FORMERLY Industries Helping Hands

THE iSCALE

# Data Science | 30 Days of Machine Learning | Day - 22

Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

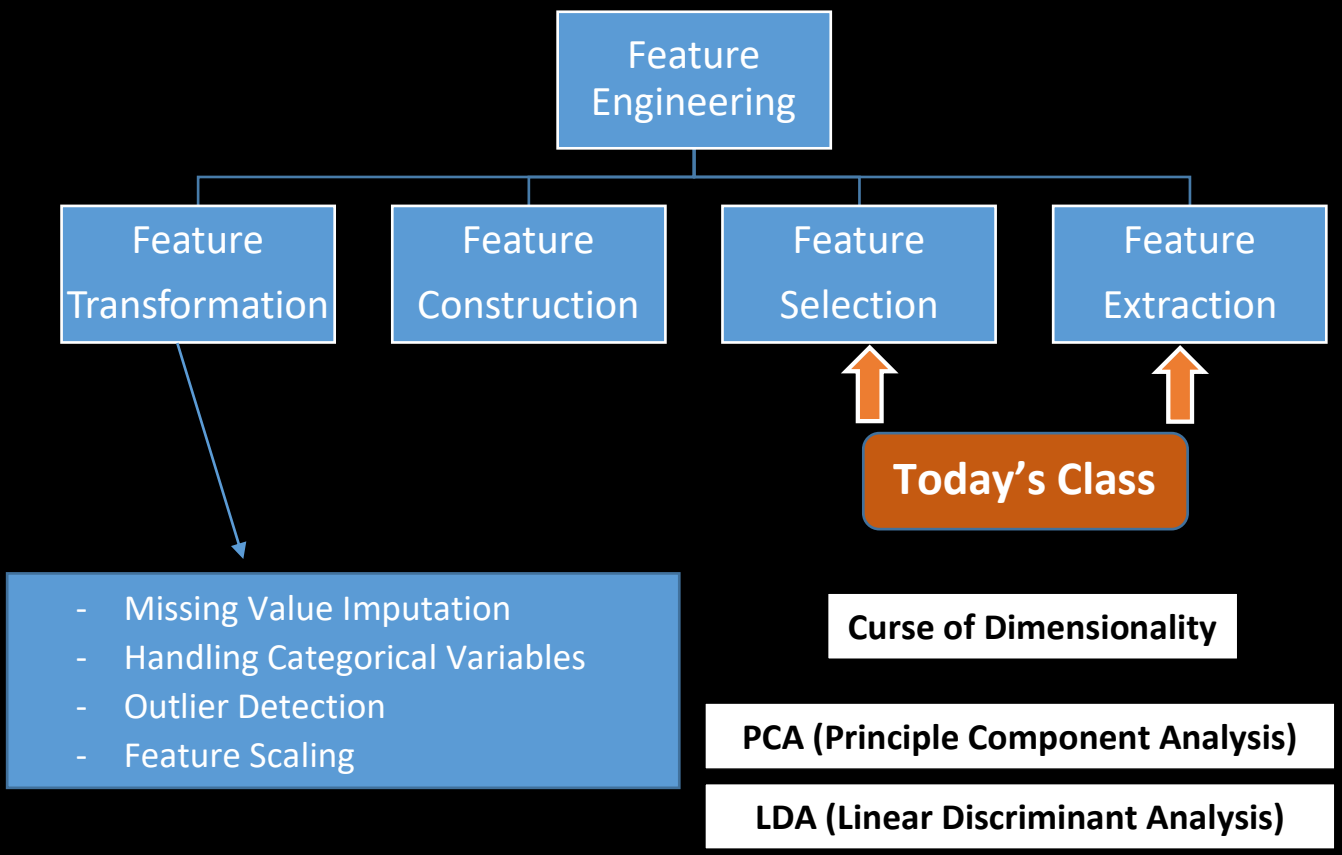
## ----Today Topics | Day 22----

**Feature Selection | Feature Extraction | Curse of Dimensionality**

---

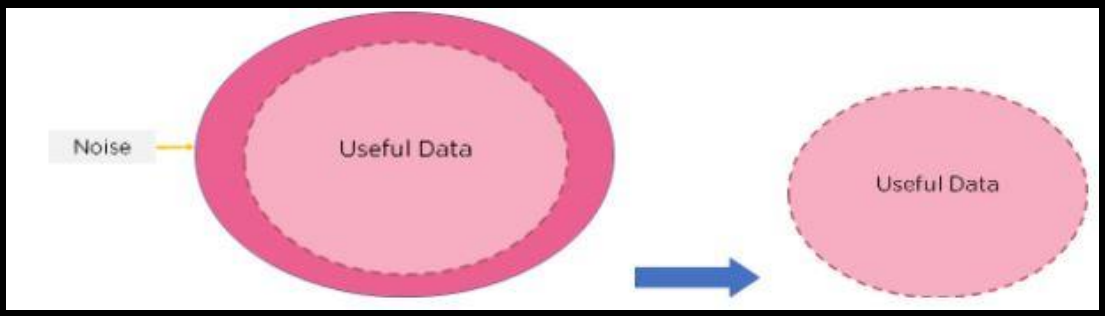
- **What is "Feature Selection" in machine learning?**
- **Why "Feature Selection" is Important?**
- **Type of Feature Selection Models.**
- **What is "Feature Extraction" in machine learning?**
- **Discuss about "Curse of Dimensionality" concept.**

- Dataset Link GitHub: https://github.com/TheiScale/30_Days_Machine_Learning/



## - What is “Feature Selection” in machine learning?

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.




