Prediction models

Seminar Data Science for Economics

Madina Kurmangaliyeva
m.kurmangaliyeva@uvt.nl

Spring 2021

Tilburg University

Population

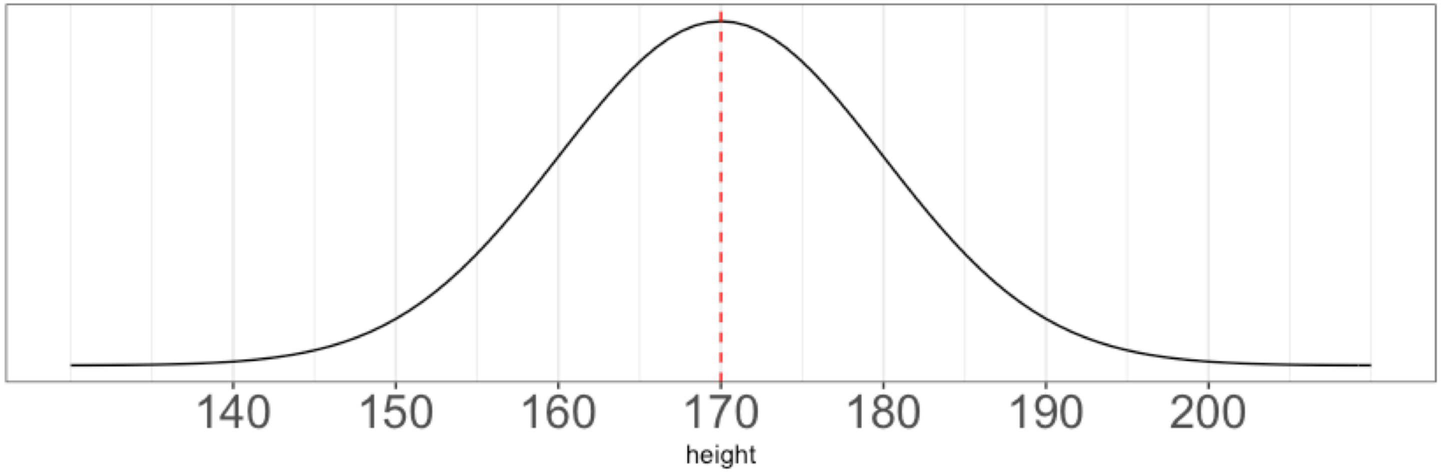
Data-generating process (DGP)

f(.)

Sample

Random sample from the DGP

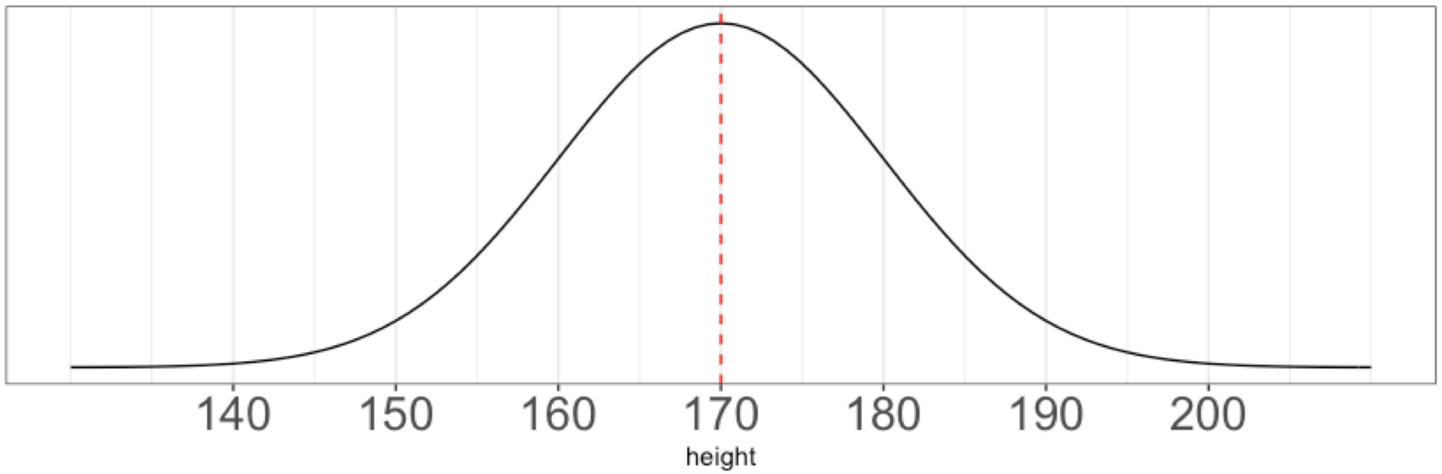
Population



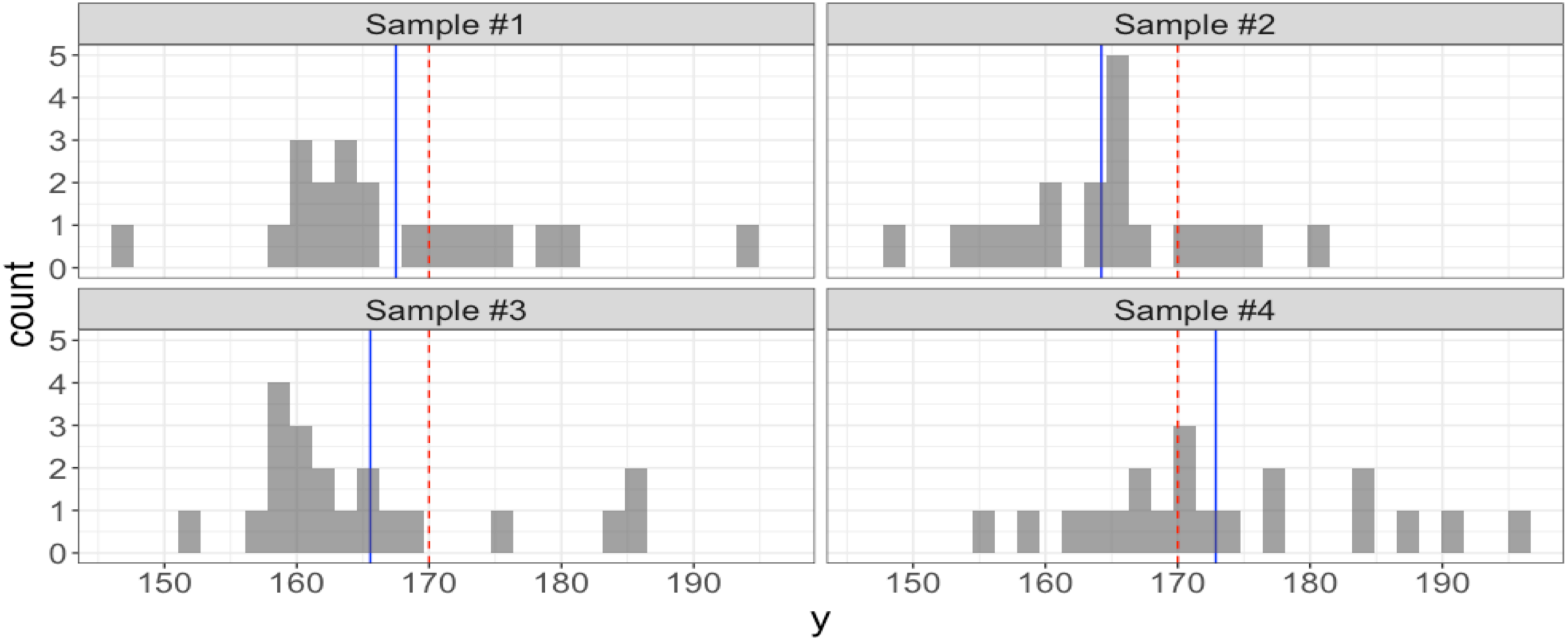
Sample

blue line is the sample mean, and red dotted line is the population mean

Population



Sample



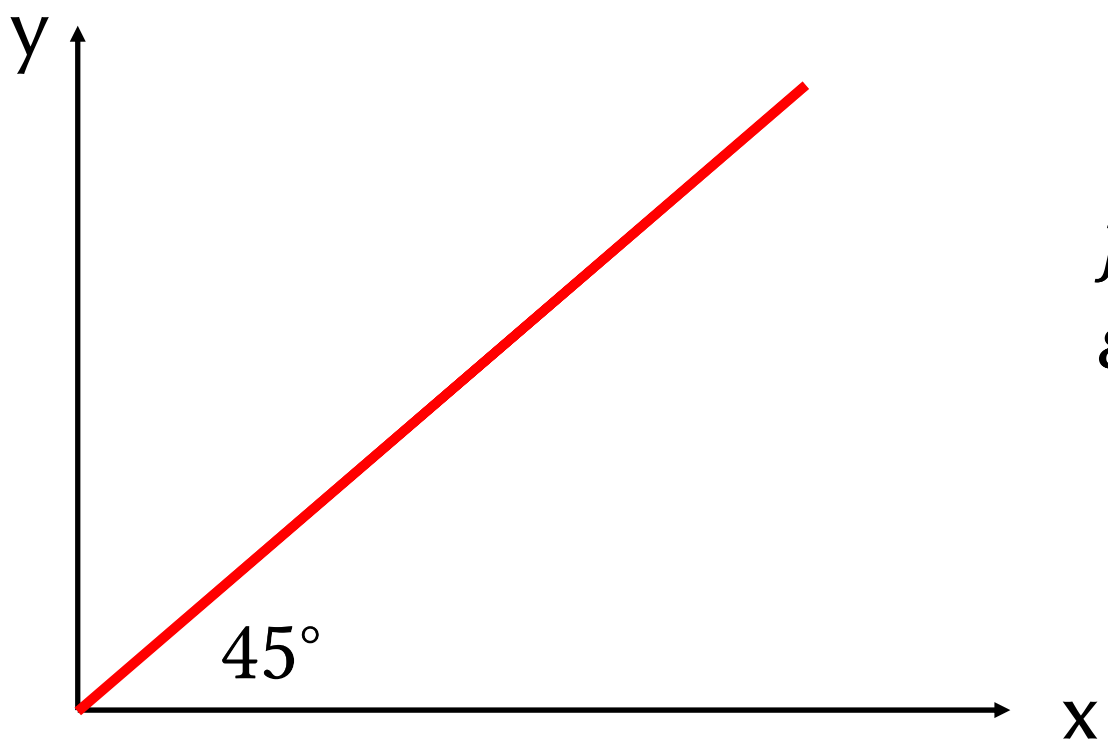
blue line is the sample mean, and red dotted line is the population mean

Population

$$y = f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

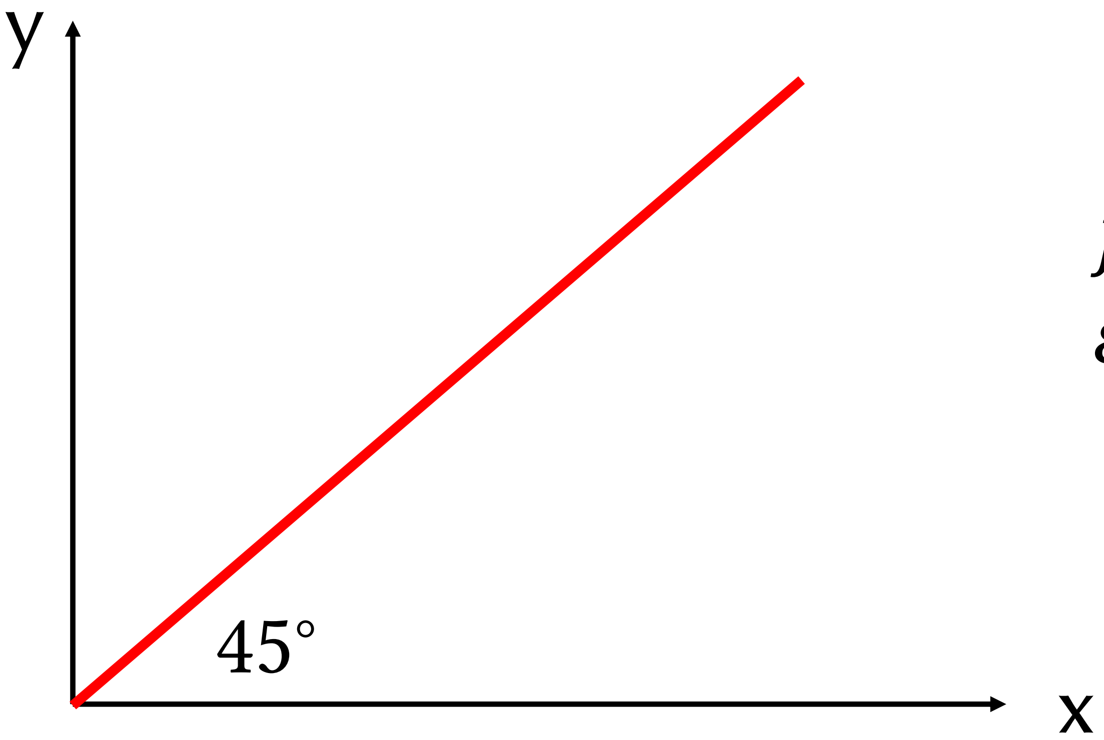
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

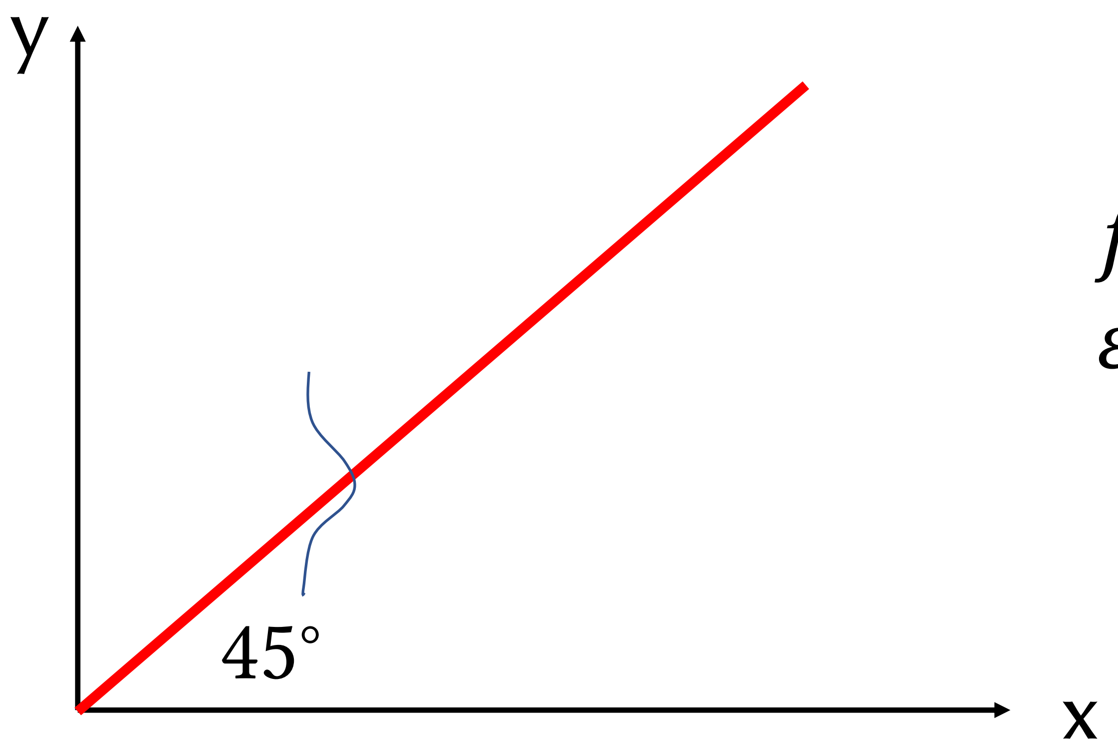
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

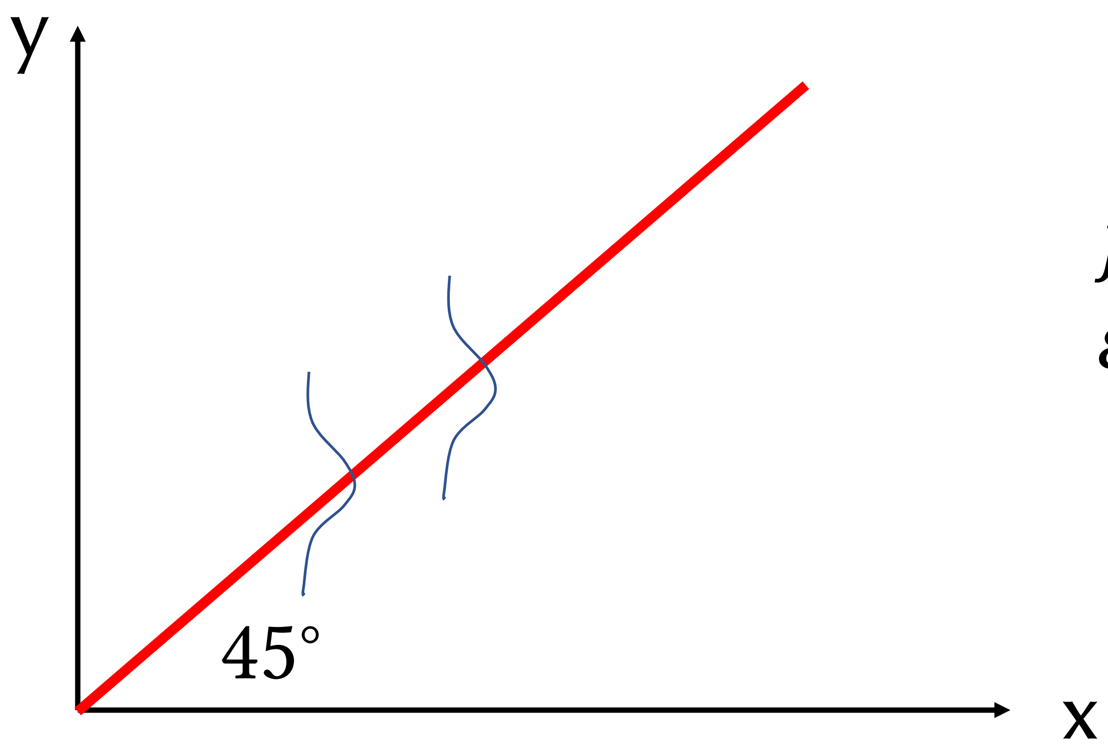
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

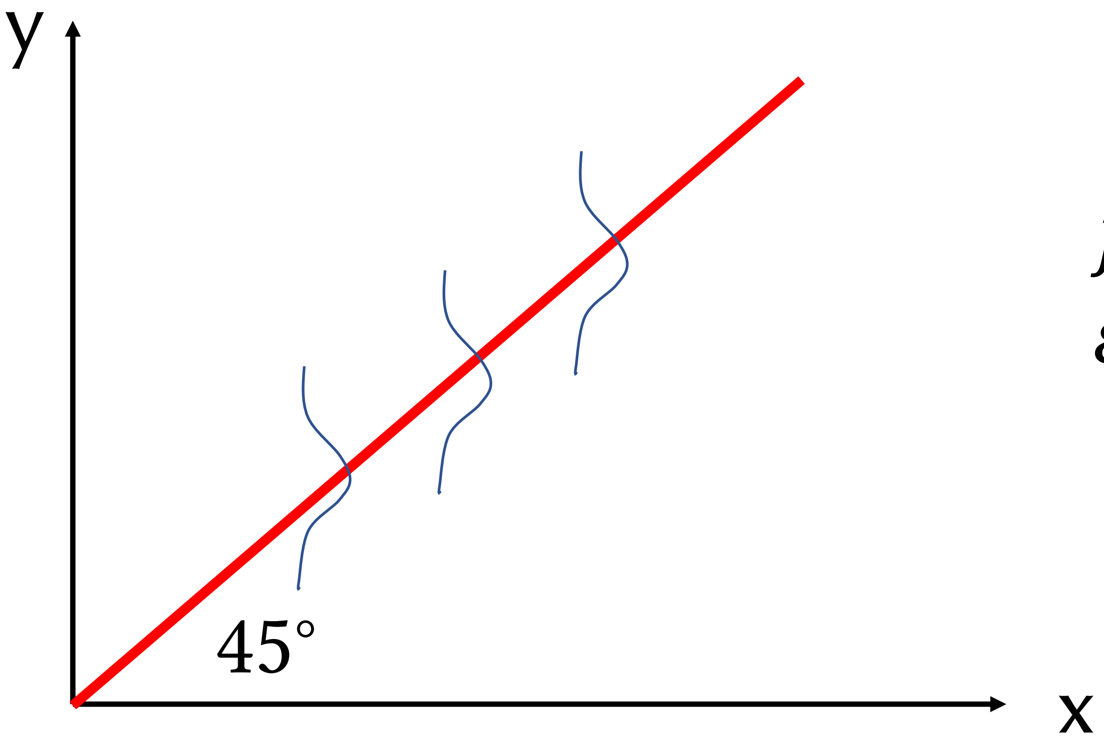
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

