香港城市大學
City University of Hong Kong

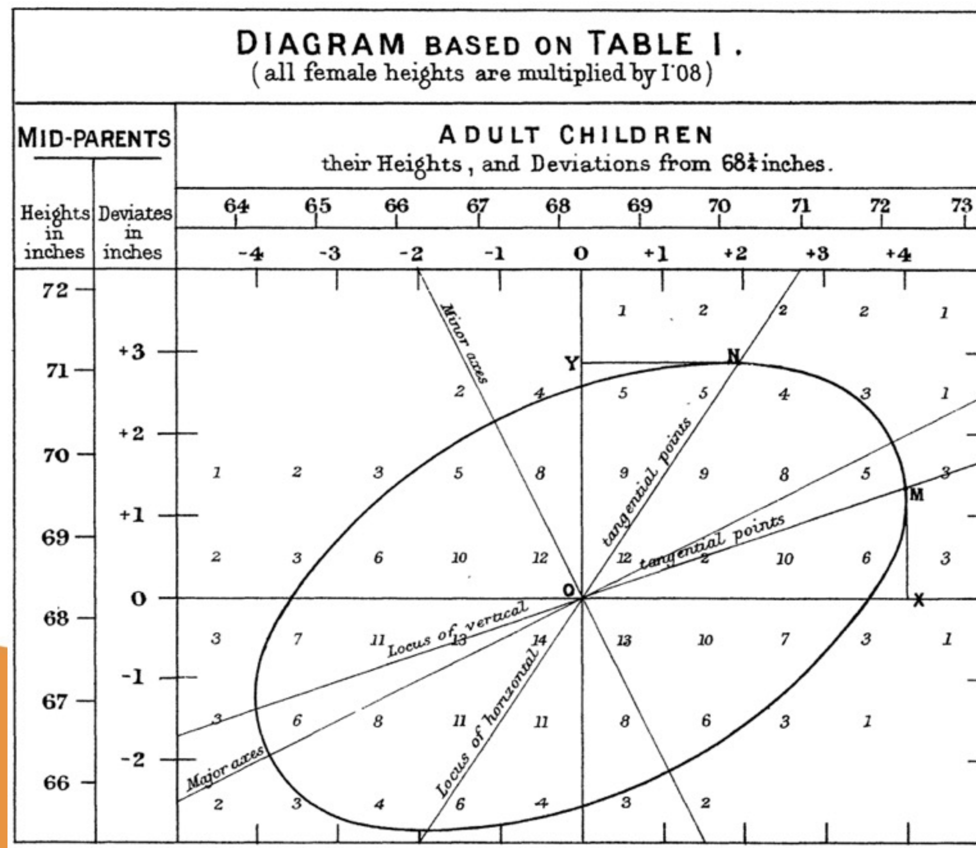
IS6400: Business Data Analytics

Regression for Prediction

專業 創新 胸懷全球
Professional · Creative
For The World

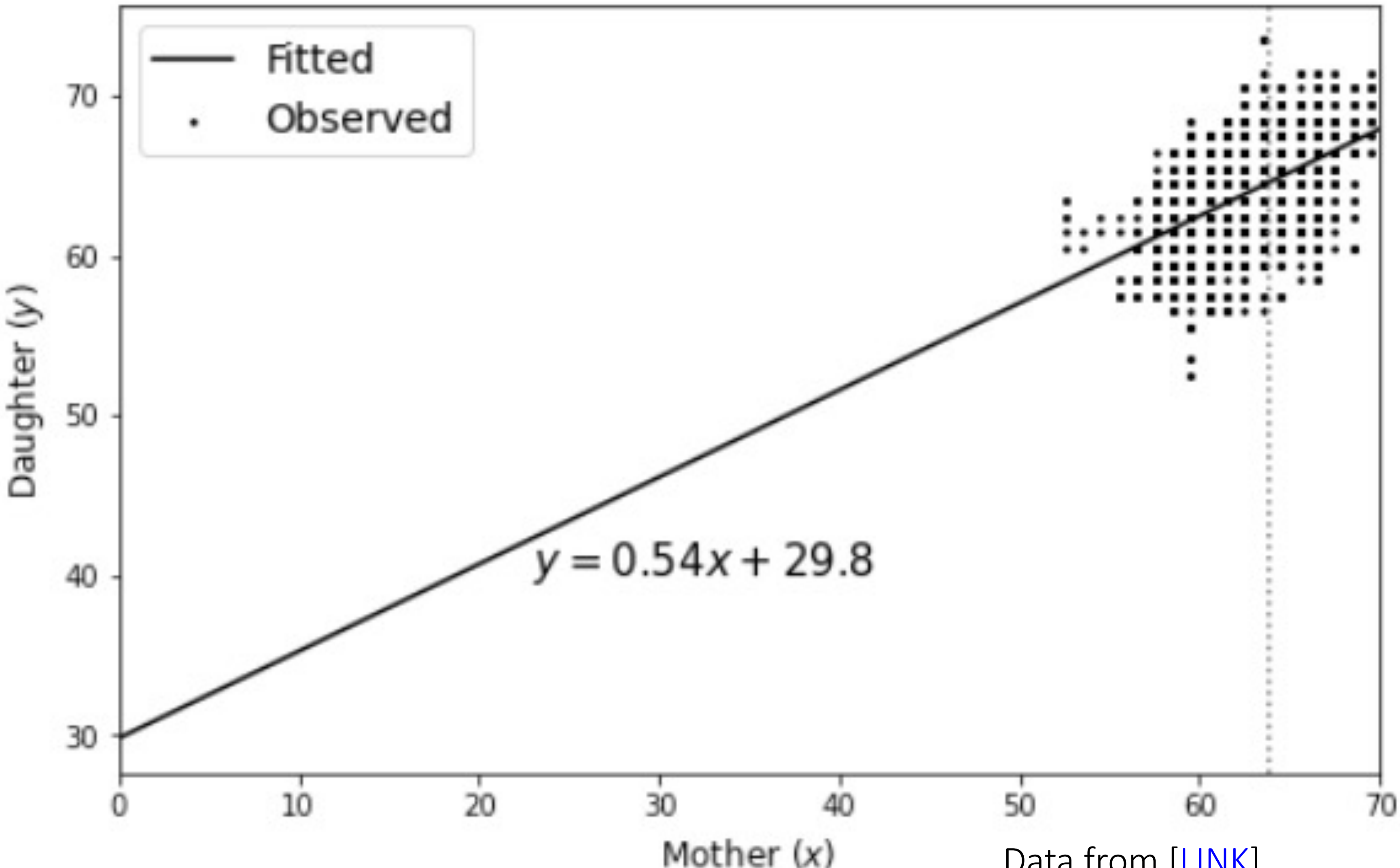
Regression: The Origin

The experiments showed further that the mean filial regression towards mediocrity was directly proportional to the parental deviation from it. This curious result was based on so many plantings,



Galton, F. (1886). Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

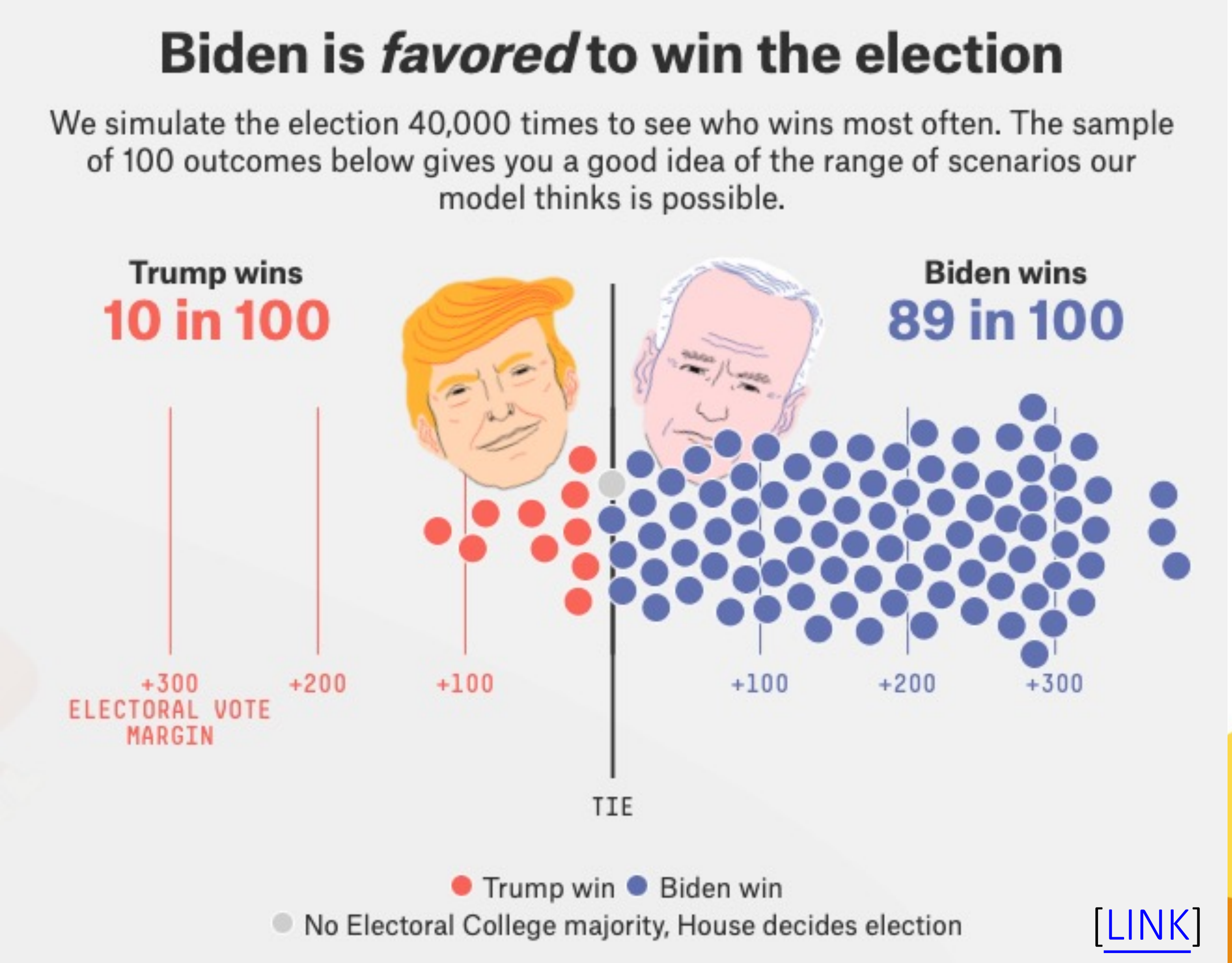
Reproducing the regression



Data from [LINK]

Regression Use Cases

- Prediction (Our focus)



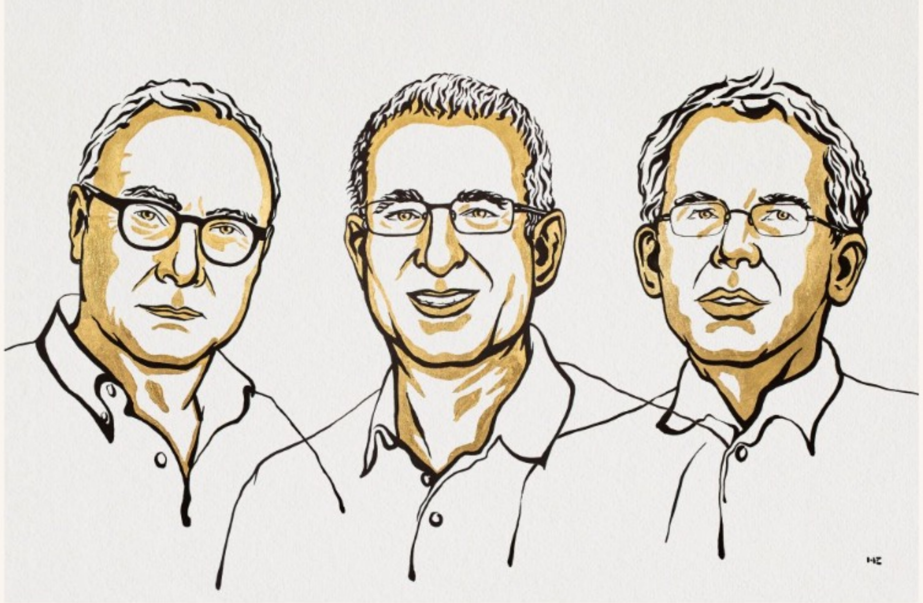
[LINK]

Regression Use Cases

- Explanation!
  - The credibility revolution

Economic sciences laureates 2021

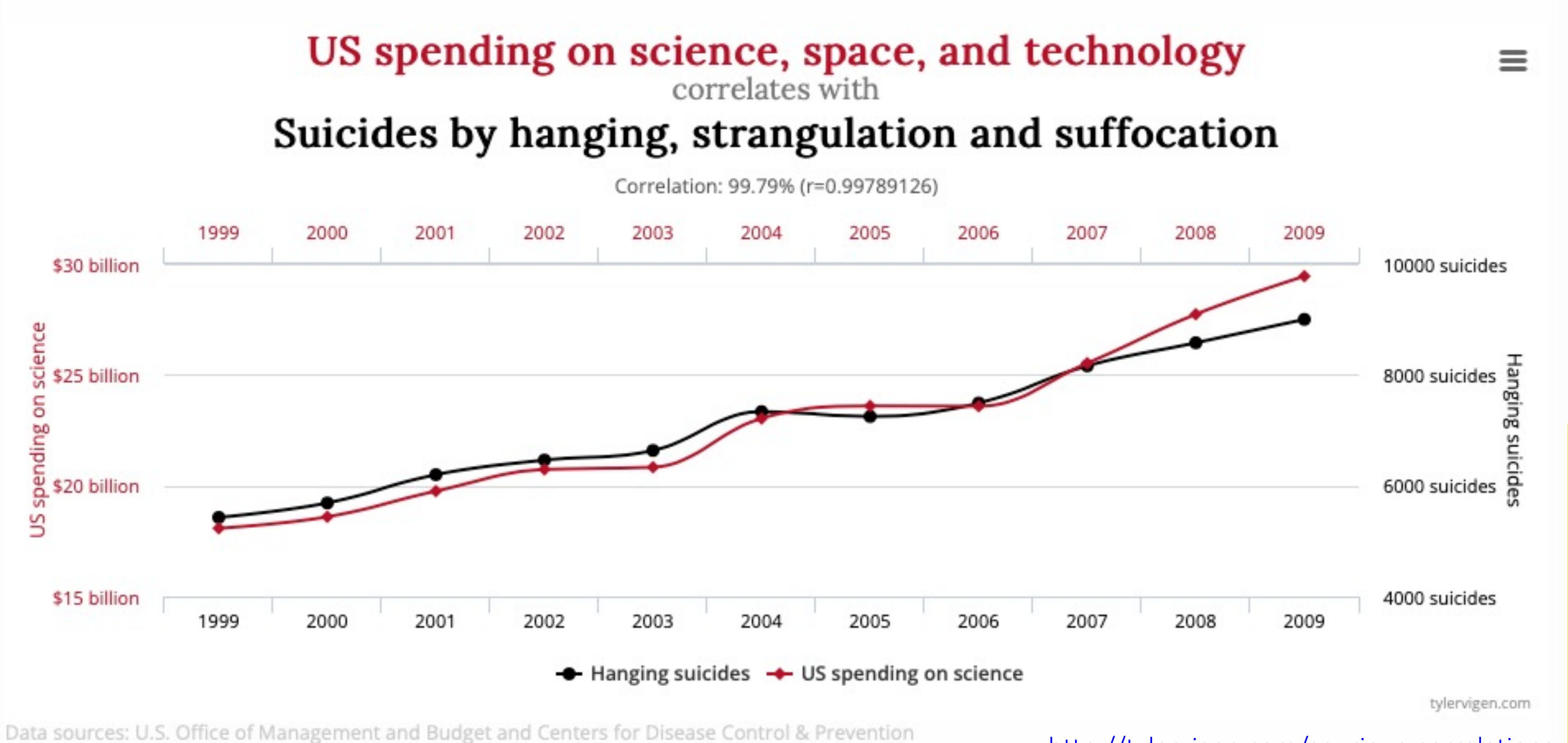
The Royal Swedish Academy of Sciences has decided to award the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2021 with one half to David Card "for his empirical contributions to labour economics" and the other half jointly to Joshua Angrist and Guido Imbens "for their methodological contributions to the analysis of causal relationships"



Ill. Niklas Elmehed © Nobel Prize Outreach.

