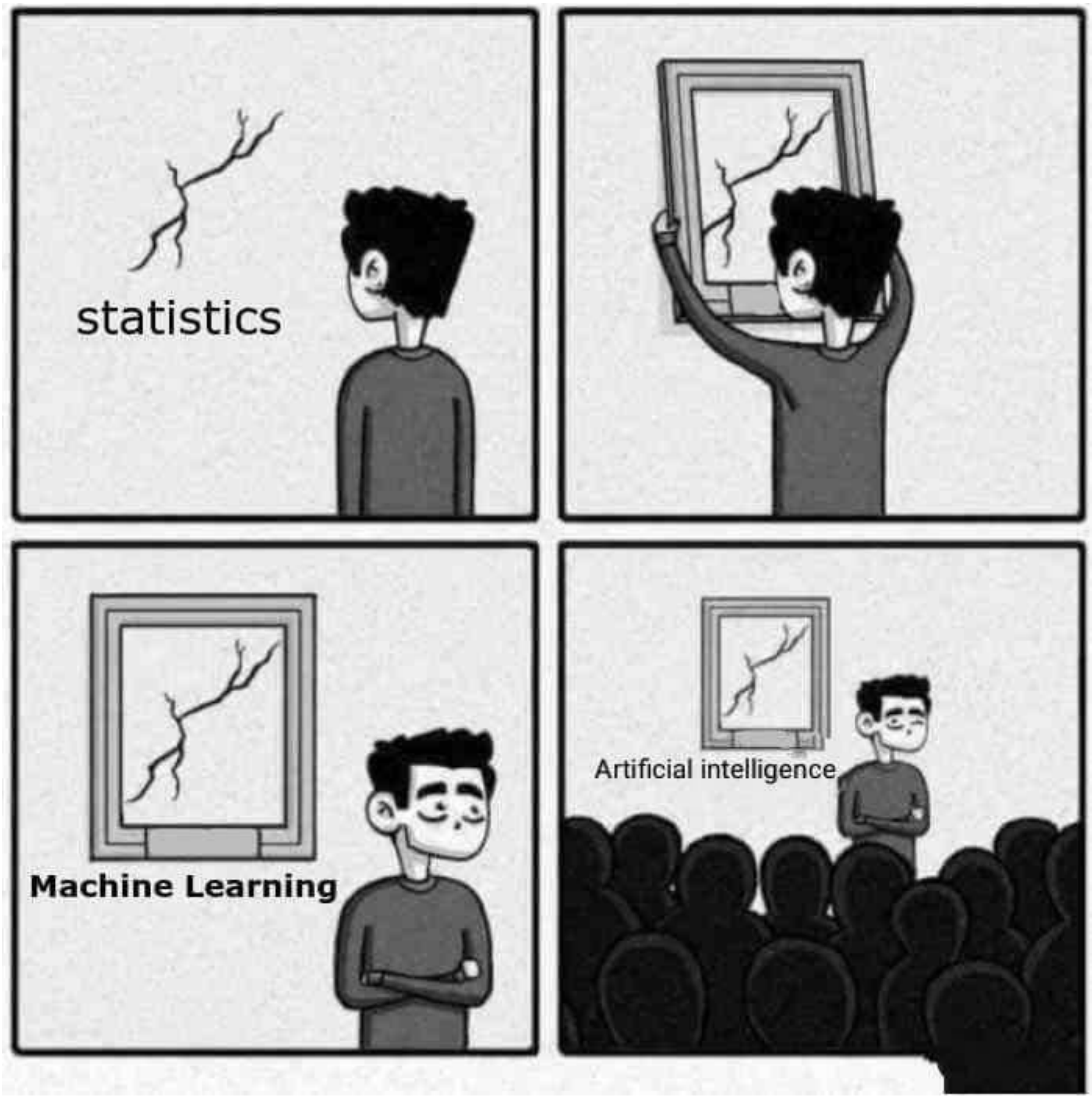
Chapter 2: Statistical Learning



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?

