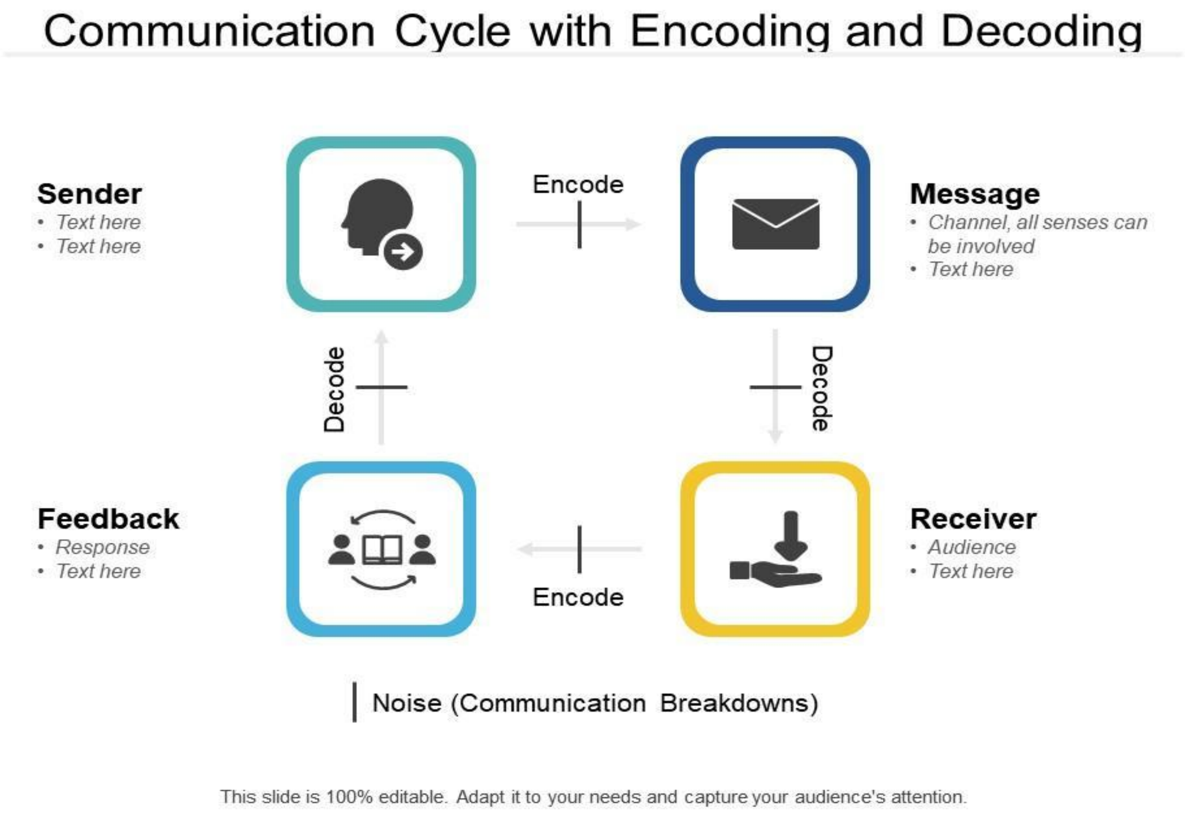
Encoding(Simple way)

