CDSF09, CDSF10, CDSF11 and CDSF12

Recap IV

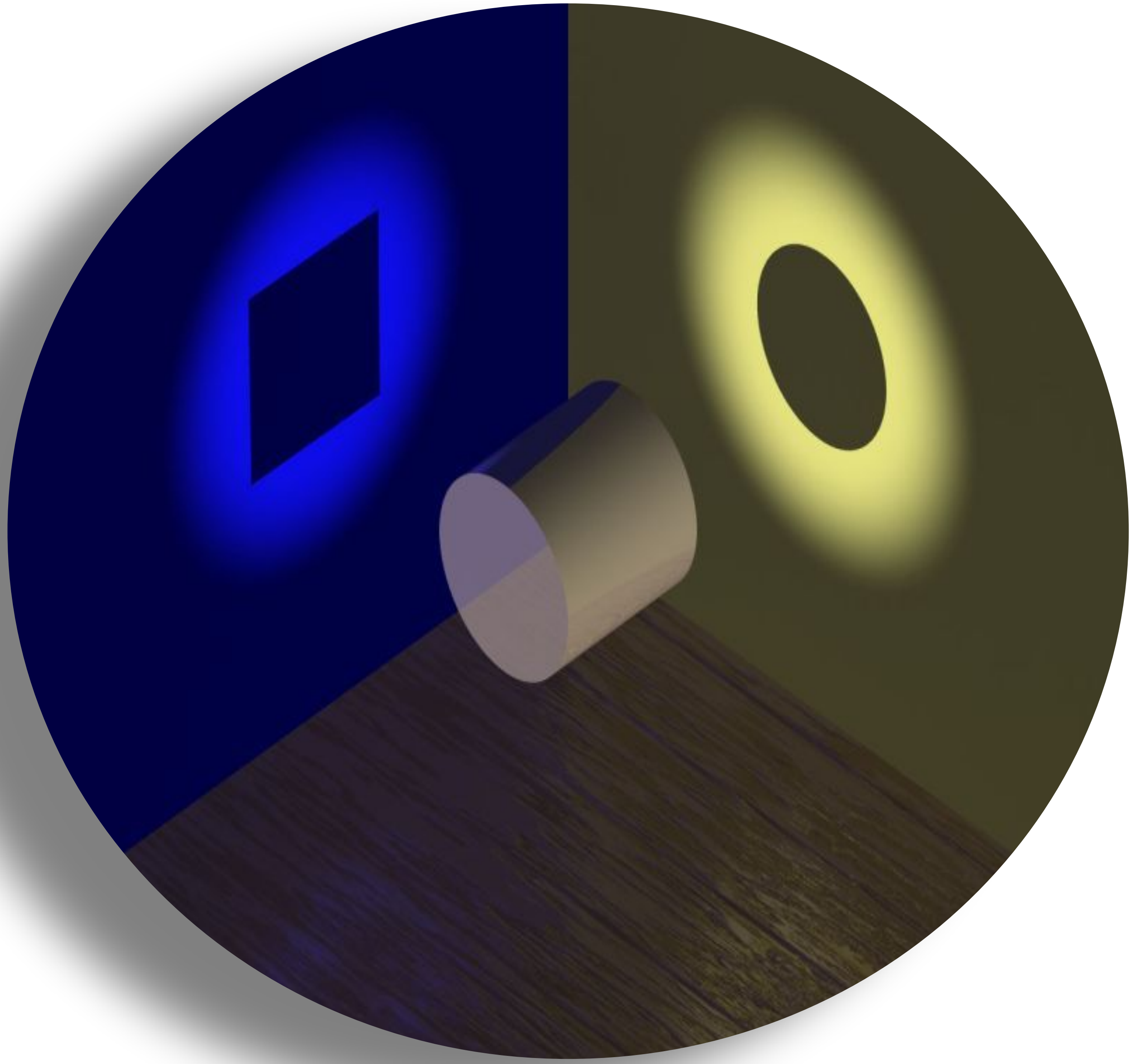
DS Academy

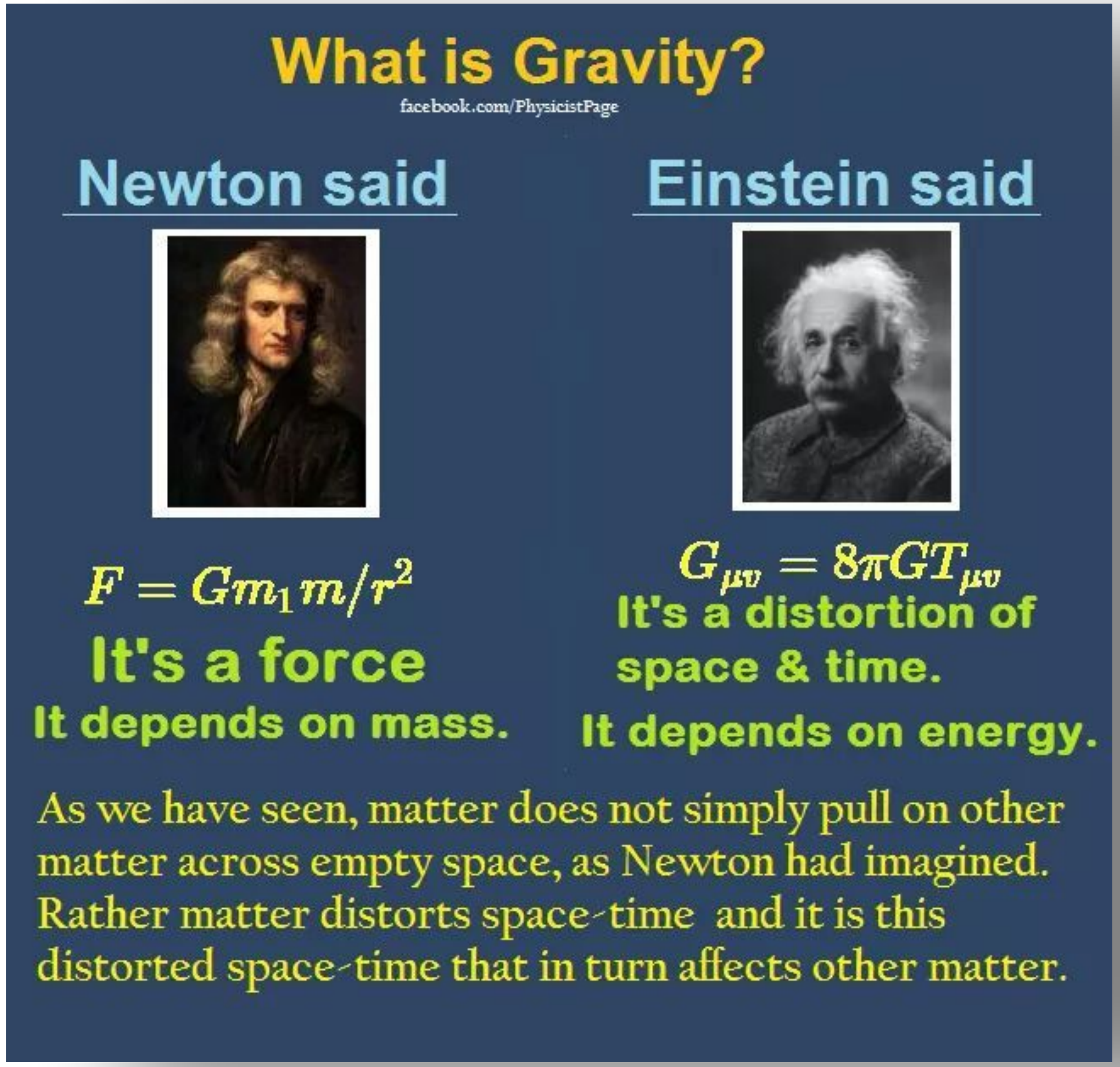
TODOS OS DIREITOS RESERVADOS

2018

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

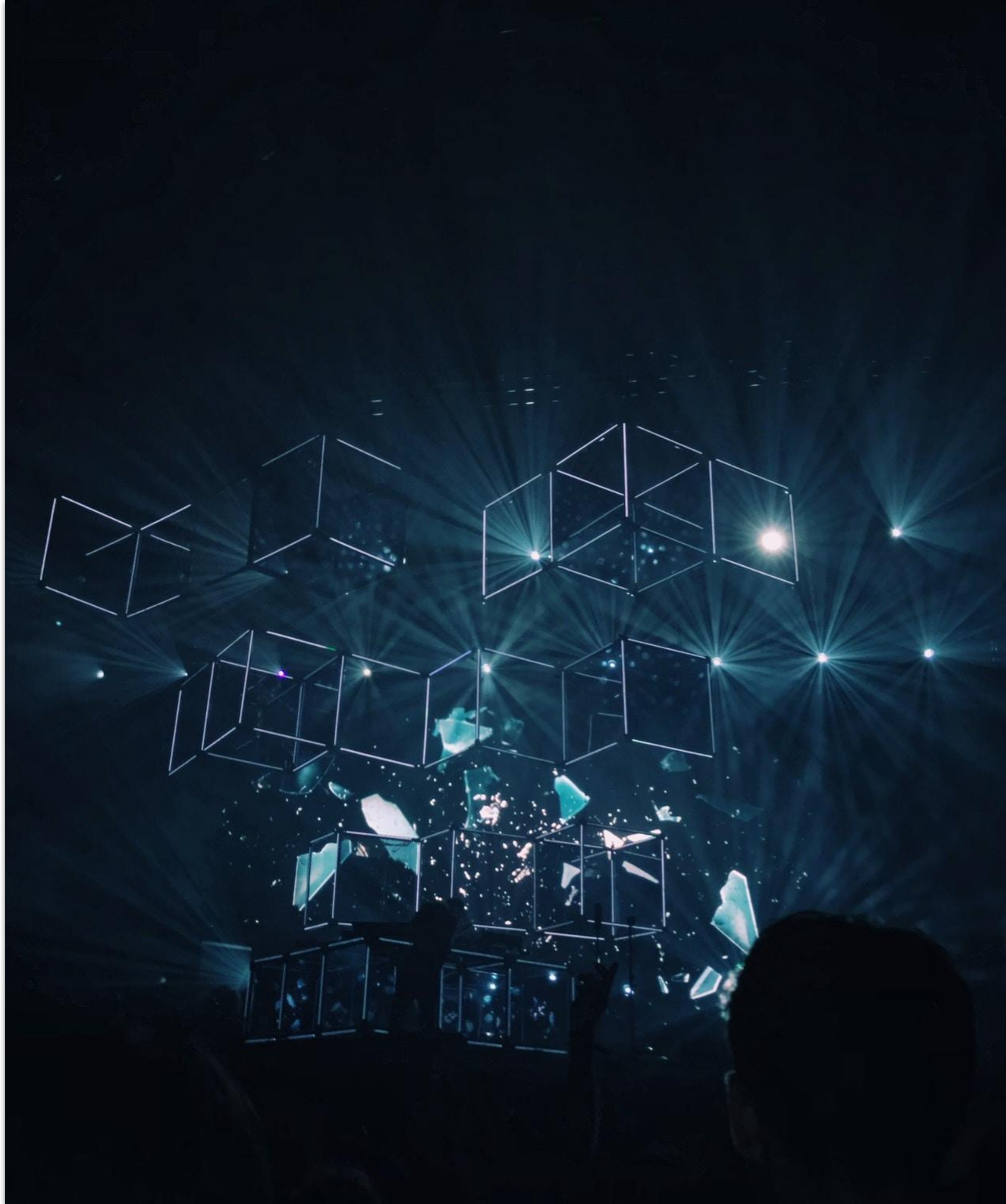
George Box





Agenda

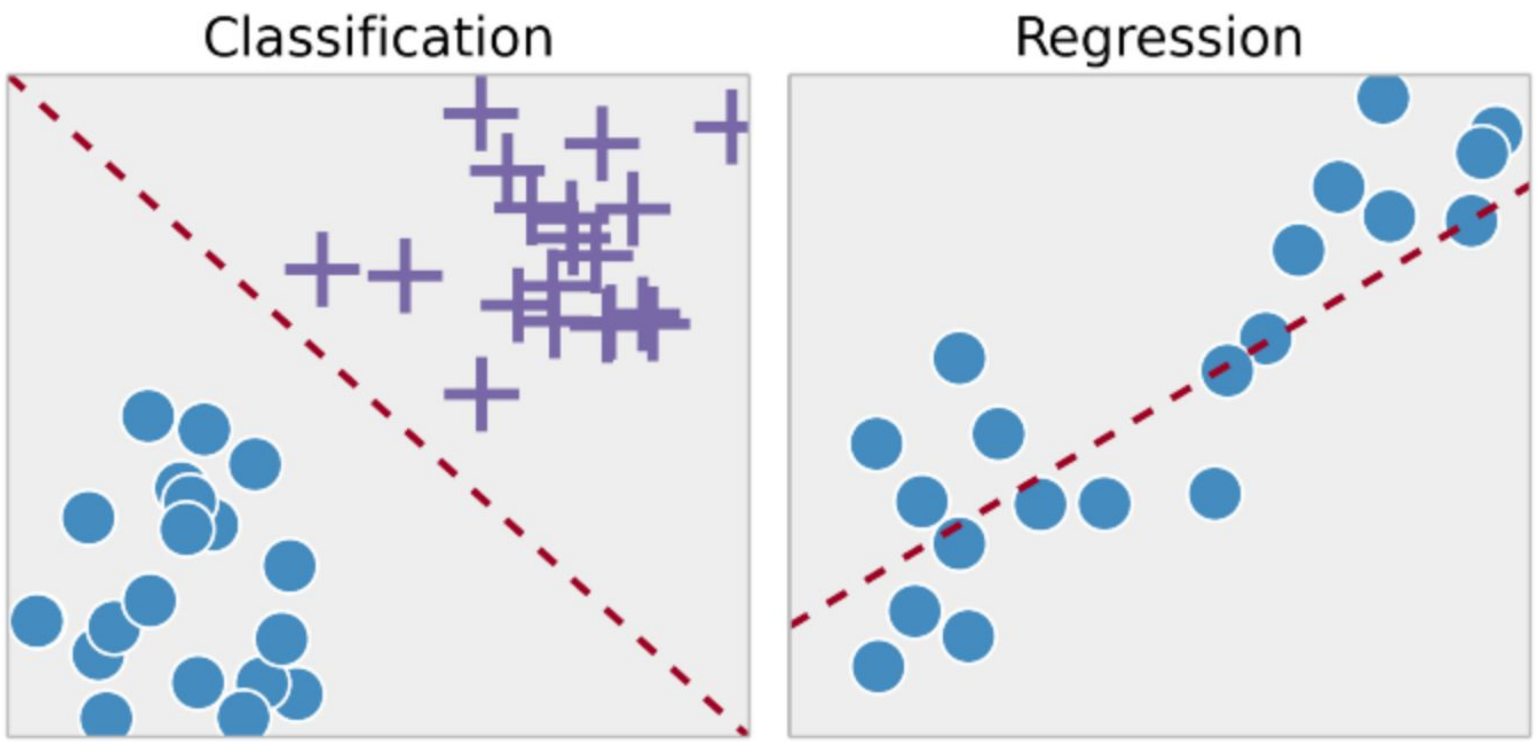
1. The Learning Problem
2. Bias and Variance
3. Train/Val/Test Split
4. Validation Techniques
5. Regularization
6. Dimensionality Reduction
7. PCA

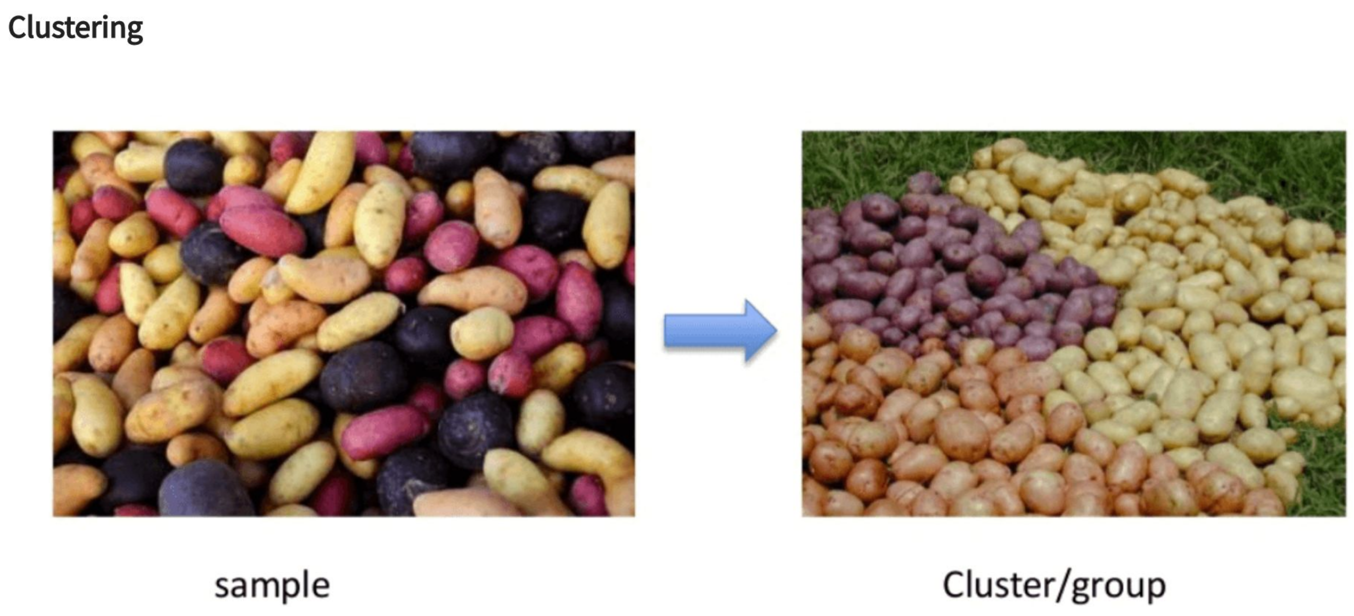


The learning problem

- **Supervised:**
  - **Regression**: to predict a single output value using training data.
  - **Classification**: to group the output inside a predefined class.

- **Unsupervised.**
  - **Clustering**: It deals with finding a structure or pattern in a collection of uncategorized data.





- **The bias error** is an error from erroneous assumptions in the learning algorithm.
  - High bias can cause an algorithm to miss the relevant relations between features and target outputs (*underfitting*).

- **The variance** is an error from sensitivity to small fluctuations in the training set.
  - High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

