Machine Learning for Data Science (CS4786)
Lecture 4

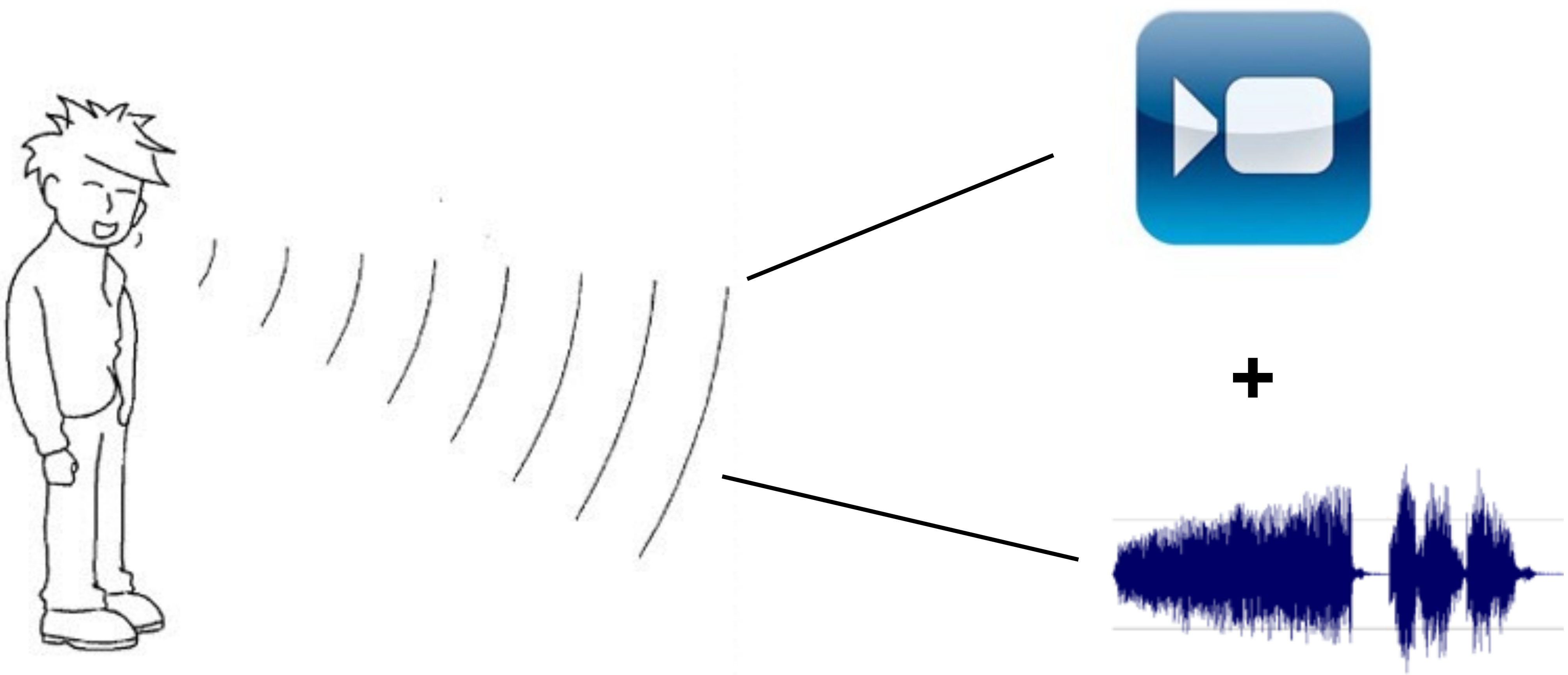
Canonical Correlation Analysis (CCA)

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2015sp/

When to use CCA?

- When we have redundancy in data.
- When the relevant information is part of the redundancy
- Same data point from two different view/sources

Example I: Speech Recognition



- Audio might have background sounds uncorrelated with video
- Video might have lighting changes uncorrelated with audio
- Redundant information between two views: the speech

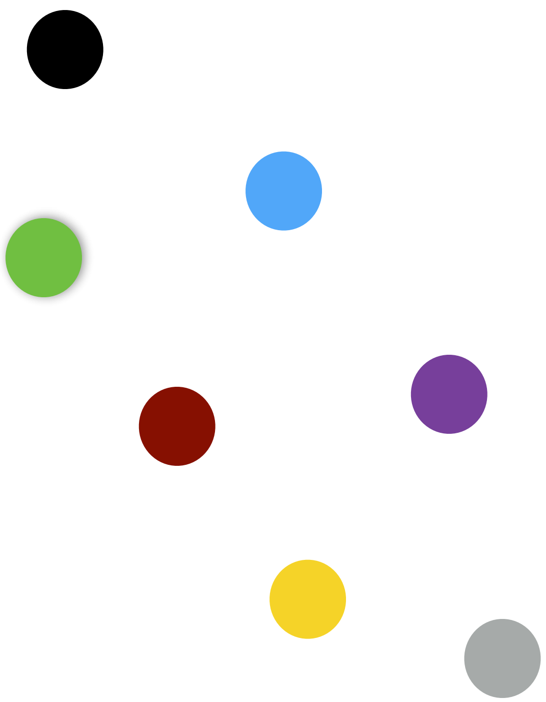
Example II: Combining Feature Extractions

- Method A and Method B are both equally good feature extraction techniques
- Concatenating the two features blindly yields large dimensional feature vector with redundancy
- Applying techniques like CCA extracts the key information between the two methods
- Removes extra unwanted information

