Supervised Learning

Model assessment vs selection

June 18th, 2021

Supervised learning

What is **statistical learning?** Preface of Introduction to Statistical Learning with Applications in R (ISLR):

> *refers to a set of tools for modeling and understanding complex datasets*

What is **supervised learning?**

Goal: uncover associations between a set of **predictor** (aka **independent** / **explanatory**) variables / features and a single **response** (or **dependent**) variable

- Ex: MLB records the batting average (hits / at-bats) for every player in each season. Can we accurately predict player X's batting average in year $t+1$ using their batting average in year t? Can we uncover meaningful relationships between other measurements (which are **predictors** / **features**) and batting average in year $t+1$ (which is the **response**)?

- Ex: We are provided player tracking data from the NFL, all x,y coordinates of every player on the field at every tenth of the second. Can we predict how far a ball-carrier will go at any given moment they have the football?

Examples of statistical learning methods / algorithms

You are probably already familiar with statistical learning - even if you did not know what the phrase meant before

Examples of statistical learning algorithms include:

- Generalized linear models (GLMs) and penalized versions (Lasso, elastic net)
- Smoothing splines, Generalized additive models (GAMs)
- Decision trees and its variants (e.g., random forest, boosting)
- Neural networks (e.g., convolutional neural networks)

Two main types of estimation **given values for predictors**:

- **Regression** models: estimate *average* value of response
- **Classification** models: determine *the most likely* class of a set of discrete response variable classes

Which method should I use in my analysis?

Depends on your goal - the big picture: **inference** vs **prediction**

Let Y be the response variable, and X be the predictors, then the **learned** model will take the form:

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

- Care about the details of $\hat{f}(X)$? $\Rightarrow$ You want to perform statistical inference
- Fine with treating $\hat{f}(X)$ as a black-box? $\Rightarrow$ Your interest is prediction

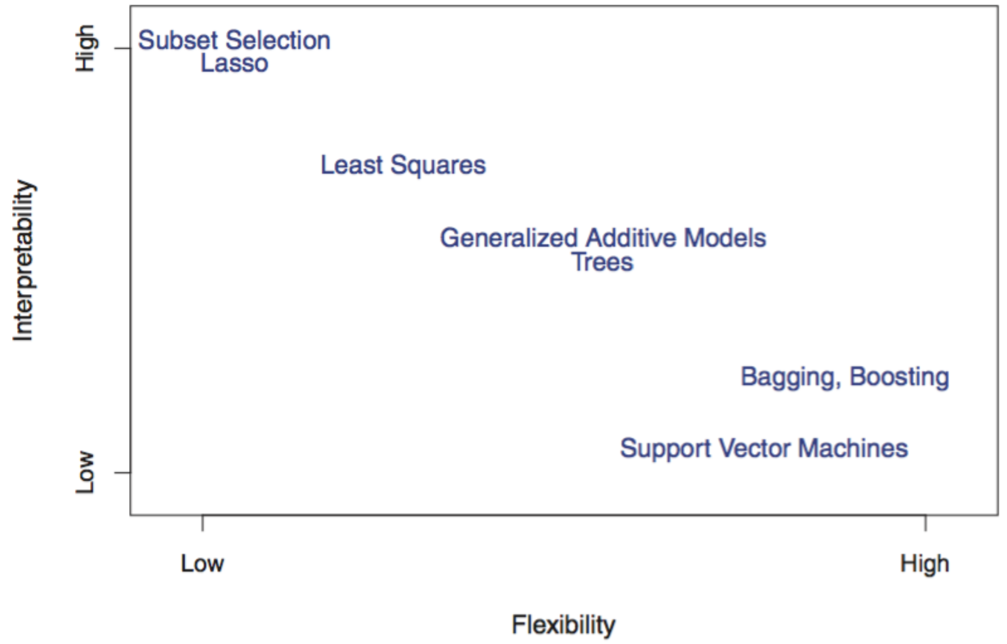
Any algorithm can be used for prediction, however options are limited for inference

Active area of research on using more black-box models for statistical inference

Model flexibility vs interpretability

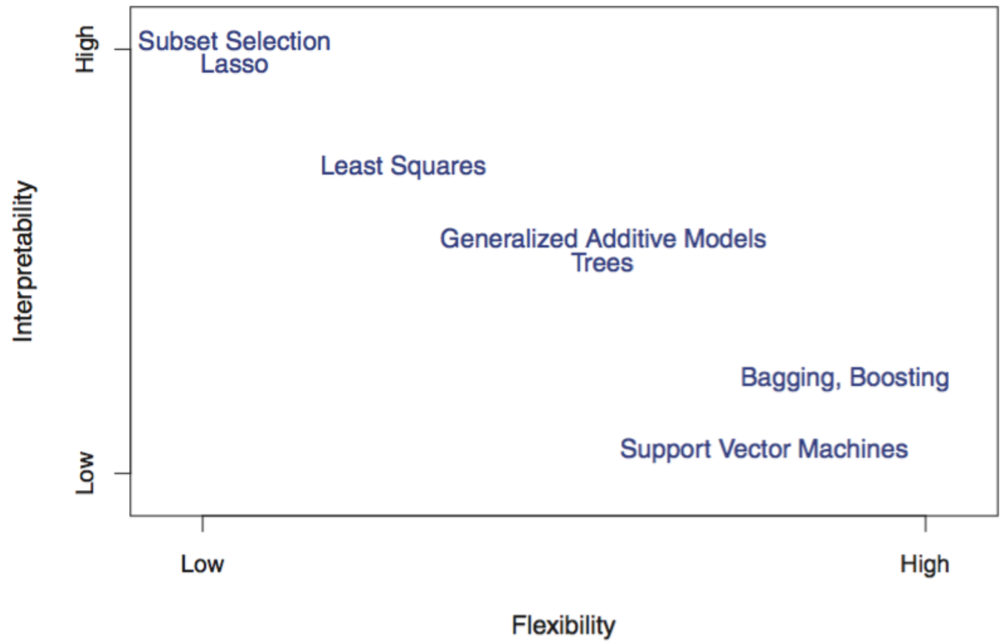
Generally speaking: **tradeoff** between a model's *flexibility* (i.e. how "wiggly" it is) and how **interpretable** it is

