EE16B
Designing Information Devices and Systems II

Lecture 9B
Finish PCA, SVD

Intro

–Last Time:

–Show procedure via AA^T

– PCA

–Today:

– Continue PCA

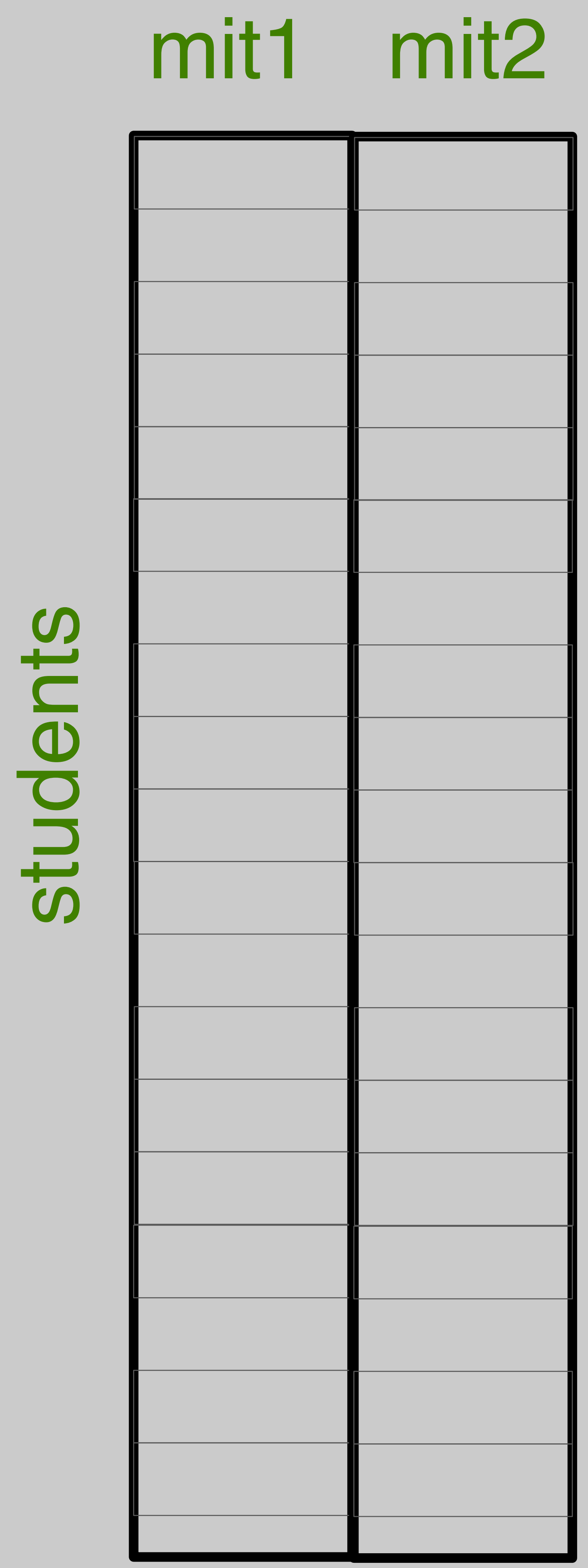
–Examples of PCA

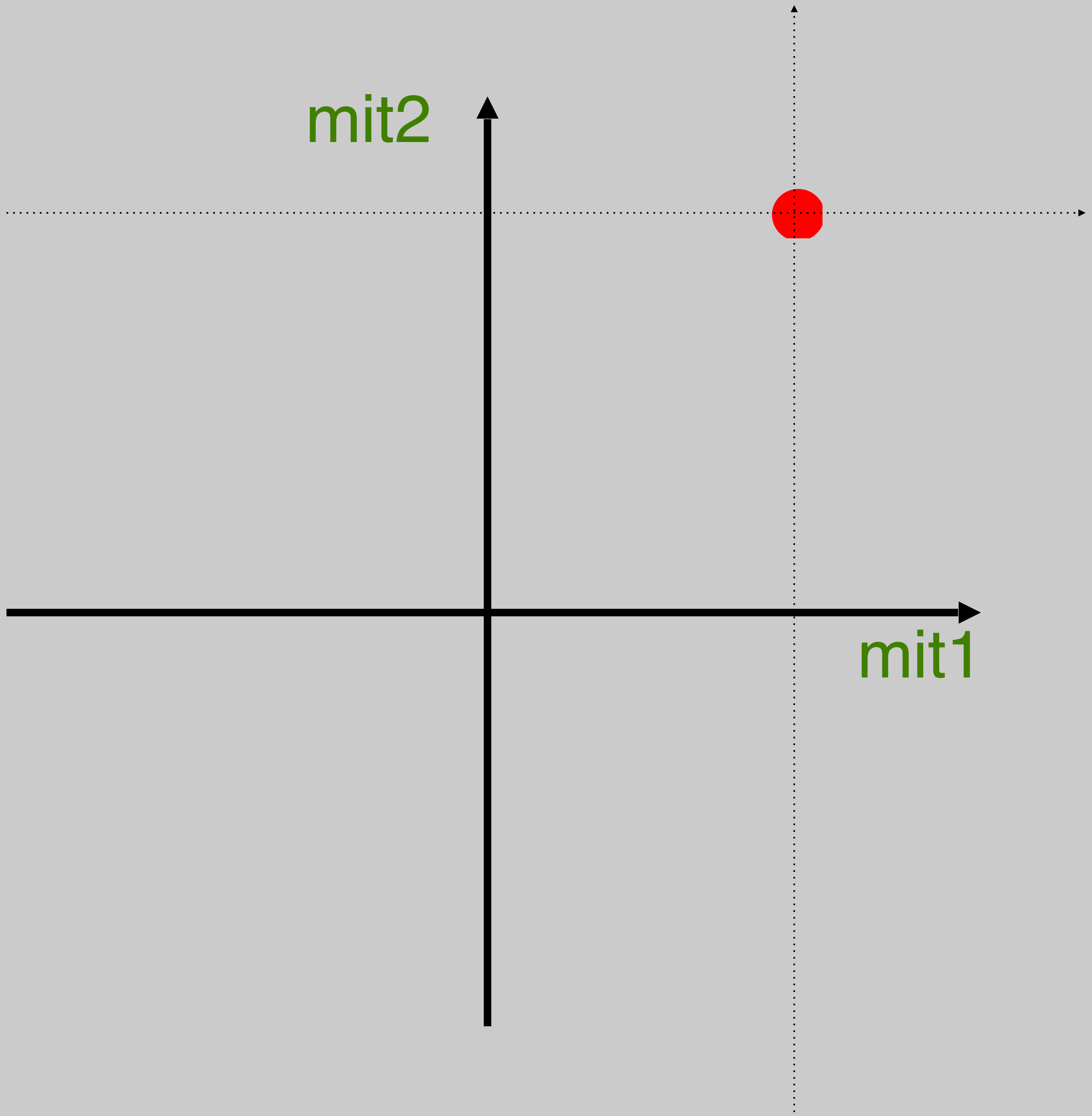
–K-means

– Continue proofs (symmetric matrices)