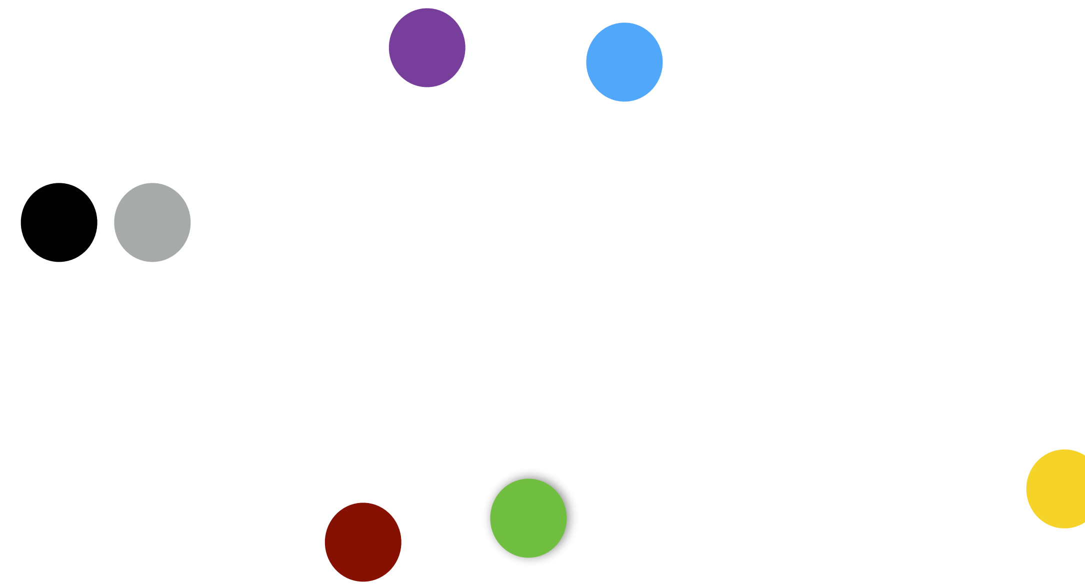
Two View Dimensionality Reduction

- Data comes in pairs $(\mathbf{x}_1, \mathbf{x}'_1), \ldots, (\mathbf{x}_n, \mathbf{x}'_n)$ where $\mathbf{x}_t$'s are d dimensional and $\mathbf{x}'_t$'s are d' dimensional
- Goal: Compress say view one into $\mathbf{y}_1, \ldots, \mathbf{y}_n$, that are K dimensional vectors
  - Retain information redundant between the two views
  - Eliminate "noise" specific to only one of the views

Which Direction to Pick?

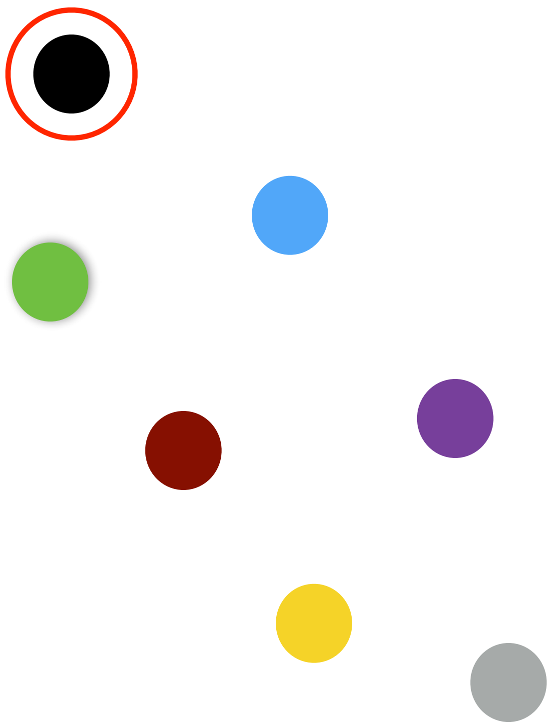


View I

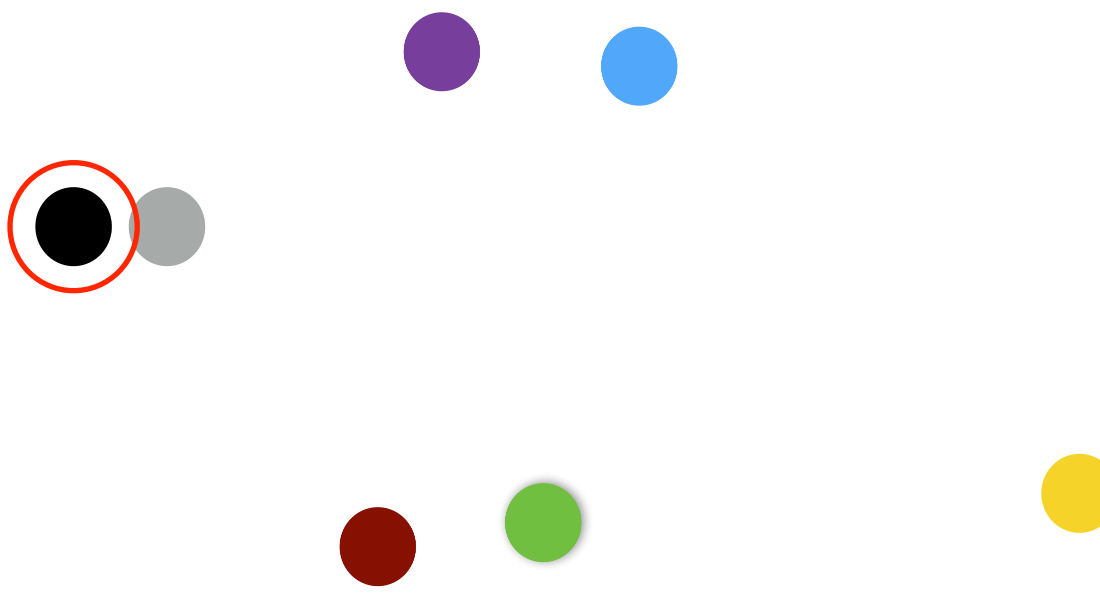


View II

Which Direction to Pick?

