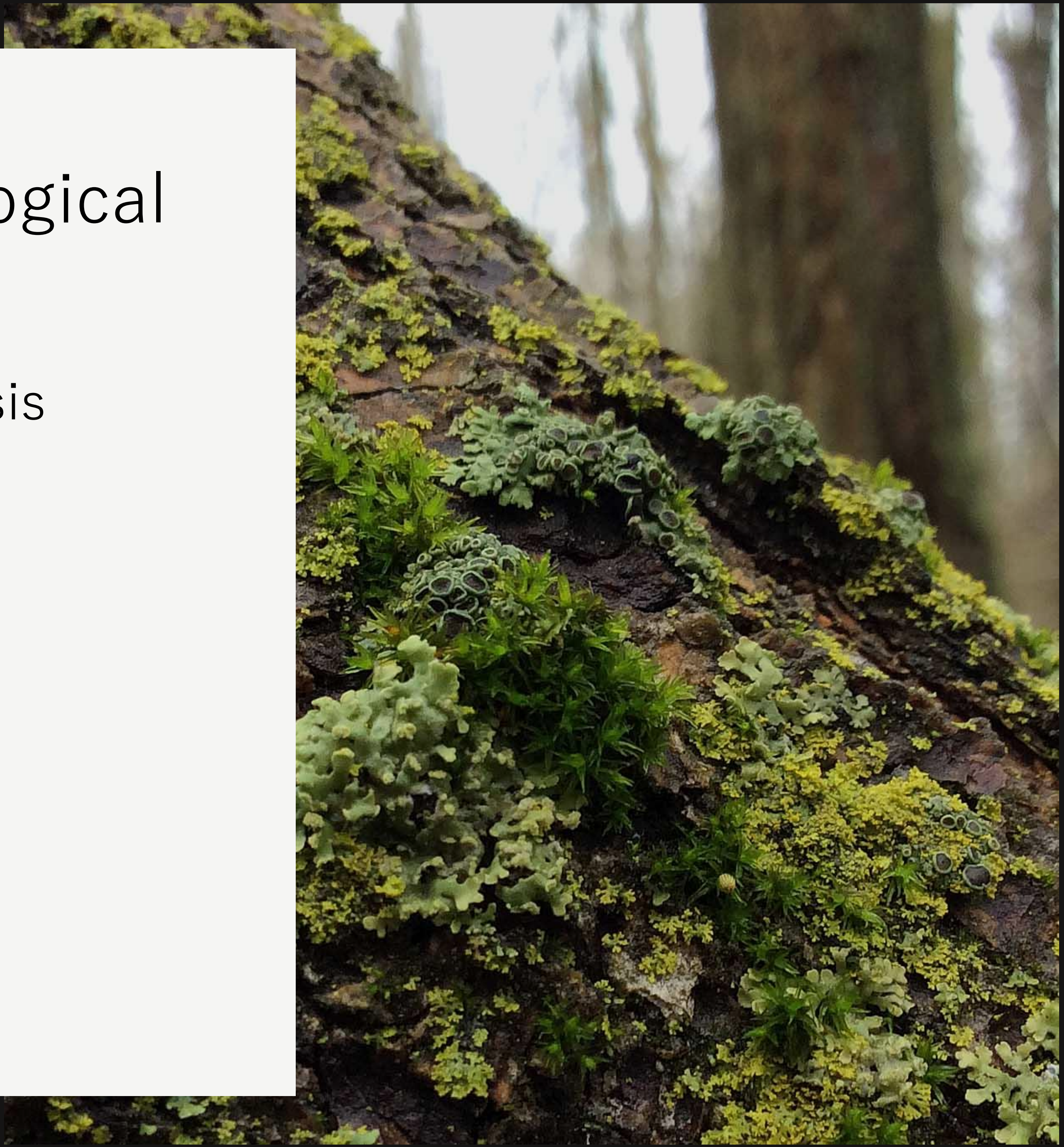
FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

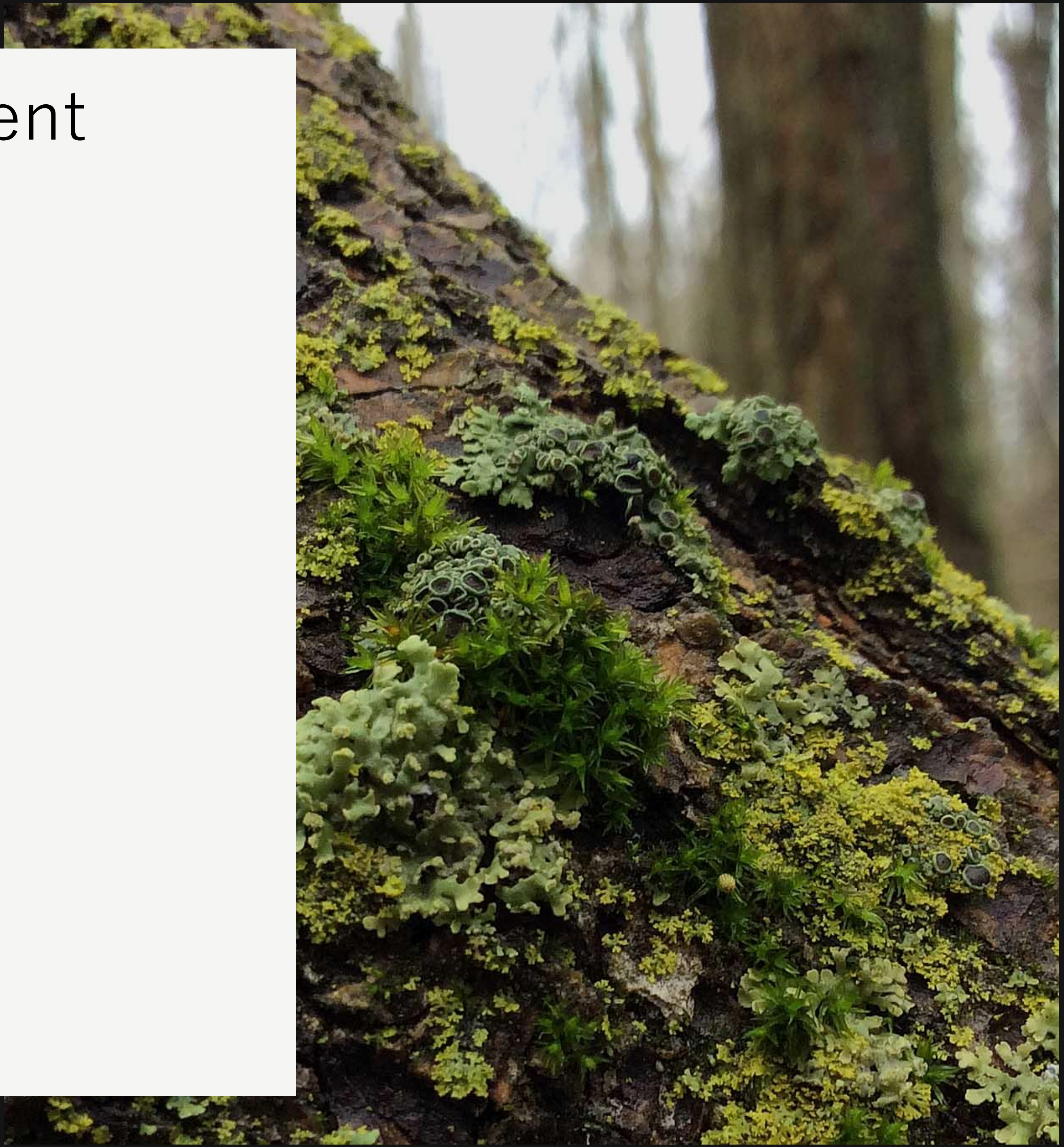
Lecture 6: Principal Component Analysis

Thursday, October 17, 2024

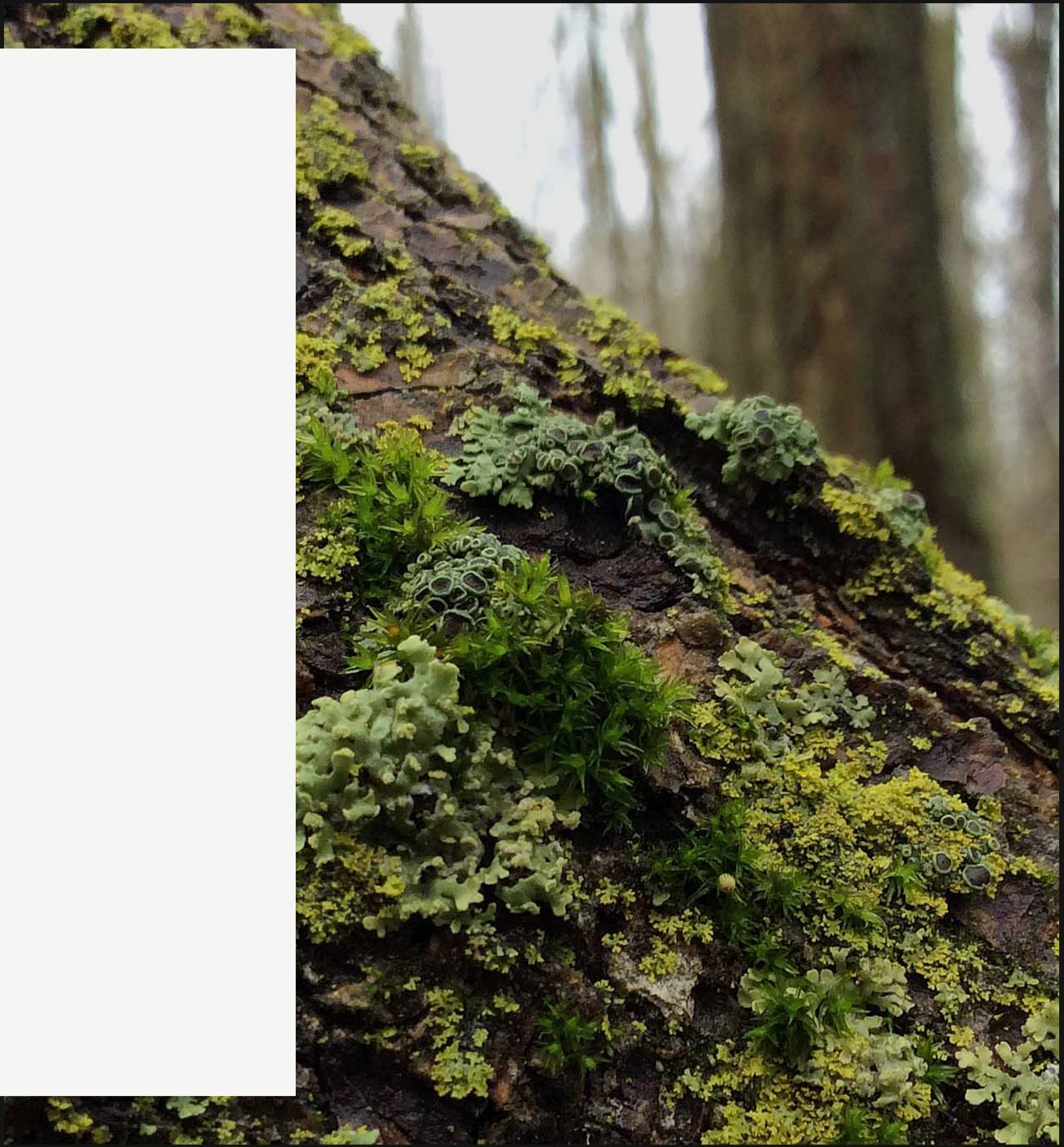


Lecture 6: Principal Component Analysis

- Dispersion Matrices
- PCA Steps
- Assessing Meaningful Components
- Limitations



Recap: Eigenvectors and Eigenvalues



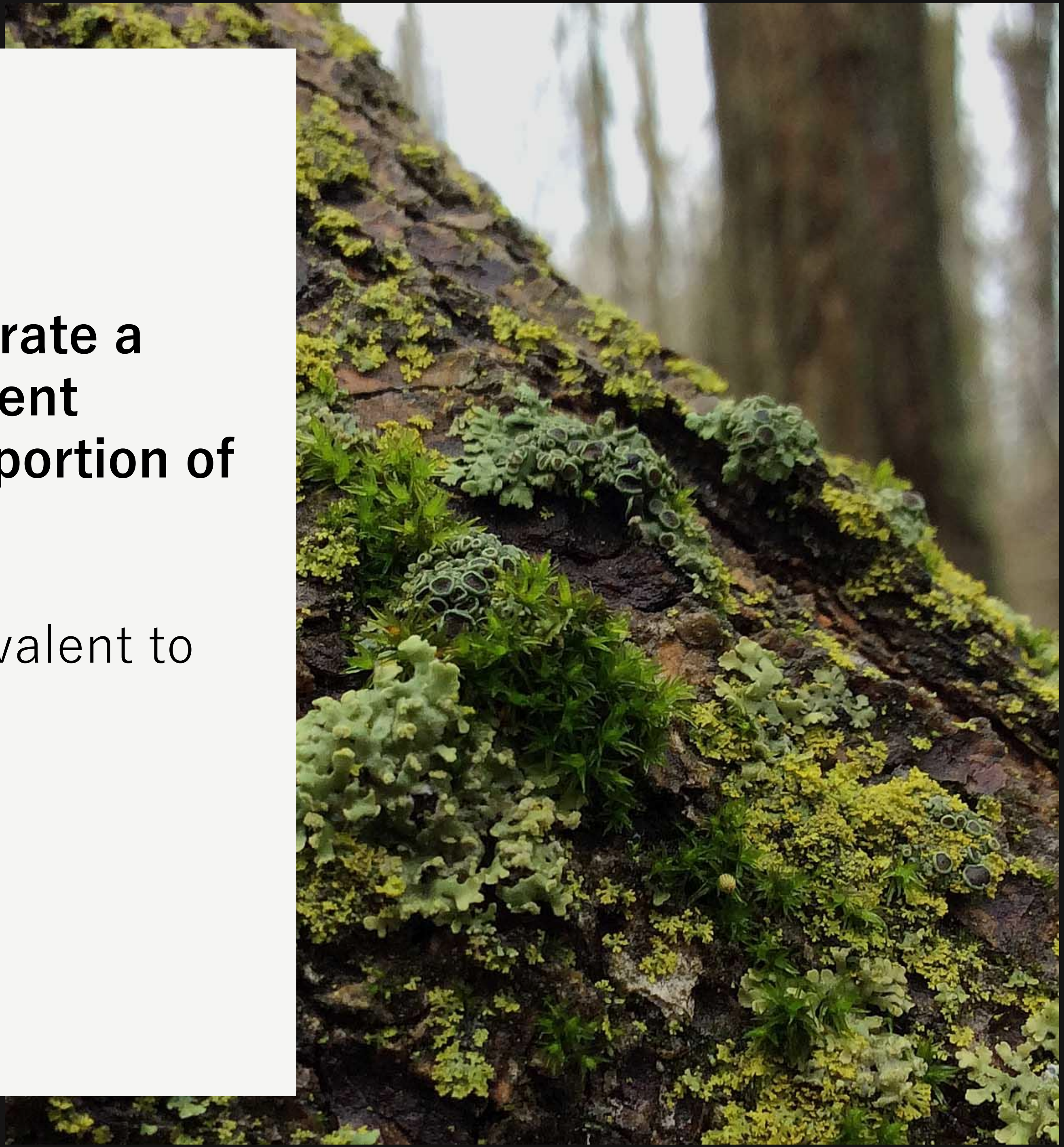
Recap: Eigenvectors and Eigenvalues

- The goal of eigenanalysis is to **generate a small number of linearly independent variables, each explaining a large portion of the variation**.

Recap: Eigenvectors and Eigenvalues

- The goal of eigenanalysis is to **generate a small number of linearly independent variables, each explaining a large portion of the variation**.

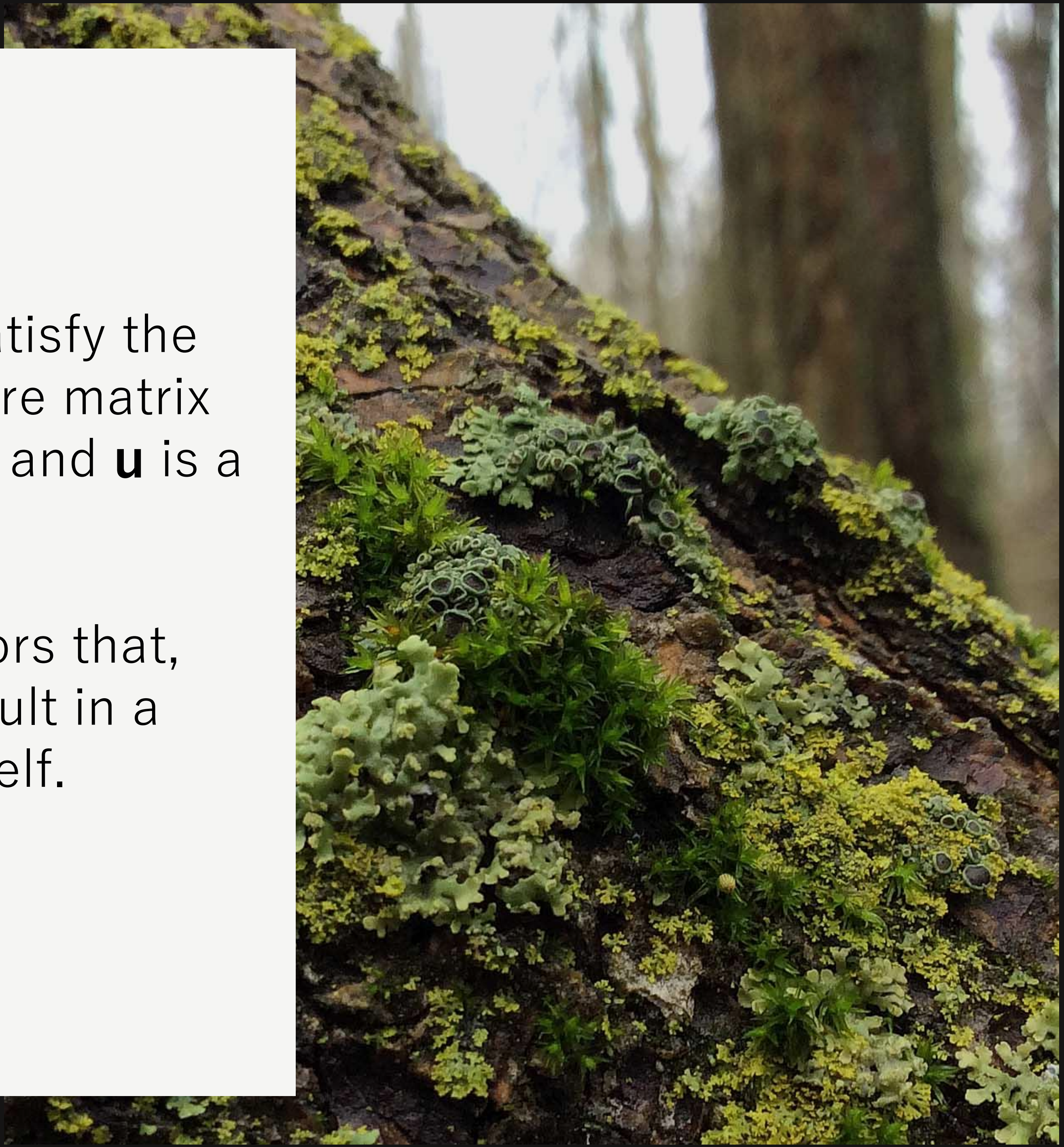
- i.e., generate a diagonal matrix equivalent to the square matrix **A**



Recap: Eigenvectors and Eigenvalues

- **Eigenvalues** ($\boldsymbol{\lambda}$) are scalars that satisfy the equation $\mathbf{Av}=\boldsymbol{\lambda}\mathbf{u}$, where **A** is a square matrix (for example, an association matrix) and **u** is a non-zero vector.

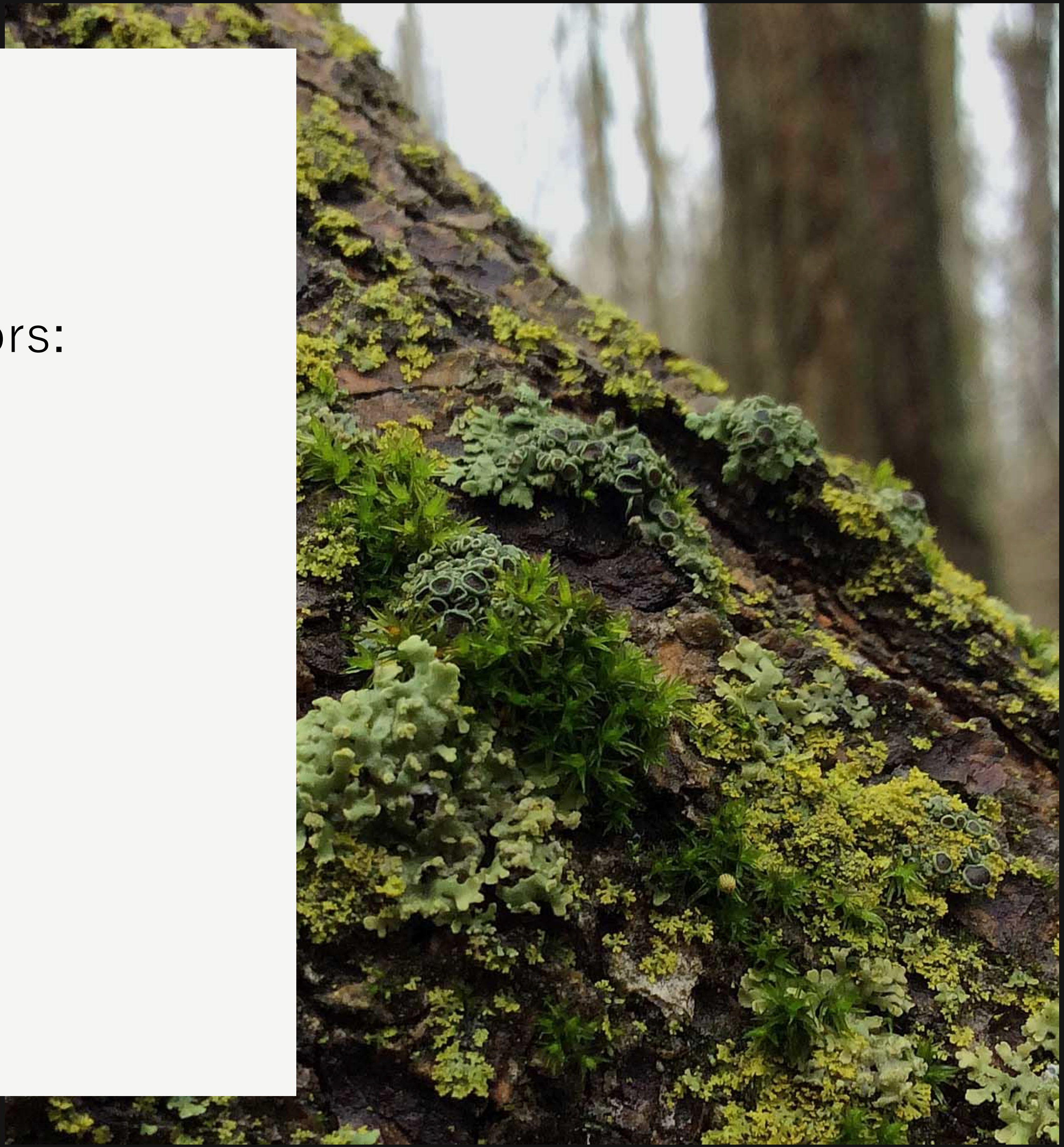
- **Eigenvectors** (**u**) are non-zero vectors that, when multiplied by the matrix **A**, result in a vector that is a scalar multiple of itself.



Recap: Eigenvectors and Eigenvalues

Solving for eigenvalues and eigenvectors:

$$\mathbf{A} = \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix}$$

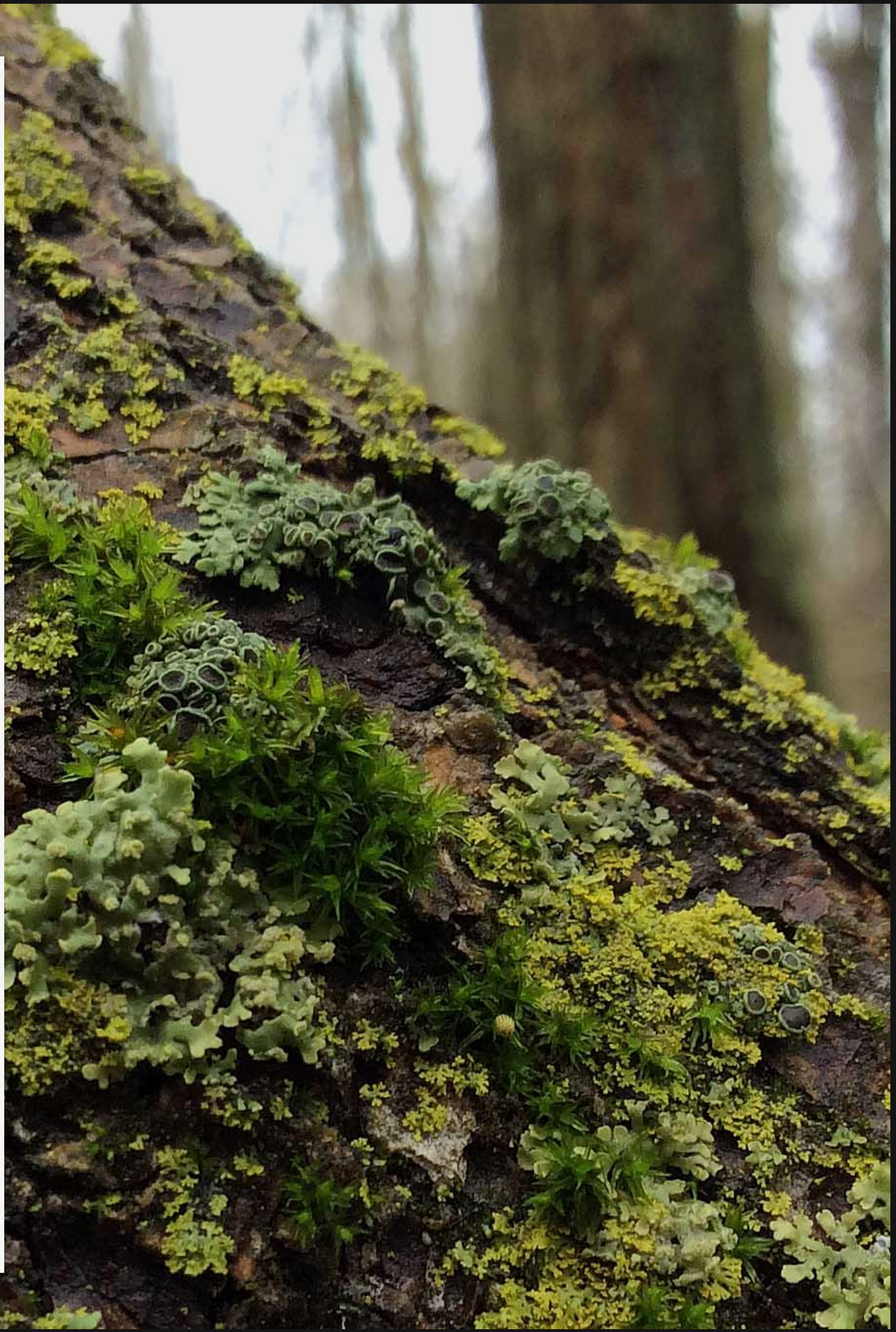


Recap: Eigenvectors and Eigenvalues

Solving for eigenvalues and eigenvectors:

1) Form the characteristic equation $|\mathbf{A} - \lambda\,\mathbf{I}| = 0$

$$|\mathbf{A} - \lambda\,\mathbf{I}| = \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = 0$$



Recap: Eigenvectors and Eigenvalues

Solving for eigenvalues and eigenvectors:

2) Solve for eigenvalues ($\boldsymbol{\lambda}$)

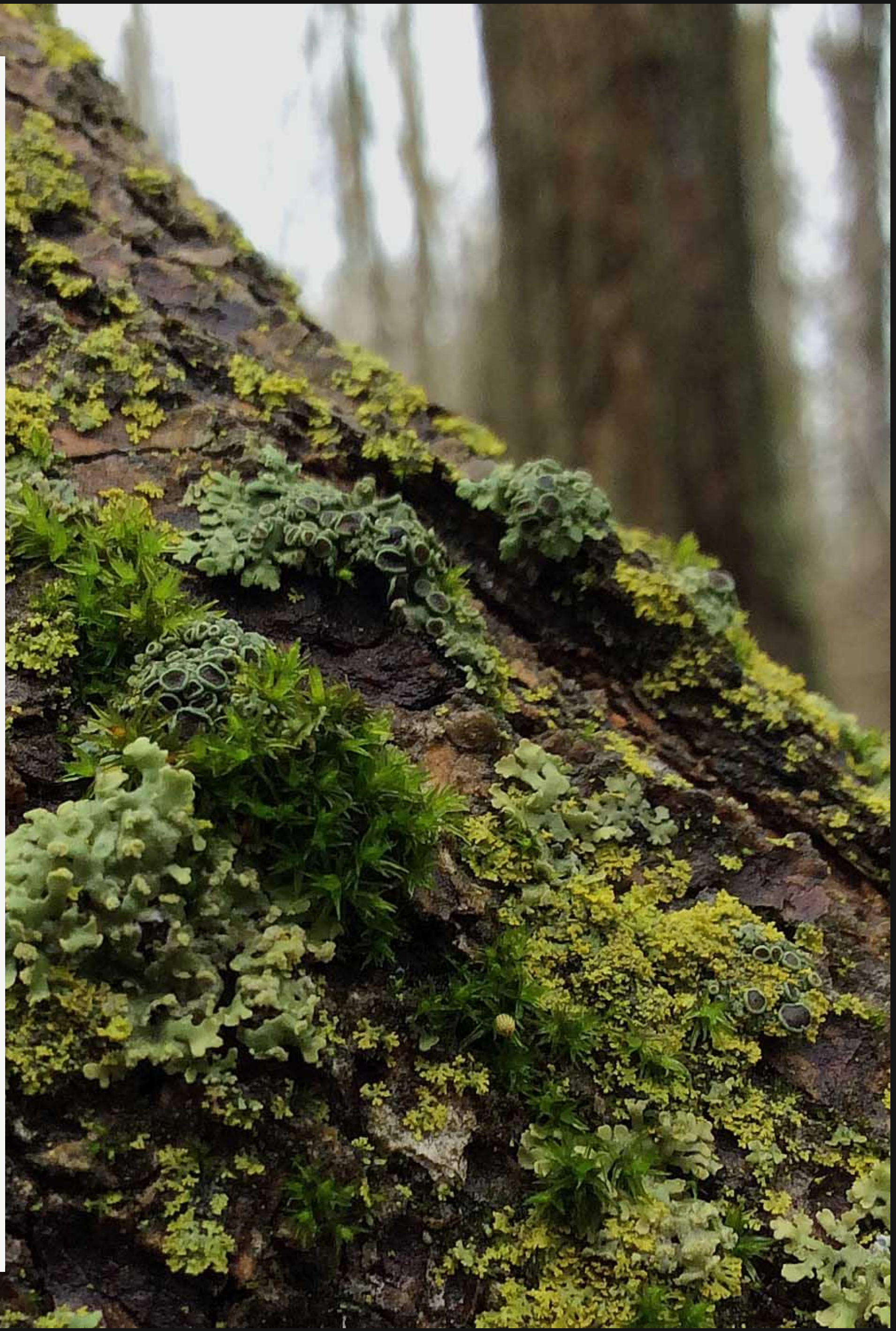
$$|\mathbf{A} - \lambda\,\mathbf{I}| = \begin{bmatrix} 4-\lambda & 1 \\ 2 & 3-\lambda \end{bmatrix} = 0$$

$$(4-\boldsymbol{\lambda}) \times (3-\boldsymbol{\lambda}) - 2 \times 1 = 0$$

$$\boldsymbol{\lambda}^2 - 7\boldsymbol{\lambda} + 10 = 0$$

$$(\boldsymbol{\lambda} - 2) \times (\boldsymbol{\lambda} - 5) = 0$$

$$\lambda_1 = 5, \lambda_2 = 2$$



Recap: Eigenvectors and Eigenvalues

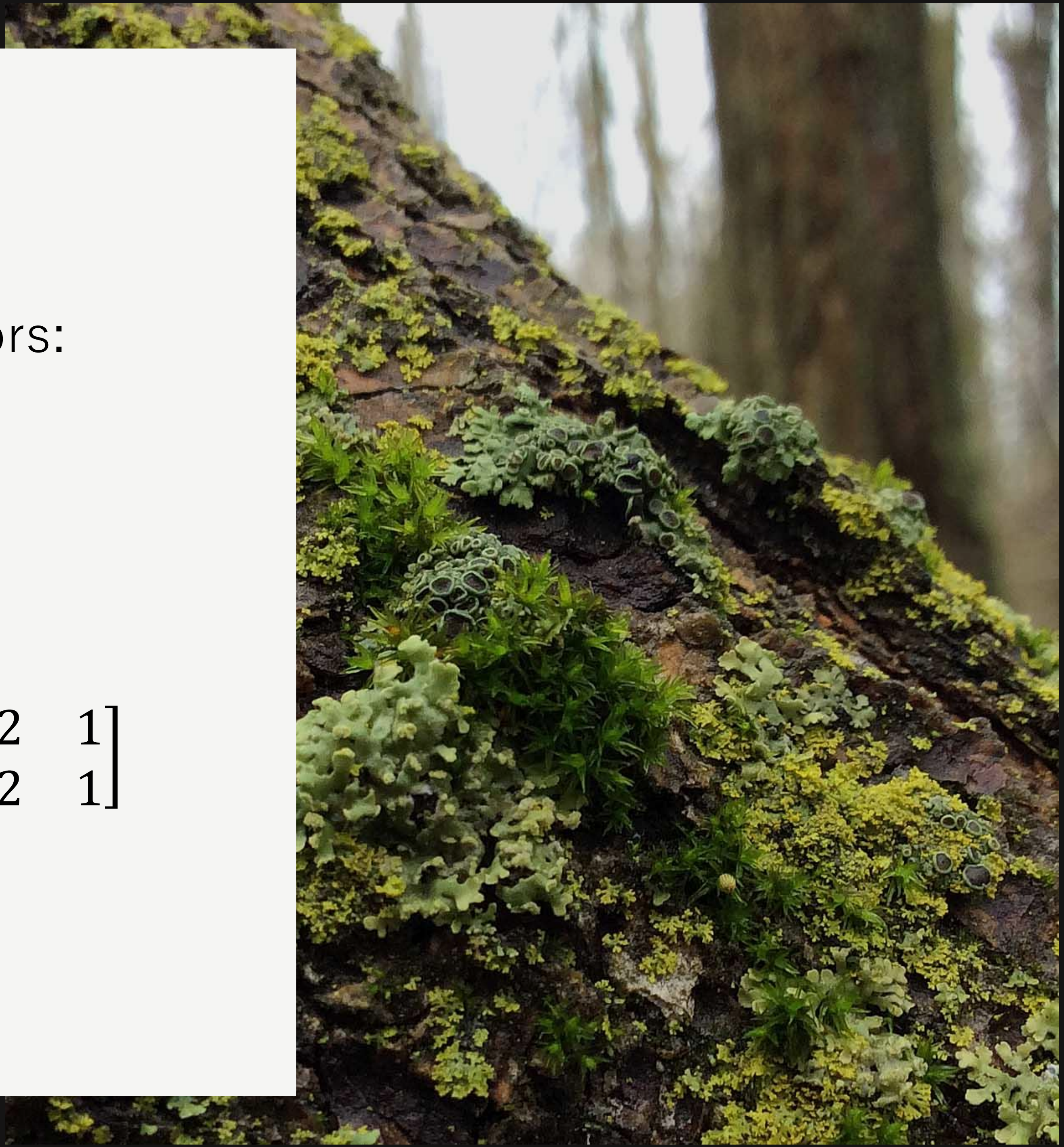
Solving for eigenvalues and eigenvectors:

3) Solve for eigenvectors ($\mathbf{u}$)

