[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: So where we left it last time is that we found some theoretical justification for why people might be able to tell you who the influential people in a network are, even if they have no idea really of the shape of the network. Simply because it happens to be that there is a natural link between how many times we hear about someone and the ability from any message originating from that person to reach a lot of people. So although it's not the same object, they are related and potentially quite closely related if they are-- once we are starting to look at processes that continue for a long time.

So the faster, now, let's put that in a context of thinking about immunization. So that was in fact only just the second time we tested it. We tested this idea for immunization was so in a large scale project undertaken in collaboration with the government of Haryana in India. And the objective was to increase for immunization in a context where immunization rate was low and the demand was quite good. The supply was quite good. They invested in a lot of infrastructure to make it easy.

Going back to the point that you made about Chile, to make it easy for people to go. And yet, it got stuck at full immunization rate in the high 40s according to the report by the parents, which are probably overestimated anyway.

So what we had in our hands was this thing that maybe gossips would work, as well as the previous experiment that I already showed you on the impact of small incentives on immunization. And so we decided that we could put the two things together, as well as a third intervention, which is very frequent.

And probably even for your own health, you get reminders of going to the doctor. If you've gone once to the eye doctor, you will get a reminder once a year that it is time to go to the eye doctor. So SMS reminder have also become quite popular.

So we wanted to try the three things. And to do that, we first had to set up an MIS. They didn't have a way of collecting immunization in a reliable way. It was done in registries, paper registries that were not collected. So the first thing we did is to work with a group here at MIT to set up a small mHealth platform. I'm saying small in that it was very barebones. It was just for immunization.

Incidentally, I thought this would be really easy. And for whatever reason, it's not. It took them a long time. And it wasn't working fast. And so it took months to develop that thing, which is really basic. It's like an Android form. There is a registration aspect. And then-- for a new kid. And then, the follow up as kids come up.

For whatever reason, it's harder than you'd think. I still don't understand why. But it's worth keeping that in mind. If you want to get into any things like that involving software, it's always harder than you think.

Maybe, also we wanted to work at scale. So we worked in seven districts, a little over 2,000 villages. In total, we had 300,000 unique children in our system and almost 500,000 vaccines. What I'm going to present today, it's all based on the administrative data from that system that we owned.

So one issue is whether that system, that data is reliable, or whether people lie about for example, whether kids that have never been immunized that are registered as being immunized. That's particularly relevant for the incentive. But because you could imagine a nurse deciding that they are going to immunize a lot of children, giving their phone number, or their cousin's phone number, or everyone, their friend's phone number, and in this way collecting the small incentives, which were air time on the phone.

So what we did there is we sent people to the field to collect-- to contact all of the families that had-- not all. Some random sample of the families that had been registered in the system. Find them, checked they existed, and then if they existed, whether they remembered going to the camp, and the kind of vaccinations they got. And the good news is that we could find everyone.

Unlike in similar exercises that have been performed from the MNREGA program or welfare program where you go to the village to look for the beneficiaries of the welfare program, and you find half of them. So we find all of them. And then, when we find them, they at least remember going to the camp. Although sometimes, they don't remember which shot they got. But then, it's not clear who is wrong at this point.

So our take away from that is by and large, the data is-- the administrative data is reliable enough. And that's what we are going to use as our measure. Yeah?

AUDIENCE: Question. When you're running an experiment with multiple treatment ops, how do you think about how many ops to have and--

ESTHER DUFLO: It's an excellent question. And I thank you very much for asking it to me. And that's basically what I'm going to talk about in five minutes. In fact, I hope you're going to read that paper one day. This is the paper. I should have put on the syllabus for you to read now. It's asking exactly that question. Instead I put the "Gossip" paper, which is fine, too. And then, it's-- but which highlights just one part of that. And not this question about-- there are many treatments there.

But let me come back to that in a minute, just close the "Gossip" part that we are done with that. So in the "Gossip" part-- so in our first gossip experiment, which in the paper you read comes second. But chronologically it was first. We just had this question of if I want to see people for fair, who should I go see? So the idea was to capture something, which is not loaded and is all about diffusion of information. And then, in that first experiment, the experiment was a bit artificial. We were trying to transfer information about a cell phone raffle.

But here, when we are now in a real world setting and in particular a health setting, there might be-- there is this trust issue, which is the people whom you might trust to tell you when the next affair is or when the next cell phone raffle is might not be the one you're going to talk to about health. And although we have a theory for saying people might be good at knowing who is good at spreading information, we don't have a similar theory to tell us whether people are good at spreading information that people will believe, that people trust.

So what we decided to do is to just test it. So we asked. People either were randomly at random seeds. Someone asked, by the way, in the paper should you have had random seeds? So we had random seeds.

Or seeds that were where we elicited gossips, where we listed the trusted people or trusted gossips. So how do we elicit any of these people? We go and find out. Talk to 15 people, and we ask them.

So for gossip, that's the question. If they share information about the music festival, street play, fair, many people would learn about it. That's because-- and then we have the theory for how, why. Could you name these people?

Then for trusted people, it's who are the people in this village that you and many villagers trust within and outside this neighborhood? That means when I give advice on something, many people believe that it's correct. And they tend to follow it.

So, of course, choosing the right fertilizer for the crop, and keeping your child healthy. So we specifically added the health example because we were going to transmit information about health.

And then finally, trusted gossip. It's a mouthful. But many people get to know about a piece of information. And among those people, who are trusted? So you need to first be a gossip, and then also be trusted.

And then, the intervention was once these people were named, so we found this result that very few people are named. And there is a lot of consensus of who they might be. We pick the five-- the common top of the 15 out of all the domination from the 15 random people we interviewed. And we go for that.

So someone was asking also whether it's pretty expensive to go visit 15 people. But it's not really expensive to go visit 15 people, especially if we don't really care who they are. Because it's not that we are-- under this theory, it's anybody would be fine. It doesn't need to be random, random. It could be just anyone. So it goes quite quickly to go visit 15 people.

