WEBVTT - https://subtitletools.com

00:00:02.770 --> 00:00:05.840 - All right, and it says the meeting is being recorded.

00:00:05.840 --> 00:00:10.480 Okay, so thanks everyone,

00:00:10.480 --> 00:00:12.510 for coming to this seminar.

00:00:12.510 --> 00:00:15.773 And I hope everyone is doing well.

00:00:16.900 --> 00:00:21.130 Today, I'm going to talk about some issues

00:00:21.130 --> 00:00:23.520 of selection bias in early analysis

00:00:23.520 --> 00:00:25.713 of the COVID-19 pandemic.

00:00:27.529 --> 00:00:30.600 You can find the manuscript on line, on arXiv,

00:00:30.600 --> 00:00:34.583 and the slides of this talk is also available on my webpage.

00:00:37.993 --> 00:00:42.100 So, here are the three collaborators,

00:00:42.100 --> 00:00:43.923 involved in this project.

00:00:44.820 --> 00:00:48.030 So Nianqiao is a PHD student at Harvard,

00:00:48.030 --> 00:00:50.180 and we kind of only met online.

00:00:50.180 --> 00:00:53.710 We never met in person, and I sort of created

00:00:53.710 --> 00:00:58.080 a dataset in January, and I wanted some help,

00:00:58.080 --> 00:01:02.730 and somehow she saw this and she said: I could help you.

00:01:02.730 --> 00:01:07.093 And we kind of developed a collaboration.

00:01:08.388 --> 00:01:12.450 And Sergio and Rajen are both, ah,

00:01:12.450 --> 00:01:16.463 lecturers in the Stats Lab in Cambridge.

00:01:17.580 --> 00:01:19.370 And I'd like to thank many, many people

00:01:19.370 --> 00:01:23.050 who have given us very helpful suggestions.

00:01:23.050 --> 00:01:25.623 This is just some of them.

00:01:28.166 --> 00:01:31.970 I'd like to begin with just saying COVID-19

00:01:31.970 --> 00:01:36.810 is personal for everyone, and what I would share

00:01:36.810 --> 00:01:41.810 is partly my story, my personal story with COVID-19.

00:01:44.150 --> 00:01:49.053 So here is a photo of me and my parents,

00:01:49.940 --> 00:01:54.940 taken last September, when I went back to China, 00:01:55.680 --> 00:01:58.130 to see my family.

00:01:58.130 --> 00:02:01.490 So both myself and my parents,

00:02:01.490 --> 00:02:05.720 we all grow up in Wuhan, China.

00:02:05.720 --> 00:02:10.368 And on a sunny day in September, we went to,

00:02:10.368 --> 00:02:12.218 well, this is the Yellow Crane Tower,

00:02:13.102 --> 00:02:15.343 a sort of landmark building in Wuhan.

00:02:16.810 --> 00:02:20.040 And the funny thing is, I think I've never been there,

00:02:20.040 --> 00:02:24.190 on top of the tower, in my entire life.

00:02:24.190 --> 00:02:27.400 And this is actually the first time I went there.

00:02:27.400 --> 00:02:32.360 This is something like if you have a famous local attraction

00:02:32.360 --> 00:02:34.993 for tourists, you actually don't go, as a local.

00:02:38.971 --> 00:02:43.470 And so, on January 23, because the epidemic

00:02:43.470 --> 00:02:48.470 was growing so fast in Wuhan, it started a lockdown.

00:02:51.890 --> 00:02:55.790 So, if we went on top of the Yellow Crane Tower,

00:02:55.790 --> 00:03:00.070 this is what we would see on a typical day,

00:03:00.070 --> 00:03:01.183 before the lockdown.

00:03:02.120 --> 00:03:05.700 And on the right, so, there's sort of what happens

00:03:05.700 --> 00:03:09.800 after the lockdown, and I liked how the journalist

00:03:09.800 --> 00:03:13.490 used sort of this gloomy weather as the background,

00:03:13.490 --> 00:03:16.780 and certainly reflected everybody's mood,

00:03:16.780 --> 00:03:17.963 after the lockdown.

00:03:20.540 --> 00:03:25.540 So, this project begins on January 29.

00:03:25.870 --> 00:03:30.360 So had a conversation with my parents over the phone,

00:03:30.360 --> 00:03:35.360 and they told me that a close relative of ours

00:03:35.420 --> 00:03:39.663 was just diagnosed with, quote/unquote, viral pneumonia.

00:03:41.000 --> 00:03:44.830 So, basically at that point, we all think that must

00:03:44.830 --> 00:03:49.413 be COVID-19, but because there was not enough tests,

00:03:50.700 --> 00:03:54.190 this relative could not get confirmed.

00:03:54.190 --> 00:03:56.350 And this prompted me to start looking

00:03:56.350 --> 00:03:58.363 through the data available at the time.

00:03:59.330 --> 00:04:02.940 But I quickly realized that the epidemiological data

00:04:02.940 --> 00:04:06.670 from Wuhan are very unreliable.

00:04:06.670 --> 00:04:10.160 And here is some anecdotal evidence.

00:04:10.160 --> 00:04:15.160 The first evidence is about inadequate testing.

00:04:15.700 --> 00:04:18.350 So actually this relative of mine could not get

00:04:18.350 --> 00:04:22.010 an RT-PCR test until mid-February,

00:04:22.010 --> 00:04:27.010 and she actually developed symptoms on about January 20.

00:04:29.020 --> 00:04:31.613 So by mid-February, she was already recovering.

00:04:33.200 --> 00:04:35.670 And she took, I think, several tests.

00:04:35.670 --> 00:04:38.280 Her first test was actually negative,

00:04:38.280 --> 00:04:40.330 and a few days later she was tested again,

00:04:40.330 --> 00:04:42.770 and the result came back positive.

00:04:42.770 --> 00:04:46.210 So there's also a lot of false negative tests.

00:04:46.210 --> 00:04:47.623 I think, in general.

00:04:48.720 --> 00:04:52.600 And another problem with the epidemiological data from Wuhan

00:04:52.600 --> 00:04:54.813 is insufficient contact tracing.

00:04:56.040 --> 00:05:01.040 So, her husband, this relative of mine's husband,

00:05:02.710 --> 00:05:07.670 he also showed COVID symptoms, but he quickly recovered

00:05:07.670 --> 00:05:11.467 from that, and in the end he was never tested for COVID.

00:05:16.750 --> 00:05:19.420 So, you can also see the insufficient testing

00:05:19.420 --> 00:05:22.330 from this incidence plot.

00:05:22.330 --> 00:05:27.330 So this is the daily confirmed cases, up until mid-February, 00:05:29.350 --> 00:05:32.960 and this is when the travel ban started,

00:05:32.960 --> 00:05:36.310 or the lockdown started, January 23,

00:05:36.310 --> 00:05:40.540 and on February 12, there was a huge spike

00:05:40.540 --> 00:05:45.540 of over 10,000 cases, much more than the previous few weeks.

00:05:50.200 --> 00:05:54.420 And the reason for that was not suddenly because people

00:05:54.420 --> 00:05:57.040 were infected on that date.

00:05:57.040 --> 00:06:00.790 It's because of a change of diagnostic criterion.

00:06:00.790 --> 00:06:02.373 So before February 12,

00:06:03.910 --> 00:06:08.260 everybody needs to have a positive RT-PCR test

00:06:09.640 --> 00:06:13.080 to be confirmed a COVID-19 case.

00:06:13.080 --> 00:06:16.070 But since February 12, because there,

00:06:16.070 --> 00:06:19.830 the health system in Wuhan was so overwhelmed,

00:06:19.830 --> 00:06:23.440 the government decided to change diagnostic criterion.

00:06:23.440 --> 00:06:28.040 So without RT-PCR tests, you can still be diagnosed

00:06:28.040 --> 00:06:32.613 with COVID-19 if you satisfy several other criteria.

00:06:33.860 --> 00:06:36.760 And this sort of change in diagnostic criteria

00:06:36.760 --> 00:06:40.790 only happened in the Hubei Province

00:06:40.790 --> 00:06:42.363 and not elsewhere in China.

00:06:44.590 --> 00:06:49.590 So a solution, if we like to avoid these problems

00:06:49.950 --> 00:06:54.530 with data from Wuhan, so one clever solution

00:06:54.530 --> 00:06:57.990 is to use cases that are reported from, sorry,

00:06:57.990 --> 00:06:59.253 exported from Wuhan.

00:07:00.670 --> 00:07:02.680 So this has two benefits.

00:07:02.680 --> 00:07:05.150 First of all, testing and contact tracing

00:07:05.150 --> 00:07:07.873 were quite intensive in other locations.

00:07:08.820 --> 00:07:12.750 So, it's reasonable to expect that a lot of the bias

00:07:12.750 --> 00:07:16.450 due to sort of under-ascertainment will be less severe

00:07:16.450 --> 00:07:18.193 if we use data from elsewhere.

00:07:20.360 --> 00:07:25.153 And also, many locations, particularly in some cities

00:07:25.990 --> 00:07:29.713 in China, published detailed case reports,

00:07:31.000 --> 00:07:33.650 instead of just case counts.

00:07:33.650 --> 00:07:36.420 And if you look at these detailed case reports there are

00:07:36.420 --> 00:07:41.420 a lot of information that can be used for inference.

00:07:44.160 --> 00:07:46.203 This is not our idea.

00:07:47.170 --> 00:07:51.110 And I think one of the, at least one of the first persons

00:07:51.110 --> 00:07:56.110 to use this design was a report from Neil Ferguson's group

00:07:56.610 --> 00:07:58.233 in Imperial College, London,

00:07:59.250 --> 00:08:02.820 and they published a report on January 17,

00:08:02.820 --> 00:08:07.280 and what it did was a simple sort of division of the number

00:08:07.280 --> 00:08:11.040 of cases detected internationally, over the number

00:08:11.040 --> 00:08:14.053 of people traveled from Wuhan, internationally.

00:08:15.280 --> 00:08:17.840 And they found that it could be

00:08:17.840 --> 00:08:22.043 over 1,700 cases by January 17, in Wuhan.

00:08:25.955 --> 00:08:30.170 So, I started this on January 29,

00:08:30.170 --> 00:08:35.170 and within about two weeks, managed to put something online.

