Digital And The Everyday: From Codes To Cloud

Prof. Bidisha Chaudhuri
Department of Multidisciplinary
International Institute of Information Technology, Bangalore

Lecture – 09
Data-driven Identities Part 0I

So, my session today is on data driven identities. So, I will just directly get in to what the topic is about without any more assumptions. So, first part that yesterday when we are talking about I think most of the sessions, and also in the introduction that I tried to touch upon is to understand that how this digital on the everyday is a socio technical assemblage, it is neither purely technical nor purely social. So, when we talk about data do you understand how is different from information, what is the difference between information and data.

Student: Information can be expected from the data.

Information can be.

Student: Expected from the.

Expected from the data.

Student: So, it is related to information data.

Data made into information ok.

Student: Data is processed information.

Data

Student: Once information is processed.

Data.

Student: It becomes data.

So in by that logic data always precedes information. So, you see that as we you know proceed with this sort of definitions of this terms data information, you see that the vagueness of the terms, and the conflict in nature of this terms that. We use them to mean

different things, but when we actually have to define them we really do not know, what exactly the difference between the 2, or where the data ends and information starts, by that logic if data precedes information, we are not really sure at what point it seizes to be data and it becomes information.

So, now if we are looking the question that I asked you that try I think about data from like a time when there was no information technology before ICTS. So, data if I just put it in a very abstract or a very general sense is basically quantification of the world, you're trying to classify, and categorize, what you see and when to try and classify and categorize what you see there is a certain meaning that you are attaching to this classificatory schemes.

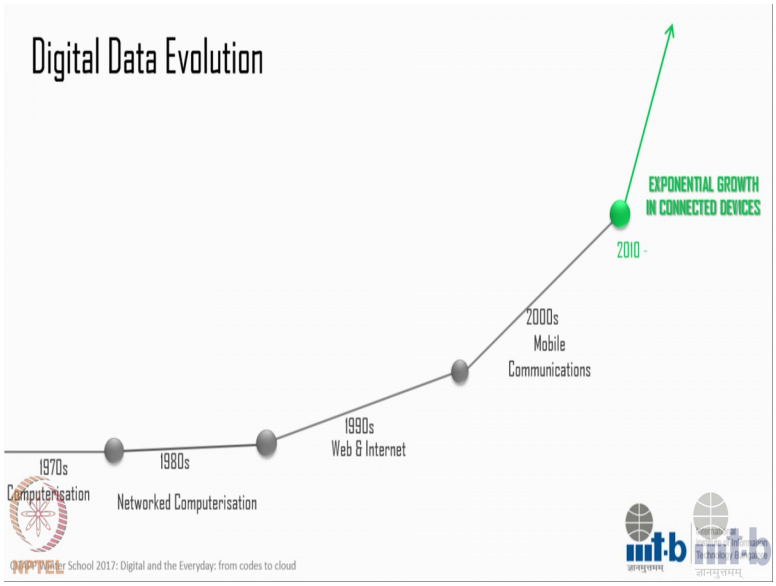
The one who is generating this data, let us say senses data. So, there are categorize on which I collect information from people. So, I ask what is your age, I ask what is your gender, I ask what is your educational qualifications, these for me are information about me. Which I am giving to someone for them to fitter into this classificatory scheme, and they have the meanings of these schemes clear in their mind, for them to convert this information into data for their tables.

So, the relationship between data and information is something which goes on in a loop. So, what is information for you might be data for someone, and what is data for someone might be information for you right, are you with me till here? So, now, if we are talking about data is quantification in a very broader way just feel free to stop me at any point, if you have any doubt. So, if you are talking about data as quantification of the world, you are trying to make sense of the world through some quantitative or very clear categories. Now does it mean that it is a very purely technical thing that when an information. So now, we are moving from information towards talking about data because that is the theme of our session.

So, if you are talking about data in a quantified or in categorized classified manner does it then become so, when we move from information to data does it, then become a very technical thing. So, then try and think about this access that we were talking about yesterday whereas, the material is on the other side and the social on the other. So, this access that we are talking about so, we first try and see how this data is both a material as well as a social thing. So, that is what we were trying to do in this session.

So, the first segment of the talk we are trying to look at what is a material aspect of data, and we are going to do it in terms of digital data. So, we are not looking at data which is pre digital, we are trying to look at data which exist in the digital space clear. And then the next part which is the data driven identity. So, when we move on to identity that is when we look at the social implications of data or how data becomes the social thing as well.

