



Lecture

LEGAL TECH

4 AI

Chiara Gallese, PhD

**MSCA Fellow, Department of Law
University of Turin**



About me

I am a law scholar working at the intersection between law, ethics, and computer sciences, with a focus on healthcare applications. I started my journey in academia as a PhD student in Japanese private international law. Then, I worked in Milan as a litigation lawyer in the finance and banking sector and as a privacy consultant. I then moved to the Netherlands and worked in the Research Data Management team at the Eindhoven University of Technology and as a member of the regulatory board at the Eindhoven MedTech Innovation Center. I came back to Italy in 2021 to work as computer science scholar in the UNI4JUSTICE project in Trieste, and I joined Unito in 2023 as a MSCA Fellow.



My Project

I am the PI of the DataCom project, part of Marie Skłodowska Curie actions.



A new EU Framework for an Ethical Re-use of Health Data

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by
the European Union**

Definition of AI

AI is a buzzword that is currently used to market different models in research and industry. In the media, it is used in exaggerated terms to boost clicks and visualizations.

The truth is that a real AI - the one that can mimic human intelligence - doesn't exist (yet), and it is not clear if it will ever will.

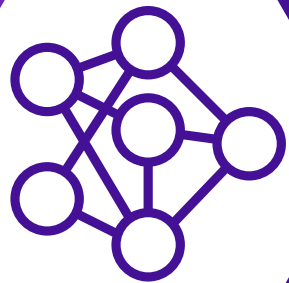
Importance Of Clear Definitions

It is important to provide a clear definition in order to understand each other and the implications of our words when we discuss the possibilities, features and risks of AI systems



So, what do we mean by “AI”?

Algorithms that perform specific tasks, that a human would be able to do in a very long time if done on paper and without a calculator: it is NOT magic.

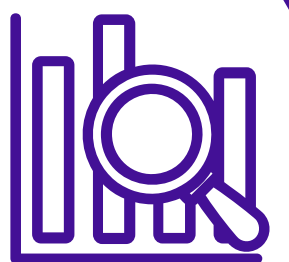
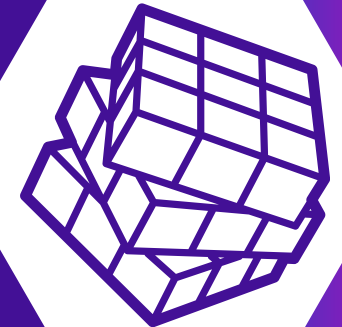


Machine learning

Supervised or unsupervised models

Logic approaches

Techniques based on logic, such as fuzzy logic

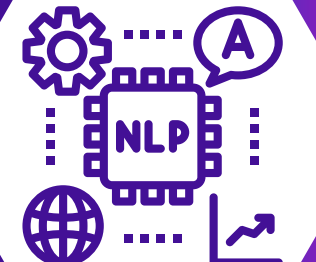


Statistical models

Models based on traditional techniques in the field of statistics

Other methods

Various techniques such as symbolic approaches for NLP

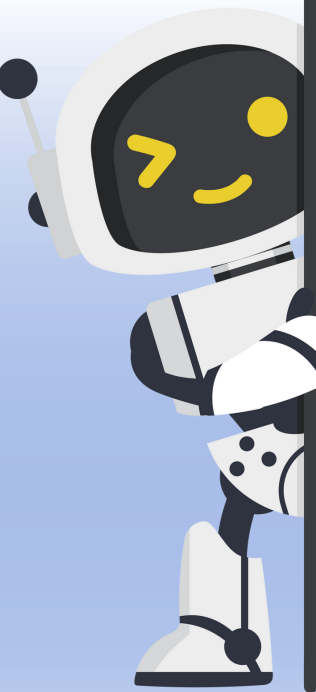


Does it make sense?

What is the difference between different models from a legal point of view?

Why is it different if Microsoft Excel calculates a function, or if another algorithm is used?

Autonomy is often cited to justify a different legal discipline.



The AI Act

According to the AI Act, AI systems used in the administration of Justice are considered **High-Risk**.

This means that they must comply with many new obligations and requirements, including data governance, that will ensure the responsible creation of the model.



Current use of AI in the legal field



Document writing

Drafting of judgments, documents, contracts, letters, proofreading



Document analysis

Summarization of judgments, topic modeling, information retrieval



Predictive systems

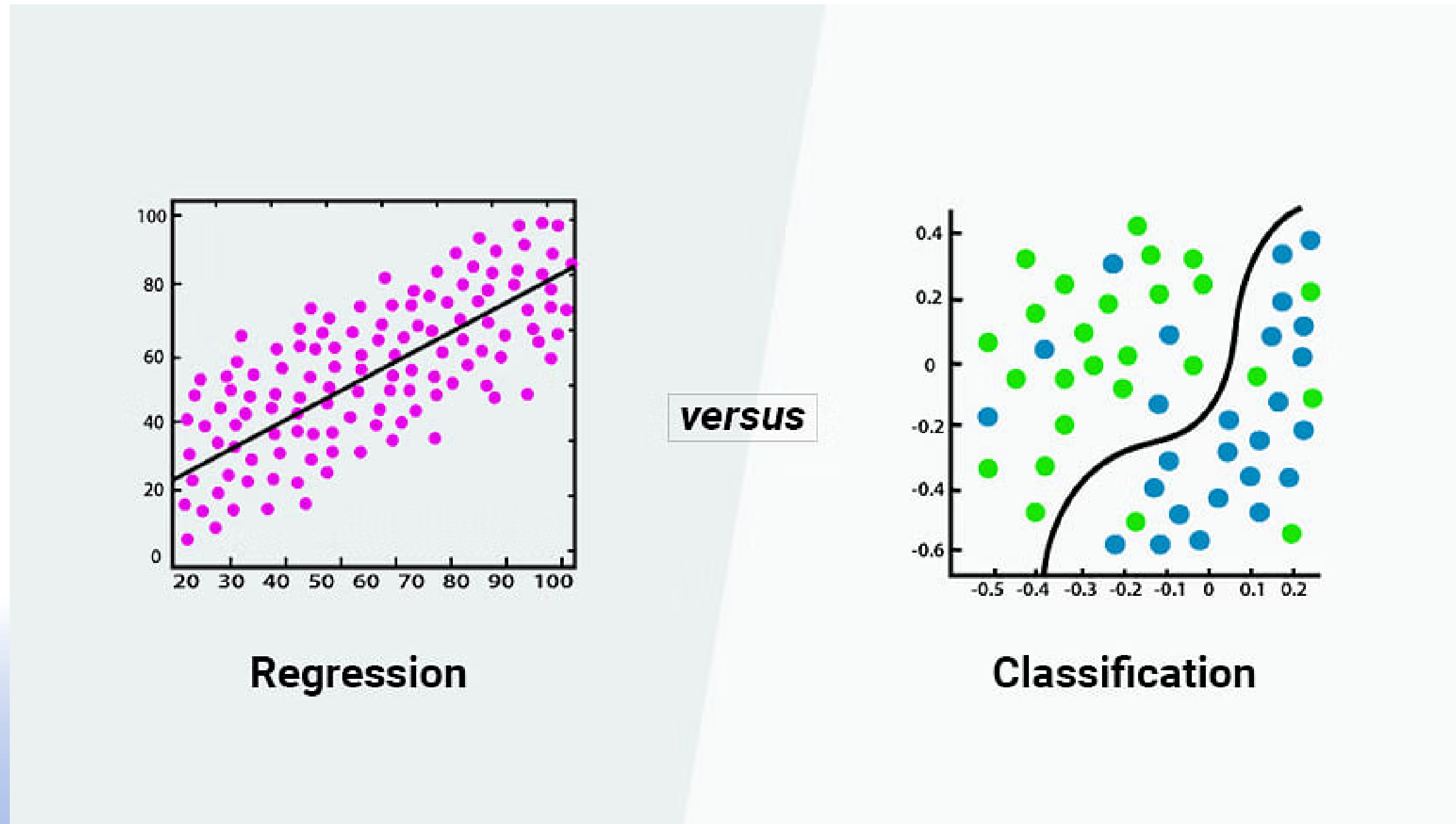
Risk assessment for lawyers and judges; predictive policing, social scoring, prediction of case outcomes