- Simpler the parametric form of the model $\Rightarrow$ the easier it is to interpret
- Hence why **linear regression** is popular in practice



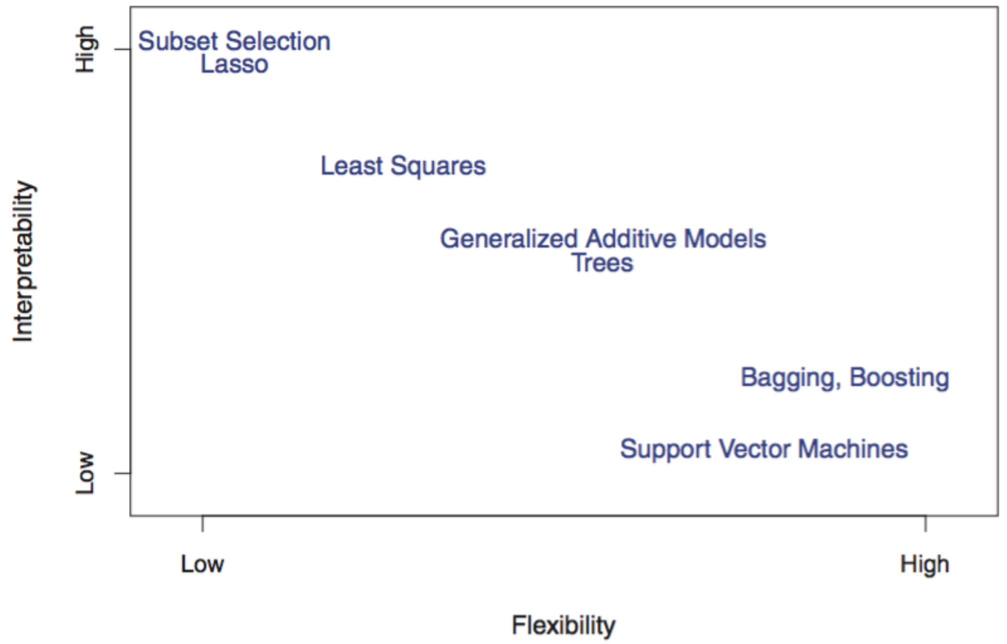
ISLR Figure 2.7

Model flexibility vs interpretability



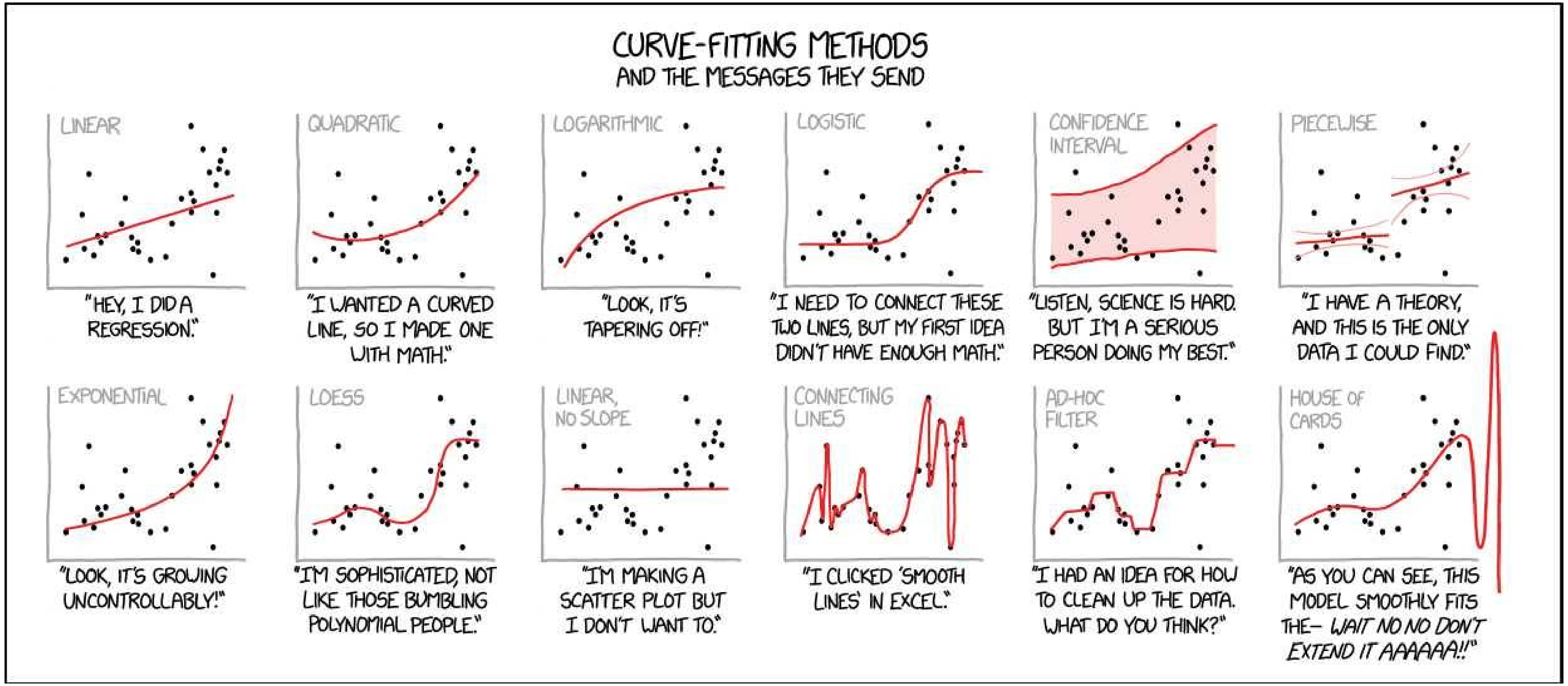
- **Parametric** models, for which we can write down a mathematical expression for $f(X)$ **before observing the data**, *a priori* (e.g. linear regression), **are inherently less flexible**
- **Nonparametric** models, in which $f(X)$ is **estimated from the data** (e.g. kernel regression)

