EE16B
Designing Information Devices and Systems II

Lecture 10B

PCA

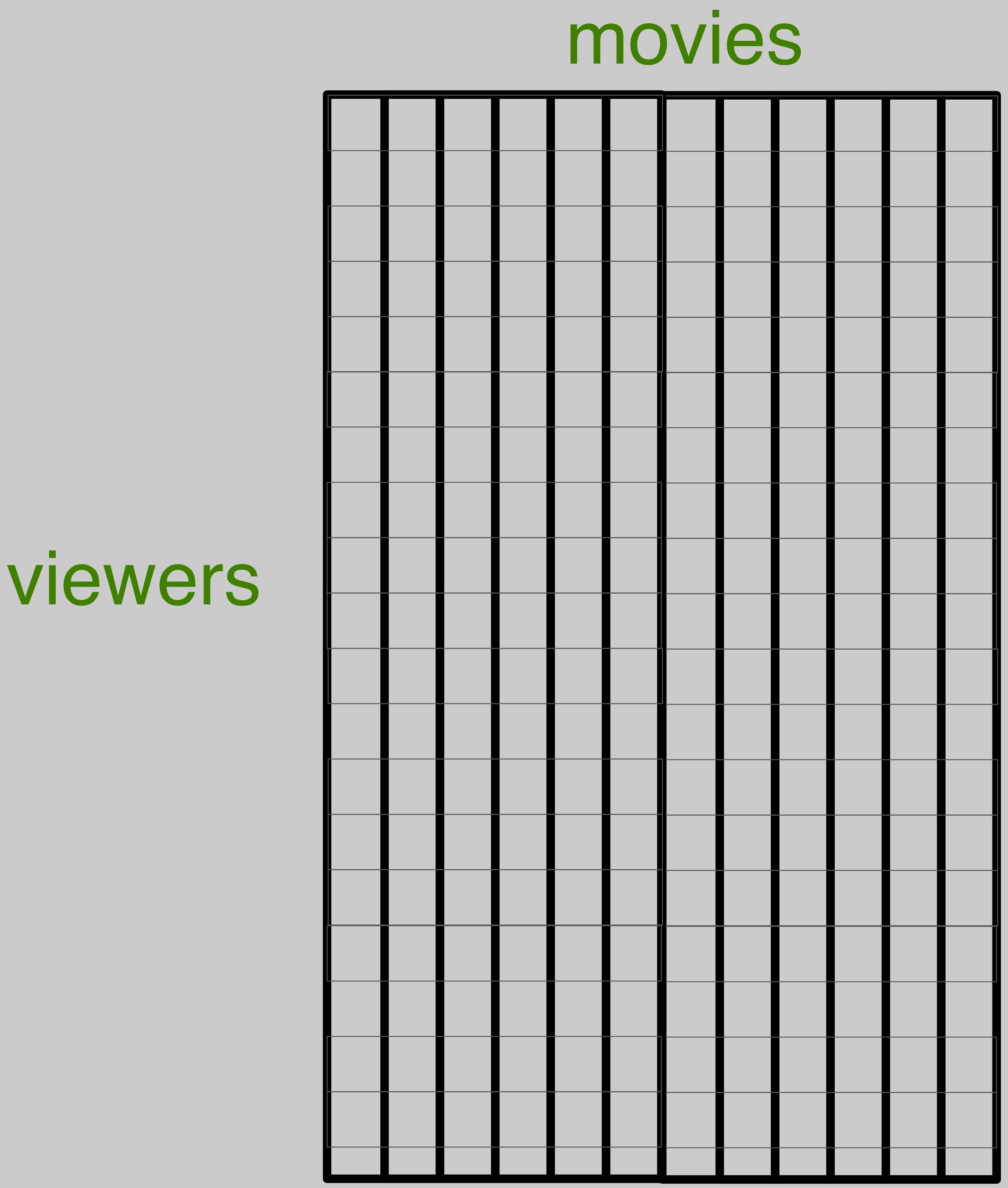
Intro

- –Last Time:
  - – Uniqueness and Geometry of SVD
  - – Finished proofs
  - – Started PCA
- –Today:
  - – Continue PCA
  - –Examples of PCA
  - – K-means (maybe)

Principal Component Analysis

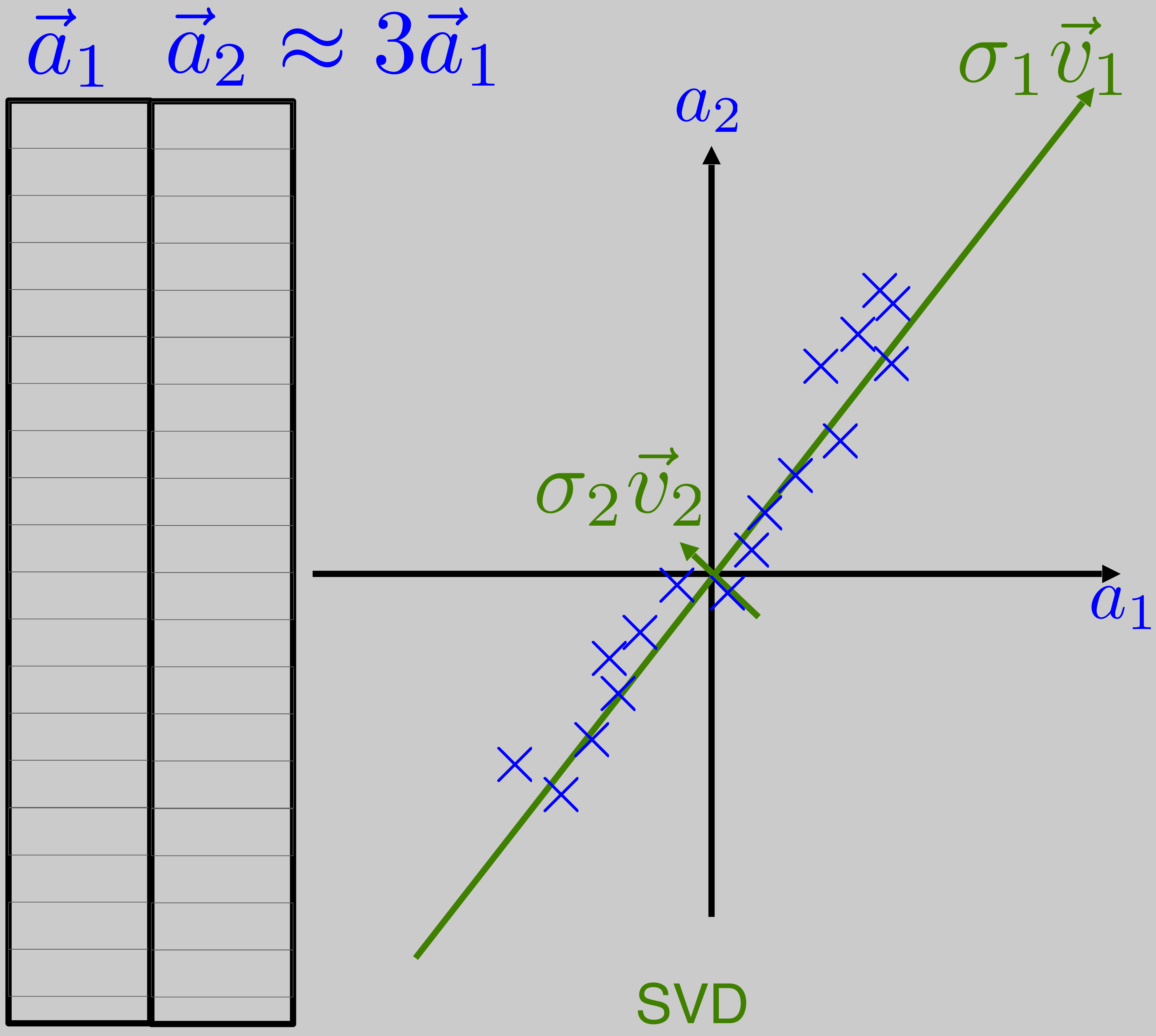
Application of the SVD to datasets to learn features

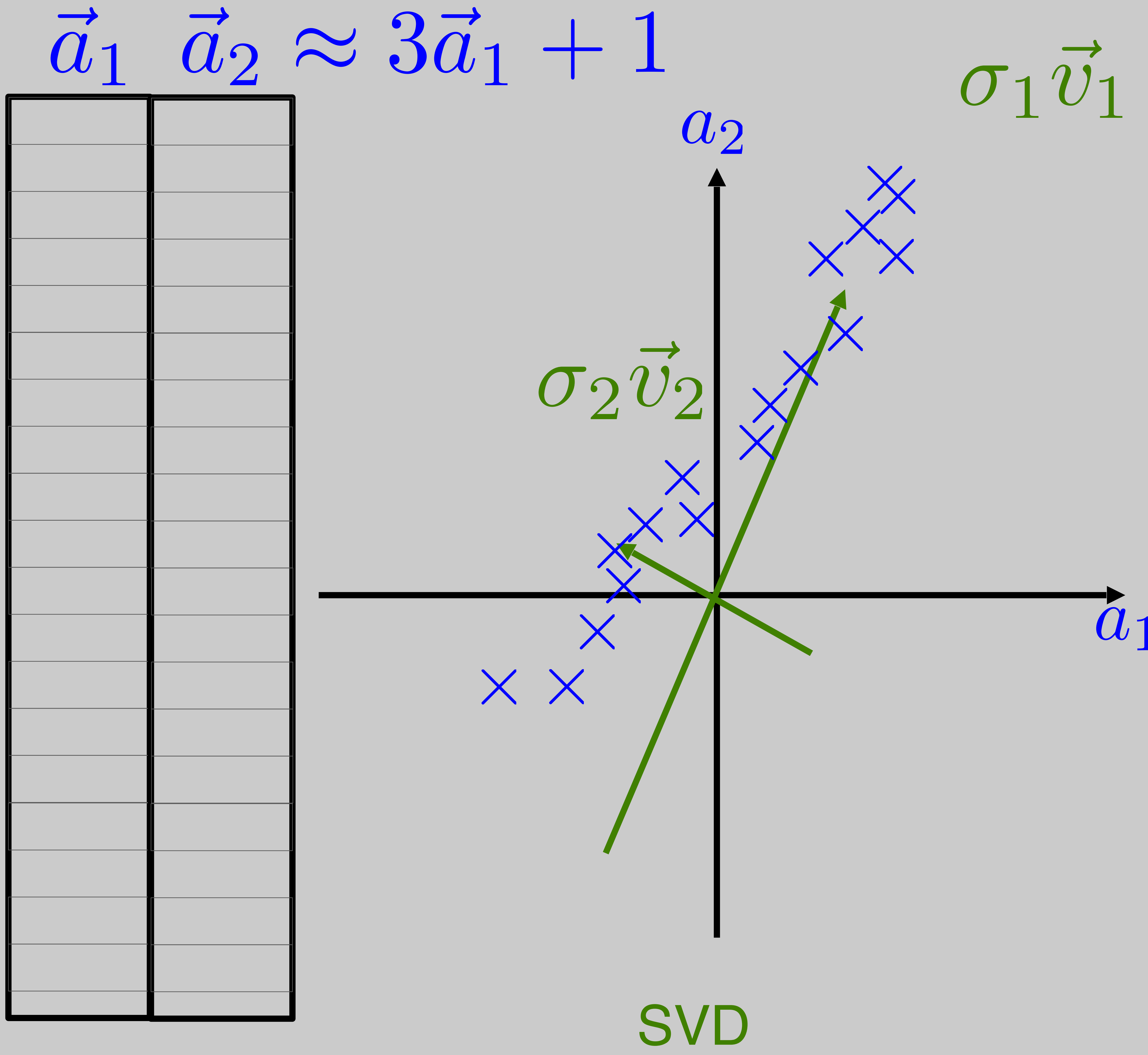
PCA is a tool in statistics and machine learning, which can be computed using SVD



Example -- PCA

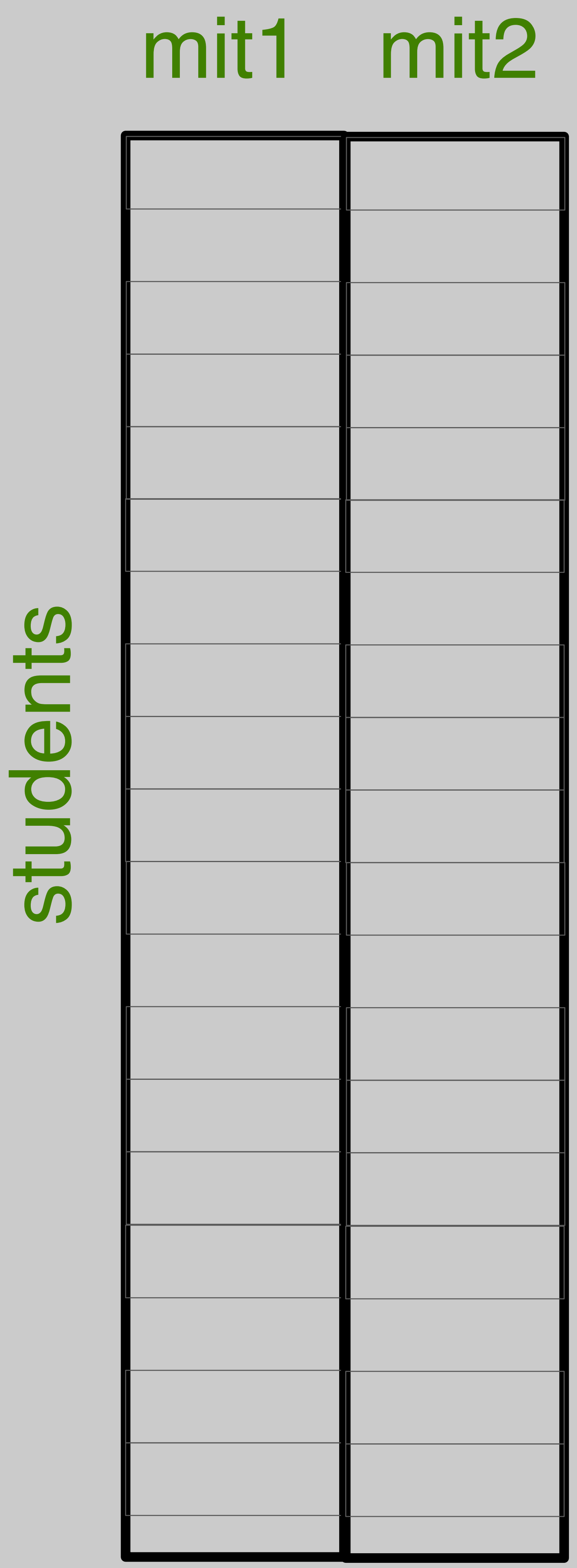
Consider data s.t.