## - Why “Feature Selection” is Important?

Machine learning models follow a simple rule: whatever goes in, comes out. If we put garbage into our model, we can expect the output to be garbage too. In this case, garbage refers to noise in our data.

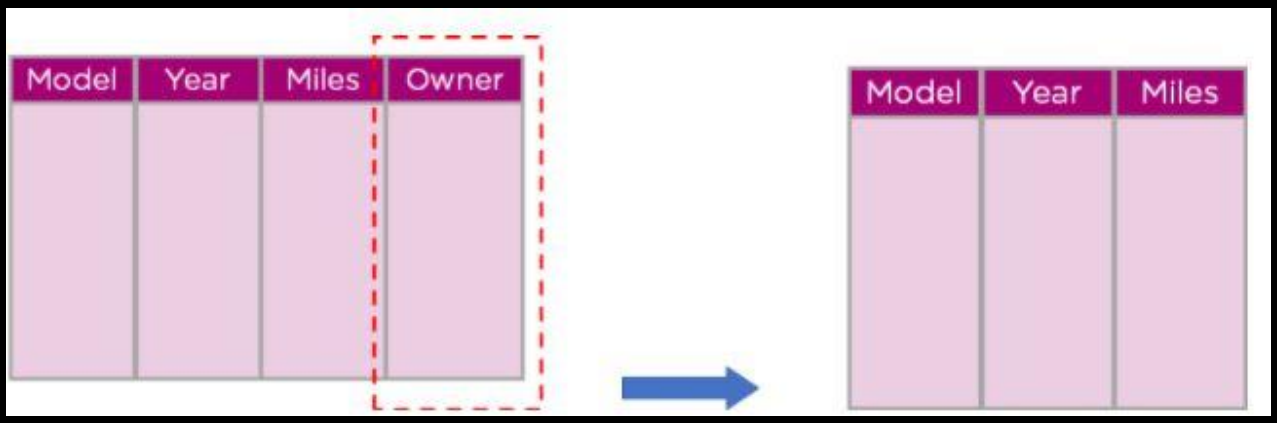
To train a model, we collect enormous quantities of data to help the machine learn better. Usually, a good portion of the data collected is noise, while some of the columns of our dataset might not contribute significantly to the performance of our model. Further, having a lot of data can slow down the training process and cause the model to be slower. The model may also learn from this irrelevant data and be inaccurate.

