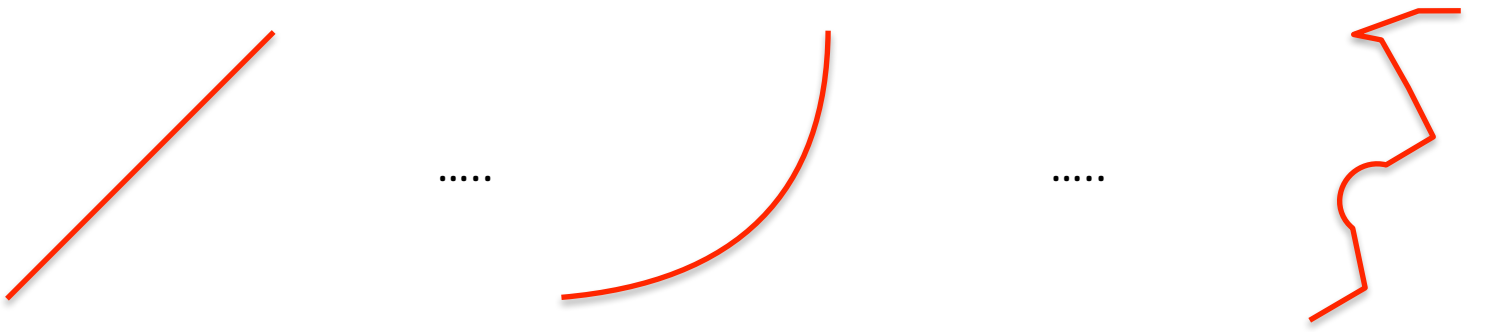
Data Science 4

TILBURG UNIVERSITY

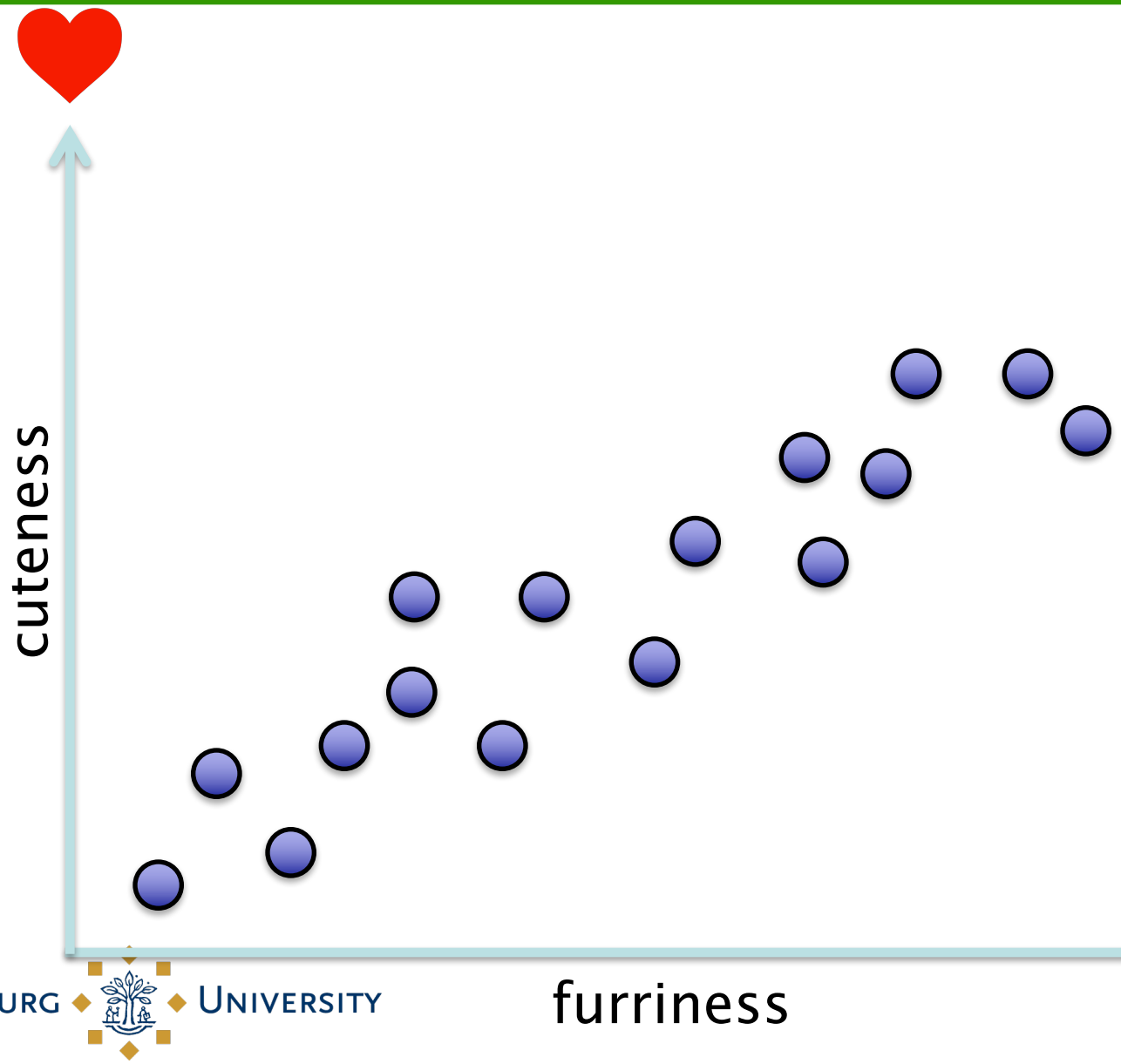
Understanding Society

Overview

- Linear Regression
- Polynomial Regression
- Nearest Neighbor Classification and Regression
- Overfitting and Underfitting



Correlation



Description versus Prediction

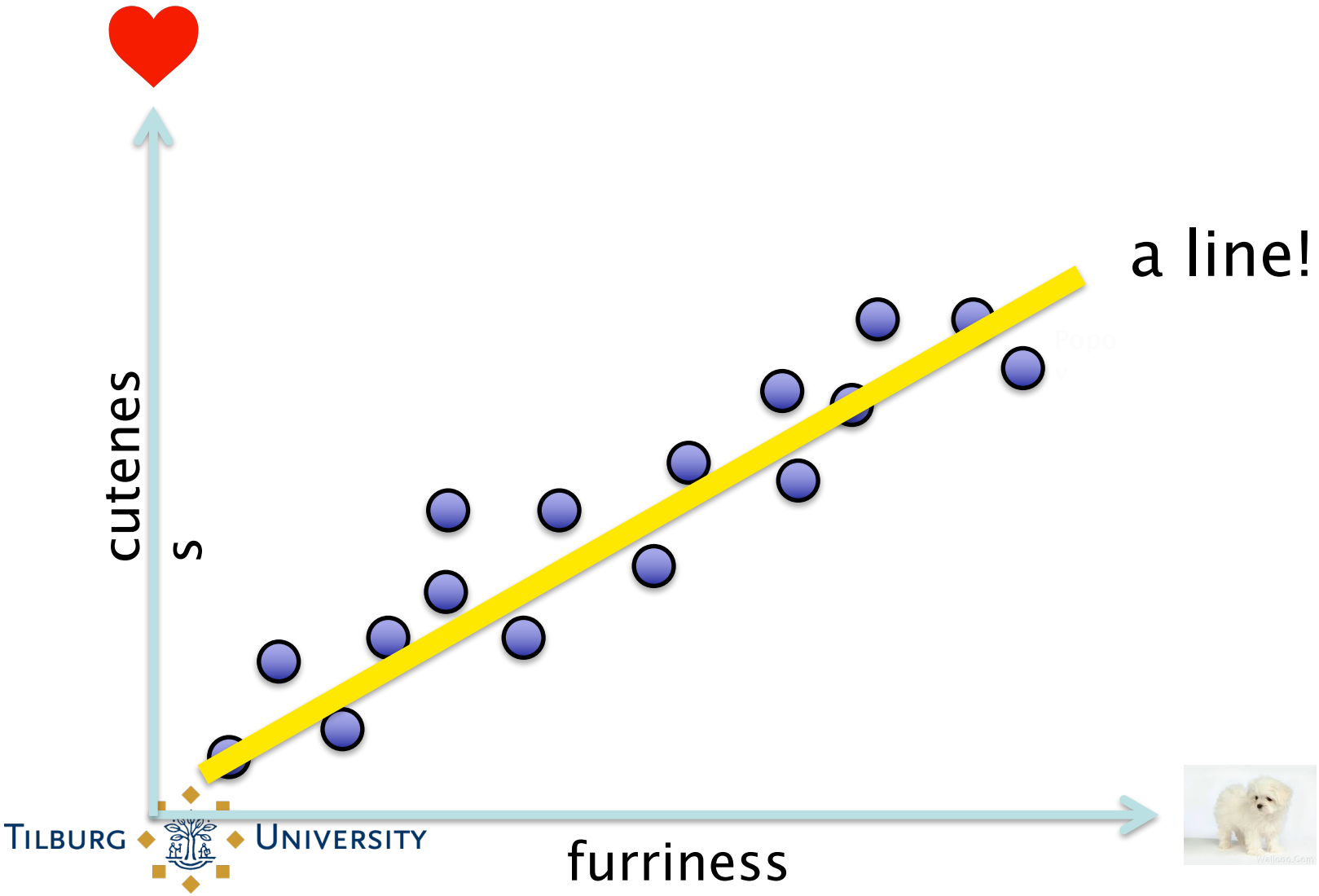
- Correlation is describing the data.
- We want to **predict** data,
- therefore we model the data

What is a model?