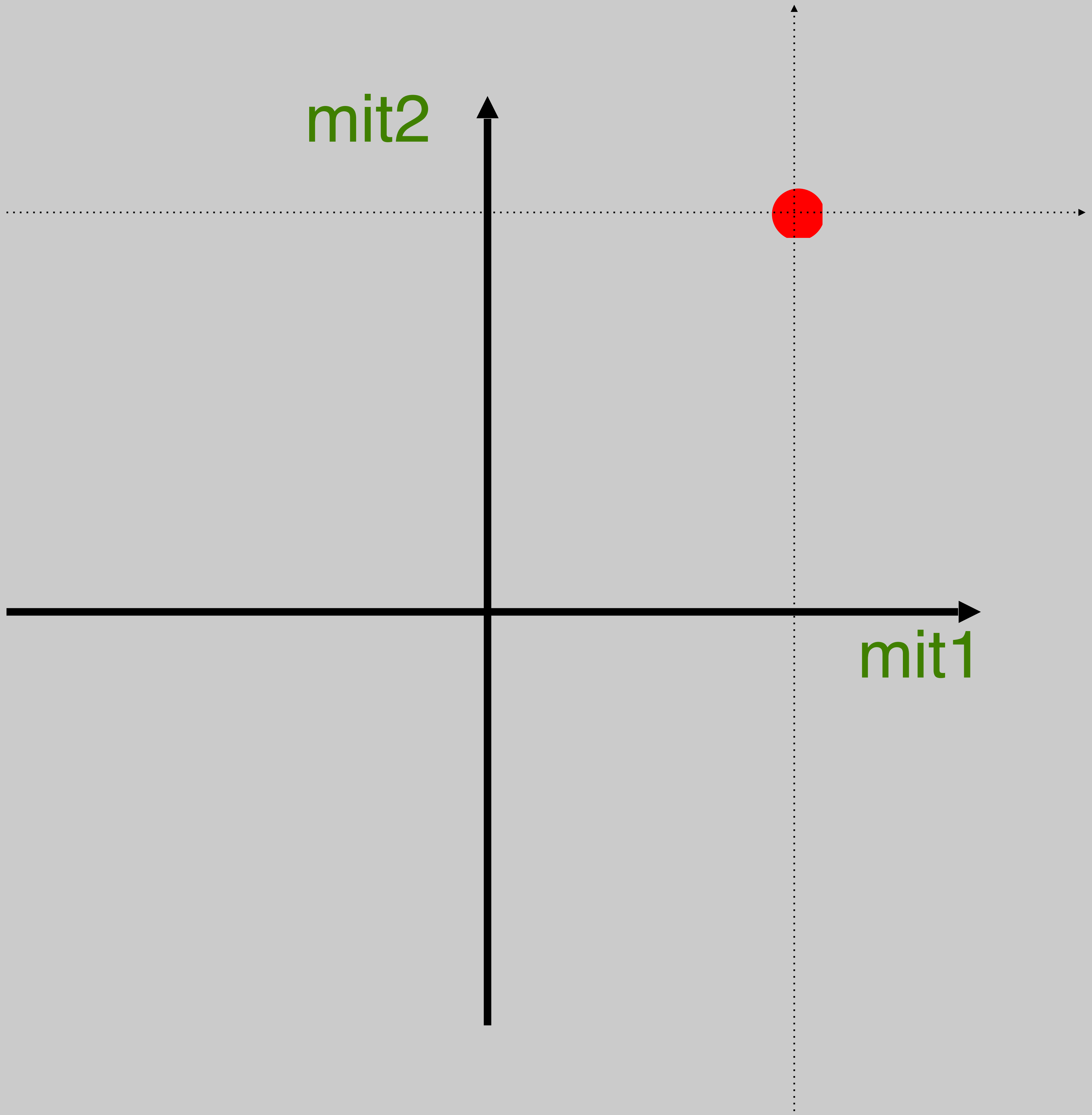




Example -- PCA

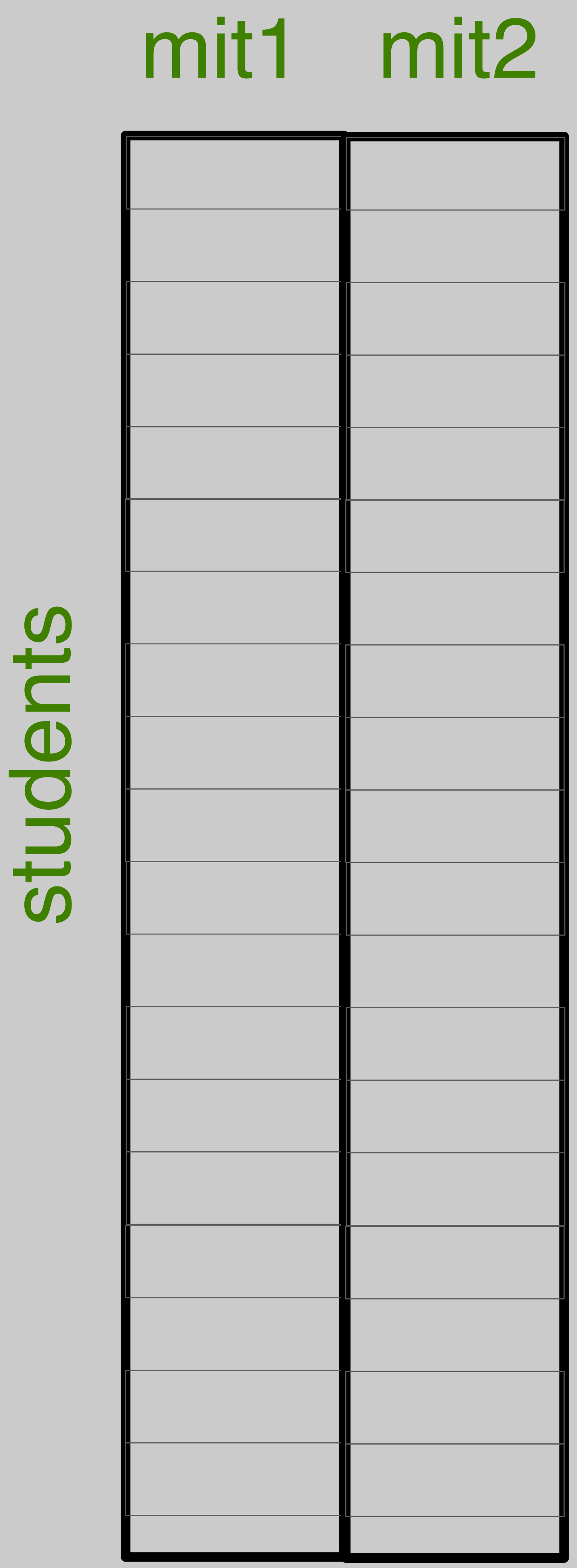
Consider miterm data

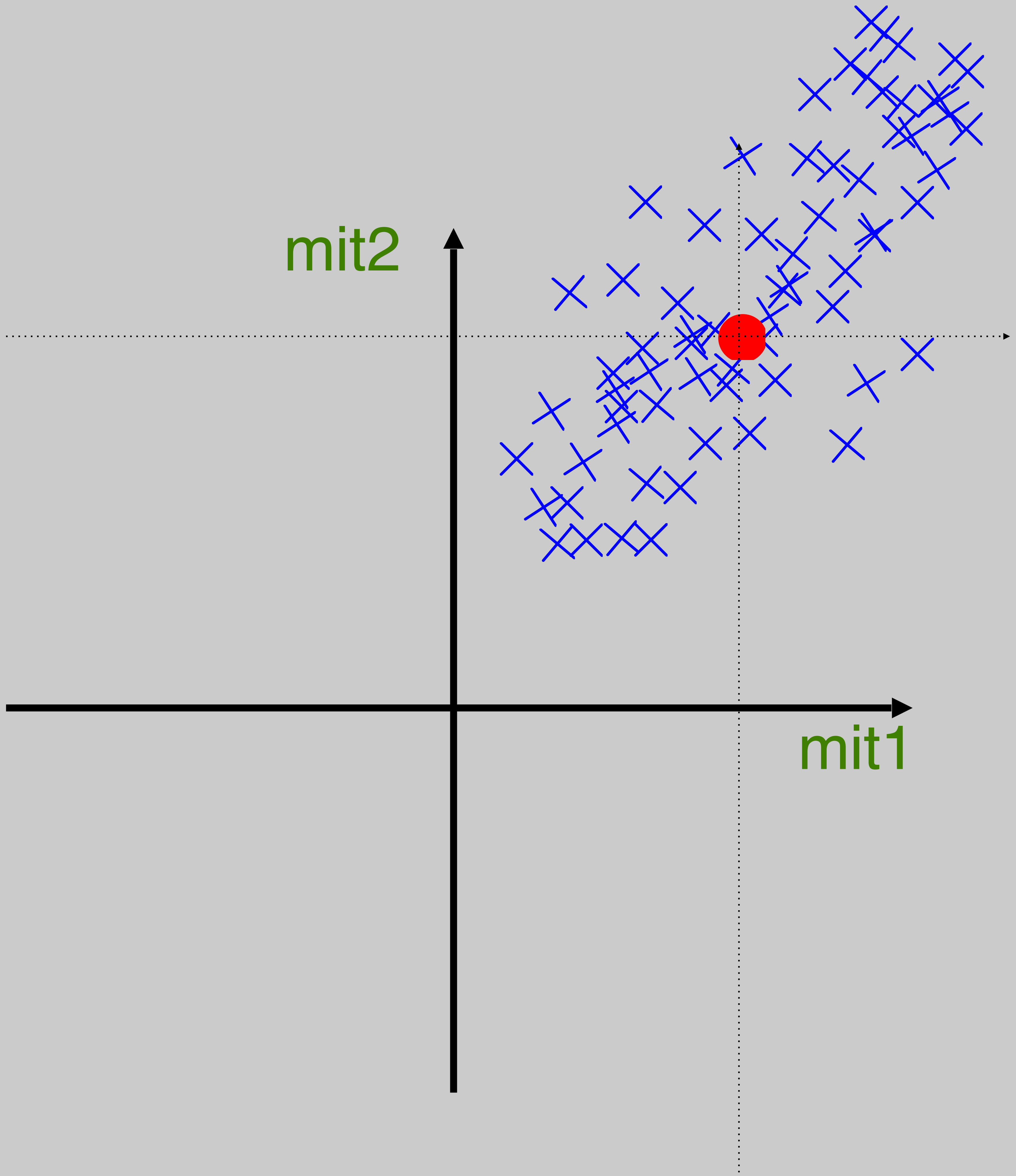




Example -- PCA

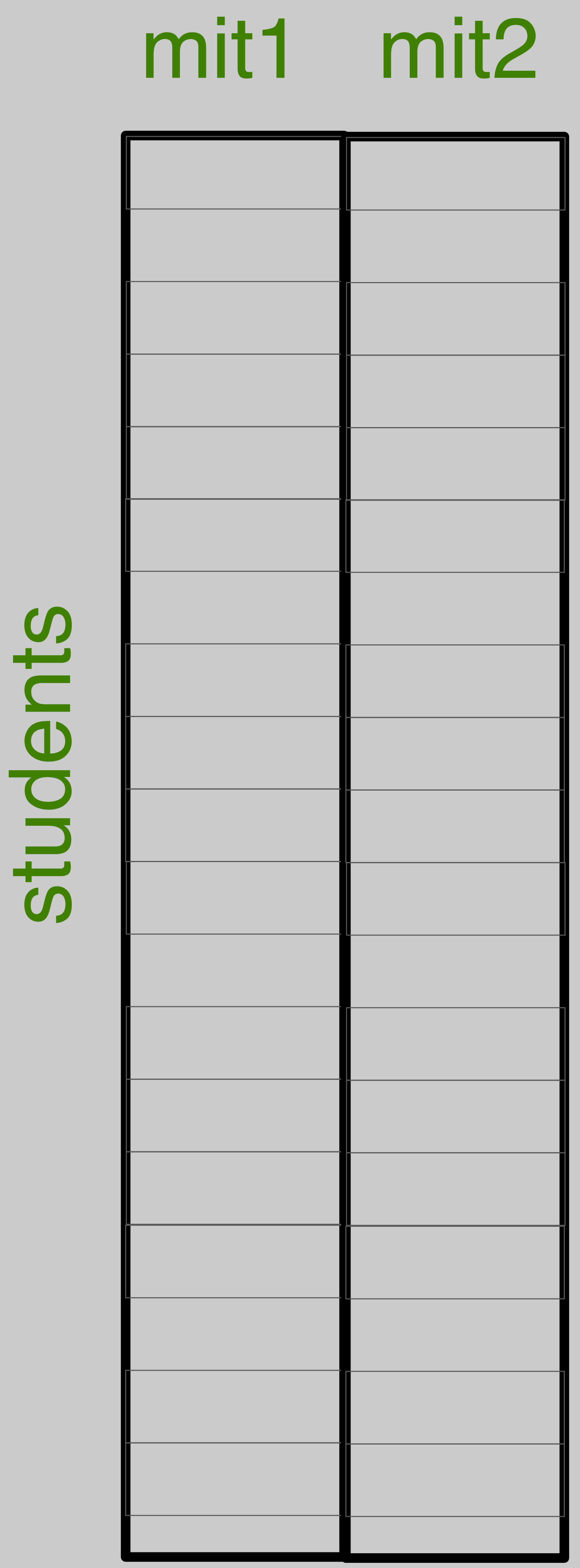
Consider miterm data

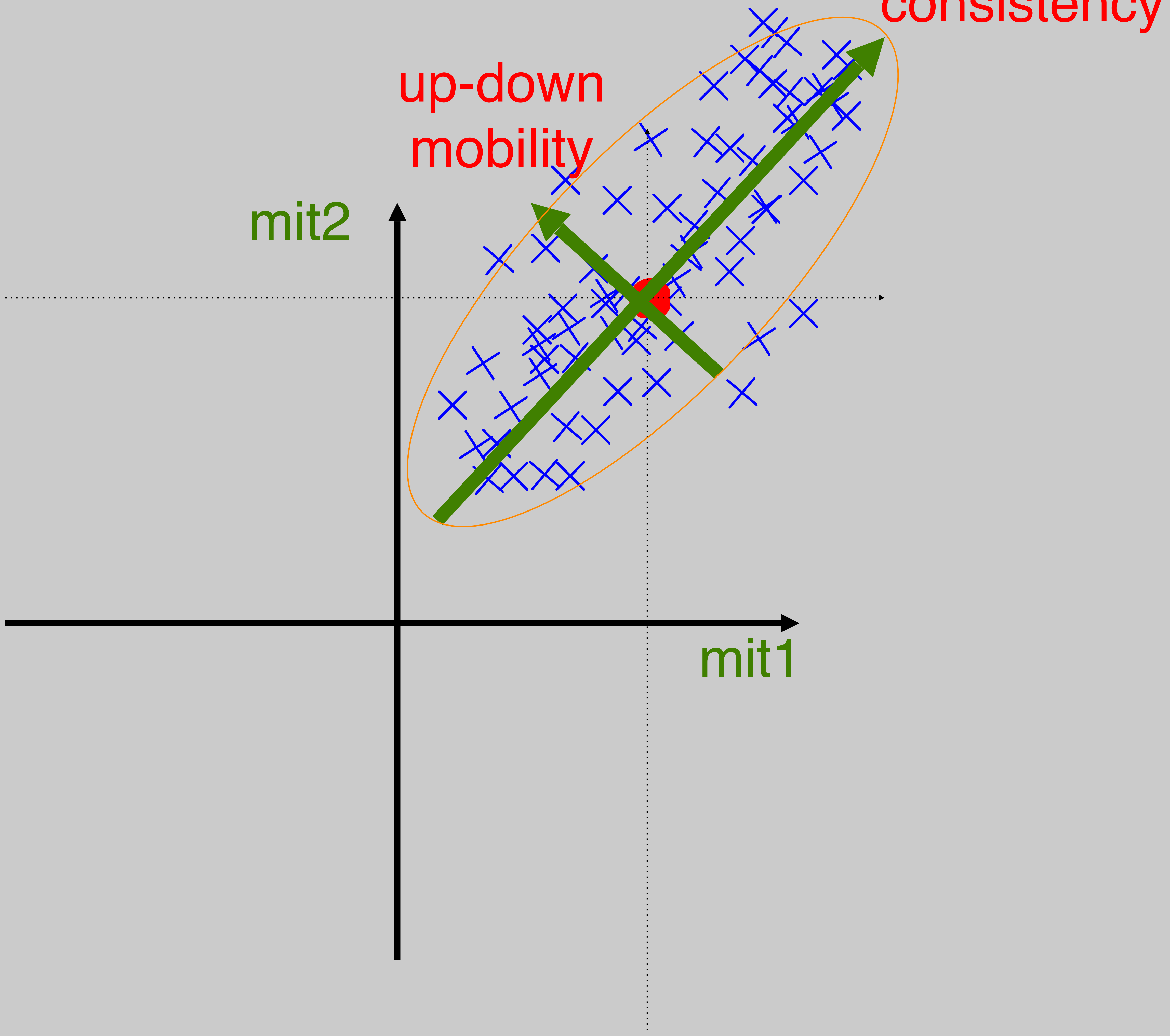




Example -- PCA

Consider miterm data



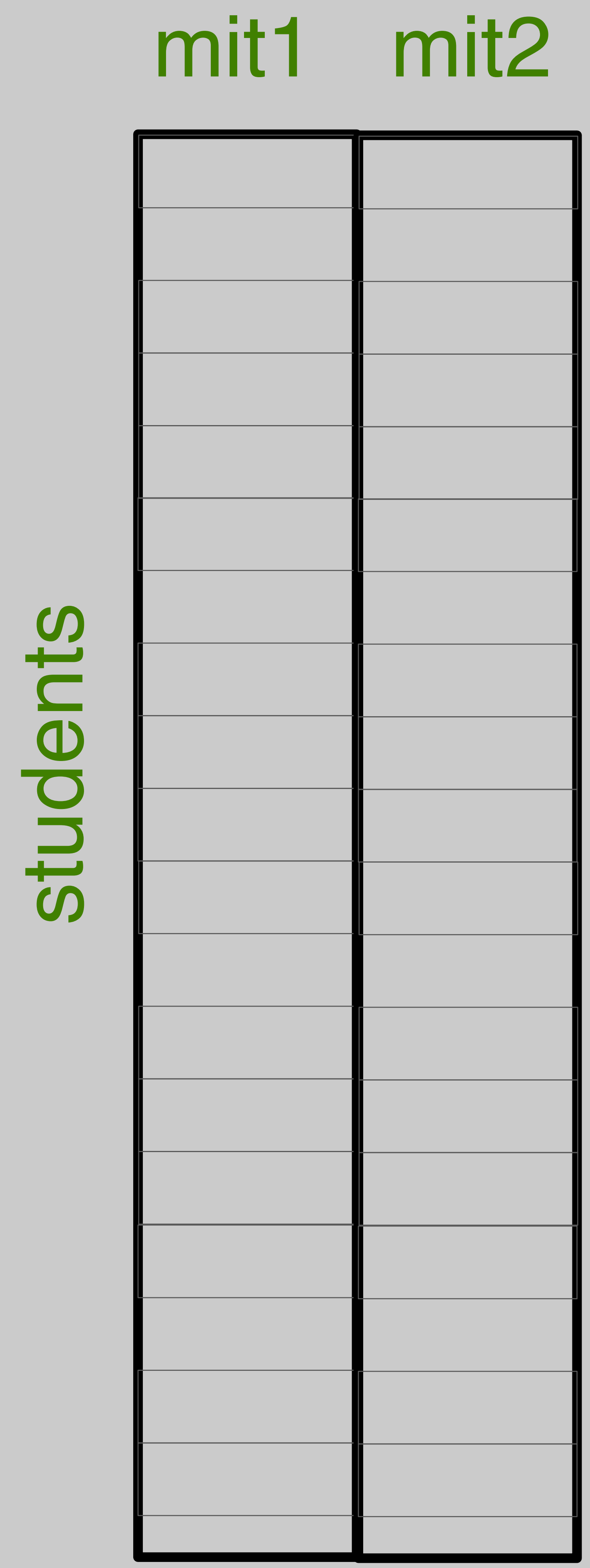


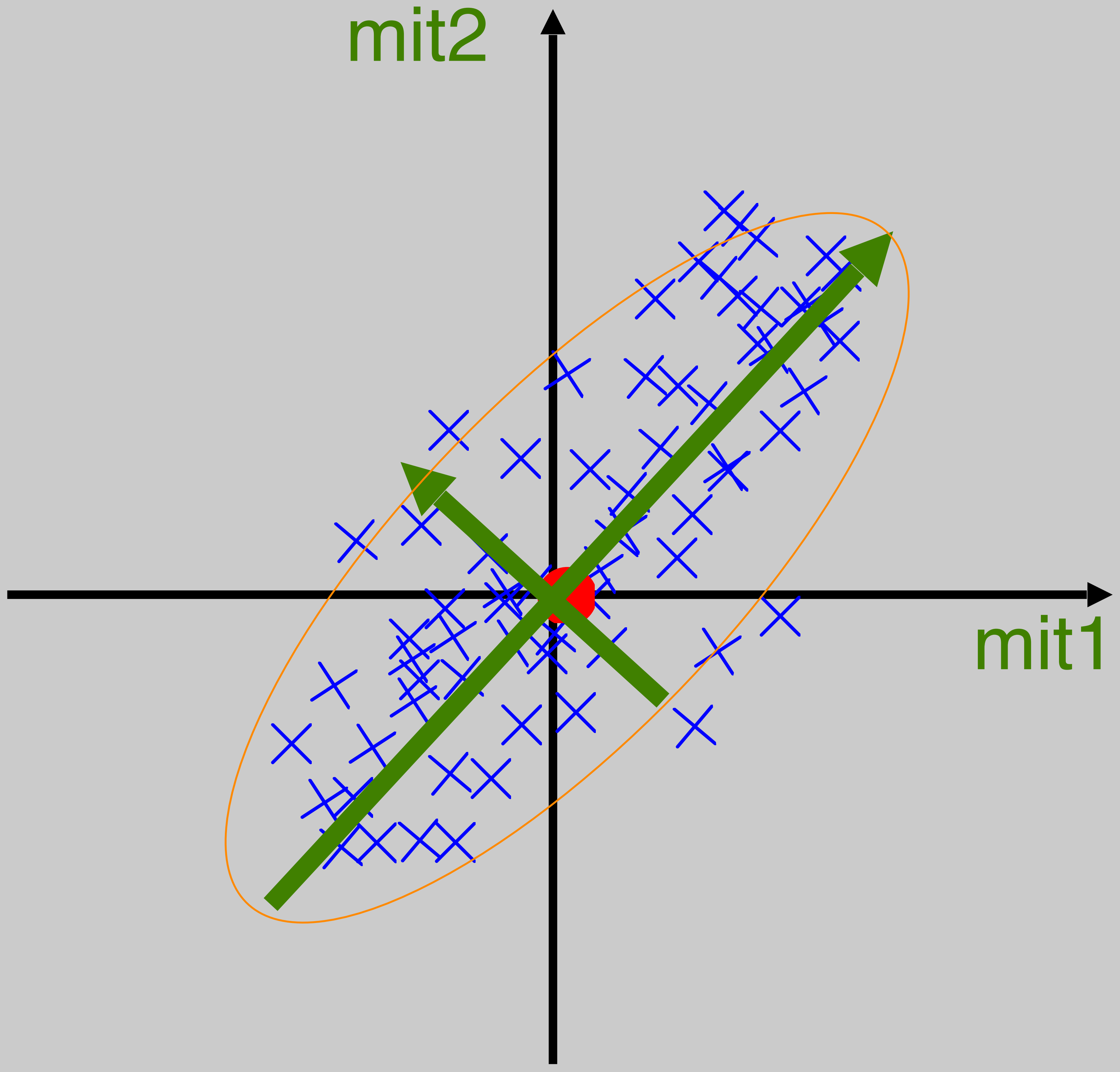
PCA Procedure

Remove averages from column of A

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

$\vec{v_i}$ are principal components!





A^TA as sample covariance matrix

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

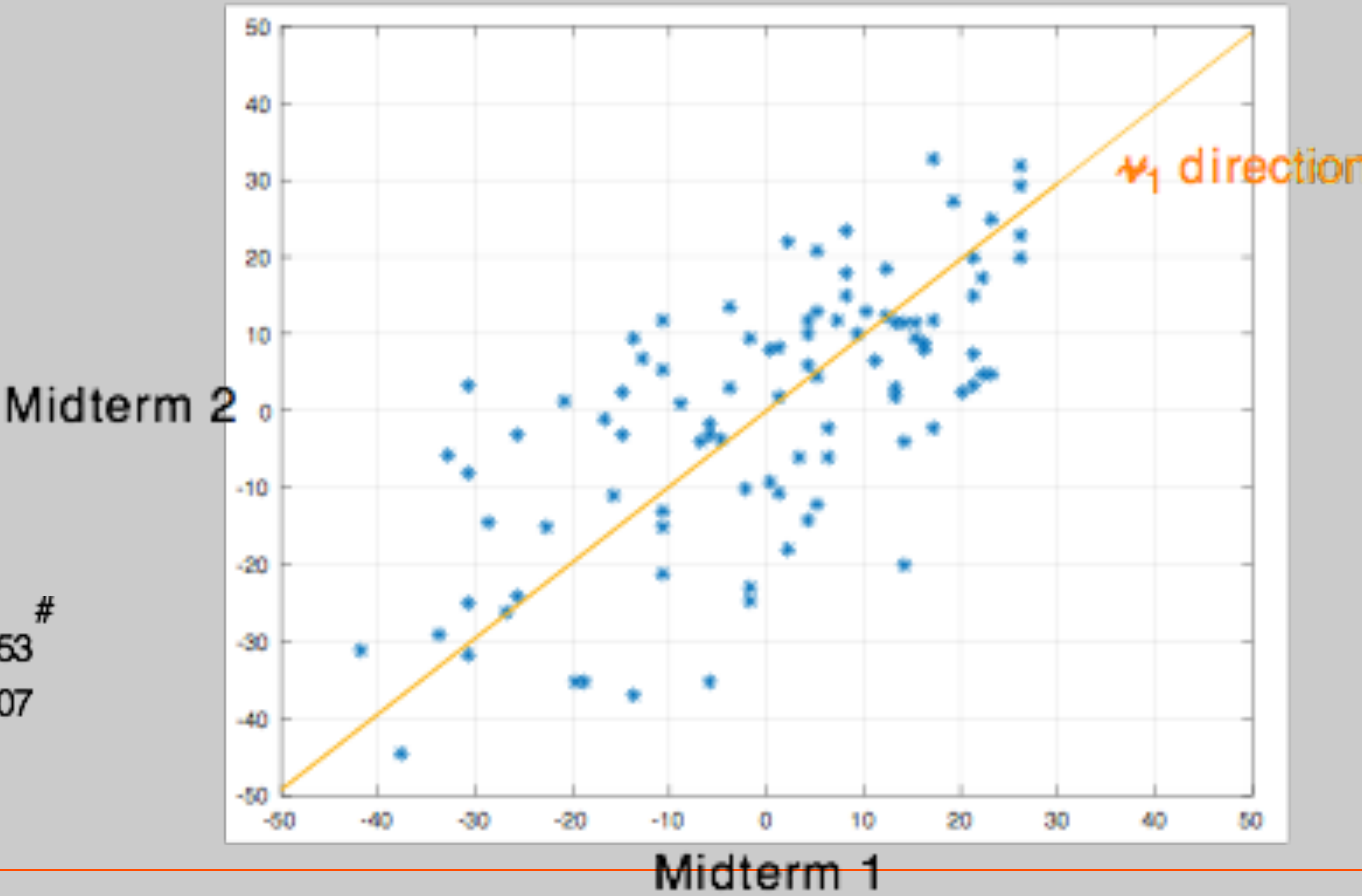
$$\tilde{A}^T\tilde{A} = (\vec{a} - a_\mu\vec{1})^T(\vec{a} - a_\mu\vec{1})$$

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

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

Sample
Variance!

