

# Warm-up as you walk in



1. [https://www.sporcle.com/games/MrChewypoo/minimalist\\_disney](https://www.sporcle.com/games/MrChewypoo/minimalist_disney)
2. <https://www.sporcle.com/games/Stanford0008/minimalist-cartoons-slideshow>
3. <https://www.sporcle.com/games/MrChewypoo/minimalist>



# 10-601 Introduction to Machine Learning

Machine Learning Department  
School of Computer Science  
Carnegie Mellon University


## PCA

Pat Virtue

# Welcome Micah!!



# Learning Paradigms

Paradigm	Data
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot) \text{ and } y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \dots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot)$
$\hookrightarrow$ Clustering	predict $\{z^{(i)}\}_{i=1}^N$ where $z^{(i)} \in \{1, \dots, K\}$
 $\hookrightarrow$ Dimensionality Reduction	convert each <u><math>\mathbf{x}^{(i)} \in \mathbb{R}^M</math></u> to <u><math>\mathbf{u}^{(i)} \in \mathbb{R}^K</math></u> with <u><math>K \ll M</math></u>
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \dots\}$
Active Learning	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and can query $y^{(i)} = c^*(\cdot)$ at a cost
Imitation Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), \dots\}$
Reinforcement Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}, r^{(1)}), (s^{(2)}, a^{(2)}, r^{(2)}), \dots\}$

# PCA Outline

- **Dimensionality Reduction**
  - High-dimensional data
  - Learning (low dimensional) representations
- **Principal Component Analysis (PCA)**
  - Examples: 2D and 3D
  - Data for PCA
  - PCA Definition
  - Objective functions for PCA
  - PCA, Eigenvectors, and Eigenvalues
  - Algorithms for finding Eigenvectors / Eigenvalues
- **PCA Examples**
  - Image Compression
  - MRI Image Reconstruction

# **DIMENSIONALITY REDUCTION**

# High Dimension Data

Examples of high dimensional data:

- High resolution images (millions of pixels)





## Examples of high dimensional data:

- Multilingual News Stories  
(vocabulary of hundreds of thousands of words)





# High Dimension Data

Examples of high dimensional data:

- Brain Imaging Data (100s of MBs per scan)

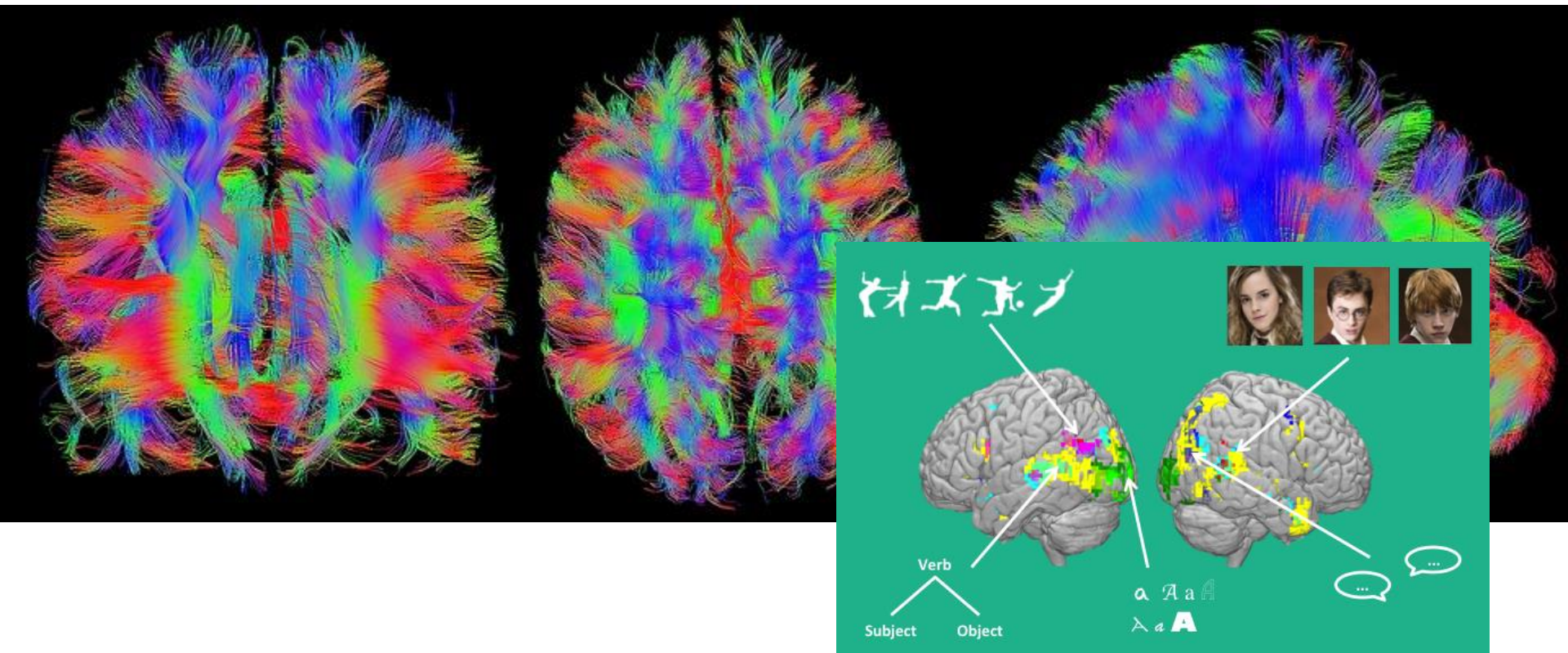


Image from (Wehbe et al., 2014)

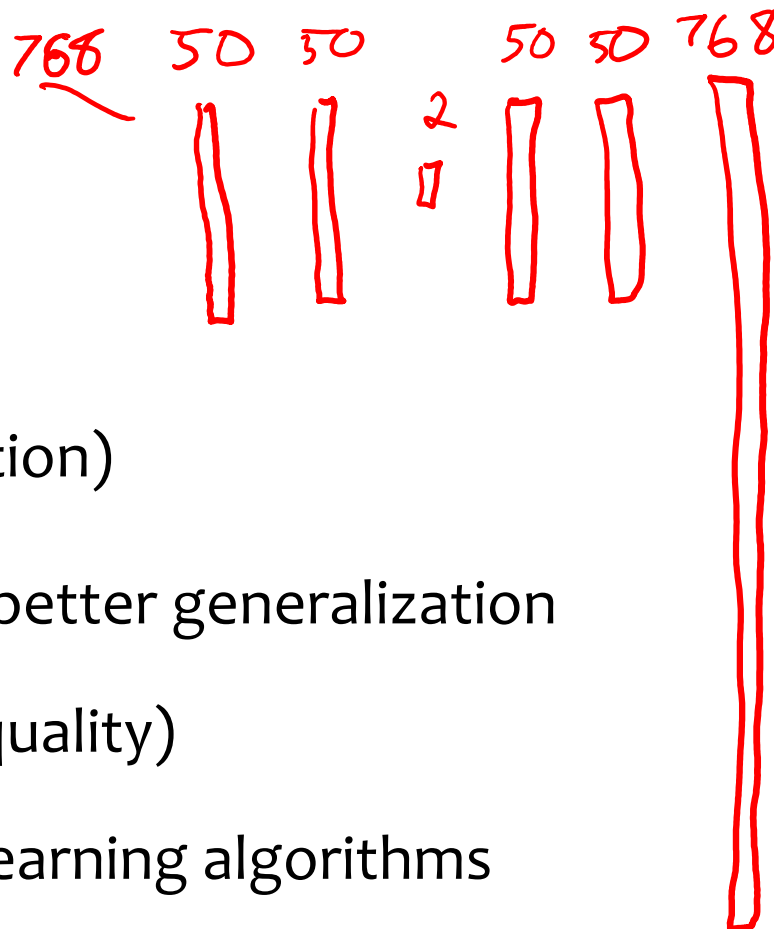
Image from <https://pixabay.com/en/brain-mrt-magnetic-resonance-imaging-1728449/>

# Learning Representations

**PCA, Kernel PCA, ICA:** Powerful unsupervised learning techniques for extracting hidden (potentially lower dimensional) structure from high dimensional datasets.