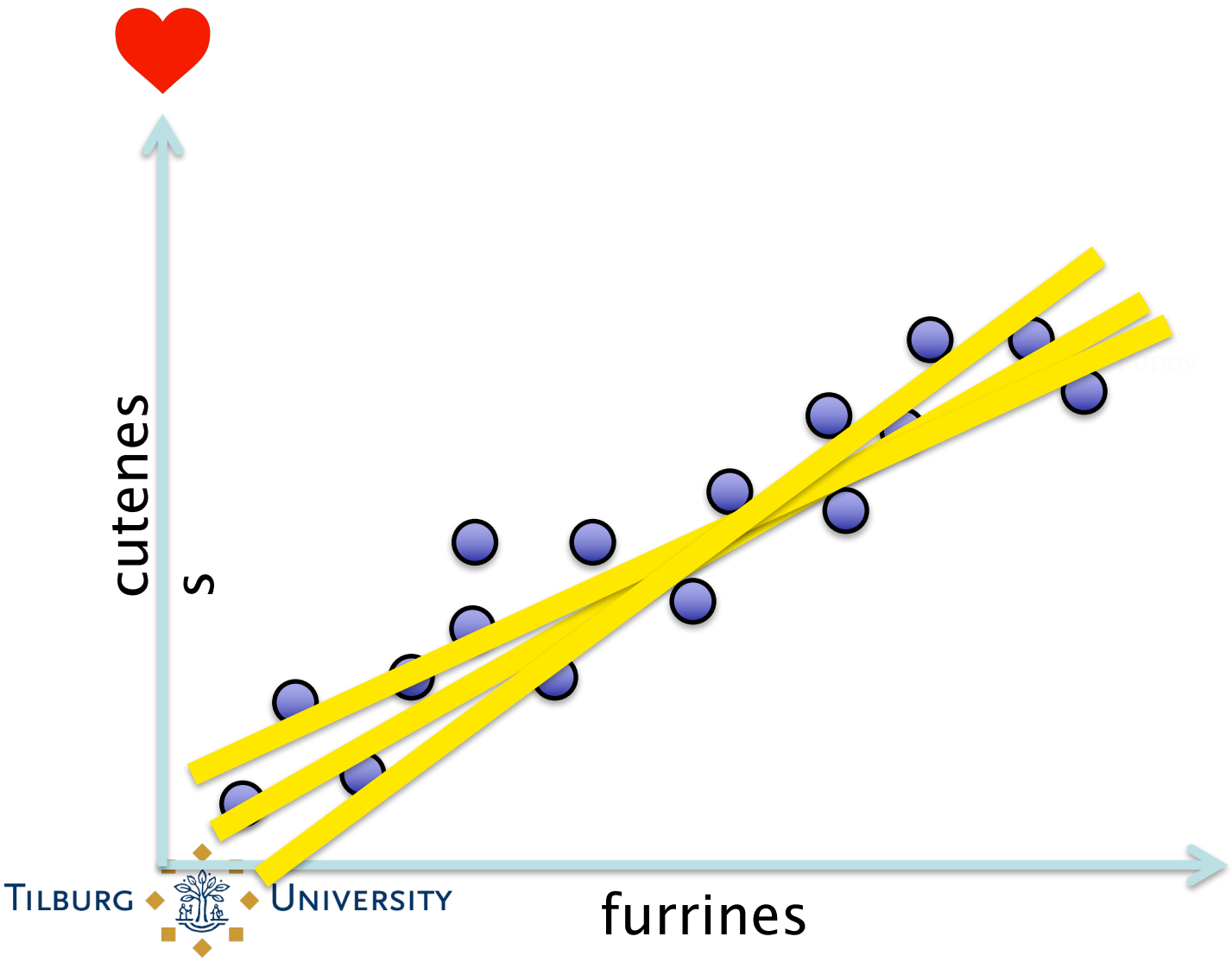
- In machine learning, a model takes the feature(s) as input and generates as output a label estimate

- Input: furriness
- Output: estimate of cuteness
- Evaluation of the model: difference between estimated and true cuteness values

What is a suitable model?

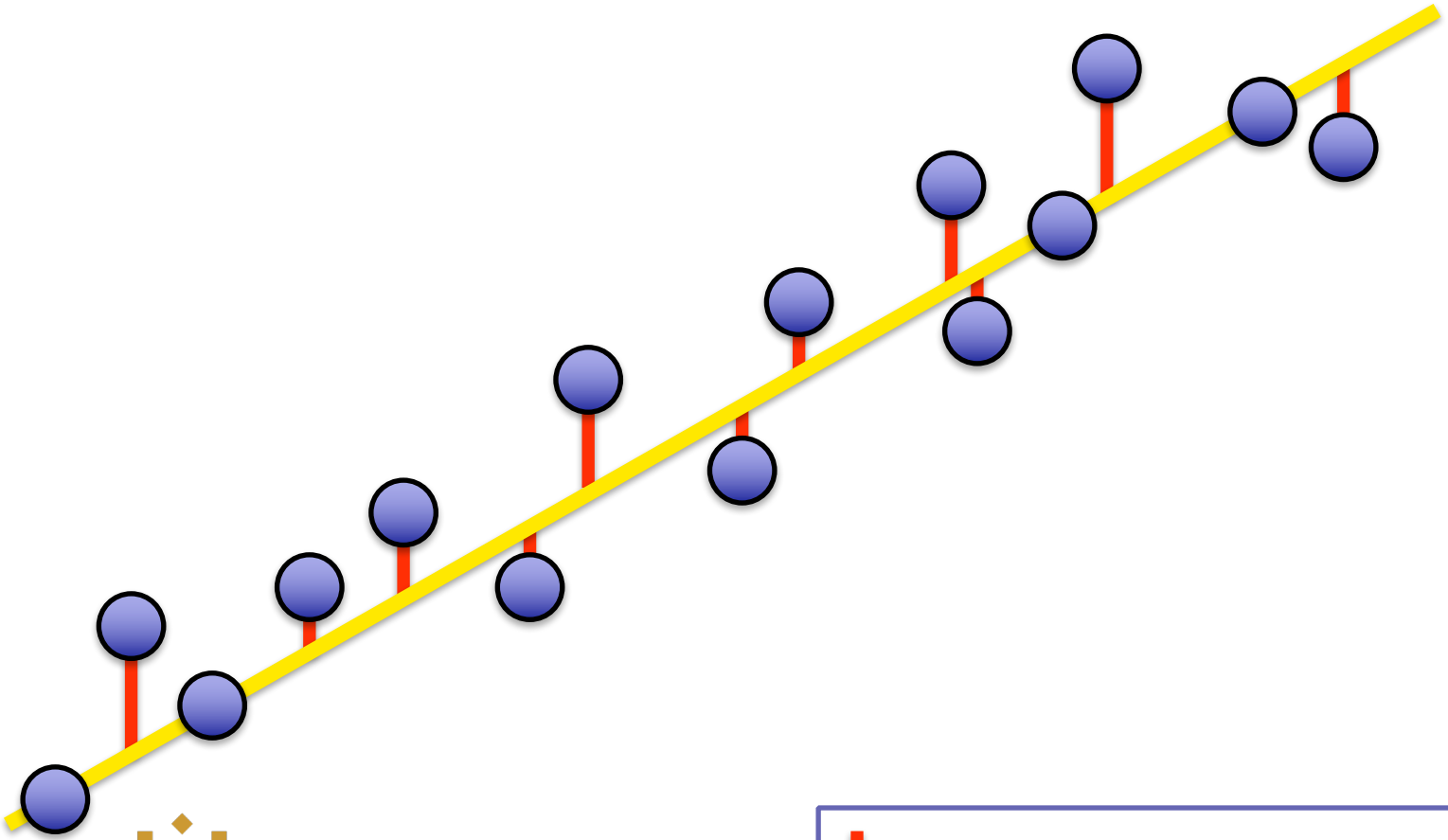


What is the best model (line)?



