EECS16B
Designing Information Devices and Systems II

Lecture 14B
Principal Component Analysis

Data Analysis with SVD

$$A \approx \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_1 \vec{v}_1^T + \cdots + \sigma_{\hat{r}} \vec{u}_{\hat{r}} \vec{v}_{\hat{r}}^T$$

effective

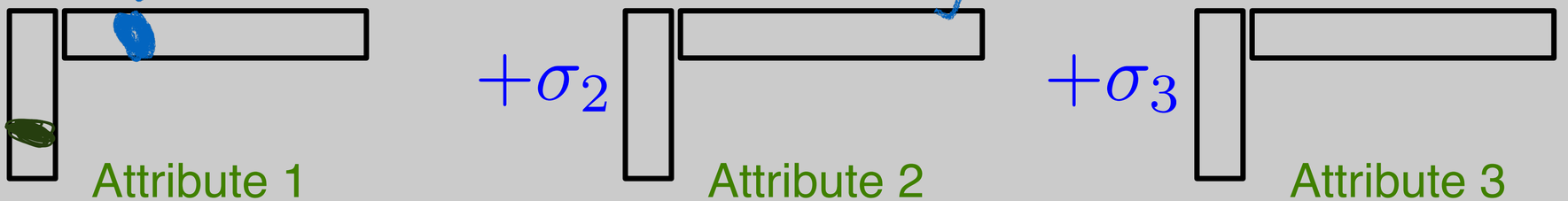
n movies ratings

m viewers

1	2	5	5	1	3	3	3	2	5	5	4	4	3	2	2	5	1	1
5	3	2	1	1	3	3	3	1	1	2	4	5	5	4	4	1	3	2
5	1	2	1	1	2	3	3	1	1	2	1	5	5	3	5	1	1	2
5	3	2	1	1	3	2	3	1	1	2	4	1	1	4	4	5	1	5
1	2	3	2	1	3	2	3	2	1	2	1	1	1	4	4	5	1	5
1	1	1	1	5	3	3	3	1	5	2	4	4	4	2	5	5	1	1

how much this movie is Fanny

$\approx \sigma_1$ $+\sigma_2$ $+\sigma_3$



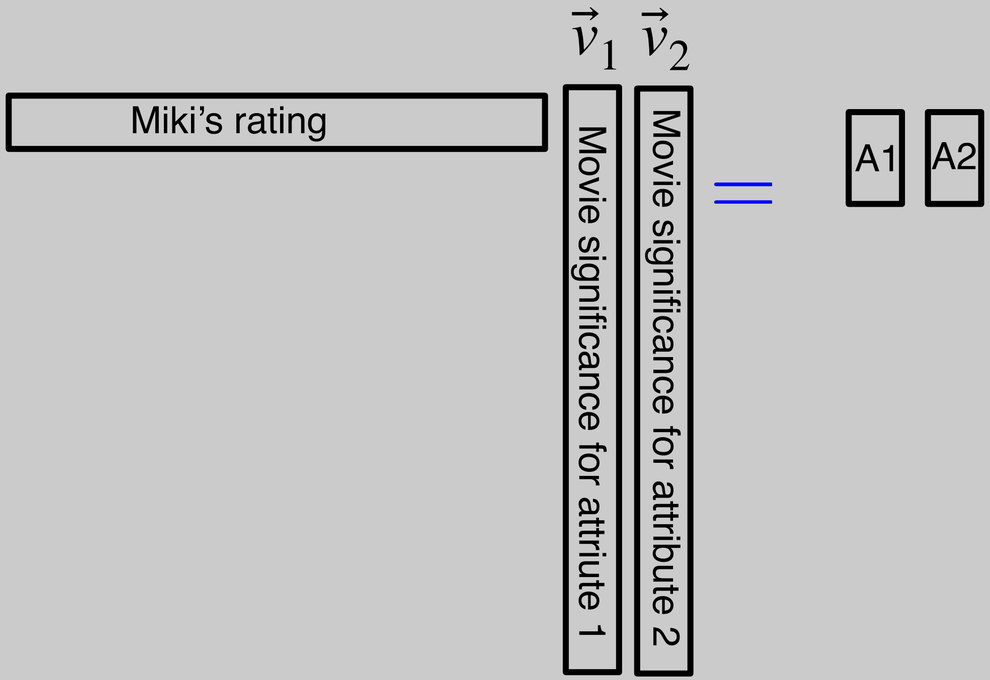
Attribute 1 “Funny”

Attribute 2 “Scary”

Attribute 3 “Dramatic”

how much Miki likes Fanny

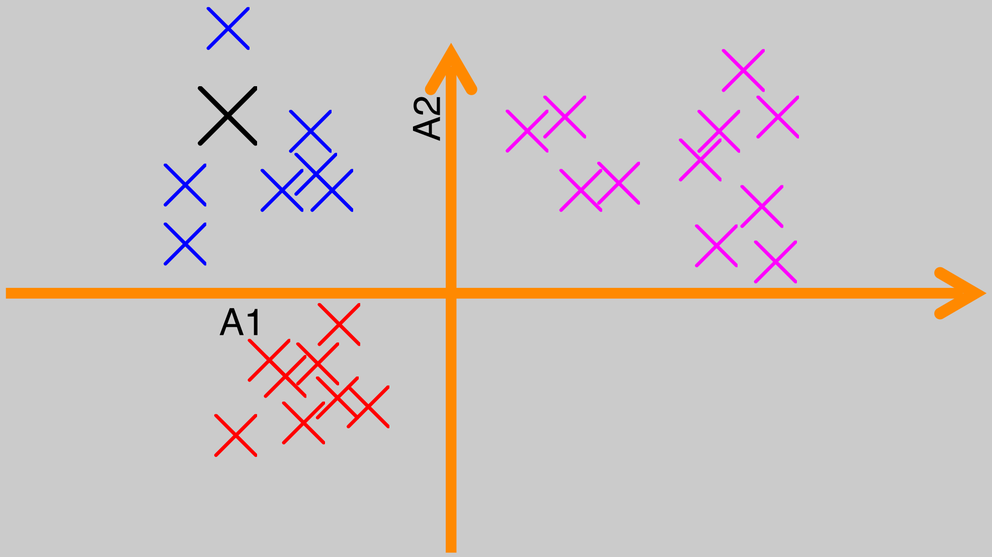
Classification with SVD



n movies

m viewers

1	2	5	5	1	3	3	3	2	5	5	4	4	3	2	2	5	1	1
5	3	2	1	1	3	3	3	1	1	2	4	5	5	4	4	1	3	2
5	1	2	1	1	2	3	3	1	1	2	1	5	5	3	5	1	1	2
5	3	2	1	1	3	2	3	1	1	2	4	1	1	4	4	5	1	5
1	2	3	2	1	3	2	3	2	1	2	1	1	1	4	4	5	1	5
1	1	1	1	5	3	3	3	1	5	2	4	4	4	2	5	5	1	1



Miki belongs to class: like A2 don't like A1

Classification with SVD

n movies

m viewers

1	2	5	5	1	3	3	3	2	5	5	4	4	3	2	2	5	1	1
5	3	2	1	1	3	3	3	1	1	2	4	5	5	4	4	1	3	2
5	1	2	1	1	2	3	3	1	1	2	1	5	5	3	5	1	1	2
5	3	2	1	1	3	2	3	1	1	2	4	1	1	4	4	5	1	5
1	2	3	2	1	3	2	3	2	1	2	1	1	1	4	4	5	1	5
1	1	1	1	5	3	3	3	1	5	2	4	4	4	2	5	5	1	1

$\vec{u}_1^T$

Viewer significance for attriute 1

$\vec{u}_2^T$

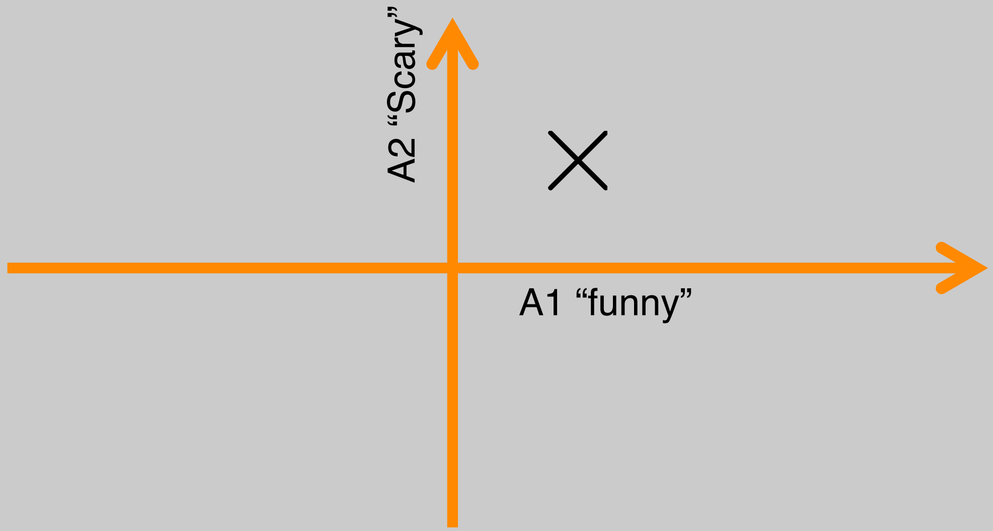
Viewer significance for attriute 2

Star Wars Ratings

=

A1

A2



Star Wars is somewhat funny and somewhat scary

Prediction with SVD

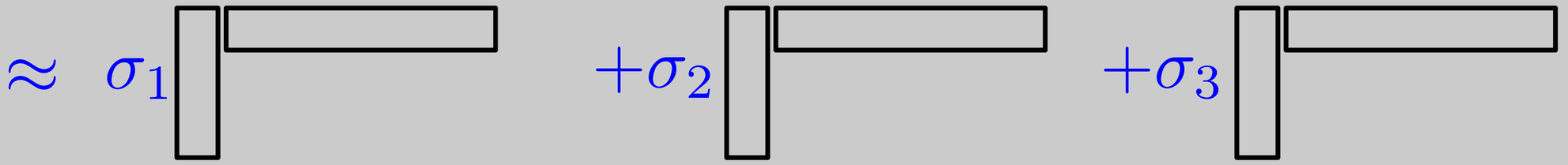
Can try to predict preferences of a new customer with few ratings

See homework!

n movies

m views

1	2	5	5	1	3	3	3	2	5	5	4	4	3	2	2	5	1	1
5	3	2	1	1	3	3	3	1	1	2	4	5	5	4	4	1	3	2
5	1	2	1	1	2	3	3	1	1	2	1	5	5	3	5	1	1	2
5	3	2	1	1	3	2	3	1	1	2	4	1	1	4	4	5	1	5
1	2	3	2	1	3	2	3	2	1	2	1	1	1	4	4	5	1	5
1	1	1	1	5	3	3	3	1	5	2	4	4	4	2	5	5	1	1
1	?	?	2	?	?	?	?	3	5	1	?	?	?	?	5	2	?	3



Attribute 1

Attribute 2

Attribute 3

Low-rank Completion

What if my database is full of “holes”?

Should be still low-rank!

n movies

m views

1		5	5	1	3		3	2	5	5	4		3	2	2	5		1	
5	3		1	1	3	3		1	1	2		5		4	4	1	3	2	
	1	2	1	1		3	3		1	2	1	5	5	3	5		1	2	
5	3		1		3	2		1		2	4	1		4		5		5	
1		3	2	1	3	2	3	2	1		1		1	4		5	1	5	
1	1			5	3	3		5	2	4			2	5	5	1	1		

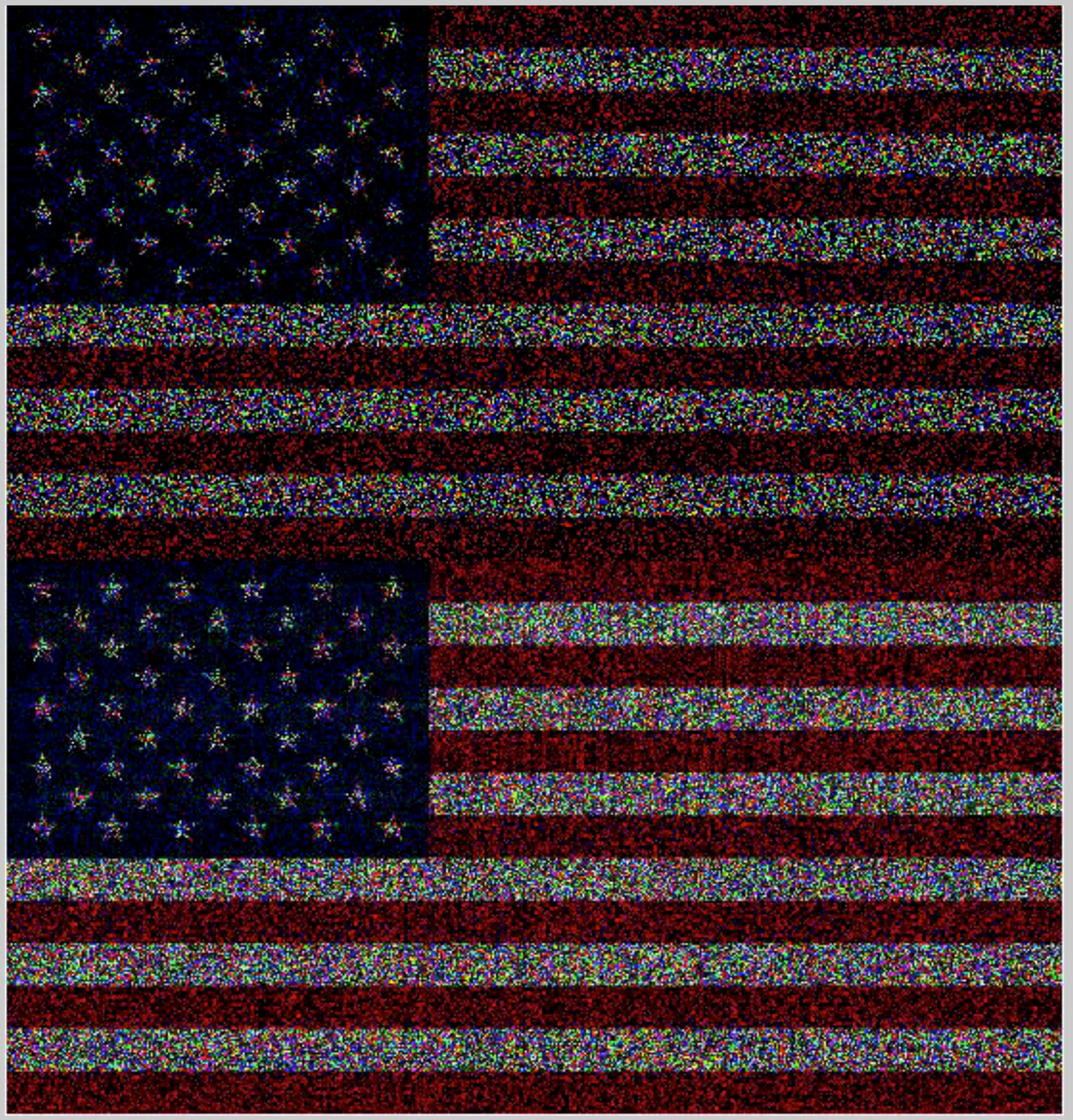
Q) Can we complete missing data?