## Example:

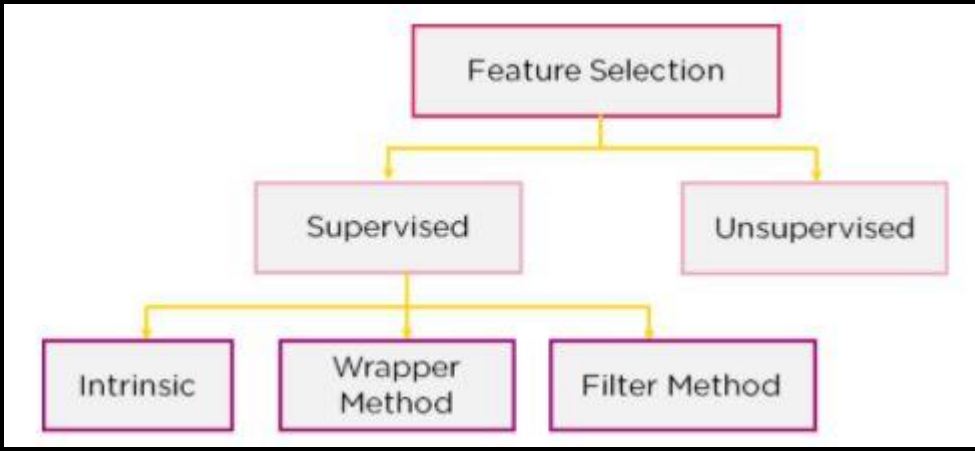
Consider a table which contains information on old cars. The model decides which cars must be crushed for spare parts.

| Model | Year | Miles | Owner |
|---|---|---|---|
| | | | |

In the above table, we can see that the model of the car, the year of manufacture, and the miles it has travelled are pretty important to find out if the car is old enough to be crushed or not. However, the name of the previous owner of the car does not decide if the car should be crushed or not. Further, it can confuse the algorithm into finding patterns between names and the other features. Hence we can drop the column.



## Type of Feature Selection Models:

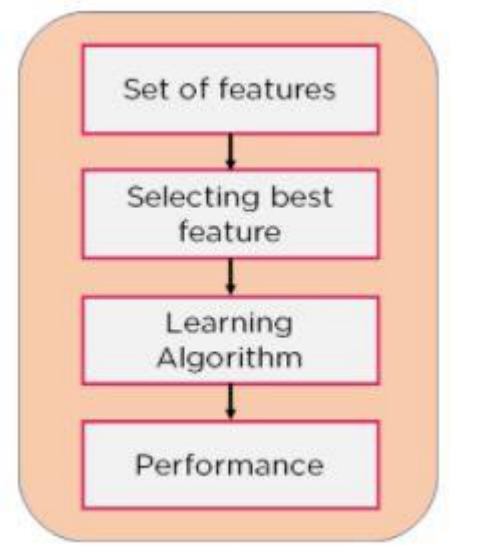



Feature selection models are of two types:

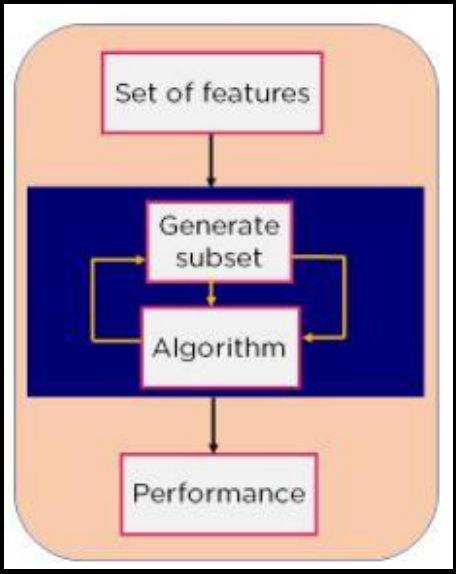
**Supervised Models:** Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model

**Unsupervised Models:** Unsupervised feature selection refers to the method which does not need the output label class for feature selection. We use them for unlabelled data.