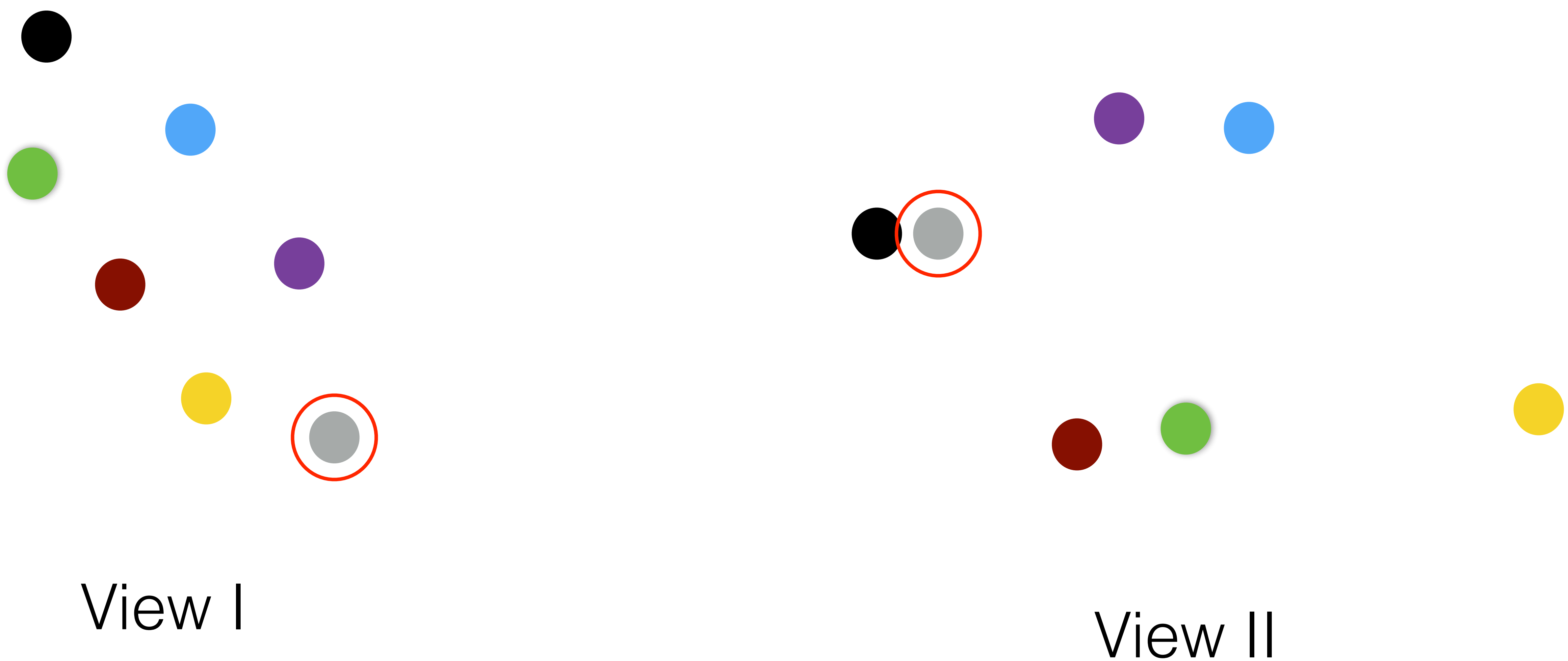


View I

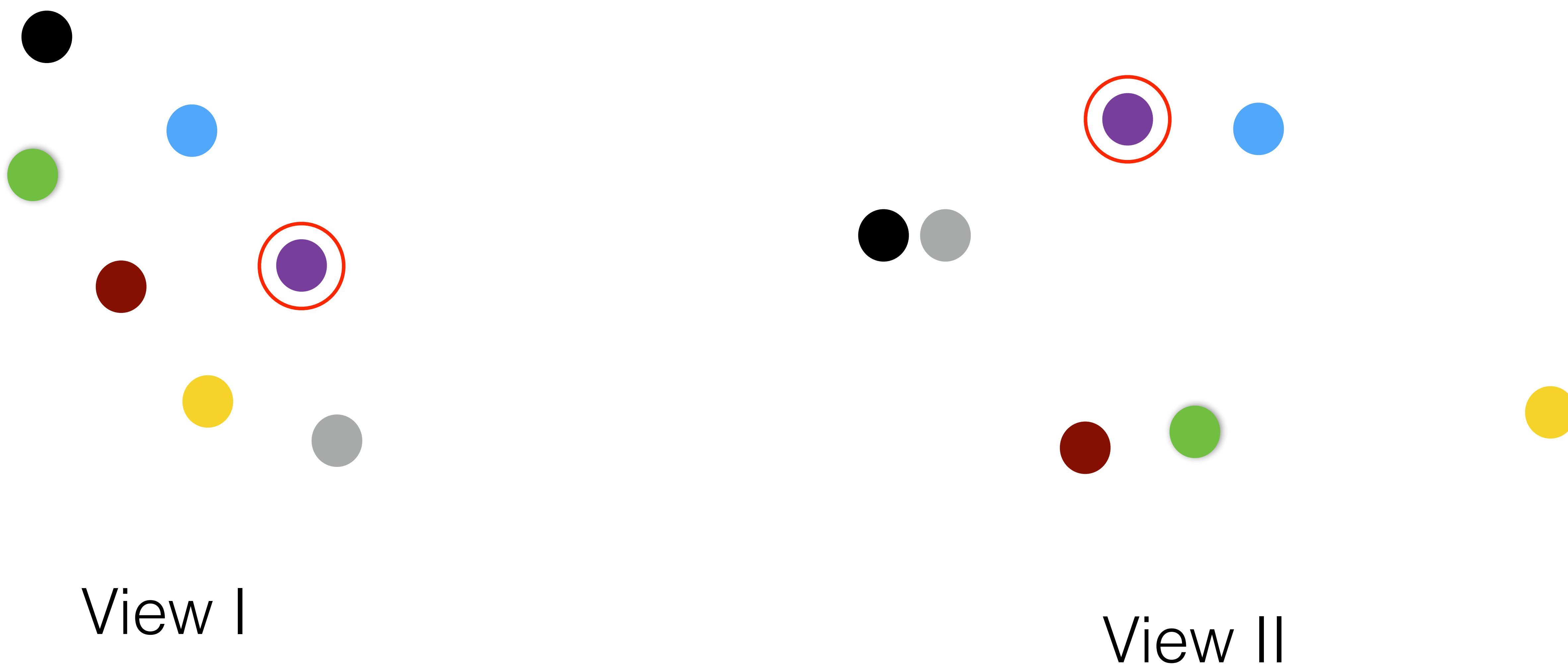


