Goodies in Statistic and ML.

DataShape, Inria Saclay

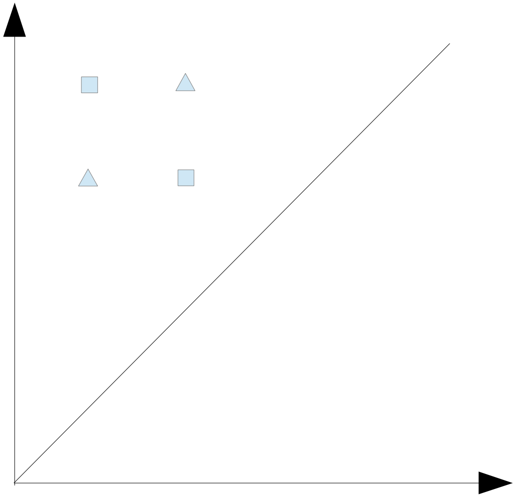
Gudhi is in Statistic and ML.

DataShape, Inria Saclay

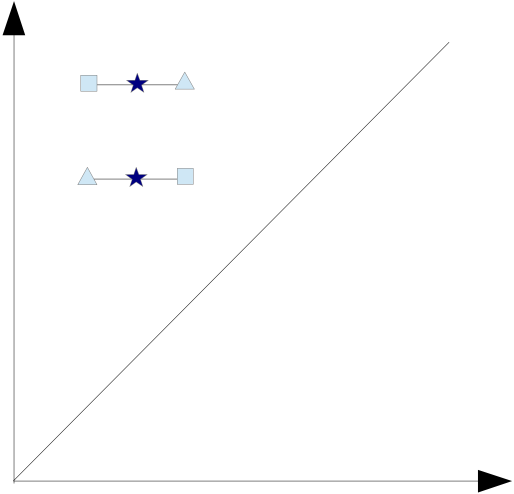
Where did I start playing with statistics?

- Analysis of time varying patterns from dynamical systems, more than 4 years ago.
- No statistical tools for persistent homology available.
- No efficient implementation of Bottleneck/Wasserstein distances available.
- Yet, there was a strong need for that in topological data analysis.

Why persistence diagrams are not sufficient?



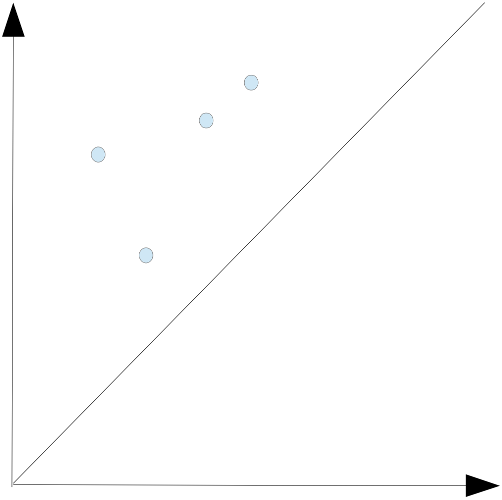
Why persistence diagrams are not sufficient?



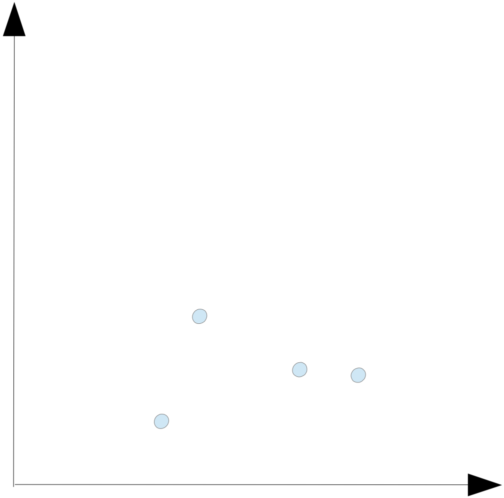
Persistence landscapes.

- Idea by Peter Bubenik.
- Very closely related to size functions used before (in dimension 0) by Bologna group.
- Lift persistence diagrams to Banach space of functions.
- This space is large enough to have well defined averages and scalar products.

