Unsupervised Learning


Predictor variables
Response variables
?
X
Y

Variables

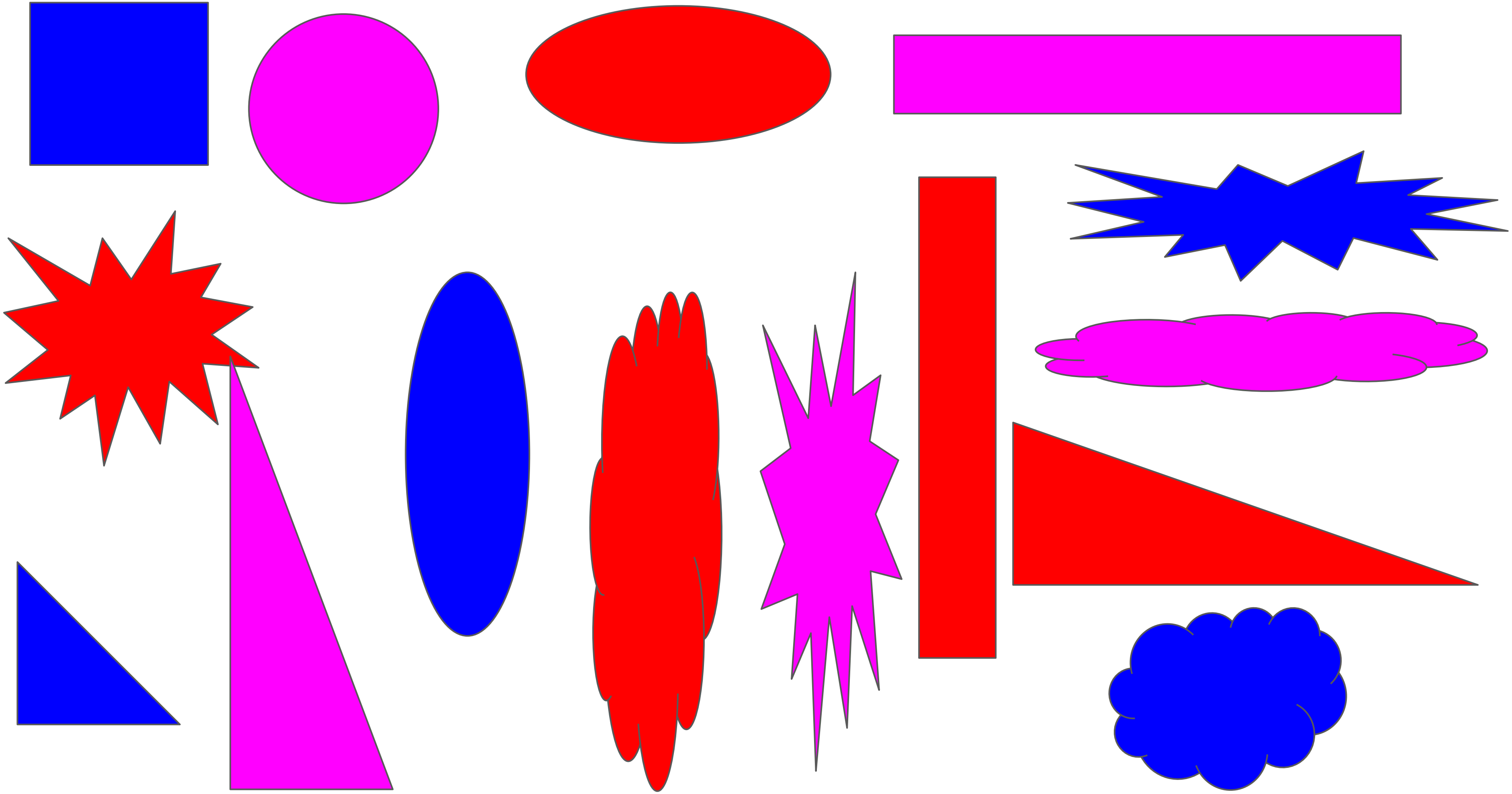
$$? \xrightarrow{?} Y$$

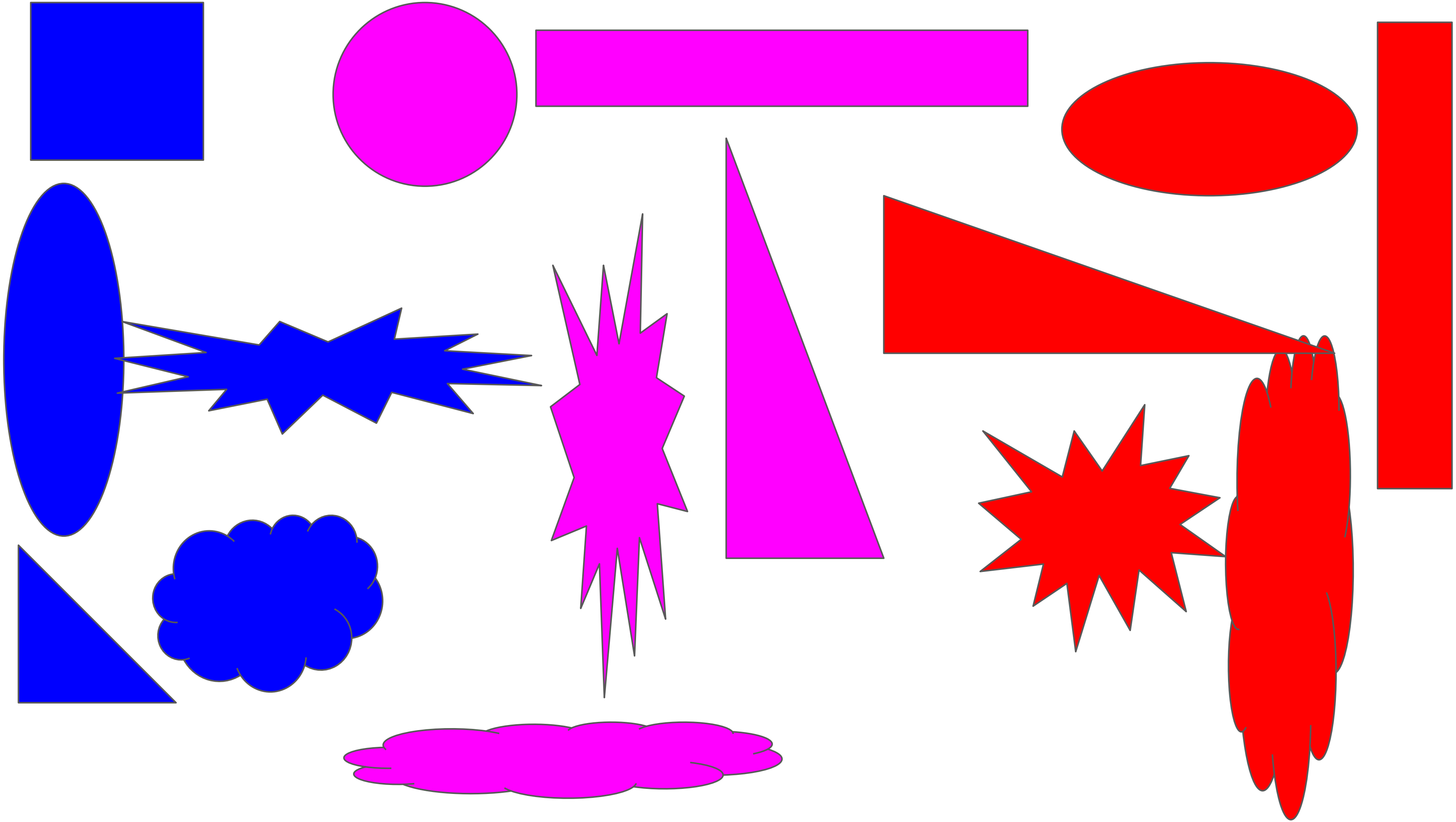
DATA: X

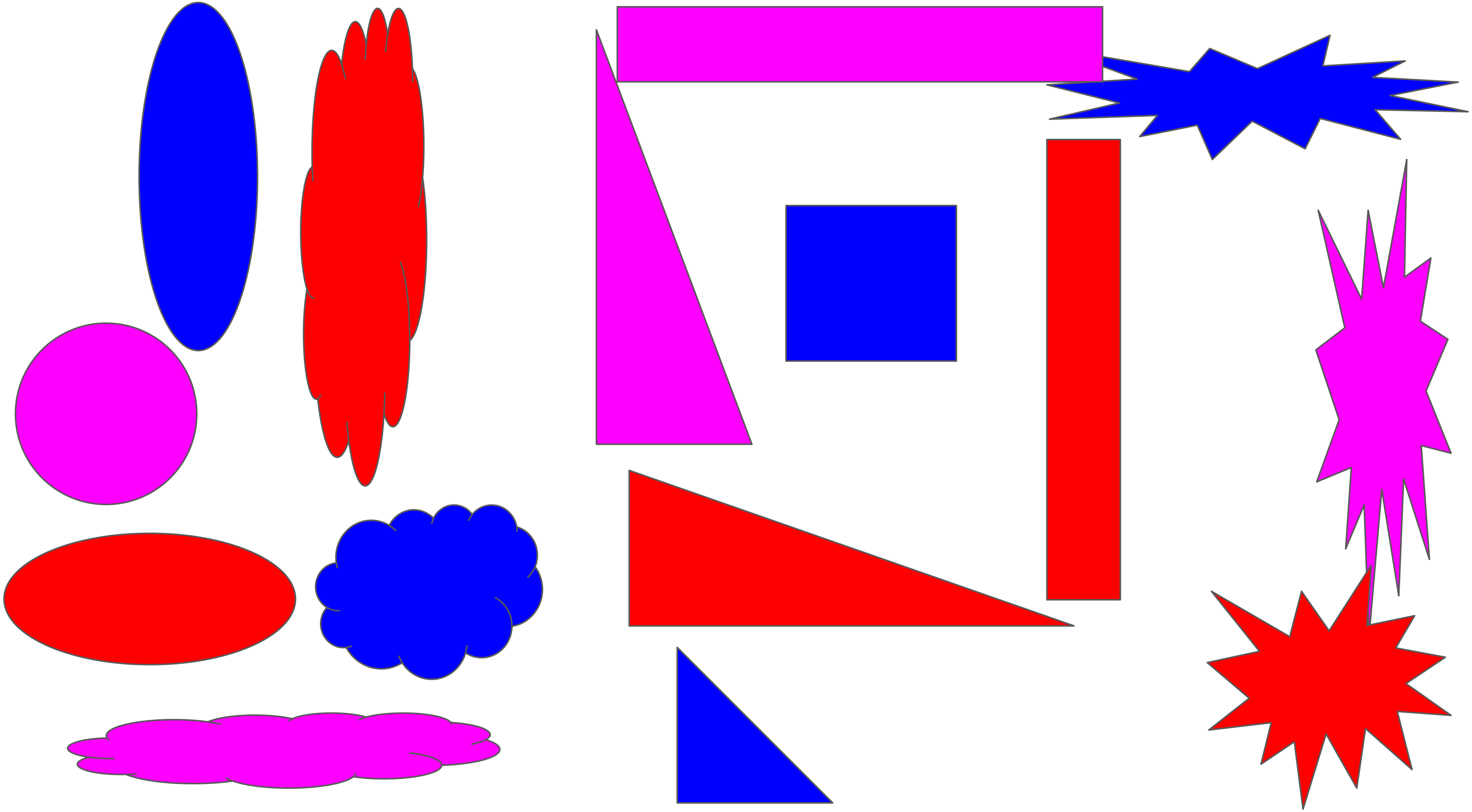
