Dimension Reduction for Big Data Analysis

Dan Shen

Department of Mathematics & Statistics
University of South Florida

danshen@usf.edu

October 24, 2014

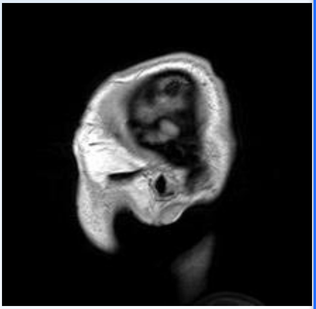
Outline

- Multiscale weighted PCA for Image Analysis
- Human brian artery tree analysis

Outline

- Multiscale weighted PCA for Image Analysis
- Human brian artery tree analysis

Image Analysis



Raw Images

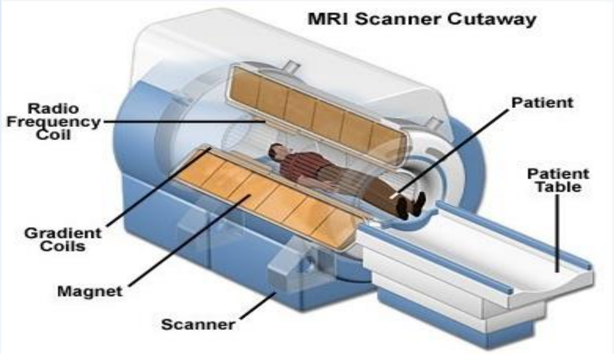
- Image Reconstruction
- Image Registration
- Image Segmentation

→ Statistical Analysis

USF
UNIVERSITY OF SOUTH FLORIDA

Image Analysis

Image Reconstruction



Fourier Transforms

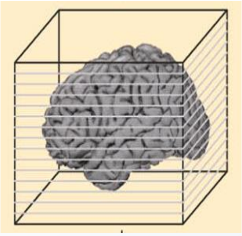


Image Registration

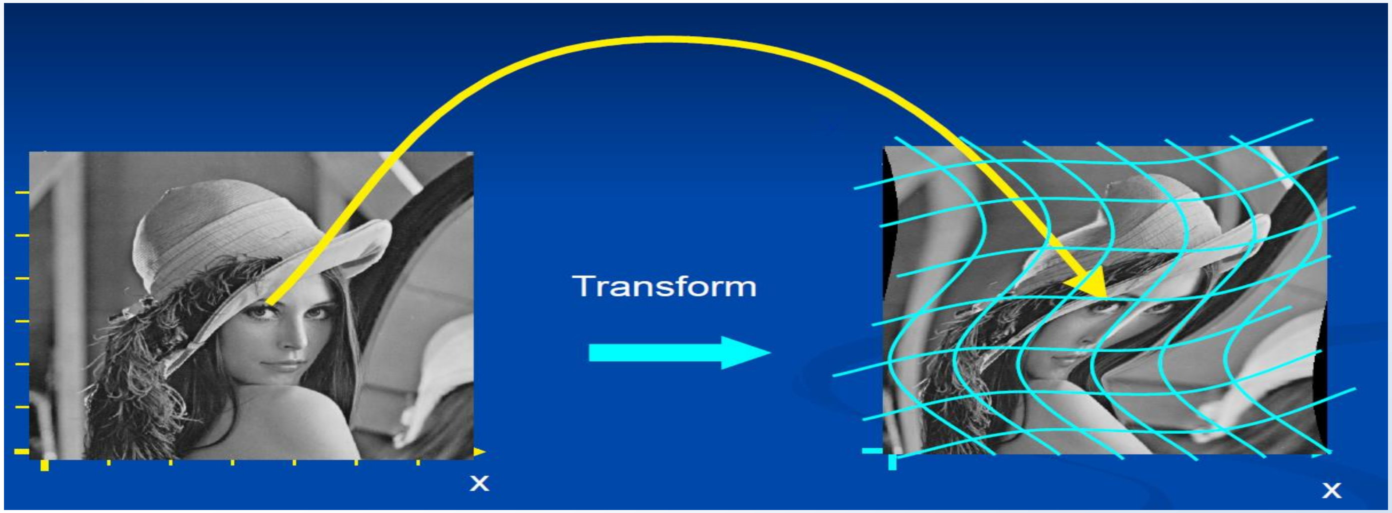
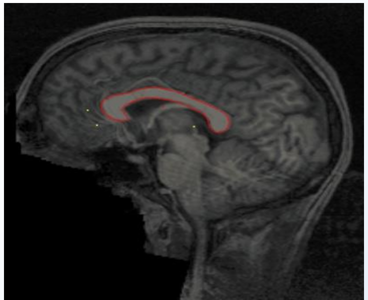
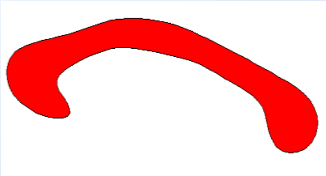
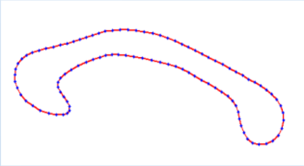


Image Segmentation







Challenges in Image Analysis

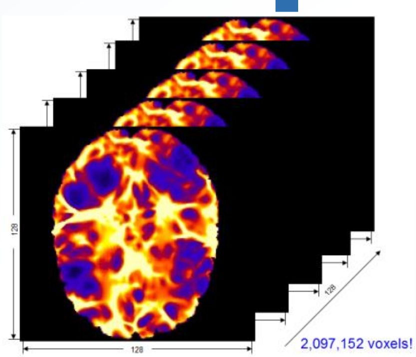
" Large p, small n" problem

$p = 563 \times 750 = 422{,}250$

$x_i^T = (\chi_{1,i}, \chi_{2,i}, \cdots \chi_{p,i})$

$$X_{p\times n} = \begin{pmatrix} \chi_{1,1} & \cdots & \chi_{1,i} & \cdots & \chi_{1,n} \\ \chi_{2,1} & \cdots & \chi_{2,i} & \cdots & \chi_{2,n} \\ & & \vdots & & \\ \chi_{p,1} & \cdots & \chi_{p,i} & \cdots & \chi_{p,n} \end{pmatrix}$$

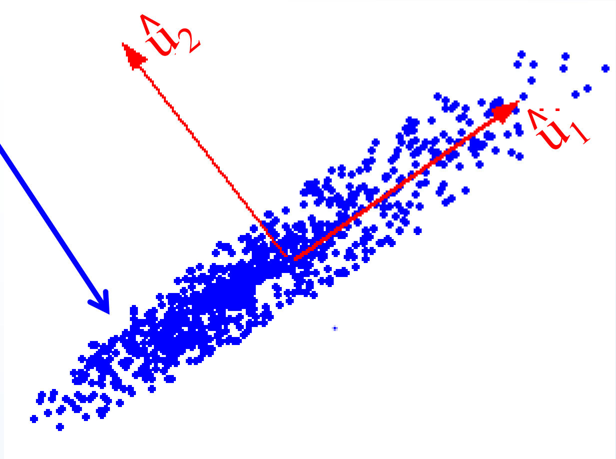
$p = 128^3 = 2{,}079{,}152$

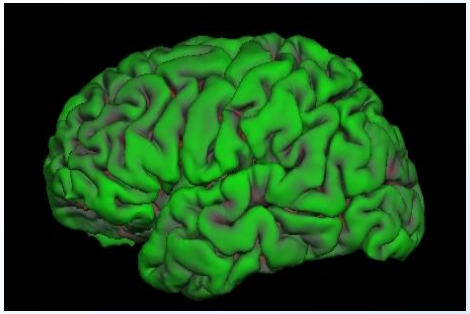


Principal Component Analysis

one data object

$$X_{p\times n} = \begin{pmatrix} \chi_{1,1} & \cdots & \chi_{1,i} & \cdots & \chi_{1,n} \\ \chi_{2,1} & \cdots & \chi_{2,i} & \cdots & \chi_{2,n} \\ & & \vdots & & \\ \chi_{p,1} & \cdots & \chi_{p,i} & \cdots & \chi_{p,n} \end{pmatrix}$$





Our Main Contribution

PCA doesn't work for high dimensional image data, what can we do ??

- Our multiscale weighted PCA will answer this question

UNIVERSITY OF SOUTH FLORIDA

ADNI Data

Alzheimer's Disease Neuroimaging Initiative (ADNI) data:

- Alzheimer's disease is a progressive, degenerative disorder that attacks the brain's nerve cells, or neurons, resulting in loss of memory, thinking and language skills, and behavioral changes.
- 390 subjects (218 normal controls and 172 AD patients)
- Download: http://adni.loni.usc.edu/

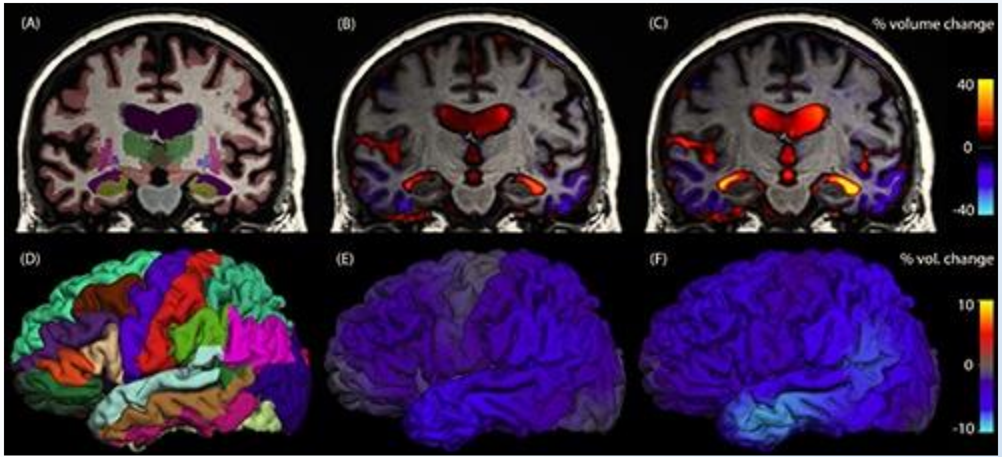
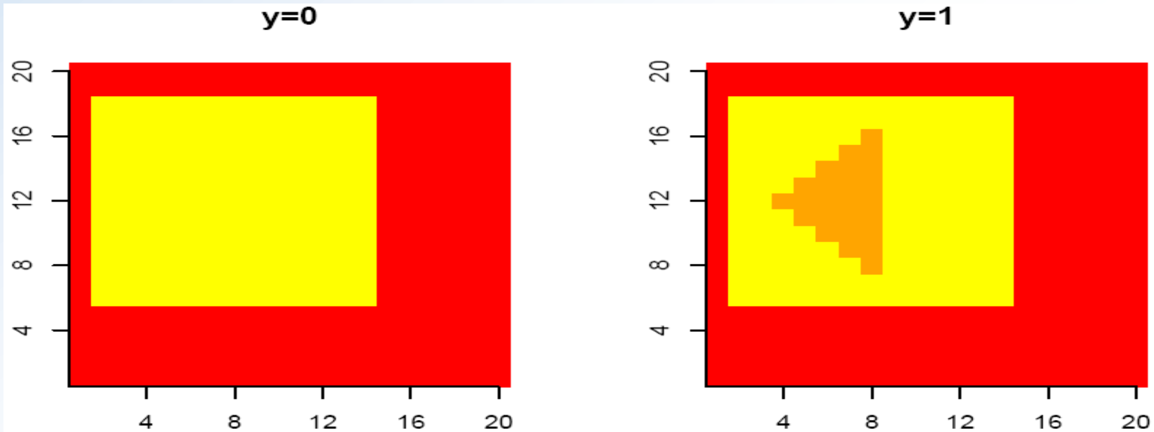
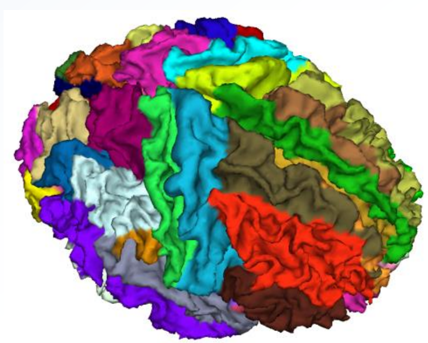
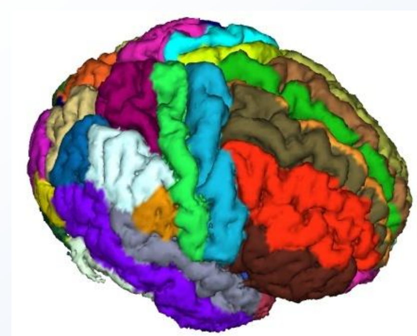


Image Classification

Underlying spatial information: features are spatially dependent







Dimension reduction becomes important and necessary for image data to improve prediction accuracy and increase classification efficiency

Limitation of PCA

- Inconsistency of PCA for large p, small n
- PCA treats all pixels/voxels equally
- PCA treats all pixels/voxels independently
- PCA doesn't consider the association with the outcome

UNIVERSITY OF SOUTH FLORIDA

Motivation

Propose Multiscale Weighted PCA (MWPCA)

- enables a selective treatment of individual features
- has the ability of utilizing the local spatial information
- takes into account the association with outcome
- integrates feature selection, smoothing, feature extraction in a single framework

PCA: Reconstruction

Find low dimensional representation of the data through minimizing the reconstruction error

$$\varepsilon = \sum_{i=1}^{n} \left\| X_i - \overline{X} - U_k a_i \right\| = \sum_{i=1}^{n} \sum_{j=1}^{p} (\tilde{x}_{i,j} - \tilde{u}_j a_i)^2, \qquad U_k^T U_k = I_k$$

where $\overline{X}$ is the mean and $U_k = (\tilde{u}_1, \ldots, \tilde{u}_p)^T$

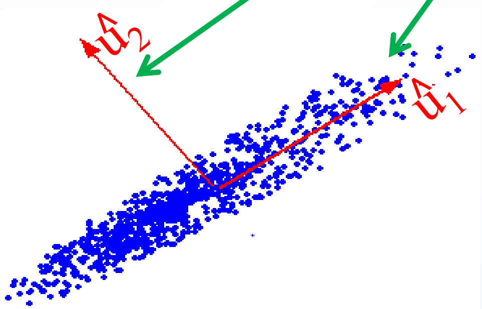
PCA: Reconstruction

Find low dimensional representation of the data through minimizing the reconstruction error

$$\varepsilon = \sum_{i=1}^{n} \left\| X_i - \overline{X} - U_k a_i \right\| = \sum_{i=1}^{n} \sum_{j=1}^{p} (\tilde{x}_{i,j} - \tilde{u}_j a_i)^2, \qquad U_k{}^T U_k = I_k$$

where $\overline{X}$ is the mean and $U_k = (\tilde{u}_1, \ldots, \tilde{u}_p)^T$

- columns of U_k are the first k principal component directions



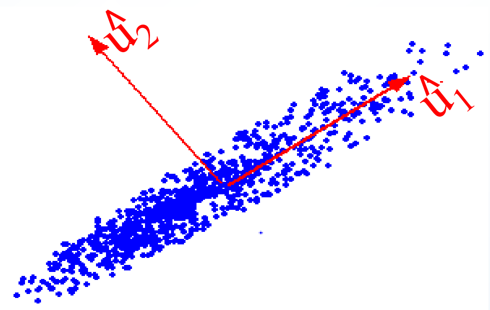
PCA: Reconstruction

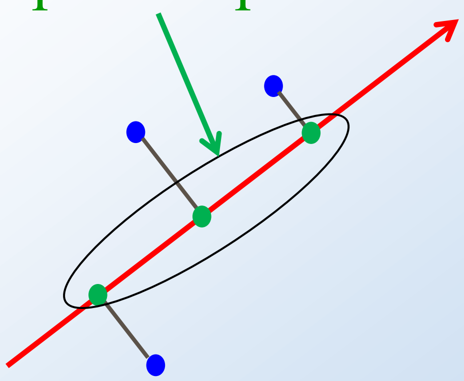
Find low dimensional representation of the data through minimizing the reconstruction error

$$\varepsilon = \sum_{i=1}^{n} \left\| X_i - \bar{X} - U_k a_i \right\| = \sum_{i=1}^{n} \sum_{j=1}^{p} (\tilde{x}_{i,j} - \tilde{u}_j a_i)^2, \qquad U_k{}^T U_k = I_k$$

where $\bar{X}$ is the mean and $U_k = (\tilde{u}_1, ..., \tilde{u}_p)^T$

- columns of U_k are the first k principal component directions
- columns of $A_k = (a_1, ..., a_n)^T$ are principal component scores



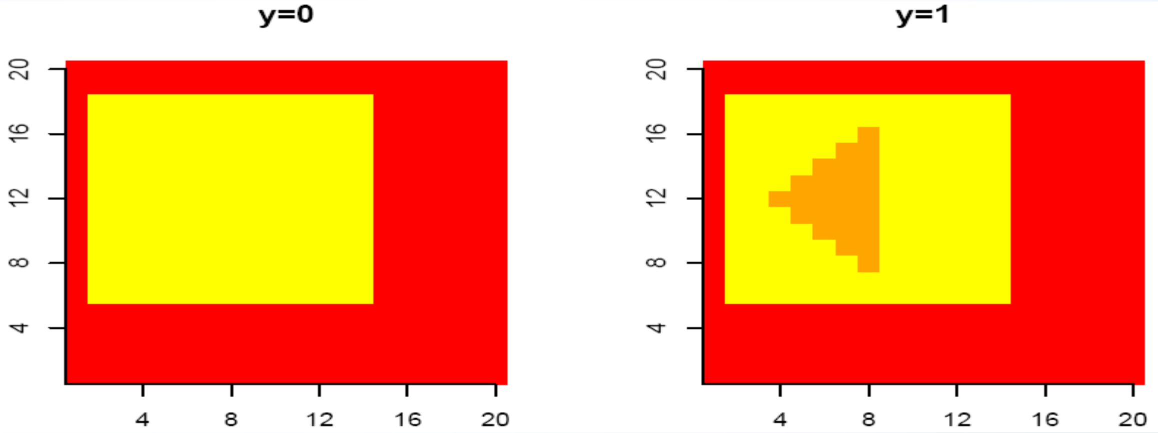


Multiscale Weighted PCA

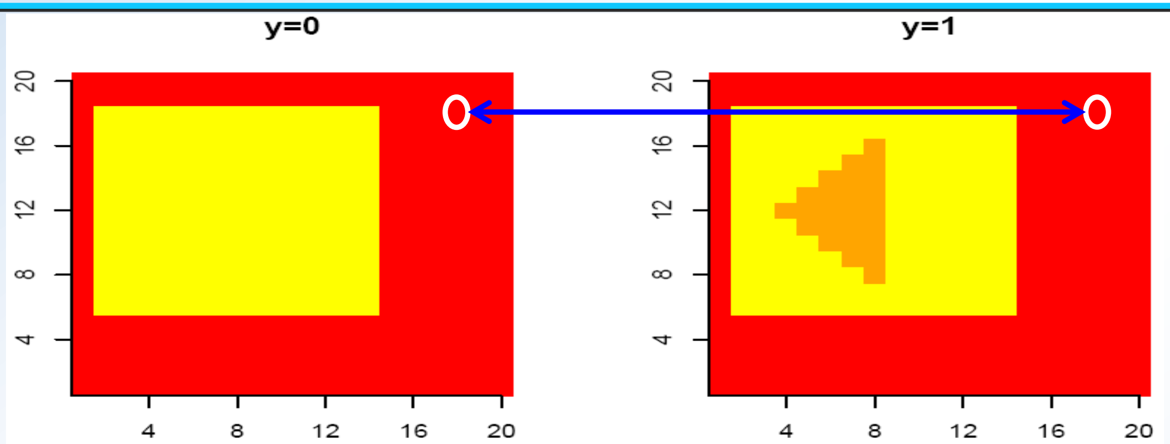
$$\varepsilon = \sum_{i=1}^{n} \sum_{j=1}^{p} w_j \sum_{d \in B(j;h)} w(j,d;h)(\tilde{x}_{i,j} - \tilde{u}_j a_i)^2$$

Two sets of weights:

- Global spatial weight: w_j for each pixel/voxel with $\sum_{j=1}^{p} w_j = p$



Global Spatial Weight



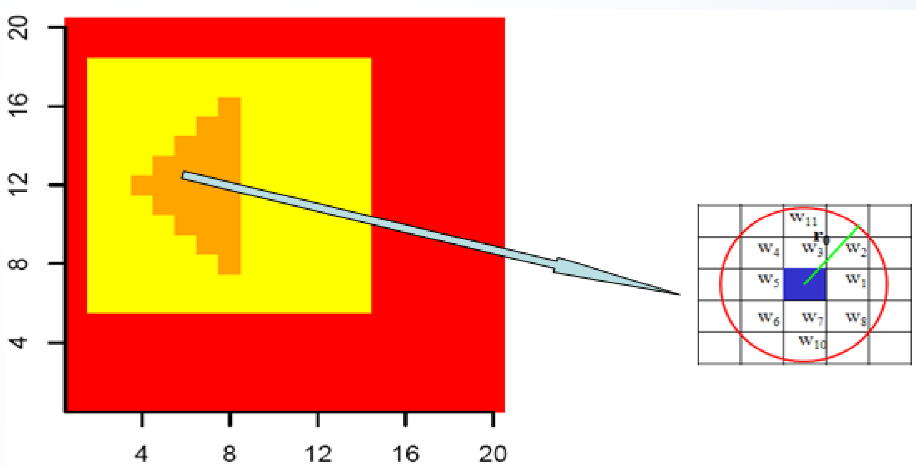
- θ_j : measure the association between the j-th pixel and the class information
  - For example: pearson correlation, test statistics, and so on.
- Define global weight : $w_j = f(\theta_j)$
  - For example: $w_j = \dfrac{p|\theta_j|}{\sum_{j=1}^{p}|\theta_j|}$

Multiscale Weighted PCA

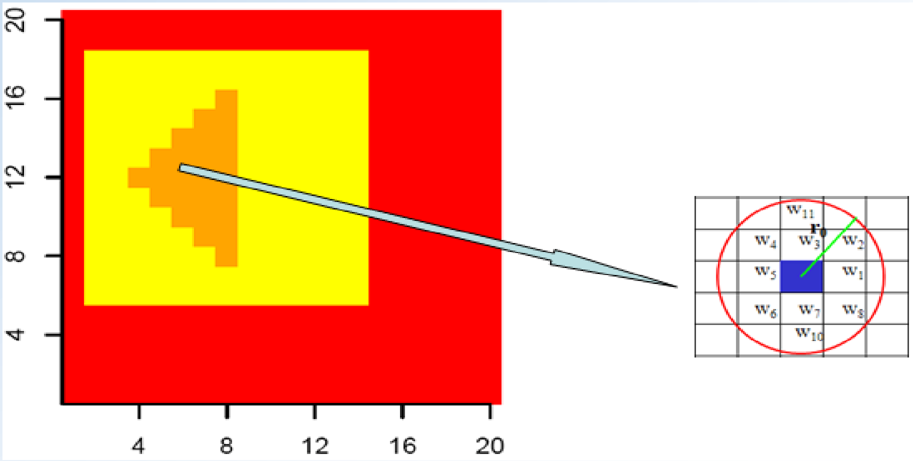
$$\varepsilon = \sum_{i=1}^{n} \sum_{j=1}^{p} w_j \sum_{d \in B(j;h)} w(j,d;h)(\tilde{x}_{i,j} - \tilde{u}_j a_i)^2$$

Two sets of weights:

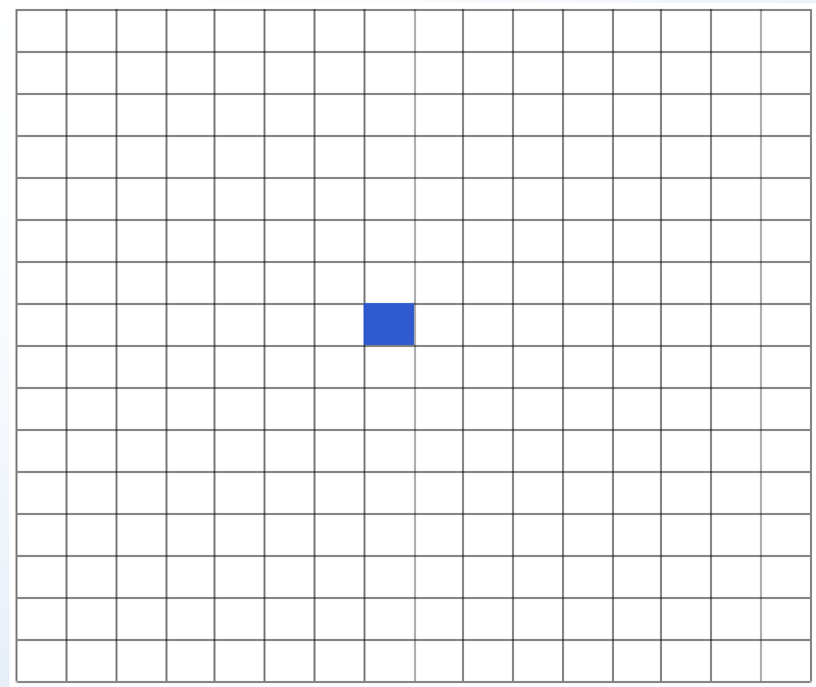
- Global spatial weight: w_j for each pixel/voxel with $\sum_{j=1}^{p} w_j = p$
- Local spatial weight: $w(j,d;h)$ for each pixel/voxel d in the neigborhood $B(j;h)$ (with radius h) of pixel/voxel j, with $\sum_{d \in B(j;h)} w(j,d;h) = 1$



