

# Twelfth Session

**Alireza Moradi**

---



| [Linkedin](#) |



| [k](#)

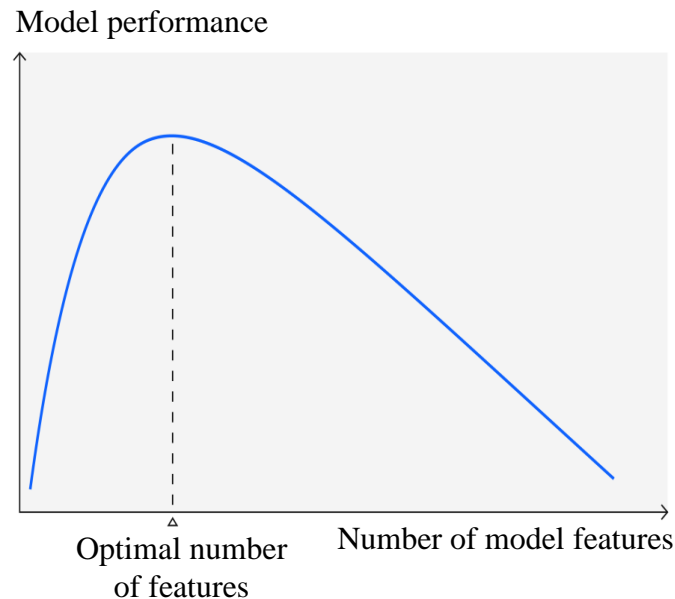
# Dimensionality reduction

---

# Dimensionality reduction

- **Dimensionality reduction** is a method for **representing a given dataset using a lower number of features (i.e., dimensions)** while still capturing the original data's meaningful properties.
- High-dimensional datasets pose several practical concerns for ML algorithms, such as **increased computation time, storage space** for big data, etc.
- But the biggest concern is perhaps decreased accuracy in predictive models. Statistical and machine learning models trained on high-dimensional datasets **often generalize poorly**.

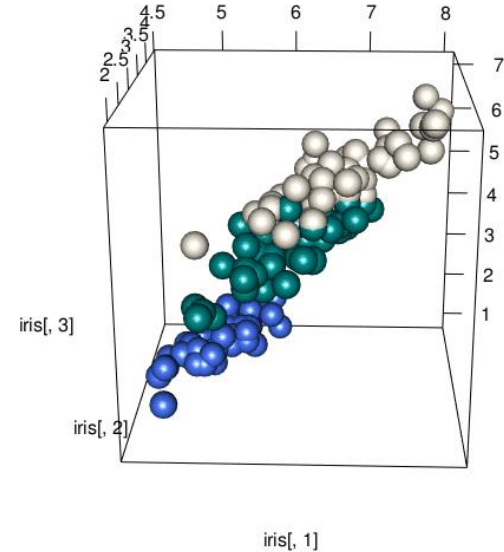
└──────────┘  
?



# Dimension reduction ; Visualization

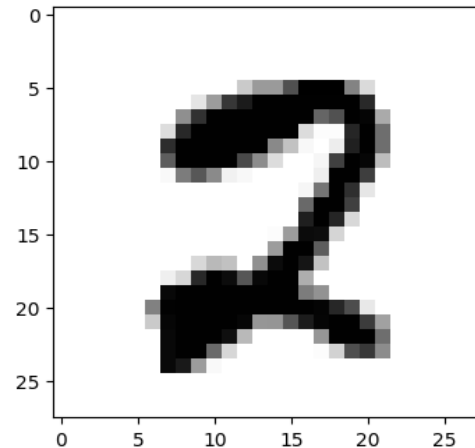
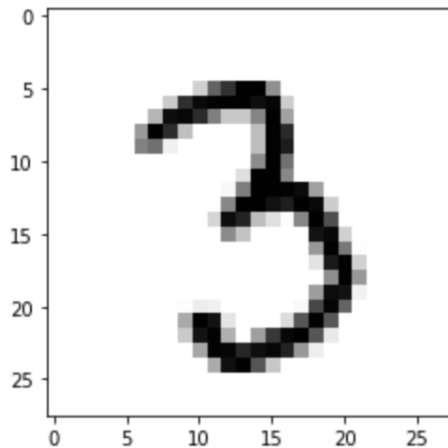
- Many data sets have dozens or even millions of dimensions. To visualize them we can do dimensionality reduction, projecting the data down to a map in **two dimensions**.
- Projecting to **three dimensions** is also possible which can then be explored interactively.

$$X_1, X_2, \dots, X_P \quad \Longrightarrow \quad X'_1, X'_2$$



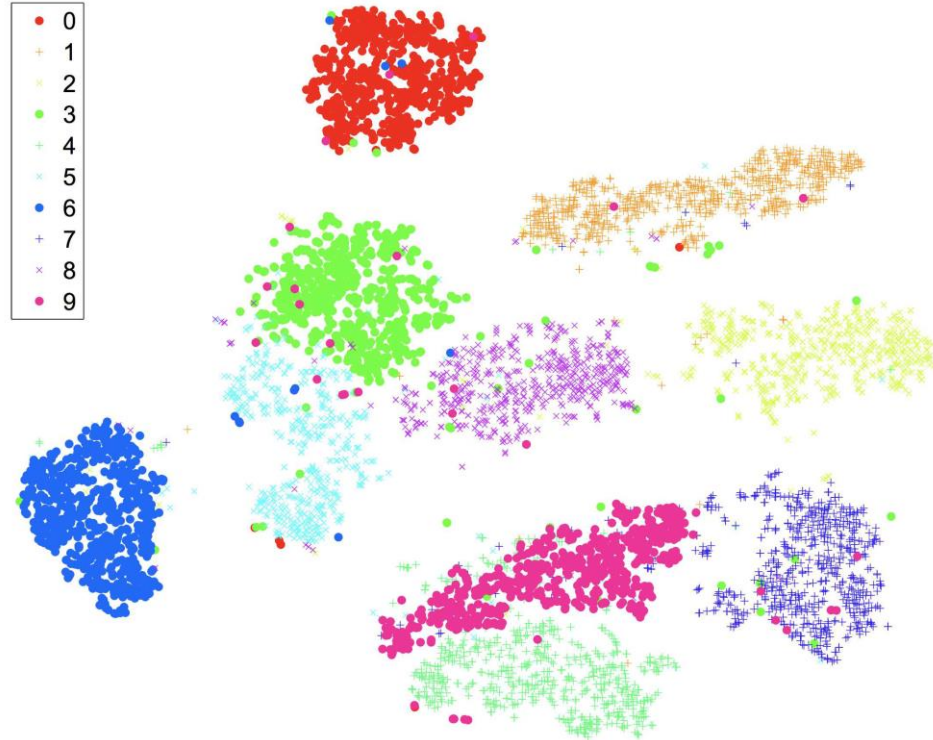
# Dimension reduction; example

- [MNIST](#) is digit recognition dataset. A collection of 60,000 images of handwritten digits, each  $28 \times 28$  pixels and thus 784 dimensions.
- The **t-SNE algorithm** finds a representation that accentuates the differences between clusters.
- t-SNE = t-distributed stochastic neighbor embedding.



# Dimension reduction; example

- A two-dimensional t-SNE map of the MNIST data set:



# Principal component analysis

---

# Principal component analysis (PCA)

- **PCA** produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance and are mutually uncorrelated.
- We aim to find the components which explain the maximum variance. This is because, we want to retain as much information as possible using these components.
- ✓ Why higher variance means more information?

StudentID	Age	Score
001	17	20
002	17	13
003	17	17



# Principal component analysis (PCA)

- The **first principal component** of a set of features  $X_1, X_2, \dots, X_p$  is the normalized linear combination of the features that has the largest variance.

