Title: Lecture - Machine Learning, PHYS 777

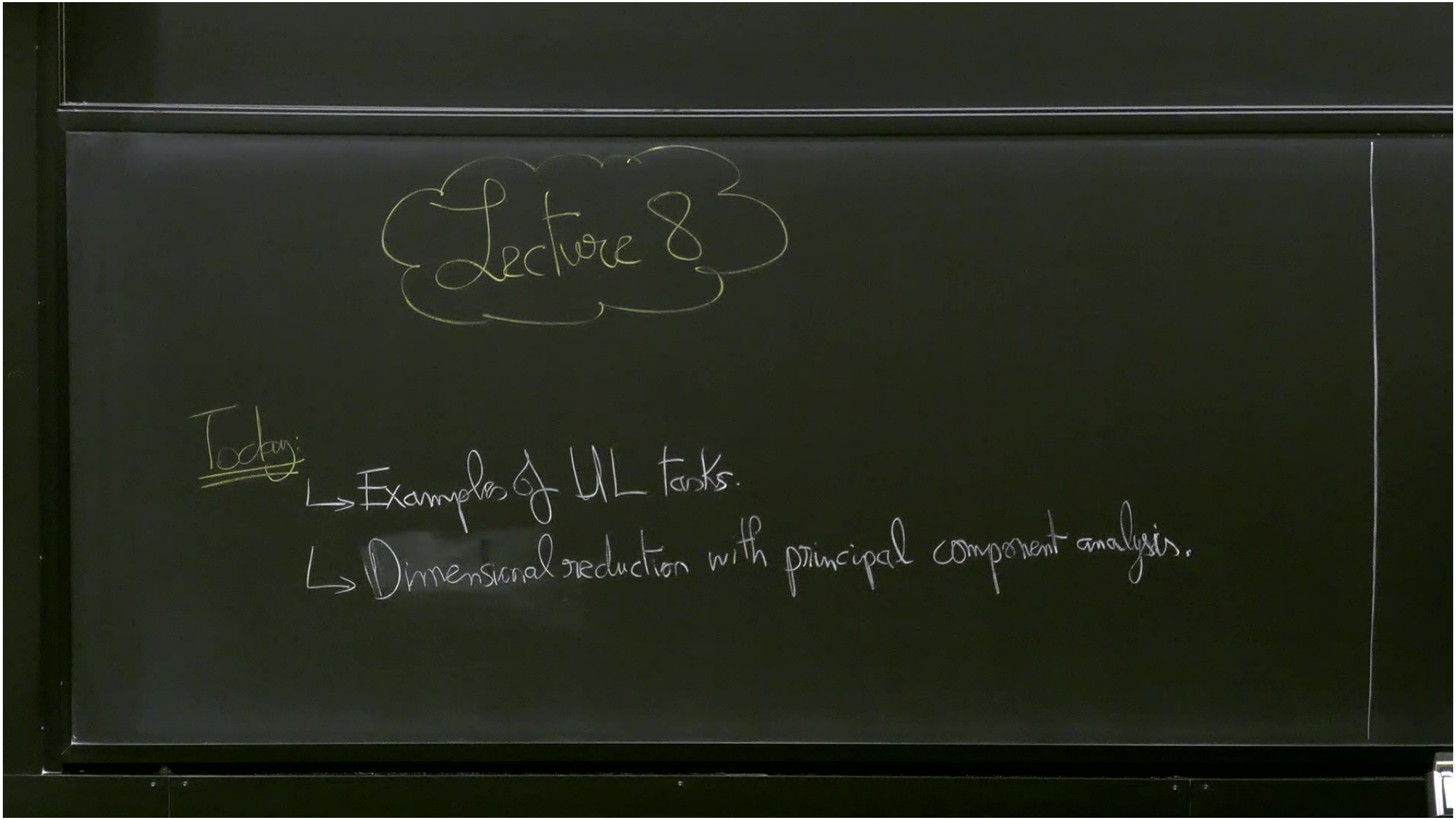
Speakers: Mohamed Hibat Allah

Collection/Series: Machine Learning (Elective), PHYS 777, February 24 - March 28, 2025

Subject: Condensed Matter, Other

Date: March 13, 2025 - 9:00 AM

URL: https://pirsa.org/25030039


Lecture 8
Today:
↳ Examples of UL tasks.
↳ Dimensional reduction with principal component analysis.

Unsupervised Learning:

Goal: learn structural properties of an unlabelled dataset.

→ Our dataset can be expressed as $D = \{\vec{X}\}$ where

each $\vec{X} = (X_1, X_2, \ldots, X_N)$

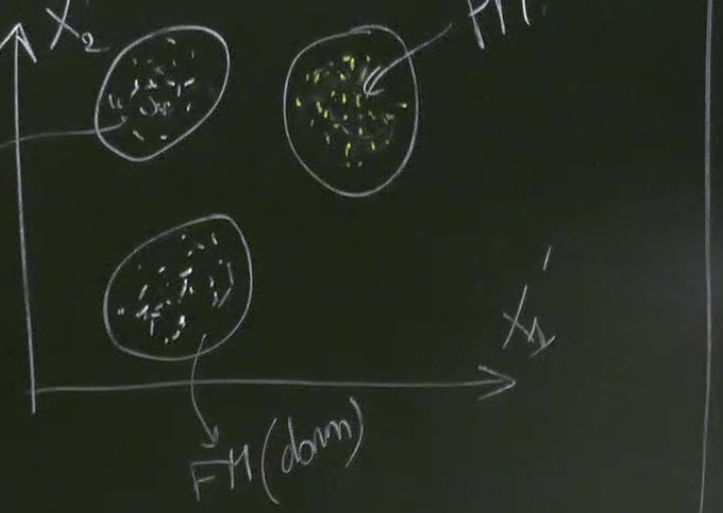
→ The dataset consists of M datapoints $D = \{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_M\}$

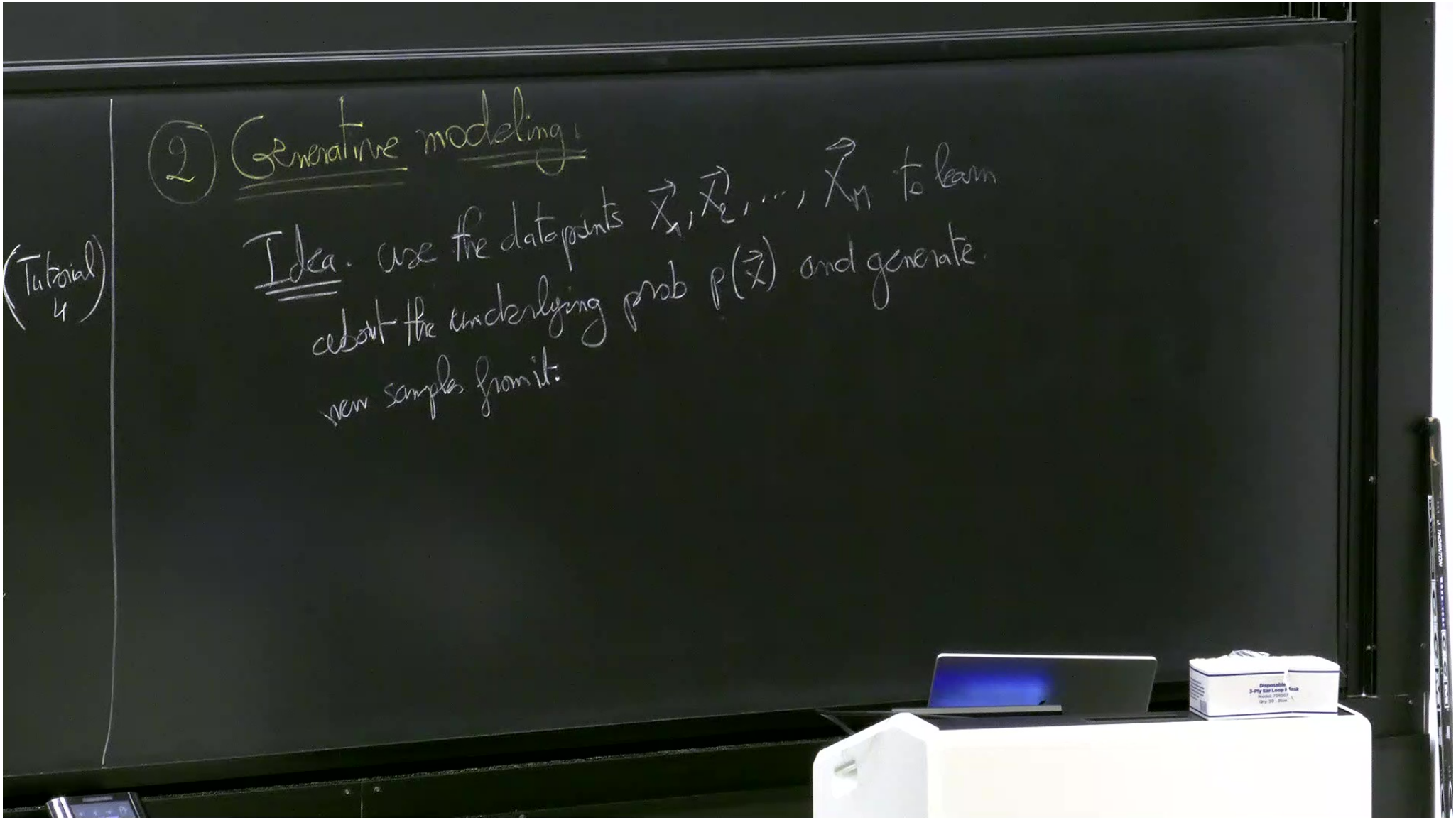
analysis.