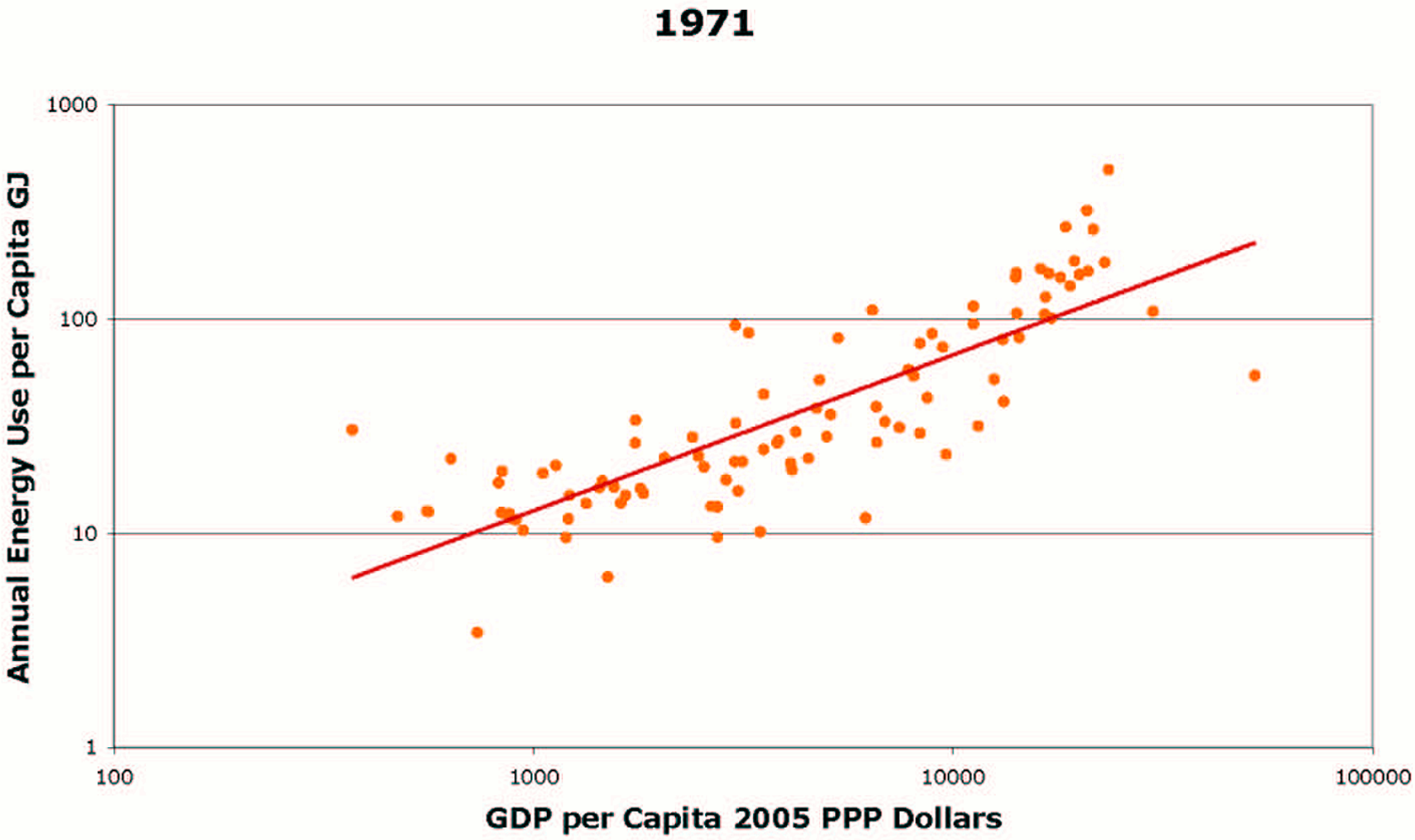
TILBURG UNIVERSITY

Answer: the best fitting line



deviations from the model = error

Can be determined automatically…



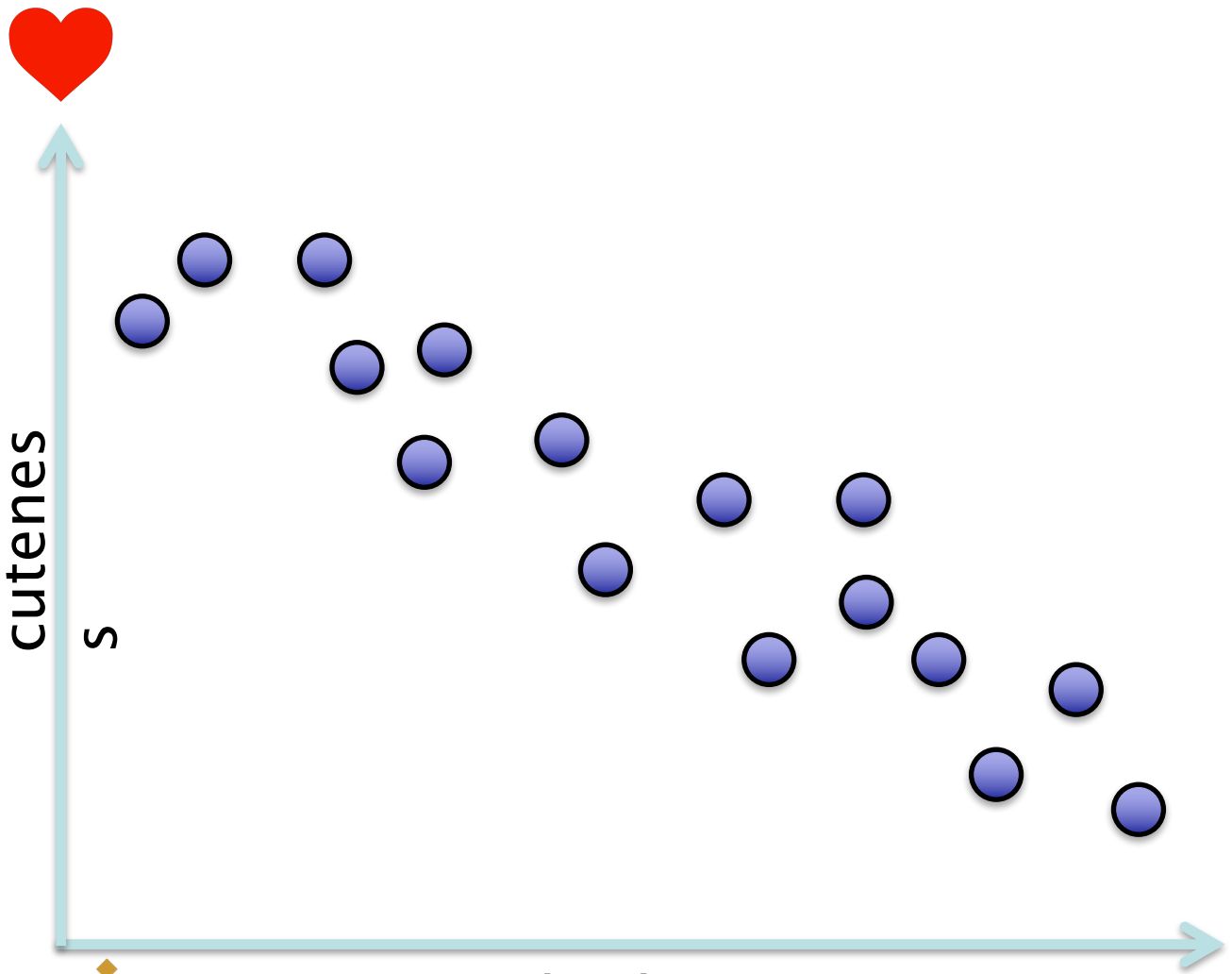
Regression Equation

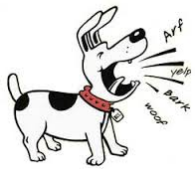
CUTENESS = **a** FURRINESS + **c**

a is the slope of the FURRINESS line —> positive value

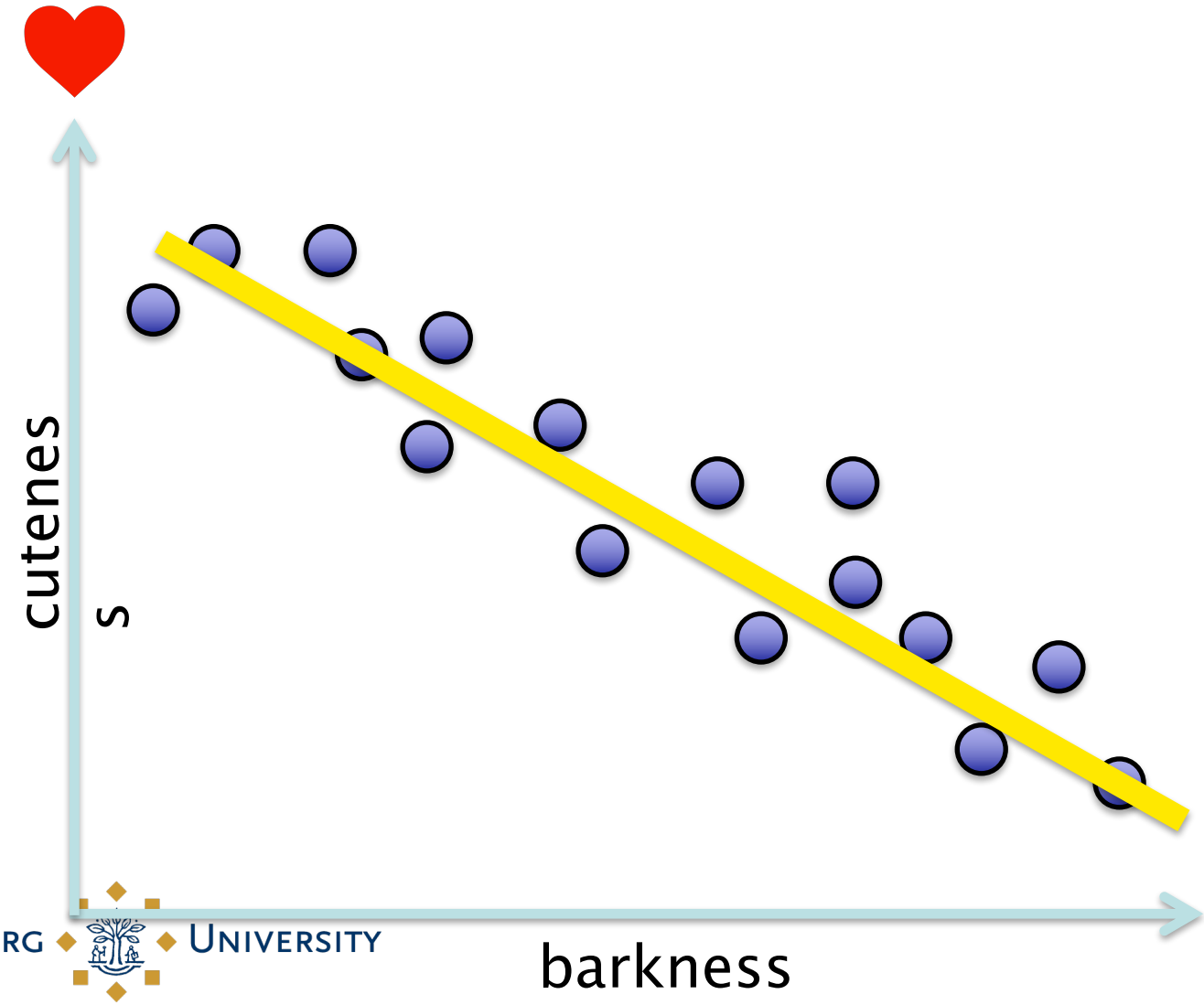
c is a constant

Adding a second feature…

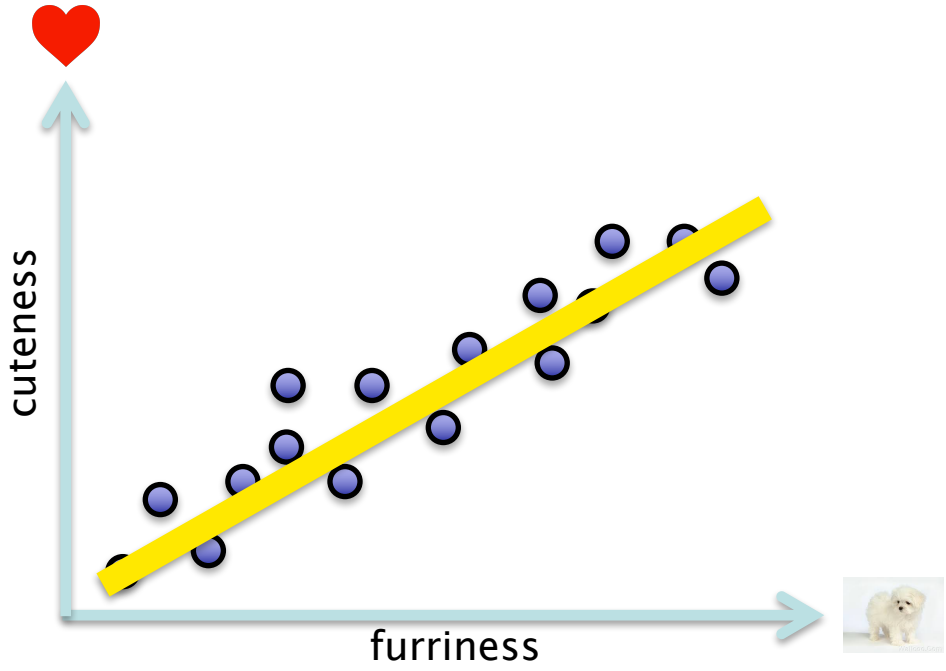


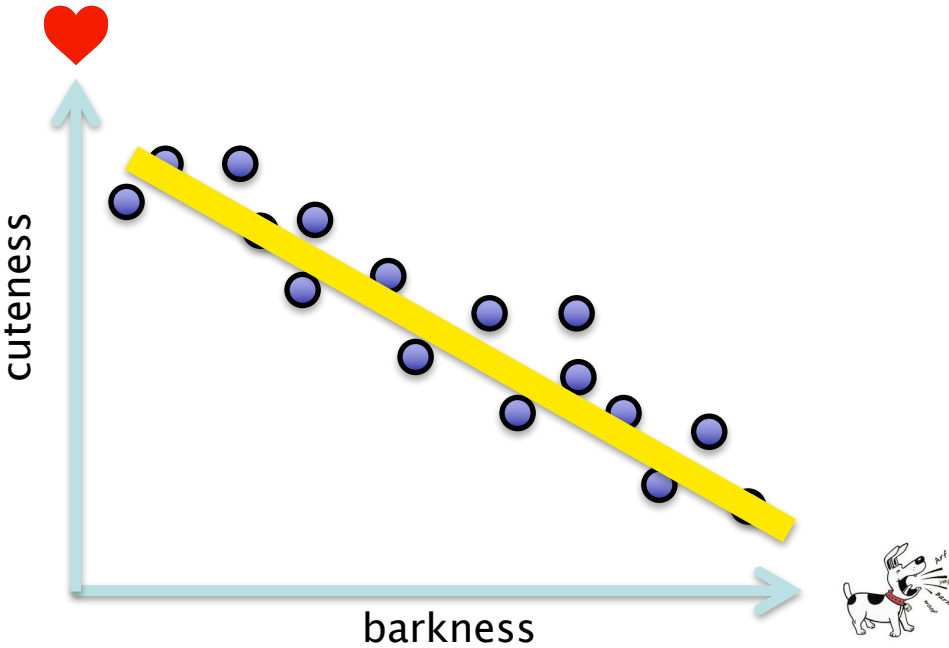


and a second model

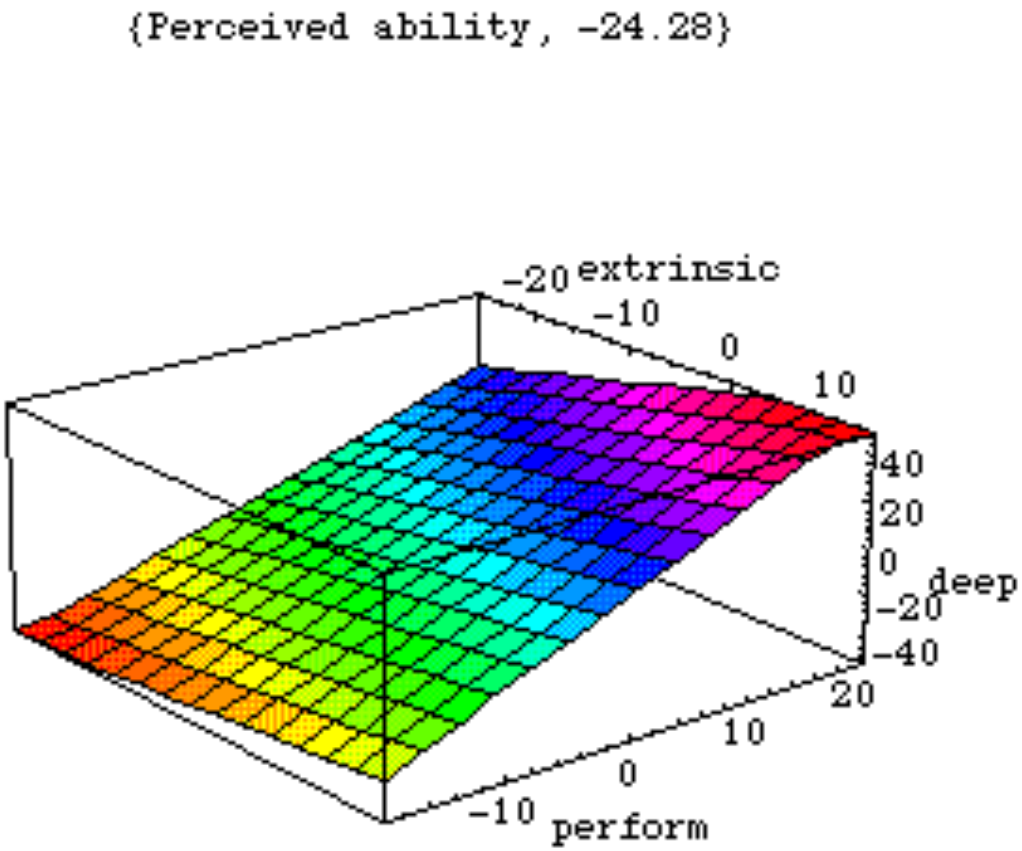


Multivariate regression





Multivariate model (regression surface) can be automatically determined



Multivariate Regression Equation

$$\text{CUTENESS} = \mathbf{a}\,\text{FURRINESS} + \mathbf{b}\,\text{BARKNESS} + \mathbf{c}$$

a is the slope of the FURRINESS line —> positive value

b is the slope of the BARKNESS line —> negative value

c is a constant

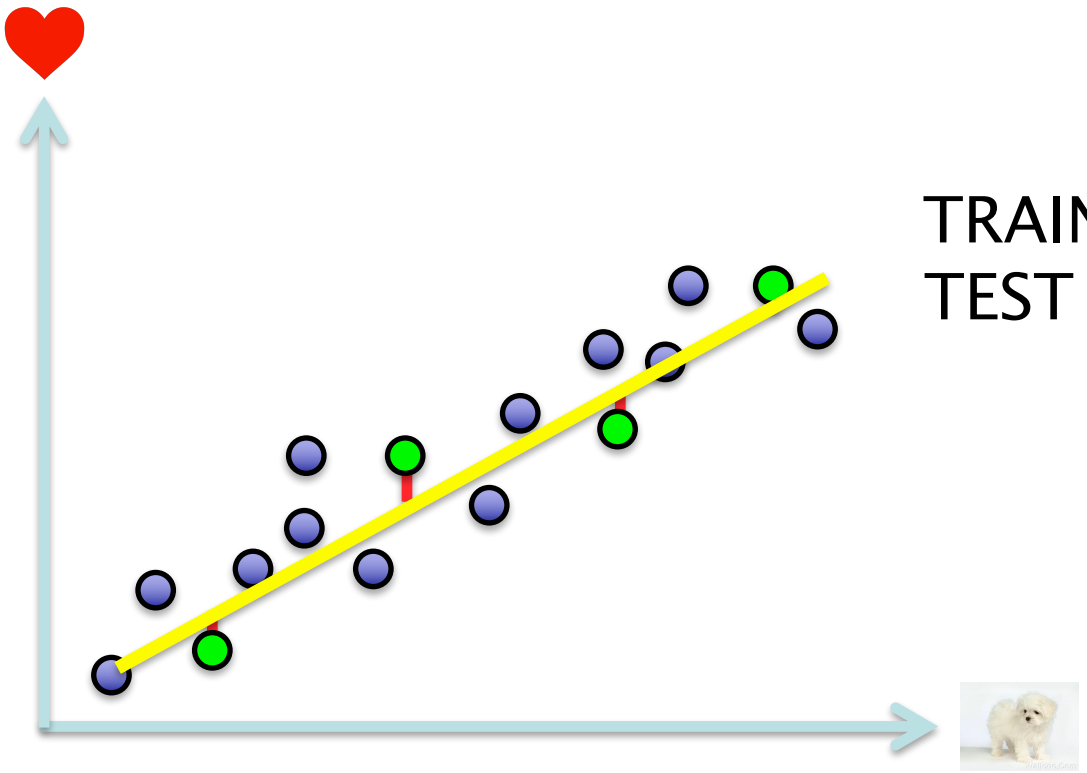