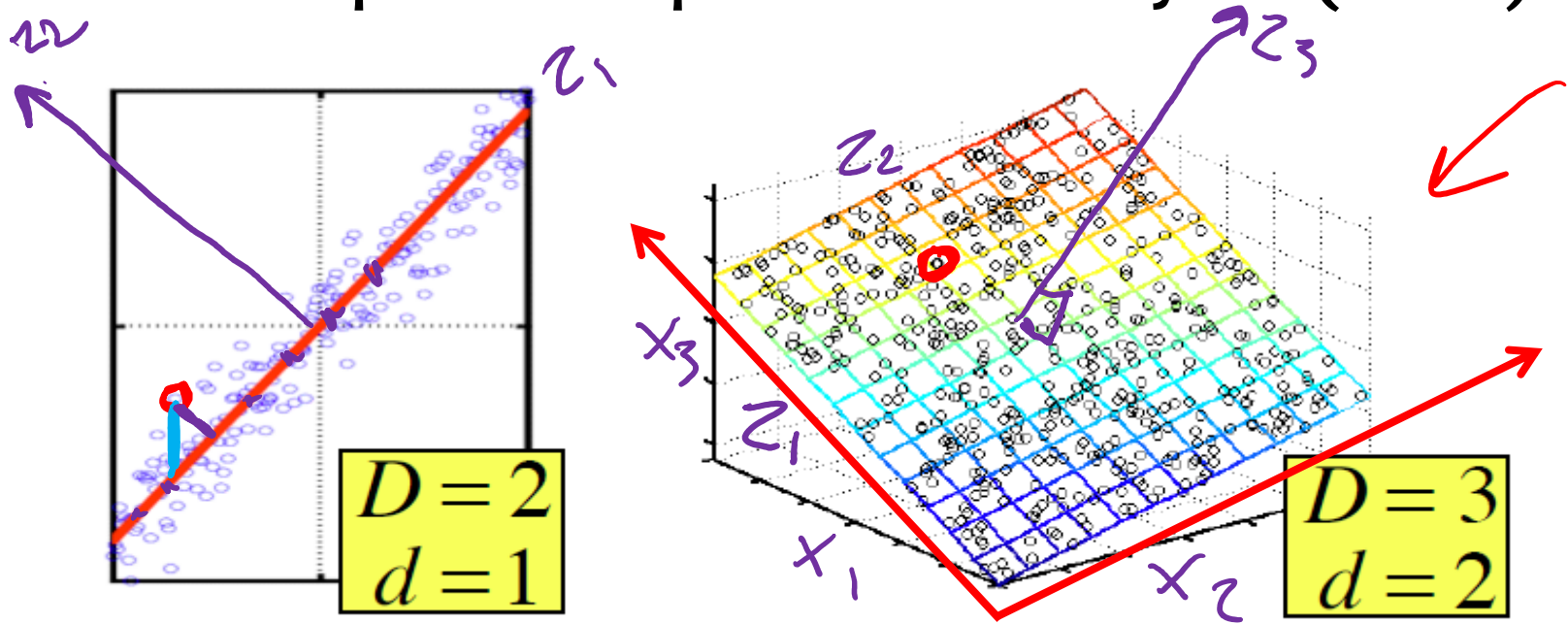
## Useful for:

- Visualization
- More efficient use of resources (e.g., time, memory, communication)
- Statistical: fewer dimensions  $\rightarrow$  better generalization
- Noise removal (improving data quality)
- Further processing by machine learning algorithms



# **PRINCIPAL COMPONENT ANALYSIS (PCA)**

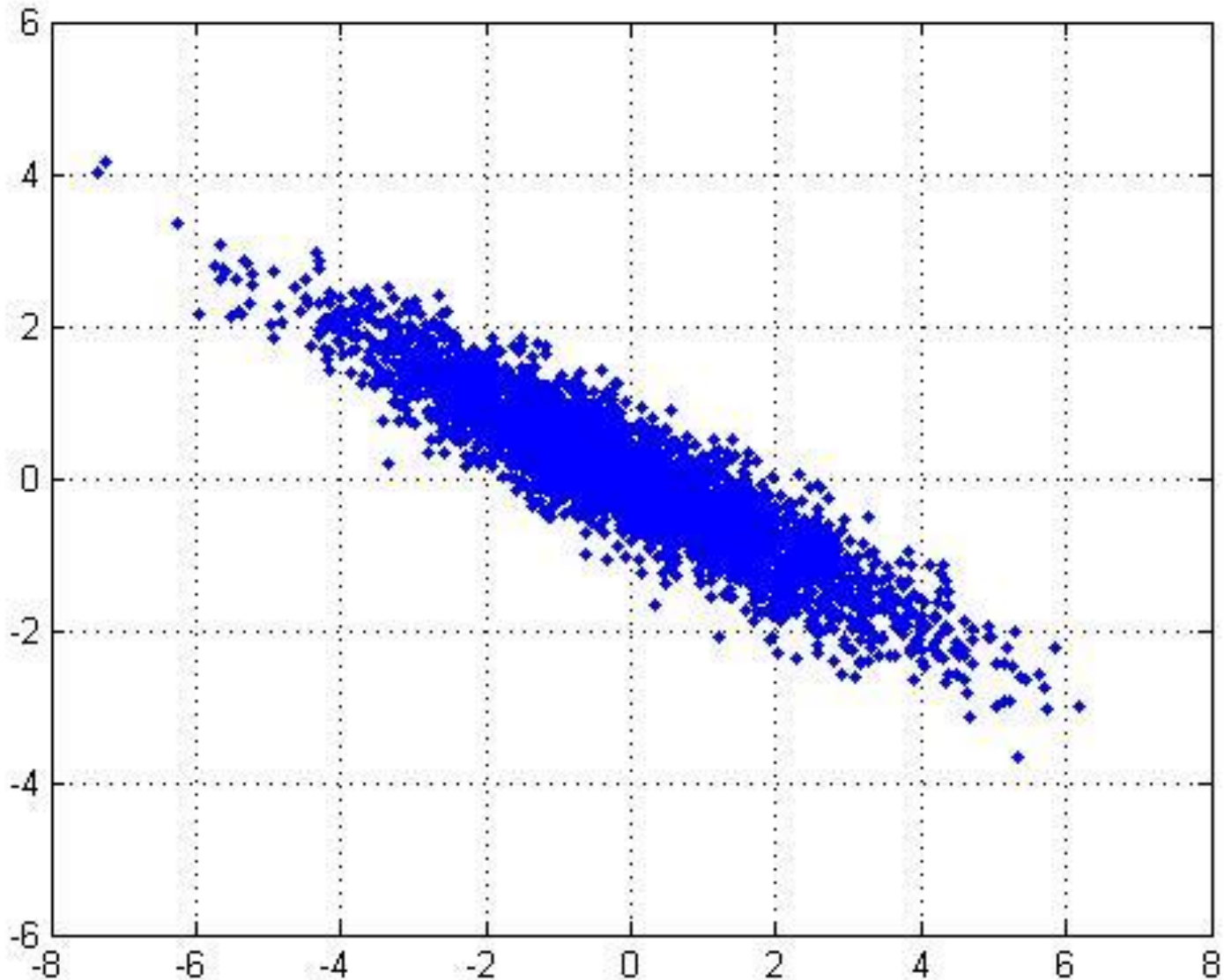
# Principal Component Analysis (PCA)



