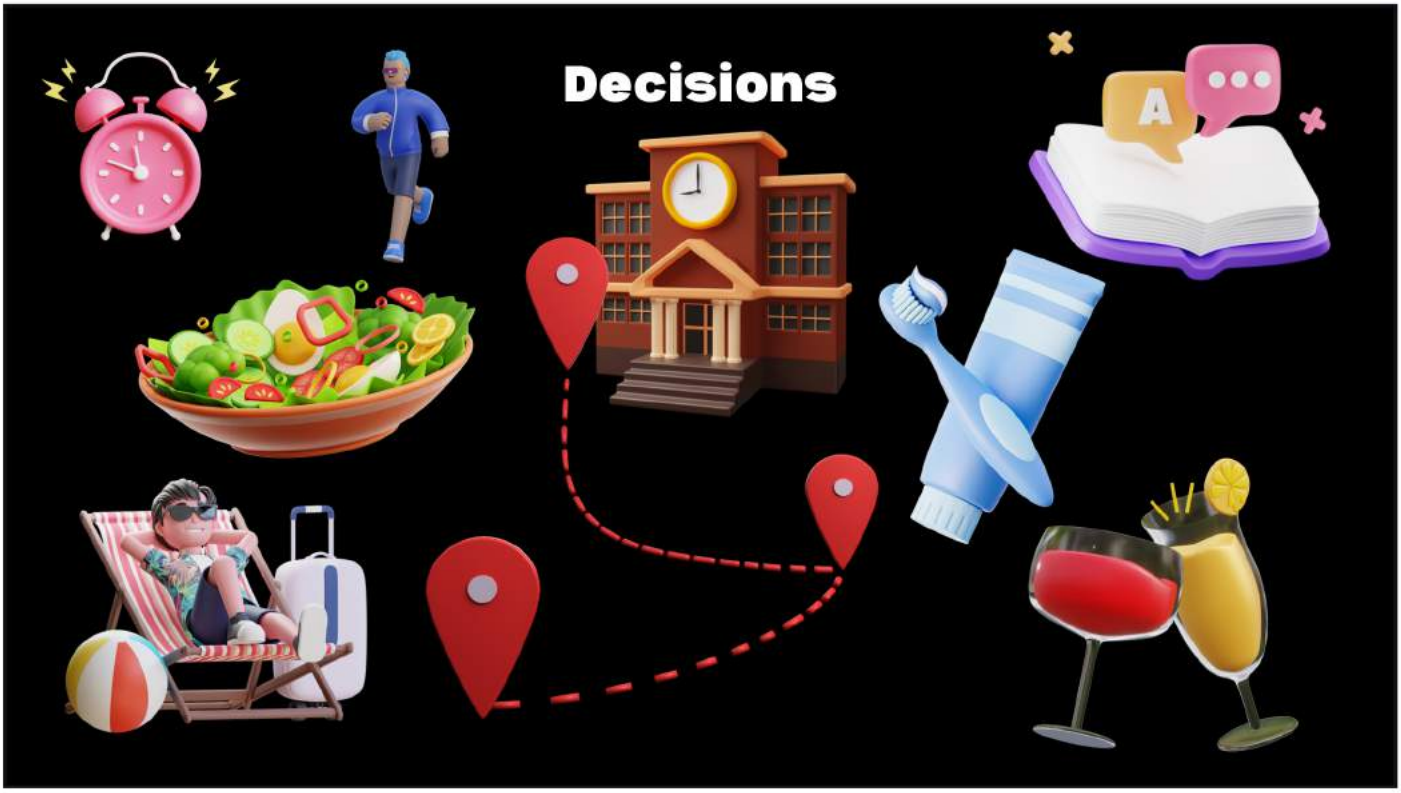
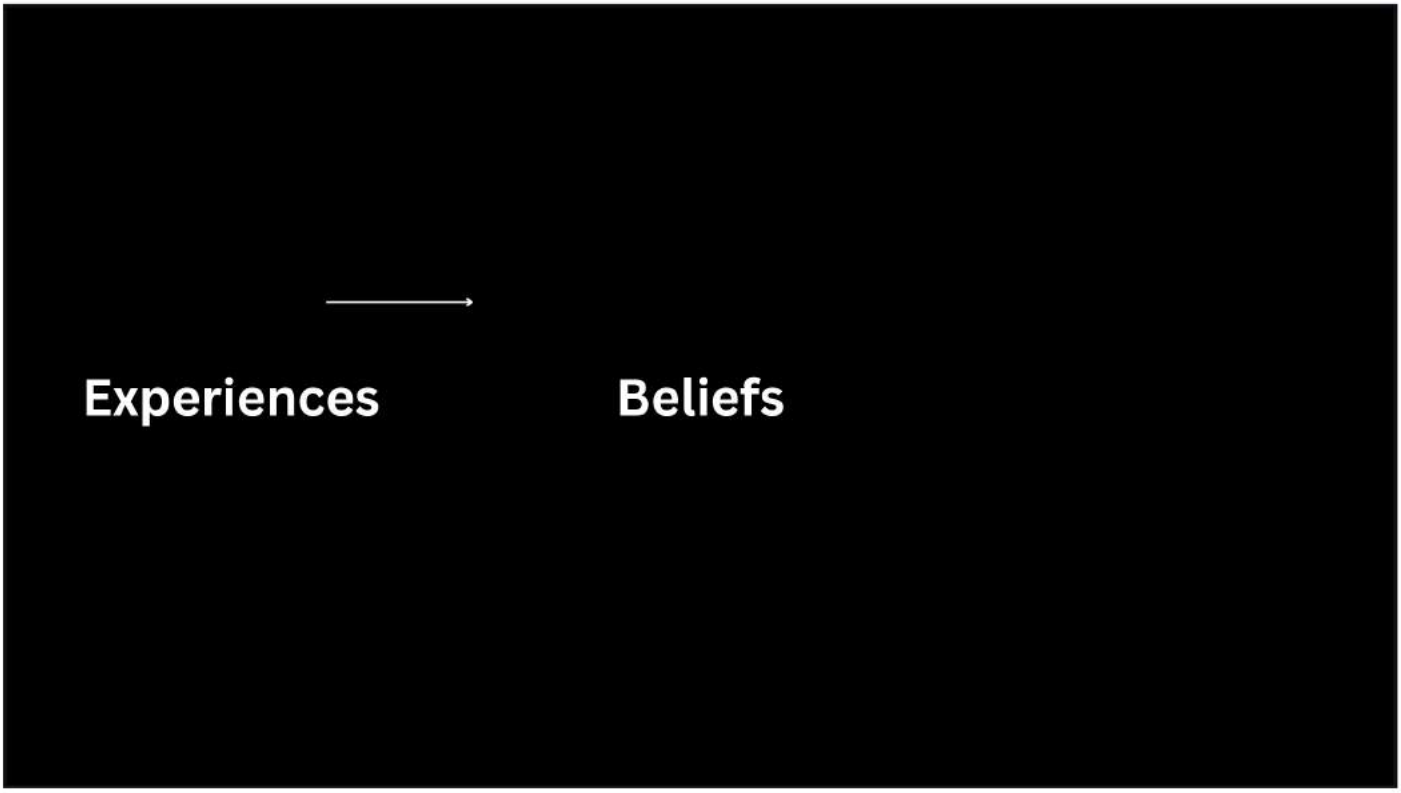
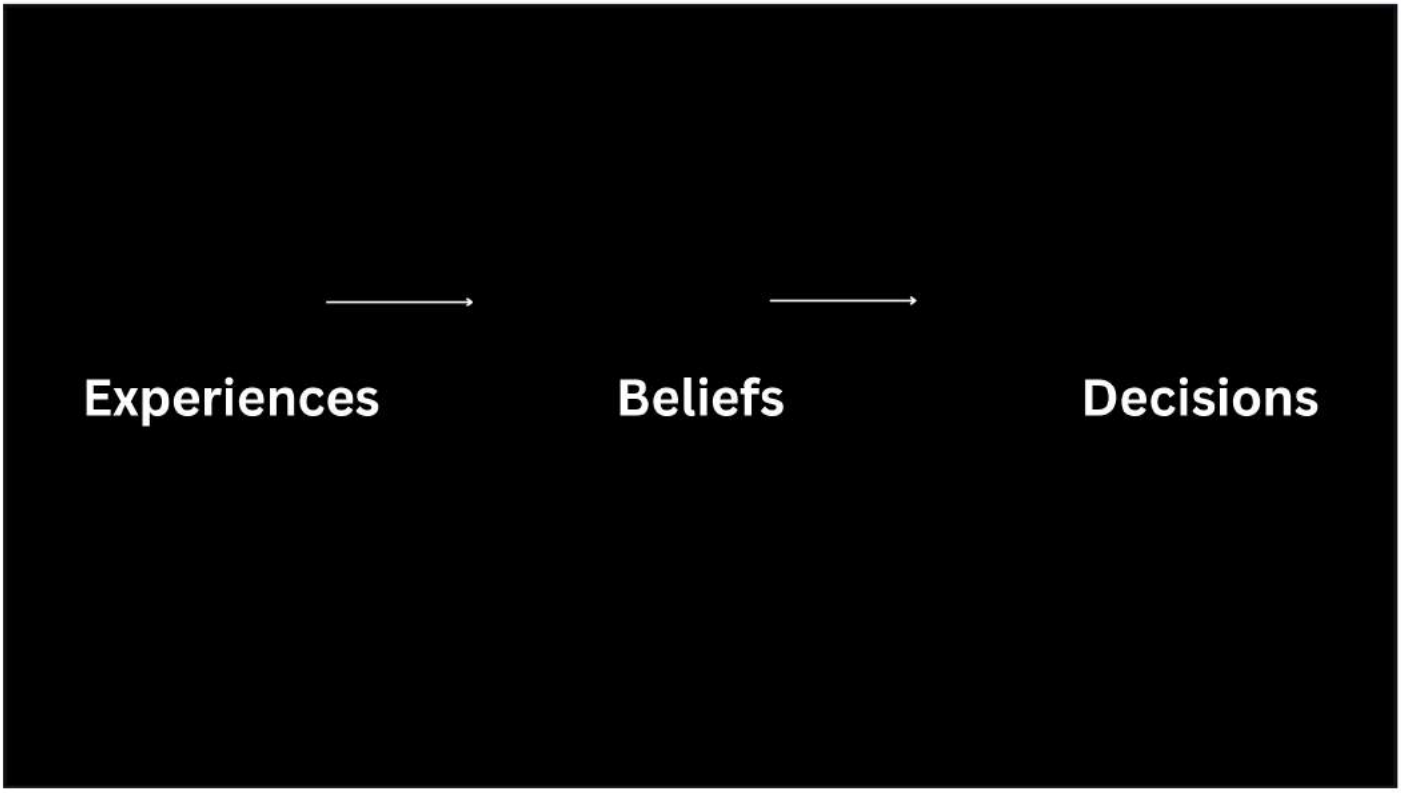
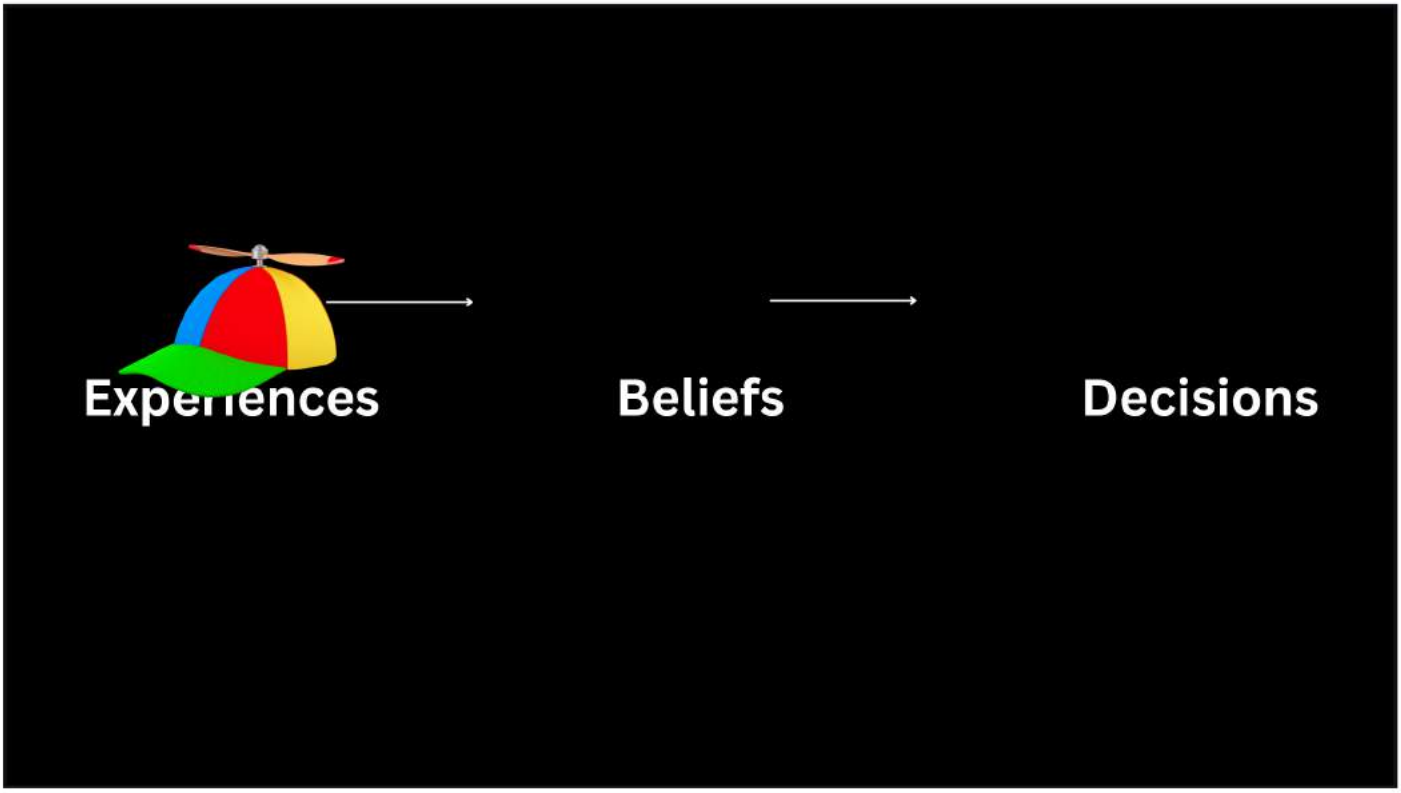
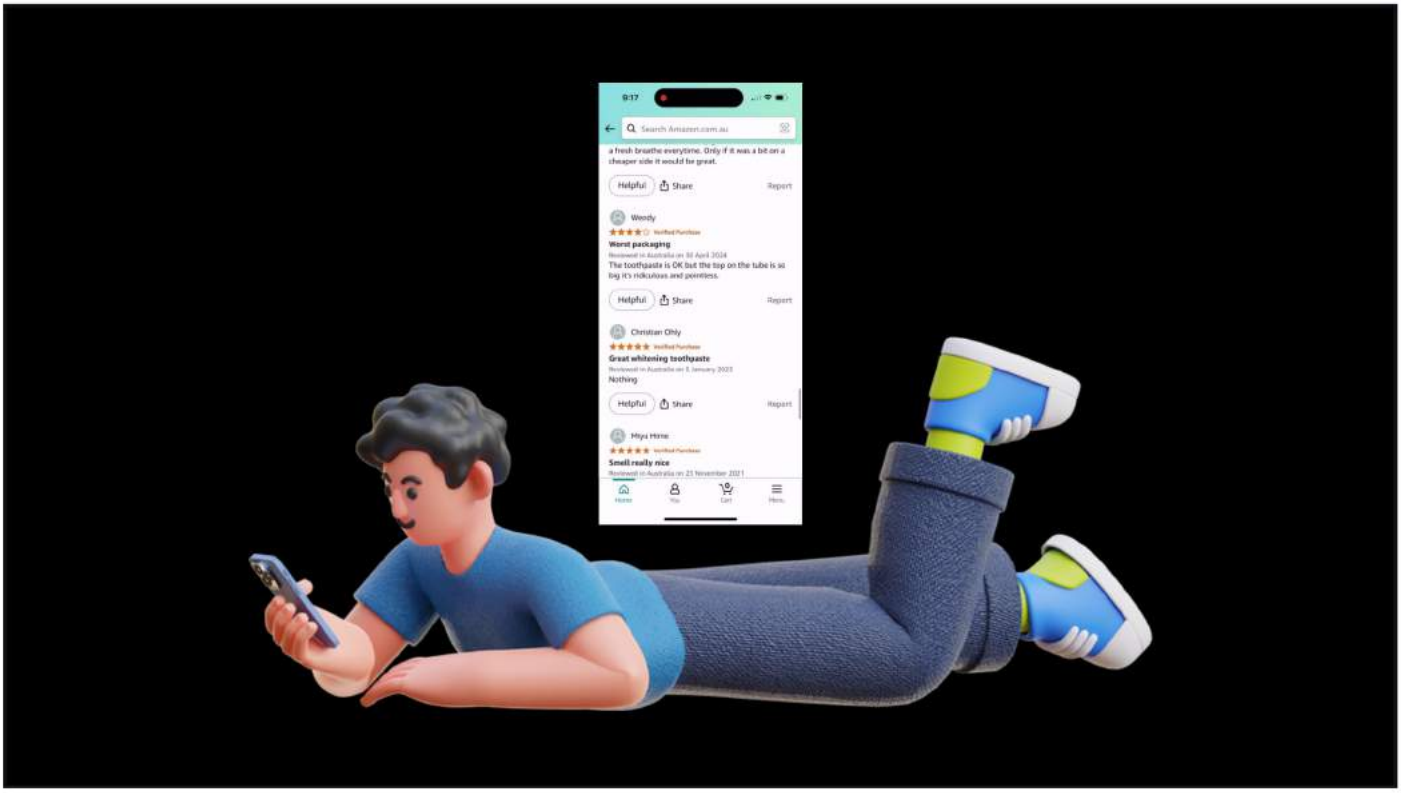