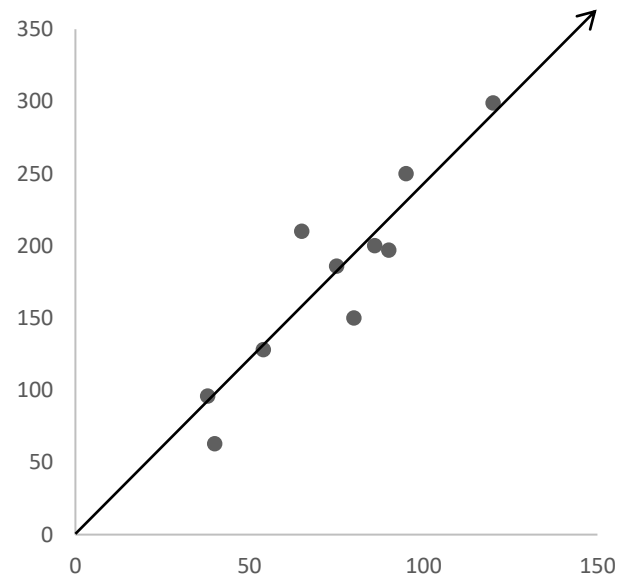
$$Z_1 = A_{11} \cdot X_1 + A_{12} X_2 + \dots + A_{1p} \cdot X_p$$

- By normalized, we mean that:

$$\sum_{j=1}^P A_{1j}^2 = 1$$

- Loadings of the first principal component:

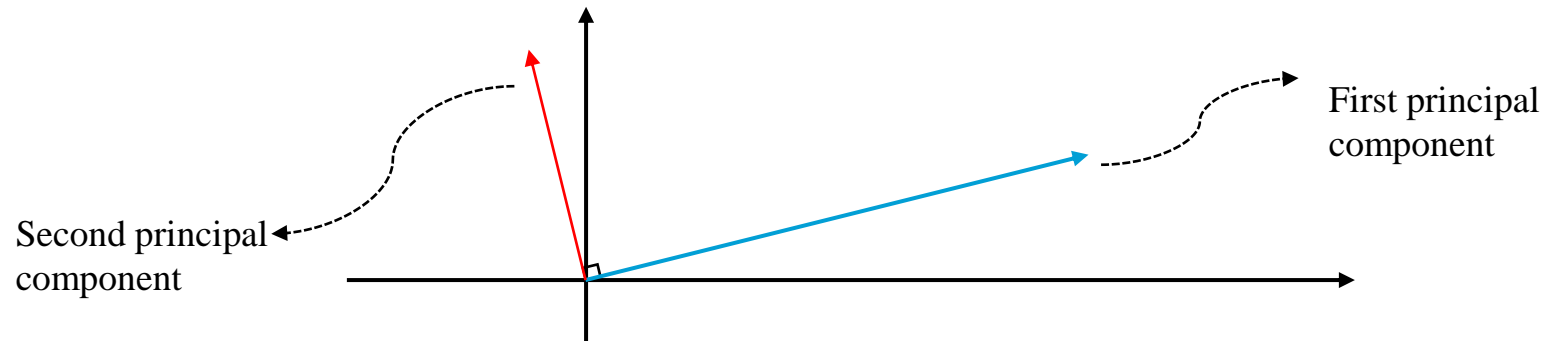
$$A_1 = [A_{11}, A_{12}, \dots, A_{1p}]$$



# Principal component analysis (PCA)

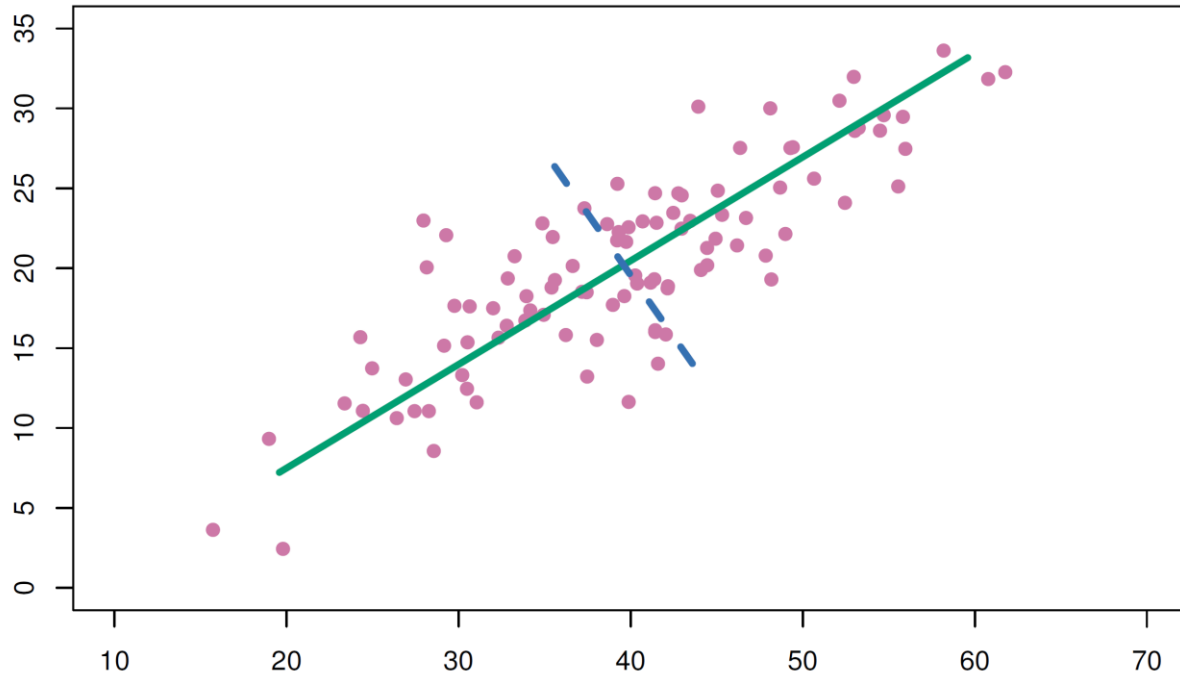
- Similarly, we can define the second principal component and all subsequent components (up to the  $j$ -th) in the same way.
- The **second principal component** is the linear combination of  $X_1, X_2, \dots, X_p$  that has maximal variance among all linear combinations that are **uncorrelated with  $Z_1$** .

$$Z_2 = A_{21} \cdot X_1 + A_{22} \cdot X_2 + \dots + A_{2p} \cdot X_p$$



# Principal component analysis (PCA)

- The **green** solid line indicates the **first principal component** direction, and the **blue** dashed line indicates the **second principal component** direction.

