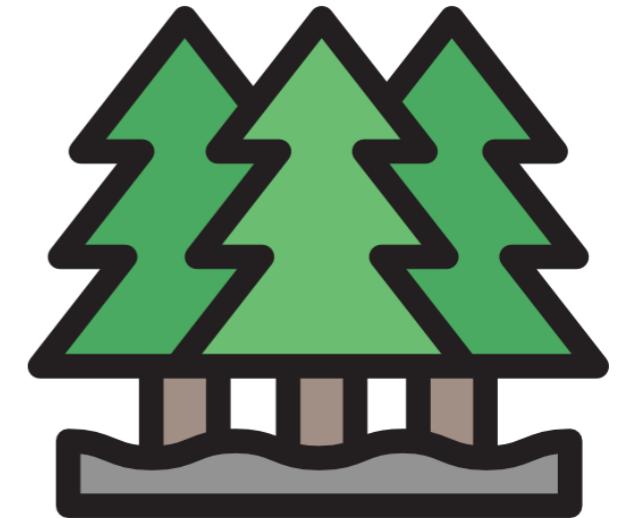


# Self-Driving Cars & Forest Ecology: Modeling for Machine Learning



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

Slides available on Twitter [@rudeboybert](https://twitter.com/rudeboybert)

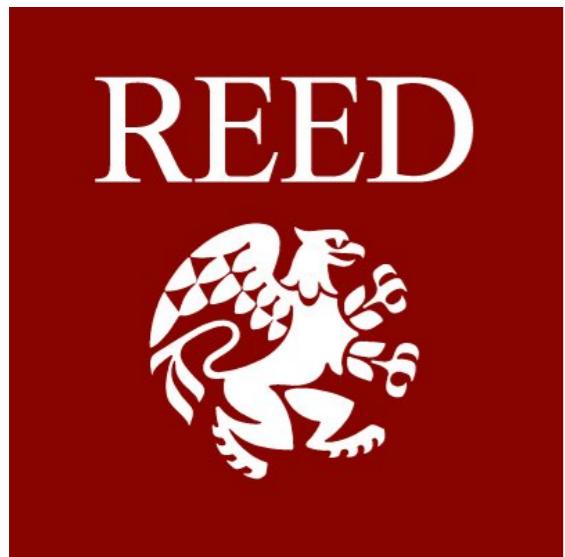
# What variables are being collected?



# Background



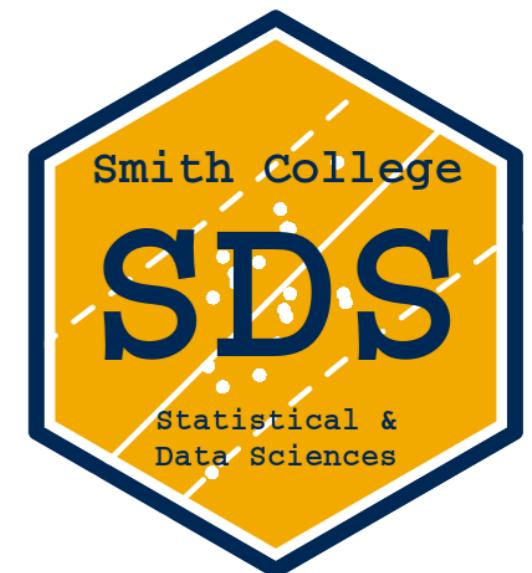
Google



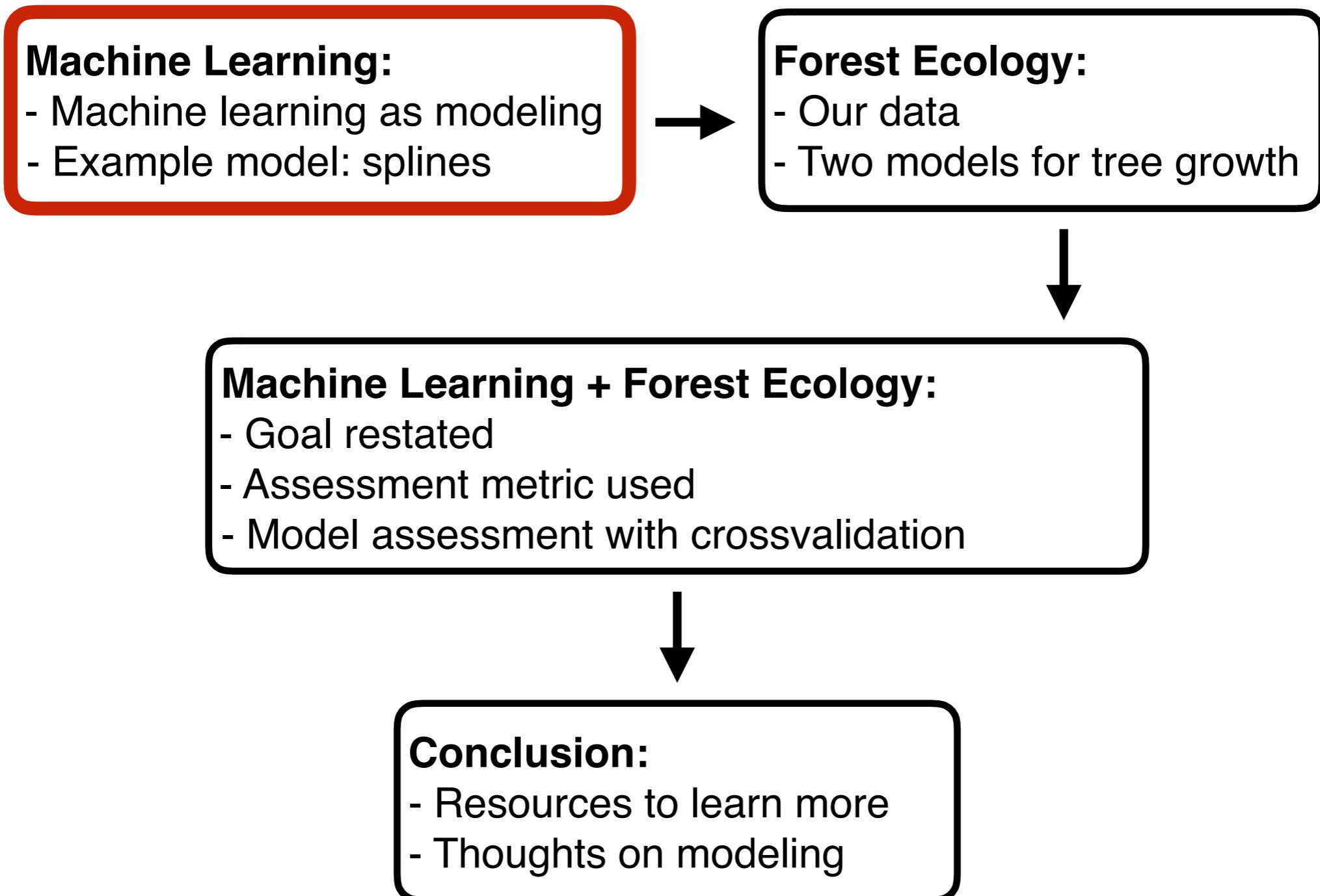
Middlebury



AMHERST  
COLLEGE



# Road Map



# Machine Learning



WAYMO

NFT

AI

Prediction!

ATCH FIX



# Machine Learning as Modeling

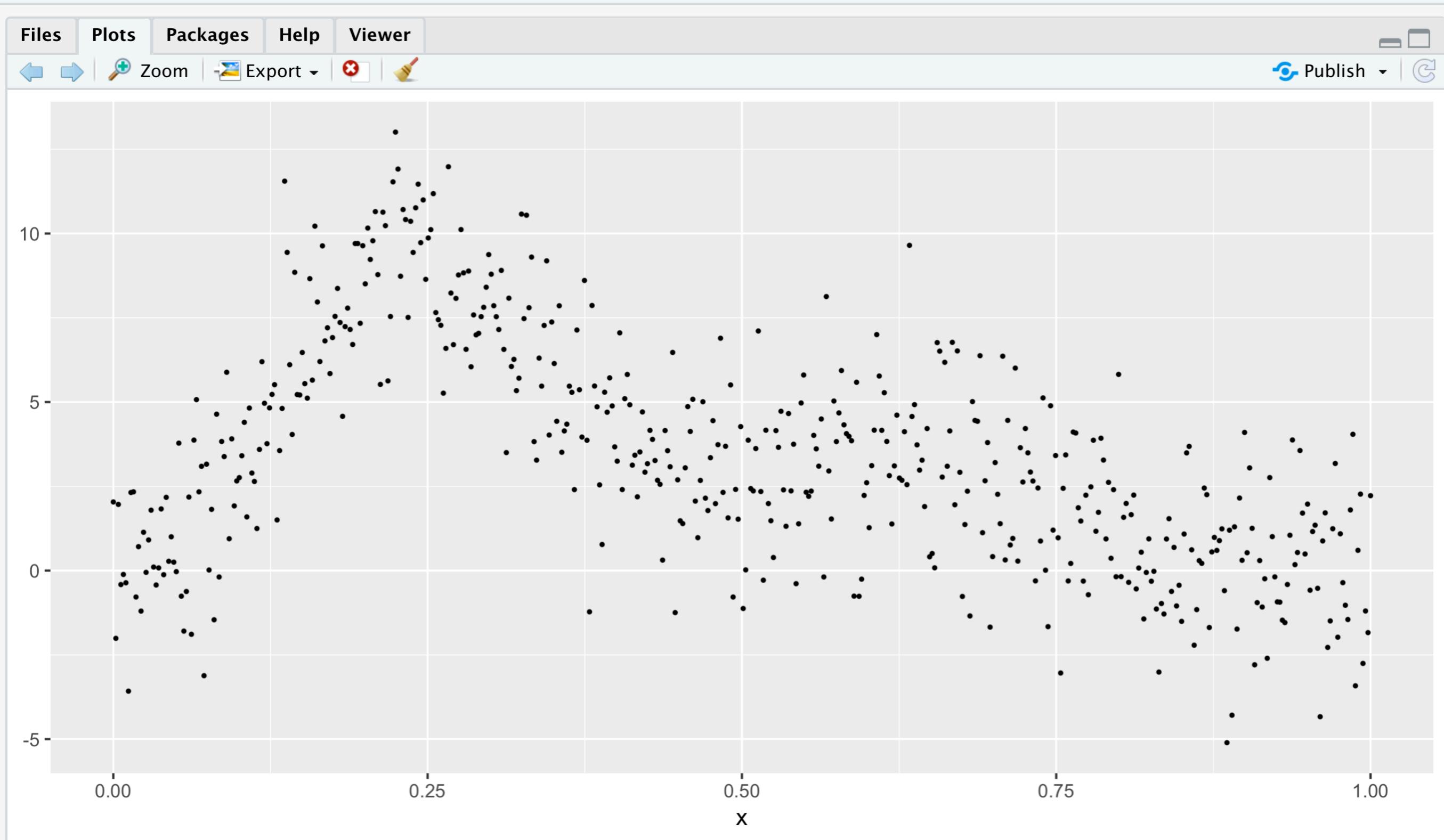
True (Unknown) Model:  $y = f(\vec{x}) + \epsilon$

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

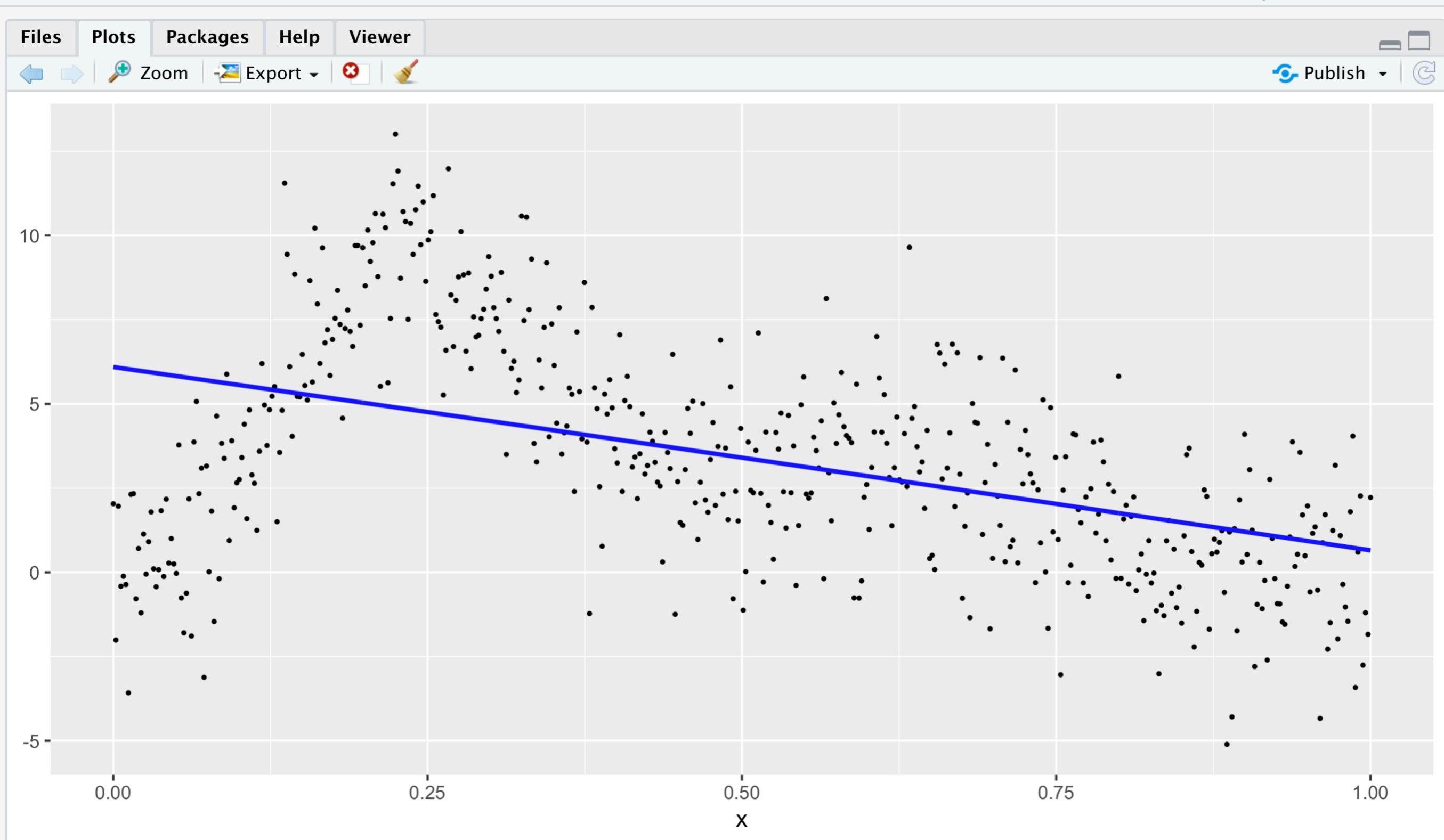
Now to the blackboard for  
Chalk Talk #1...



