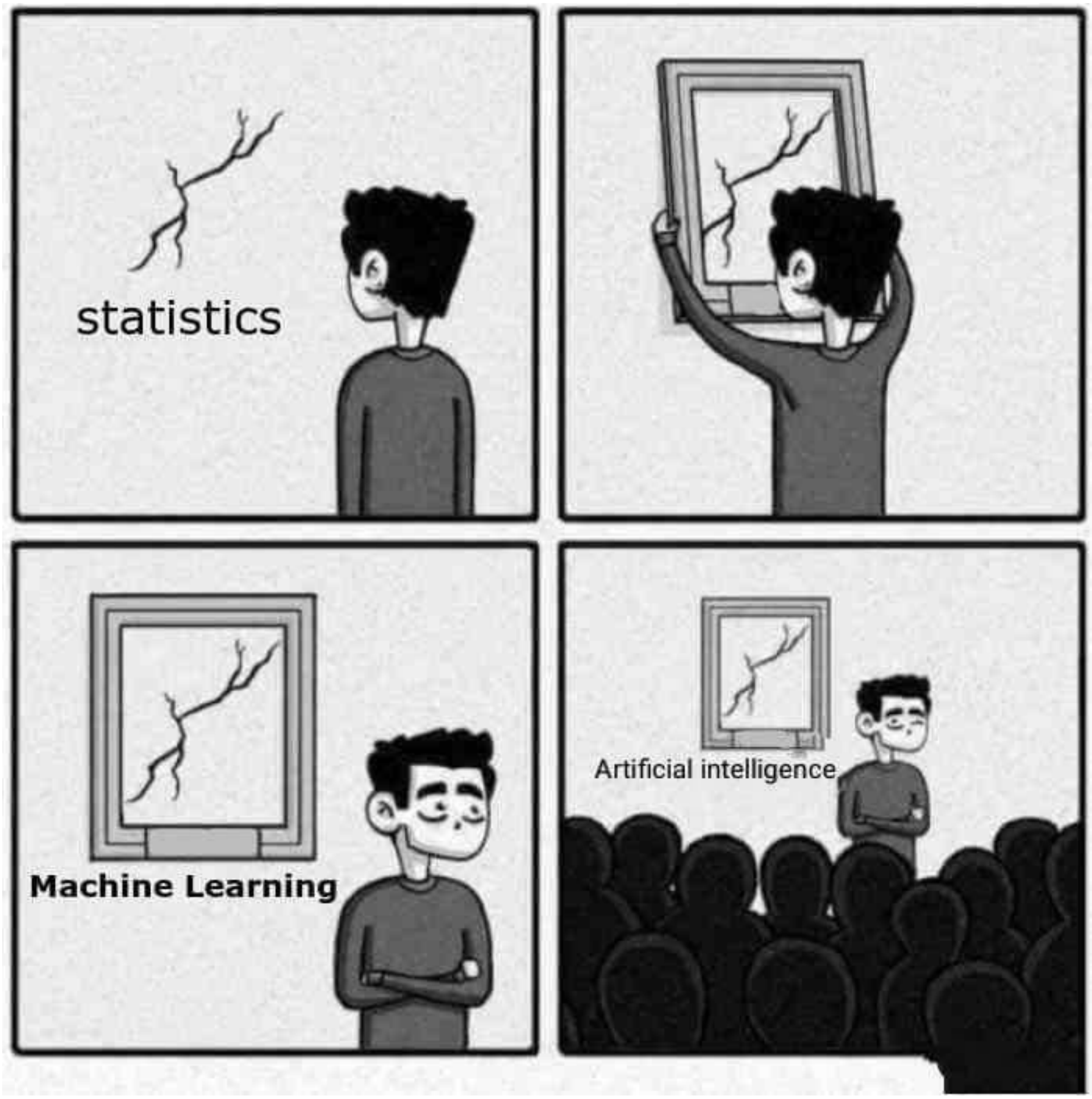
Chapter 2: Statistical Learning



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?