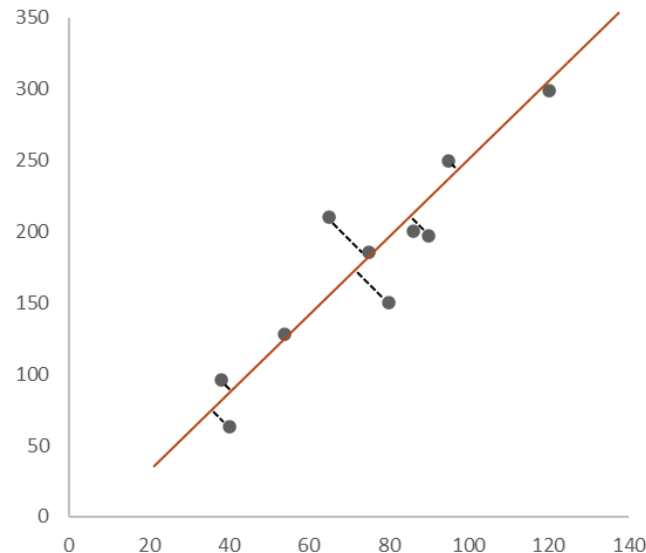


# Principal component analysis (PCA)

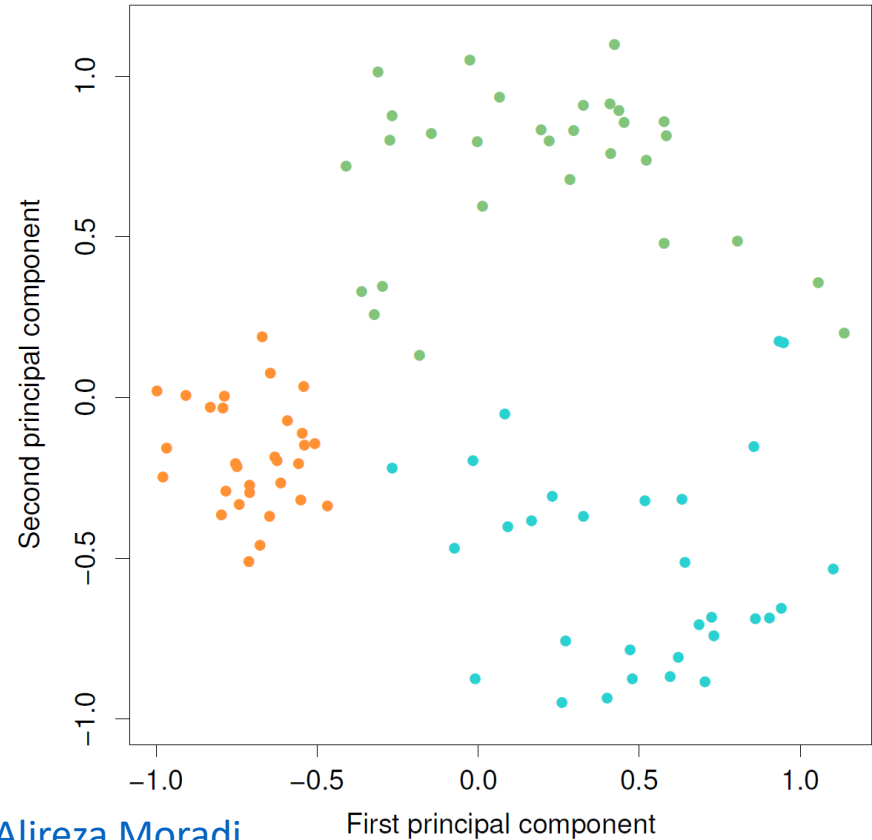
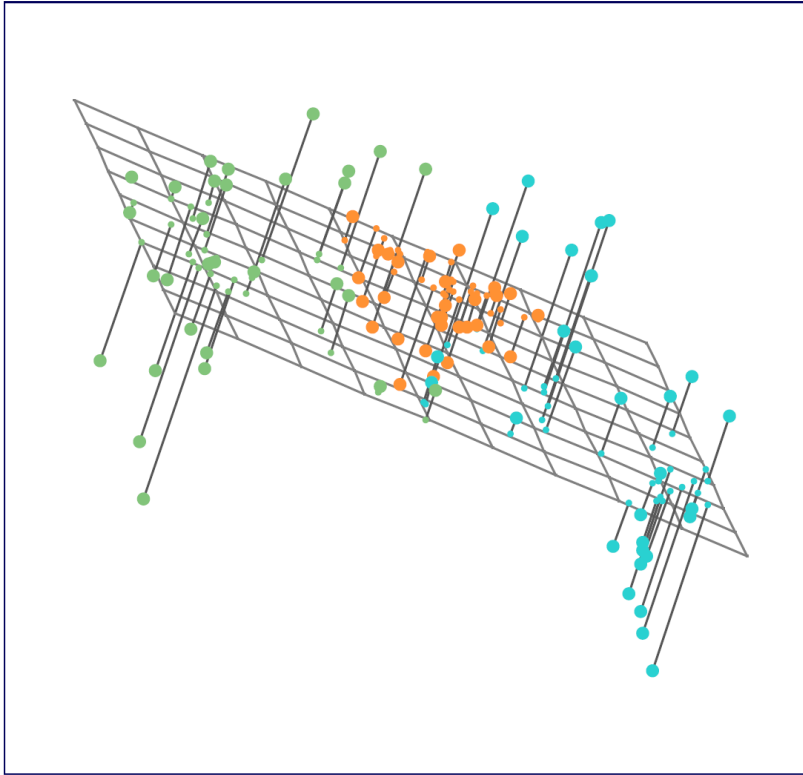
- **PCA** find **the line closest to the observations**.
- The first principal component has a very special property: it defines **the line in p-dimensional space that is closest to the n observations**.
- We are using average squared Euclidean distance as a measure of closeness:

$$d(a, b) = \sum_{j=1}^p (a_j - b_j)^2 \leftarrow \text{---}$$

- ✓ What do the first two principal components show?

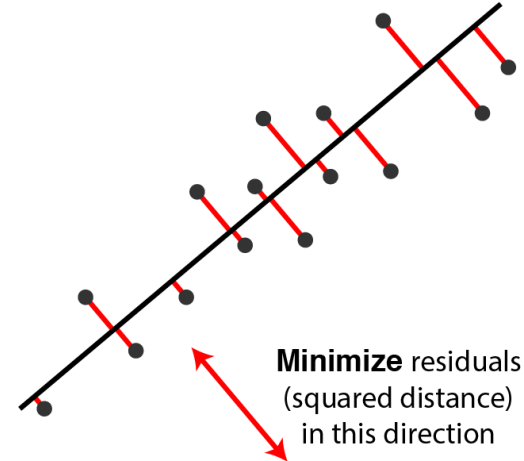
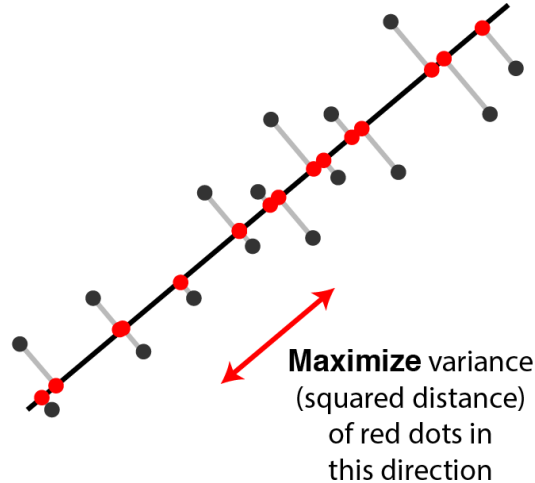