Automated tasks

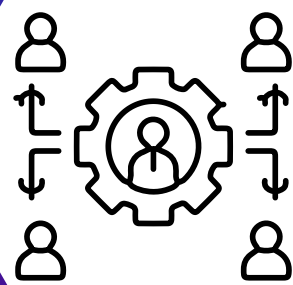
Automating administrative tasks such as planning, meetings, customer care

But how does AI work?



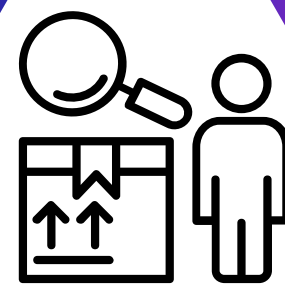
Let's start with a simple example of two different set of problems that we want to solve

Supervision level



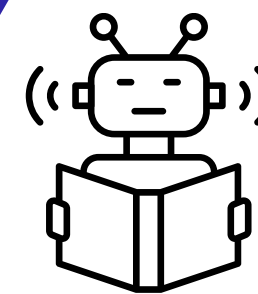
Supervised

In supervised learning, a computer is given lots of examples for which the answer (label) is known. Each example helps the computer to "learn" the patterns or features that are important for making a decision. Once the learning phase is over, the computer can use what it has learned to make predictions or decisions about new data it has never seen before.



Semi-supervised

In semi-supervised learning, the algorithm learns from a dataset that includes both labeled and unlabeled data, usually mostly unlabeled. This approach is used when acquiring a fully labeled dataset is expensive or infeasible, and it leverages a large amount of unlabeled data better to understand the structure of the underlying data distribution or to make better predictions.