*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*

Receiver
- *Audience*
- *Text here*

Feedback
- *Response*
- *Text here*

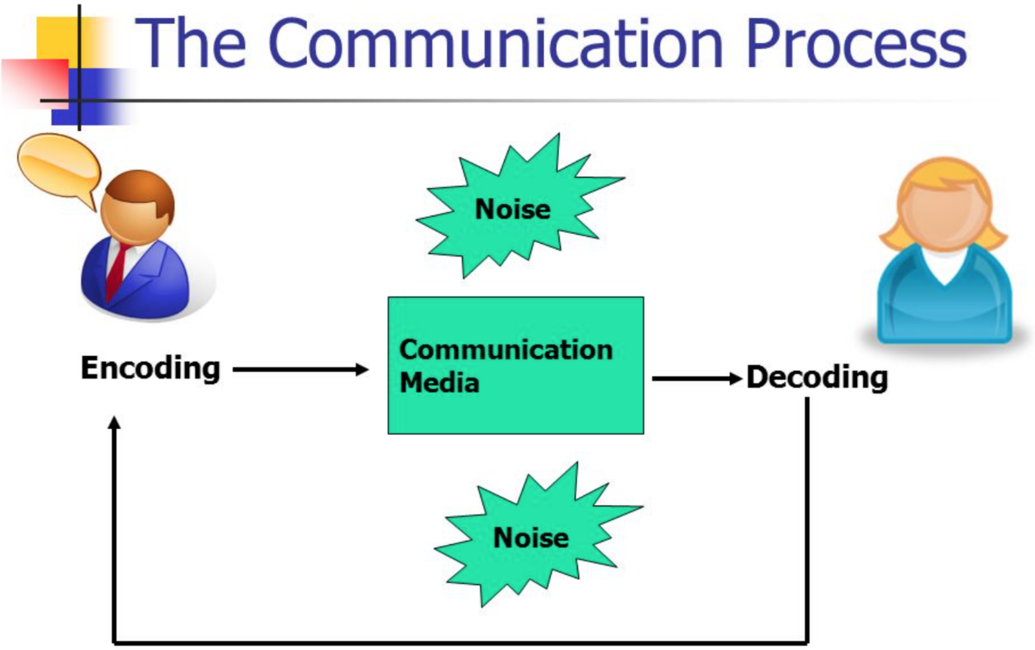
Noise (Communication Breakdowns)

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.

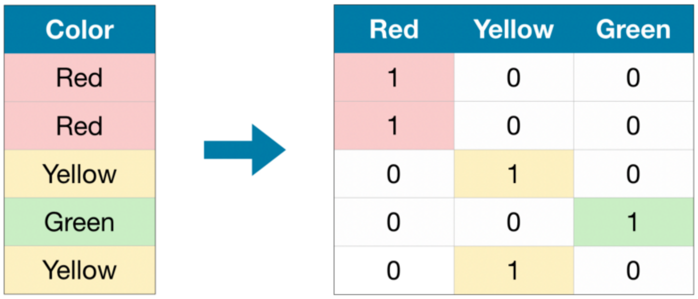
- What is an example of encoding?



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.



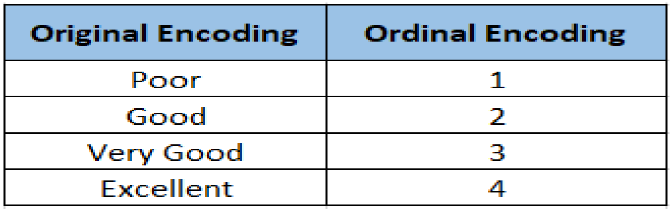
Color
Red
Red
Yellow
Green
Yellow

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.