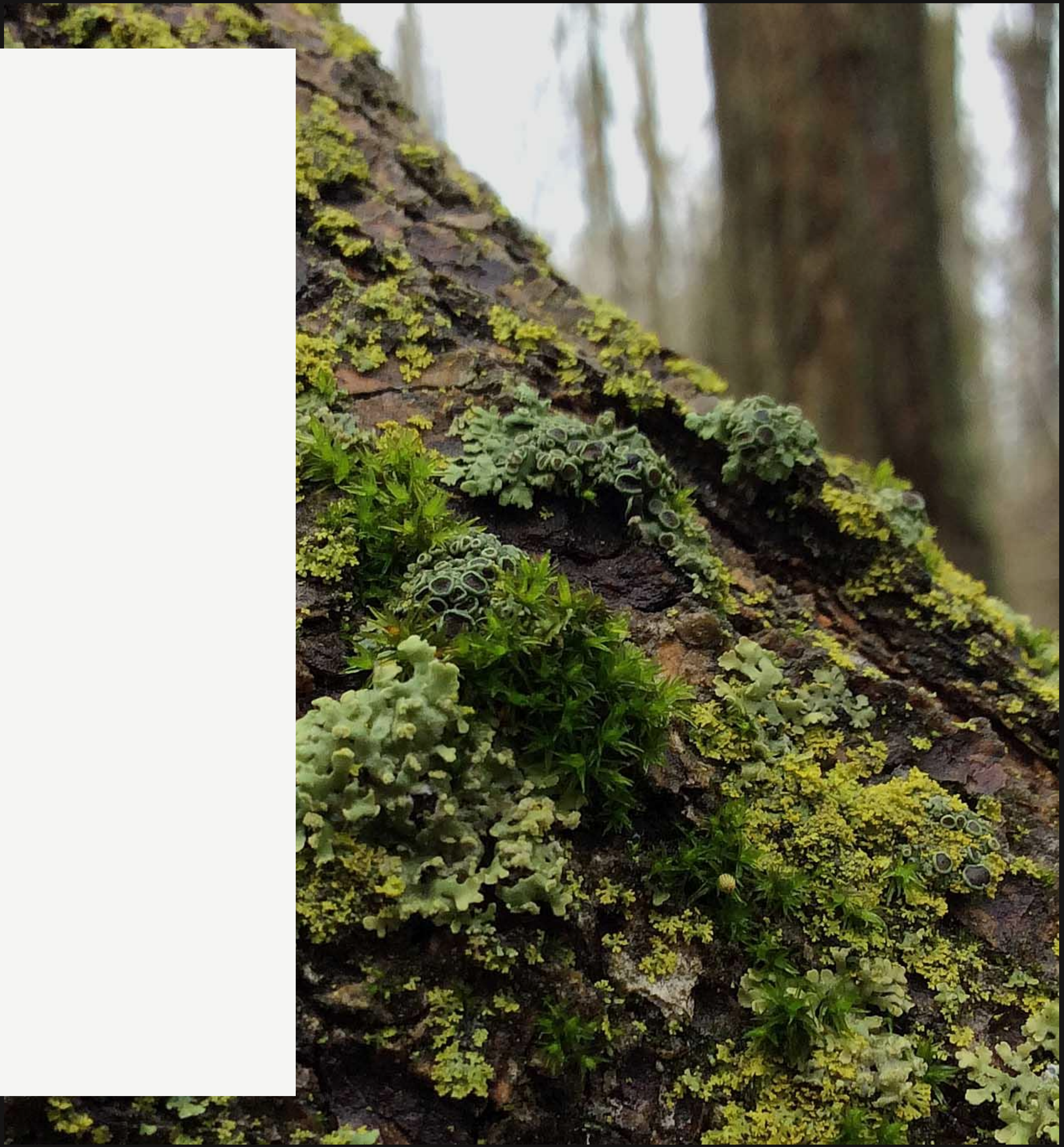
$$(\mathbf{A} - \boldsymbol{\lambda}\mathbf{I})\mathbf{u} = 0$$

$$(\mathbf{A} - \boldsymbol{\lambda}_1\mathbf{I}) = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \quad (\mathbf{A} - \boldsymbol{\lambda}_2\mathbf{I}) = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$$

$$\mathbf{u}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$



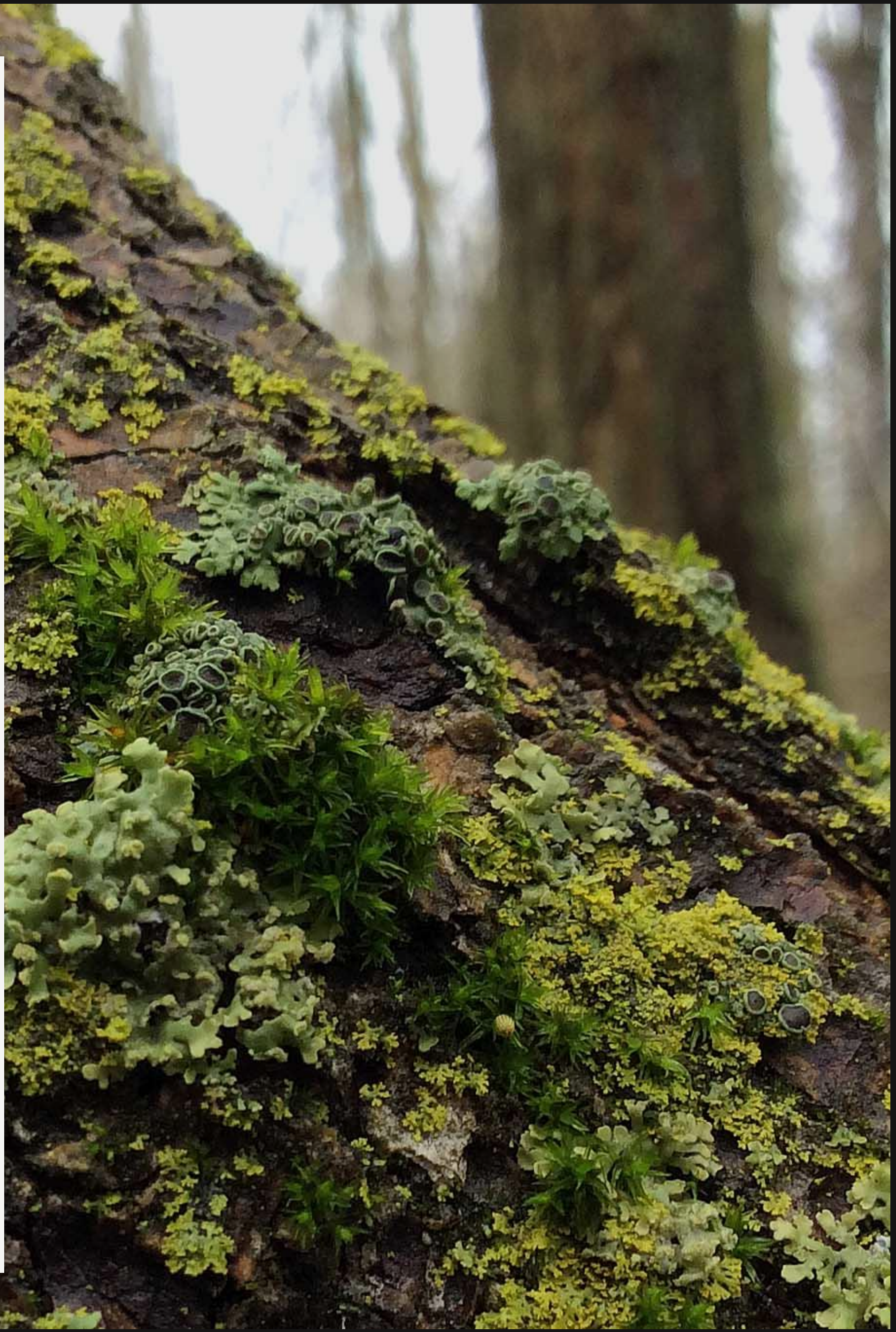
Intro: Ordination



Intro: Ordination

Intro: Ordination

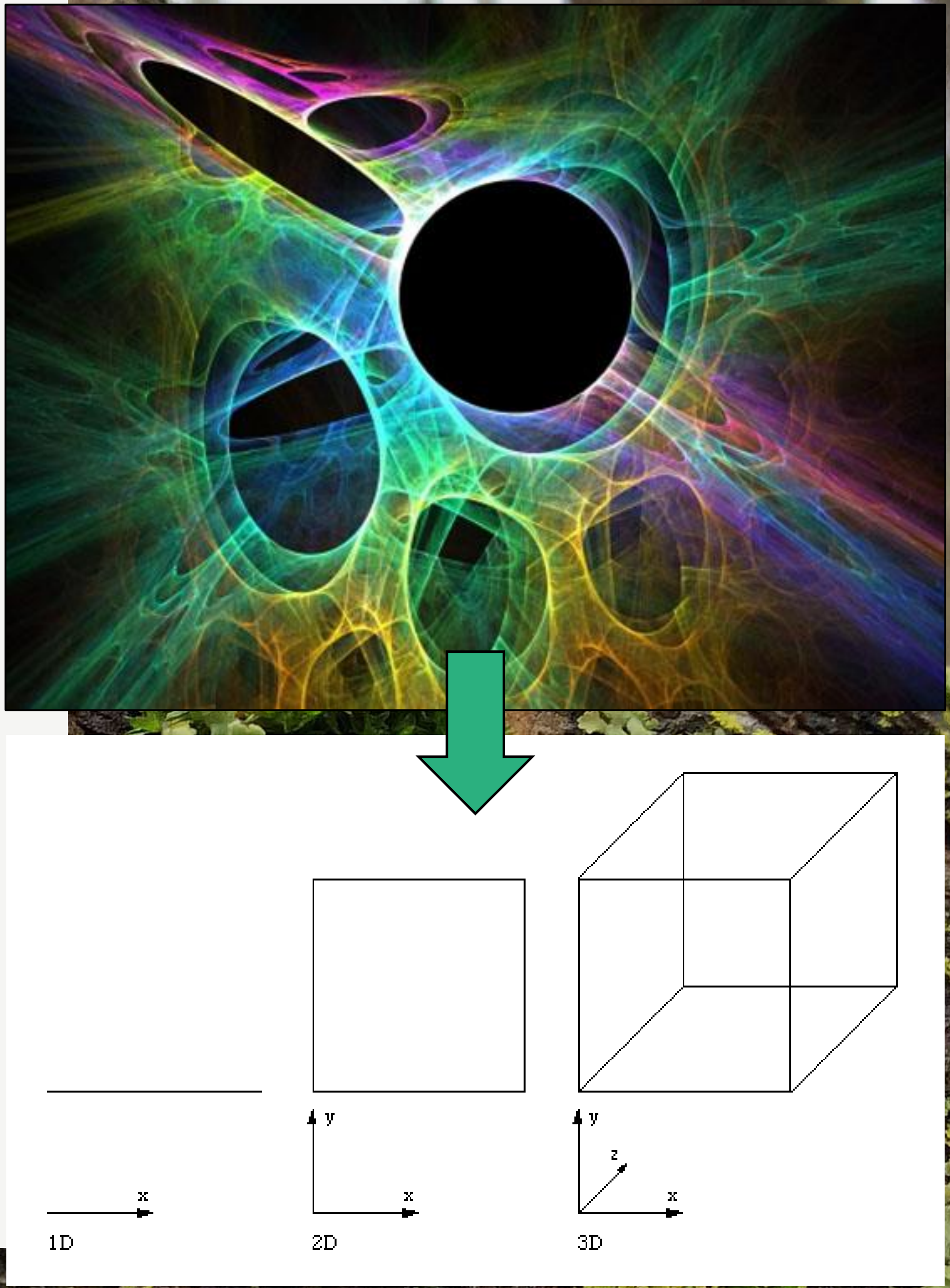
Multidimensional statistics are designed to account for the covarying nature of ecological data.



Intro: Ordination

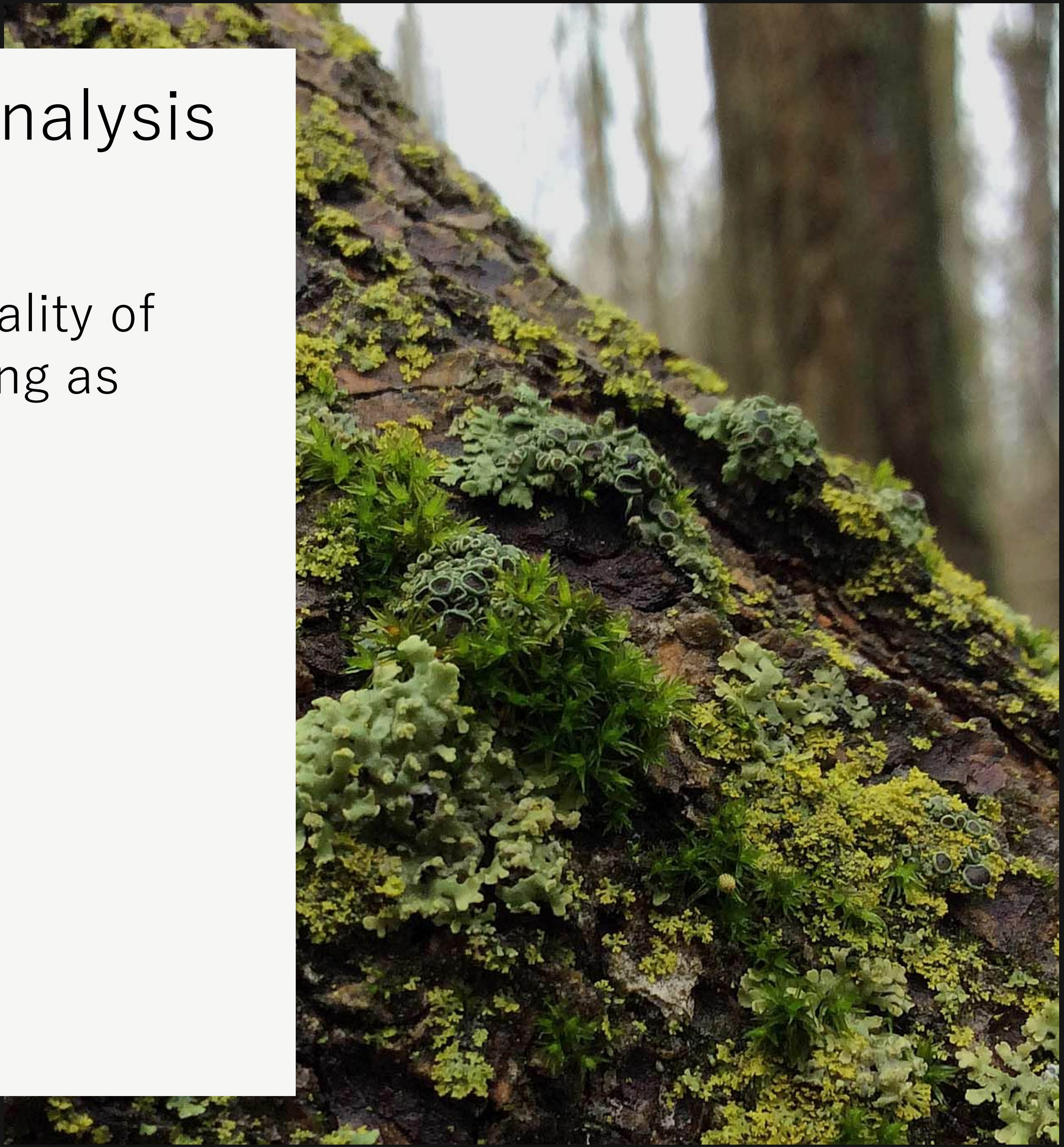
Multidimensional statistics are designed to account for the covarying nature of ecological data.

Ordination (or **gradient analysis**) is an <u>exploratory</u> technique that simplifies large ecological datasets by representing them in a reduced number of dimensions.



Intro: Principal Component Analysis

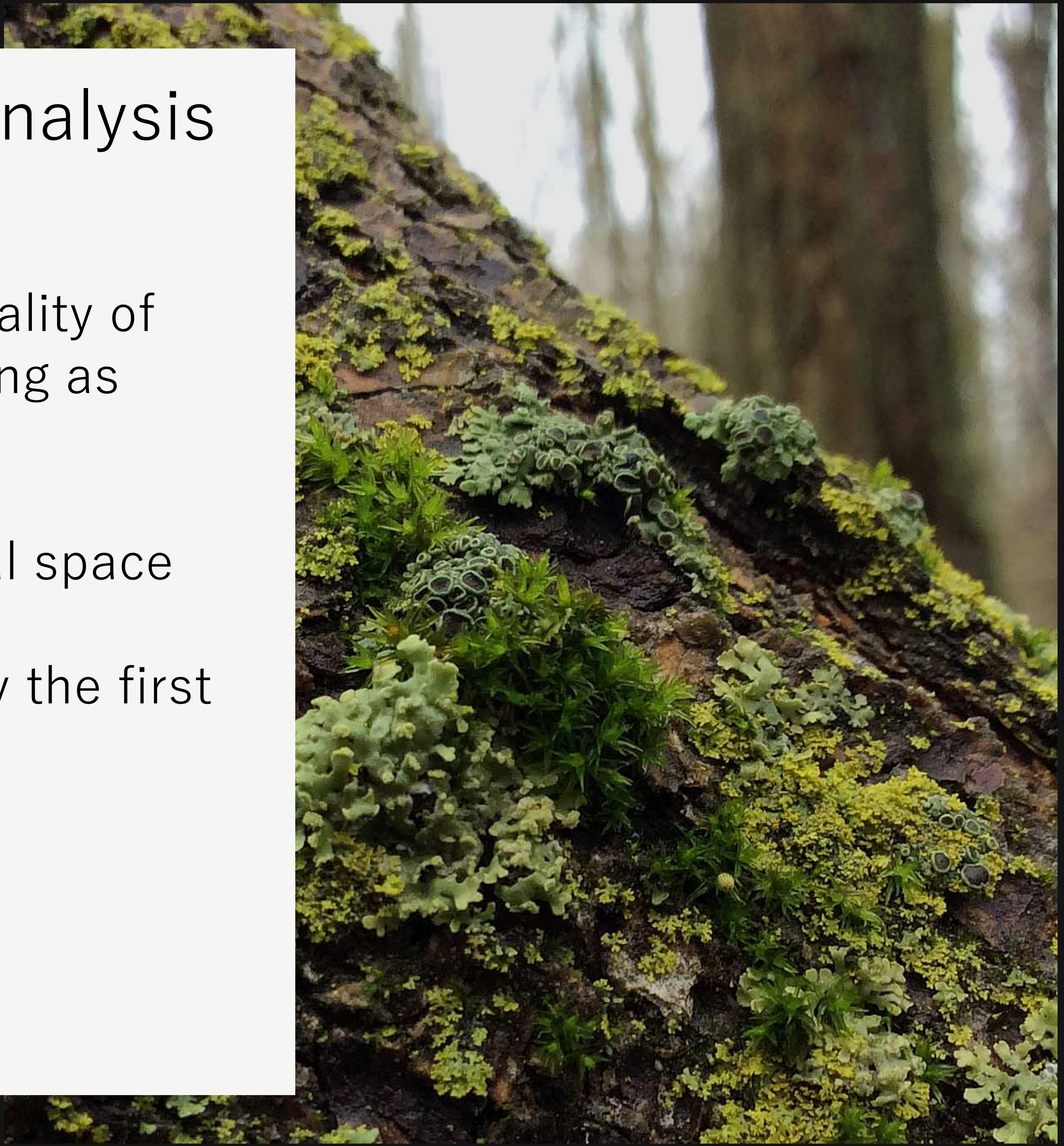
Principal Component Analysis uses eigenanalysis to reduce the dimensionality of large, ecological datasets while retaining as much information as possible.



Intro: Principal Component Analysis

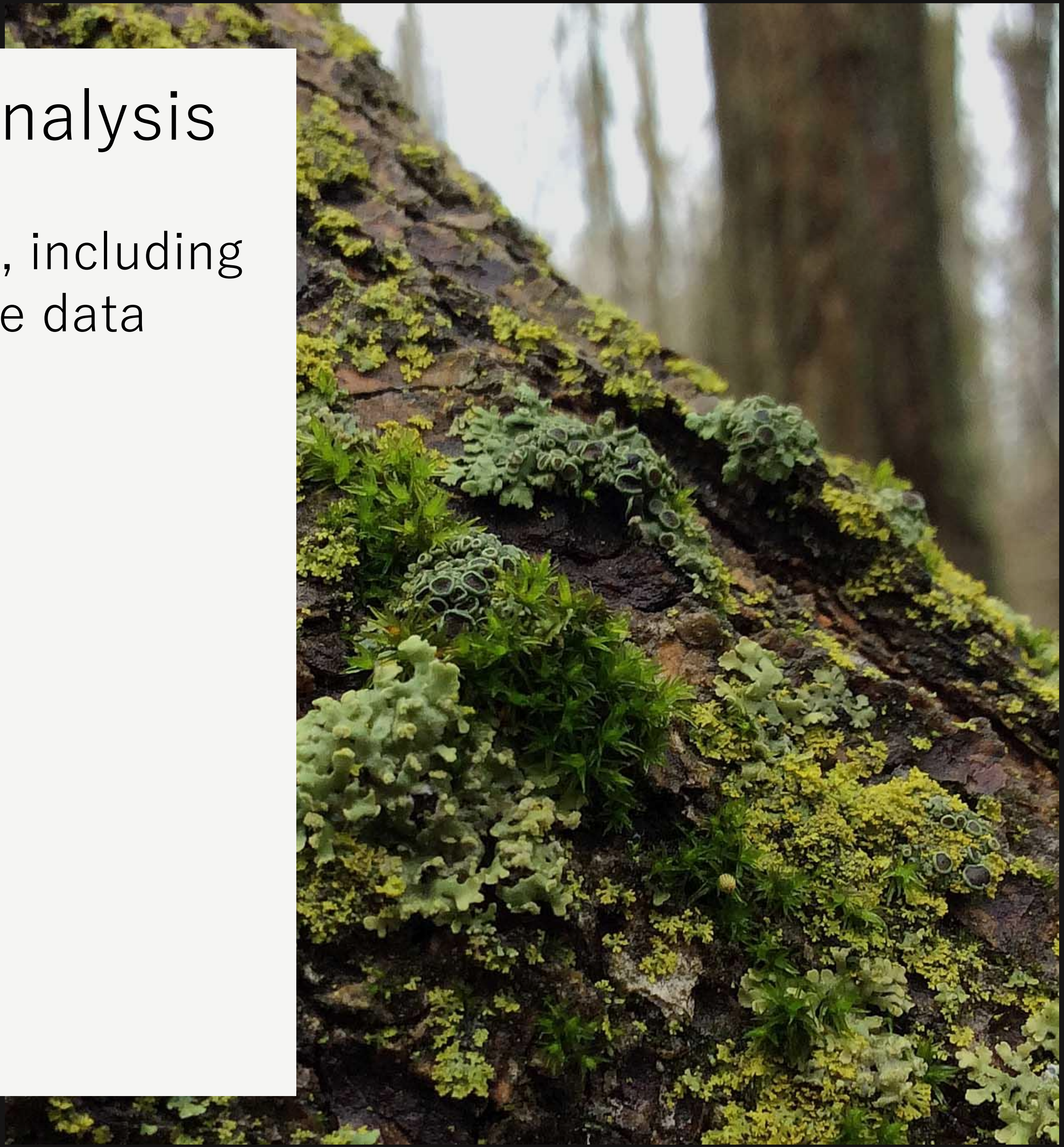
Principal Component Analysis uses eigenanalysis to reduce the dimensionality of large, ecological datasets while retaining as much information as possible.

- Re-projects data in multidimensional space
- Maximizes the variance explained by the first principal axes (eigenvectors)



Intro: Principal Component Analysis

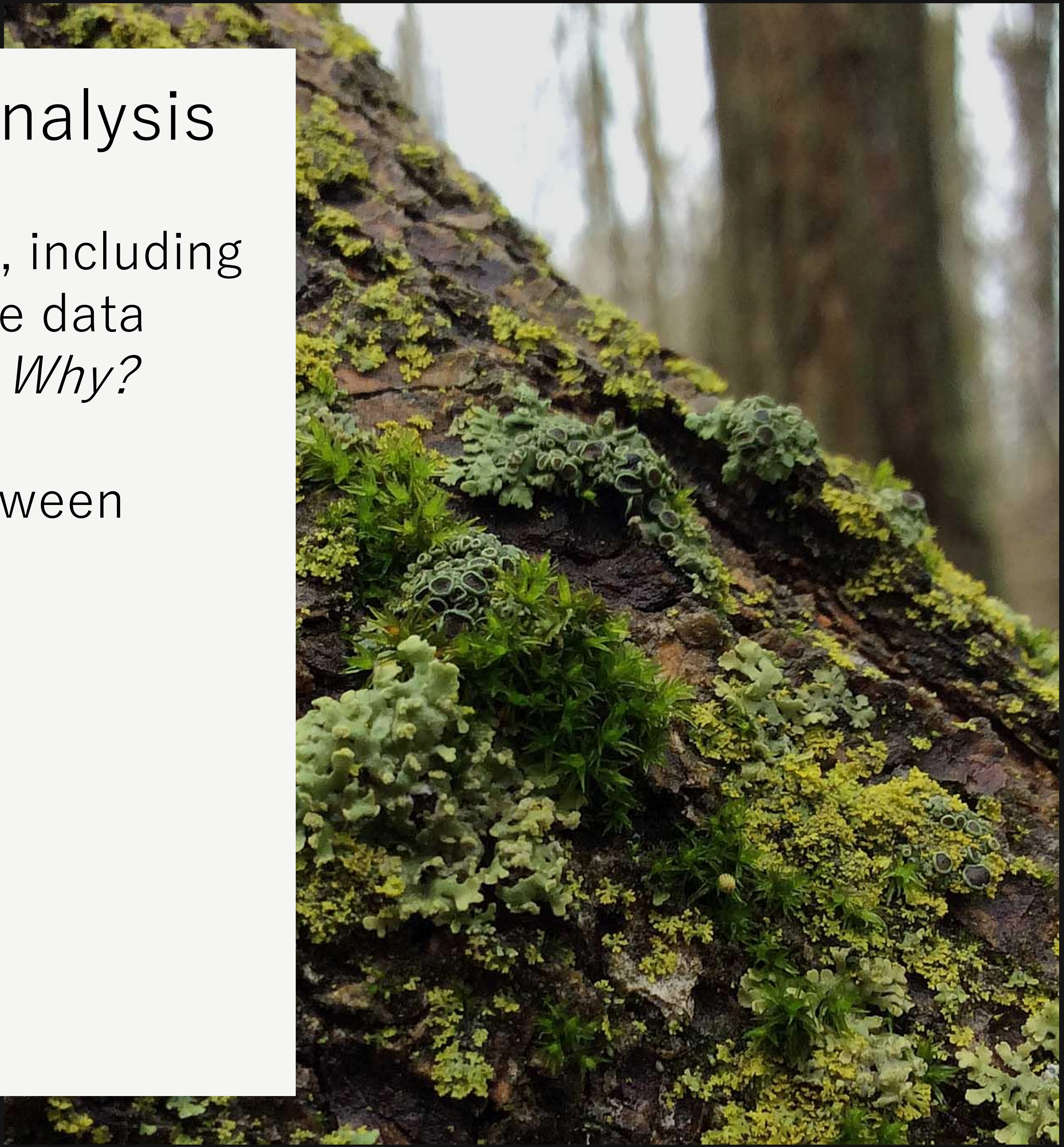
Many methods of multivariate analysis, including PCA, perform better when the response data distributions are **multivariate normal**.



Intro: Principal Component Analysis

Many methods of multivariate analysis, including PCA, perform better when the response data distributions are **multivariate normal**. *Why?*

1. PCA Assumes the relationships between variables are linear

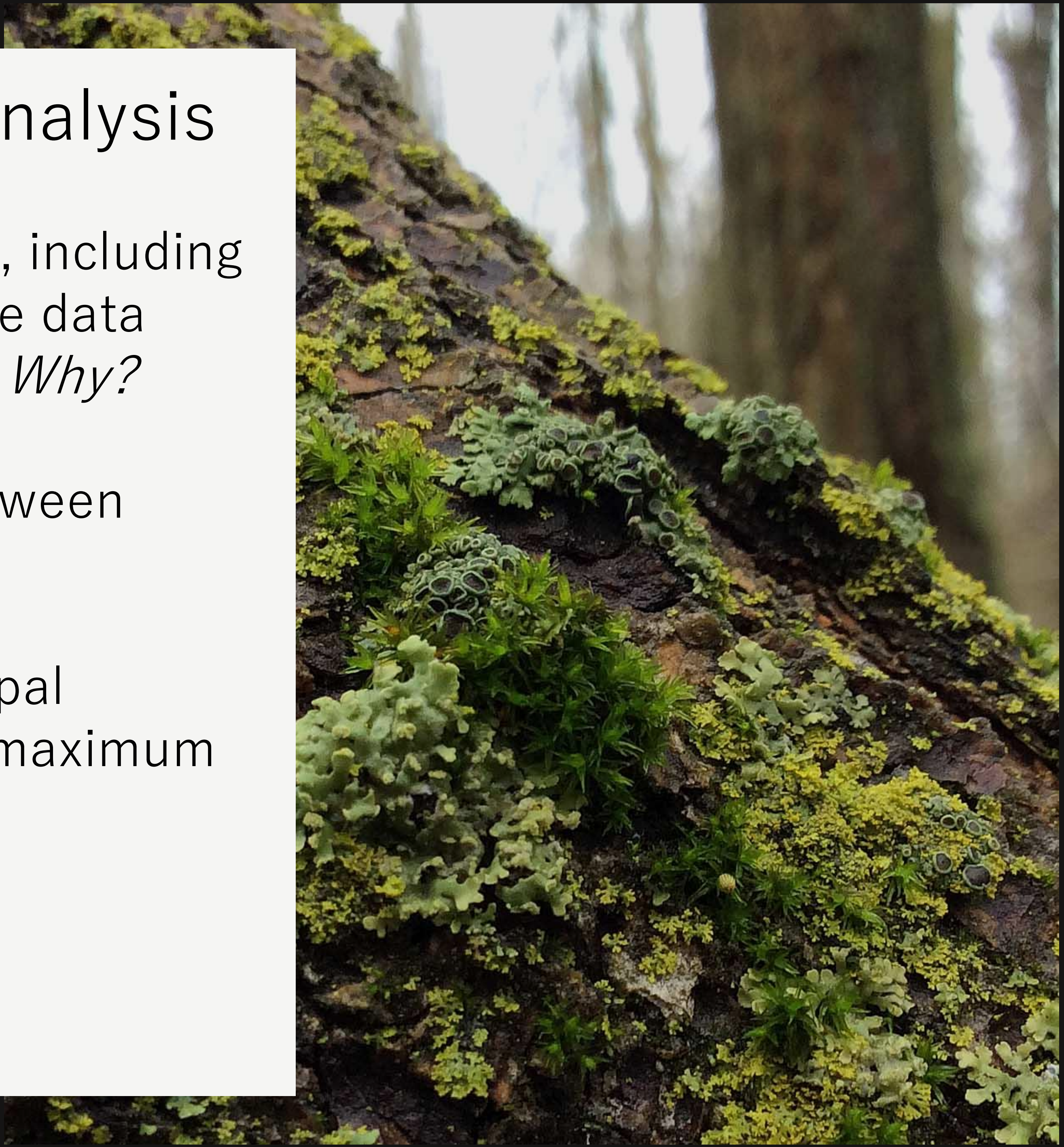


Intro: Principal Component Analysis

Many methods of multivariate analysis, including PCA, perform better when the response data distributions are **multivariate normal**. *Why?*

1. PCA Assumes the relationships between variables are linear

2. PCA depends on aligning the principal components with the directions of maximum variability



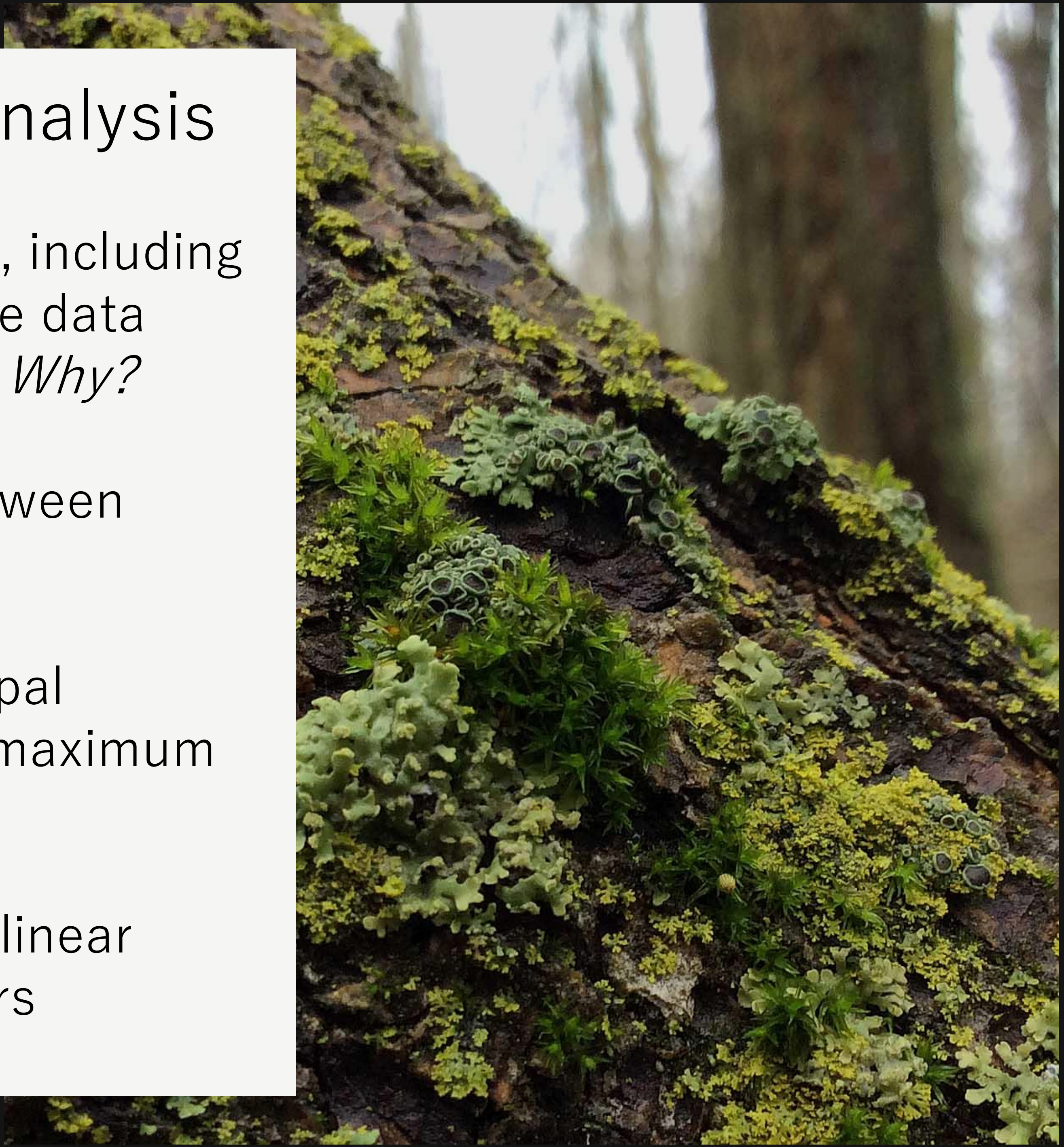
Intro: Principal Component Analysis

Many methods of multivariate analysis, including PCA, perform better when the response data distributions are **multivariate normal**. *Why?*

1. PCA Assumes the relationships between variables are linear

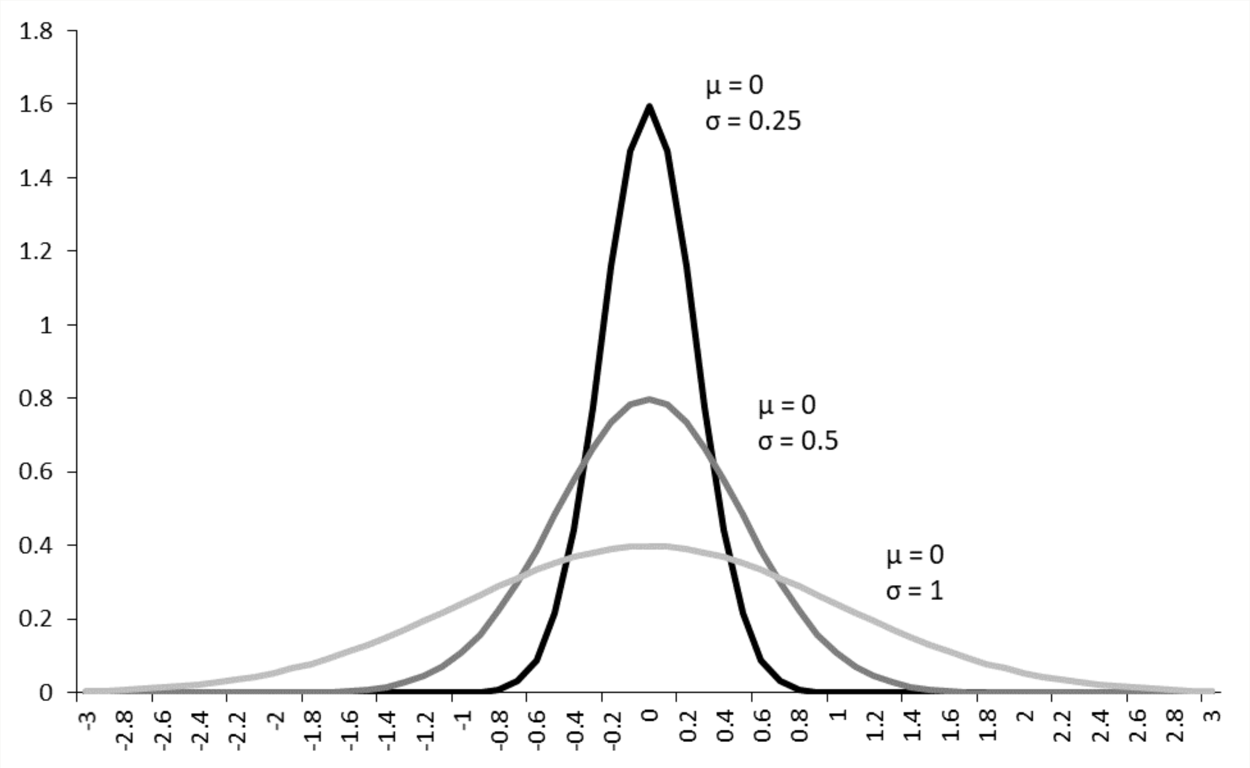
2. PCA depends on aligning the principal components with the directions of maximum variability

3. Interpretation is influenced by non-linear relationships, skewness, and outliers



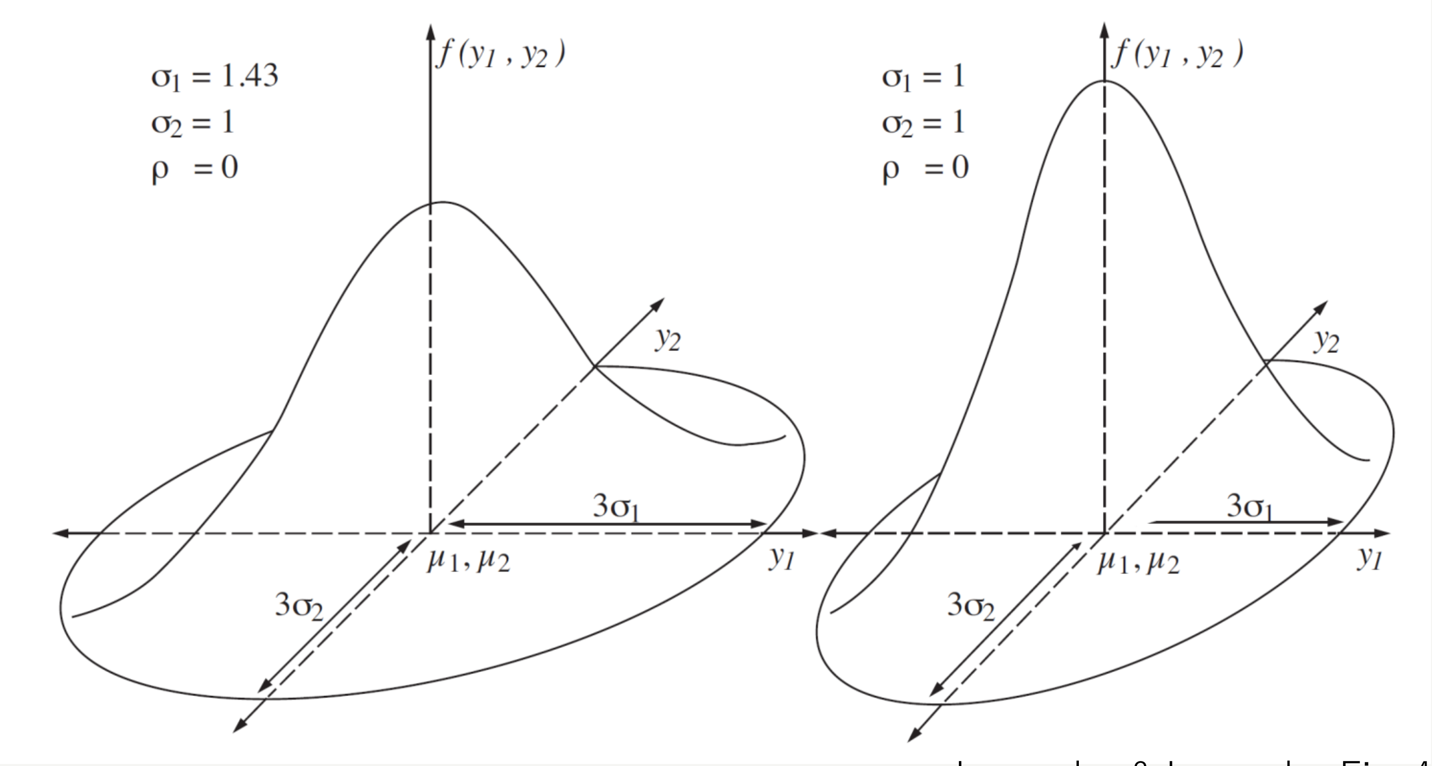
Intro: Principal Component Analysis

Univariate normal distribution: Only requires mean (μ) and standard deviation (σ).