That said, we are now interested in the scale up of this program, which I'm short cutting, but which turned out to be effective and cost-effective. And now, we actually exactly stumbled upon this issue, which is for us, it's pretty easy to go see 15 people, and record the answer, and be done in our gossips. But for the government if you want to do that for every single village, it becomes a thing.

So who is going to do it? And how they are going to do it? And are they going-- if we ask the frontline health worker to do it the ashes, are they just going to make up the names? And so basically, it's a huge stumbling block.

So in fact, when I read your question I was like, well, it's not that expensive. But then I say, well, actually, it is logistically a nightmare. And so what we are trying to do in the scale up of that is we are not doing that anymore. We are asking the parents when they-- since now they have a little better mHealth intervention, we are asking the parents to tell us. So that's what we were hoping to do is to using-- to add a module to their registration camp when people come to say, hey, by the way, do you know someone who is the gossip?

And so we thought that would be very smart. But that didn't-- so that was the idea. That it's going to be very cheap. It would in fact be very cheap, but it involves modifying the software, which turns out to be a huge nightmare. And getting the nurse to use the software and to use that modified version. So we've also been stuck on that for some month.

So the current version-- so this was version 2. So the current version is that we have a large database of phone numbers because there is a huge health insurance programs where many people are in need, and they have phone numbers. And these health insurance people have a call center. They call them for bias. People call them to ask for insurance stuff, and they call people in turn.

And these people in the call center can do a survey, can do a short survey to a few people in each village on the phone. So that's where we are now. That's what seems feasible. Stay tuned, I tell you because we haven't yet tried and therefore, failed. But that might be an issue. Anyway so that seems simple. But it's not that simple.

But once we find them, then we go in and visit them, try to enroll them in the program, and we tell them that we will send them these messages. So we have incentive and non-incentive villages. We saw that the incentive villages, they were well-- they would be well-placed to mention that there is an incentive program. That's a concrete piece of information people probably don't know because the advertising of the incentive program is not huge. And then, in no incentive villages it's just a general immunization is good for you, go ahead and do it.

So here is the basic results. I'll tell you about more results when I get to Kayla's questions in more details. But the basic result is that if we look at measles that's going to be my main outcomes because that's the last vaccination in the sequence. That's the log number of people who receive the measles vaccine in any monthly camp in a village. So here, it's in logs. Later, I'm going to show you results in levels. But for some reason in these people, we show them in logs. And it's 18% higher in places in a village that have a gossip.

These two, the trusted and trusted gossip are not-- neither of them is actually significant. They're also not significantly different. So I can't say that it's worth. But what is relevant is that it's no better. And in particular, you might think that the trusted gossip would be obviously better because it's both a gossip and it's trusted, but of course, when you're trying to find a trusted gossip, you lose some on the gossip dimension.

So if in fact, it turns out that you don't get the most gossipy gossip potentially because they are not necessarily the most trusted. So if in fact, the trust dimension is not that important compared to repeating the thing, then you might lose something. So we can say that they look-- they all look fairly similar.

And then, these are the incentive results. But I'll come to that in a second.

So putting it all together. And so that's the other paper written on the same experiment. So the first one really just was more for use this experiment just for the gossip aspect. But then, we exactly have-- Kyla asked this question-- is there are a lot of treatment here. And why are there a lot of treatment? There are a lot of treatment because we are not, in this case, we didn't go to this project for testing a particular theory.

Sometimes, you go on and you have something very simple in mind, very clear in mind. You want to test a particular theory, or you want to test whether intervention A works. But our mandate here with the gun was a bit different, which is that can you come up with the best possible intervention for us in terms of increasing immunization rates based only on the men intervention? At least they are taking care of the supply. So we had the supply sorted out.

But based on what you know on the literature, what's the best we can do both for increasing immunization overall, and also in terms of cost effectiveness and cost per dollar? And when we started brainstorming with them, the uncaring thing on social-- the bracelet was not there yet. So we didn't have that part of it. But what we had was the idea of cash incentive, which we had found to be effective in Udaipur. The reminders, which everybody is using, which has been endorsed by the Indian Academy of Pediatrics, and this new idea of gossip.

And then, so there is evidence from the rest of the literature that each of these strategy may improve take up. But we don't know which one is the most effective, which one is the most cost effective, how they should be used. For example, in Udaipur we tried one type of incentive, lentils, and then an increase, which was slopy and the equivalent of maybe about $1 a shot into this term, in PPP term.

Maybe we could get less. Maybe you could just use less money. Maybe it doesn't need to be slopy. With SMS, do you need to SMS everyone? Or you SMS some people with the gossip. As you know, we didn't know if we should use gossips, or trusted people, or trusted gossips. So they are different types of variants. And then you start asking yourself whether the combination of these variants would work well.

So for example, the gossip might work better when there is incentive as well because there is a clear message to convey than not. So at least this one was potentially a clear hypothesis. But it's a mess. There is a jumble of things. And so here, it comes to Kyla's question is, how do you proceed? So in the literature, the way you typically get to that, and that's the best way we have most of the time, is you are collecting estimates from a lot of different papers and putting them on a common scale.

So that's really the Campbell Collaborative and the Cochrane Review is what they do for health intervention, and now also for some social science intervention. They're basically saying, on vitamin A, we have 20 studies. And we can put them on a scale. In fact, once you have that, you can try to do meta analysis like the ones that we found for the-- we discussed for the [INAUDIBLE] paper.

J-PAL is doing that, trying to accumulate the impact of different intervention on something, and then put them on a common scale to find out what seems to be most cost effective to increase test scores, or attendance, or anything you're interested in. The problem, of course, is the population intervention vary considerably across studies. So it's comparable or not. When you test different intervention in different contexts, you can't get to the interaction, which makes learning-- if you want to answer a question in particular, in this case with high error, which is the view was that whatever you tell us, we're going to try and scale it up.

The idea in experiment there is not one experiment with one treatment. It's to run a single, large RCT, which has everything you can think of.

The problem is that even though the experiment, this experiment was pretty large-- so it's 2,000 villages. But in fact, 900 villages where we did the census, where we could do the gossips. But it's not infinite. And the number of possible combinations is large. So what do you do?