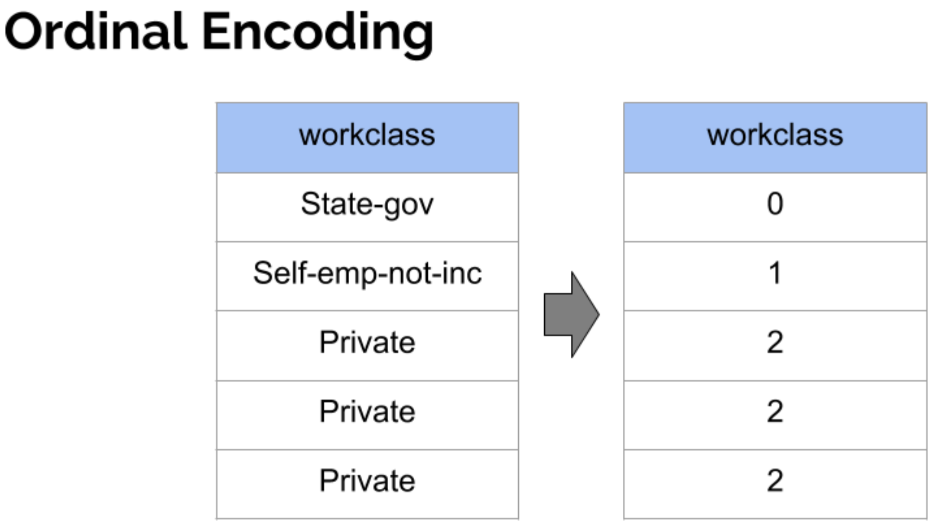
Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

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.

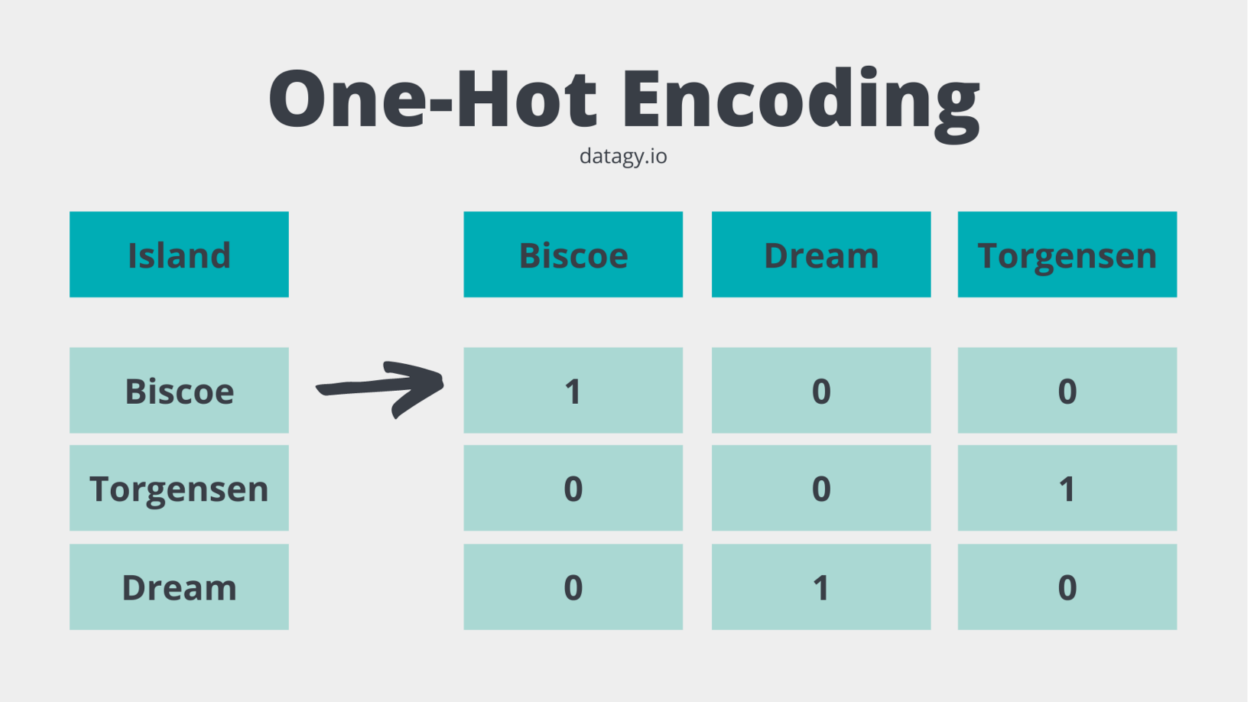
For some variables, an ordinal encoding may be enough. The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship.It is a natural encoding for ordinal variables. For categorical variables, it imposes 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

For categorical variables where no ordinal relationship exists, the integer encoding may not be enough, at best, or misleading to the model at worst.

