CSE 424
Big Data

Algorithms of Big Data Analytics

Slides 5

Instructor: Asst. Prof. Dr. Hüseyin ABACI

Outline

- Statistical Analysis
  - Summarization
  - A/B Testing
  - Regression
  - Classification
- Machine Learning
  - Classification (Supervised)
  - Clustering (Unsupervised)
- Filtering
  - Content Filtering
  - Collaborative Filtering
  - Similarity Measures
- Recommendation Systems
  - Matrix Factorisation
  - Alternating Least Squares (ALS)
- Performance Measures of Metrics

Thomas Erl, Wajid Khattak, and Paul Buhler, Big Data Fundamentals Concepts: Drivers & Techniques, 2016
Bahga, Arshdeep, and Vijay Madisetti. Big data science & analytics: A hands-on approach. VPT, 2016.

Statistical Analysis

- Statistical analysis uses statistical methods based on mathematical formulas as a means for analyzing data.
- This type of analysis is commonly used to describe datasets via summarization, such as providing the counts, mean, median, min, max or TopN of statistics associated with the dataset.
- Along side to A/B testing, analysis can also be used to infer patterns and relationships within the dataset, such as regression and correlation.

Statistical Analysis - Summarization

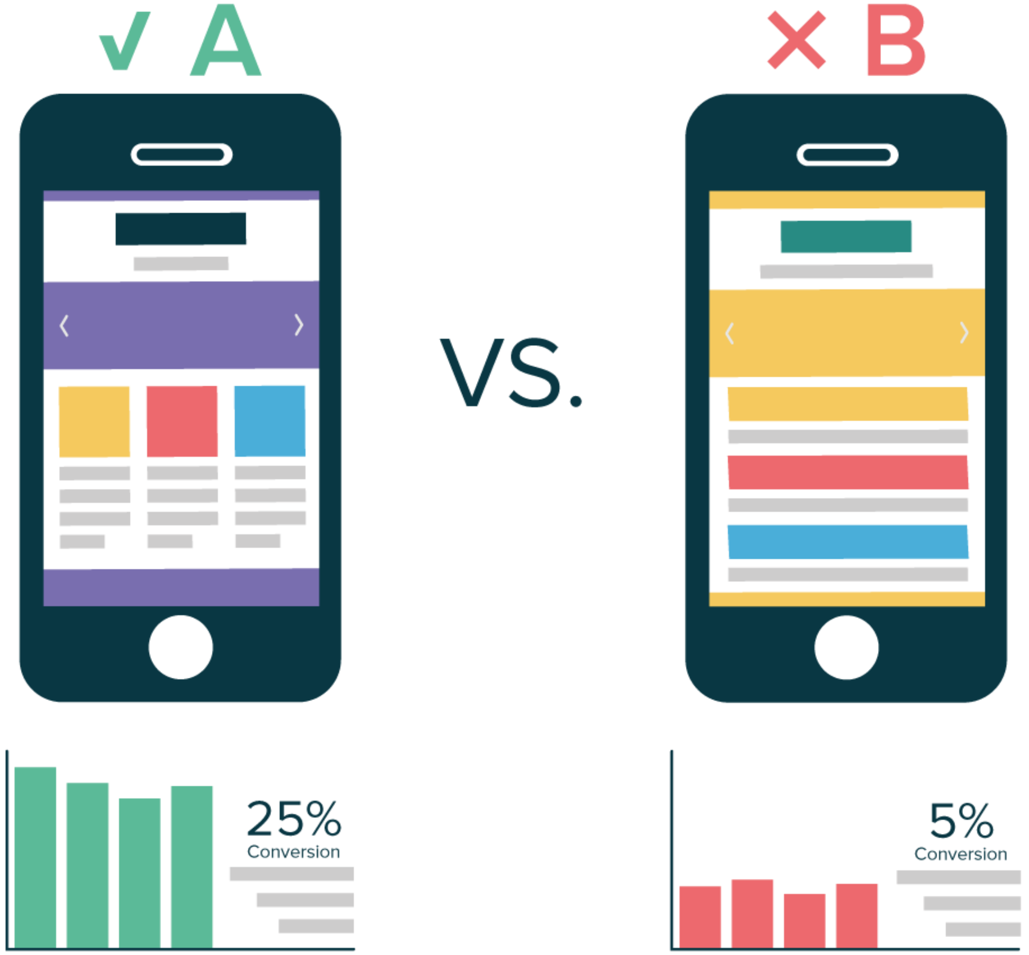
- Numerical summarization are used to compute various statistics such as counts, maximum, minimum, mean, etc.
- These statistics help in presenting the data in a summarized form.
- For example, computing the total number of likes for a particular post, computing the average monthly rainfall or finding the average number of visitors per month on a website.
- Most commonly the MapRaduce programing model is used for big data summarization.
- MapReduce is best suited for descriptive analytics and the basic statistics computational tasks because the operations involved can be done in parallel.

Summarizations

- **Count:** Calculates occurrences of a entry in big data. For example, number of words in a text (word count), number of clicks per link, number of likes per item etc.
- **Max/Min:** Similar to count, in this method summarization finds maximum and minimums in big data. For example, the best selling products etc.
- **TopN:** Finds top N entry in big data. For example, finds what is the top selling 10 products in the shop, 5 most liked movies etc.

A/B Testing

- A/B testing, also known as split or bucket testing, compares two versions of an element to determine which version is superior based on a pre-defined metric. The element can be a range of things.



https://retail.strands.com
/products/ab-testing/

A/B Testing (cont.)

- Although A/B testing can be implemented in almost any domain, it is most often used in marketing.
- For example, in order to determine the best possible layout for an ice cream ad on Company A’s Web site, two different versions of the ad are used. Version A is an existing ad (the control) while Version B has had its layout slightly altered (the treatment).
- Both versions are then simultaneously shown to different users:
  - ➢ Version A to Group A
  - ➢ Version B to Group B
- The analysis of the results reveals that Version B of the ad resulted in more sales as compared to Version A.

A/B Testing (cont.)

- Sample questions can include:
  - *Is the new version of a drug better than the old one?*
  - *Do customers respond better to advertisements delivered by email or postal mail?*
  - *Is the newly designed homepage of the Web site generating more user traffic?*

Correlation

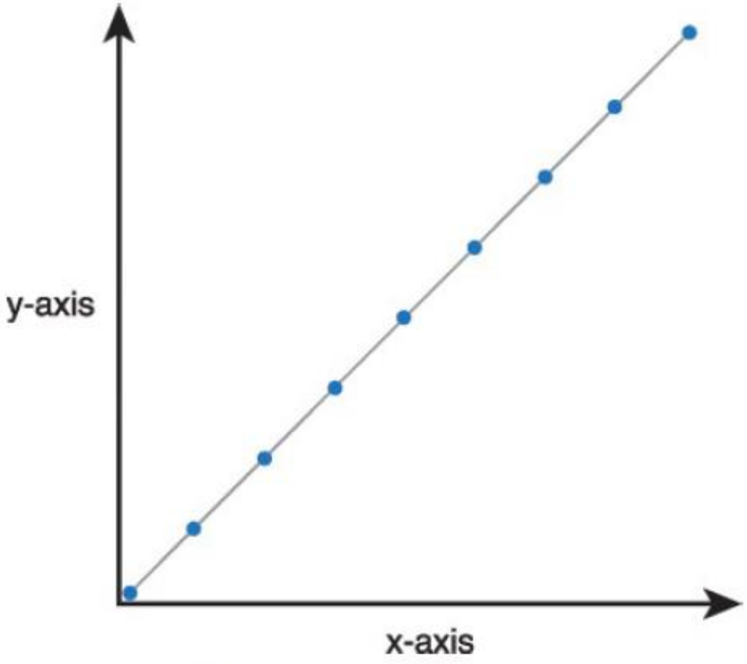
- Correlation is an analysis technique used to determine whether two variables are related to each other. If they are found to be related, the next step is to determine what their relationship is.
- For example, the value of Variable A increases whenever the value of Variable B increases. We may be further interested in discovering how closely Variables A and B are related, which means we may also want to analyze the extent to which Variable B increases in relation to Variable A’s increase.

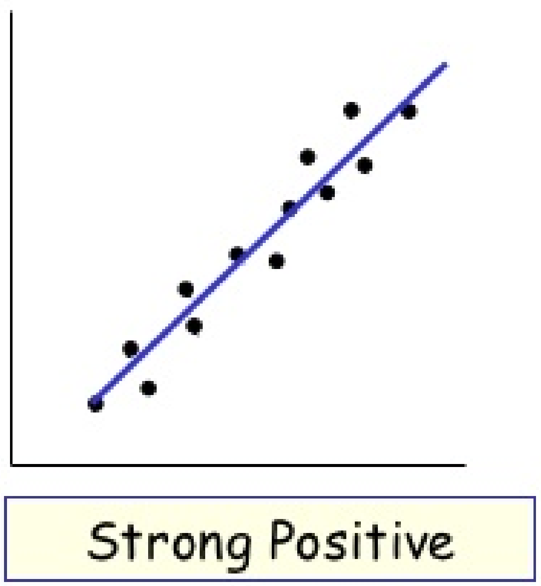
Correlation (cont.)

- Correlation is therefore commonly used for data mining where the identification of relationships between variables in a dataset leads to the discovery of patterns and anomalies. This can reveal the nature of the dataset or the cause of a phenomenon.
- When two variables are considered to be correlated they are aligned based on a linear relationship. This means that when one variable changes, the other variable also changes proportionally and constantly.

Correlation (cont.)

- Correlation is expressed as a decimal number between –1 to +1, which is known as the correlation coefficient. The degree of relationship changes from being strong to weak when moving from –1 to 0 or +1 to 0.
- Below figure shows a correlation of +1, which suggests that there is a strong positive relationship between the two variables.

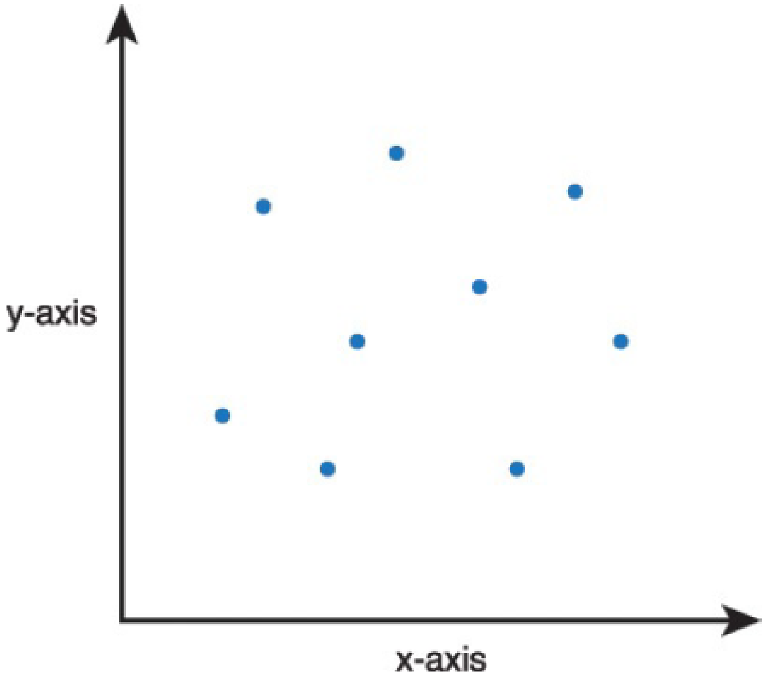




When one variable increases, the other also increases and vice versa.

Correlation (cont.)

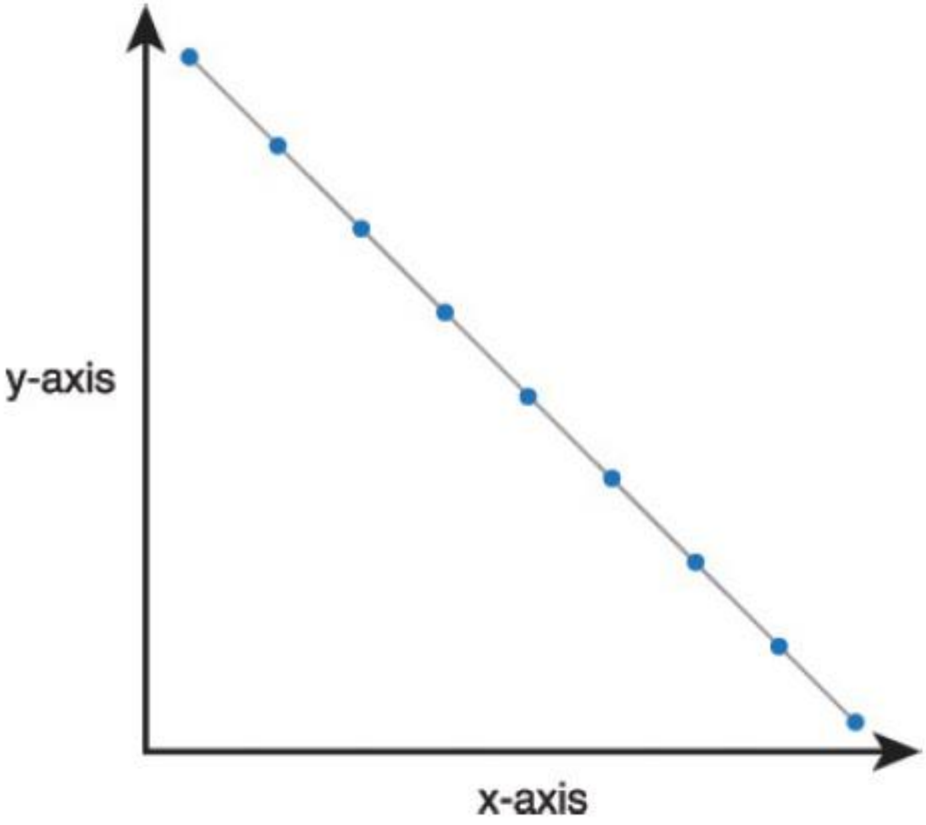
- Figure below shows a correlation of 0, which suggests that there is no relationship at all between the two variables.