Bias vs. Variance

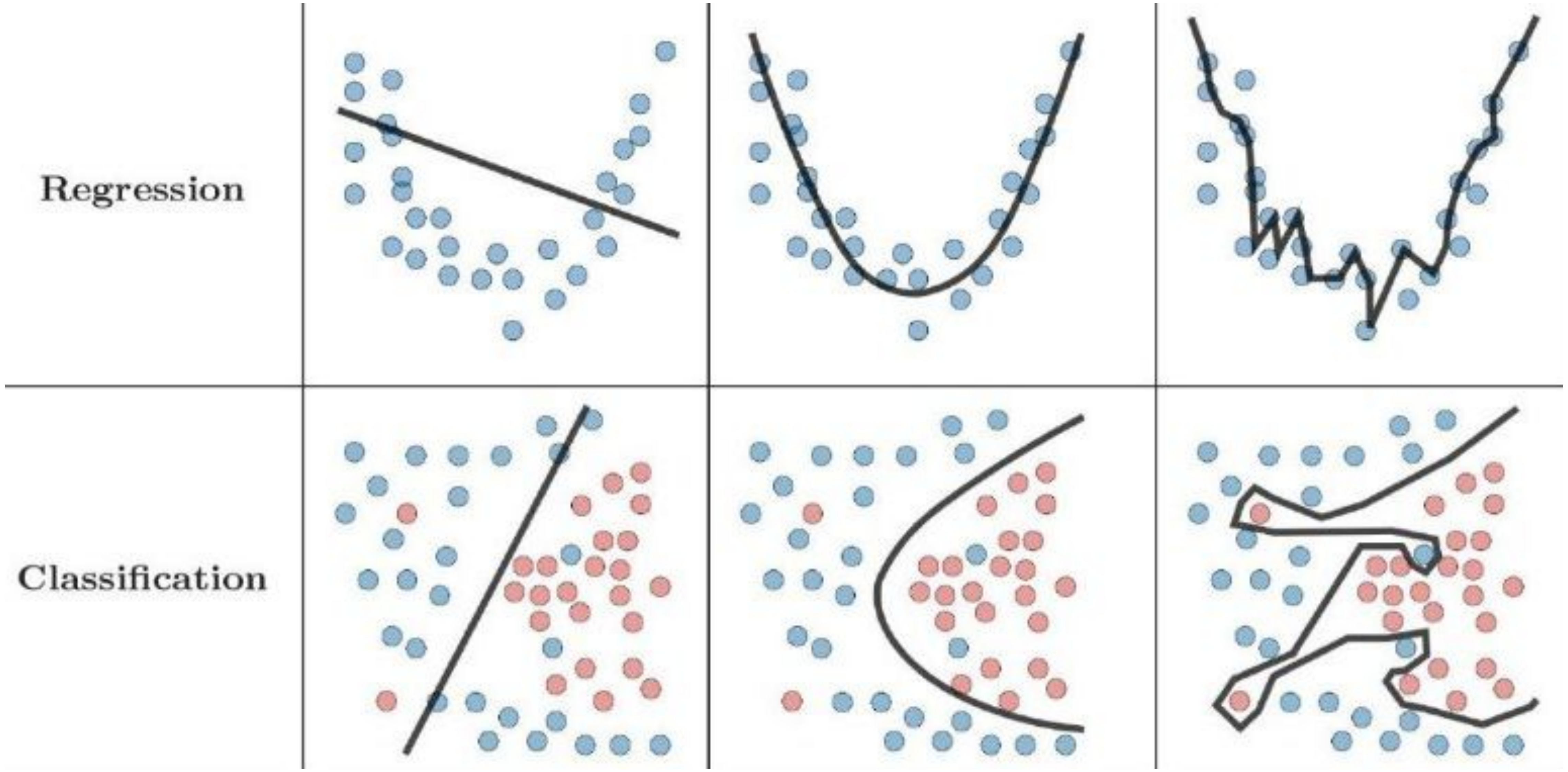


Image from TowardsDataScience

Bias vs. Variance

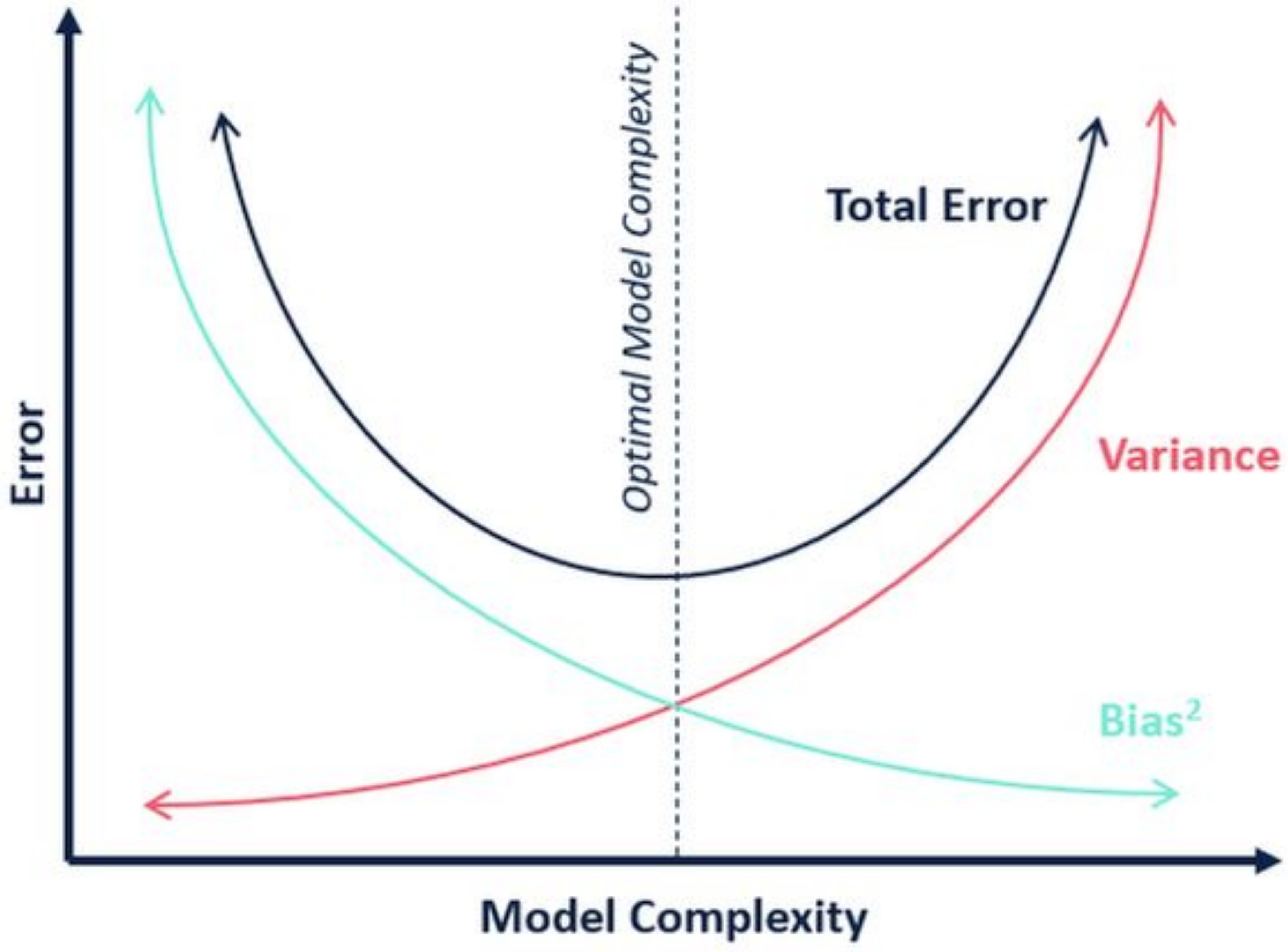


Image from Alteryx Community

Train / Validation / Test

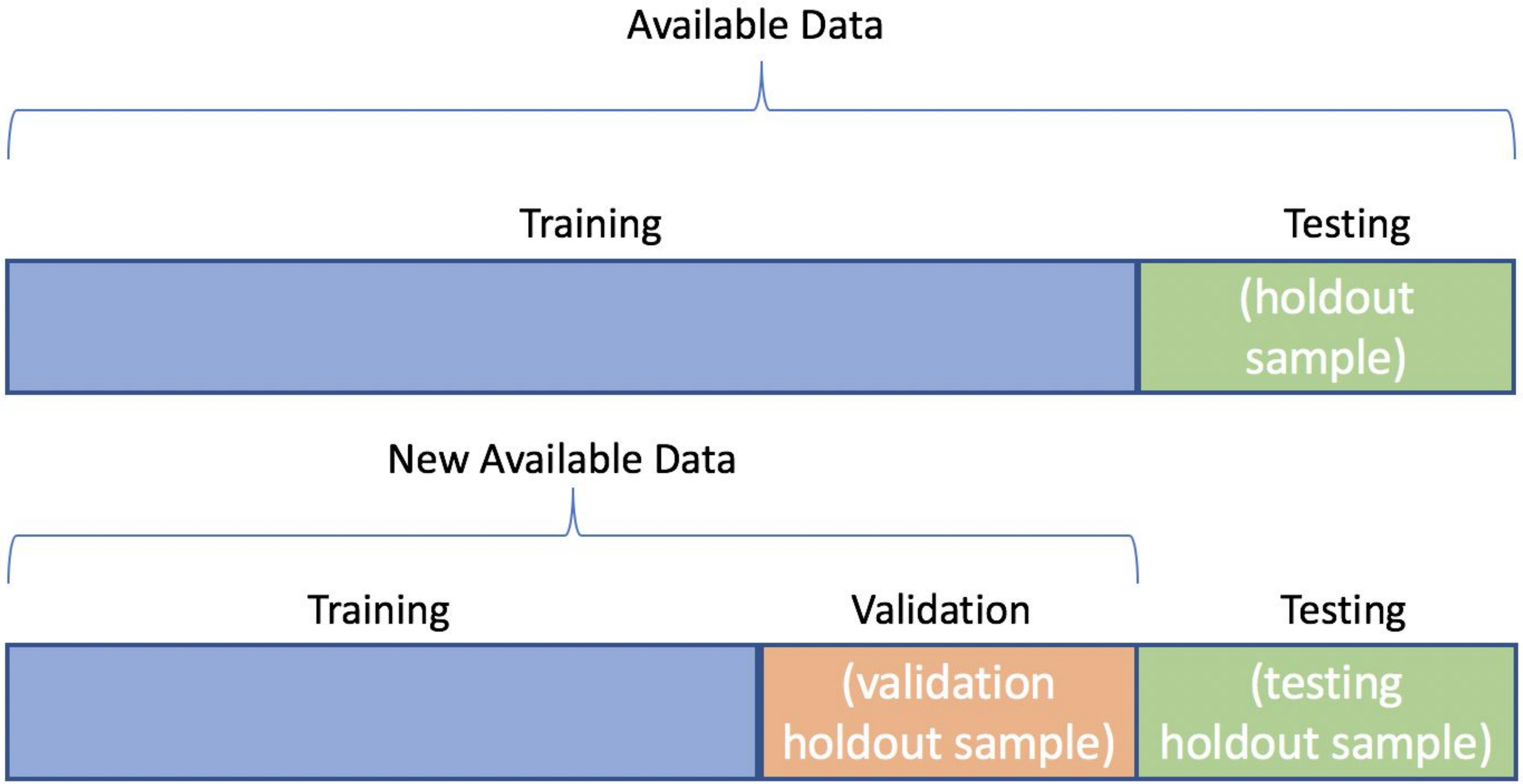


Image from datacamp.com