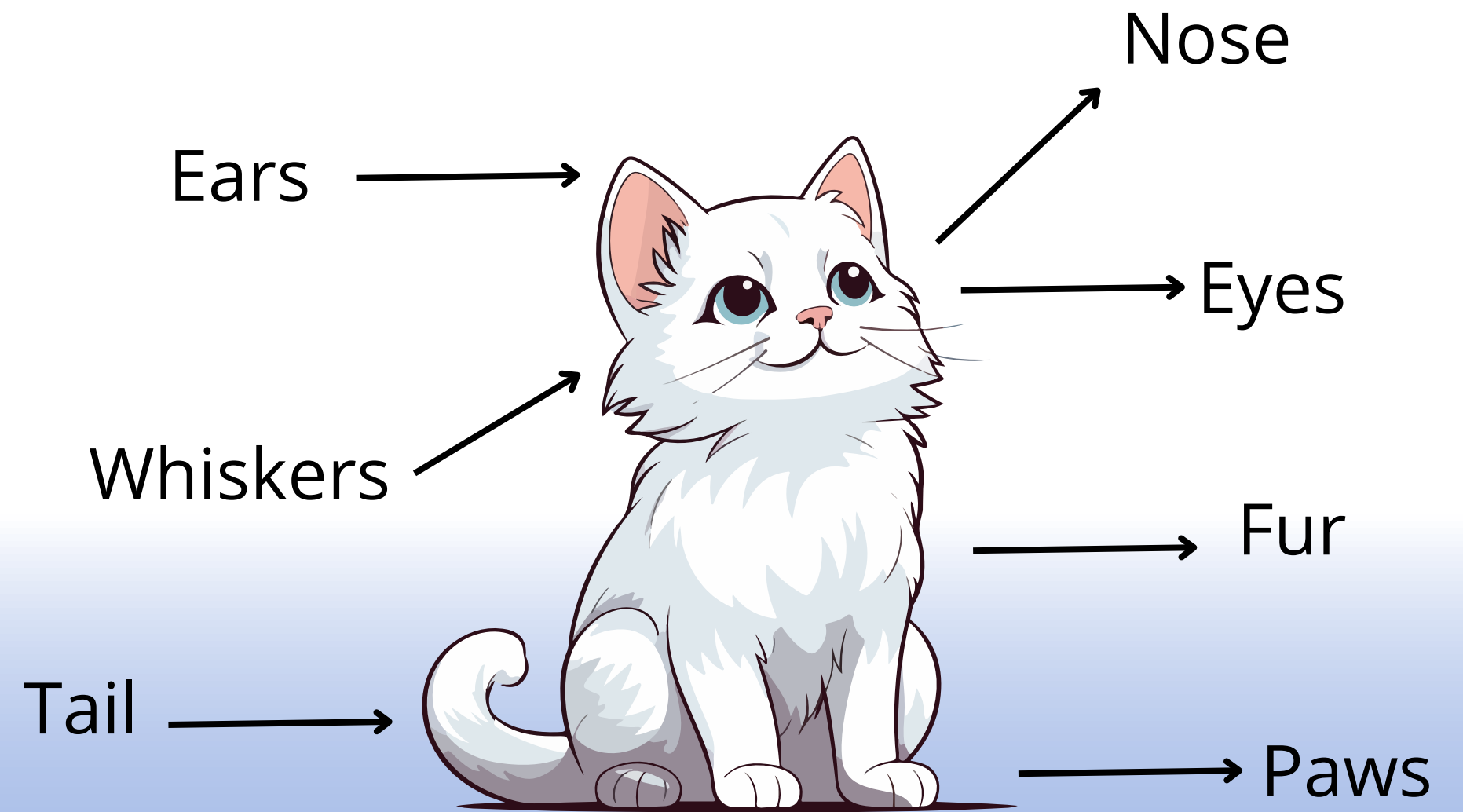


Unsupervised

Unsupervised learning is a type of algorithm used to draw inferences from datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The algorithms themselves infer the natural structure present within a set of data points.

Humans label items often so that we can learn to recognize them!

How does the kid put the right label on a new cat? She recognizes its feature:



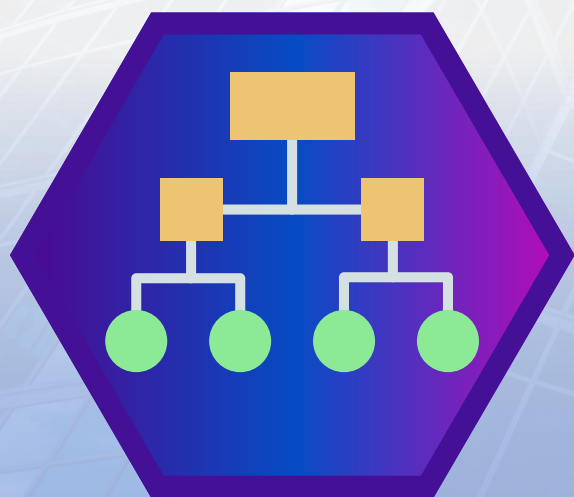
Cat!

Regression VS Classification



Regression

Regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices.



Classification

Classification finds functions that help divide the dataset into classes based on various parameters. When using a Classification algorithm, a computer program learns from the training dataset and categorizes the data into various categories. This includes the so-called “logistic regression.”

Regression

Let's start with a very simple example.

We want to **predict** how much a lawsuit will cost to the client, and we have the following **data** from previous lawsuits:

- the total fee for each case, and
- the total number of hours spent studying the case documents