(Refer Slide Time: 05:27)



So, just to put thing in perspective is that what I am trying to do here, is to map out the trajectory through which digital data has sort of travelled. So, these are not really exact timelines I just tried to do it in a you know 10 years sort of a format. So, it is not like they are setting stone. So, the first stage is about when the computerization is taking place. So, there is certain kind of data that you are generating in digital get space, then from there you have network computer.

So, emails and stuff like that and starting not email so much, but you know shared computer network, and the communication that exist within that and from there of course, the ware and the internet and the email so the different kinds of data that are getting generated, from that we are moving which is you know sort of the scenario that we are all used to by now, there is something called mobile communications which is basically looking at.

The data that is generated through mobile phones, communications, and then towards end, when the with the advent of smart phones what we see this extremely connected devices. So, if you add the mobile communication with internet, and the web, and everything that existed before that. So, all these things are connected now. So, the amount of data that we are going to generate can you imagine how; how much data we generate, any idea.

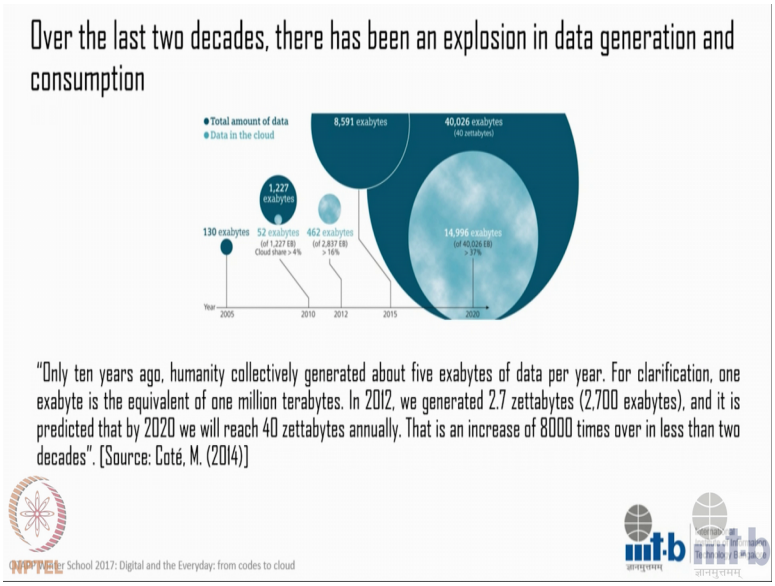
Student: (Refer Time: 07:10).

That the kind of data that exist in the digital space term.

Student: (Refer Time: 07:12).

Yeah. So, I myself I am not very like you know clear about this.

(Refer Slide Time: 07:18)



Terms like hexa byte, peta bytes, zeta bytes like who is bigger than whom kind of a thing there is a greater than relation series, there goes on write. So, this is something so, in 10 years ago humanity collectively generate about 5 hexa bytes of data per year, and you can see what is that hexa bytes. So, our files are usually you know what we see when we are operating its megabytes or kilobyte.

Student: No.

So, this is what the volume that we cannot even imagine how the volume that we are talking about here, you can see the measurement of what hexa byte basically these zeta byte. So, we are literally talking about zeta bytes now. And so by 2015 will have this much data that is a prediction that we will have, and the darker 1. So, what why suddenly he has you know such a large volume of data. Who is producing this data?

Student: Consumers yes a consumer.

That is consumers and so, what is going to happen to this all this data where do they stay.

Student: Datacenters.

Datacenters anybody else, where do they stay.

Student: Cloud.

I here very soft murmurs I need clearer voice.

