1 What is AI and its importance for businesses

okay So in uh in this video I'll I'll give the lecture for week one which is
um which is populated with all this content here Uh and uh from that I I
crafted some goals uh obviously we can change this uh and
this is a draft version of week one Uh again all the the figures images and
slides they are very drafty Um but I hope that with this we can kind of uh
better differentiate the se the sections of week one and what could be taken to
on
online learning um kind of self-study learning or not
So I'm going to start the course I I will going to teach as if I am giving a class in a
in a in a uh in a classroom
with students Um and uh this will generate kind of a script uh usually
this online courses you have the script uh you know very
um flowing very well uh and here in here because it's a lecture uh you will see
lots of a and and and stopping points and comments but this is an
underlying material for uh a more a more kind of movie script
after Okay So let's uh let's start the the course
I'm going to start the I'm going to start the the
uh the lecture now Okay So just one second
Okay Okay Welcome to the course In this module we will
explore artificial intelligence what it really means how it learns
and by doing that we will face several examples of how AI is being used across
industries uh to create value We know that AI is everywhere
uh powering recommendation systems on Netflix or YouTube optimizing
logistics
for Amazon or even in our cell phones uh it's it's for sure everywhere and
it's it's becoming vital of any modern business So what is artificial
intelligence Artificial intelligence is about getting computers to
do things that uh require human intelligence So understanding language
is an example reasoning is another
example For examp uh navigating the physical world learning predicting the
future All of these are underlying tasks that requires human intelligence and
that when we set computers to do it we're talking about artificial intelligence
AI is being uh is seen increasingly everywhere and
for sure is uh said to be the next phase of digital
transformation If we look at the um at the
history in the late 1990s the internet uh was the digital technology that
transformed businesses Then cloud computing came in Then mobile computing
in the late 2000s with the internet of things And now AI is the digital
technology that has the potential to transform businesses So companies
that were slow to react to the emergency of this uh digital technologies

uh were left behind uh and therefore from experience we know that companies now need to react uh to this new emergence of
AI So what are the implications for
businesses First of all the f the the f uh the first thing is to know that even if you're not in an industry that is IT related
um this this digital transformation will affect your business because AI is everywhere and it's a general technology that is being used in several different domains So managers need to understand this technology its
application and uh make changes to
uh to embark on this uh digital trend
These changes can be uh different business models uh changing technology infrastructure
organizational process uh processes and the culture of the business uh as well

So to let's start diving deep into

AI So AI is a fancy name This was my dog Sorry AI is a fancy name again uh to say
that a computer is doing what a human would be doing right There are uh there's a taxonomy of AI uh and usually experts agree in this that there are two different uh big groups or two different approaches to AI The first approach is expert systems and the second approach is learning systems
So expert systems are AI So computers are doing what humans would do but they
are hardcoded So their knowledge comes from a human typing in the
rules One example of this is uh diagnostic systems in healthcare where doctors create a set of rules from their knowledge they hardcode
that in the computer and then the computer kinds of gives the diagnos uh the diagnosis What the characteristics of expert systems is that it relies on this hard-coded rules uh written by domain experts But this uh has
some um some uh challenges and limitations So it struggles it it
struggles with uh un unexpected situations or ambiguity when these rules are um when these rules are overlapping for
example So it is inspired in in the human intelligence right of the of reasoning with logic Uh but it's hardcoded
uh there are more modern examples of rule-based uh AI systems for example uh credit score engines that banks use uh uses or ellegibility for insurance claims So we have predefined rules that determine the outcomes and we will code
these rules Sometimes we need the computer because to to aggregate all these rules with a human mind would be very difficult Uh so we're not saying that these systems are not good They're just limited in some aspects
Uh so for example an AI chess player and
uh uh could could be an expert system that has all the rules of chess inserted

into it So if something changes in the
program uh in the in the environment in the game that is not hardcoded the AI
will have problem problems with it Now the other branch of artificial
intelligence is the machine learning branch or learning
systems So this branch enables computers to learn from data and improve their
performance on tasks over time without being explicitly programmed
for the task So you want the computer to
uh to perform a task and you don't you won't explicitly
tell the computer how to do it You will just make the computer learn from first
principles Um so this branch of artificial
intelligence can generalize to new tasks and solve things that we don't know how
to solve We don't know the rules Uh and it was loosely inspired uh and and
validated by neuroscience Okay So uh we can we can give some examples
of machine learning uh mo of machine learning systems or learning systems
uh when a machine wants to predict whether a customer will purchase a
product right It can it can analyze the browsing behavior the time spent on the
product page and past transactions to estimate the likelihood
of of uh of purch of p of the client purchasing a product
So there is no rule there is no kind of specific hard-coded model that uses
browsing behavior uh time spent in in web pages and the transactions history
to estimate the the likelihood of conversions We actually need data to
learn So um there are several other examples
that we will come across but again we are enabling computers to learn from the
data uh and perform the task So in this case the computer will
be giving several and several and several examples of customers and it will
extract from the data a knowledge
that kind of combines all of this uh features of browsing behavior uh time
spent on product pages and past transactions to estimate the likelihood
of a conversion So this will be the main focus of this
course So this course is mainly focused in machine learning Um and also uh in
technical AI
literature sometimes you will see that expert systems are called symbolic AI
and learning systems are called machine learning or um statistic statistical
learning

Welcome to the course. In this module, we'll explore artificial intelligence—what it really means, how it learns, and how it's being used across industries to create value.

AI is everywhere: powering recommendation systems on Netflix and YouTube, optimizing logistics for Amazon, and embedded in our smartphones. It has become essential to modern business.

So, what is artificial intelligence?

Artificial intelligence is about getting computers to do things that require human intelligence. Understanding language, reasoning, navigating the physical world, learning, and predicting the future are all tasks that require human intelligence. When we program computers to perform them, we call it artificial intelligence.

AI is increasingly present in our lives and is considered the next phase of digital transformation. In the late 1990s, the internet transformed businesses. Then came cloud computing, followed by mobile computing and the Internet of Things in the 2000s. Now, AI is the emerging digital technology poised to transform businesses.

Companies that were slow to respond to past digital transformations were left behind. We know from experience that businesses must act quickly in response to AI.

What are the implications for businesses?

First, even if you're not in an IT-related industry, this transformation will still affect your business. AI is a general-purpose technology being applied across many domains. Managers need to understand the technology, its applications, and make the necessary changes to embrace this digital shift. These changes may involve new business models, updated technology infrastructure, revised organizational processes, and changes in company culture.

Let's start diving deeper into AI.

AI is a term used to describe computers performing tasks that humans typically do. There is a common taxonomy of AI, and experts usually agree on two major categories: **expert systems** and **learning systems**.

Expert systems involve computers doing what humans would do, but with hardcoded rules. The knowledge in these systems is entered manually by humans. For example, in healthcare, doctors create diagnostic rules based on their expertise, and these are programmed into the system. The computer then uses these rules to provide diagnoses.

A key characteristic of expert systems is their reliance on fixed, hardcoded logic written by domain experts. However, these systems face challenges when dealing with unexpected situations or overlapping rules. While inspired by human reasoning and logic, they are limited in flexibility.

Modern examples of rule-based AI systems include credit scoring engines used by banks or systems that assess eligibility for insurance claims. These systems rely on predefined rules to determine outcomes. Computers are essential in these contexts because aggregating and applying all these rules manually would be too complex.

Although effective, these systems have limitations. For instance, an AI chess player could function as an expert system by using the rules of chess. But if the game environment changes in a way that wasn't anticipated or programmed, the system will fail to adapt.

The other branch of artificial intelligence is **machine learning**, or **learning systems**. These systems allow computers to learn from data and improve their performance over time without being explicitly programmed for each specific task.

Instead of coding step-by-step instructions, the machine learns from examples and patterns in the data. This approach allows the system to generalize to new tasks and solve problems for which no clear rules exist. It is loosely inspired by neuroscience and has been validated in practice.

An example of a learning system is a machine that predicts whether a customer will purchase a product. It can analyze browsing behavior, time spent on a product page, and past transactions to estimate the likelihood of purchase. There are no hardcoded rules for this. Instead, the system learns from data.

This process involves feeding the machine many examples of customer behavior. From that data, it learns to combine features—such as browsing patterns, time on site, and transaction history—to predict conversions.

This course will primarily focus on machine learning. In technical literature, you may also encounter the terms **symbolic AI** for expert systems and **statistical learning** or **machine learning** for learning systems.

2. Understanding data

Okay So because we're focusing on AI uh on learning systems uh AI
learning systems or AI machine learning systems we need to understand
data I'm sorry data So let's talk a little bit more about
data So data are examples from which the AI
learns A data set is a collection of examples So here you can see this is a
collection of examples of a customer purchase history So you have the order
ID what the person purchased what category it was on uh the code of that
particular product the name of the seller and the purchase price So this is a data
set It's a
collection of samples or examples of things that
happened and that the AI can learn from This is a numerical uh example of
data but data can be images uh it could be uh text any form of input
that can be processed by computers Okay Um another data another example
would be uh weather data So which includes temperature humidity and wind
Uh and uh it is used to predict the the forecast for tomorrow for example
So we are uh so as you can see here
uh the what the the in the examples or
data points they are called we can call it attribute uh we can call it instances So
instances so the first line the first row is one is one purchase u
example and it's an instance this is one one

data point of that purchase So every row in a data set is a
instance or an example or a a sample
point The the information that comes in the data
set in the columns are called attributes
or features Attributes or features
So the the category for example of the p uh of the product purchased is an
attribute or a feature of this particular data
set So different instances will have different values for each one of these
attributes Another example would be for uh for example the uh weather data
set where you know if it's cloudy if it's
windy the humidity uh or the the the air
pressure and this is for day one day two
day three So each day is an instance and the values of the
characteristics or attributes or features are appear in the columns for
each row So you can see here in this image
some of the attributes are numerical some of the attributes are categorical
or text For example um the the category
in on of the product uh uh purchased it could be text it could be
numbers it can be images uh data needs to be processed by
computers So in the end of the day this data will be transformed
into into something uh uh in in the form of an input that can be processed by a
computer Here is a is an example of an image An image can also be part of a
data set Uh for example this is the instance number one of a data set So
it's an image and an image is is composed of several um uh little squares
You can think about it as a a a a grid of very small um squares And each of
these squares have different intensities of red of green and then and of blue So
for example this first uh this image here is the first row of a data set
where it has several columns The the attribute R1 is the value or the
intensity of red of pixel one intensity of green of pixel one and intensity of
uh blue of pixel one then pixel 2 then pixel 3 and so on So you will have a a
data set with several columns and you could have a 100 images and you would
have 100 rows Okay Now uh data
sets are again the these collection of examples where machine learning will
will learn from Okay But not all data is useful right So you have to have high
quality well-prepared data sets so that your machine learning model can learn
Okay Just uh as an example a ride sharing app might collect data such as
pickup location drop off location type of the day uh time of the day driver
rating the fair uh you uh the fair used But if if your data is inconsistent if
you have several missing key variables um the AI won't learn right how
to for example uh predict the closest or the the best option of fair of the the
closest driver and the best option of of fair for you

Because we're focusing on AI learning systems—or machine learning systems—we need to understand **data**.

