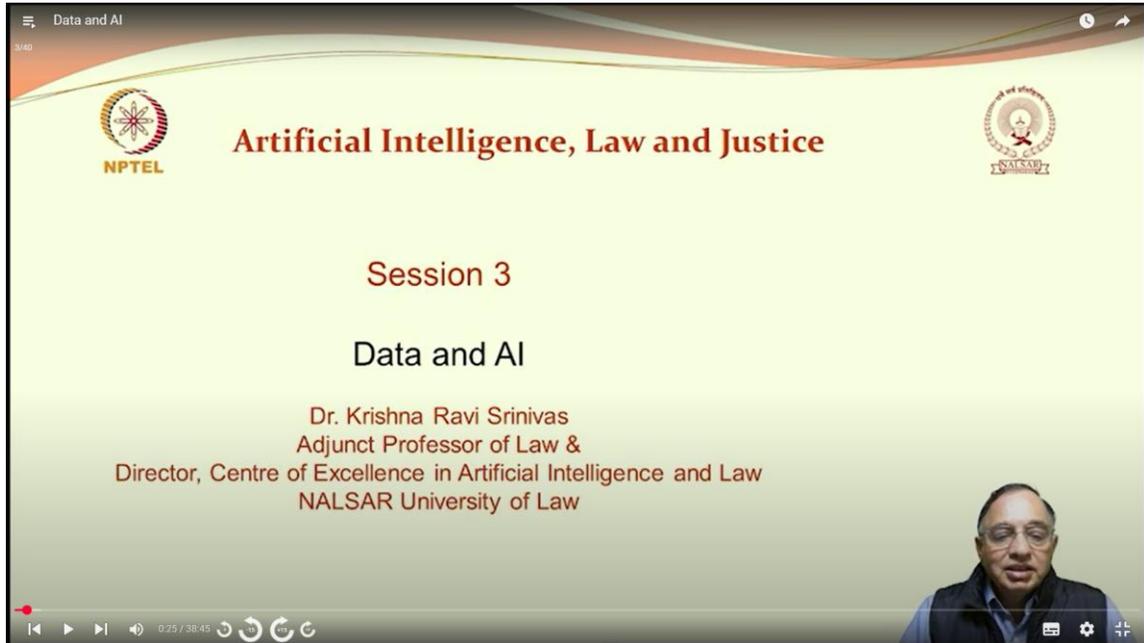


**Course Name – Artificial Intelligence, Law and Justice**  
**Professor Name – Dr. Krishna Ravi Srinivas**  
**Department Name – Center of Excellence in Artificial Intelligence and Law**  
**Institute Name – NALSAR University of Law**  
**Week – 01**  
**Lecture – 03**



The screenshot shows a video player interface. At the top left, it says 'Data and AI' and '3/40'. The main slide has a light green background with a wavy top border. On the left is the NPTEL logo, and on the right is the NALSAR University of Law logo. The title 'Artificial Intelligence, Law and Justice' is centered in a large, bold, dark red font. Below it, 'Session 3' is written in a smaller, dark red font. Underneath that, 'Data and AI' is written in a black font. The presenter's name and title are listed in a smaller, dark red font: 'Dr. Krishna Ravi Srinivas, Adjunct Professor of Law & Director, Centre of Excellence in Artificial Intelligence and Law, NALSAR University of Law'. In the bottom right corner, there is a small video window showing the presenter, a man with glasses wearing a dark vest over a light blue shirt. At the bottom of the slide, there is a video player control bar with a red progress bar, a play button, a volume icon, and a timestamp of '0:25 / 38:45'.

Artificial Intelligence Law and Justice Session 3 - Session 3 is on data and Artificial Intelligence. In this session, we will deal with the key concepts in data and AI, and then an understanding of these concepts will help us understand better not only the relationship between data and AI but also the pervasiveness of the clear description between data and AI, which is something that is inseparable; today, we cannot conceptualise AI without data, but data on its own has so many other significances. However, when it comes to AI, it has a special significance. How that happens, we will have a look at it.

The image is a screenshot of a video lecture. At the top left, it says 'Data and AI'. The main title is 'Recap-Rule of Law' in a large, bold, red font. Below the title is a bulleted list of topics: 'Development of the Rule of Law and its key features.', 'Contextualization of Rule of Law for India and Indian Constitution.', 'Some important cases on Rule of Law', 'Recent analysis on Rule of Law', and 'Role of Digital Technologies in Rule of Law'. The NPTEL logo is on the left and the IIT Bombay logo is on the right. A video player interface is at the bottom, showing a progress bar at 1:08 / 38:45 and a small video feed of the speaker in the bottom right corner.

