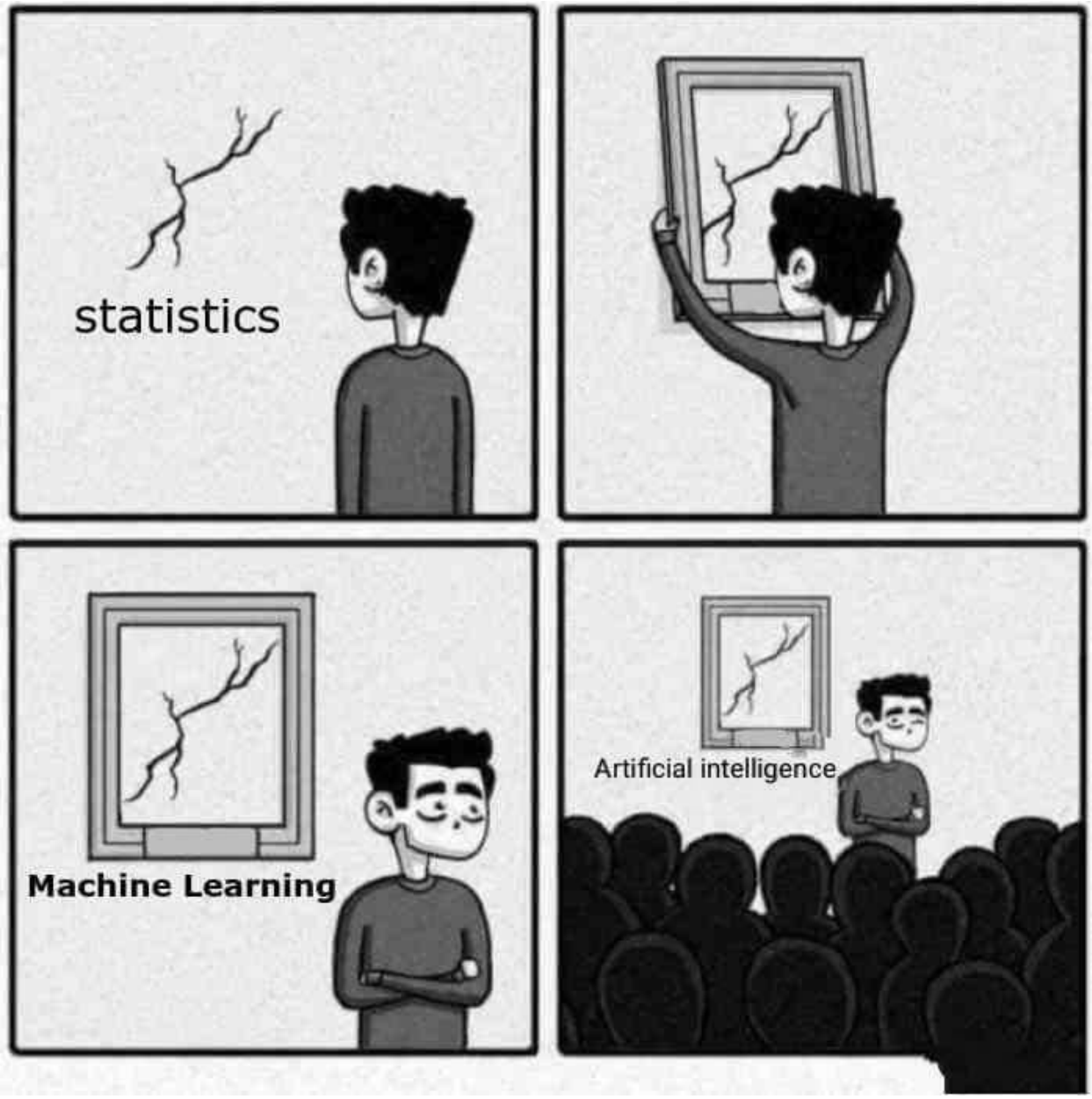
Chapter 2: Statistical Learning



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

statistical machine learning is more than just statistics and more than just machine learning.