So the first thing you could do is to say, well, it's too large. That's going to be too many things to look at. You're never going to be power to look at every single cell of interaction. So run very few treatments. There is a paper by Karthik Muralidharan and others that point out that once you have interaction, if you have, say, two treatments and their interaction, it gets messy to ignore the interaction, which is what a lot of people do. And then, therefore, it's confusing.

So the McKenzie solution is just don't do it. Run one treatment. That's nice. But that means, which one? How will you pick which one? Or you could include all combos, but potentially lose power because you have a lot of different cells. And the cells can become quite small. I'll tell you exactly how small in our case, for example, when I get to that.

So in practice, what people do is sometimes run many experiments, and then pull exposed if it looks like the interactions are not significant. But then that poses a problem of statistics, which is if you first run all interactions then remove them when they are not significant, then it's not clear how to read the test statistics because it might just have been a fluke that it was or wasn't significant. So we can't interpret these statistics unless we go with a clear frame in mind.

So in the literature, there is a lot of push towards going simple and pre-analysis plan. And therefore, going towards one or saying in advance exactly what you're going to run. The issue is that it's assuming the conclusion. So this is an image from *The Matrix.* Well, The Oracle, so this is The Oracle problem, The Oracle point. It's that if you know what it is that you want to test, then first of all, there is a question of whether you should test it anyway.

But you run the experiment to justify your prior. So you didn't come here to make the choice. You've already made it. You have to understand why you made it. So that's the issue with Oracle. And in a lot of cases, for example, in this case, we didn't have an Oracle. So that's the problem. So that's the tension that Kyla's question was asking.

And then there is another problem, which is when you-- even if you had a large, large, large, large, large, large set of sample size, so you could run many, many things individually, you can think of them as all of the combos, and variants, and variant combos, or variants as individual treatment. And then, you tell the government this is the best policy. You go for it. This is the effect you're going to get for it.

OK. Now, that's nice, except that it will make the policy is chosen to be the best in that sample. So by definition, it performed pretty well in that sample. And it could be a fluke. It's like the best basketball player. If you rank people by basketball player, the best basketball player is someone who's on the day you were ranking has scored many, many, many goals. But it might not be that much best overall. It's just that day he performed so well that it looks absolutely fabulous. That's not his-- you might be very disappointed once you get him to your team that on most days is not as fabulous.

Similarly, for our policy, if you picked it among a bigger range of policy to be the best, it will look really great. By definition, it performs really well in your small sample. The question is, you're giving it credit for the epsilon. And so that's not a conceptually difficult problem. There's a very nice paper by Isaiah and co-authors that looks at that. And basically, this just can be corrected. And so this is-- they have a paper called "Inference on Winners."

Now, the problem is the more policies you are trying to test against, and the more similar they are to each other, the more you're going to be penalized by the inference on winners. So naturally, because to be the best you'd have to be particularly lucky.

So that means that that's one more reason to not have a lot and lot and lot of treatments in principle, but then, we still would sometimes would like to have a lot, a lot of treatments because we really don't know. And we would like to run one experiment and have a coherent answer that doesn't assume the solution in advance.

AUDIENCE: Esther?

ESTHER DUFLO: Yeah?

AUDIENCE: So I'm curious. This is random, right? It's if I happen to run?

ESTHER DUFLO: Yeah. So it's just you happen to-- in that sample, you happen to do quite well. So just imagine, for example, a setting where in fact, all of the policies have the same effect. You're still going to rank them according to the epsilon. So that's what we have to get rid of.

AUDIENCE: So this is not like the scale up problem?

ESTHER DUFLO: That's not the scale up problem. It's just I have tested a lot of policy. I'm picking the best. And therefore, and it is a standard thing. That's exactly what we're going to do here. That's exactly what we want to do here. And that's a common way to think about policy problems. What's the best? And then, the best is also the one who performed the best in that sample. And hence, that has benefited from the most fluke by construction.

So this is correctable. This is pretty easy to correct, actually. There is nothing-- it's just very classical statistics in there. The problem is the correction is going to penalize you for having a lot of options you're trying against each other, as well as for having options that are in fact, similar. Yep?

AUDIENCE: Yeah, I was just thinking that-- well, it wouldn't be the same because according to the performance against people control, that between each other because I don't know if I can see that two interventions are one is--

ESTHER DUFLO: Yeah, so the control here doesn't-- the control here is the same for everyone. So when you're going against-- if you do the treatment effect, you're going to do the treatment effect against the control. And you're saying intervention A are the biggest treatment effect, B as the second biggest treatment effect, et cetera. They're all ranked against the same control.

AUDIENCE: Yeah, but then, for example, the first, the best run of the second run between each other, not be-- not statistically. If I'm comparing those two, they could be-- I couldn't rule out that one is better than the other.

ESTHER DUFLO: So that's possible. But it's in that case, you don't know which is the best. But even if they are, it might be because it is in part because your first one happened to do well in the sample compared to the second one.

AUDIENCE: If they are?

ESTHER DUFLO: Suppose they are statistically different. It could still be in part because the first one happened to have a very good draw of the epsilons in the treatment, and the two didn't. So that's the problem. It's not the problem that you might think they are different, but in fact they are not statistically different, which is a good point. But even if they are, it could be just for a fluke performance in that sample.

So it's not conceptually difficult. But it's not difficult to deal with at all, even in its data. But it's just that how much you penalize yourself. It depends on how many policies you run against each other, which makes sense because you have more ignorance. So that's the tension. That's why your question needed more than a 1 minute answer. It's what do you do?

So what we did is that we tried. Basically, we tried to navigate getting the best of both worlds by saying, look, most policies don't work anyways. And most policy variation might not matter. So for example, dosage. Dosage might not be that relevant in the gossip world. The fact that the gossip is trusted, or trusted gossip might be similar. We don't know whether it's similar. You say maybe it is.

So maybe a lot of my possible treatments can be eliminated from my consideration set. And therefore, I can run many, many treatments and get rid of anything that doesn't work at all. And then, pull together what is similar. And then, be left only with intervention that seems promising at the end of the day and estimate my regression in that set of promising policies.

