Factorial Methods

K. Gibert[1,2]

[1] *Department of Statistics and Operation Research*

[2] *Knowledge Engineering and Machine Learning group*
Universitat Politècnica de Catalunya, Barcelona

Master Oficial en Enginyeria Informàtica
Universitat Politècnica de Catalunya

Factorial Methods

- Find the isomorph transformation from original space

  keeps the adjacency relationships among variables

- Results expressed in a ficticious space

- Might produce interpretation problems

- Methods
  - PCA (Principal components analysis)
  - Simple correspondence analysis
  - Multiple correspondence analysis

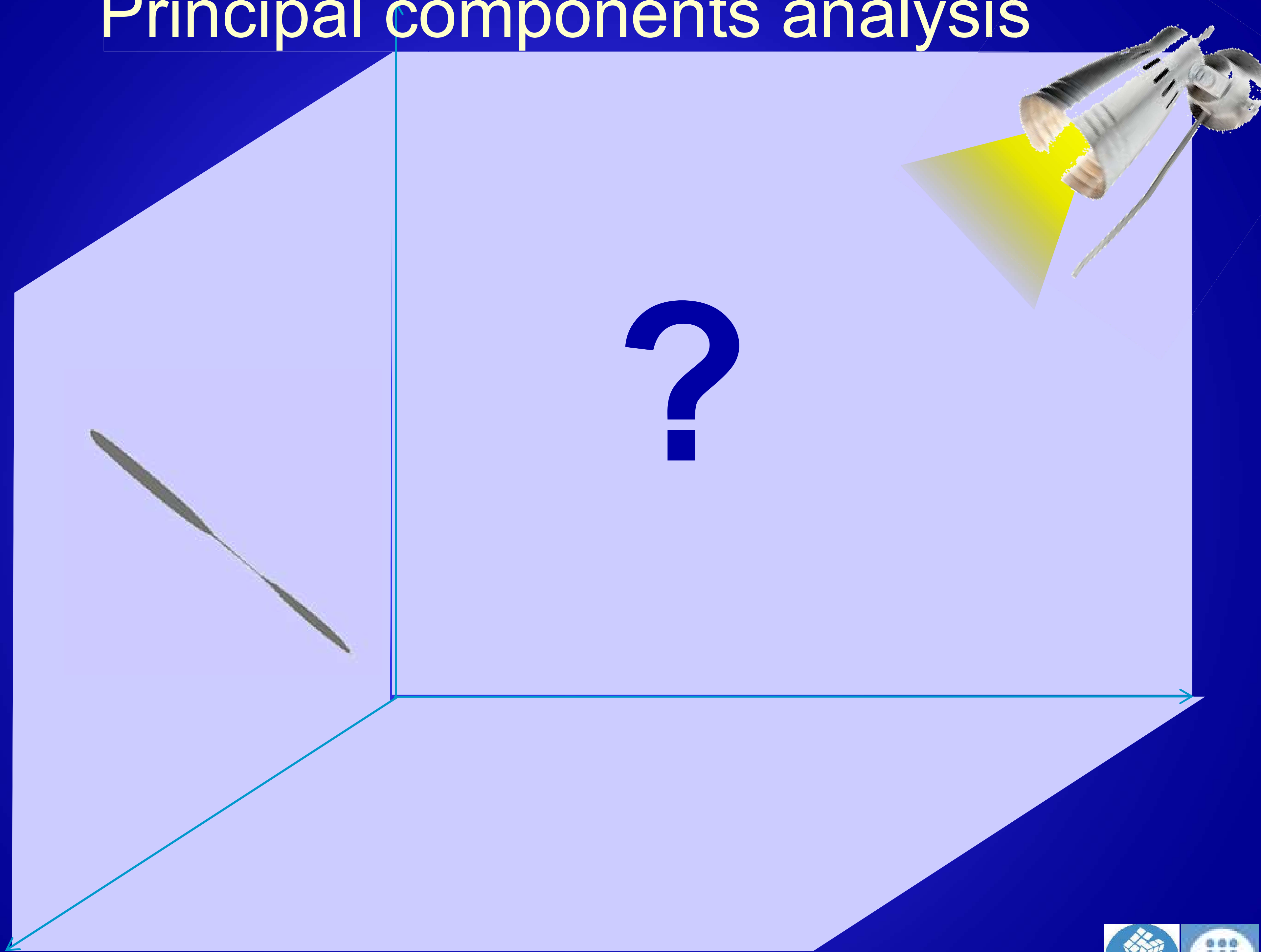
Factorial Methods

- Principal Components Analysis
  - Only numerical variables
  - Find the most informative projection planes

    (factorial planes)

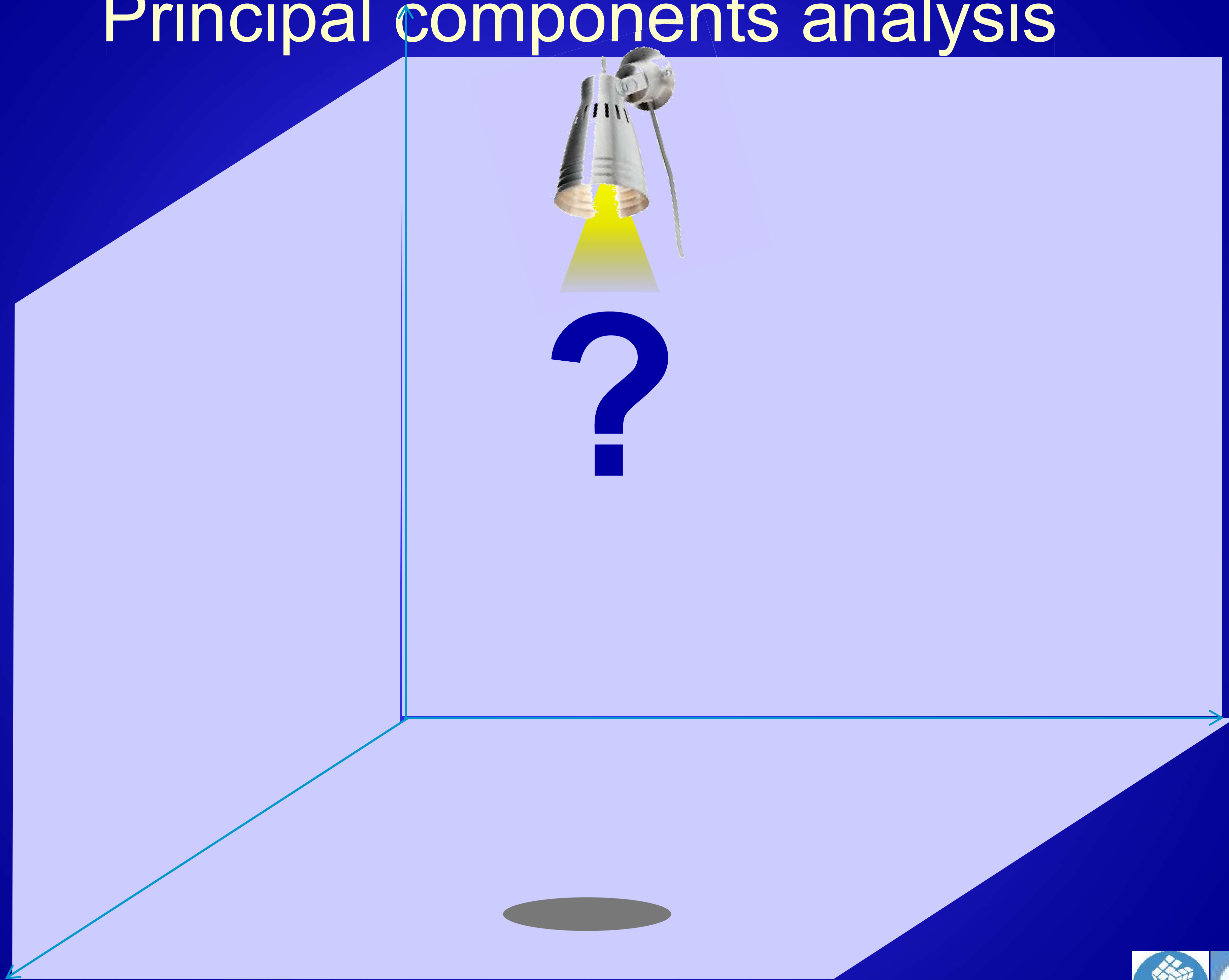
Example "Copas"