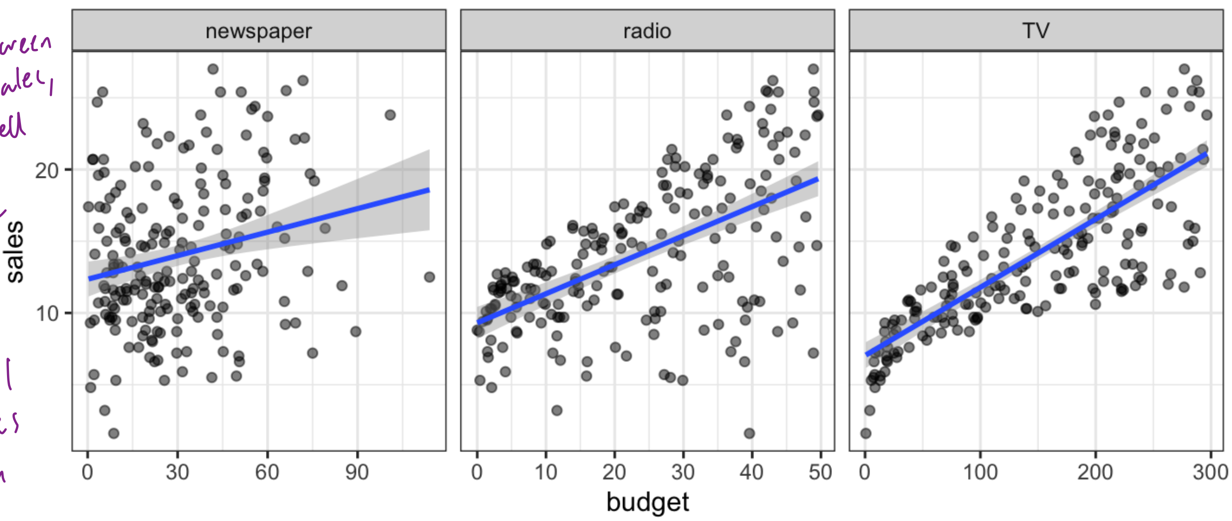
A scenario: We are consultants hired by a client to provide advice on how to improve sales of a product.

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 advertising and sales, maybe we can tell our client how to advertise to increase sales
⇒ develop an accurate model to predict sales based 3 media 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 – observing quantitative variable Y and p predictors $X_1, X_2, \ldots, X_p$

Assume there is some relationship between predictors and Y.

unknown, fixed

random error term mean 0 and independent of X.

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

systematic information that X provides about Y

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

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

$$\hat{Y} = \hat{f}(X)$$

estimate of f

* remember error averages 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 $\hat{f}$ was estimated perfectly we would still have some error because $\hat{Y} = \hat{f}(X)$ but Y is still 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 predicting Y.

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

expected value of squared difference between predicted & actual Y

variance of error term.

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

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

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 $X_1, \dots, X_p$ 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.

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

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

We may be interested in the following questions:

1. Which predictors are associated with the response?
   often only a small fraction of predictors are substantially associated w/ Y $\Rightarrow$ identifying those can be useful.

2. What is the relationship between the response and each predictor?
   some predictors may have a positive (or negative) relationship w/ Y.

3. Can the relationship between Y and each predictor be adequately summarized w/ a linear equation or is the relationship more complicated?

To return to our advertising data,

inferential questions:

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

predictive question:

- What can I expect sales to be if we spend \$200k on TV and \$0 on newspaper and 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 approaches can provide accurate predictions, but are much less interpretable (inference is very challenging or impossible).

"training data"

1.2 How do we estimate f?

We have observed n different data points and want to estimate (train) f w/ $\hat{f}$

Goal:

apply a statistical learning method to the training data in order to estimate unknown function 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. Make an assumption about the shape of f.

   e.g. $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

   parameters

2. Use training data to fit or "train" this model

   e.g. estimate $\beta_0, \beta_1, \ldots, \beta_p$ w/ ordinary least square (one of many choices).

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

Why?

This simplifies the problem of estimating f because its usually easier to estimate a set of parameters than fit some arbitrary function f.

Dis advantage:

What if the model we choose is very different than the shape of f?

Then the estimate (and predictions) will be poor.

We could try a more flexible model, but this means more parameters and can lead to "overfitting" $\Rightarrow$ fitting errors in training data too closely!

Non-parametric

shape

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.

technical term.

Why?