When one variable increases, the other may stay the same, or increase or decrease arbitrarily.

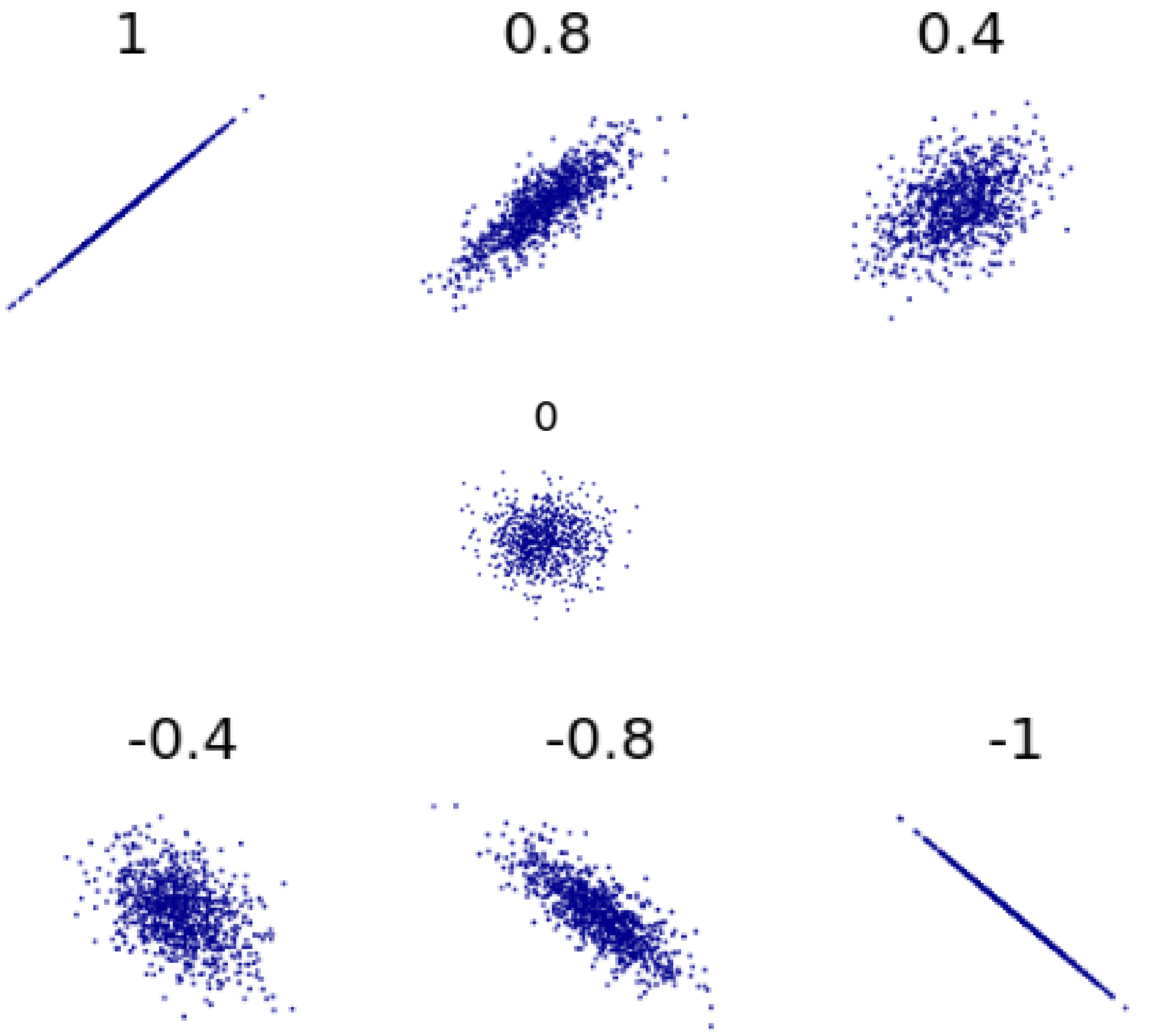
Correlation (cont.)

- In figure below, a slope of –1 suggests that there is a strong negative relationship between the two variables.



When one variable increases, the other decreases and vice versa.

Correlation (cont.)



Correlation (cont.)

- For example, managers believe that ice cream stores need to stock more ice cream for hot days, but don't know how much extra to stock. To determine if a relationship actually exists between temperature and ice cream sales, the analysts first apply correlation to the number of ice creams sold and the recorded temperature readings. A value of +0.75 suggests that there exists a strong relationship between the two. This relationship indicates that as temperature increases, more ice creams are sold.

Correlation (cont.)

Further sample questions addressed by correlation can include:

- *Does distance from the sea affect the temperature of a city?*
- *Do students who perform well at elementary school perform equally well at high school?*
- *To what extent is obesity linked with overeating?*

Regression

- The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset**.**
- As a sample scenario, regression could help determine the type of relationship that exists between **temperature, the independent variable**, and **crop yield, the dependent variable**.
- Applying this technique helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variable. When the independent variable increases, for example, does the dependent variable also increase? If yes, is the increase in a linear or non-linear proportion?

Regression (cont.)

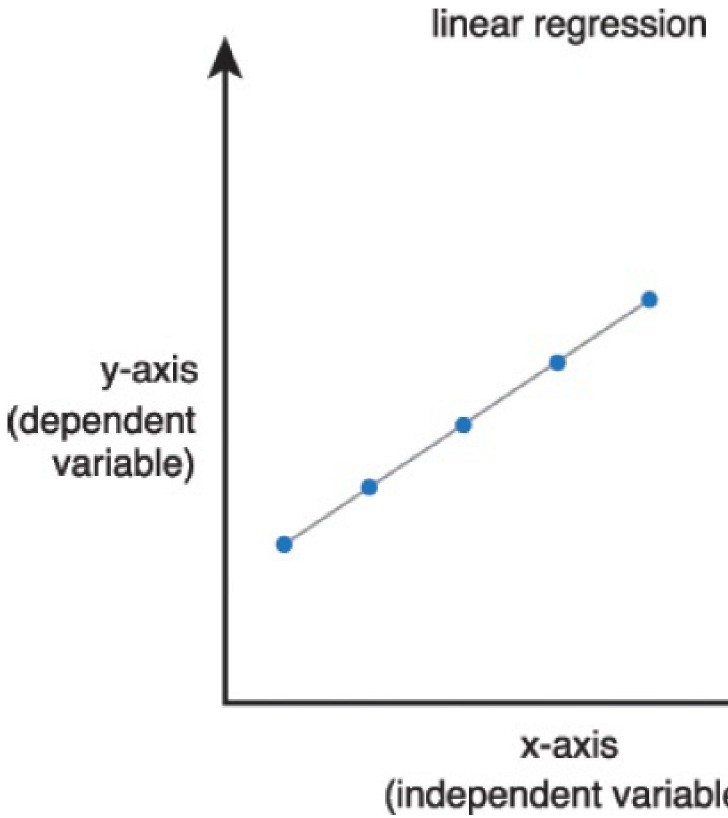
- For example, in order to determine how much extra stock each ice cream store needs to have, the analysts apply regression by feeding in the values of temperature readings. These values are based on the weather forecast as an independent variable and the number of ice creams sold as the dependent variable. What the analysts discover is that 15% of additional stock is required for every 5-degree increase in temperature.

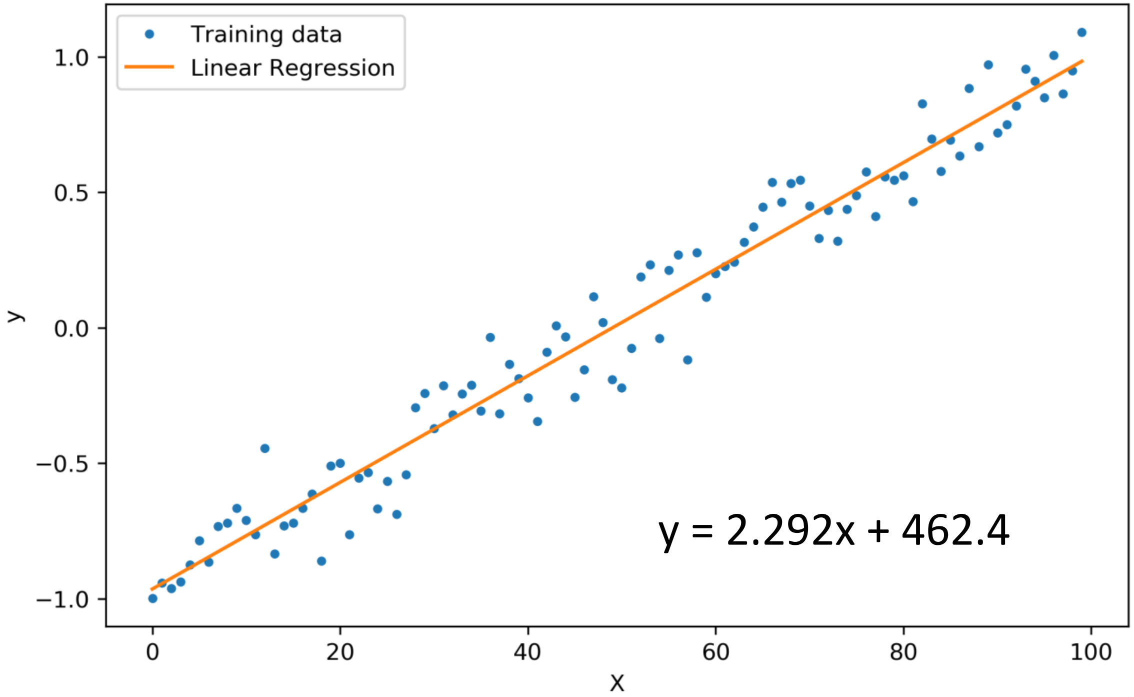
Regression (cont.)

- More than one independent variable can be tested at the same time. However, in such cases, only one independent variable may change, while others are kept constant. Regression can help enable a better understanding of what a phenomenon is and why it occurred. It can also be used to make predictions about the values of the dependent variable.

Linear Regression

- Linear regression represents a constant rate of change, as shown in below figure.





Linear Regression

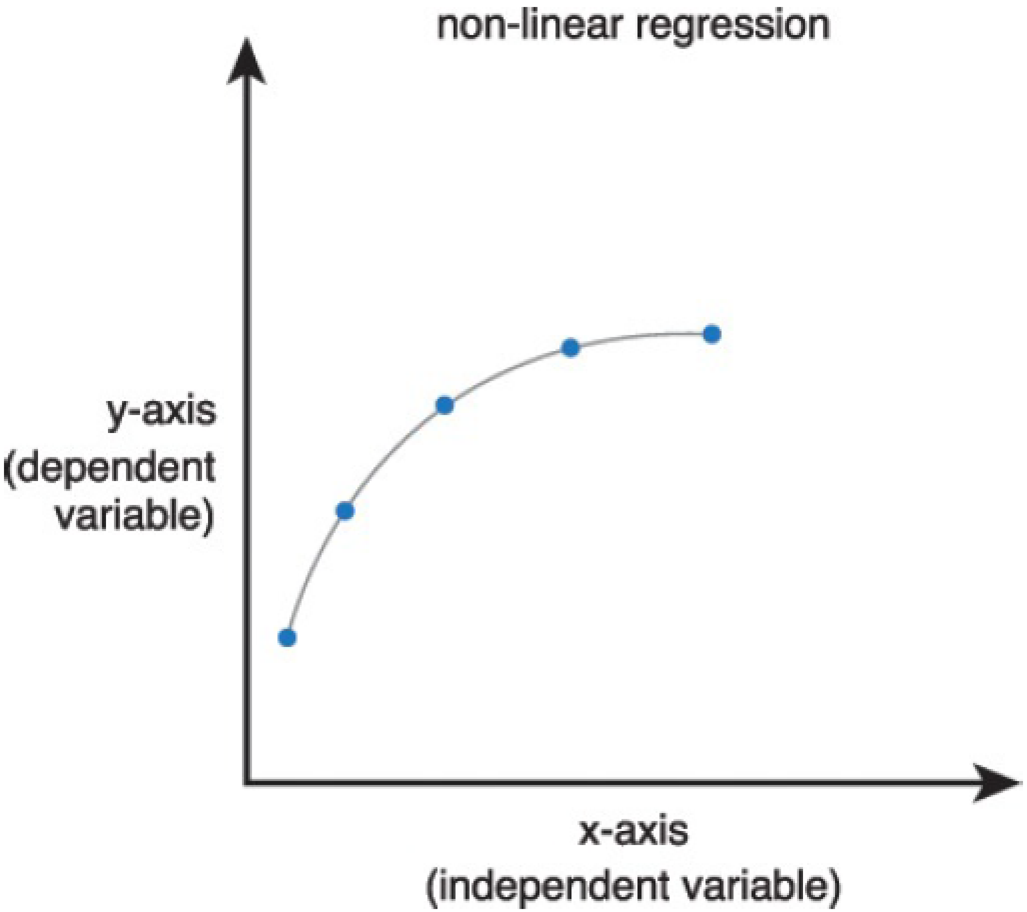
- In linear regression a dependent variable ***y*** is modelled as a linear combination of the independent variables ***x***.
- In regression, the goal is to learn a function h(x) from the training set, which can predict the values of y. The function h(x) is called the hypothesis.
- For linear regression,

$$h(x) = \sum_{i=0}^{n} \theta_i x_i$$

where θ_0, θ_1, ..., θ_n are the parameters.