View II

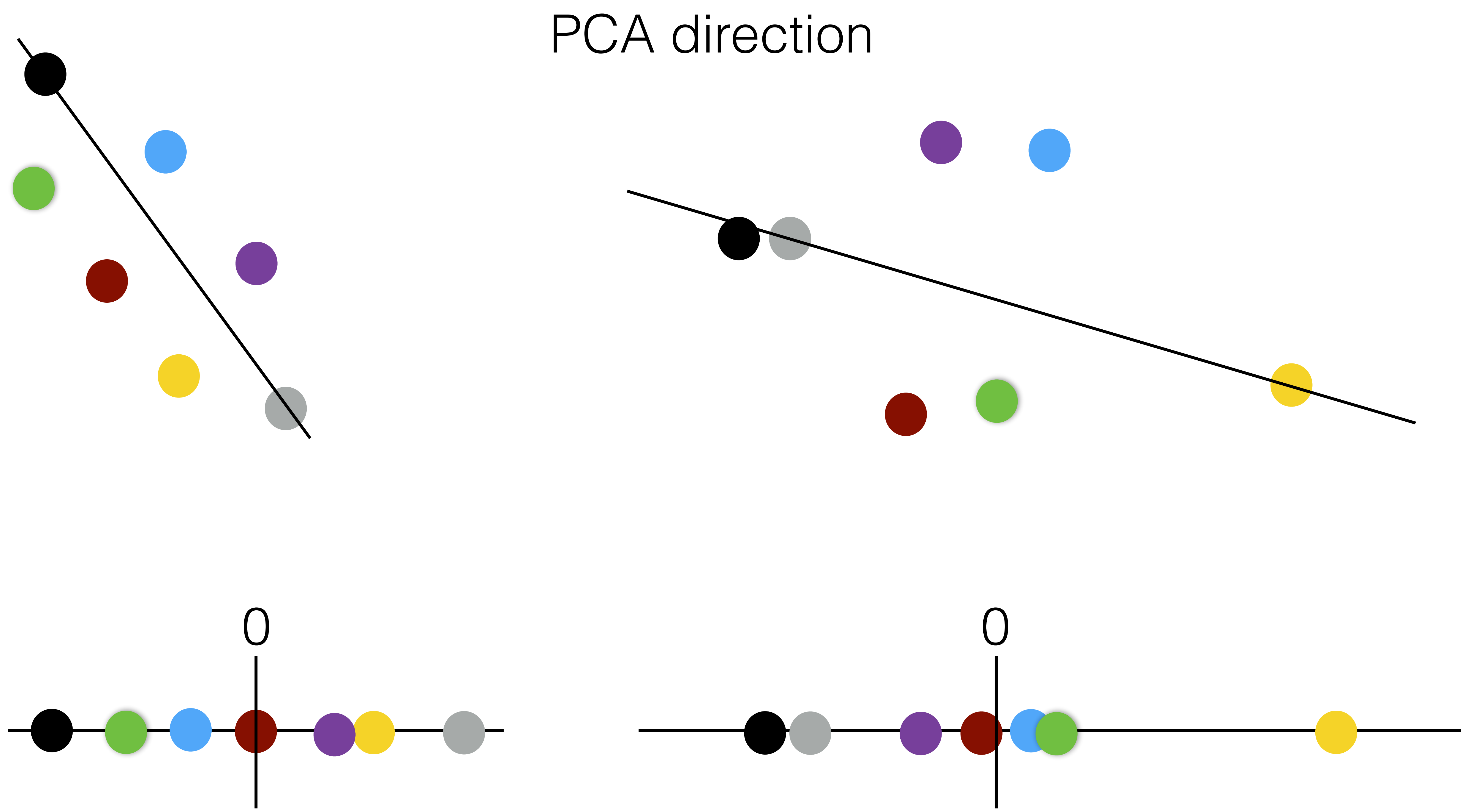
Which Direction to Pick?