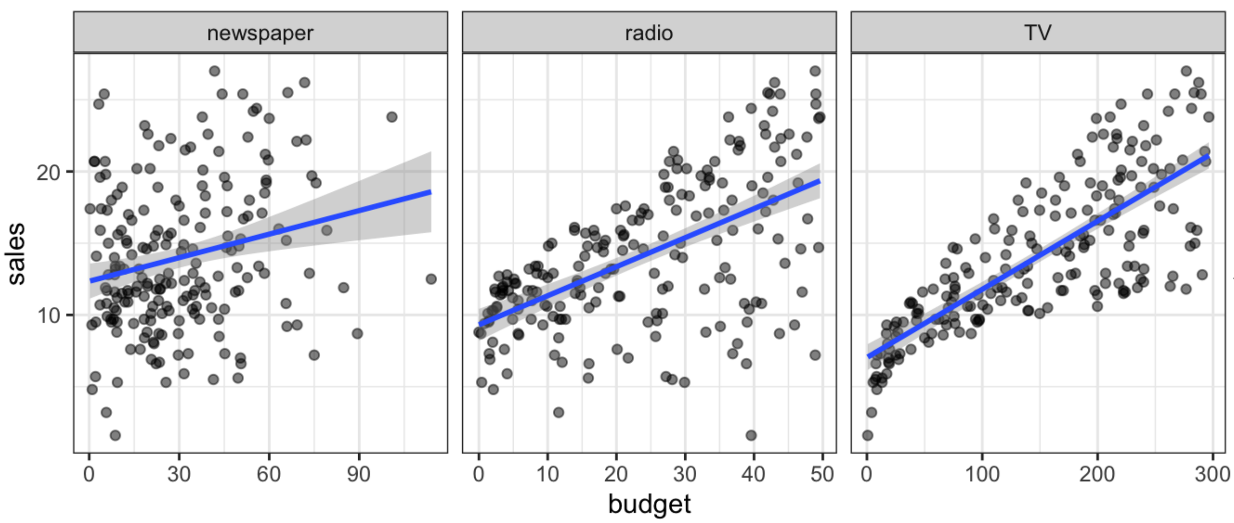
A scenario: We are consultants hired by a client to provide advice on how to improve sales of a product.

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 btw/ ads & sales we can tell our client how to advertise to increase sales.
$\Rightarrow$ develop an accurate model to predict sales based on 3 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 $X_1, \dots, X_p$

Assume there is some relationship between predictors and response,

random error, mean 0 and independent of X

fixed but unknown

$$Y = \underbrace{f}_{\text{systematic information that } X \text{ provides about } Y}(X) + e.$$

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

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

(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 of f, but we can reduce error by an using an appropriate stat learning method to estimate f.

irreducible: Even if $\hat{f}$ was estimated perfectly we would still have some error because 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 predicting Y.

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

expected value of squared difference between predicted and actual y

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

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 w/ response?

2. What is relationship btw response and each predictor?

3. Can relationship be adequately summarized by a linear equation or is it more complicated?

To return to our advertising data,

Inference:
- 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 advertising?

prediction:
- What can I expect sales to be if spend \$200k on TV and \$0 on newspaper & radio?

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

e.g. Linear models allow for interpretable inference but maybe not the most accurate prediction.

non-linear approaches can provide accurate predictions but much less interpretable.

1.2 How do we estimate f?

"training data" train

We have observed n different data points and 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. Make an assumption about shape of f.

   parameters

   e.g. $f(x) = \beta_0 + \beta_1 x_1 + \dots \beta_p x_p$.

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

   e.g. estimate $\beta_0, \beta_1, \dots, \beta_p$ with ordinary least squares (one of many options).

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

Why?

This simplifies the problem of estimating f.

<u>Disadvantage:</u>

What if shape we choose in not similar to f?

Then the estimate (and predictions) will be poor.

We can try more flexible model, this means more parameters

can lead to <u>overfitting</u>: 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.

Why?

Advantage

- fit wide range of possible shapes
- no restrictions on shape => can't assume wrong shape.

Disadvantages

- haven't simplified the problem!

=> need a lot of data.

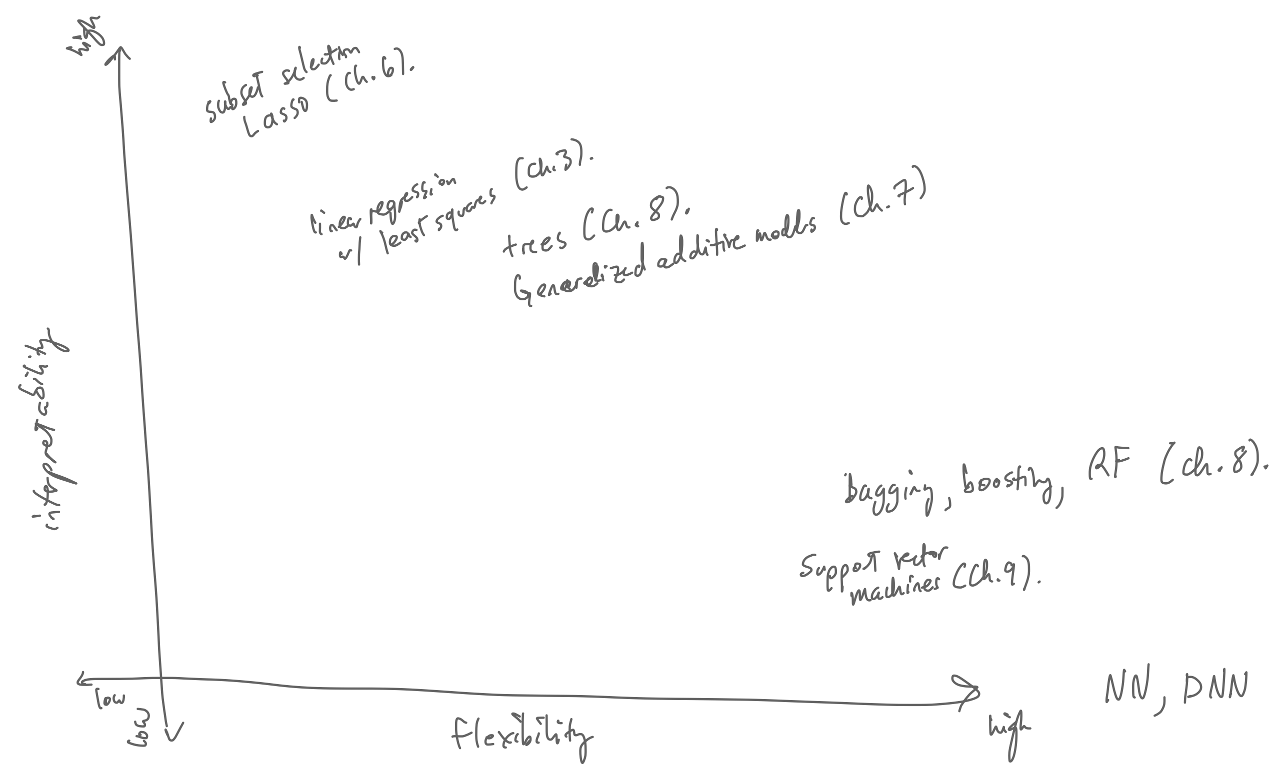
e.g. splines (ch. 7)

