# MATH 342W / 642 / RM742 Spring 2025 HW #1

Sergio E. Garcia Tapia

Friday 7th February, 2025

## Problem 1

These are questions about Silver's book, the introduction and chapter 1.

(a) [easy] What is the difference between *predict* and *forecast*? Are these two terms used interchangably today?

To forecast means to conjecture that something will occur based on previous observations, information, or experience that a person may have. According to Silver, on page 5, the terms are used interchangeably today. He mentions that the term *predict* had a more otherworldly meaning, such as a fortune teller telling you that they see something in the future without necessarily any evidence or scientific justification.

(b) [easy] What is John P. Ioannidis's findings and what are its implications?

On page 11, Silver shares that Ioannidis concluded it would not be possible to successfully carry out many of the experiments that are posted in peer-reviewed journals related to biomedical research. It implies that while the theoretical results in research may be useful for modeling the world, we ought to be careful in the predictions that we make based on them. If we misinterpret the data that we gather or make invalid assumptions in our model, then this will turn affect the accuracy of our scientific predictions.

(c) [easy] What are the human being's most powerful defense (according to Silver)? Answer using the language from class.

According to Silver, it is the fact that we "simplify the world in ways that confirm our biases". In other words, when we are awash with information, we make interpretations that are convenient to us as if we wanted to validate a manufactured "truth", even if the data says otherwise.

(d) [easy] Information is increasing at a rapid pace, but what is not increasing?

According to Silver, the amount of *useful* information is not increasing, since he argues that most of it is "noise". As a result, we have to spend a lot of energy sifting through volumes of data, especially given that we retain much less than we are awash with.

(e) [difficult] Silver admits that we will always be subjectively biased when making predictions. However, he believes there is an objective truth. In class, how did we describe the objective truth? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, t, z_1, \ldots, z_t, \delta, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1 \cdot}, \ldots, x_{n \cdot}$, etc.

In class we talked about creating mathematical models as abstractions for reality. In this context, the objective truth is represented by reality, such as the data that we measure and that results from a system phenomenon in reality. We assume that the universe is explained mathematically, so that a model can be described by the functional relationship $y = t(z_1, \ldots, z_t)$ that remains unknown to us. Our models of reality are based on observations captured by a set $\mathbb{D}$ of input-output pairs of type $(x_k, y_k)$ such that $y_k = f(y_k)$, where $x_1, \ldots, x_p$ are features that act as proxies for $z_1, \ldots, z_t$, and $f$ is the best functional relationship we can get from the inputs.

(f) [easy] In a nutshell, what is Karl Popper's (a famous philosopher of science) definition of *science*?

A discipline where hypotheses are falsifiable, in the sense that they can be "tested in the real world by means of a prediction". In other words, it should be possible to design an experiment and gather data that allows us to determine whether our assertions are valid or not.

(g) [harder] Why did the ratings agencies say the probability of a CDO defaulting was 0.12% instead of the 28% that actually occurred? Answer using concepts from class.

The ratings agencies assumed that the probability of their customers defaulting on their CDO was largely independent. However, in reality they were largely correlated as a result of housing prices rising.

(h) [easy] What is the difference between *risk* and *uncertainty* according to Silver's definitions?

Risk refers to a situation in which there is a nontrivial probability that it will not result in a favorable outcome. Uncertainty is when we are aware of the existence of risk, but we are unable to accurately quantify the probability associated with said risk.

(i) [difficult] How does Silver define *out of sample*? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, z_1, \ldots, z_t, \delta, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1 \cdot}, \ldots, x_{n \cdot}$, etc. WARNING: Silver defines *out of sample* completely differently than the literature, than practitioners in industry and how we will define it in class in a month or so. We will explore what he is talking about in class in the future and we will term this concept differently, using the more widely accepted terminology. So please forget the phrase *out of sample* for now as we will introduce it later in class as something else. There will be other such terms in his book and I will provide this disclaimer at these appropriate times.