Which Direction to Pick?

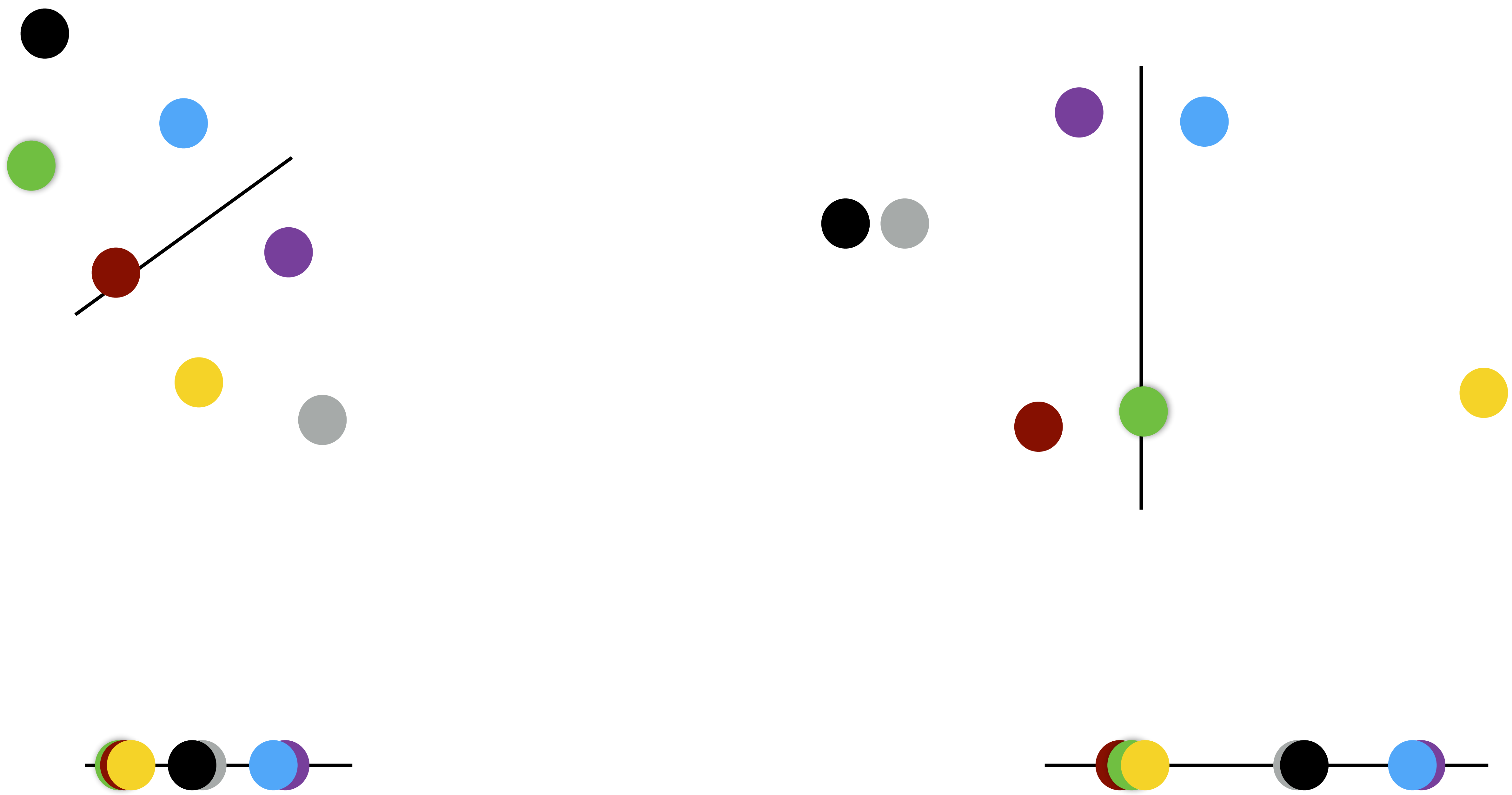


WHICH DIRECTION TO PICK?



$$\text{Average dot product} = \text{covariance small}$$

Which Direction to Pick?



