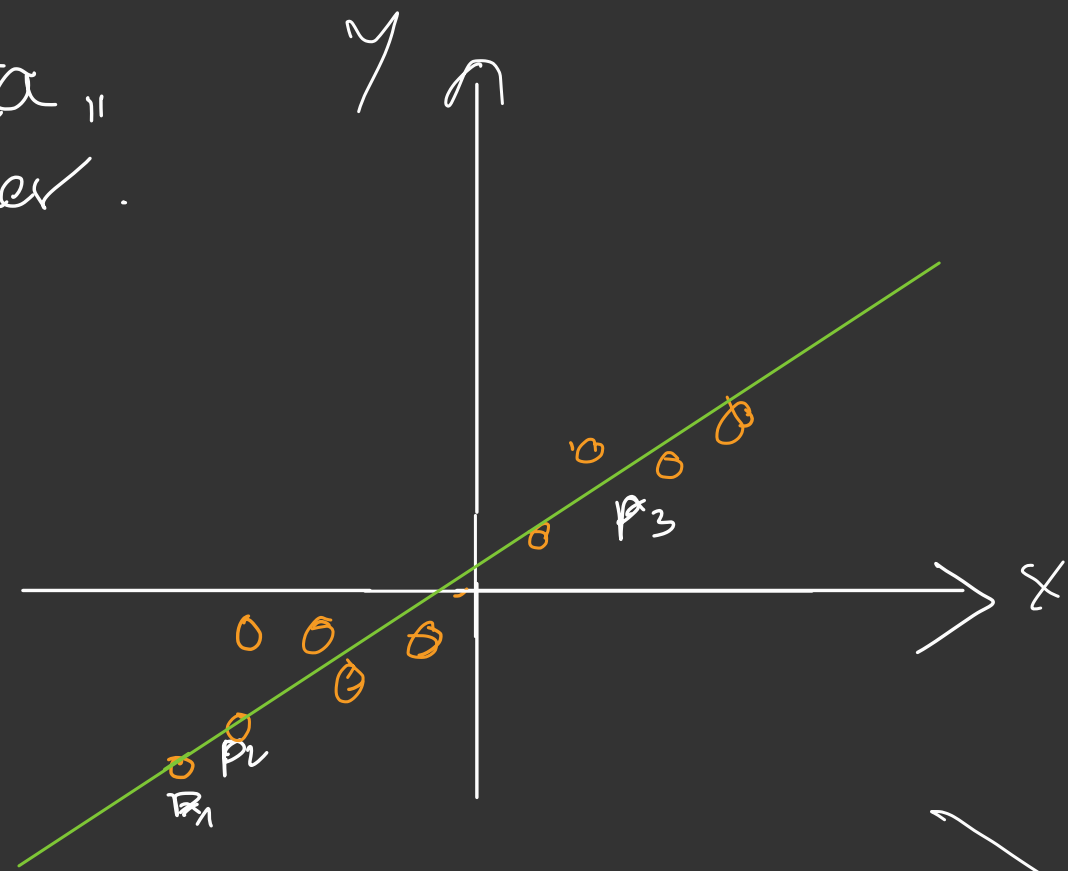


Unsupervised Learning: "Understand data" without teacher.

Principal Component Analysis (PCA)

Example: 2D dataset $S = (x_i, y_i)_{i=1}^n$

$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} \in \mathbb{R}^{n \times 2}$$



