DSSC Course Project

The third segment of the Data Science and Scientific Computing core course is entirely centered around one big project, which you will work on continuously for the entire length of the segment. The aim of this project is to combine three important aspects of data science into one complete and closed loop:

1. Data generation
2. Visualization
3. Quantitative data analysis

The first step creates interesting and complicated data. The next step gives us some qualitative analysis of the data by visualizing it on the computer. The third step is quantitative analysis, extracting useful numbers from the data. In practice, this process often loops back to the beginning — the data analysis often suggest ways to change the data generation step, which in turn leads to more analysis, etc. Each of these steps are explained in detail below.

1. Data generation

Many scientists generate data not through real-world experiment, but through mathematical models or computer simulations. This first step of the project involves writing a computer program (a mathematical model, a simulation, etc) which takes in a few parameters and outputs data.

Finding a suitable model to study is not an easy task. Some models generate data that are trivially simple and too easy to analyze, and some others generate data that are overwhelmingly complicated and pointless to analyze. We supply several suggestions for where to start in this project in the list below. This is not an exhaustive list, and we encourage you to deviate from this list and come up with your own pet problem to study. The most important aspect is that the output should be complicated enough that visualization and data analysis (steps 2 and 3) are interesting, non-trivial, and likely to provide insights into the data. The program you create should also run fast enough to generate plenty of data for you to analyze before the midterm project report.

Grading considerations: Your reports and presentations should clearly explain the program you write for data generation, emphasizing why the problem is interesting and how you implemented it. The data generation method should work reliably. Convince us that the method consistently generates valid data by safeguarding it against computational errors, excessive numerical noise, and nonsense parameters. We give extra consideration (bonus points) to models that are more difficult to implement in code. We also give extra consideration to programs that generate their data more efficiently than a straightforward approach. If you took extra effort to make your code run fast (through analytical or computational speed-ups), then call our attention to this fact in your project report and presentation.

2. Visualization

Once we have our data, we begin to make sense of it by visualizing it. Good, meaningful scientific visualization requires creative and non-trivial decisions, because
the data is often complicated and high-dimensional. The qualitative analysis achieved through visualization often suggests trends in the data that merit further investigation, or suggests interesting questions to ask. For this project, to make visualization interesting, we expect your data to be at least two dimensional. You should have many choices of which data to visualize and how to visualize it. You can visualize individual simulation runs, and you can visualize entire summaries of all simulations. The visualizations you submit for this project should paint a clear picture of the data, they should be non-obvious, they should show creativity, they should suggest a starting point for a discussion about your data, and they should suggest some clear avenues for a quantitative investigation of your data.

We expect you to investigate a number of techniques for visualizing high-dimensional data: which variables should you visualize? What is the most informative way to scale your axes? Can you use color or geometry in an insightful way? Does it help to animate your data? Do visualizations of vector fields, trajectories, or surfaces make sense for your data?

Grading considerations: You will submit your 2 best visualizations of your data and justify your decisions and techniques in both the written report and presentation. Bonus points will be given to particularly challenging, creative, and/or insightful visualizations.

3. Quantitative data analysis
Besides drawing insight from visualization, we can also draw insight through quantitative analysis. This often comes in the form of regression (finding the best simple function that explains the output data) or classification (deciding which input parameters lead to which types of output). After you extract these numbers or classifications from your data, you should be able to discuss why the results confirm your expectations, or why they are surprising (and what further questions they inspire).

For this project, we expect you to give us some quantitative results for questions that are particularly important or interesting for your model. An exact answer to a boring question is less important than an approximate answer to an interesting question.

Grading considerations: Focus on 2 different questions for your quantitative data analysis, and discuss them in your written reports and presentations. Convince us that you answered the problems carefully and accurately. Bonus points will be given to particularly thorough analyses, particularly challenging analysis techniques, and particularly insightful approaches.

Milestones and due dates:
- Due 30 May, 2018
  - Project proposal report:
    A 1-2 page document explaining the problem you wish to study. Explain the
data-generation technique in detail, discuss what will make the resulting data interesting, and report preliminary results. Speculate on how you plan to visualize and analyze the data.

- Project proposal presentation:
  A 7-minute presentation explaining the problem you wish to study. Explain the data-generation technique in detail, discuss what will make the resulting data interesting, and report preliminary results. Speculate on how you plan to visualize and analyze the data.