Data refers to the examples from which AI learns. A **dataset** is a collection of examples. For instance, this could be a collection of customer purchase history records. You might have fields like order ID, product purchased, category, product code, seller name, and purchase price. This is a dataset: a structured collection of records representing real events that the AI can learn from.

While this is a numerical example, data can also be images, text, or any form of input that computers can process.

Another example is **weather data**, which might include temperature, humidity, and wind speed. This type of data could be used to predict the weather forecast for the next day.

Each row in a dataset represents an individual **instance** or **example**. For example, one row could represent a single purchase event. These are sometimes also referred to as **data points** or **samples**.

The information in each column is called an **attribute** or **feature**. For example, the product category is a feature of the dataset. Different instances will have different values for each feature.

In a weather dataset, each row might represent one day. Columns would include attributes like cloud cover, wind, humidity, and air pressure.

Some attributes are **numerical**, such as purchase price or temperature. Others are **categorical**, such as product category or weather condition (e.g., cloudy, sunny). Attributes can also be **text** or **images**. Regardless of format, all data must eventually be transformed into a numerical form that a computer can process.

For example, an image can also be an instance in a dataset. Images consist of many small squares (pixels), and each pixel contains intensity values for red, green, and blue. The dataset would then include columns for the intensity of each color channel for each pixel. If you have 100 images, you would have 100 rows in the dataset.

In summary, **datasets** are collections of examples from which machine learning models learn. However, not all data is useful. You need high-quality, well-prepared datasets for effective learning.

For example, a ride-sharing app might collect data such as pickup location, drop-off location, time of day, driver rating, and fare amount. But if the data is inconsistent or contains many missing values, the AI model will struggle to learn. In that case, it won't be able to accurately predict the best driver or optimal fare for a user.

2.1 From Raw Data to Actionable insights and Data Driven decision making

Okay So continuing in uh uh talking about data the we are overwhelmed by data and
uh the true value of data lies in what we can learn from it Right So uh the
goal is to take raw data and transform it into something
more useful uh information for example uh or
predictions predictions about what might happen next uh and that can and and that
can be used in the real world So more useful information you could um you
could have um uh some kind of important information
informing your decision making making predictions Predictions about something
that you want uh to uh to forecast for example or predicting something that's
going to happen next based on the information stored
uh uh sitting behind this raw data
So um the this a large amount of
data actually uh makes us uh kind of
find this um important informations this transformations and uh of of raw data
into information predictions uh through a iterative and exploratory
process So we have the data we start explore exploring this data We find uh
important correlations between attributes for example
uh or attributes that depends on each other
and this will actually uh will serve for an actionable insight
you can have an actionable insight and this will um change the course of your
decisions for example in businesses So uh let me give you an
example So you you go to a supermarket checkout Uh you have a loyalty card and
they'll give you um uh uh they'll give you some kind of
coupons and because they have your uh loyalty um they have your name address
and they have access to all sorts of demographic data about you and people
like you right So these coupons are individually
crafted uh for you and you will get
bargains and um they will sell more
So actually they know what you've bought They know your demographic information
your profile and they'll they'll take this data and they'll analyze it They'll
include uh this this data from maybe
thousands of millions of people just like you They'll do experiments to find
um given what you've bought today and your profile
uh what should be the next the next coupon to send you by email So it's kind
of a a mechanism for individual prices and it's it's very beneficial You get
the bargain and the super market sells more
So um this this process of getting the

data exploring the data finding important um uh important information or useful information in the data and then do an actionable having an actionable insight where you can inform decision making is part of what we call it's part of a bigger um area uh of of a bigger
field called data science So data science is uh not only
transforming this raw data into um um into insights and into information but it's also the um the whole process
So the process of storing the data of uh
cleaning the uh of acquiring the data of cleaning the data of storing the data of then analyzing the data uh and and trying to to figure out um trans the uh and trying to figure out um good information and insights from the data
It also encompasses the the the the uh building for example
predictions prediction models of what what can happen next Um so I'll give you an example Imagine a telecom company Uh they they have they are they want to um
they want to try to to um learn
to oh god they they want to try to understand
um when customers cancel their um their plans
and and they want also to recommend given this insight that they are kind of uh wanting and they're kind of looking for recommend retention strategies So the data science process will collect clean store the data It will use what we call data analytics and data mining to find this uh kind of hidden uh information and correlations between all the attributes and features finding the most important ones right So maybe the um the the clients that had um a decrease in their uh in their calls the number of calls they will be
more prone uh to cancel their plans So
you have a bunch of data sets raw data and you with data analytics and data mining you will kind of uh extract important information important correlations important features of the data um that regards the cancellation of plans and then you can build machine learning models AI models to predict uh the the what's the probability of cancellations given the number of um phone calls for example and also you can get this and transform this into a recommendation and uh escalate this to the business management
uh to kind of change policies and strategies for
retention So in all this process uh you can generate reports and data visualization These are very important skills for this um this
exploration uh uh of of the data science
um field inside businesses
Um it's it's so data mining just to uh because you come you will come across these terms uh in uh uh in AI sometimes you hear about
data mining sometimes you hear about data analytics uh and data science um think about u uh data mining and
analytics as analyzing the data the the current
data uh trying to understand why the data is that way um through statistics

through histograms um sometimes through machine learning models
Data mining is also uh the this process of finding important features important correlations important information that um that is related to whatever your aspect you're analyzing Um so in the case of our telecom company
uh the data analytics would be opening this data understanding each column of the data each attribute working on um cleaning this data doing
um histograms and understanding why it's that way getting it prepared to then mine this data in terms of finding uh interesting features that relates to cancellations and then uh use machine learning and then after that you can use machine learning models AI models to to do even more right predictions and and recommendations so all of this cycle is the data science it's kind of
encompasses data data mining and data analytics I also I always think that it's it's very good to make a um it's very good to make a uh uh
parallel with a chef that is working in a kitchen So the analytics and data mining is like analyzing the fridge what you have what ingredients you have and why you have that And uh the data mining is kind of
picking the the ingredients identifying the most promising elements
right So for example one uh tomatoes uh goes very well with
basil And so you would figure out how they might work together and the best ones And the data science is the full culinary journey It covers everything
from choosing where to get the ingredients to cook uh uh and plating
the dish Um it it even u uh
incompasses predictions on how well it suits the dinner's preferences
So data mining and analytics equips you with the raw materials and insights and data science is all this this whole process of bringing together uh the experience of of leveraging data Now these distinctions are are not always rigid right So they're they're used really loosely They they they um overlap right So analytics overlap with with uh with data mining concepts and and they use the same tools in data sciences Uh they data scientists uh often rely on anal analytics um to validate their models Uh they use similar tools and
they have similar skills Um so there are significant a significant overlap but it's important to kind of make uh this uh this uh distinctions because you will

come across this uh terminologies

Continuing our discussion on data: we are surrounded by it, and its true value lies in what we can learn from it.

The goal is to take raw data and transform it into something more useful—such as information or predictions—that can be applied in real-world scenarios. Useful information can help support decision-making, and predictions can forecast what might happen next based on the patterns in the data.

Large volumes of data allow us to uncover valuable insights through an exploratory and iterative process. We explore data, identify correlations between attributes, and detect dependencies. These discoveries can lead to **actionable insights**, which may influence business decisions.

For example, when you check out at a supermarket and use a loyalty card, you receive personalized coupons. These are generated based on your demographic profile and purchase history. The supermarket aggregates this data, along with data from millions of other customers, and runs experiments to determine what coupons to offer next. It's a system of individualized pricing: you receive tailored bargains, and the store increases sales.

This process—of gathering and analyzing data, generating insights, and making decisions—is part of a broader field called **data science**.

Data science involves more than just transforming raw data into insights. It includes the entire workflow: acquiring, cleaning, storing, and analyzing data to uncover meaningful information. It also includes building models that predict future events.

For instance, consider a telecom company that wants to understand why customers cancel their plans and to develop strategies to retain them. The data science process begins with collecting, cleaning, and storing customer data. Then, using **data analytics** and **data mining**, analysts search for hidden patterns and relationships—such as a decrease in the number of calls being a predictor of cancellation.

Once these patterns are identified, machine learning models can be developed to predict the likelihood of customer churn. The insights gained can inform retention strategies and guide management decisions. The output of this process often includes reports and **data visualizations**, which are key tools in the business setting.

You'll often hear the terms **data mining**, **data analytics**, and **data science** used in AI. Here's how to distinguish them:

- **Data analytics** focuses on examining the current data to understand patterns, trends, and distributions—often through statistics, charts, and basic analysis.
- **Data mining** goes deeper by finding meaningful features, correlations, and hidden patterns relevant to specific questions—like identifying which variables predict cancellations.
- **Data science** encompasses both analytics and mining, plus the full pipeline: gathering data, preparing it, analyzing it, building predictive models, and deploying those models.

An analogy: imagine a chef in a kitchen. **Analytics and data mining** involve checking what ingredients are in the fridge and deciding which are the most promising. **Data science** is the full culinary experience: choosing the ingredients, cooking the meal, plating it, and even predicting how much the diners will enjoy it.

While these distinctions exist, in practice they often overlap. Data scientists use analytics to validate models and apply similar tools and techniques as analysts. Still, understanding the terminology is important because you will encounter these terms frequently.

2.1.1 Fuel for AI – Big data

So
uh here is a a slide that I'm not sure if I'm I'm going to uh insert It's about
uh types of analytics I'm not sure if it it's good or not So in in the next uh
kind of interaction of the script I will decide if this will be in the u in the
script or not just have to think a little bit more about it
So so um going going for moving forward
um the the the value of data is learning from it and and it okay learning is fine
but you actually have to act on them So for for for businesses um it's very
important that managers know how to interpret findings and ask the right
questions right so that they can apply the insights and the recommendations and
and understand the predictions of all this data science process to real
business problems So it is very important to managers to to be
um to be digitally savvy w with data literacy and also have the domain
expertise of their uh of their industry or of their um niche
Um so not not every uh pattern in data for example is meaningful or uh
correlations or uh uh kind of uh insights Um so we have
to to uh analyze this uh managers have to understand
uh the the power and the limitations of
data analysis Sometimes data analysis comes with some um of uh with some uh
uh correlations um and relationships between attributes
uh that are totally nonsense right And it might it might be just a sporious
thing So and sometimes data the data scientists that are working on it they
don't have the expertise on the domain to kind of uh trash that out and say oh
this is not good So managers have to understand uh this process of data
science to um to understand the limitations of data
and then um have this scrutiny to to select
what's coming from the data analytics Um often the companies have uh data
scientists uh a a chief data scientist or a data architect and they uh this
professionals they they manage the the data flow they will they will design the
infrastructure and we will talk about
infrastructure in a uh uh just in a in a bit Um and they will also make choices
um make choices of um technologies they're they're going to
use for managing their data So managing their data meaning uh capturing
storing
organizing the data We will talk about this as well And also they will choose
analytical tools um platforms frameworks for example if
they if they're if they're going to use uh third-party softwares with uh with

