Self-Driving Cars & Forest Ecology: Modeling for Machine Learning

Albert Y. Kim

Assistant Professor

[Statistical & Data Sciences](), Smith College

Sigma Xi, The Scientific Research Honor Society

Tuesday 2019/12/12

What variables are being collected?

Road Map

Machine Learning:
- Machine learning as modeling
- Example model: splines

Forest Ecology:
- Our data

Machine Learning + Forest Ecology:
- Goal restated
- Assessment model performance

Conclusion:
- Thoughts on modeling

Machine Learning

Prediction!

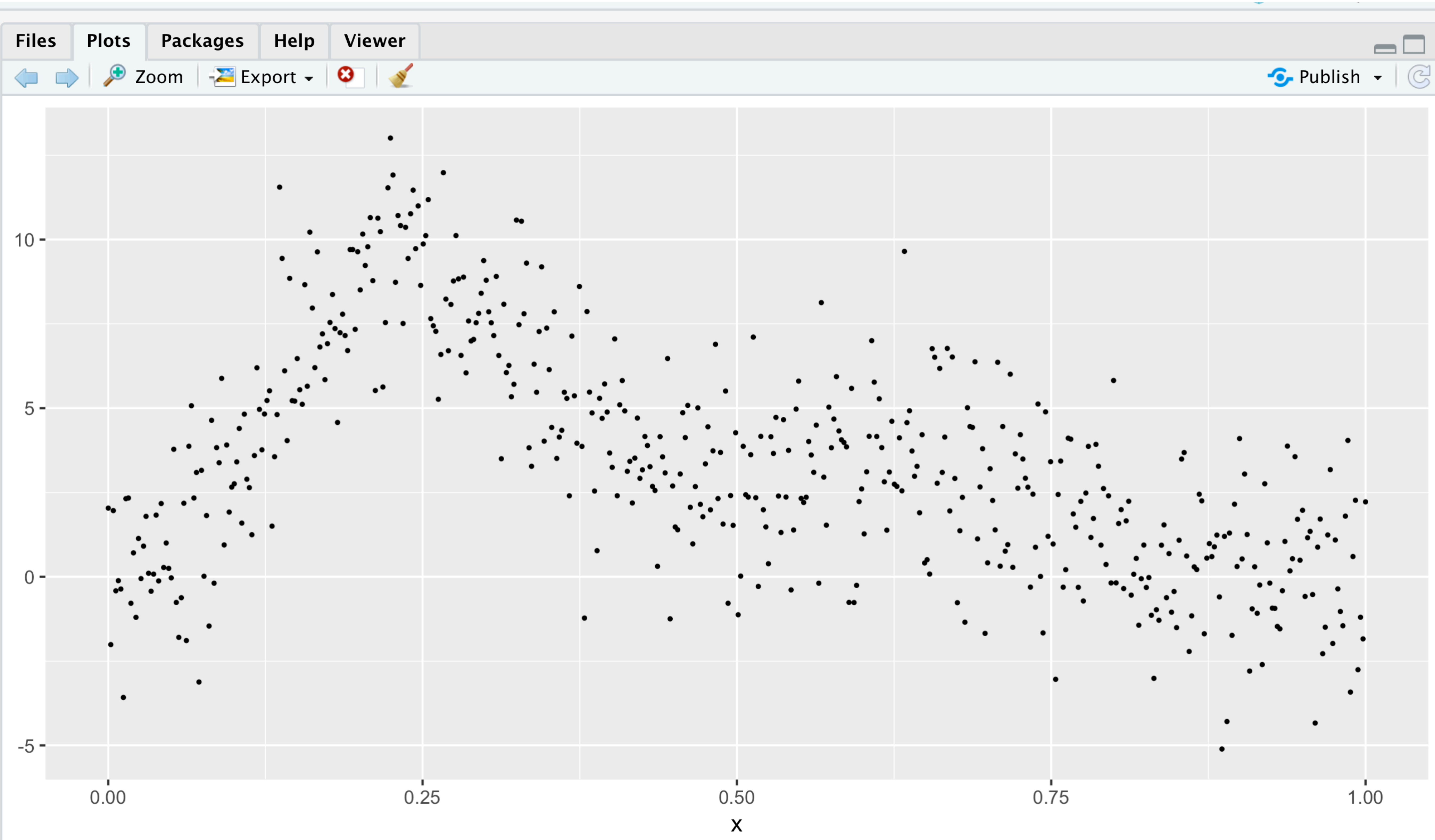
Machine Learning as Modeling

True (Unknown) Model: $y = f(\vec{x}) + \epsilon$

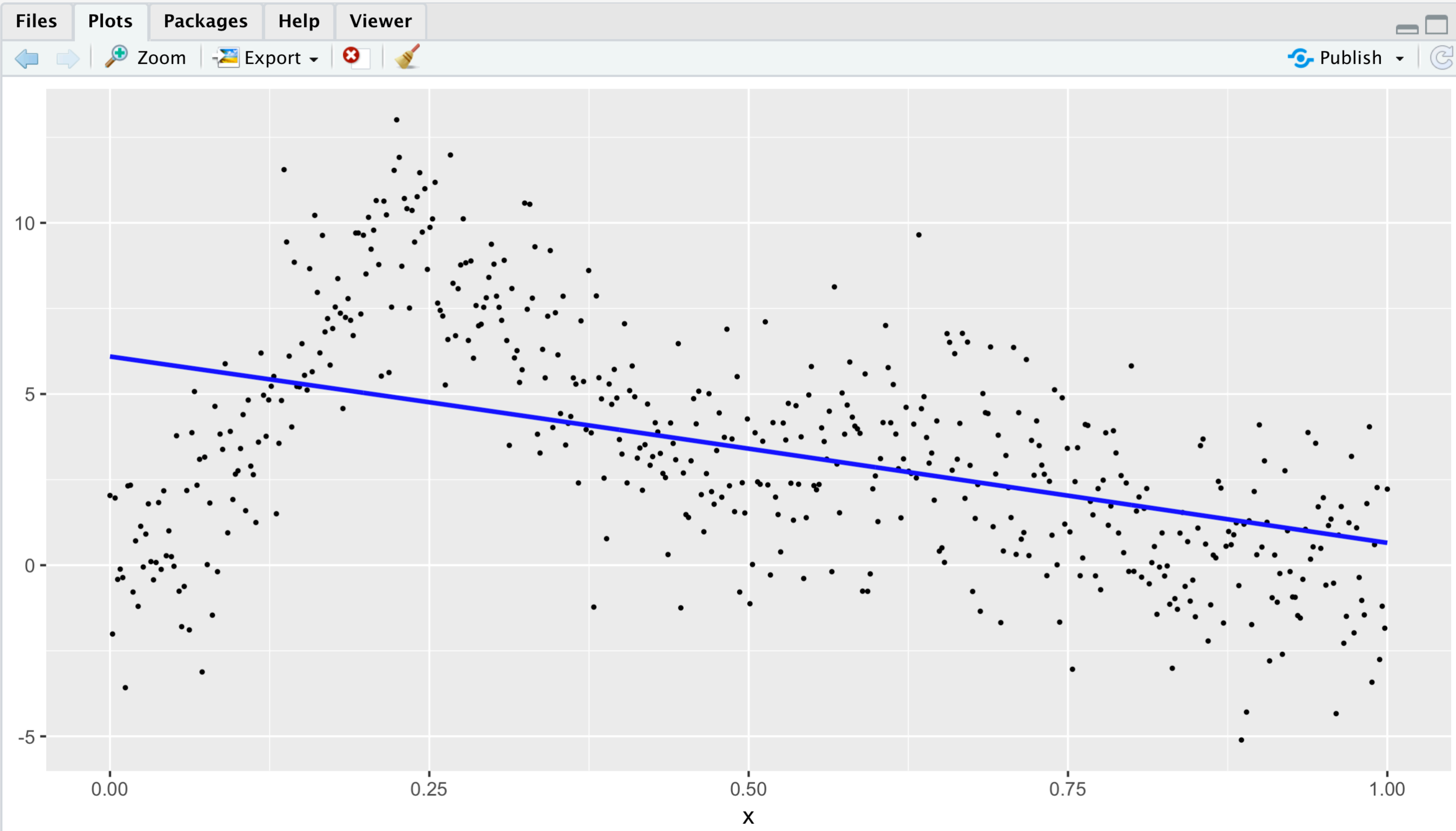
Approximated Model: $\hat{y} = \hat{f}(\vec{x})$