$$\downarrow \text{Using “properties” of X}$$

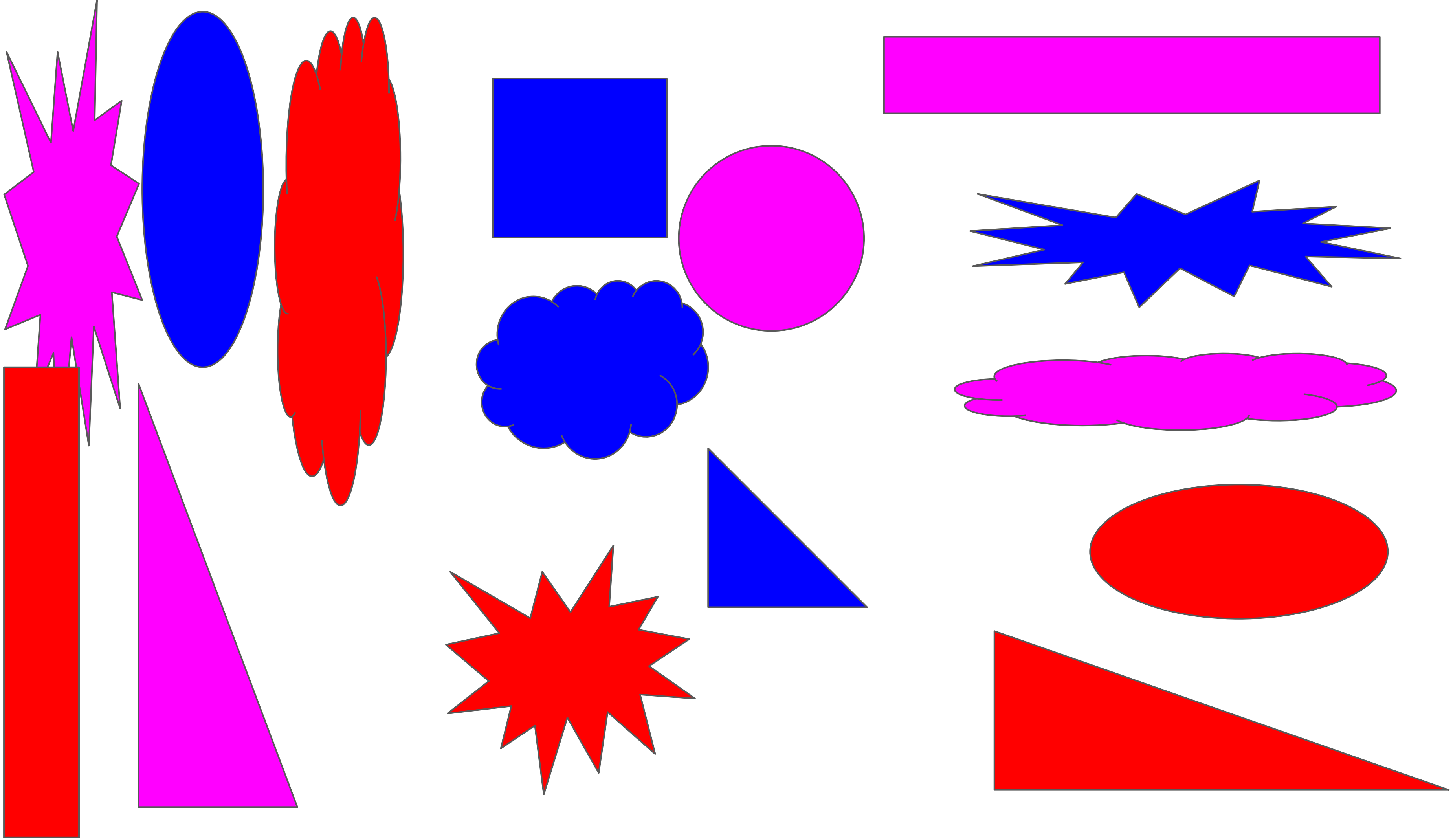
MODEL: $\boldsymbol{\beta}_{0,}\ \boldsymbol{\beta}_{1,}\ \boldsymbol{\beta}_{2,}\ \ldots,\ \boldsymbol{\beta}_n$

Let’s start with a dataset









What should these properties be?