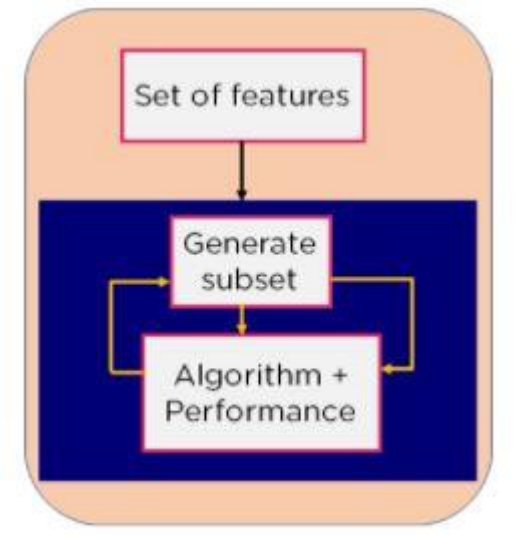
1. **Filter Method:** In this method, features are dropped based on their relation to the output, or how they are correlating to the output. We use correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly. Eg: Information Gain, Chi-Square Test, Fisher's Score, etc.




2. **Wrapper Method:** We split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again. It forms the subsets using a greedy approach and evaluates the accuracy of all the possible combinations of features. Eg: Forward Selection, Backwards Elimination, etc.



3. **Intrinsic Method:** This method combines the qualities of both the Filter and Wrapper method to create the best subset.



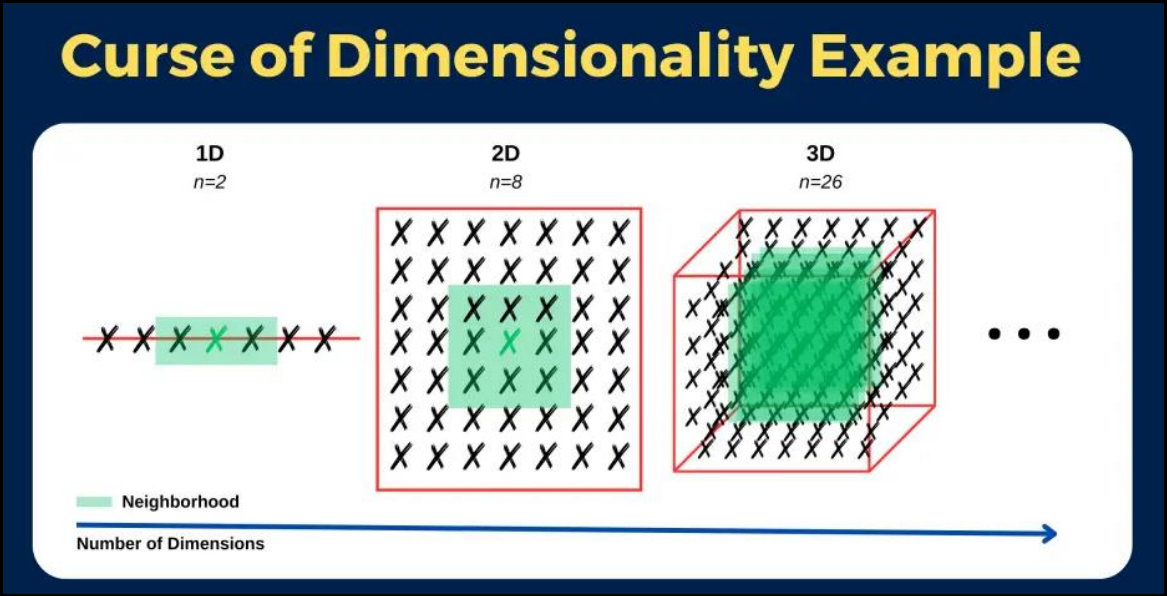

## The Curse of Dimensionality:

### What are dimensions?

**In the context of data analysis and machine learning, dimensions refer to the features or attributes of data. For instance, if we consider a dataset of houses, the dimensions could include the house's price, size, number of bedrooms, location, and so on.**

### How does the curse of dimensionality occur?

**As we add more dimensions to our dataset, the volume of the space increases exponentially. This means that the data becomes sparse. Think of it this way: if you have a line (1D), it's easy to fill it with a few points. If you have a square (2D), you need more points to cover the area. Now, imagine a cube (3D) - you'd need even more points to fill the space. This concept extends to higher dimensions, making the data extremely sparse.**




## Higher Dimension & Data Sparsity:

**As mentioned, data becomes sparse, meaning that most of the high-dimensional space is empty. This makes clustering and classification tasks challenging.**

