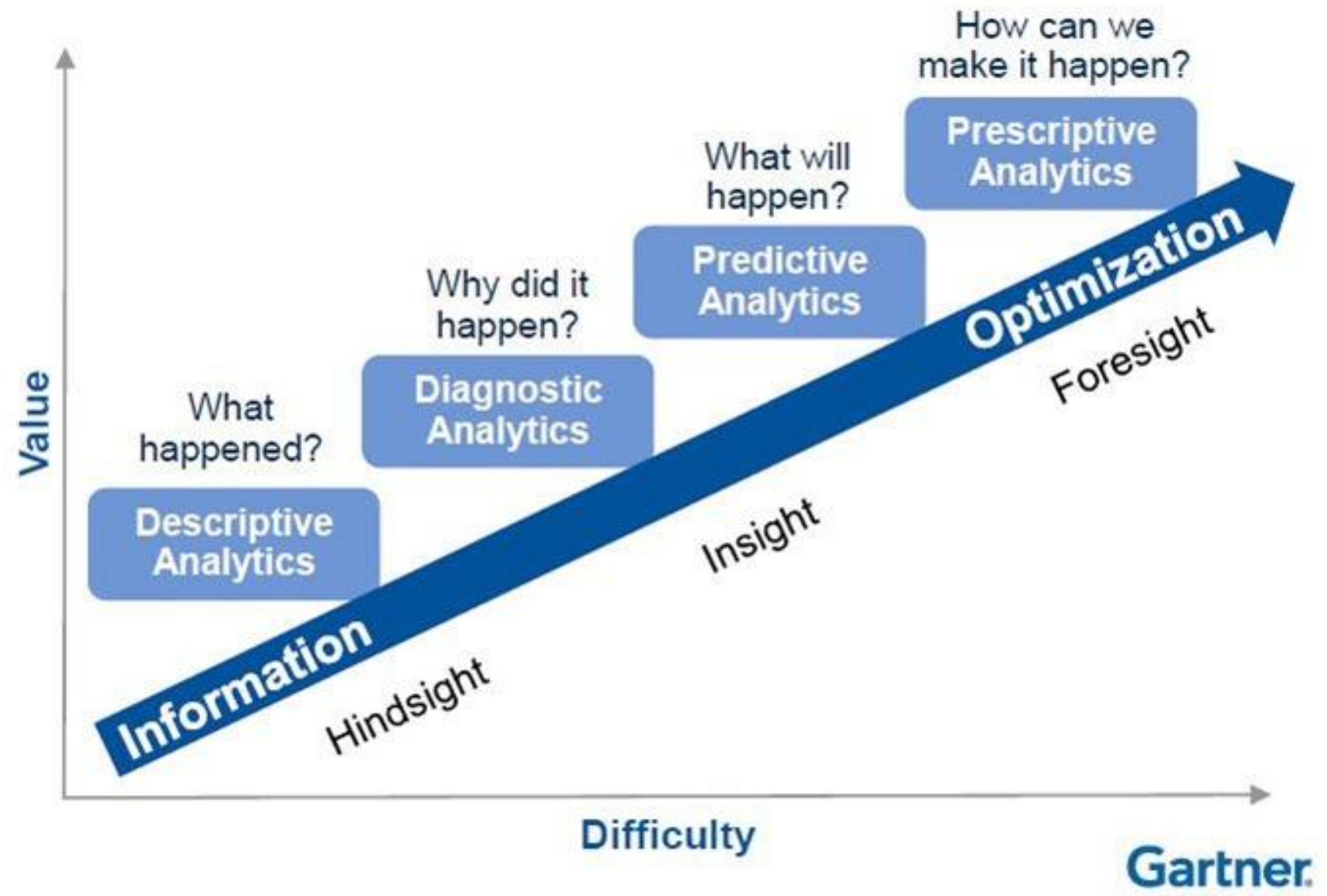
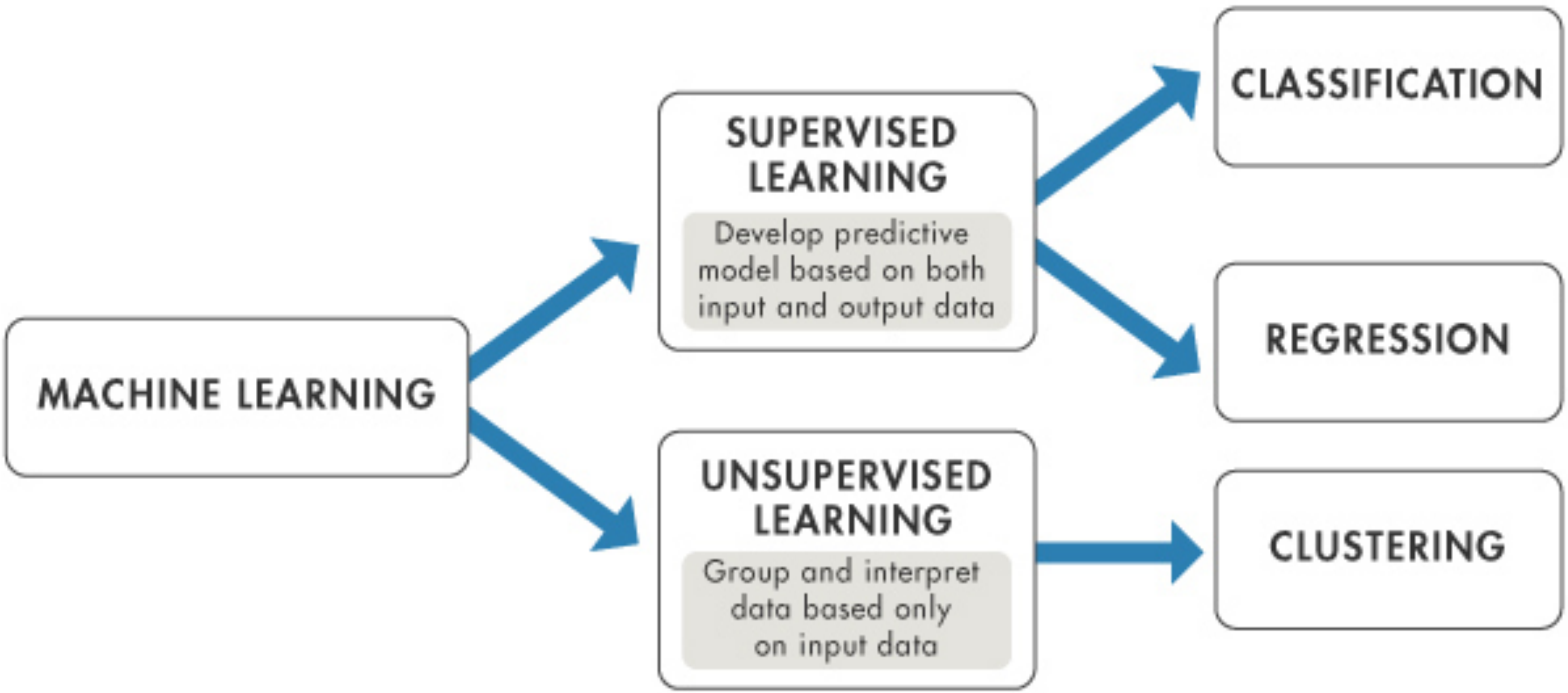