- Well defined
  - No ambiguity
- Computable
  - Always has an answer

Data Centric and multidimensional

Variables:	**var_1**	**var_2**	**var_3**	**…**	**var_{n-1}**	**var_n**
Samples:						
$sample_1$						
$sample_2$						
$sample_3$						
…						
$sample_{m-1}$						
$sample_m$						

Almost always very multidimensional

$N_{variables} >> M_{samples}$

- Peptides in CE-MS
- Genotypes after sequencing
- Species in a given area
- Transcripts in a tissue
    - Transcripts in a cell

Techniques in unsupervised learning

Hodgepodge of nonsense!!!!

Nearest Neighbours
- K nn
- Fixed-radius nn
- Approximate nn

Principal Component Analysis

Cluster Analysis
- Connectivity based CA
- Centroid based
- Grid based

Multidimensional Scaling
- Principal coordinate analysis
- Metric
- Non metric
- Generalized multidimensional scaling

Neural Network Autoencoders
- Classic
- Variational
- Denoissing

Mixture Models
- Gaussian
- Categorical
- Multivariate

Neighbour Embedding
- TSNE
- UMAP

What we will do today

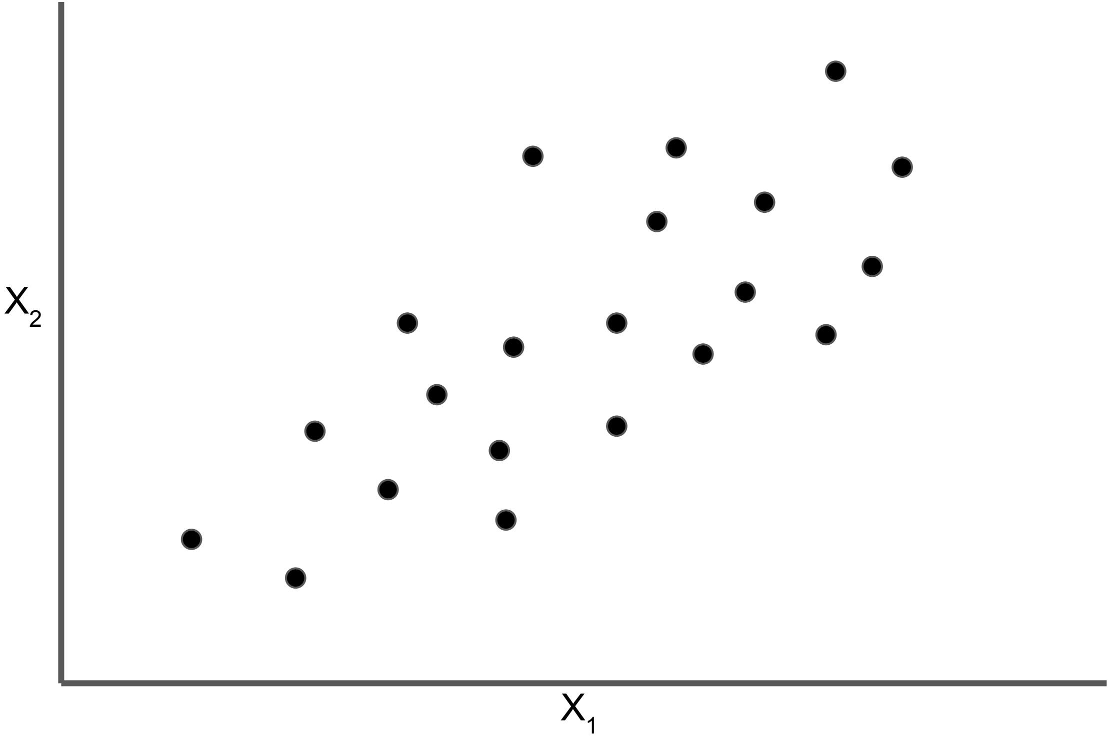
One technique - PCA

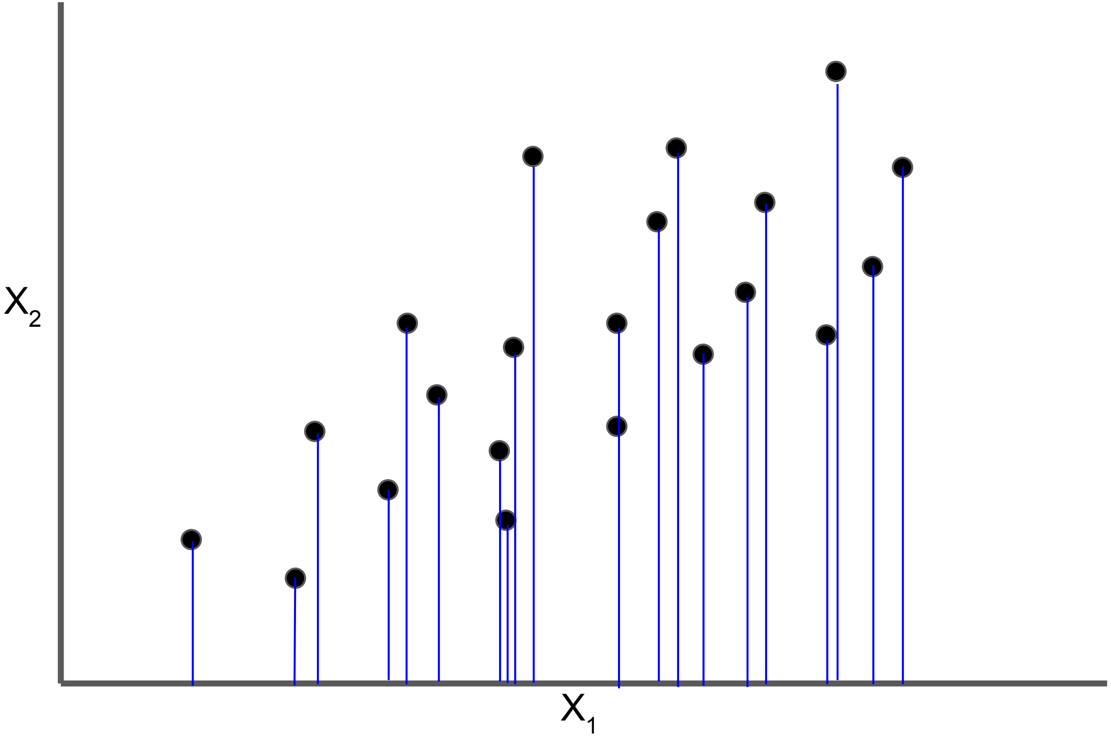
- Cover it broadly, without brushing off the important details.
- Set the groundwork for other techniques.
- Not learn hot to “compute” a PCA, but understand its principles.