Advantage:

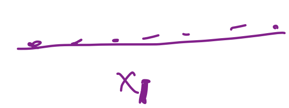
- fit a wider range of possible shapes for f.
- no restrictions on shape ⇒ we can't assume the wrong shape of f!

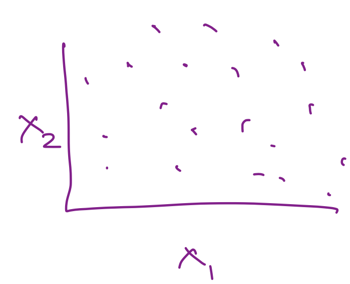
Disadvantage:

- they don't reduce the problem!

⇒ need a lot of data.

e.g. splines (ch. 7).





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.

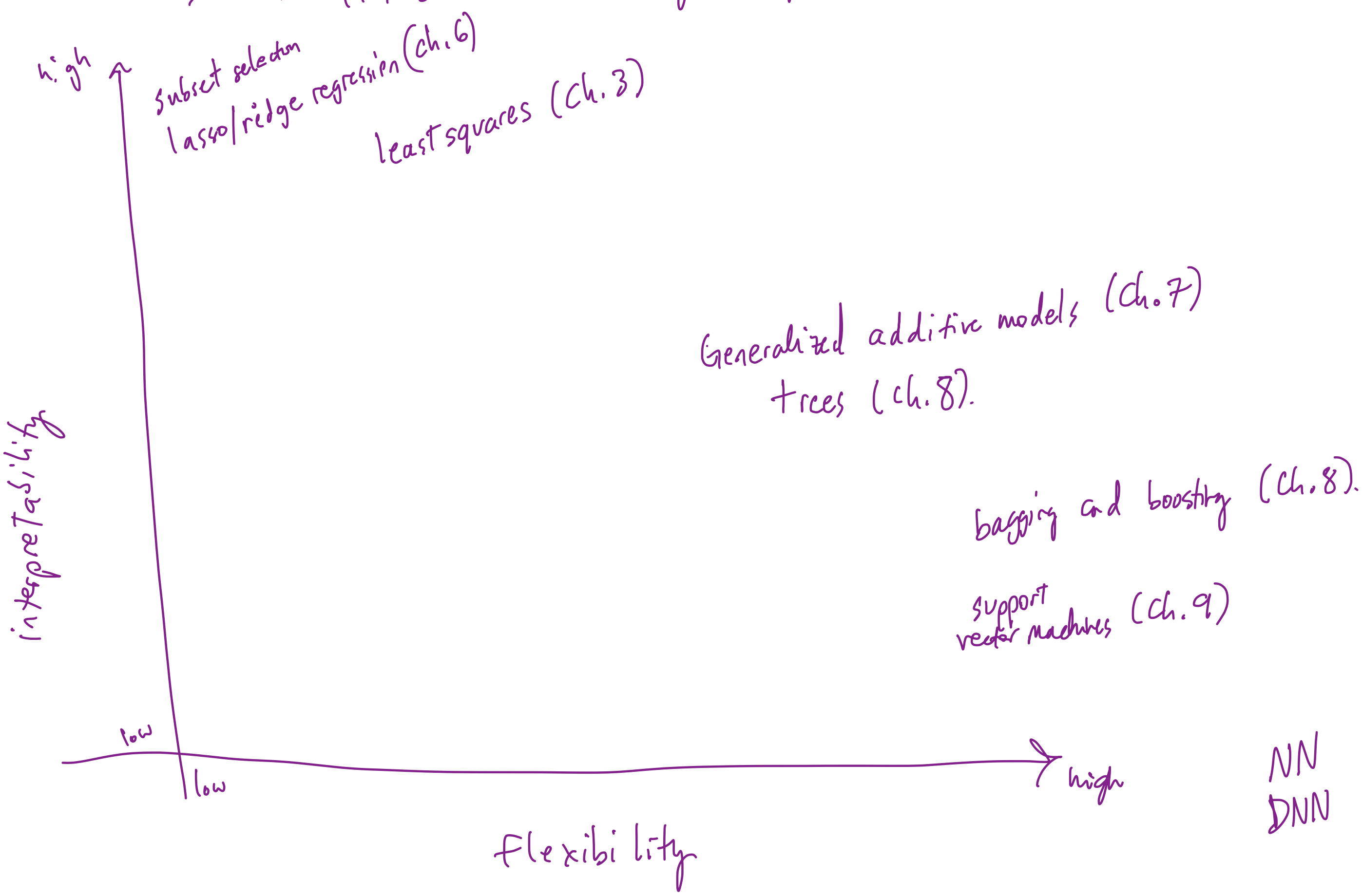
e.g. linear regression vs. splines

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

- If we care about inference, restrictive models are interpretable.

