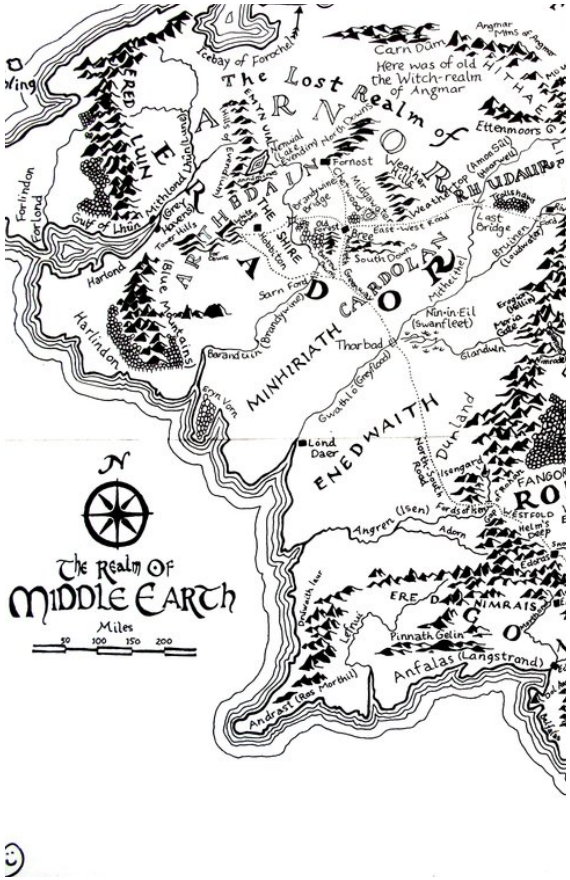
Where Are We?



- Our fundamental unit of text analysis is the document term matrix.
- We can compare documents using various distance measures and metrics.
- Now cover some more descriptive measures, dealing with diversity, complexity and style of content.
- And think more seriously about the nature of the sampling process that produces the texts we see, and what to do about it.

Complexity of Text

Lexical Diversity

- Recall that the elementary components of a text are called tokens. These are generally words, but they may also include numbers, sums of money, etc.
- The types in a document are the set of unique tokens.
- Thus we typically have many more tokens than types, because authors repeat tokens.
- We can use the type-to-token ratio (TTR) as a measure of lexical diversity. This is:

$$\boxed{TTR = \frac{\text{total types}}{\text{total tokens}}}$$

- For example, authors with limited vocabularies will have a low lexical diversity.

Uh oh. . . Tabloid vs Broadsheet

NEW YORK POST

NEWS

Iraqi troops retake key government complex in Ramadi

By Associated Press December 28, 2015 | 6:34am | Updated

Members of Iraq's elite counter-terrorism service secure a neighborhood in the city of Ramadi.

Photo: Getty Images

MORE ON: ISIS

BAGHDAD — Iraqi military forces on Monday retook a strategic government complex in the city of

$$TTR = \frac{250}{491} = 0.51$$

The New York Times

Obama's 'Boots on the Ground': U.S. Special Forces Are Sent to Tackle Global Threats

Japan and South Korea Settle Dispute Over Wartime 'Comfort Women'

ITINERARIES T.S.A. Moves Closer to Rejecting Some State Driver's Licenses for...

MIDDLE EAST

Iraqi Forces Retake Center of Ramadi From ISIS

By FALIH HASSAN and SEWELL CHAN DEC. 28, 2015

Iraqi soldiers at the Anbar police headquarters in Ramadi, Iraq, on Monday, after seizing a government complex from the Islamic State. Ahmad Al-Rubaye/Agence France-Presse — Getty Images

Email

Share

BAGHDAD — Iraqi forces said on Monday they had seized a strategic government complex in the western city of Ramadi from the Islamic State after a fierce weeklong battle, putting them on the verge of a crucial victory following a brutal seven-month occupation of the city by the extremist group

$$TTR = \frac{428}{978} = 0.43$$

Hmm...

- Unexpected, and mostly product of different text lengths: shorter texts tend to have fewer repetitions (of e.g. common words).
- So make denominator non-linear:
- 1954 Guiraud index of lexical richness : $\boxed{R = \frac{\text{total types}}{\sqrt{\text{total tokens}}}}$
- So NY Post: $\frac{250}{\sqrt{491}} = 11.28$; NYT: $\frac{428}{\sqrt{978}} = 13.68$.

$\rightarrow$ has been augmented—Advanced Guiraud—to exclude very common words.

Measurement of Linguistic Complexity

- over a hundred years of literature on measurement of 'readability'—general issue was assigning school texts to pupils of different ages and abilities.

- Flesch (1948) suggests *Flesch Reading Ease* statistic

$$FRE = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

based on $\hat{\beta}$s from linear model where y = average grade level of school children who could correctly answer at least 75% of mc qs on texts. Scaled s.t. a document with score of 100 could be understood by fourth grader (9yo).

- Kincaid et al later translate to US School grade level that would be (on average) required to comprehend text.

Readability Guidelines

- In practice, estimated FRE can be outside $[0, 100]$.

- However, in general...

Score	School level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

Examples

Score	Text
-800	Molly Bloom's (3.6K) Soliloquy, *Ulysses*
42	mean economics article; judicial opinion
45	life insurance requirement (FL)
48	*New York Times*
65	*Reader's Digest*
67	*Al Qaeda* press release
77	Dickens' works
80	children's books: e.g. *Wind in the Willows*
90	death row inmate last statements (TX)
100	this entry right here.

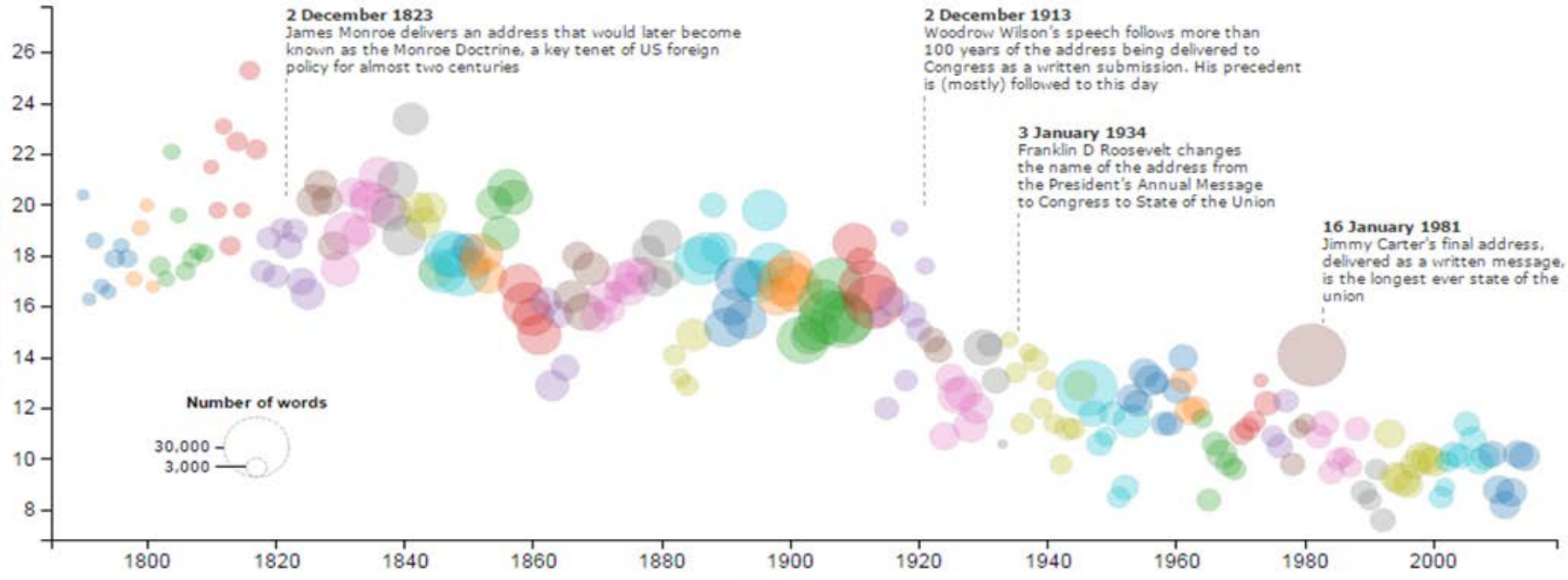
Notes

- Flesch scoring only uses syllable information: no input from rarity or unfamiliarity of word.
- E.g. "`Indeed, the shoemaker was frightened`" would score similarly to "`Forsooth, the cordwainer was afeared`"
- Widely used because it 'works', not because it is justified from first principles
- One of many such indices: Gunning-Fog, Dale-Chall, Automated Readability Index, SMOG. Typically highly correlated.
- Surprisingly little effort to describe statistical behavior of estimator: sampling distribution etc.

The state of our union is ... dumber:

How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union



Dale-Chall, 1948

- yields grade level of text sample.

$$DC = 0.1579 \times (\text{PDW}) + 0.0496 \times \left(\frac{\text{total words}}{\text{total sentences}} \right)$$

- Where PDW is percentage of difficult words,
- And a 'difficult' word is one that does not appear on Dale & Chall's list of 763 (later updated to 3000) 'familiar' words.
- E.g. `about`, `back`, `call`, etc.

Some Substantive Concerns about Readability Measures

- Designed for education
- Tested/validated on children
- Designed for readers in 1940/50s
- No uncertainty around estimates since sampling distribution not known

Style and Stylometrics

Mystery of *The Federalist Papers*

- 85 essays published anonymously in 1787 and 1788
- Generally agreed that Alexander Hamilton wrote 51 essays, John Jay wrote 5 essays, James Madison wrote 14 essays, and 3 essays were written jointly by Hamilton and Madison.
- That leaves 12 that are disputed.

Mosteller and Wallace, 1963/4

In essence, they...