Principal components analysis

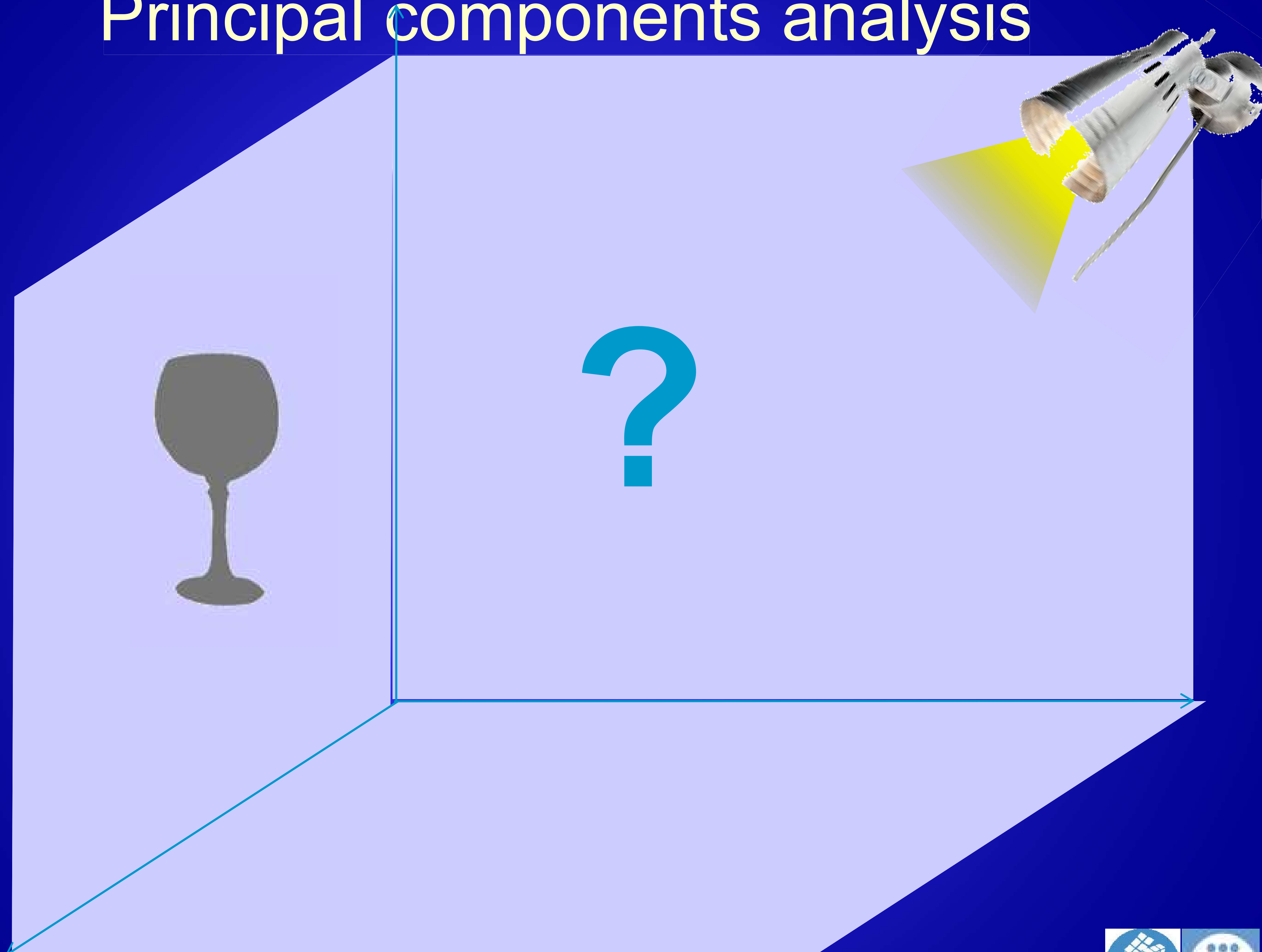


©K. Gibert

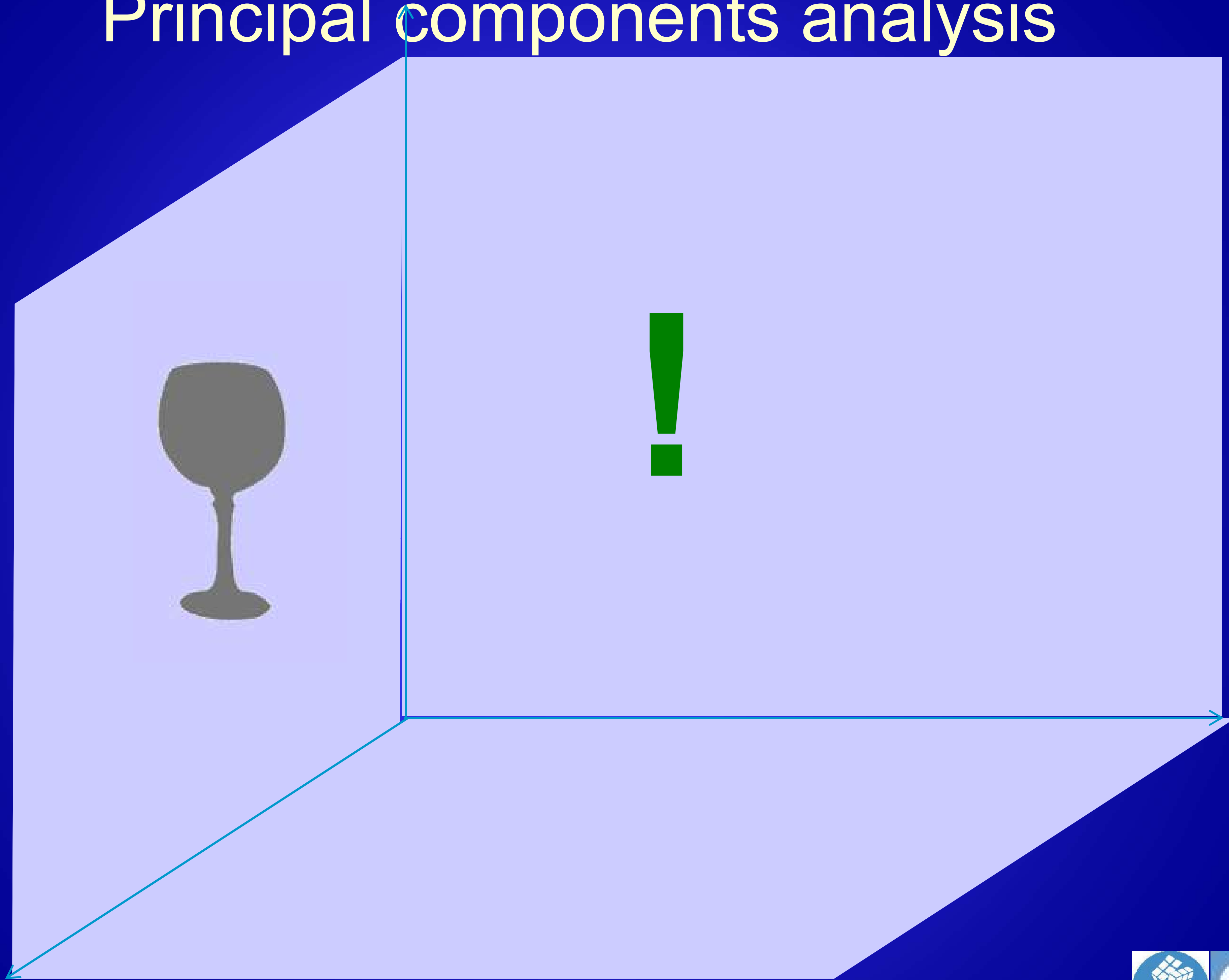
Principal components analysis



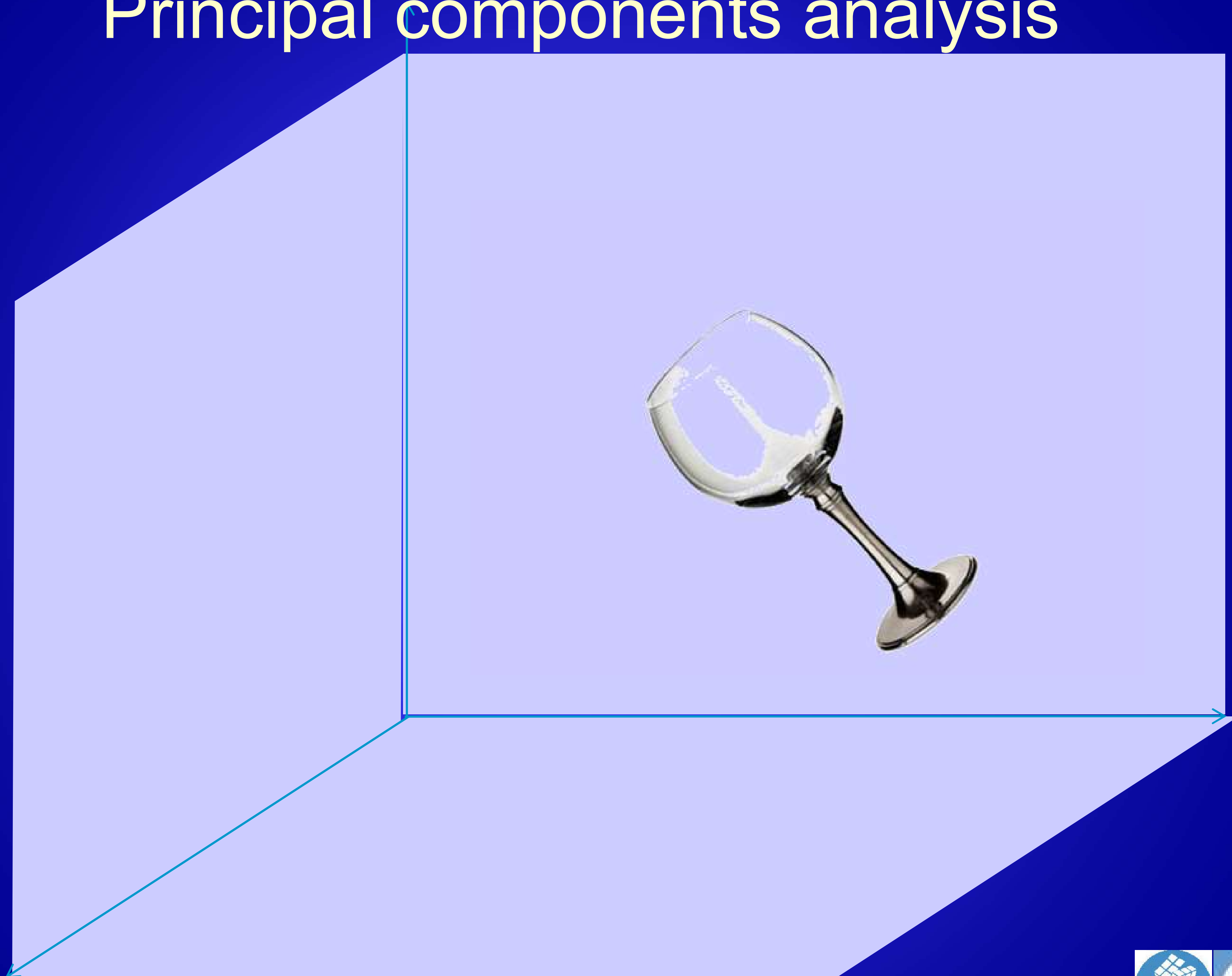
Principal components analysis



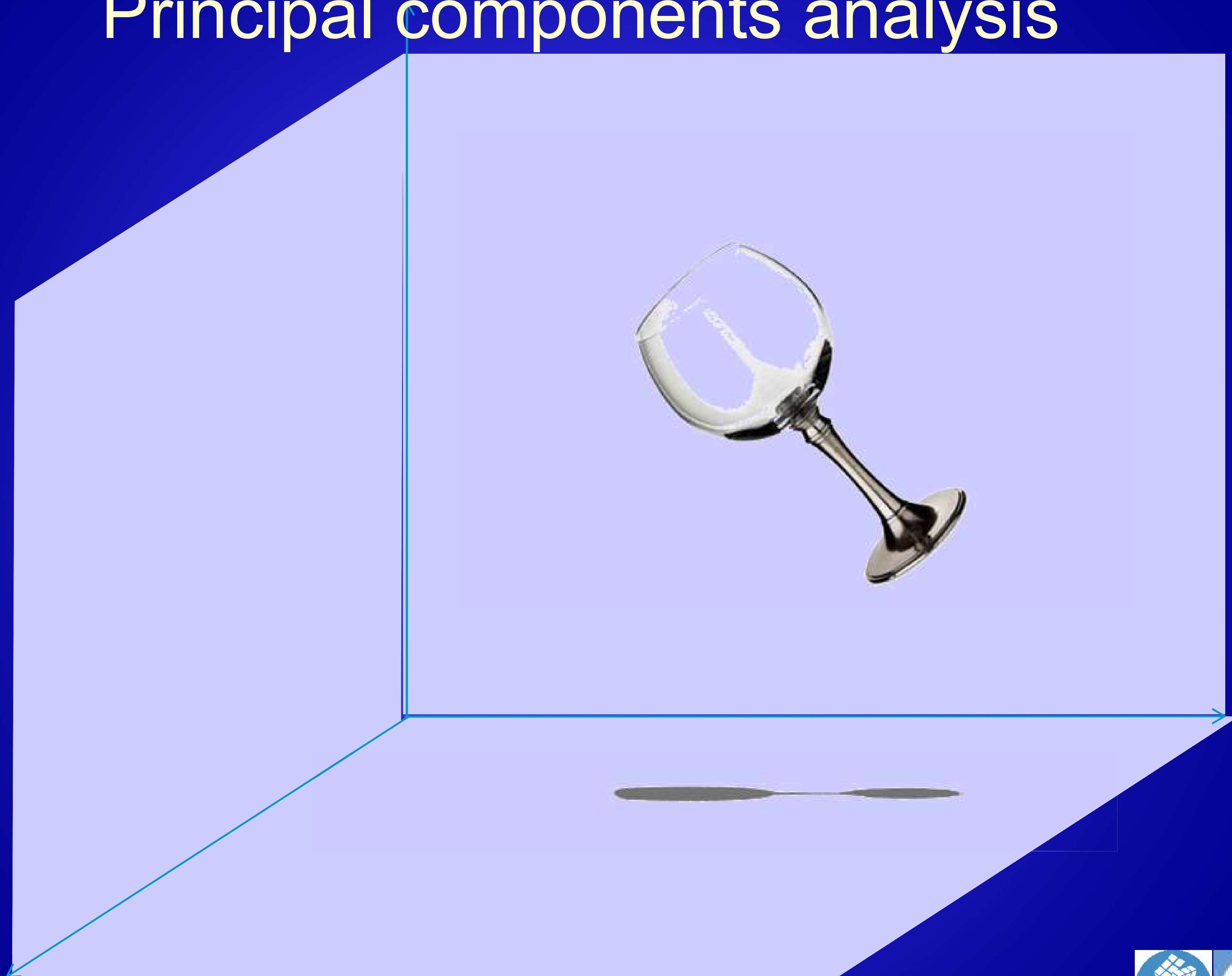
Principal components analysis



Principal components analysis

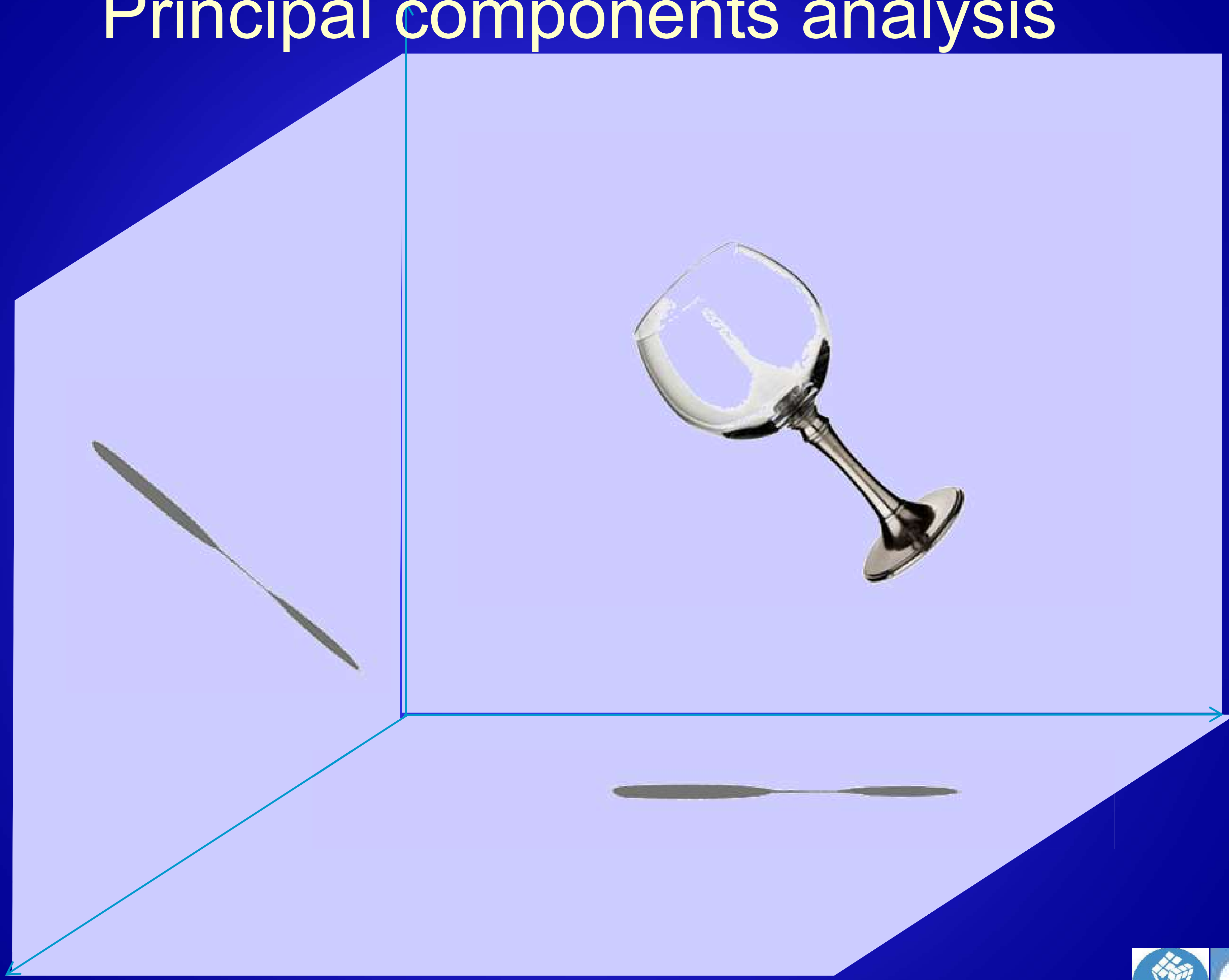


Principal components analysis

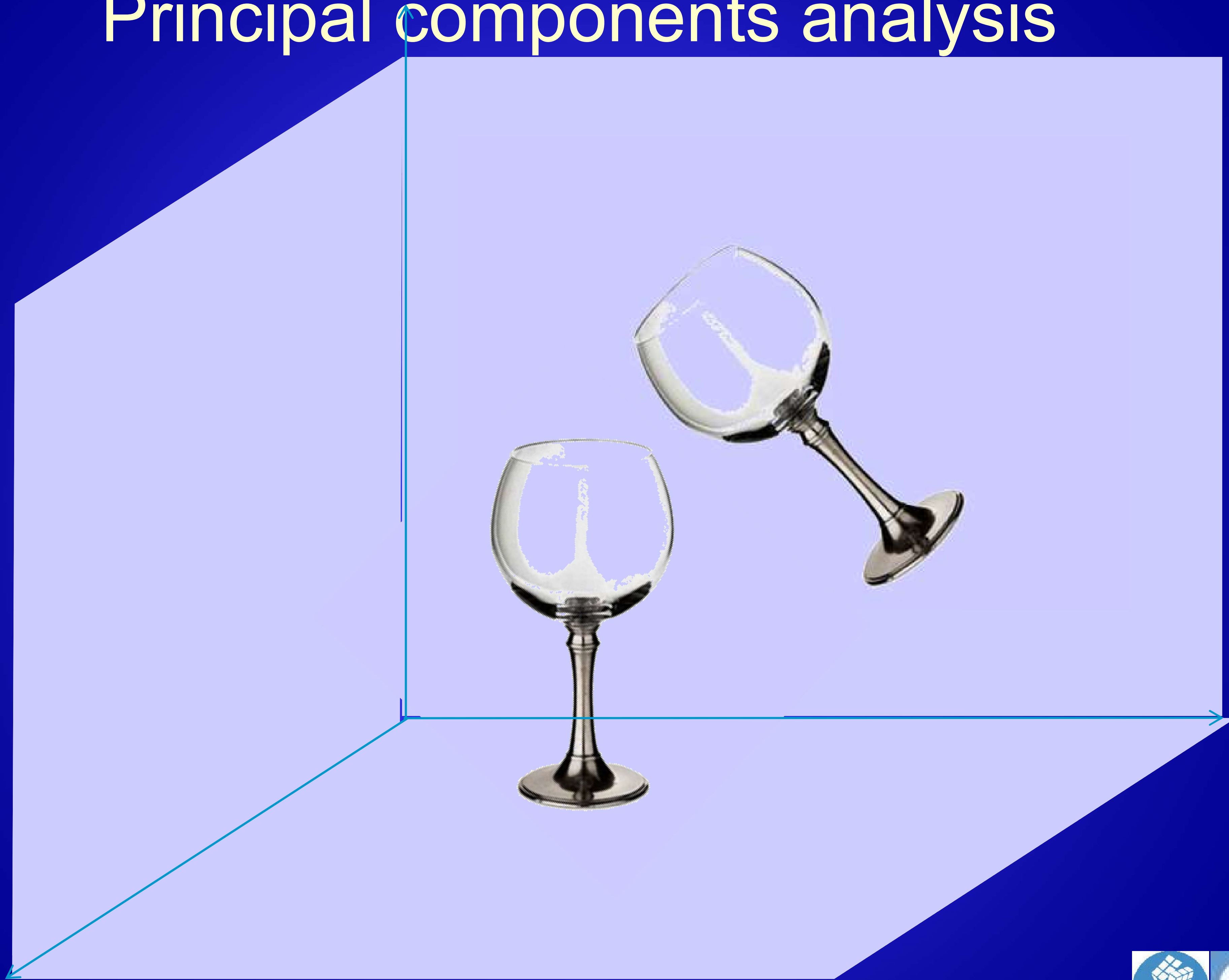


©K. Gibert

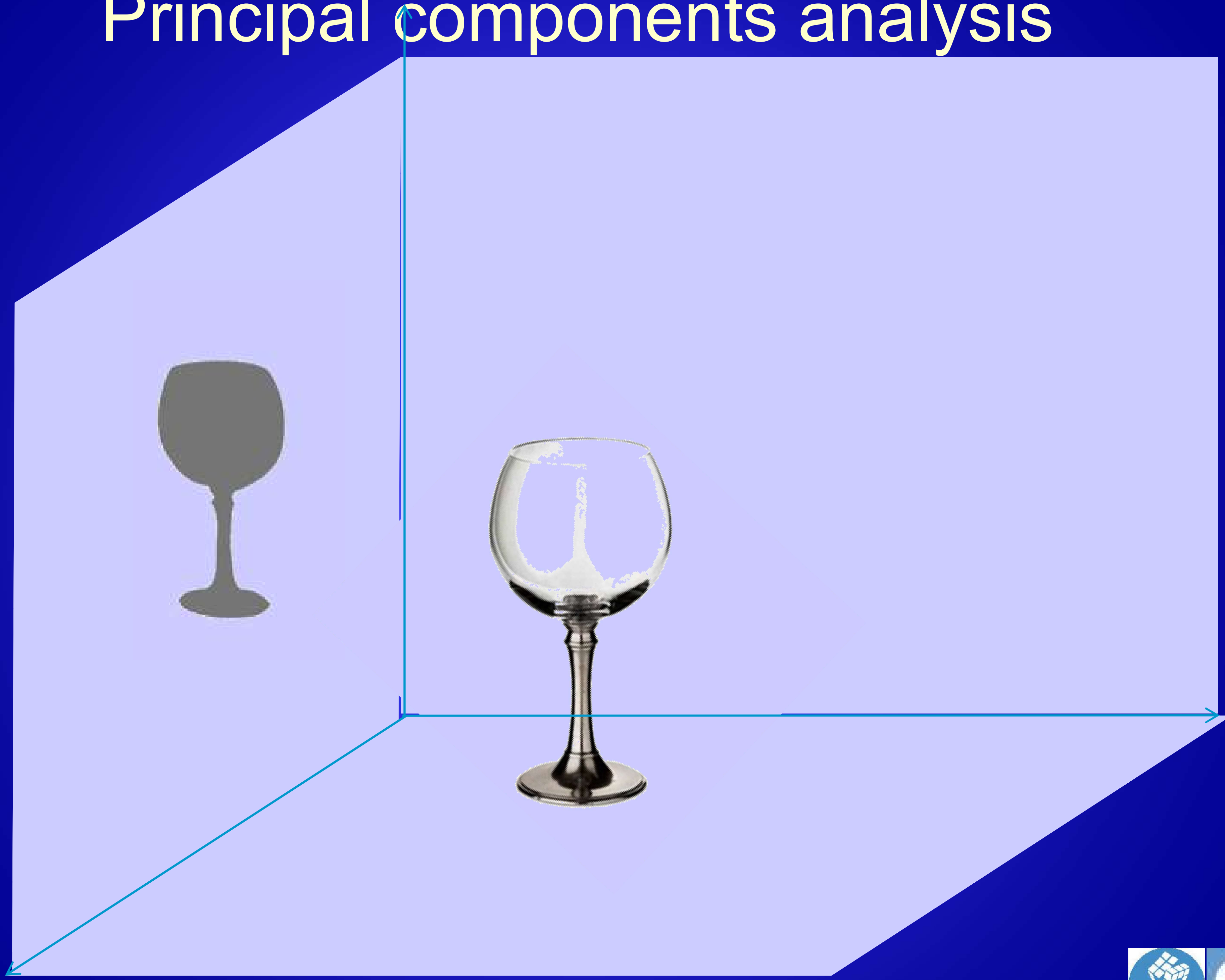
Principal components analysis



Principal components analysis

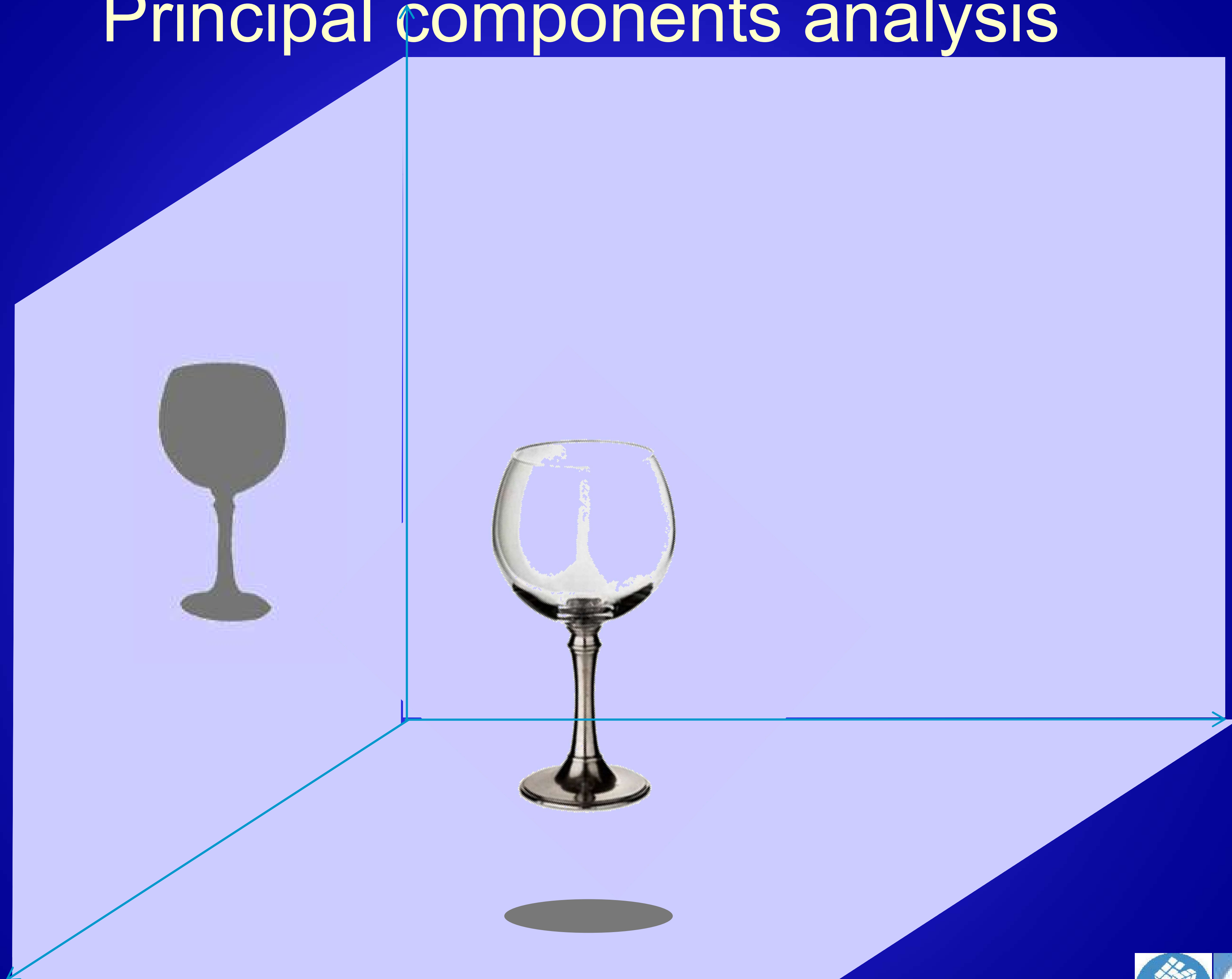


Principal components analysis



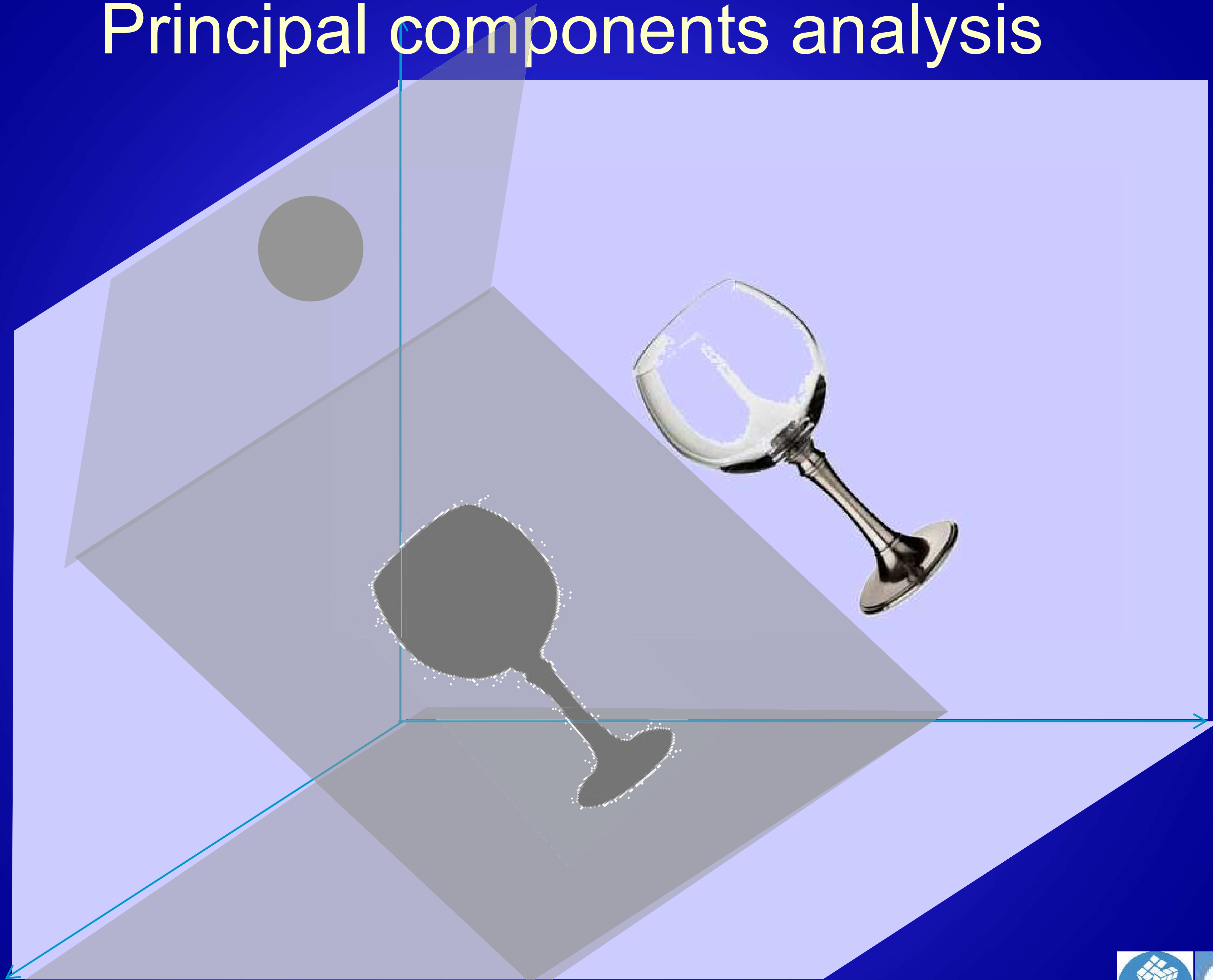
©K. Gibert

Principal components analysis

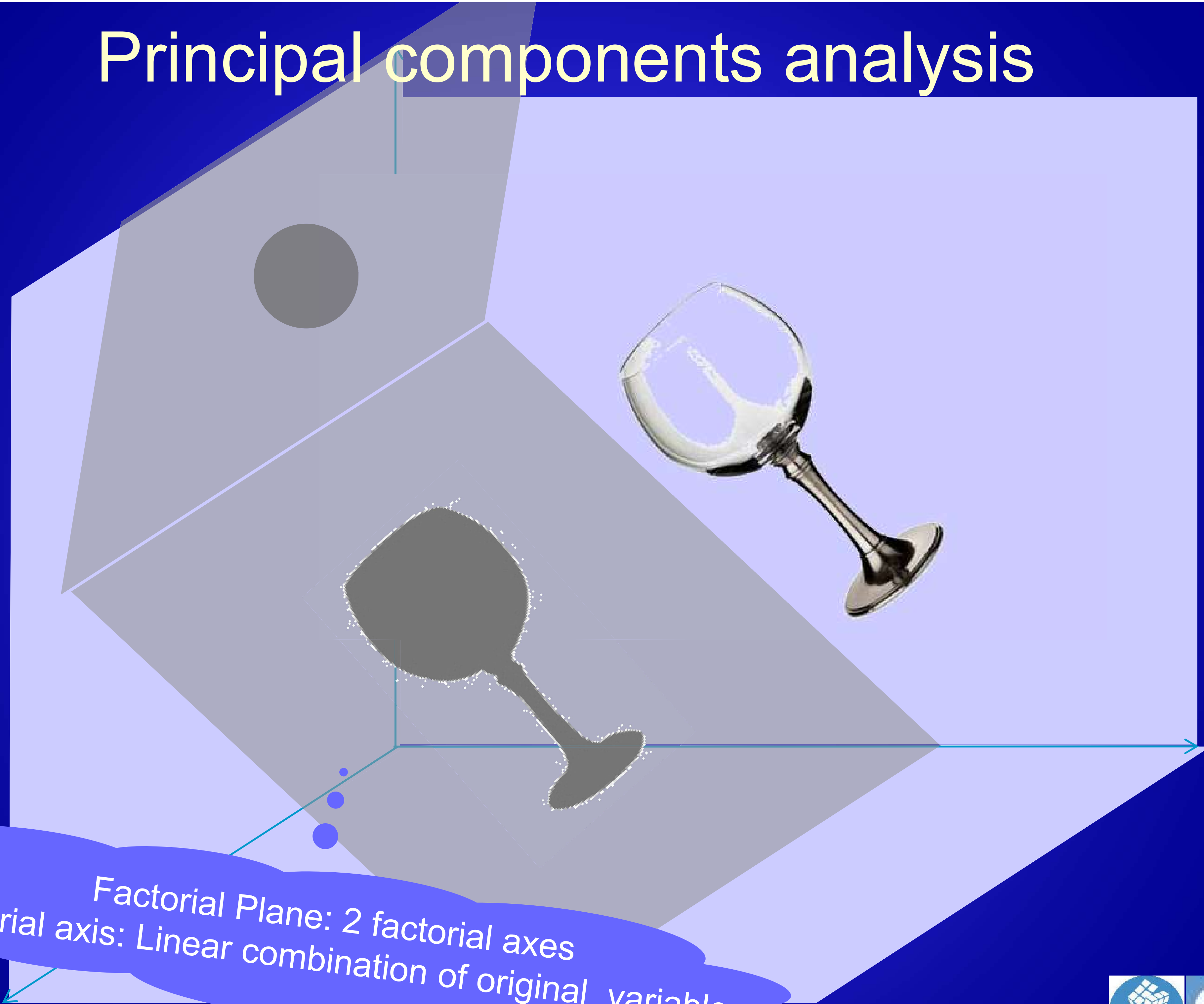


©K. Gibert

Principal components analysis

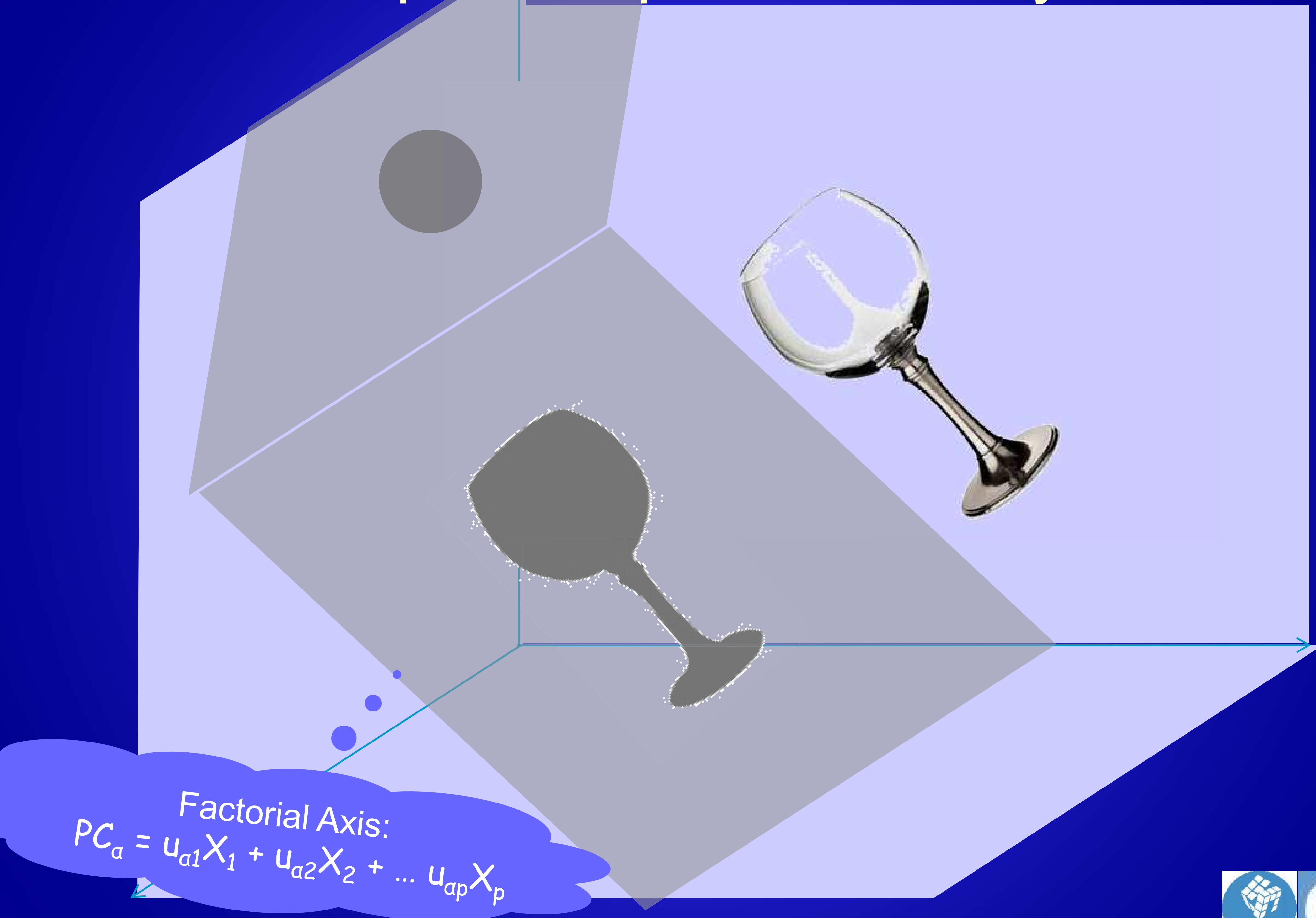


Principal components analysis



Factorial Plane: 2 factorial axes
Factorial axis: Linear combination of original variables