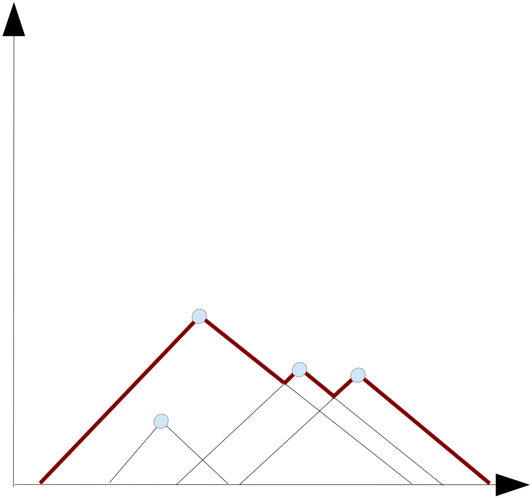
Persistence landscapes.



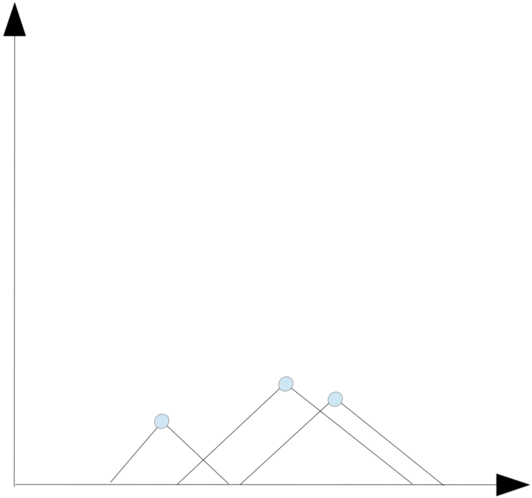
Persistence landscapes.



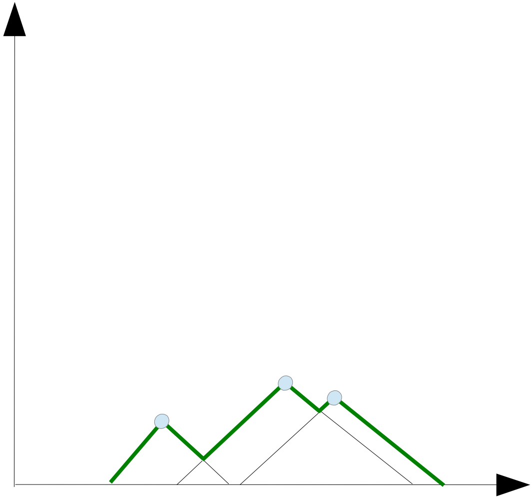
Persistence landscapes.



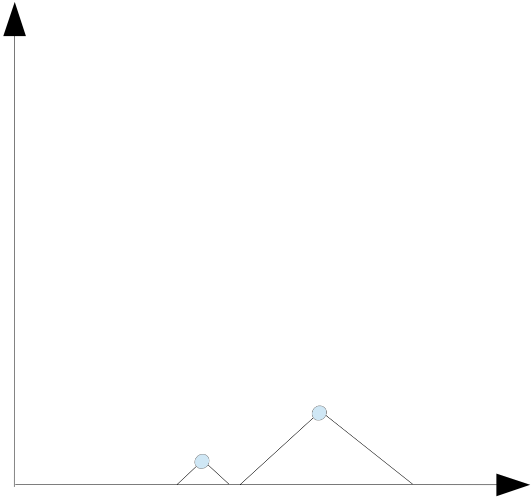
Persistence landscapes.



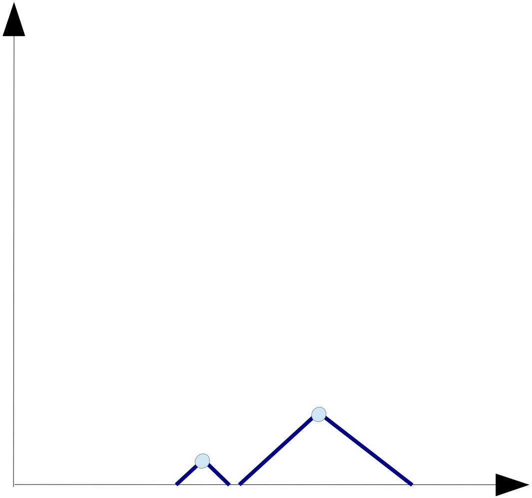
Persistence landscapes.



