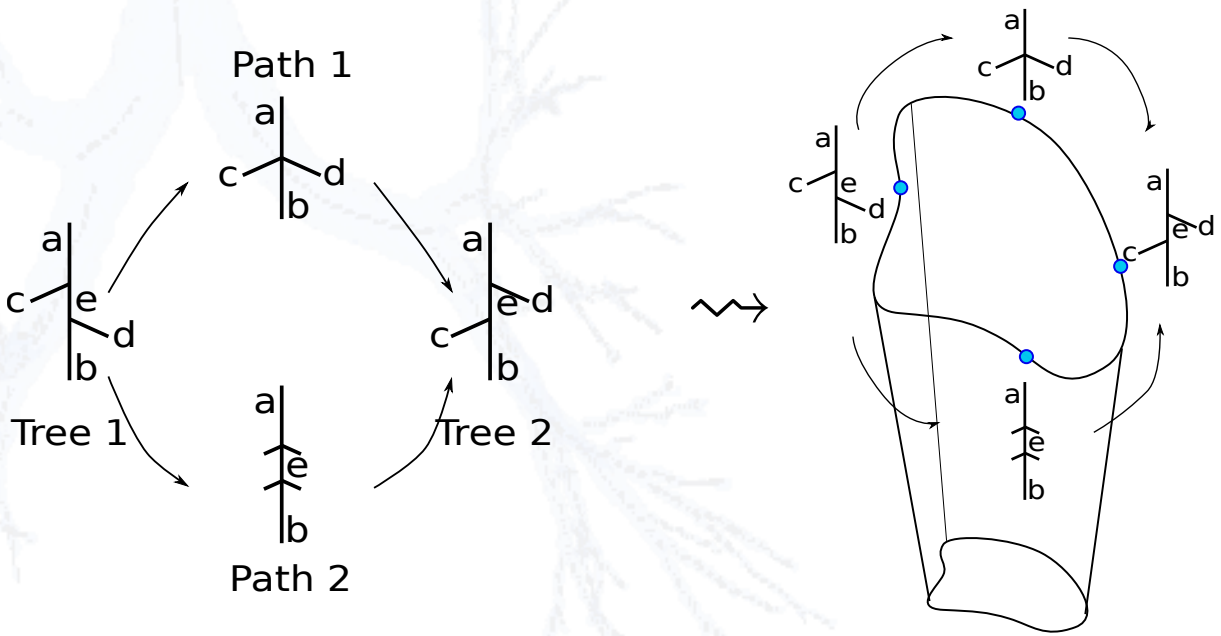
What would a path-connected space of deformable trees look like?



- Easy: Trees with same topology in their own "component"
- Harder: How are the components connected?
- Solution: glue collapsed trees, deforming one topology to another
- $\rightsquigarrow$ Stratified space, self intersections

A space of tree-like shapes: Intuition

The tree-space has conical "bubbles"

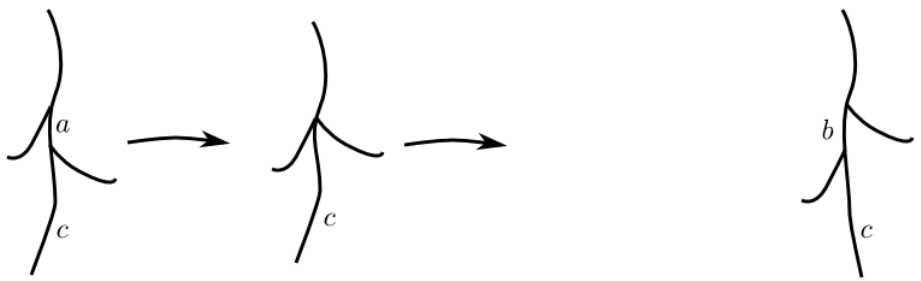


Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a nonlinear metric space
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.

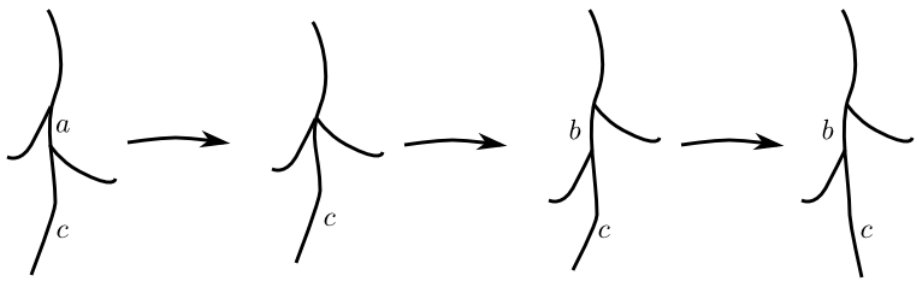
Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a nonlinear metric space
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



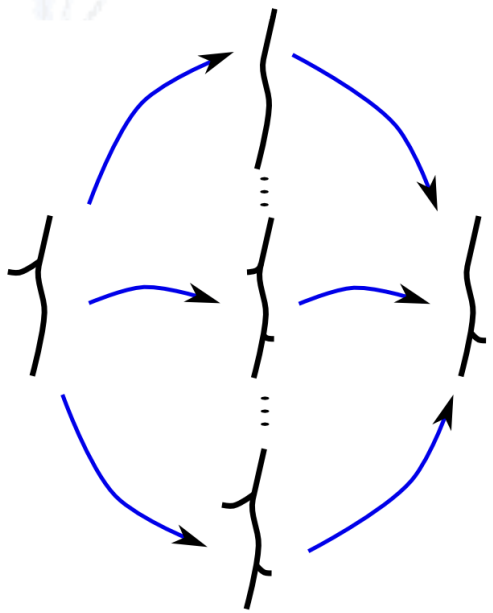
Classical example: Tree edit distance (TED)

- TED is a classical, algorithmic distance
- tree-space with TED is a nonlinear metric space
- dist(T_1, T_2) is the minimal total cost of changing T_1 into T_2 through three basic operations:
- Remove edge, add edge, deform edge.



Classical example: Tree edit distance (TED)

- Tree-space with TED is a geodesic space, but almost all geodesics between pairs of trees are non-unique (infinitely many).



- Then what is the average of two trees? Many!
- Tree-space with TED has everywhere unbounded curvature.
- TED is *not* suitable for statistics.

Classical example: Tree edit distance (TED)

Many state-of-the-art approaches to distance measures and statistics on tree- and graph-structured data *are* based on TED!

- Ferrer, Valveny, Serratosa, Riesen, Bunke: Generalized median graph computation by means of graph embedding in vector spaces. Pattern Recognition 43 (4), 2010.
- Riesen and Bunke: Approximate Graph Edit Distance by means of Bipartite Graph Matching. Image and Vision Computing 27 (7), 2009.
- Trinh and Kimia, Learning Prototypical Shapes for Object Categories. CVPR workshops 2010.

Classical example: Tree edit distance (TED)

The problems can be "solved" by choosing specific geodesics.
OBS! Geometric methods can no longer be used for proofs, and one risks choosing problematic paths.