Example -- PCA

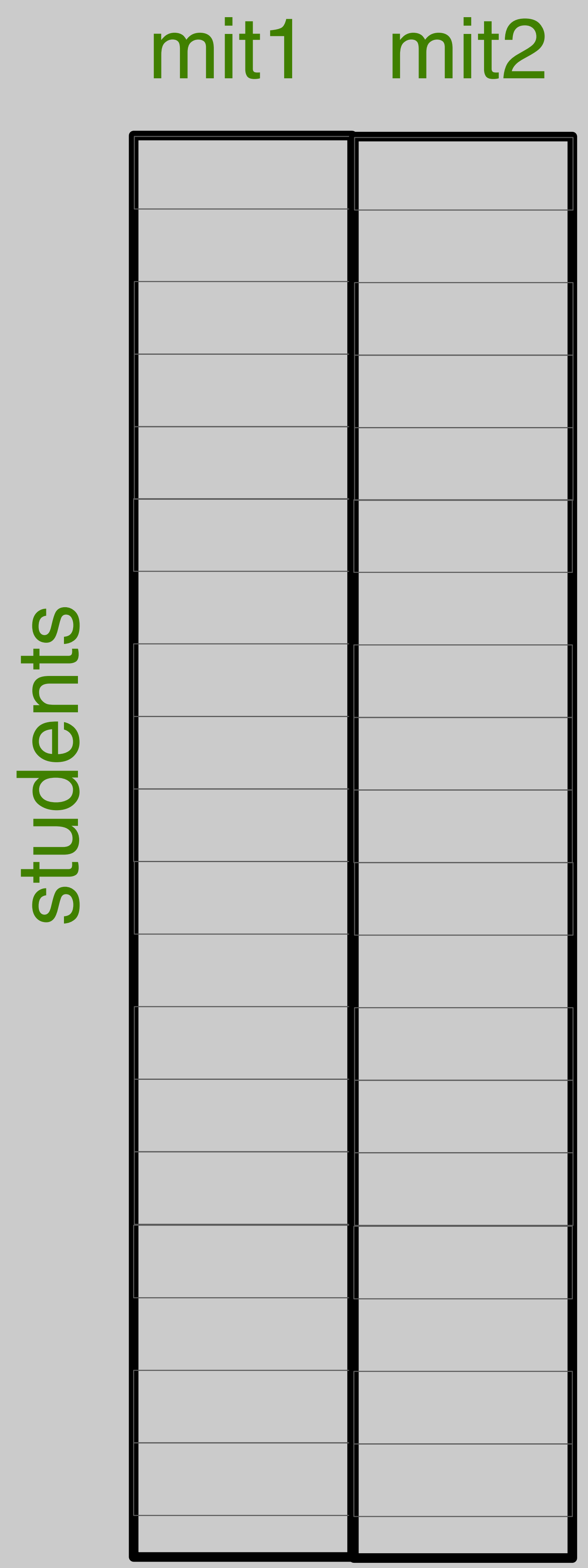
Consider miterm data

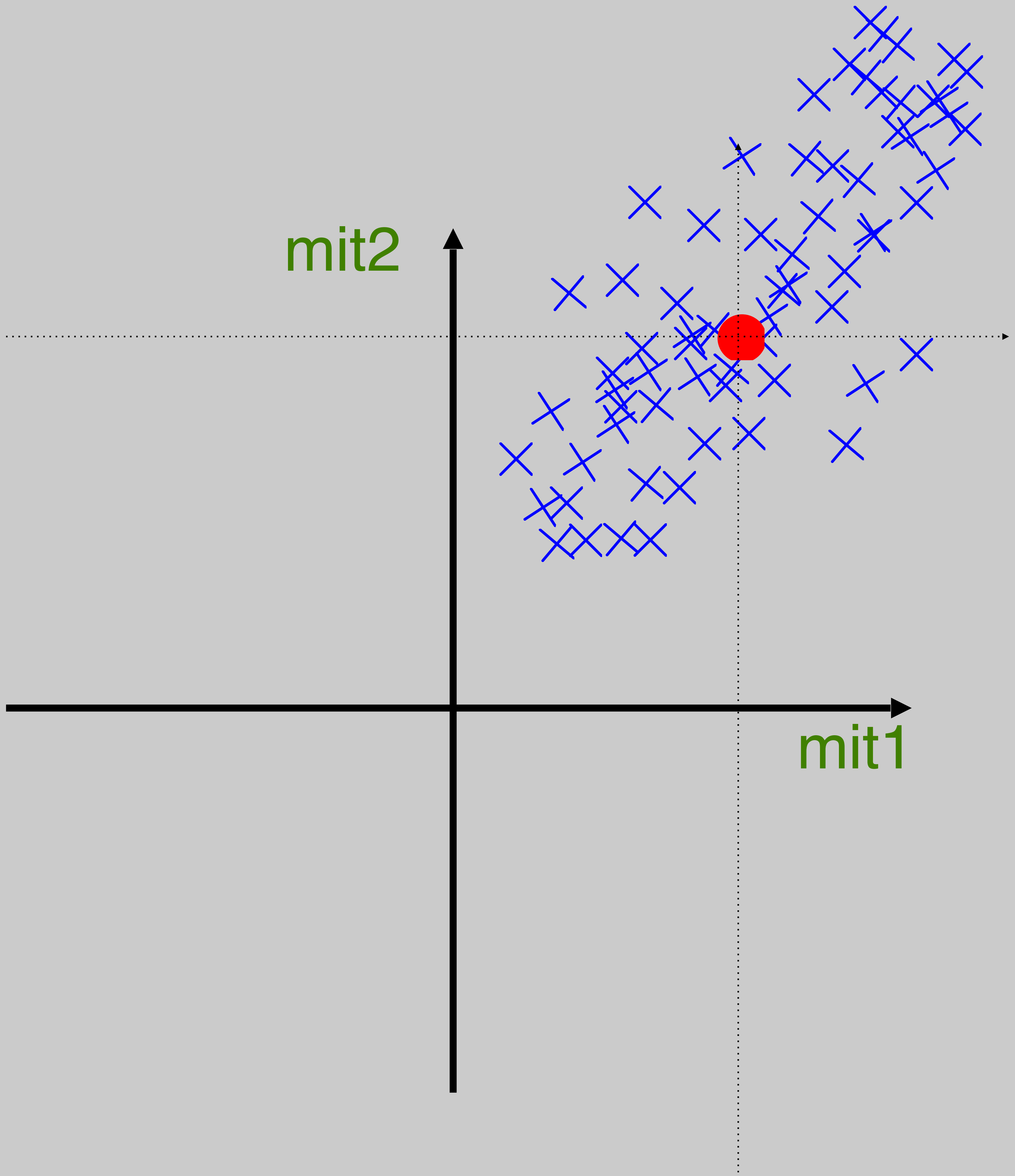




Example -- PCA

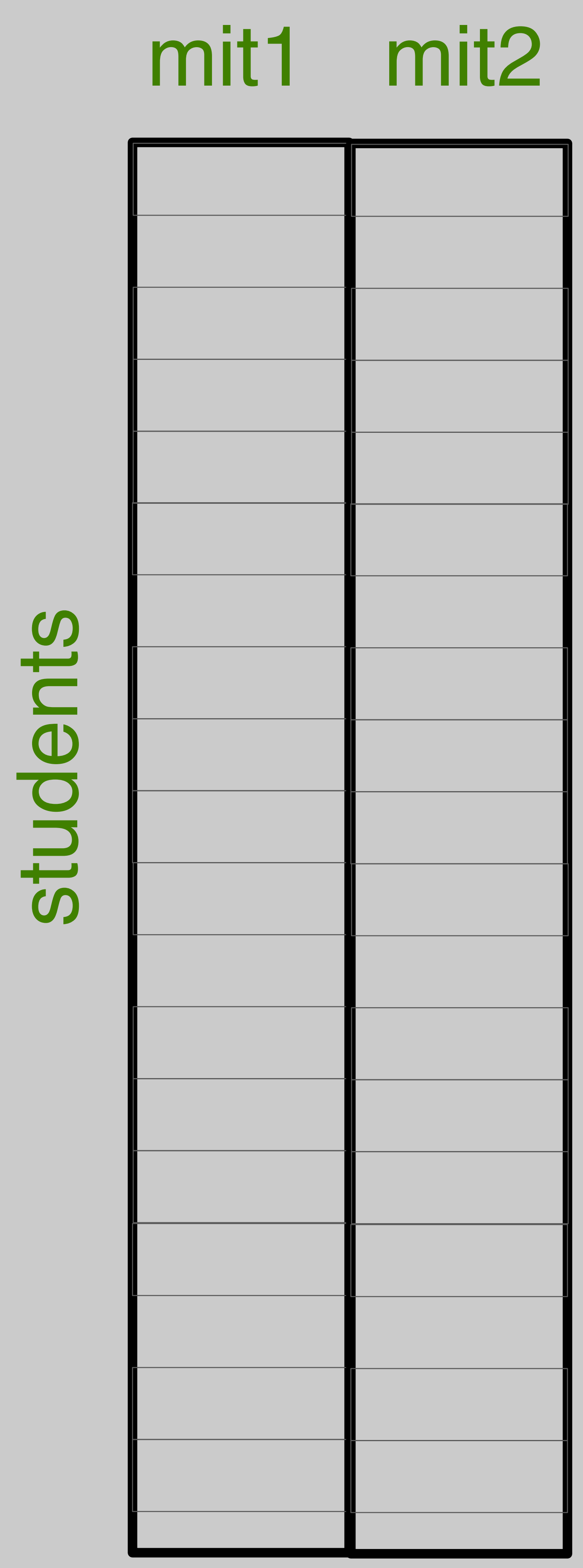
Consider miterm data

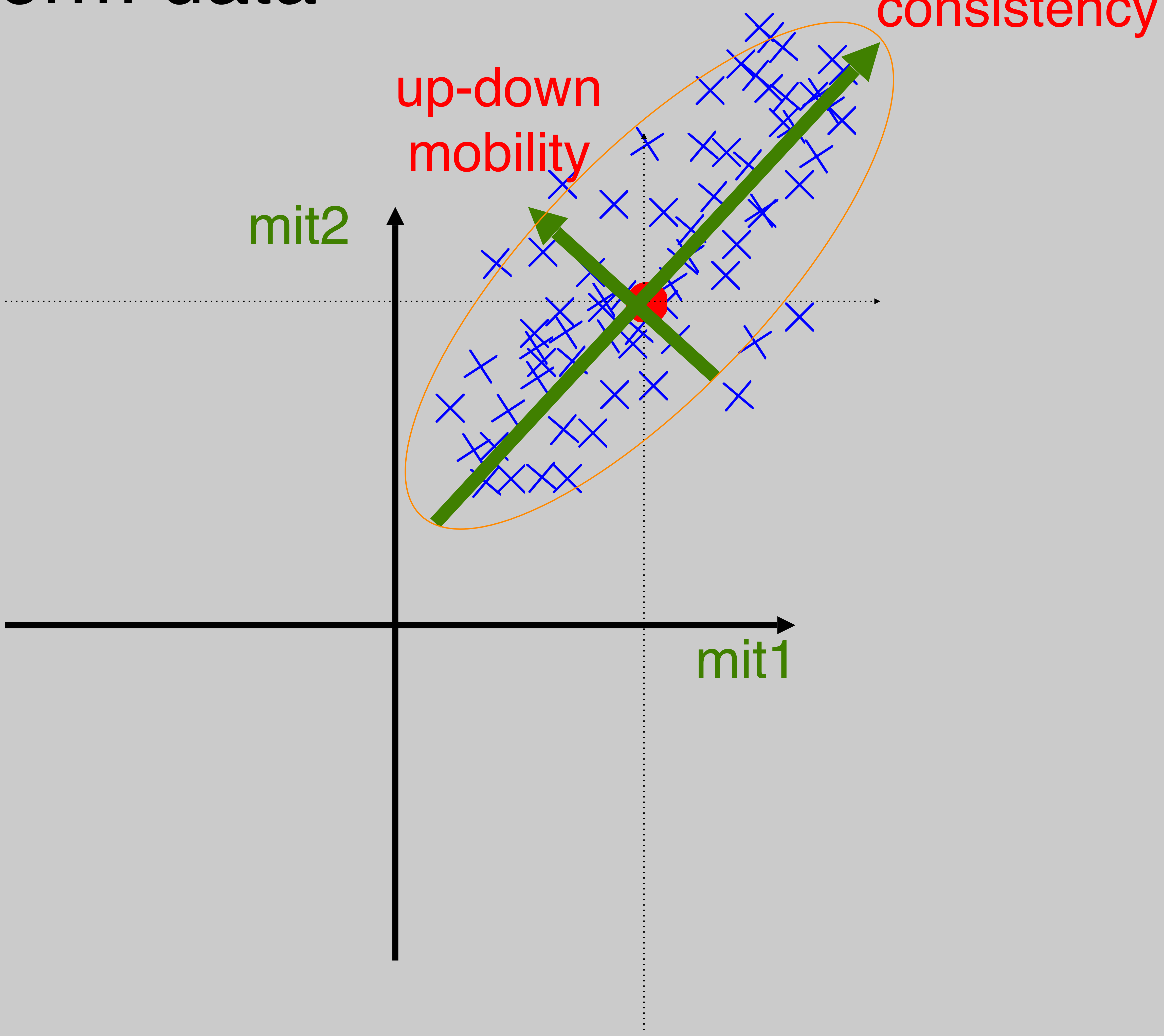




Example -- PCA

Consider miterm data

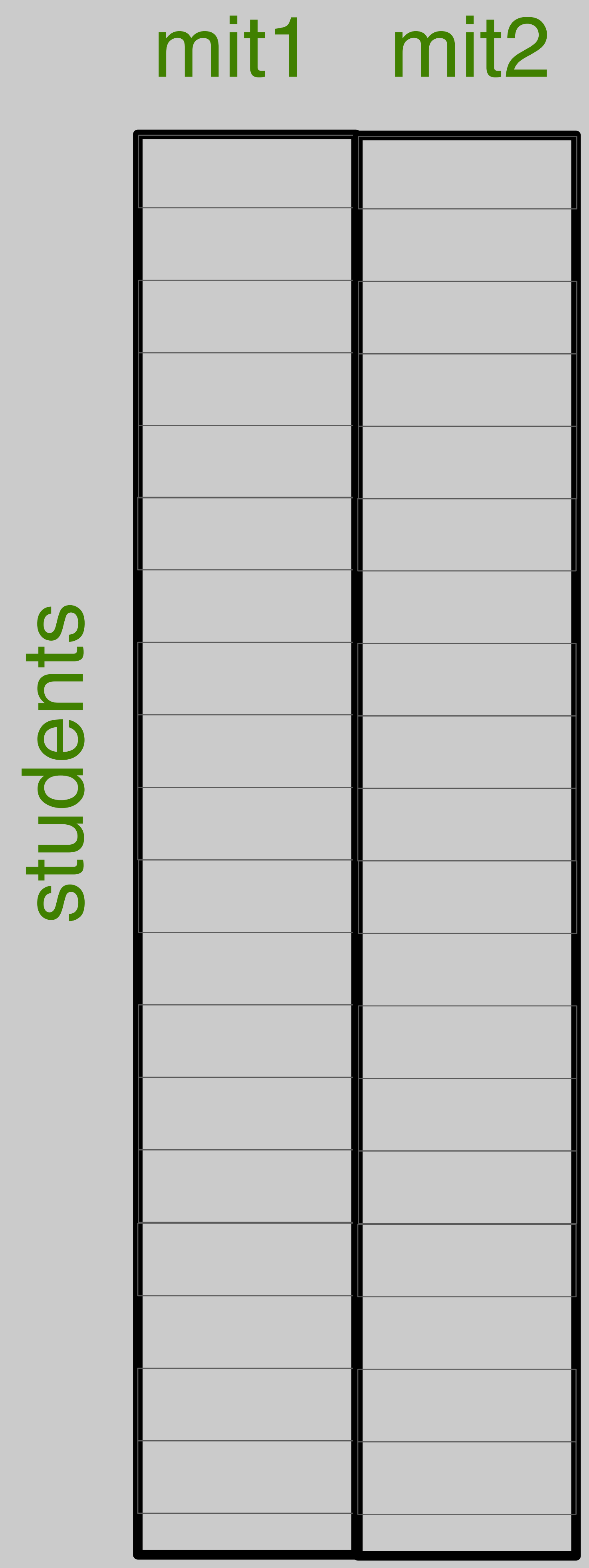


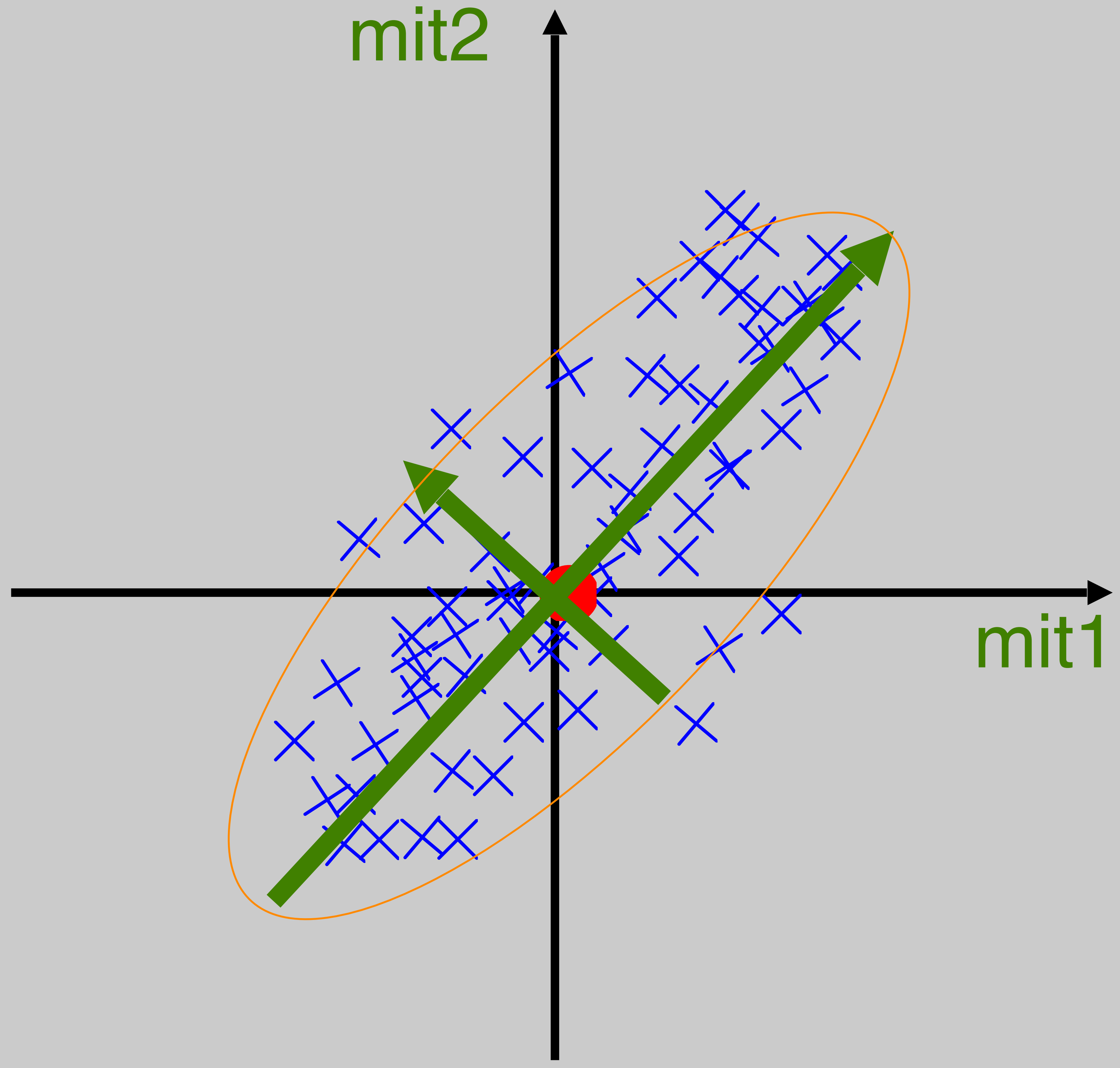


PCA Procedure

Remove averages from column of A

From $A^T A$, find $\sigma_i, \quad \vec{v_i}$





$\vec{v_i}$ are principal components!

A^TA as sample covariance matrix

$$A = \vec{a} \qquad a_\mu = \frac{1}{N}\sum_{i=0}^{N-1} a_i \qquad \tilde{A} = \vec{a} - a_\mu \vec{1}$$

$$\tilde{A}^T\tilde{A} = (\vec{a} - a_\mu\vec{1})^T(\vec{a} - a_\mu\vec{1})$$

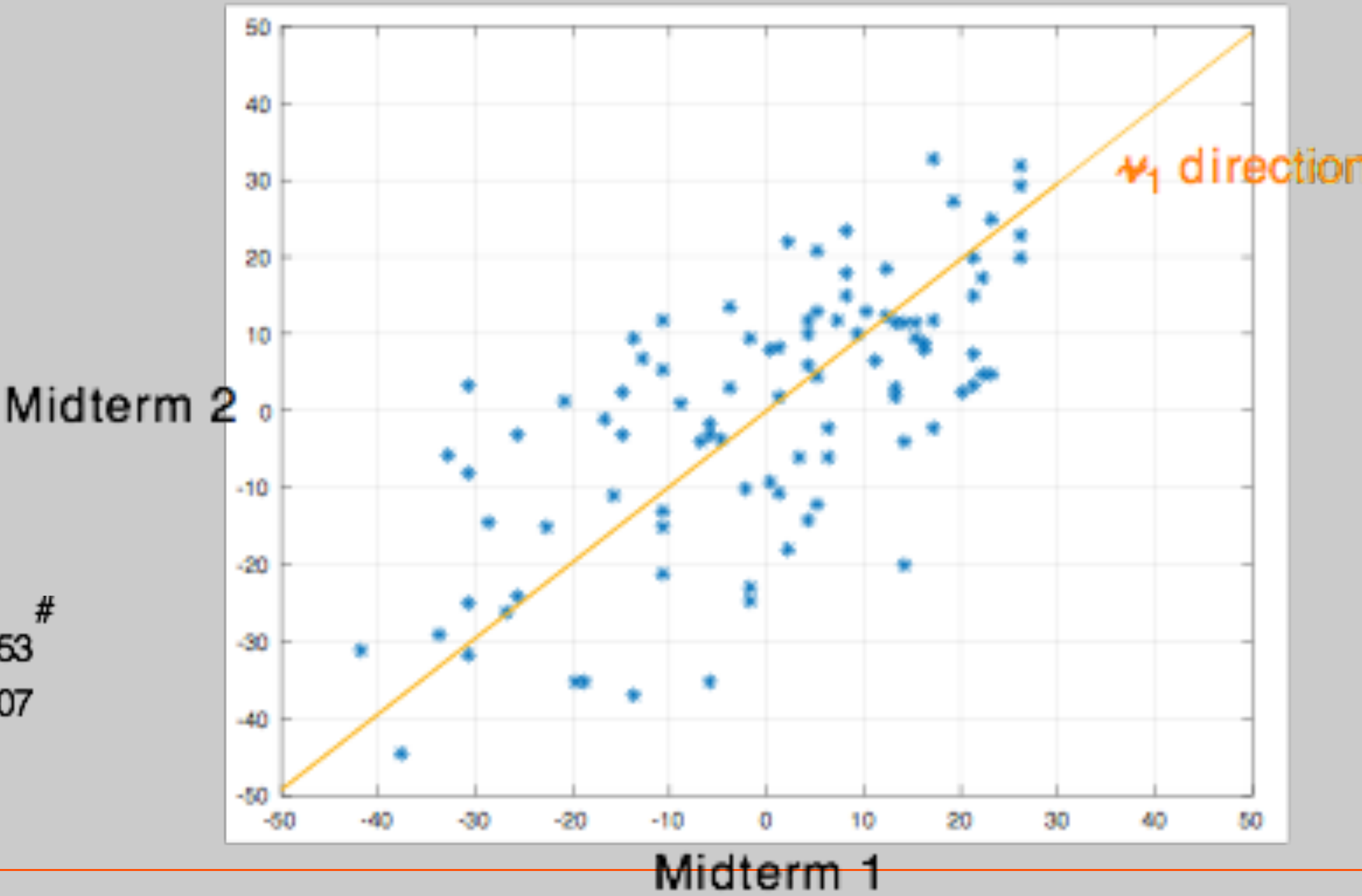
$$= \vec{a}^T\vec{a} - 2Na_\mu^2 + Na_\mu^2 \quad = \vec{a}^T\vec{a} - Na_\mu^2$$

$$\frac{1}{N}\tilde{A}^T\tilde{A} = \frac{1}{N}\vec{a}^T\vec{a} - a_\mu^2 \; = \frac{1}{N}\sum_{i=0}^{N-1} a_i^2 - a_\mu^2 \quad = a_\sigma^2$$

Variance!