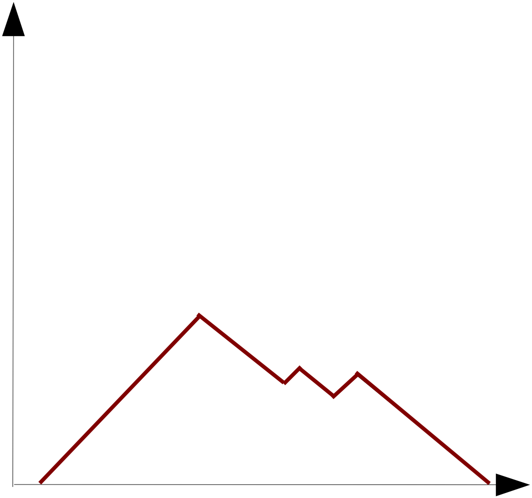
Persistence landscapes.



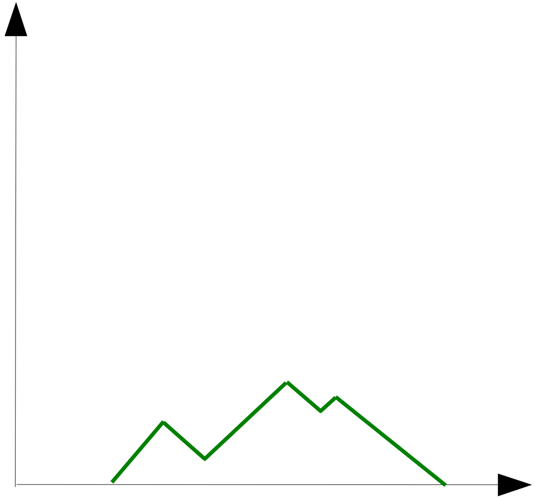
Persistence landscapes.



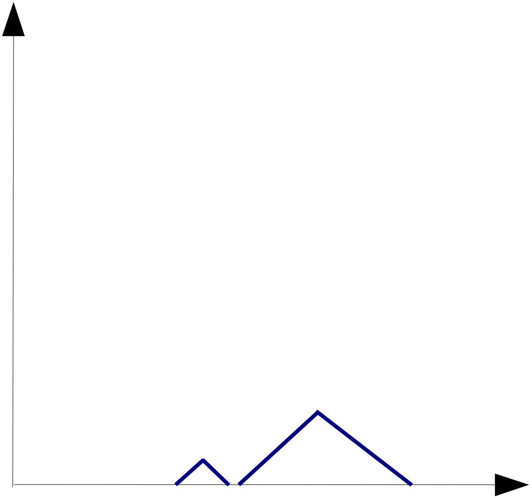
Persistence landscapes.



Persistence landscapes.



Persistence landscapes.



Persistence landscapes.

- Bottleneck stability.
- Averages.
- L^p distances.
- Scalar products.
- Various ways to vectorize.

Persistence landscape toolbox.

- Computations of distance matrix.
- Computation of averages landscapes.
- Standard deviation.
- Computations of integrals.
- Moments computations.
- Permutation test.
- T-test, anova.
- Classifiers.

Persistence landscape toolbox.

- In almost all the cases, we used only a few property of the landscapes.
- And it was not important at all that we use landscapes.
- Let us have a look at a concrete example.

Permutation test example.

Input: Two collections of persistence diagrams $c_1, ..., c_n$ and $d_1, ..., d_n$.

Output: p-value of a statement that averages of $c_1, ..., c_n$ and $d_1, ..., d_n$ are different.

Convert them to your favourite representation $\mathcal{A}$.

counter $= 0$.

$C =$ average of $c_1, ..., c_n$, $D =$ average of $d_1, ..., d_n$.

for N times **do**

  $B = \{c_1, ..., c_n, d_1, ..., d_n\}$.

  Shuffle B, and divide to B_1 and B_2.

  if $d(B_1, B_2) > d(C, D)$ **then**

    Increment *counter*.

return $\frac{counter}{N}$.

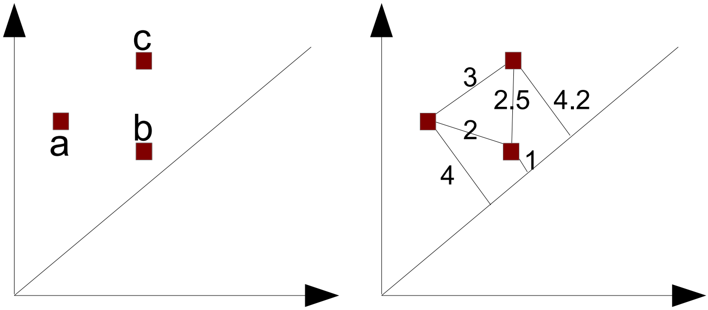
What do we need to do statistics?