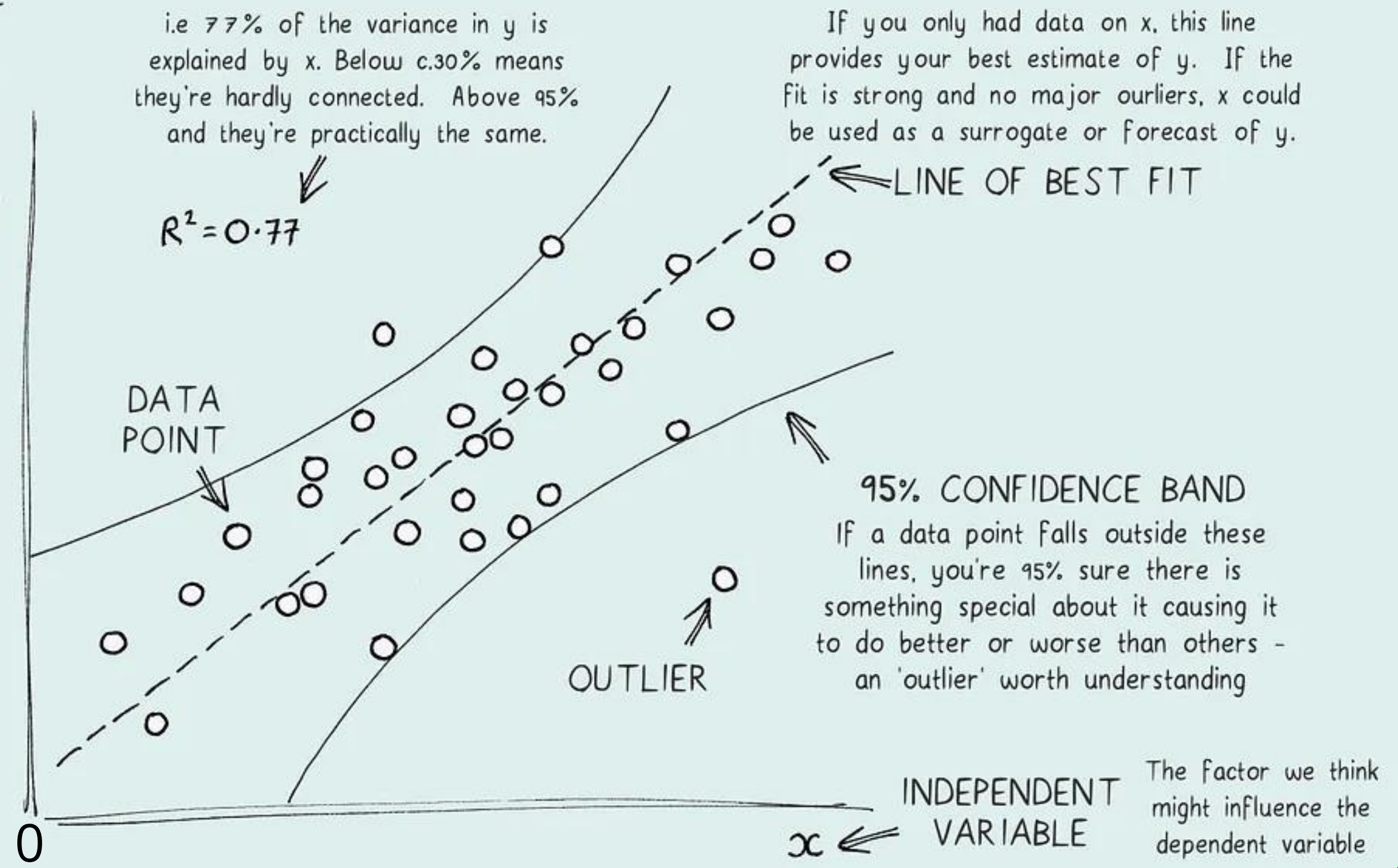
LINEAR REGRESSION

The thing we want to explain
DEPENDENT VARIABLE

Cost of lawsuit

y

The more x increases,
the more y increases



Hours per week spent analyzing documents

Classification

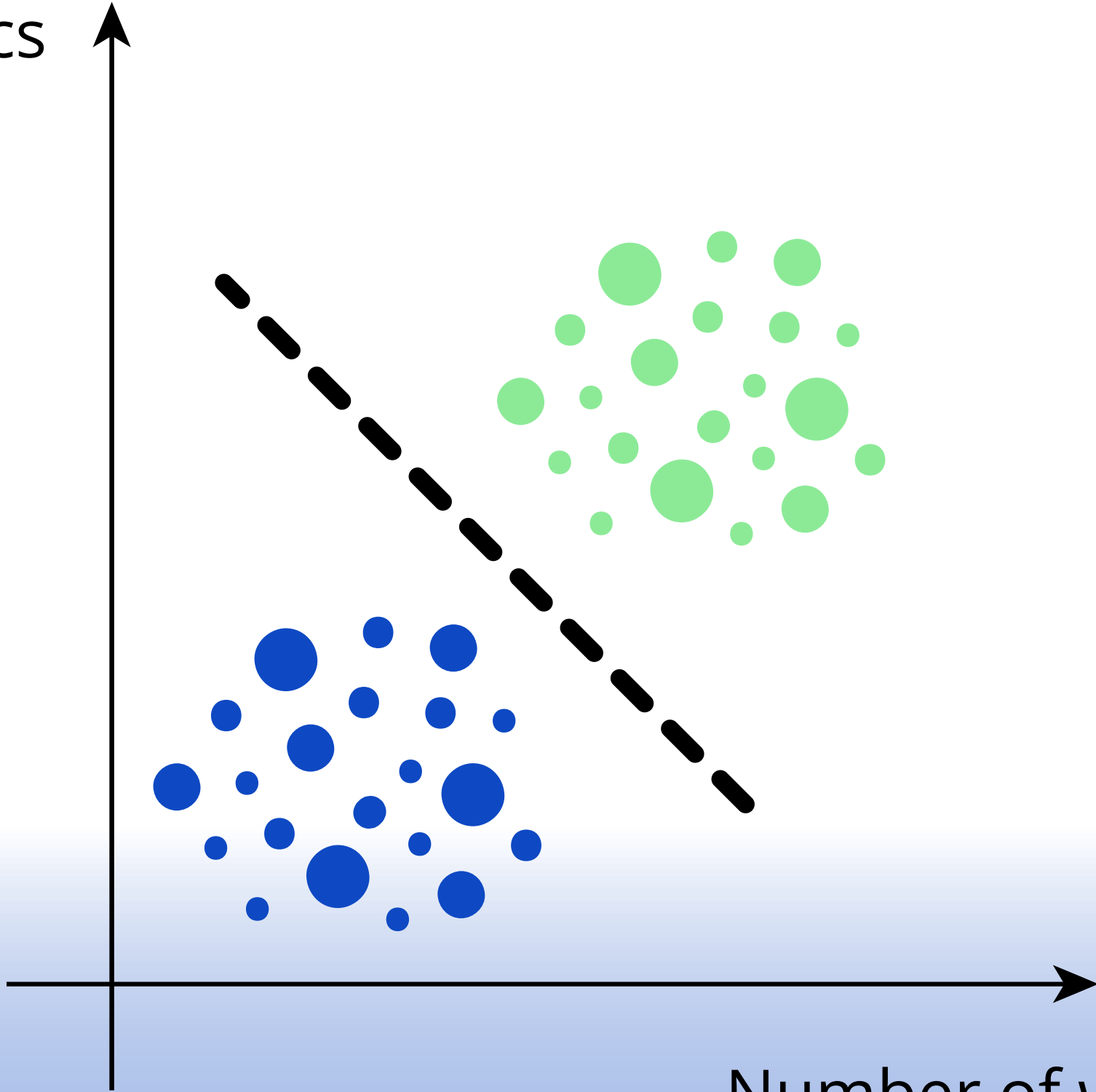
We want to predict if a client will be a loser or a winner.

We have the data from previous cases:

- number of supporting documents, and
- number of witnesses



Number of docs








- Winners
- Losers

Number of witnesses

Generative AI

Generative AI, like ChatGPT, uses a combination of different techniques, such as neural networks and statistical models. Based on their training data, they predict what word/sentence comes usually after a certain word/sentence:

| Input | Previsione | Probabilità | No probability = random |
|------------|--|-------------|-------------------------|
| Ciao, come |  | stai | 90% |
| |  | va | 70% |
| |  | butta | 60% |
| |  | sai | 40% |
| |  | sedia | 5% |



The importance of the training dataset

An AI system is only as good as its training data. **Garbage in = garbage out**



How data is collected

There are many concerns regarding the source of data.

Is the data copyrighted?

Is the data obtained legally?

Is there consent from the author?



How data is preprocessed

Before the data is used to train AI, it is pre-processed, that is, modified to make it suitable for the training-
For example, labeling, cleaning, augmenting, downsizing, correcting, structuring, and changing the format.

This can introduce new biases.



