DATA 420 Notes – Jan. 16, 2024

1 Administuff

- Class website: `http://stephendavies.org/data420`
- Syllabus
- Roll call
- Harmony Peura – grader
- Reading check #1

2 Overview

2.1 D.G.P.'s and "Data Mining in reverse"

Paradoxically, in Data Science *we normally don't care about the data.* We care about the real-world phenomenon that generated the data. This "phenomenon" can be thought of as a **system** or a **data generating process (D.G.P.)**

Data Mining (DATA 419) is about starting with the data, and using it to infer properties of the D.G.P.

Mod & Sim (this class) is about starting with the D.G.P. instead.

Two things that blow my mind about this whole field:

- It's possible to (meaningfully!) study a hurricane without a hurricane. (If this weren't true, mod & sim would be impossible.)
- The results of that study can be surprising and unpredictable. (If this weren't true, mod & sim would be pointless.)

2.2 Important terms

system "A coherently organized set of interconnected elements that achieves something." – Dana Meadows

model A simplified (a.k.a. "wrong") view of a system. ("All models are wrong. Some are useful." – George Box)

simulation An experiment performed on a model.

parameter An aspect of the simulation (quantitative or qualitative) that we can "tune" or "dial" to change its behavior.

complexity Neither regular and predictable, nor entirely chaotic, but somewhere in the middle, reflecting organization and purpse.

adaptive A system whose components change their behavior in response to their environment.

emergence A "surprising" phenomenon at the macro-level which is not easily traceable to micro-behaviors of the system. "The whole is more than the sum of its parts."

2.3 The purpose of modeling

1. To predict
2. To control
3. To understand

One can position various models on a continuum* from an **abstract model** at one extreme to a **fascimile** at the other. A fascimile is very precise model calibrated to a specific real-world data set. The goal is to try and replicate the behavior of the modeled system precisely so that predictions can be made. (Think economic forecasting. An example prediction might be: "if we reduce the top U.S. tax rate from 39.6% to 37% next year, we estimate the GDP to rise by 1.2%.)

On the other pole are abstract models designed to expose the general behavioral patterns of systems with certain structures. Such patterns are sometimes called "**stylized facts**." (Examples might be "the introduction of birth control in a society leads to less poverty" or "increased online communication leads to more entrenched political polarization.") Here the goal is not to try and calibrate the system to an actual precise scenario, but to understand what kind of behavior we might generally expect from other complex systems with its characteristics.

2.4 The micro/macro dichotomy

The "micro level" describes how individual elements in a complex system behave.

- In a market, how does a trader decide whether to buy or sell?
- In a nuclear reaction, how likely is a neutron to strike another neutron?
- In a social network, what influences a person to propagate information?

The "macro level" describes the resulting large-scale behavior of the entire system.

- Will the market reach an equilibrium, and if so how quickly, and what will the prices settle to?
- Will a chain reaction be sustained, and if so how much energy will it produce?
- Will a piece of news "go viral," and if so how rapidly?

Much of mod & sim is concerned with the question of *what macro-level behaviors arise from certain micro-level behaviors?* In many cases, this turns out to be extremely non-obvious, and yet it is at the heart of understanding anything complex.

*Suggested by Nigel Gilbert, a giant in the field.

2.5 Mod & sim paradigms

The "field" of mod & sim is large, and full of independent approaches from diverse fields. Depending on what you count, there have been principally *five* main "paradigms"[†]:

- **System Dynamics (SD)**
- Microsimulation (μsim)
- Cellular Automata (CA)
- Discrete-Event Simulation (DES)
- **Agent-Based Modeling (ABM)**

The boldface items are the ones we'll be focusing on this semester.

3 Python development

We'll be writing all our programs in Python this semester. To do this, you need an environment and toolset that lets you write and execute `.py` files.

If you already have such an environment and are comfortable with it, please continue to use it. (Note that if you're missing certain key Python libraries, like NumPy/SciPy or Matplotlib, you'll have to download and install those to your environment.) If you do not, the first assignment (already posted on the class web page's "XP" tab) has a pointer to the download page.

To Do

Complete the "Python workout" assignment, due Monday. This will involve getting your Python environment set up, if you've not already done so.

Crack open Nate Silver's amazing book and read the preface and introduction. **If your reading stamina is sub-par, do *not* try to do this all in one sitting. Split it into at least three manageable chunks.**

[†]There are also some other things that potentially fall under the mod & sim umbrella. Certain statistical techniques, often called "Monte Carlo" methods, are able to compute probabilistic estimates of intractable analytic expressions. Too, things like Markov chains can be considered "models", and can be simulated. We won't be covering those in this course.