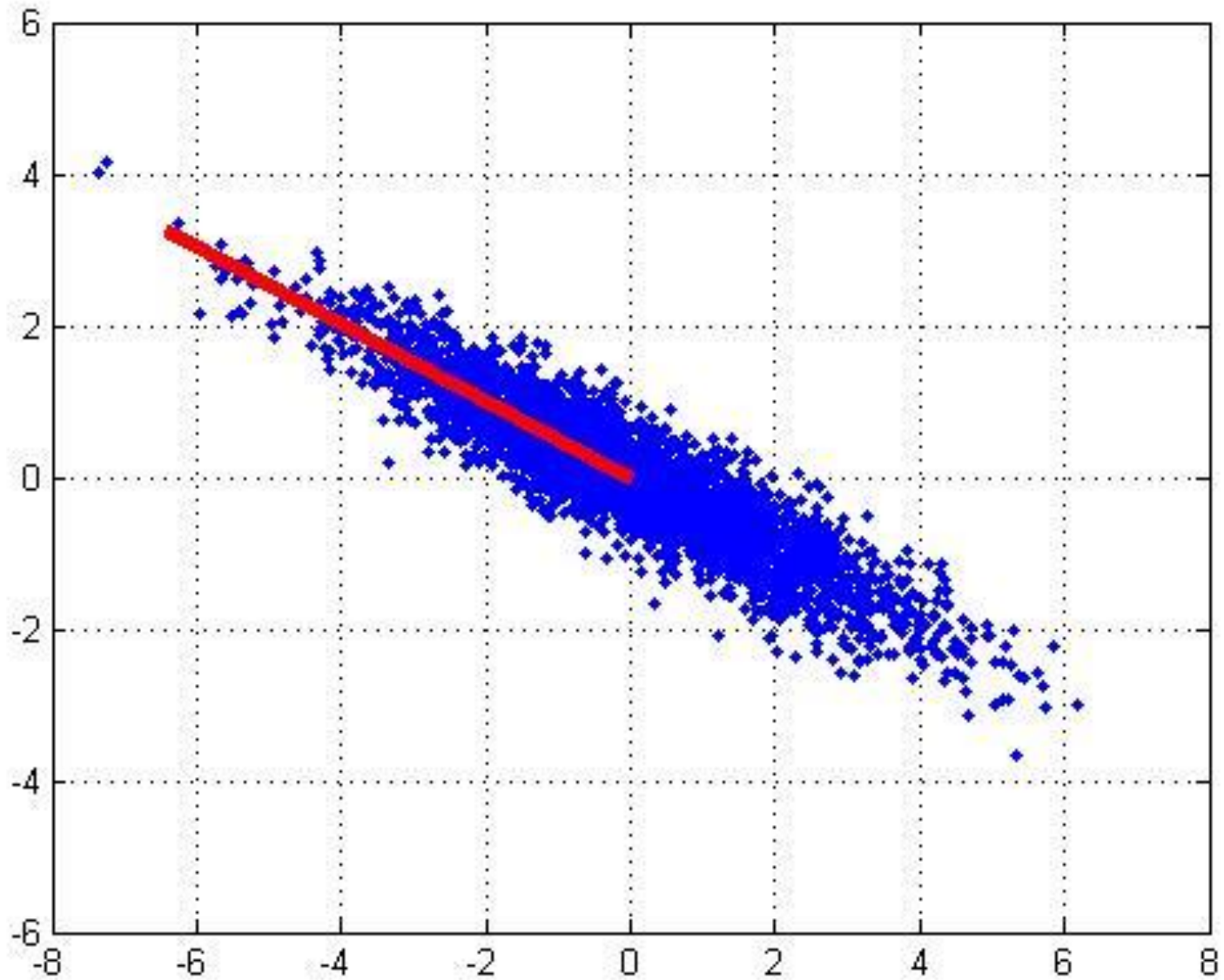
In case where data lies on or near a low  $d$ -dimensional linear subspace, axes of this subspace are an effective representation of the data.

Identifying the axes is known as [Principal Components Analysis](#), and can be obtained by using classic matrix computation tools (Eigen or Singular Value Decomposition).

# 2D Gaussian dataset

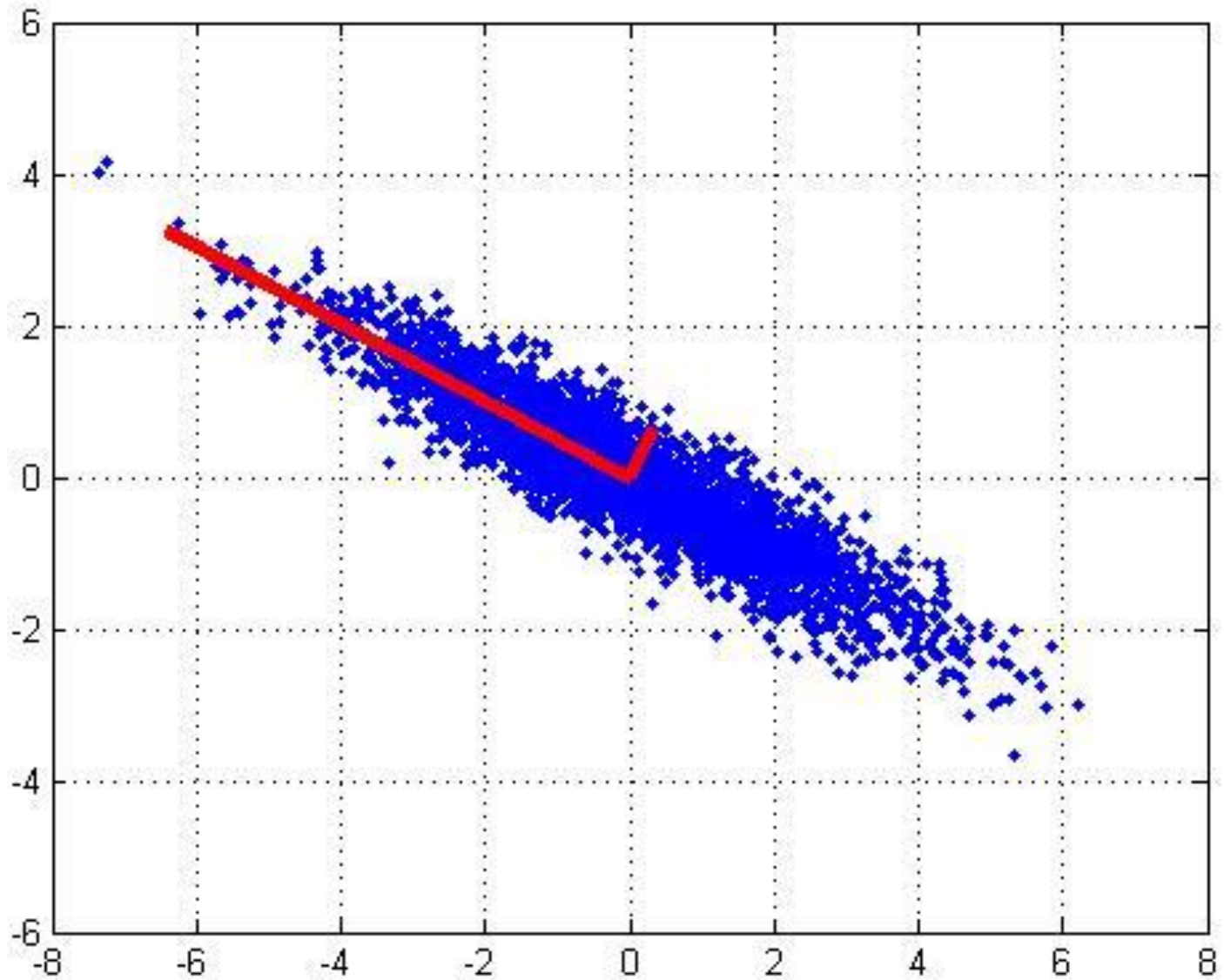


# 1<sup>st</sup> PCA axis





# 2<sup>nd</sup> PCA axis





# Growth Plate Imaging

## Growth Plate Disruption and Limb Length Discrepancy



8 year-old boy with previous fracture and  
4cm leg length discrepancy

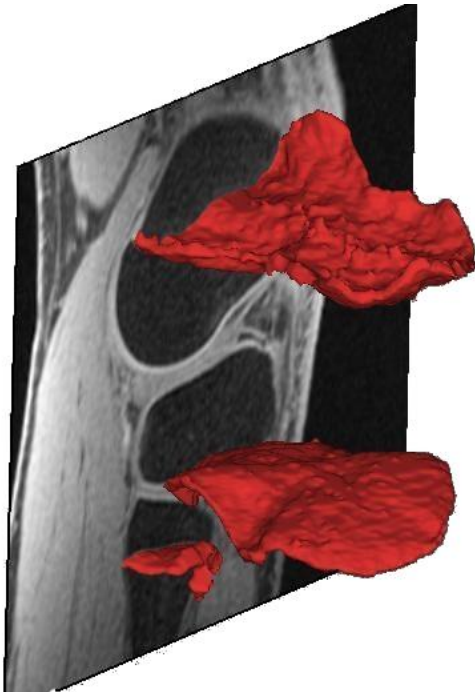


Images Courtesy  
H. Potter, H.S.S.