Example midterm

$$\frac{1}{93} A^T A = \begin{bmatrix} 297.69 & 202.53 \\ 202.53 & 292.07 \end{bmatrix}$$

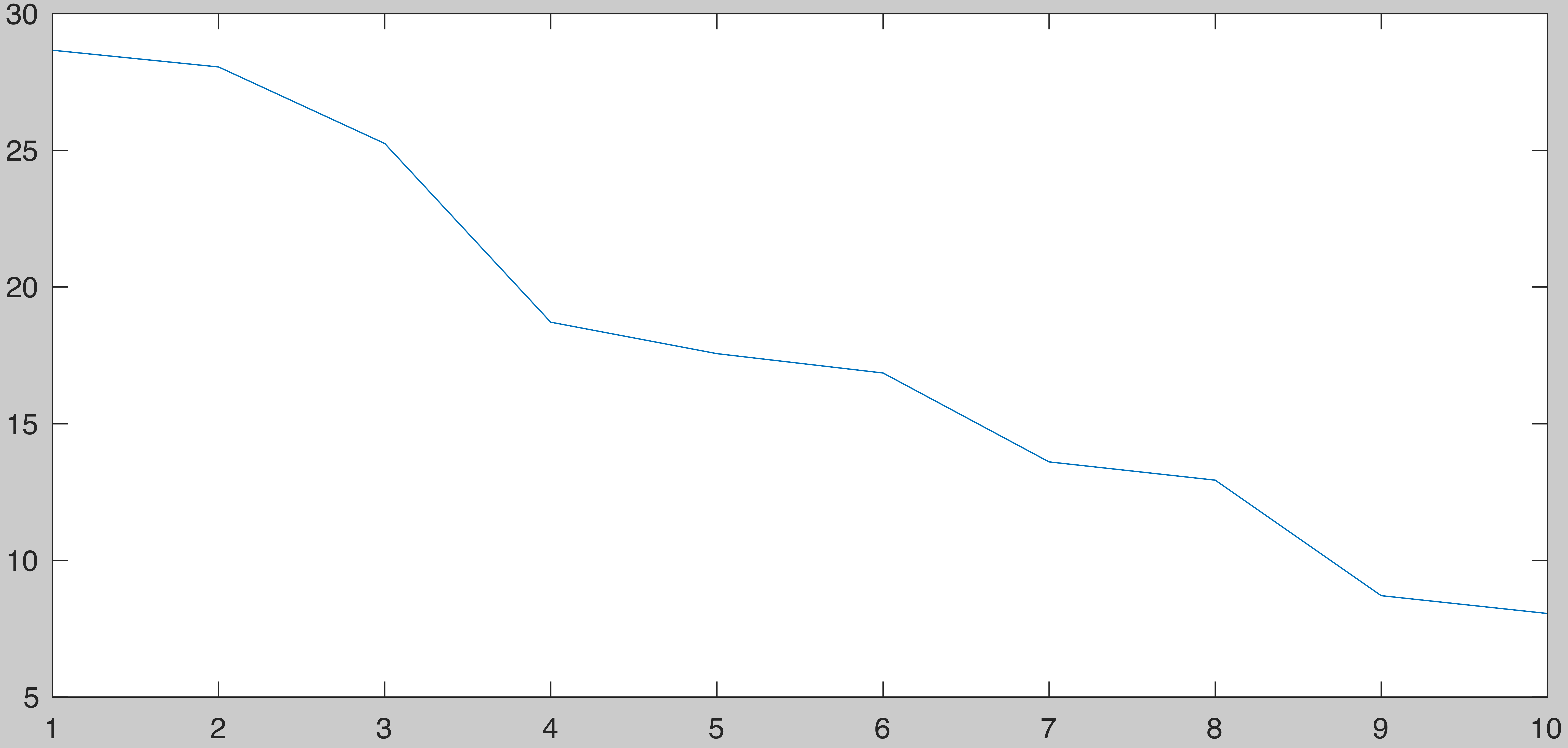


Mid Semester Survey Results

1) How’s going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture

Data Science

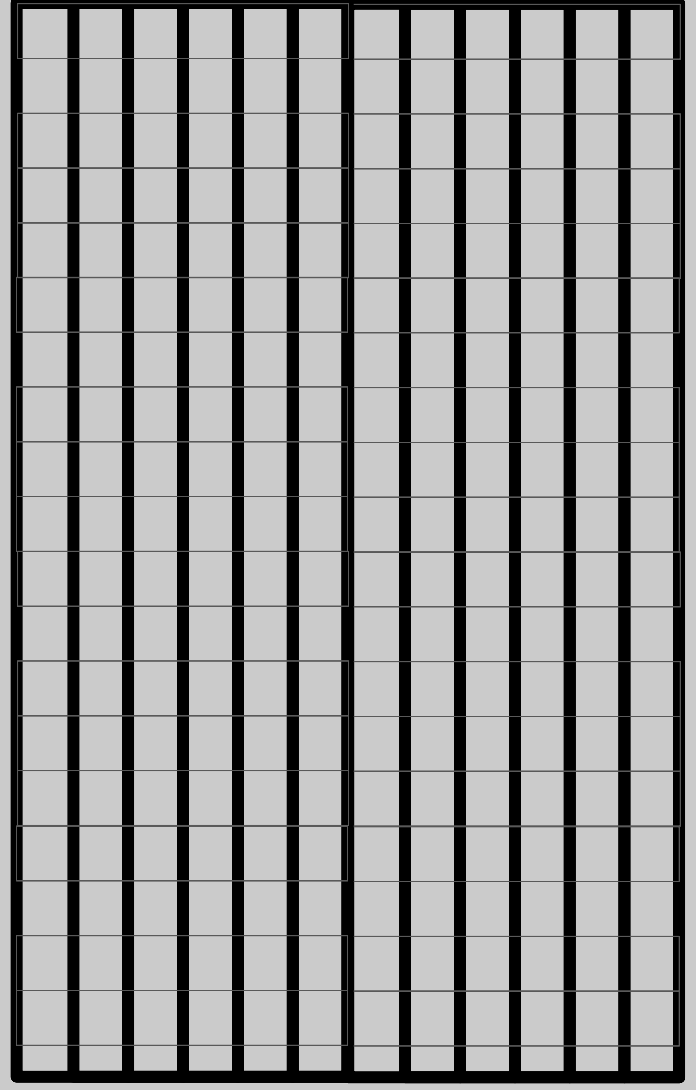
Singular values

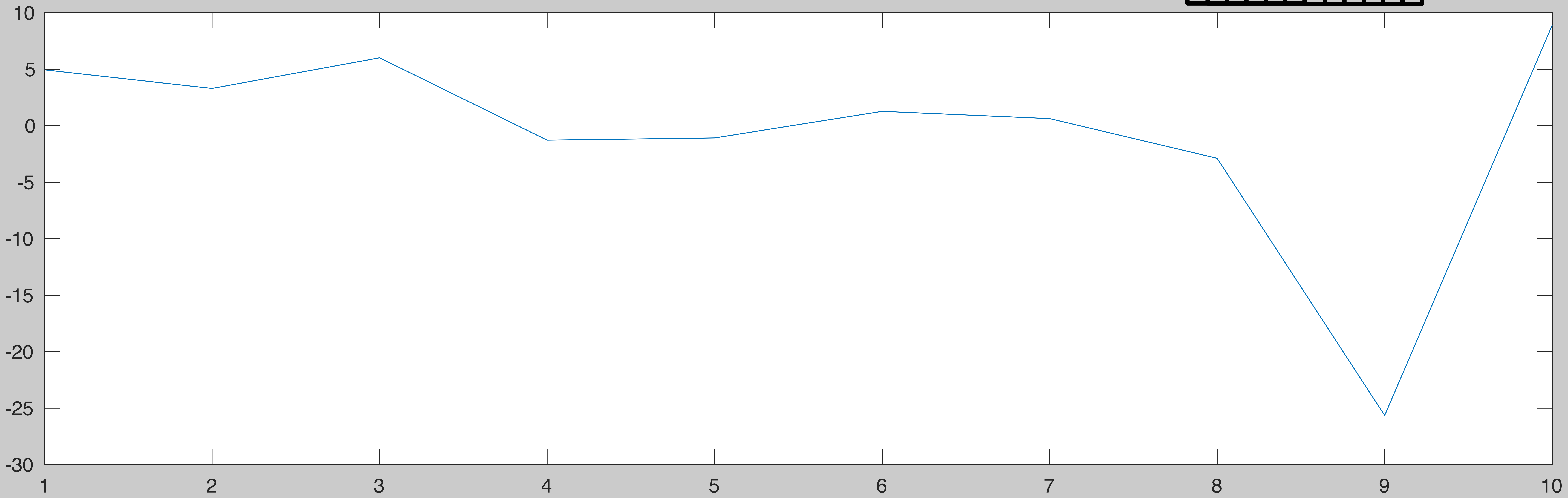


Data Science

$$A^T \vec{u}_1$$

1) How's going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture

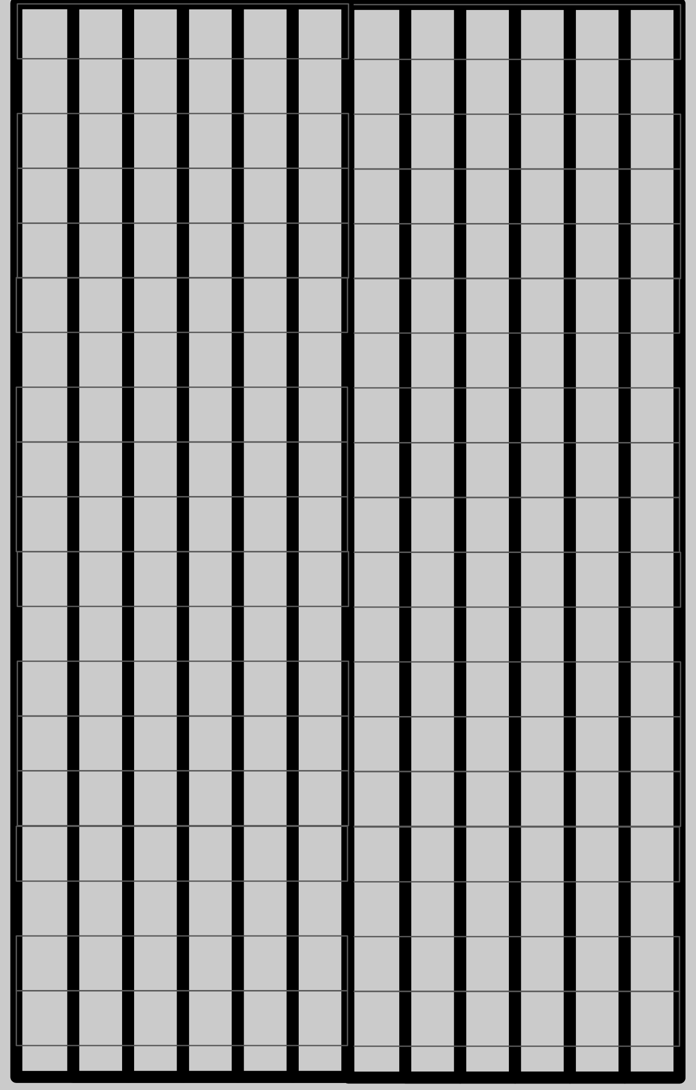




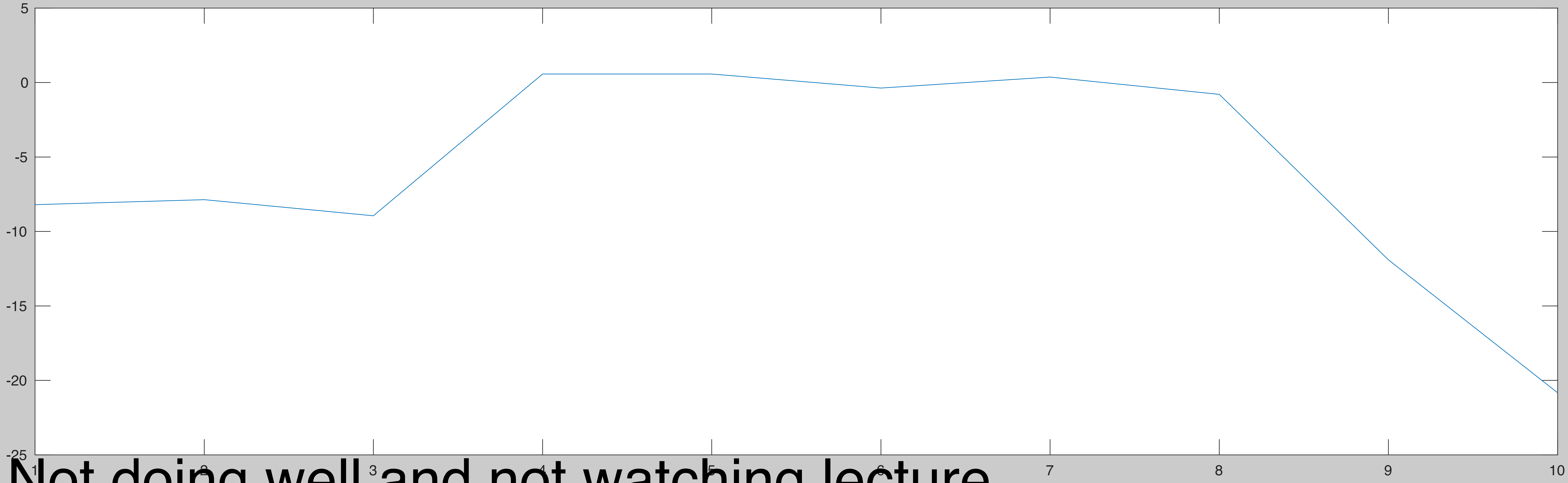
Correlated mostly with 16A

Data Science

1) How's going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture



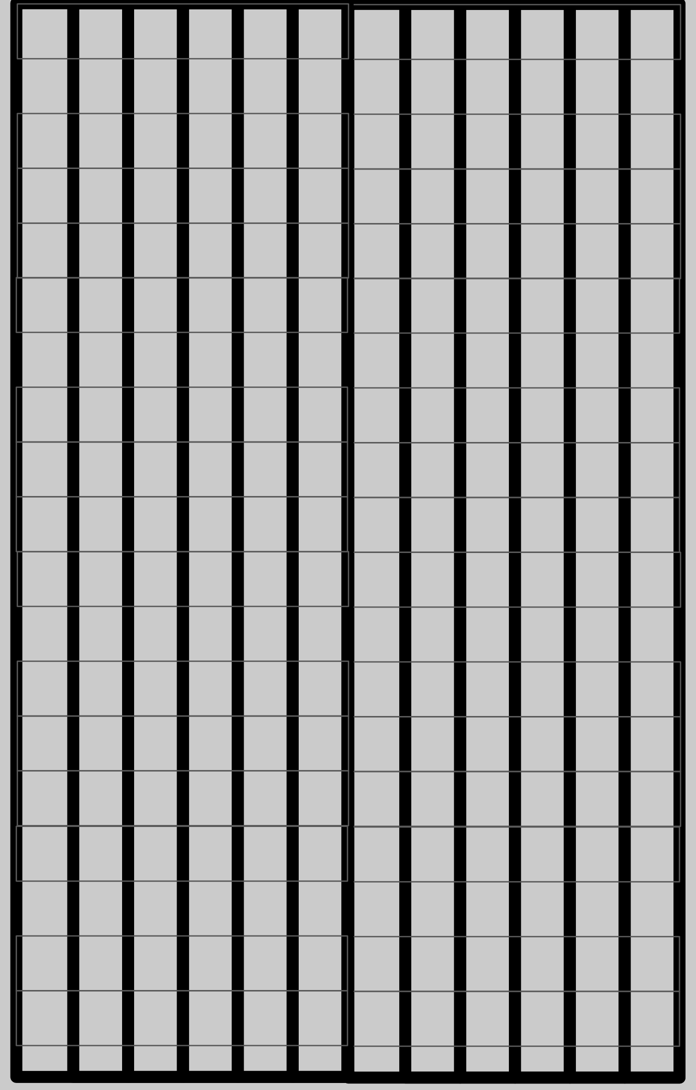
$$A^T \vec{u}_2$$



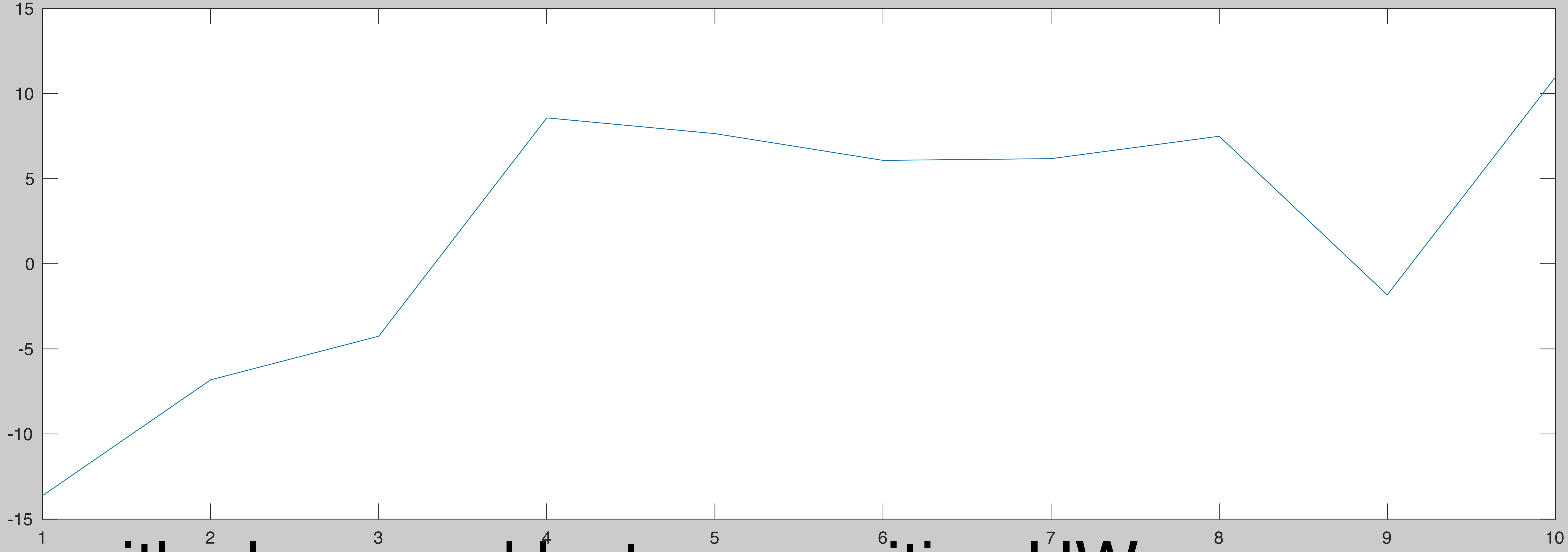