① Dimensional reduction:

Idea: reduce the dimensionality. N of D by finding important features or correlation in D. (Tutorial 4)

$N \longrightarrow N' = 2$ or 3.



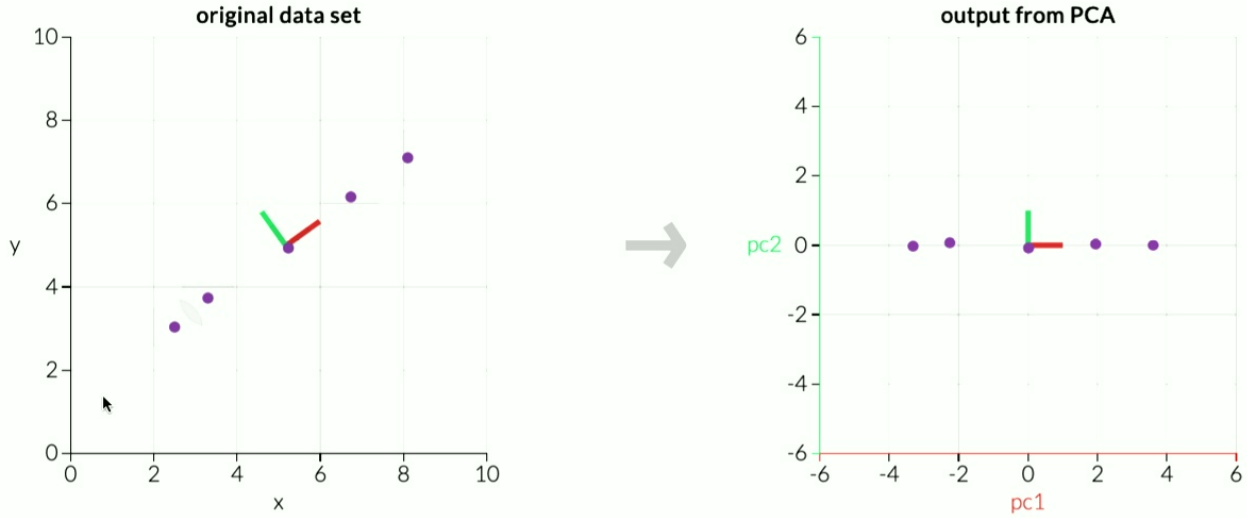

(Tutorial 4)
(2) Generative modeling:
Idea. use the datapoints $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_M$ to learn about the underlying prob $p(\vec{x})$ and generate new samples from it.

Principal component analysis (PCA).

Idea. Generate a lower dimensional representation of the N-dim datapoints by applying a linear transformation to the dataset $D = \{\vec{x}\}$.

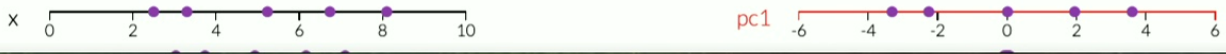
anything physical, they're combinations of height and weight called "principal components" that are chosen to give one axes lots of variation.

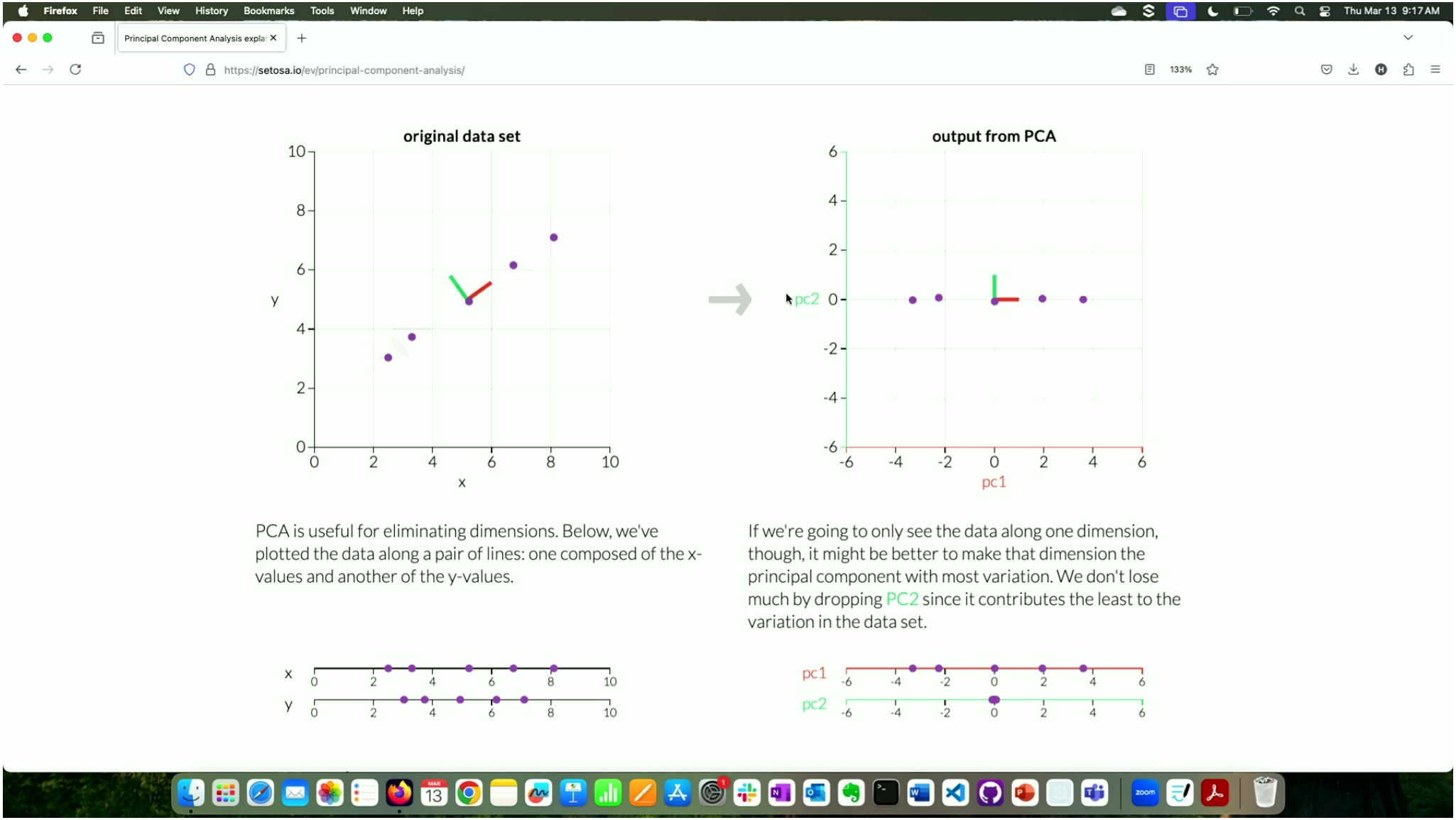
Drag the points around in the following visualization to see PC coordinate system adjusts.



PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.

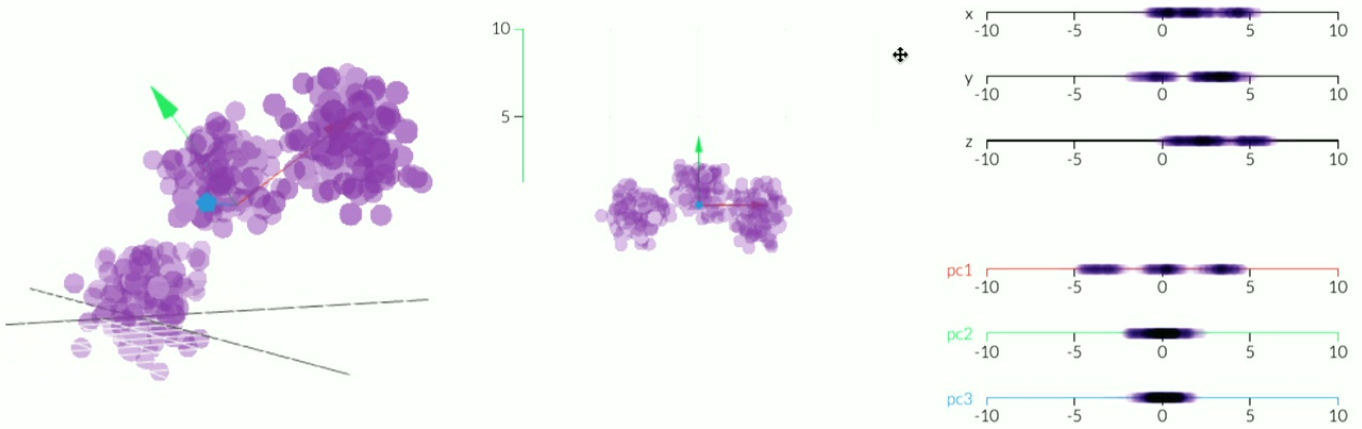
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.




Firefox File Edit View History Bookmarks Tools Window Help
Thu Mar 13 9:17 AM
Principal Component Analysis expla
https://setosa.io/ev/principal-component-analysis/
133%
original data set
output from PCA
y
x
pc2
pc1
PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.
x
y
pc1
pc2

