

HUDK 4050: CORE METHODS IN EDM

In the news



Survey: More Teacher Training Needed for Ed Tech Tools

The Edtech Entrepreneurs Disrupting The Way We Learn

Forbes



Grant to Fund Ed Tech Implementation Sharing Prototype

Encouraging Tech Grads to Give Back

**INSIDE
HIGHER ED**

Events

Title	Date - Time	Location
<u>Designing LA for Humans with Humans</u>	10/16 - 12:00pm	Online
Chip Paucek, Co-Founder of 2U	10/17 - 5:00pm	Low Library 207
<u>Data Science Institute Fall Scholars Programs</u>	Deadline: 10/21	
Where on Earth is AI Headed?	10/25 - 2:00pm	Davis Auditorium
<u>Columbia Nano Day</u>	10/29	Schapiro CEPSR
<u>Robotics to Retrain & Restore Human Movements</u>	11/1 - 12:00pm	NWB Rm 1406
<u>Science Communication Workshop</u>	11/20 - 9:30am	Low Library

A2 Due Thursday
Before Class

Essential Problem

- Dimensionality Reduction
 - Feature selection: select a subset of dimensions
 - Feature extraction: transform lots of dimensions into fewer dimensions
- Why?
 - As a form of insight
 - Avoid “Curse of Dimensionality”

Curse of Dimensionality

Sparsity: The more dimensions that we add, the more comparisons we are missing

	Stats	Cog Psy
Amy	3	2
Chen	2	2
Asif	1	3

Possible Combinations

3 - 3
3 - 2
3 - 1
2 - 3
2 - 2
2 - 1
1 - 3
1 - 2
1 - 1

Curse of Dimensionality

Sparsity: The more dimensions that we add, the more comparisons we are missing

	Stats	Cog Psy	Socio- logy	Crit Theory	Wood- work	Data Sci	Music	Design
Amy	3	2	1	1	3	2	2	2
Chen	2	2	2	3	1	3	2	3
Asif	1	3	3	7	3	2	1	1

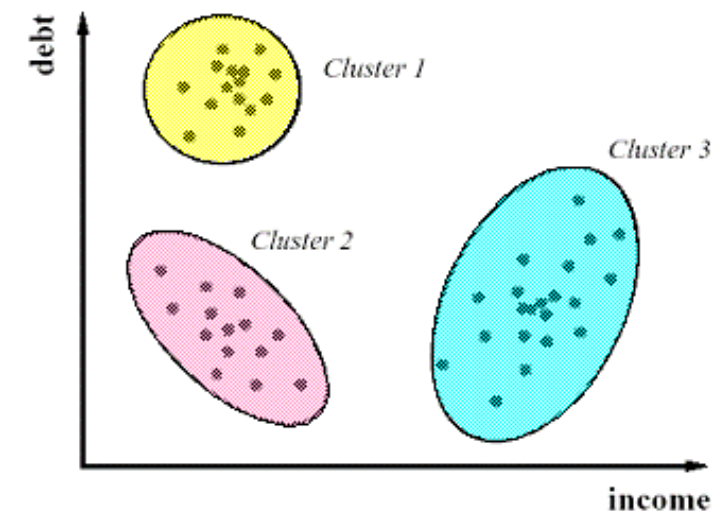
How to reduce dimensions?