Non-Linear Regression

- Non-linear regression represents a variable rate of change, as shown in below figure.

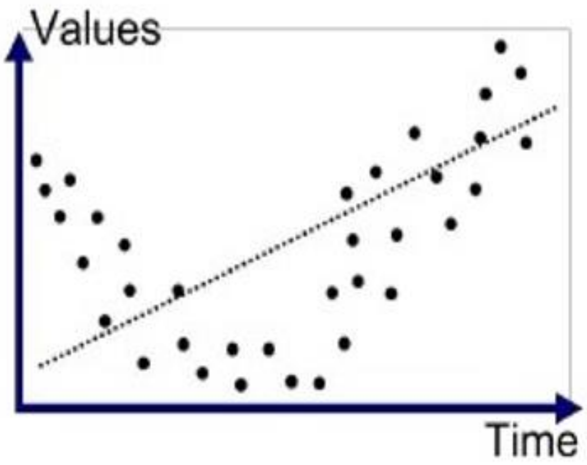


Regression (cont.)

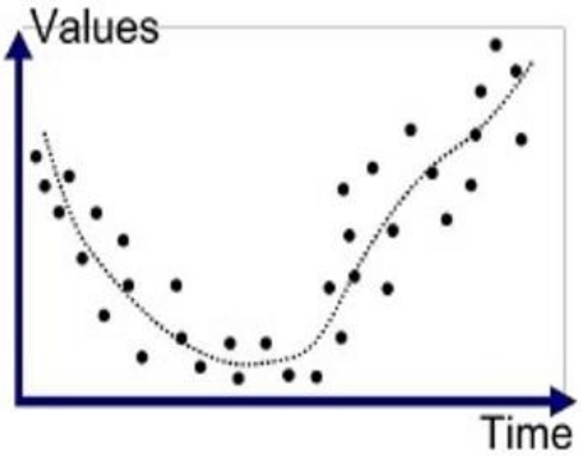
Sample questions can include:

- *What will be the temperature of a city that is 250 miles away from the sea?*
- *What will be the grades of a student studying at a high school based on their primary school grades?*
- *What are the chances that a person will be obese based on the amount of their food intake?*

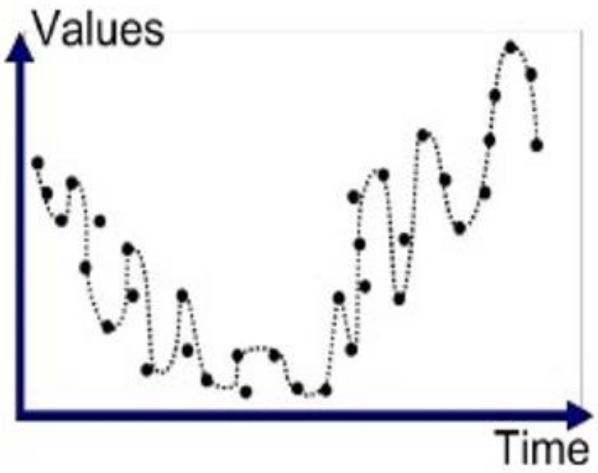
Overfitting / Underfitting



Underfitted



Good Fit/Robust



Overfitted

Correlation vs. Regression

- Regression and correlation have a number of important differences.
- Correlation does not imply causation. The change in the value of one variable may not be responsible for the change in the value of the second variable, although both may change at the same rate. This can occur due to an unknown third variable, known as the ***confounding factor***. Correlation assumes that both variables are independent.
- Regression, on the other hand, is applicable to variables that have previously been identified as dependent and independent variables and implies that there is a degree of causation between the variables. The causation may be direct or indirect.

Correlation vs. Regression (cont.)

- Within Big Data, correlation can first be applied to discover if a relationship exists. Regression can then be applied to further explore the relationship and predict the values of the dependent variable, based on the known values of the independent variable.

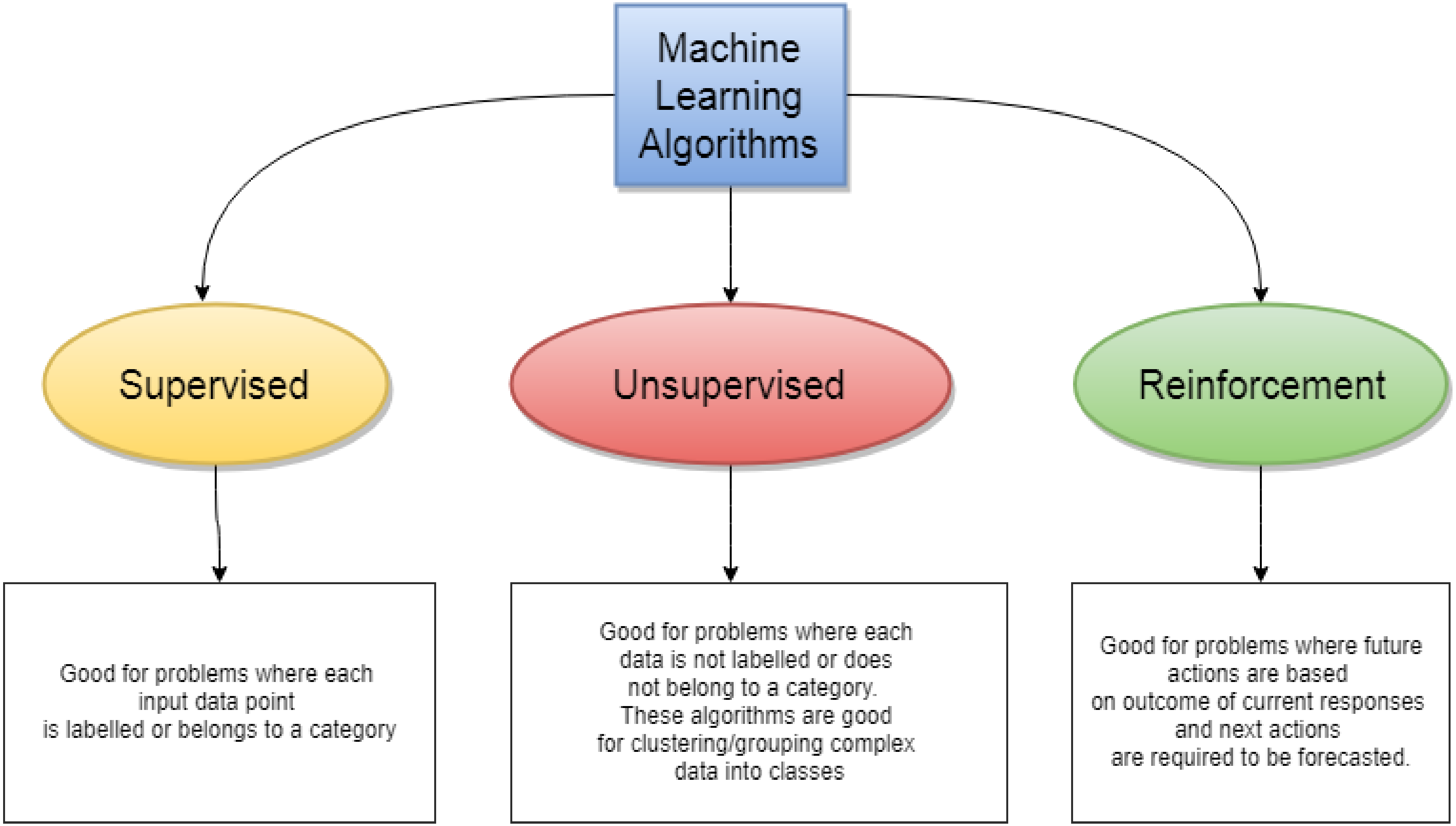
Machine Learning

- Humans are good at spotting patterns and relationships within data. Unfortunately, we cannot process large amounts of data very quickly. Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how.
- Human knowledge can be combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning.

Data Mining vs. Machine Learning

- **Data Mining** is about using **Statistics** as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon. Data Mining builds intuition about what is really happening in some data and is still little more towards math than programming, but uses both.
- **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes. Math is the basis for many of the algorithms, but this is more towards programming.
- **Data Mining** explains patterns while **Machine Learning** predicts with models.

Machine Learning



Machine Learning Algorithms

Supervised

Unsupervised

Reinforcement

Good for problems where each input data point is labelled or belongs to a category

Good for problems where each data is not labelled or does not belong to a category. These algorithms are good for clustering/grouping complex data into classes

Good for problems where future actions are based on outcome of current responses and next actions are required to be forecasted.

Predict next value

Identify clusters

Learn from mistakes

Classification

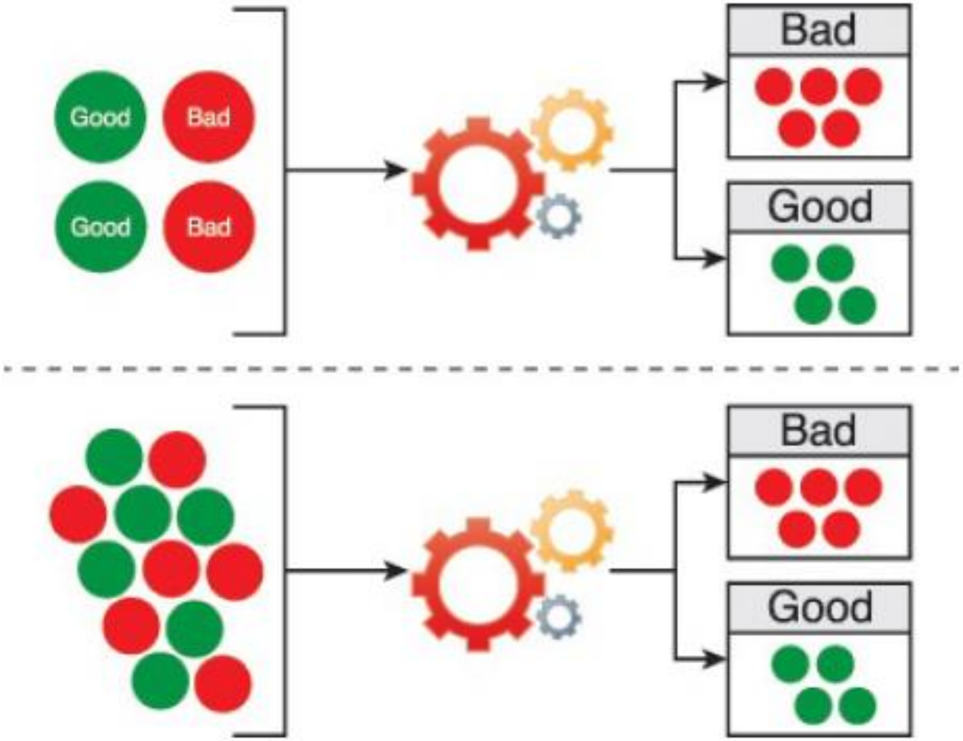
- Classification is the process of categorizing objects into predefined categories.
- Classification is achieved by classification algorithms that belong to a broad category of algorithms called supervised machine learning.
- Supervised learning involves inferring a model from a set of input data and **known responses** (labels) to the data (training data) and then using the inferred model to predict responses to new data.

Classification (cont'd)

- There are various types of classification approaches for big data analytics including:
- **Binary classification:** Binary classification involves categorizing the data into two categories. For example, classifying the sentiment of a news article into positive or negative, classifying the state of a machine into good or faulty, classifying the health test into positive or negative, etc.
- **Multi-class classification:** Multi-class classification involves more than two classes into which the data is categorized. For example, classify a set of images of fruits which may be oranges, apples, or pears.
- **Document classification:** Document classification is a type of multi-class classification approach in which the data to the classified is in the form of text document. For example, classifying news articles into different categories such as politics, sports, etc.

Classification (cont'd)

- It consists of two steps:
  1. The system is fed training data that is already categorized or labeled, so that it can develop an understanding of the different categories.
  2. The system is fed unknown but similar data for classification and based on the understanding it developed from the training data, the algorithm will classify the unlabeled data.



