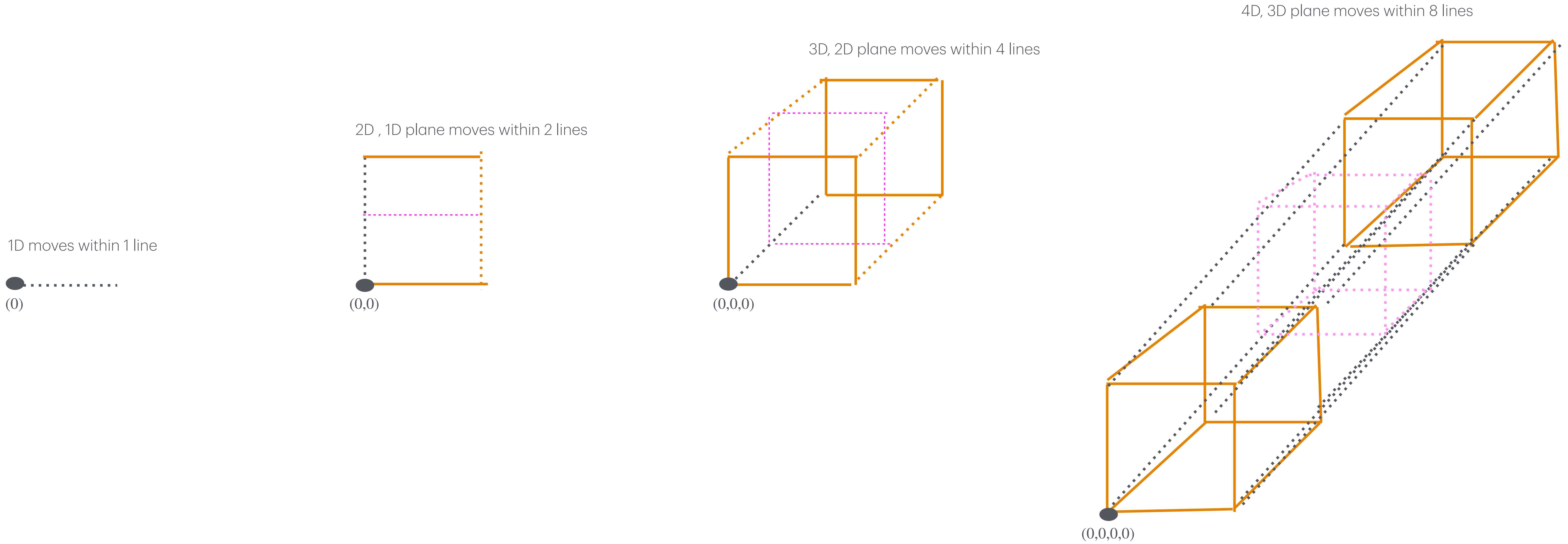
Dimension Reduction

Dimension

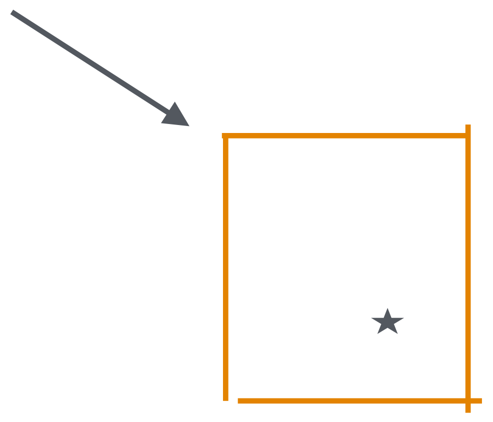


(0,0,0,0)