Mean, median, mode

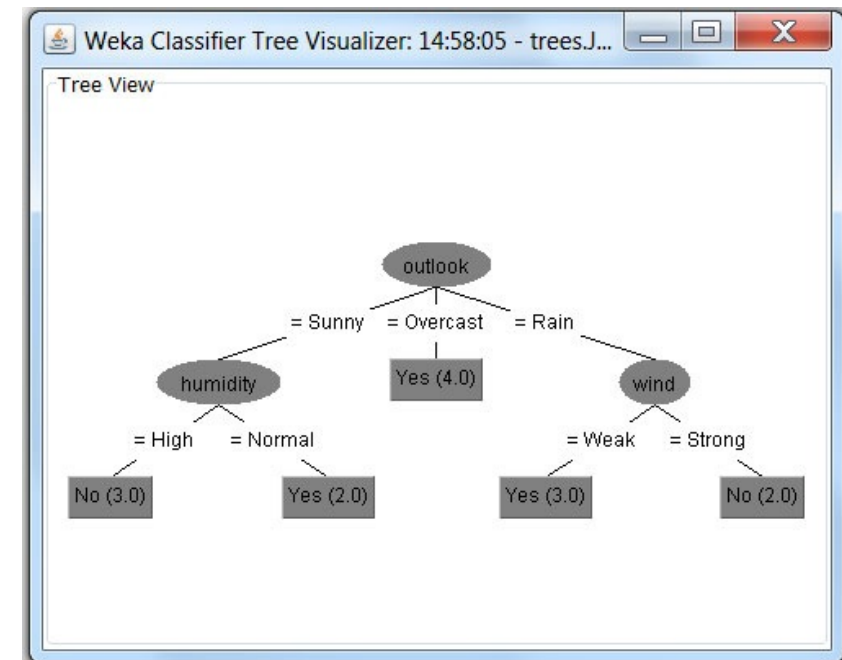
Principle Component
Analysis

Google

Cluster Analysis



Decision Tree



Dimensionality Reduction

Cluster Analysis

Grouping stuff

By Variables

ID	Var1	Var2	Var3
A			
B			
C			
D			

ID	Var2
A	
B	
C	
D	

Selection