VICHACK
A peek into the future
CITADEL | Securities
REA Group
Jane Street

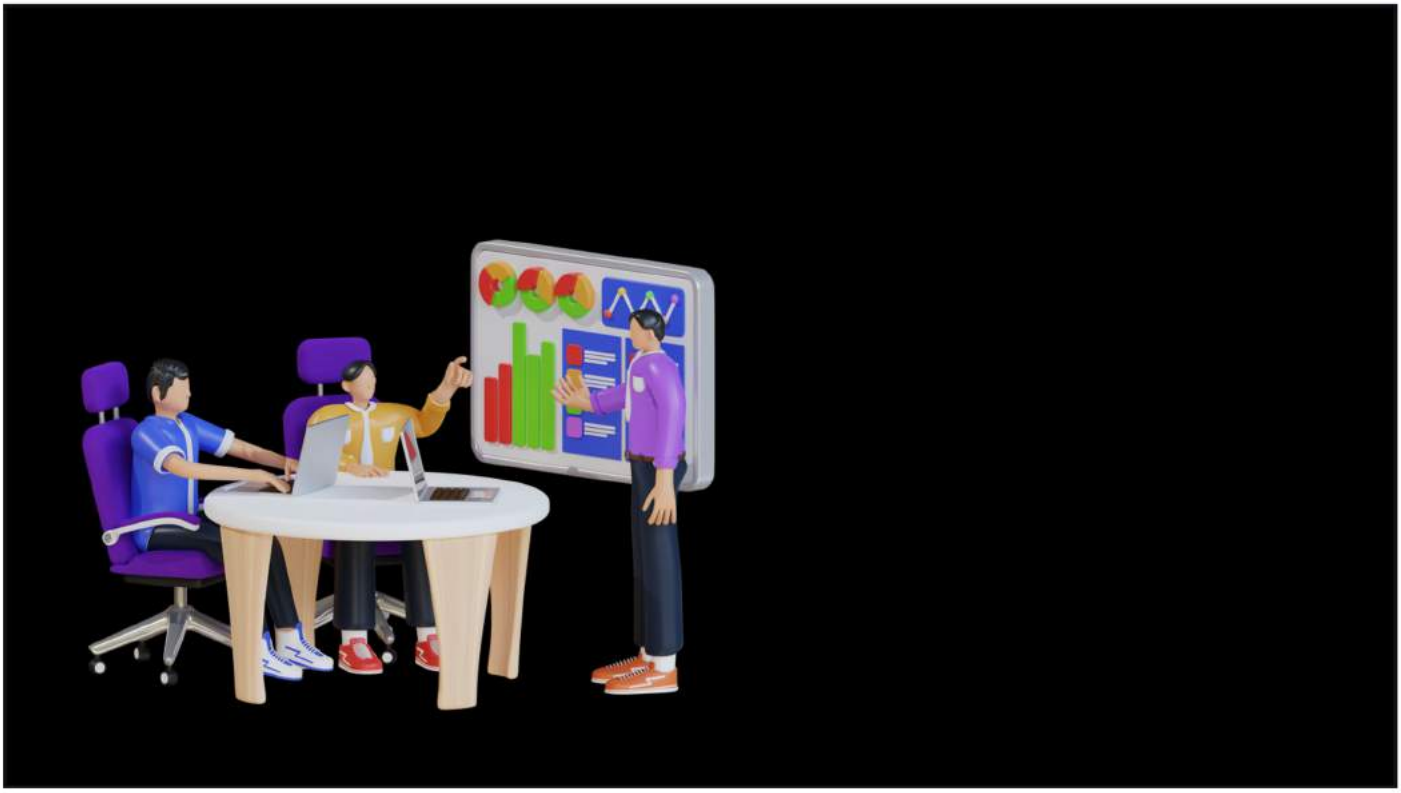

Decisions
A


Experiences
Beliefs


Experiences
Beliefs
Decisions

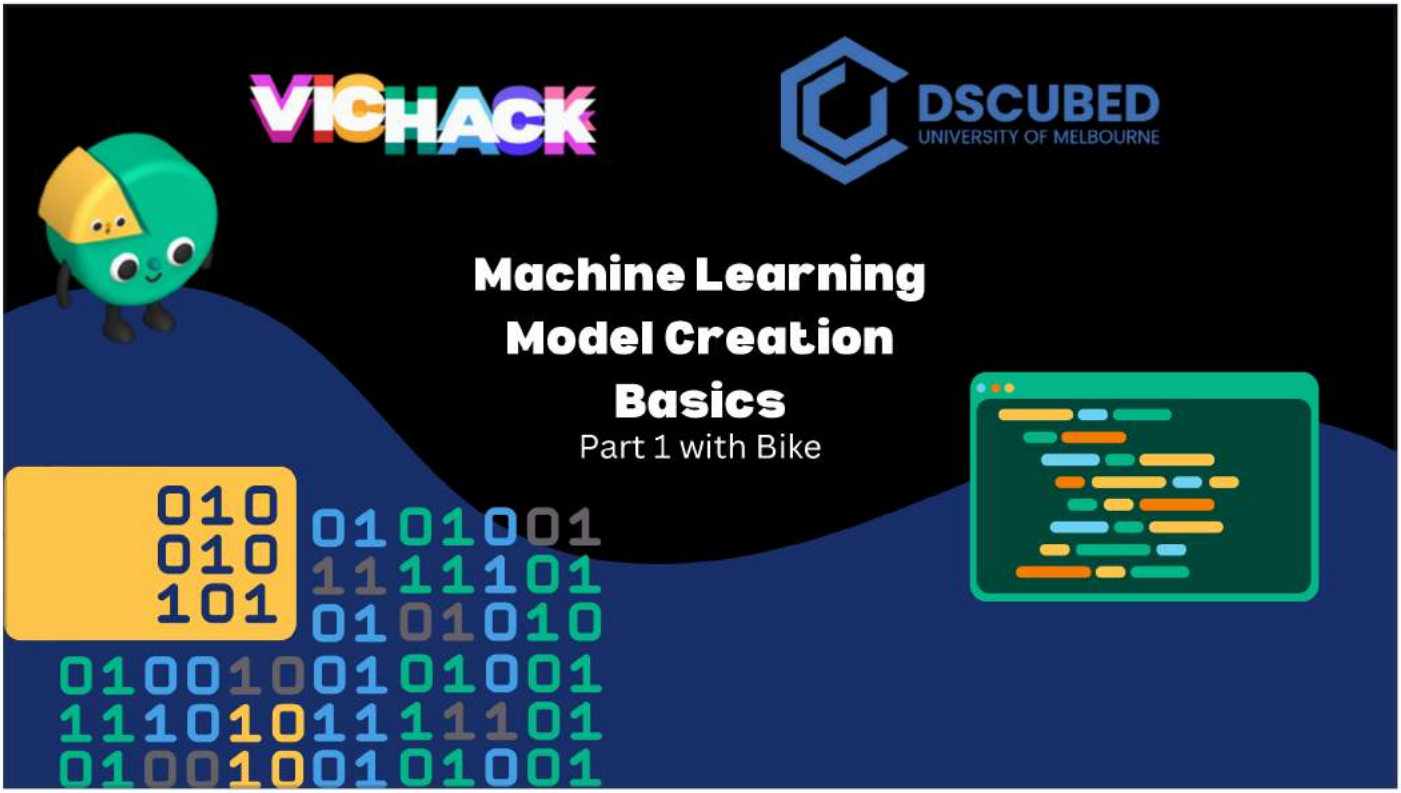

Experiences
Beliefs
Decisions

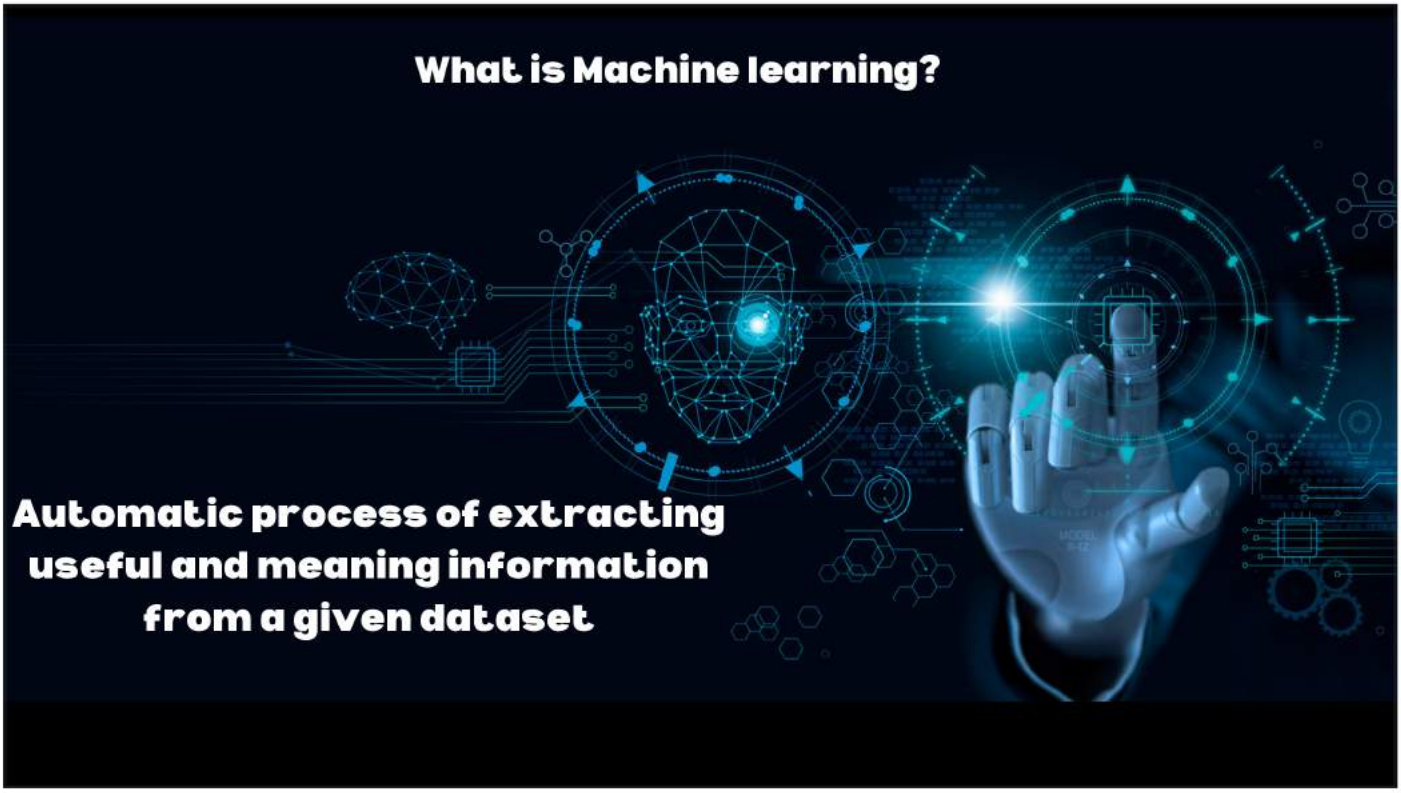

a fresh breathe everytime. Only if it was a bit on a cheaper side it would be great.
Helpful
Share
Report
Weedy
Worst packaging
The toothpaste is OK but the top on the tube is so big it's ridiculous and pointless.
Helpful
Share
Report
Christian Ohly
Great whitening toothpaste
Nothing
Helpful
Share
Report
Miyu Hime
Smell really nice



Great Data Science, Great Decisions

A peek into the future

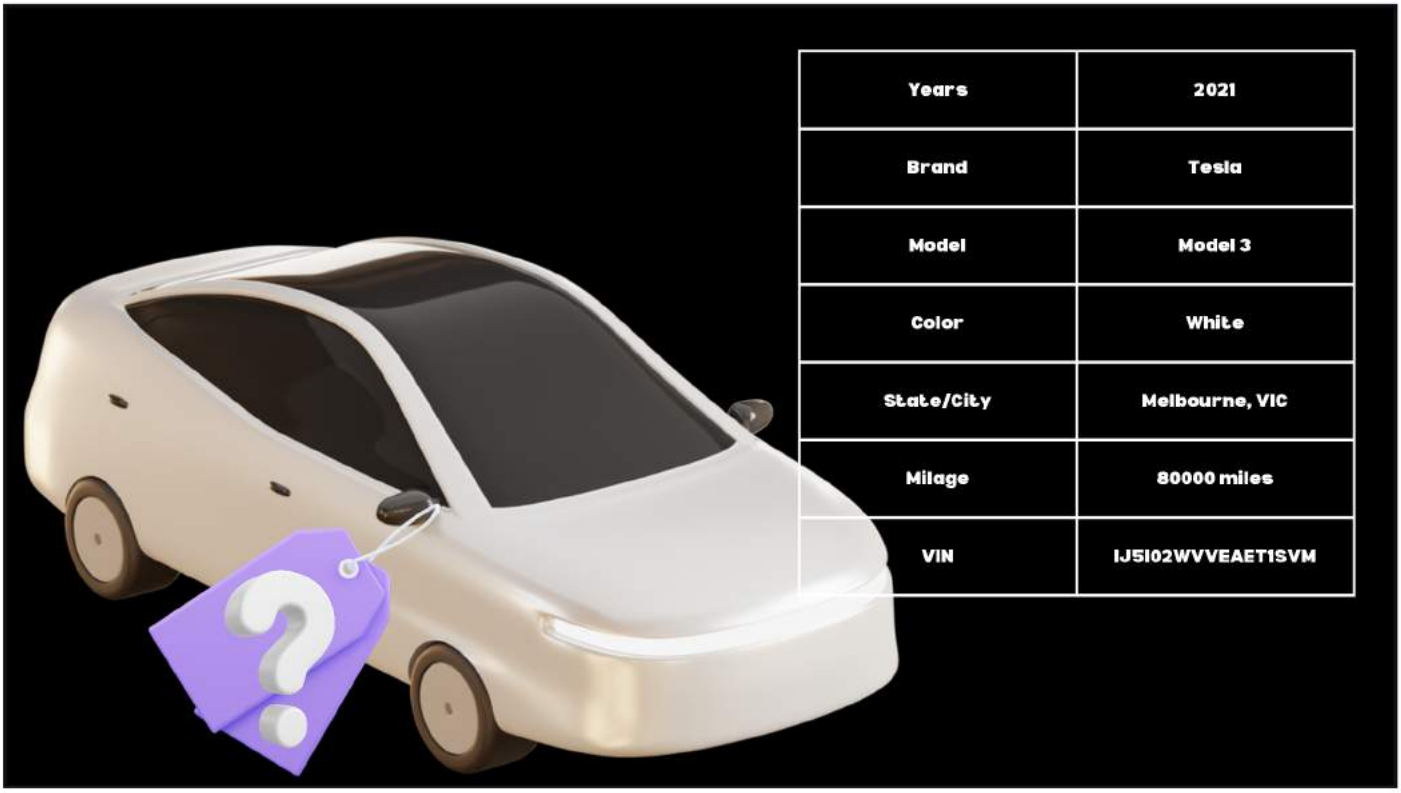

VICHACK
DSCUBED
UNIVERSITY OF MELBOURNE
Machine Learning
Model Creation
Basics
Part 1 with Bike
010
010
101
01010001
11111101
01010101
0100100101001
1110101111101
0100100101001


What is Machine learning?
Automatic process of extracting useful and meaning information from a given dataset