Forcing an ordinal relationship via an ordinal encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a **one-hot encoding** can be applied to the ordinal representation. This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable.


One-Hot Encoding
datagy.io

Island
Biscoe
Torgensen
Dream

Biscoe	Dream	Torgensen
1	0	0
0	0	1
0	1	0

Each bit represents a possible category. If the variable cannot belong to multiple categories at once, then only one bit in the group can be "on." *This is called one-hot encoding …* — Page 78, Feature Engineering for Machine Learning, 2018.

PIPELINE

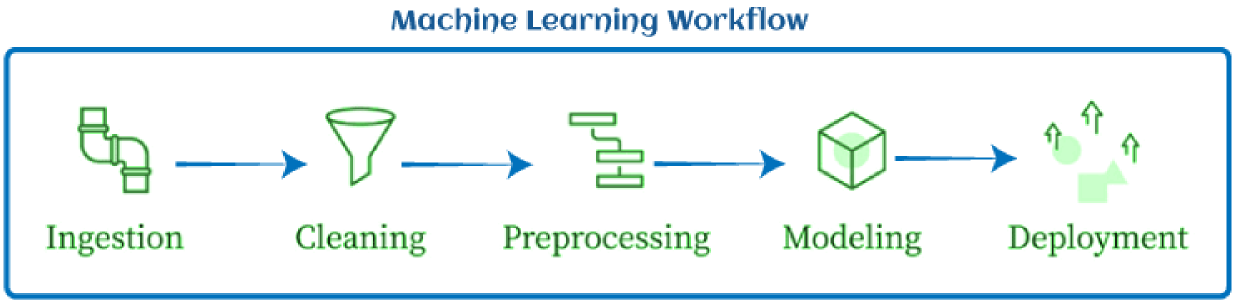
- What is Machine Learning Pipeline?

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.***

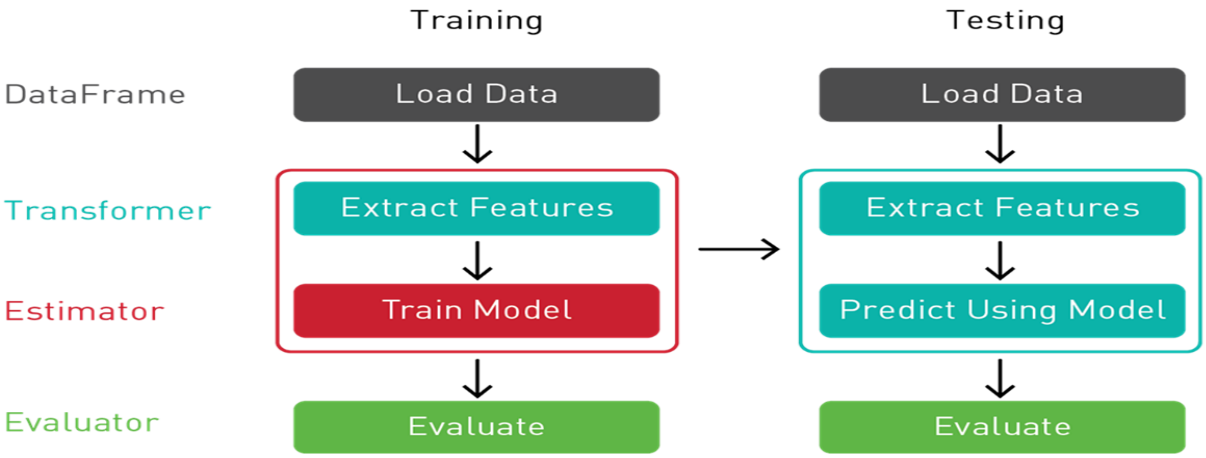
We can understand it with an example.

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.*