ID	Var2+3
A	
B	
C	
D	

Extraction

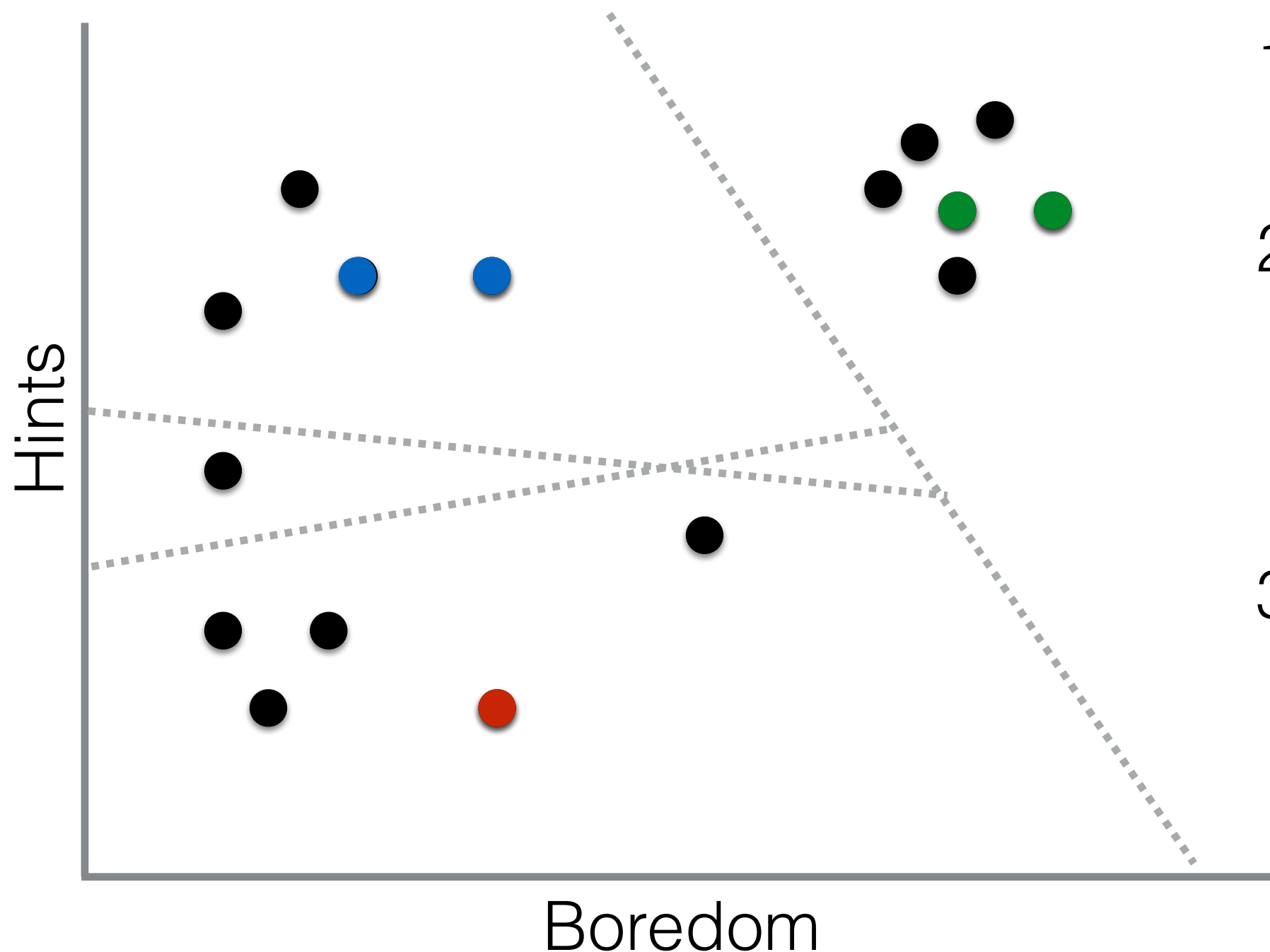
By People



ID	Var1	Var2	Var3
A			
C			

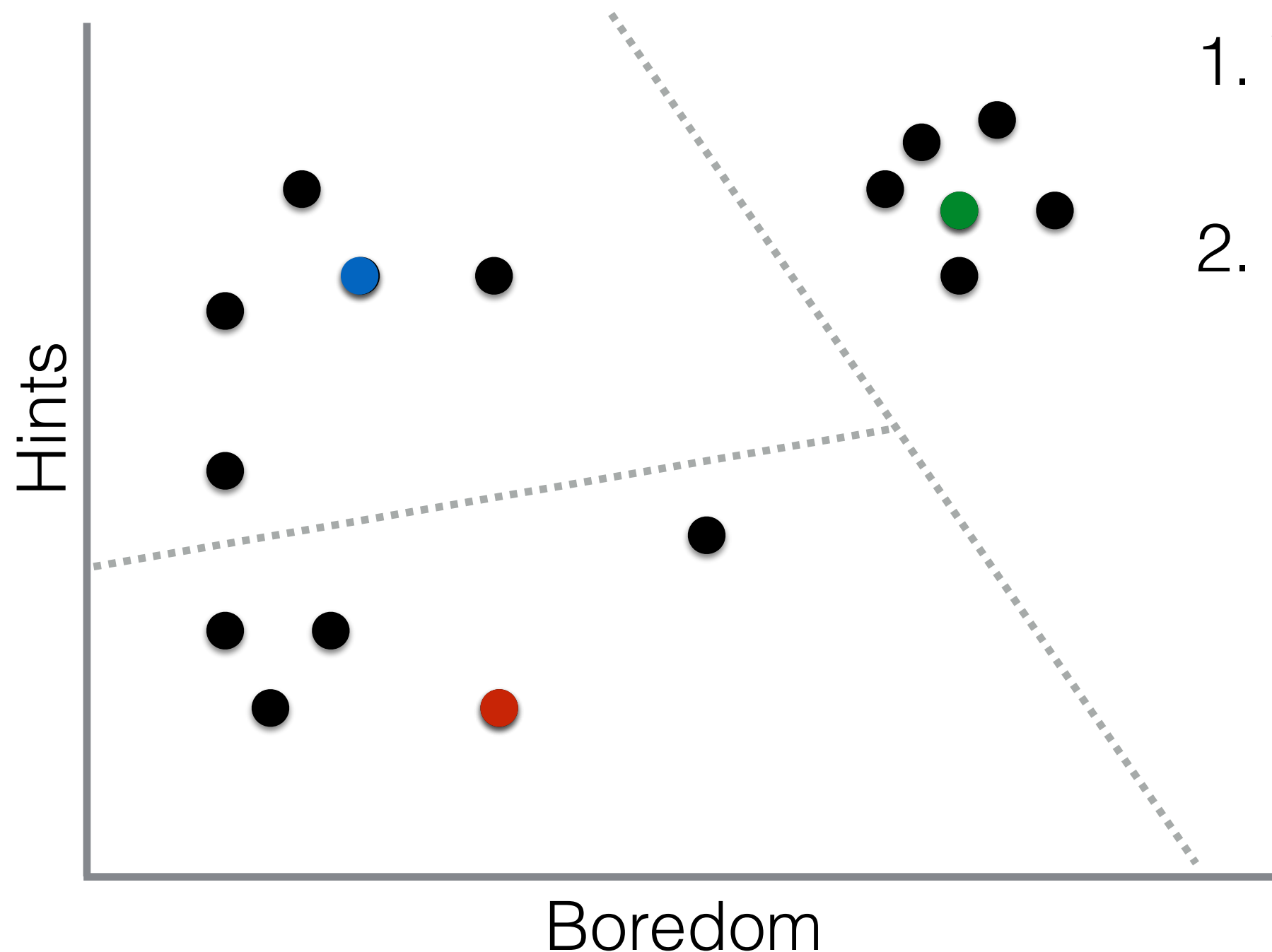
ID	Var1	Var2	Var3
B			
D			

Cluster Analysis: K-means



1. Select some random points
2. Associate those points with closest other points