A) Sometimes! Very recent mathematical and practical results show you can.
Keywords: Compressed Sensing, Low-rank completion, robust PCA

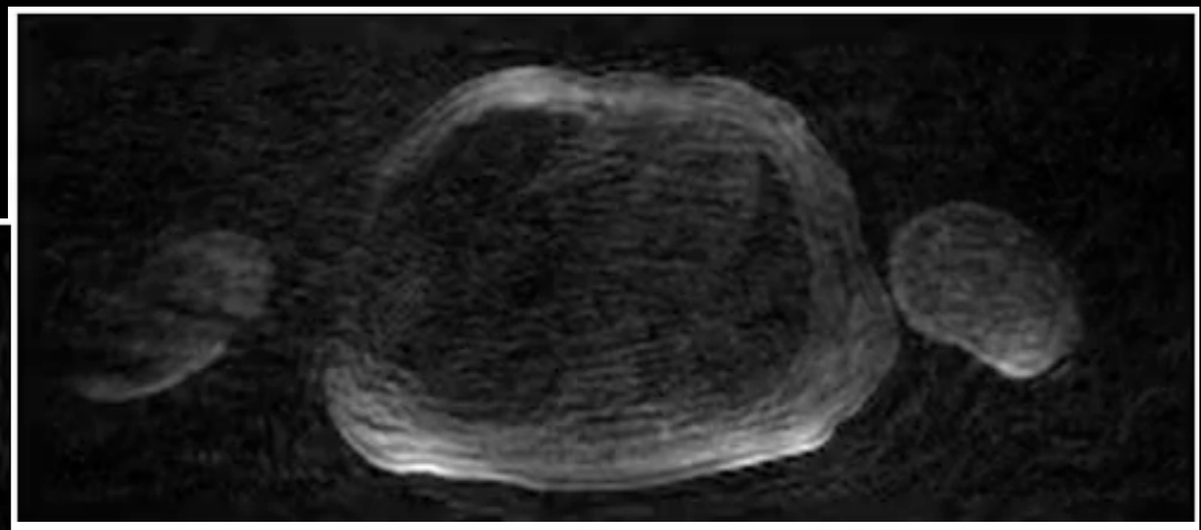
E. Candes and B. Recht, Foundation of Computational Mathematics, 2009;9:717

Low-Rank Recovery from 20% pixels

- Algorithm for low-rank completion:
  - flag_hat = flag
  - Compute [U,S,V] = svd(flag_hat)
  - $$\text{flag}_{\text{hat}} = \sum_{i=0}^{6} \sigma_i \vec{u}_i \vec{v}_i^T$$
  - update missing pixels in flag from flag_hat
  - repeat (250 times here)

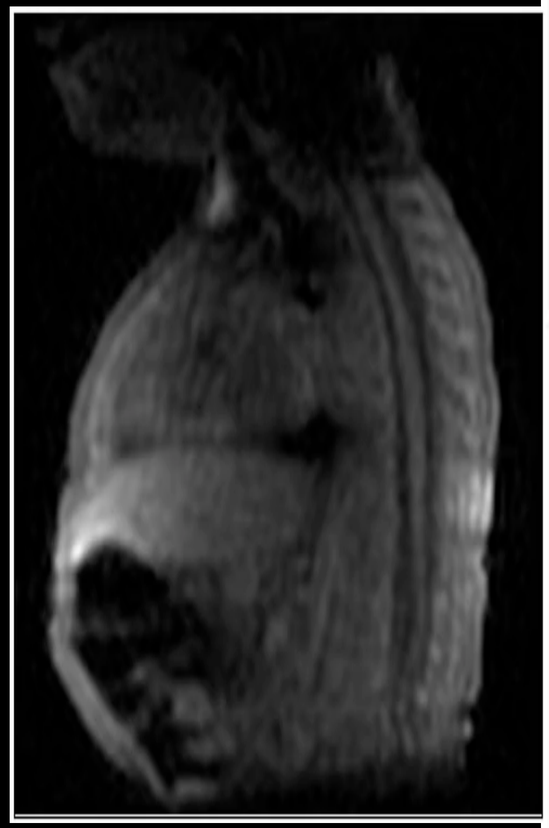


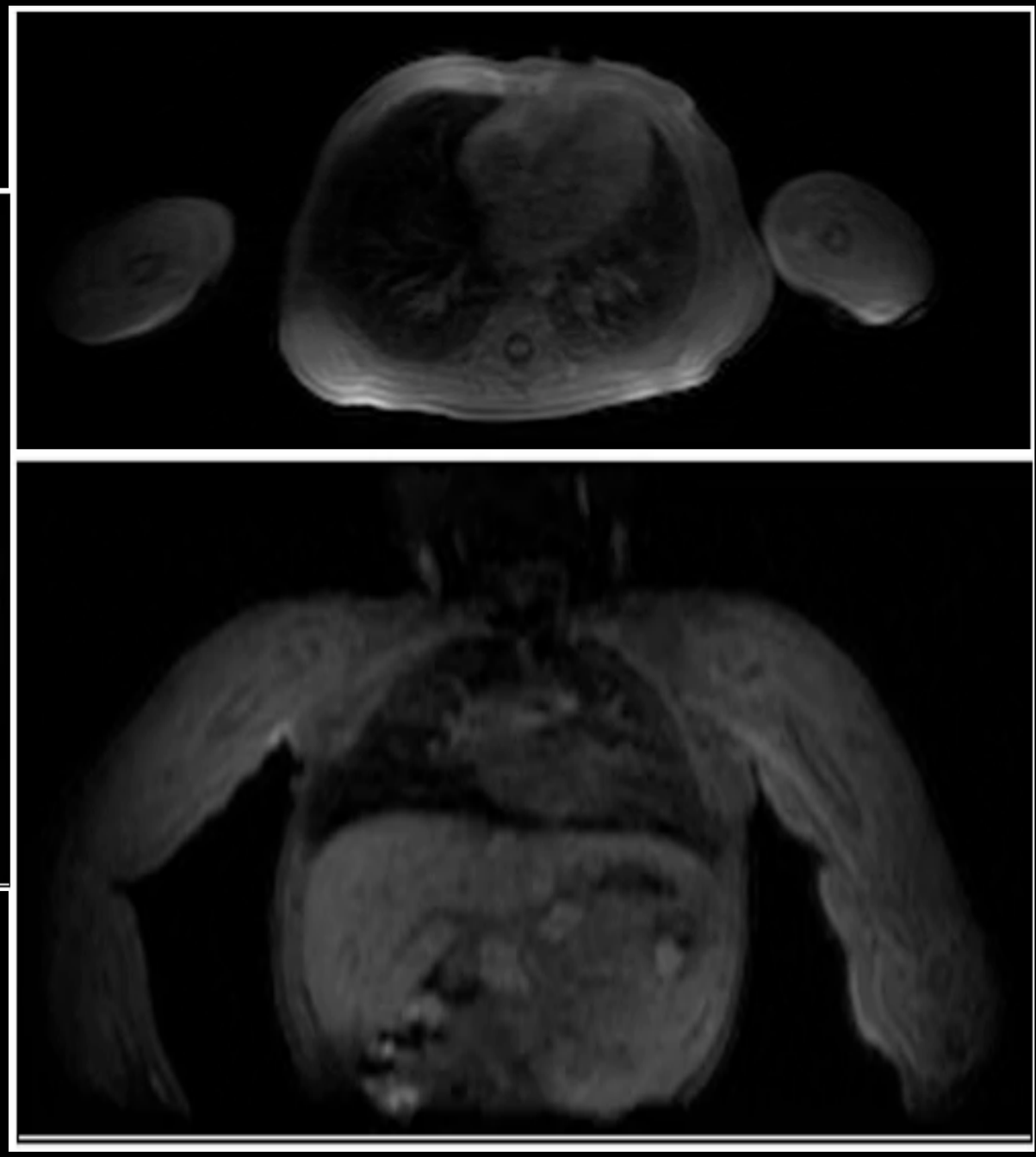
Dynamic MRI
Missing Data



~1.5 x 1.5 x 3 mm^3

Dynamic MRI Low-Rank



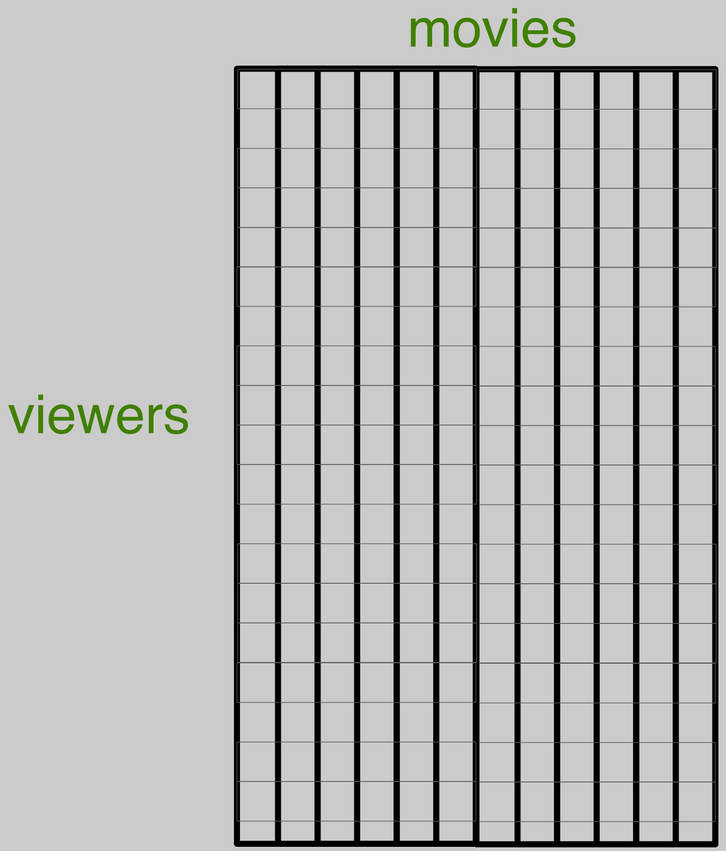


~1.5 x 1.5 x 3 mm^3

Principal Component Analysis

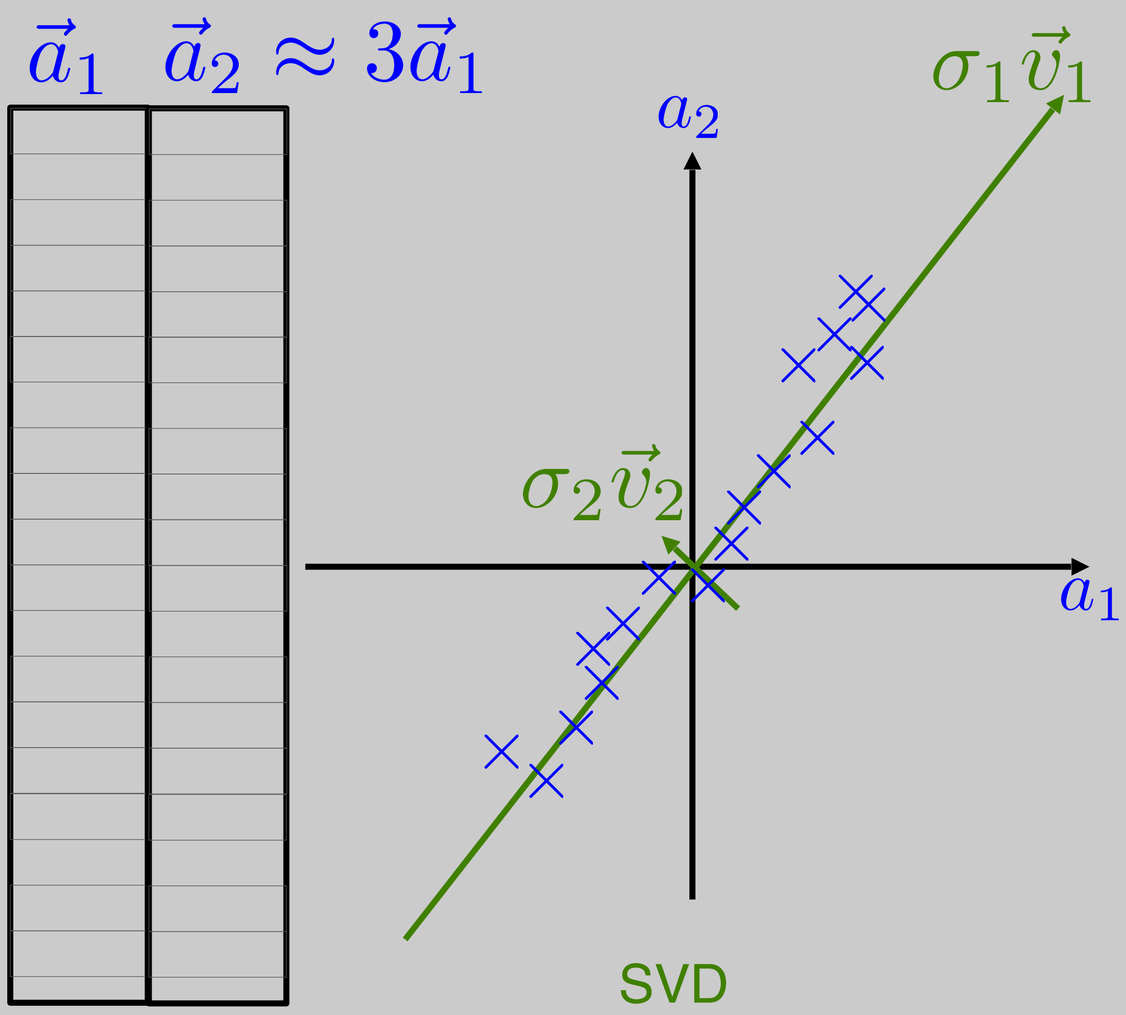
Application of the SVD to datasets to learn features