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. eg. linear regression vs. splines.

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

- If interested in inference, restrictive models are more interpretable.
- flexible models lead to complicated estimates of f, so difficult to understand/explain.
- too much flexibility can also lead to overfitting.



2 Supervised vs. Unsupervised Learning

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

Supervised

for each observation of predictors x_i, $i=1,\dots,n$ there is an associated response y_i

goal: fit model that relates response to predictors

↳ prediction or inference.

methods: linear regression, logistic regression, GAMs, trees, boosting, RF, SVM, etc

Unsupervised

for each observation $i=1,\dots,n$ we have a vector of measurements $\underline{x}_i$ but no response y_i.

goal: learn about structure of data.

methods: principal components, clustering.

Semi-supervised

we have some responses y_i, but not all n of them.

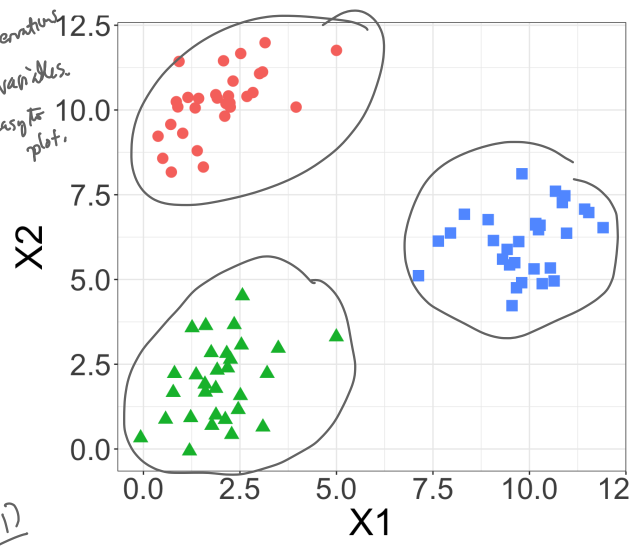
What's possible when we don't have a response variable?

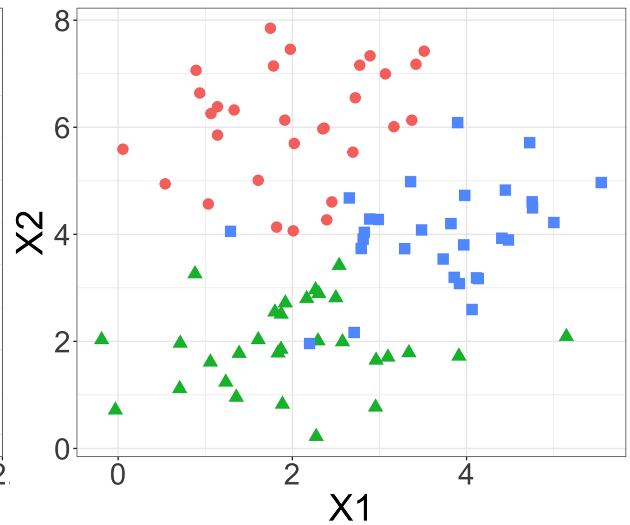
- We can seek to understand the relationships between the variables, or
- We can seek to understand the relationships between the observations.

"cluster analysis"

goal: based on $x_1, \ldots, x_n$, discern if they fall into distinct groups.

$n = 90$ observations of $p = 2$ variables.
3 groups, easy to plot.





when $p > 2$ leads to $\frac{p(p-1)}{2}$ distinct scatter plots. may need more automated approaches. (Ch. 10).

well separated groups, easy to cluster.

groups overlapping, will be 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?

Maybe it is 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 the scope of this class.

3 Regression vs. Classification

Variables can be either quantitative or categorical.

numeric values

one of K diffrent 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 either (depends).

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 can use quantitative or categorical predictors.