Train / Validation / Test

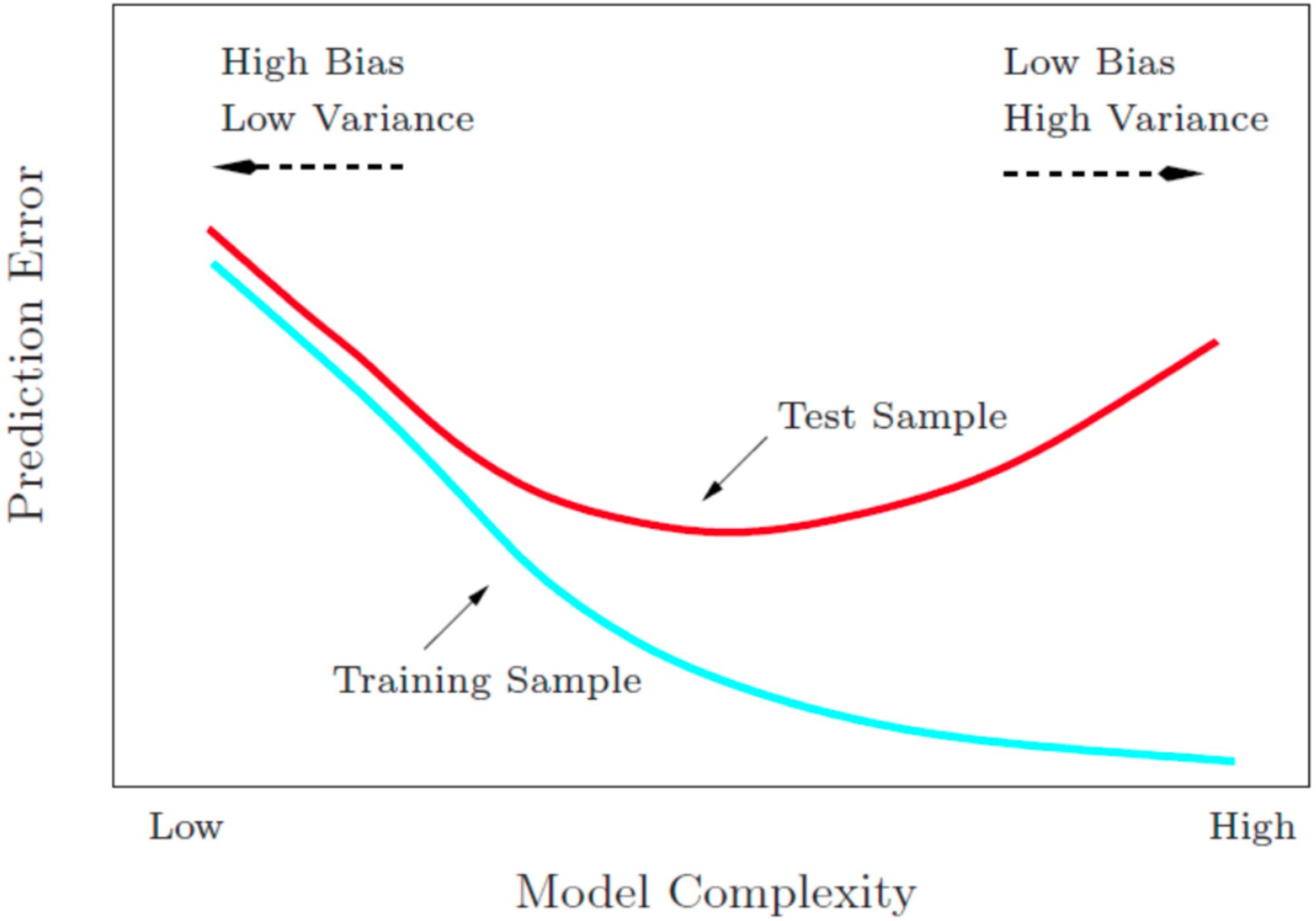
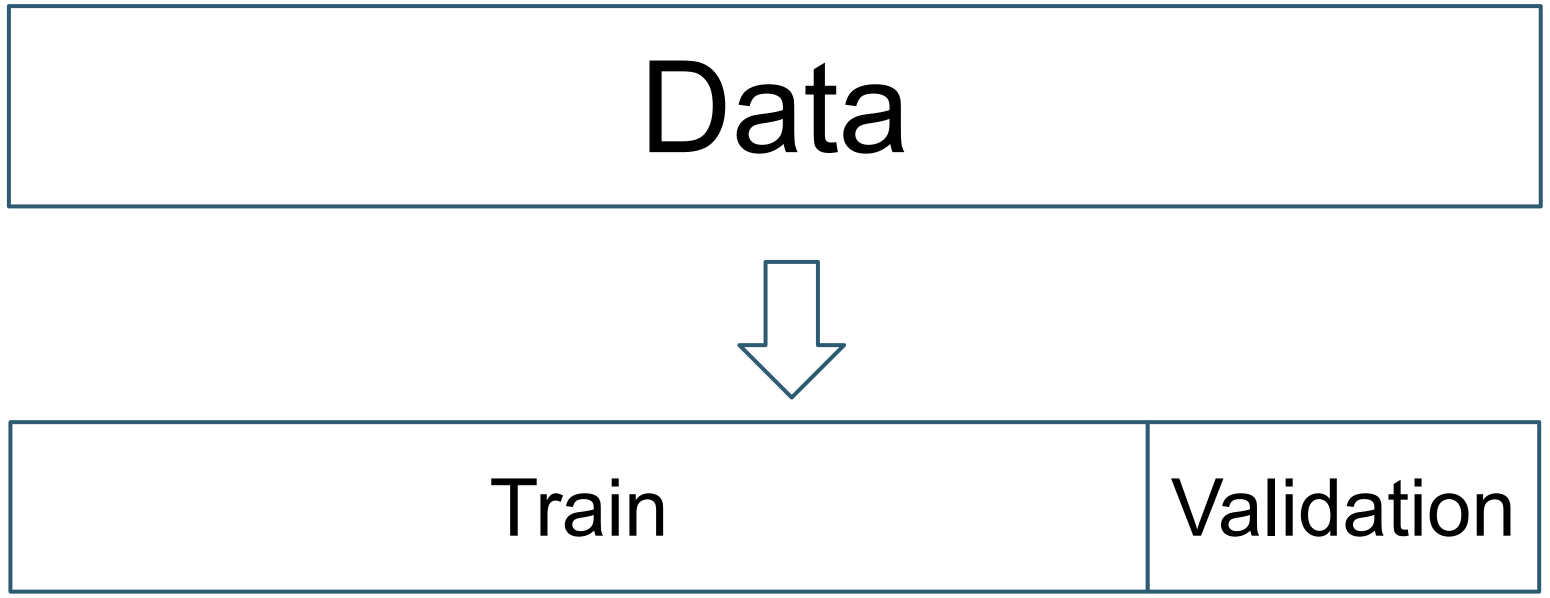


Image from datacamp.com

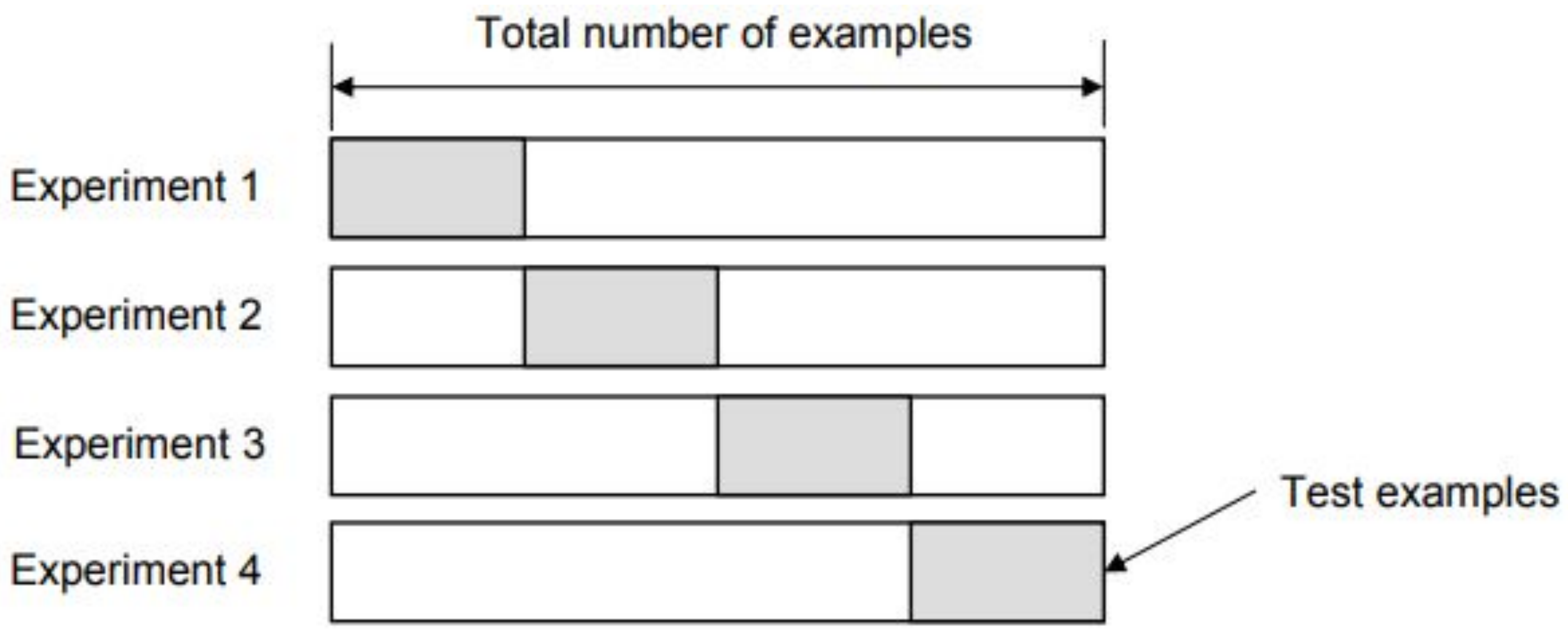
Holdout: The data is split into two different datasets labeled as a training and a validation dataset.