How to measure the prediction power?

- Linear regression is still model fitting, i.e., description rather than prediction

Prediction with Linear Regression

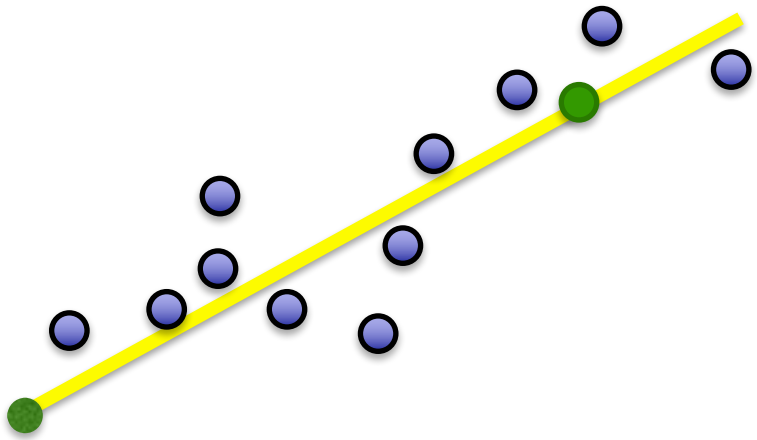
Fit the model to a subset of the data (training)

Test the predictions on the remaining data

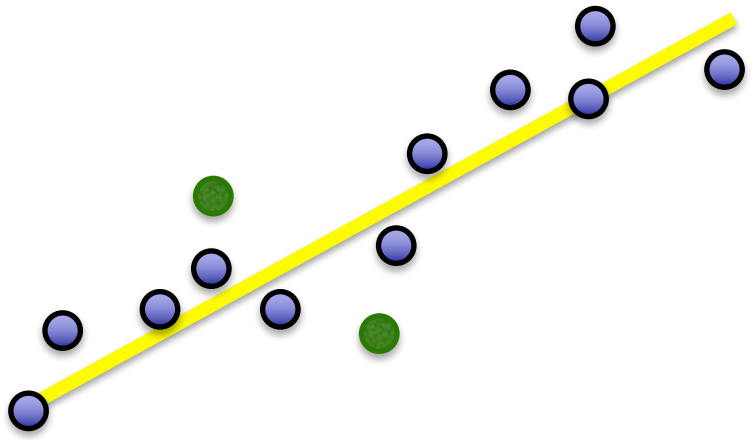


Training/Test Set: Selection Bias

- The selection of the training and test sets can affect the test results
  - positively —> test set contains “easy” instances
  - negatively —> test set contains “hard” instances



“easy” test instances



“hard” test instances

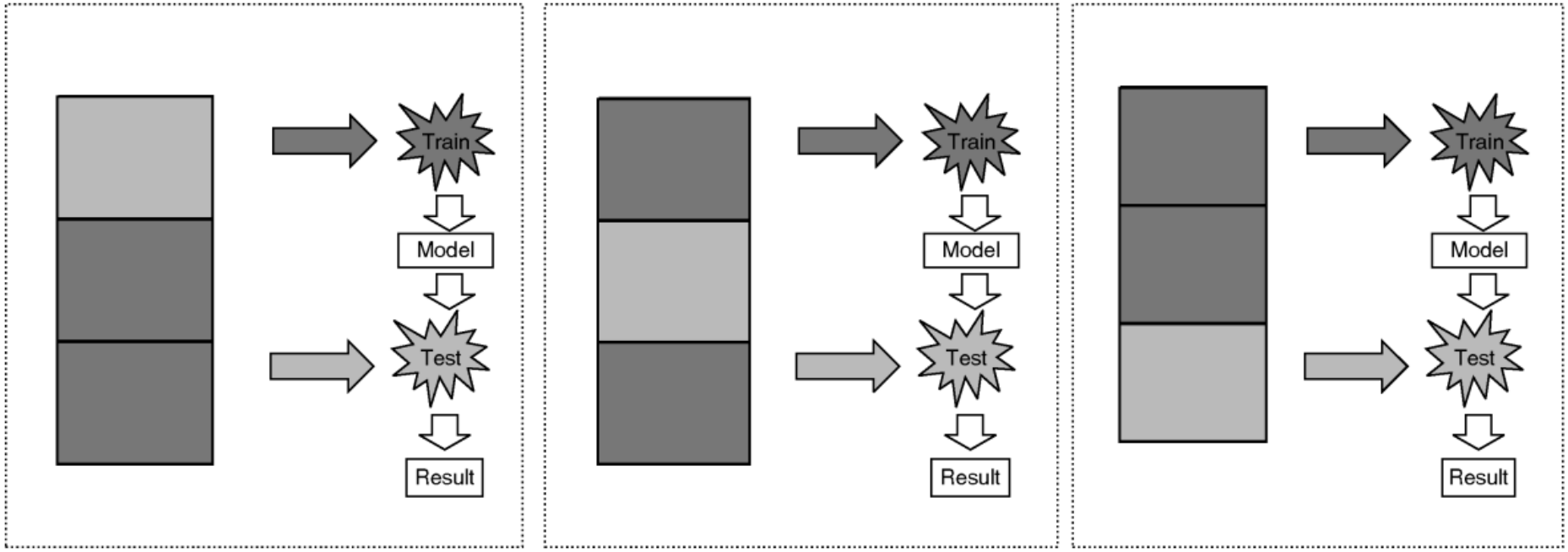
Cross Validation Evaluation Procedure

- Procedure to remove the training/test set selection bias

General idea:

- Perform multiple classification or regression experiments with the dataset.
- In each experiment, select a different partitioning of the dataset into training and test sets.
- Average over the prediction accuracies obtained in the experiments.