Regression Use Cases

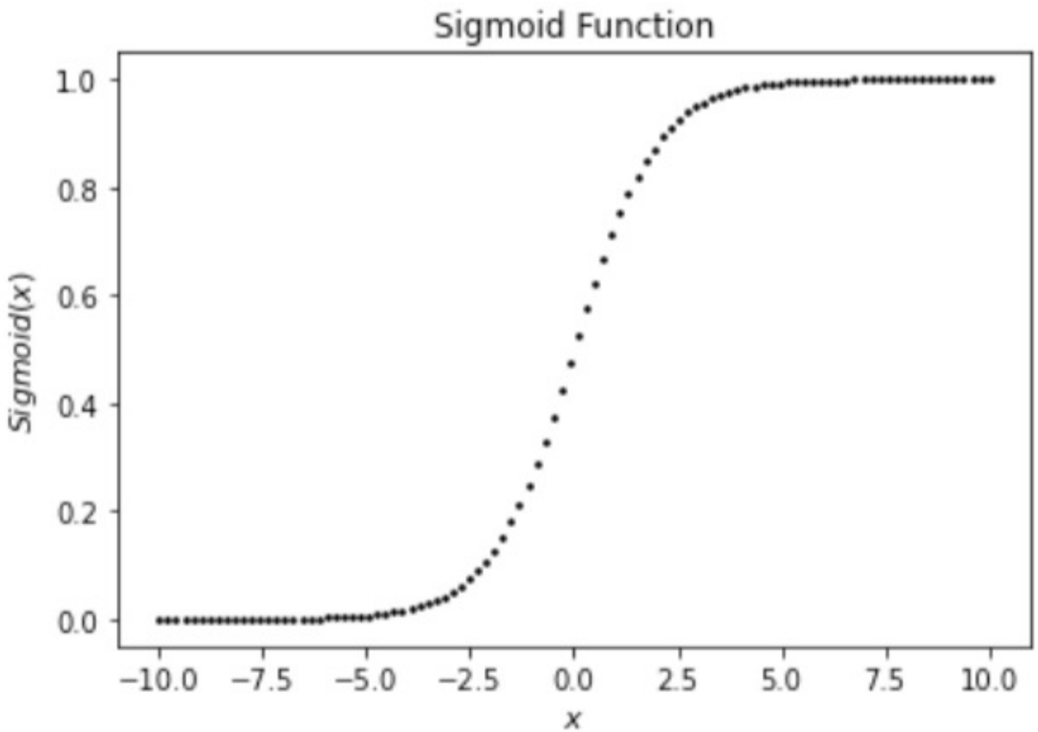
- Explanation!
  - Spurious correlation



http://tylervigen.com/spurious-correlations

Regression vs. Classification?

- Regression for continuous outcome
- Classification for discrete outcome
- Yet…



$$Sigmoid(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \in (0,1)$$

Starting the journey…

- The anatomy of simple linear regression:

Predictor or feature

Continuous outcome

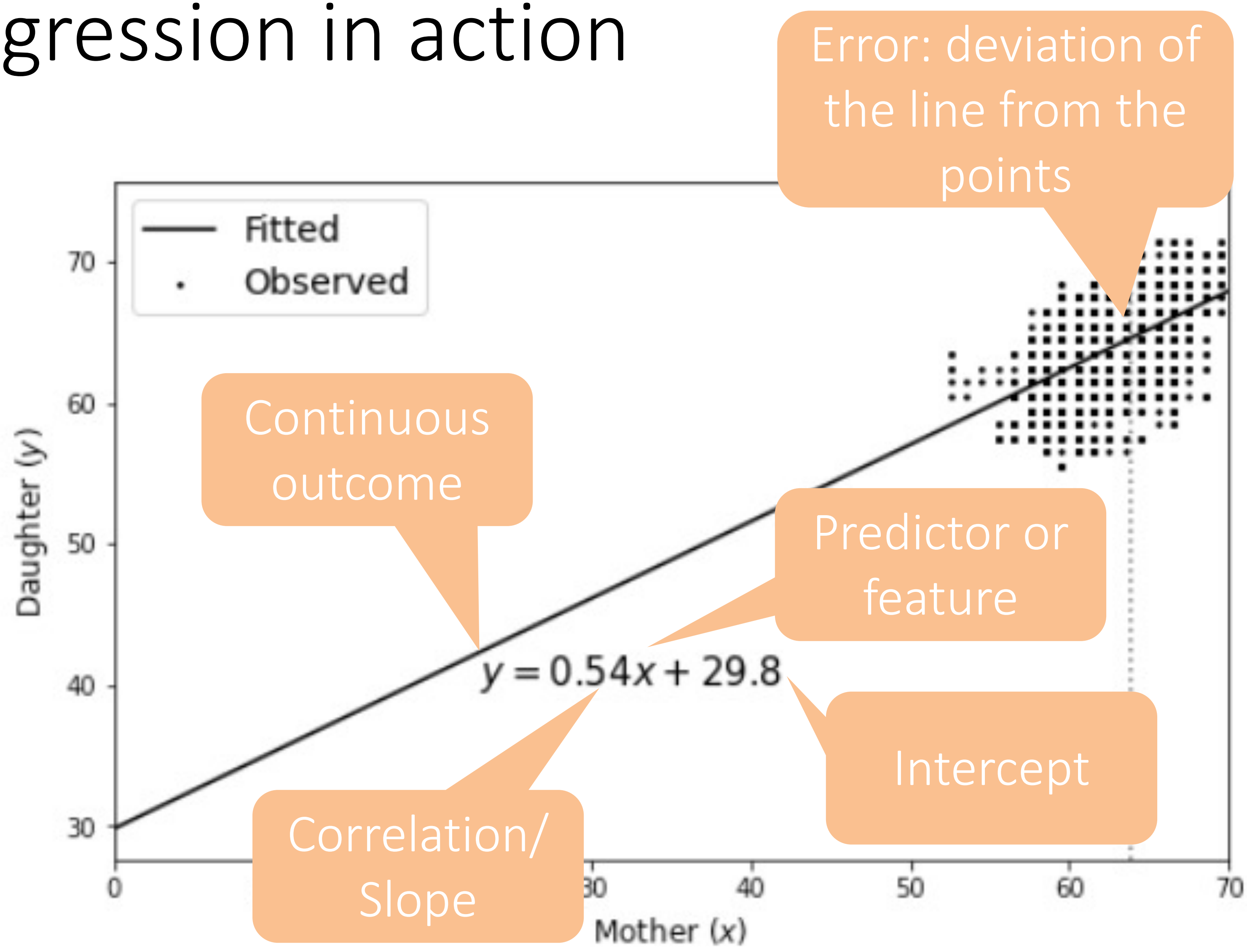
$$y = \beta_1 x + \beta_0 + \epsilon$$

Random noise with zero-mean

Correlation/ Slope

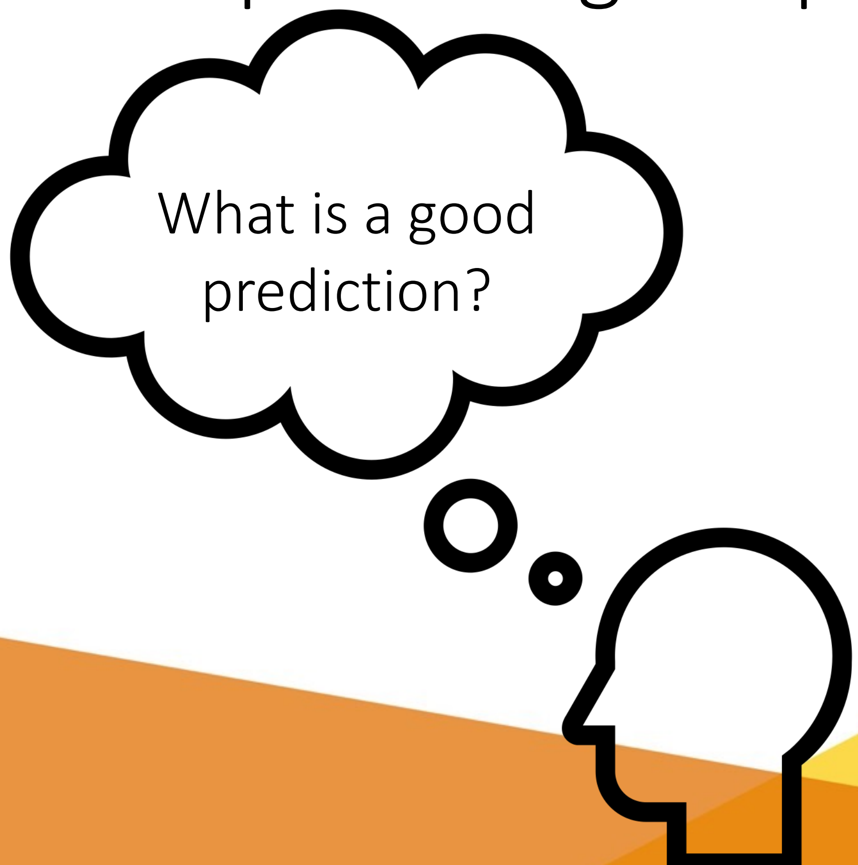
Intercept

Regression in action



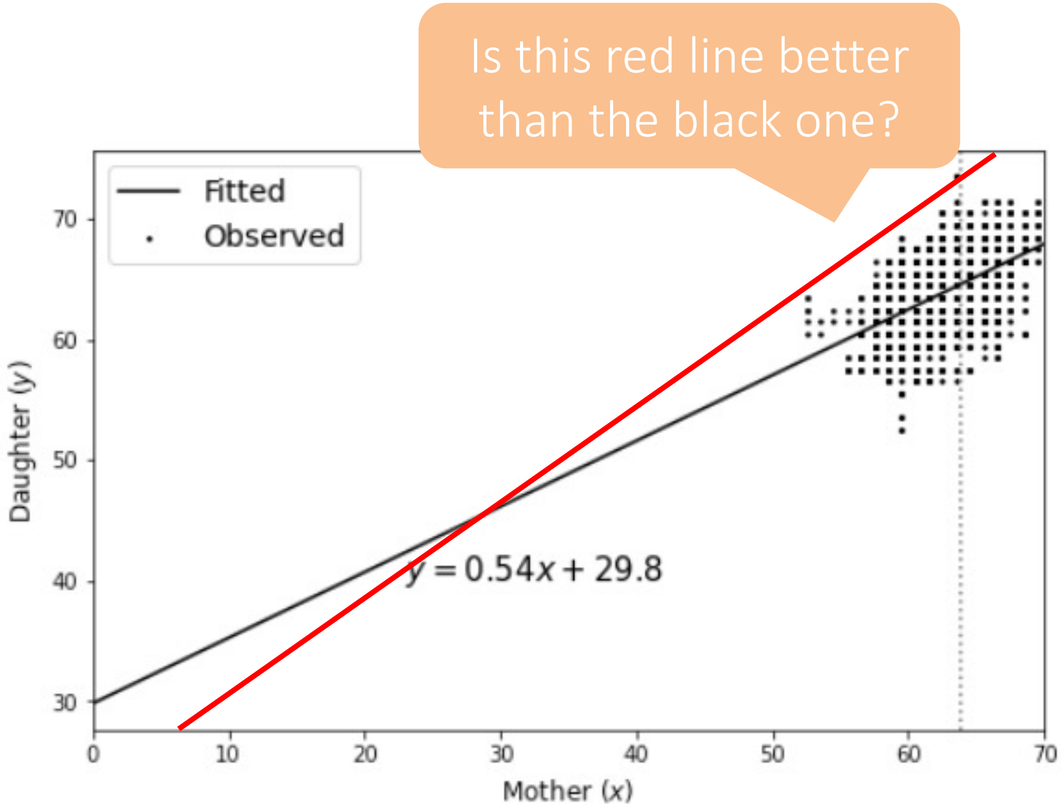
But how do we get there?

- How do we pick proper values of the parameters to produce good predictions?



Measuring prediction performance

- Qualitatively, the closer our predicted values to true values are, the better we perform.



Error measurement

- Intuitively, we can just contrast the predicted and true values to obtain their numerical difference.

$$\epsilon_i = y_{\mathrm{i}} - (\beta_1 x_i + \beta_0)$$

$$\mathrm{Err} = \sum_i \epsilon_i$$

Is this appropriate?

Error measurement: Squared Error

- But errors of different signs may cancel out!

$$\epsilon_i = y_{\mathrm{i}} - (\beta_1 x_i + \beta_0)$$

$$\mathrm{Err} = \sum_i \epsilon_i^2$$

Let's square it then!