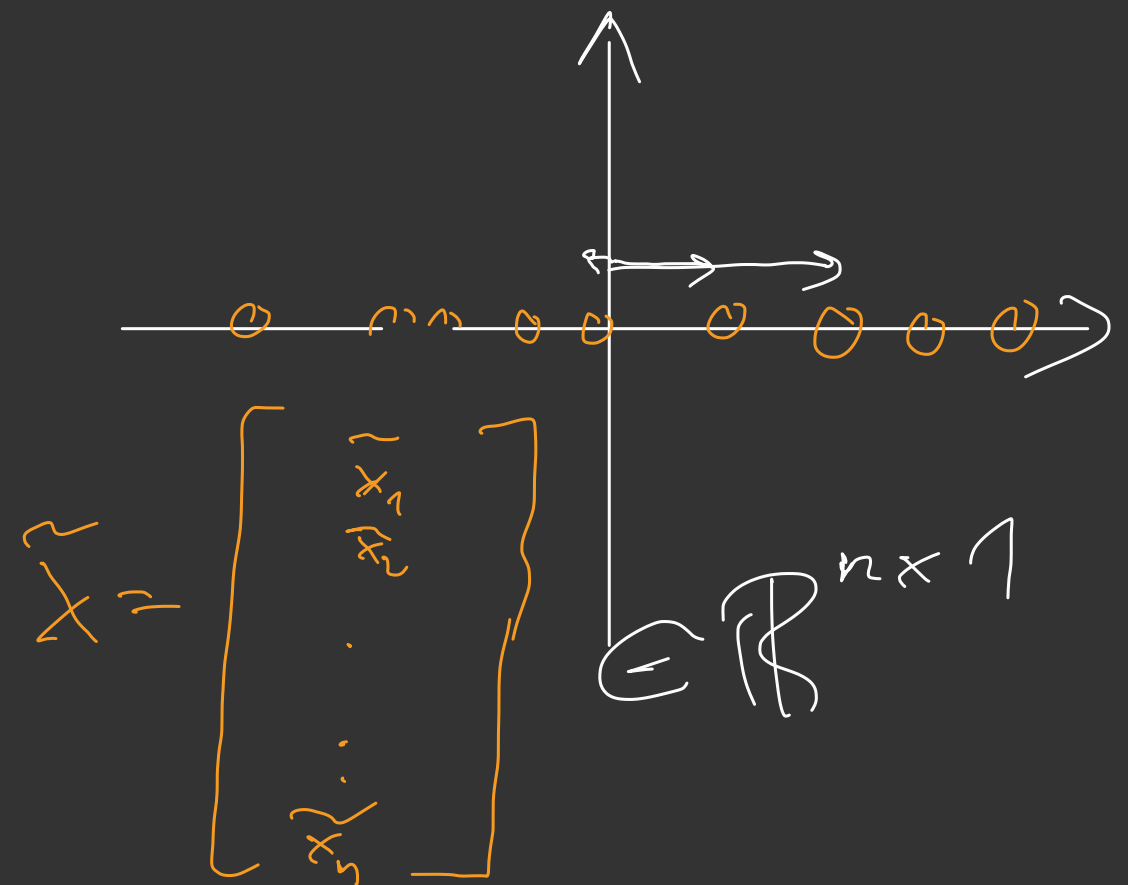
PCA

Looks like "almost" a line, can we (as an approximation) represent S in 1D?

Idea: Dimension reduction for

▷ interpretability

▷ downstream computational savings



Human genetics

Single Nucleotide Polymorphisms: the most common type of genetic variation in the genome across different individuals.

They are **known** locations at the human genome where **two** alternate nucleotide bases (**alleles**) are observed (out of A, C, G, T).

SNPs



Matrices including thousands of individuals and hundreds of thousands if SNPs are available.