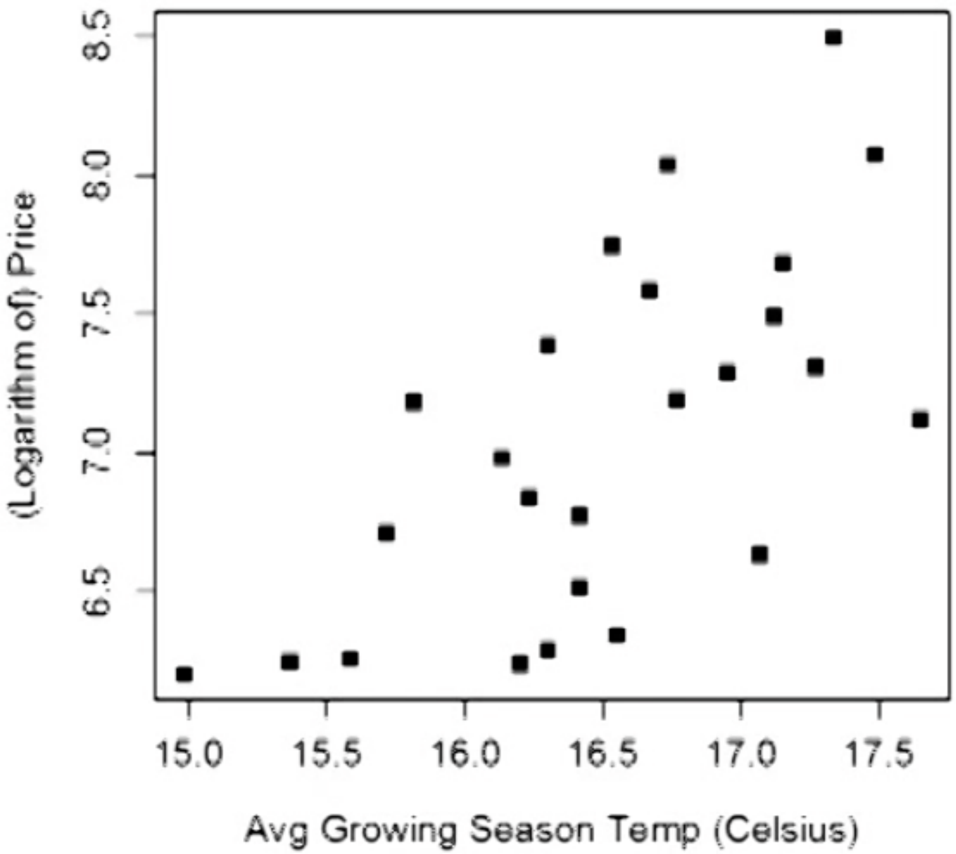
Linear Regression