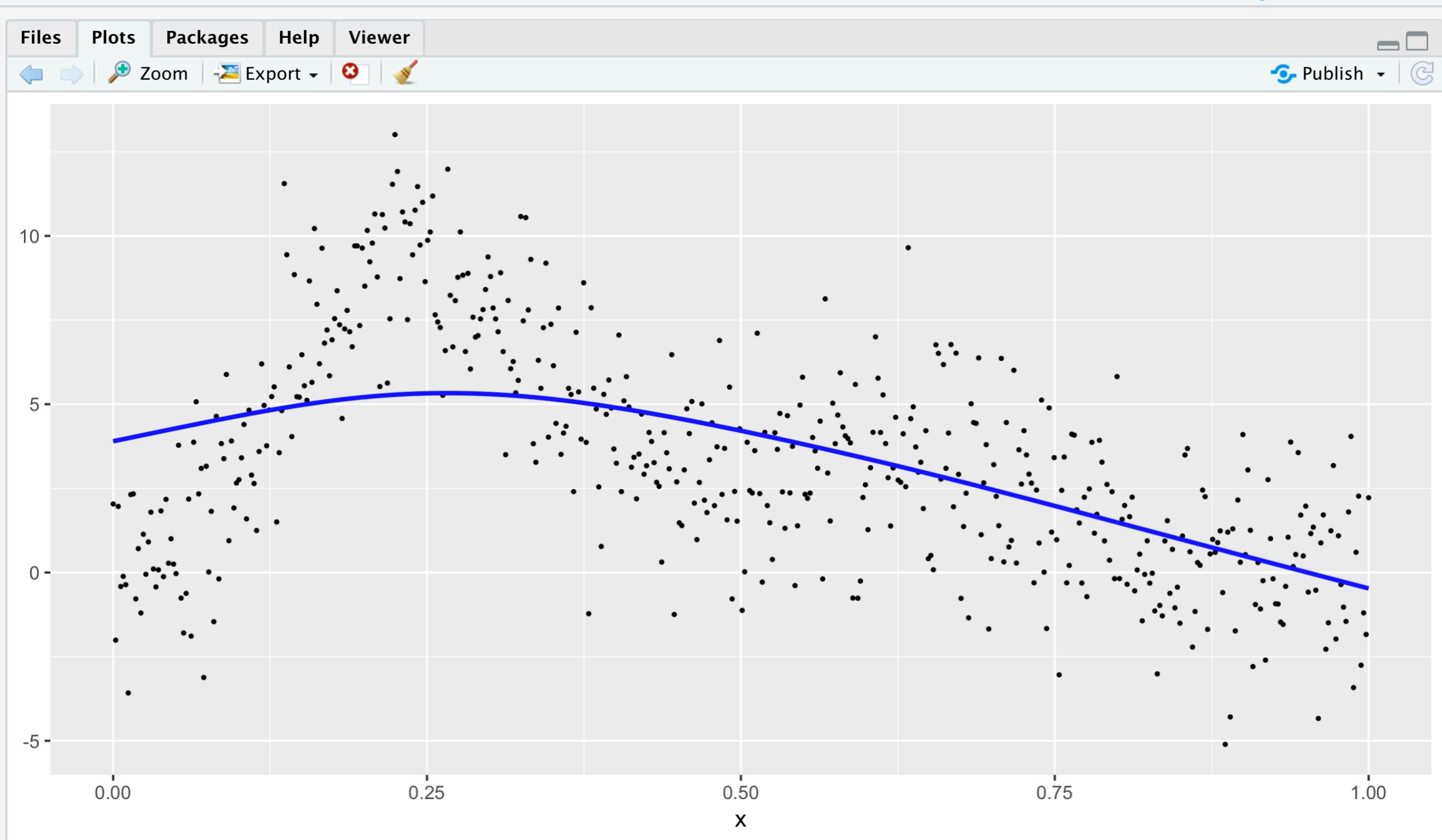
# Given Data $(x, y)$ from unknown $f(x)$



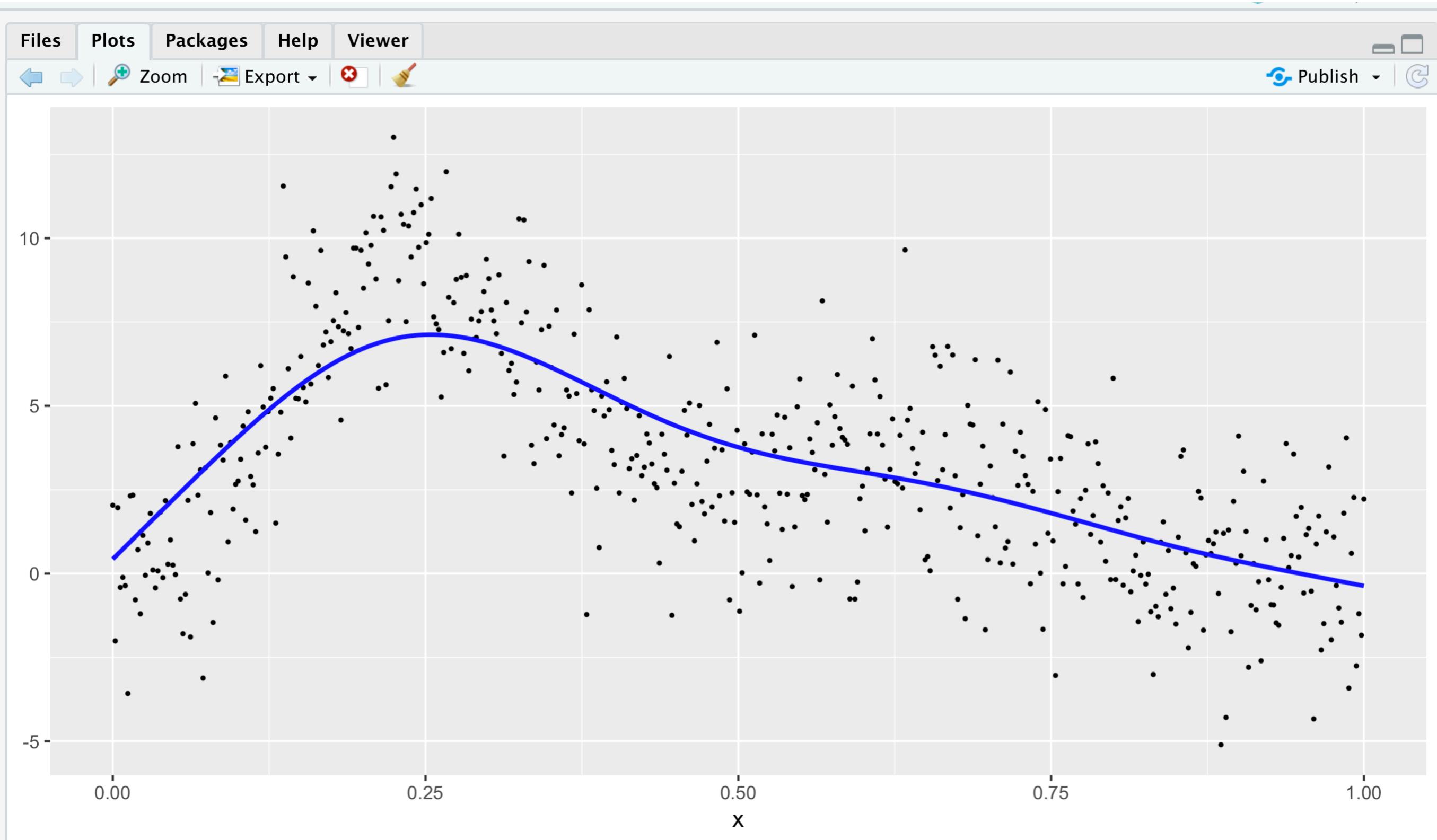
# Approximate (i.e. “fit”) a Model $\hat{f}(x)$



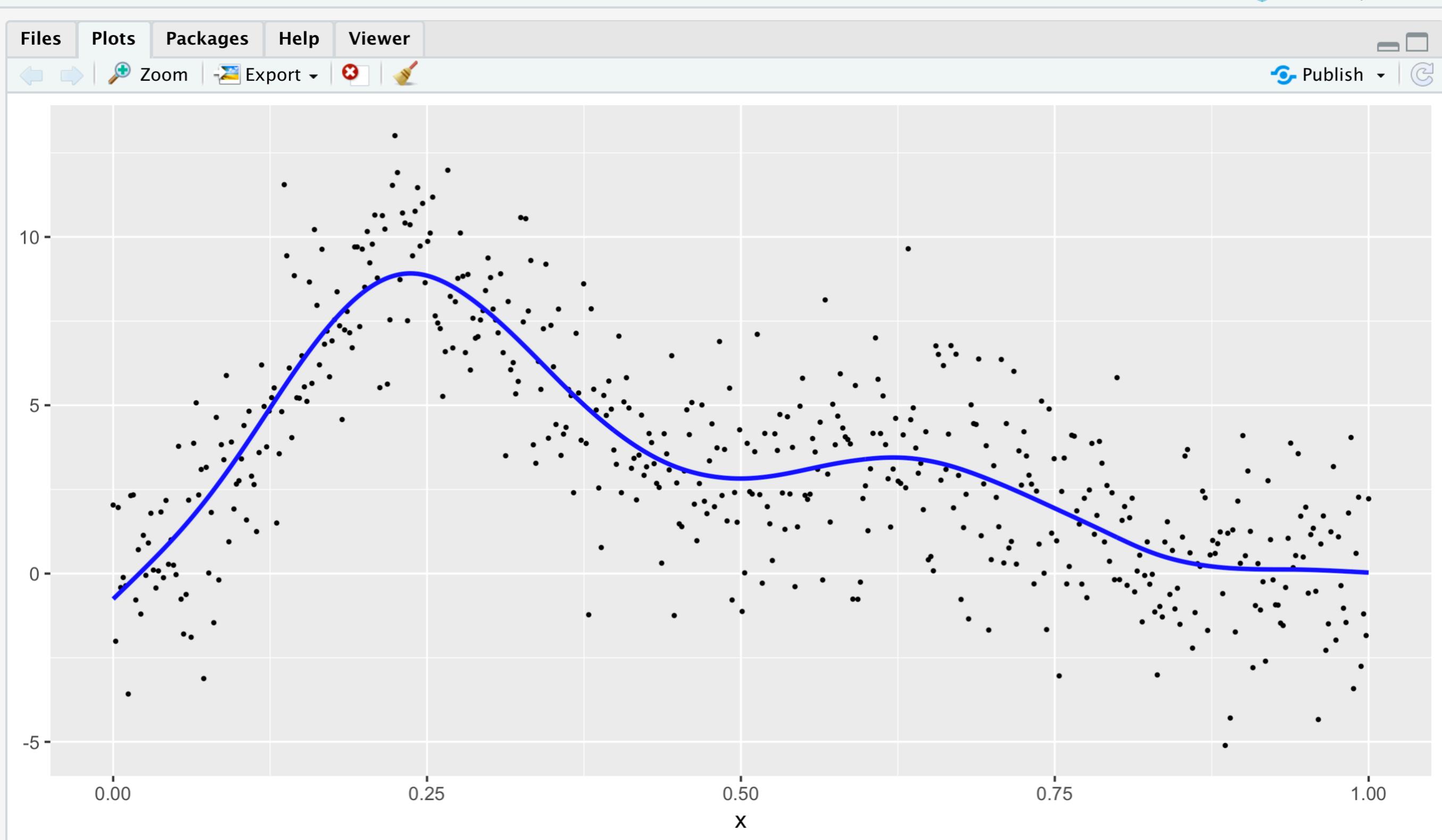
# How about this $\hat{y} = \hat{f}(x)$ ?



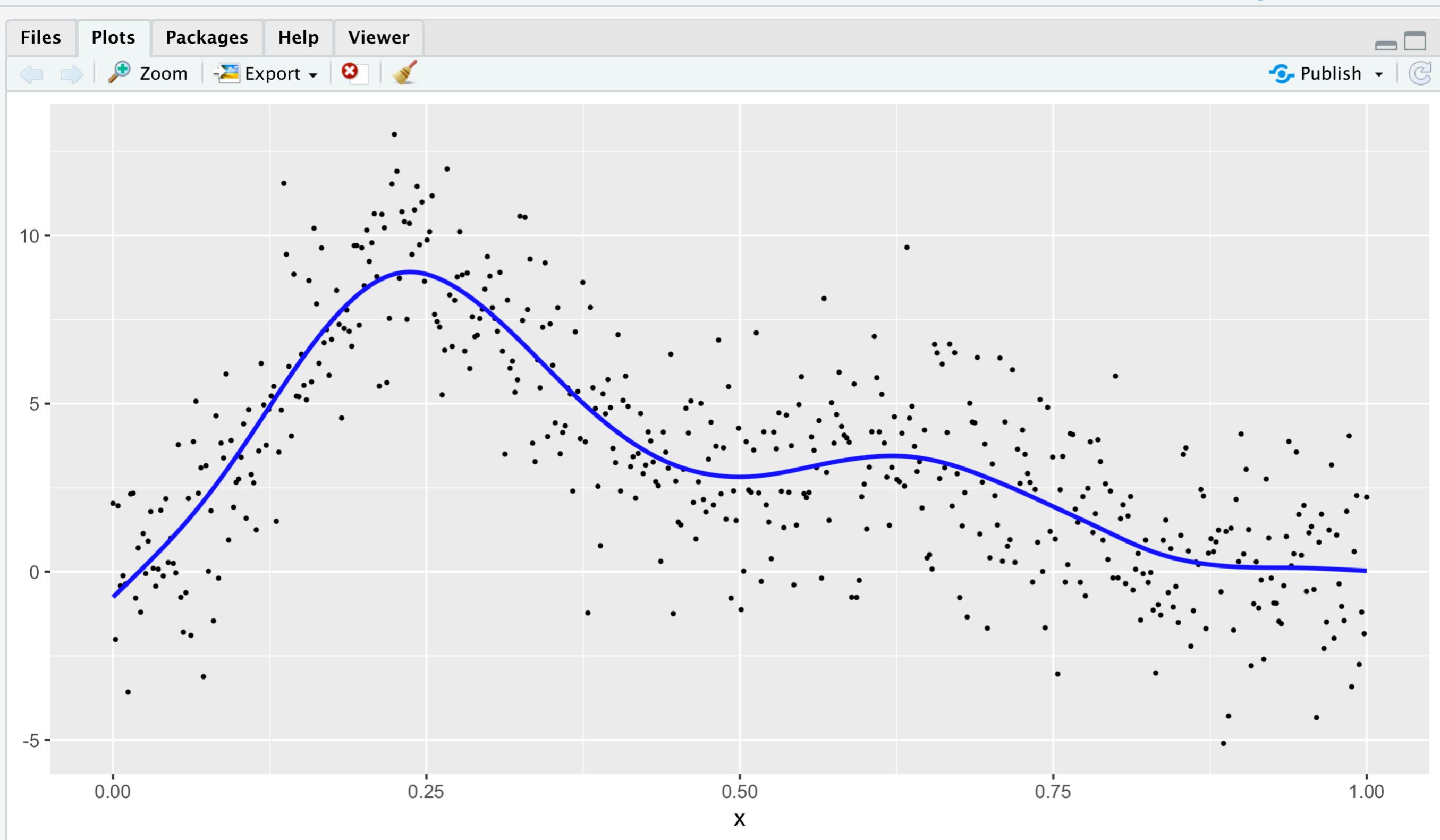
# How about this $\hat{f}(x)$ ?



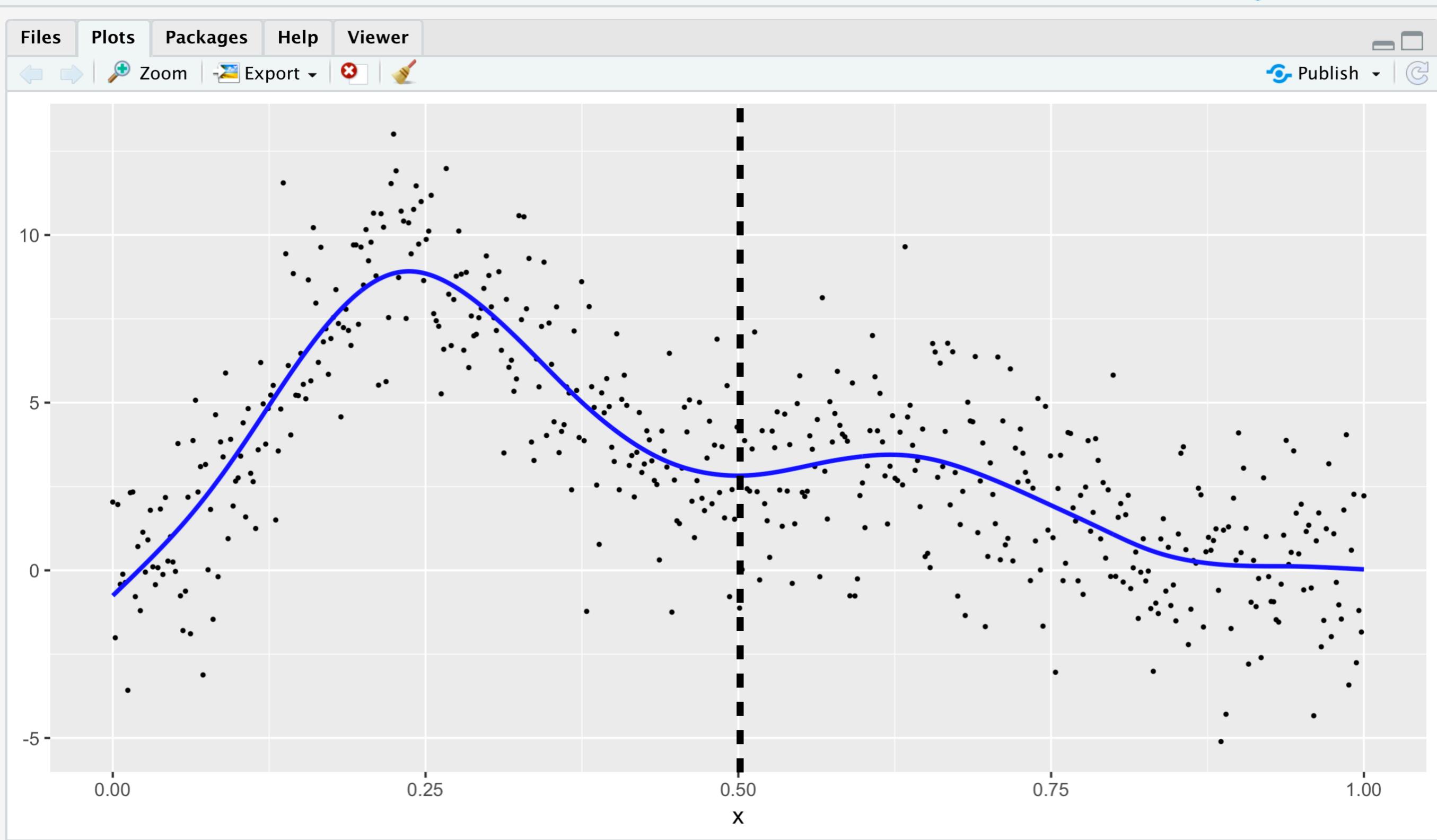
# How about this $\hat{f}(x)$ ?