Silver uses out-of-sample to mean that we apply our previous experiences to a given situation when in fact there are conditions that render our assumptions invalid. Our approximation $g$ of the functional relationships $f$ and $t$ are computed using the historical data $\mathbb{D}$ (e.g. driving sober). However, if we then try to use $g$ on future data that does not result from the same phenomenon (e.g. driving drunk), then the predictions from $g$ may not apply.

(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

By "precision", Silver means that data obtained from carrying out an experiment is consistent, where values obtained are relatively close to one another (clustering together). In "Introduction to Probability, Statistics, and Random Processes", Pishro-Nik, explains that variance is a measure of how spread out the distribution of a random variable is. Thus, higher precision as used by Silver corresponds to lower variance.

Silver thinks of high accuracy as obtaining a result in a model that closely behaves like the phenomenon it is trying to describe in real life. In his book, Pishro-Nik introduces the notion of bias in the context of estimators for quantities related to a random variable. For example if we have a random variable $X$ for which we are interested in the mean, an estimator is used to obtain a sample that has a distribution that is identical to $X$ and the sample is used to estimate the mean. The bias of the estimator is then used to describe how far on average it is from the real value. In this case, the model is the estimator, and the real-life phenomenon is the real value (for example, of the mean). A bias of 0 corresponds to high accuracy as meant by Silver.

## Problem 2

Below are some questions about the theory of modeling.

(a) [easy] Redraw the illustration of Earth and the table-top globe except do not use the Earth and a table-top globe as examples (use another example). The quadrants are connected with arrows. Label these arrows appropriately.

See Figure 1, where I have drawn the phenomenon of flying, and how airplanes are modeled after birds doing this activity. I labeled the horizontal axes with possible inputs such as the weight of the board, the drag force they experience, and so on, and the output consists of their flying altitude and speed.

(b) [easy] Pursuant to the fix in the previous question, how do we define *data* for the purposes of this class?

Data refers to observations or measurements we make while observing a phenomenon manifest itself in real life.

(c) [easy] Pursuant to the fix in the previous question, how do we define *predictions* for the purposes of this class?

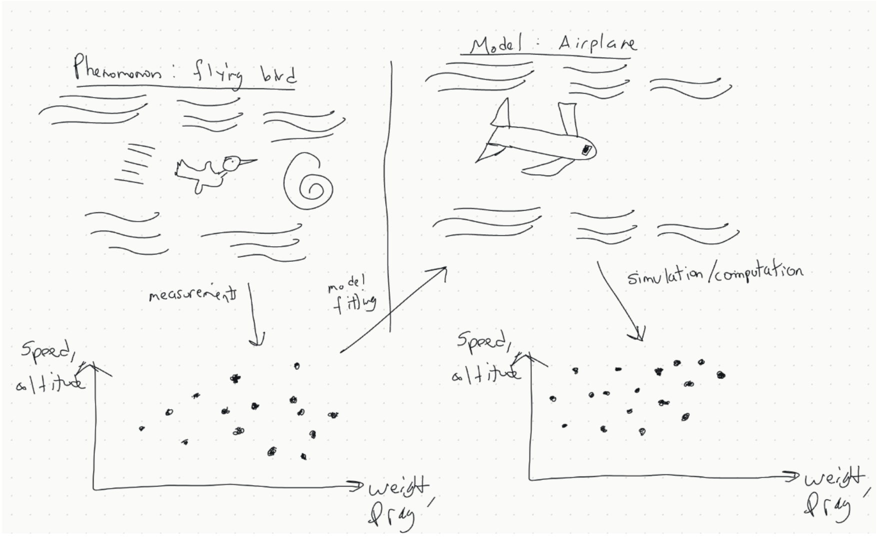

Figure 1: Modeling the phenomenon of birds flying with airplanes.

Predictions are inferences that we make after experimenting with the data that we observe. We obtain them from our model and they are meant to be approximations to real data.

(d) [easy] Why are "all models wrong"? We are quoting the famous statisticians George Box and Norman Draper here.

Models attempt to capture essential characteristics of phenomena, but they make simplifying assumptions in order to make it feasible to perform meaningful analysis. Whether knowingly or unknowingly, they do not account for complex processes and interactions that are present in the real-life phenomenon.

