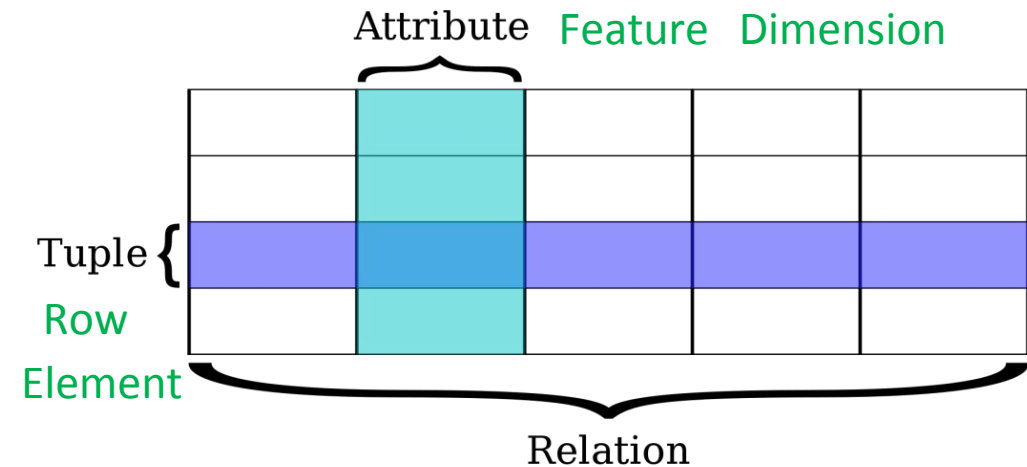# Linear Algebra

Praphul Chandra

Math Recap

- Numbers (Scalars)
  - Add, Subtract, Multiply, Divide
  - Identities : Zero, One
  - Size, distance

- Vectors
  - Add, Subtract, Multiply, Divide
  - Identities : Zero, One
  - Size, distance

- Matrices
  - Add, Subtract, Multiply, Divide (Decomposition)
  - Identities : Zero, One
  - Size, distance

Attribute Feature Dimension

Tuple {

Row

Element

Relation

$$\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip}) \in \mathbb{R}^p$$
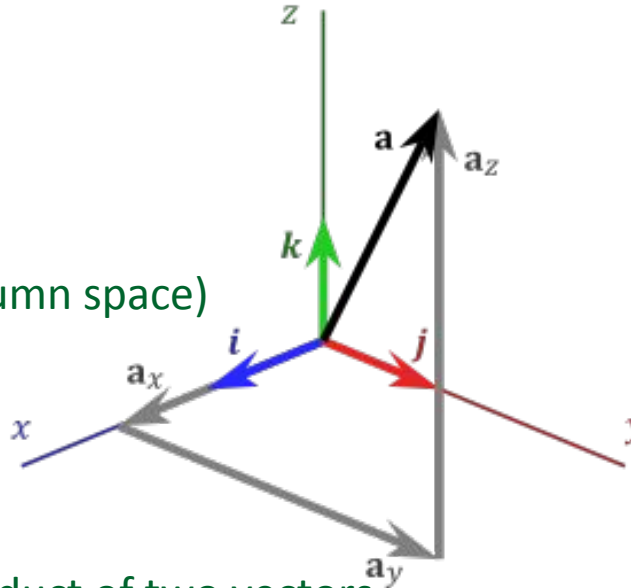
$$X \in \mathbb{R}^{n \times p}$$

Vectors

- In Data
  - Each row in X
  - Each column in X (Define the column space)
- Operations
  - Add, Subtract : component wise
- Multiply
  - Inner Product : Element wise product of two vectors
  - a.k.a. scalar product a.k.a. dot product
- Norm
  - Measure of how "big" a vector is

$$\begin{pmatrix} 3 \\ -2 \\ 6 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 3 \\ -5 \end{pmatrix} = 3 \times 2 + (-2) \times 3 + 6 \times (-5) = 6 - 6 - 30 = -30.$$

$$\mathbf{w} \in \mathbb{R}^P, \mathbf{x} \in \mathbb{R}^P$$

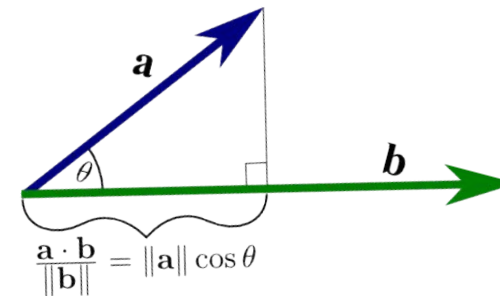$$\mathbf{w}^T \mathbf{x} = \sum_{j=1}^{P} w_j x_j = \langle \mathbf{w}, \mathbf{x} \rangle$$

$$\|x_i\|_1 = \sum_p |x_j|$$

$$\|x_i\|_2 = \left(\sum_p |x_j|^2 \right)^{\frac{1}{2}}$$

$$\|x_i\|_z = \left(\sum_p |x_j|^z \right)^{\frac{1}{z}}$$
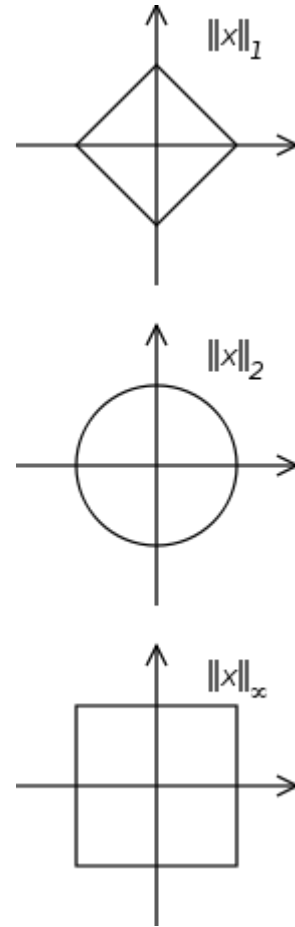
$$\|x_i\|_\infty = \max_p |x_j|$$

$$\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{b}\|} = \|\mathbf{a}\| \cos \theta$$

$$\begin{aligned}
\mathbf{w}^T \mathbf{x} &= \|\mathbf{w}\|\|\mathbf{x}\| \cos \theta \\
\theta = 90 &\Rightarrow \mathbf{w}^T \mathbf{x} = 0 \\
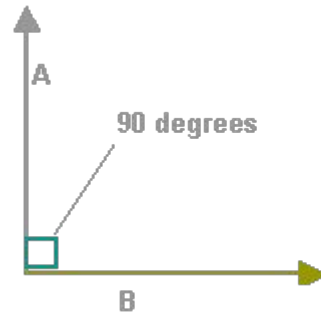\theta = 0 &\Rightarrow \mathbf{w}^T \mathbf{x} = \|\mathbf{w}\|\|\mathbf{x}\| \\
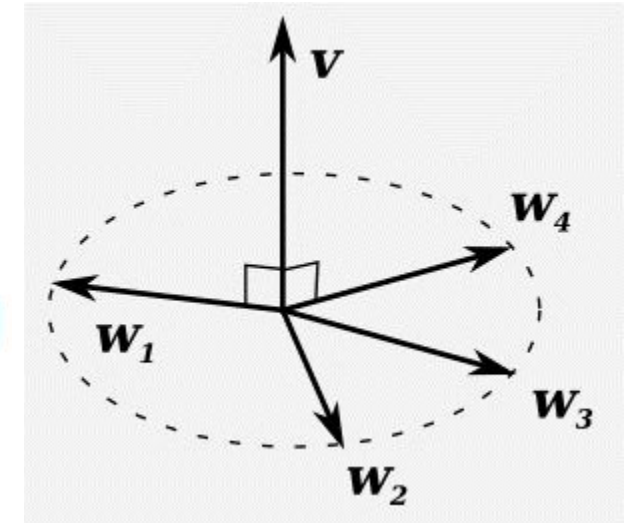\mathbf{w}^T \mathbf{w} &= \|\mathbf{w}\|^2
\end{aligned}$$

Independent Vectors

- Orthogonal vectors
  - "perpendicular"
  - Dot Product is zero

$$w^T x = \|w\|\|x\| \cos \theta$$
$$\theta = 90 \Rightarrow w^T x = 0$$
$$\theta = 0 \Rightarrow w^T x = \|w\|\|x\|$$
$$w^T w = \|w\|^2$$