- Distances.
- Averages.
- Scalar product.
- Vectorization.
- Confidence bounds.

Other representations of persistence.

- Persistence landscapes on a grid (simplified representation used in TDA R-package).
- Persistence vectors (by M. Cariere, S. Oudot and M. Ovsjanikov).
- Various types of "put a (weighted) kernel in every point of persistence diagrams" distributions:
  - Persistence Stable Space Kernel, by J. Reininghaus, U. Bauer, R. Kwitt.
  - Persistence Weighted Gaussian Kernel by G. Kusano, K. Fukumizu, Y. Hiraoka.
  - Persistence Images by Chepushtanova, Emerson, Hanson, Kirby, Motta, Neville, Peterson, Shipman, Ziegelmeier.

(Truncated) Vectors of distances.

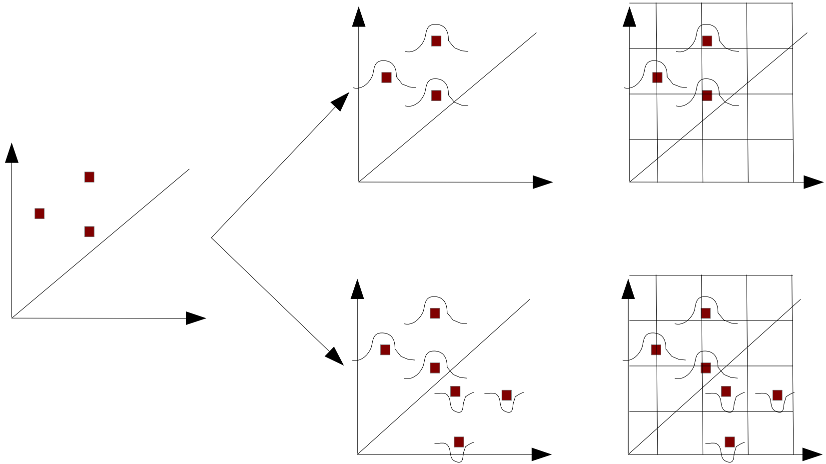


(3,1,1)

(Truncated) Vectors of distances, statistical operations.

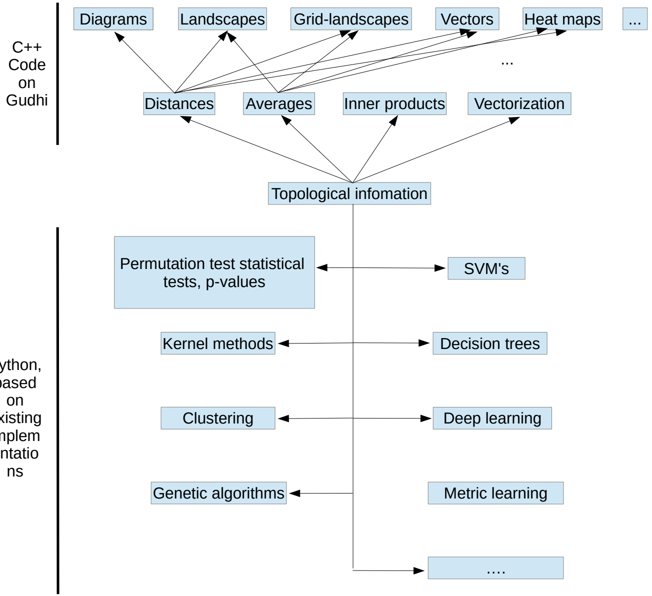
1. Point-wise averages.
2. Max, l^p distances.
3. Various projections to $\mathbb{R}$ are possible.
4. Scalar products of vectors well defined.
5. Vectorization is for free.

Distributions on diagrams.



Distributions on diagrams.

1. In any comparisons, grid sizes have to be comparable.
2. Distances and averages possible to define.
3. W-1 stable.
4. Vectorization possible.
5. Real-valued function possible to define.


C++ Code on Gudhi
Diagrams
Landscapes
Grid-landscapes
Vectors
Heat maps
...
...
Distances
Averages
Inner products
Vectorization
Topological infomation
Python, based on existing implementations
Permutation test statistical tests, p-values
SVM's
Kernel methods
Decision trees
Clustering
Deep learning
Genetic algorithms
Metric learning
....

