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

Albert Y. Kim
Assistant Professor
Statistical & Data Sciences, Smith College

Slides available on Twitter @rudeboybert

What variables are being collected?

Background

Road Map

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

Forest Ecology:
- Our data
- Two models for tree growth

Machine Learning + Forest Ecology:
- Goal restated
- Assessment metric used
- Model assessment with crossvalidation

Conclusion:
- Resources to learn more
- Thoughts on modeling

Machine Learning

WAYMO

NETFLIX

Prediction!

TCH FIX

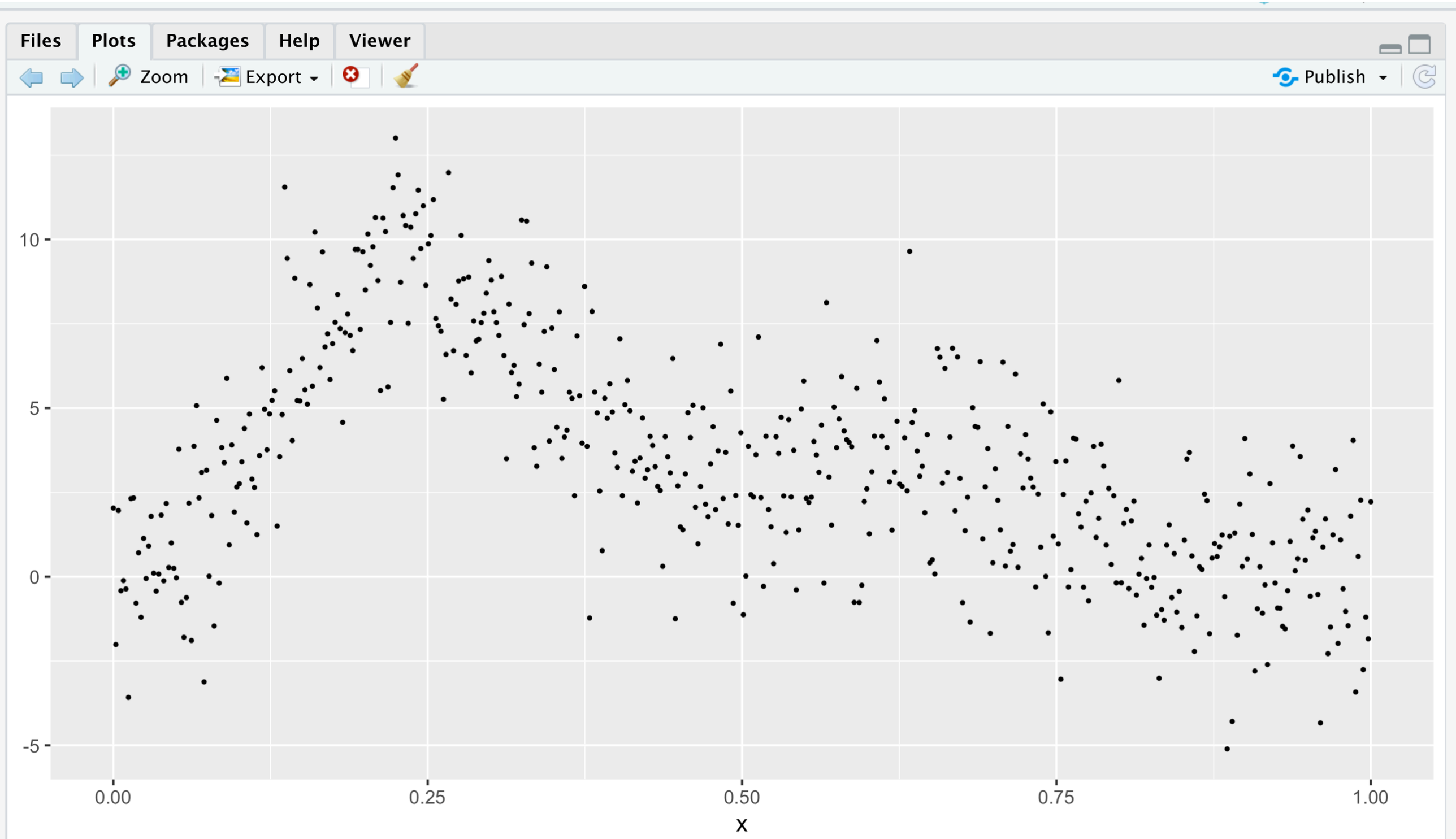
Machine Learning as Modeling

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

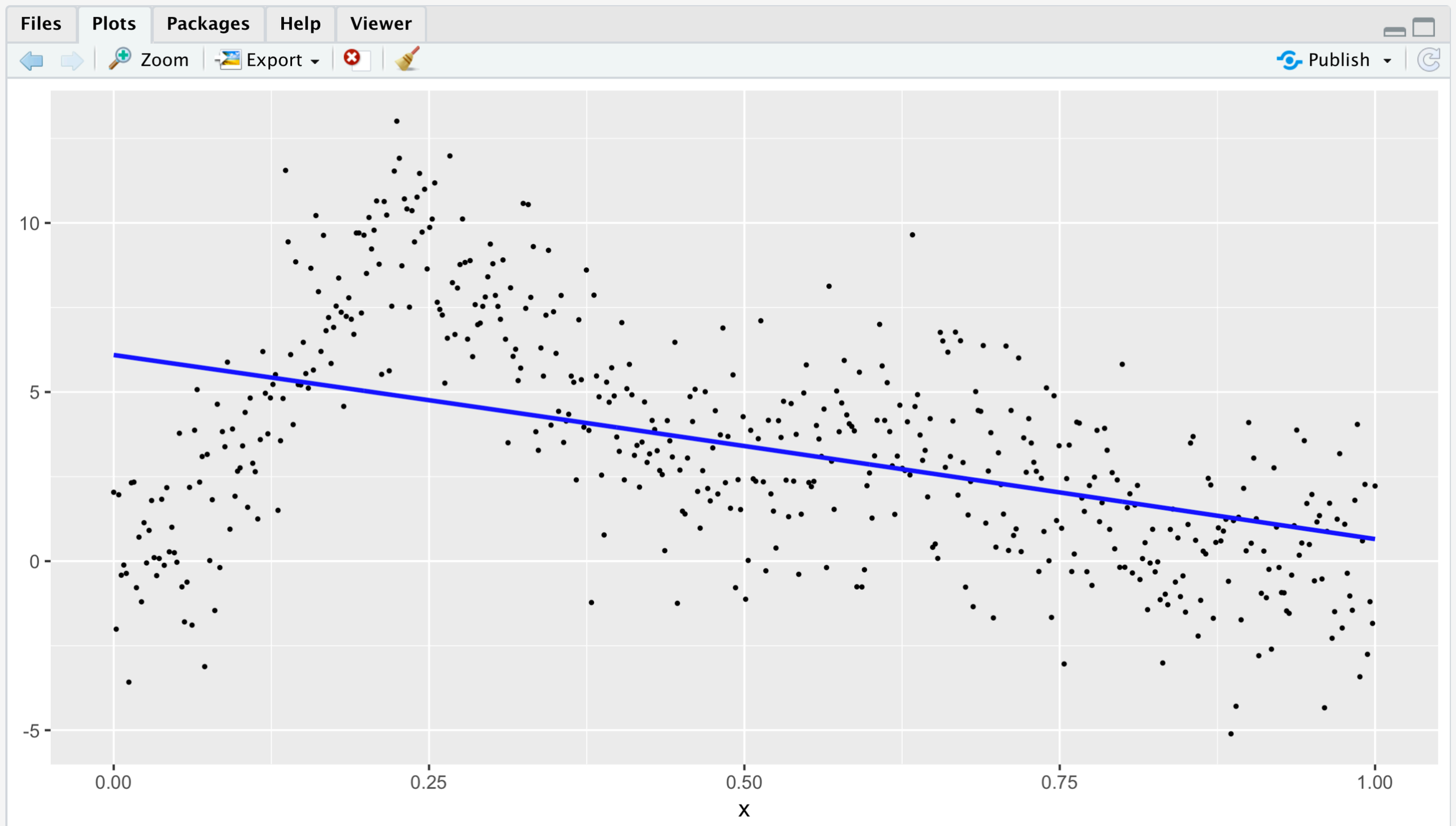
Approximated Model: $\hat{y} = \hat{f}(\overrightarrow{x}) = f(\overrightarrow{x})$

Now to the blackboard for Chalk Talk #1…

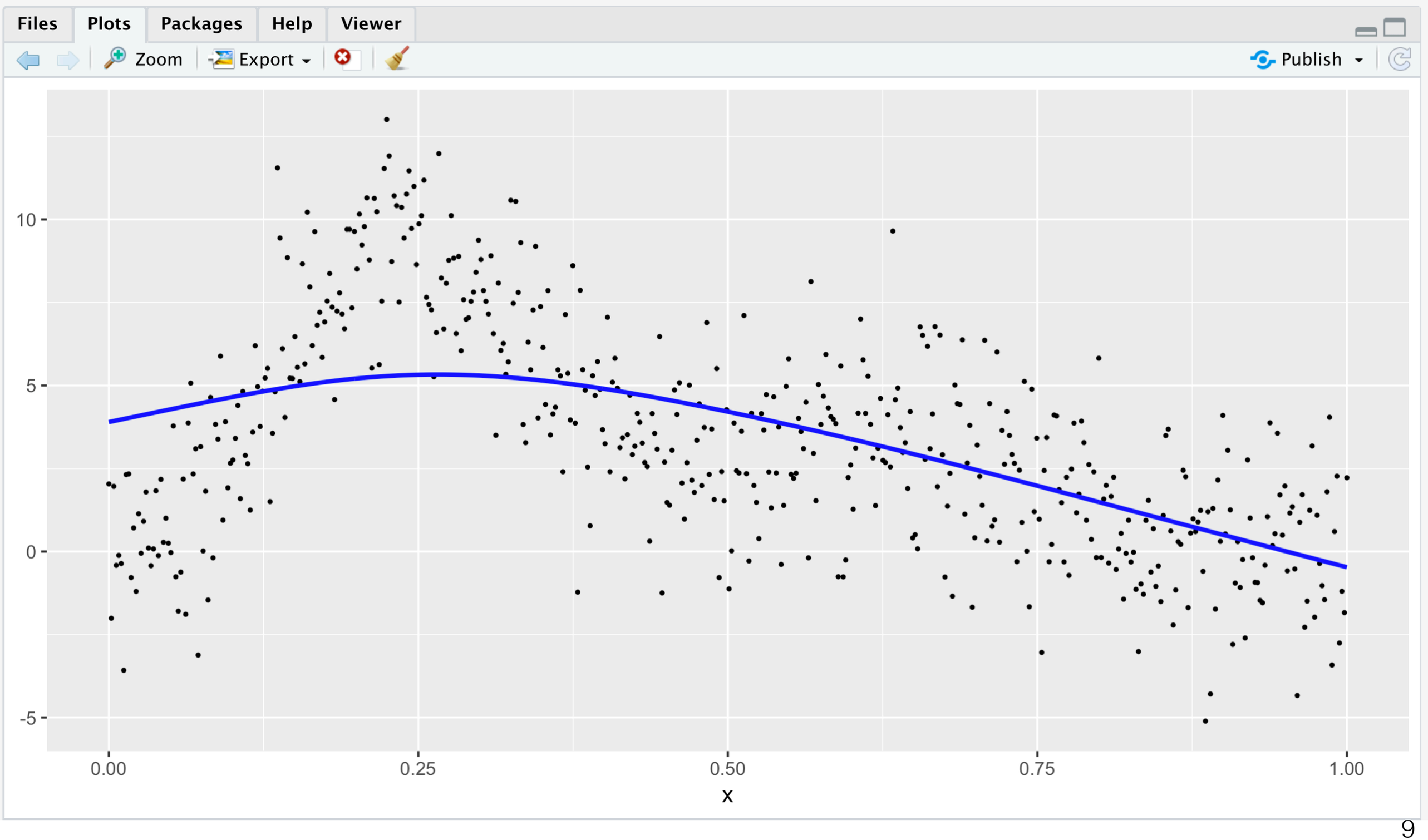
Given Data (x, y) from unknown $f(x)$



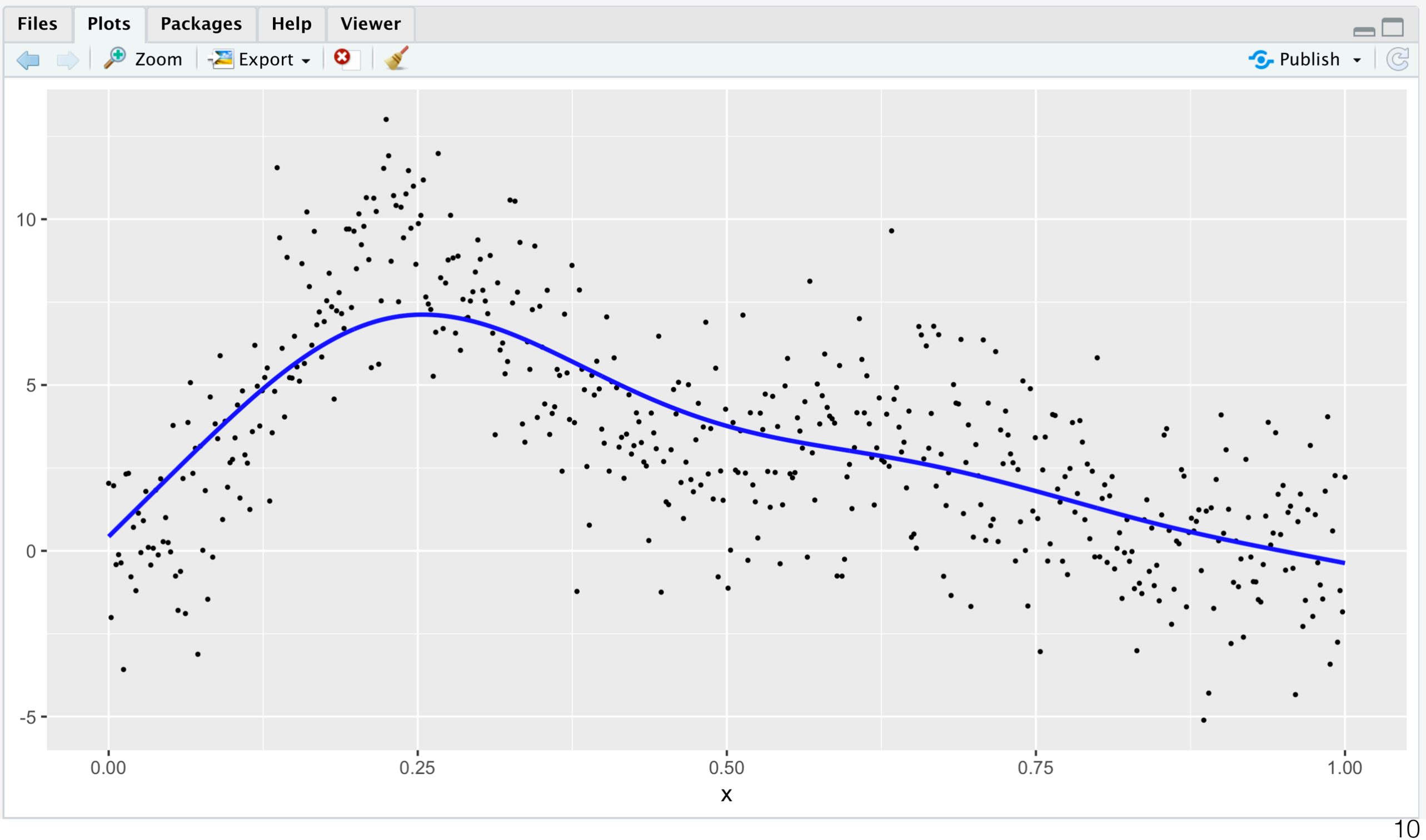
Approximate (i.e. "fit") a Model $\hat{f}(x)$