Classification (cont’d)

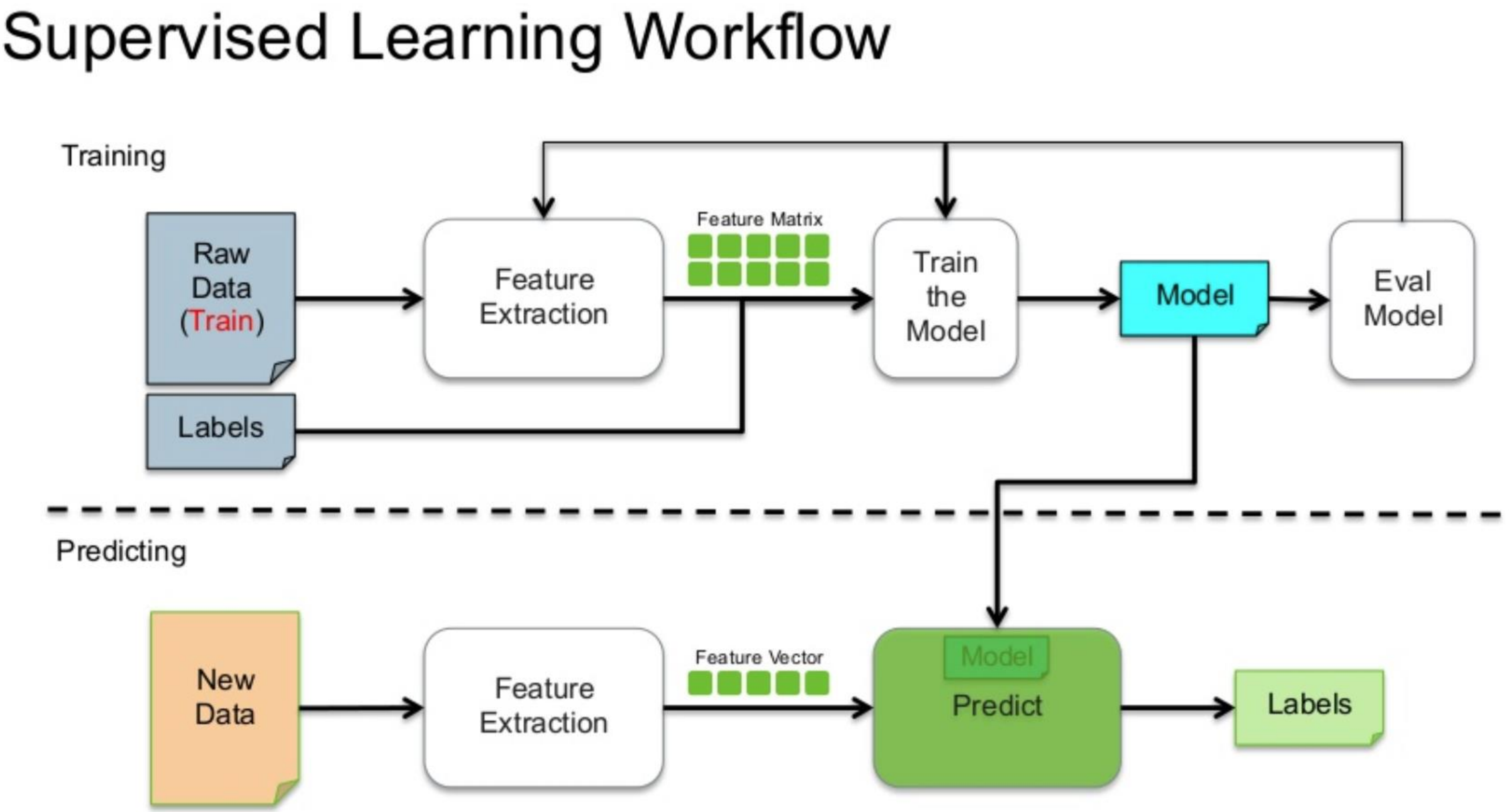
- For example, a bank wants to find out which of its customers is likely to default on loan payments. Based on historic data, a training dataset is compiled that contains labeled examples of customers that have or have not previously defaulted. This training data is fed to a classification algorithm that is used to develop an understanding of “good” and “bad” customers.
- Finally, new untagged customer data is fed in order to find out whether a given customer belongs to the defaulting category.

Classification (cont’d)

Sample questions can include:

- *Should an applicant’s credit card application be accepted or rejected based on other accepted or rejected applications?*
- *Is a tomato a fruit or a vegetable based on the known example of fruit and vegetables?*
- *Do the medical test results for the patient indicate a risk for a heart attack?*

Classification (cont'd)



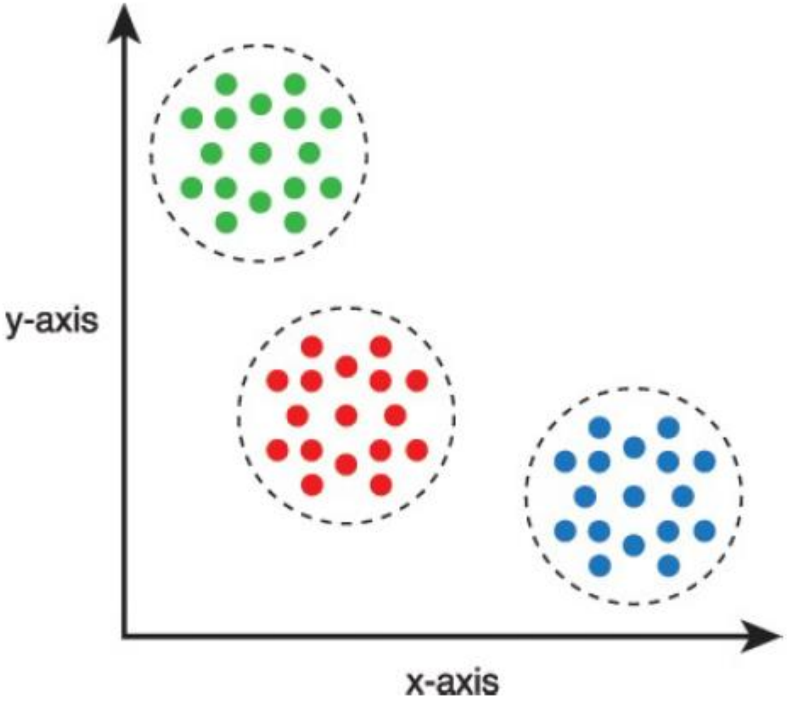
Supervised Learning Workflow

Clustering (Unsupervised Machine Learning)

- Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties. There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings. How the data is grouped depends on the type of algorithm used. Each algorithm uses a different technique to identify clusters.
- Clustering is generally used in data mining to get an understanding of the properties of a given dataset. After developing this understanding, classification can be used to make better predictions about similar but new or unseen data.

Clustering (Unsupervised Machine Learning) (cont.)

- Clustering can be applied to the categorization of unknown documents and to personalized marketing campaigns by grouping together customers with similar behavior.



Clustering (Unsupervised Machine Learning) (cont.)

- For example, a bank wants to introduce its existing customers to a range of new financial products based on the customer profiles it has on record. The analysts categorize customers into multiple groups using clustering. Each group is then introduced to one or more financial products most suitable to the characteristics of the overall profile of the group.

Clustering (Unsupervised Machine Learning) (cont.)

Typical cluster models include:

- Connectivity models
- Centroid models
- Distribution models
- Density models
- Subspace models
- Graph-based models

Clustering (Unsupervised Machine Learning) (cont.)

Sample questions can include:

- *How many different species of trees exist based on the similarity between trees?*
- *How many groups of customers exist based upon similar purchase history?*
- *What are the different groups of viruses based on their characteristics?*

Filtering

- Filtering is the automated process of finding relevant items from a pool of items. Items can be filtered either based on a user’s own behavior or by matching the behavior of multiple users.

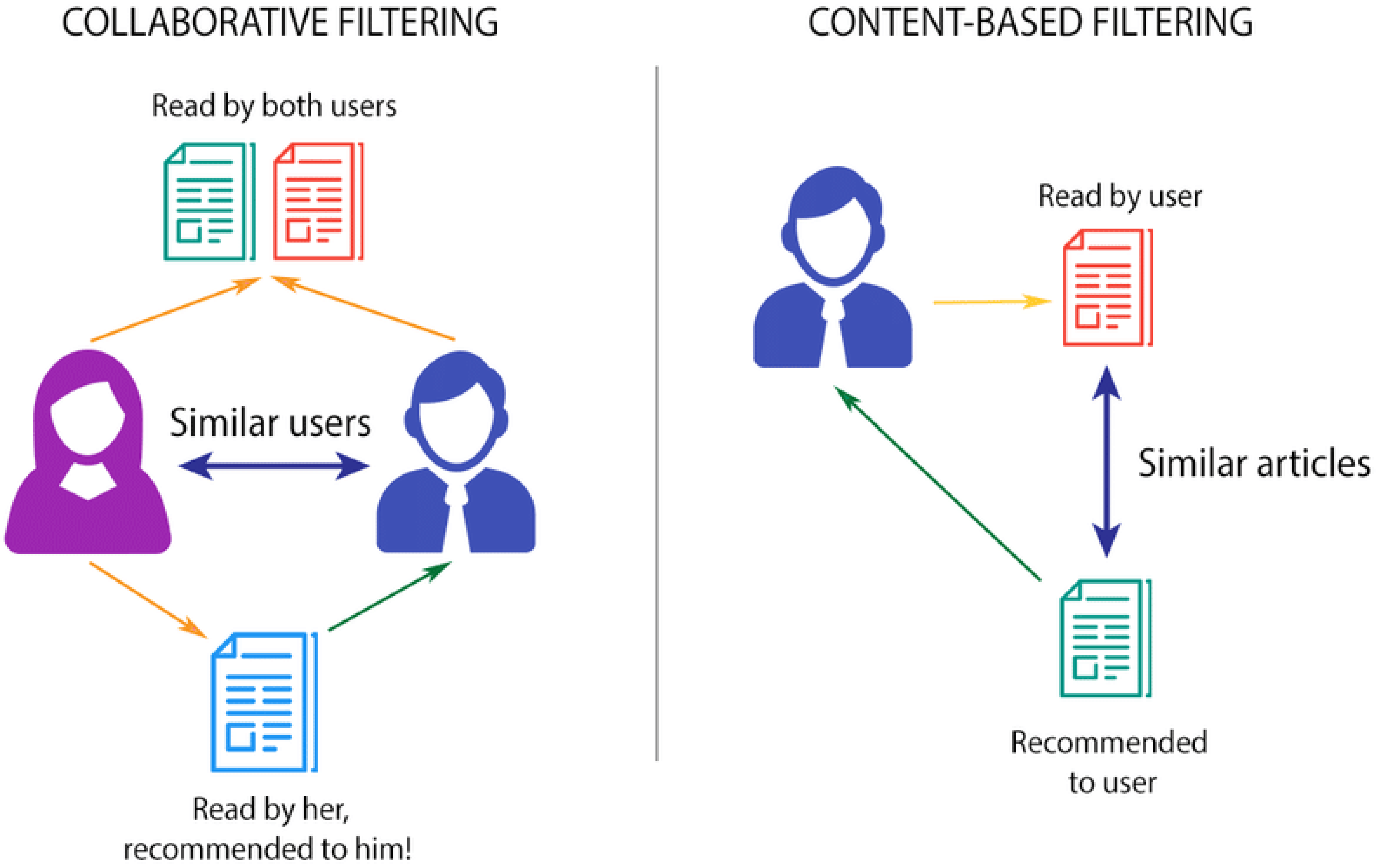
Sample questions can include:

- *How can only the news articles that a user is interested in be displayed?*
- *Which holiday destinations can be recommended based on the travel history of a vacationer?*
- *Which other new users can be suggested as friends based on the current profile of a person?*

Filtering (cont.)

- Filtering is generally applied via the following two approaches:
  - Collaborative filtering
  - Content-based filtering

Filtering – Simple Example



https://www.themarketingtechnologist.co/building-a-recommendation-engine-for-geeksetting-up-the-prerequisites-13/

Collaborative Filtering (cont.)

- A common medium by which filtering is implemented via the use of a recommender system. **Collaborative filtering** is an item filtering technique based on the collaboration, or merging, of a user’s past behavior with the behaviors of others. A target user’s past behavior, including their likes, ratings, purchase history and more, is collaborated with the behavior of similar users. Based on the similarity of the users’ behavior, items are filtered for the target user.
- **Collaborative filtering** is solely based on the similarity between users’ behavior. It requires a large amount of user behavior data in order to accurately filter items. It is an example of the application of the law of large numbers.

Collaborative Filtering (cont.)

- Collaborative filtering approaches are of two types:
  - **Memory-based approach**: There are two types of memory-based approaches: user-based collaborative filtering and item-based collaborative filtering.
    - User-based collaborative filtering finds users similar to a given user and recommends the items they have liked.
    - Item-based collaborative filtering finds items similar to the items a user has previously liked.
  - The similarity between users (in user-based collaborative filtering) or items (in item-based collaborative filtering) is calculated using the users' ratings of the items.

Collaborative Filtering (cont.)

- **Model-based approach**: In model-based collaborative filtering approach (e.g. ALS), a model of user ratings is built first and then the model is used to make predictions. Method adopts a probabilistic approach and predicts the user ratings for the items which the user has not rated.

Collaborative Filtering (cont.)

https://en.wikipedia.org/wiki/Recommender_system#/media/File:Collaborative_filtering.gif

Content-based Filtering

- Content based filtering finds similar items, vector representation of two items, similarity measures (such as cosine similarity) or neighbourhood methods (such as clustering methods) are used. This approach requires the items to have certain meaningful features which can be used for computing similarity.
- Content based filtering is about answering the following question: for a certain item, what are the items most similar to it? Here, the precise definition of similarity is dependent on the model involved.