What is in the data

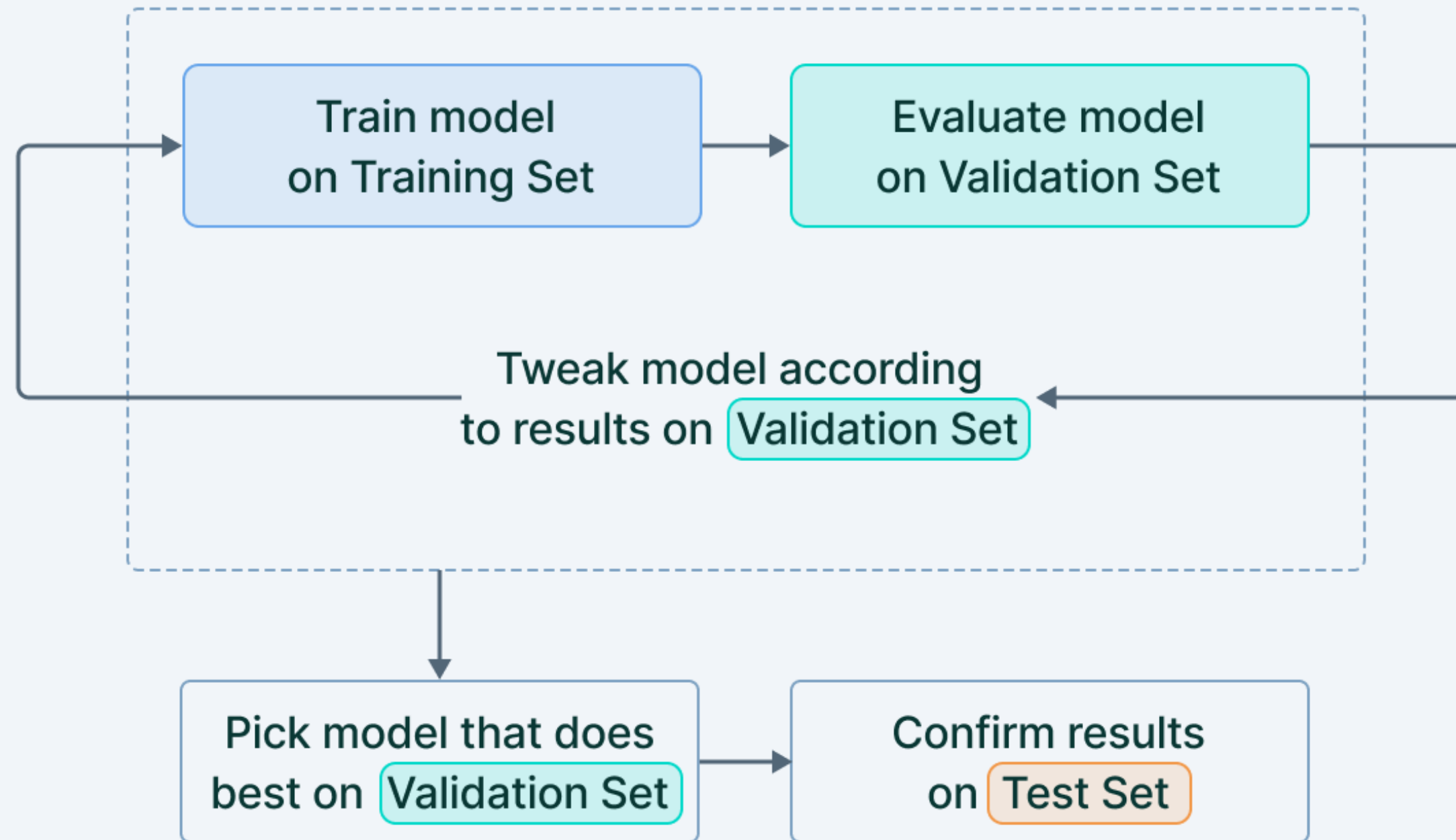
Is the data complete?

Is there illegal content?

Is the data suitable to respond to the problem we want to solve?

Is the data of high quality?

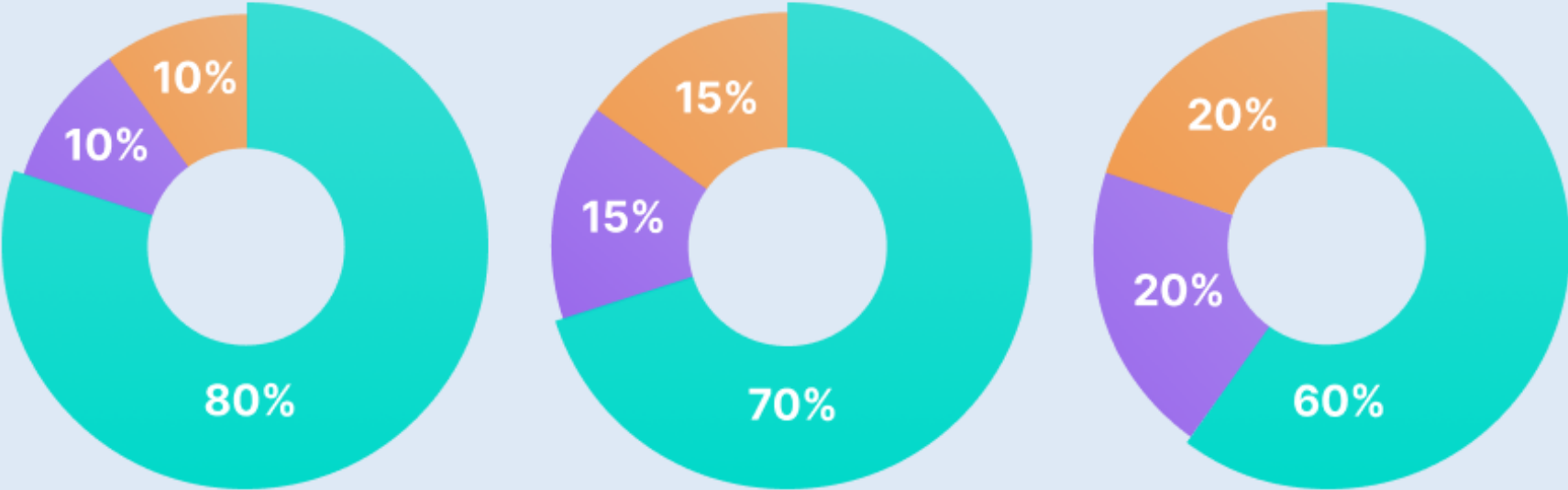
Training data/validation/test



Source: <https://www.v7labs.com/>

V7 Labs

Training Data Needs



Source: <https://www.v7labs.com/>

1

Examples

We want to predict if a case will succeed or not. We collected the following **features** from previous cases:

- Identity of the judge - some judges are more lenient than others
- City - some city Courts are stricter
- Type of proceeding - some topics are more sensitive
- Number of parties - huge proceedings are more complicated
- Number of supporting documents - more evidence increases the chances
- Time of the hearing - after lunch, judges are happier
- Number of years of experience of lawyers - experience is an important factor that can influence the chances of winning
- Number of witnesses - more testimony increases the chances
- Gender of involved parties - males are more aggressive during the trial

What's wrong?