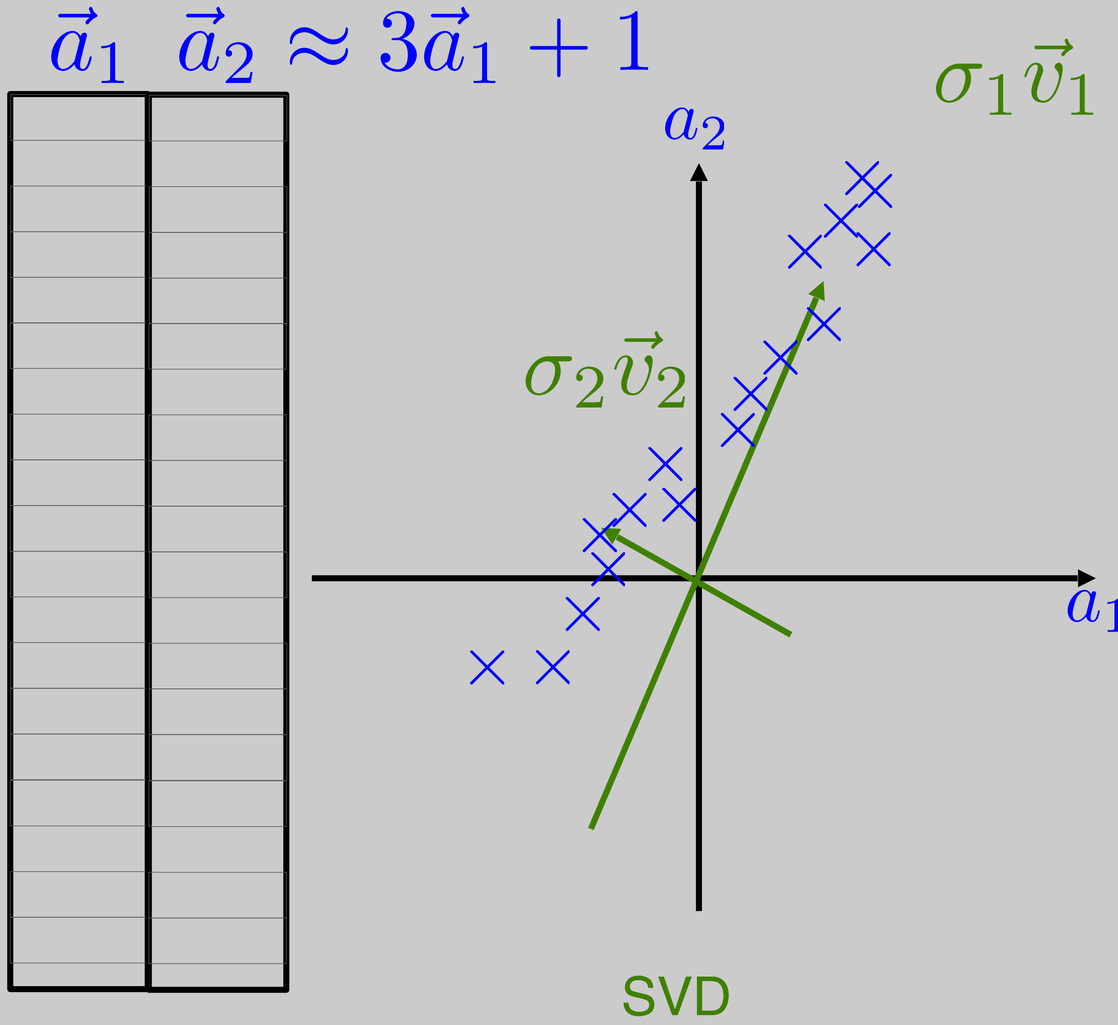
PCA is a tool in statistics and machine learning, which can be computed using SVD



Example -- PCA

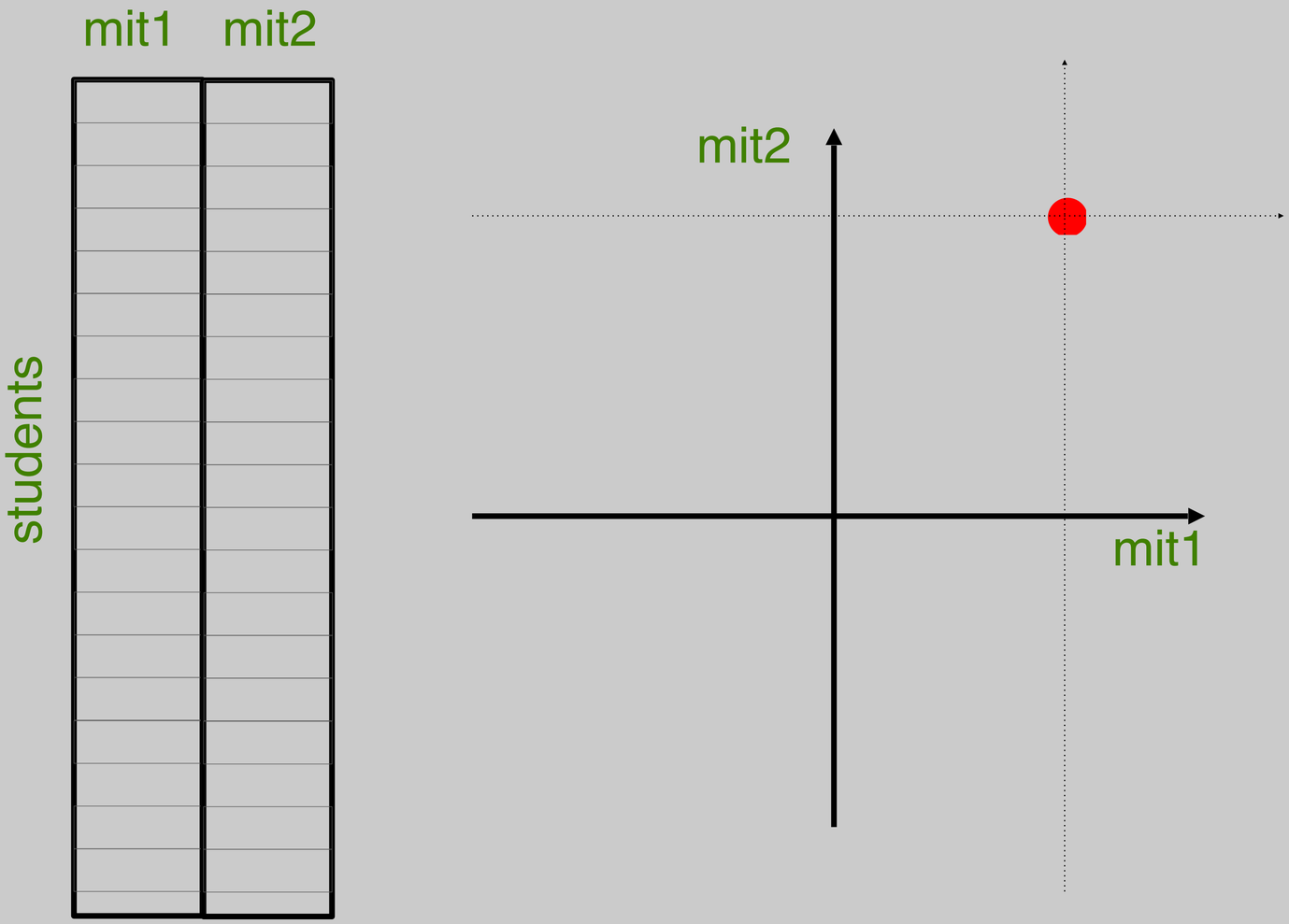
Consider data s.t.





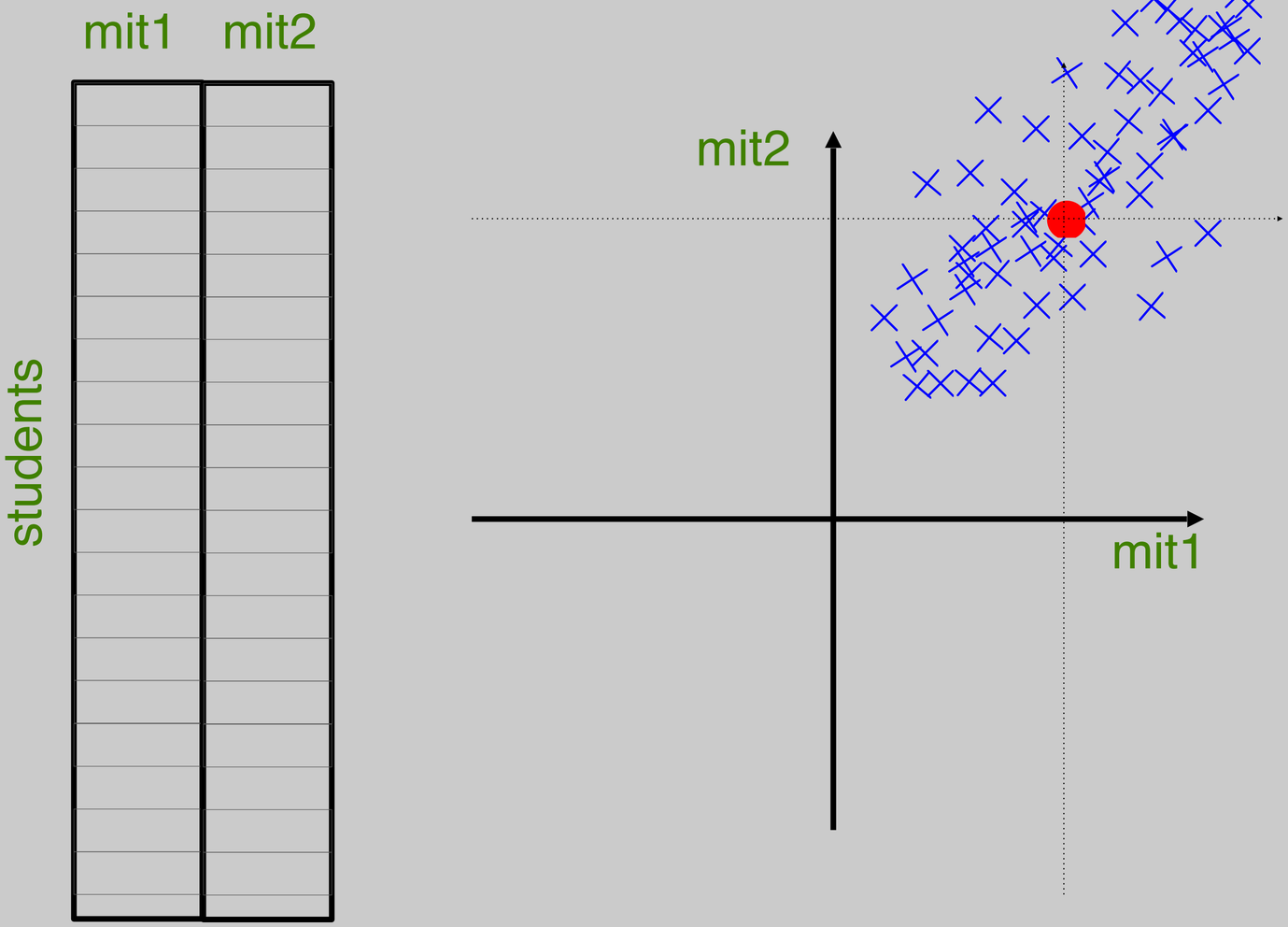
Example -- PCA

Consider miterm data



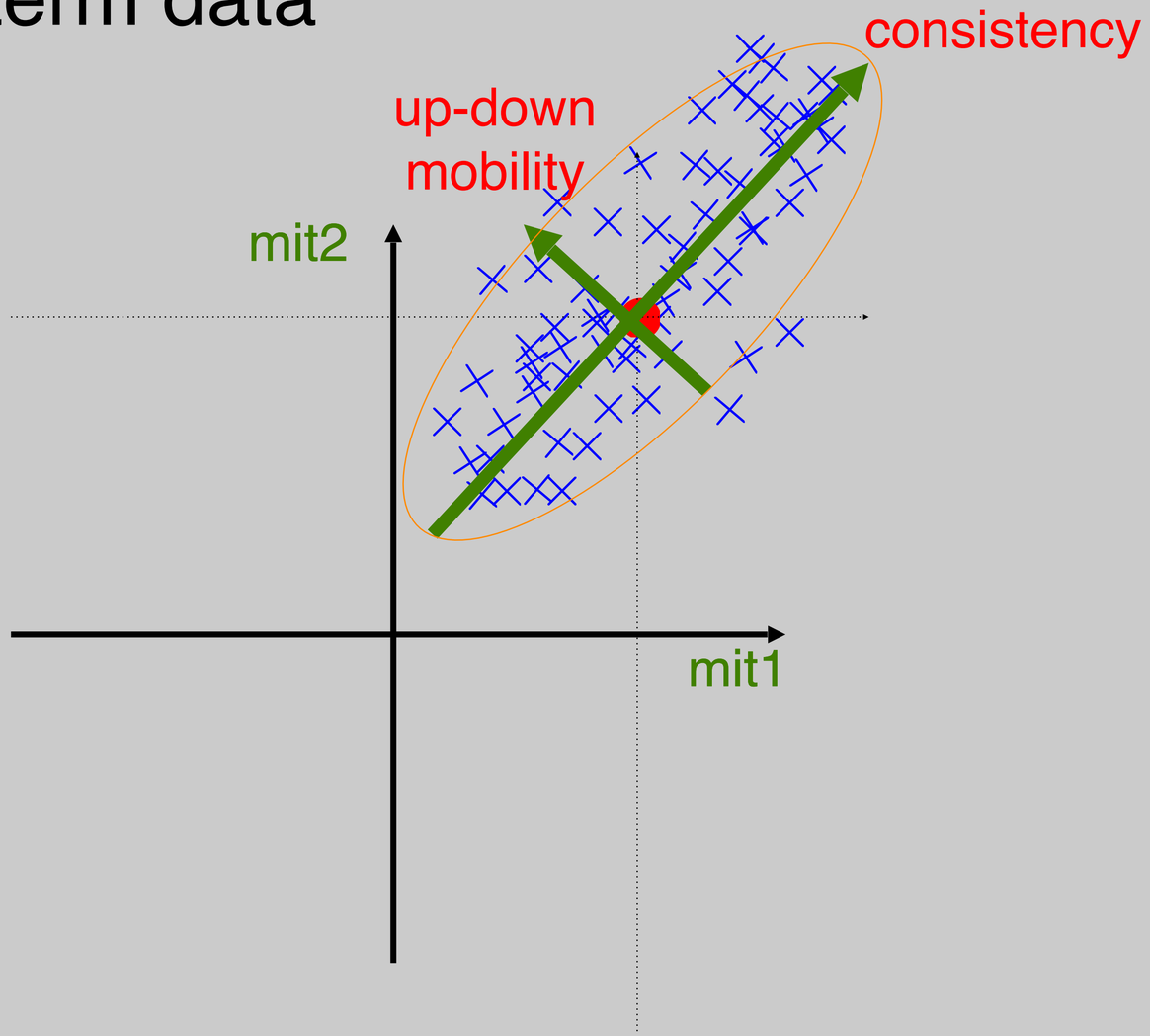
Example -- PCA

Consider miterm data



Example -- PCA

Consider miterm data

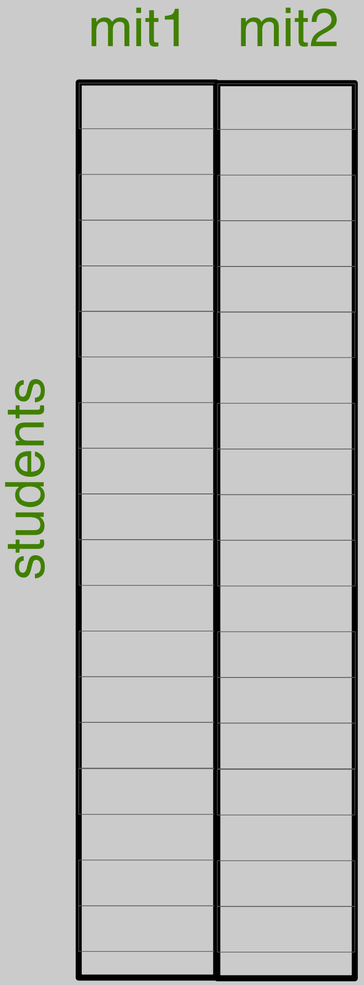


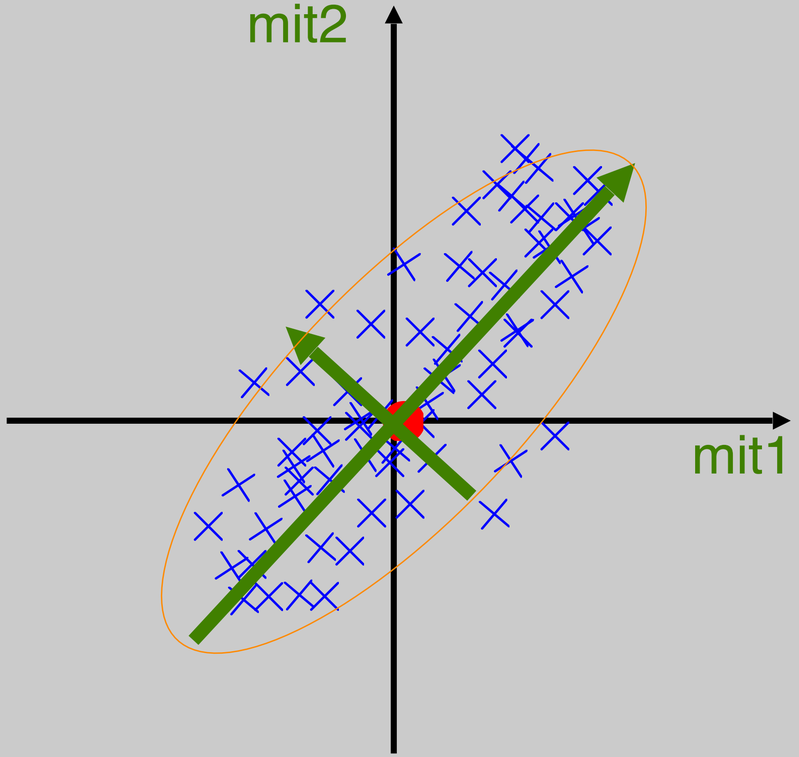
PCA Procedure

Remove averages from column of A

From A^TA, find σ_i, $\vec{v_i}$

$\vec{v_i}$ are principal components!