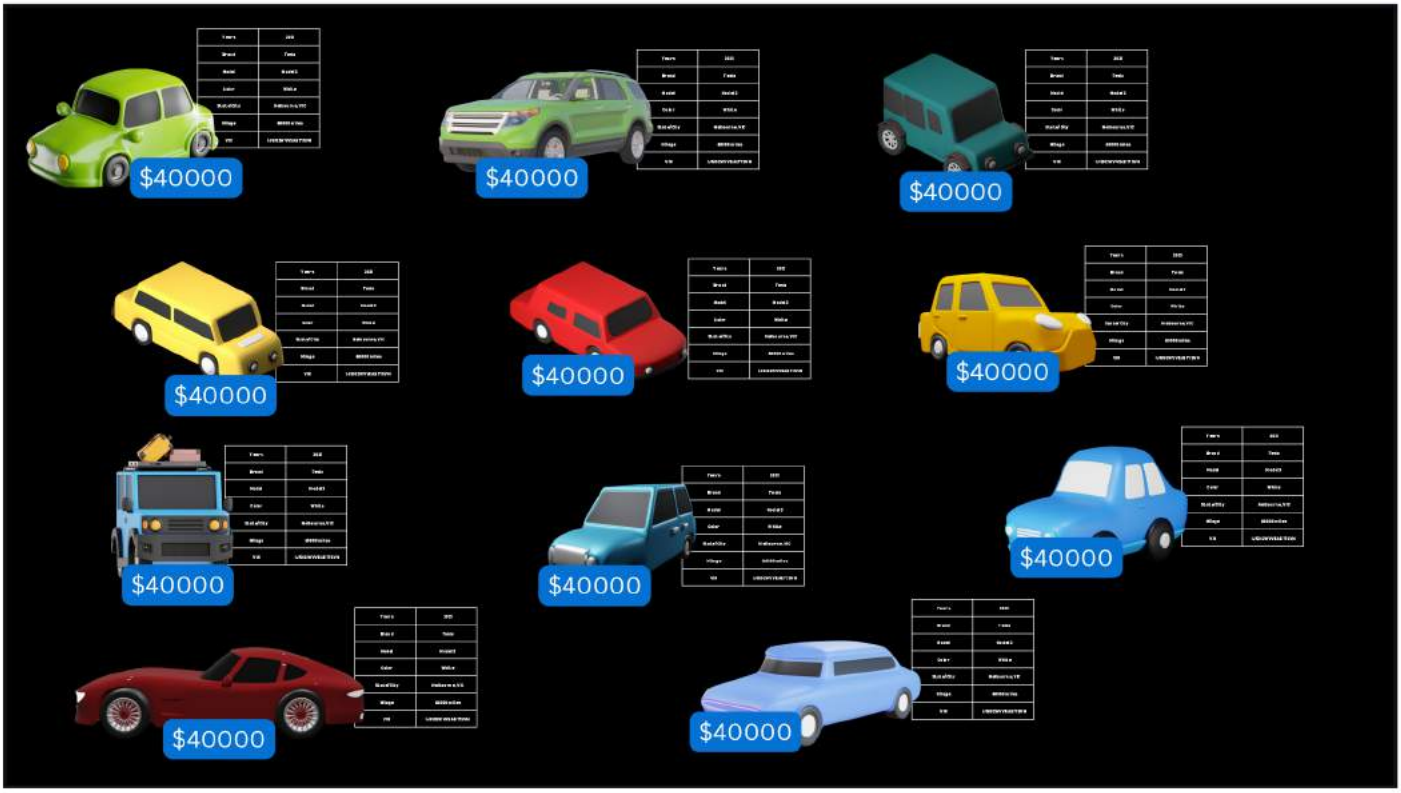

Supervised
Unsupervised

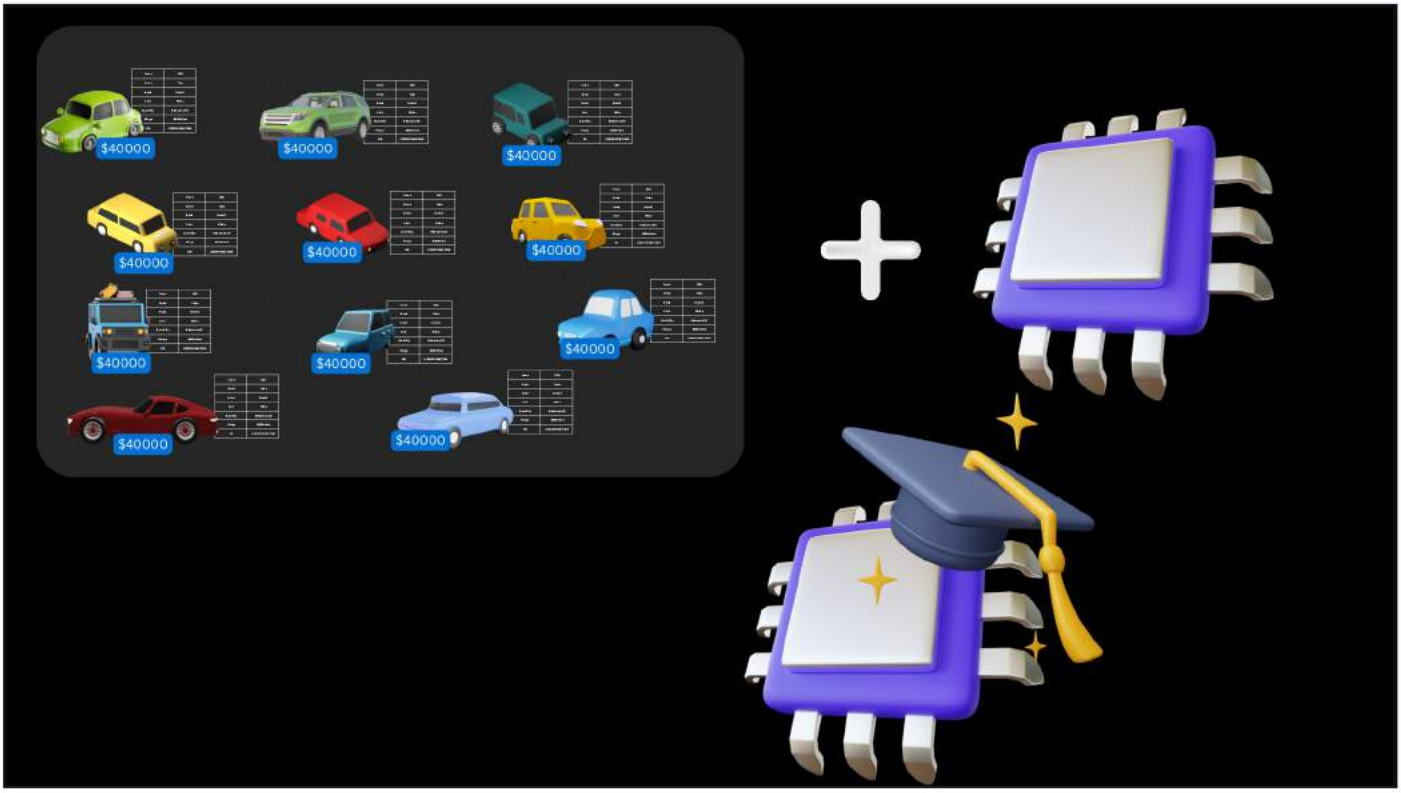
Supervised

Learns from features -> predict outcome variable

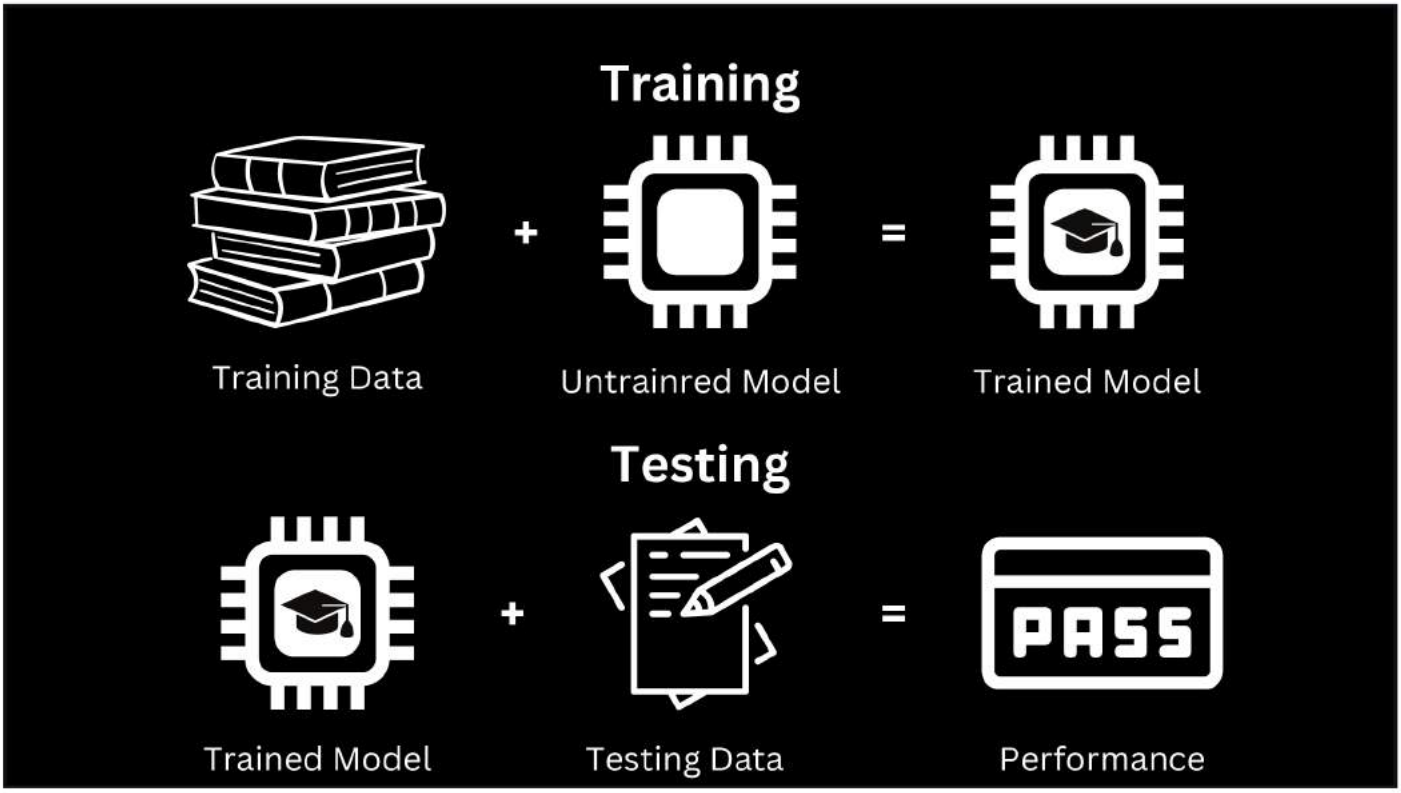


Years	2021
Brand	Tesla
Model	Model 3
Color	White
State/City	Melbourne, VIC
Milage	80000 miles
VIN	IJ5I02WVVEAET1SVM

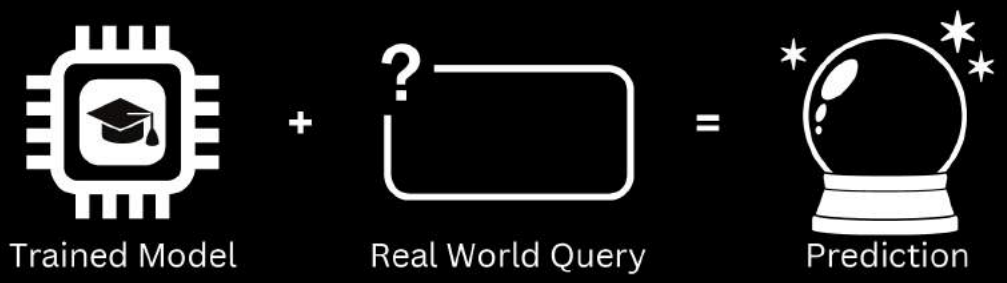

$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000


$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000
$40000

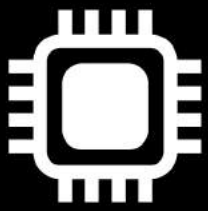

$40000


Training
+
=
Training Data
Untrainred Model
Trained Model
Testing
+
=
PASS
Trained Model
Testing Data
Performance

Usage



What happens inside?



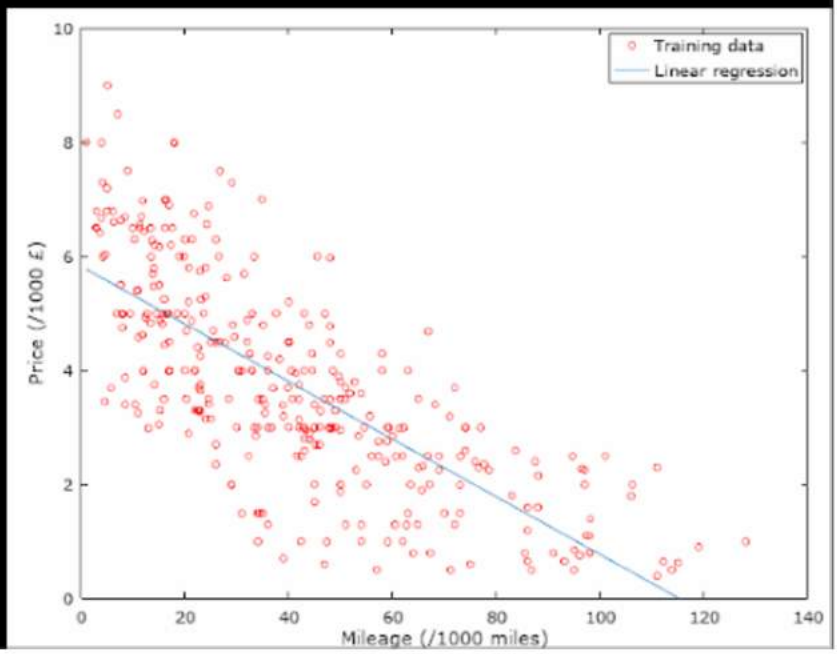
But after a 2 min ad break!

Regression type problems

Predict car price with mileage

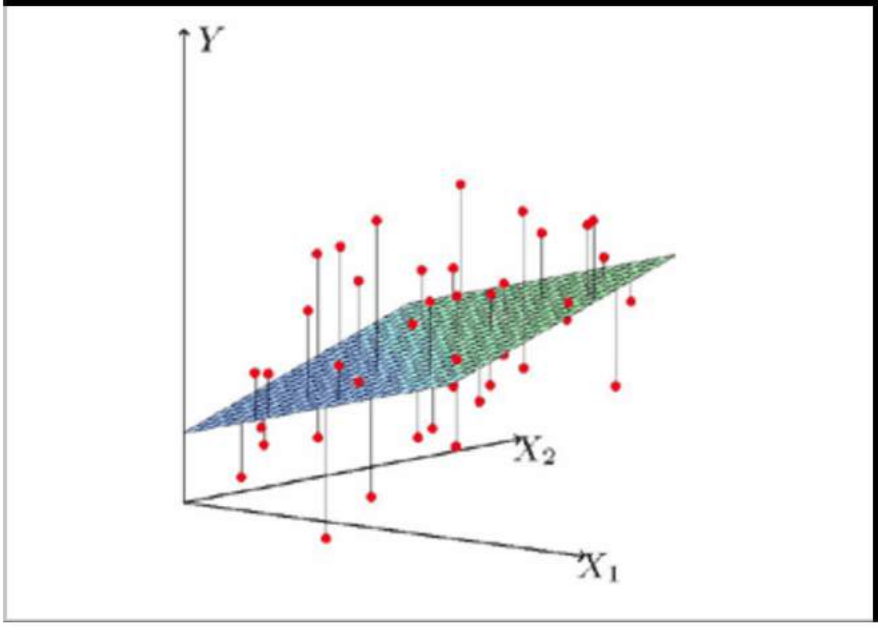
$$y = m * x + c$$

$$\text{Price} = m * \text{milage} + c$$



Predict price with both year and mileage

y = B1*Mileage + B2*Year + B0



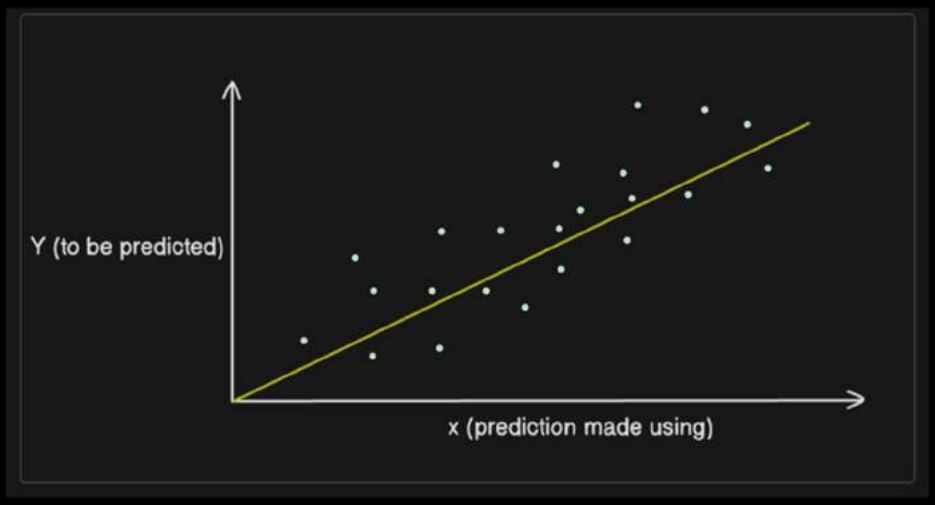
- Y = Price
- B1, B2 = gradient/weight
- B0 = Intercept
- X1 = Mileage
- X2 = Year