00:08:36.520 --> 00:08:39.940 Which we also used internationally confirmed cases

00:08:39.940 --> 00:08:41.783 to estimate epidemic growth.

00:08:43.896 --> 00:08:48.200 And what we used were 46 coronavirus cases

00:08:48.200 --> 00:08:53.200 who traveled from Wuhan and then were subsequently confirmed

00:08:53.280 --> 00:08:57.103 in six Asian countries and regions.

00:08:58.560 --> 00:09:02.410 And the main result was that the epidemic was doubling

00:09:02.410 --> 00:09:04.540 in size every 2.9 days.

00:09:06.120 --> 00:09:09.760 And we used the Bayesian analysis, and the 95 percent

00:09:09.760 --> 00:09:11.920 critical interval was two to 4.1.

00:09:13.710 --> 00:09:17.410 And of course, when I was writing this article,

00:09:17.410 --> 00:09:21.970 I was mostly just working on this dataset that we collected,

00:09:21.970 --> 00:09:26.620 very hard and (muttering), thinking about what model

00:09:26.620 --> 00:09:29.113 is suitable for this kind of data.

00:09:30.250 --> 00:09:34.023 And just before I posted this pre-print,

00:09:34.023 --> 00:09:37.860 I realized there was a similar article

00:09:37.860 --> 00:09:42.860 that already published in The Lancet, on January 31.

00:09:44.730 --> 00:09:49.557 And what's really puzzling is they used almost the same data

00:09:50.970 --> 00:09:54.110 and very similar models, but somehow reached

00:09:54.110 --> 00:09:56.603 completely different conclusions.

00:09:57.720 --> 00:10:01.840 So they used data from December 31 to January 28,

00:10:01.840 --> 00:10:05.050 that are exported from Wuhan internationally.

00:10:05.050 --> 00:10:06.620 And they would like to infer the number

00:10:06.620 --> 00:10:08.303 of infections in Wuhan.

00:10:09.570 --> 00:10:11.630 And one of the main results,

00:10:11.630 --> 00:10:16.480 which was this epidemic doubling time, was 6.4 days,

00:10:16.480 --> 00:10:21.040 and the 95 percent critical interval was 5.8 to 7.1.

00:10:21.040 --> 00:10:24.060 So that's drastically different from ours.

00:10:24.060 --> 00:10:26.290 So again, ours was 2.7, within two to four,

00:10:28.963 --> 00:10:30.380 and this was 6.4.

00:10:32.880 --> 00:10:35.710 And this is talking about the doubling time.

00:10:35.710 --> 00:10:39.780 So the doubling time of six days versus three days,

00:10:39.780 --> 00:10:42.870 that's sort of really, really different.

00:10:42.870 --> 00:10:45.480 And the confidence intervals, the credible intervals

00:10:45.480 --> 00:10:46.713 didn't even overlap.

00:10:48.870 --> 00:10:51.073 So I was really puzzled by this.

00:10:52.450 --> 00:10:56.713 And before I tell you what I think,

00:10:57.710 --> 00:11:00.790 how the Lancet paper got it wrong,

00:11:00.790 --> 00:11:02.720 I'd like to just show you this plot.

00:11:02.720 --> 00:11:05.250 You probably have seen this many times before,

00:11:05.250 --> 00:11:10.250 in news articles, which is just sort of a logarithm

00:11:10.350 --> 00:11:15.337 of the total cases versus the days, ah,

00:11:16.384 --> 00:11:20.500 or some time, zero, for each country.

00:11:21.410 --> 00:11:25.530 And what you see is for both the total number of cases

00:11:25.530 --> 00:11:27.080 and the total number of deaths,

00:11:29.200 --> 00:11:34.200 it sort of grew about 100-fold in the first 20 days.

00:11:35.110 --> 00:11:36.360 At least among these countries

00:11:36.360 --> 00:11:39.743 that were most hard-hit by COVID-19.

00:11:41.540 --> 00:11:44.817 And if you just use that as a variable of estimate,

00:11:44.817 --> 00:11:47.497 of the doubling time, that corresponds

00:11:47.497 --> 00:11:50.247 to a doubling time of three days.

00:11:51.889 --> 00:11:56.039 Of course, this is sort of very kind of anecdotal,

00:11:56.039 --> 00:12:01.000 because this data were not collected in a very careful way,

00:12:01.000 --> 00:12:03.530 and the amount of cases were not reported,

00:12:03.530 --> 00:12:05.880 but this is just to show you that perhaps

00:12:06.967 --> 00:12:11.967 the doubling time of 6.4 days was a bit just, too long.

00:12:14.184 --> 00:12:16.920 So, towards the end of the talk,

00:12:16.920 --> 00:12:19.760 I'll tell you what we think led

00:12:20.890 --> 00:12:22.663 to these very different results.

00:12:23.830 --> 00:12:28.830 Just some spoilers, so the crucial difference

00:12:29.710 --> 00:12:32.870 is that the Lancet study actually did not

00:12:32.870 --> 00:12:37.033 take into account the travel ban on January 23.

00:12:37.890 --> 00:12:39.390 And that actually had a very, 00:12:39.390 --> 00:12:44.390 very circumstantial selection effect on the data.

00:12:44.510 --> 00:12:48.263 And this will be made precise later on in the talk.

00:12:52.770 --> 00:12:54.400 So, for the rest of the talk,

00:12:54.400 --> 00:12:56.960 I'll first give you an overview of selection bias.

00:12:56.960 --> 00:13:00.760 So no math, just sort of an outline of what kind

00:13:00.760 --> 00:13:04.570 of selection bias you could encounter in COVID-19 studies.

00:13:04.570 --> 00:13:08.460 Then I'll talk about how we sort of overcome them,

00:13:08.460 --> 00:13:11.790 by sort of collecting the dataset very carefully

00:13:11.790 --> 00:13:15.093 and building a model very carefully.

00:13:16.810 --> 00:13:19.180 And then I'll talk about why

00:13:20.070 --> 00:13:22.137 the Lancet study I just mentioned

00:13:22.137 --> 00:13:25.513 and some other early analysis were severely biased.

00:13:26.460 --> 00:13:29.150 If there is time, I will tell you a little bit

00:13:29.150 --> 00:13:31.733 about our Bayesian nonparametric model.

00:13:33.160 --> 00:13:35.970 And then I'll give you some lessons

00:13:35.970 --> 00:13:38.403 I learned from this work.

00:13:40.440 --> 00:13:42.290 So selection bias.

00:13:42.290 --> 00:13:45.590 So we identified at least five kinds

00:13:45.590 --> 00:13:48.790 of selection bias in COVID-19 studies.

00:13:48.790 --> 00:13:52.590 So the first one is due to under-ascertainment.

00:13:52.590 --> 00:13:55.820 So this may occur if symptomatic patients

00:13:55.820 --> 00:13:59.253 do not seek healthcare, or could not be diagnosed.

00:14:00.090 --> 00:14:04.130 So essentially, all studies using cases confirmed

00:14:04.130 --> 00:14:06.763 when testing is insufficient,

00:14:08.060 --> 00:14:10.550 would be susceptible to this kind of bias.

00:14:10.550 --> 00:14:12.693 And there is no cure to this.

00:14:13.830 --> 00:14:18.830 It may lead to varied kind of direction and magnitude

00:14:20.640 --> 00:14:25.573 of bias, and basically what we can do is to, 00:14:27.450 --> 00:14:31.930 to think about a clever design to avoid this problem,

00:14:31.930 --> 00:14:36.807 to focus on locations where the testing is intensive.

00:14:41.960 --> 00:14:46.283 The second bias is due to non-random sample selection.

00:14:47.690 --> 00:14:50.760 So, basically this means that the cases included

00:14:50.760 --> 00:14:53.713 in the study are not representative of the population.

00:14:56.080 --> 00:15:01.080 So this essentially applies to all studies,

00:15:02.800 --> 00:15:05.910 because detailed information about COVID-19 cases

00:15:05.910 --> 00:15:09.963 are usually sparse; they're not always published.

00:15:11.280 --> 00:15:14.100 But especially for studies that do not have a clear

00:15:14.100 --> 00:15:18.930 inclusion criterion, and if they just sort of simply

00:15:18.930 --> 00:15:23.930 collect data out of convenience, then there could be

00:15:24.770 --> 00:15:27.763 a lot of non-random sample selection bias.

00:15:29.830 --> 00:15:33.460 And again, statistical models are not really gonna help you

00:15:33.460 --> 00:15:34.880 with this kind of bias.

00:15:34.880 --> 00:15:39.880 You'd use, you'd follow some protocol for data collection,

00:15:40.040 --> 00:15:44.060 and you would exclude some data that do not meet

00:15:44.060 --> 00:15:45.723 the sample inclusion criterion.

00:15:46.790 --> 00:15:51.790 Even when that may, leads to inefficient estimates.

00:15:57.490 --> 00:16:00.020 The third bias is due to the travel ban.

00:16:00.020 --> 00:16:04.343 This is kind of my spoiler about that Lancet study.

00:16:05.770 --> 00:16:09.150 So basically, outbound travel from Wuhan

00:16:09.150 --> 00:16:14.150 to anywhere else was banned from January 23 to April eight.

00:16:15.930 --> 00:16:20.930 So if the study analyzed cases exported from Wuhan,

00:16:21.034 --> 00:16:26.034 then they're susceptible to this selection defect.

00:16:26.740 --> 00:16:30.660 And this would usually lead to underestimation

00:16:30.660 --> 00:16:34.919 of epidemic growth, and the reason is that, so,

00:16:34.919 --> 00:16:36.900 the epidemic is growing very fast,

00:16:36.900 --> 00:16:40.535 but then you essentially can't observe cases

00:16:40.535 --> 00:16:44.490 that were supposed to leave Wuhan after January 23.

00:16:44.490 --> 00:16:46.770 So if you just wait for a long time,

00:16:46.770 --> 00:16:50.450 and then look at the epidemic curve among the cases

00:16:50.450 --> 00:16:54.580 exported from Wuhan, it may appear that, ah,

00:16:54.580 --> 00:16:57.650 it sort of dies down a little bit,

00:16:57.650 --> 00:17:01.010 but that's not because of the epidemic being controlled.