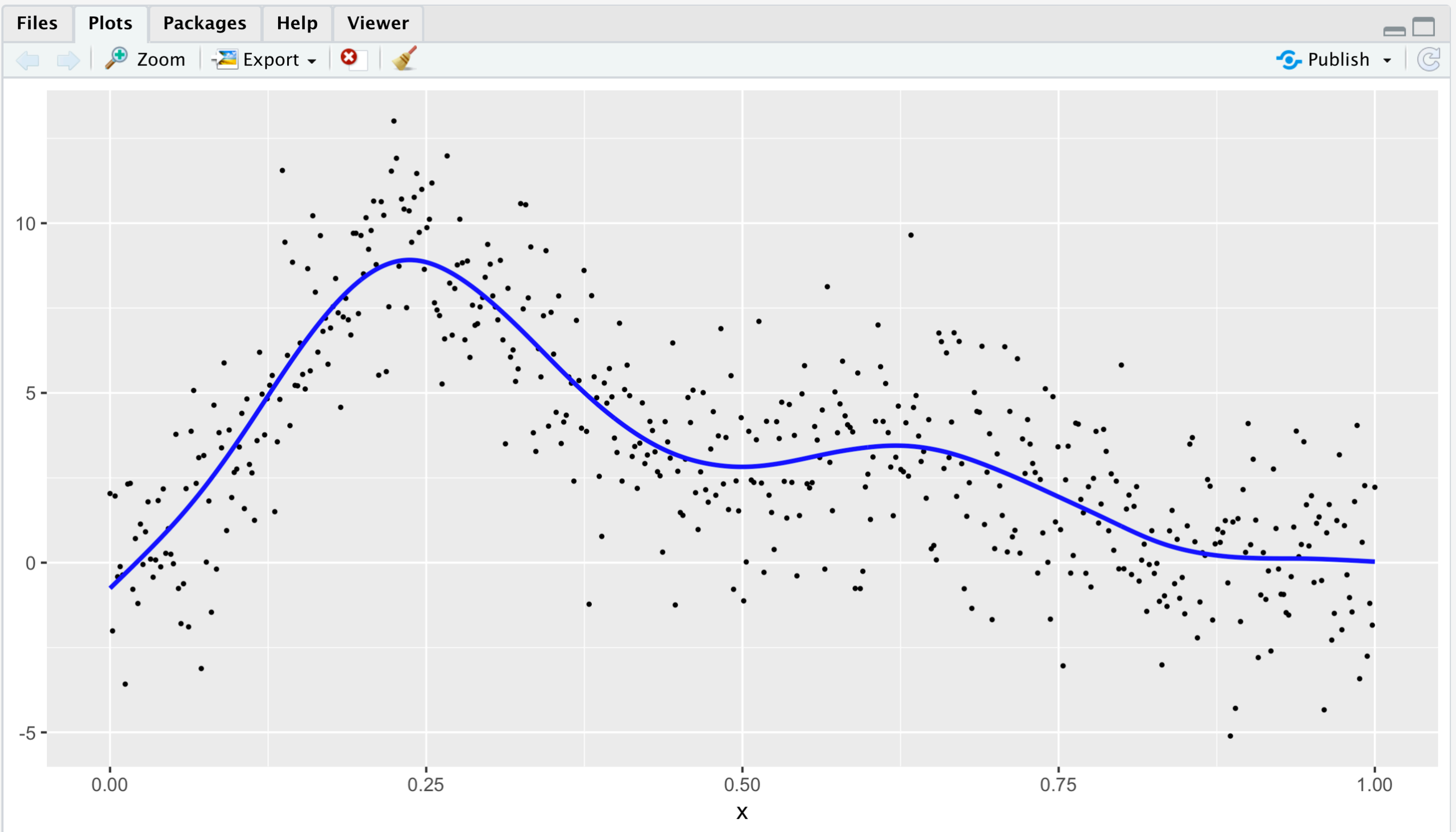
How about this $\hat{y} = \hat{f}(x)$?



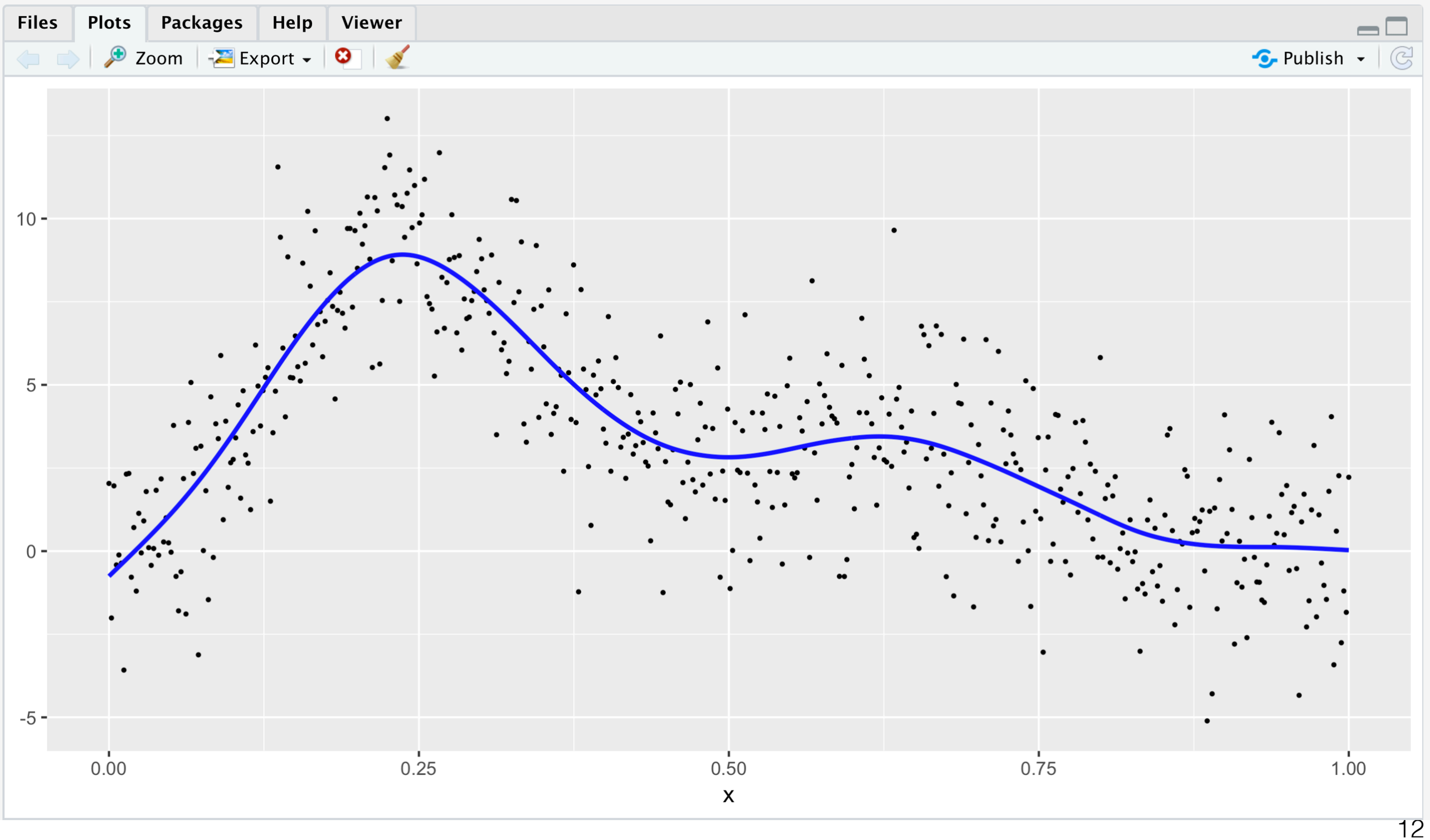
How about this $\hat{f}(x)$?



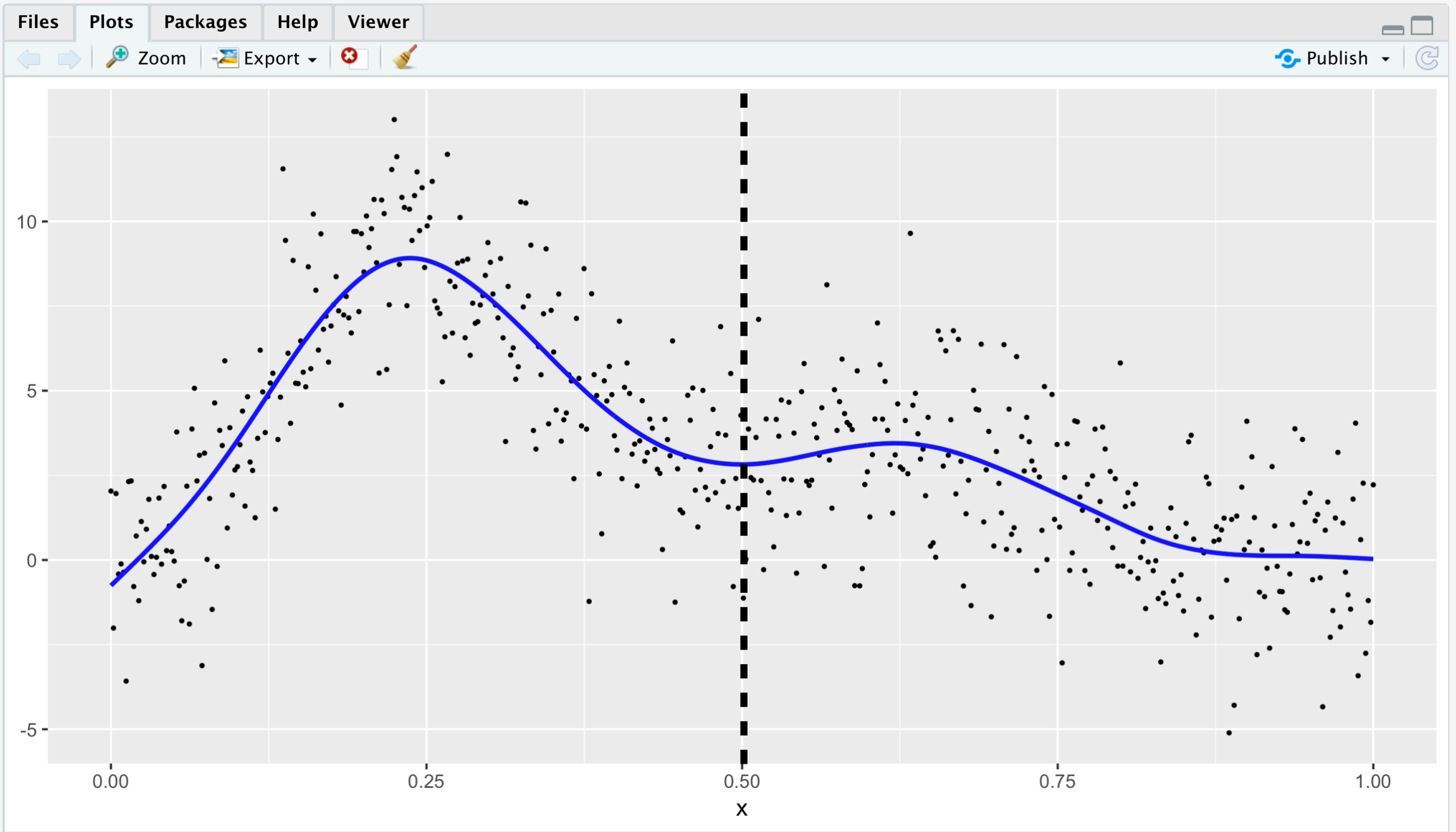
How about this $\hat{f}(x)$?



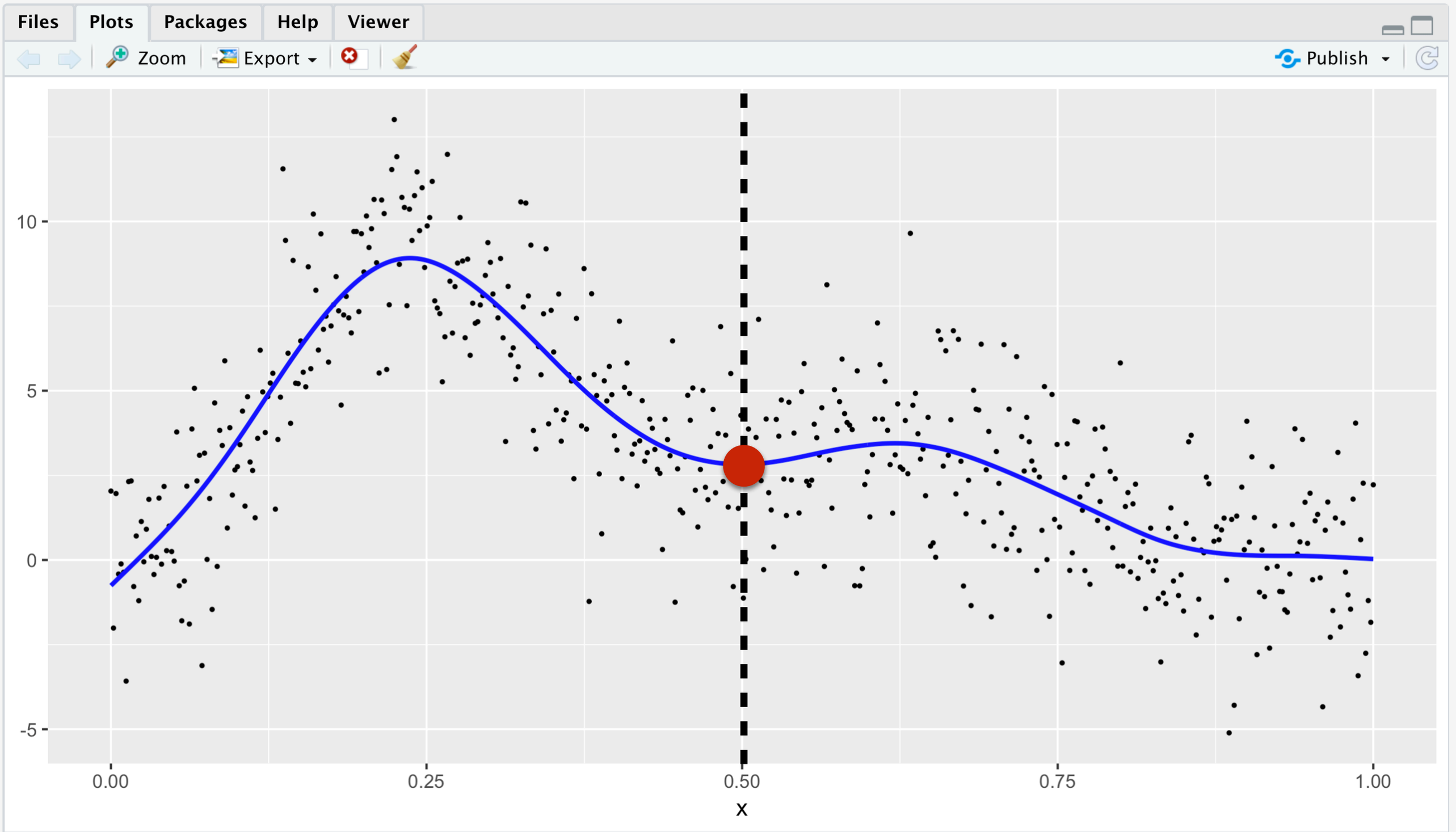
What does this $\hat{f}(x)$ predict for $x = 0.5$?



