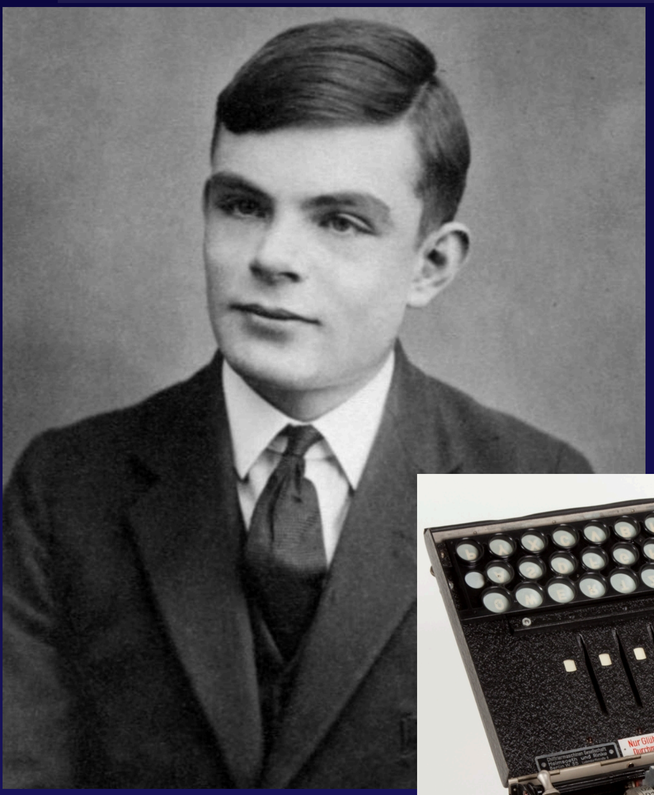
A CLOUD GURU

Categorical Encoding

Brock Tubre
INSTRUCTOR

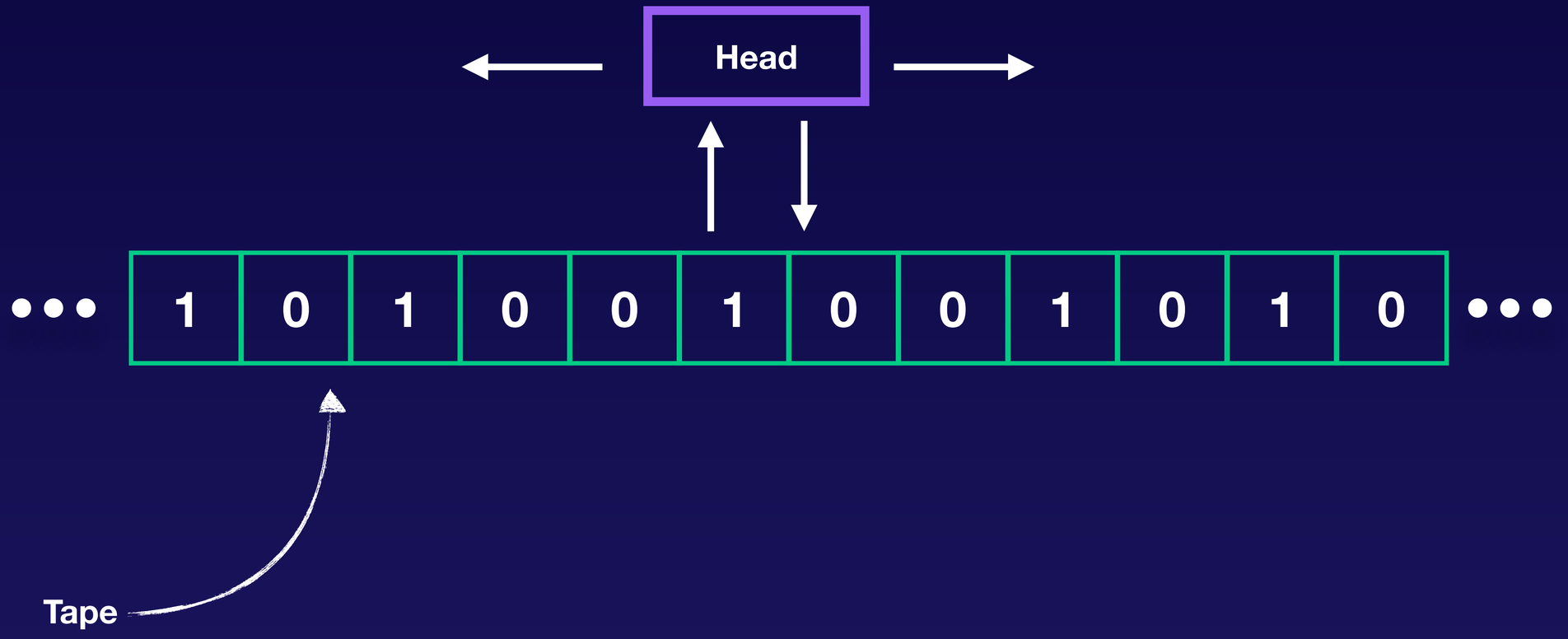


Alan Turing

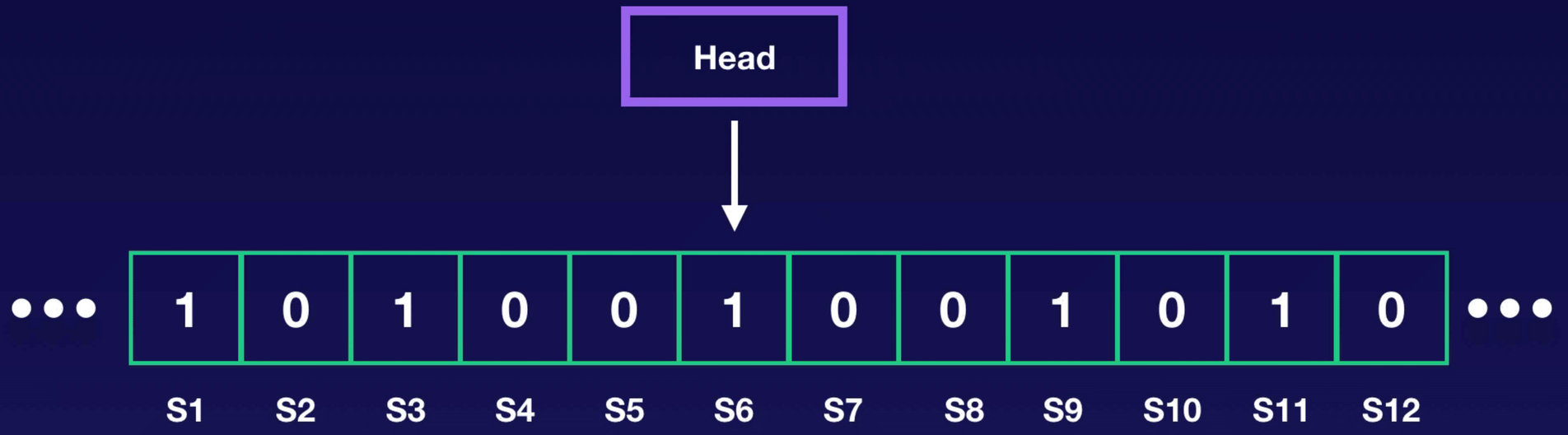
Educator, Mathematician, PhD (1912–1954)

An English scientist in mathematics and cryptology who helped intercepted coded messages and decipher the messages. Founding father of computer science and artificial intelligence.

Turing Machine



Turing Machine



Present State	Symbol Scanned	Write	Move	New State
S6	1	0	Right	S7

Categorical Encoding

Categorical Encoding

Categorical encoding is the process of manipulating categorical variables when ML algorithms expect numerical values as inputs.

Changing category values in our dataset to numbers.

Categorical Encoding

categorical variable = categorical feature = discrete feature

When to encode?