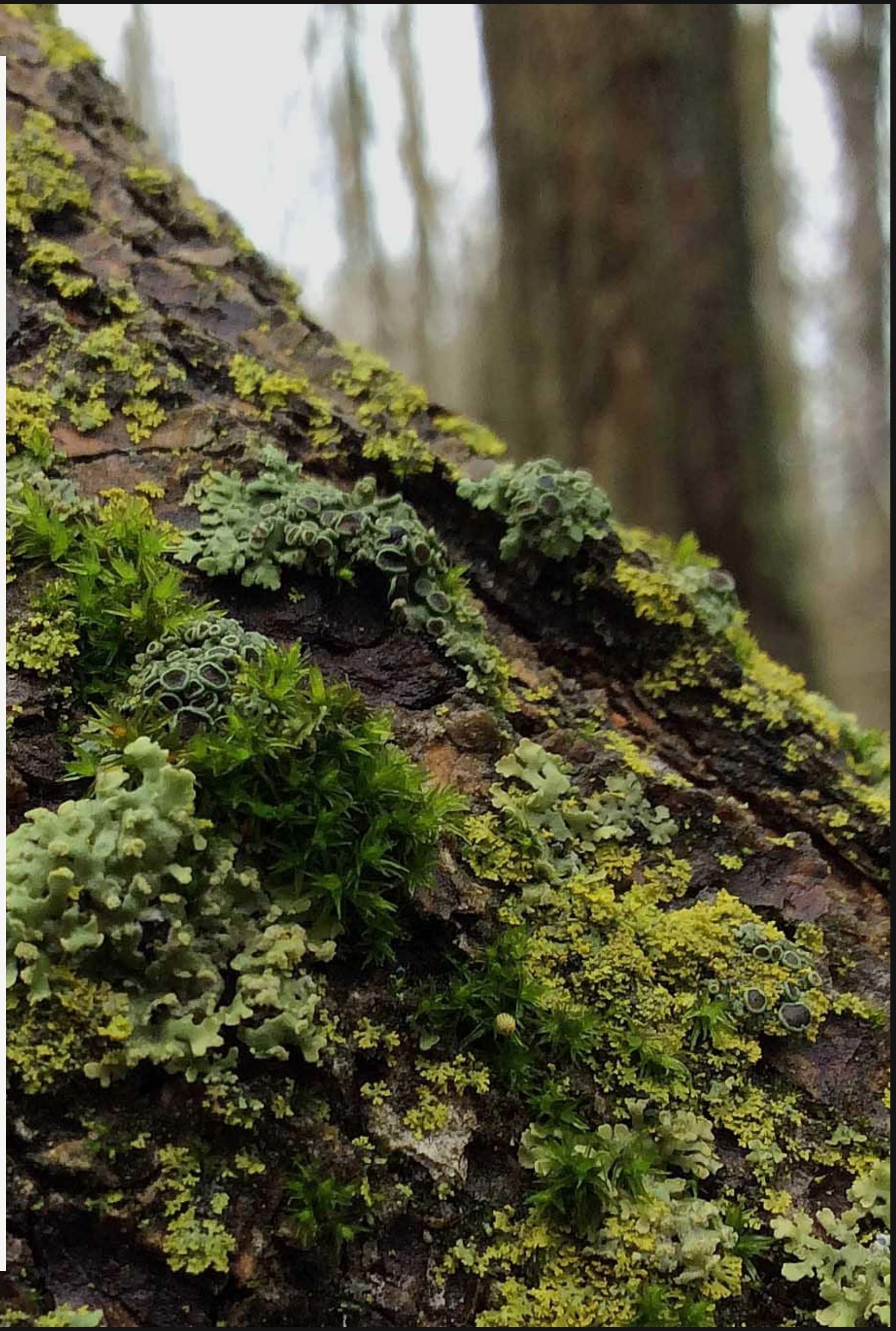


Intro: Principal Component Analysis

Multivariate normal distribution: Requires mean (μ), standard deviation (σ), and correlation (ρ).

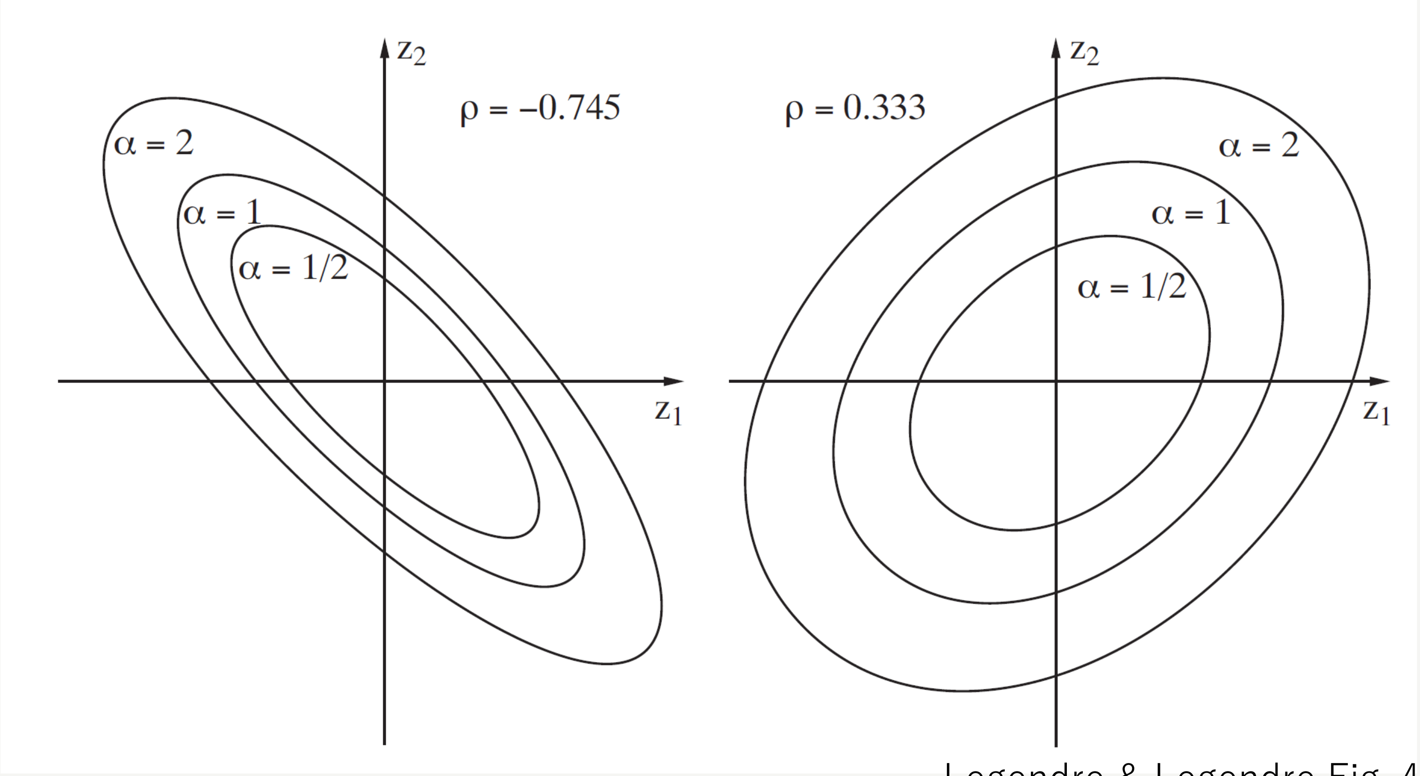


Legendre & Legendre Fig. 4.6

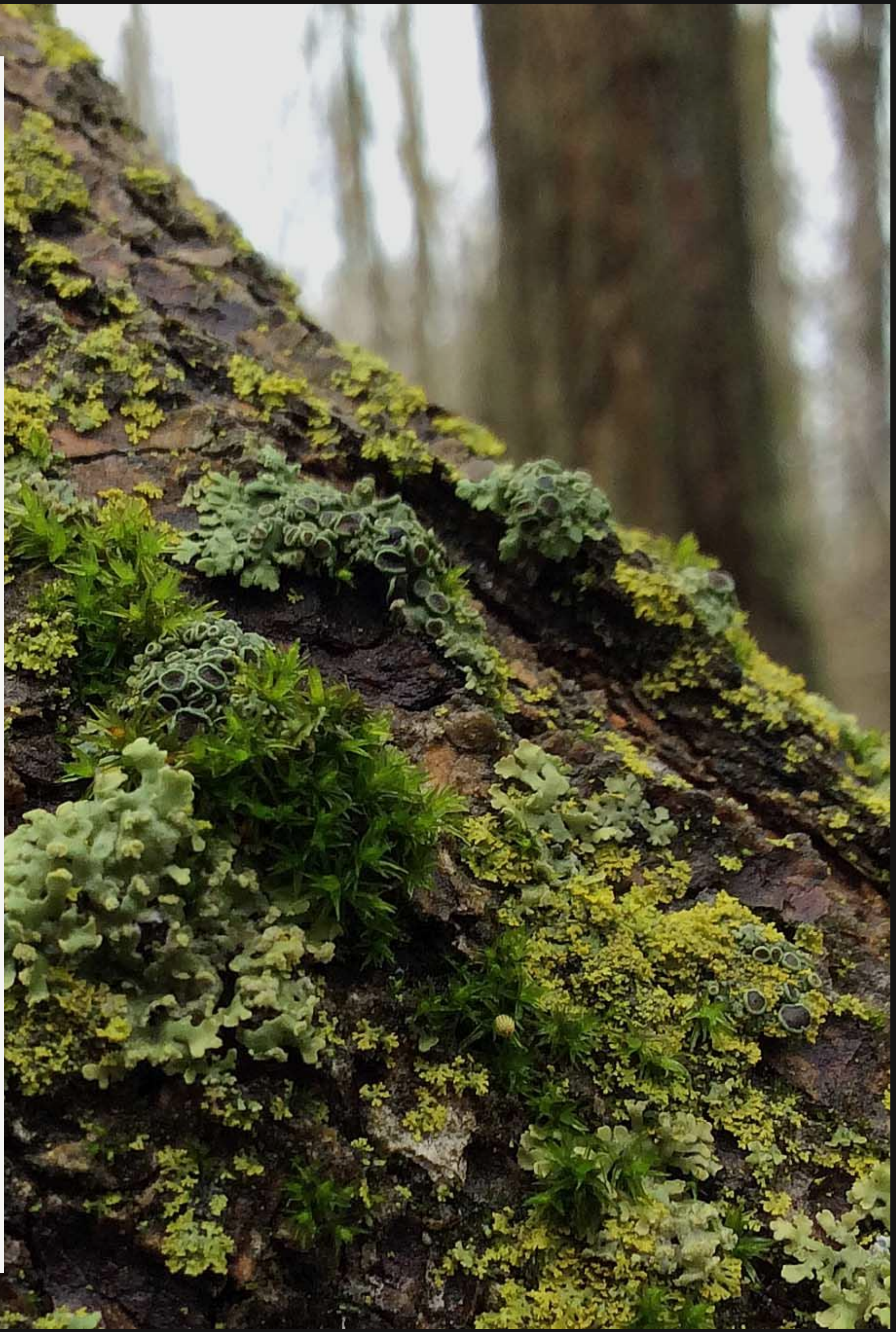


Intro: Principal Component Analysis

Multivariate normal distribution: Requires mean (μ), standard deviation (σ), and <u>correlation</u> (ρ).

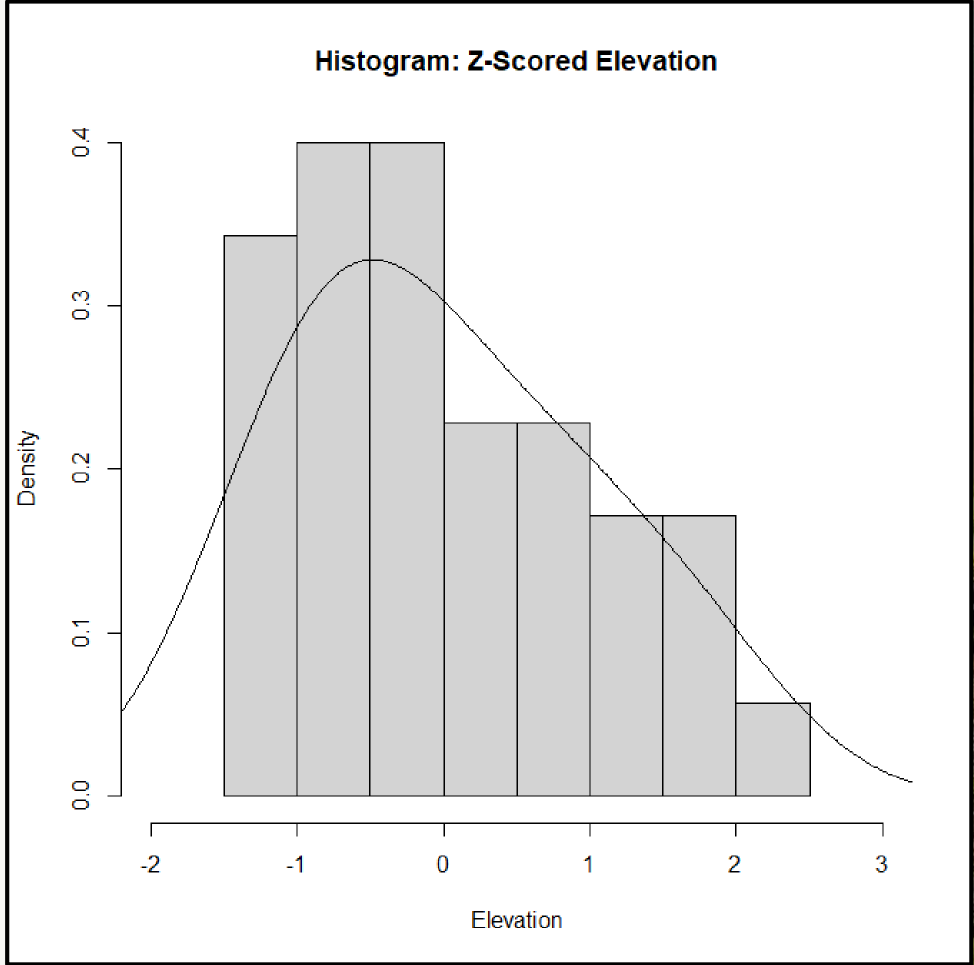


Legendre & Legendre Fig. 4.7



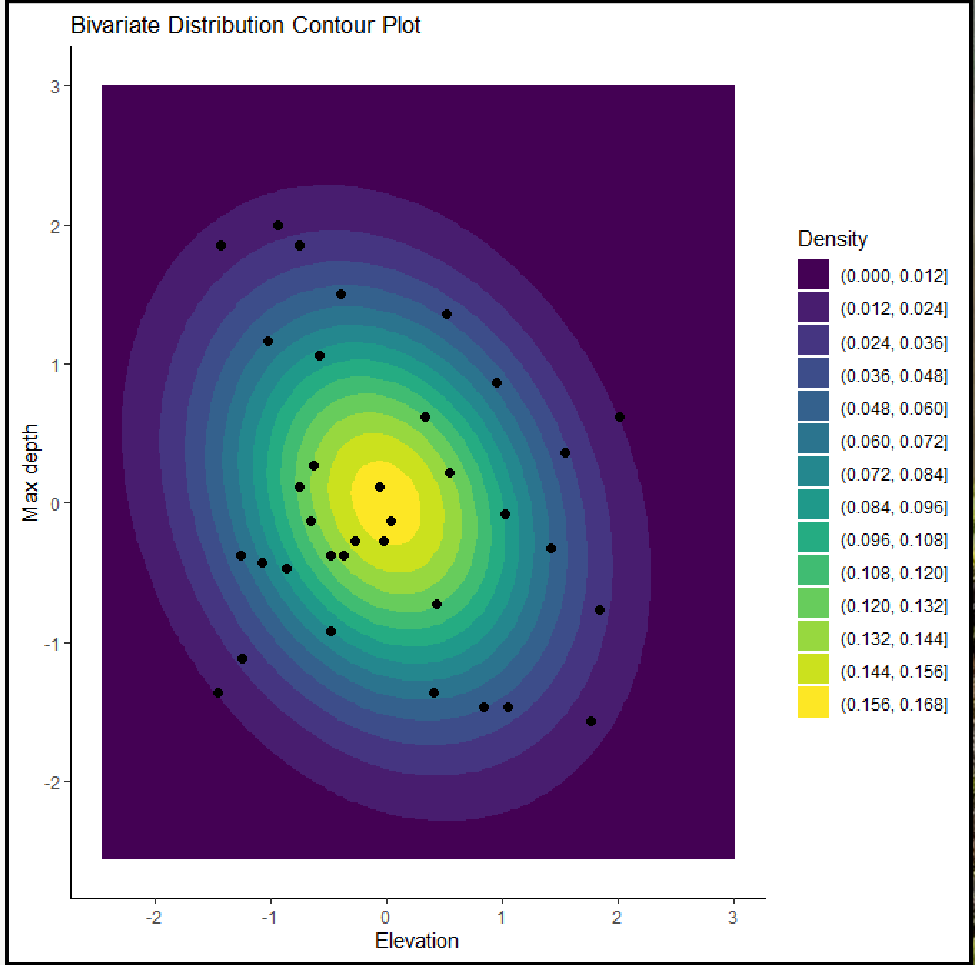
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



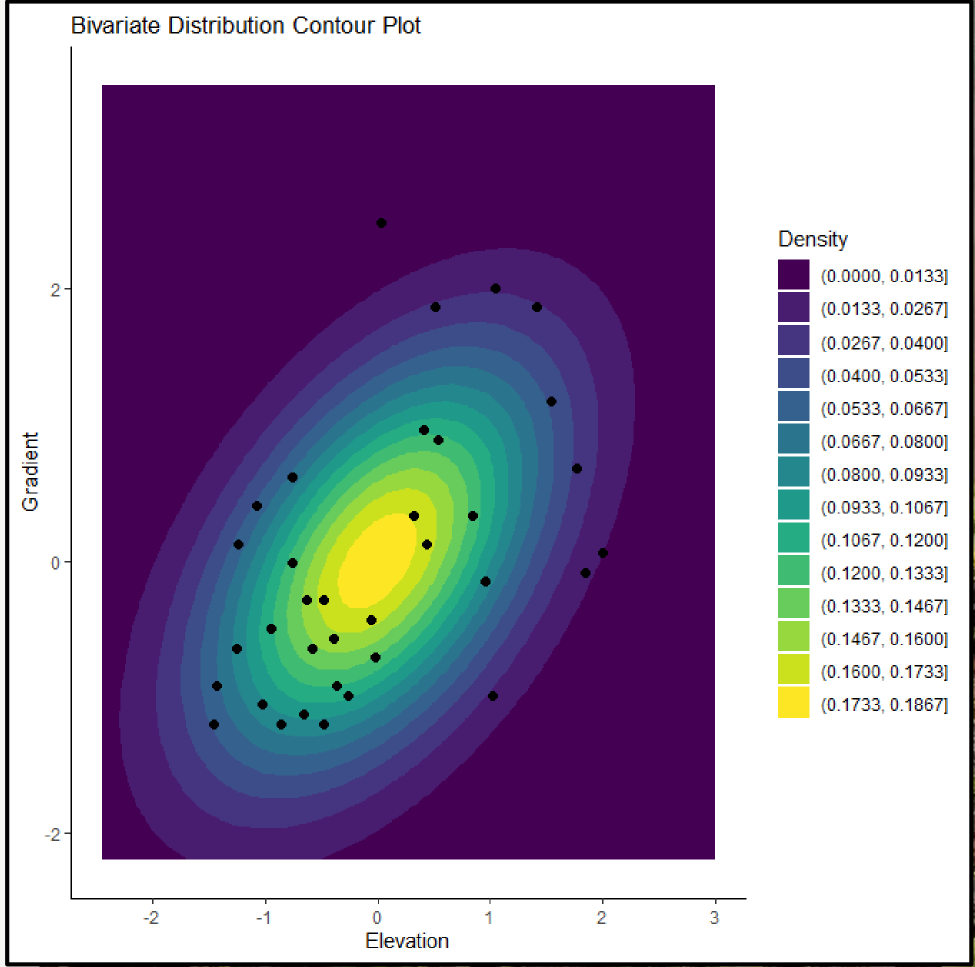
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



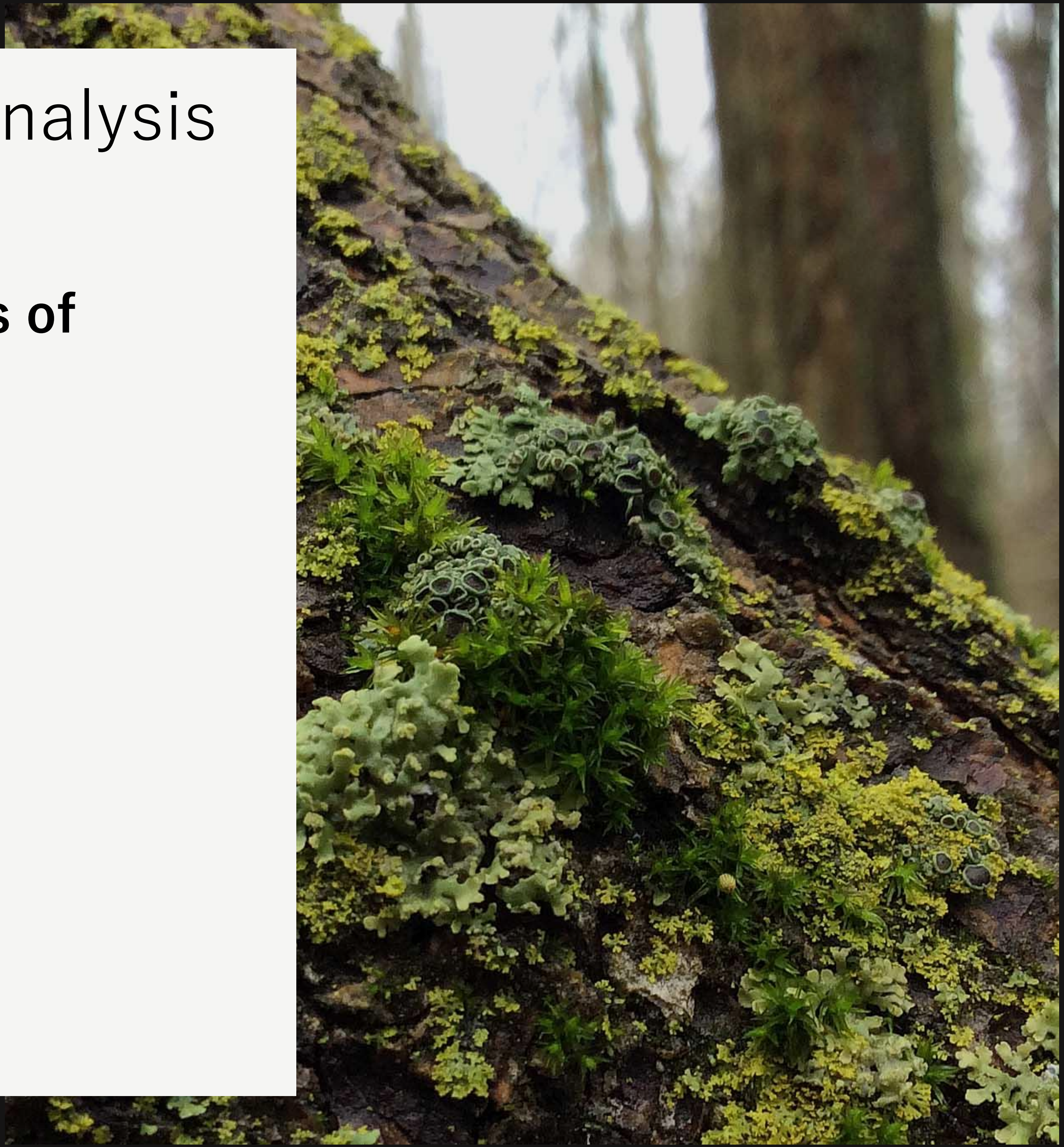
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



Intro: Principal Component Analysis

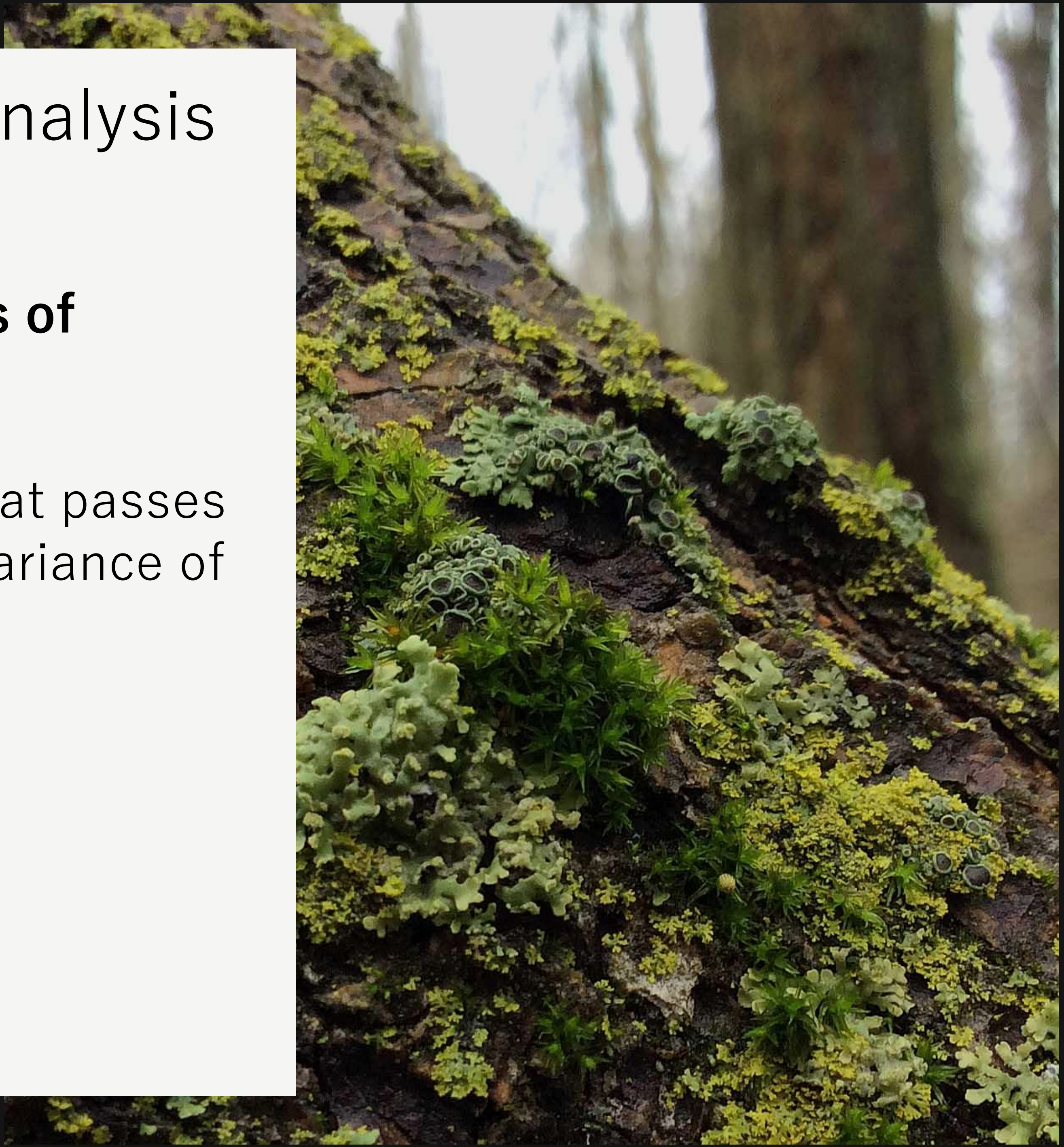
PCA depends on aligning the principal components (axes) with the **directions of maximum variability**.



Intro: Principal Component Analysis

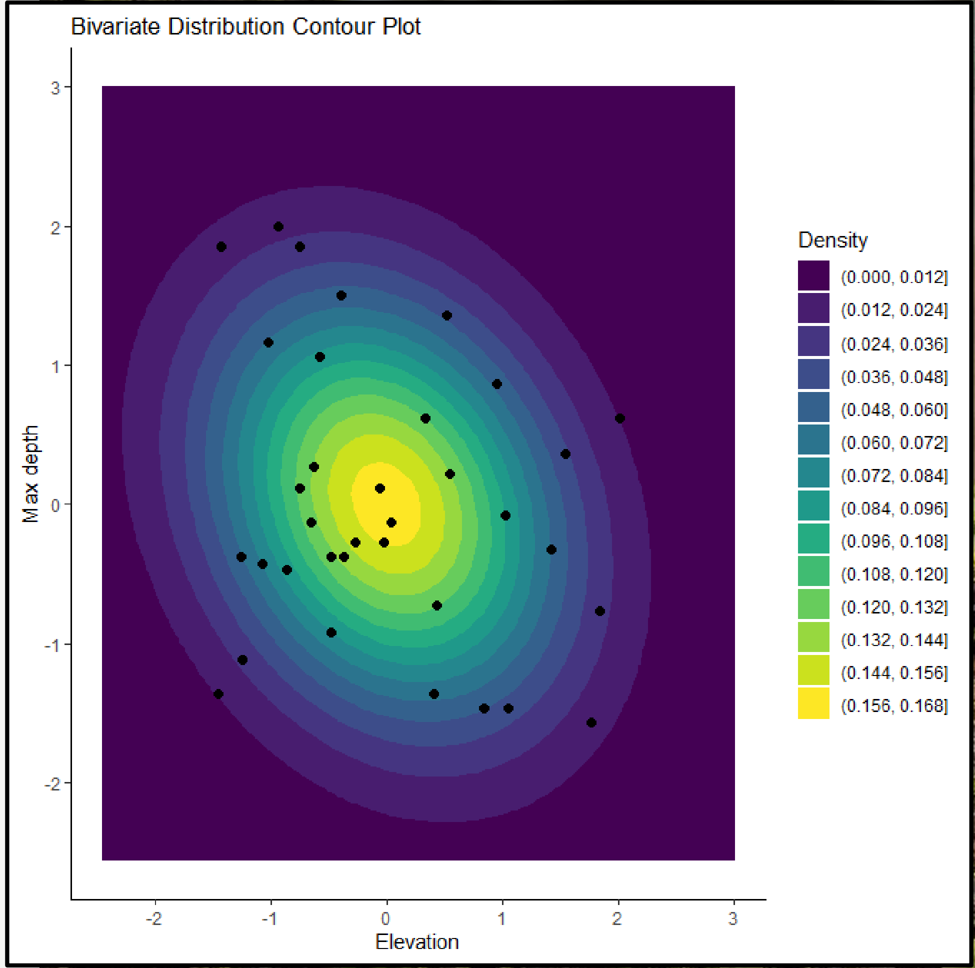
PCA depends on aligning the principal components (axes) with the **directions of maximum variability**.

- The first **principal axis** is the line that passes through the dimension of greatest variance of the ellipsoid.