3. Move the selected point to the mean point in the cluster

Cluster Analysis: K-means



1. Very sensitive to starting values
2. Not good at dealing with complex shapes

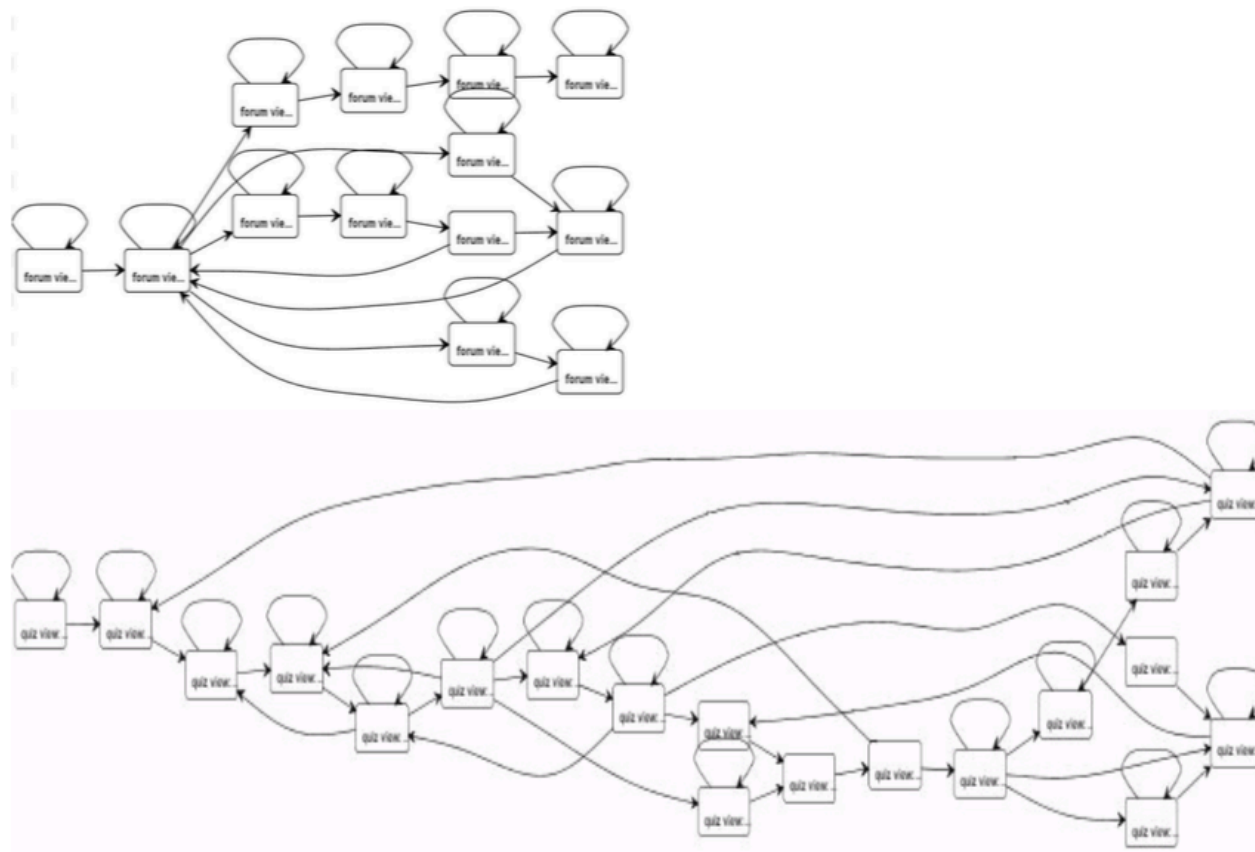


Figure 2. Heuristic net of all students.