Big Data Analytics

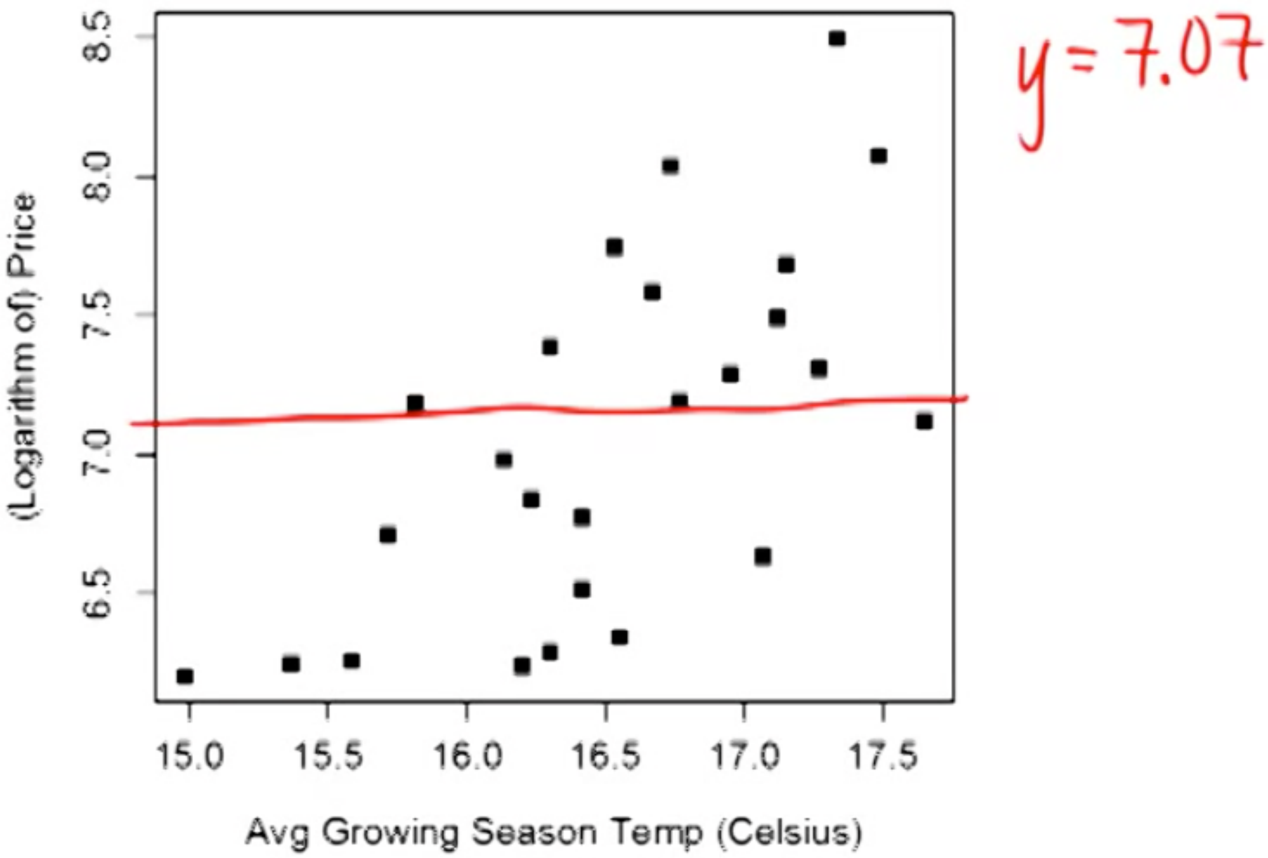



MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

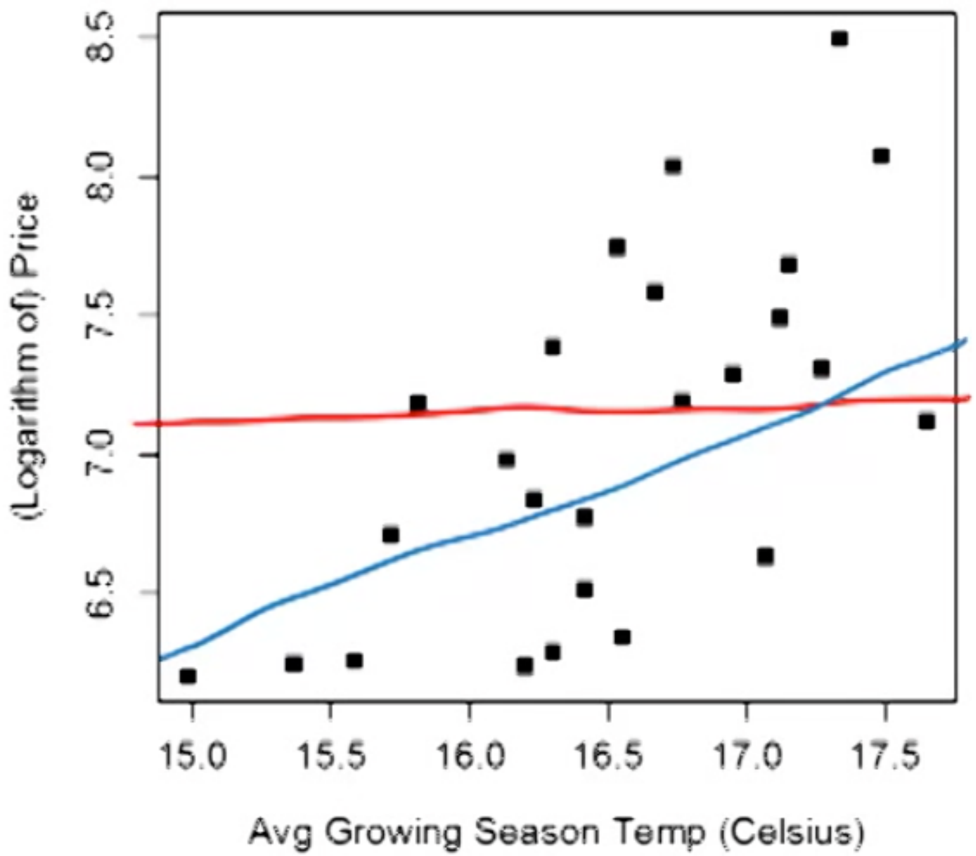
One-Variable Linear Regression



One-Variable Linear Regression



One-Variable Linear Regression



$y = 7.07$

$y = 0.5(AGST) - 1.25$

The Regression Model

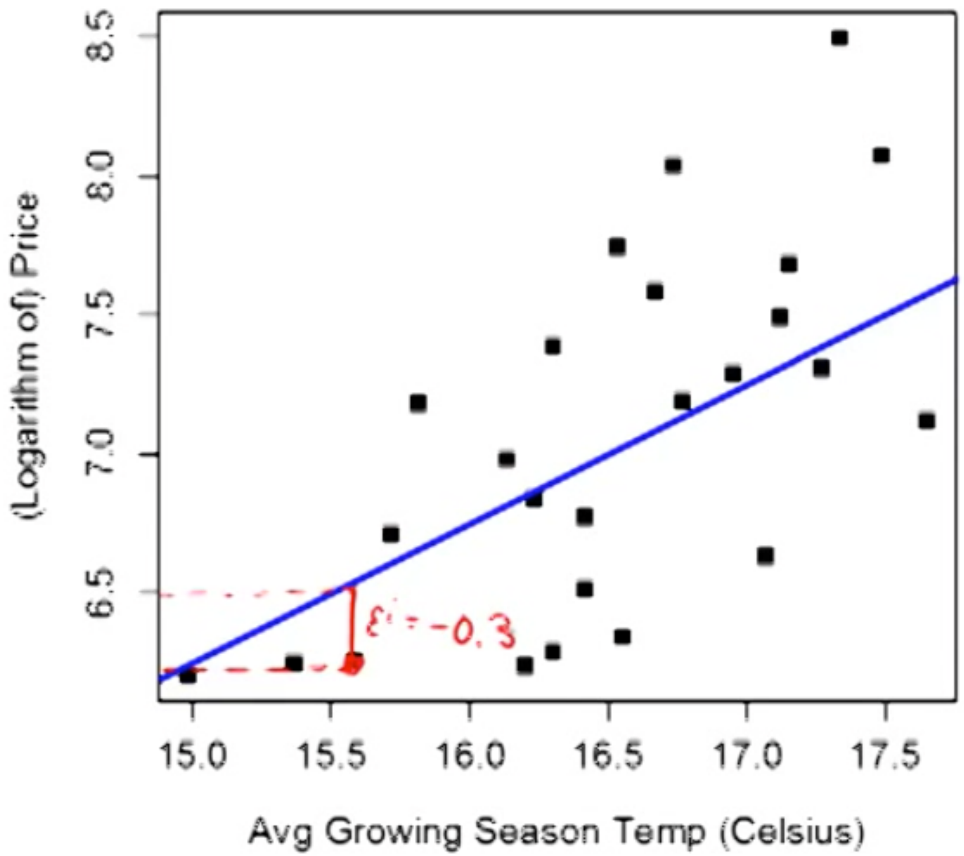