Additional features.

1. Topological inference.
2. Distance to measure.

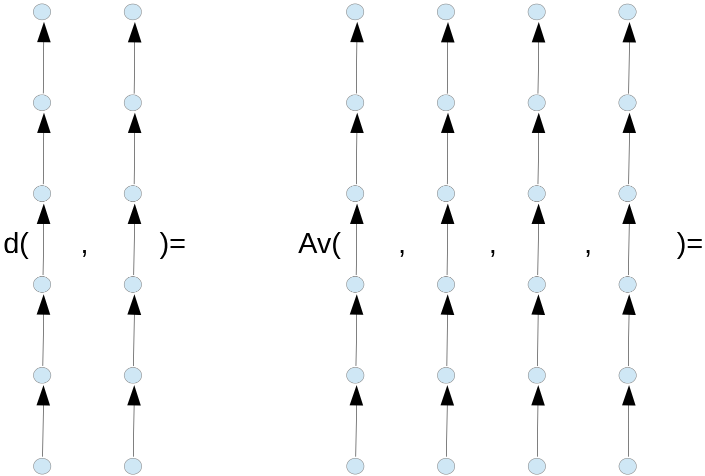
Looking forward, time varying data.

- Quite often our data are time–varying.
- In each time step we are given a scalar value function.
- But filtration is changing (continuously).
- Multi dimensional persistence.... no...
- Methods for time varying data.
- Note that we cannot go back in time.

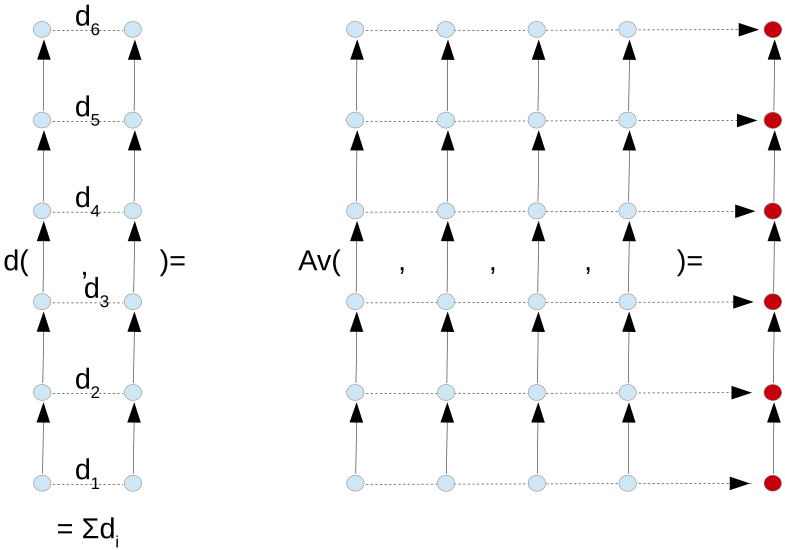
Time varying data.

- Suppose we know only the data from the constitutive time steps.
- We do not know how they were transformed to each other.

Distances and averages.



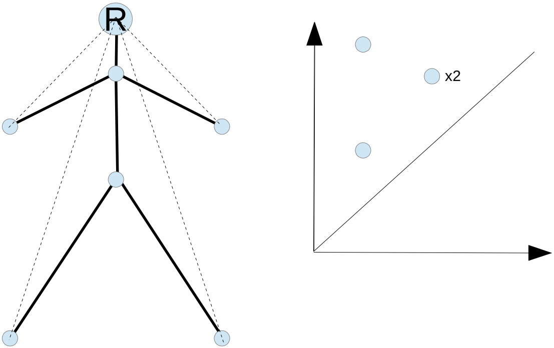
Distances and averages.



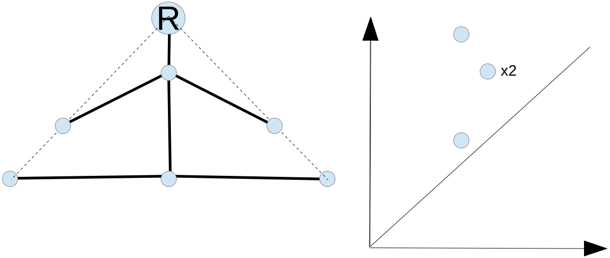
Topological process.