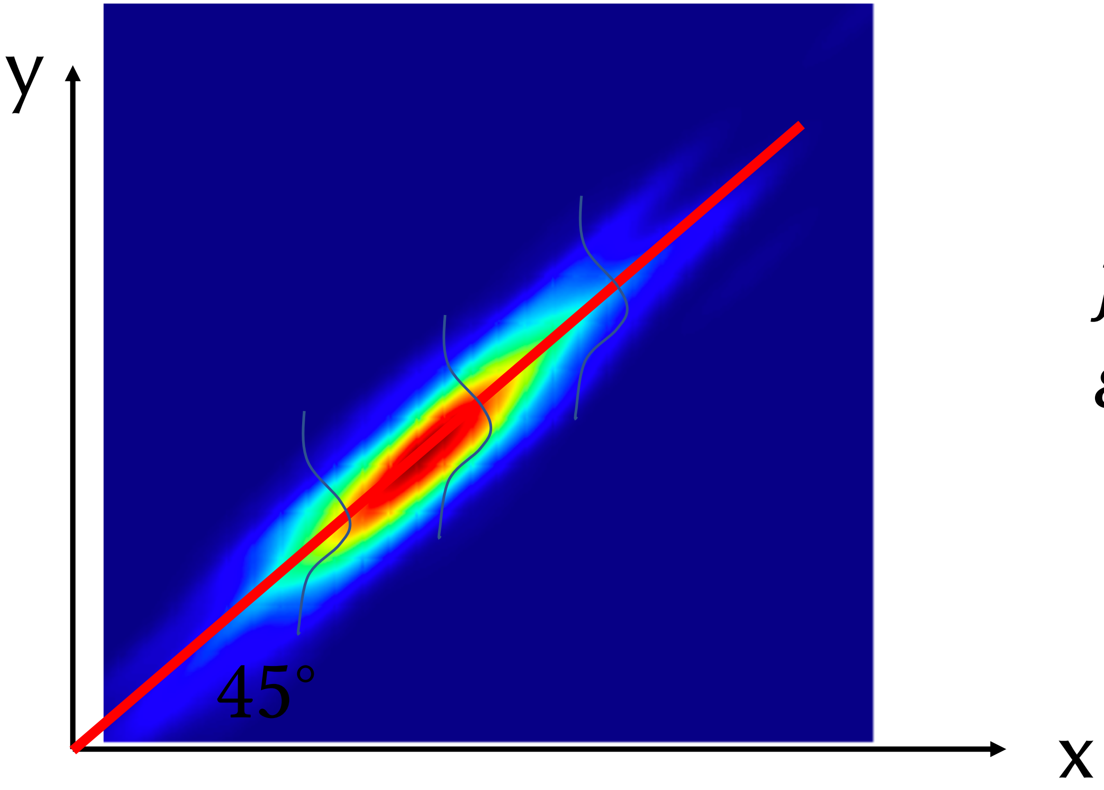
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

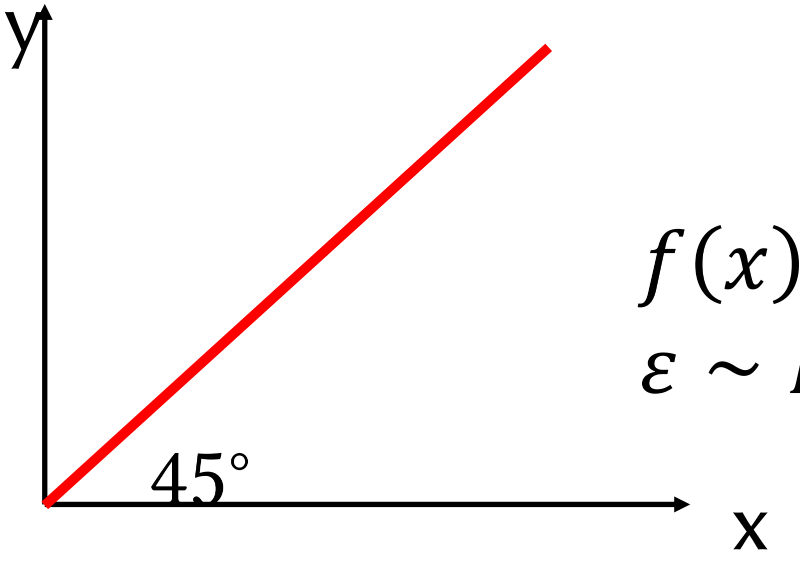
Population



$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

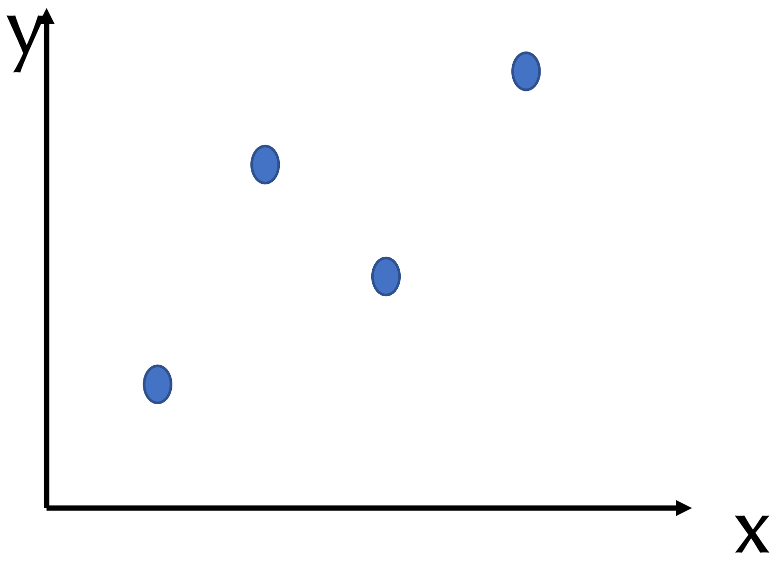
Sample

Population



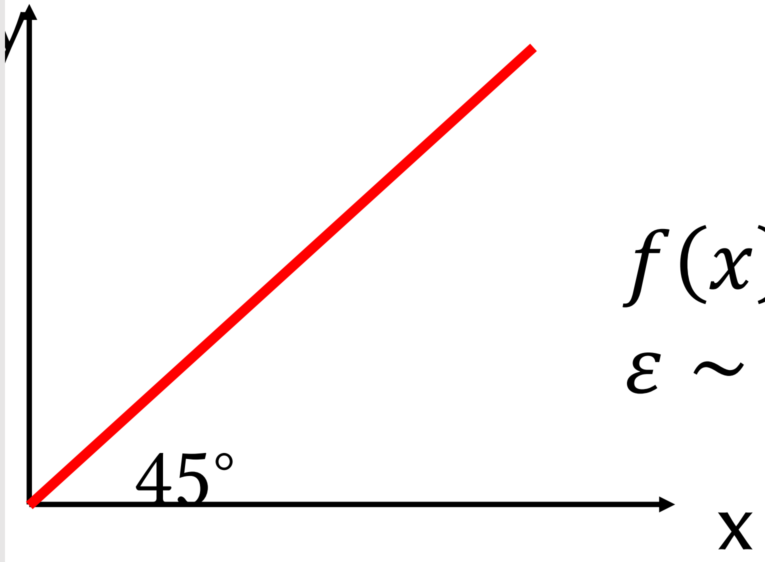
$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample



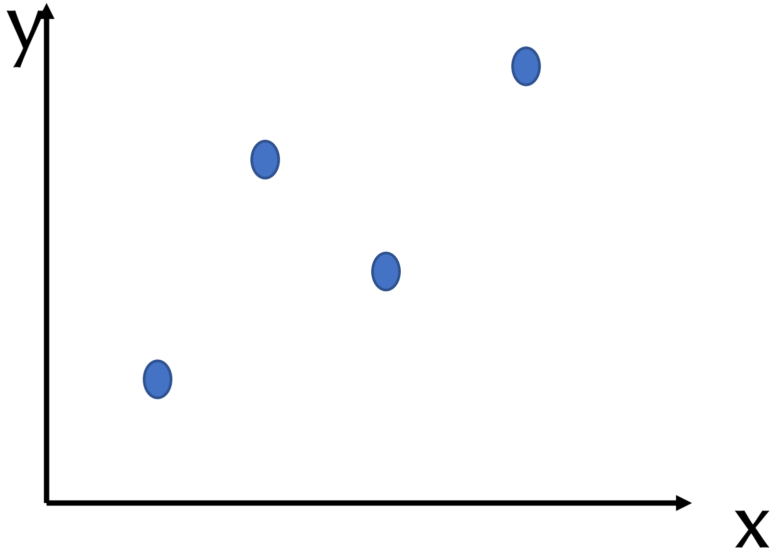
If we want to predict y from x there is nothing better than to know the true DGP relation between y and x,
but we do not know it

Pc

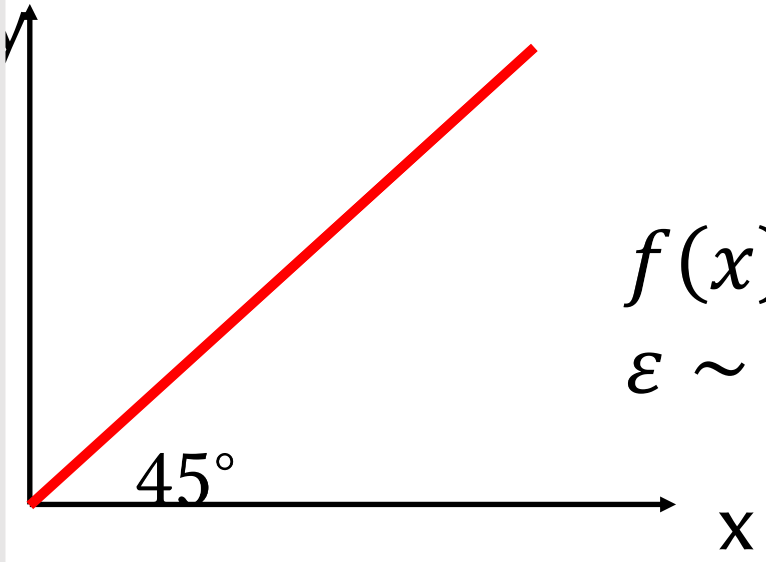


$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

Sample

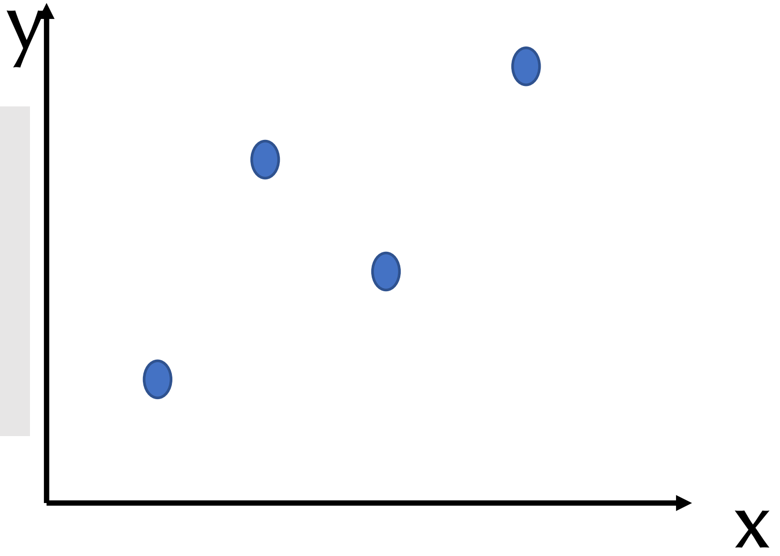


If we want to predict y from x there is nothing better than to know the true DGP relation between y and x,
but we do not know it

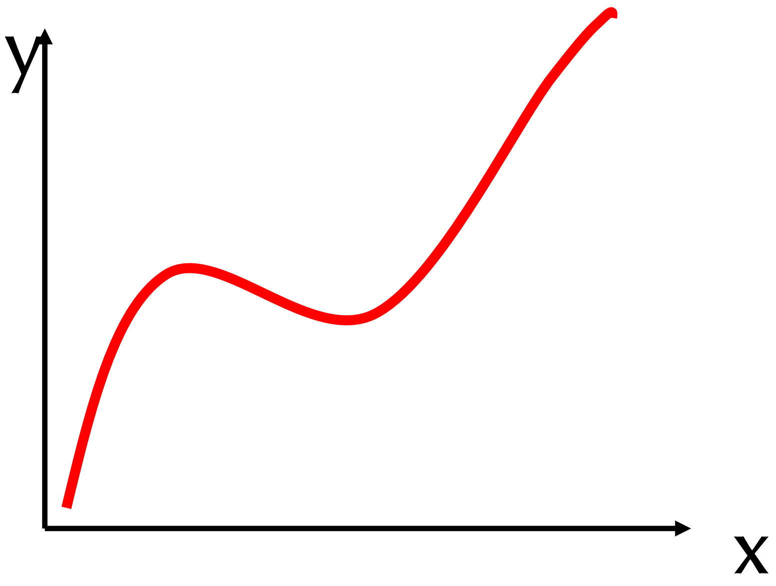


$$f(x) = x + \varepsilon,$$
$$\varepsilon \sim N(0,1)$$

All we got is this sample, from which we try to approximate the DGP

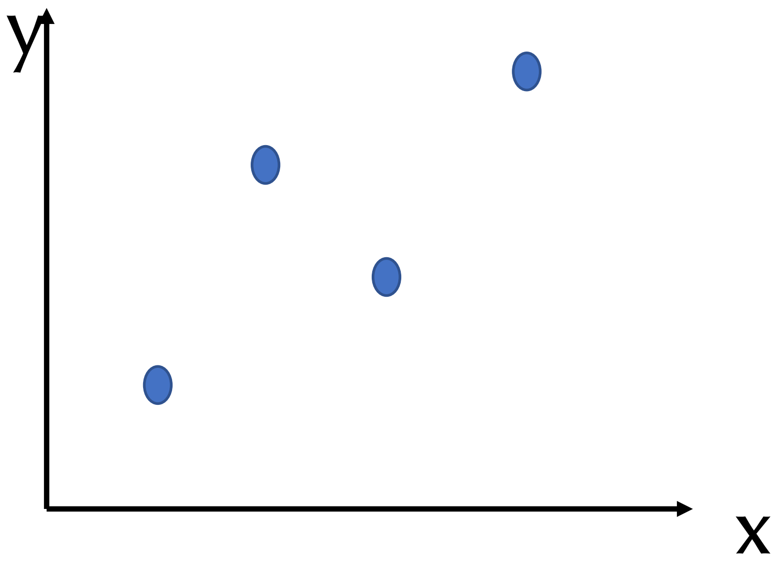


Population



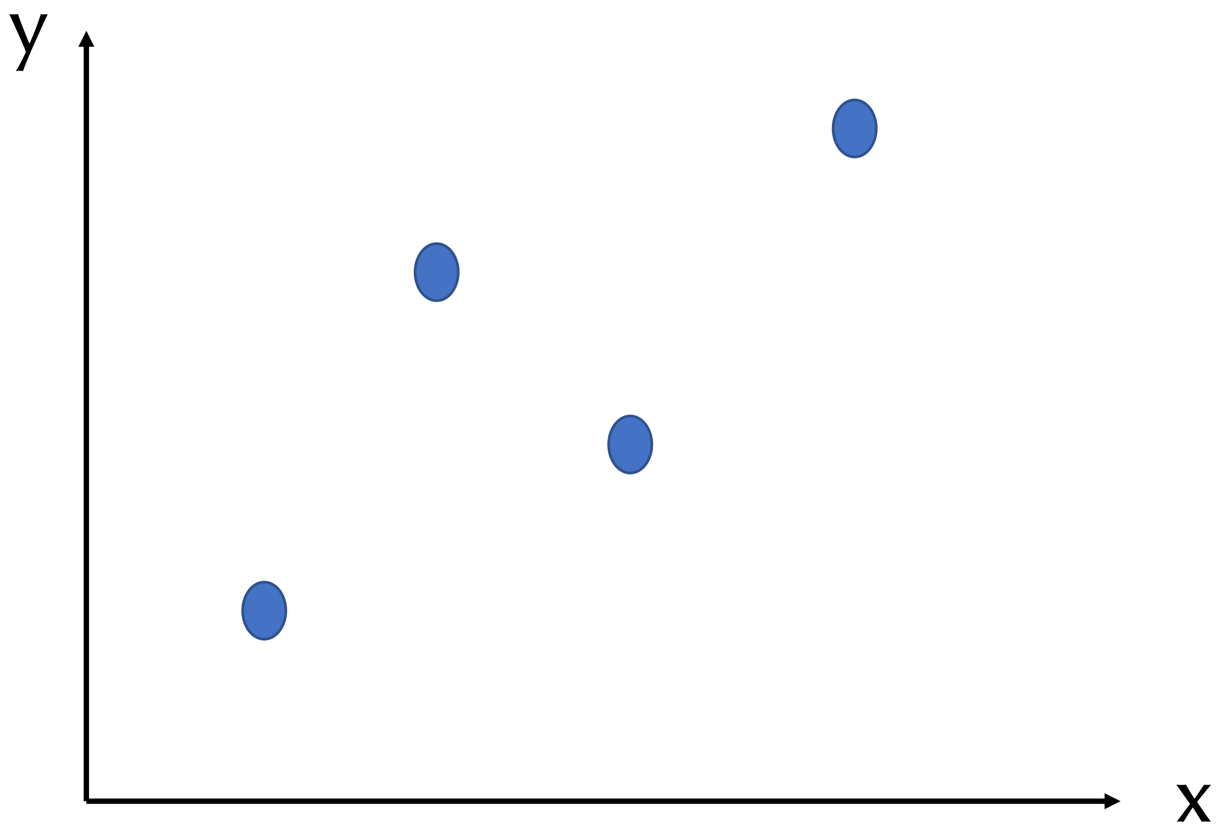
But the same sample could have been generated by a different DGP

Sample



Sample

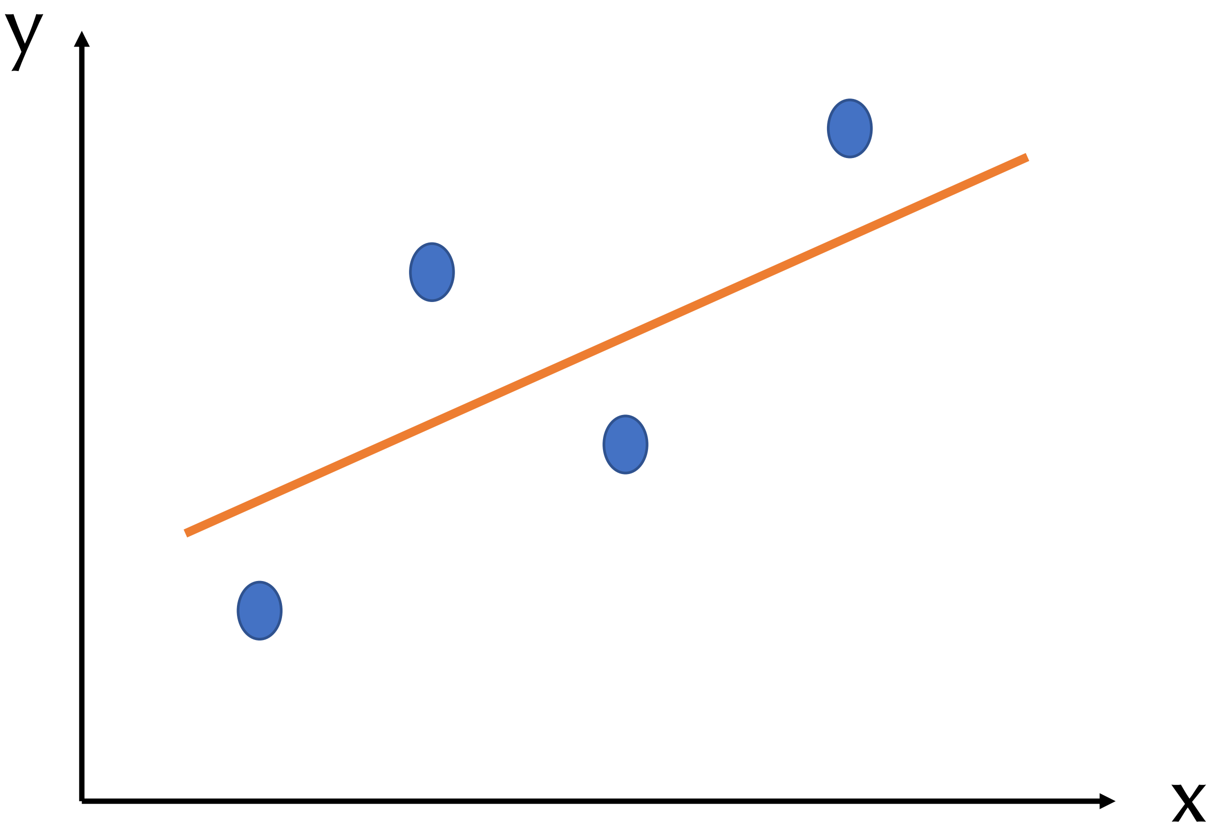
How do we approximate the DGP?



Sample

By a straight line?