Prior dimension capable of moving its plane in next dimension

2D

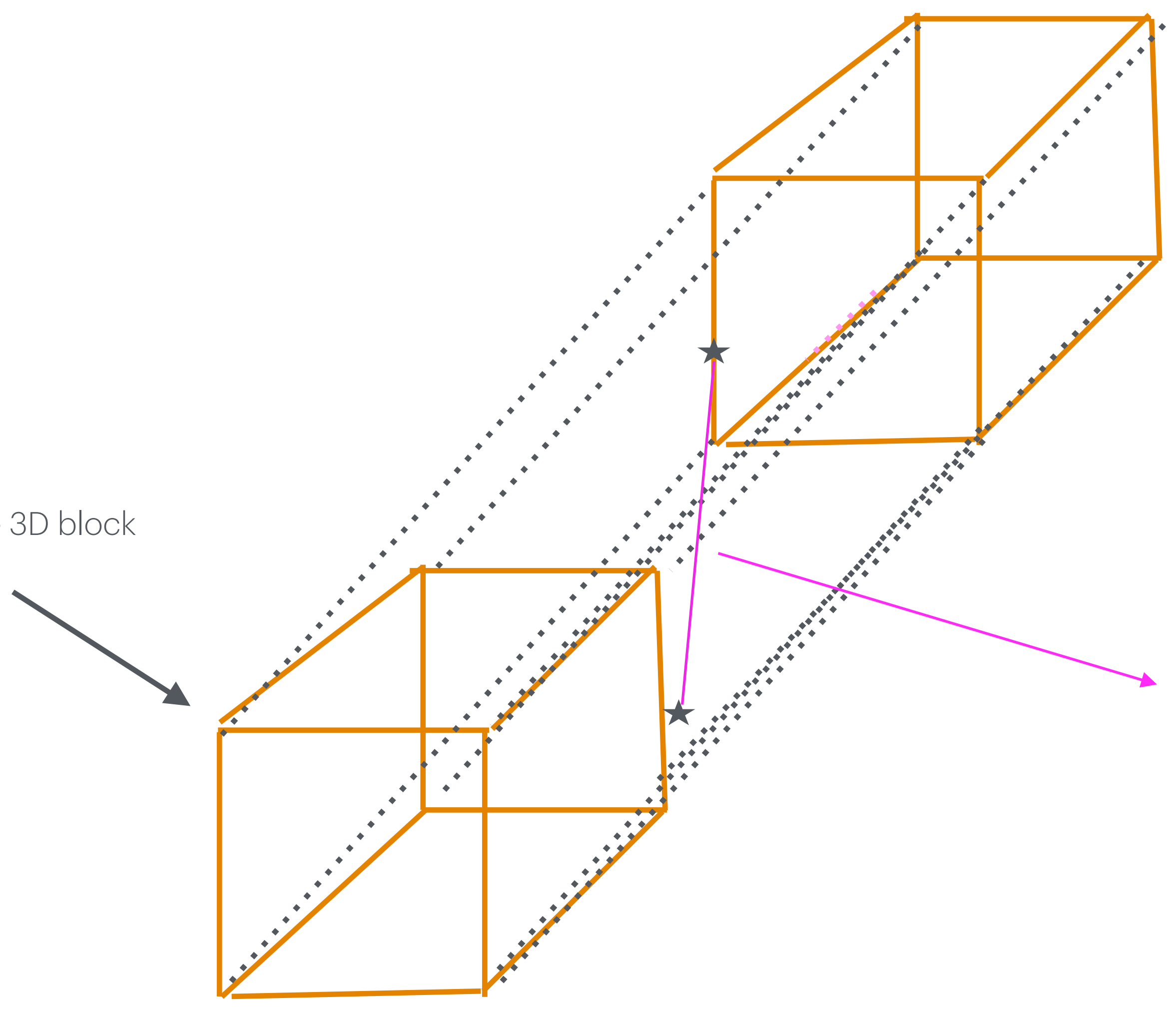
Border is this 2D plane



Lower probability of a random point touching the border

4D

Border is this entire 3D block



Higher probability of a random point touching the border

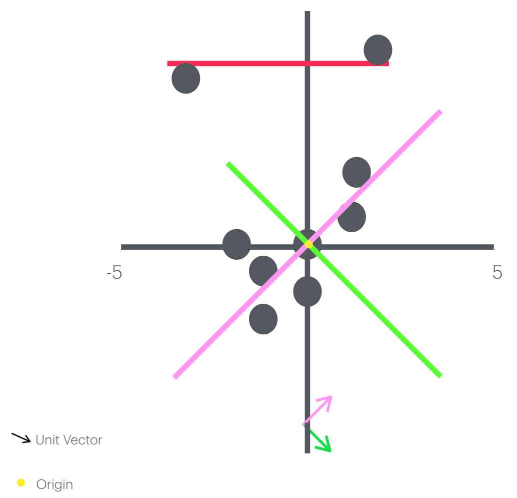
Because random samples are likely to be near borders. The distance between two randomly picked points in higher dimensions will be large

In higher dimensions, training instances are likely to be far away from another

Predictions aren't so reliable because higher dimensions have so much 'ether' between instances. Patterns are such high dimensions are almost impossible to find

In order for a model to fit a high dimensional dataset it must overfit because patterns in higher dimensions are so hard to find (think of a model fitting random noise, not able to generalize(fit) a noise pattern)