- This can be a *60/40* or *70/30* or *80/20* split.
- Stratification can be used to balance the sets.



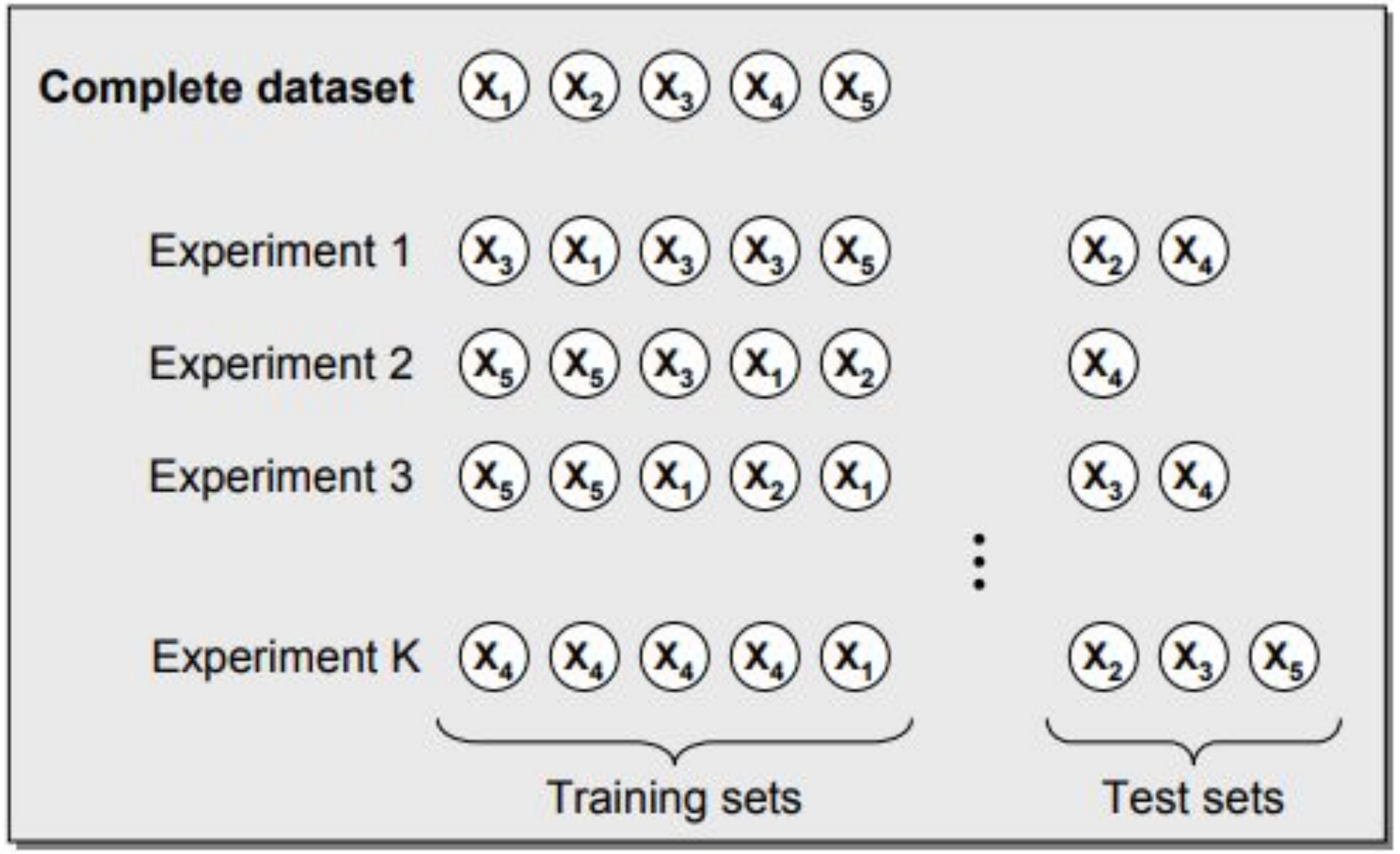
K-Fold Cross-Validation

- k-1 folds are used for training and the remaining one is used for testing.
- The error rate of the model is average of the error rate of each iteration.



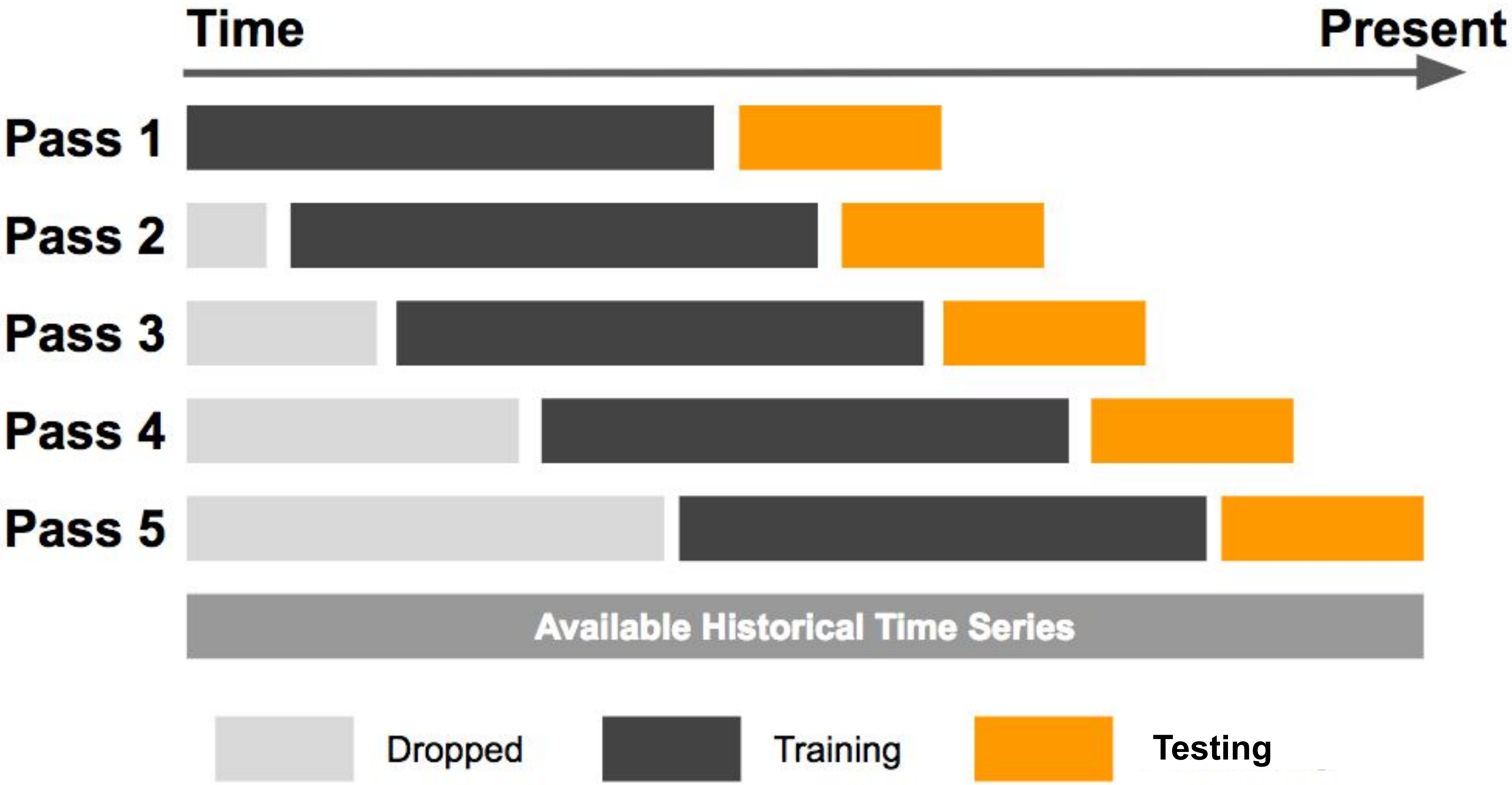
Bootstrapping

- The training dataset is randomly selected with replacement.
- The not used data is used for the test set.
- The error rate of the model is average of the error rate of each iteration.



Rolling Cross-Validation

Rolling Cross-Validation

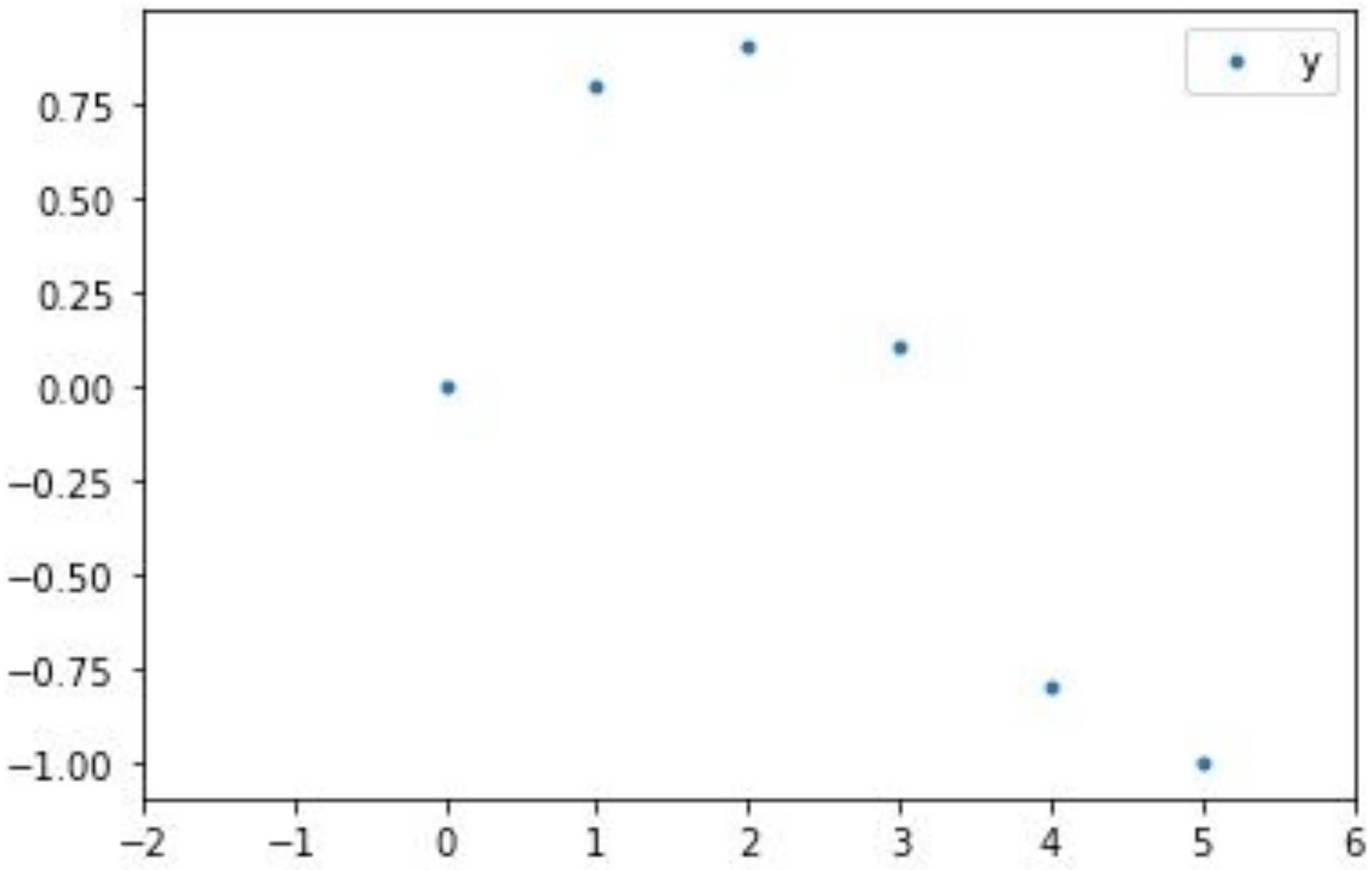


Regularization is the process of adding information in order to prevent overfitting.