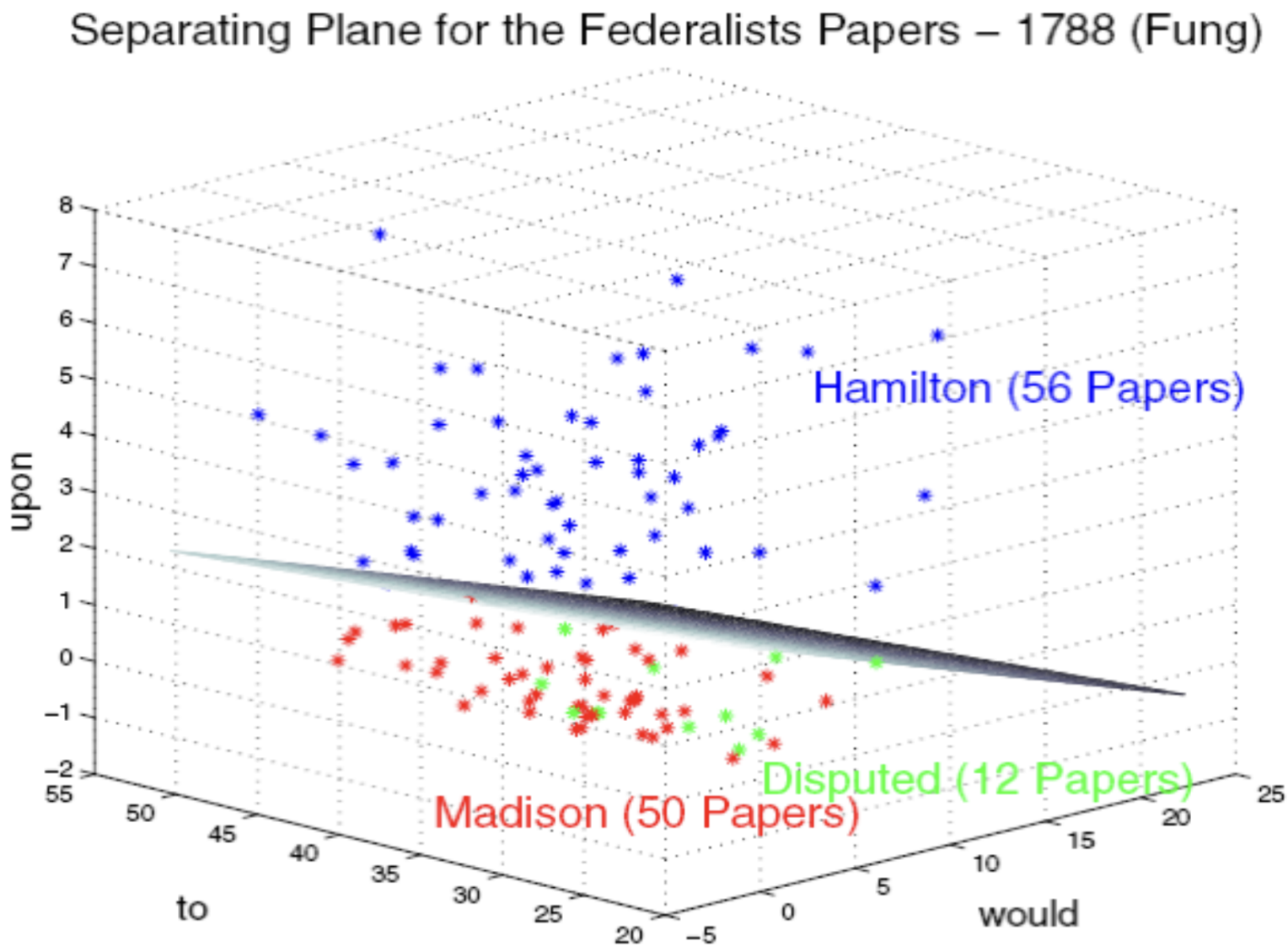
- Count word frequencies of function words (`by`, `from`, `to`, etc.) in the 73 essays with undisputed authorship
- Then collapse on author to get word frequencies specific to the authors
- They ask "if rates of function word usage are constant within authors for these documents, which author was most likely to have written essay x given the observed function word usage of these authors on the other documents?"

More Details

a	been	had	its
one	the	were	all
but	has	may	only
their	what	also	by
have	more	or	then
when	an	can	her
must	our	there	which
and	do	his	my
things	who	any	down
if	no	so	this
are	even	in	not
some	to	with	as
every	into	now	such
up	would	at	for
is	of	than	upon
your	be	from	it
on	that	was	will
should			

- May think that sentence length distinguishes authors, but Hamilton and Madison "practically twins" on this.
- Use function words—conjunctions, prepositions, pronouns— for two (related) reasons:
  1. authors use them unconsciously
  2. therefore, don't vary much by topic.
- NB: typically assume one instance of a function word is independent of the next, and use is fixed over a lifetime (and constant within a given text).

→ wrong, but models relying on these assumptions discriminate well (see Peng & Hengartner on e.g. Austen v Shakespeare)

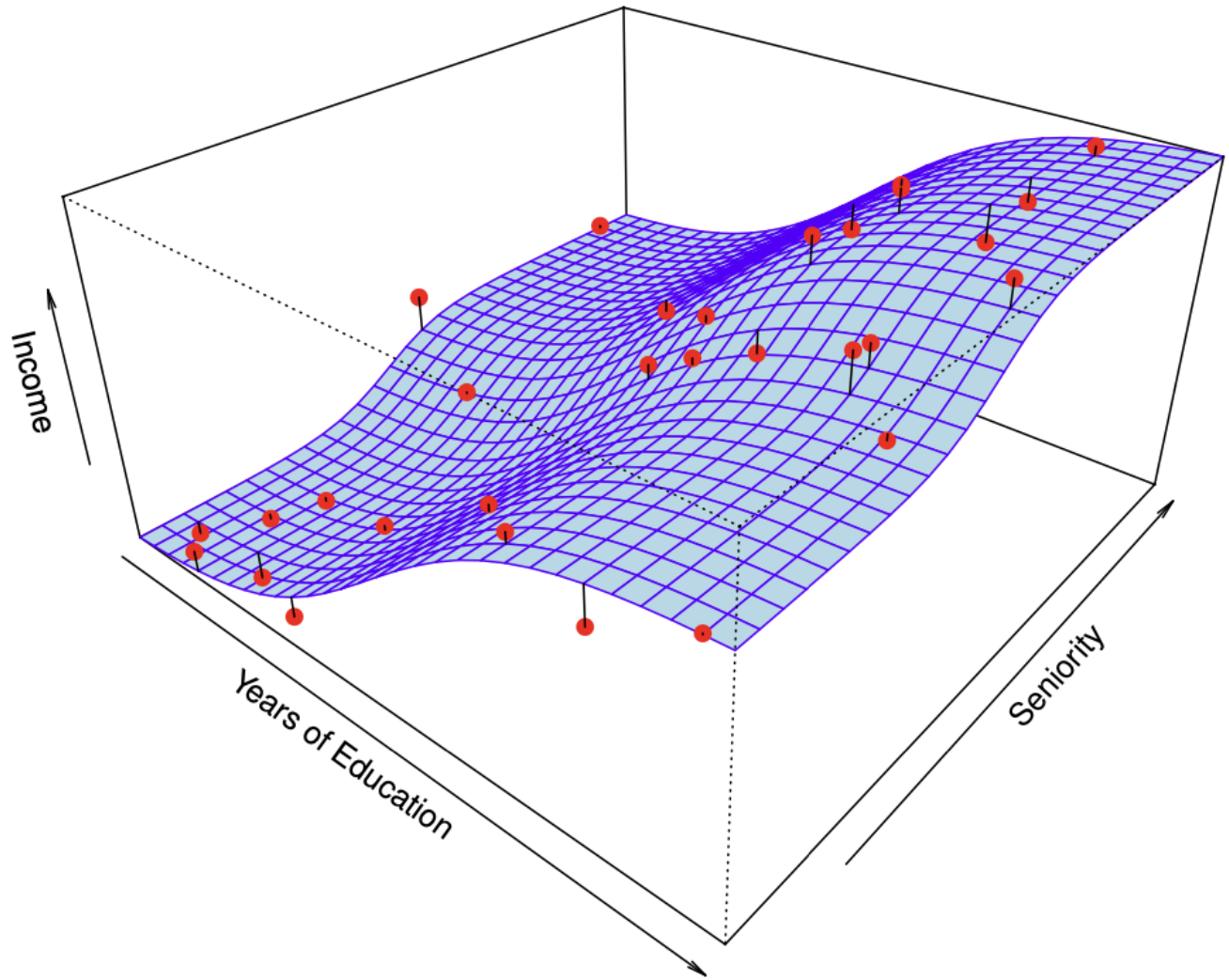

Separating Plane for the Federalists Papers – 1788 (Fung)
upon
to
would
Hamilton (56 Papers)
Madison (50 Papers)
Disputed (12 Papers)
8
7
6
5
4
3
2
1
0
−1
−2
55
50
45
40
35
30
25
20
−5
0
5
10
15
20
25

Statistical Learning, Cross-Validation, and Sampling

We start with some underlying "true" model of reality that we want to discover something about using a sample of observations drawn from that true model.