All features

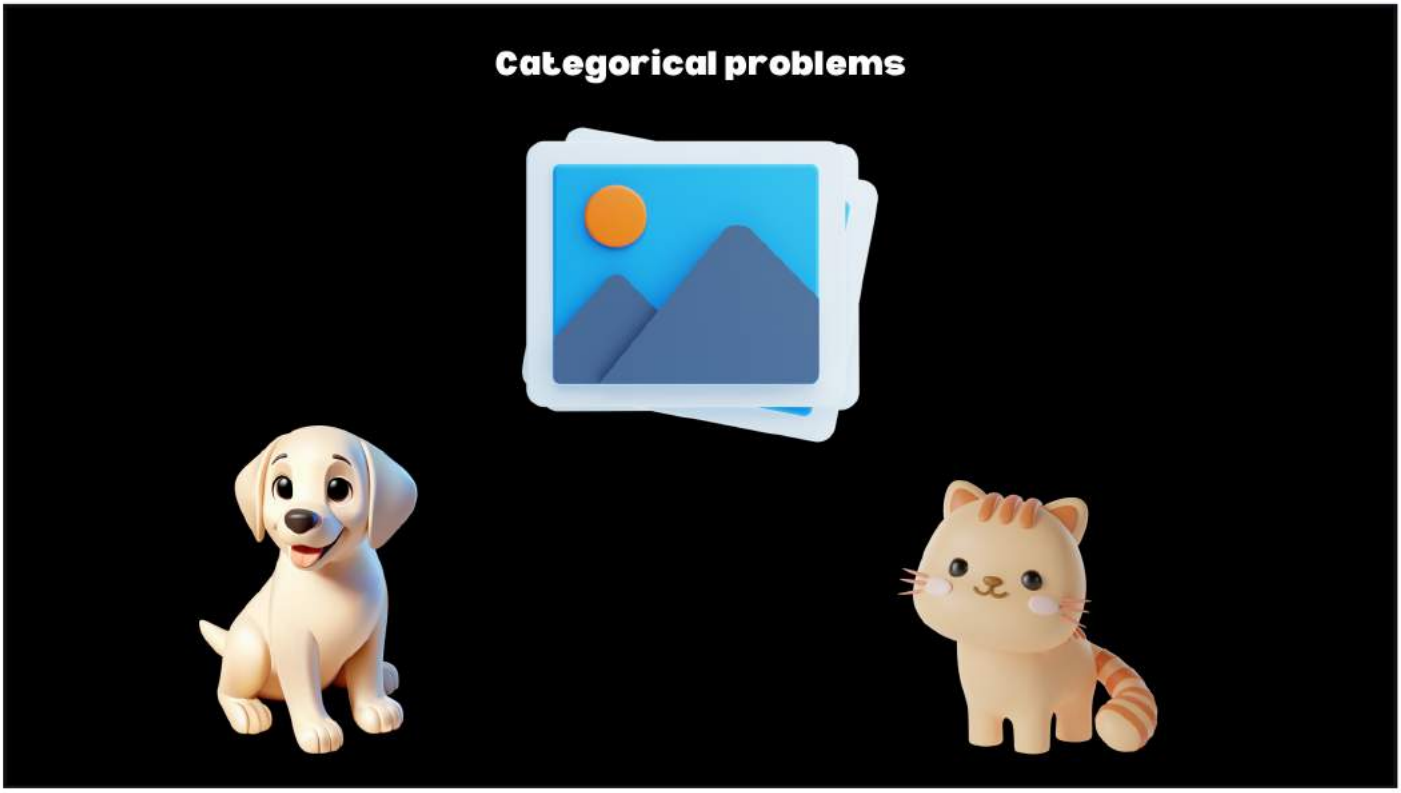
- Year
- Brand
- Model
- Colour
- State/City
- Mileage
- VIN

Linear Regression ML Model

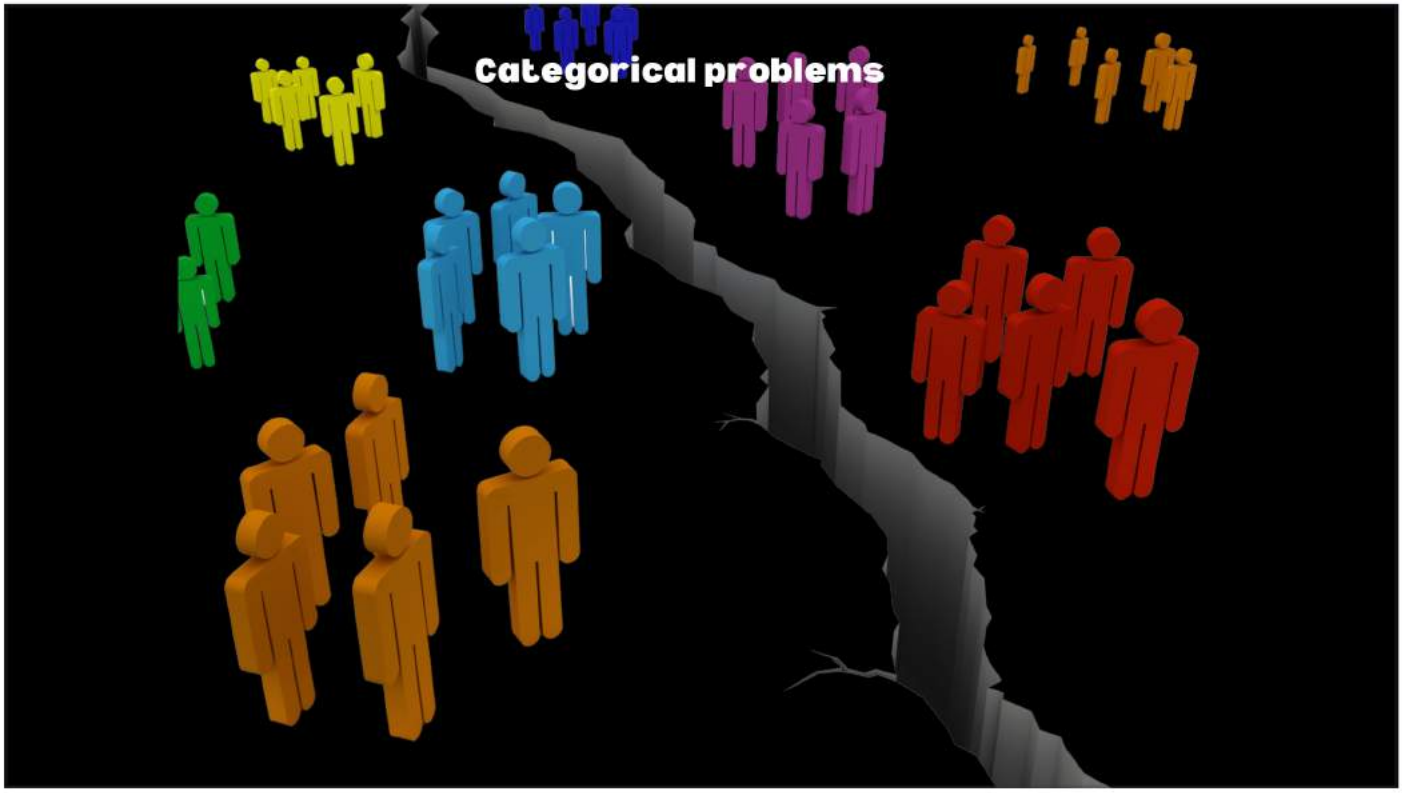
Equation: $y = b_0 + x_1 b_1 + x_2 b_2 + \ldots + x_n b_n$



For your **x**, figure out the best **b** to predict **y** using this equation


Categorical problems

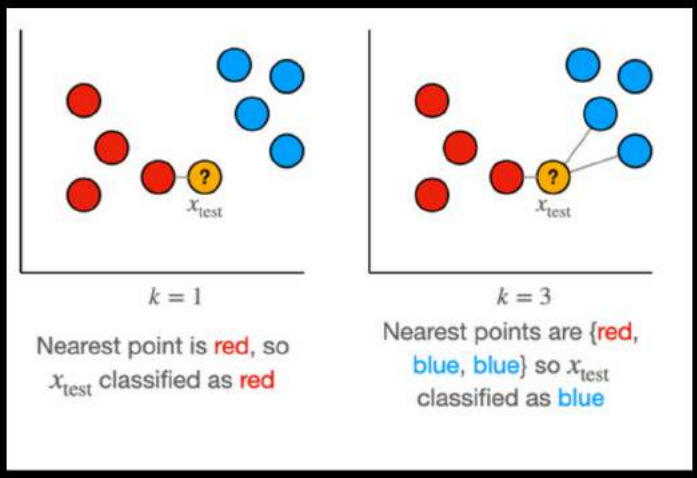

Ticket

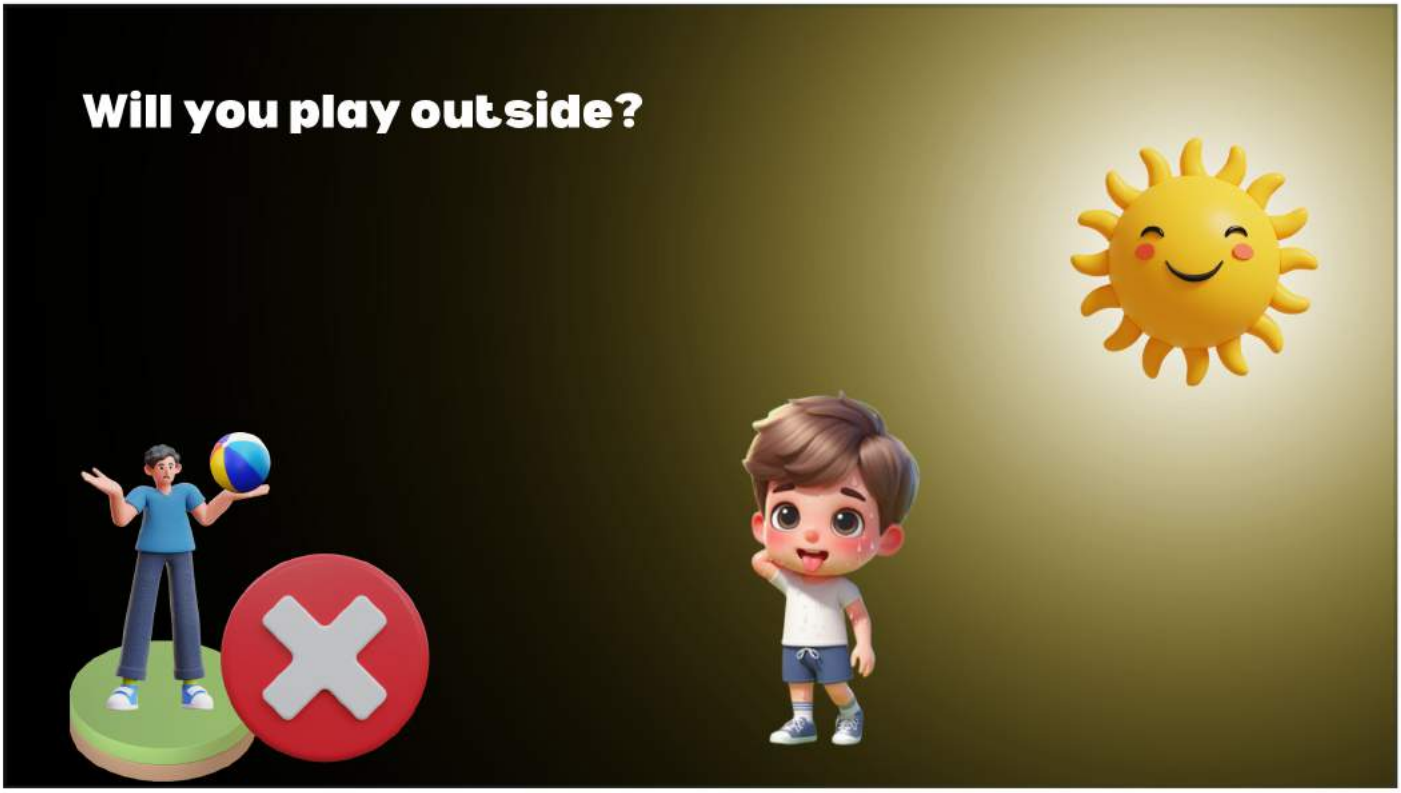

Categorical problems

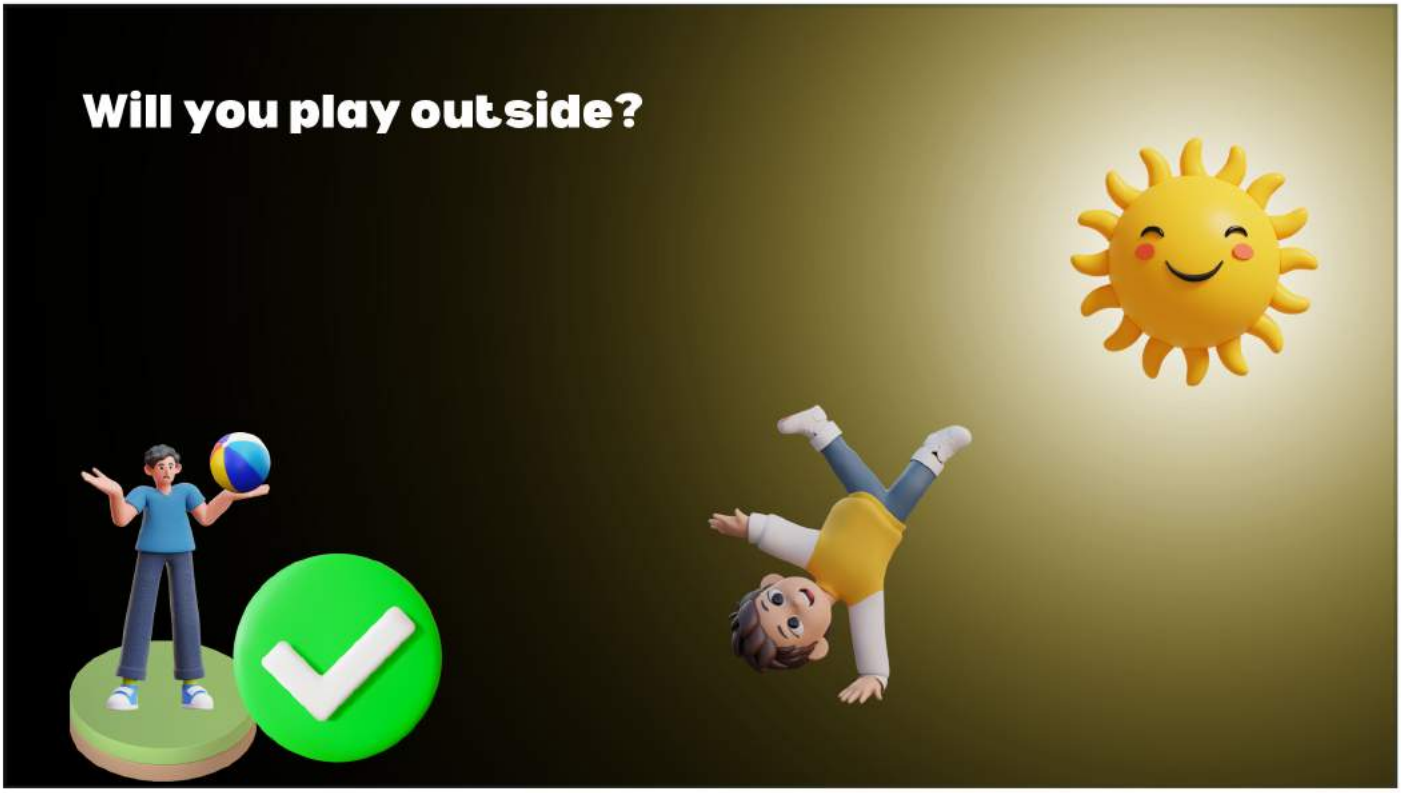
Think a little about it...