3-fold Cross Validation



Cross-Validation. Figure 1. Procedure of three-fold cross-validation.

k-fold Cross Validation

- Often k is defined as a suitable (large) number

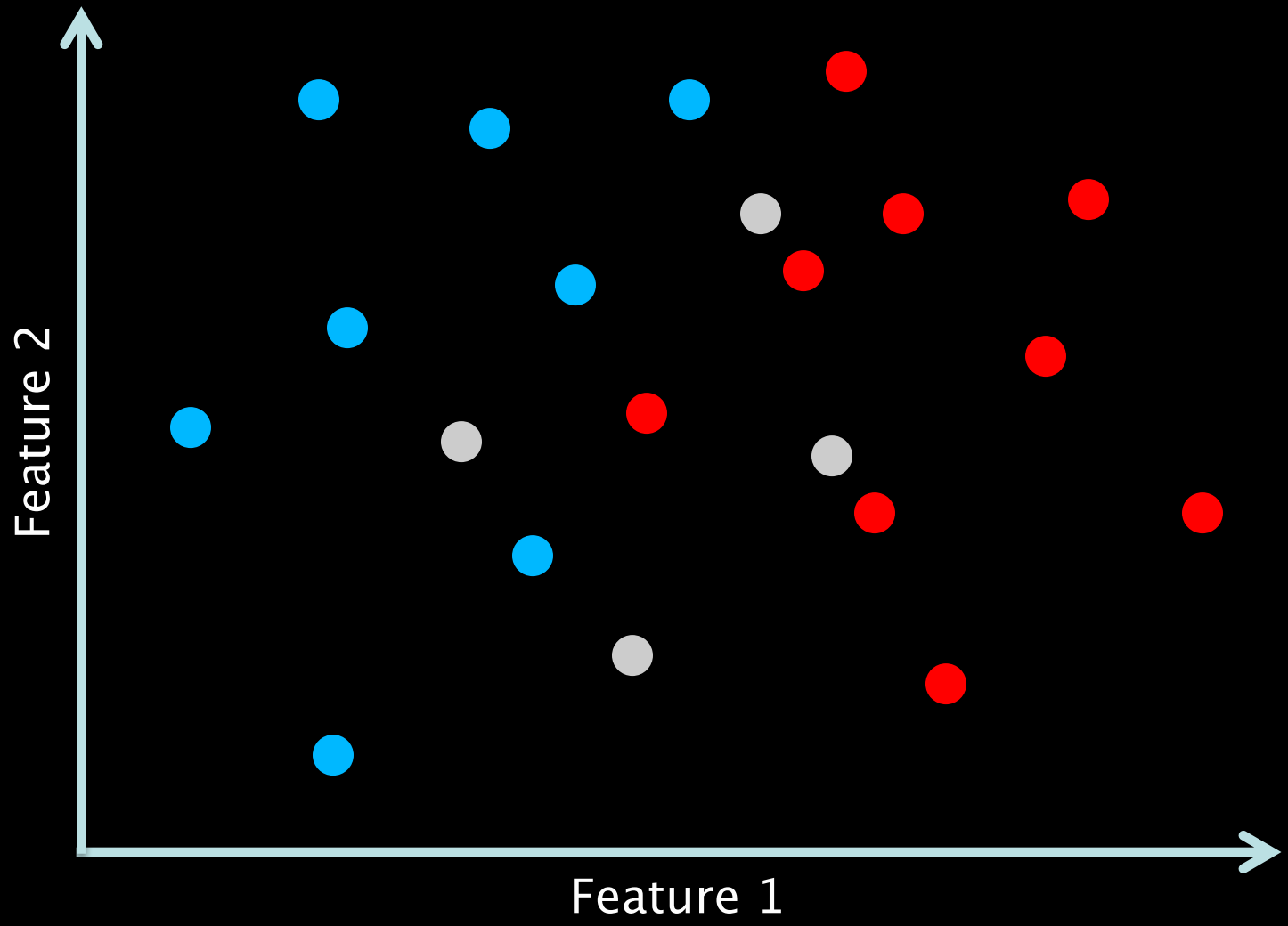
- Examples
- Dataset of 1000 instances
  —> k=10, because 10-fold cross validation yields training sets of 900 instances and test sets of 100 instances.

- Dataset of 777 instances
  —> k=7, resulting in 666/111 partitioning

- Leaving-one-out cross (LOO) validation:
  k = number of instances

Nearest Neighbour Classification

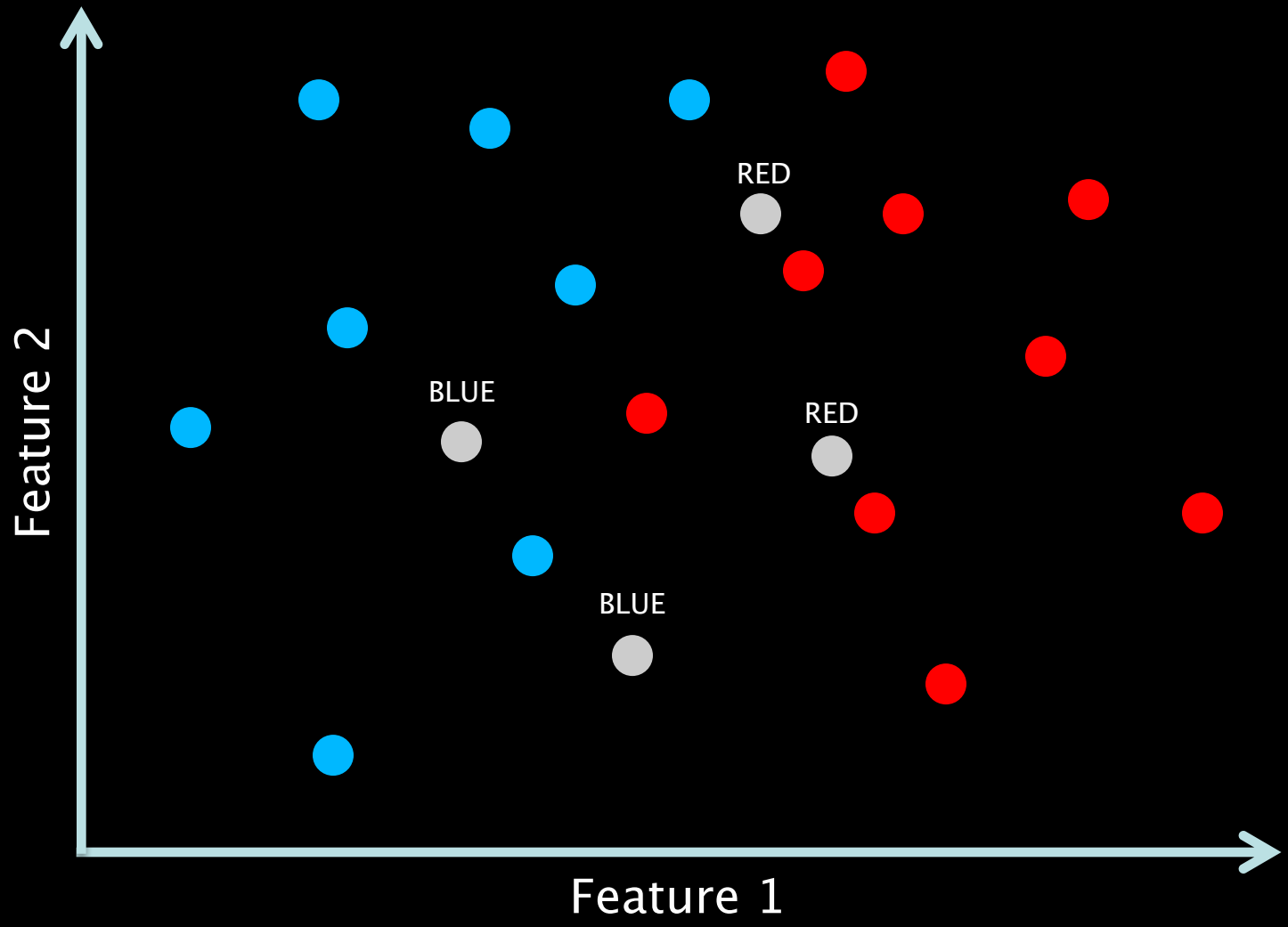
Two Classes (blue & red)



Nearest-neighbour classifier

- Given a set of labeled instances (training set), new instances (test set) are classified according to their nearest labeled neighbour

Nearest Neighbour Classification



k-NN

- In the k-NN classifier, the parameter k represents the number of labeled neighbours considered