$$Y = f(X) + \epsilon.$$

An example of the "true" model and our sample of observations

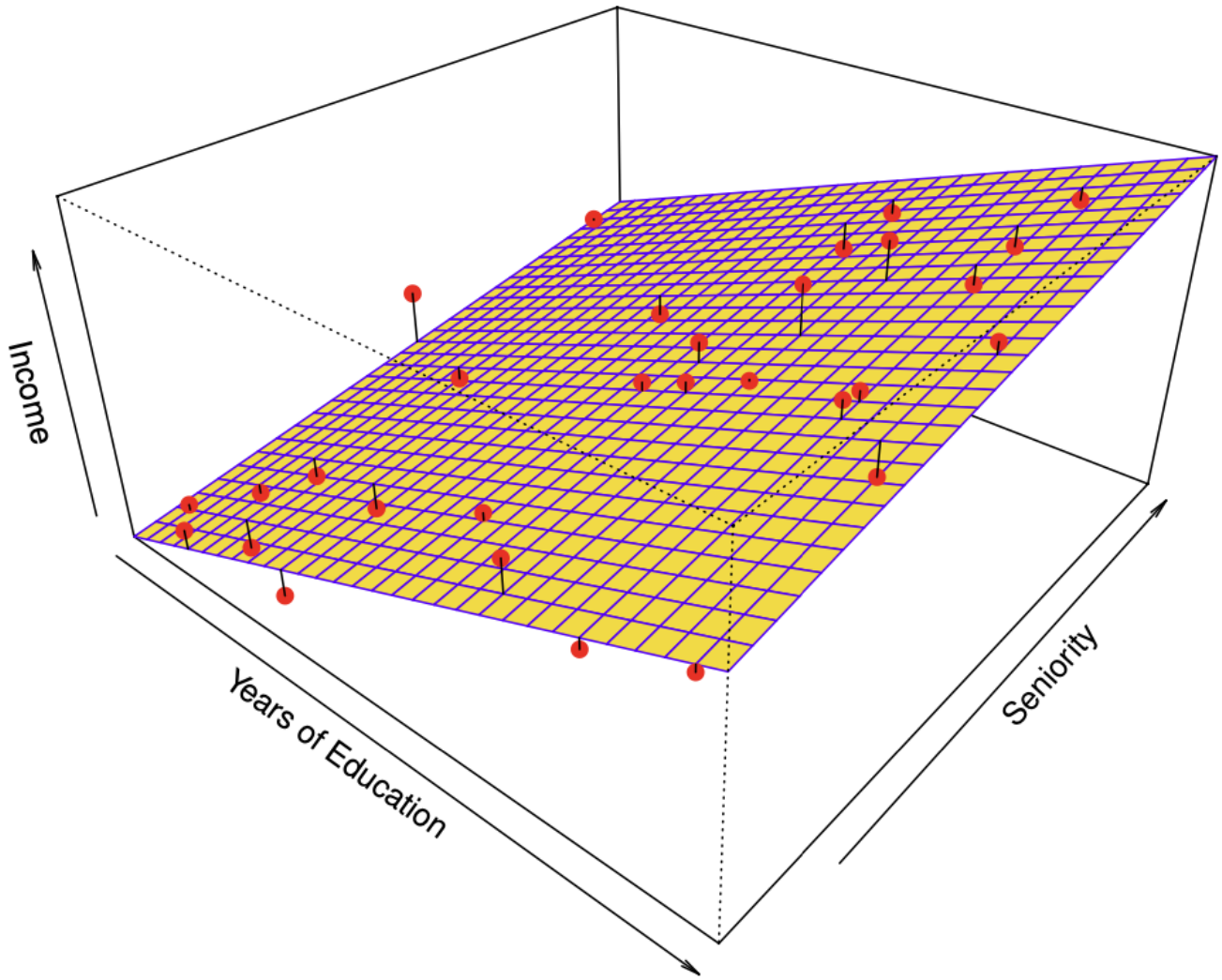


We want an estimate of the "true" model based on our sample

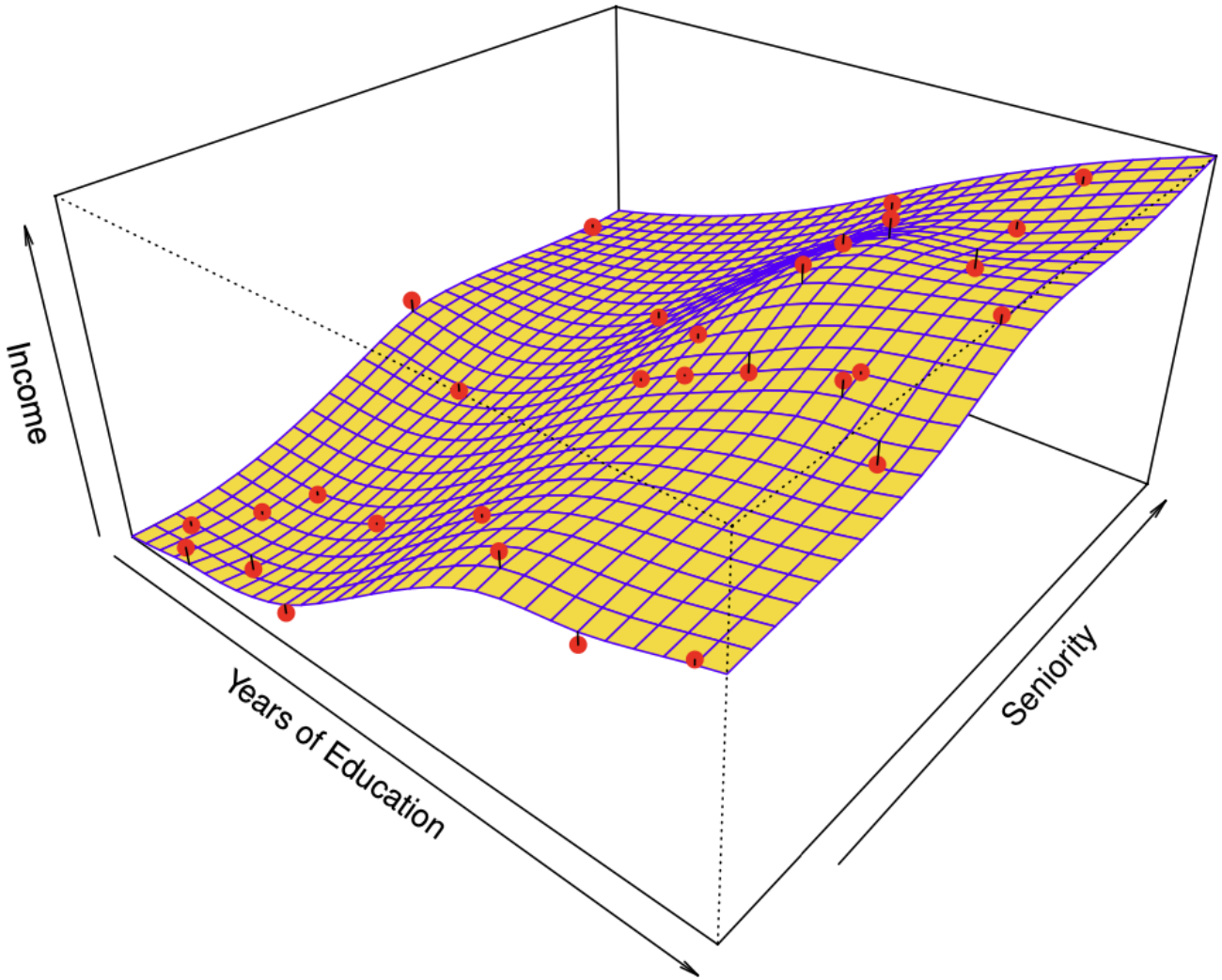
$$\hat{Y} = \hat{f}(X),$$

$\rightarrow$ But remember, we want our estimated model to make good out of sample predictions!

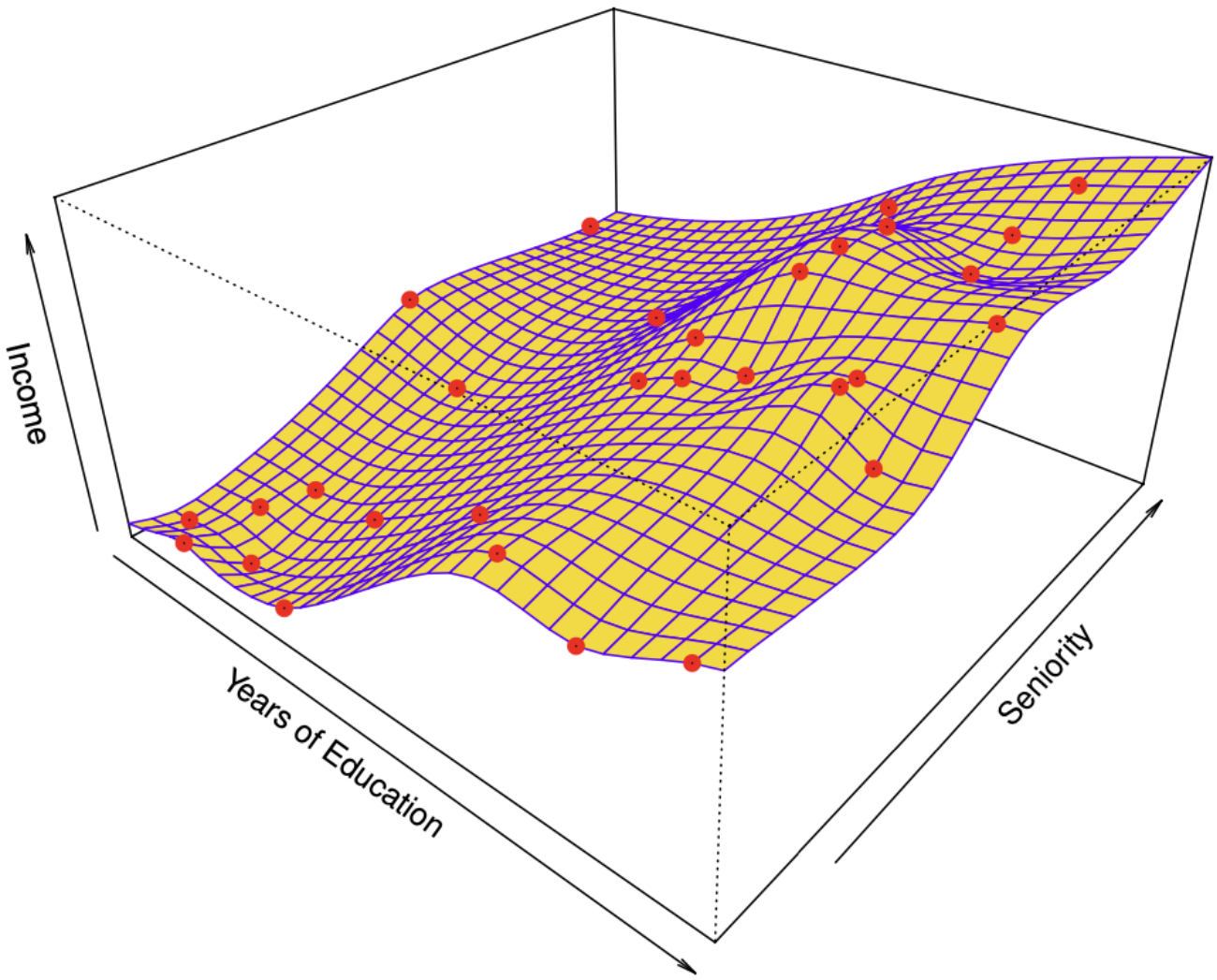
An example of a simple parametric estimate of $f()$: OLS



A more flexible model...



A very flexible model: AKA "overfitting the data"

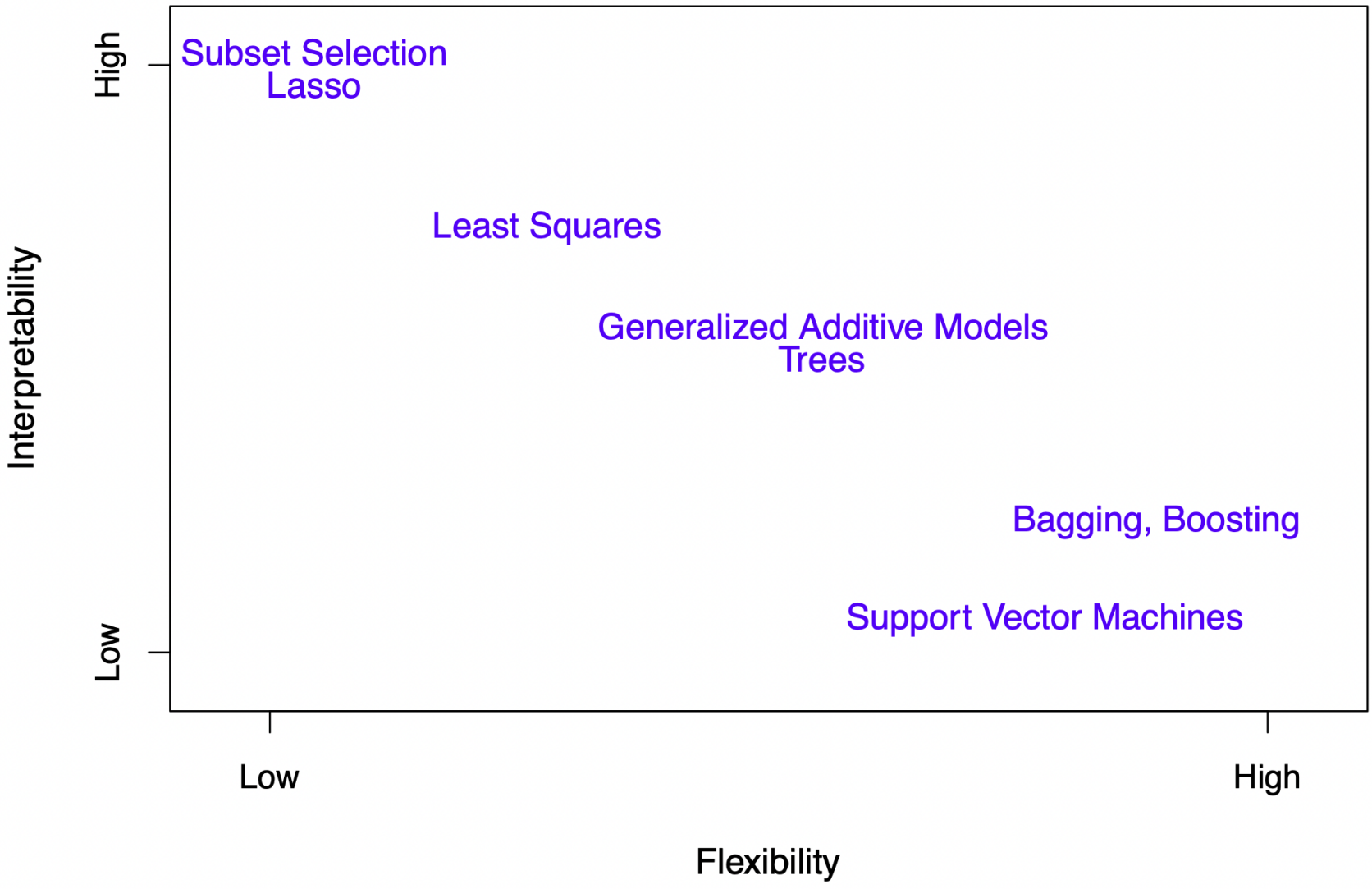


We can never eliminate all variance from our model...