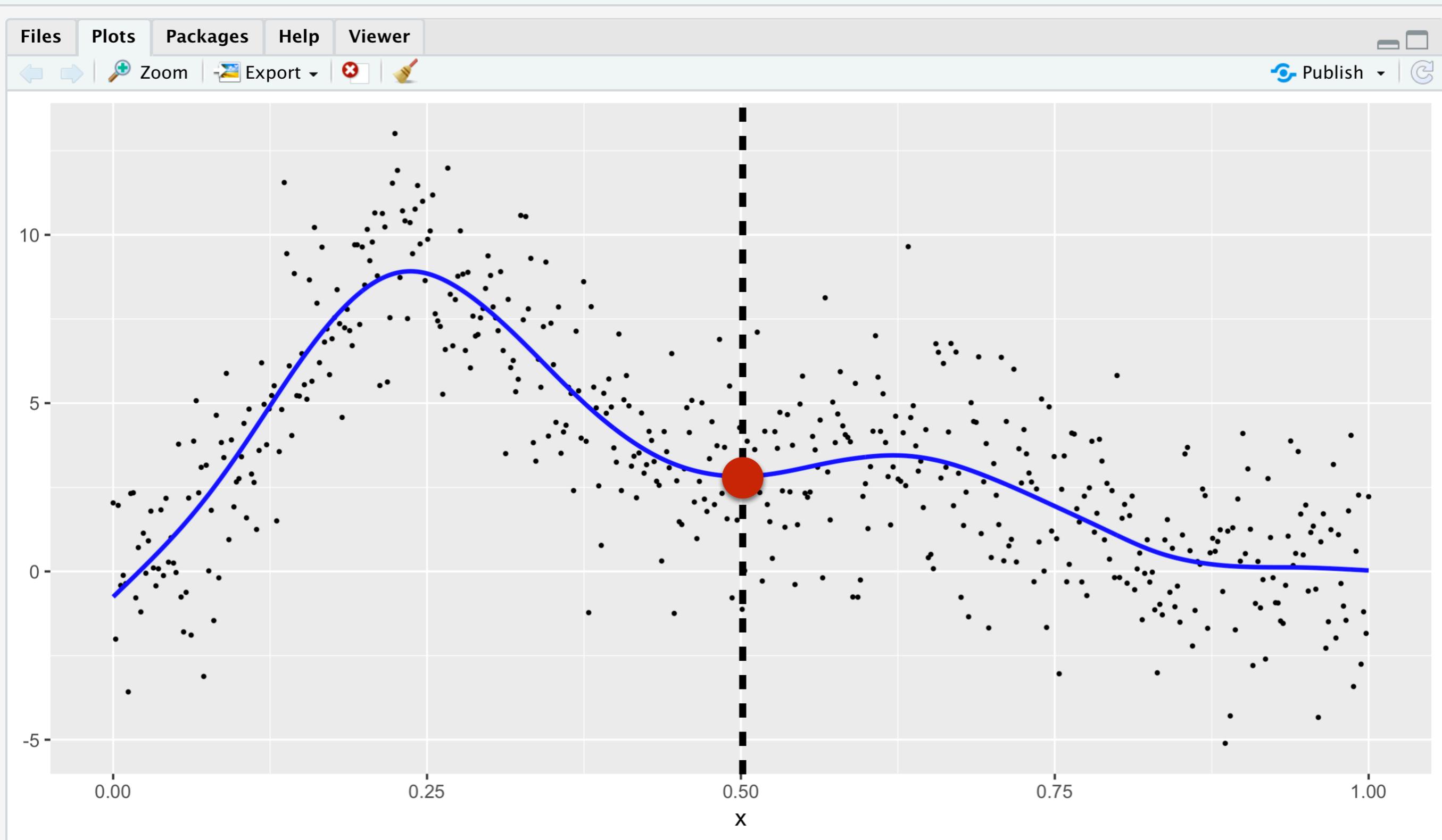
# What does this $\hat{f}(x)$ predict for $x = 0.5$ ?



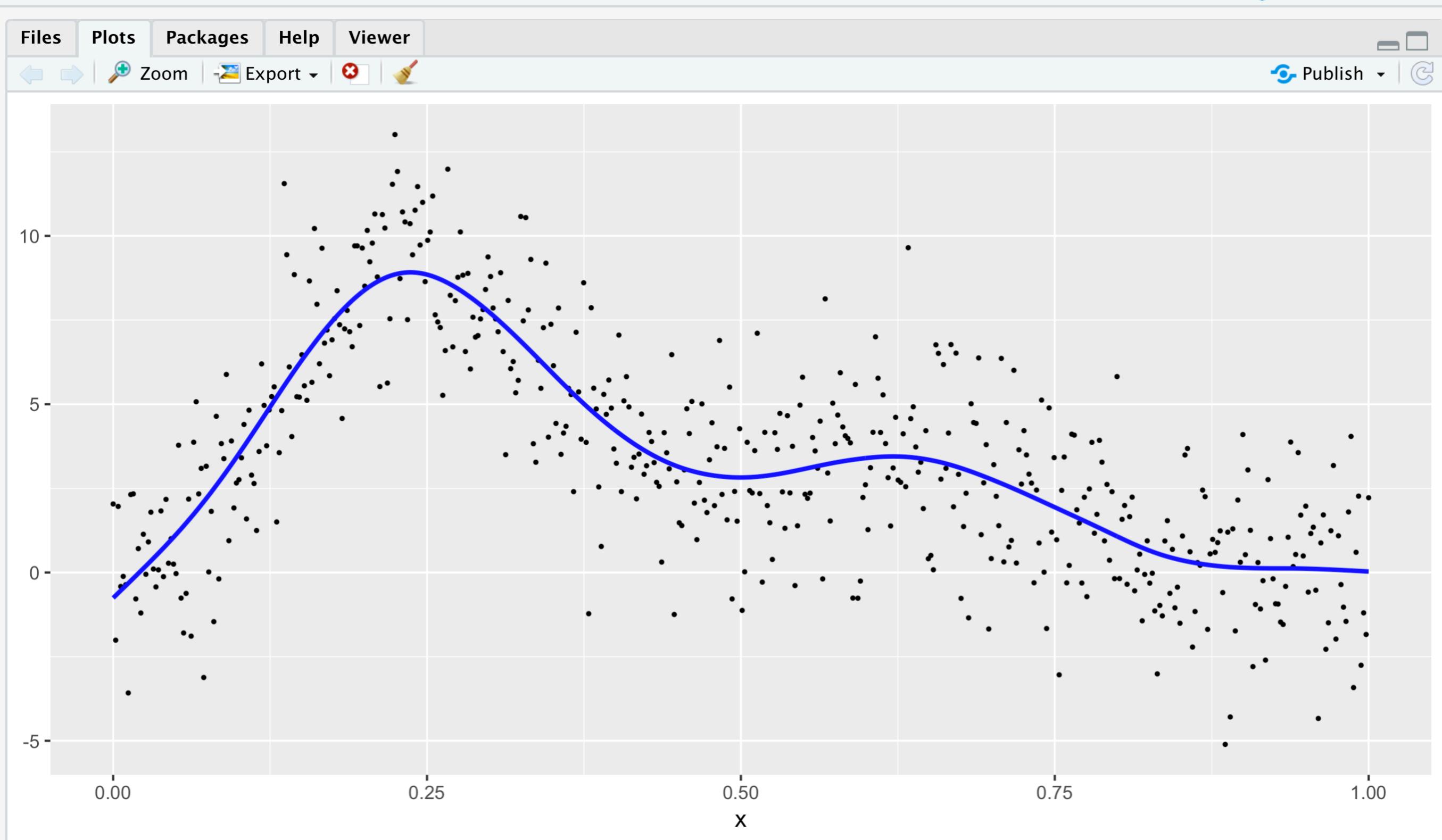
# What does this $\hat{f}(x)$ predict for $x = 0.5$ ?



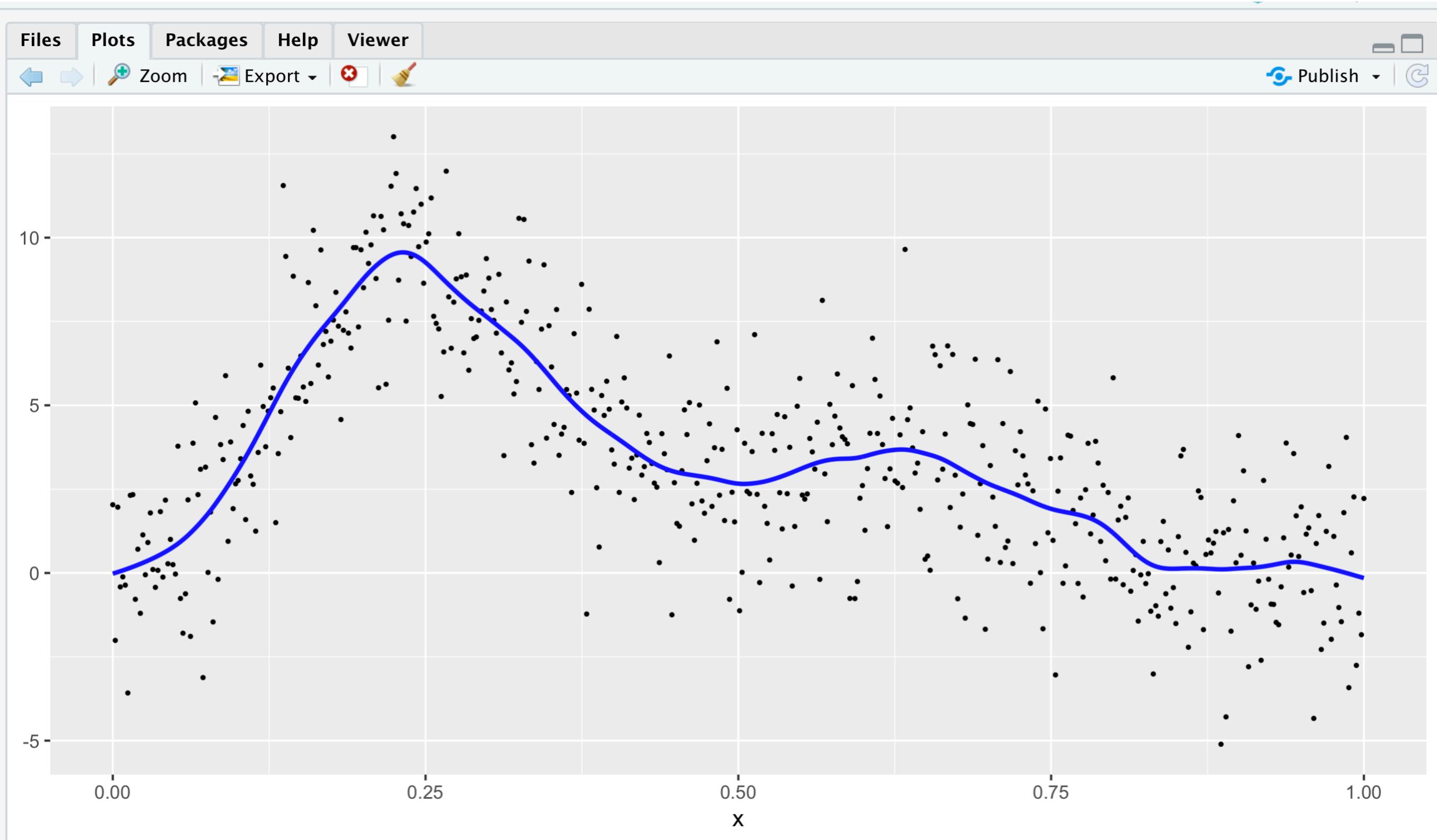
# What does this $\hat{f}(x)$ predict for $x = 0.5$ ?



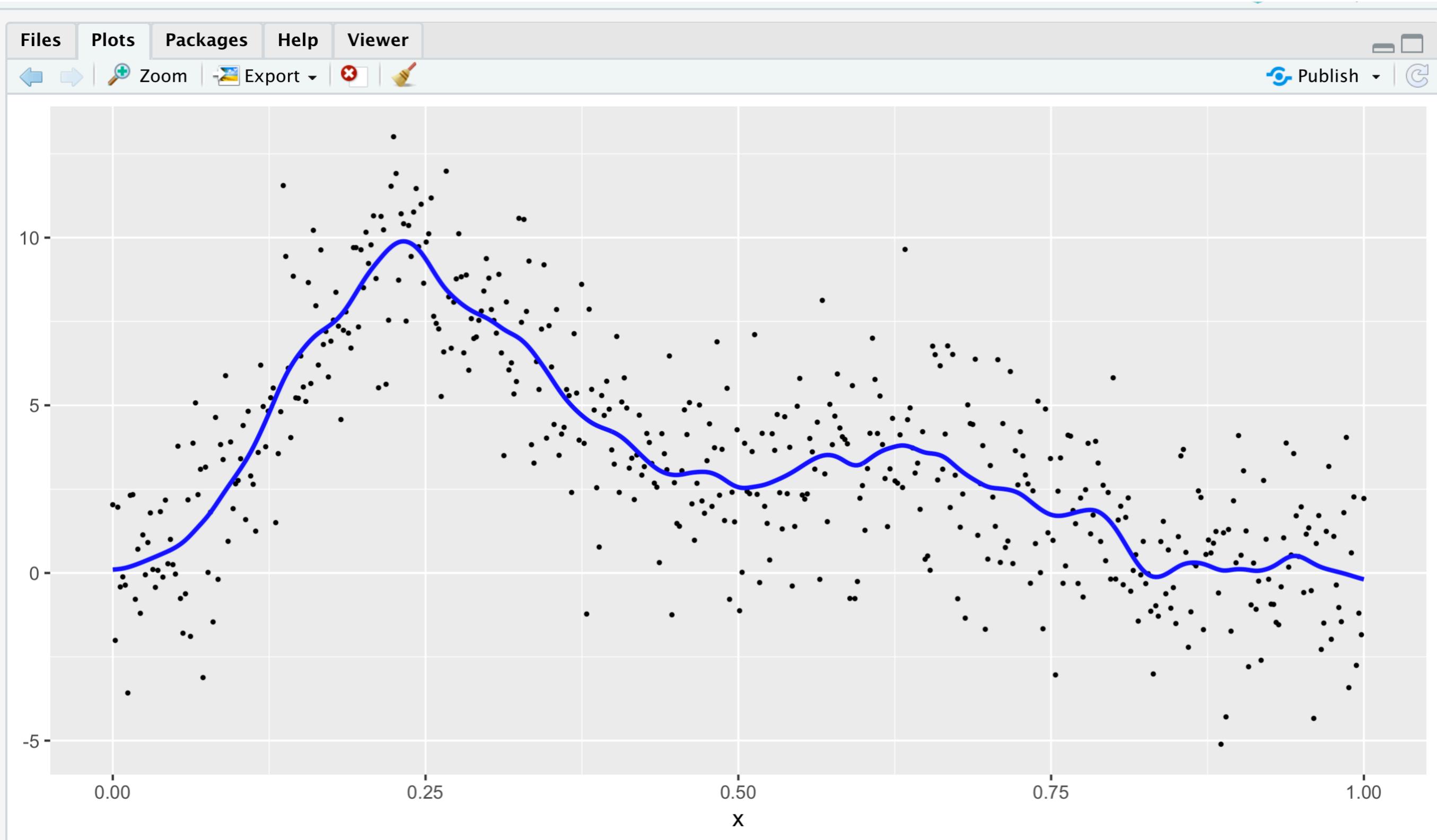
Ok, great. But instead of this  $\hat{f}(x)$  ...