Model flexibility vs interpretability



- If your goal is prediction $\Rightarrow$ your model can be as arbitrarily flexible as it needs to be
- We'll discuss how one estimates the optimal amount of flexibility shortly...

Looks about right...


CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND
LINEAR
"HEY, I DID A REGRESSION."
QUADRATIC
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."
LOGARITHMIC
"LOOK, IT'S TAPERING OFF!"
LOGISTIC
"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."
CONFIDENCE INTERVAL
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."
PIECEWISE
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."
EXPONENTIAL
"LOOK, IT'S GROWING UNCONTROLLABLY!"
LOESS
"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."
LINEAR, NO SLOPE
"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."
CONNECTING LINES
"I CLICKED 'SMOOTH LINES' IN EXCEL."
AD-HOC FILTER
"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"
HOUSE OF CARDS
"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE- WAIT NO NO DON'T EXTEND IT AAAAAA!!"

Model assessment vs selection, what's the difference?

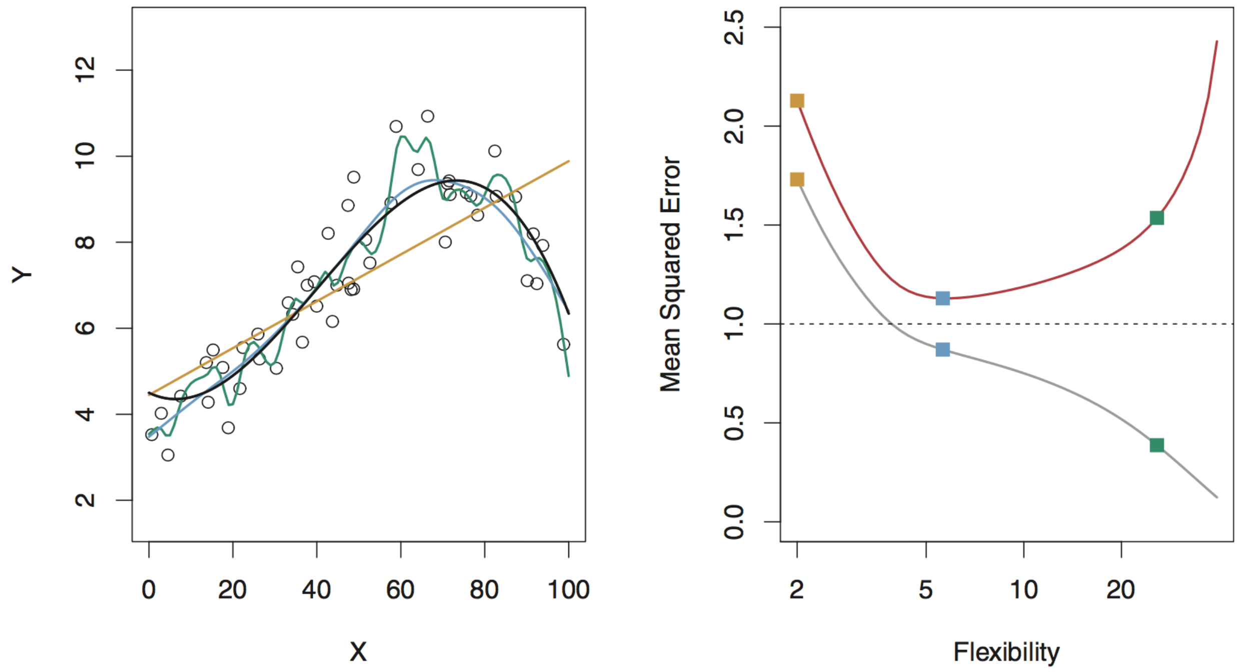
Model assessment:

- **evaluating how well a learned model performs**, via the use of a single-number metric

Model selection:

- selecting the best model from a suite of learned models (e.g., linear regression, random forest, etc.)