We choose methods based on data AND our goals.

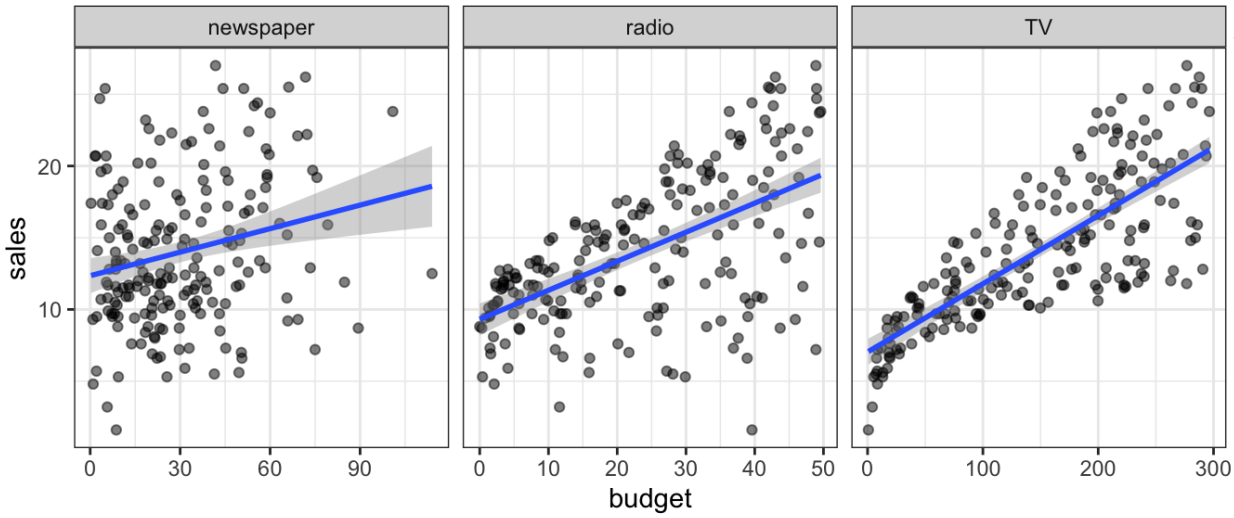
1 What is Statistical Learning?

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

X_1	X_2	X_3	Y
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
⋮			

$n = 200$

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 ads and sales we can tell client how to advertise to increase sales.
⇒ develop an accurate model to predict sales based on ad budgets.

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

advertising budgets.
X_1 – TV
X_2 – radio
X_3 – newspaper

output variable "response", "dependent variable"

Y – sales

More generally – Observe quantitative variable Y and p predictors $X_1, \ldots, X_p$.

Assuming there is some relationship between predictors $X_1, \ldots, X_p$ and response Y.

$$Y = f(X) + e.$$

f: fixed but unknown.

e: random error term, mean 0 and independent of X.

$f(X)$: systematic information that X provides about Y

f can involve more than 1 input variable (e.g. TV, radio, 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.

our 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

prediction for Y $\rightarrow \hat{Y} = \hat{f}(X)$. (remember error averages to 0).

$\hat{f}$: estimate of f

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 very 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. we can reduce error by using an appropriate statistical learning method to estimate it.

irreducible: Even if $\hat{f}$ was a perfect estimate we would still have some error since $\hat{y} = \hat{f}(x)$, but Y is a function of e! We cannot reduce this, no matter how well we estimate f.

Why? e contains unmeasured variables that could be useful for predicting Y, (also measurement error)

Consider an estimate $\hat{f}$ and predictors X (fixed).

expected value of squared difference b/w predicted and actual Y.

$$E[(Y - \hat{y})^2] = E\left[(f(x) + e - \hat{f}(x))^2\right]$$

$$= \underbrace{[f(x) - \hat{f}(x)]^2}_{\text{reducible.}} + \underbrace{\text{Var}(e)}_{\text{irreducible}}$$

variance of error term.