1. **Performance Decreases**
2. **Computation**

**Dimensionality Reduction**

**Feature Section**

- Forward Selection
- Backward Elimination

**Feature Extraction**

- PCA (Principal Component Analysis)
- LDA (Linear Discriminant Analysis)

FORMERLY Industries Helping Hands

THE iSCALE

## Day 22 | Data Curious Minds

# How Netflix uses Data Science

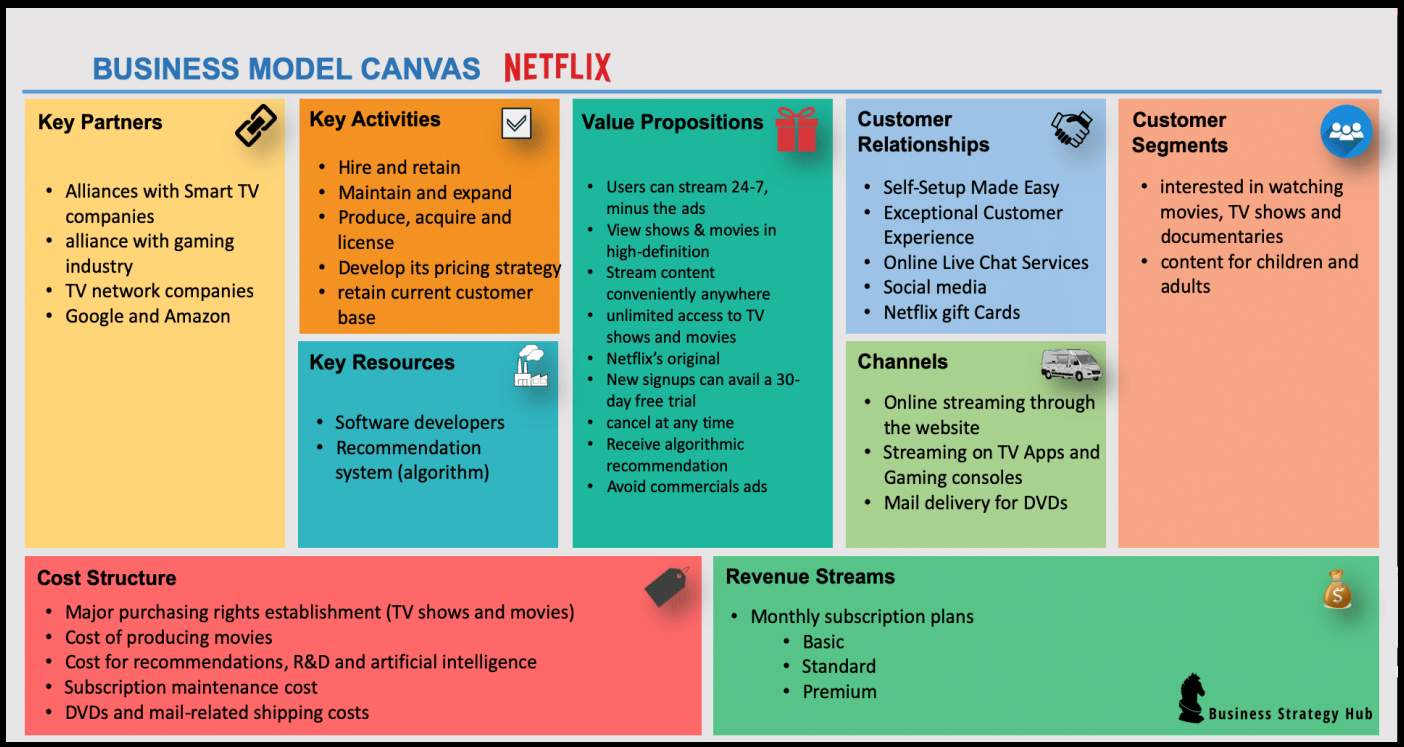
Netflix ended 2022 with around 231 million global paid subscribers