  *Do not go over the allotted time!* After the presentation, we will have up to 5 minutes for additional questions and discussion.

- Due 11 June, 2018:
  - Midterm report:
    A 2-3 page document broken up into 3 sections (Data generation, Visualization, and Quantitative data analysis). The first section should be just about finalized, as you should already have a completed data generation program by this point. Feel free to copy all or parts of this section from your previous report. The next two sections should have well-justified decisions and a detailed discussion. They do not need completed visualizations or analyses yet, but they should have intelligent speculations about what we should expect to see when the project is completed.

  - Midterm presentation:
    A 7-minute presentation re-introducing the problem, and discussing your choices for visualization and data analysis in more detail. Discuss what you hope to clarify or discover with your visualizations and analyses. Briefly report any visualization or quantitative analyses if you have them.

    *Do not go over the allotted time!* After the presentation, we will have up to 5 minutes for additional questions and discussion.

- Due 20 June, 2018:
  - Final report:
    A 3-4 page document broken up into 3 sections (Data generation, Visualization, and Quantitative data analysis). Feel free to copy parts of your previous reports. Showcase and discuss your two visualizations, and submit any supplementary material if necessary (like animations or 3D visualization programs). The quantitative analysis section should discuss the method and results in detail.

    We will give generous extra credit to projects which have gone through all three steps more than once (if the data analysis inspired questions which required you to generate more data or change your model, which inspired new analyses, etc.).

  - Final presentation:
    A 15-minute presentation re-introducing the problem, with a detailed emphasis on your visualization and data analysis. Convince the audience that the analyses were worthwhile by pointing out any surprises or insights that they gave you.

    *Do not go over the allotted time!* After the presentation, we will have up to about 5 minutes for additional questions and discussion.
Example projects

Below are some example models and simulations that we think are suitable for this project. Feel free to adapt them to what you think is interesting, or to propose something completely different!

Boids
This project simulates emergent crowd behaviour by using simple flocking rules. The term 'boids' referring to bird-like objects was introduced at SIGGRAPH 1987 by Craig Reynolds. This idea was further developed i.a. as steering behaviours. Implement original flocking rules (separation, alignment and cohesion) as well any additional rules you think might be interesting, and analyze behaviour of a flock (eg. breaking) for different setups. We can ask many questions about these simulated flocks:

- How does the initial geometry of the flock affect the final result?
- What tools can we use to analyze the flock shapes?
- What if some of the agents are predators, and the rest are prey?
- What if the flocks operate in 3D instead of 2D? What if they are mixed (e.g. owls operate in 3D while mice are constrained to 2D)?
- What if we add add "seeking goal" behaviors, like finding food? What causes/prevents them from achieving their goals, and how quickly?
- What are the most useful visualizations of these results?

Link:
- The original animations

Iterated Prisoner's dilemma
Simulate several members of a population interacting with one another. We can model each interaction with a simple "game" called a prisoner's dilemma. Each agent can either "cooperate" or "defect": When both agents cooperate, they both receive a reward; when both defect, they are both punished; when they both choose different strategies, the cooperator is punished, and the defector is rewarded. Many social interactions can be modeled by iterating this process many times (called an iterated prisoner's dilemma), nicely illustrated by this interactive website. We can ask many interesting questions about the results of different strategies and scenarios:

- What is the best strategy for winning an iterated prisoner's dilemma?
- How does cooperation evolve within a population?
- What if each agent remembers its interactions? How do you model this memory?
- What if each agent can observe other agents interacting?
- What if each agent interacts more with certain ones (like its neighbors) than others, instead of interacting with everyone equally?
- What are the effects of different rewards and punishments?
What are the effects of random miscommunication (accidental cooperation or defecting)?
What are the best ways to visualize these effects?

Population dynamics (Lotka-Volterra)
Lotka-Volterra model describes the population dynamics of species competing for resources and interacting as predators and prey. The simplest version of this can be modeled with a coupled set of differential equations:

\[
\frac{da}{dt} = \alpha + \gamma ab \\
\frac{db}{dt} = \beta + \delta ab
\]

which describes the dynamics of two population sizes \(a\) and \(b\) as the populations grow and interact with each other. The greek letters denote growth rates of the population, and the mixed \(ab\) terms describe what happens when different members of each population encounter each other: positive numbers mean the population grows, and negative numbers mean it shrinks. Simple models have already been well-studied, but there are many wide open questions if we make the model more complex:

- What if we add a “carrying capacity” term that limits the growth of each population (by limiting the amount of food available)? What if this carrying capacity changes over time (like with seasons, or as a function of something else)?
- What if the location of each population member depends on where it is in space? What if we model individual interactions discretely or probabilistically? Instead of modeling this as a system of ordinary differential equations, we can model it as a partial differential equation, as a flock of discrete agents, or as a social network graph.
- Can you think of any situations where particularly dramatic bifurcations (i.e. “catastrophes”) occur? Think population collapse, extinction, plagues, etc.
- What are the most compelling ways to visualize this data?