Not doing well and not watching lecture

Data Science

1) How’s going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture



$$A^T \vec{u}_3$$

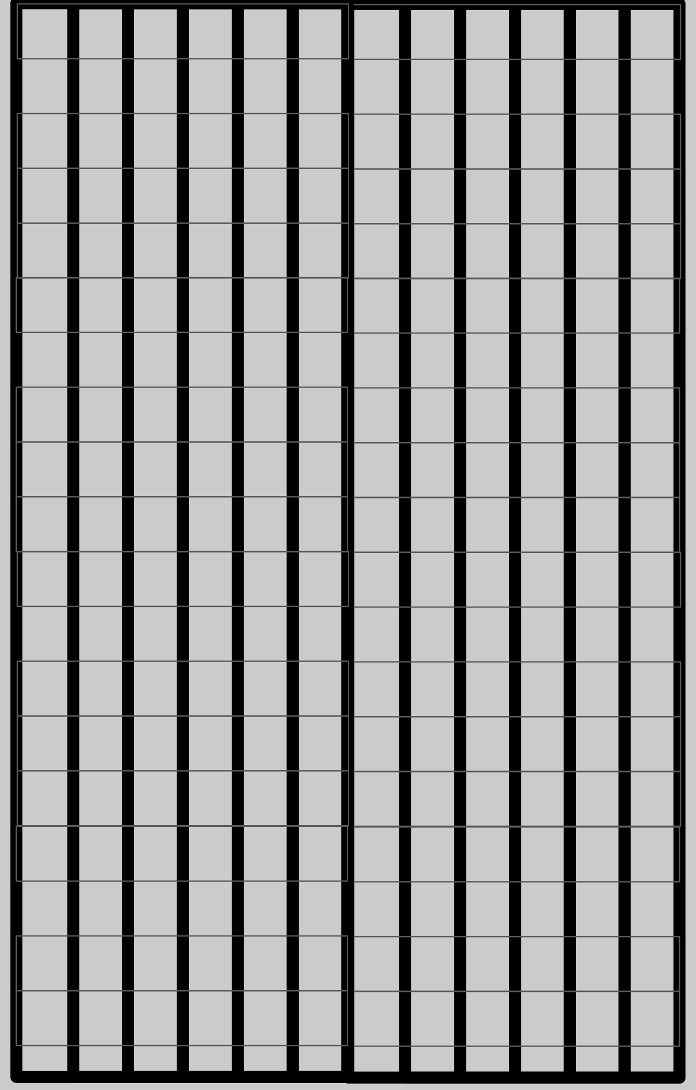


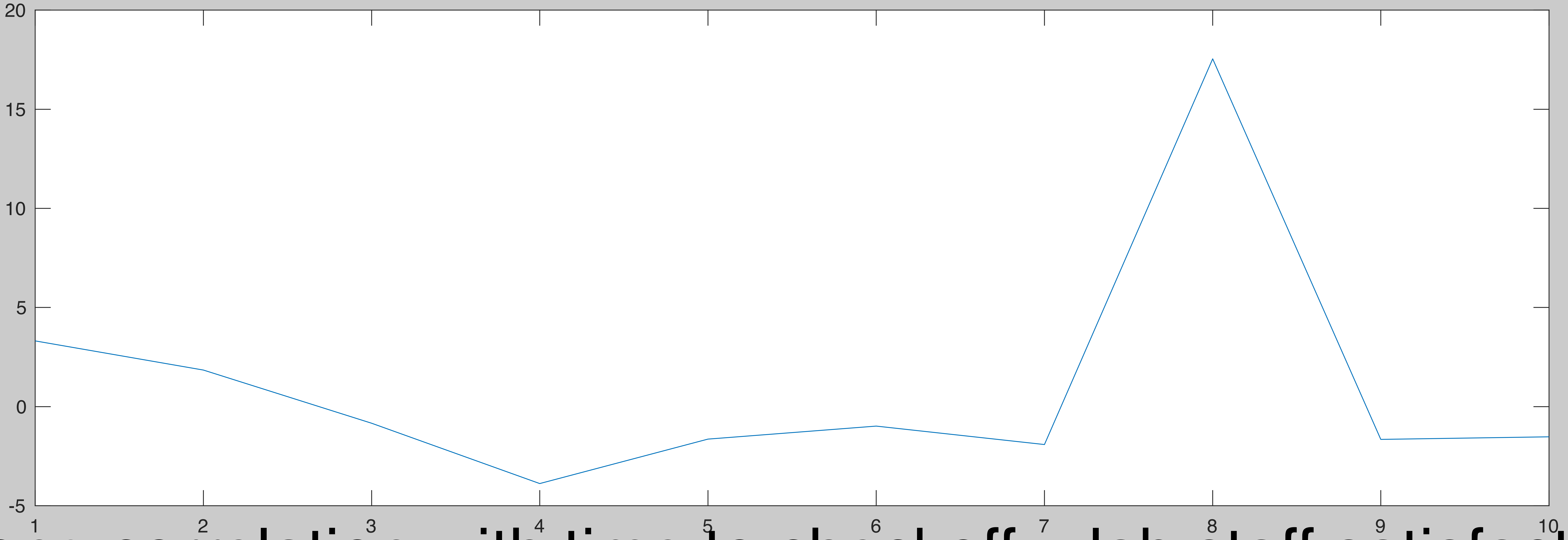
Negative with class, and lecture, positive HW

Data Science

$$A^T \vec{u}_4$$

1) How's going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture



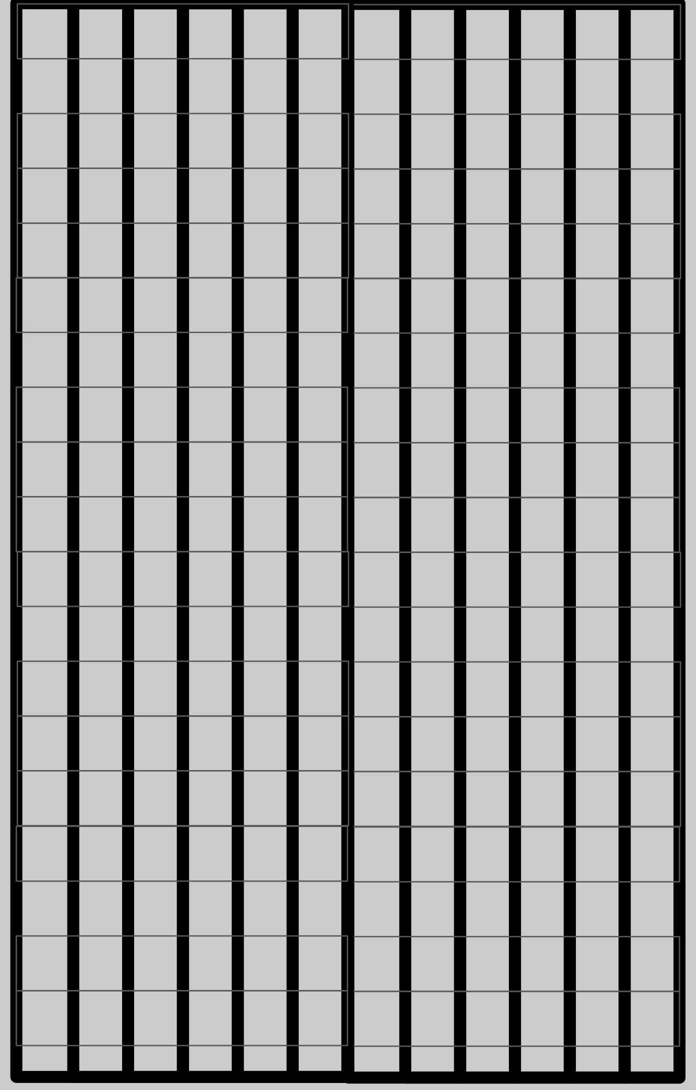


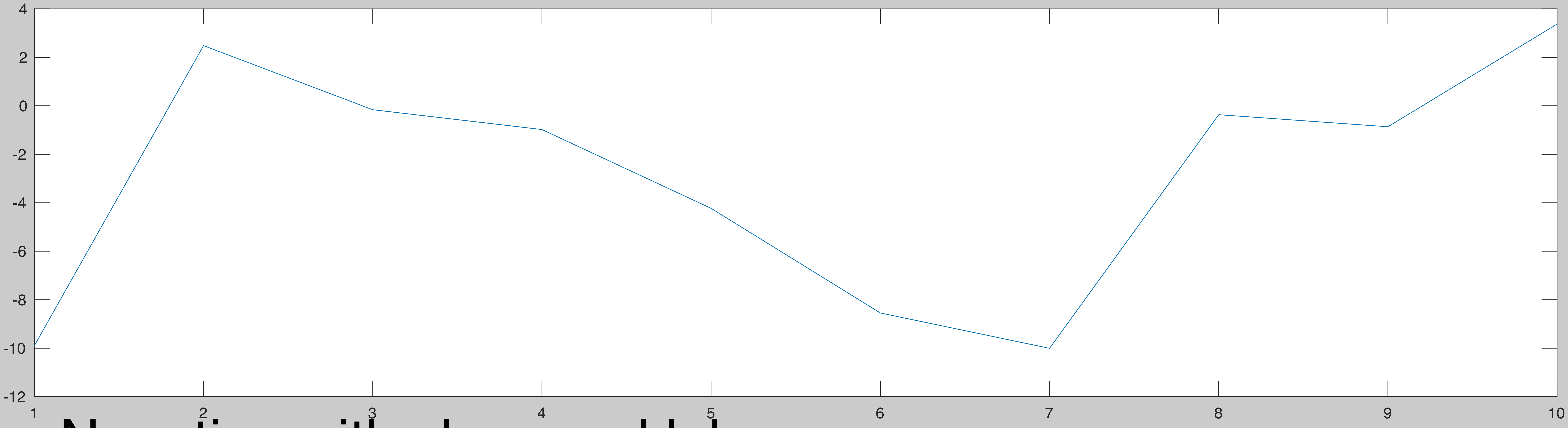
Strong correlation with time to checkoff – lab staff satisfaction?

Data Science

$$A^T \vec{u}_5$$

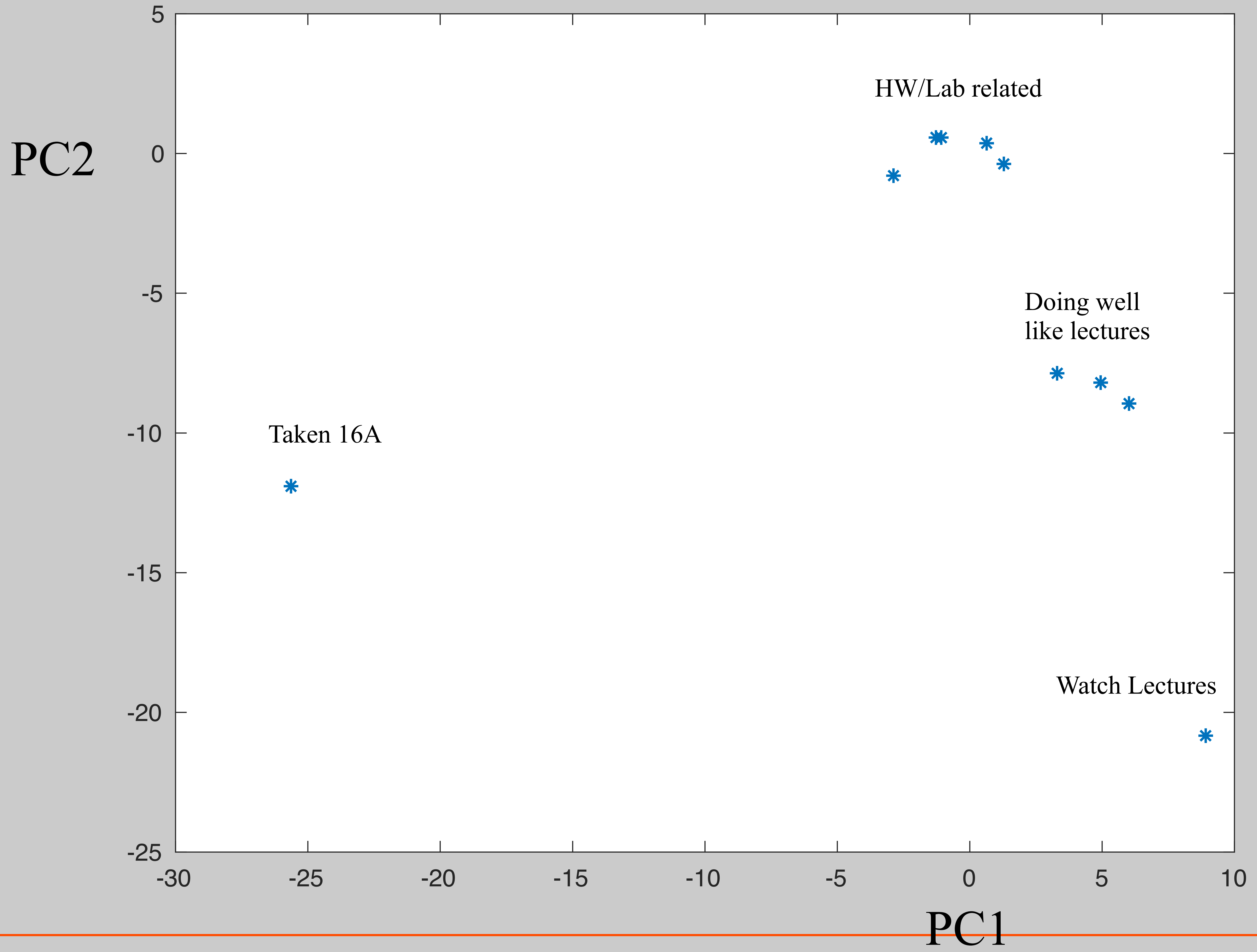
1) How's going
2) Circuits lecture
3) Systems lecture
4) Homework difficulty
5) Length of HW
6) Lab difficulty
7) Length of lab
8) Time to checkoff
9) 16A
10) Attend lecture



