AI ETHICS

Data Science against COVID-19: The Valencian Experience

By Nuria Oliver, Co-Founder and Scientific Director, ELLIS Alicante Unit

In this module, we explore the ELLIS Alicante Foundation's Data-Science for COVID-19 team's work in the Valencian region of Spain. The team was founded in response to the pandemic in March 2020 to assist policymakers in making informed, evidence-based decisions. The team tackles four different work areas: modeling human mobility, building computational epidemiological models, predictive models on the prevalence of the disease, and operating one of the largest online citizen surveys related to COVID-19 in the world. This lecture explains the four work streams and shares lessons learned from their work at the intersection between data, AI, and the pandemic.

Lecture Transcript

0:01 Hello, I'm Nuria Oliver. I am co-founder and director of the ELLIS Unit Alicante Foundation called the Institute of Humanity Centric AI. Today I'm going to be sharing with you an experience on using data and AI to support better policymaking in the Valencian region of Spain. This has been done in the context of the Data-Science for COVID-19 team, a team that I have led, or I led between March of 2020 and April of 2022, and composed of scientists of all levels from students to full professors from a variety of universities and research centers in the Valencian region of Spain, which is located on the east of Spain by the Mediterranean coast.

0:49 The main purpose for this team, which was created at the very beginning of the pandemic, is and has been to assist the policymakers and particularly the presidency of the region into making more informed decisions related to the pandemic. Decisions that would be informed by the evidence captured by the data. Fundamentally, the main goal has been in filling the gap that there is between where the data is and where the policymakers are. To fill this gap, we first structured our work into four different areas. First of all, human mobility modeling, because an infectious disease that is transmitted from human to human doesn't become a pandemic if people don't move. So modeling human mobility is of utmost importance. The second work stream has been on building computational epidemiological models, models that would predict the number of COVID-19 cases, not only under the current situation, but under different potential scenarios. The third area has focused on building predictive models mostly of hospital occupancy, intensive care occupancy, and also of the prevalence of the disease. And the last area consists of a very large scale online citizen survey called the COVID-19 Impact Survey, which has over 720,000 answers and is one of the largest surveys in the world related to COVID-19.

2:20 But even with the output of all these different modules, there is still a gap between where the results of these work streams are and where the policymakers are. And I say in one of the key elements for the success of our team has been this layer here in the sort of like tan color of 'results interpretation, aggregation, preparation and

translation' into political insights and actionable items that has been mainly done by Director General working for the President of the region, but belonging to the team in collaboration with me.

3:00 The team is a very multidisciplinary team, with members that have a variety of areas of expertise depending on the work stream. So the mobile data team has a lot of expertise on data wrangling, data visualization using sort of like GIS tools. The computational epidemiological team has a lot of expertise on computer modeling and machine learning. The predictive models team has also a lot of background on machine learning and statistics. And the citizen science team has also background in human computer interaction.

3:34 We've been working very hard in a very agile and intense way. I've organized daily meetings for many months at 9:30 am. And I've been writing reports that I've shared directly with the presidency of the government of the region every day, at around 5pm. Everyone obviously has signed the corresponding NDAs and code of ethics. We share the code using GitHub, and we have a very active Slack channel that we've been using throughout the pandemic. This is an example of one of our online meetings. We actually had never met in-person most of us and we had never worked together most of us either. We have an official website, which is in Spanish and in Valencian on the official website of the government of the Valencian region of Spain.

4:24 Something that is important to highlight is that our experience on our team has been fairly unique in the world, because there are a lot of challenges related to creating such a team. First of all there are capacity, awareness and also like detail and difficulties in the governments and the public authorities that make it very difficult for a team like this to exist and be sustained over time. There are also concerns and difficulties about accessing the data. Of course, there are obvious concerns about the privacy and data protection. During the beginning of our work, we work very closely with the Data Protection Officer at the government. But in any case, we never analyzed any personal data or any individual data. We analyze mostly publicly available data. There is also a gap between where the research outputs are and where the operational projects and the decisions are made. And one of the challenges is how to fill that gap. And of course, if all these challenges are not met, when you are facing a pandemic, where you have to make decisions very quickly, constantly, it is very difficult to leverage this idea of using data to support policymaking if you don't have already all these elements in place.