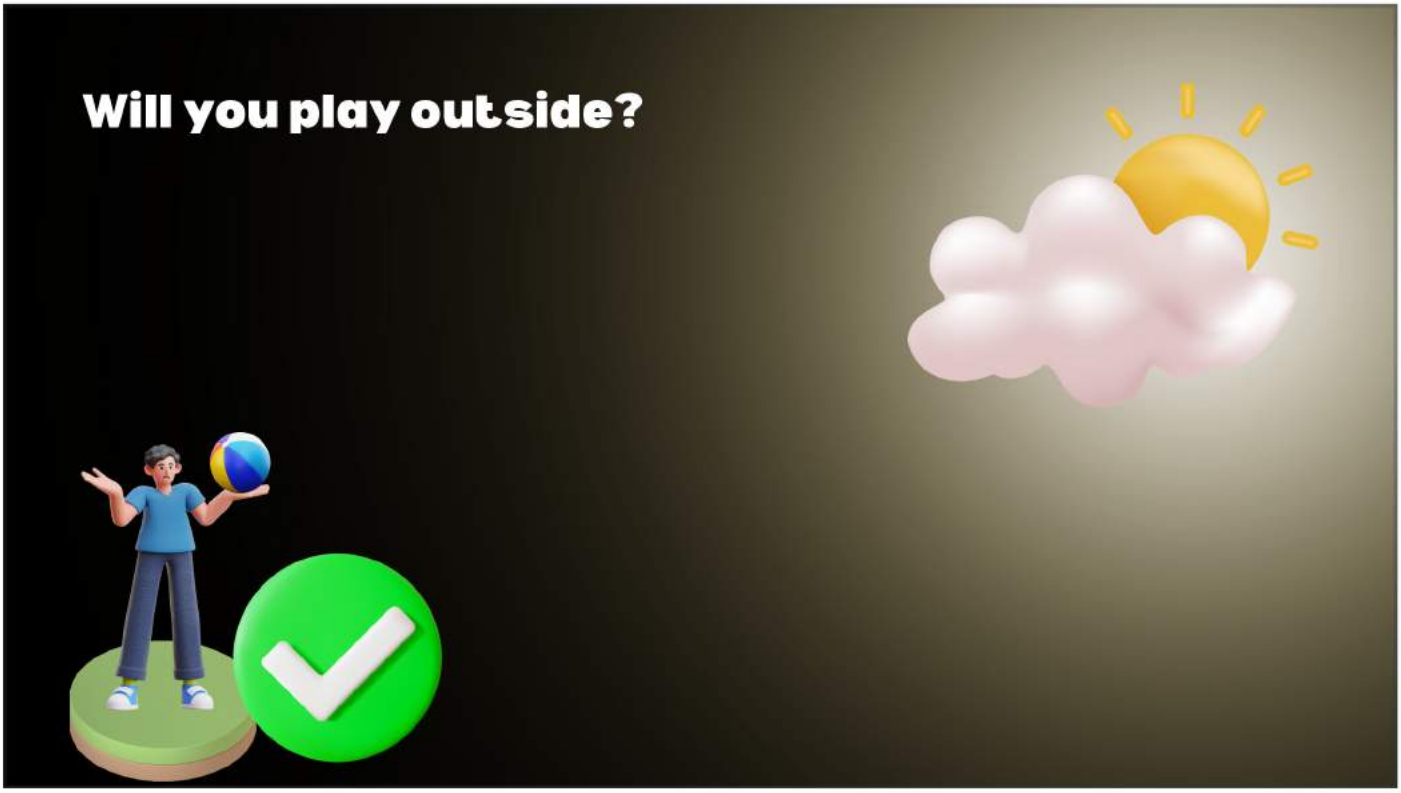
K-NN: K Nearest Neighbours ML Model

Classify a new instance using closest data points




Will you play outside?


Will you play outside?

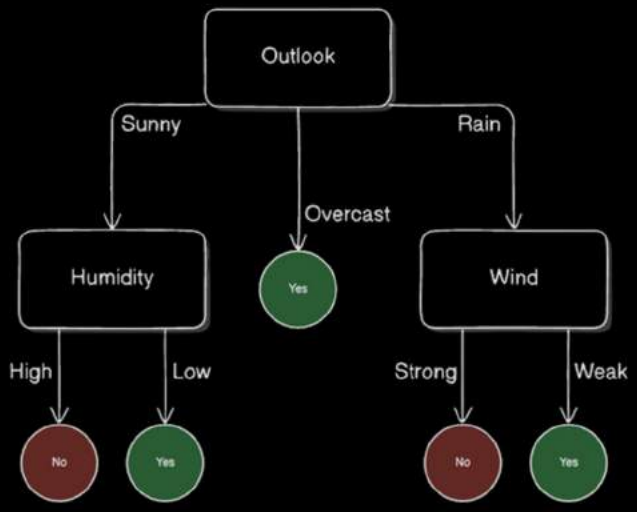

Will you play outside?


Will you play outside?


Will you play outside?

Decision Trees ML MODEL

Question: Will you play outside if
Outlook= Sunny; Humidity= high; Wind= Weak



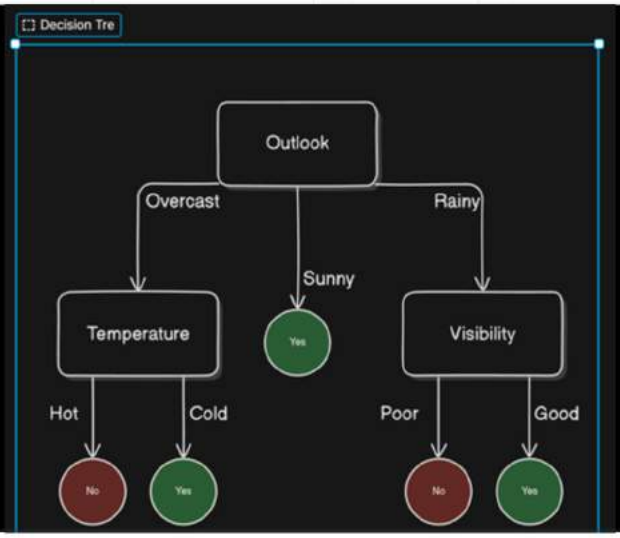
Your Turn!!!

Outcome variable: Play Tennis

Outlook	Temperature	Visibility	PlayTennis
Overcast	Cold	Good	Yes
Overcast	Hot	Good	No
Sunny	Hot	Good	Yes
Rainy	Cold	Good	Yes
Rainy	Cold	Poor	No

Answer:

Outlook	Temperature	Visibility	PlayTennis
Overcast	Cold	Good	Yes
Overcast	Hot	Good	No
Sunny	Hot	Good	Yes
Rainy	Cold	Good	Yes
Rainy	Cold	Poor	No



Un-supervised

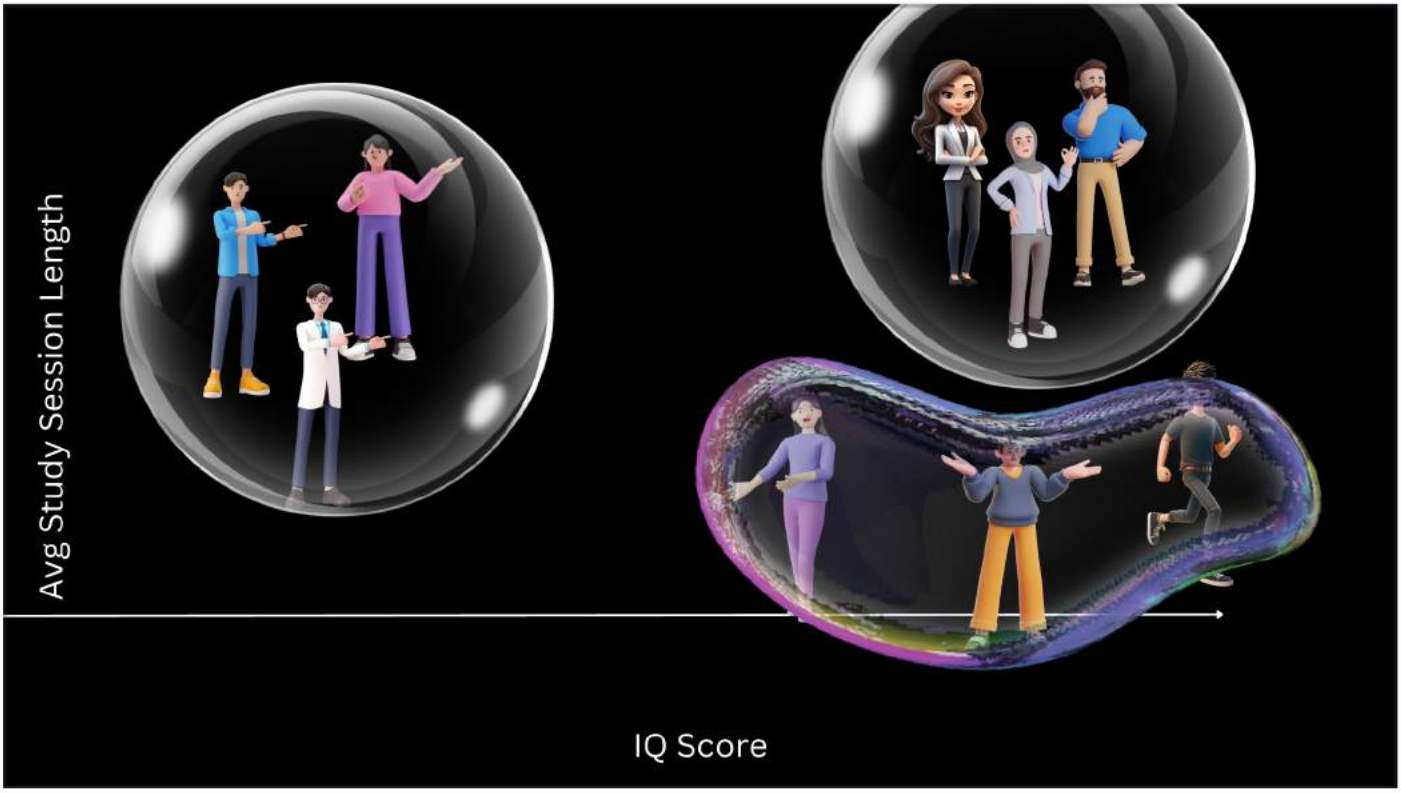
No set goal or something to predict

Model learns distinct structure and hidden patterns in data

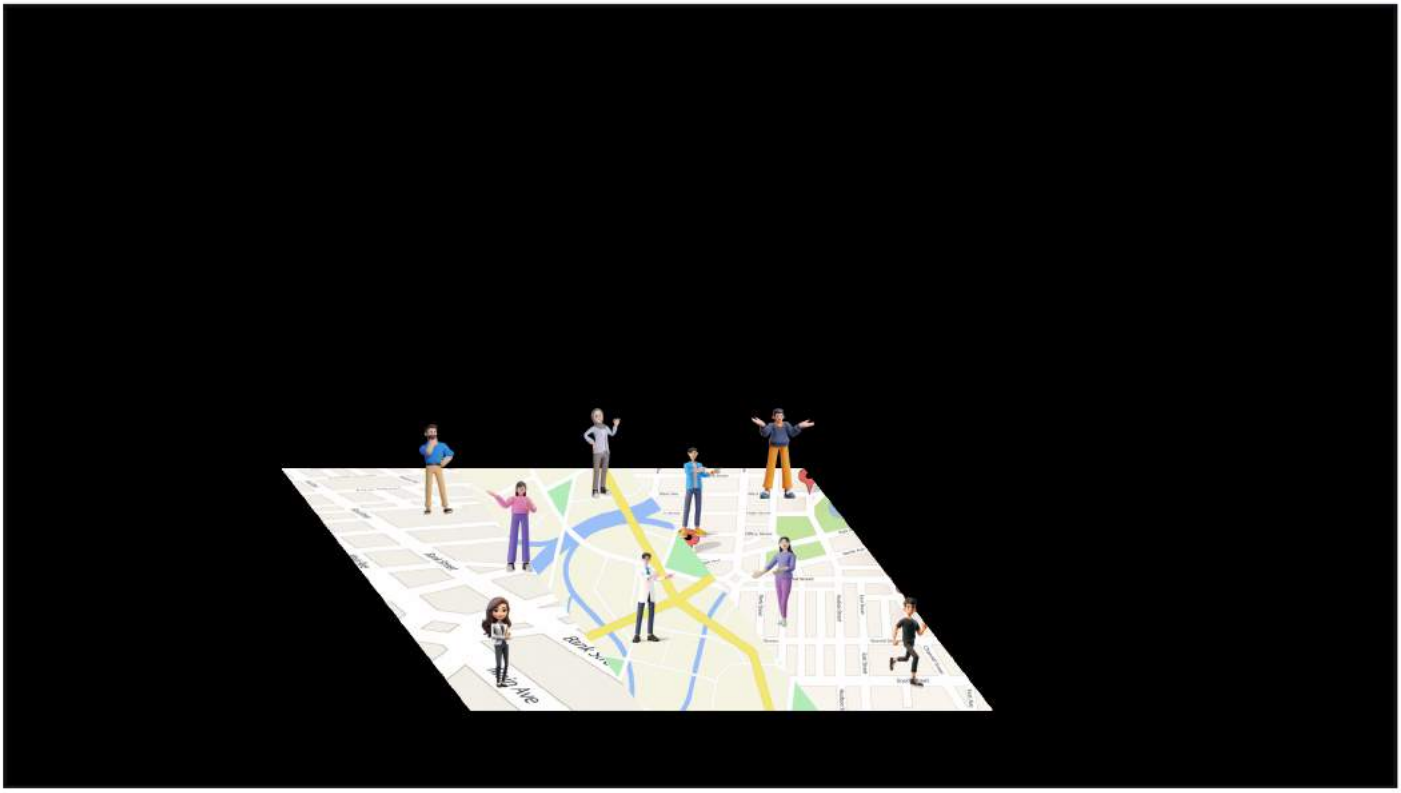
Examples: finding groups/ clusters that are similar within populations


Study Groups

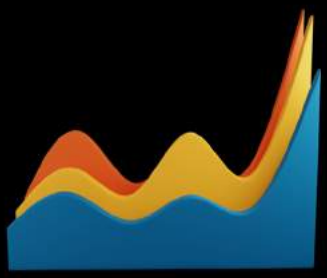

Avg Study Session Length
IQ Score


Avg Study Session Length
IQ Score


Avg Study Session Length
IQ Score




IQ



And a lot more!

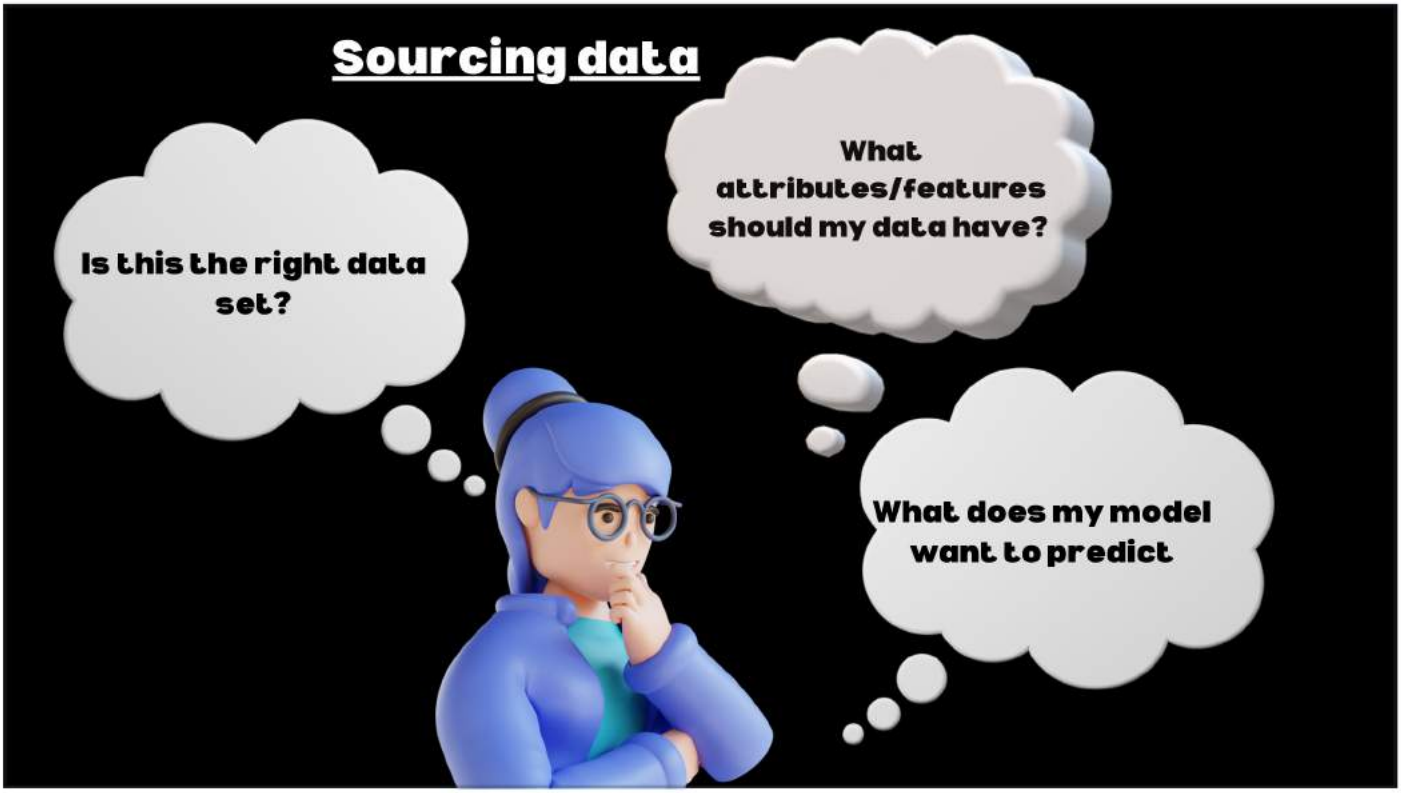
Supervised: Linear Regression, Logistic Regression, KNN, SVM, Naive Bayes, Random Forests, XGBoost, Neural Networks

Unsupervised: PCA, K-Means Clustering, Hierarchical Clustering, Association Rule Mining

Enough theory!!!

