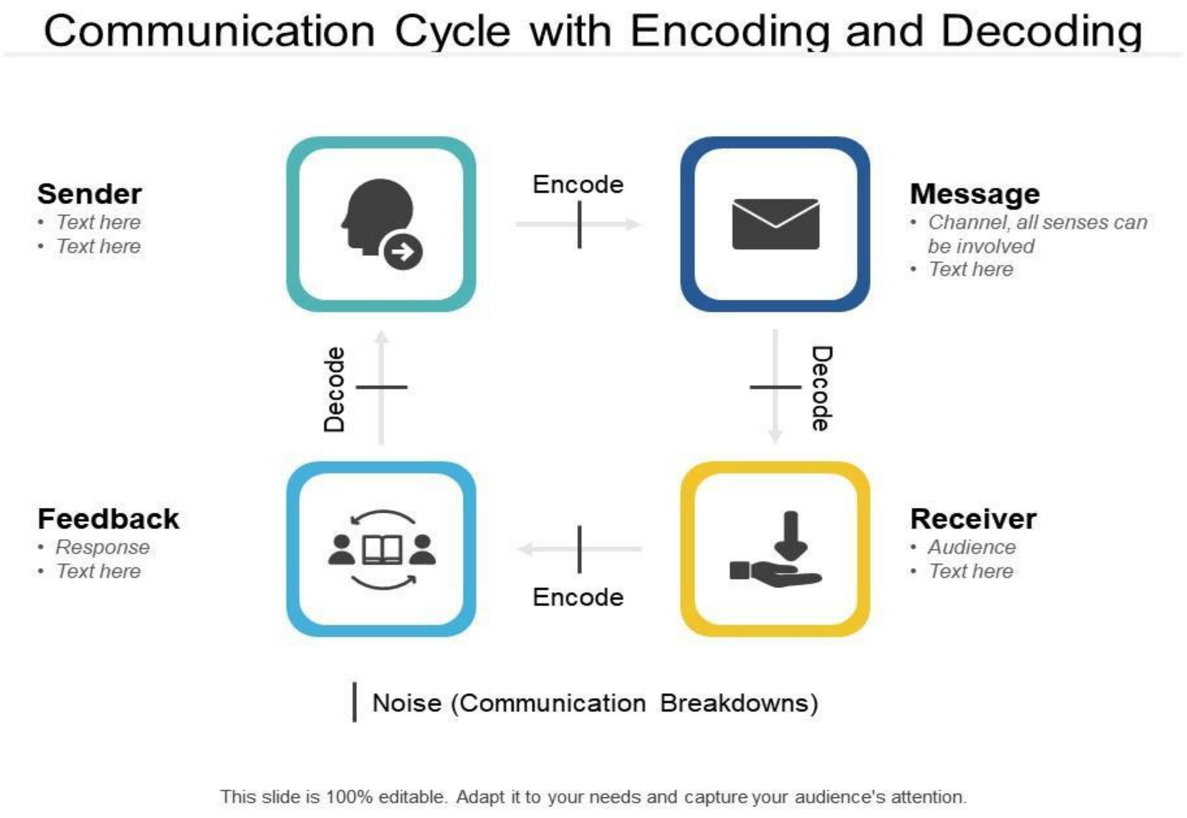
Encoding(Simple way)

The process of conversion of data from one form to another form is known as ***Encoding****.* It is used to transform the data so that data can be supported and used by different systems. Encoding works similarly to converting temperature from centigrade to Fahrenheit, as it just gets converted in another form, but the original value always remains the same. Encoding is used in mainly two fields:

- **Encoding in Electronics:** In electronics, encoding refers to converting analog signals to digital signals.
- **Encoding in Computing:** In computing, encoding is a process of converting data to an equivalent cipher by applying specific code, letters, and numbers to the data.