# Principal component analysis (PCA)



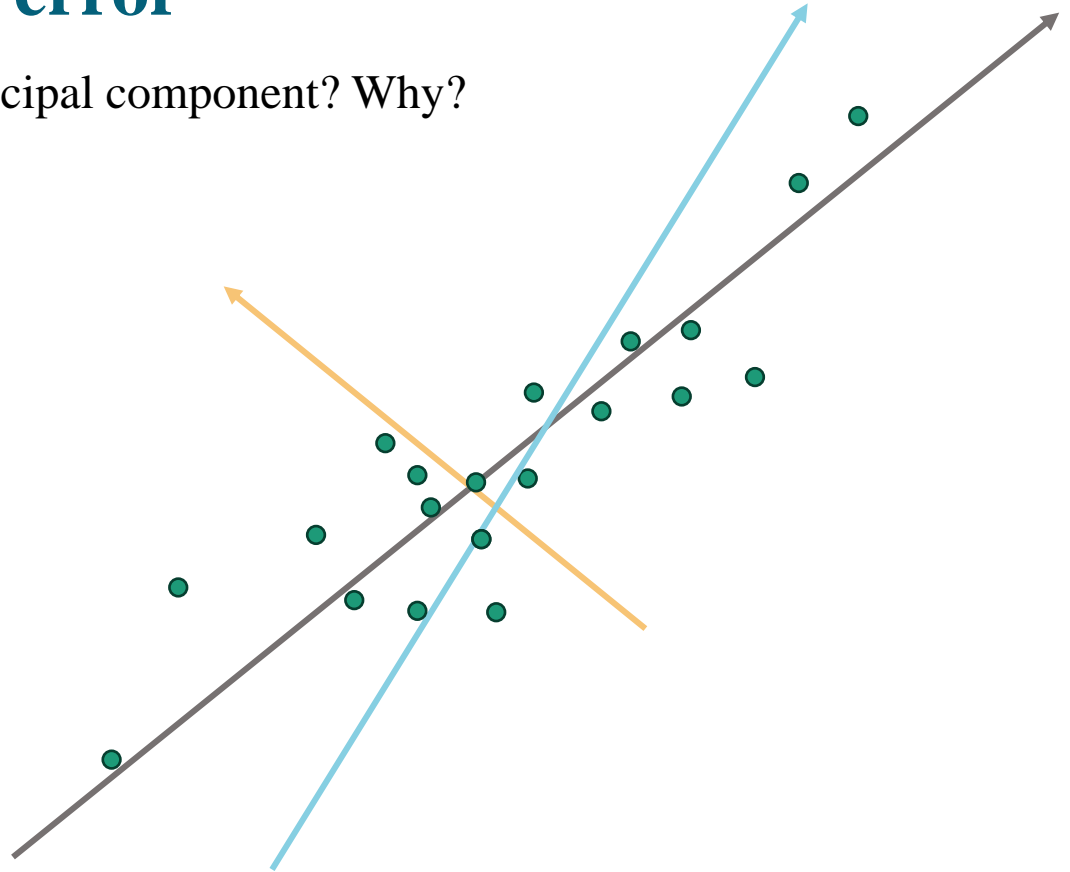
# PCA; reconstruction error

- PCA projects the data onto a subspace which **maximizes the projected variance**, or equivalently, **minimizes the reconstruction error**.



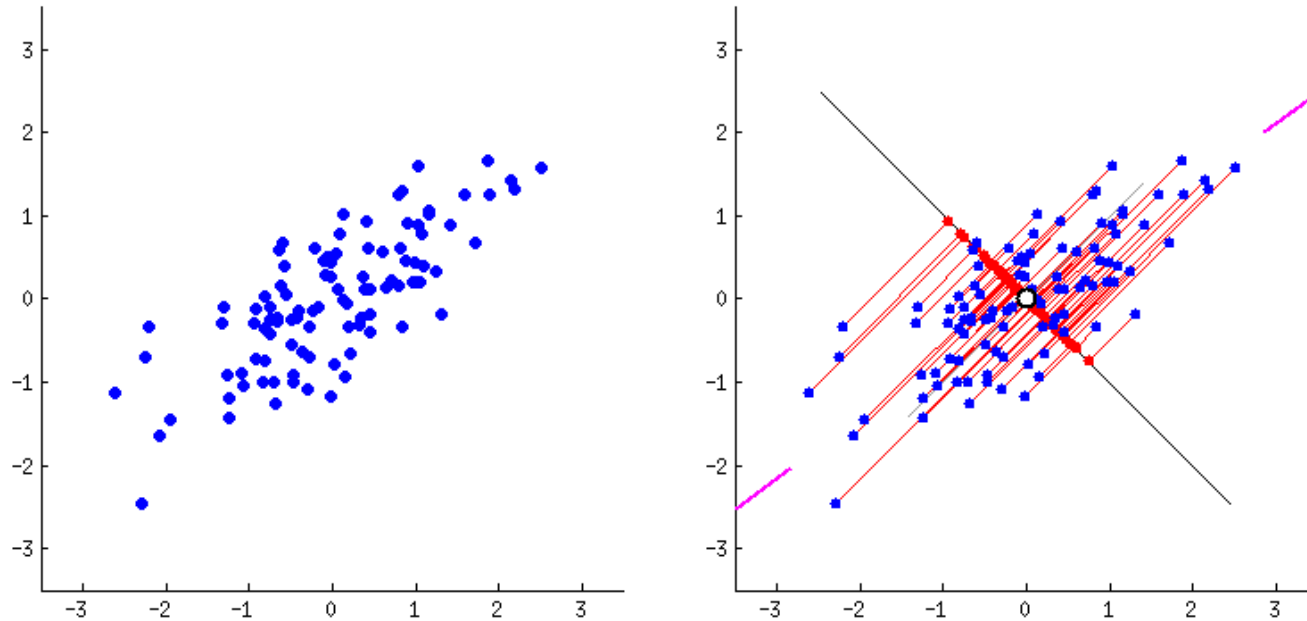
# PCA; reconstruction error

✓ Which one is better as first principal component? Why?



# PCA; reconstruction error

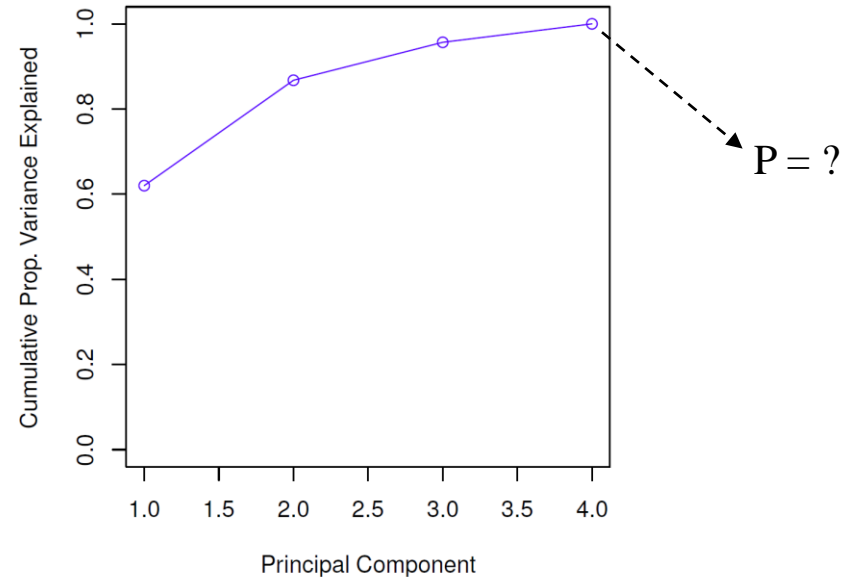
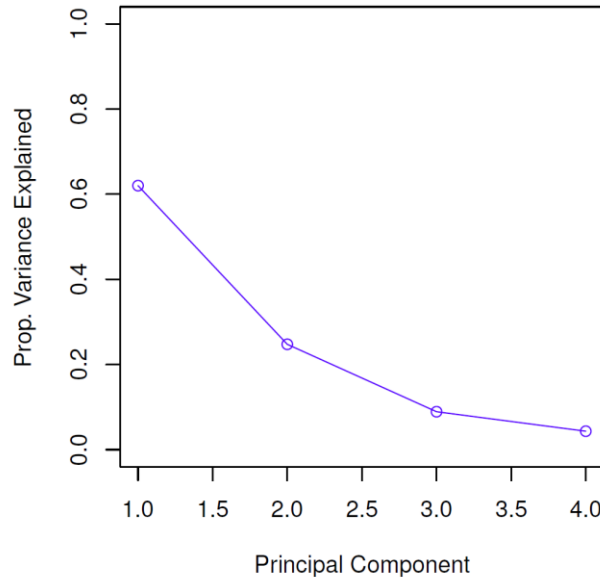
- An illustration of minimizing the reconstruction error to obtain first principal component.