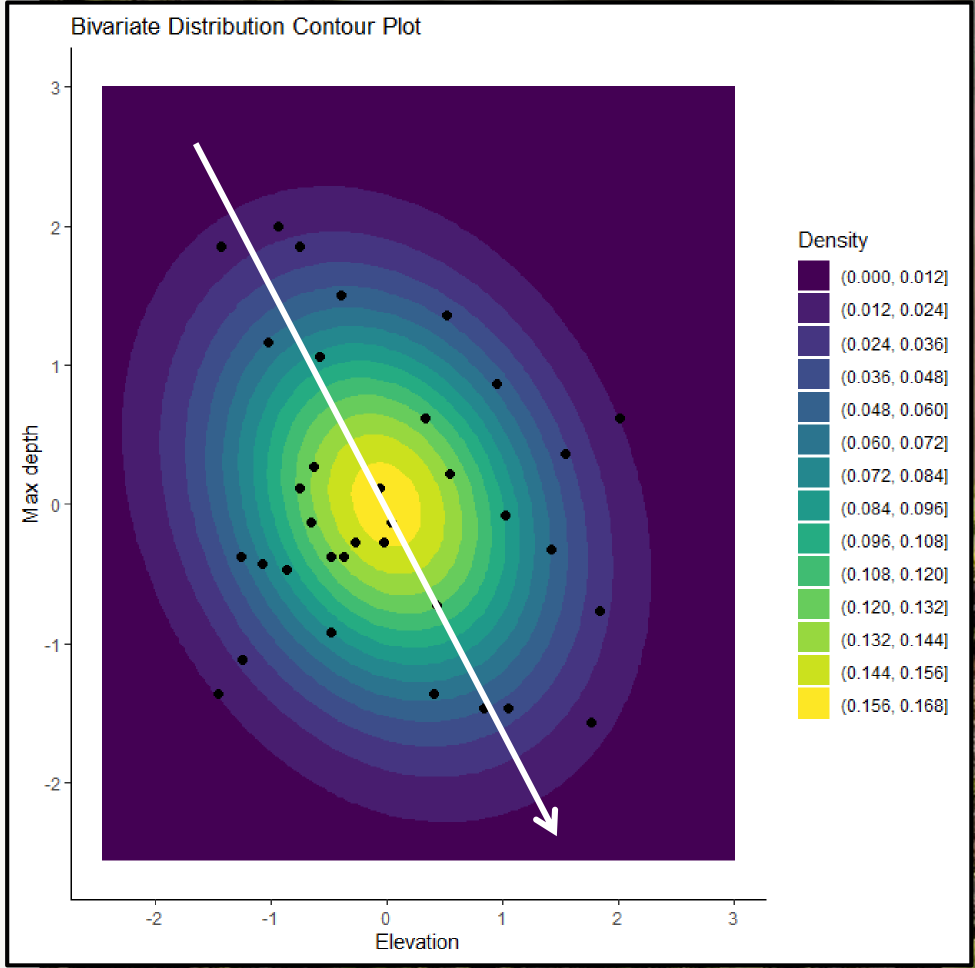
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



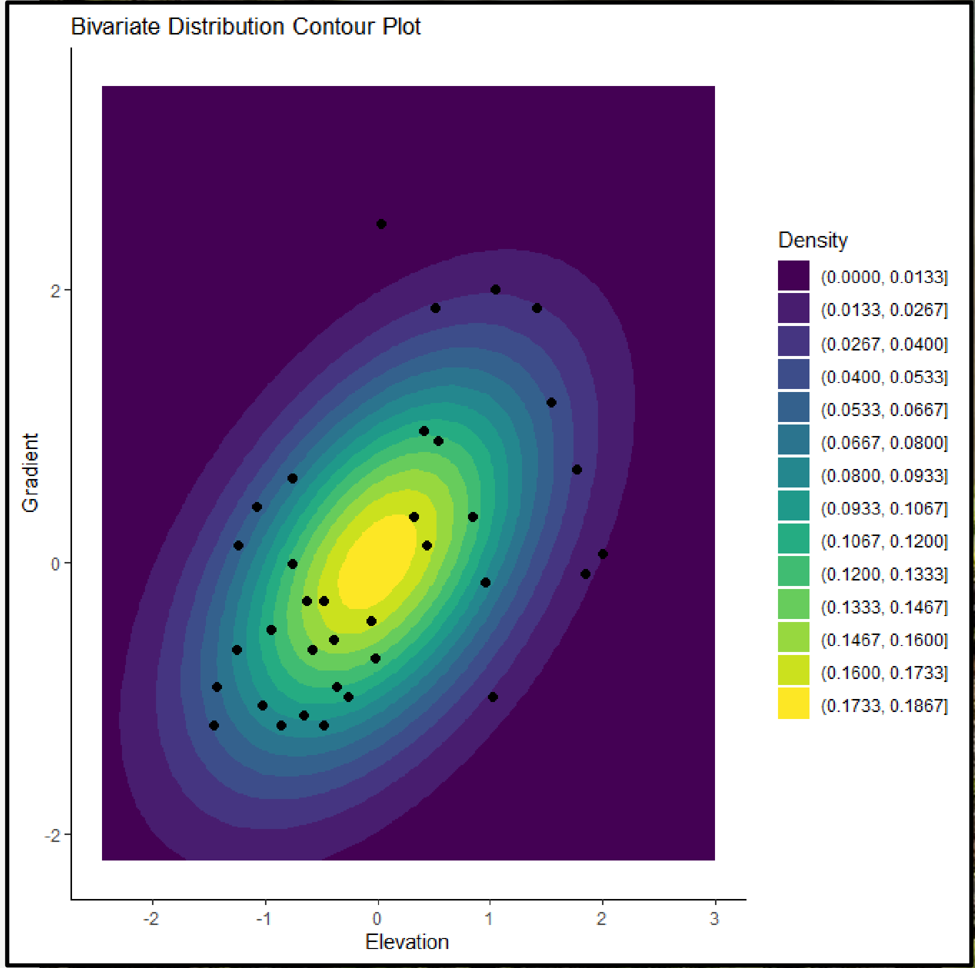
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



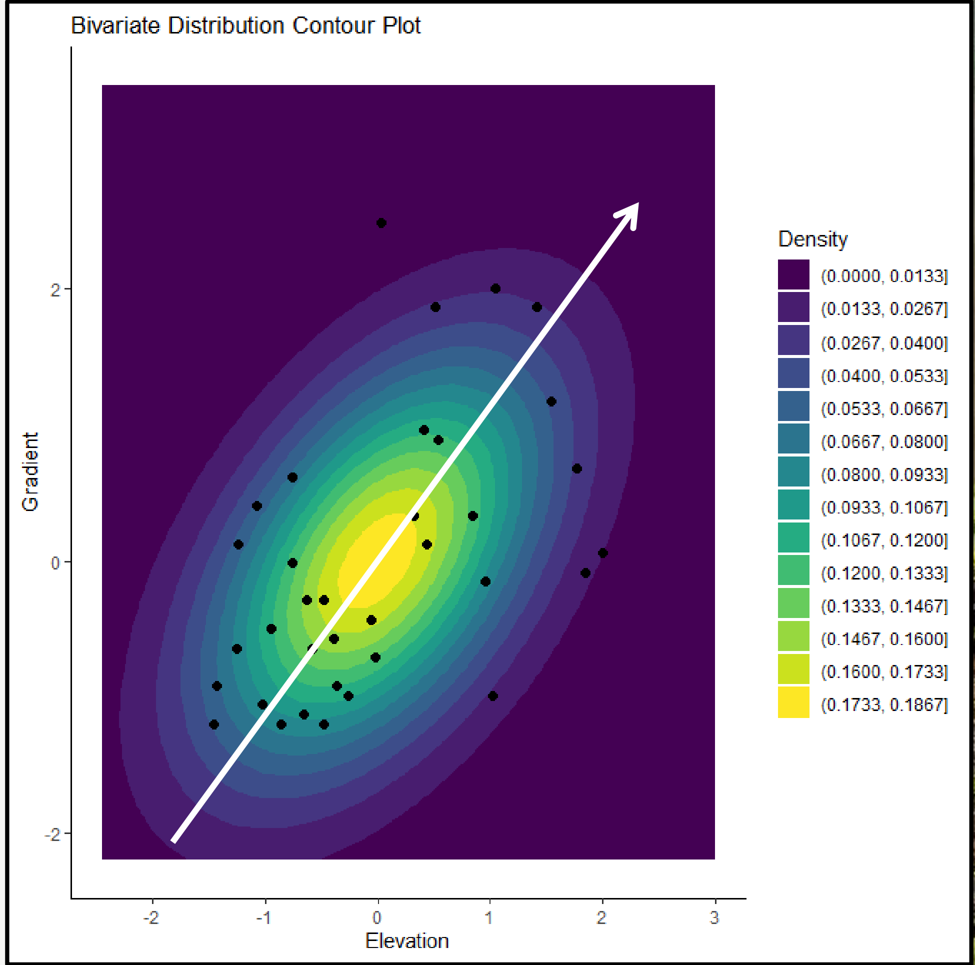
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



Intro: Principal Component Analysis

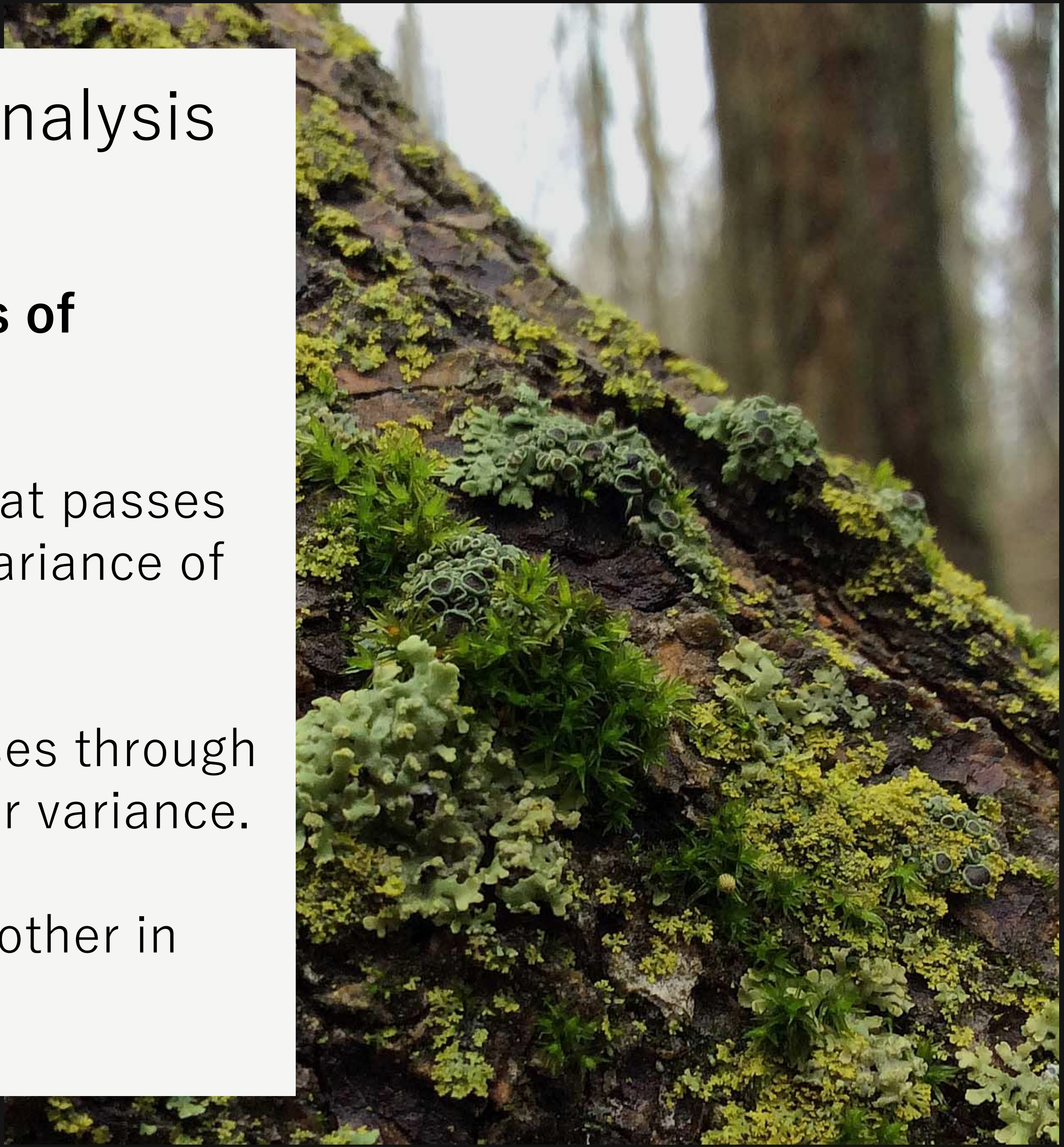
Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



Intro: Principal Component Analysis

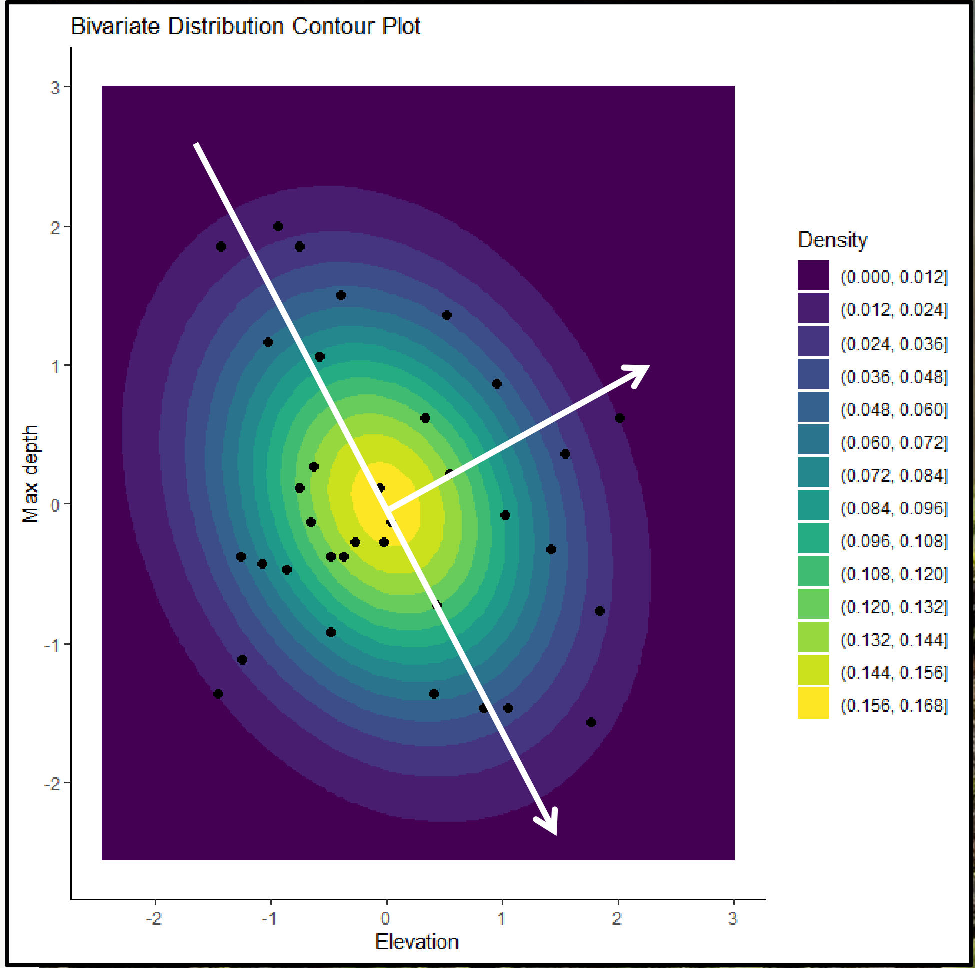
PCA depends on aligning the principal components (axes) with the **directions of maximum variability**.

- The first **principal axis** is the line that passes through the dimension of greatest variance of the ellipsoid.
- Each subsequent principal axis passes through dimensions of successionally smaller variance.
- All axes are <u>perpendicular</u> to one another in hyperspace.



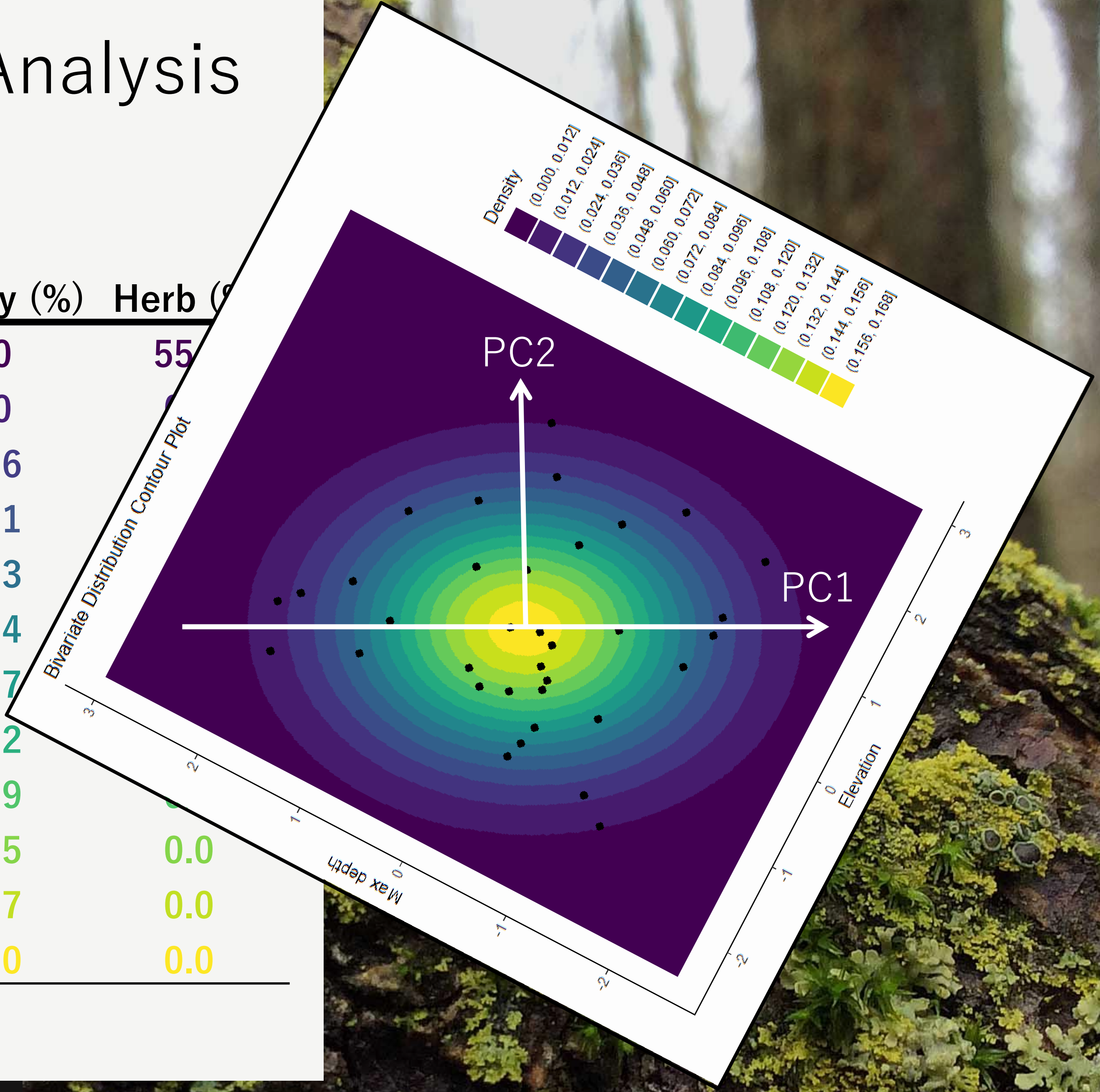
Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

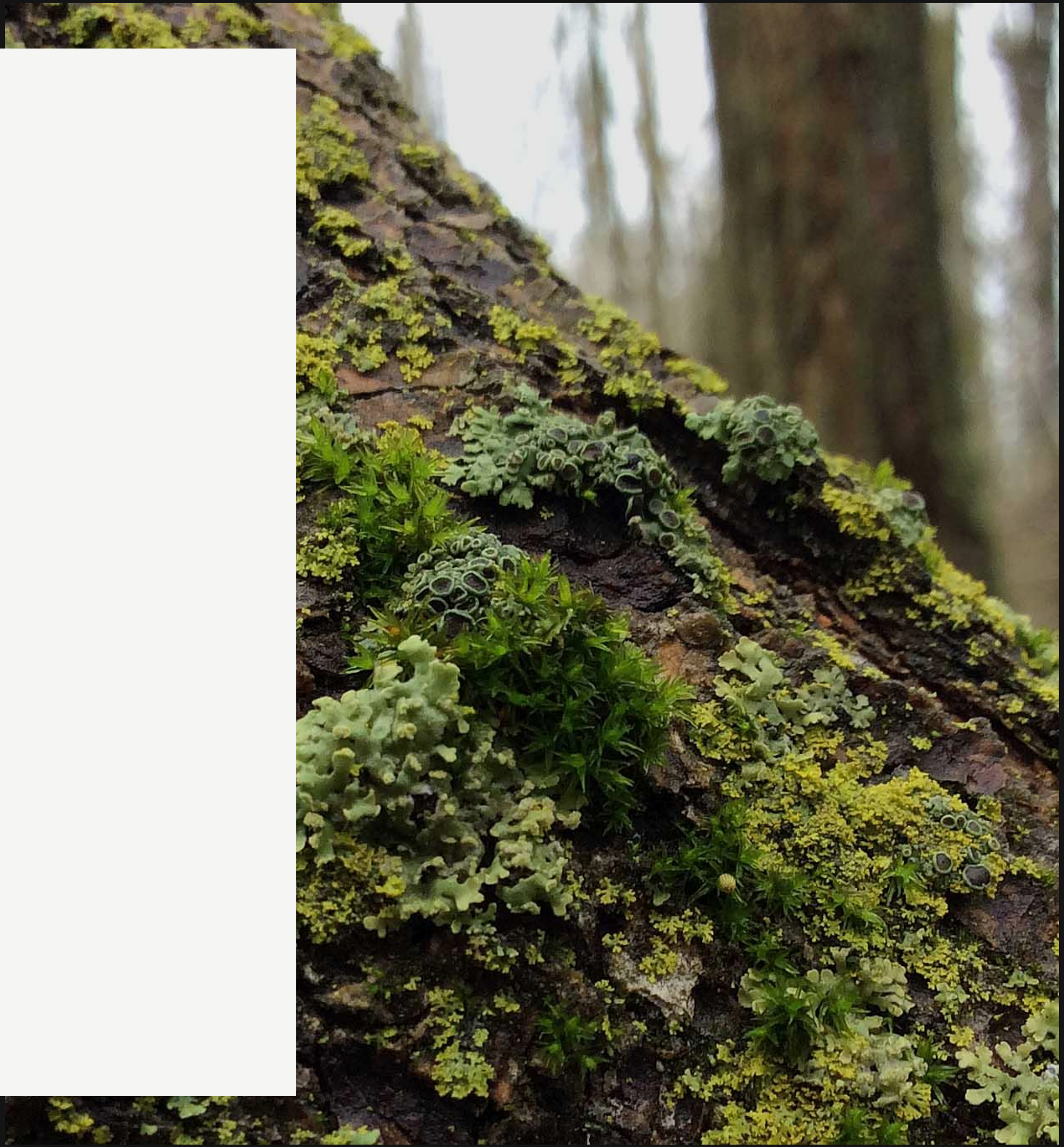


Intro: Principal Component Analysis

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%
Silvies-11	0.45	0.3	1439	0.0	55
Silvies-34	0.78	1.1	1487	0.0	
Silvies-02	0.71	0.4	1372	29.6	
Silvies-15	0.40	0.2	1471	41.1	
Silvies-07	0.50	1.3	1547	52.3	
Silvies-08	0.40	0.6	1492	51.4	
Silvies-22	0.42	0.9	1555	54.7	
Silvies-18	0.42	0.5	1510	46.2	
Silvies-12	0.52	3.2	1658	51.9	
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0

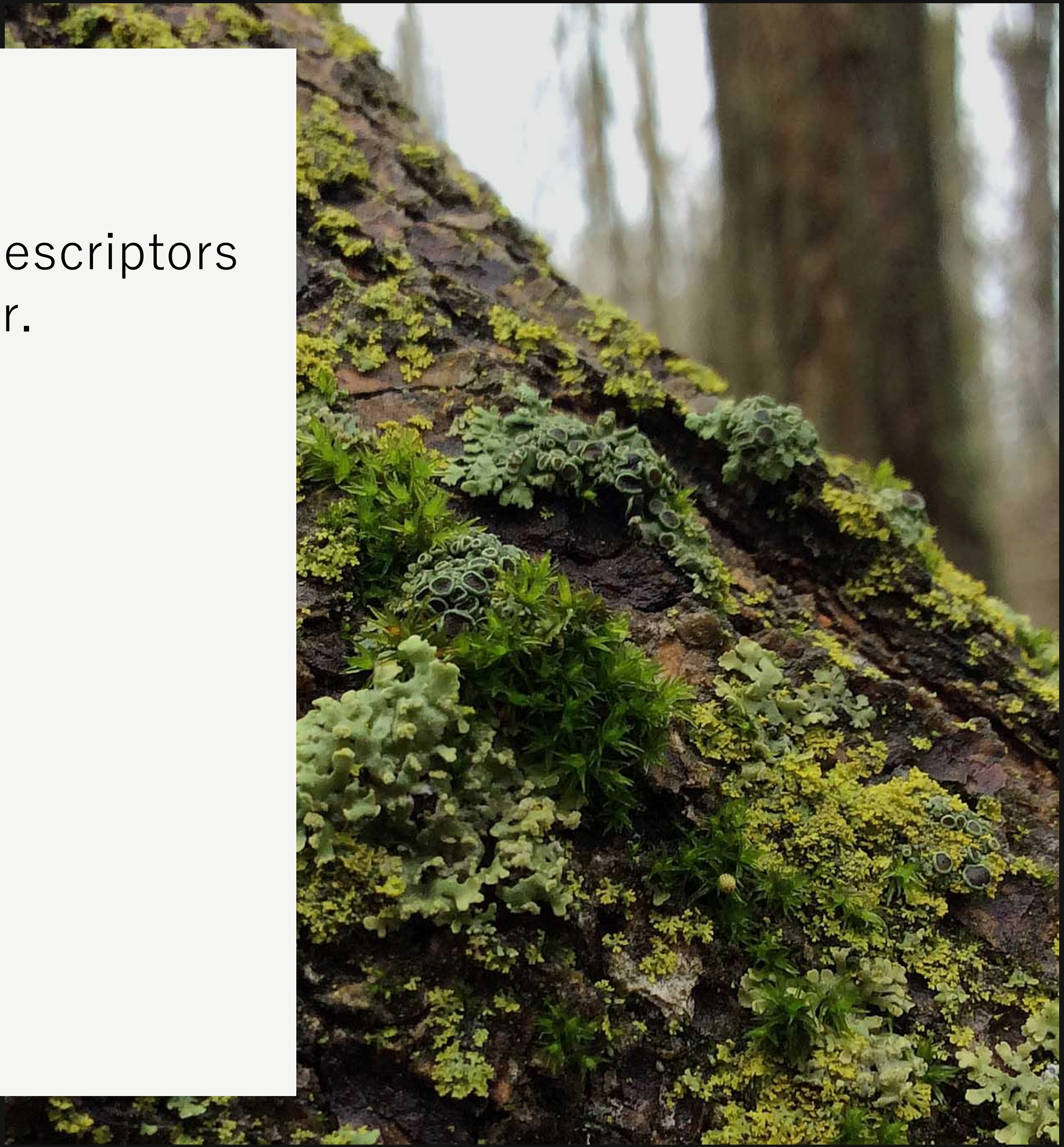


Dispersion Matrices



Dispersion Matrices

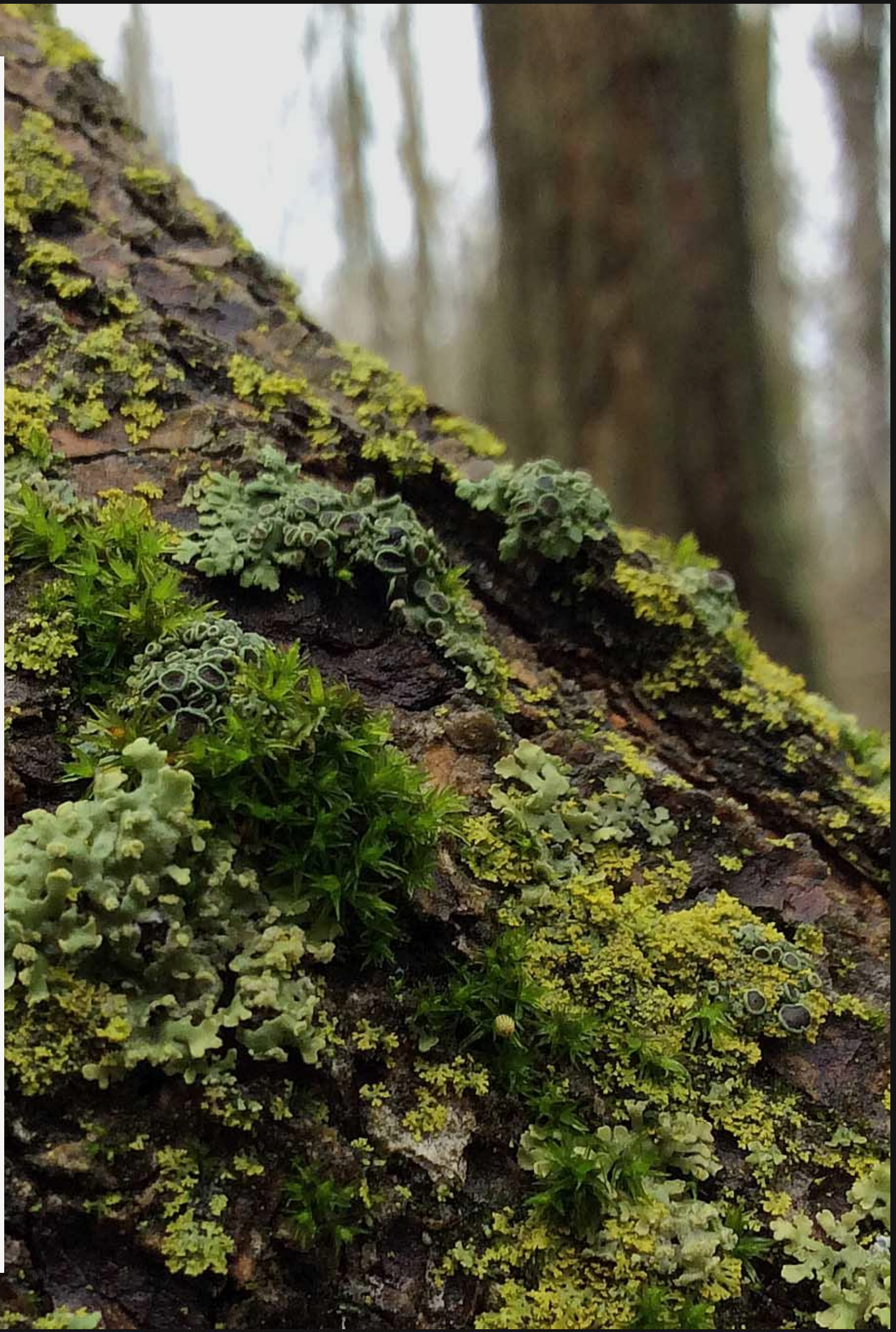
Univariate statistics assume that the descriptors are linearly independent of one another.



Dispersion Matrices

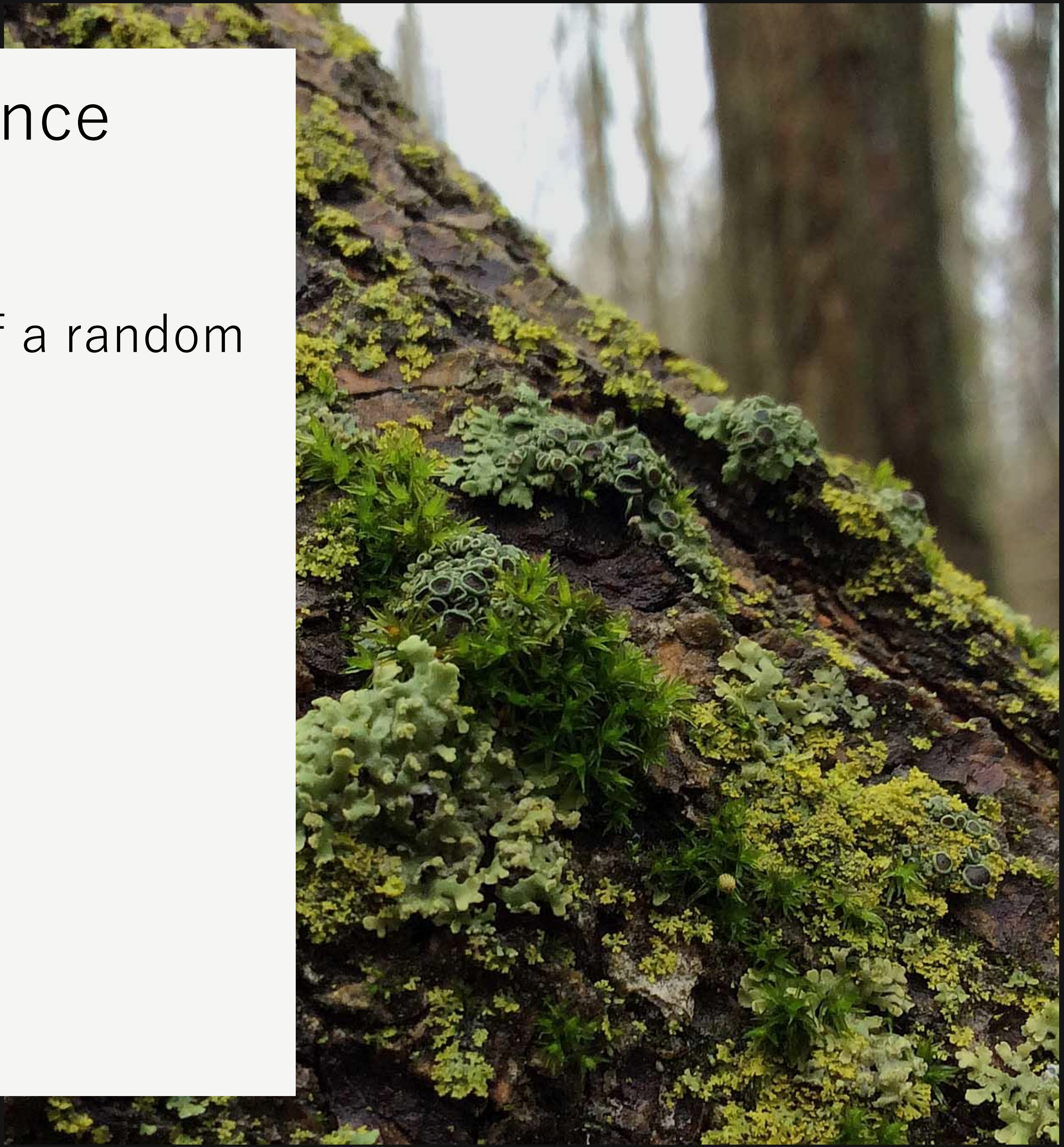
Univariate statistics assume that the descriptors are linearly independent of one another.

Multivariate methods account for the dependence among descriptors.



Dispersion Matrices: Covariance Matrix

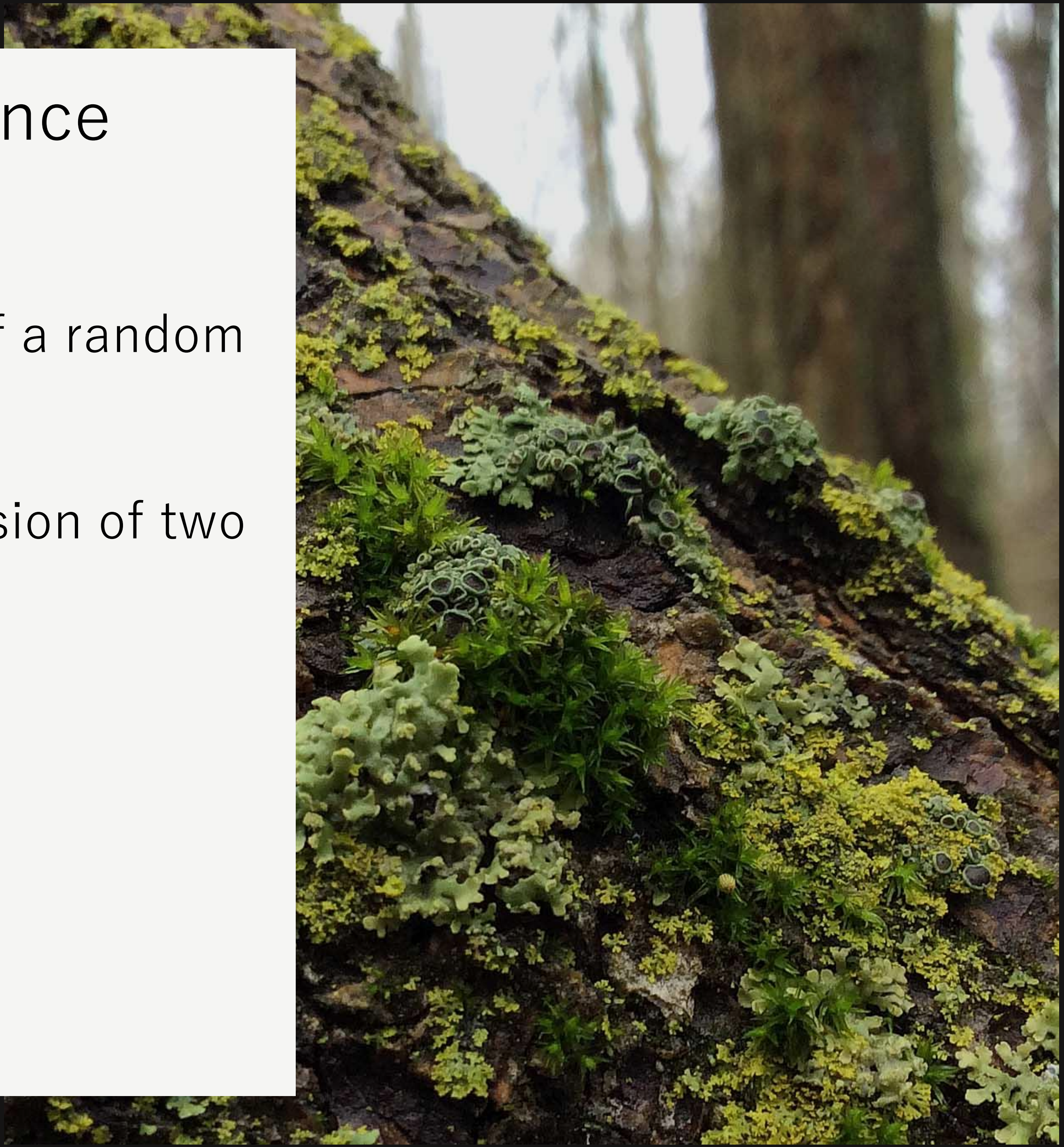
Variance is a measure of dispersion of a random variable around its mean.



Dispersion Matrices: Covariance Matrix

Variance is a measure of dispersion of a random variable around its mean.

Covariance measures the joint dispersion of two random variables around their means.