- One-variable regression model

$$y^i = \beta_0 + \beta_1 x^i + \epsilon^i$$

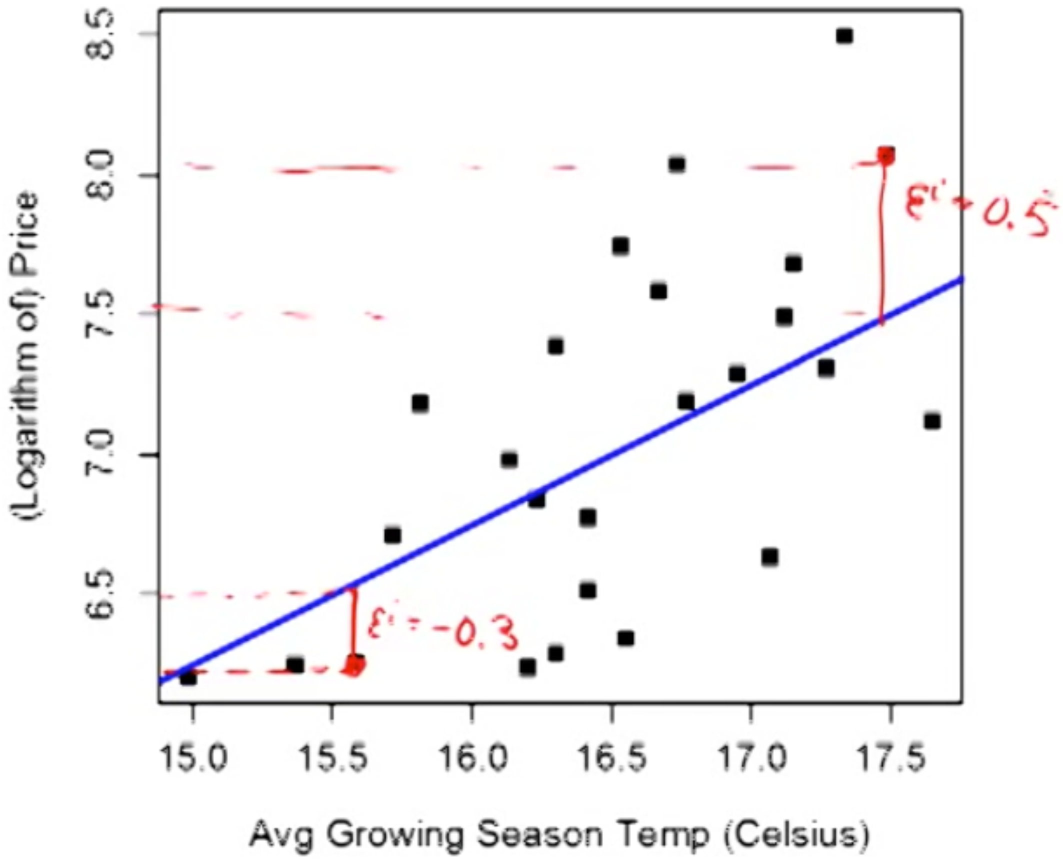
y^i = dependent variable (wine price) for the i^{th} observation
x^i = independent variable (temperature) for the i^{th} observation
ϵ^i = error term for the i^{th} observation
β_0 = intercept coefficient
β_1 = regression coefficient for the independent variable