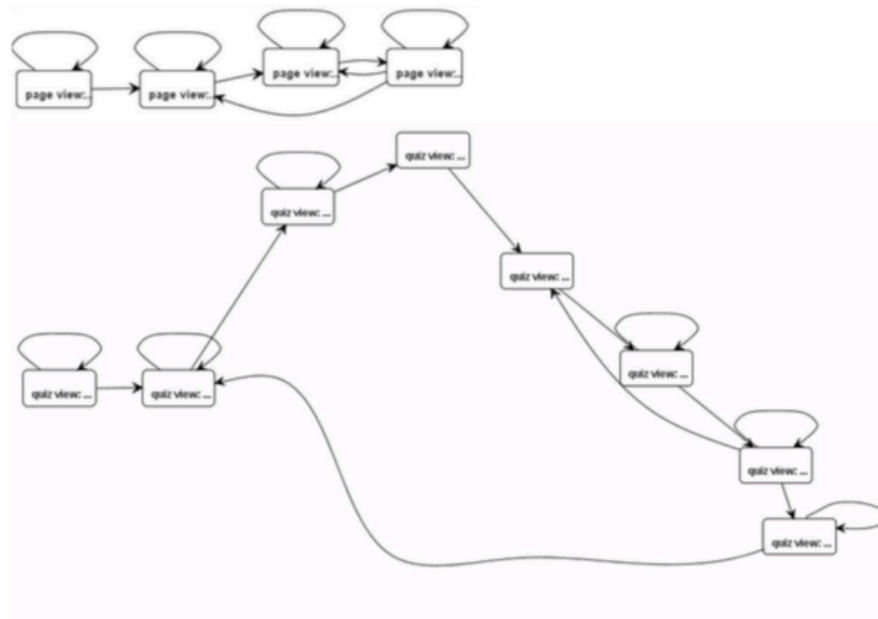
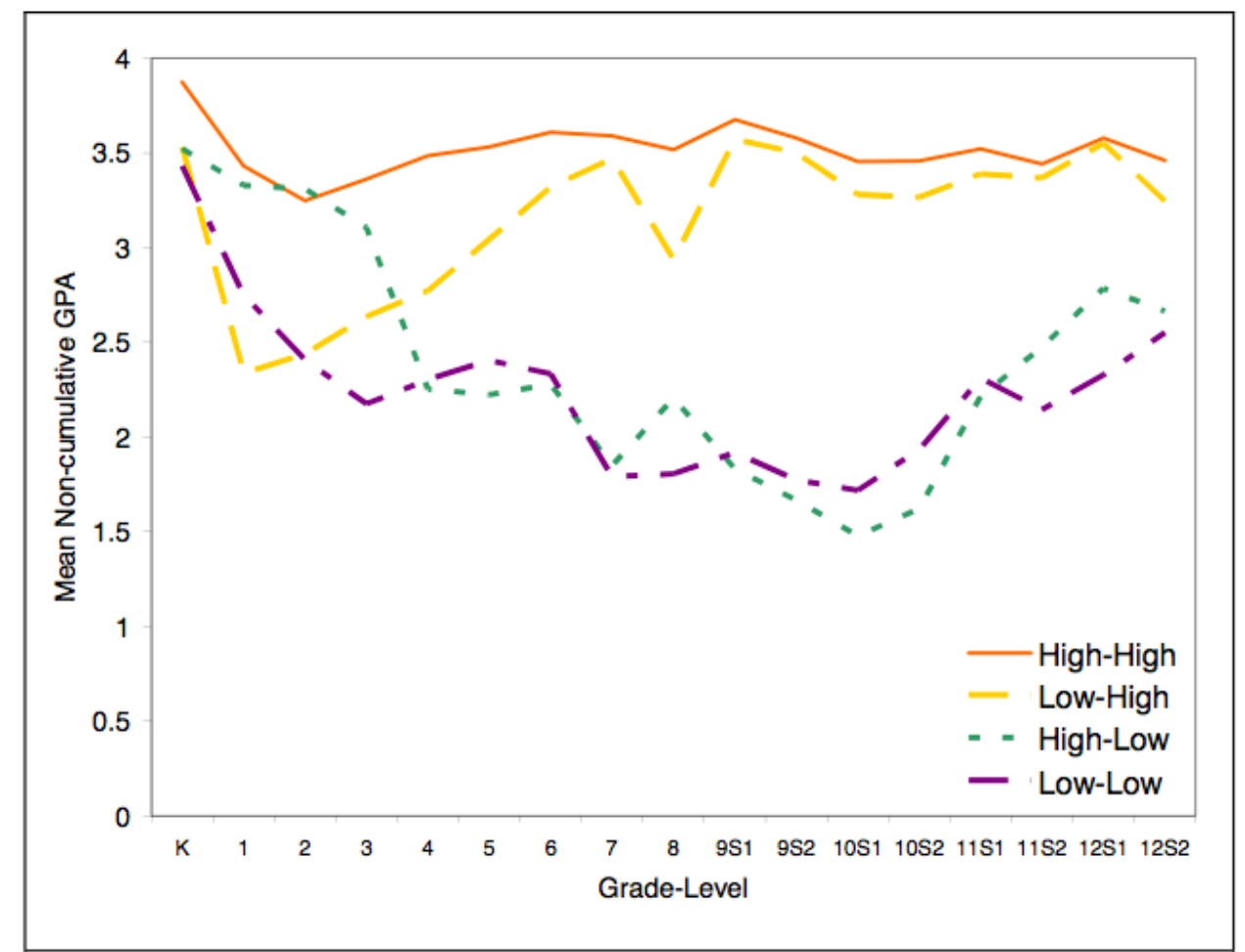