00:17:01.010 --> 00:17:02.713 That's because of the travel ban.

00:17:03.770 --> 00:17:07.620 And fortunately this bias, you can correct for it

00:17:07.620 --> 00:17:10.100 by deriving some likelihood function

00:17:10.100 --> 00:17:13.103 tailored for the travel restrictions.

00:17:15.390 --> 00:17:19.580 The fourth bias is ignoring, is due to ignoring

00:17:19.580 --> 00:17:24.580 the epidemic growth, and basically if you think about people

00:17:24.670 --> 00:17:29.110 who have been in Wuhan before January 23,

00:17:29.110 --> 00:17:31.310 they're much more likely to be infected

00:17:31.310 --> 00:17:36.310 towards the end of their exposure period than early,

00:17:36.920 --> 00:17:40.023 and that's because the epidemic was growing quickly.

00:17:41.730 --> 00:17:44.940 So, there are many studies, or I should say

00:17:44.940 --> 00:17:48.100 there are several studies of the incubation period

00:17:48.100 --> 00:17:52.210 that simply treat infections as uniformly distributed

00:17:52.210 --> 00:17:55.963 over the patients' exposure period to Wuhan.

00:17:56.860 --> 00:17:59.390 And this will lead to overestimation

00:17:59.390 --> 00:18:01.630 of the incubation period.

00:18:01.630 --> 00:18:03.750 Because actually, the infection time is much,

00:18:03.750 --> 00:18:08.750 much closer to sort of the end of their exposure.

00:18:11.110 --> 00:18:14.660 And this is also a bias that can be corrected for,

00:18:14.660 --> 00:18:19.143 by doing statistical analysis carefully.

00:18:20.870 --> 00:18:25.420 The fifth and last bias is due to right-truncation.

00:18:25.420 --> 00:18:29.710 So this happens in early analysis because,

00:18:29.710 --> 00:18:34.603 to sort of win time to battle for this epidemic,

00:18:35.890 --> 00:18:38.490 and to publish sort of fast.

00:18:38.490 --> 00:18:42.800 So as you all know, there's a race for publications

00:18:42.800 --> 00:18:47.500 about COVID-19; a lot of people sort of truncated

00:18:47.500 --> 00:18:50.663 the dataset before a certain time,

00:18:51.520 --> 00:18:53.940 but by that time the epidemic maybe

00:18:53.940 --> 00:18:56.393 was still quickly growing or evolving.

00:18:58.076 --> 00:19:01.340 And this could lead to some right-truncation bias.

00:19:03.476 --> 00:19:06.660 And this generally would lead to underestimation

00:19:06.660 --> 00:19:08.263 of the incubation period.

00:19:09.740 --> 00:19:13.060 So this is, so incubation period, I forgot to mention,

00:19:13.060 --> 00:19:18.060 is just the time between infection to showing symptoms.

00:19:19.710 --> 00:19:22.420 So, right-truncation would lead to underestimation

00:19:22.420 --> 00:19:26.090 of incubation period, because people with longer

00:19:26.090 --> 00:19:31.090 incubation period may not have showed symptoms

00:19:31.110 --> 00:19:35.563 by the time that these datasets were collected.

00:19:37.980 --> 00:19:42.980 So the solution to this is we need to both collect cases

00:19:44.870 --> 00:19:47.570 that meet the selection criterion, and continue

00:19:47.570 --> 00:19:52.570 that data collection until a sufficiently long time.

00:19:54.370 --> 00:19:58.820 Or, you derive some likelihood function to correct

00:19:58.820 --> 00:20:00.250 for the right-truncation.

00:20:00.250 --> 00:20:02.573 So we'll go over this later.

00:20:03.760 --> 00:20:05.623 So just to recap,

00:20:07.030 --> 00:20:10.640 so on a very high level, there are at least five

00:20:10.640 --> 00:20:14.750 kinds of biases in COVID-19 analysis.

00:20:14.750 --> 00:20:19.750 And if you read sort of article pre-prints or use articles,

00:20:19.990 --> 00:20:23.680 I think you will find some kind, I mean,

00:20:23.680 --> 00:20:27.833 some resemblance of these biases in many studies.

00:20:30.410 --> 00:20:34.413 And the keys to avoid selection bias is basically,

00:20:35.310 --> 00:20:38.070 I mean, this is simple in words,

00:20:38.070 --> 00:20:40.260 but you just do everything carefully.

00:20:40.260 --> 00:20:42.050 You design the study carefully,

00:20:42.050 --> 00:20:45.030 and collect the sample carefully,

00:20:45.030 --> 00:20:47.270 and analyze the data carefully.

00:20:47.270 --> 00:20:51.310 But the reality, of course, is not that simple.

00:20:51.310 --> 00:20:54.770 And what I will show below, it's an example

00:20:54.770 --> 00:20:59.770 of our try to eliminate or to reduce selection bias,

00:21:02.010 --> 00:21:03.333 as much as possible.

00:21:05.950 --> 00:21:10.120 So, let me tell you the dataset we collected.

00:21:10.120 --> 00:21:15.120 So we found 14 locations in Asia,

00:21:16.340 --> 00:21:20.650 some are international, so Japan, South Korea, Taiwan,

00:21:20.650 --> 00:21:23.050 Hong Kong, Macau, Singapore.

00:21:23.050 --> 00:21:26.710 Some are sort of in mainland China.

00:21:26.710 --> 00:21:29.793 So there are several cities in mainland China.

00:21:30.720 --> 00:21:35.720 So all these locations have published detailed case reports

00:21:35.740 --> 00:21:37.703 from their first local case.

00:21:39.730 --> 00:21:43.480 So, most of the Chinese locations, I mean,

00:21:43.480 --> 00:21:46.280 they were done with the first wave of the epidemic

00:21:46.280 --> 00:21:47.493 by the end of February.

00:21:49.250 --> 00:21:54.240 So Japan, Korea and Singapore saw some resurgence

00:21:54.240 --> 00:21:57.410 of the epidemic later on, and eventually,

00:21:57.410 --> 00:22:02.340 they did not publish detailed case reports.

00:22:02.340 --> 00:22:07.340 But for our purposes, these locations all published

00:22:07.370 --> 00:22:10.850 detailed reports before mid-February,

00:22:10.850 --> 00:22:15.330 and that's about three weeks after the lockdown of Wuhan.

00:22:15.330 --> 00:22:18.980 So it's pretty much enough to find out

00:22:18.980 --> 00:22:21.203 all the Wuhan exported cases.

00:22:24.360 --> 00:22:27.993 So just to give you a sense of the kind of data

00:22:27.993 --> 00:22:31.790 that we collected, this is sort of all

00:22:31.790 --> 00:22:35.997 the important columns in the dataset,

00:22:35.997 --> 00:22:40.283 and the particularly important columns are marked in red.

00:22:42.340 --> 00:22:46.653 So, we collected, there was a case ID,

00:22:48.600 --> 00:22:53.600 where the case lived, the gender, the age,

00:22:54.270 --> 00:22:57.220 whether they had known epidemiological contact

00:22:57.220 --> 00:23:01.930 with other confirmed cases, whether it has

00:23:01.930 --> 00:23:04.563 known relationship with other confirmed cases.

00:23:06.540 --> 00:23:09.250 This is sort of an interesting column

00:23:09.250 --> 00:23:14.250 that basically we like to find out what cases were

00:23:15.200 --> 00:23:20.200 exported from Wuhan, but that's, of course, not recorded.

00:23:20.210 --> 00:23:25.210 I mean you can only infer that from what has been published.

00:23:26.560 --> 00:23:28.440 So this is an attempt to do that.

00:23:28.440 --> 00:23:31.540 So this column, outside column means that,

00:23:31.540 --> 00:23:35.120 whether we think the data collector thinks

00:23:35.120 --> 00:23:37.573 this case is transmitted outside Wuhan.

00:23:38.810 --> 00:23:42.900 So most of the time, this is relatively easy to fill.

00:23:44.700 --> 00:23:47.243 For example, if you've never been to Wuhan,

00:23:47.243 --> 00:23:49.870 this entry must be yes.

00:23:49.870 --> 00:23:52.260 But sometimes, this can be a little bit tricky.

00:23:52.260 --> 00:23:56.390 For example, this person, the fifth case in Hong Kong,

00:23:56.390 --> 00:24:00.080 is the husband of the fourth case in Hong Kong,

00:24:00.080 --> 00:24:03.053 and they traveled together from Wuhan to Hong Kong.

00:24:04.470 --> 00:24:09.470 So it's unclear if this case is transmitted

00:24:11.000 --> 00:24:14.163 in or outside Wuhan, so we put a ”likely” there.

00:24:15.640 --> 00:24:20.430 And the other information are some dates,

00:24:20.430 --> 00:24:24.803 the beginning of stay in Wuhan, the end of stay in Wuhan,

00:24:25.760 --> 00:24:28.780 the period of exposure, which would equal to

00:24:30.140 --> 00:24:32.629 beginning to the end of stay in Wuhan,

00:24:32.629 --> 00:24:35.310 for Wuhan exported cases,

00:24:35.310 --> 00:24:38.373 but can be different for other cases.

00:24:40.530 --> 00:24:44.120 When the person, when the case arrived at a final location

00:24:44.120 --> 00:24:47.520 where they are confirmed a COVID-19 case.

00:24:47.520 --> 00:24:49.563 When the person showed symptoms.

00:24:51.250 --> 00:24:53.780 When did they first go to a hospital,

00:24:53.780 --> 00:24:58.580 and when were they confirmed a COVID-19 case.

00:24:58.580 --> 00:25:03.580 So we collected about 1,400 cases with all this information.

00:25:04.690 --> 00:25:09.347 And overall, I think our dataset was relatively high

00:25:11.350 --> 00:25:16.350 in quality, and most of the cases had known symptom onset

00:25:18.370 --> 00:25:21.833 dates; only nine percent of them have that entry missing.

00:25:26.560 --> 00:25:27.683 So,

00:25:29.600 --> 00:25:33.200 so one important step after this is to find out

00:25:33.200 --> 00:25:37.210 which cases are actually exported from Wuhan.

