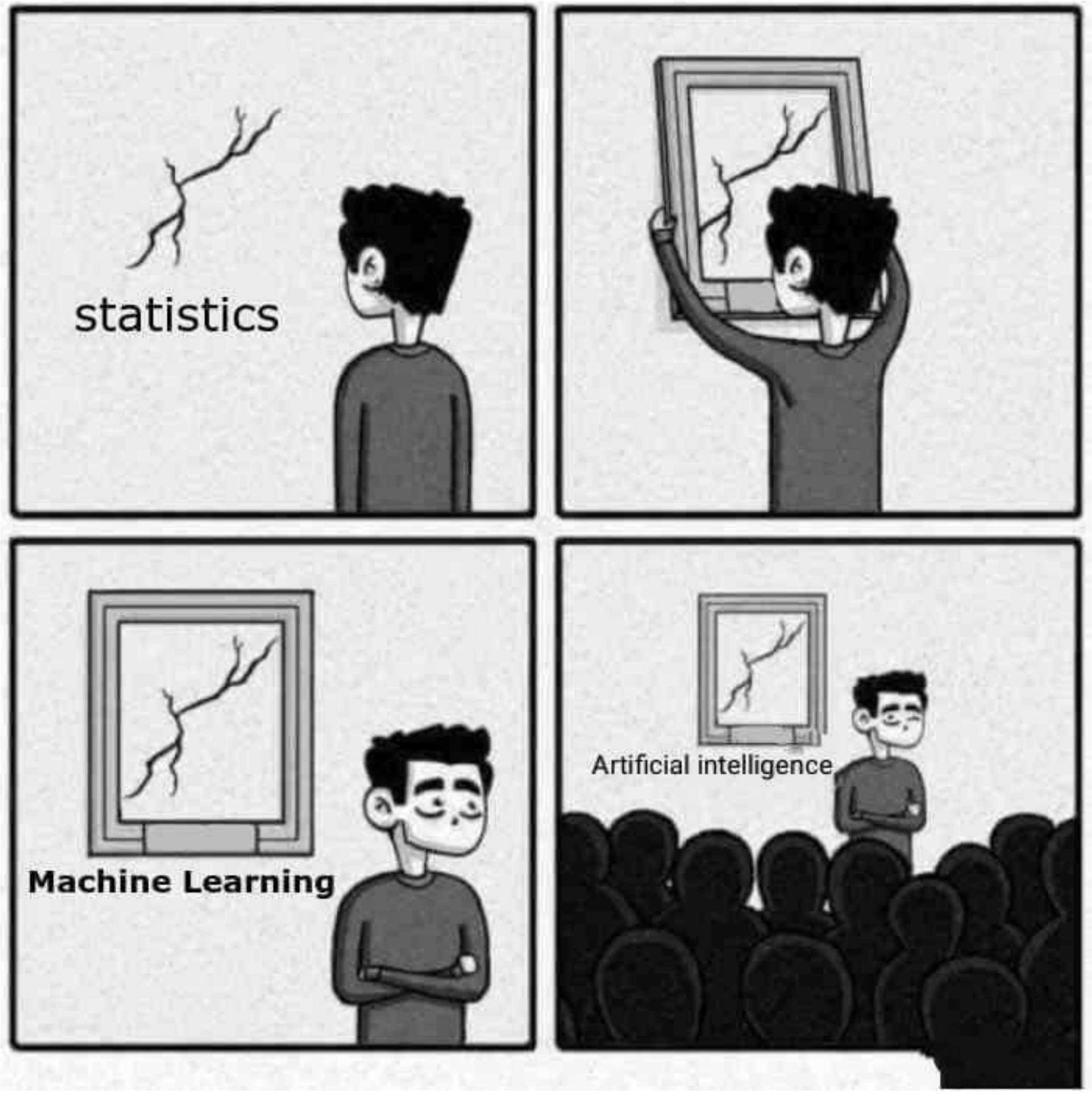
Chapter 2: Statistical Learning



Credit: https://www.instagram.com/sandserifcomics/

- Statistical machine learning is more than just statistics and it is more than just machine learning.
- we choose methods based on data AND our goals.