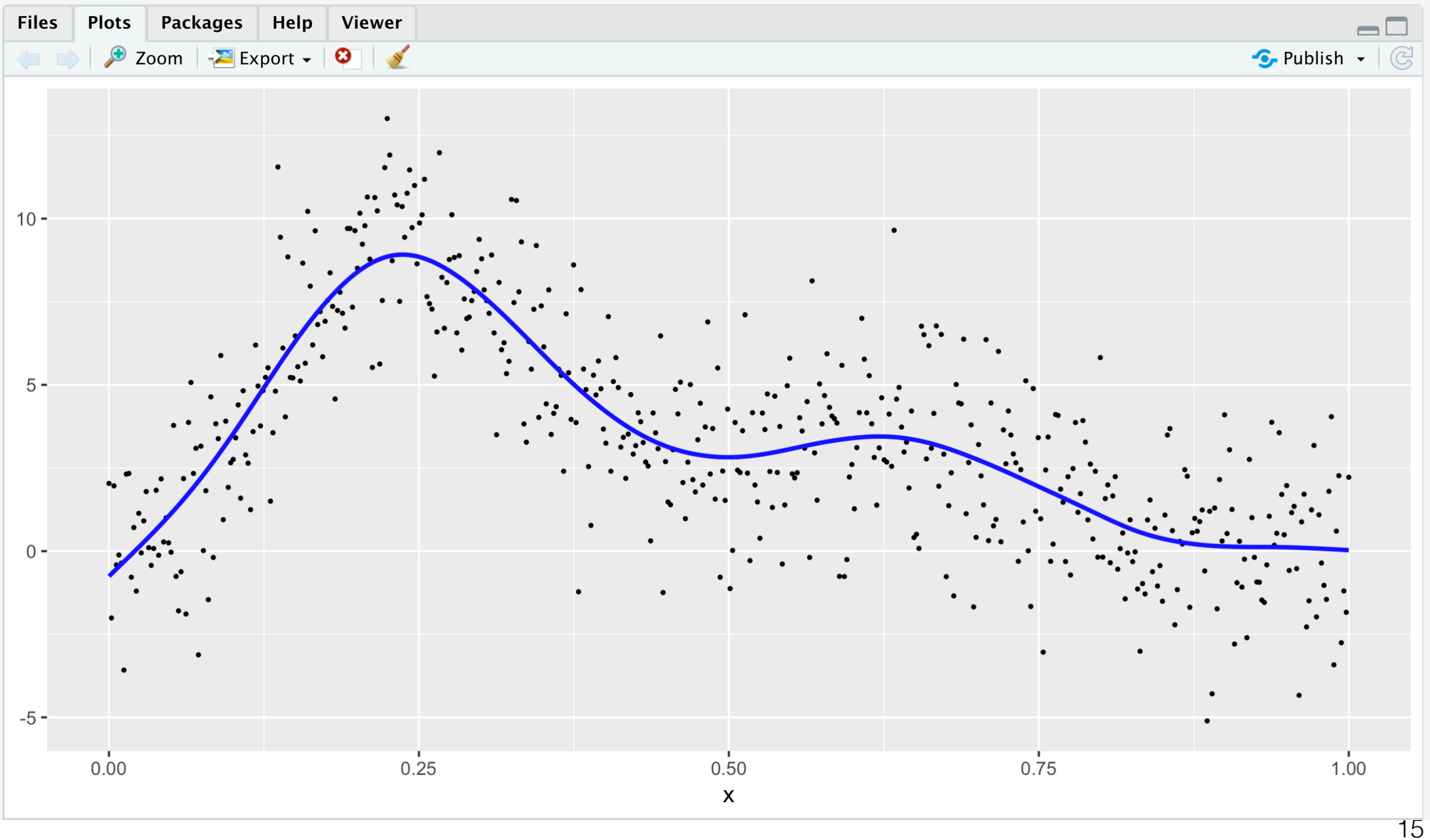
What does this $\hat{f}(x)$ predict for $x = 0.5$?



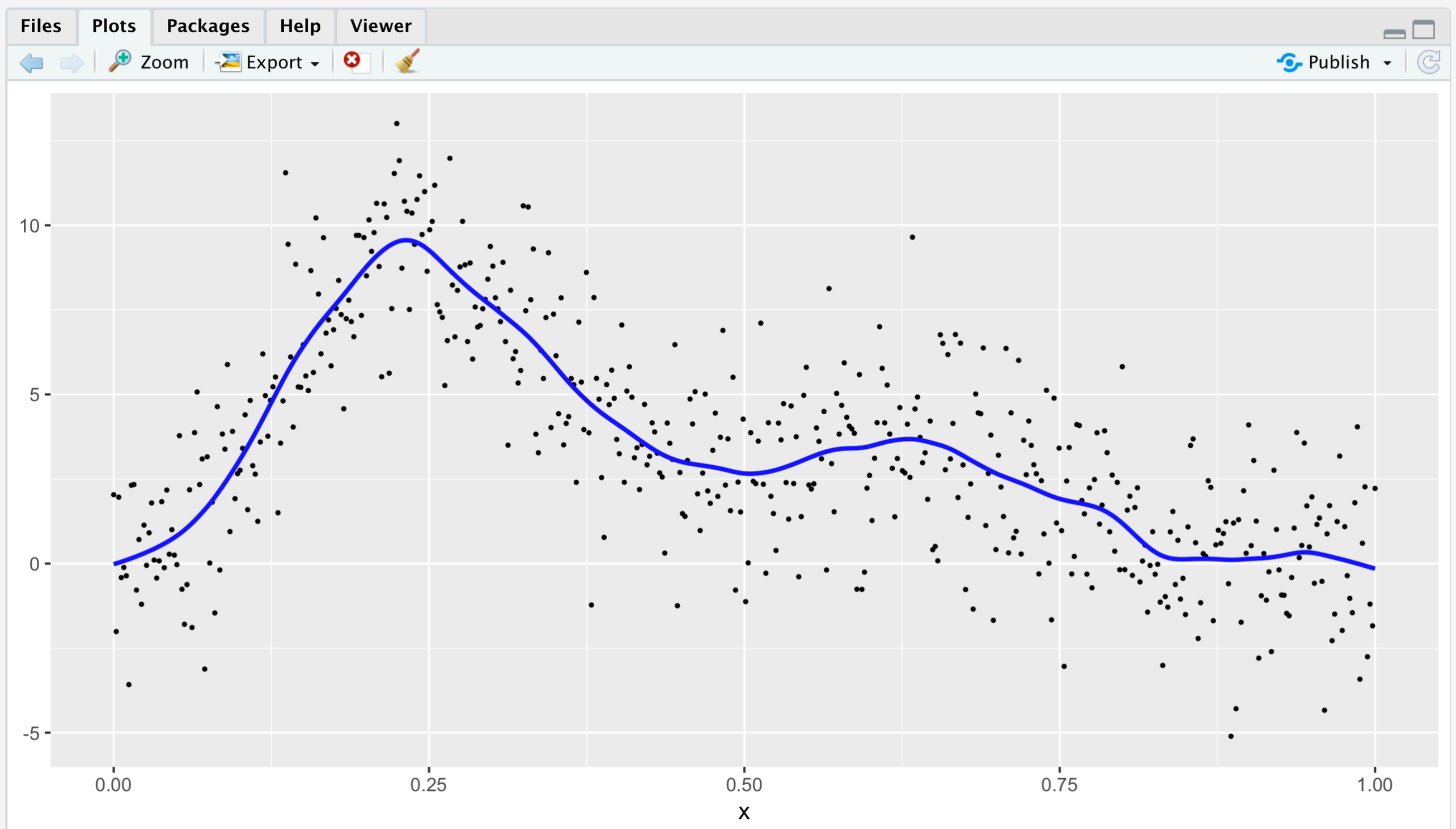
What does this $\hat{f}(x)$ predict for $x = 0.5$?



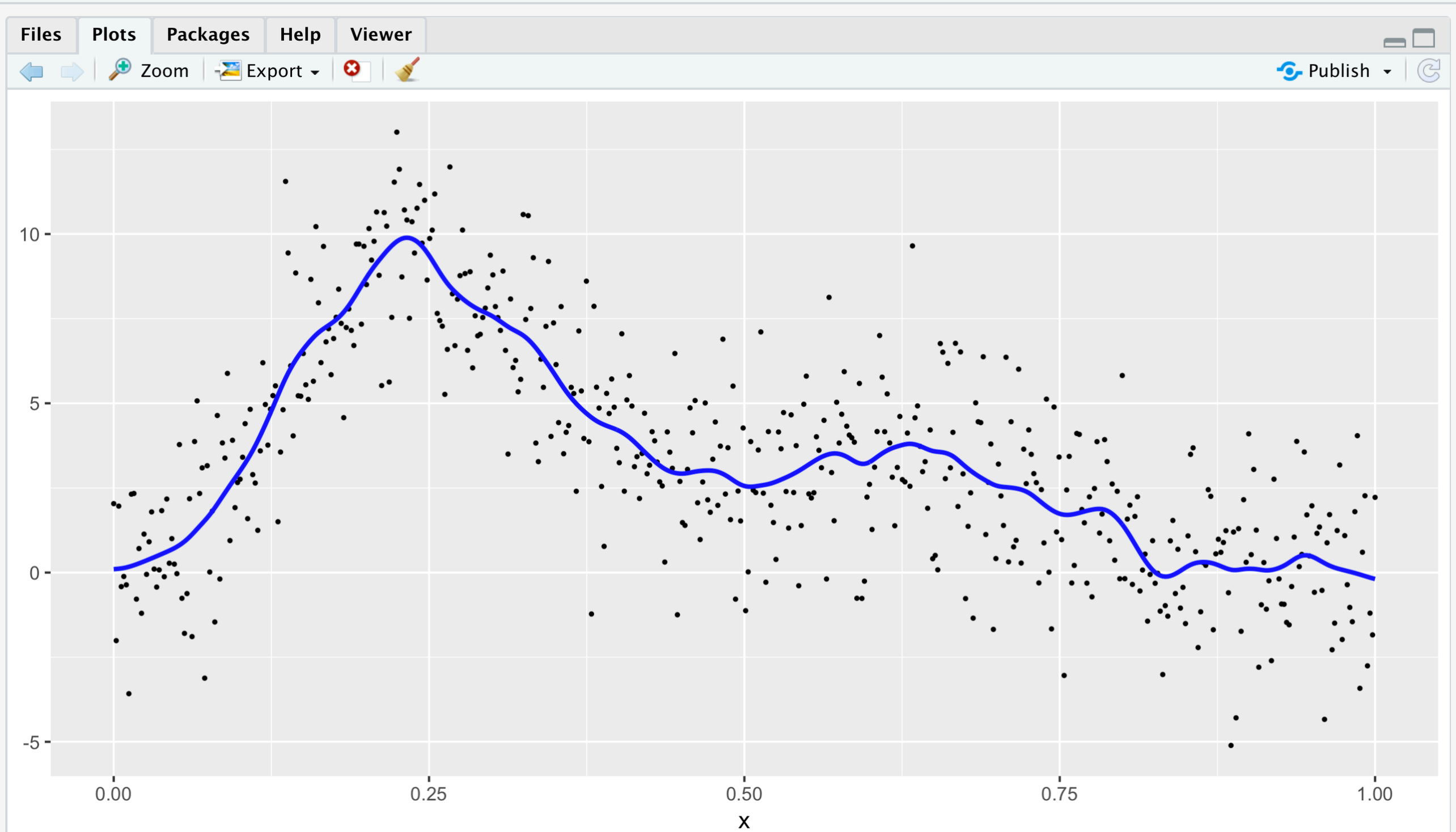
Ok, great. But instead of this $\hat{f}(x)$ …



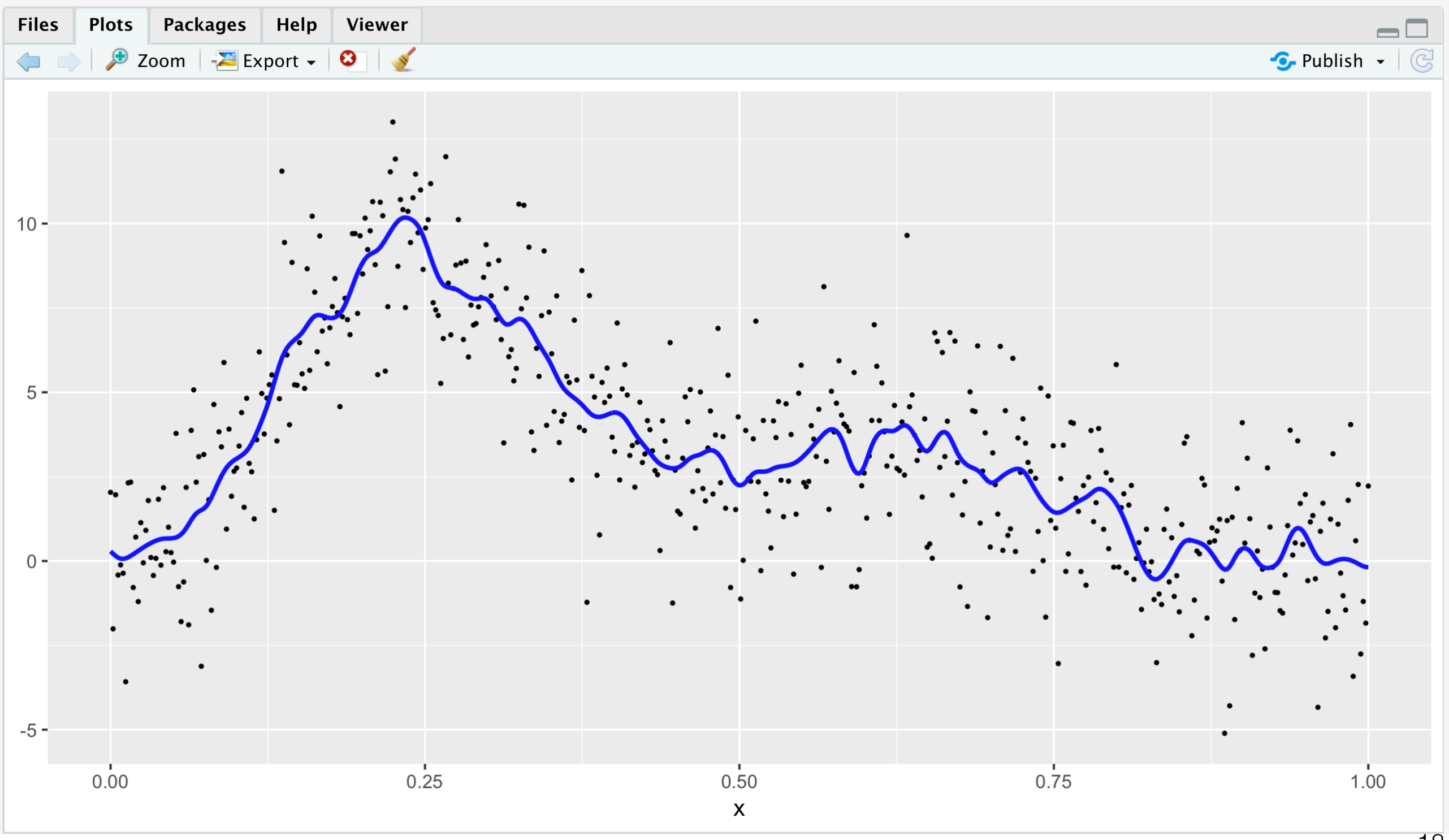
How about this $\hat{f}(x)$?