00:25:37.210 --> 00:25:41.210 So I've been using this terminology from the beginning

00:25:41.210 --> 00:25:45.360 of the talk, but basically the case is Wuhan exported

00:25:45.360 --> 00:25:49.800 if they are infected, if they were infected in Wuhan.

00:25:49.800 --> 00:25:51.513 And then confirmed elsewhere.

00:25:53.030 --> 00:25:58.000 So we had a sample selection criterion

00:25:58.000 --> 00:26:02.700 to discern a Wuhan exported case.

00:26:02.700 --> 00:26:04.643 I'm not going to go over it in detail,

00:26:06.110 --> 00:26:08.500 but basically the principle we followed

00:26:08.500 --> 00:26:13.500 is that we would only consider a case as Wuhan exported

00:26:14.200 --> 00:26:18.770 if it passed a beyond a reasonable doubt test.

00:26:18.770 --> 00:26:21.360 So basically, if we think there is a reasonable doubt

00:26:21.360 --> 00:26:24.603 that the case could be infected elsewhere,

00:26:25.760 --> 00:26:30.023 then we would say: let's exclude that from the dataset.

00:26:31.100 --> 00:26:34.603 So this eventually gives us 378 cases.

00:26:38.880 --> 00:26:41.333 Next I'm gonna talk about the model we used.

00:26:45.810 --> 00:26:48.280 So the model is called: BETS.

00:26:48.280 --> 00:26:52.720 It's named after sort of four key epidemiological events.

00:26:52.720 --> 00:26:56.170 The beginning of exposure, the end of exposure,

00:26:56.170 --> 00:27:00.690 time of transmission, which is usually unobserved,

00:27:00.690 --> 00:27:02.813 and the time of symptom onset, S.

00:27:06.240 --> 00:27:11.240 So what we will do below is we'll first define the support

00:27:12.600 --> 00:27:15.633 of these variables, so we call that P.

00:27:17.120 --> 00:27:22.120 Which is basically represents the Wuhan exposed population.

00:27:24.240 --> 00:27:26.823 So this is the population we would like to study.

00:27:28.120 --> 00:27:31.420 We will then construct a generative model

00:27:31.420 --> 00:27:33.133 for these random variables.

00:27:33.980 --> 00:27:37.483 Basically, for everyone in the Wuhan exposed population.

00:27:38.900 --> 00:27:42.060 Then, to consider the sample selection,

00:27:42.060 --> 00:27:45.570 we'll define a sample selection set, D,

00:27:45.570 --> 00:27:49.193 that corresponds to cases that are exported from Wuhan.

00:27:50.930 --> 00:27:53.770 Then finally we will derive likelihood functions

00:27:53.770 --> 00:27:55.973 to adjust for the sample selection.

00:27:56.900 --> 00:28:00.630 So essentially, what we're trying to infer is

00:28:00.630 --> 00:28:05.140 the disease dynamics in the population, P,

00:28:05.140 --> 00:28:09.543 but we only have data from this sample, D.

00:28:10.990 --> 00:28:13.930 So here's a lot of work that needs to be done

00:28:13.930 --> 00:28:16.163 to correct for that sample selection.

00:28:19.740 --> 00:28:23.147 So intuitively, this population P are just all people

00:28:23.147 --> 00:28:28.147 who have stayed in Wuhan, between December first

00:28:29.410 --> 00:28:34.410 and January 24, so anyone who has been in Wuhan

00:28:35.540 --> 00:28:38.660 for maybe even just a few hours,

00:28:38.660 --> 00:28:43.660 they would count as someone exposed to Wuhan.

00:28:45.320 --> 00:28:50.320 And I'm going to make some conventions to simplify

00:28:50.900 --> 00:28:52.963 this set, P, a little bit.

00:28:53.900 --> 00:28:57.913 So B equals to zero has a special meaning.

00:28:59.120 --> 00:29:02.120 So, so zero is the time zero,

00:29:02.120 --> 00:29:05.580 which is 12 AM of December one.

00:29:05.580 --> 00:29:10.580 And it means that they actually started their stay in Wuhan

00:29:10.810 --> 00:29:14.653 before time zero, so they live in Wuhan essentially.

00:29:15.700 --> 00:29:20.390 And B greater than zero means these other cases

00:29:21.300 --> 00:29:25.470 visited Wuhan sometime in the middle of this period,

00:29:25.470 --> 00:29:26.913 and then they left Wuhan.

00:29:29.060 --> 00:29:33.450 So E equals to infinity means that the case did not arrive

00:29:33.450 --> 00:29:35.860 in the 14 locations we are considering

00:29:35.860 --> 00:29:38.893 before this lockdown time, L.

00:29:40.890 --> 00:29:42.370 So for the purpose of our study,

00:29:42.370 --> 00:29:45.150 we did not need to differentiate between people who

00:29:45.150 --> 00:29:48.710 have always stayed in Wuhan past time L,

00:29:48.710 --> 00:29:51.900 or people who left Wuhan before time L,

00:29:51.900 --> 00:29:54.090 but went to a different location

00:29:55.100 --> 00:29:57.103 other than the ones we are considering.

00:29:58.400 --> 00:30:02.420 So T equals to infinity means that the cases

00:30:02.420 --> 00:30:05.840 were not infected during their stay in Wuhan.

00:30:05.840 --> 00:30:08.240 So this could be infected outside Wuhan,

00:30:08.240 --> 00:30:10.573 or it could be they were never infected.

00:30:11.830 --> 00:30:15.950 And S equals to infinity means that the case

00:30:15.950 --> 00:30:18.680 did not show symptoms of COVID-19,

00:30:18.680 --> 00:30:22.040 and it can simply be, they were never infected.

00:30:22.040 --> 00:30:27.040 Or the case was actually tested positive for COVID-19,

00:30:27.490 --> 00:30:31.933 but never showed symptoms, so it's, they're asymptomatic.

00:30:33.710 --> 00:30:37.870 So under these conventions, this is the set,

00:30:37.870 --> 00:30:41.420 this is the support for this population, P.

00:30:41.420 --> 00:30:44.030 So B is between zero and L,

00:30:44.030 --> 00:30:47.380 E is between B and L or infinity,

00:30:47.380 --> 00:30:50.620 T is between B and E, which means that they are,

00:30:50.620 --> 00:30:53.530 in fact, in Wuhan, or infinity.

00:30:53.530 --> 00:30:55.627 And S is between T and infinity,

00:30:55.627 --> 00:30:57.933 and S can be equal to infinity.

00:31:00.440 --> 00:31:03.133 So now we have defined this population, P.

00:31:04.400 --> 00:31:08.463 And now let's look at a general model,

00:31:09.400 --> 00:31:13.403 a data-generated model for this population.

00:31:15.020 --> 00:31:18.270 So, by the basic law of probability,

00:31:18.270 --> 00:31:21.160 we could decompose the joint distribution

00:31:21.160 --> 00:31:25.460 of BETS into these four, and the first two

00:31:25.460 --> 00:31:27.210 are the distribution of B and E.

00:31:27.210 --> 00:31:29.520 They are related to travel.

00:31:29.520 --> 00:31:32.370 The second one, sorry, the third one is the distribution

00:31:32.370 --> 00:31:34.790 of T given B and E.

00:31:34.790 --> 00:31:37.920 So that's about the disease transmission.

00:31:37.920 --> 00:31:40.600 And the last one is the distribution of S,

00:31:40.600 --> 00:31:44.583 given BET, and that's related to disease progression.

00:31:46.900 --> 00:31:49.610 So we need to make two basic assumptions,

00:31:49.610 --> 00:31:54.490 and they are important because we would like to infer

00:31:54.490 --> 00:31:56.500 what's going on in the population P,

00:31:56.500 --> 00:32:01.500 from the sample T, from these Wuhan exported cases.

00:32:01.960 --> 00:32:05.290 So we need to sort of make assumptions

00:32:05.290 --> 00:32:08.320 so we can make that extrapolation.

00:32:08.320 --> 00:32:10.510 So the first assumption, we assume it's about

00:32:10.510 --> 00:32:14.100 this disease transmission, and it basically means

00:32:14.100 --> 00:32:17.163 that the disease transmission is independent of travel.

00:32:18.350 --> 00:32:22.490 So there is a basic sort of function that's independent

00:32:22.490 --> 00:32:25.803 of the travel that's growing over time.

00:32:27.000 --> 00:32:31.293 And then there's the rest of the points mass at infinity.

00:32:32.790 --> 00:32:36.840 This T function, so, it will appear later on.

00:32:36.840 --> 00:32:38.883 It's the epidemic growth function.

00:32:40.160 --> 00:32:43.240 The second assumption is that the disease progression

00:32:43.240 --> 00:32:45.203 is also independent of travel.

00:32:46.420 --> 00:32:49.470 So, what's assumed here is basically

00:32:49.470 --> 00:32:54.350 that there is one minus mu of the infections,

00:32:56.460 --> 00:33:00.390 that are asymptomatic in that they didn't show symptoms.

00:33:00.390 --> 00:33:03.180 The amount of people who showed symptoms,

00:33:03.180 --> 00:33:07.060 the incubation period, which is just S minus T,

00:33:07.060 --> 00:33:09.233 follows this distribution, H.

00:33:10.870 --> 00:33:13.890 Okay, so H is the density of the incubation period,

00:33:13.890 --> 00:33:15.783 for symptomatic cases.

00:33:17.360 --> 00:33:21.223 And this whole distribution does not depend on B and E.

00:33:23.880 --> 00:33:26.320 So these are sort of the two basic assumptions

00:33:26.320 --> 00:33:28.163 that we relied on.

00:33:29.920 --> 00:33:32.490 There are two further parametric assumptions

00:33:32.490 --> 00:33:37.410 that were useful to simplify the interpretation,

00:33:37.410 --> 00:33:39.003 but they can be relaxed.

00:33:41.390 --> 00:33:45.040 So the next, one assumption is the epidemic

00:33:45.040 --> 00:33:49.253 was growing exponentially before the lockdown.

00:33:51.040 --> 00:33:53.600 And then that, the other assumption is that the incubation

00:33:53.600 --> 00:33:58.020 period is gamma-distributed, okay?