Mahoney & Prineas 2015

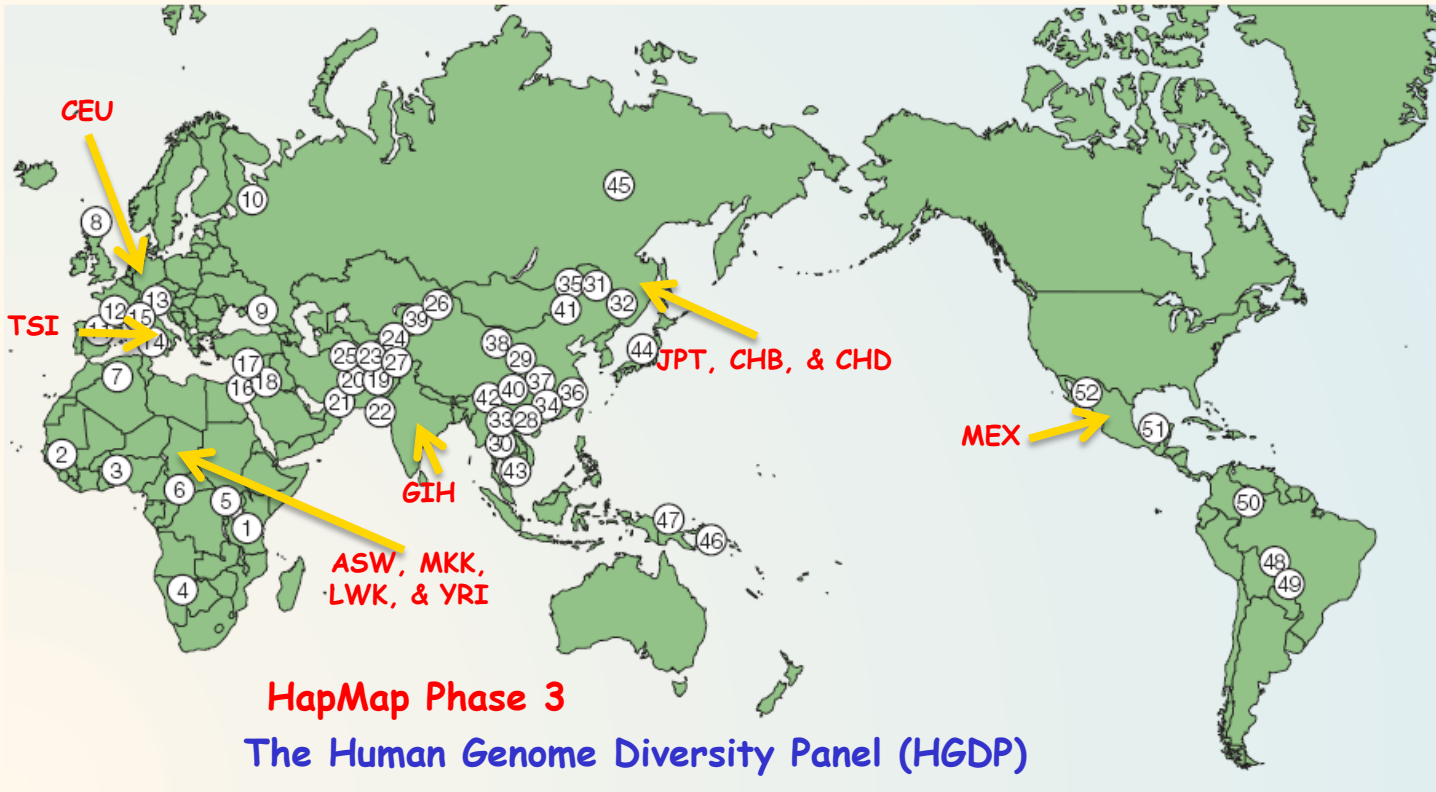
HGDP data

- 1,033 samples
- 7 geographic regions
- 52 populations

HapMap Phase 3 data

- 1,207 samples
- 11 populations

We will apply SVD/PCA on the (joint) HGDP and HapMap Phase 3 data.



Africans

- 1 Bantu
- 2 Mandenka
- 3 Yoruba
- 4 San
- 5 Mbuti pygmy
- 6 Biaka
- 7 Mozabite

Europeans

- 8 Orcadian
- 9 Adygei
- 10 Russian
- 11 Basque
- 12 French
- 13 North Italian
- 14 Sardinian
- 15 Tuscan

Western Asians

- 16 Bedouin
- 17 Druze
- 18 Palestinian

Central and Southern Asians

- 19 Balochi
- 20 Brahui
- 21 Makrani
- 22 Sindhi
- 23 Pathan
- 24 Burusho
- 25 Hazara
- 26 Uygur
- 27 Kalash