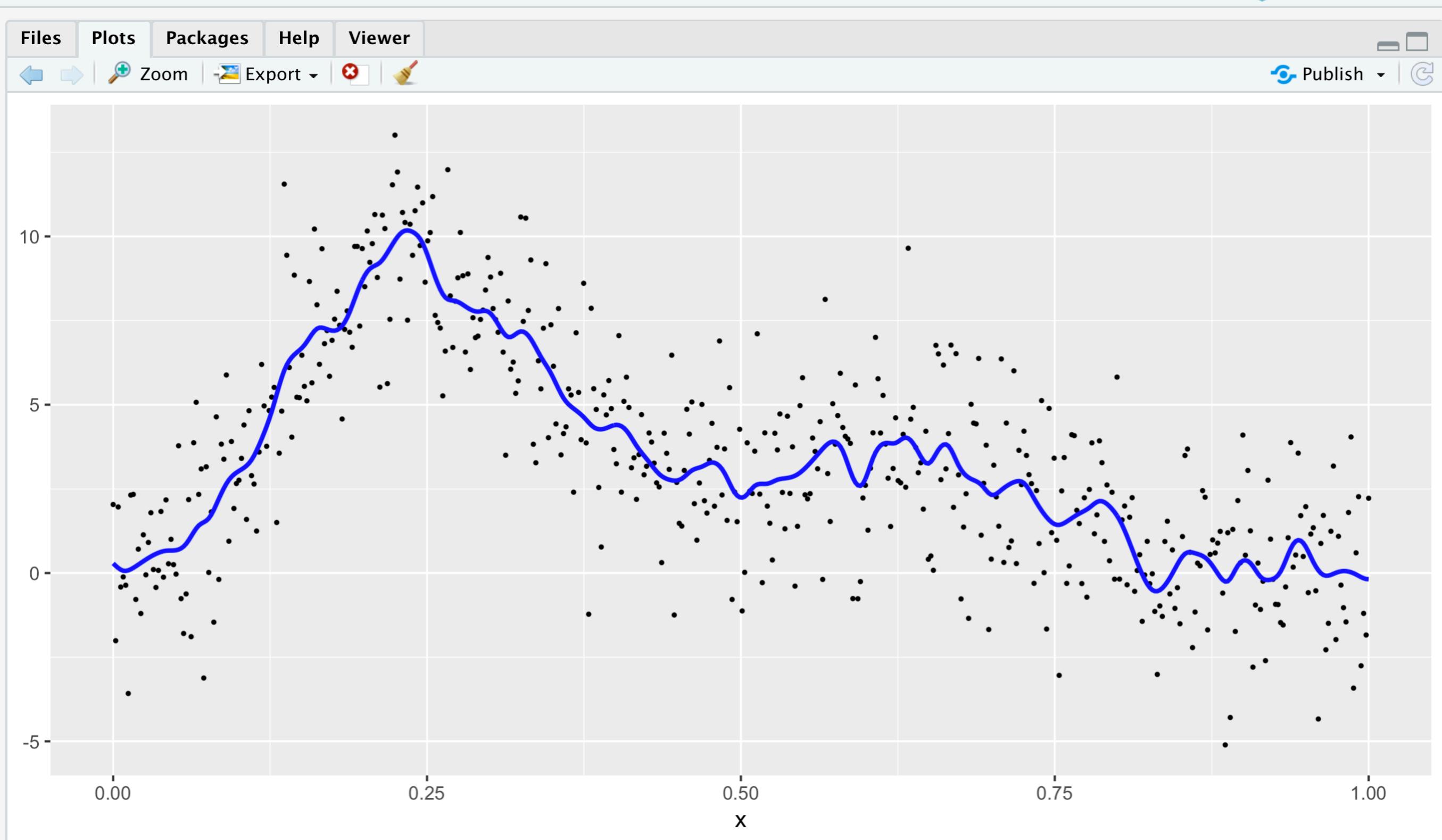
# How about this $\hat{f}(x)$ ?



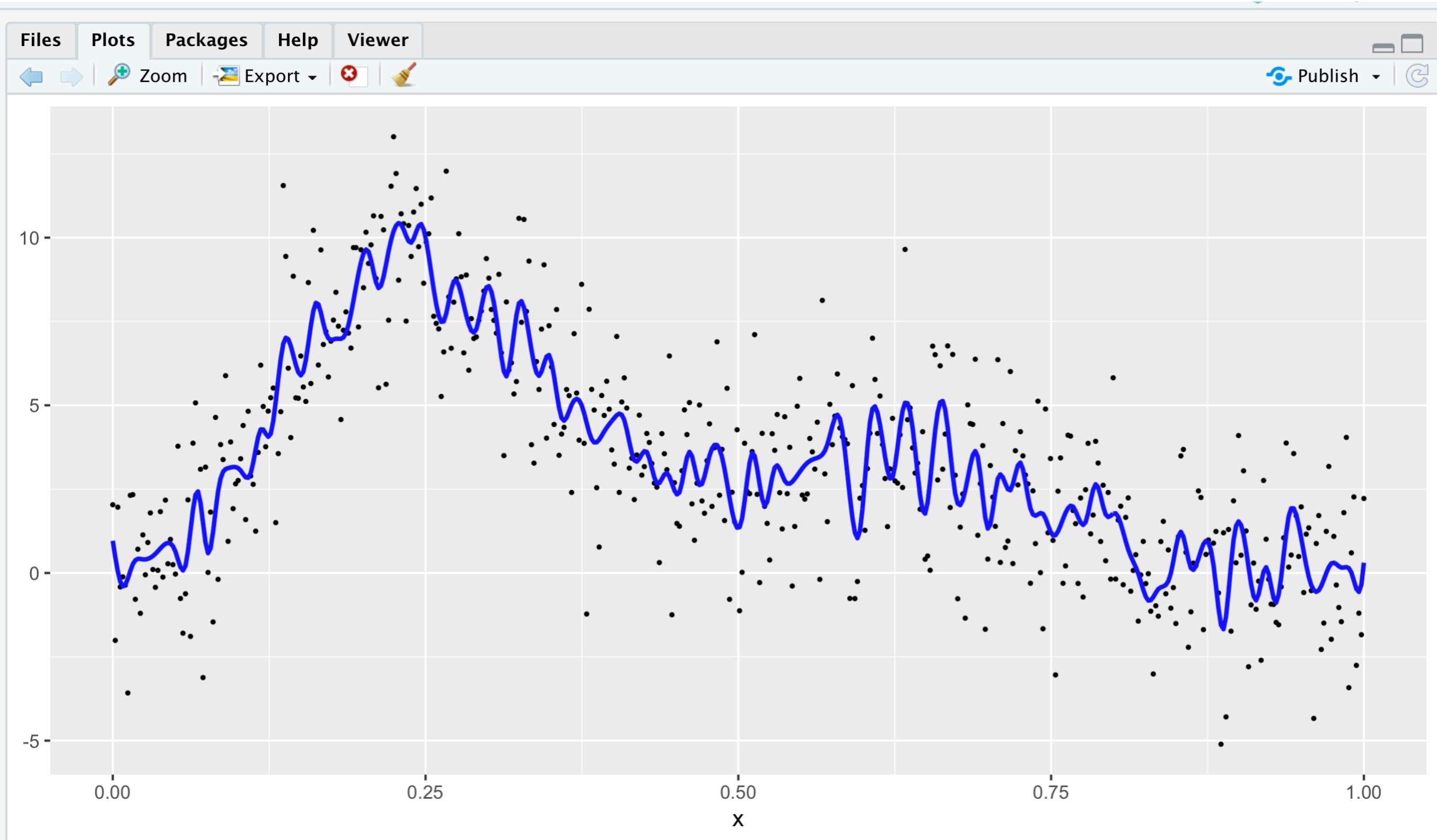
# How about this $\hat{f}(x)$ ?



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



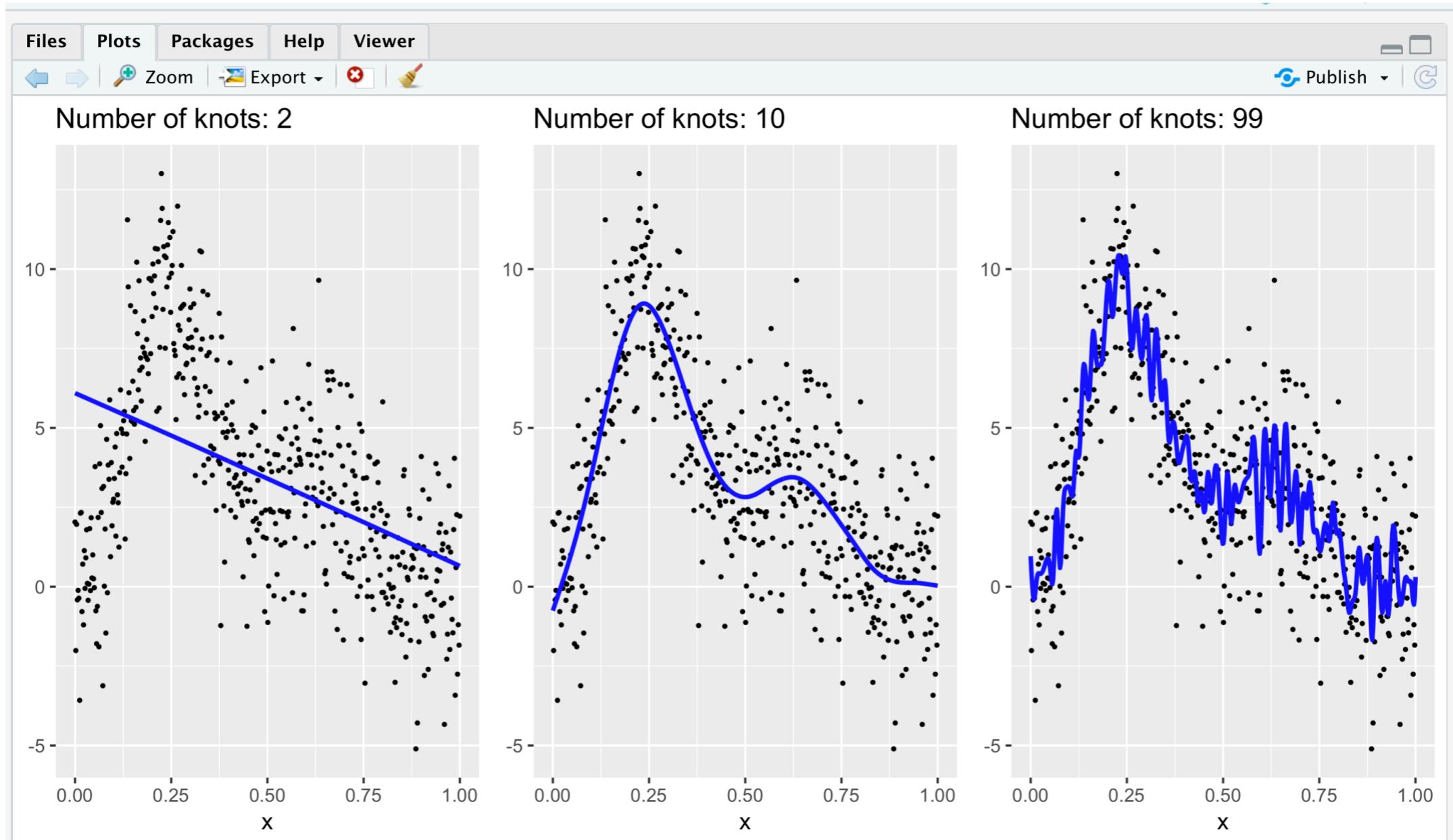
# How about this $\hat{f}(x)$ ?



# 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 $\hat{f}(x)$

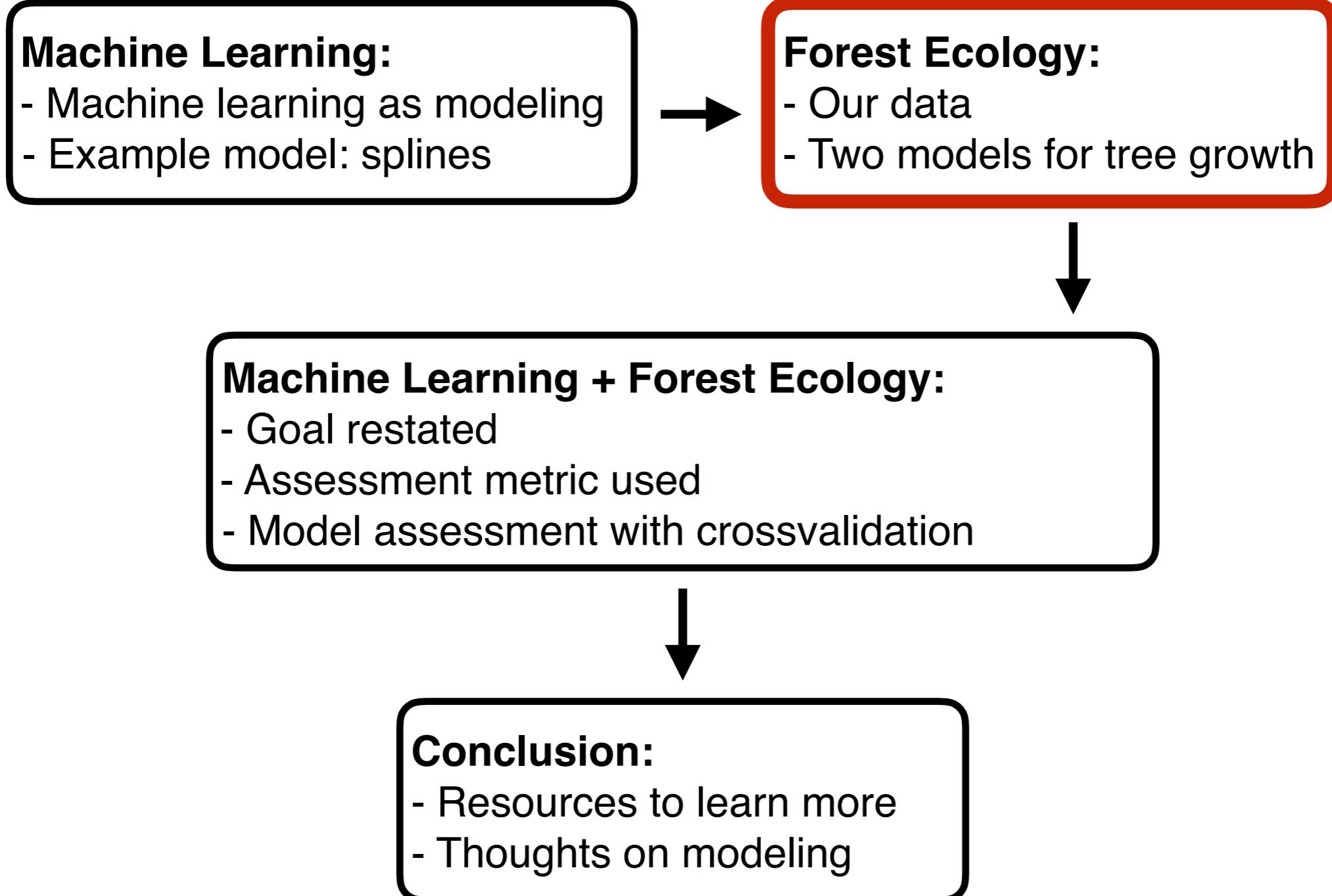


Underfit!