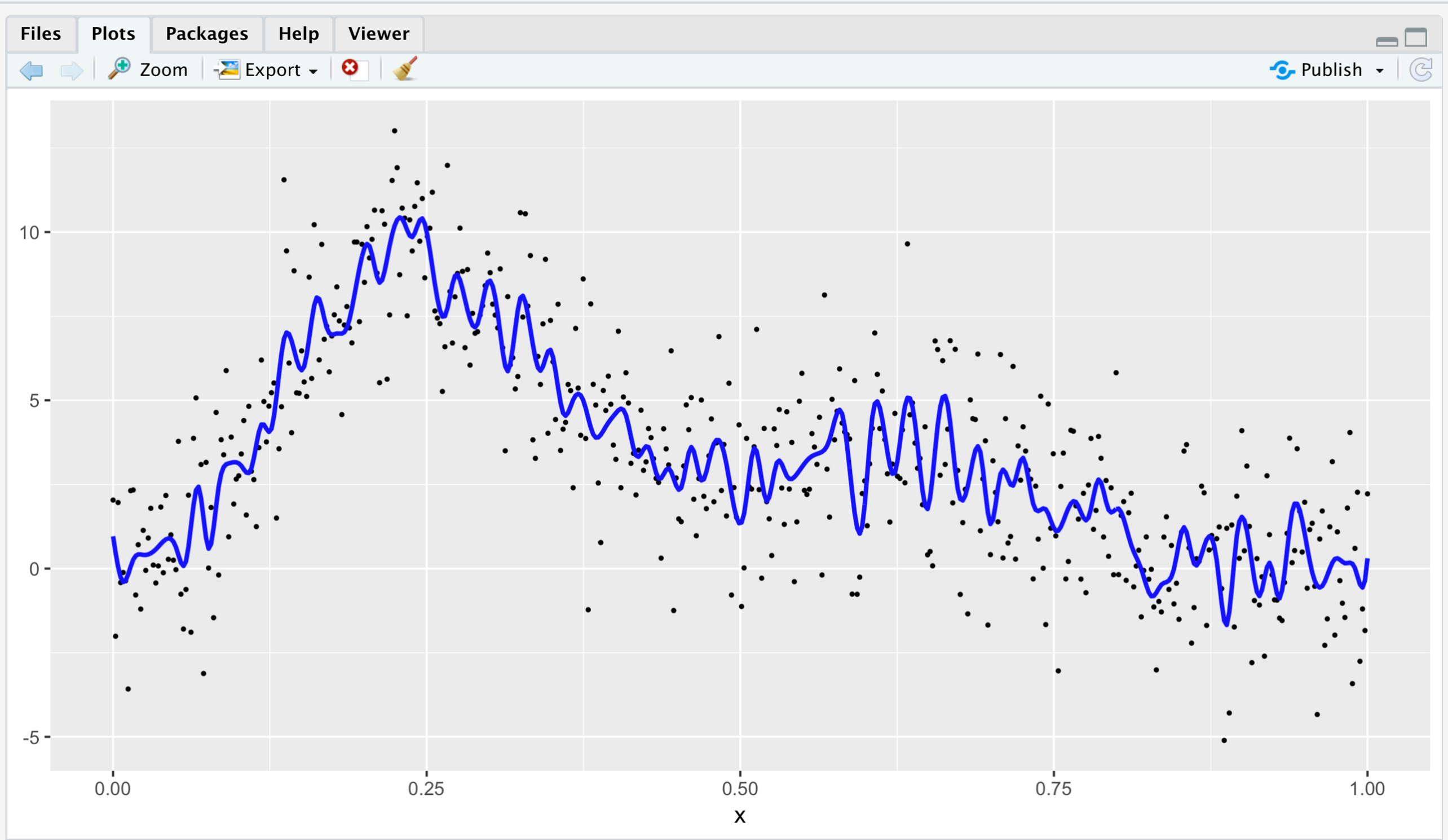
How about this $\hat{f}(x)$?



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



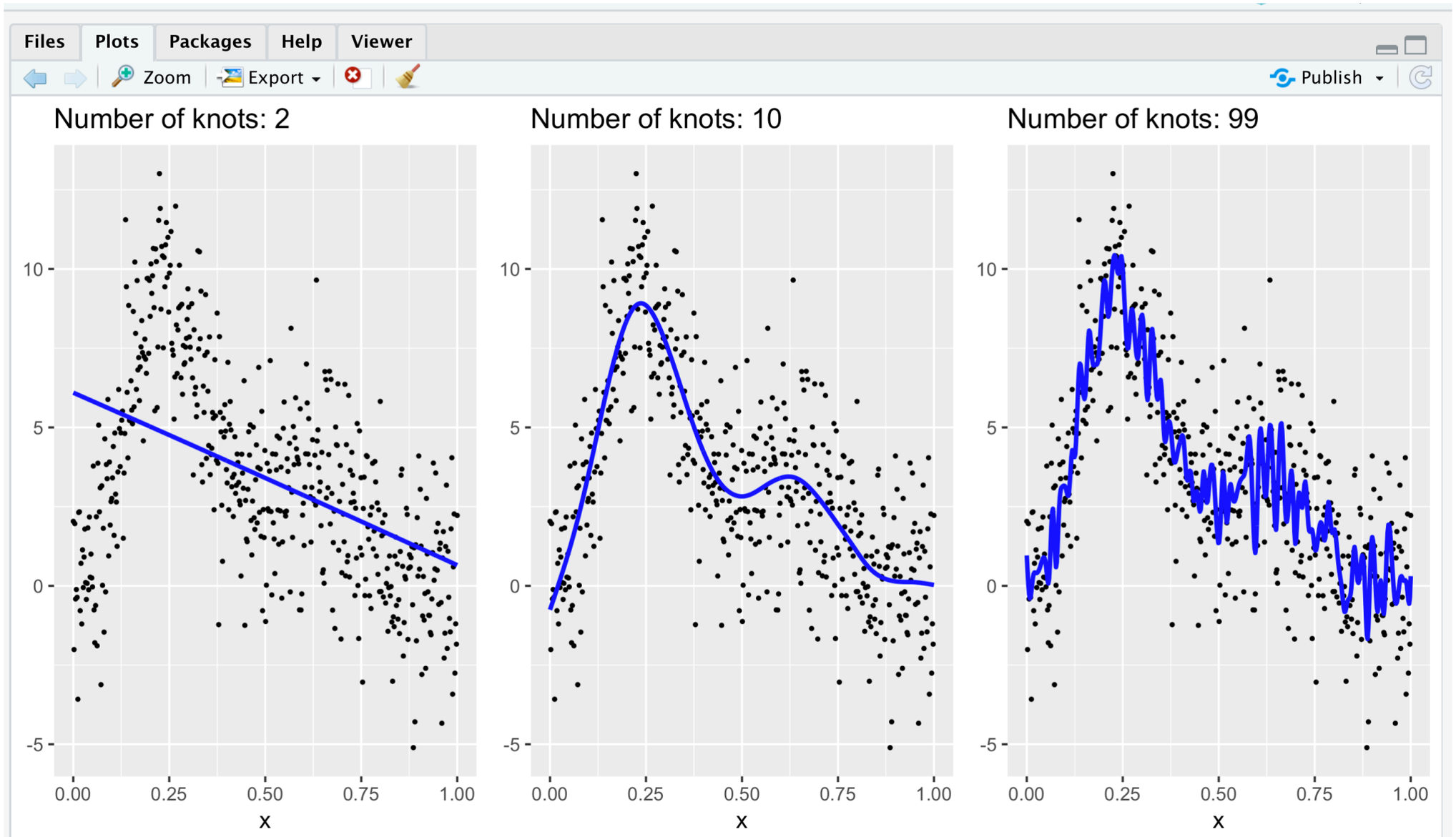
How about this $\hat{f}(x)$?



Model Fitting Method: (Cubic) Splines

- Splines use linear algebra to find the blue curve $\hat{f}(x)$ that **minimizes** the (squared) vertical distances between:
  - the predicted $\hat{y} = \hat{f}(x)$
  - the observed y
- Amount of “wiggle” is dictated by user using the "number of knots”
- In other words, "number of knots” controls the **complexity of the model**

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



Underfit!

"Just right!"

Overfit!

Road Map

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

→

Forest Ecology:
- Our data
- Two models for tree growth

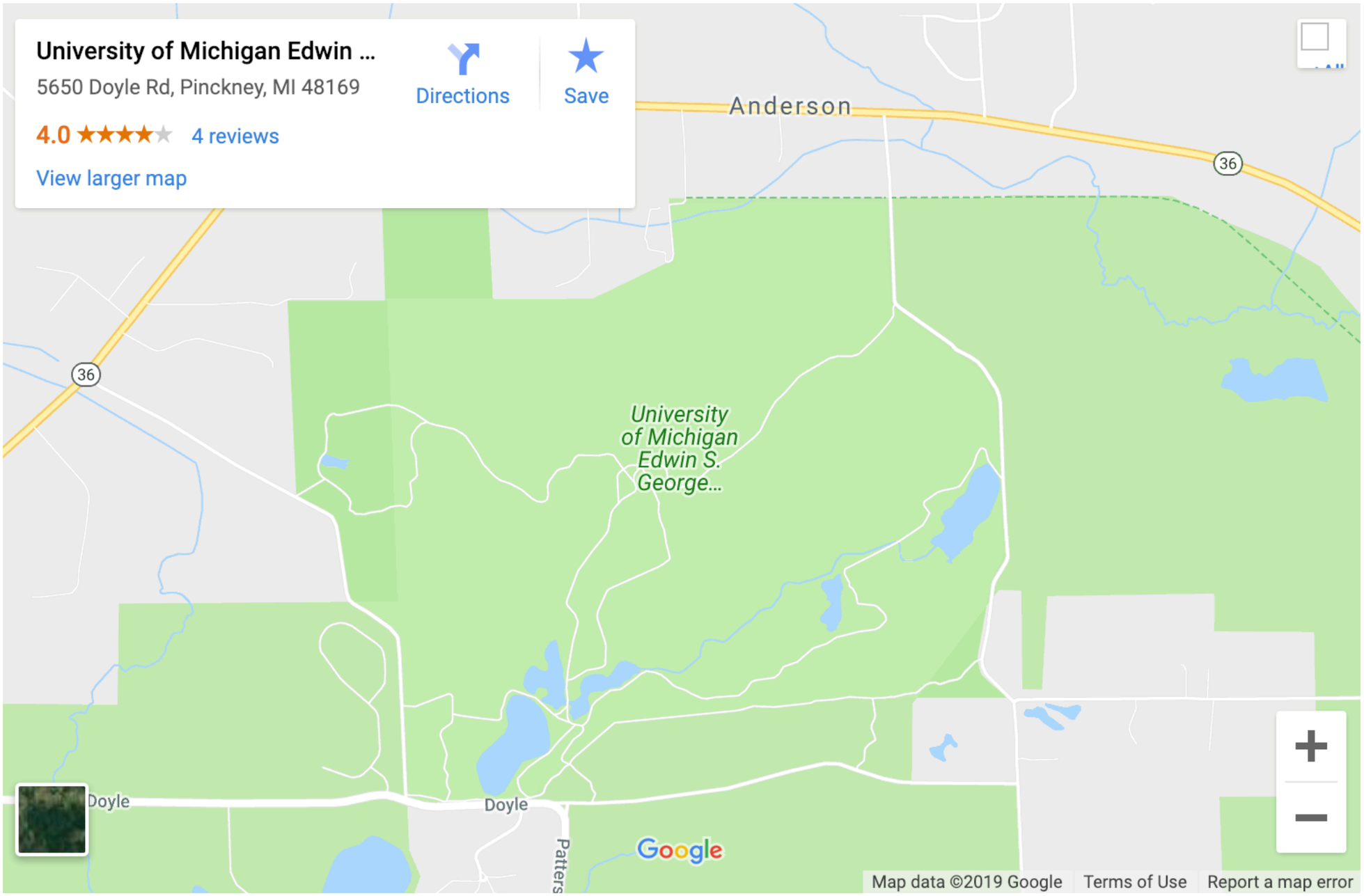
↓

Machine Learning + Forest Ecology:
- Goal restated
- Assessment metric used
- Model assessment with crossvalidation