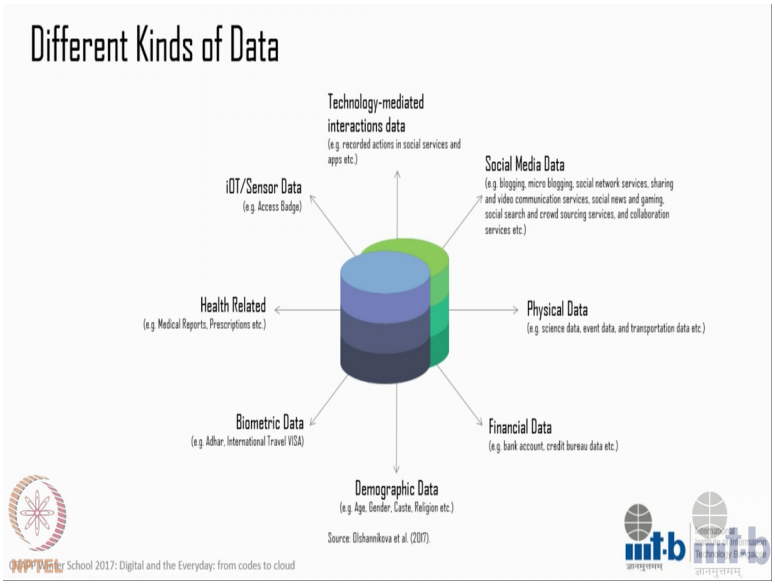
Student: Clouds.

Clouds. Where are those clouds?

Student: Servers.

So you see of course, your that with clouds we are seeing more and more percentage of the that exist in the digital spaces moving to cloud, we will come to what is cloud and how it operates in a bit, but this is the prediction by 20, 20. So, about 37 to 40 percent of our data that we generate on the digital space will be in the cloud.

(Refer Slide Time: 09:33)



Now what are these data that are giving generated what kinds of data that we are talking about right.

So, this is of course, again not exhaustive, this is just to give you a sense of the kind of data that are generated and that exist. So, first is of course, the technology mediator interaction data that we generate on our daily basis when we use an app, or a service, online, or through our mobile, then of course social media data, which is our social networking sites, blogging, cloud source platforms, all these that we do on a daily basis in the digital in any of the digital platforms. Then of course, there is this IOT and sensor data, which are this RFID codes that are you know attached. So, everybody understands. What is IOT internet of things is?

Student: Yes.

I do not need to explain no. So, all these data now this all of you know what this term the big data. So, this is what pretty much talking about when we are talking about big data, mostly the upper three sections of the data. Now of course, there is this physical data your bank account data credit bureau, then we a have biometric data which is a travel informations, or like when you do immigration cross border immigrations, or now something like an Adhaar or any other biometric information that you have given. Demographic data would be something like a sensor.

So, all these data now going back to the first slide remembers looking at the you know the trajectory. So, of course, with computerizations and stuff like that all these data has some point become digitized, those who are not like mediated through the technologies itself.

So, physical data are also now becoming digitized and they as somehow getting all connected that is what we are talking about this exponential growth, in the way we have connected devices, and because of that the volume of data that exist in the digital space now. Now in this kinds of data, there are three types in thinking about it data from a technical point of view one is the structured data semi structured data, and unstructured data anybody from CS background want to explain, what are these no.

Student: B 4 unstructured data addressing social message for like that like in various.

So, these are unstructured data.

Student: Unstructured data.

This will be the unstructured data. So, what would be the structured data?

Student: Might be like a is equal.

Proper levels.

Student: Proper.

So, where will this structured data an example here would be.

Student: That one be demographic data.

This one.

Student: Label data.

Label data ok.

Student: So, each and every religion may be asked.

Yes.

Student: So, we can place any categorize it, and some of it is like been unstructured data, we do not have labels to it. So, this is like each I would say the social media data.

Um.

Student: So, that is like.

Yes. So, this one and a semi structured.

Student: Semi structured is we are playing groups, we have some groups available some like a somebody pass through a some clustering algorithms. So, that we obtain some relation between them, that is like semi structured.

Ok.

Student: Specific IOT pre structured, or semi structured depending on when or which that is coming out.

So IOT could be both semi structure, and it is a structured data. So, people who do not come from CS background do you understand now the difference between the 3, if not please raise your hand, because it is not easy to understand.

Student: So, the essence is may be like structured data happening is cluster are labeled on like mean.

Yes.

Student: Since we have data which is important in spectrum.

It is not unorganized.

Student: Ha.

So, structure does not, unstructured does not mean it is not organized.

Student: Ok.

So, I will try and break it in to more simple words. So, structured data so when we say structure data we basically mean it as within a frame work of a computational world so, a

computational world would be what codes, programming languages, and the algorithms. So, with all these what happens is structure data is only machine readable.

Student: Right.

