Week 1, video 1

Intro to EDM

Why EDM now?

Which tools to use in class

Big Data in Education

This textbook

- In this MOOC, you'll learn methods used for exploring big data in education

Two communities

- International Educational Data Mining Society
  - First event: EDM workshop in 2005 (at AAAI)
  - First conference: EDM2008
  - Publishing JEDM since 2009
- Society for Learning Analytics Research
  - First conference: LAK2011
  - Journal of Learning Analytics (founded 2012)

Two communities

- Joint goal of exploring the "big data" now available on learners and learning
- To promote
  - New scientific discoveries & to advance learning sciences
  - Better assessment of learners along multiple dimensions
    - Social, cognitive, emotional, meta-cognitive, etc.
    - Individual, group, institutional, etc.
  - Better real-time support for learners

EDM/LA is…

- "… escalating the speed of research on many problems in education."
- "Not only can you look at unique learning trajectories of individuals, but the sophistication of the models of learning goes up enormously."

Arthur Graesser, Editor,
Journal of Educational Psychology

EDM/LA is…

- “… great.”

- Me

EDM and LAK

- Despite the area's newness, we've learned a few things about key problems

- This course is about methods that have been found to be useful for those problems by EDM/LAK researchers

Where do methods come from?

- Some of the methods would be familiar to someone with a background in Data Mining or Machine Learning
- Some of the methods would be familiar to someone with a background in Psychometrics or traditional Statistics
- You don't have to have either of these backgrounds to get something out of the course
  - Pick and choose what you find most useful

A few words for data miners

- You'll find that there are some current trends in data mining that aren't represented
- Some of those haven't gotten here yet
- Some of those haven't been very useful yet
- I'll be focusing on the methods of broadest usefulness, not coolest newestness

A word of note

- Just because a method is more recent or produces more complex models does not mean it's better

- With complex real-world data, more complex approaches tend to over-fit more to the noise in the data or the biases in the training sample

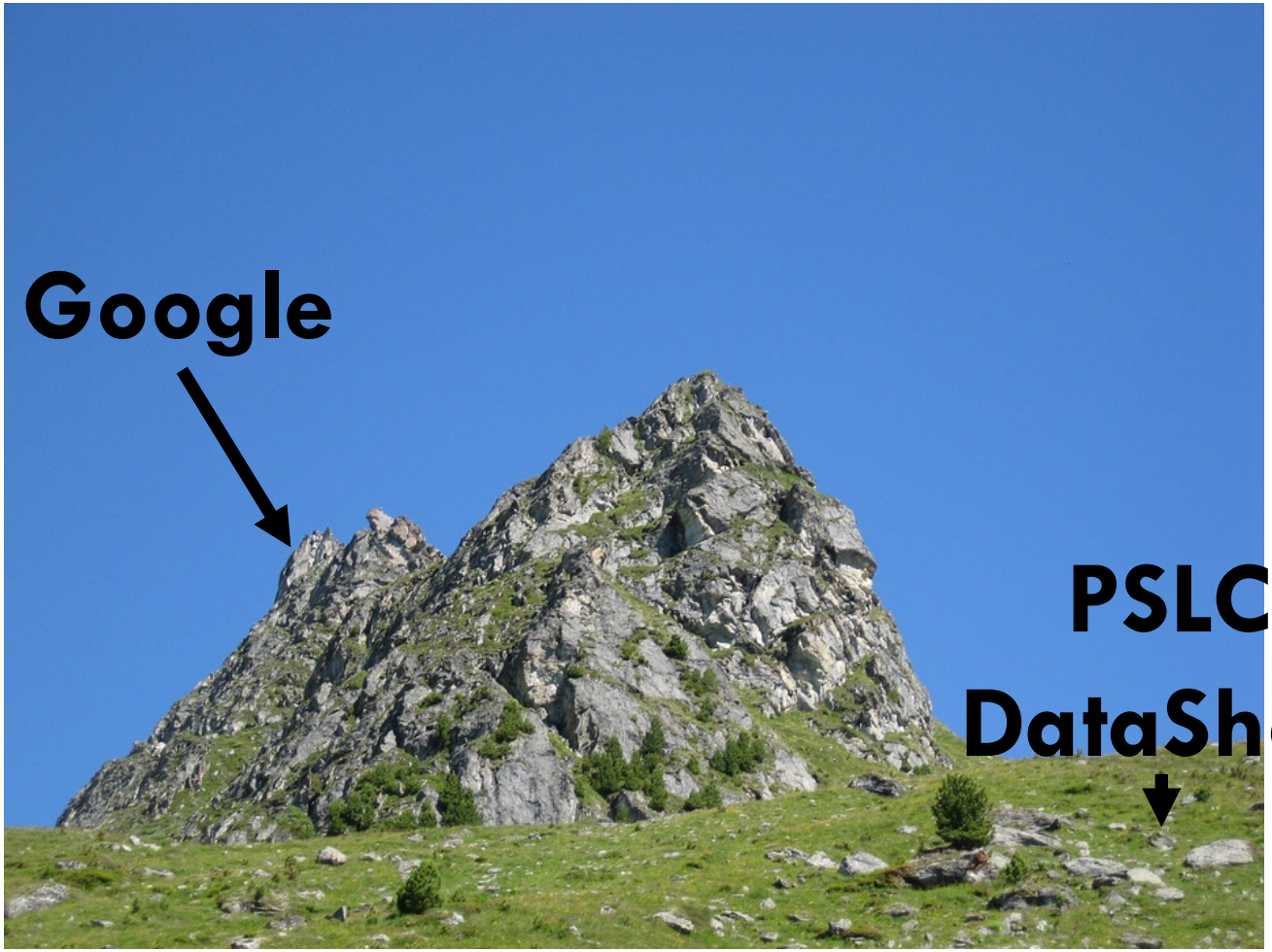
  (Hand, 2006, *Classifier Technology and the Illusion of Progress*)