5:55 So what have we done? I'm going to quickly share a summary of some of the results and the type of work that we have done in each of these work streams. Starting with a mobile data analysis. In this area, we wanted to be able to answer questions such as: if the confinement measures implemented by the government had impacted the mobility and to which degree, what kind of mobility was impacted, and also, if the reduction in mobility was enough to contain the spread of the virus. This is a visualization of the type of data that we've analyzed. So we were declare the pilot region in Spain by Vice President Calviño to be able to access aggregated anonymized human mobility data that was shared with us by the National Office of Statistics who had collected that data through a collaboration with the three largest telcos in Spain: Telefonica, Vodafone, and Orange. And this is a visualization of the mobility in the region of Spain in March of 2020 during the beginning of the pandemic.

7:00 With this data, we analyze different elements of mobility. The first one was the success or not of the stay at home campaign. To do that, we measure, and we show here the percentage of population that never left their area of residence for more than two hours in a day, which is the kind of data that we had. Each of these irregularly shaped regions that are seen on the map are the lowest level of granularity that we had access to the data.

7:33 And as you can see, throughout the first wave of infections of COVID-19, in the spring of 2020, there was a huge reduction in the mobility of the population reflected in a very large percentage of the population that stayed in their area of residence. You see most of the map is green and green corresponds to at least 80% of the population never leaving their area of residence during the day for more than two hours. We can also see the evolution of the impact of the confinement measures on The Stay At Home campaign, where we see that at the very beginning of the pandemic, in mid-March of 2020, before the labor confinement measures were implemented, the map appears as a yellow and orange so the there was around 80 or 70% of the population staying in their area of residence and this percentage significantly increased during the weeks in Spain, where we had no labor mobility. And as you can see, the map became completely green, meaning that more than 90% of the population never left their area of residence for more than two hours during the day.

8:45 To support public policymaking beyond these different areas, it is also important to understand and carry out the analysis in what is called the health zones. The health zone is a geographic area that is served by a hospital. There are 24 health zones in the Valencian region of Spain. And here we show the percentage of the population that stayed in their area of residence in weekends and during the week for each of the different health zones in the Valencian region. And as you can see, for all of them, even during the week, more than 80% of the population stayed in the area of residence. We also studied the impact of the confinement measures on labor mobility given that labor mobility is one of the biggest components of human mobility. And we found that on average, there were 60% less people outside of their area of residence during working hours during the confinement period of Spain in the first wave of COVID-19 when compared to a reference day in November of 2019.

9:56 Of course to support policymaking, it's also very important to develop intuitive visualizations that could ease the understanding and the use of the data and the analysis done on the data. And we built this visualization for the government so they could click on any municipality and understand the levels of incoming mobility and outgoing mobility and labor mobility as a percentage of population that has stayed at home, and so forth, and the terrain to which degree the confinement measures are working. Beyond understanding the direct impact on the mobility, we also performed our community analysis on the mobility data to identify what is called mobility communities. These are geographic regions that have a lot of inner community, inner mobility, but don't have a lot of mobility with other regions. Through this analysis, we identified 14 Mobility communities in the Valencian region that could be used to identify the optimal way to do, for example, partial confinements.

11:05 The second work stream has focused on building computational epidemiological models. These are computer models that enable us to answer questions such as: how many infected individuals would be under different possible scenarios of different confinement measures or non pharmaceutical interventions. And of course, these models will enable us to determine whether the confinement measures are enough or stricter measures will need to be implemented. We developed three different computational epidemiological models. The first one is, Xi (ξ) metapopulation model, is a classical computational epidemiological model that divides the population into four states, which correspond to the acronym of the name - SEIR. So the population is originally in a susceptible state because no one has been exposed before to COVID-19. So everyone is supposed to be susceptible of getting the disease with a certain probability (beta - β). Some fraction of the population is exposed to the virus with a certain probability (sigma - σ), those that are exposed become infectious and infected. And with another probability (gamma - γ), they are retired from the system because they recover or they die. This model is

given by a set of differential equations with certain parameters to characterize these different probabilities that had been estimated already at the beginning of the pandemic in the literature for COVID-19. And we adjusted those parameters to fit well the data observed in the Valencian region of Spain. We use this model every day to make predictions on the number of COVID-19 cases and the number of active COVID-19 cases throughout the pandemic.

