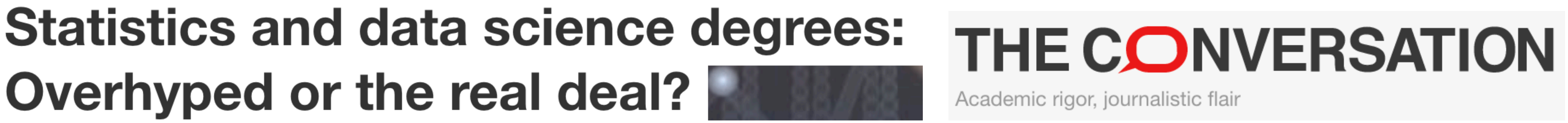
# HUDK 4050: CORE METHODS IN EDM

10/23/18 2:25 PM

# In the news



Statistics and data science degrees: Overhyped or the real deal?

THE CONVERSATION

Academic rigor, journalistic flair

ScienceDaily®

Scientists use AI to develop better predictions of why children struggle at school

EdSurge

Libraries Look to Big Data to Measure Their Worth—And Better Help Students

THE TECH EDVOCATE

WHERE LEARNING ANALYTICS GO WRONG

Daily Herald

Suburban Chicago's Information Source

Roboticist trains AI to write fortunes -- and things get weird

# Events

| Event | Date | URL |
|---|---|---|
| NYAS: Healthcare in the Era of Big Data | October 24-25 | https://www.nyas.org/events/2018/healthcare-in-the-era-of-big-data-opportunities-and-challenges/?utm_source=The+New+York+Academy+of+Sciences&utm_campaign=f0807f47cb-eNews_October_2018-10-18&utm_medium=email&utm_term=0_cba25b11d2-f0807f47cb-184577937&mc_cid=f0807f47cb&mc_eid=cfeeec7fb2 |
| Cross-device User Clustering at Adobe | 5:30 November 29 | https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e28-655b8cee-0165-5dd5c72b-00001287events@columbia.edu&recurrenceId= |
| DSI: Towards Better Reinforcement Learning for High Stakes Domains | 5:30pm November 1 | https://www.eventbrite.com/e/new-york-data-science-seminar-series-emma-brunskill-stanford-tickets-51551174952 |
| Data Law in a Global Digital Economy | November 9 | https://www.guariniglobal.org/data-law) |
| NYAS: Deep Learning to Accelerate Drug Development | November 13 | https://www.nyas.org/events/2018/deep-learning-to-accelerate-drug-development-an-nyc-symposium/?utm_source=The+New+York+Academy+of+Sciences&utm_campaign=f0807f47cb-eNews_October_2018-10-18&utm_medium=email&utm_term=0_cba25b11d2-f0807f47cb-184577937&mc_cid=f0807f47cb&mc_eid=cfeeec7fb2 |
| People centric approach to optimize Data Science, Commercial impact and Leadership | 10:30am November 14 | https://events.columbia.edu/cal/event/eventView.do?b=de&calPath=%2Fpublic%2Fcals%2FMainCal&guid=CAL-00bb9e24-655b8449-0165-5e0ea7e9-00001957events@columbia.edu&recurrenceId= |
| Machine Learning Innovation Summit | December 12-13 | https://www.theinnovationenterprise.com/summits/machine-learning-innovation-summit-new-york-2018 |

# 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 | Sociology | Crit Theory | Woodwork | 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 |