Eastern Asians

- 28 Han (S. China)
- 29 Han (N. China)
- 30 Dai
- 31 Daur
- 32 Hezhen
- 33 Lahu
- 34 Miao
- 35 Oroqen
- 36 She
- 37 Tujia
- 38 Tu
- 39 Xibo
- 40 Yi
- 41 Mongola
- 42 Naxi
- 43 Cambodian
- 44 Japanese
- 45 Yakut

Oceanians

- 46 Melanesian
- 47 Papuan

Native Americans

- 48 Karitiana
- 49 Surui
- 50 Colombian
- 51 Maya
- 52 Pima

Cavalli-Sforza (2005) *Nat Genet Rev*

Rosenberg et al. (2002) *Science*

Li et al. (2008) *Science*

The International HapMap Consortium
(2003, 2005, 2007) *Nature*

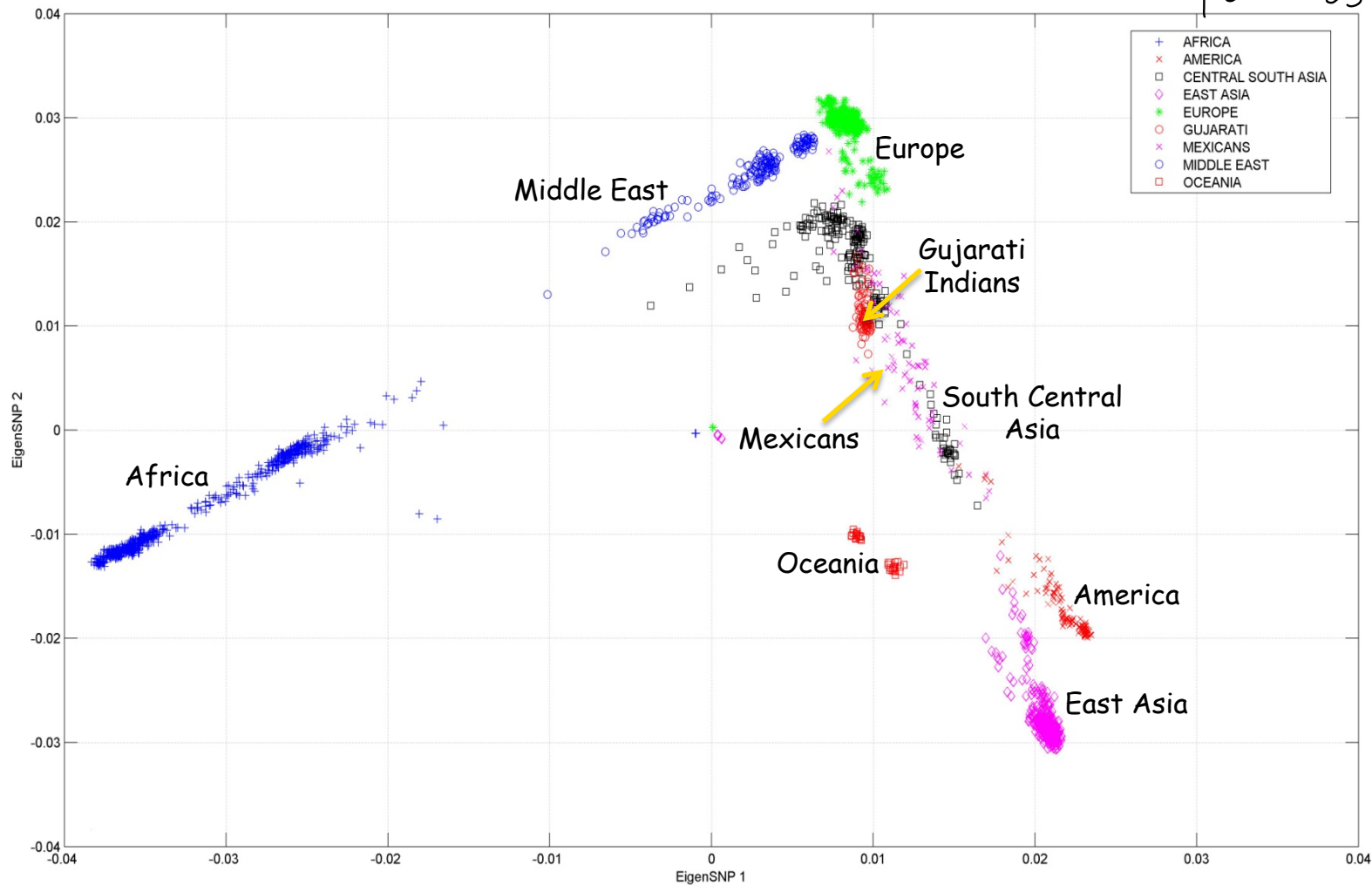
Matrix dimensions:

2,240 subjects (rows)

447,143 SNPs (columns)

Dense matrix:

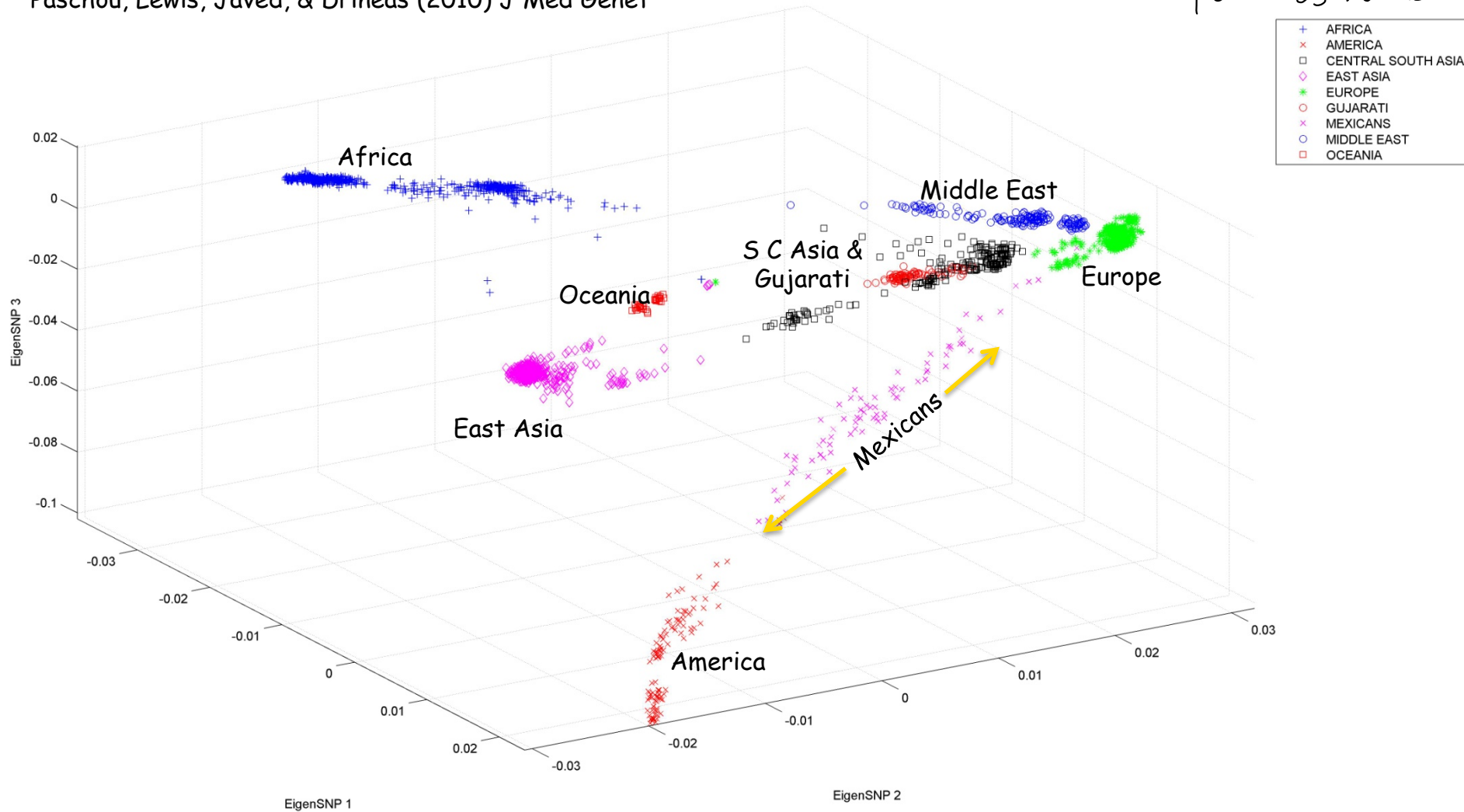
over one billion entries



- Top two Principal Components (PCs or eigenSNPs)

(Lin and Altman (2005) *Am J Hum Genet*)

- The figure renders visual support to the "out-of-Africa" hypothesis.
- Mexican population seems out of place: we move to the top three PCs.