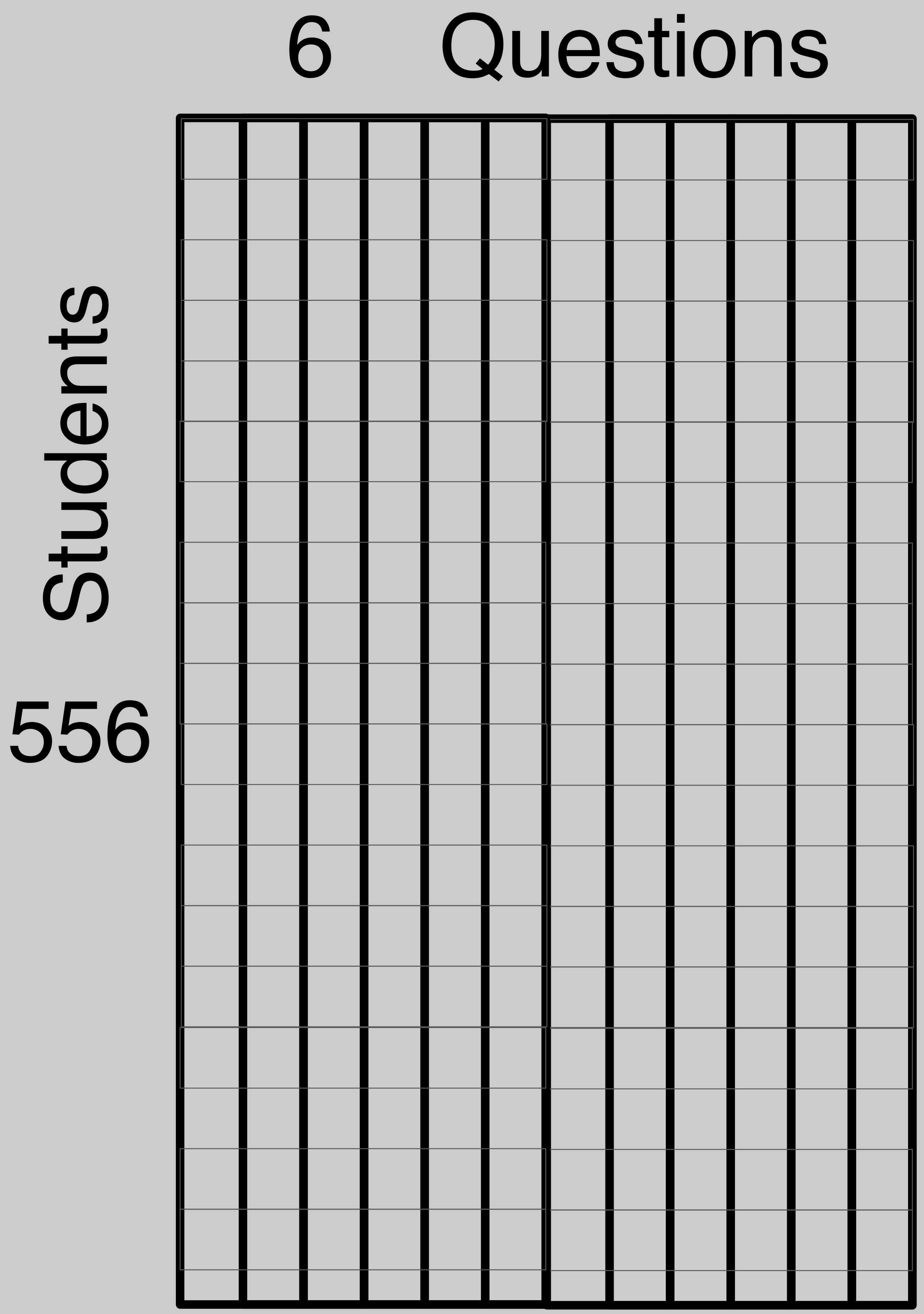
Example midterm



$$\frac{1}{93}A^TA = {}^{\text{"}}\begin{matrix} 297.69 & 202.53 \\ 202.53 & 292.07 \end{matrix}^{\#}$$

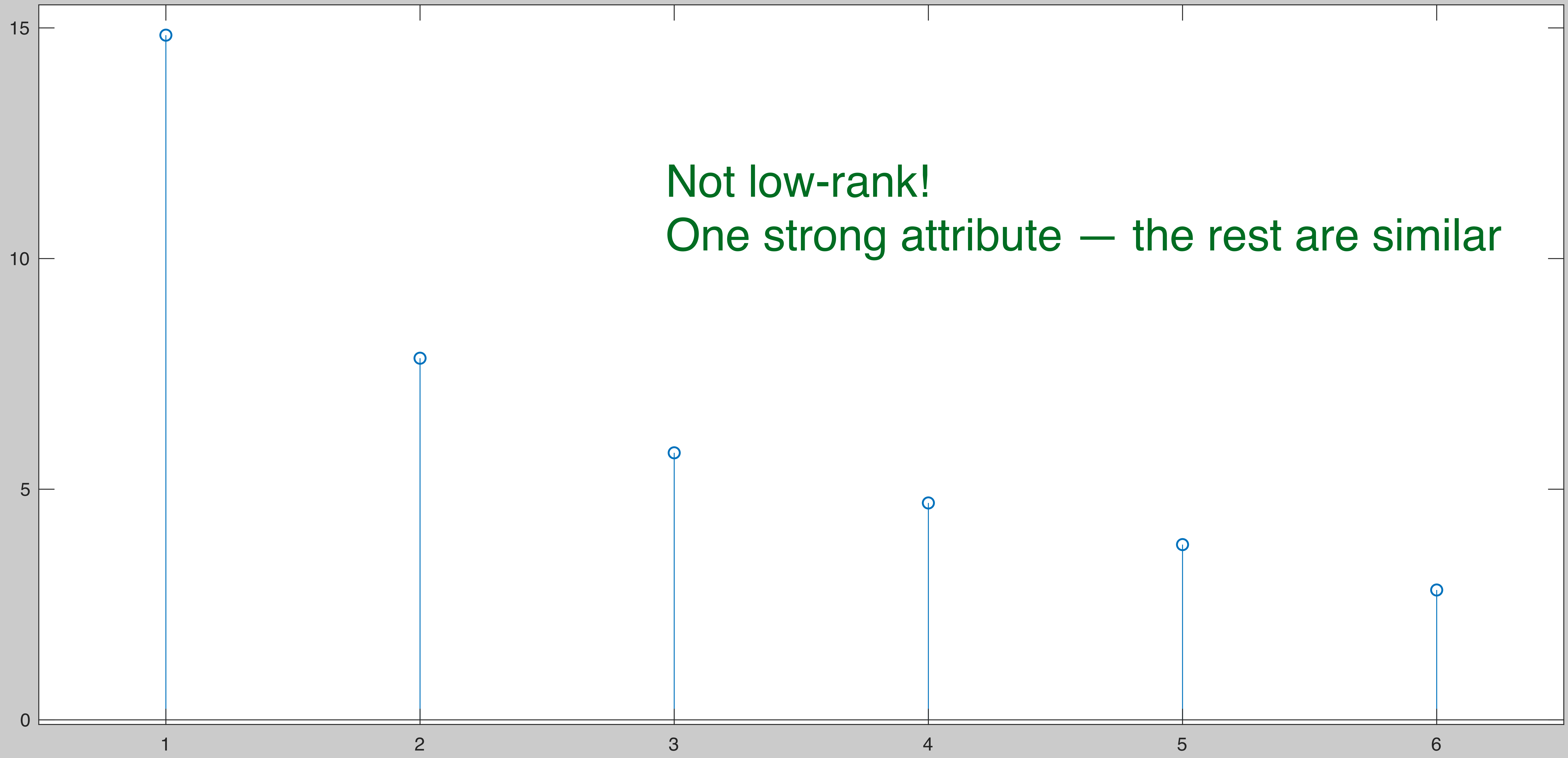
Mid Semester Survey Results

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH



Data Science

Singular values

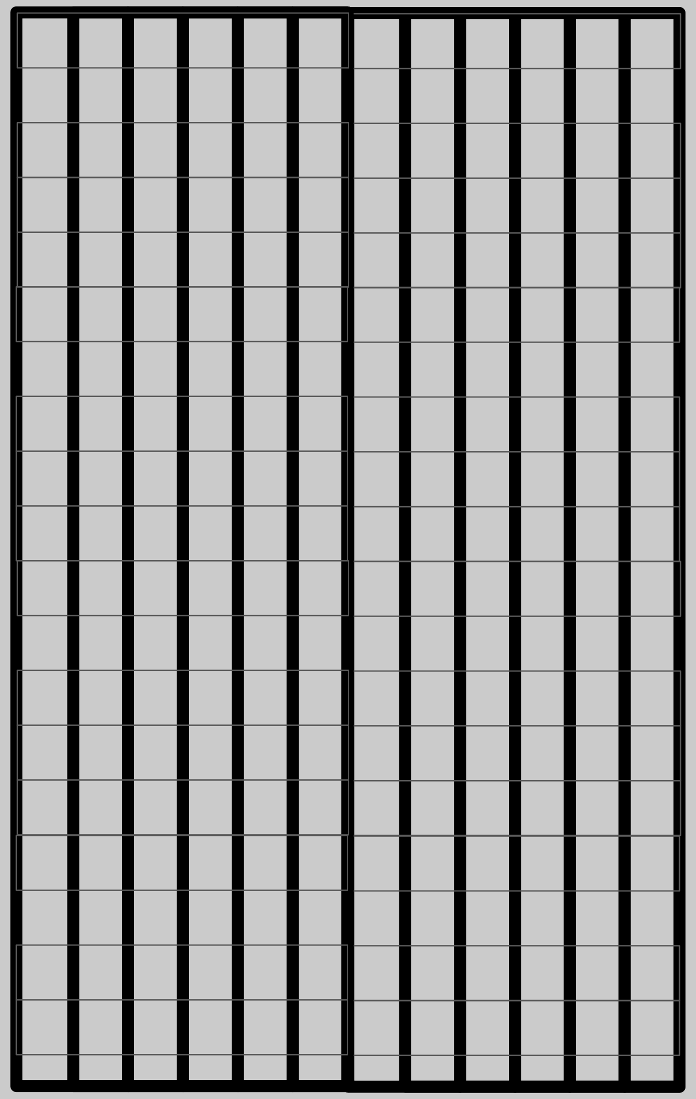


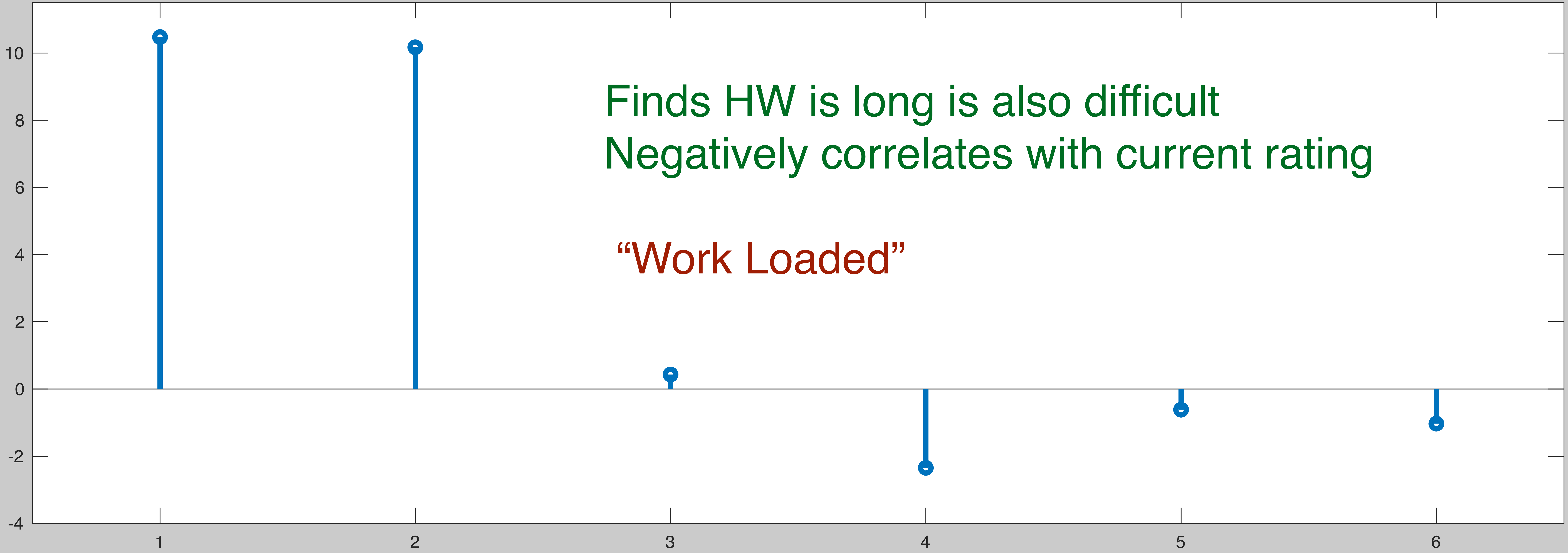
Not low-rank!
One strong attribute — the rest are similar

Data Science

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

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH





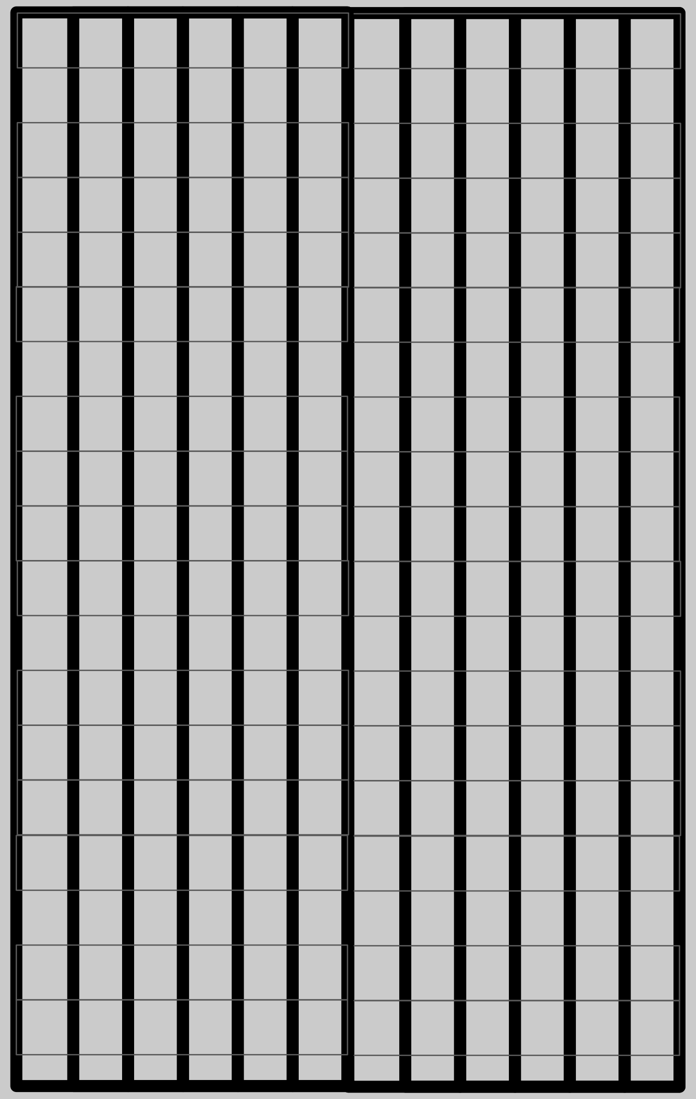
Finds HW is long is also difficult
Negatively correlates with current rating

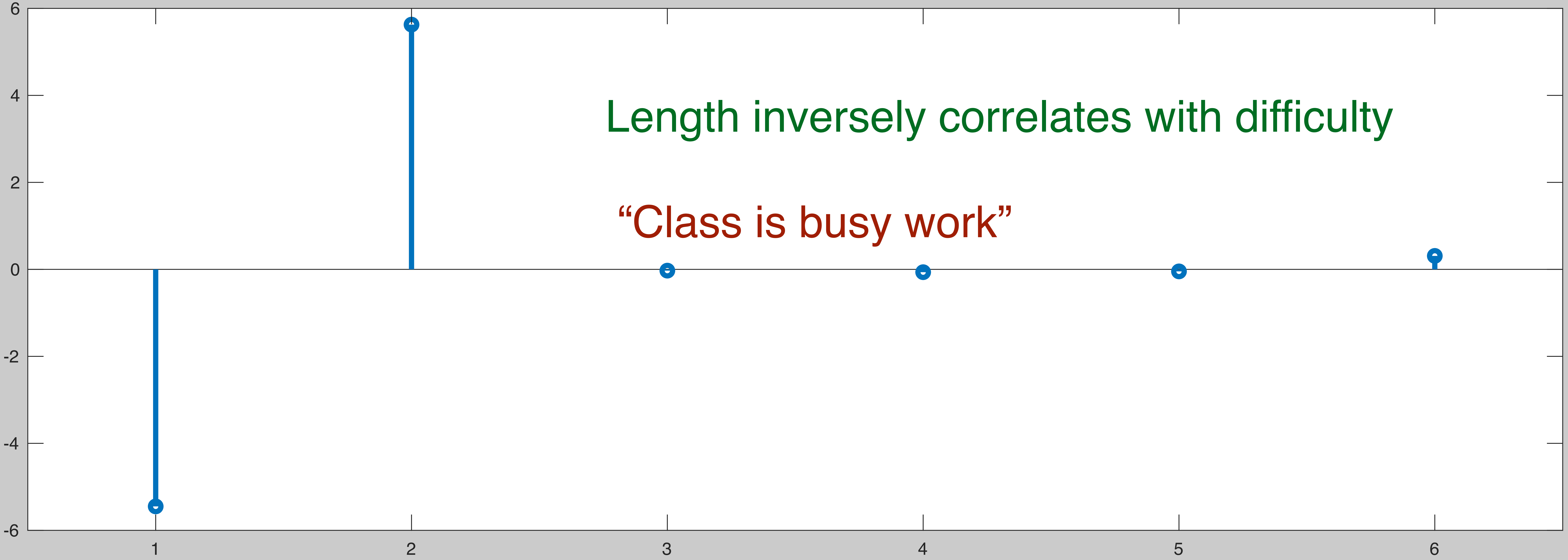
“Work Loaded”