1 What is Statistical Learning?

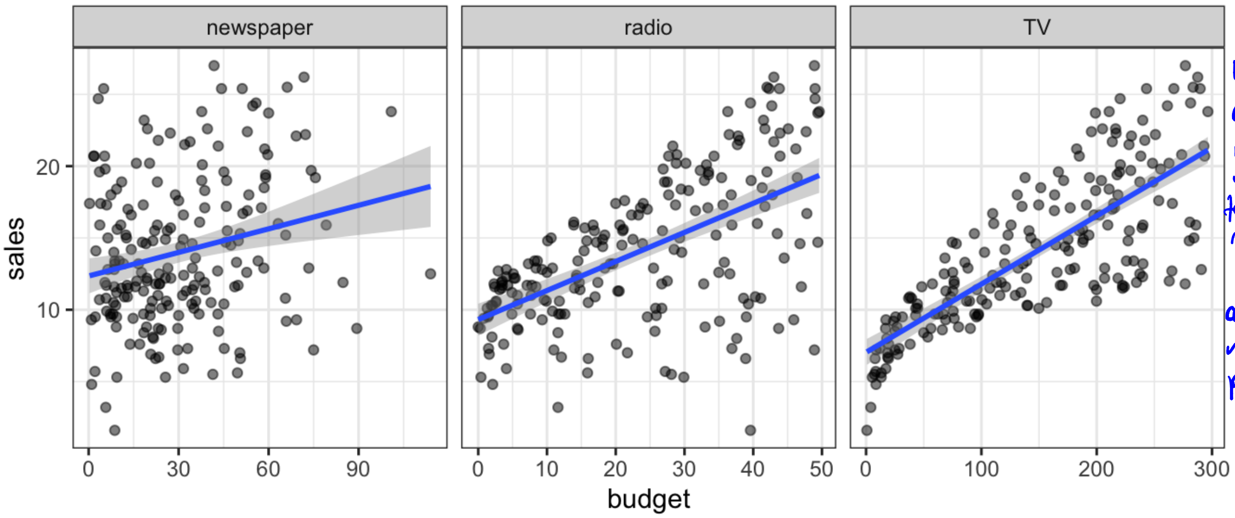
A scenario: We are consultants hired by a client to provide advice on how to improve sales of a product.

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5

4 markets

n=200

We have the advertising budgets for that product in 200 markets and the sales in those markets. It is not possible to increase sales directly, but the client can change how they budget for advertising. **How should we advise our client?**



If there is an association between sales and advertising, maybe we tell our client how to advertise to increase sales ⇒ develop an accurate model to predict sales based on 3 budgets.

input variables "predictors", "features", "independent variables"

advertising budgets

X_1 - TV

X_2 - radio

X_3 - Newspaper

y **output variable** "response", "dependendent variable"

sales

More generally – observe quantitative variable Y and p predictors $X_1, X_2, \ldots, X_p$

assume there is some relationship between predictors and Y.

$$Y = \underbrace{\overset{\text{fixed but unknown}}{f}(X)}_{\text{systematic information that } X \text{ provides about } Y} + \overset{\text{random error term, mean 0 and independent of } X}{e}.$$

f can involve more than one input variable (e.g. TV, radio, and newspaper)

Essentially, *statistical learning* is a set of approaches for estimating f.

1.1 Why estimate f?

There are two main reasons we may wish to estimate f.
goals for an analysis

Prediction

In many cases, inputs X are readily available, but the output Y cannot be readily obtained (or is expensive to obtain). In this case, we can predict Y using

$$\text{prediction for } Y \rightarrow \hat{Y} = \underset{\text{estimate of } f}{\hat{f}}(X)$$

remember errors e averages out to 0

In this case, $\hat{f}$ is often treated as a "black box", i.e. we don't care much about it as long as it yields accurate predictions for Y.
exact form not as important

The accuracy of $\hat{Y}$ in predicting Y depends on two quantities, *reducible* and *irreducible* error.

reducible: $\hat{f}$ is not a perfect estimate for f, but we can reduce error by using an appropriate statistical learning method to estimate it.

irreducible: Even if we estimated f perfectly (with $\hat{f}$) we would still have some error because $\hat{y} = \hat{f}(X)$ but Y is a function of e! We cannot reduce this no matter how well we estimate f.