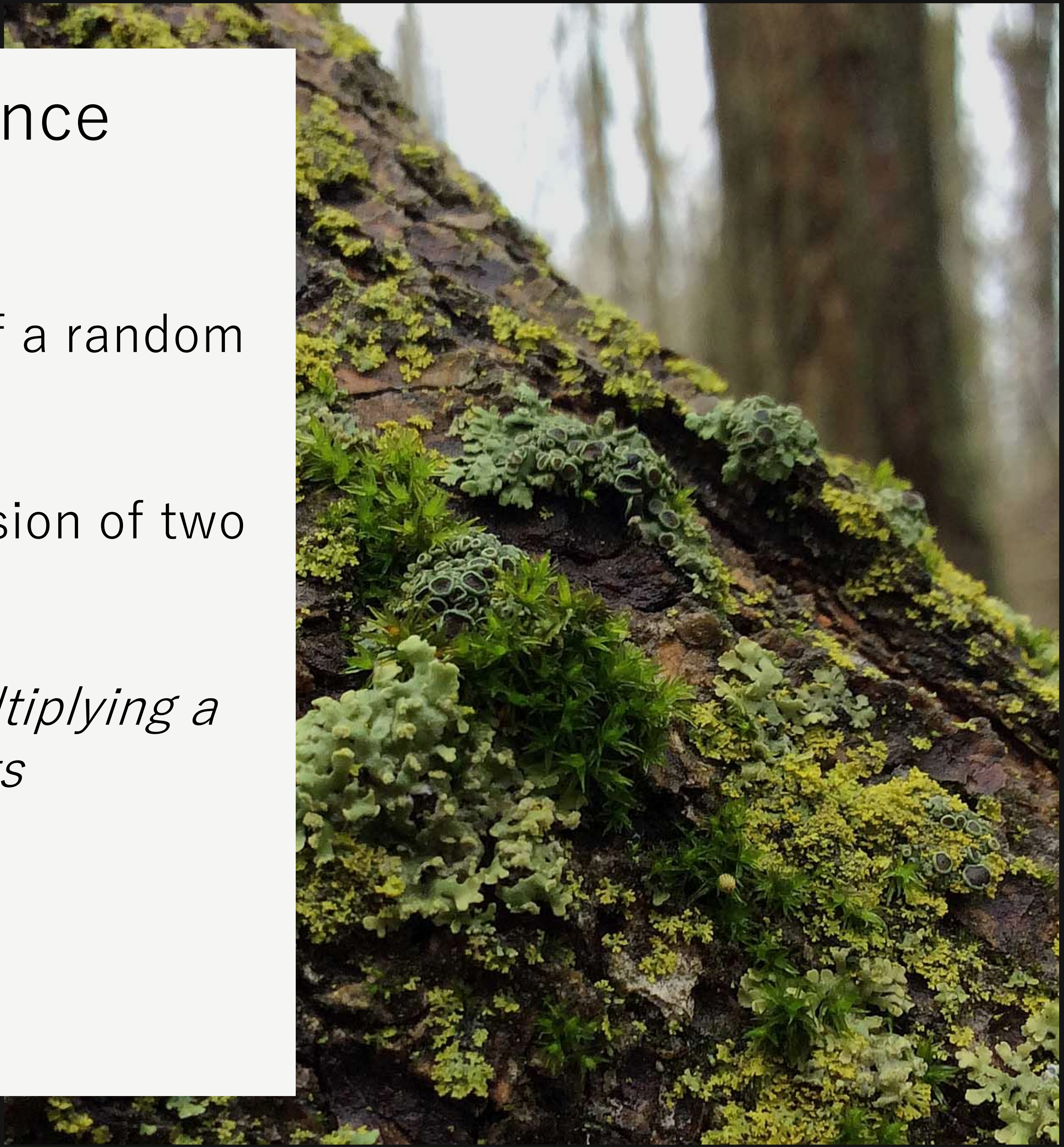
Dispersion Matrices: Covariance Matrix

Variance is a measure of dispersion of a random variable around its mean.

Covariance measures the joint dispersion of two random variables around their means.

A covariance matrix is obtained by multiplying a matrix of column-centered data with its transpose.

$$\mathrm{cov}(\mathbf{Y}) = \mathbf{S} = \frac{1}{n-1}[y - \bar{y}]'[y - \bar{y}]$$

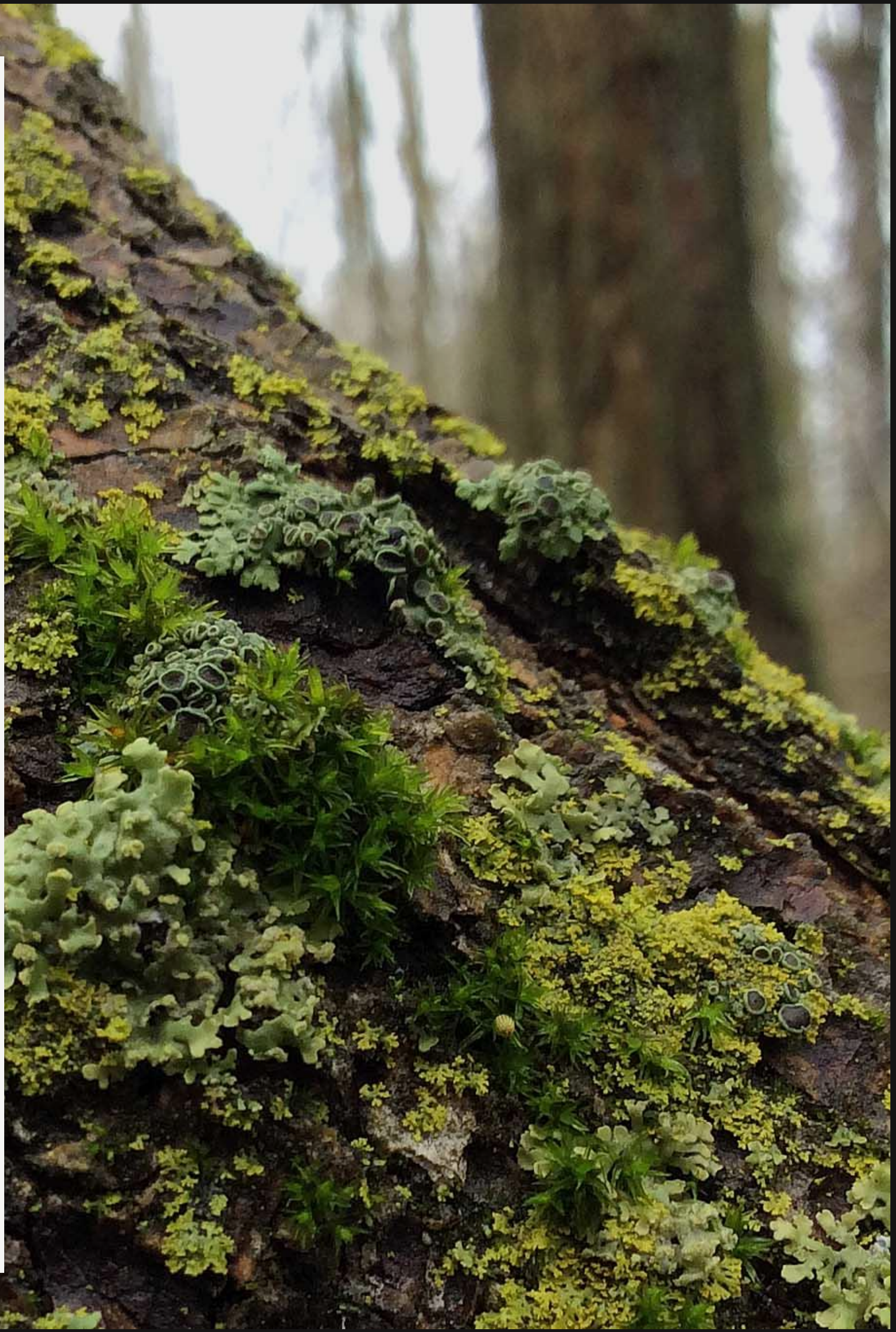


Dispersion Matrices: Covariance Matrix

Variance is a measure of dispersion of a random variable around its mean.

Covariance measures the joint dispersion of two random variables around their means.

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$



Dispersion Matrices: Covariance Matrix

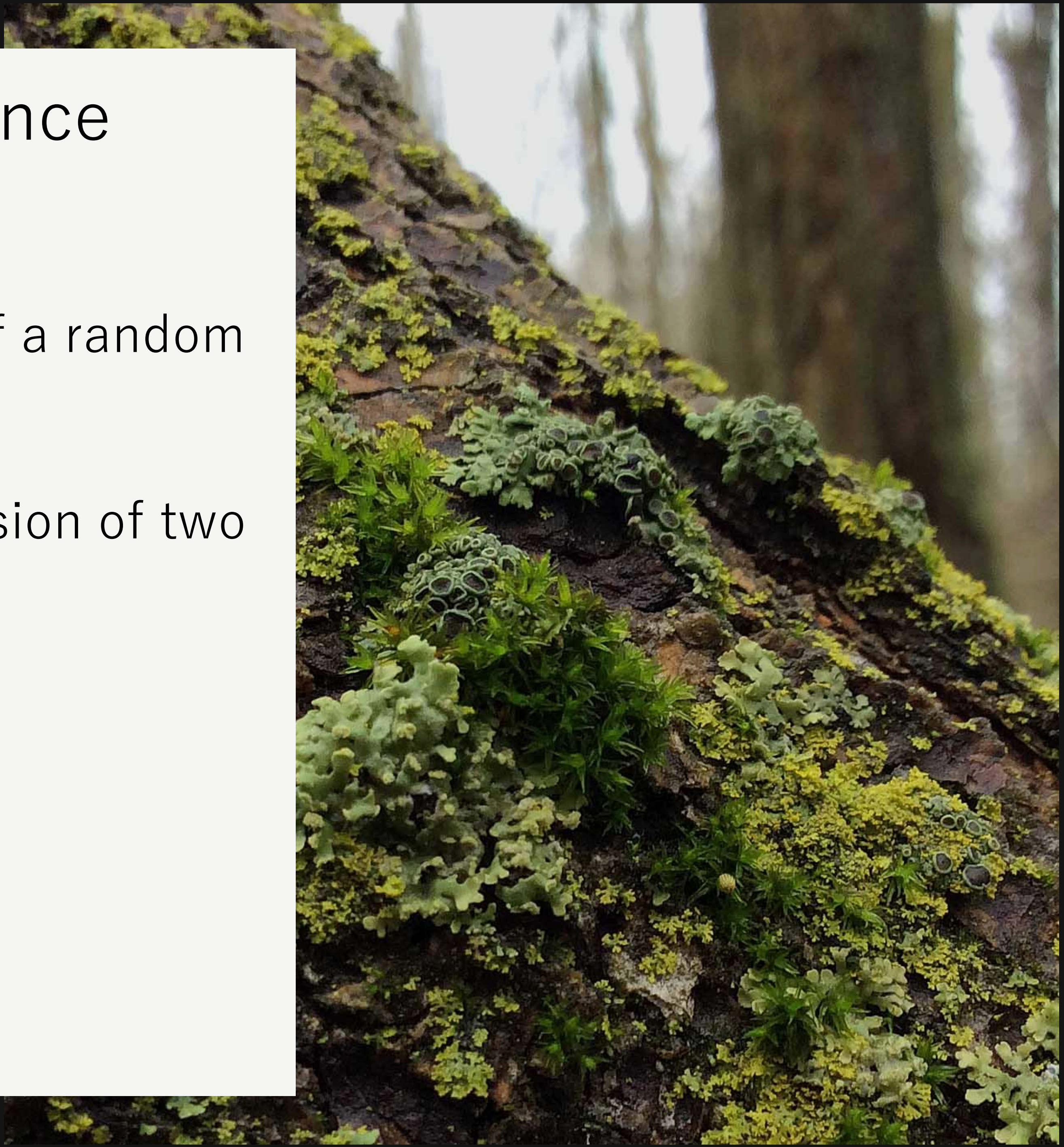
Variance is a measure of dispersion of a random variable around its mean.

Covariance measures the joint dispersion of two random variables around their means.

Variance $\mathbf{y}_1$

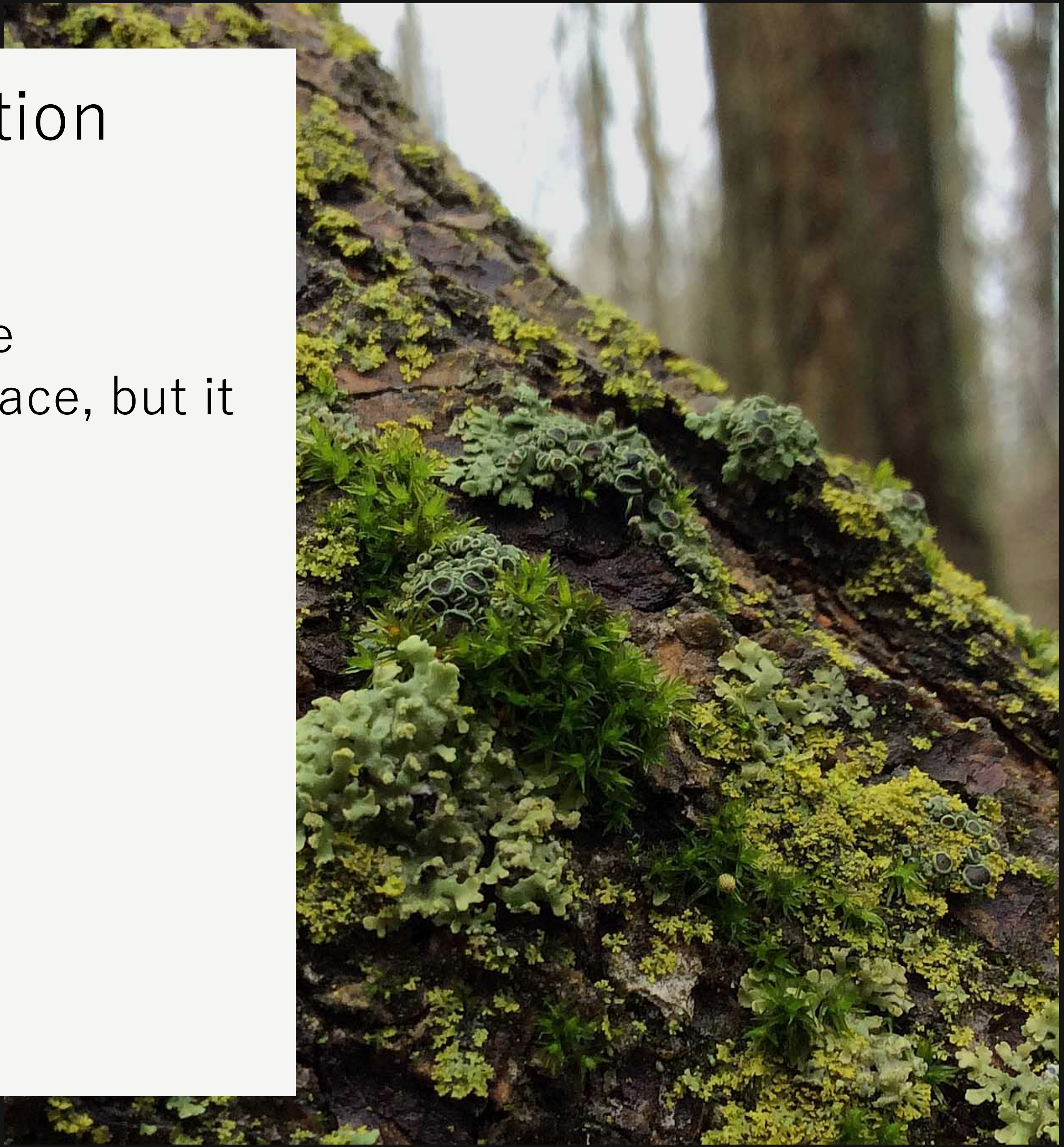
Covariance $\mathbf{y}_1\ \mathbf{y}_2$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$



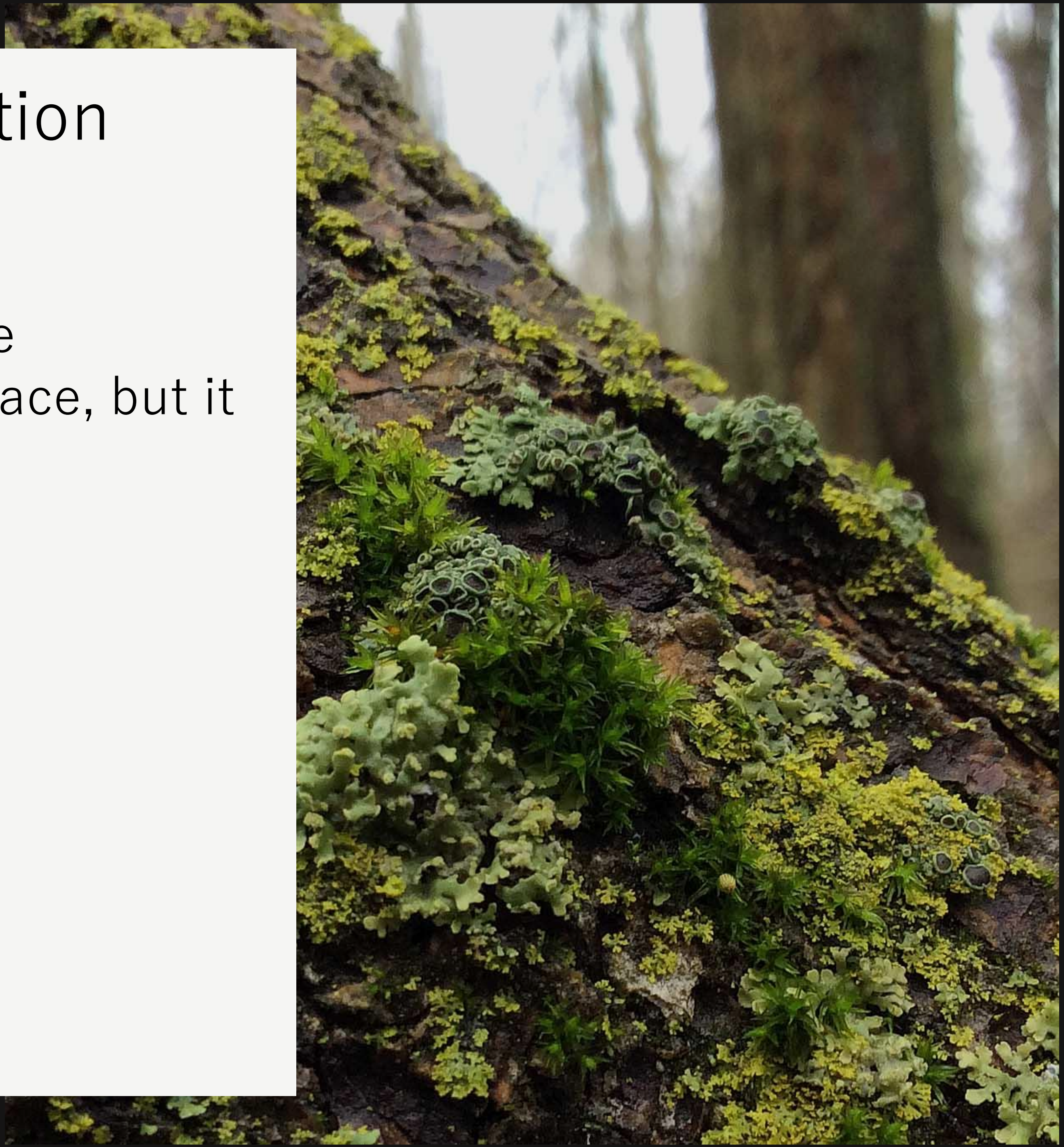
Dispersion Matrices: Correlation Matrix

Covariance provides information on the orientation of the data in descriptor space, but it *does not* quantify the intensity of that relationship.



Dispersion Matrices: Correlation Matrix

Covariance provides information on the orientation of the data in descriptor space, but it *does not* quantify the intensity of that relationship.

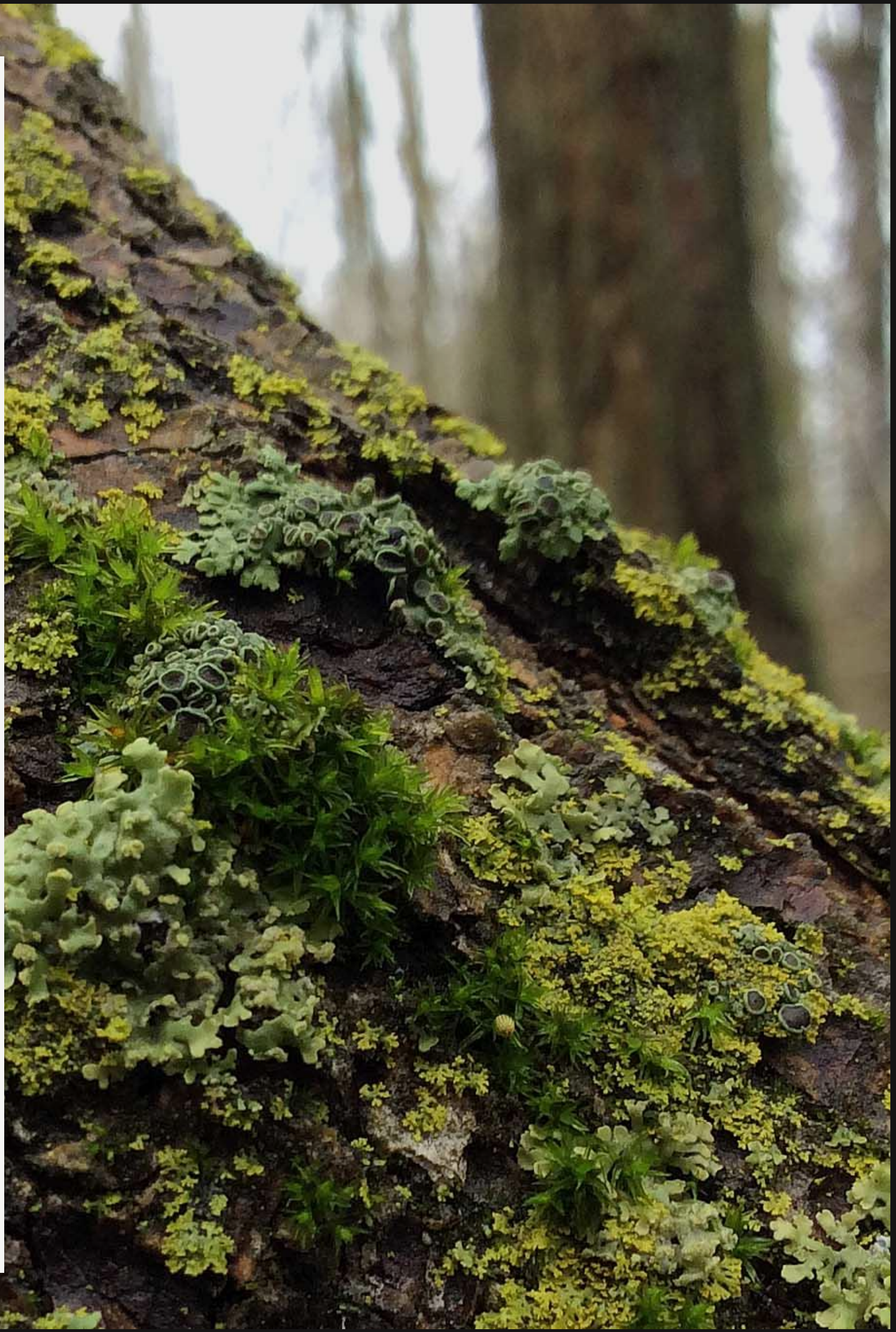


Dispersion Matrices: Correlation Matrix

Correlation is the measure of <u>dependence</u> between two variables.

A correlation matrix is the dispersion matrix of the standardized variables.

$$\mathrm{cor}(\mathbf{Y}) = \mathbf{R} = \frac{1}{\mathrm{n}-1}\left[\frac{\mathrm{y}-\bar{\mathrm{y}}}{s_y}\right]'\left[\frac{\mathrm{y}-\bar{\mathrm{y}}}{s_y}\right]$$

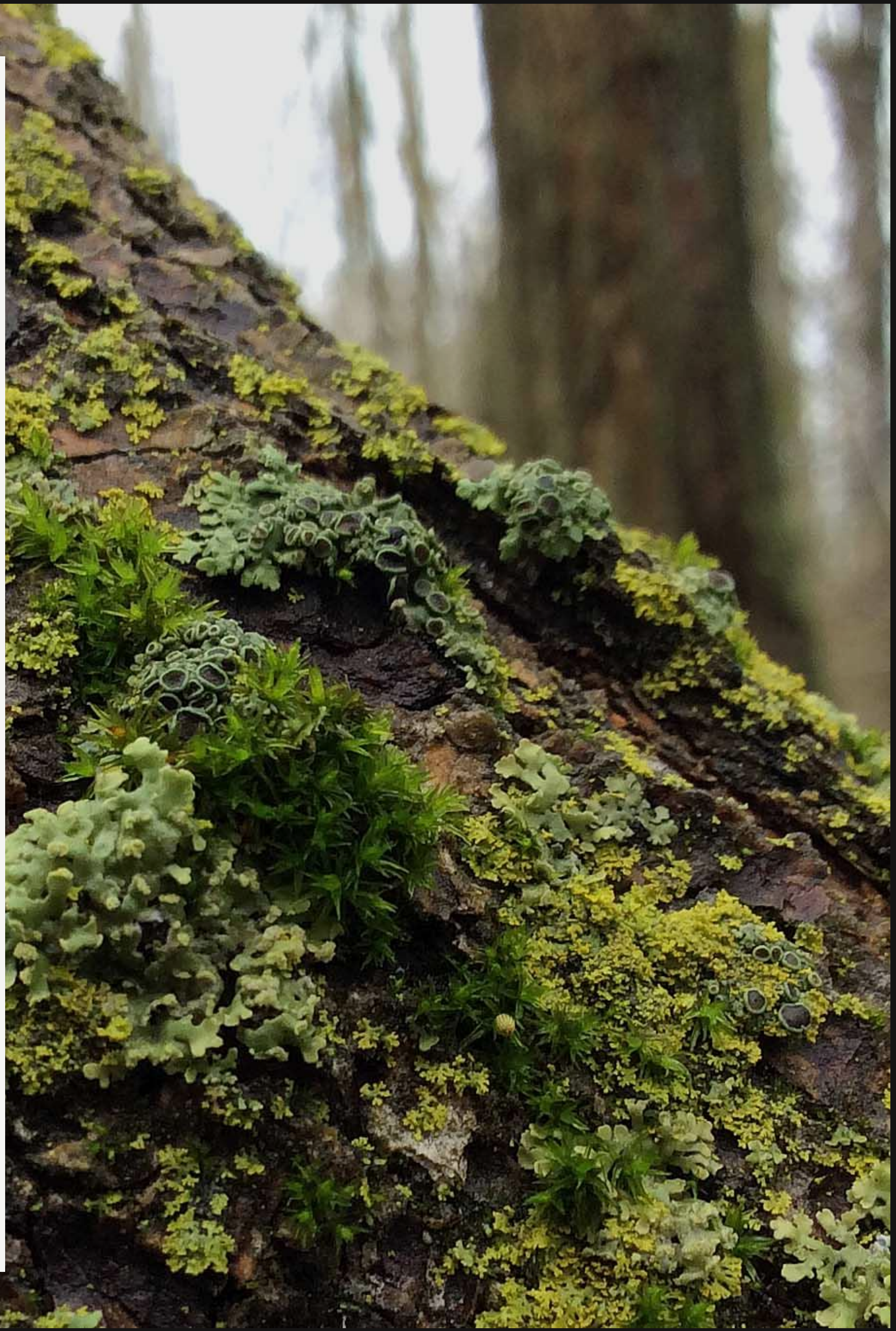


Dispersion Matrices: Correlation Matrix

Correlation is the measure of <u>dependence</u> between two variables.

A correlation matrix is the dispersion matrix of the standardized variables.

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$



Dispersion Matrices: Correlation Matrix

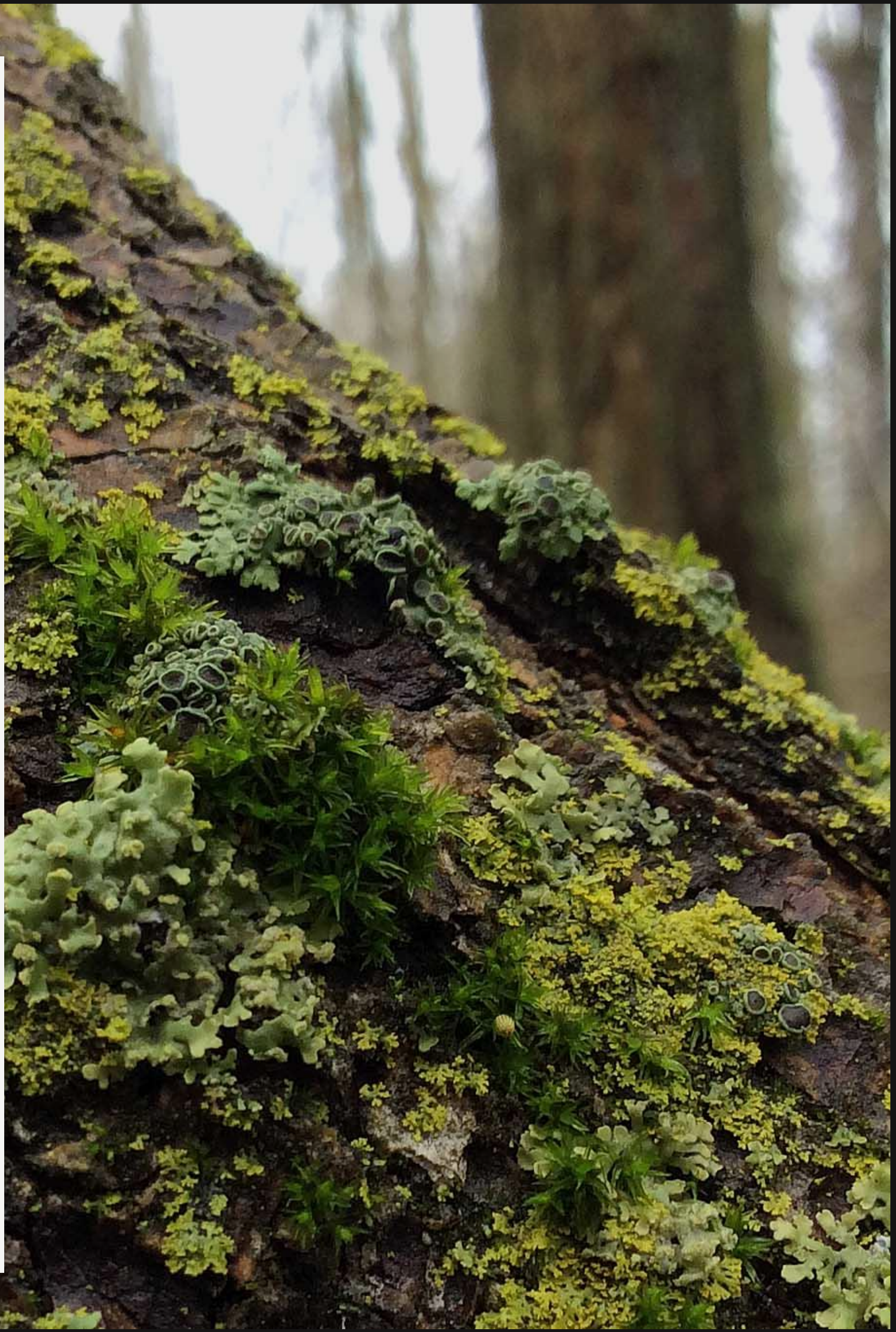
Correlation is the measure of dependence between two variables.

A correlation matrix is the dispersion matrix of the standardized variables.

Correlation ranges from -1 to 1, where 0 indicates linear independence

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

The comparison of any descriptor with itself is complete dependence

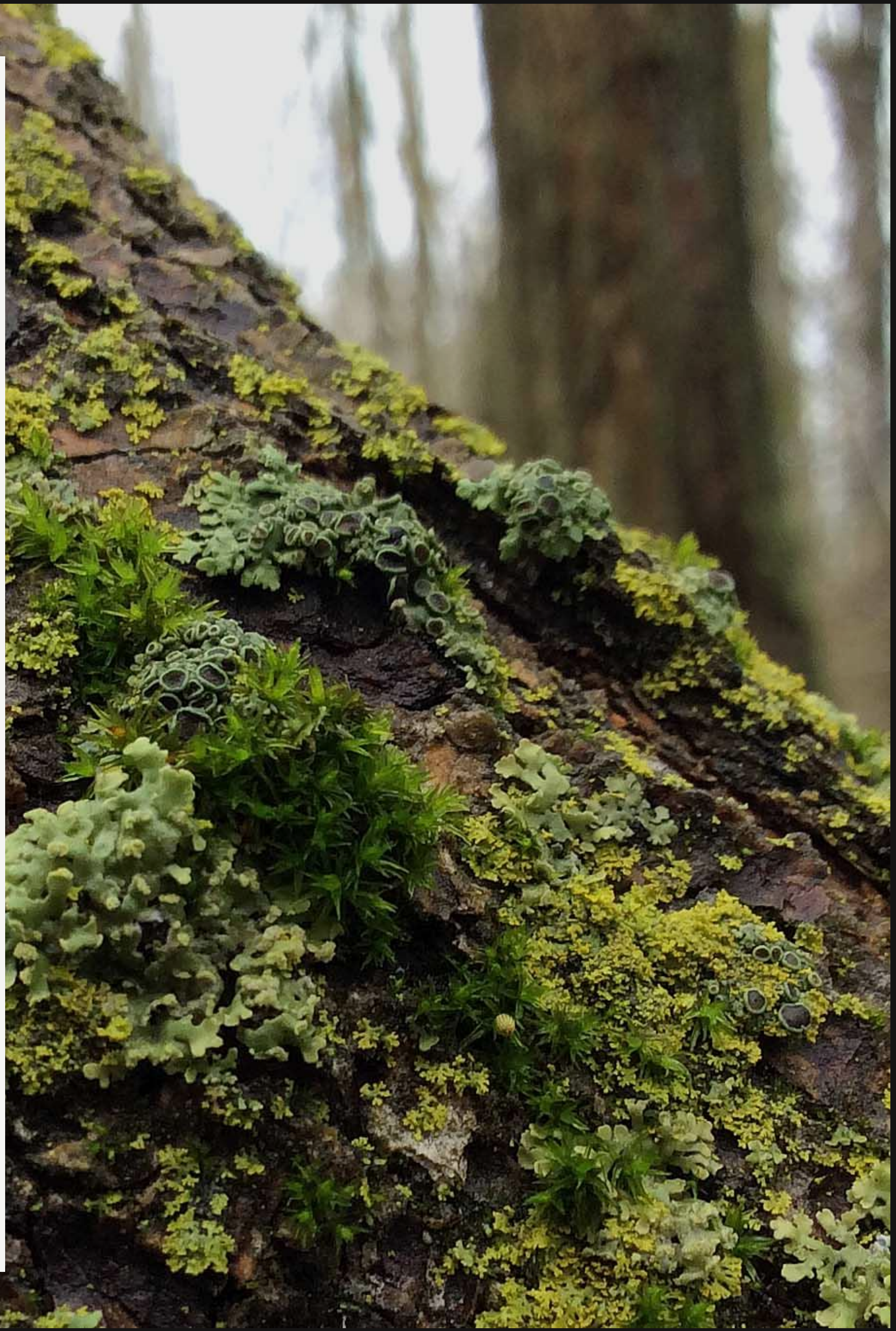


Dispersion Matrices: Covariance or Correlation?

Covariance:

- Data on the same scale
- The magnitude of the data matters

Variables with larger variances will dominate the first few PCs.



Dispersion Matrices: Covariance or Correlation?