Now that we have the error…

- What do we want from a regression model?

$$\mathcal{J}(\beta_1, \beta_0) = \frac{1}{2}\Sigma_i^N {\epsilon_i}^2 = \frac{1}{2}\Sigma_i^N [y_i - (\beta_1 x_i + \beta_0)]^2$$

$$min\mathcal{J}(\beta_1, \beta_0) = min\frac{1}{2}\Sigma_i^N [y_i - (\beta_1 x_i + \beta_0)]^2$$

Least square!

For math convenience only.

How do we infer the parameters?

- How do we get betas' to minimize the error?
- Intuitively, we could search through some space…

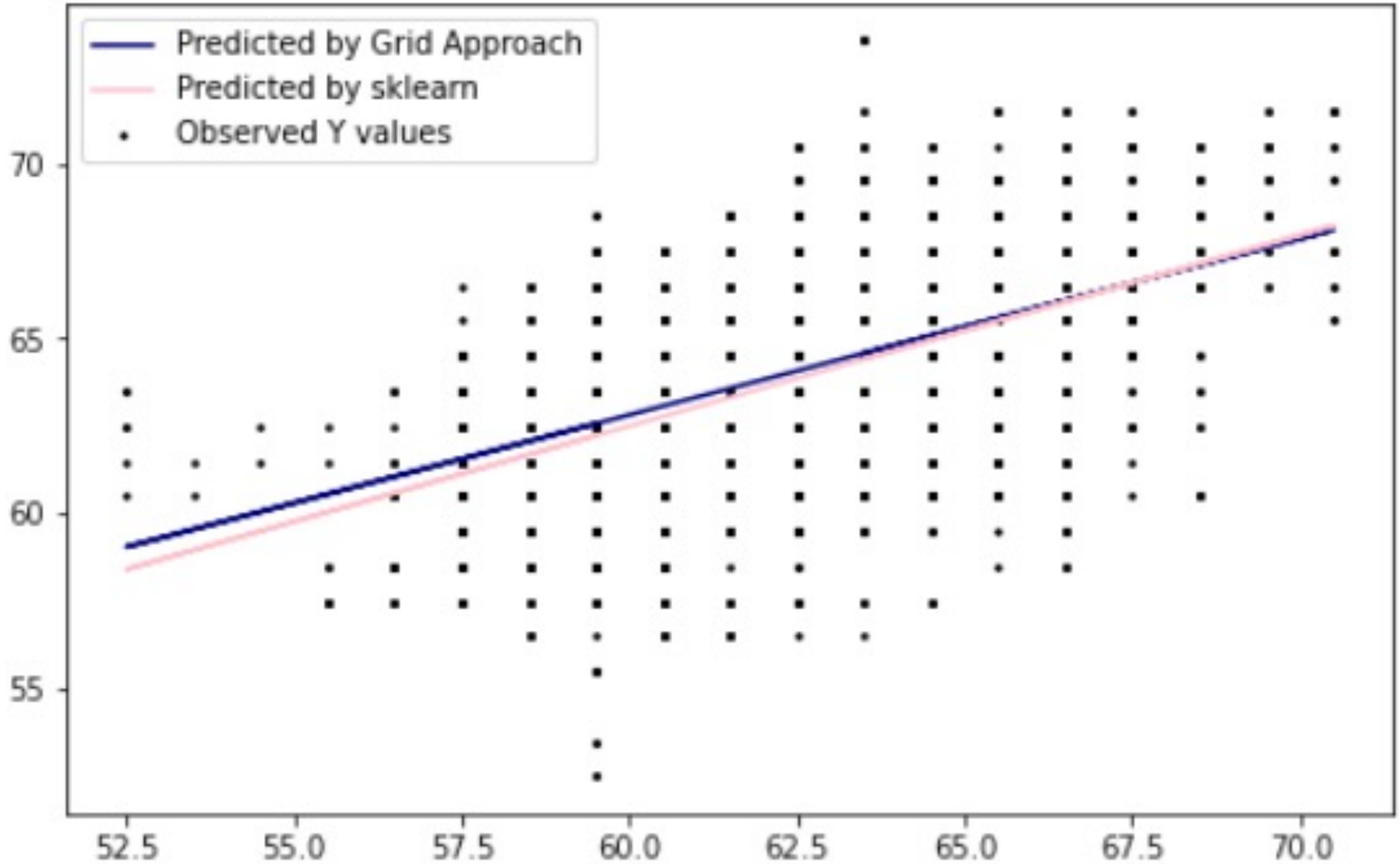
Intuitive Approach: Grid Search

- We make a ”grid” of candidate values

	β_0					
β_1	[100,100]	[100,99]	…	…	…	[100,1]
	…	…				…
			…			
				…		
					…	
	[1, 100]	[1, 99]	…	…	…	[1,1]

Compute MSE on each of these cells. The parameter values with the smallest MSE are considered “good” parameters.

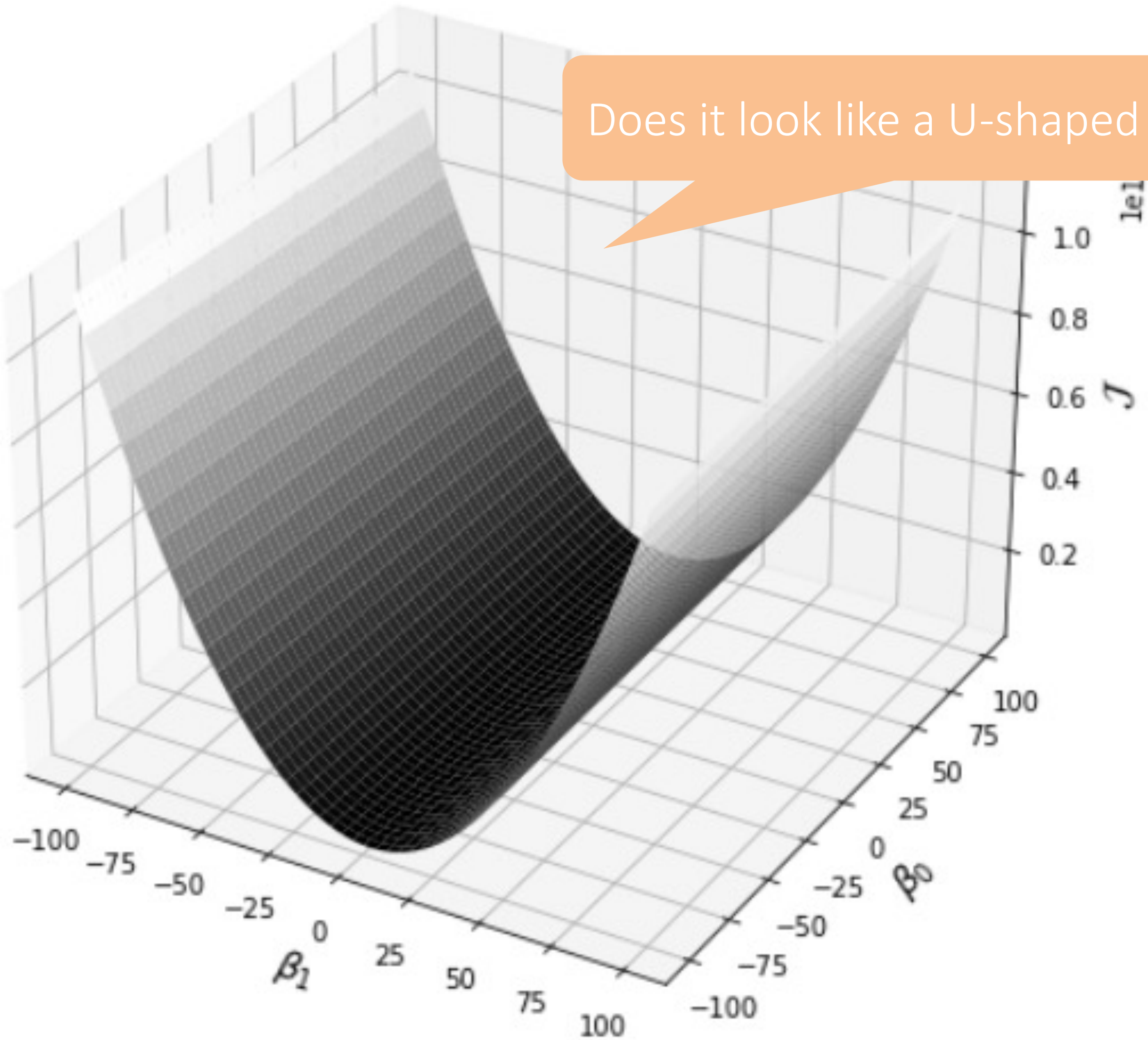
Intuitive Approach: Grid Search



But this is inefficient…

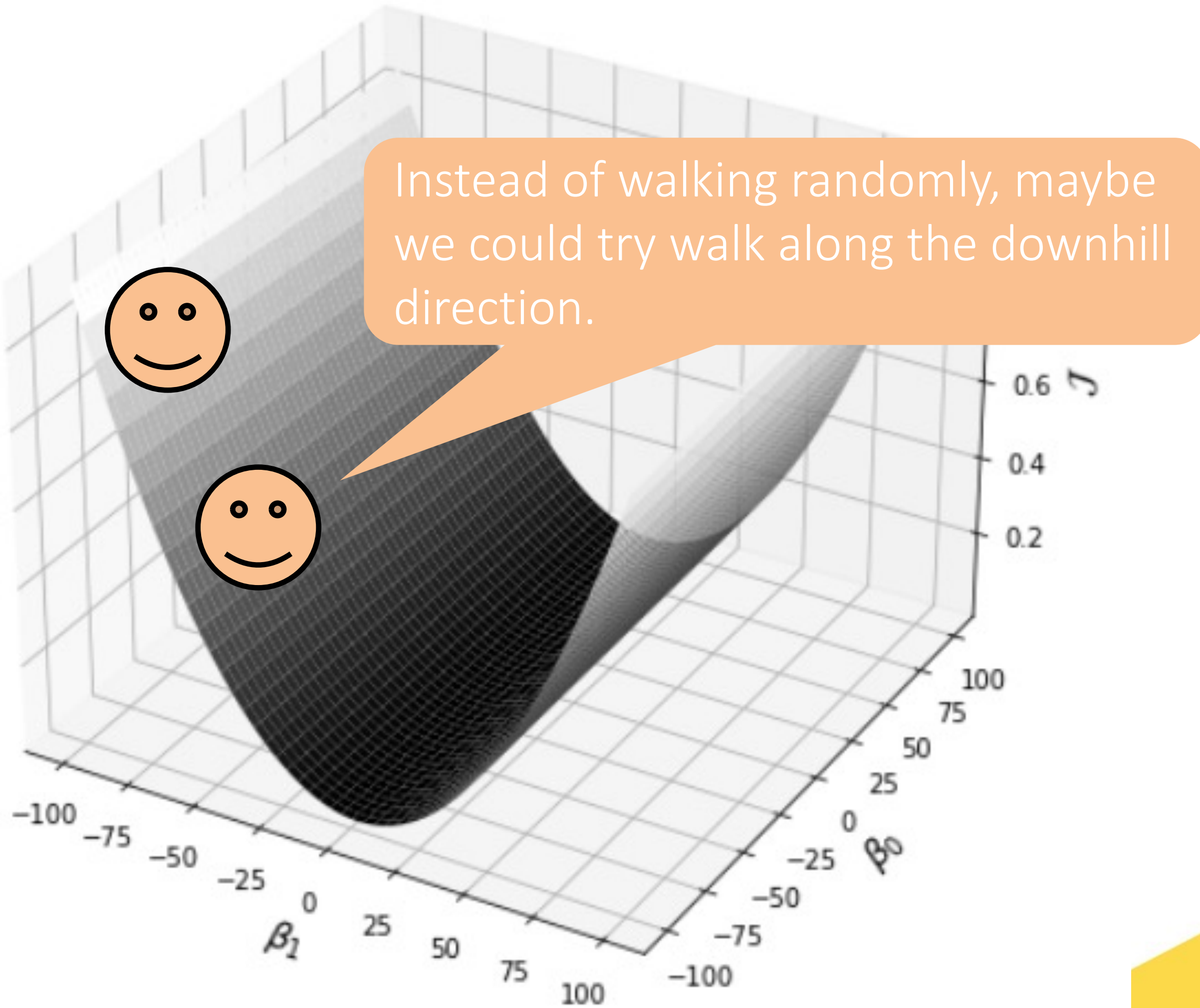
- What if the true parameters happen to be out of the grid?
- What if we have 10 parameters in this model?
  – This will make our search space too high.

What do we learn from grid search?



Does it look like a U-shaped curve?