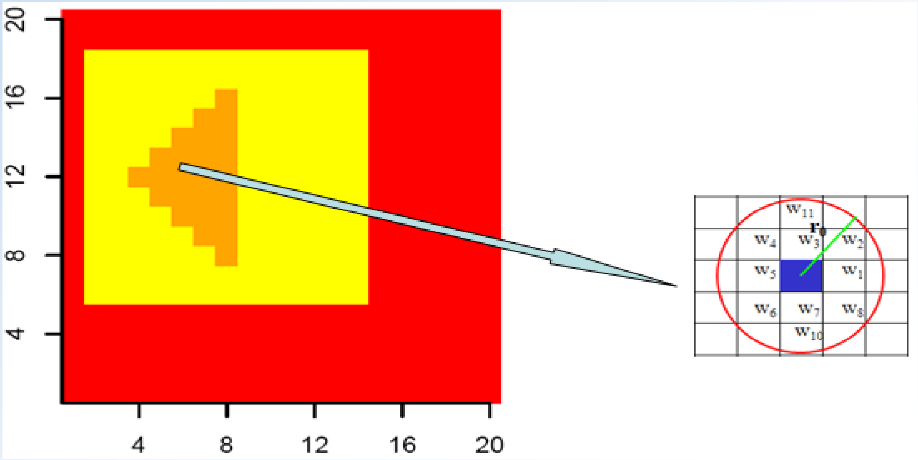
Local Spatial Weight



- Weight adaptation
- Stopping

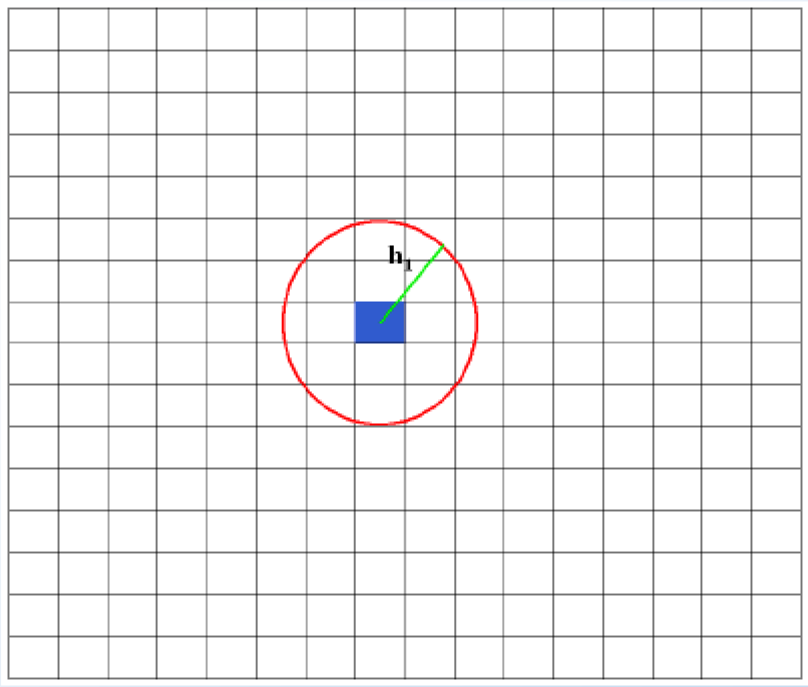


Local Spatial Weight

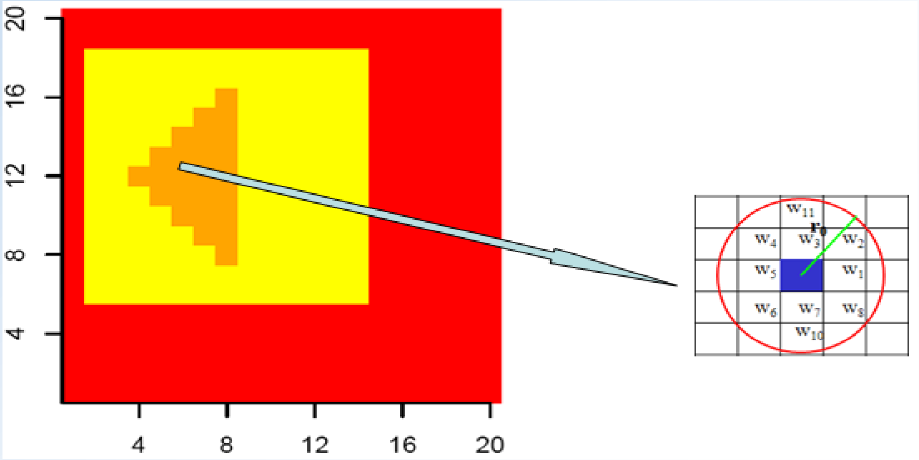


- Weight adaptation

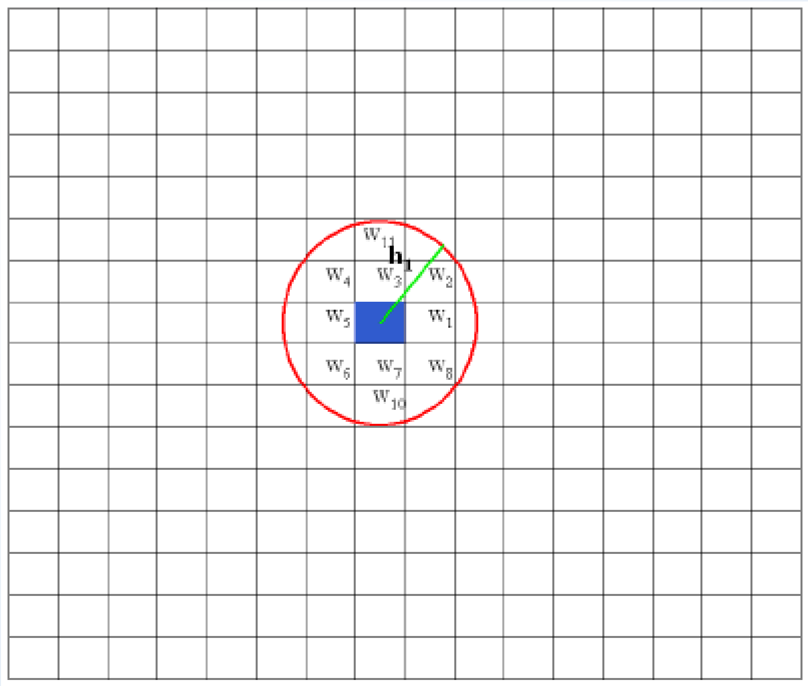
- Stopping



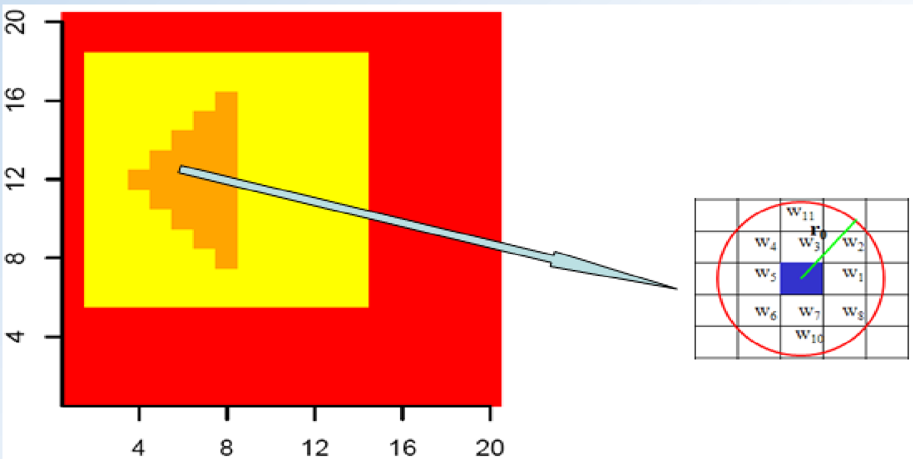
Local Spatial Weight



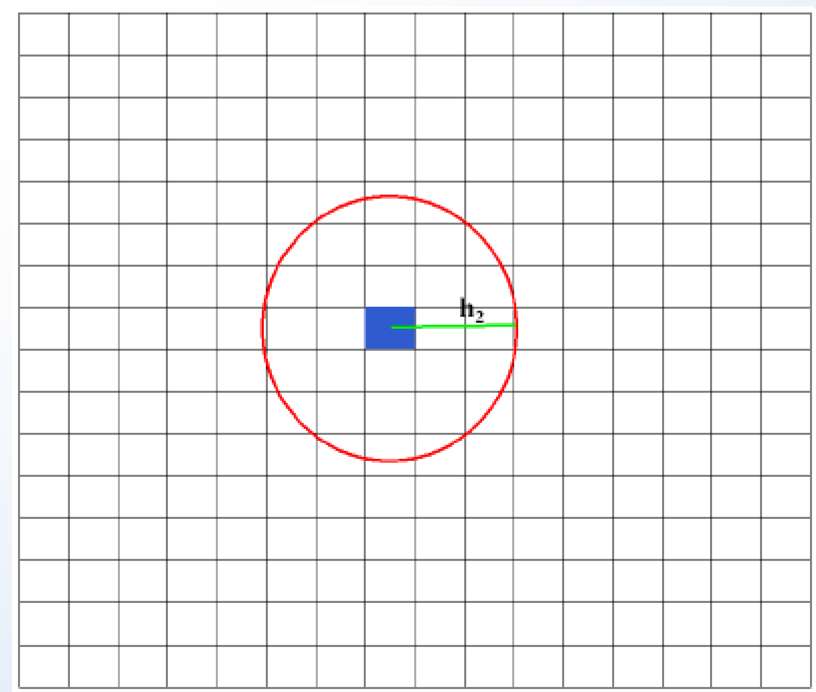
- Weight adaptation
- Stopping



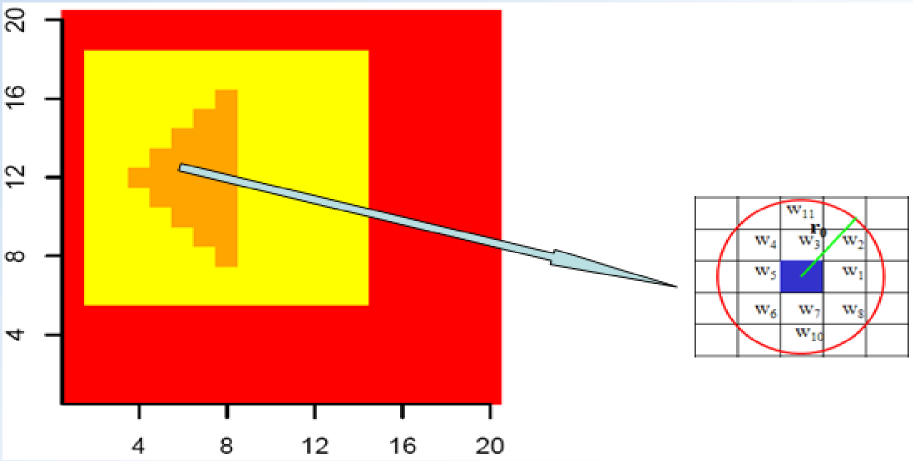
Local Spatial Weight



- Weight adaptation
- Stopping

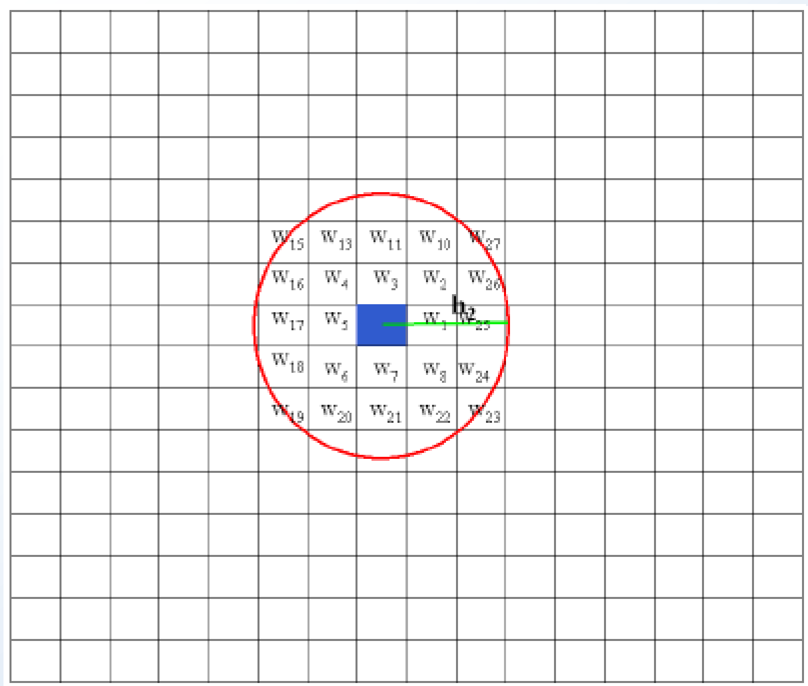


Local Spatial Weight

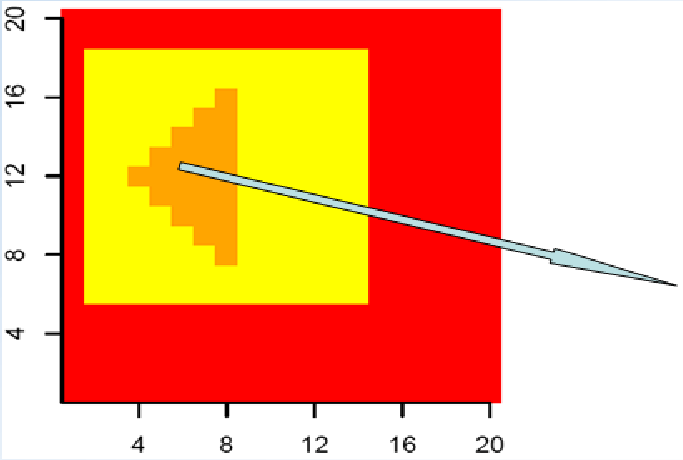


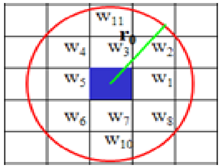
- Weight adaptation

- Stopping

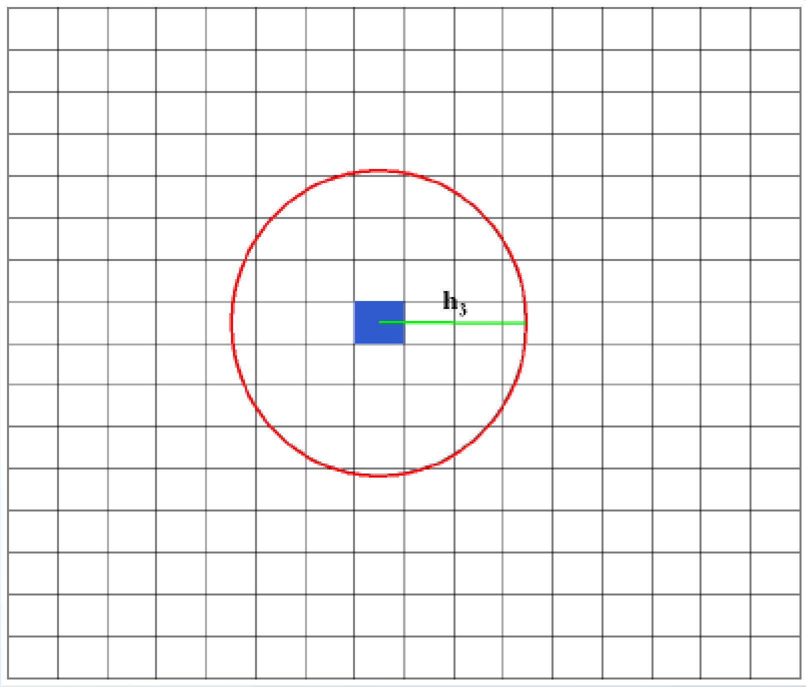


Local Spatial Weight

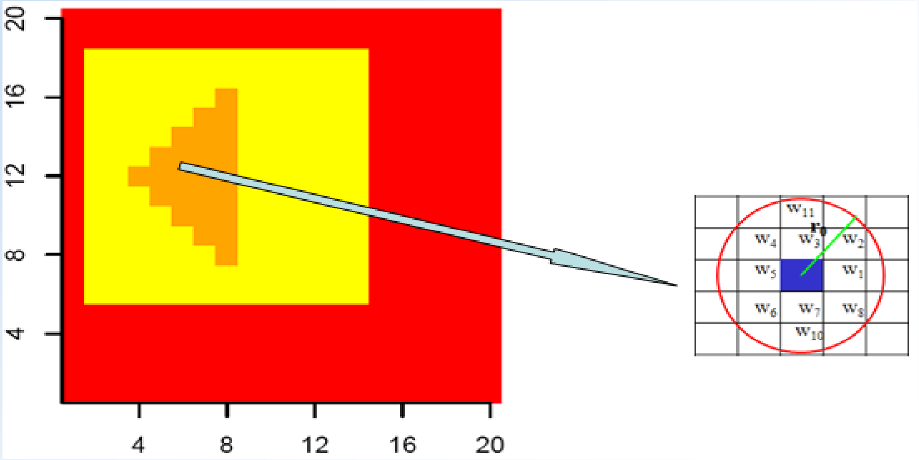




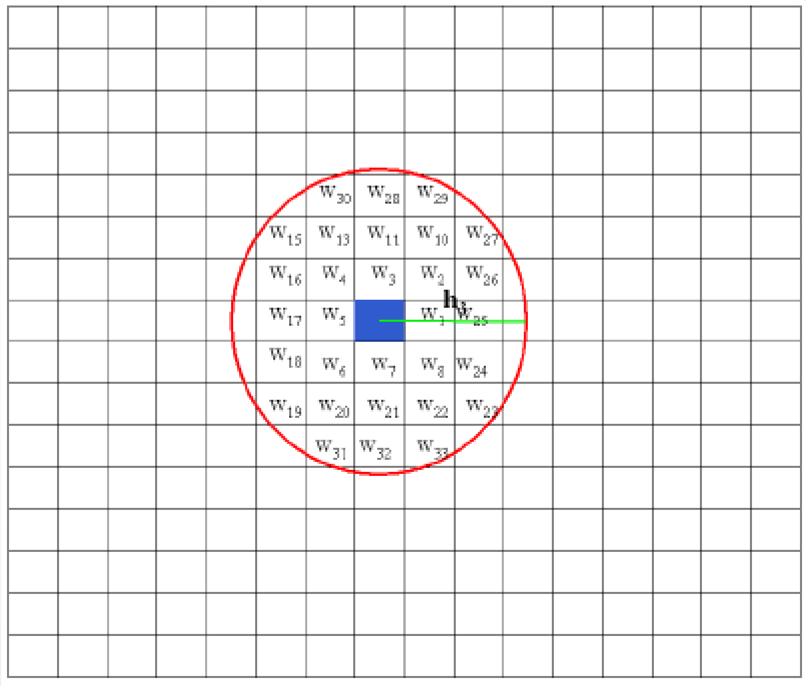
- Weight adaptation
- Stopping



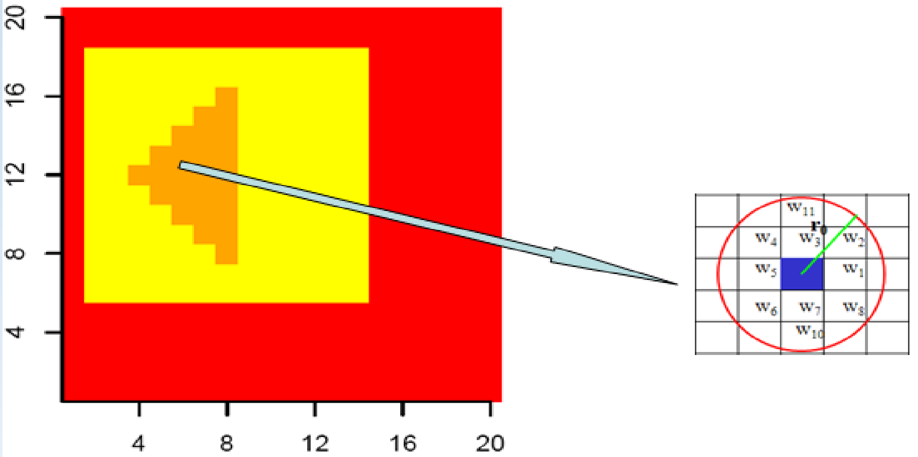
Local Spatial Weight



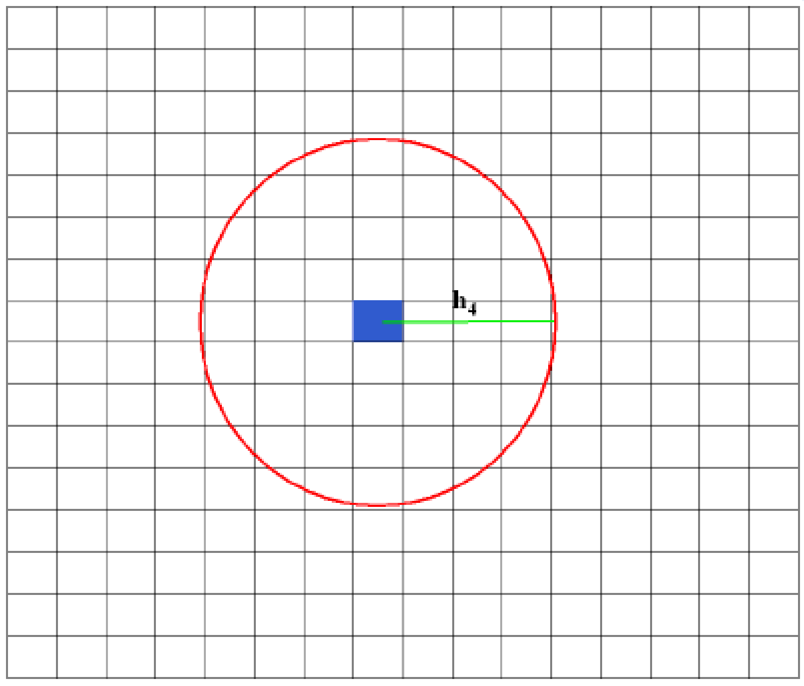
- Weight adaptation
- Stopping



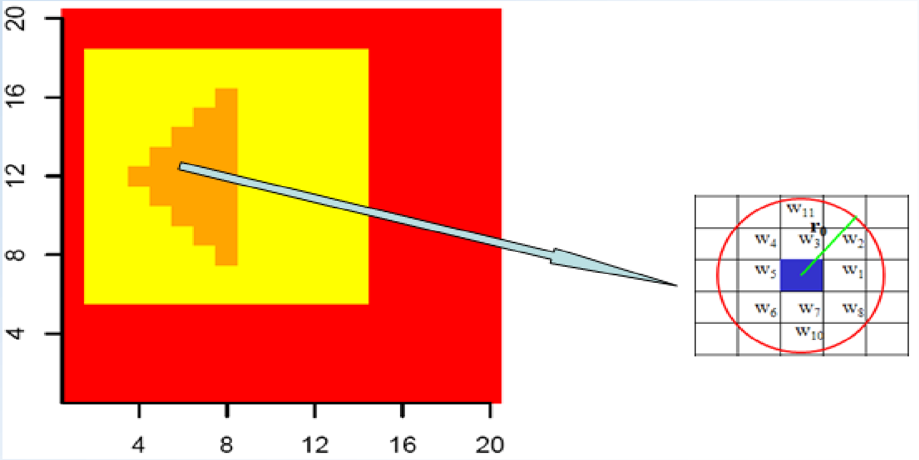
Local Spatial Weight



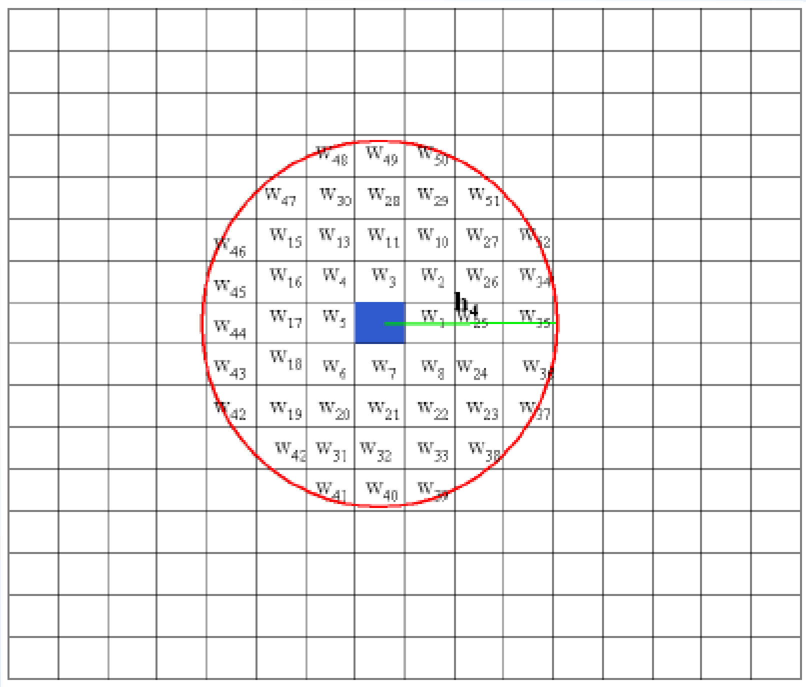
- Weight adaptation
- Stopping



Local Spatial Weight



- Weight adaptation
- Stopping



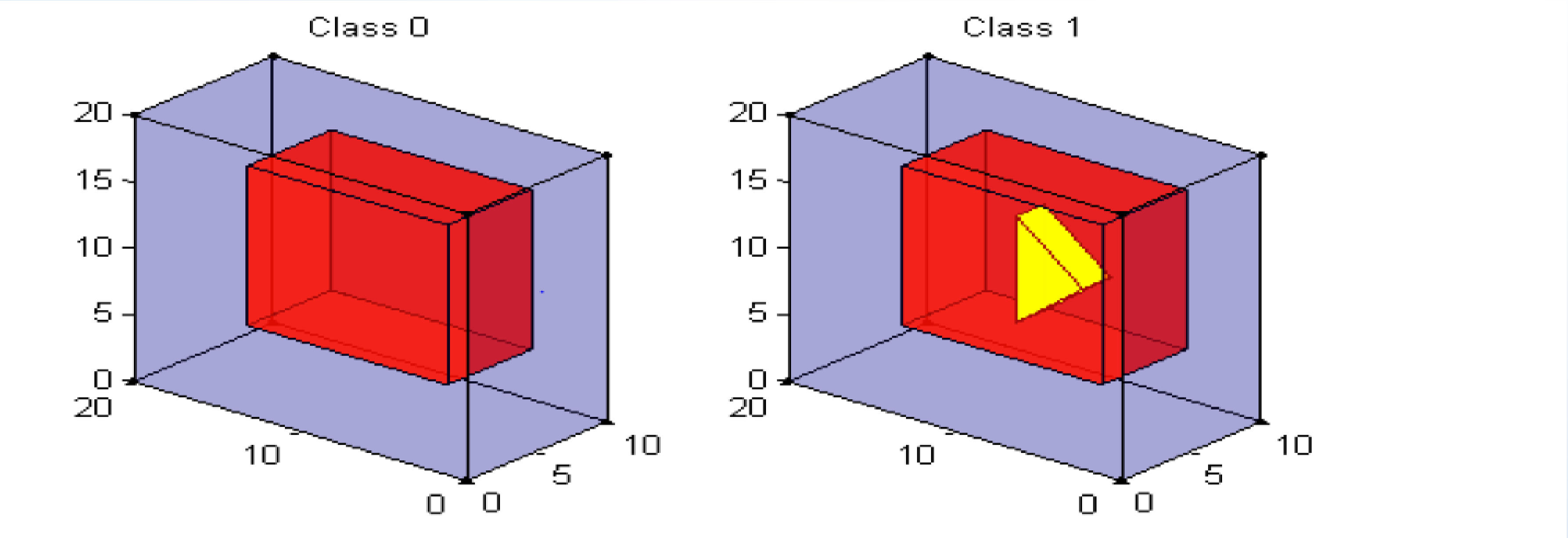
Local Spatial Weight

$$w(j,d;h) = K_{loc}(D_1(j,d)/h)K_{st}(D_2(j,d)/C_n)$$

where $K_{loc}(u)$ and $K_{st}(u)$ are two decreasing kernel functions.

- Distance kernel $K_{loc}(u)$: more weights on the closer voxels
- Similarity kernel $K_{st}(u)$: more weights on the similar voxels

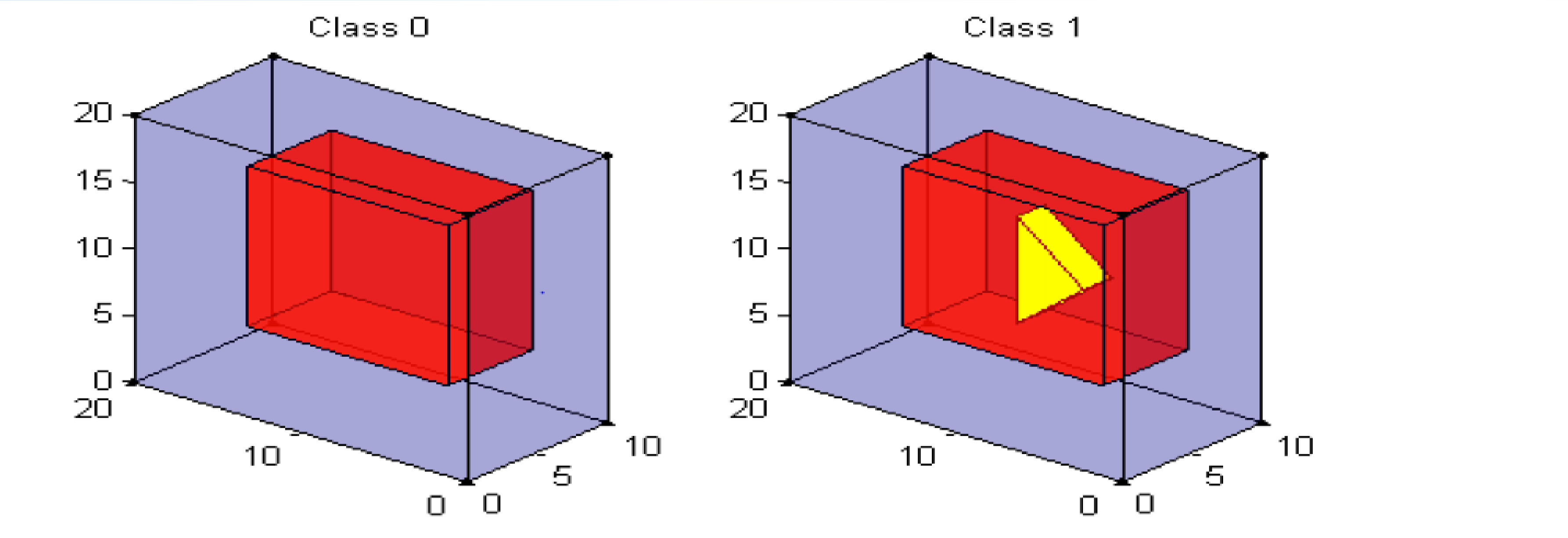
Simulation



Generate two group simulation images

- First group contains 40 images, whose true image is from class 0
- Second group contains 60 images, whose true image is from class 1