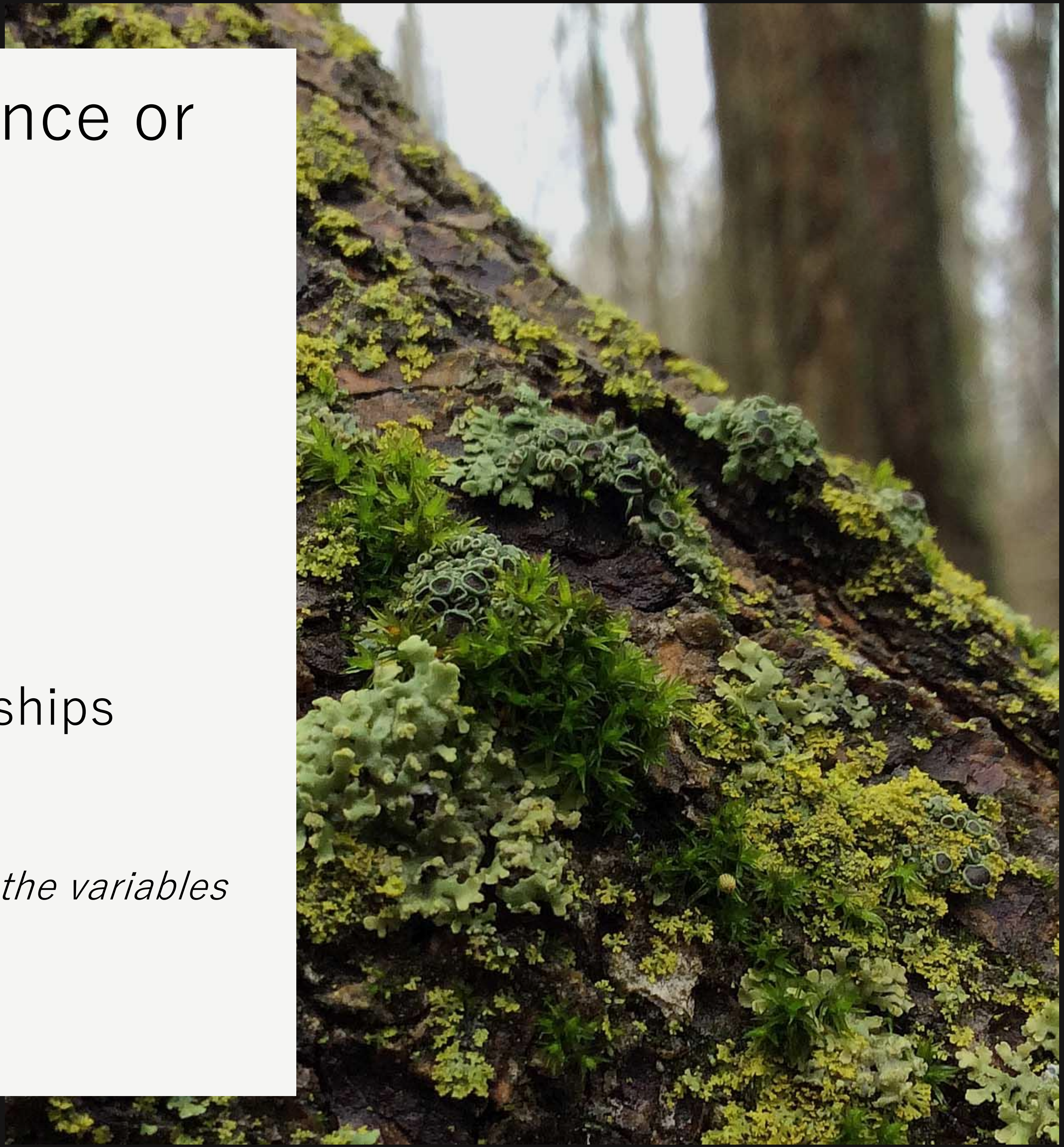
Covariance:

- Data on the same scale
- The magnitude of the data matters

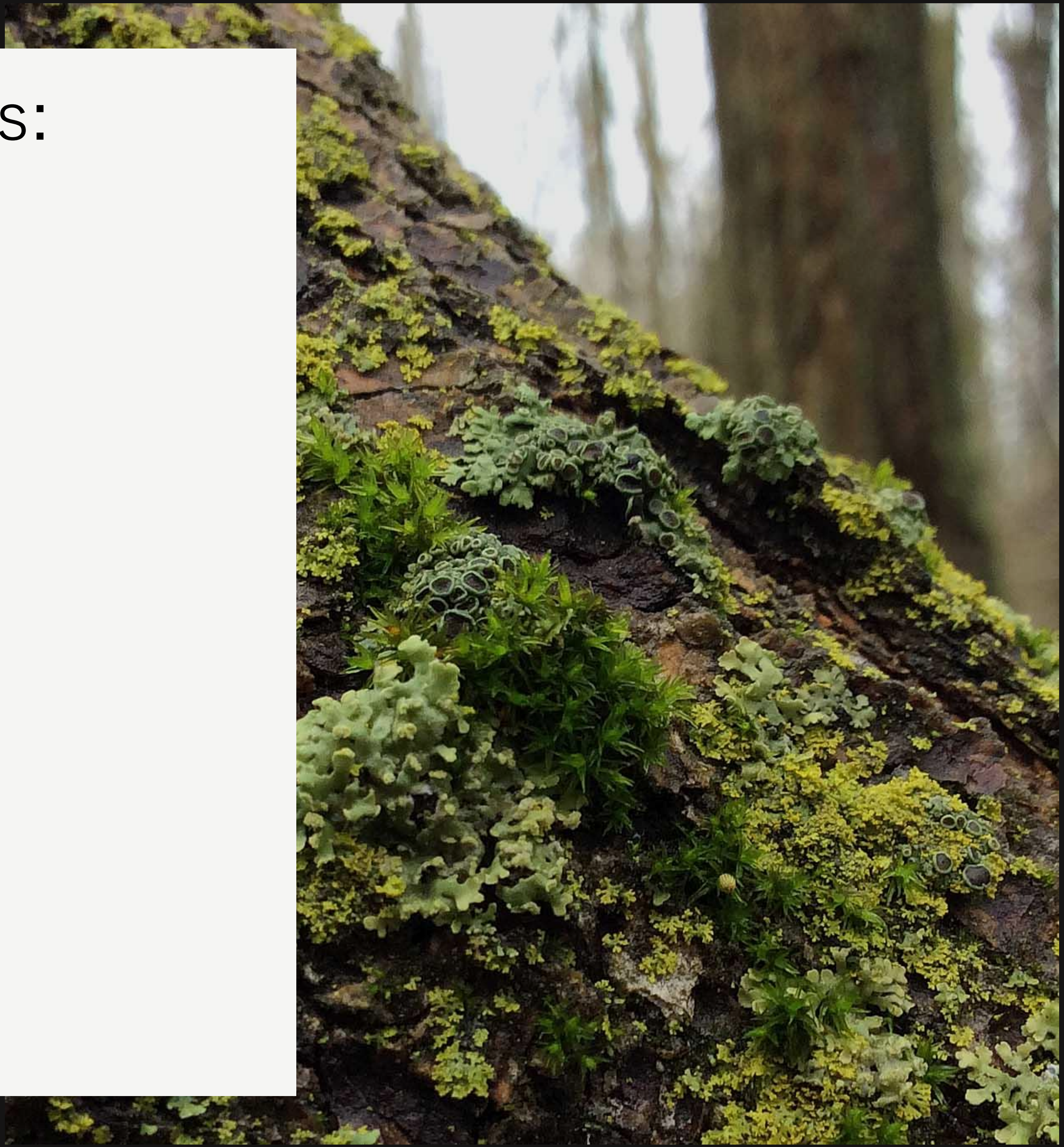
Correlation:

- Variables are on different scales
- Interest is in understanding relationships regardless of scale

The analysis focuses on the relationships between the variables rather than their absolute magnitudes.

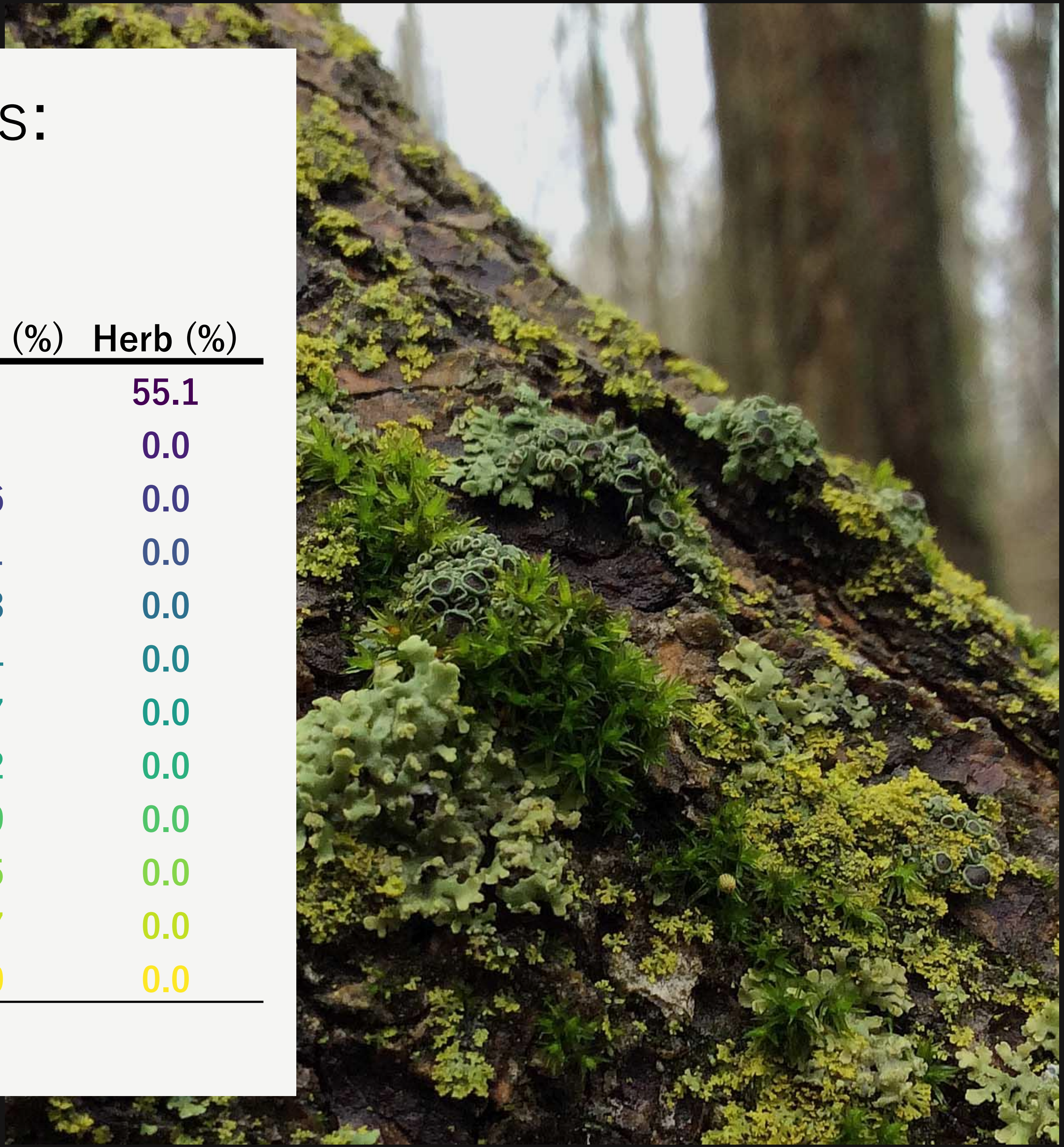


Principal Component Analysis: Process and Steps



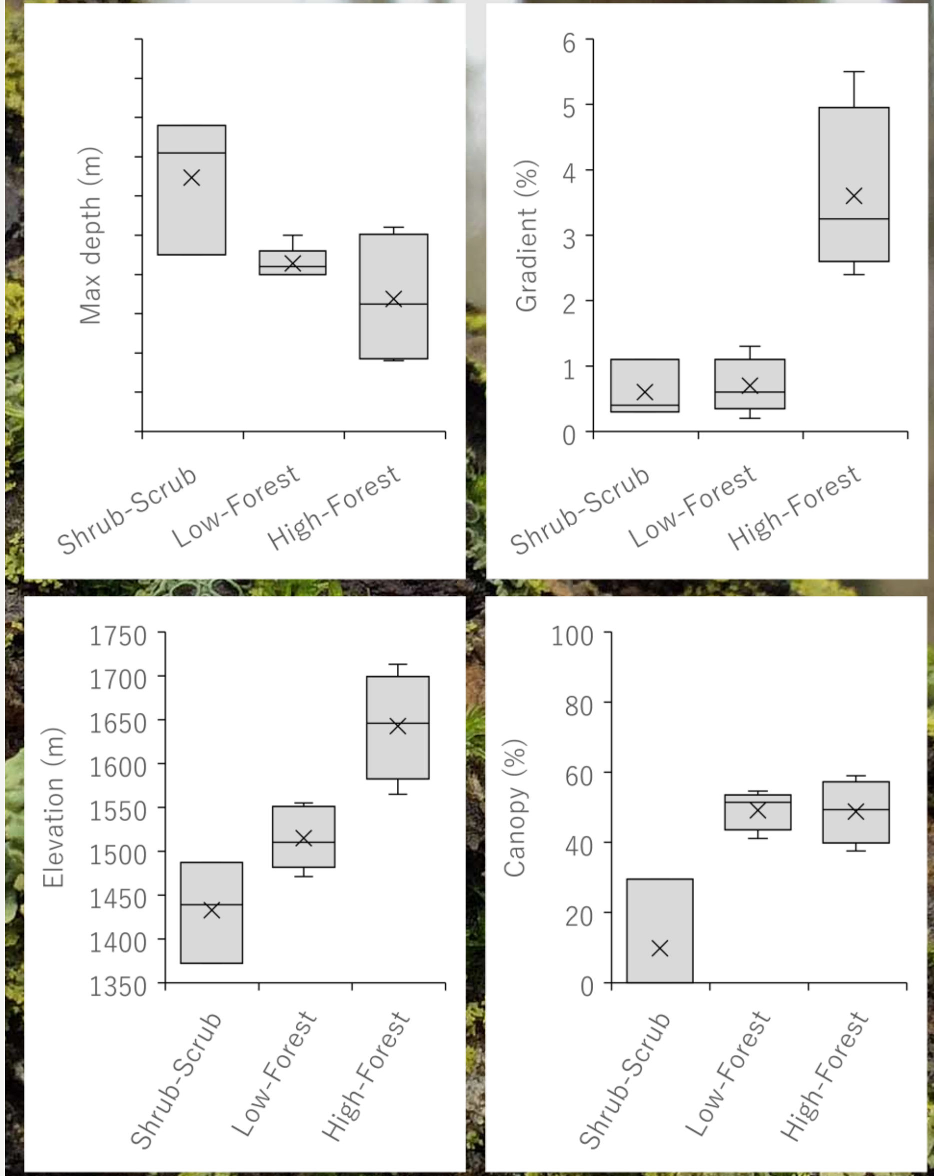
Principal Component Analysis: Process and Steps

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



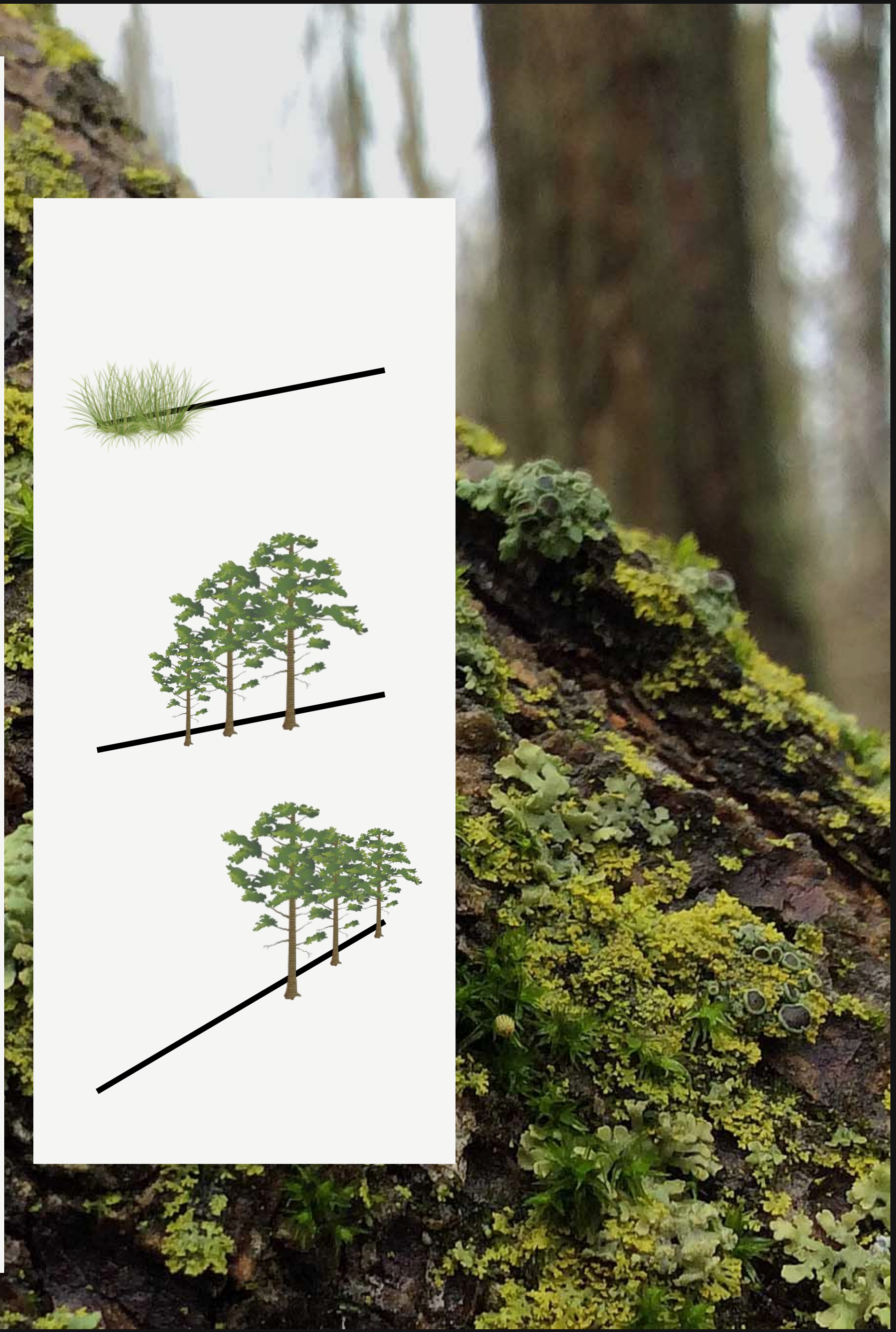
Principal Component Analysis: Process and Steps

Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



Principal Component Analysis: Process and Steps

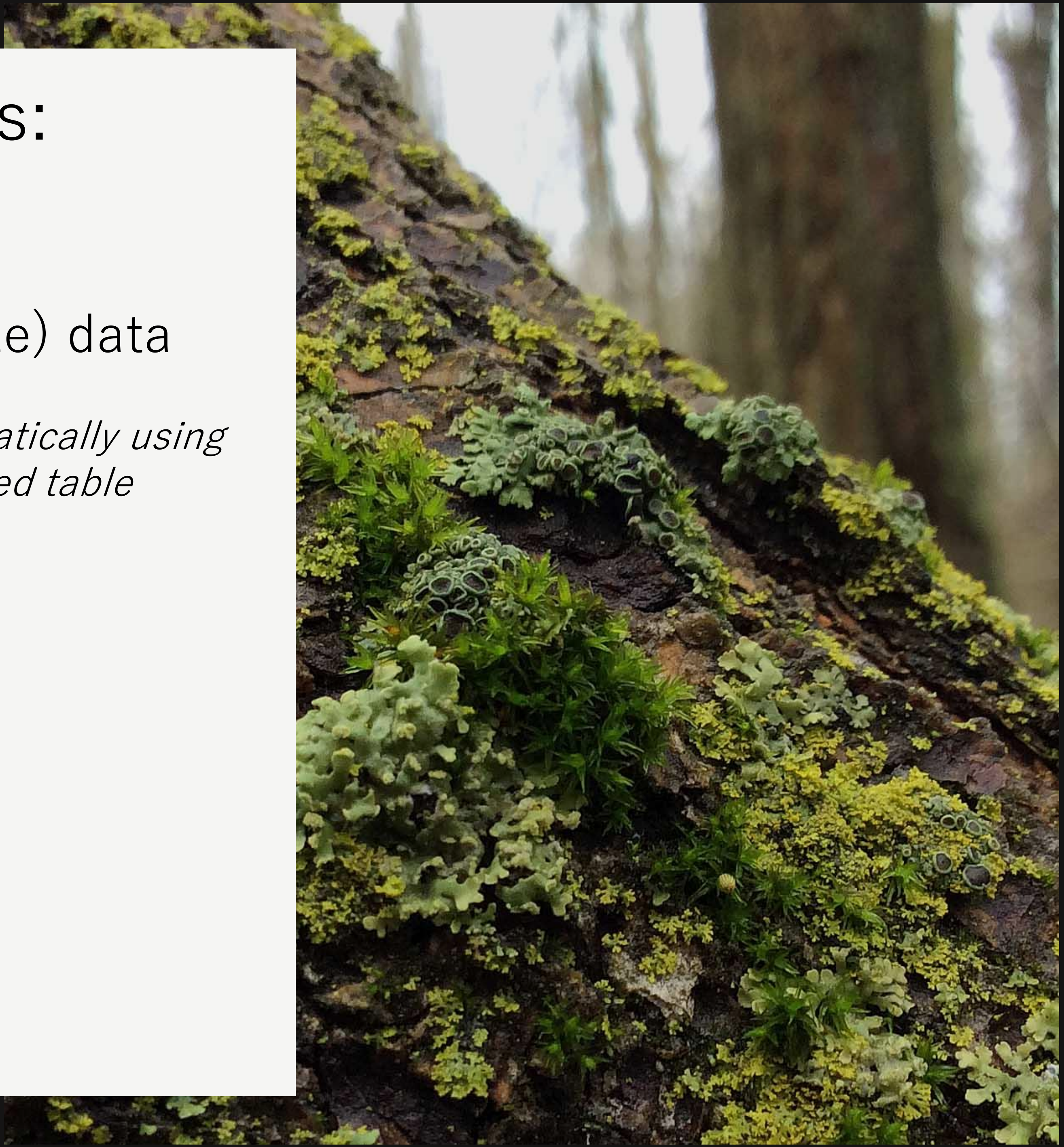
Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	0.45	0.3	1439	0.0	55.1
Silvies-34	0.78	1.1	1487	0.0	0.0
Silvies-02	0.71	0.4	1372	29.6	0.0
Silvies-15	0.40	0.2	1471	41.1	0.0
Silvies-07	0.50	1.3	1547	52.3	0.0
Silvies-08	0.40	0.6	1492	51.4	0.0
Silvies-22	0.42	0.9	1555	54.7	0.0
Silvies-18	0.42	0.5	1510	46.2	0.0
Silvies-12	0.52	3.2	1658	51.9	0.0
Silvies-21	0.18	2.4	1713	37.5	0.0
Silvies-05	0.45	5.5	1565	46.7	0.0
Silvies-03	0.20	3.3	1634	59.0	0.0



Principal Component Analysis: Process and Steps

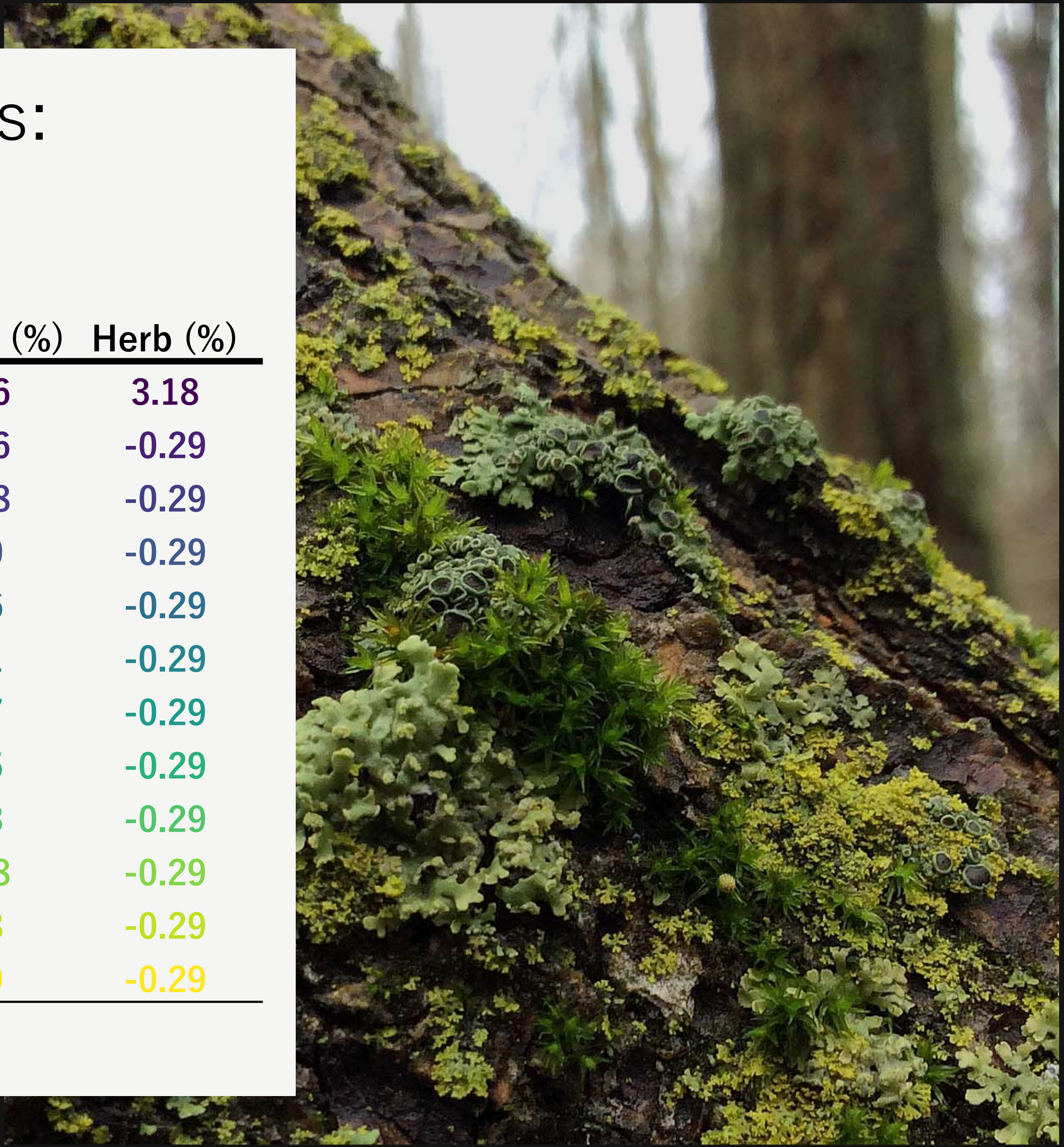
Step 1) Column center (and standardize) data

Note: The ‘prcomp()’ function in R does this automatically using the ‘scale’ prompt, but we will show the standardized table anyway for clarity.



Principal Component Analysis: Process and Steps

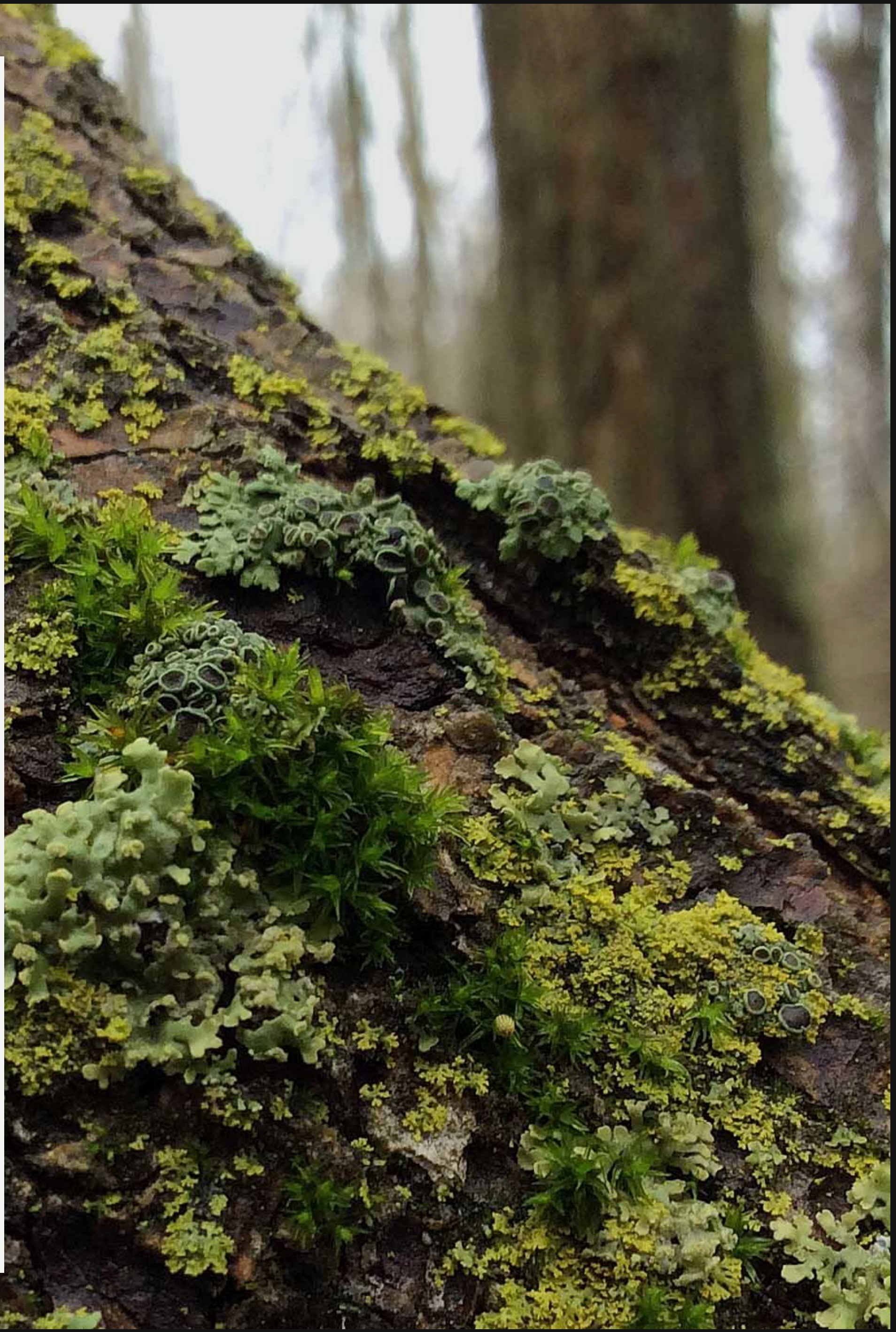
Site ID	Max Depth (m)	Gradient (%)	Elevation (m)	Canopy (%)	Herb (%)
Silvies-11	-0.01	-0.82	-1.01	-1.96	3.18
Silvies-34	1.90	-0.33	-0.52	-1.96	-0.29
Silvies-02	1.50	-0.76	-1.71	-0.48	-0.29
Silvies-15	-0.30	-0.88	-0.68	0.10	-0.29
Silvies-07	0.28	-0.21	0.10	0.66	-0.29
Silvies-08	-0.30	-0.64	-0.46	0.61	-0.29
Silvies-22	-0.19	-0.45	0.19	0.77	-0.29
Silvies-18	-0.19	-0.70	-0.28	0.35	-0.29
Silvies-12	0.39	0.95	1.25	0.63	-0.29
Silvies-21	-1.58	0.46	1.82	-0.08	-0.29
Silvies-05	-0.01	2.36	0.29	0.38	-0.29
Silvies-03	-1.47	1.01	1.00	0.99	-0.29



Principal Component Analysis: Process and Steps

Step 2) Generate covariance or correlation matrix, **S**, (i.e., the dispersion matrix of descriptors)

Site ID	**Max Depth (m)**	**Gradient (%)**	**Elevation (m)**	**Canopy (%)**	**Herb (%)**
Depth	**1.00**	**-0.27**	**-0.64**	**-0.52**	**0.00**
Gradient	**-0.27**	**1.00**	**0.64**	**0.35**	**-0.26**
Elevation	**-0.64**	**0.64**	**1.00**	**0.49**	**-0.32**
Canopy	**-0.52**	**0.35**	**0.49**	**1.00**	**-0.62**
Herb	**0.00**	**-0.26**	**-0.32**	**-0.62**	**1.00**



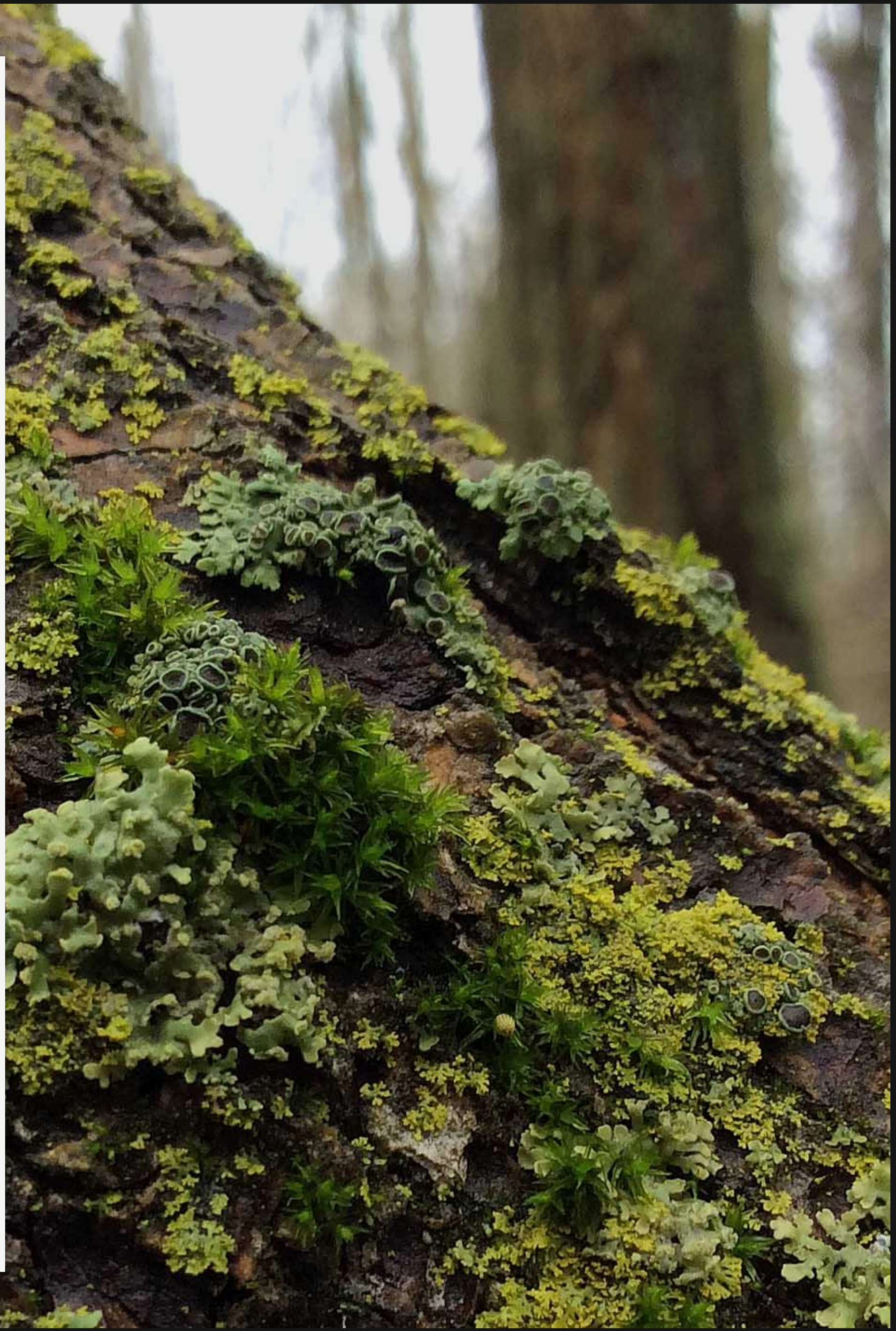
Principal Component Analysis: Process and Steps

Step 3) Solve the characteristic equation:

$$|\mathbf{S} - \lambda_k \mathbf{I}| = 0$$

to find the eigenvalues

$$\left|\begin{bmatrix} 1.00-\lambda_1 & -0.27 & -0.64 & -0.52 & 0.00 \\ -0.27 & 1.00-\lambda_2 & 0.64 & 0.35 & -0.26 \\ -0.64 & 0.64 & 1.00-\lambda_3 & 0.49 & -0.32 \\ -0.52 & 0.35 & 0.49 & 1.00-\lambda_4 & -0.62 \\ 0.00 & -0.26 & -0.32 & -0.62 & 1.00-\lambda_5 \end{bmatrix}\right| = 0$$