- A technique to improve the generalizability of a learned model.

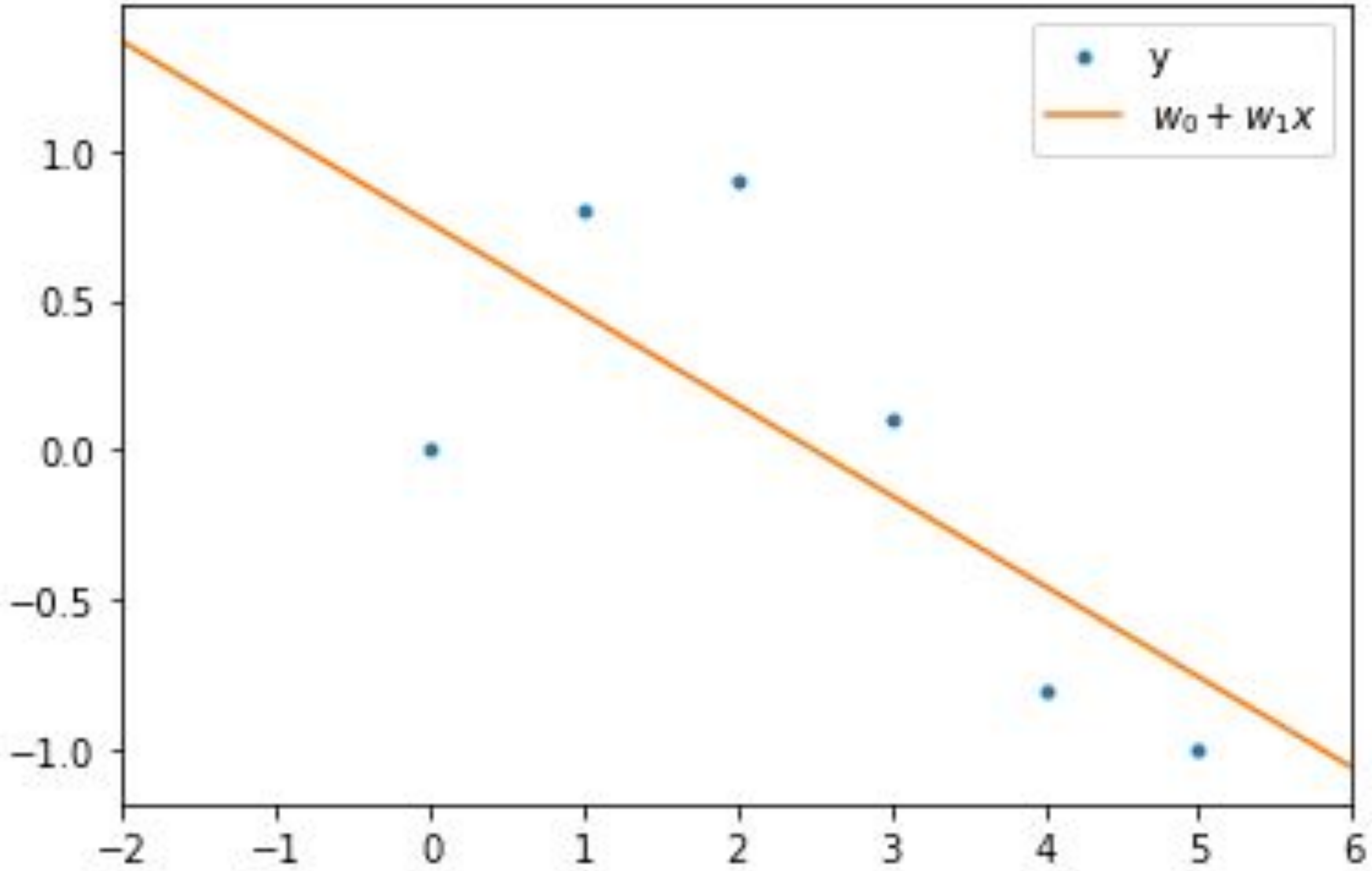
The process of learning

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \ldots\right)^2$$



The process of learning

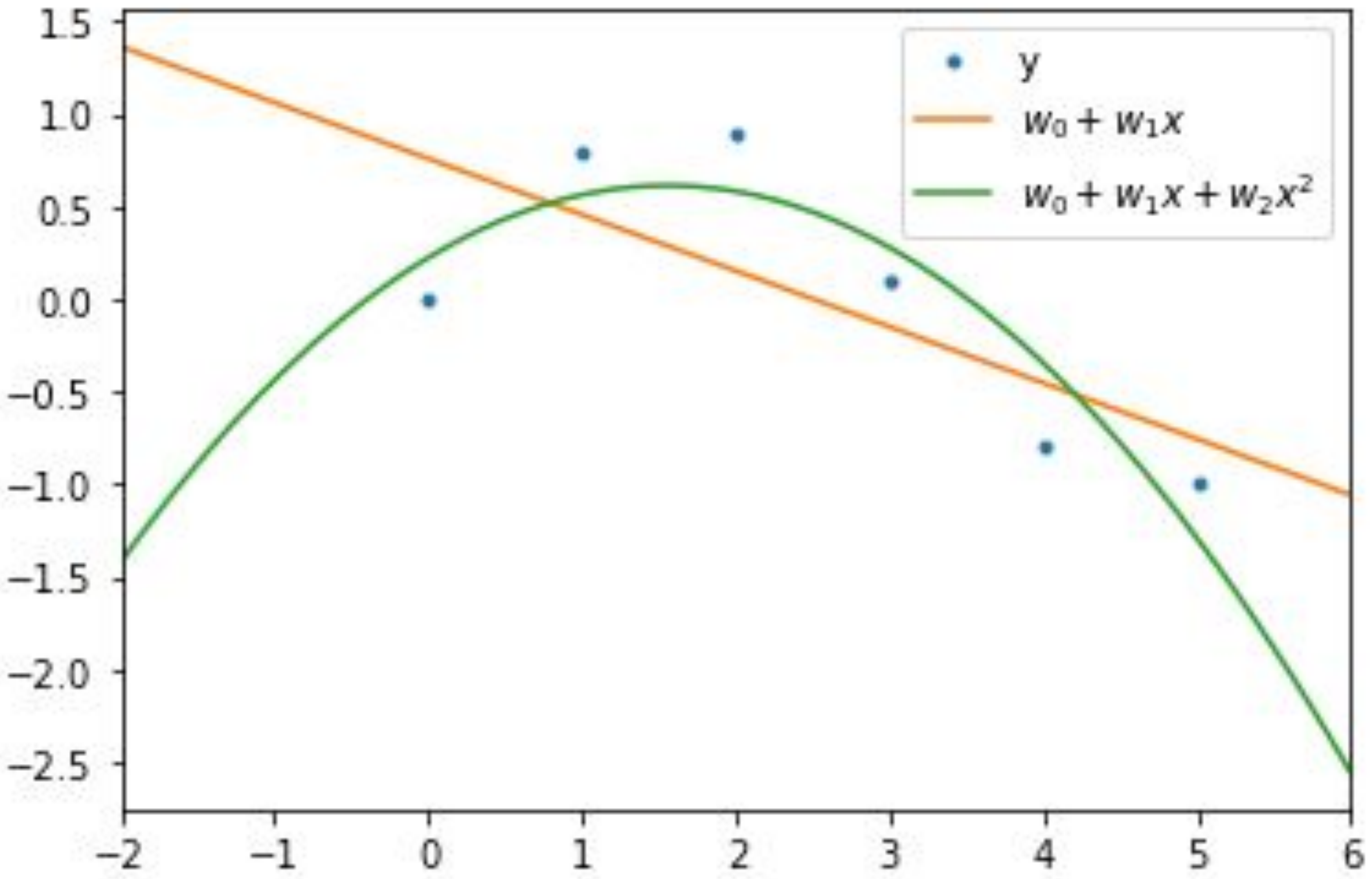
$$S = \sum_{i=0}^{N} (y_i - (w_0 + w_1 x_i))^2$$



The process of learning

$$S = \sum_{i=0}^{N} (y_i - (w_0 + w_1 x_i)^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2\right)^2$$

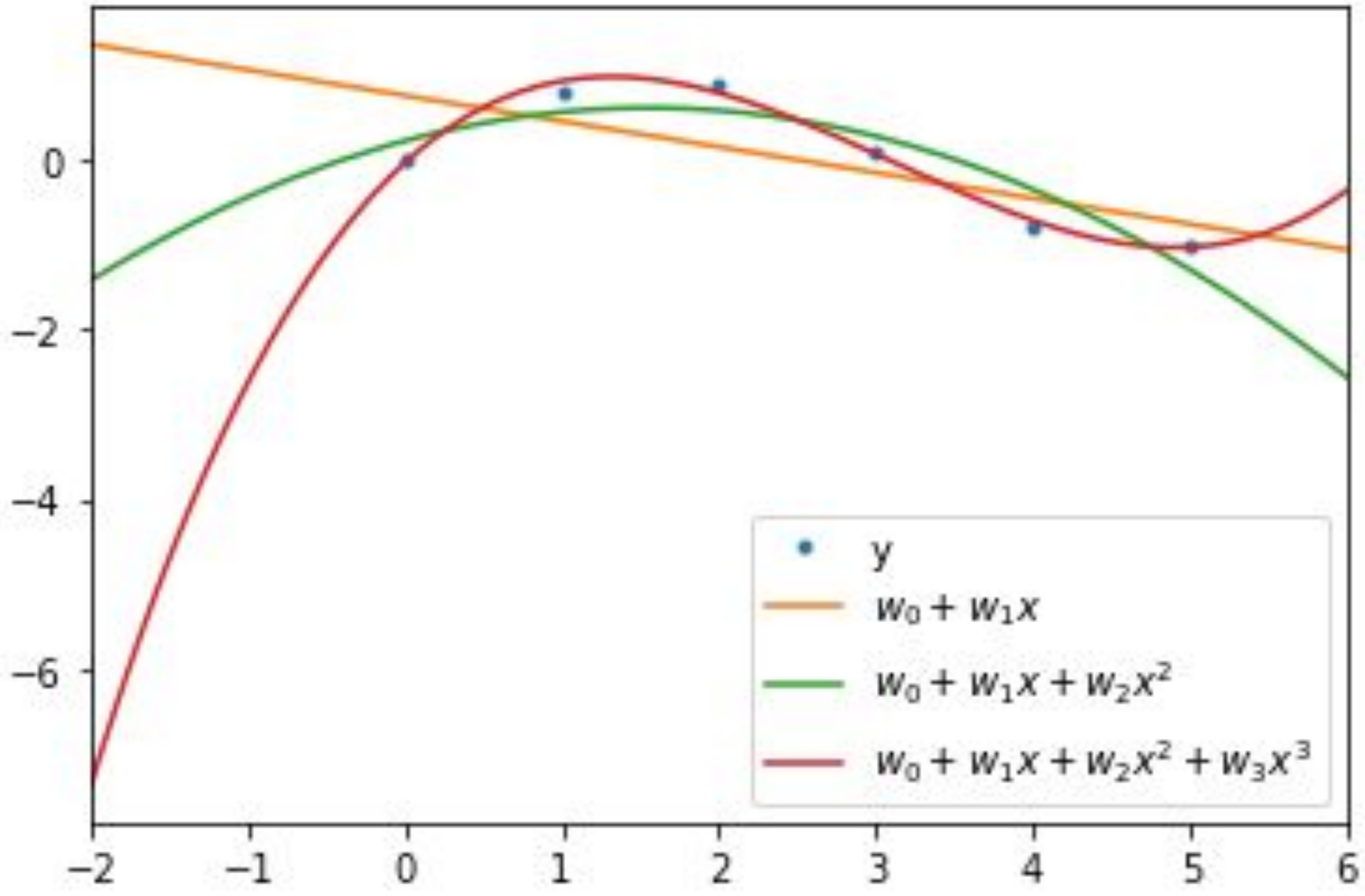


The process of learning

$$S = \sum_{i=0}^{N} (y_i - (w_0 + w_1 x_i))^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2)\right)^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3)\right)^2$$