- Independent Vectors
  - No correlation among the vectors.
  - If v1 and v2 are independent,
    - v2 cant be expressed as a linear combination of v1
  - If v1, v2 and v3 are independent,
    - v2 cant be expressed as a linear combination of v1, v3
  - Orthogonal ➜ Independence
    - Other way is not true
  - In n dimensions, n linearly independent vectors can span the entire space:
  - These vectors are called bases

- Independent Vectors : Data Interpretation
  - How many attributes are dependent on others (can be expressed as linear combinations)
  - The number of linearly independent rows/columns is the rank
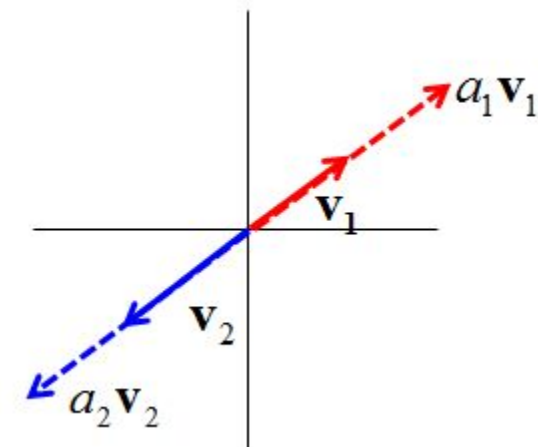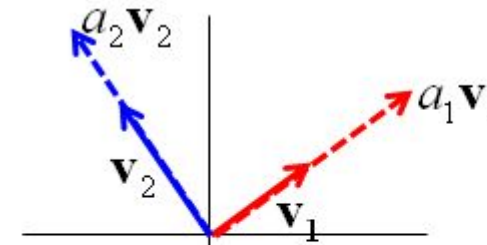
Welcome to the Matrix

The Matrix

- Data as Matrix
  - Row & Column

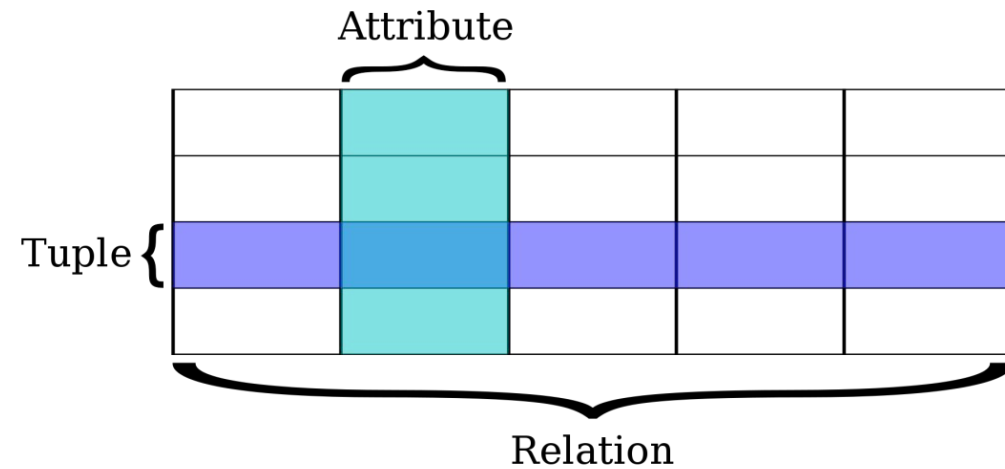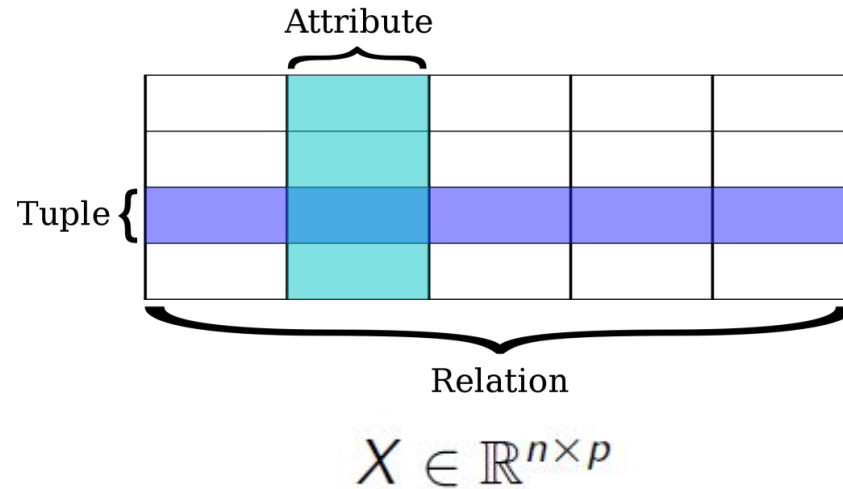- Functions as Matrix
  - Transformations
  - Linearity

- Graphs as Matrix
  - Relationships

The Matrix

- In Data
  - The Data Matrix X

- Operations
  - Add, Subtract : component wise

- Norm
  - Measure of how "big" a matrix is

- Multiply
  - ...

Attribute

Tuple {

Relation

$$X \in \mathbb{R}^{n \times p}$$

$$\|x_i\|_1 = \sum_p |x_j|$$

$$\|x_i\|_2 = \left(\sum_p |x_j|^2 \right)^{\frac{1}{2}}$$

$$\|x_i\|_z = \left(\sum_p |x_j|^z \right)^{\frac{1}{z}}$$

$$\|x_i\|_\infty = \max_p |x_j|$$

Matrix 1 **Matrix 2** **Matrix 1 + 2**

$$\begin{bmatrix} 10 & 0 \\ -4 & 5 \end{bmatrix} + \begin{bmatrix} -6 & 3 \\ 1 & -7 \end{bmatrix} = \begin{bmatrix} 4 & 3 \\ -3 & -2 \end{bmatrix}$$

2 x 2 　　2 x 2 　　2 x 2

$$X \in \mathbb{R}^{n \times p}$$

$$\|X\|_1 = \max_j \sum_{i=1}^{n} |x_{ij}| : \text{Max abs. col. sum}$$

$$\|X\|_\infty = \max_i \sum_{j=1}^{p} |x_{ij}| : \text{Max abs. row sum}$$

$$\|X\|_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{p} |x_{ij}|^2 \right)^{\frac{1}{2}} = \|X\|_F :$$

Frobenius Norm

Matrix Multiplication : Row x Column

- Matrix x Matrix = Matrix
  - Each element in C is a product of a row from A & column from B
  - Watch the sizes!

$$\begin{bmatrix} * & * & * & * \\ 4 & 2 & 3 & 7 \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} * & * & 1 & * \\ * & * & 1 & * \\ * & * & 0 & * \\ * & * & 2 & * \end{bmatrix} = \begin{bmatrix} * & * & * & * \\ * & * & 20 & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}$$

$$c_{23} = \sum_k a_{2k}.b_{k3} = 4.1 + 2.1 + 3.0 + 7.2 = 20$$

$$c_{ij} = \sum_k a_{ik}.b_{kj}$$
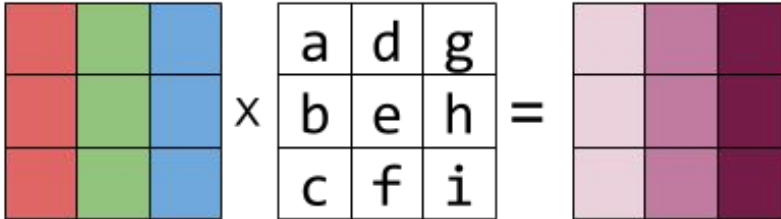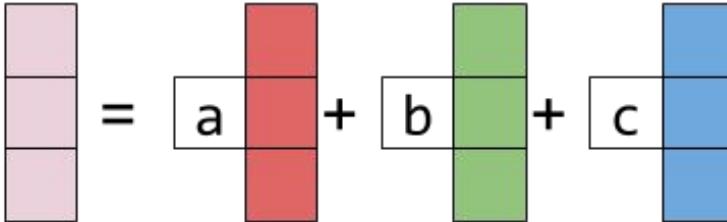
Matrix Multiplication : Column View
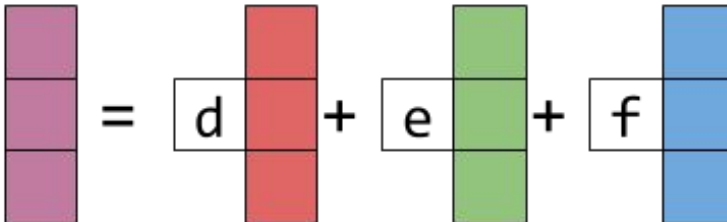
- Each column in C is a linear combination of the columns in A;