Communication Cycle with Encoding and Decoding

Sender
- *Text here*
- *Text here*

Message
- *Channel, all senses can be involved*
- *Text here*

Feedback
- *Response*
- *Text here*

Receiver
- *Audience*
- *Text here*

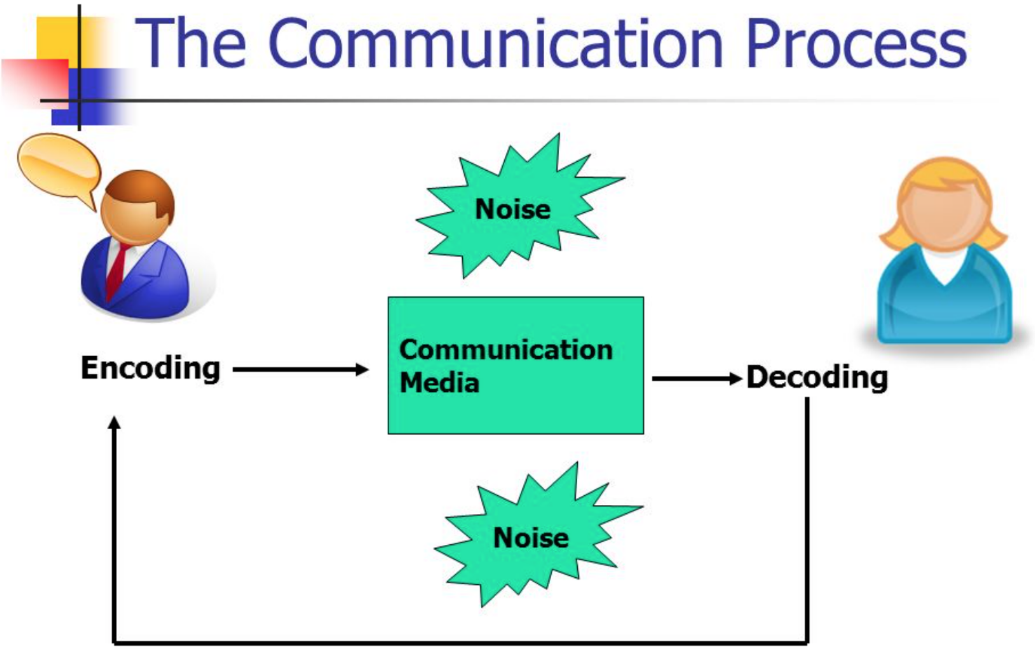
Noise (Communication Breakdowns)

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

- What do you mean encoding?

In computers, encoding is the process of putting a sequence of characters **(letters, numbers, punctuation, and certain symbols)** into a specialized format for efficient transmission or storage. Decoding is the opposite process -- the conversion of an encoded format back into the original sequence of characters.

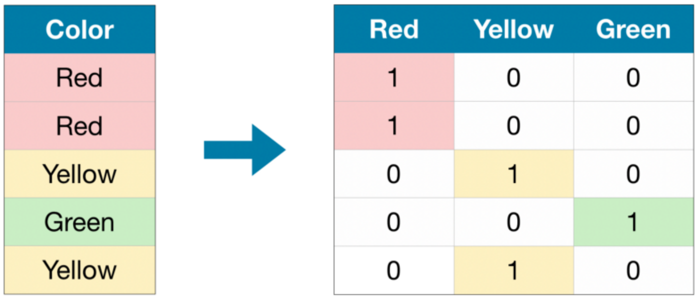
- What is an example of encoding?



For example, you may realize you're hungry and encode the following message to send to your roommate: **“I'm hungry. Do you want to get pizza tonight?”** As your roommate receives the message, they decode your communication and turn it back into thoughts to make meaning.

1. Nominal Variable

When we have a feature where variables are just names and there is no order or rank to this variable's feature.