↳ flexible methods can lead to complicated estimates of f so that it is difficult to understand how any individual predictor is associated w/ the response.

In some settings we only care about prediction accuracy
⇒ more flexible model may be prefered.



2 Supervised vs. Unsupervised Learning

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

Supervised:

for each observation of predictors $\underline{x}_i$, $i = 1, \dots, n$ there is an associated response y_i

goal: fit a model that reflects the relationship between response and predictors.

maybe for inference or prediction.

methods: OLS regression, logistic regression, LASSO, GAM, boosting, SVM, etc.

Unsupervised

for each observation $i = 1, \dots, n$ we have a vector of measurements $\underline{x}_i$,

but no response y_i.

e.g. cancer example from Ch. 1.

goals: clustering.

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. cluster analysis.

based on observations $\underline{x}_1, \dots, \underline{x}_n$ discern if fall into distinct groups.

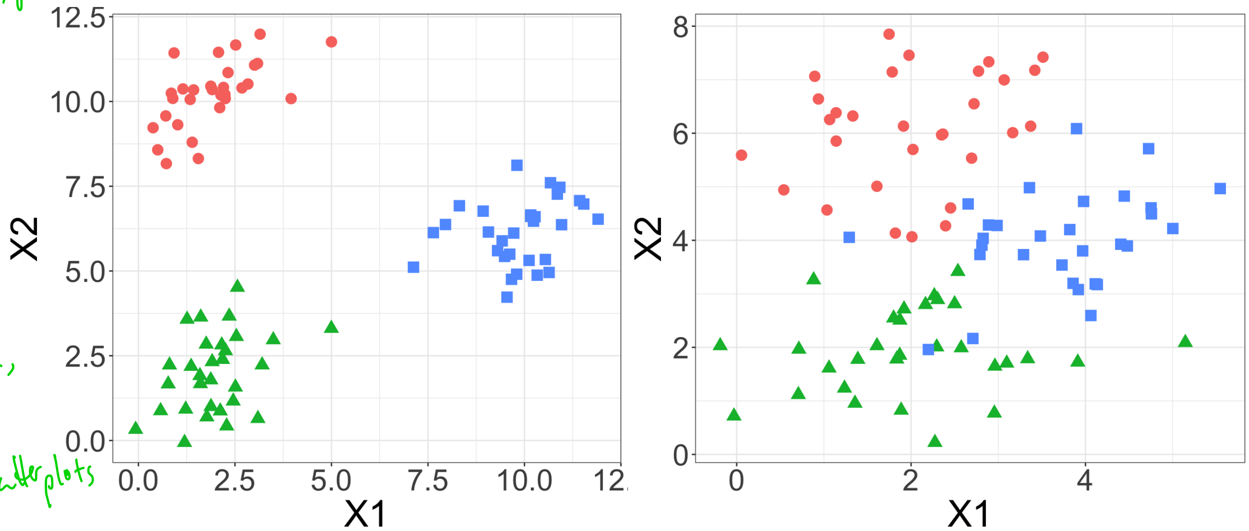
easy to plot

n=90 of p=2

3 true groups.

When p>2, leads to $\frac{p(p-1)}{2}$ scatterplots

may need more automated methods (ch. 10).



Well separated groups would be easy to cluster

overlapping groups harder to cluster.

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?

Missing values

Maybe it's expensive to collect y but not x.

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

"semi-supervised" methods.

outside scope of this class.

3 Regression vs. Classification

response

Variables can be either quantitative or categorical.

numeric

one of K differet classes or categories.

Examples –

Age quantitative.

Height quantitative.

Income quantitative.

Price of stock quantitative.

Brand of product purchased categorical

Cancer diagnosis categorical

Color of cat categorical

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

"regression" "classification"

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

most methods in this course can use quantitative or categorical predictors.