How to make our own ML model?

But after a 5 min break!


Sourcing data
Is this the right data set?
What attributes/features should my data have?
What does my model want to predict

Example

Aim: Predict a cars price

Dataset 1:

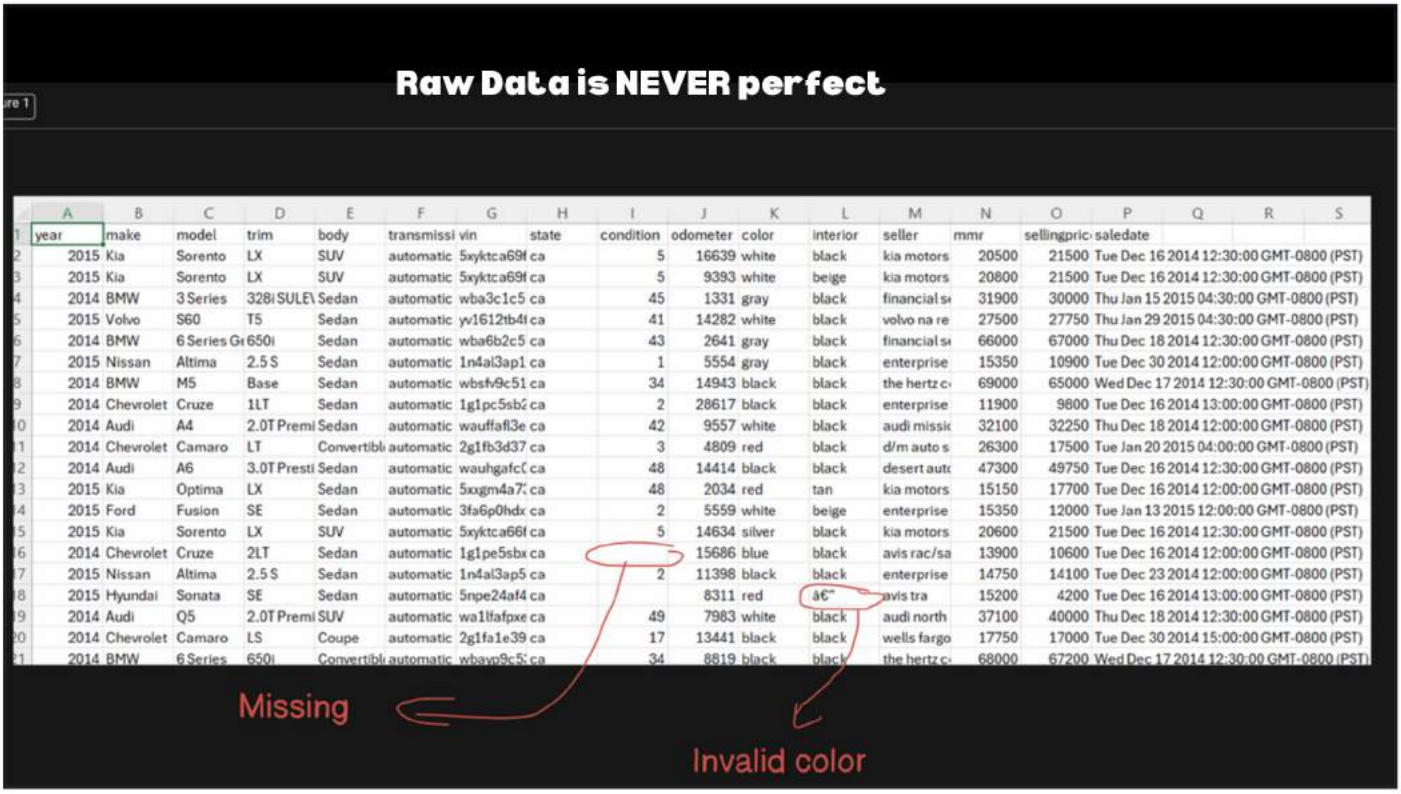
Feature	Type	Description
Price	Integer	The sale price of the vehicle in the ad
Years	Integer	The vehicle registration year
Brand	String	The brand of car
Model	String	model of the vehicle
Color	String	Color of the vehicle
State/City	String	The location in which the car is being availabie for purchase
Mileage	Float	miles traveled by vehicle
Vin	String	The vehicle identification number is a collection of 17 characters (digits and capital letters)
Title Status	String	This feature included binary classification, which are clean title vehicles and salvage insurance
Lot	Integer	A lot number is an identification number assigned to a particular quantity or lot of material from a single manufacturer.For cars, a lot number is combined with a serial number to form the Vehicle Identification Number.
Condition	String	Time

Dataset 2:

- ID : Identification number
- MS : Member state
- Mp : Manufacturer pooling
- VFN : Vehicle family identification number
- Mh : Manufacturer name EU standard denomination
- Man : Manufacturer name OEM declaration
- MMS : Manufacturer name MS registry denomination
- TAN : Type approval number
- T : Type
- Va : Variant
- Ve : Version
- Mk : Make
- Cn : Commercial name
- Ct : Category of the vehicle type approved
- Cr : Category of the vehicle registered
- m (kg) : Mass in running order complete vehicle
- Mt : WLTP test mass

Are all of these features useful?

Feature	Type	Description
Price	Integer	The sale price of the vehicle in the ad
Years	Integer	The vehicle registration year
Brand	String	The brand of car
Model	String	model of the vehicle
Color	String	Color of the vehicle
State/City	String	The location in which the car is being available for purchase
Mileage	Float	miles traveled by vehicle
Vin	String	The vehicle identification number is a collection of 17 characters (digits and capital letters)
Title Status	String	This feature included binary classification, which are clean title vehicles and salvage insurance
Lot	Integer	A lot number is an identification number assigned to a particular quantity or lot of material from a single manufacturer.For cars, a lot number is combined with a serial number to form the Vehicle Identification Number.
Condition	String	Time

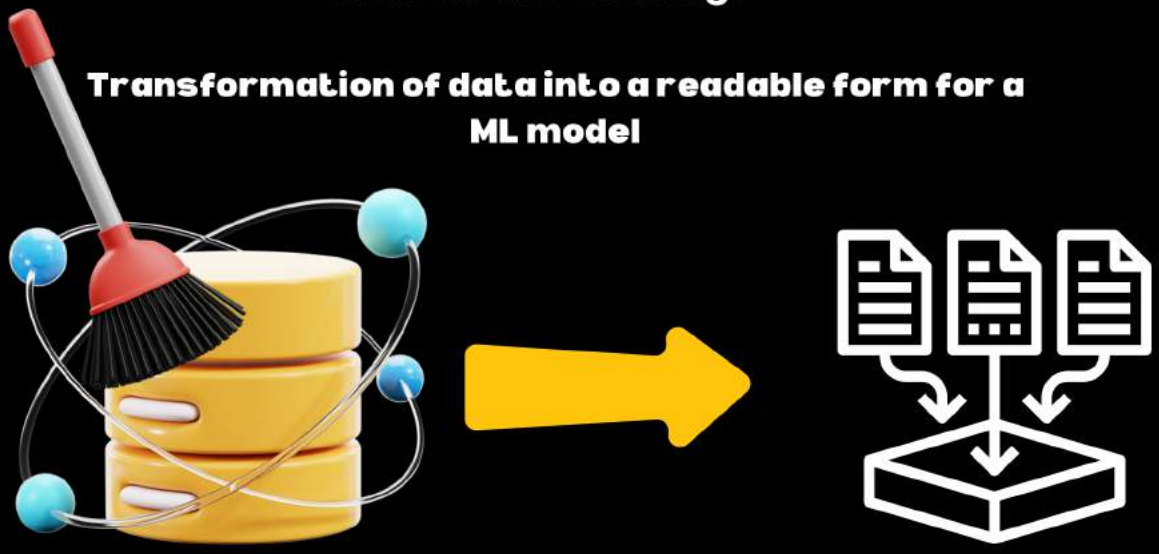


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	year	make	model	trim	body	transmissi	vin	state	condition	odometer	color	interior	seller	mmr	sellingpric	saledate			
2	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69f	ca	5	16639	white	black	kia motors	20500	21500	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)			
3	2015	Kia	Sorento	LX	SUV	automatic	5xyktca69f	ca	5	9393	white	beige	kia motors	20800	21500	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)			
4	2014	BMW	3 Series	328i SULE	Sedan	automatic	wba3c1c5	ca	45	1331	gray	black	financial s	31900	30000	Thu Jan 15 2015 04:30:00 GMT-0800 (PST)			
5	2015	Volvo	S60	T5	Sedan	automatic	yv1612tb4	ca	41	14282	white	black	volvo na re	27500	27750	Thu Jan 29 2015 04:30:00 GMT-0800 (PST)			
6	2014	BMW	6 Series G	650i	Sedan	automatic	wba6b2c5	ca	43	2641	gray	black	financial s	66000	67000	Thu Dec 18 2014 12:30:00 GMT-0800 (PST)			
7	2015	Nissan	Altima	2.5 S	Sedan	automatic	1n4al3ap1	ca	1	5554	gray	black	enterprise	15350	10900	Tue Dec 30 2014 12:00:00 GMT-0800 (PST)			
8	2014	BMW	M5	Base	Sedan	automatic	wbsfv9c51	ca	34	14943	black	black	the hertz c	69000	65000	Wed Dec 17 2014 12:30:00 GMT-0800 (PST)			
9	2014	Chevrolet	Cruze	1LT	Sedan	automatic	1g1pc5sb2	ca	2	28617	black	black	enterprise	11900	9800	Tue Dec 16 2014 13:00:00 GMT-0800 (PST)			
10	2014	Audi	A4	2.0T Premi	Sedan	automatic	wauffafl3e	ca	42	9557	white	black	audi missi	32100	32250	Thu Dec 18 2014 12:00:00 GMT-0800 (PST)			
11	2014	Chevrolet	Camaro	LT	Convertibl	automatic	2g1fb3d37	ca	3	4809	red	black	d/m auto s	26300	17500	Tue Jan 20 2015 04:00:00 GMT-0800 (PST)			
12	2014	Audi	A6	3.0T Presti	Sedan	automatic	wauhgafc(	ca	48	14414	black	black	desert aut	47300	49750	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)			
13	2015	Kia	Optima	LX	Sedan	automatic	5xxgm4a7:	ca	48	2034	red	tan	kia motors	15150	17700	Tue Dec 16 2014 12:00:00 GMT-0800 (PST)			
14	2015	Ford	Fusion	SE	Sedan	automatic	3fa6p0hdx	ca	2	5559	white	beige	enterprise	15350	12000	Tue Jan 13 2015 12:00:00 GMT-0800 (PST)			
15	2015	Kia	Sorento	LX	SUV	automatic	5xyktca66f	ca	5	14634	silver	black	kia motors	20600	21500	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)			
16	2014	Chevrolet	Cruze	2LT	Sedan	automatic	1g1pe5sbx	ca		15686	blue	black	avis rac/sa	13900	10600	Tue Dec 16 2014 12:00:00 GMT-0800 (PST)			
17	2015	Nissan	Altima	2.5 S	Sedan	automatic	1n4al3ap5	ca	2	11398	black	black	enterprise	14750	14100	Tue Dec 23 2014 12:00:00 GMT-0800 (PST)			
18	2015	Hyundai	Sonata	SE	Sedan	automatic	5npe24af4	ca		8311	red	â€"	avis tra	15200	4200	Tue Dec 16 2014 13:00:00 GMT-0800 (PST)			
19	2014	Audi	Q5	2.0T Premi	SUV	automatic	wa1lfafpxe	ca	49	7983	white	black	audi north	37100	40000	Thu Dec 18 2014 12:30:00 GMT-0800 (PST)			
20	2014	Chevrolet	Camaro	LS	Coupe	automatic	2g1fa1e39	ca	17	13441	black	black	wells fargo	17750	17000	Tue Dec 30 2014 15:00:00 GMT-0800 (PST)			
21	2014	BMW	6 Series	650i	Convertibl	automatic	wbayp9c5:	ca	34	8819	black	black	the hertz c	68000	67200	Wed Dec 17 2014 12:30:00 GMT-0800 (PST)			

Data Cleaning

What is Data Cleaning?

Transformation of data into a readable form for a ML model

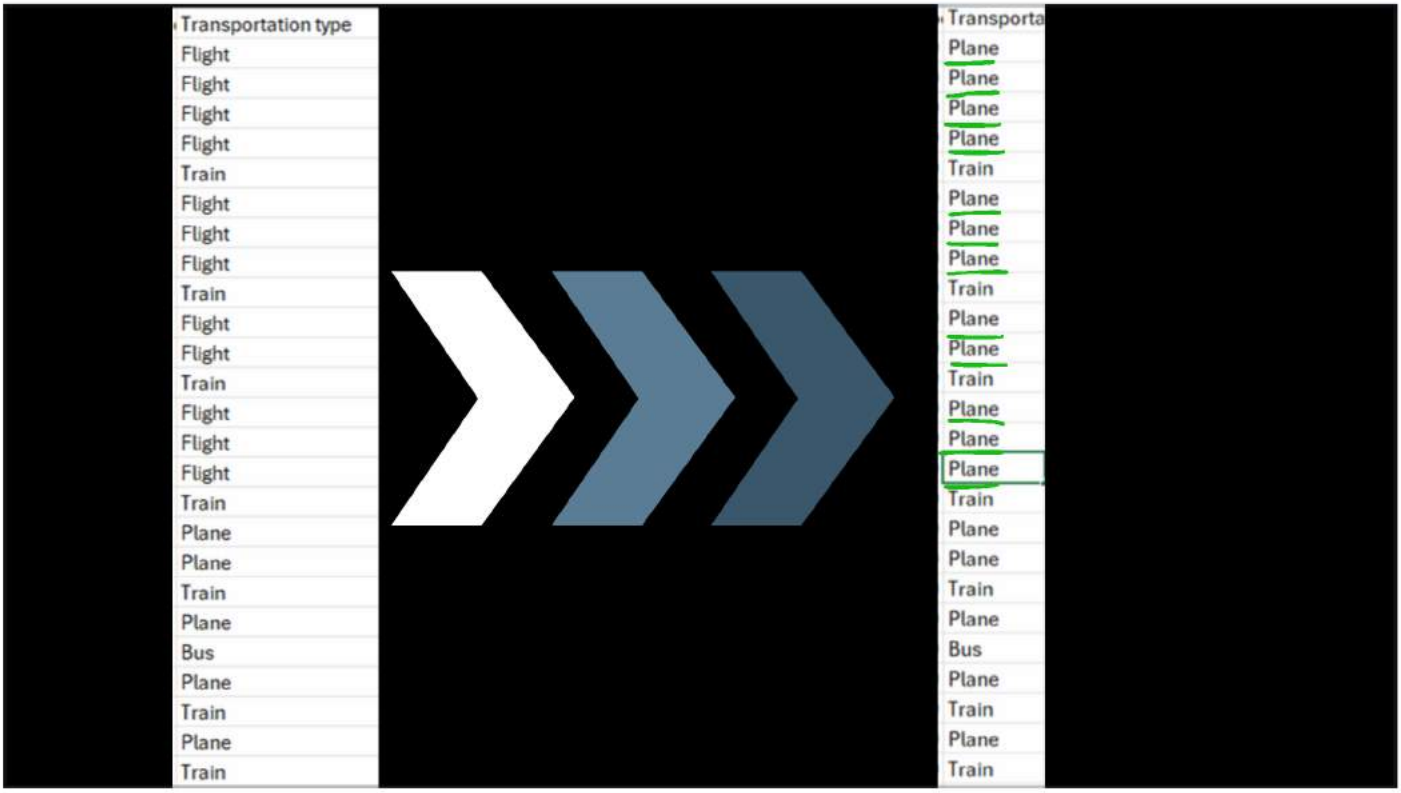


Problems with data:

Unique values: Flight, Plane, Bus, Train, Care Rental.