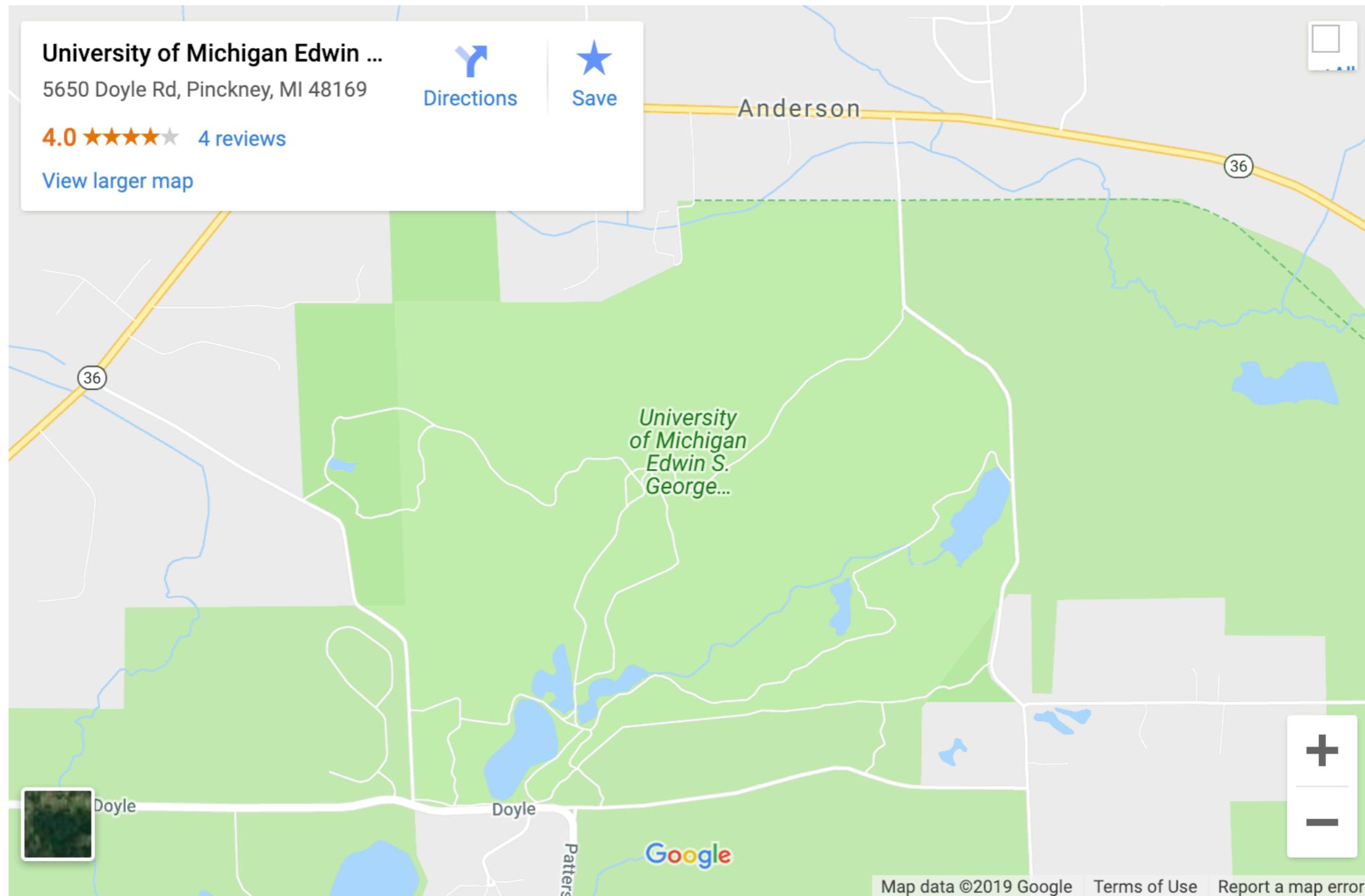
“Just right!”

Overfit!

# Road Map

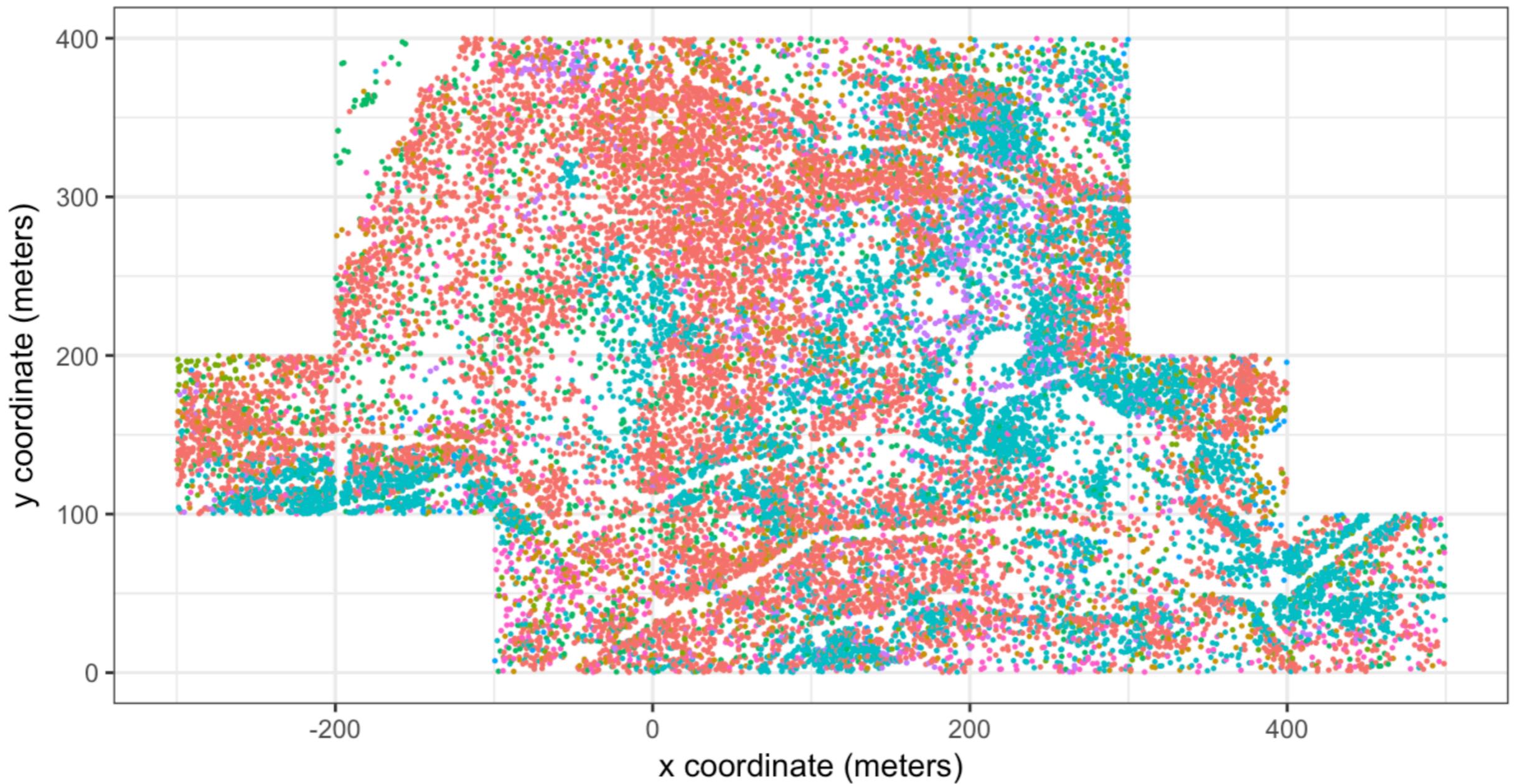


# Data: 2008 & 2014 Censuses of Trees



# Data: 2008 Snapshot

Spatial distribution of top 8 species

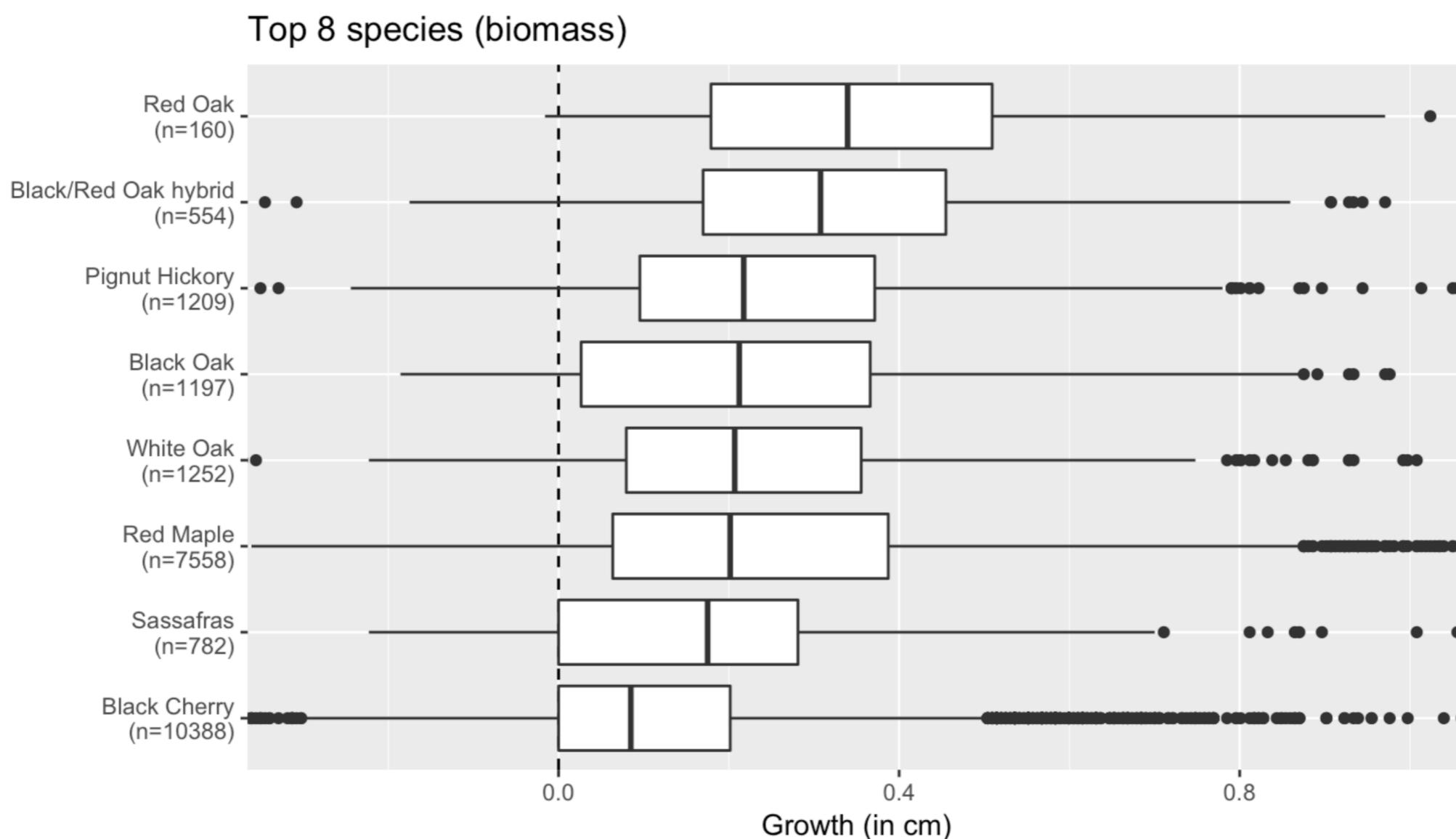


# Recall our Variables!



$y$ : Outcome Variable = Avg Annual Growth

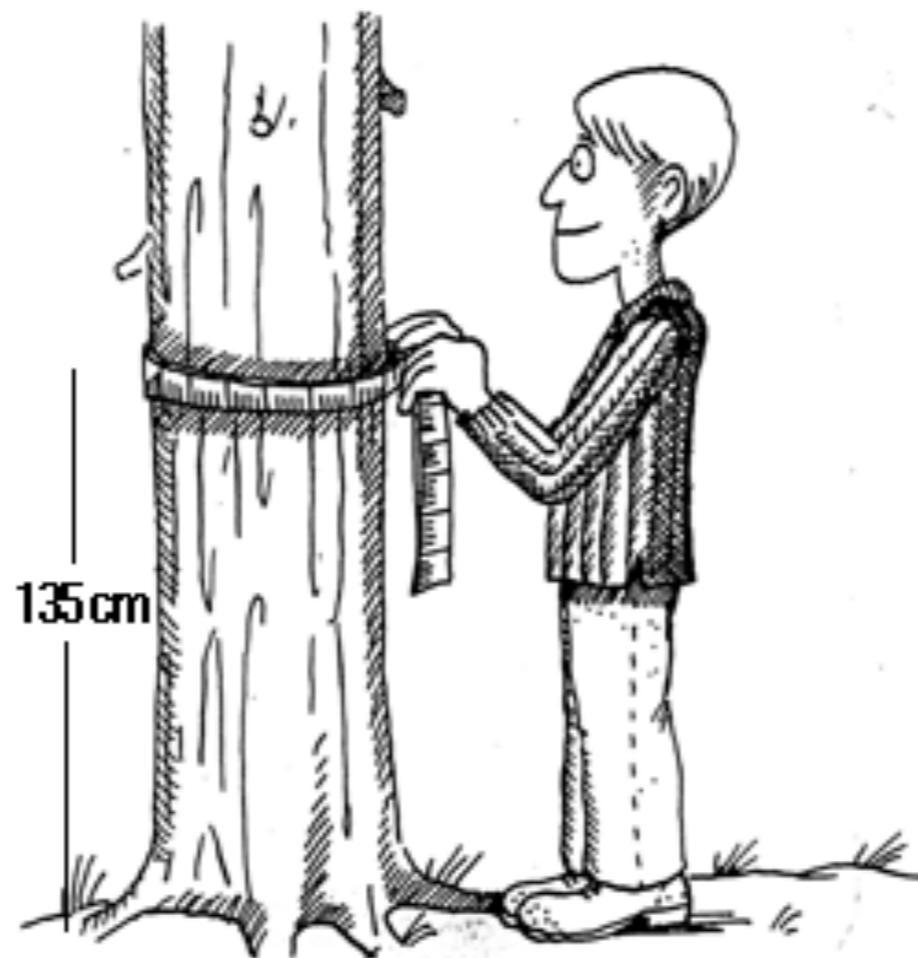
Observed average annual growth of trees 2008-2014



# Predictor Variables $\vec{x}$

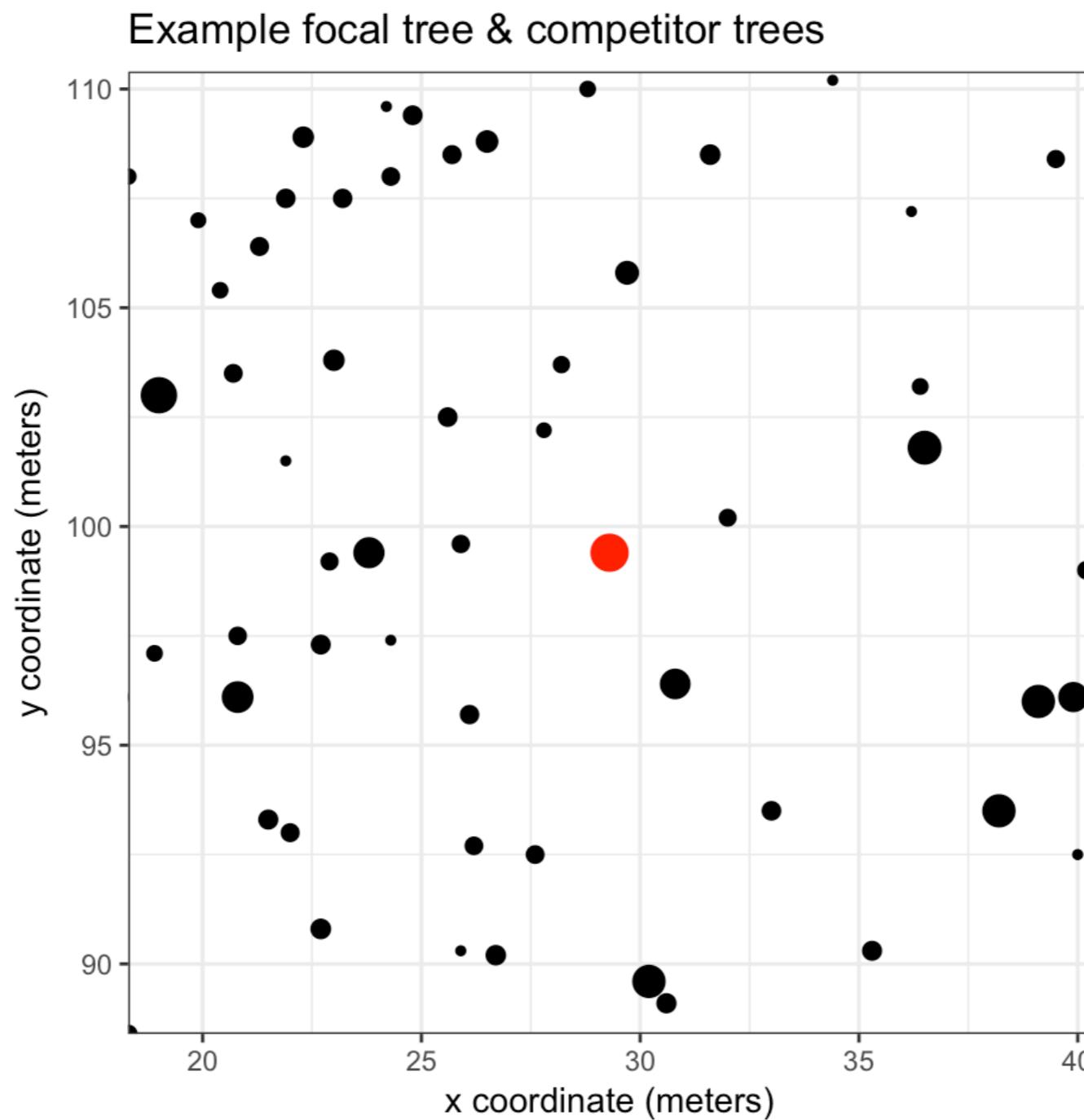
$x_1$  : Species of tree

$x_2$  : Size of tree (diameter at breast height)