# Proportion Variance Explained

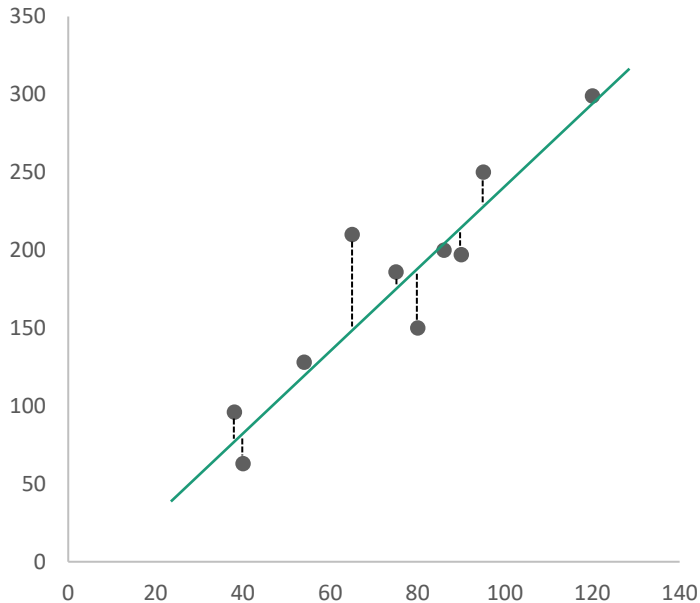
- To understand the **strength of each component**, we are interested in knowing the **proportion of variance explained (PVE)** by each one.



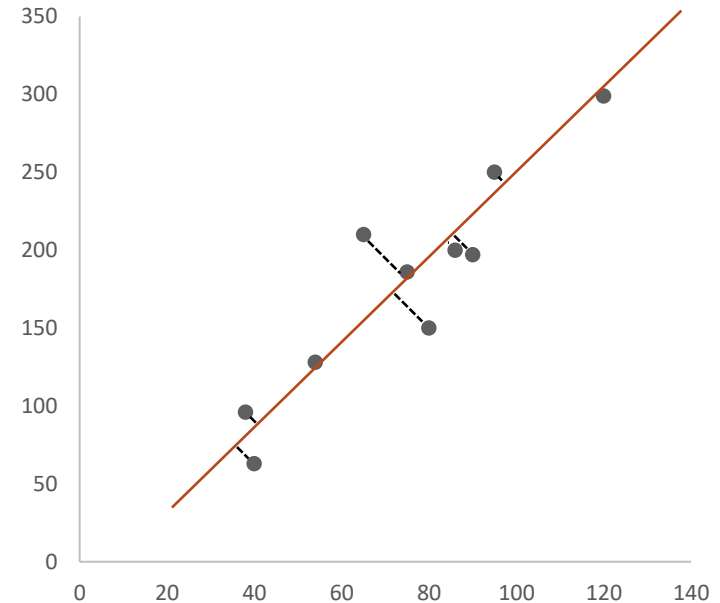
- ✓ Suppose the PVE for 1st PC is 20%. Is this a satisfactory level of explanation?

# PCA vs Linear regression

✓ What is the difference?



Linear regression



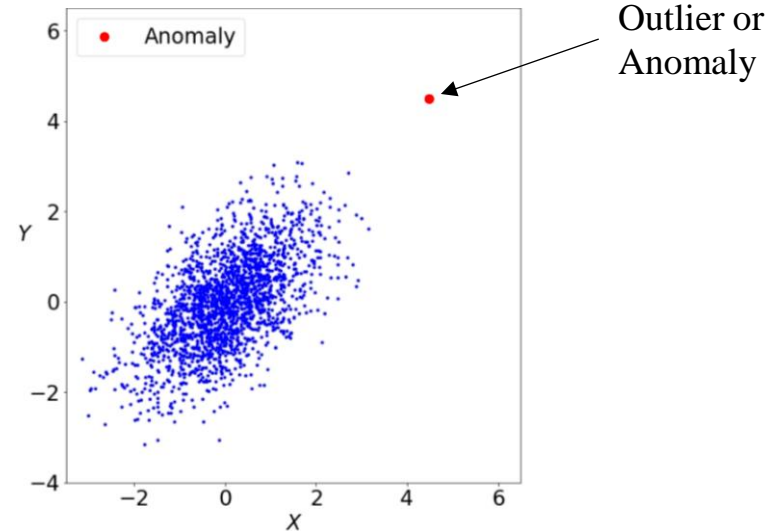
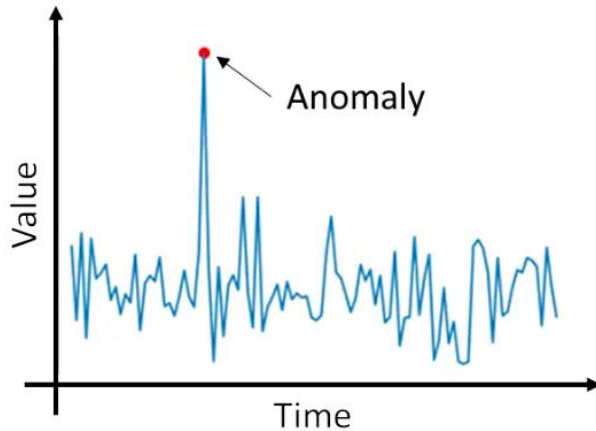
PCA

# Anomaly detection

---

# Anomaly detection

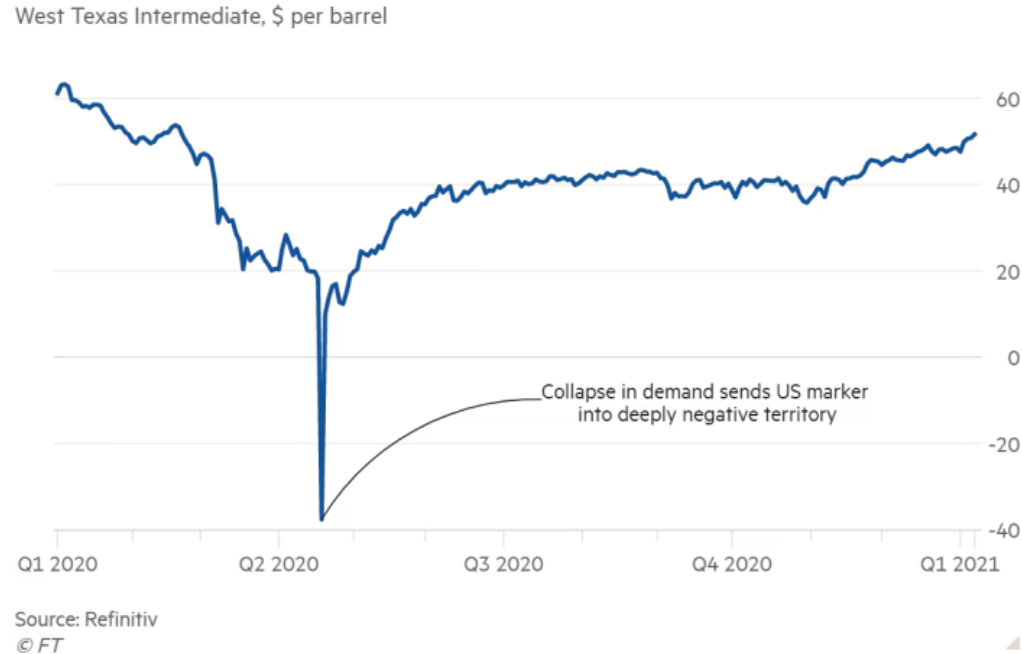
- Assume that a given statistical process is used to generate a set of data objects. An **outlier** is a data object that **deviates significantly from the rest of the objects**, as if it were **generated by a different mechanism**.