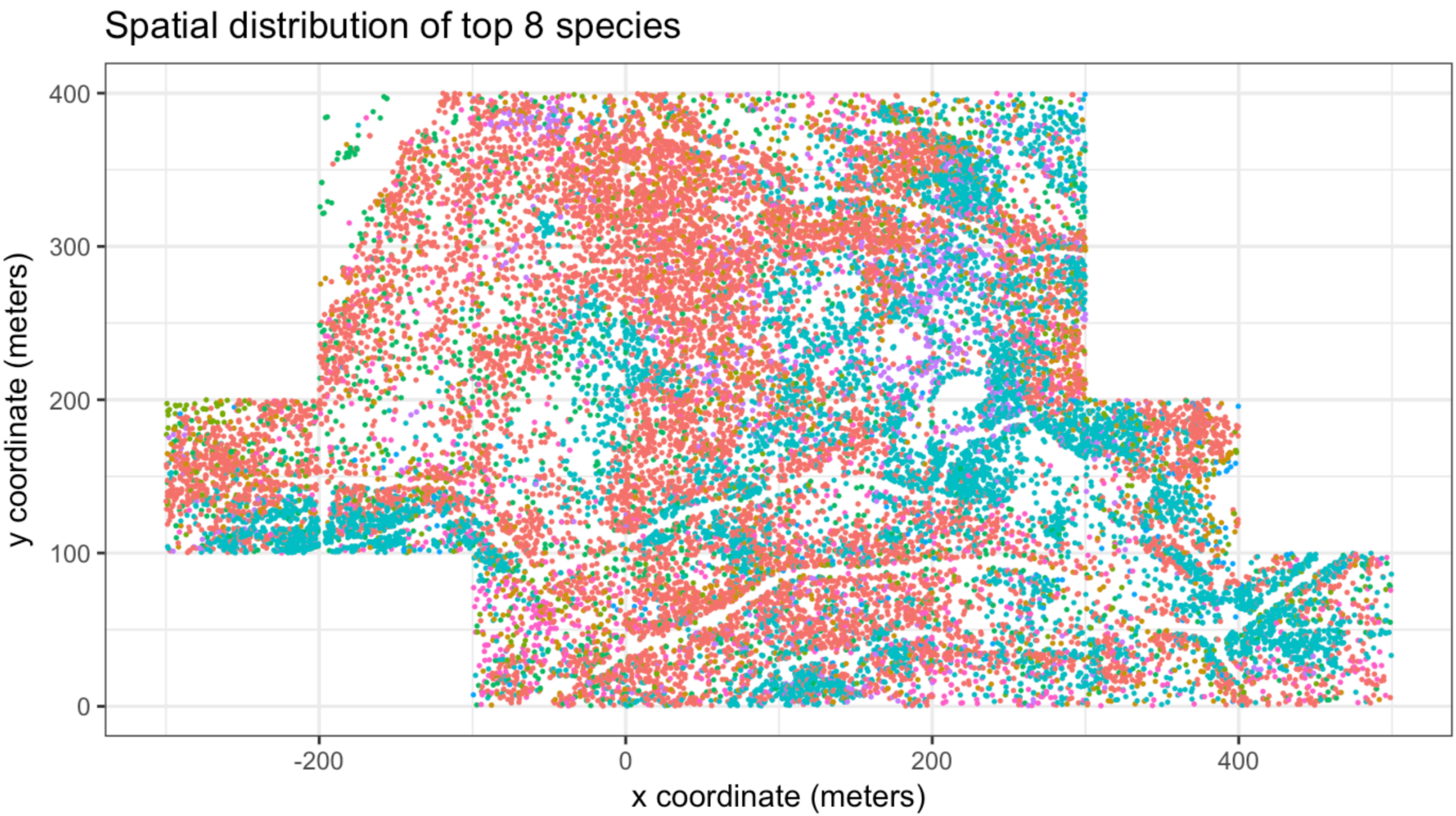
↓

Conclusion:
- Resources to learn more
- Thoughts on modeling

Data: 2008 & 2014 Censuses of Trees



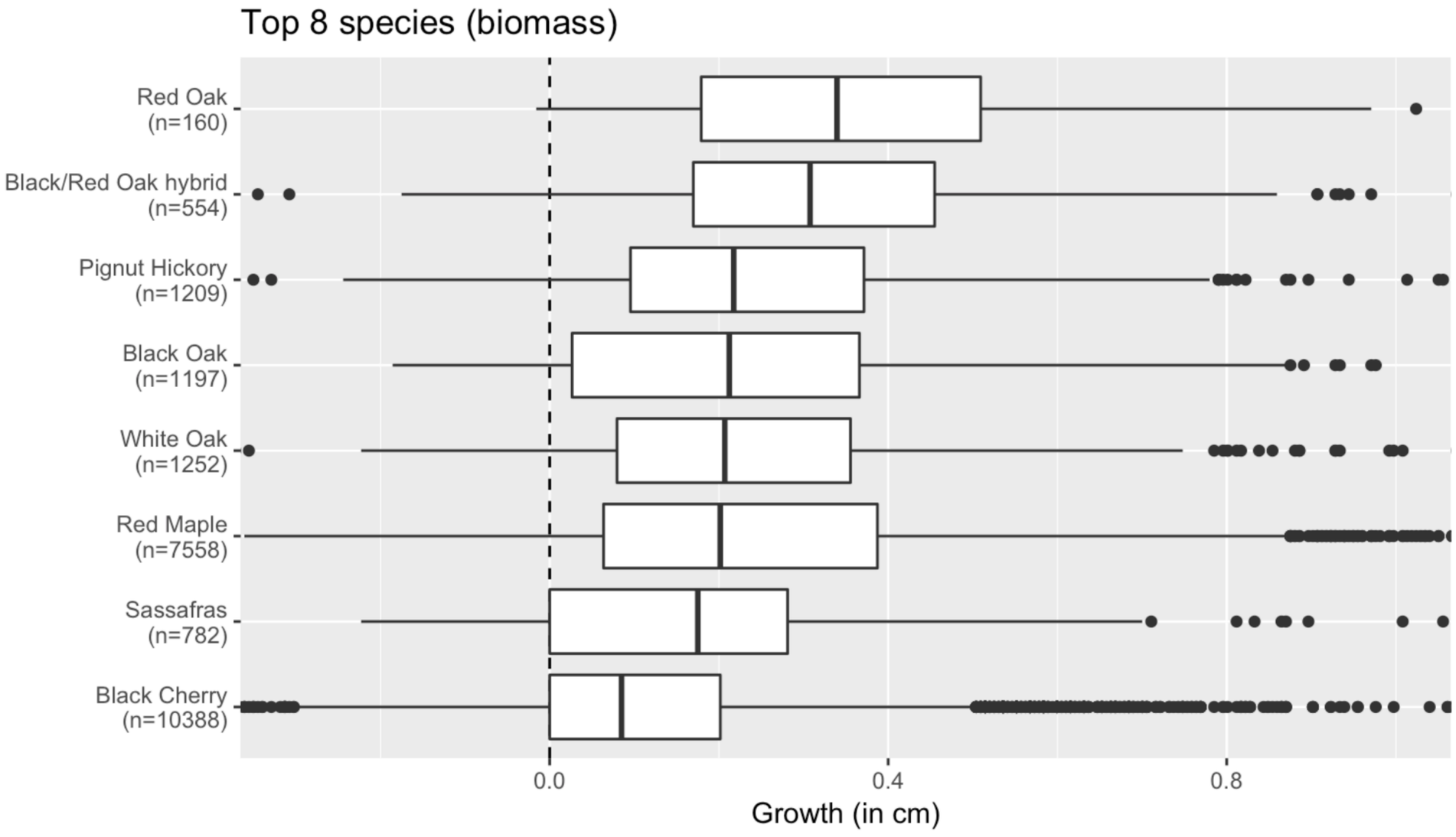
Data: 2008 Snapshot



Recall our Variables!

y : Outcome Variable = Avg Annual Growth

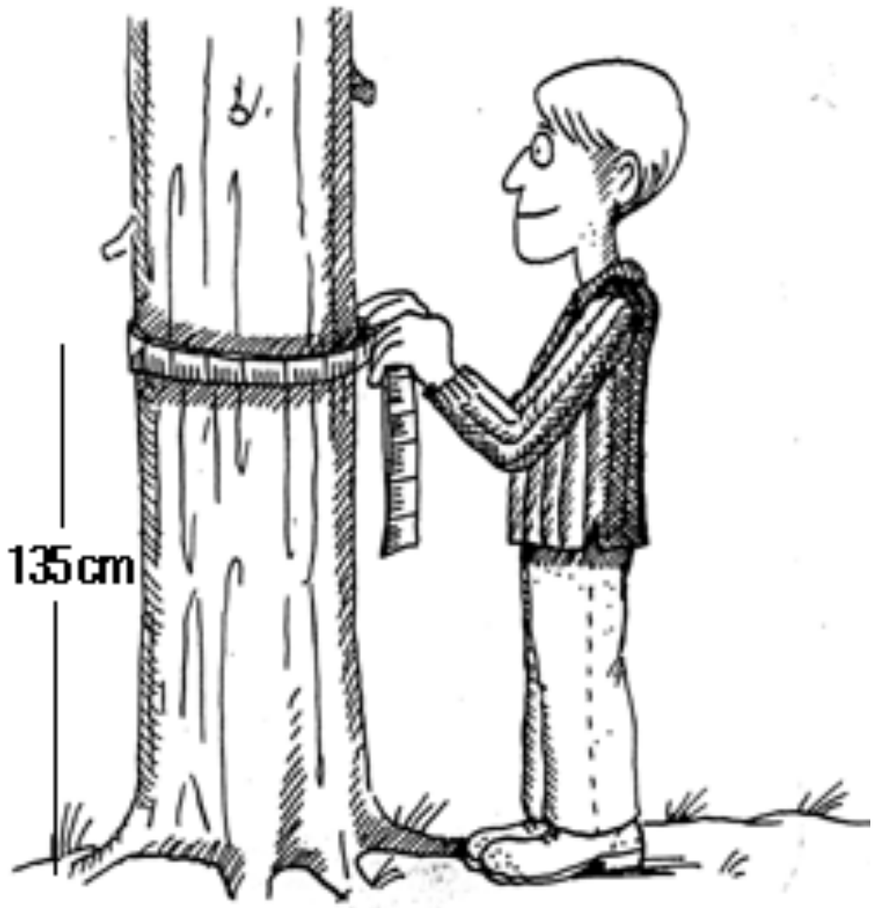
Observed average annual growth of trees 2008-2014



Predictor Variables $\overrightarrow{x}$

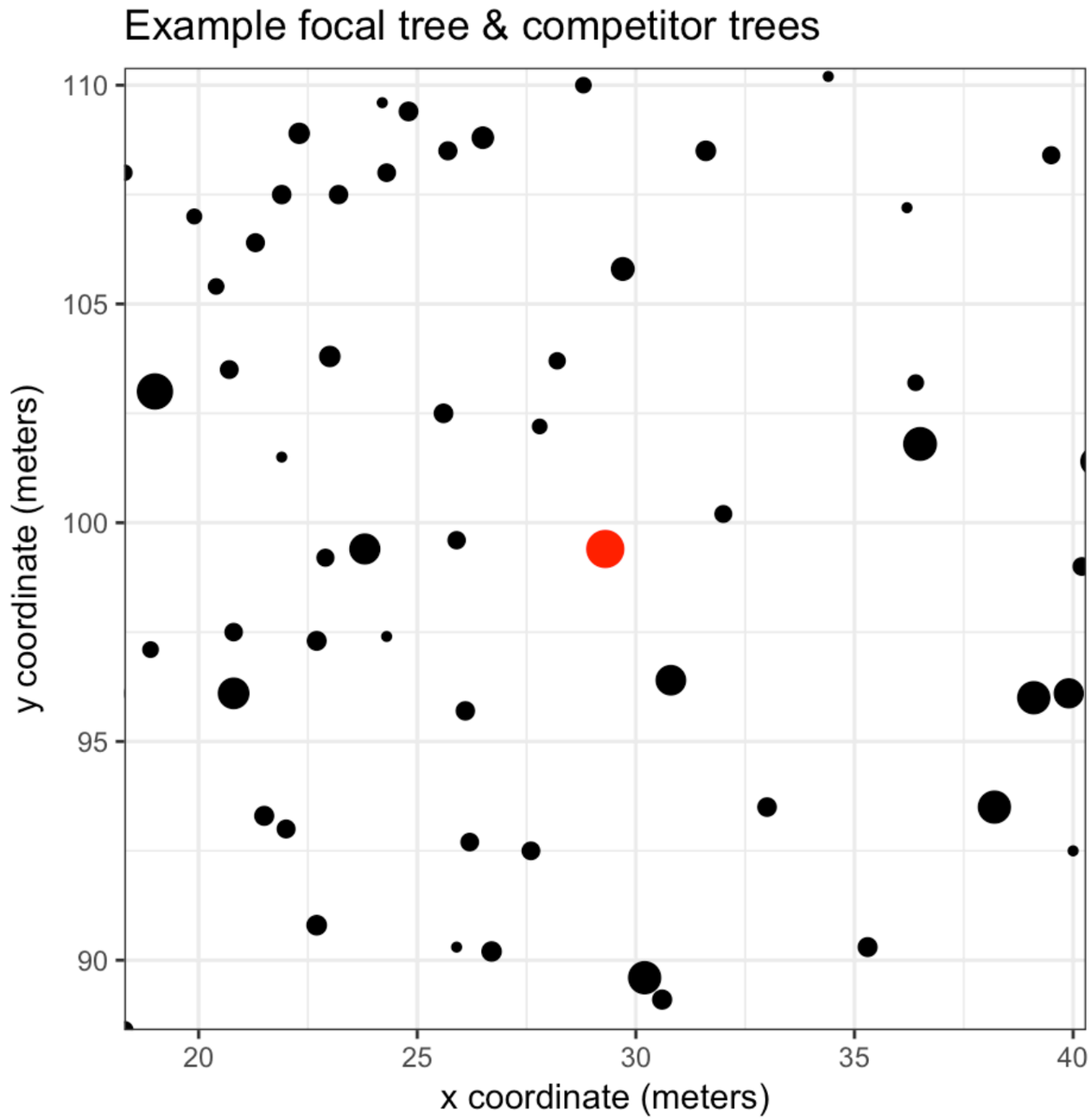
x_1 : Species of tree

x_2 : Size of tree (diameter at breast height)