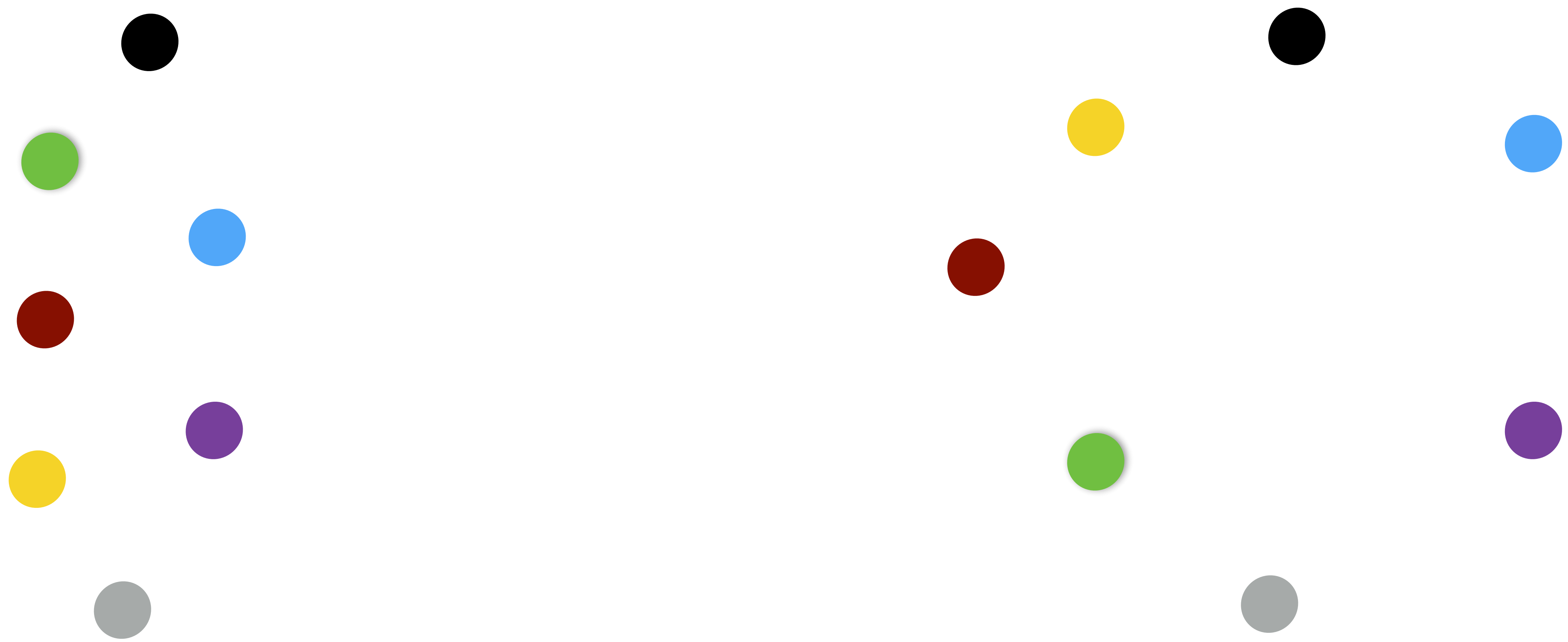
Direction has large covariance

Why not Maximize Covariance

$$\text{Say } \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t[2] \cdot \mathbf{x}'_t[2] > 0$$

Scaling up this coordinate we can blow up covariance

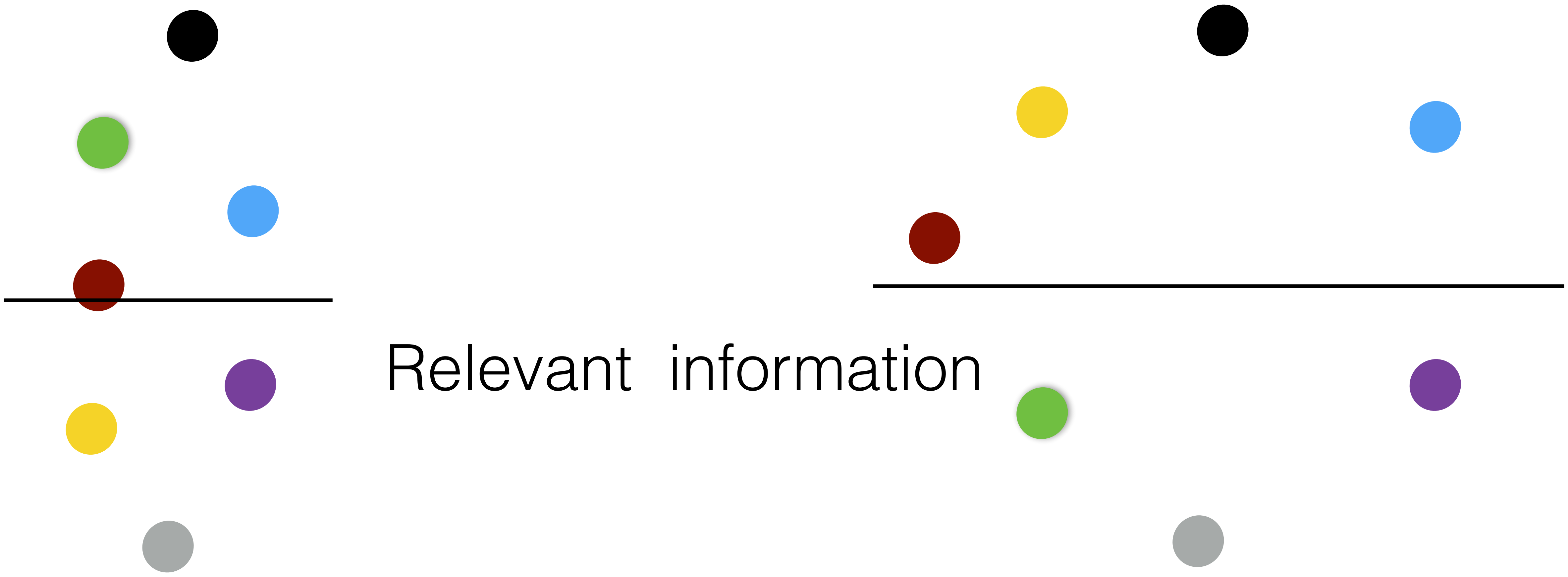
Why not Maximize Covariance



$$\text{Say } \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t[2] \cdot \mathbf{x}'_t[2] > 0$$

Scaling up this coordinate we can blow up covariance

Why not Maximize Covariance



$$\text{Say } \frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t[2] \cdot \mathbf{x}'_t[2] > 0$$

Scaling up this coordinate we can blow up covariance

Basic Idea of CCA

- Normalize variance in chosen direction to be constant (say 1)
- Then maximize covariance
- This is same as maximizing "correlation coefficient" (recall from last class).