- Anomaly detection**, sometimes called **outlier detection**, is a process of finding patterns or instances in a dataset that deviate significantly from the expected or “normal behavior.”

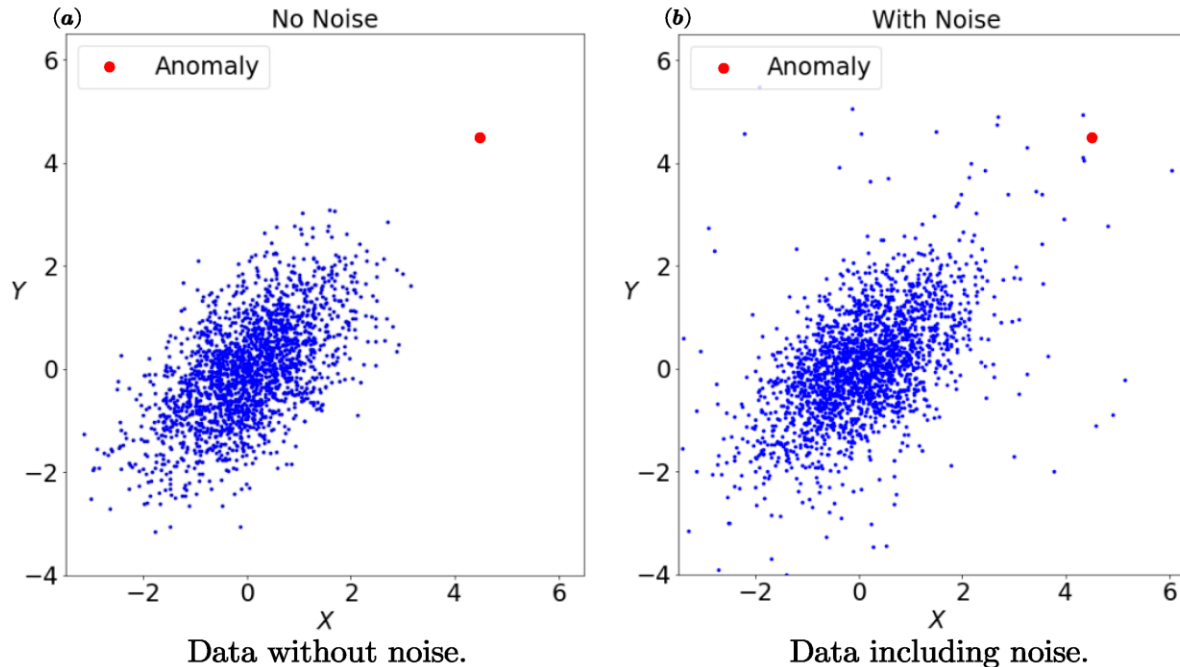
# Anomaly detection; example

- Example: Negative price for WTI in second quarter of 2020



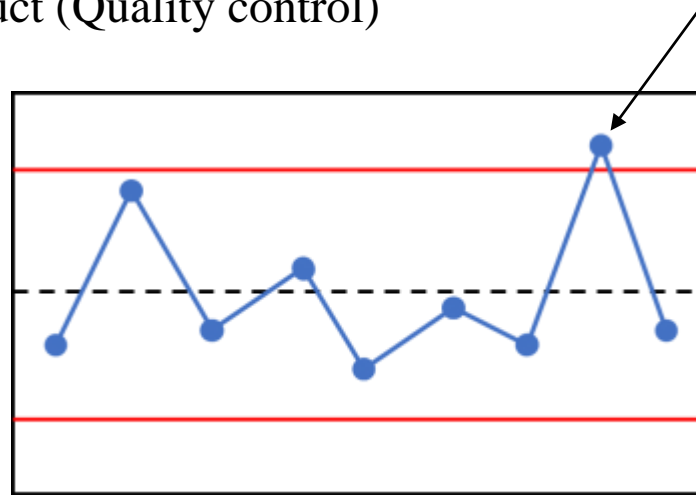
# Anomaly vs noise

- One of the main challenges in outlier detection is finding ways to distinguish outliers from noise.



# Anomaly detection; applications

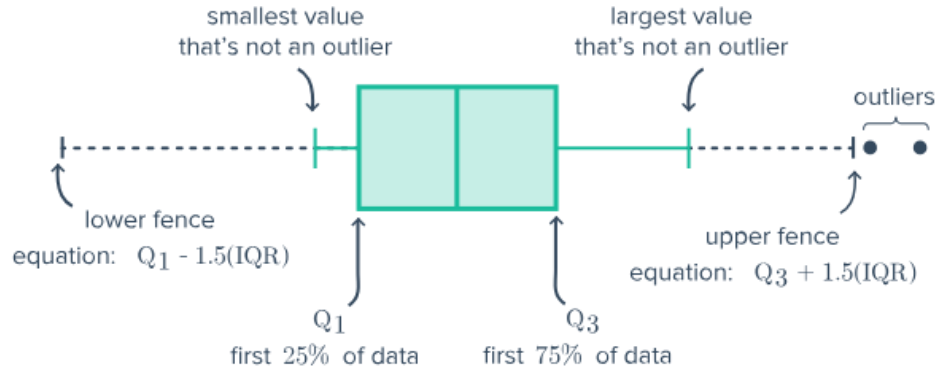
- **Fraud detection** like Financial transactions fraud:  
Normal: Routine purchases and consistent spending by an individual in Tehran.  
Outlier: A massive withdrawal from Tabriz from the same account.
- **Manufacturing:**  
Finding new previously unseen defects in product (Quality control)  
Industrial equipment monitoring
- **Security applications:**  
Monitoring machines in a data center  
Anomalous Login



# Interquartile range method (IQR)

- The upper and lower fences represent the cut-off values for upper and lower outliers in a dataset.

- They are calculated as:
- $$\begin{cases} \text{Lower fence} = Q_1 - (1.5 \times IQR) \\ \text{Upper fence} = Q_3 + (1.5 \times IQR) \end{cases}$$

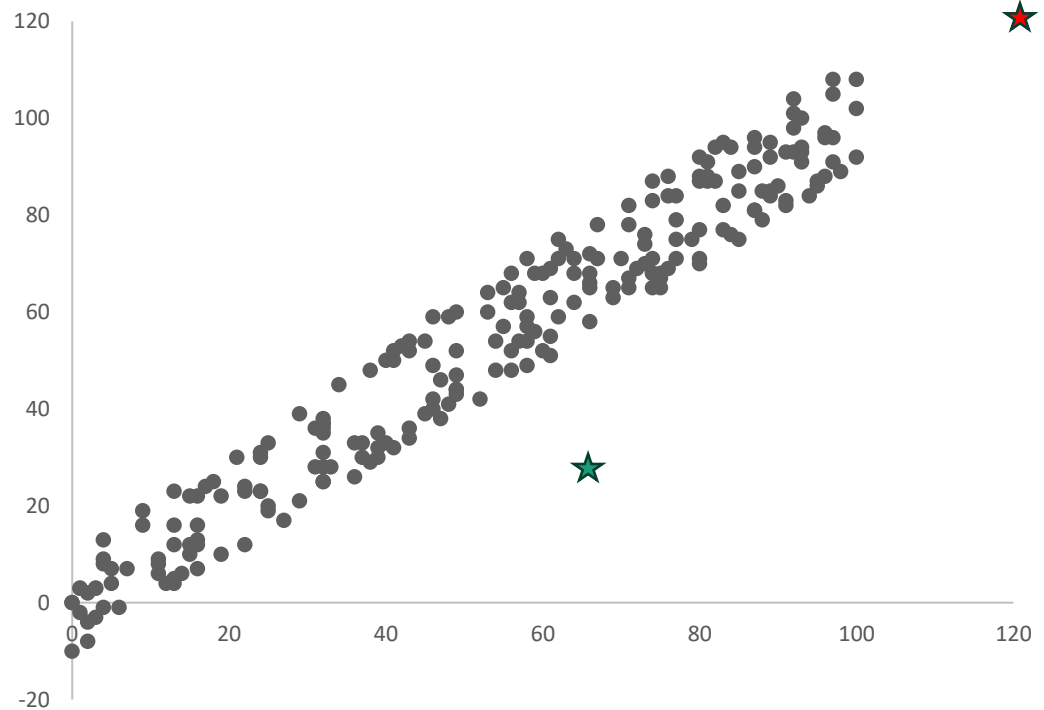


- ✓ Is this method applicable to handling multiple features simultaneously?