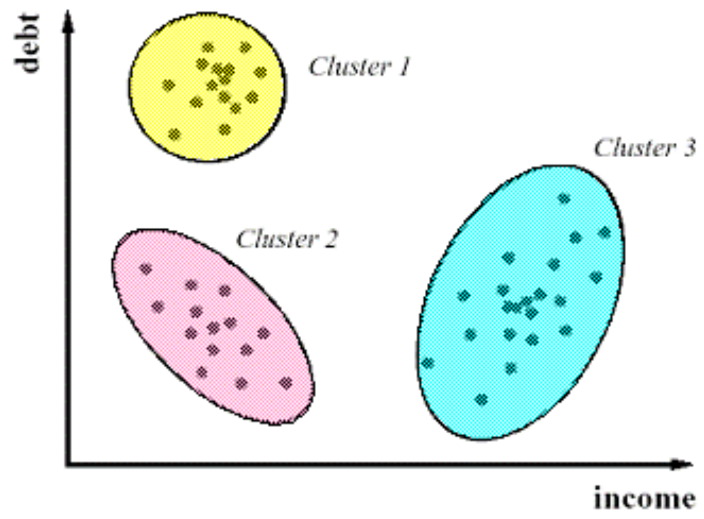
# How to reduce dimensions?

Mean, median, mode

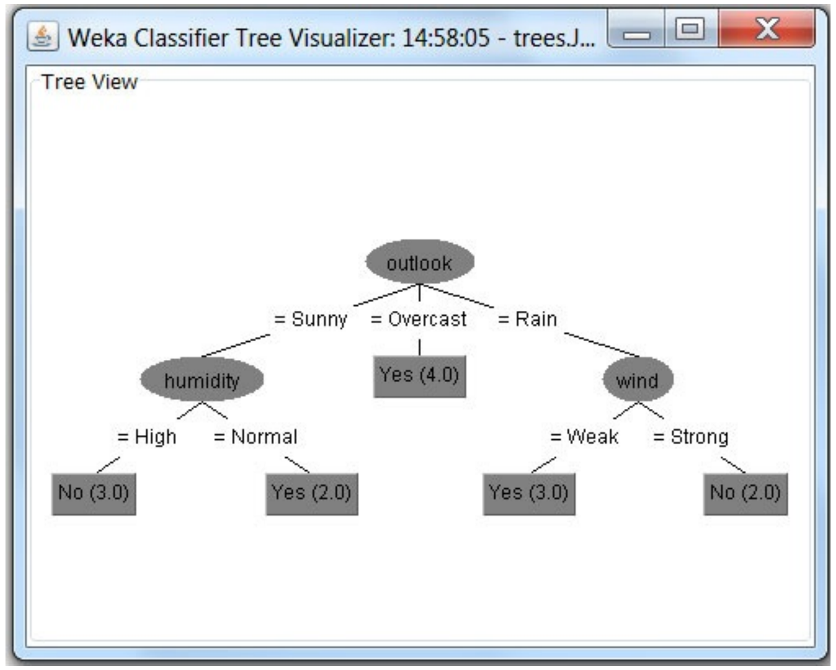
Principle Component Analysis

Google

Cluster Analysis



Decision Tree


Weka Classifier Tree Visualizer: 14:58:05 - trees.J...