- Combination specified by columns in B
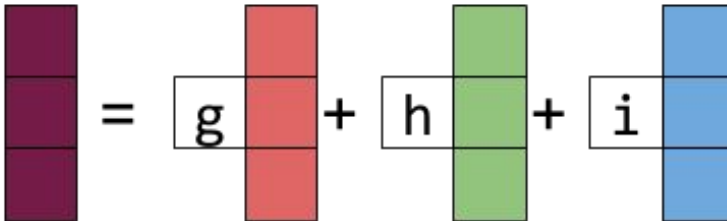
$$
\begin{bmatrix} 2 & 0 & 7 & 1 \\ 4 & 2 & 3 & 7 \\ 1 & 2 & 5 & 6 \\ 3 & 8 & 0 & 0 \end{bmatrix} \begin{bmatrix} * & * & 1 & * \\ * & * & 1 & * \\ * & * & 0 & * \\ * & * & 2 & * \end{bmatrix} = \begin{bmatrix} * & * & 4 & * \\ * & * & 20 & * \\ * & * & 15 & * \\ * & * & 11 & * \end{bmatrix}
$$

$$
C_{*3} = 1\begin{bmatrix} 2 \\ 4 \\ 1 \\ 3 \end{bmatrix} + 1\begin{bmatrix} 0 \\ 2 \\ 2 \\ 8 \end{bmatrix} + 0\begin{bmatrix} 7 \\ 3 \\ 5 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 1 \\ 7 \\ 6 \\ 0 \end{bmatrix}
$$

Matrix Decomposition (~ Division ~ Factorization)

- Matrix decomposition
  - Express a matrix as a the product of two matrices : C = BA
  - a.k.a. Matrix factorization
  - Recall prime number factorization (The fundamental theorem of Algebra)

- Why decompose a data matrix?
  - Express the matrix in different basis
  - X = XI : Native (Naïve) Basis : (Multiplicative) Identity Matrix : I
  - XI = UV : Data (X) expressed in Native Basis (I) is expressed as a linear combination (V) in a new basis (U)
  - Desirable: Orthogonal (Independent) basis : uncorrelated features

The notion of Basis

- Basis of a Vector Space
  - A set of linearly independent vectors such that every vector in the vector space is a linear combination of this set.
  - Dimension : number of linearly independent vectors which span the space

- Rank of a Matrix
  - Dimension of the basis for the column space of X.
  - How many linearly independent vectors capture the information provided in the p columns?

- Basis Transformation
  - Express given data in some other basis
  - Dimensionality Reduction: Can we drop some columns in X without loss of information?
    - E.g. if they are correlated with other columns?
    - $k \leq p$: reduction of rank can be lossy or lossless
  - Expansion : Data may be "separable" (classification) in higher dimensions (kernel, SVM)
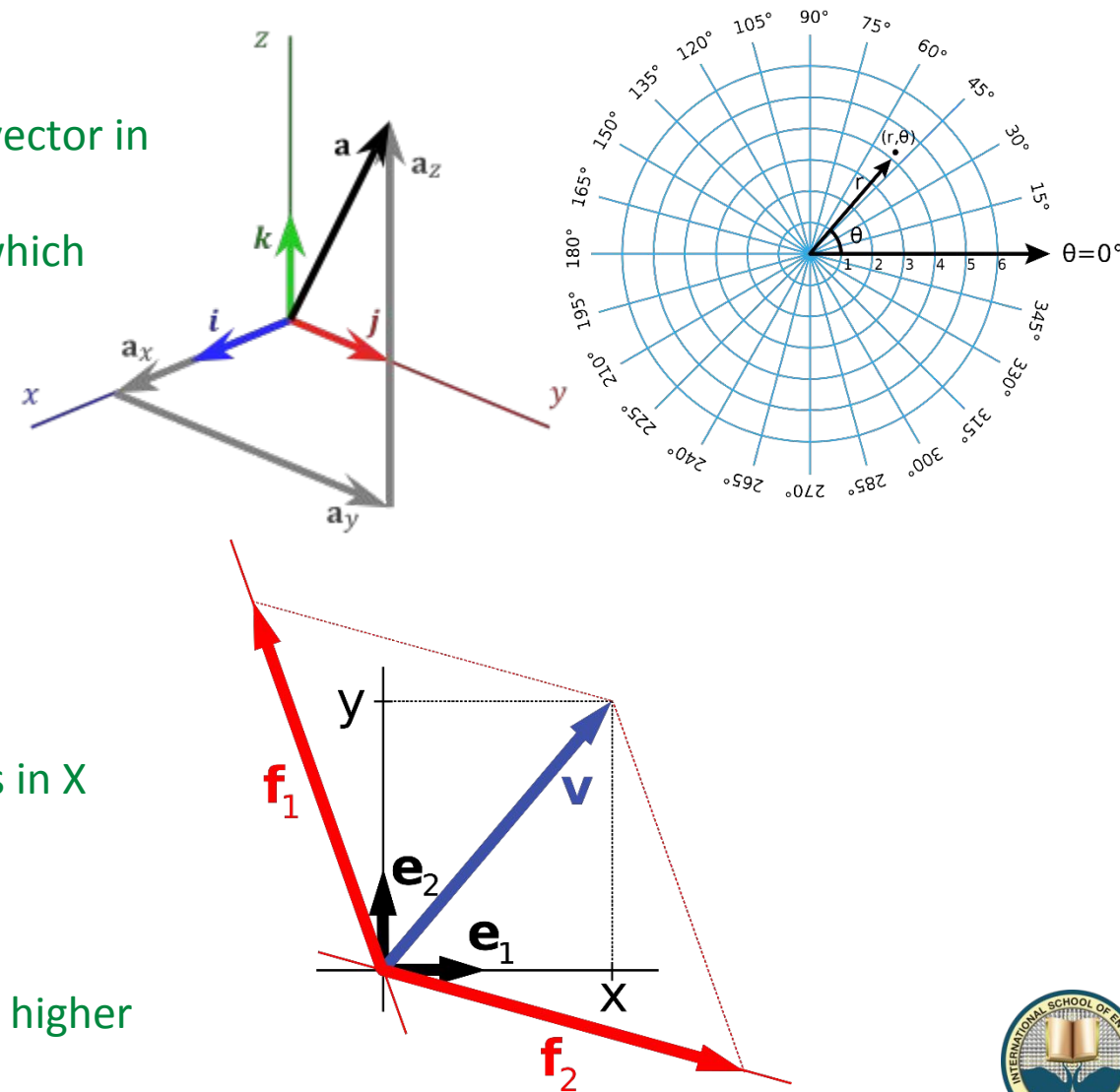
# Basis Transformation

Praphul Chandra

Basis Transformation Motivation : Example

name	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	2	8122	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.1	50.42	15.31	46.26	4.72	63.44	276.4	5	8036	Decastar
ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268	7	8004	Decastar
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.1	8	7995	Decastar
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.6	4.92	52.33	262.1	9	7802	Decastar
HERNU	11.37	7.56	14.41	1.86	51.1	15.06	44.99	4.82	57.19	285.1	10	7733	Decastar
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.1	4.72	55.4	282	11	7708	Decastar
NOOL	11.33	7.27	12.68	1.98	49.2	15.29	37.92	4.62	57.44	266.6	12	7651	Decastar
BOURGUIGN	11.36	6.8	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.7	13	7313	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.9	69.71	282	2	8820	OlympicG
Karpov	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11	3	8725	OlympicG
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.4	58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.9	55.39	278.05	5	8343	OlympicG
Zsivoczky	10.91	7.14	15.31	2.12	49.4	14.95	45.62	4.7	63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.8	57.76	264.35	7	8237	OlympicG
Bernard	10.69	7.48	14.8	2.12	49.13	14.17	44.75	4.4	55.27	276.31	9	8225	OlympicG
Schwarzl	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.1	56.32	273.56	10	8102	OlympicG
Pogorelov	10.95	7.31	15.1	2.06	50.79	14.21	44.6	5	53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3	14.77	1.88	50.3	14.34	44.41	5	60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.6	64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6	14.31	1.98	48.68	14.23	41.1	4.92	51.77	278.1	6	8030	Decastar
Nool	10.8	7.53	14.26	1.88	48.81	14.8	42.05	5.4	61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5	51.53	274.21	19	7926	OlympicG