Not altogether satisfactory: the principal components are linear combinations of all SNPs, and - of course - can not be assayed!

Can we find **actual SNPs** that capture the information in the singular vectors?

Formally: **spanning the same subspace.**

Setting of PCA: Let $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times k}$ Ex: n : nr. of subjects
 k : nr. of SKPs.

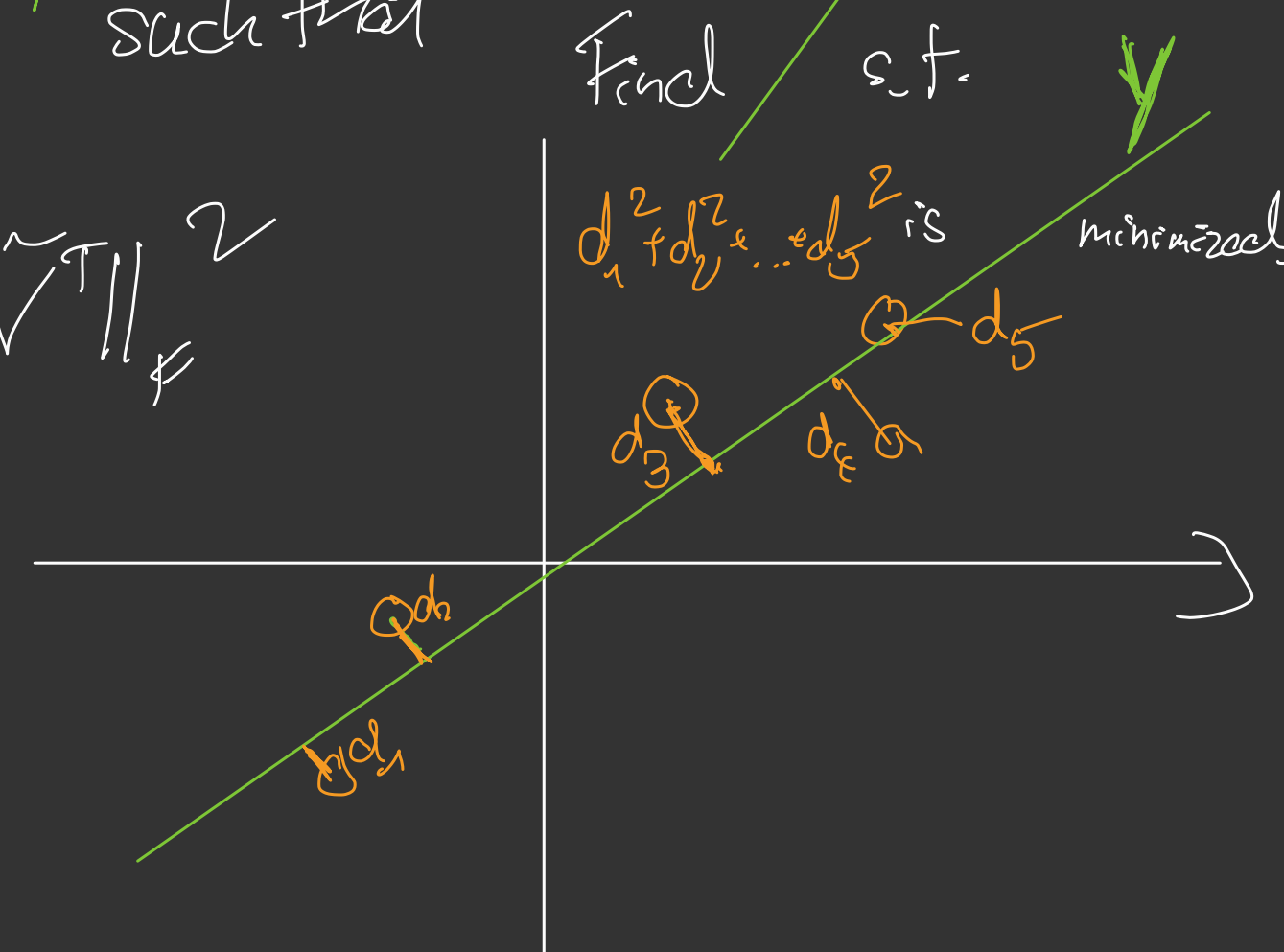
be a matrix of n data points with k features.