We can walk faster, can't we?



$$\frac{\partial \mathcal{J}}{\partial \beta_1} = \frac{\Sigma_i^N x_i(\beta_1 x_i + \beta_0 - y_i)}{N} = -\frac{\Sigma_i^N x_i \epsilon_i}{N}$$

$$\frac{\partial \mathcal{J}}{\partial \beta_0} = \frac{\Sigma_i^N (\beta_1 x_i + \beta_0 - y_i)}{N} = -\frac{\Sigma_i^N \epsilon_i}{N}$$

Walking down the hill

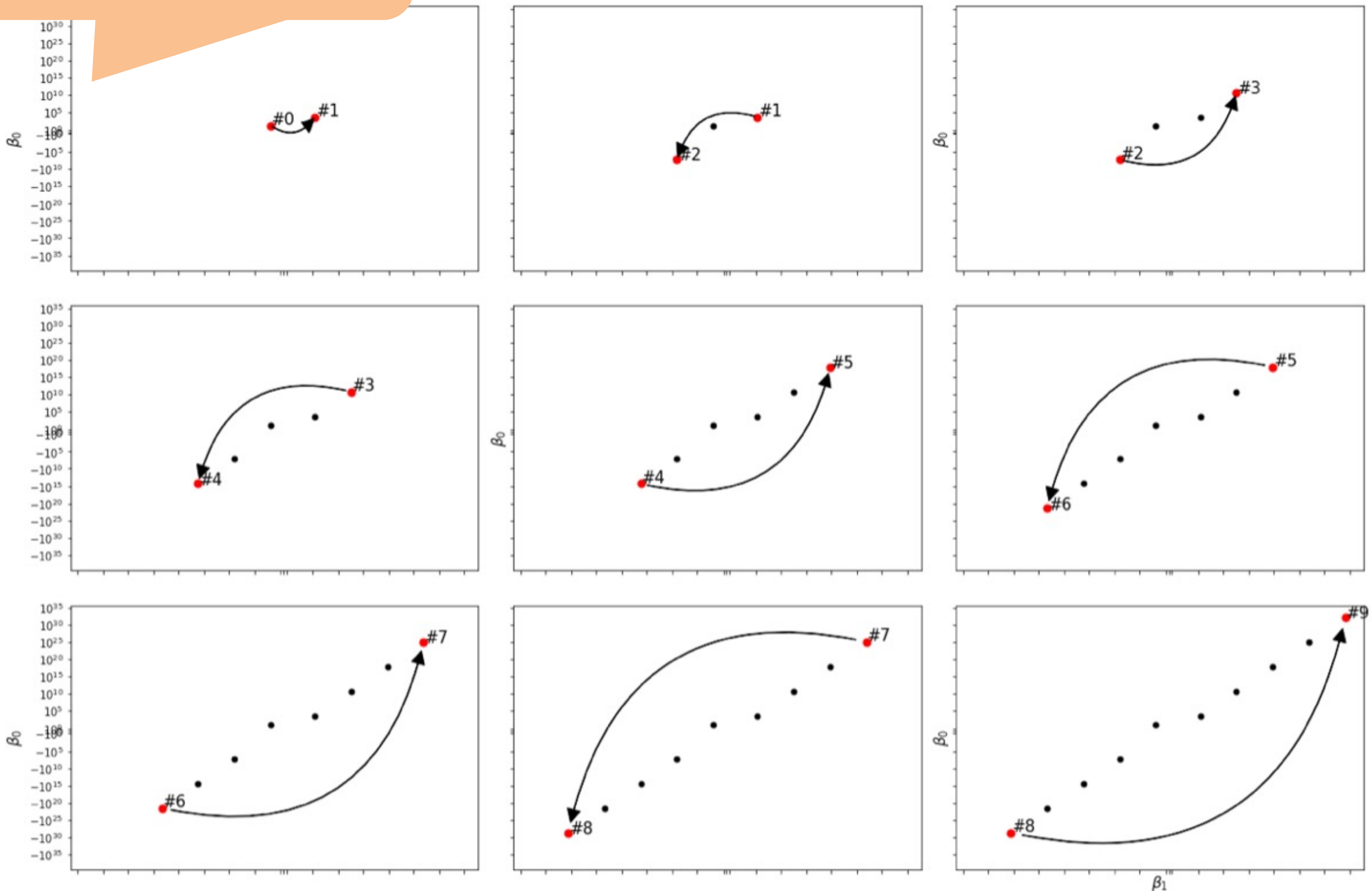
- Gradient Descent:

$$\beta_1 := \beta_1 - \frac{[-\Sigma_i^N x_i \epsilon_i]}{N} \qquad \beta_0 := \beta_0 - \frac{[-\Sigma_i^N \epsilon_i]}{N}$$

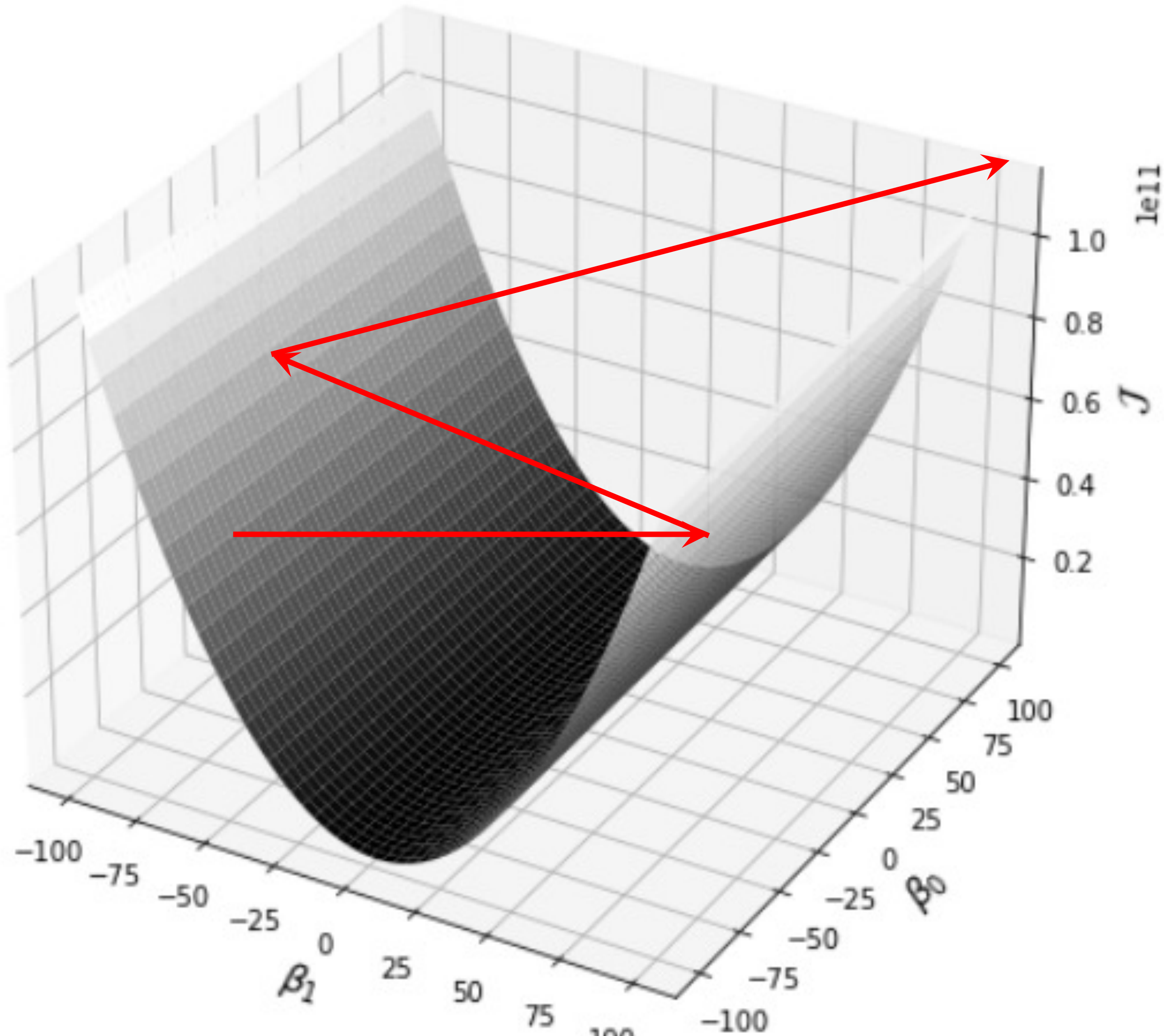
- Iterative update of the parameters to push them walking down the hill.

However, we need to walk slowly

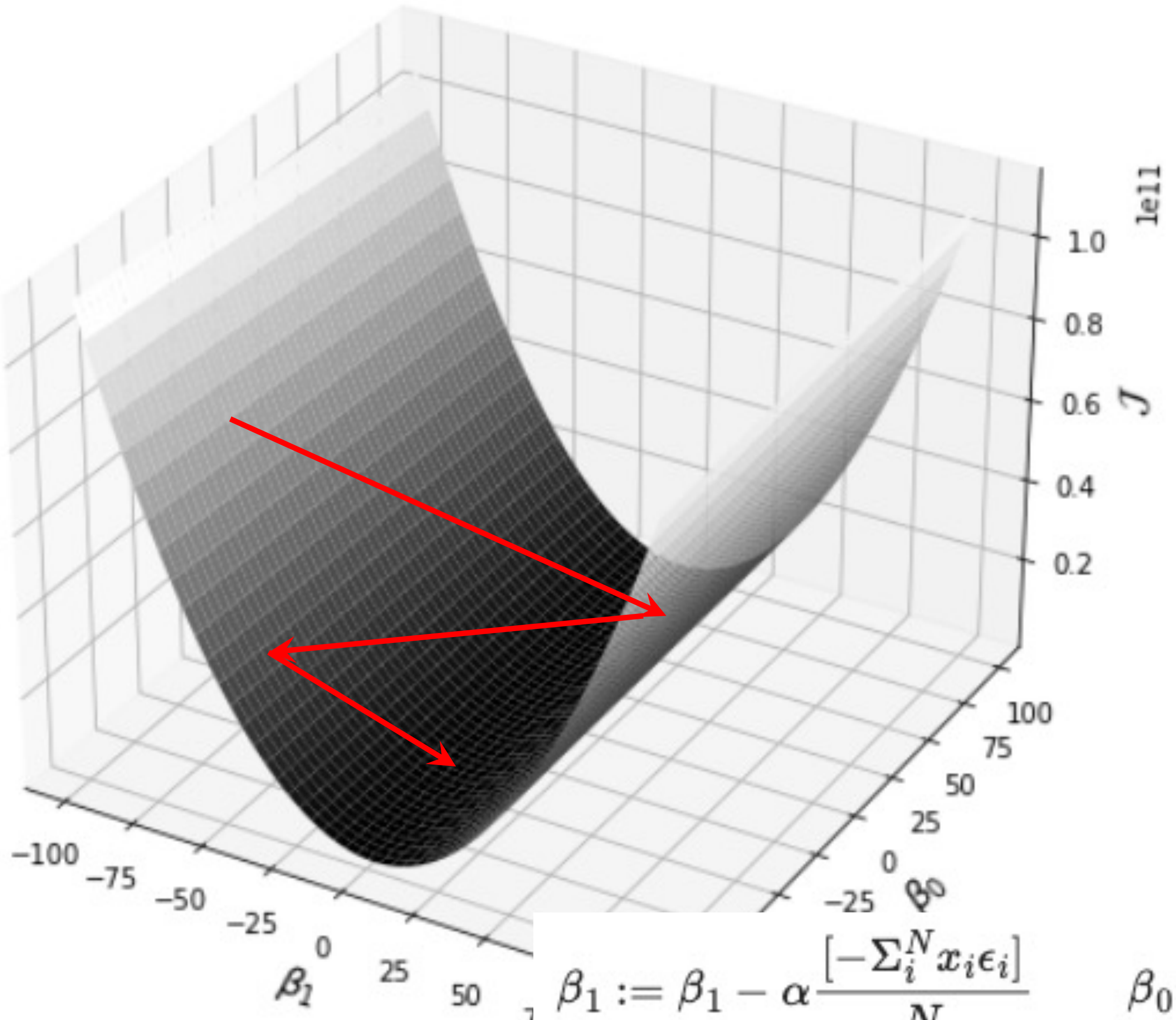
Throughout the iterations, the parameter values are diverging.