PCA



Axis with small variance

Axis with highest variance (c_1

Axis perpendicular to the highest variance (largest amount of remaining variance) (c_2

Find Principal

$$\mathbf{X} = \mathbf{U\Sigma V^T}$$

$X =$ Training Data

$SVD - SingleValueDecomposition$

$$\mathbf{V} = \begin{pmatrix} | & | & & | \\ c_1 & c_2 & \dots & c_n \\ | & | & & | \end{pmatrix}$$

$c_n =$ unit Vectors where principal components lives (**line on the top left figure** c1_1 — > $y = c_1 \cdot x_1$)

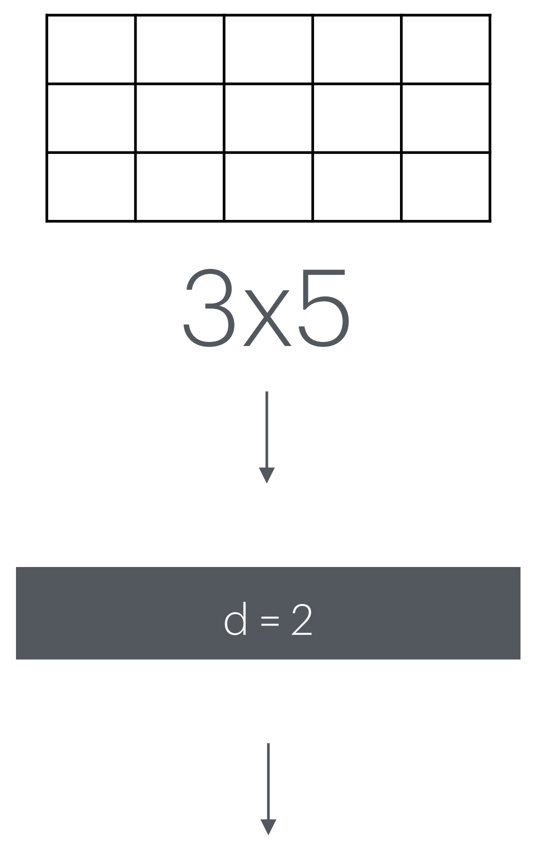
$n =$ features

Dataset must be centered around the origin when using PCA

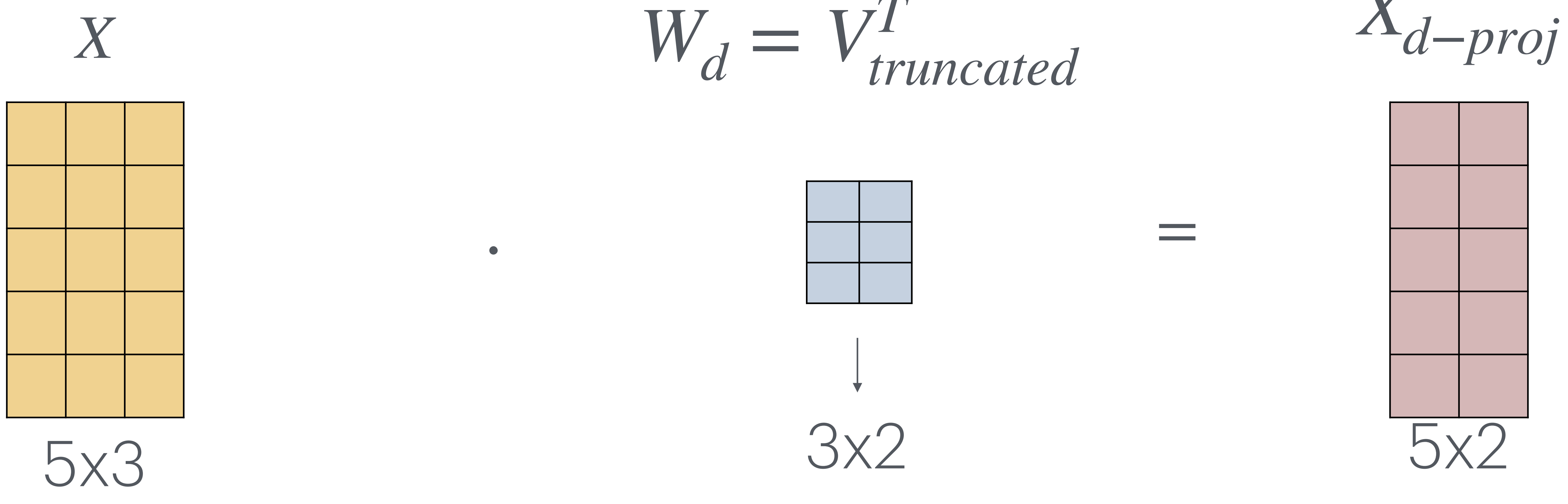
Reduced Projection

$$X_{d-proj} = XW_d$$

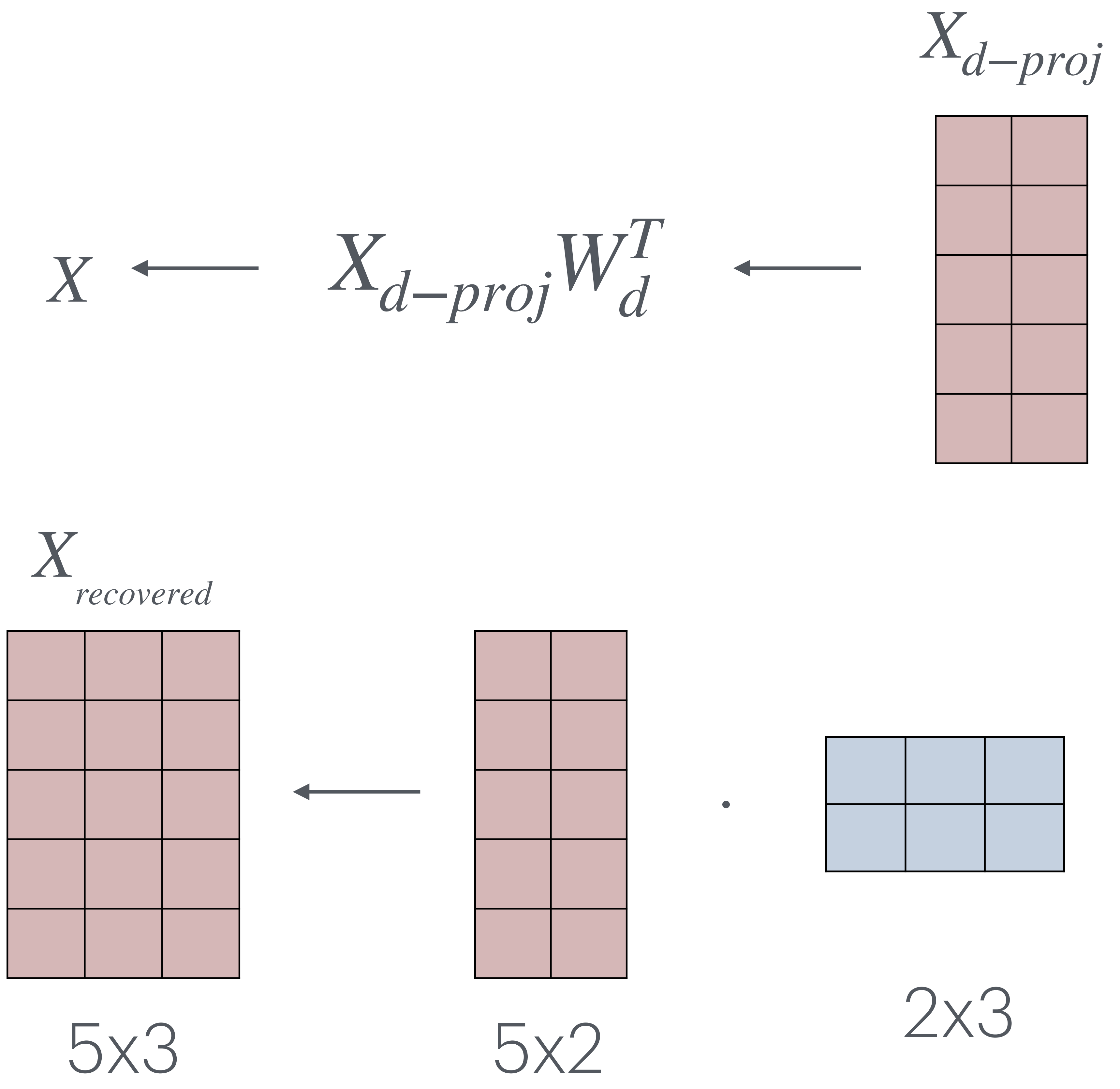
$$V^T = \begin{bmatrix} - & c_1 & - \\ - & c_2 & - \\ - & c_3 & - \\ - & \vdots & - \\ - & c_n & - \end{bmatrix}$$



$$W_d = V^T_{truncated}$$



Recover Reduced Projection



TSNE MNIST