So what do I mean by pooling? We're only going to pool-- and that's why cannot pool in some setting and not get the full play of the problem that Karthik Muralidharan identified, which is if I pool ex-post when the interaction is negative, that's a little-- that's fishy.

We are only going to pool-- or we are only going to ask the data whether they want to pool things that are dosages of the same or variance of the same treatment. So I'm putting in some theory. It's not really theory, some amount of structure there to say, well, any incentives-- I want to try a different incentive because who? Knows they might differ. But I'm going to allow the data to pool them if it turns out that all the effects are the same.

So example, high and low incentive will be allowed to pool. Slope and flat, and incentive will be allowed to pool. Slope and flat, and high and low will be allowed to pool. And so that creates a treatment profile of any incentives. And similarly, any gossips. And then, any reminder. A lot of reminders versus fewer reminders.

So I'm telling the data that they can pull. I'm going to permit the data they can pool that way. And then, you run a LASSO on that. You run into a small problem, which is that when you run a LASSO on that, by definition, you have a lot of correlation between. You're going to structure. So think of writing the regression in order to-- allowing pooling instead of have now run all of your different possible treatments and interactions. Instead of running one regression for each possible treatment, you run a regression of the form, any incentive, and then marginal effect of high incentive.

So then that might be-- that might become zero, as opposed to high incentive, low incentive, and then you test the difference. Why it's relevant to write it this way? It's because you are going to run a LASSO on this regression. And LASSO is going to penalize you for having a lot of terms. And it's going to just ask you to cull a bunch of terms saying they are completely irrelevant. They shouldn't even be there.

So if you had two policies, high and low incentive, that in fact had the same effect, LASSO would want them both if they were both effective. If you write it as any incentive and an incentive-- the extra effect of high versus low, then LASSO is going to get rid of that. So it will allow get rid of it.

But there is a problem with doing that, which is as soon as you do that, you create a lot of correlation between the variable because-- and by definition, any incentive and incentive high are very correlated because incentive high will have a one every time. Incentive will have a one every time. Incentive one has high as the one. And in addition, it will have some zeros when the incentive is, in fact low.

So by construction, you're going to have a lot of collision, which is precisely when you cannot run LASSO. Intuitively, it's that the intuition for it being that LASSO is not going to know what to pick if the two things are correlated. Statistically, it fails. Irrepresentability from TVA, which is saying you cannot-- the irrepresentability is that you can only run Lasso if no variable can be approximated very well by a combination of other ones.

So but there is a solution for this problem, which is a preconditioning of the variable that basically organizes them. It penalizes you for doing that, but it works.

And then, once you've done that, you run your LASSO. The LASSO is going to select some variable. It's going to kick out everything that is ineffective. And it's going to also kick out things that are marginal difference that are not pertinent. And once you have that, you have very few variables left. And you can do your post-processing. That is running, which is basically running your less regression happily on a very small set of variables.

And now, you can run your Andrews correction without being penalized by the fact that there was a lot of variables in the first place. And you go ahead. So that strategy works. And it's nicely compatible.

You adjust-- you don't have a flu case that's too high because you've-- with Andrews, it's if you wanted to really say, this is going back to Salu's point, which is suppose you have high incentive and low incentive. And in fact, they have the same effect. They are not statistically different. So one of them will come out ahead and will be penalized a lot by the fact that it's very close to the other one.

But in fact, if you know that they should be pooled, then you won't have that problem. You'll just have incentives. So that's why we run it in these steps.

So if we go back with how many treatments we have, so incentives are crossed two by two, by high and low, and flat and slope. Reminders have to. Smaller, a few versus a lot. The ambassadors of random, plus information seed, plus trusted seed, plus I think there's one from the board information times trust. That makes four.

So in total, that's 75 unique policy combinations. Everything is interacted with everything. So it's 75 unique policy combinations. So with 914, which are highly used, this is useful. This is a sample that where we do a minor in incentive, but the gossip is only done in 900. The resulting sample size would actually be quite low.

That's how everything is crossed with everything. It doesn't really matter. So we could start. We start by doing-- if we start by doing the naive thing of not interacting with everything, you run this regression. So this is a dummy for each thing. It's already giving you some interesting results. We already saw the gossip results.

And then, in terms of incentives it looks like the high slope incentive works. Although, maybe, you can say that it's so statistically different from low slope. High slope giving flat incentive doesn't seem to work. SMS, there's not significant impact. This isn't in the entire sample. And that's in the smaller sample.

So if you just looked at that you would say, well, gossip looked effective, perhaps as effective as giving high slope incentive. And otherwise, we are not very sure what will pool, what will not pool. And of course, we have nothing to say about the interaction from looking in that.

So that's the procedure. You define what's allowed to pool potentially. So in our case, it's pretty simple because it's by category of intervention. Once you do that, you define-- you specify the regression in this marginal way. So to me, that looks very intuitive. But in fact, it's a bit tricky when you have a lot of possibilities.

If two by two, it's pretty clear. There's only one way to do it. If you start having many, potentially a large number of different treatments and different treatment profiles, it takes a bit of combinatorics to say how to make sure not to forget some, and how they should be pooled, and why it still works. And that's described in the paper.

Then, you use the Puffer transformation to make it LASSO compatible. So that's just a way of de-- basically changing the expression of the variable to make them orthogonal to each other. It doesn't come for free, by the way. It adds noise to the data, as it should. Then you run your-- so you make your LASSO, then your post LASSO, and finally the winner curse correction.

So that's the pooling specification, looks something like that where you have, for example, SMS, any SMS, and then the marginal effect of SMS, any slope, and then whether it's high, any flat incentive, and whether it's a flat incentive. Random is separate. Info Hub is separate. Trust is separate. And then Trusted Info is on top of trust. And then, all of the interactions that are in the axes of all of these things. So that's the main effect, as in the main treatment profile.

So you see that we already took some decisions in some sense because for example, we don't allow-- we don't do an any-- Random is its own. We don't allow random to pool with the rest because we think that seed is different than-- a chosen seed is different than a random seed. So there is already a little bit of theory in there. Theory in there, structured in there, even in doing just that.