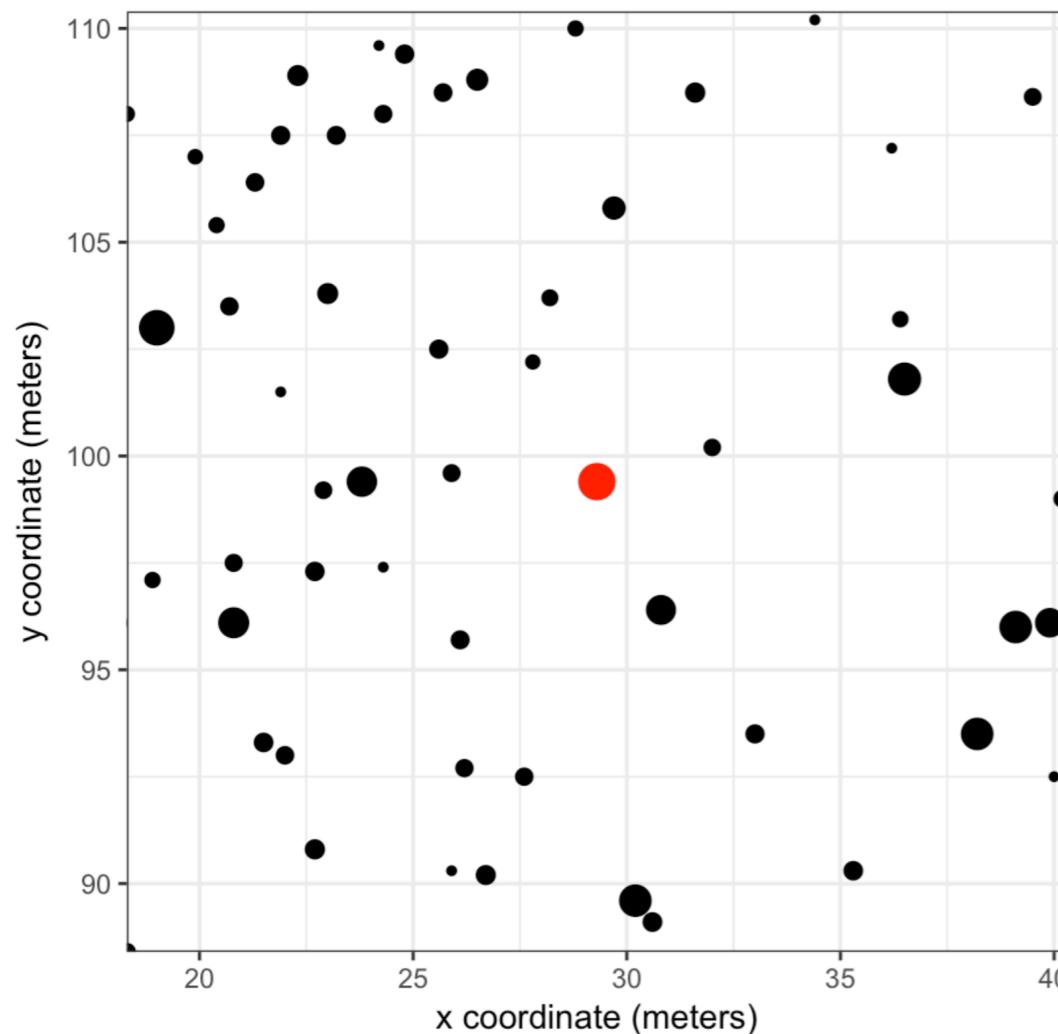
# Predictor Variables

$x_3$  : Number and size of competitor trees (biomass)

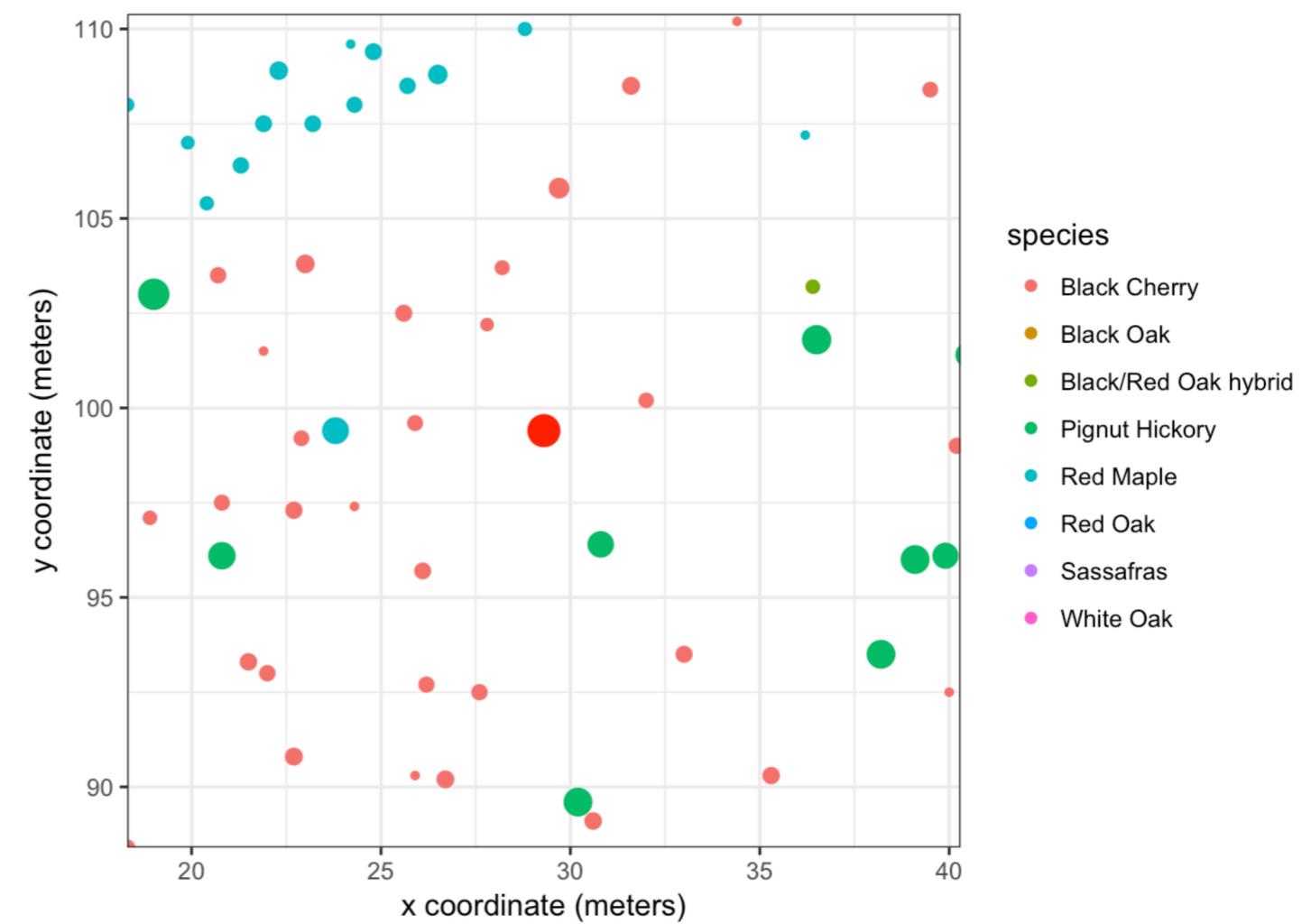


# Two Models of Competition

Example focal tree & competitor trees



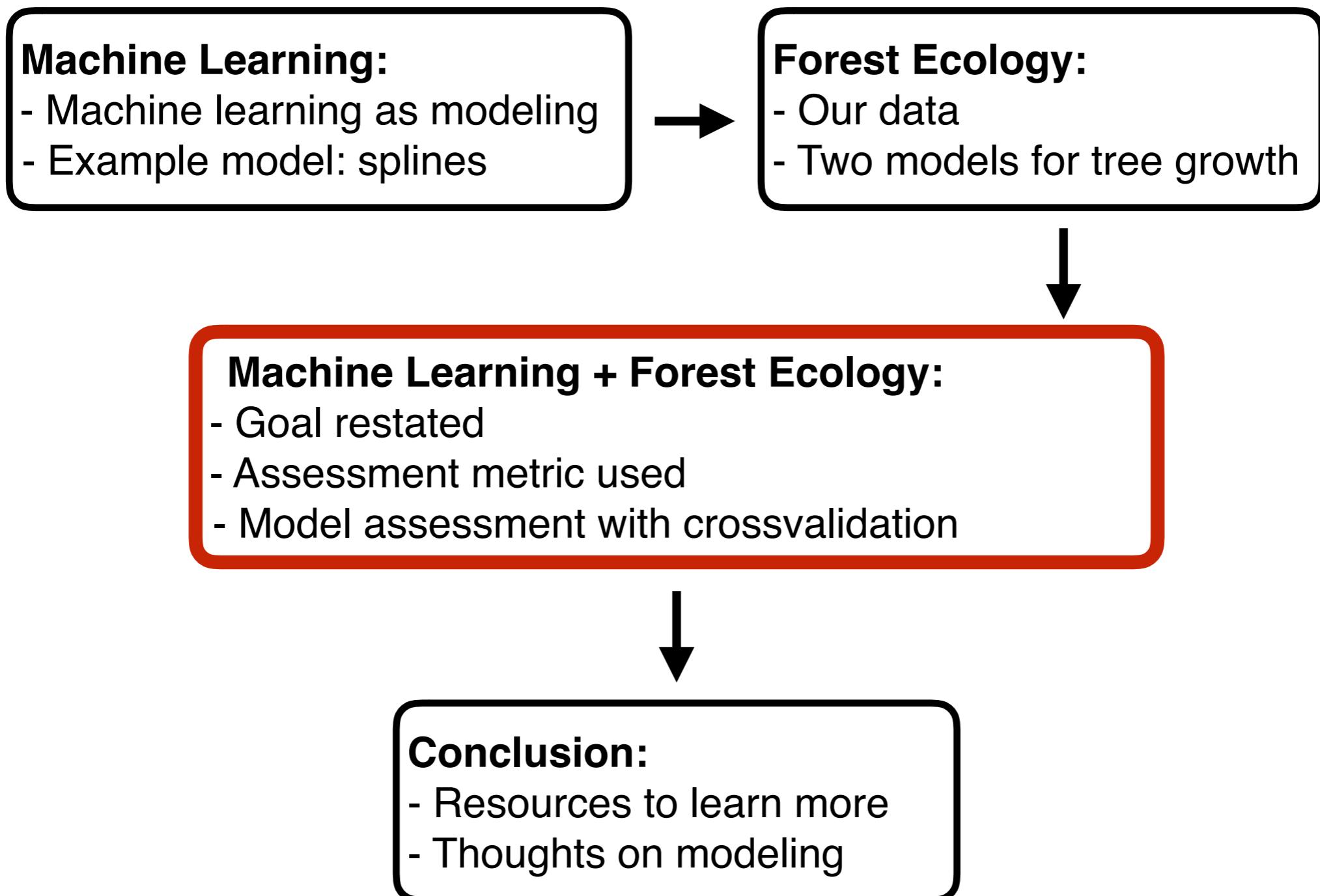
Example focal tree & competitor trees



Which model is better?

Yea or nay on distinguishing competitor species?

# Road Map



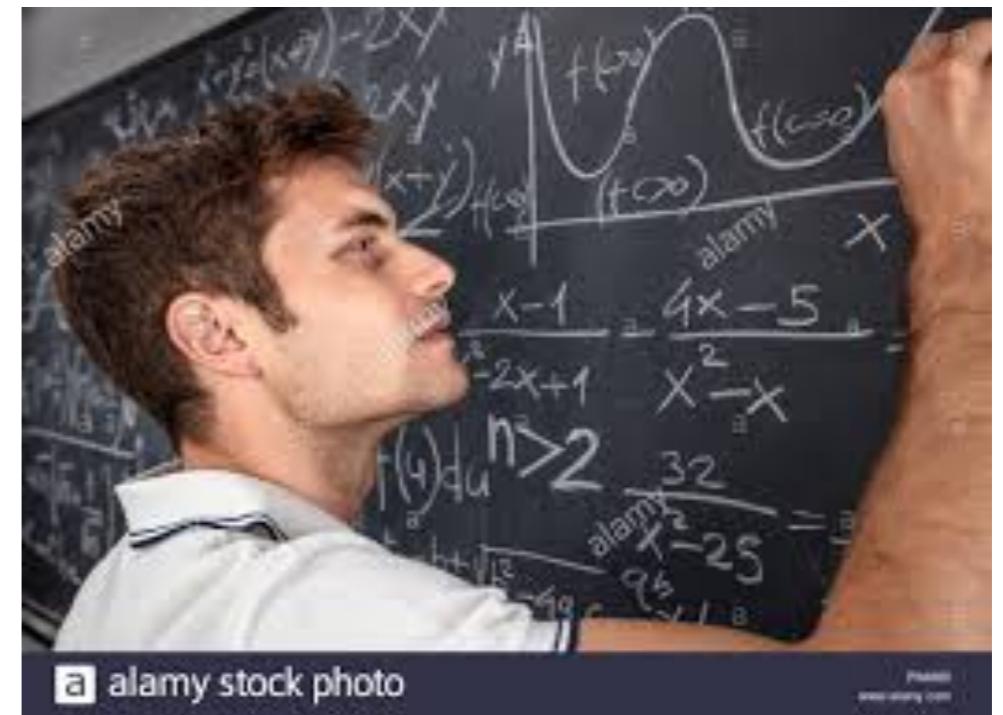
# Machine Learning & Forest Ecology

- **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...



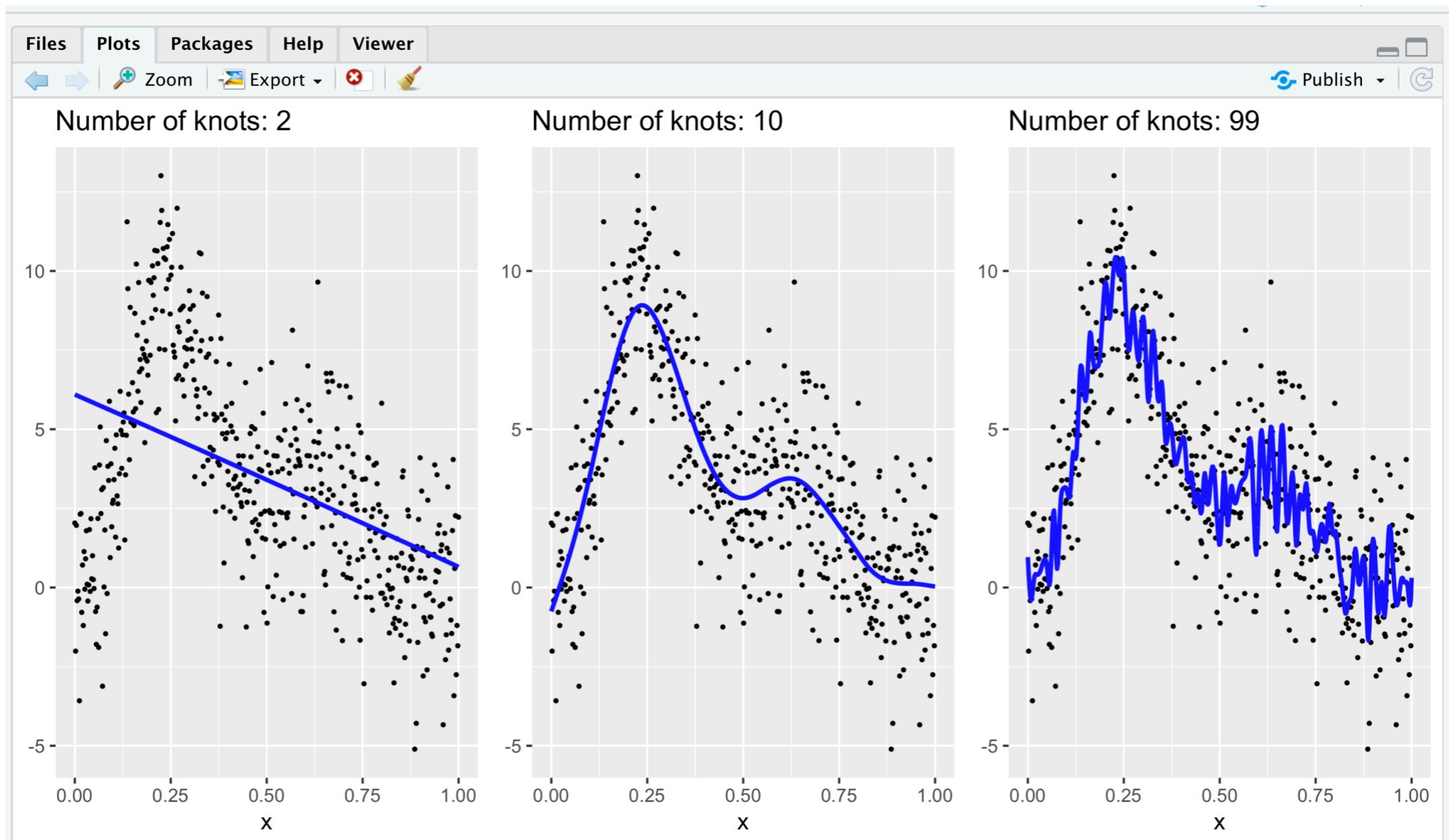
# Mean Squared Error

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

The screenshot shows the Kaggle Leaderboard for the "Google Analytics Customer Revenue Prediction" competition. The competition is a "Featured Prediction Competition" with \$45,000 in prize money. It is an RStudio competition with 1,104 teams and has one month left. The current leader is Marwen Sallem with a score of 0.0000, achieved using the "Root Mean Squared Error" metric. The competition has received 19 views.