Simulation



Classification Error	PCA	SPCA	WPCA	MWPCA
K-NN	0.338 (0.071)	0.152 (0.050)	0.186 (0.055)	0.027 (0.025)
SVM	0.327 (0.078)	0.159 (0.055)	0.215 (0.067)	0.028 (0.026)

ADNI Data

Alzheimer's Disease Neuroimaging Initiative (ADNI) data:

- 390 subjects (218 normal controls and 172 AD patients)

Classification Error	**PCA**	**SPCA**	**WPCA**	**MWPCA**
K-NN	0.382 (0.028)	0.343 (0.045)	0.344 (0.052)	0.227 (0.041)
SVM	0.329 (0.029)	0.313 (0.043)	0.310 (0.042)	0.215 (0.032)

Outline

- Multiscale weighted PCA for Image Analysis
- Human brian artery tree analysis

Data Background

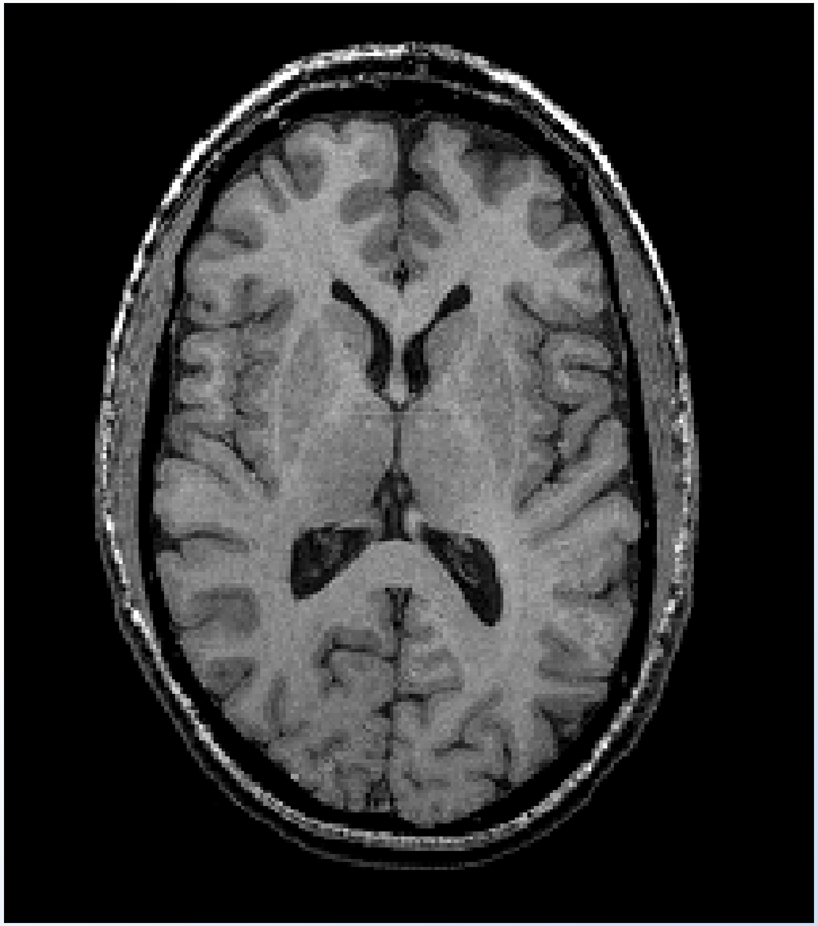
Each Data "Point":

- Tree of Brain Arteries
- For One Person
- Collected by Liz Bullitt

Blood vessel tree data

One Person

- MRI view
- Single Slice
- From 3-d Image



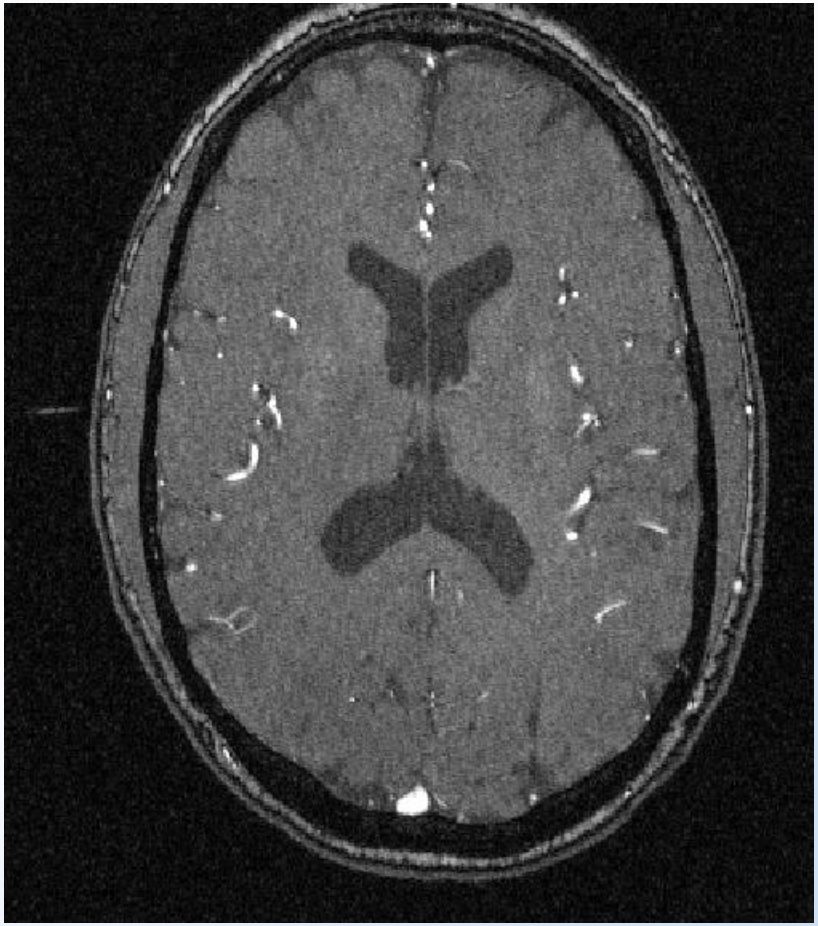
UNIVERSITY OF SOUTH FLORIDA

Blood vessel tree data

One Person’s brain:

- MRA view
- Finds blood vessels

  (show up as white)
- Track through 3d

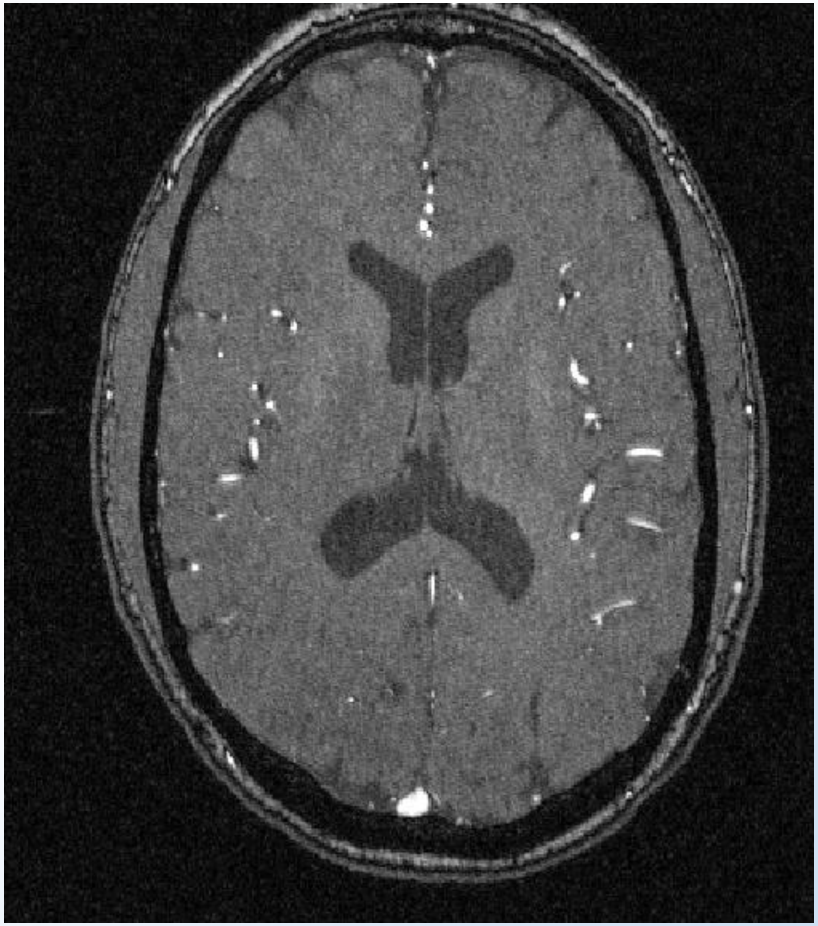


Blood vessel tree data

One Person’s brain:

- MRA view
- Finds blood vessels

  (show up as white)
- Track through 3d

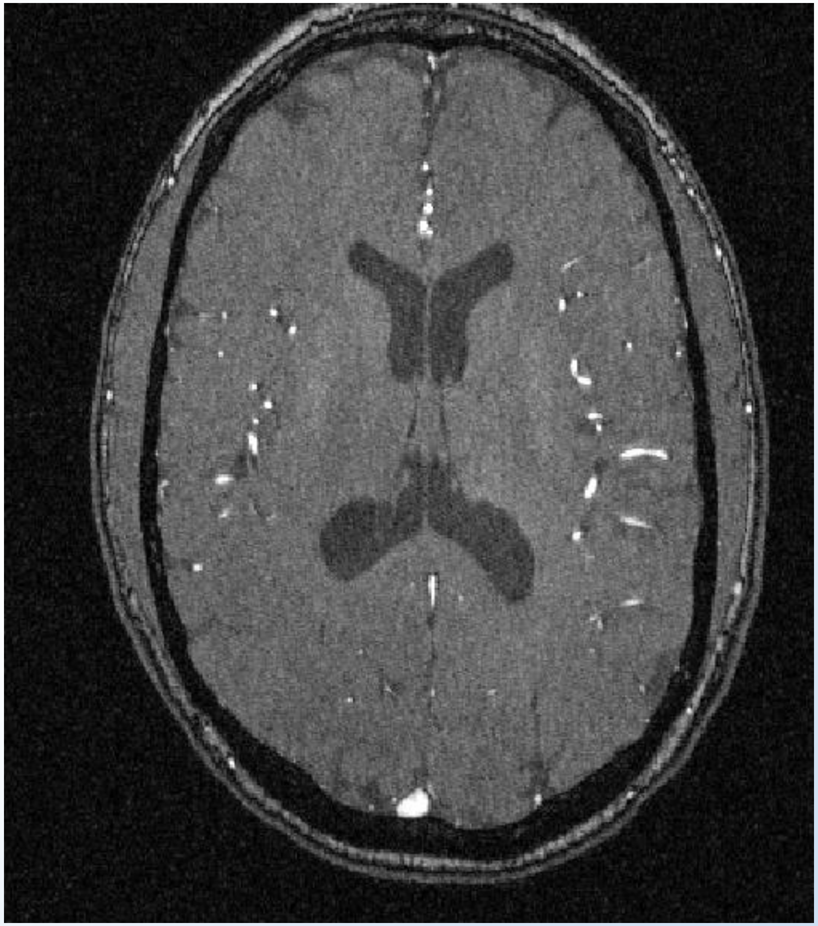


Blood vessel tree data

One Person's brain:

- MRA view
- Finds blood vessels

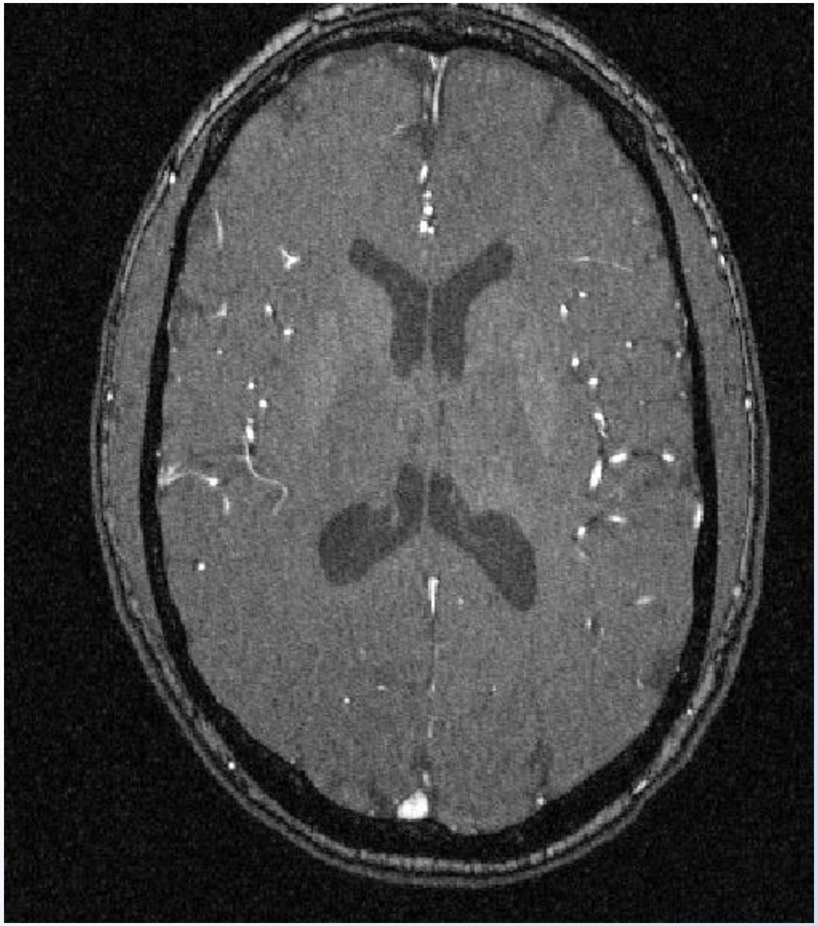
  (show up as white)
- Track through 3d



Blood vessel tree data

One Person's brain:

- MRA view
- Finds blood vessels
  (show up as white)
- Track through 3d

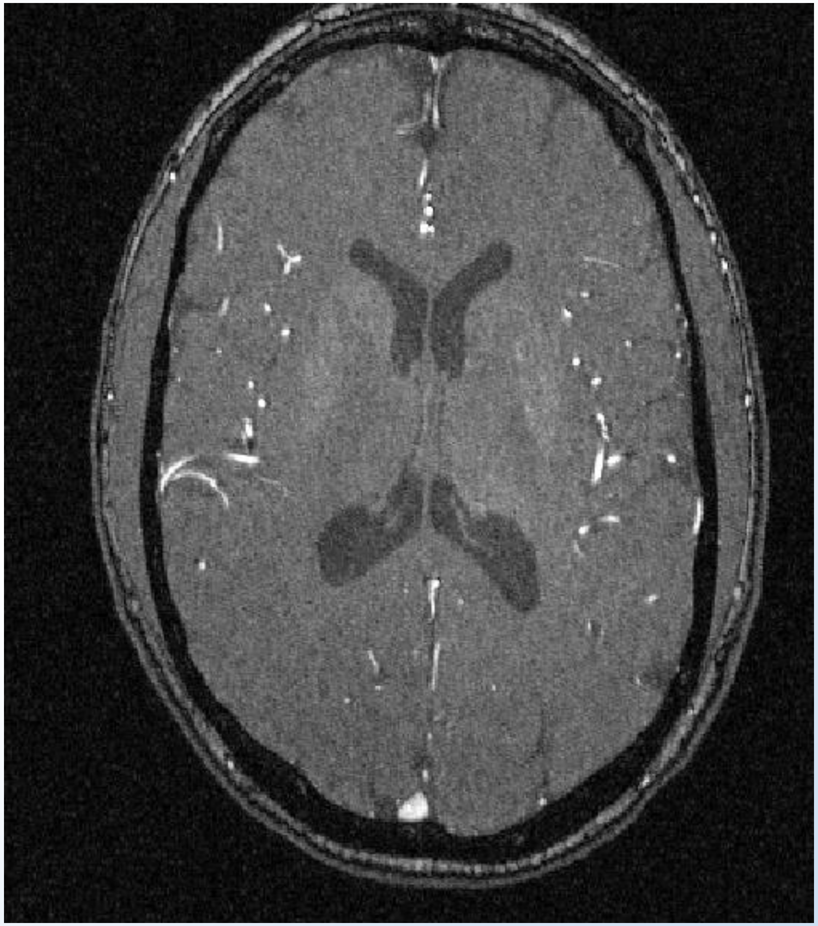


Blood vessel tree data

One Person's brain:

- MRA view
- Finds blood vessels

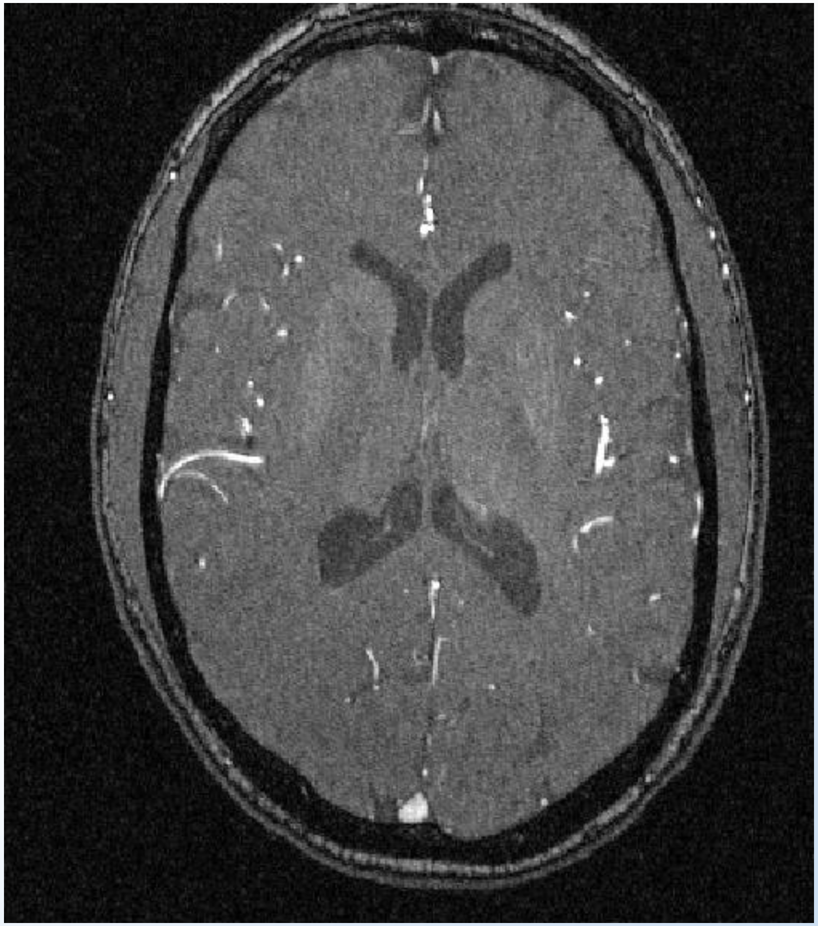
  (show up as white)
- Track through 3d



Blood vessel tree data

One Person's brain:

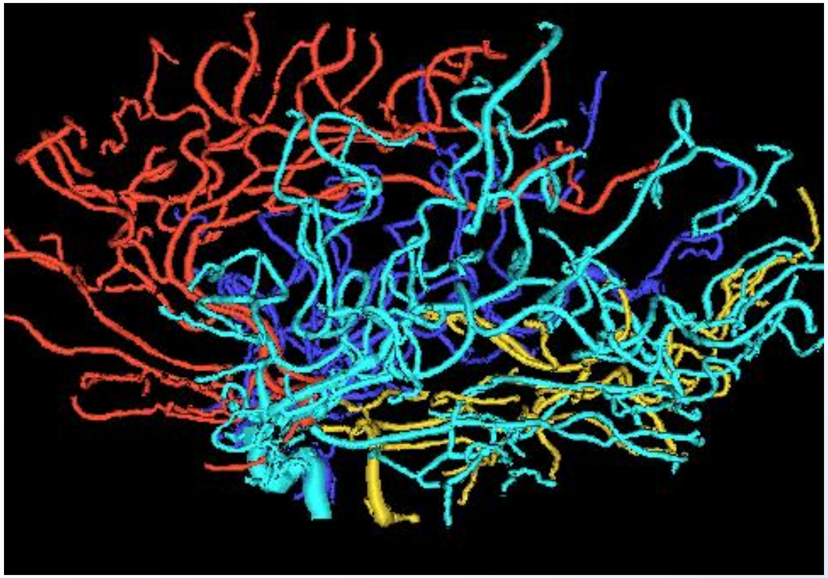
- MRA view
- Finds blood vessels
  (show up as white)
- Track through 3d



Blood vessel tree data

One Person's brain:

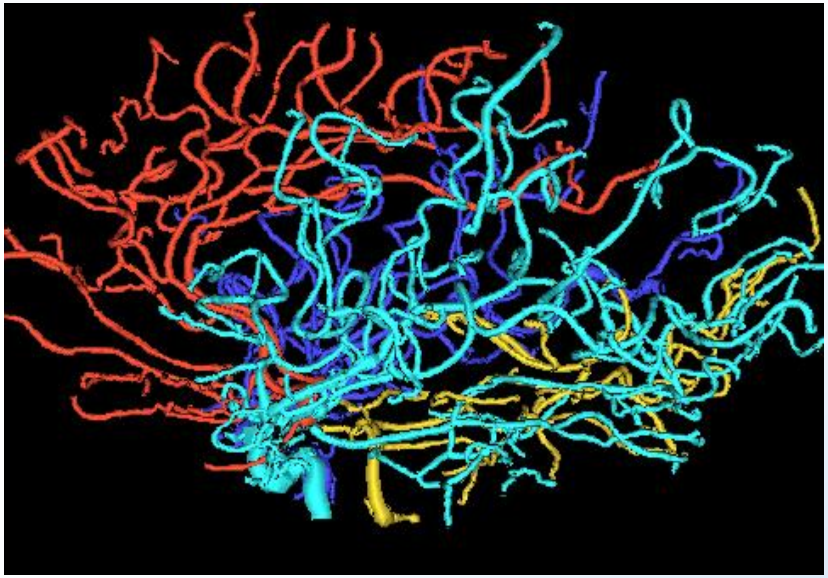
- From MRA
- *Segment* tree
- of vessel segments
- Using *tube tracking*
- Bullitt and Aylward (2002)



Blood vessel tree data

One Person's brain:

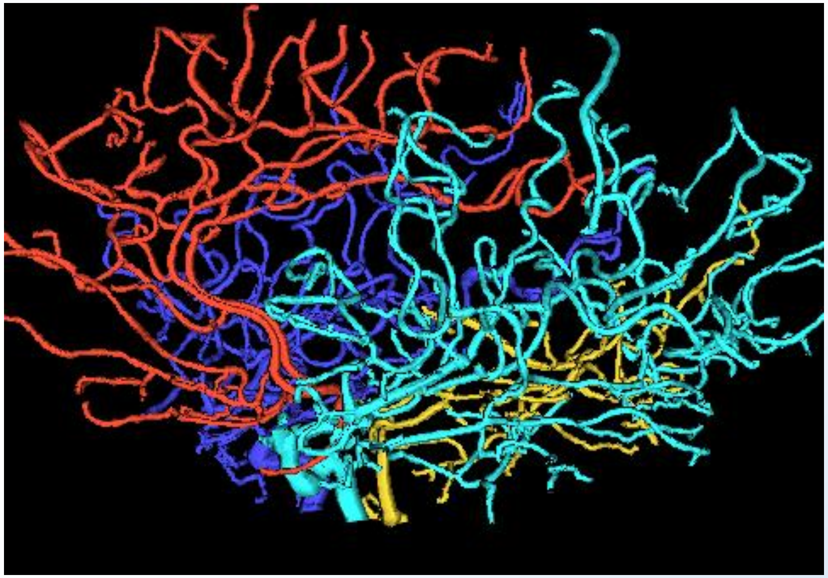
- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

One Person’s brain:

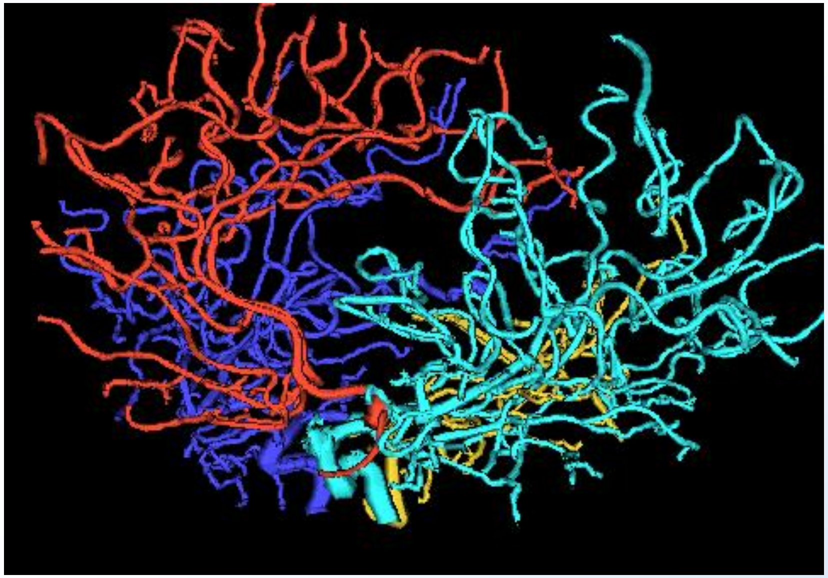
- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

One Person's brain:

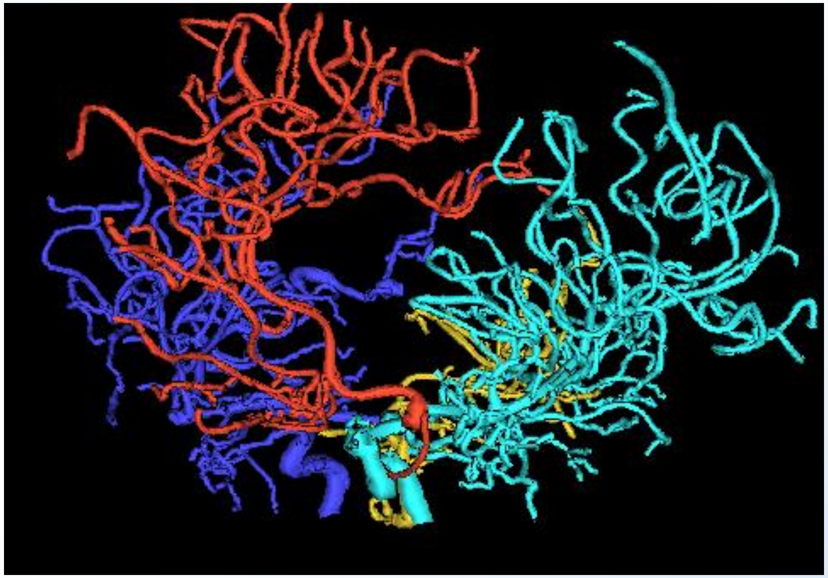
- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

One Person's brain:

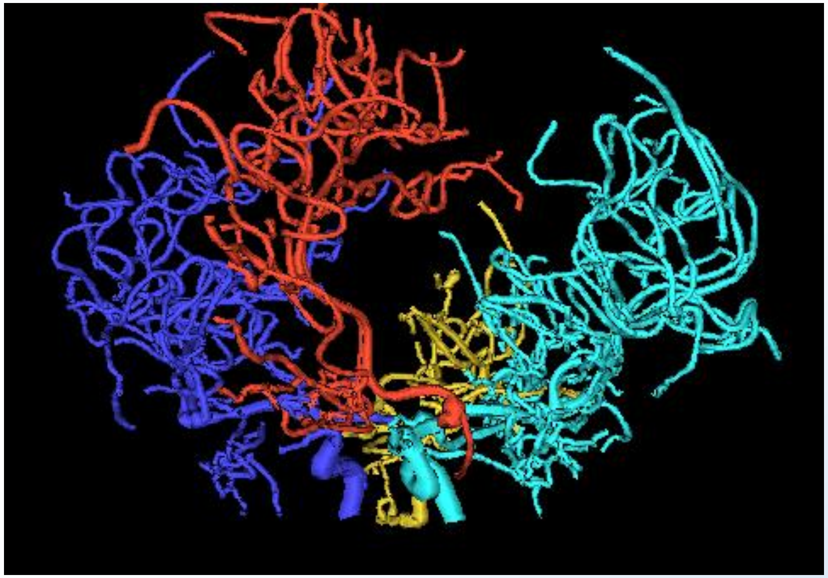
- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



Blood vessel tree data

One Person's brain:

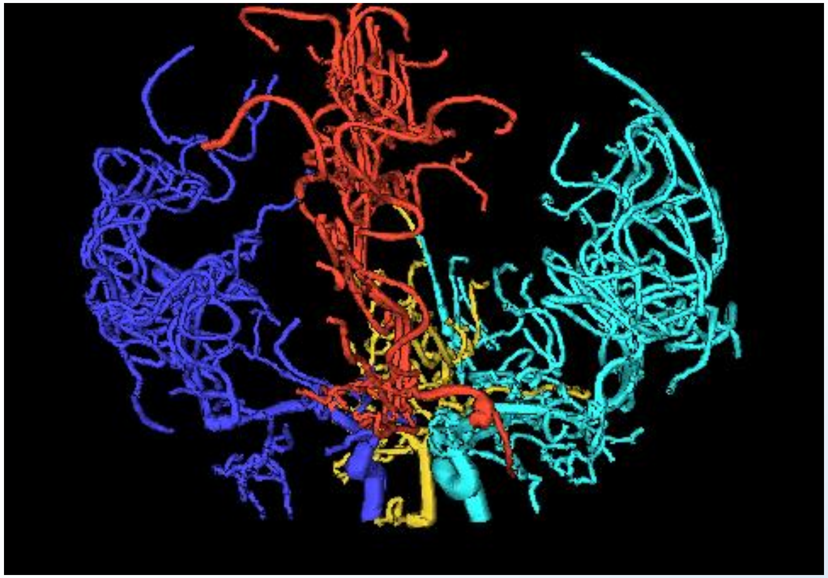
- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



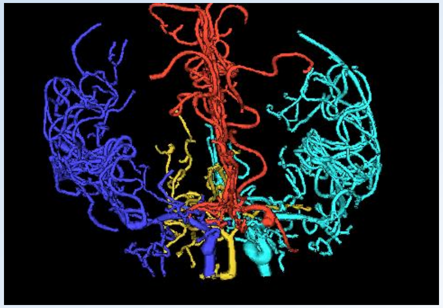
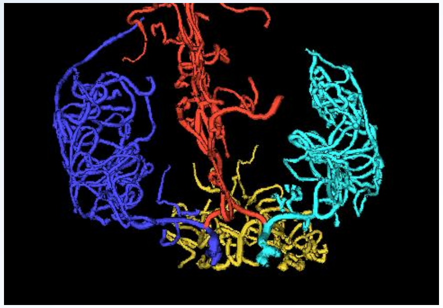
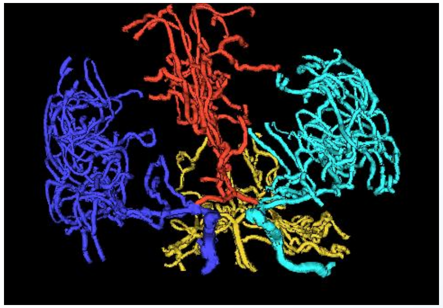
Blood vessel tree data

One Person’s brain:

- From MRA
- Reconstruct trees
- in 3d
- Rotate to view



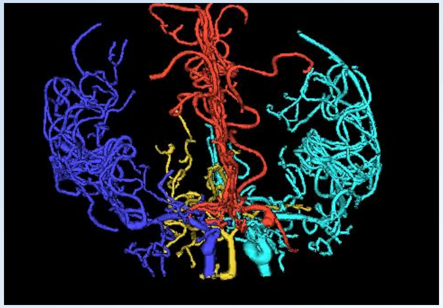
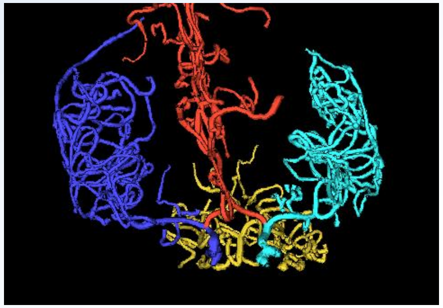
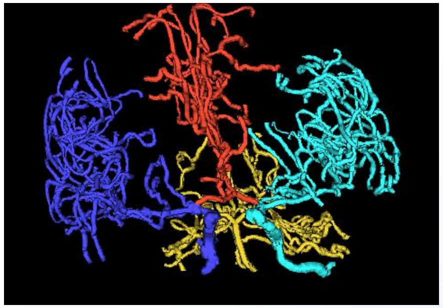
Blood vessel tree data

 ,  , ... , 

- n=98
- Statistical goals:

  1. Structure of Population (understand variation)
  2. Gender difference (Classification)
  3. Age difference
  4. Build model

Blood vessel tree data

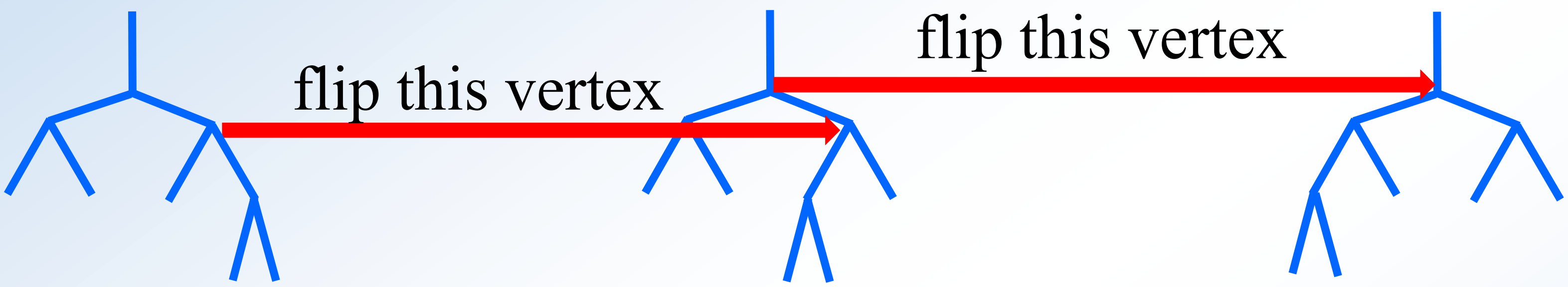
 ,  , … , 

- n=98

- Statistical goals:

  1. Structure of Population (understand variation)
  2. Gender difference (Classification)
  3. Age difference
  4. Build model

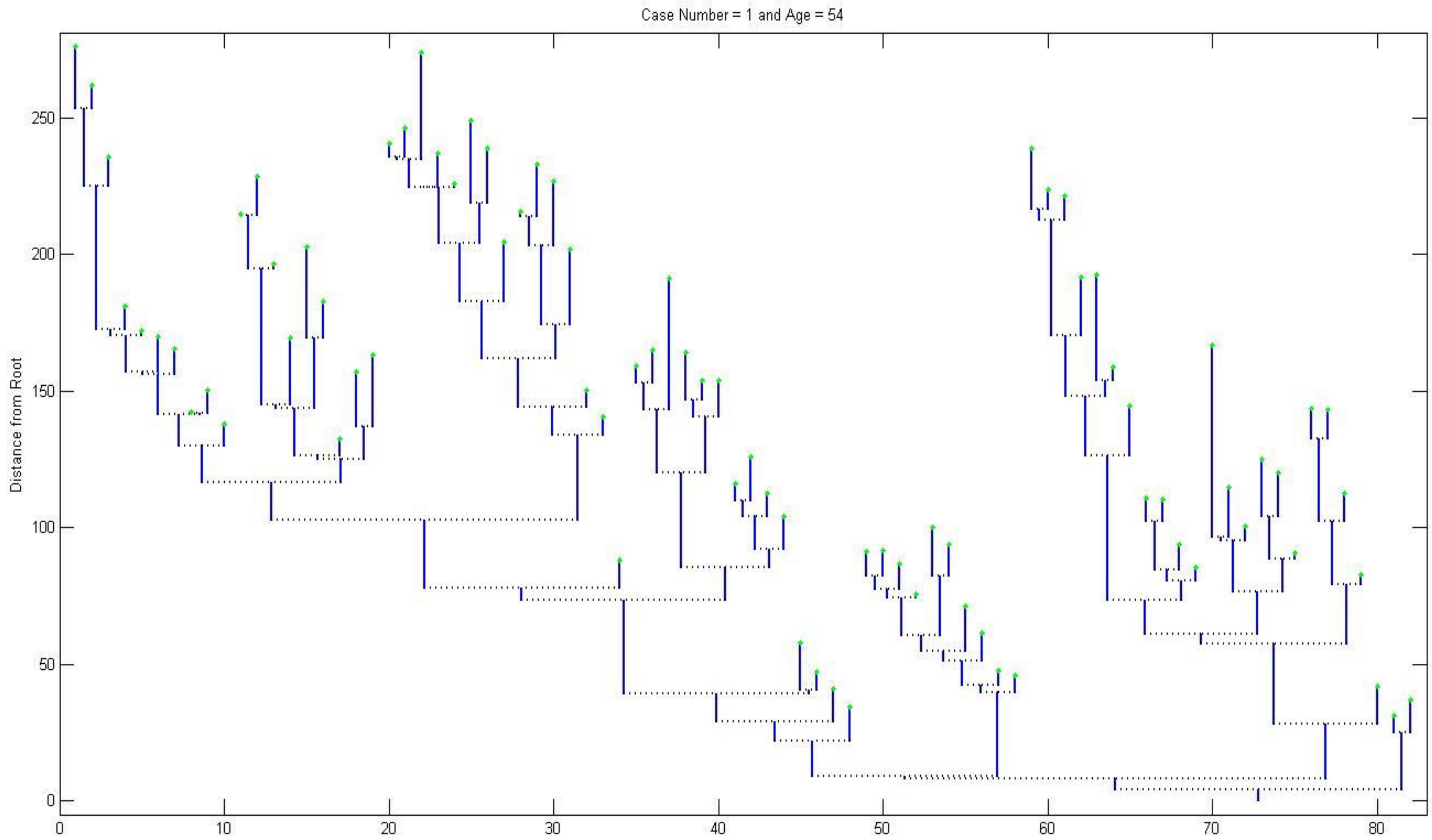
Descendant Correspondence



- Embed 3-d tree in 2-d
- More descendants to the left

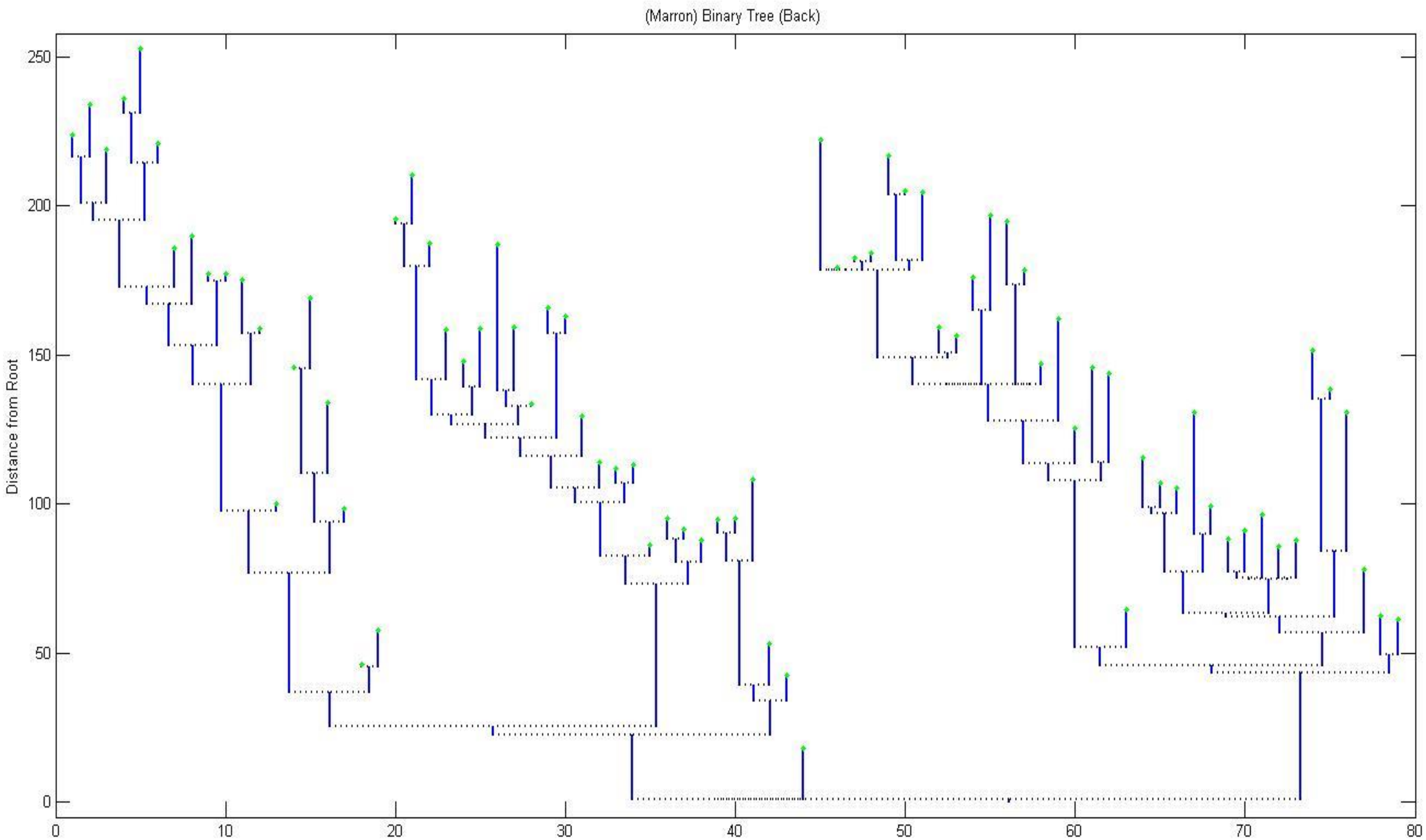
Individual Back Tree

Descendant Correspondence with Branch Length



Marron's Back Tree

Descendant Correspondence with Branch Length

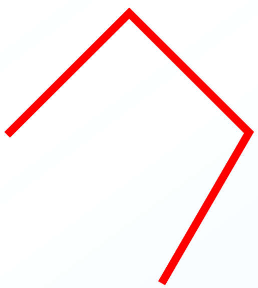


Dyck Path Representation

Example 1, Assume that we have three following trees

Tree 1

Tree 2

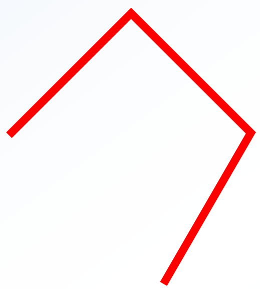


Tree 3

Support Tree: union of trees

Tree 1

Tree 2



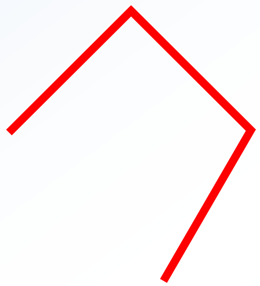
Tree 3

Tree 1

Support Tree: union of trees

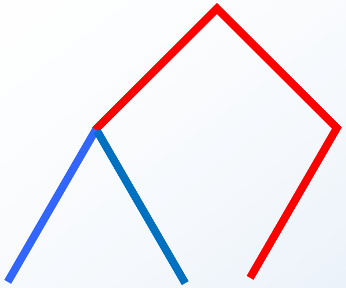
Tree 1

Tree 2



Tree 3

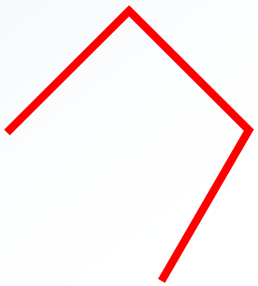
Tree 1,2



Support Tree: union of trees

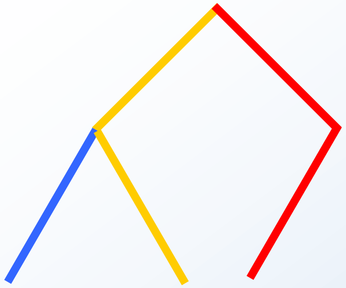
Tree 1

Tree 2



Tree 3

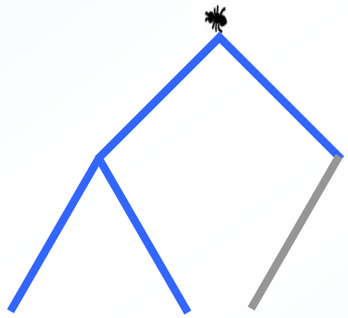
Tree 1,2,3



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

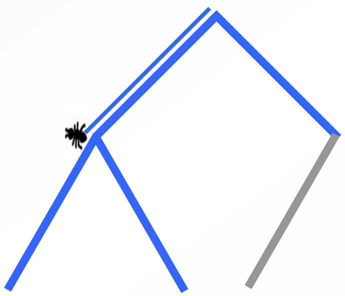
Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