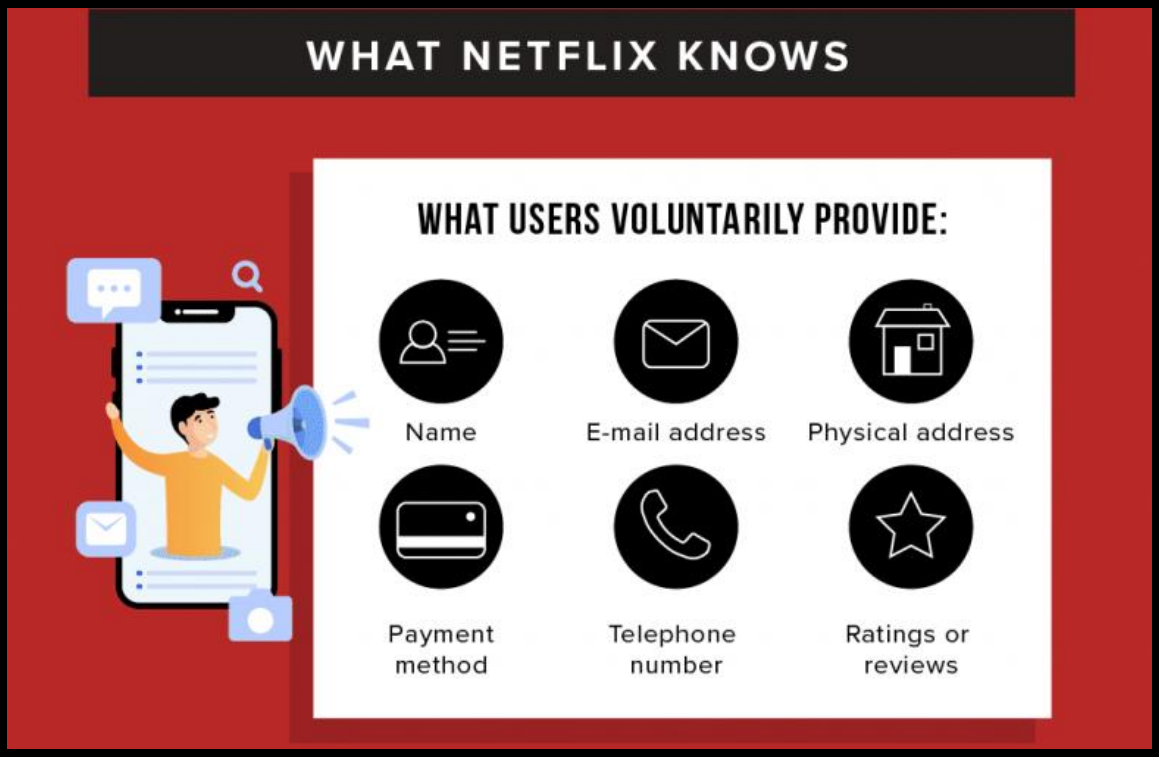
Number of employees: 12,800 (2022)

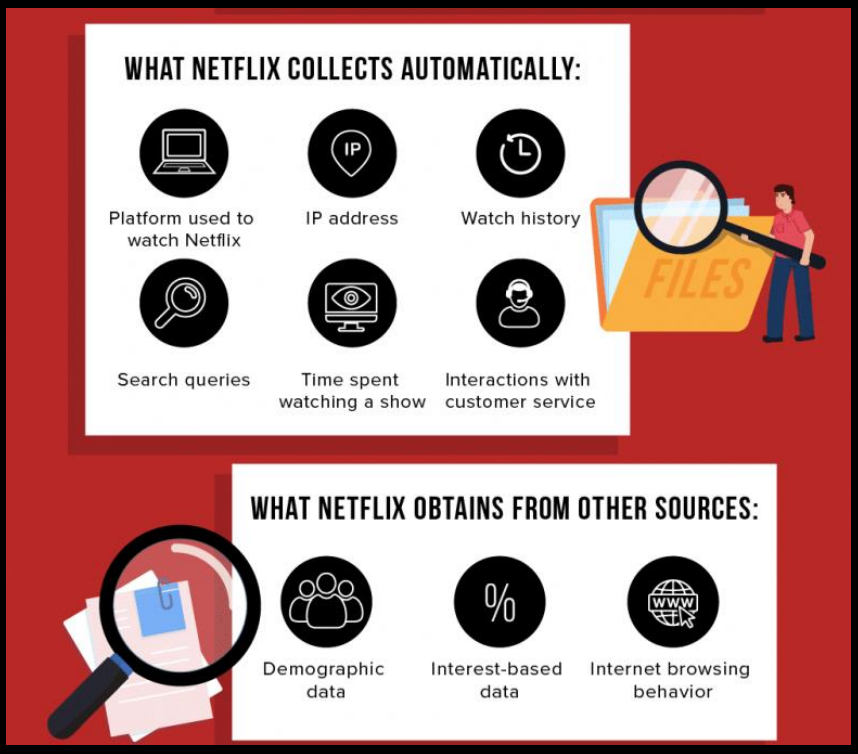
Revenue: 3,160 crores USD (2022)