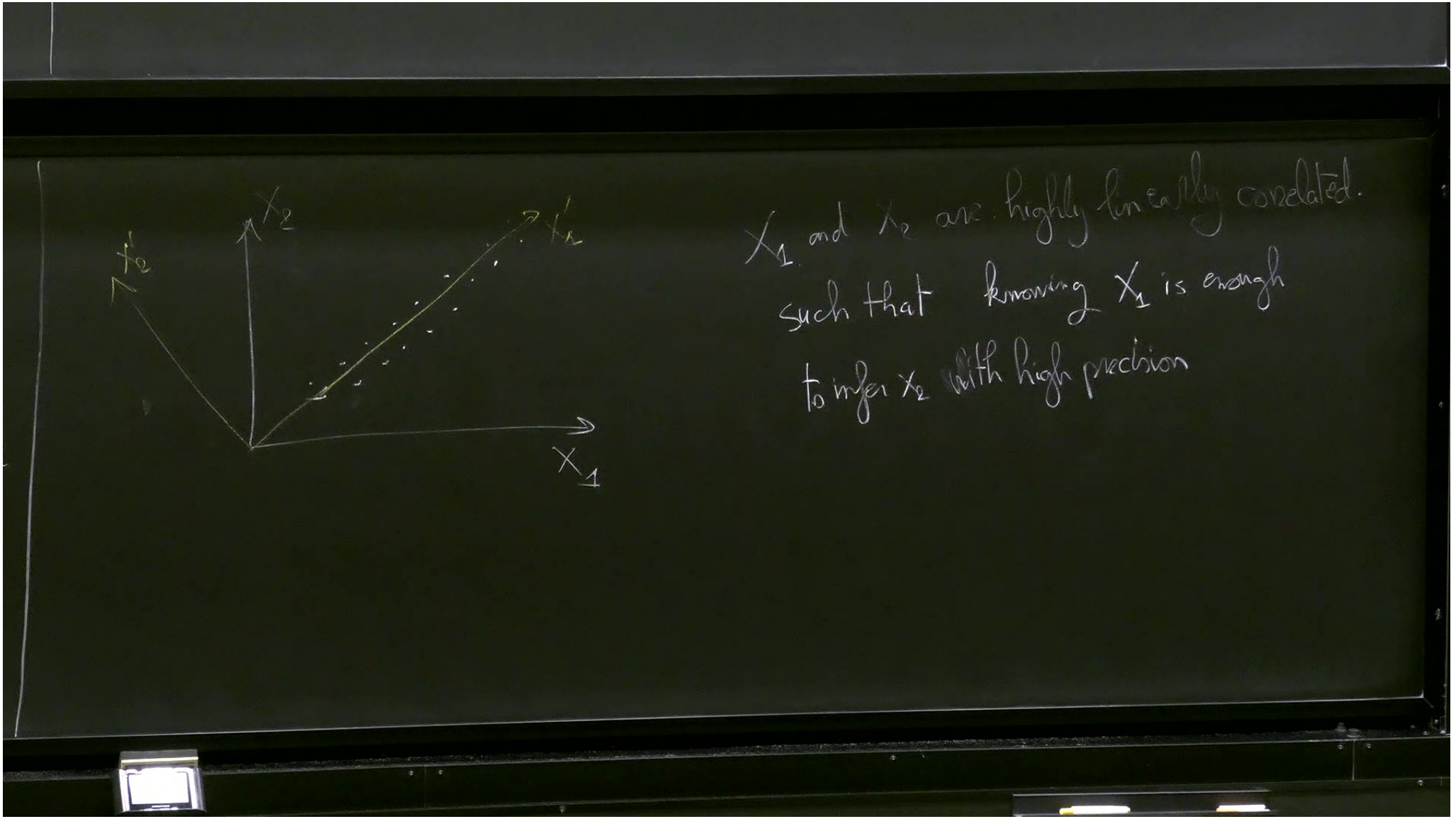
3D example

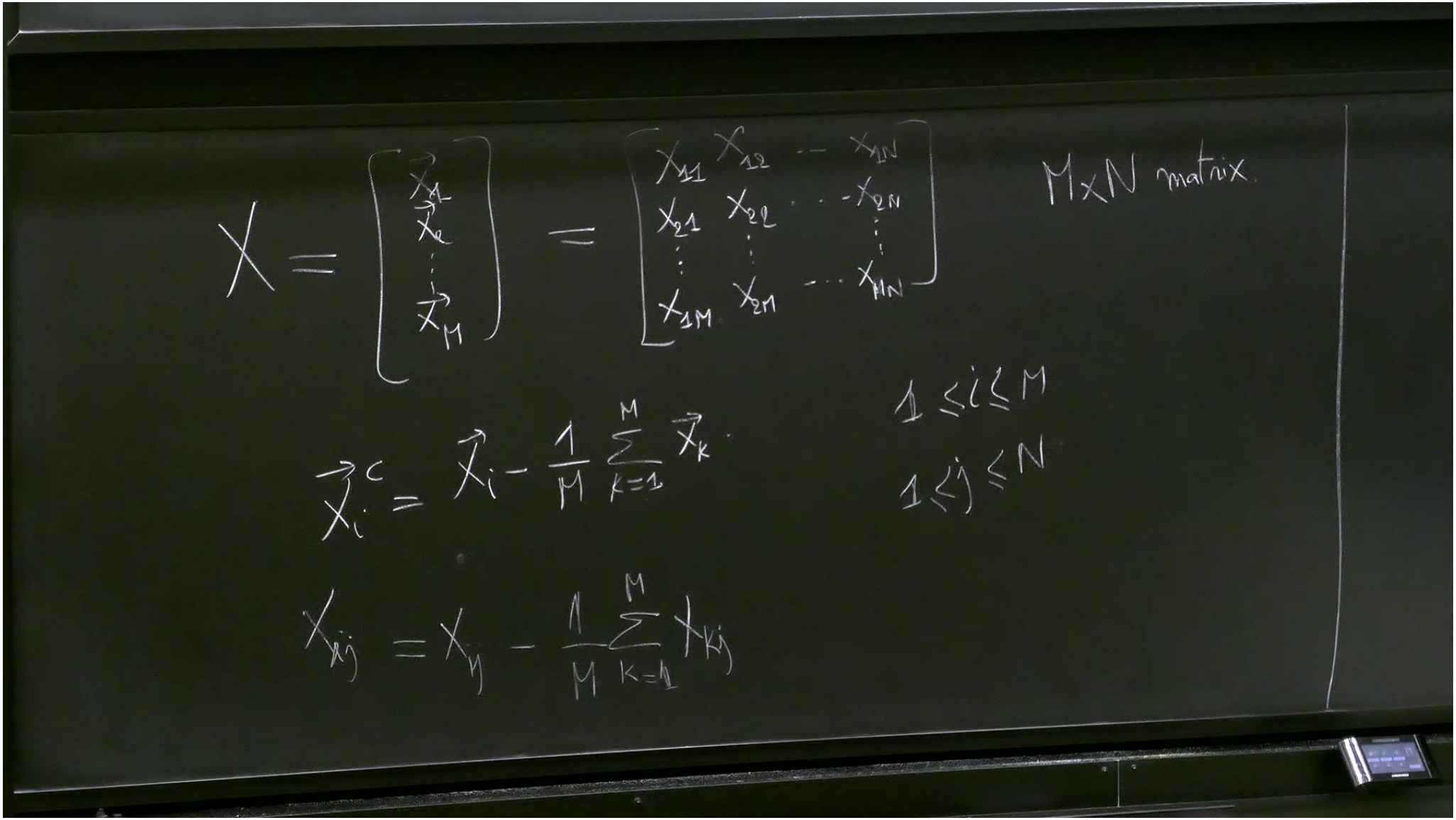
With three dimensions, PCA is more useful, because it's hard to see through a cloud of data. In the example below, the original data are plotted in 3D, but you can project the data into 2D through a transformation no different than finding a camera angle: rotate the axes to find the best angle. To see the "official" PCA transformation, click the "Show PCA" button. The PCA transformation ensures that the horizontal axis PC1 has the most variation, the vertical axis PC2 the second-most, and a third axis PC3 the least. Obviously, PC3 is the one we drop.

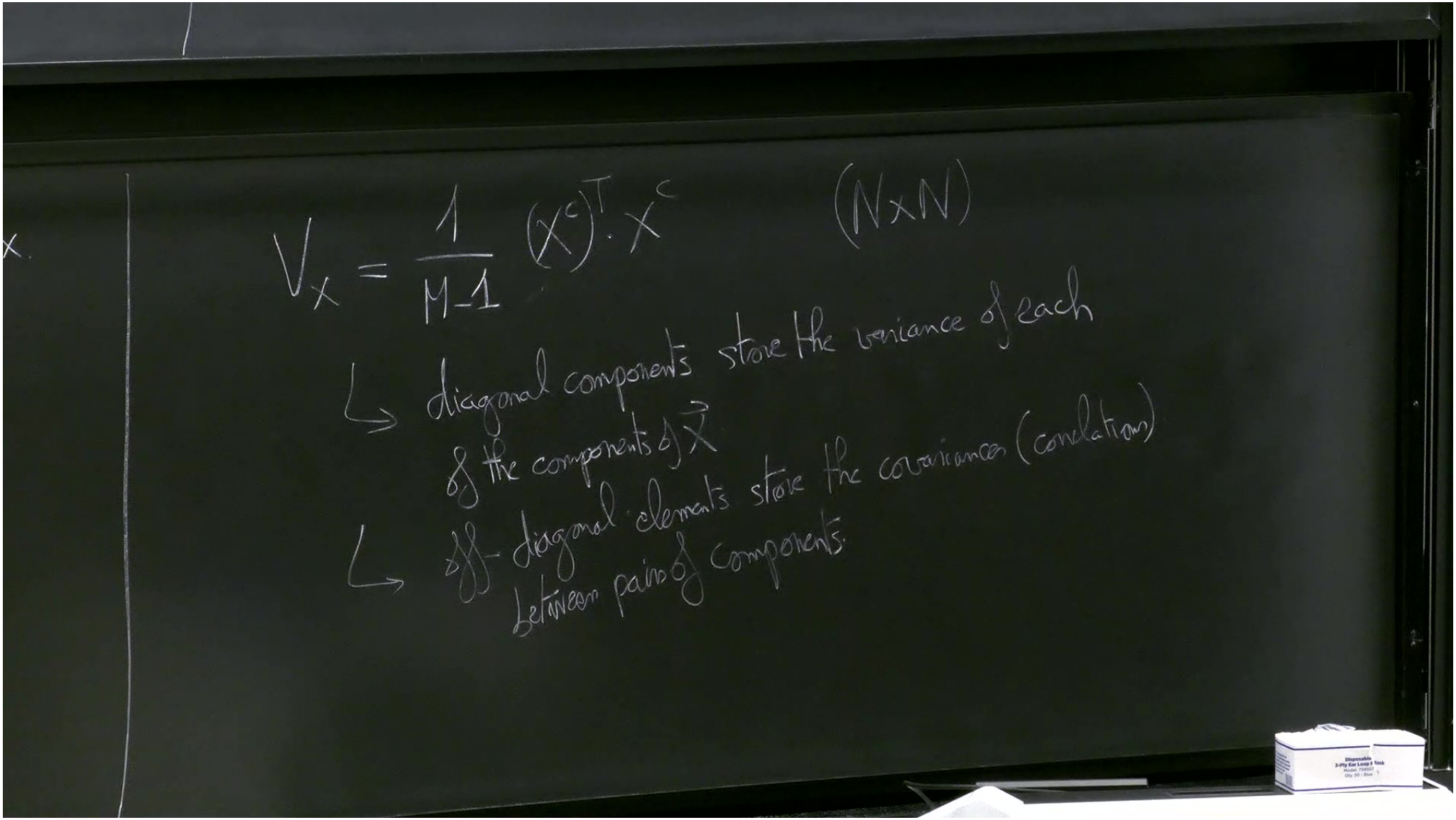


show PCA reset

Eating in the UK (a 17D example)


x_2'
x_2
x_1'
x_1
X_1 and X_2 are highly linearly correlated.
such that knowing X_1 is enough
to infer X_2 with high precision


$X = \begin{bmatrix} \vec{x}_1 \\ \vdots \\ \vec{x}_M \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & & x_{2N} \\ \vdots & \vdots & & \vdots \\ x_{1M} & x_{2M} & \cdots & x_{MN} \end{bmatrix}$
$M \times N$ matrix.
$\vec{x}_i^{\,c} = \vec{x}_i - \frac{1}{M} \sum_{k=1}^{M} \vec{x}_k$
$1 \le i \le M$
$1 \le j \le N$
$x_{ij} = x_{ij} - \frac{1}{M} \sum_{k=1}^{M} x_{kj}$