Problem	Algorithm	Encoding
Predicting the price of a home	Linear Regression	Encoding necessary
Determine wether given text is about sports or not	Naive Bayes	Encoding not necessary
Detecting malignancy in radiology images	Convoluted Neural Network	Encoding necessary

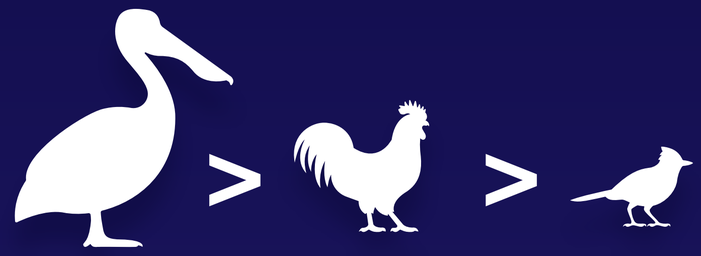
Categorical Encoding Examples

Color: { green, purple, blue }

Evil: { true, false }

ID	Name	Evil
1	Luke	false
2	Leia	false
3	Han	false
4	Vadar	true

Size: { L > M > S }



Categorical Encoding Examples

Nominal - order does not matter

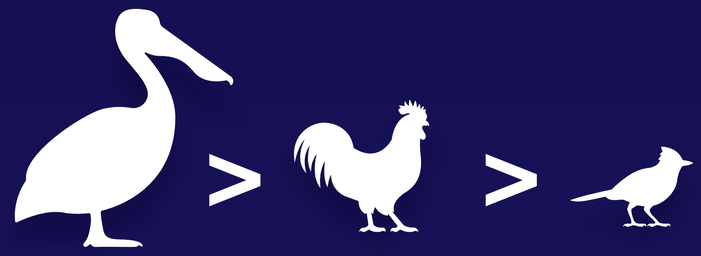
Color: { green, purple, blue }

Evil: { true, false }

ID	Name	Evil
1	Luke	false
2	Leia	false
3	Han	false
4	Vadar	true

Ordinal - order does matter

Size: { L > M > S }



Encoding Categorical Features

Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	N
2	house	4	2988	L	243566	Y
3	house	3	1877	N	125700	N
4	condo	5	3876	M	345000	Y
5	apartment	2	1250	N	120900	Y

Encoding Categorical Features

Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	N
2	house	4	2988	L	243566	Y
3	house	3	1877	N	125700	N
4	condo	5	3876	M	345000	Y
5	apartment	2	1250	N	120900	Y

Encoding Categorical Features

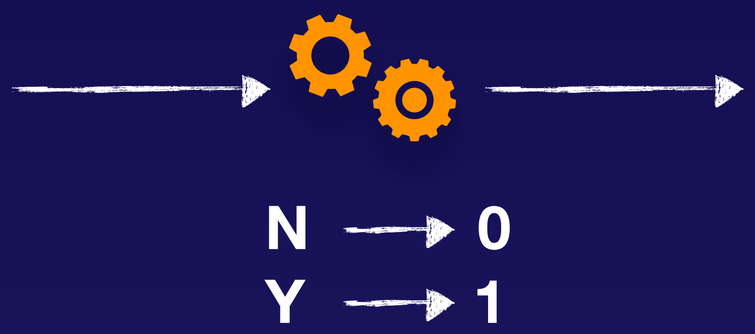
Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	N
2	house	4	2988	L	243566	Y
3	house	3	1877	N	125700	N
4	condo	5	3876	M	345000	Y
5	apartment	2	1250	N	120900	Y

Encoding Categorical Features

Loan_Approved
N
Y
N
Y
Y



Loan_Approved
0
1
0
1
1

Encoding Categorical Features

Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	0
2	house	4	2988	L	243566	1
3	house	3	1877	N	125700	0
4	condo	5	3876	M	345000	1
5	apartment	2	1250	N	120900	1