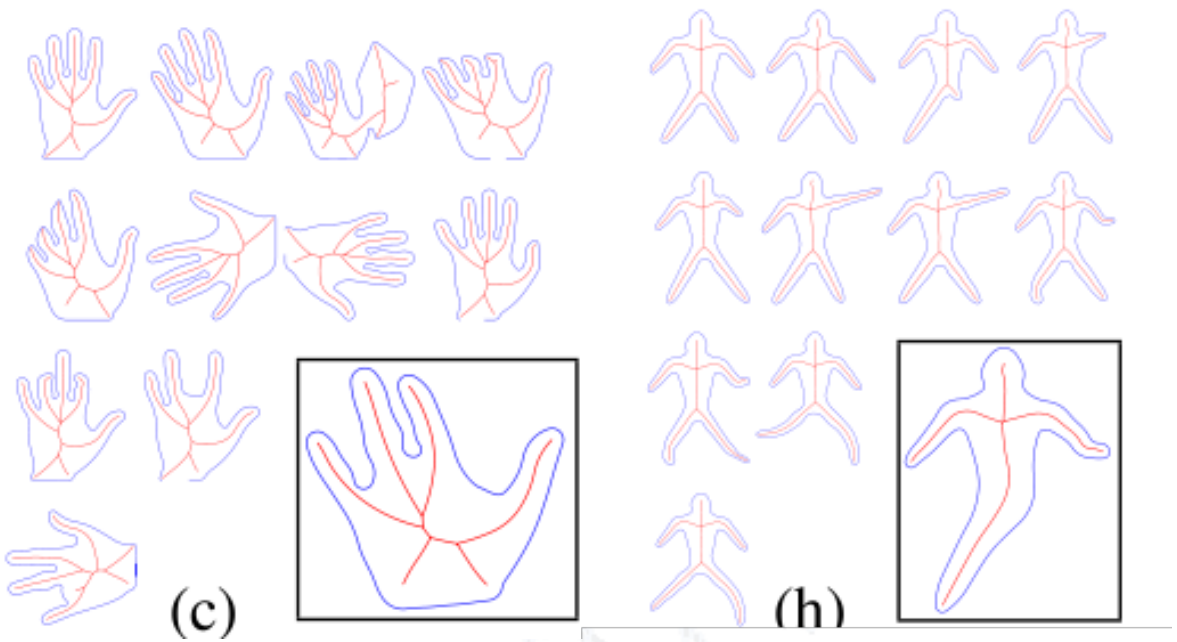


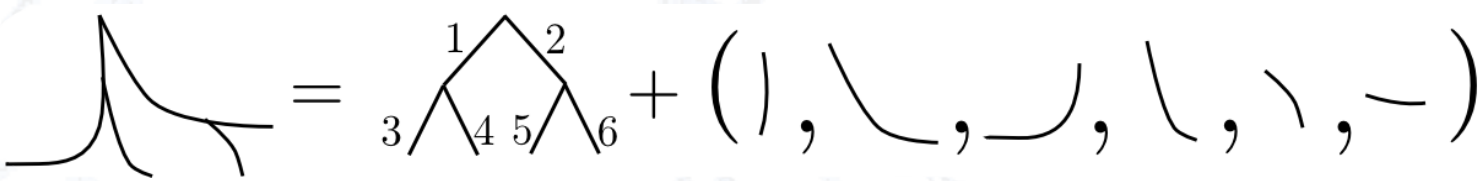
Figure: Trinh and Kimia (CVPR workshops 2010) compute average shock graphs using TED with the simplest possible choice of geodesics.

Build a tree-space: Tree representation

How to represent geometric trees mathematically?

Tree-like (pre-)shape = pair $(\mathscr{T}, x)$

- $\mathscr{T} = (V, E, r, <)$ rooted, ordered/planar binary tree, describing the tree topology (combinatorics)
- $x \in \bigoplus_{e \in E} A$, each coordinate in an attribute space A describing edge shape



Build a tree-space: Tree representation

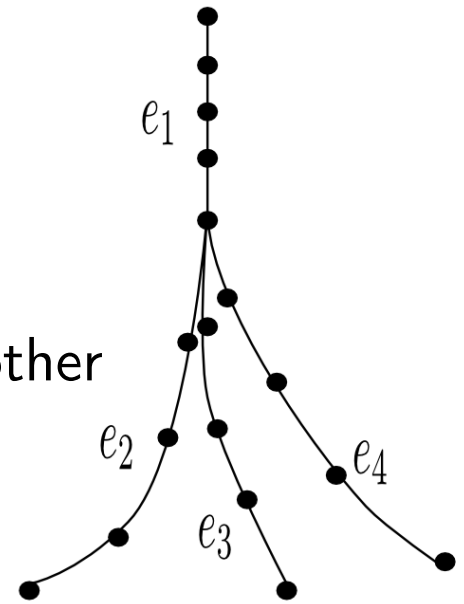
We are allowing collapsed edges, which means that

- we can represent higher order vertices
- we can represent trees of different sizes using the same combinatorial tree $\mathscr{T}$

(dotted line = collapsed edge = zero/constant attribute)

Build a tree-space: Tree representation

- Edge representation through landmark points:
- Edge shape space is $(\mathbb{R}^d)^n$, $d = 2, 3$.
- (For most results, this can be generalized to other vector spaces)



The space of tree-like preshapes

First: $\mathscr{T}$ an infinite, ordered (planar), rooted binary tree

Definition

Define the space of tree-like *pre*-shapes as the direct sum

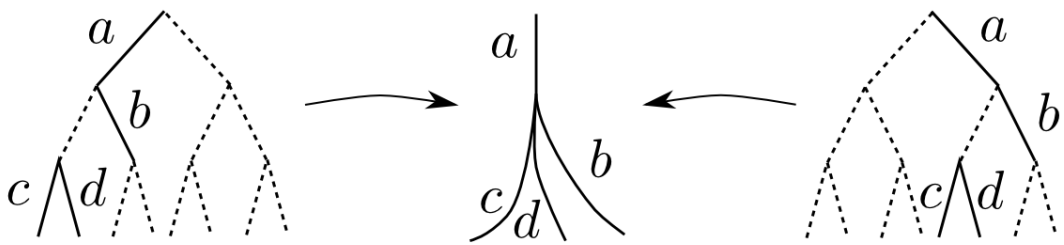
$$\bigoplus_{e \in E} (\mathbb{R}^d)^n$$

where $(\mathbb{R}^d)^n$ is the edge shape space.

This is just a space of *pre-shapes*.