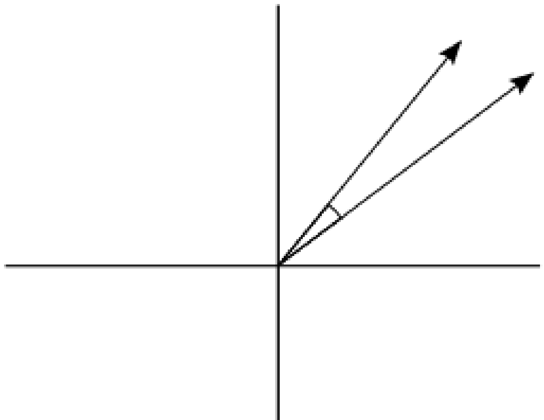
Content-based Filtering – Similarity Measures

- In most cases, similarity is computed by comparing the **vector representation of two items** using some similarity measure to produce a single value.
- Common similarity measures include
  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance
  - Jaccard Similarity
  - Cosine Similarity

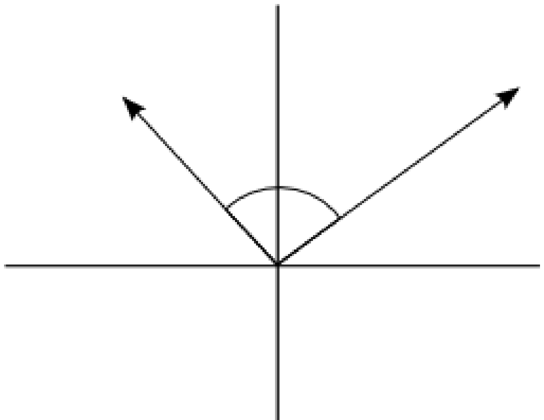
Similarity Measures – Cosine Similarity

- Cosine similarity is a measure of the angle between two vectors in an n-dimensional space. It is computed by first calculating the dot product between the vectors and then dividing the result by a denominator, which is the norm (or length) of each vector multiplied together (specifically, the L2-norm is used in cosine similarity). In this way, cosine similarity is a normalized dot product.
- The cosine similarity measure takes on values between -1 and 1. A value of 1 implies completely similar, while a value of 0 implies independence (that is, no similarity).
- This measure is useful because it also captures negative similarity, that is, a value of -1 implies that not only are the vectors not similar, but they are also completely opposite (dissimilar).

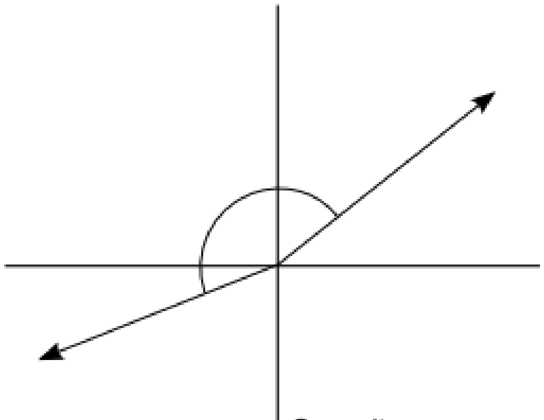
Similarity Measures – Cosine Similarity



Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%



Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%



Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%

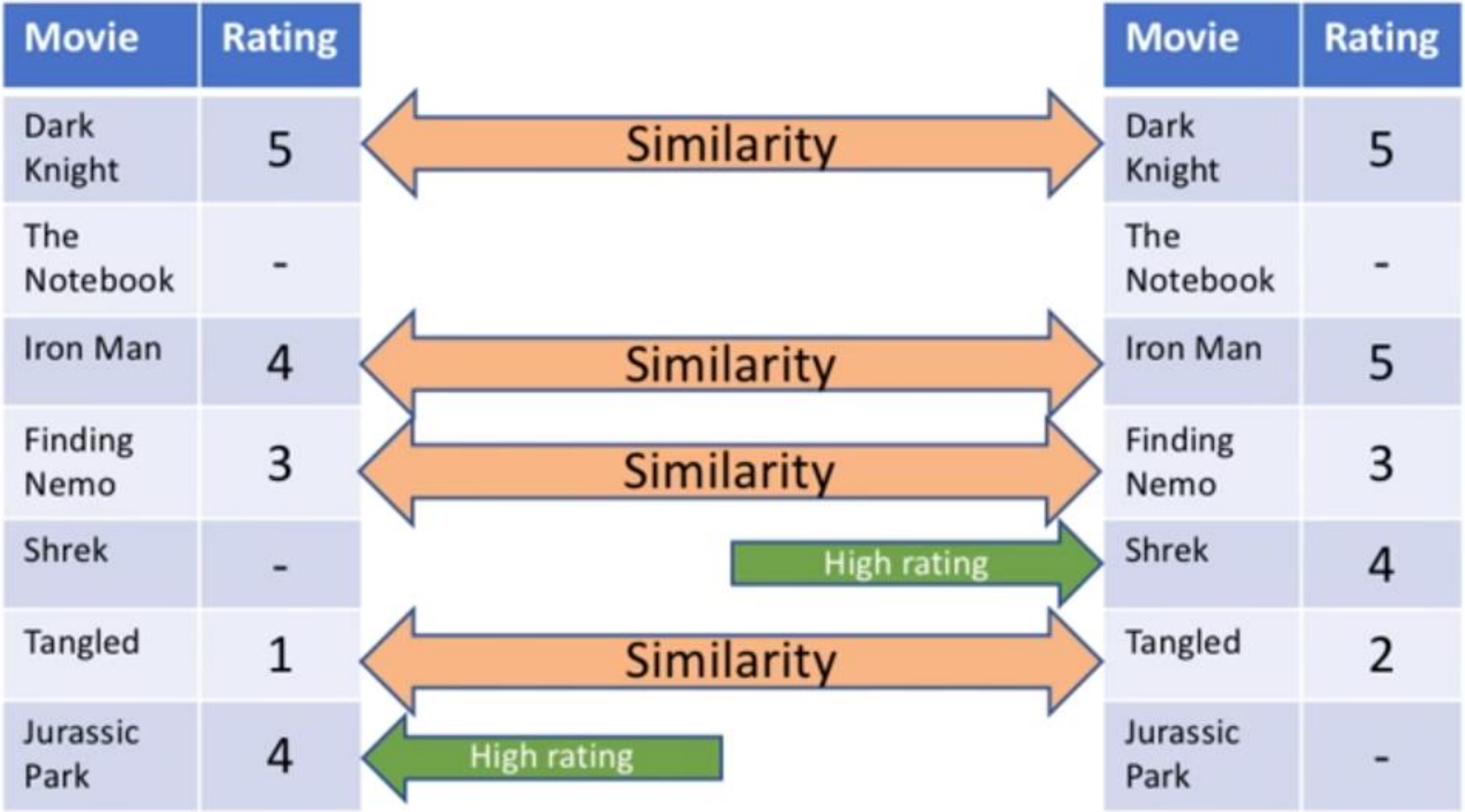
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

Recommendation Systems

- Recommendation systems are used in a wide range of applications (such as e-Commerce, social networking, or content delivery applications), to recommend new products or new content to the users (such as movies, books, Web pages etc.).
- Content-based and Collaborative filtering approaches used for recommendation systems are as follows:
- It may also be based on a hybrid of both collaborative filtering and content-based filtering to fine-tune the accuracy and effectiveness of generated suggestions.

Movie Recommendation Systems - Collaborative

- Let us now look at an example of a system for making movie recommendations using the collaborative filtering approach.



User A

Movie	Rating
Dark Knight	5
The Notebook	-
Iron Man	4
Finding Nemo	3
Shrek	-
Tangled	1
Jurassic Park	4

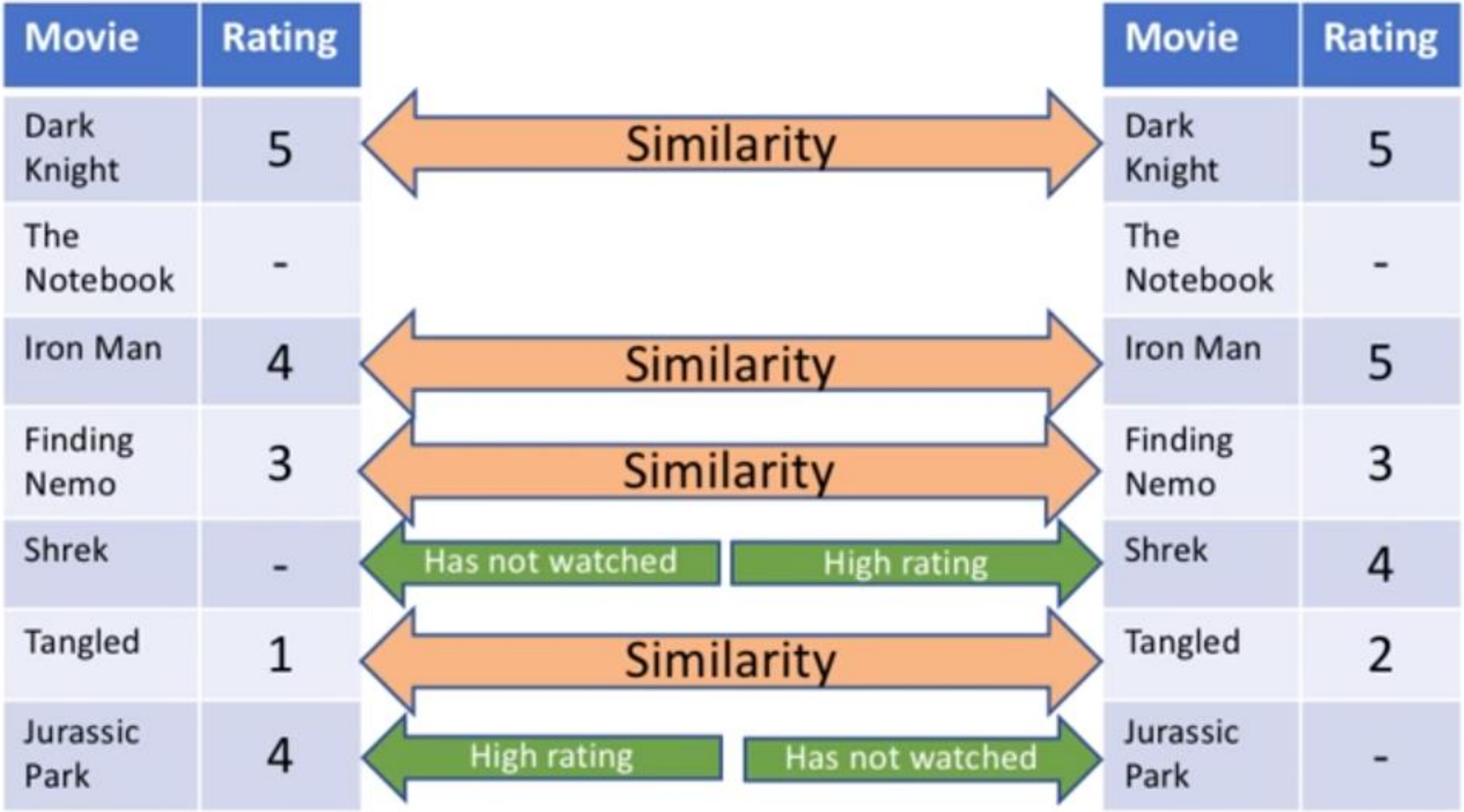
User B

Movie	Rating
Dark Knight	5
The Notebook	-
Iron Man	5
Finding Nemo	3
Shrek	4
Tangled	2
Jurassic Park	-

Movie Recommendation Systems (cont'd)

User A

User B



Movie	Rating
Dark Knight	5
The Notebook	-
Iron Man	4
Finding Nemo	3
Shrek	-
Tangled	1
Jurassic Park	4

Movie	Rating
Dark Knight	5
The Notebook	-
Iron Man	5
Finding Nemo	3
Shrek	4
Tangled	2
Jurassic Park	-

Similarity

Similarity

Similarity

Has not watched High rating

Similarity

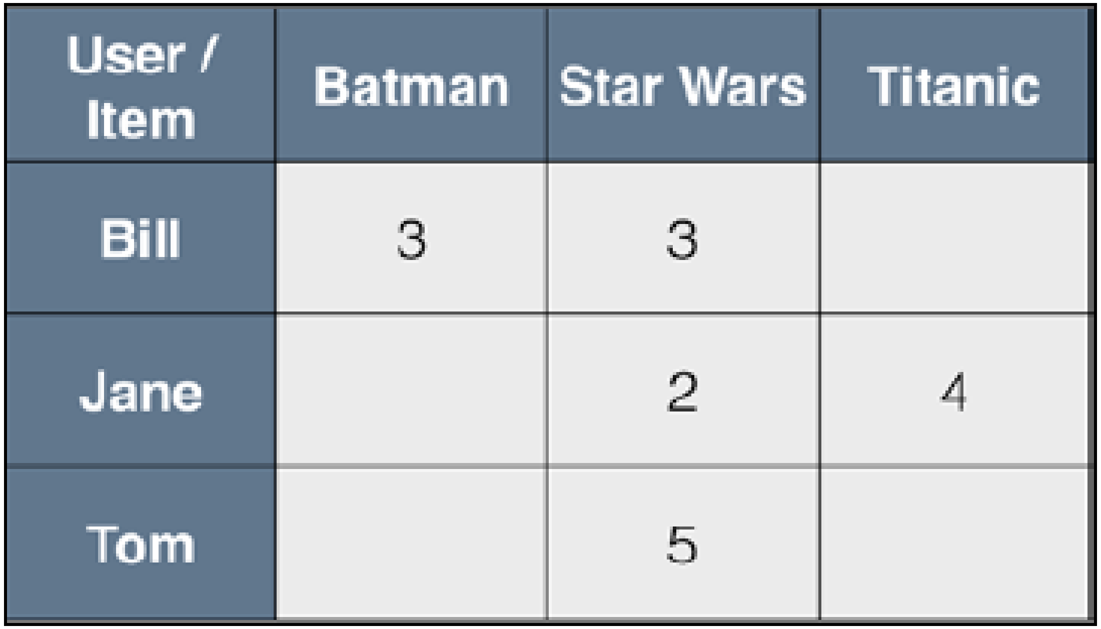
High rating Has not watched

Matrix Factorization

- Spark's recommendation models includes an implementation of matrix factorization, we will focus our attention on this class of models, they perform extremely well in collaborative filtering.
- When we deal with data that consists of preferences of users that are provided by the users themselves. This includes, for example, ratings, thumbs up, likes, and so on that are given by users to items.