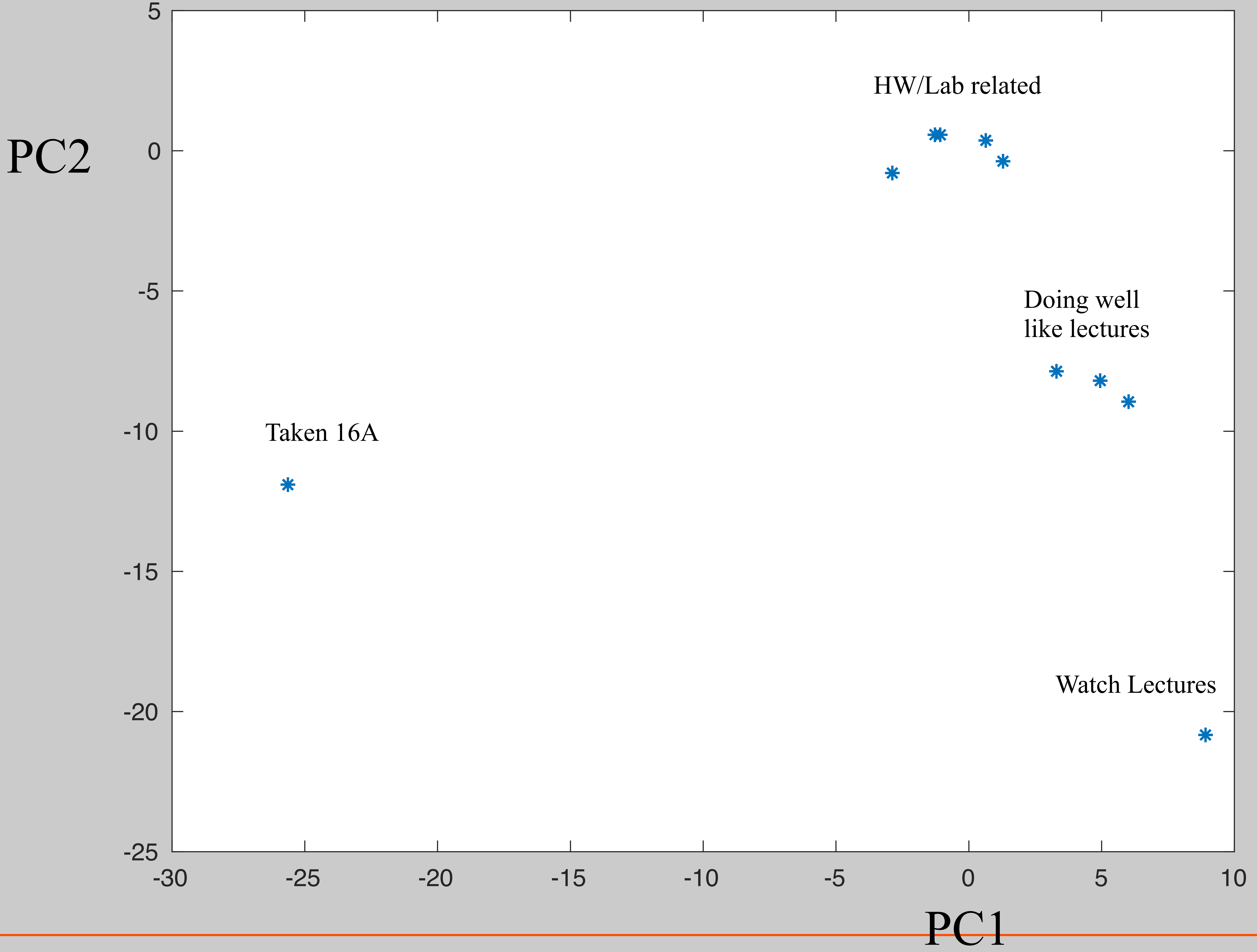


Negative with class and lab

Data Science

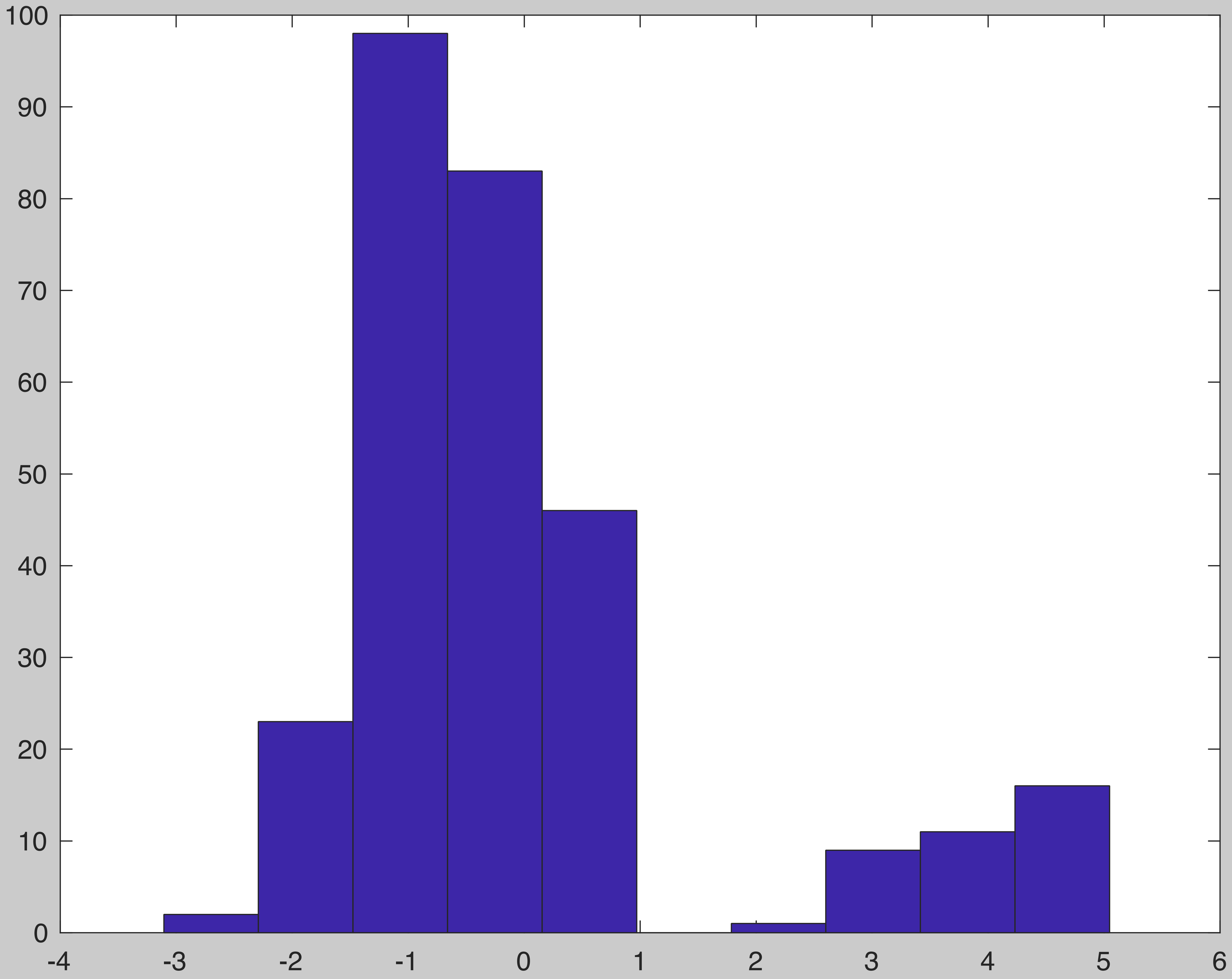


Data Science

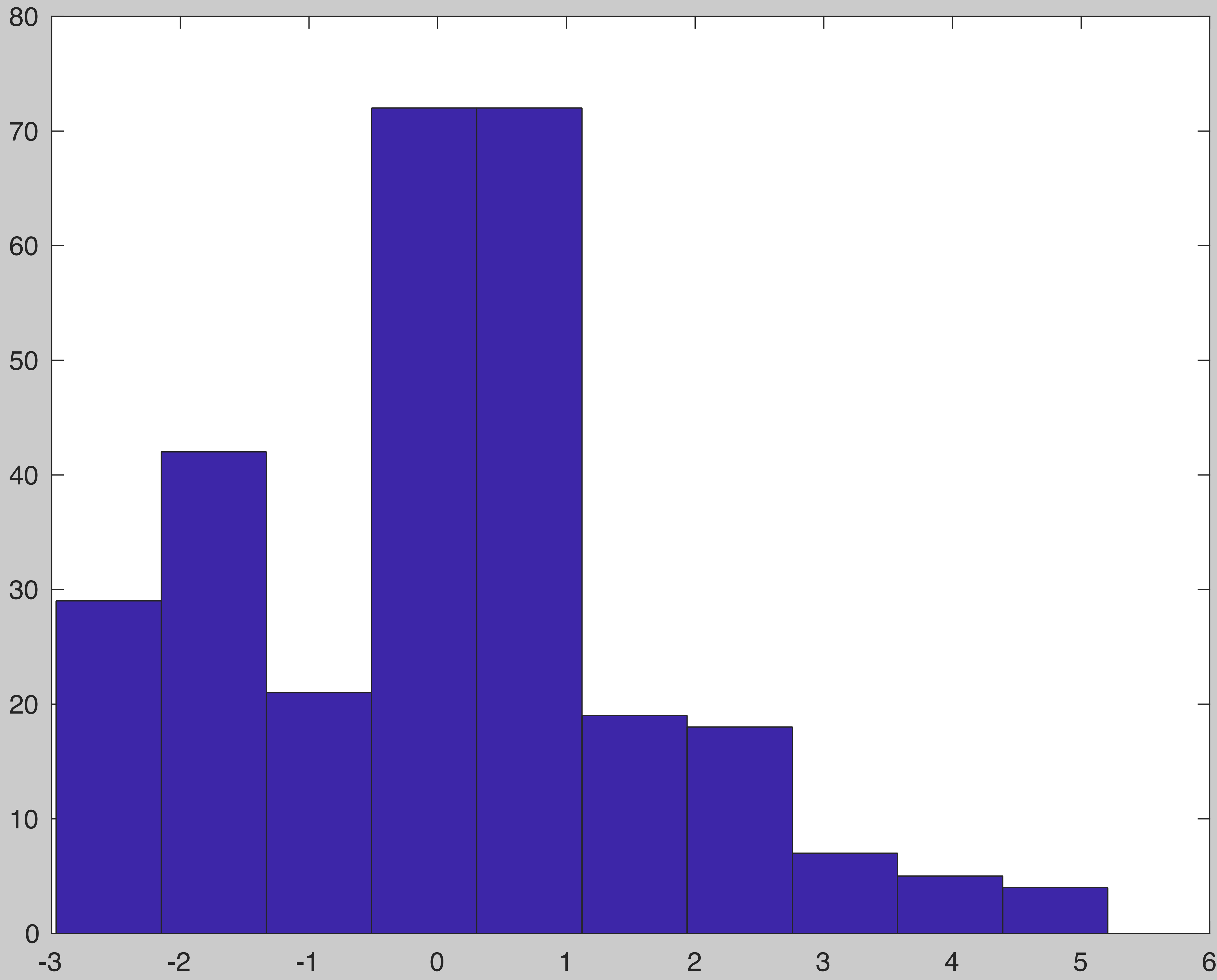


Data Science

$A\vec{v}_1$



$A\vec{v}_2$



PCA in Genetics Reveals Geography

Genes mirror geography within Europe
Nature **456**, 98-101 (6 November 2008)

Study:

Characterized genetic variatios in 3,000 Europeans from 36 Countries

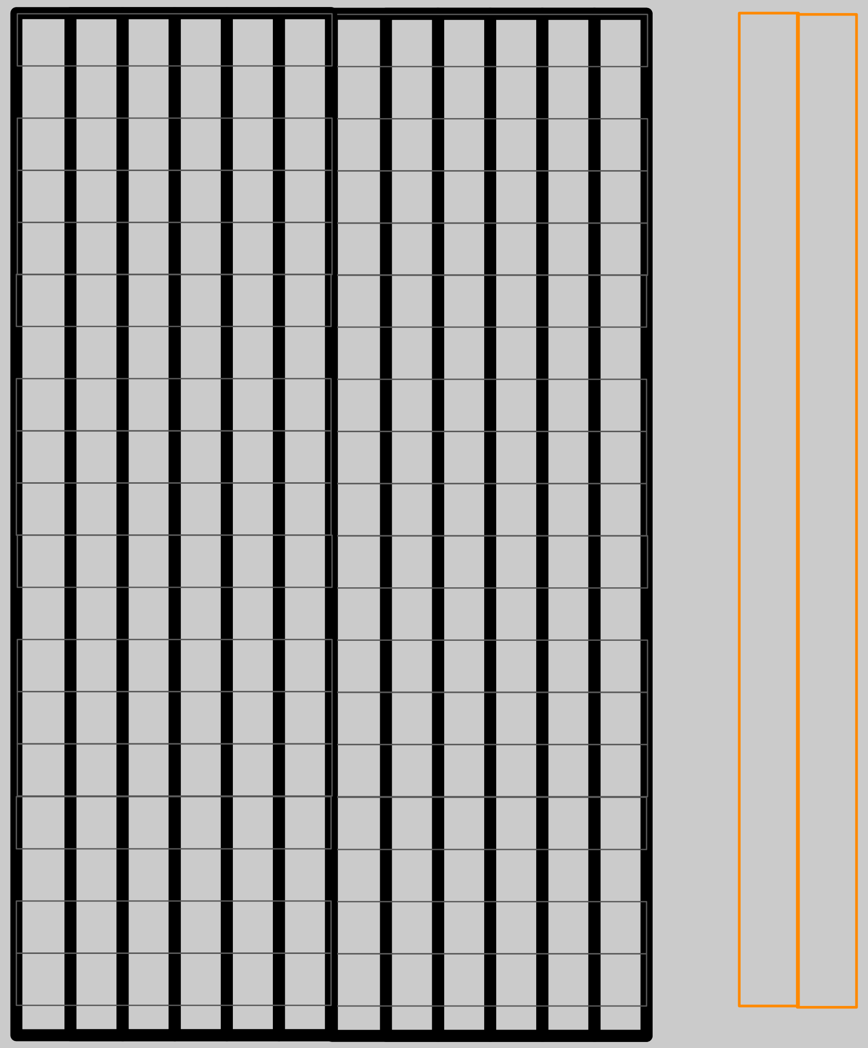
Built a matrix of 200K SNPs (single nucleotide polymorphisms)