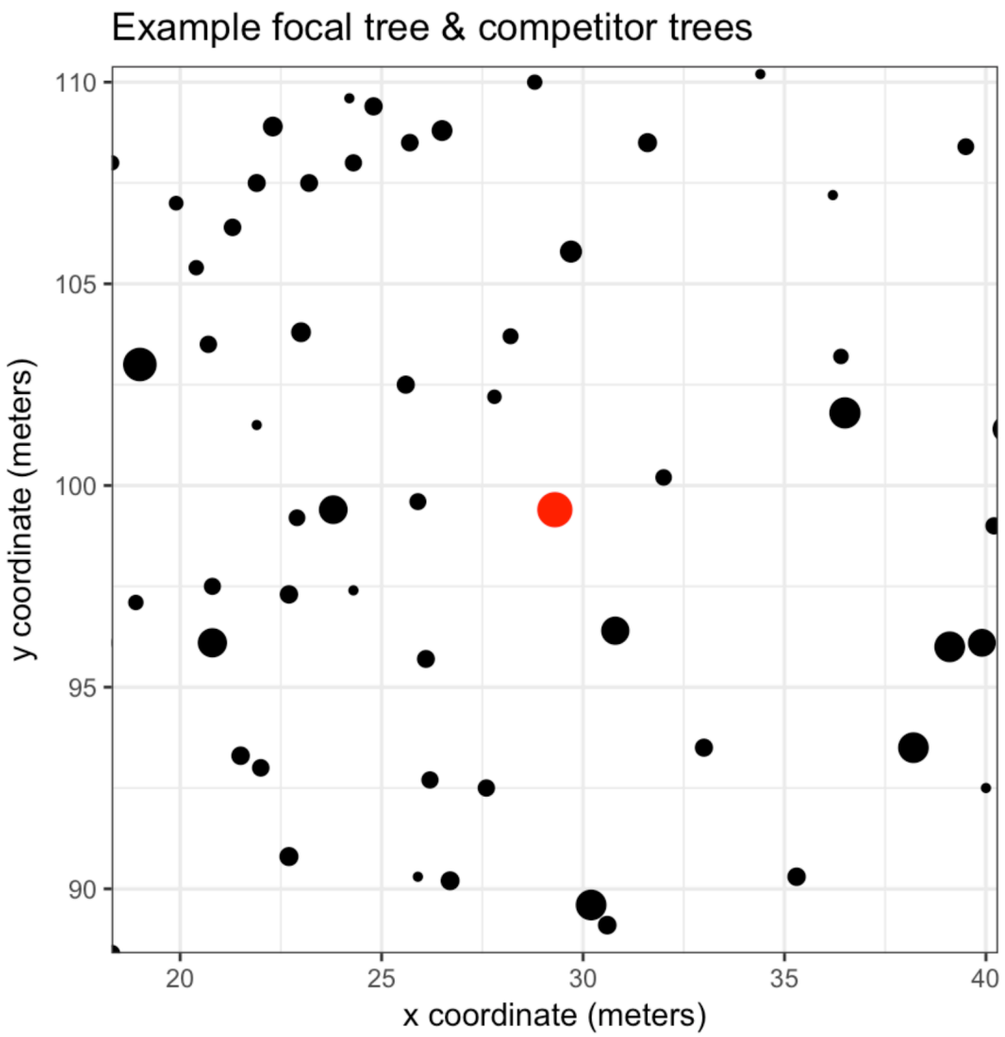


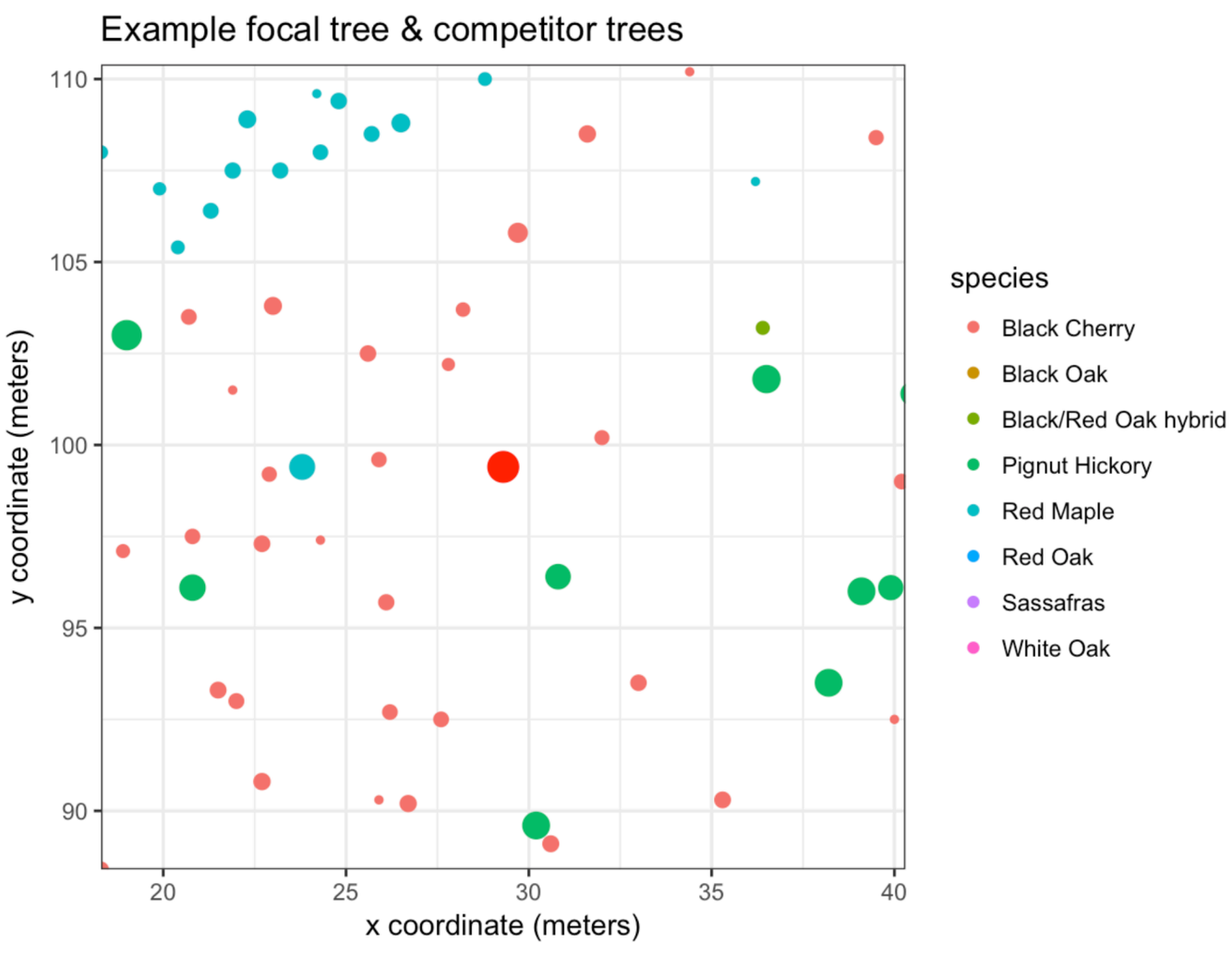
Predictor Variables

x_3 : Number and size of competitor trees (biomass)



Two Models of Competition





Which model is better?
Yea or nay on distinguishing competitor species?

Road Map

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

Forest Ecology:
- Our data
- Two models for tree growth

Machine Learning + Forest Ecology:
- Goal restated
- Assessment metric used
- Model assessment with crossvalidation

Conclusion:
- Resources to learn more
- Thoughts on modeling

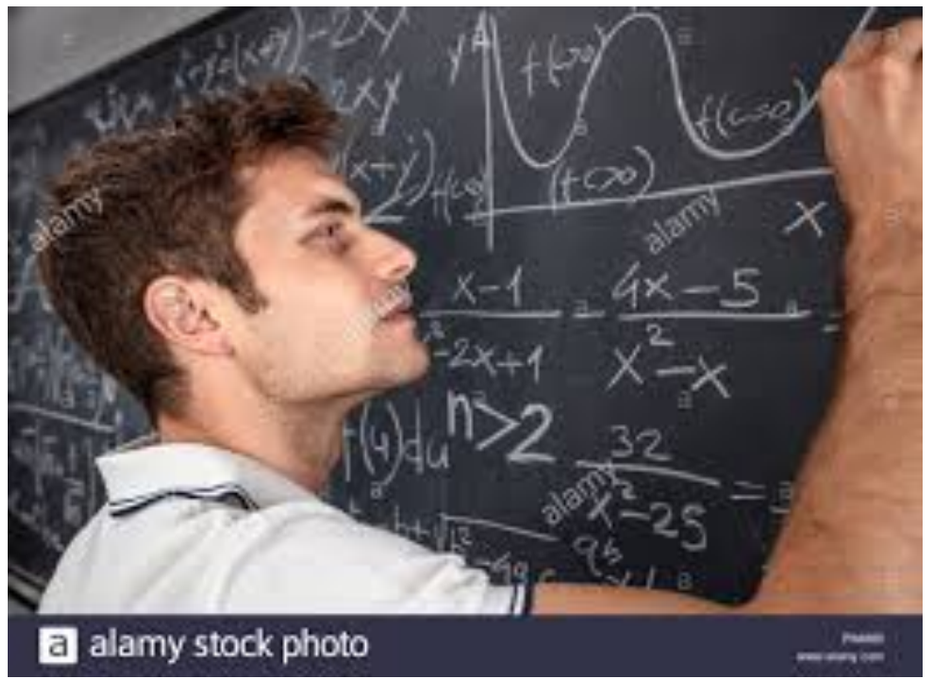
Machine Learning & Forest Ecology

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

Model Assessment Metric

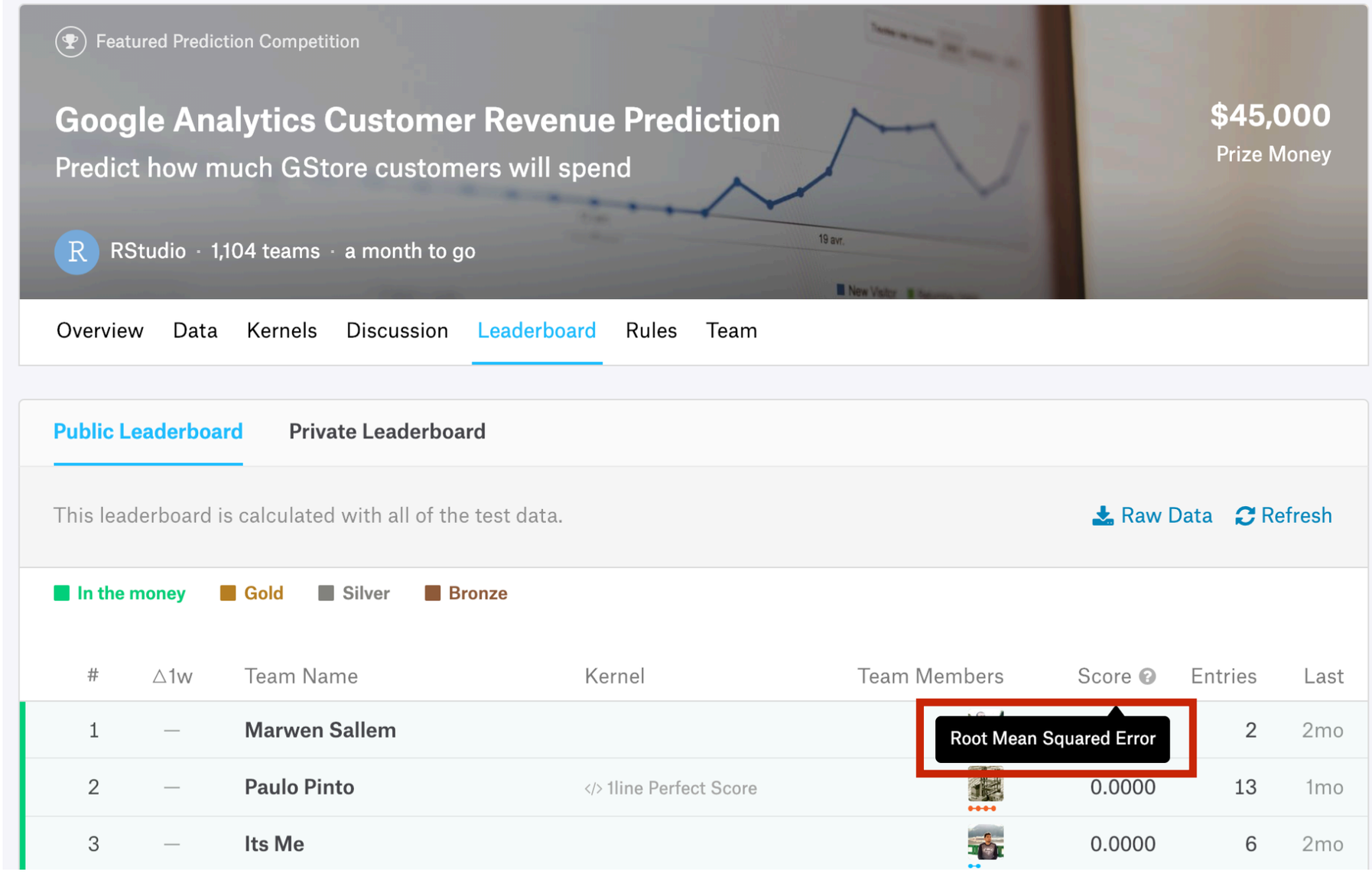
- Question: “How good is our model?”
- Answer: “This is answered using the **Mean Square(d) Error** metric!”

Back to the blackboard for Chalk Talk #2…

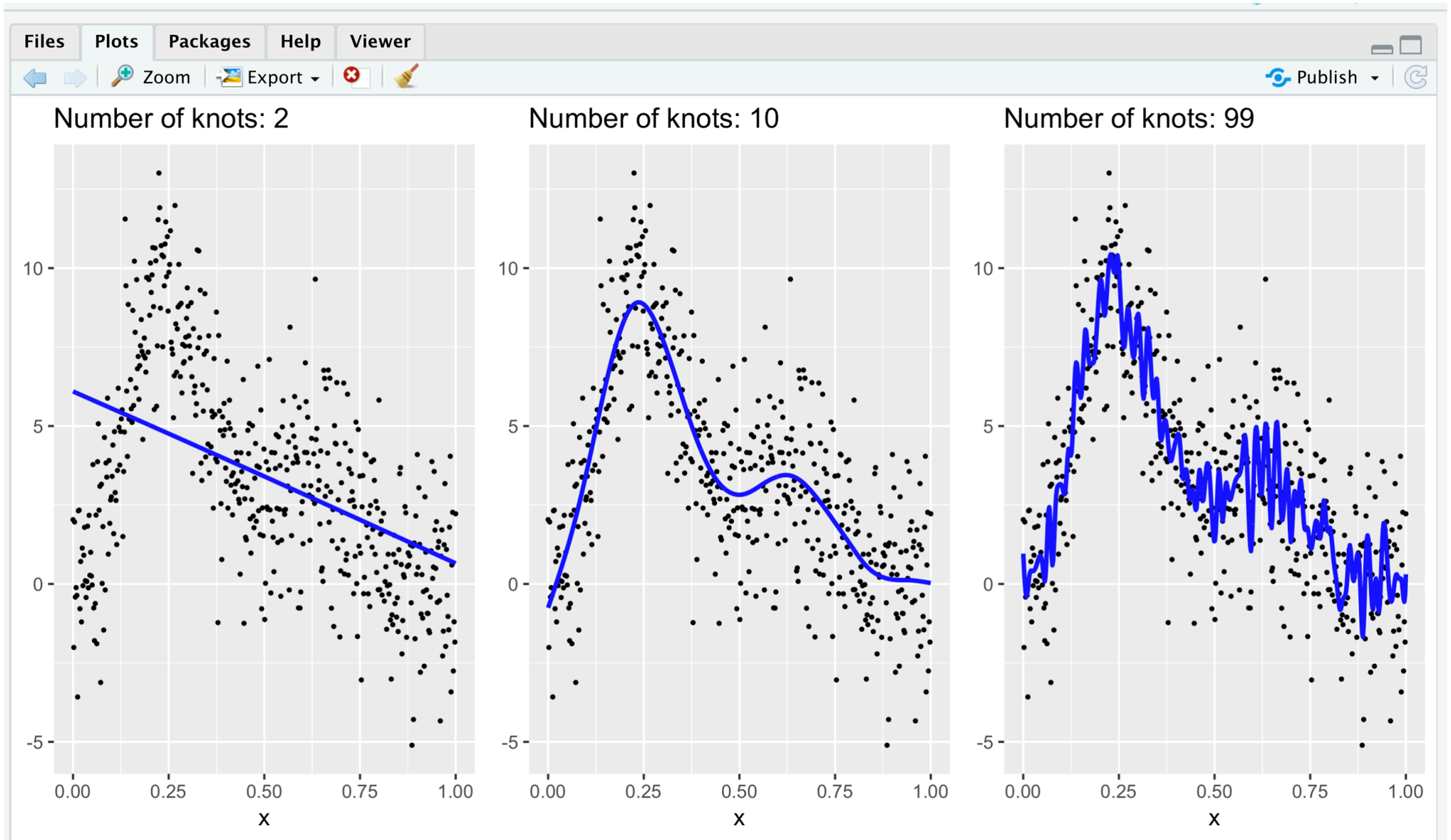


Mean Squared Error

On Machine Learning predictive modeling competition site [Kaggle]():



Hold up! What about underfitting vs overfitting?