(e) [harder] Why are "[some models] useful"? We are quoting the famous statisticians George Box and Norman Draper here.

In spite of being wrong, models can provide insights that are still applicable. A model does not need to exactly match the phenomenon that it is based on, as long as its provides a close enough approximation for the purpose that one is trying to achieve. We use models to try to make predictions, and though they are certainly wrong, they can still be helpful for that.

(f) [harder] What is the difference between a "good model" and a "bad model"?

One difference is the level of ambiguity. A good model results in positive validation, where the predictions closely match the real-life data. A good model is also operational, with inputs and outputs that are measurable, so that we can form falsifiable hypotheses. A good model is unambiguous, with clear inputs and outputs, as well as the hypothesized relationship that connects them.

## Problem 3

We are now going to investigate the famous English aphorism “an apple a day keeps the doctor away” as a model. We will use this as springboard to ask more questions about the framework of modeling we introduced in this class.

(a) [easy] Is this a mathematical model? Yes / no and why.

It is not a mathematical model because it is ambiguous. It appears to be a model for human health, but the parameters that describe a person’s health are not completely clear, nor are other aspects such as how many apples or the type of apple.

(b) [easy] What is(are) the input(s) in this model?

A given person, a given date, and whether the person ate apples on that date.

(c) [easy] What is(are) the output(s) in this model?

A given date and whether the person went to the doctor.

(d) [harder] How good / bad do you think this model is and why?

It’s not great because it supposes that apple consumption is strongly related to an individual’s health. Under this model, someone who does not eat apples at all (perhaps because they dislike them) would need to visit the doctor often. It would be more accurate if we instead identified and measured specific nutrient of apples and the expected health benefit. The model also does not account for the specific reason why someone is visiting the doctor. For example, a person might visit the doctor often even though they eat a lot of apples, but their visit might be unrelated to their apple consumption.

(e) [easy] Devise a metric for gauging the main input. Call this $x_1$ going forward.

The number of apples eaten since the age of 2.

(f) [easy] Devise a metric for gauging the main output. Call this $y$ going forward.

The number of visits to the doctor’s office since the age of 2.

(g) [easy] What is $\mathcal{Y}$ mathematically?

$\mathcal{Y} = \mathbb{N} \cup \{0\}$, the set of natural numbers and 0.

(h) [easy] Briefly describe $z_1, \ldots, z_t$ in English where $y = t(z_1, \ldots, z_t)$ in this *phenomenon* (not *model*).

The causal information $z_1, \ldots, z_t$ describes a person’s health condition. It consists of a person’s medical history, symptoms they are showing, their diet, physical condition, and more.

(i) [easy] From this point on, you only observe $x_1$. What is the value of $p$?

$p$ is the number of features. Since we only consider $x_1$, we have $p = 1$.

(j) [harder] What is $\mathcal{X}$ mathematically? If your information contained in $x_1$ is non-numeric, you must coerce it to be numeric at this point.

$\mathcal{X} = \mathbb{N} \cup \{0\}$, since it accounts for the number of apples a person has eaten. It is the input space or covariate space.

(k) [easy] How did we term the functional relationship between $y$ and $x_1$? Is it approximate or equals?

In this class we are assuming that phenomena can be described mathematically. We said that $y = f(x_1) + \delta$, where $\delta$ is the error due to ignorance, and $f$ is the best functional relationship between the features and the response.

(l) [easy] Briefly describe *supervised learning*.

Supervised learning is an approach to learning from data where we take examples observed as a result of real-life phenomena, meaning that we have correct outputs corresponding to certain inputs. The result is a rule that we use to predict real-life phenomena based on inputs that do not come from the examples.

(m) [easy] Why is *supervised learning* an *empirical solution* and not an *analytic solution*?

The empirical solution $g$ in supervised learning is obtained by considering the behavior in a set of examples $\mathbb{D}$. The trouble is that we do not know the behavior outside of $\mathbb{D}$, and there may be multiple analytic solutions $f$ that agree with $g$ on $\mathbb{D}$. The supervised learning approach only tries to predict behavior outside of $\mathbb{D}$, not match it exactly, so the solution that it produces cannot be analytic.