# Growth Plate Imaging

## Growth Plate Disruption and Limb Length Discrepancy

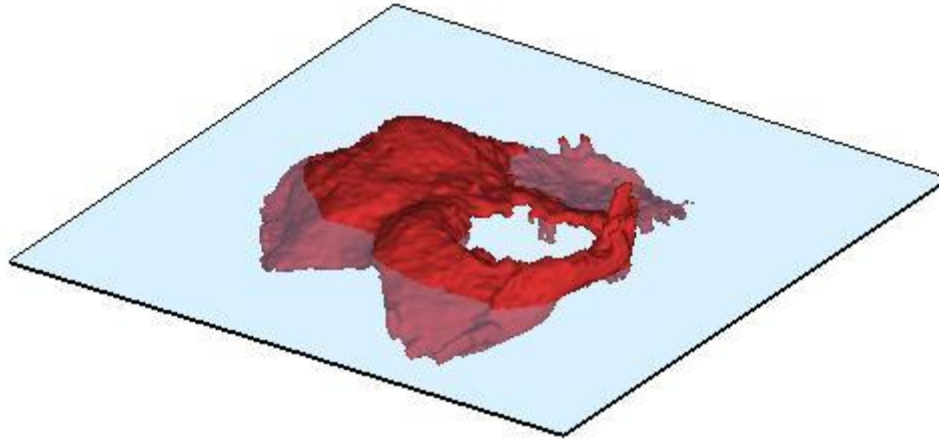
8 year-old boy with previous fracture and  
4cm leg length discrepancy



Images Courtesy  
H. Potter, H.S.S.

# Growth Plate Imaging

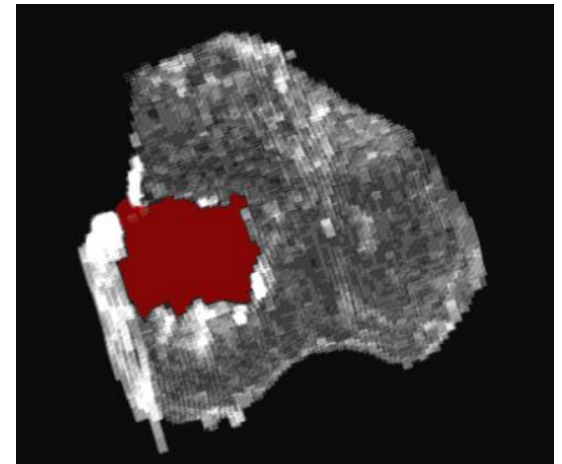
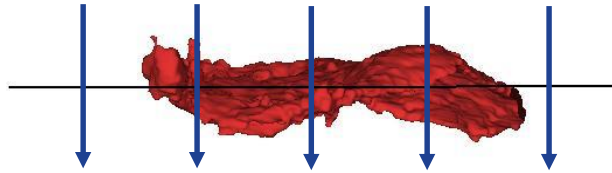
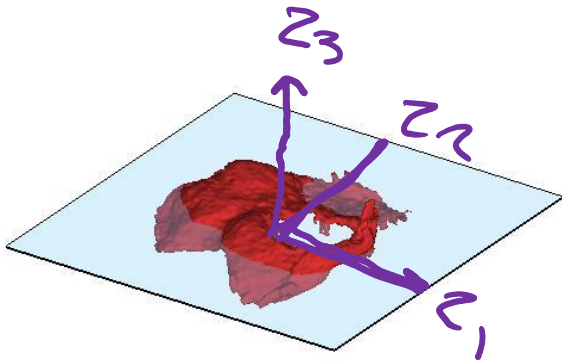
## Area Measurement



# Growth Plate Imaging

## Area Measurement

$$M = 3$$
$$K = 2$$



Flatten Growth Plate to Enable 2D Area Measurement

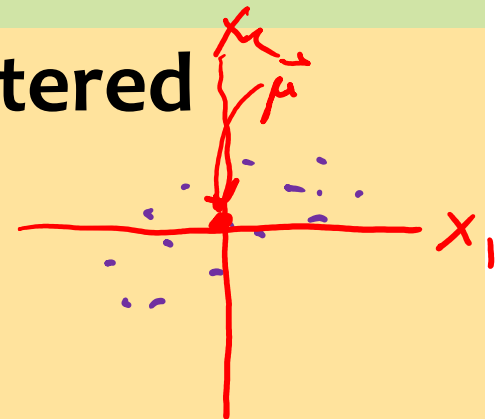
# Data for PCA

$$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

We assume the data is **centered**

$$\vec{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} = \mathbf{0}$$



**Q:** What if  
your data is  
**not** centered?

**A:** Subtract  
off the  
sample mean

# Sample Covariance Matrix

The sample covariance matrix is given by:

$$\Sigma_{jk} = \frac{1}{N} \sum_{i=1}^N \underbrace{(x_j^{(i)} - \mu_j)}_{\uparrow} \underbrace{(x_k^{(i)} - \mu_k)}_{\uparrow}$$

Since the data matrix is centered, we rewrite as:

$$\Sigma = \frac{1}{N} \mathbf{X}^T \mathbf{X}$$

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(N)})^T \end{bmatrix}$$

# Projections

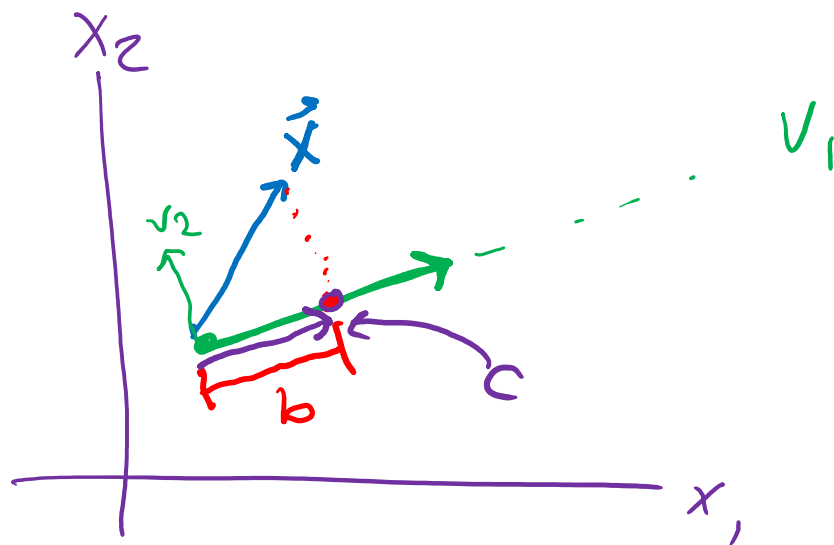
Quiz: What is the projection of point  $x$  onto vector  $\underline{v}$ , assuming that  $\underline{\|v\|_2 = 1}$ ?