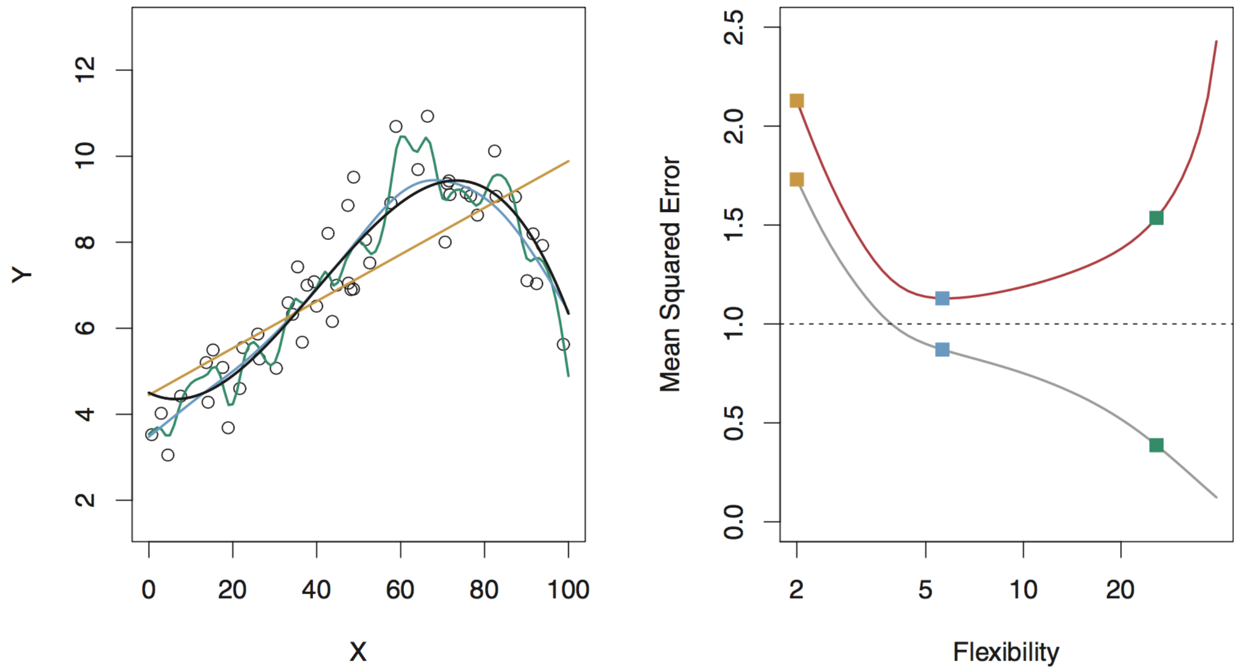
Model **flexibility** (ISLR Figure 2.9)



- Left panel: intuitive notion of the meaning of model flexibility
- Data are generated from a smoothly varying non-linear model (shown in black), with random noise added:

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

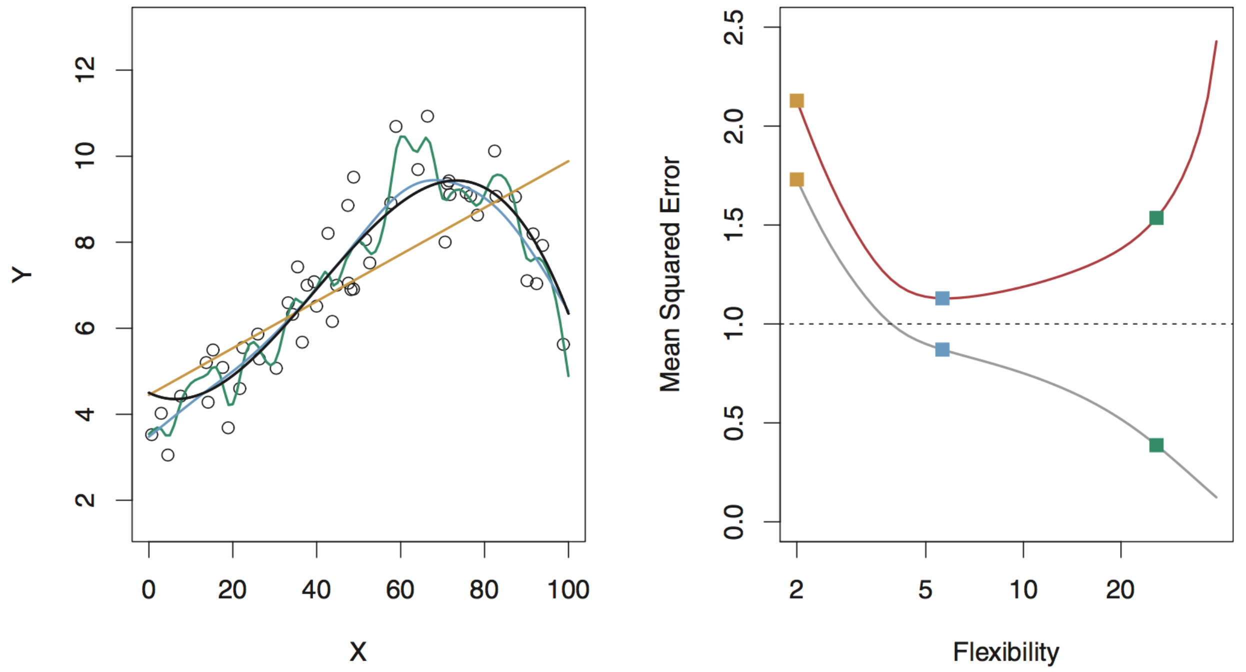
Model **flexibility**



Orange line: an inflexible, fully parametrized model (simple linear regression)

- **Cannot** provide a good estimate of $f(X)$
- Cannot **overfit** by modeling the noisy deviations of the data from $f(X)$

Model flexibility



Green line: an overly flexible, nonparametric model

- **It can** provide a good estimate of $f(X)$, **BUT** it goes too far and overfits by modeling the noise

This is NOT generalizable: bad job of predicting response given new data NOT used in learning the model

So... how do we deal with flexibility?