Maximizing Correlation Coefficient

- Say $\mathbf{w}_1$ and $\mathbf{v}_1$ are the directions we choose to project in views 1 and 2 respectively we want these directions to maximize,

$$\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)\cdot\left(\mathbf{y}'_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right)$$

s.t. $\frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t[1]\right)^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{y}'_t[1] - \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}'_t[1]\right) = 1$

where $\mathbf{y}_t[1] = \mathbf{w}_1^\top\mathbf{x}_t$ and $\mathbf{y}'_t[1] = \mathbf{v}_1^\top\mathbf{x}'_t$

Canonical Correlation Analysis

- Assume data in both views are centered : $\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t = \mathbf{0}$, $\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}'_t = \mathbf{0}$
  Hence $\frac{1}{n}\sum_{t=1}^{n} \mathbf{y}'_t[1] = \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_t[1] = 0$

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{y}_t[1] \cdot \mathbf{y}'_t[1]$$

$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} (\mathbf{y}_t[1])^2 = \frac{1}{n}\sum_{t=1}^{n} (\mathbf{y}'_t[1])^2 = 1$$

Canonical Correlation Analysis

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top \mathbf{x}_t \cdot \mathbf{v}_1^\top \mathbf{x}_t'$$

$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} (\mathbf{w}_1^\top \mathbf{x}_t)^2 = \frac{1}{n}\sum_{t=1}^{n} (\mathbf{v}_1^\top \mathbf{x}_t')^2 = 1$$

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top \mathbf{x}_t \mathbf{x}_t'^\top \mathbf{v}_1$$

$$\text{subject to } \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_1^\top \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_1 = \frac{1}{n}\sum_{t=1}^{n} \mathbf{v}_1^\top \mathbf{x}_t' \mathbf{x}_t'^\top \mathbf{v}_1 = 1$$

Canonical Correlation Analysis

- Hence we want to solve for projection vectors $\mathbf{w}_1$ and $\mathbf{v}_1$ that

$$\text{maximize } \mathbf{w}_1^\top \Sigma_{1,2} \mathbf{v}_1$$
$$\text{subject to } \mathbf{w}_1^\top \Sigma_{1,1} \mathbf{w}_1 = \mathbf{v}_1^\top \Sigma_{2,2} \mathbf{v}_1 = 1$$

- Writing Lagrangian taking derivative equating to 0 we get

$$\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\mathbf{w}_1 = \lambda^2 \Sigma_{1,1}\mathbf{w}_1 \quad \text{and} \quad \Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\mathbf{v}_1 = \lambda^2 \Sigma_{2,2}\mathbf{v}_1$$

or equivalently

$$\left(\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}\right)\mathbf{w}_1 = \lambda^2 \mathbf{w}_1 \quad \text{and} \quad \left(\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}\right)\mathbf{v}_1 = \lambda^2 \mathbf{v}_1$$

CCA Algorithm

- Write $\tilde{\mathbf{x}}_t = \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}'_t \end{bmatrix}$ the $d + d'$ dimensional concatenated vectors.
- Calculate covariance matrix of the joint data points

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}$$

- Calculate $\Sigma_{1,1}^{-1}\Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$. The top K eigen vectors of this matrix give us projection matrix for view I.

- Calculate $\Sigma_{2,2}^{-1}\Sigma_{2,1}\Sigma_{1,1}^{-1}\Sigma_{1,2}$. The top K eigen vectors of this matrix give us projection matrix for view II.