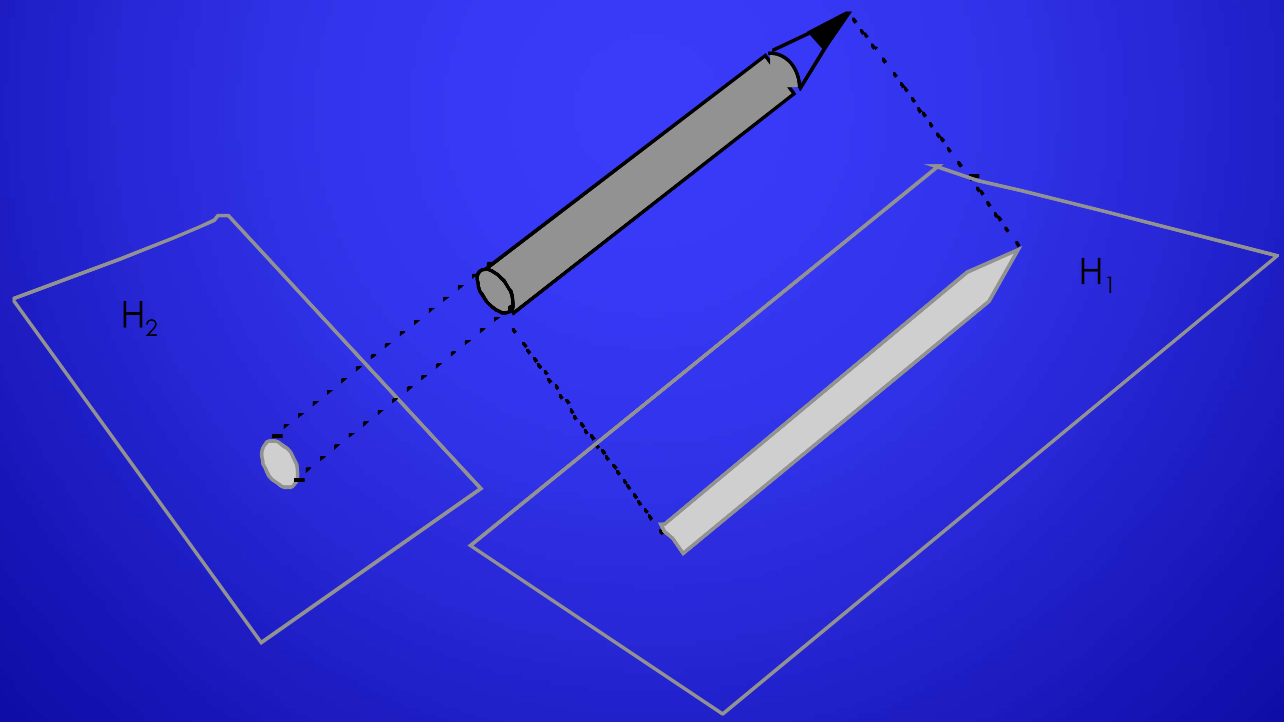
Principal components analysis



Principal components analysis

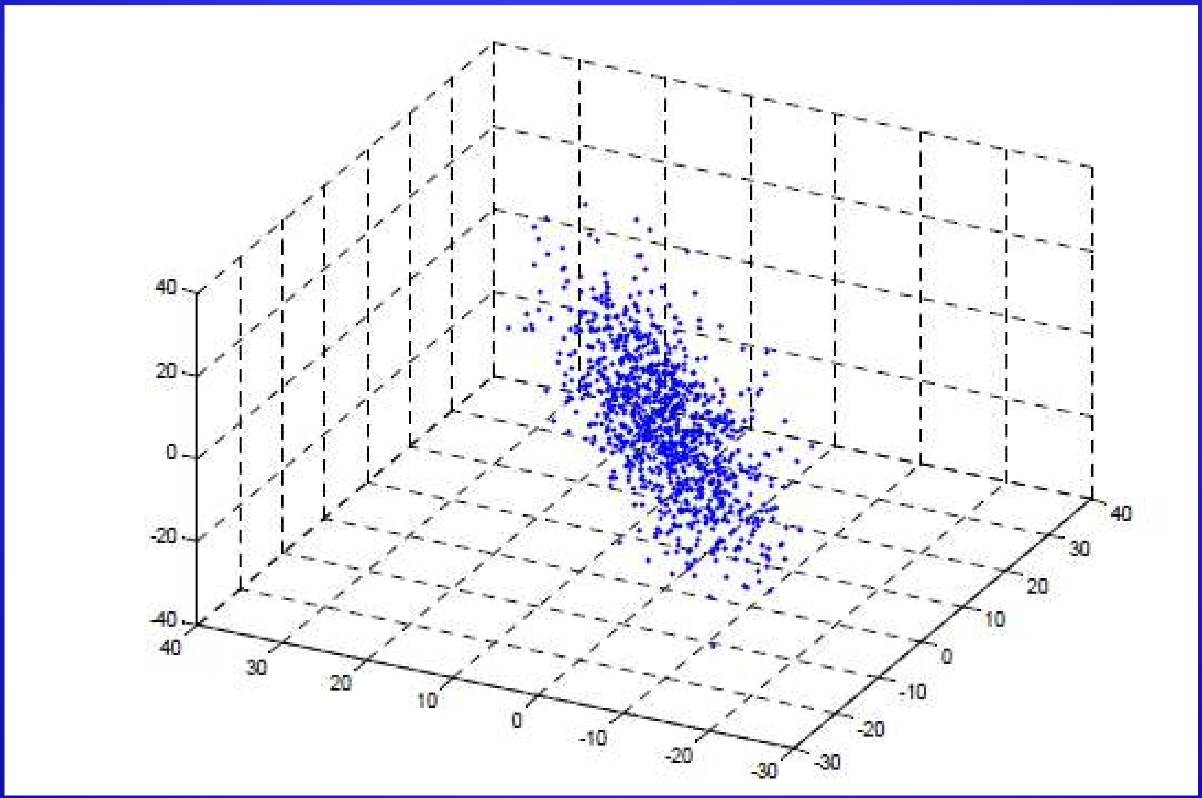
- Purpose:
  - To project the cloud of points upon a subspace (plane) retaining as much original cloud information.

    (see video)



Principal components analysis

- Find the most informative projection planes of data cloud *(factorial planes)*



Factorial Methods

- Output: K factors rotating original X variables
- Factors: Linear combinations of original variables

Several uses:

- As an associative data mining method:
  analyze relationships among variables
  Project variables and modalities and find associations

Factorial Methods

- Output: K factors rotating original X variables
- Factors: Linear combinations of original variables

Several uses:

- As an associative data mining method to analyze relationships among variables
  Project variables and modalities and find associations

- As a preprocessing method for elicitation of latent variables