So, if you look at this data without knowing the codes, you will be not be able to or the knowing the language, you will not be able to understand. So, only machine readable data from the computational point of view is called structured data. Now semi structured data is something which is encoded in programming languages in a way that it with some markers, and tags that it can be red by both the man and the machine.

So, this markup languages they have you would.

Student: (Refer Time: 11:02).

Extend.

Student: XML.

Extendable.

Student: HTML.

Something of that sort yeah.

Student: X.

Yeah. So, this markup language is a HTML would be one example of that then there is this thing is XTML.

Student: XML

XML Right.

Student: XML.

Yes. So, all this languages they are basically these are markup languages, which basically make it this your web data the digital data they encode it in such a set a way that it is readable by the man, and the machine in the same way. Now the unstructured data that is where your social media data, that is where you are all this app that you use, or the

games that you play stuff like that. So, things that we do, on a daily basis on our let us say on a smart phones are unstructured data, because for it to be machine readable it needs to be worked upon. The machine cannot immediately read it as of now.

Student: So, it is basically like the code defined from the point of view of the machine.

It is not that is what I said it is not unorganized.

Student: Yeah

So, of course when you are doing all those things there is a logic the way it will be structured, but in the computational world this would be something like a raw data that comes, which needs to be worked upon clear.

Student: Mam, biometrics if even man can also get like human can gets 1.

Which one.

Student: Biometric.

No you cannot.

Student: I am saying that.

Yes. So, no the point is that what is biometric, it is basically takes your the biological data from your body, and then encodes into some informational system.

Student: Right.

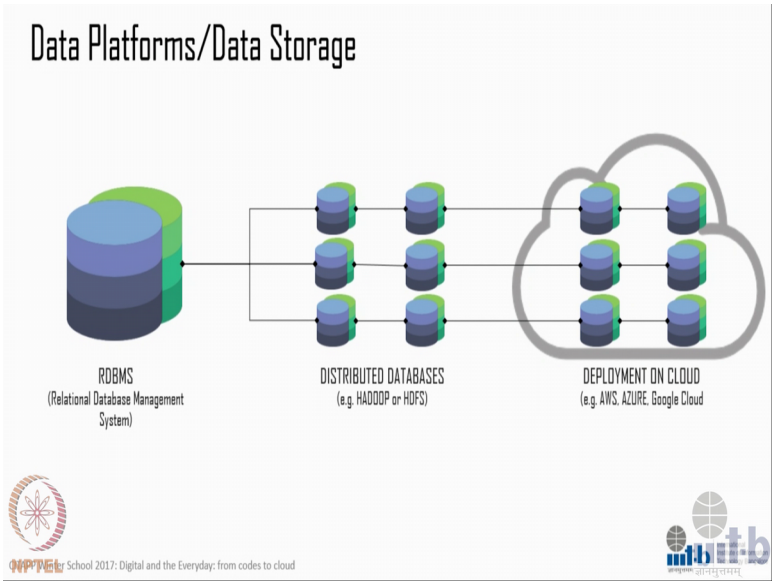
How you do that encoding. So, it is not immediately readable by a man that is that is why it is called the biometric data, there is a your that is why the metric comes from see you take bodily data, and the moment you say it is data its quantified. So, I am quantifying the features of your body, in a digital space how you quantify in bits, it is a combination of 0 and 1 am I correct.

Student: Yes.

So, that is what it. So, that is what pretty much simply it mean for interdisciplinary crowd that how what are the different types of data that exist, the semi structured the

structured, and the unstructured. Now what happens, we have now why we are talking about all these things, why we are that that this kind of data that we are talking about. So, what has changed for us to talk about, to be able to talk about all this things in the same environment?

(Refer Slide Time: 18:15)



Everybody understands what is relational database?

Student: No.

No. Anybody form CS want to take a attempt.

Student: Any one.

Relational database look any anybody who has a background.

Student: I am a computer science course.

Yes you want to explain what is relational database?

Student: There are actually 3 data model sale hierarchical relational network hierarchy relational.

Um

Student: Where is hierarchical next is network there is a relational.

Ok.

Student: Relation is stable relational that is; obviously, it is in terms of table it has got field does not this like a field plot derivative field is also found as the column. Column or field row or column, and field is also called an attribute row a collection of tables constitute a relational database model.

And, why do we what is the value of relational database?

Student: What do you mean by relation data mean?