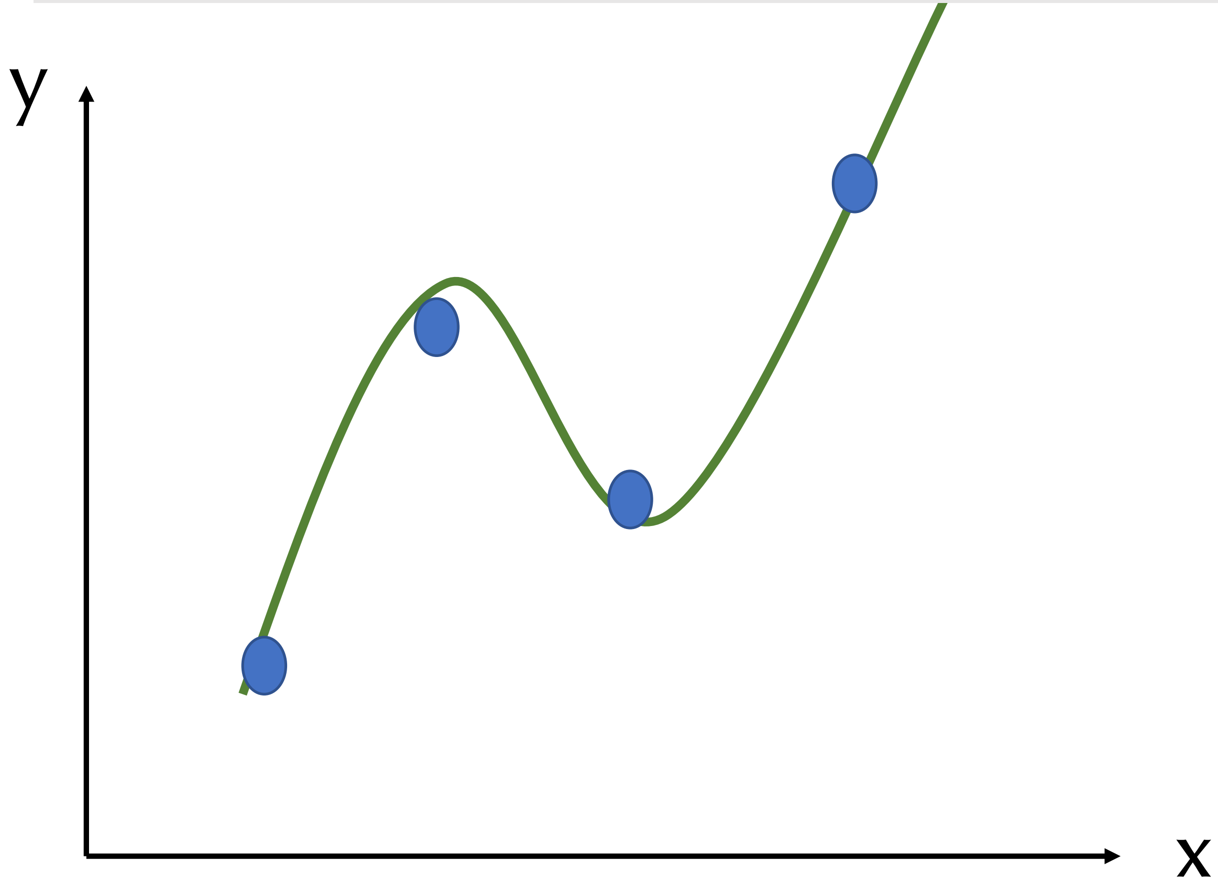
OLS $y_i = \alpha + \beta x_i + e_i$



Sample

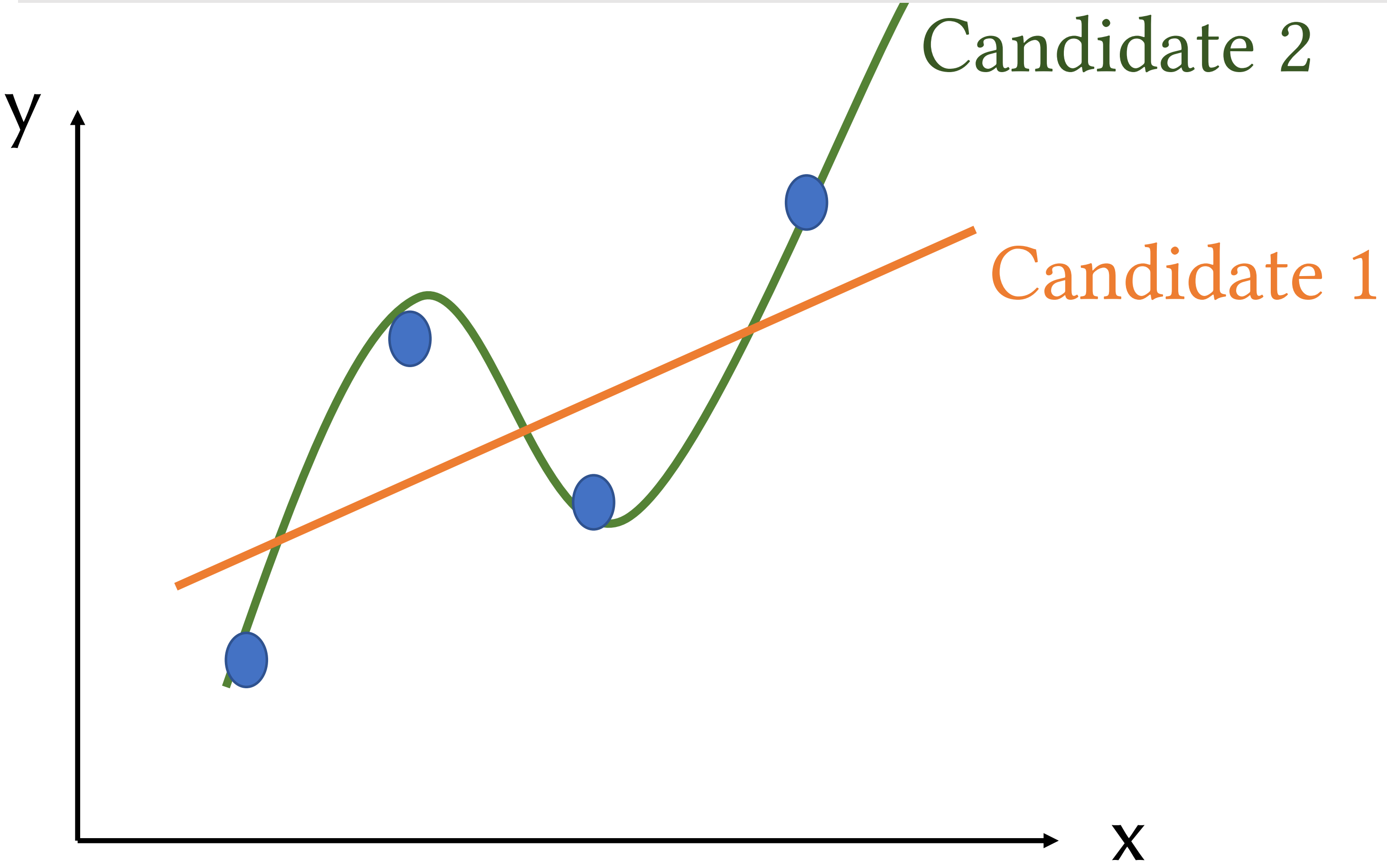
Or by a higher order polynomials?

OLS $y_i = \alpha + \beta x_i + \gamma x_i^2 + \delta x_i^3 + e_i$

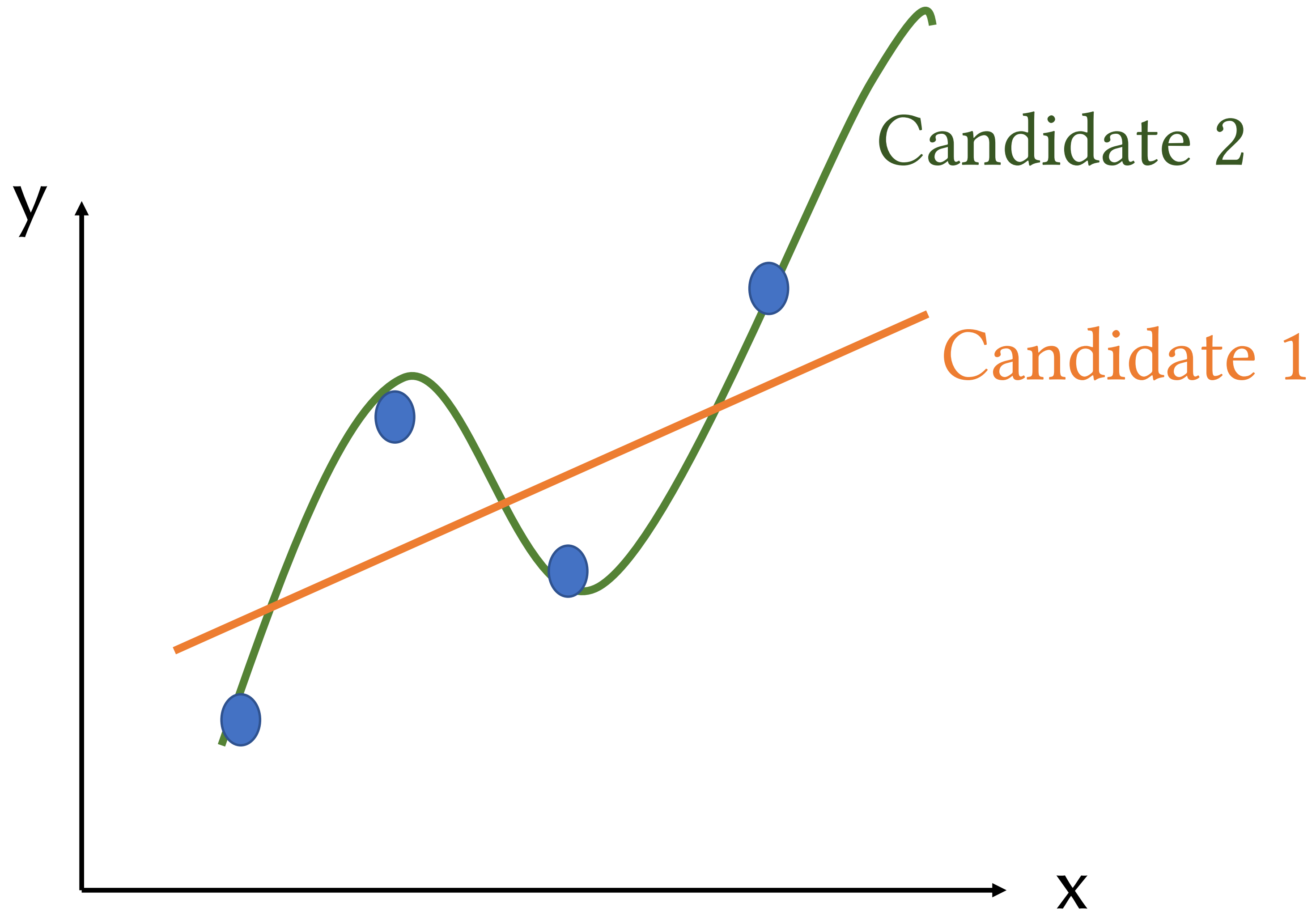


Sample

So we have different candidate approximations to the DGP



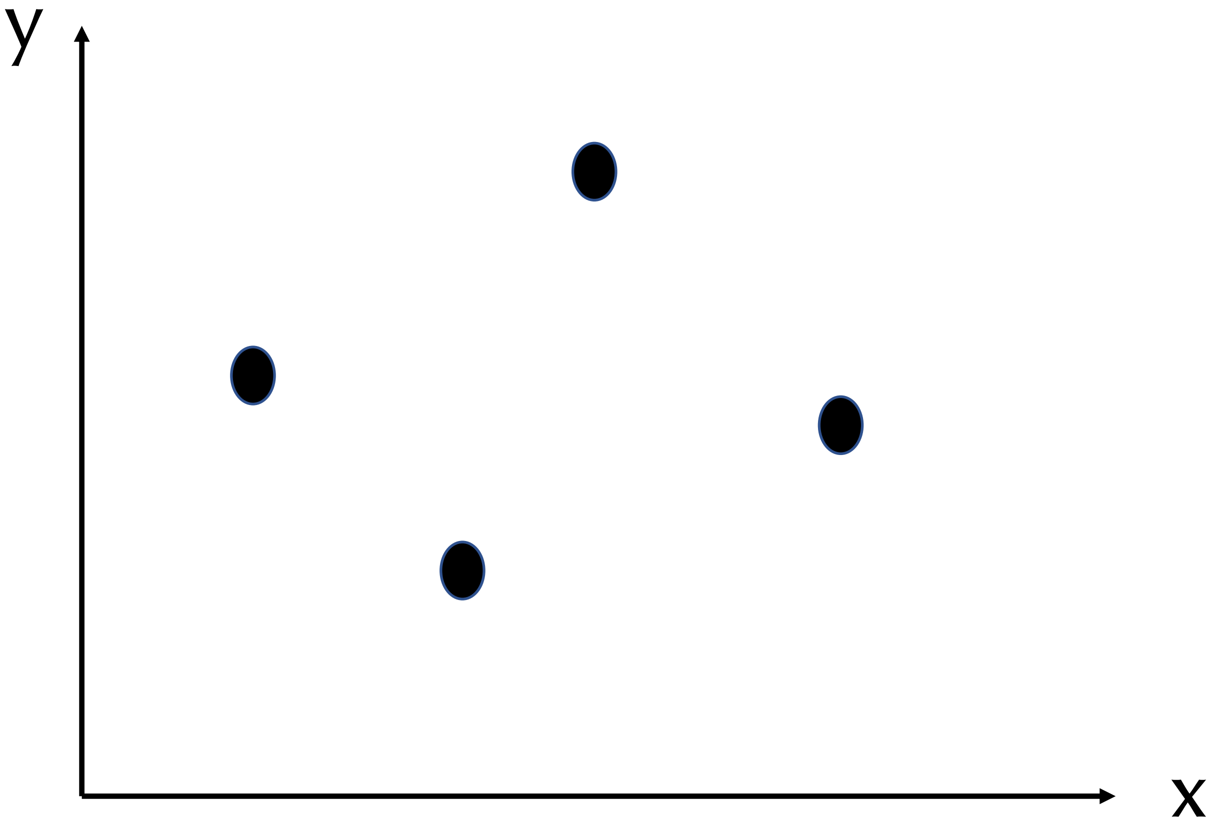
Sample



Which one is better?

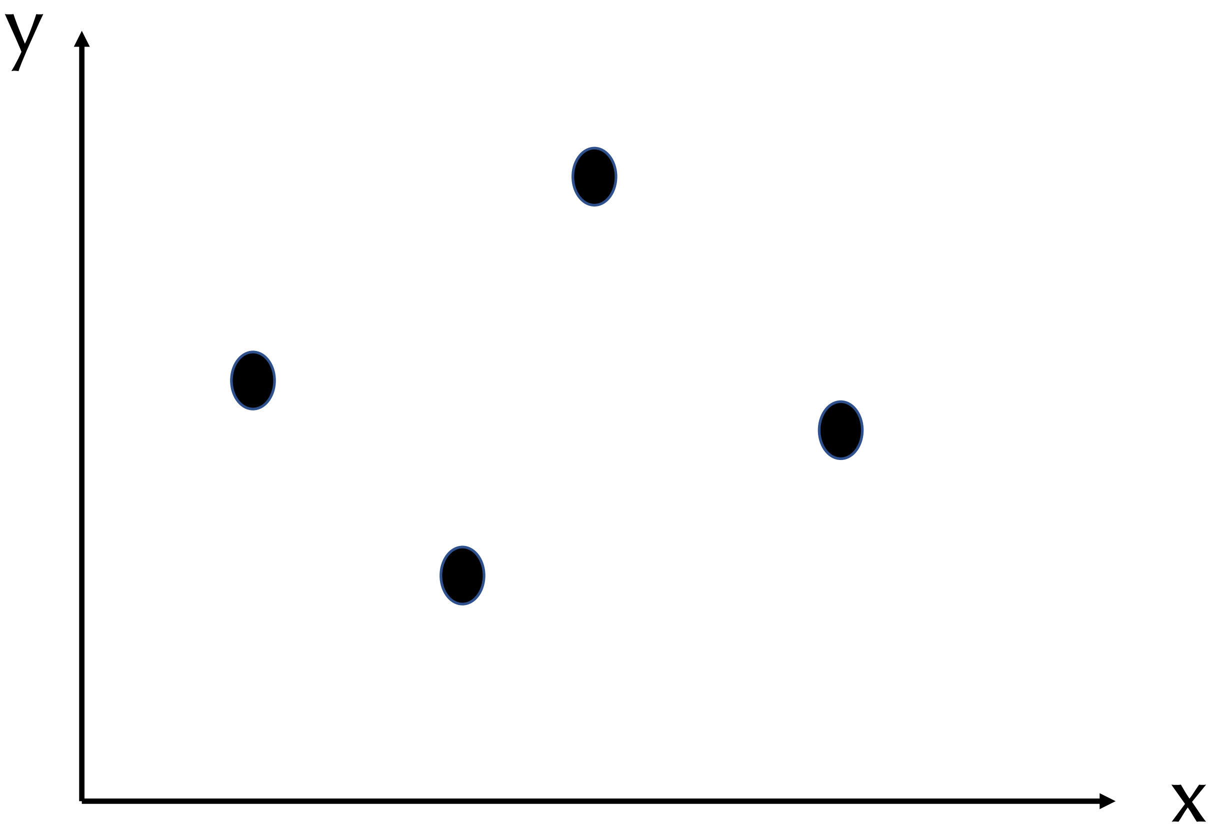
Sample

Luckily we have another (independent) random sample from the same DGP



Sample

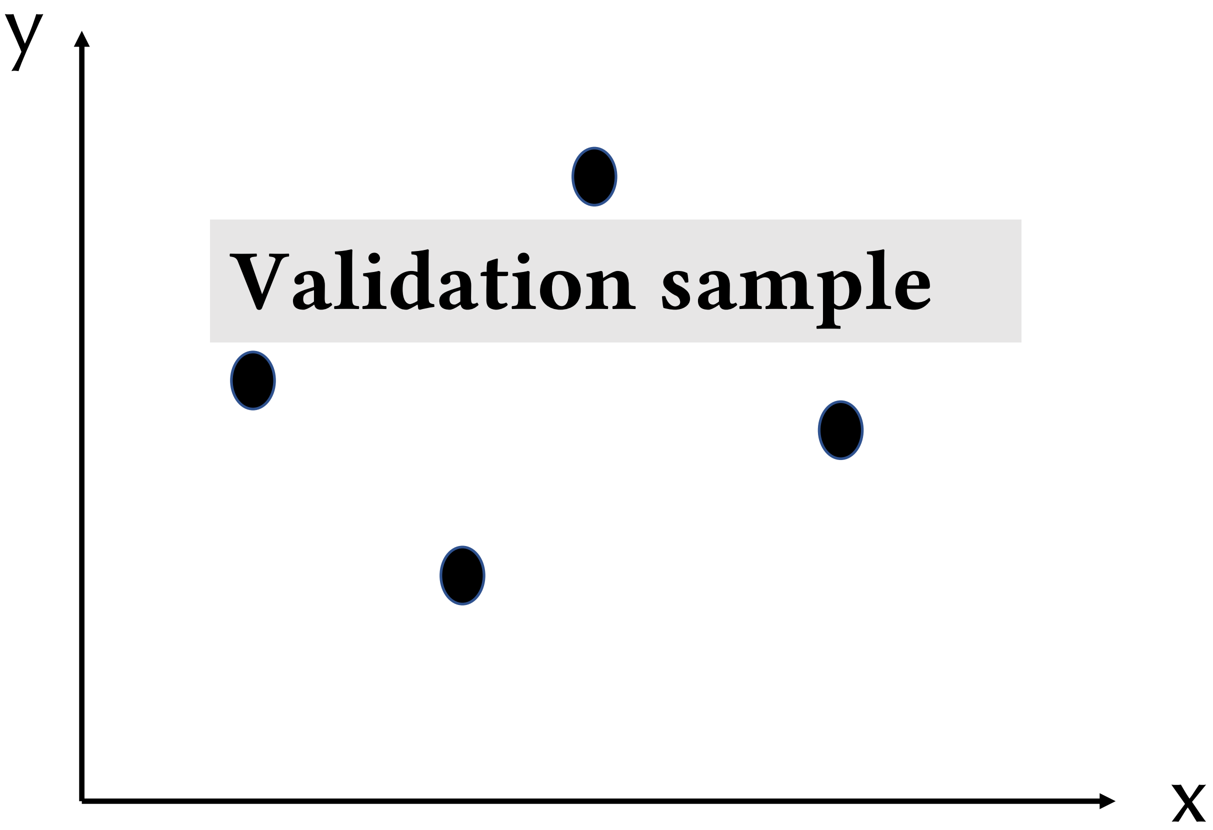
Luckily we have another (independent) random sample from the same DGP



So we can test which candidate is better at predicting! Horse race between the two models