The system finds **relations** in the data and has a high accuracy for our test datasets. However, we did not include the following features in our training:

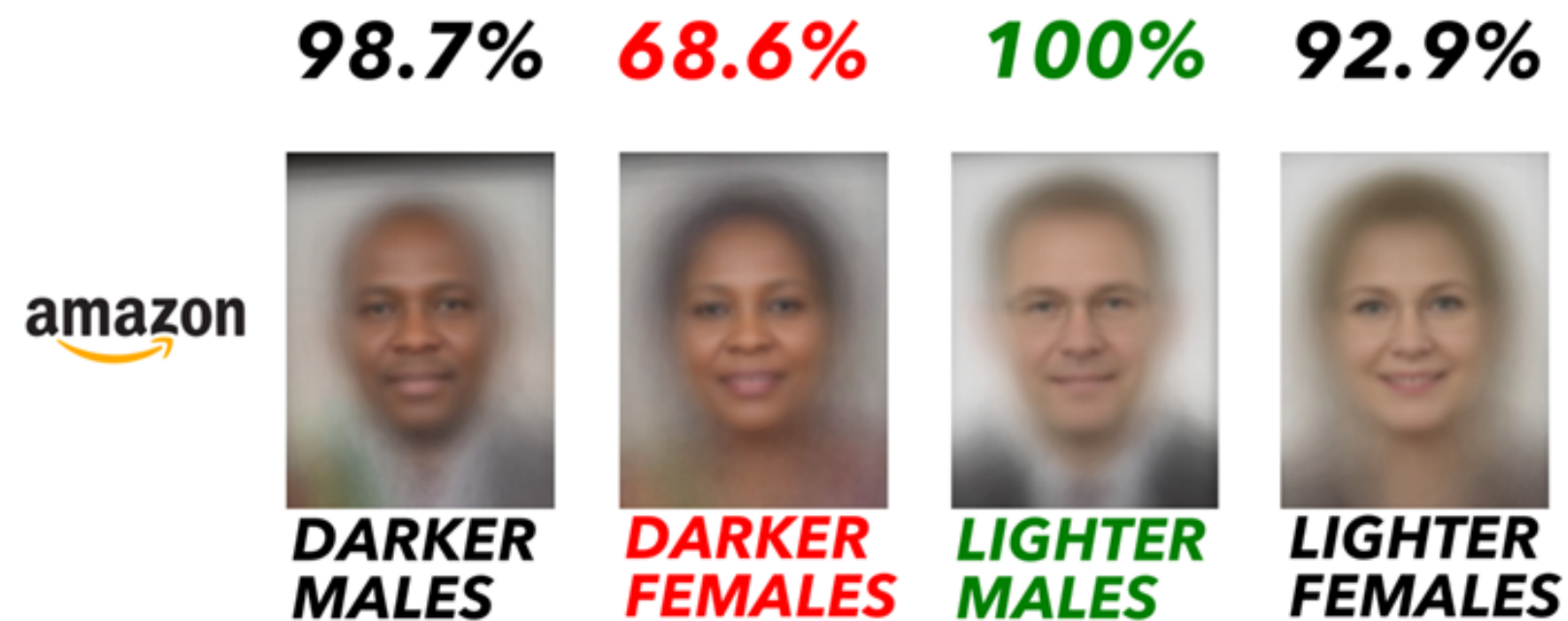
- nationality of claimant - there is a prejudice against certain nationalities and not knowing the language is a barrier to an effective defense
- age of the claimant - older people are considered more reliable
- gender of the judge - female judges might have a different perspective
- gender of the lawyer - female lawyers are not taken as seriously as male
- information about statutory limitation - if close to the expiring date, the chances of rejection are higher
- presence of children - important in family proceedings
- concurring criminal proceedings - can influence the outcome of the civil proceeding

In addition, we only had a **certain amount** of data:

- Name of the judge - 70% of data are from male judges
- City - 60% of data are from large cities
- Type of proceeding - 50% of proceeding are from commercial topics, 20% from labour, 20% from family, and 10% from others
- Number of parties - 30% of proceedings involved only 2 parties
- Number of supporting documents - 80% of cases had more than 20 documents
- Time of the hearing - 90% of hearings are in the morning
- Number of years of experience of lawyers - 90% of data are from experienced lawyers
- Number of witnesses - 80% of cases has only one witness
- Gender of involved parties - 90% of clients are male

If we do not include **sufficient information** in our training data, the system will not be able to generalize when a different case is concerned, and the accuracy will be much lower when we do not have enough examples represented in our training data.

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark



Amazon Rekognition Performance on Gender Classification

Source of image: Deborah Raji and Joy Buolamwini paper in 2019

Spurious correlations are another important problem in AI. Especially when deep learning is concerned, or in the case of unsupervised learning, there is the tendency to include in the training dataset all information that is available. The bigger the dataset is, the better. However, the system can learn from correlated data which is not always correct.

Explain the Prediction

Source of images: 10.13140/RG.2.2.22973.31207



Feature selection

We need to choose all the **variables** in the dataset that are useful for building a model. Adding redundant variables reduces the model's generalization capability and may also reduce the overall accuracy of a classifier. Furthermore, adding more variables to a model increases its overall **complexity**.

Feature = variable = attribute

01

Relevant

Relevant to the problem that is to be solved - it has to convey the right information

03

Not redundant

Sometimes, some features correlate heavily with others, so they become redundant

02

Complete

It is important to include all relevant features

04

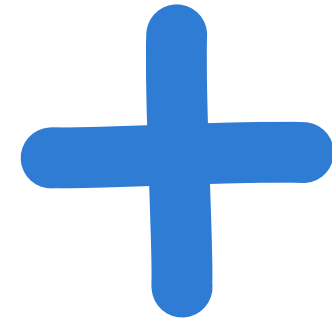
Appropriate

Each feature should have an appropriate number of instances (examples)