Matrix Factorization – (cont'd)

- We can take these ratings and form a two-dimensional matrix with users as rows and items as columns. Each entry represents a rating given by a user to a certain item.
- Since in most cases, each user has only interacted with a relatively small set of items, this matrix has only a few non-zero entries (that is, it is very sparse).

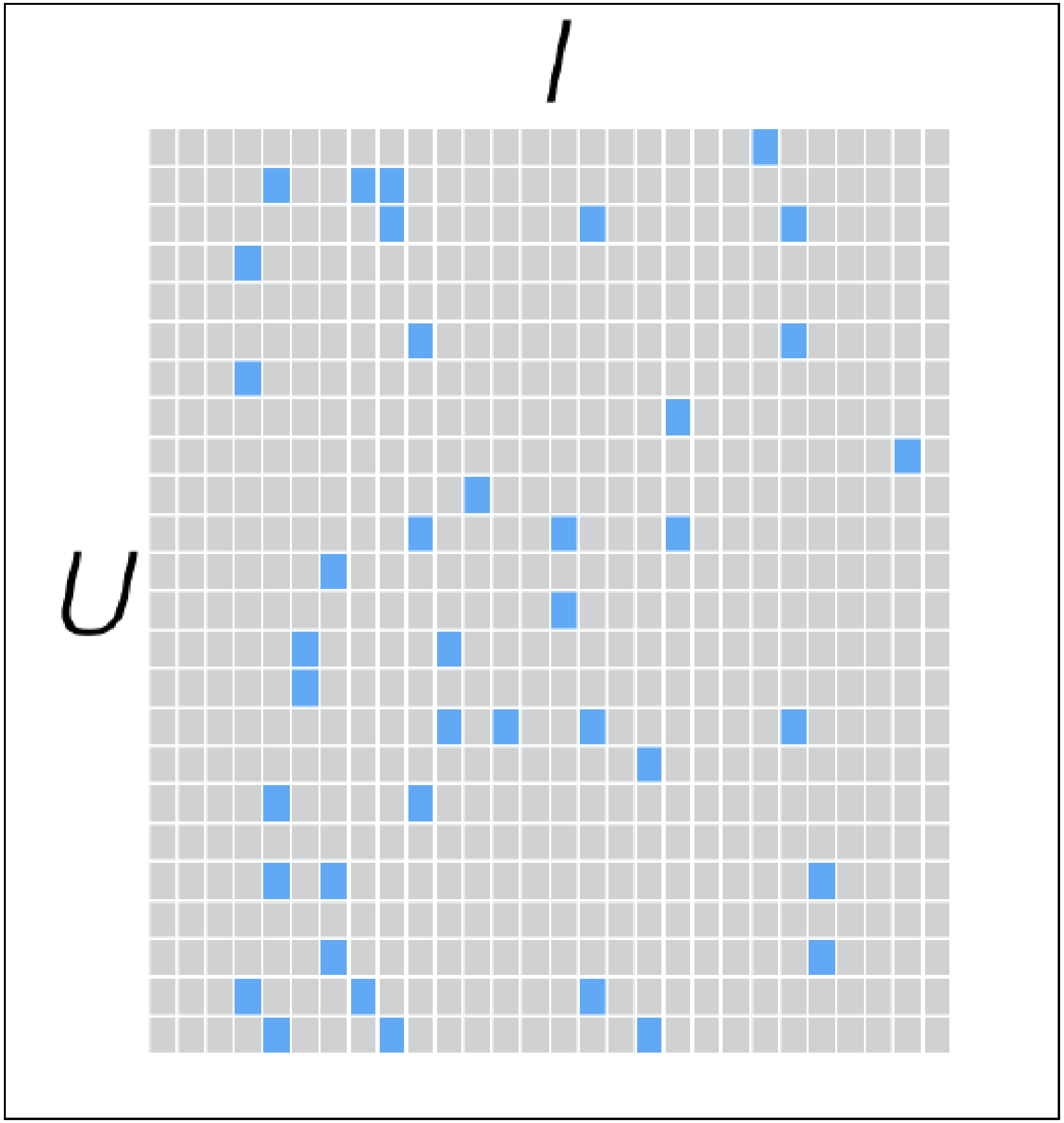


User / Item	Batman	Star Wars	Titanic
Bill	3	3	
Jane		2	4
Tom		5	

A simple movie-rating matrix

Matrix Factorization

- Matrix factorization (or matrix completion) attempts to directly model this user-item matrix by representing it as a product of two smaller matrices of lower dimension.
- Thus, it is a dimensionality-reduction technique. If we have U users and I items, then our user-item matrix is of dimension U x I and might look something like the one shown in the diagram:



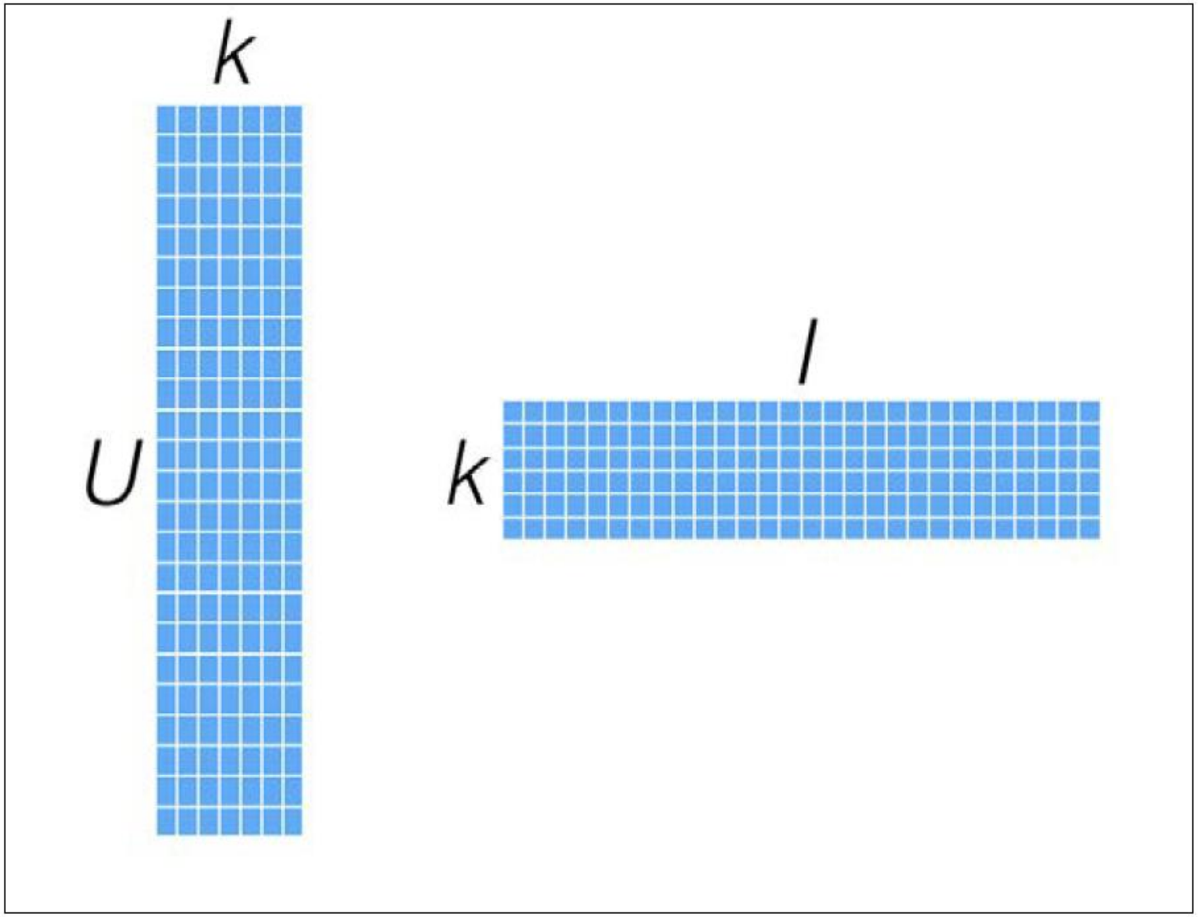
A sparse ratings matrix

Matrix Factorization

- If we want to find a lower dimension (low-rank) approximation to our user-item matrix with the dimension k, we would end up with two matrices: one for users of size U x k and one for items of size I x k. These are known as factor matrices.

- If we multiply these two factor matrices, we would reconstruct an approximate version of the original ratings matrix.

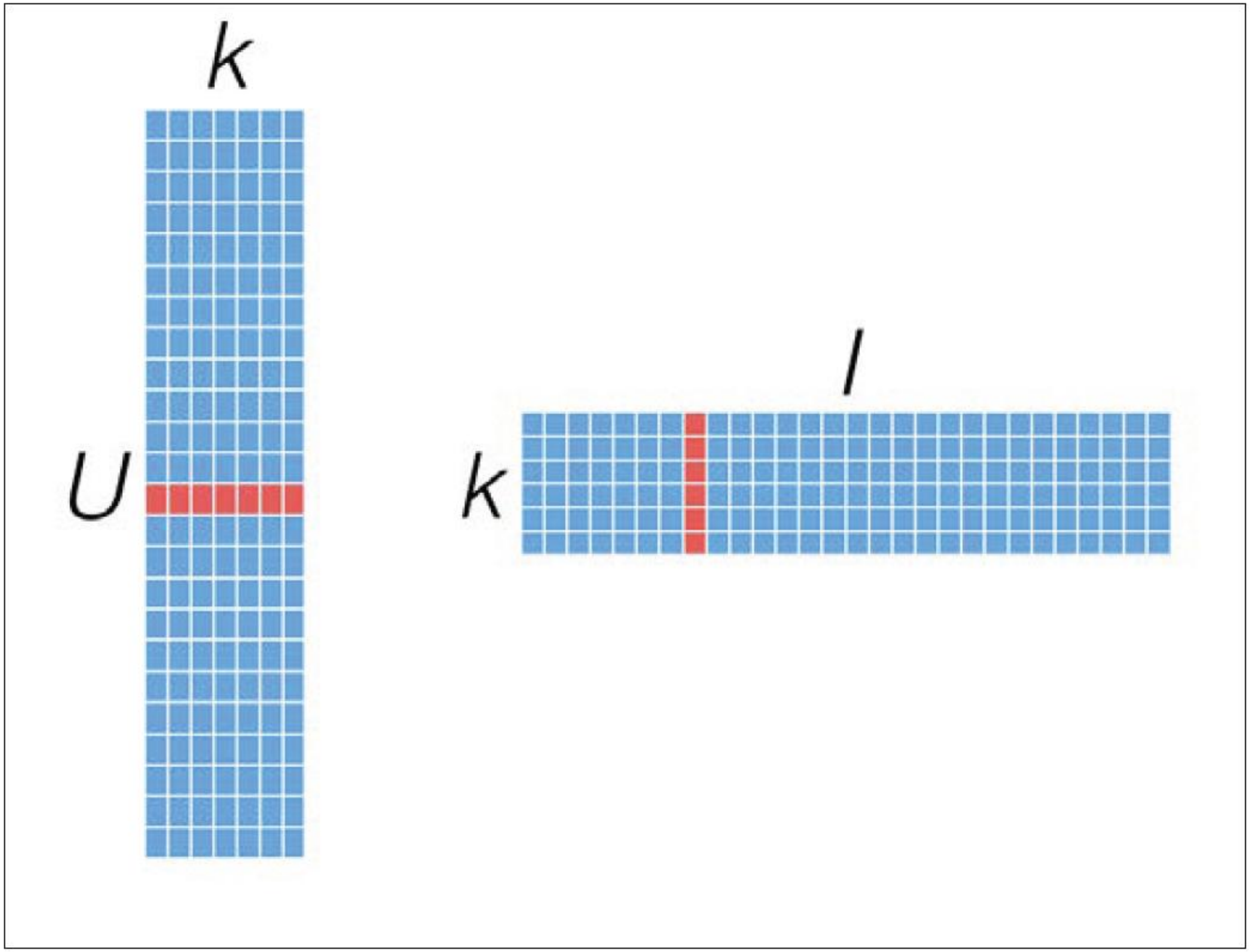
- Note that while the original ratings matrix is typically very sparse, each factor matrix is dense, as shown in the diagram:



The user- and item-factor matrices

Matrix Factorization

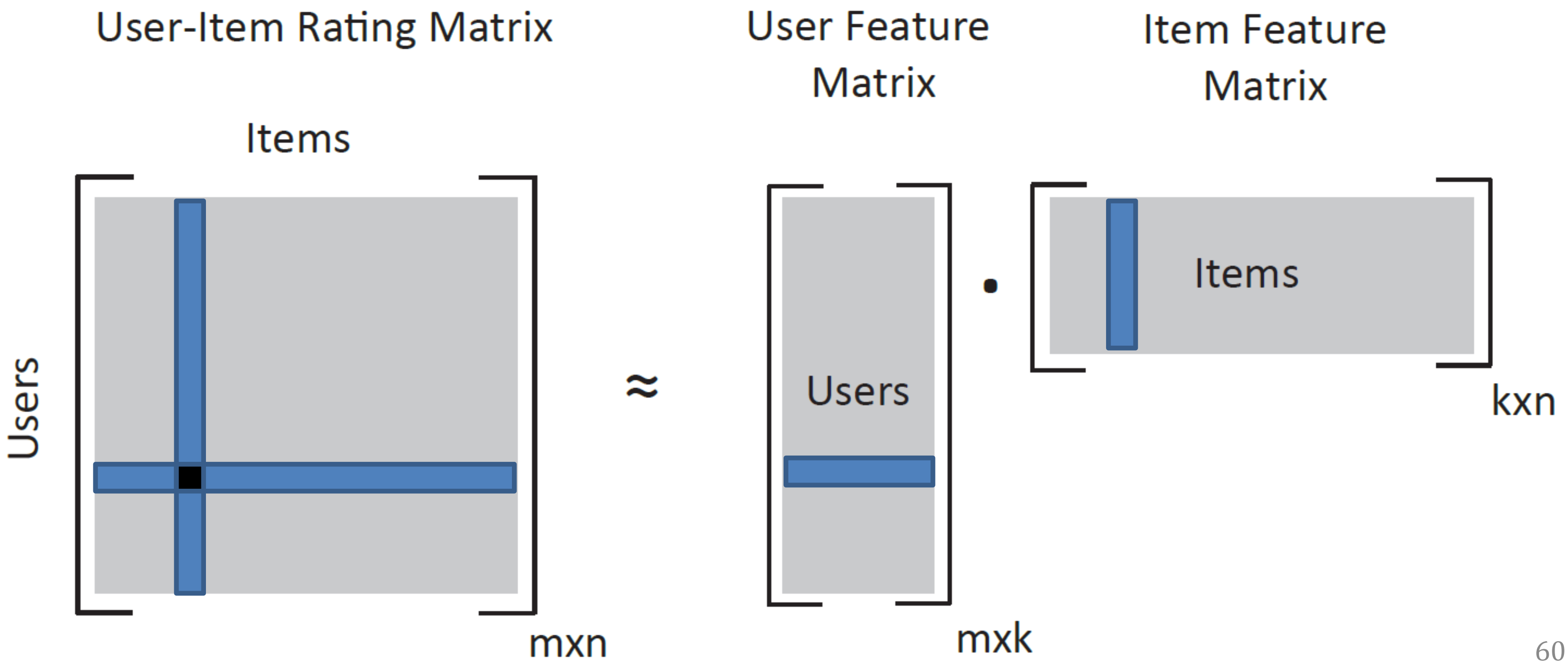
- This is illustrated with the highlighted vectors in the following diagram:



Computing recommendations from user- and item-factor vectors

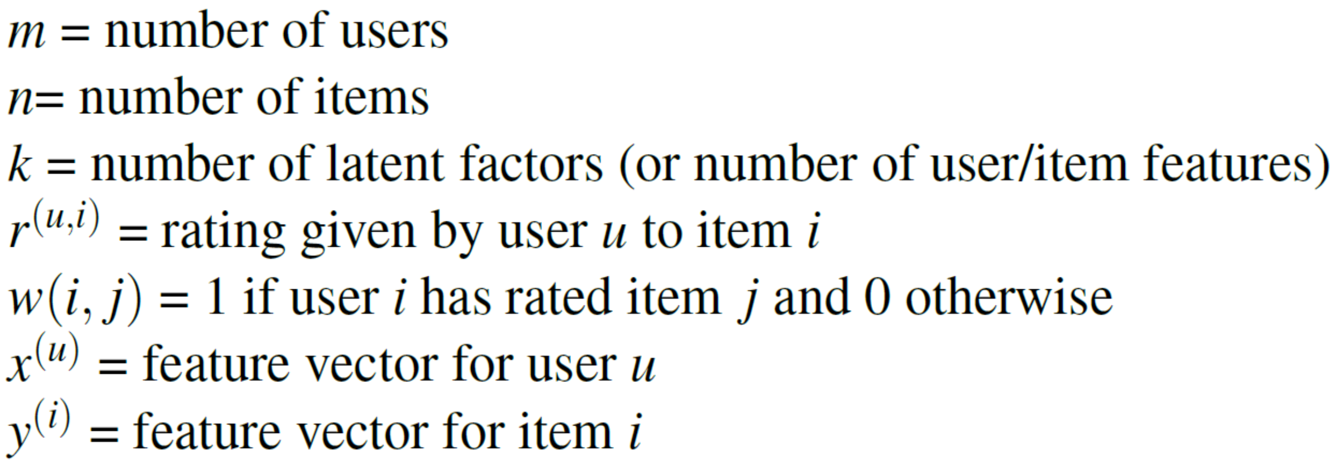
Alternating Least Squares (ALS)

- In this section, we will describe a model-based collaborative filtering approach based on Alternating Least Squares (ALS) algorithm.
- ALS works by iteratively solving a series of least squares regression problems. In each iteration, one of the user- or item-factor matrices is treated as fixed, while the other one is updated using the fixed factor and the rating data. Then, the factor matrix that was solved for is, in turn, treated as fixed, while the other one is updated. This process continues until the model has converged (or for a fixed number of iterations).



Alternating Least Squares (ALS)

- Let us formulate the collaborative filtering problem. Let



m = number of users

n= number of items

k = number of latent factors (or number of user/item features)

$r^{(u,i)}$ = rating given by user u to item i

$w(i, j)$ = 1 if user i has rated item j and 0 otherwise

$x^{(u)}$ = feature vector for user u

$y^{(i)}$ = feature vector for item i

- Figure shows (in previous slide) a user rating matrix where each row belongs to a user and the columns are the ratings given to items. Given the user-item rating matrix, the learning objective is to learn the user and item latent features (that represent the user preferences and item features).
- In other words, given an *m x n* dimensional user-item matrix, we want to factorize the matrix into an *m x k* matrix (user feature vector) and *k x n* matrix (item feature vector).

Alternating Least Squares (ALS)

To learn the user features $(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$ for all users and item features $(y^{(1)}, y^{(2)}, \ldots, y^{(n)})$ for all items, we can define the cost functions for $x^{(u)}$ and $y^{(i)}$ as follows:

$$J(x^{(1)}, x^{(2)}, \ldots, x^{(m)}) = min_{x^{(u)}} \frac{1}{2} \sum_{u=1}^{m} \sum_{i:w(i,j)=1} (x^{(u)T} y^{(i)} - r^{(u,i)})^2 + \frac{\lambda}{2} \sum_{u=1}^{m} \sum_{l=1}^{k} (x_l^{(u)})^2$$

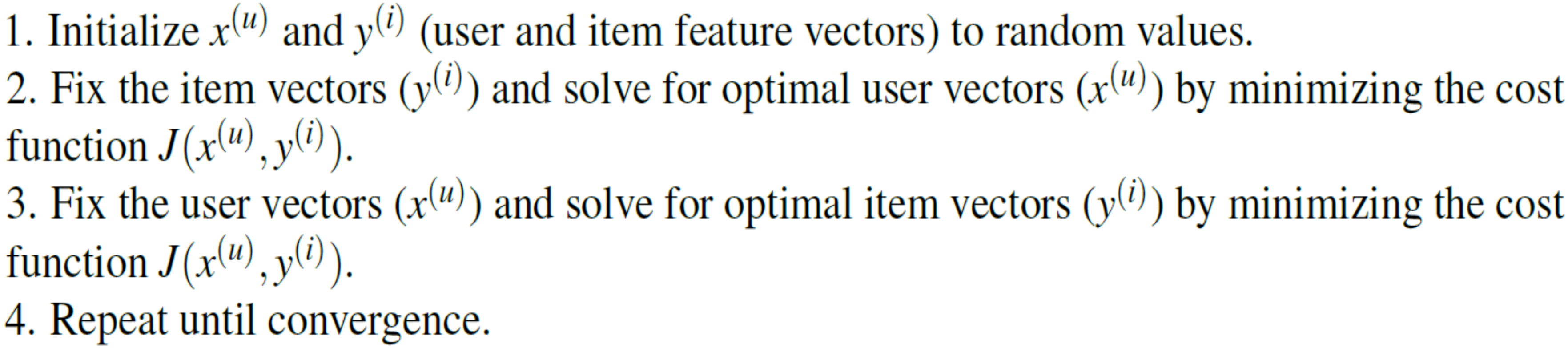
$$J(y^{(1)}, y^{(2)}, \ldots, y^{(n)}) = min_{y^{(i)}} \frac{1}{2} \sum_{i=1}^{n} \sum_{u:w(i,j)=1} (x^{(u)T} y^{(i)} - r^{(u,i)})^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{l=1}^{k} (y_l^{(i)})^2$$

where λ is the regularization parameter which is added to prevent over-fitting of data. The cost functions for $x^{(u)}$ and $y^{(i)}$ can be combined as follows:

$$J(x^{(1)}, \ldots, x^{(m)}, y^{(1)}, \ldots, y^{(n)}) = min_{(x^{(u)}, y^{(i)})} \frac{1}{2} \sum_{(u,i):w(i,j)=1} (x^{(u)T} y^{(i)} - r^{(u,i)})^2 + \frac{\lambda}{2} \left(\sum_{u=1}^{m} \sum_{l=1}^{k} (x_l^{(u)})^2 + \sum_{i=1}^{n} \sum_{l=1}^{k} (y_l^{(i)})^2 \right)$$

Alternating Least Squares (ALS)

- To solve this optimization problem, the Alternating Least Squares (ALS) algorithm can be used. The ALS algorithm is summarized as follows:



1. Initialize $x^{(u)}$ and $y^{(i)}$ (user and item feature vectors) to random values.
2. Fix the item vectors ($y^{(i)}$) and solve for optimal user vectors ($x^{(u)}$) by minimizing the cost function $J(x^{(u)}, y^{(i)})$.
3. Fix the user vectors ($x^{(u)}$) and solve for optimal item vectors ($y^{(i)}$) by minimizing the cost function $J(x^{(u)}, y^{(i)})$.
4. Repeat until convergence.

Performance Evaluation Metrics

- For a binary classification problem (with two classes Positive Class 1 and Negative Class 0) we can have four possible cases:
- (1) For a Positive class if the prediction is Positive then this is a

TruePositive - TP,

- (2) For a Positive class if the prediction is Negative then this is a

FalseNegative - FN,

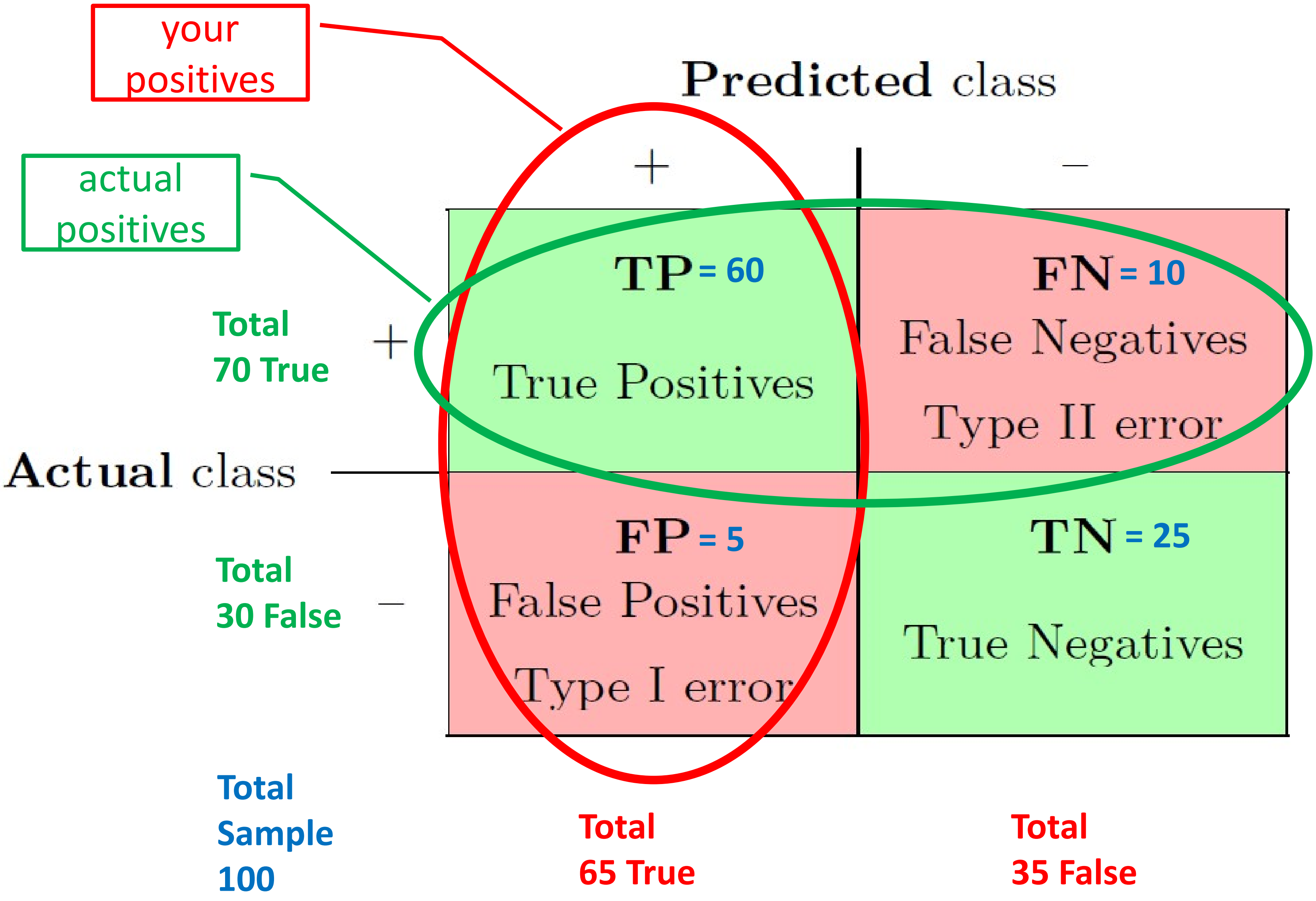
- (3) For a Negative class if the prediction is Negative then this is a