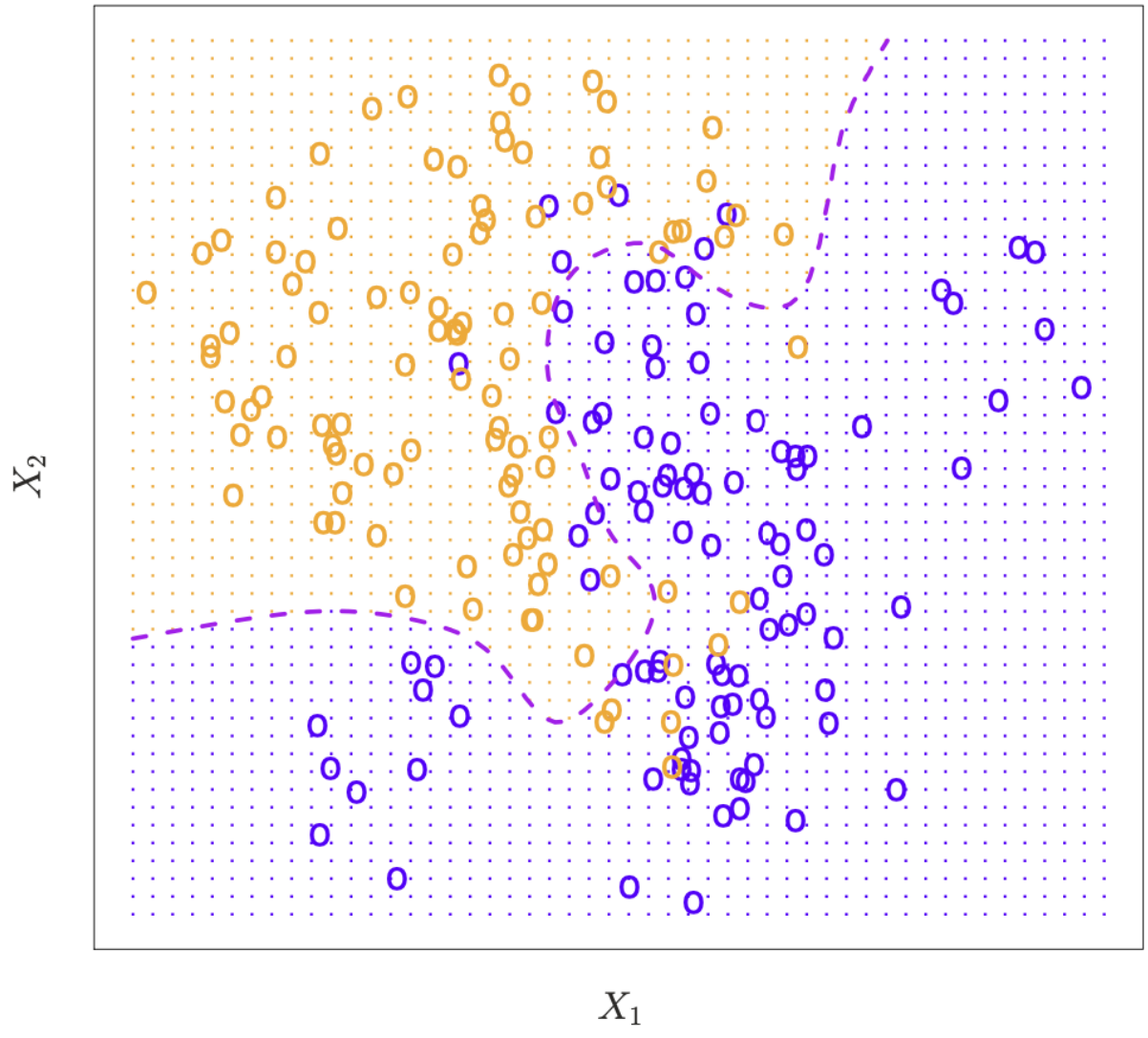
$$\begin{aligned} \mathrm{E}(Y - \hat{Y})^2 &= \mathrm{E}[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Some trade-offs

- Prediction accuracy versus interpretability.
  — Linear models are easy to interpret; thin-plate splines are not.
- Good fit versus over-fit or under-fit.
  — How do we know when the fit is just right?
- Parsimony versus black-box.
  — We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.


Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
High
Low
Interpretability
Low
High
Flexibility

The best we can do is the Bayes Estimator

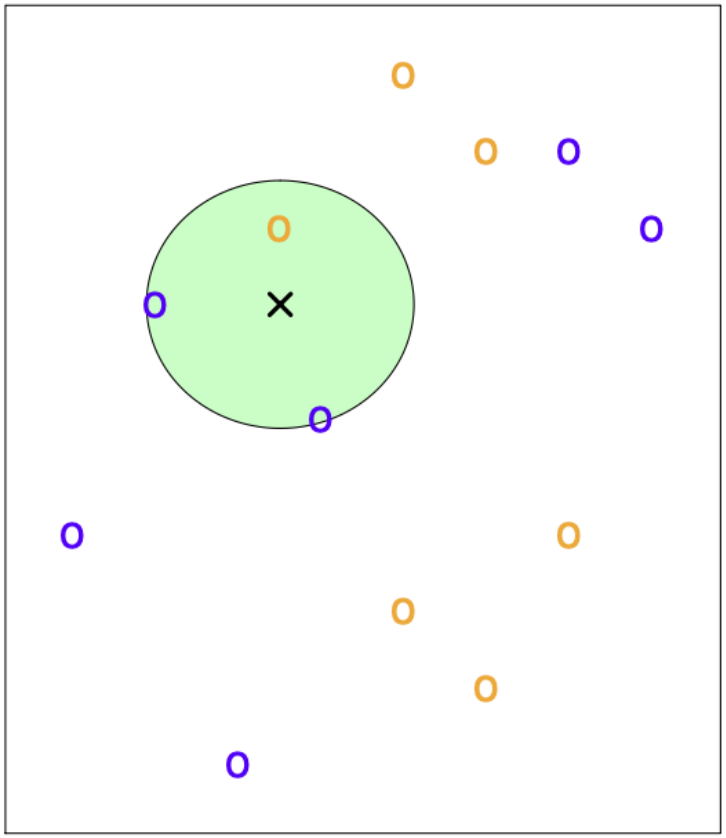


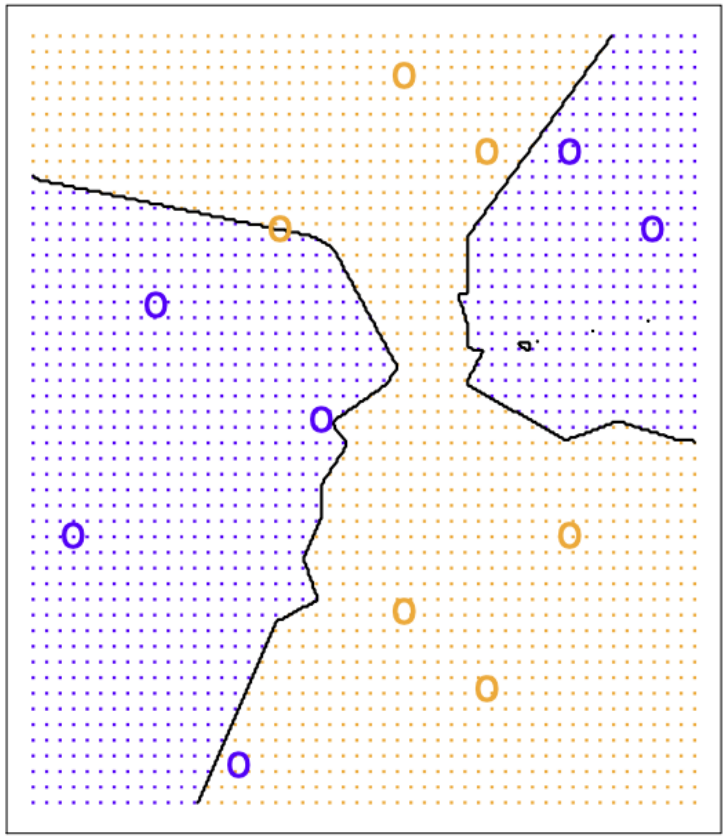
The Bayes Estimator for a Classification Problem

$$\Pr(Y = j | X = x_0)$$

$\rightarrow$ But we don't know the "true" underlying conditional distribution of Y given X.