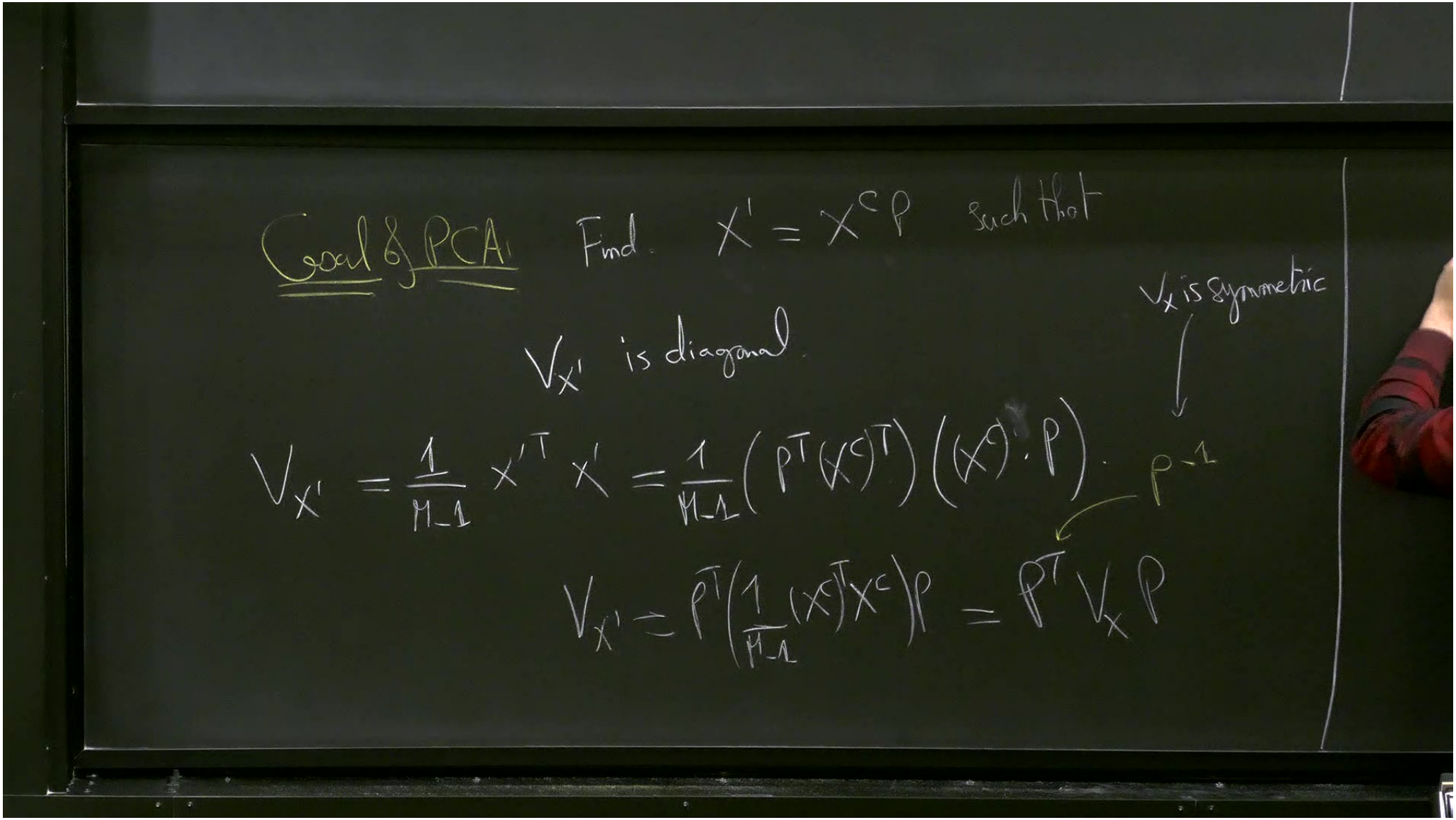

$V_x = \frac{1}{M-1} (X^c)^T \cdot X^c$ $(N \times N)$
↳ diagonal components store the variance of each of the components of $\vec{X}$
↳ off-diagonal elements store the covariances (correlations) between pairs of components.

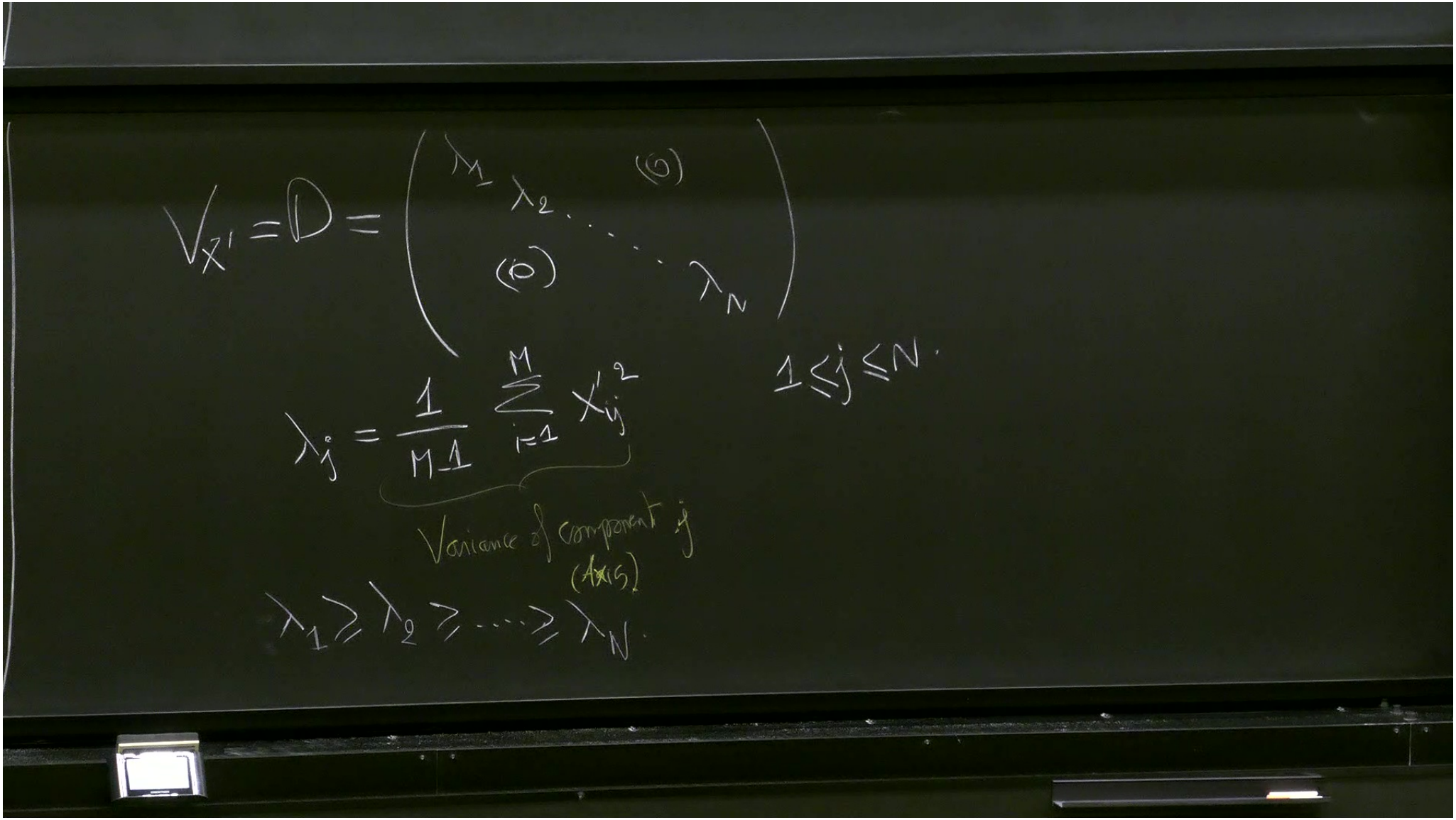
Goal of PCA: Find $X' = X^c P$ such that