- The best model (choice of coefficients) has the smallest error terms

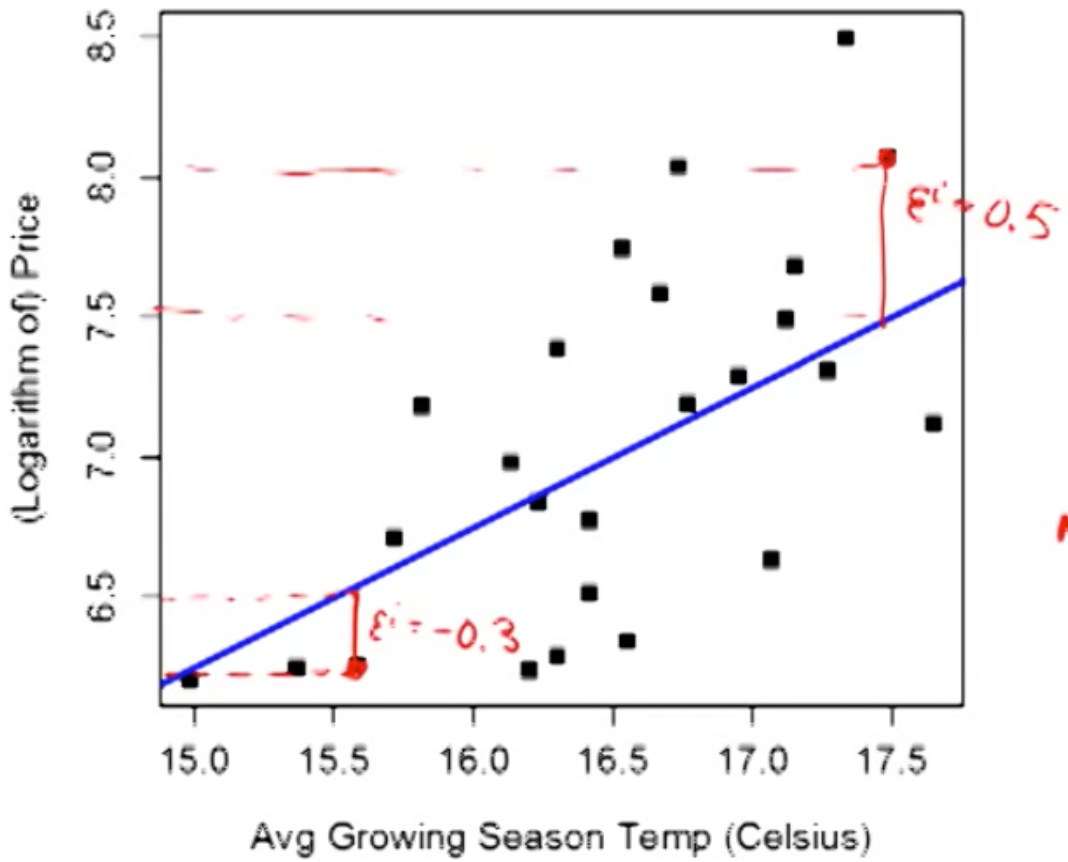
Selecting the Best Model



Selecting the Best Model



Selecting the Best Model

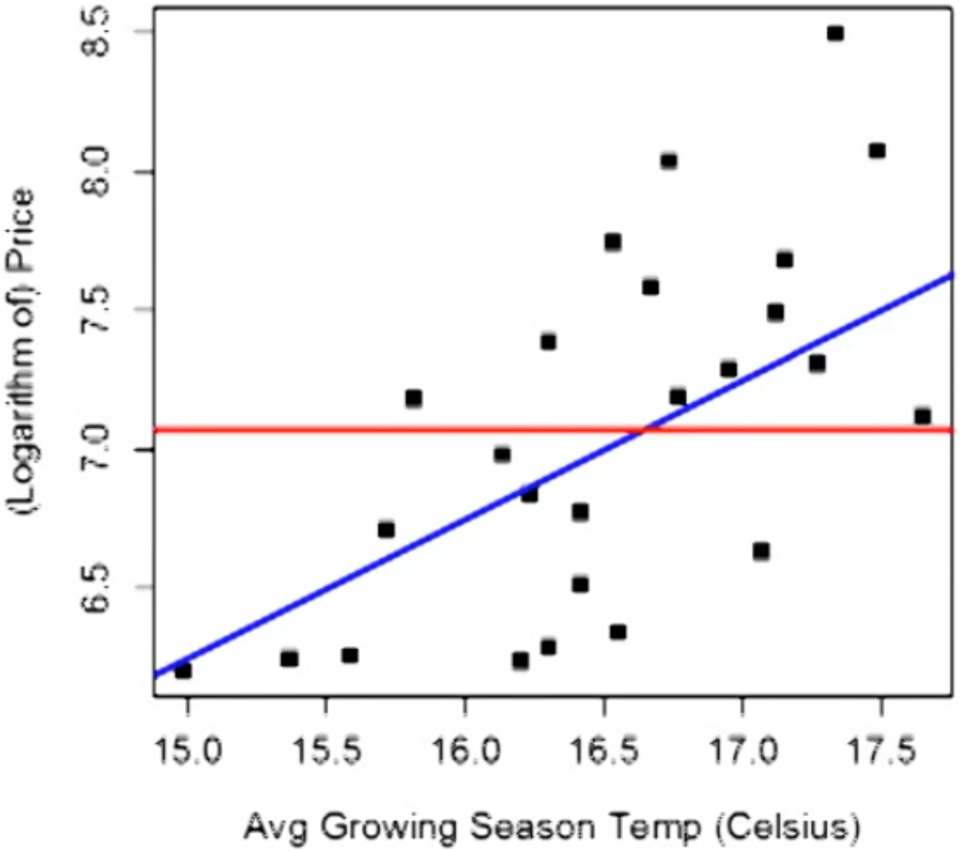


SSE = Sum of Squared Errors

$= (\varepsilon^1)^2 + (\varepsilon^2)^2 + \ldots + (\varepsilon^N)^2$

N = # data points

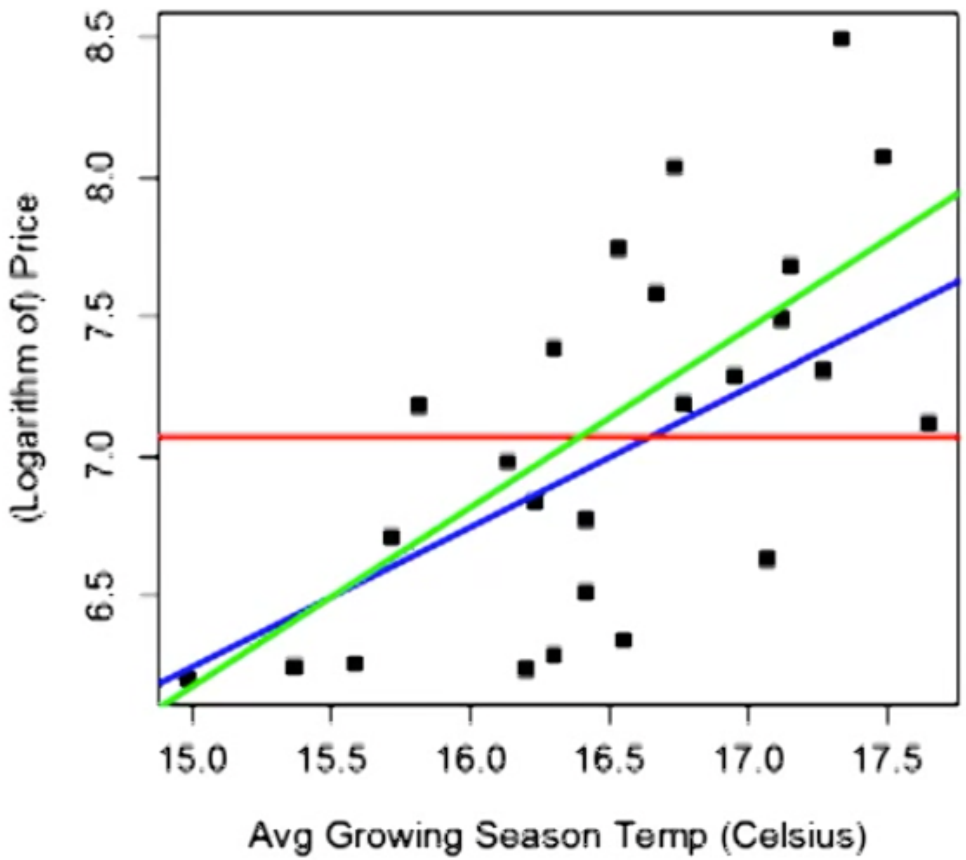
Selecting the Best Model