What makes data “big”?

- Laney (2000) “The Three Vs”

- Volume
  - How much total data?
- Velocity
  - How fast is data coming in?
    (and how fast do you have to handle it?)
- Variety
  - Incompatible formats, non-aligned data structures, inconsistent data semantics

Is educational data big?



Public domain image from https://pixabay.com/p-215119/?no_redirect

Not that big?

- But the name of the course is big data in education!

Not that big?

- Big data in education ***is*** big
  - Big by comparison to most classical education research
  - Big compared to common data sets in many domains

- But it's not human genome project or google big

It *is* big enough

- That differences in r^2 of 0.0019 routinely come up as statistically significant
  (Wang, Heffernan, & Beck, 2011; Wang & Heffernan, 2013)

I will talk about statistical significance

- Sometimes
- But it will not be a focus of the class

I will talk about statistical significance

- Sometimes
- But it will not be a focus of the class

- Also: statisticians note, terminology is sometimes conflicting between stats and data mining/machine learning
  - I'll highlight particularly annoying cases where they emerge

Types of EDM/LA method

(Baker & Siemens, 2014; building off of Baker & Yacef, 2009)

- Prediction
  - Classification
  - Regression
  - Latent Knowledge Estimation
- Structure Discovery
  - Clustering
  - Factor Analysis
  - Domain Structure Discovery
  - Network Analysis
- Relationship mining
  - Association rule mining
  - Correlation mining
  - Sequential pattern mining
  - Causal data mining
- Distillation of data for human judgment
- Discovery with models

Prediction

- Develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables)

- Which students are off-task?
- Which students will fail the class?

Structure Discovery

- Find structure and patterns in the data that emerge "naturally"
- No specific target or predictor variable

Relationship Mining

- Discover relationships between variables in a data set with many variables

Discovery with Models

- Pre-existing model (developed with EDM prediction methods… or clustering… or knowledge engineering)

- Applied to data and used as a component in another analysis

Why now?

- Why didn't EDM emerge in the early 1980s, like bioinformatics?

A lot of reasons

- One of the key ones: not enough data
  - In the 1980s, collecting educational data was highly resource-intensive and difficult to scale
  - Much of the data that was easily collectible was purely summative in nature
  - Getting data on learning processes and learner behaviors, in field settings, required methods like
    - Quantitative field observations
    - Video recordings
    - Think-Aloud studies
  - None of which scale easily

Fast-forward to today

- Lots of standardized exams
  - Still summative in nature
- But lots of students now use internet-based educational software in class
  - Can be used to get at learning processes and learner behaviors
  - At a fine-grained scale (can log behavior at a second by second level)
  - Data acquisition is very scalable
- And there are these things called MOOCs which you may have heard of….

PSLC DataShop

(Koedinger et al, 2008, 2010)

- World's leading public repository for educational software interaction data
- >250,000 hours of students using educational software
- >30 million student actions, responses & annotations
  - Actions: entering an equation, manipulating a vector, typing a phrase, requesting help
  - Responses: error feedback, strategic hints
  - Annotations: correctness, time, skill/concept

Tools

- There are a bunch of tools you can use in this class.
  - RapidMiner is one tool you will need to learn in this course
    - Accessible to non-programmers
    - A large proportion of the power of Python or R
  - There is a walkthrough with instructions for getting started

Closing thoughts

- EDM/LAK methods emerging for big data in education
- In this class, you'll learn the key methods and how to use them for
  - Promoting scientific discovery
  - Driving intervention and improvements in educational software and systems
- Strengths & weaknesses of methods for different applications
- Is your analysis trustworthy? Is it applicable?