Value in the sense that why would you use what is the.

Student: Importance.

Importance that use value how you use it.

Student: We easily reorganized as and when we.

Student: Keys are there various keys are there.

So, you can run.

Student: Reorganize the data as, and when we need it the way we need it.

So everybody understands what is relational database, if not please raise your hand, if you need more explanation its fine.

Student: Yeah we need explanation.

You need more explanation.

Student: Mam both places things like schematic database not there, but still not our peoples little more in compared to a schematic other databases.

Yes. So, I agree with you I will come to that in a bit. So, a relational database is something let us say let me try and put it in a very simple example, let us see you have lots of your exams are over, you have the whole semester full of notes, which are in physical form somewhere like papers that you taking notes like all of you are taking notes now. And you have certain ways to mark that which class, or which course, this

notes belongs to and your trying to organize at the end of the semester. So, you got different file boxes, and you want to put this paper in to some categories to organize.

Now, in when you try and do that so, you have created this levels so, this is course one, this is course two, this is course three. Now, when you try an organize all your scattered notes, and papers according to this boxes, you realize that you cannot make a decision which box this particular piece of paper should go to you feel like, this can go to box 1, but it can also go to box 2.

Now what you do? So, you create a marker that is what he was talking about that you create attributes to which you can say that I am putting it here, but if need I can also use it for 2. So, you can re organize your database in a way that depending on your need, you can run a query, and get these data in the particular function that you need, the what task that you need to perform on the basis of that you can recall these data, or the recall this papers does that.

Student: Does it yeah. So, some essence that the like, he was mentioning at the fields performance.

Yes.

Student: Does it predefined.

Exactly.

Student: Mechanism or schema whatever it might be called in , but let us say for example, it is a rule on big data cluster there is.

Yeah. So, that is will come to I just yes, we will come to that. So, I just for people who do not have a background on all this terms, I just wanted to make it clear that what do we mean by when we are talking about relational database. So, now going to.

Student: Sorry.

Yes.

Student: Let me let me speak here of an animals. So, let us say walk in to a super market. So, you have all the items are labeled numbered, and products based on each category of

it what happened is called a clear of a spoon or whatever, then you walk into a warehouse you will still have may be an RFID on each or some building or number.

Yes.

Student: But to an average common person in a large warehouse, it is the same maintenance talking you will have no idea.

Yes.

Student: But they because of using the RFID, and some of the labels they have for each say common ID, I have may.

Do not look at them.

Student: Many make sense of it.

Yes.

Student: And in organizer now.

Yes.

Student: Next to a set of books, there may be a you know a code of book.

Exactly.

Student: But this is our horizontal.

Hm.

Student: So, workship makes no sense, but then it is all software.

Yes.

Student: So, in a supermarket I will is enclosed example through an RDMS. A large retailer warehouse Amazon warehouse or where it is massive, but it is completely it looks among next. It is probably more of a big data.

Yeah.

Student: Analogy may be.

Yeah, no I think it is a good analogy. So, relational database is something which made this knows the relationships possible between different categories of data. So, you not categorize them as some predetermined levels, but you also create relationship between these levels, because these levels are not always you realize when you try to use this data this categories are not water tight compartments.

So, there overlaps between these categories and in order for this overlaps to make sense for you to be able to use the data that you have gathered. So, that you create relationship between these categories. So, that is in very simplistic term am I sorry, I am like you know putting it in a very simplistic way, if there is a lot of complicated mechanisms are go behind organizing this relational database, but that is what it means, like that is what it does.

Now, what happens as again I am going back to the previous slides, now when you are creating this you know, all these data, and the hex bytes the volume that we are talking about, what happens a relational database is not able efficiently able to integrate all this data. So, then we are moving to something called distributed databases. So, you are pretty much saying rather than keeping it in one you know table, you are making breaking the data into small parts and distributing it, and that is something has anybody heard Google file system.

Student: Yes mam.

So, that is of the ways examples of distributed databases, yes.

Student: In spite we tend for big mix data. So, does it mean that hadoop should remain in its distributed or database only.

So, hadoop can be both hadoop can be in the distributed, I will come to that just give stay with me, we will explain this whole slide is about that. So, when you are doing this. So, you are basically so, Google file system is one way of doing that. So, you try and integrate all that data that is being generated, let us say on a Google search page you are trying to distribute it in a way. So, that you integrate you can basically what the relational database was trying to see that create this relations, and understand this relations in a

more effective way, you could not do it with this management system anymore so, you are moving toward distributed system.