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	Marwen Sallem			Root Mean Squared Error 0.0000	2	2mo
2	—	Paulo Pinto	</> 1line Perfect Score		0.0000	13	1mo
3	—	Its Me			0.0000	6	2mo

# Hold up! What about underfitting vs overfitting?



Underfit!

"Just right!"

Overfit!

# How? Using Validation Set Approach

1 2 3

n

Split your data into:



7 22 13

91

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

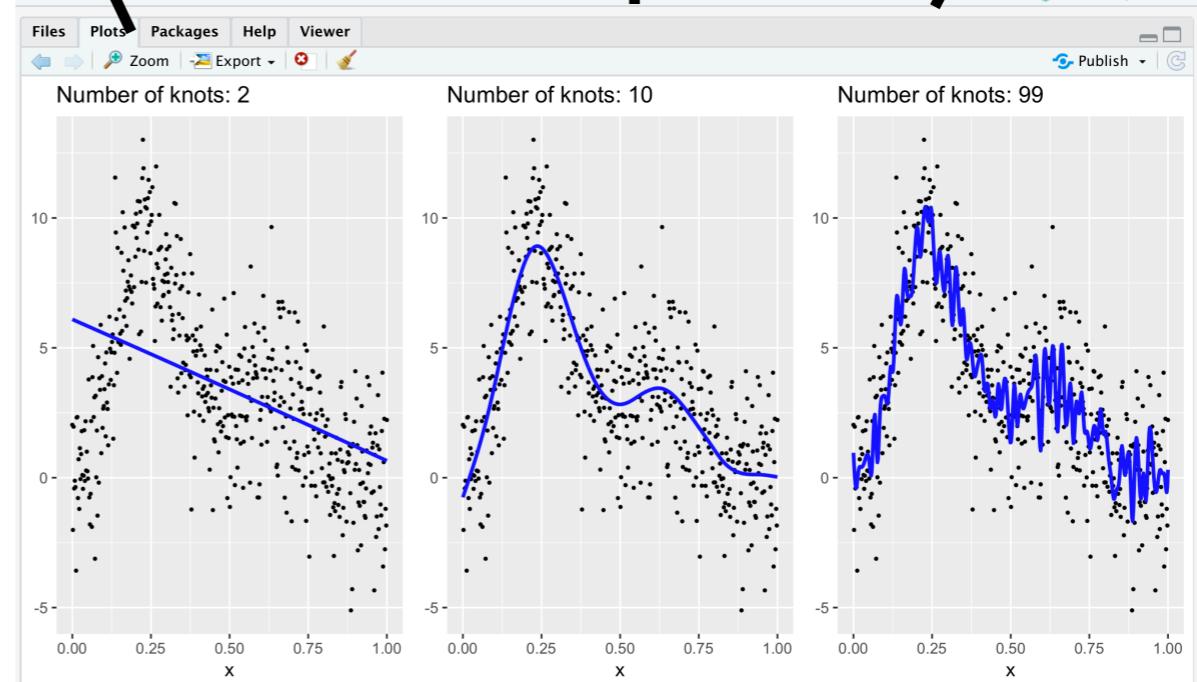
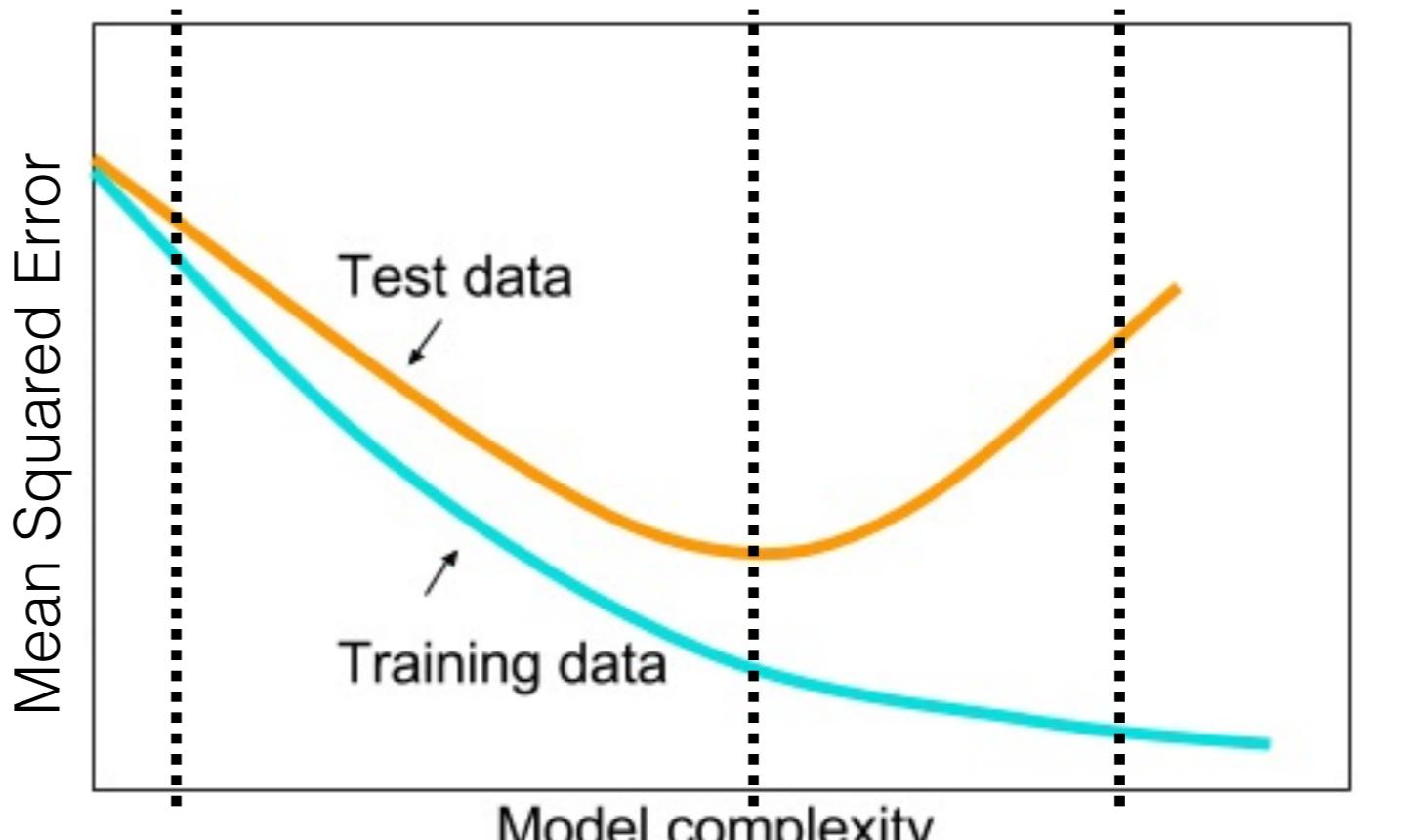
Fit your model on  
the ***training*** data

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

Underfit!

“Just right!”

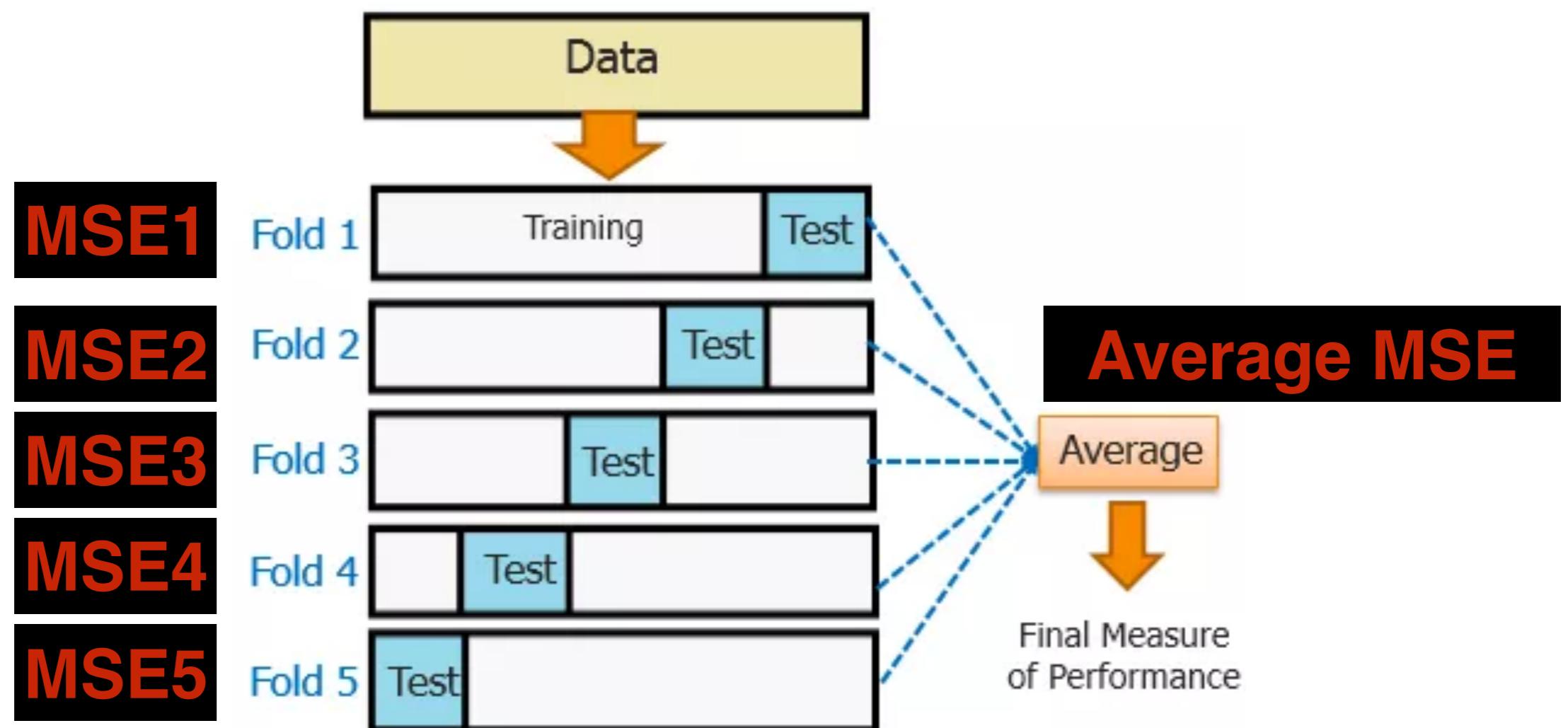
Overfit!



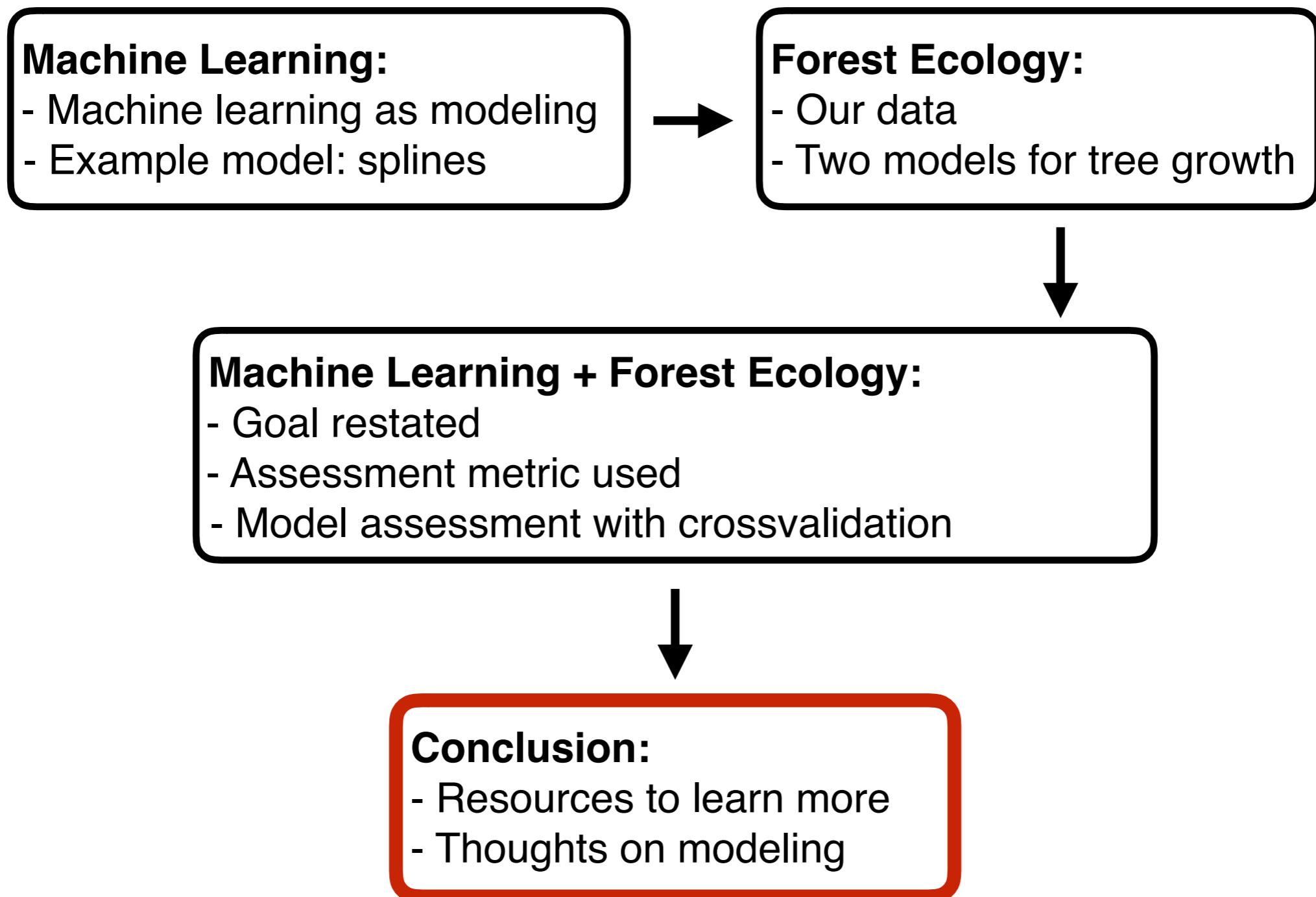
Recall for splines,  
the # of knots controls  
the **model complexity**

# Generalization: 5-Fold Crossvalidation

Repeat validation training/test set split 5 times:

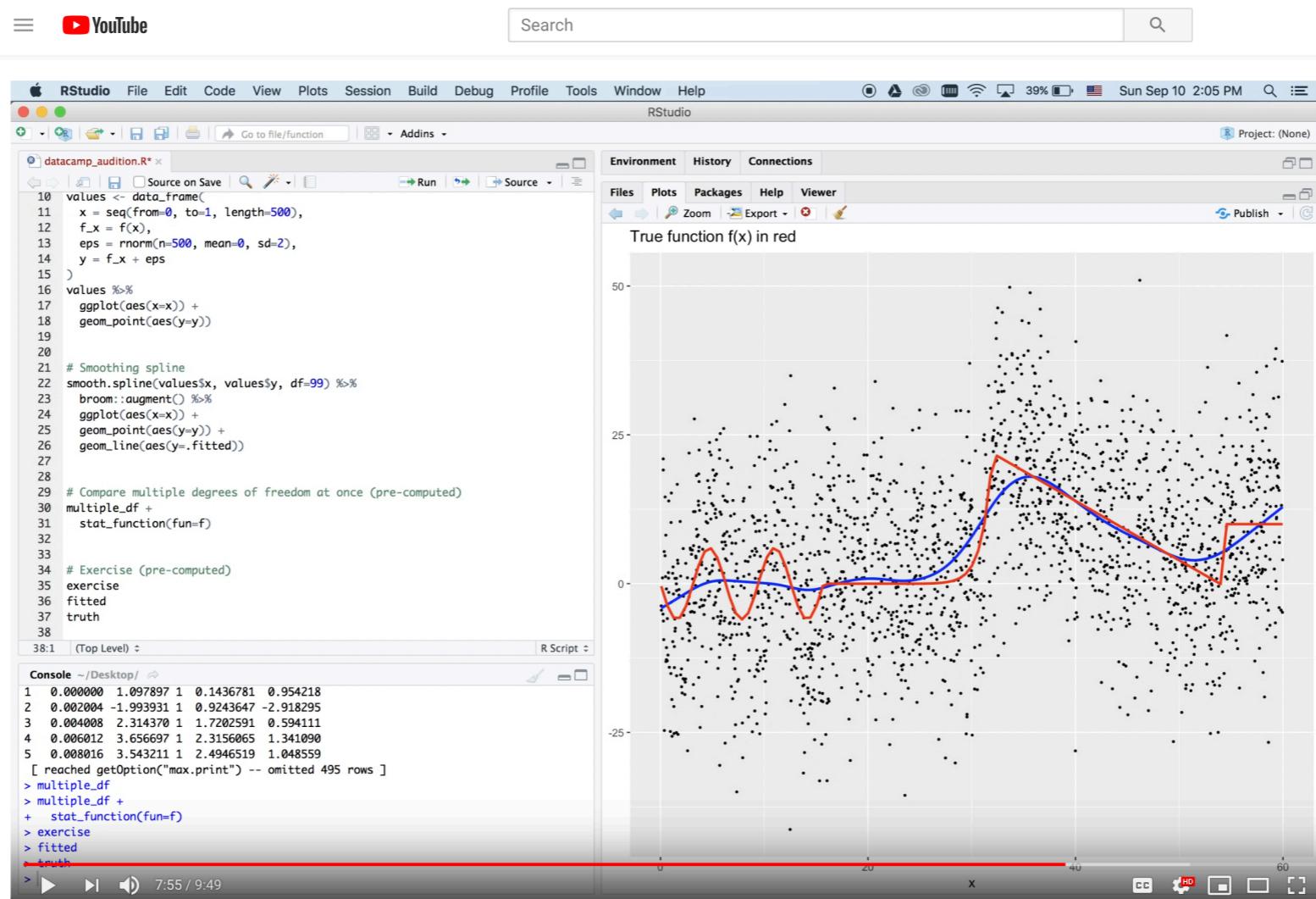


# Road Map



# 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](https://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:

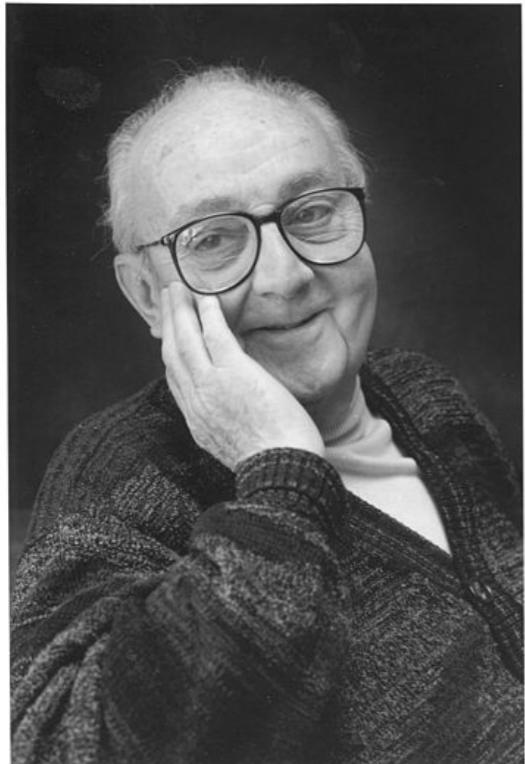
Scenario:

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

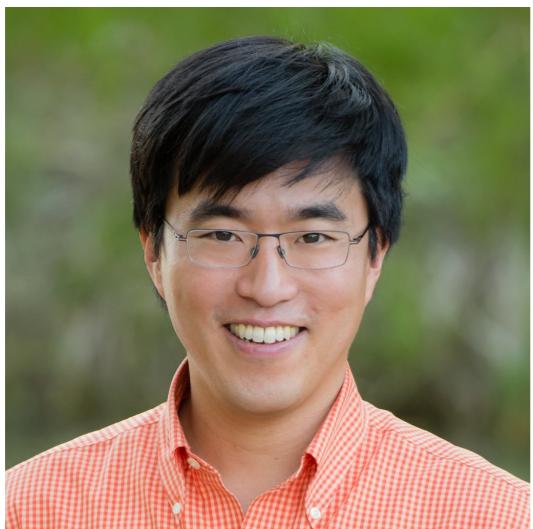
... might not be the  
same for these folks:



# To Close: Two Quotes on Modeling



“All models are wrong,  
but some are useful.”  
George Box



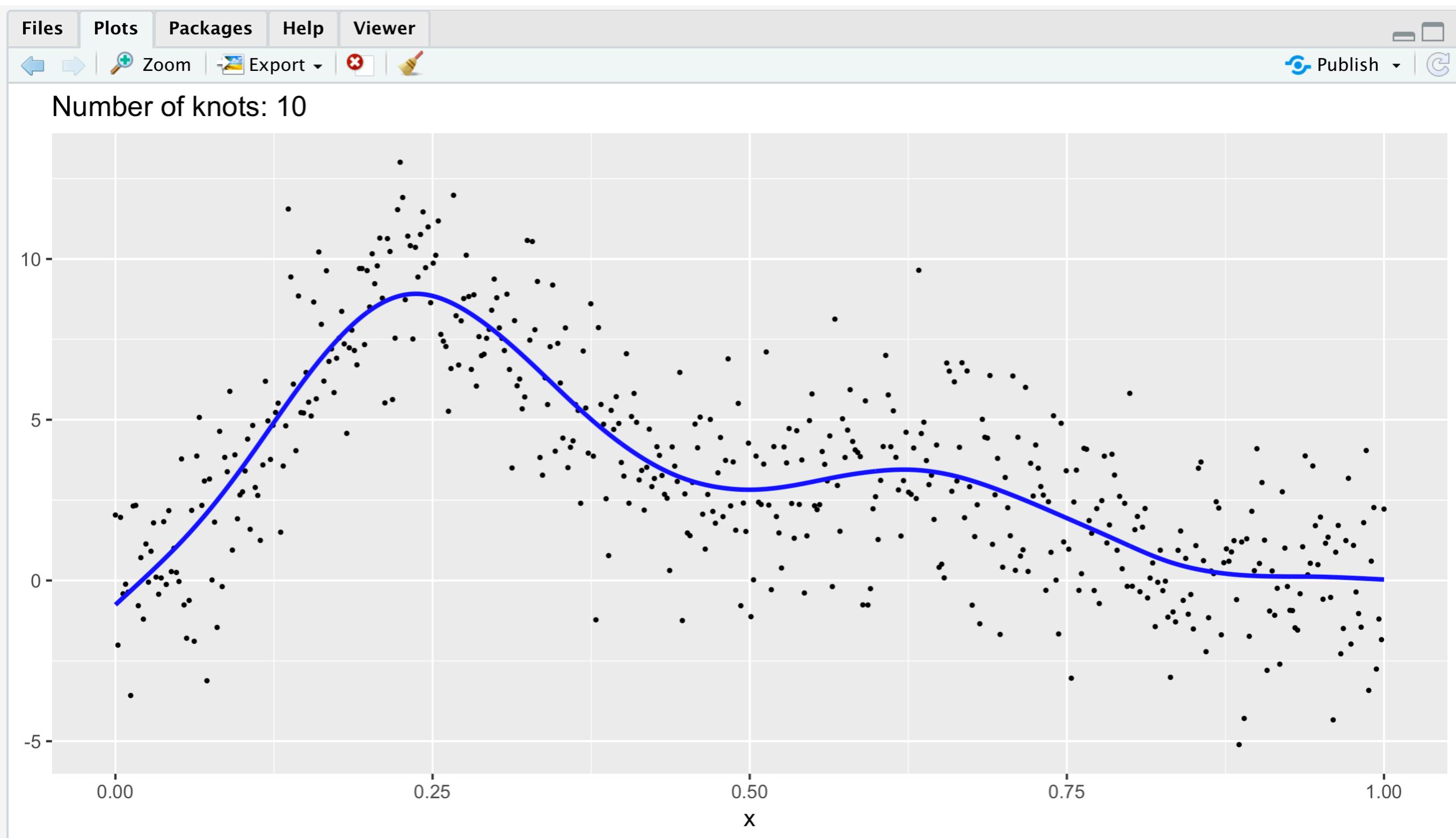
“WTF is up with your  
 $\hat{f}(x)$ ?” @rudeboybert

Thanks!

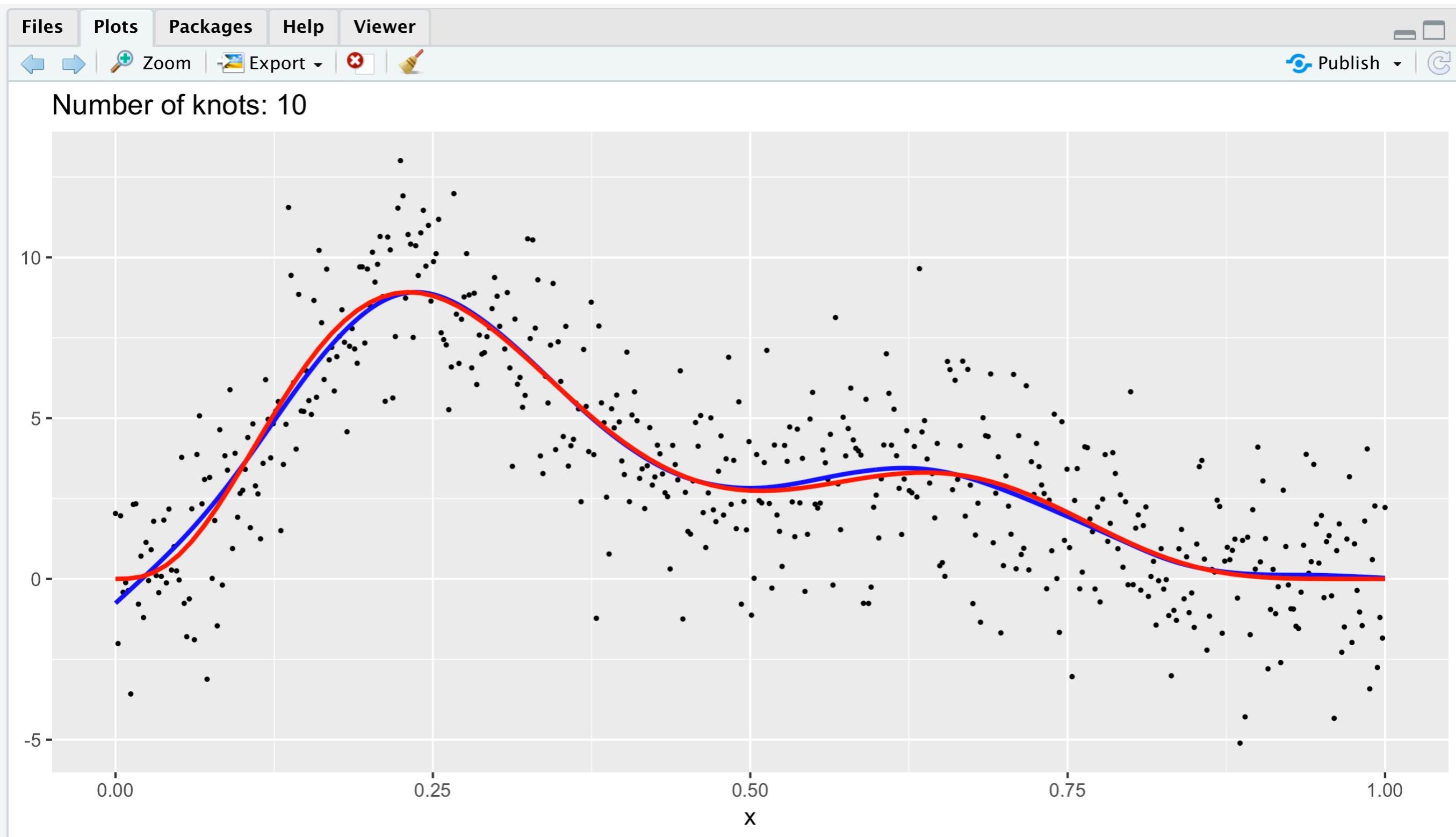
Before I go: A “Wizard of Oz” Reveal...



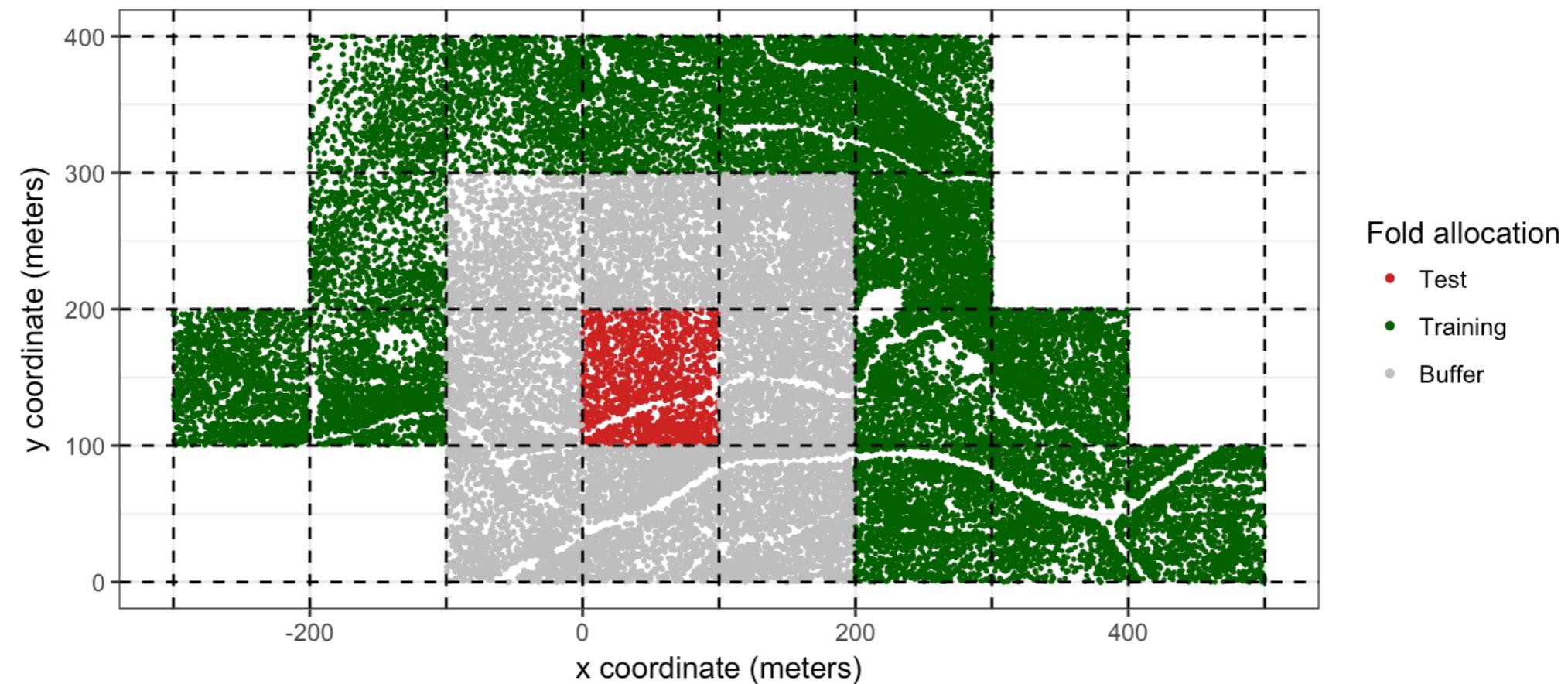
# Our approximated $\hat{f}(x)$ was pretty close...



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

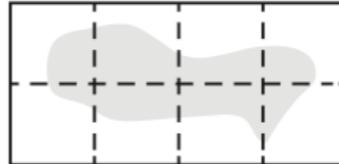
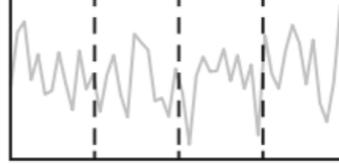
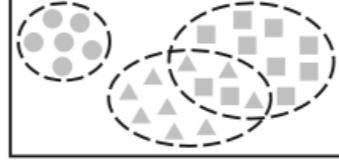


# Our Data is Spatial: Spatial Crossvalidation



# Resource 3: Paper

“Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure” [Roberts \(2017\)](#)

Dependence structure	Parametric solution	Blocking	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	