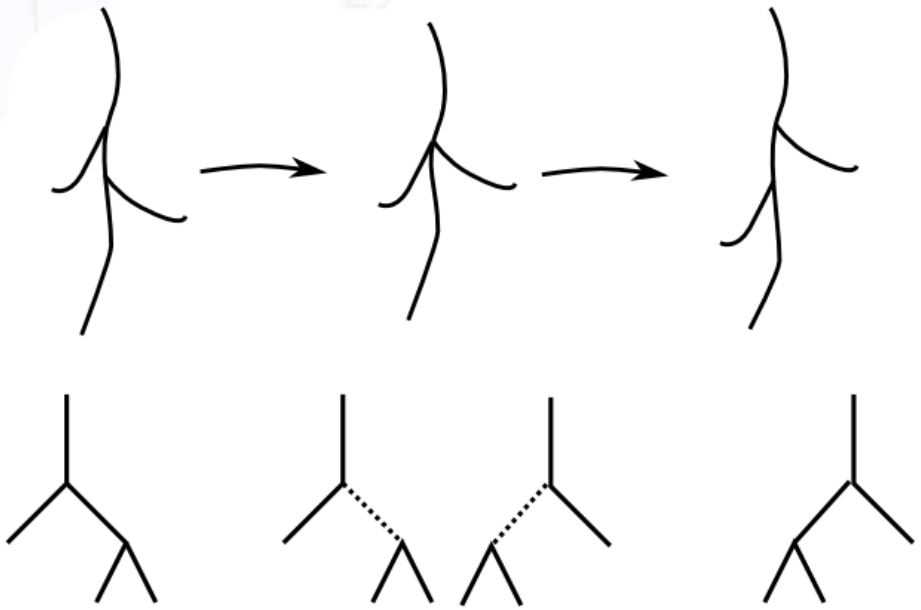
From pre-shapes to shapes

Many shapes have more than one representation



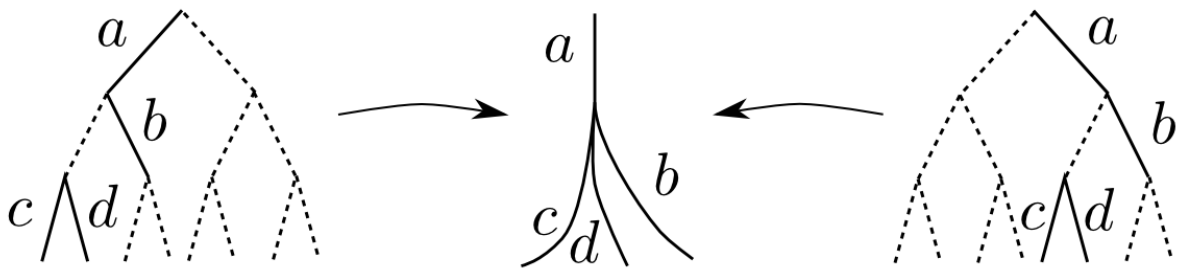
From pre-shapes to shapes

Not all shape deformations can be recovered as natural paths in the pre-shape space:



Shape space definition

- Start with the pre-shape space $X = \bigoplus_{e \in E} (\mathbb{R}^d)^n$.
- Define an equivalence $\sim$ by identifying points in X that represent the same tree-shape.

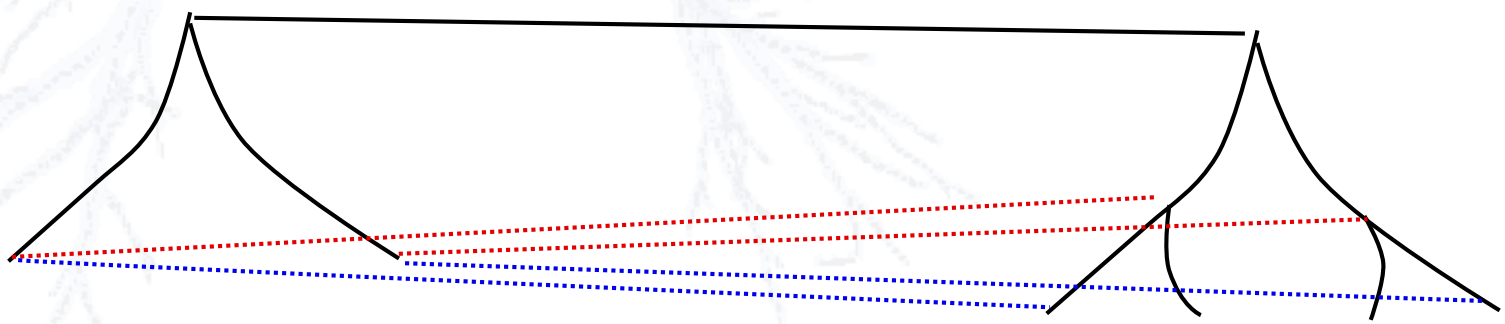


- This corresponds to identifying, or gluing together, subspaces $\{x \in X | x_e = 0 \text{ if } e \notin E_1\}$ and $\{x \in X | x_e = 0 \text{ if } e \notin E_2\}$ in X.
- The space of ordered (planar) tree-like shapes $\bar{X} = X/\sim$ is a folded vector space.

Shape space definition

Remark

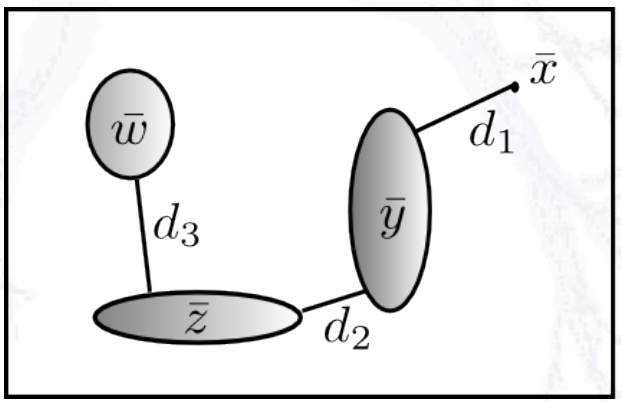
- Tree-shape definition a little unorthodox: we do not factor out scale and rotation of the tree.
- Our data (segmented airway trees) are incomplete; the number of segmented branches is unstable and depends on the health of the patient.



Definition of metric on tree-space

Given a metric d on the vector space $X = \bigoplus_{e \in E} (\mathbb{R}^d)^n$ we define the quotient pseudometric $\bar{d}$ on the quotient space $\bar{X} = X/\sim$ by setting

$$\bar{d}(\bar{x}, \bar{y}) = \inf \left\{ \sum_{i=1}^{k} d(x_i, y_i) | x_1 \in \bar{x}, y_i \sim x_{i+1}, y_k \in \bar{y} \right\}. \qquad (1)$$



$$\bar{d}(\bar{x}, \bar{w}) = d_1 + d_2 + d_3$$

Theorem

The quotient pseudometric $\bar{d}$ is a metric on $\bar{X}$. □

Definition of metric on tree-space

- Two metrics on $\bar{X}$ from two product norms on $X = \bigoplus_{e \in E} (\mathbb{R}^d)^n$:

$$\begin{aligned} l1 \text{ norm:} \quad & d_1(x, y) = \sum_{e \in E} \|x_e - y_e\| \\ l2 \text{ norm:} \quad & d_2(x, y) = \sqrt{\sum_{e \in E} \|x_e - y_e\|^2} \end{aligned}$$

- $\bar{d}_1$ = Tree Edit Distance
- Terminology: $\bar{d}_2$ = QED (Quotient Euclidean Distance) metric.