So what do we find when we do all this? So very few things survive LASSO. Only four combinations survive Lasso, in terms of increase of immunization. So this, we just do it very simple. It's a number. It's number of shots given-- number of measles shots given in the village every month. So the average, it's 7.32 in the control group. What appears to be the most effective is to do everything.

So that's even after all this work, maybe you could have known it. But that was not obvious. And all of the incentives pool. Now, all of the slopy incentives pool. So high and low incentive pull. The Info Hub pool. So the trusted gossip and trusted gossip pull together. And the SMS pool.

So it's pretty nice that the variance seems to make no difference. So you could go with-- if you want to pick the cheapest, you could pick low, slopy incentive SMS, low level of SMS, and gossips because that pools with the more ambitious version of it. And that increases immunization level by four, relative to control, which was a 7.32. So it's a pretty large proportional increase.

The other things-- so the other thing that also works is something very similar, but using the trusted seed instead of the info hubs. That's also significant. These other two, although they were selected by LASSO and the post Lasso appears to be insignificant on their own, although you can make them-- you can set a difference, which is no seed at all, and high slope, and SMS.

And then, this one is-- actually, it has a negative impact. It's selected by LASSO, but it does have a negative impact here.

Once you look at-- yeah?

AUDIENCE: [INAUDIBLE] I guess two questions. That line, does that mean that a whole lot of treatments are pooled in with control?

ESTHER DUFLO: Yes, all of the treatments are pooled with the control. I mean, LASSO doesn't want them. So basically, they are. They become--

AUDIENCE: OK. And then on that last [INAUDIBLE]

ESTHER DUFLO: So LASSO selects some variable, and then you post-- you run your OLS regression. And then, it doesn't turn out to be significant. LASSO penalizes you, basically. What LASSO does is that it penalizes you for every variable you introduce. And then, it lets you keep-- so at some point, depending on how-- depending on the penalty level you choose, it says how many variables you should introduce. And then, you introduce them by the fact that they contribute more to the variance.

To this case, it's after that day, which doesn't mean they are that-- it means these other things that have even smaller coefficients.

Also, once you rerun the regression with only the four, you get different coefficient. You don't get the same coefficient as when you run if you had attempted to run the full regression with OLS because all of the rest is sent to control.

AUDIENCE: [INAUDIBLE] even then just changing your [INAUDIBLE] was trusted [INAUDIBLE] and that means that [INAUDIBLE]

ESTHER DUFLO: This one to this one. Yeah. I don't think the trusted-- yeah, I don't know. I don't think that the trust-- it's the trusted one or not trusted. So let me think. Yeah, it's just trusted by the-- trusted info on their own.

Yeah, I don't have a great interpretation for why they seem to pull with the other one. But they produce no effect on their own. This is, by the way, that's the result we had earlier. That the trusted gossip on their own seems to have no effect. But they have no effect in a way that it's not significantly different from putting them with the other info hubs.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yeah, exactly.

Oh, incidentally, when you run this regression, you rerun it now as in each of them is an individual package. So it's the effect of-- the combo of each of these is the effect of the combo policy, not the marginal effect.

When you look at immunizations per dollars, LASSO selects a few more things. But most of them are negative. Why? Because you are spending money on these programs. And we are now comparing to the control group. So you could well say that you are perfectly willing to spend money on getting more kids immunized even if the cost per immunization increases. But it could also be, like we had in the pool setting, that there is some treatment where the cost per immunization decreases because there are fixed cost of running your program.

So in the cost per immunization, we include both fixed cost and variable cost. Any intervention like incentives, et cetera, increases the variable cost. But means that this fixed cost is spread over a larger group of people. So in the-- that's why in the Udaipur pool results we had that seemingly paradoxical result. That it's cheaper to give incentive, not to give them.

But it didn't have to be the case. And in fact, in the entire sample, all of the incentive results, all of the incentive effects are selected because they have this very negative treatment effect. They are expensive, the cost per dollar. But what emerges? And in fact, in this case is the only policy that is cost-effective compared to doing nothing, in the sense that it increased the cost on immunization per dollar is the doing SMS combined within info hubs.

So it's gossip plus SMS. It makes sense. It's cheap to do. And therefore, the marginal cost is not very large to the extent they are impacts. The impacts means that all your fixed infrastructure is used more effectively.

So now, we can do the best policy correction. So if you remember, we had an initial effect of about four for the combination for the best policy, in case of number of shots, which was the combo of everything. So we penalize it for being close to the next one. And we arrived, therefore, at the effect of an increase in immunization compared to control of about three.

Four million shots per dollar. Since there is only one policy that has a positive impact, it doesn't get penalized. It's on its own. So the penalty impact is the impact.

So that's a summary of the results. But I think we went through this already. So what's the policy prescription? So at some point, all these people from high end actually, they came here. And the delegation will present them the results. And so obviously, it would be-- they could do gossip SMS everywhere. It's cheap, and it's effective, and it's cost effective. So fine.

But if they're also interested in increasing-- in some places doing-- really having a big push on immunization. They are willing to pay for it. The most effective policy is the combination. So then, are there places they can identify where this policy would be effective? And that's the second paper that you read, which is trying to do this predictive medicine, which is can I figure out a group of people for whom this intervention will work?

And the issue is predictive medicine is even more than with multiple treatment, is this possibility of specification searching, which is exposed. You will always find a group for which it looks like the treatment was working. So that's why the FDA does not allow you to report any subgroup that were not pre-specified. You just cannot. Otherwise, you say, well, my drug doesn't work on average. But it worked on people which have read charts on the day of the trial, and then you can always come up with a reason.

So in economics, we are a little less stressed about that probably because there are less financial incentive reasons to push a particular results. And there is always-- we have more tolerance, I think, for theorizing ex-post about results, partly because we like theory, and we should. And we've learned something about the world, and even if the subgroups are experts, you may have learned something about them. So it would be a bit sad not to use them.