machine learning and statistical p analysis power or if they're going to
use for example Python libraries uh to help turn this uh complex data
sets into insights and then decision decisions So
um again uh we we want this insights to turn into action to create decisions
inside businesses that uh for sure and this this figure is overloaded uh I'm
not sure what this means matrix for example but but um for sure the data lit
literacy here uh and the commitment with with this digital um uh technology uh
has to has to exist to kind of um to kind of fuel the data uh driven
decisions a little bit more about what what
happened in in the world in this past uh uh years So um what happened is now
we
have what we call big data right So we
uh in the past years data was generated um by having everything going digital
Our smartphones our refrigerator um
people and things that are connected uh uh all over the web and every time that
we connect something um we generate a piece of data Okay
So every time we check out an item at the supermarket every t time we swipe
our credit card every time we send an email uh or even if you uh even uh you
can consider the keystrokes on your keyboard every time we make a phone call
um or or walk through a sec a past a security camera all of this will
generate a little bit of data Okay So and because all of these devices all of
these equipments are connected and are tracking and logging all of this
information Now in the past years in the past years uh thanks to all
of this uh devices wearables social media and internet of things we have a
capac we have a uh amount of data a volume of data that is very big
And this kind of went uh in parallel or together with the computing capacity So
the the the data increased and the capacity to store data has also
increased So that's why we're seeing this boom in machine learning and that's
why we are in the golden era of AI where AI is trending because we can have data
to learn this uh models Okay So
um now before we had some data some companies stored a little bit of data
here and there from their operations but after all this uh digital technologies
um we have actually what we call big data So we have data that has volume
that has um a a rich v variety of data It could
be uh structured data unstructured data multimedia it could be videos it could
be text um and you you we're talking about
terabytes or or pabytes of of data Um also this this data
um because of our comput computing capacity can uh we can exchange or
stream this datas very quickly and um because of the amount of data um there's
lots of inconsistencies missing values ambiguity ambig ambiguities
So we have everything now in a scale that we didn't have before Okay Uh don't
worry uh we're not this is not the focus of the course um of big data per se If
you want to know more about big data and its challenges in business
infrastructure for example

or or its computational challenges you can check out the extra
resources and and for sure this data is uh is is more uh rich in terms of
information um but it's also more challenging to
um to manage Okay So speaking of managing all this data um we continue
talking about data First of all um in in what means to have all this data for
your machine learning models uh for your AI models to work Well it means that
you
need to manage you have to collect the data So you have to collect you have to
clean the data you have to store you have then to extract knowledge
of the data and and and organize the data and
make the data data pertinent for extracting insights of that data Okay So
the initial step to accomplish this in a in a um in a a business is through what
we call database management systems So database management systems We
also call
this only databases It it's it's common to just call this as databases So um
databases are structured collections of data So there the the data set sits
inside this databases where you can find this information and retrieval of
information is possible Companies use this information um uh to manage their
their companies use several tools to manage their databases such as SQL uh or
big data platforms uh like snowflake or da d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d d dab bricks or something like that
Um we the the the core aspect of this course is not on uh data data
infrastructure um or information systems infrastructure for businesses but if you
want to learn more we will leave uh a extra resource for you
So this databases um they they have different characteristics Uh they're
operation databases So for businesses so this operation
databases they're usually stored locally for each operation of the company of the
company and other more comprehensive databases
uh where you have um uh information from
several operations and more historical data and
and and and and aggregation of data from all your processes They are usually uh
stored in big data management systems in the cloud
and in what we call uh data warehouses So here is uh
warehouses warehouses uh for this uh this
databases So you have I will give you an example
Think about Airbnb Every booking review photo location
uh uh every information is part of their data right
So um and they will use this data For example
um if a customer logs in and see a past past bookings a operation database is
accessed right So it's it's the operations of uh recent bookings It's
part of of a faster um with faster access database uh that
usually is stored for that particular operation Now
Airbnb they have a centralized for sure they have a a a data warehouse where
they have a bunch of uh aggregate data on all its operation all historical

informations and this data can be used for the purpose of extracting knowledge from this this history from this past informations
So these the uh for example if um Airbnb
wants to recommend um uh for example uh
listing and adjusts prices It will for sure create a machine
learning model for example that uses historical data not only from past bookings but from uh from uh for example u um characteristics or profile of the user to make this recommendations So it will it will need this kind of aggregate um pieces of information

Here is a slide on types of analytics. I'm still considering whether to include it in the final script, so it may be revised later.

Moving forward—the value of data lies in learning from it, but learning alone isn't enough. Businesses must **act** on that learning. It is essential that managers know how to interpret findings, ask the right questions, apply insights and recommendations, and understand predictions from the data science process to solve real business problems.

Managers need to be **digitally savvy**, with both **data literacy** and **domain expertise** in their industry. Not every pattern or correlation in data is meaningful. Some may be spurious. Data scientists, while skilled in analysis, might not have the necessary domain knowledge to filter out irrelevant or misleading patterns. Managers must understand the power—and limitations—of data analysis to critically evaluate the outputs.

Organizations often employ data scientists, chief data scientists, or data architects. These professionals manage data flow, design infrastructure, and make technological decisions related to **capturing**, **storing**, **organizing**, and **analyzing** data. They also select analytical tools and frameworks, choosing between third-party software with built-in machine learning capabilities or custom solutions using, for example, Python libraries. The goal is to transform complex datasets into actionable insights that inform decision-making.

It's critical to turn insights into action to support data-driven decisions. While some visual tools or diagrams might be complex, the key takeaway is that **data literacy** and commitment to digital transformation are crucial to driving these decisions.

Let's step back and look at how we got here. In recent years, we've entered the age of **big data**. As the world digitized—through smartphones, appliances, and internet-connected devices—huge volumes of data began to be generated. Every online transaction, credit card swipe, email, phone call, or step past a security camera produces a small piece of data.

Thanks to wearable devices, social media, and the Internet of Things, we now have **massive volumes** of data. This explosion in data coincided with increases in computing power, enabling us to store, process, and analyze it efficiently. This convergence explains why we're now in a "golden era" of AI—where machine learning thrives because the required data is finally available.

Previously, companies collected small amounts of operational data. Now, digital technologies have created **big data**, characterized by:

- **Volume**: massive datasets measured in terabytes or petabytes.
- **Variety**: structured, unstructured, text, images, and multimedia.
- **Velocity**: real-time or high-speed data streaming.
- **Veracity**: data often includes inconsistencies and missing values.

Managing data at this scale is complex, and while this course doesn't focus specifically on **big data infrastructure**, you'll find additional resources provided if you'd like to explore this further.

Let's talk about **managing data** for machine learning and AI models. Managing data involves collecting, cleaning, storing, organizing, and transforming it into formats that support insight generation.

The foundational step in this process is using **Database Management Systems (DBMS)**—commonly referred to simply as **databases**. These systems store structured collections of data, allowing efficient retrieval and management. Businesses use a variety of tools to manage databases, such as **SQL** or cloud-based **big data platforms** like **Snowflake** or **Databricks**.

Although this course doesn't focus on data infrastructure, you'll find optional materials to explore this area further.

Databases vary in type:

- **Operational databases** store real-time data from daily business operations. These are often local and optimized for quick access.
- **Data warehouses** store aggregated and historical data, typically in the cloud, for large-scale analytics and machine learning applications.

For example, take **Airbnb**. Every booking, review, photo, and location generates data. If a customer logs in to view recent bookings, the system accesses an **operational database** for real-time information.

At the same time, Airbnb maintains **centralized data warehouses** containing historical and aggregated data from across its platform. This data supports insight generation and predictive modeling.

For instance, Airbnb might use machine learning models to recommend listings or adjust prices. These models would be trained on historical data—such as past bookings and user profiles—requiring access to large, integrated datasets stored in data warehouses

3. What are machine learning models

3.1 Types and tasks of machine learning models

So the ex so let's talk now
about machine learning models Let's dive dive into
uh AI right so the machine learning models that we're talking about that we
that is the um the theme of this course and now that we know um that
machine learning models are used inside this data science cycle this data
science field um where we're going to leverage data to have insights and to
create tools uh that accomplishes humanlike tasks Um we
need to kind of understand it a little bit more what it is and how it works
Machine learning models are computational models um mathematical
models that can be uh that that that sits inside computers that ultimately
performs a task We saw that uh AI perform uh human uh intelligence tasks
for so for example um classifying um something
predicting something reasoning So the the task the underlying task of a model
depends on what you need the model to do But a model will receive inputs and it
will do
this task this prediction classification or reasoning and will
output an answer for you Okay Um and
because we are talking about machine learning models these models will learn
from data sets they will learn how to do this with data sets Okay So again we
will uh the the model will learn to recognize for example a pattern
um and and output and uh a a underlying
task It will uh learn how to predict something it will and then we'll output
this something Okay so in the end of the day and here
are some examples I could present several photos
of dogs and cats and here we could have uh a figure with that dogs and cats here
um I would present a figure with dogs and cats to a machine that would
classify uh if that particular image is a dog or
is a cat So this this human related task this human um intelligence task that
we're performing is what we call classification We could also uh be
wanting to predict the stock market So
you will have inputs as the actual values of the stock market and the

outputs will be the forecast of the stock markets or the inputs could be the the characteristics of the day um pressure uh air pressure humidity uh if it's windy or not and so on And you will learn to predict uh to forecast the weather tomorrow Okay So that's what a machine learning model do It learns to do this mapping from input to output So if we if we understand the output that uh this output depends on the input So y is actually a function of f The machine learning model is learning without any hard coding how to approximate this mapping from the input to the output So it's actually learning how to map the u for example pixels of images and its several features into uh a cat right so the the uh the output is cat or dog for example so there's a mapping that we are learning okay now for a uh we could we could use the another example is an email spam detector Okay So you could feed in uh the uh features of an email uh for example number of URLs and if that person um is in your contact list or not and your model will uh reason if that input is a spam or a not spam Okay So in the end of the day your model is learning this mapping is it's it's kind of um learning how to transform from your outputs to a uh from your inputs to the output that you want for this PAM email example for uh the the learning of this uh model will be made from data from data sets that you already have So let's say that you have several example emails of um so email one 2 3 and and so on and here's the number of URLs if the person is your in your contact or not um in your contact list and other features So what we're going to do is we're going to use this this this information this kind of uh knowledge that is sitting behind all of this sitting behind this data set and and through learning we will uh approximate this functions so that when a new email comes in and you give the number of URLs of this particular new email and if it if that is in your contacts or not and all the other features this machine will reason for you if this email is a spam or not a spam Okay So in machine learning a model will will approximate this function It will identify this relationships between input and output um to make decisions to make predictions to make classifications um on new information and the way it will learn will be on information that in past information that sits in data sets Okay So let's continue studying machine learning We saw that machine learning models are models mathematical representations or computer that sits in computers So computer models that given some inputs will accomplish a task The way the machine learning learns the type of learning under the AI under the machine learning model is actually very important There are four different types of learning and they are very related to the task you want to to accomplish So for example supervised learning is one of the first method meth methods we have actually supervised

unsupervised self-supervised and reinforcement learning Some experts leave self-supervised out of this list and kind of um uh place them uh just just beneath like inside the group of unsupervised learning But I like to highlight that because this is the the learning method that that is um um used in most large language models uh the models that are um that are used in in chat GPT for example So supervised learning is when we train a model we we make it learn using data but the data is labeled So you have input output what we call input output pairs So we are telling the machine what the correct answer answers are This is like teaching a student with with the answer keys And unsupervised learning is the opposite You don't have any label data You're not telling the machine what is the correct answer The model has to figure out patterns or clusters in its own Self-supervised learning is in between both So the model creates its own labels for the data and reinforcement learning is when an AI system learns by interacting with the environment So um the super the the the types of machine learning will they're very related to the task you want to accomplish
So let's dive into um each one of these and uh learn that learn how to identify this tasks For example here classation and regression are tasks that you can accomplish using supervised learning Clustering anomaly detection you can accomplish using unsupervised learning classification regression and natural language processing tasks Um which is which is summarization translation you can accomplish using self-supervised learning And if you want to optimize um you can accomplish optimization using uh models that learn via reinforcement learning So let's start talking about supervised learning Okay So supervised learning

Let’s now talk about **machine learning models**—a central topic in this course.