Theorem

Let $\bar{d} = \bar{d}_1$ or $\bar{d}_2$. Then $(\bar{X}, \bar{d})$ is a geodesic space. □

Unordered trees

- Give each tree a random order
- Denote by G the group of reorderings of the edges (in $\mathscr{T}$) that do not alter the connectivity of the tree.
- The space of spatial/unordered trees is the space $\bar{\bar{X}} = \bar{X}/G$
- Give $\bar{\bar{X}}$ the quotient pseudometric $\bar{\bar{d}}$.
- $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$ chooses the order that minimizes $\bar{\bar{d}}(\bar{\bar{x}}, \bar{\bar{y}})$.

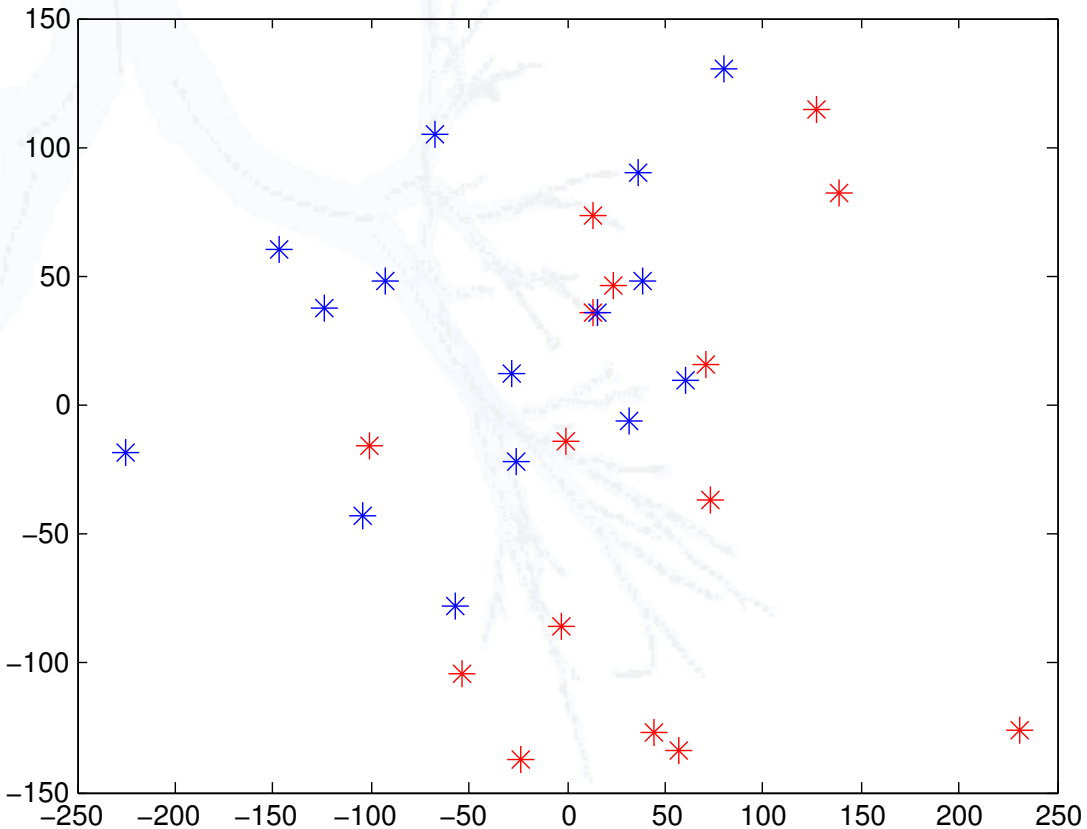
Theorem

For the quotient pseudometric $\bar{\bar{d}}$ induced by either $\bar{d}_1$ or $\bar{d}_2$, the function $\bar{\bar{d}}$ is a metric and $(\bar{\bar{X}}, \bar{\bar{d}})$ is a geodesic space.

Distances between airways?

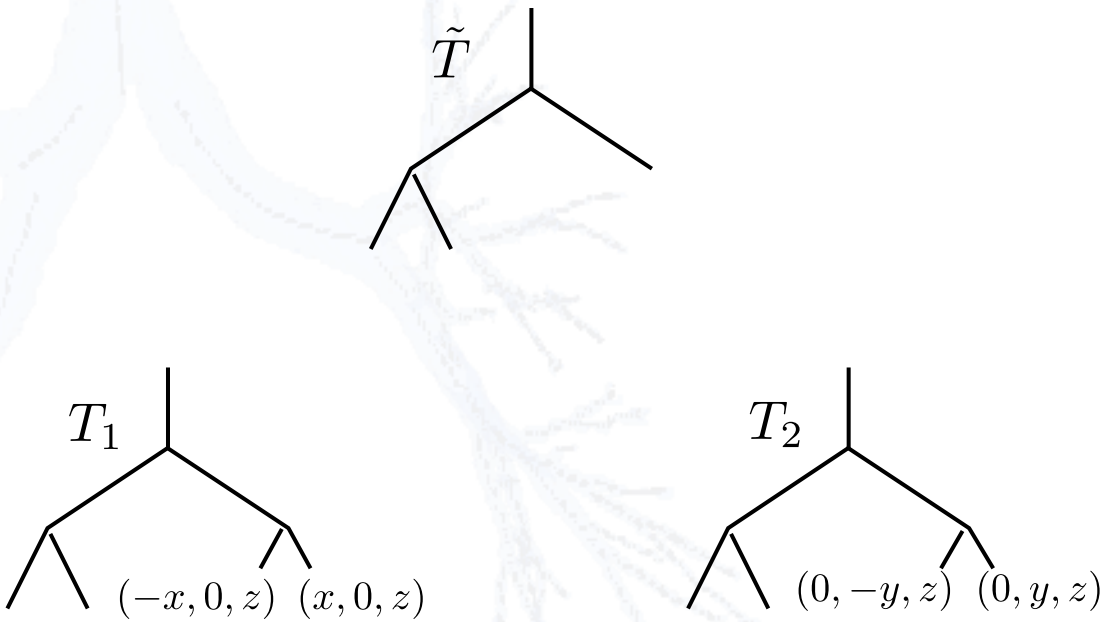
Evaluation of metric:

Approximate geodesic distances between 30 airways of healthy individuals and individuals with moderate COPD.