Now to the blackboard for Chalk Talk #1…

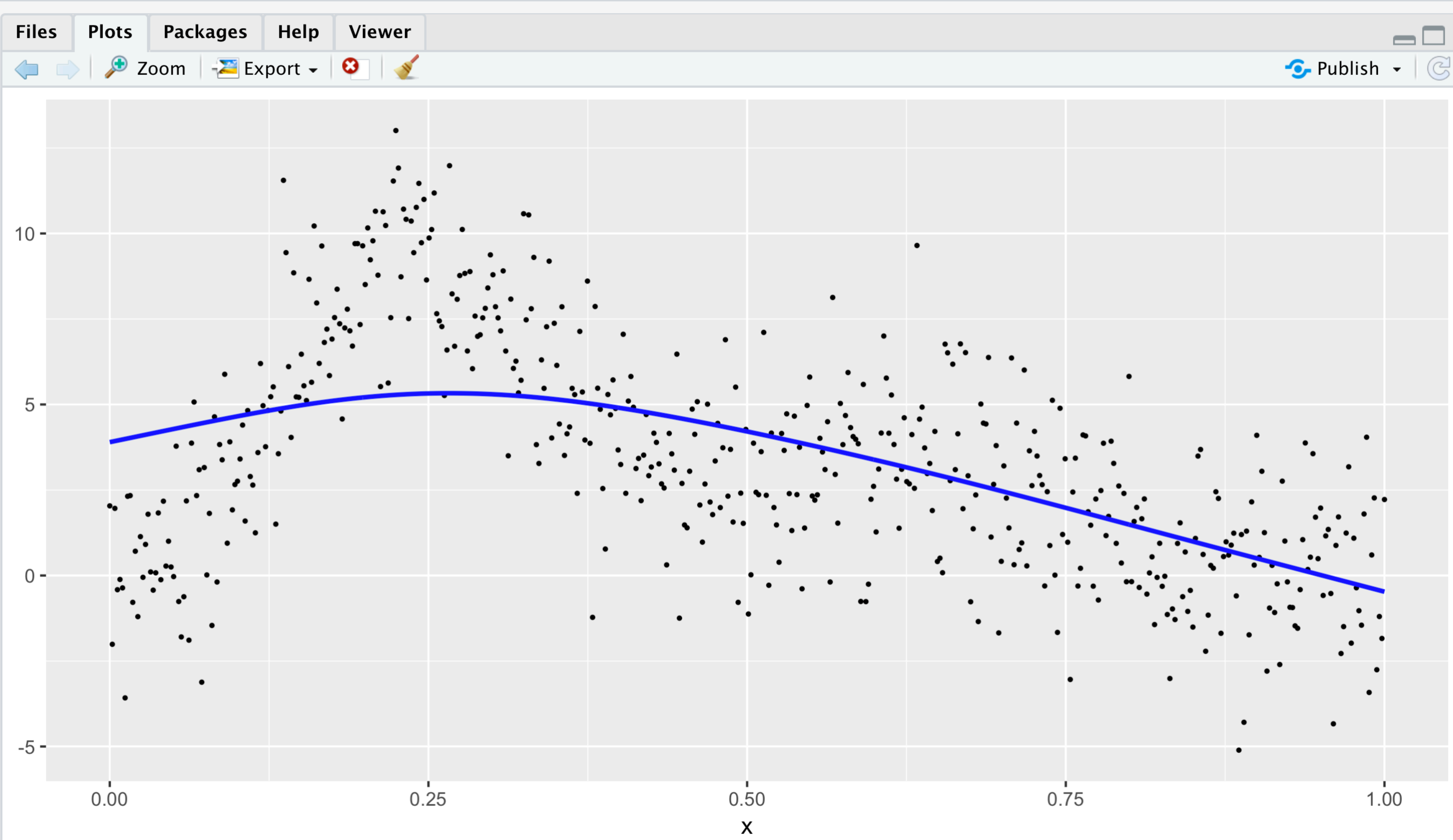
Given Data (x, y) from "unknown" $f(x)$



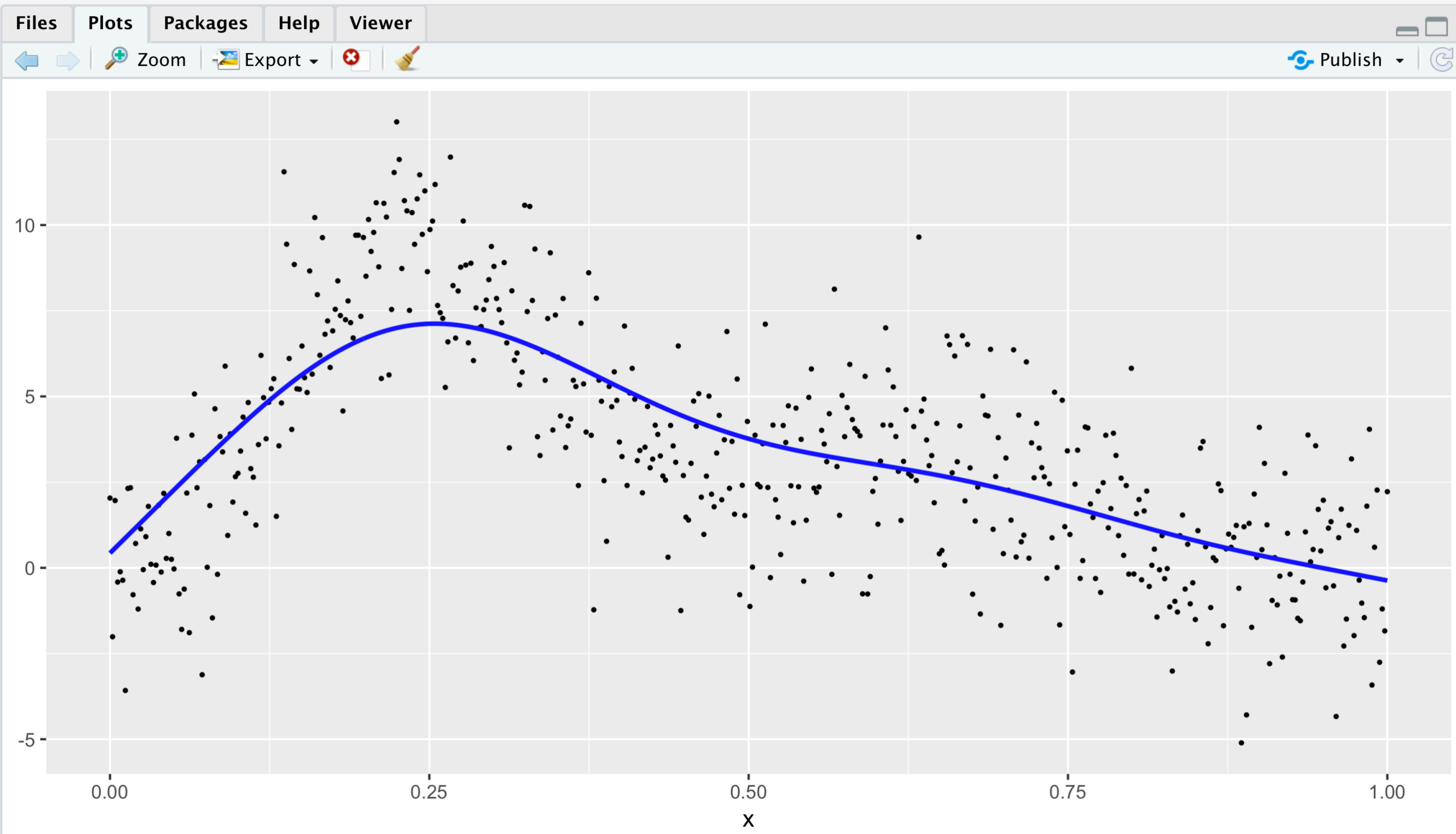
Approximate (i.e. “fit”) a Model $\hat{f}(x)$



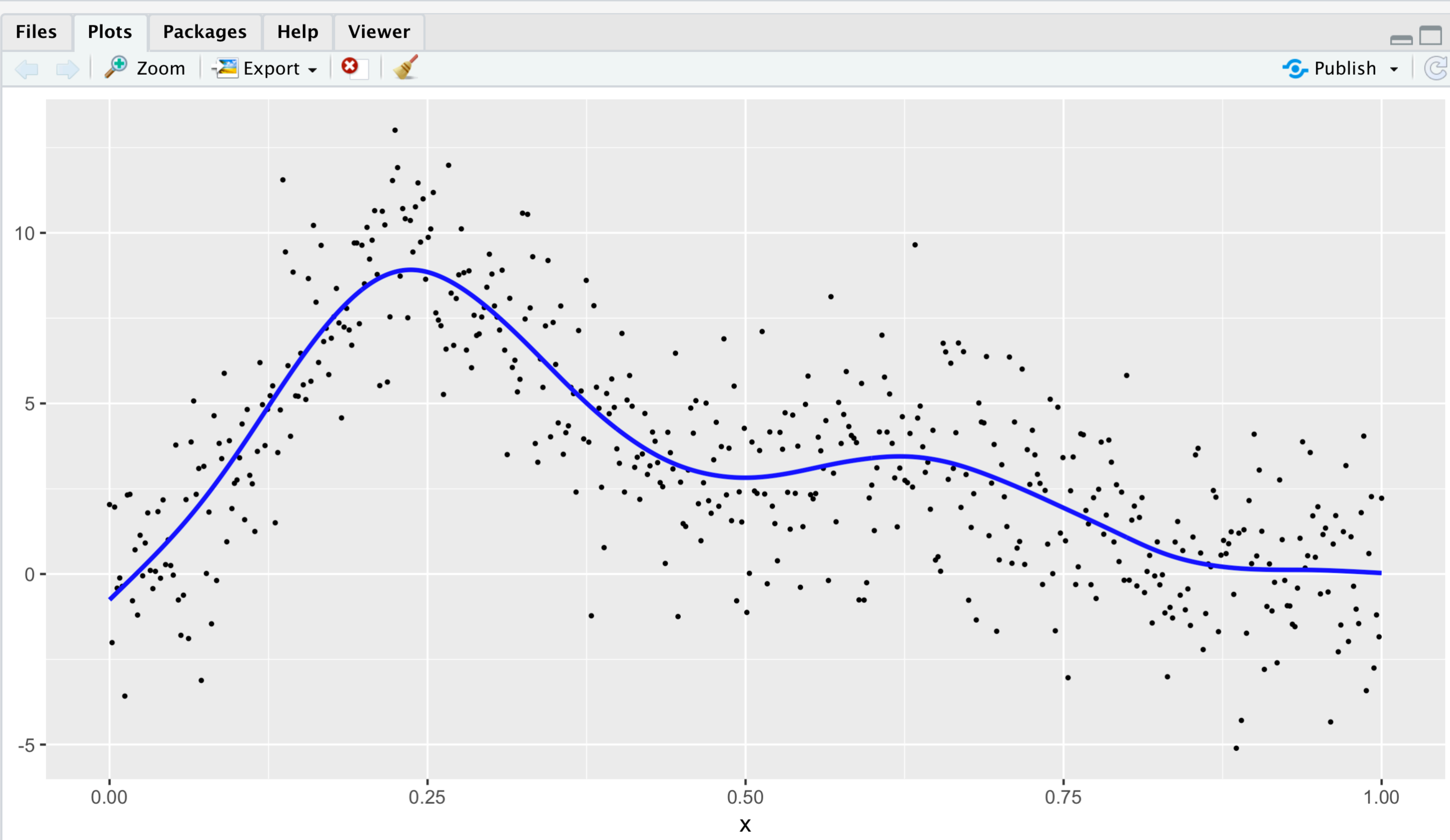
How about this $\hat{y} = \hat{f}(x)$?



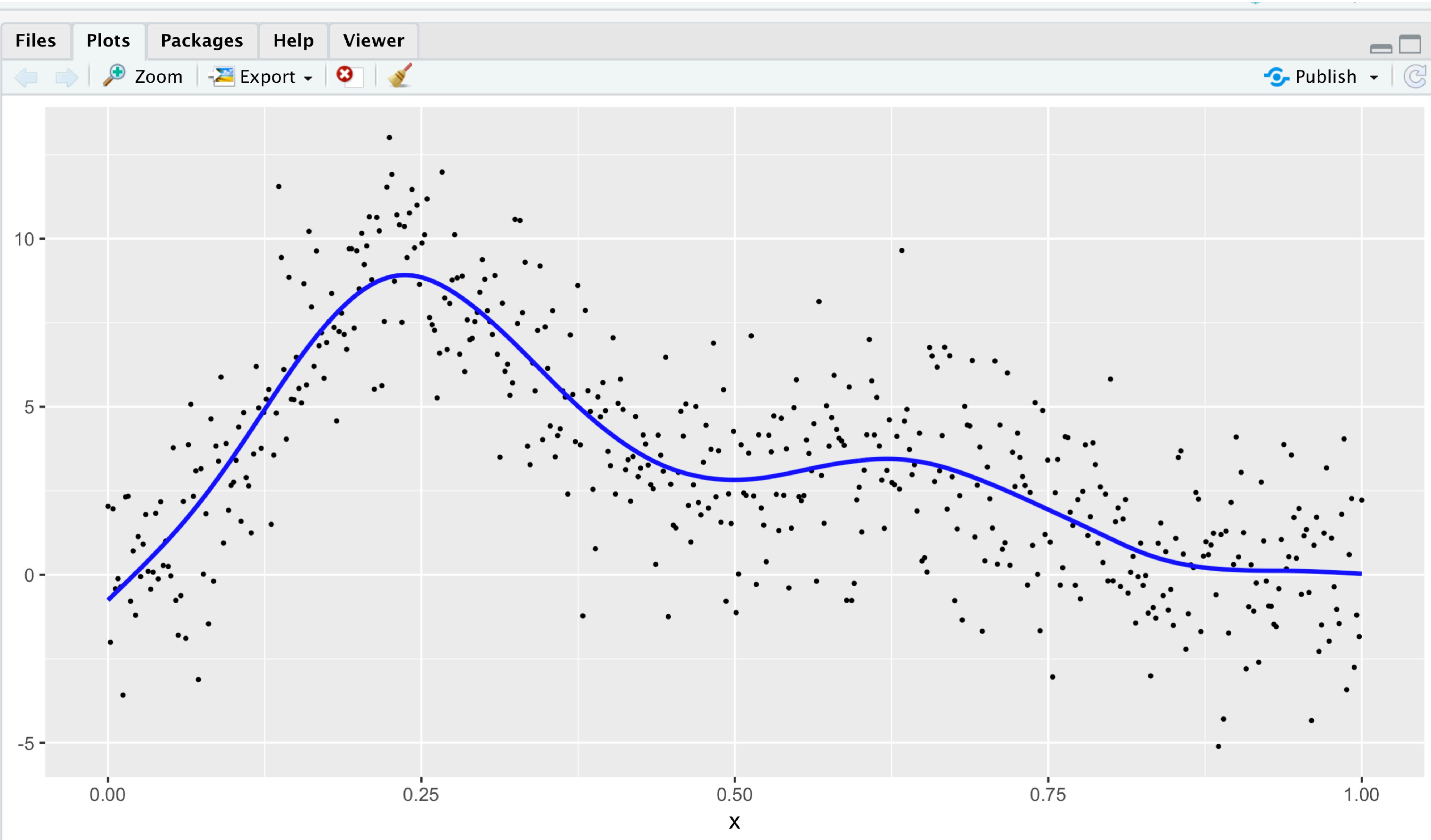
How about this $\hat{f}(x)$?