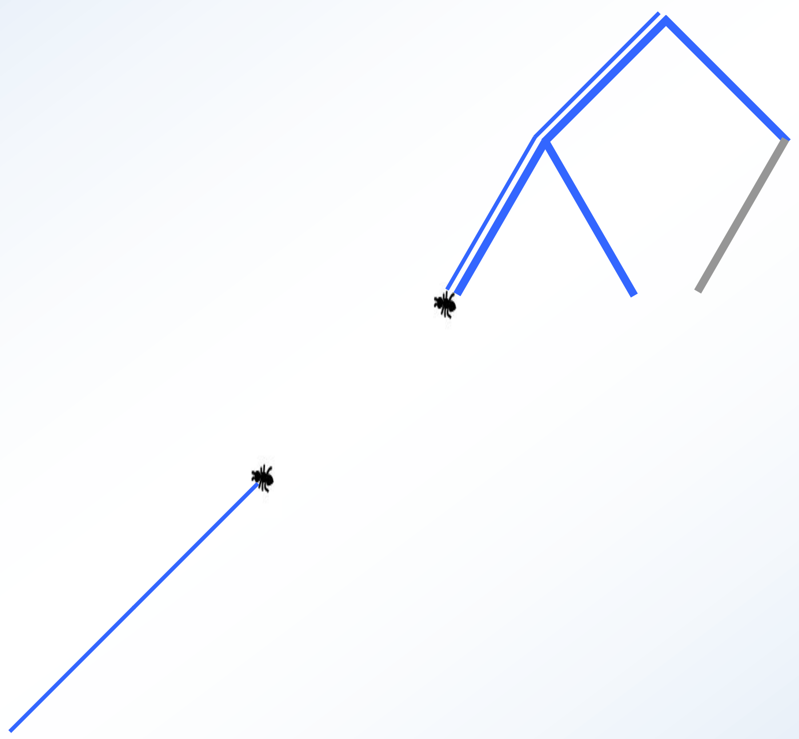
Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

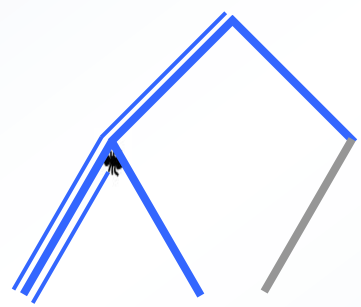
Tree 1/ Support Tree

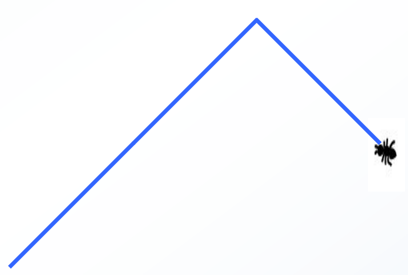


Dyck Path Representation

Now, we show how to transform the first tree as a curve.

Tree 1/ Support Tree

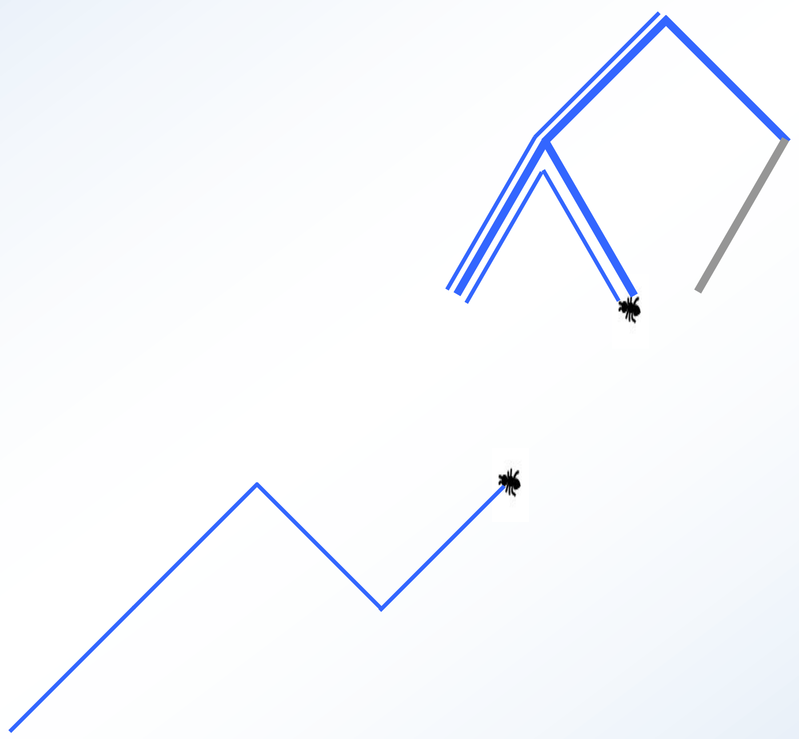




Dyck Path Representation

Now, we show how to transform the first tree as a curve.

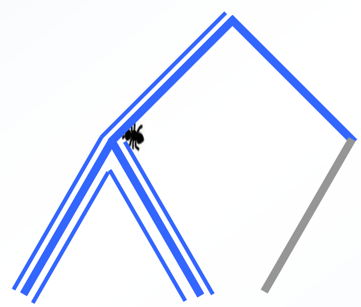
Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

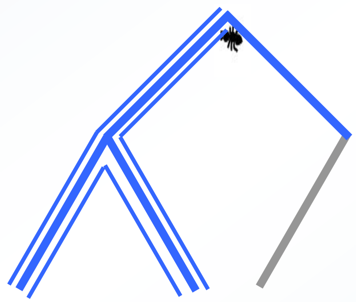
Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the first tree as a curve.

Tree 1/ Support Tree

Dyck Path Representation

Now, we show how to transform the first tree as a curve.

Tree 1/ Support Tree

Dyck Path Representation

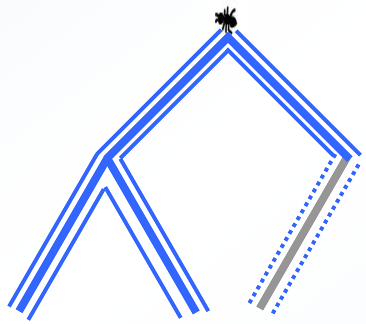
Now, we show how to transform the first tree as a curve.

Tree 1/ Support Tree

Dyck Path Representation

Now, we show how to transform the first tree as a curve.

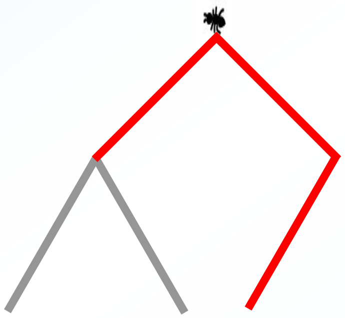
Tree 1/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

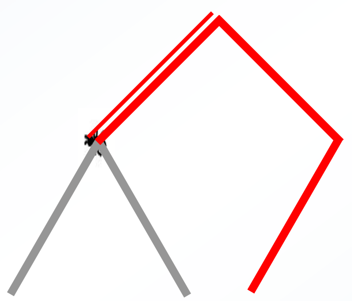
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

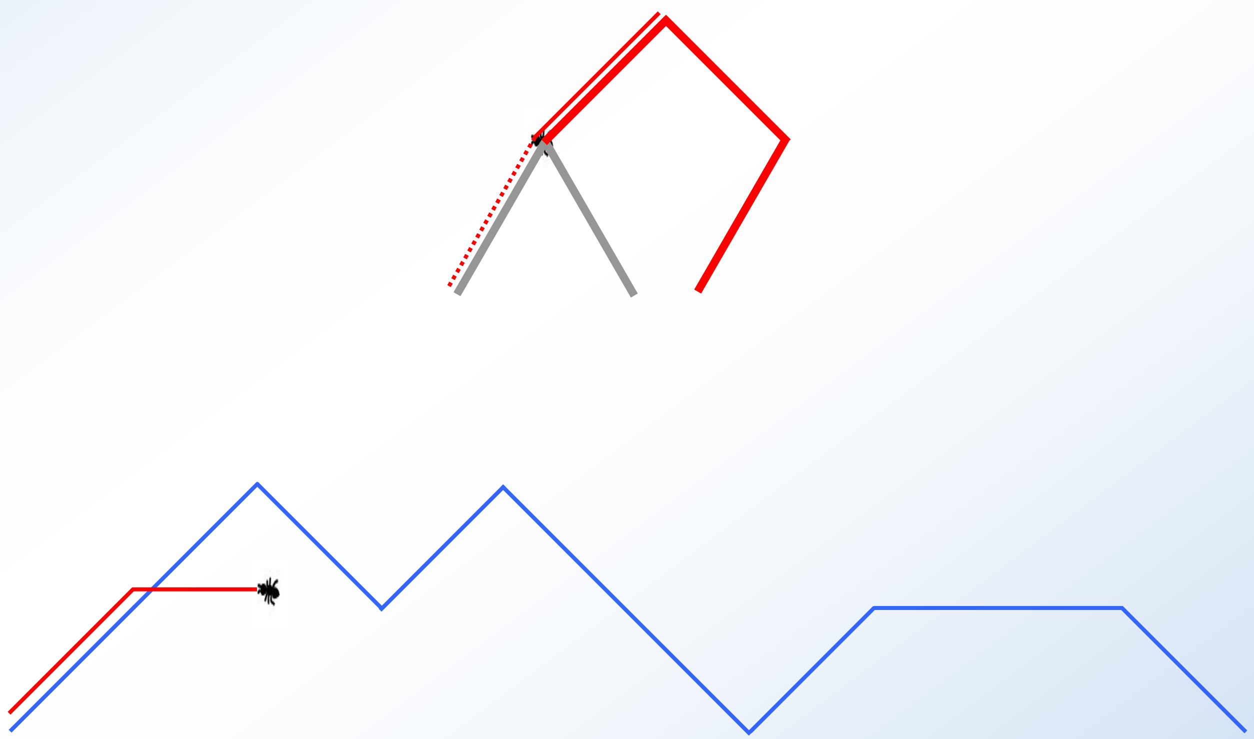
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

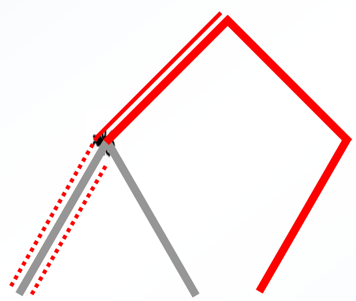
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

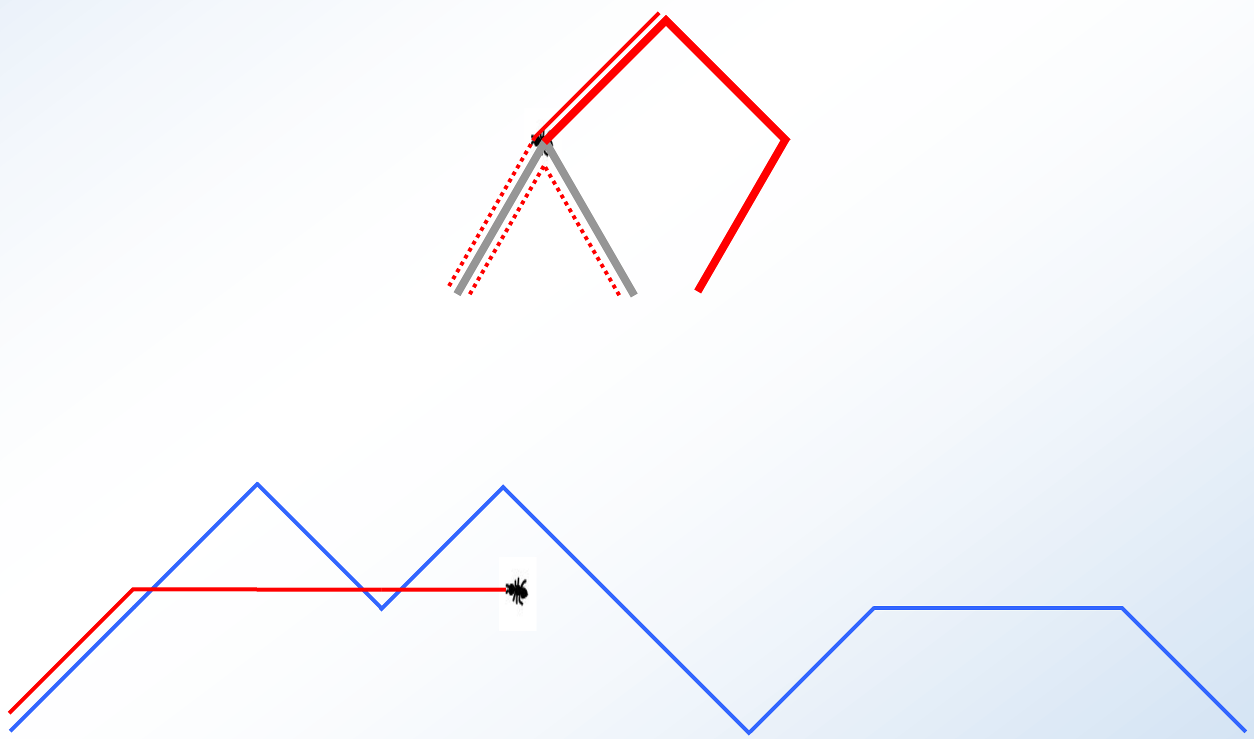
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

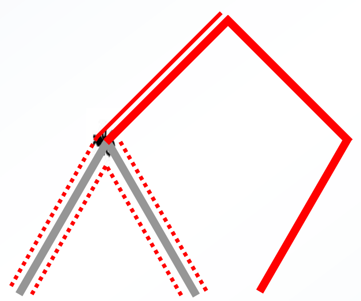
Tree 2/ Support Tree

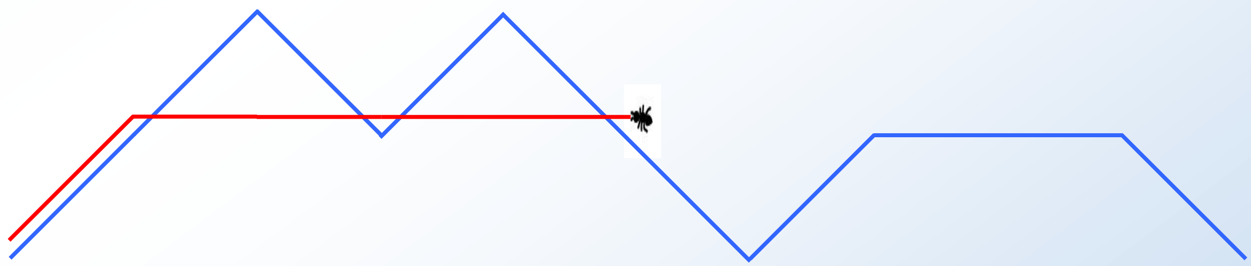


Dyck Path Representation

Now, we show how to transform the second tree as a curve.

Tree 2/ Support Tree

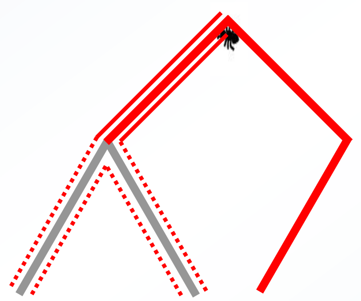


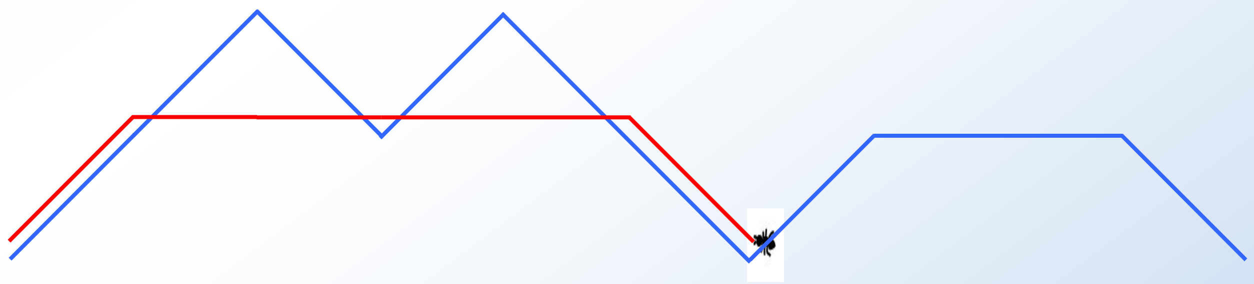


Dyck Path Representation

Now, we show how to transform the second tree as a curve.

Tree 2/ Support Tree

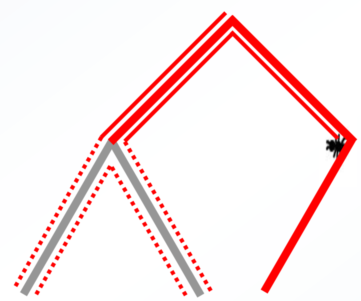


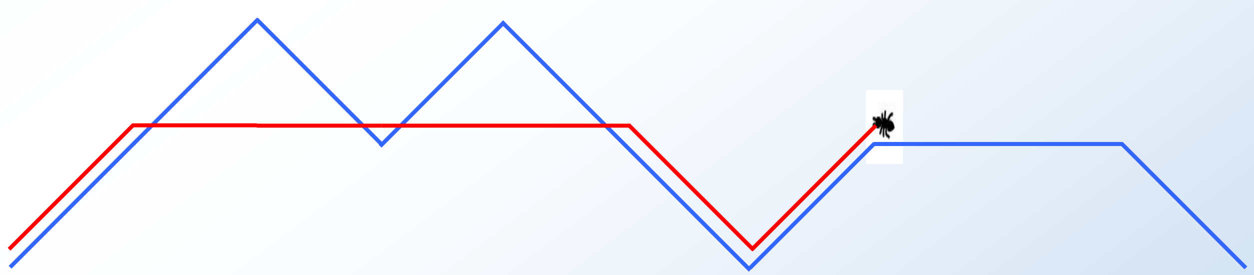


Dyck Path Representation

Now, we show how to transform the second tree as a curve.

Tree 2/ Support Tree

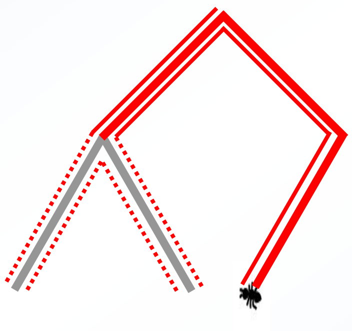




Dyck Path Representation

Now, we show how to transform the second tree as a curve.

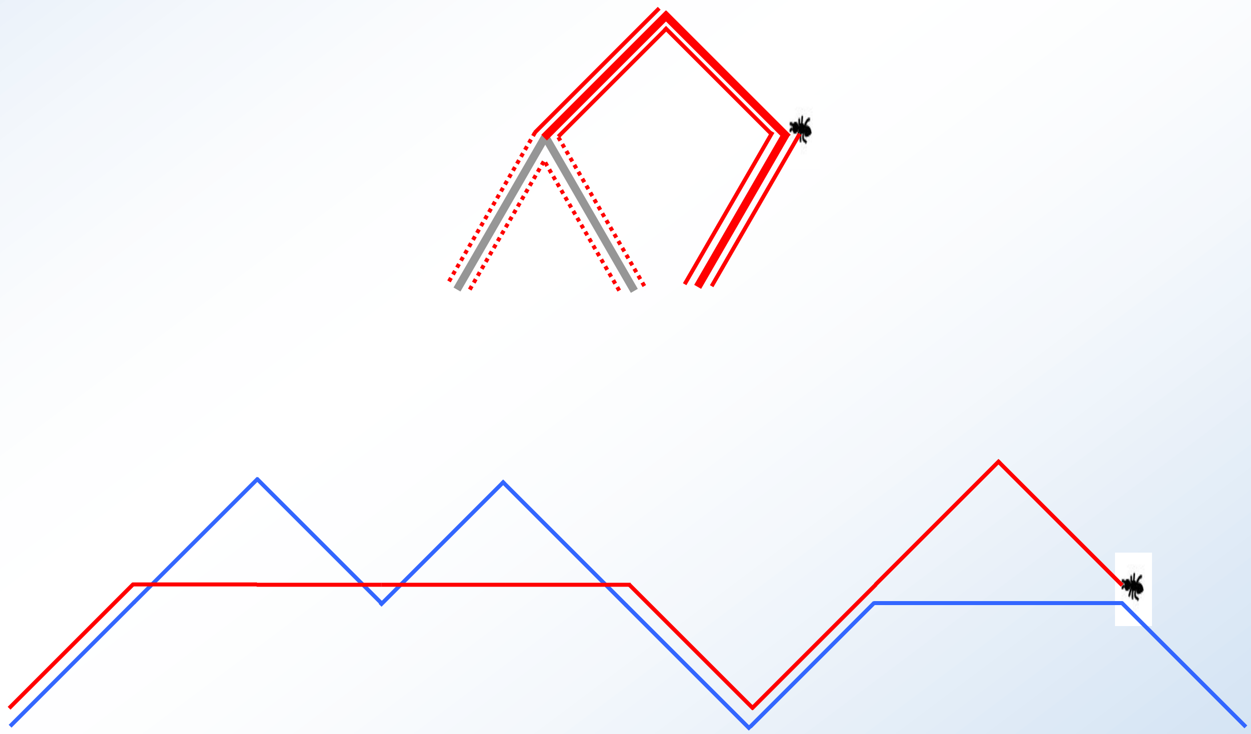
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

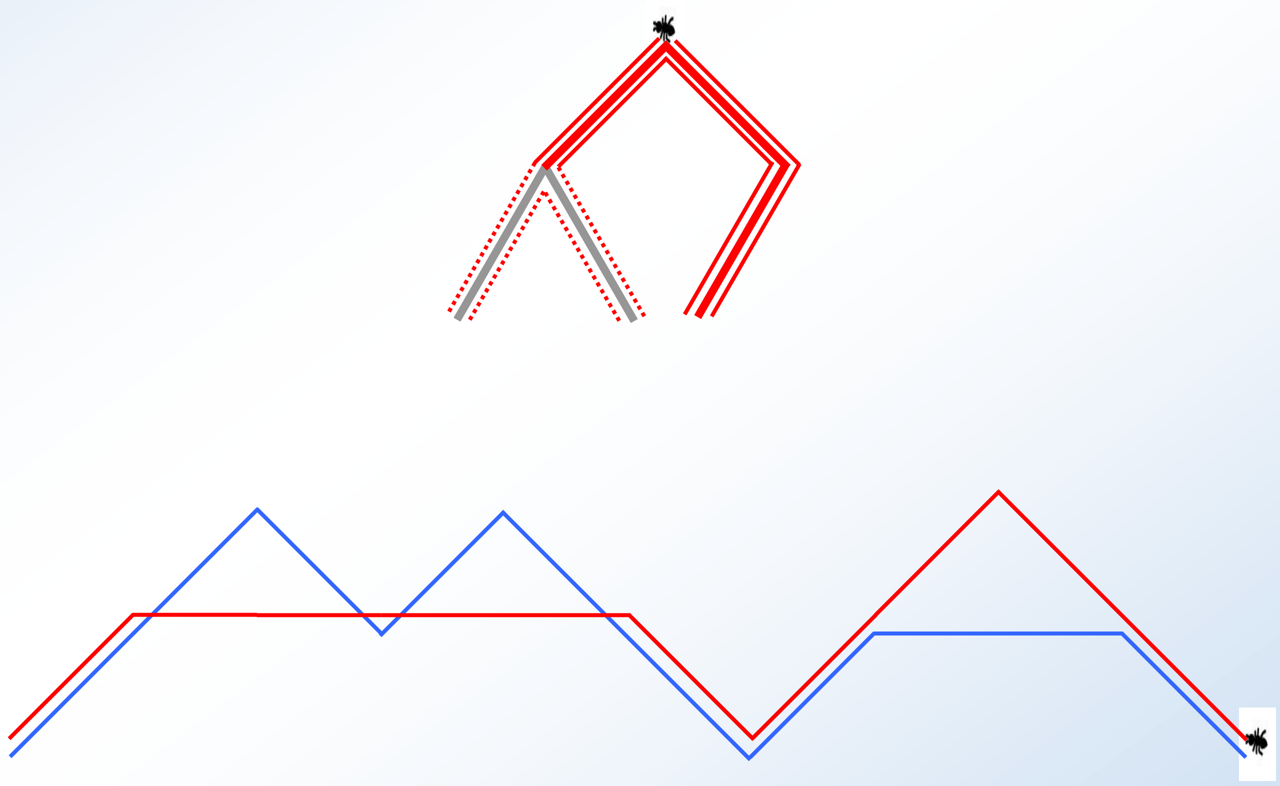
Tree 2/ Support Tree



Dyck Path Representation

Now, we show how to transform the second tree as a curve.