Principal Component Analysis: Process and Steps

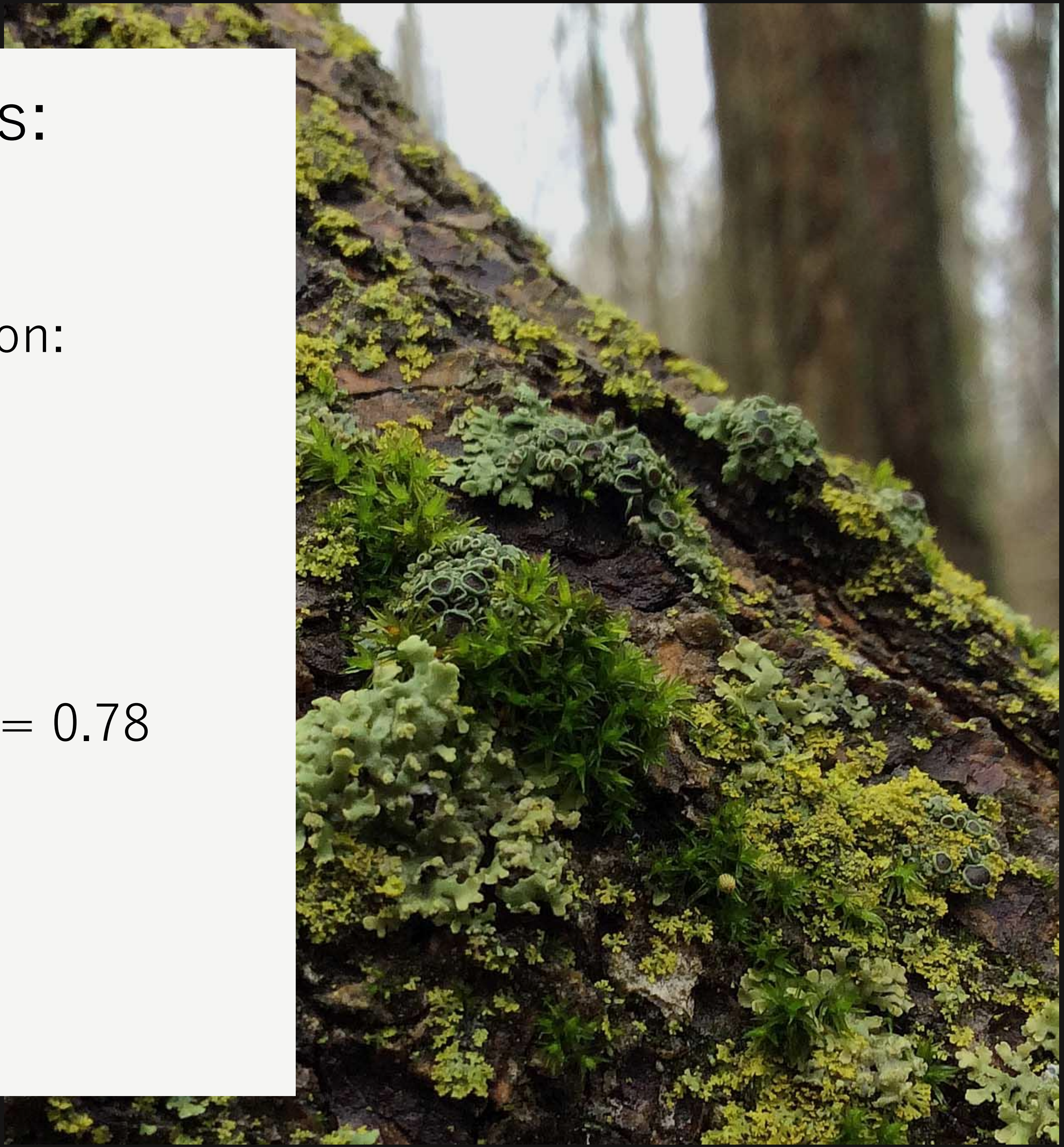
Step 3) Solve the characteristic equation:

$$|\mathbf{S} - \lambda_k \mathbf{I}| = 0$$

to find the eigenvalues

$$\lambda_1 = 2.69 \qquad \lambda_2 = 1.09 \qquad \lambda_3 = 0.78$$

$$\lambda_4 = 0.31 \qquad \lambda_5 = 0.12$$

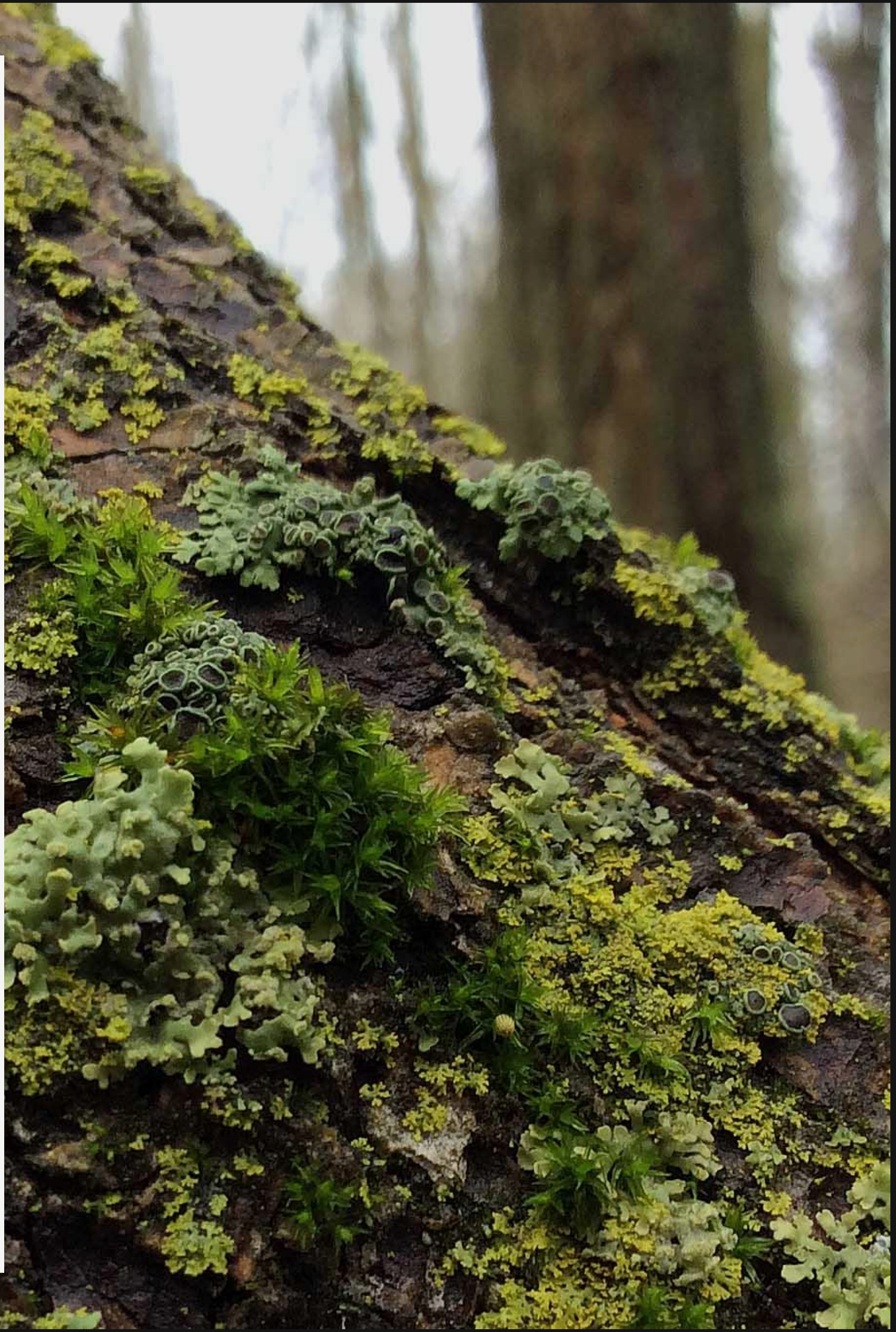


Principal Component Analysis: Process and Steps

Step 4) Solve for eigenvectors (i.e., **loadings**)

$$(\mathbf{S} - \boldsymbol{\lambda}\,\mathbf{I})\mathbf{u} = 0$$

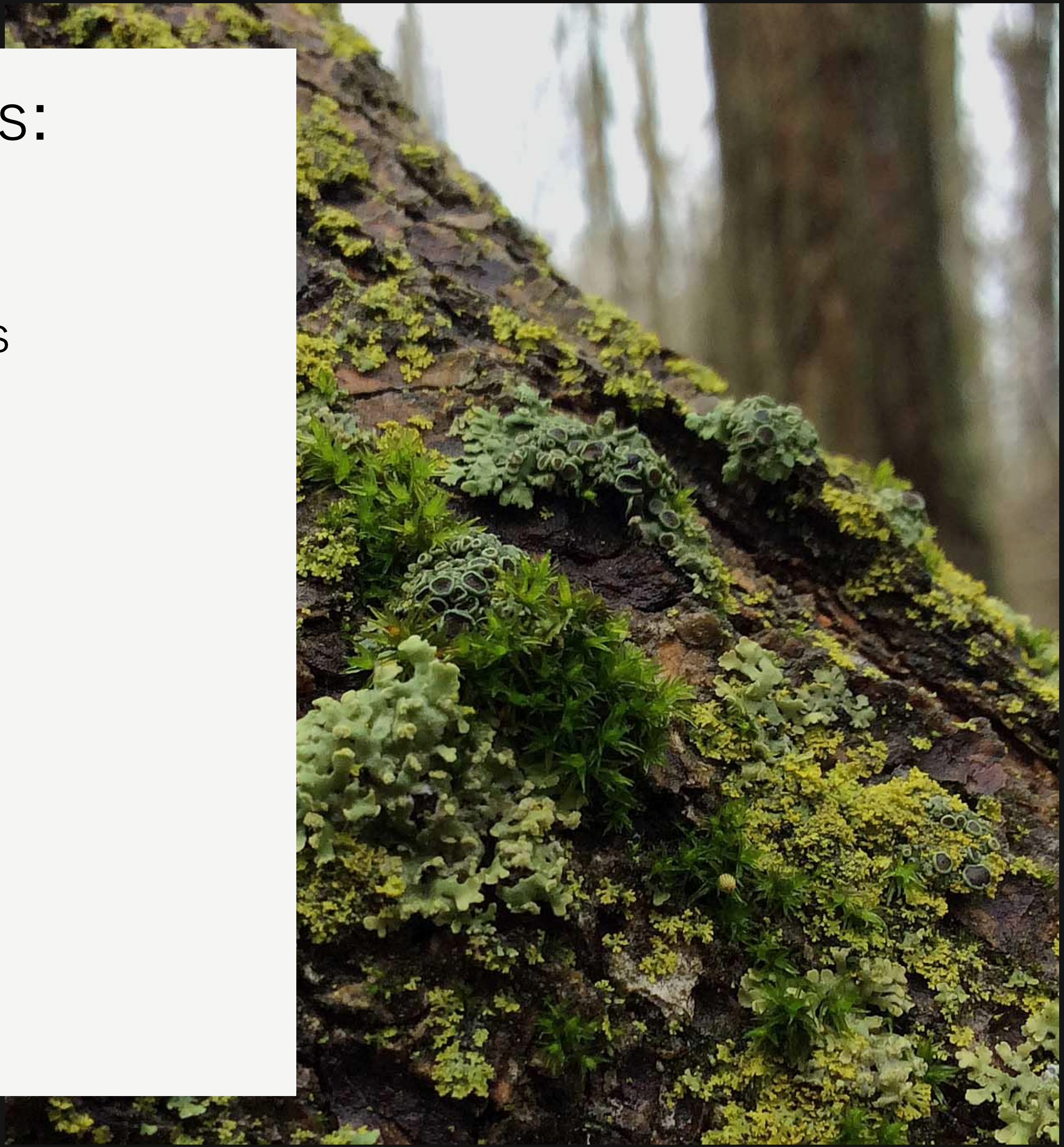
Site ID	PC1	PC2	PC3	PC4	PC5
Max Depth	0.42	-0.52	0.50	0.01	-0.56
Gradient	-0.42	0.12	0.74	-0.45	0.23
Elevation	-0.53	0.26	0.19	0.64	-0.47
Canopy	-0.50	-0.29	-0.41	-0.53	-0.48
Herbaceous	0.35	0.75	0.01	-0.35	-0.44



Principal Component Analysis: Process and Steps

Step 5) Compute principal components

$$\mathbf{F} = \mathbf{Y}_c\mathbf{U}$$

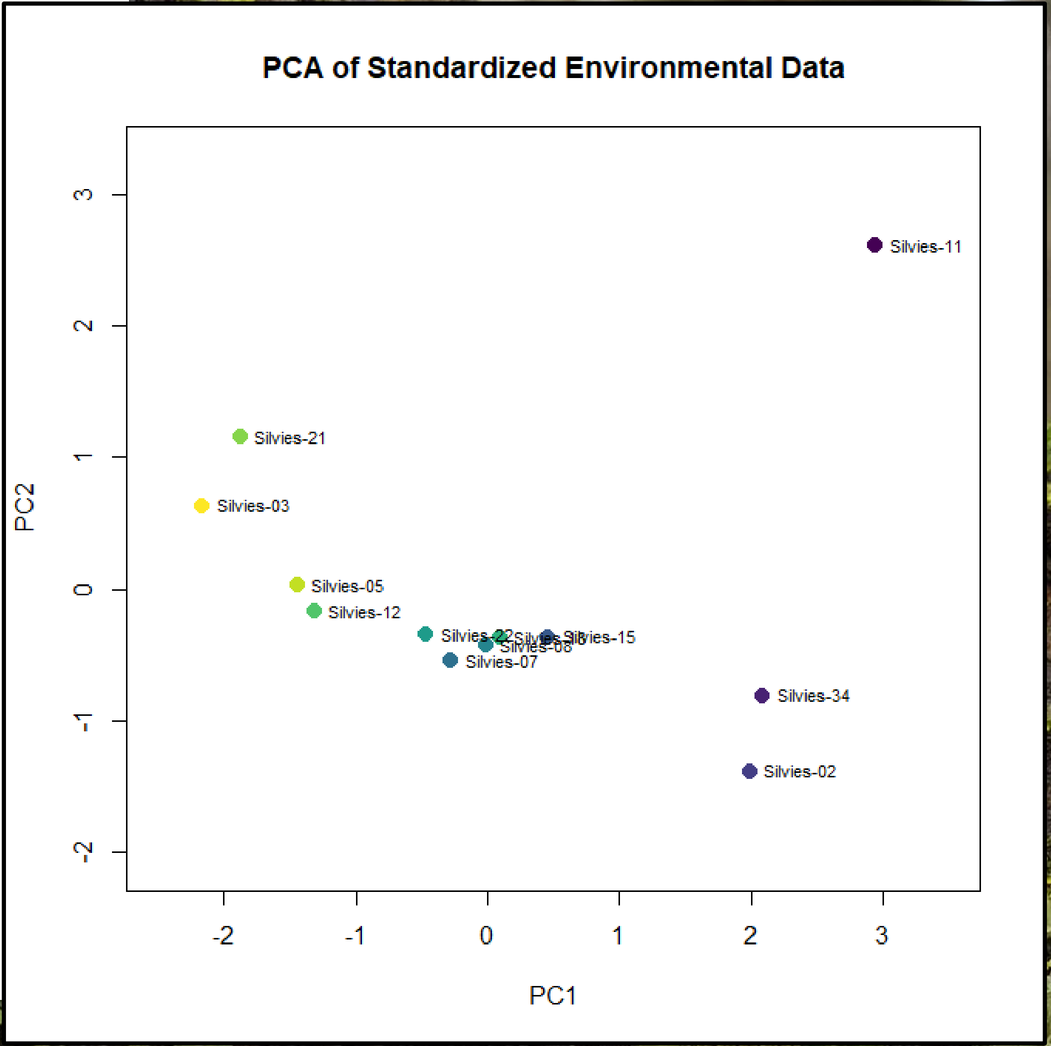


Principal Component Analysis: Process and Steps

Step 5) Compute principal components

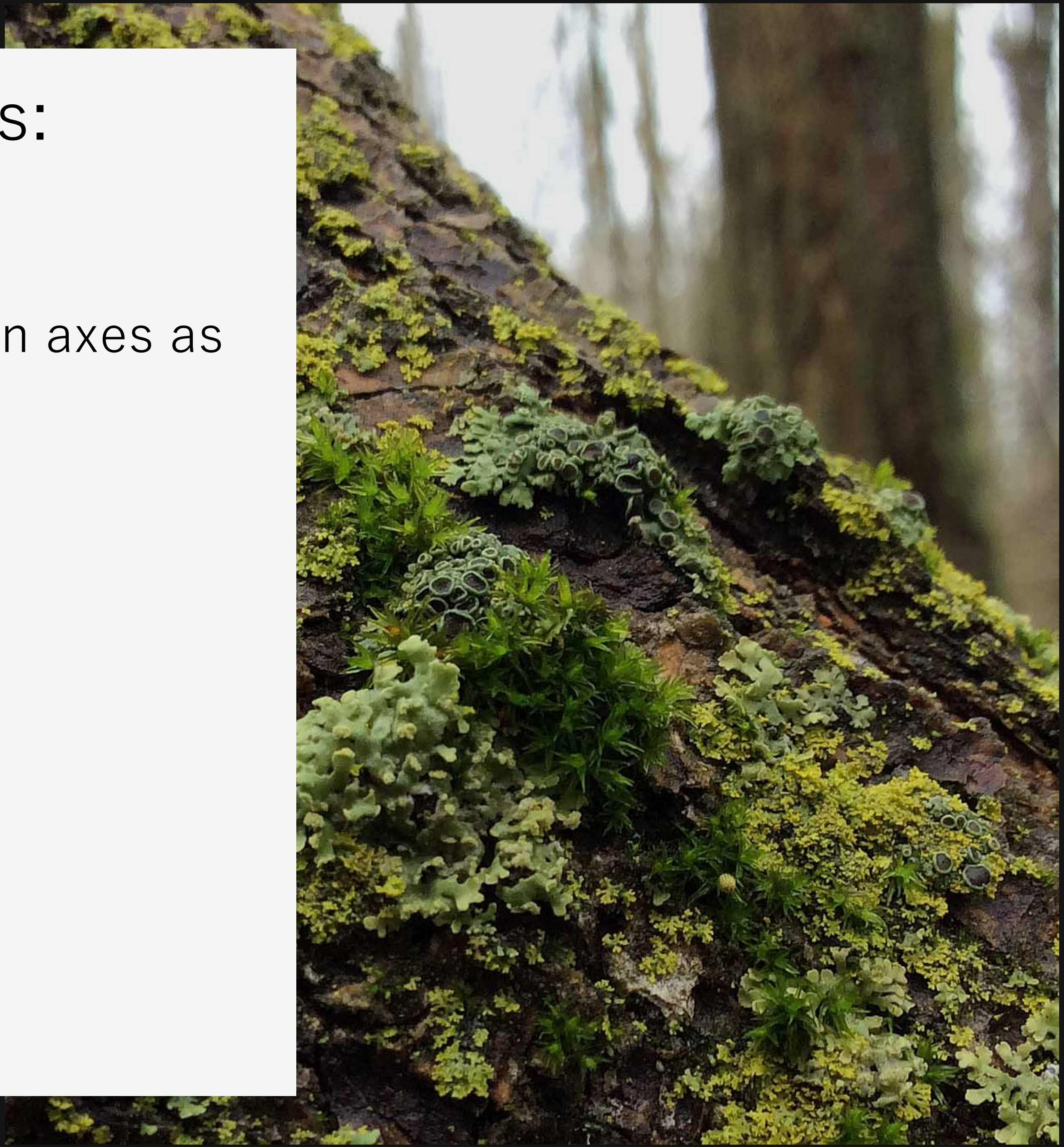
$$\mathbf{F} = \mathbf{Y}_c\mathbf{U}$$

Also referred to as “scores”



Principal Component Analysis: Process and Steps

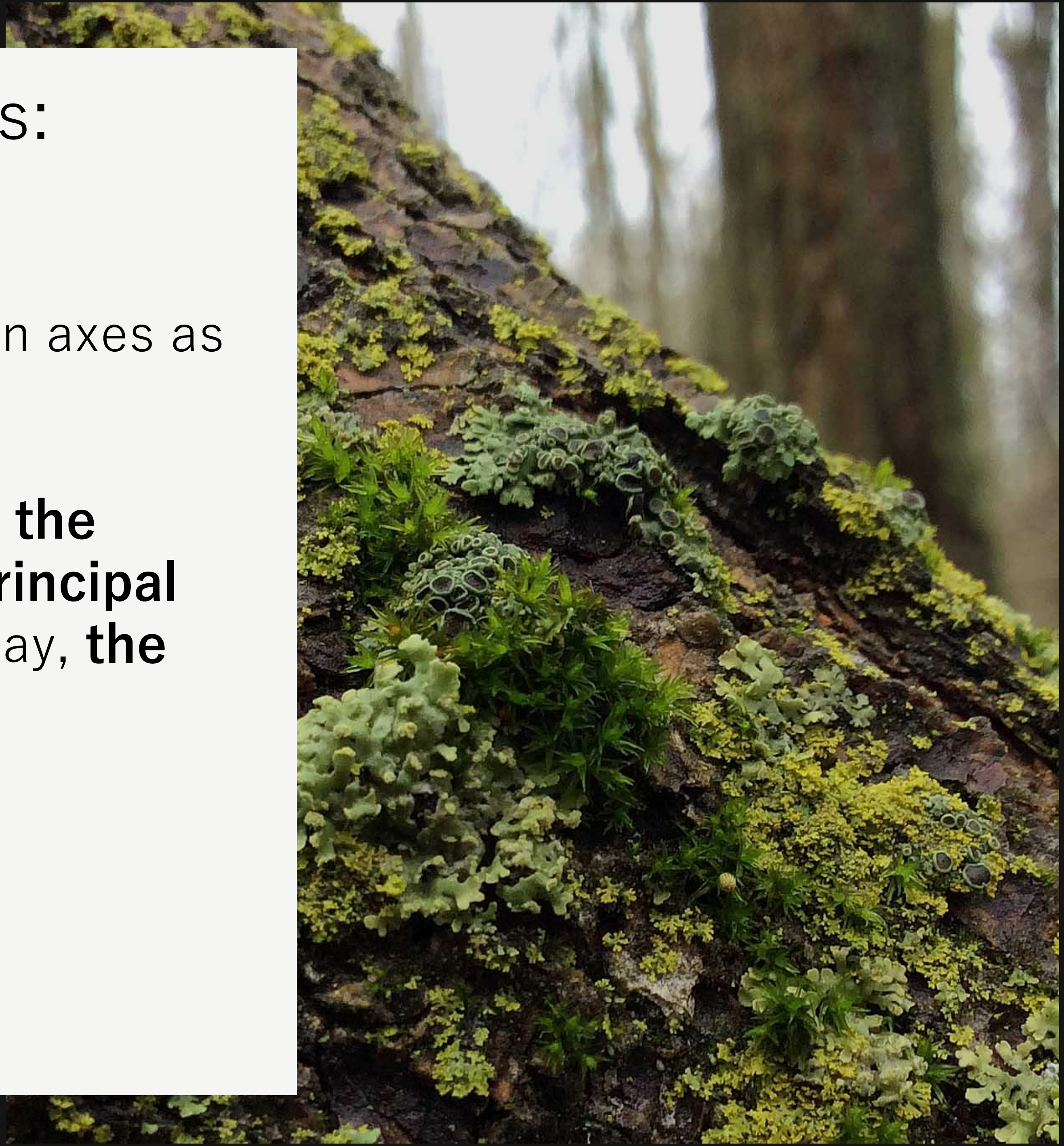
Step 6) Scale loadings of descriptors on axes as appropriate



Principal Component Analysis: Process and Steps

Step 6) Scale loadings of descriptors on axes as appropriate

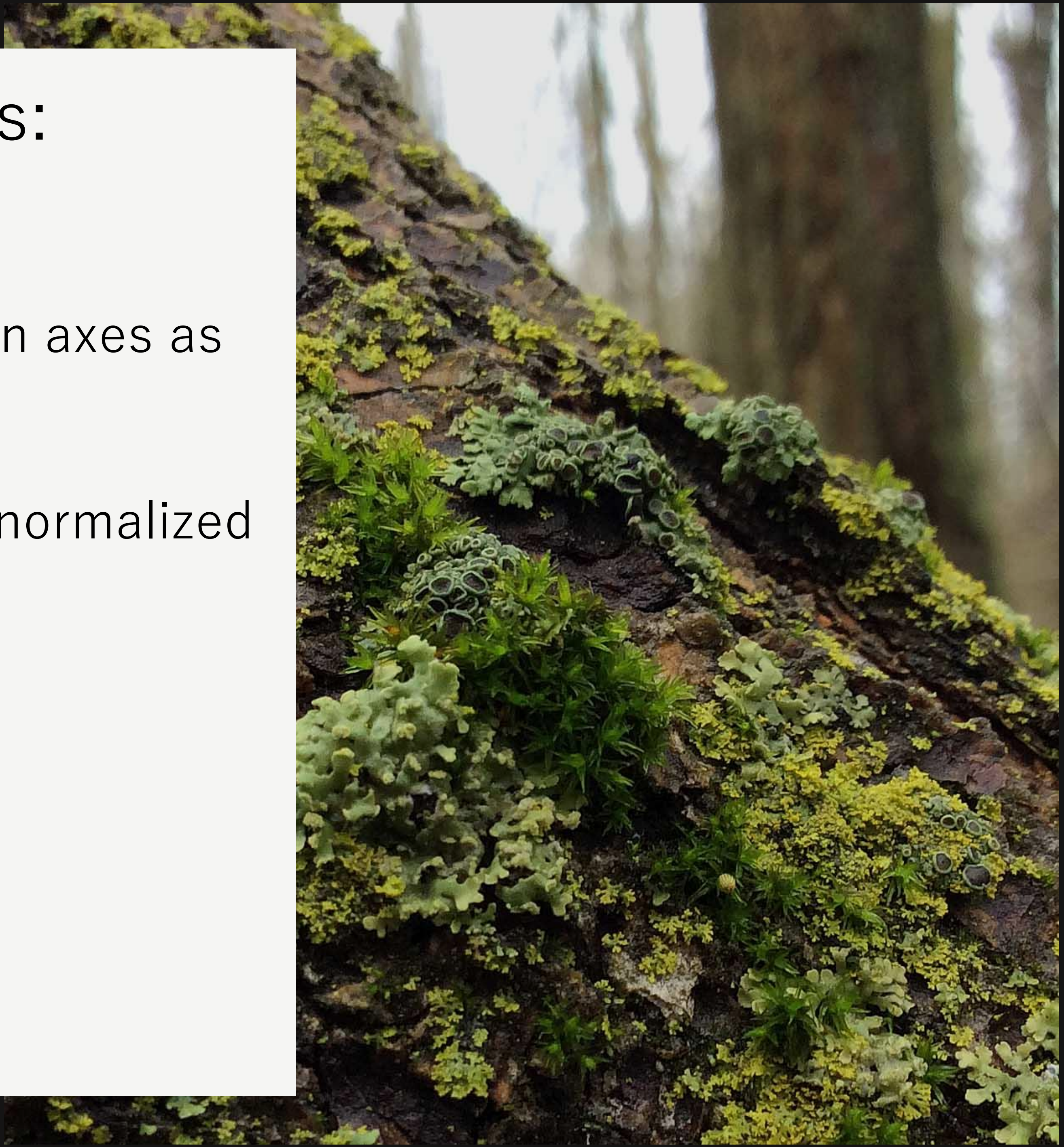
Provides information about **the role of the descriptors in the formation of the principal components** and, if scaled a certain way, **the relationships among the descriptors themselves**.



Principal Component Analysis: Process and Steps

Step 6) Scale loadings of descriptors on axes as appropriate

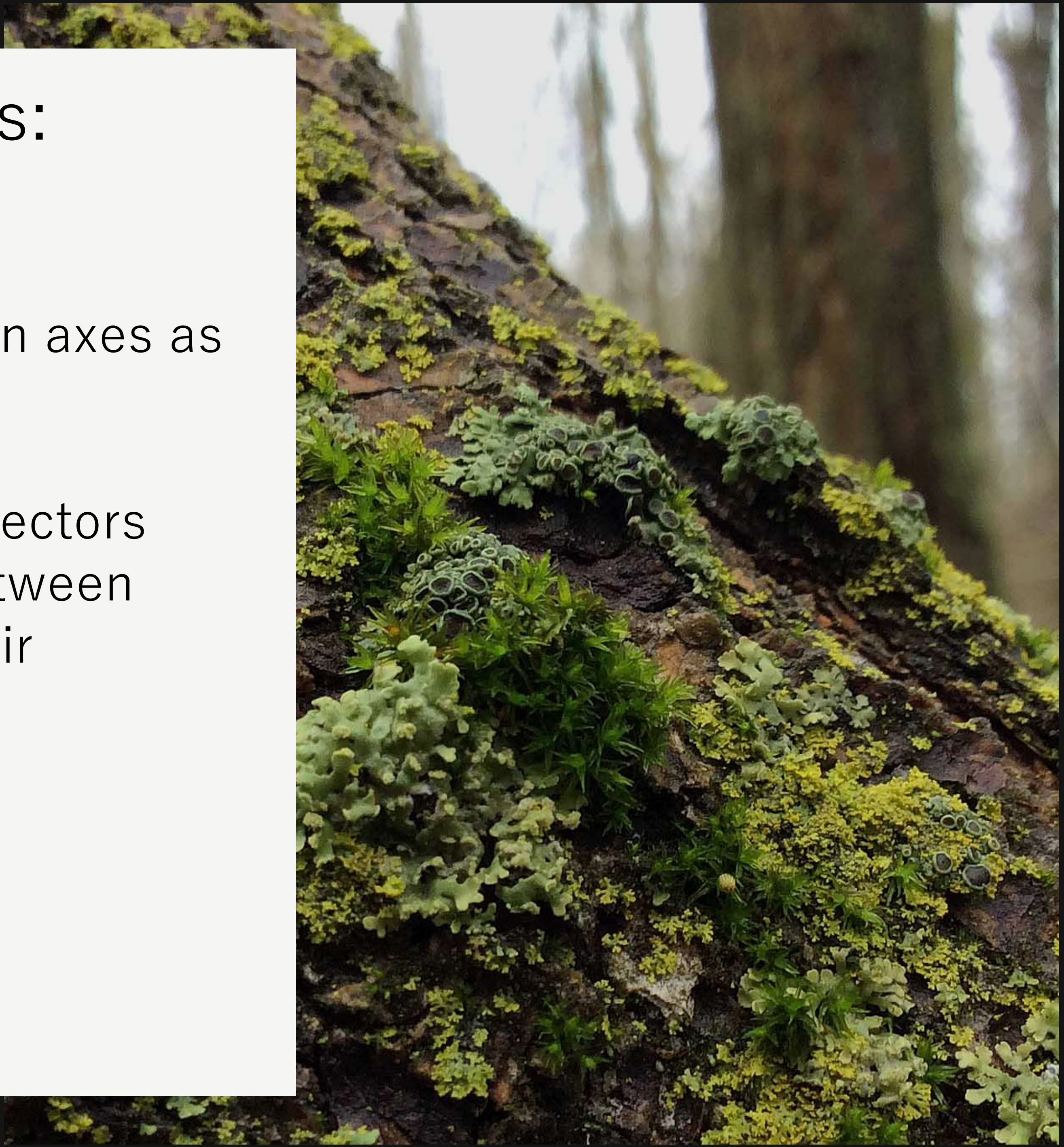
Scaling Method #1: Eigenvectors are normalized to unit length 1



Principal Component Analysis: Process and Steps

Step 6) Scale loadings of descriptors on axes as appropriate

Scaling Method #2: Scales the eigenvectors such that the cosines of the angles between descriptor-axes are proportional to their covariances.

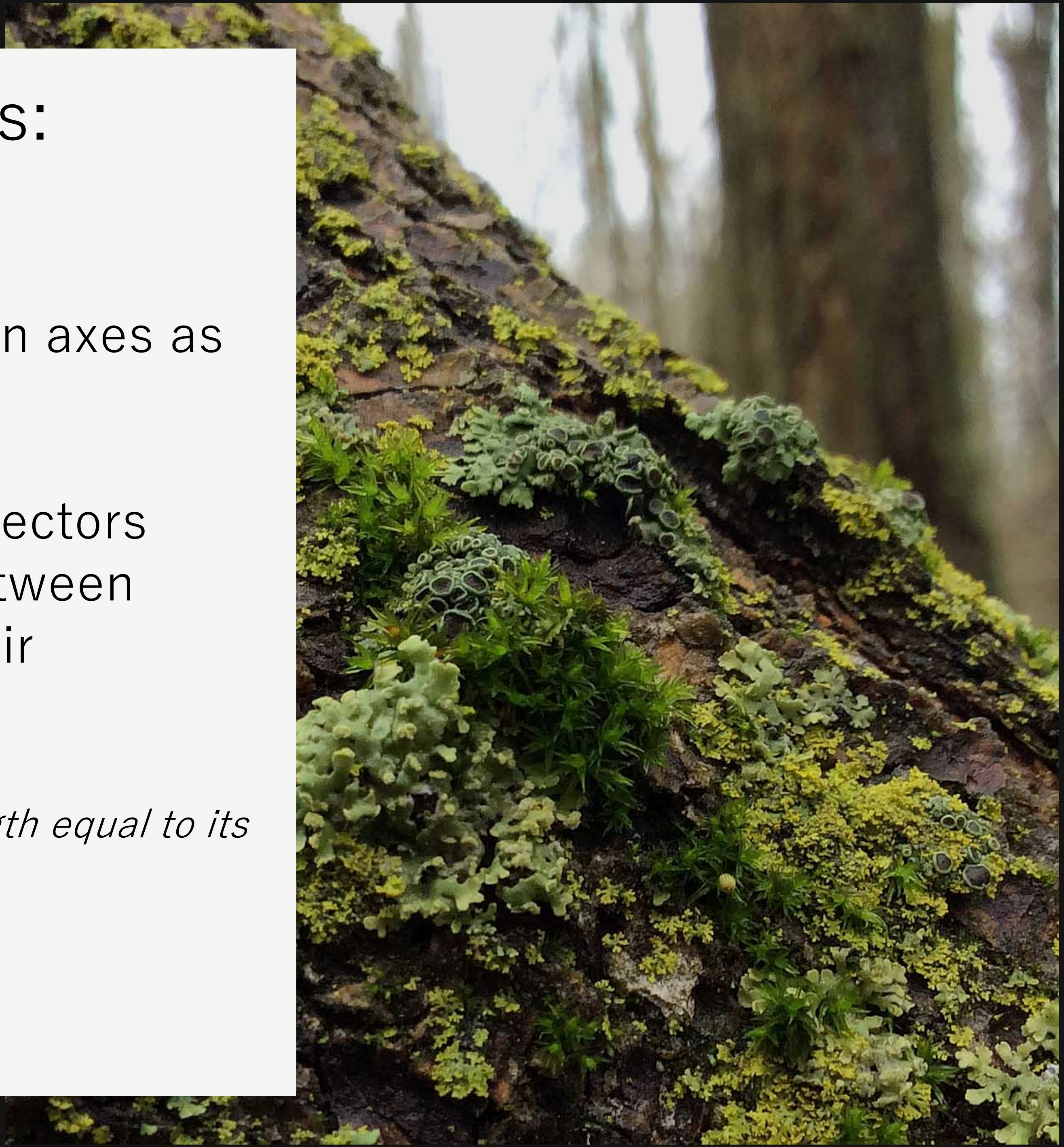


Principal Component Analysis: Process and Steps

Step 6) Scale loadings of descriptors on axes as appropriate

Scaling Method #2: Scales the eigenvectors such that the cosines of the angles between descriptor-axes are proportional to their covariances.

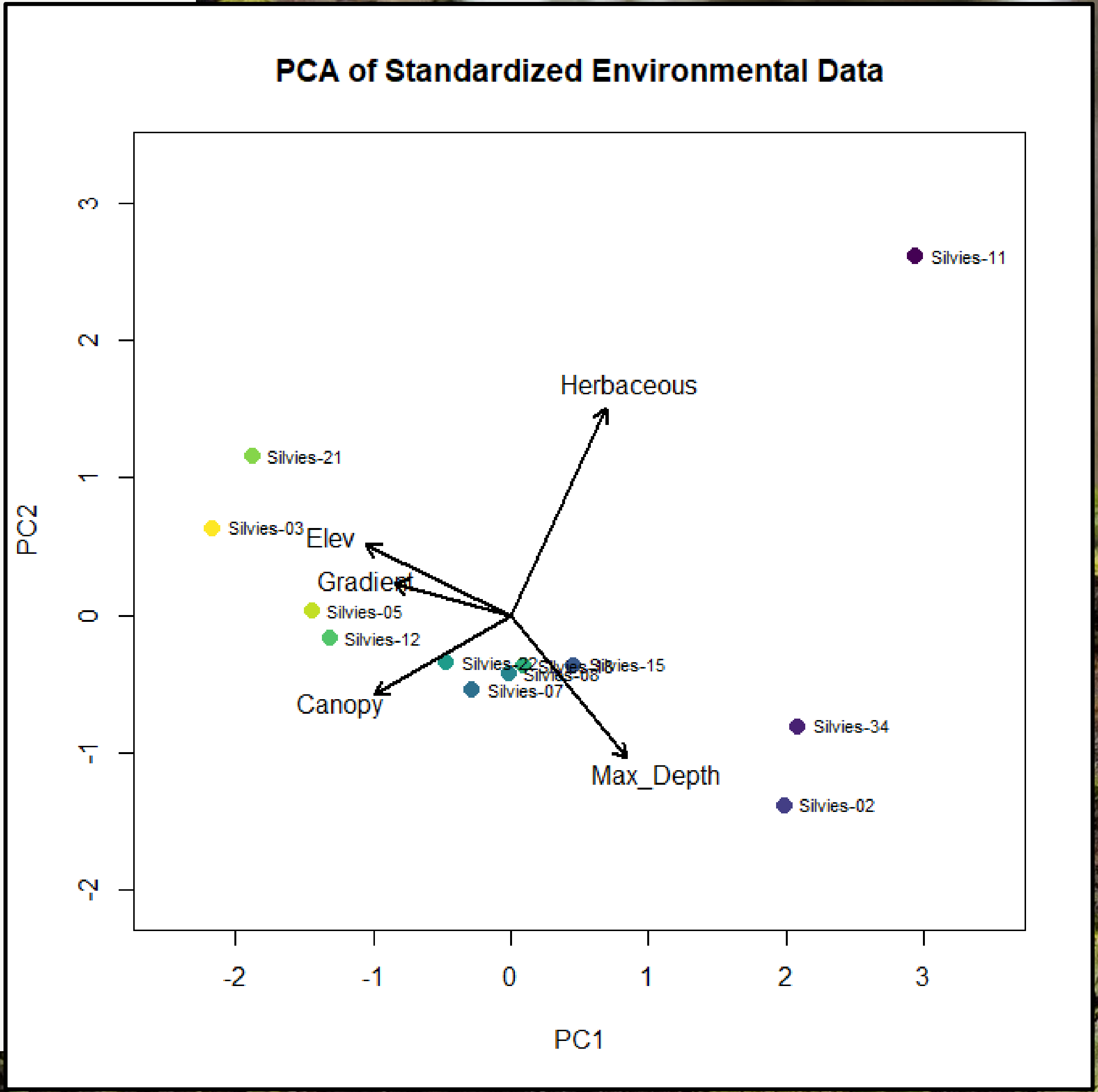
Accomplished by scaling each eigenvector to a length equal to its standard deviation.