So before doing that, let us have a recap of the last class. In the last class, we discussed the rule of law. We mentioned the development of the rule of law, starting from the Greek times and how it has developed. We also underscored its key features. Then we went into an in-depth discussion on how the rule of law has been contextualized in the Indian context, particularly through the constitutional scheme of things, as well as by other means, and the role of the Supreme Court in underscoring the importance of the rule of law, as well as affirming it, expanding the scope, and ensuring that the rule of law is well understood and implemented in this country through different cases. So, the cases that we saw also gave us an understanding that the Supreme Court is an institution that upholds the rule of law in more than one way.

Besides discussing some of these cases on the rule of law, we also saw how the recent analysis of the rule of law, particularly the idea that the rule of law is nothing but a contract between law enforcers, lawmakers, people who are associated with different organs of the state and its institutions, and the common citizens, shows how the rule of law can be the cement or a contract that binds both of them, binding in the sense that both have to abide by the rule of law. So, when we also saw that the rule of law has a lot of dynamism inbuilt, it is an evolving agile concept, but these days the real challenge to the rule of law comes not from the idea of law or from the constitutional scheme of things, but from the rise of digital technologies, which we said we would be discussing in detail in class in the subsequent sessions. Rather, we also looked at the way digital technologies can impact the rule of law both positively and negatively, and then how we deal with that impact on the rule of law from digital technologies.

Data and AI

NPTEL

## Why Data Matters in/for AI

- Data can be considered as the resource for AI to 'train' machines
- AI algorithms need access to data to perform tasks
- For machine learning and other purposes huge quantity of data is needed in AI



3:18 / 38:45



So, in this class, we will first look at why data matters for AI. Now, data can be defined broadly as anything that we can record, store, visualise, or understand. Societies have been dealing with data right from time immemorial. For example, even the edicts on the stones are different edicts or data. The earlier codes of conduct or the earlier codes of laws written in different ways in different media are also data, but so is the data that could be scattered from different sources of the earlier periods. But data can be considered a resource for AI to train machines. As we saw in the first class, machines need to be trained; they need to undergo some training, and then data is fed to them so that they can perform that training.

And without data to train the machines, they will not be able to understand, nor will they be able to do things the way we want them to. So, data is a key resource for AI. But more than the machines themselves, artificial intelligence algorithms also need to access data to perform tasks. Why? Because algorithms can scrutinise something and then give a decision, can process enormous quantities of data and give decisions, or can arrive at some conclusion, but primarily, without data, they cannot do anything. For example, the decision tree looks fine; it has a whole lot of branches, yes or no, but unless I input something, you won't get anything out of it, or unless you are able to provide sufficient data, and that data is also of good quality, the algorithms will not work the way we want them to work, or we will not be able to use them to the fullest extent we desire.

So, for algorithms, we need access to data so that they can perform the task. But the amount of data required for AI purposes is huge. The reason it is huge is that we are expecting too much from AI, and as AI applications are built more and more, they become widespread and are applied across different sectors in different areas of life. Access to data, or more importantly, the availability of data, is becoming important, so AI, in one sense, has an insatiable appetite for data, but in another sense, the demand for data has also arisen because our modern society itself is built on datafication. So, data today is not abstract data; data today is not data that is scattered around; data today is not

data which we can simply say, "Oh, unless you see it, nobody is going to see it," or "Unless we do something with it, nobody is going to make use of it," it is just blank; nothing will happen out of it.

So, data that is inscribed in different media will have value when it is put to use, and then value addition will happen. Data, when digitised, can be easily shared, manipulated, and spread across the world. Data flows can happen across the globe in a few seconds, so data is something like oil or a resource that can flow across the globe in no time. When we have this understanding of the digitisation of data, we will also know why we need to pay attention to some concepts, including datafication, society, and AI.

The image shows a video player interface for a lecture titled "Data and AI". The slide content includes the NPTEL logo, the title "Datafication, Society and AI", and a list of bullet points. A small inset image shows two people interacting with a digital data visualization. A video feed of the speaker is visible in the bottom right corner.

NPTEL

## Datafication, Society and AI

- Day and day out we consume, generate and share data knowingly or unknowingly
- Earlier societies too had generated and used data and processed them
- With progress in technology the capacity to deal with data increased
- But digital technologies have enabled generation of data, organizing them and using them in a scale that was not possible before