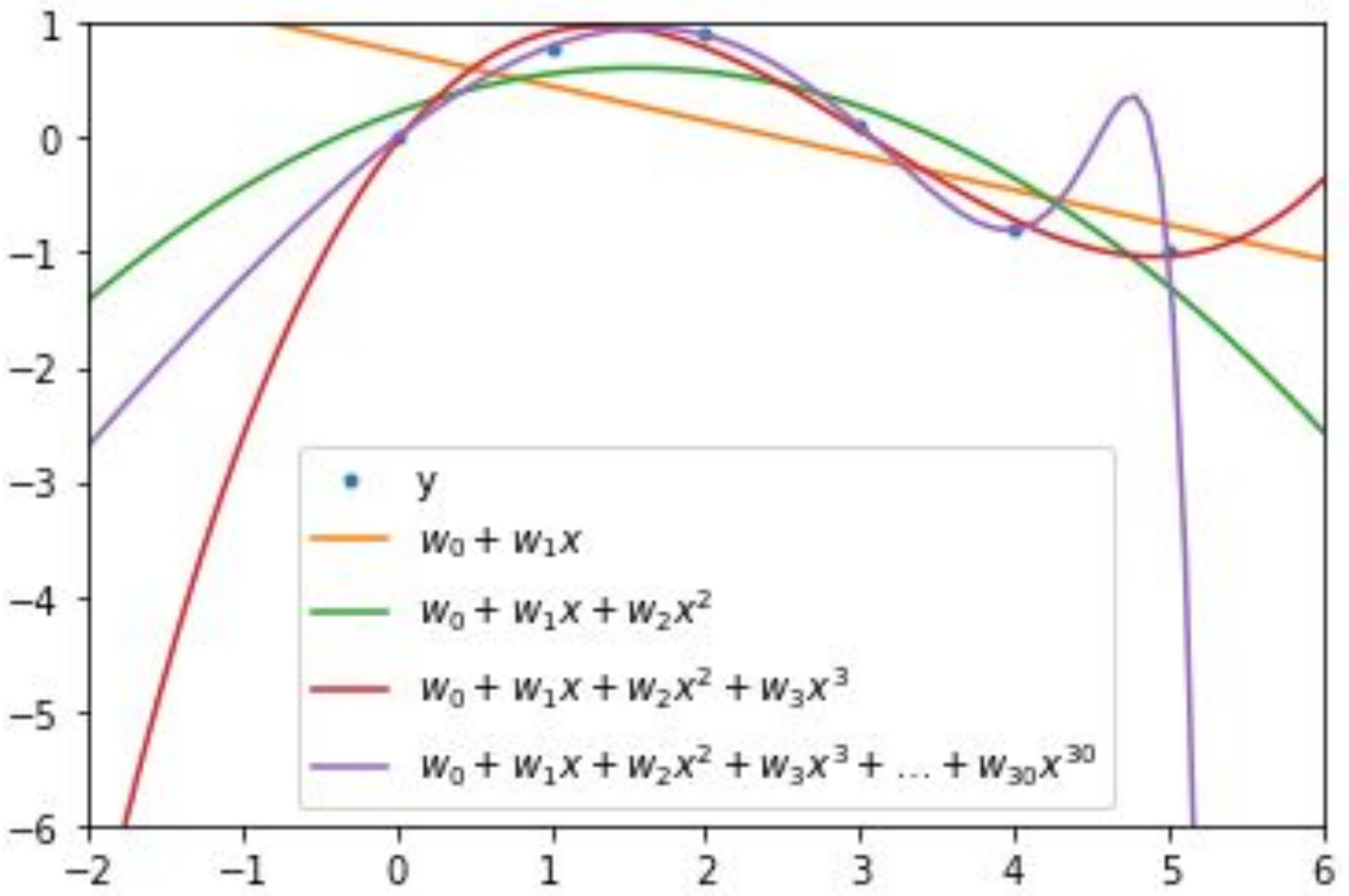
The process of learning

$$S = \sum_{i=0}^{N} (y_i - (w_0 + w_1 x_i))^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2\right)^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3\right)^2$$

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \cdots + w_{30} x_i^{30}\right)^2$$



Regularization

- L1 Regularization or **Lasso Regularization**

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \cdots + w_p x_i^p\right)^2 + \lambda \sum_{j=1}^{p} |w_j|$$

- L2 Regularization or **Ridge Regularization**

$$S = \sum_{i=0}^{N} \left(y_i - (w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 + \cdots + w_p x_i^p\right)^2 + \lambda \sum_{j=1}^{p} w_j^2$$

Comparison

L1 Regularization	L2 Regularization
Computational inefficient	Analytical Solution
Sparse outputs	Non-sparse outputs
Built-in Feature selection	No Feature selection

Dimensionality Reduction is the process of reducing the number of features under consideration by obtaining a set of principal variables

- **Feature selection**: try to find a subset of the input variables
  - the filter strategy
  - the wrapper strategy (e.g. search guided by accuracy)
  - embedded strategy (L1 can be used for that).
- **Feature projection:** transforms the data in the high-dimensional space to a space of fewer dimensions.
  - PCA
  - Autoencoders
  - NMF

Recap III - Pearson Correlation

$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

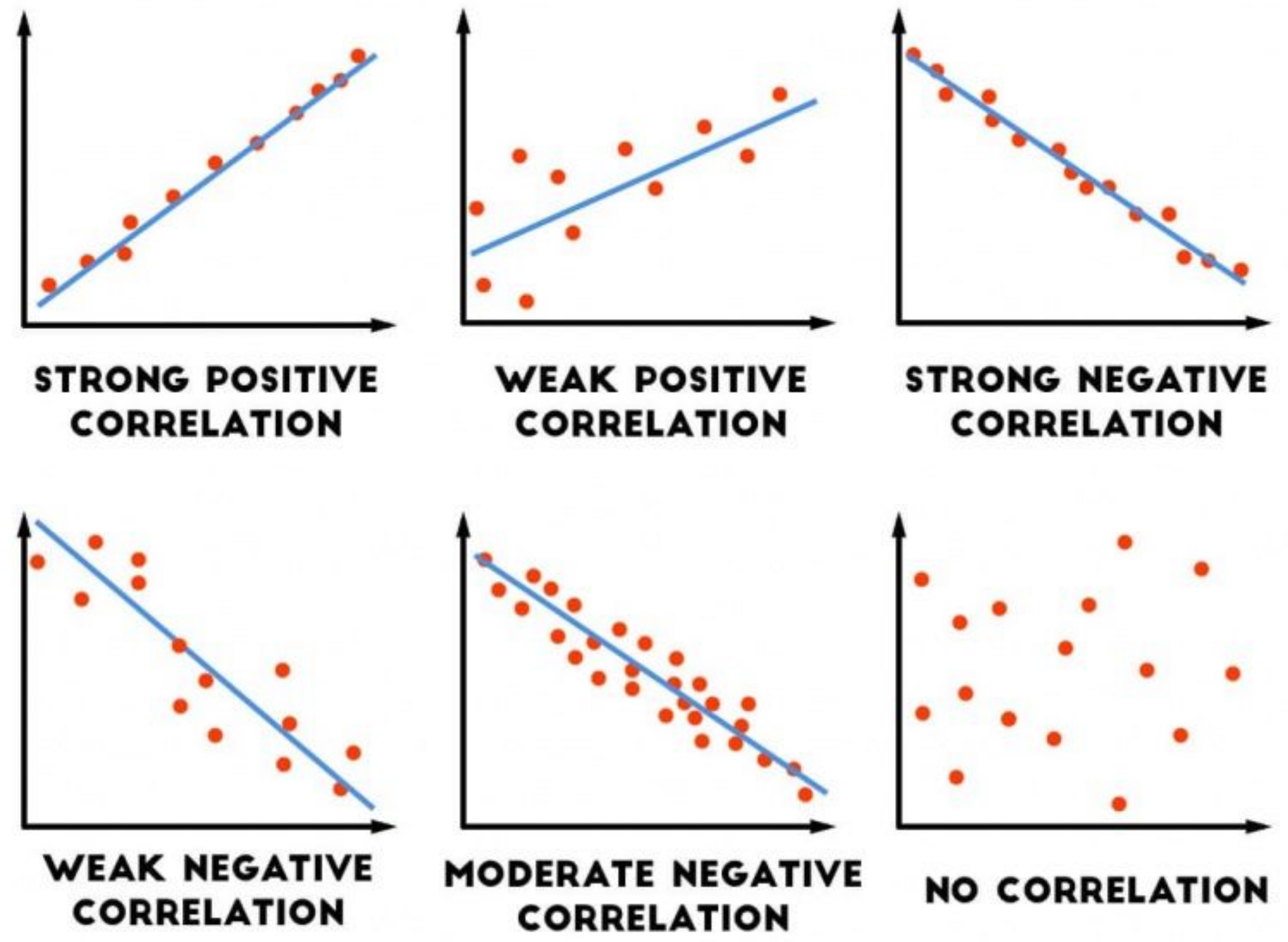


Image from Andres Robotics and Science

What about Cov(x,y,z)?

$$\mathrm{Cov}(x,y) = \mathrm{Cov}(y,x)$$
$$\mathrm{Cov}(x,z) = \mathrm{Cov}(z,x)$$
$$\mathrm{Cov}(z,y) = \mathrm{Cov}(y,z)$$

$$\mathrm{Cov}(x,y) = \frac{\Sigma(x_i-\sigma_x)(y_i-\sigma_y)}{N}$$

$$\mathrm{Cov}(x,z) = \frac{\Sigma(x_i-\sigma_z)(z_i-\sigma_z)}{N}$$

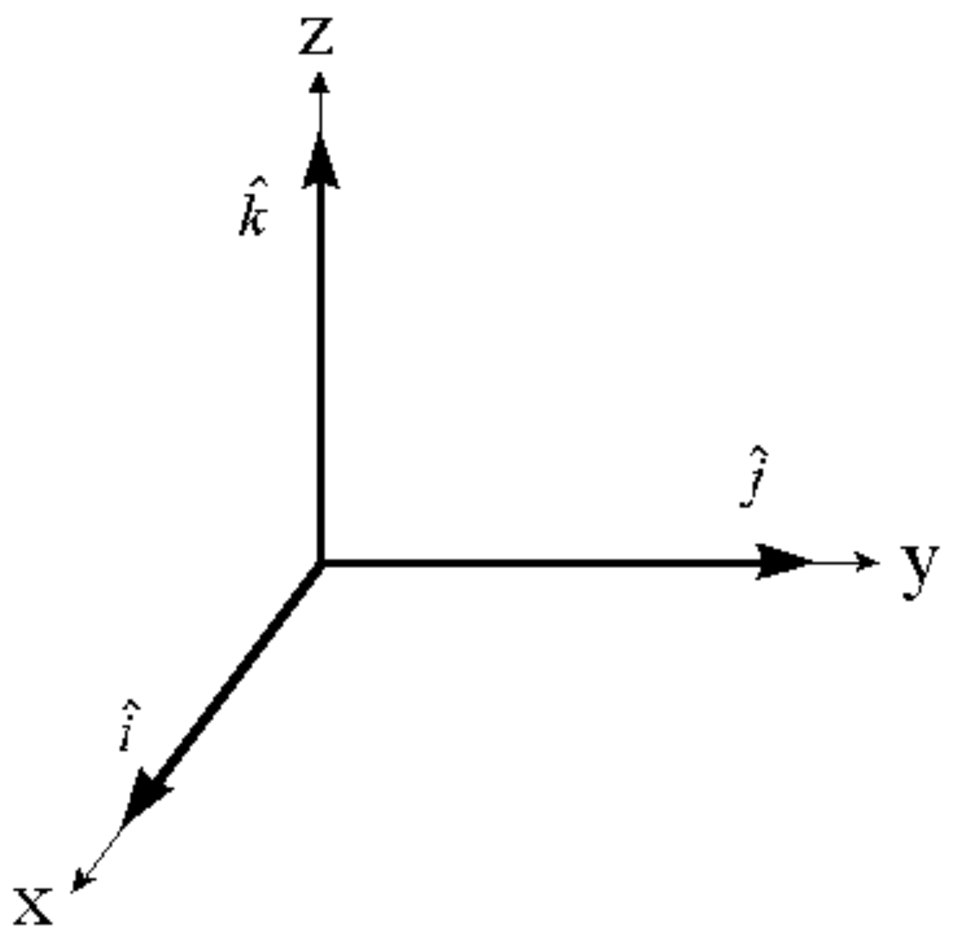
$$\mathrm{Cov}(z,y) = \frac{\Sigma(z_i-\sigma_z)(y_i-\sigma_y)}{N}$$

What is x, y and z?