A^TA as sample covariance matrix

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

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

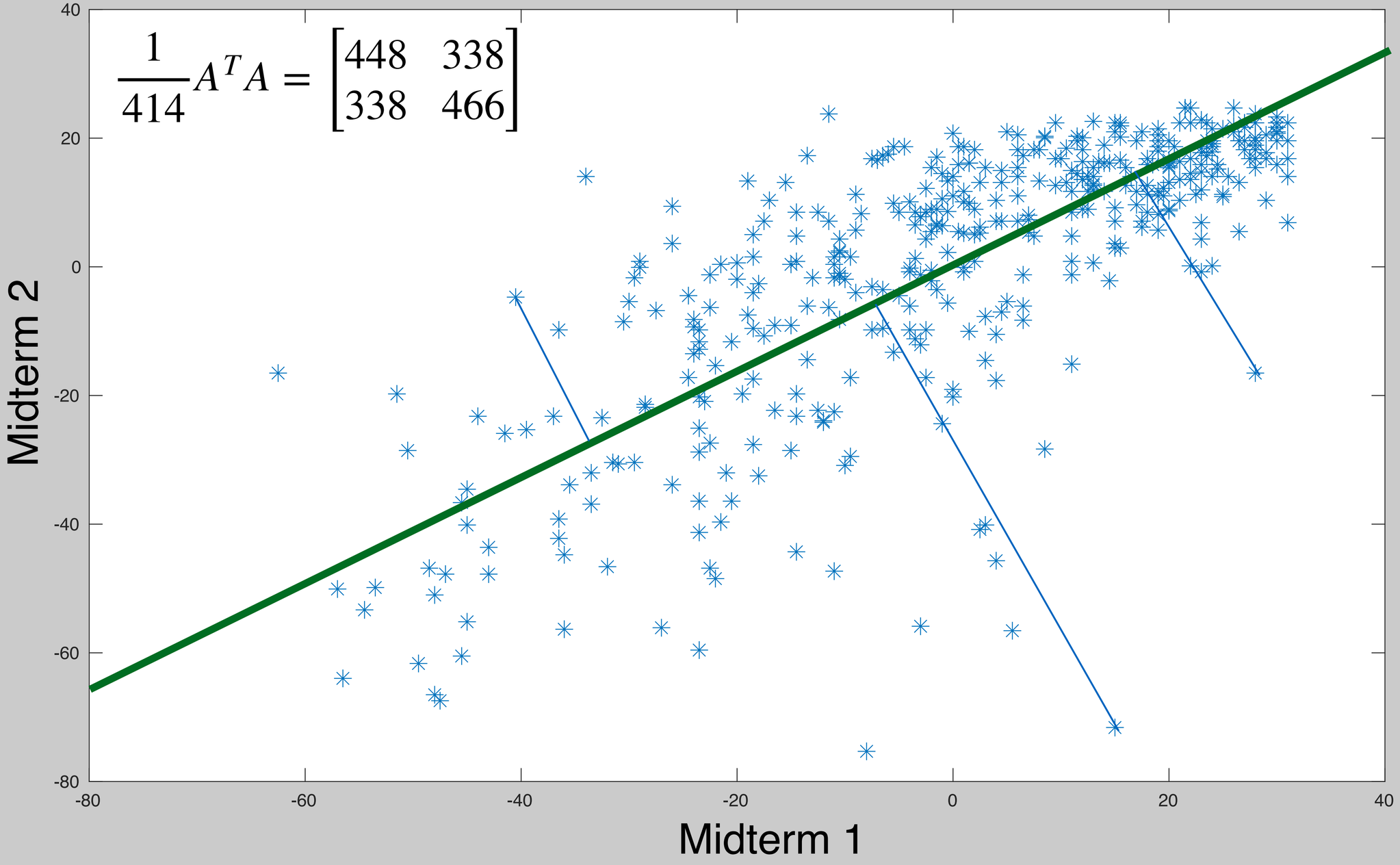
$$= \vec{a}^T\vec{a} - 2Na_\mu^2 + Na_\mu^2 \quad = \vec{a}^T\vec{a} - Na_\mu^2$$

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

Sample Variance!

Example midterm

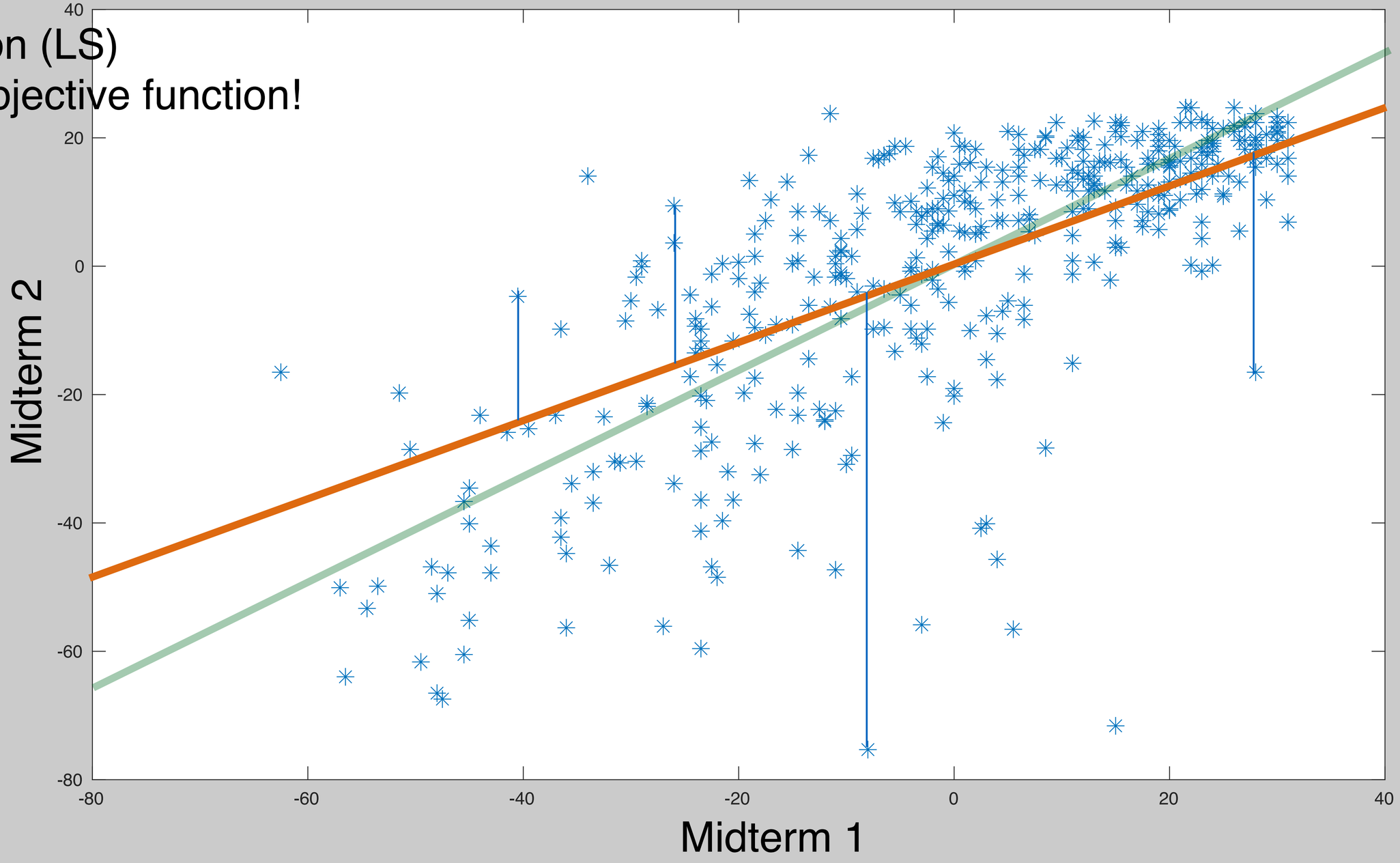
$$\frac{1}{414}A^T A = \begin{bmatrix} 448 & 338 \\ 338 & 466 \end{bmatrix}$$



Example midterm

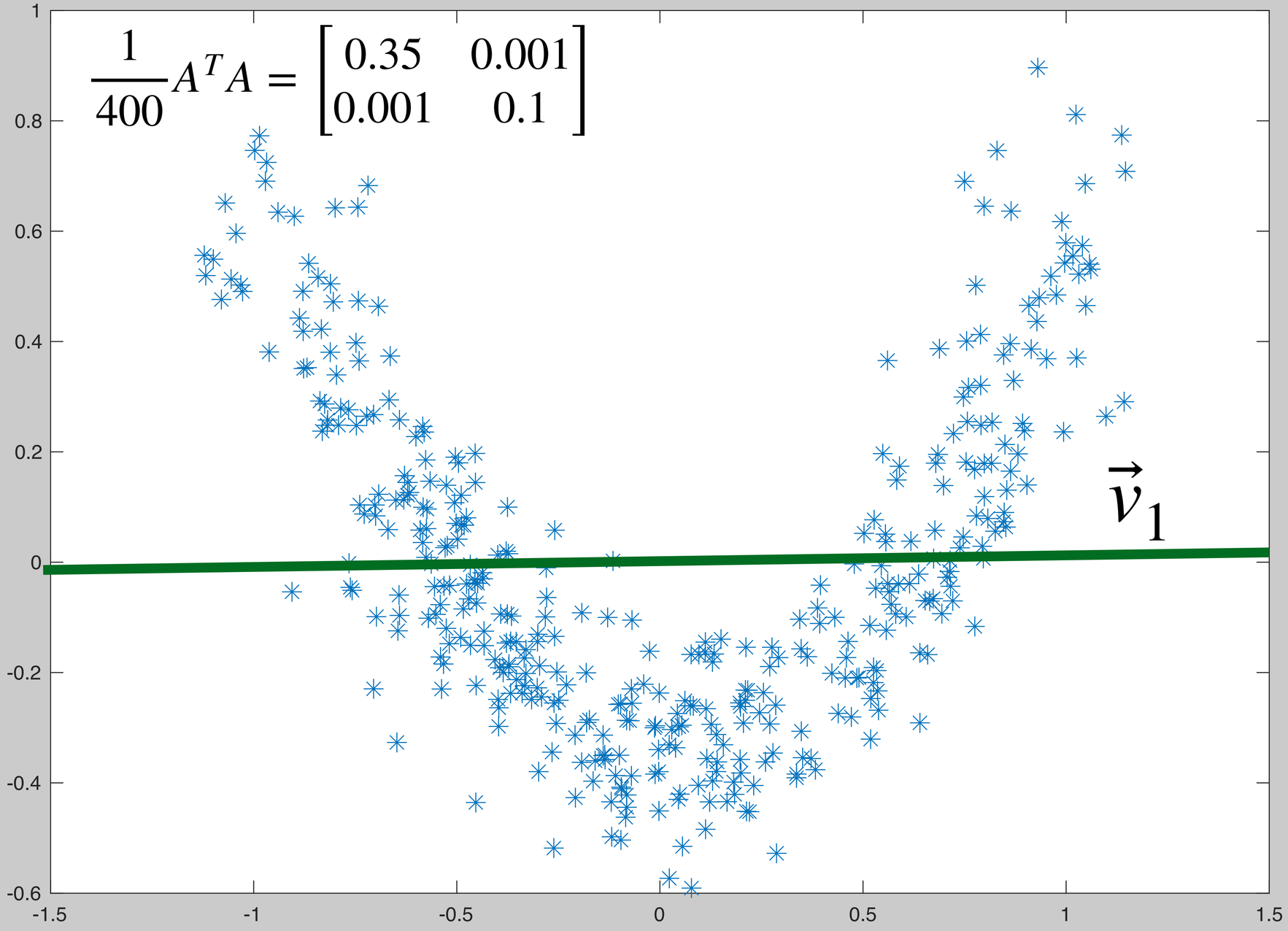
Linear regression (LS)

Not the same objective function!