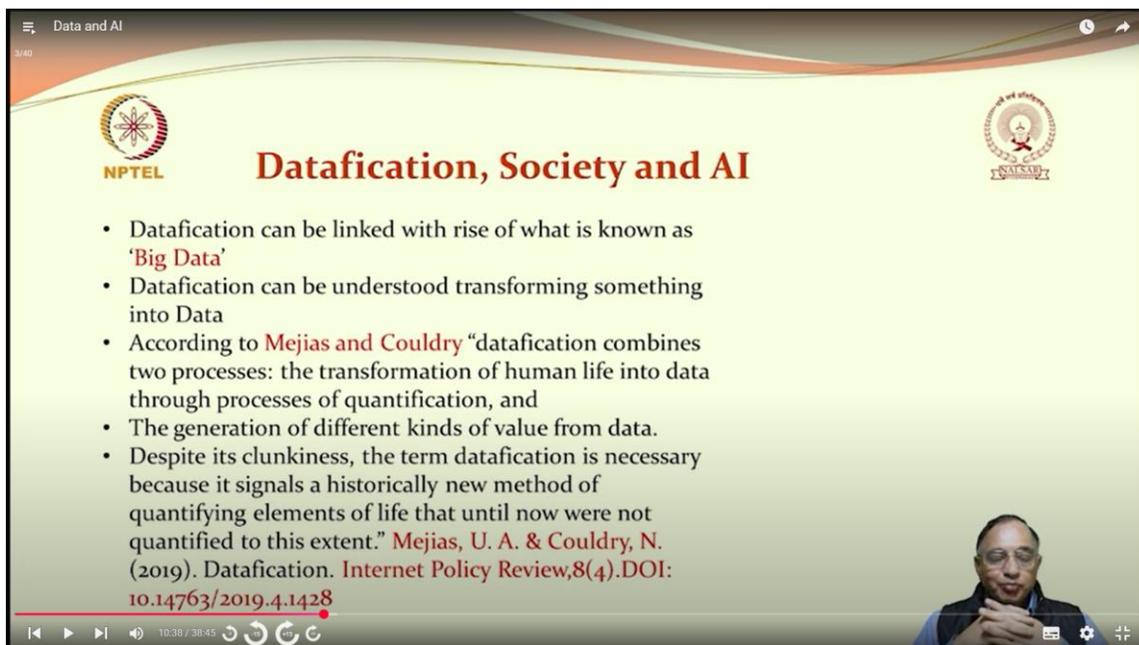
6:56 / 38:45

So, what is datafication? Day in and day out, we all consume data. When I see a newspaper, I consume data. When I go to the railway station, I book tickets and give data to the booking agent or the person who is doing the booking. That data pertains to my travels. Based on that data, he or she allocates me a seat and then gives me a ticket that again has data. So, day in and day out, we consume that data; we generate data.

For example, when I go out and take a selfie, my mobile phone records that data; it not only records my image, but it also records a couple of other things, like where it was shot, at what time it was shot, and which phone was used to take it. So, data is generated simultaneously whenever we do something, and that data generation occurs whether we are aware of it or not. So, in that sense, we are consuming, generating, and sharing data on a day-to-day basis, whether we know it or unknowingly, in different ways through different media and then through different systems of doing things. But earlier societies also generated a lot of data and then used that data and processed it with the available technologies. So, you need technology not only to generate data, not only to store data, not only to process data, but also to make use of it, to disseminate it, to share it, to guard it, to hoard it, and for various other purposes.

So, what has made the huge difference now? The huge difference that has been made is the progress in technology. In fact, historians of data or datafication have traced how data has played different roles in different contexts and times. When censuses were organised and taken, enormous amounts of data were generated, and then when governments expanded the scope of tax and commerce, they found various means and ways to levy and collect taxes and to control people's movement, whether when they tried to migrate or when they tried to move from one place to another. They not only collected data but also processed it and used it to monitor, manipulate, and govern people, surveilling them in different ways. Thus, data again became a very powerful means to control, understand, and guide society in one way or another.

So, when we talk about that, we need to look at the historical evolution of data capture through different means and how it has been used by different institutions in different contexts. But the paradigm shift has occurred through digital technologies because they have enabled the generation of data, organizing it, and using it on a scale that was unprecedented, one that was not even imagined before. So, the 20th century and 21st century is more or less enormous, given the enormous power of digital technologies for the datafication of society, as well as the implementation of datafication in our own daily lives. We are there not only as consumers and producers of data but also as part of the larger framework or larger web through which the data flows.



The image shows a video player interface for a lecture titled "Data and AI". The slide content is as follows:

**Datafication, Society and AI**

- Datafication can be linked with rise of what is known as 'Big Data'
- Datafication can be understood transforming something into Data
- According to **Mejias and Couldry** "datafication combines two processes: the transformation of human life into data through processes of quantification, and
- The generation of different kinds of value from data.
- Despite its clunkiness, the term datafication is necessary because it signals a historically new method of quantifying elements of life that until now were not quantified to this extent." **Mejias, U. A. & Couldry, N.** (2019). Datafication. *Internet Policy Review*,8(4).DOI: [10.14763/2019.4.1428](https://doi.org/10.14763/2019.4.1428)

The video player shows a progress bar at 10:38 / 38:45 and a small video feed of the speaker in the bottom right corner.

Datafication can be linked to the rise of what is known as big data. Big data means different things in different contexts to different people, but it can also be understood as data that is available in huge quantities in a way that it can be put to use. Value addition can happen, and more importantly, value addition can be derived. Furthermore, the data processing itself, by different means, can not only enhance value but also share the value through different means. So, to put datafication in a different way, we can simply say that datafication occurs when we transform something into data. For example, the image that my mobile phone captures is just an image.