Now, Google file system is one example which is proprietary, hadoop is something which is open source platform. So, other people can come to hadoop and tweak it that is what a platform would mean, that tweak it based on their need. So, now, hadoop or something like a distributed data base system might exist in a cluster in a center sort of a environment, it can also exist in a cloud. So, what then happens that, sorry how what does it mean when you move from here to here. So, if you do it in a sort of clustered way, and if you move it to a cloud, why do we need to move it to a cloud? Hadoop is a open source platform, which basically then host all these data. So, for example, yahoo and it is distributed so, it has multiple nodes in multiple places.

Student: Ok.

So, like face book use hadoop, yahoo use hadoop, the many other you know big place use hadoop. So, why do we need to move to the cloud? What do we gain?

Student: It can be accessed can be accessed to any limit.

That is also distributed. So, distributed data base management is the key here. So, that can be done in a clustered environment, and in a cloud environment both. And in both the environment it can be accessed sorry.

Student: What is the difference of what is the exactly if the distributed database.

So, you are basically now breaking the data.

Student: Yeah.

Into smaller chunks, and storing it in different places, you do not have a centralized database system any more.

Student: Even that cloud also.

So, this is the logic in which the database is organized the distributed database, it can be hosted in a cluster environment, it can be hosted in a cloud environment.

Student: If we can show some example.

Yes.

Student: Where distributed database is compulsory that is more, that one simplified in understanding, and in but for distributed database, the data cannot be stored any such examples.

Yes I will come to that. So, the distributed the difference between the relational to the distributed, is that there is no logical difference between storing and computing any more. So, storing and computing this two functions are sort of merged, does it make sense no. So, here you store the data that is the data base, that you are storing it, and when you have run a query that is when you compute the data. Here this is possible simultaneously. So, the storing and computing can go on together simultaneously. No. Ask me. What does not make sense where is it unclear?

Student: non travelling complications all just.

Yes it is not, what you can. What you do? It is about it enables you to do.

Student: Basically it can be any smooth teaching of a process.

Yes of course, so this is a evolution. So, we move from RDMS to a distributed.

Student: I think from a function from like, you know using data aspiration aspiration.

Yes if that simplifies what you mean of course, that would be something yes.

Student: Then how difficult to retrieve it from RDMS.

It is not retrieval. So, we have moved on form that to this that is the analogy he was giving that in a super market, and the warehouse. So, as you increase the volume, the super market is much contains space; you could operate with that RDMS, because the table, that you have the columns that you have created in your table just think about the super market in a table form. So, there are finite numbers of table columns that you have created, and so the there are finite numbers of relations that you can create between these columns.

Now, when you go to a warehouse, which is ten times bigger than that the volume has increased exponentially so, you cannot operate with that it is you can, but it will be very

inefficient way of doing it. So, it is the volume of the data, and your need to relate the data to integrate and aggregate different data columns that makes move from RDMS to distributed, does that make sense. So, what happens? So, first there is a storage thing that you are solving by moving that in now you have broken it down so, you can store a large volume. Functionally speaking what he mention, now you can also not only store, but compute at the same time in this database system.

Now, this cannot happen in a structured in a clustered environment, this the same logic can run on a cloud environment. So, who are the big players in cloud like Amazon web services? So, whatever you do, all that data that you are doing generating through using your smart phones, so these are biggest, players Amazon web services, AZURE, and Google cloud. So, they basically hold like three of them together store 100 percent of the data that the volume that we saw that 40 percent of the data that is generated recede in the cloud. So, that 40 percent is this 3. Shall we move on?

Student: What is the difference between the cluster environment, and the cloud environment?

So, cluster environment means that I can have the distributed database, but all of it can reside in let us say electronic city.

Student: Means what?

So, there is a still a physical angle to it yes, in the cloud you do not need that. So, what is happening, the it is becoming more and more virtual, when you are moving to a cloud platform yes.

Student: To understand it what is politics of it like?

That we will come to later first we are focusing on the material aspect of it how does it work, because not all of us in the room understand how it exactly it works right. So, let us get a hang on how technically it works, will come to the politics of it that is the next part of our presentation.