# Interquartile range method (IQR)

- ✓ Which anomaly point is discoverable by IQR method?
- Effective anomaly detection in our case requires **considering all features simultaneously**.

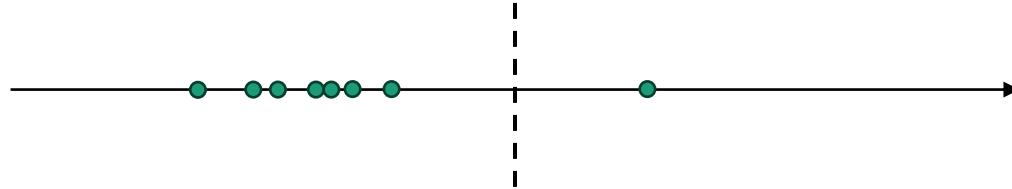


# Isolation Forest

---

# Isolation forest

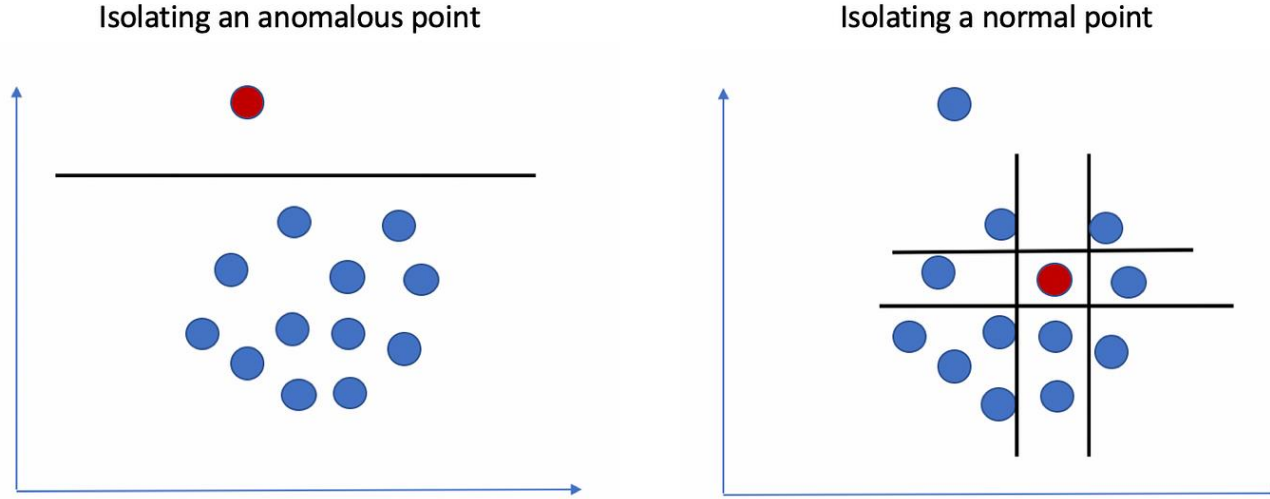
- **Isolation Forest** is an unsupervised machine learning algorithm for anomaly detection.
- As the name implies, Isolation Forest is **an ensemble method** (like random forest). In other words, **it use the average of the predictions by several decision trees when assigning the final anomaly score** to a given data point.
- Main idea:



- The approach employs binary trees to detect anomalies.

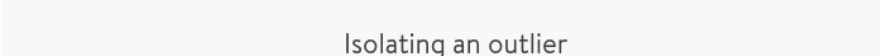
# Isolation tree

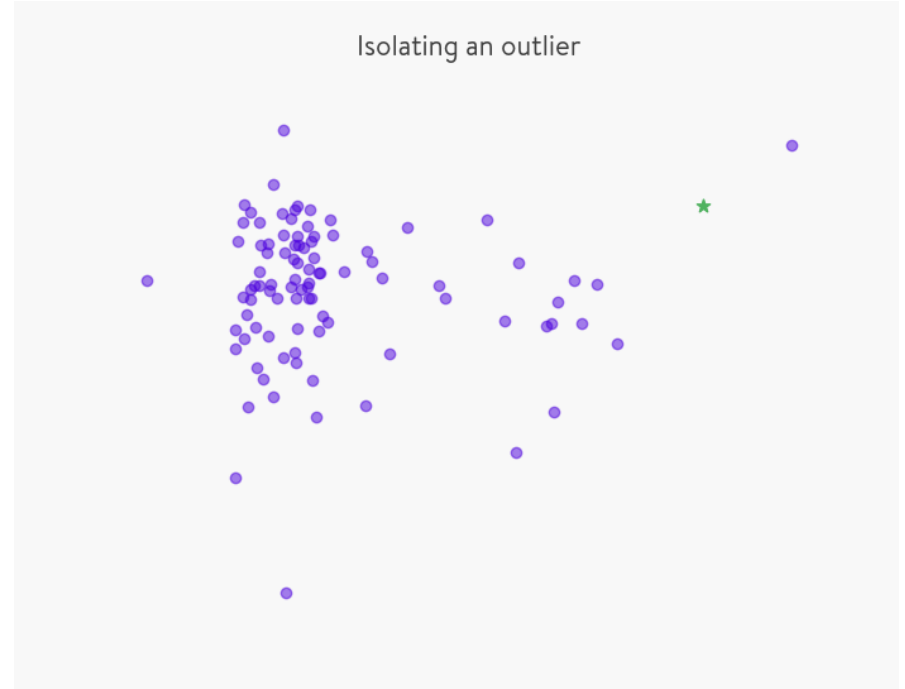
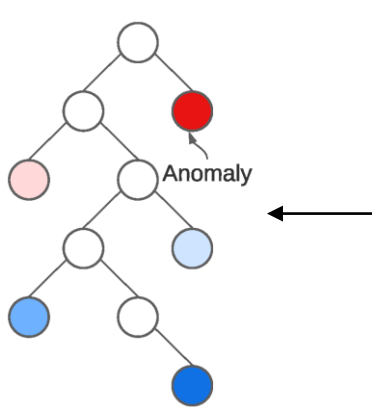
- The **isolation tree** algorithm **selects a random dimension** and **randomly splits the data along that dimension**.



- On average, an **anomalous data** point is going to be isolated in a bounding box at a **smaller tree depth** than other points.