But then datafication occurs when that image is stored in some system, analysed, and its various characteristics noted, and particularly the camera that was used, the time at which it was taken, and the geographical location where it was captured are also taken into account. It means that datafication occurs in this context. So, when we transform something into data, the data can be either tangible or intangible. That doesn't make a difference. Datafication occurs when this transformation takes place.

According to Mejjas and Couldry, data application combines two processes. One, they transform human life into data through the process of quantification. Two, the generation of different kinds of values from data. Now that the transformation of human life into data happens through different processes. For example, through ways of measuring our health parameters, through sensors, through different ways of collecting taxes, income taxes, customs taxes, and other methods, all these things are also part of our lives, which again results in quantification.

Then the second thing happens: different kinds of value arise from data, so the data application concepts may be found to be very vague and undefinable, meaning anything and everything, but that is not the problem that matters here. What matters is that it historically signals a new method of quantifying elements of data that, until now, were not quantified to this extent. For example, DNA mapping. DNA mapping has unleashed the power to map the DNA to our genome and then to understand the scope of the DNA, the potential to predict what diseases could arise, whether the genome has some abnormalities, or if it is a normal genome. Now, when we analyse the genomic data, we can learn many things from it, although the person from whom the data has been extracted may not even be aware of this or may not even realise that this data holds so much value; it can predict their future in one sense or another and has huge ramifications for their life.

So, when we talk about that, we mean we are quantifying the elements of life that we were not able to quantify to this extent, in the sense that in the previous era, we didn't have access to genomic data to quantify it in terms of the potential to have a specific disease or condition that could develop in later years. So, this sort of quantification was not available, and would not be done before. So datafication happens when we do this sort of quantification. So, this datafication also means that we can add value, reduce value, interpret it, combine it with different methods of datafication, and then obtain a 360-degree view of a person's life by combining various sources, including their bank financial transactions, health records, official records, educational records, and other means. So datafication happens when data is combined, put together, values are added, extracted, and shared.

Data and AI

NPTEL

## Datafication and AI

- The relationship between AI and Data is two fold
- AI needs huge quantity of data and hence datafication enables development and deployment of AI
- But AI also facilitates datafication by creating fresh data and outputs and makes it an efficient and wide spread process
- So AI is inseparable from Datafication



15:08 / 38:45

Having said that, we should know why datafication is important for AI, because the relationship between datafication and AI is twofold. This is again a very simplistic picture. The picture is much more complex. AI, as we said, needs a huge quantity of data, and datafication enables the development and deployment of AI. In other words, AI, as a steam engine, cannot run unless "the data, which is the core, is fed into it." So, AI without data is just AI. It won't do anything meaningful, nor can whatever it does make any sense. But AI has also facilitated datafication because what does AI ultimately do? AI generates a lot of outputs; it analyses data and then does so in a very efficient and widespread process. And then, when AI does it, it also facilitates datafication because AI does not process data and then stops with that. AI processes data; it adds a lot of value, using AI data to come up with outputs to help us in decision-making or arrive at decisions.

All of this facilitates datafication. So, AI is inseparable from datafication. So datafication, data, and AI are very closely linked, and in fact, we cannot think of datafication without AI today. Nor can we think of AI today without data. Of course, data can exist without AI. Of course, datafication can also happen through various means without using AI. But as society becomes more and more integrated through AI, and as society depends more and more on AI and different technologies, the digitisation process is bound to lead to datafication on an unprecedented scale, a scale that we would have never imagined or even thought of a few decades ago.

Data and AI

NPTEL

## Metadata

- Metadata is data about data
- A digital image often has information about location, date and time and phone/device used to capture
- Unstructured data has little value however huge it may be
- But structured data or information can be used in ML
- They can be evaluated, analysed and compared by algorithms



17:13 / 38:45

Metadata: what exactly is metadata? Put simply, metadata is data about data; in other words, metadata tells us about the data. Metadata can be a thing that gives you information combined together to provide a better idea. For example, the metadata about a person's health can come from a normal physical examination, which includes blood and urine samples, temperature, and other factors, but it can also include genomic data.

So, when the metadata is data about data, we talk of the metadata that has much more value than plain raw data, which is just a discrete digital element. As we said earlier, it has information about location, date and time, and the phone device used to capture it. All these get embedded in the digital image itself, and there is no need to capture them separately. So, in one sense, this is an example of metadata, but the problem here is that unstructured data has little value, however large it may be. For example, unstructured data on the population of a city or a population of a ward, which has merely unclassified data of citizens, their addresses, their ages, and their incomes, will not make much sense.