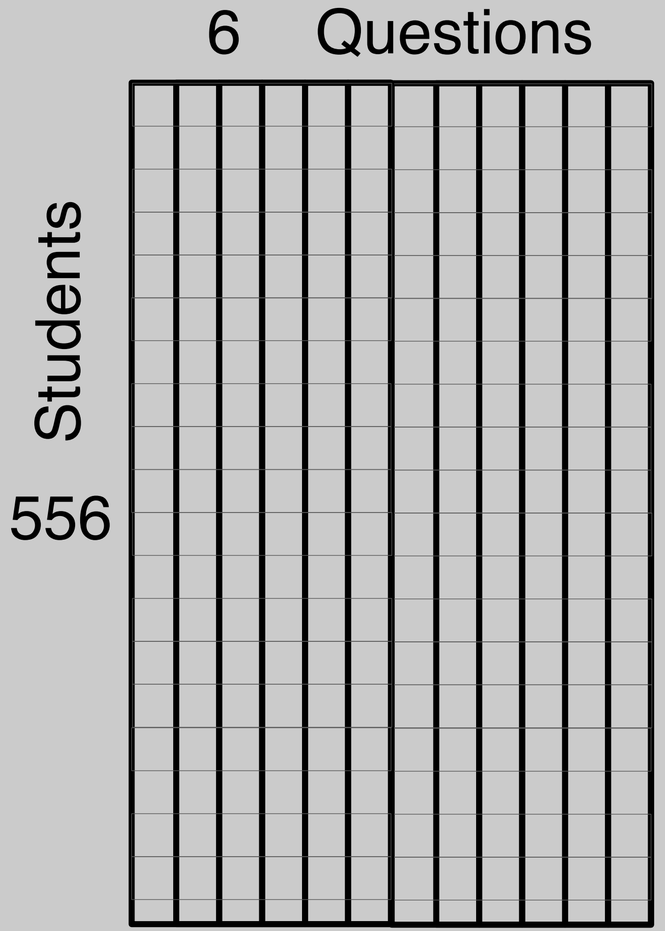
PCA captures linear correlations

Linear subspace does not capture the low dimensionality of a one dimensional manifold



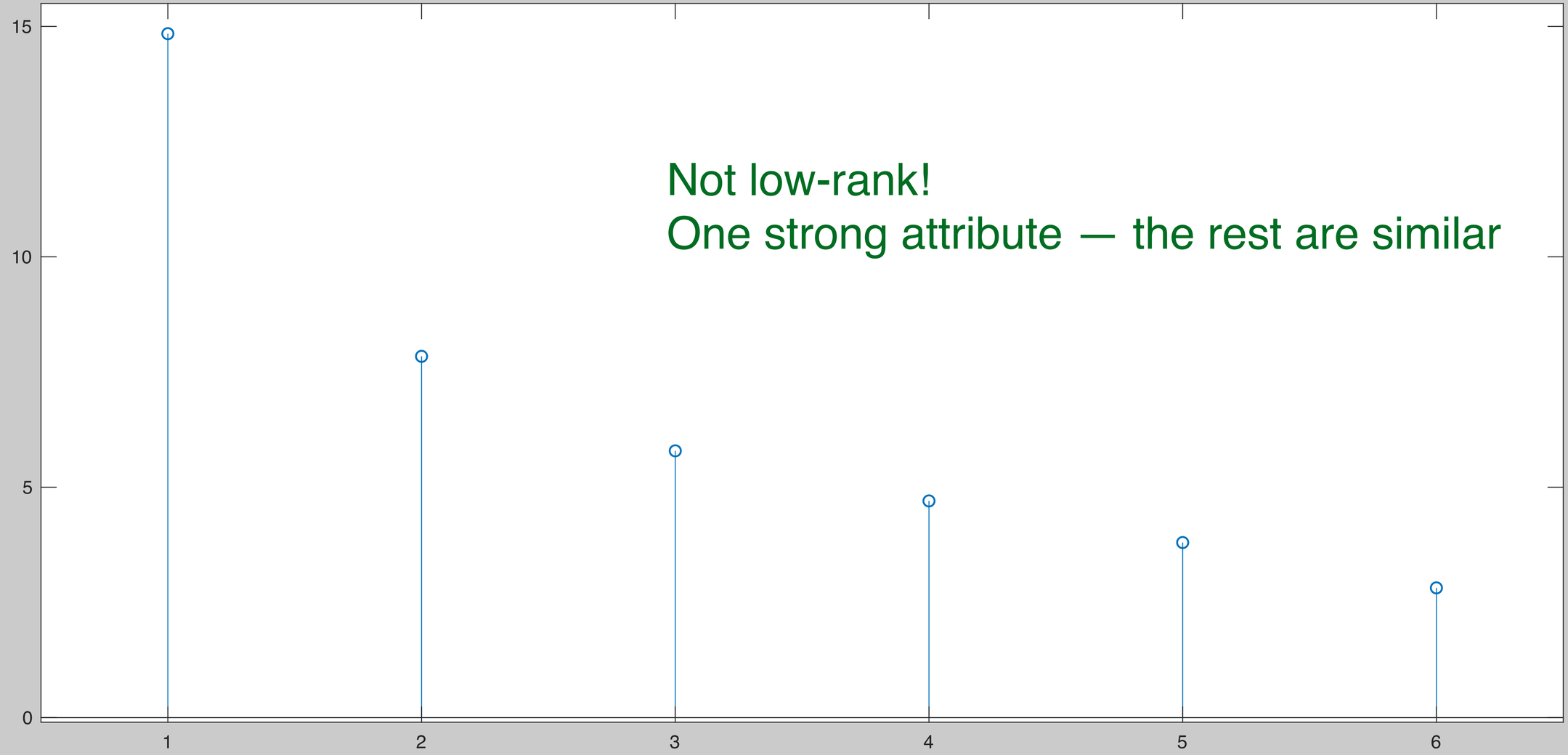
Mid Semester Survey Results (2020)

1) HW difficulty
2) HW Length
3) Lab hour/week
4) Class rating
5) Class expectation rating
6) Comfortable attending OH



Data Science

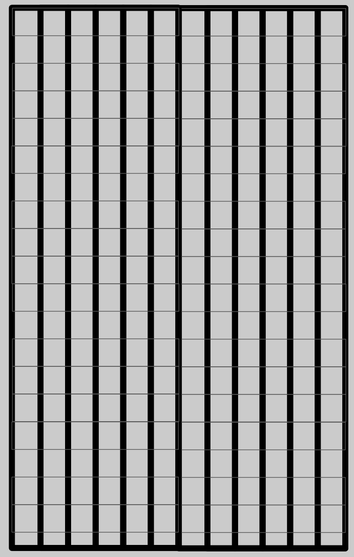
Singular values

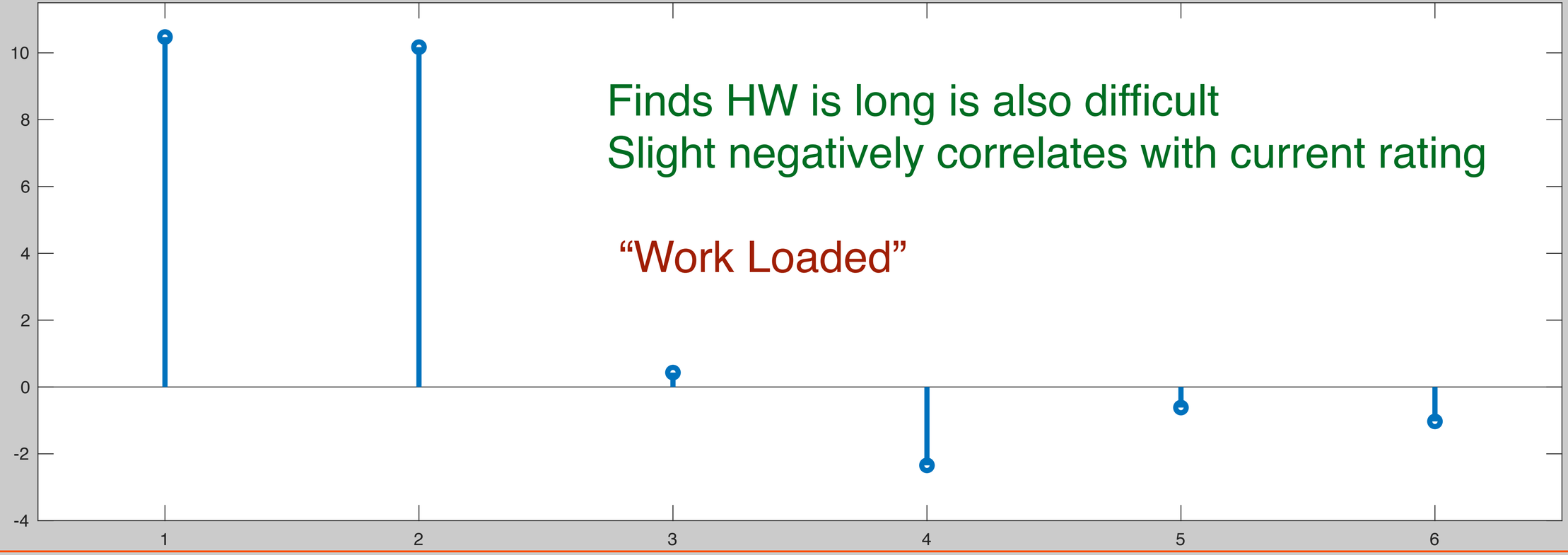


Data Science

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

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





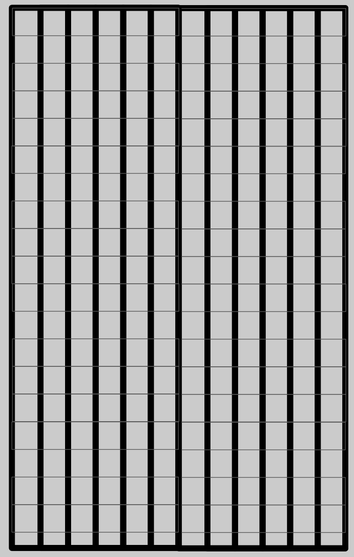
Finds HW is long is also difficult
Slight negatively correlates with current rating

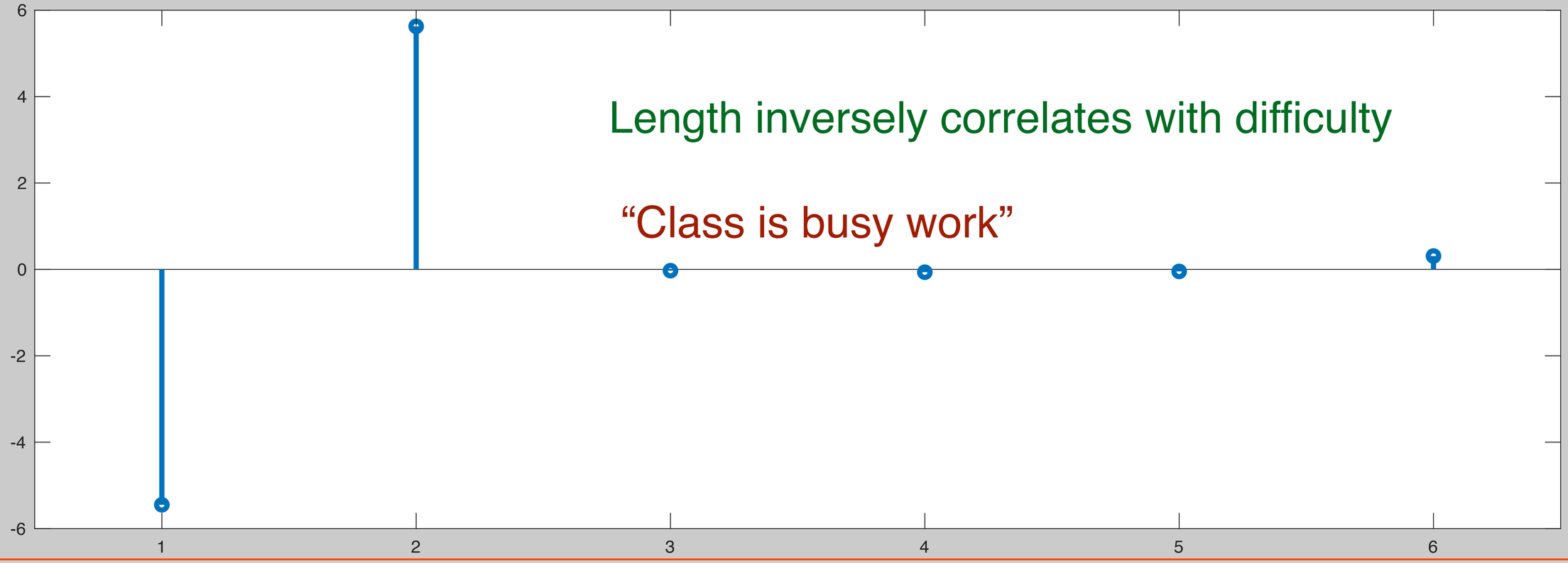
"Work Loaded"