Data X

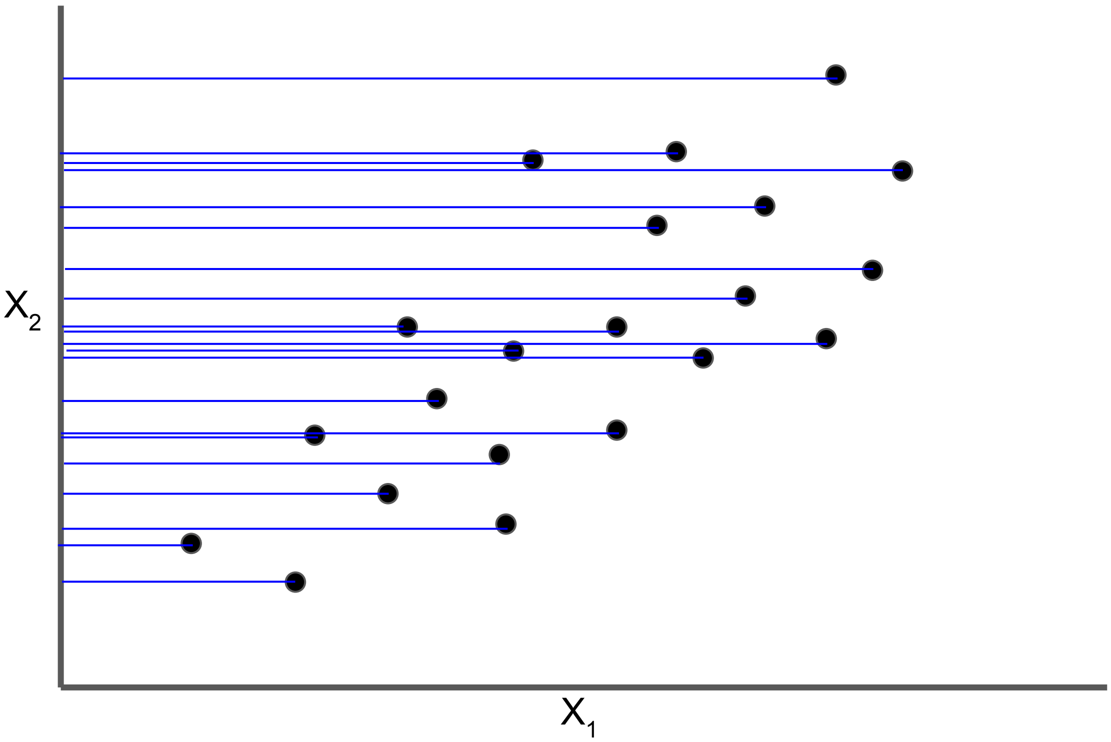
	x_1	x_2
$sample_1$		
$sample_2$		
$sample_3$		
…		
$sample_{m-1}$		
$sample_m$		

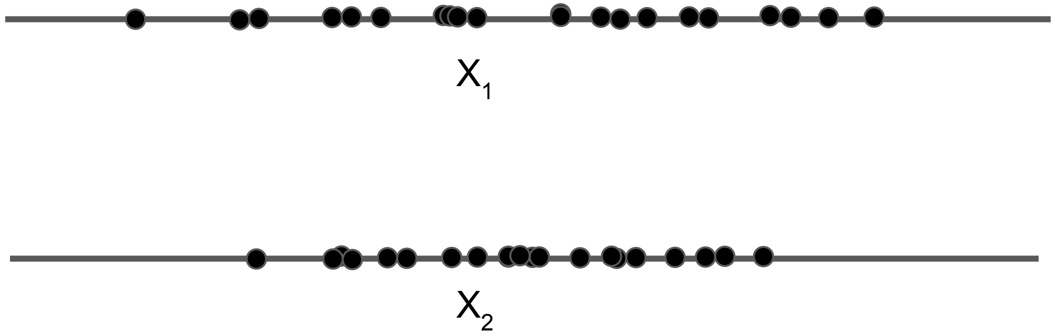
We have some data $X \in \mathbb{R}^2$





X2
X1


X_2
X_1

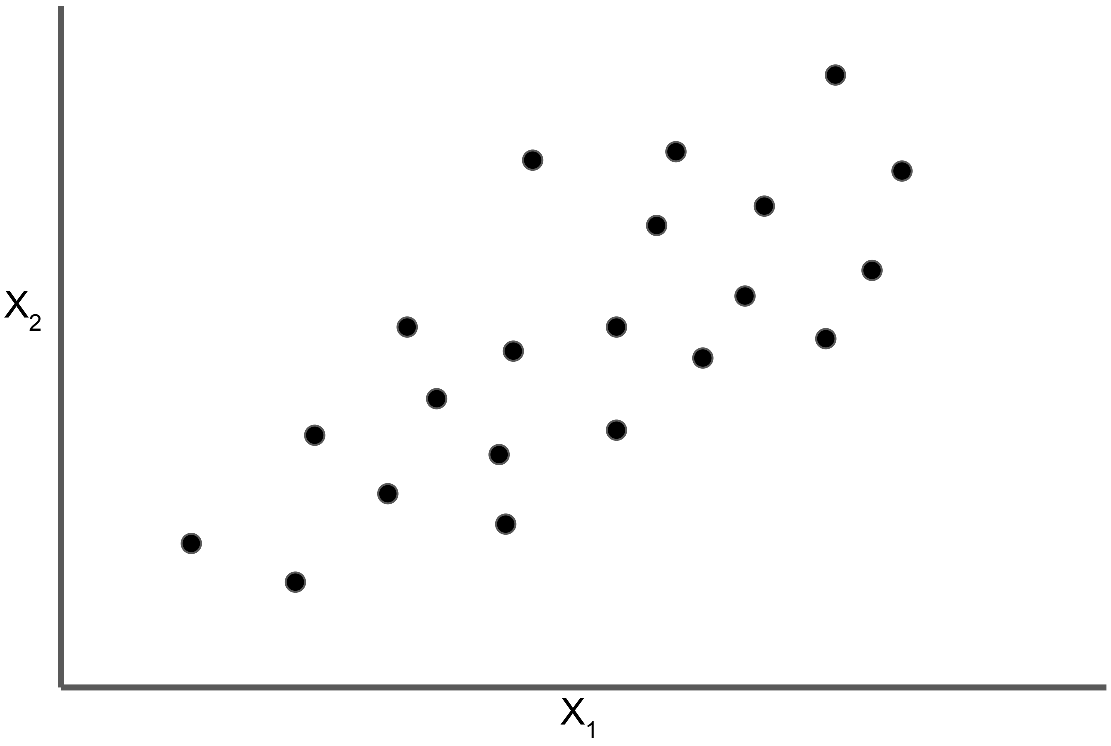


$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

var(x1) and var(x2)

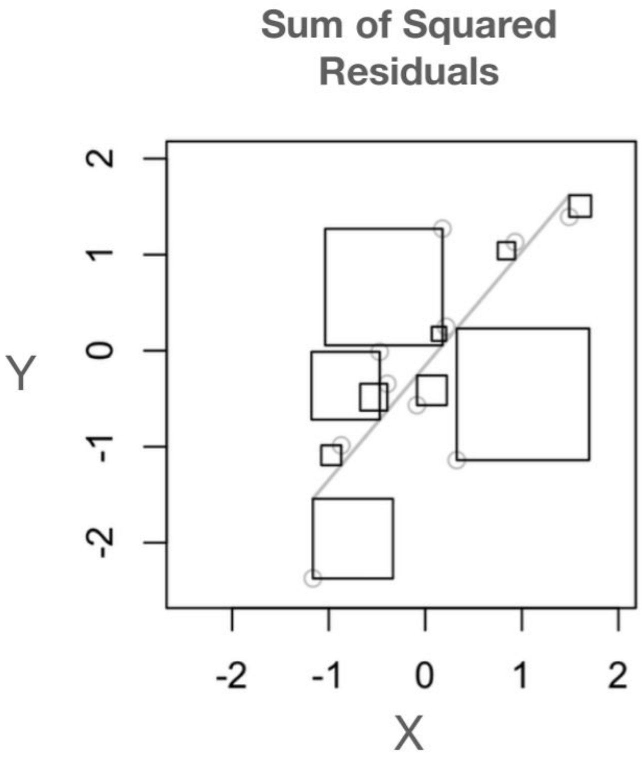
Total variance in the sample:

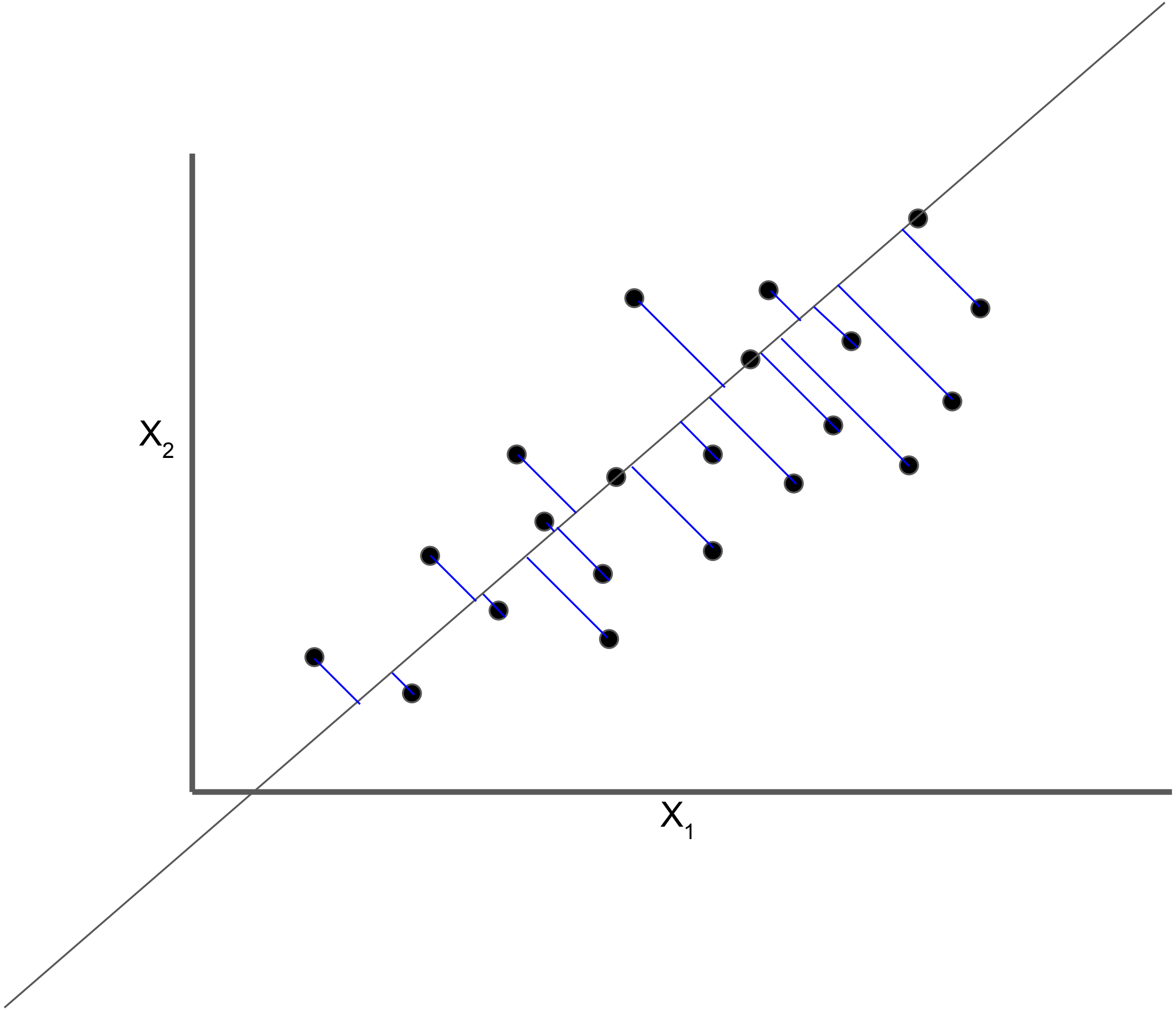
var(x1) + var(x2)