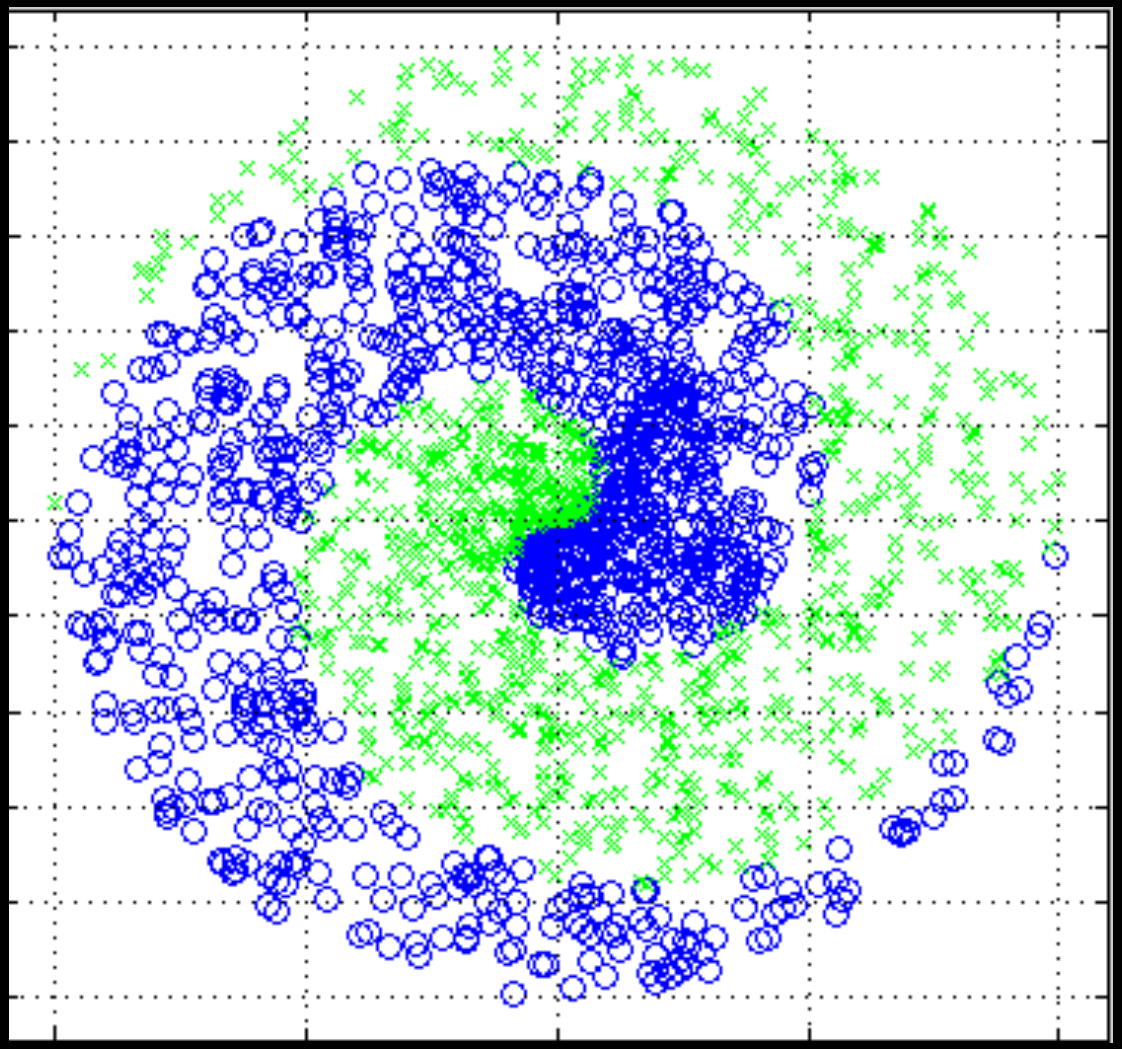
  k = 3: test examples are assigned the labels of the (majority of the) 3 nearest neighbours

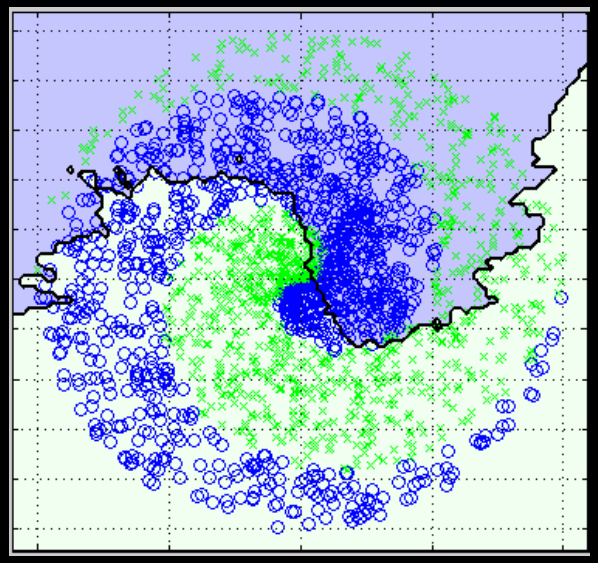
  k = N: test examples are assigned the labels of the (majority of the) 3 nearest neighbours

  For even N in case of an equal number of nearest neighboring labels of two classes: flip a coin

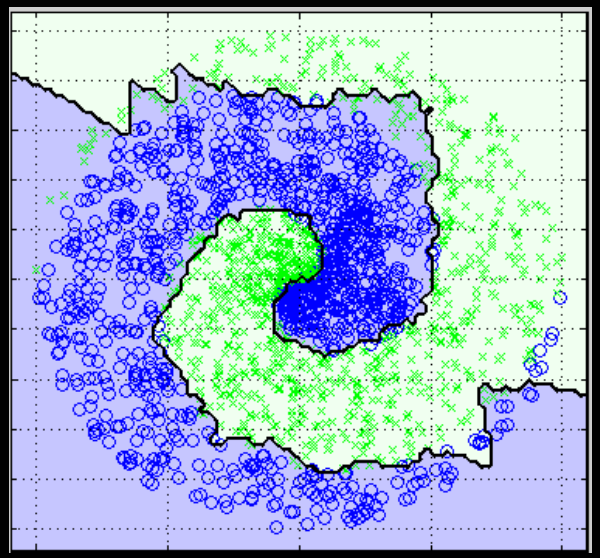
k determines the "model complexity"

- The model in k-NN is the decision boundary that separates the classes (In regression, the model is the line that fits the data)
- Smaller k leads to more complex decision boundaries

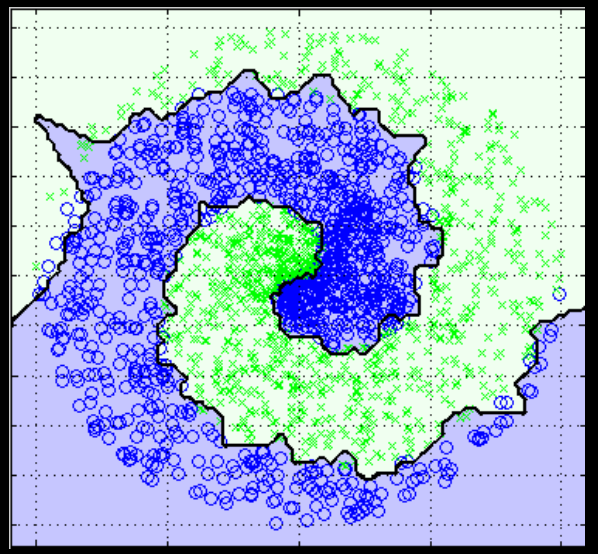




k = 100



k = 10



k = 1

Increasing model complexity

How to determine model complexity?

- Depends on complexity of the separation between the classes
- Start with the simplest model (large k in kNN), and increase complexity (smaller k)

Polynomial Regression

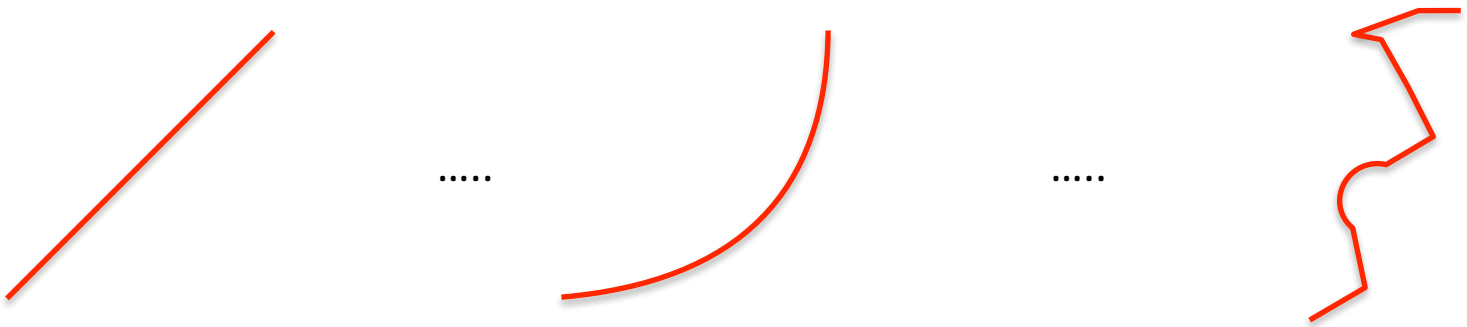
$$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0,$$

The building blocks are formed by the xn (the larger n, the more complex the shape). The values of the a's determine the presence of the building blocks.

Polynomial Regression

- Increasing the complexity (flexibility) of the regression equation, leads to more complex models

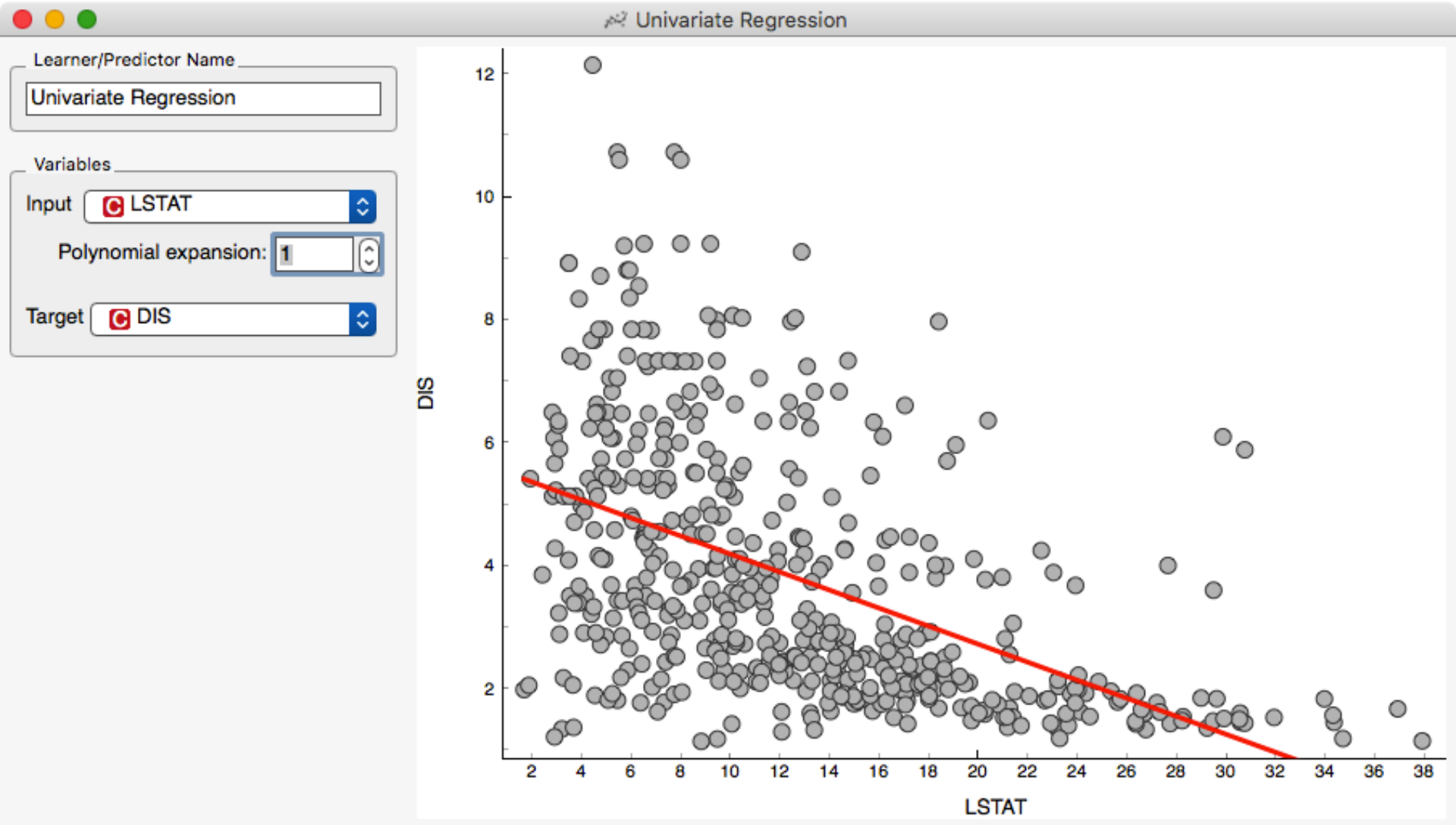
$$\text{CUTENESS} = \mathbf{a}\,\text{FURRINESS} + \mathbf{b}\,\text{BARKNESS}^2 + \mathbf{c}$$