However, we need to walk slowly



However, we need to walk slowly

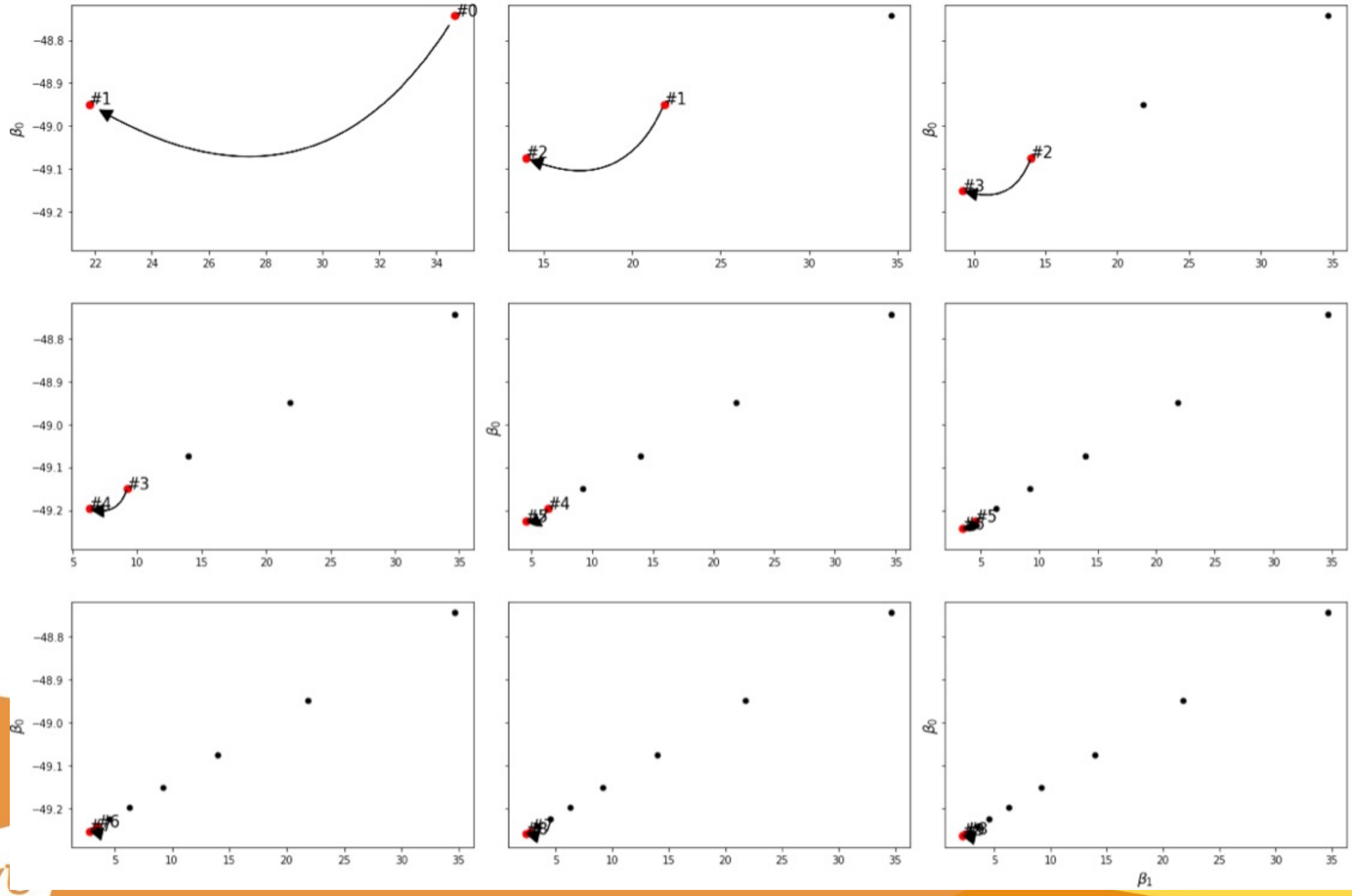


Learning rate

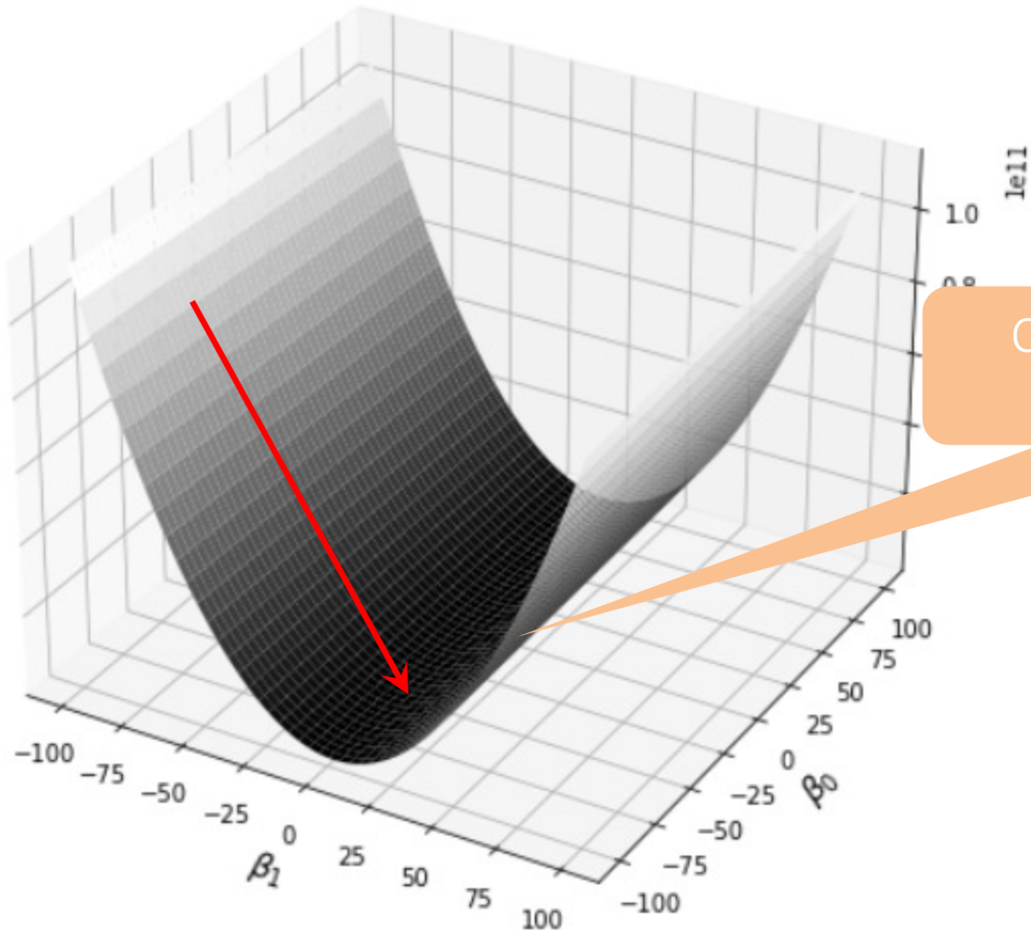
$$\beta_1 := \beta_1 - \alpha \frac{[-\Sigma_i^N x_i \epsilon_i]}{N}$$

$$\beta_0 := \beta_0 - \alpha \frac{[-\Sigma_i^N \epsilon_i]}{N}$$

Now it works better…



Can we directly go to the best point?



Convex! There is a global optimum!

A Probabilistic Perspective (Optional)

$$p(y|\theta) = p(y|x,\ \beta_1,\ \beta_0)$$

$$\downarrow$$

$$y = \beta_1 x + \beta_0 + \epsilon \qquad\qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\searrow \quad \swarrow$$

$$y|x,\ \beta_1,\ \beta_0 \sim \mathcal{N}(\beta_1 x + \beta_0, \sigma^2)$$

$$\downarrow$$

$$p(y|\theta) = p(y|x,\ \beta_1,\ \beta_0)$$
$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big[- \frac{(y_i - \beta_1 x_i - \beta_0)^2}{2\sigma^2} \Big]$$

A Probabilistic Perspective (Optional)

$$p(y|\theta) = p(y|x,\ \beta_1,\ \beta_0)$$
$$= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y_i - \beta_1 x_i - \beta_0)^2}{2\sigma^2}\right]$$

Log-transformation to turn products into summations.

$$l(\theta) = \log \mathcal{L}(\theta) = \log p(y|\theta)$$
$$= \sum_i \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_i \frac{(y_i - \beta_1 x_i - \beta_0)^2}{2\sigma^2}$$
$$= -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2$$
$$= -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{\sigma^2} \mathcal{J}$$

MLE is equivalent to Least Square!

A Probabilistic Perspective (Optional)

$$\frac{\partial l}{\partial \beta_1}$$

$$\frac{\partial l}{\partial \beta_0}$$

$$\frac{\partial l}{\partial \sigma}$$

Set partial derivatives to zeros.

$$\beta_1 = \frac{\sum_i x_i y_i - \bar{y}\sum_i x_i}{\sum_i x_i^2 - \bar{x}\sum_i x_i}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\sigma^2 = \frac{\sum_i (y_i - \beta_1 x_i - \beta_0)^2}{N}$$

See notebook for the comparison on the estimated parameters between sklearn's output and our manual computation

Recap: Parameter Estimation

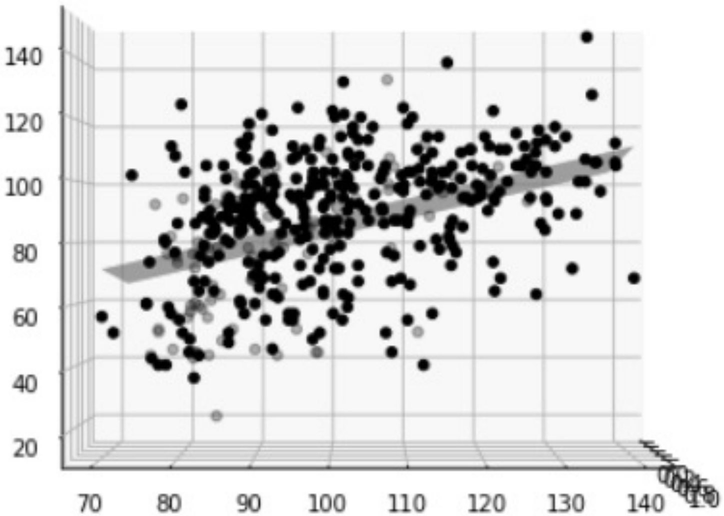
- What are we doing when we train a prediction mode?

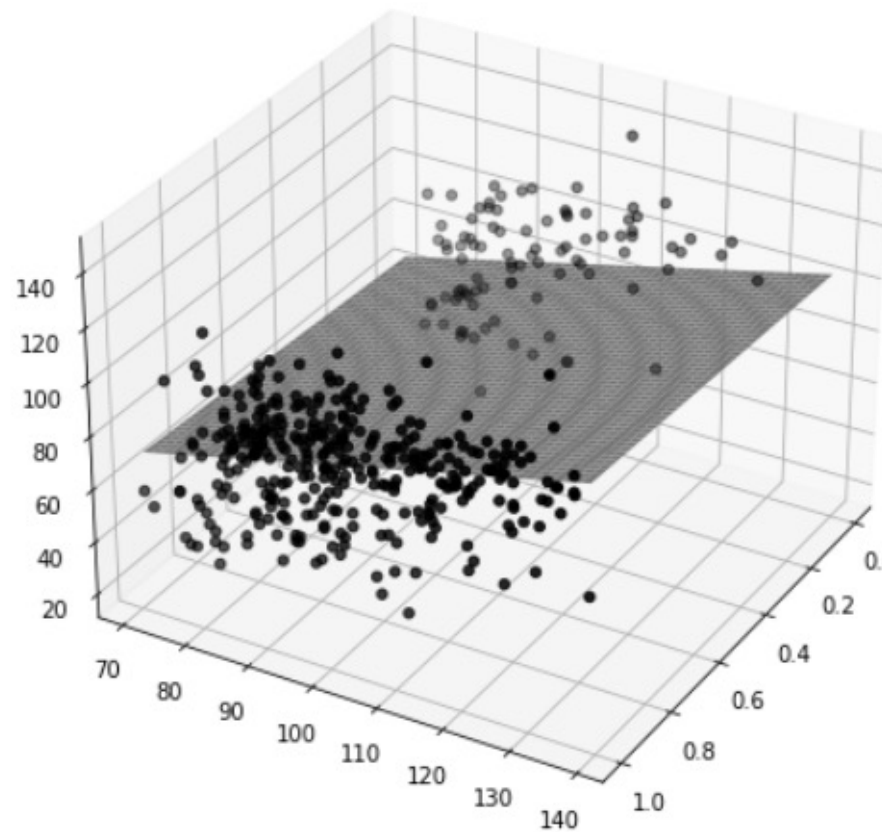
We are walking on the parameter space, looking for those who achieve the lowest error possible.

- This applies to various machine learning models for prediction.

Multiple Linear Regression

$$\text{IQ} = 5.95 \times \text{MomHighSchool} + 0.56 \times \text{MomIQ} + 25.73$$





Caveat in Interpretation

$$\text{IQ} = 5.95 \times \text{MomHighSchool} + 0.56 \times \text{MomIQ} + 25.73$$

	MomHS	MomIQ
count	434.00	434.00
mean	0.79	100.00
std	0.41	15.00
min	0.00	71.04
25%	1.00	88.66
50%	1.00	97.92
75%	1.00	110.27
max	1.00	138.89

Caveat in Interpretation

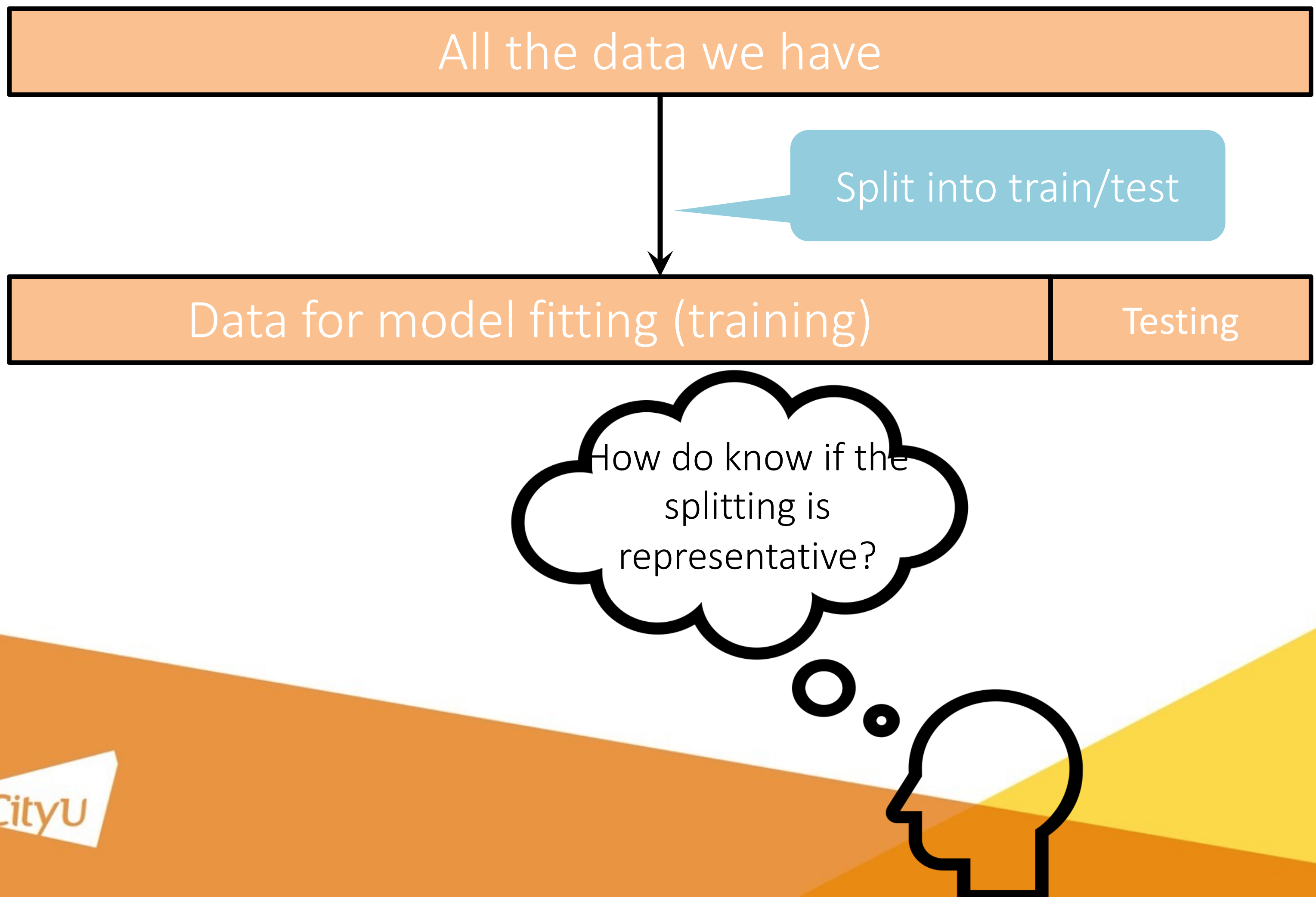
$$\mathrm{IQ} = 5.95 \times \mathrm{MomHighSchool} + 0.56 \times \mathrm{MomIQ} + 25.73$$

	MomHS	MomIQ
count	434.00	434.00
mean	0.79	100.00
std	0.41	15.00
min	0.00	71.04
25%	1.00	88.66
50%	1.00	97.92
75%	1.00	110.27
max	1.00	138.89