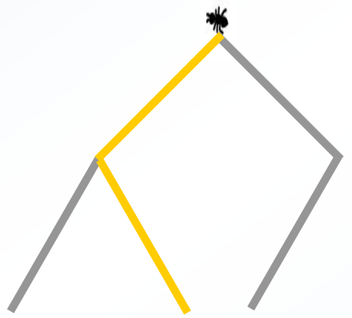
Tree 2/ Support Tree

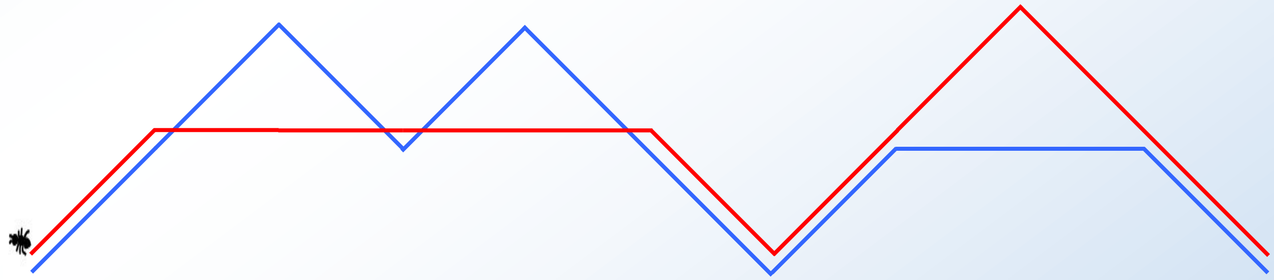


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

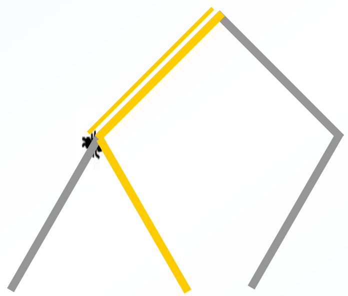


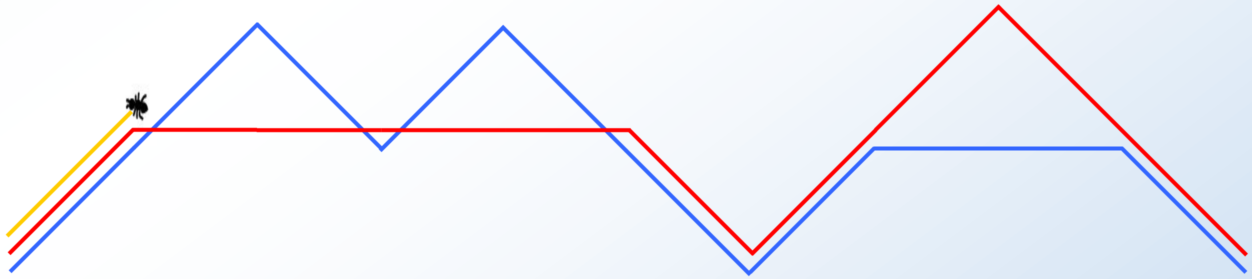


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

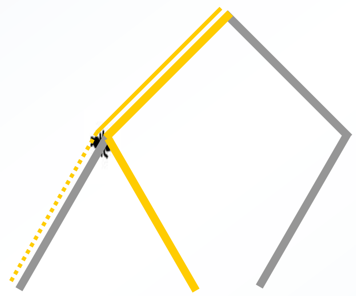


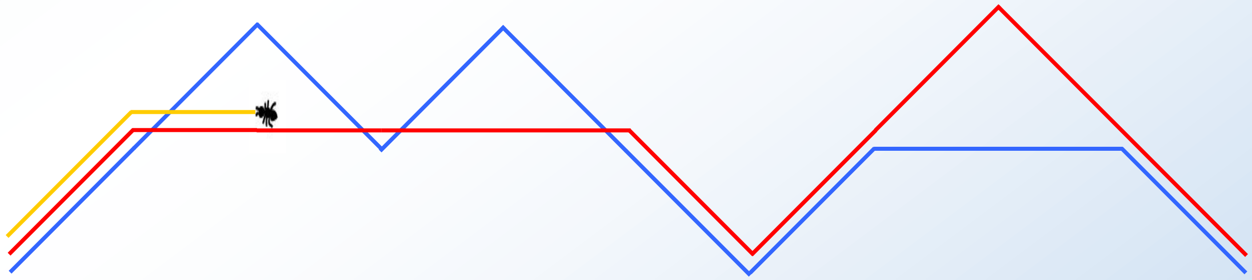


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

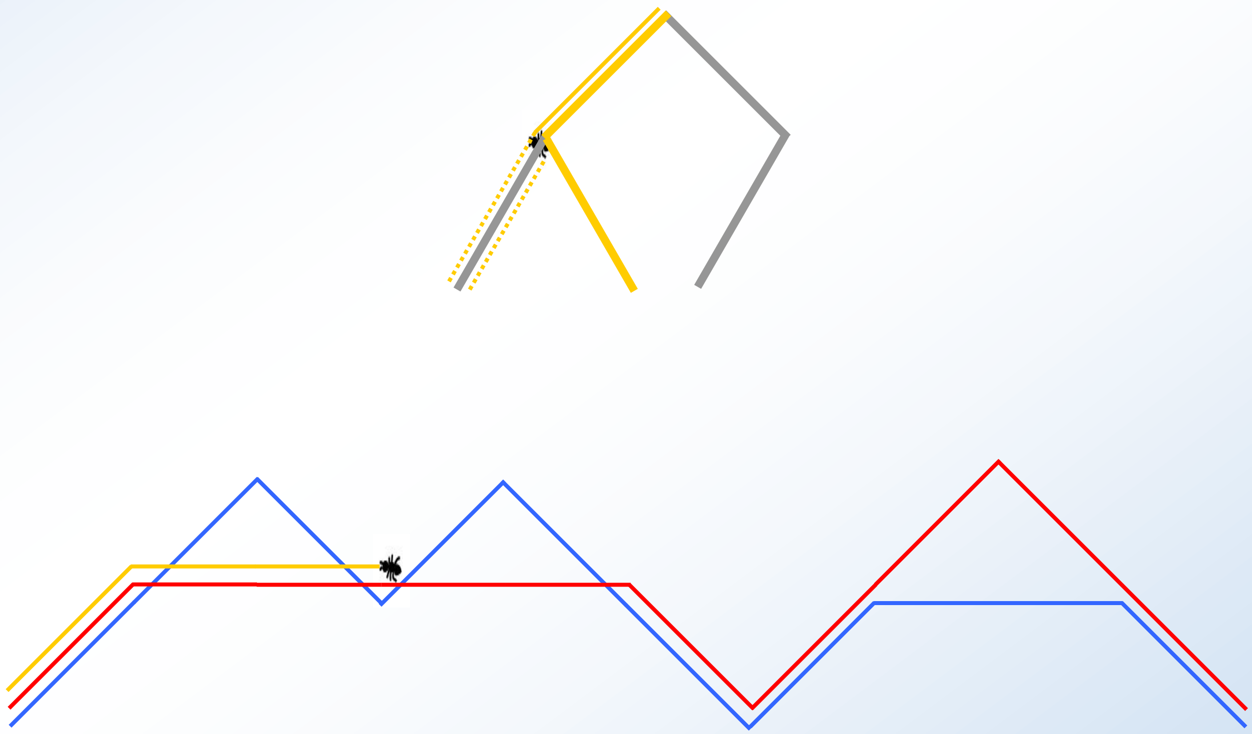




Dyck Path Representation

Now, we show how to transform the third tree as a curve.

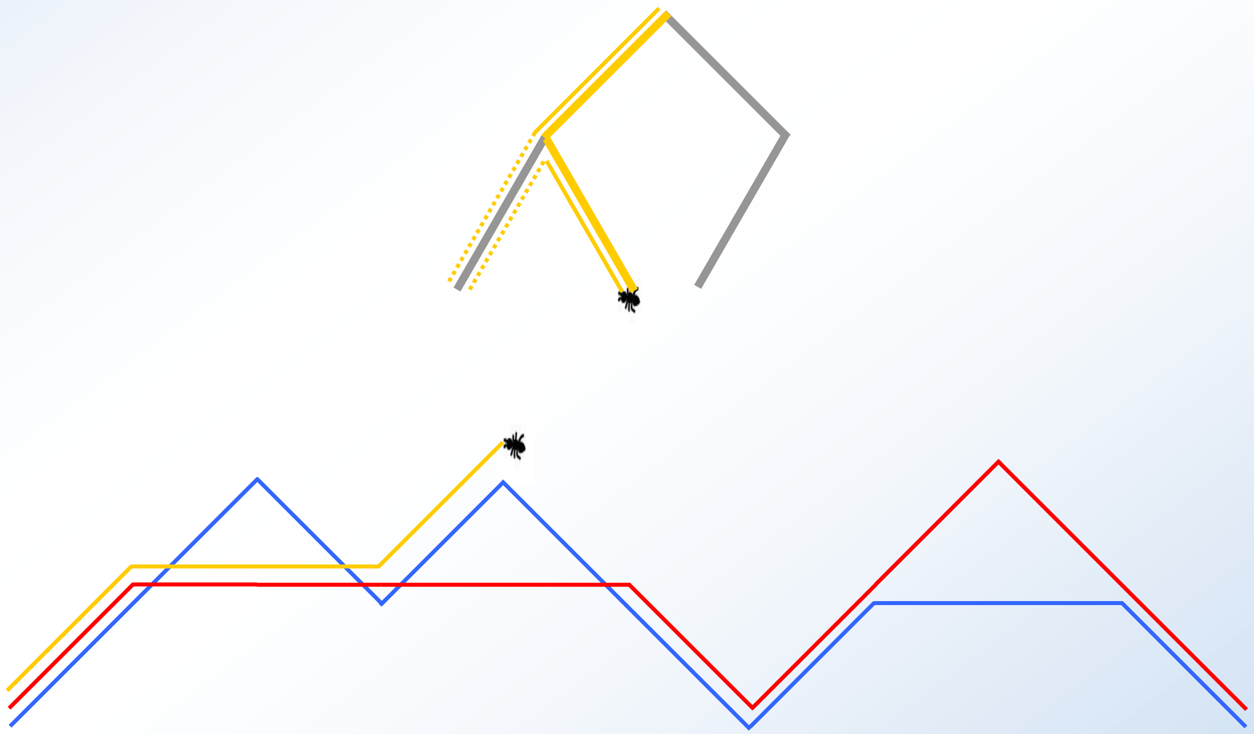
Tree 3/ Support Tree



Dyck Path Representation

Now, we show how to transform the third tree as a curve.

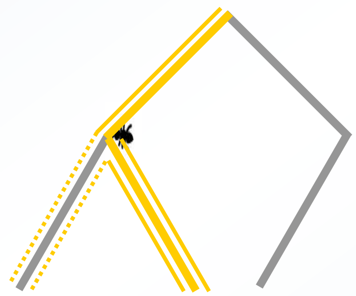
Tree 3/ Support Tree

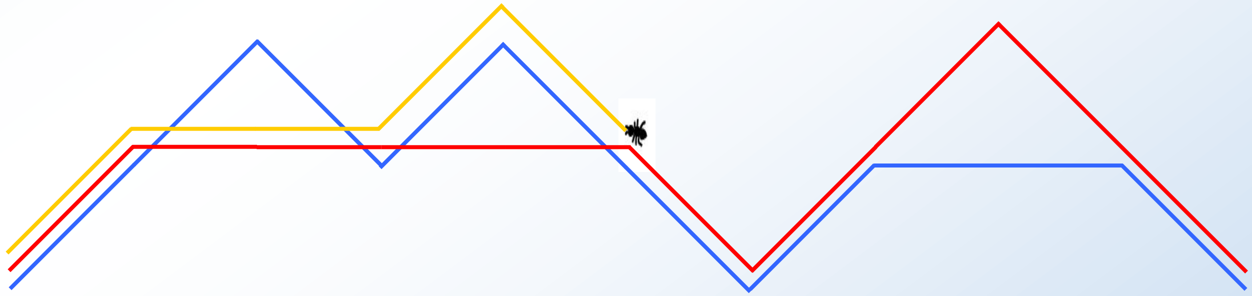


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

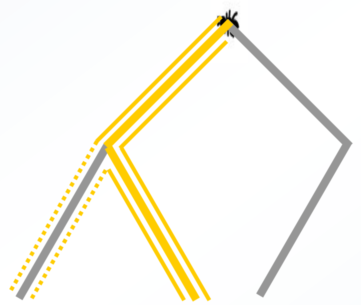


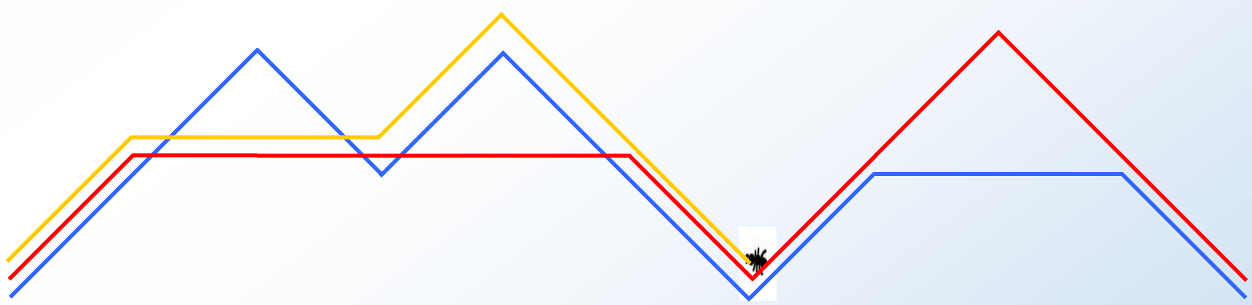


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

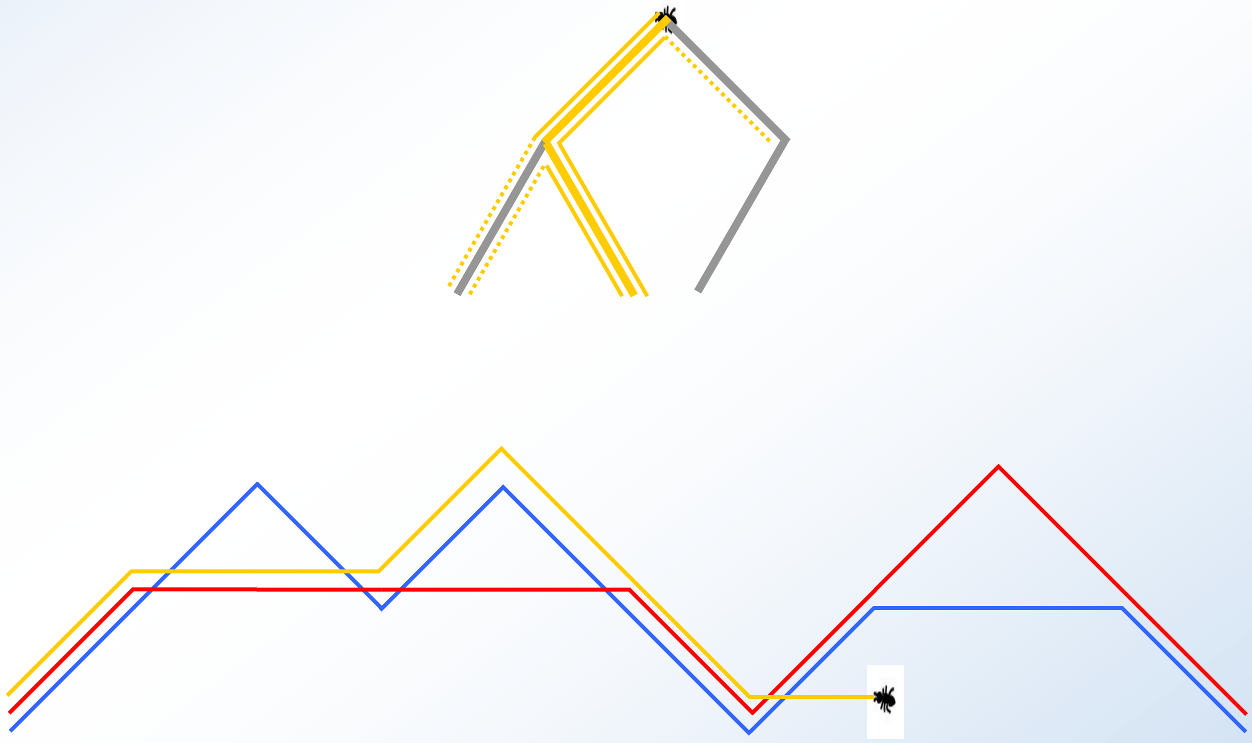




Dyck Path Representation

Now, we show how to transform the third tree as a curve.

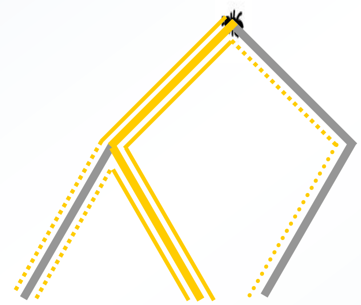
Tree 3/ Support Tree

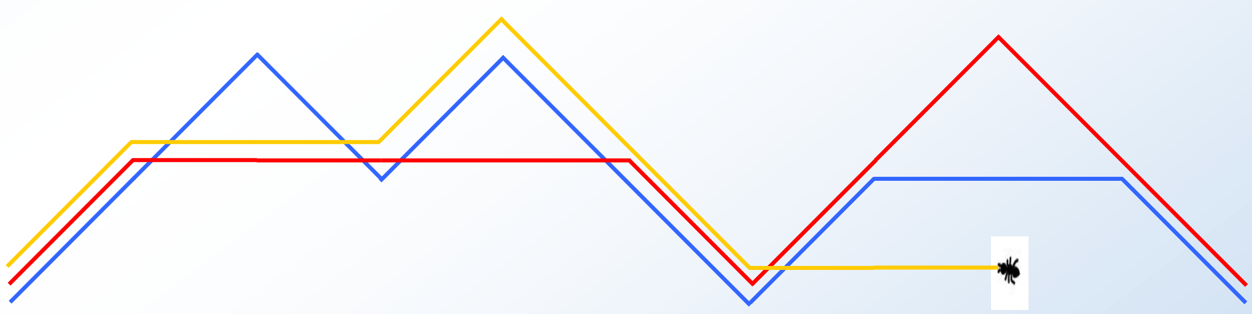


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

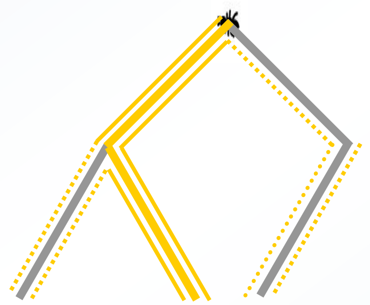


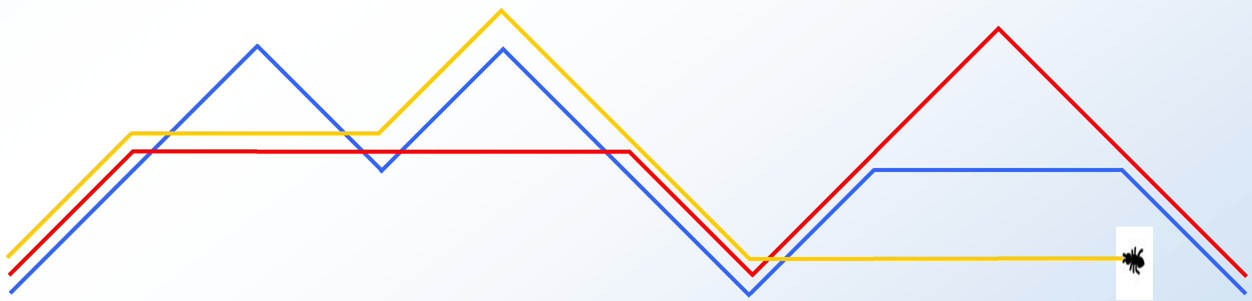


Dyck Path Representation

Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree

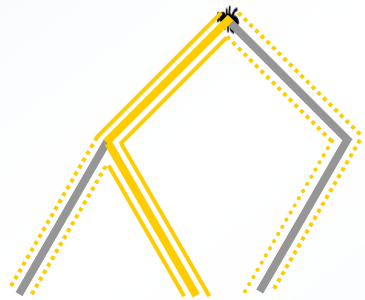


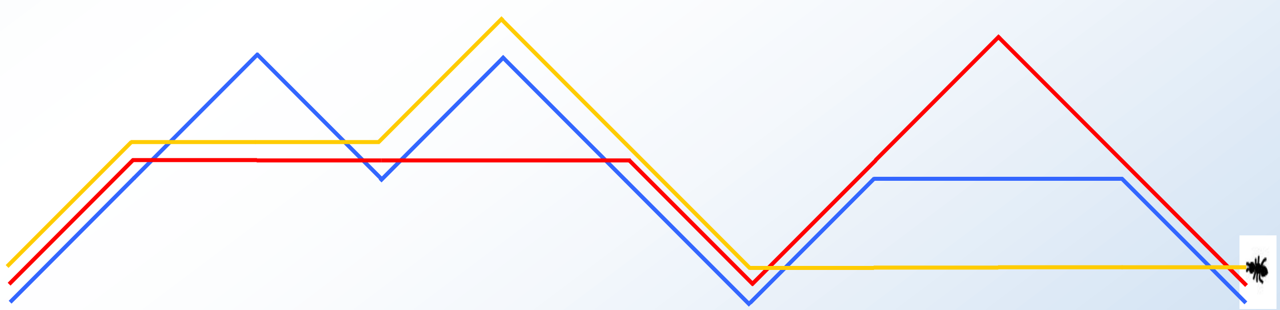


Dyck Path Representation

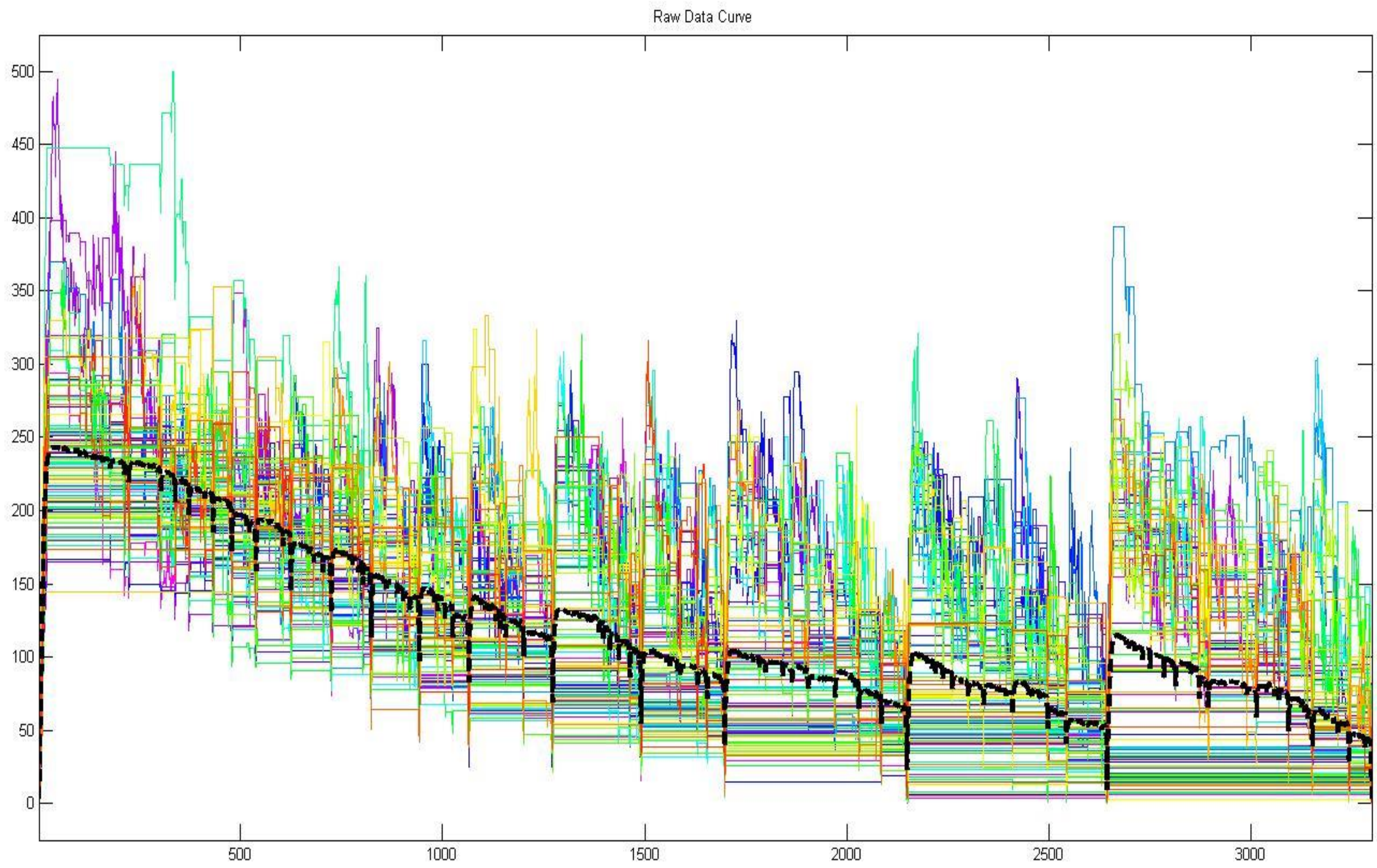
Now, we show how to transform the third tree as a curve.