(n) [harder] From this point on, assume we are involved in supervised learning to achieve the goal you stated in the previous question. Briefly describe what $\mathbb{D}$ would look like here.

The set $\mathbb{D}$ is a collection of input-output pairs $(x_k, y_k)$, perhaps collected by conducting a survey, where $x_k$ is the number of apples that a given person has eaten since age 2, and $y_k$ is the number of times that person has been to the doctor since age 2.

(o) [harder] Briefly describe the role of $\mathcal{H}$ and $\mathcal{A}$ here.

$\mathcal{H}$ is the hypothesis set of functions $h$ that are thought to closely approximate the behavior of the true analytic function $f$. The set encapsulates our assumptions about the functional form we think describes $f$ appropriately. $\mathcal{A}$ is the learning algorithm that we apply to arrive at an empirical solution $g$ from the hypothesis set $\mathcal{H}$ that best approximates $f$ by using the examples $\mathbb{D}$ observed from the real-life phenomenon.

(p) [easy] If $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$, what should the domain and range of $g$ be?

It should be the same domain as the functions in $\mathcal{H}$, and these in turn have the same domain as the real-life phenomenon $f : \mathcal{X} \to \mathbb{Y}$. Hence the domain is $\mathcal{X}$, the the range is $\mathcal{Y}$.

(q) [easy] Is $g \in \mathcal{H}$? Why or why not?

Yes. The set $\mathcal{H}$ has all of the hypothesis functions under considerations, those that we believe may conceivably well-approximate $f$, and the algorithm $\mathcal{A}$ picks what it believes is the best one among them.

(r) [easy] Given a never-before-seen value of $x_1$ which we denote $x^*$, what formula would we use to predict the corresponding value of the output? Denote this prediction $\hat{y}^*$.

We would use $\hat{y}^* = g(x^*)$.

(s) [harder] $f$ is the function that is the best possible fit of the phenomenon given the covariates. We will unfortunately not be able to define "best" until later in the course. But you can think of it as a device that extracts all possible information from the covariates and whatever is left over $\delta$ is due exclusively to information you do not have. Is it reasonable to assume $f \in \mathcal{H}$? Why or why not?

No, it is not reasonable. We only know information about the function $f$ in $\mathbb{D}$, and there are many functions, all with different properties, which may match $f$ at such input-output pairs. It's possible that the assumptions we make lead to a hypothesis set $\mathcal{H}$ that contains hypothesis functions that accurately approximate $f$ on $\mathbb{D}$ but do not do well outside of $\mathbb{D}$.

(t) [easy] In the general modeling setup, if $f \notin \mathcal{H}$, what are the three sources of error? Copy the equation from the class notes. Denote the names of each error and provide a sentence explanation of each. Denote also $e$ and $\mathcal{E}$ using underbraces / overbraces.

In class we said that

$$\begin{aligned} y &= t(z_1, \ldots, z_t) \\ &= f(x_1, \ldots, x_p) + \delta \\ &= h^*(x_1, \ldots, x_p) + \mathcal{E} \\ &= g(x_1, \ldots, x_p) + e \end{aligned}$$

The three sources of errors are:

- *Ignorance error* $\delta = t - f$: It is the error incurred by the fact that our features $x_1, \ldots, x_p$ cannot possibly capture all the information implied by the true drivers $z_1, \ldots, z_t$.
- *Misspecification error* $\mathcal{E} - \delta = f - h^*$: It is the error incurred by choosing a hypothesis set $\mathcal{H}$ that may not correctly capture the functional behavior of $f$. A more expansive $\mathcal{H}$ can lead to a decrease in this type of error.
- *Estimation error* $h^* - g$: Error incurred by not having enough data examples in $\mathbb{D}$. We can decrease this by increasing the amount of data (i.e. increasing $n$).