Data Science

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

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH





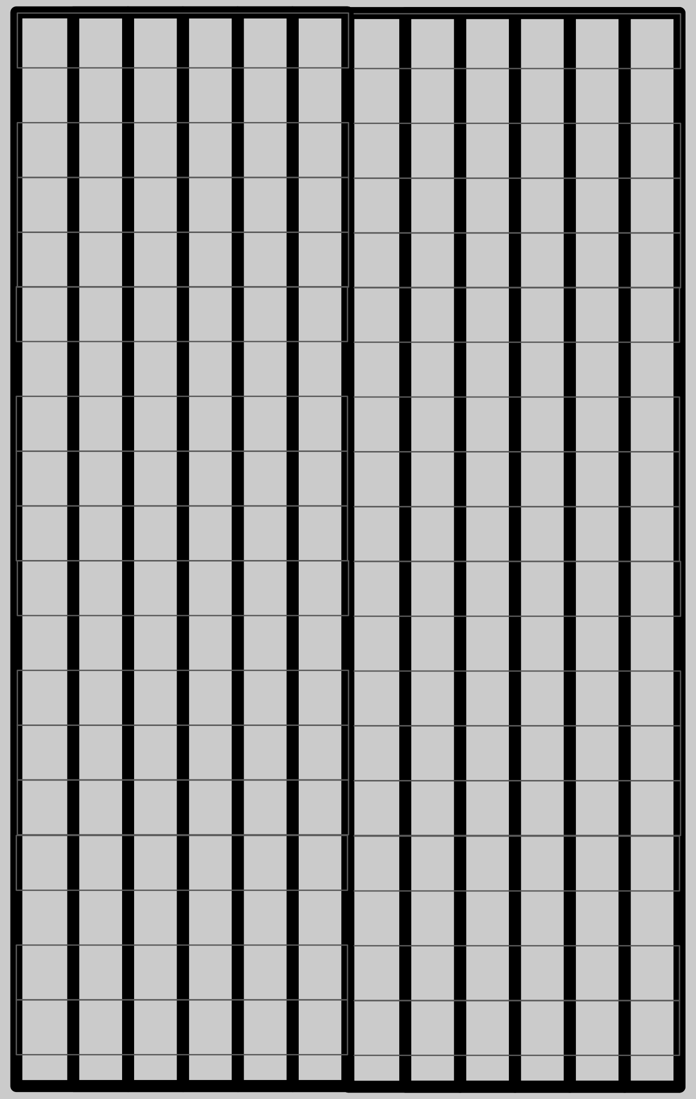
Length inversely correlates with difficulty

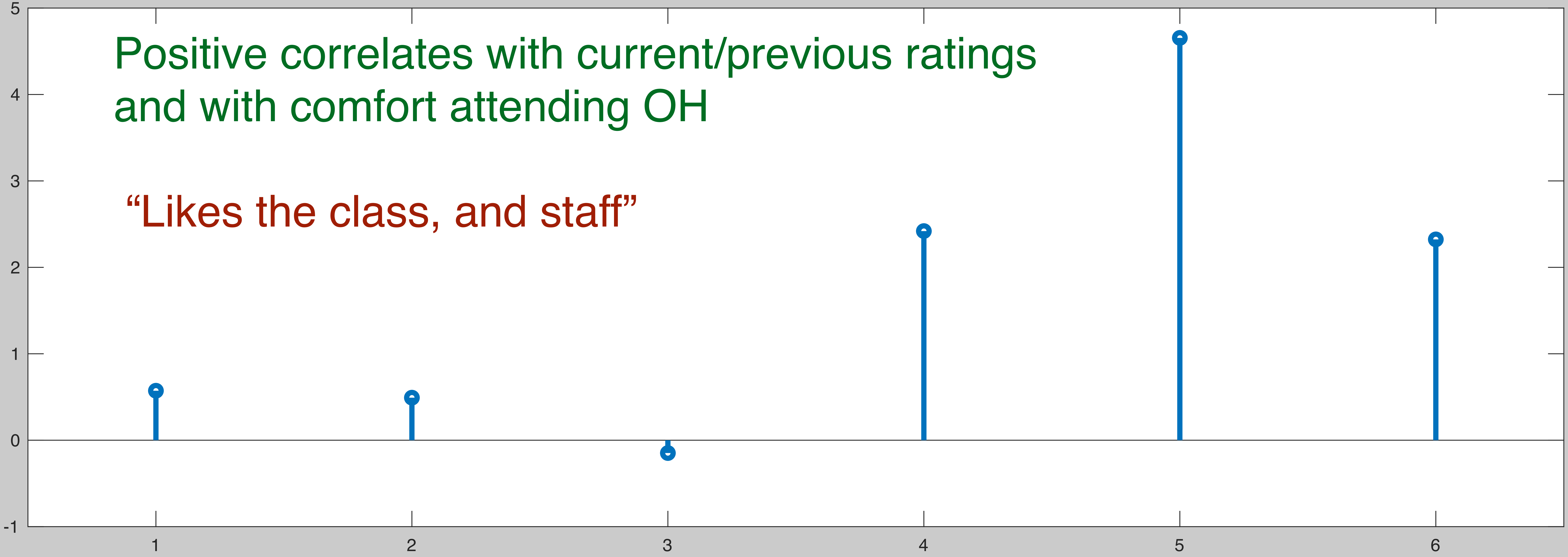
“Class is busy work”

Data Science

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

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH





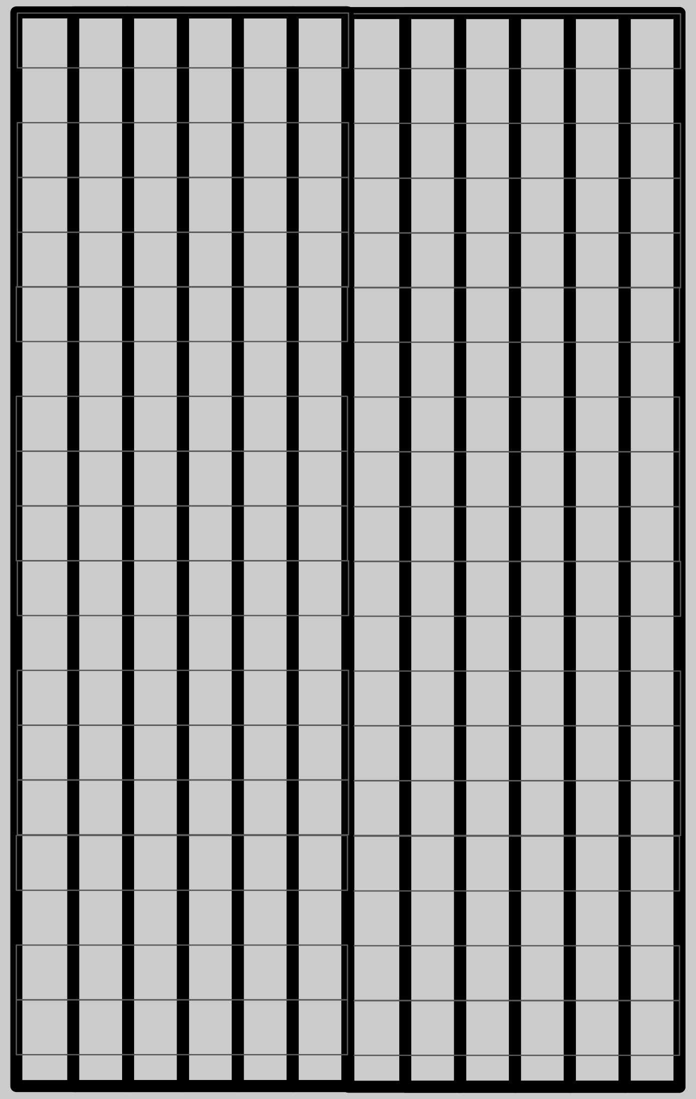
Positive correlates with current/previous ratings and with comfort attending OH

“Likes the class, and staff”

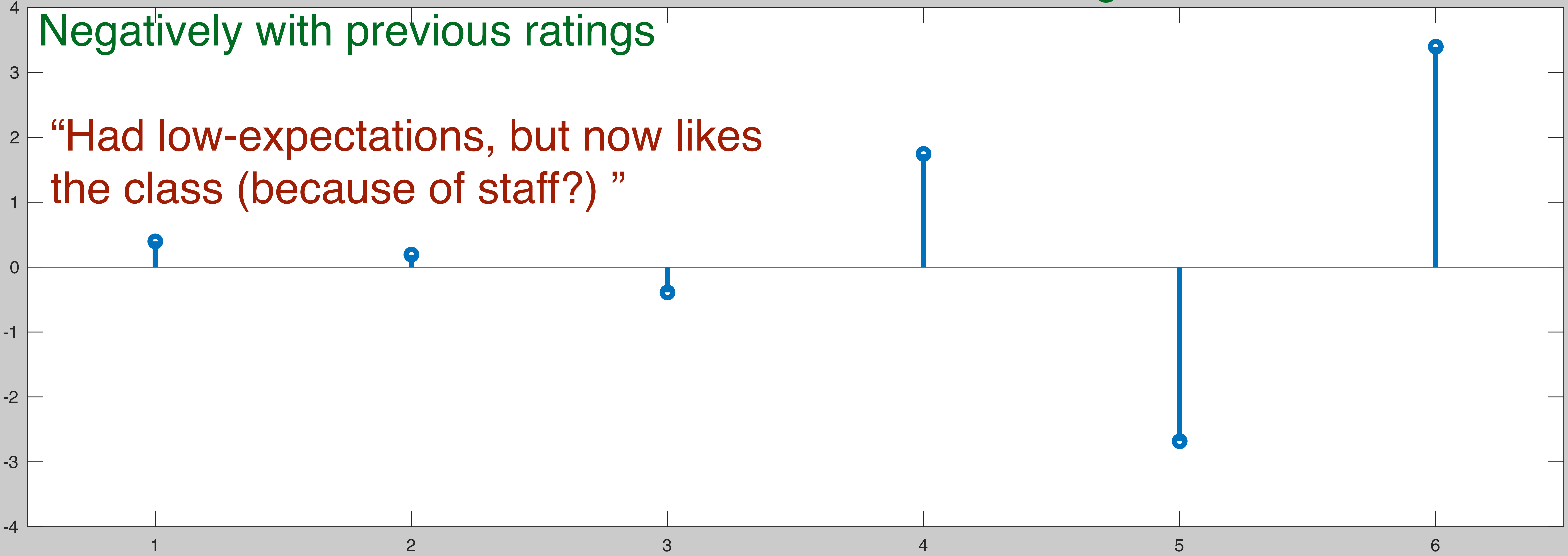
Data Science

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

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH



Positive correlates with current & with comfort attending OH



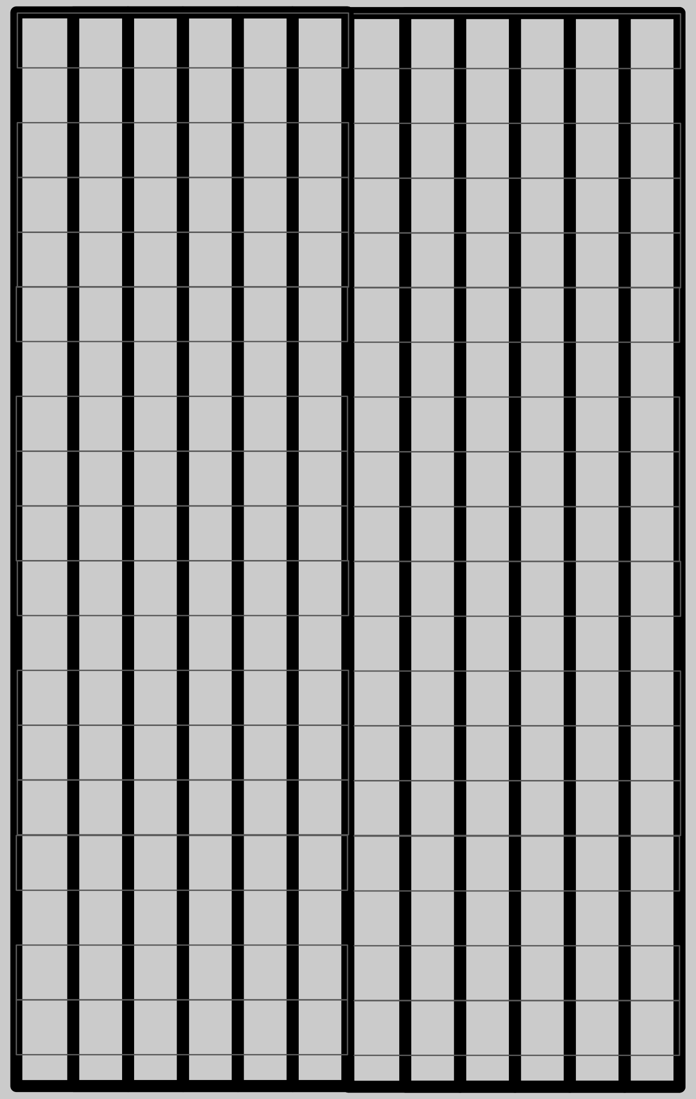
Negatively with previous ratings

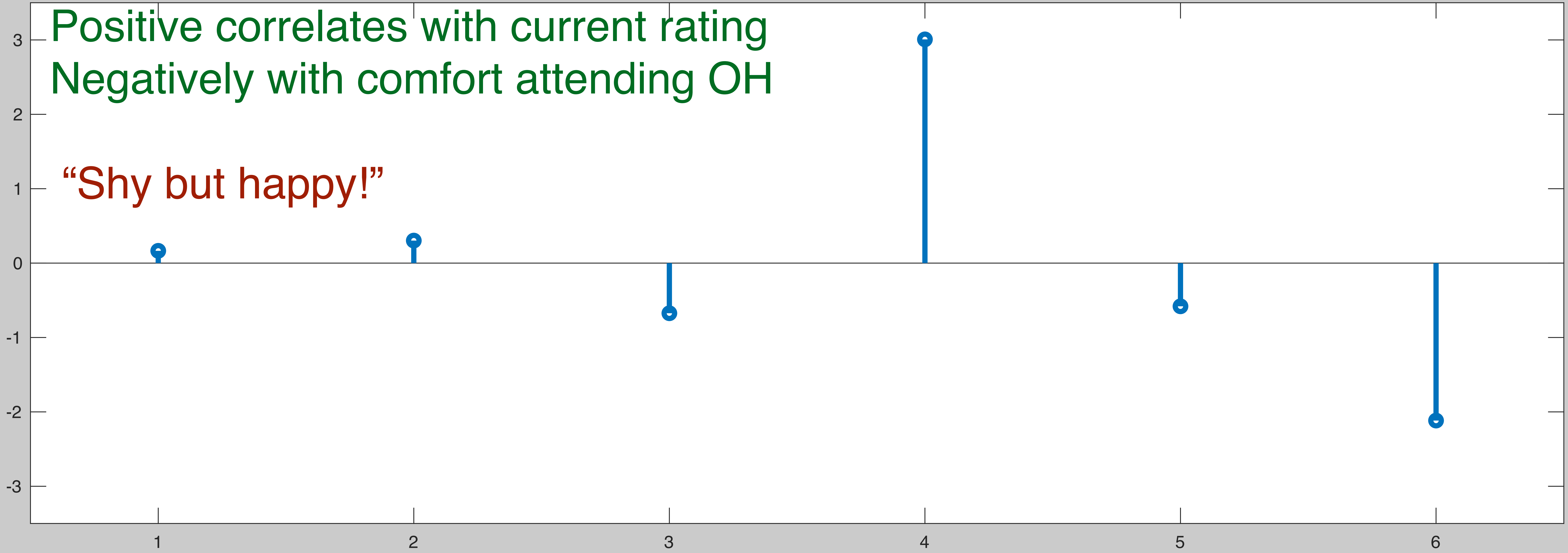
“Had low-expectations, but now likes the class (because of staff?) ”

Data Science

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

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH





Positive correlates with current rating
Negatively with comfort attending OH

“Shy but happy!”