Now that we understand machine learning's role within the broader data science cycle—where data is leveraged to generate insights and build tools that perform human-like tasks—we can look more closely at what machine learning models are and how they work.

Machine learning models are computational and mathematical models that sit inside computers and perform specific tasks. As we’ve discussed, AI enables computers to carry out tasks that typically require human intelligence—such as classification, prediction, and reasoning. A model receives inputs, performs one of these tasks, and produces an output.

Because we're discussing machine learning, these models **learn** how to perform tasks from data. They learn to recognize patterns and use them to produce outputs such as predictions or classifications.

For example, if we present images of dogs and cats, a machine learning model can learn to classify whether a new image is of a dog or a cat. This is a **classification** task.

Another example is predicting the stock market. The model would take historical data as input and produce a forecast as output. Similarly, weather prediction involves input variables such as air pressure, humidity, and wind conditions to forecast tomorrow's weather.

In each of these cases, the machine learning model learns a **mapping** from inputs to outputs. Mathematically, we can say that the output **y** is a function **f(x)** of the input **x**, where **f** is the model. The model approximates this function using data, without requiring hardcoded instructions.

Consider another example: an **email spam detector**. The model might take features like the number of URLs in the email, whether the sender is in your contact list, and other attributes. It will then classify the email as "spam" or "not spam."

To train this model, we need **labeled data**—examples of past emails with features and their corresponding labels (spam or not spam). The model uses these examples to learn patterns and relationships so that, when it encounters a new email, it can accurately classify it based on learned rules.

In short, a machine learning model:

- Learns relationships between inputs and outputs from data
- Makes predictions, classifications, or decisions
- Uses past examples to generalize to new situations

Now, let's talk about **types of learning** in machine learning. These define how the model learns from data. There are four main types:

1. **Supervised Learning**: The model is trained on labeled data—input-output pairs where the correct answers are known. This is like teaching with an answer key. Tasks such as classification and regression typically use supervised learning.
2. **Unsupervised Learning**: The model is trained on **unlabeled** data. It must find patterns or groupings in the data on its own. Examples include clustering and anomaly detection.
3. **Self-Supervised Learning**: This is a hybrid approach. The model generates its own labels from raw data. This technique is commonly used in large language models, like those behind ChatGPT, and supports tasks like summarization and translation.

4. **Reinforcement Learning**: The model learns by interacting with an environment. It receives feedback in the form of rewards or penalties and uses this feedback to improve its performance over time. This approach is often used for optimization tasks.

Each type of learning is suited to different tasks:

- **Supervised learning**: Classification, regression
- **Unsupervised learning**: Clustering, anomaly detection
- **Self-supervised learning**: Language tasks such as summarization and translation
- **Reinforcement learning**: Optimization and decision-making in dynamic environments

Let's now take a closer look at **supervised learning**.

3.2. Supervised Learning

So supervised learning is a
foundational approach uh where models learn from what we call labeled examples
So it requires labeled data uh what we call input and output data
Okay So uh the the data comes with a label
meaning the true target or outcome is
provided for the machine right as anam
as examples So let's say here you want to train your
your machine learning to predict um whether photographs of uh Alan Touring uh whether photographs are are of Alan Touring or not So the way you do it is you collect several several data several photos of Alan Touring and you also provide a label saying yes this first photo is Alan Turing the second photo is also Alan Touring and the third photo is also Alan Touring And the
task is um for example in this case is the
classification task where giving a new instance a new unseen data a new photo your machine will take that input which is the photo and then it will predict in this in this um case classify if that photo is of Alan Touring or
not So in the end of the day we use several data to train the
same thing as the learning process So you train your data set your sorry your model and this model is deployed so that you can use and test this model in unseen data So let me show you a very uh
curious video of a real world application of a classification task So
let me open this website You can see a video and this is a video of um of a a Tesla auto driving system and you can see that what it's seeing is a truck What is happening is there's a truck on in front of the car uh that holds uh transporting um traffic lights and the the Tesla auto

driving system have learned through several examples and
photographs of of um of traffic lights
to identify that as a traffic light And so this is crazy because what you see in the Tesla monitor is that like there are several traffic lights coming towards you but it's because it learned to classify that particular thing as
traffic light So somebody showed several traffic lights and said look this is a traffic light this is a traffic light and insisted uh showing lots of traffic lights so that your model now when it sees a traffic light will output a traffic light for you So it's it's doing what it's supposed to do um uh in this real world system Another application would be for example recognizing tumors in X-ray So
you you you the input for your model is now X-ray images and the output of your model is uh cancer or not cancer
So and you can also just predict a simple um thing as if it's going to rain or not rain given um this um u inputs for example humidity
pressure and uh uh and and the and the and the characteristics of
the day if it's windy or not for example So let's um let's go a little
bit further in this concepts So when I say it requires training data and output data I'm telling you that the the data set used to train So the data set used for learning has to have columns For example
here we have days where humidity were was 1.2 For example this first row humidi humidity was 1 point uh 1.2 2 and the pressure the air pressure was 1 um sorry one and but it also has to have a
column where you're giving the label you're giving the target or you're giving the answer you're kind of teaching the machine learning you're showing a data where the where the machine learning can know the correct answer so for this day here day one
um given this values of humidity and pressure it rained This is a historical data set So this is past data Day two the humidity was the same
but the pressure changed and there was no rain and so on So you have several data
data points okay in your data set So what are you going to do is you are going to um
to train the the model to predict if
the if the if within this input so your model will receive an input's value of humidity and pressure and it will label rain or not rain
Okay And this is a classification task So the prediction when we're talking about classification task it means that the prediction is
discrete right Like it's it's a class It's really a class It could be one or zero It could be rain no rain It could be um
um if you're trying to predict the the profile of a consumer consumer A B or C So there are distinct classes or discrete classes So this this is what we call a classification a classification uh task
Now the the goal as we said in in earlier uh this this week is that the

model uh it it needs to work in unseen data So if a new day comes in you will deploy the the um your model So you don't have the
answer You don't have right um the answer It's it's it's it's it's not
something that went you don't you didn't store that in your database This is actually happening now So you're you're
testing or deploying your model in this in this new instance and that's where I want it to to to output rain or no rain
So we have part we have data that needs to be used for
training Uh so we're teaching the model and some the unseen data will be uh used
to deploy the model Okay And and we will talk about training and testing uh with a little bit more details um soon
Another example is for example uh if you want to predict the marital status of a person giving the age and the income So the ML algorithm learns to map right So from the features from from the input the or what we call um
attributes um so oops so from the attributes attributes
um will it will learn to predict this target So when a new person come comes
in right so an unseen person comes in that was not used to train my model So for example Louisa my age and my income
and if I am married or not So what I want is my model to exactly output this Okay Now um this this kinds of data sets they
need to have the label and how can we get this data sets
with the label Well you you have to have this in information for example here you have
personal information that you can be retrieving from a bank data set or whatever Um but if you're talking about a model a
machine learning classification model for images
um you will need of for example images of a dogs of of dogs and cats You will have to have a a human labeling that photos
right So um you will have to make think about a spreadsheet where you
have all the photos from Alan Touring Somebody will have to check if that photos are from really from Alan Touring
and will create a column that is uh the kind of the output or the target of Alan Touring or not Alan Touring Alan Touring or not Alan Touring So this is called labeling and it's a process that it's difficult it's expensive Um obviously
for data sets like this one this comes for granted but for other data sets labeling is a very um tedious and
expensive process and it's actually a major bottleneck uh in uh um um in supervised
learning Um so um uh what is this I remember So
um because of that um you can imagine that
uh how how smart Google is for example or or other uh big tech companies where
they ask you to check it to verify that you're a human but checking selecting

images of buses for example So you're labeling for free um this kind uh data for for Google for example Okay So let's think about
um what a model a classification of a
model is Okay So oops
sorry Let me go back So a classification model for the
rain no rain example is this giving if if we were to plot that table
right So this table here let me go back really quick of humidity pressure and the label If we were to plot this points So for example data point one right is 1.2 and one and it's rain So it's let's say that it's it's here So this is data 0.1 It has a value uh 1 1.2 of humidity and a value of um so
it has a value of 1.2 two of humidity one of pressure So it lies here and it rains So in this in this um uh picture I'm depicting rain as
um red points and no rain as blue points
So what what we want is a
classification is a model will draw a line It will so you want a
classification task So in this case this line is is a is my model because this this line here is actually capable of separate this data points Okay So a simple um a simple line in this case is a classification model
So after I learned how to separate after I kind of so the learning process is how to position this this line in this
um in this instance space right or in the input space So this is the space of inputs and the model is positioned here where it learned
uh to separate So um the the goal is to when a new data
point comes in for example this is the new the unseen data
point the the model will be able you will kind of ask the model okay so is this unseen data point rain or no rain so it will get the
coordinates and because it's just a line it will be sitting underneath the line and the and the model will be able to
say okay underneath the line is no rain it's blue it's no rain so this classificator here uh this sorry classifier here oh I'm killing the English language uh is just um what is above the line is rain what is below the line is uh no rain so in the end of the day this classifier will and output and reason and predict your class The the the the training mechanism of h how do we kind of learn this line We will talk and discuss it later but just so that you can visualize this Um this is a process that usually starts random So you start with a random classifier and as it learns from the data that is being made of that is made available to uh the training process um this this classifier kind of gets better and better and better in in predicting in in separating class A from B or rain
and no rain So the this is the learning this is uh representing the learning mechanism and how the learning mechanism is actually shaping the the classification the the classifier and obviously that several
classifiers would work right um you can see classifier number one number two and