Tree 3/ Support Tree





Dyck Path Curves (Back Tree)



Dyck Path Curves

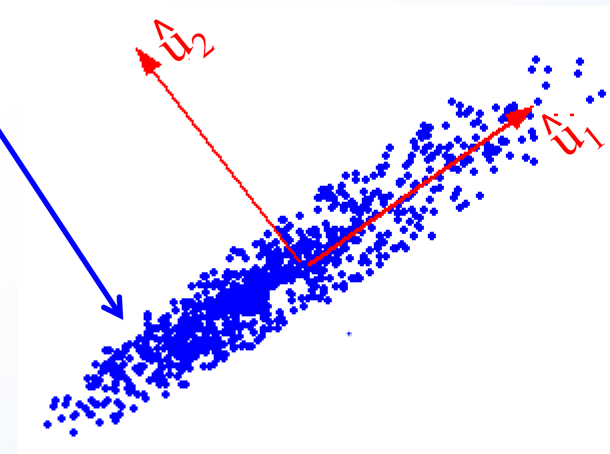
Properties:

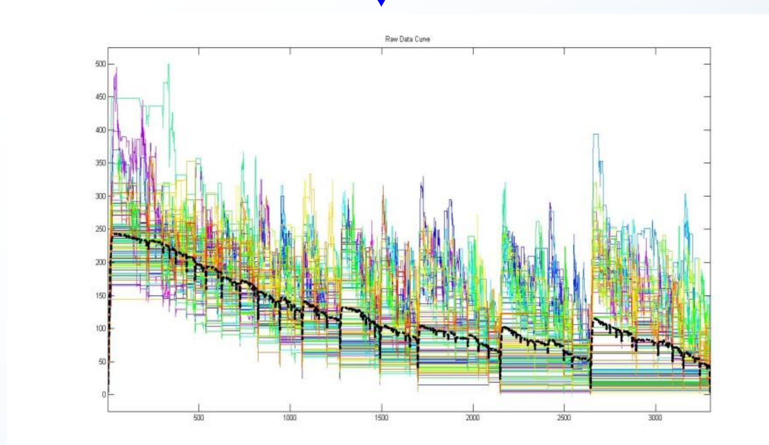
- Flat curve segments correspond to missing branches
- Rainbow color corresponds to age
  ranging from magenta (for young) to red (for old)
- The left part is taller than the right part
  the descendant correspondence
- The range of x-value is twice of the branch number
  every branch is passed twice - Dyck Path

Principal Component Analysis

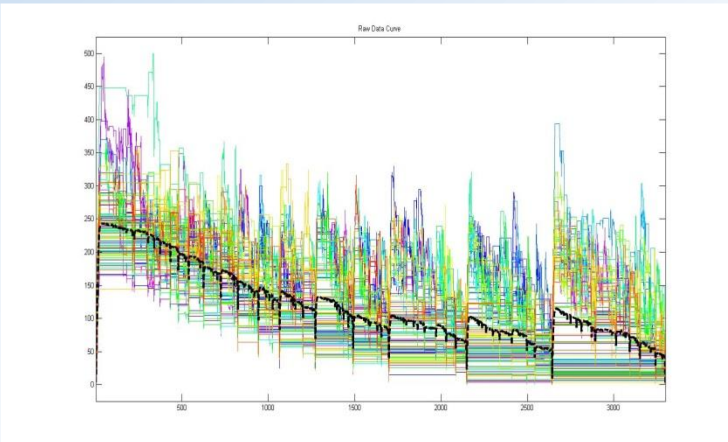
one data object

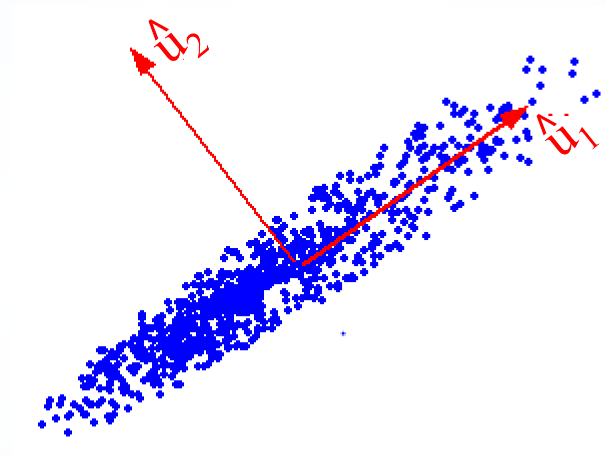
$$X_{p\times n} = \begin{pmatrix} \chi_{1,1} & \cdots & \chi_{1,i} & \cdots & \chi_{1,n} \\ \chi_{2,1} & \cdots & \chi_{2,i} & \cdots & \chi_{2,n} \\ & & \vdots & & \\ \chi_{p,1} & \cdots & \chi_{p,i} & \cdots & \chi_{p,n} \end{pmatrix}$$

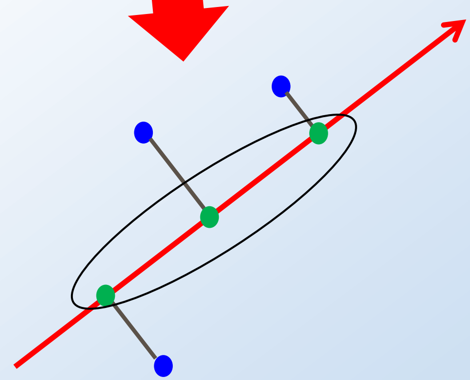




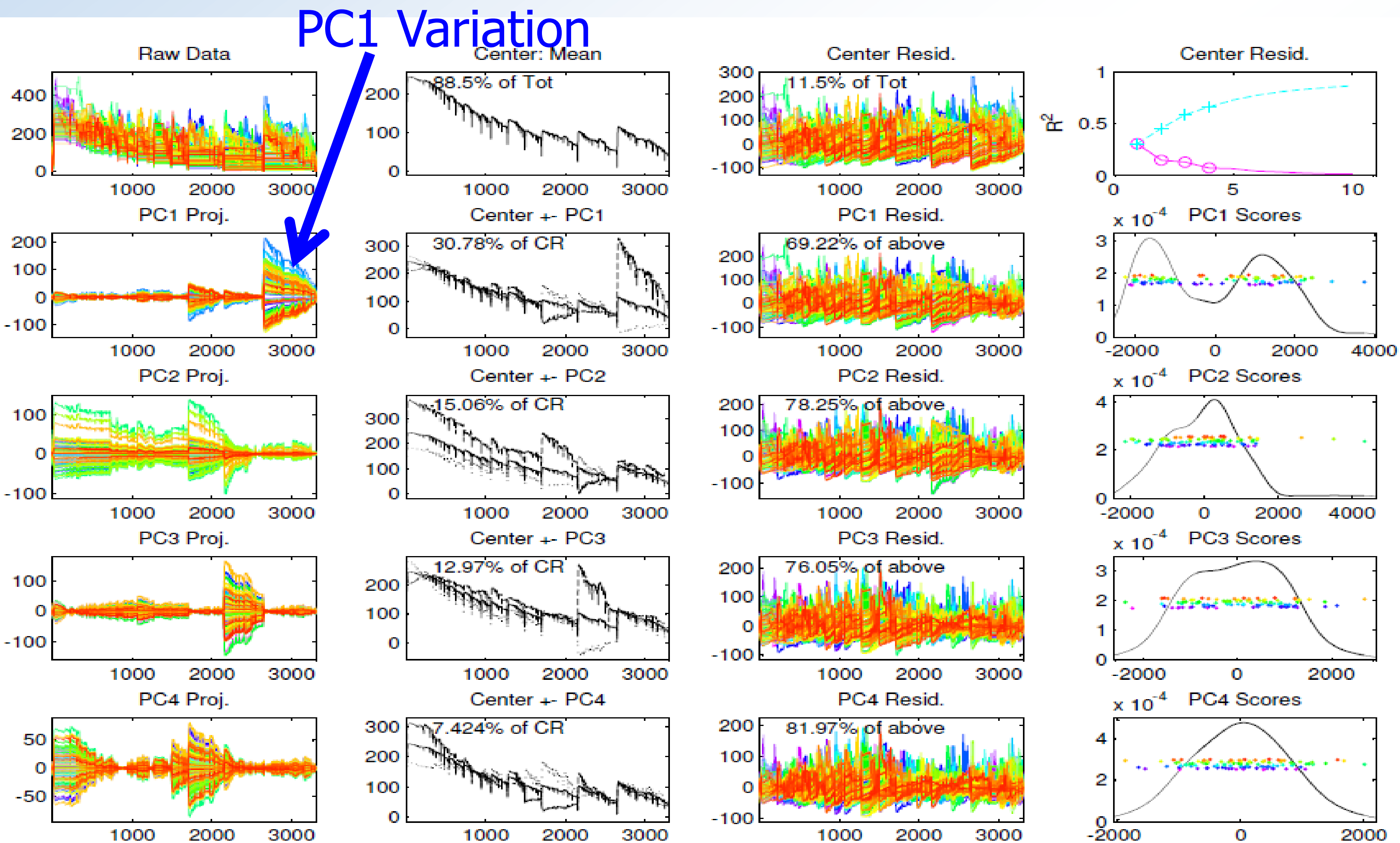
Principal Component Analysis



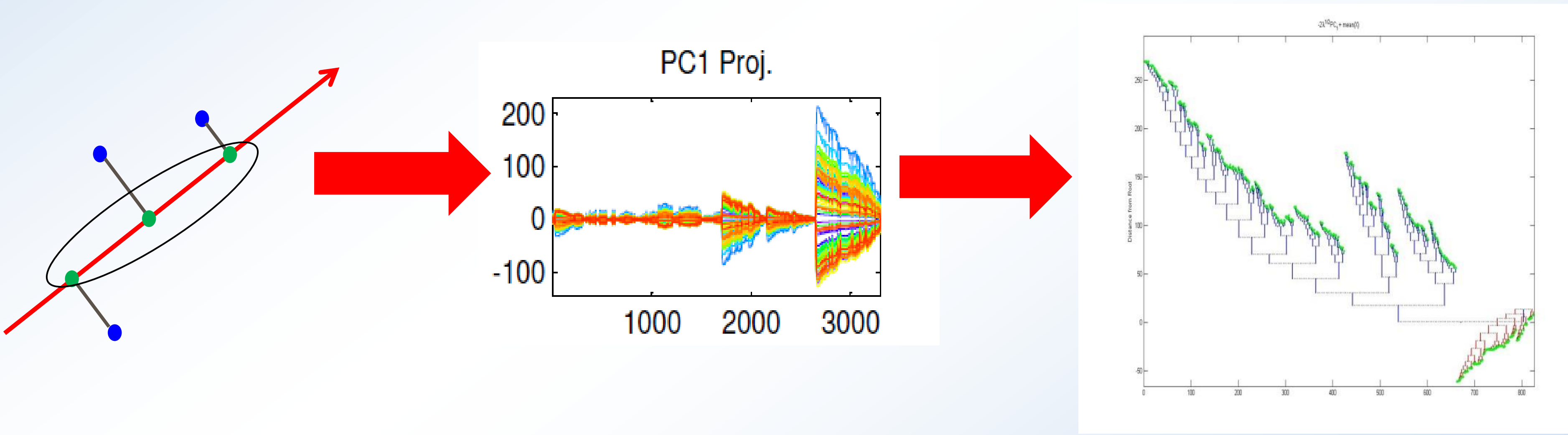




PCA of the Dyck Path Curves (Back Tree)

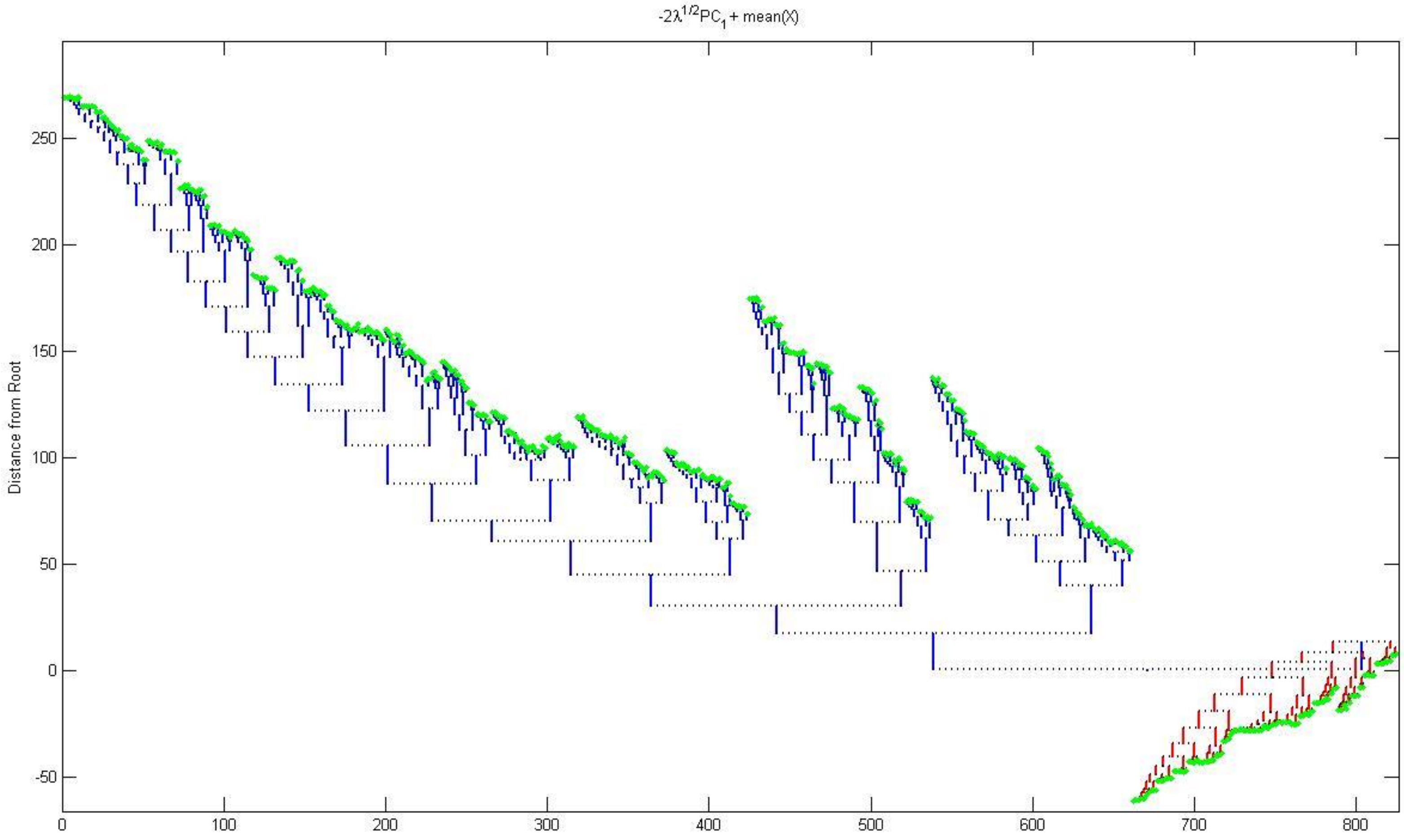


Tree interpretation of the PC direction

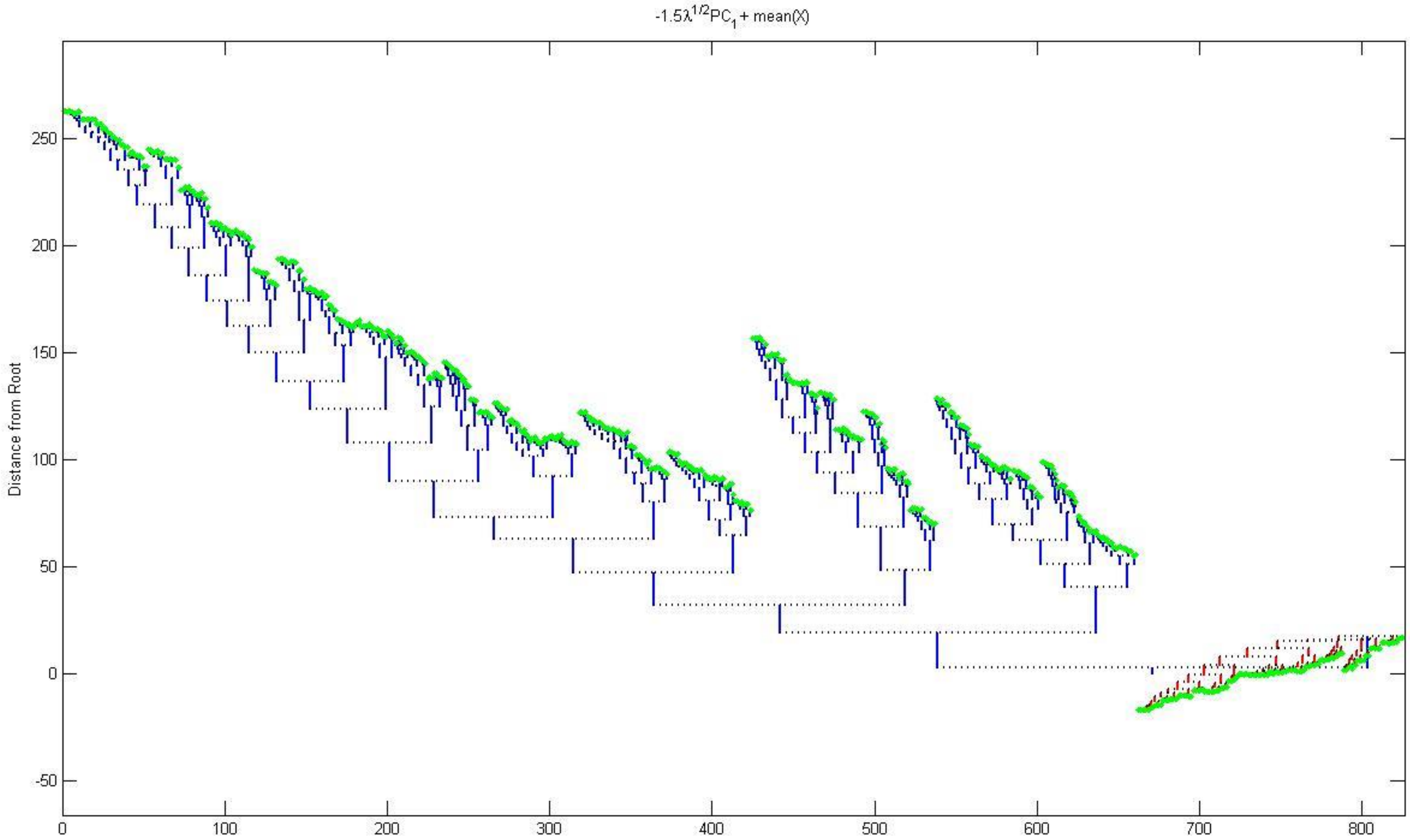


UNIVERSITY OF SOUTH FLORIDA

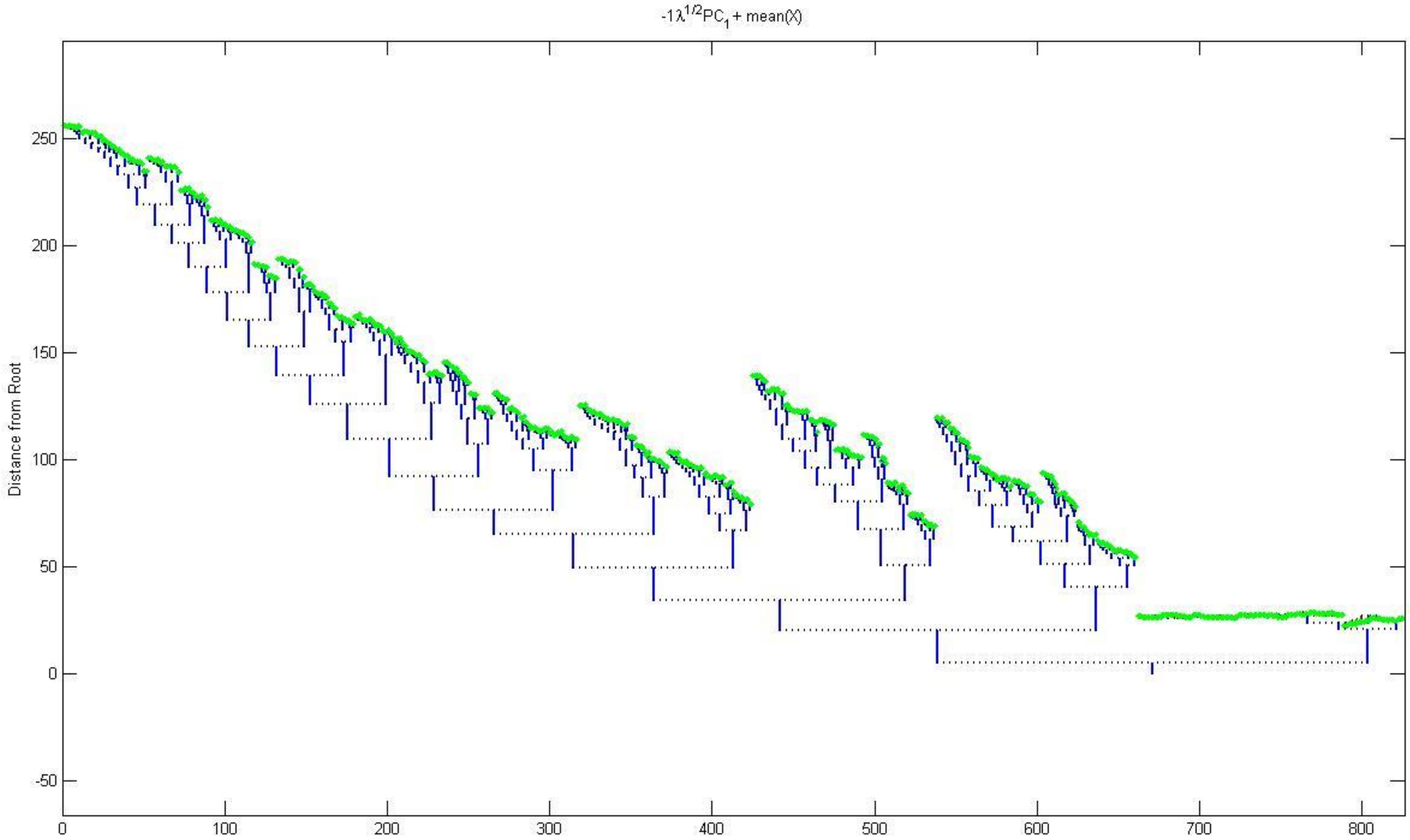
PC1 Direction (Back Tree)