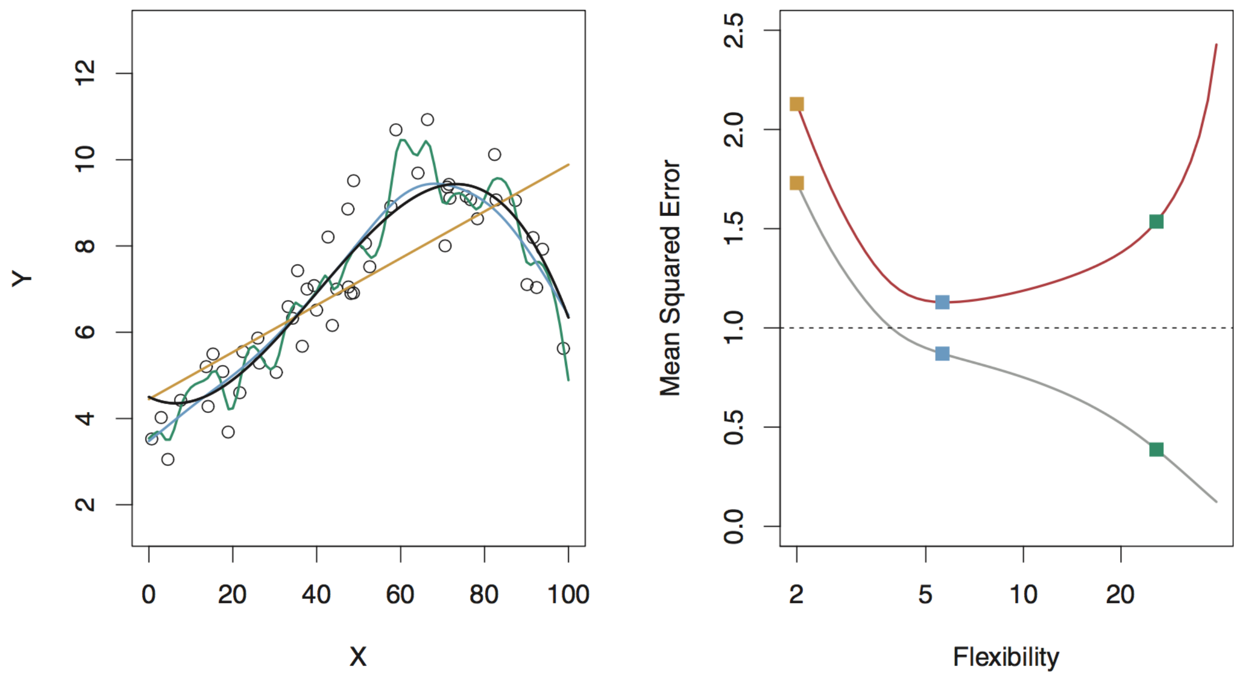
GOAL: We want to learn a statistical model that provides a good estimate of $f(X)$ **without overfitting**

There are two common approaches:

- We can **split the data into two groups**:
    - **training** data: data used to train models,
    - **test** data: data used to test them
    - By assessing models using "held-out" test set data, we act to ensure that we get a **generalizable(!)** estimate of $f(X)$
- We can **repeat data splitting k times**:
    - Each observation is placed in the "held-out" / test data exactly once
    - This is called **k-fold cross validation** (typically set k to 5 or 10)

k-fold cross validation is the preferred approach, but the tradeoff is that CV analyses take $\sim k$ times longer than analyses that utilize data splitting

Model assessment



- Right panel shows **a metric of model assessment**, the mean squared error (MSE) as a function of flexibility for both a training and test datasets
- Training error (gray line) **decreases as flexibility increases**
- Test error (red line) decreases while flexibility increases **until** the point a good estimate of $f(X)$ is reached, and then it **increases as it overfits to the training data**

Brief note on reproducibility

An important aspect of a statistical analysis is that it be reproducible. You should…

1. Record your analysis in a notebook, via, e.g., `R Markdown` or `Jupyter`. A notebook should be complete such that if you give it and datasets to someone, that someone should be able to recreate the entire analysis and achieve the exact same results. To ensure the achivement of the exact same results, you should…

2. Manually set the random-number generator seed before each instance of random sampling in your analysis (such as when you assign data to training or test sets, or to folds):

```
set.seed(101)     # can be any number...
sample(10,3)      # sample three numbers between 1 and 10 inclusive
```

```
## [1]  9 10  6
```

```
set.seed(101)
sample(10,3)      # voila: the same three numbers!
```

```
## [1]  9 10  6
```

Model assessment metrics

Loss function (aka *objective* or *cost* function) is a metric that represents **the quality of fit of a model**

For regression we typically use **mean squared error (MSE)**

- quadratic loss: squared differences between model predictions $\hat{f}\left(X\right)$ and observed data Y

$$\text{MSE} = \frac{1}{n}\sum_{i}^{n}(Y_i - \hat{f}\left(X_i\right))^2$$

For classification, the situation is not quite so clear-cut

- **misclassification rate (MCR)**: percentage of predictions that are wrong
- **area under curve (AUC)** (we'll come back to this later)
- interpretation can be affected by **class imbalance**:
  - if two classes are equally represented in a dataset, an MCR of 2% is good
  - but if one class comprises 99% of the data, that 2% is no longer such a good result…

Back to model selection

Model selection: picking the best model from a suite of possible models

- Picking the best covariance constraints (e.g. *VVV*) or number of clusters with BIC
- Picking the best regression model based on **MSE**, or best classification model based on **MCR**

