MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS


Anastasios Sidiropoulos
Massachusetts Institute of Technology

Krzysztof Onak
Massachusetts Institute of Technology

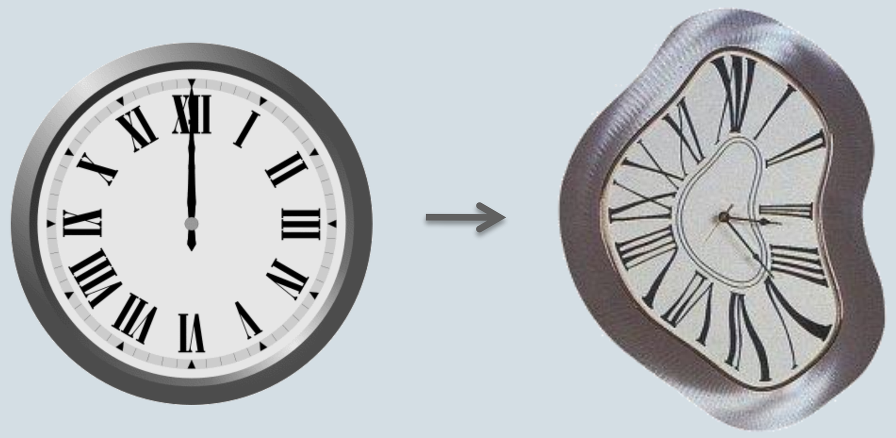
Algorithmic Embeddings into Low-Dimensional Spaces

What is a metric embedding?

An embedding of an input metric space into a host metric space is a mapping that sends each point of the input space to a point of the host space. Such a mapping has low distortion if the geometry of the resulting space approximates the geometry of the input space.

Formally, the distortion of an embedding f:X→X', from a space M=(X,D), into a space M'=(X',D'), has distortion c if for any pair of points x,y in X,

$$D(x,y) \leq D'(f(x),f(y)) \leq c \cdot D(x,y)$$



Goal

Distortion quantifies the extend to which an embedding preserves the original distances. Naturally, it is desired to obtain embeddings with the smallest possible distortion. We study the approximability of the minimum distortion embedding into low-dimensional Euclidean spaces.

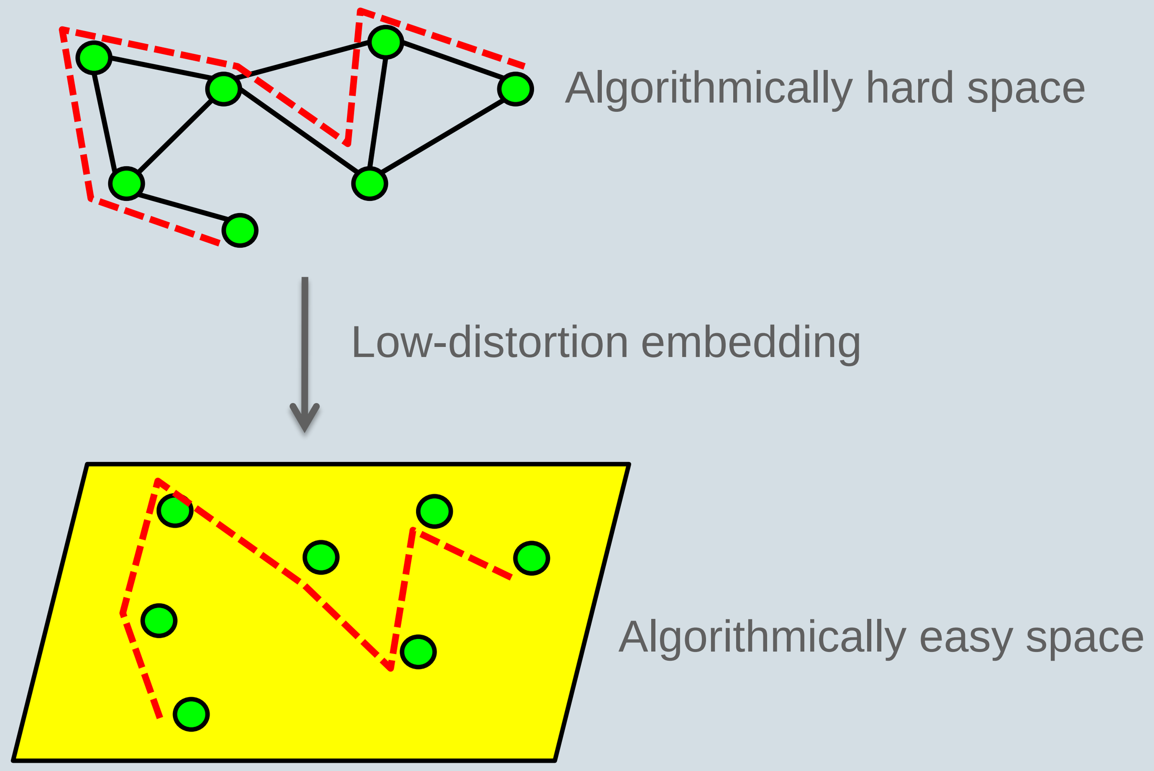
Motivation

Geometric interpretation. Using geometric techniques for the analysis of general metrical data.