SSE = 10.15

SSE = 6.03

Selecting the Best Model



SSE = 10.15

SSE = 6.03

SSE = 5.73

Other Error Measures

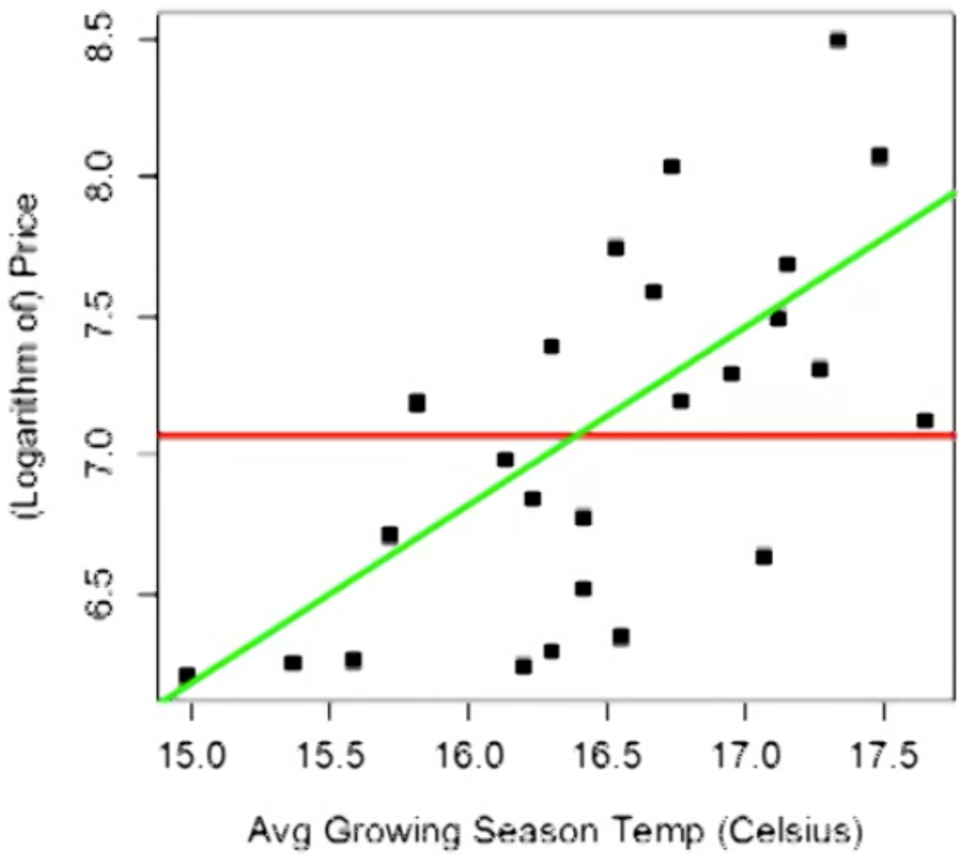
- SSE can be hard to interpret
  - Depends on N
  - Units are hard to understand

- Root-Mean-Square Error (RMSE)

$$RMSE = \sqrt{\frac{SSE}{N}}$$

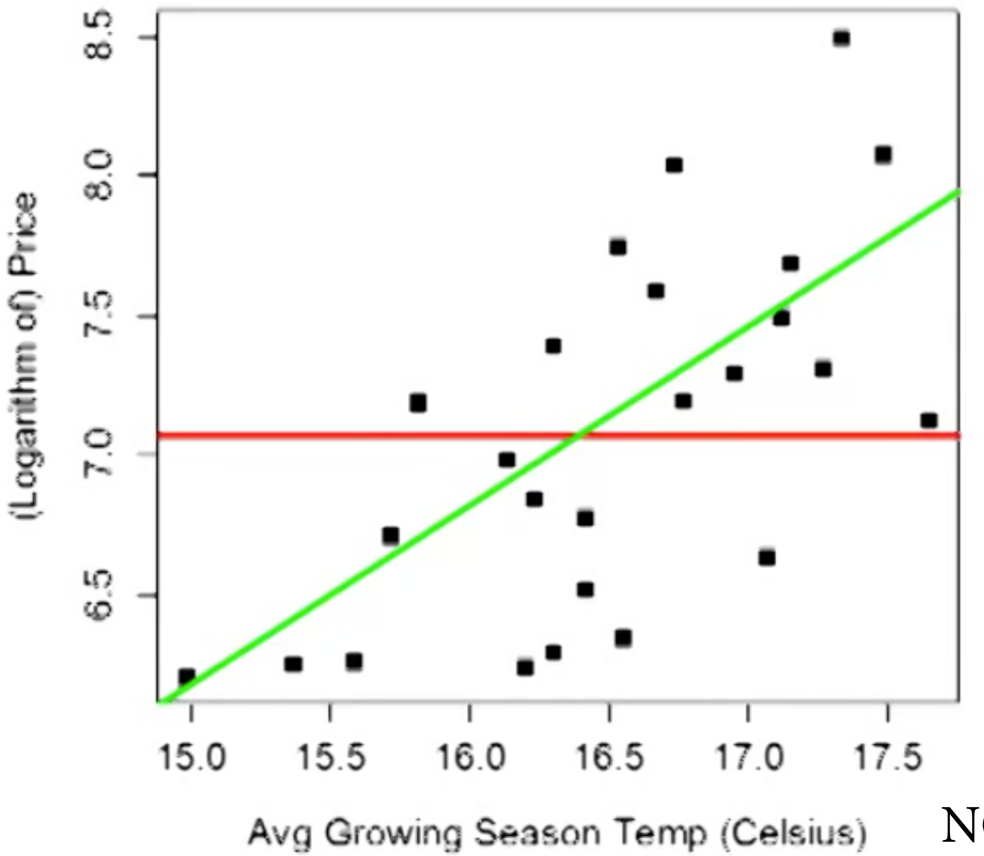
- Normalized by N, units of dependent variable

R^2



- Compares the best model to a "baseline" model
- The baseline model = mean of data does not use any variables
  - Predicts same outcome (price) regardless of the independent variable (temperature)

R^2



$SSE = 5.73$

$SST = 10.15$ = Total Sum of Squares

$$R^2 = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{5.73}{10.15}$$

$$= 0.44$$

NOTE: This identify is true only when the model is optimal: minimizing the SSE.