Why? e contains unmeasured variables that would be useful for prediction of Y

We will focus on techniques to estimate f with the aim of reducing the reducible error. It is important to remember that the irreducible error will always be there and gives an upper bound on our accuracy.

Inference

Sometimes we are interested in understanding the way Y is affected as $X_1, \dots, X_p$ change. We want to estimate f, but our goal isn't to necessarily predict Y. Instead we want to understand the relationship between X and Y.

We may be interested in the following questions:

1.

2.

3.

To return to our advertising data,

Depending on our goals, different statistical learning methods may be more attractive.

1.2 How do we estimate f?

Goal:

In other words, find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y). We can characterize this task as either *parametric* or *non-parametric*

Parametric

1.

2.

This approach reduced the problem of estimating f down to estimating a set of *parameters*.

Why?

Non-parametric

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead we seek an estimate of f tht is as close to the data as possible without being too wiggly.

Why?

1.3 Prediction Accuracy and Interpretability

Of the many methods we talk about in this class, some are less flexible – they produce a small range of shapes to estimate f.

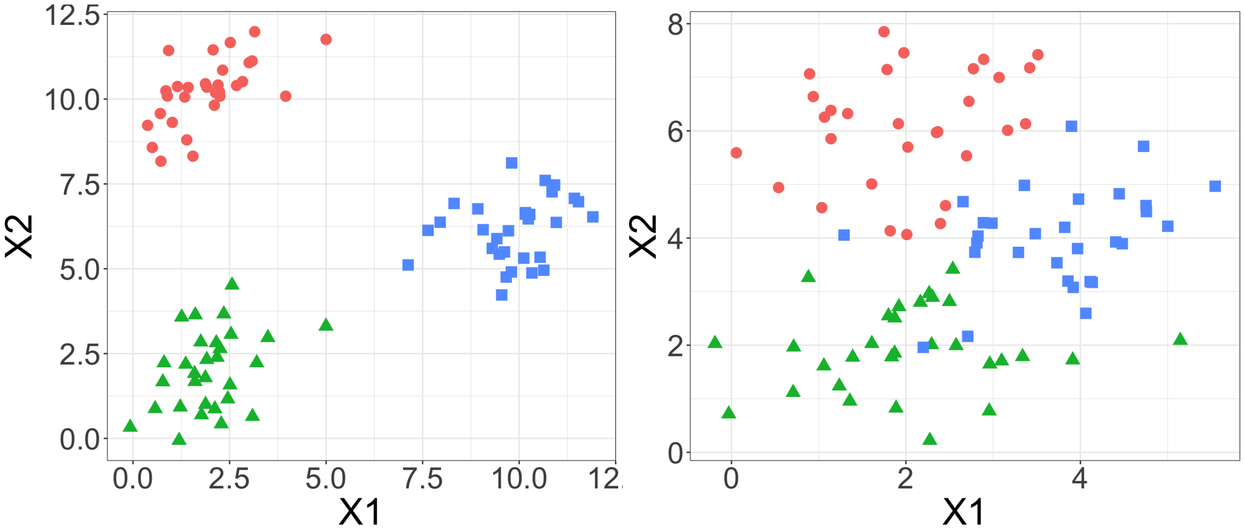
Why would we choose a less flexible model over a more flexible one?

2 Supervised vs. Unsupervised Learning

Most statistical learning problems are either *supervised* or *unsupervised* –

What's possible when we don't have a response variable?

- We can seek to understand the relatopnships between the variables, or
- We can seek to understand the relationships between the observations.



Sometimes it is not so clear whether we are in a supervised or unsupervised problem. For example, we may have $m < n$ observations with a response measurement and $n - m$ observations with no response. Why?

In this case, we want a method that can incorporate all the information we have.

3 Regression vs. Classification

Variables can be either quantitative or categorical.

Examples –

Age

Height

Income

Price of stock

Brand of product purchased

Cancer diagnosis

Color of cat

We tend to select statistical learning methods for supervised problems based on whether the response is quantitative or categorical.

However, when the predictors are quantitative or categorical is less important for this choice.