Molecular dynamics
We can model the behavior of matter at the molecular level by literally simulating every molecule. Molecular dynamics models each molecule as a particle, and each particle experiences forces based on its proximity to all other molecules in the system (like electric attraction and repulsion). We can ask many interesting questions about these systems:

- What conditions cause different states of matter (solid crystals vs. disordered gases)?
- What types of forces are necessary to create stable matter
- What are the best ways to visualize these results?
- Which types of changes (like orientation-dependent forces or more complicated potentials) lead to which emergent effects?
- Can you simulate a chemical interaction?
- What kinds of quantitative data (like reaction rates) can you extract from this simulation?
Astrophysics simulation
The movement of stars and galaxies takes a long time, and we usually are not around to observe interesting astrophysical effects. However, we can extract a lot of useful information from simulations. A simulation of multiple bodies attracted to each other by gravity can lead to all sorts of interesting questions like:

- What conditions cause certain structures (orbits, galaxies, etc) to form?
- What kinds of quantitative analyses can you perform to show that these structures are indeed stable?
- What types of topological structures can you observe (clusters, webs, surfaces, ...)?
- What conditions cause a stable galaxy or star to tear apart?
- How sensitive are your results to the type of numerical integrator you use? How important is conservation of energy for your results?
- What other forces can you add, and how do they affect the results?
- What are the most interesting ways to visualize your results?

Neuron simulation
According to Kirchoff’s law, the total current, $I$, flowing across a patch of a cell membrane is the sum of the membrane capacitive current and all the ionic currents

$$I = C \frac{dV}{dt} + I_{Na} + I_{K} + I_{L}$$

which relates the voltage across a synapse to the current due to the flow of sodium (Na), potassium (K), and some voltage leak through the membrane (L). These current terms are generally going to be functions of the voltage $V$, making this a non-linear ordinary differential equation. Please see chapter 2 of this book available online for an extensive discussion of different neural models that lead to all sorts of interesting questions:

- What kinds of current models are necessary for different behaviors?
- What kinds of limit behaviors (sinks, saddles, cycles) can we observe?
- How do the results depend on the initial conditions?
- What interesting visualizations can you show?
- What kinds of bifurcations can you create through small changes in parameters?

Diffusion Simulation
Many biological and physical processes rely on diffusion (chemical diffusion across a synapse, spread of genes throughout a population, viscosity in a turbulent fluid, etc). Behavior of 1D or 2D diffusion is relatively simple to analyze in a simple environment, but it can get difficult to analyze if we have complicated boundaries or a detailed environment (like diffusion rates that change in space or time). Simulate a diffusion equation

$$\frac{\partial}{\partial t} a = \frac{\partial}{\partial x} (D(x, t) \frac{\partial}{\partial x} a)$$

where $D$ is the rate of spread, and $a$ is the substance that is diffusing. Many different effects can happen if we vary the initial conditions and the diffusion rate.

- How do the initial conditions affect the results?
- How do boundaries and diffusion rates affect the results?
- What if the diffusion rate changes over time?
- What is the best way to visualize your results?
- What are some other terms (forces, reactions, etc) can you add to make the problem more interesting?
**Something else**
The previous project ideas are only suggestions. Do you have any interesting data generation and analysis questions that excite you, but are not on this list? Any ideas from your own research or rotations? Feel free to reuse some ideas from your previous work and use this project to build up some aspects that were previously lacking. If you have a simulator that creates relevant and important data, you can use this opportunity to explore novel visualization and quantitative analysis. However, please keep in mind that you are expected to work put in the same work as the other students; you should spend twice as much effort on the visualizations or quantitative analyses if you put less effort into the data generation portion of the project.

It is usually not advisable to start with a powerful tool (like a fancy data analysis technique) and search for a problem to show it off. The best results tend to come from techniques and visualizations that are optimized for the data.