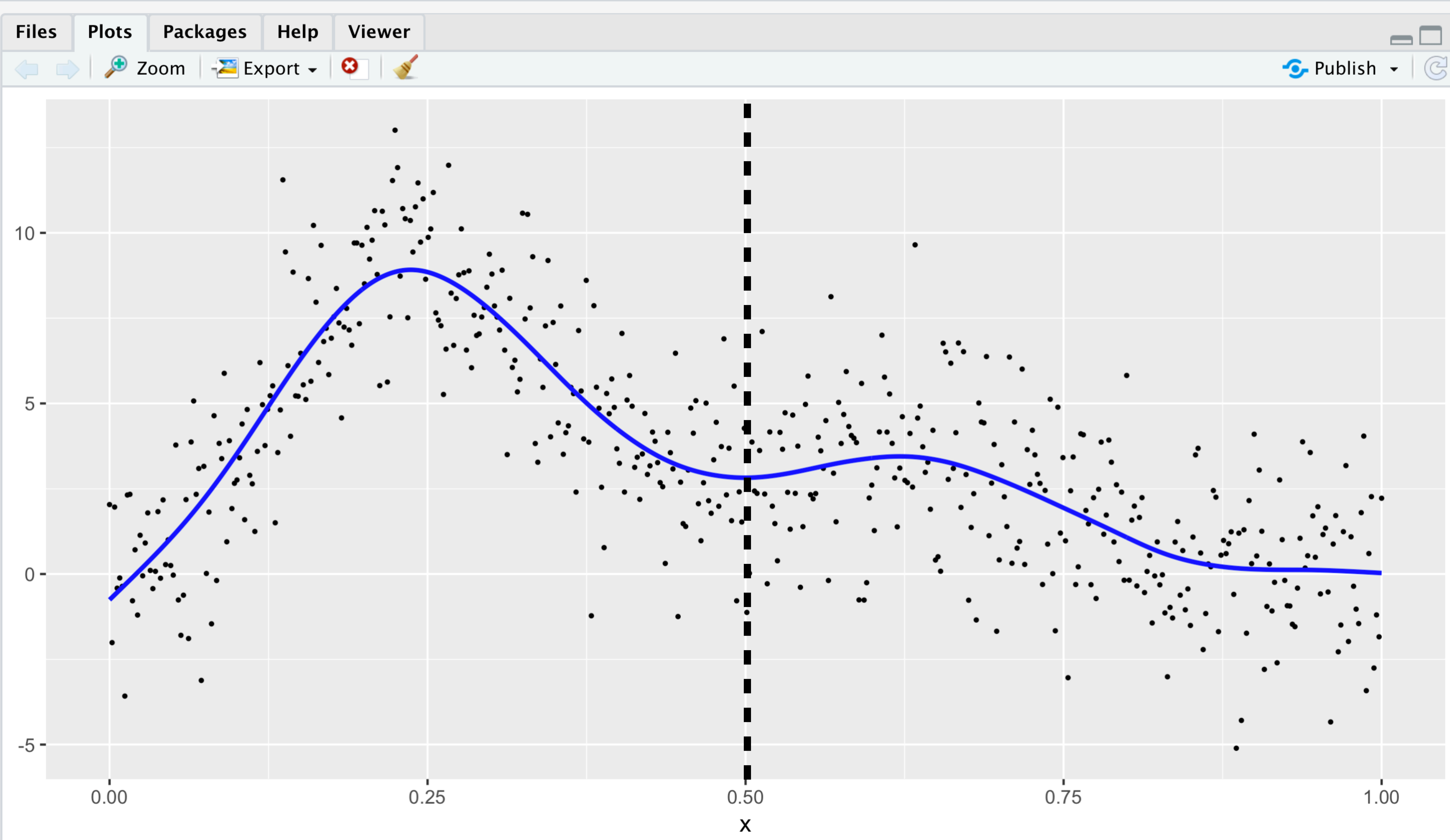
How about this $\hat{f}(x)$?



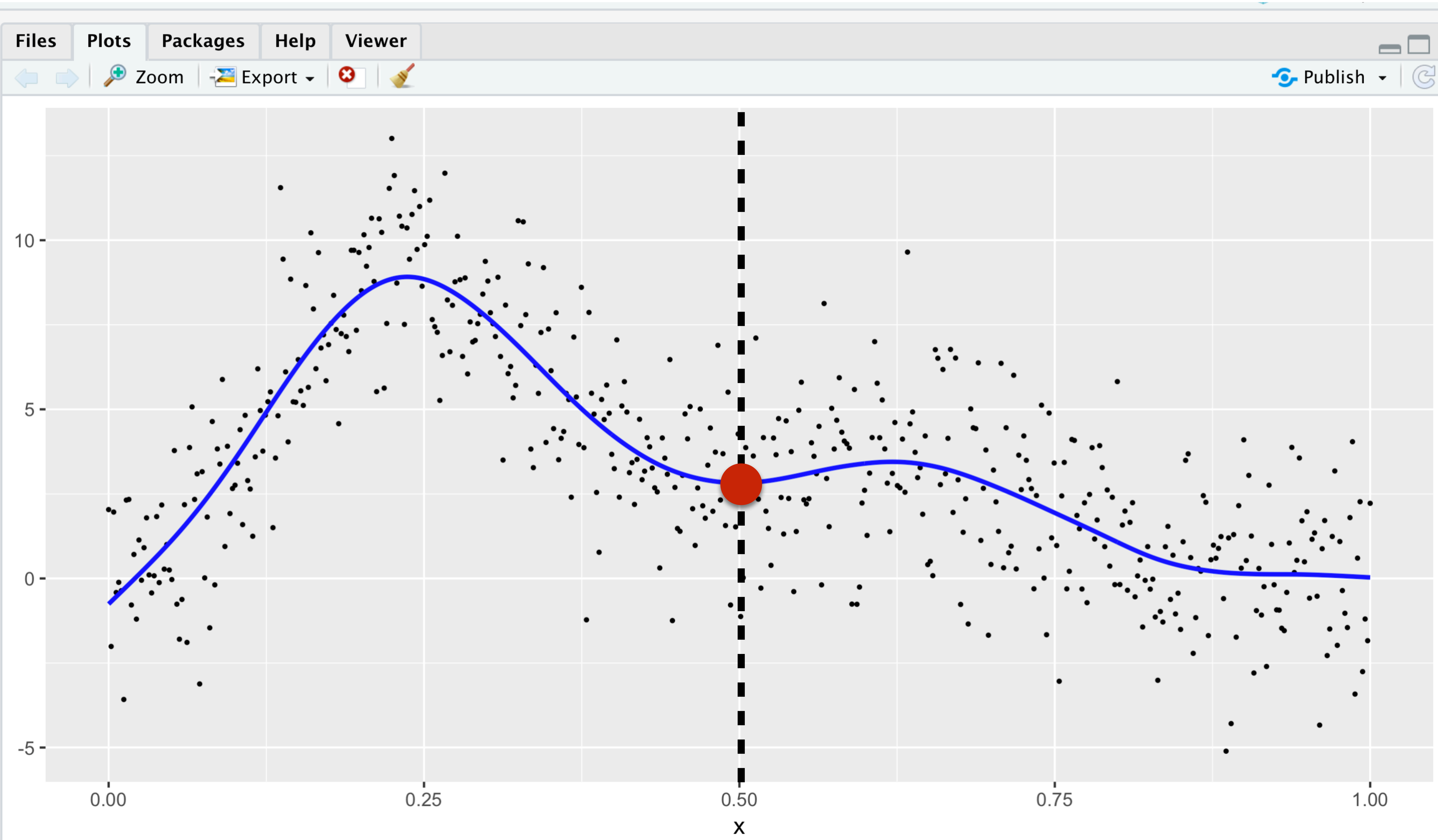
What does this $\hat{f}(x)$ predict for $x = 0.5$?



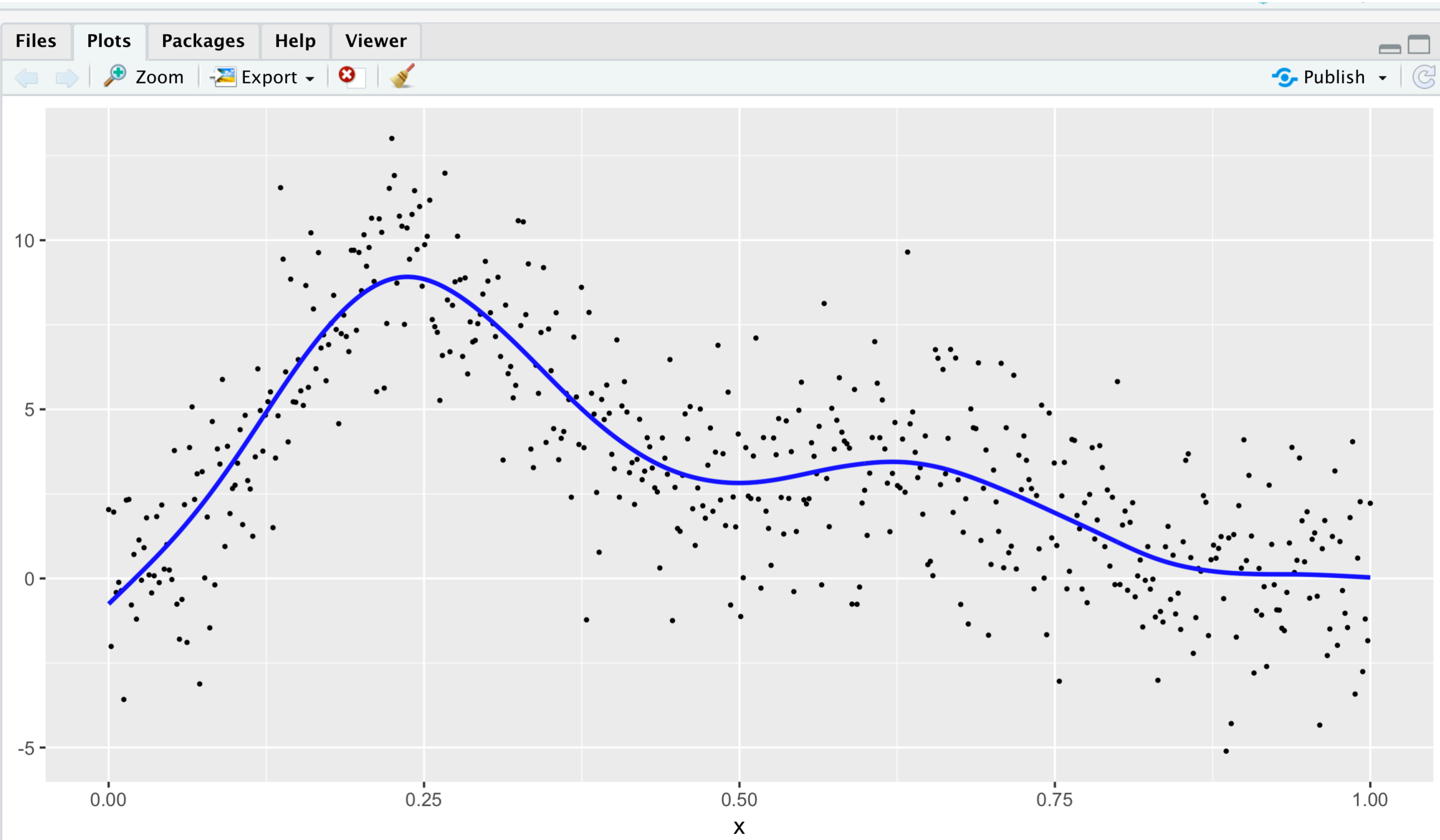
What does this $\hat{f}(x)$ predict for $x = 0.5$?



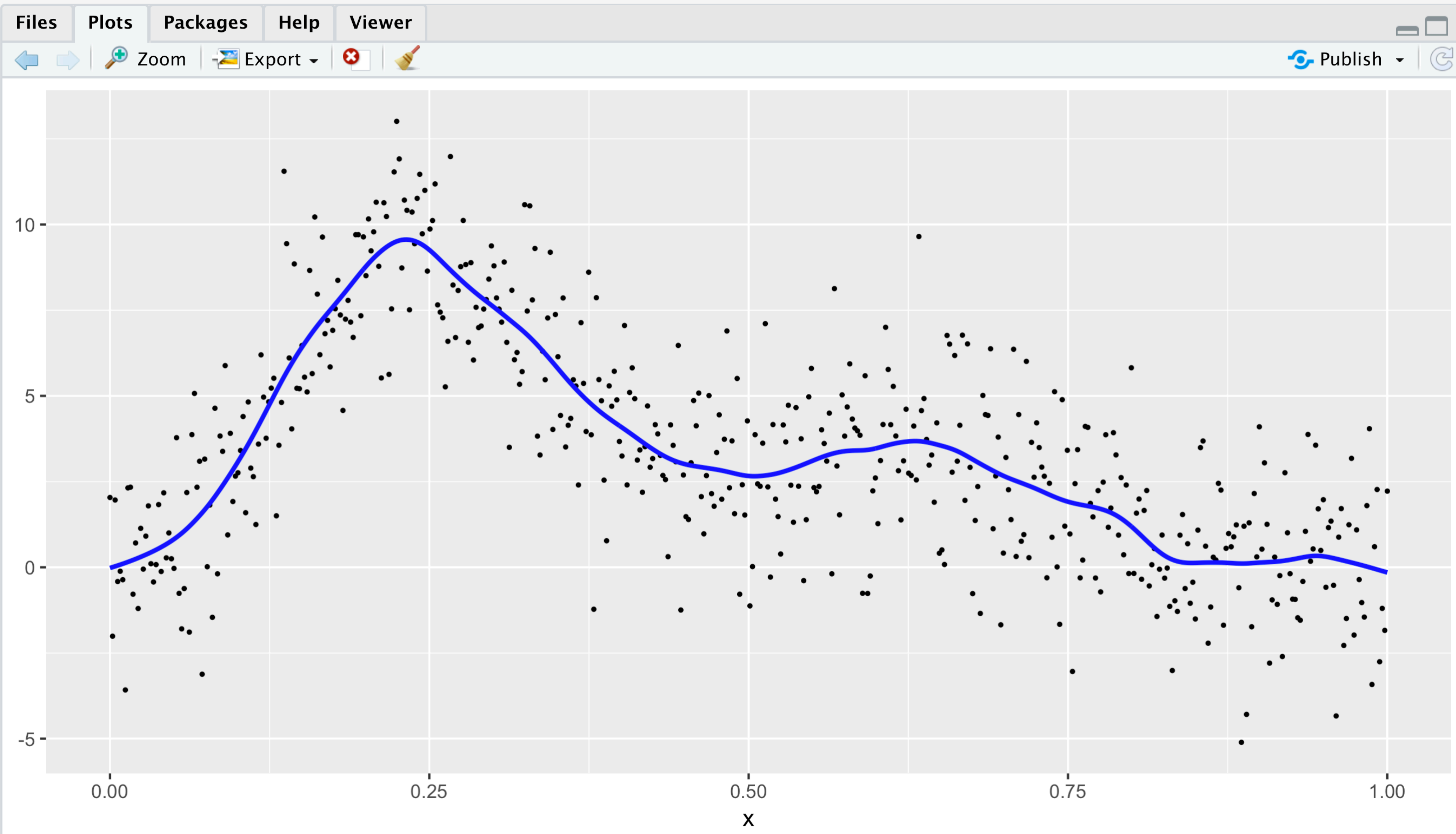
What does this $\hat{f}(x)$ predict for $x = 0.5$?