- Which **athletes** have similar capability? (Clustering)
- Which **sports** require the same underlying **skills**? (Hard to "name")
  - Intuition: Athletes good at sport-X tend to be good at which other sport? (Quantify)
  - Can we uncover the underlying skill dimensions?
  - Each sport can be represented as a skill combination.
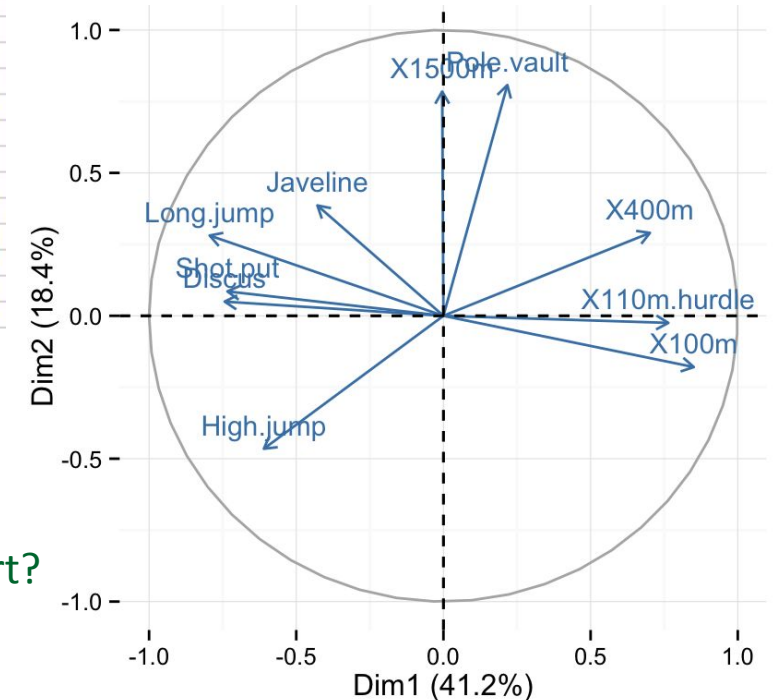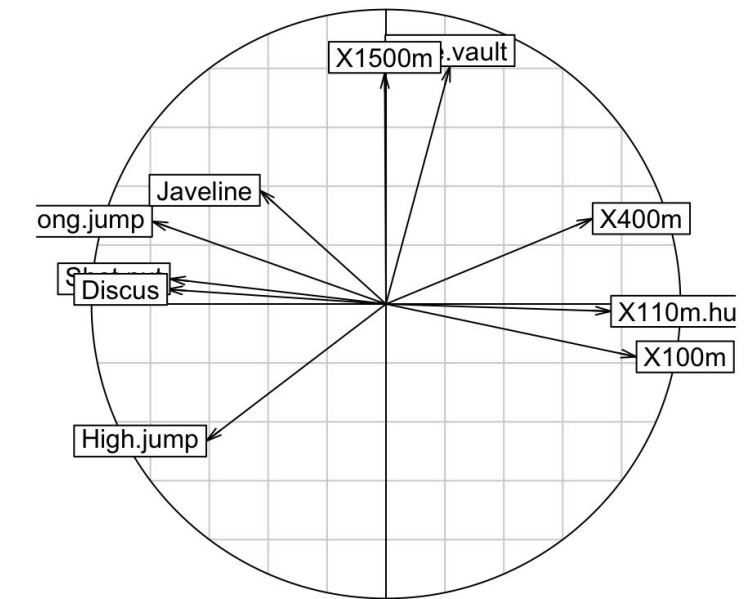
An Example (cont'd)
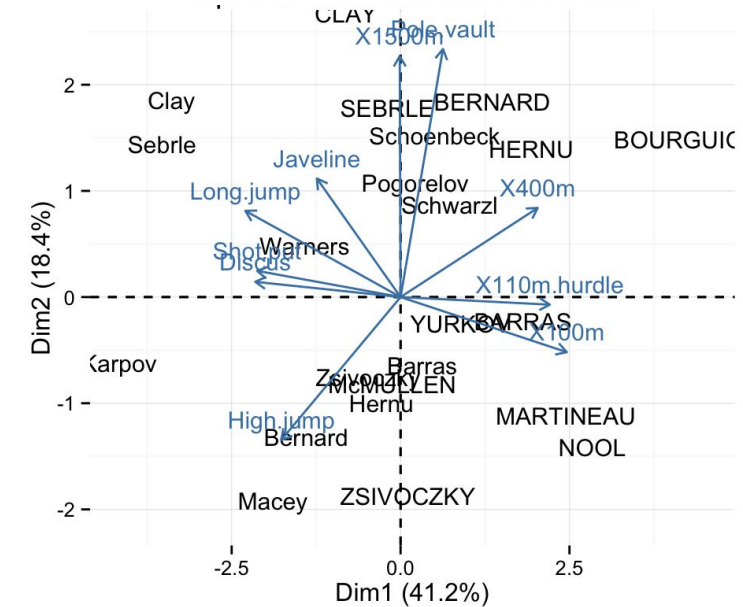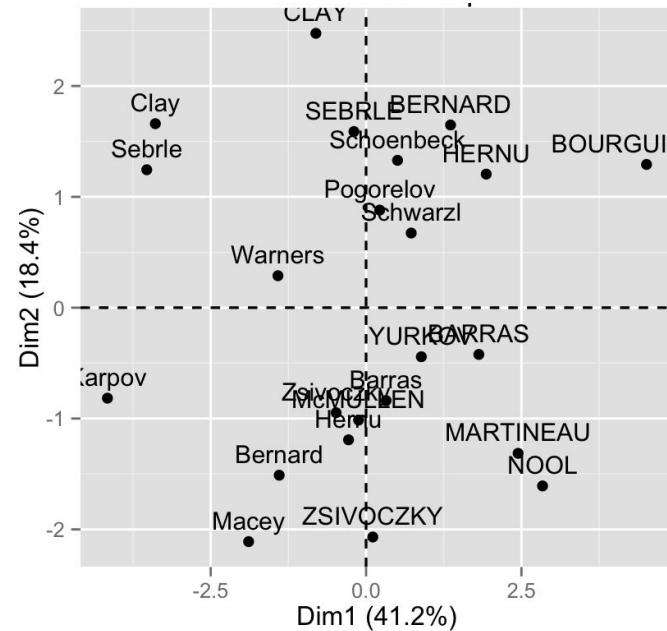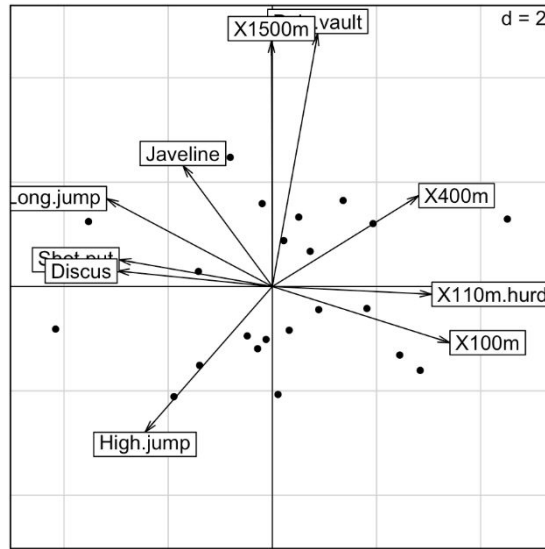
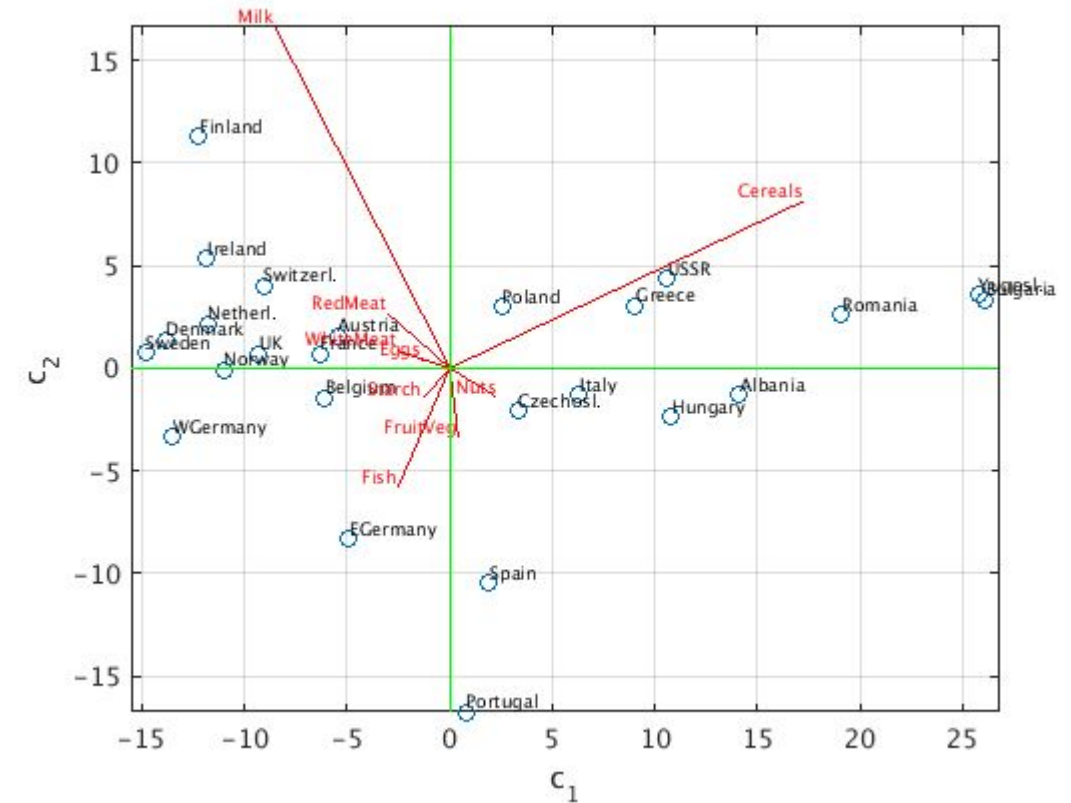- Each **sport** can be represented as a **skill** combination.
  - Interpretation: Based on athletes' performance, athletes who tend to do well in sport-A also do well in sport-B