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

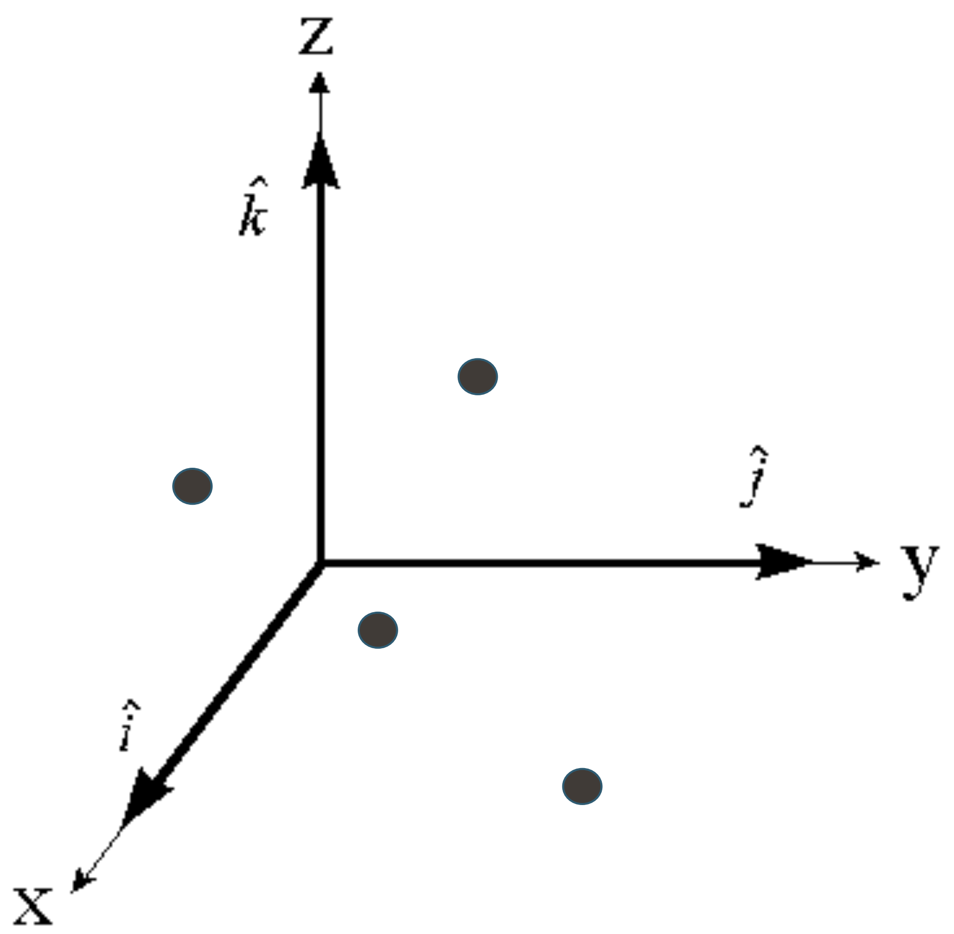
$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}$$



X, Y, Z as a Matrix

$$M_{n,3} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \cdots & \cdots & \cdots \\ x_n & y_n & z_n \end{bmatrix}$$

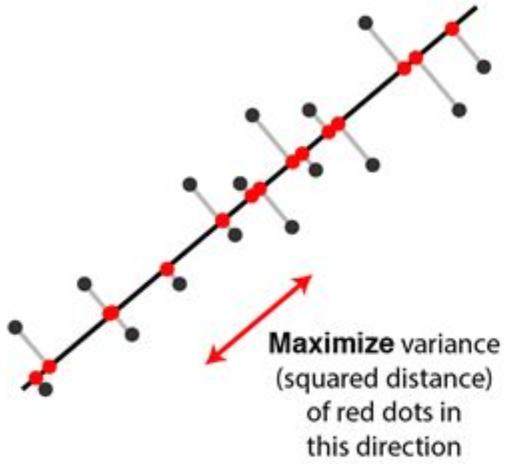


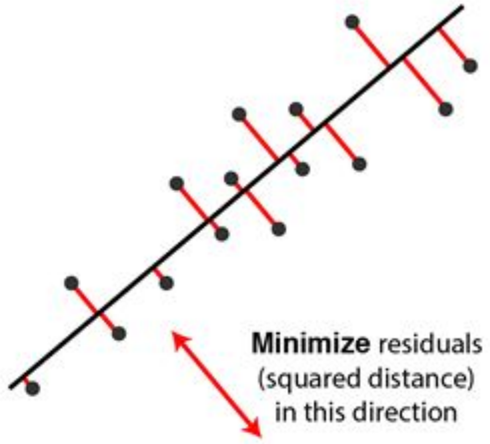
Covariance Matrix

	X	Y	Z
X	var_x	cov_{xy}	cov_{xz}
Y	cov_{xy}	var_y	cov_{yz}
Z	cov_{xz}	cov_{yz}	var_z

Example:

From 2 dimensions to 1





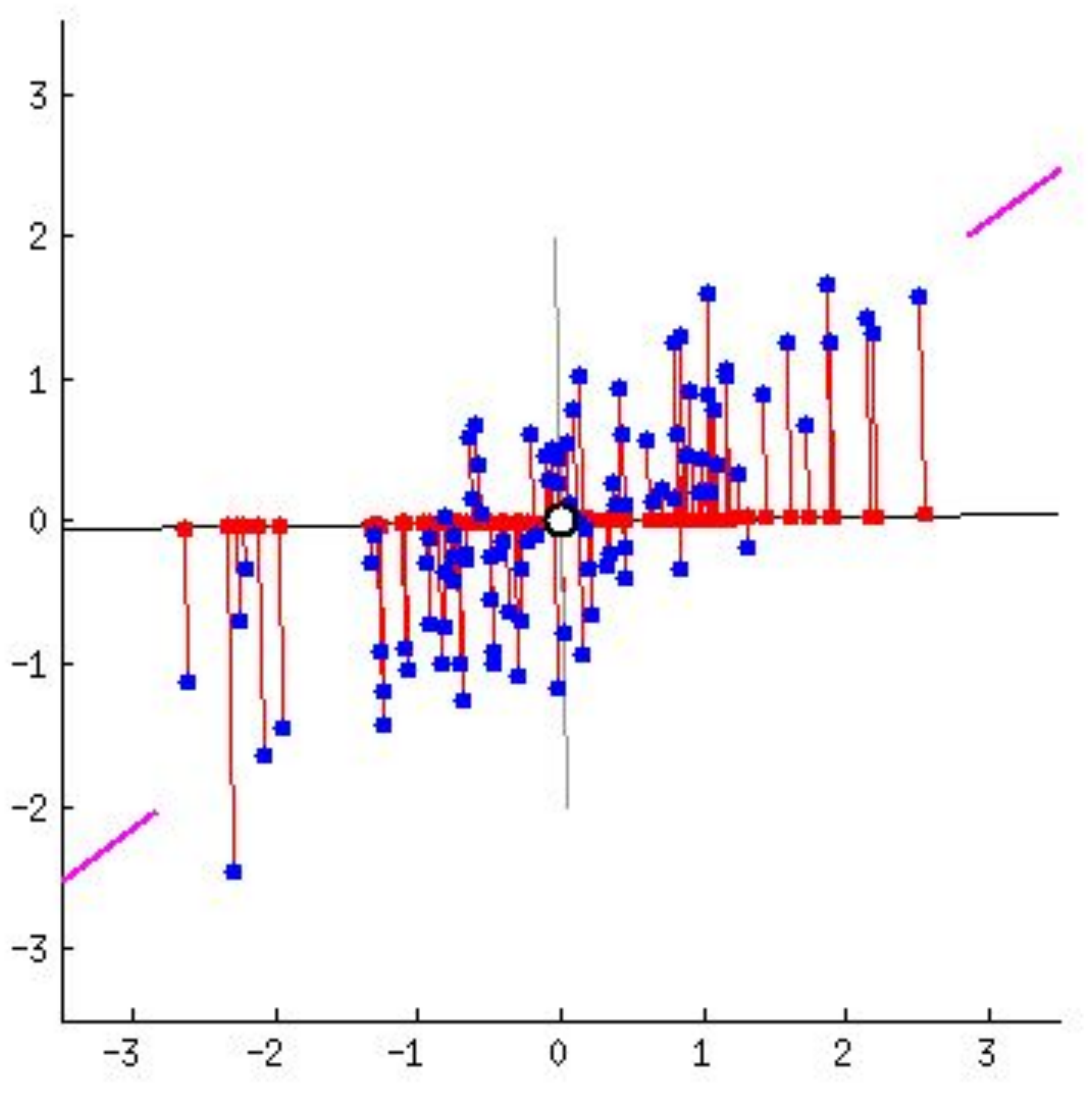


Image from [stackexchange]()

Example:

From 2 dimensions to 1

BEST LINE
Best summarizes
information

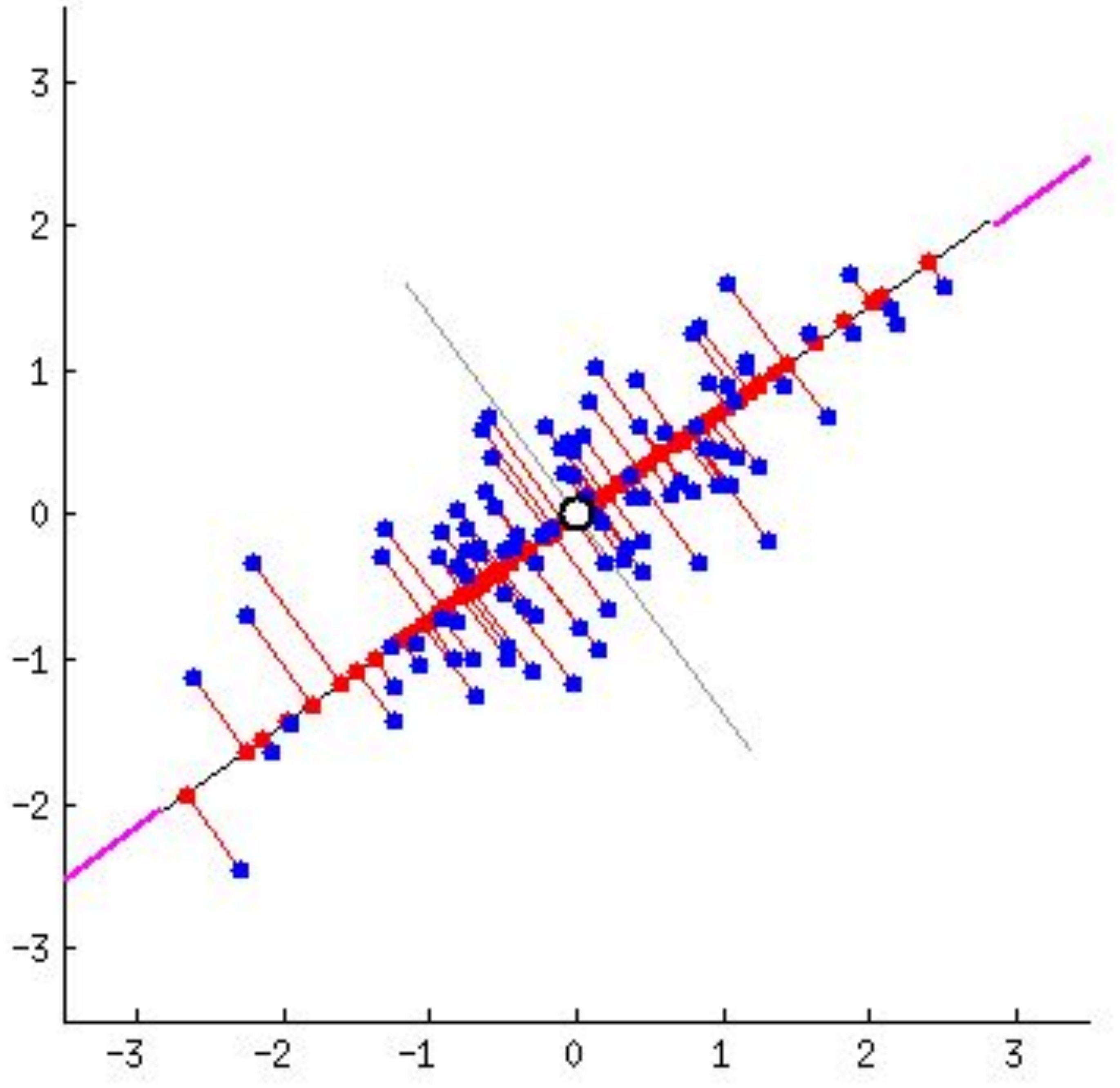
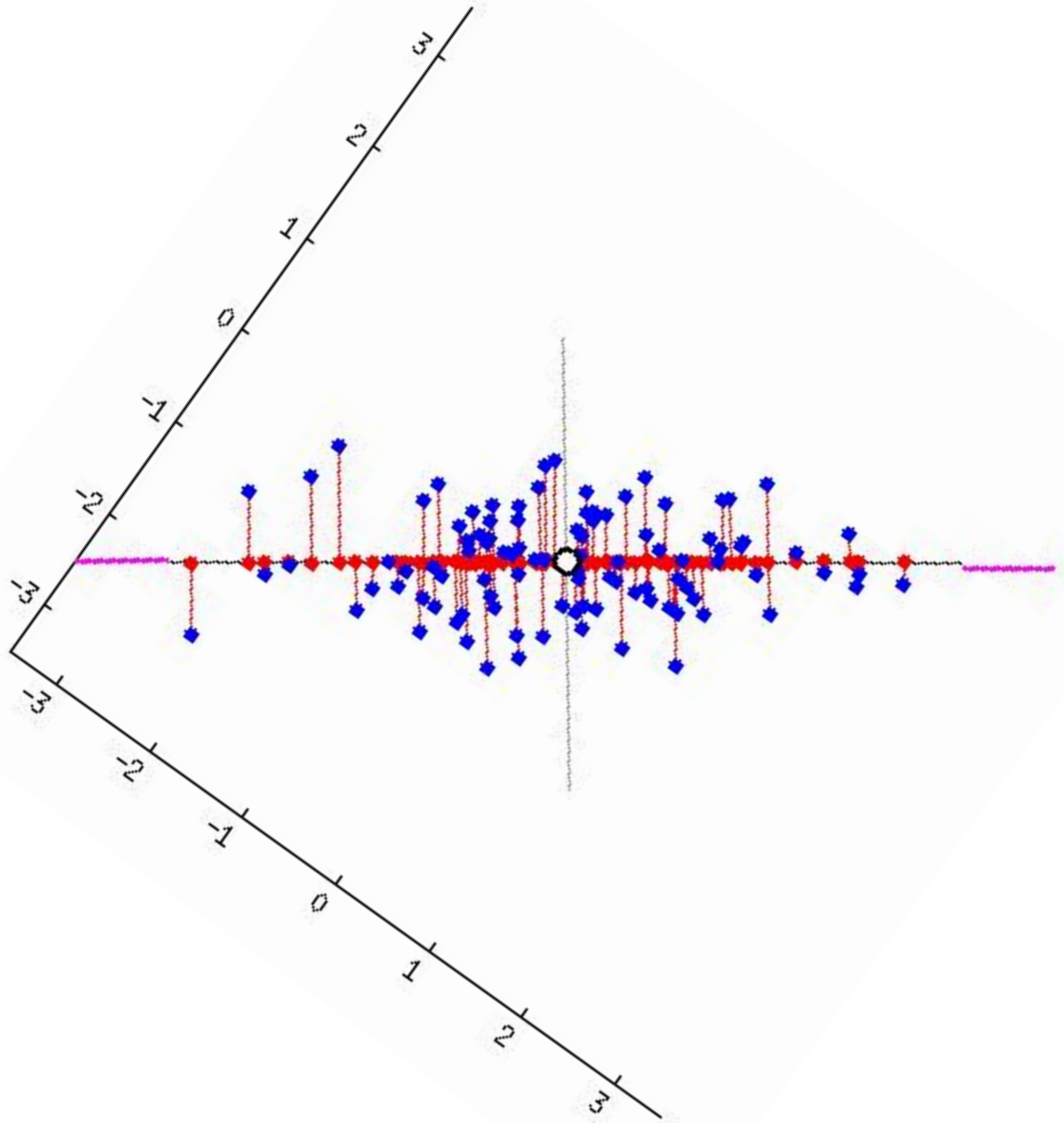