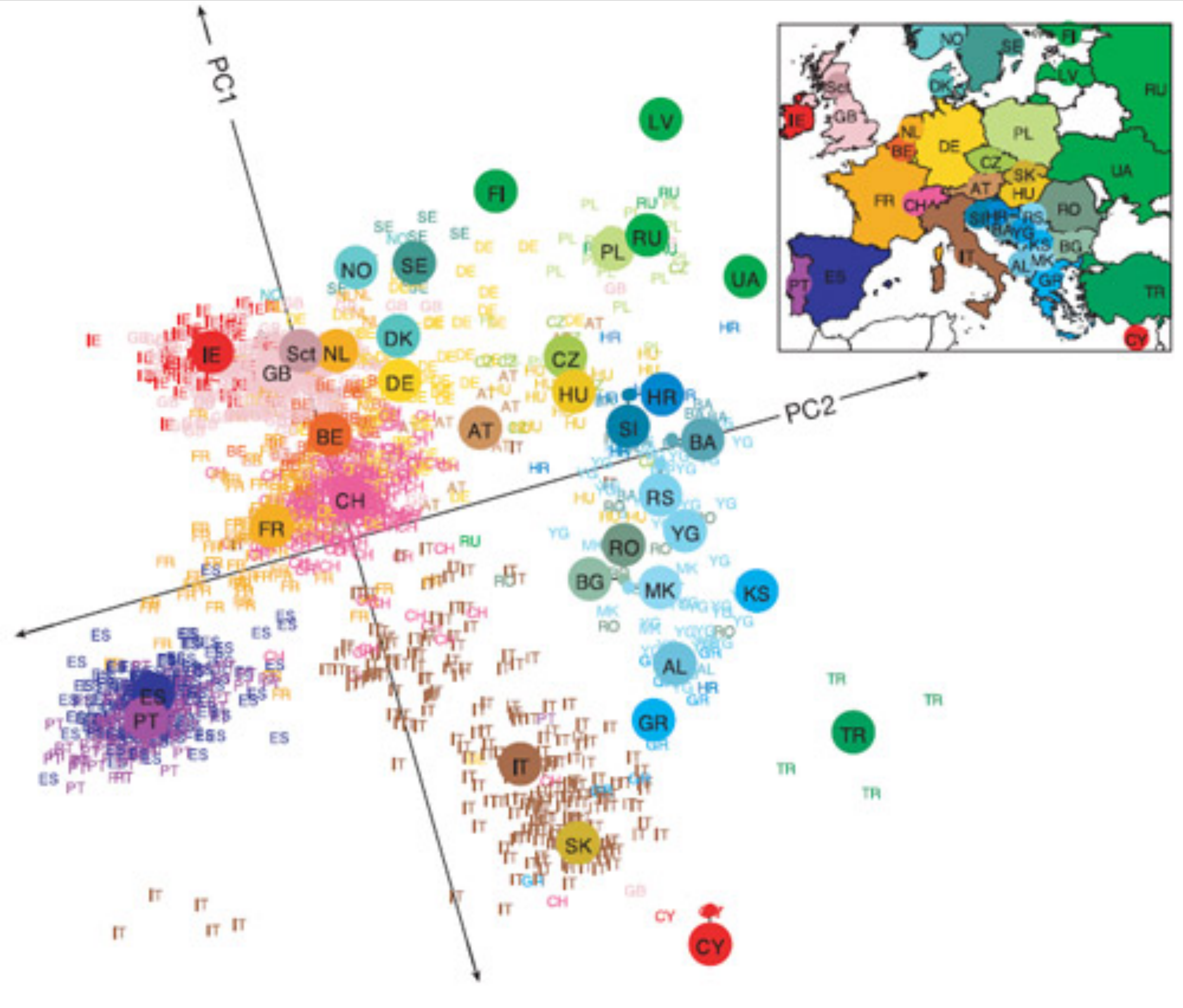
Computed largest 2 principle components

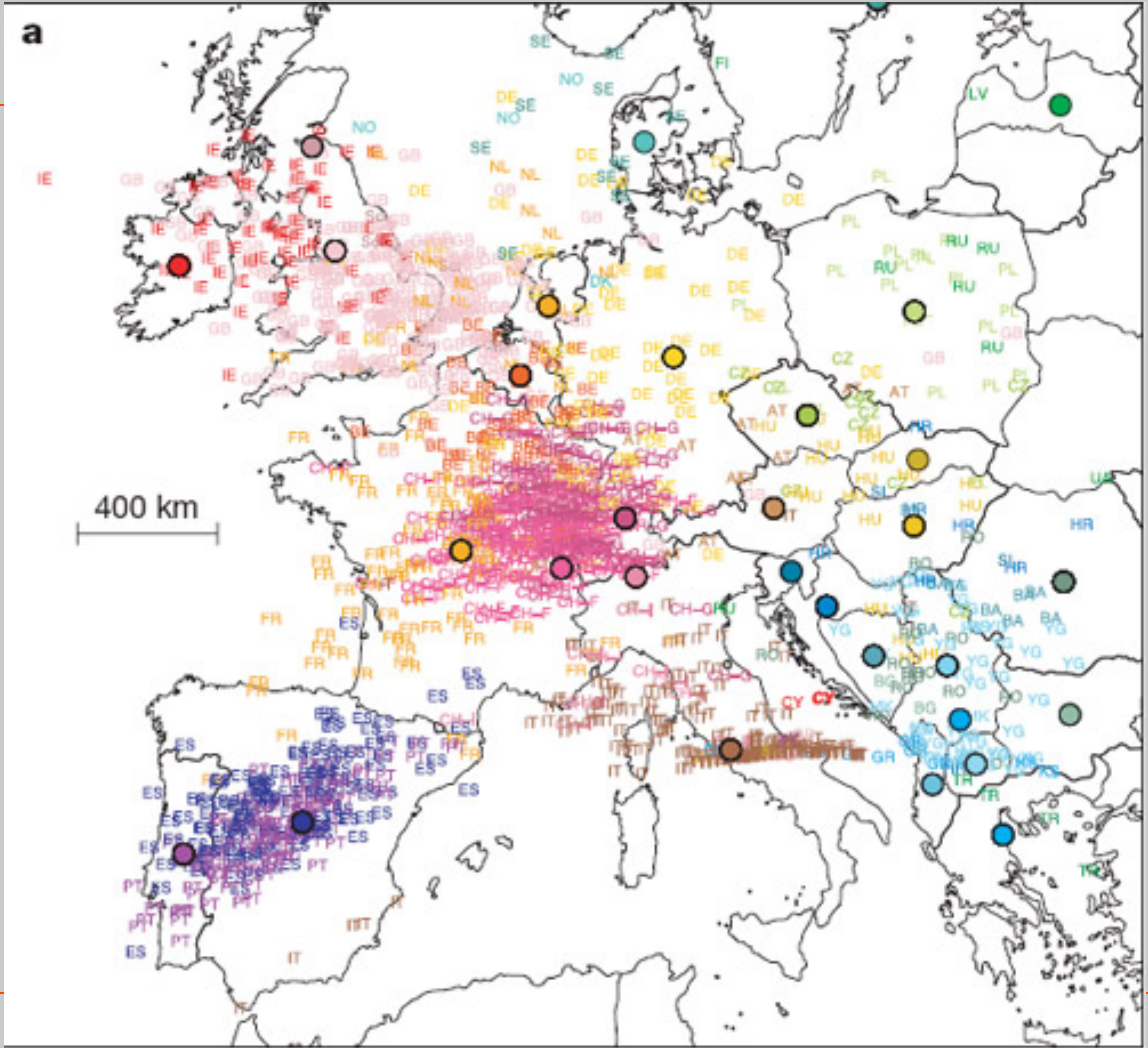
Projected subjects on 2 dimentional data

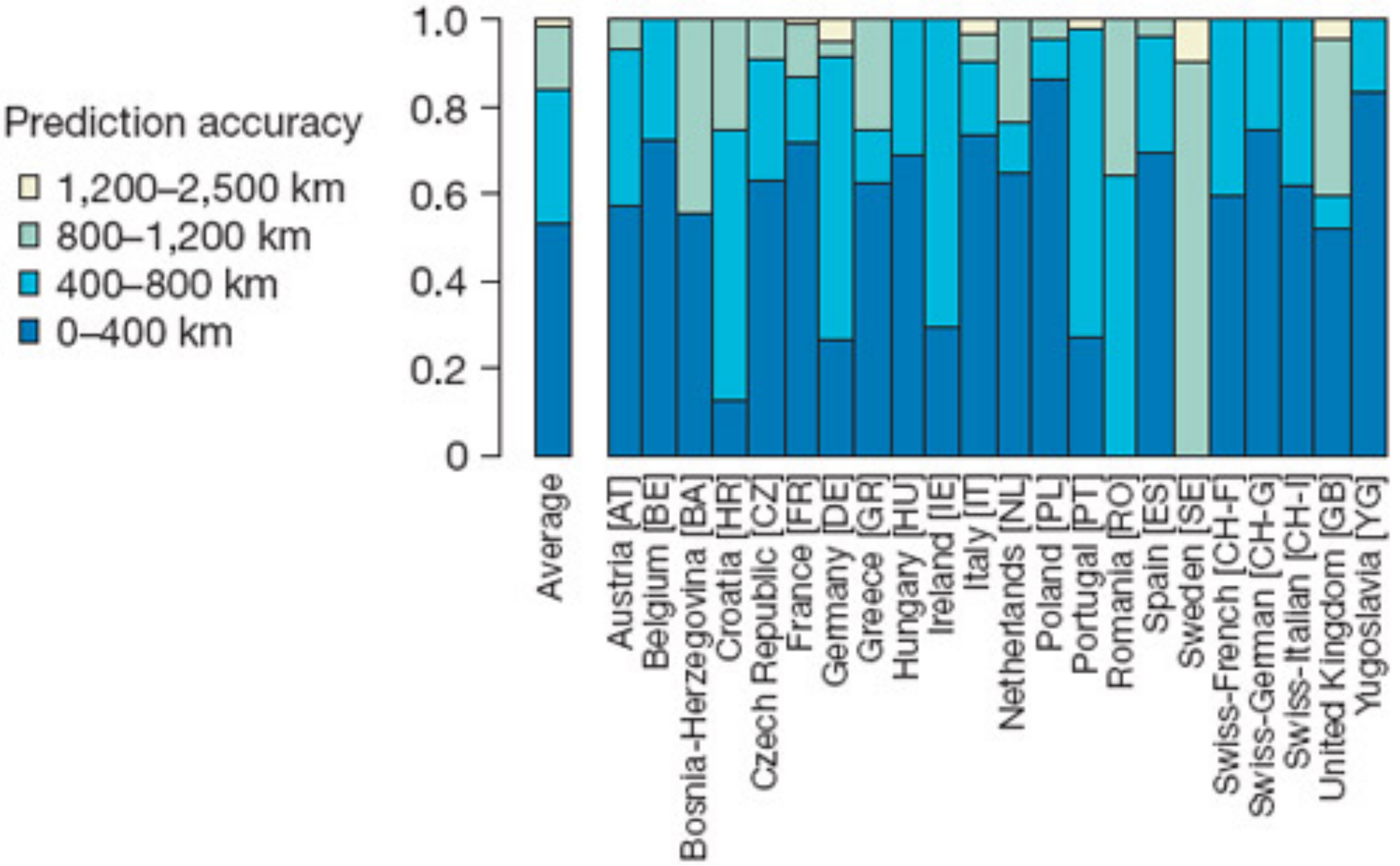
Overlayed the result on the map of Europe

$A\vec{v}_1 \quad A\vec{v}_2$




PC1
PC2


a
400 km



Prediction accuracy
1,200–2,500 km
800–1,200 km
400–800 km
0–400 km
1.0
0.8
0.6
0.4
0.2
0
Average
Austria [AT]
Belgium [BE]
Bosnia-Herzegovina [BA]
Croatia [HR]
Czech Republic [CZ]
France [FR]
Germany [DE]
Greece [GR]
Hungary [HU]
Ireland [IE]
Italy [IT]
Netherlands [NL]
Poland [PL]
Portugal [PT]
Romania [RO]
Spain [ES]
Sweden [SE]
Swiss-French [CH-F]
Swiss-German [CH-G]
Swiss-Italian [CH-I]
United Kingdom [GB]
Yugoslavia [YG]

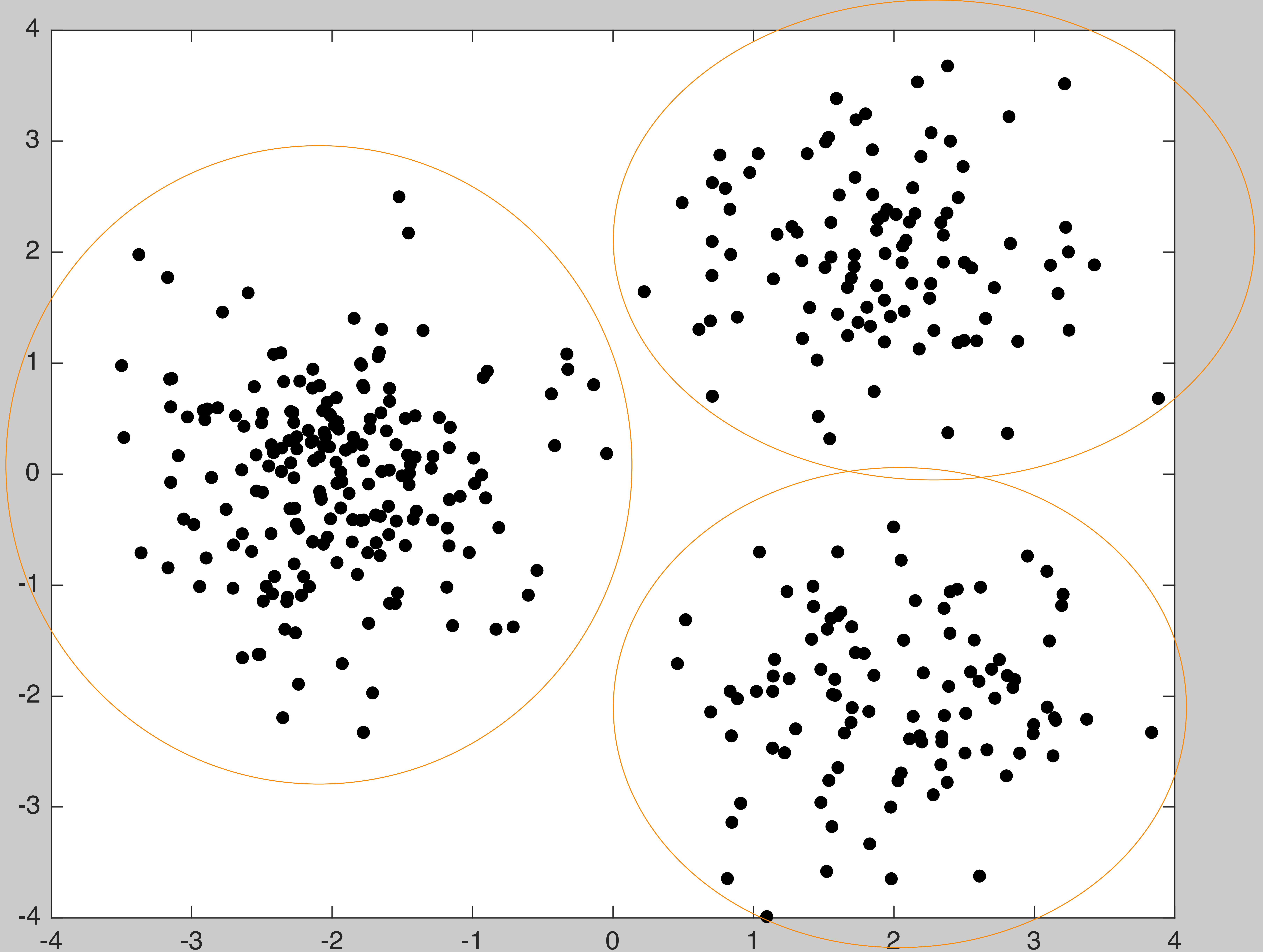
Interesting conclusions

“The results have implications for a lot of biomedical research. Many scientists are scanning entire genomes on a hunt for SNPs that affect a person’s risk of diseases like cancer or their reaction to drugs. Novembre says that researchers who are running these “whole-genome studies” need to bear in mind where their sample has come from. Even if a study looks at a small and seemingly related parts of Europe, it would have to adjust for any geographical influences in the genetic variations it uncovers.”

http://phenomena.nationalgeographic.com/2008/09/01/european-genes-mirror-european-geography/

23 and me

Michael Lustig		100%
●	European	98.9%
●	Middle Eastern & North African	0.9%
●	East Asian & Native American	< 0.1%
●	Unassigned	0.2%