number three the three of them would would correctly classify the training instances um and they would um also be fairly
reasonable in classifying a new data point So let's say that the new data point falls here um all three of them
would classify A If it falls here all three of them would classify B obviously that um some of them will if if a data point for example falls here um classifier number one would make a mistake because everything that's uh below the line would be B um but in in
uh in other on conversely if a
point B falls here um sorry a point A falls here uh the the third one would make a mistake So and you would compare this classifiers and choose a better a better
one We will discuss several of these mechanisms next Okay So we saw the classification
task Now now uh just as a reminder that um a
a line in this case for uh two-dimensional problems and when I say dimension I I am talking about the number of attributes So in this case we have attribute H and attribute P So we can put in the in the graph um in a 2D graph a line is a proper uh classifier but there are classifiers that are way more complex right so you could have uh uh kind of arbitrary shapes that that are way more complicated than this line so this example is a very uh basic one let's talk about other predictive class a task So it's a prediction that we're doing So it it falls under the umbrella of supervised learning Um but it's called regression Okay So it also requires labeled data So
the input and the output data And um it it's it actually
um the model will learn how to um to
predict not a class not a discrete um not a discrete
thing or or group but a number So in
this case when we're talking about regression tasks we're talking about predicting numerical values uh um
um sorry uh continuous values So numerical
continuous values it's not discrete It's not different classes Um it it could be whatever number it is here It's not a predetermined class A B or C or one or zero It could have any number any numerical continuous value So let's say that we want to predict a um uh we want to map um the
um for a company We want to um we want to answer a question of how money spent
advertising predicts money earned in sales Right So let's say that we have this data set and uh we we have um
uh the amount of uh money spent in light advertisement and the amount of money
spent in aggressive advertisement Um and we want to predict how much the the the money spent in this advertisements predicts how we we're
going to earn in sales Okay So in this case uh what we want is to kind of fit

that function um where our model will receive an input
and we'll predict the so the um the money earned in sales Money
earned in sales is what we want to predict and the input is the um the
the money spent in advertising
could that could be the sum of both light and aggressive Okay So um what
what we're going to do what our what our uh model is doing is we're going to
learn this function a function uh that
predicts y uh as as a function of of the input
Okay So uh we want to come with um we want to
come with with with f with this function and regression The most basic model for
regression is also a line Um it it's what we call linear regression Um so the
the goal is not separate between types of observations but to predict the number
of a particular uh a numerical number of the of a particular observation So uh
linear regression is the most common um model So given here the the money
spent in advertising if we if we were to to do a
graphic we would see the money in advertising and the money earned in
sales So if we put this points in the graph we have um this this is the input
what we call the input space or the instances space So you're kind of depicting
all of this
um all of the points Okay So in in this case we have one 2 3 4 5 6 7 8 9 I'm
showing only three but we have nine points of this data set
um what what we're going to do for example linear regressor would kind of
would we would come up with this line with this linear equation
um that would would help us predict the
value So when a new data point comes in right where you
know oh I've I've I had 0.5 um um,000
uh spent in light advertisement and.5 so a total of one uh in aggressive
advertisement If you want to predict how much money you will earn in the sales
you can go and put this unseen uh point in uh to to the test with your model So
your model let's say that your your point your new point
um um comes comes here right Uh so an
amount of money of one okay and the
amount of adver the money earned it will be given by the model so oops by the
model by by the line right so by this function f okay so let's say oh it will
be $1.2 2 million Okay Or 1.3 So a linear regression is the most
um simple method of regression tasks But we
have several regression um tasks We have for example polomial
regression So some sometimes you can see that here the line is not capturing it's
capturing the trend but it's not capturing all the points You could for
example uh fit not a line but a polomial
like that Right So so something like I will have to erase this but uh you could you
could um come up with a with a model that is a polomial
uh fitting Okay And we will discuss this what types
of models uh that we have um much more in depth

um in in the next weeks
Okay So now let's uh so again the the how can I come up
with this line How can I come up with this polomial For example let's think
about the line during training doing the learning Uh I will have to find the
parameters of that line So the intercept of the line and the slope of the line that
um um that that is able the best uh
the classifiers that are able to capture this this this numerical trend Okay So
again we will have to kind of during training we're kind of getting better
We're kind of adjusting adjusting the slope um adjusting the line adjusting
the slope adjusting the intercept uh up until a point where we will choose the
red one here because it is the best model Okay
to finalize uh supervised models Um supervised uh supervised
learning
models We're going to talk about forecasting Um forecasting also
um lands um resides on the superver uh supervised learning methods because it
requires labeled data So one column of your data set has to be the
target And um for examples are to predict um
um to predict the uh a value of a stock in a stock market or um a a
uh a trajectory or something that is kind of happening in the future
So the the way to do it is exactly the same in terms of um of uh getting um
offering to the model a data set of inputs attributes that are inputs and
attribute that is the target So for example here we have a curve and we have
the curve up until this point So um um this this curve or this
trajectory will be used for this for example let's say that this is the value of a
stock uh this will be used for a
um for training the model and what we want ultimately is the output after that
So we're going to output the model after that So the the value of the stock after
this point in time So one way to do it is with the data that you have up until
this point you organize a data set in this following way Say that you have
several points um x1 x2
x3 x4 x5 x6
x7 x8 and so on So you will organize the table like this
A first instance of your data set will be x1
x2 and x3 So you're going to take this three points as input and you will predict
the
fourth So the target will be oops this is uh this is x4
Okay As a second instance you are
um you will kind of select you won't use x1 anymore You will use the point x2 x3
and x4 All of these points are his uh you have they are sitting in your data
set to predict x5 So to predict x5 So
um and so on in the next round you won't use x2 you will actually use x3 x4 and
x5 to predict x6 right so that would be the third
instance so what is your uh model learning your model learning is learning
from uh uh inputs x1 um x2 x3 to input

x4 for x x of t x of t + 1 x of t + 2 to
input uh x t + 5 uh sorry oh my god let me put
this way um let me just rephrase this so your model will be learning and let me use a a proper notation you want your your your um
um your model will be learning the next point in time giving the previous point the pre one previous of that and two actually two previous of that So think about the first line the first row here you're predicting x4 based on x1 x2 and x3 So that's what you're doing you're predicting and you're you're going to learn how to do that with all
these points that you have up until this um at this time After that you stop the training and then you deploy your model to kind of forecast other uh points So forecasting is very important and um we actually uh uh call it regressive models because you need points in the past uh to kind of construct your supervised learning data set Um and they have several implica uh
several um uses in um in economics for example
Okay

Let's talk about **supervised learning**, a foundational approach where models learn from labeled examples. This means the data used includes both the input variables and the correct output or **target**. The model is shown the right answers during training.

For example, suppose we want to train a machine learning model to recognize whether photographs are of Alan Turing. We collect a dataset of images, and for each one, we provide a label—"Alan Turing" or "Not Alan Turing." The model learns to classify a new, unseen photo as either Alan Turing or not. This is a **classification** task.

The model is trained on this labeled dataset and is then deployed to make predictions on new, unseen data.

A real-world example is Tesla's self-driving system. In a video, the system misclassifies a truck transporting traffic lights as actual traffic lights. This occurred because the model had learned, from many examples, to recognize traffic lights. When it saw multiple lights on the truck, it classified them accordingly. This illustrates how machine learning models behave based on their training.

Another application is identifying tumors in X-ray images. The input is the image, and the output is a classification: "cancer" or "no cancer." Similarly, we might predict rain based on input variables like humidity, air pressure, and wind conditions.

To train such a model, the dataset must include both inputs and the target label. For instance, we might have a dataset with rows showing daily weather measurements—humidity, pressure—and a label indicating whether it rained. The model learns the relationship between the inputs and the output (rain or no rain).

This is still a classification task. The prediction is **discrete**, meaning the output belongs to a set of distinct categories—rain or no rain, class A or B, 1 or 0.

The goal is for the model to work on new data. When a new day comes, we provide the input data, and the model predicts whether it will rain. This unseen data wasn't part of the training set; the model is now being deployed to make real-time predictions.

Training and testing will be discussed in more detail later, but the principle remains: use labeled data to teach the model, and test it on new data.

Another example is predicting marital status based on age and income. The algorithm learns to map the input features (age and income) to the output (e.g., "married" or "single").

Labeled data can be easy to obtain for structured datasets—like from banks—but more challenging for image classification. In those cases, humans must label each photo. For example, someone must review images of Alan Turing and mark them as "Alan Turing" or "Not Alan Turing." This labeling process is time-consuming, expensive, and often a bottleneck in supervised learning.

Companies like Google cleverly use human verification tasks—like identifying buses in images—to gather labeled data at scale.

Now, let's revisit the rain example. Suppose we plot humidity and pressure on a 2D graph. Each point represents a day, labeled "rain" or "no rain." A simple **classifier** could be a line that separates the data points by class. This line is the model. If a new data point falls above the line, it might be labeled "rain"; if below, "no rain."

This classifier is learned during training. Initially, the model starts with a random line, and as it learns from the training data, it adjusts the line to better separate the classes. Multiple classifiers might work equally well on training data, and their performance can be compared on unseen data.

Note: a straight line works well in simple, two-attribute problems, but more complex datasets might require nonlinear classifiers.

Let's move to another **supervised learning** task: **regression**.

Regression involves predicting a **continuous numerical value**, rather than a discrete class. It still requires labeled data—input and output pairs.

For instance, a company may want to know how advertising spending predicts sales. The inputs might be dollars spent on light and aggressive advertising, and the output is revenue generated. A **regression model** learns this mapping.

The simplest regression model is a **linear regression**, which fits a line to the data. The model predicts sales based on input advertising spend. If a new point is introduced, the model uses the line to predict the outcome.

Sometimes, a linear model isn't sufficient. A **polynomial regression** might better capture the relationship if the data follows a curve. We'll explore different regression models more deeply in future sessions.

To train a regression model, the learning algorithm adjusts the line (or curve) to best fit the data. It tunes parameters such as the slope and intercept to minimize prediction error.

Now let's consider **forecasting**, another task under supervised learning.

Forecasting involves predicting future values based on past data. For instance, we may want to predict a stock's price tomorrow using historical data.

One method is to restructure the data into overlapping sequences. For example:

- First row: inputs = x_1, x_2, x_3 → predict x_4
- Second row: inputs = x_2, x_3, x_4 → predict x_5
- Third row: inputs = x_3, x_4, x_5 → predict x_6

And so on. The model learns how to predict the next point based on a sequence of past values.

This is called a **regressive supervised model**, as it requires historical points to predict future ones. Forecasting is widely used in economics and other fields where time-dependent predictions are critical.