One Solution: K Nearest Neighbors





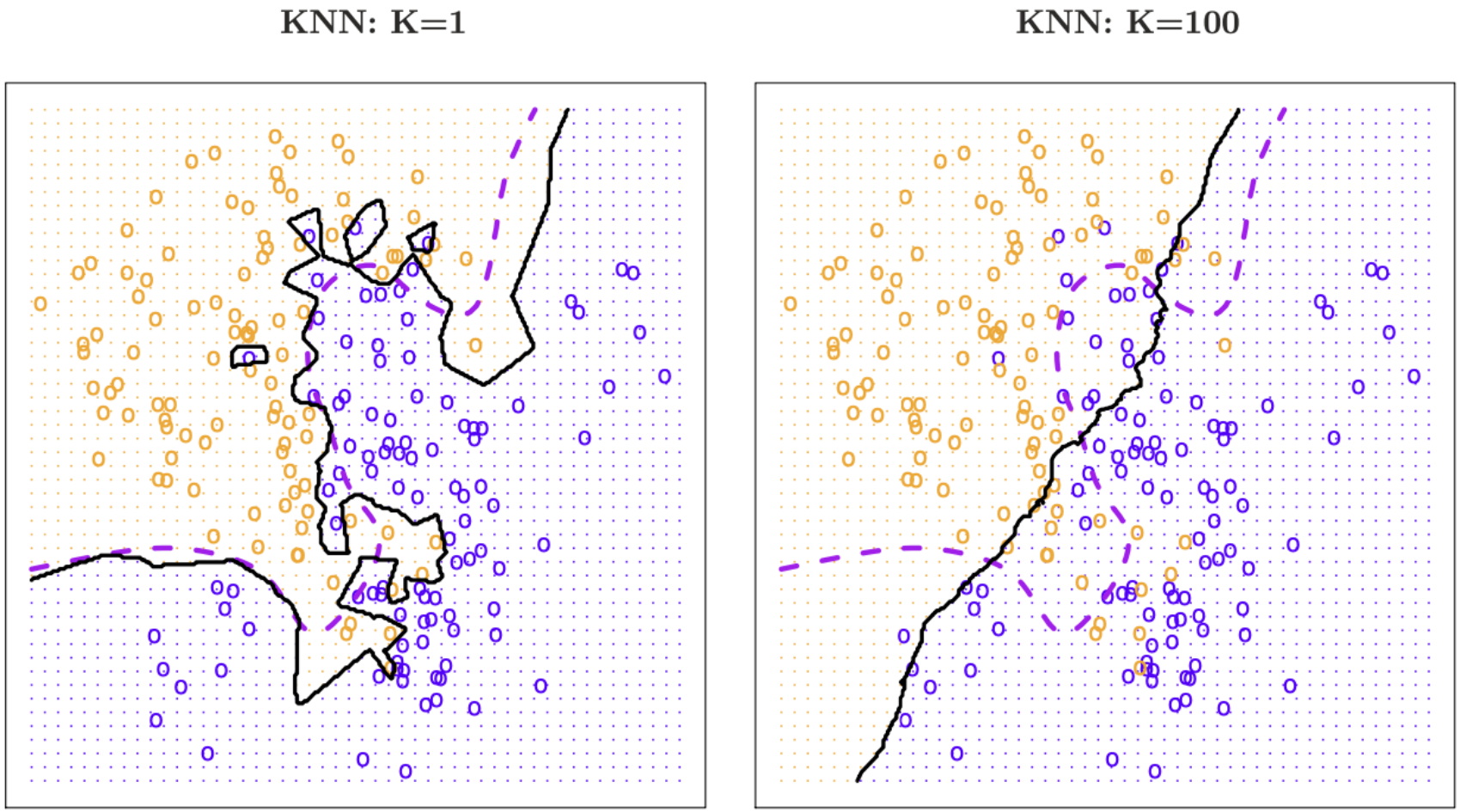
The details. . .

given observation to the class with highest *estimated* probability. One such method is the *K-nearest neighbors* (KNN) classifier. Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

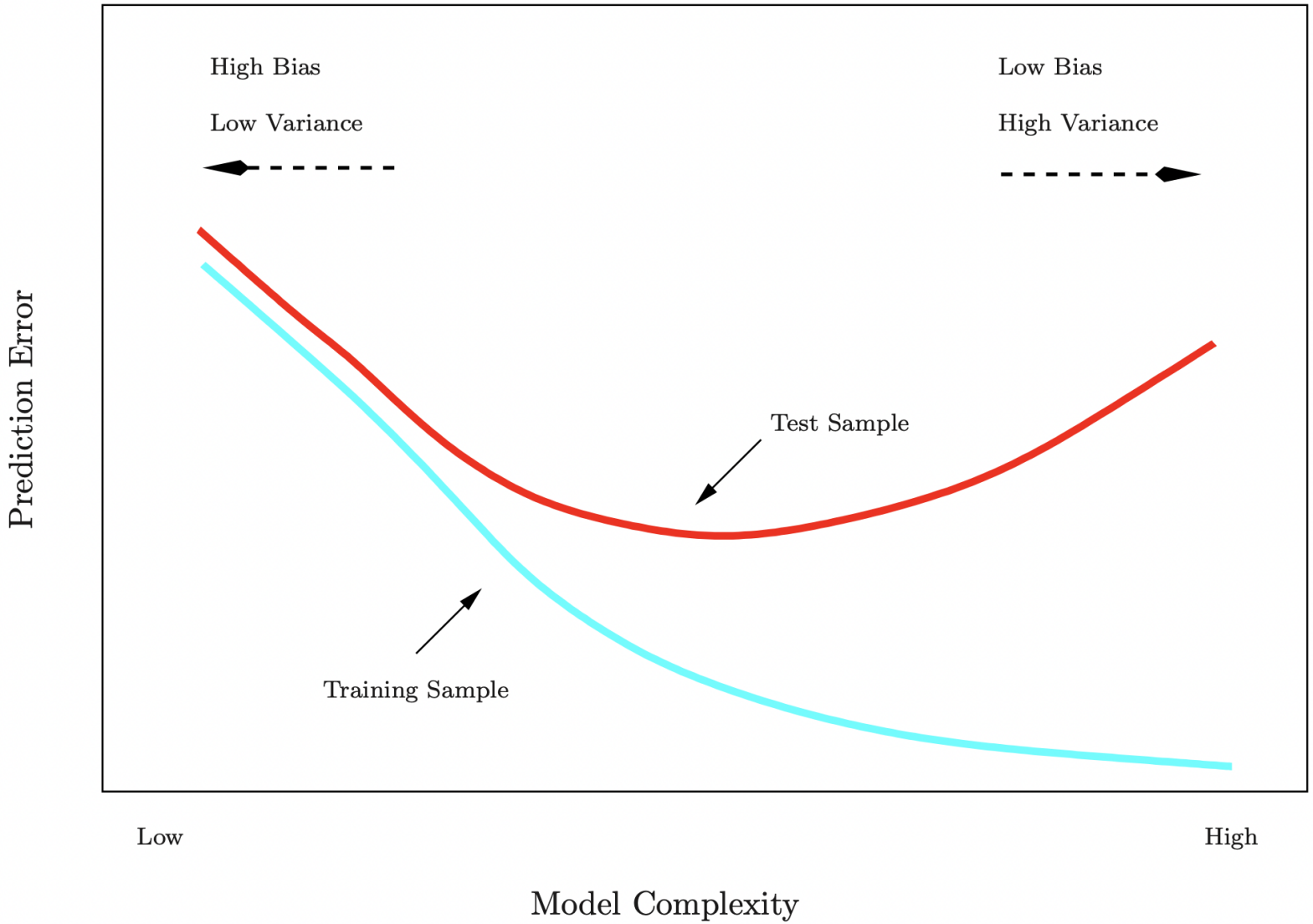
$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j). \qquad (2.12)$$

Finally, KNN classifies the test observation x_0 to the class with the largest probability from (2.12).

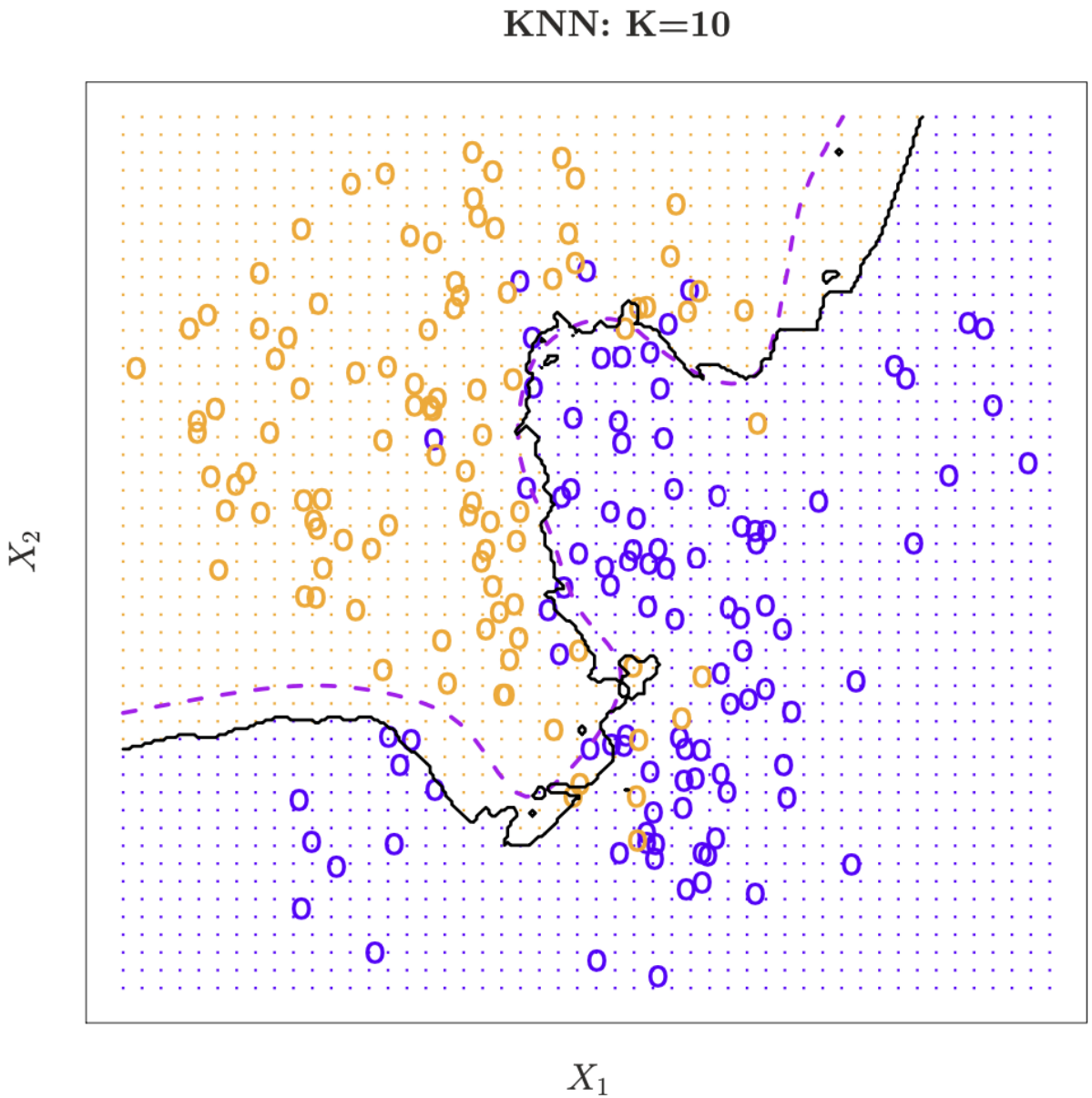
The choice of K is related to how well we can minimize the prediction error in the training vs. the test data. . .