# 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 | |

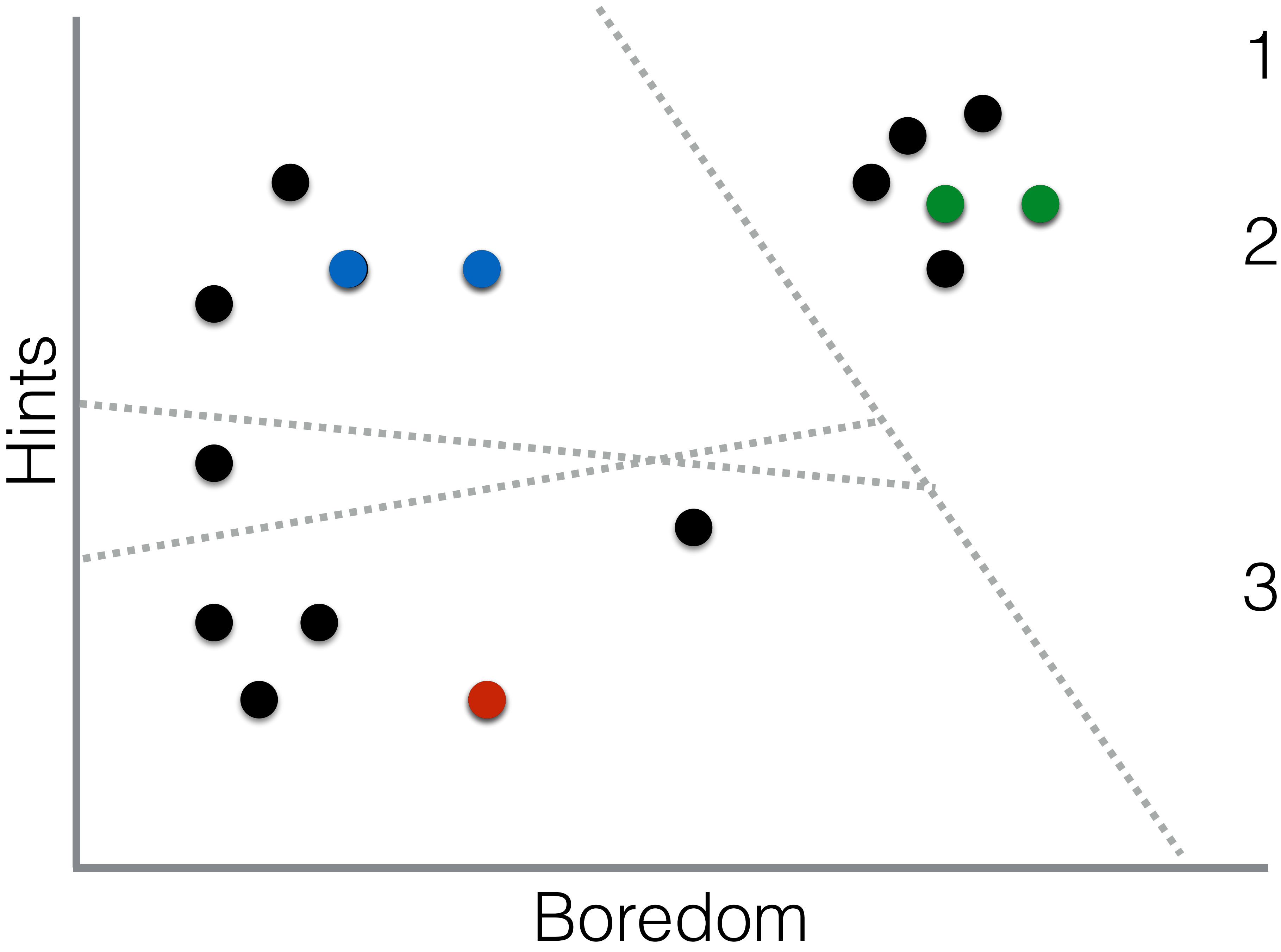
Extraction

## By People

| ID | Var1 | Var2 | Var3 |