Data Science

$$A^T \vec{u}_6$$

1) HW difficulty

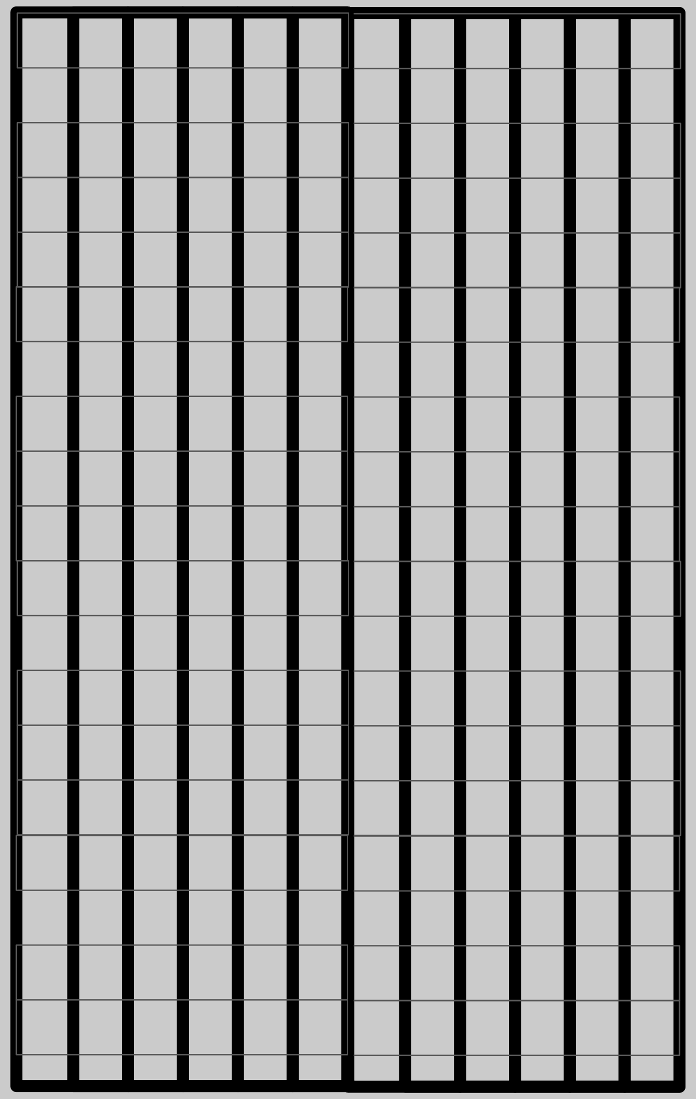
2) HW Length

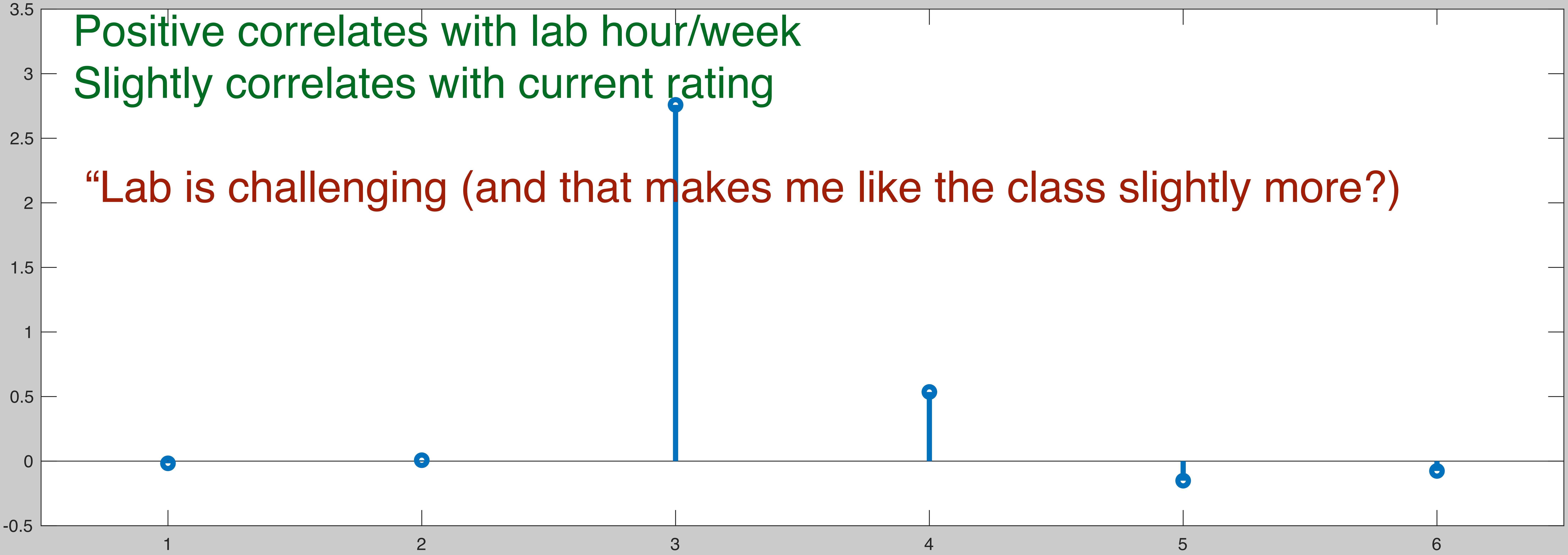
3) Lab hour/week

4) Current rating

5) Previous Rating

6) Comfortable attending OH



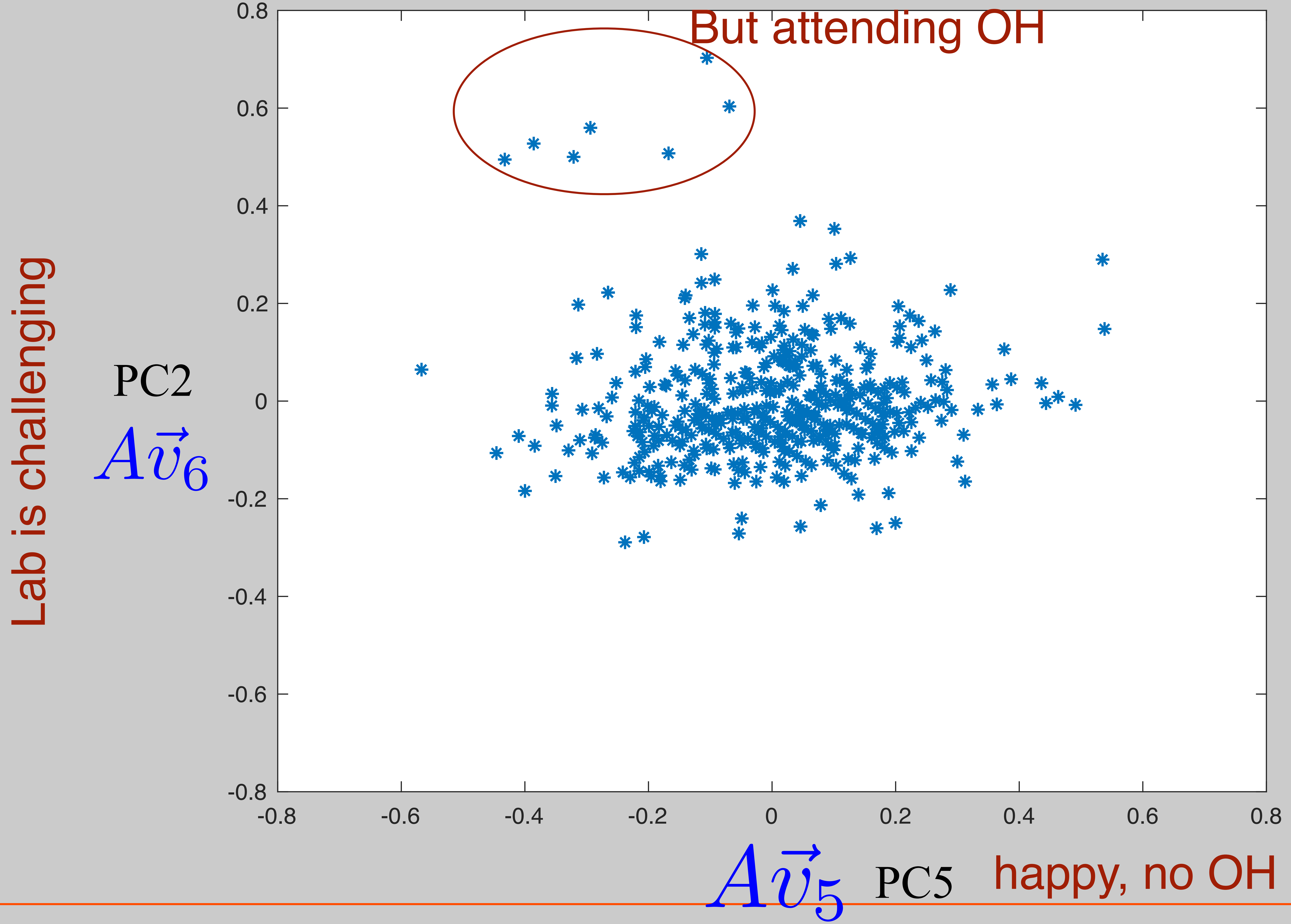


Positive correlates with lab hour/week
Slightly correlates with current rating

“Lab is challenging (and that makes me like the class slightly more?)

Data Science

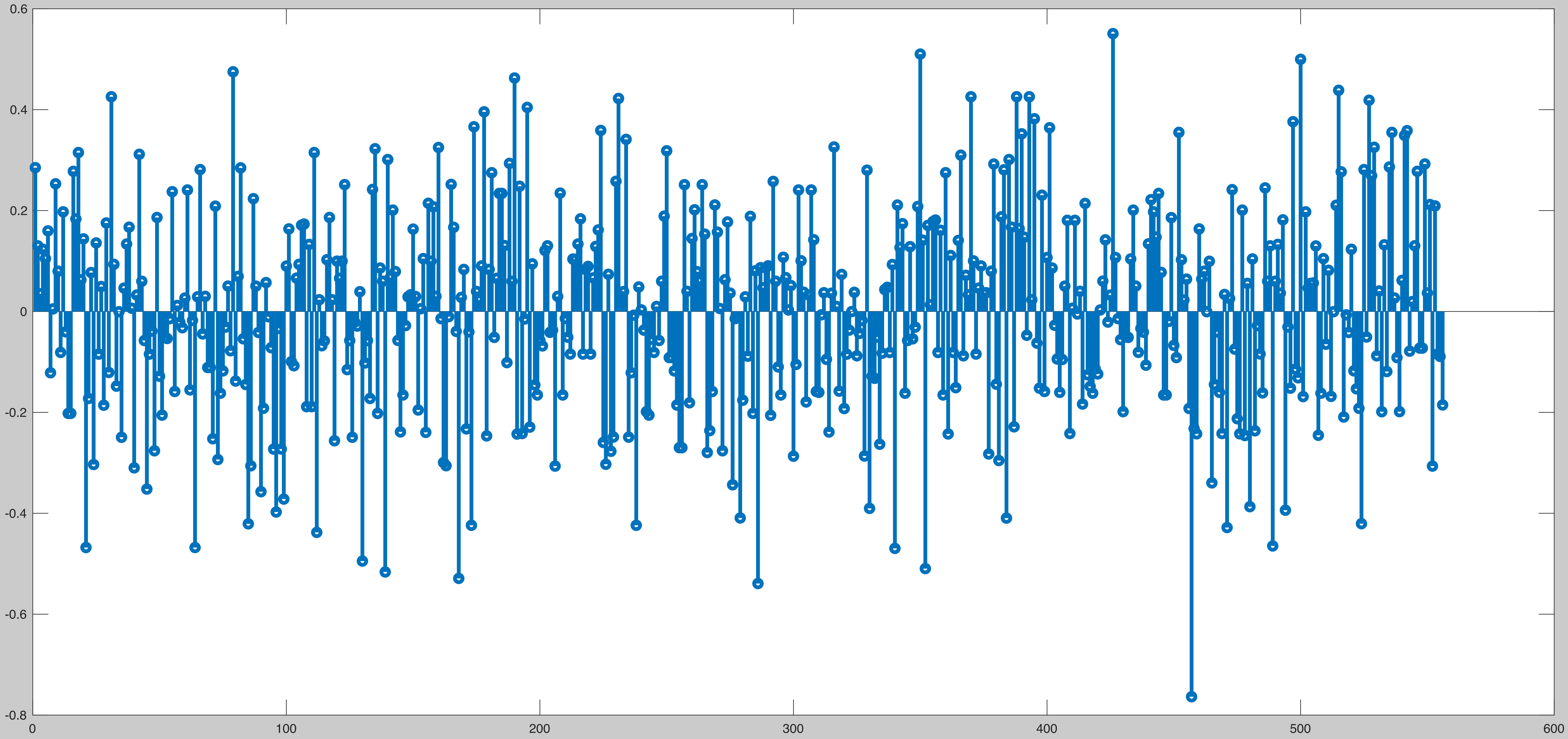
Lab is very challenging
Unhappy
But attending OH



happy, no OH

Data Science

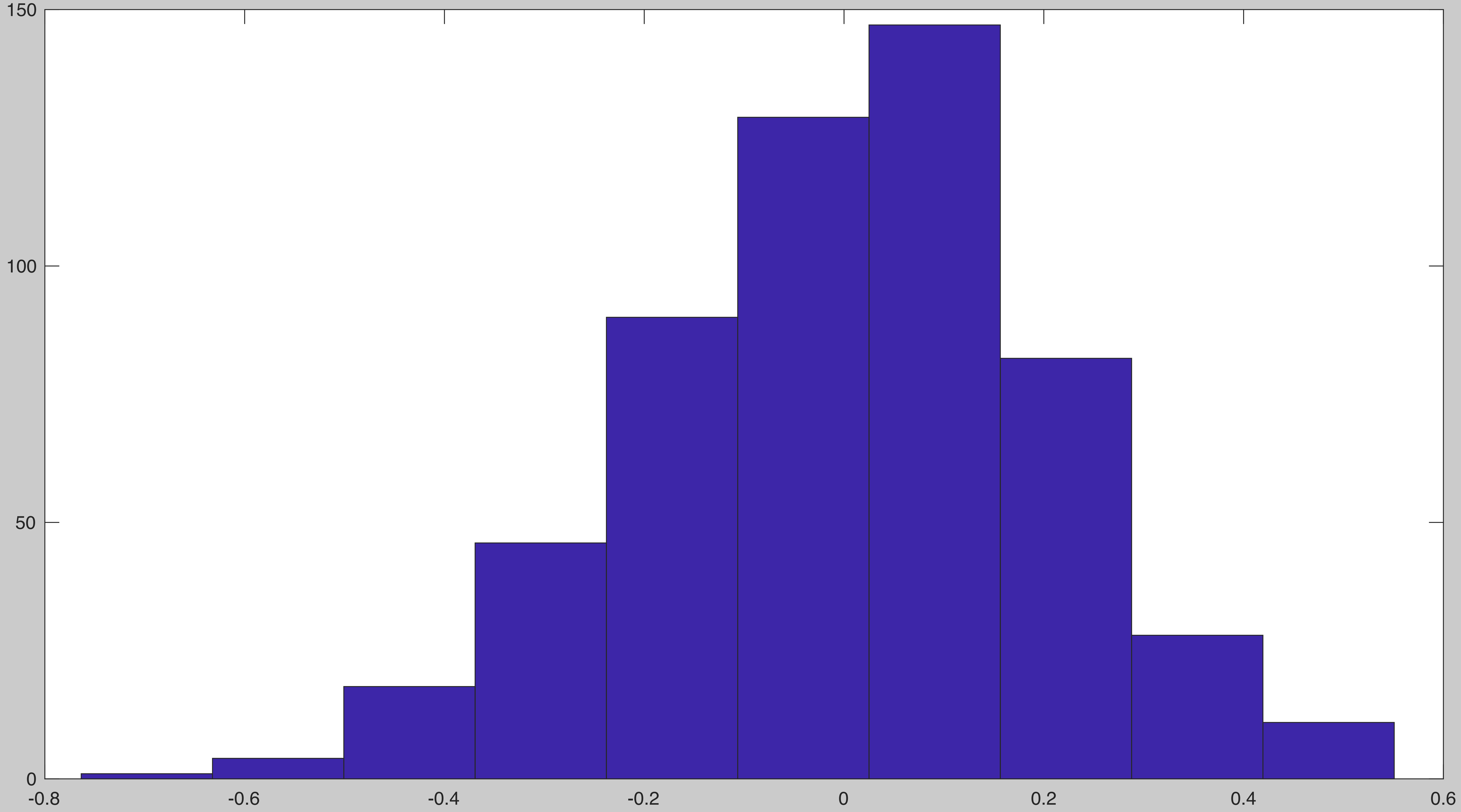
$A\vec{v}_4$ Had low-expectations, but now likes the class



Students

Data Science

Histogram of $A\vec{v}_4$



Had low-expectations, but now like the class

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

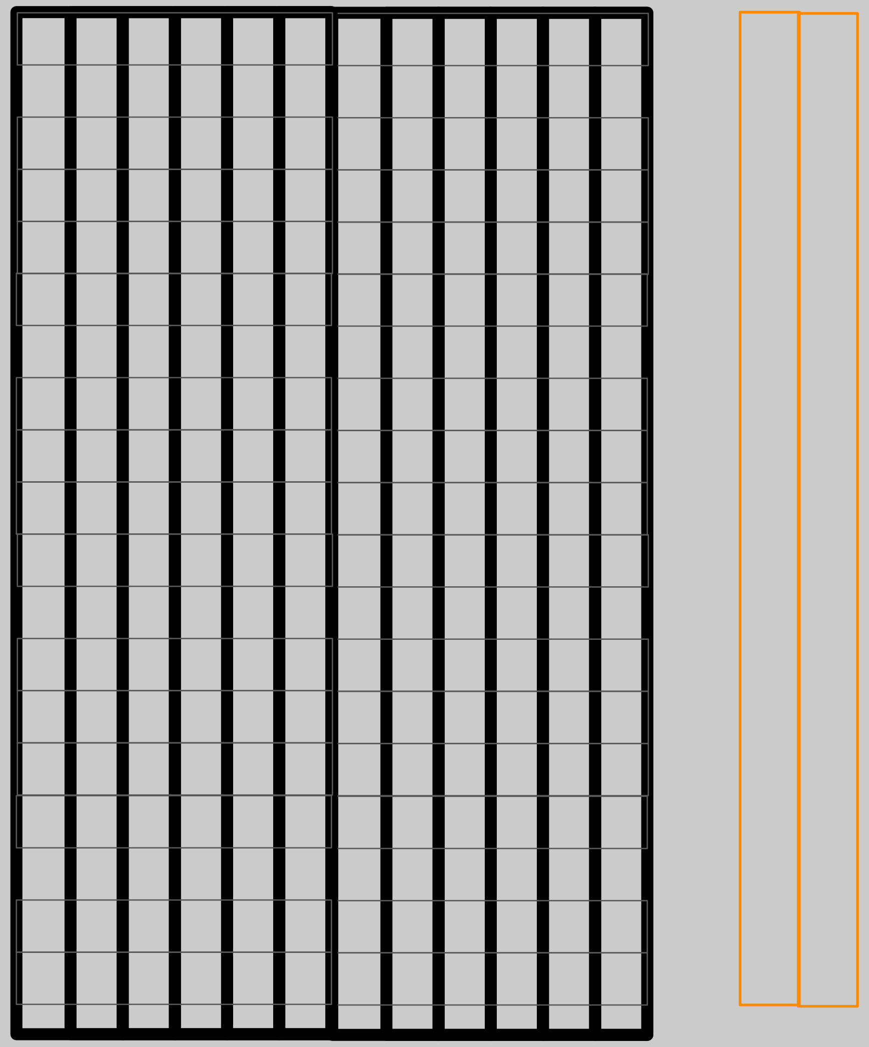
Built a matrix of 200K SNPs (single nucleotide polymorphisms)

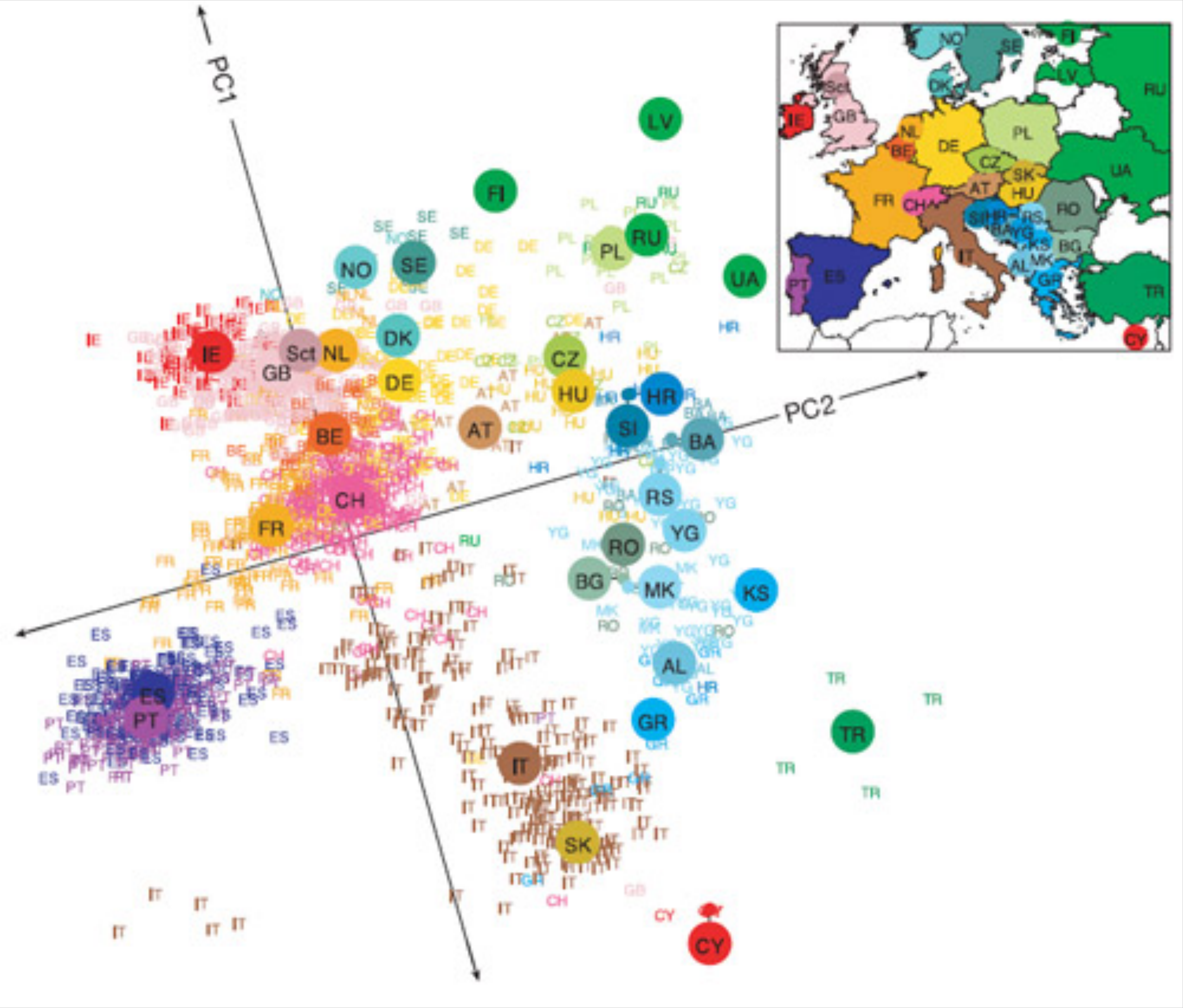
Computed largest 2 principle components

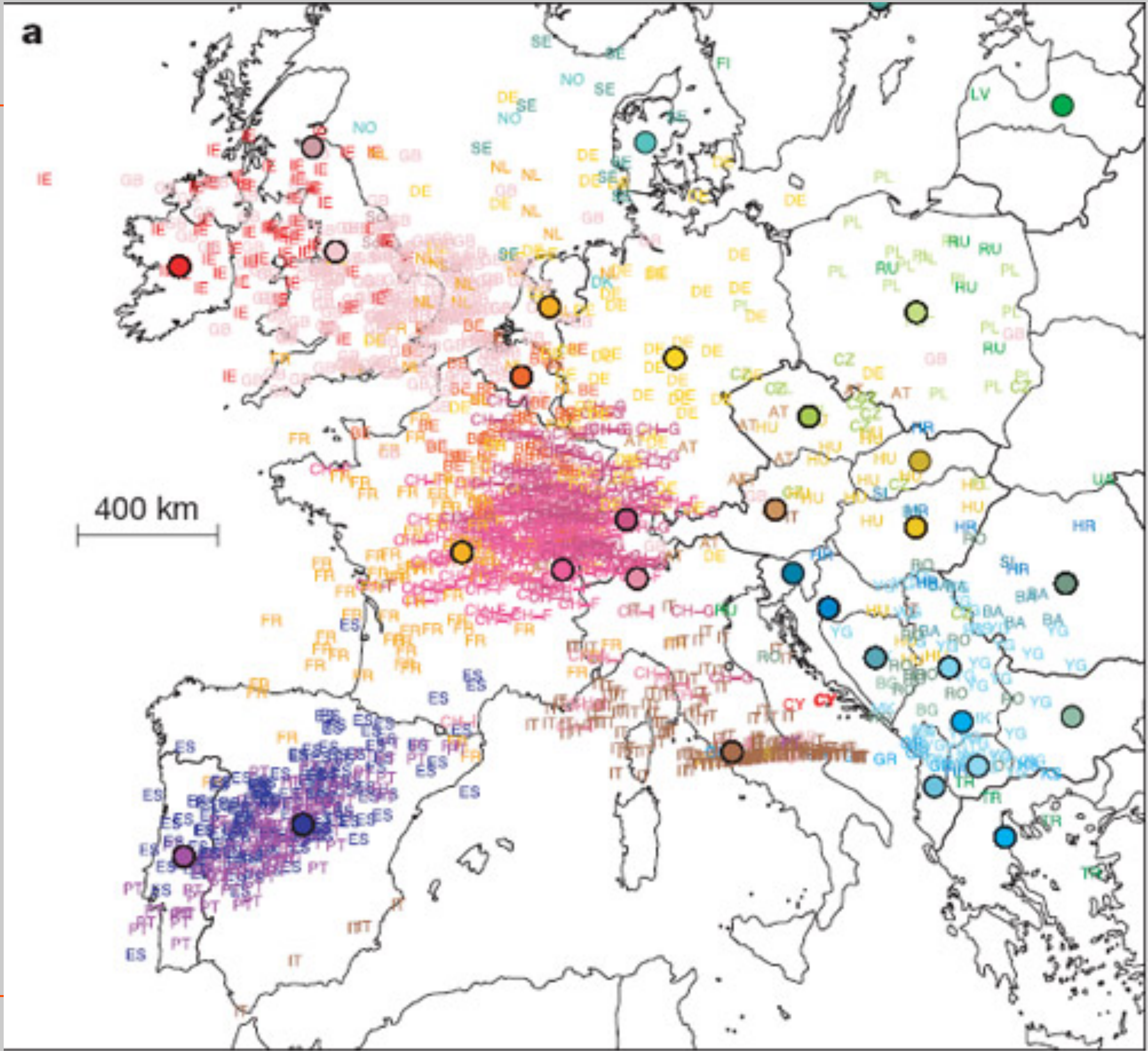
Projected subjects on 2 dimentional data