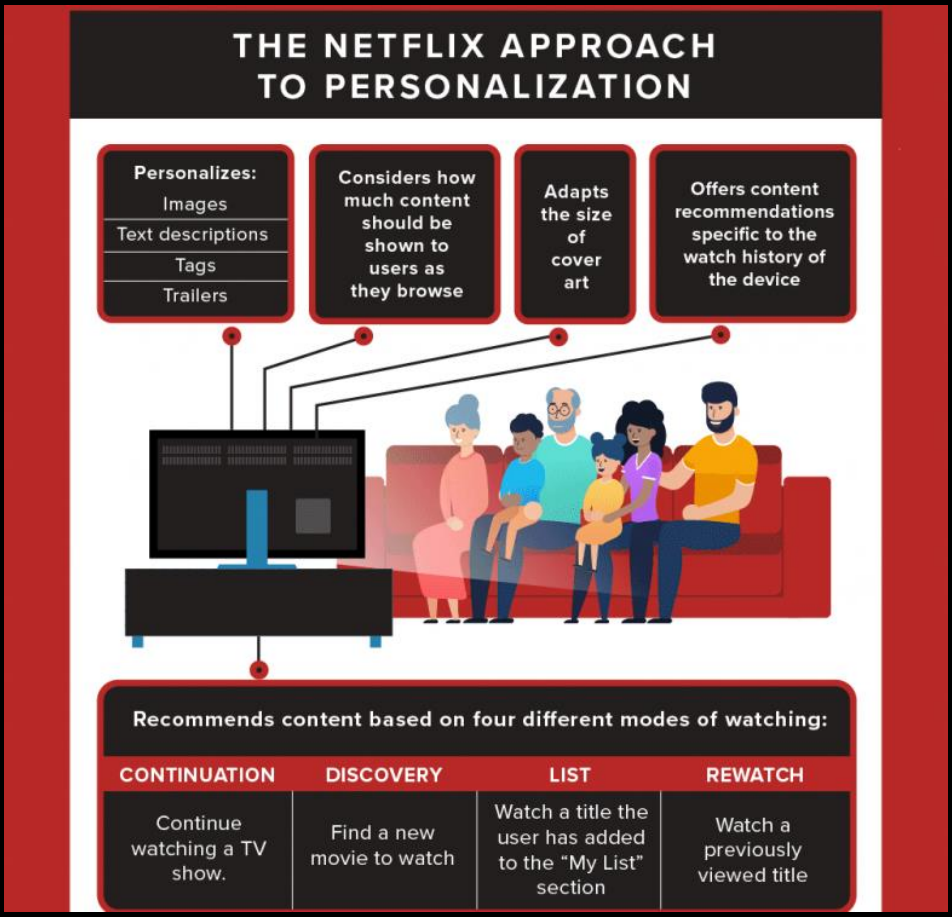
Founded: 29 August 1997, Scotts Valley, California

Founders: Reed Hastings, Marc Randolph




WHAT NETFLIX KNOWS
WHAT USERS VOLUNTARILY PROVIDE:
Name
E-mail address
Physical address
Payment method
Telephone number
Ratings or reviews