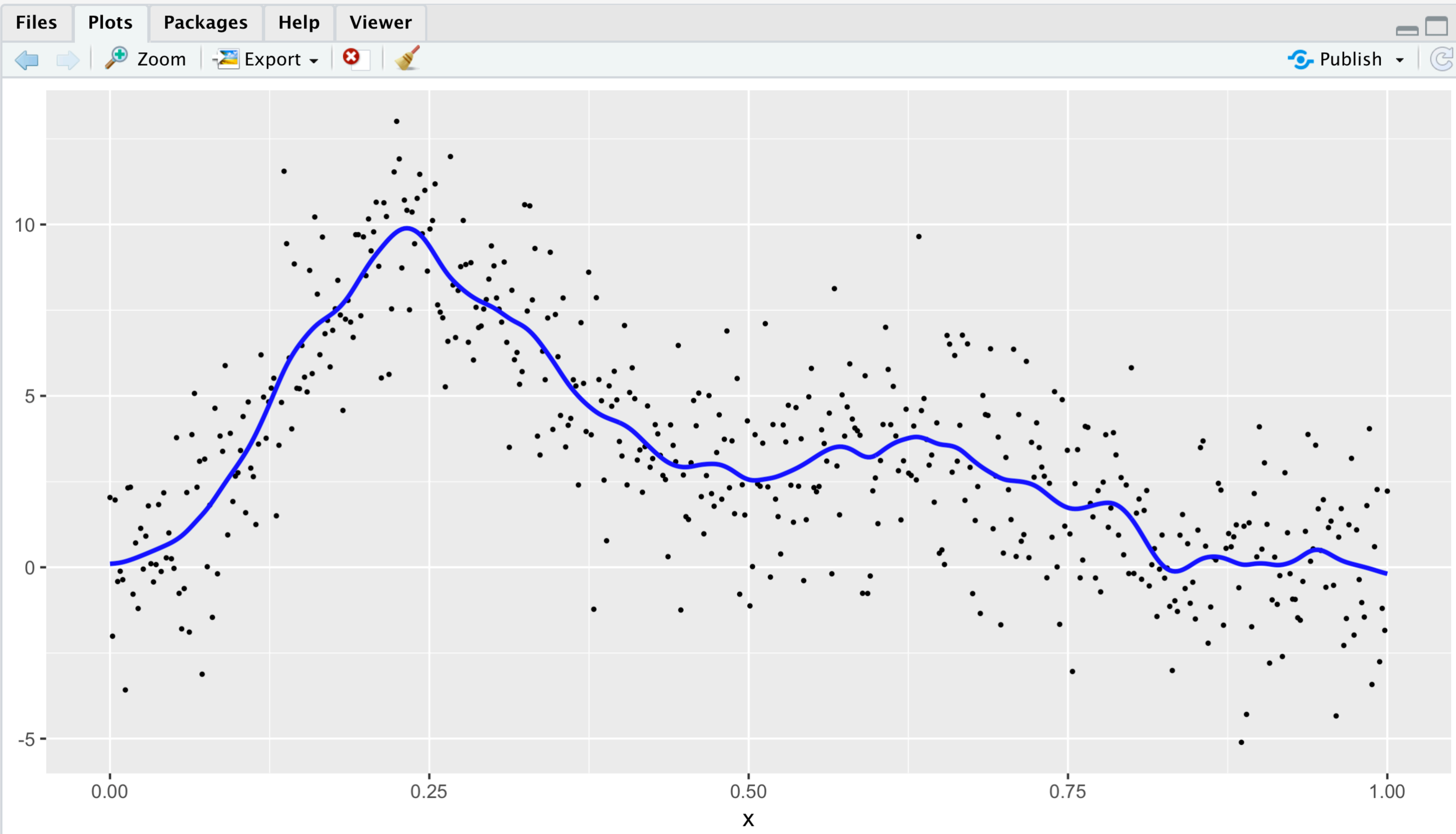
Ok, great. But instead of this $\hat{f}(x)$ …



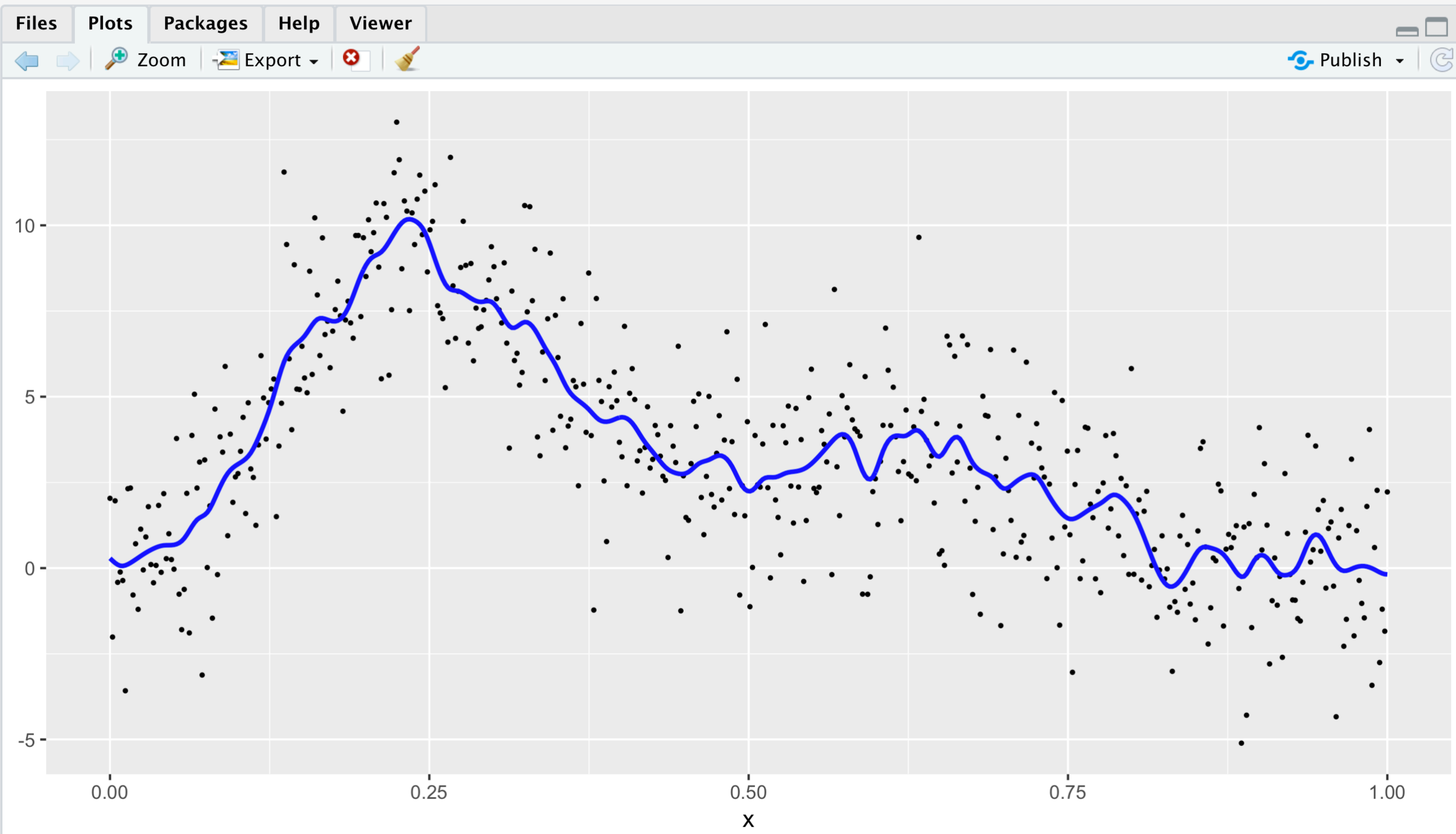
How about this $\hat{f}(x)$?



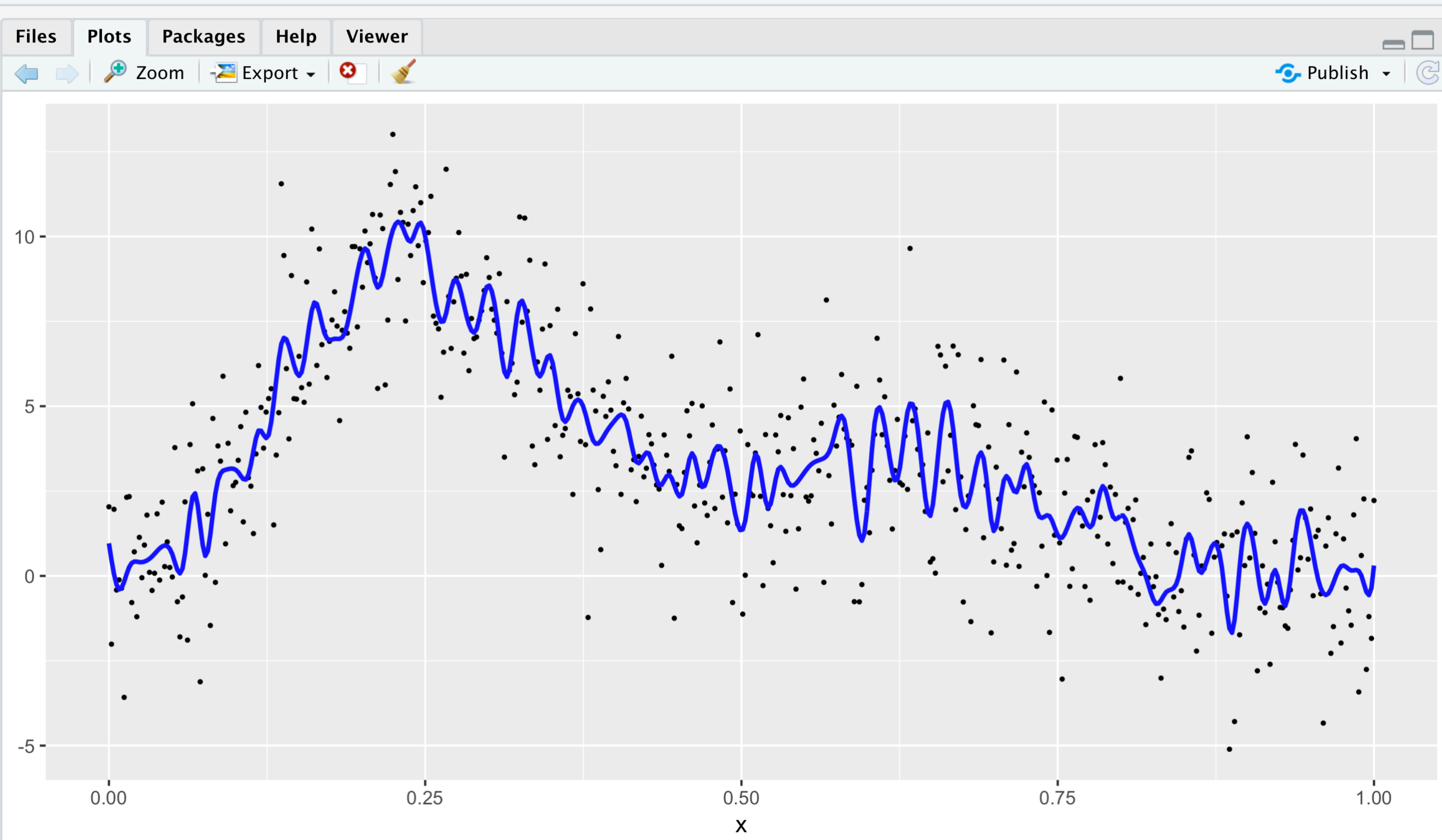
How about this $\hat{f}(x)$?