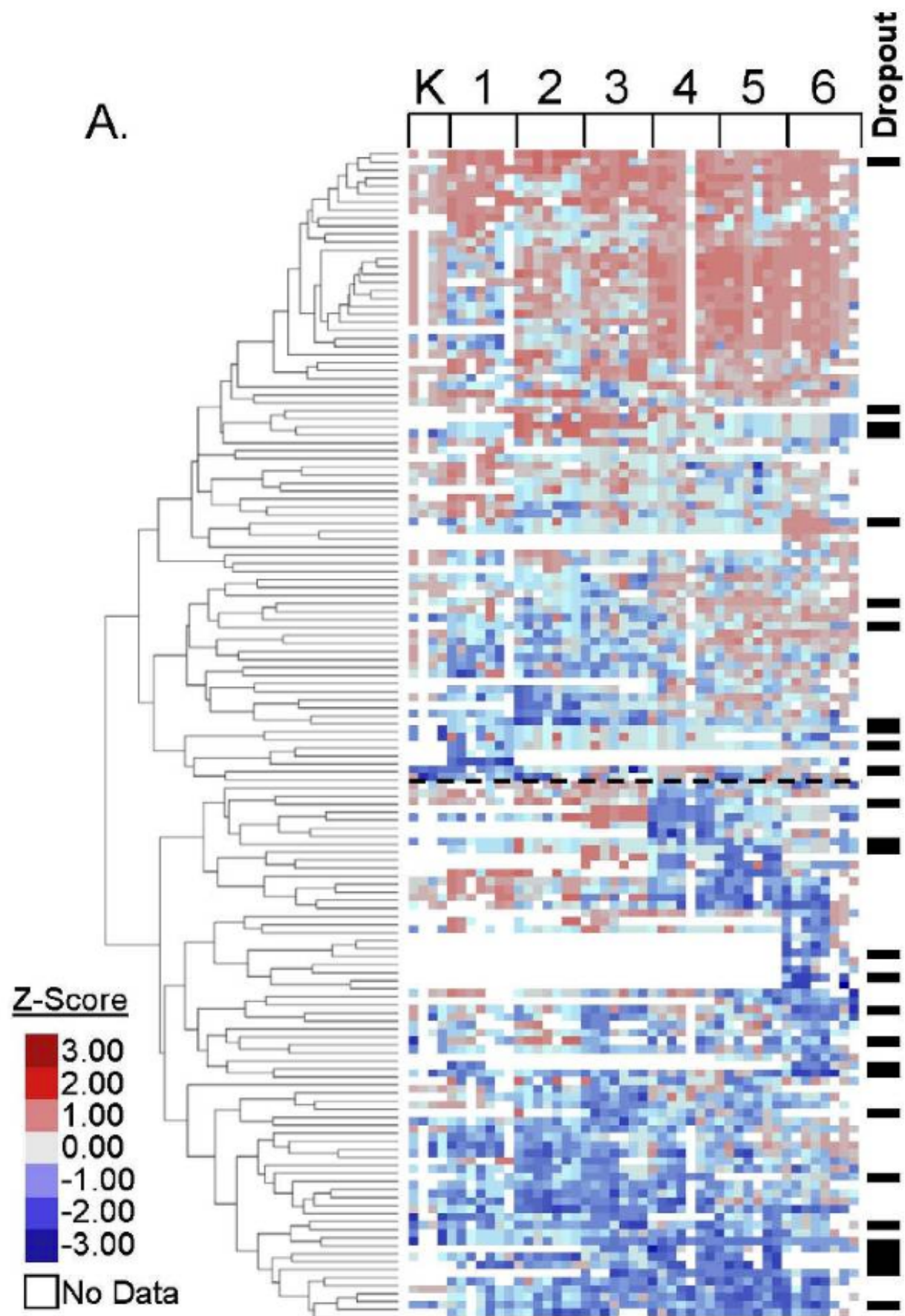


