Lecture 2: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia*

January 28, 2025 (last updated March 8, 2025)

Models are abstractions or approximations to *reality* (absolute truth, system, or a phenomenon).

For example, a "model airplane" is, of course, a model for an actual airplane. A map is a model for the real streets in a city. Quotes can also be thought of models, for example, the quote

> *Early to bed, early to rise makes a man healthy, wealthy, and wise.*

is a model for human success. A 20th century statistician George Box once said

> *All models are wrong, but some are useful.*

See Figure 1, depicting the Earth and a model to it. The goals of models are:

(1) **Predictions**: This is the goal of this class.

(2) **Explanation**: This is the focus of other courses such has MATH 343.

In **reality**, we can perform **measurements** if we have the right tools, and we will obtain **data** (see Figure 2). From the model, you can make **predictions** (see Figure 3). A natural question is: are our predictions **accurate**? To determine this, we compare the predictions to the data that we gathered and we perform **validation**. However, a very important fact to remember is that *predictions are not reality*! Data is the result of a system phenomenon in reality. If the predictions are close to it, then we have done well.

*Based on lectures of Dr. Adam Kapelner at Queens College. See also the course GitHub page.

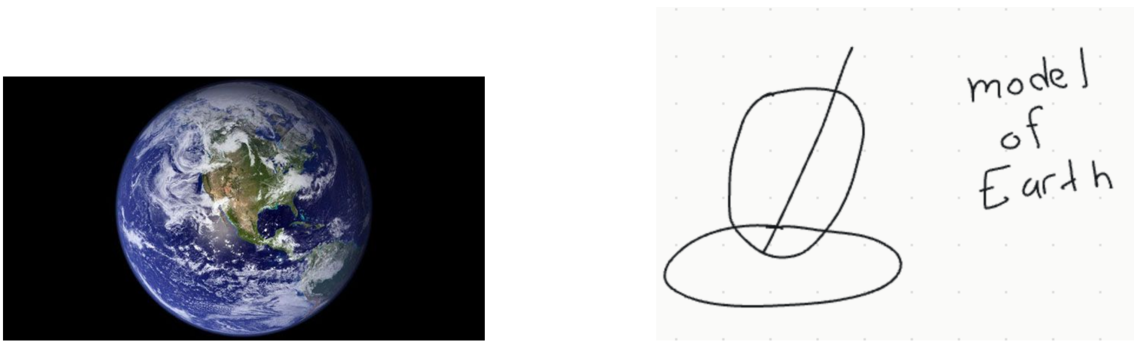

Figure 1: The Earth and a model of the earth (image by NASA)

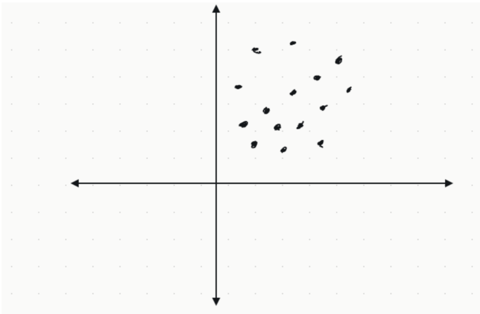

Figure 2: Data measurements in some phenomenon.

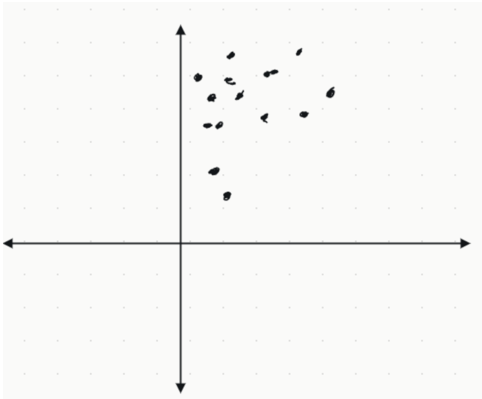

Figure 3: Predictions based on model to some phenomenon.

Consider again the quote on human success. We can think of it as as a function:

$$\begin{bmatrix} \text{health} \\ \text{wealth} \\ \text{wisdom} \end{bmatrix} = f\left(\begin{bmatrix} \text{bedtime} \\ \text{waketime} \end{bmatrix}\right)$$

For now, we will say f is **deterministic**. A problem with the formulation above is the fact that the english language is ambiguous. Bedtime is ill-defined, as well as health, wisdom, wealth, and waketime. THe model from the quote does not tell us how to measure the inputs or outputs.

How can we make this model exact? We need a **metric**, i.e., a way to exactly, unambiguously numerically measure all inputs and outputs. For this model, we need 5 metrics:

(1) **Bedtime**: Perhaps we can say seconds, average all bedtimes since some age, such as 22.

(2) **Waketime**: We can use a similar metric as bedtime.

(3) **Health**: Medical literature invented a quality-of-life metric that we can use.

(4) **Wealth**: Maybe your net worth, or rate of change of your financial state.

(5) **Wisdom**: Maybe the number of books you have read.

As you can see, coming up with metrics is difficult. Nevertheless, once you have defined metrics, these become scalars, and we can get on to modeling. What we do learn from this model is that it is not very operational. We don't say "bad", however, because it is not "bad" until we use it to make predictions.

Mathematical models are a subset of all models. Mathematical models will be the focus of this course because they are ambiguous. Mathematical models go back to about 2000 BCE. When you do science, you are building models, and we have built amazingly accurate mathematical models, such as:

$$F = ma \quad \implies \quad a = \frac{F}{m}$$

This model uses the mass of an object and the force applied to it as inputs, and computes the acceleration of the object as an output. Another famous model is $E = mc^2$. The universe is explicable mathematically, regardless of whether you believe it. For the purposes of this class:

- Phenomena *are* explicable mathematically.
- One output for the functions in our models.

We will often use the following notation:

$$y = t(z_1, z_2, \ldots, z_t)$$

Their meaning is:

- $z_1, z_2, \ldots, z_t$ are the true, causal input information.
- y is the output, phenomenon, response, outcome, or dependent variable.
- t is the exact functional relationship.