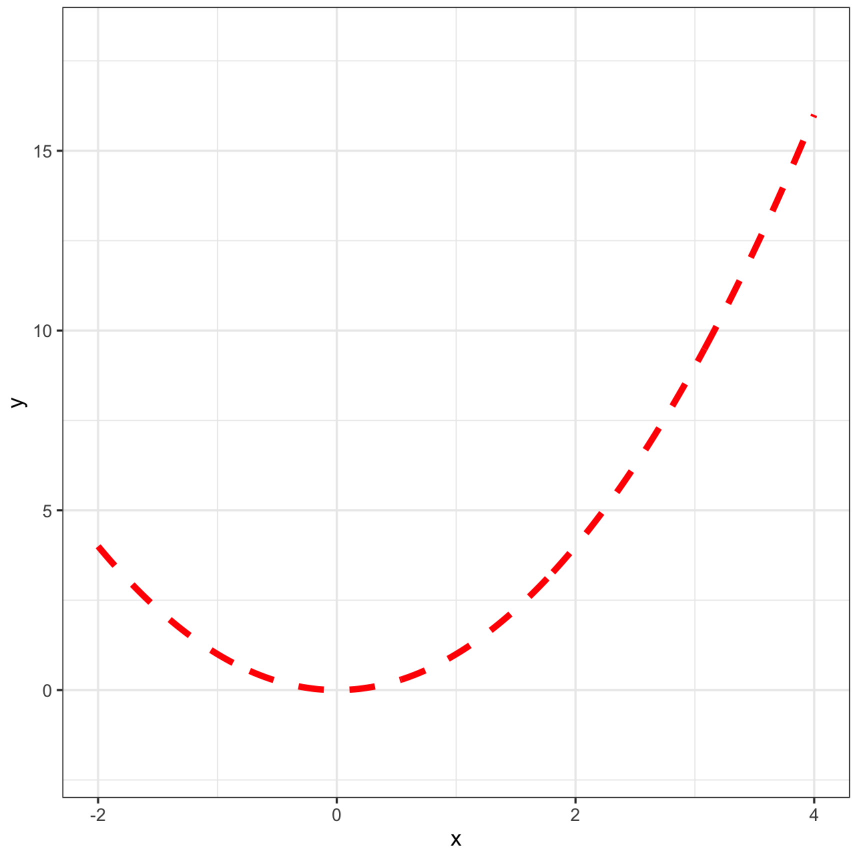
Two things to keep in mind:

1. **Ensure an apples-to-apples comparison of metrics**
    - every model should be learned using **the same training and test data**! Do not resample the data between the time when you, e.g., perform linear regression and vs you perform random forest.
2. **An assessment metric is a random variable**, i.e., if you choose different data to be in your training set, the metric will be different.

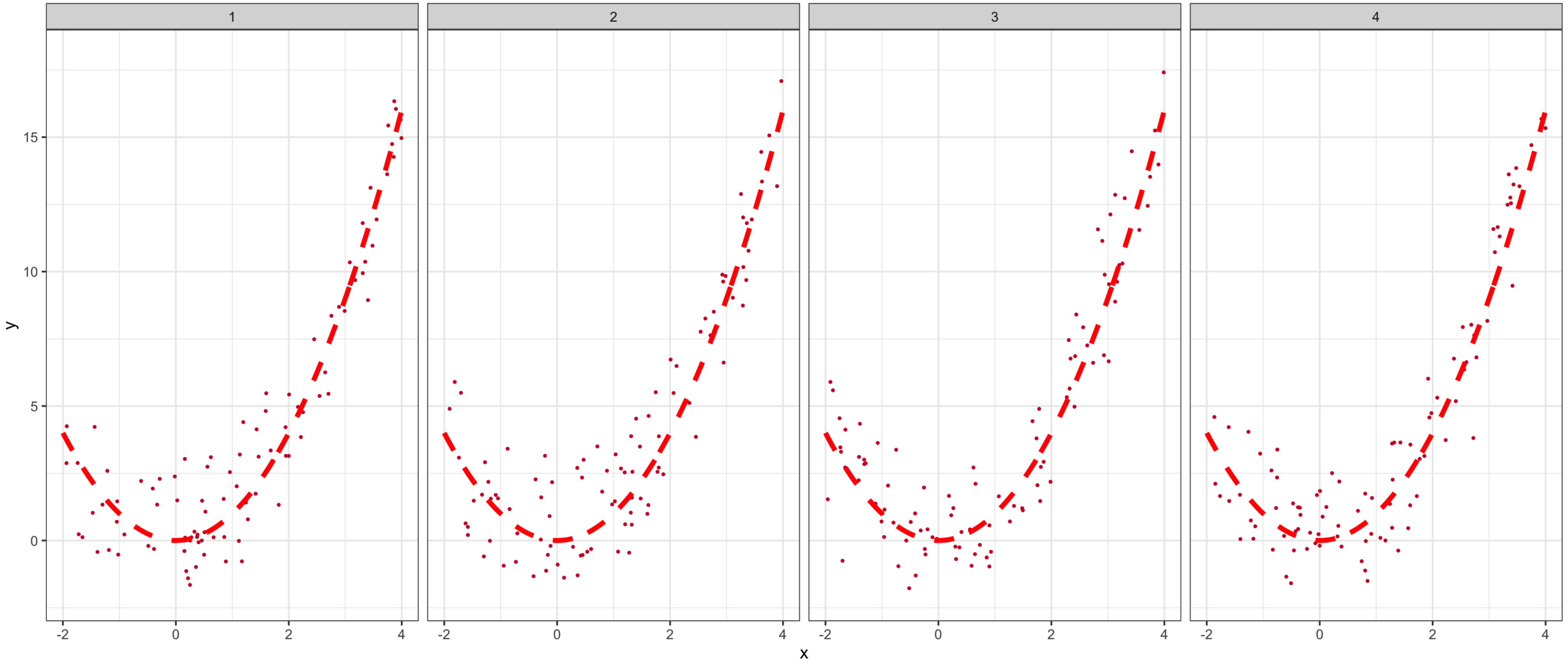
For regression, a third point should be kept in mind: **a metric like the MSE is unit-dependent**