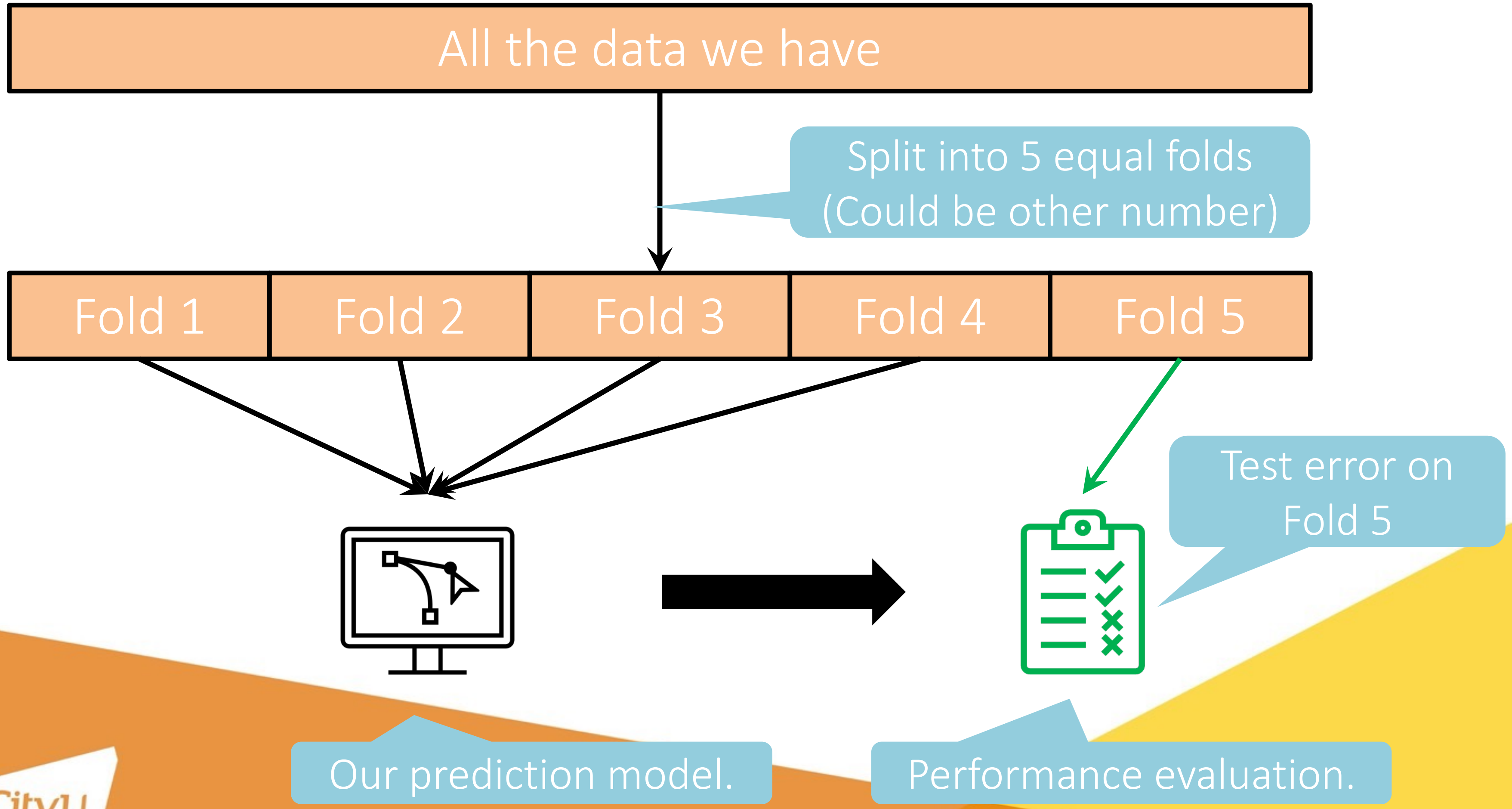
Suggestion: Standardize the dataset such that the interpretation becomes:
"1 SD increase in X is associated with some SD increase in Y"

Caveat: When we have train/test split, we need to standardize test data based on the train data.