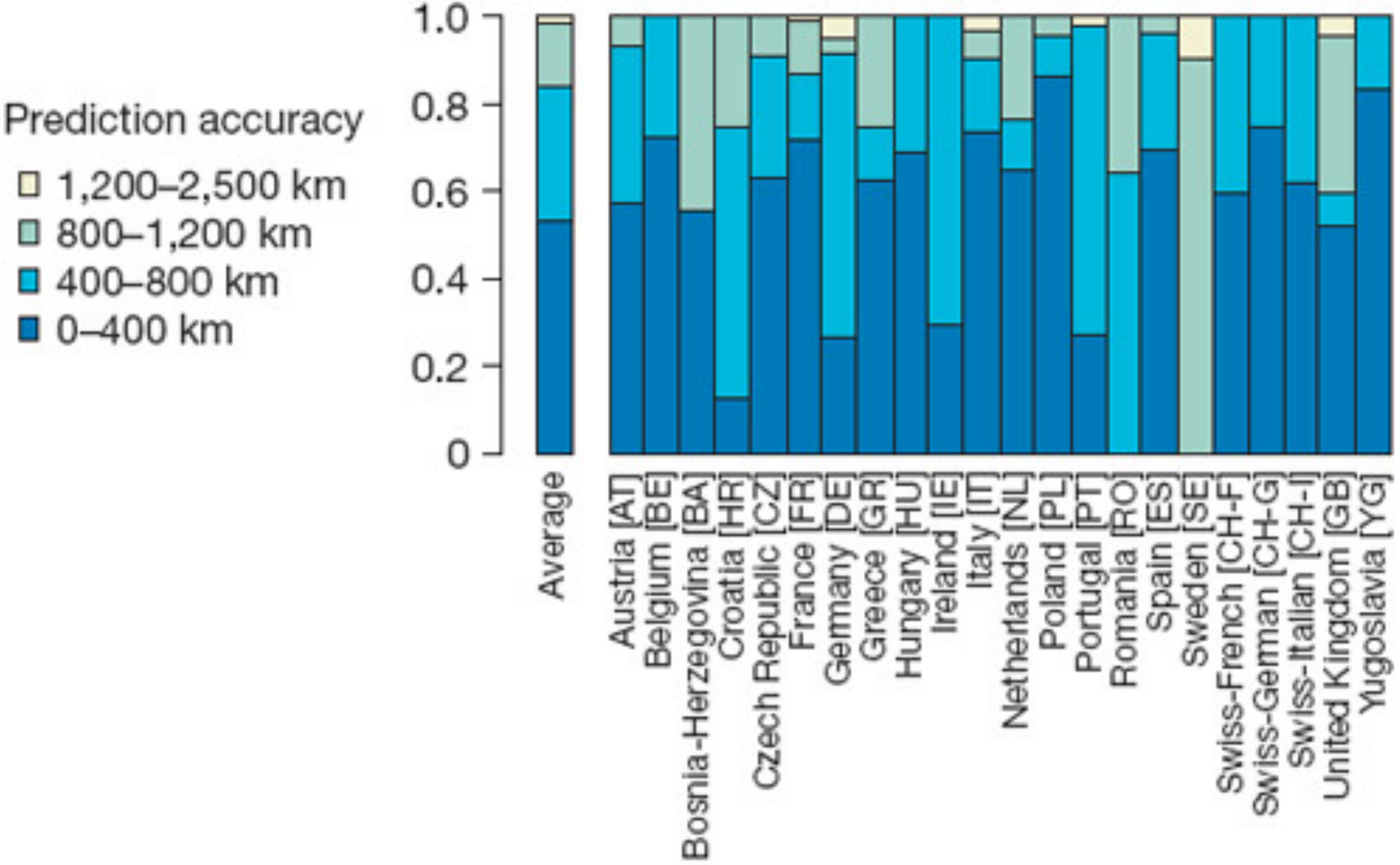
Overlayed the result on the map of Europe

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




PC1
PC2
LV
FI
RU
PL
UA
NO
SE
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
GR
TR
ES
PT
IT
SK
CY


a
400 km

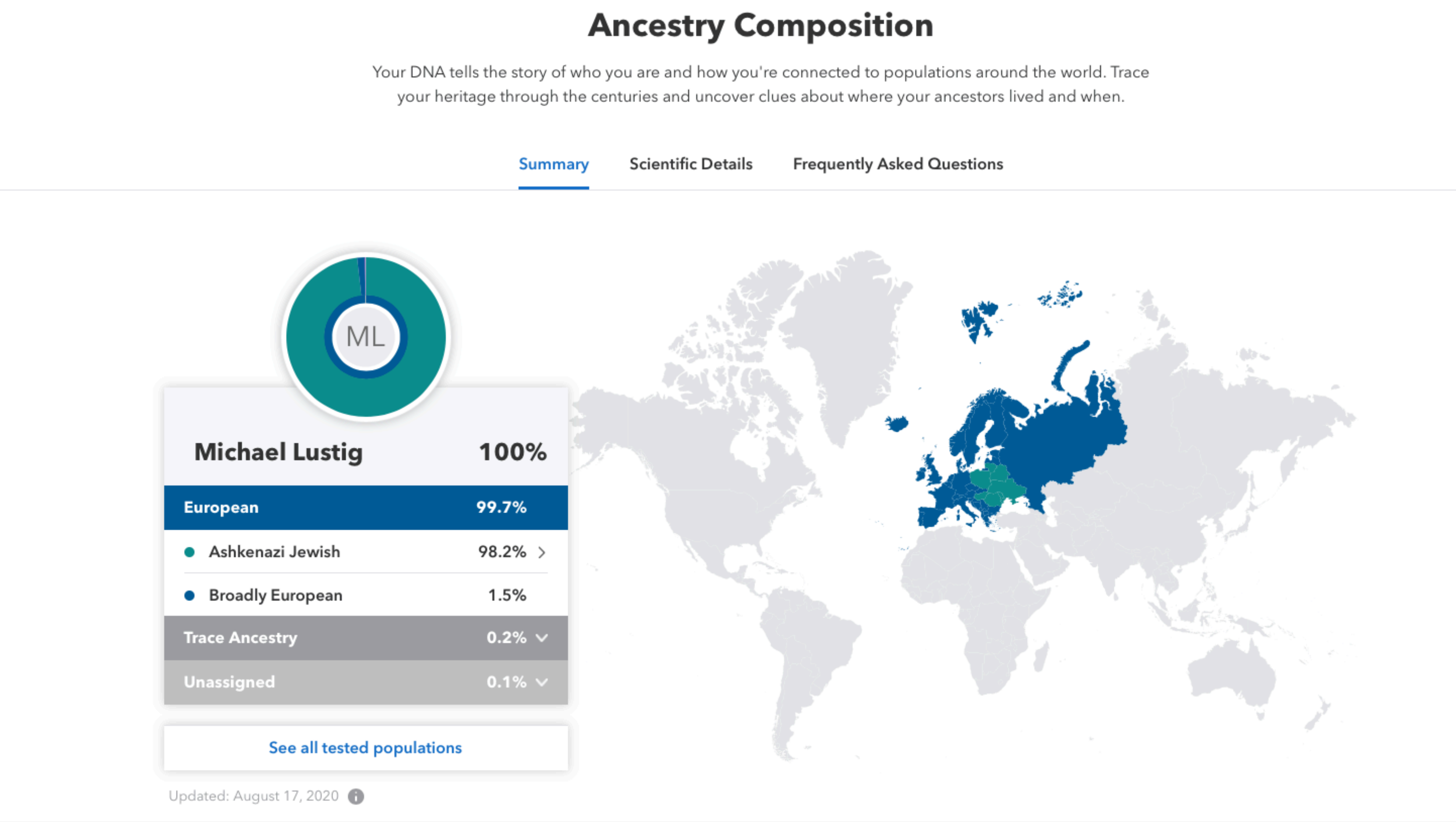

Prediction accuracy
1,200–2,500 km
800–1,200 km
400–800 km
0–400 km
1.0
0.8
0.6
0.4
0.2
0
Average
Austria [AT]
Belgium [BE]
Bosnia-Herzegovina [BA]
Croatia [HR]
Czech Republic [CZ]
France [FR]
Germany [DE]
Greece [GR]
Hungary [HU]
Ireland [IE]
Italy [IT]
Netherlands [NL]
Poland [PL]
Portugal [PT]
Romania [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



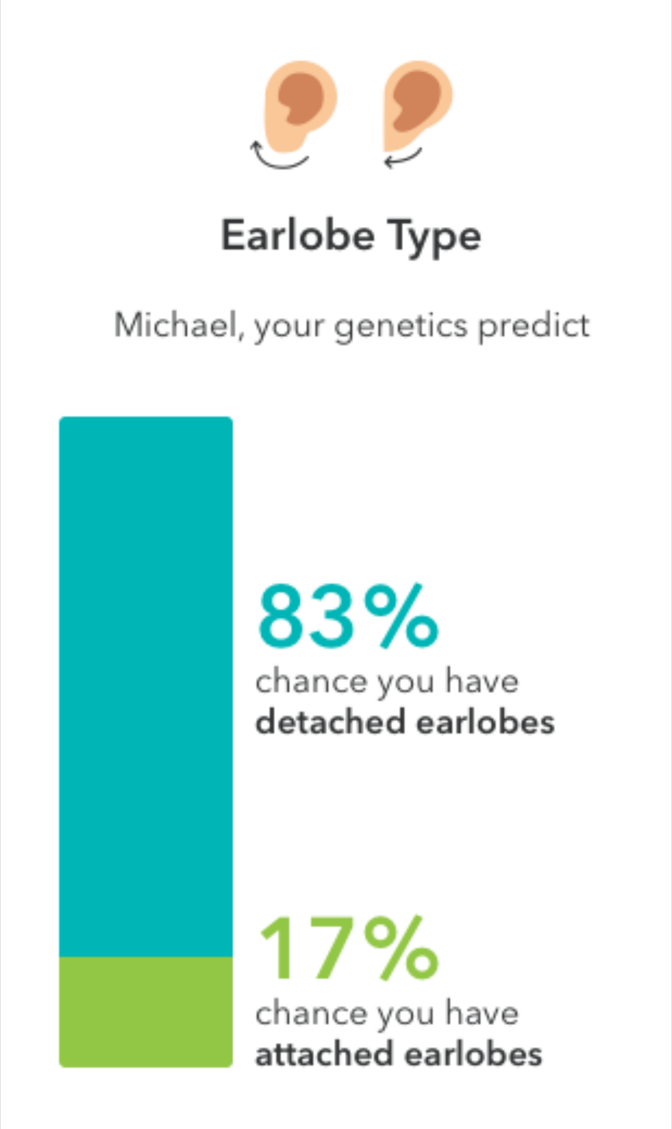
Ancestry Composition

Your DNA tells the story of who you are and how you're connected to populations around the world. Trace your heritage through the centuries and uncover clues about where your ancestors lived and when.

Michael Lustig	100%
European	99.7%
● Ashkenazi Jewish	98.2% ›
● Broadly European	1.5%
Trace Ancestry	0.2% ⌄
Unassigned	0.1% ⌄

Updated: August 17, 2020

Physical features

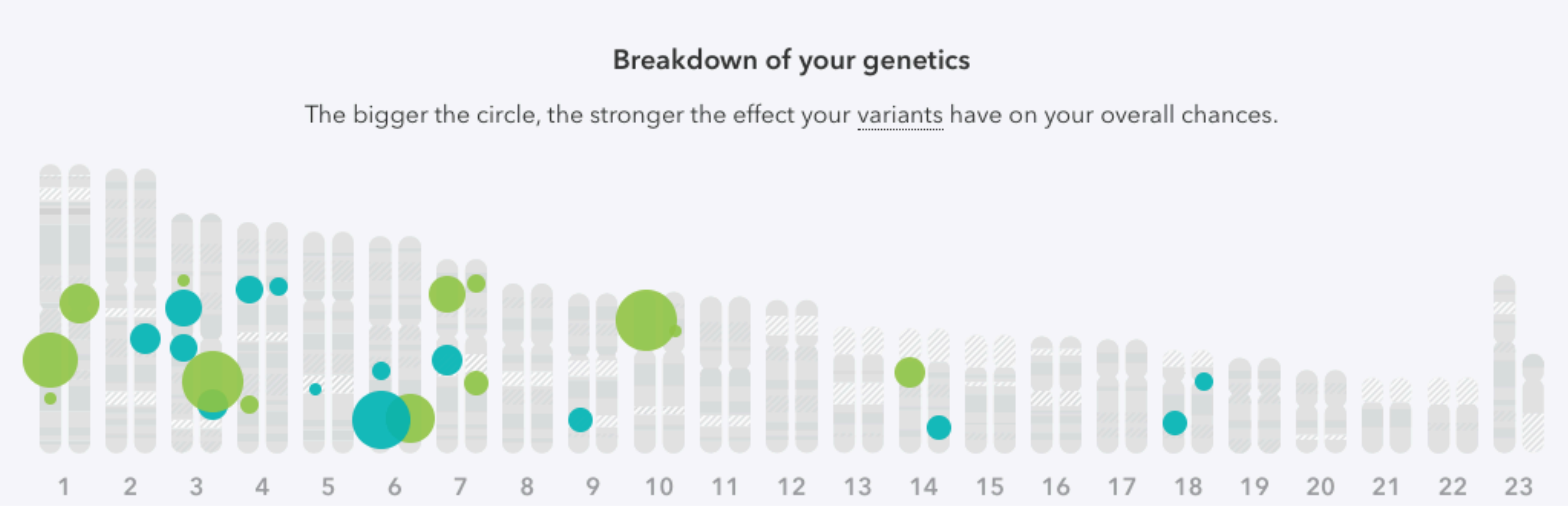


Earlobe Type

Michael, your genetics predict

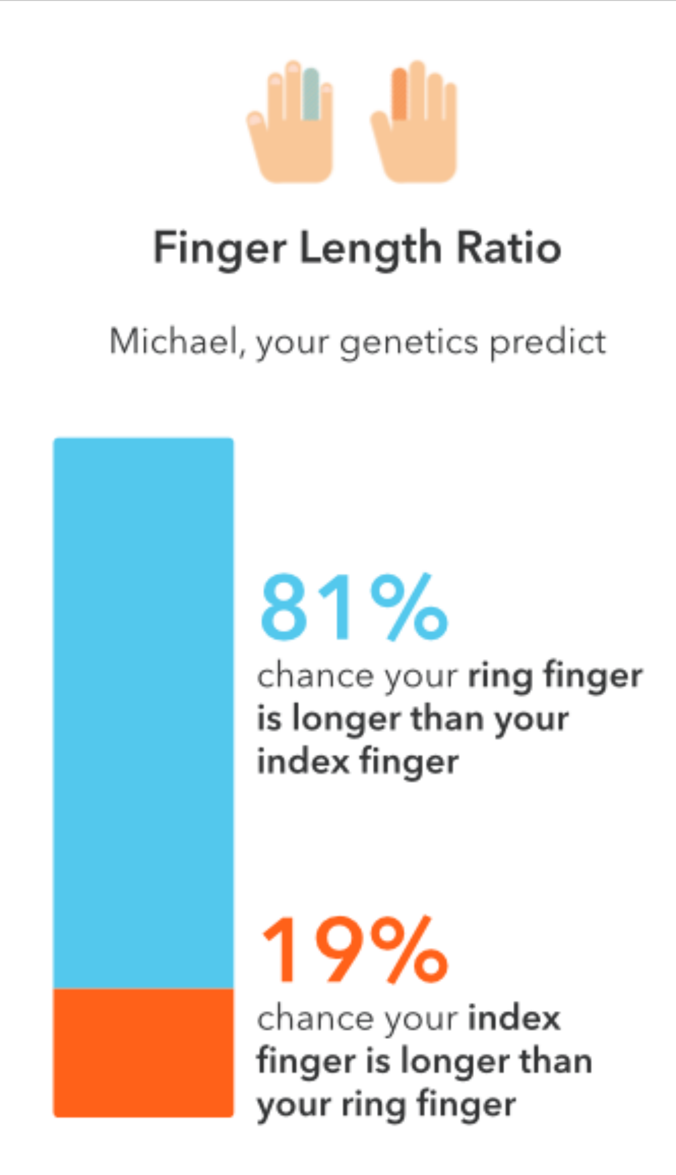
83% chance you have **detached earlobes**

17% chance you have **attached earlobes**



Breakdown of your genetics

The bigger the circle, the stronger the effect your variants have on your overall chances.

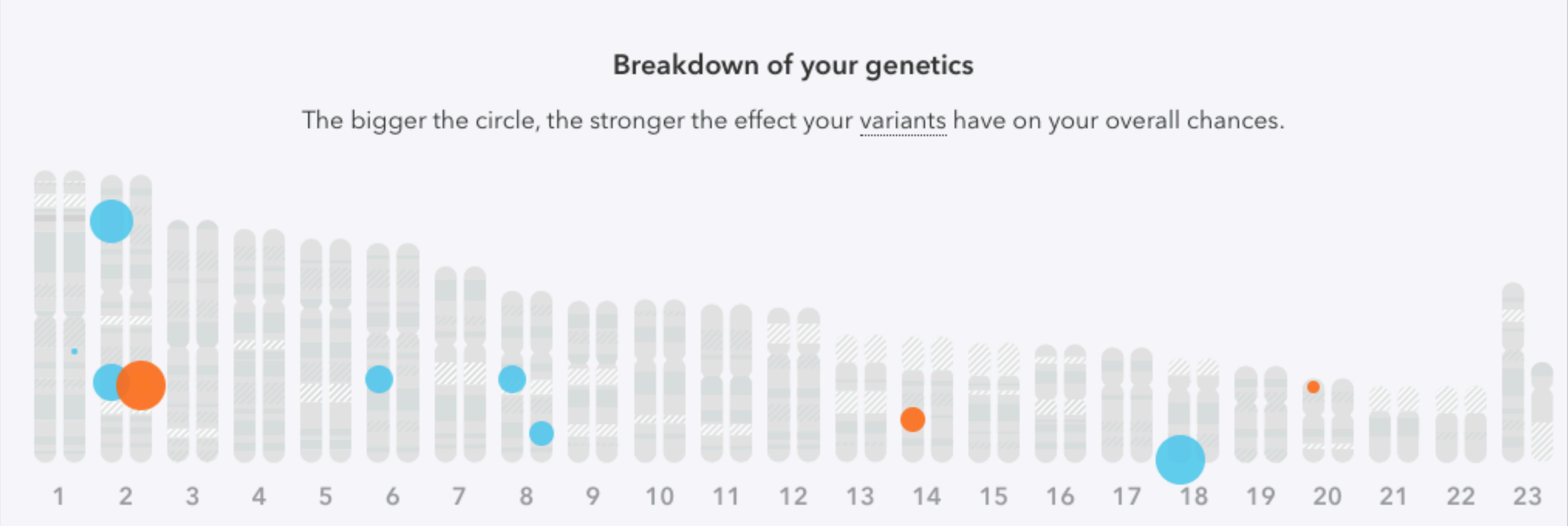


Finger Length Ratio

Michael, your genetics predict

81% chance your **ring finger is longer than your index finger**

19% chance your **index finger is longer than your ring finger**



Breakdown of your genetics

The bigger the circle, the stronger the effect your variants have on your overall chances.