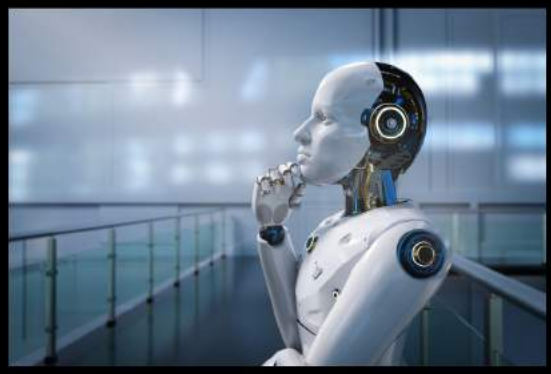
Same meaning

What might be the problem with this?


Transportation type
Flight
Flight
Flight
Flight
Train
Flight
Flight
Flight
Train
Flight
Flight
Train
Flight
Flight
Flight
Train
Plane
Plane
Train
Plane
Bus
Plane
Train
Plane
Train
Transporta
Plane
Plane
Plane
Plane
Train
Plane
Plane
Plane
Train
Plane
Plane
Train
Plane
Plane
Plane
Train
Plane
Plane
Train
Plane
Bus
Plane
Train
Plane
Train

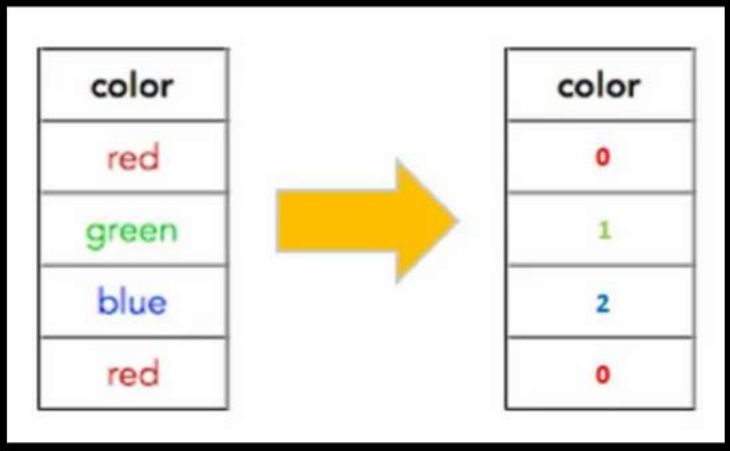
ML Models can only take on numeric values

Turn all nominal data -> numerical



Solution: Label encoding

What is Label encoding?



Different types of encoding

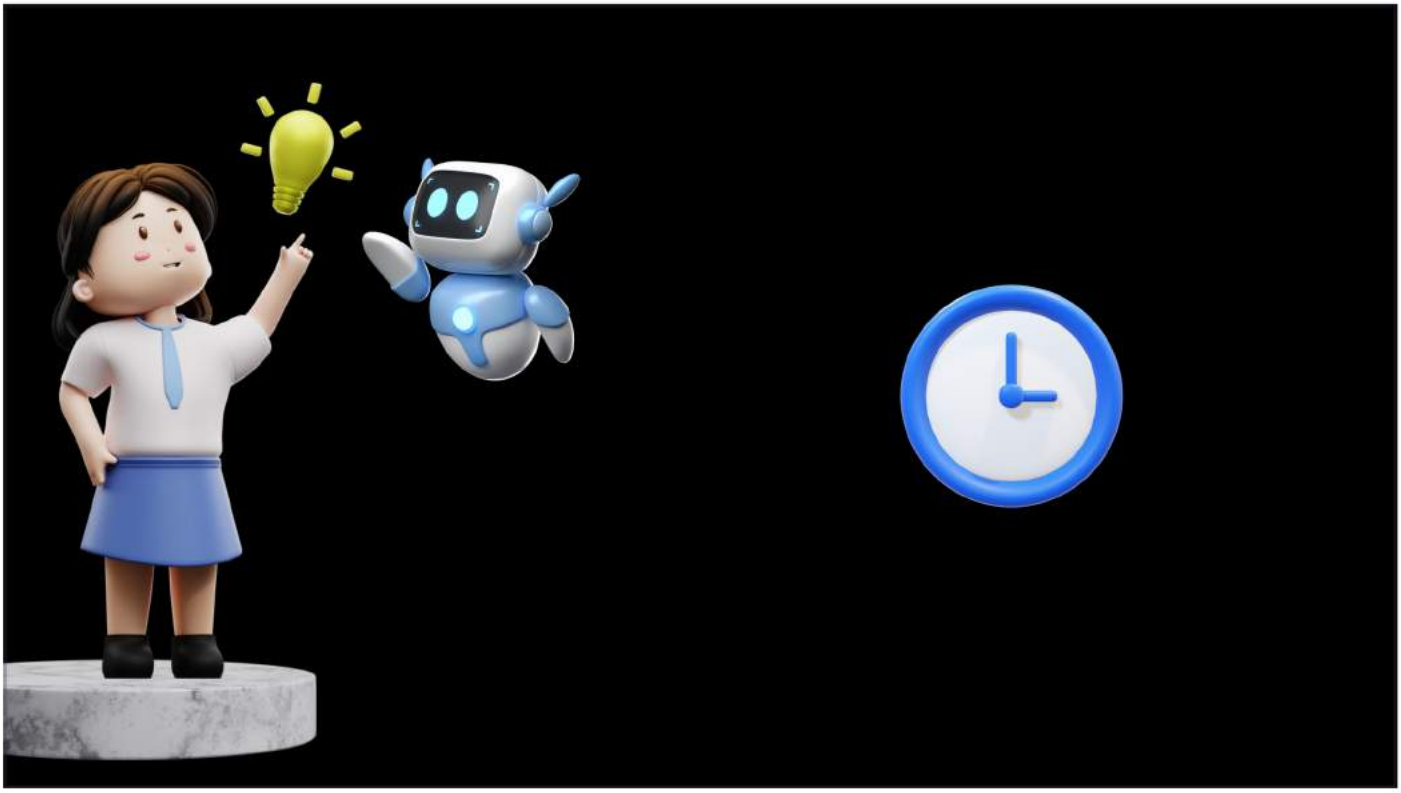
Frequency encoding

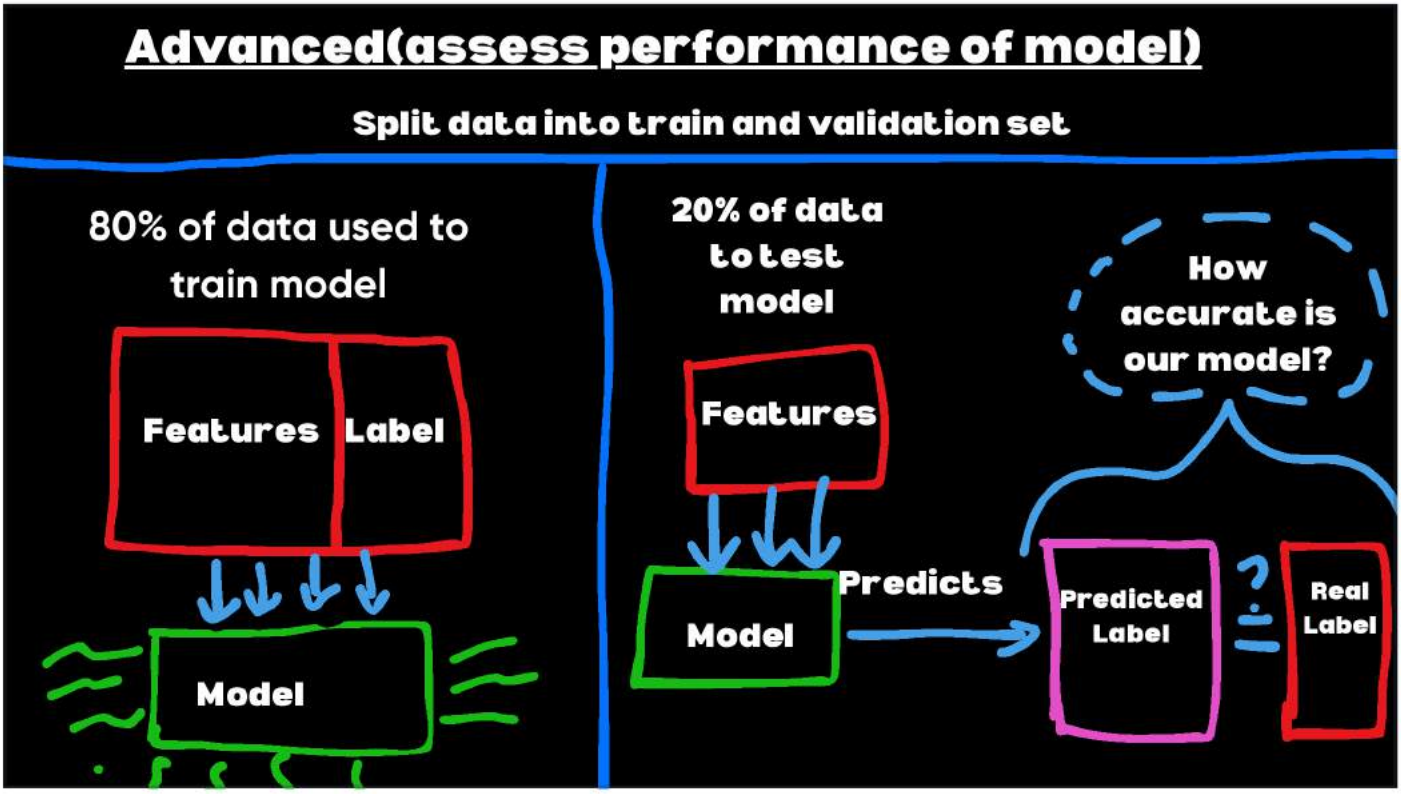
Ordinal encoding

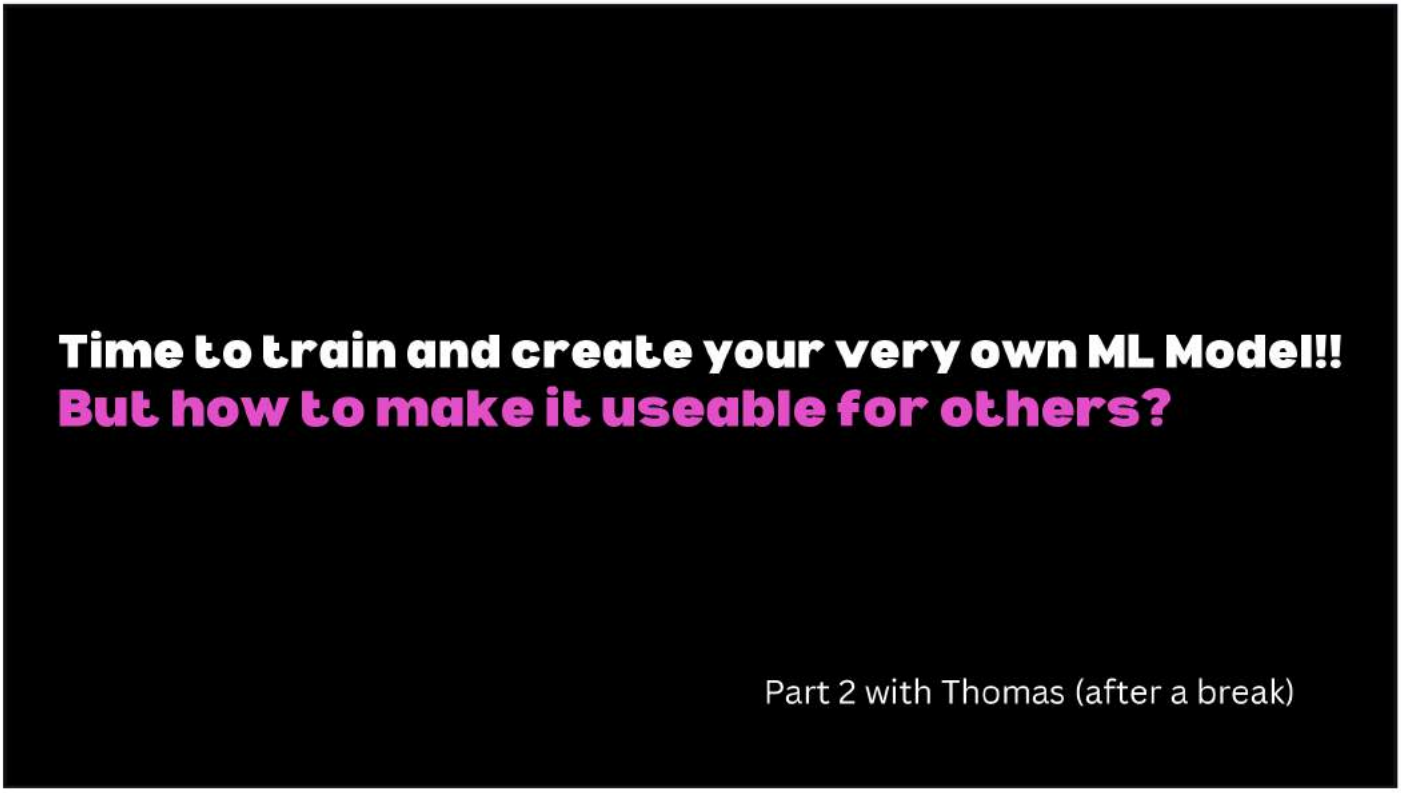
Mean encoding

Many more!!: https://medium.com/@brandon93.w/converting-categorical-data-into-numerical-form-a-practical-guide-for-data-science-99fdf42d0e10#:~:text=Frequency%20Encoding,with%20high%20cardinality%20categorical%20data.

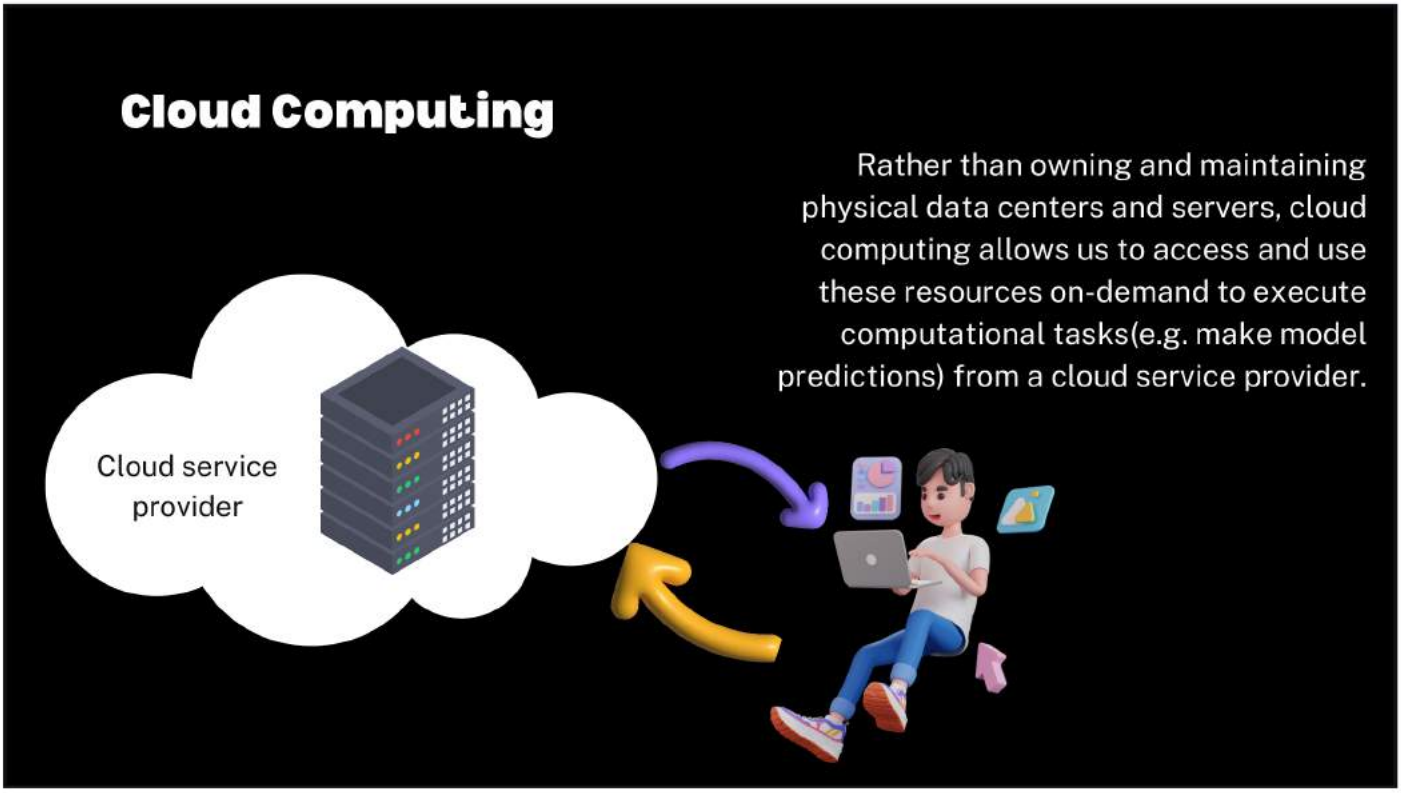
Long day, but are we done yet?