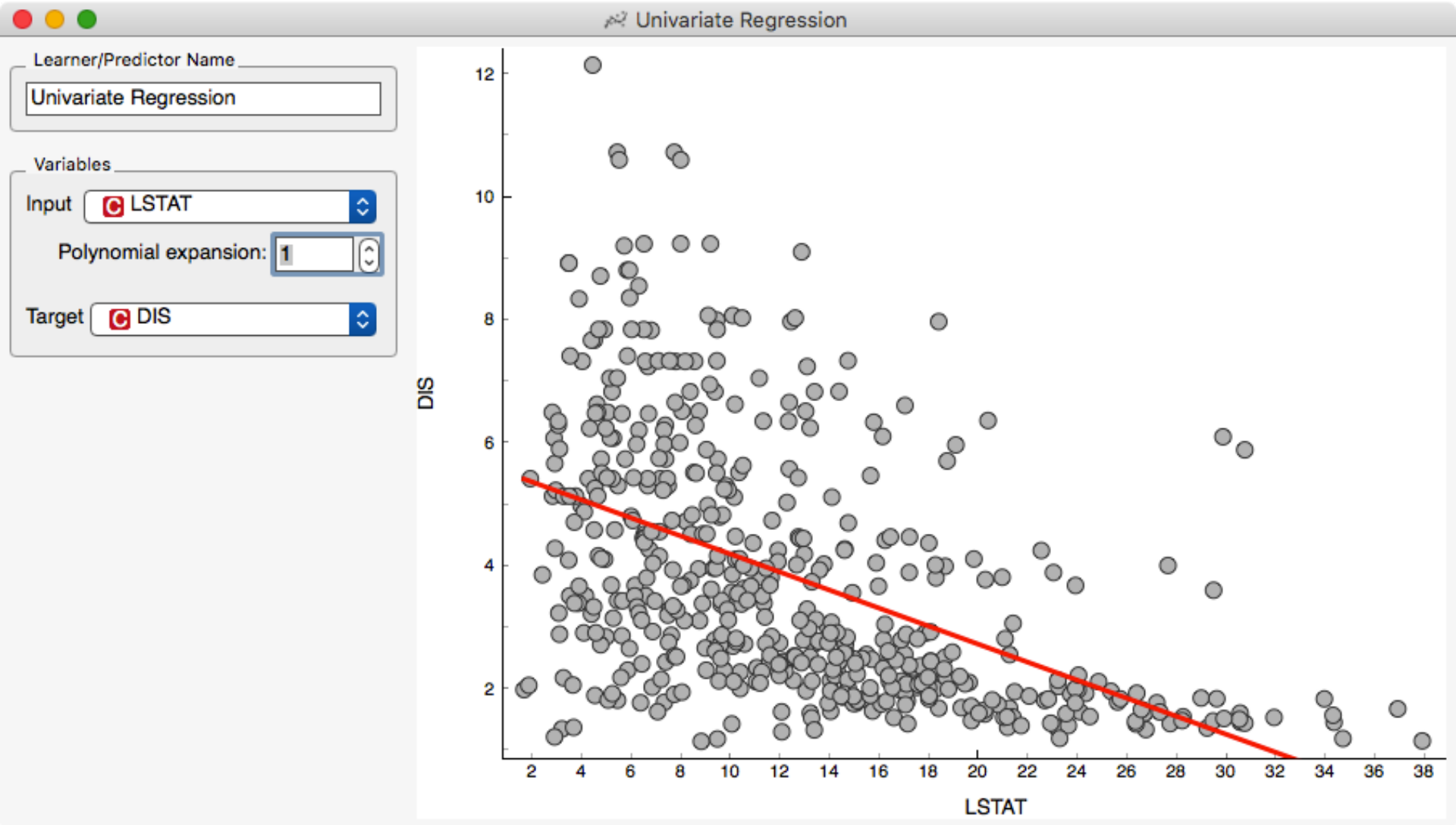
simple model complexity complex

Tilburg University

Polynomial Expansion (complexity parameter)