So that's the exact same-- it's not exactly the same trade off as before, but there is a trade off of that nature. So going all the way to the medicine standard and say, we are just going to allow a very few pre-analysis plan and penalize multiple comparison with q-values or other forms of adjusting for multiple comparisons. Then, you would get a bit boring. You wouldn't be able to get as much information.

And often, we have a lot of data. And we would like to use it, but we would like to use it in a disciplined way. Hence, machine learning has the advantage of potentially giving you some rules, some algorithm that you can specify in advance, that you could put in your pre-analysis plan. I'm going to use so-and-so's machine learning technique to look for possible-- for whether they are heterogeneous effect and who seemed to be the most affected.

So in principle, the hope is you get the best of both worlds because you have the ability to look for differences in a statistically disciplined way. And you won't do too much data mining. The problem is minor problem, machine learning is really not a tool for saying. It's a tool for prediction. It's a tool for predicting what's going to happen, but it's not a tool to interpret any differences in treatment effect we find.

And in particular, in the absence of very strict conditions to be met. We don't know how to estimate the causal average treatment effect in a way that is interpretable. So the idea that, OK, I'm going to-- the full predictive medicine ideal for anybody with any set of x characteristics, and I can predict their treatment effect. We just don't know how to do that, unless we are willing to be super restrictive.

And then, Susan [INAUDIBLE] has a paper with Steven Wagner with enough under sufficiently harsh conditions, you can do things like trees to give an answer to this question. But these conditions are pretty hard, and they are not really testable. So what can you do?

Well, you can say, let's give up on estimate the whole function. Can we say at least a few things? So we give up on estimating the full condition of a treatment effect as opposed unrestricted as a function of all those x's that people or villages might have. But I want to answer a few simple questions.

Number one, is there any heterogeneity? So that's one parameter. Number two, if there is any-- or is it just any heterogeneity that I think is there is in fact, not there, number two if there is any heterogeneity, can I put-- can I make say, five groups where I estimate the treatment effect for these five groups. So it's now five parameters, still a small number of parameters.

And then, maybe one more step is saying, well, now that I have these five groups, maybe I can describe the characteristics of these five groups. It's not going to be causal. It's going to be predictive. But maybe it's interesting, anyway. So that's what was in the paper was Victor. We use the machinery of the Bell machine learning that Victor Whitney had co-developed that we already discussed in the Ghana paper to apply to this problem.

Not important to answer to understand the detail of how it works. The most important thing is what I already told you, which is very clear. Reduce your ambitions.

The second most important thing is as in the first thing, as in that first paper, use sample splitting. And basically, think, for example, is there any heterogeneity? How is it going to work? Well, I split my sample in two. In half of the sample, I use any machine learning tool I like to come up with my best guess for the causal average treatment effect.

That is, think of it if you think of it in some version of LASSO, for example. It's interacting all my x's and all of the interaction possible with the treatment. And then, that gives in principle a prediction for each person of what the treatment effect should be. I know that this is unlikely to be great because it's full of flux, and noise, and stuff like that that are going to be wrongly attributed to the effect of the x's. But I still do it.

You could use LASSO, random forest, neural network, anything you like to make that prediction for each combination of covariates. And then, in the sample that you left behind, so note that it's different from sample splitting as it's used in ML typically because for example, in order to do even the exercise that I'm doing, usually, you will need sample splitting to tune the parameter and all that. So that's not-- that's sample splitting.

My sample splitting is that in my half of the sample, I put my best foot forward to estimate the thing. And then, in the half of the sample that I kept, I'm going to regress the treatment effect on this characteristic and see whether I get just linearly. And say, if there is any effect, I should find one.

If the two effect was as estimated in-- if the two heterogeneity in effect was as estimated in my machine learning sample, in the test sample I should find a coefficient of one when I try to predict it for each person. It should line up in the same way. And if in fact, it was all noise, it's not going to line up in the same way. So I'm going to find a big fat zero.

Empirically, in my experience of trying this, most of the time you stop there because you find zero. That where you think that you're heterogeneity is maybe you even told yourself some nice stories about why it would be heterogenous. And in fact, once you do it in the other sample, and you regress the predicted treatment effect on the actual treatment effect for the person with those characteristics, you don't find any difference.

If you do, then you can say, oh, great. Then now, I can separate my groups into who is the most likely to be affected by the treatment, or by quintile, and run and calculate the treatment effect for the people who are predicted to be the least affected, and the second least affected, and the third least affected. And so I make the separation based on my machine learning sample. And then, I estimate the treatment effect in that group.

So this is a treatment effect in that group in the leftover sample. And then once you've done that once, you can do it many, many, many times, many of the splits many times. And then, you even get a standard, therefore, for that effect. That reflects both the actual noise you have in your data set and the fact that your sample split. So each sample still introduce its own noise that needs to be taken into account. Yep?

AUDIENCE: What exactly is the difference between what you do in the training set and the test set? So in the training set, it's the first time.

ESTHER DUFLO: You will stimulate the treatment-- you will stimulate the treatment effect.

AUDIENCE: Interactive regression.

ESTHER DUFLO: The interactive regression. For example, think of it this way. In the second set you predict for each person what's there as a function of your estimate in the z, what their treatment effect would have been. And, then you regress.

For example, you run a linear regression of why your outcome. So think of running the following regression. On your outcome, on treatment, treatment interaction interacted with this predicted effect and the predicted effect. So this is now a linear coefficient. And just asking, are the people that I predict to have the largest treatment effect, do they in fact have a larger treatment effect?

That linear parameter is of no interest in and of itself. It just say, it should be-- ideally, if that estimate was very, very precise, that coefficient would be one. If that estimate is junk, that estimate would be zero. That's the main use of that linear thing.

Then, what you're doing here is similar, but dummified, dummified version. You basically on the basis of the estimate in your machine learning sample, you've made five groups of the people who are predicted to be the most affected based on their X characteristics and their baseline characteristics. And then, in your test sample, you are actually estimating the treatment effect for the people who have these characteristics, this combination of characteristics. In that quintile of the people who are predicted to have the lowest treatment effect.

So again, if it was all junk, it would just be a line because the people who have the lowest treatment effect in your machine learning sample, in fact, don't in real life. Makes sense? And then, so you can do that.