Sample

Luckily we have another (independent) random sample from the same DGP

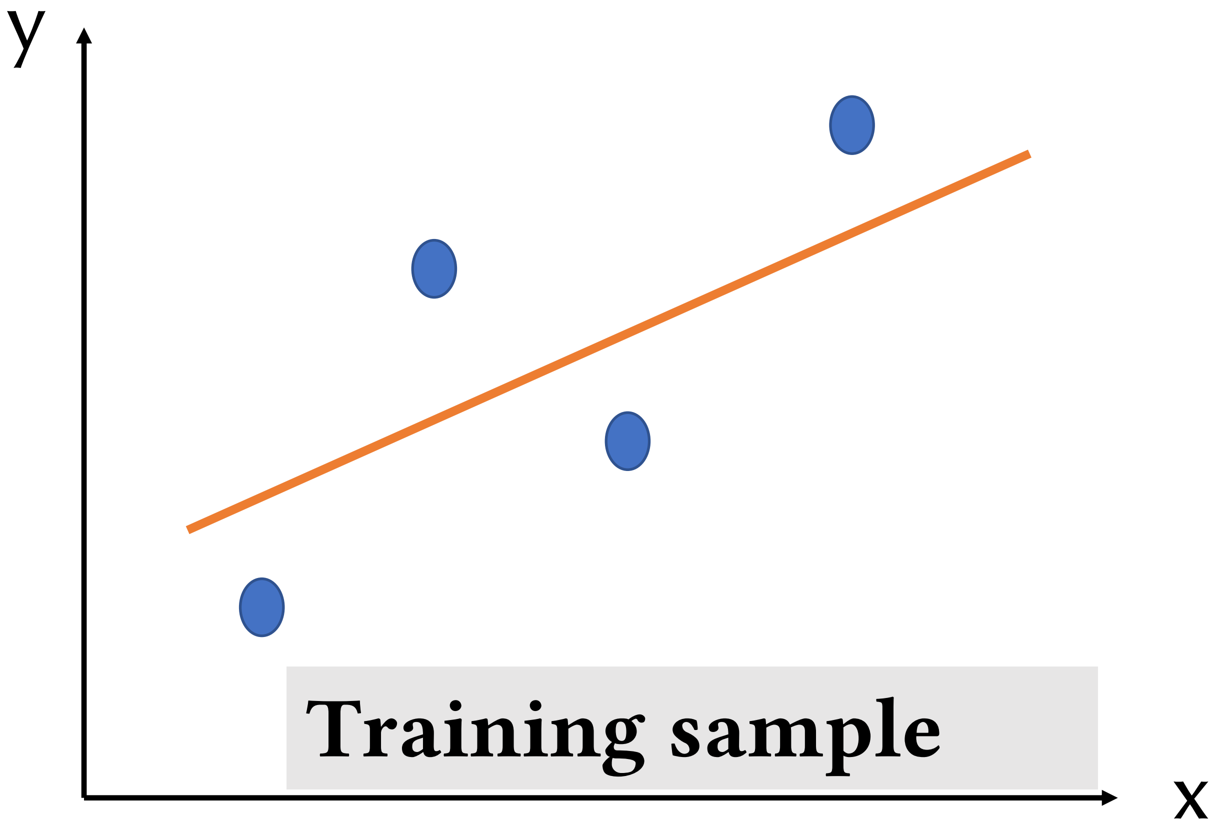


So we can test which candidate is better at predicting! Horse race between the two models

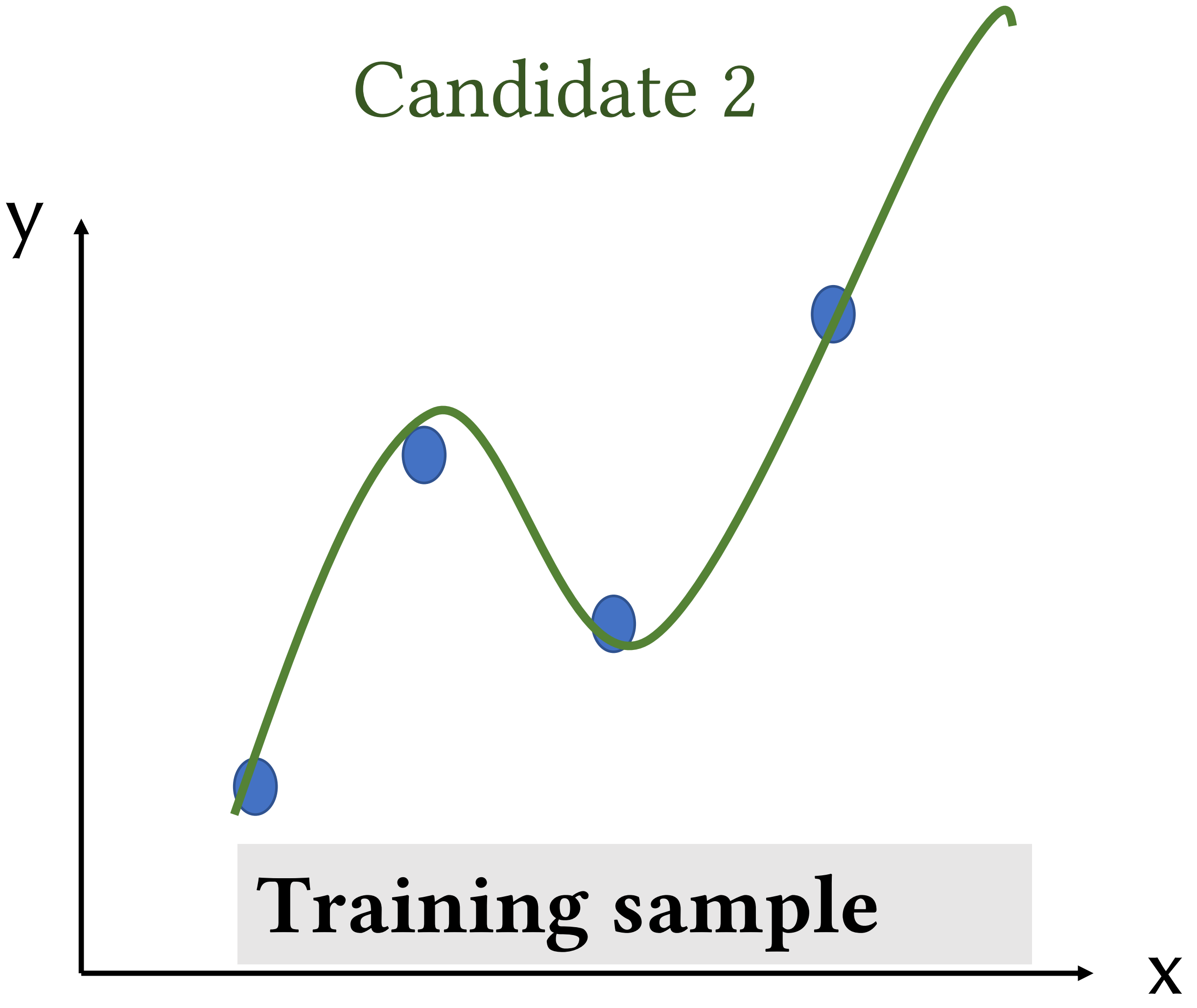
Sample

Which one is better?

Candidate 1

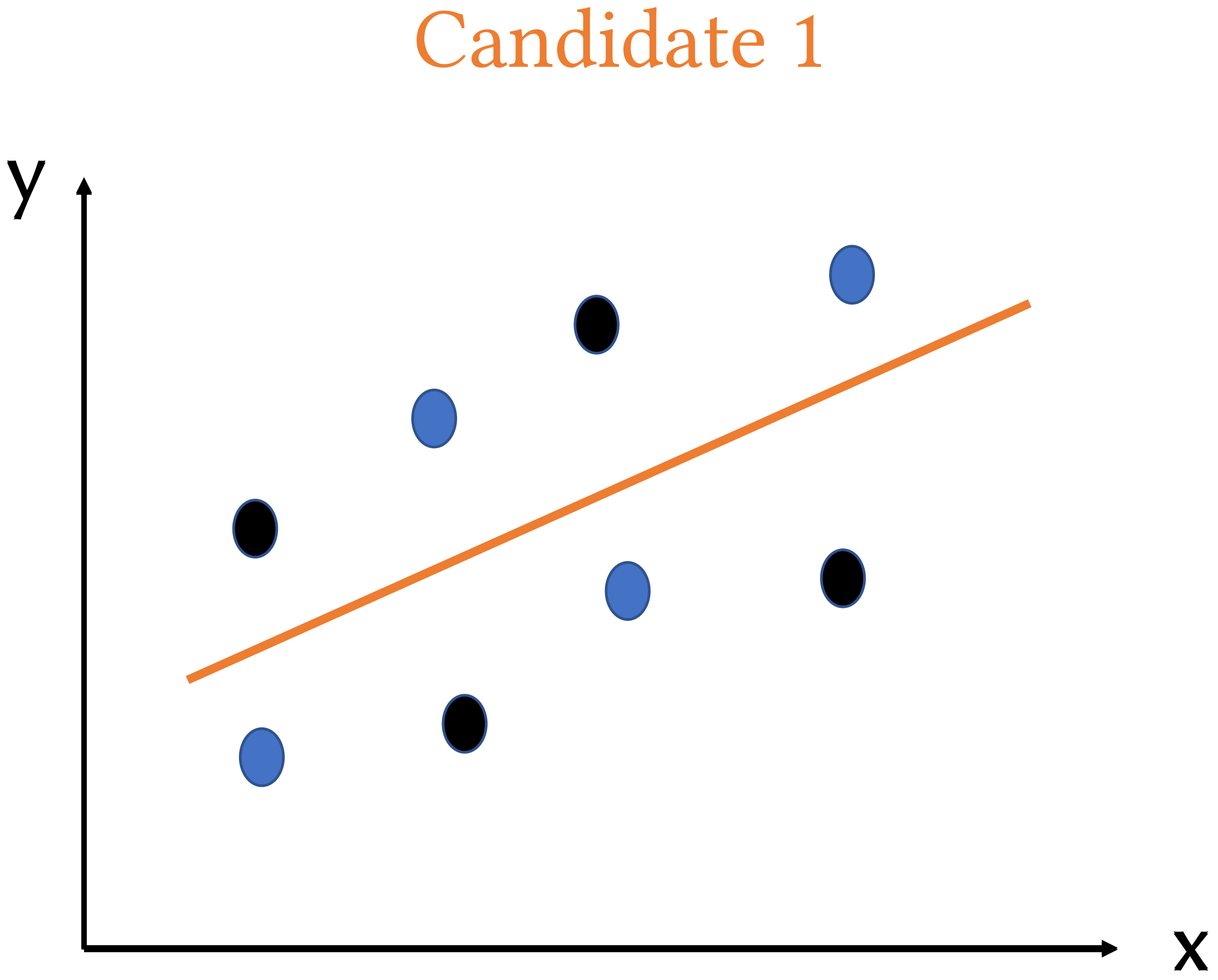


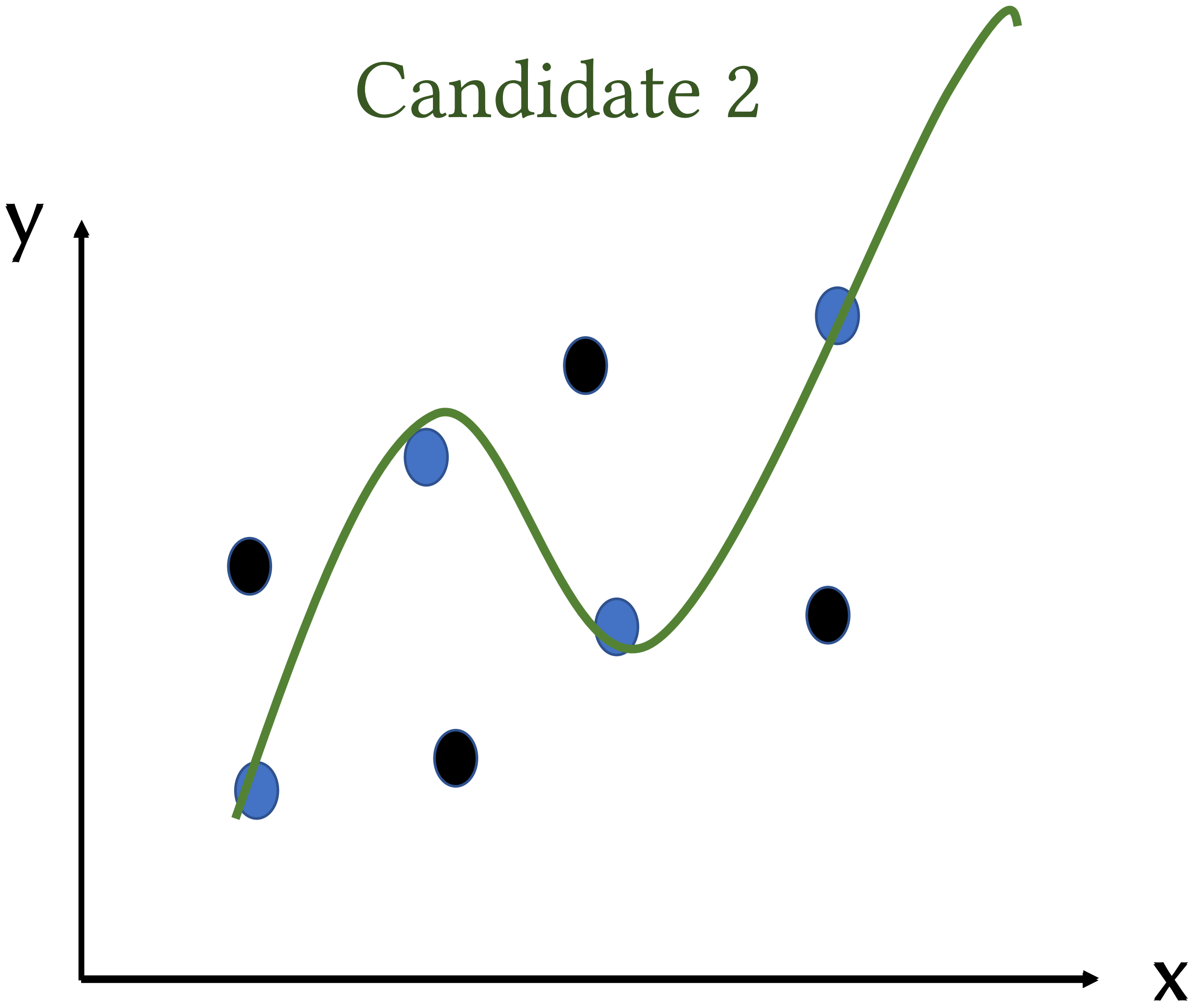
Candidate 2



Sample

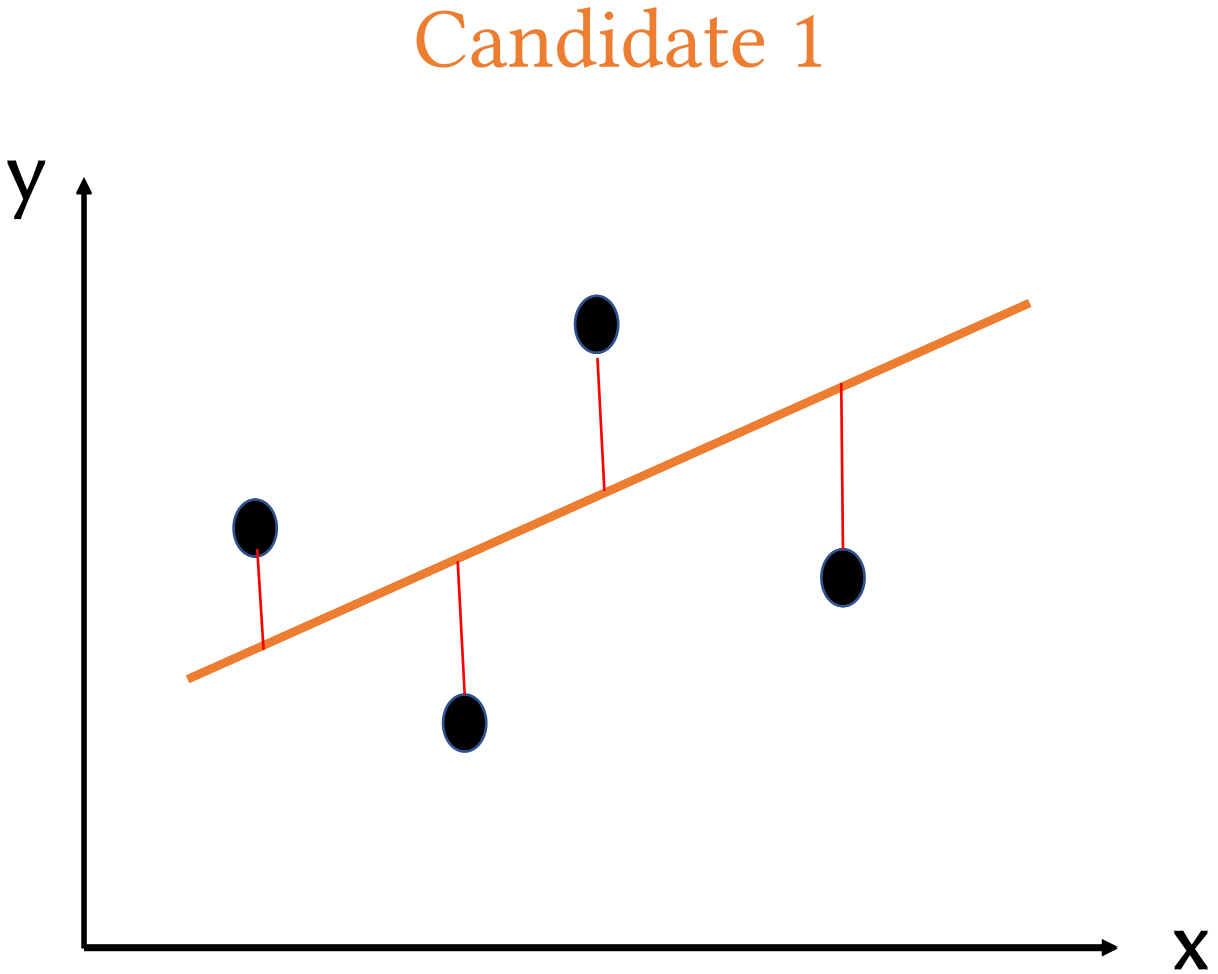
Which one is better?

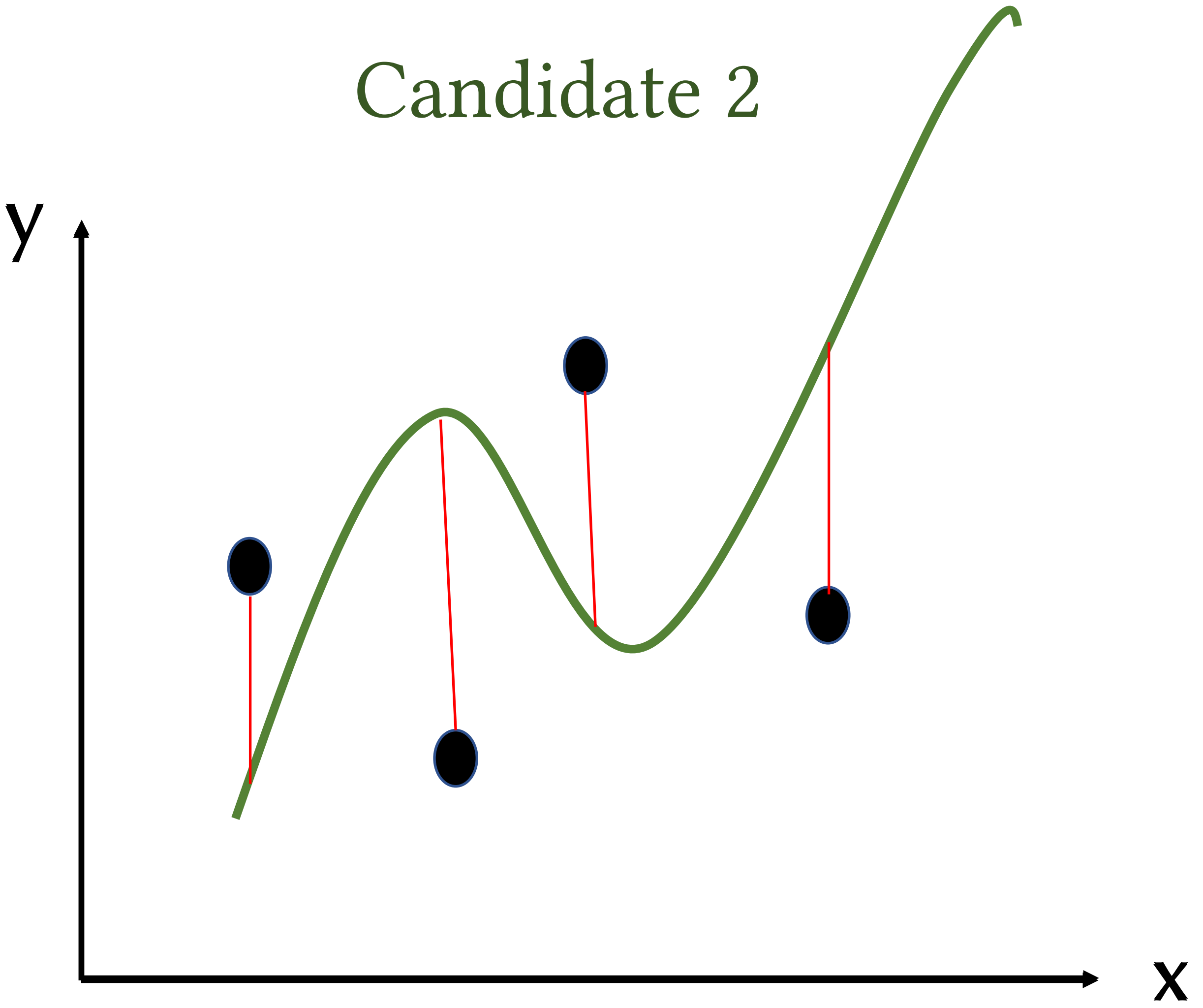




Sample

Which one is better?