CCA Demo

The Tall, the Fat and the Ugly

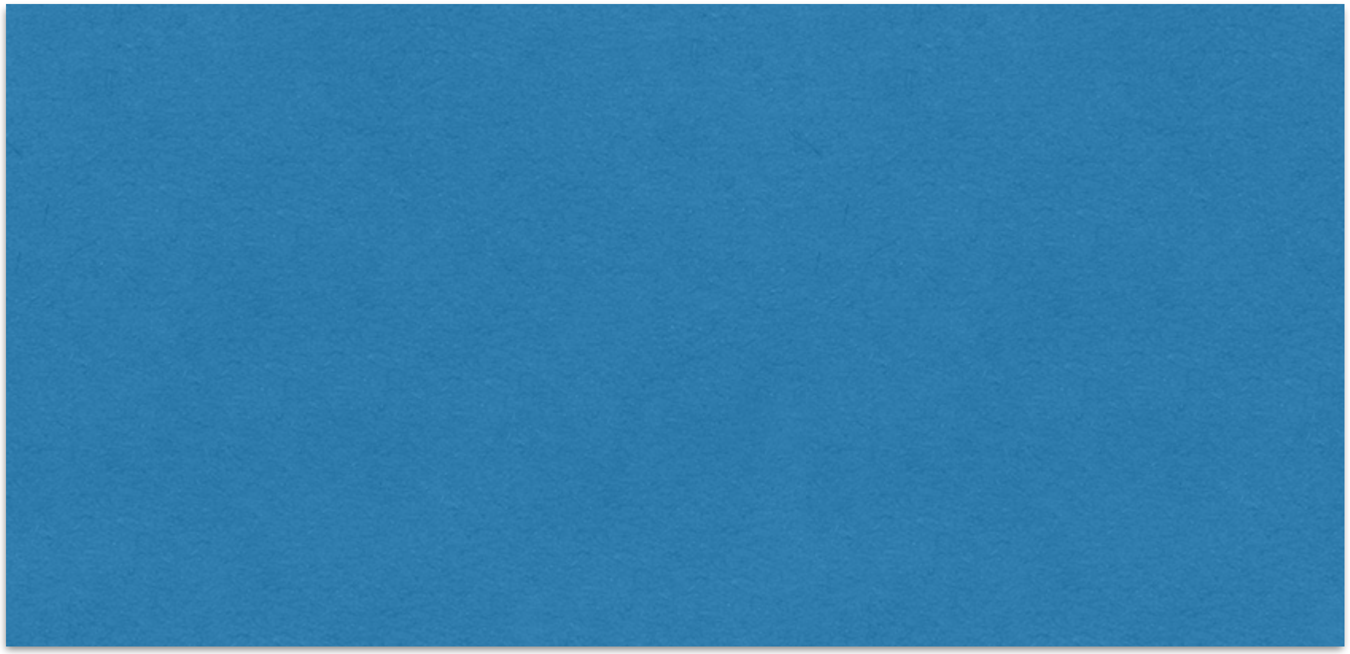
The Tall, THE FAT AND THE UGLY

$X =$ 

- If d small, calculate covariance matrix
  - PCA of the single view
  - CCA for concatenated view
- Do eigen decomposition of $d \times d$ matrix, computationally easy

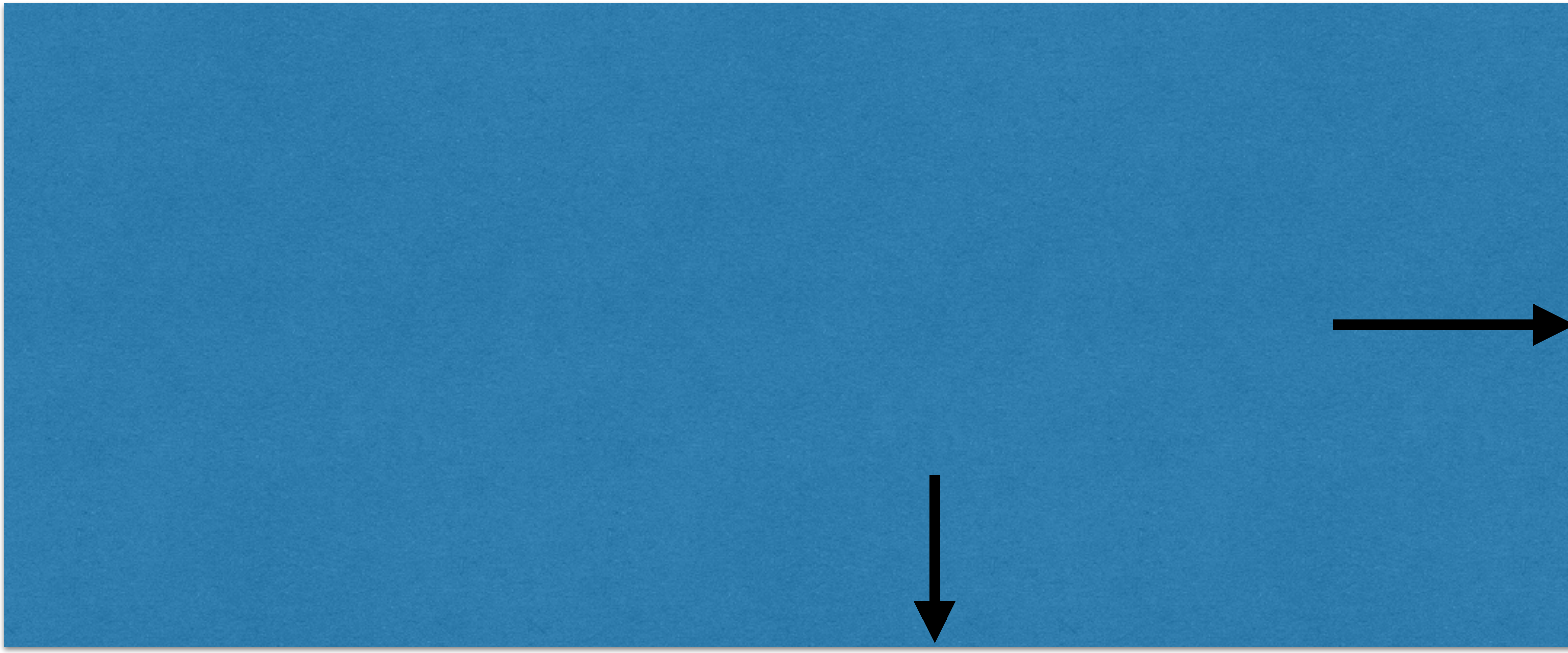
The Tall, **the Fat** and the Ugly

$X =$



- If d large by $d \times n$ manageable, directly do Singular Value Decomposition (SVD) of data matrix

The Tall, the Fat and **the Ugly**

$X =$ 

- d and n so large we can't even store in memory
- Only have time to be linear in n

I there any hope?