  Project active and illustrative variables/individuals on first/second factorial plan and interpret factors (find latent variables)

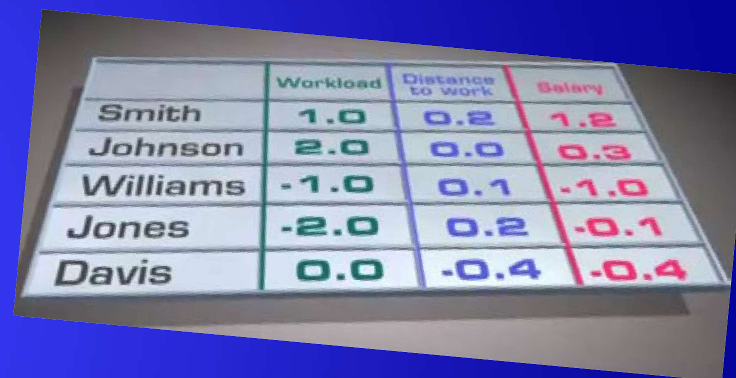
- As a preprocessing method for multidimensionality reduction

Factorial Methods

Data	Factorial Method
Continuous variables	Principal Component Analysis PCA
Contingency table	(Simple) Correspondence Analysis CA
Categorical variables	Multiple Correspondence Analysis MCA

Factorial Methods

- Principal Components Analysis
  - Only numerical variables
  - Find the most informative projection planes (factorial planes, maximize projected inertia)



	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

Given <X,M,D>

- A data matrix X (nxp) centered
- A matrix of individuals weights D (nxn)
- Assume euclidean metrics to compare individuals ($M= \mathbb{I}_p$)

Si les dades estan centrades l'angle entre dues variables projectades coincideix amb la correlació entre elles