- Each **athlete** can be represented as a **skill** combination
  - Interpretation : Based on athletes' performance, athlete-U is similar to athlete-V
  - Interpretation : Based on athletes' performance, athlete-U is likely to do well at sport-B
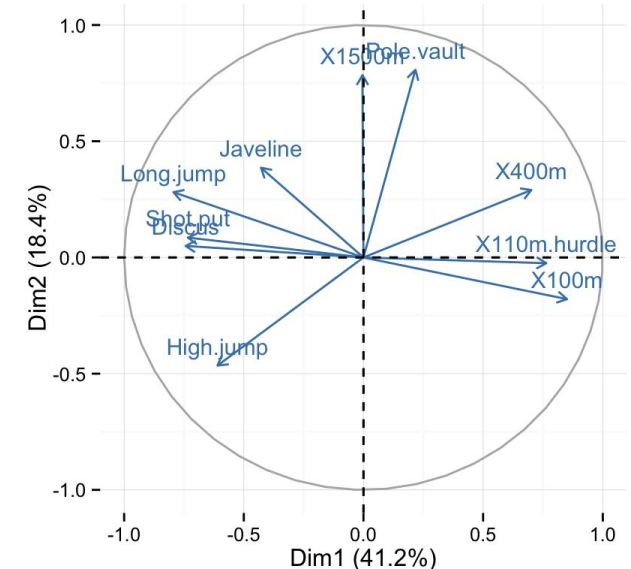
Another Example

T =

	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	FruitVeg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechosl.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
EGermany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherl.	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerl.	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
WGermany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugosl.	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

- The first dimension increases with cereal consumption, and decreases with white & red meat consumption.
- The second dimension increases with milk consumption, and decreases with fish consumption.
- Visualizing high dimensional data in 2D space such that similar items appear closer (Clustering ?)

Basis Transformation

- Native (Naïve) feature space
  - Reflects the structure of the data in which it was captured
  - dimension = feature = axis = column of raw data
  - Each observation is represented (as a vector) in this feature space

- Dimensions (Features) can be combined to create "new" dimensions
  - Create a new column by "combining" columns of raw data
  - We now have new dimensions and thus a new feature space
  - dimension = feature = axis ≠ column of raw data
  - Observations can be represented in this new space (as vectors)
  - Native (naïve) dimensions can be represented in this new space (as vectors)

- Basis transformation
  - Process which changes the bases / axes / dimensions in which the observations are expressed
  - dimension = feature = axis ≠ column of raw data
  - a.k.a. dimension transformation

- Dimensionality reduction
  - # of new dimensions < # of native dimensions *(e.g. drop some (or all) of the original dimensions)*
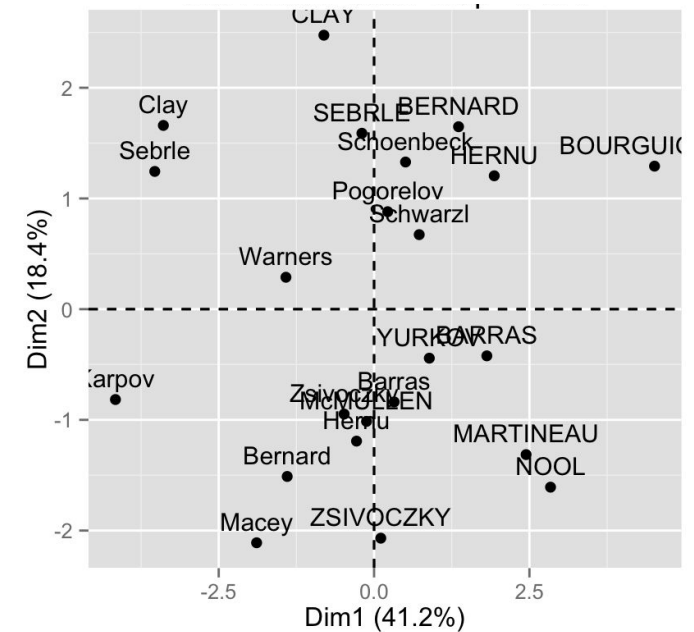
Optimal Bases : Desiderata for Dimensionality Reduction

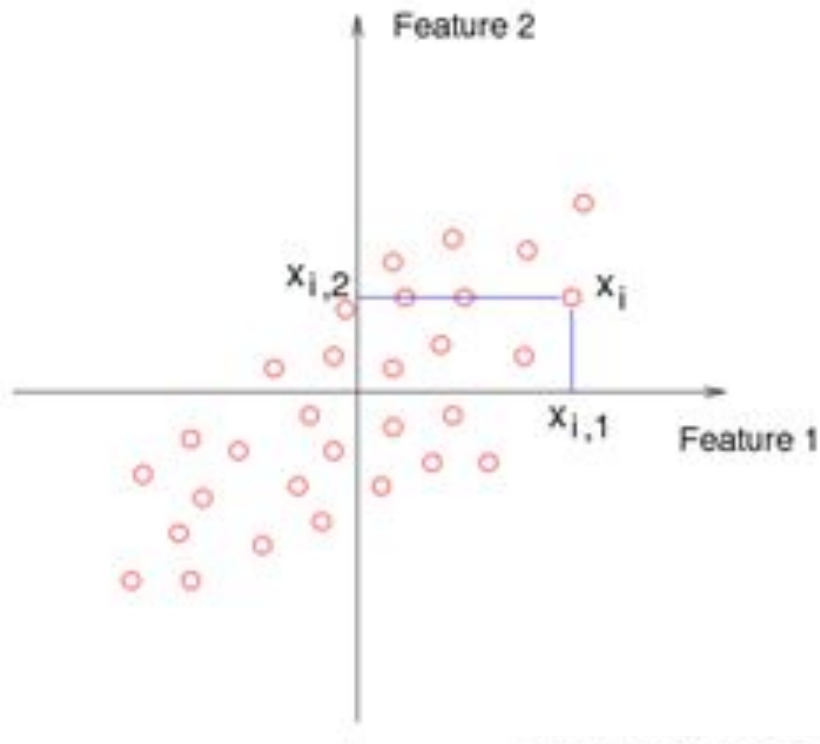|	RedMeat	WhiteMeat	Eggs	Milk	Fish	Cereals	Starch	Nuts	FruitVeg
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechosl.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
EGermany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherl.	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzerl.	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
WGermany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
Yugosl.	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

- Among many possible "new" bases, which one to choose?
  - What is our objective?