Training- versus Test-Set Performance



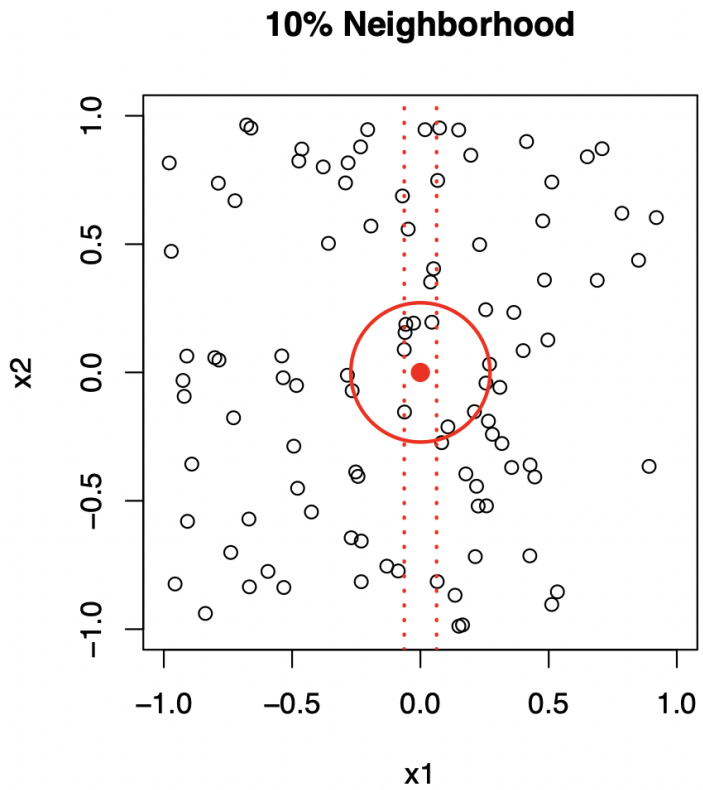
This is a nice compromise. . . But how do we figure out the right N systematically?

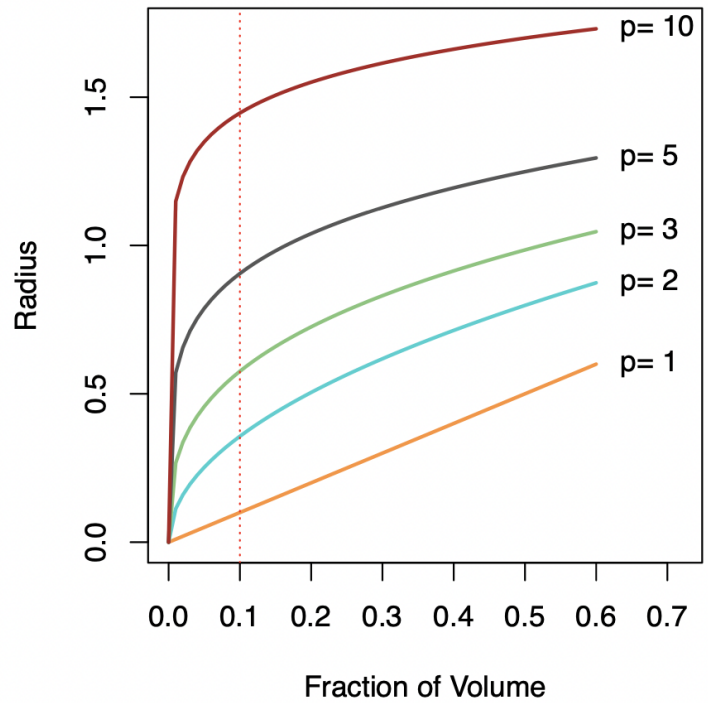


The Curse of Dimensionality

- Nearest neighbor averaging can be pretty good for small p — i.e. $p \leq 4$ and large-ish N.
- We will discuss smoother versions, such as kernel and spline smoothing later in the course.
- Nearest neighbor methods can be *lousy* when p is large. Reason: the *curse of dimensionality*. Nearest neighbors tend to be far away in high dimensions.
    - We need to get a reasonable fraction of the N values of y_i to average to bring the variance down—e.g. 10%.
    - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.

The curse of dimensionality





Sampling and Uncertainty

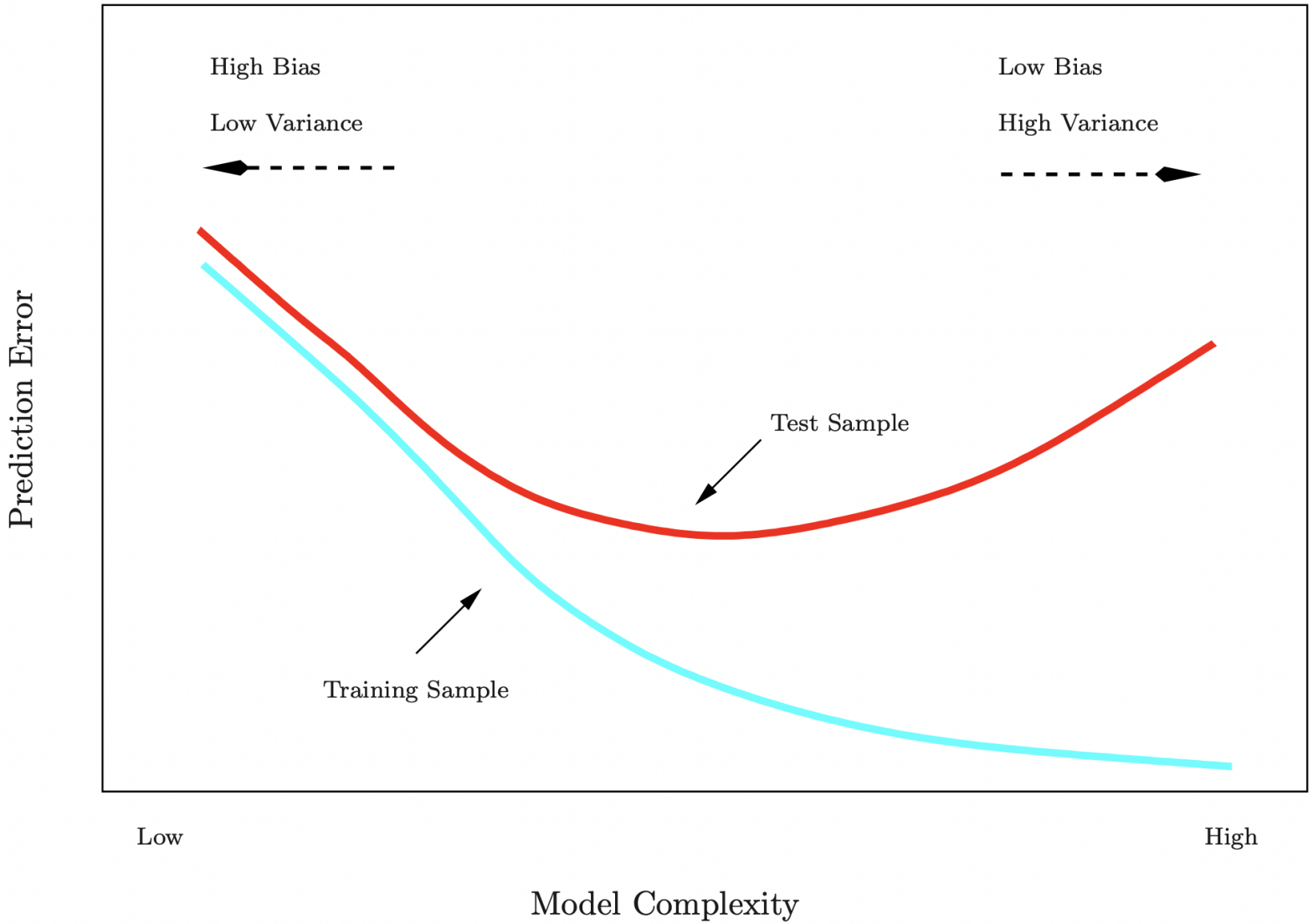
Cross-validation and the Bootstrap

- In the section we discuss two *resampling* methods: cross-validation and the bootstrap.
- These methods refit a model of interest to samples formed from the training set, in order to obtain additional information about the fitted model.
- For example, they provide estimates of test-set prediction error, and the standard deviation and bias of our parameter estimates

Training Error versus Test error

- Recall the distinction between the *test error* and the *training error:*
- The *test error* is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the *training error* can be easily calculated by applying the statistical learning method to the observations used in its training.
- But the training error rate often is quite different from the test error rate, and in particular the former can *dramatically underestimate* the latter.

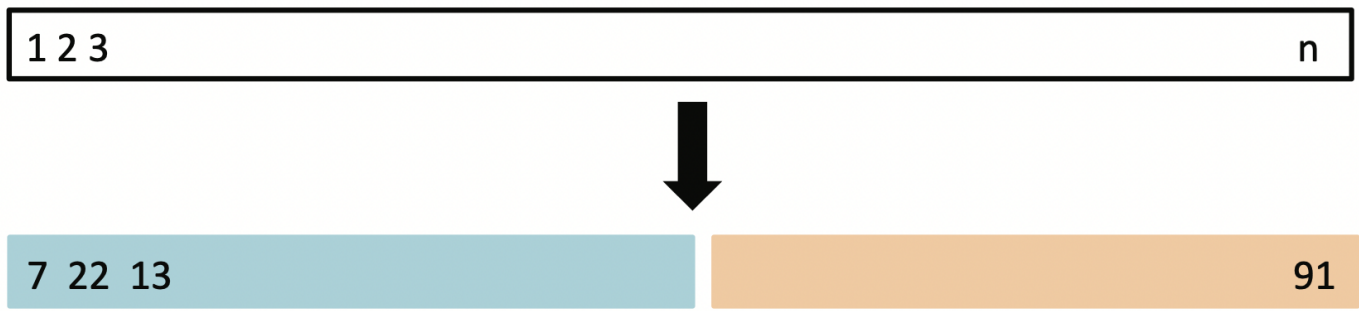
Training- versus Test-Set Performance



Validation-set approach

- Here we randomly divide the available set of samples into two parts: a *training set* and a *validation* or *hold-out set*.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an estimate of the test error. This is typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative (discrete) response.