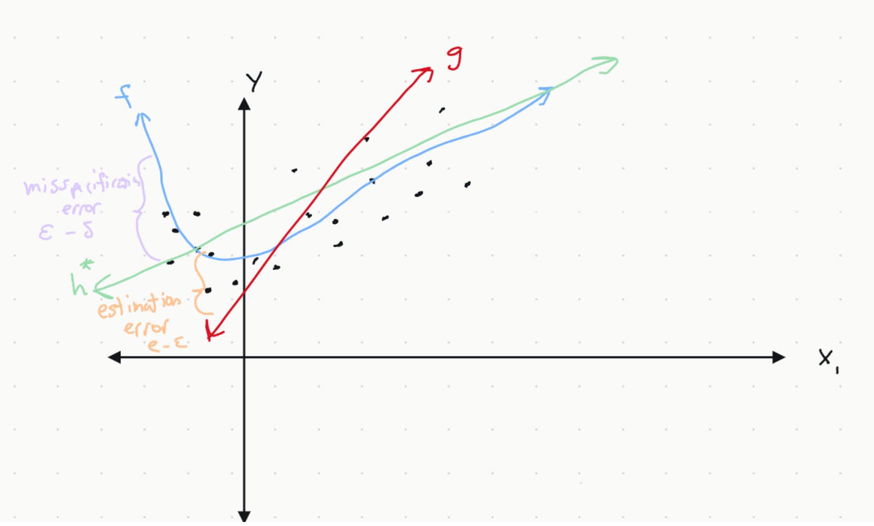

Figure 2: Plot of $f$, $h^*$, and $g$ for a general model (Question 3v).

(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

- *Ignorance error*: Increase the number of relevant features (increase $p$).
- *Misspecification error*: Choose a more expansive $\mathcal{H}$.
- *Estimation error* $h^* - g$: We can decrease this by increasing the amount of data (i.e. increasing $n$), or by improving the algorithm $\mathcal{A}$.

(v) [harder] In the general modeling setup, make up an $f$, an $h^*$ and a $g$ and plot them on a graph of $y$ vs $x$ (assume $p = 1$). Indicate the sources of error on this plot (see last question). Which source of error is missing from the picture? Why?

See Figure 2. The plot is missing $\delta$, the ignorance error. That error is due to using insufficiently-many features to capture the information from the true drivers. We cannot include them in the plot because aside from it not being possible to know $t$ and the $z$'s, it would not be possible to plot them $t$ in the same graph since it is a function of the $z$'s and not the features $x$'s.

(w) [easy] What is a null model $g_0$? What data does it make use of? What data does it not make use of?

A null model uses the mode statistic to determine the most common value. It is a constant function whose output is that most repeated value. It makes use of the outputs $\vec{\mathbf{y}}$ from $\mathbb{D}$, and it does not make use of the inputs $\vec{\mathbf{x}}_i$ from $\mathbb{D}$. The null model misclassifies any point whose output (response) does not equal the mode. It serves a reference for performance of our algorithm. Since the null model does not take any of the features into account, a good model that takes into account the features ought to have a smaller misclassification error than the null model. Otherwise, either our algorithm is poor, or our choice of features is poor.

(x) [easy] What is a parameter in $\mathcal{H}$?

It is a quantity that helps to identify a candidate function in $\mathcal{H}$, which our algorithm tries to compute. For example, in the threshold model, the parameter is a real number $\theta$ that determines the smallest value for which our model would predict 1 as the response.

(y) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. What would the null model $g_0$ be and why?

The null model would be

$$g_0 = \text{Mode}[\vec{\mathbf{y}}], \quad \vec{\mathbf{y}} \in \mathcal{Y}$$

That is, it is a constant function whose value is either always 0 or always 1, depending on which is more frequent among the values in our data set $\mathbb{D}$.

(z) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. If we use a threshold model, what would $\mathcal{H}$ be? What would the parameter(s) be?

We would have

$$\mathcal{H} = \{\mathbb{I}_{x \geq \theta} : \theta \in \mathbb{N} \cup \{0\}\}$$

Here, $\mathbb{I}$ is the indicator function. The parameter is $\theta$, the minimum number of apples (the threshold) needed to determine whether the doctor “kept away" or not (the response).

(aa) [easy] Give an explicit example of $g$ under the threshold model.