~~A.  $vx$~~

B.  $v^T x = b$

C.  $(\underline{v^T x})v = c$

D.  $v^T x x^T v$





# Principal Component Analysis (PCA)

## *Whiteboard*

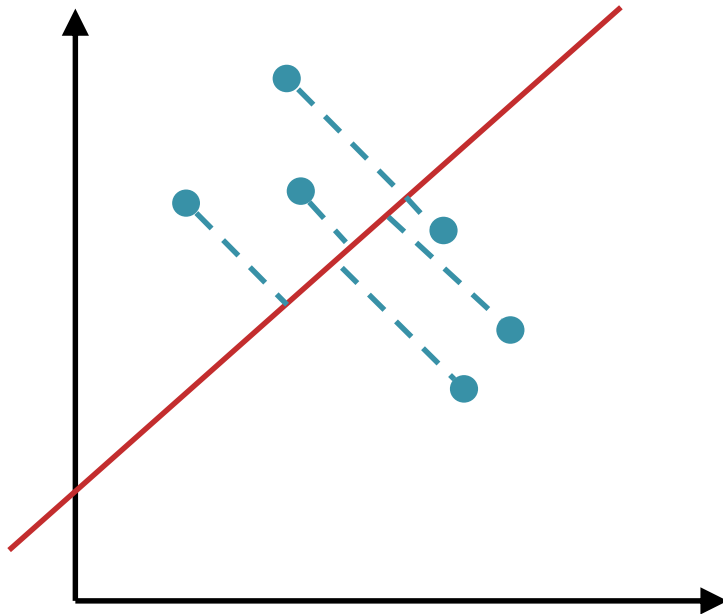
- PCA Sketch
- Objective functions for PCA

# Maximizing the Variance

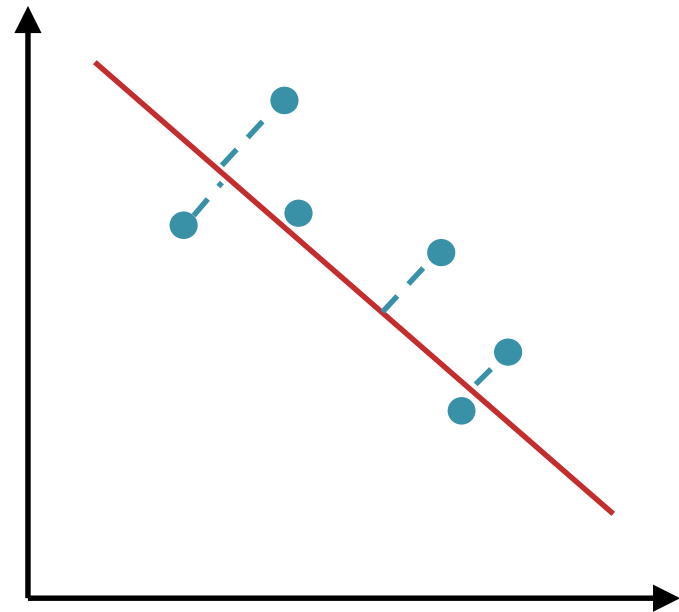
**Quiz:** Consider the two projections below

2. Which maximizes the variance?
3. Which minimizes the reconstruction error?

Option A



Option B



2)

3)

# Principal Component Analysis (PCA)

## *Whiteboard*

- PCA, Eigenvectors, and Eigenvalues
- Algorithms for finding Eigenvectors / Eigenvalues

# PCA

## Equivalence of Maximizing Variance and Minimizing Reconstruction Error

**Claim:** Minimizing the reconstruction error is equivalent to maximizing the variance.

**Proof:** First, note that:

$$\|\mathbf{x}^{(i)} - (\mathbf{v}^T \mathbf{x}^{(i)})\mathbf{v}\|^2 = \|\mathbf{x}^{(i)}\|^2 - (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (1)$$

since  $\mathbf{v}^T \mathbf{v} = \|\mathbf{v}\|^2 = 1$ .

Substituting into the minimization problem, and removing the extraneous terms, we obtain the maximization problem.

$$\mathbf{v}^* = \operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}^{(i)} - (\mathbf{v}^T \mathbf{x}^{(i)})\mathbf{v}\|^2 \quad (2)$$

$$= \operatorname{argmin}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}^{(i)}\|^2 - (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (3)$$

$$= \operatorname{argmax}_{\mathbf{v}: \|\mathbf{v}\|^2=1} \frac{1}{N} \sum_{i=1}^N (\mathbf{v}^T \mathbf{x}^{(i)})^2 \quad (4)$$

$$(5)$$

# PCA: the First Principal Component

To find the first principal component, we wish to solve the following constrained optimization problem (variance minimization).

$$\mathbf{v}_1 = \underset{\mathbf{v}: ||\mathbf{v}||^2=1}{\operatorname{argmax}} \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} \quad (1)$$

So we turn to the method of Lagrange multipliers. The Lagrangian is:

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1) \quad (2)$$

Taking the derivative of the Lagrangian and setting to zero gives:

$$\frac{d}{d\mathbf{v}} (\mathbf{v}^T \boldsymbol{\Sigma} \mathbf{v} - \lambda(\mathbf{v}^T \mathbf{v} - 1)) = 0 \quad (3)$$

$$\boldsymbol{\Sigma} \mathbf{v} - \lambda \mathbf{v} = 0 \quad (4)$$

$$\boldsymbol{\Sigma} \mathbf{v} = \lambda \mathbf{v} \quad (5)$$

Recall: For a square matrix  $\mathbf{A}$ , the vector  $\mathbf{v}$  is an **eigenvector** iff there exists **eigenvalue**  $\lambda$  such that:

$$\mathbf{A} \mathbf{v} = \lambda \mathbf{v} \quad (6)$$

# SVD for PCA

For any arbitrary matrix  $\mathbf{A}$ , SVD gives a decomposition:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \quad (1)$$

where  $\mathbf{\Lambda}$  is a diagonal matrix, and  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal matrices.

Suppose we obtain an SVD of our data matrix  $\mathbf{X}$ , so that:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \quad (1)$$

Now consider what happens when we rewrite  $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$  terms of this SVD.

$$\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^T\mathbf{X} \quad (2)$$

$$= \frac{1}{N}(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)^T(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T) \quad (3)$$

$$= \frac{1}{N}(\mathbf{V}\mathbf{\Lambda}^T\mathbf{U}^T)(\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T) \quad (4)$$

$$= \frac{1}{N}\mathbf{V}\mathbf{\Lambda}^T\mathbf{\Lambda}\mathbf{V}^T \quad (5)$$

$$= \frac{1}{N}\mathbf{V}(\mathbf{\Lambda})^2\mathbf{V}^T \quad (6)$$

Above we used the fact that  $\mathbf{U}^T\mathbf{U} = \mathbf{I}$  since  $\mathbf{U}$  is orthogonal by definition.

We find that  $(\mathbf{\Lambda})^2$  is a diagonal matrix whose entries are  $\Lambda_{ii} = \lambda_i^2$  the squares of the eigenvalues of the SVD of  $\mathbf{X}$ . Further, both  $\mathbf{X}$  and  $\mathbf{X}^T\mathbf{X}$  share the same eigenvectors in their SVD.

Thus, we can run SVD on  $\mathbf{X}$  without ever instantiating the large  $\mathbf{X}^T\mathbf{X}$  to obtain the necessary principal components more efficiently.