Encoding Categorical Features

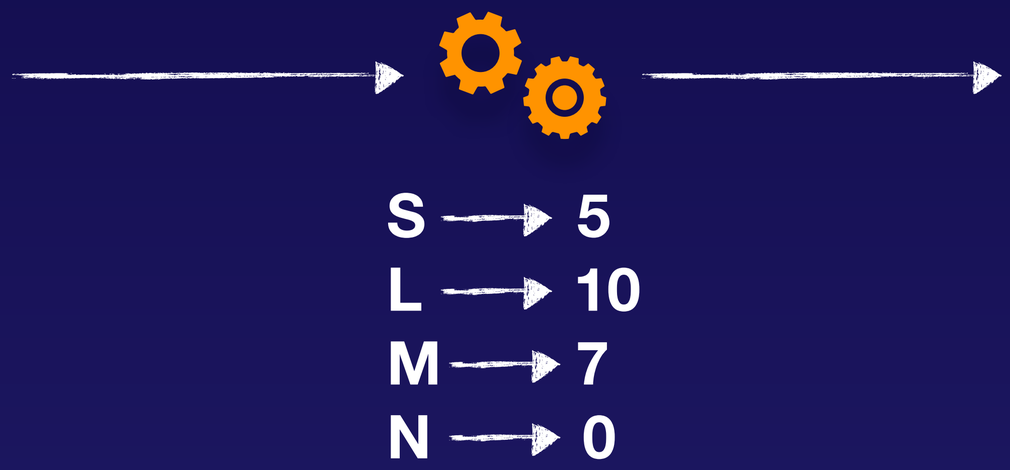
Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	250555	0
2	house	4	2988	L	243566	1
3	house	3	1877	N	125700	0
4	condo	5	3876	M	345000	1
5	apartment	2	1250	N	120900	1

Encoding Categorical Features

Pool_Size
S
L
N
M
N



Num_Pool_Size
5
10
0
7
0

Encoding Categorical Features

Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Num_Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	5	250555	0
2	house	4	2988	L	10	243566	1
3	house	3	1877	N	0	125700	0
4	condo	5	3876	M	7	345000	1
5	apartment	2	1250	N	0	120900	1

Encoding Categorical Features

Problem

Will a home loan be approved?

ID	Type	Bedrooms	Area	Pool_Size	Num_Pool_Size	Price	Loan_Approved
1	condo	2	2432	S	5	250555	0
2	house	4	2988	L	10	243566	1
3	house	3	1877	N	0	125700	0
4	condo	5	3876	M	7	345000	1
5	apartment	2	1250	N	0	120900	1

Encoding Categorical Features

Type
condo
house
house
condo
apartment

condo → 1
house → 2
apartment → 3

Num_Home_Type
1
2
2
1
3

Encoding Categorical Features

Type
condo
house
house
condo
apartment

Num_Home_Type
1
2
2
1
3

condo < house < apartment

Encoding Categorical Features

Encoding nominal variables to integers is a bad idea!

Type
condo
house
house
condo
apartment

Num_Home_Type
1
2
2
1
3

condo < house < apartment

One-hot Encoding

One-hot Encoding

Transforms nominal categorical features and creates new binary columns for each observation.

Adding columns to your dataset of 1’s and 0’s.

One-hot Encoding

Type
condo
house
house
condo
apartment

one-hot encoding

Type_condo	Type_house	Type_apartment
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1

One-hot Encoding

Type
condo
house
house
condo
apartment

one-hot encoding

Type_condo	Type_house	Type_apartment
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1

One-hot Encoding

Type
condo
house
house
condo
apartment

Type_condo	Type_house	Type_apartment
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1

One-hot Encoding

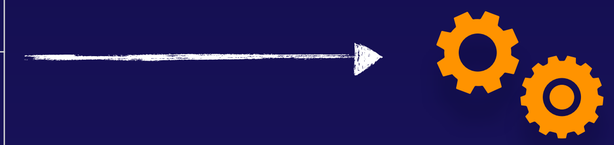
Type
condo
house
house
condo
apartment

one-hot encoding

Type_condo	Type_house	Type_apartment
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1

One-hot Encoding

Type
condo
house
house
condo
apartment



one-hot encoding

Type_condo	Type_house	Type_apartment
1	0	0
0	1	0
0	1	0
1	0	0
0	0	1

One-hot Encoding

1

One-hot or not?

One-hot encoding is not always a good choice when there are many, many categories.

2

Grouping

Using techniques like grouping by similarity could create fewer overall categories before encoding.

3

Mapping Rare Values

Mapping rare values to “other” can help reduce overall number of new columns created.

Categorical Encoding Summary

1

ML Algorithm Specific

In general, categorical encoding is used when the ML algorithm can not support categorical data.

2

Text into Numbers

We must find a way to turn text attributes into numeric attributes within our datasets.

3

There is No “Golden Rule”

There is no “golden rule” on how to encode your categories (or transform your data in general).

4

Many Different Approaches

There are many different approaches and each approach can have a different impact on the outcome of your analysis.