Lowest Complexity (p.e.=1): Underfitting



Bit more complexity (p.e.=2): Best fit

Univariate Regression

Learner/Predictor Name

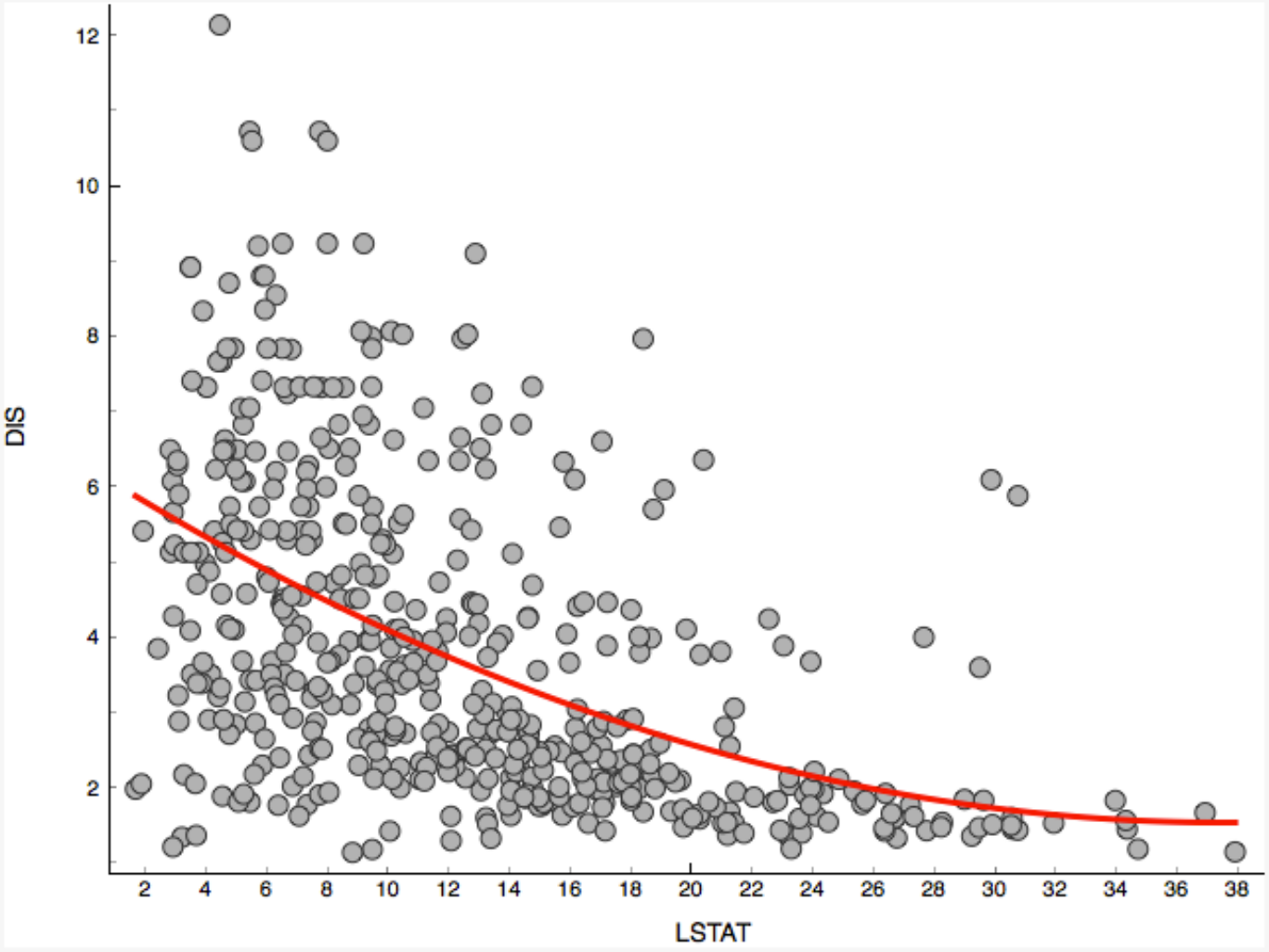
Univariate Regression

Variables

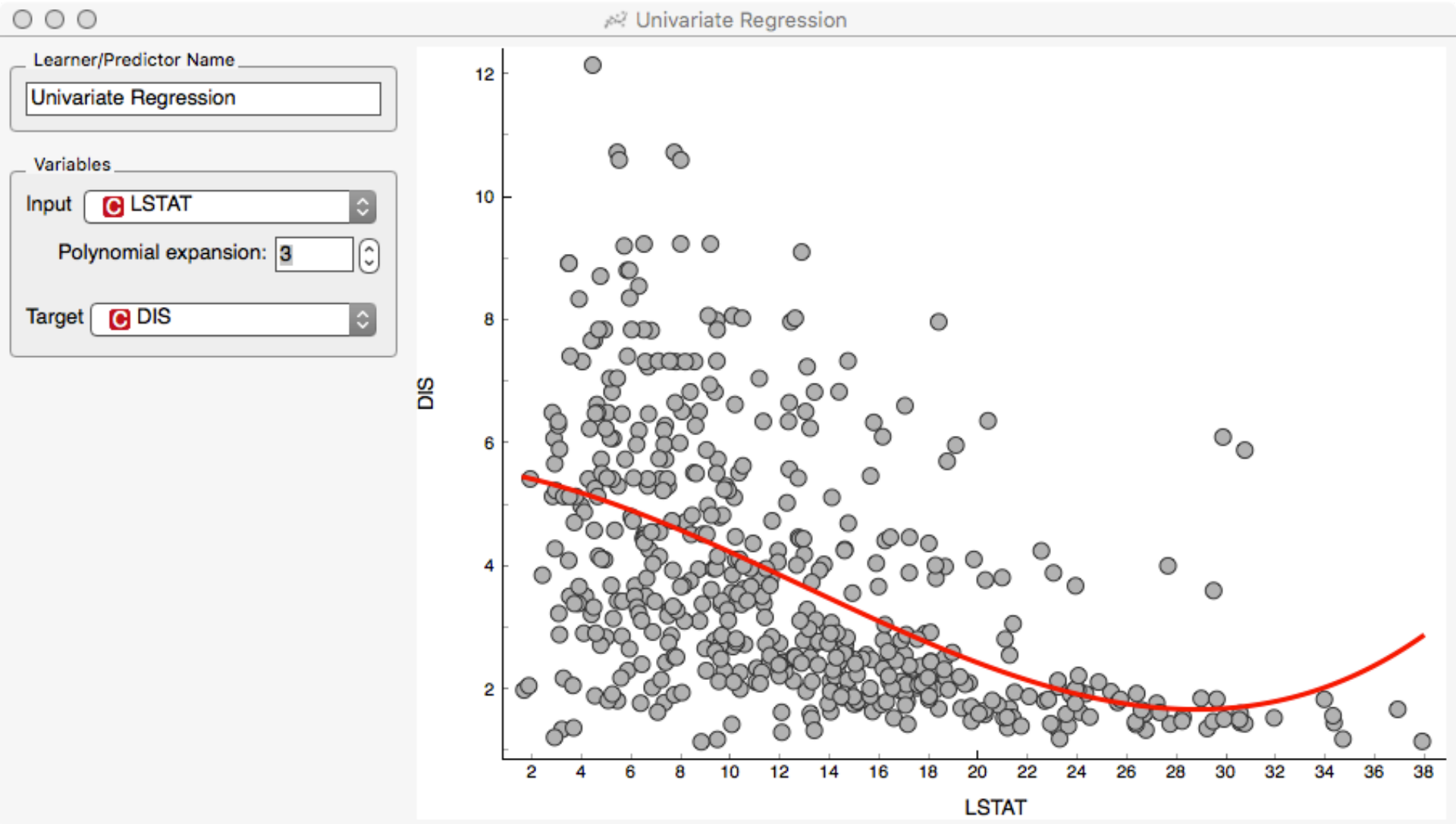
Input: LSTAT

Polynomial expansion: 2

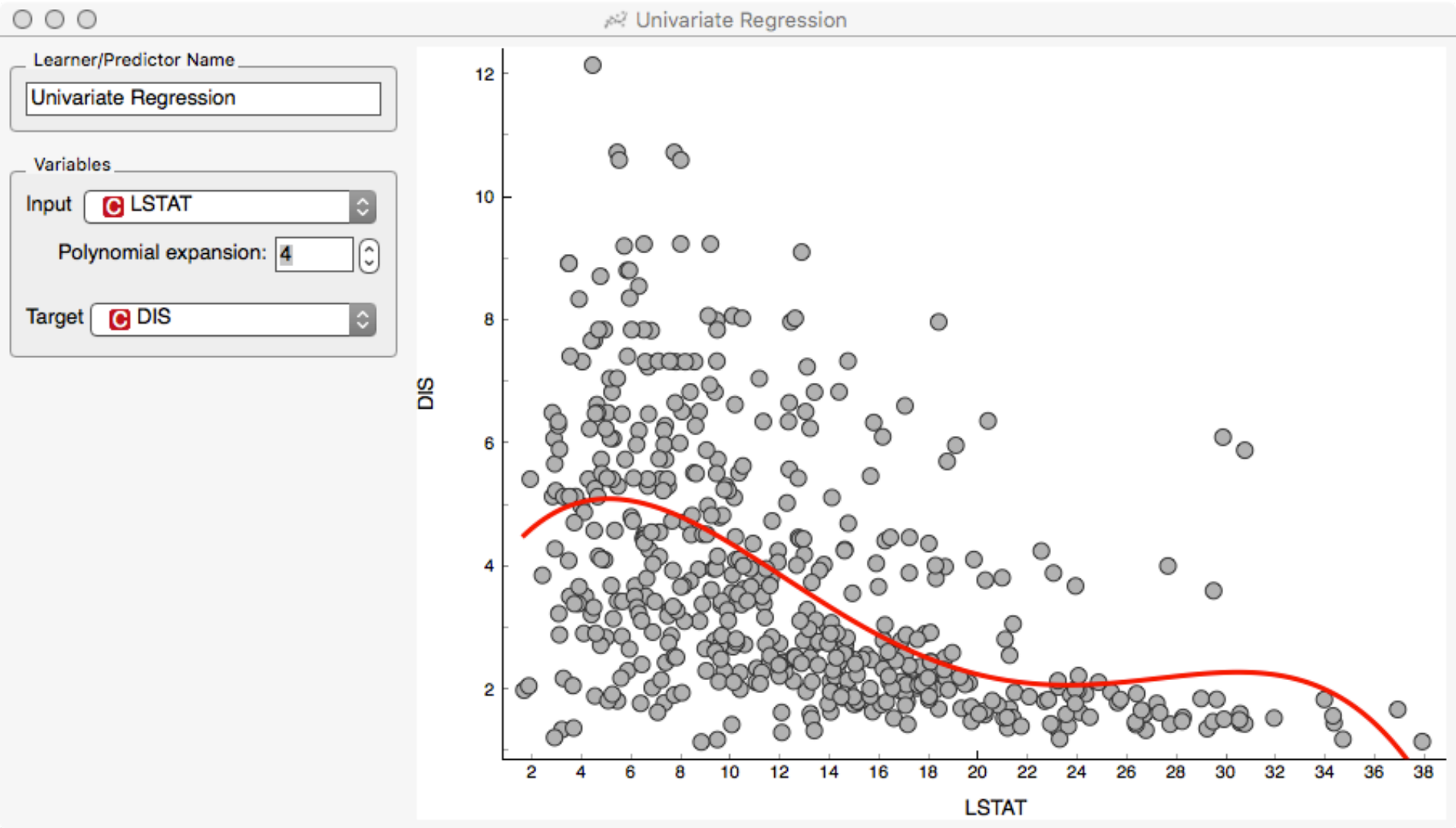
Target: DIS



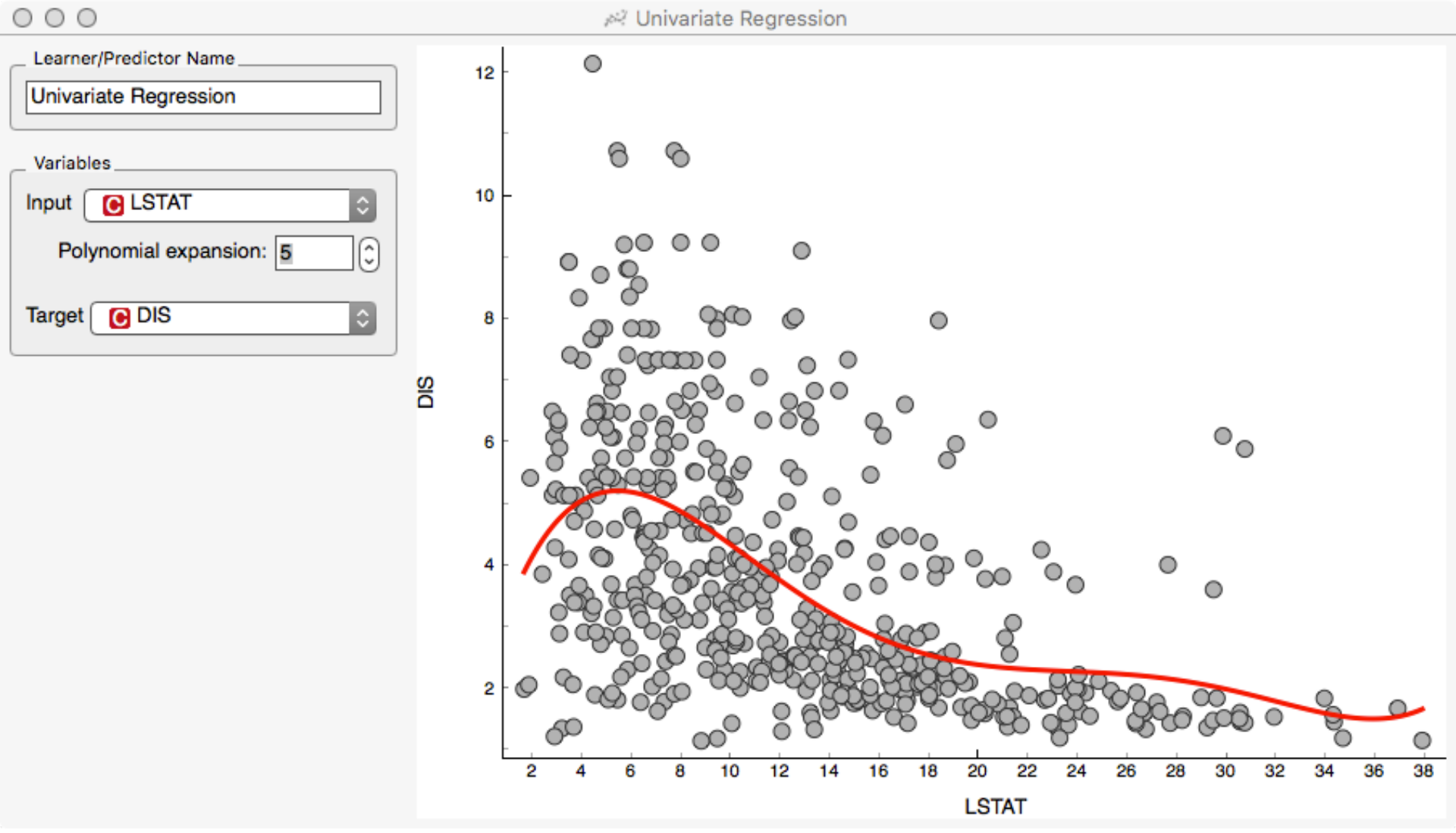
Even more complexity (p.e.=3): Overfitting



More complexity (p.e.=4): Overfitting



Large complexity (p.e.=5): Overfitting



Underfitting and Overfitting

- The relation between the complexity of the induced model and underfitting and overfitting is a crucial notion in machine learning

- **Underfitting:**
  - The induced model is not complex (flexible) enough to model the data
  - performs badly both on training and test set

- **Overfitting**
  - The induced model is too complex to model the data (tries to fit noise)
  - performs better on training set than on test set

Orange data fitting (no prediction!)

- Load the housing.tab dataset
- Univariate Regression widget

- Select two different features/attributes
- Vary the polynomial expansion (0 is a straight line) and try to identify underfitting, fitting, and overfitting by means of visual inspection

Housing.tab

- Concerns housing values in suburbs of Boston.

- Attribute Information:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

Orange prediction
Nearest Neighbour regression