Exploring KC models in DataShop: with a side of clustering and AFM

Michael Wixon (mwixon@wpi.edu)
Learning Sciences and Technologies
Worcester Polytechnic Institute

Daniel Seaton (dseaton@mit.edu)
Department of Physics
Massachusetts Institute of Technology

Mentors: Nan Li and Zach Pardos

Problem Statement

- Human Built KC Models Require Effort
- PCA Built KC Models are Hard to Interpret
- I wanted Interpretable KC Models more Efficiently

Data Specific Solution

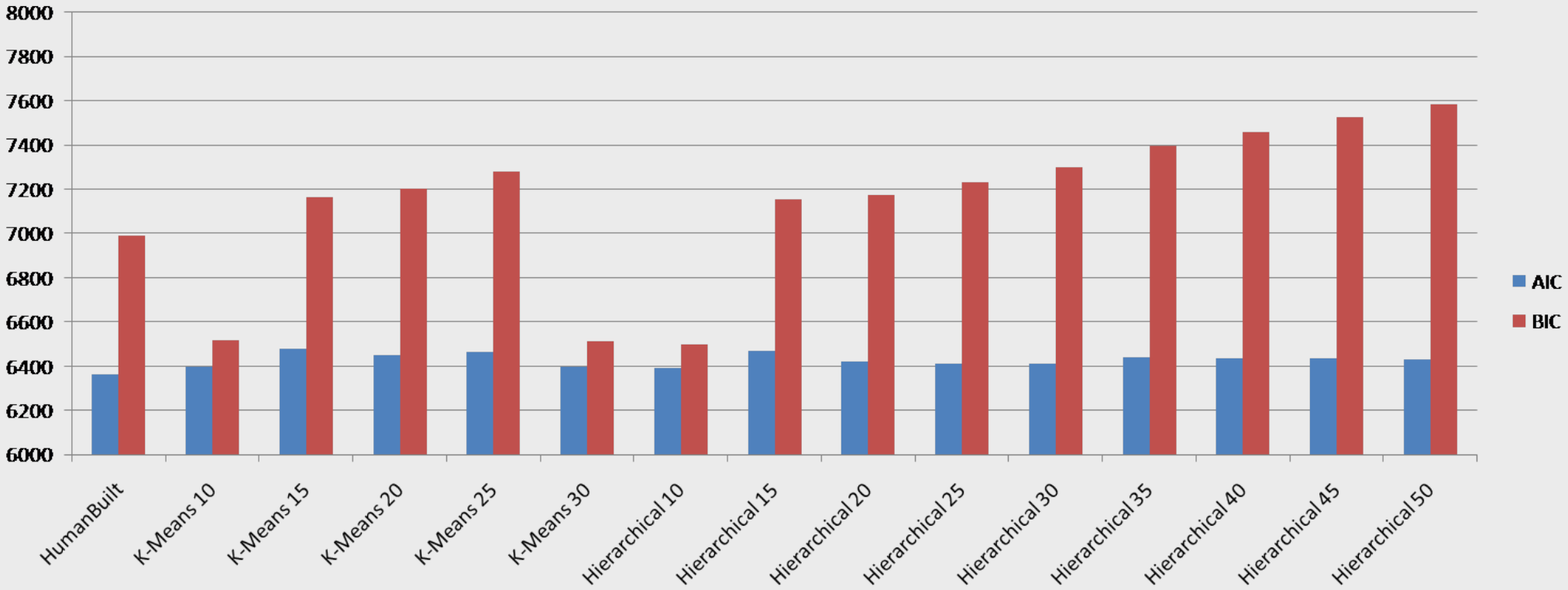
Algebra Dataset example: -4y + 8 = (3/x) … not too general

- Convert to simplified form: : -nv + n = (n/v)
- Look for Unigrams, Bigrams & Trigrams eg: -, -n, -nv
- Create Binary Matrix of which problems include which unigrams, bigrams, & trigrams

	n	nv	-n	(n	+n=
-4y + 8 = (3/x)	1	1	1	1	1
x = 7/8 – 1	1	0	1	0	0
6 + 1 = y/8 * 4	1	0	0	0	1
9/(-4x) = 9 + x/3	1	1	1	0	0
y/6 = 9/6	1	0	0	0	0

- Run Cluster Analysis on that Binary Matrix
- Use Clusters as KC Model

Qualified Success



Cluster/KC	1	3	12	13	18
Follows Form	n=n/v	n/v*v=nv	n/n=nv/n (often common denominator)	v=n/n	(-nv) No other *,/,or v in equation

Philosophy and Future Work

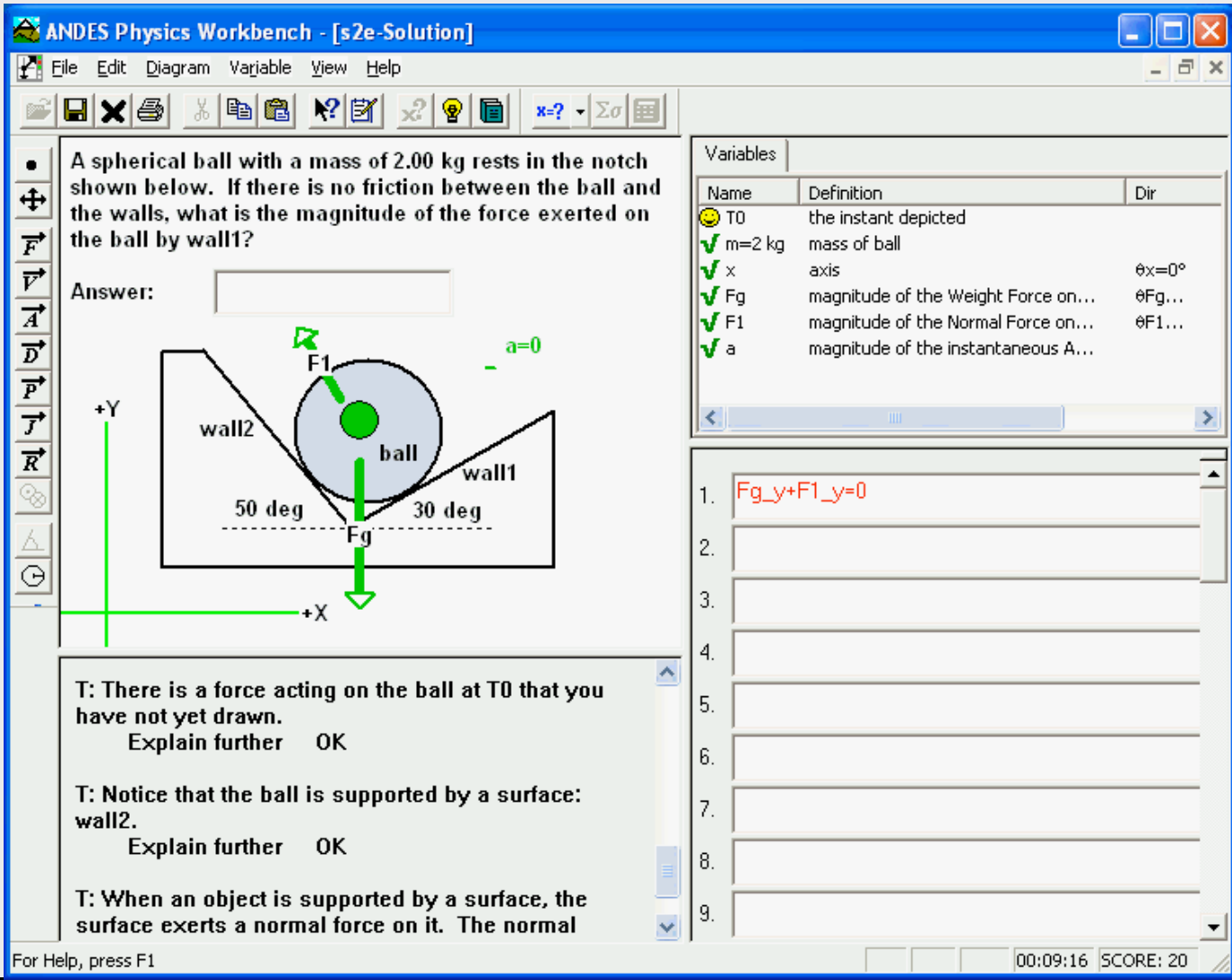
- Synthesis: Apply LFA to Use Best of Each KC Model
- Extension:
  - Target areas where human judgment doesn't perform well
  - Apply automated methods to fill those gaps
- Proposed Future Work
  - Find KCs which perform "badly"
  - Apply PCA to that subset

Acknowledgements

- Thanks to Ken Koedinger who suggested the idea of generating KC models from clustered ngrams
- Thanks to mentors Nan Li and Zach Pardos whose help made this project possible
- Thanks to CMU's LearnLab and DataShop for providing tools to facilitate these analyses and a vibrant learning environment

Exploring Physics KCs for the Andes tutor

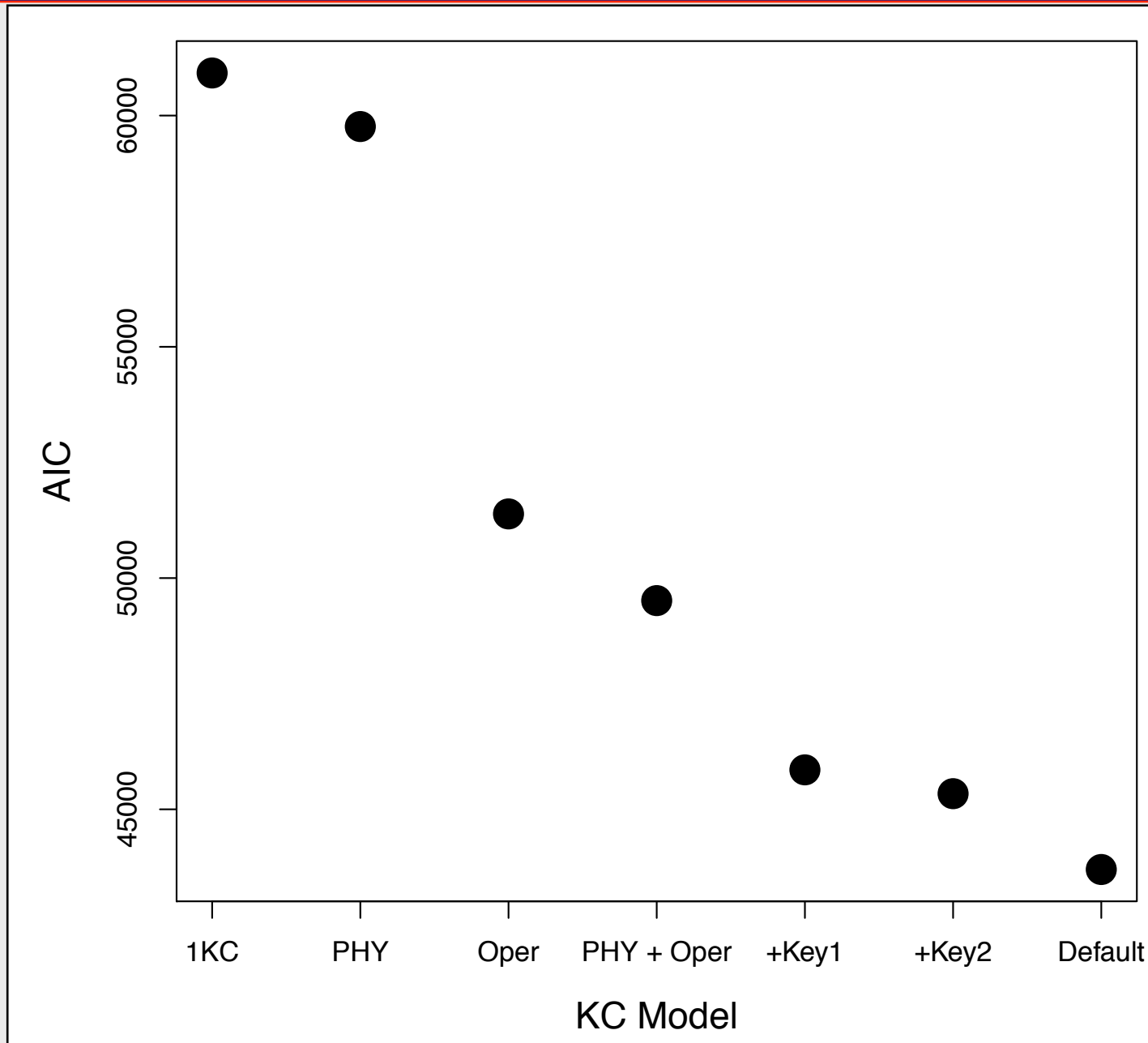
- Andes is an intelligent tutor system designed to help students with physics homework: USNA Physics - Fall 2008, Fall 2009
- Hands on tutor that takes students through all problem solving steps
- Majority of the Andes data sets have not been analyzed, and models depend on tasks
- Attempted to add physics information extracted from step and problem names



Michael Wixon and Daniel Seaton

Exploring Physics KCs for the Andes tutor

1KC:	Single KC
PHY:	by physics topic
Oper	Andes task (not step)
PHY+ Oper	concatenate two above
Oper +Key1	refine Oper with Key1
Oper +Key2	refine Oper with Key2
Default	Tasks and physics



Michael Wixon and Daniel Seaton

Exploring the Additive Factor Model (AFM)

- Interested in finding ways to incorporate time-based features into predictions of learning
- AFM is in a class of logistic regression techniques widely used in educational prediction (was great to get some valuable experience!)

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik}$$

- Adjusted AFM in the R language to include time-based features, and explored the outcomes of the prediction

Koeding, McLaughlin, & Stamper (2012)
Cen, Koedinger, & Junker (2006)

Exploring the Additive Factor Model (AFM)

- Geometry96-97 DataShop data set
- Included duration feature (time in sec)
- Also investigated ways of discretizing time

Model	AIC
AFM	5084.1
AFM + time	5082.0
AFM + KC:time	5083.8

- Time in this context does not make large differences. Will continue to investigate, including how best to incorporate various time measures into such models.

Thanks! and Future Work

- Future work:
  - Prepare online course data from MIT for DataShop
  - Continue exploring AFM and logistic regression models

- Thanks to Ken Koedinger and Nan Li for access to AFM model exploration code in R!!!

Learn Lab Workshop
Ken Koedinger
John Stamper
ALL STAFF!!!
Nan Li
Zach Pardos

WPI
Janice Gobert
Ryan Baker
Michael Sao Pedro

MIT
David E. Pritchard
Saif Rayyan
Yoav Bergner

Exploring the Additive Factor Model (AFM)

- Interested in finding ways to incorporate time-based features into predictions of learning
- Adjusted AFM in the R language to include time-based features, and explored the outcomes of the prediction

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik}$$

Given:
p_{ij} = probability student i gets step j correct
Q_{kj} = knowledge component k needed for step j
T_{ik} = opportunities student i has had to practice k
Estimated:
θ_i = proficiency of student i
β_k = difficulty of KC k
γ_k = gain for each practice opportunity on KC k

Koeding, McLaughlin, & Stamper (2012)
Cen, Koedinger, & Junker (2006)