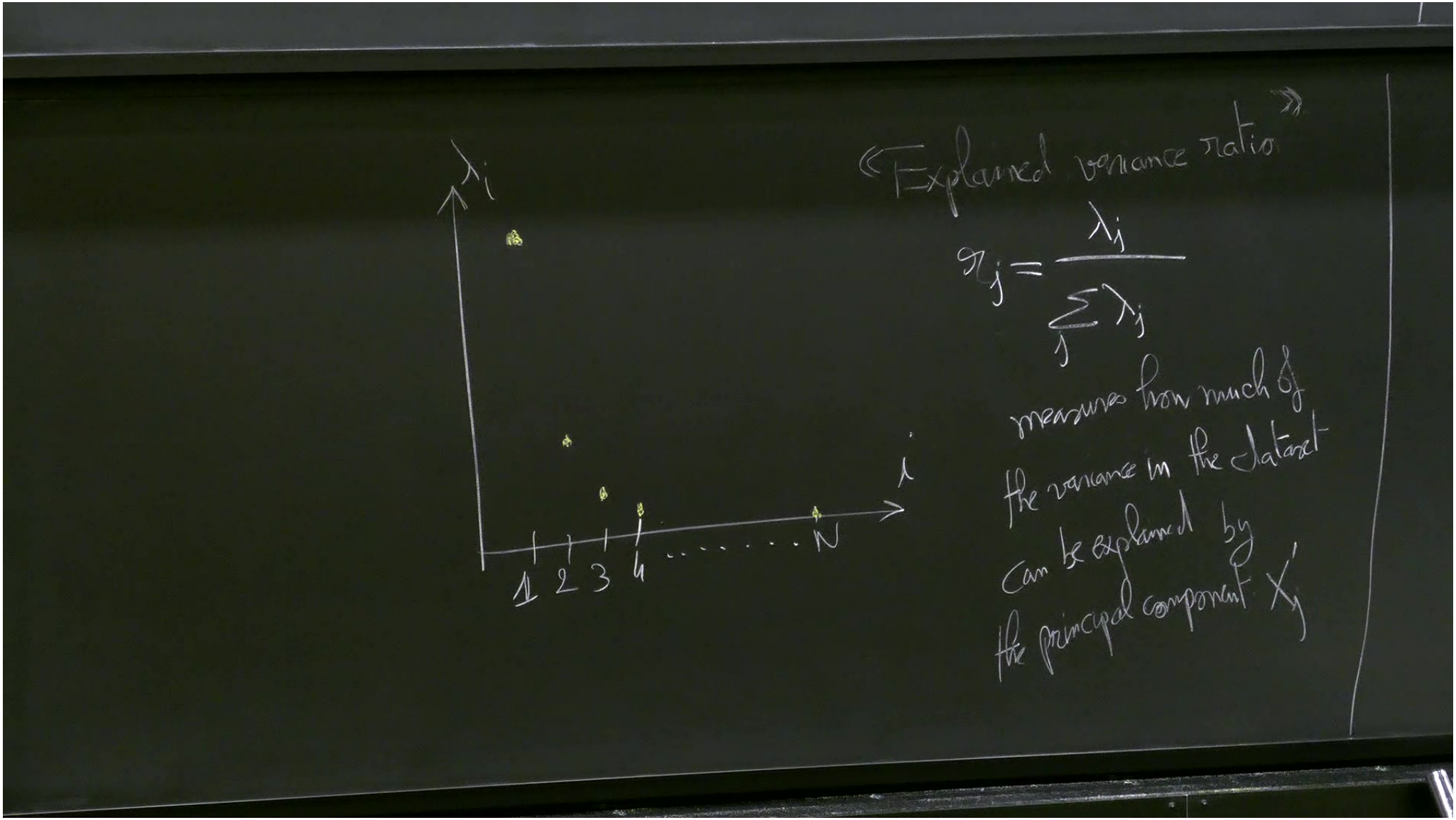
$V_{X'}$ is diagonal.

$$V_{X'} = \frac{1}{M-1} X'^T X' = \frac{1}{M-1} \left(P^T (X^c)^T \right) \left((X^c) P \right).$$

$$V_{X'} = P^T \left(\frac{1}{M-1} (X^c)^T X^c \right) P = P^T V_X P$$


Goal of PCA: Find $X' = X^c P$ such that
$V_{X'}$ is diagonal.
V_X is symmetric
$V_{X'} = \frac{1}{M-1} X'^T X' = \frac{1}{M-1} (P^T (X^c)^T)(X^c P)$
P^{-1}
$V_{X'} = P^T \left(\frac{1}{M-1} (X^c)^T X^c \right) P = P^T V_X P$

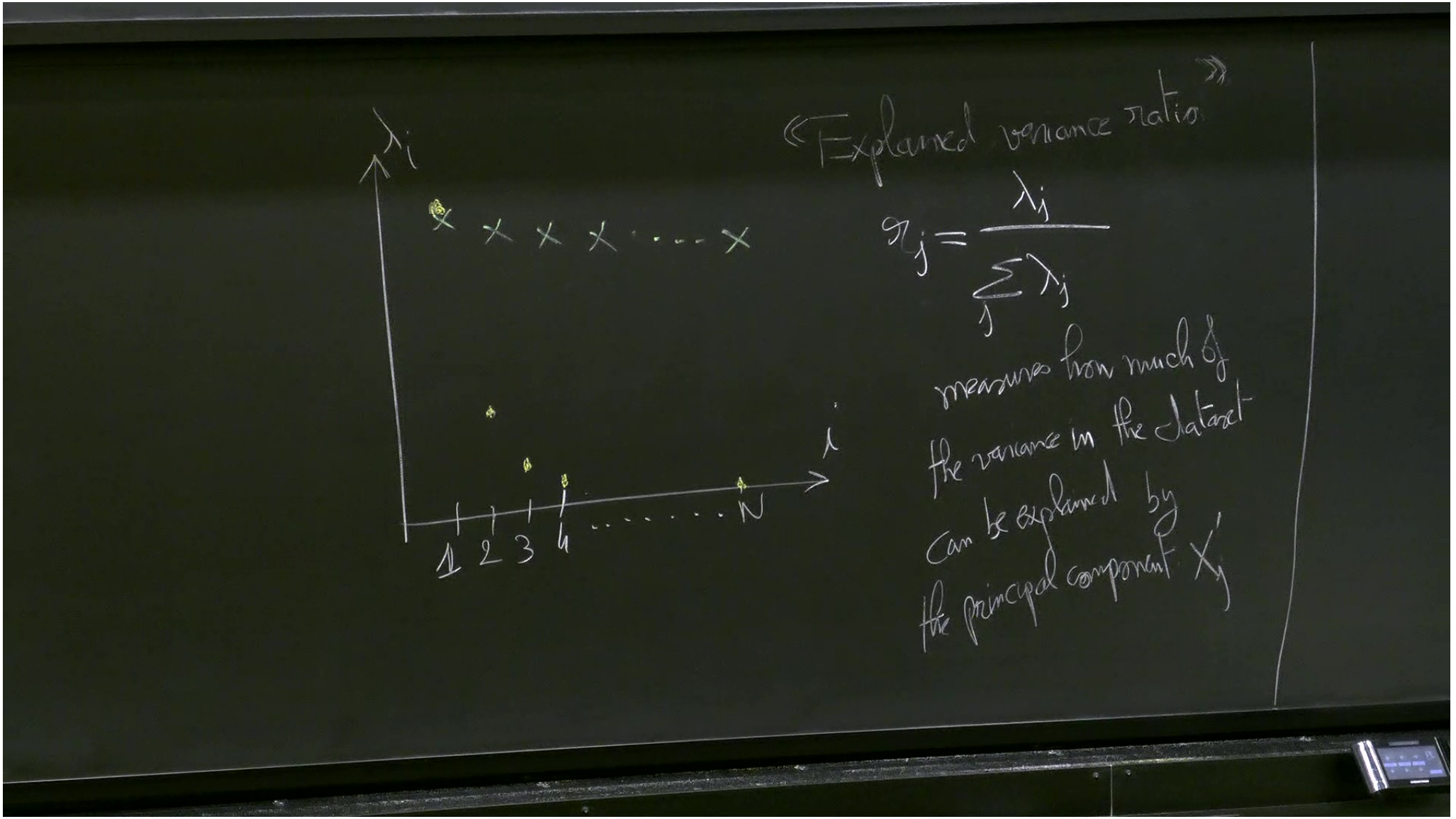

$V_{X'} = D = \begin{pmatrix} \lambda_1 & & (0) \\ & \lambda_2 & \\ & & \ddots \\ (0) & & \lambda_N \end{pmatrix}$
$\lambda_j = \frac{1}{M-1} \sum_{i=1}^{M} X'^2_{ij}$
$1 \le j \le N.$
Variance of component j
(Axis)
$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N.$