3.3 UnSupervised Learning

So now what we're going to do we're going to talk about
um we're going to talk about unsupervised learning So different differently from the supervised learning
Unsupervised learning uh has data obviously because it learns from data but you don't have the target you don't
have any feedback of the target or the patterns uh that you want So unsupervised learning is very it's related to the to the task of clustering
and with clustering you can do several things for example you uh with anomaly detection So uh what what is clustering So clustering is kind of separating things into sim similar groups Uh and that is what ultimately unsupervised

learning does Anomaly detection is just an application of kind of a clustering where um normal things are grouped into
similar um similar things are grouped in this normal group and something that is an anomaly is kind of separated from that Um but so let's let's take a look at what um what a what a a clustering means Let
me get rid of this because I was kind of practicing So here we have a data set where where we have uh name age and the
income I don't have a target I I I'm not trying to predict from the age the income So it it's one might say oh this is the target I want to kind of do a regression from age to income No that's not what I want to do I just want the a machine learning model an AI to find characteristics of this
data similar characteristics on this data So for example what we can see in if we plot this data where we have age in the x axis and income in the y- axis is that there are there are people that have low uh age they're younger and have low income There are people that are old and
have uh low income And for sure there are people that are old and have higher income Okay So if you if you want to try to sell something this could be informing your business that okay look target the the customers that are old Why Well because younger people for sure don't have high income right But you didn't tell your machine learning to explicitly do this you kind of you kind of inputed the age the income So the the attributes or the or the inputs and the output of your model is just a similarity grouping similarity
uh grouping Okay the way to train this
uh we will see this in deta a little bit more in details but your algorithm starts with like like points that that
uh are equivalent to the center of these groups that it's that that are um that the model is finding And with all this data they
um you're not teaching it but it's kind of discovering So you start with kind of different uh centrids and I'm I I will have to use different colors because of the groups have different colors So it's
a blue it's a green and it's a red or orange I don't know I think it's red
So you start with different kind of um representations for these groups and while the machine learning is assessing the data the training data it will be it will better position the centroidids and when it's all finished training you will have one um centrid here representing the red
group a blue one there representing the um uh blue group and the green will kind of fall here So then you uh you can see
that um you will have you can kind of measure the the distance of the of the points of each row that are depicted as blue dots here to the centrids and decide if it's green if it belongs to the green to the red and to the blue
Don't uh don't be scared if you're um you're not completely understanding what I'm talking about We're going to we're
going to take a look at this uh clustering mechanisms in more
detail

Let's now talk about **unsupervised learning**.

Unlike supervised learning, **unsupervised learning** uses data that does **not** include labeled targets or expected outcomes. The model learns from the structure and patterns in the data itself, without being explicitly told what to predict.

Unsupervised learning is closely associated with the task of **clustering**, and one common application of clustering is **anomaly detection**. In anomaly detection, most data points form coherent groups (or clusters), and anything that doesn't fit well into a group is flagged as an anomaly.

So, what does **clustering** mean?

Clustering involves grouping similar items together based on their attributes. The machine learning model identifies these natural groupings without prior labels. This is the essence of unsupervised learning.

Let's consider a simple example. Suppose you have a dataset containing individuals' **name**, **age**, and **income**. You're not trying to predict income based on age. Instead, you're asking the model to discover patterns or similarities in the data.

If you plot age on the x-axis and income on the y-axis, you may see that:

- Younger individuals tend to have lower income.
- Older individuals may have either low or high income.

From this, you might infer useful business insights—for instance, targeting high-income older individuals as potential customers. But this insight wasn't programmed or labeled in the data; the model found the grouping **on its own**.

During clustering training, the algorithm begins with some assumptions—typically random **centroids** representing the centers of the groups. These centroids are often initialized arbitrarily, and through iterations, the algorithm improves their positions to better represent the actual groupings in the data.

For example, a clustering algorithm might start with three randomly placed centroids: one red, one blue, and one green. As it processes the data, it adjusts these centroids based on the positions of nearby data points.

Eventually, the training process converges, and you end up with well-defined clusters:

- One centroid for the red group
- One for the blue group
- One for the green group

Each data point is then assigned to the group whose centroid it is closest to—based on a distance metric like Euclidean distance.

In the end, this results in groups of similar individuals. The model doesn't know what the groups "mean" in human terms—it simply organizes the data based on similarity.

If this process is still a bit unclear, don't worry—we'll explore clustering mechanisms in greater detail later on

3.4 Reinforcement Learning

So let's talk about reinforcement learning Reinforcement learning is the is um um also uh
um a branch of of machine learning and it we could say that it's it's a newer branch of machine learning that has recently been um um accepted as a branch of machine learning Um it gained momentum in the late 20 uh uh 2010 for example Uh it's uh it's it's it's the basis of the success of deep mind uh
chess engine the alpho uh engine and um the alpha fold which is the the the system that predicts protein proteins um structures um and that gave the Nobel prize to um Danny Savis from from from Google
Um but the way it works it operates on a a
fundamentally it operates on a fundamentally different principle So instead of learning from uh labeled
examples supervised um kind of learning or finding patterns in an unlabeled data such as uh unsupervised learning It learns from what we call interaction and feedback Right So it reinforcement learning is is more like training a pat
and it is very related to the way humans um uh learn by experience Right So think about a baby uh learning how to crawl it in first attempts are are failures Um but at with practice and and and um and exploring the environment uh the baby will will uh kind of figure it out So what we call is that the model learns through trial and error Okay And uh the way we um the learn happens is that it this this model gets rewarded for uh good decisions right and penalized for uh poor decisions So for example the baby if um it tries to crawl uh with only one hand it will it will be kind of penalized because it will the baby will fall and hit his head his her head Um but as soon as the baby kind of
um get an action right of putting one arm after the other it will get a reward because it will it will get to where it wants Okay He or she wants So
um it's it's a particularly uh powerful task and it it's usually um it's usually used for optimization when you're so if you think about again the baby crawling what the baby is doing is optimizing its decision to complete the the the task Okay So in businesses you could think about maybe optimizing business strategies uh when you don't have um uh

uh domain knowledge So lots of domain law knowledge in terms of rules It's a very complex environment um it's um you don't have labeled data and you don't have um you want to kind of know what is the
best strategy So this models can come up with an idea of what is good or a bad um
decision uh outcome um decision and and kind of a um strategy
Usually machine learning courses don't cover reinforcement learning as basic machine learning um like
um theory uh because it's still fairly uh
uh new and and and and niche but we will
talk a little bit more about it because it is part of large language models at some extent So here's the way it works
uh uh in reinforcement learning we learn the learning is uh held from experience as a result uh of of exploration of what we call the agent or the model So your model makes an action in an given environment and it will be uh penalized or or will be rewarded So you can have a a positive reward or a negative reward um and the the results of your actions will change the state So I'll give you an example Uh let's say that this yellow agent this model wants to achieve a green square
here The environment is this um oh I'm not showing I didn't change
the slide Sorry So um here again it's the action environment a agent uh part where I um described here and here's the example So so the yellow agent wants to
achieve the green square and it has to move in this lattice Um we don't we're not we're not
programming it uh we're not telling it the rules We're not hard coding it in any way The yellow the yellow agent will learn
from exploring the environment The environment here is this puzzle is this board is this lattice of squares and um and the actions are the movement So for example let's say the uh oops So let's say that the agent tries to move right because there's a wall there Uh well the the red squares means that um it's it's a wall meaning that it it cannot do it cannot do this move It cannot move towards that direction So
the the agent will bump into this wall and will receive uh or a punishment or a very small reward Right So this is an experience and he will keep that in a data set and will learn from this experience that going to the right in
this particular state is not possible Okay Um so it took an action go to the right in a particular state which is the first uh first position of the board to the to the uh lower left and it got a reward and a new state The new state is I didn't move at all I continue The new state is the same as before Then let's say the uh the agent moved um um the agent moved
here to here right So it it went up and uh the action is going up The in the new state is this new position So now I'm in the this this uh square here and I wasn't rewarded at all So I keep exploring So I I kind of keep that

experience in my data set Now I for example let's say that I go to the right and again this is my new state I wasn't rewarded But say that I try to go up now So when I try to go up I won't get it I won't do it So why Because there's a wall So if I try to go up okay my I will
receive a reward The the I won't change this the state will will kind of be the same and I um um after my action of trying to go up
um but I will be penalized if I do if I try to go down uh
the same thing And so it it it keeps exploring
Okay Uh it keeps exploring up until a point Eventually it it will find a way uh um to get to to the to the green point So it will find a trajectory that takes him to the goal and once he hits the goal he gets a big reward Okay So for example he came here um so he was exploring during learning and keeping all of this data in a data set and he made these kinds of actions and then he landed here Once he landed here he got a,000 points reward Okay So after all this playing this one round of playing and getting to the goal you have several data points of actions and its rewards and then you can use this data set to train So the training mechanisms of reinforcement learning are very very cons uh complex Okay we're not going to to go into details about reinforcement learning but the concept must be very um but the concept it it's good to have the concept clear of what it's it does uh because again it's part of larger language models at some uh for some
extent to some extent and um so what why are we talking about optimization Well
this agent learned through exploring um how to to win the game So it's kind of doing uh if you keep exploring maybe you will find one way another way another way and ultimately you can find better ways than others to
um to uh accomplish your goal It depends on how much exploration the agent does
But what we call this a um um an optimization kind of action right So one example of reinforcement learning is the uh play playing video games So uh Google
has several papers on um agents and models that play it Atari games for example or the game of Doom uh where they don't teach anything They just they just kind of um release the agent with the model randomly exploring the environment getting rewards which are the video game kind of points or lives and after using this all this data from several runs the the the agent will kind of optimize its actions to win the games

Let's now talk about **reinforcement learning**.

Reinforcement learning is a distinct branch of machine learning that has gained traction in recent years. It became especially prominent in the late 2010s through breakthroughs like **DeepMind's AlphaGo** and **AlphaFold**, the latter of which predicted protein

structures with such success it contributed to a Nobel Prize awarded to Demis Hassabis at Google.

Reinforcement learning operates on fundamentally different principles compared to supervised and unsupervised learning. Instead of learning from labeled data or pattern discovery in unlabeled data, reinforcement learning is based on **interaction and feedback**.

It mirrors how humans and animals learn from experience. For example, consider how a baby learns to crawl. Early attempts may fail, but through exploration and trial-and-error, the baby eventually discovers the correct sequence of movements. The baby receives **positive reinforcement** (success) for effective actions and **negative reinforcement** (failure or discomfort) for poor ones.

Similarly, in reinforcement learning, a model—called an **agent**—learns to make decisions through trial-and-error in a given **environment**. The agent receives **rewards** for good decisions and **penalties** for poor ones.

This makes reinforcement learning especially powerful for **optimization** problems, where the model is trying to discover the best strategy in complex or uncertain environments where labels are not available.

Although most introductory machine learning courses don't cover reinforcement learning in depth, it is increasingly important—especially in modern AI systems like **large language models**, which integrate components of it.

How Reinforcement Learning Works

In reinforcement learning, the model (agent) interacts with the environment by taking actions. Each action leads to a **new state**, and the agent receives a **reward** (positive or negative) based on the outcome. Over time, the agent learns to associate actions with better outcomes.

Example:
Suppose a yellow agent must reach a green square on a board. The environment is a grid of squares, and some positions are blocked (walls).

- Initially, the agent doesn't know anything. It tries moving right—but there's a wall, so it receives a small or negative reward and remains in the same position.
- It records this experience: *state* → *action* → *new state* → *reward*.
- Next, it tries moving up. It changes position and adds that experience.
- Over many such interactions, the agent builds a dataset of experiences and begins to **learn** which actions lead to better results.

Eventually, after enough exploration, the agent discovers a **path to the goal** (the green square). Upon reaching the goal, it receives a **large reward** (e.g., +1000). This feedback reinforces the sequence of actions that led to success.

Over multiple rounds of play, the agent collects many such experiences and can begin to **optimize** its strategy—not just finding *a* path, but finding the **best** path to reach the goal more efficiently.

This process is why reinforcement learning is often associated with optimization. The agent learns not only how to succeed but how to succeed *better* over time.

Applications

A well-known application of reinforcement learning is in **video games**. For instance, Google has developed agents that learn to play Atari games or complex 3D environments like **Doom**. These agents are not explicitly programmed with rules. Instead, they interact with the game, receive scores or penalties, and gradually optimize their actions based on accumulated experience.