Advanced(assess performance of model)
Split data into train and validation set
80% of data used to train model
Features
Label
Model
20% of data to test model
Features
Model
Predicts
Predicted Label
Real Label
How accurate is our model?



Thank you Bike for giving a great introduction of those basic machine learning models and concepts. Hello everyone, I'm Thomas and I'm going to talk about the model deployment to AWS. After you have built your own amazing model predicting the travel transportation cost, car prices or even stock market price with a very high accuracy, how to present your model to your audience and make it accessible and useable for the users? Here comes the model deployment. We will put the model to the cloud and authorised people can ask the cloud to predict the outcome variable using the model through API calls. That'll be what we are going to do in the second part of the workshop and you will need this deployment skills in the hackathon.



First of all, what is cloud computing? Cloud computing is essentially a group of data centers scattering around the globe to help perform computational tasks and run our software projects. Before the era of cloud computing, we may have to build and maintain physical servers at our home, which is tedious. Now we reserve some virtual machines in the data center of the cloud provider to run our projects by just clicking some buttons on the cloud provider console, and your personal website is on the cloud and everyone can get to it.

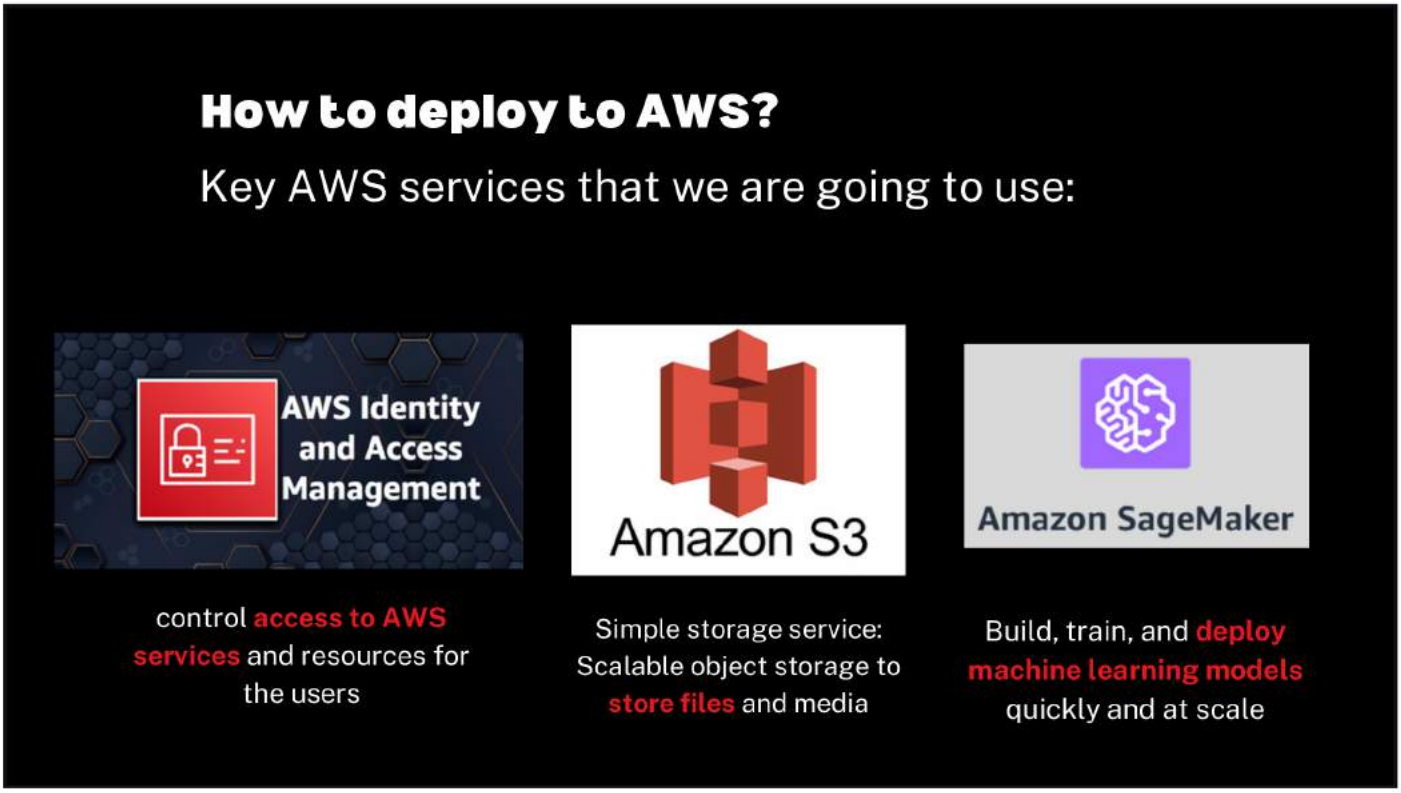
Model Deployment on AWS

AWS (Amazon Web Services)
offers a wide range of powerful cloud services for businesses and individuals

Examples: Compute services(EC2, lambda), storage services(S3), database services(RDS), networking(VPC), security and identity services(IAM) and many more...

Industry leader in cloud computing

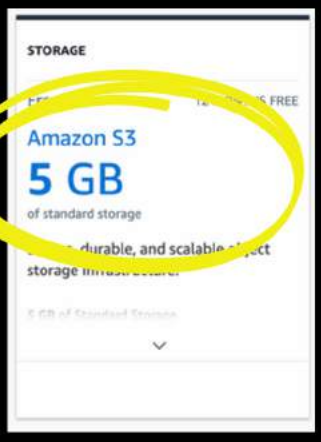
And we are using one of the cloud service provider, AWS, the Amazon Web Services, which provides a wide range of powerful cloud services for businesses and individuals. Have anyone of you heard about these cloud providers before? AWS is the industry leader in cloud computing. It leads over 30% in the market share in the world and once you unlock the power of AWS, you can build almost any software projects you can think of in the cloud. On the slide, you can see some of the popular and services such as the compute services EC2 to help us create and manage virtual machines, storage services S3 to store files and media on the cloud, virtual private cloud VPC to create a private network and many more. There are over 200 services in AWS, I understand it may be a bit overwhelming to learn AWS in just 1 but it will be fine because we are just introducing the basis and using only 3 services.

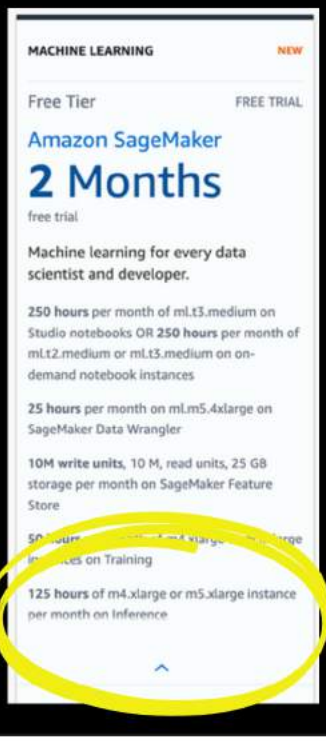


These are the 3 services that we are going to use on AWS. FIrst of all, the IAM,

No Free Lunch But Free Tier!

Discover on your own:
https://shorturl.at/rRKvW



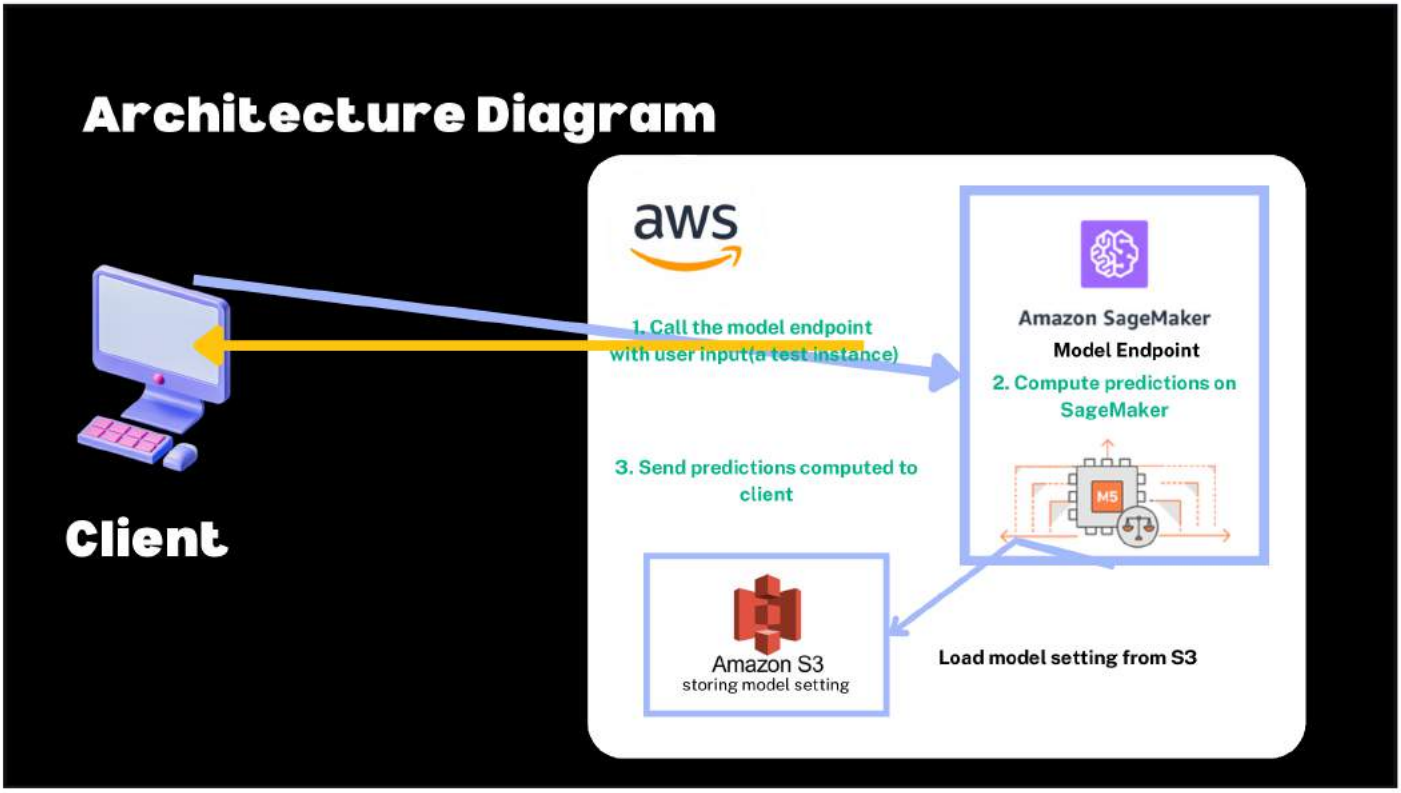


P.S. Credit or debit card required in account registration

The demo won't incur any cost :))

Around 5 days of free model inference computation

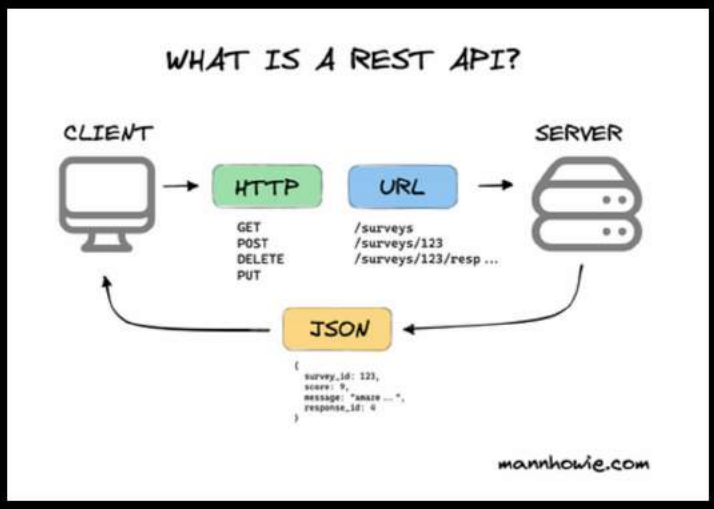
Architecture Diagram



What is Endpoint

A point of entry in a communication channel when 2 systems are interacting.
Technical: A specific URL that an API(Application Programming Interface) exposes to interact with a server

SageMaker provides a REST API endpoint with the POST method for users to make requests with input in the request body and receive the output



Live Demo of Model Deployment

Demo instructions included in Notion Blog Post

Improvements(1) - API Gateway

Tutorial: https://shorturl.at/e69Y2

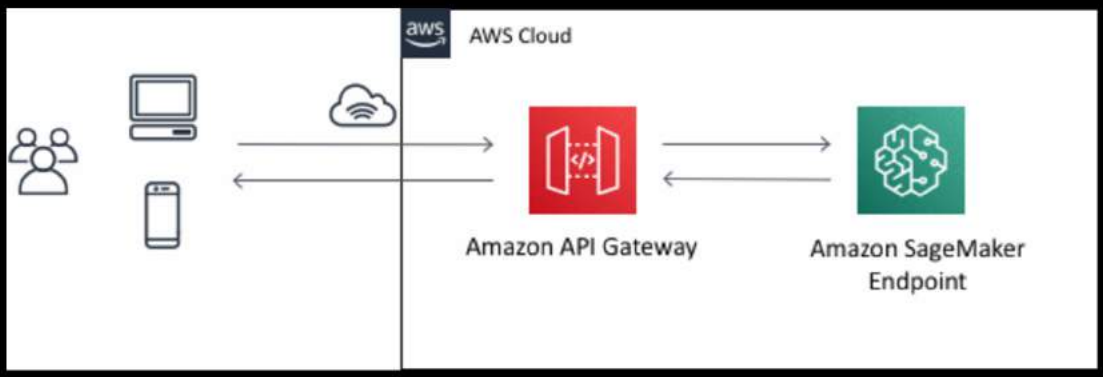
API Gateway is a fully managed service to create, maintain, monitor, and secure APIs at any scale.

Integrating with API Gateway can:

- Create a client-facing REST API and hide the underlying Amazon SageMaker runtime inference API
- Supports authentication and authorization of client requests
- Manages client requests through the use of throttling, rate-limiting, and quota management
- Uses firewall features as provided by AWS WAF
- ...

In addition to the techs we used in the demo, we can integrate other powerful AWS services into SageMaker to improve the architecture and make it more reliable. The first one will be the API Gateway, which is a fully-managed service dedicated to create, maintain, monitor and secure APIs at any scale. We can create a GET endpoint on top of the SageMaker endpoint to first process the test instance and pass it to SageMaker for inference. Then, we can transform the prediction result from SageMaker before sending it to the client, which can hide the underlying SageMaker runtime inference API from the frontend. In addition, by connecting to API gatway, we can take advantages of other more AWS service such as the firewall features AWS WAF to improve the security of our endpoint. More advanced and sophisticated authentication and authorisation can also be introduced with API gateway as well. However, this is just some advanced improvement on top and you'll be fine if you just follow the instructions on Notion. If you want more challenges, go for this integration with API gateway.

Architecture Diagram with AWS API Gateway

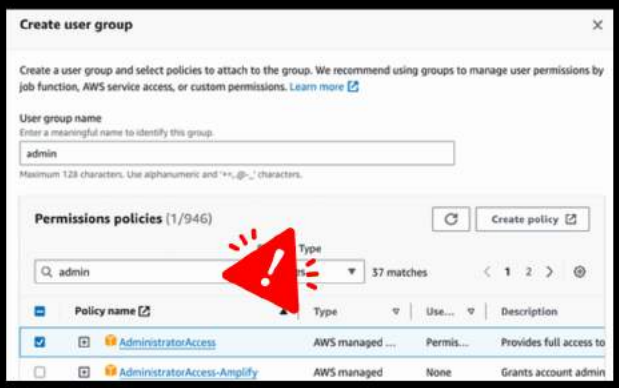


This is the architecture diagram of the improvement in the previous slide.

Improvements(2) - IAM

Least-privilege permissions

Only grant the minimum permissions required to perform a task to users, roles and user groups, enhancing the security of the system.



The second improvement we can adopt is the least-privilege permissions. Remeber when we create a user group in IAM for our computer to access the AWS cloud, we have added the full administratorAccess policy to the user group. However, this permission is much stronger than what we really need to perform the tasks in demo. Instead, we should only grant the minimum permissions required to perform a task to users, roles, user groups, in order to improve the security of the system. And this practice on cloud computing is called the least-privilege permissions.

That's the last slide of our workshop. Thank you so much for attending it and I hope you will VicHack.