«Explained variance ratio»

$$r_j = \frac{\lambda_j}{\sum_j \lambda_j}$$

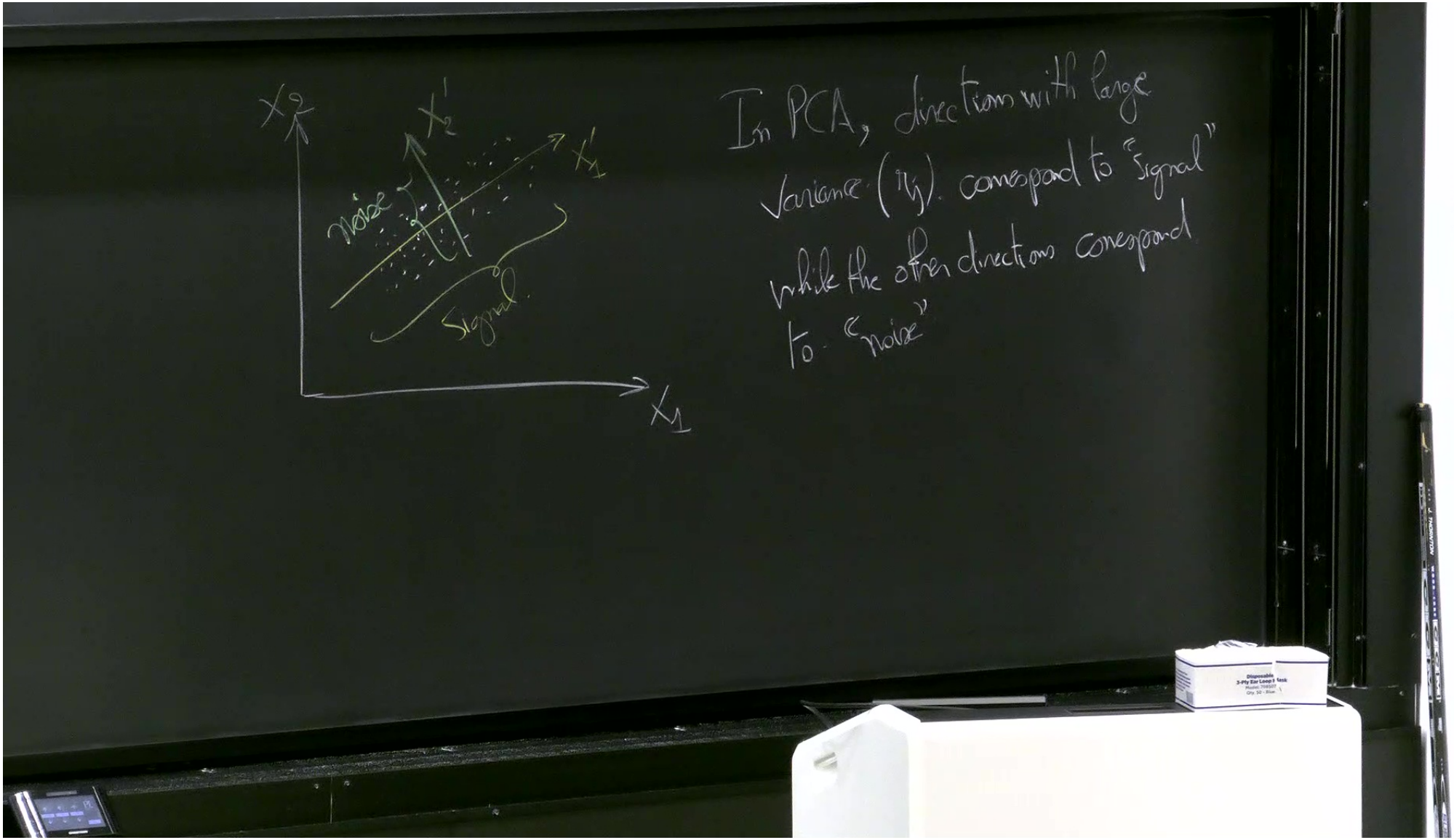
measures how much of the variance in the dataset can be explained by the principal component X_j



«Explained variance ratio»

$$r_j = \frac{\lambda_j}{\sum_j \lambda_j}$$

measures how much of the variance in the dataset can be explained by the principal component X_j


noise
Signal
In PCA, directions with large
variance (λ_i) correspond to "signal"
while the other directions correspond
to "noise"

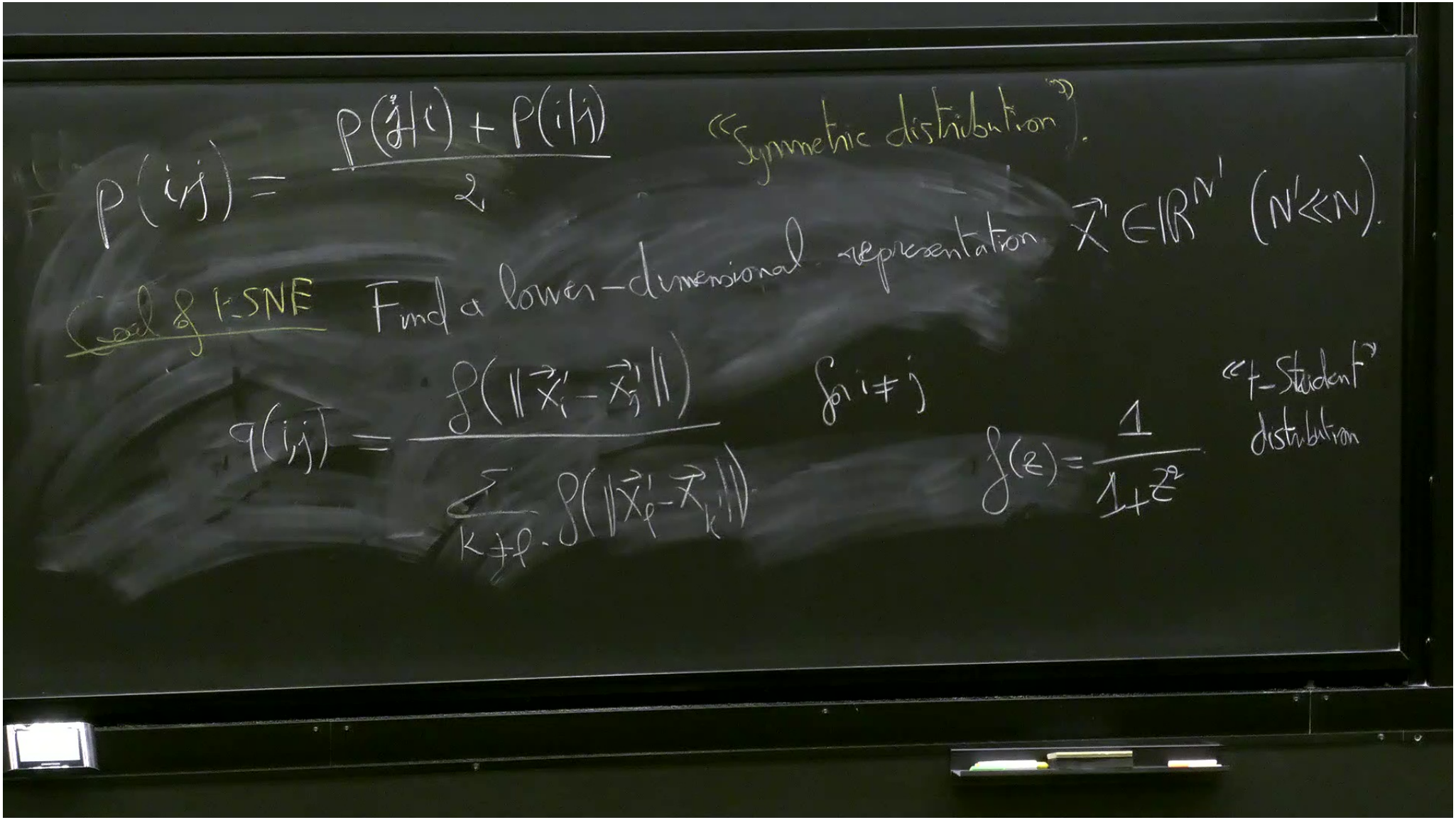
Dimensional reduction with t-SNE (t-distributed Stochastic neighbour embedding)