12:53 The second model that we built is an agent-based epidemiological model. This model is an individual model, which means that it models each individual in the region which in the Valencian region of Spain will be 4.9 million people. Each individual is modeled by an agent, which has an age and a gender, and some behavior and a contact matrix. And these agents can also be in the four different states of susceptible, exposed, infected and recovered. And with this system, once the parameters are tuned, you can run simulations and see how the pandemic curve would evolve. We use this system as well every day to obtain predictions on the number of COVID-19 infections, the number of hospitalized patients, intensive care unit occupancy and death.

13:45 And finally, in the context of the XPrize, COVID-19 XPrize Pandemic Response Challenge, we decided to participate and create a team called the Valencia IA4COVID-19. And to participate in this challenge, which took place between November of 2020 and March of 2021, we developed a different computational epidemiological model using deep neural networks. In this case, we have two banks of deep neural networks particularly LSTMs. On the top we can see the LSTMs that model the number of COVID 19 cases in 236 regions and countries in the world. And on the bottom we see the layer that models the interventions, the non pharmaceutical interventions or confinement measures. And the predictions that the model provides are the result of combining the predictions provided by the top bank of models and the bottom layer. This predictor performed pretty well during the competition. It was the third best model in the main rank globally and the first in the countries for Europe and Asia during the competition. But more importantly, we were able to use this model starting in December of 2020, which meant right on time for the third wave of infections of COVID-19 in the Valencian region of Spain, which took place right after Christmas of 2020/2021.

15:21 As you can see on the top graphs, the yellow dashed line is the ground truth, the real number of cases, the blue are the predictions by our model and the red are the predictions by a state of the art model. As you can see, our model predicted the third wave of infections really well. We worked during 2021 in expanding this model to include vaccination. And what you can see on the bottom are the predictions that our model did to predict the number of COVID-19 cases during the sixth wave of infections that took place right at the beginning of 2022. And again, you can see that our model, in this case is the red line predicted very accurately the sixth wave of infections when compared to a state of the art model shown in gray.

16:08 Moreover, in the context of the COVID-19 competition, the XPrize competition, we had to add a new layer to our architecture, which is shown here in pink, which is an automatic prescriptor of non pharmaceutical interventions or policies. The main challenge was to create a system that would recommend up to 10 different policies for each country and region in the world, that would have the ultimate optimal trade off between the cost of implementing such a policy, the economic costs and the social cost, and the number of COVID-19 cases that would result if such a policy were to be applied. To do that, we developed different strategies and algorithms to come up with the 10 recommendations that would be on the Pareto front on this two dimensional space, where on one axis, we have the cost of the interventions and on the other axis, we have the number of COVID 19 cases that

will result. We build some Tableau visualizations of our prescriptor of interventions. So the policymakers could use them, they could click on a particular country, and they could see the different potential policies, and the number of COVID-19 cases that will result from applying such policies. Both the COVID-19 case predictor and this prescriptor of interventions were part of the solution that we provided for the XPrize Pandemic Response Challenge and we were declared the world winners of the challenge, which offered great external validation to our work. We also have published a couple of papers on this topic, and we won best paper award in ECML PKDD 2021.

18:04 In the summer, at the end of the spring of 2020, we performed an analysis using the agent based model on the impact of contact tracing. We know that contact tracing was going to be of critical importance to try to contain a new wave of infections and we did some simulations to understand what percentage of the infected individuals would need to be contact trace to be able to flatten the curve. And according to our simulations, even if around 40% of infected individuals were contact traced, of course, assuming they would all isolate, then you could really reduce the peak of the number of infections. Since the spring of 2021, when the vaccines became widely available, we incorporated vaccination both on the deep learning based model that we use for the XPrize competition and also on the individual agent based model. And we've been running simulations since then using the vaccination.