X_2
X_1

Correlated variables

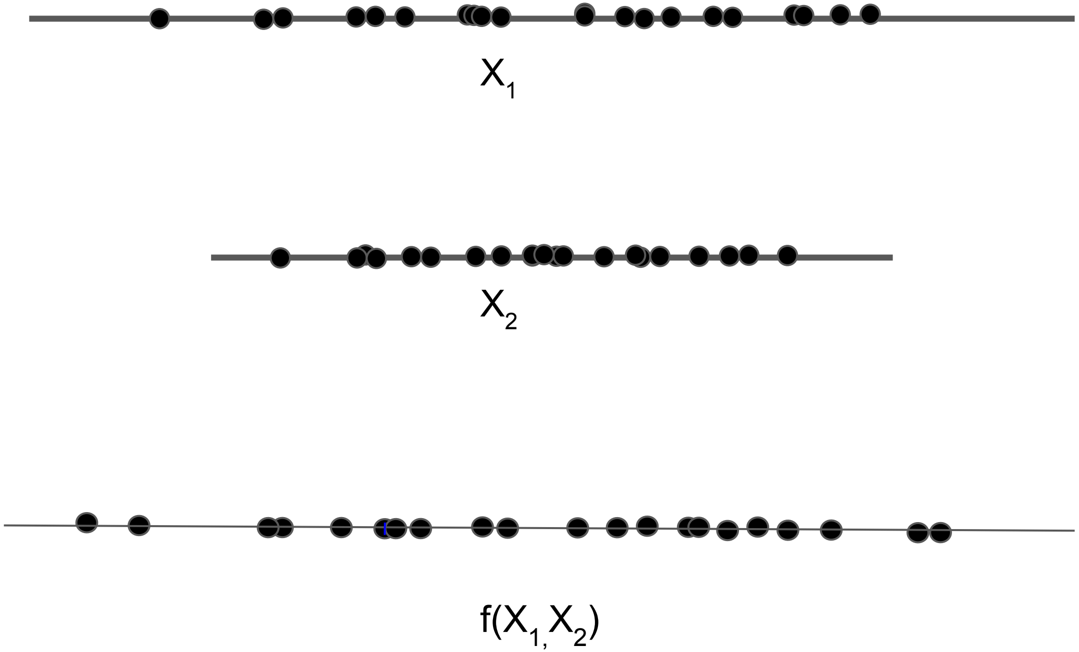




$f(X_1, X_2)$

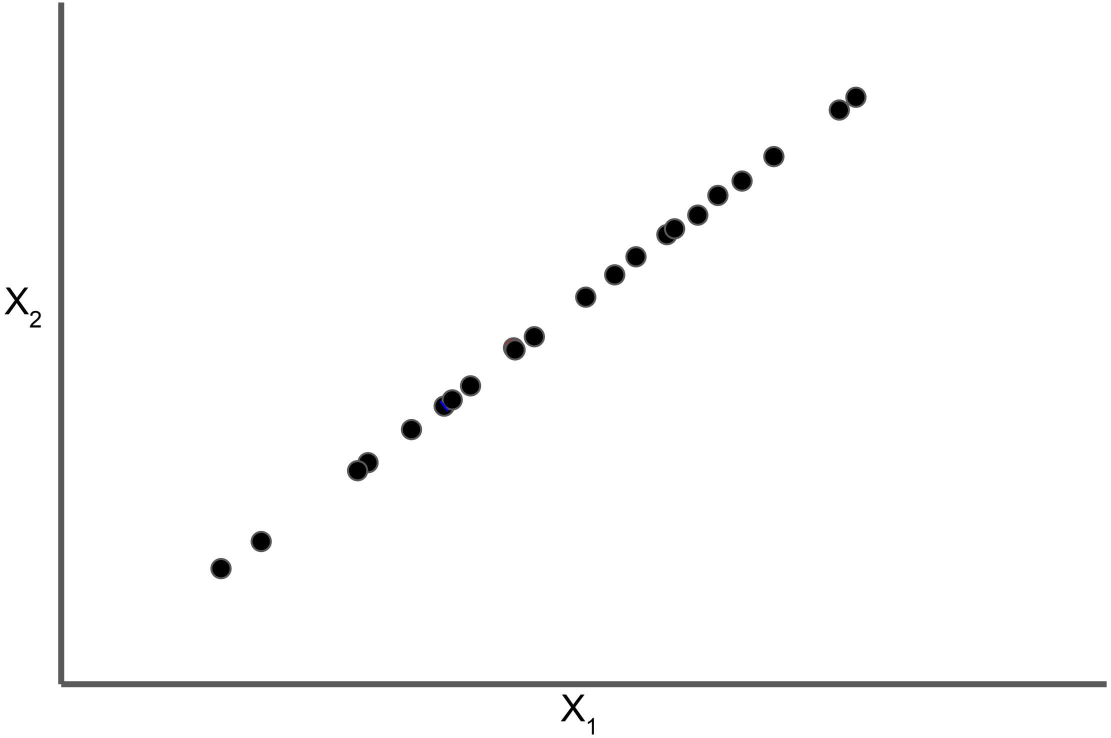
$f(X_1, X_2)$ must be a linear function

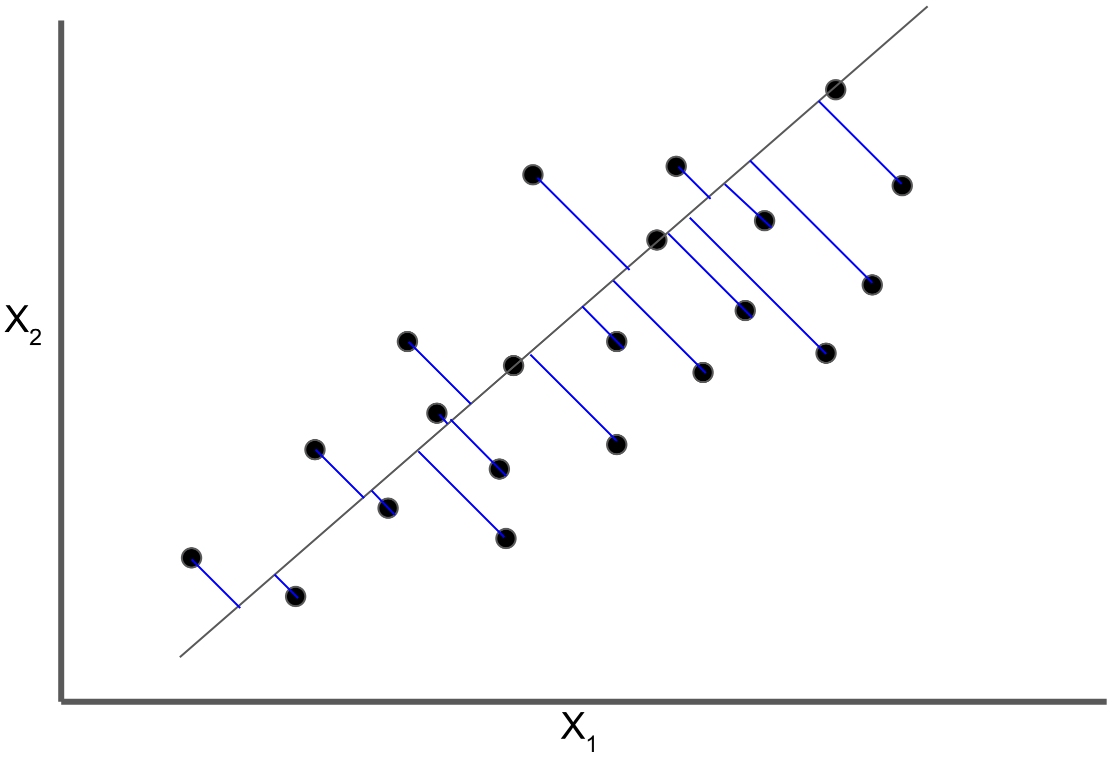
Addition
Multiplication