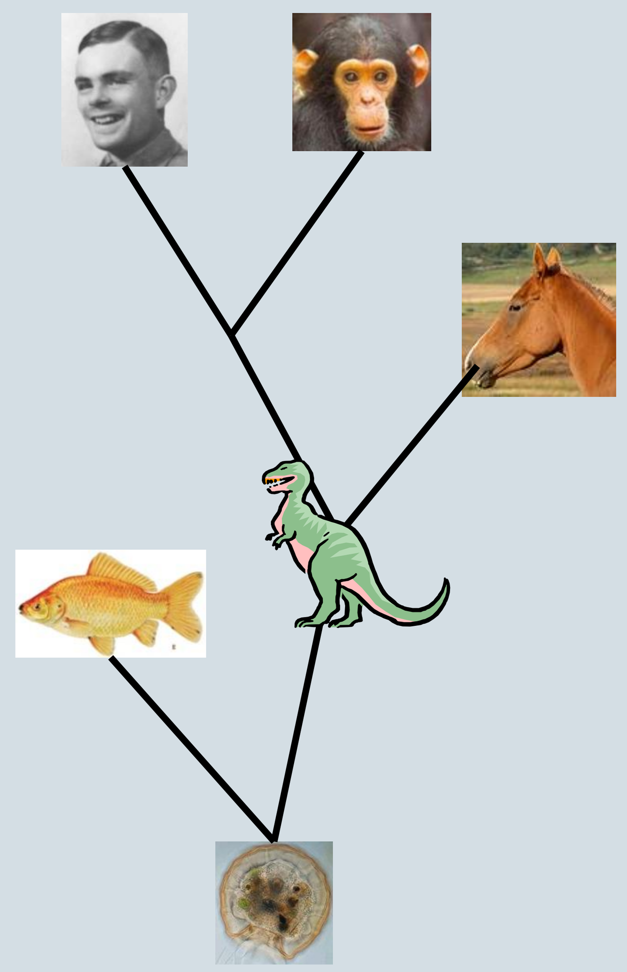
Succinct data representation. Embedding into low-dimensional spaces. Metrical compression.

Visualization via embedding into the plane. Multi-dimensional scaling.

Metric optimization. Embedding into algorithmically simple spaces.



Phylogenetic reconstruction can be expressed as the problem of minimum-distortion embedding into tree metrics.



Ultrametrics

A metric space is called an ultrametric if for any triple of points x,y,z,

$$D(x,y) \leq \max\{D(x,z), D(z,y)\}$$

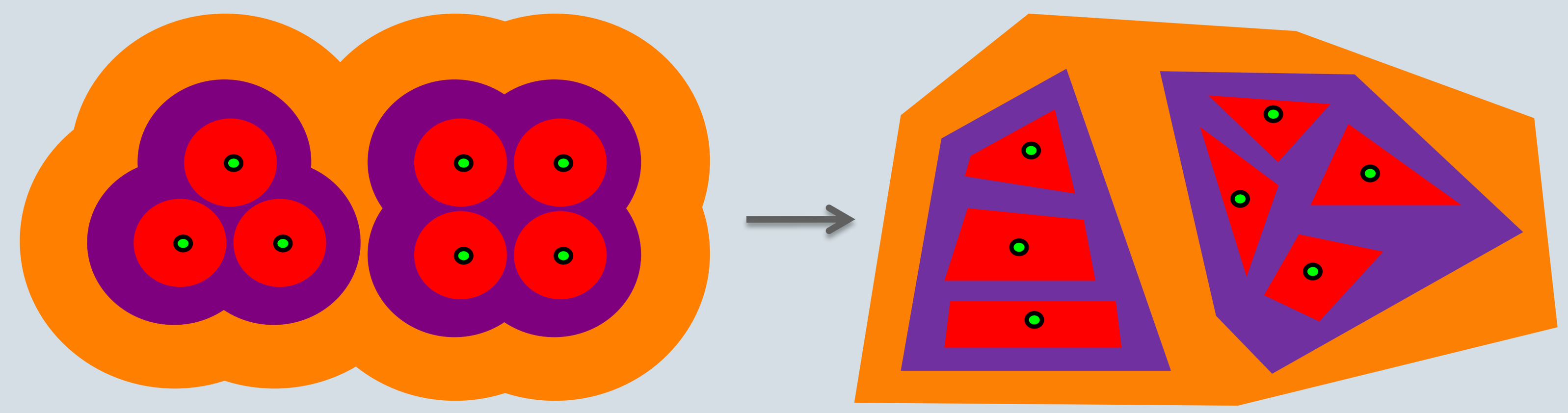
Ultrametrics can be used to model hierarchies, and chronograms. For example, consider the metric space defined on the set of all biological species, where the distance between two species A and B is the time since A and B diverged genetically. It is easy to see that the resulting metric space is an ultrametric.