00:33:58.020 --> 00:34:02.233 So there's some parameters, kappa, R and alpha, beta.

00:34:04.650 --> 00:34:08.233 So, don't worry about nuisance parameter mu,

00:34:09.214 --> 00:34:11.990 which is the proportion of asymptomatic cases.

00:34:11.990 --> 00:34:15.720 And kappa, which is some baseline transmission.

00:34:15.720 --> 00:34:19.183 So it turns out that they would be canceled

00:34:19.183 --> 00:34:22.530 in the likelihood function, so they won't appear

00:34:22.530 --> 00:34:24.243 in the likelihood function.

00:34:25.480 --> 00:34:28.300 And (muttering) these parametric assumptions,

00:34:28.300 --> 00:34:32.290 they can be relaxed and they will be relaxed

00:34:32.290 --> 00:34:36.713 in the Bayesian parametric analysis, if I can get to there.

00:34:38.370 --> 00:34:42.440 But essentially, these are very useful assumptions

00:34:42.440 --> 00:34:47.403 that allow us to derive formulas explicitly.

00:34:50.345 --> 00:34:53.653 So I have covered the full data BETS model

00:34:55.650 --> 00:34:57.820 for the population P.

00:34:57.820 --> 00:35:01.233 Now we need to look at what we can observe.

00:35:02.200 --> 00:35:06.950 So what we can observe are people in B

00:35:06.950 --> 00:35:10.873 that satisfy three additional restrictions.

00:35:11.760 --> 00:35:14.980 The first restriction is that the transmission

00:35:14.980 --> 00:35:19.980 is between their exposure to Wuhan.

00:35:22.790 --> 00:35:26.870 The second restriction is that the case needs to leave

00:35:26.870 --> 00:35:28.893 Wuhan before the lockdown time, L.

00:35:30.730 --> 00:35:33.460 The third restriction is that the case

00:35:33.460 --> 00:35:35.263 needs to show symptoms.

00:35:36.160 --> 00:35:37.853 So S is less than infinity.

00:35:39.380 --> 00:35:41.260 So some of the locations we considered

00:35:41.260 --> 00:35:45.830 did report a few asymptomatic cases, but overall,

00:35:45.830 --> 00:35:50.140 asymptomatic ascertainment was very inconsistent.

00:35:50.140 --> 00:35:53.763 So we only considered cases who showed symptoms.

00:35:56.140 --> 00:36:01.140 So this gives us the set of samples

00:36:01.293 --> 00:36:03.883 that we can observe in our data.

00:36:09.022 --> 00:36:14.022 So, which likelihood function should we use?

00:36:14.500 --> 00:36:16.810 For a moment, let's just pretend that the time

00:36:16.810 --> 00:36:19.830 of transmission, T, is observed.

00:36:19.830 --> 00:36:24.830 So if we had samples, ID samples from the population, P,

00:36:25.100 --> 00:36:28.800 then we could just use this product of the density

00:36:28.800 --> 00:36:33.800 of BETS as a likelihood function.

00:36:33.820 --> 00:36:36.350 But this is not something we should use,

00:36:36.350 --> 00:36:39.590 because we actually don't have samples from P.

00:36:39.590 --> 00:36:44.590 We have samples from D, so what we should do is to condition

00:36:45.620 --> 00:36:50.620 on the selection set, D, and use this likelihood function,

00:36:52.300 --> 00:36:56.310 which is basically just the density divided by the

00:36:56.310 --> 00:37:01.310 probability that someone is selected in this set, D.

00:37:03.950 --> 00:37:06.543 Okay, this is called unconditional likelihood,

00:37:07.420 --> 00:37:10.603 to contrast with the conditional likelihood.

00:37:11.450 --> 00:37:14.044 So, in unconditional likelihood,

00:37:14.044 --> 00:37:18.160 we consider the joint distribution of B, E, T, and S.

00:37:18.160 --> 00:37:20.200 But in the conditional likelihood,

00:37:20.200 --> 00:37:24.900 we consider the conditional distribution of T and S,

00:37:24.900 --> 00:37:26.350 given B and E.

00:37:26.350 --> 00:37:29.030 So this is the conditional distribution of the disease

00:37:29.030 --> 00:37:32.440 transmission and progression, given the travel.

00:37:32.440 --> 00:37:34.543 So this treats travel as fixed.

00:37:35.430 --> 00:37:37.530 So to compute this conditional likelihood,

00:37:38.459 --> 00:37:42.113 we need further conditions on B and E, okay?

00:37:48.210 --> 00:37:52.060 But in reality, the time of transmission, T, is unobserved,

00:37:52.060 --> 00:37:55.330 so we cannot directly use the likelihood function,

00:37:55.330 --> 00:38:00.330 as on the last slide, so one possibility is to treat T

00:38:01.100 --> 00:38:05.153 as a latent variable and use, for example, an EM algorithm.

00:38:06.700 --> 00:38:09.823 The way we chose is to use an integrated likelihood.

00:38:11.372 --> 00:38:13.510 That just sort of marginalized

00:38:13.510 --> 00:38:17.343 over this unobserved variable, T.

00:38:19.200 --> 00:38:22.710 So, the unconditional likelihood is the product

00:38:22.710 --> 00:38:26.250 over the cases of the integral

00:38:26.250 --> 00:38:29.733 of the density function over T.

00:38:31.070 --> 00:38:34.380 And the conditional likelihood is just a product

00:38:34.380 --> 00:38:39.193 of the integral of the conditional distribution of T and S,

00:38:40.210 --> 00:38:41.043 over T.

00:38:44.750 --> 00:38:48.690 So, the reason we sort of considered both

00:38:48.690 --> 00:38:51.160 the unconditional likelihood and conditional likelihood

00:38:51.160 --> 00:38:55.050 is that the unconditional likelihood is a little bit

00:38:55.050 --> 00:39:00.000 more efficient, because it also uses information

00:39:00.000 --> 00:39:05.000 in this density, BE, given your selected.

00:39:05.840 --> 00:39:08.090 So that contains a little bit of information.

00:39:09.228 --> 00:39:12.040 But a conditional likelihood is more robust.

00:39:12.040 --> 00:39:17.040 So, because it does not need to specify how people traveled,

00:39:17.940 --> 00:39:22.163 so it is robust to misspecifying those distributions.

00:39:24.000 --> 00:39:28.883 So I'll stop here and take any questions up to now.

00:39:35.620 --> 00:39:37.283 Is this clear to everyone?

00:39:39.570 --> 00:39:41.663 If so, I'm gonna proceed.

00:39:44.850 --> 00:39:49.010 Okay, so under these four assumptions

00:39:49.010 --> 00:39:52.580 that I introduced earlier, you can sort of compute

00:39:52.580 --> 00:39:56.680 the explicit forms of the conditional likelihood functions.

00:39:56.680 --> 00:39:59.390 I'm not gonna go over the detailed forms,

00:39:59.390 --> 00:40:01.940 but I just want to point out that first of all,

00:40:01.940 --> 00:40:04.420 as I mentioned earlier, this does not depend on

00:40:04.420 --> 00:40:07.203 the two nuisance parameters, mu and kappa.

00:40:08.190 --> 00:40:12.380 And second of all, this actually reduces to a likelihood

00:40:12.380 --> 00:40:17.380 function that's previously derived in this paper in 2009

00:40:18.970 --> 00:40:21.940 by setting this R equals to zero.

00:40:21.940 --> 00:40:24.010 So R equals to zero means that the epidemic

00:40:24.010 --> 00:40:27.953 was not growing, so it's mostly a stationary epidemic.

00:40:29.970 --> 00:40:34.970 So that's reasonable for maybe influenza, but not for COVID.

00:40:39.500 --> 00:40:42.340 So for unconditional likelihood, we need to make

00:40:42.340 --> 00:40:45.530 further assumptions about how people traveled,

00:40:45.530 --> 00:40:48.730 the assumption we used was just a very simple,

00:40:48.730 --> 00:40:50.500 sort of a uniform assumption,

00:40:50.500 --> 00:40:52.080 uniform distribution assumption,

00:40:52.080 --> 00:40:54.990 that assumes that the travel was stable

00:40:54.990 --> 00:40:57.673 in the period that we considered.

00:40:58.790 --> 00:41:00.410 And we use those assumptions,

00:41:00.410 --> 00:41:05.410 we can derive the closed form unconditional likelihood.

00:41:06.320 --> 00:41:09.000 There's a little bit of approximation that's needed,

00:41:09.000 --> 00:41:14.000 but that's very, very reasonable in this case.

00:41:18.250 --> 00:41:21.540 So, I'd like to show you the results

00:41:21.540 --> 00:41:24.020 that fit in these parametric models.

00:41:24.020 --> 00:41:27.510 So what we did is we obtained point estimates

00:41:27.510 --> 00:41:31.650 of the parameters by maximizing the likelihood functions

00:41:31.650 --> 00:41:35.730 I just showed you, and then we obtained 95 percent

00:41:35.730 --> 00:41:38.453 confidence intervals, by a likelihood ratio test.

00:41:40.520 --> 00:41:45.133 So, what you can see is broadly, over different locations,

00:41:46.100 --> 00:41:50.023 the estimated doubling time was very consistent.

00:41:51.830 --> 00:41:55.110 Also cross-conditional and unconditional likelihood,

00:41:55.110 --> 00:42:00.110 so the doubling time was about two to 2.5 days.

00:42:01.180 --> 00:42:06.180 And the median incubation period is about four days,

00:42:06.980 --> 00:42:09.390 but there is a little bit of variability

00:42:11.226 --> 00:42:12.883 in the estimates.

00:42:13.980 --> 00:42:16.220 It turns out that the variability is mostly

00:42:16.220 --> 00:42:19.353 because of the parametric assumptions that we used.

00:42:20.800 --> 00:42:24.543 And then the 95 percent quantile is about,

00:42:26.880 --> 00:42:29.150 12 to 14 days.

00:42:29.150 --> 00:42:31.200 Or if you consider the sampling variability,

00:42:31.200 --> 00:42:33.853 that is about 11 to 15 days.

00:42:34.840 --> 00:42:39.660 Okay, but broadly speaking, across the different locations,

00:42:39.660 --> 00:42:44.660 they seem to suggest very similar answers.