Data has to be structured based on some parameters, or it has to be sorted and organised according to certain criteria or, for that matter, on the basis of some rationale or logic. Census data or data about a person's ward can be structured according to age, gender, income, or sub-classification in the sense that if they are in a municipality, which ward or which part of the city they are in; thus, structured data is data that can provide information that can be used in machine learning, and when structured data is fed into an algorithm, it can be evaluated, analysed, and compared by the algorithm. For example, when I give structured data about the population of a city to an algorithm and then ask the algorithm to tell me how many of those citizens are entitled to vote, it can do that because I am providing it with structured data, and I am also giving it the rationale to explain whether a person can vote or not, so this sort of metadata or structured data is very useful and relevant for AI and the use of algorithms because raw, unstructured data on its own will not have much value for such uses.

Data and AI

3/40



## Data Quality, Algorithms

- However algorithms need data is accurate, reliable and fit for the purpose
- Hence quality of data including authenticity and accuracy makes a difference
- If data is of poor quality or un reliable then it can result in 'GIGO' outcome
- GIGO is Garbage In Garbage Out



20:15 / 38:41

So obviously, data quality is very, very important. But, as we said, algorithms need data that is accurate, reliable, and fit for the purpose. When we say fit for the purpose, we mean that the structured data itself, which the algorithm has to use, should be sufficient for the purpose. For example, if the algorithm has to simply decide whether the person has to vote or is entitled to cast a vote, that particular information should be available. It does not make sense to fit the algorithm when we want it to, reflect the data that pertains to income but then does not provide information about the age of the individuals we want to assess in terms of eligibility to vote. Hence, the quality of the data, including the authenticity and accuracy, makes a huge difference. Normally, the authenticity and accuracy have to be vetted by someone or assured before the algorithms step in, try to do the number crunching, and come up with outputs, indications, or the way we want them to decide.

Put it in other words: if the data is of poor quality or under-labelled, then it can result in a GIGO outcome. GIGO outcome means garbage in, garbage out. Let us understand one thing very clearly. If we feed garbage to the system, the AI system, or the algorithms, we can expect only garbage out of it because the algorithms will look at it as data fed to them. On their own, they can learn and do many things, but if the fundamental problem is that the data itself is garbage, the output cannot be anything but garbage. So, if we understand that the output is going to be GIGO, we need to recognise that data quality is something to which we need to pay attention.

Data and AI

NPTEL

## Data Quality, Algorithms

- There can be issues with data – errors, irrelevance, inaccurate and inadequate
- Data may have biased information and may not be representational as needed
- If data from clinical trials does not have the right type of data because clinical trial did not include women, rural population or their representation was inadequate
- Then Algorithm that uses that data for training might result in faulty or wrong inferences




22:14 / 38:45

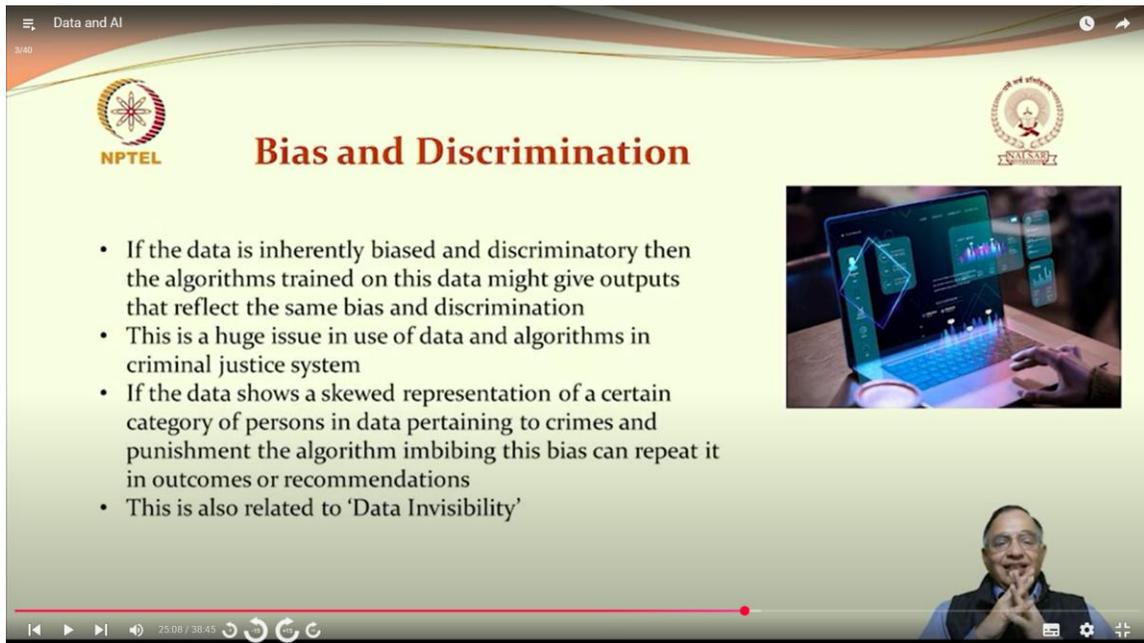
There can be many issues with data quality in various ways; for example, the data can have multiple errors. The census data, or the data that pertains to citizens of a city, can have many errors; certain columns might not be filled out, or might be wrong. For example, when we say gender, the column could be blank, which could indicate an error, or the data that has been collected could be irrelevant. The data is mixed with the data of humans, non-humans, and animals in census data, or data from a location will not have much relevance.