# Principal Component Analysis (PCA)

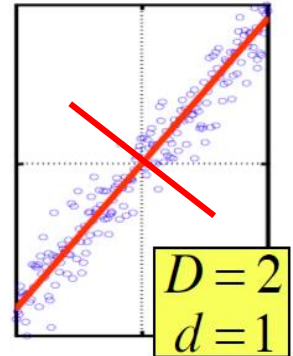


# Principal Component Analysis (PCA)

$(X^T X) \mathbf{v} = \lambda \mathbf{v}$ , so  $\mathbf{v}$  (the first PC) is the eigenvector of sample correlation/covariance matrix  $X^T X$

Sample variance of projection  $\mathbf{v}^T X^T X \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Thus, the eigenvalue  $\lambda$  denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

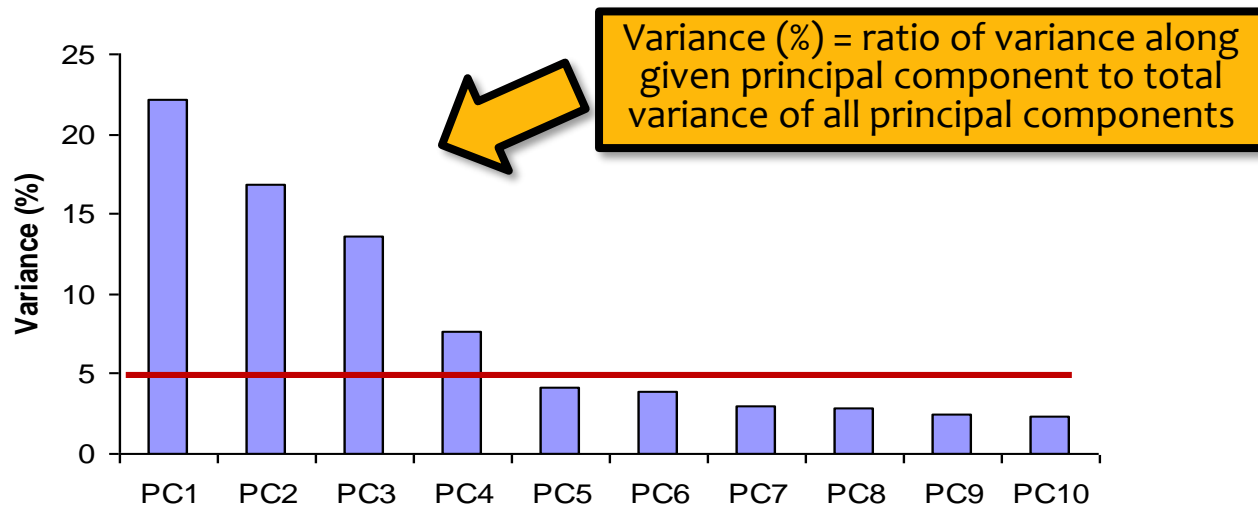


Eigenvalues  $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$

- The 1<sup>st</sup> PC  $\mathbf{v}_1$  is the the eigenvector of the sample covariance matrix  $X^T X$  associated with the largest eigenvalue
- The 2nd PC  $\mathbf{v}_2$  is the the eigenvector of the sample covariance matrix  $X^T X$  associated with the second largest eigenvalue
- And so on ...

# How Many PCs?

- For  $M$  original dimensions, sample covariance matrix is  $M \times M$ , and has up to  $M$  eigenvectors. So  $M$  PCs.
- Where does dimensionality reduction come from?  
Can ignore the components of lesser significance.



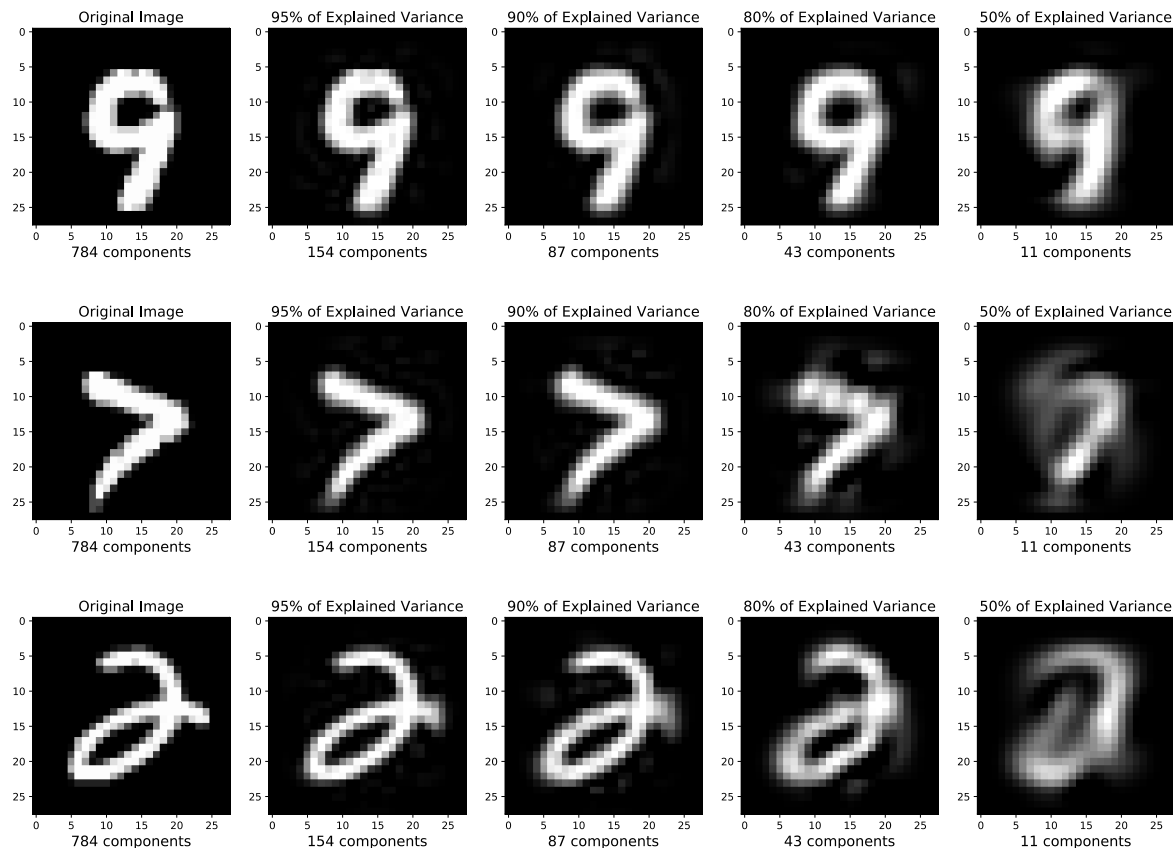
- You do lose some information, but if the eigenvalues are small, you don't lose much
  - $M$  dimensions in original data
  - calculate  $M$  eigenvectors and eigenvalues
  - choose only the first  $D$  eigenvectors, based on their eigenvalues
  - final data set has only  $D$  dimensions