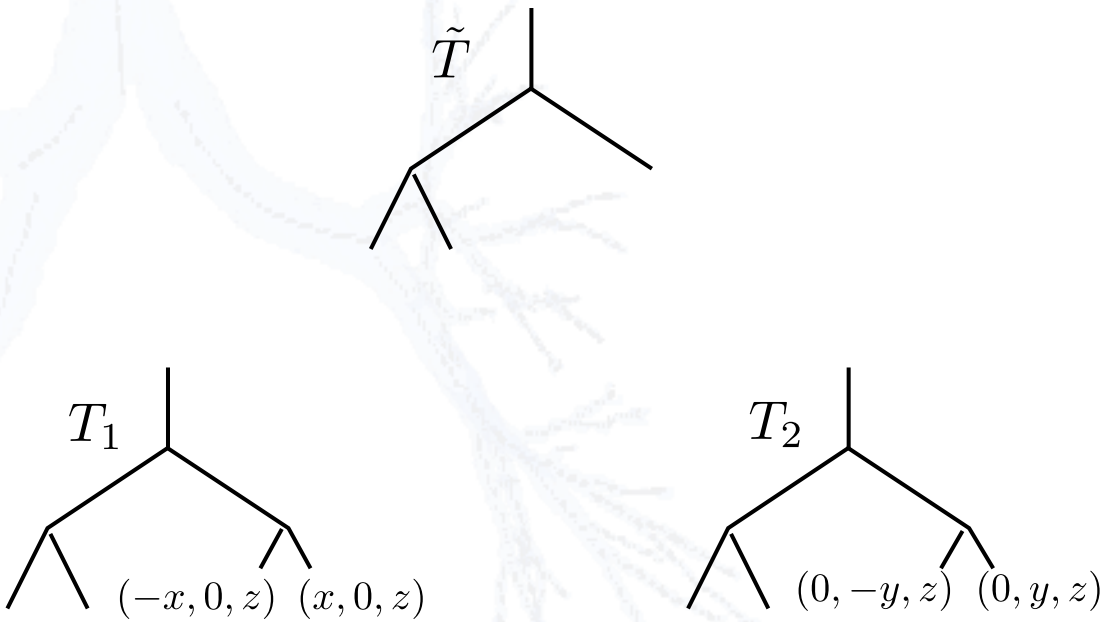
Curvature of shape space?

- This space has everywhere unbounded curvature!



Curvature of shape space?

- This space has everywhere unbounded curvature!



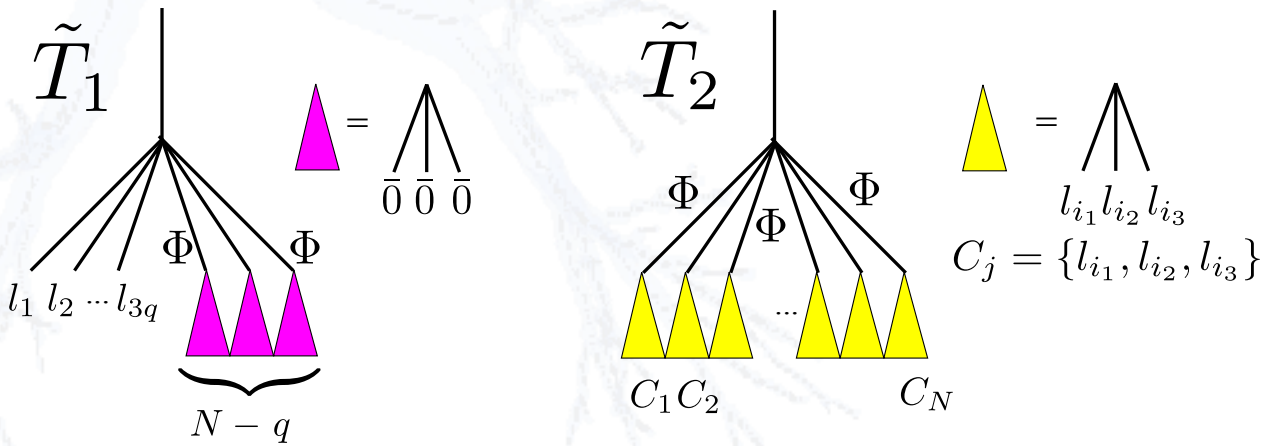
- Oh dear.

Complexity of computing geodesics?

Assume edge attributes have dimension > 1
(for dim $= 1$, Scott Provan).

Theorem

Computing QED geodesics is NP complete.

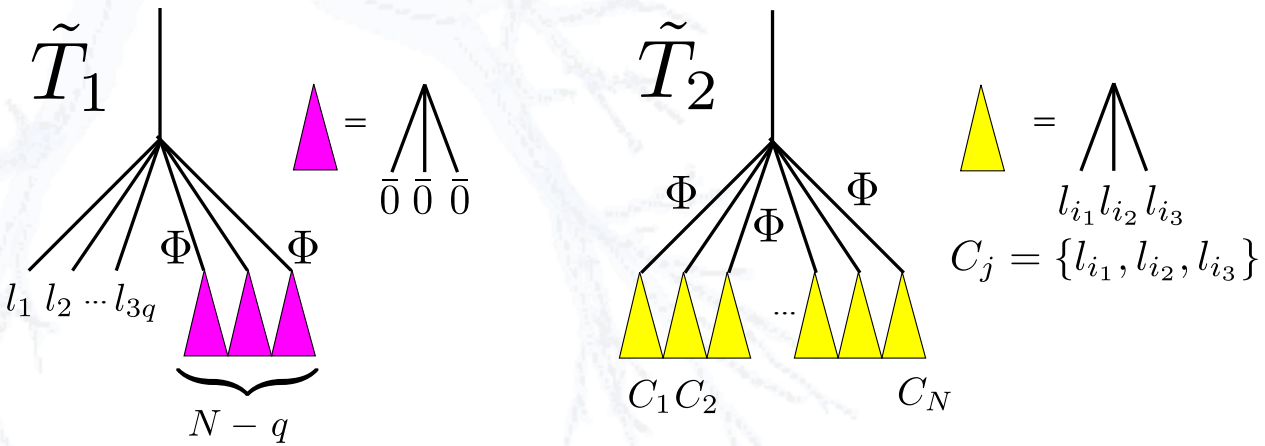


Complexity of computing geodesics?

Assume edge attributes have dimension > 1
(for dim $= 1$, Scott Provan).

Theorem

Computing QED geodesics is NP complete.



Oh dear.

First set of assumptions:

- **Assume: underlying rooted, ordered, binary tree $\mathscr{T}$ is finite.**
- **Study the new shape space $\bar{X}$.**

Curvature of shape space

Theorem

- Consider $(\bar{X}, \bar{d}_2)$ and $(\bar{\bar{X}}, \bar{\bar{d}}_2)$, ordered/unordered tree-shape space.
- At generic points, the space is locally $CAT(0)$.
- Its geodesics are locally unique at generic points.
- At non-generic points, the curvature is unbounded. □

Curvature of shape space

In fact, curvature is one of:

- $+\infty$
- 0
- $-\infty$

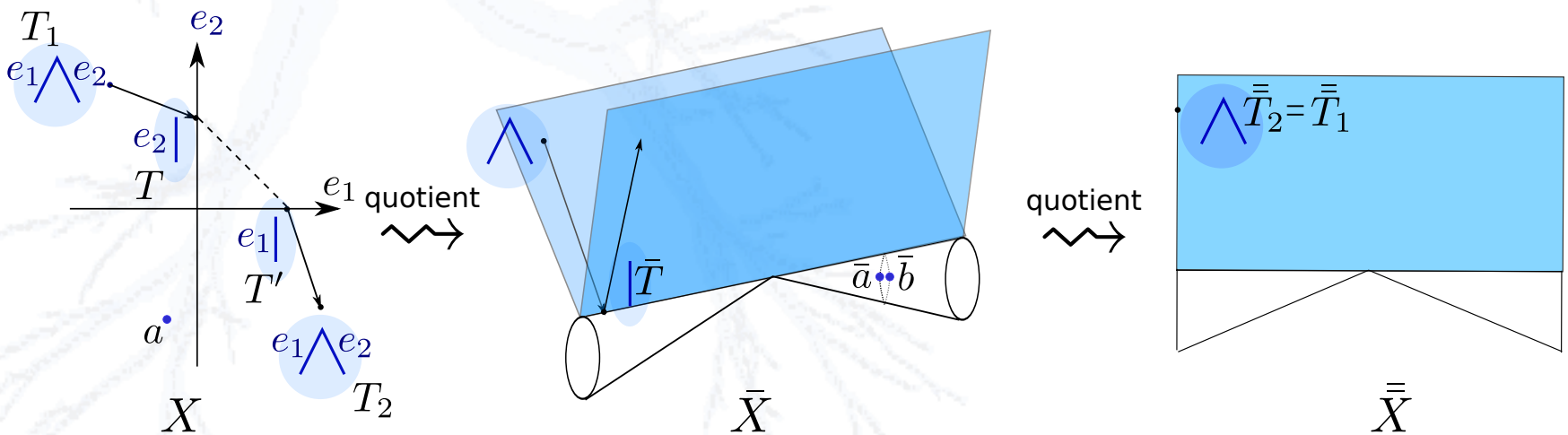
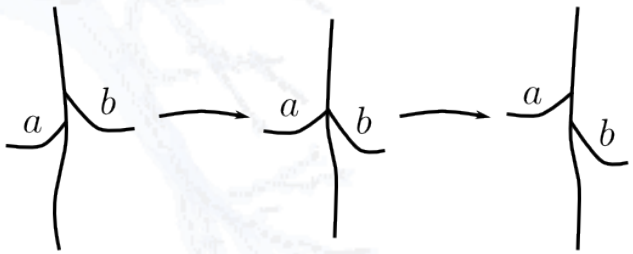


Figure: Space of ordered/unordered trees with at most 2 edges

How much better did it get?

- How local is "local"?
- The $CAT(0)$ neighborhoods in $\bar{X}$ now mainly consist of trees with the topology $\mathscr{T}$.
- Transitions such as

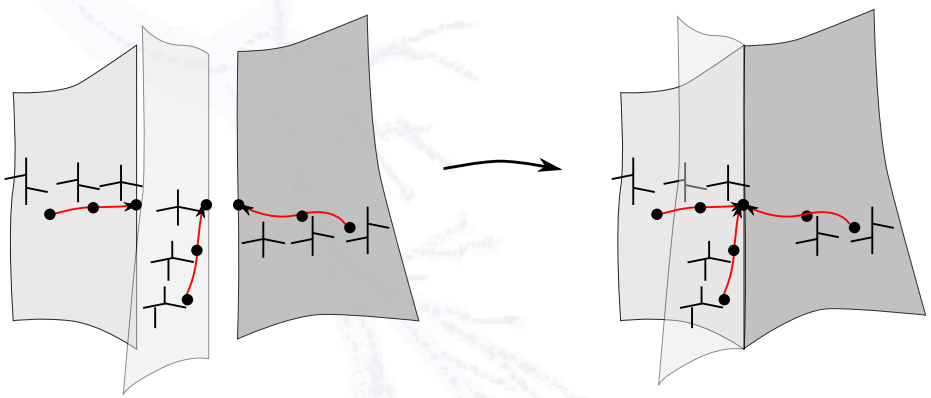


do not take place in $CAT(0)$ neighborhoods.

- Computational complexity? Still NP complete.

Second set of assumptions:

- **Restrict to:** all representations of certain restricted tree topologies.
- **Example 1:** Restrict to the set $\bar{\bar{X}}_N$ of trees with N leaves.



- **Example 2:** Restrict to all topologies occuring in airway trees.

How much better did it get?

- The $CAT(0)$ neighborhoods are now larger and can contain different top-dimensional tree topologies.
- Computational complexity? Still NP complete.

We can compute means!

Leaf vasculature data:

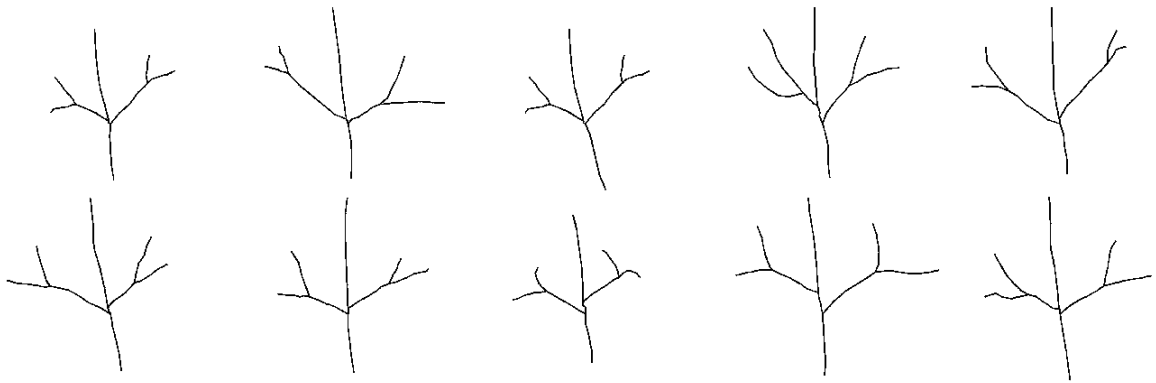


Figure: A set of vascular trees from ivy leaves form a set of planar tree-shapes.

Figure: a): The vascular trees are extracted from photos of ivy leaves. b) The mean vascular tree.

We can compute means!

The mean upper airway tree[1]

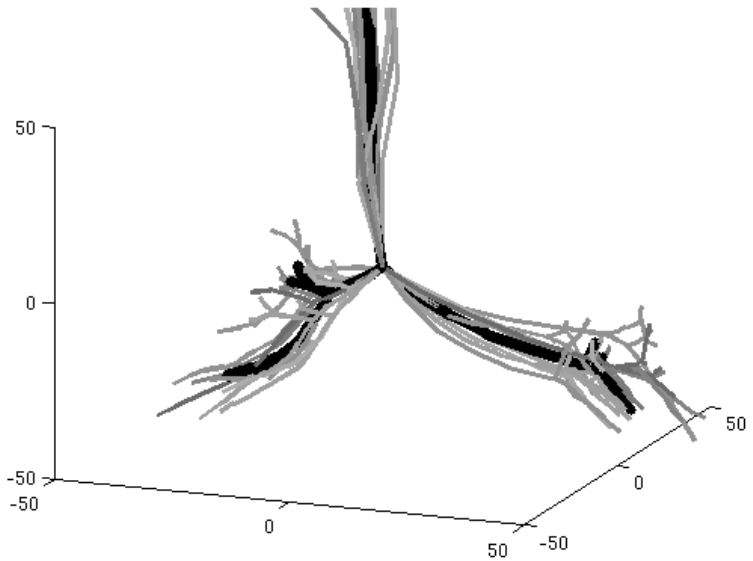


Figure: A set of upper airway tree-shapes along with their mean tree-shape.

[1] Feragen et al, *Means in spaces of treelike shapes*, ICCV2011

We can compute means!

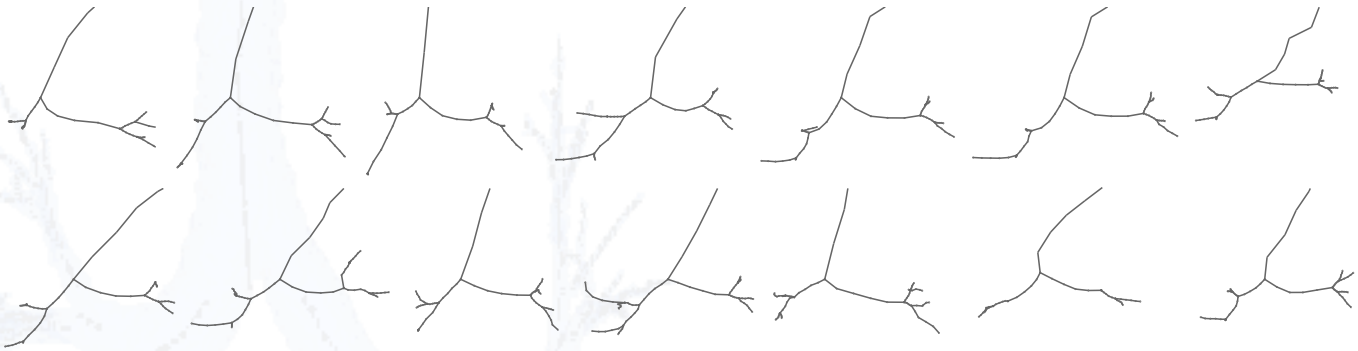


Figure: A set of upper airway tree-shapes (projected).[1]

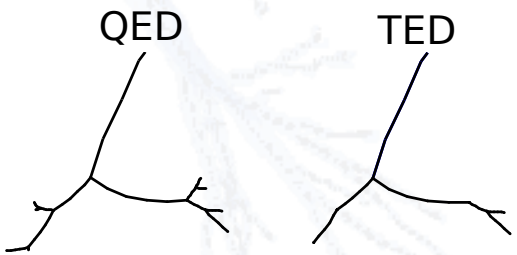


Figure: The QED and TED (algorithm by Trinh and Kimia) means.

[1] Feragen et al, *Towards a theory of statistical tree-shape analysis*, submitted

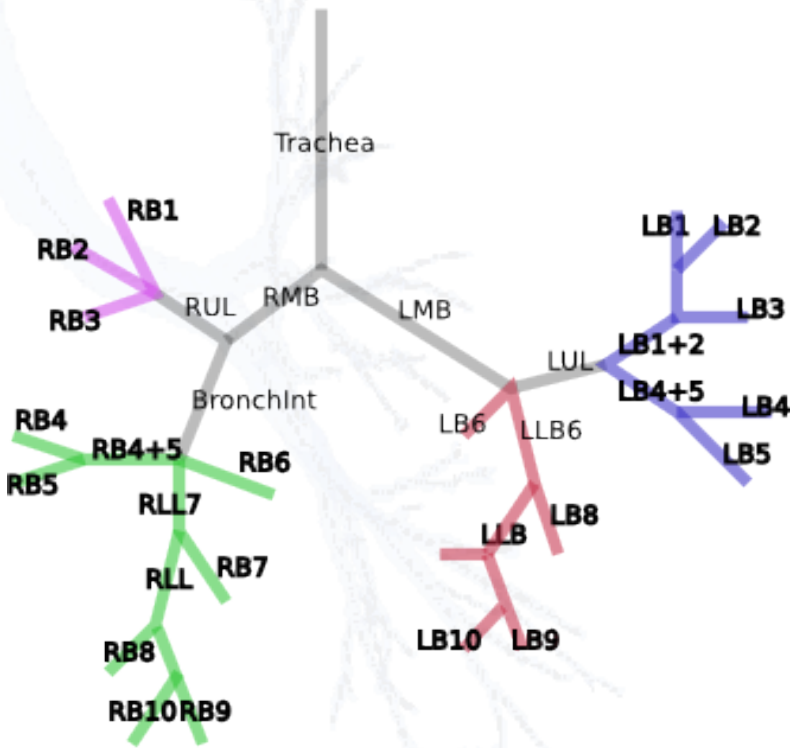
aasa@diku.dk,

Third (a) set of assumptions

- **Order** your edges: Left-right for each set of siblings
- For the $L1$ distances (TED), there are now polynomial time algorithms for distance.
- Open question: QED?
- Too restrictive for our data!

Property of airways

The first 6-8 generations of the airway tree are "similar" in different people.

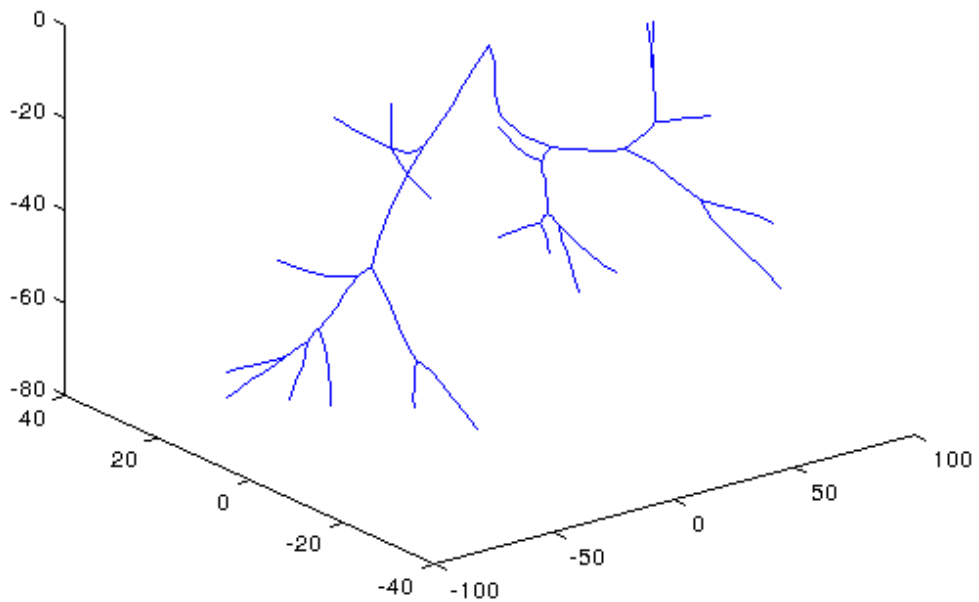


NB!: Not all present in all people; **not all present in all segmentations.**

Third (b) set of assumptions

- **Label** the "leaves" of your trees and insist that all trees have the same leaf label set.
- Vector version of the phylogenetic tree-space.
- Polynomial time algorithms
- **Also:** Factor out leaf labels via leaf permutation group $\rightsquigarrow$ NP complete.

Mean airway tree

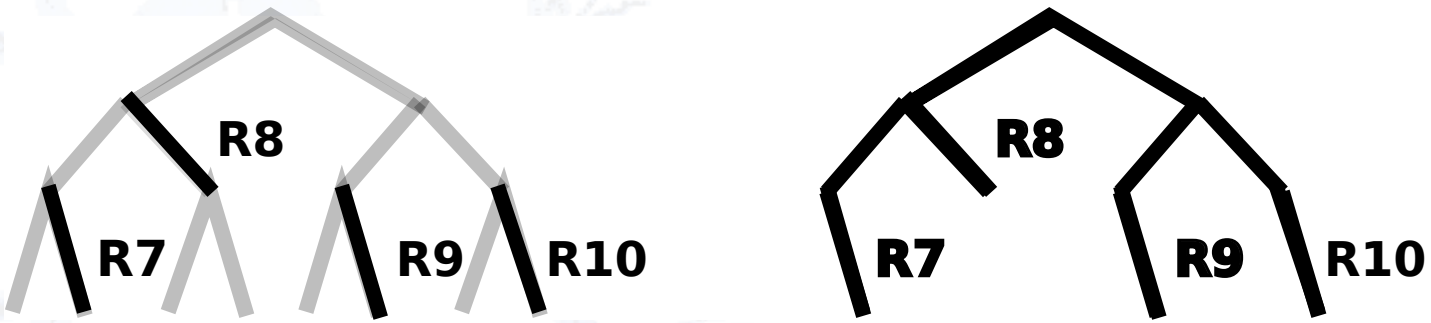


Joint with Megan Owen.

Heuristic for geodesic airway branch labeling[2]

Idea:

- Generate leaf label configurations and the corresponding tree spanning the labels



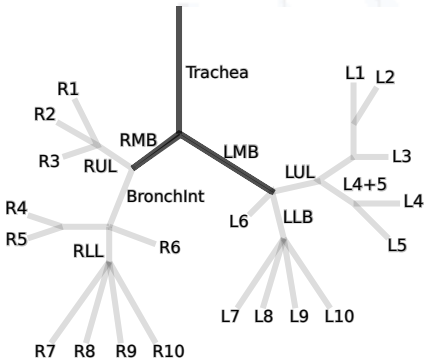
- Evaluate configuration in comparison with training data using geodesic deformations between leaf-labeled airway trees (Owen, Provan)

[2] F., Petersen, Owen, Lo, Thomsen, Dirksen, Wille, de Bruijne, *A hierarchical scheme for geodesic anatomical labeling of airway trees*, MICCAI 2012.

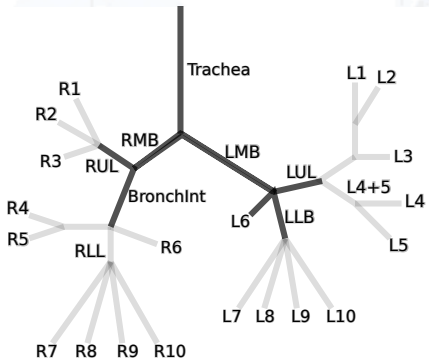
Heuristic for geodesic airway branch labeling[2]

Idea:

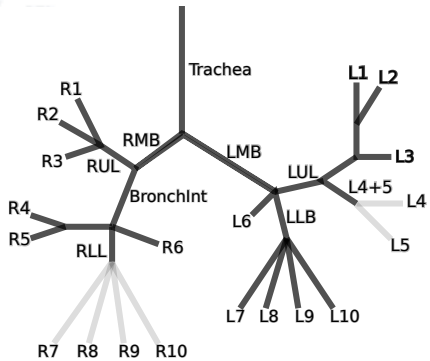
- Make tractable using a hierarchical labeling scheme



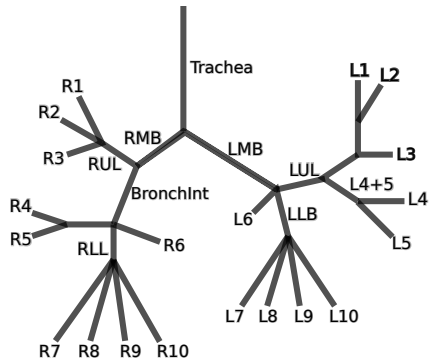
search 3 generations



search 2 and 2 generations



search 2, 2, 2 and 3 generations



search 3 and 2 generations

[2] F., Petersen, Owen, Lo, Thomsen, Dirksen, Wille, de Bruijne, *A hierarchical scheme for geodesic anatomical labeling of airway trees*, MICCAI 2012.

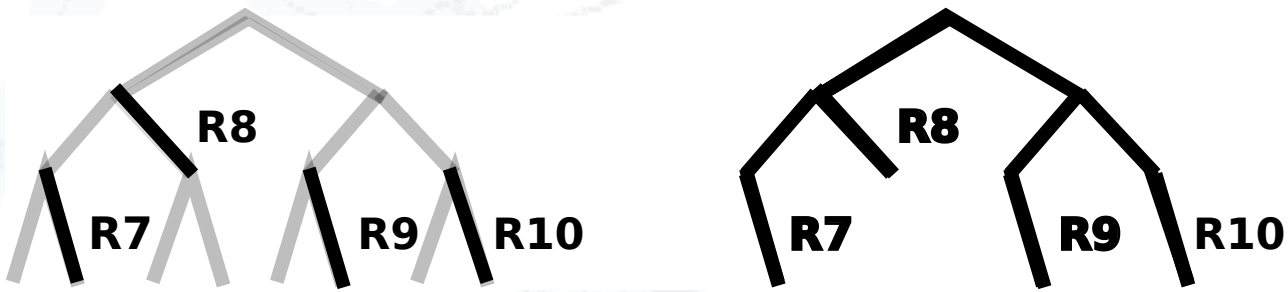
Heuristic for geodesic airway branch labeling[3]

- 40 airway trees from 20 subjects with different stages of COPD, hand labeled by 3 experts in pulmonary medicine.
- All 20 segmental labels were assigned (segmental = most distal branches) at an average success rate of 72.8%.
- **Performance:** as good as the performance of an expert in pulmonary medicine.
- Not significanly correlated with stage of COPD.

[3] Feragen et al, *A hierarchical scheme for geodesic anatomical labeling of airway trees*, MICCAI 2012.

What did we just do?

- Used leaf-labeled tree distance
- Coincides with QED distances **when**
  - QED geodesic induces same leaf matching as the leaf labelings
  - Everything below the leaves is dropped
  - Everything dropped above the leaves is considered noise



- Used right, this can be used as a heuristic to compute unordered, unlabeled tree distance via phylogenetic tree distance
- Heuristic takes care of noise above the leaf level.

Conclusion

- Unlabeled, unordered tree distances are NP hard
- Unlabeled, unordered trees live in spaces of unbounded curvature
- Adding assumptions gives some bounds on curvature and complexity, but decreases the ability to represent the data

Open question

- Conjecture 1: Given a tree T_1, for a generic tree T_2, there is a unique geodesic joining them
- Conjecture 2: $CAT(0) \rightsquigarrow$ generic $CAT(0)$ – unique means for generic datasets?