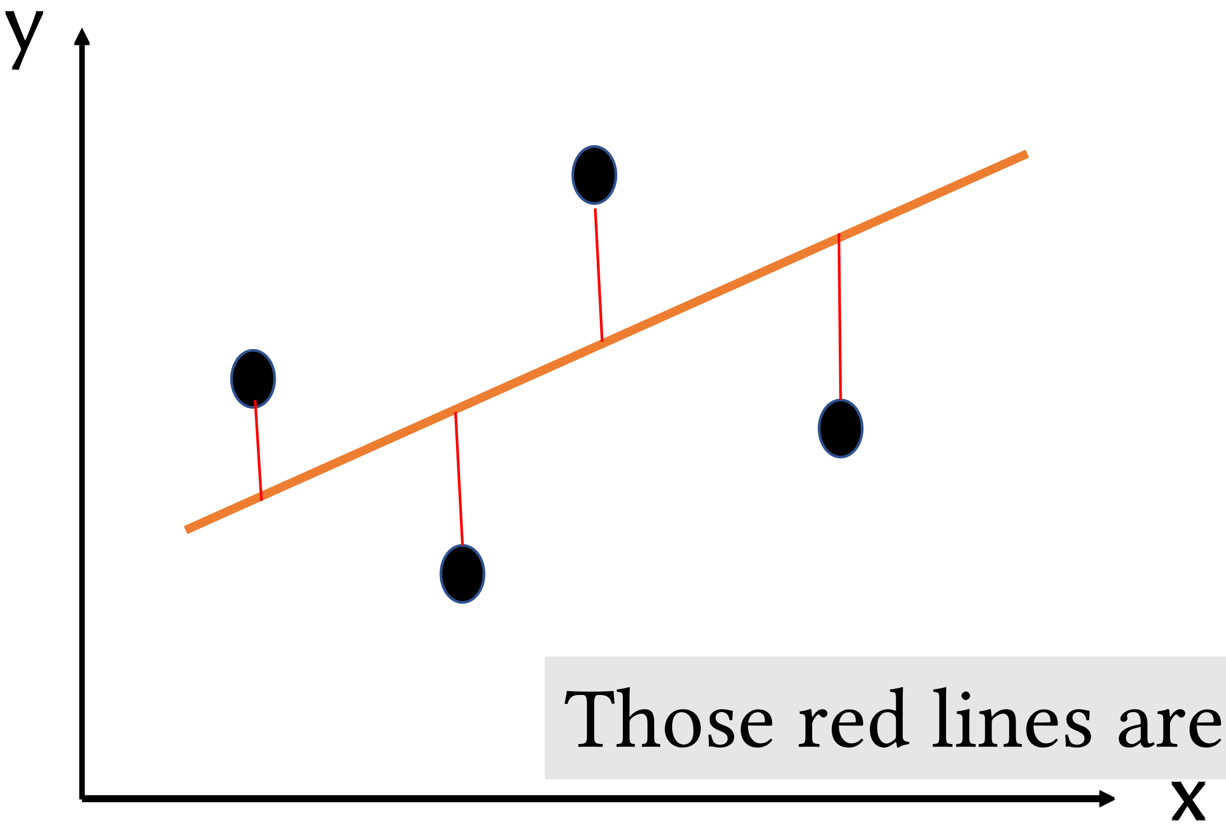




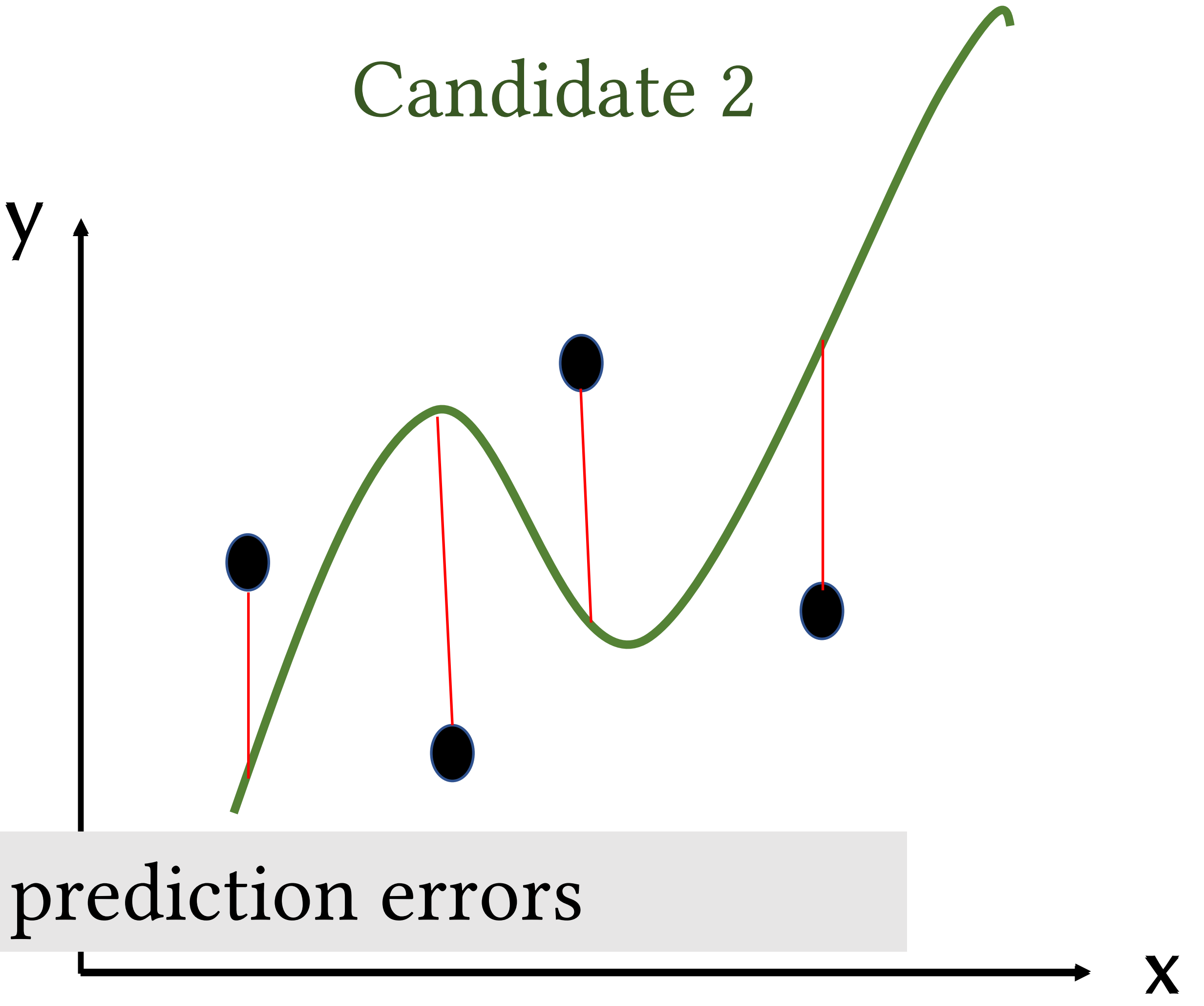
Sample

Which one is better?

Candidate 1



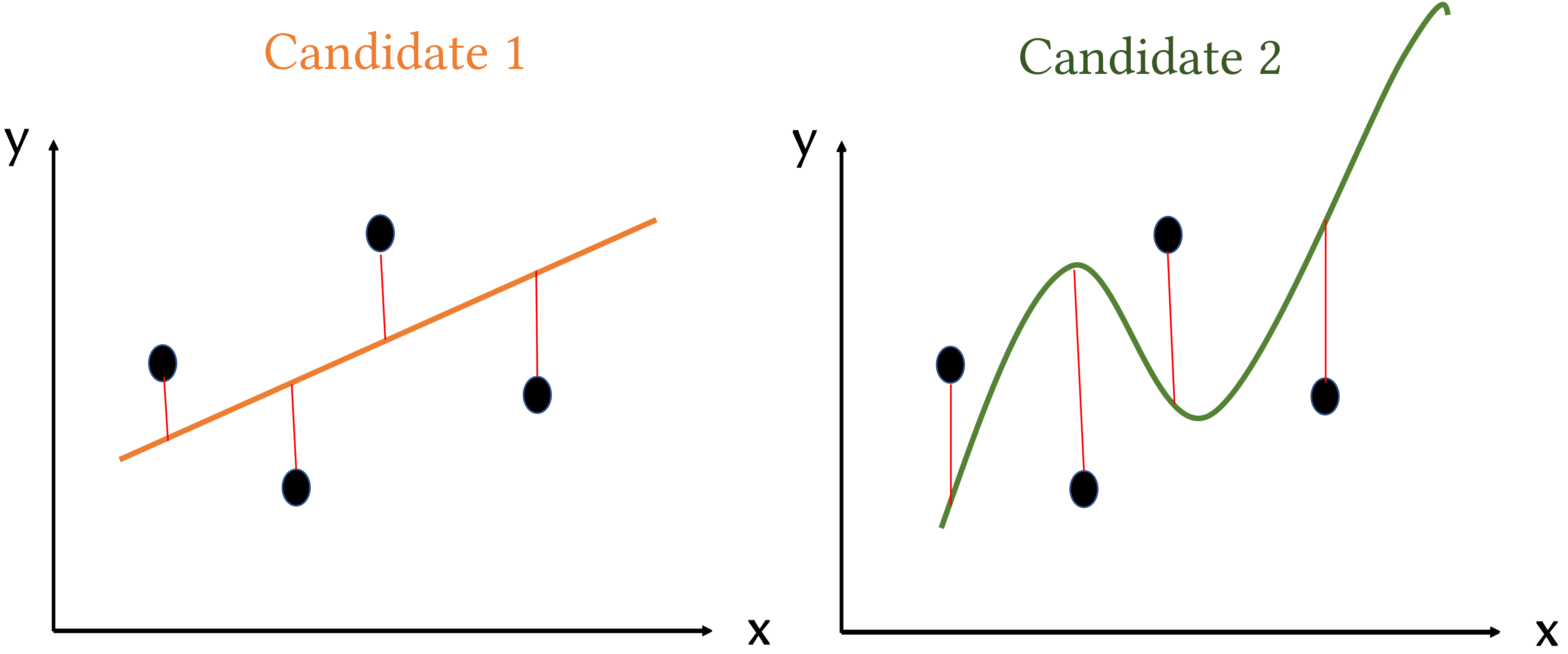
Candidate 2



Those red lines are prediction errors

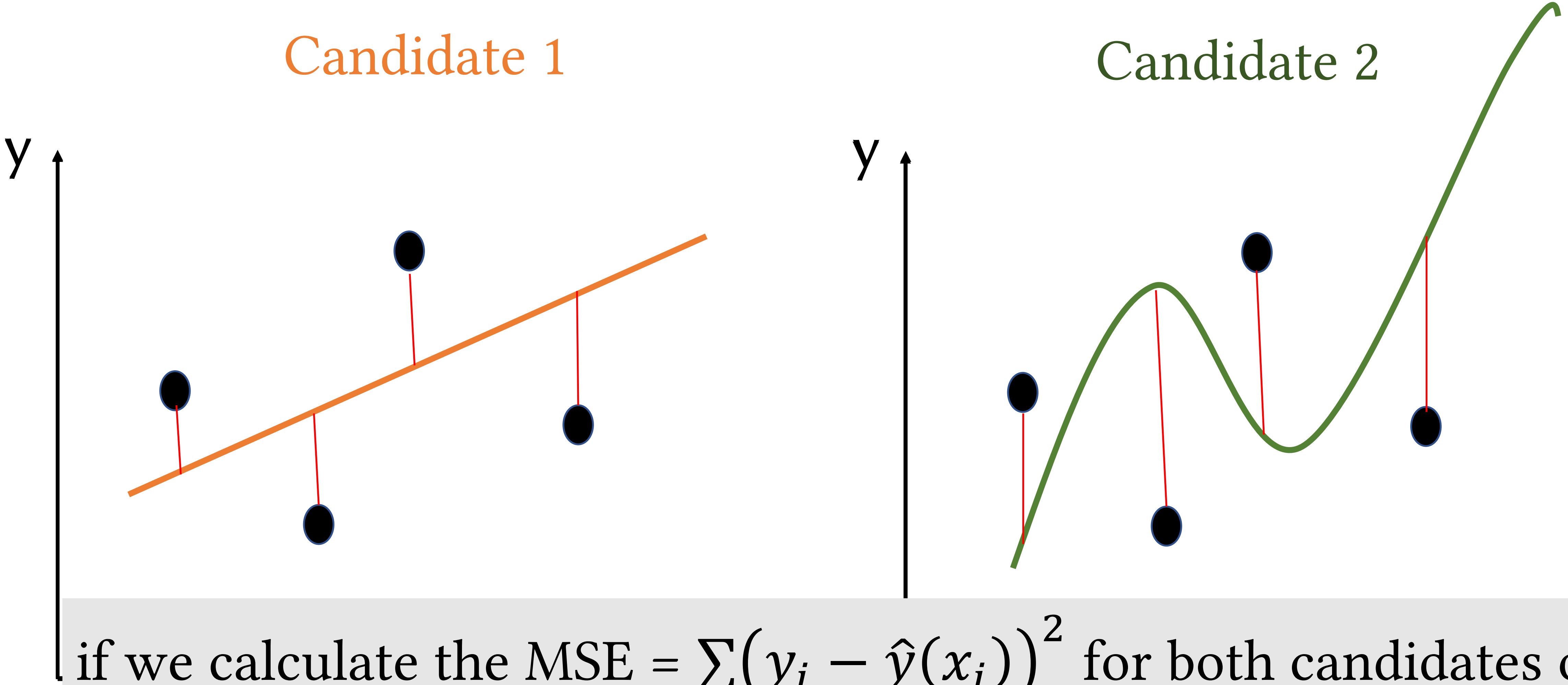
Sample

Clearly Candidate 1 is better than Candidate 2



Sample

Clearly Candidate 1 is better than Candidate 2

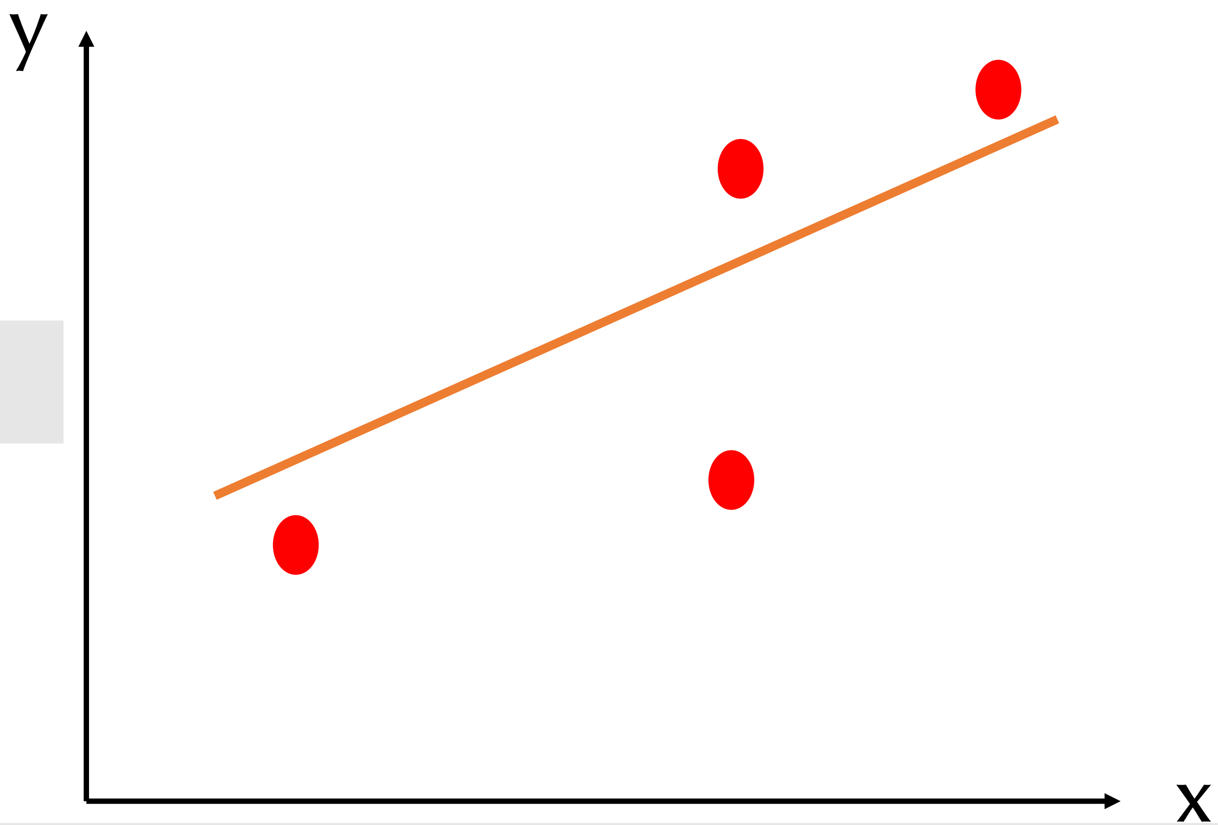


if we calculate the MSE = $\sum\left(y_i - \hat{y}(x_i)\right)^2$ for both candidates on the new sample, Candidate 1 will have lower MSE

Sample

What are the expected prediction errors from Candidate 1?

Test sample

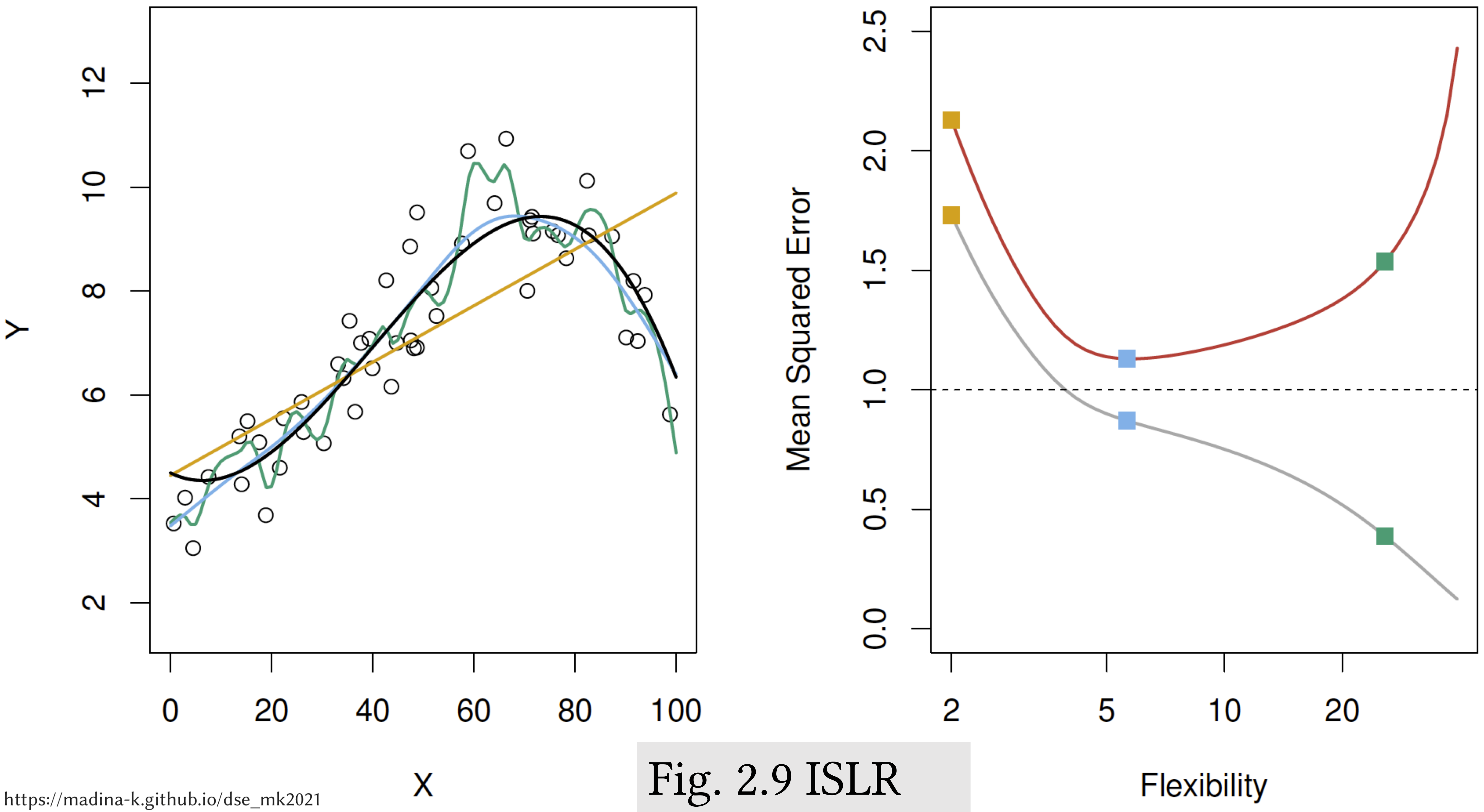


For this we need yet another sample (so far untouched by training or validation)

In general, all prediction tasks have the same basic steps:

1. Train a **less flexible model** on a **training sample**
2. Train a **more flexible model** on the training sample
3. Compare MSE of model 1 and 2 on the **validation sample** and choose the one with the smallest MSE
4. Calculate MSE of the chosen model on the **test sample**. $\sqrt{\text{MSE}}$ is the expected prediction error of your best prediction model.