Labeled VS non labeled Classification

Word1

Word2

Word3

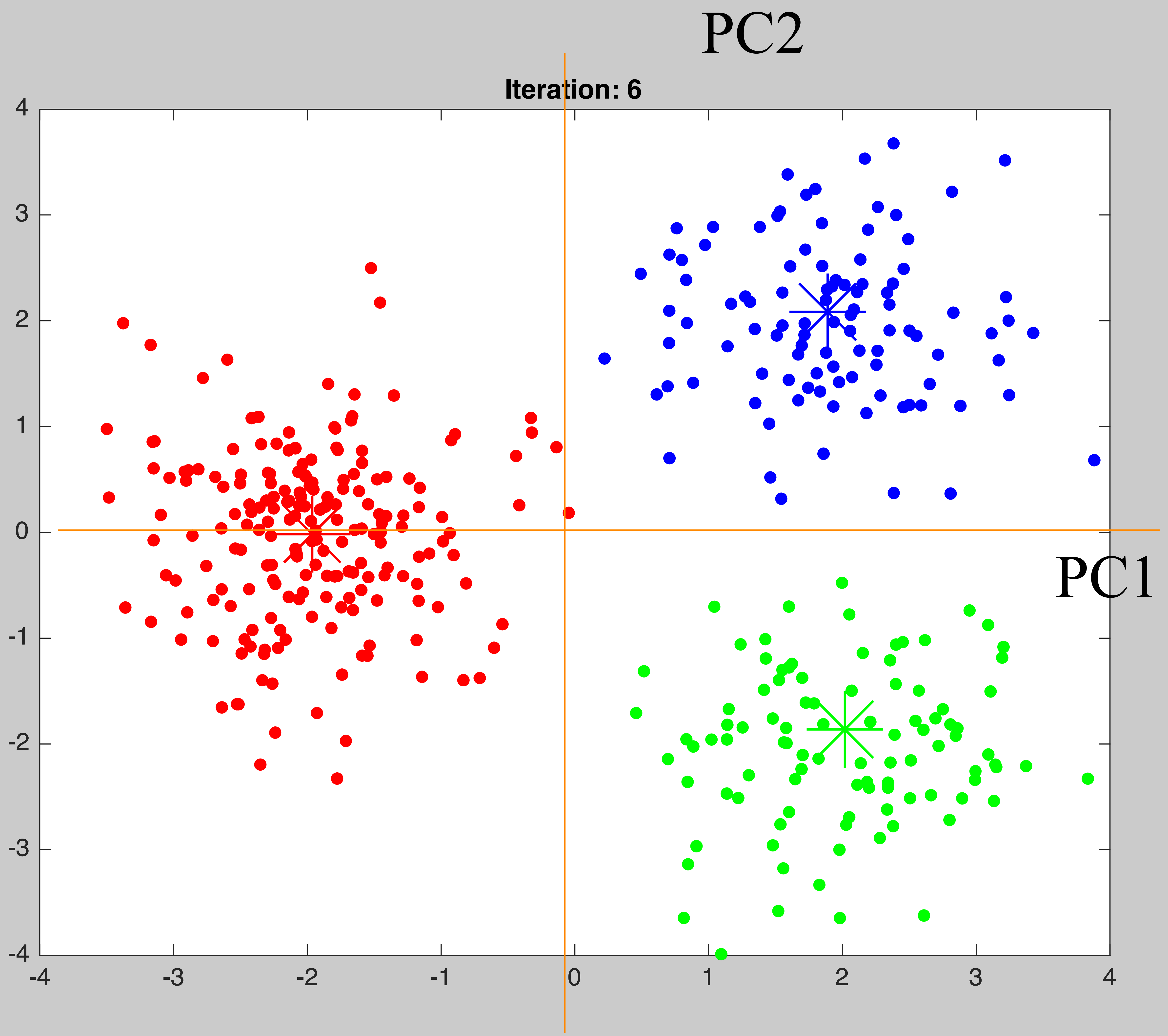
Word4

Word5

Word6

Word7

Word8



Labeled VS non labeled Classification

Word1

Word2

Word3

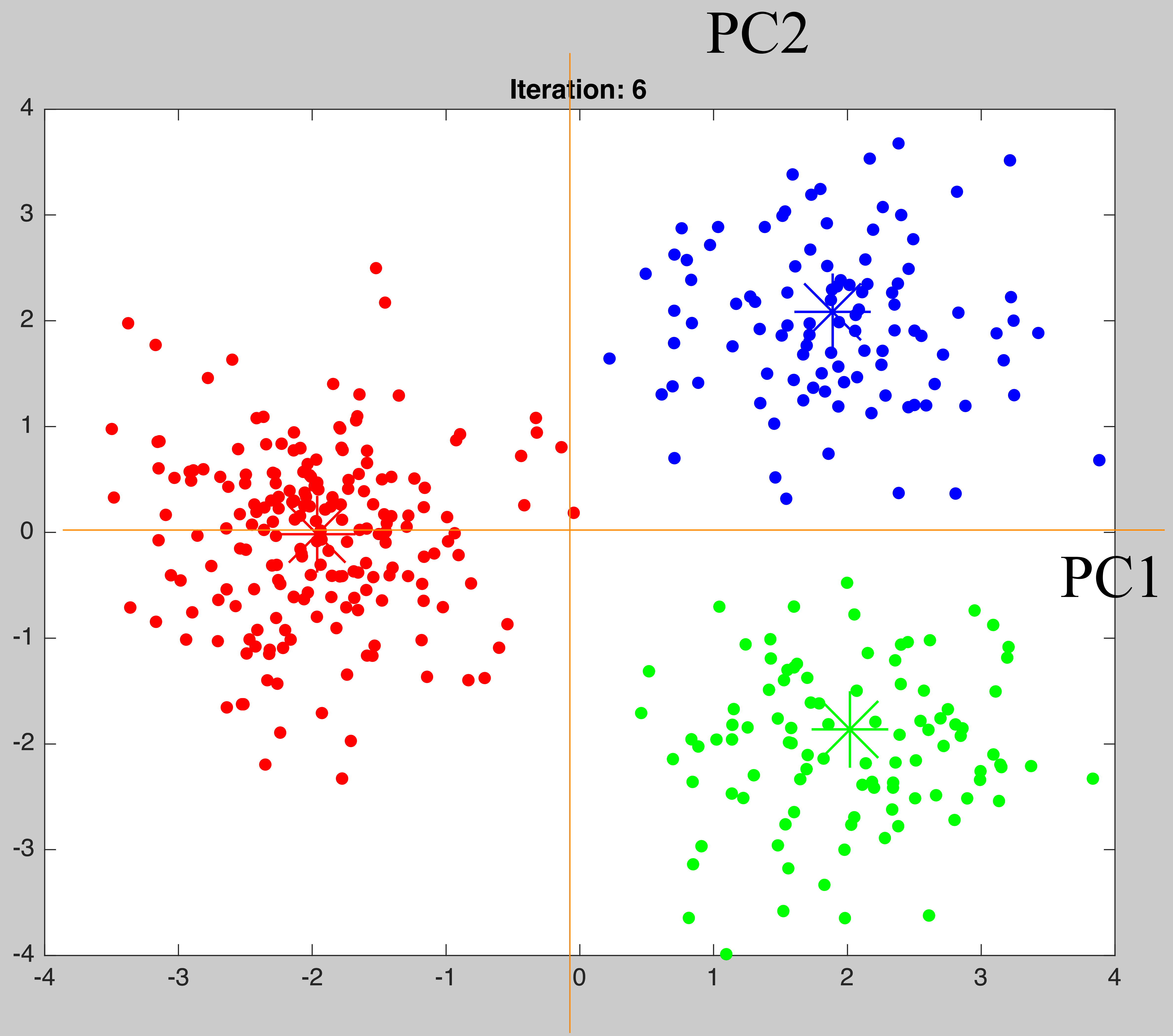
Word4

Word5

Word6

Word7

Word8



Labeled VS non labeled Classification

“Banana”

“Banana”

“Banana”

“Mango”

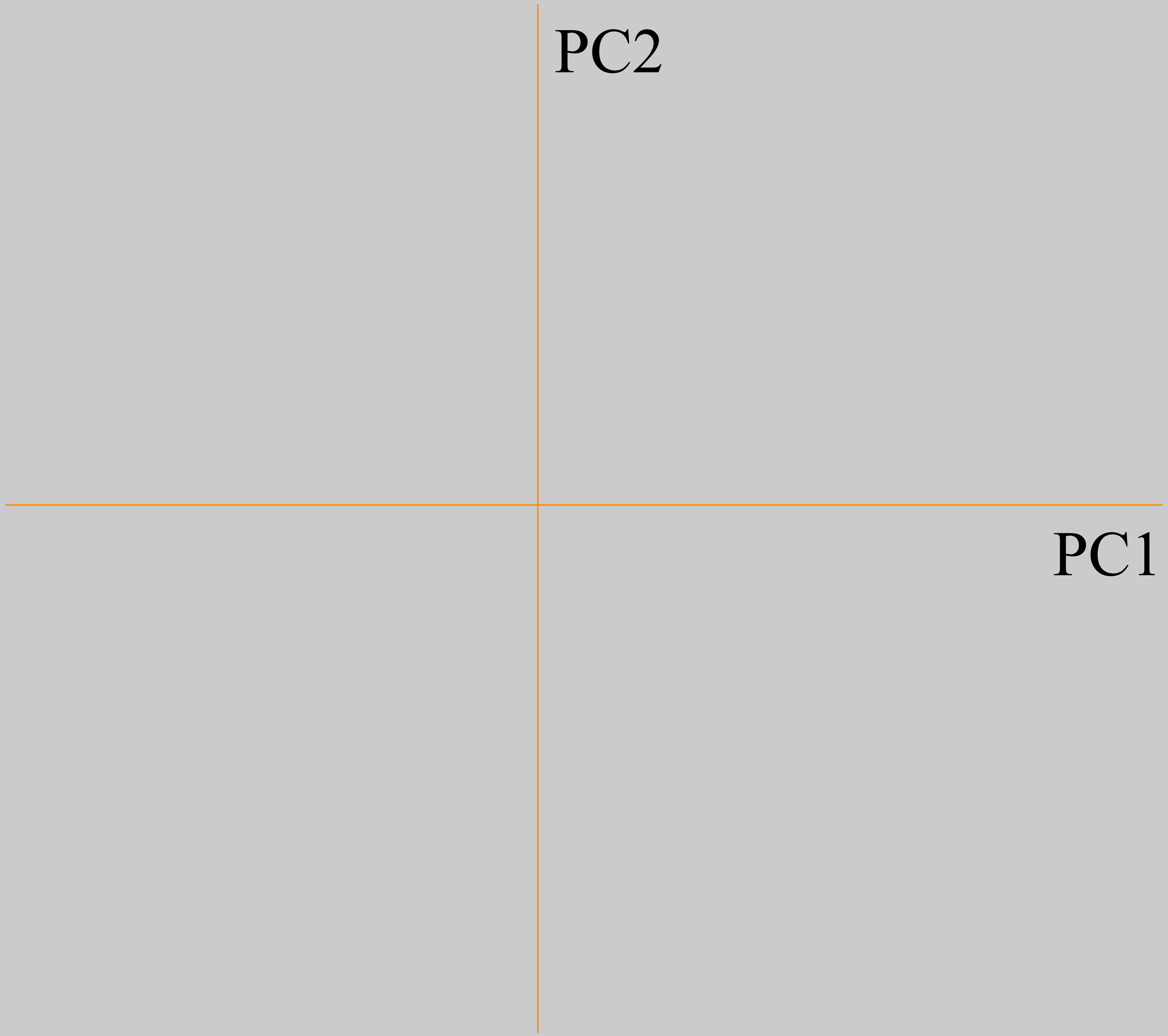
“Mango”

“Mango”

“Chop”

“Chop”

“Chop”



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

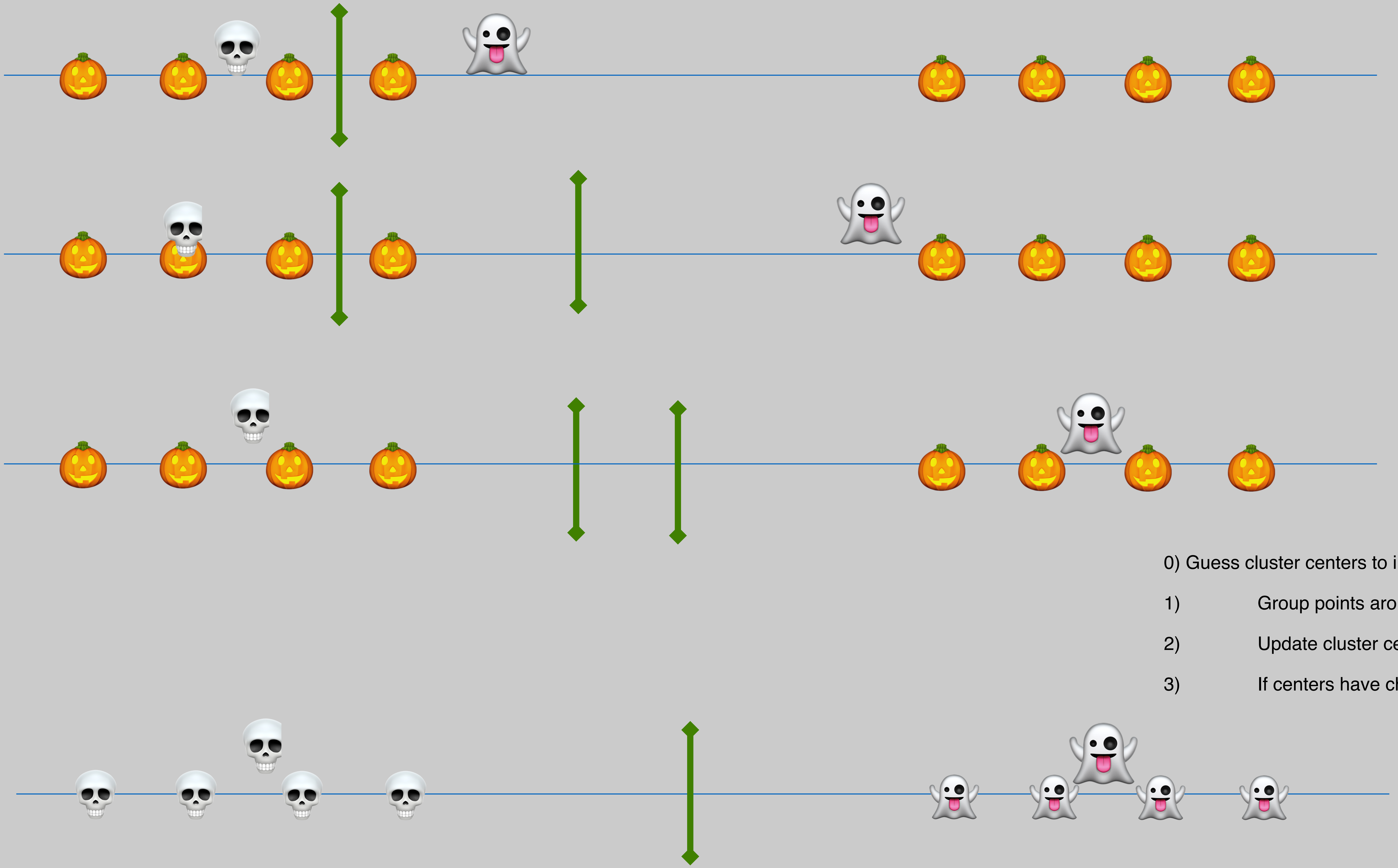
1) Group points around nearest center

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 $\vec{m}_1, \vec{m}_2, \cdots, \vec{m}_k$

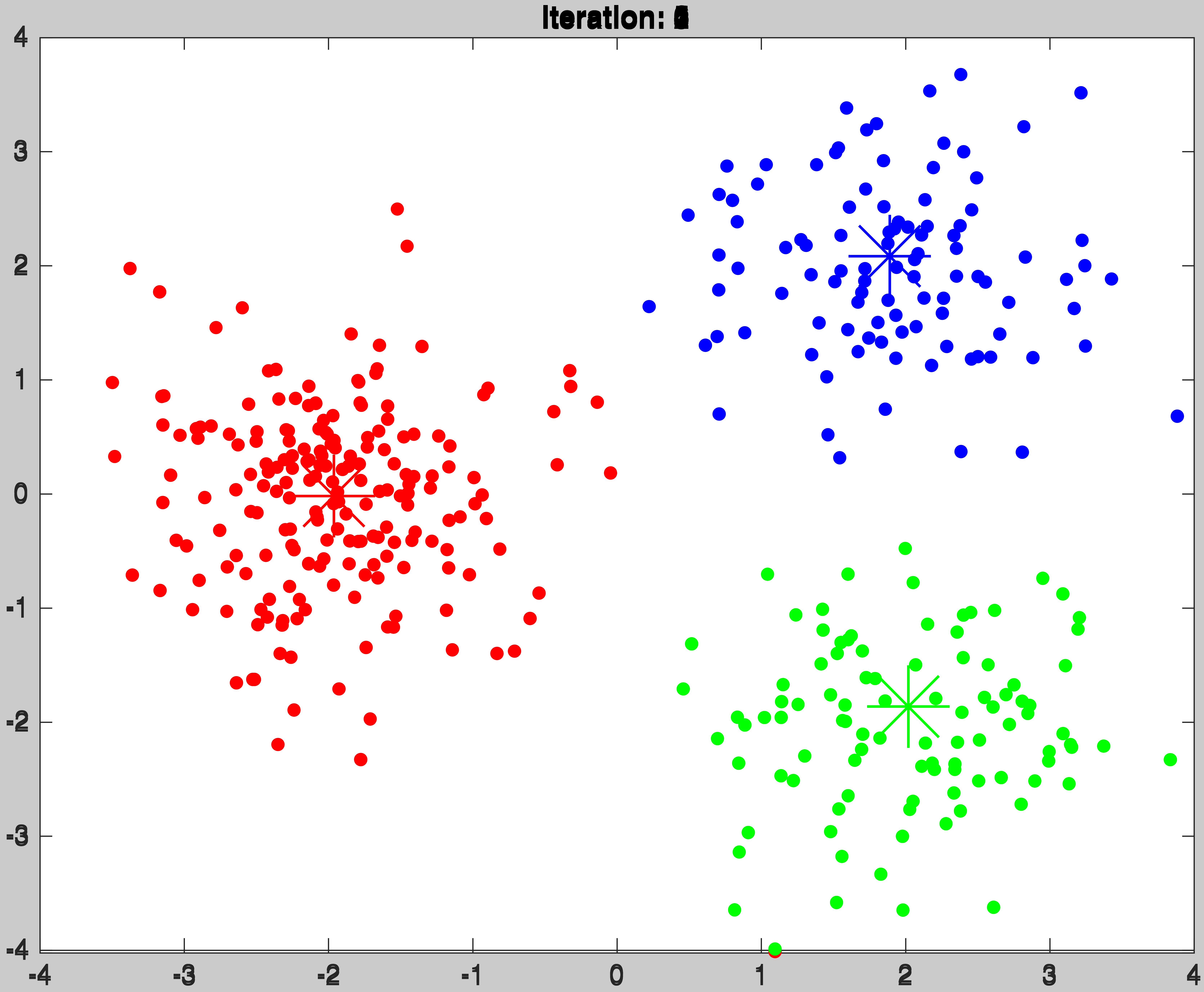
1) Assign points to cluster: point $\vec{x}$ goes to cluster i if,