$$\mathrm{var}(f(X_{1,}X_2)) > \mathrm{var}(x1) \text{ or } \mathrm{var}(x2)$$

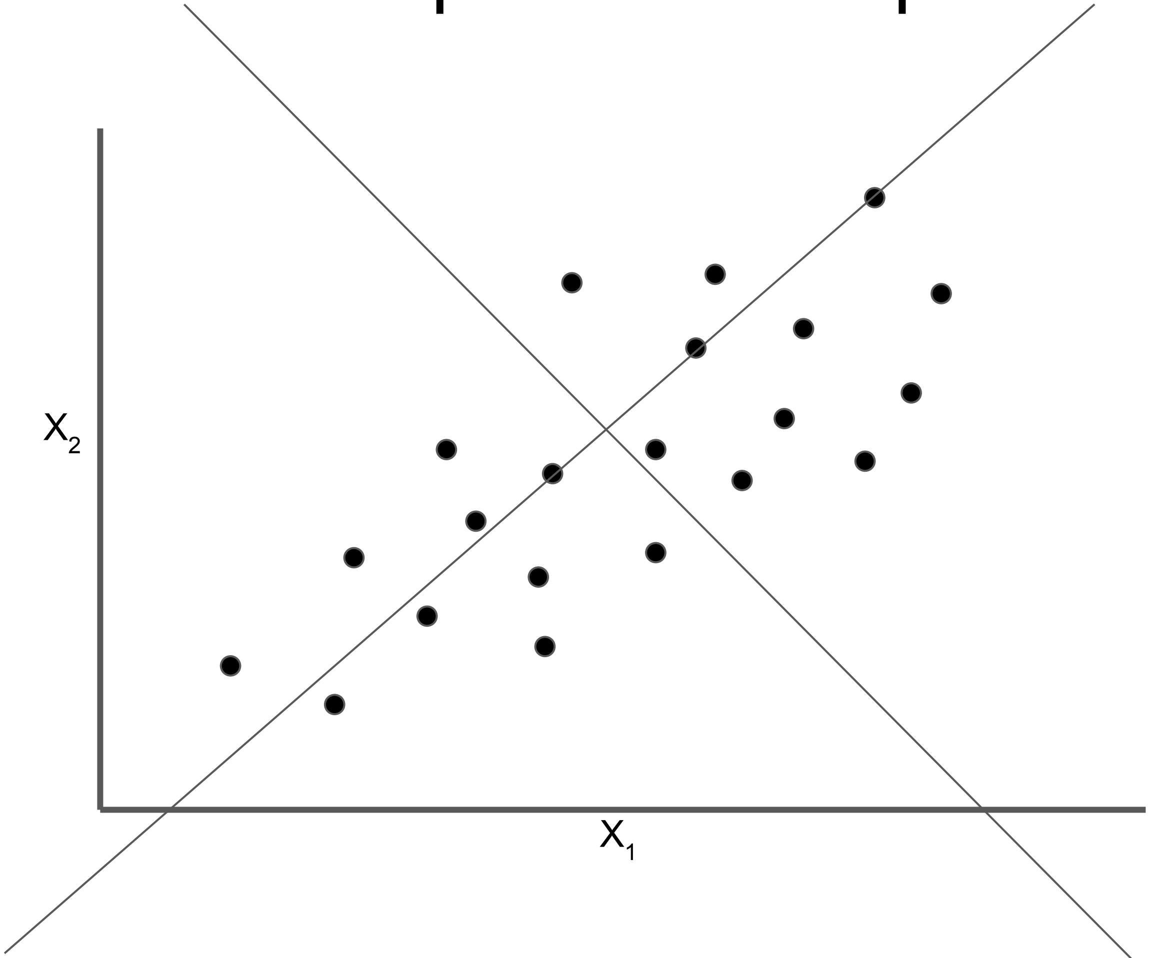
We have missing information



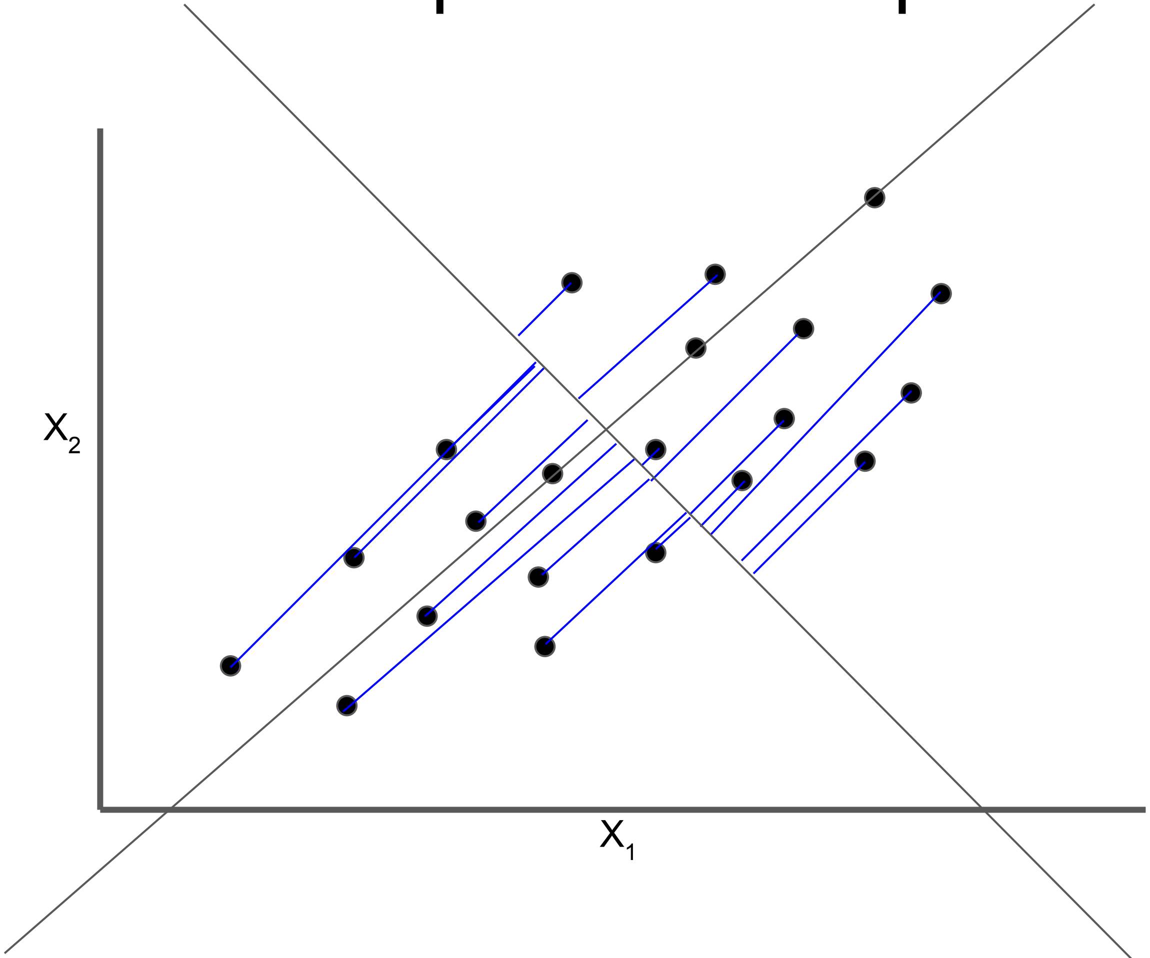


X_2
X_1

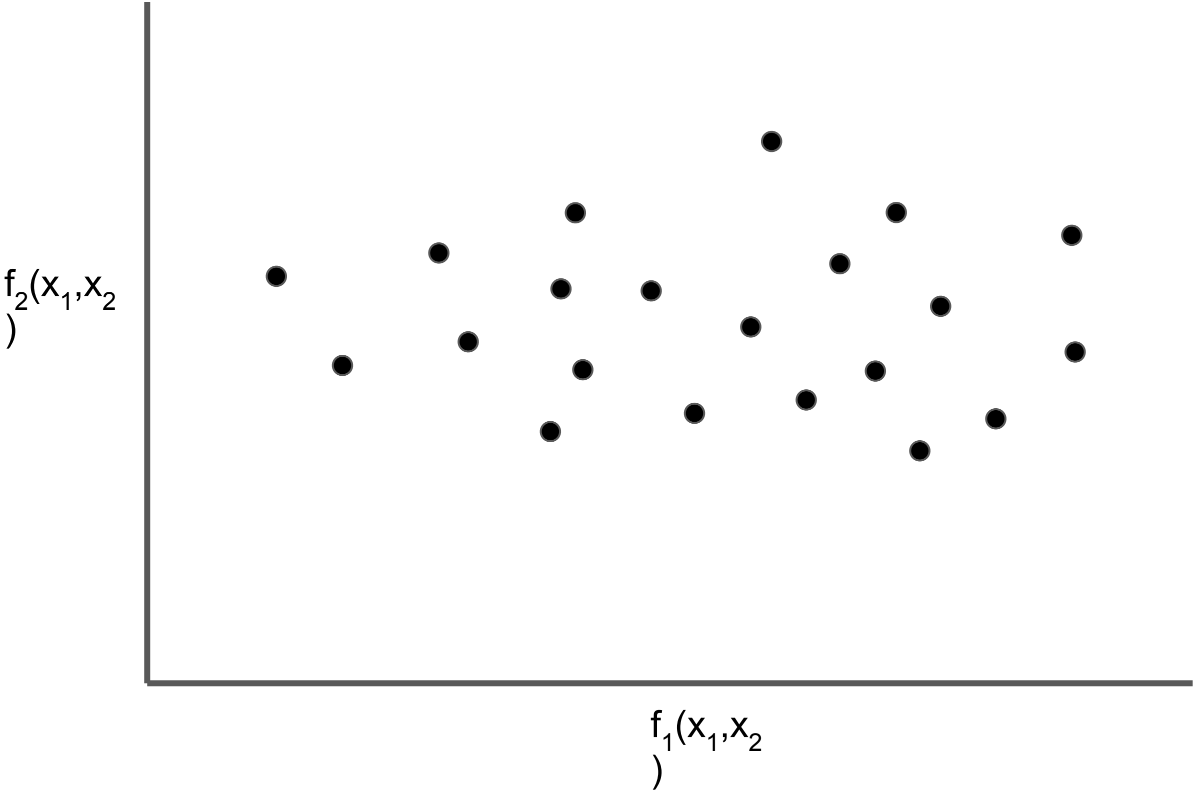
We need a new “independent component”