TrueNegative - TN,

- (4) For a Negative class if the prediction is Positive then this is a

FalsePositive - FP,

Confusion Matrix



your positives

actual positives

		Predicted class	
		+	−
Actual class	+	**TP** = 60 True Positives	**FN** = 10 False Negatives Type II error
	−	**FP** = 5 False Positives Type I error	**TN** = 25 True Negatives

Total 70 True

Total 30 False

Total Sample 100

Total 65 True

Total 35 False

Performance Evaluation Metrics - TPR

- The performance of classification algorithms can be evaluated using the following metrics:
- **True Positive Rate (TPR)/ Sensitivity / Recall:** True Positive Rate (TPR) also called Sensitivity or Recall is the fraction of the positives which are classified correctly. Recall is defined as the number of true positives divided by the sum of true positives and false negatives (that is, the number of examples that were in class 1, but were predicted as class 0 by the model). We can see that a recall of 1.0 (or 100 percent) is achieved if the model doesn't miss any examples that were in class 1 (that is, there are no false negatives).

$$TPR = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

Performance Evaluation Metrics - Precision

- **Precision:** Precision is the fraction of objects that are classified correctly.
- In the binary classification context, precision is defined as the number of true positives (that is, the number of examples correctly predicted as class 1) divided by the sum of true positives and false positives (that is, the number of examples that were incorrectly predicted as class 1). Thus, we can see that a precision of 1.0 (or 100 percent) is achieved if every example predicted by the classifier to be class 1 is, in fact, in class 1 (that is, there are no false positives).
- Precision is defined as,

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

Performance Evaluation Metrics – Precision vs Recall

- Generally, precision and recall are inversely related; often, higher precision is related to lower recall and vice versa. To illustrate this, assume that we built a model that always predicted class 1. In this case, the model predictions would have no false negatives because the model always predicts 1; it will not miss any of class 1. Thus, the recall will be 1.0 for this model. On the other hand, the false positive rate could be very high, meaning precision would be low (this depends on the exact distribution of the classes in the dataset).

Performance Evaluation Metrics – Precision vs Recall

- Precision and recall are not particularly useful as standalone metrics, but are typically used together to form an aggregate or averaged metric. Precision and recall are also dependent on the threshold selected for the model.
- Intuitively, below some threshold level, a model will always predict class 1. Hence, it will have a recall of 1, but most likely, it will have low precision. At a high enough threshold, the model will always predict class 0. The model will then have a recall of 0, since it cannot achieve any true positives and will likely have many false negatives. Furthermore, its precision score will be undefined, as it will achieve zero true positives and zero false positives.

Performance Evaluation Metrics - TNR

- **True Negative Rate (TNR)/ Specificity**: True Negative Rate (TPR) also called Specificity is the fraction of the negatives which are classified correctly.

$$TNR = \frac{TrueNegative}{(TrueNegative + FalsePositive)}$$

Performance Evaluation Metrics - FPR

- **False Positive Rate (FPR):** False Positive Rate (FPR) is defined as,

$$FPR = \frac{FalsePositive}{(FalsePositive + TrueNegative)}$$

Performance Evaluation Metrics – Accuracy and F1-Score

- **Accuracy:** Accuracy is defined as,

$$Accuracy = \frac{(TruePositive + TrueNegative)}{(TruePositive + TrueNegative + FalsePositive + FalseNegative)}$$

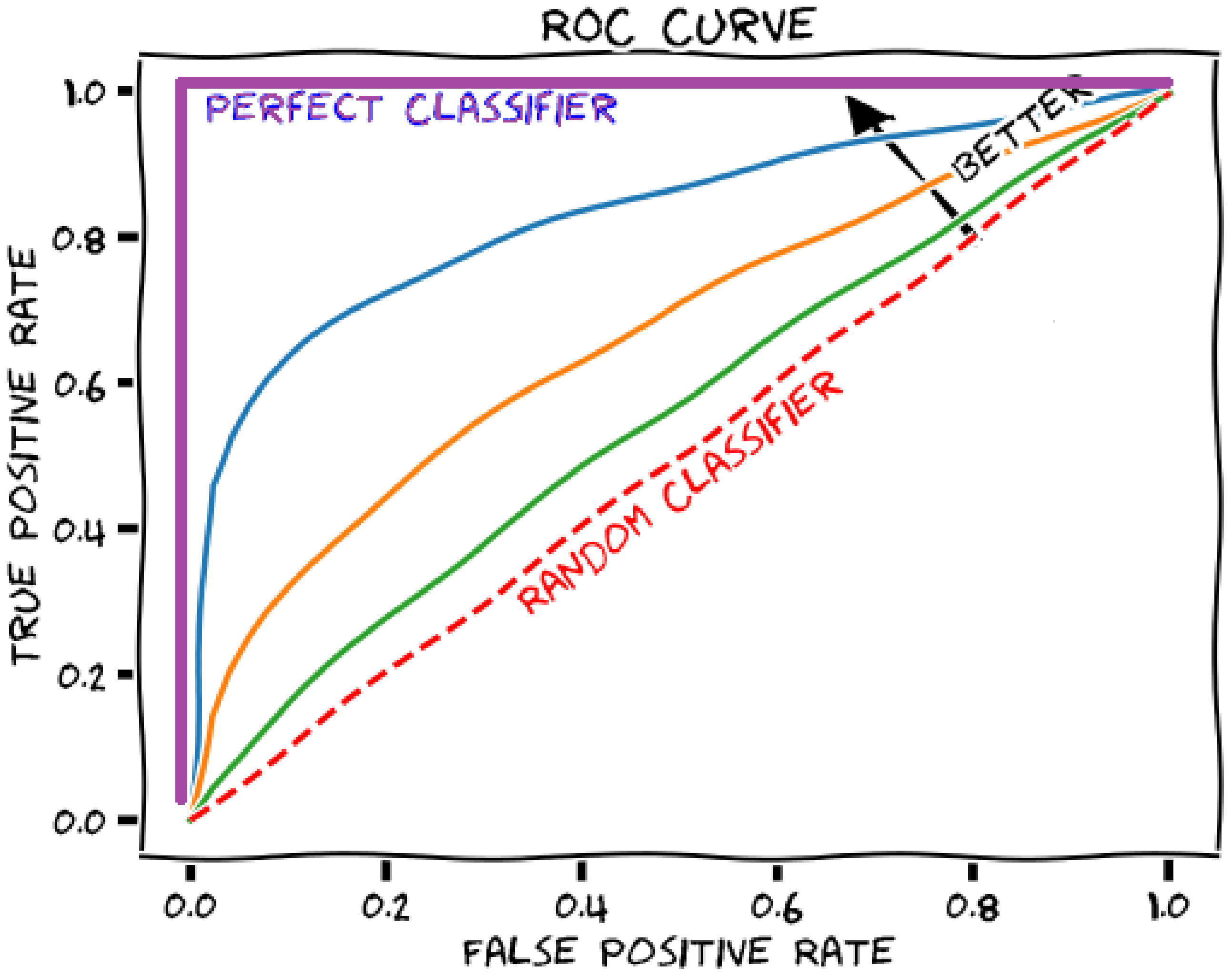
- **F1-score:** F1-score is a measure of accuracy that considers both precision and recall. F1-score is the harmonic means of precision and recall given as,

$$F1 - Score = \frac{2(Precision)(Recall)}{(Precision + Recall)}$$

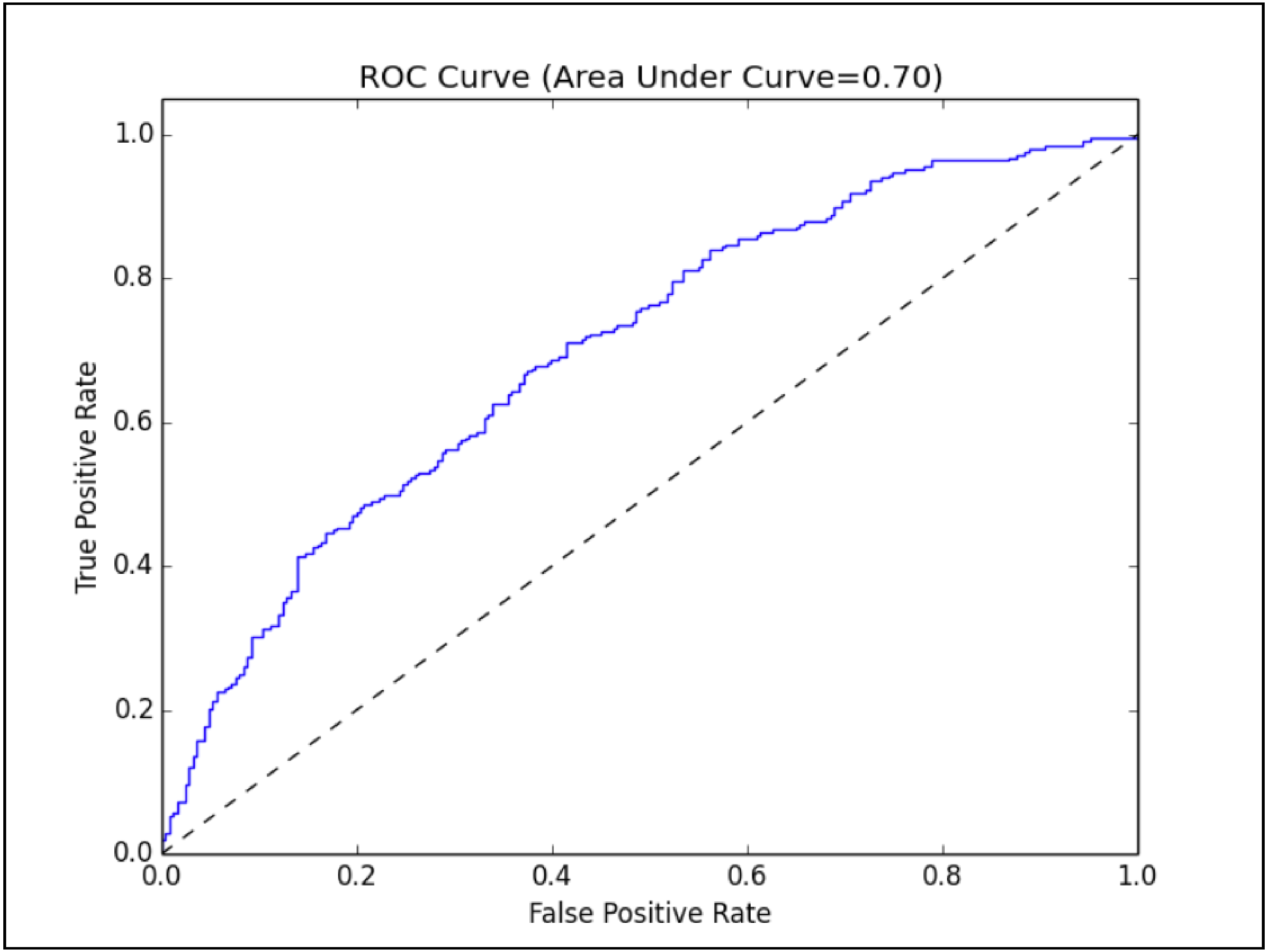
Performance Evaluation Metrics – ROC and AUC

- **Receiver Operating Characteristics (ROC) Curve:** ROC curve is the plot of the True Positive Rate (TPR) versus the False Positive Rate (FPR). For different values of the discrimination threshold (threshold for the probability above which we choose a positive class), we get a number of pairs of (TPR, FPR) values.
- In a manner similar to precision and recall, the ROC curve (plotted in the following figure) represents the classifier's performance tradeoff of TPR against FPR, for different decision thresholds. Each point on the curve represents a different threshold in the decision function for the classifier.
- **Area Under Curve (AUC):** AUC is the area under the ROC curve.

Performance Evaluation Metrics – ROC



Performance Evaluation Metrics – ROC and AUC



The ROC curve

Performance Evaluation Metrics - MSE

- **Mean Squared Error (MSE):** Mean Squared Error is the mean of the sum of the square of the errors between the estimated and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(h(x^{(i)}) - y^{(i)})^2$$

Performance Evaluation Metrics – R-Squared

- **Coefficient of Determination (R^2):** Coefficient of Determination also called R^2 or R-Squared, is a measure of how well the model is able to explain the variation of the data. R^2 is defined as,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - h(x^{(i)}))^2}{\sum_{i=1}^{n}(y^{(i)} - \mu)^2} = 1 - \frac{SSE}{SST}$$

- where SSE is the residual sum of squares and SST is the total sum of squares. R^2 varies between 0 and 1. R^2=1 means that the model explains all the variability of the data around its mean.