- an MSE of 0.001 in one analysis context is not necessarily better or worse than an MSE of 100 in another context

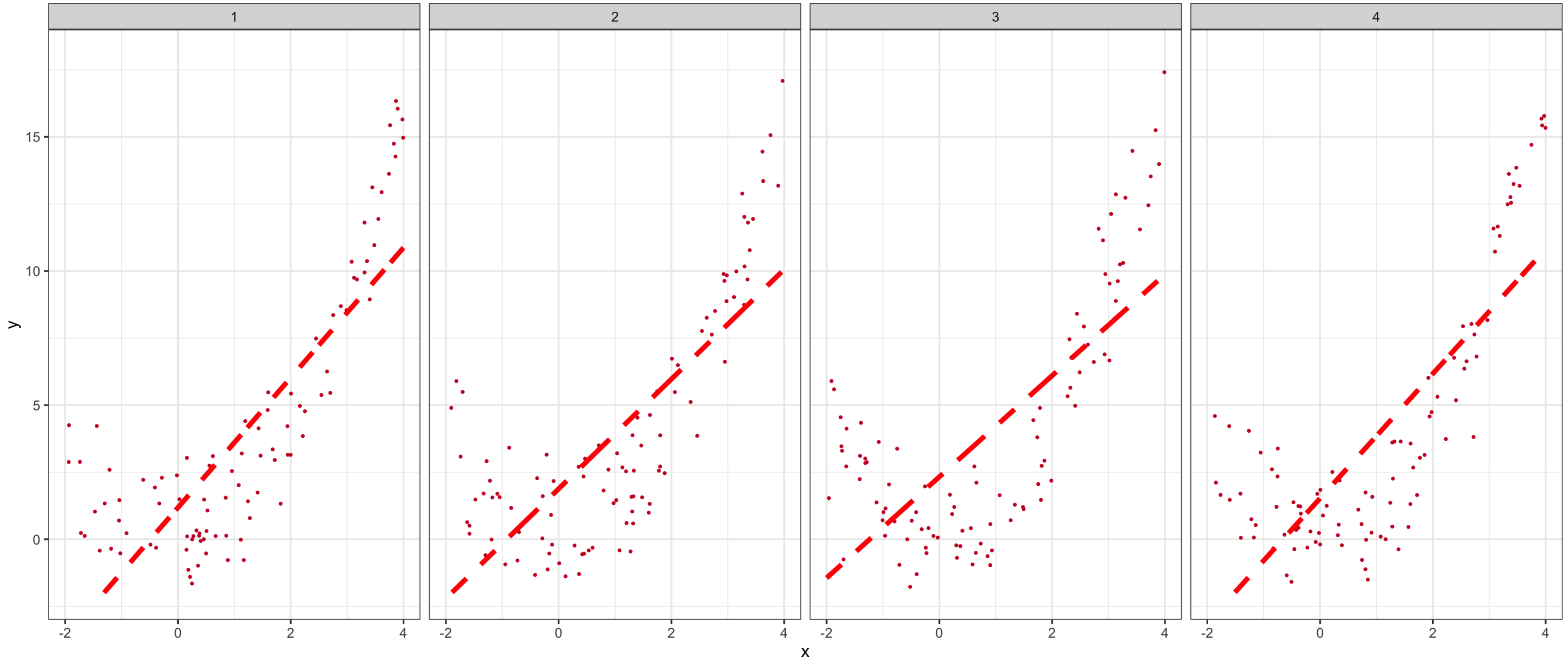
An example **true** model



The repeated experiments...