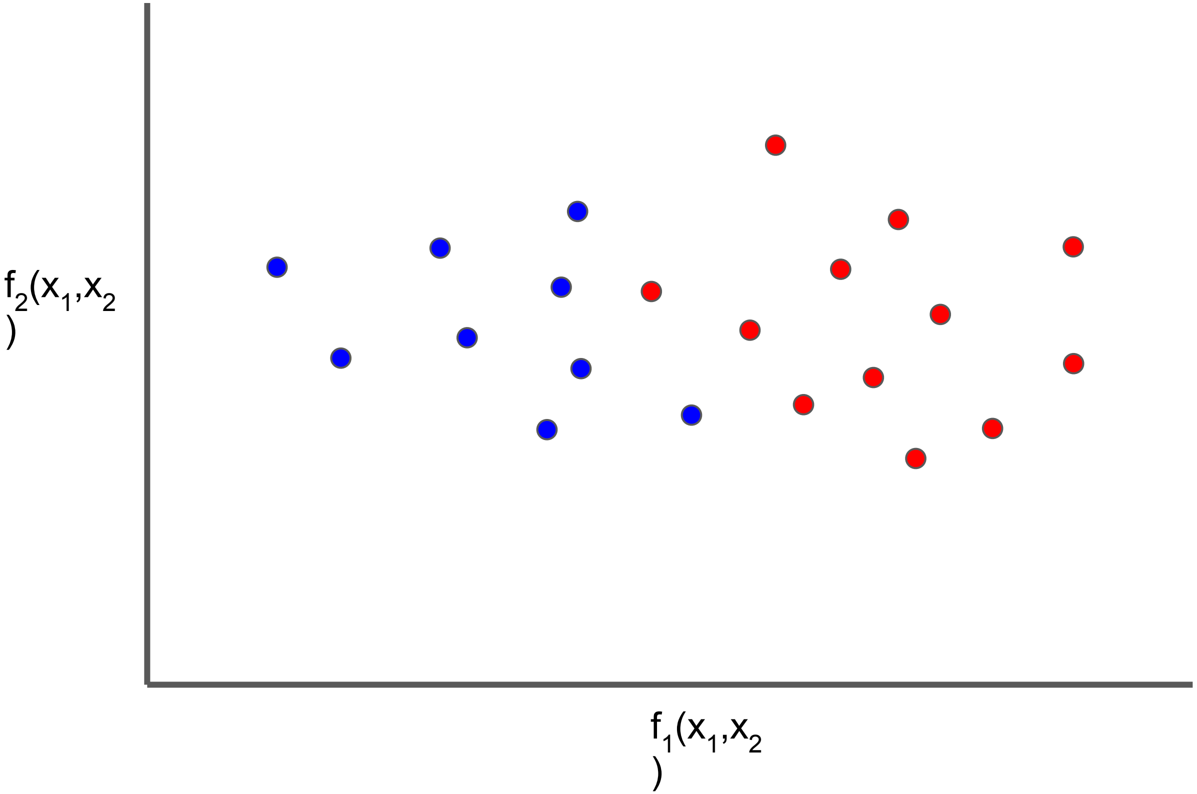


We need a new “independent component”



$$\mathrm{var}(f_1(X_{1,}X_2)) + \mathrm{var}(f_2(X_{1,}X_2)) == \mathrm{var}(x1) + \mathrm{var}(x2)$$


$f_2(x_1,x_2)$
$f_1(x_1,x_2)$

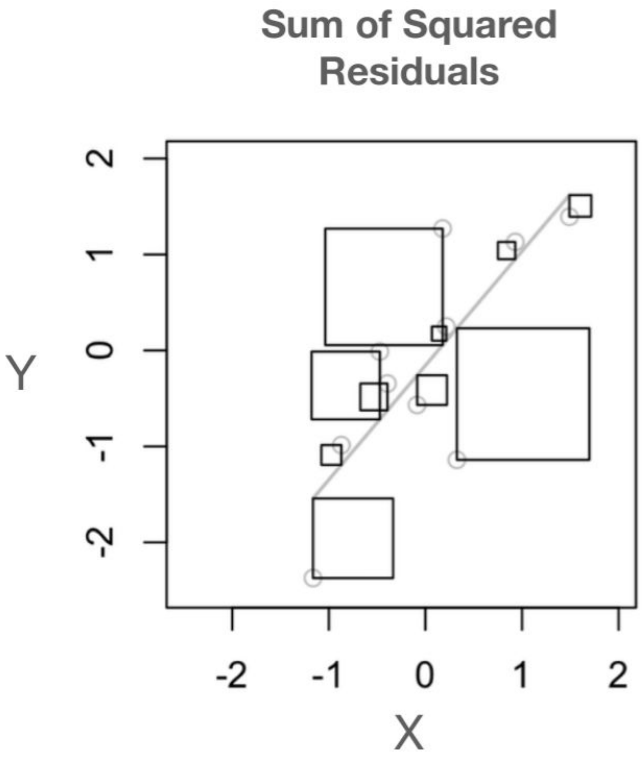


$$var(f_1(X_{1,}X_2)) + var(f_2(X_{1,}X_2)) == var(x1) + var(x2)$$

Two dimensions are cool, but we have bigger problems that that.

Let's define the “constraints”

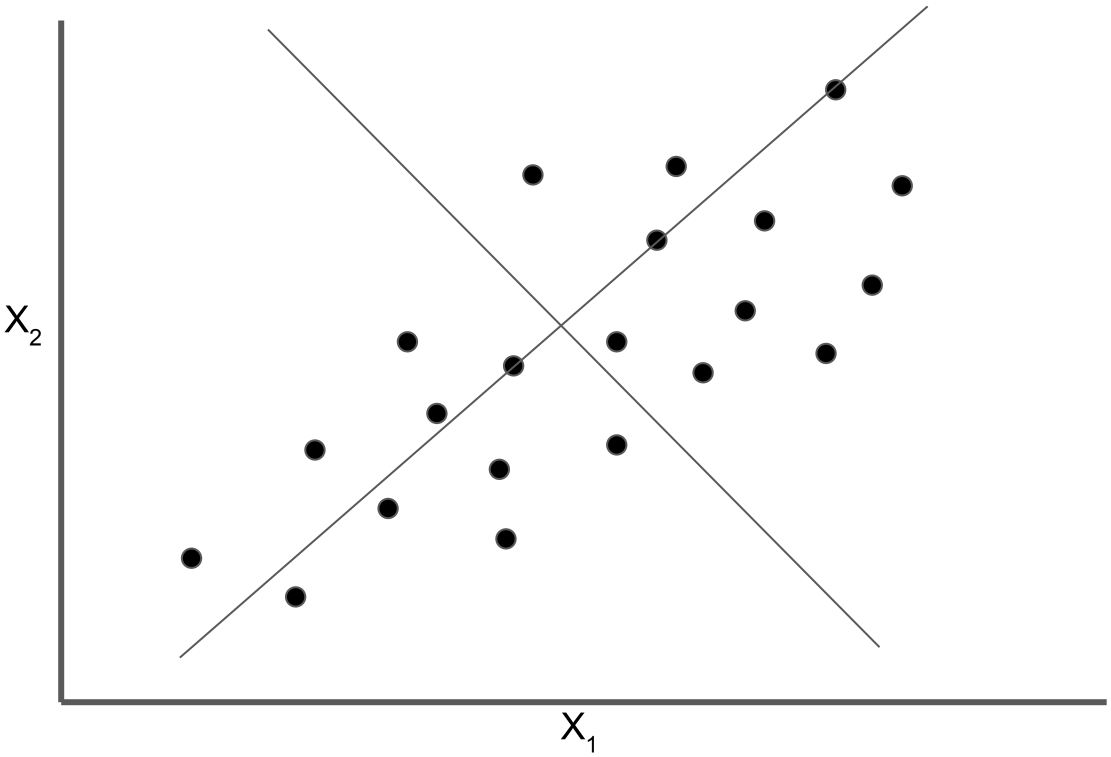
Constrains



$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

but our estimate $\hat{y}_i$ is simply a linear function of x_i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$


X_2
X_1

Constrains

Maximum variance

Linear independence

How to compute?

Nothing in linear algebra makes sense except in the light of eigen-analysis.

Matrices as linear transformation

The covariance matrix

	var$_1$	**var**$_2$	**var**$_3$	**…**	**var**$_{n-1}$	**var**$_n$
var$_1$	S^2(**var**$_1$)					
var$_2$	COV(**var**$_1$, **var**$_2$)	S^2(**var**$_2$)				
var$_3$	COV(**var**$_1$, **var**$_3$)		S^2(**var**$_3$)			
…	…			…		
var$_{n-1}$	COV(**var**$_1$, **var**$_{n-1}$)				S^2(**var**$_{n-1}$)	
var$_n$	COV(**var**$_1$, **var**$_n$)					S^2(**var**$_n$)