00:42:47.170 --> 00:42:50.830 So, just to summarize, the initial doubling time

00:42:50.830 --> 00:42:54.013 seems to be between two to 2.5 days.

00:42:54.939 --> 00:42:56.840 Median incubation period is about four days,

00:42:56.840 --> 00:43:00.623 and 95 percent quantile is about 11 to 15 days.

00:43:02.600 --> 00:43:05.430 So, those sort of were our results,

00:43:05.430 --> 00:43:07.520 using the parametric model.

00:43:07.520 --> 00:43:12.470 And next I'm going to compare it with some other

00:43:12.470 --> 00:43:17.470 earlier analysis, and give you a demonstration,

00:43:17.640 --> 00:43:21.110 or an argument of why some of the other early analysis

00:43:21.110 --> 00:43:23.210 were severely biased.

00:43:23.210 --> 00:43:26.890 So first, let's look at this Lancet paper that I mentioned

00:43:26.890 --> 00:43:30.380 in the beginning of the talk that estimated doubling time.

00:43:30.380 --> 00:43:34.143 So the doubling time they estimated was 6.4. days.

00:43:36.610 --> 00:43:41.610 So, what happened is these authors used a modified

00:43:43.630 --> 00:43:48.090 SEIR model, so the SEIR model is very common

00:43:48.090 --> 00:43:51.310 in epidemic modeling, so the modified that model

00:43:51.310 --> 00:43:55.140 to account for traveling, but they did not account

00:43:55.140 --> 00:43:56.463 for the travel ban.

00:43:58.180 --> 00:44:03.180 So, basically to sort of simplify what's going on, 00:44:05.340 --> 00:44:08.810 what they essentially did is they used the density

00:44:08.810 --> 00:44:12.783 of the symptoms as in the population P,

00:44:14.760 --> 00:44:18.997 so they fitted this density, but they fit it using, ah,

00:44:19.990 --> 00:44:24.763 samples from the set D.

00:44:25.920 --> 00:44:29.420 So it is quite reasonable to assume that the incidence

00:44:29.420 --> 00:44:34.360 of symptom onset was growing exponentially in the population

00:44:34.360 --> 00:44:36.650 that is exposed to Wuhan.

00:44:36.650 --> 00:44:41.650 So given P, this distribution, margin distribution of S,

00:44:41.660 --> 00:44:46.640 was perhaps growing exponentially before the lockdown.

00:44:46.640 --> 00:44:49.260 But we don't actually have samples from P.

00:44:49.260 --> 00:44:50.913 We have a sample from D.

00:44:52.331 --> 00:44:57.331 So, we actually can derive the density of S and D,

00:44:58.570 --> 00:45:01.543 and that looked very different from exponential growth.

00:45:02.720 --> 00:45:06.320 So, basically the intuition is that if you look at

00:45:06.320 --> 00:45:09.230 the distribution of the transmission, T,

00:45:09.230 --> 00:45:13.200 it is growing exponentially, but it also has this effect,

00:45:13.200 --> 00:45:17.070 this exponential RT times L minus T.

00:45:17.070 --> 00:45:20.270 So basically, if you are transmitted on time T,

00:45:20.270 --> 00:45:24.560 then you only have the time between T to L

00:45:24.560 --> 00:45:28.610 to leave Wuhan and be observed by us.

00:45:28.610 --> 00:45:31.920 Okay, so that's why it's not only exponential growth,

00:45:31.920 --> 00:45:36.920 but there's also a decreasing trend, L minus T,

00:45:39.239 --> 00:45:41.853 for the distribution of the time of transmission.

00:45:42.980 --> 00:45:45.340 So from the time of symptom onset,

00:45:45.340 --> 00:45:47.193 it's just the time of transmission, 00:45:48.117 --> 00:45:52.300 convolved with the distribution of the incubation period.

00:45:52.300 --> 00:45:55.780 And that has this form that is approximately

00:45:55.780 --> 00:45:59.640 an exponential growth, and then times this term,

00:45:59.640 --> 00:46:03.260 that is L plus some quantity that depends

00:46:03.260 --> 00:46:07.913 on the incubation period and the epidemic growth, minus S.

00:46:09.530 --> 00:46:12.393 So this is a term that is not considered,

00:46:13.443 --> 00:46:17.520 in this simple exponential growth model.

00:46:17.520 --> 00:46:20.723 Which is basically what's used in that Lancet paper.

00:46:23.260 --> 00:46:26.410 Okay, so to illustrate this,

00:46:26.410 --> 00:46:29.340 what I'm showing you here is a histogram

00:46:29.340 --> 00:46:34.293 of the symptom onset of all the Wuhan exported cases,

00:46:35.210 --> 00:46:37.000 who are also residents of Wuhan.

00:46:37.000 --> 00:46:41.363 So they stayed from December first to January 23.

00:46:42.640 --> 00:46:46.300 What you see is that it was kind of growing very fast,

00:46:46.300 --> 00:46:49.100 perhaps exponentially in the beginning,

00:46:49.100 --> 00:46:53.253 but then it slows down around the time of the lockdown.

00:46:54.650 --> 00:46:59.500 Okay, so the orange curve is the theoretical fit

00:46:59.500 --> 00:47:03.910 that we obtained in the last slide,

00:47:03.910 --> 00:47:07.730 using the maximum likelihood estimator of the parameters.

00:47:07.730 --> 00:47:10.283 So it fits the data quite will.

00:47:11.610 --> 00:47:16.590 So what happened, I think, with the Lancet paper is,

00:47:16.590 --> 00:47:19.750 so the basically stopped about January 28th,

00:47:19.750 --> 00:47:22.930 so it's about here, and they essentially tried to fit

00:47:22.930 --> 00:47:27.930 an exponential growth from the beginning to January 28.

00:47:29.420 --> 00:47:32.550 And that would lead to much faster growth

00:47:32.550 --> 00:47:37.527 than fitting the whole model to account for the selection.

00:47:41.120 --> 00:47:41.953 Okay.

00:47:43.510 --> 00:47:46.260 So that's about epidemic growth.

00:47:46.260 --> 00:47:48.560 Next I will talk about several studies

00:47:48.560 --> 00:47:50.633 of the incubation period.

00:47:51.750 --> 00:47:56.580 So, these studies are susceptible to two kinds of biases.

00:47:56.580 --> 00:48:01.060 One is that some of them ignore the epidemic growth,

00:48:01.060 --> 00:48:03.900 so instead of using this likelihood function,

00:48:03.900 --> 00:48:05.918 this conditional likelihood function,

00:48:05.918 --> 00:48:08.430 to just fit this R is equal to zero,

00:48:08.430 --> 00:48:10.220 and then they use this likelihood function

00:48:10.220 --> 00:48:12.763 that was derived in the early paper.

00:48:15.490 --> 00:48:19.890 The other bias is sort of right-truncation.

00:48:19.890 --> 00:48:22.070 And basically, they kind of stopped

00:48:22.070 --> 00:48:24.450 the data collection early and only used cases

00:48:24.450 --> 00:48:29.230 confirmed by then, so people with long incubation periods

00:48:29.230 --> 00:48:32.620 are less likely to be included in the data,

00:48:32.620 --> 00:48:35.903 so that leads to underestimation of the incubation period.

00:48:37.649 --> 00:48:40.400 And a solution to this is you can actually derive

00:48:40.400 --> 00:48:43.190 the likelihood with additional conditioning events,

00:48:43.190 --> 00:48:45.070 that S is equal, sorry,

00:48:45.070 --> 00:48:48.420 less than or equal to some threshold, M.

00:48:48.420 --> 00:48:52.240 Suppose you stop the data collection a week after M,

00:48:52.240 --> 00:48:55.970 and you say: perhaps we have all, find out all the cases

00:48:55.970 --> 00:48:58.510 who showed symptoms beforehand.

00:48:58.510 --> 00:49:00.423 We can use this likelihood function.

00:49:01.710 --> 00:49:03.850 I'm not gonna show you the exact form,

00:49:03.850 --> 00:49:07.850 but basically you need to further divide by, ah,

00:49:10.140 --> 00:49:13.720 the probability of S less than or equal to M,

00:49:13.720 --> 00:49:18.070 and you can obtain closed-form expression for this

00:49:18.070 --> 00:49:20.203 under our parametric assumptions.

00:49:21.540 --> 00:49:23.373 Using integration by parts.

00:49:25.408 --> 00:49:29.460 So, I'd like to show you an experiment

00:49:29.460 --> 00:49:33.150 to illustrate this selection bias.

00:49:33.150 --> 00:49:37.690 So in this experiment, we kind of stop the data collection

00:49:37.690 --> 00:49:42.690 between any day from January 23 to February 18,

00:49:42.930 --> 00:49:47.930 and we fitted sort of this parametric BETS model,

00:49:48.070 --> 00:49:50.610 using one of the following likelihood.

00:49:50.610 --> 00:49:53.820 So this is the likelihood that treats R equals to zero,

00:49:53.820 --> 00:49:56.270 so it's adjusted for nothing,

00:49:56.270 --> 00:49:59.420 and this is the likelihood derived earlier

00:49:59.420 --> 00:50:00.923 and used in other studies.

00:50:02.010 --> 00:50:04.760 This is the likelihood function that adjusts for the growth,

00:50:04.760 --> 00:50:08.330 so R is treated as an unknown parameter.

00:50:08.330 --> 00:50:12.370 And this is the likelihood on the last slide that adjusted

00:50:12.370 --> 00:50:16.310 for both the growth and the right-truncation,

00:50:16.310 --> 00:50:20.560 as less than or equal to M.

00:50:20.560 --> 00:50:23.450 So the point estimates are obtained by MLEs,

00:50:23.450 --> 00:50:25.180 and the confidence intervals are obtained

00:50:25.180 --> 00:50:26.693 by nonparametric Bootstrap,

00:50:27.740 --> 00:50:32.203 and we compared our results with three previous studies.

00:50:36.080 --> 00:50:41.080 So this is, basically summarizes this experiment.

00:50:42.040 --> 00:50:43.470 This is a little bit complicated,

00:50:43.470 --> 00:50:48.100 so let me walk you through slowly.