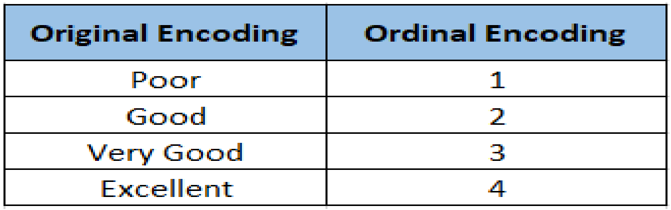
Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

For example: City of person lives in, Gender of person, Marital Status, etc…
In the above example, We do not have any order or rank, or sequence. All the variables in the respective feature are equal. We can't give them any orders or ranks. Those features are called **Nominal features**.

2. Ordinal Variable

When we have a feature where variables have some order/rank.



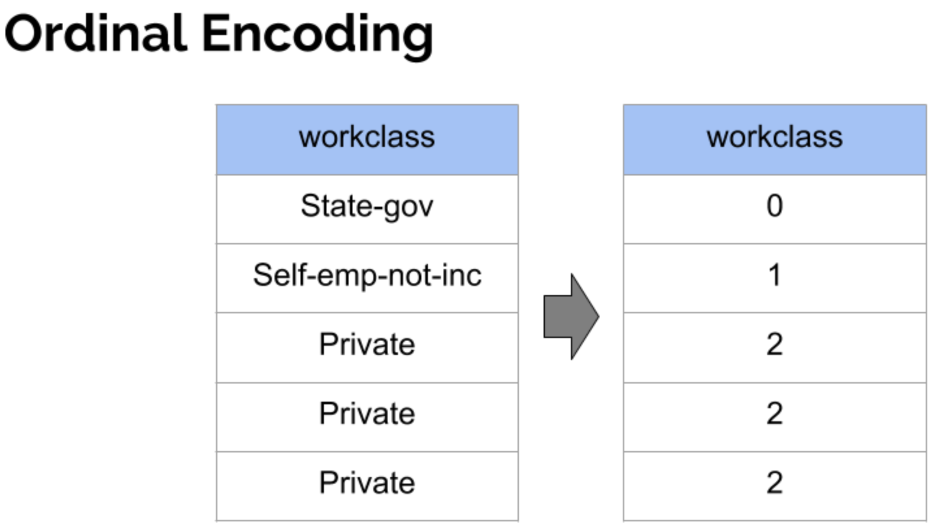
Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

For example: Student's performance, Customer's review, Education of person, etc. In the above example, we have orders/ranks/sequences. We can assign ranks based on student's performance, based on feedback given by customers, based on the highest education of the person. Those features are called **Ordinal features**.

Ordinal Encoding

In ordinal encoding, each unique category value is assigned an integer value.

For example, *"red"* is 1, *"green"* is 2, and *"blue"* is 3.



Ordinal Encoding

workclass
State-gov
Self-emp-not-inc
Private
Private
Private

workclass
0
1
2
2
2

This is called an **ordinal encoding or an integer encoding** and is easily reversible. Often, integer values starting at zero are used.

For some variables, an ordinal encoding may be enough. The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship.It is a natural encoding for ordinal variables. For categorical variables, it imposes an ordinal relationship where no such relationship may exist. This can cause problems and a one-hot encoding may be used instead.