- Observations can be represented in this new space (as vectors)
  - Similar observations should appear closer to each other
  - Similarity defined in terms of distance in the feature space
  - Dis-Similar observations should appear further from each other
  - The new dimension should "emphasize" the dissimilarity
  - The new dimension should capture the most variance
  - If observations vary significantly in one of the native dimensions, this dimension should be retained / emphasized
  - If observations do not vary in one of the native dimensions, this dimension should be dropped / de-emphasized

- Delete or Modify?
  - Drop a feature (Reduce Dimension)
    - Drop "height" ➔ Project data onto weight axis
    - All people with the same weight treated identically
  - Change the axes : a new "height-weight" dimension
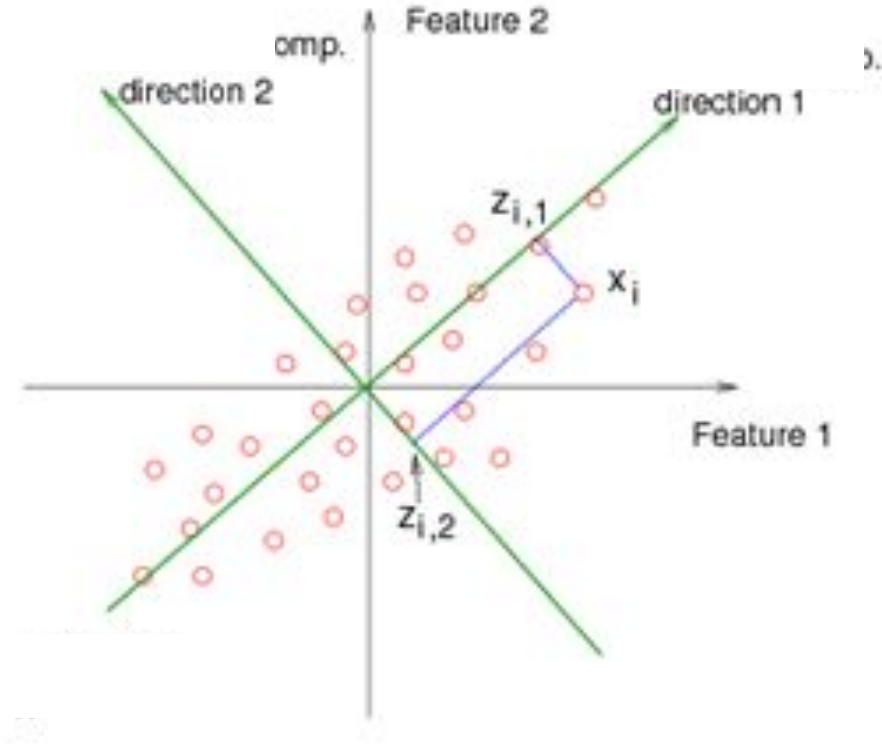   - A dimension of maximum variance explains the essence of the data.

Principal Components as the Optimal basis

- Direction-1
  - Data rows (elements) vary most from each other
  - Most discriminatory power
  - Specified as a linear combination of Feature-1 & 2

- Direction-2
  - Data rows (elements) vary "2nd most"
  - Loss in discriminatory power minimal if features correlated
  - Seed idea of dimensionality reduction!

Principal Components as the Optimal basis (cont'd)

- Naïve basis
  - Reflect the method the data was gathered
  - Not necessarily optimal for analyzing the data
  - As many dimensions as there are variables

- Idea-1 : No variance ➔ No value
  - Get rid of dimensions (features) in which variance is l

- Idea-2 : Covariance ➔ Change basis
  - Combine features to create new pseudo-features
  - Pseudo-features are in the direction of maximum vari
  - Caution: Loss of interpretability

- Change basis so as to
  - Maximize variance (Interesting basis)
  - Find a lower dimensional representation to spot patterns (SOM)
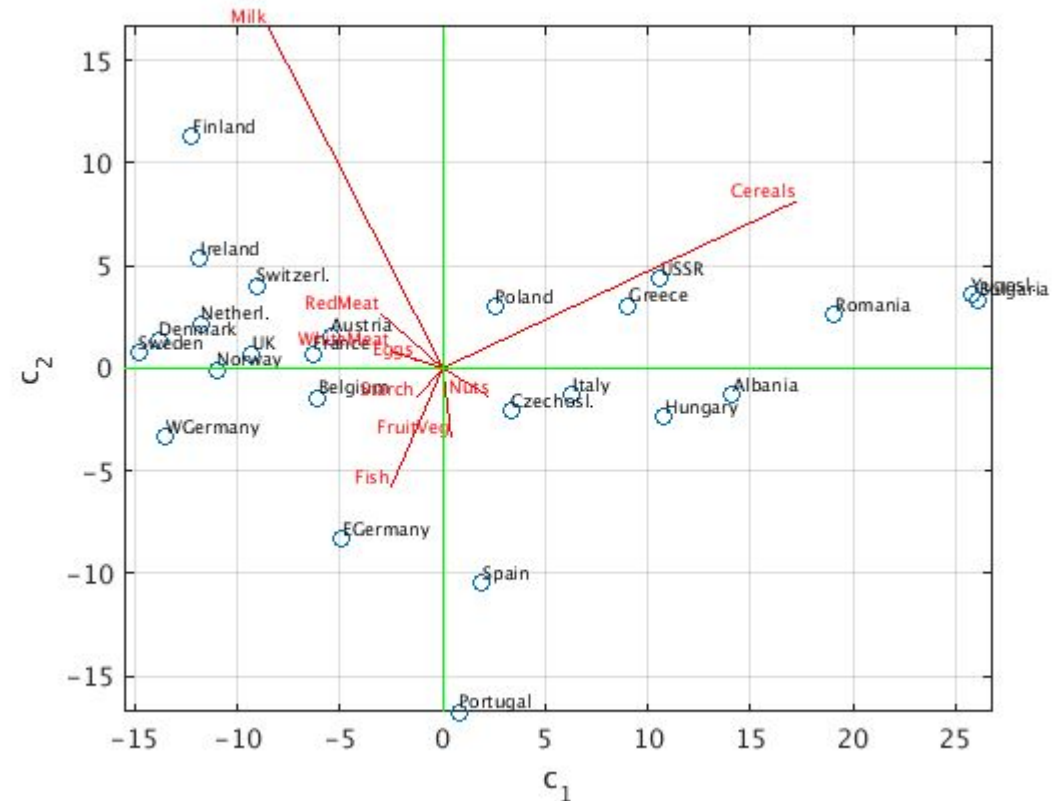
Optimal basis & Principal Components

- The optimal basis
  - is a linear combination of the naive basis (rotate & stretch)
  - maximizes the variance in the data

- The dimension of the optimal basis
  - Is the number of principal components ($k \leq p$)
  - Can be reduced by dropping directions where variance is "low"

- Top-k principal components
  - Approximate the data
  - Explain what percentage of variance
  - Preserve the structure in the data (The Elephant)

- How is this different from linear regression?
  - LR : Determine a line of best fit to a data set,
  - PCA: Determine several orthogonal lines of best fit to the data set.