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

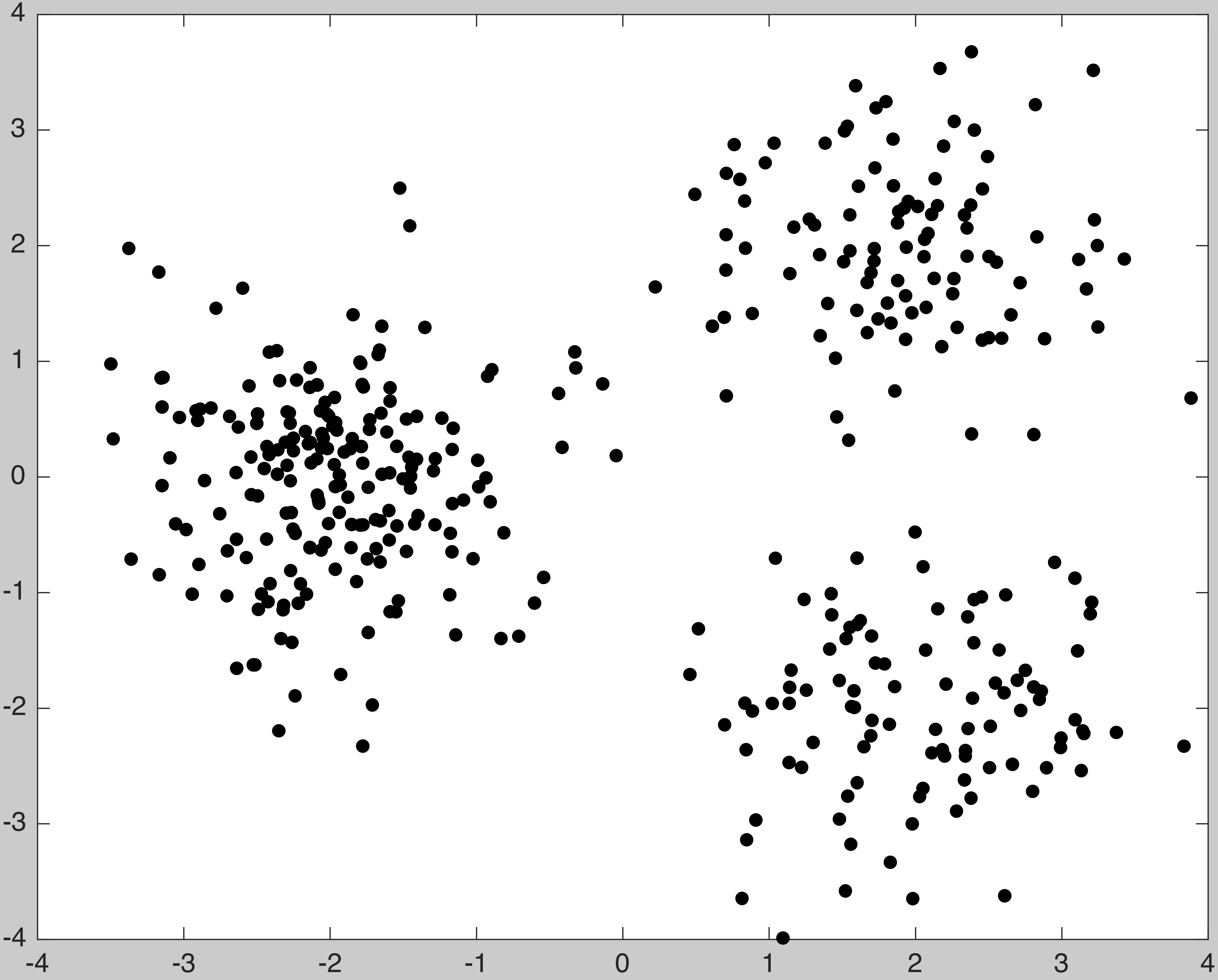
2) Let S_i be the set of samples in cluster i recompute cluster centers:

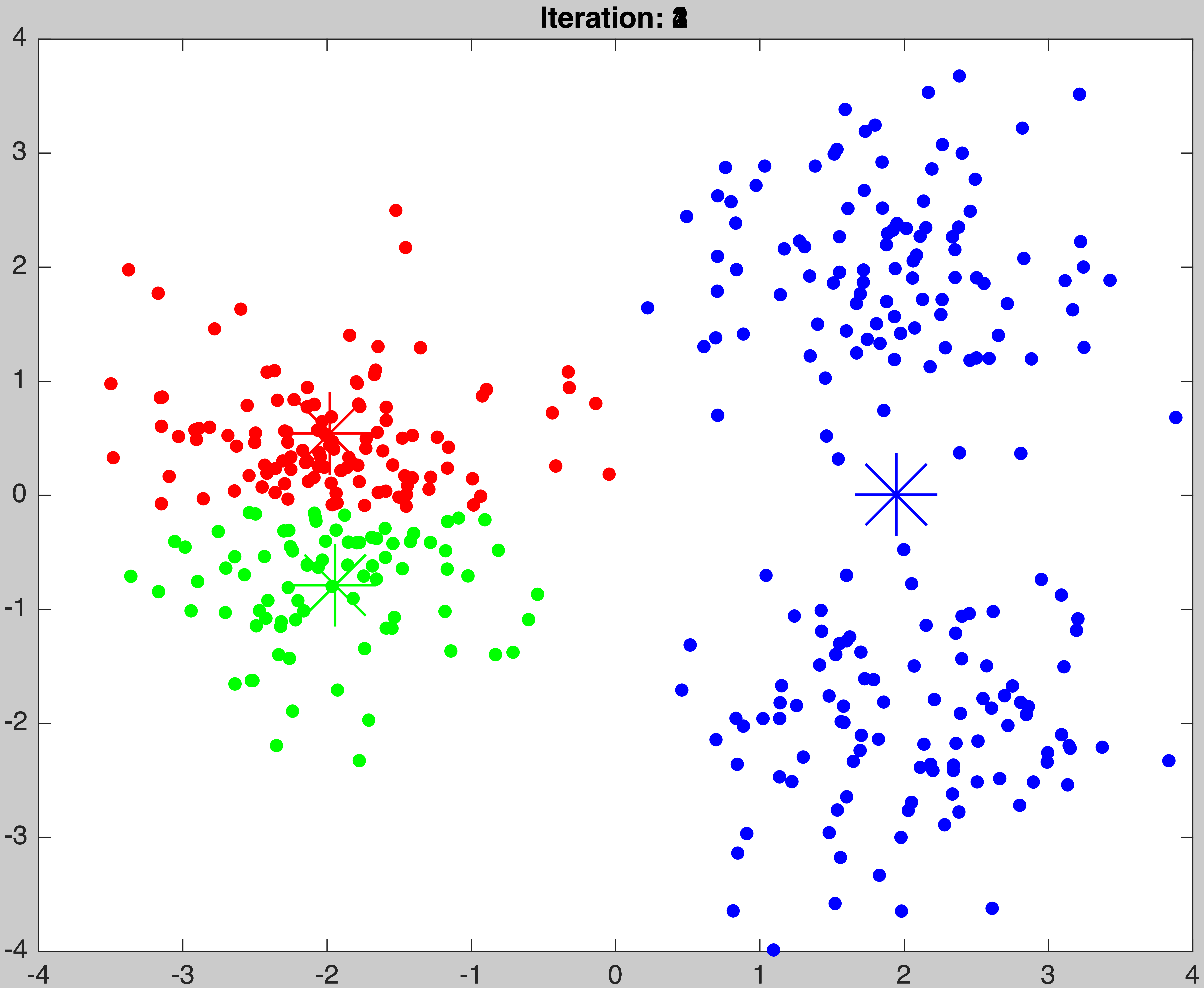
$$\vec{m}_i = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}$$

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



EE16B M. Lustig, EECS UC Berkeley


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


Iteration: 2
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

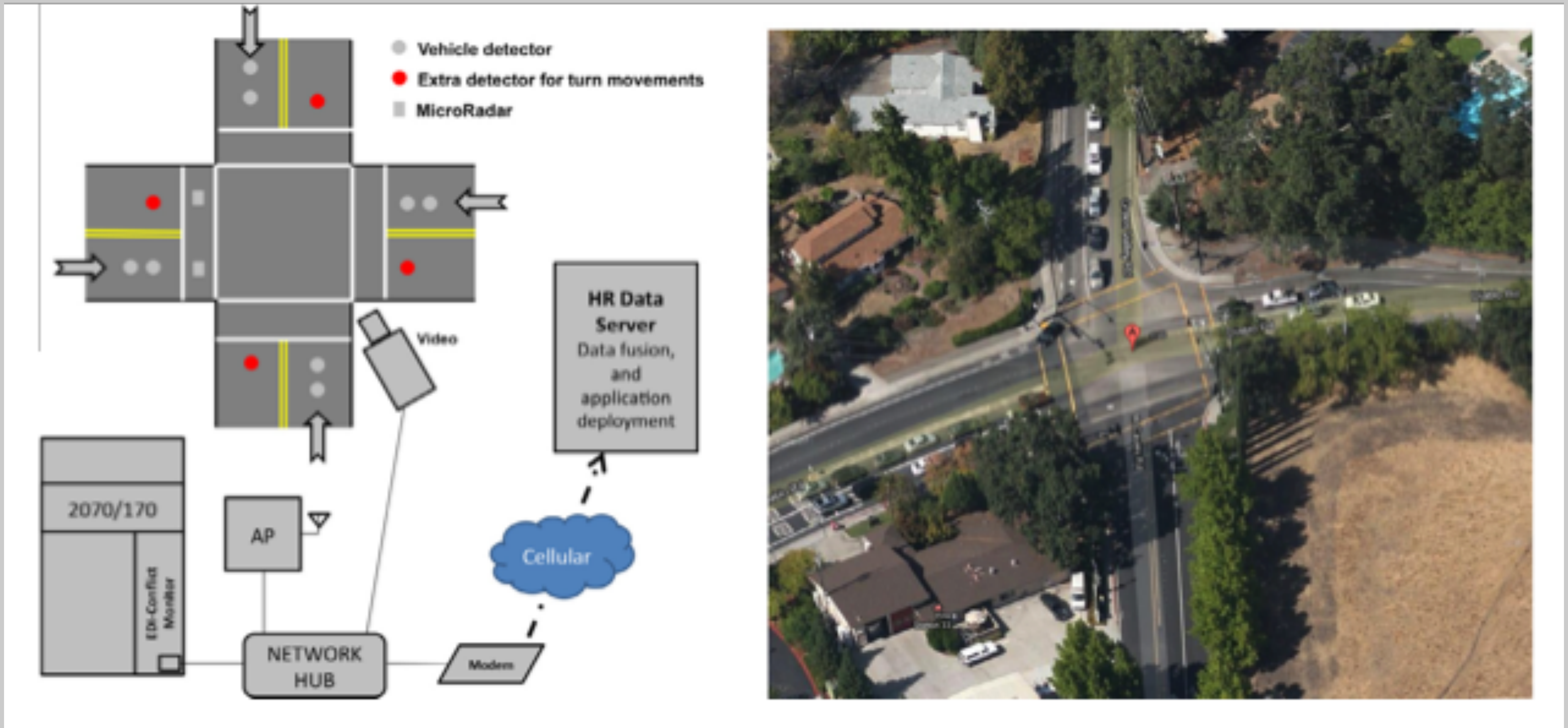
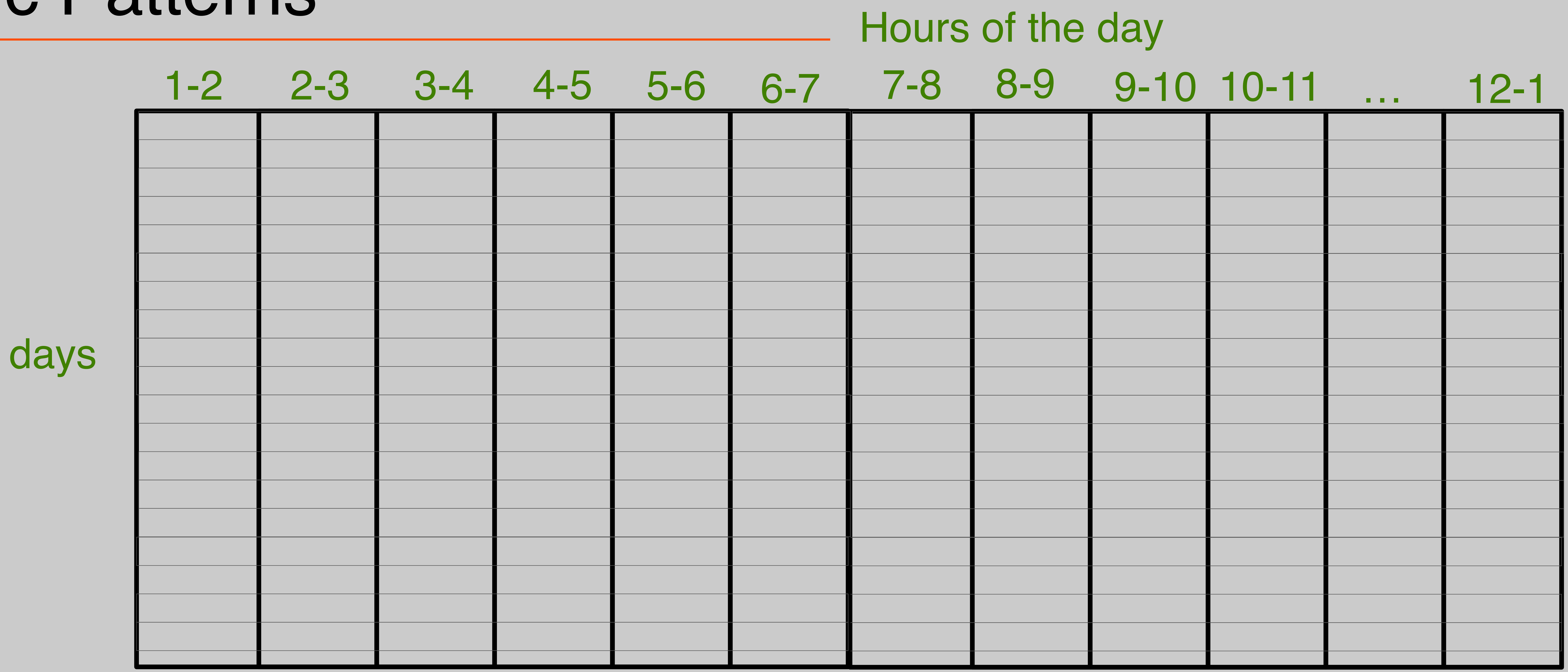


Fig. 1. HR system in Danville, CA.

Traffic Patterns



What would k-means cluster to?

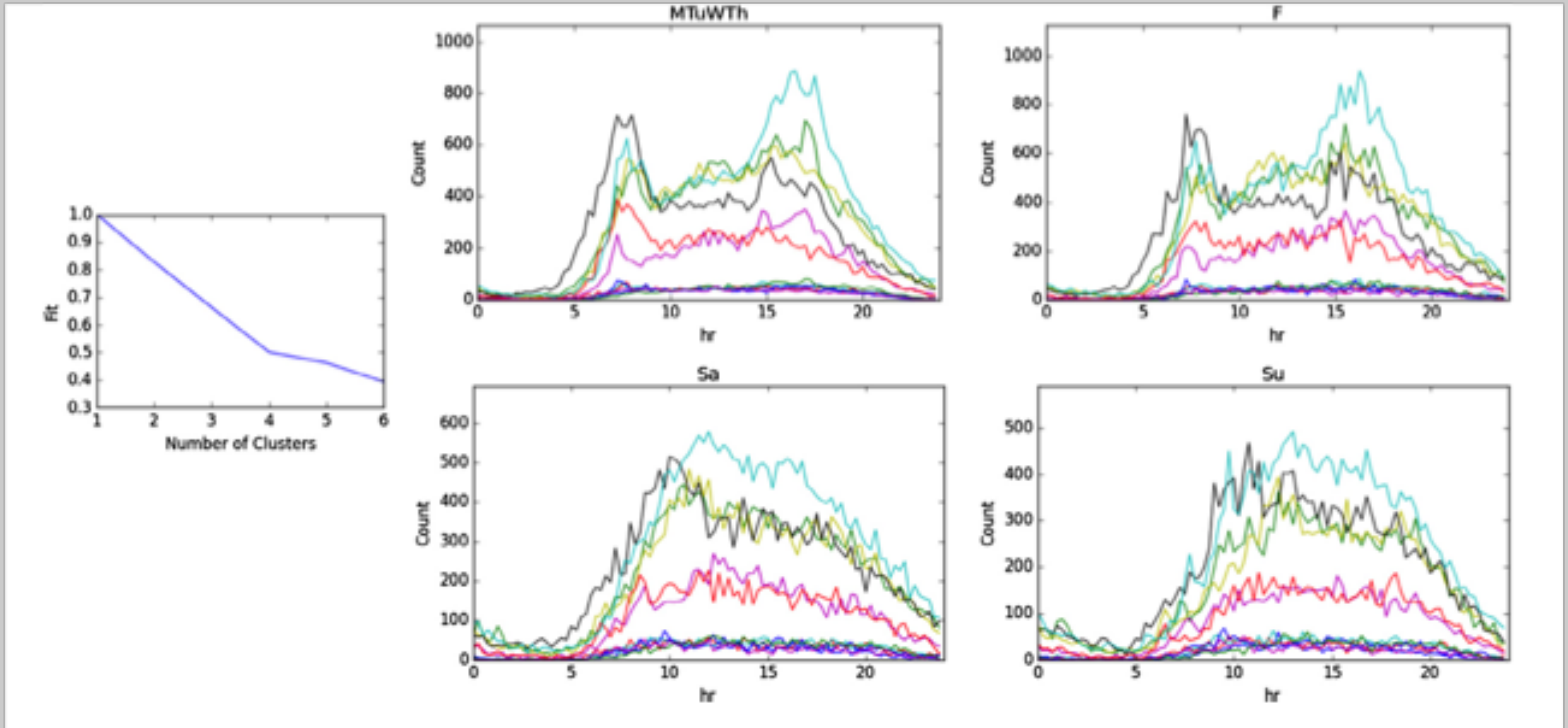


Fig. 5. Clustering of daily data for Dec 2014 to May 2015 in an intersection in Beaufort, SC.

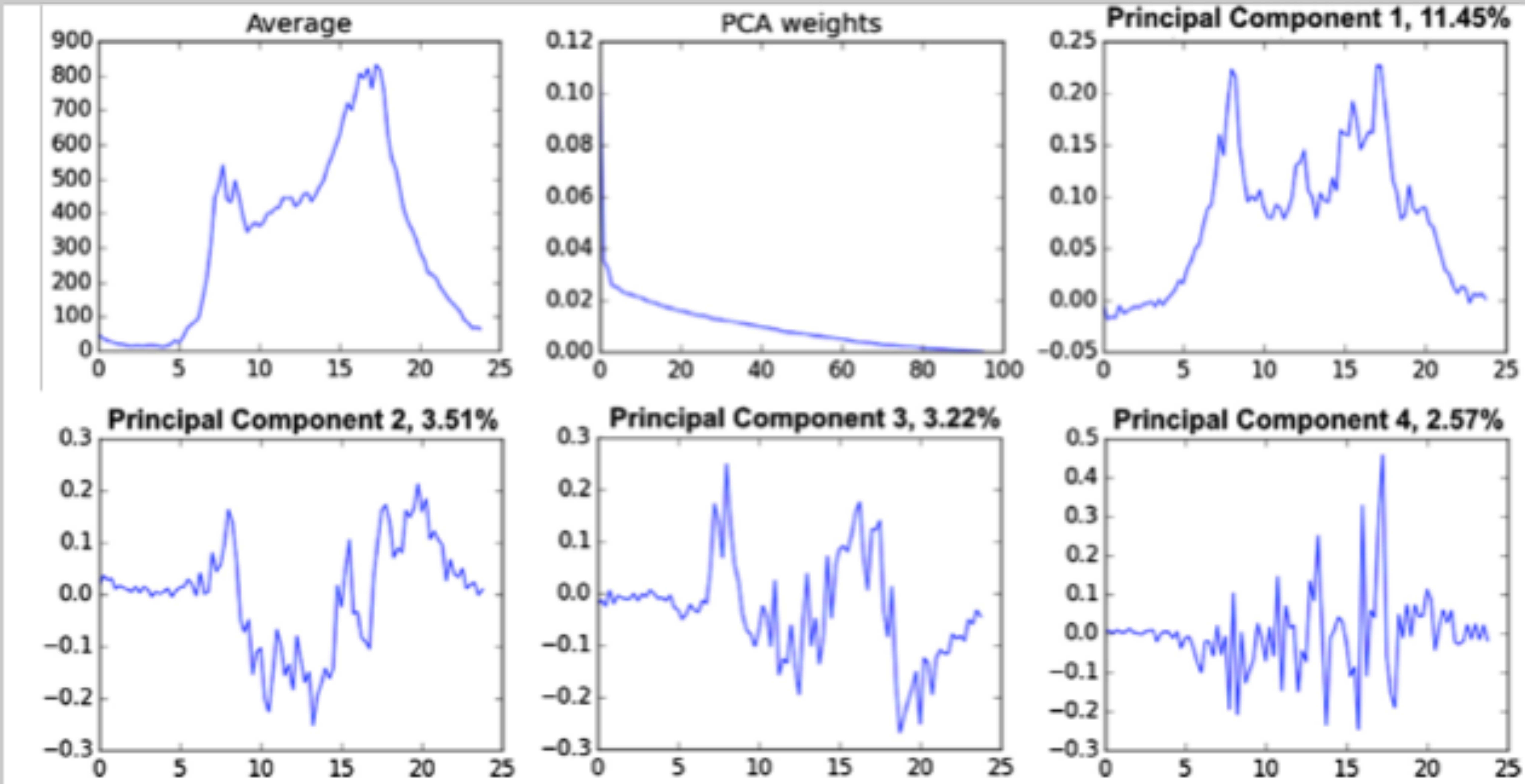


Fig. 8. Four PCA components of the North–South through movement and the average. The *x*-axis is hours.