# PCA EXAMPLES

# Projecting MNIST digits

## Task Setting:

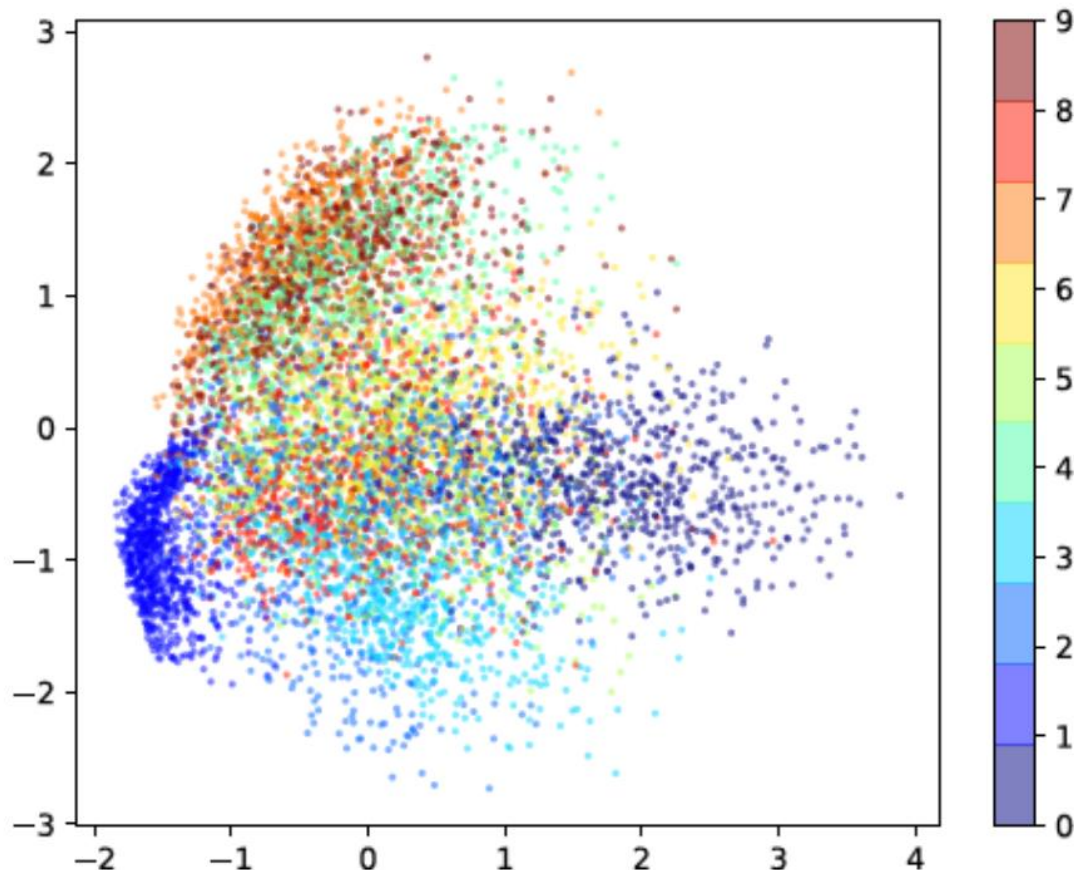
1. Take 28x28 images of digits and project them down to K components
2. Report percent of variance explained for K components
3. Then project back up to 28x28 image to visualize how much information was preserved



# Projecting MNIST digits

## Task Setting:

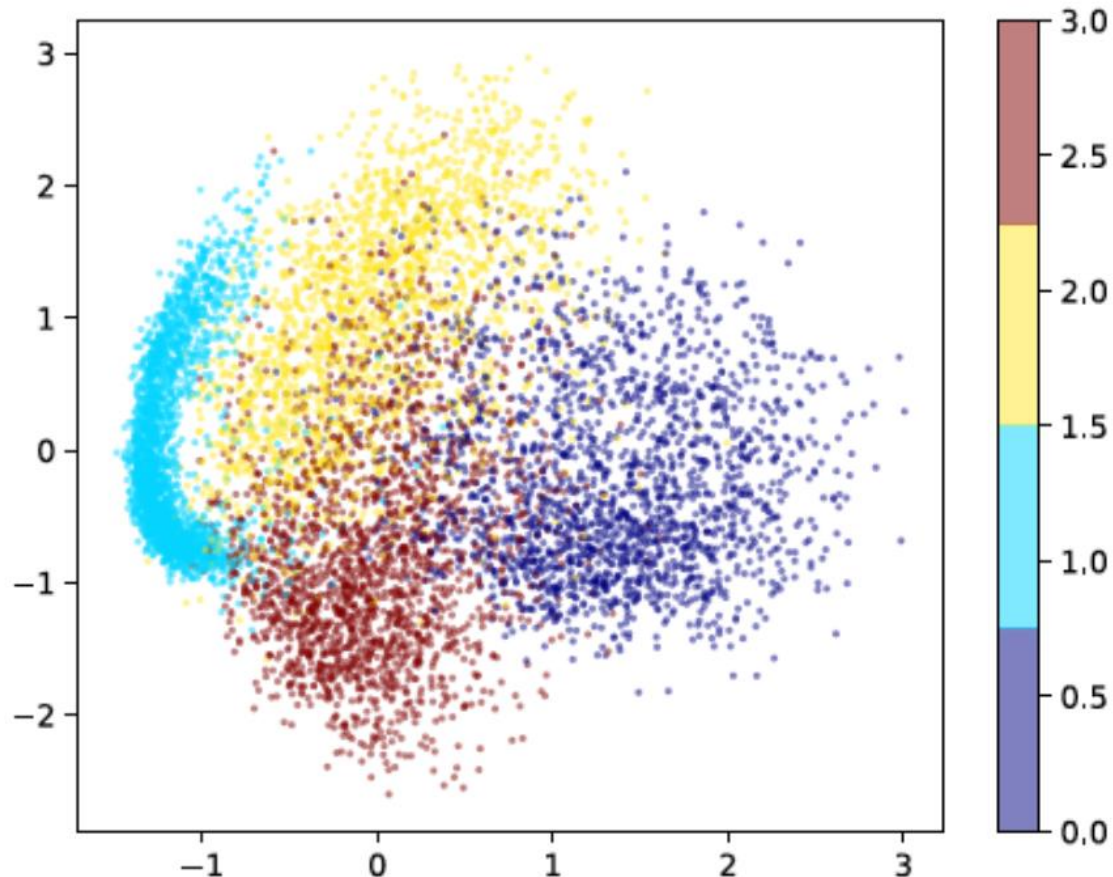
1. Take 28x28 images of digits and project them down to 2 components
2. Plot the 2 dimensional points



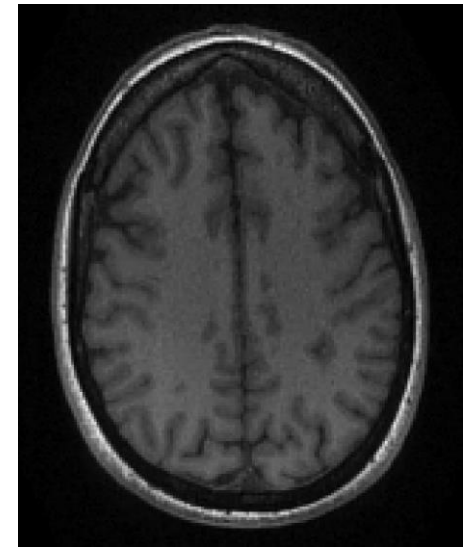
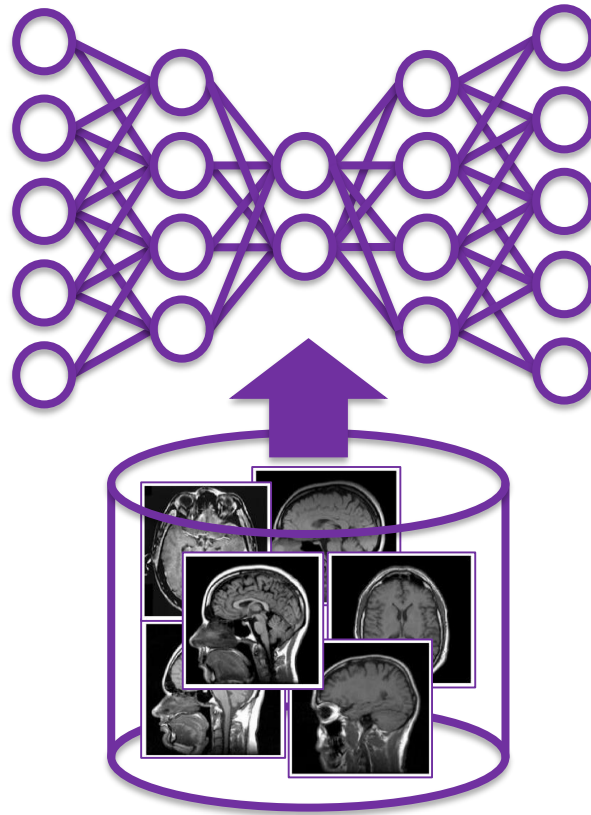
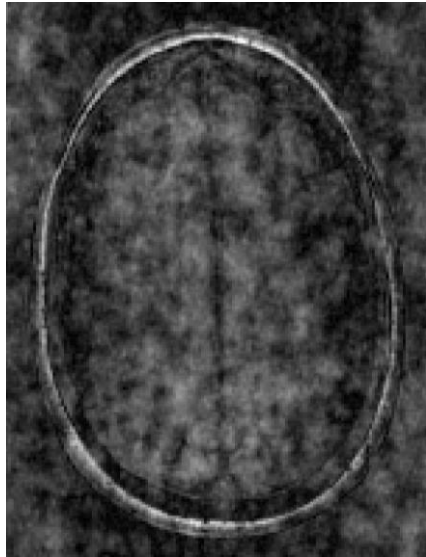
# Projecting MNIST digits