19:09 The fair workstream consists of building predictive models, mostly predictive models of hospital occupancy, intensive care occupancy, and also, we built a model to infer the prevalence of the disease. We built different kinds of models, including deep neural network based models also using LSTMs, which we use to predict the intensive care occupancy as shown on the blue graph here.

19:40 And last but not least, we deployed citizen science part of this project through an online anonymous survey called the COVID-19 Impact Survey. We launched this survey in March of 2020 because there were a lot of important questions related to the pandemic that we couldn't really respond to, because we didn't have the right data sources. For example, what is the social contract behavior that people have even during the confinement measures? What is the resilience of the population towards the confinement? What's the prevalence of symptoms? What's the testing availability? What's the emotional impact of the pandemic on people's lives? What kind of individual protection measures do people take during the pandemic? Is contact tracing working? Is the app for contact tracing working? I mean, there were so many questions that we couldn't really answer. So we decided to design the shortest possible survey that would enable us to answer such questions. The survey originally had 26 questions, we expanded it in the summer of 2021 to include six more questions on social isolation. So it has 32 questions right now. It is still online. It's been translated to many different languages. And it was launched on March 28, 2020. We obtained an incredible response to the survey. In the first 40 hours, we collected over 140,000 answers, mostly from Spain. So we decided to very quickly analyze the data, share the data, and also prepare a scientific paper sharing the main insights and results from analyzing the data. And we did that and we published that in April of 2020.

21:28 We also built two visualizations of the survey data, the one on the left using ArcGIS and the one on the right, using Tableau. And we've been, we've kept updating these visualizations on a weekly basis, so everyone could understand you know what the impact of the pandemic was on people's lives. We've published several papers describing some of the main insights resulting from the analysis of the data. For example, we compared the

impact of the control and mitigation strategies used in Italy and Spain, because we have a very large sample of responses also from Italy, from the survey. Or, in another study, we looked at which ones are the key factors that determine people's willingness or unwillingness to be confined during the pandemic. And what are some of the results of this survey, so just want to share some of the perhaps more interesting findings. The first one is about the emotional impact of the pandemic. This is a large sample of over 380,000 answers. And what we have found week after week throughout the entire pandemic is that the most impacted group, psychologically, is the youth and particularly young woman who are the ones that report the highest levels of stress. Over 50% of the young females, aged 18 to 29, report having a stress level that they consider detrimental to their health. Females are represented in this red color and the males in blue color.

23:12 But also loneliness has been very high among the youth, much higher even than among other age groups. When we look at the perception of the government measures, what we observe, and we look at the yellow graph here, which would be the percentage of people that consider that the government should do more, we find a very strong correlation between having a wave of infections and the population demanding more measures. Throughout the entire pandemic, the most popular answer has been that the government should do more.

23:45 When we look at what kinds of protection measures people adopted in Spain, the most popular measure because there was a face mask mandate was wearing masks. And the percentages were extremely high in the upper 80s or 90s, until the restrictions were lifted. And then as you can see here, this is data between January and March of 2022, the percentage of people wearing masks went down a lot. There is a very big gender difference where females, represented in red, are significantly more compliant with all the protection measures when compared to males.

24:27 The willingness to get vaccinated in Spain has been very, very high throughout the pandemic. In fact, the data that I'm showing here is very late in the pandemic when pretty much everyone was already vaccinated. In earlier time periods, over 90% of both male and female respondents would demonstrate a willingness to get vaccinated. It's also interesting to see that indoor ventilation has been the least adopted protection measure throughout the pandemic despite the importance of having very good ventilation to protect, limit the spread of the virus. When we look at the perception of risk of different activities through the entire pandemic, this is data between May of 2020 and March of 2022, we find that the activity that consistently has been considered the safest in terms of risk of infection of COVID-19 is doing sports individually. And the activity that has also been considered the least safe throughout the entire pandemic is flying by plane, even though there aren't really that many reported outbreaks in planes. It is interesting the evolution of the perception of the safety of the beach, which is marked here in red, where we see that in the winter months, people don't consider that it's safe to go to the beach but as soon as the summer months arrive, then a significantly larger percentage of people consider that it’s safe going to the beach.