Matrix $M^{1/2} X'DXM^{1/2}$

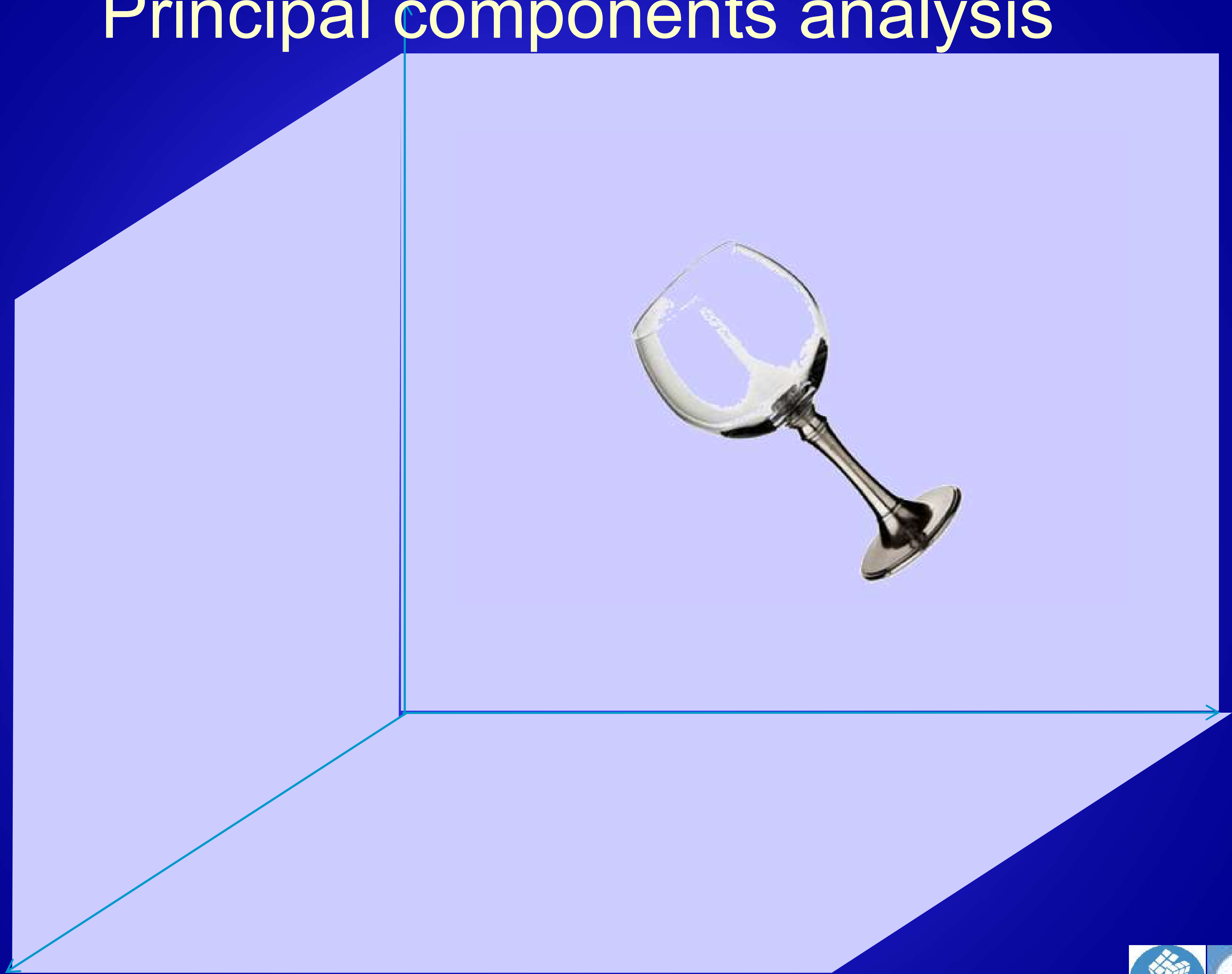
- Product of data with the two metrics
- Simetric,
- Semidefinite
- Catches relationships and opositions of data

Factorial Methods

Given triplet <X,M,D>, diagonalize $M^{1/2} X'DXM^{1/2}$

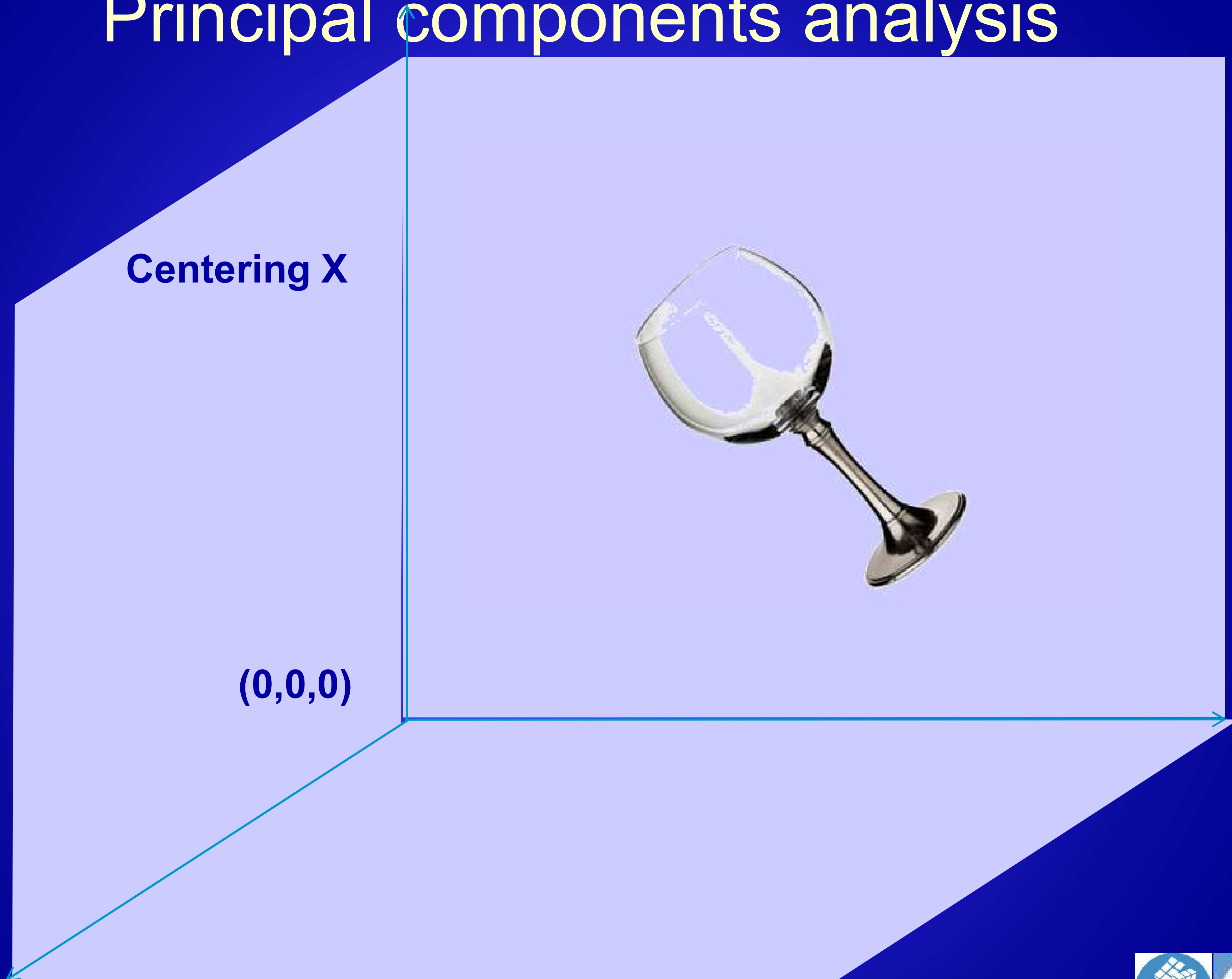
Data	Factorial Method	X	M	D
Continuous variables	PCA	Centered data matrix	$\mathbb{I}_p$	$\mathbb{I}_n$
Contingency table (n_{ij})	CA	$F=(n_{ij}/n_i)$	$diag(1/f_j)$	$diag\ (f_i)$
		$G=(n_{ij}/n_j)$	$diag(1/f_i)$	$diag\ (f_j)$
Categorical variables	MCA	$F=(f_{ij}/(f_i/\sqrt{f_j}))$	$\mathbb{I}_p$	$diag\ (f_i)$
		Burt table	$\mathbb{I}_{n+p}$	$diag(n_{ij})$