## Task Setting:

1. Take 28x28 images of digits and project them down to 2 components
2. Plot the 2 dimensional points



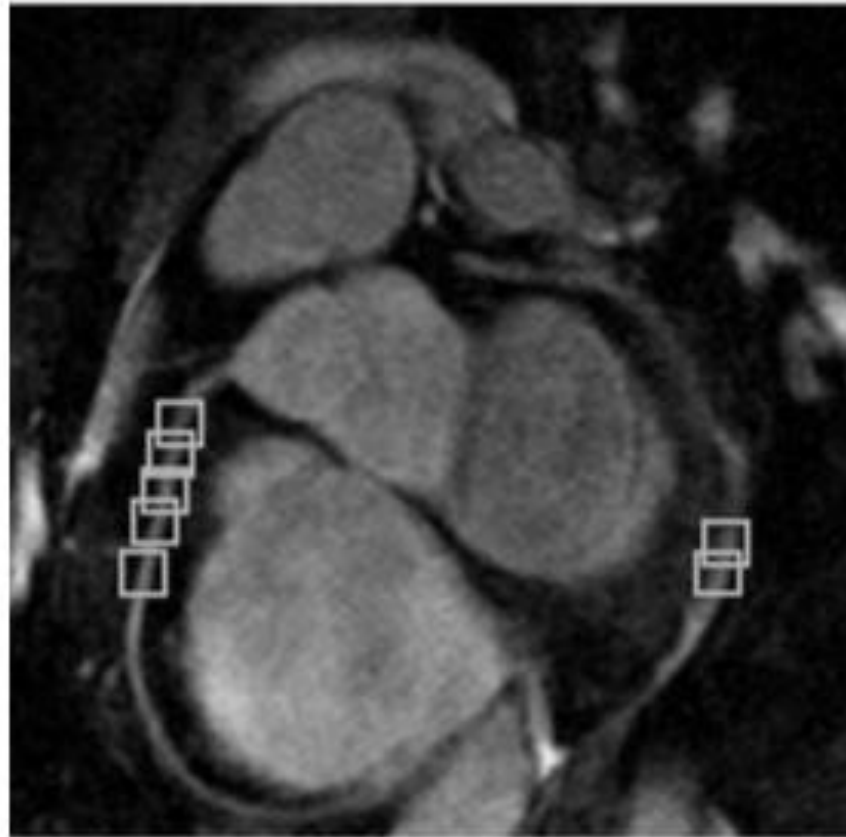
# MRI Image Reconstruction





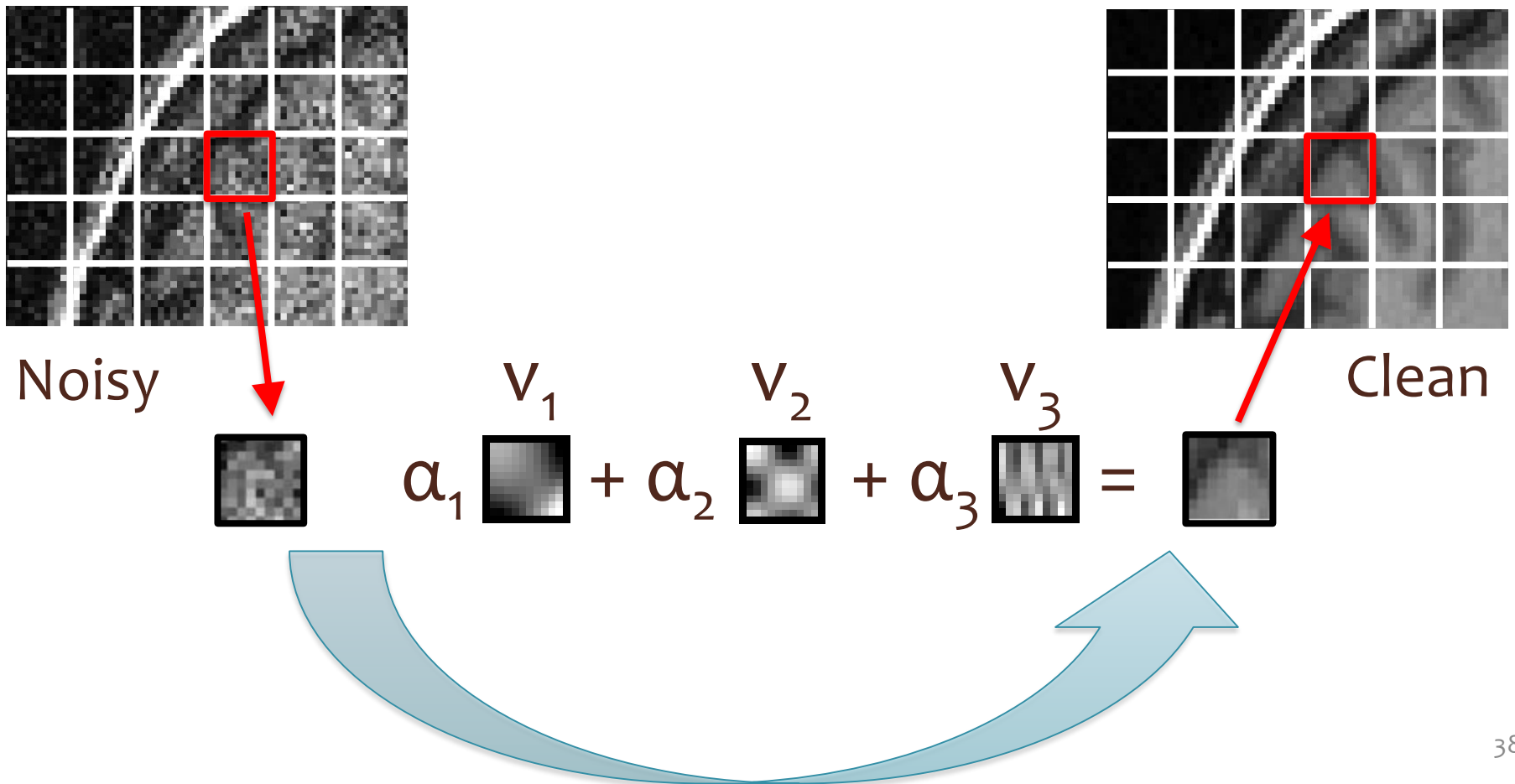
# MRI Image Reconstruction

Lots of redundant structure at patch level



# MRI Image Reconstruction

- Image Denoising



# MRI Image Reconstruction

Quiz: If I have a 10x10 patch from a noisy image, which number of principal components kept will allow more noise in the resulting patch?

A. 100

B. 10

# Learning Objectives

## Dimensionality Reduction / PCA

*You should be able to...*

1. Define the sample mean, sample variance, and sample covariance of a vector-valued dataset
2. Identify examples of high dimensional data and common use cases for dimensionality reduction
3. Draw the principal components of a given toy dataset
4. Establish the equivalence of minimization of reconstruction error with maximization of variance
5. Given a set of principal components, project from high to low dimensional space and do the reverse to produce a reconstruction
6. Explain the connection between PCA, eigenvectors, eigenvalues, and covariance matrix
7. Use common methods in linear algebra to obtain the principal components