Underfit!

"Just right!"

Overfit!

How? Using Validation Set Approach

Fit your model on ***training*** data

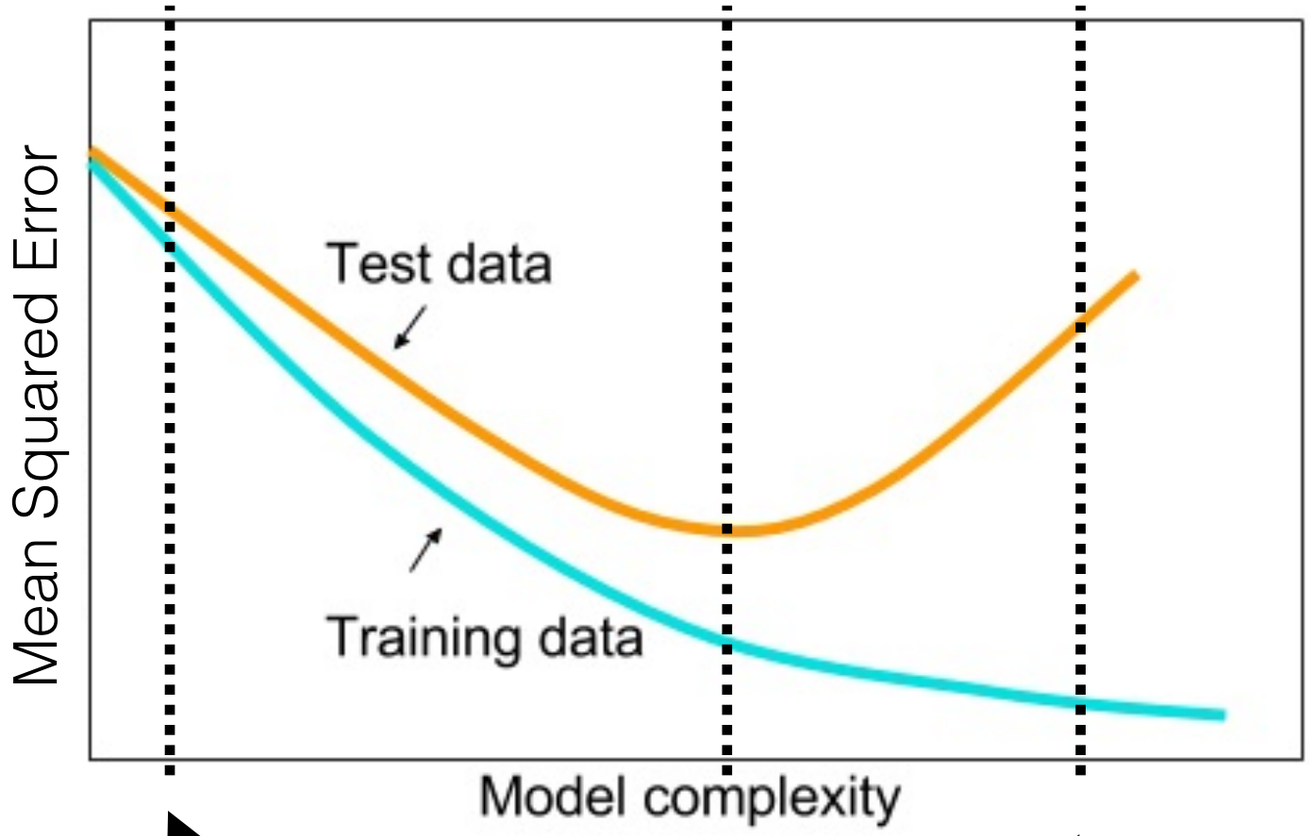
Assess your model on ***test*** data

One last time to blackboard for Chalk Talk #3…

Typical Mean Square Error Performance

Fit your model on the ***training*** data

Assess your model's MSE on the ***test*** data

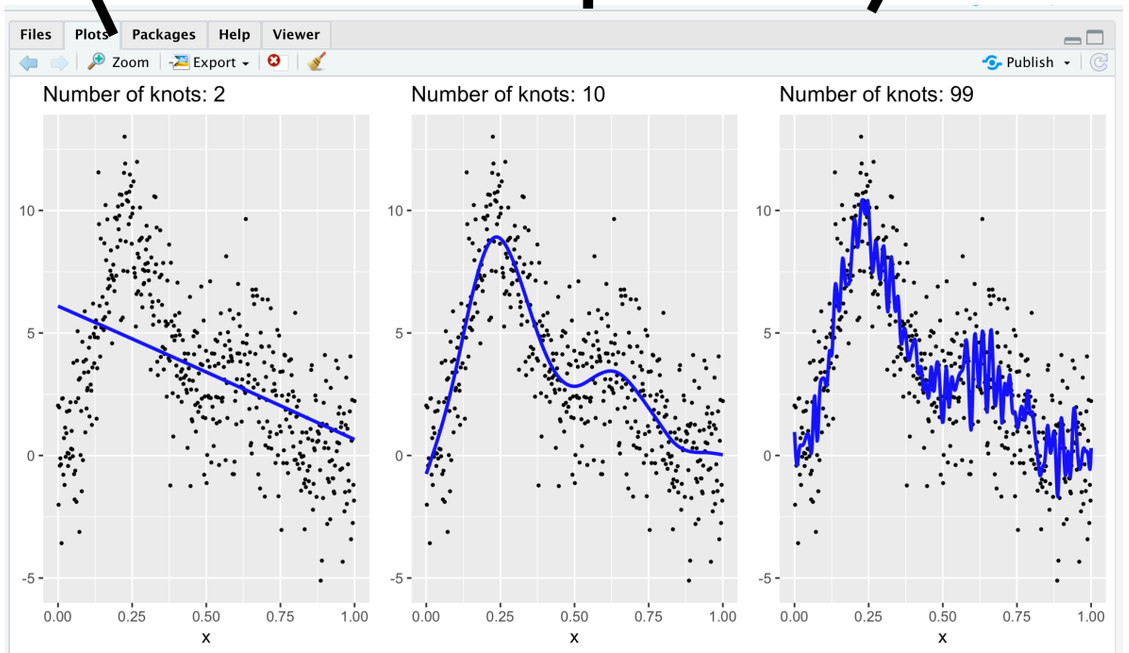


Underfit!

"Just right!"

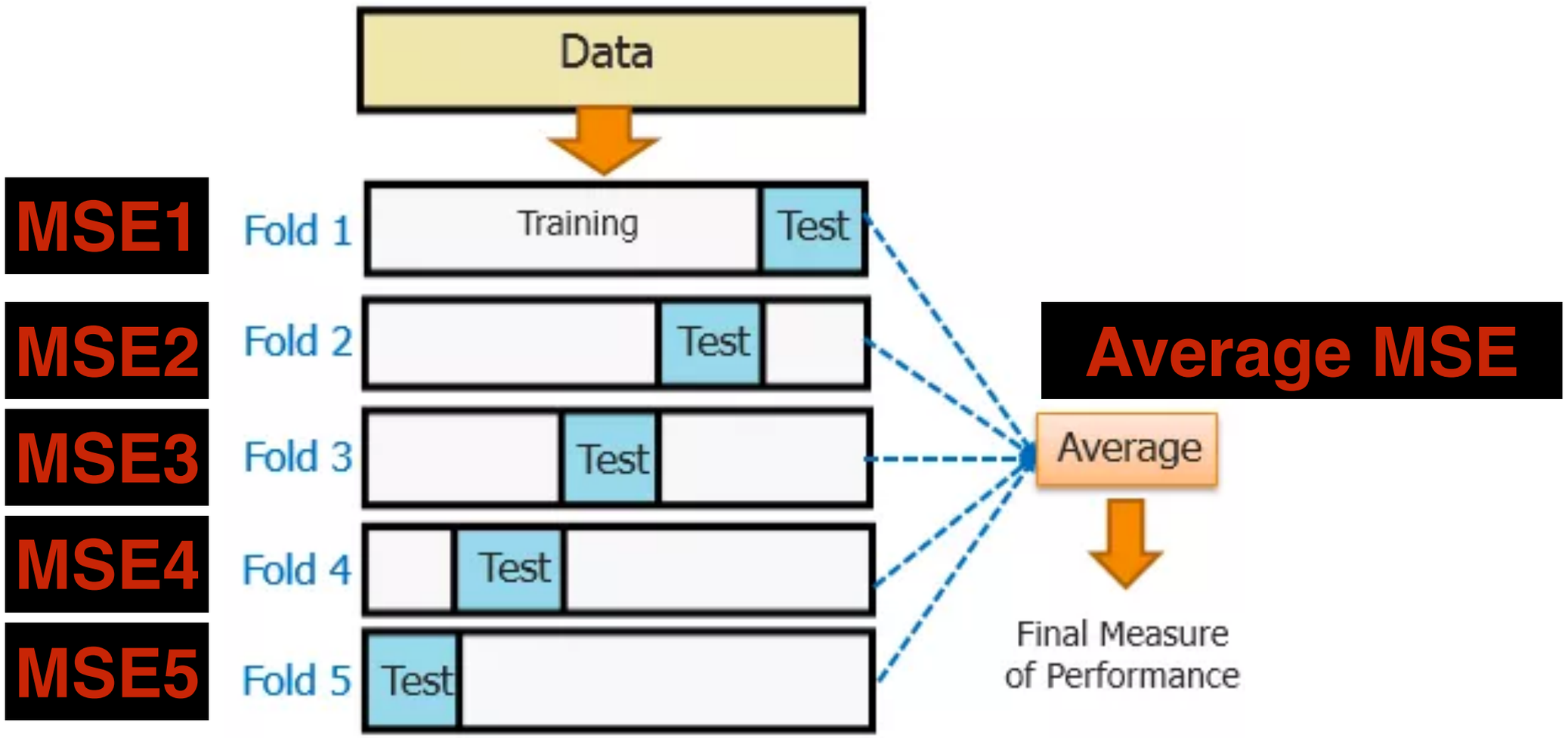
Overfit!

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
- Two models for tree growth

↓

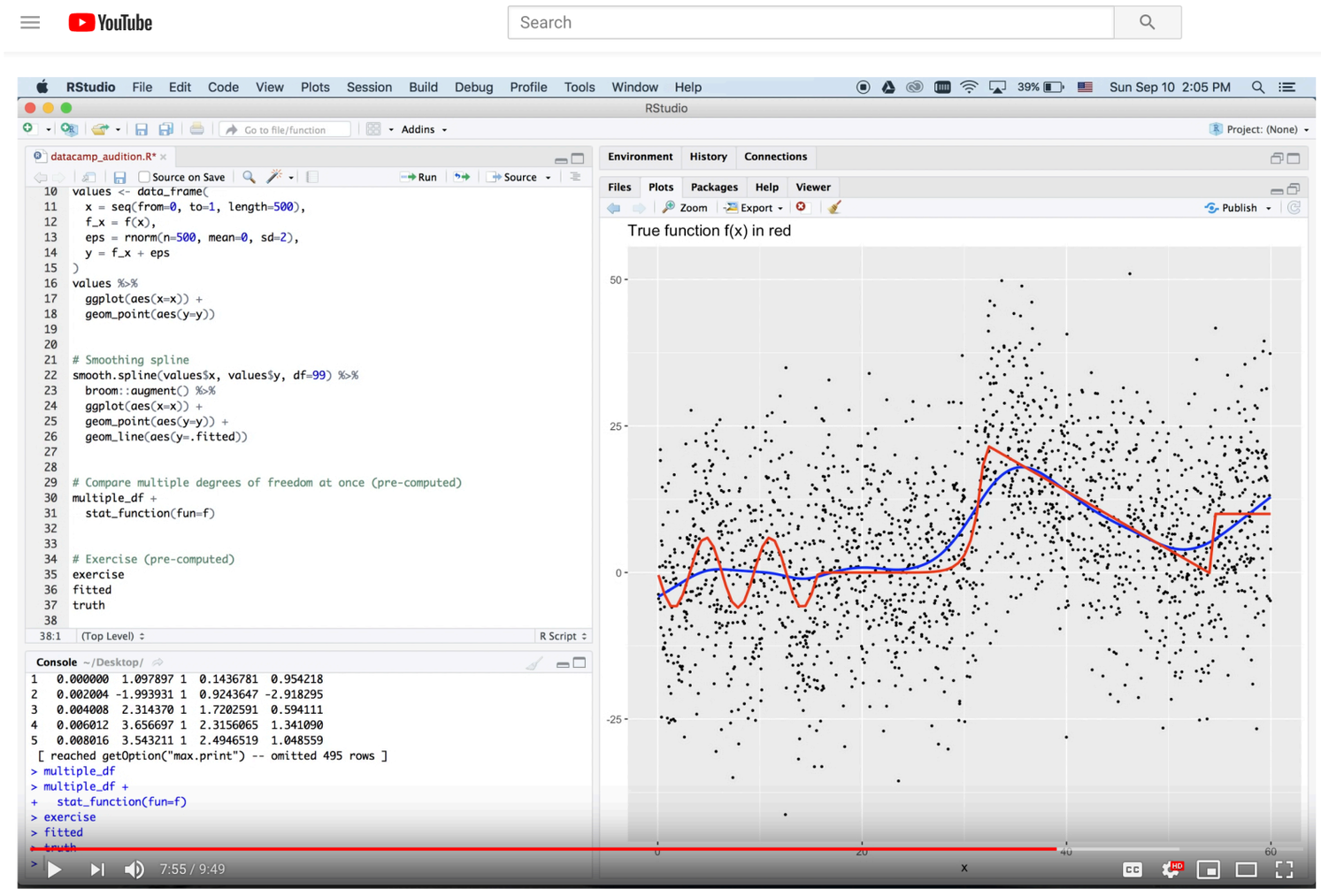
Machine Learning + Forest Ecology:
- Goal restated
- Assessment metric used
- Model assessment with crossvalidation

↓

Conclusion:
- Resources to learn more
- Thoughts on modeling

Resource 1: Intro to Splines Video

IMO splines are among the gentlest intro models to learning ML with!