Although we won't explore reinforcement learning algorithms in depth in this course, it's important to understand the core concept: **learning through interaction and feedback to optimize outcomes**.

3.5 Self Supervised learning

Okay Now finally let's talk about
self-supervised learning Again this is uh as reinforcement learning it's it's
it's a very dense um technical um
um subject but it's important because it's part of the large language model So
self-super supervised learning is a kind of a
um mix between supervised and unsupervised learning So it's a kind of
a mix between both worlds and the the
way it the the tasks associated with this kinds of this kind of learning are
what we call natural language processing tasks So for example text
summarization

translations uh questions and answers sentiment analysis which which is a classification
task So actually self-supervised learning can also do classification and regression and do also NLP tasks Um so
NLP is is tasks that we accomplish uh with language right So summarizing texts translating texts uh answering questions So the GPTs
uh large language models um so the GPTs which have the uh the name of their models or large language models are uh based on self-supervised
learning Other tasks that use self-supervised learning is image and v
video generation So nowadays you can go to any large language model and you can
ask to generate an image or a video based on your words uh or a description that you're giving it It for sure this um
um these models are using some kind of self-supervised learning Okay And so when we say when we talk about self-supervised learning we're
talking about models that have two different stages The
um the first stage of the model is to
uh take unlabeled data Right So uh the the self-supervised
uh learning is very good for when you have immense amounts of of unlabeled data For example text all the text in the worldwide web in the internet they don't have label They're just words right So you you don't have label you could kind of do a supervised kind of
task if you want to classify something but you will have some labeler to um to label them So self-supervised learning has two
different kinds of of of tasks uh uh two
uh two underlying steps The first step is to take this amount this great amount of unsupervised data For example you have images of a cat of cats and dogs but you don't have the label You just have the images You don't have that last column with the targets Um but somehow
the first part is um of the self
supervised learning is kind of generate automatically generate pseudo labels Pseudo labels So you take unsupervised data and
you have a mechanism to uh call it pseudo label So in this first step the mechanism will kind of create pseudo labels for all the cats and a different pseudo label for for the dogs So in the end of the day uh if you think about a data set of several images or uh what you're doing is you're you're you're you're kind of putting cats closer to each other dogs closer to each
other and more distance from the cat And for example helicopter more even more
distance because it's not even an animal So you're kind of having good representations of unlabeled data And then after that you're you're you are um using a supervised learning technique to accomplish a task Okay So if let's say that after doing pseudo labels and un un um so you don't have the labels I have to take this out uh but you kind of separated them in in this representation

the better representations then you can come here and say okay given this pseudo
representations given the pseudo labels um um given this this uh better better representation find an animal So then you will use a
supervised way So you have uh you will have a
target a target column and here you will have the representation or the coordinates of the images Right So let's say they have x and y
values and you say well this one this one this one and this one So I don't know all XY is positive Um they are
animals and animals We have four of them
there And then we have one that has a X that is
negative Uh and let me put five here This is not animal So you will kind of generate a supervised data set and your classifier will be very it
it will make um um lots of difference if you have this representation of the data uh more organized because then it's easier for your uh cla uh classifier to see that if a if the xcoordinate is negative then it's not an animal Okay So
it makes the super uh the supervision task way more efficient and that's what is um underneath the large language models
Okay Um don't think about you don't need to think this about images You if you substitute everything that I told you with words um it's the same thing So imagine that
you're organizing words in a coordinate and you have cat cat dog dog word word dog word cat word helicopter Uh the word dog and word cat will be will be closer to each other and the word helicopter is uh distant apart So if you want to do a prediction of the next possible word or a translation or a summarization
um you can use a this you can kind of come up with a supervised data set that uses this uh this better representation
uh and do a more effective task So in the end of the day we have a uh step one you you will generate pseudo labels in order to get better representations of data better features that represents the data So the features are not uh they're the data is not kind of not organized Now you're organizing and you're going to get this features to kind of come up
with a supervised data set to accomplish some kinds of task
Okay so we we talked about all of this um
supervised unsupervised self-supervised and reinforcement learning method uh types Here is a is a summary for for you

Now finally, let's talk about self-supervised learning. This is, like reinforcement learning, a dense technical subject, but it's important because it's part of large language models.

Self-supervised learning is a kind of mix between supervised and unsupervised learning. The tasks associated with this kind of learning are what we call natural language processing tasks—for example, text summarization, translation, question answering, and

sentiment analysis, which is a classification task. Self-supervised learning can also do classification, regression, and NLP tasks.

NLP refers to tasks accomplished using language—summarizing texts, translating texts, answering questions. GPTs—large language models—are based on self-supervised learning.

Other tasks that use self-supervised learning include image and video generation. Today, you can ask a language model to generate an image or a video based on a description. These models are using some form of self-supervised learning.

In self-supervised learning, there are two stages. The first stage is to take unlabeled data. Self-supervised learning is useful when you have large amounts of unlabeled data—for example, all the text on the internet. This text isn't labeled; it's just words. You could try to use supervised learning, but you would need human labelers.

Self-supervised learning has two steps. First, take a large amount of unsupervised data—like images of cats and dogs without labels—and automatically generate pseudo-labels. You don't have that final column with targets, but the model creates pseudo-labels for cats, and different ones for dogs. In the end, you're organizing the data so that cats are closer to each other, dogs are closer to each other, and something like a helicopter is further away. You're creating better representations of unlabeled data.

Then, in the second step, you use supervised learning to accomplish a task. After organizing the data into better representations, you use the pseudo-labels to define a supervised dataset. For example, given these representations with x and y values, you can define that certain areas represent animals and others don't. You can then train a classifier more effectively.

This makes the supervised task more efficient, and this is what underlies large language models.

You can think of this same idea with words. If you substitute the image examples with words, the process is the same. Words like "cat" and "dog" are close together. A word like "helicopter" is far from them. If you want to predict the next word, or do translation or summarization, you use this representation to do it better.

In the end, the process is:

1. Generate pseudo-labels to get better representations of the data.
2. Use those features to build a supervised dataset and accomplish a specific task.

So we've now covered supervised, unsupervised, self-supervised, and reinforcement learning methods. Here's a summary for you.

4 The importance of data processing

Now as I said here in the um in the last part that we we want to we need to represent our data effectively Okay So
this will uh leads us to the next item that is the importance of having good features in your data Okay So let's say that
um uh you want to have a classifier or uh you want to have
um a regression model whatever um whatever kind of model that that you're working with right so let's say that
um just one second that I lost my comments ments I lost it Where is
it I lost my comments Where is
it I don't know where my comments are Um okay So uh so
the Oh I found it Oops Oh
man Okay So let's say that you're building a a
classifier of uh for example images and the output is saying if it's a car or not a car So it we have to have fe the features the attributes of this of this data set they they they must be relevant For example if one of the attributes is color it will be very hard for the classifi if if you're working only with
color and and uh and height of the of the of
the of the object in the image This will be very hard because like cars and motorcycles have kind of the same height and they all are red or or or or gray or black or green So you you need to have some features that are very important Okay So we need to extract features and this is crucial for the the the the success of the model Okay So the the model needs to kind of capture the important aspects of the
data that is related to the prediction it's doing right So again color and height will won't do it So the the uh there's a field called feature a feature extraction or feature selection that experts use machine learning algorithms as well to kind of accomplish it they can use also statistical algorithms to kind of see to find what what attributes are better So for example one
one attribute would be the width right So because motorcycles are kind of thin and cars h they are they're they they have a larger width that would be
a a um a a feature that is more relevant to the classification uh model Okay Um so during data mining you you kind of have to discover which ones are most important kind of what features are most important and sometimes having lots of
data will make you will make you see uh things that um things that are weren't clear

before So you you never thought about that but using algorithms you can check which column is more relevant to the task that you're
accomplishing Okay And uh just to uh uh wrap this part of feature extraction um I I want to I want to talk about a uh what we call feature
engineering Sometimes uh it's also called it's it's also part of feature extraction but um but you can also um hear this term as
feature engineering So what what is feature engineering So sometimes um in the in this process of discovery and in the process of extracting
features that are relevant to the task it might be that you need to create new features So new attributes or combine existing
raw attributes uh in order to to have something that is um
meaningful for your uh for your machine learning task So here's an example uh maybe you are um tracking uh purchases and there's a
column uh in the data uh the is the date of the purchase Now if you're if you want to kind of build a prediction of whether the sales are um
increasing or decreasing or even if you're not predicting you don't have like a a model of prediction but maybe
you just want to explain some events Maybe it's splitting this feature into two new ones where one is the day of the week and the other one is a uh is is a is a feature uh telling you if it's a holiday or not
So you you you created this new features and these new
features for sure would be important for
um for the success of the model And again this feature extraction and feature engineering are aspects that are so important and they can be a difference between a um a model that is
successful and a model that um uh that
fails or uh has not does not perform
um uh well Usually um there's a step um um in uh
like before you could u use this data to train your machine learning models to accomplish tasks uh that we call data prep-processing So
um the data prep-processing step is the the analytics data mining and feature engineering steps So is when you understand what you have you deal with mi missing values Um you deal with kind of uh applying filters to the data removing outliers Um then finding which filters and
important correlations um inside the the the data um there are
inside inside the data You come up with good features if you need to you create them and um so this this is called pre-processing
and feature engineering like creating new data is is is a very
um um it's very common when the data is not structured So here's a concept that is important Uh a structured data is the one that that each column fits a purpose So for example you have name age income and the uh marital status So each figure
of uh of of a row uh each figure will um
um different figures will kind of appear in this in the columns of each

row um and and they are the same for each instances Now there there is data that is not structured They're not organized that way So for example a Twitter um um post
um or a a book that's data as well You can use that right for example to uh for uh accomplishing natural language processing But the data is not a structure uh uh it's just a bunch of
words um that you cannot separate in in columns right So when you have this kinds of data you will you have to kind of engineer and try to organize or structure the data as much as possible This is a time consuming and challenging process So in businesses this is something that takes time and often
requires uh data analysts data scientists to work this pre-processing
uh task right Uh so they spend a lot of time and is for sure an art Um uh it you have to have you have to have like experienced ex uh experts that know what they're doing and kind of the ways they
can go through to um engineer good data and do a grow a great pre-processing step inside the pre-processing step And um we have feature engineering that we just talked about Sometimes you have to scale the features Uh sometimes you have
to do dimensionality reduction data cleaning data augmentation Some of them are kind of
straightforward Feature engineering we discussed Data cleaning is when you have you want to get rid of outliers or kind
of um um some some data that has uh you know like unreadable characters something that is kind of um it needs a cleaning data augmentation you you deal with missing values or the generation of new um instances Uh sometimes you have
to generate synthetic data uh to because you don't have enough data
and we will talk about synthetic data um uh in in the in the in the next
weeks Uh sometimes we will you have to do feature scaling and dimensionality reduction feature scale and dimensionality reduction We will also see uh in the next week when we're
talking about uh evaluation and how we
uh what are the aspects that kind of affect the goodness of an algorithm

Now let’s talk about the importance of effectively representing data. This leads us to the importance of having **good features** in your data.