Data Science

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

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

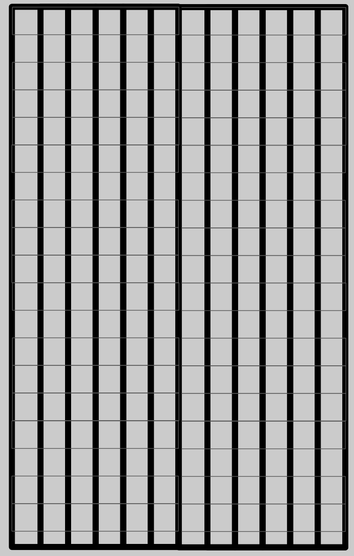


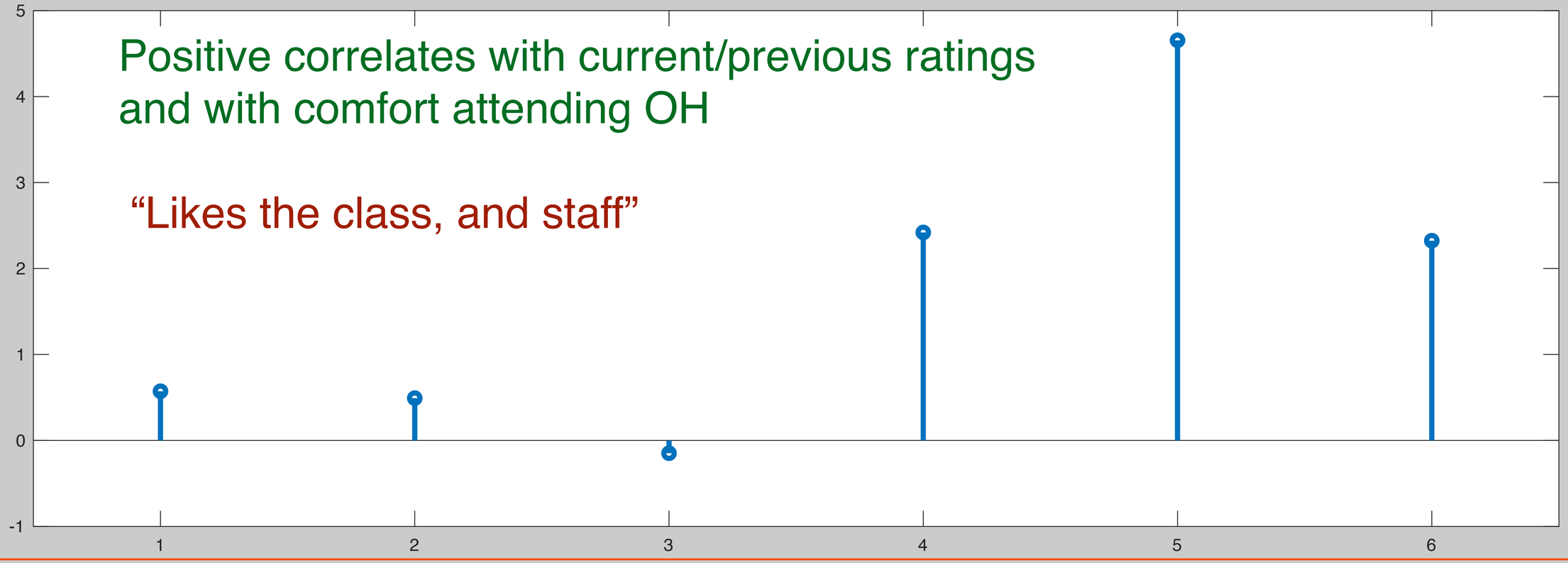


Data Science

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

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

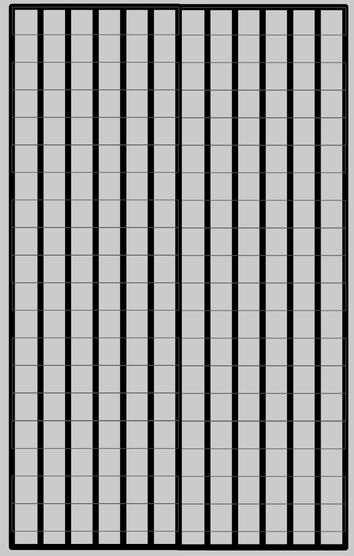




Data Science

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

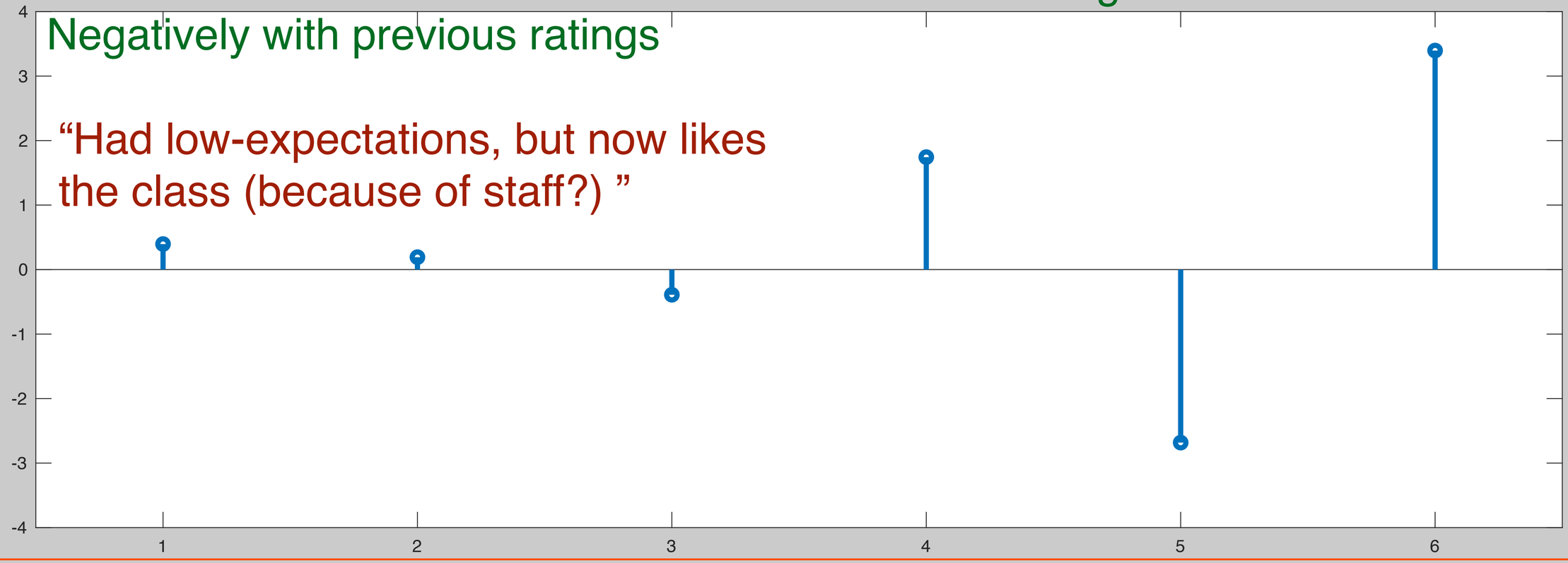
1) HW difficulty
2) HW Length
3) Lab hour/week
4) Class rating
5) Expectation Rating
6) Comfortable attending OH



Positive correlates with current & with comfort attending OH

Negatively with previous ratings

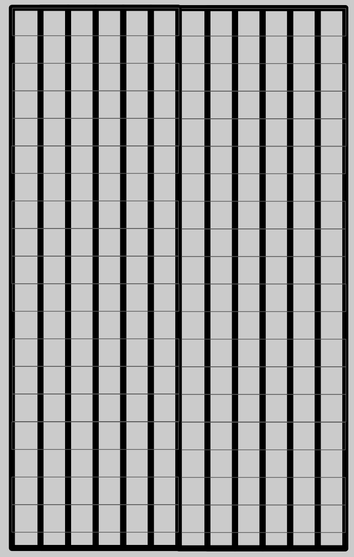
"Had low-expectations, but now likes the class (because of staff?) "

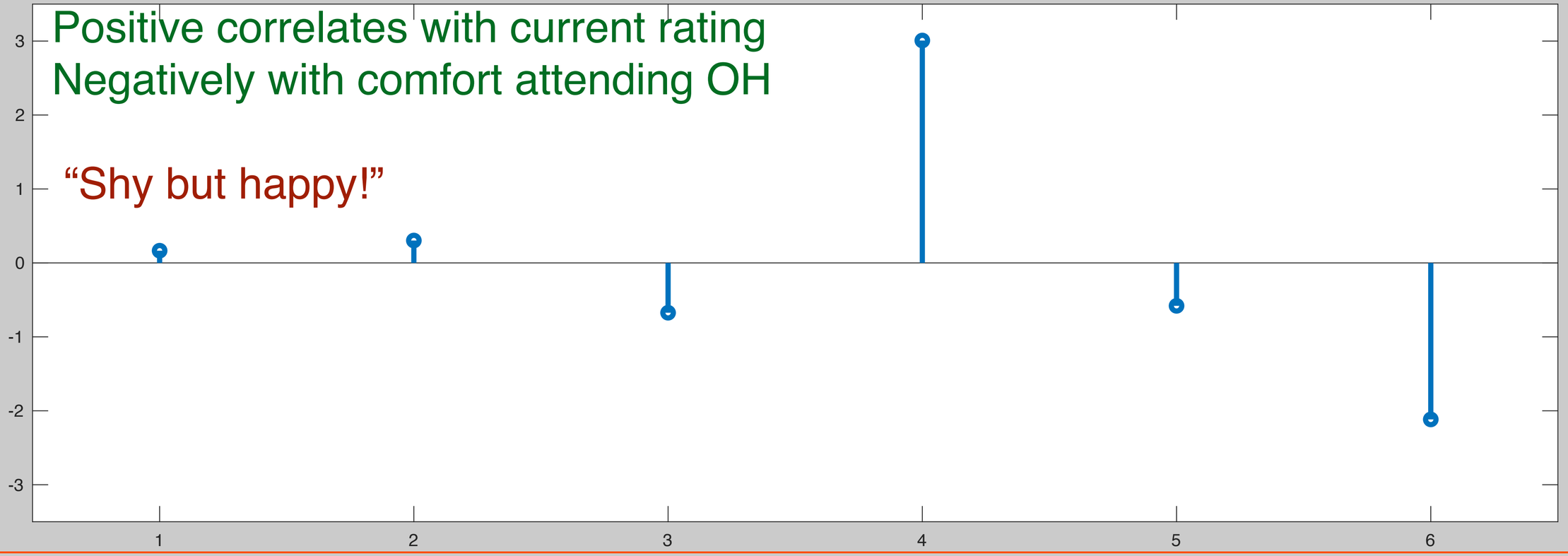


Data Science

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

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

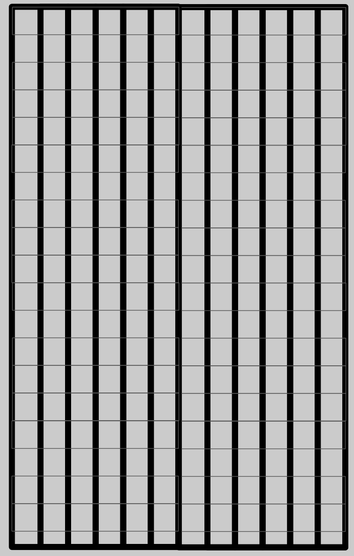


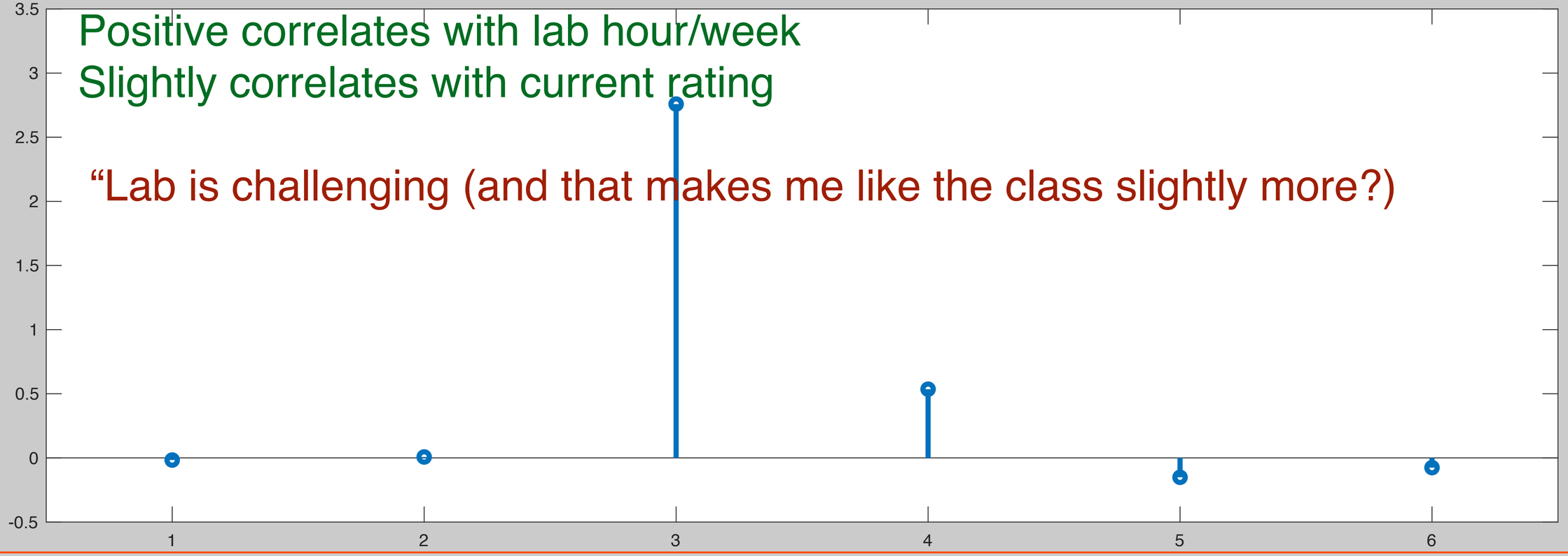


Data Science

$A^T \vec{u}_6$

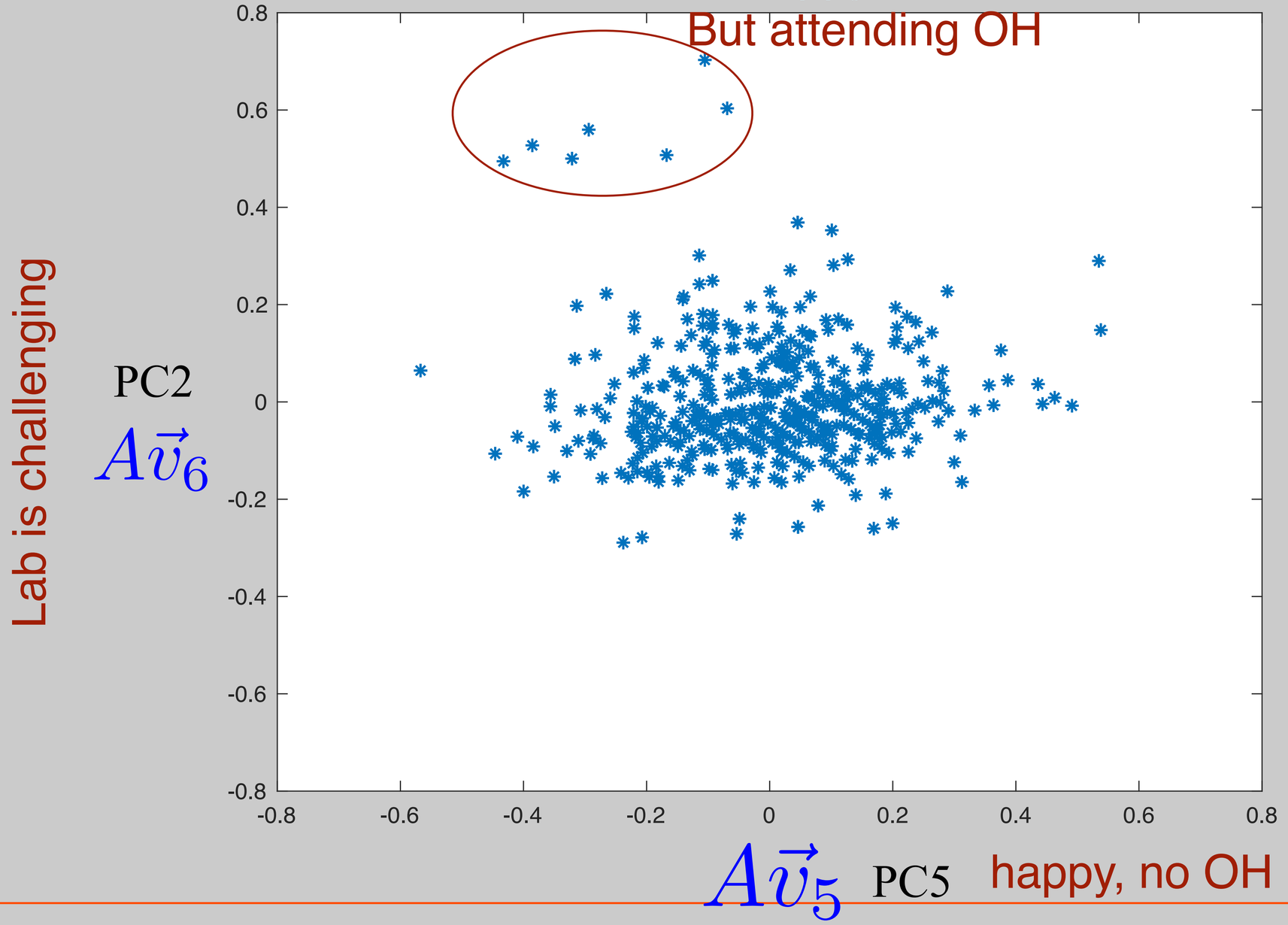
1) HW difficulty
2) HW Length
3) Lab hour/week
4) Current rating
5) Previous Rating
6) Comfortable attending OH