00:50:48.100 --> 00:50:50.420 So there are three likelihood functions we used.

00:50:50.420 --> 00:50:53.530 One adjusts for nothing; that's the orange.

00:50:53.530 --> 00:50:56.630 The one is adjusted only for growth,

00:50:56.630 --> 00:51:00.183 and the ones that adjusted for both growth and truncation.

00:51:02.004 --> 00:51:03.550 Okay, so what you can immediately see

00:51:03.550 --> 00:51:06.803 is that if we adjusted for, ah,

00:51:07.940 --> 00:51:12.370 if we adjusted for nothing, then this is much larger

00:51:13.900 --> 00:51:16.653 than the other estimates.

00:51:17.510 --> 00:51:20.440 So actually, if you adjusted for nothing,

00:51:20.440 --> 00:51:23.330 and if you sort of used our entire data set,

00:51:23.330 --> 00:51:26.930 the median incubation period would be about nine days.

00:51:26.930 --> 00:51:30.600 And the 95 percent quantile would be about 25 days.

00:51:30.600 --> 00:51:32.393 So that's just way too large.

00:51:35.280 --> 00:51:38.420 And if you ignored right-truncation, for example,

00:51:38.420 --> 00:51:42.660 if you used this likelihood function we derived earlier,

00:51:42.660 --> 00:51:47.660 that only accounts for growth, you underestimate

00:51:47.880 --> 00:51:51.170 the incubation period in the beginning, as expected,

00:51:51.170 --> 00:51:54.123 but you slowly converge to this final estimate.

00:51:56.850 --> 00:51:59.980 And if you use this likelihood function and adjust for both

00:51:59.980 --> 00:52:02.960 growth and truncation, you actually get

00:52:02.960 --> 00:52:07.583 some quite sensible results by the end of January.

00:52:09.050 --> 00:52:13.530 So, it has large uncertainty, but it's roughly unbiased,

00:52:13.530 --> 00:52:17.383 and it kind of eventually converges to that estimate.

00:52:18.220 --> 00:52:21.780 The same estimate that we obtained

00:52:23.210 --> 00:52:26.333 using the blue curve, but using the full data.

00:52:28.440 --> 00:52:29.273 Okay.

00:52:30.430 --> 00:52:35.430 So, for the sake of time, I think I'll skip the part

00:52:35.600 --> 00:52:37.763 about Bayesian nonparametric inference.

00:52:39.620 --> 00:52:42.610 One thing that's a little bit interesting, I think,

00:52:42.610 --> 00:52:47.610 is there seems to be some difference between men

00:52:48.210 --> 00:52:51.160 and women in their incubation period.

00:52:51.160 --> 00:52:53.790 So these are sort of the posterior mean

00:52:53.790 --> 00:52:58.790 and posterior credible intervals for nonparametric

00:53:00.744 --> 00:53:04.350 incubation period, and you can see that men

00:53:04.350 --> 00:53:09.193 seem to develop symptoms quicker than women.

00:53:11.200 --> 00:53:13.890 So, that's a little bit interesting,

00:53:13.890 --> 00:53:17.960 and maybe, I mean, I'm not a doctor,

00:53:17.960 --> 00:53:22.170 but it could be related to the observation

00:53:22.170 --> 00:53:24.410 that men seem to be more susceptible,

00:53:24.410 --> 00:53:28.393 and die more frequently than women.

00:53:30.900 --> 00:53:33.193 So let's, let me conclude this talk.

00:53:34.480 --> 00:53:39.480 So these are some conclusions we found about COVID-19,

00:53:39.570 --> 00:53:42.760 using our dataset and our model.

00:53:42.760 --> 00:53:47.760 Initial doubling time in Wuhan was about two to 2.5 days.

00:53:49.880 --> 00:53:52.450 The median incubation period is about four days,

00:53:52.450 --> 00:53:55.010 and the proportion of incubation period

00:53:55.010 --> 00:53:57.523 above 14 days is about five percent.

00:53:59.530 --> 00:54:03.370 There are a number of limitations for our study.

00:54:03.370 --> 00:54:07.050 For example, we used the symptom onset reported

00:54:07.050 --> 00:54:11.060 by the patients and they are not always accurate.

00:54:11.060 --> 00:54:13.310 There could be behavioral reasons for people

00:54:13.310 --> 00:54:16.513 to report a later symptom onset.

00:54:17.720 --> 00:54:21.340 Even though these locations are intensive in their testing

00:54:21.340 --> 00:54:24.910 and contact tracing, some degree of under-ascertainment

00:54:24.910 --> 00:54:26.520 is perhaps inevitable.

00:54:28.365 --> 00:54:32.173 As I have shown you, in our dataset collection,

00:54:34.200 --> 00:54:36.200 discerning the Wuhan exported case

00:54:36.200 --> 00:54:39.010 is not a black and white decision.

00:54:39.010 --> 00:54:42.810 We used this beyond a reasonable doubt kind of criterion,

00:54:42.810 --> 00:54:46.250 but that's one criterion you can apply.

00:54:47.137 --> 00:54:50.640 And the crucial assumptions, we put the first

00:54:50.640 --> 00:54:52.960 two assumptions, which means that the travel

00:54:52.960 --> 00:54:56.910 and disease are independent, and that can be violated.

00:54:56.910 --> 00:55:01.910 For example, if I, if people tend to cancel

00:55:02.110 --> 00:55:05.453 their travel plans when feeling sick.

00:55:08.600 --> 00:55:11.750 Nevertheless, I think I have demonstrated some very

00:55:11.750 --> 00:55:16.750 compelling evidence for selection bias in early studies.

00:55:17.000 --> 00:55:22.000 Some of the biases you can correct by designing the study

00:55:25.280 --> 00:55:28.720 more carefully, some require more sophisticated

00:55:28.720 --> 00:55:30.623 statistical adjustments.

00:55:32.700 --> 00:55:36.880 And basically, I think the conclusion is:

00:55:36.880 --> 00:55:39.713 you should make un-calculated BETS.

00:55:40.690 --> 00:55:43.600 So, we should always carefully design the study

00:55:43.600 --> 00:55:46.713 and adhere to our sample inclusion criteria.

00:55:47.800 --> 00:55:52.800 And the statistical inference should not be based

00:55:52.850 --> 00:55:55.090 on some intuitive calculations,

00:55:55.090 --> 00:55:57.840 but should be based on first principles.

00:55:57.840 --> 00:56:00.470 So in this study, we kind of went back all the way

00:56:00.470 --> 00:56:03.823 to defining the support of random variables.

00:56:04.660 --> 00:56:06.787 So that's sort of statistics 101.

00:56:08.340 --> 00:56:11.200 But that's actually, it's extremely important.

00:56:11.200 --> 00:56:14.510 So I found it really helpful to start all the way

00:56:14.510 --> 00:56:19.510 from the beginning and develop a generative model.

00:56:20.200 --> 00:56:23.943 And that avoids a lot of potential selection biases.

00:56:25.320 --> 00:56:28.610 So the final lesson I'd like to share from this whole study

00:56:28.610 --> 00:56:33.610 is that I think this demonstrates the data quality

00:56:33.630 --> 00:56:37.540 and better design are much more important

00:56:37.540 --> 00:56:40.003 than data quantity and better modeling,

00:56:42.060 --> 00:56:44.043 in many real data studies.

00:56:45.500 --> 00:56:47.440 Thanks for the attention,

00:56:47.440 --> 00:56:49.883 and I'll take any questions from here.

00:56:51.080 --> 00:56:52.823 - Thanks to you for the nice talk.

00:56:53.750 --> 00:56:56.523 Does anyone have questions for Qingyuan?

00:57:00.250 --> 00:57:02.343 So Qing, I think someone, ah,

00:57:03.880 --> 00:57:06.063 yeah, Joe sent you a question.

00:57:07.290 --> 00:57:08.910 - Okay.

00:57:08.910 --> 00:57:12.080 - Are there any information in datasets of whether patient

00:57:12.080 --> 00:57:13.763 is healthcare worker?

00:57:15.184 --> 00:57:18.510 - No, these are not usually healthcare workers.

00:57:18.510 --> 00:57:20.690 These are exported from Wuhan, so they're usually

00:57:20.690 --> 00:57:24.310 just people who traveled maybe for sightseeing,

00:57:24.310 --> 00:57:28.190 or for the Chinese New Year, they traveled from Wuhan

00:57:28.190 --> 00:57:31.893 to other places and were diagnosed there.

00:57:34.300 --> 00:57:38.300 - Right, so also he has another question,

00:57:38.300 --> 00:57:41.330 Joe has another question also: how can we evaluate

00:57:41.330 --> 00:57:45.073 the effectiveness of social distancing and mask guidelines?

00:57:48.720 --> 00:57:53.480 - I think this study we did was not designed

00:57:53.480 --> 00:57:57.150 to answer those questions.

00:57:57.150 --> 00:57:59.623 We did have a very, ah,

00:58:01.120 --> 00:58:02.810 sort of preliminary analysis.

00:58:02.810 --> 00:58:07.540 So we broke the study period into two parts.

00:58:07.540 --> 00:58:11.820 So on January 20, it was confirmed publicly

00:58:11.820 --> 00:58:15.570 that the disease was human-to-human transmissible,

00:58:15.570 --> 00:58:20.290 so we broke the period into two parts:

00:58:20.290 --> 00:58:25.030 those before January 20 and those after January 20.

00:58:25.030 --> 00:58:27.490 But the after period is just three days.

00:58:27.490 --> 00:58:32.490 So January 21, 22, 23, and we found that if we fit

00:58:32.510 --> 00:58:36.030 different growths to these two periods, the second period,

00:58:36.030 --> 00:58:40.443 it seemed that the growth was substantially slower.

00:58:42.410 --> 00:58:47.410 The growth, the exponent R is not quite zero,

00:58:48.190 --> 00:58:49.870 but it's close.

00:58:49.870 --> 00:58:52.220 So it seems that the knowledge of sort

00:58:52.220 --> 00:58:56.213 of human-to-human transmissibility and the fact that,

00:58:57.570 --> 00:59:00.570 I think, masks are probably much more,

00:59:00.570 --> 00:59:03.310 were much more available in Wuhan,

00:59:03.310 --> 00:59:07.890 people started to do some social distancing

00:59:07.890 --> 00:59:10.570 right after January 20.