Assume without loss of generality that X is centered, i.e. $\frac{1}{n}(1, 1, \dots, 1)^T X = (0, \dots, 0)$
 (average value of each feature = 0).

Goal: For $r \leq k$, find r -dimensional subspace $Y \subset \mathbb{R}^k$, $\dim(Y) = r$
 and orthonormal basis $V = \begin{bmatrix} v_1 & \dots & v_r \end{bmatrix} \in \mathbb{R}^{k \times r}$ such that

$V = \underset{\tilde{V} \in \mathbb{R}^{k \times r}: \tilde{V}^T \tilde{V} = I_r}{\text{argmin}} \frac{1}{n} \|X - X \tilde{V} \tilde{V}^T\|_F^2$

"minimize sum of squares of distances between points and projected points"



Properties of PCA: ▷ Obtain (approximate) low-dimensional representations

$$\tilde{Z} := X V = \begin{bmatrix} -\tilde{z}_1 \\ \vdots \\ -\tilde{z}_m \end{bmatrix} \in \mathbb{R}^{n \times \tau}$$

of the data points $\{x_1, \dots, x_n\}$.

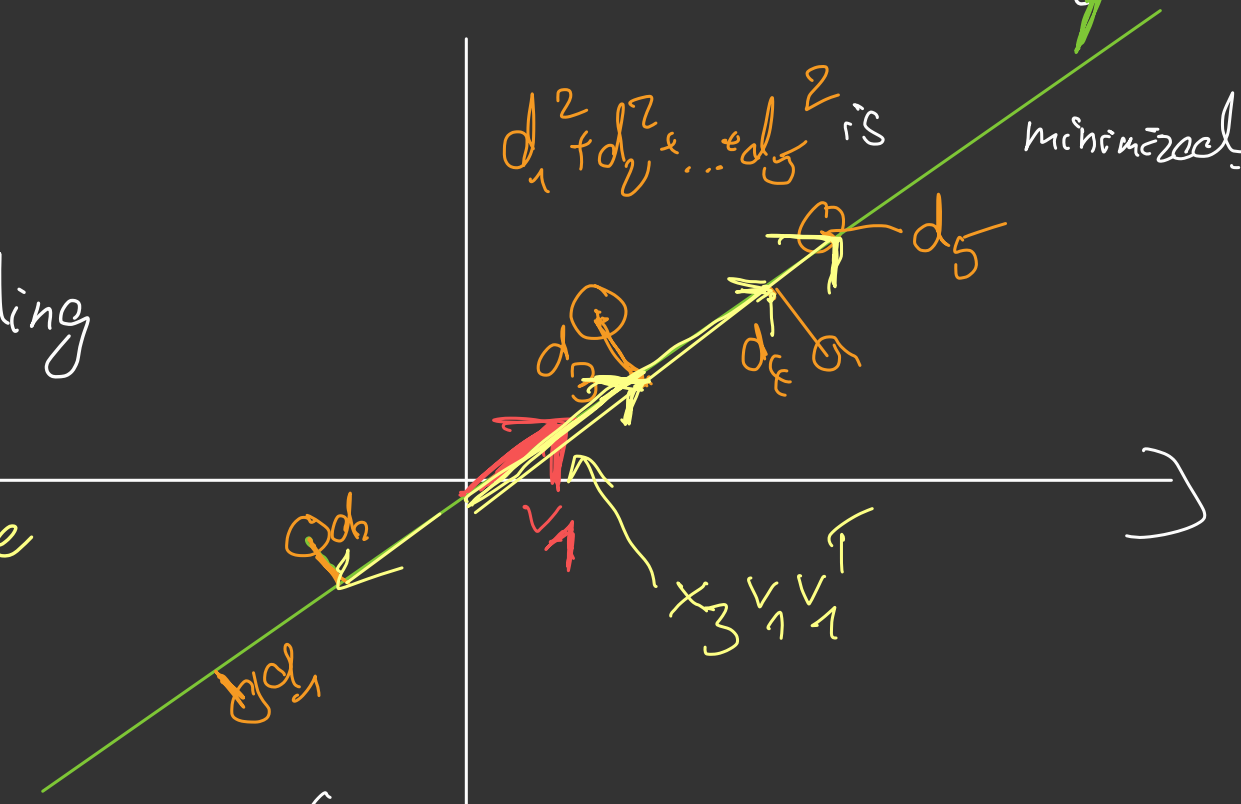
$$E_x: \begin{matrix} n & k \\ 2240 & \times & 447143 \\ \hline n & \tau \\ 2240 & \times & 9 \end{matrix}$$

▷ Memory savings: $\mathcal{O}(n \tau)$ instead of $\mathcal{O}(nk)$ parameters