Figure 3. Heuristic net of fail students.

Bogarín, Romero, Cerezo,
Miguel & Sánchez-Santillán
(2014)



Bowers (2010)

K-means Gotchas

- Assumes there are clusters to find - it will find clusters regardless of whether there are any or not
- Does not work on some shapes (Like PB&J need an even spread)
- Need uniform scale (uniform variance) - larger scale will swamp smaller scale
- Doesn't work on categorical data of more than two categories (and the scale may be difficult to interpret)
- Can get stuck on local minima (need to run iterations)
- Too easy

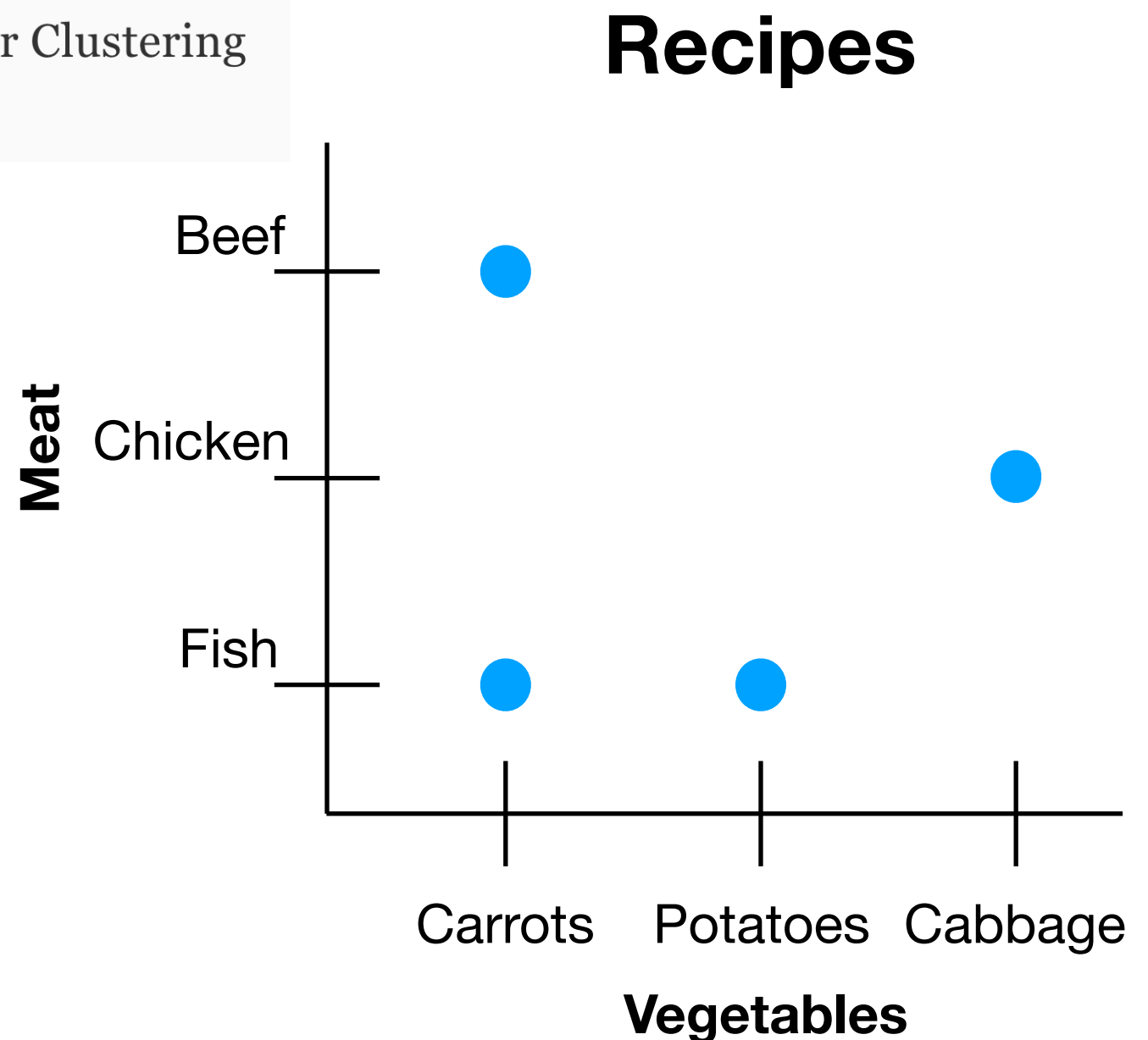
K-modes

[Data Mining and Knowledge Discovery](#)

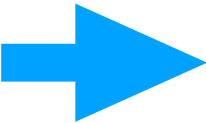
September 1998, Volume 2, [Issue 3](#), pp 283–304 | [Cite as](#)

Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values

- Same as K-means, but uses the modal value of a vector
- Similarity



K-modes

- Put A2 data into the format opposite 
- install the `klaR` package
- `kmodes(df, number of modes, iter.max = 10, weighted = FALSE)`
- Color the vertices in your network diagram according to cluster

student	HUDK4050	HUDK4011	HUDK5053
A	Yes	No	No
B	Yes	No	Yes
C	Yes	No	Yes
D	No	Yes	No

K-modes

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes

1. Randomly choose cluster

“model student”

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes

student	C1	C2
S1	2	1
S2	0	3
S3	2	1
S4	2	1
S5	3	0
S6	1	2
S7	1	2

2. Calculate “distance” for each student from model

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes

student	C1	C2
S1	2	1
S2	0	3
S3	2	1
S4	2	1
S5	3	0
S6	1	2
S7	1	2

3. Allocate to closest cluster

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes

student	C1	C2
S1	2	1
S2	0	3
S3	2	1
S4	2	1
S5	3	0
S6	1	2
S7	1	2

4. Calculate Cluster Mode

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes
C1	Yes	Yes	Yes/No
C2	No	No	No

student	C1	C2
S1	2	1
S2	0	3
S3	2	1
S4	2	1
S5	3	0
S6	1	2
S7	1	2

5.Repeat Step 2 with Modal Clusters...

student	HUDK4050	HUDK2020	HUDK5011
S1	Yes	Yes	No
S2	No	No	No
S3	Yes	No	Yes
S4	Yes	Yes	No
S5	Yes	Yes	Yes
S6	No	Yes	No
S7	No	No	Yes
C1	Yes	Yes	Yes
C2	No	No	No

student	C1	C2
S1	1	2
S2	3	0
S3	1	2
S4	1	2
S5	0	3
S6	2	1
S7	2	1

**Repeat steps until
no change.**

K-modes Gotchas

- Assumes there are clusters to find - it will find clusters regardless of whether there are any or not
- Does not work on some shapes (Like PB&J need an even spread)
- Need uniform scale (uniform variance) - larger scale will swamp smaller scale
- Can get stuck on local minima (need to run iterations)
- Too easy

Cluster Survey

[http://bit.ly/
hudk405019CLUSTER](http://bit.ly/hudk405019CLUSTER)

Cluster Activity

core-methods-in-edm/
class-activity-6