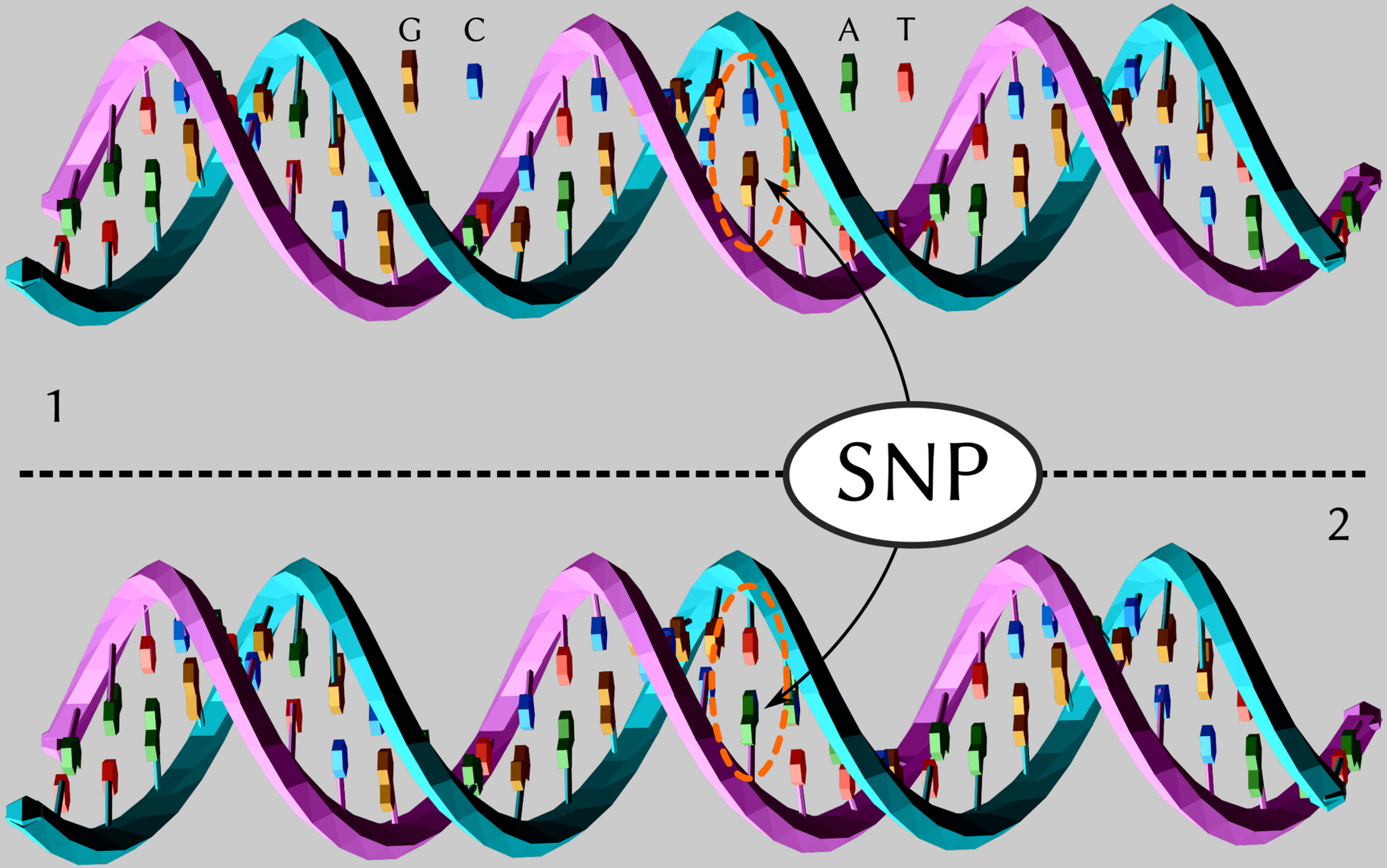


Data Science

Lab is very challenging
Unhappy
But attending OH



Genetic Data



PCA in Genetics Reveals Geography

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

Study:

Characterized genetic variations in 3,000 Europeans from 36 Countries

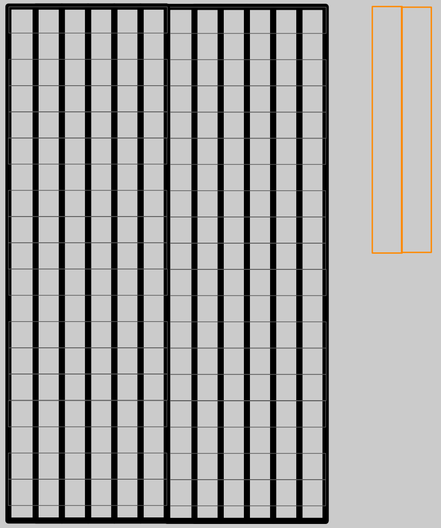
Built a matrix of 200K SNPs (single nucleotide polymorphisms)

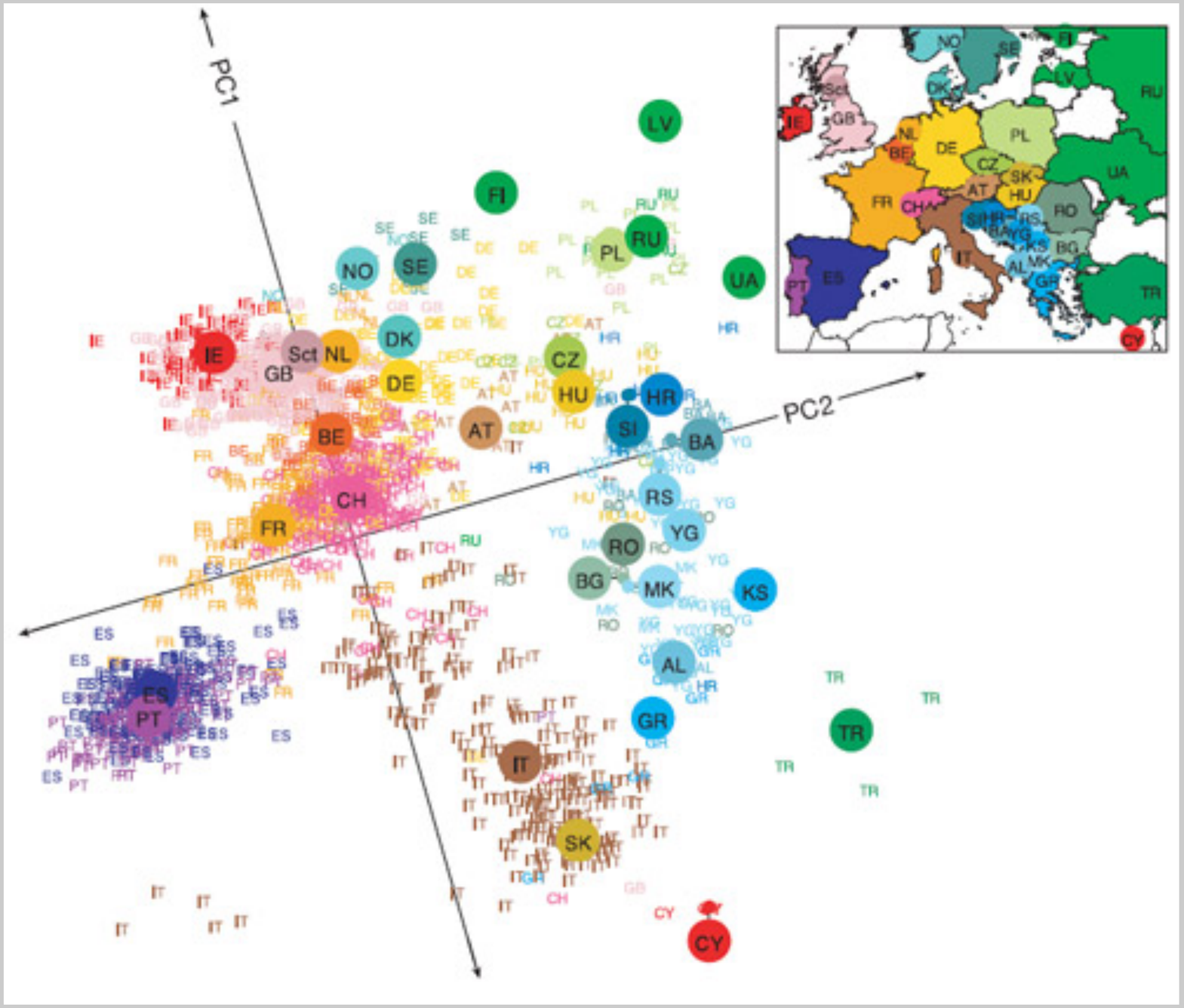
Computed largest 2 principle components

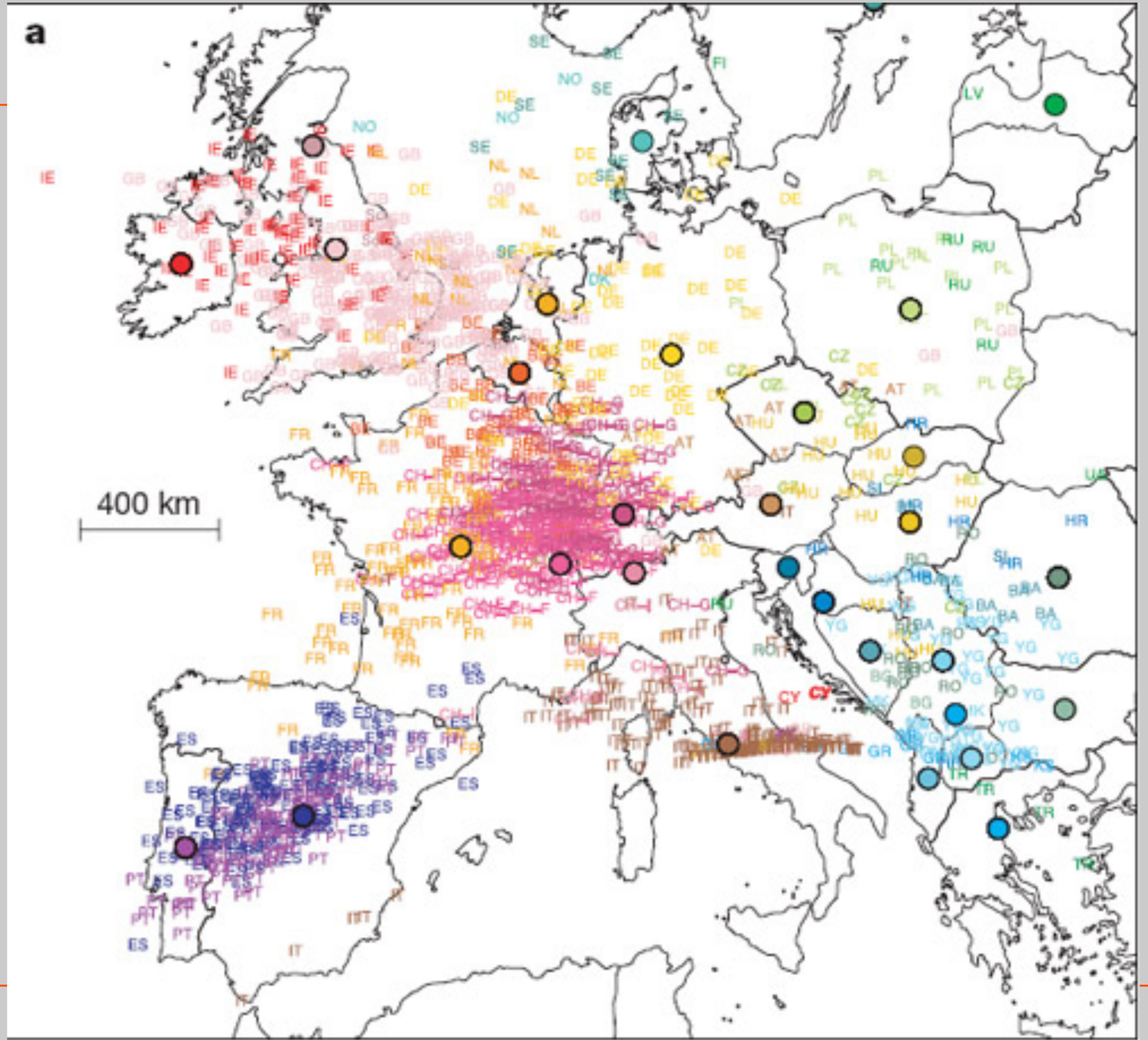
Projected subjects on 2 dimentional data

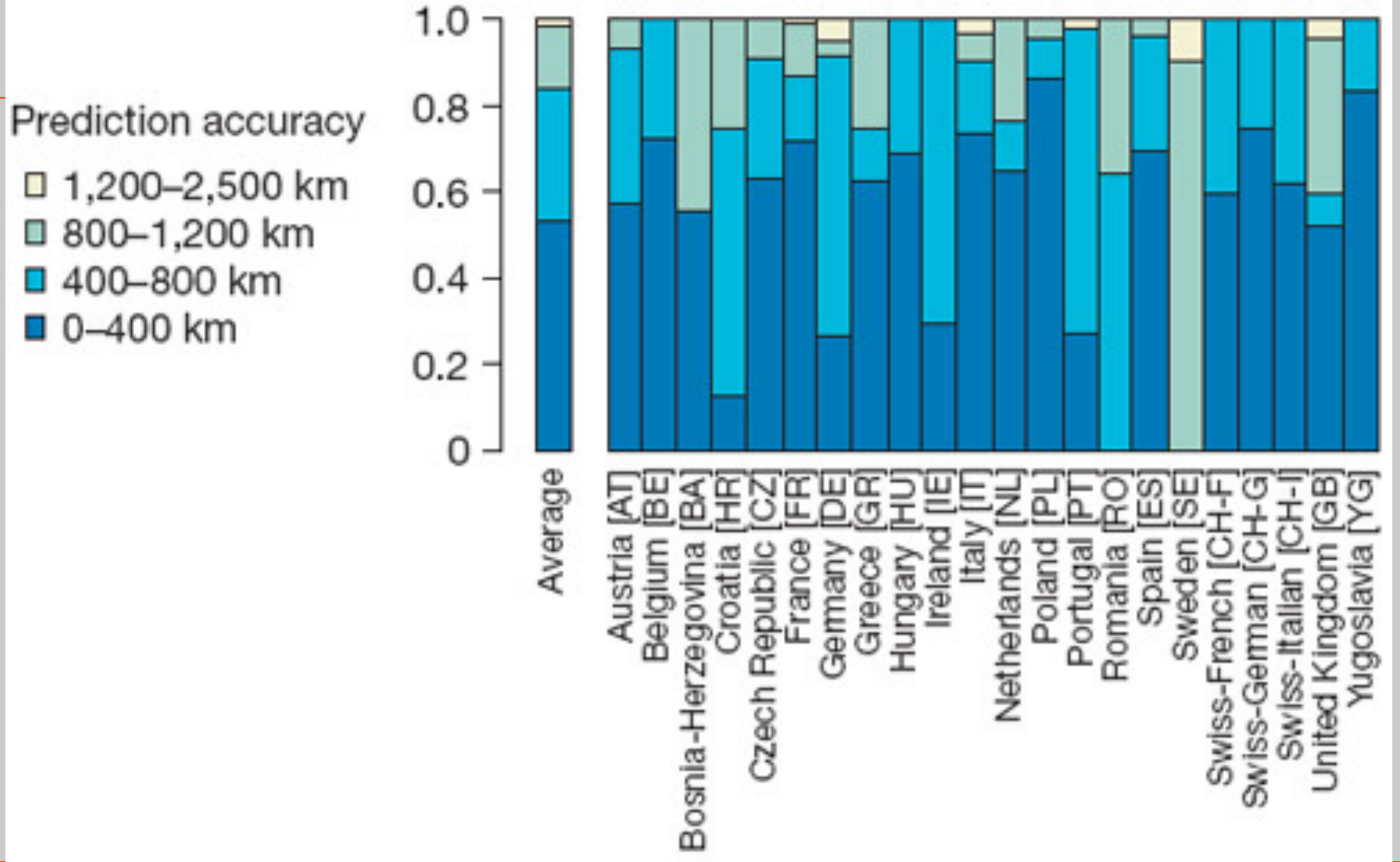
Overlayed the result on the map of Europe