- The representation of a process is a time series (a vector) of persistence diagrams.
- I call this time series a *topological process*.
- All the statistical operations can be done coordinate–wise.
- We may however have more information.

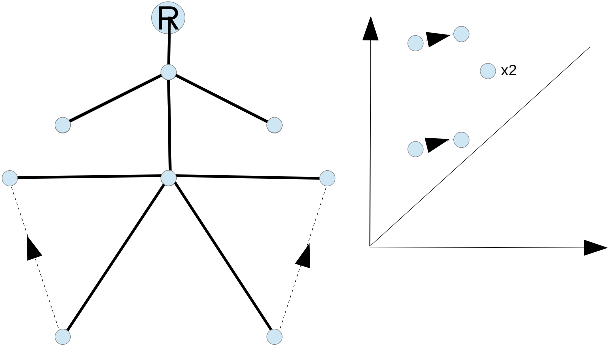
Full transformation.



Full transformation.



Full transformation.



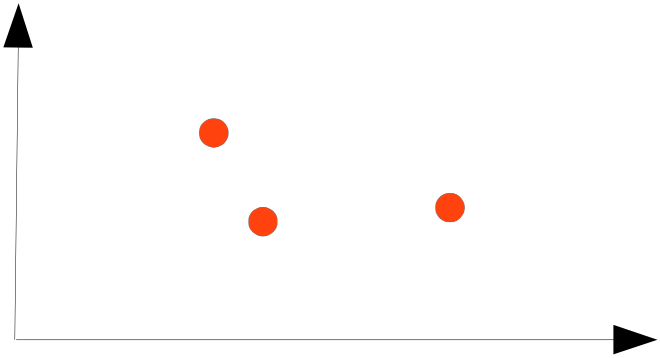
Time-varying data statistics.

- Diagrams, points $\rightarrow$ paths (vines and vineyards).
- (Dynamic) landscapes (updating of structure is needed).
- Gaussian kernel–based representations (we get 3d instead of 2d distribution).
- Persistence vectors changing in a sooth way.
- ...

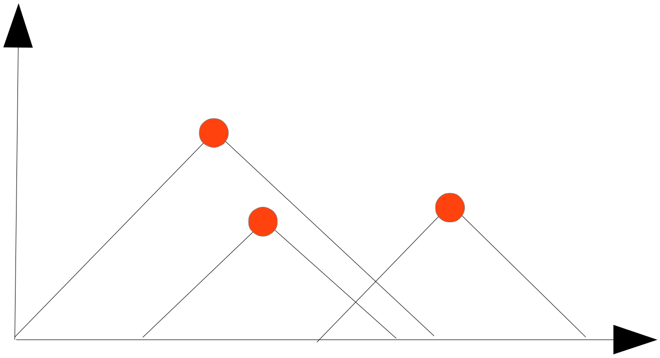
Time-varying trees statistics, vines and vineyards.

- Continuously time-varying persistence diagram gives us a vineyard.
- Standard bottleneck and Wasserstein distances defined by integrals of standard distances.
- Mean vineyard can be defined in analogy to Frechet mean of two diagrams.
- See phd thesis of Liz Munch.

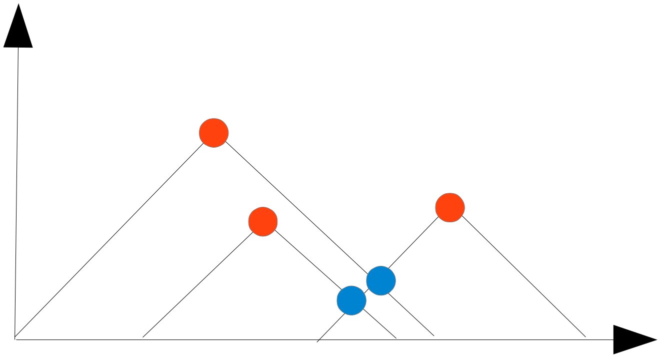
Time-varying trees statistics, landscapes.