So here, these are the results for the immunization sample. So they are exceptionally nice in that they actually-- something comes out of it. It's not always the case, as I said. And what you find is very, very strong heterogeneity to the point that the places that are most affected are really affected. You have huge treatment effect. And the places which are not, which are least effective, have negative treatment effect.

So then, what you can do is two things. First, I will discuss that for a minute of why is it even possible to have negative treatment effect. And secondly, you could say that the last thing that we do is to say, well, we can do a classification analysis, which is to say, what are the characters-- this is like a bunch of these interacted with each other in a complicated way.

But if I just run a summary statistics, just a summary statistics table of the sum characteristic of these different groups, do they look different, the groups? So there you have to be careful not to overinterpret this difference because again, they're not causal. But at least it gives you a description.

And what we find that's pretty interesting is that the places that are the most affected are the ones where immunization rates were very, very low to start with. And the places that were the least affected are the places that have high immunization rates to start with. So I remember I first presented that in the bank, in the World Bank and my friend Jishnu Das, whose work we are going to talk a lot about in the rest of today and next week.

I was like, come on. You cannot write a paper where you do this whole thing and you just tell me the people who are most affected by-- and immediately forgot about the people whose immunization level were low to start with. And first I was like, yeah, yeah, you're right. That sounds a bit silly. But then I thought about it, and it's like, why?

Even the people with high immunization rate, it's not that they have super-high immunization rate. They have plenty of scope to improve. So it's not mechanical. They are not bounded. They are not at all in the range where they would be bounded. So nothing says that the impact of intervention are concave, in general. In fact, in many cases, it's for example, in the ultra poor program, the people who benefit the most are the richest of the ultra poor.

So I don't think it is mechanical, but maybe it is. It's true that a little bit like the first paper, it sounds a bit sad after all this exercise to come up with something obvious. And then, the other thing that is clearly not mechanical, is that you get a negative impact. Which might be the most interesting-- or is probably the most interesting and scary results of here.

So why do you think the result might be-- why do you think some people might be negatively affected? Some villages and therefore, some people might be negatively affected by a program like that.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Do you mean they were already doing something and--

AUDIENCE: If other people are positioning themselves [INAUDIBLE]

ESTHER DUFLO: It's the number of shots that are given in the month that's the outcome. So there might be crowding out of other things. But here, for example, if people come from other villages, they would be counted. This is including any crowding of other people from other villages. Yep.

AUDIENCE: One hypothesis, I guess, that was mentioned briefly in the paper which is that some people are very intrinsically motivated to get vaccinated. And once you introduce this external encouragement, then there's maybe some behavioral explanation of actually I don't want to do it anymore.

ESTHER DUFLO: Yes, maybe this is one of these example of weaknesses as a series of paper. One is called "Finding the Price" that shows that when a school penalizes parents, gives the fine to people for parents when they collect their kids late at daycare, the number of parents who send their kids-- who pick up their kids late actually increases because before, people thought, oh, this is the deadline. I have to pick up my kids.

And now, it's you're actually sending a signal that you can do it against some price. Other example--

AUDIENCE: Another reason why all this stuff happened was if everyone in the village had already been treated with the [INAUDIBLE] people for vaccinations, and everyone else needs lots of information [INAUDIBLE]. But with financial incentive, they might wonder like, why is the government or whoever it is trying to get money [INAUDIBLE] people--

ESTHER DUFLO: Yeah, exactly. So the people-- so some version of it, which behavioral or not could be that people were already doing it. And then, suddenly, there is an incentive. Why do they need me? Why do they need to pay me to do that? It must be really bad if they need to pay me to do that. I shouldn't want to do it by myself.

In the context of the COVID 19 vaccine, for example, in discussion in France of making it mandatory or quasi-mandatory, which is what it is now, it certainly wasn't. It was there that if you make something mandatory, then it says-- it sends a signal that it's not something you should desire. And instead, in fact, people were extremely suspicious of the vaccine in France. It was one of the highest-- before the vaccination campaign even started, it was one of the highest suspicion countries.

And then, the beginning of the campaign was totally butched. It was really slow. In the first two weeks, they immunized 12 people. Since they immunized some woman called Morrisette in an old age home person. They immunized 12 people in two weeks. It was a disaster. The whole thing was about there is not going to be enough vaccine, and we are not going to be able to do it.

And then, the view on whether people wanted to get vaccinated changed. And one possible interpretation is that that became to be seen like something rare. That you should desire because there is not enough of it. And there is not-- and these idiots of the government, they are not able to deliver this good that we really need. And therefore, your mistrust in the government paradoxically, I think, played in there in the favor of generating a consensus more than anything else.

And then, I'm not saying they did that on purpose. I know they didn't. And I don't know whether my explanation is correct, but it could be. What I know they did do on purpose is to not have a mandate, or anything looking like that, or incentives until quite late because they were worried about that backlash effect. That if you put an incentive, people are just going to not do it just to spite the government.

And here, that might be a bit like that. It's why do they need to pay me to do the thing? I was going to do it anyway, but it must be good if they pay me to do it. And that's consistent with the fact that the effects are the smallest in places that have a lot of immunization already because it leaves a bunch of people who would otherwise have done it who can be unconvinced.

A lot of people-- whereas, it was in the effect in the places where nobody gets vaccinated. There is nobody who was intrinsically motivated to start with, so your extrinsic motivation is helping it.

AUDIENCE: In this particular setting, do you have any evidence or prior to this intervention when people were getting immunized, was it previously associated with the government? So could it be that now the government wants to give me the-- previously, I was having to go to my doctor. There wasn't as much of a link between immunizations and the government wanting you to do it. So maybe it's specifically about, I'm mistrustful of the government. But not necessarily of medicine. But now that the government is imposing itself, I don't know if--

ESTHER DUFLO: I don't know because the incentive was done as part of the government structures anyway. So it could be an instance of keep the government out of my Medicare. That is, suddenly people realized the importance of the government in earning their care, but I'm not sure. It could be.