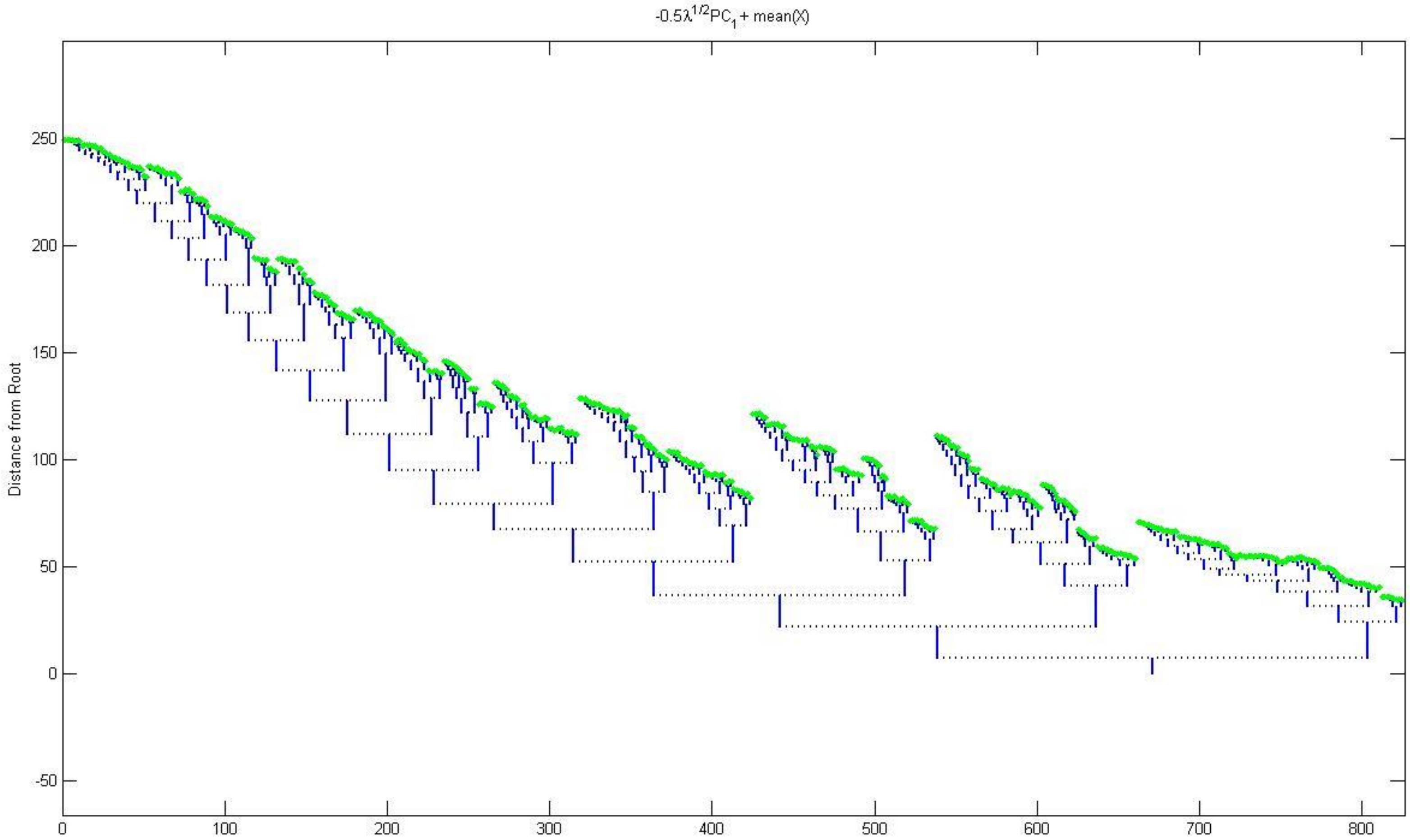
PC1 Direction (Back Tree)



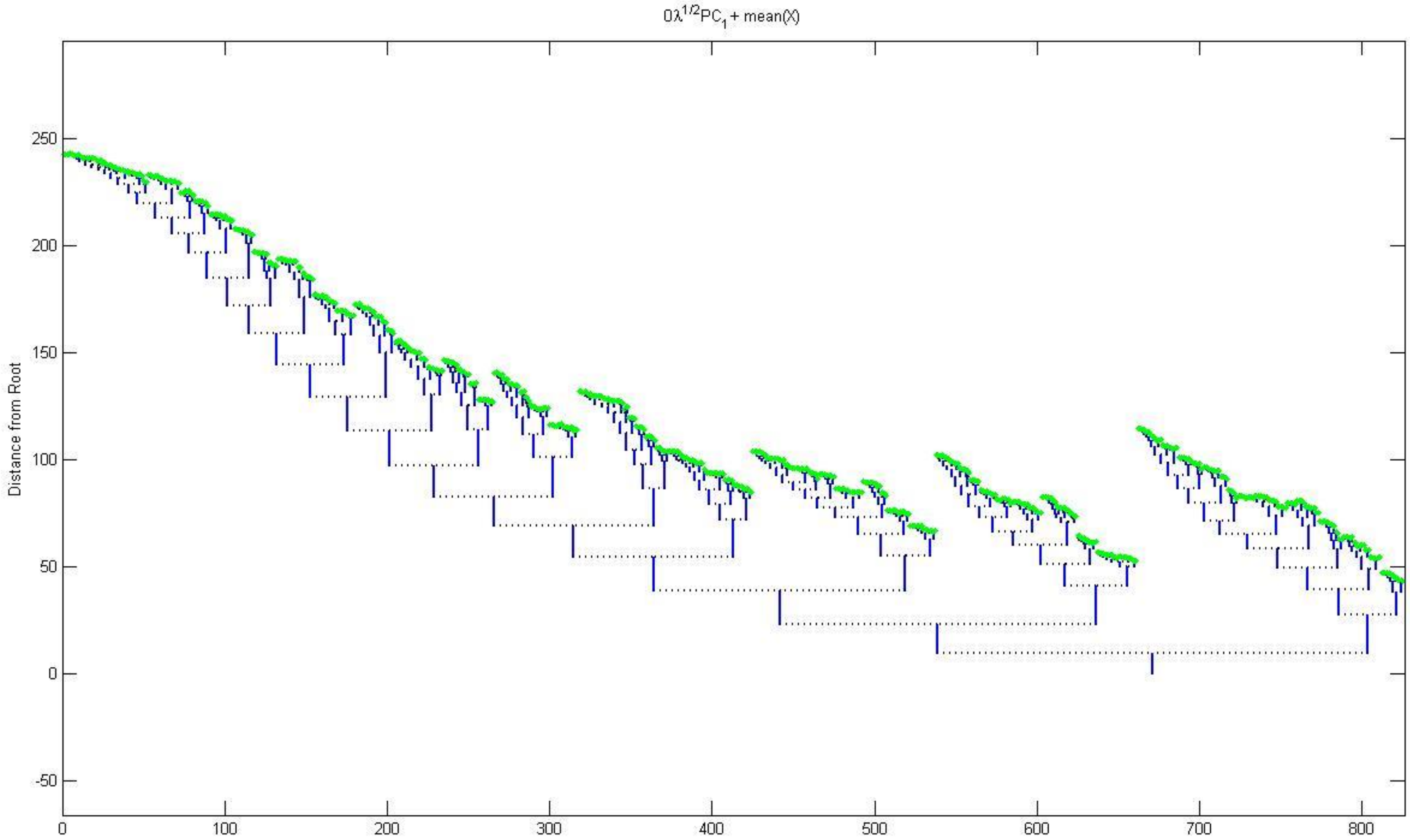
PC1 Direction (Back Tree)



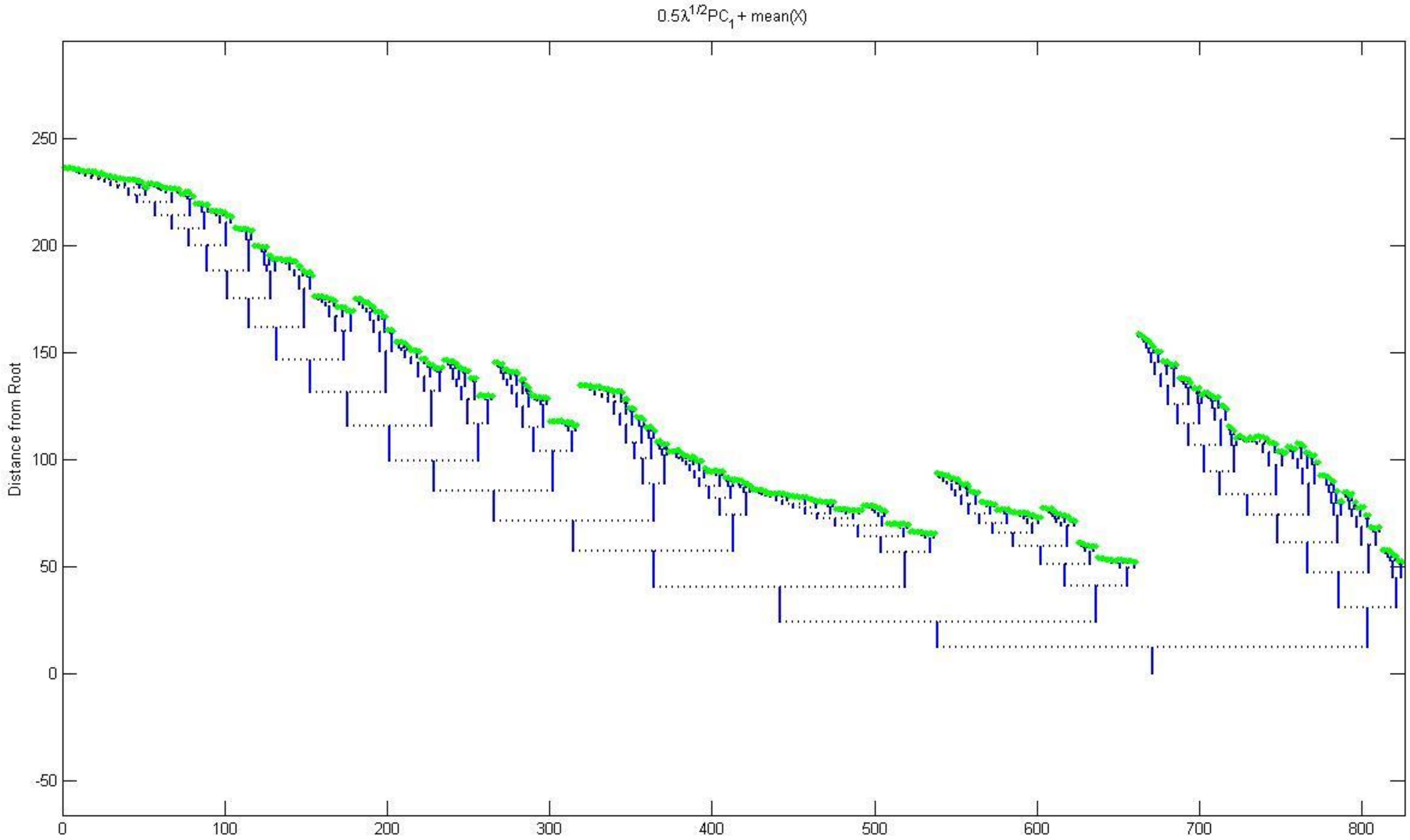
PC1 Direction (Back Tree)



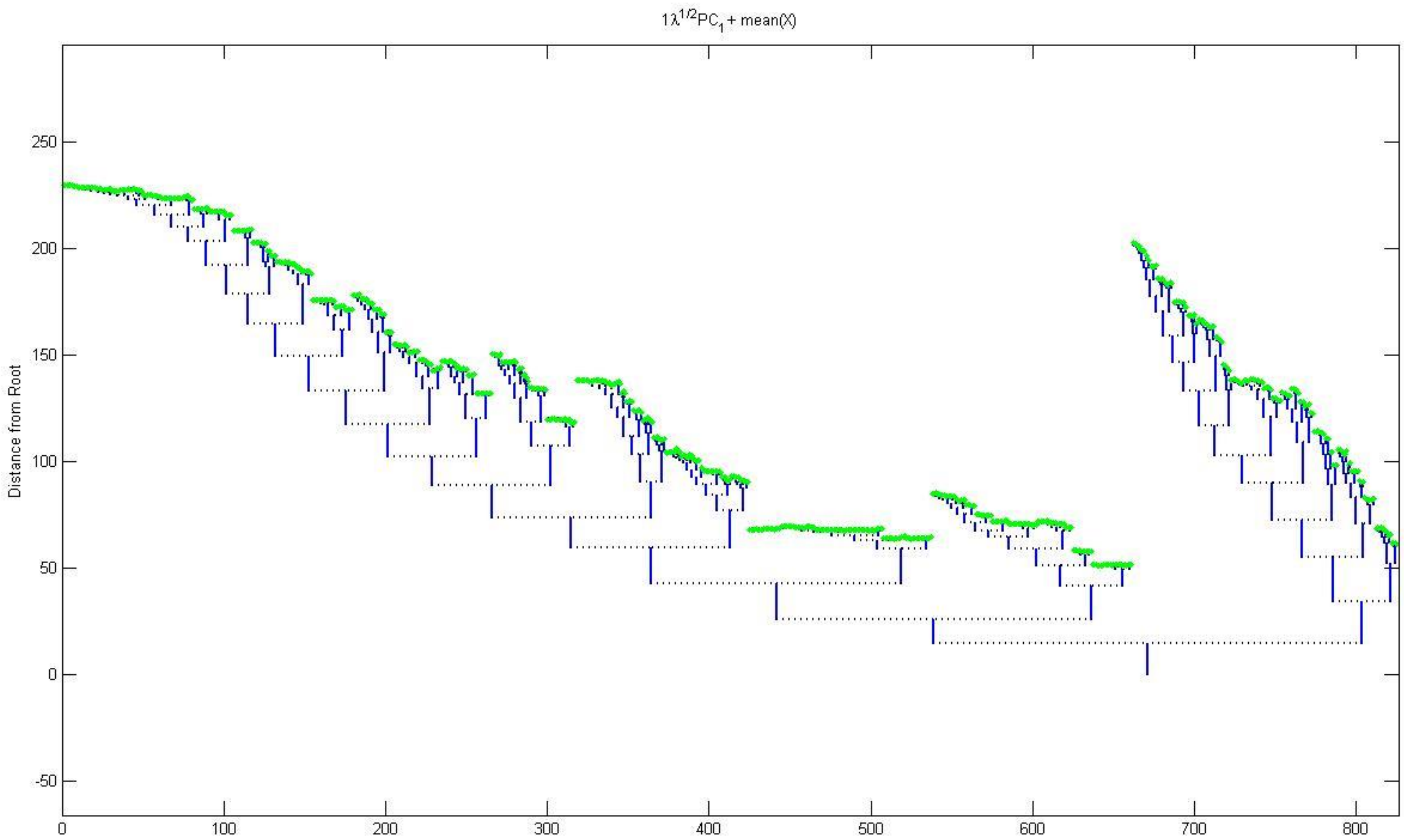
PC1 Direction (Back Tree)



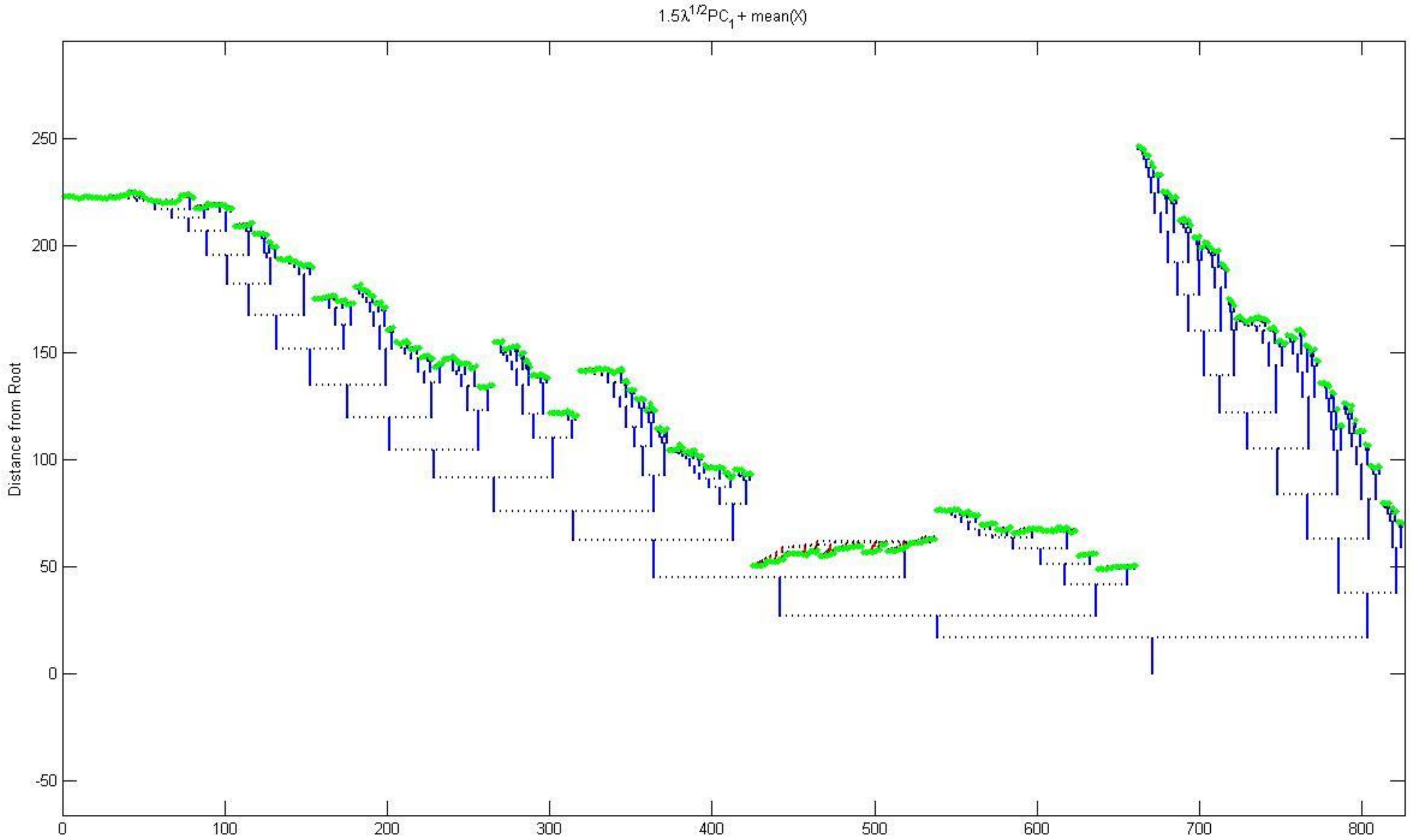
PC1 Direction (Back Tree)



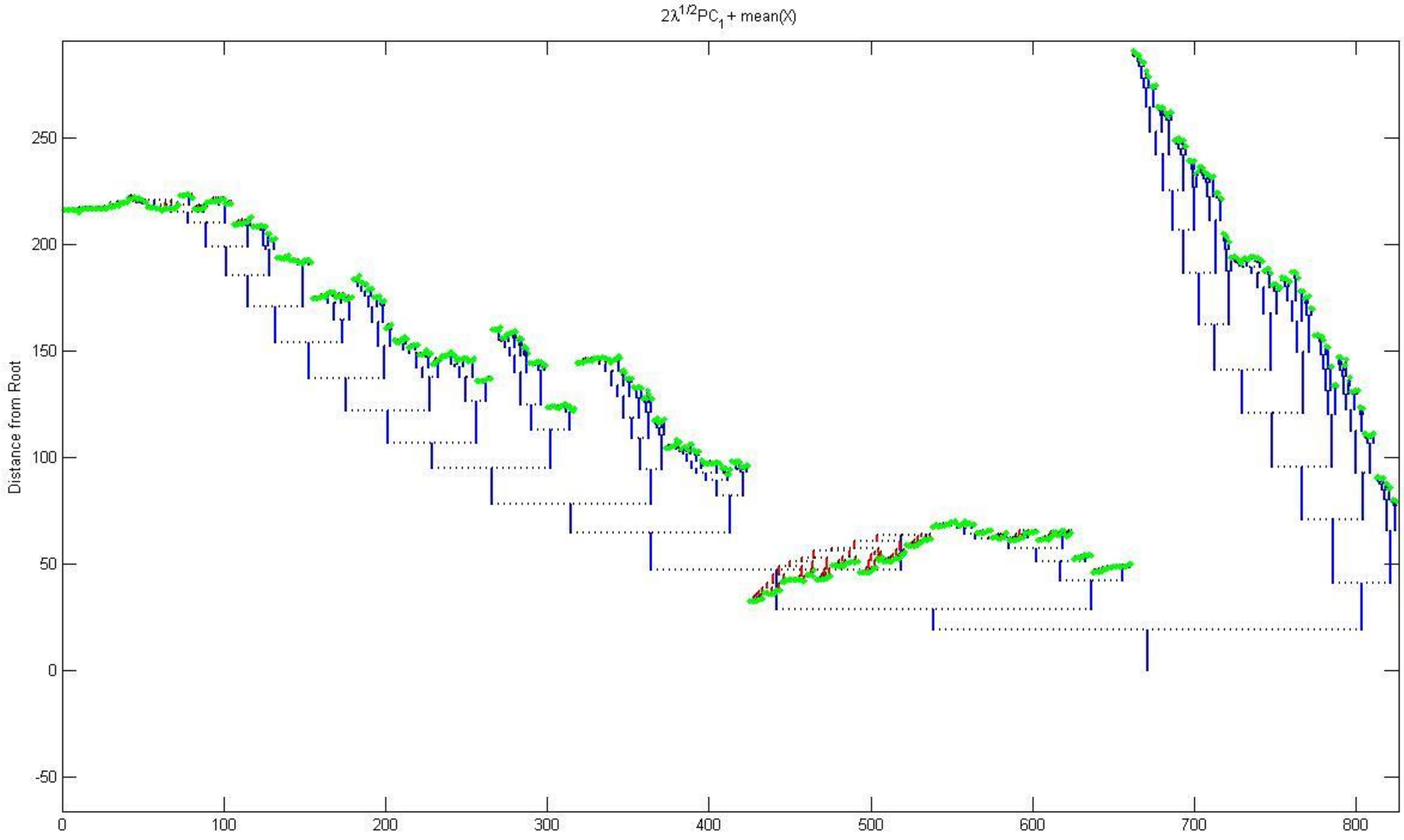
PC1 Direction (Back Tree)



PC1 Direction (Back Tree)



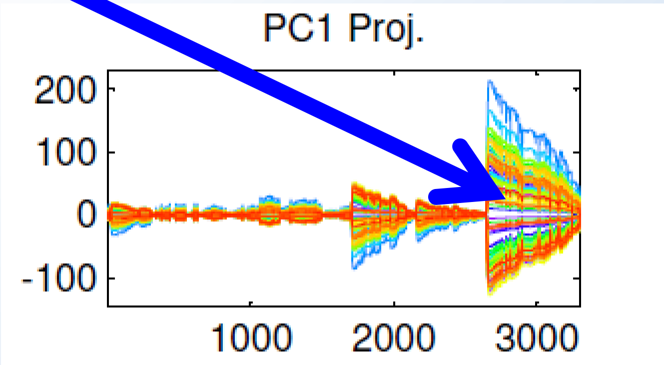
PC1 Direction (Back Tree)



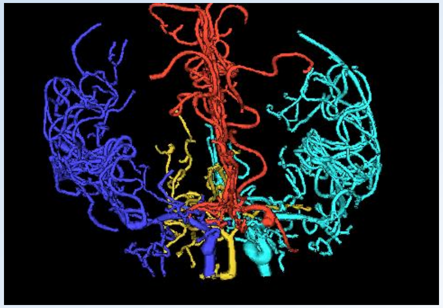
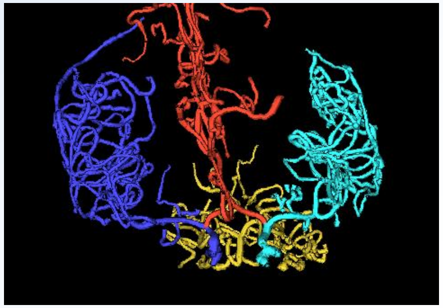
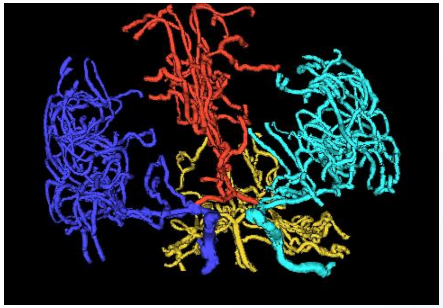
Tree Interpretation of the PC direction (Back Tree)

Summary :

- Main variation: banches in the right part of the binary trees
- Reflects the result from the PCA of the Dyck path curves



Blood vessel tree Analysis

 ,  , … , 

- n=98

- Statistical goals:

  1. Structure of Population (understand variation)

  2. Gender difference (Classification)

  3. Age difference

  4. Build model

Blood vessel tree Analysis

Dan Shen, et al (2014). **Functional data analysis of tree data objects**, (Featured Article) Journal of Computational and Graphical Statistics, 23, 418-238.

Thank You !