2500 winning judgments



Training data



500 winning judgments

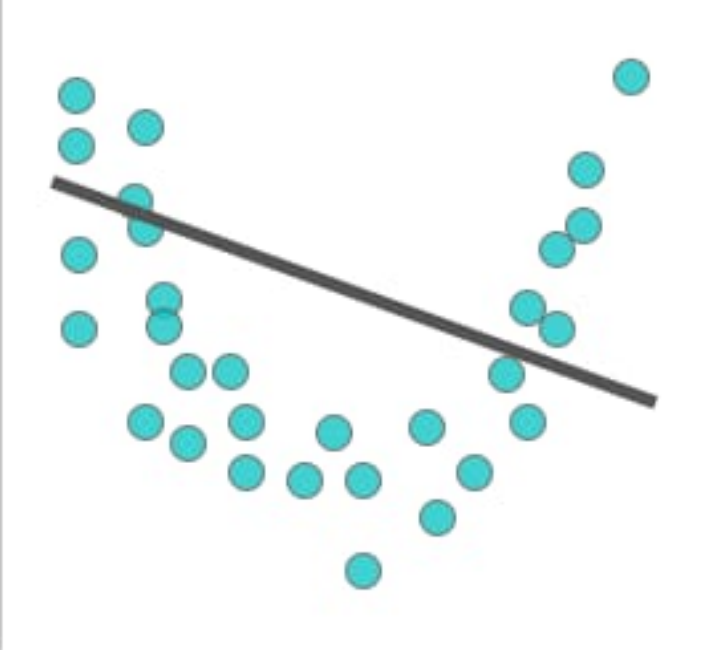
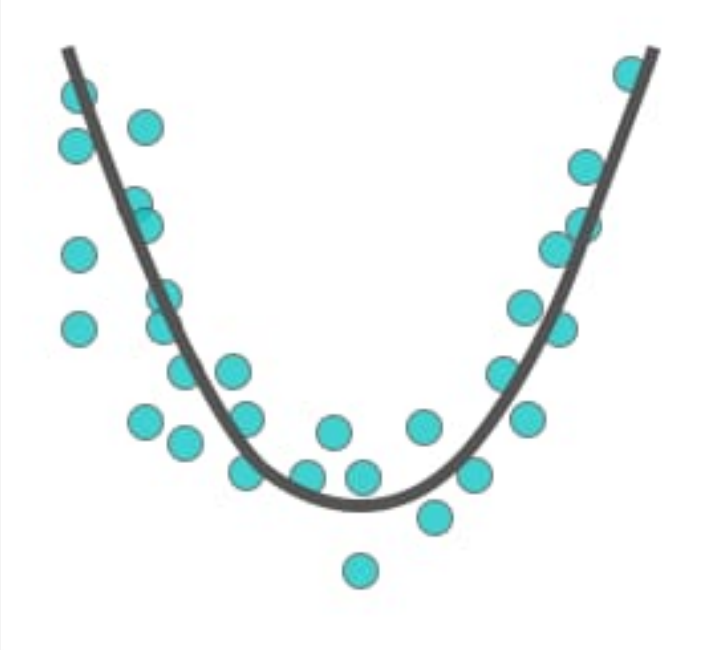
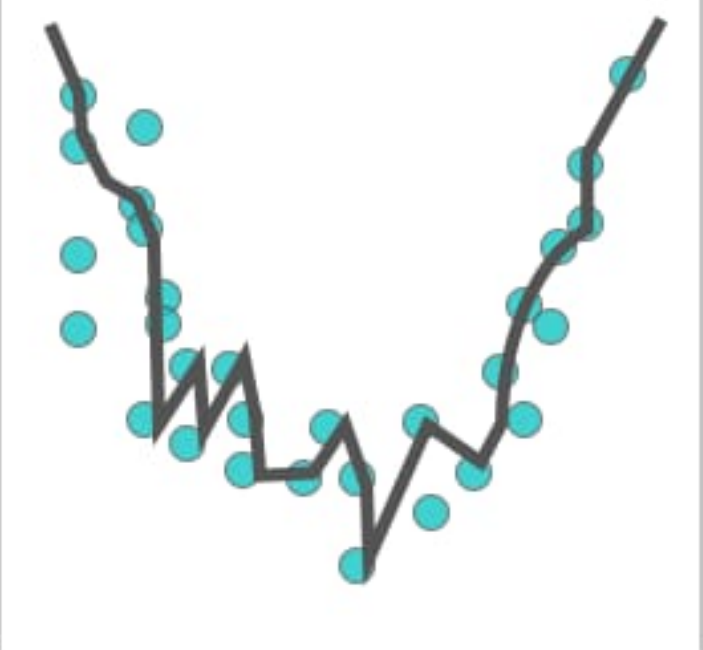
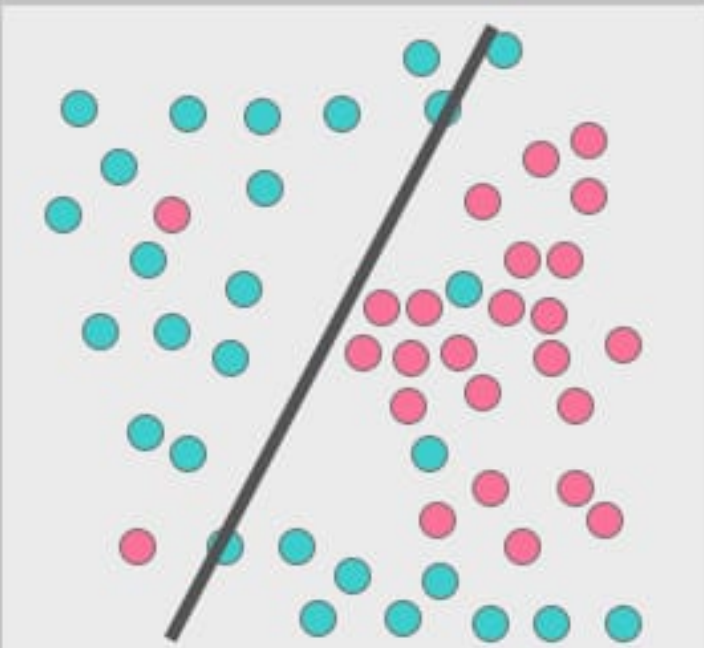
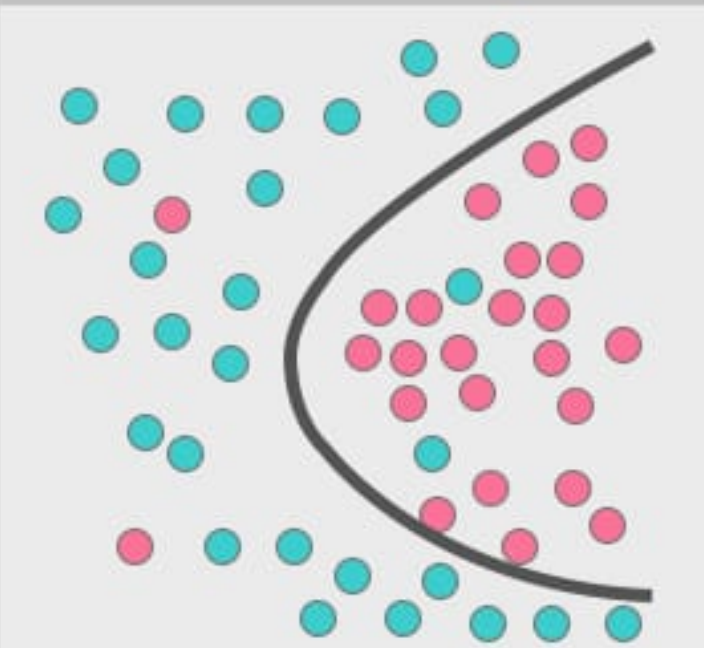
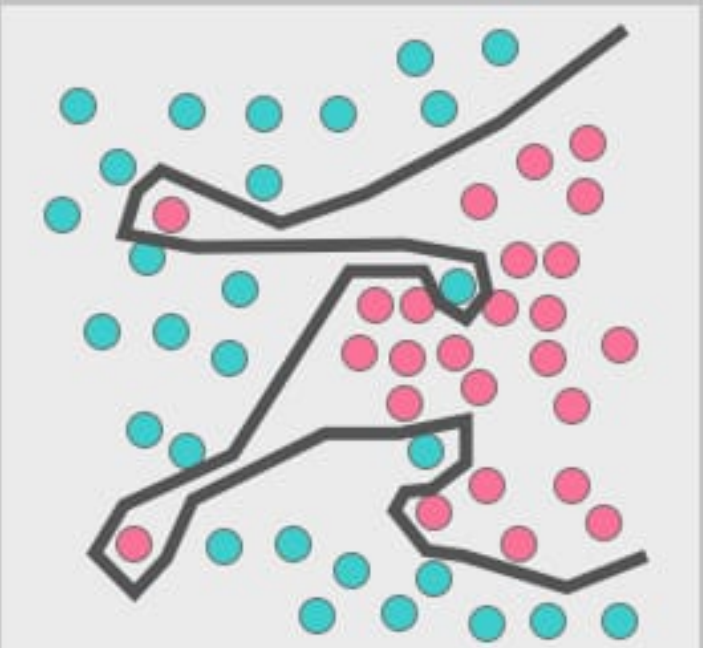