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.

almost always unknown in practice.

Inference

Sometimes we are interested in understanding the way Y is affected as $X_1, \ldots, X_p$ 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.

i.e. how Y changes as a function of $X_1, \ldots, X_p$

$\Rightarrow \hat{f}$ no longer a black box! We need to know its form.

We may be interested in the following questions:

1. Which predictors are associated w/ the response?
   often only a small fraction are substantially associated w/ response $\Rightarrow$ identifying important predictors can be useful.
2. What is the relationship btw/ response and each predictor?
   some predictors may have a positive (or negative) relationship w/ Y.
3. Can the relationship w/ Y and each predictor be adequately summerized by a linear relationship or is it more complicated?

To return to our advertising data,

Inference questions:

- Which media contribute to sales?
- Which media generate the biggest boost in sales? given
- How much increase in sales is associate w/ a $\checkmark$ increase in TV ads?

prediction question:

- What can I expect sales to be if we sped \$200k on TV ads and \$0 on newspaper & radio?

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

E.g. linear models allow for simple and interpretable inference but may not yield most accurate predictions.

highly nonlinear models can provide accurate predictions, but much less interpretable.

(inference is often challenging or impossible!)

1.2 How do we estimate f?

"training data"

"training"

We have observed n different data points & want to estimate f w/ $\hat{f}$

Goal:

apply a statistical learning method to training data to estimate unknown f.

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.

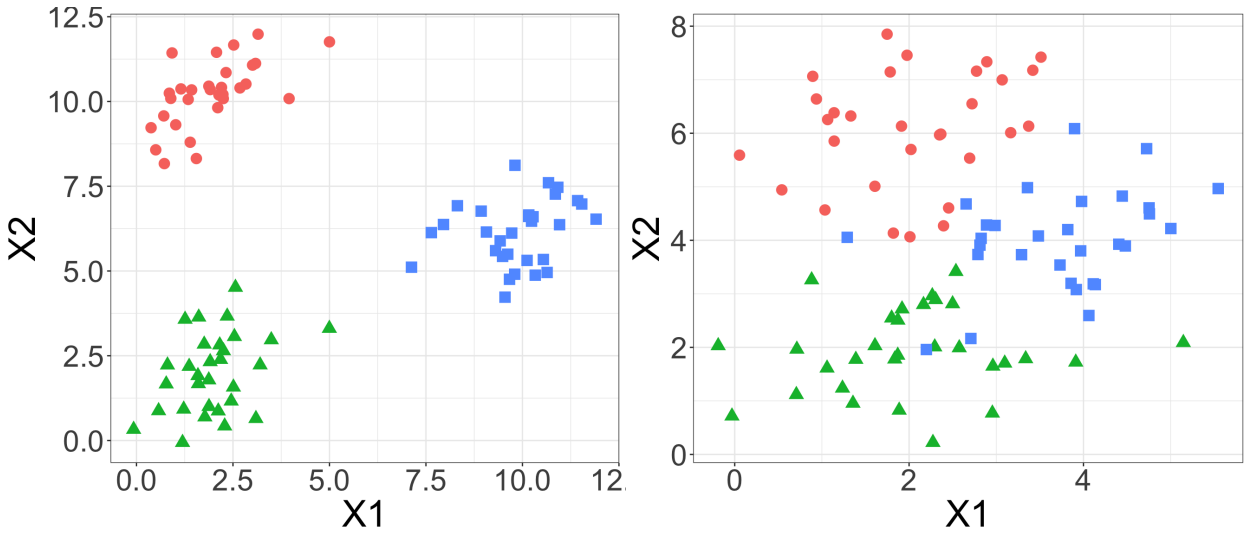
Why would we choose a less flexible model over a more flexible one?

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.