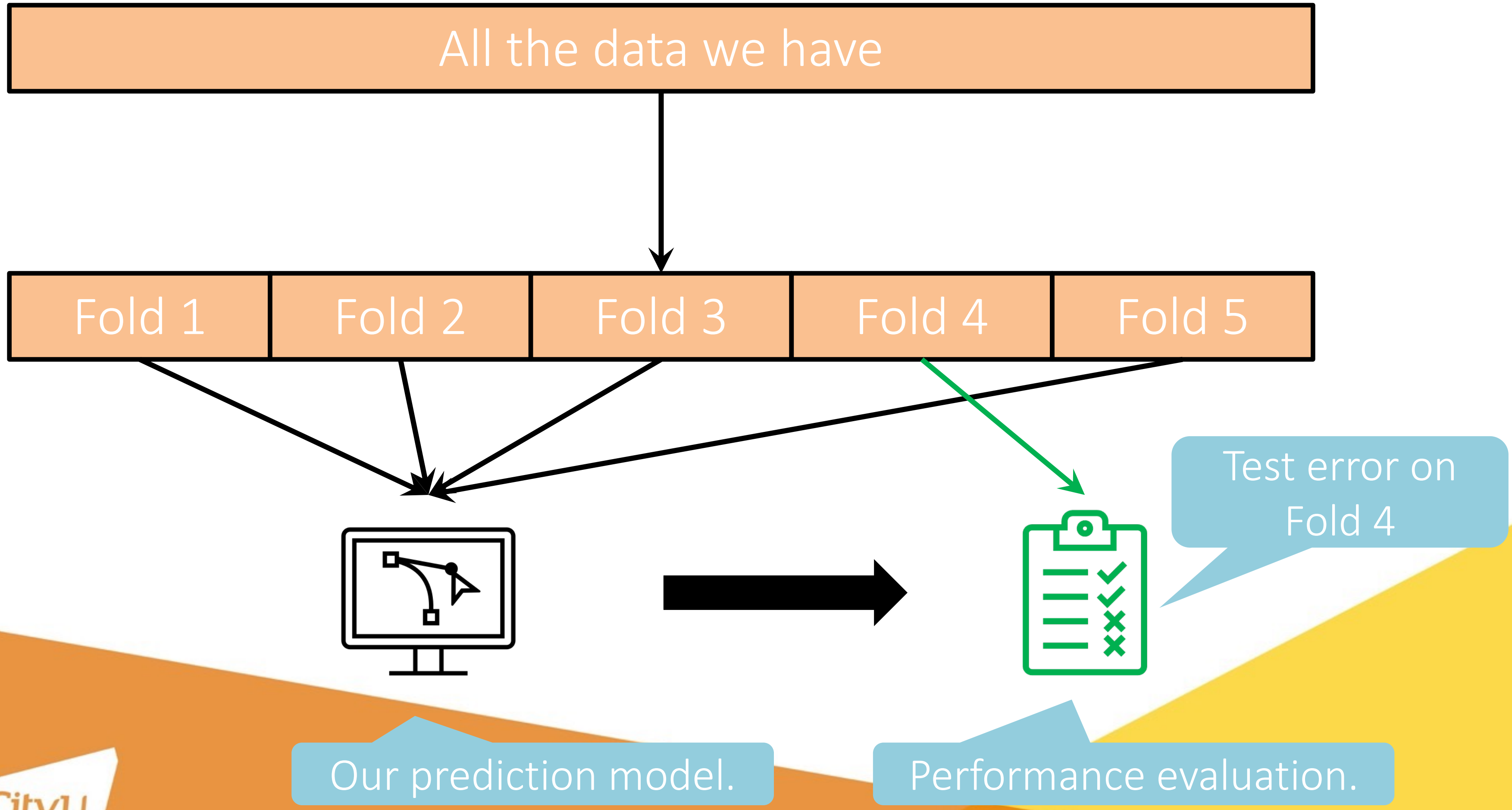
Prediction Performance Evaluation



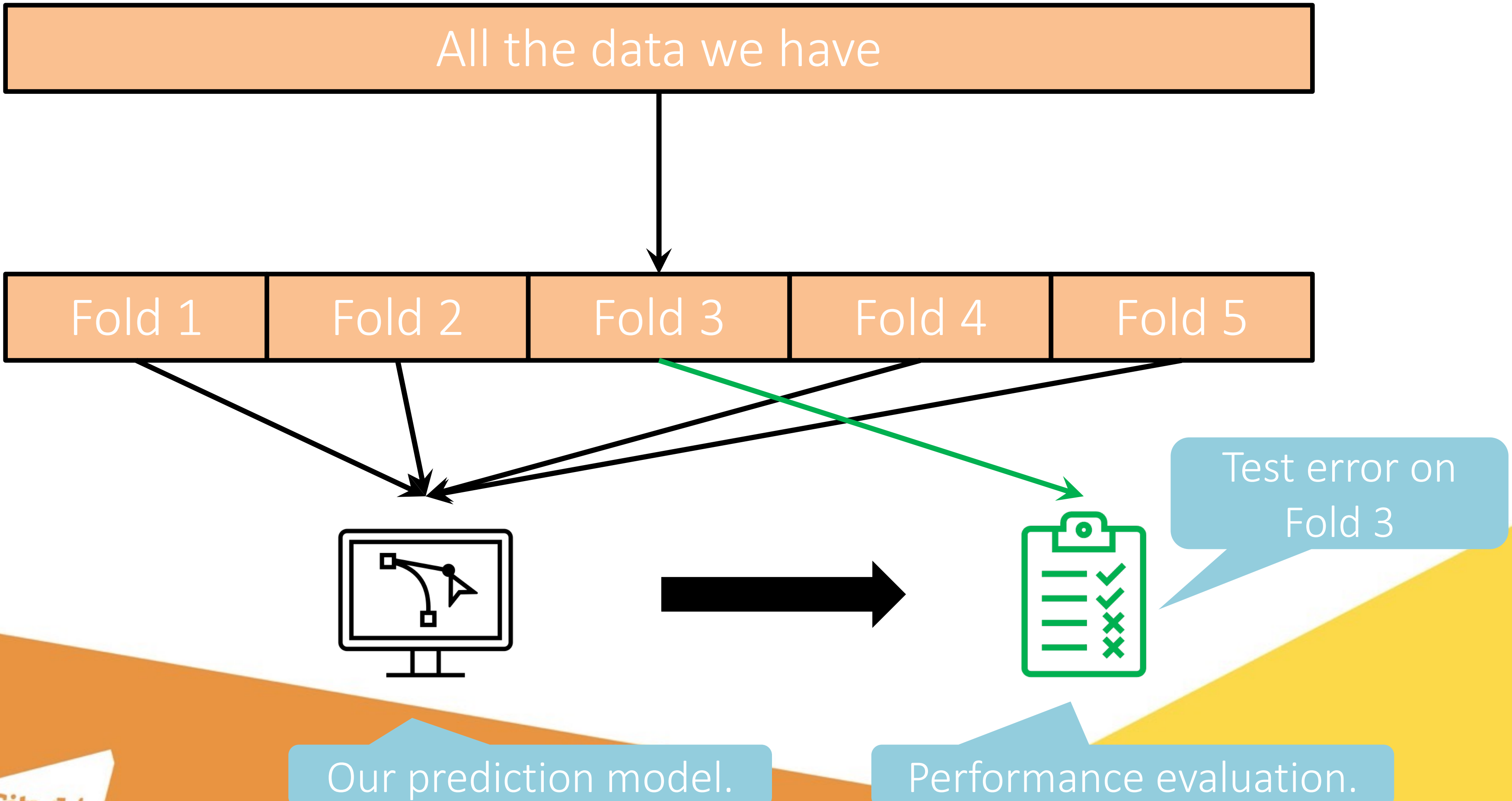
(K-Fold) Cross Validation



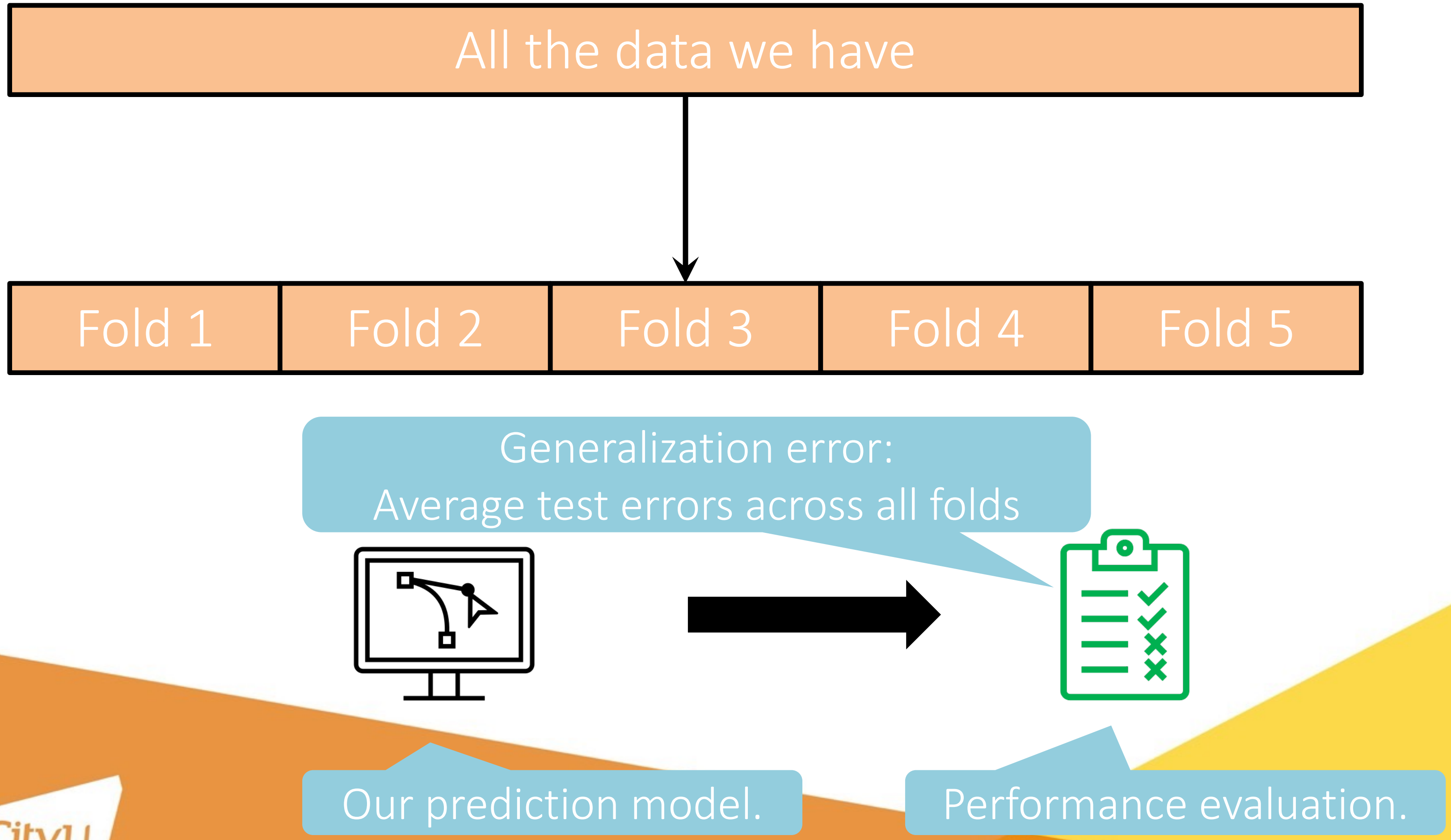
(K-Fold) Cross Validation



(K-Fold) Cross Validation



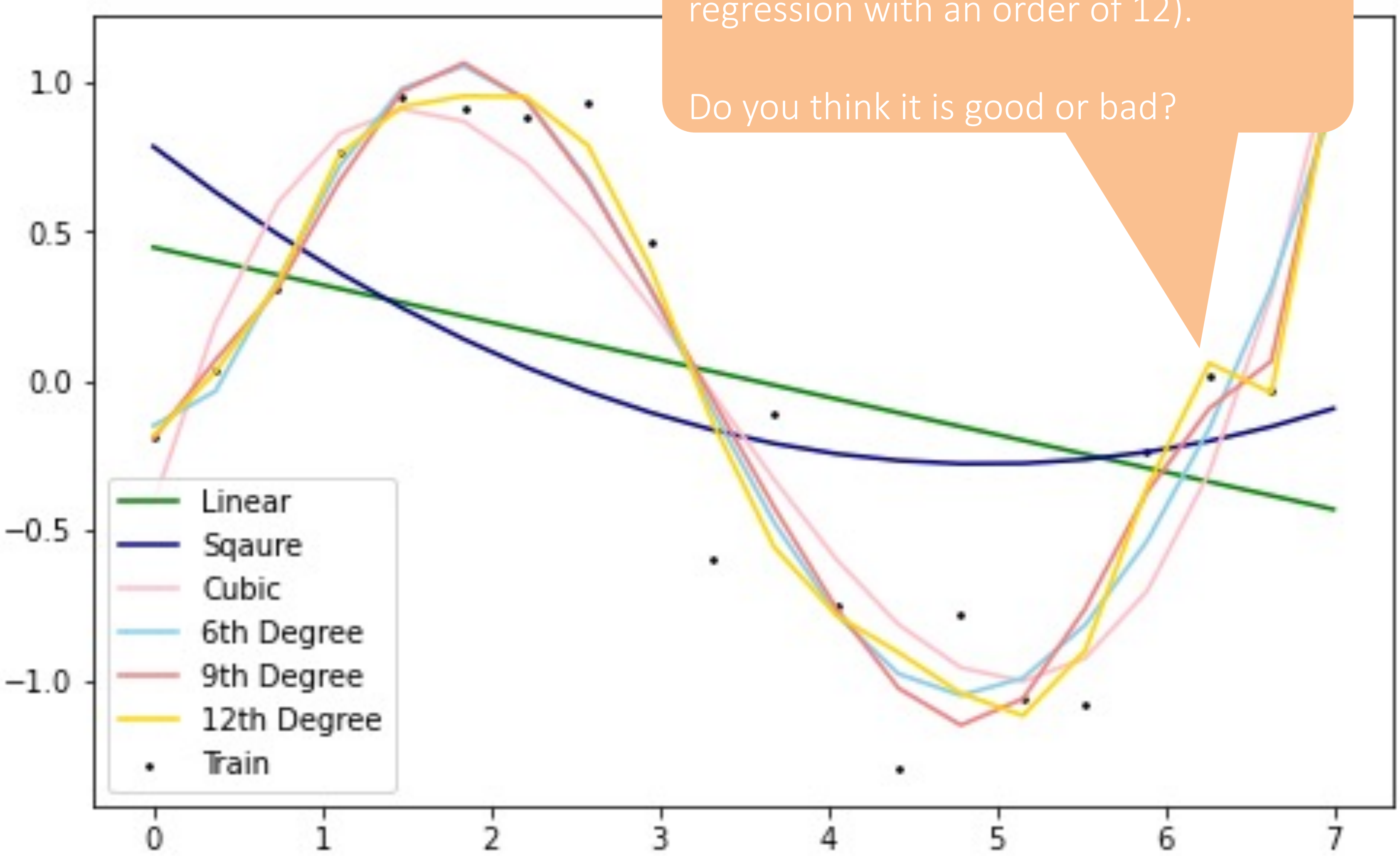
(K-Fold) Cross Validation



Overfitting: Fit the pattern or noise?

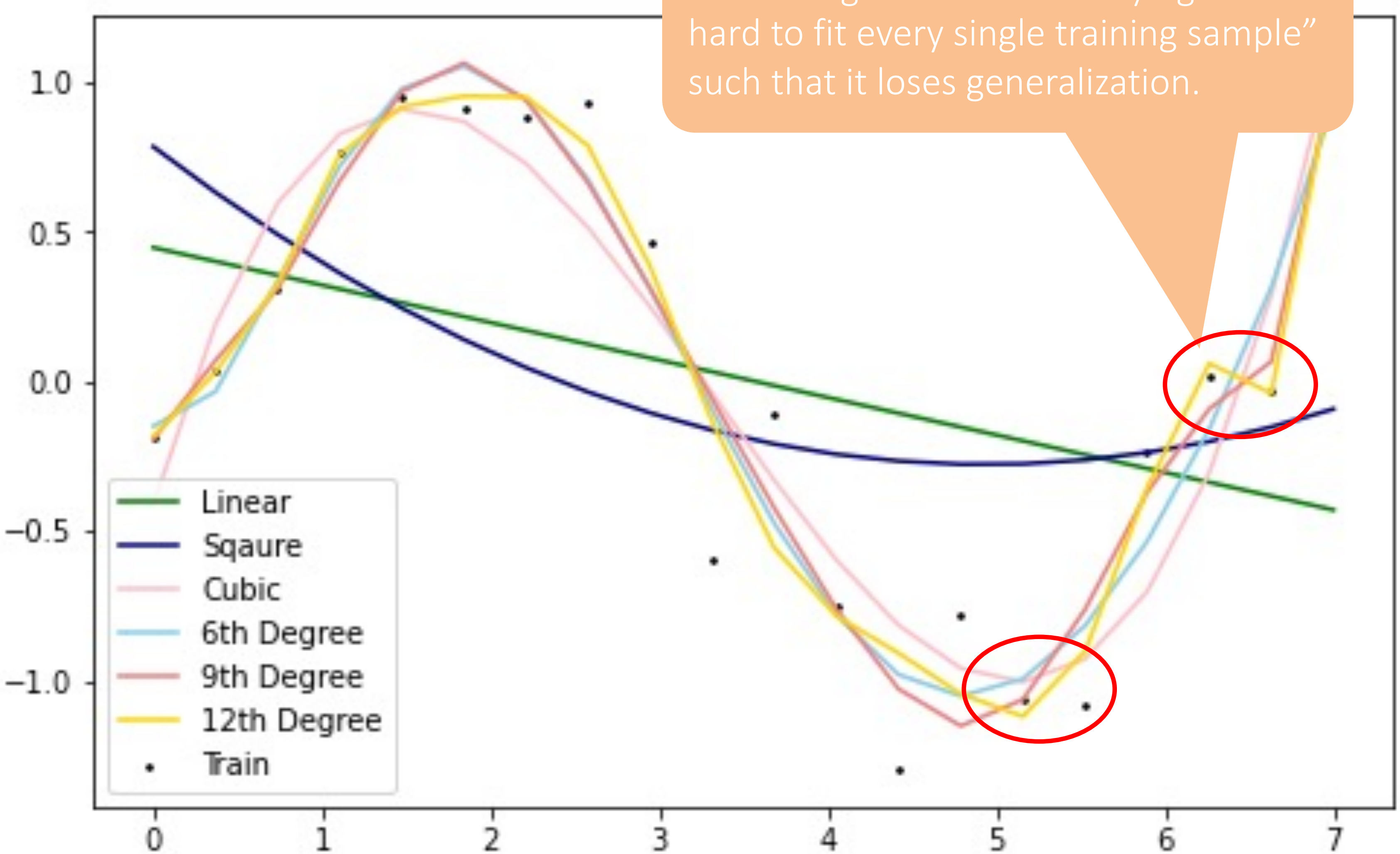
Look at the orange line (polynomial regression with an order of 12).

Do you think it is good or bad?

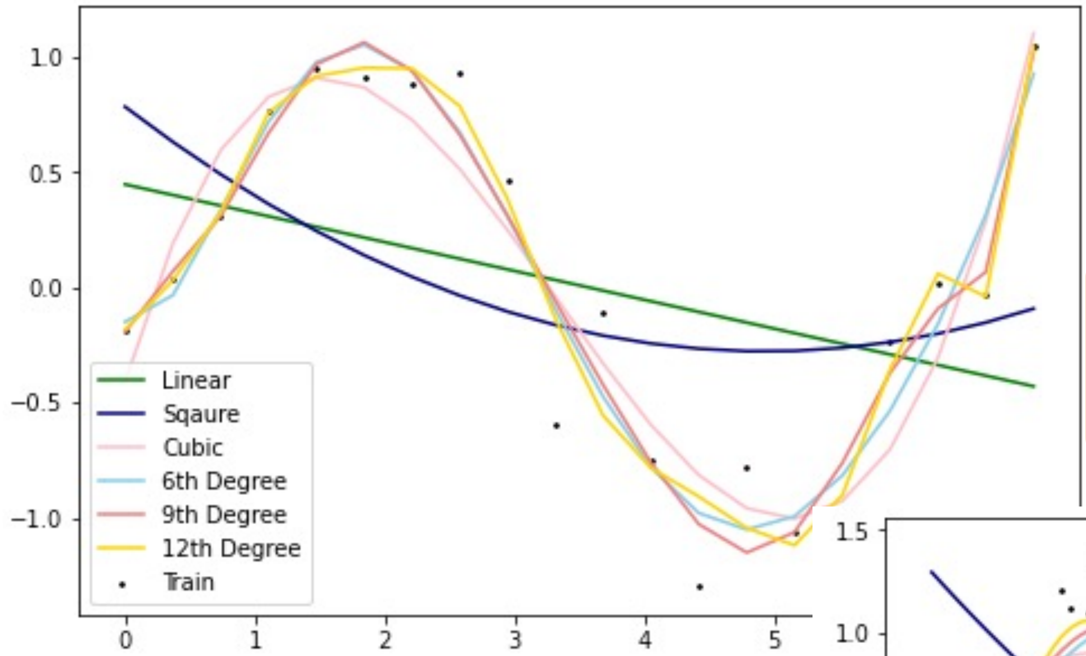


Overfitting: Fit the pattern or noise?

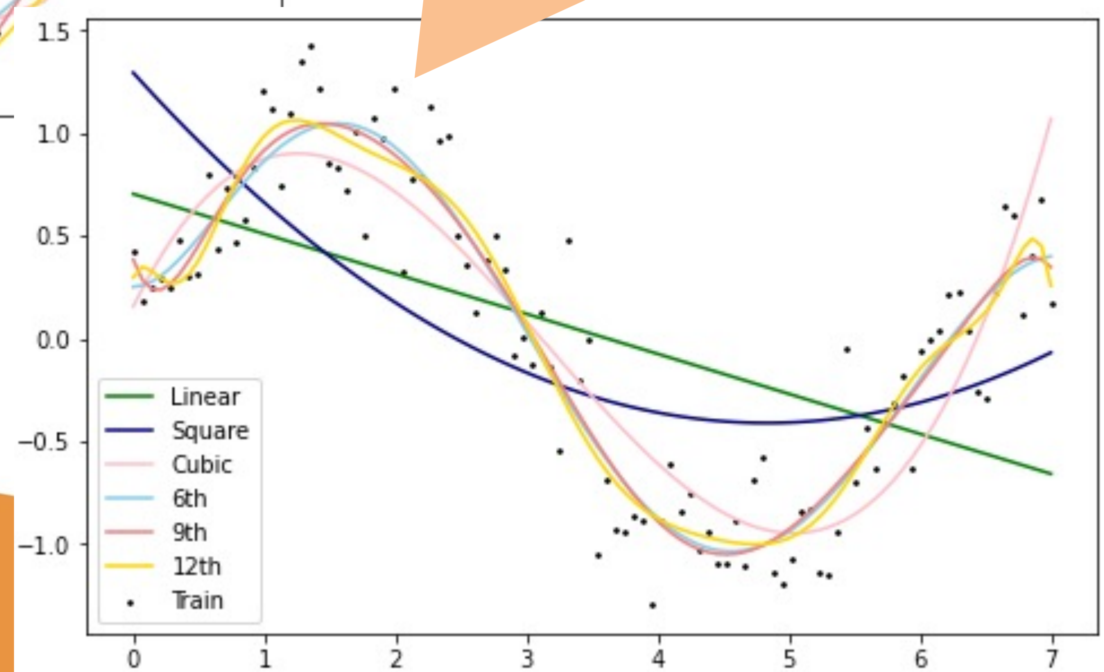
Overfitting: the model is ”trying too hard to fit every single training sample” such that it loses generalization.



Overfitting Solution 1: More data!

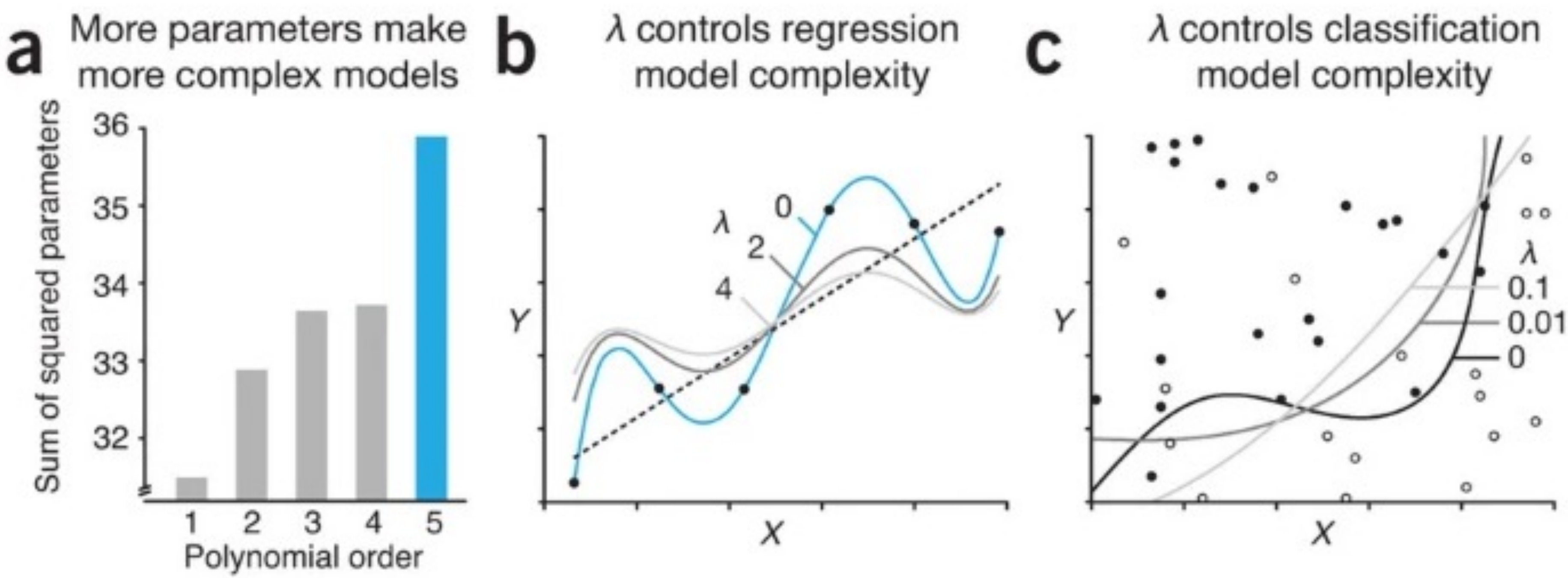


Complex models look “smoother” when more data is given.