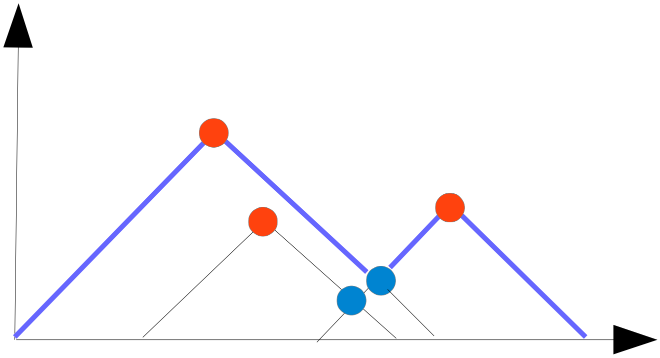
Time-varying trees statistics, landscapes.



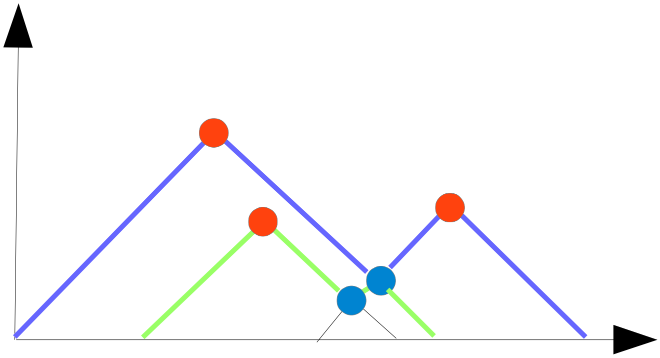
Time-varying trees statistics, landscapes.



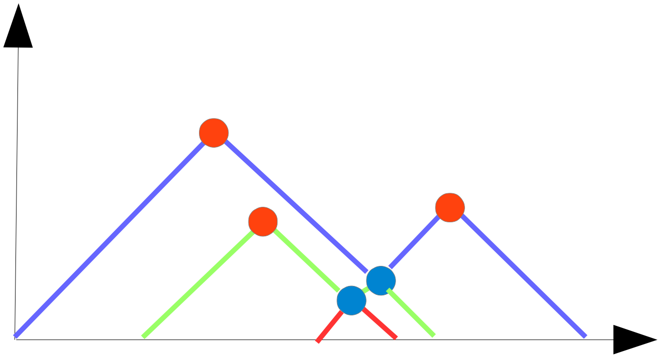
Time-varying trees statistics, landscapes.



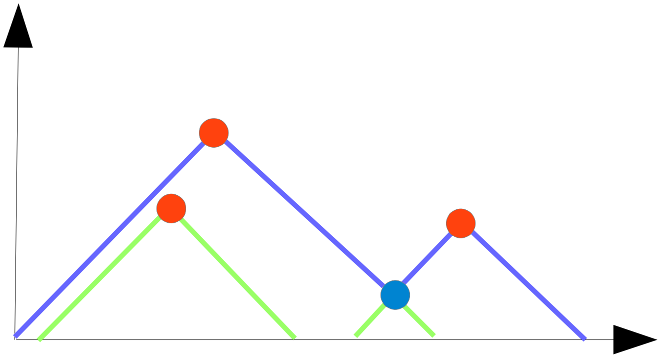
Time-varying trees statistics, landscapes.



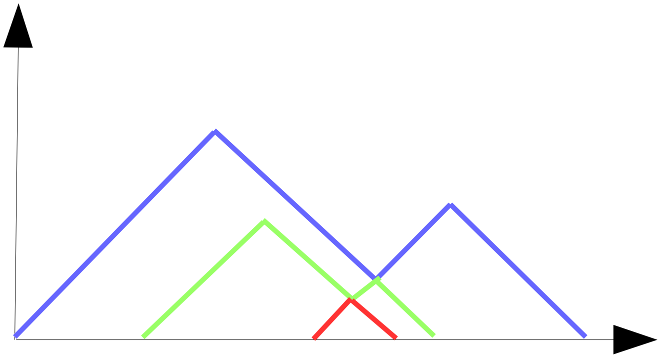
Time-varying trees statistics, landscapes.



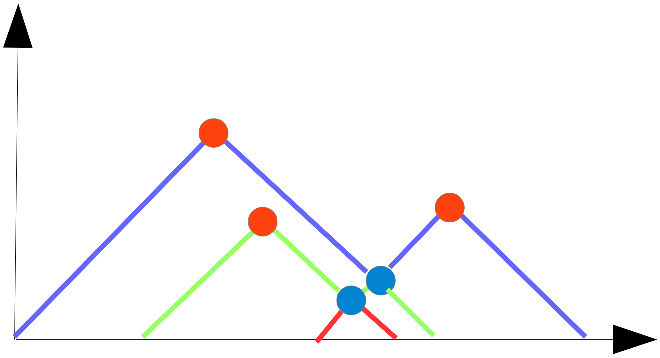
Time-varying trees statistics, landscapes.