It could be inaccurate in terms of other parameters, and it could be inadequate in the sense that the data we need to fulfil a purpose, meet an objective, or achieve a goal should be very relevant and sufficient for that purpose. And then, more importantly, the data can have biased information or may not be representative as needed. Now, the bias in information can occur from different sources. It can occur from the process of data collection itself. For example, when we collect data from clinical trials, if the data does not include the right type of data, it is often because the clinical trial data did not include women, rural populations, or their representation was inadequate.

The health data will not be sufficient; it could be wrong and is biased, not because the data per se is biased, but because the process was biased. There were exclusions; there were certain categories of people who were left out, and the representation was not adequate. Put in another way, in a normal society, the distribution between genders will be more or less 50 to 50. But if the clinical trial did not adhere to that norm and the clinical trial data had only 70% men and 30% women, or vice versa, the representation was insufficient and inadequate.

So, it is poor data, or the data quality is poor. But when an algorithm uses that data for training, it could result in faulty inferences. So, the algorithm, underlying logic can result in faulty inferences or wrong inferences because the data fed was wrong. So, the training

purpose would not have been met fully or adequately when the data suffers from serious quality issues, including representation, bias, and discriminatory data that is built in.



The screenshot shows a video player with a slide titled "Bias and Discrimination". The slide includes the NPTEL logo and the NALSAR Hyderabad logo. The main content is a list of four bullet points. To the right of the text is an image of a person's hands typing on a laptop with a futuristic digital interface overlaid. At the bottom right of the slide, there is a small video inset of a man speaking. The video player controls at the bottom show a progress bar at 25:08 / 38:45.

- If the data is inherently biased and discriminatory then the algorithms trained on this data might give outputs that reflect the same bias and discrimination
- This is a huge issue in use of data and algorithms in criminal justice system
- If the data shows a skewed representation of a certain category of persons in data pertaining to crimes and punishment the algorithm imbibing this bias can repeat it in outcomes or recommendations
- This is also related to 'Data Invisibility'

So, what is bias and what is discrimination? When we say bias, in one sense, it means that leaving out certain people or giving more importance to certain people, or there is a lack of equality; there is a lack of proportional representation in data, so if the data is inherently biased and discriminatory, then the algorithm might give outputs that reflect the same bias and discrimination. For example, when we have such data where the population being studied is 70% men and 30% women, and then the data is analysed to look into anaemia as a disease or as a health indicator, since the data itself is not fully representative, bias will arise.

Data might give a wrong inference that women are more prone to anaemia because we have collected only 30% of the samples from women, not 50%. Only when women contribute 50% of the samples can the data interpretation be somewhat reasonable and adequate enough to make sense. But this is a huge issue in the use of data and algorithms in the criminal justice system, which is a major problem in the use of AI and algorithms in the criminal justice system. Why? Because the data would be based on a biased sense of classification, there would be bias in terms of collection, bias in terms of labelling, and other parameters. So when data shows a skewed representation of a certain category of persons in data pertaining to crimes and punishment, the algorithm that imbibes this, or the algorithms that are trained on that, will take that bias; they will be fed biased data and will repeat it, saying that people from a specific location or people from a specific group are more prone to commit a crime, more prone to violent behaviour, or, in other contexts, may say that they are more prone to default in repaying loans or more prone to default when they use a credit card, they do not pay the amount that is due from them on time.

So, the recommendations and outcomes could be biased, not because the algorithm per se is wrong or biased, but because the data that was fed into it was biased. There is also this

whole concept of data invisibility when the data collection process itself makes certain people or certain categories of people invisible because data collection was not done in those regions or locations, or because the data collection process itself was not fully representative and made certain people invisible. For example, in a city of 100 wards, the data collection should be done for all 100 wards. If the data collection is restricted to 80 wards, the population of 20 wards is being made invisible. But if the data collectors assumed or never told that the data collection was the full and complete data for that city, the invisible portion or the people who were made invisible, were not part of it, then bias and discrimination are inbuilt; they emanate from the very process of data collection itself. Forget about the other biases and discriminations that could arise from other factors.

The image is a screenshot of a video lecture slide. At the top left, it says 'Data and AI' and '3:40'. The slide features the NPTEL logo on the left and the NALSAR logo on the right. The title 'FAIR Data' is centered in a large, bold, red font. Below the title is a bulleted list of points. To the right of the list is an illustration of a person in a blue uniform interacting with a digital interface. At the bottom right, there is a small video feed of a man speaking. The bottom of the slide shows a video player interface with a progress bar and various control icons.