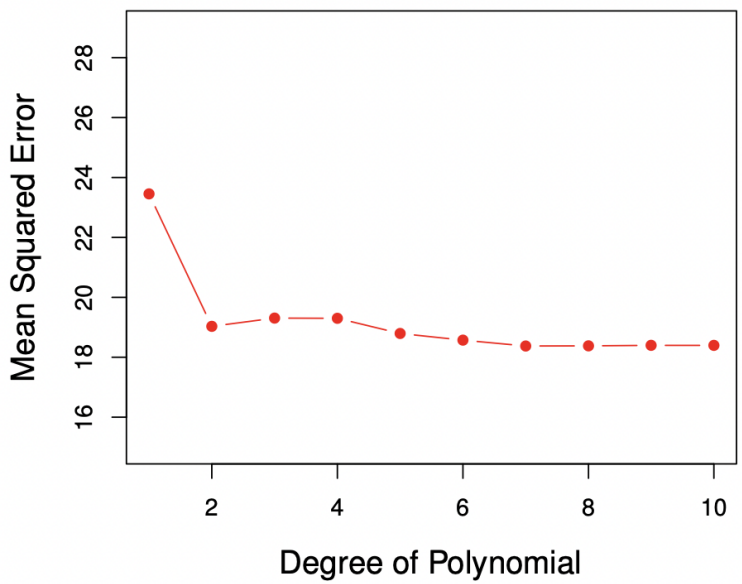
The Validation process

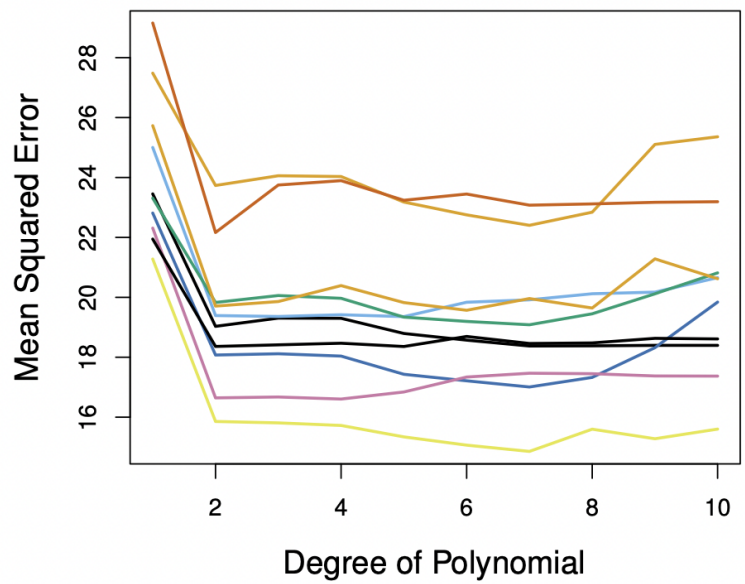


A random splitting into two halves: left part is training set, right part is validation set

Example: automobile data

- Want to compare linear vs higher-order polynomial terms in a linear regression
- We randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations.





Drawbacks of validation set approach

- the validation estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.
- In the validation approach, only a subset of the observations — those that are included in the training set rather than in the validation set — are used to fit the model.
- This suggests that the validation set error may tend to *overestimate* the test error for the model fit on the entire data set.

K-fold Cross-validation

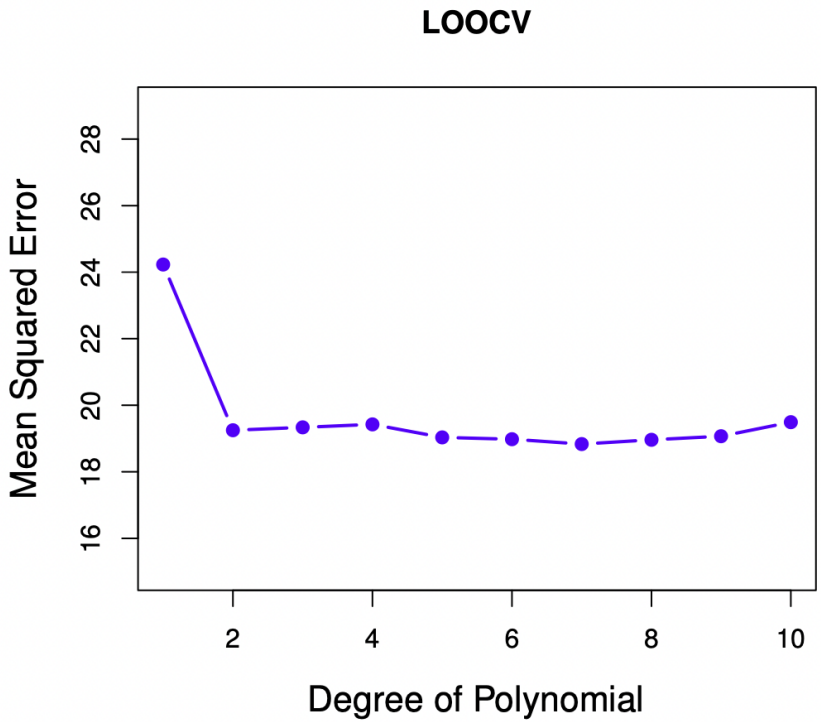
- *Widely used approach* for estimating test error.
- Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts. We leave out part k, fit the model to the other $K-1$ parts (combined), and then obtain predictions for the left-out kth part.
- This is done in turn for each part $k = 1, 2, \ldots K$, and then the results are combined.

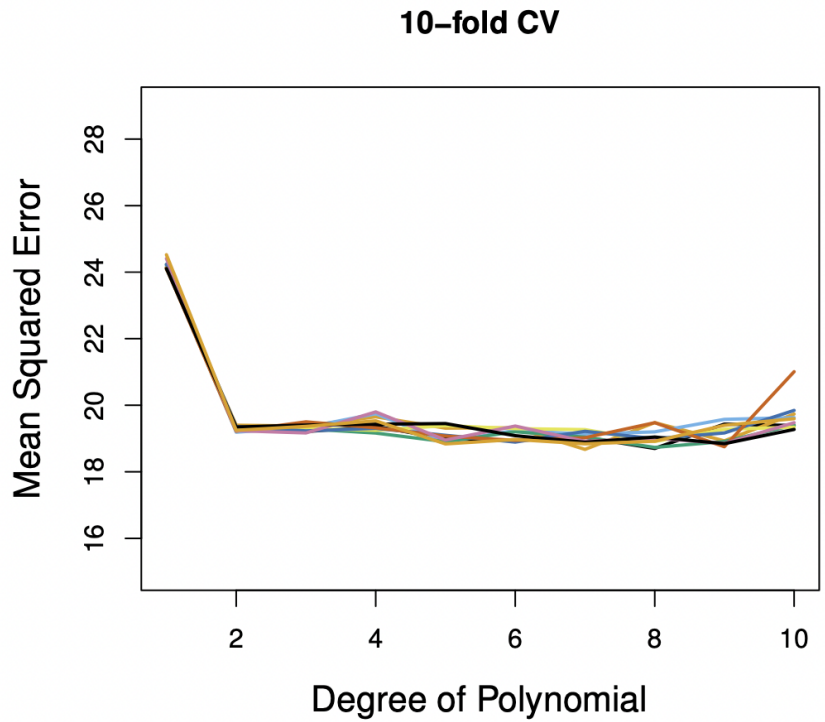
K-fold Cross-validation in detail

Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Validation	Train	Train	Train	Train

Auto data revisited





Other issues with Cross-validation

- Since each training set is only $(K-1)/K$ as big as the original training set, the estimates of prediction error will typically be biased upward. *Why?*
- This bias is minimized when $K = n$ (LOOCV), but this estimate has high variance, as noted earlier.
- $K = 5$ or 10 provides a good compromise for this bias-variance tradeoff.

Sampling and Uncertainty

To now, we've been concerned with point estimates

e.g. the lexical diversity of a story is 0.43, the FRE of a speech is 55.2 etc

But this is unsatisfying: it says nothing of the variance of such estimates, yet surely we are more certain of an estimate from a long text (or many texts) than we are from a short text.

e.g. how much would you trust a one word response vs a 1000-word speech from a member of parliament in terms of its complexity or policy content?

$\rightarrow$ think a little more systematically about the sampling distribution of a statistic.

Sampling Distributions: Reminder

Suppose we are interested in the population mean, μ and we our we use the sample mean $\bar{x}$ as our estimator of it.

We want the sampling distribution of sample mean, i.e. the distribution of the sample mean for all possible samples we *could have* drawn.

$\rightarrow$ important, because it tells us the uncertainty around our statistic, and we can use that to produce e.g. confidence intervals and make statements about the statistical significance of differences between means of different groups.

The Bootstrap

- The *bootstrap* is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

Where does the name came from?

- The use of the term bootstrap derives from the phrase *to pull oneself up by one's bootstraps*, widely thought to be based on one of the eighteenth century "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:

  The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

- It is not the same as the term "bootstrap" used in computer science meaning to "boot" a computer from a set of core instructions, though the derivation is similar.

A simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y, respectively, where X and Y are random quantities.
- We will invest a fraction α of our money in X, and will invest the remaining $1 - \alpha$ in Y.
- We wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $\text{Var}(\alpha X + (1 - \alpha)Y)$.
- One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

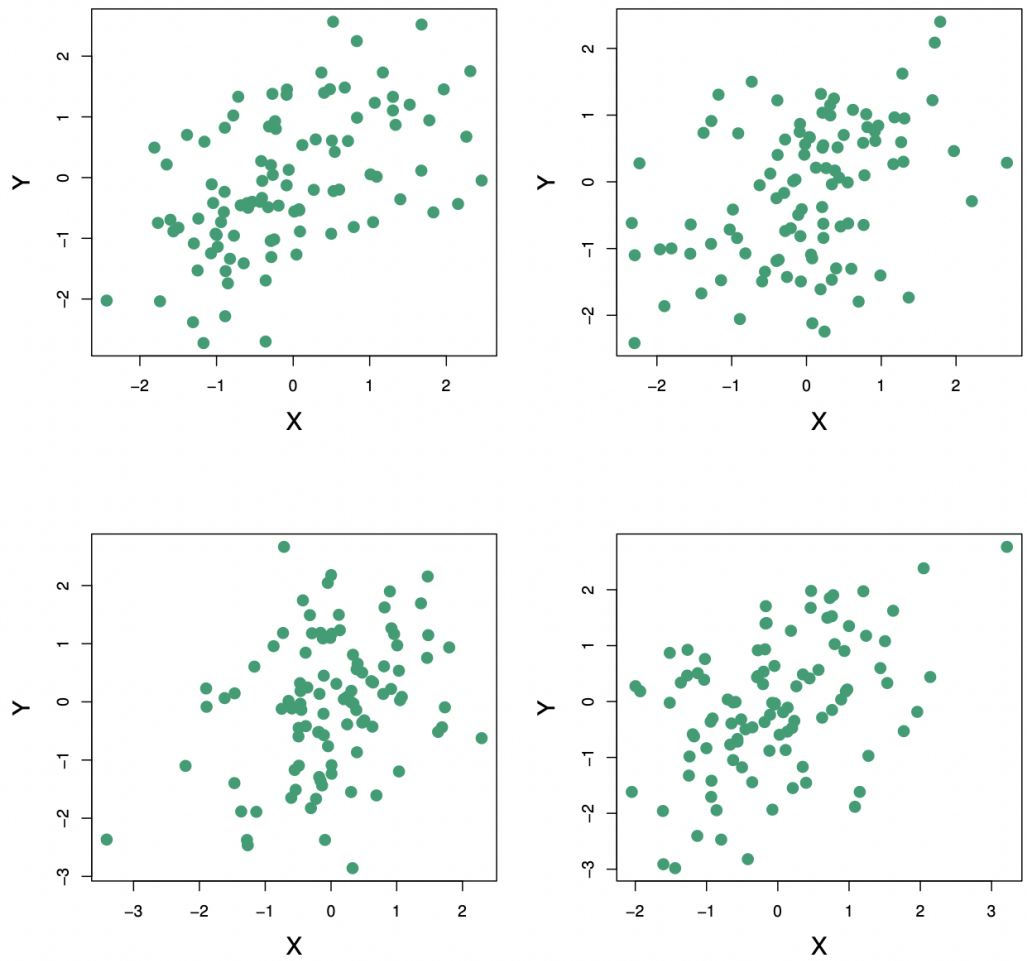
where $\sigma_X^2 = \text{Var}(X), \sigma_Y^2 = \text{Var}(Y)$, and $\sigma_{XY} = \text{Cov}(X, Y)$.