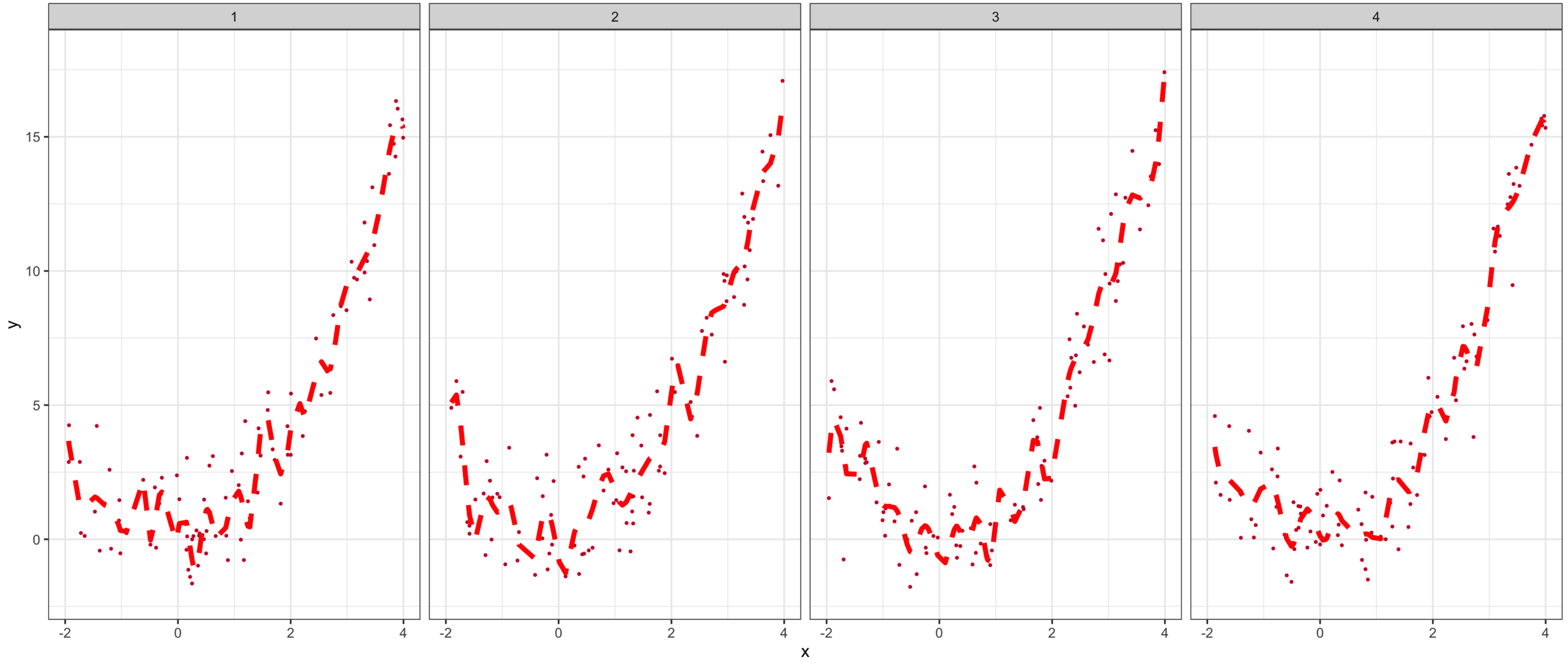
The linear regression fits



Look at the plots. For any given value of x:

- The *average* estimated y value is offset from the truth (**high bias**)
- The dispersion (variance) in the estimated y values is relatively small (**low variance**)

The spline fits



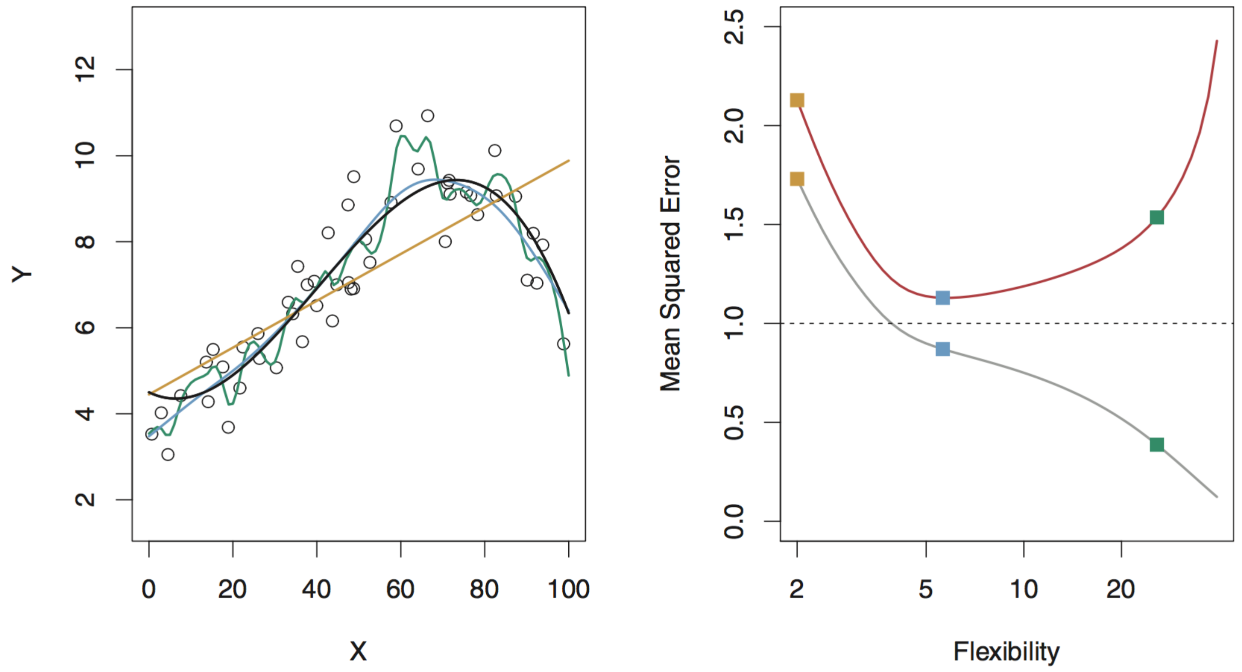
Look at the plots. For any given value of x:

- The *average* estimated y value approximately matches the truth (**low bias**)
- The dispersion (variance) in the estimated y values is relatively large (**high variance**)

Bias-variance tradeoff

"Best" model minimizes the test-set MSE, where the **true** MSE can be decomposed into:

$$\mathrm{MSE} = (\mathrm{Bias})^2 + \mathrm{Variance}$$



Towards the left: high bias, low variance. Towards the right: low bias, high variance.

Optimal amount of flexibility lies somewhere in the middle