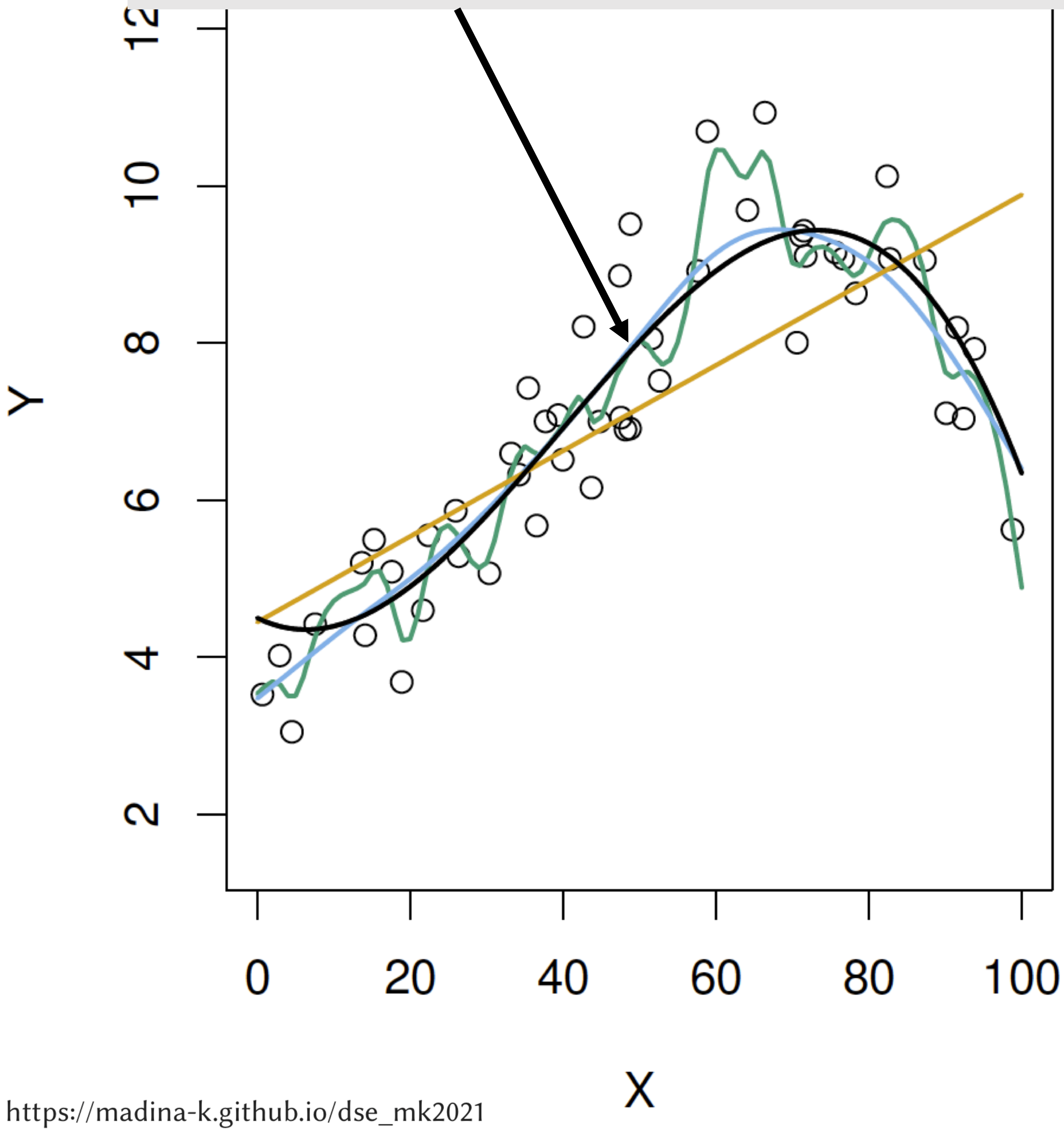
Do prediction errors usually increase or decrease with more flexible models?

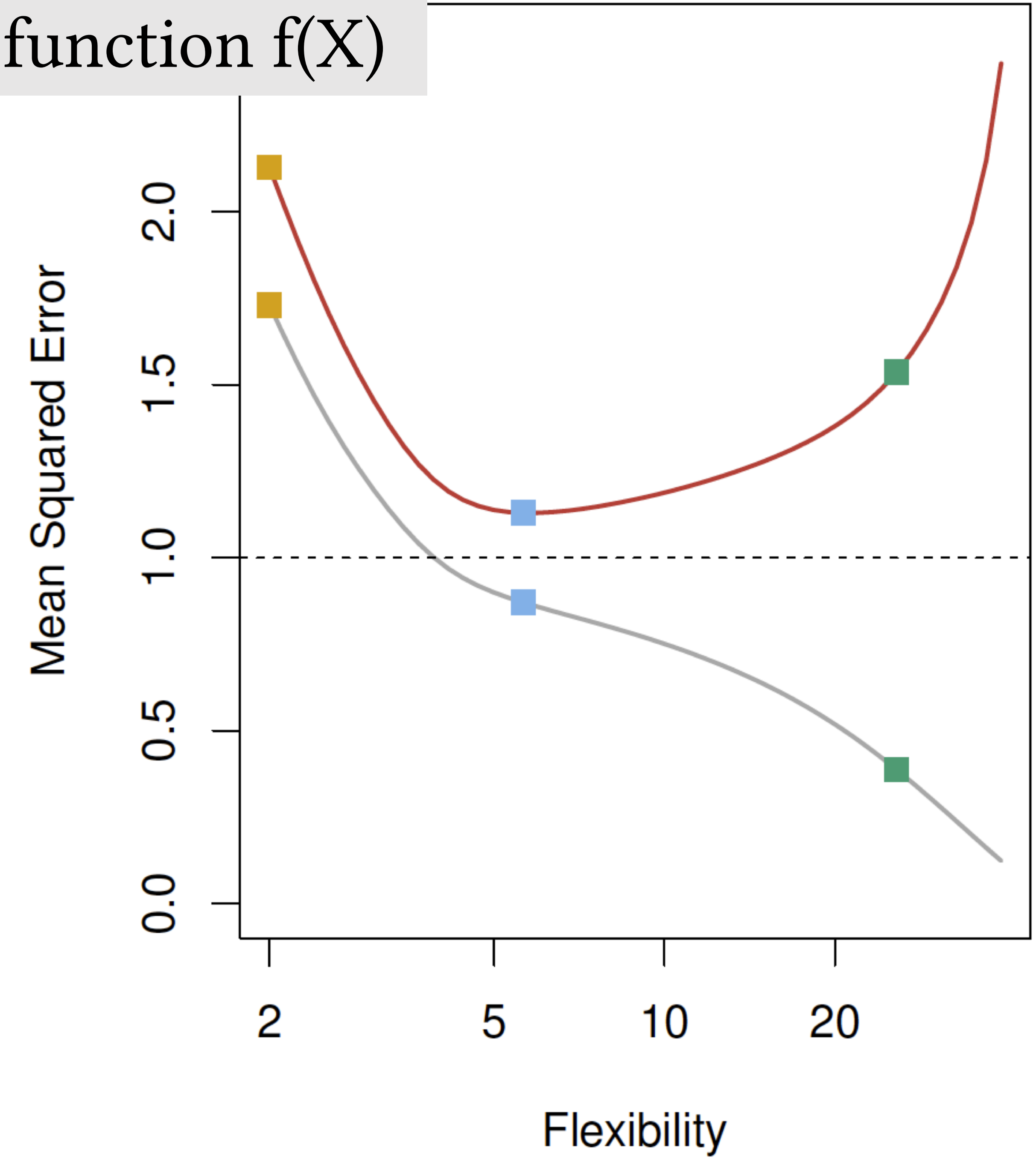


12
10
8
6
4
2
Y
0
20
40
60
80
100
X
2.5
2.0
1.5
1.0
0.5
0.0
Mean Squared Error
2
5
10
20
Flexibility

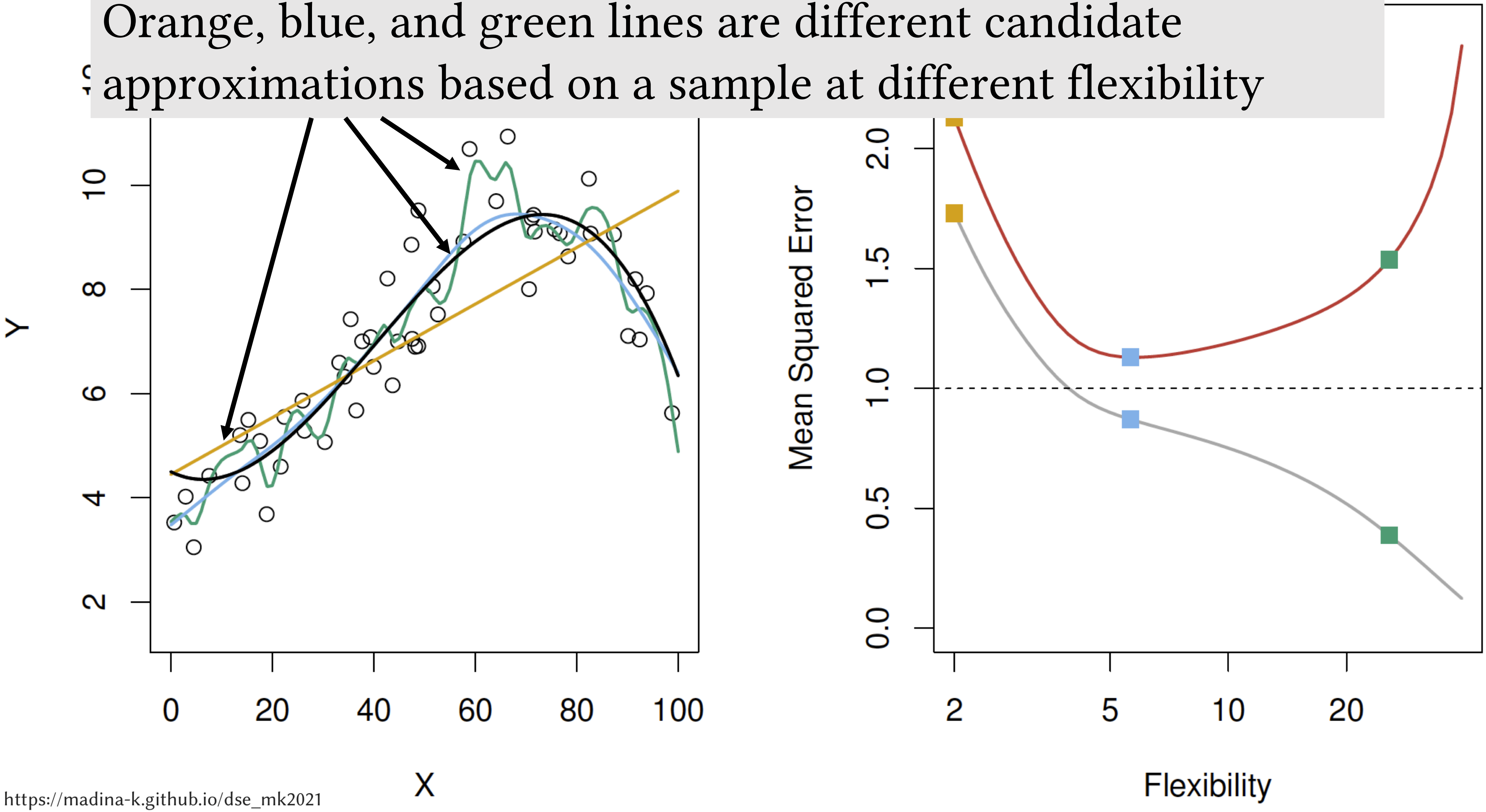
Fig. 2.9 ISLR

Black line is the true population function f(X)

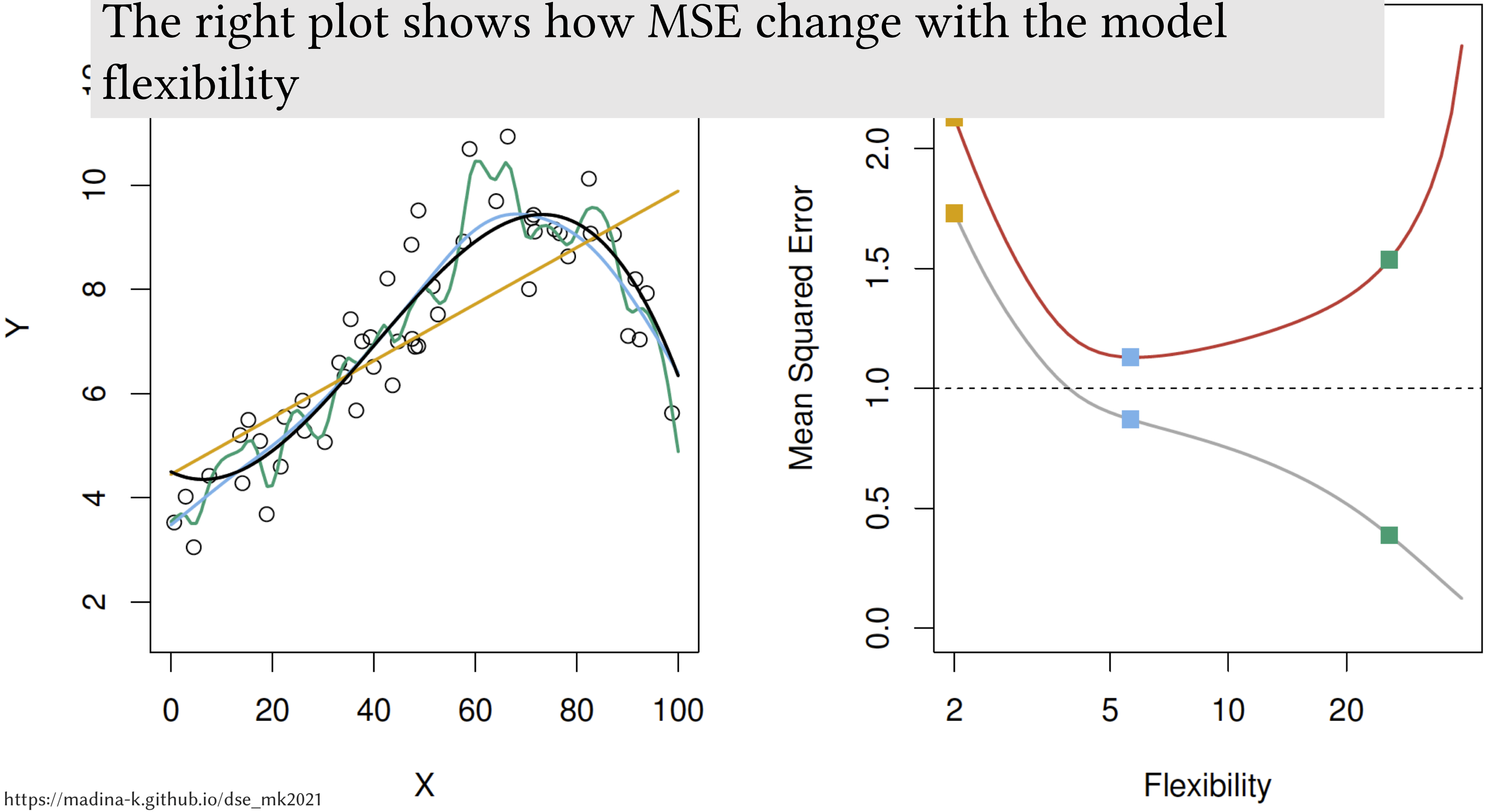




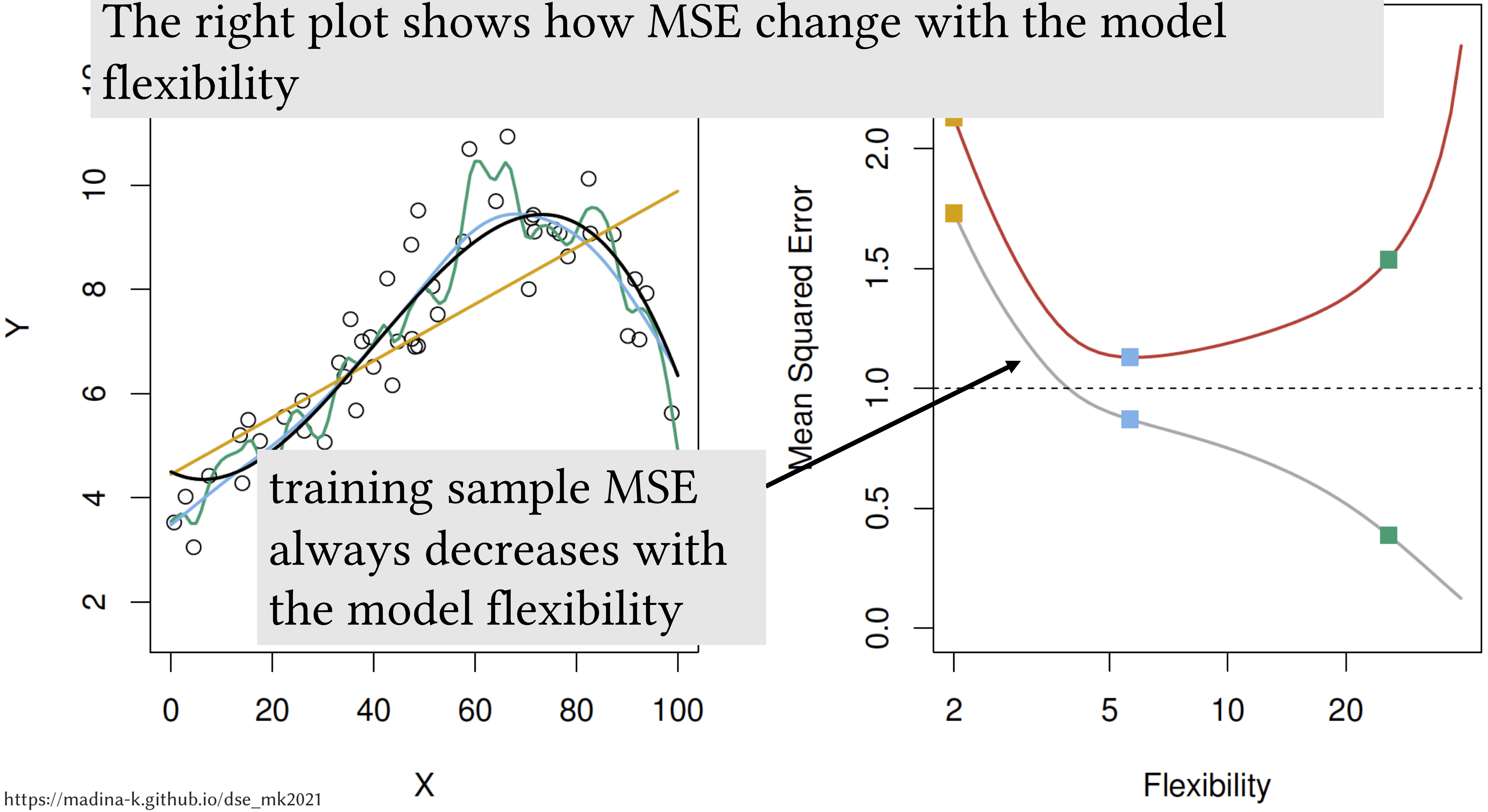
Orange, blue, and green lines are different candidate approximations based on a sample at different flexibility