Algebra: orthonormal transform
Geometry: axis rotation
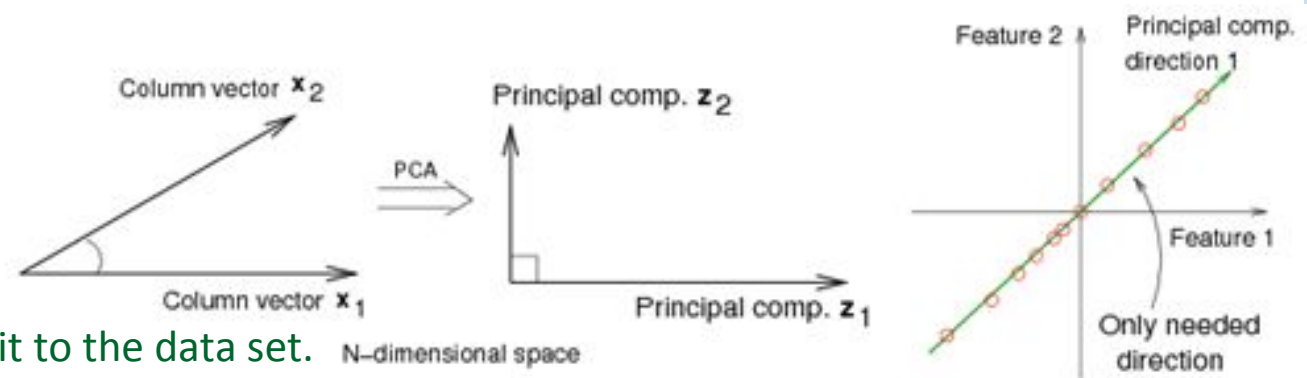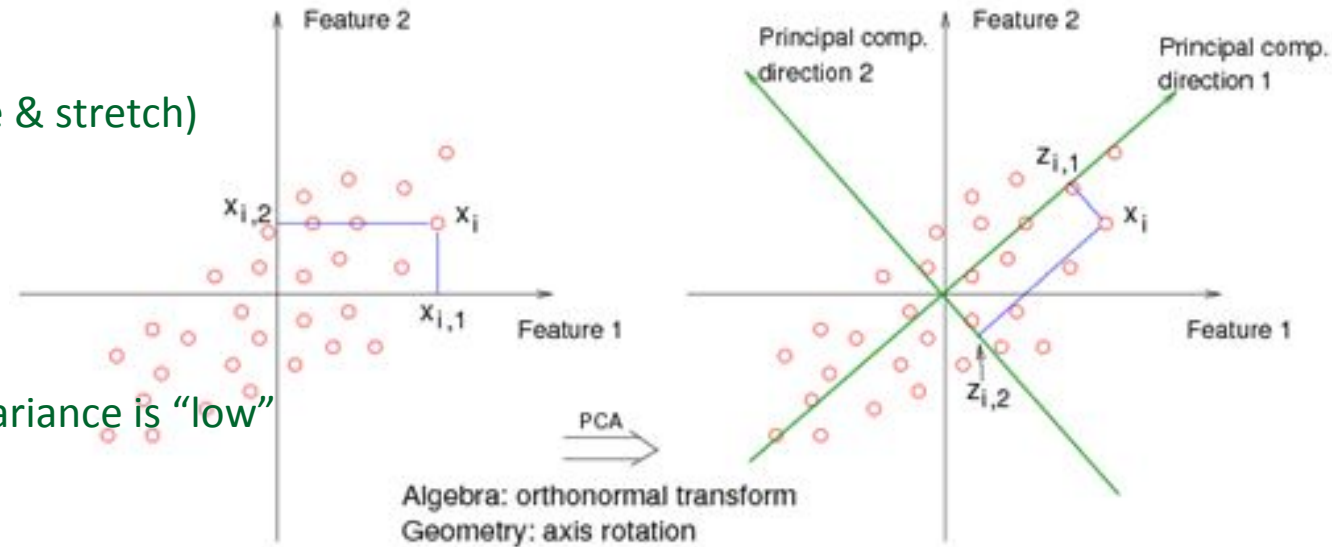
An Example : revisited

name	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m	Rank	Points	Competition
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.7	1	8217	Decastar
CLAY	10.76	7.4	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.5	2	8122	Decastar
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.1	4	8067	Decastar
YURKOV	11.34	7.09	15.19	2.1	50.42	15.31	46.26	4.72	63.44	276.4	5	8036	Decastar
ZSIVOCZKY	11.13	7.3	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268	7	8004	Decastar
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.1	8	7995	Decastar
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.6	4.92	52.33	262.1	9	7802	Decastar
HERNU	11.37	7.56	14.41	1.86	51.1	15.06	44.99	4.82	57.19	285.1	10	7733	Decastar
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.1	4.72	55.4	282	11	7708	Decastar
NOOL	11.33	7.27	12.68	1.98	49.2	15.29	37.92	4.62	57.44	266.6	12	7651	Decastar
BOURGUIGN	11.36	6.8	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.7	13	7313	Decastar
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5	70.52	280.01	1	8893	OlympicG
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.9	69.71	282	2	8820	OlympicG
Karpov	10.5	7.81	15.93	2.09	46.81	13.97	51.65	4.6	55.54	278.11	3	8725	OlympicG
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.4	58.46	265.42	4	8414	OlympicG
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.9	55.39	278.05	5	8343	OlympicG
Zsivoczky	10.91	7.14	15.31	2.12	49.4	14.95	45.62	4.7	63.45	269.54	6	8287	OlympicG
Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.8	57.76	264.35	7	8237	OlympicG
Bernard	10.69	7.48	14.8	2.12	49.13	14.17	44.75	4.4	55.27	276.31	9	8225	OlympicG
Schwarzl	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.1	56.32	273.56	10	8102	OlympicG
Pogorelov	10.95	7.31	15.1	2.06	50.79	14.21	44.6	5	53.45	287.63	11	8084	OlympicG
Schoenbeck	10.9	7.3	14.77	1.88	50.3	14.34	44.41	5	60.89	278.82	12	8077	OlympicG
Barras	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.6	64.55	267.09	13	8067	OlympicG
KARPOV	11.02	7.3	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.2	3	8099	Decastar
WARNERS	11.11	7.6	14.31	1.98	48.68	14.23	41.1	4.92	51.77	278.1	6	8030	Decastar
Nool	10.8	7.53	14.26	1.88	48.81	14.8	42.05	5.4	61.33	276.33	8	8235	OlympicG
Drews	10.87	7.38	13.07	1.88	48.51	14.01	40.11	5	51.53	274.21	19	7926	OlympicG

Finding the principal components

Eigen & Singular Value Decomposition

Matrix Decomposition

- Why decompose a data matrix?
  - Express the matrix in different basis
  - X = XI : Native (Naïve) Basis
  - XI = UV : Data (X) expressed in Native Basis (I) is expressed as a linear combination (V) in a new basis (U)
  - Desirable: Orthogonal basis : uncorrelated features

- Principal Component Analysis
  - Aim is to create new columns (dimensions, features, axes) which are a **linear combination of the existing columns**
  - X = UV : Column view of multiplication: **X is a linear combination (V) of columns of U**

- Dimensionality Reduction
  - What if we drop some columns of U?
  - Not equivalent to dropping some columns of X.
  - Effectively drop those new dimensions which have the least variance (If we can rank the columns of U by variance)
  - Instead of p columns in the native basis, preserve only k columns in new basis : find the "best" k-column rank approximation of X

PCA Dimensionality Reduction using Singular Value Decomposition

- PCA
  - U contains the principal **components** : directions aligned with variance
  - S contains a measure of how much **variance** is explained by each of the new components
  - V contains the linear combinations of these new components (**loadings**) to recover data in original basis

$$X = USV^T$$
$$= u_1 s_1 v_1^T + u_2 s_2 v_2^T + \ldots + u_p s_p v_p^T$$
$$\approx u_1 s_1 v_1^T + \ldots + u_k s_k v_k^T$$

$$X \approx U \; S \; V^T$$

X: n x p   U: n x k   S: k x k V^T: k x p
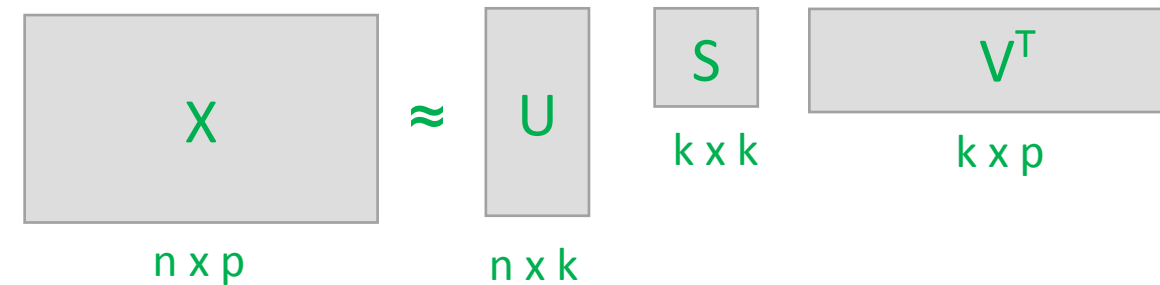
$$u_1, u_2, \ldots, u_k$$

Orthonormal basis for the columns of X…

$$s_1 \geq s_2 \geq \ldots s_k$$

…sorted in order of explained variance

- SVD obtains the best low-rank approximation of X
 - k < p : Which k to keep?

$$\min_{rank_{A_k} \leq k} ||A - A_k||_F^2$$

$$||X||_2 = \left(\sum_{i=1}^{n} \sum_{j=1}^{p} |x_{ij}|^2 \right)^{\frac{1}{2}} = ||X||_F$$

Principal Component Analysis : The view from Linear Algebra

- How to find the principal components of a data matrix X?

- Linear Algebra provides very powerful mathematical tools
  - To analyze, manipulate a matrix
  - To make computations efficient

- The principal components are the singular values of the data matrix.
  - What?