How about this $\hat{f}(x)$?



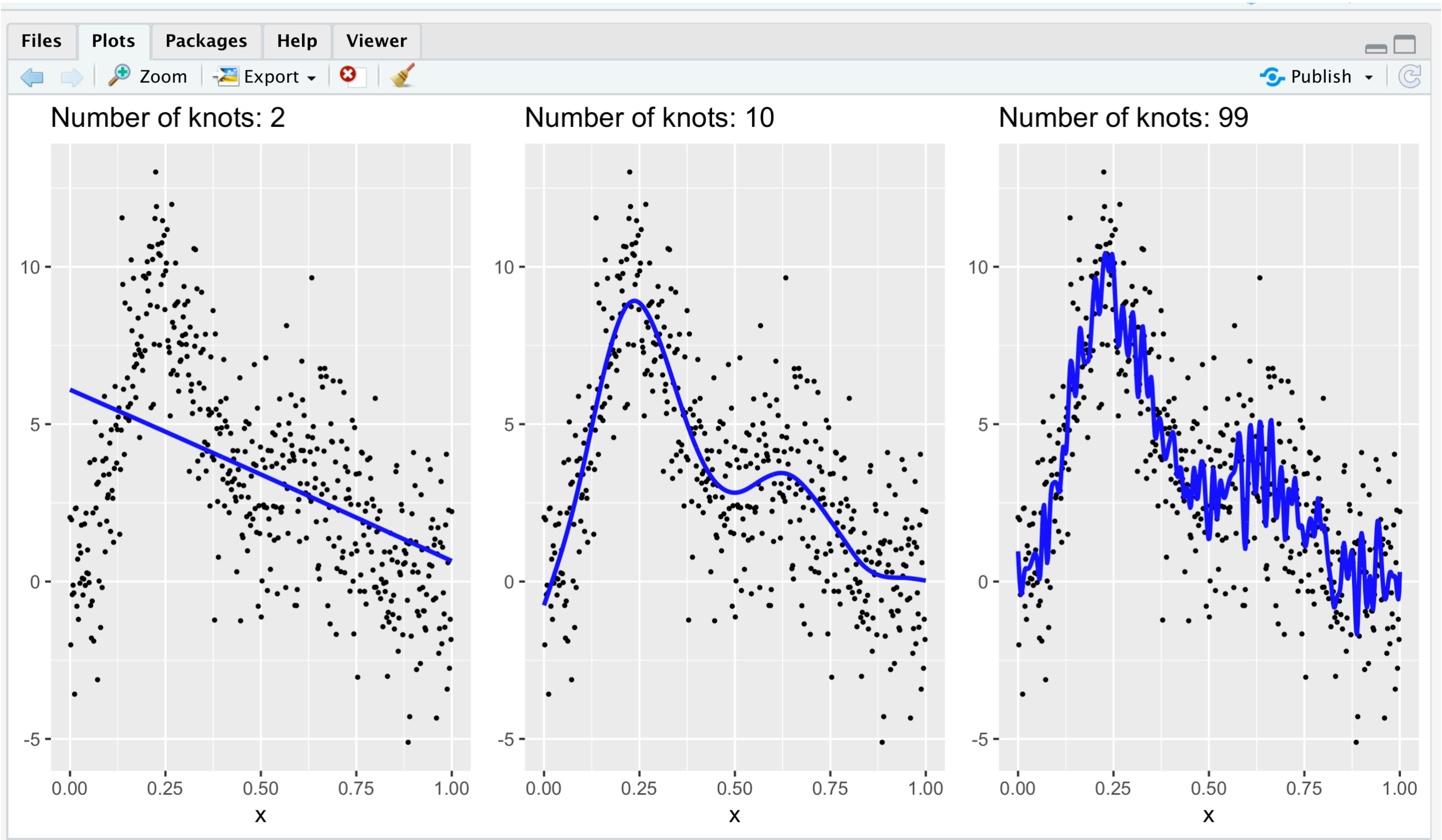
How about this $\hat{f}(x)$?



Model Fitting Method: (Cubic) Splines

- Splines fit the blue curve $\hat{f}(x)$ that **minimizes** the (squared) vertical distances between all:
  - predicted points $\hat{y} = \hat{f}(x)$ and
  - observed points y
- Amount of “wiggle” is the **complexity of the model**
- Occam’s Razor

Three Different $\hat{f}(x)$



Underfit! "Just right!" Overfit!

Road Map

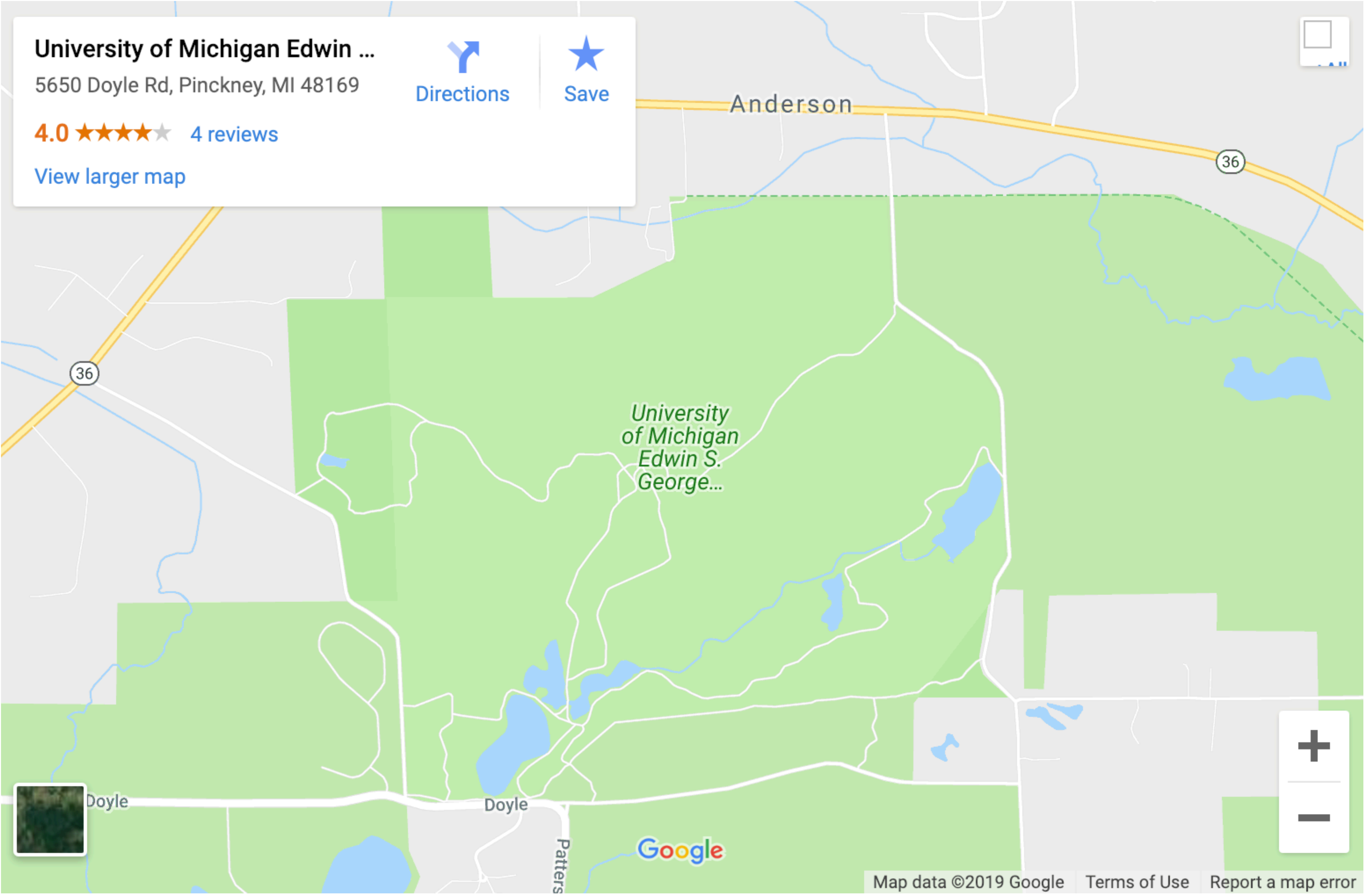
Machine Learning:
- Machine learning as modeling
- Example model: splines

Forest Ecology:
- Our data

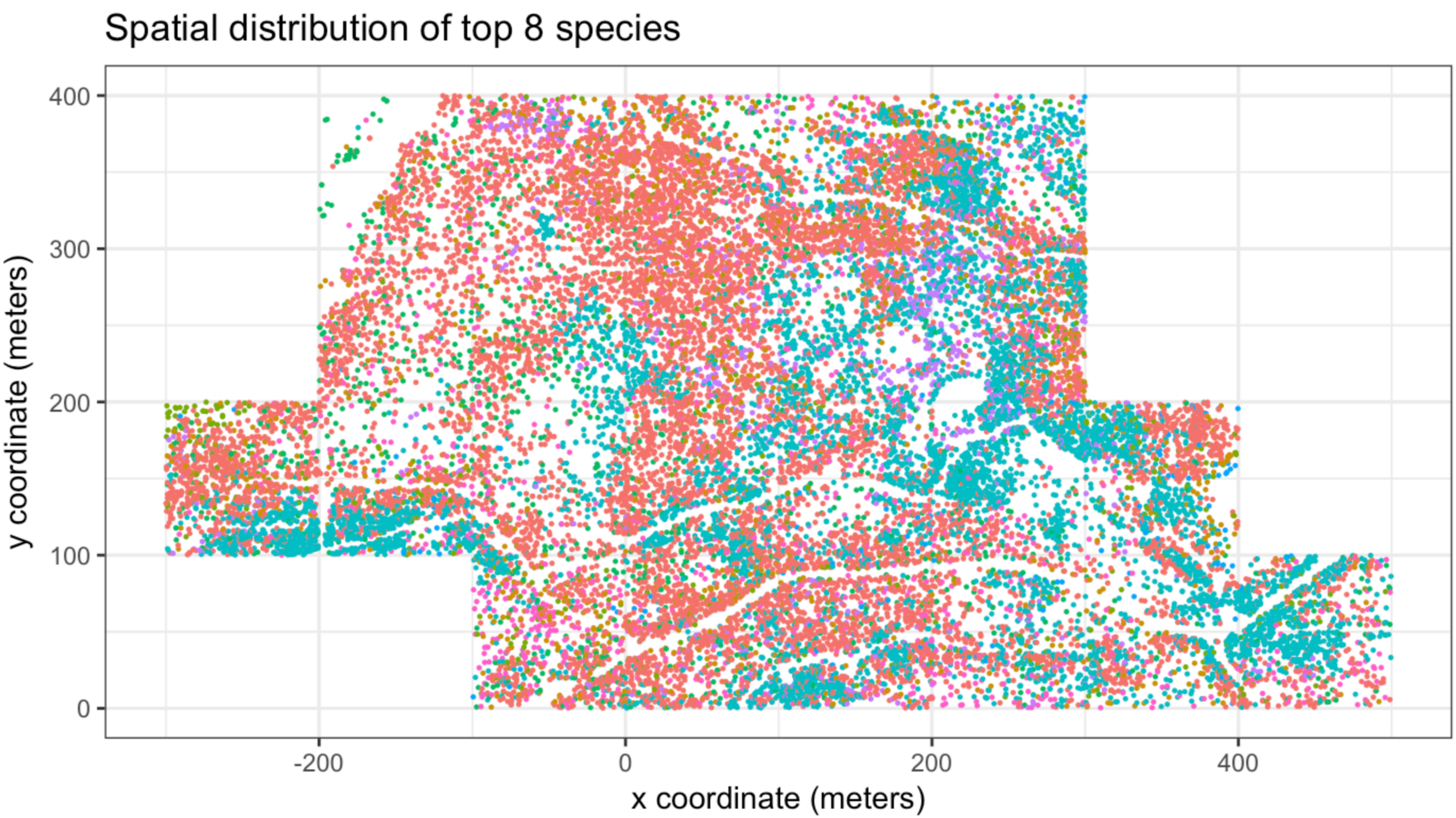
Machine Learning + Forest Ecology:
- Goal restated
- Assessment model performance