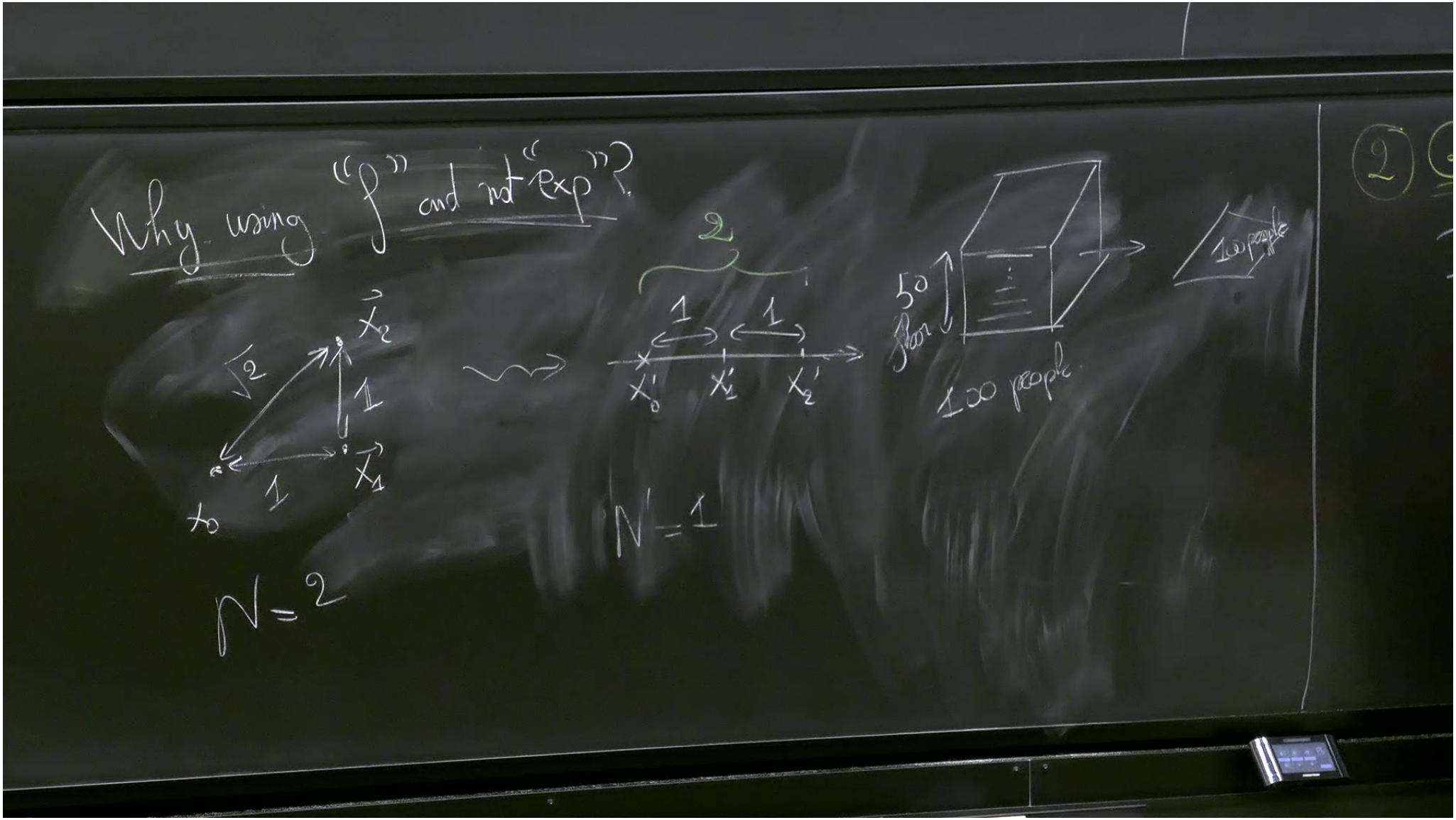
$$p(j|i) = \frac{\exp\left(-\|\vec{x}_i - \vec{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\vec{x}_i - \vec{x}_k\|^2 / 2\sigma_i^2\right)} \qquad (i \neq j)$$

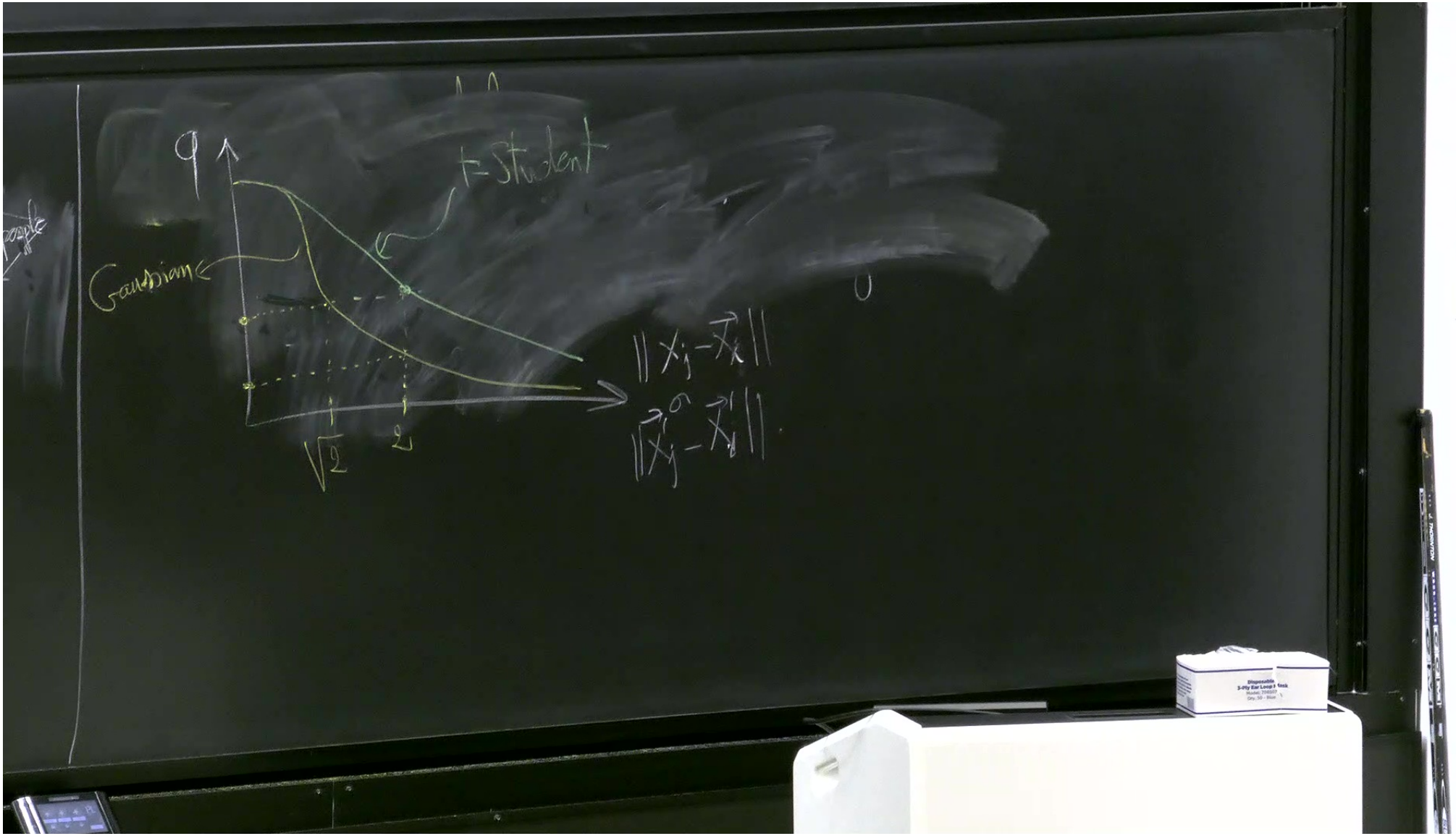
↓

The probability that $\vec{x}_i$ would choose $\vec{x}_j$ as its neighbour...

σ_i are related to the perplexity (Hyperparameter) → See HW 2.


$p(i,j) = \frac{p(j|i) + p(i|j)}{2}$
"Symmetric distribution"
Goal of t-SNE: Find a lower-dimensional representation $\vec{x}' \in \mathbb{R}^{N'}$ $(N' \ll N)$.
$q(i,j) = \frac{f(\|\vec{x}'_i - \vec{x}'_j\|)}{\sum_{k \neq l} f(\|\vec{x}'_k - \vec{x}'_l\|)}$ for $i \neq j$
$f(z) = \frac{1}{1+z^2}$
"t-Student" distribution


Why using "ρ" and not "exp"?
$\vec{x}_2$
$\sqrt{2}$
1
1
$\vec{x}_1$
x_0
N = 2
2
1
1
x'_0
x'_1
x'_2
N = 1
50 floor
100 people
100 people


t-Student
Gaussian
$\sqrt{2}$
2
$\|\vec{x}_j - \vec{x}_i\|$
$\|\vec{x}_j^0 - \vec{x}_i^0\|$