Example continued

- But the values of σ_X^2, σ_Y^2, and σ_{XY} are unknown.
- We can compute estimates for these quantities, $\hat{\sigma}_X^2$, $\hat{\sigma}_Y^2$, and $\hat{\sigma}_{XY}$, using a data set that contains measurements for X and Y.
- We can then estimate the value of α that minimizes the variance of our investment using

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}.$$

Example continued



Each panel displays 100 *simulated returns for investments* X *and* Y. *From left to right and top to bottom, the resulting estimates for* α *are* 0.576, 0.532, 0.657, *and* 0.651.

Example continued

- To estimate the standard deviation of $\hat{\alpha}$, we repeated the process of simulating 100 paired observations of X and Y, and estimating α 1,000 times.
- We thereby obtained 1,000 estimates for α, which we can call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$.
- The left-hand panel of the Figure on slide 29 displays a histogram of the resulting estimates.
- For these simulations the parameters were set to $\sigma_X^2 = 1, \sigma_Y^2 = 1.25$, and $\sigma_{XY} = 0.5$, and so we know that the true value of α is 0.6 (indicated by the red line).

Example continued

- The mean over all 1,000 estimates for α is

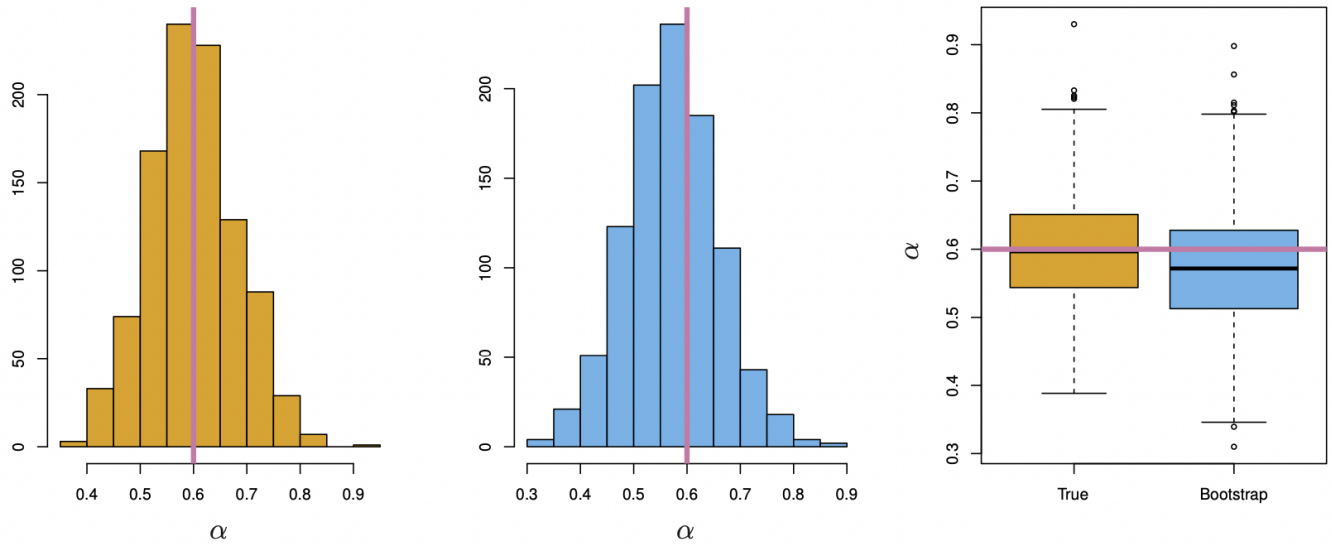
$$\bar{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996,$$

very close to $\alpha = 0.6$, and the standard deviation of the estimates is

$$\sqrt{\frac{1}{1000-1} \sum_{r=1}^{1000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083.$$

- This gives us a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) \approx 0.083$.
- So roughly speaking, for a random sample from the population, we would expect $\hat{\alpha}$ to differ from α by approximately 0.08, on average.

Results

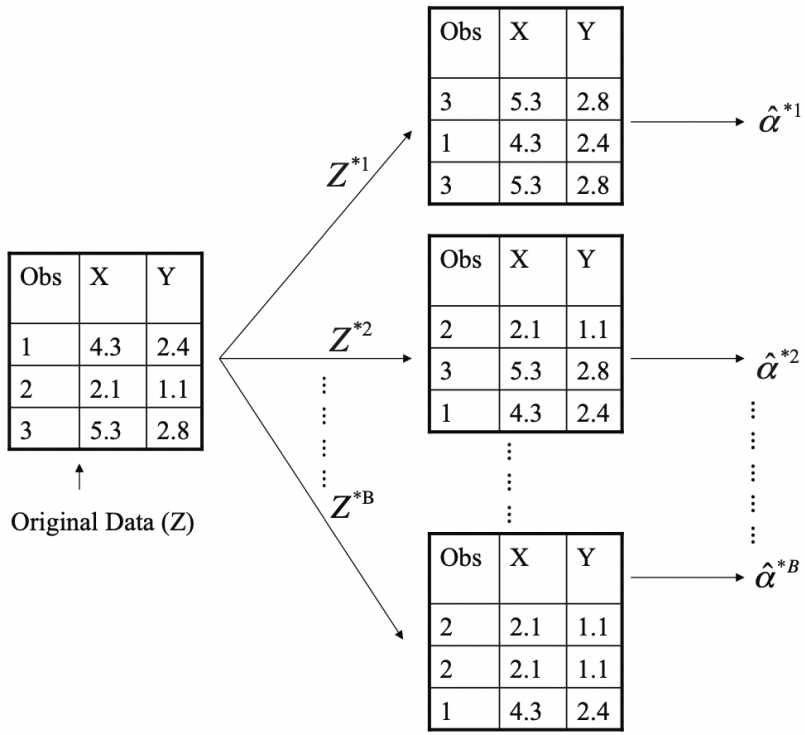


Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. *Center:* A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set. *Right:* The estimates of α displayed in the left and center panels are shown as boxplots. In each panel, the pink line indicates the true value of α.

Now back to the real world

- The procedure outlined above cannot be applied, because for real data we cannot generate new samples from the original population.
- However, the bootstrap approach allows us to use a computer to mimic the process of obtaining new data sets, so that we can estimate the variability of our estimate without generating additional samples.
- Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set *with replacement*.
- Each of these "bootstrap data sets" is created by sampling *with replacement*, and is the *same size* as our original dataset. As a result some observations may appear more than once in a given bootstrap data set and some not at all.

Example with just 3 observations



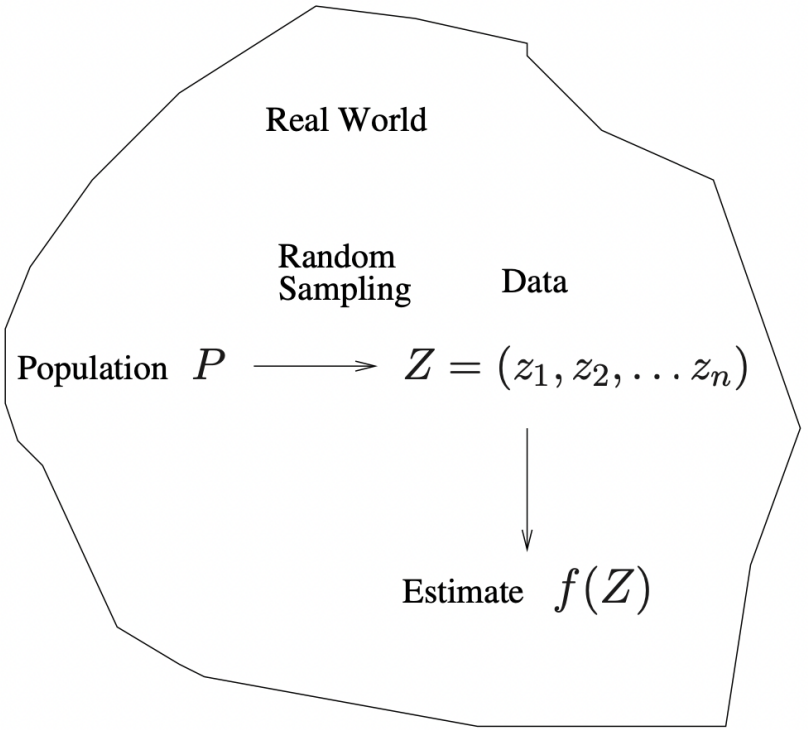
A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α

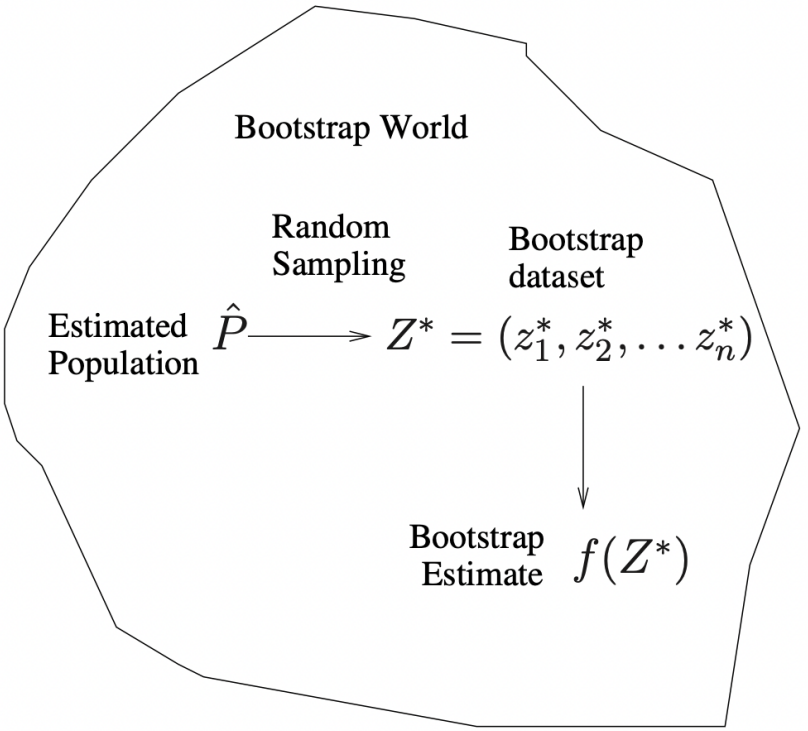
- Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$
- This procedure is repeated B times for some large value of B (say 100 or 1000), in order to produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$.
- We estimate the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}.$$

- This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set. See center and right panels of Figure on slide 29. Bootstrap results are in blue. For this example $\mathrm{SE}_B(\hat{\alpha}) = 0.087$.

A general picture for the bootstrap





Other uses of the bootstrap

- Primarily used to obtain standard errors of an estimate.
- Also provides approximate confidence intervals for a population parameter. For example, looking at the histogram in the middle panel of the Figure on slide 29, the 5% and 95% quantiles of the 1000 values is (.43, .72).
- This represents an approximate 90% confidence interval for the true α.