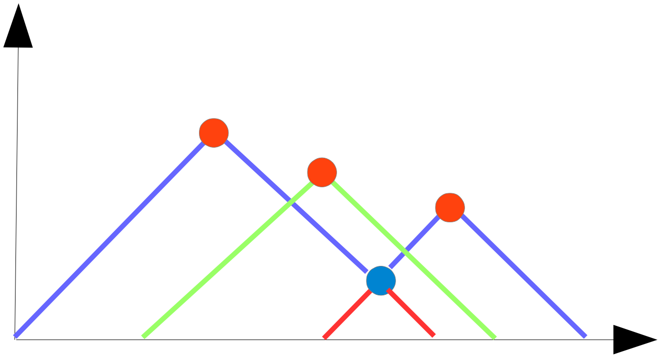
Time-varying trees statistics, landscapes.



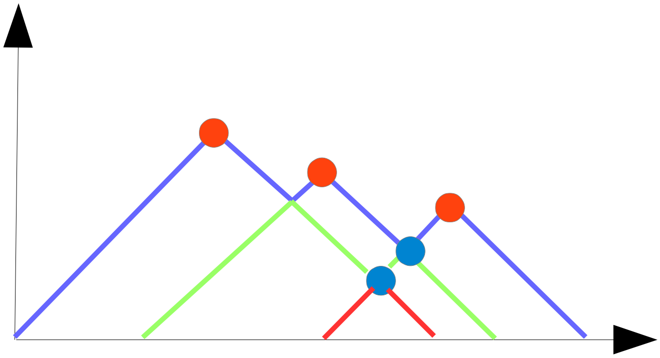
Time-varying trees statistics, landscapes.



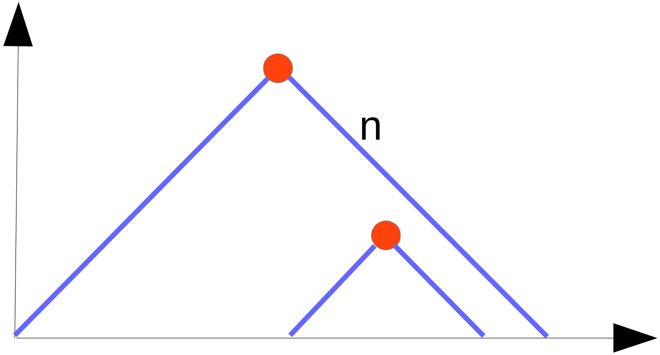
Time-varying trees statistics, landscapes.



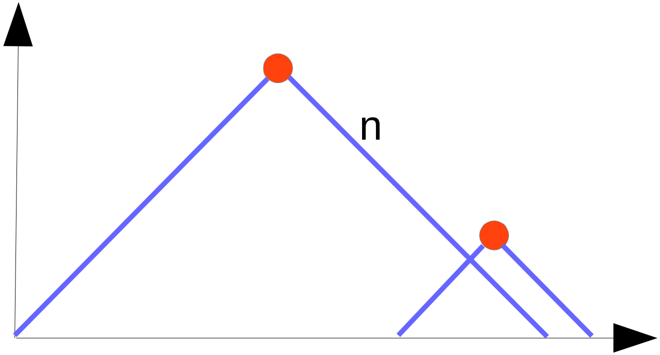
Time-varying trees statistics, landscapes.



Time-varying trees statistics, landscapes, non generic points.



Time-varying trees statistics, landscapes, non generic points.



Time-varying trees statistics, landscapes.

- The points are moving in a continuous way.
- Therefore intersection of line segments used to create landscapes moves in continuous way.
- New intersections may be created.
- Old intersections may disappear.

Time-varying diagrams statistics.

- In this case, the Gaussian kernel (with whatever mean and stdiv) travels along wines in vineyard.
- That gives continuous distribution in $\mathbb{R}^3$.
- Distances and averages are the standard ones from the L^p space.

Finally.

Let us have some goodies!

Thank you for your time!

DataShape Team, INRIA, Saclay and Sophia-Antipolis
contact: pawel.dlotko, vincent.rouvreau @ inria.fr