Embedding ultrametrics into R^d

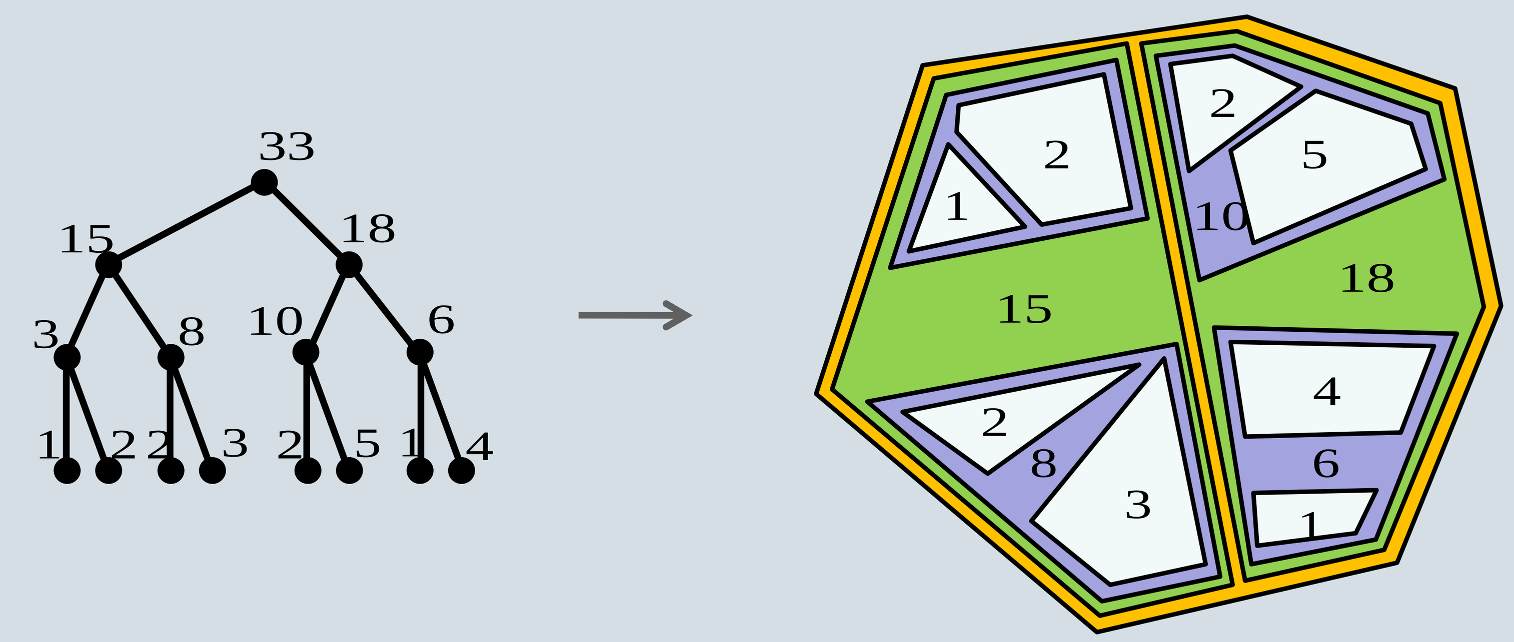
We develop an efficient approximation algorithm for embedding ultrametrics into d-dimensioinal Euclidean space, minimizing the distortion. Our algorithm computes an embedding with distortion $O(OPT \cdot \log^{O(1)} n)$, where OPT is the optimal distortion.

The main idea is to try to match certain structural properties of the optimal embedding, that can be computed efficiently. To that extend, we characterize the optimal embedding as the solution to a certain system of geometric constrains on convex sets in R^d.

Based on this observation, we reduce the problem on the existence of certain hierarchical partitions of R^d into a set of convex sets, each having small aspect ratio. Formally, we show the following. Let T be a tree with weights on the vertices, such that the weight of a vertex is at least the sum of the weights of its children. Then, we can compute for each vertex v of T a convex set A_v in R^d, such that



- The set of each vertex is contained in the set of its parent.
- The sets of vertices that are not on the same branch, are disjoint.
- Each set has small aspect ratio.



MADALGO – Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation