

Winter 2022 SySc 531
DATA MINING WITH INFORMATION THEORY (DMIT)
Online Synchronous Tues 5:30-7:20

MODEL almost ANY DATA!

- Do you have data in spread-sheet (rectangular) format, i.e., where variables are columns and cases* are rows?
- Might the variables have non-linear relations or complex interaction effects?
- Do you want to do exploratory modeling to discover these relations?

For example, if you have a set of input variables that might predict an output variable, where you don't know what the predictive relations are, do you want to discover these relations? If you answer 'yes' to the above questions, take this data mining course.

* Cases could be members of a population where variables are their attributes; or cases could be time points (or locations in space) at which variables are measured.

DMIT is a *project-based* course that offers you an opportunity to use information theoretic methods to analyze data. These methods are implemented in a software package named Occam, developed at PSU, which will be the analytical tool used in the course. The theory underlying these methods is taught in SySc 551/651 Discrete Multivariate Modeling (DMM), but *this* course (DMIT) is *stand-alone*, and does *not* have DMM as a prerequisite.

In DMIT, only the theory needed to understand the inputs and outputs of Occam will be presented. Occam's algorithms will not be explained. This is to make it possible for you to do exploratory modeling on data of interest to you without having first to master the underlying theory. Information about Occam, papers that illustrate its methodology (called RA: reconstructability analysis), and the Occam user manual are at the instructor's Selected Works (SW) site: https://works.bepress.com/martin_zwick/ in the Discrete Multivariate Modeling category; you see this category when you go to this site.)

If you want to understand RA theory, you can take DMM later, or you could look at the optional text at the Booksore: Krippendorff, Klaus (K). *Information Theory: Structural Models for Qualitative Data*. Series: Quantitative Applications in the Social Sciences, Paper # 62, Sage Publications, Beverly Hills, California, 1986 (ISBN 0-8039-2132-2, paperback) or read the tutorial "Overview" paper, available at the SW site and Canvas..

Required course readings from the SW site, also in Canvas, are: (i) the Occam manual, (ii) research papers and presentation slides discussed in class, and any additional material selected by the Instructor; there may also be videos you will need to watch.

Prerequisites: (a) Basic probability and statistics or machine learning (e.g., Math 105 or Stat 243 or equivalent), (b) access to data, ideally data that you *know something about and want to analyze*. For students who do not have their own data, the instructor will provide information on possible data sources (but will not be able to provide actual data). Analyzing data that *you* are interested in is *greatly* preferable to analyzing someone else's data that you don't care about.

Course requirements & Grades:

See the assignments described later in this syllabus. Grades will be based primarily on the final project report at the end of the course (75%), secondarily on the slide presentation of the project results (20%), and on class participation (5%, a grade tie-breaker, if needed). Reports must be **emailed** to the instructor and course TA as Word or pdf files on or before their due dates; similarly the presentation should be emailed as a PowerPoint file.

Late submission of the draft report (Assignment #4, due February 15) will receive limited or no feedback.

If your final course grade is right on the cusp between two adjacent possible grades, late submission of the final research paper or slide presentation (Assignment #5, due March 8) will likely cause your grade to be the lower of the two.

The draft and final reports (Assignments #4 and #5) should be single documents; any supplemental information should be in appendices of this document, not in separate files. Reports and presentations may be publicly shared, but if you want your report and/or presentation *not* to be shared with anyone or to be shared only in a limited way, you need to make an explicit request to the instructor for your work not to be shared.

Expectations about class attendance:

This course used to be given live twice a week, but has been reformatted to make it easier for people working full time to take it. Although classes will be recorded and available via Canvas, it is desirable that you attend the first three lectures because you'll get more from attending (remotely) than just watching the recording.

Ideally, you should attend all the Tues synchronous sessions, but at the very least you are expected to attend on Jan 18th to give your 5 minute project description, on Feb 1st to give your 5 minute progress report, and Mar 8th & 15th to make your 10 minute presentation and watch/hear others present. Even though your presentation will be short, everyone is entitled to an audience. Your remote attendance is expected to be live, i.e., you should be visible on the Zoom screen. Just participating via audio with a screen showing only your name is not conducive to a satisfactory class experience.

Teaching Assistant:

Alexander York (AY) (alexYork@pdx.edu) will be a teaching assistant for this course. You'll be able to interact with him via email, and Zoom time with him will be scheduled.

Course Outline [**In red: specific things that you need to do**]

Email assignments to MZ & AY (zwick@pdx.edu , alexYork@pdx.edu)

Weeks [1-3] Lectures providing the main knowledge you need in this course.

Weeks [1-4] Get your data, get familiar with Occam by repeating demo runs, prepare Occam input file & do a run on it

**[1] Jan 4 Introduction to DMIT: presentation on data mining with Occam.
After this class: watch video demo of Olson's binning program.**

**[2] Jan 11 More information on the RA methodology & Occam software;
Four demo Occam runs. If no time for this in class, watch the video**

**[3] Jan 18 Prototype RA research paper & presentation
Submit Assignment #1
Give a 5 min description of your project to the class
Repeat for yourself (not to be submitted) the four demo Occam runs.**

Additional information on RA/Occam will likely be given in all classes.

**[4] Jan 25 Students work on projects, with assistance from Instructor & TA.
You should have an Occam input file for your data no later than this day, have repeated the four demo Occam runs, and have done at least a first Occam run on your own data. Submit Assignment #2.
Assistance on projects by MZ will be *mainly* in class, with some limited email assistance. Assistance also provided by AY, to be scheduled.**

**[5] Feb 1 Continued work on projects
Submit Assignment #3 (one page progress report)
Give a 5 min progress report to the class**

[6] Feb 8 Continued work on projects

**[7] Feb 15 Continued work on projects
Submit Assignment #4: draft project report for comments/guidance.
Late papers will not receive comments.**

[8] Feb 22 Project work continues

[9] Mar 1 Project work continues

**[10] Mar 8 Submit Assignment #5: final report + slides
Lateness may affect grade
10 min slide presentations of projects**

[11] Mar 15 10 min slide presentations of projects (continued)

Email all assignments to both MZ & AY (zwick@pdx.edu , alexYork@pdx.edu)

Assignment #1 (due no later than Jan 18)

Include the following information:

- (a) Short (tentative) project title
- (b) *Very* short statement of objective of analysis (not more than a sentence or two about the *substantive* research question you want to explore)
- (c) Source of the data, e.g., web site, a PSU faculty member
- (d) Number of variables (#columns in spreadsheet)
- (e) Sample size (#rows in spreadsheet)
- (f) Whether or not you expect to use the binning utility program
- (g) Whether you plan to do directed or neutral analysis or both (directed-only is standard)

If there are questions you cannot answer, omit the items, but send answers later, as soon as you can. If you don't actually have your data set in hand, submit this assignment later, as soon as you have your data in hand. There's no point in giving us guesses about such questions as number of variables, sample size, etc.

Assignment #2 (due no later than Jan 25)

Include the following information:

- (a) Whether you've repeated the four demo Occam runs in Jan 4 slide presentation.
 - (b) Whether you have an Occam input file for your data
 - (c) Whether you've done at least a loopless Occam run on your Occam input file.
- If you haven't done all of these before Jan 25, email MZ&AY when you have done them.

Assignment #3 (due no later than Feb 1)

Short summary (maximum length: one page) of the status of your project.

Assignment #4 (due no later than Feb 15)

A *draft* of your final project report, as close as you can come to Assignment #5.

Assignment #5 (due no later than March 8)

Submit a final report plus slides for a 10 minute presentation. The report should be close to the type and quality that could be submitted to a scientific meeting. It should include: introduction (include a short discussion of the subject area and problem), data description (include details about binning of continuous variables), methods (strategy & steps in the RA analysis), results, and discussion. *You do not, however, need to include in your report any explanation of RA methodology.* In a typical project, the goal will be to find one or more good models by using the Search action of Occam, considering variable-based models without loops, variable-based models with loops, and state-based models. The Fit action of Occam should then be used to show what the selected model says.