|---|---|---|---|
| A | | | |
| C | | | |

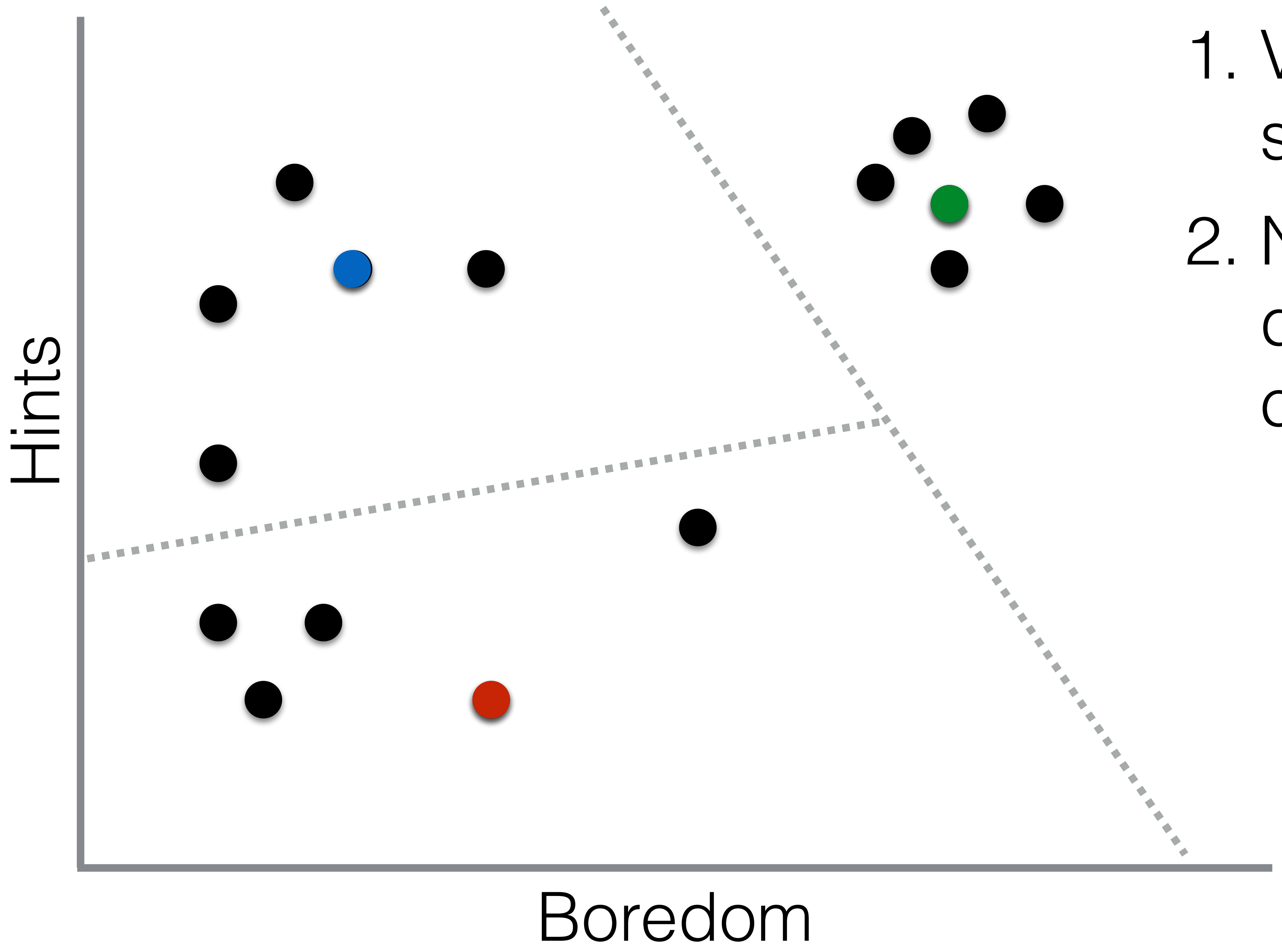
| ID | Var1 | Var2 | Var3 |
|---|---|---|---|
| B | | | |
| D | | | |