25:59 A very important element in the TTI - Test Trace Isolate Control Strategy for the pandemic is the capacity of people to self isolate, particularly the capacity of the positive cases to self isolate. And what we have been observing through the entire pandemic is that roughly 50% of the population aged 59 and younger report not being able to self isolate if they had to. This is a very large percentage that we've been sharing with the government this information hoping that they will implement programs to reduce the burden of self isolation on the population so people could more easily self isolate. When we look at the reasons why people cannot self

isolate, the most common reason is because of sharing the household. But then we observe age differences and gender differences, we see that it is the youth those aged 18 to 29 years old, the ones that are more likely to report not being able to self isolate because of psychological reasons, including the fear of stigmatization. And then we also observe a very significant age difference among those that report not being able to self isolate due to having to take care of children, where it is the female respondents aged 30 to 59, the ones that report with the highest probability not being able to self isolate, because of taking care of children, with a much larger in a much larger percentage that their male counterparts of the same age range.

27:46 And finally, we also obtained some data on the effectiveness of the contact tracing app. Most European, most countries in the world deployed some kind of contact tracing app to support the contact tracing efforts. However, there hasn't been a lot of evidence on the effectiveness of such an app. In our survey, we asked people if they discovered being infected with COVID-19 because of the app. And this is the data that we obtained. From a sample of over 139,000 people, 42,000 people reported having the app installed, so there's a very high adoption rate of roughly 30%. Out of these 42,000 people, 7.7% reporting having had a positive contact with a contact or close contact with a positive case, however, only 2.6% of those reported having discovered it via the app, which will be 86 people. Of those 86 people only 27 got tested, and only seven tested positive. So from an original sample of 139,000 people, the app helped identify seven positive cases.

28:59 More recently, we have carried out a study on the impact of the pandemic on social isolation. Social isolation is a very big worry in public health and being submitted to social confinement measures for almost two years or for two years, we hypothesize that it probably had a big impact on the social isolation of the population. So in the summer of 2021, we added six questions in the survey to measure social isolation. It was the short version of the Lubben Social Isolation Scale. And we recently published this paper where we report the really worrisome results where we identify that on average, more than a quarter of the Spanish population is socially isolated, and this percentage goes up to 30% for those aged between 40 and 59. We also observe a huge impact of the economic situation of people in the risk of social isolation, and also a milder, yet significant impact of the psychological impact of the pandemic on the probability of people being socially isolated.

30:18 So what have we learned after two years of this intense work at the intersection between data, AI, and the pandemic? So, the most important lesson that we've learned is that a pandemic is not a public health challenge. It is a societal challenge that requires holistic approaches and solutions. And in particular, we've learned that we have an opportunity to create a virtuous cycle between these three elements: data, people and technology, and processes and policies to be able to move towards a situation where the policies and the processes are measured and evaluated quantitatively, but also are inspired and informed by the data. Data that will need to be systematically captured and analyzed to be able to understand where we are, why we are, where we are, and where we might be going. People and technology. So both the right numbers of experts in a variety of fields from healthcare and social personnel to contact tracers, teachers, researchers, but also empowered with the right technology to be able to do the work properly. And of course, if we analyze the data, and we have all these human resources, we want the result of the work to impact the policies and the processes that we define. So it will be very important to close the loop and to really inform the public policies and measure the impact of the public policies through the analysis of the data, and design public policies that respond to underlying needs in society. For example, if we know that roughly 50% of the population, age 59, and younger and unable to self isolate, it would seem important to deploy programs to support self isolation. If we know that the youth are the demographic

group, which is the most impacted psychologically by the pandemic, it will be very important to deploy psychological support programs for the youth and so forth.

32:39 Reflecting on this intersection between data, technology and society together with the Data-Pop Alliance and the Vodafone Institute, we wrote this publicly available paper where we reflect on how technology and data have been useful in the pandemic, but also put forward six recommendations so in future pandemics or future, similar situations, we can make a better use of data and technology for the people and by the people. We've also extensively published most of the work that I have presented in this presentation. And I encourage you to take a look at these publications if you want to know more. And we've given many, many talks in different events, and we've also had very extensive media coverage of our work. With this, I thank you very much for your interest and I hope that you found my presentation to be relevant and interesting to you. Thank you very much.