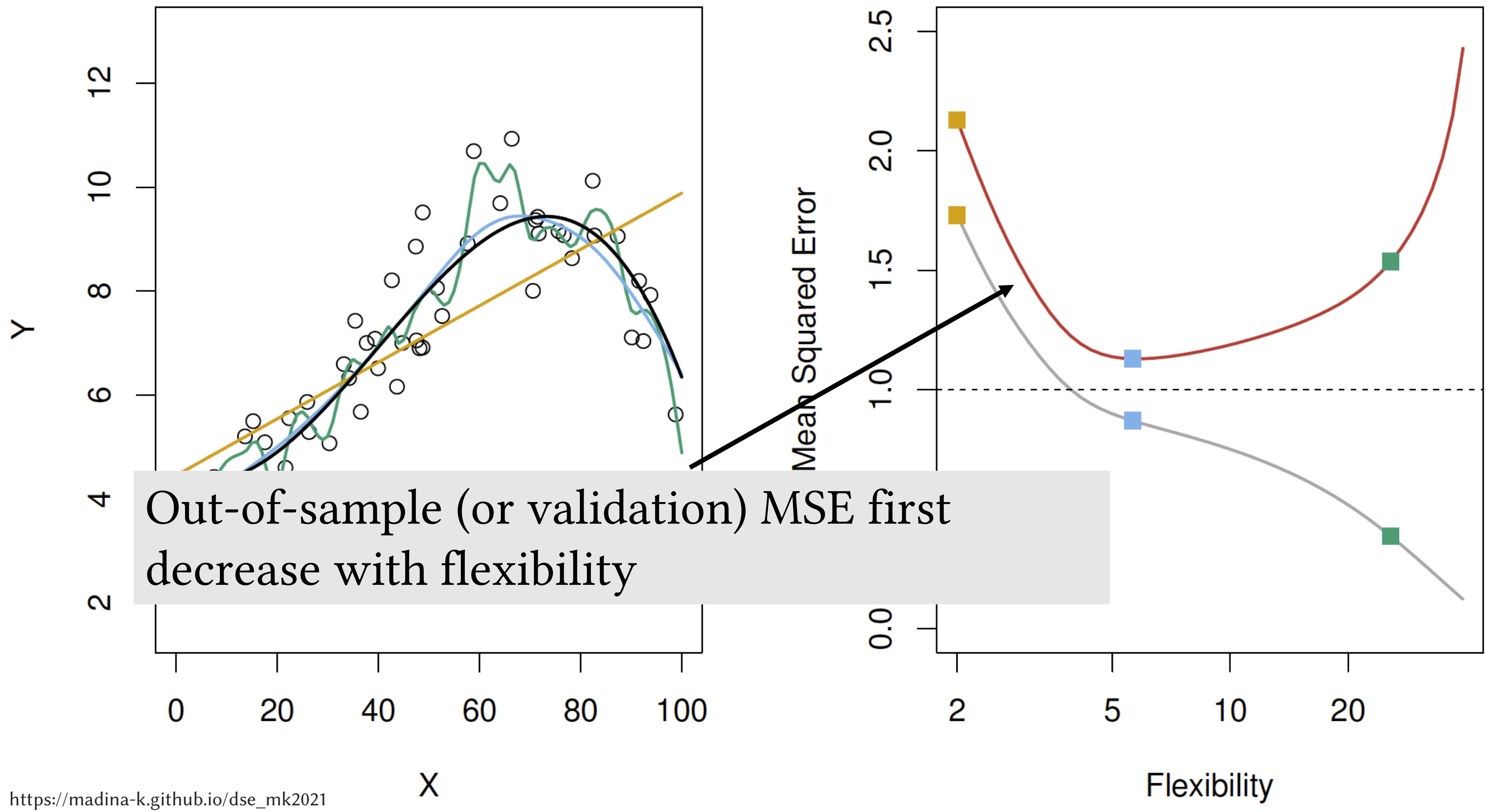
The right plot shows how MSE change with the model flexibility

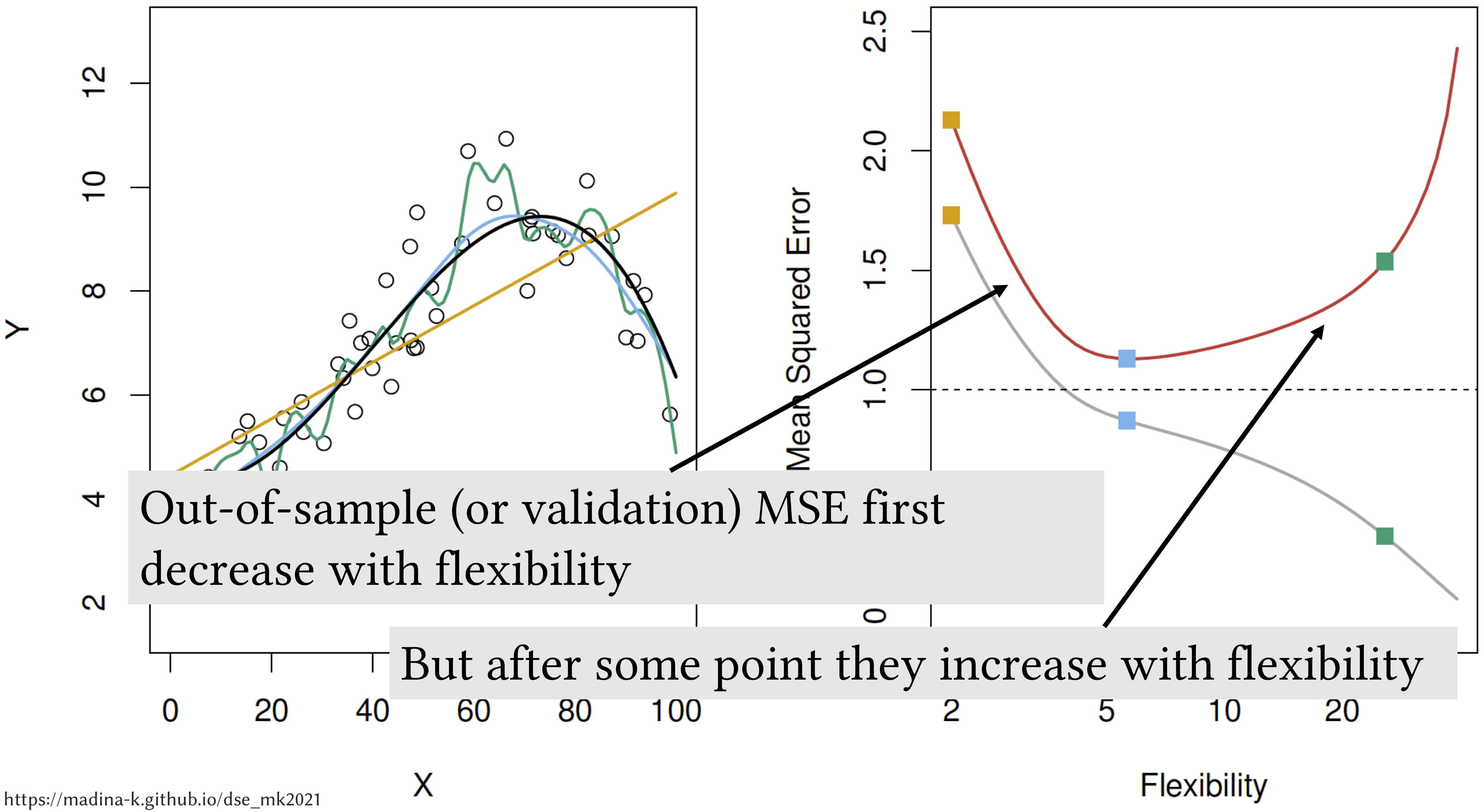


The right plot shows how MSE change with the model flexibility



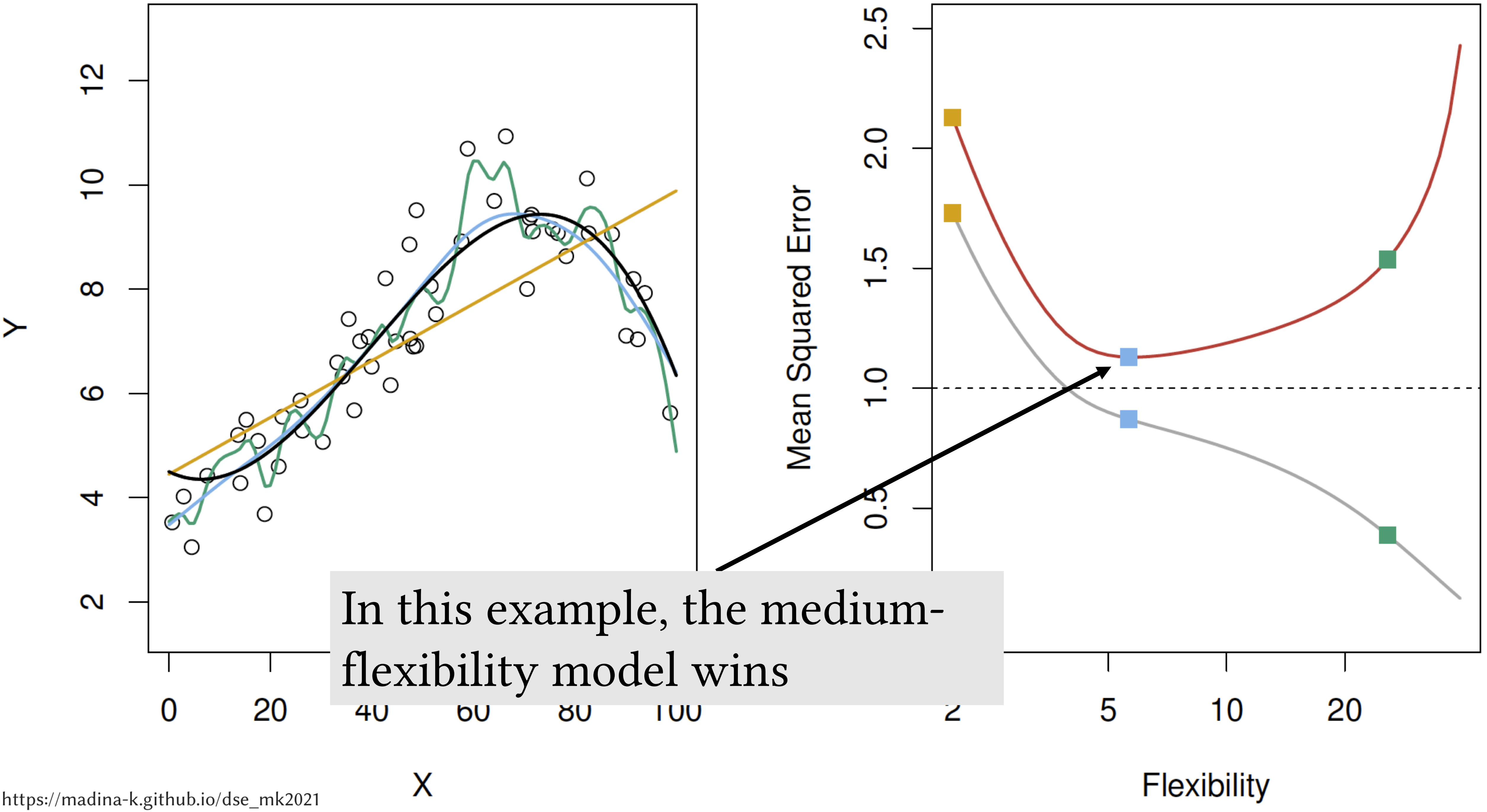
training sample MSE always decreases with the model flexibility


Y
X
Mean Squared Error
Flexibility
Out-of-sample (or validation) MSE first decrease with flexibility



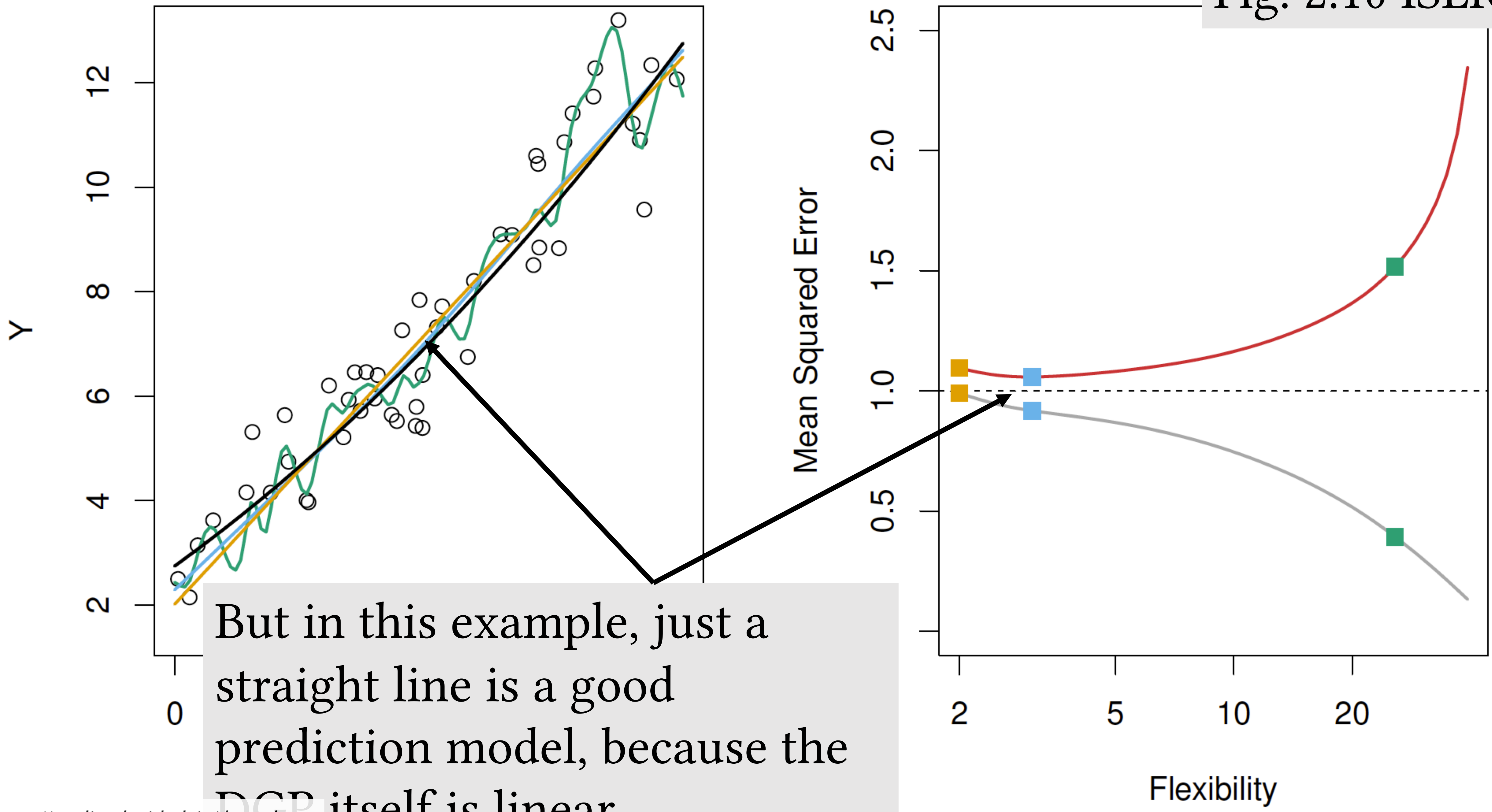
Out-of-sample (or validation) MSE first decrease with flexibility

But after some point they increase with flexibility



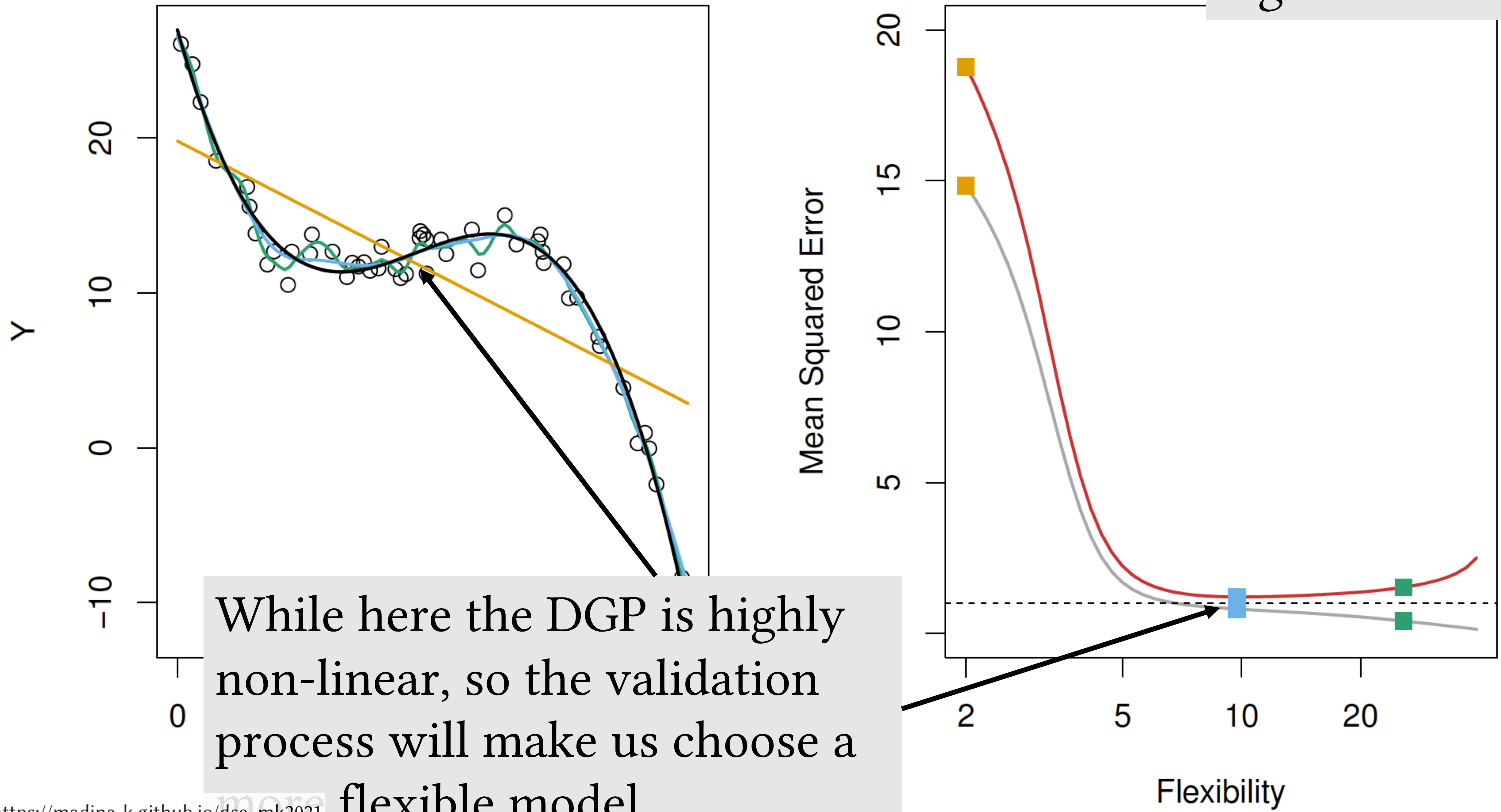
In this example, the medium-flexibility model wins

Fig. 2.10 ISLR



But in this example, just a straight line is a good prediction model, because the DGP itself is linear

Fig. 2.11 ISLR



While here the DGP is highly non-linear, so the validation process will make us choose a more flexible model

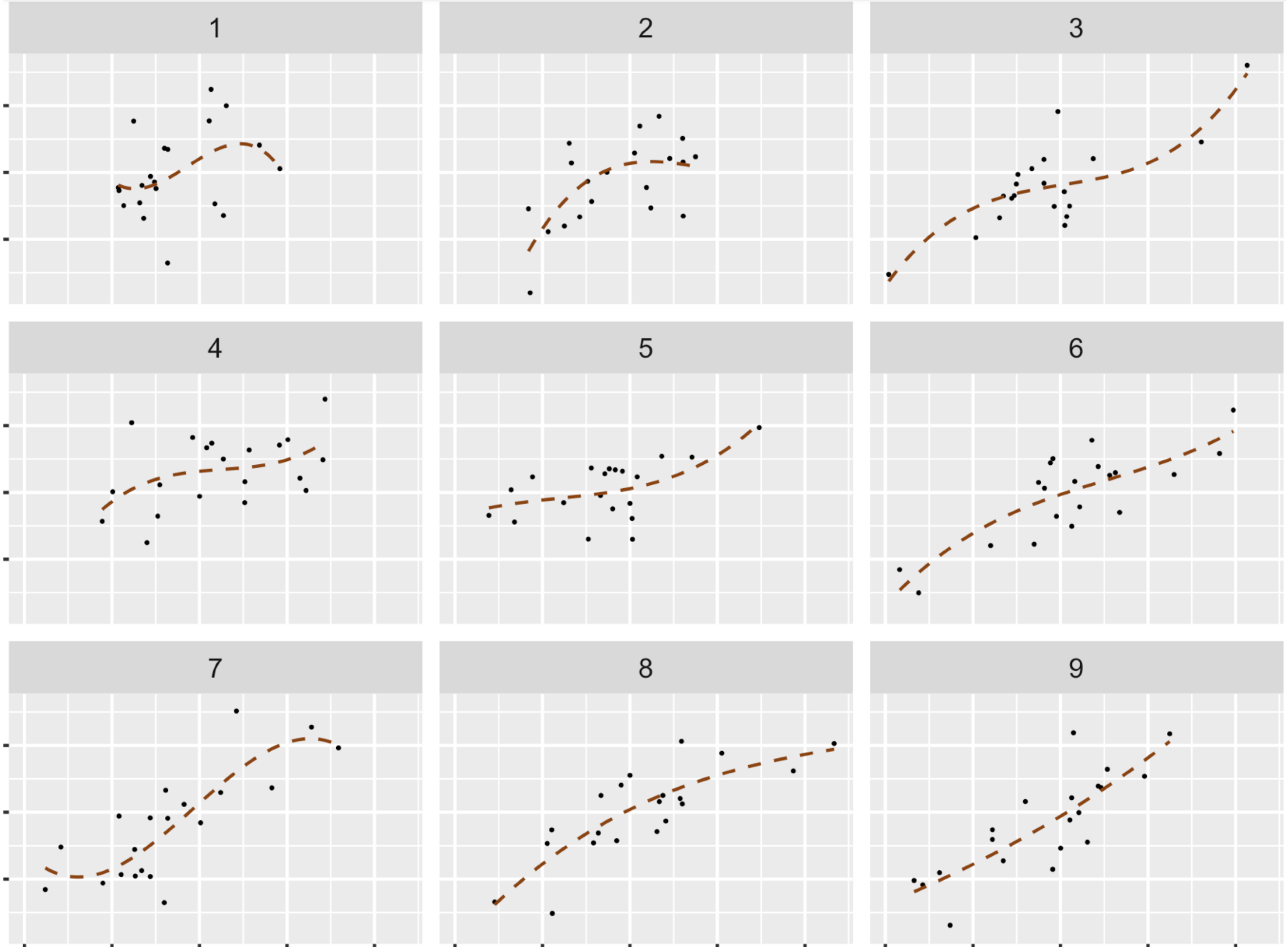
Why is out-of-sample MSE U-shaped?