We could have $g = \mathbb{I}_{x \geq 2}$. That is, if you eat 2 or more apples, you keep the doctor away. Otherwise, you're due for a visit to the doctor.

## Problem 4

As alluded to in class, modeling is synonymous with the entire enterprise of science.

In 1964, Richard Feynman, a famous physicist and public intellectual with an inimitably captivating presentation style, gave a series of seven lectures in 1964 at Cornell University on the “character of physical law”. Here is a 10min excerpt of one of these lectures about the scientific method. Feel free to watch the entire clip, but for the purposes of this class, we are only interested in the following segments: 0:00-1:00 and 3:48-6:45.

(a) [harder] According to Feynman, how does the scientific method differ from learning from data with regards to building models for reality? (0:08)

In the scientific method, a hypothesis is invalid as soon as the experiment fails to validate it. However, in learning from data, we expect that it will differ from data, and care more so about how closely the hypothesis approximates what occurs in nature.

(b) [harder] He uses the phrase "compute consequences". What word did we use in class for "compute consequences"? This word also appears in your diagram in 2a. (0:14)

We used the word "predictions", which are obtained from the model as a means of approximating data.

(c) [harder] When he says compare consequences to "experiment", what word did we use in class for "experiment"? This word also appears in your diagram in 2a. (0:29)

We used the words "data" and "measurements", which are obtained from observations of the real phenomenon in nature.

(d) [harder] When he says "compare consequences to experiment", which part of the diagram in 2a is that comparison?

The validation step, where we compare the data we observed to the simulations (predictions).

(e) [difficult] When he says "if it disagrees with experiment, it's wrong" (0:44), would a data scientist agree/disagree? What would the data scientist further comment?

A data scientist will agree because, as George Box said, "all models are wrong". However, the data scientist would insist that useful information can nevertheless be obtained. The data scientist might still use the information obtained from the simulations to aid in making predictions, but while being careful to remember that they are not the absolute truth.

(f) [difficult] [You can skip his UFO discussion as it belongs in a class on statistical inference on the topic of $H_0$ vs $H_a$ which is *not* in the curriculum of this class.] He then goes on to say "We can disprove any definite theory. We never prove [a theory] right... We can only be sure we're wrong" (3:48 - 5:08). What does this mean about models in the context of our class?

The fact that models will never be the absolute truth. It's possible that our model performs well, meaning that it is accurate when comparing to the observed data, and that the prediction that it makes turn out to match reality accurately. However, many underlying assumptions we make to construct such a model disregard complex dynamics of the phenomena they are based on, so we should be careful not to forget that they are limited.

(g) [difficult] Further he says, "you cannot prove a *vague* theory wrong" (5:10 - 5:48). What does this mean in the context of mathematical models and metrics?

It means that when we are creating a model, we must be precise and unambiguous in our formulation. There must be a way to test how the model performs so that it's possible to perform validations to ascertain what parts of it are accurate and what parts of it are not.

(h) [difficult] He then he continues with an example from psychology. Remeber in the 1960's psychoanalysis was very popular. What is his remedy for being able to prove the vague psychology theory right (5:49 - 6:29)?

He declares the need to specify ahead of time exactly how much love is "not enough' and how much love is "overindulgent".

(i) [difficult] He then says "then you can't claim to know anything about it" (6:40). Why can't you know anything about it?

Because it would not be testable or measurable. If the hypothesis about hating the mother has to do with how much love they received, then it should be possible to quantify that. If it is not quantifiable, it's not possible to measure the extent to which the amount of love made a difference.

Just to demonstrate that this modeling enterprise is all over science (not just Physics), I present to you the controversial theoretical political scientist John Mearsheimer. He's all over youtube and there's nothing special about this video that I will link here about Can China Rise Peacefully? Feel free to watch the entire clip, but for the purposes of this class, we are only interested in the following segments referenced in the questions which has nothing to do with China, only his theory of "power politics".

(j) [difficult] Is Mearsheimer's model of great power politics / international relations (i.e., modern history) 9:35-17:22 simple or complicated? Explain.

(k) [difficult] Summarize his ideas about limitations of his theory from 39:18-40:00 using vocabulary from this class.