Corresponding R code at bit.ly/rudeboybert_splines

Resource 2: DataCamp Pathway

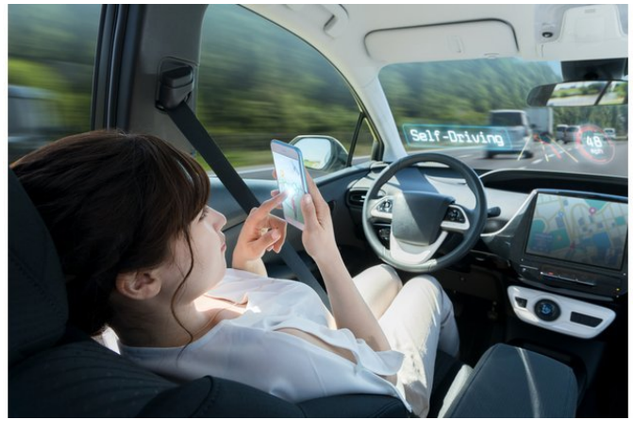
1. Build your tidyverse data science toolbox with Introduction to the Tidyverse.
   In particular data viz and data wrangling.

2. Just enough modeling theory & exercises with Modeling with Data in the Tidyverse.
   In particular Ch4 “Validation Set Prediction framework”, the bridge between modeling and…

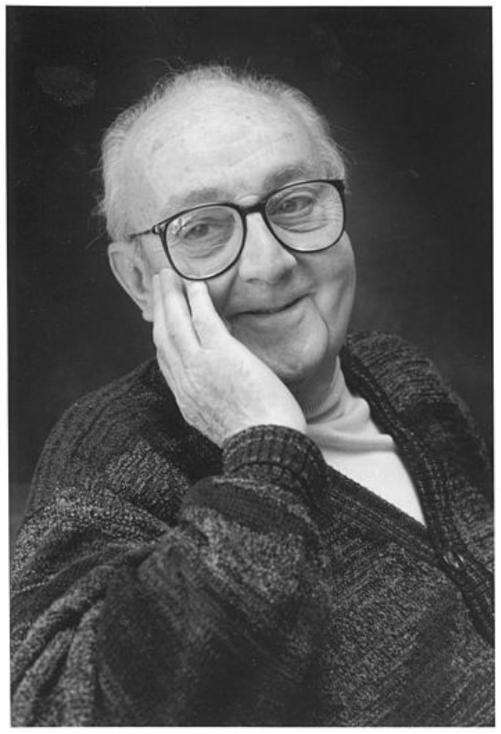
3. Machine learning methods with Machine Learning in the Tidyverse

Closing thoughts

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



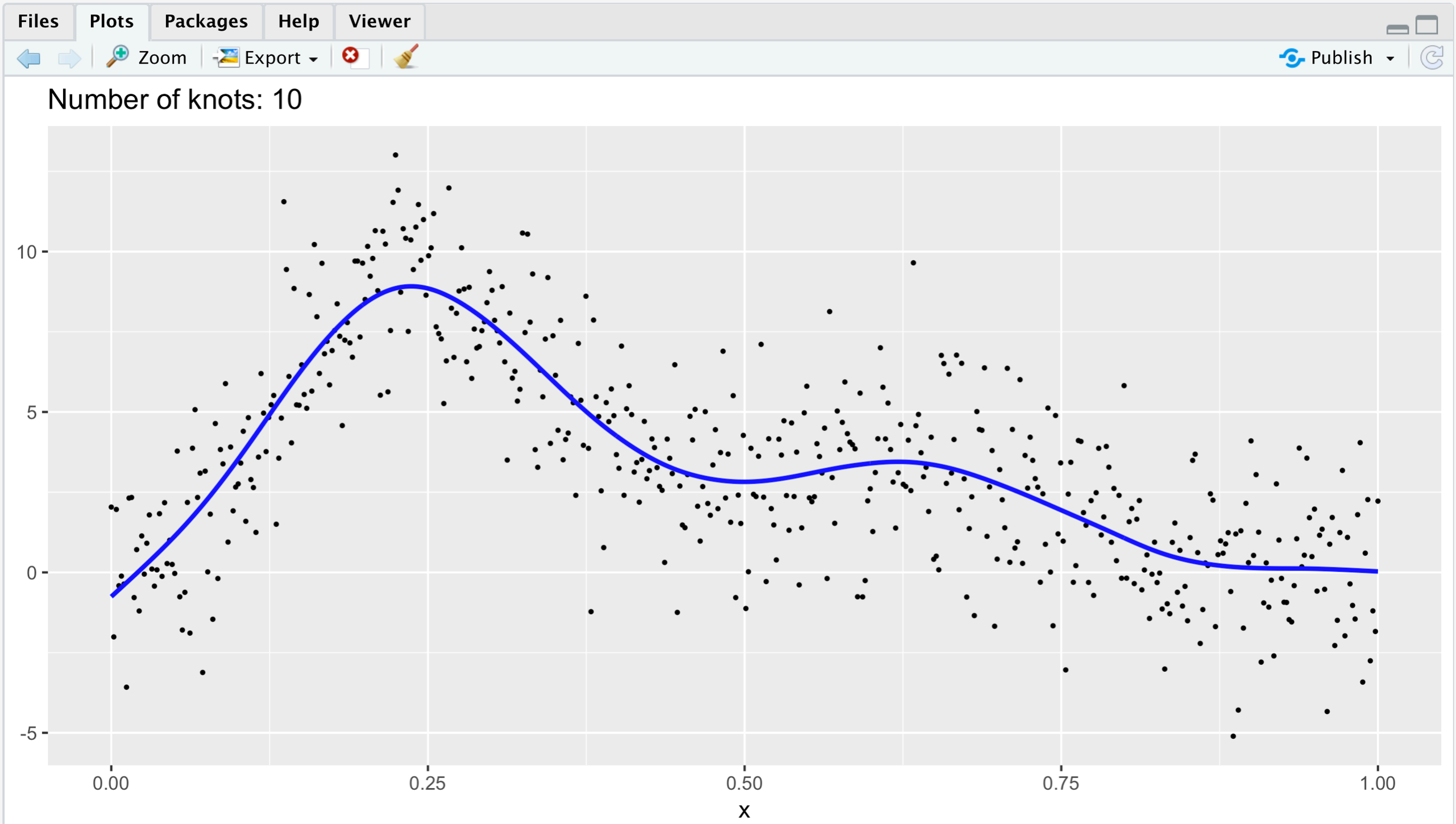
"All models are wrong,
but some are useful."
George Box

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

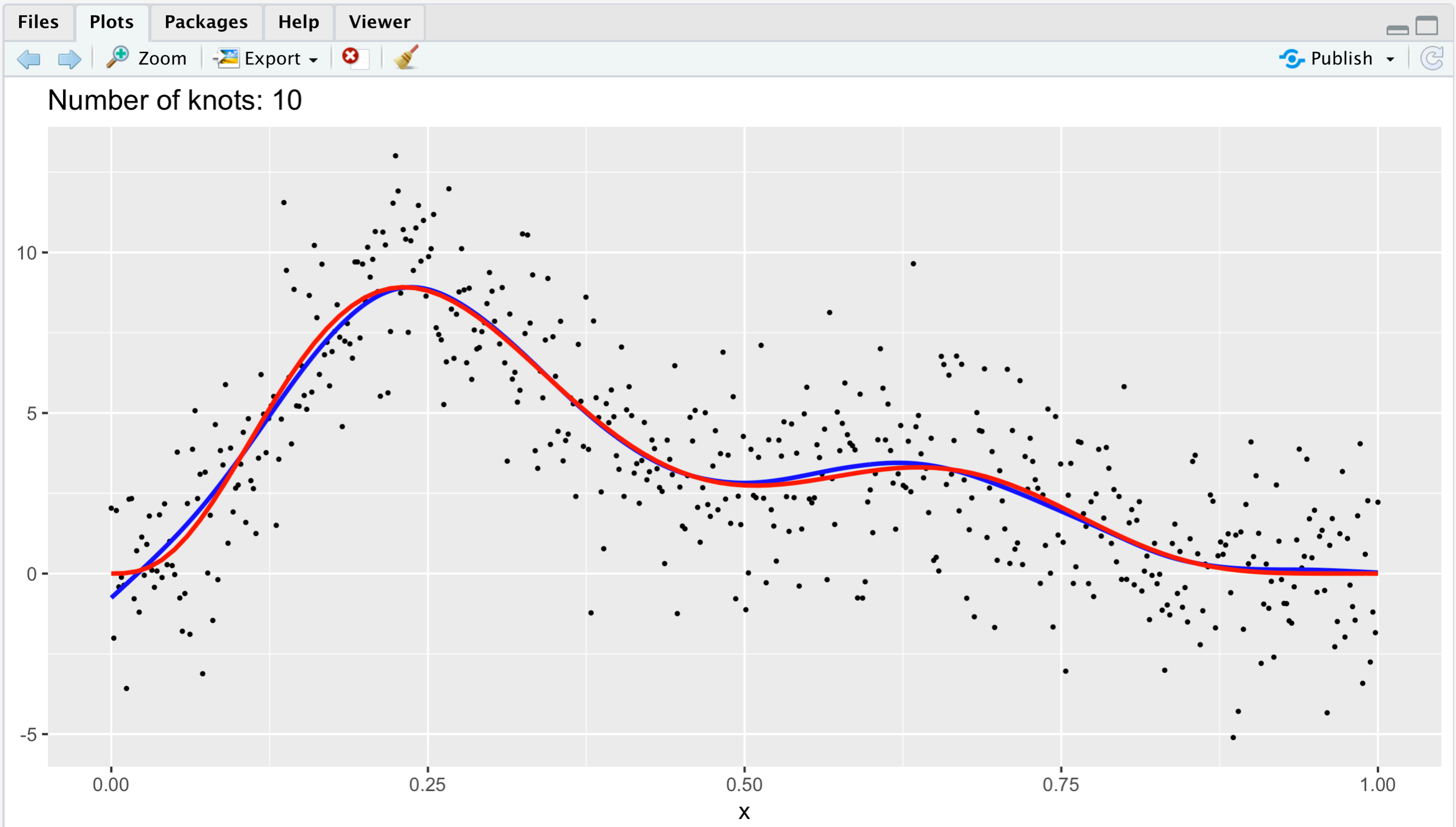
Thanks!

Before I go: A “Wizard of Oz” Reveal…

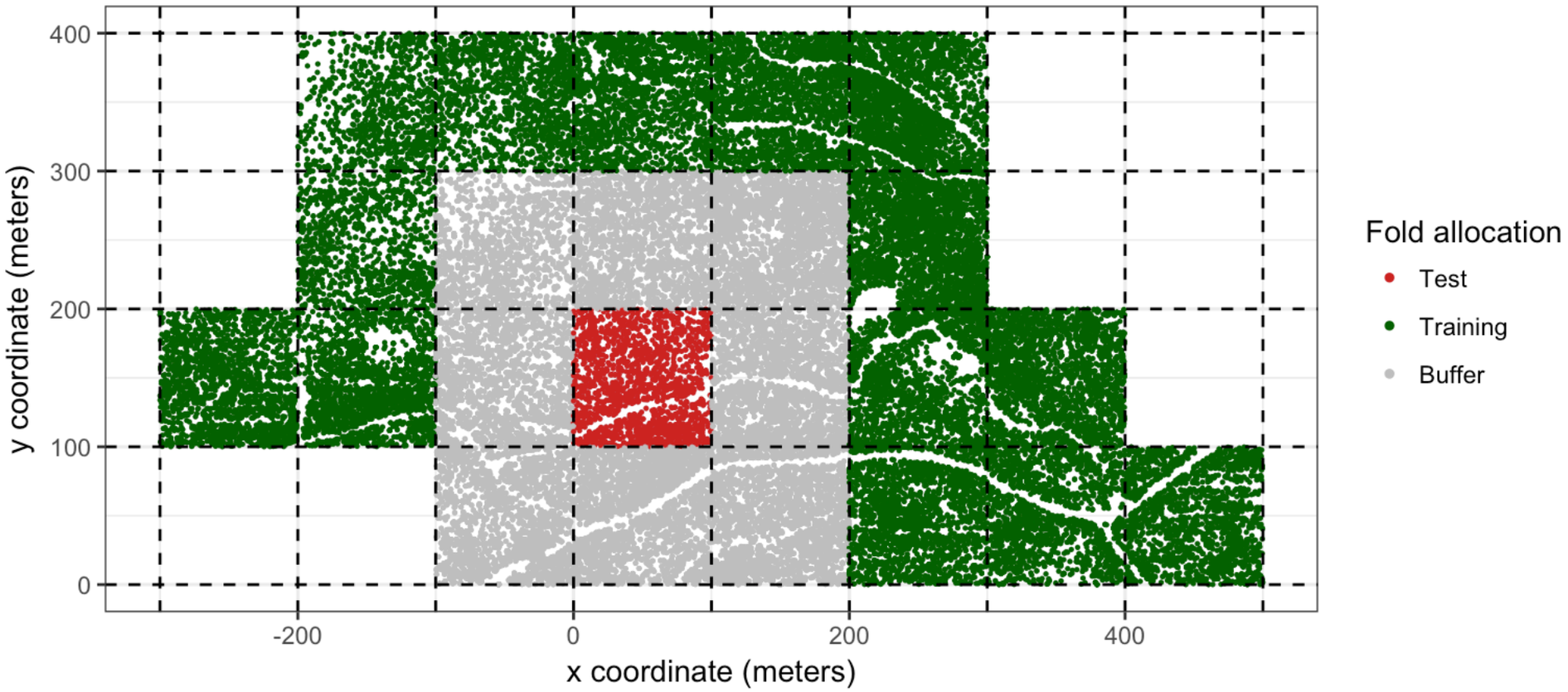
Our approximated $\hat{f}(x)$ was pretty close…



… to the *true* model $f(x) = 0.2x^{11}(10(1-x))^6 + 10(10x)^3(1-x)^{10}$

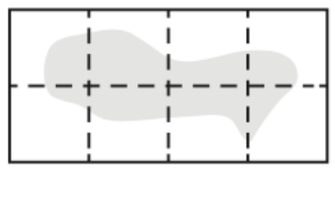
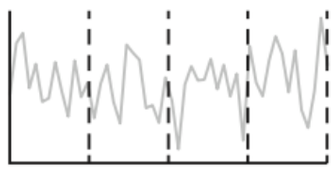
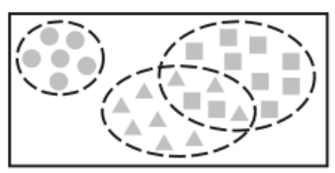


Our Data is Spatial: *Spatial* Crossvalidation



Resource 3: Paper

"Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure" Roberts (2017)

Dependence structure	Parametric solution	Blocking	Blocking illustration
Spatial	Spatial models (e.g.CAR, INLA, GWR)	Spatial	
Temporal	Time-series models (e.g.ARIMA)	Temporal	
Grouping	Mixed effect models (e.g. GLMM)	Group	
Hierarchical / Phylogenetic	Phylogenetic models (e.g. PGLS)	Hierarchical