Conclusion:
- Thoughts on modeling

Data: 2008 & 2014 Censuses of Trees



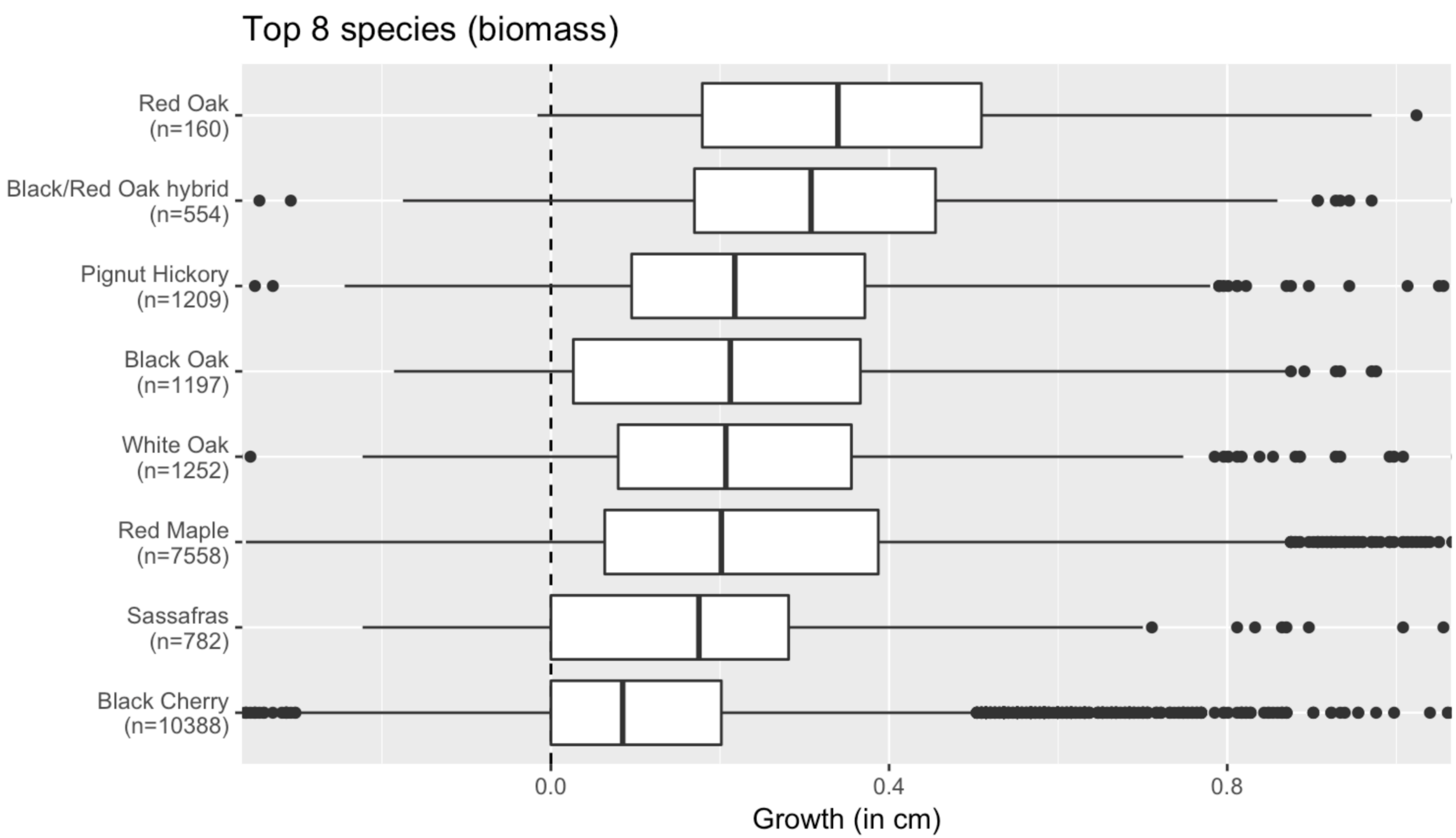
Data: 2008 Snapshot



Recall our Variables!

y : Outcome Variable = Avg Annual Growth

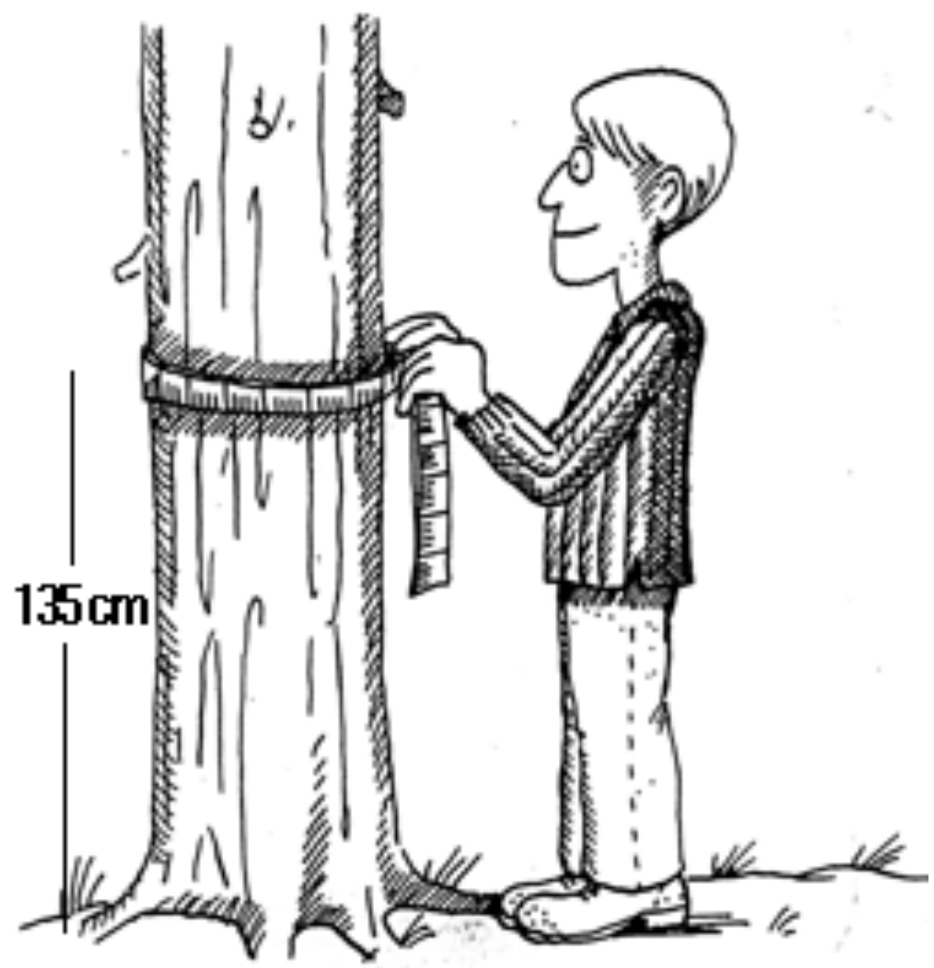
Observed average annual growth of trees 2008-2014



Predictor Variables $\vec{x}$

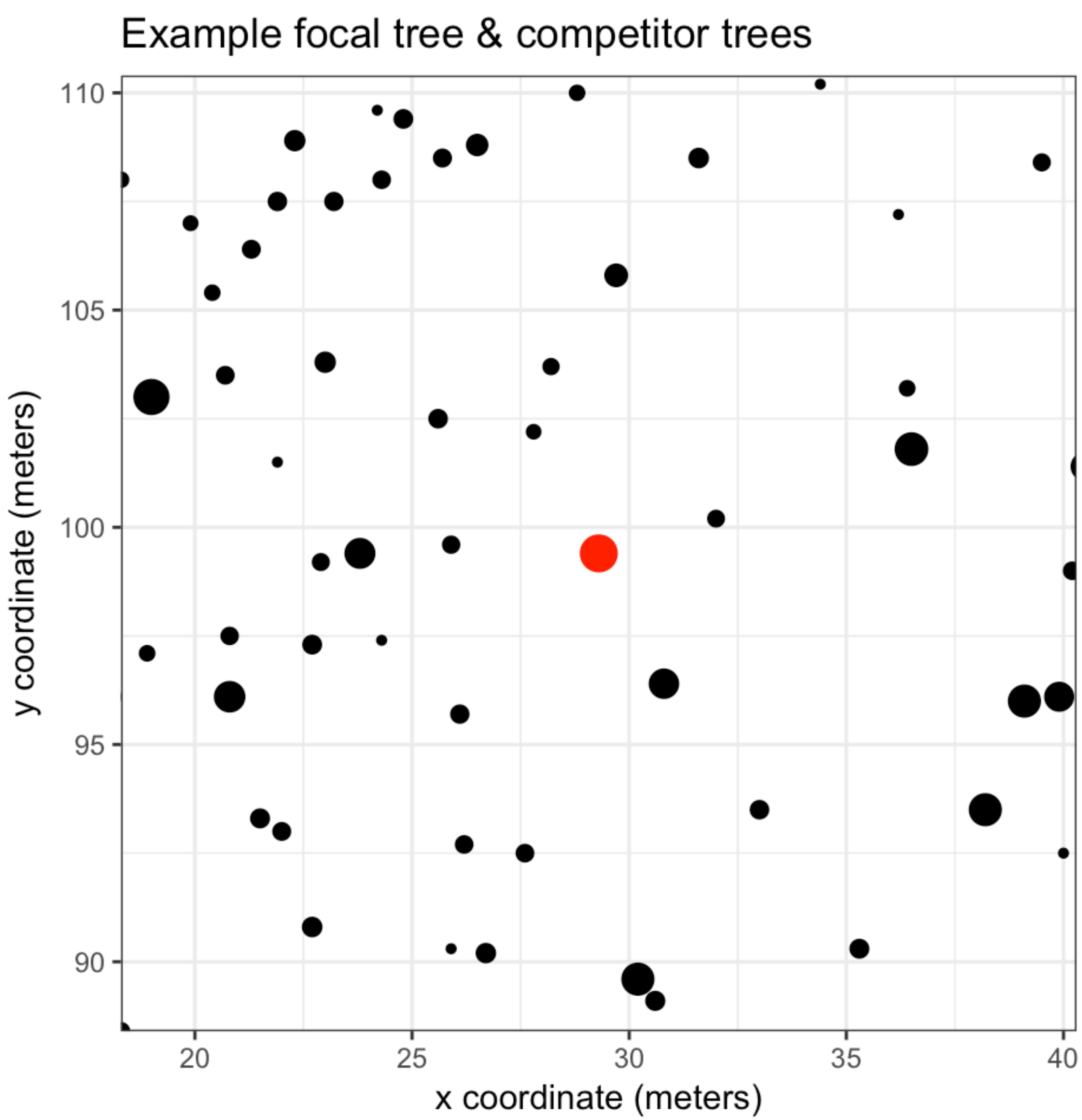
x_1 : Species of tree

x_2 : Size of tree (diameter at breast height)



Predictor Variables

x_3 : Number and size of competitor trees (biomass)