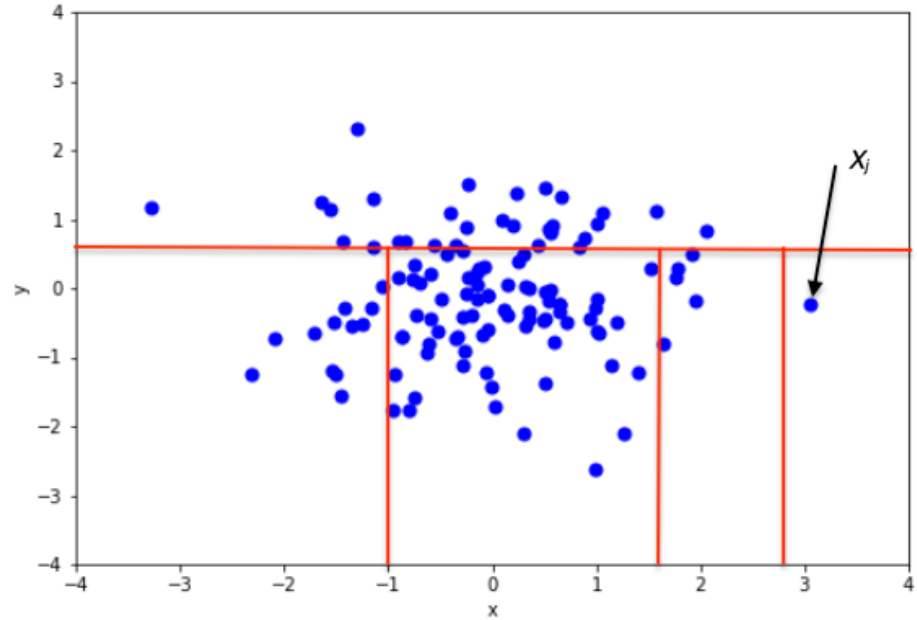
# Isolation tree

- Isolation Forests were built based on the fact that **anomalies** are the data points that are “**few and different**”.
  - Each tree in an Isolation Forest is called an **Isolation Tree. (iTree)**
- 
- The image shows a scatter plot with a light gray background. The title 'Isolating an outlier' is centered at the top in a dark gray font. There is a cluster of approximately 10 purple circular data points in the lower-left quadrant. A single green star is located in the lower-right quadrant, representing an outlier. A single purple circular data point is also located in the upper-right quadrant, separate from the main cluster.



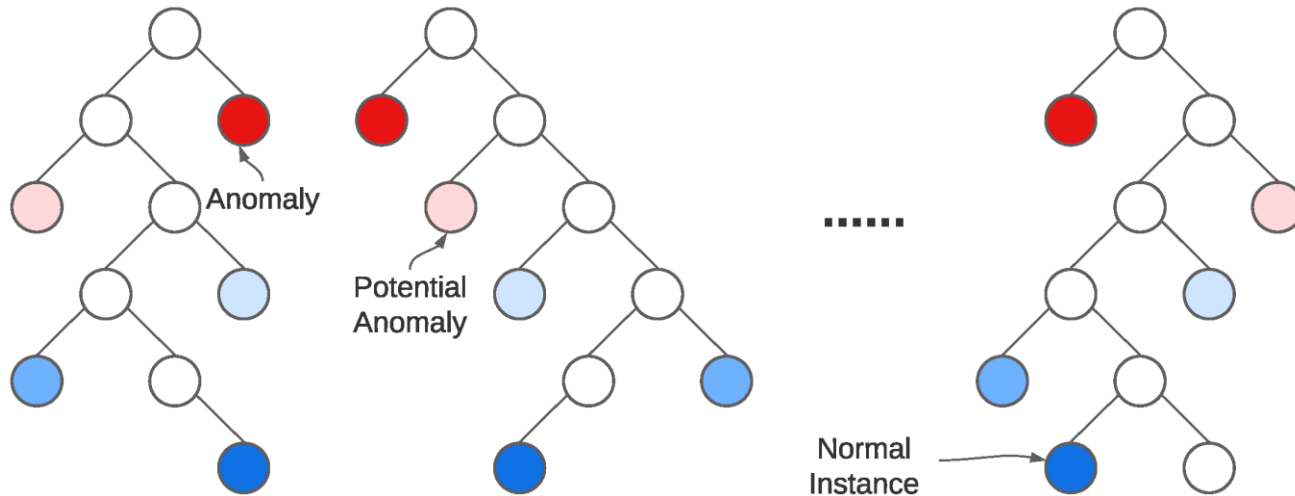
# Isolation tree; example

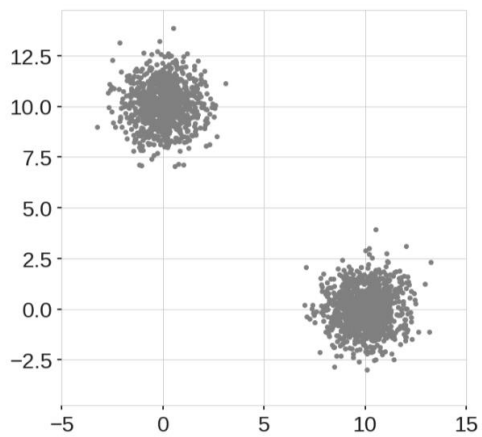
✓ What is the isolation tree here?



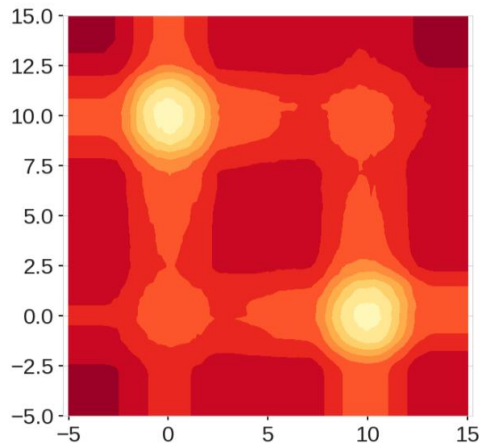
# Isolation forest; Anomaly score

- An **anomaly score** is assigned to each of the data points based on the depth of the tree required to arrive at that point.
- This score is an aggregation of the depth obtained from each of the iTrees. How?

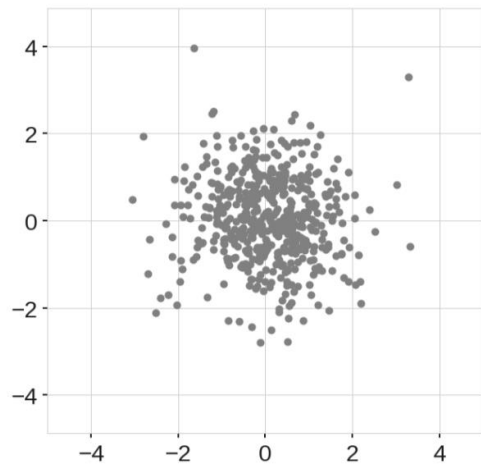
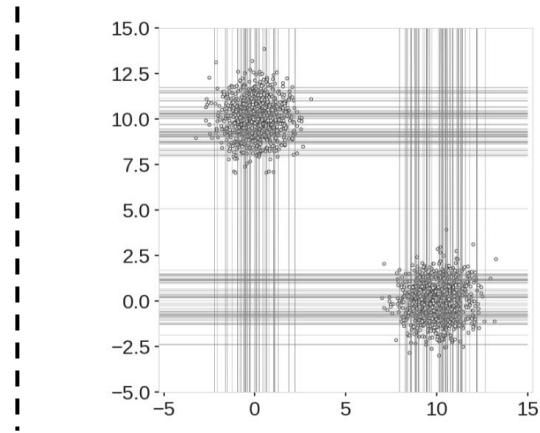




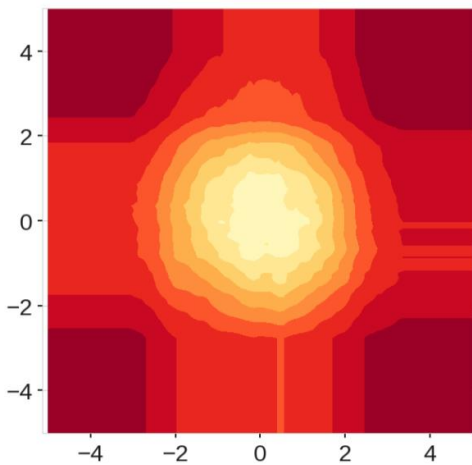
(a) Two normally distributed clusters



(b) Anomaly Score Map



(a) Normally Distributed Data



(b) Anomaly Score Map