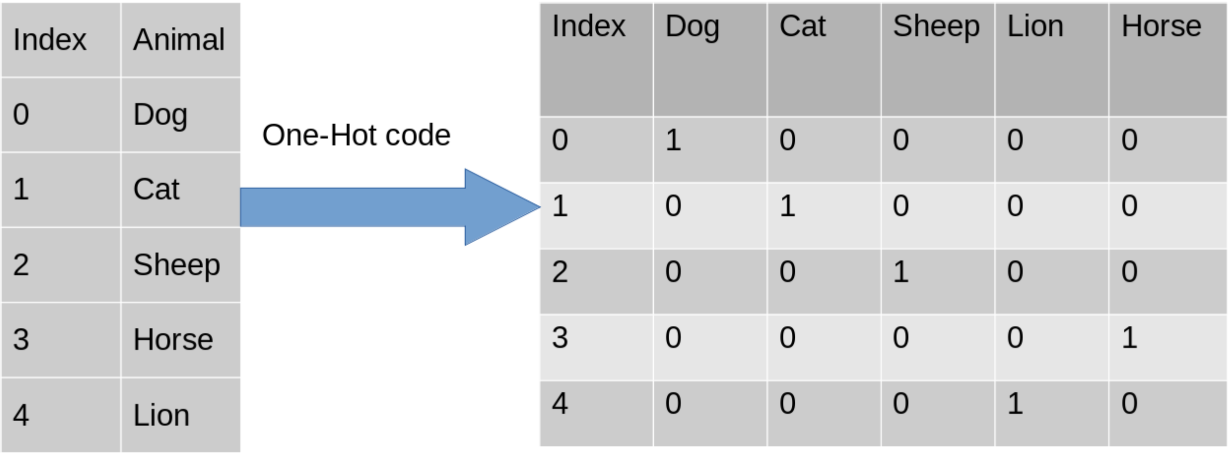
One-Hot Encoding

We use this categorical data encoding technique when the features are nominal(do not have any order). In **one hot encoding**, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, **0 represents the absence, and 1 represents the presence of that category**.

These newly created binary features are known as **Dummy variables**. The number of dummy variables depends on the levels present in the categorical variable. This might sound complicated.

Let us take an example to understand this better.

Suppose we have a dataset with a category animal, having different animals like Dog, Cat, Sheep, Cow, Lion. Now we have to one-hot encode this data.



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

After encoding, in the second table, we have dummy variables each representing a category in the feature Animal. Now for each category that is present, we have 1 in the column of that category and 0 for the others.

Label Encoding

In label encoding in Python, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4).

- When should we use label encoding?

Use LabelEncoder **when there are only two possible values of a categorical features**. For example, features having value such as yes or no. Or, maybe, gender feature when there are only two possible values including male or female.

Binary Encoding

Initially categories are encoded as Integer and then converted into binary code, then the digits from that binary string are placed into separate columns. for eg: for 7 : 1 1 1. This method is quite preferable when there are more number of categories.

Dummy Encoding

Dummy encoding also **uses dummy (binary) variables**. Instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding uses k-1 dummy variables.

Hash Encoding

Hash Encoding represents the categorical data into numerical value by the hashing function. Hashing is often used in data encryption or data comparison, but the main part is still similar — transform one feature to another using hashing function.

PIPELINE

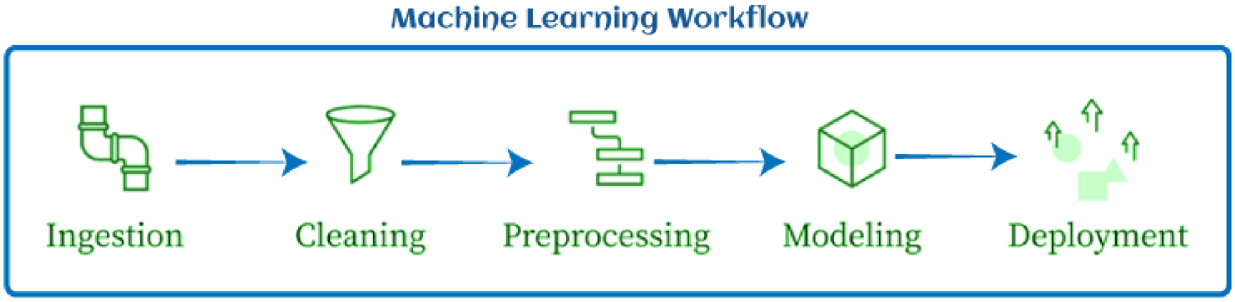
- What is Machine Learning Pipeline?