NPTEL

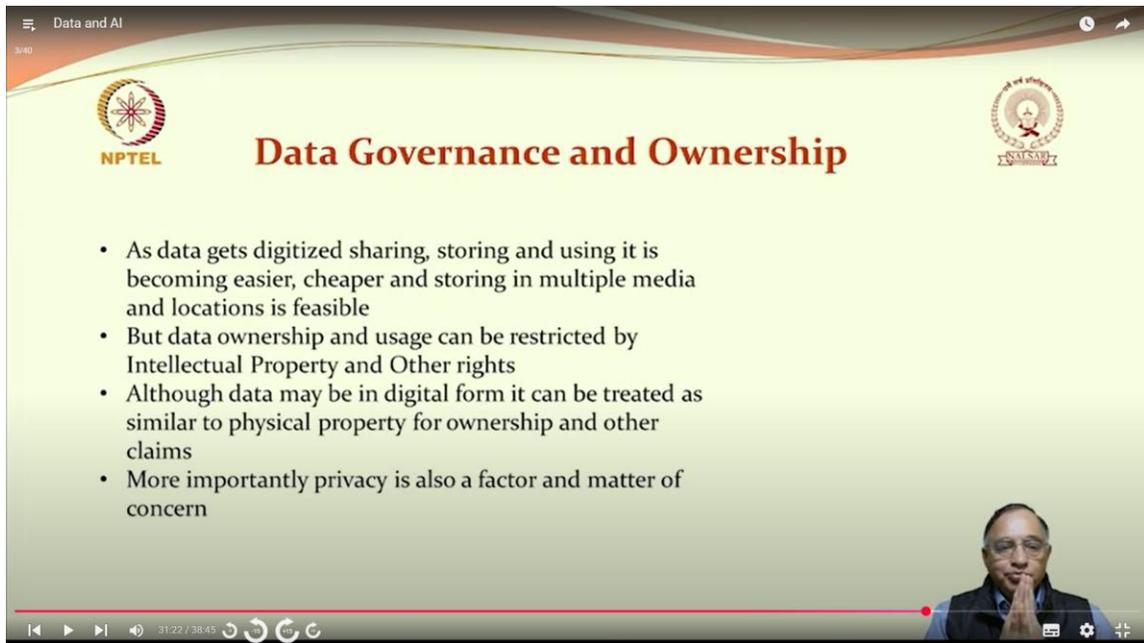
## FAIR Data

- The issues could also be due to lack of digitization or inadequate digitization
- So when using and accessing digital data it is better to ask 'FAIR'
- Is it Findable
- Is it Accessible
- Is it Interoperable, and,
- Is it Reusable
- Developing adequate quality controls, check lists and criteria for use of data is desirable

So, how do we address that? One way to address that is by going through a criterion called FAIR. As we said, it could be due to a lack of digitization or inadequate digitization. So, we look at the data, whether it is metadata or raw data, through four lenses or four criteria.

First, is it findable? Can we find the data that is truly representative? That is the first criterion. Is it accessible? In the sense that, is the data accessible to the persons, to the algorithms, or to the AI systems when it is needed? Is it interoperable? When we talk about interoperability, the data should be able to be fed, understood, processed, and then an outcome could be derived from it through different systems that are based on different operating systems and different AI systems, in the sense that can we use the same data in the same way across systems, the data which was used for one purpose should be utilized for another purpose through the different systems. This means that systems are able to communicate with each other when they use the same data. Is it reusable? When we say reusable, we mean can we use the data for different purposes? For example, the data collected from a society would be made reusable for different purposes in terms of criteria such as income, location, age, and gender. But if the data are not fully

representational or incomplete, it will not be fully interoperable. It will not be fully reusable. So, to do that, we need to have adequate quality controls at different stages of data collection, data checking, data processing, and data labelling. Then we need to have a different checklist and criteria for the use of data. It is here that the four FAIR principles have a very prominent role to play.



The image is a screenshot of a video lecture slide. At the top left, there is a menu icon and the text 'Data and AI'. Below it is the number '3/40'. The slide features two logos: NPTEL on the left and NATA on the right. The main title is 'Data Governance and Ownership' in a large, bold, red font. Below the title is a bulleted list of four points. At the bottom right, there is a small video inset showing a man speaking. At the bottom, there is a video player control bar with a progress indicator at 31:22 / 38:45 and various playback controls.

NPTEL

## Data Governance and Ownership

- As data gets digitized sharing, storing and using it is becoming easier, cheaper and storing in multiple media and locations is feasible
- But data ownership and usage can be restricted by Intellectual Property and Other rights
- Although data may be in digital form it can be treated as similar to physical property for ownership and other claims
- More importantly privacy is also a factor and matter of concern

So, when we discuss this, we get into the concept called data governance. As data gets digitised, sharing, storing, and using it becomes easier and cheaper in multiple media and in different locations across the globe or within a country. But data ownership and usage can be restricted due to intellectual property. Because those who capture the data might claim that they have intellectual property rights in terms of copyright, database rights, or a combination of exclusive copyrights through them. And then there are other rights. For example, different laws can give different users or stakeholders other rights. The rights include the right to privacy, the right to access to health data, and the right to control access to others in terms of sensitive data, including facial and health information.