Test data



100% accuracy,
always winning!

The system will always tell you that you will win the case

| | Underfitting | Just right | Overfitting |
|------------------------------------|--|---|---|
| Symptoms | <ul style="list-style-type: none"> • High training error • Training error close to test error • High bias | <ul style="list-style-type: none"> • Training error slightly lower than test error | <ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance |
| Regression illustration |  |  |  |
| Classification illustration |  |  |  |

Source of image: [10.11591/ijece.v12i4.pp4243-4252](https://doi.org/10.11591/ijece.v12i4.pp4243-4252)

As shown, feature selection and collected data are **crucial**. If you miss to collect the right amount of data and to include in your AI model, you will have a **biased model**, even if you do not realize this. This is extremely **dangerous** in the legal field, where fundamental rights - such as the right to defense, to personal freedom, and to cross-examination - are at stake.

In addition, data **based on the past** might not be suitable to predict the future, e.g., after the **law changes**.

To overcome such risks, it is essential that a **domain expert** is involved in the creation of AI models.



**Do you trust programmers
with your life?**

2

Examples

We want to use ChatGPT to analyze legal documents and draft our trial papers.

Can we do that?

Chat GPT has the following problems:

- 1) We don't know the full training dataset
- 2) It does not have internet access
- 3) The training data in Italian is probably very small
- 4) Its analysis is only limited to what is inside its training data
- 5) The final output depends on the prompt
- 6) It does not have reasoning capabilities
- 7) It is not safe for privacy
- 8) Its style is very fluffy and recognizable

Try it yourself asking it to solve one of the bar exam cases!

'Humiliated' NY lawyer who used ChatGPT for 'bogus' court doc profusely apologizes

By [Kyle Schnitzer](#) and [Priscilla DeGregory](#)

Published June 8, 2023 | Updated June 9, 2023, 5:22 p.m. ET

According to the New York Times, the lawyer behind the brief, Steven A. Schwartz, told the judge he had asked ChatGPT to verify its work and it offered reassurances ...

Example of child support prediction

Un sistema per calcolare agevolmente gli assegni di mantenimento per coniuge e figli

I DATI DELLA PRATICA (SEMPRE VISIBILI)

PULSANTI PER ENTRARE NELLE SCHEDE DEI CALCOLI

MONITOR DI CONTROLLO

LA STRUTTURA ORGANIZZATIVA DI REMIDA FAMIGLIA

Struttura delle schede di calcolo

- Determinazione dei redditi di Lui e di Lei
 - Determinazione dell'assegno per il coniuge
 - Determinazione misura assegno
 - Personalizzazione della misura assegno
 - Calcolo per risparmio/ aggravio fiscale
 - Determinazione dell'assegno per i figli
 - Determinazione misura contributo
 - Calcolo assegno con correttivo per tempi di permanenza
 - Calcolo assegno con correttivo per accudimento
 - Calcolo per spese straordinarie

Conteggi coerenti ed affidabili tratti da un'analisi di 2500 casi reali

Enorme risparmio di tempo per i conteggi

Miglioramento dell'immagine di modernità ed efficienza dello studio legale.

Risultati immediati da sottoporre al cliente

ReMida

Example of case law analysis



Lisia

Can we still use AI?

AI can be a useful assistant, but it cannot be a substitute for a human. If you know your subject, you will notice how flawed it can be. You cannot trust its output unless you built it yourself for a very specific problem and application.



Know your model

You need to know the training corpus and the feature selection performed by the programmer. Otherwise, you cannot be sure that the result is not biased.



Human Oversight

You need to apply your own judgement when employing AI models. Is it suitable for the problem you want to solve? Is the output reliable based on your experience? Or is it biased? Is your case very particular?

Conclusions

The use of technology is very important for progress. Even when jobs are lost because of AI, new positions become available thanks to it. The accessibility of AI at all levels, at a reasonable cost, makes legal professions more democratic. Processes become faster and more efficient.

However, attention should be paid to the risks of misuse and biases. AI is built by humans, often biased humans or people with little knowledge of the legal field. This poses a number of concerns that we need to consider before using any kind of technology, and it is crucial to understand how AI works in order to become able to employ it correctly.

By understanding AI's inner functioning and actively working towards mitigating biases, we can exploit technology's full potential in the legal field while ensuring fairness and accuracy.



10/04/2024

Thank You

For Your Attention

Get in touch with me:



Contact to my email

chiara.gallese@unito.it



<https://www.aiandlaw.eu>

<https://www.datacomproject.eu>