A ***Machine Learning pipeline is a process of automating the workflow of a complete machine learning task***. It can be done by enabling a sequence of data to be transformed and correlated together in a model that can be analyzed to get the output. A typical pipeline includes **raw data input, features, outputs, model parameters, ML models, and Predictions**. Moreover, an ML Pipeline contains multiple sequential steps that perform everything ranging from data extraction and pre-processing to model training and deployment in Machine learning in a modular approach. It means that *in **the pipeline, each step is designed as an independent module, and all these modules are tied together to get the final result.***

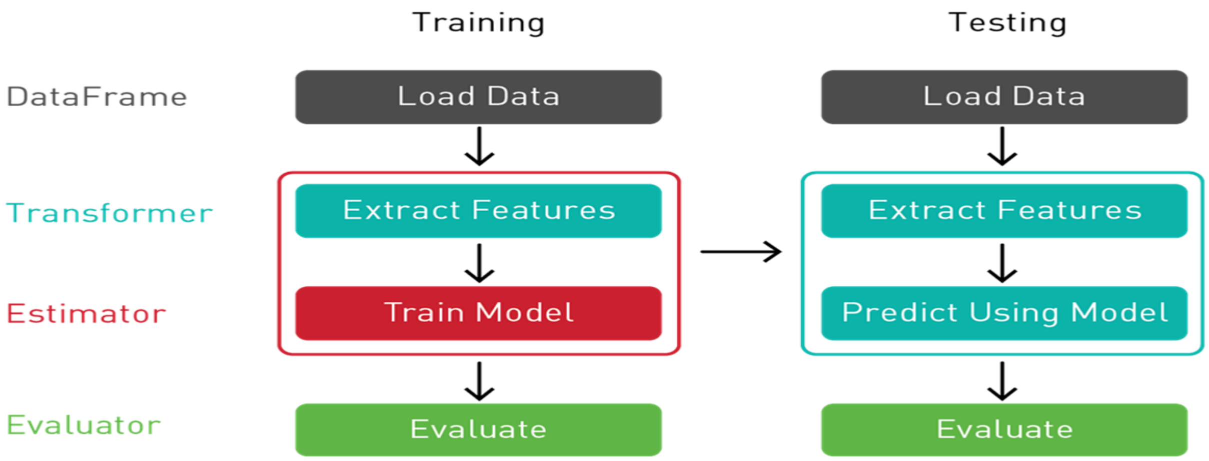
We can understand it with an example.

Building any ML model requires a huge amount of data to train the model. As data is collected from different resources, it is necessary to clean and pre-process the data, which is one of the crucial steps of an ML project. However, whenever a new dataset is included, we need to perform the same pre-processing step before using it for training, and it becomes a time-consuming and complex process for ML professionals.

To solve such issues, **ML pipelines** can be used, which can remember and automate the complete pre-processing steps in the same order.



- Why do we use pipeline in machine learning?



A machine learning pipeline is used **to help automate machine learning workflows**. They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative.

1. **Transformer:** It takes a dataset as an input and creates an augmented dataset as output. *For example, A tokenizer works as Transformer, which takes a text dataset, and transforms it into tokenized words*.
2. **Estimator:** An estimator is an algorithm that fits on the input dataset to generate a model, which is a transformer. *For example, regression is an Estimator that trains on a dataset with labels and features and produces a logistic regression model.*

References:

1. https://pbpython.com/categorical-encoding.html
2. https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/
3. https://ecampusontario.pressbooks.pub/commbusprofcdn/chapter/1-2/
4. https://askanydifference.com/difference-between-encoding-and-decoding/
5. https://danielmiessler.com/study/encoding-encryption-hashing-obfuscation/
6. https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/
7. https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
8. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
9. https://towardsdatascience.com/4-categorical-encoding-concepts-to-know-for-data-scientists-e144851c6383
10. https://medium.com/analytics-vidhya/what-is-a-pipeline-in-machine-learning-how-to-create-one-bda91d0ceaca
11. https://valohai.com/machine-learning-pipeline/
12. https://www.datarobot.com/blog/what-a-machine-learning-pipeline-is-and-why-its-important/