Road Map

Machine Learning:
- Machine learning as modeling
- Example model: splines

Forest Ecology:
- Our data

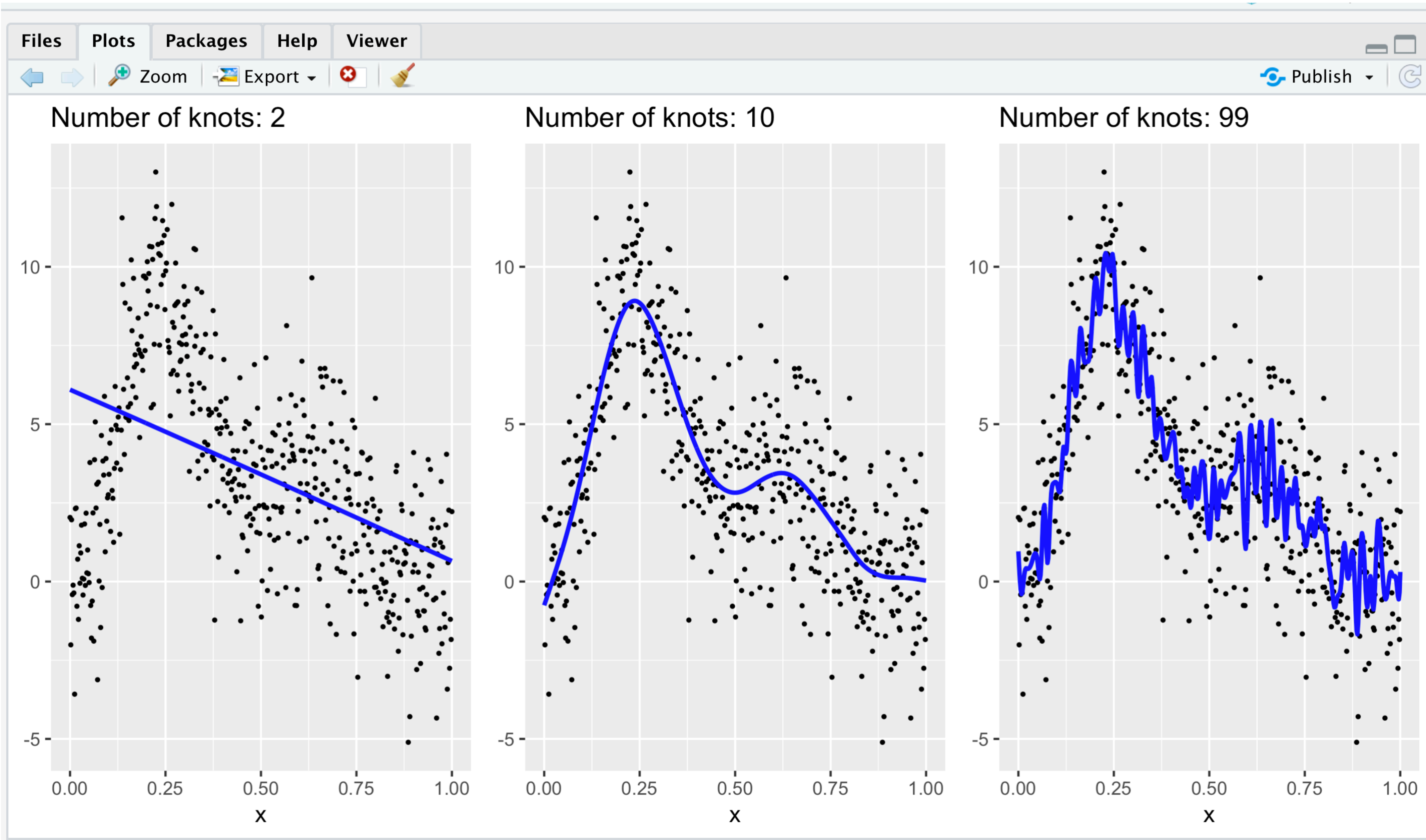
Machine Learning + Forest Ecology:
- Goal restated
- Assessment model performance

Conclusion:
- Thoughts on modeling

Machine Learning & Forest Ecology

- **Goal of Modeling:** Fit models $\hat{f}(x)$ that best approximate the true (unknown) model $f(x)$
- **Goal of Machine Learning:** Fit models that best "predict" the outcome variable
- **My goal:** Fit models that best predict the growth of trees
- **Tools**: The same machine learning tools and framework as self-driving cars

Issue of underfitting vs overfitting?



Underfit! “Just right!” Overfit!

Validation Set Approach

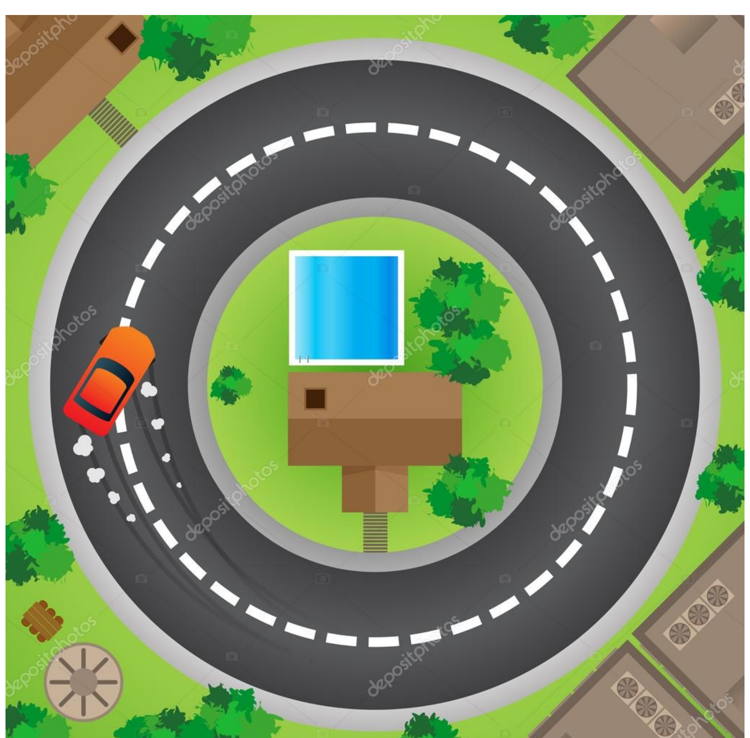
1 2 3 n

Split your data into:

7 22 13 91

Fit/train model on ***training*** data

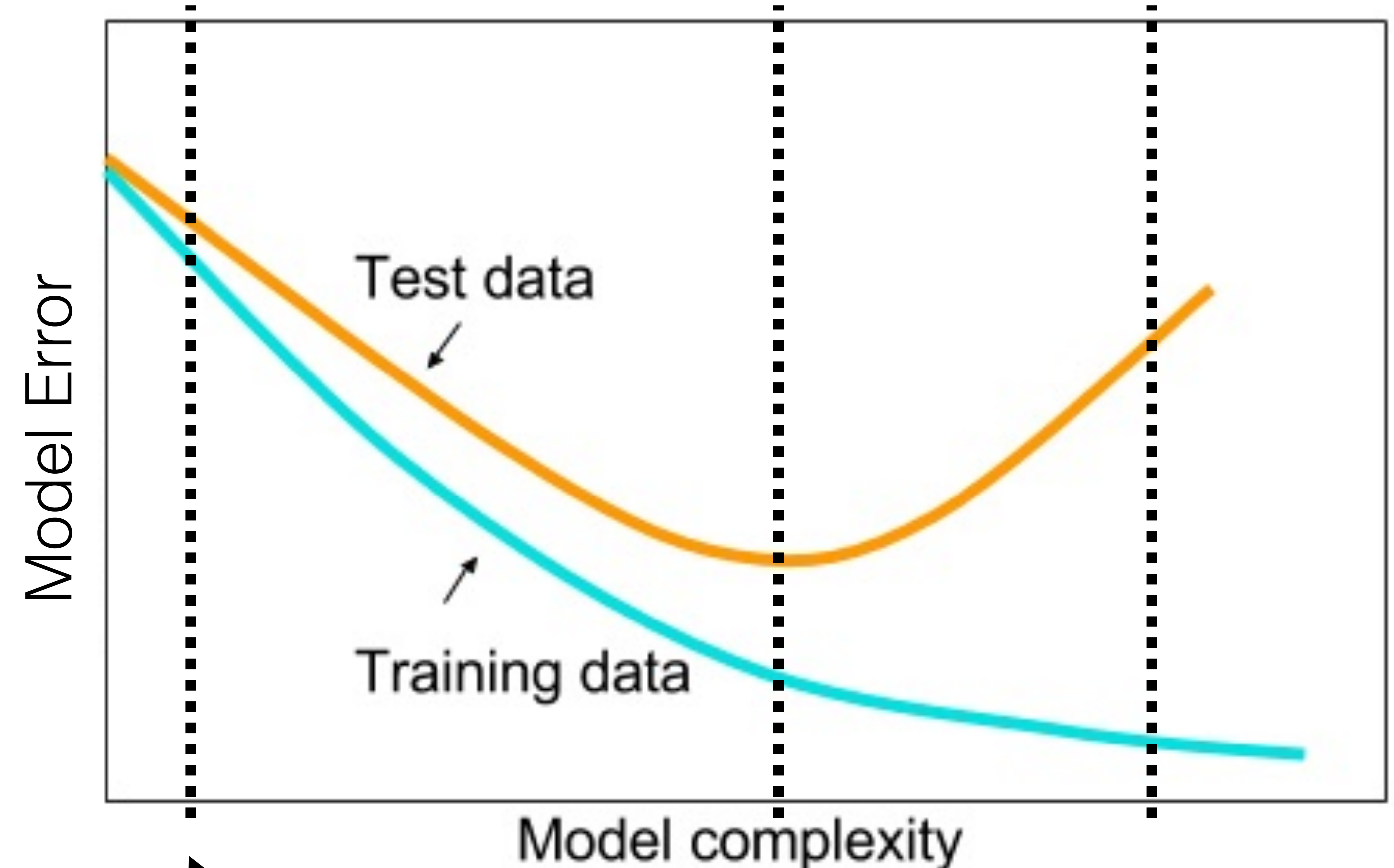
Assess model on independent ***test*** data



Typical Model Performance

Fit/train model on the ***training*** data

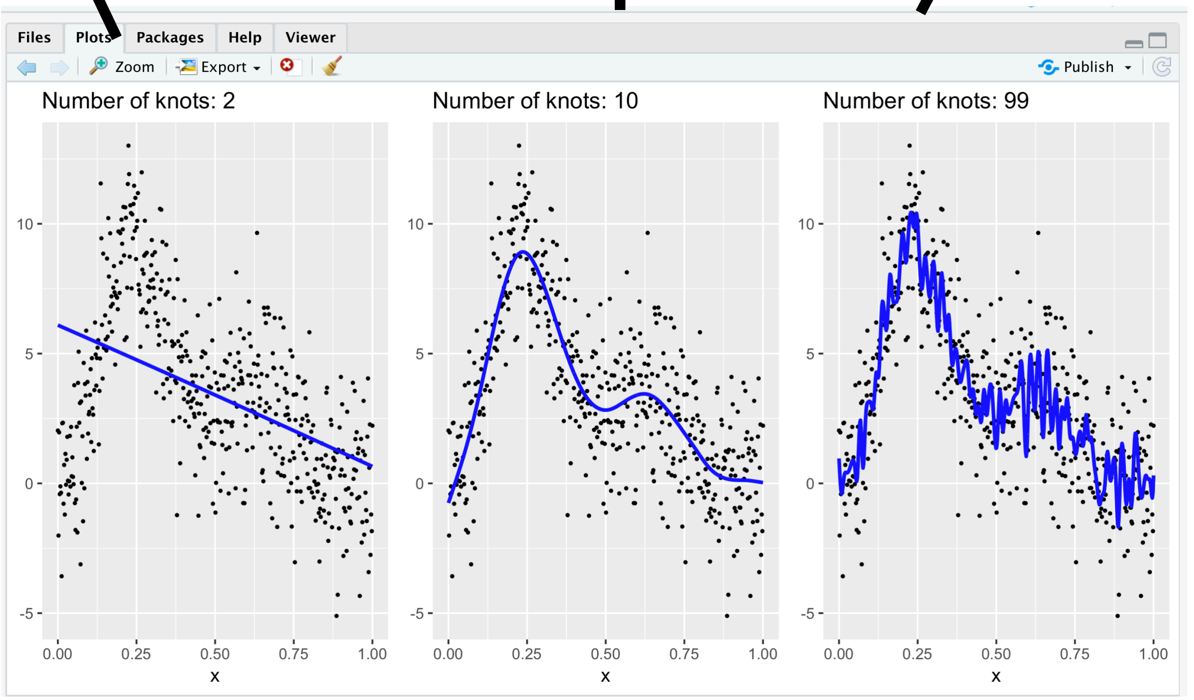
Assess your model's "error" on independent ***test*** data



Underfit!

"Just right!"