Principal components analysis

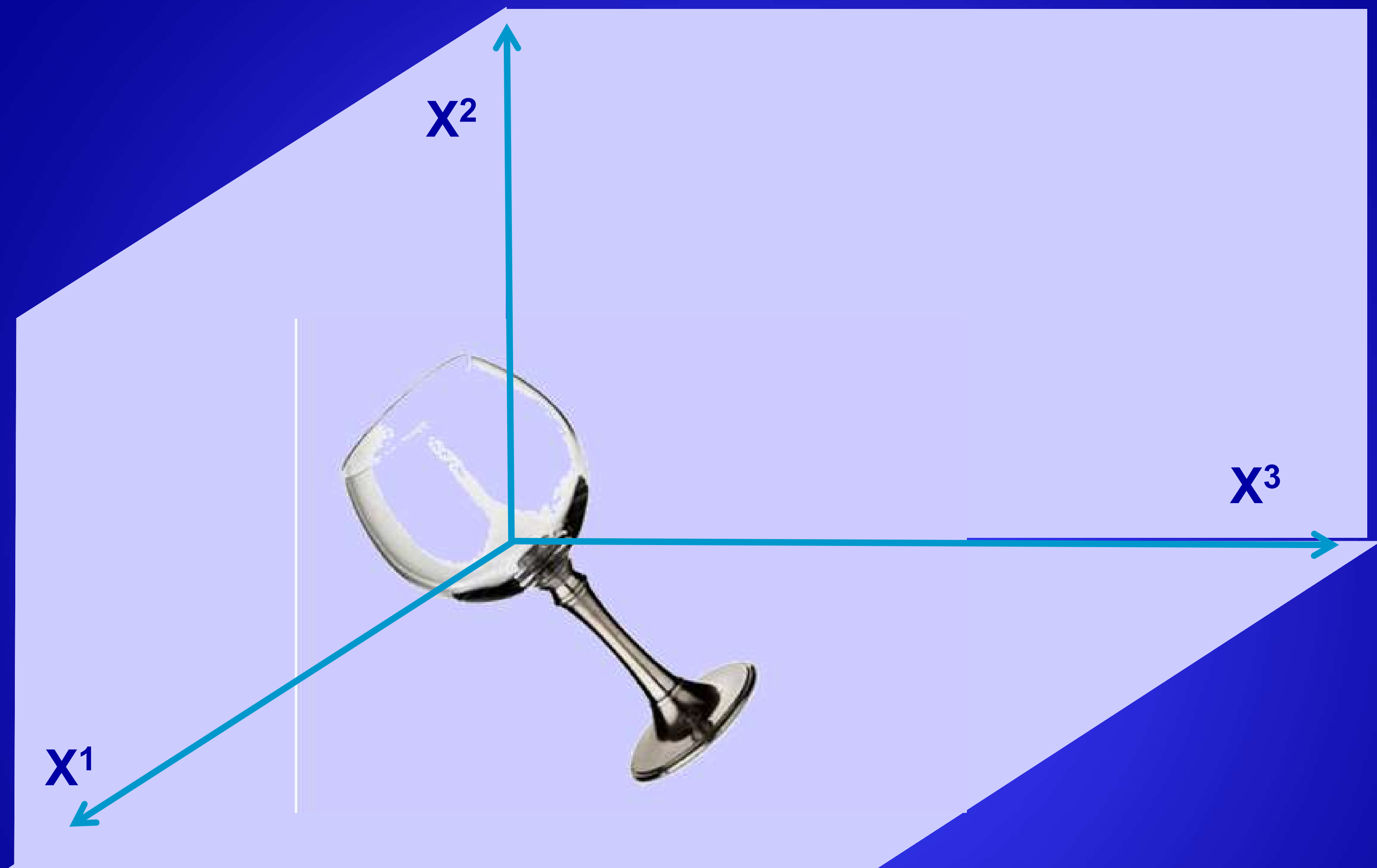


©K. Gibert

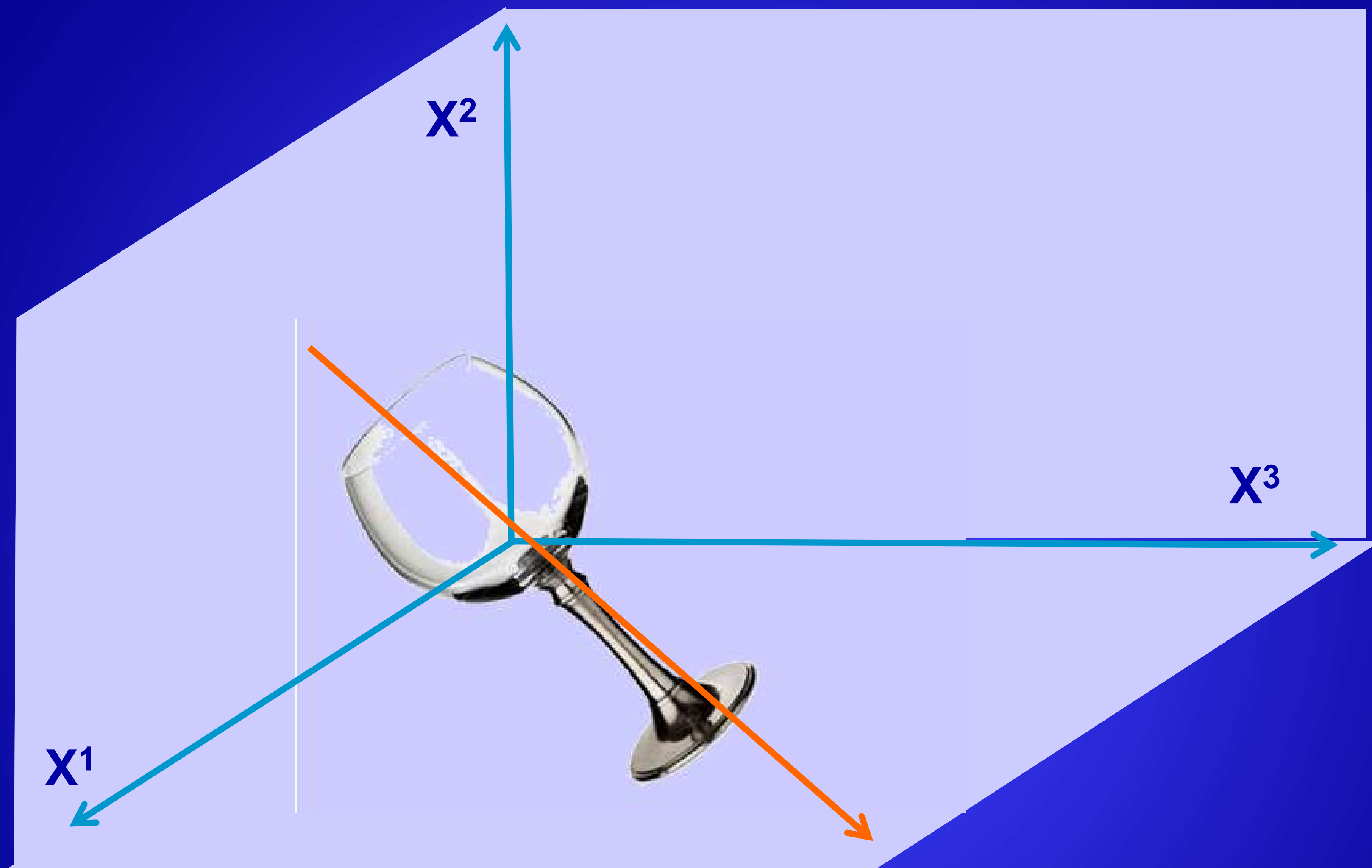
Principal components analysis



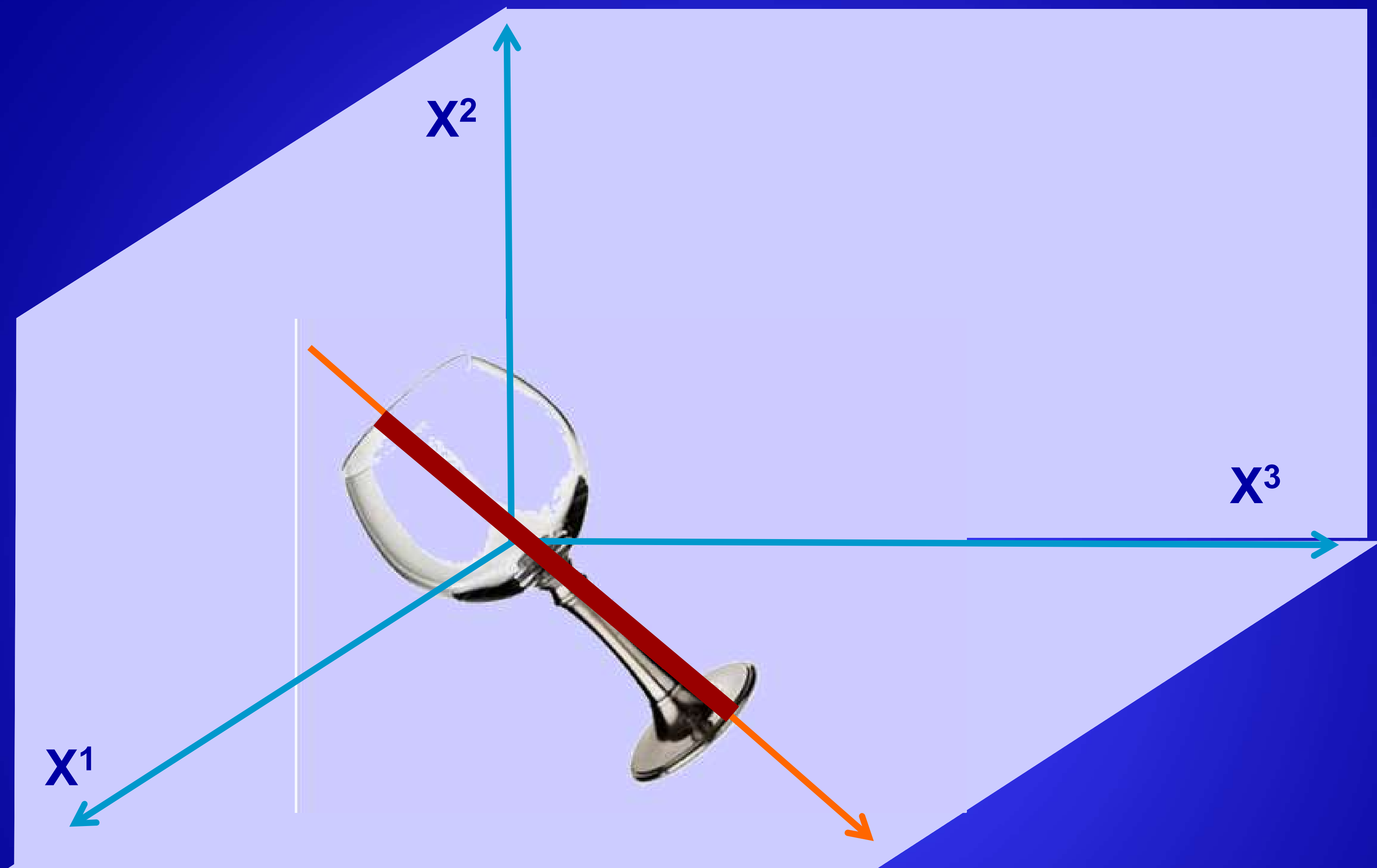
Principal components analysis



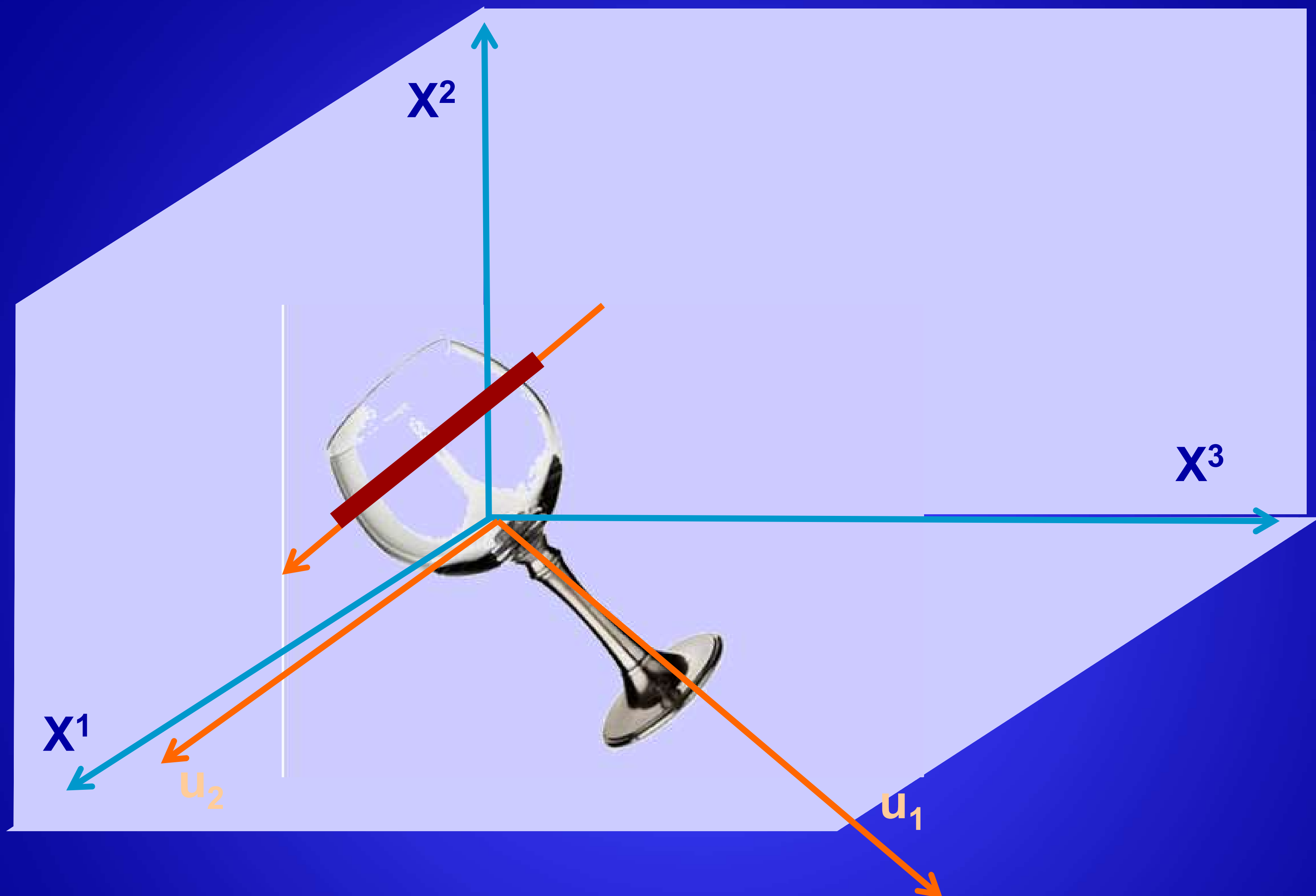
Principal components analysis



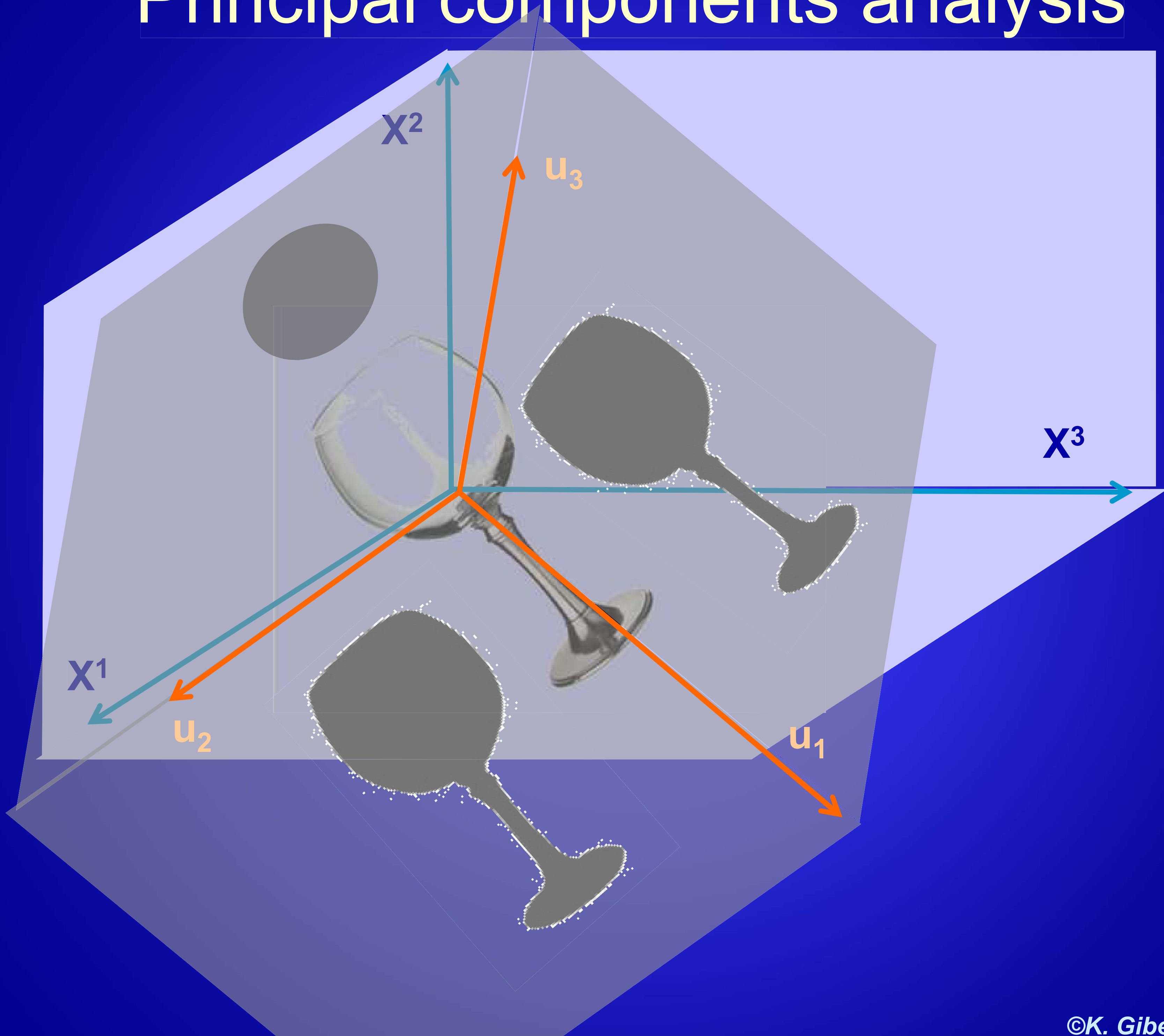
Principal components analysis



Principal components analysis

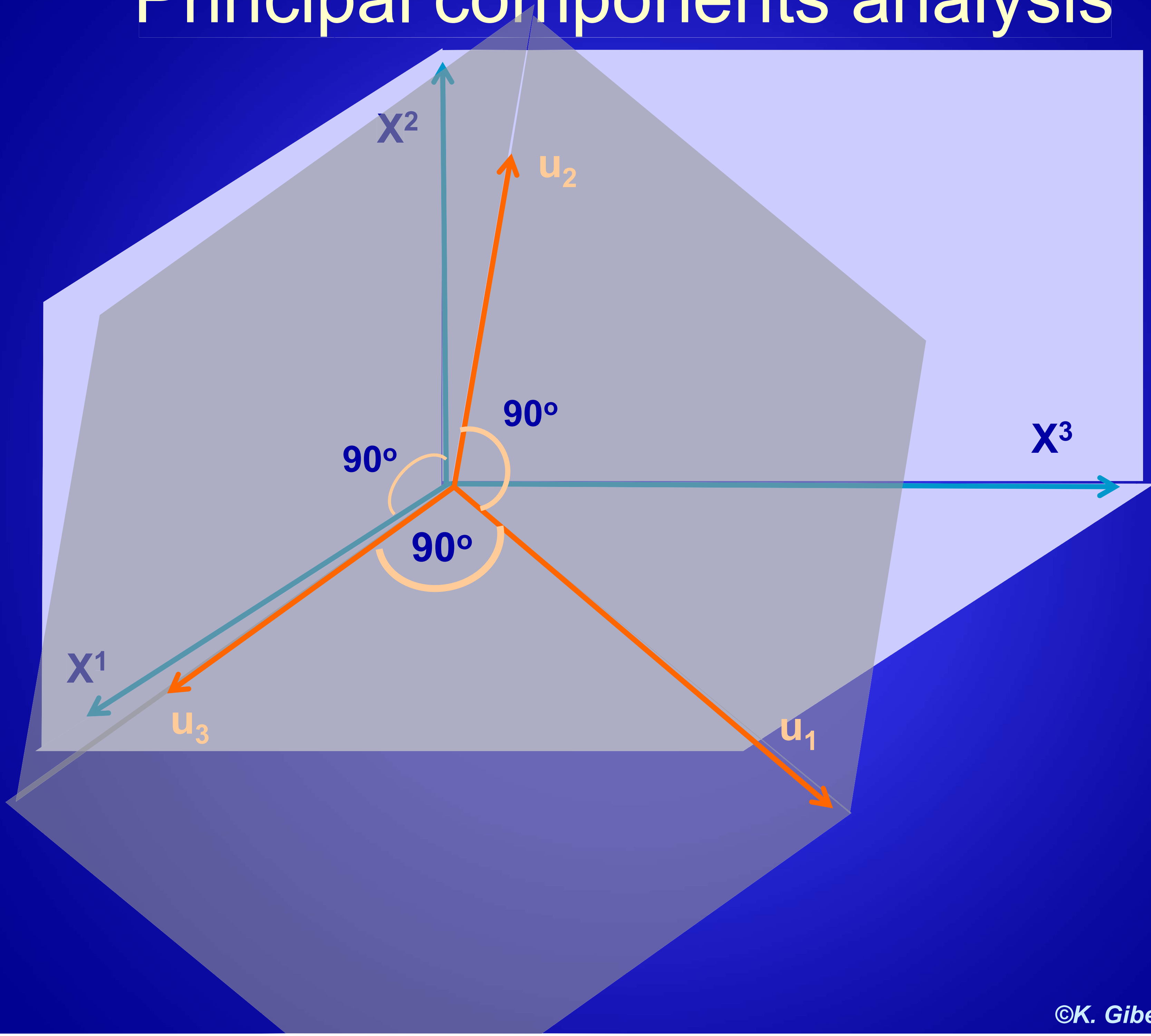


Principal components analysis

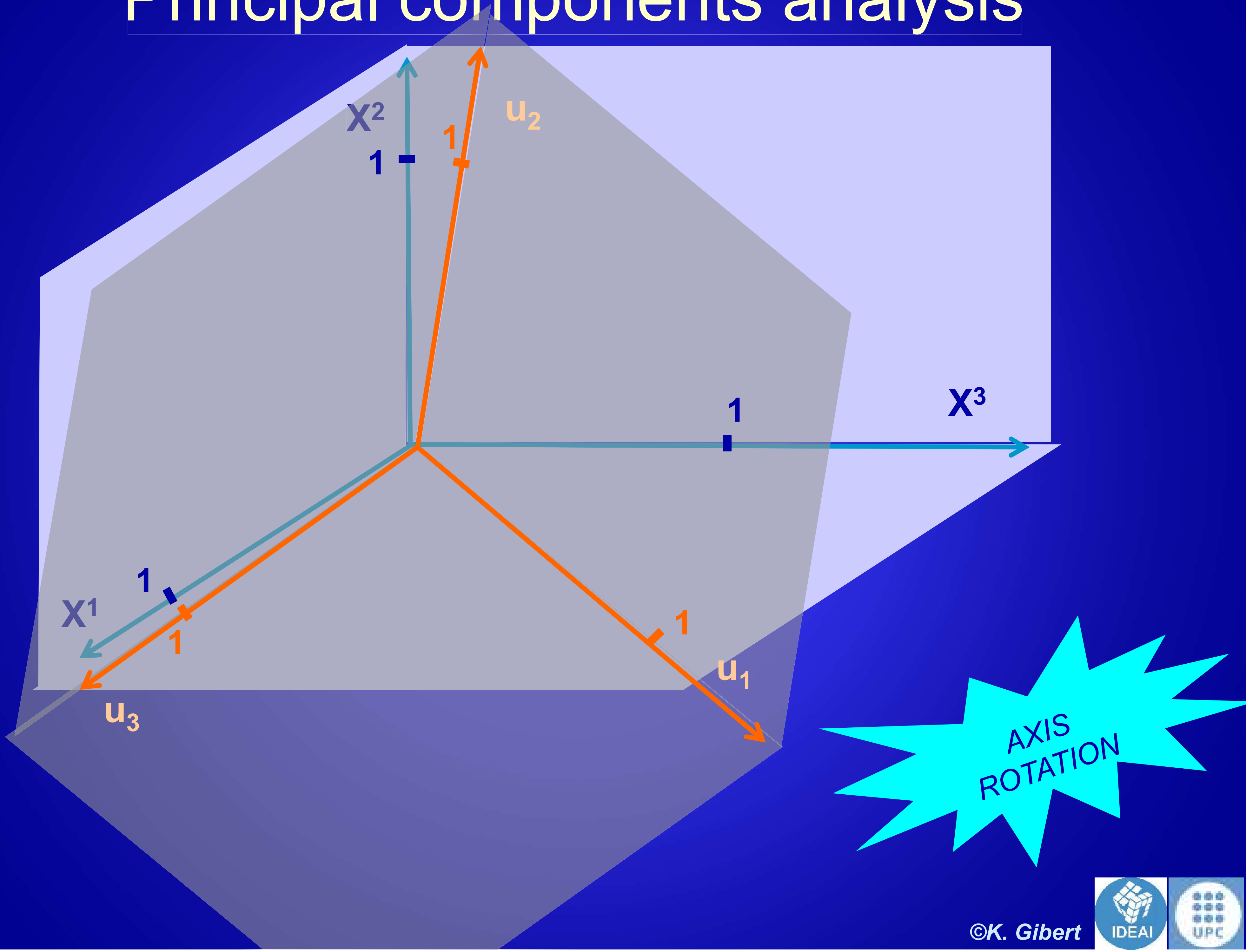


©K. Gibert

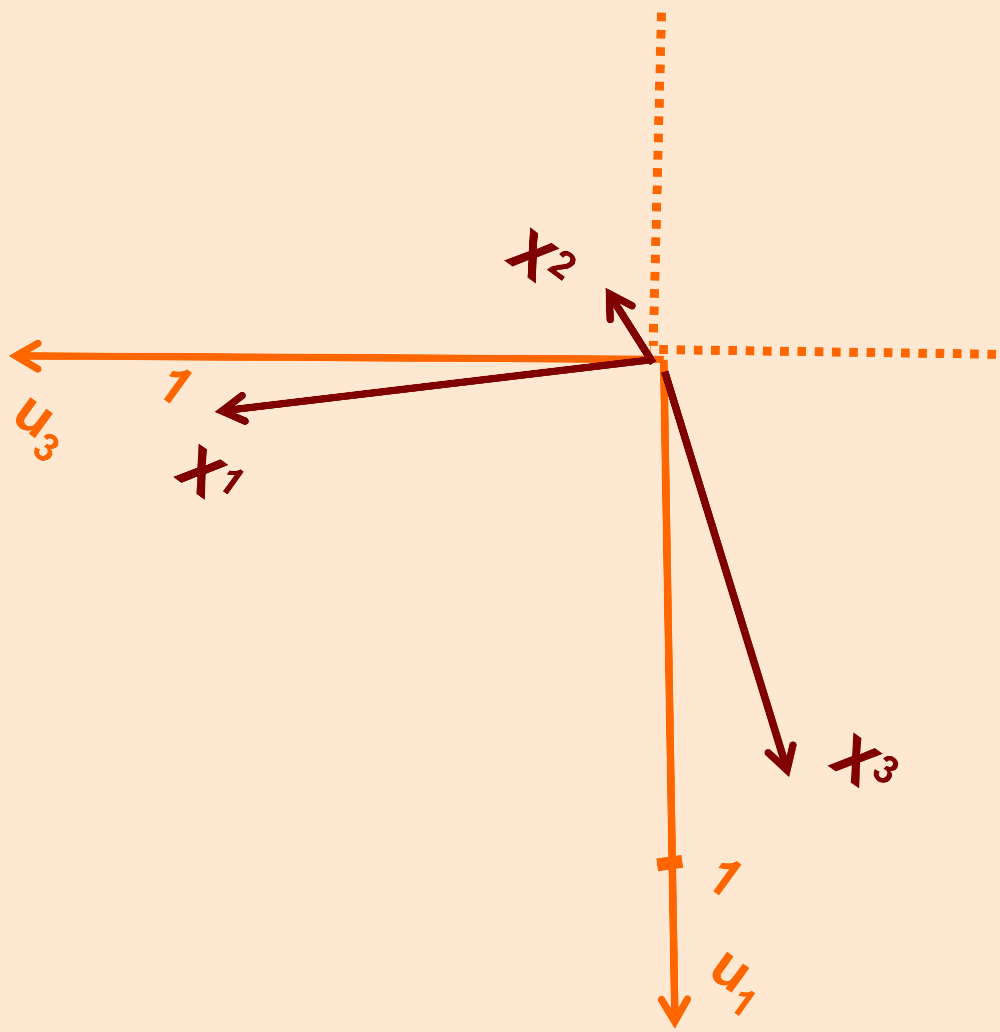
Principal components analysis



Principal components analysis



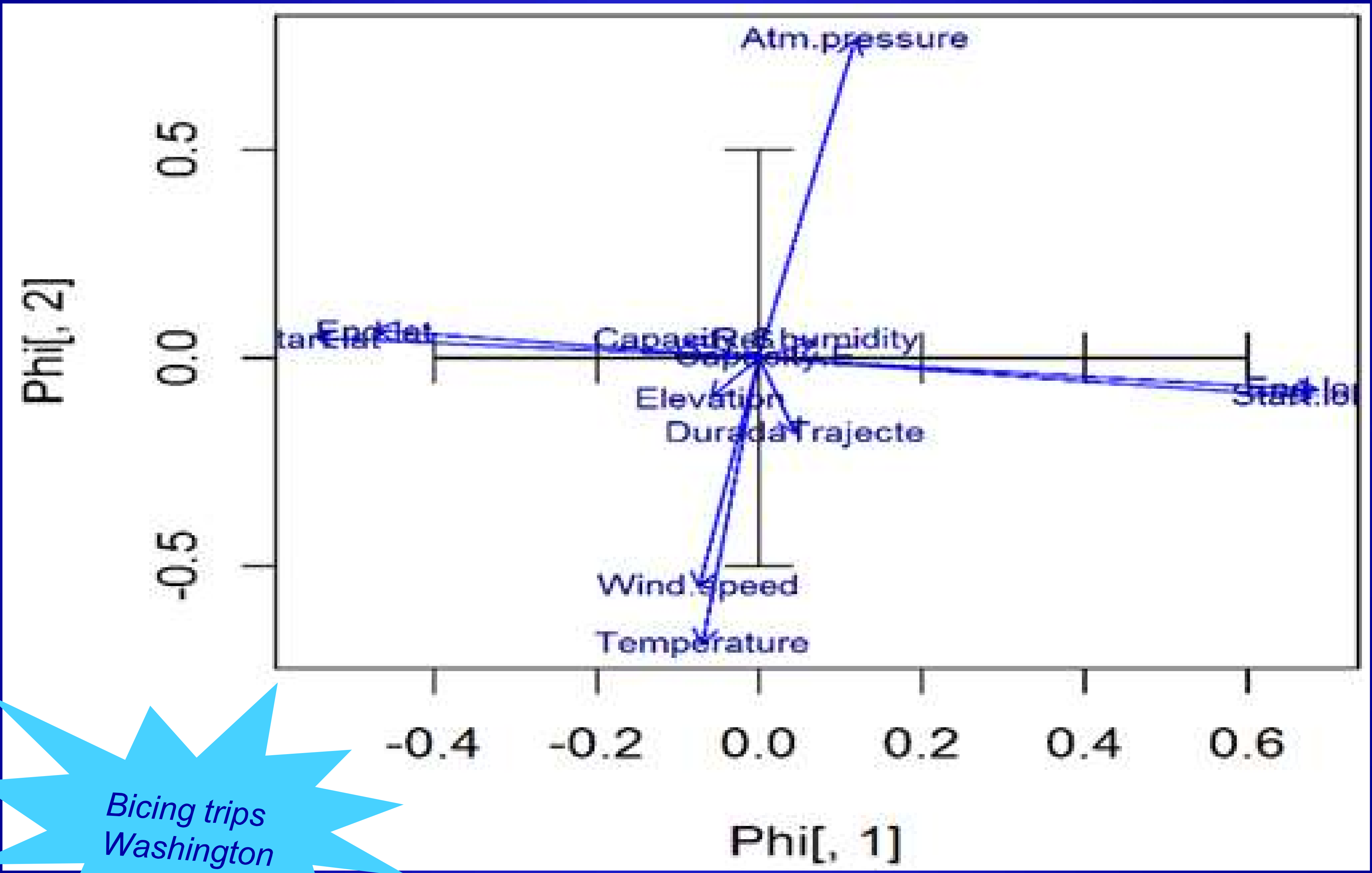
Principal components analysis



Angles linked with Association

Small angles : correlation

Principal components analysis



Bicing trips Washington

Principal components analysis

Variables	**Meaning**
Start.date	Date of the beginning of the trip
End.date	Date of the arrival
Durada.Trajecte	Transit's total duration