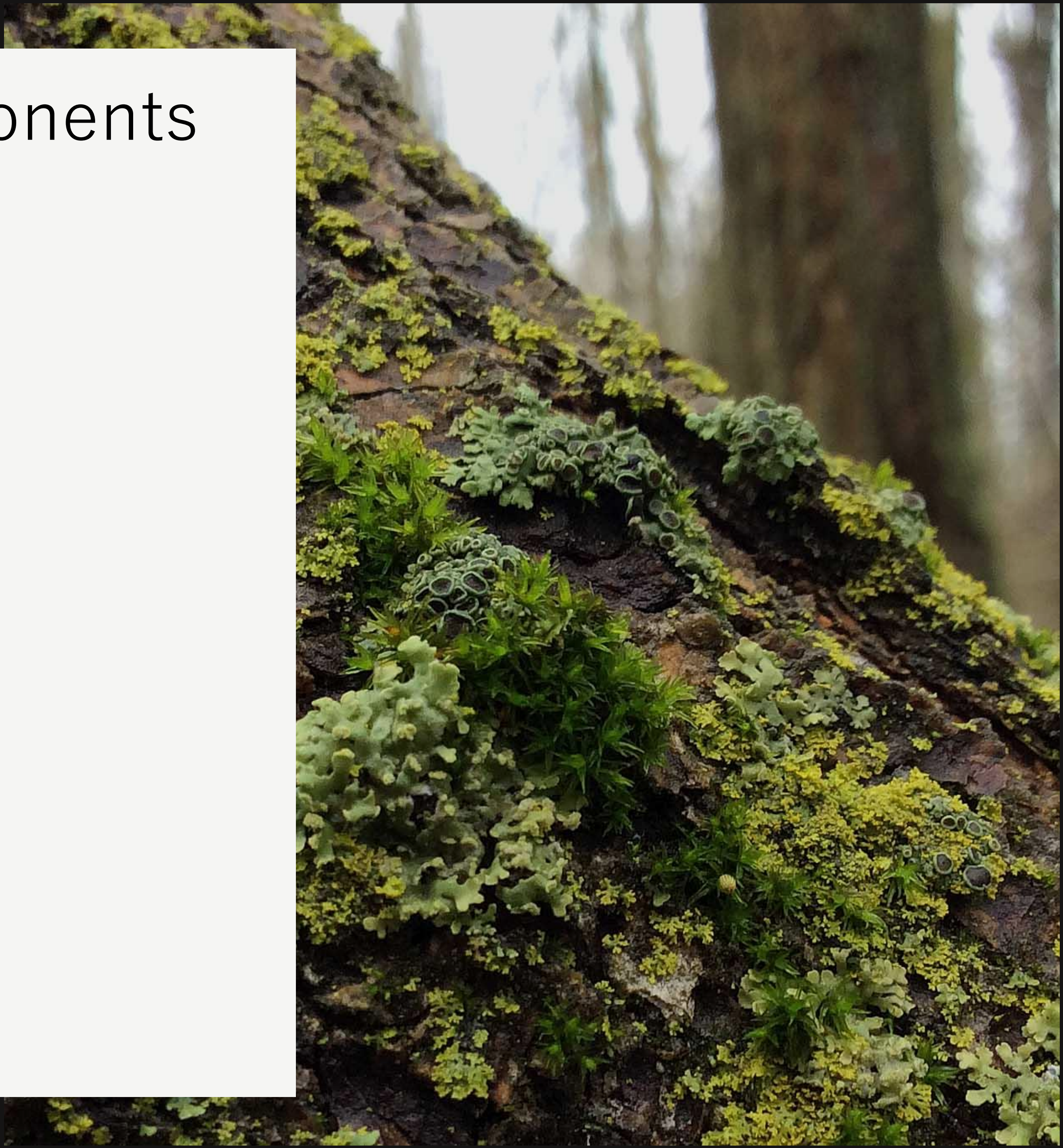
Principal Component Analysis: Process and Steps

Step 7) Visualize using a PCA biplot

Scaling 2: correlation biplot

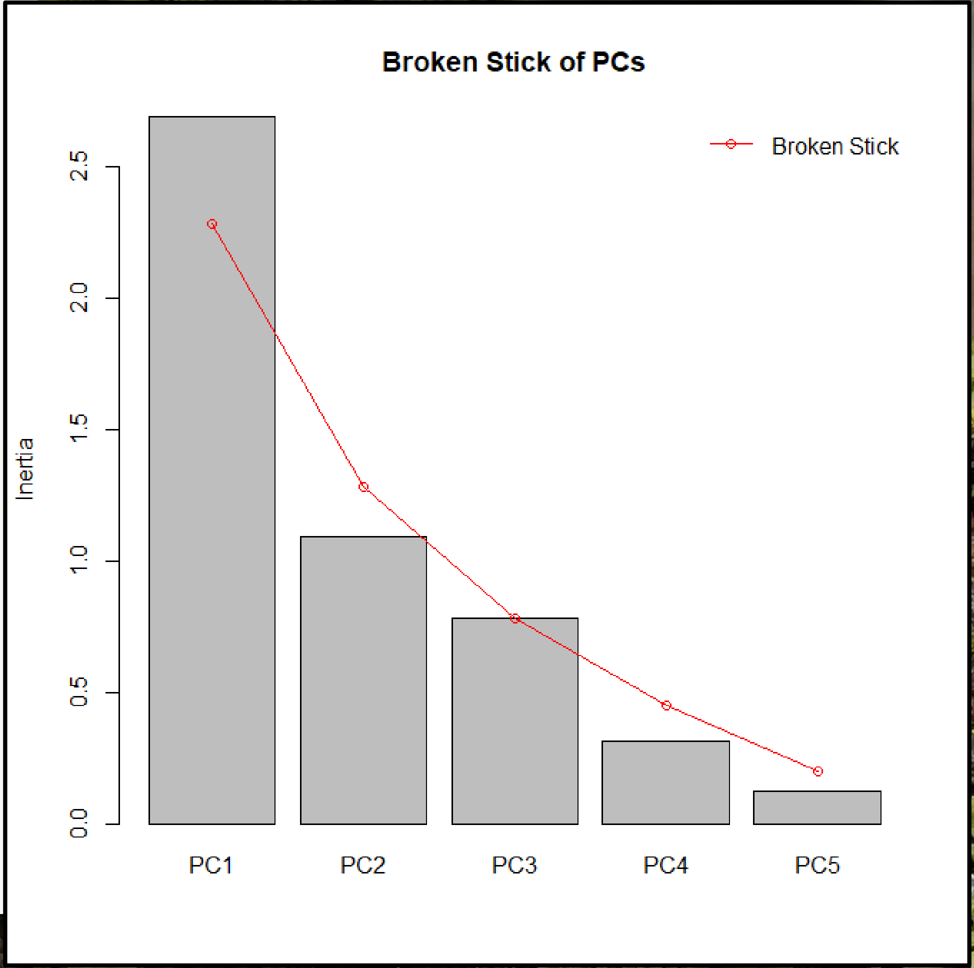


Assessing Meaningful Components

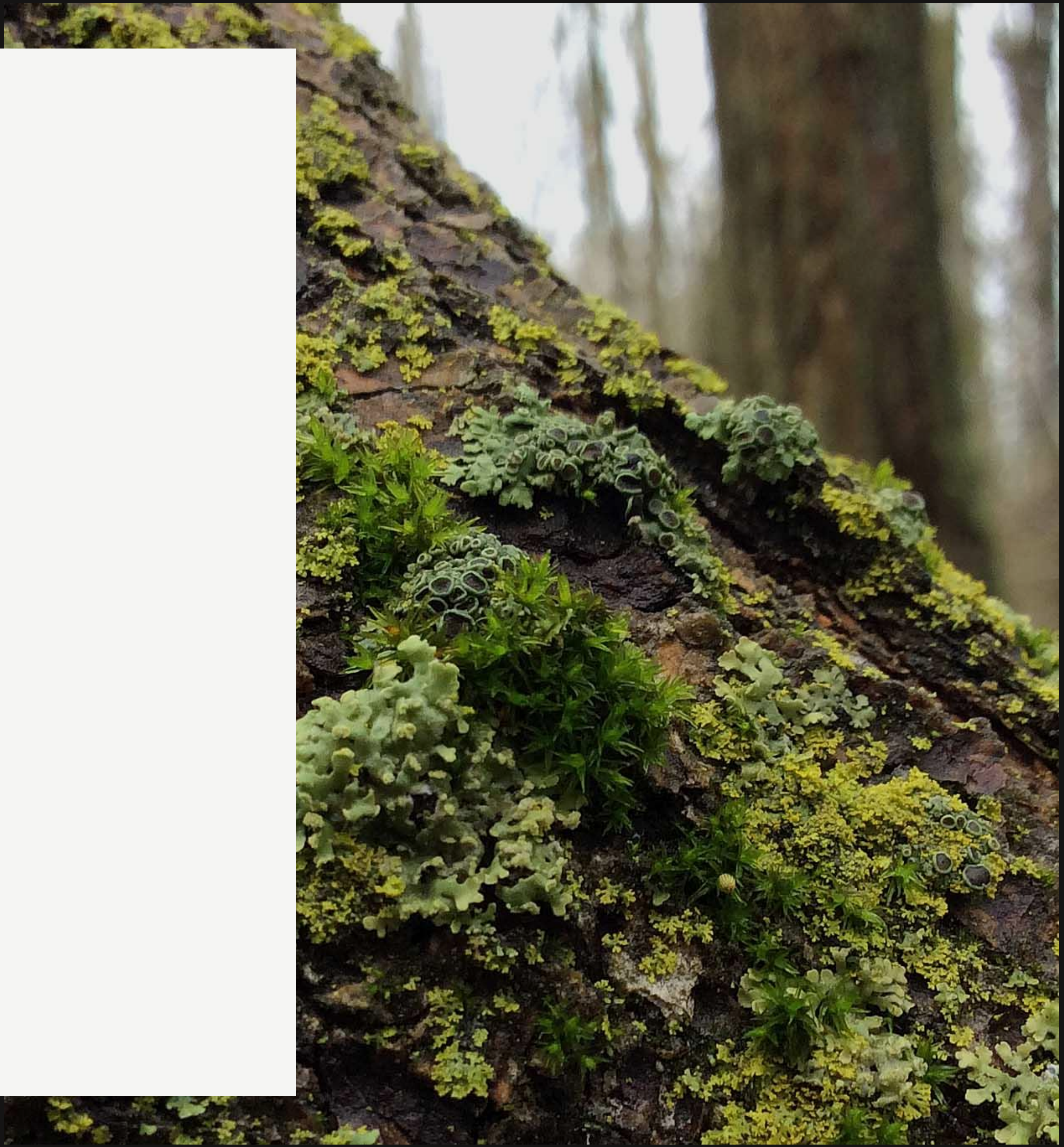


Assessing Meaningful Components

The **broken stick** model identifies principal axes that explain a fraction of variance as small as or smaller than would be predicted by chance.



Limitations



Limitations

PCA: To use or not to use?

- Optimal use calls for normalization of the data

Limitations

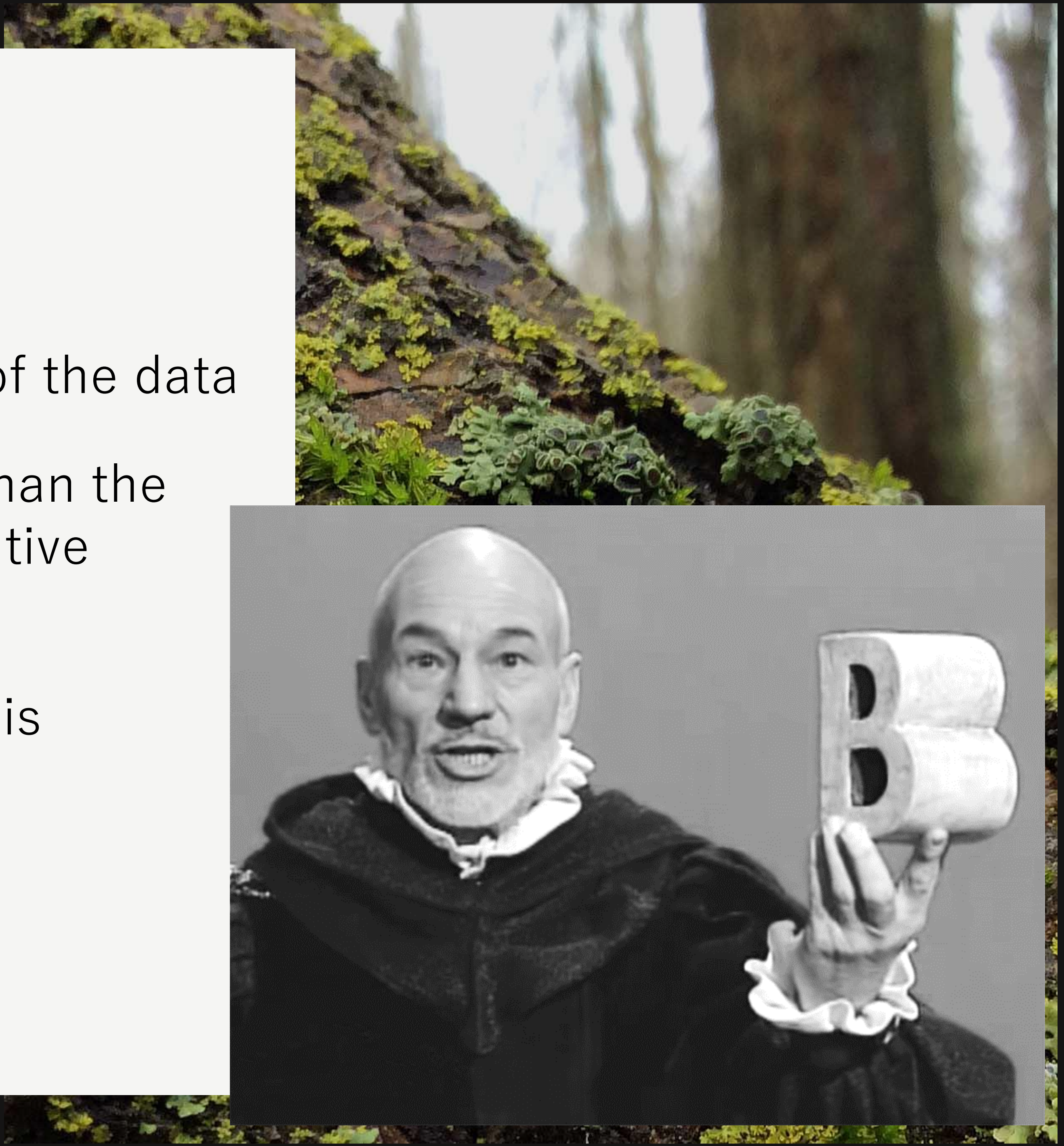
PCA: To use or not to use?

- Optimal use calls for normalization of the data
- If the number of objects is smaller than the number of descriptors ($n < p$), negative eigenvalues will occur

Limitations

PCA: To use or not to use?

- Optimal use calls for normalization of the data
- If the number of objects is smaller than the number of descriptors ($n < p$), negative eigenvalues will occur
- PCA is not useful for R-mode analysis



Limitations

PCA: To use or not to use?

- Optimal use calls for normalization of the data
- If the number of objects is smaller than the number of descriptors ($n < p$), negative eigenvalues will occur
- PCA is not useful for R-mode analysis
- PCA cannot incorporate multi-state descriptors

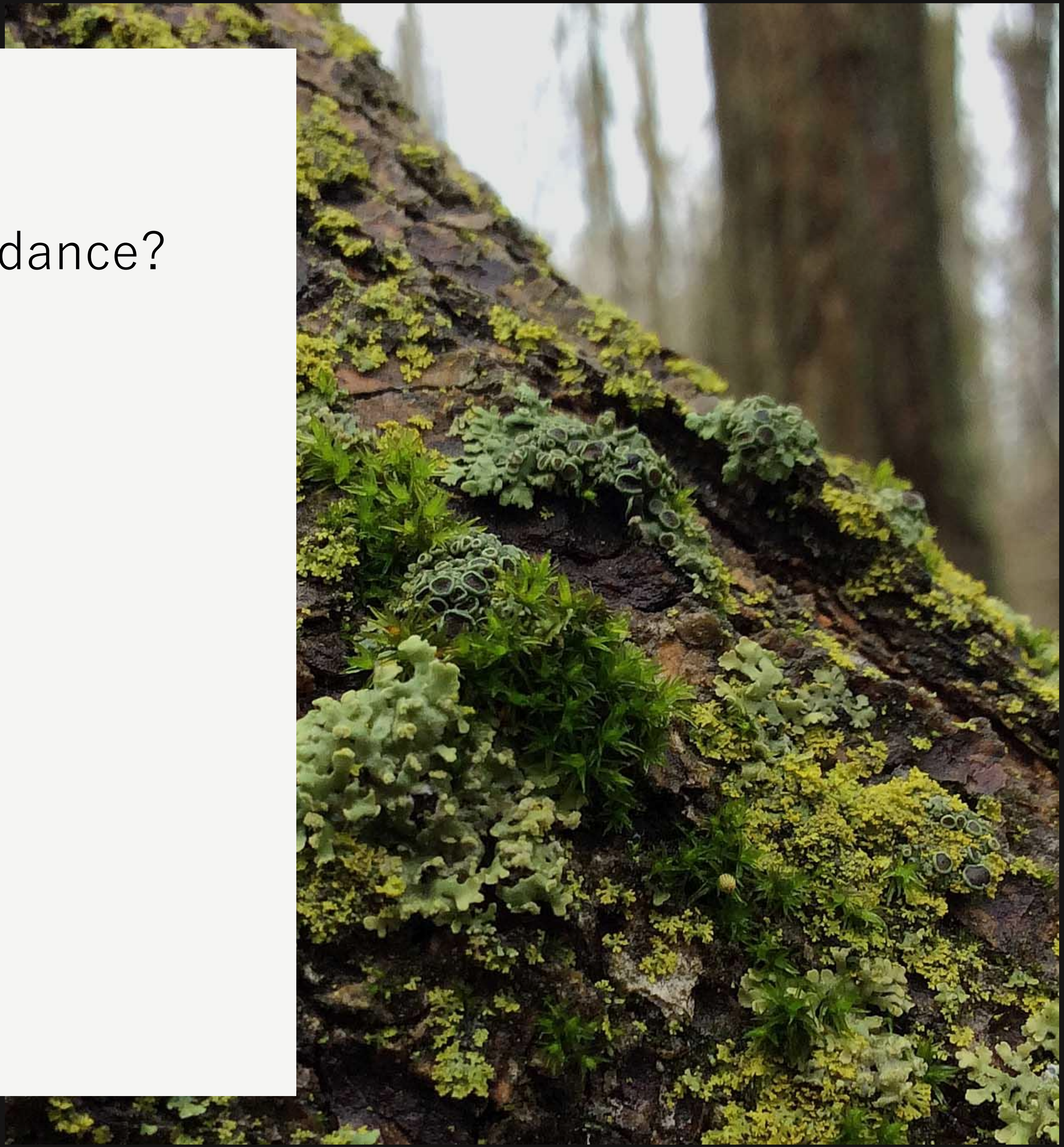
Limitations

PCA: To use or not to use?

- Optimal use calls for normalization of the data
- If the number of objects is smaller than the number of descriptors ($n < p$), negative eigenvalues will occur
- PCA is not useful for R-mode analysis
- PCA cannot incorporate multi-state descriptors
- Watch out for the double zero problem!

Limitations

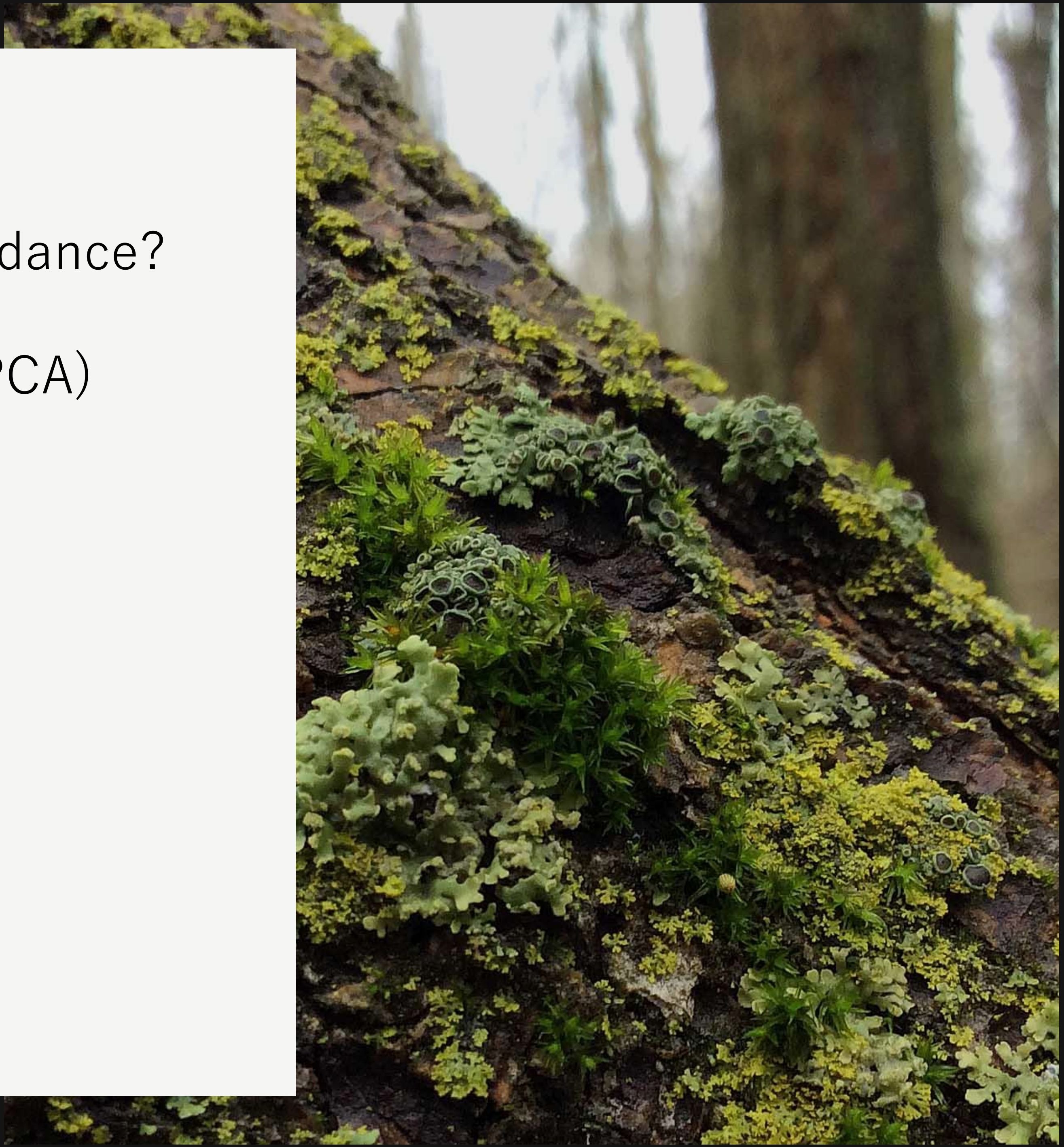
But can PCA be used for species abundance?



Limitations

But can PCA be used for species abundance?

Enter: Transformation-based PCA (tbPCA)

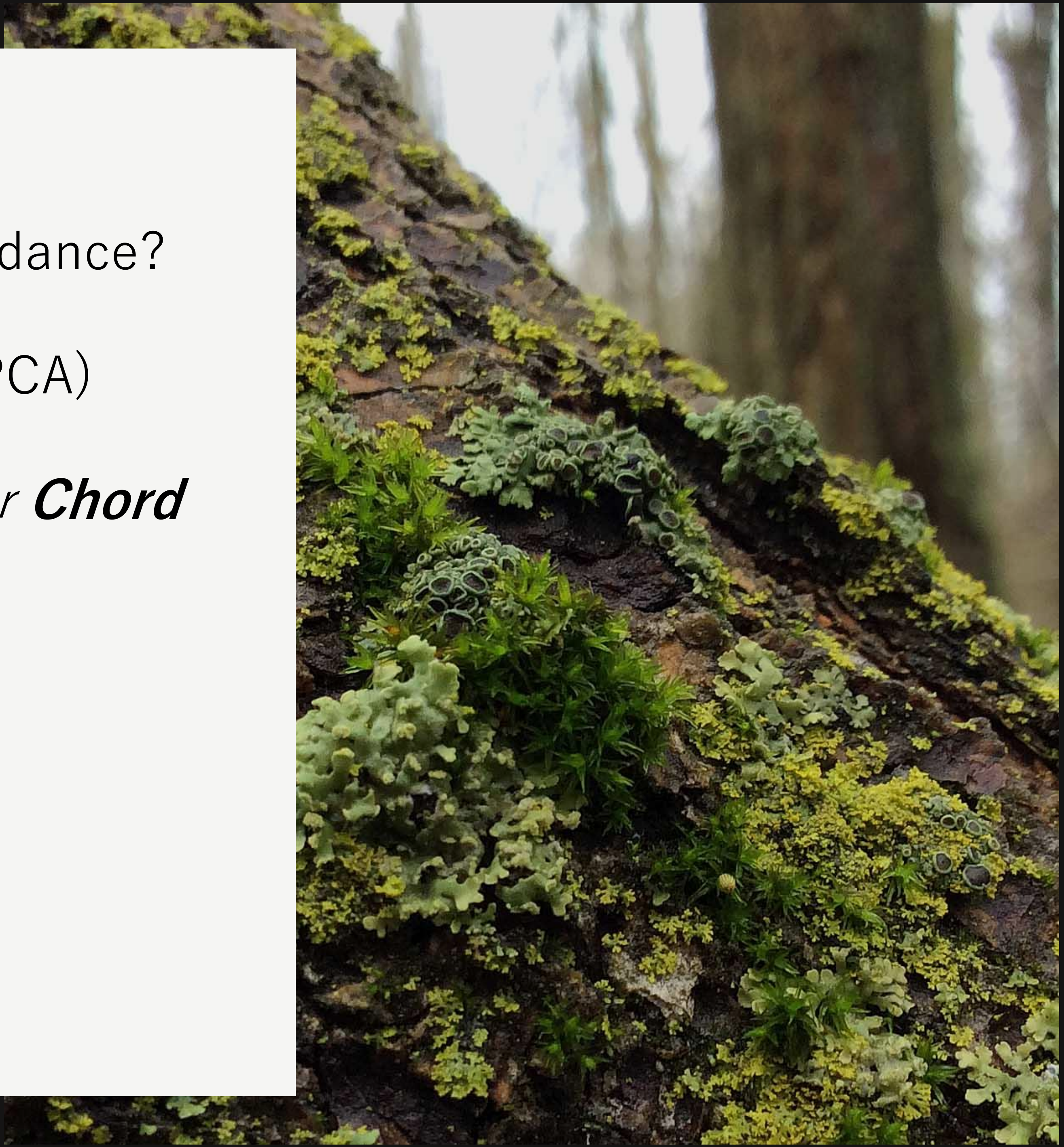


Limitations

But can PCA be used for species abundance?

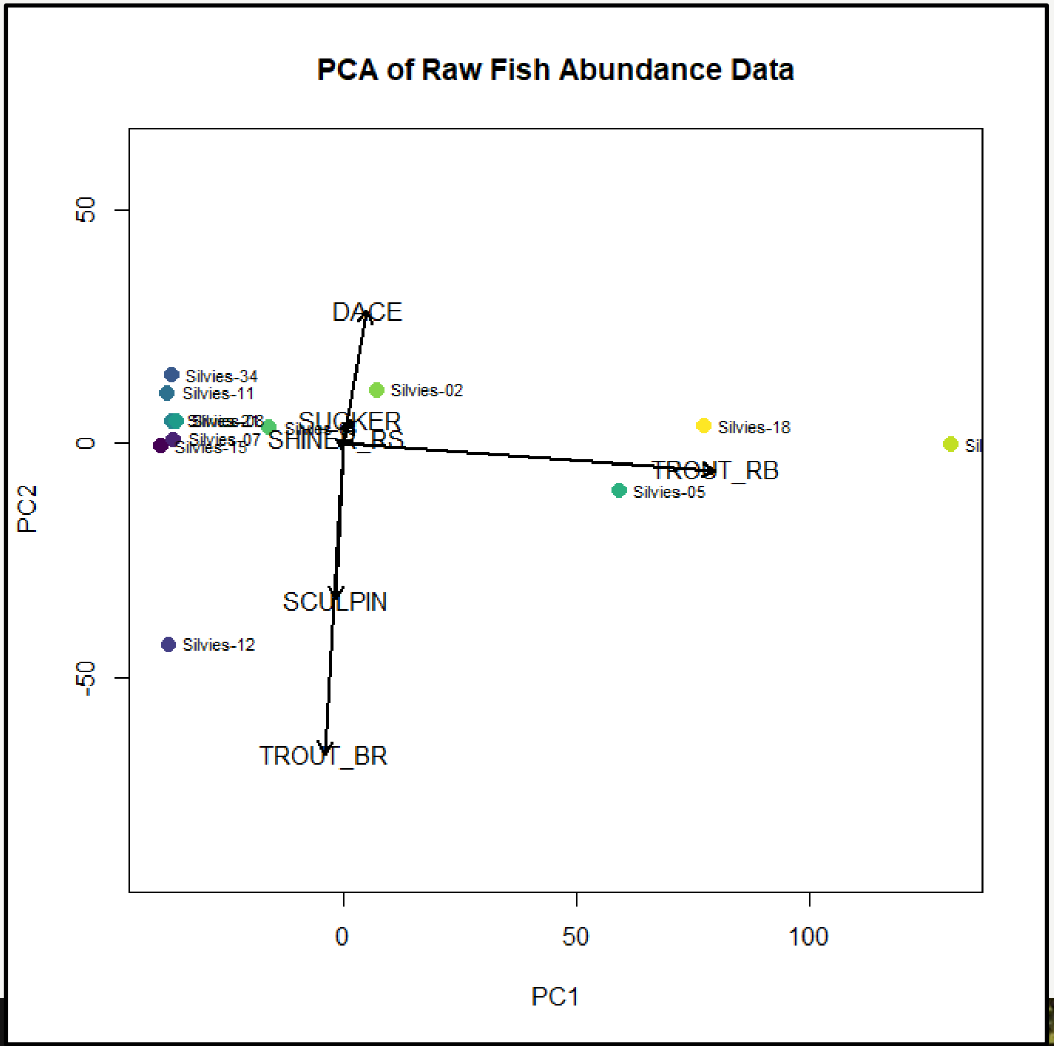
Enter: Transformation-based PCA (tbPCA)

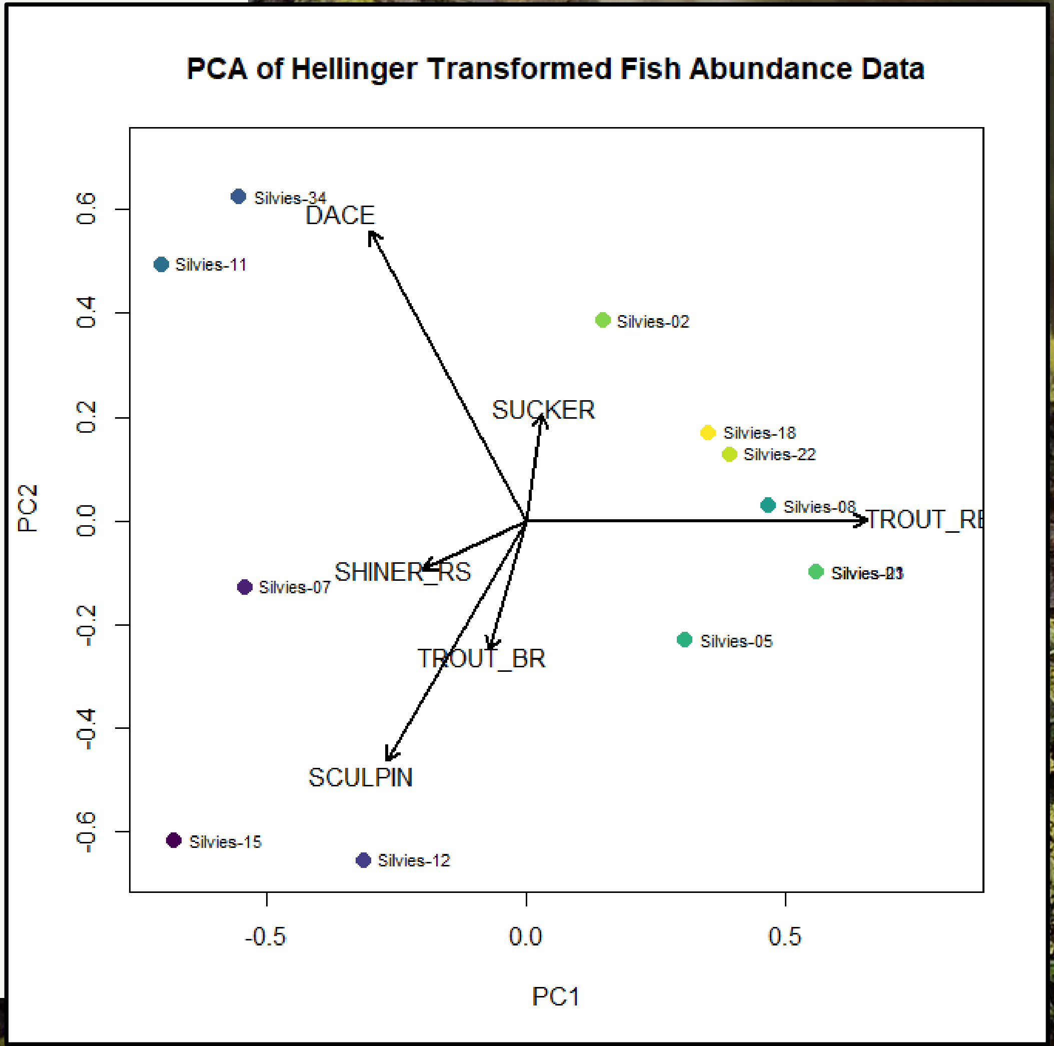
*Usually conducted using a **Hellinger** or **Chord** transformation*



Limitations

Transformation-based PCA (tbPCA)



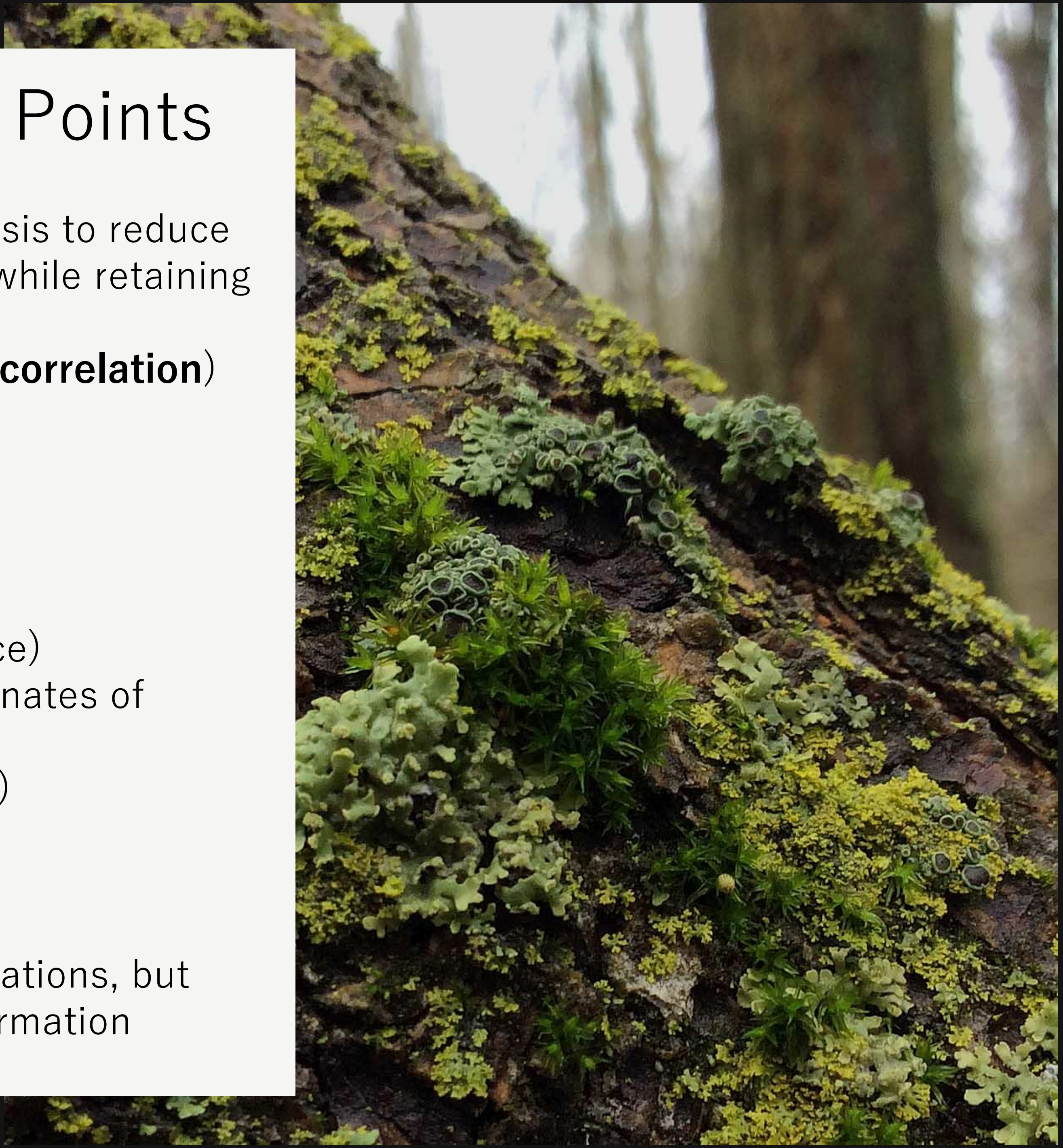


Conclusion: Summary of Key Points

- **Principal Component Analysis** uses eigenanalysis to reduce the dimensionality of large, ecological datasets while retaining as much information as possible
    - Carried out on a **dispersion** (**covariance** or **correlation**) matrix

- Steps:
    1. Column center (and standardize) data
    2. Compute dispersion matrix
    3. Solve for eigenvalues (i.e., explained variance)
    4. Solve for eigenvectors (i.e., **loadings**, coordinates of principal axes)
    5. Compute principal components (i.e., **scores**)
    6. Scale eigenvectors
    7. Visualize using PCA biplot

- Species abundance data violates many PCA limitations, but can be overcome with Hellinger or Chord transformation



Questions?