▷ $V = \begin{bmatrix} | & & | \\ v_1 & \dots & v_\tau \\ | & & | \end{bmatrix}$: Matrix w/ eigenvectors corresponding to τ largest eigenvalues of empirical covariance matrix

$$M = \frac{1}{n-1} \sum_{i=1}^n x_i x_i^T = \frac{1}{n-1} X^T X \in \mathbb{R}^{k \times k}$$

▷ PCA finds orthogonal directions of maximal variance (interpretation favored in statistics)



$$V = \underset{\tilde{V} \in \mathbb{R}^{k \times \tau}: \tilde{V}^T \tilde{V} = I_\tau}{\text{argmin}} \frac{1}{n} \|X - X \tilde{V} \tilde{V}^T\|_F^2 = \frac{1}{n} \|X^T X\|_F^2 - \text{tr}\left(\tilde{V}^T \frac{1}{n} X^T X \tilde{V}\right)$$

$$\text{Var}_x (x^T v_1) = \mathbb{E} [(x^T v_1)^2] = \frac{1}{n} \sum_{i=1}^n \frac{1}{n} \|x_i^T v_1\|^2 = \frac{1}{n} \sum_{i=1}^n v_1^T x_i x_i^T v_1$$

Terminology: ▷ Columns v_i of $V = \begin{bmatrix} | & & | \\ v_1 & \dots & v_r \\ | & & | \end{bmatrix}$

"Principal directions / components / axes"

▷ Columns $x v_i$ of $XV \in \mathbb{R}^{n \times r}$

"Principal components" / "Scores"

▷ λ_i : i -th eigenvalue of M : "Variance explained by i -th PC"

▷ $\sqrt{\lambda_i} v_i$: " i -th loading"