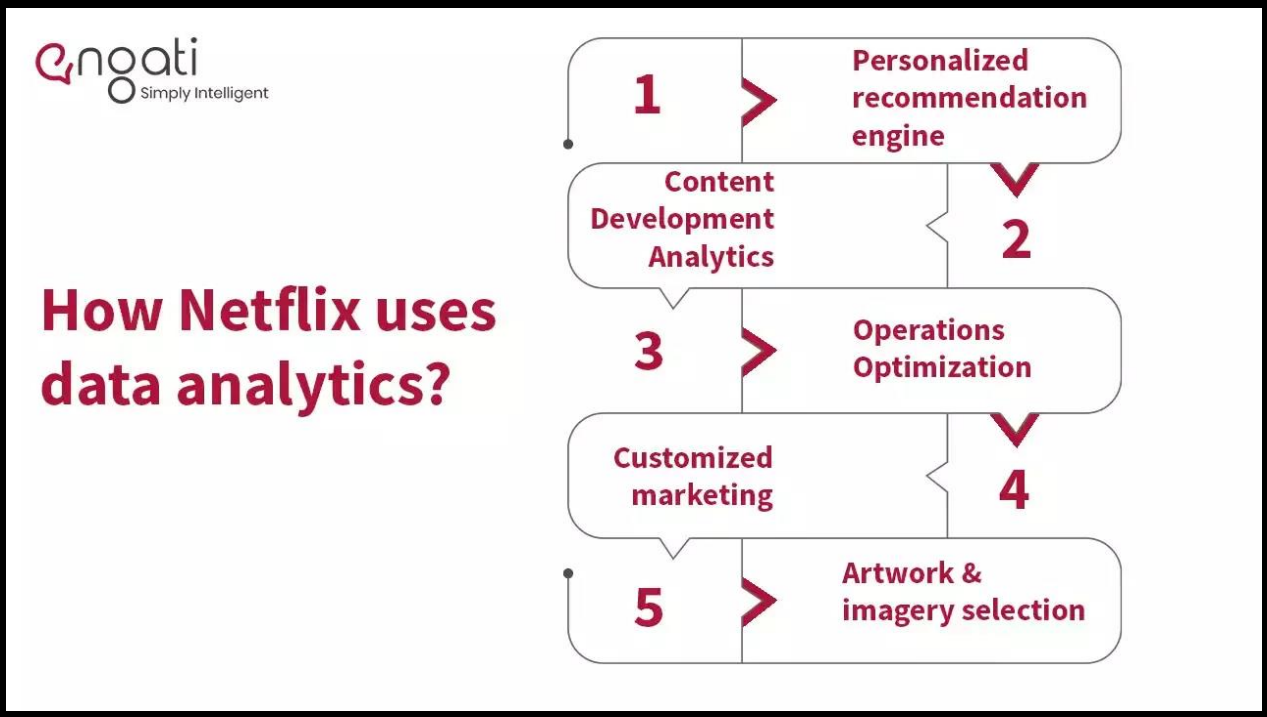

WHAT NETFLIX COLLECTS AUTOMATICALLY:
Platform used to watch Netflix
IP address
Watch history
Search queries
Time spent watching a show
Interactions with customer service
FILES
WHAT NETFLIX OBTAINS FROM OTHER SOURCES:
Demographic data
Interest-based data
Internet browsing behavior




## Personalized movie recommendation

Netflix collects information about your viewing habits, including the date and time you watched a show or movie. This data can be used to recall you based on the device you used to manage the show or movie. It can also be used to rate what you watched. Netflix also keeps track of what movies and shows the users watch to analyze various aspects of their customers' behavior, such as their viewing habits. This data is then used to create a personalized viewing experience for each customer by offering the most relevant content for each individual

- What day you watch content
- What time you watch content
- The device on which the content was watched
- How the nature of the content?
- Searches on the platform
- Portions of content that got re-watched
- Whether content was paused, rewind, or fast forward
- User location data
- When you leave content
- The ratings given by the users
- Browsing and scrolling behavior



Netflix's success highlights the value of data analytics because it presents an incredible insight into user's preferences allowing them to make smart decisions that deliver maximum ROI on their choices.

"Where there is data smoke, there is business fire."
— Thomas Redman

The beauty of Netflix is on the 28th of October they push a button and the film will be in 190 countries at the same moment in 17 languages.


# Aesthetic Vision Analysis


Power Bi Project Video Link:

https://youtu.be/ds69YbxuGVE