AUDIENCE: You get a result like this and you're communicating it back to the call center. Do you translate it back into [INAUDIBLE]

ESTHER DUFLO: Yes, so the way that-- this one translates nicely. Basically, you have a table like that, except you can express it in graphs to say, well, you get the biggest effect in places where there is a lot of vaccination already. These are five dimensions of vaccination, but you can pick just one. And none of the other-- and then, that's the one thing that comes out.

And for example, for the government, very important to tell them that it's not about religion. This is Haryana. That's like Hindu. The government is like Hindu dominated. So they are very prompt to say that this is are Muslims who don't want to get immunized. So you can show this effect doesn't vary by Muslim, Hindu. That kind of things.

But you're right. That's why we want to get to the [INAUDIBLE] to give something that is concrete, actionable, understandable. But they are still not doing it, just full one. It's not there. We're very keen to scale up the gossip plus SMS. But this one is more complicated. The fact that it cost money, so even though another important message for policy is that if you focus on the places that are the most-- that have the largest effect, it becomes cost-effective in those places.

Not tremendously so. It's not terribly-- it's not terribly much cheaper than the control, which is what we do in which we find in Udaipur. But in the top groups, you cannot project that you don't lose any money per shot compared to not doing the incentives. So that's also why it's important to have that analysis because the policy message becomes a little bit more subtle. It's A, you don't want to do it where it's going to backfire. But B, in places where it actually works, you might as well do it. It's not even going to cost you more per immunization.

But I totally take the point that. My empirical observation is that that's a message that's harder to convey than the simple message. This is what you can do everywhere. This is cheap to do, and go for it.

So all that means that-- so first of all, there is a lot we don't fully understand for health. But it's really a great set of topics. Many of them still open. And then, it makes for a pretty complicated supply response. In an environment where the demand is complicated-- has all of these things going on, the trust, the hyperbolic discounting, the misinformation, et cetera. You might expect that's where the supply is going to be hard to discipline.

So we don't really have time to cover all of the supply today. But I can just start a little bit, and then continue. I do want to teach about health care supply, so I'm going to continue later. Where do we start?

So the motivation is from Arrow. Health care is a credence good with a lot of asymmetric information. The provider knows more than the patient. For all of the reason we described, that's clearly true. Learning is very difficult. And there are a lot of externalities.

So even if you knew everything, you will have fairly low demand for some goods, like prevention. And then, demand too high for some other goods, like antibiotic. Even if you understand everything, just because the antibiotic is going to help you now, whereas the immunization is going to help all other people. So you cannot rely on the individuals to choose the right outcome.

For all of these reasons, unregulated private health care will tend to treat badly, and to over provide some medication and to under provide some medication. Hence, our conclusion is not talking about India. It's talking about the US, but that applies to India. It is the general social consensus, clearly, that the laissez-faire solution for medicine is intolerable. We agree. Except that-- I'm going to skip that. Well, maybe not.

So indeed, learning about the quality of health care you receive is very difficult, both because of what we discussed of self-limiting diseases, and because of externalities. As a result, people have a very poor understanding of their health. And they have a poor understanding of whether they are getting good health care. So we discussed self-limited diseases already.

And that leads potentially to overuse of antibiotic or [INAUDIBLE] for COVID because you think that they have been effective, when in fact, it's just that the disease became later. The other problem is externalities. That's when a lot of people get immunized or get dewormed, nobody gets sick. And you don't get less sick because you got the immunization or the deworming.

The result is that there is a lot of-- it's very easy for a provider to take advantage of the situation to give very, very bad care. And in particular, to give very, very, very bad medicines. In an environment where many people take malaria treatment even when they don't have malaria, and people are very naive about the model, there is very strong incentive to sell fake treatment because usually, you can get away with it.

So to investigate that, Bjorkman, Svensson, and Yanakizawa-Drott sent mystery shopper to buy a anti-malaria drug, the new one, ACT, artemisinin combination therapy at 99 markets in Uganda, and then test them in the lab. And what they find is that 37% of the shop sells some counterfeits, that the price is unrelated with whether it's counterfeits or not, and that the share-- the proportion of counterfeits is increasing in the share of naive consumers or people who misunderstand malaria.

So they ask people questions about what you think causes malaria. So some people think that rain causes malaria. And some people know that it's mosquitoes. And there are mosquitoes when it rains. And the more of these naive household you have, the more counterfeits there is. People are more likely to have-- sorry. This is it here. The more naive people you have, the more likely it is that a place sell counterfeit malaria.

And that's related to the fact that the more naive people are, the more people are convinced that the drug they get is good, perhaps because they tried it, and it worked as far as they're concerned. So what they do in this paper is that they ask themselves whether a good quality can drive bad ones. So they work with BRAC, the same of ultra poor fame and otherwise extremely reputable and competent organization, who now doesn't work just in Bangladesh, but also in Uganda and other East African countries.

And BRAC sets up pharmacies, which sells bonded malaria medicine of good quality. And their question is, is it going to lead the other people to disappear? And the question-- the answer is in two steps. On average, the quality in the shop increases, and the price of the malaria medicine decreases. So on average, the competition effect, in fact, works out. If you introduce this bonded product, people will flock to that. And then, put pressure on the other guys. And they will tend to either go for good medicine and reduce the prices.

However, this works less in villages where they are more naive customers. So when you interact-- you use this table, when you interact the fraction the NGO sells drug on average, there is less likely to be fake drug in the village. If there are more native people as we saw before, that's this one, they are more fake drugs. And then, if there is more native people in the village, the treatment effect is less strong.

So number one, when you have more native people in the village who don't understand the model, you have more fake drugs. Number two, it makes it difficult to-- more difficult to address it by putting good drugs. So that's just to say that-- to comfort the point that Arrow made. It's hard to learn about health and provider take advantage of it.

And therefore, the private health care sector is poor countries is really awful. And I'll tell you a bit more about how awful it is next time. So we'll stop here.

I think we'll have time to finish that in about a half lecture, which still will give us a one and a half lecture on gender. So that will be fine.

But on Monday, Ben is teaching. So you're going to leave that in the fridge for a minute, for one lecture, and we'll go back together.