Anonymisation: theory and practice

Hi! My name is Mark Elliott. I'm from Manchester University. I'm part of the National Centre for Research Methods (NCRM) and I lead the UK anonymisation network.

Today I'm going to talk to you about some basic concepts which make up the idea of anonymization. Ok so what is an anonymisation? Anonymisation is a process by which personal data are rendered non-personal. There's a very precise definition it ties it into the legal notion of personal data so we need to know what personal data is. There's definition of personal data within the data protection act so its data that'll relate to a living individual who can be identified either from those data or most data in other information. This definition is quite precise there is some variance between the definition in different pieces of legislation and across different jurisdictions but the notion of identifiability remains central and so that's the notion we're going to focus on when we're thinking about an organization.

Now I'm going to talk briefly about some terms and first pair of terms i want to talk about is anonymisation and de-identification both of which you may have heard of. Now they do with different parts of the data protection acts definition of personal data. De- identification tackles directly from those data i.e. it's about preventing somebody being able to recognize somebody directly from the data so for example from their name and address. Anonymisation on the other hand tackles the other part of the definition indirectly from those data and other information. It's a more complex idea and anonymisation is a deeper concept. De-identification is actually basically quite simple you simply remove or obscure in some way those direct identifiers. With anonymisation the issue is being able to decide where the set of indirect identifiers are sufficient to enable somebody to be identified. Ok so four uses of the term anonymisation that you may have heard of. The first is absolute anonymisation which essentially means that the zero risk of every identification under any circumstances. Those are the particularly useful use of the term because in order to reach the state of zero risk real identification you effectively have data which is itself of no use. It's very limited for little value. Formal anonymisation is the other side of the coin. It's just the identification so stripping away of those direct identifiers or possibly replacing them by pseudonyms and this is not sufficient. Because there remains the possibility that within the data there are indirect identifiers which enable somebody who wishes to re-identify.

Statistical anonymisation attempts to measure the risk of an identification happening and to control that risk. So here we're in this middle ground now between these two extremes of absolute and formal anonymisation and here we are allowing the possibility that re identification could occur and we're measuring the risk of it. We're not insisting that our data are absolutely anonymized but we're trying to do something in terms of reducing the risk. Now the disadvantage of statistical anonymisation is it tends to be very focused on the properties of the data. The fourth category functional anonymisation acknowledges the value of the statistical approach but also takes into account the environment in which the data exists. We're now going to go on talking more detail about this more holistic approach to anonymisation.

Ok here are some principles: the assertion is that anonymization is not primarily about the data. Anonymisation is about what we call data situations. The data situations arise from data interacting with data environments. So now we're introducing a new term the data environment. So is there a formal definition of the term data environment. So it's a set of formal and informal structures processes and mechanisms agents that either act on data

providing interpretable context for data or define control and interact with those data. What does this mean in practice? Data environments consists of agents normally people infrastructure particularly security infrastructure government's processes and most particularly other data. Data environments tend to be layered so they have data environments within data environments and partition and the security infrastructure that exists for example on a server partitions the data on that server from data in the outside world and we can think about that server existing inside that bigger data environment which is the global data environment which contains all the possible data in the world.

So data environments are complex items. In order to understand whether data is sufficiently anonymized for meeting or legal requirements and other ethical constraints you might be under you need to understand that it's about the relationship between the data that you have and the data environment that you're considering putting that data in. So essentially cannot decide whether data are safe to share or release or not by looking the data alone. You have to consider the whole of the data situation.

So anonymization is a process the goal of which is to produce safe data. Now you could get overly focused on that. It'll only make sense to even be doing this if what you're trying to do is to produce useful data and that was the point that i made right in the beginning of those for definitions about the problem with the absolute anonymization definition. We're trying to produce useful data so what we will be doing is balancing that notion of utility of useful data with the idea of trying to produce safe data.

Now zero risk is not a realistic possibility if you're going to produce useful data because that was my criticism of the absolute anonymization definition. The measures then you put in place to manage the risk should be proportionate to that risk and it's likely impact. So you're going to do something to the data either to its environment in terms of where its place and how you're going to keep it and what government you have around it or you going to manipulate the data to itself in order to reduce those risk the risk of real identification. So here we have defined the basic concepts that you need to understand anonymisation.

On the next video the anonymization decision-making framework, we describe a practical approach for carrying that out. Thank you