- The principal components are the eigenvectors of the covariance matrix of the data.
 - What??
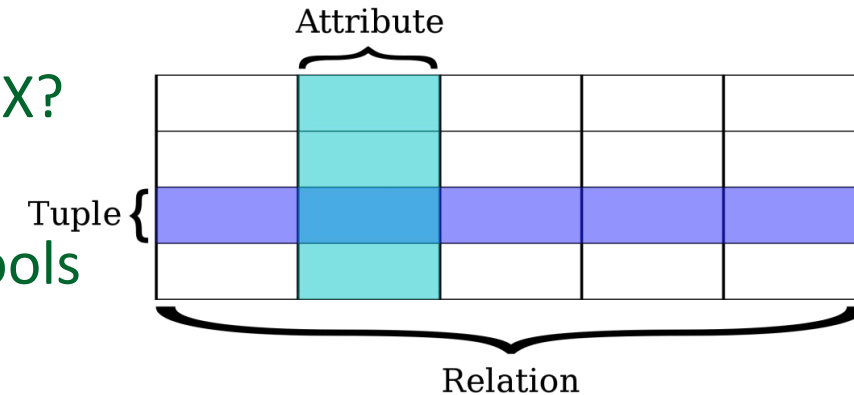
Principal Component Analysis: Summary

- Principal components of X
  - Determine the "directions" / "coordinate-system" / "basis" in which the data displays the most variance
  - Provide an approach to reduce the dimensionality of data (k ≤ p)
  - This lower dimensional space can be used to visualize clusters (Elephant). Why?
  - Optimal-k : Top-k principal components explain what percent of variance? Scree pl

$$X \approx U \, S \, V^T$$

X: $n \times p$
U: $n \times k$
S: $k \times k$
V^T: $k \times p$
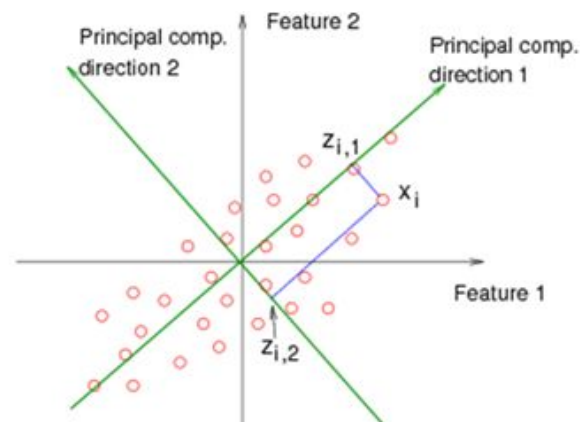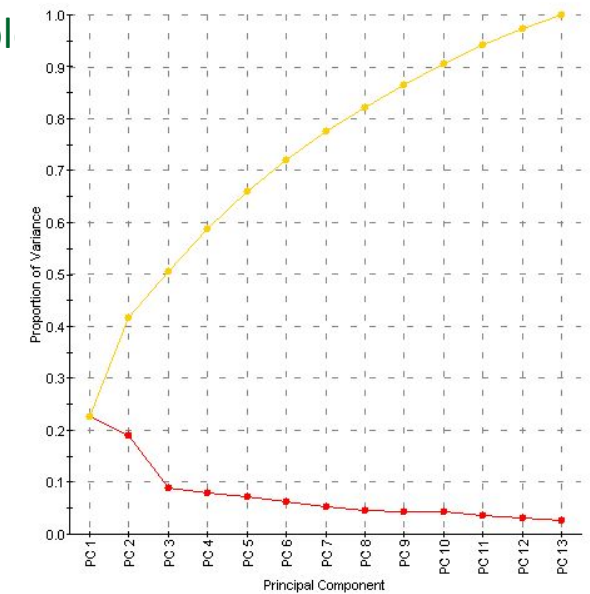
- Can be found by Singular Value Decomposition of X
  - X must be normalized (column-wise) before SVD (since different columns may be expressed in different units)
  - Equivalent to finding the eigenvectors of the covariance matrix (X^TX)
  - A non-parametric approach to dimensionality reduction
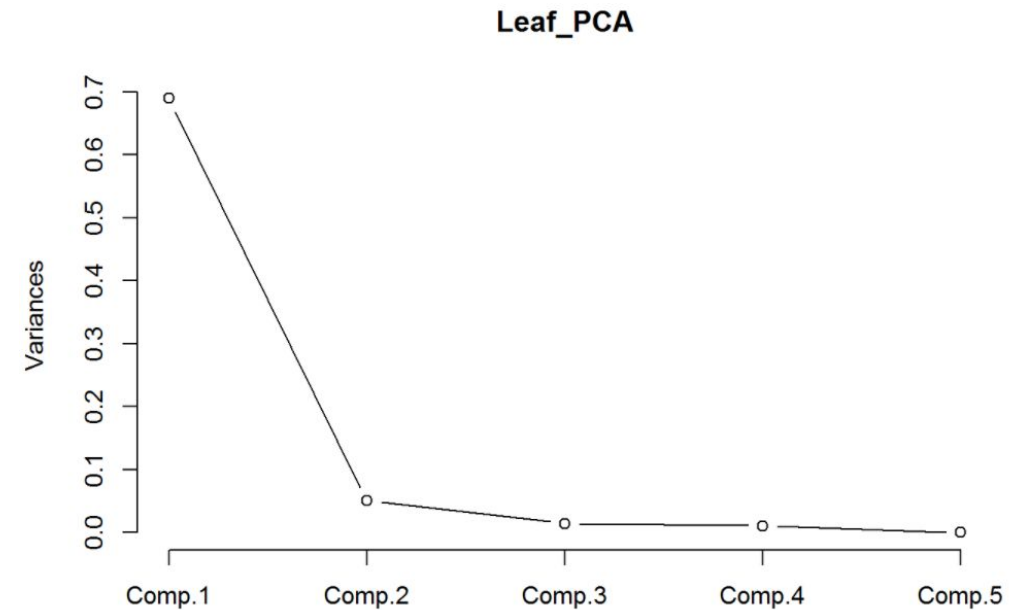
Example

```
summary(Leaf_PCA)
```

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation      0.8302248 0.22418865 0.11987329 0.1035367
## Proportion of Variance 0.9013599 0.06572552 0.01879107 0.0140183
## Cumulative Proportion  0.9013599 0.96708539 0.98587647 0.9998948
##                               Comp.5
## Standard deviation      0.0089705579
## Proportion of Variance 0.0001052315
## Cumulative Proportion  1.0000000000
```

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
...					
150	5.9	3.0	5.1	1.8	virginica

```
screeplot(Leaf_PCA, type = 'lines')
```


http://environmentalcomputing.net/principal-components-analysis/

PCA: Example (cont'd)

```
loadings(Leaf_PCA)
```

- Loadings
  - Correlations between the principal components and the original variables (Pearson's r).
  - Values closest to 1 (positive) or -1 (negative) will represent the strongest relationships, with zero being uncorrelated.

- Scores

```
              Comp.1        Comp.2        Comp.3        Comp.4
 [1,] -2.684125626  -0.319397247   0.027914828   0.0022624371
 [2,] -2.714141687   0.177001225   0.210464272   0.0990265503
 [3,] -2.888990569   0.144949426  -0.017900256   0.0199683897
 [4,] -2.745342856   0.318298979  -0.031559374  -0.0755758166
 [5,] -2.728716537  -0.326754513  -0.090079241  -0.0612585926
 [6,] -2.280859633  -0.741330449  -0.168677658  -0.0242008576
 [7,] -2.820537751   0.089461385  -0.257892158  -0.0481431065
 [8,] -2.626144973  -0.163384960   0.021879318  -0.0452978706
 [9,] -2.886382732   0.578311754  -0.020759570  -0.0267447358
[10,] -2.672755798   0.113774246   0.197632725  -0.0562954013
...
[150,]  1.390188862   0.282660938  -0.362909648  -0.1550386282
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Total_length          0.772  0.244          0.582
## Petiole_length        0.458 -0.169  0.647 -0.586
## Leaf_length           0.320  0.428 -0.627 -0.564
## Width1         0.949  0.160 -0.215 -0.163
## Width2         0.300 -0.259  0.826  0.400
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

PCA : Summary

- original_data =~ approximation = (scores * loadings) * scale + center

- where:
    - `scores` are the coordinates in your new orthogonal base
    - `loadings` are the directions of the new axis in the old base
    - `scale` are the scaling applied to the dimensions
    - `center` are the coordinates of the new base origin in the old base

- Assumptions
    - Scaling & Centering (as pre-processing or argument)
    - Outliers impact
    - Linearity

Q?

Praphul Chandra