Interpreting R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

- R^2 captures value added from using a model
  - $R^2 = 0$ means no improvement over baseline
  - $R^2 = 1$ means a perfect predictive model
- Unitless and universally interpretable
  - Can still be hard to compare between problems
  - Good models for easy problems will have $R^2 \approx 1$
  - Good models for hard problems can still have $R^2 \approx 0$

Available Independent Variables

- So far, we have only used the Average Growing Season Temperature to predict wine prices

- Many different independent variables could be used
  - Average Growing Season Temperature
  - Harvest Rain
  - Winter Rain
  - Age of Wine (in 1990)
  - Population of France

Multiple Linear Regression

- Using each variable on its own:
  - $R^2 = 0.44$ using Average Growing Season Temperature
  - $R^2 = 0.32$ using Harvest Rain
  - $R^2 = 0.22$ using France Population
  - $R^2 = 0.20$ using Age
  - $R^2 = 0.02$ using Winter Rain
- Multiple linear regression allows us to use all of these variables to improve our predictive ability

The Regression Model

- Multiple linear regression model with k variables

$$y^i = \beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_k x_k^i + \epsilon^i$$

y^i = dependent variable (wine price) for the i^{th} observation

x_j^i = j^{th} independent variable for the i^{th} observation

ϵ^i = error term for the i^{th} observation

β_0 = intercept coefficient

β_j = regression coefficient for the j^{th} independent variable

- Best model coefficients selected to minimize SSE

Adding Variables

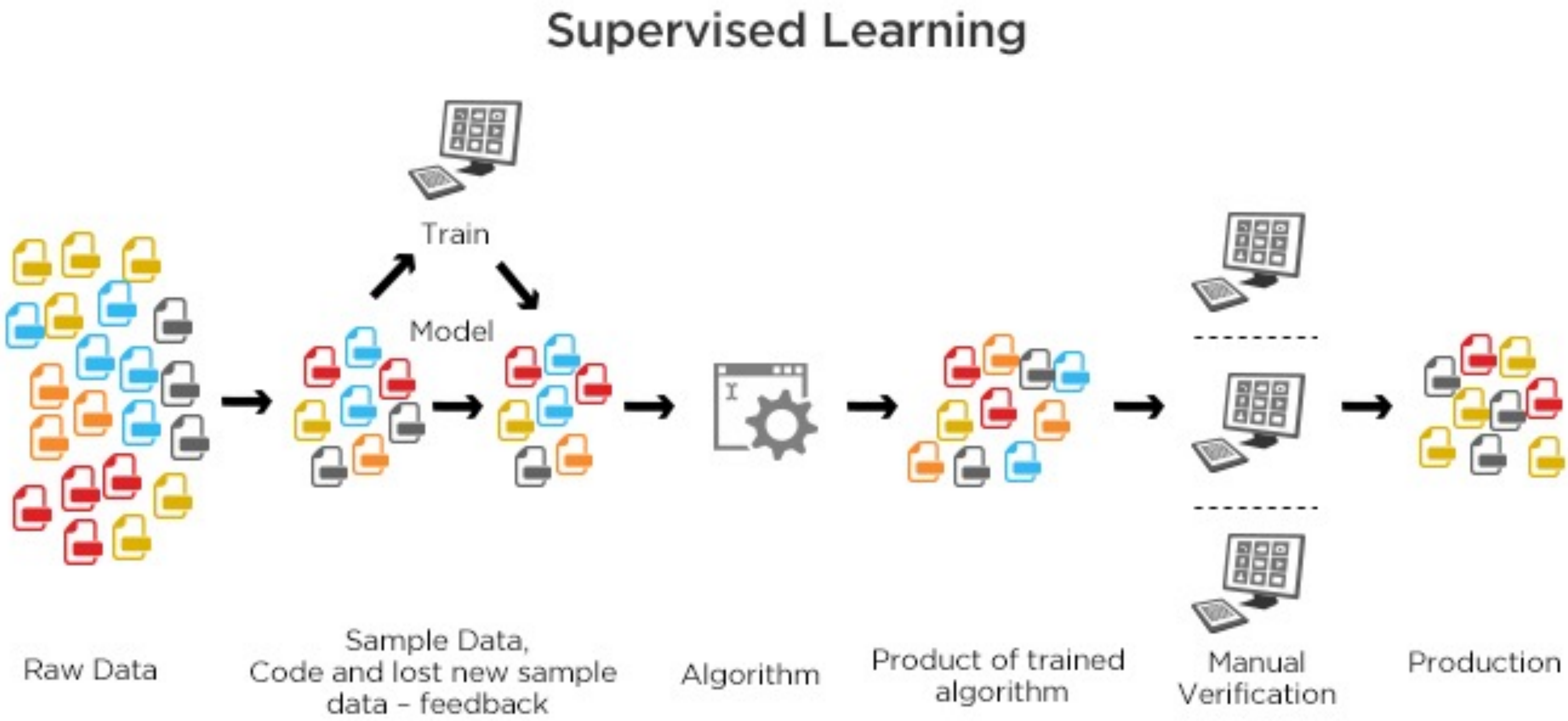
Variables	R^2
Average Growing Season Temperature (AGST)	0.44
AGST, Harvest Rain	0.71
AGST, Harvest Rain, Age	0.79
AGST, Harvest Rain, Age, Winter Rain	0.83
AGST, Harvest Rain, Age, Winter Rain, Population	0.83

- Adding more variables can improve the model
- Diminishing returns as more variables are added

Selecting Variables

- Not all available variables should be used
  - Each new variable requires more data
  - Causes *overfitting:* high R^2 on data used to create model, but bad performance on unseen data

Choosing appropriate variables among all available ones to get the best performance is an important problem, but beyond the scope of this talk.



Supervised Learning
Train
Model
Raw Data
Sample Data,
Code and lost new sample
data - feedback
Algorithm
Product of trained
algorithm
Manual
Verification
Production