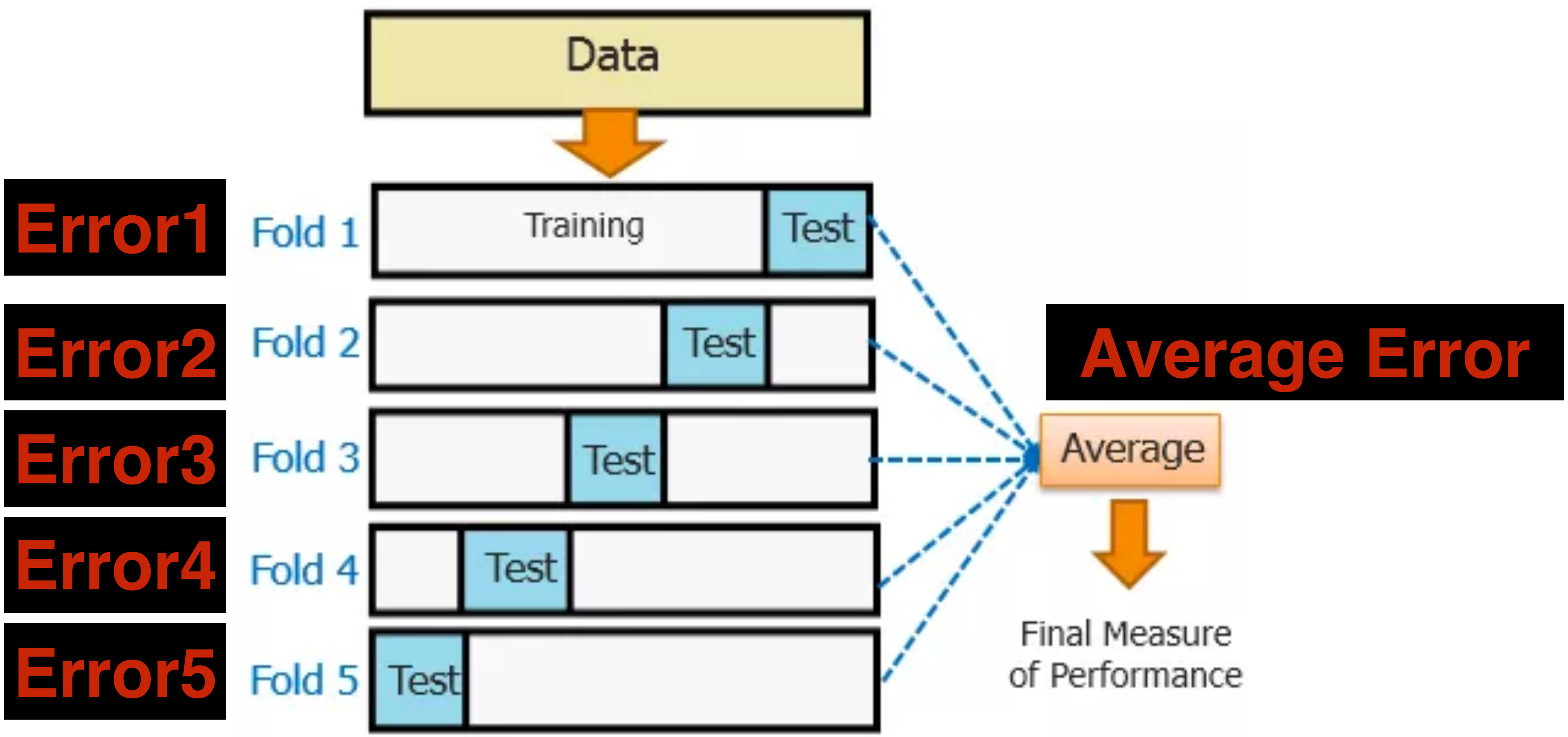
Overfit!



Recall for splines, the # of knots controls the **model complexity**

Generalization: 5-Fold Crossvalidation

Repeat validation training/test set split 5 times:



Road Map

Machine Learning:
- Machine learning as modeling
- Example model: splines

Forest Ecology:
- Our data

Machine Learning + Forest Ecology:
- Goal restated
- Assessment model performance

Conclusion:
- Thoughts on modeling

$\hat{f}(x)$ Model Fitting

Daniela Witten
@daniela_witten

Follow

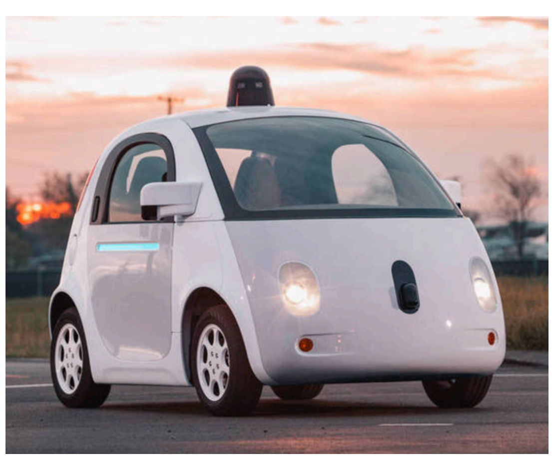
Perfect gym for a statistician

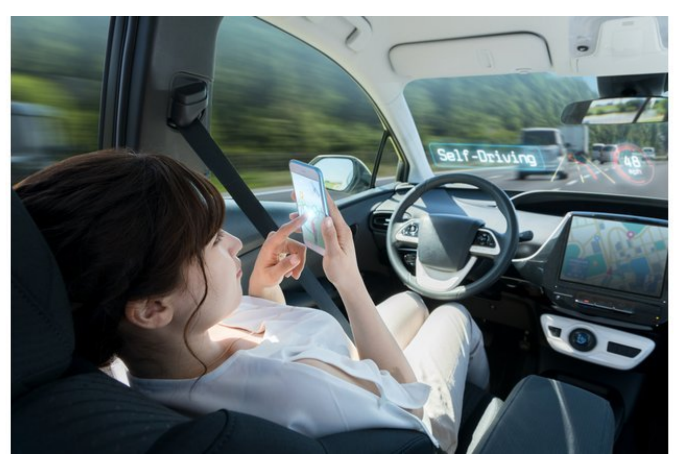
Modeling is not as objective as you think

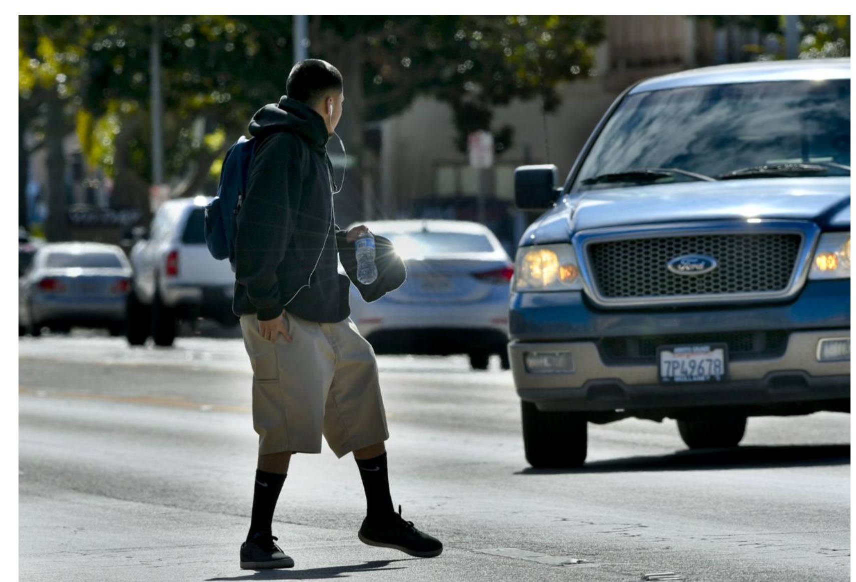
Scenario:

What they think is an “appropriate” model...

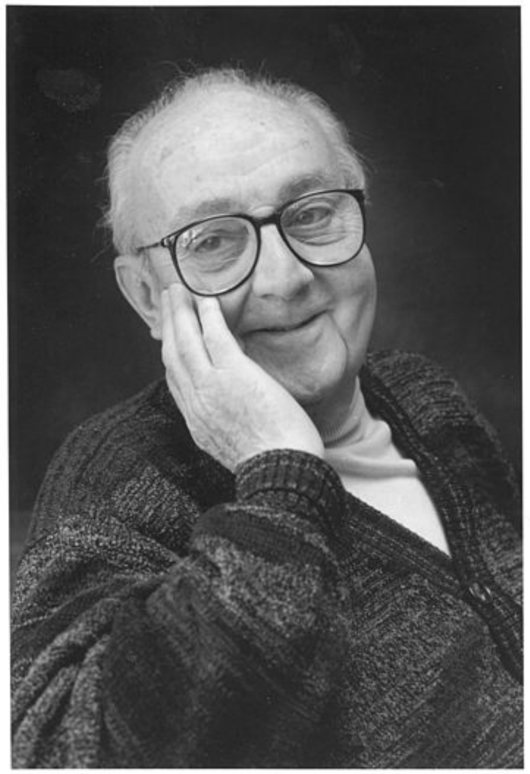
… might not be the same for these folks:







To Close: Two Quotes on Modeling



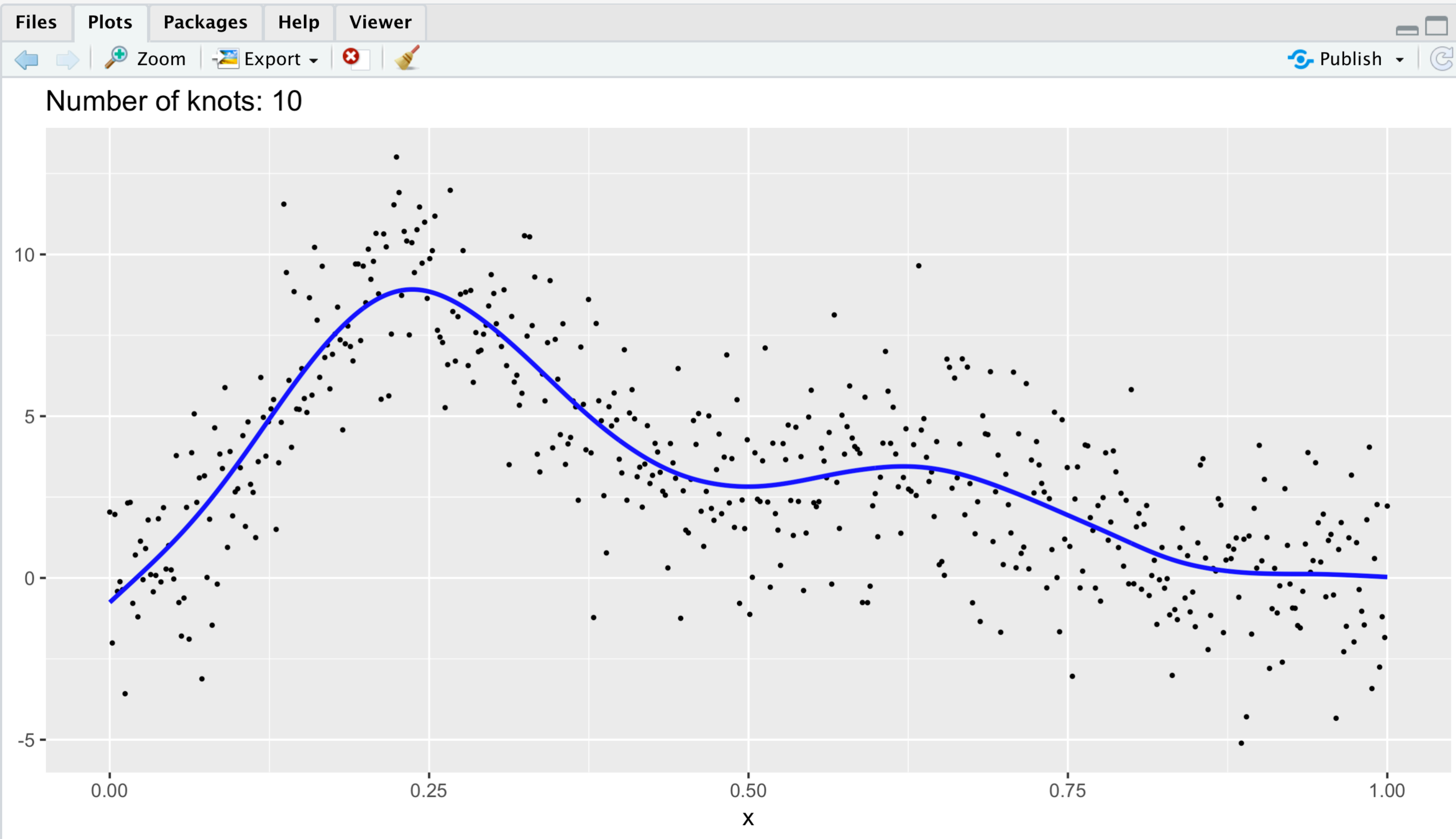
"All models are wrong, but some are useful."
George Box

"WTF is up with your $\hat{f}(x)$?" @rudeboybert

Thanks!

Before I go: A “Wizard of Oz” Reveal…

Our approximated $\hat{f}(x)$ …



… was pretty close to the *true* model:

$$f(x) = 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$$