Capacity.S	Bike capacity of the origin station
Capacity.E	Bike capacity of the destination station
Elevation	Difference in altitude between the stations of arrival and origin
Start.long	Starting station's longitude according to the CSR WGS84
End.long	Ending station's longitude according to the CSR WGS84
Temperature	Air temperature
Rel.humidity	Air relative humidity
Wind.speed	Wind speed
Atm.pressure	Atmospheric pressure

Trajectes de bicing Washington

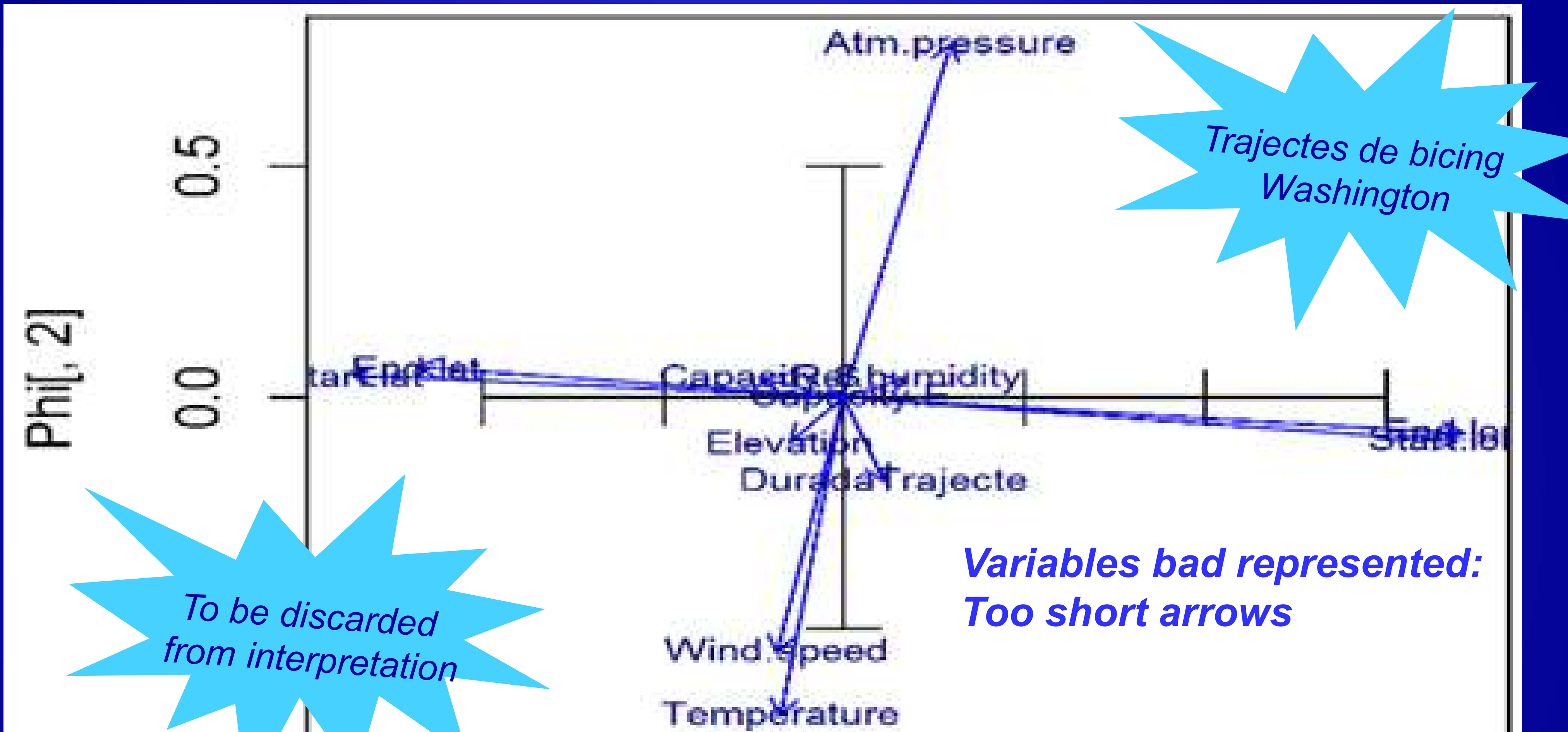
Principal components analysis

Process to interpret a factorial map

- Forget about variables bad represented in the factorial plan
- Which are the variables with relevant direct contribution to Factor in Axis X (eg. PCA1)?
- Which are the variables with relevant inverse contribution to Factor in Axis X (eg. PCA1)
- (later introduce info on qualitative variables as well)
- Analyze profiles opposed in two extremes of Axis X
- Induce a label for the Factor that represents the concept