4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

k-means

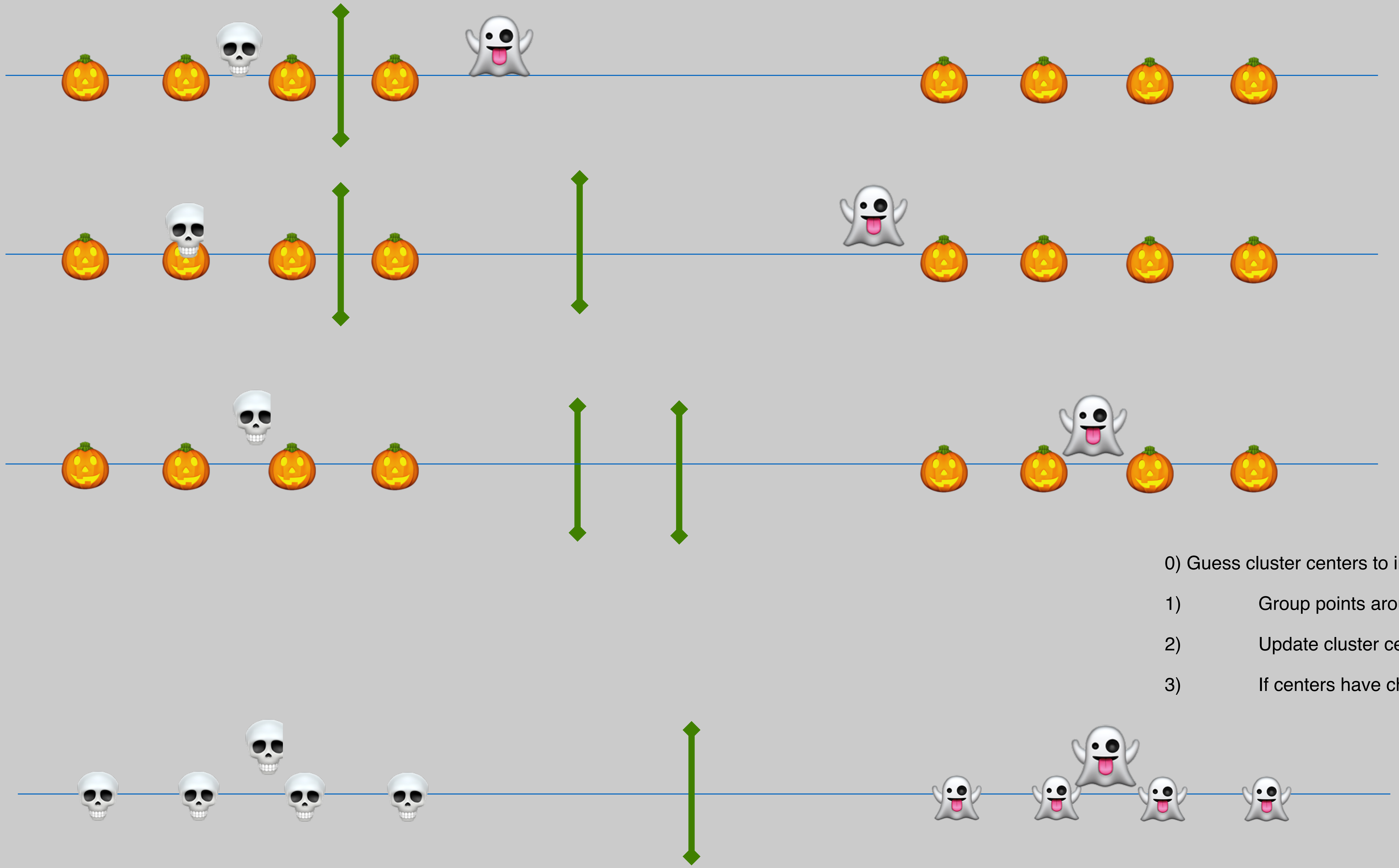
Given: $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m \in \mathrm{R}^n$

Partition them into k << m groups

0) Guess cluster centers to initialize
1) Group points around nearest center
2) Update cluster centers by averaging within group
3) If centers have changed, repeat 1-3

k-means 1D example

$$n = 1, m = 8, k = 2$$



0) Guess cluster centers to initialize

1) Group points around nearest center

2) Update cluster centers by averaging within group

3) If centers have changed, repeat 1-3

General k-means Algorithm

0) Initialize k cluster centers $\vec{m}_1, \vec{m}_2, \cdots, \vec{m}_k$

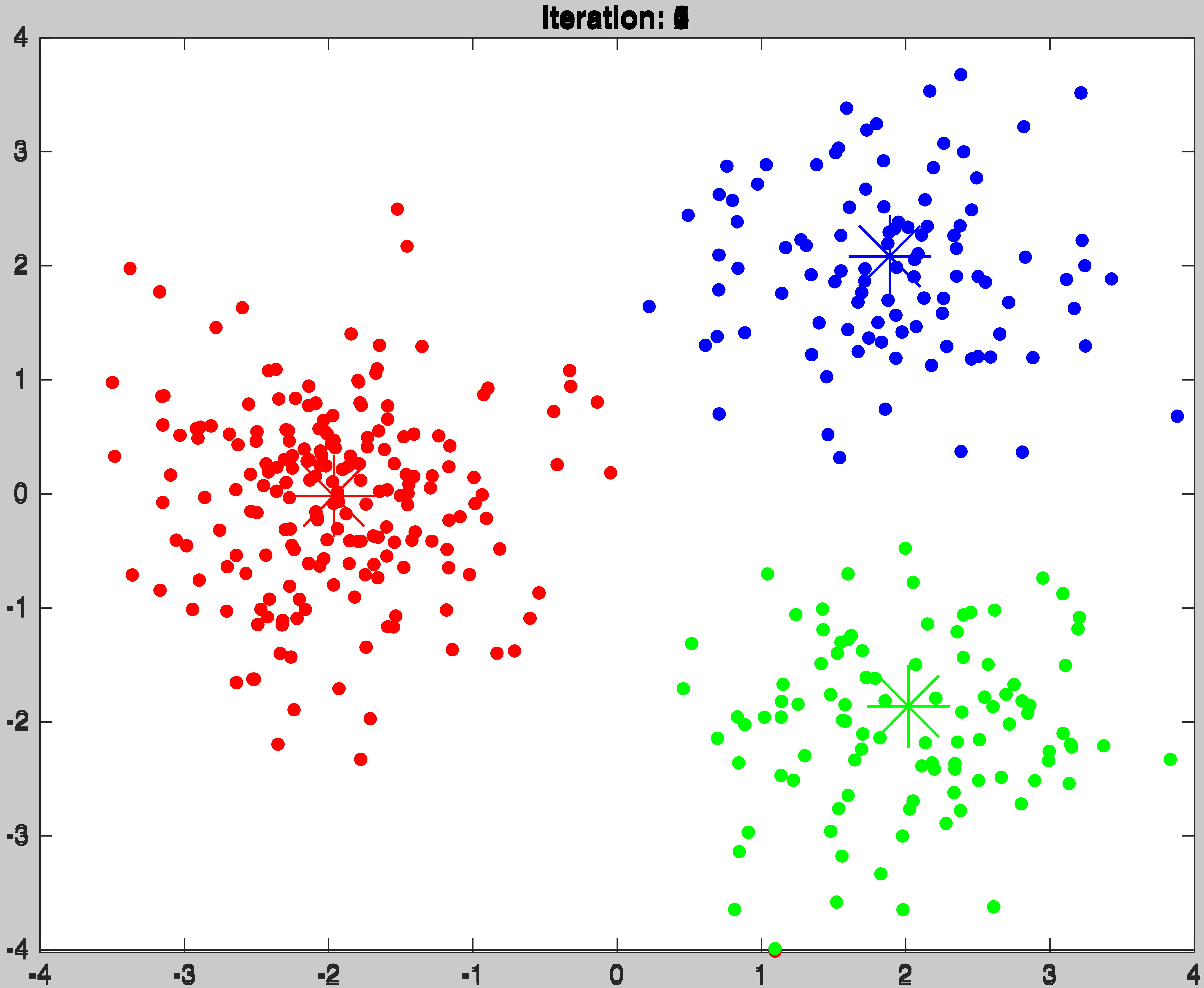
1) Assign points to cluster: point $\vec{x}$ goes to cluster i if,

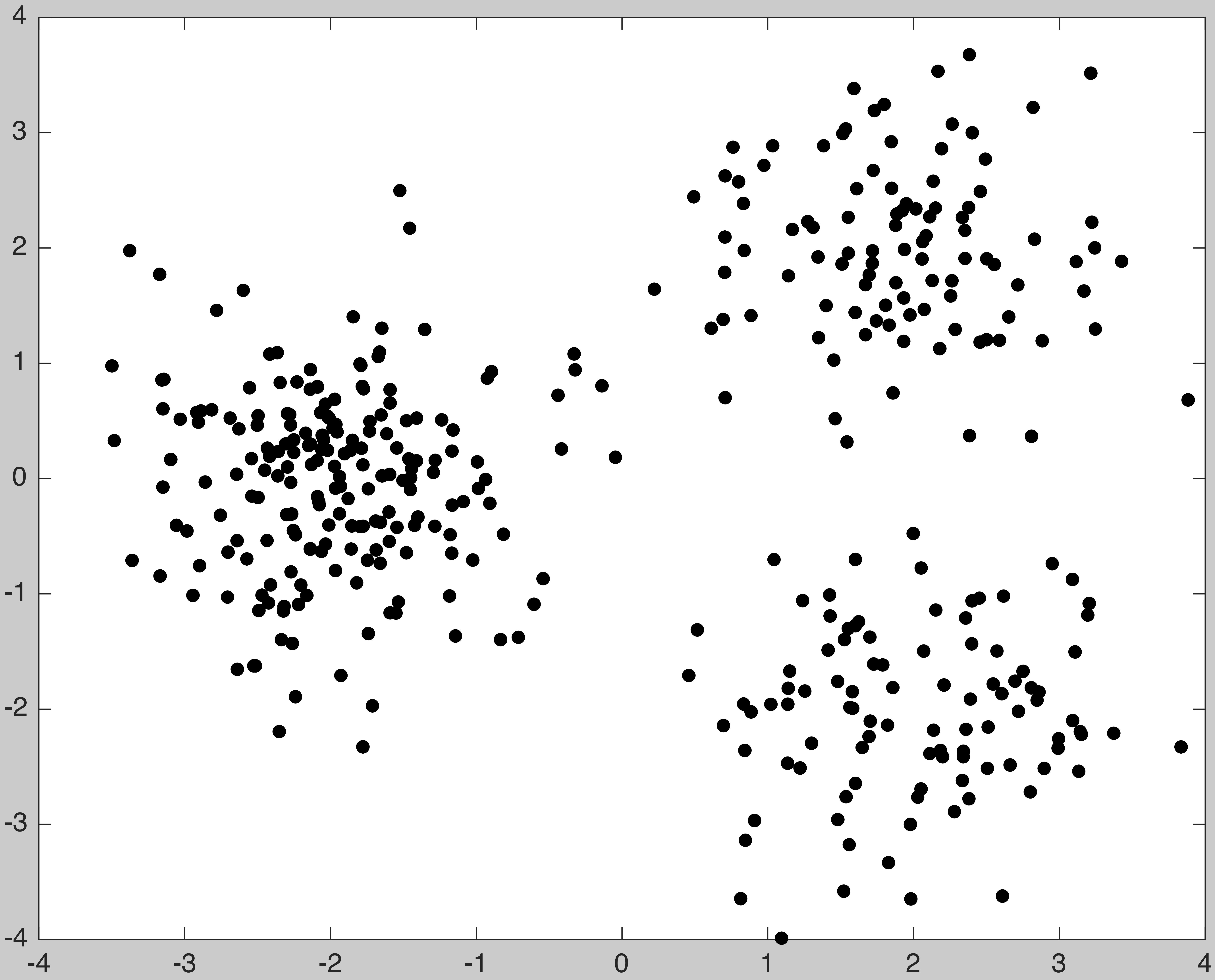
$$||\vec{x} - \vec{m}_i|| < ||\vec{x} - \vec{m}_j|| \quad \forall j \neq i$$

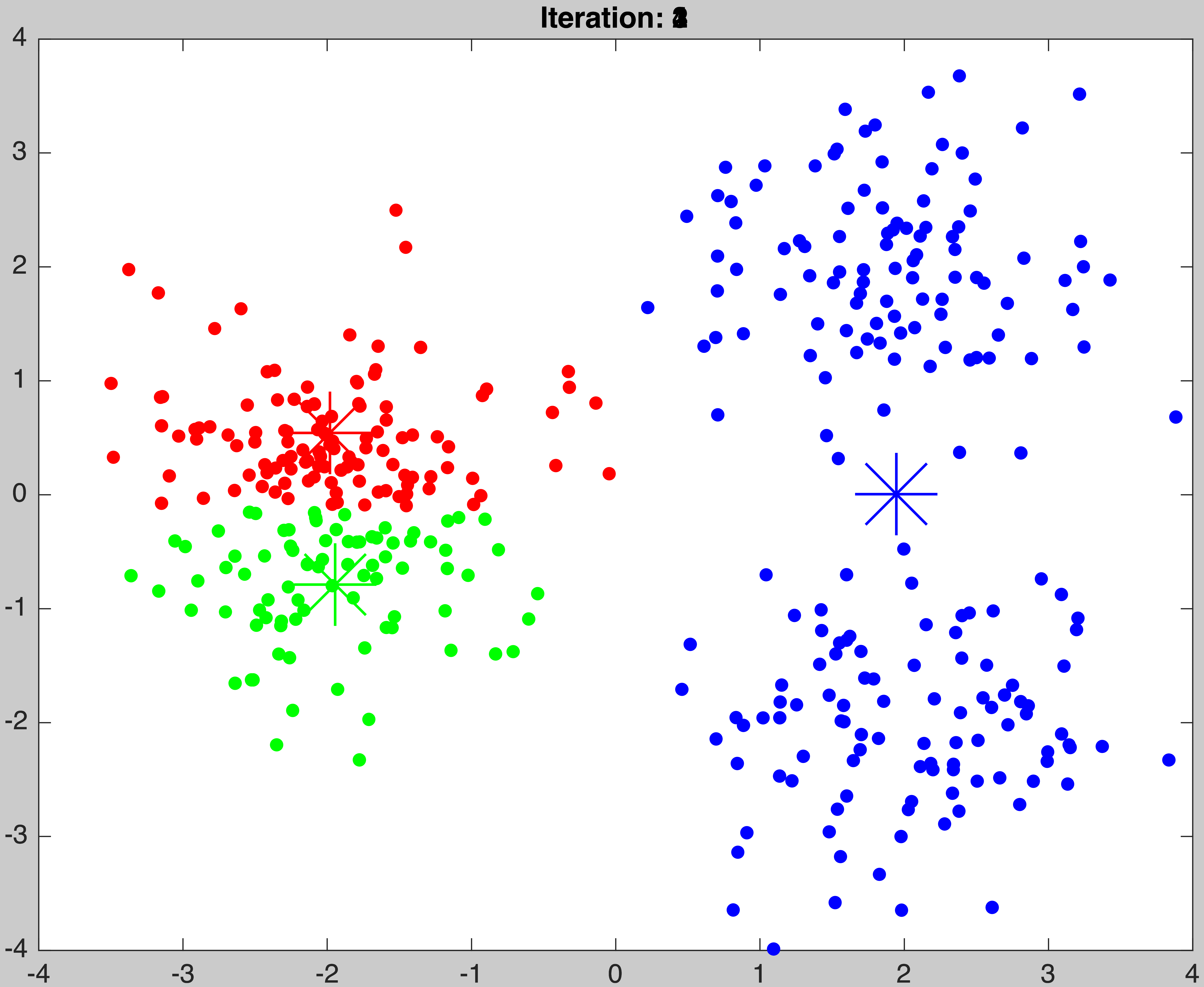
2) Let S_i be the set of samples in cluster i recompute cluster centers:

$$\vec{m}_i = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}$$

3) If any m_i has changed, repeat 1-3


Iteration: 5
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


Iteration:
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

Objective Function

Find the clustering of $\vec{x}_1, \cdots, \vec{x}_m$ into sets $S_1, \cdots, S_k$ which minimizes:

$$D = \sum_{i=1}^{k} \sum_{\vec{x} \in S_i} ||\vec{x} - \mu_i||$$

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} \vec{x}$$

While the algorithm decreases the objective, the objective is non-convex and can be stuck on local mimima.