But data is expensive

- Can we mitigate this issue by tweaking our model?



Source: https://www.nature.com/articles/nmeth.4014

Regularization to the rescue

- A model is more flexible with more parameters, which is often associated greater parameter magnitude

$$\mathcal{J}(\beta_1, \beta_0) = \frac{1}{2}\Sigma_i^N {\epsilon_i}^2 = \frac{1}{2}\Sigma_i^N [y_i - (\beta_1 x_i + \beta_0)]^2$$

$$\downarrow$$

$$\mathcal{J}(\beta) = \tfrac{1}{2}\Sigma_i^N [y_i - (\textstyle\sum_j \beta_j x_{ij} + \beta_0)]^2 + \textstyle\sum_j \|\beta_j\|^2$$

Ridge regression (i.e., linear regression with L-2 norm regularization)

Regularization: Lasso and Elastic Net

$$\mathcal{J}(\beta) = \frac{1}{2}\Sigma_i^N[y_i - (\sum_j \beta_j x_{ij} + \beta_0)]^2 + \sum_j \|\beta_j\|$$

Lasso regression (i.e., linear regression with L-1 norm regularization)

Look them up in scikit-learn!

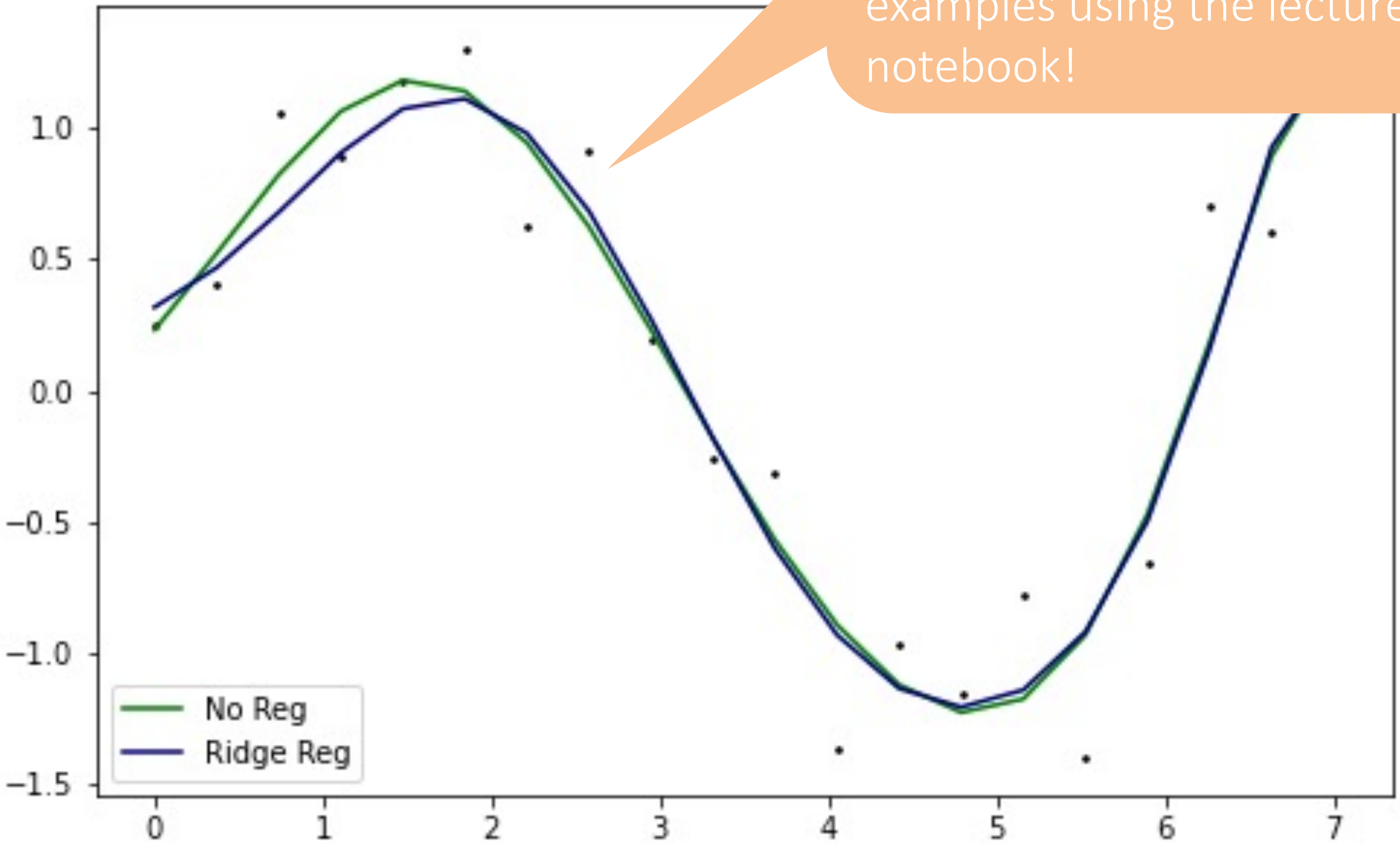
$$\mathcal{J}(\beta) = \frac{1}{2}\Sigma_i^N[y_i - (\sum_j \beta_j x_{ij} + \beta_0)]^2 + \sum_j \|\beta_j\| + \sum_j \|\beta_j\|^2$$

Elastic net regression (i.e., linear regression with L-1and L-2 norm regularization)

How does it work?

Regularization makes the model less specific to the noise!

You are encouraged to try more examples using the lecture notebook!

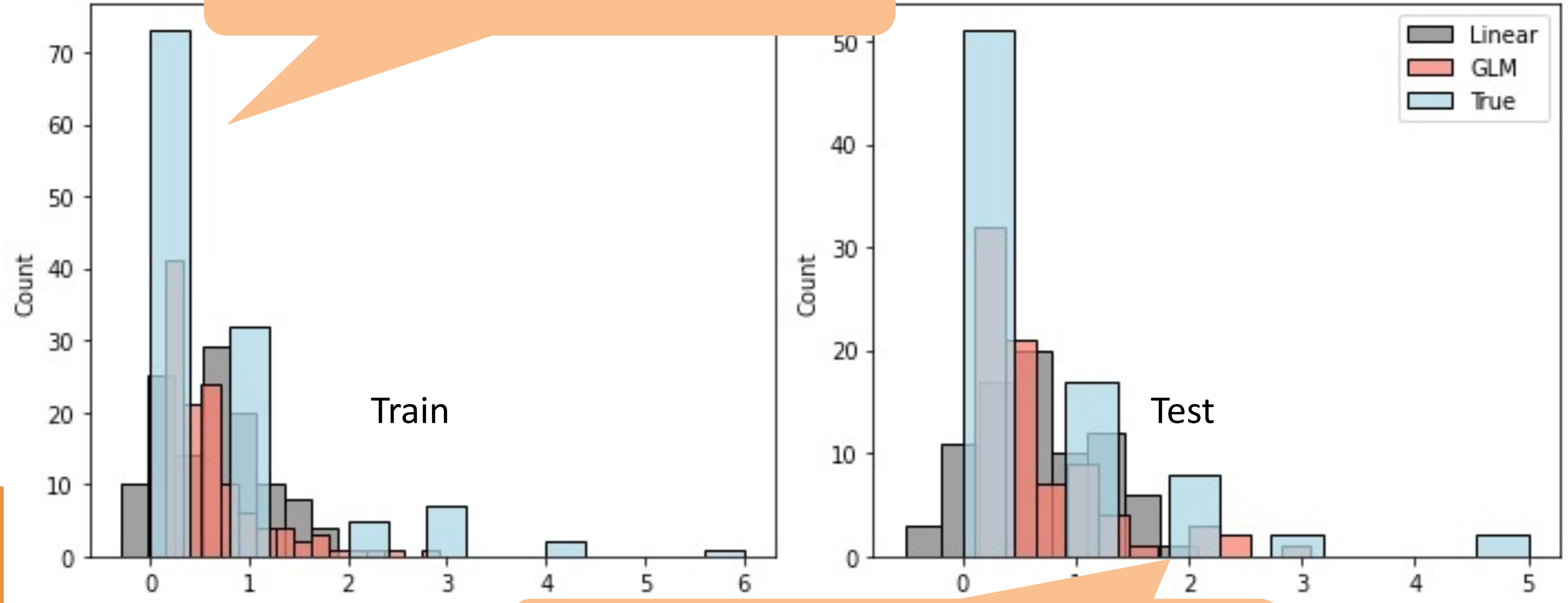


Advanced topics: GLM

- GLM: Generalized Linear Model
- Useful when target variable is not real number.
  - Examples: Count data (car accidents, elections, etc.) or binary target (two-class prediction)

Example: Poisson reg. (count data)

True data is discrete count.



Obvious Poisson regression better fits the distribution

Advanced topics: Non-parametric

- Parametric models: a fixed set of model parameters regardless of sample size.

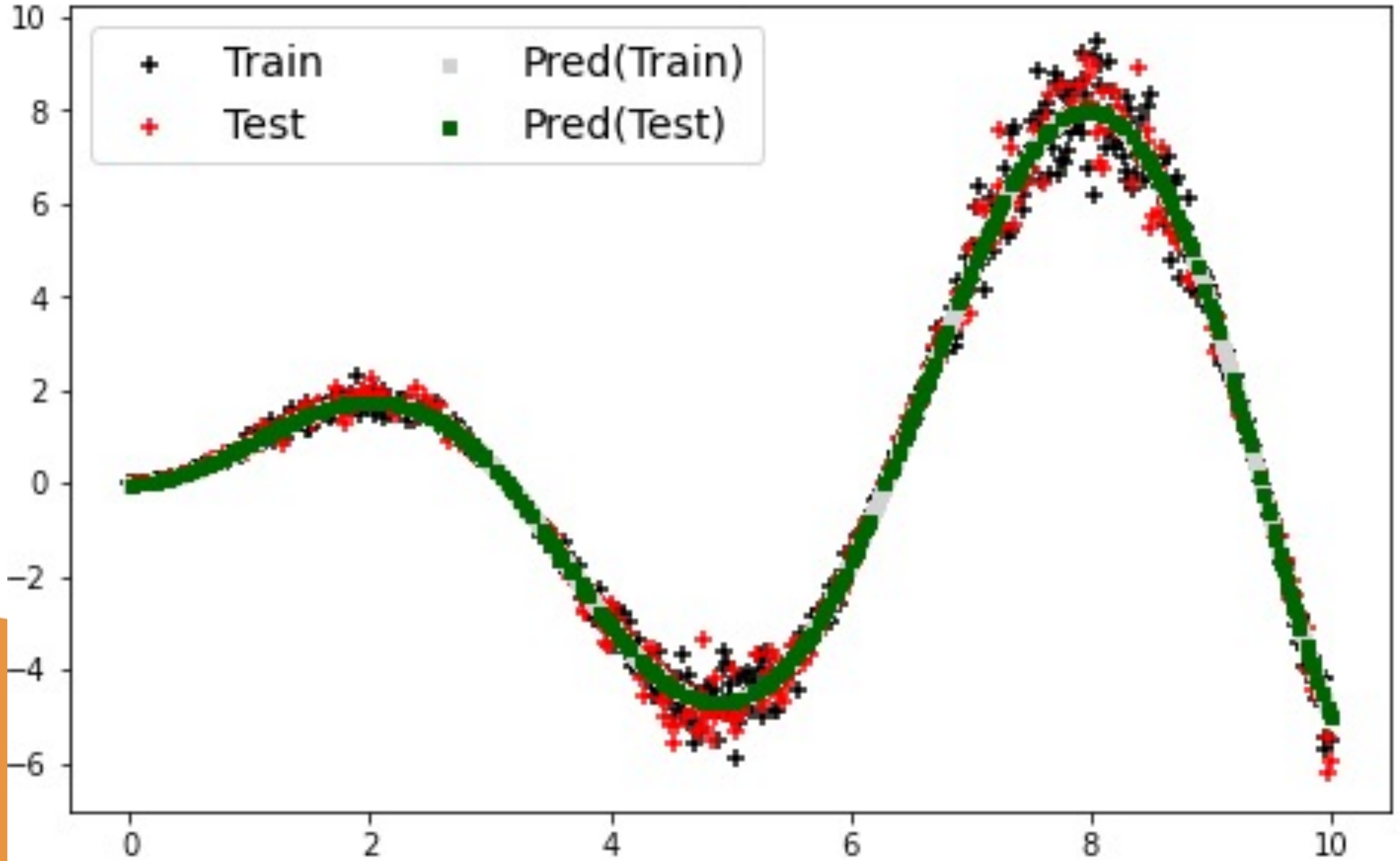
- E.g., a linear regression with 4 variables will have 4 slopes and 1 intercept (as well as the error term variance) no matter how many rows we have

Advanced topics: Non-parametric

- Non-Parametric models: a dynamic number of model parameters given different sample sizes.

- E.g., Gaussian Process
  - It models the target using a mixture of Gaussian distributions.
  - Good tutorial on this if interested (optional):
    - https://distill.pub/2019/visual-exploration-gaussian-processes/

Example: GP regression



Suggested References

- Bishop, Christopher M. (2006). Pattern recognition and machine learning. New York :Springer.
- Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press.
- Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

香港城市大學
City University of Hong Kong

The End

專業 創新 胸懷全球
Professional · Creative
For The World