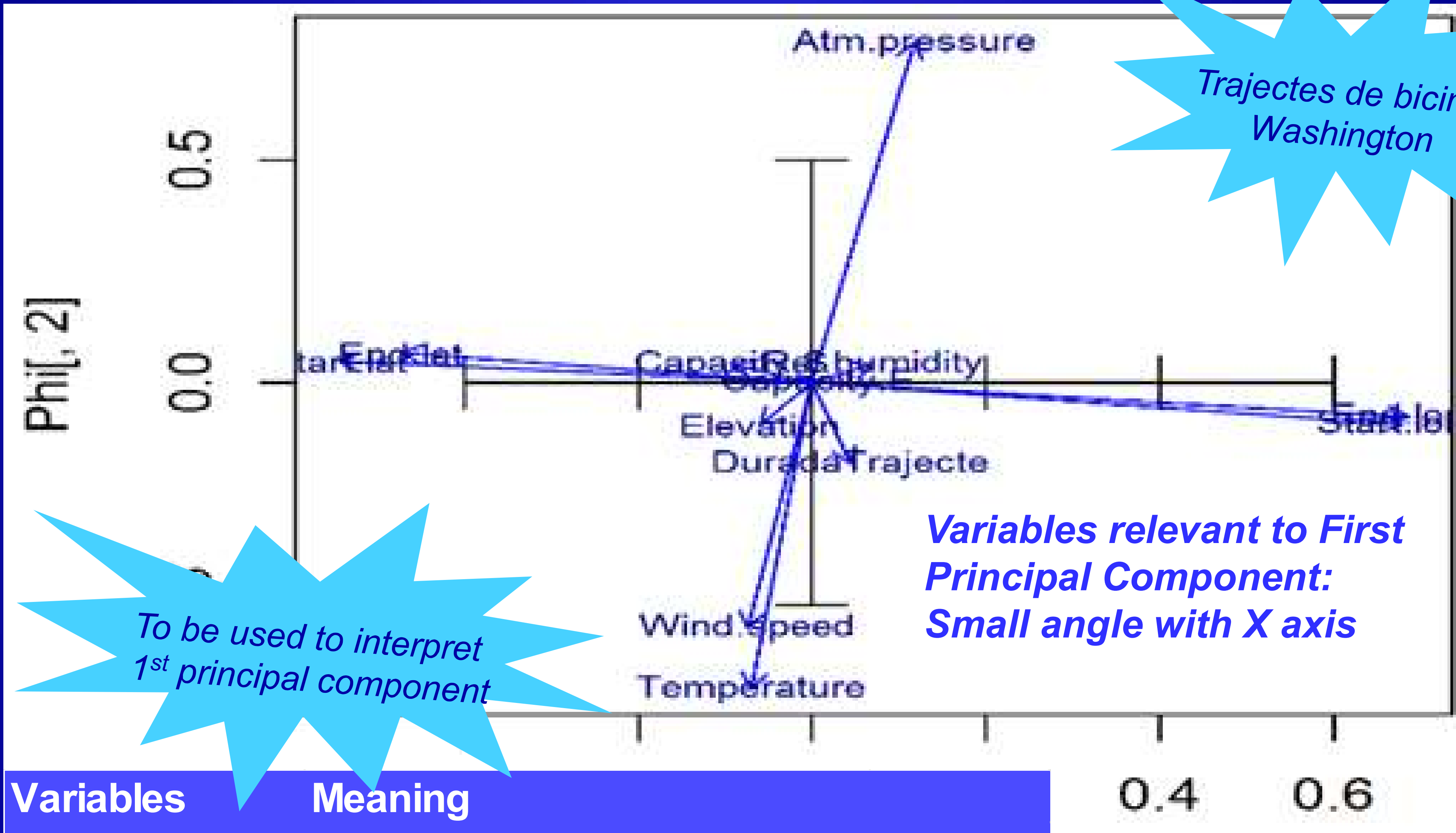
- Repeat with Factor in Axis Y

Principal components analysis



Variables	Meaning
Durada.Trajecte	Transit's total duration
Capacity.S	Bike capacity of the origin station
Capacity.E	Bike capacity of the destination station
Elevation	Difference in altitude between the stations of arrival and origin
Rel.humidity	Air relative humidity

Principal components analysis



Trajectes de bicing Washington

Variables relevant to First Principal Component: Small angle with X axis

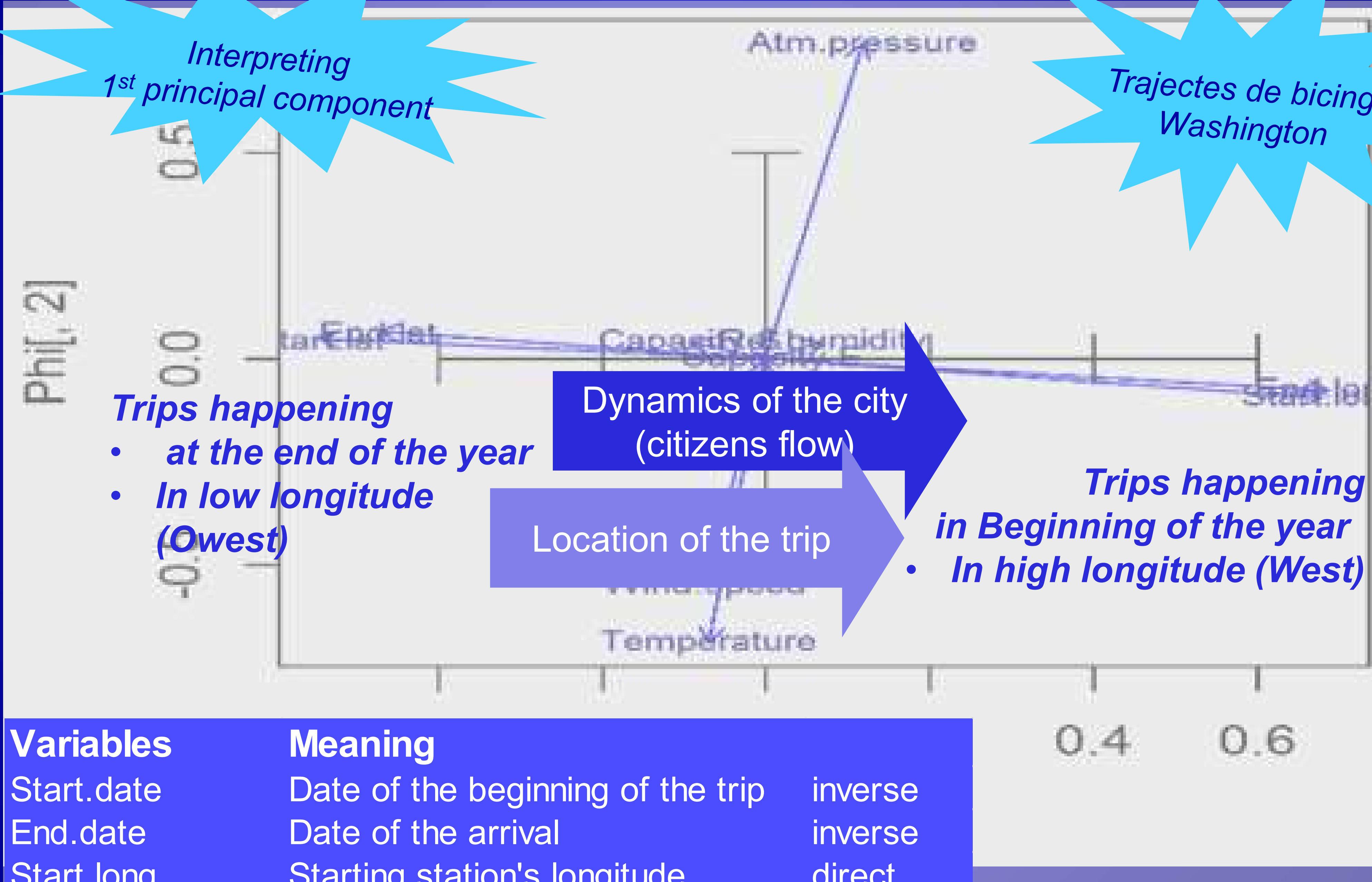
To be used to interpret 1st principal component

Variables	**Meaning**	
Start.date	Date of the beginning of the trip	inverse
End.date	Date of the arrival	inverse
Start.long	Starting station's longitude	direct
End.long	Ending station's longitude	direct

Principal components analysis

Interpreting
1st principal component

Trajectes de bicing
Washington



Trips happening
- *at the end of the year*
- *In low longitude (Owest)*

Dynamics of the city (citizens flow)

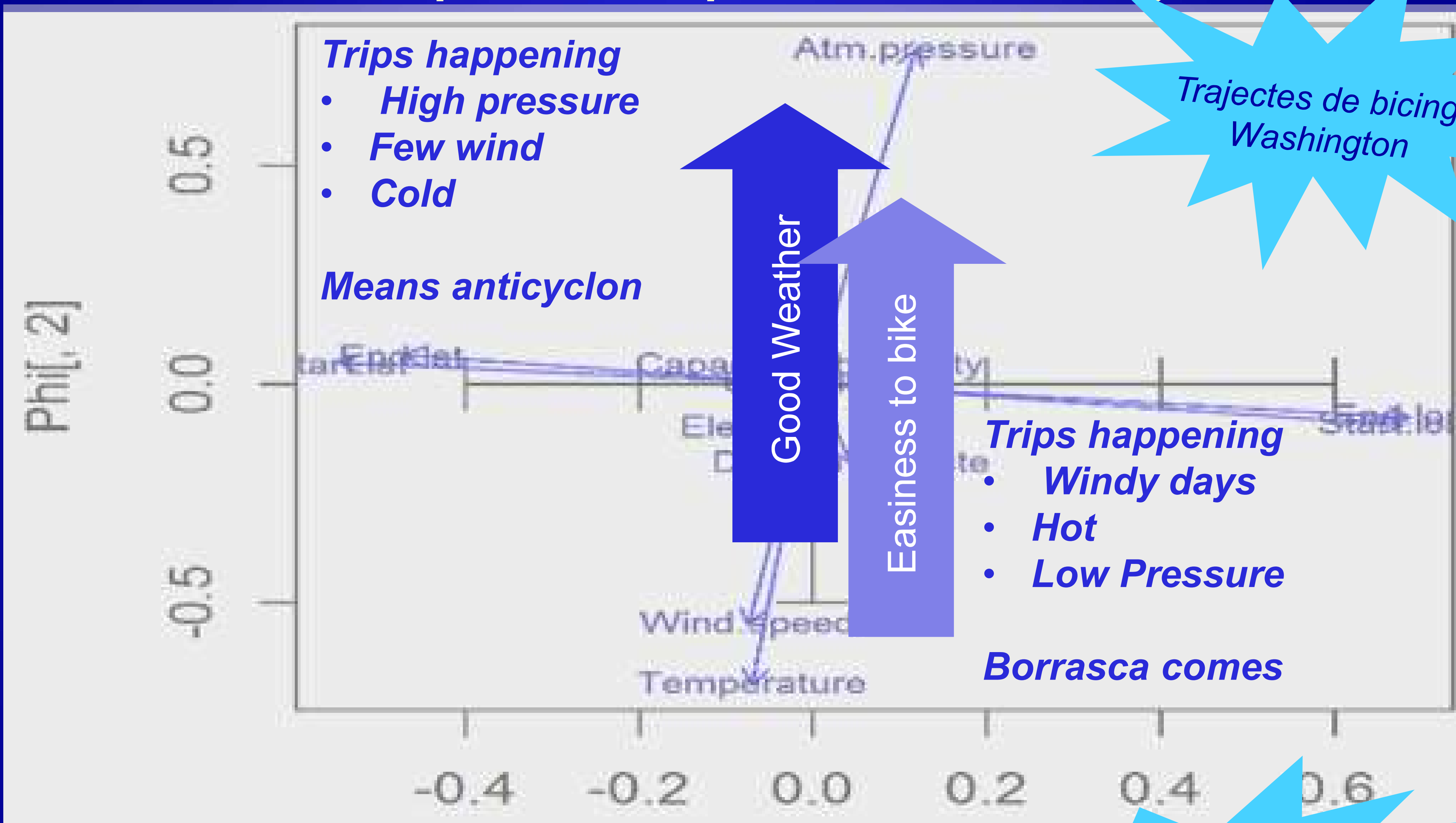
Location of the trip

Trips happening
in Beginning of the year
- *In high longitude (West)*

Variables	Meaning	
Start.date	Date of the beginning of the trip	inverse
End.date	Date of the arrival	inverse
Start.long	Starting station's longitude	direct
End.long	Ending station's longitude	direct

IDEAI UPC

Principal components analysis



Variables	Meaning	
Temperature	Air temperature	invers
Rel.humidity	Air relative humidity	invers
Atm.pressure	Atmospheric pressure	direct

Interpreting 2nd principal component

Factorial Methods

- Principal Components Analysis
  - Output: K factors rotating original X variables
  - Factors: Linear combinations of original variables

Several uses:

- As an associative data mining method to analyze relationships among variables

  Project variables and modalities and find associations

- As a preprocessing method for elicitation of latent variables

  Project active and illustrative variables/individuals on first/second factorial plane and interpret factors (find latent variables)

- As a preprocessing method for multidimensionality reduction

  Select more informative factors κ<<p *(accumulate 80% inertia)*

  Reduce data matrix to selected factors

  Alternative, keep variables mainly contributing to selected factors (smaller angles with factorial axis)