Why does the out-of-sample MSE have a U shape?

Because of bias-variance trade-off

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} - [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

The same polynomial function fit in different samples will have slightly different shape each time



https://madina-k.github.io/dse_mk2021

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} + \underbrace{[Bias(\hat{f}(x_0))]^2}_{\text{errors from simplifying a complex DGP with a simple model}} + Var(\epsilon)$$

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} + \underbrace{[Bias(\hat{f}(x_0))]^2}_{\text{errors from simplifying a complex DGP with a simple model}} + \underbrace{Var(\epsilon)}_{\text{Irreducible error}}$$

(We cannot do anything about it)

Model flexibility affects variance and bias

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} + \underbrace{[Bias(\hat{f}(x_0))]^2}_{\text{errors from simplifying a complex DGP with a simple model}} + \underbrace{Var(\epsilon)}_{\text{Irreducible error}}$$

$$\underbrace{E(y_0 - \hat{f}(x_0))^2}_{\text{expected out-of-sample MSE}} = \underbrace{Var(\hat{f}(x_0))}_{\text{the degree to which the shape of the model varies in different samples}} + \underbrace{[Bias(\hat{f}(x_0))]^2}_{\text{errors from simplifying a complex DGP with a simple model}} + \underbrace{Var(\epsilon)}_{\text{Irreducible error}}$$

More flexible models will have higher variance, but a lower bias

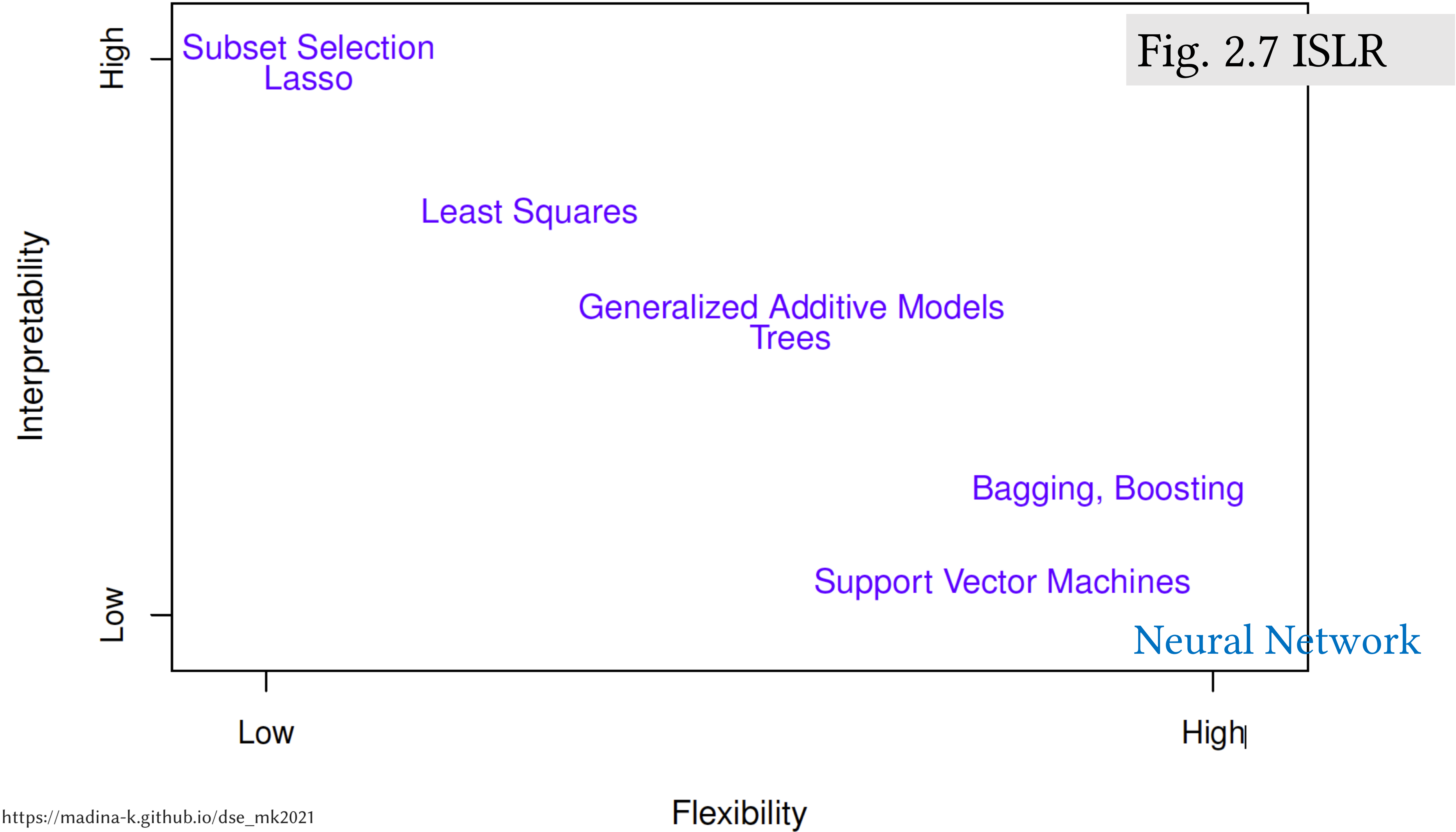


Fig. 2.7 ISLR
Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
Neural Network
Interpretability
High
Low
Flexibility
Low
High

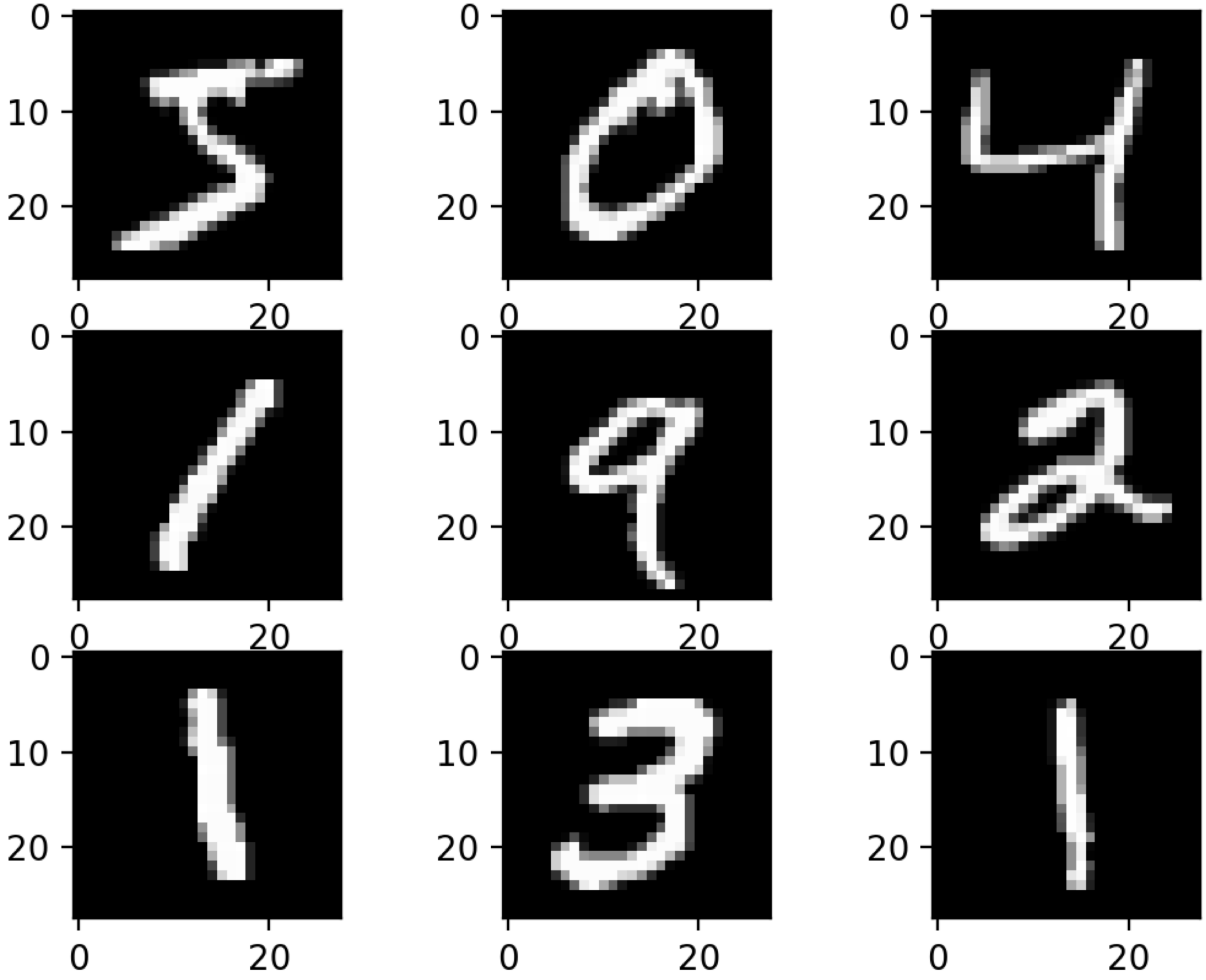
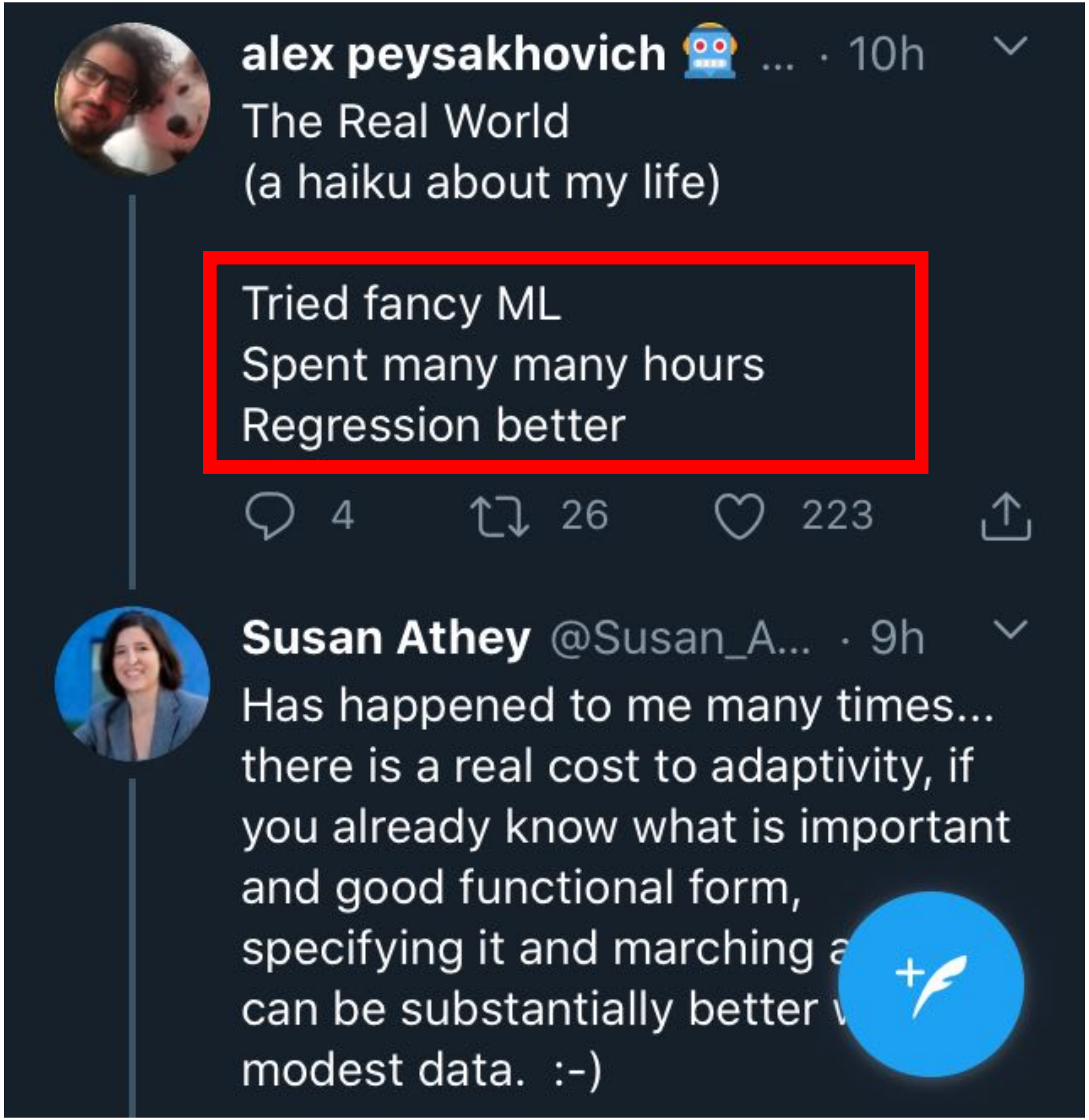
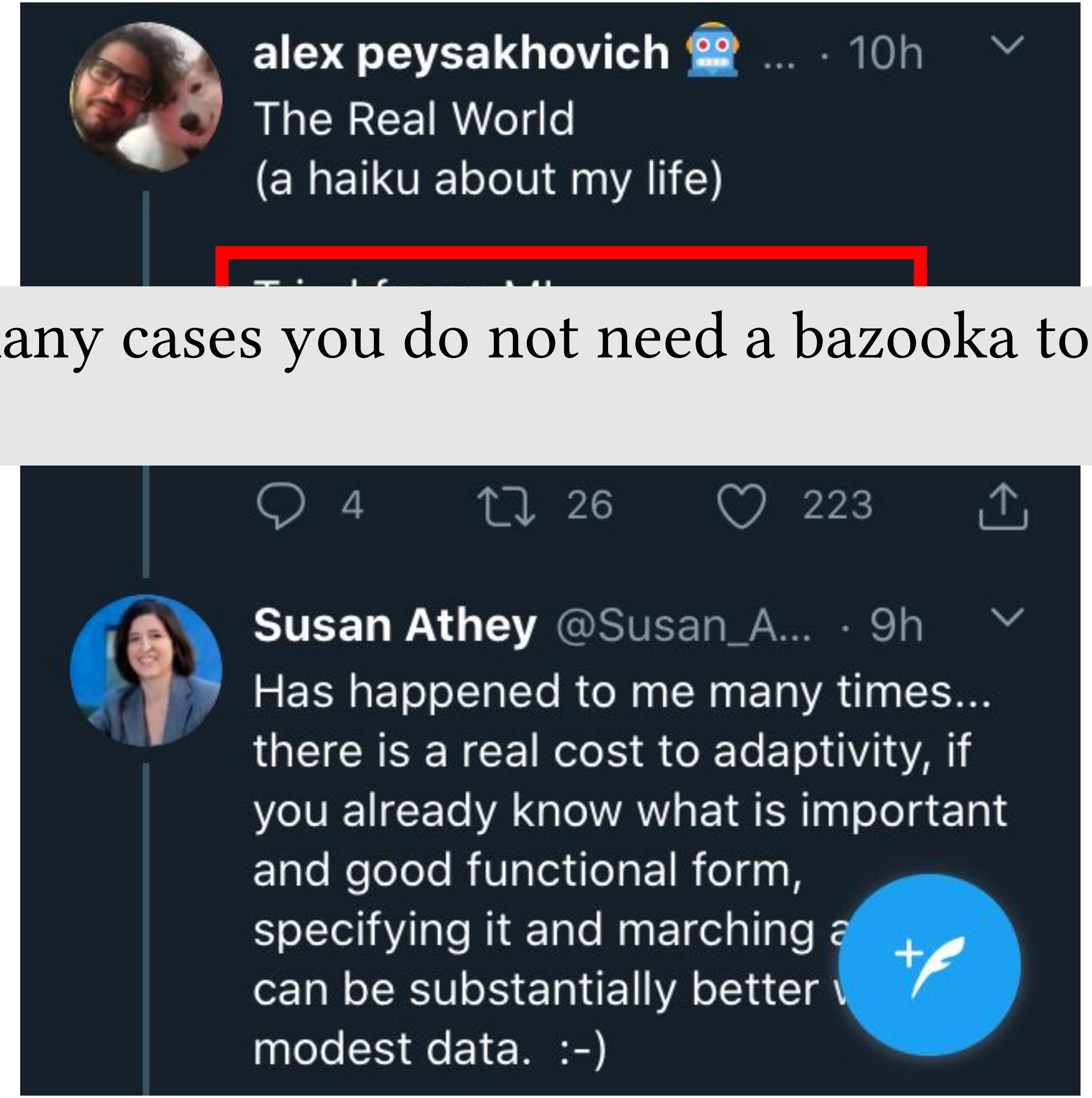


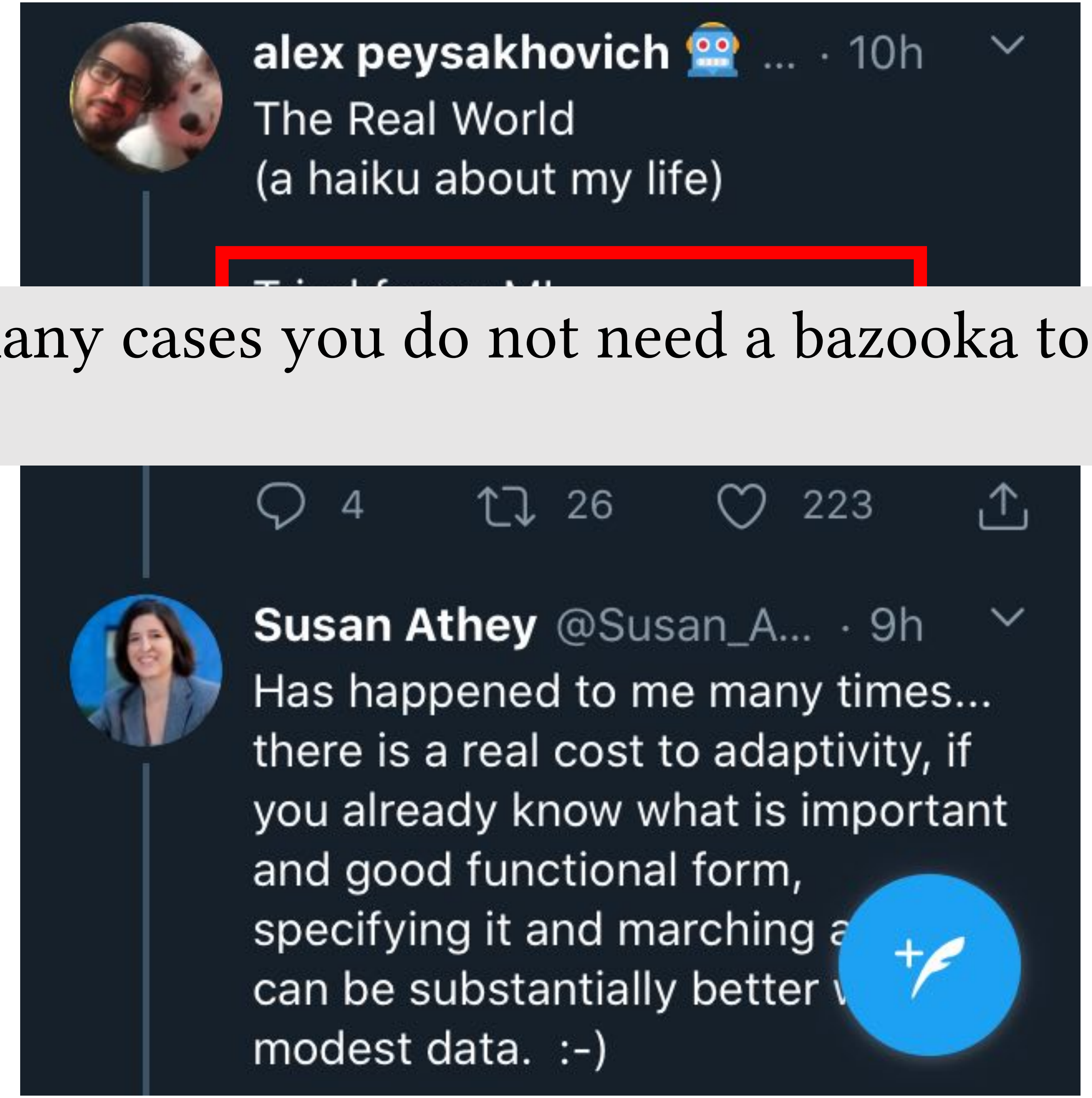
Image recognition is an example of a highly non-linear DGP. That's why more flexible models shine for this task

https://madina-k.github.io/dse_mk2021


alex peysakhovich ... · 10h
The Real World
(a haiku about my life)
Tried fancy ML
Spent many many hours
Regression better
4
26
223
Susan Athey @Susan_A... · 9h
Has happened to me many times...
there is a real cost to adaptivity, if
you already know what is important
and good functional form,
specifying it and marching
can be substantially better
modest data. :-)



i.e., in many cases you do not need a bazooka to kill a mouse



i.e., in many cases you do not need a bazooka to kill a mouse

Disclaimer: no animals were harmed in the making of this video