General problem is N-P Complete

Management of intersections with multi-modal high-resolution data ☆,☆☆

Ajith Muralidharan [1], Samuel Coogan [2], Christopher Flores, Pravin Varaiya *

Sensys Networks, Inc, Berkeley, CA 94710, United States

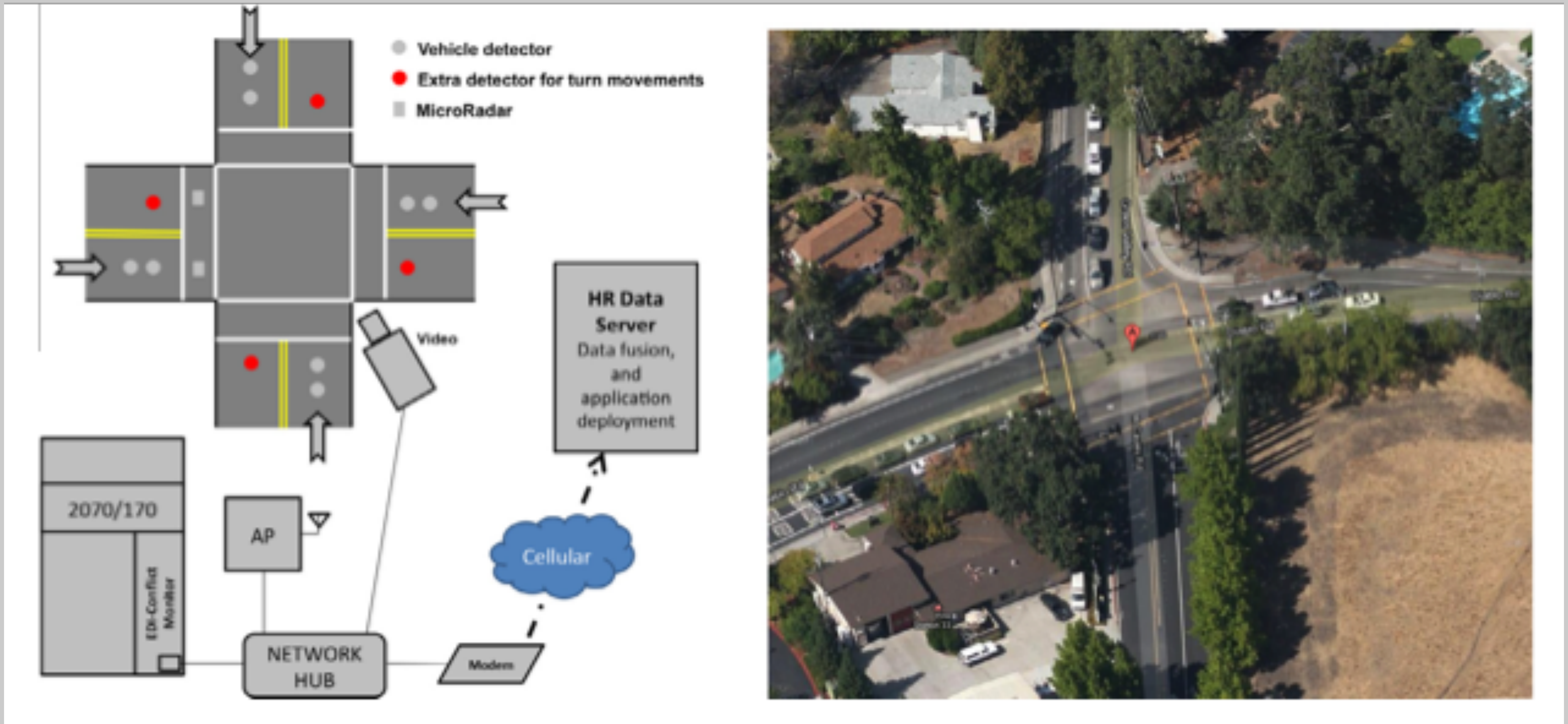
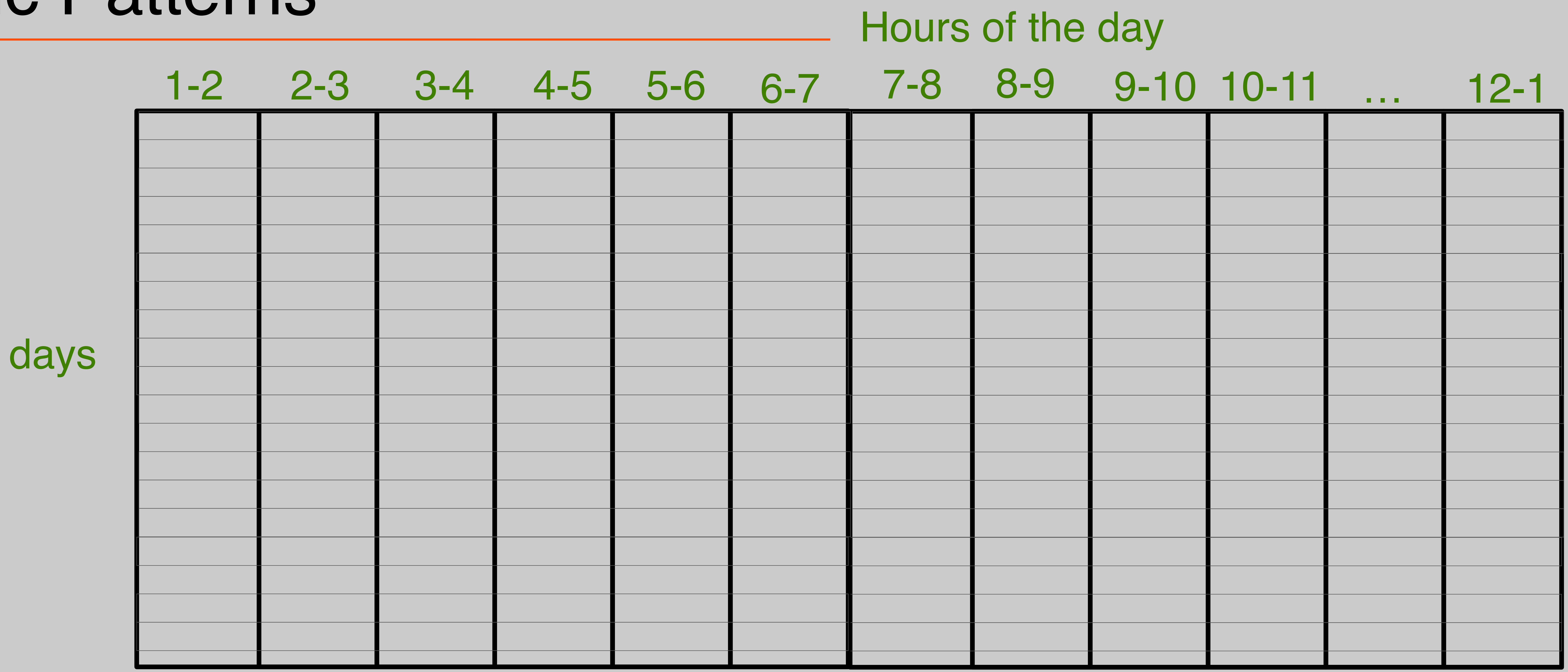


Fig. 1. HR system in Danville, CA.

Traffic Patterns



What would k-means cluster to?

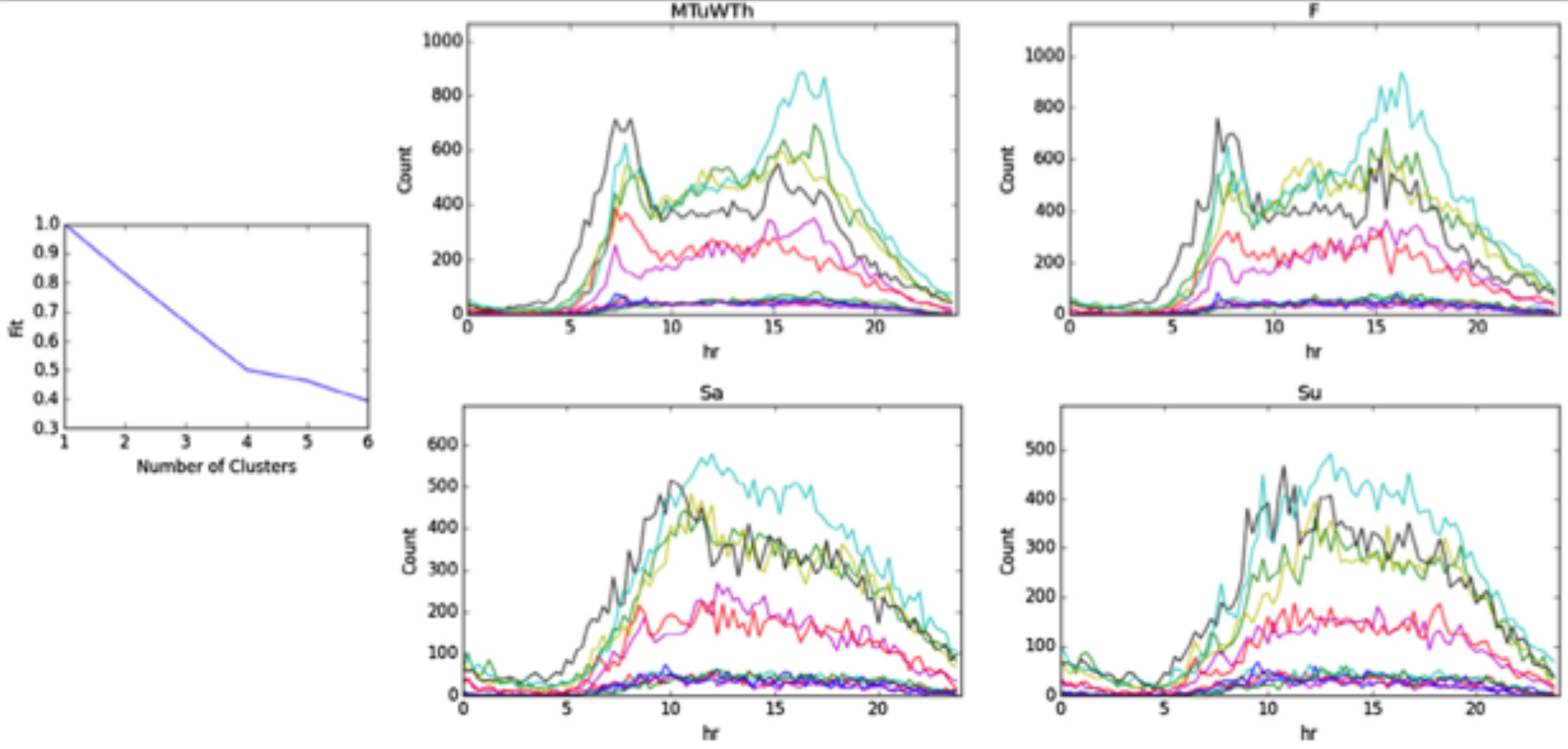


Fig. 5. Clustering of daily data for Dec 2014 to May 2015 in an intersection in Beaufort, SC.

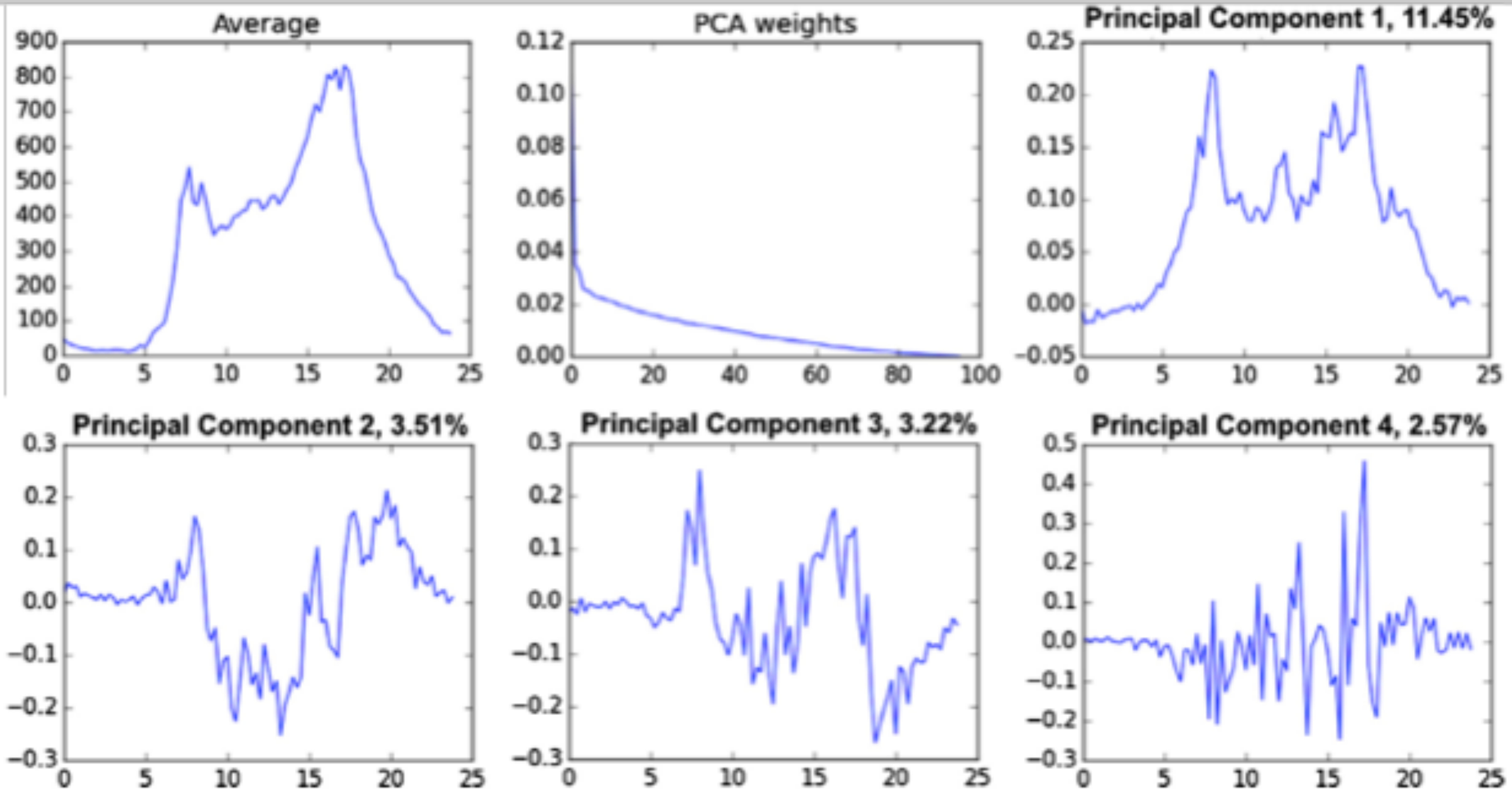


Fig. 8. Four PCA components of the North–South through movement and the average. The *x*-axis is hours.

Symmetric Matrices

We assumed before that,

A^TA has only real eigenvalues, r of them are positive and the rest are zero
A^TA has orthonormal eigenvectors (to be proven next time)

For symmetric matrices: $Q^T = Q$

$$(AB)^T = B^T A^T$$

$$(A^T A)^T = A^T A$$

$$(AA^T)^T = AA^T$$

Properties of Symmetric Matrices

1) A real-valued symmetric matrix has real eigenvalues and eigenvectors

$$Qx = \lambda x \qquad \lambda = a + ib \qquad \overline{\lambda} = a - ib$$

Somehow we need to use the symmetric and real-ness property of Q to show that b==0

$$Q\overline{x} = \overline{\lambda}\overline{x}$$

$$\overline{x}^T Q = \overline{\lambda}\overline{x}^T$$

$$\overline{x}^T Q x = \overline{\lambda}\overline{x}^T x \qquad \overline{x}^T Q x = \lambda \overline{x}^T x$$

$$\overline{\lambda}\overline{x}^T x = \lambda \overline{x}^T x \quad \Rightarrow \lambda = \overline{\lambda} \Rightarrow \lambda \in \mathrm{R}$$

Properties of Symmetric Matrices

$$Qx = \lambda x$$

$$\underbrace{(Q - \lambda I)}_{\text{real}}x = 0$$

So x is real as well

✅A real-valued symmetric matrix has real eigenvalues and eigenvectors

Properties of Symmetric Matrices

2) Eigenvectors of a symmetrix matrix can be chosen to be orthonormal

Choose two distinct eigenvalues and vectors $\lambda_1 \neq \lambda_2$

$$Qx_1 = \lambda_1 x_1 \qquad\qquad Qx_2 = \lambda_2 x_2$$

$$x_2^T Qx_1 = \lambda_1 x_2^T x_1 \qquad\qquad x_1^T Qx_2 = \lambda_2 x_1^T x_2$$

$$(\lambda_1 - \lambda_2) x_2^T x_1 = 0$$

$$\lambda_1 \neq \lambda 2 \Rightarrow x_2^T x_1 = 0$$

✅ Eigenvectors of a symmetrix matrix can be chosen to be orthonormal

Positiveness of Eigenvalues

3) If Q can be written as Q = R^TR for real R, then Q is positive semidefinite – eigenvalues greater of equal to zero

$$Qx = \lambda x$$

$$R^T R x = \lambda x$$

$$x^T R^T R x = \lambda x^T x$$

$$(Rx)^T (Rx) = \lambda x^T x$$

$$||Rx||^2 = \lambda ||x||^2 \quad \Rightarrow \lambda \geq 0$$

✅ If Q can be written as Q = R^TR for real R, then Q is positive semidefinite – eigenvalues greater of equal to zero