Image from stackexchange

Example:

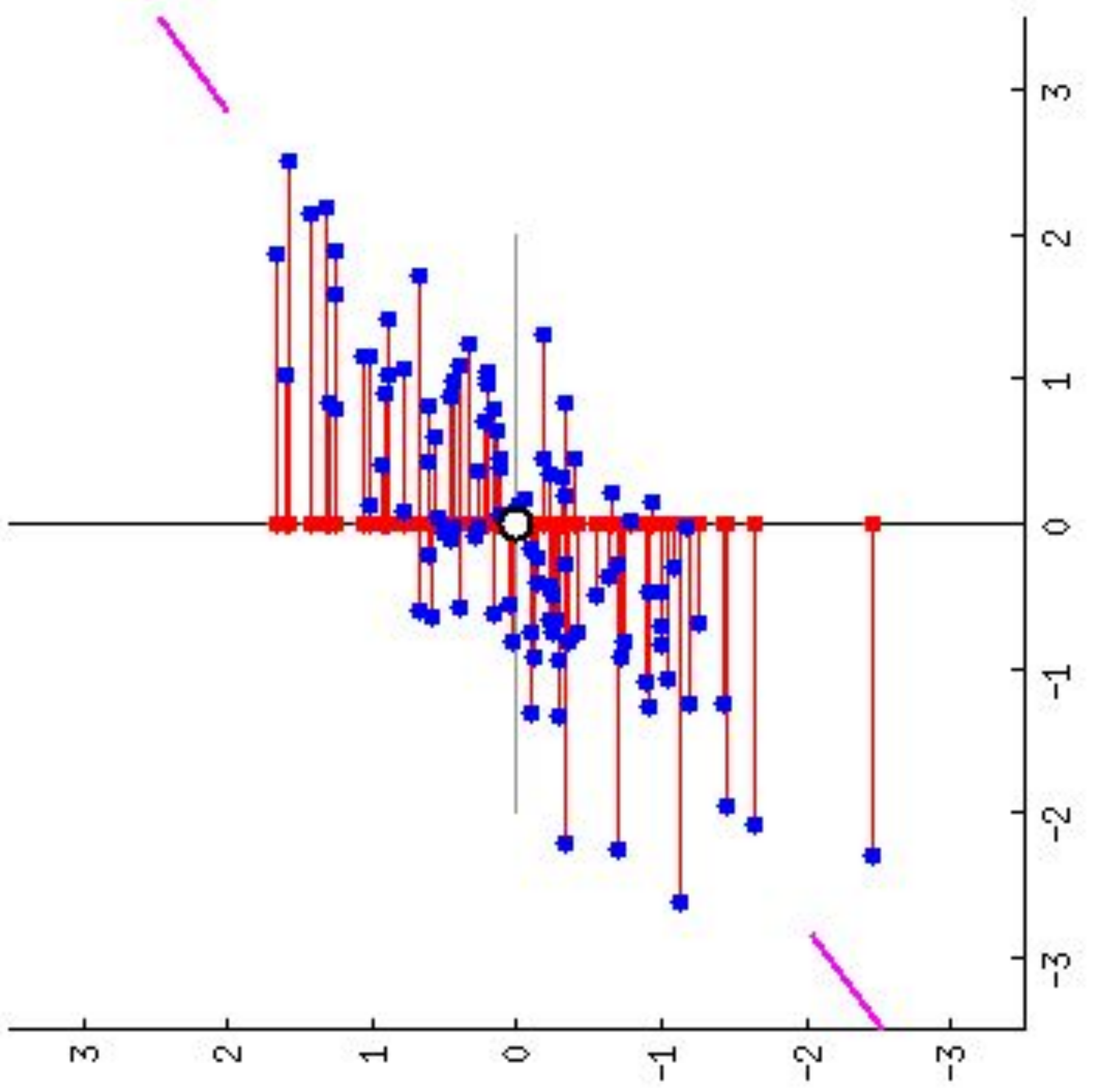
From 2 dimensions to 1

BEST LINE
Best summarizes information



Example:

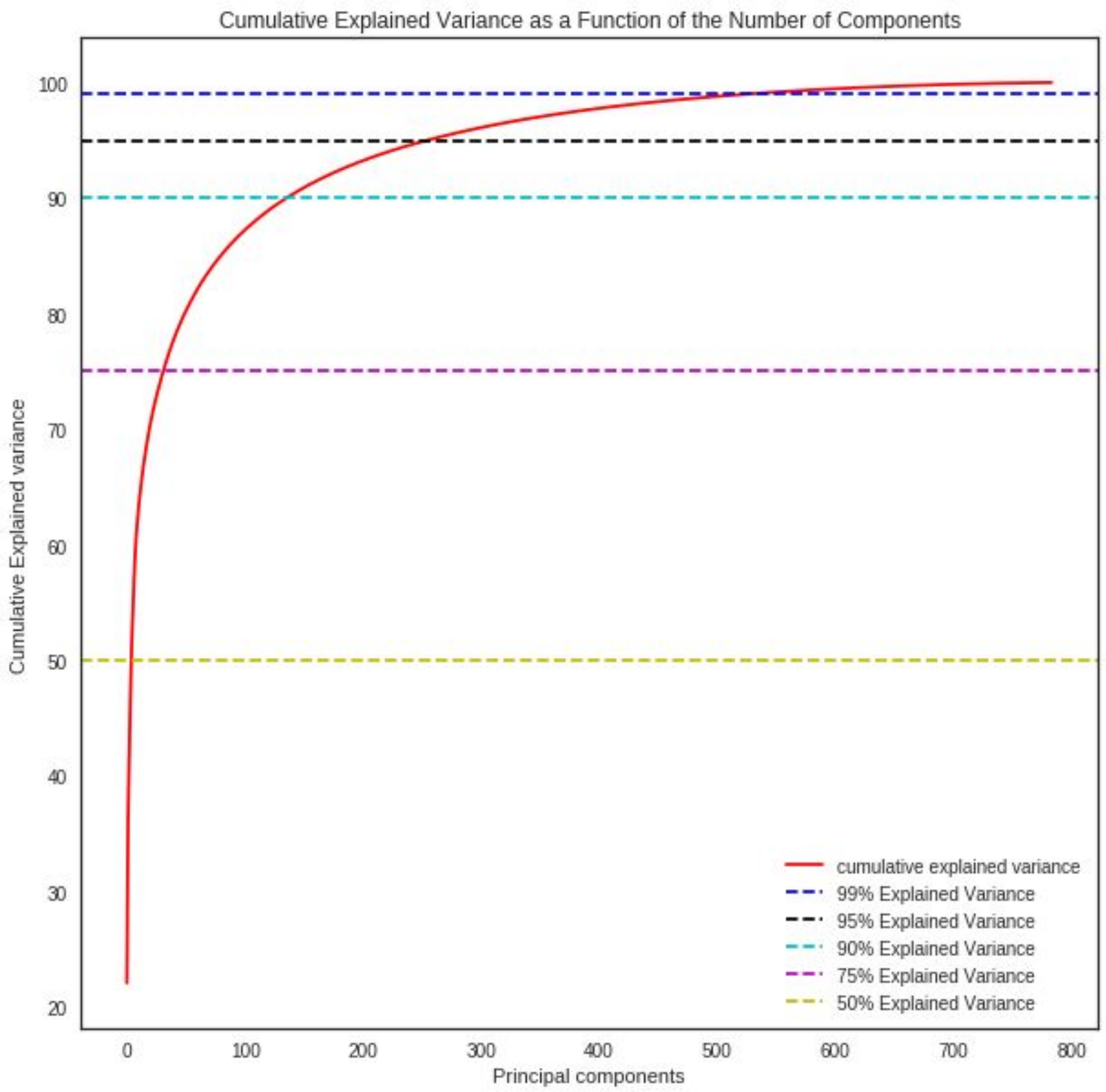
From 2 dimensions to 1



BAD LINE
Too much loss

Image from [stackexchange]

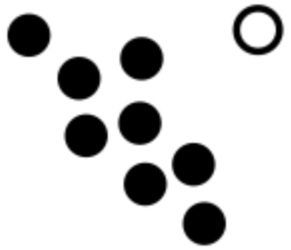
PCA - *optimal number of principal components*



Limitations

Model performance

PCA can lead to a reduction in model performance on datasets with no or low feature correlation or does not meet the assumptions of linearity.



Outliers

PCA is also affected by outliers, and normalization of the data needs to be an essential component of any workflow.

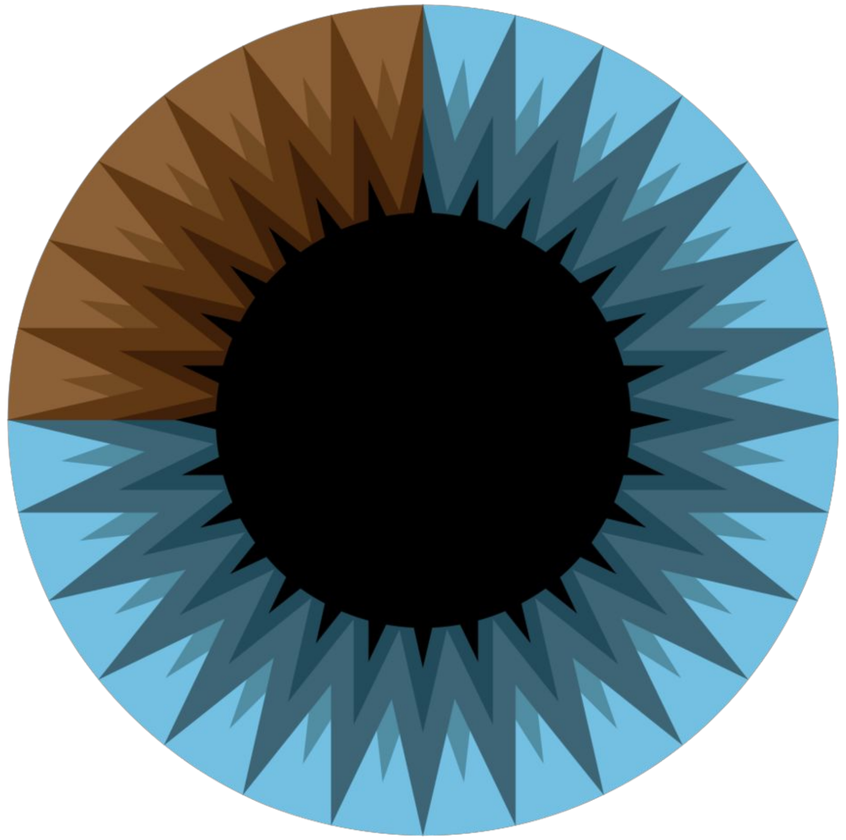
Interpretability

Each principal component is a combination of original features and does not allow for the individual feature importance to be recognized.

Useful Resources

Useful Resources

Stat Quest



3 Blue 1 Brown

[Odds](#)
[Odds ratio](#)

[Linear Algebra Playlist](#)

THANK YOU

Tecnologia + Conhecimento são nosso DNA.

O sucesso do cliente é o nosso sucesso.

Valorizamos gente boa que é boa gente.

#SOMOSTOTVERS

totvs.com

company/totvs

@totvs

fluig.com