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




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

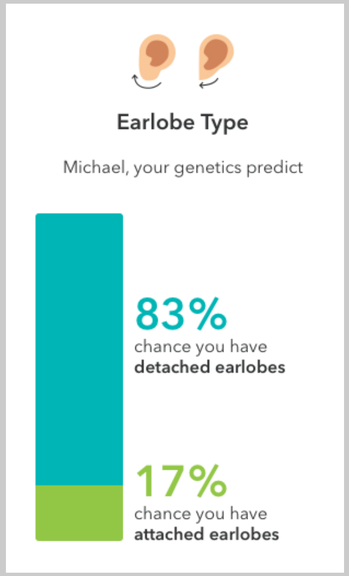
Summary | Scientific Details | Frequently Asked Questions

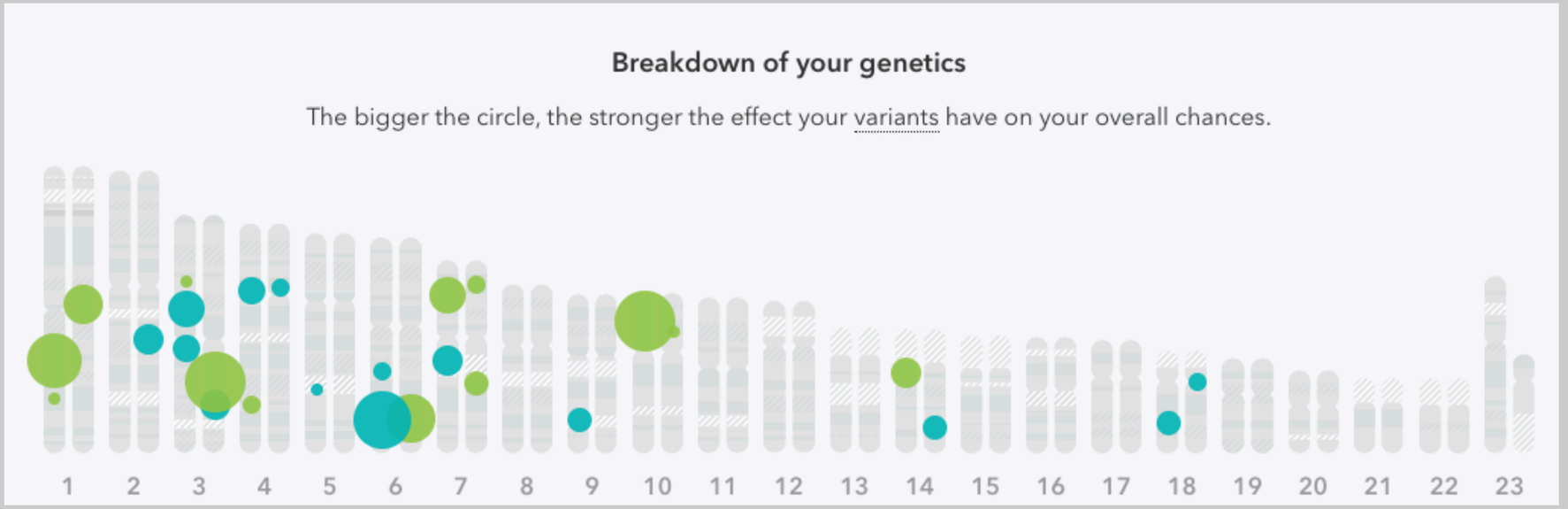
Michael Lustig	100%
European	**99.7%**
Ashkenazi Jewish	98.2% >
Broadly European	1.5%
Trace Ancestry	**0.2%** v
Unassigned	**0.1%** v

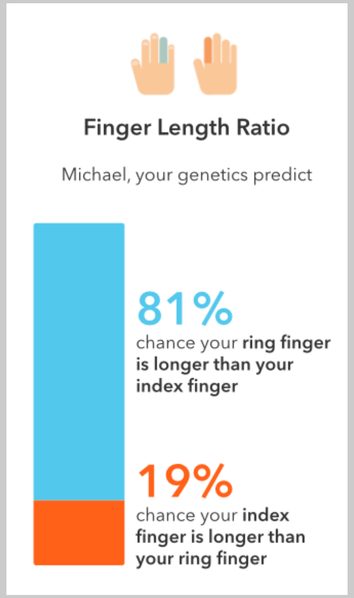
See all tested populations

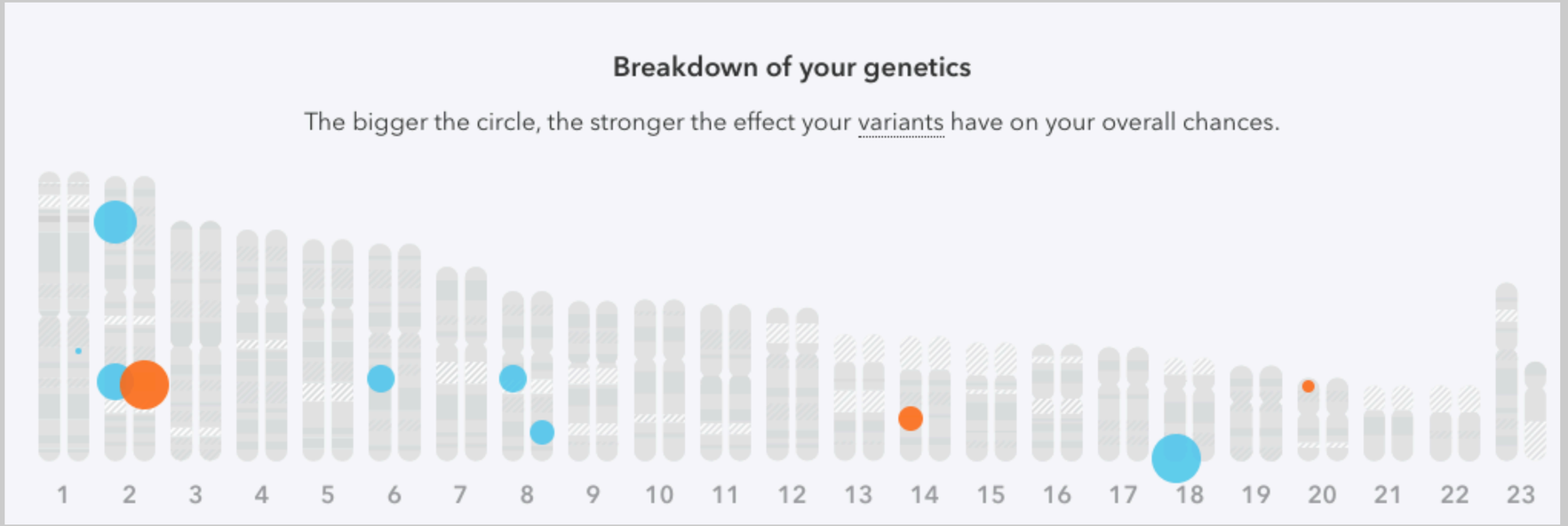
Updated: August 17, 2020

Physical features









Labeled VS non labeled Classification

Word1

Word2

Word3

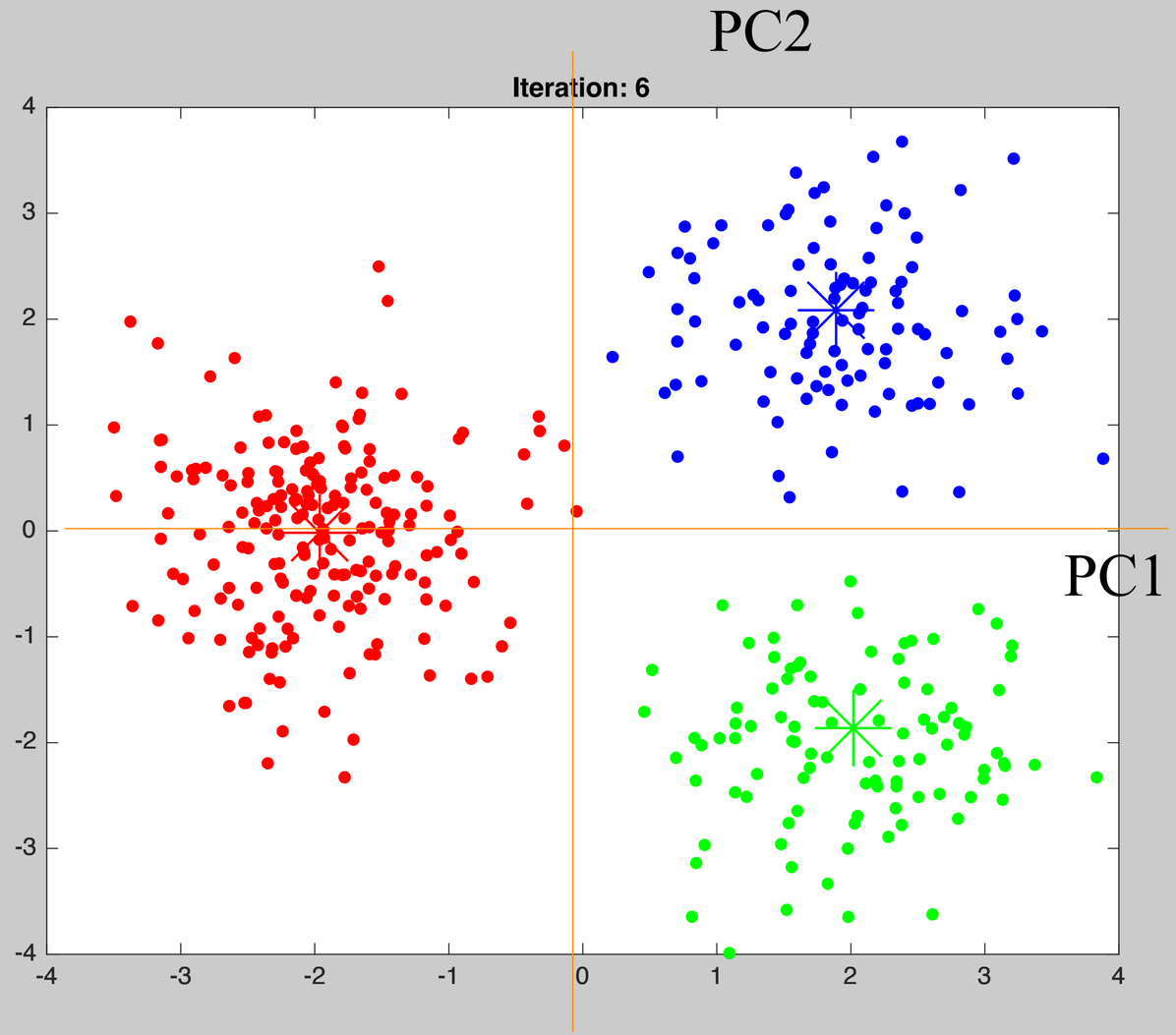
Word4

Word5

Word6

Word7

Word8



Labeled VS non labeled Classification

Word1

Word2

Word3

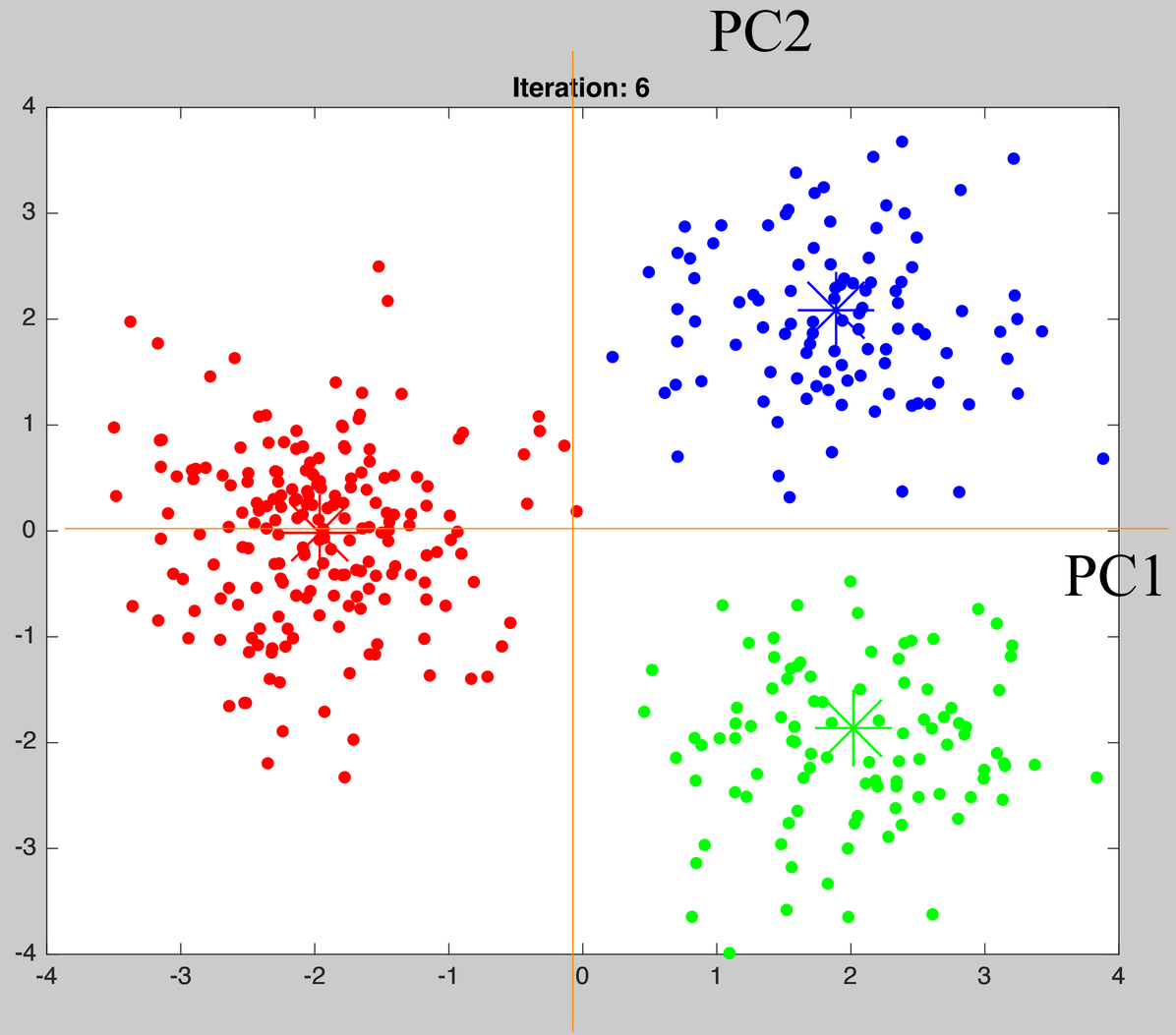
Word4

Word5

Word6

Word7

Word8



Labeled VS non labeled Classification

“Banana”

“Banana”

“Banana”

“Mango”

“Mango”

“Mango”

“Chop”

“Chop”

“Chop”