Suppose you're building a classifier or a regression model. The features—the attributes in your dataset—must be relevant. For example, if you’re building an image classifier that detects whether an image contains a car, and you only use features like **color** and **height**, the model may struggle. Many objects like cars and motorcycles can be the same height or color. So those features alone won’t help the model distinguish them.

You need features that are truly important. This is where **feature extraction** comes in. The model must capture the aspects of the data that are truly related to the prediction task. In the car versus motorcycle example, **width** might be a better feature, because motorcycles are typically narrower than cars.

There's an entire field of study dedicated to this, and machine learning practitioners use statistical or algorithmic methods to find the most relevant features. During **data mining**, you can discover which features matter most. Having more data can also help reveal relationships that weren't previously obvious. Algorithms can evaluate which columns are most relevant to the task you're trying to accomplish.

In addition to feature extraction, we also talk about **feature engineering**. This involves **creating new features** or **combining existing ones** to make your model more effective.

For example, if you're tracking purchases and one of the columns is a **date**, you might want to split that into two new features: the **day of the week** and whether it's a **holiday**. These new features could be much more informative for your machine learning task.

Feature extraction and engineering can make the difference between a model that performs well and one that fails.

This process is part of a broader step called **data preprocessing**. Before training a model, you need to prepare your data properly. Data preprocessing includes:

- Data analytics
- Data mining
- Feature engineering
- Handling missing values
- Removing outliers
- Filtering and transforming data
- Identifying important correlations

If needed, you create new features as part of this process.

Feature engineering is especially important when working with **unstructured data**. Structured data means each column has a consistent meaning—like name, age, income, marital status. Each instance fits into that format.

But unstructured data, like a **Twitter post** or a **book**, doesn't fit into neat columns. It's still usable—for instance, in NLP tasks—but you first need to **engineer structure** from it. That might involve tokenizing text, detecting sentence structure, or other techniques. This step is **time-consuming and challenging**.

In businesses, a lot of effort goes into this preprocessing phase. Data analysts and scientists often spend significant time preparing data before model training. It's an iterative, detail-oriented process that relies heavily on expertise.

Within preprocessing, you'll often perform:

- **Feature engineering** (as discussed)
- **Feature scaling**
- **Dimensionality reduction**
- **Data cleaning**
- **Data augmentation**

Feature scaling ensures numerical features are on similar ranges.
Dimensionality reduction simplifies high-dimensional datasets while retaining essential information.
Data cleaning removes outliers, fixes formatting issues, and corrects inconsistencies.
Data augmentation might involve generating synthetic data to supplement sparse datasets.

We'll explore **synthetic data** in upcoming sessions. Feature scaling and dimensionality reduction will also be covered next week when we discuss model evaluation and factors that influence algorithm performance.

5. Evaluation

So
that's where uh I'm going to wrap up Uh so
after when you after kind of um doing the data science cycle uh
storing data cleaning data do pre pre-processing um then you kind of feed this data to
your model then then the the model is trained and then you will deploy the model Uh we have to remember that
several models can tackle diff uh the same task So for example classification tasks can be accomplished by a random
forest model or neural networks decision trees So don't worry we'll all see it what that what kinds of models uh uh
what the these are Um but you have to select a model that is kind of good uh that is performing well in your task So data science is very much about experimentation you sometimes come up with several different models and you will choose the best one amongst them So
in the end of the day we will have to compare models using some kind of metric uh a
metric that corresponds to the goodness of a model and we will select the performance uh we will select the best model giving the performance of that of

the models uh um against this uh metric and using a common test data set
So as I said uh said we're going to understand uh in more details the
metrics of each uh uh uh kind of the the
metrics used to evaluate models depending on their on their task the
task they're performing but I will
uh explain to you the training and testing data division so that we can uh
understand a little bit more uh the evaluation
process So the the training sometimes also so the learning sometimes known as
training and also sometimes known as fitting is that process of adjusting a
model's parameter right to find a best model So if you if you think about the
the the linear regression is trying to find uh like the line the slope of the
line and where it intercepts u until you have what we call a best fit
So this training data is the the the data that is sitting on your data set
past data Uh it can be labeled or unlabeled depending on your uh task the
task that you're accomplishing But the model is learning
from this training data Okay Now the uh
the test data this this similar test data is a test it's a collection of data
that you you separate you we call it you a hold out you
separate so that you could evaluate the goodness of your uh of of of your model
and even compare it with other with other models So um let
me let me uh just go back to the data set where we wanted to predict the um if
it's going to rain or not Um so rain or no rain for several uh
instances Okay Um
and we had humidity values and pressure values This was the
target Um so we had several values here and we could just go back to that um
figure Um and let's say so the way we do it is
we we hold out a a a quantity and an amount of data that we know a labeled
data that exists in our data set set We hold out and we call that a test data
Usually is a smaller portion of our our data set some uh usually 10% of the data
5% of the data of the data points of the instances But what you're going to do is
you're going to train your model here in in the in the uh kind of instances of
day one and day two And on day three and and day three and day four you're not
going to use for training you're going to hold out and you're going to
deploy your your trained model here as a test So for example if you're doing this
classification one one uh metric and we will see metrics um uh in details could
be the accuracy So the number of um correct predictions that you had over
the total number of predictions So what you're going to do is you're going to
test you're going to get this data
um marine and rain So you're going to you're going to get this hold out data
this hold out data in which you know what the the model should predict and
then you're going to calculate the goodness metric So the goodness metric is
calculated exactly in this
test uh data set and then you can uh deploy you can you can kind of test this
um uh this data you can you can test or evaluate

uh several models So you can you can uh for example train different neural networks uh you can train an SVM or a decision
tree So you can compare the performance of this four um different models using a
metric in this test data set because you know what
it should be outputting and so you can calculate the
goodness or how good your um algorithm is your model is But don't worry we will see this in details um in the next week So um we will uh uh
pick up exactly in in this uh evaluation part of the course
So in the next weeks we will going we're we we will uh pick up here in evaluation and we will open aspects of the training and testing so that we can understand better the evaluation of models we will okay so I will leave uh

Let's wrap up by discussing what happens after completing the data science cycle—storing, cleaning, preprocessing the data—and feeding it into the model. Once the model is trained, the next step is deployment.

It's important to remember that multiple models can solve the same task. For example, classification can be handled by random forests, neural networks, or decision trees. You'll need to select a model that performs well for your specific task.

Data science involves a lot of experimentation. You may train several different models and choose the best one based on how they perform. To do this, we use evaluation metrics to compare models and select the best one using a common **test dataset**.

We'll cover evaluation metrics in detail for each type of task, but let's start with the basic idea of training and testing data.

The learning process—also known as **training** or **fitting**—is the adjustment of a model's parameters to best fit the data. For example, in linear regression, you adjust the slope and intercept until the best-fitting line is found. This is done using the **training data**, which is typically past data. It may be labeled or unlabeled depending on your task.

The **test data** is a separate portion of the dataset, held out specifically to evaluate the performance of the model. This data is not used during training. Holding out data helps us assess how well the model generalizes to new, unseen data.

For example, consider a dataset used to predict whether it will rain or not, based on variables like **humidity** and **pressure**. You might have several days' worth of data. You can hold out some of it—say 10%—as test data. You train your model on data from Day 1 and Day 2, and then evaluate it using data from Day 3 and Day 4.

Let's say you're using classification. A common evaluation metric is **accuracy**, which is the number of correct predictions divided by the total number of predictions. By

comparing the model's predictions on the test data to the actual outcomes, you can calculate this metric.

This process allows you to evaluate and compare multiple models. You might train different neural networks, a support vector machine (SVM), and a decision tree, then compare their performance on the same test dataset using a consistent metric.

This testing process gives you insight into how good each model is at generalizing. We'll explore these metrics and evaluation techniques in more detail in the upcoming weeks.

In the next session, we'll continue with **model evaluation** and dive deeper into the aspects of training, testing, and comparing models.

6. An overview of machine learning building blocks and workflow

to finish I'll leave you here with a with a little bit of a of a a building block of AI So first you collect the
data we pre-process this data uh we we engineer features or extract features Then we train and we evaluate models We
uh we will so you you you can train uh tune we will differentiate training and tuning um as well But understand training and tuning like as kind of
fitting the model finding the the the best parameters of your model as we saw with uh the intercept or the slope of a cur of a of a line And then you will evaluate this models and you will choose a model to deploy right So you will deploy you will do your classification task or your forecasting tax or task or your summarization using chat GPT for example or other LLM that kind of won the evaluation and then you just keep monitoring uh and you can always kind of perform this as a cycle you can collect more data and you kind of can
uh do this uh in in a kind kind of feedback loop where you all you're
always getting your your your models better All right thank you so much and uh this is week one

Um just uh to kind of wrap up here as a
comment uh I'm not sure if I should already talk about what metrics uh like give more
details um regarding the metrics I have this slide here where like oh if you're doing classification and I would define accuracy here um and if you're doing regression the
mean squared error or if I I mean it it is more specific uh to each model So I

was thinking about talking about this in the in the next
weeks So this is something Harshida for for Harshida now something that I would
love to have your feedback uh of what you think it would be good One thing that I noticed after the like I finished
and I have some so many comments to kind of uh um uh change here
but when I'm when I'm starting to explain the the tasks I am definitely
going to not use I'm going to use training and not test yet because I I
kind of I I never um defined this I would just I would just say that we're
deploying in unseen data right so I will go this way like
deploying in unseen data for all the types of uh of machine learnings um
uh everywhere I kind of um
show data sets right I will do that in that way
And after that when I come so so so then
when I'm talking about evaluation I can define what is the test
the training and the test and you know like the percentage of the data split
but again the way I did it uh so I have to to just be careful with the script
not to say test before this slide side um because then I'm going to explain the
the test set the the division and next the week two I will start in evaluation
and I will talk about the t the divisions on uh training validation and
test So so I'm I'm going to go in more depth and also explain the metrics but
the way I did right now I didn't explain how the training like the loss
function like how to adjust this uh models
um and what kind of metric we use
to kind of measure the goodness So Harshida if you think I should uh change
something here let me know Uh it's it's kind of different ways that we
could uh convey the message Let me know if it's understandable or if you think
we could we had to change that uh the video if you kind of um um it's
it's very long but I'm talking very slowly and kind of uh uh
brainstorming and also I think like if we uh with a script where you you're
reading you're not gagging or um it's it's more clean you have the images to
show like a professional thing I would say that the like the the video would
reduce a thought So it might be uh that you will give me a feedback like no tell
us a little bit more about the metrics at this point because it's not understandable
Uh include something and
remove something So that's the kinds of uh feedback that I want from you

To finish, here is a summary of the **AI building blocks**:

1. **Collect the data**
2. **Preprocess the data**
3. **Engineer or extract features**

4. **Train and evaluate models**

During training, you adjust or fit the model by finding the best parameters, such as the slope or intercept in a linear model. After evaluation, you **select the best-performing model** and **deploy it** for tasks like classification, forecasting, or summarization—such as using ChatGPT or other large language models.

Once deployed, you **monitor performance** and repeat the cycle. You may continue collecting new data and re-training to improve the model over time. This creates a **feedback loop** that helps continuously refine your models.