# Cluster Analysis: K-means

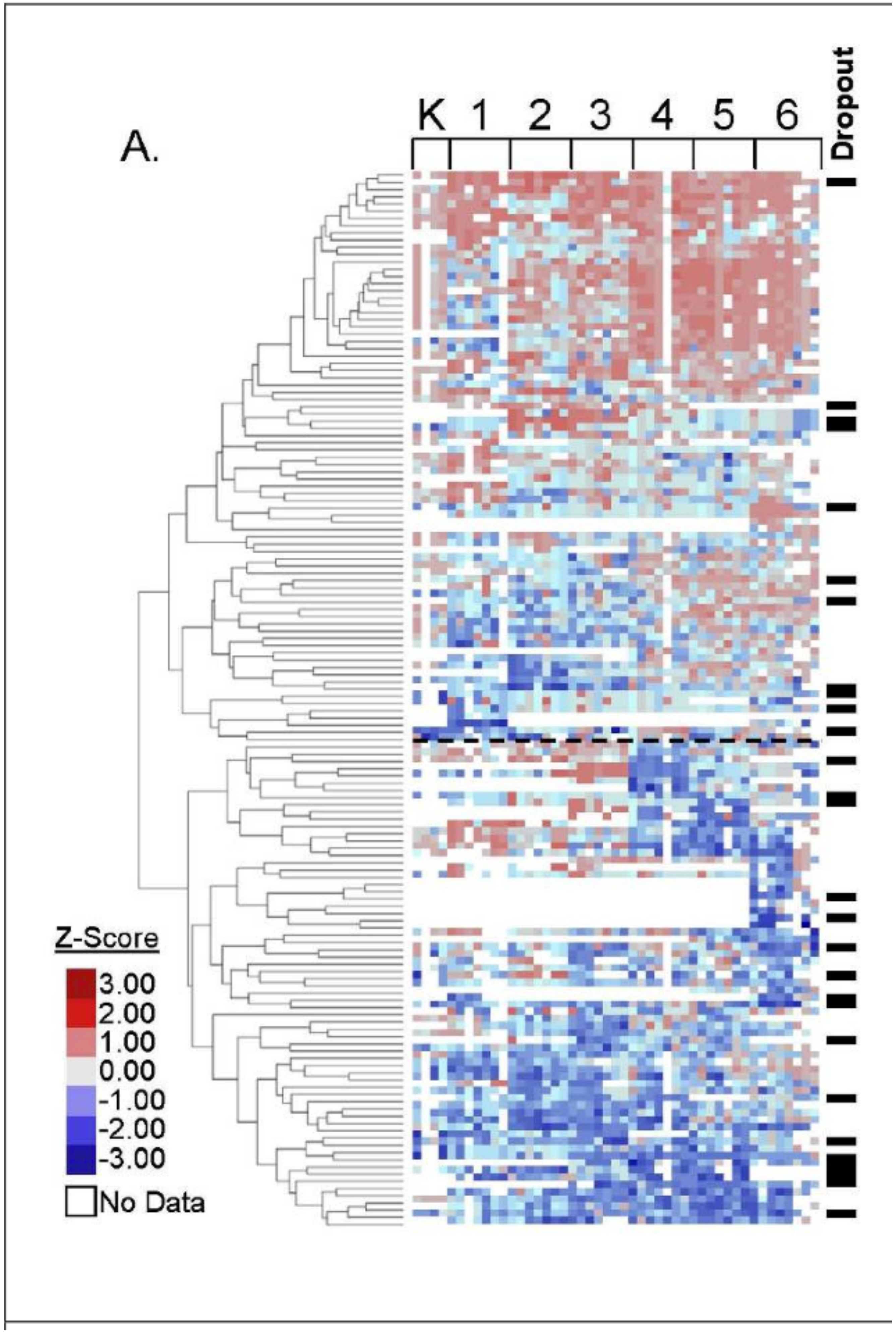


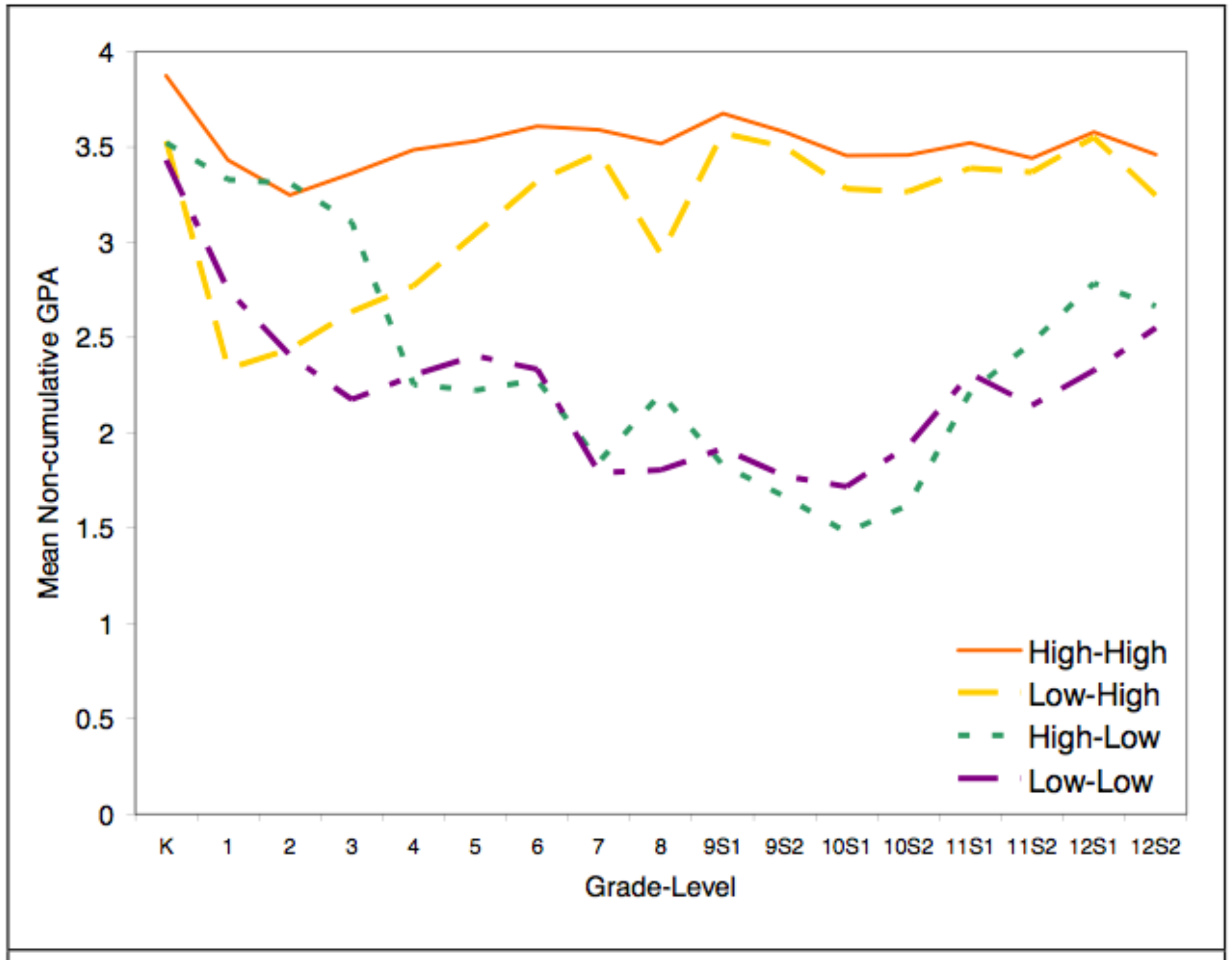
1. Select some random points
2. Associate those points with closest other points
3. Move the selected point to the mean point in the cluster

# Cluster Analysis: K-means



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


A.
K
1
2
3
4
5
6
Dropout
Z-Score
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
No Data


Mean Non-cumulative GPA
4
3.5
3
2.5
2
1.5
1
0.5
0
K
1
2
3
4
5
6
7
8
9S1
9S2
10S1
10S2
11S1
11S2
12S1
12S2
Grade-Level
High-High
Low-High
High-Low
Low-Low

Bowers (2010)

# 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 | | | |

# Feature Extraction

- Principal Component Analysis
    - Variance
    - Covariance
    - Matrix algebra

# Cluster Survey

bit.ly/HUDK4050-cluster

# Cluster Activity

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