But data in digital form can be treated as fiscal property for ownership and for other purposes. So, digitization per se may not give unbridled access to a person who wants to access it because data control, regulated through different laws and institutional mechanisms, can create layers and layers of data ownership and control mechanisms, so the data, irrespective of its physical or digital format, may be controlled through different means, such as logs or other ways, so that ownership and control are established. More importantly, as we will see in future discussions, privacy is a huge matter and a matter of concern; it is also a factor. So, privacy can determine whether data access can be enabled or not. In other words, when a person, entity, or institution fears that privacy rights will be violated when the data is processed or made available, it can step in and impose a barrier. It can put either a partial or a full stop to ensure that the privacy concerns are addressed adequately.

Data and AI

3:40

NPTEL

University of Hyderabad

## Data Governance, Law and AI

<https://barc.com/data-governance/>

33:52 / 38:45

Now we will come to this in a different way. Data mean different things for different institutions for different purposes. For a firm that is engaged in data governance at a micro level, data governance means different things. Data governance here means one thing: data architecture. Is the data architecture sufficient, and are we able to build good data models that can make use of this data? Do we have sufficient infrastructure to store them and to operationalise them? And then do we have sufficient data protocols, data access regimes, and sufficient hardware and software to ensure that there is data security that is built in? And then do we have the capacity to integrate data and to make it interoperable? And then, do we have sufficient data content management for it? Then, do we have metadata management schemes in place so that the references and mastering of data are there, and more importantly, do we have data warehousing and business intelligence capabilities that can make the best use of the data? And more importantly, do we know how to build metadata or metadata bases, and then how do we ensure that our data governance is in accordance with legal norms, adheres to legal principles, and meets all the criteria set by the government and other institutions? So, at a micro level of an institution, data governance means different things. But fundamentally, let us also understand that it means only one thing: data governance has to be in accordance with the law, has to adhere to established legal principles and legal norms, and should comply with the data governance norms set by different institutions. So, at a micro level, this is what data governance means. But at a macro level and at a different level, it is very broad.

The screenshot shows a video player interface. At the top left, it says 'Data and AI' and '3:40'. The main content area has a light green background with the NPTEL logo on the left and the NALSAR logo on the right. The title 'Data Governance, Law and AI' is centered in a large, bold, red font. Below the title, there is a bulleted list of four points. At the bottom right, there is a small video feed of a man with glasses speaking. The video player controls at the bottom show a progress bar at 35:37 / 38:45 and various playback icons.

Data and AI  
3:40

NPTEL

## Data Governance, Law and AI

NALSAR

- In a broader sense Data Governance means governing data as a resource through policies, strategies, rights, responsibilities and institutions.
- Data Protection Laws are key component in this
- The Digital Personal Data Protection Act in India, GDPR in Europe are examples
- There are data governance principles that are applicable depending upon the context, use, type and value of data

35:37 / 38:45

Here we look at data as a resource on a macro level. So, it has to be governed through different policies, strategies, rights, responsibilities, and accountabilities, and then liabilities will pitch in. So, in a broader sense, it is a broader picture we need to look at. In this, the data protection laws play a key component and a major role. Because ultimately, they are the ones who govern how data should be protected by different stakeholders and then by the government itself. For example, in India, the Digital Personal Data Protection Act is being enforced, but not fully, because the rules have not been framed completely. So, the Digital Personal Data Protection Act, or DPDP Act, is the primary act that governs data governance in India. In the context of Europe, the General Data Protection Regulation (GDPR) is the binding regime that is applicable to all data being processed in Europe or processed through different institutions in Europe. So, these two are the primary examples of data governance. Then the data governance principles are applicable depending on the context, use, type, and value of data. In this context, data also means spatial data, images, and various other data. So broadly, data governance can be impacted by different data governance regimes that can enable, restrict, facilitate, control, or play more than one role.

The image is a screenshot of a video player. The video content shows a slide with the following text:

**Next Session**

- AI in Law and Justice in India focusing primarily on initiatives of Governments including that of Courts

**Thank You**

In the bottom right corner of the video frame, a man with glasses and a dark vest over a light blue shirt is speaking. The video player interface includes a top bar with 'Data and AI' and a time indicator '3:40'. The bottom bar shows playback controls and a progress bar at '37:53 / 38:45'.

So, data governance, in fact, is a very complex thing, and this complex thing will be looked at in different sessions in this course in different ways. For example, we will look at the relationship between artificial intelligence systems and facial recognition systems, how both interact, and what sort of restrictions are needed when facial recognition systems, enabled by AI, can perform many tasks. So, in the next session, we will look at the law and justice in India, focusing primarily on the initiatives of the government, including those of the courts, in the use of AI in the judicial system in India. Thank you.