00:59:10.570 --> 00:59:14.320 I think that seemed to play a role.

00:59:14.320 --> 00:59:16.680 But that's very, very preliminary,

00:59:16.680 --> 00:59:21.577 and I think there are a lot of good studies about this now.

00:59:24.550 --> 00:59:26.400 - Donna has a question.

00:59:26.400 --> 00:59:31.400 Donna, do you want to say what your question is?

00:59:32.180 --> 00:59:33.140 - [Donna] Yeah, sure, thanks.

00:59:33.140 --> 00:59:36.230 That was a very interesting and clear talk.

00:59:36.230 --> 00:59:39.850 I really appreciated the way you carefully went through,

00:59:39.850 --> 00:59:43.907 step by step, to show-- (audio distorting)

00:59:47.449 --> 00:59:49.650 Who aren't doing that, I feel.

00:59:49.650 --> 00:59:53.770 But my question was, it was still hard for me to tell

00:59:53.770 --> 00:59:58.770 to what extent your estimates were identifiable

00:59:59.420 --> 01:00:04.270 due to assumptions and to what extent the data

01:00:04.270 --> 01:00:07.293 made the estimates fairly identifiable.

01:00:08.640 --> 01:00:11.533 - Yeah so essentially, I mean, selection bias,

01:00:12.430 --> 01:00:17.040 usually you cannot always avoid it, unless you

01:00:17.040 --> 01:00:22.000 make some kind of missing at random type of assumption.

01:00:22.000 --> 01:00:24.650 Here, we don't have a random selection.

01:00:24.650 --> 01:00:26.950 It's more like a deterministic selection,

01:00:26.950 --> 01:00:30.060 and we can quantify that selection event,

01:00:30.060 --> 01:00:35.060 but still, as you said, I think these are great questions

01:00:36.590 --> 01:00:41.321 to sort of disentangle the nonparametric assumptions

01:00:41.321 --> 01:00:44.246 needed for identification and the parametric assumptions

01:00:44.246 --> 01:00:45.463 needed for sort of better and easier inference.

01:00:50.600 --> 01:00:52.740 I don't have a formal result,

01:00:52.740 --> 01:00:56.350 but my feeling is the first two assumptions

01:00:56.350 --> 01:00:59.920 that are assumed, sort of the independence of travel

01:00:59.920 --> 01:01:04.920 and disease, that's sort of essential to the identification.

01:01:07.160 --> 01:01:11.687 And then later on, the assumptions are perhaps relaxable.

01:01:13.710 --> 01:01:15.430 So we did try to relax those

01:01:15.430 --> 01:01:17.973 in the Bayesian nonparametric analysis.

01:01:19.400 --> 01:01:23.827 But that's not a proof, so that's my, ah,

01:01:25.360 --> 01:01:26.763 best guess at this point.

01:01:28.290 --> 01:01:29.290 - [Donna] Thank you.

01:01:32.445 --> 01:01:36.833 - From Casey, said, ah, the estimates,

01:01:38.763 --> 01:01:41.480 people have estimated about five to 80 percent

01:01:41.480 --> 01:01:45.570 of asymptomatic infections, and isn't that a limitation

01:01:45.570 --> 01:01:47.760 of your model that you did not account

01:01:47.760 --> 01:01:49.770 for asymptomatic carriers?

01:01:49.770 --> 01:01:52.300 And if so, how can we possibly model for it,

01:01:52.300 --> 01:01:55.130 given the large range of estimates?

01:01:55.130 --> 01:01:59.000 So this is actually a feature of our study,

01:01:59.000 --> 01:02:03.030 because we actually had a, let's see,

01:02:05.220 --> 01:02:10.220 we had a term for the asymptomatic transmission.

01:02:14.270 --> 01:02:19.150 So, but that's just that parameter was canceled.

01:02:19.150 --> 01:02:22.340 So this parameter, mu, or one minus mu,

01:02:22.340 --> 01:02:26.910 is the proportion of asymptomatic infections.

01:02:29.480 --> 01:02:34.283 But then because we only observed cases who are,

01:02:35.570 --> 01:02:39.010 who showed symptoms, so actually in likelihood,

01:02:39.010 --> 01:02:41.683 this parameter mu got canceled.

01:02:42.970 --> 01:02:46.100 So, of course the reason we could cancel that mu

01:02:46.100 --> 01:02:47.890 is because of this assumption, too,

01:02:47.890 --> 01:02:52.650 that S is independent of the travel.

01:02:52.650 --> 01:02:54.850 So that's important.

01:02:54.850 --> 01:02:57.867 But once you assume that you actually, ah,

01:02:59.670 --> 01:03:03.113 sort of don't need to worry about asymptomatic transmission, 01:03:04.080 --> 01:03:07.990 and on the other hand, this dataset, or this whole method

01:03:07.990 --> 01:03:11.270 also provides more information about the proportion

01:03:11.270 --> 01:03:12.823 of asymptomatic infection.

01:03:15.290 --> 01:03:16.970 Hopefully that'll answer your question.

01:03:16.970 --> 01:03:18.760 - [Casey] Yeah, thanks; so you account for it

01:03:18.760 --> 01:03:22.673 by saying it's not really significant, in your estimate?

01:03:23.640 --> 01:03:25.810 - Yeah, so in the likelihood, you will get canceled.

01:03:25.810 --> 01:03:27.760 So it doesn't appear in the likelihood.

01:03:27.760 --> 01:03:30.440 So the likelihood of the data does not depend

01:03:30.440 --> 01:03:35.440 on how much are asymptomatic, because we only look

01:03:35.640 --> 01:03:37.973 at cases who are symptomatic.

01:03:39.220 --> 01:03:41.430 So this incubation period that we estimated

01:03:41.430 --> 01:03:44.010 are also the incubation period among

01:03:44.010 --> 01:03:45.660 those people who showed symptoms.

01:03:46.540 --> 01:03:48.670 - [Casey] So it's an elegant way of sidestepping

01:03:48.670 --> 01:03:51.023 the question, (laughing) in a way.

01:03:52.100 --> 01:03:55.830 - Well, it's not a sidestep, it's sort of,

01:03:55.830 --> 01:03:59.700 it's a limitation of this design.

01:03:59.700 --> 01:04:03.862 So the whole design should be robust

01:04:03.862 --> 01:04:06.730 to asymptomatic transmission, and it also gives

01:04:06.730 --> 01:04:10.603 no information about asymptomatic transmission.

01:04:11.710 --> 01:04:13.230 - [Casey] Yeah, I was really impressed at the way

01:04:13.230 --> 01:04:17.253 you took on that Lancet article and just really, ah,

01:04:18.370 --> 01:04:20.390 it was really impressive; what a great talk.

01:04:20.390 --> 01:04:21.730 Thank you so much.

01:04:21.730 --> 01:04:22.580 - Well thank you.

01:04:26.820 --> 01:04:28.790 - Hi Qing I have a question.

01:04:28.790 --> 01:04:32.990 So you mentioned before that because the measurements

01:04:32.990 --> 01:04:37.260 inside of Wuhan are the, or the, ah,

01:04:37.260 --> 01:04:38.550 the measurements that we have inside Wuhan,

01:04:38.550 --> 01:04:41.760 the numbers aren't very accurate due to various reasons.

01:04:41.760 --> 01:04:46.450 So I'm wondering that if you calculate the doubling time

01:04:46.450 --> 01:04:49.920 using the data for Wuhan city,

01:04:49.920 --> 01:04:52.800 and then take into, that uses the measurements

01:04:52.800 --> 01:04:57.756 before they changed the criterion for when it's counted

01:04:57.756 --> 01:05:01.750 as a confirmed case, and using the data before, say,

01:05:01.750 --> 01:05:03.720 you locked down, but taking into consideration

01:05:03.720 --> 01:05:06.970 that the data, you only looked at data.

01:05:06.970 --> 01:05:10.130 So you only looked at the confirmed cases before that date.

01:05:10.130 --> 01:05:12.538 Will you get a similar measurement,

01:05:12.538 --> 01:05:16.280 a similar estimate as if you're using the traveling data,

01:05:16.280 --> 01:05:17.823 or it is much worse?

01:05:18.880 --> 01:05:23.880 - Yeah, people have done an analysis on the data from Wuhan.

01:05:25.990 --> 01:05:28.730 What I would like to point out is that this figure

01:05:28.730 --> 01:05:33.545 is only the number of new, confirmed cases.

01:05:33.545 --> 01:05:36.170 So what is usually done in epidemic analysis

01:05:36.170 --> 01:05:40.010 is they don't look at the number of confirmed cases,

01:05:40.010 --> 01:05:44.690 but the number of cases who showed symptoms on a certain day

01:05:44.690 --> 01:05:49.690 because that's usually less variable, less noisy,

01:05:50.830 --> 01:05:55.420 than this sort of confirmation,

01:05:55.420 --> 01:05:59.170 because of the problem about confirmation.

01:05:59.170 --> 01:06:04.170 So people have done that, and I don't see a doubling time

01:06:05.560 --> 01:06:10.560 estimation from that; there was a journal paper on that.

01:06:12.766 --> 01:06:17.560 And there was also a very interesting comment on it

01:06:17.560 --> 01:06:20.520 that criticized some of its methodology.

01:06:20.520 --> 01:06:24.590 I didn't see a doubling time estimate.

01:06:24.590 --> 01:06:28.117 So they seemed to focus on the R-naught of the epidemic.

01:06:31.330 --> 01:06:34.090 I actually had thought about that as well,

01:06:34.090 --> 01:06:36.950 and we, in this study I have presented,

01:06:36.950 --> 01:06:40.133 I intentionally avoided to estimate R-naught.

01:06:40.980 --> 01:06:45.803 Because I think there was a lot of issues with, ah,

01:06:46.870 --> 01:06:51.700 finding out the unbiased estimate of the serial interval,

01:06:51.700 --> 01:06:54.273 which is very important in estimating R-naught.

01:06:56.270 --> 01:07:01.270 So, this estimate we found is not directly comparable

01:07:04.840 --> 01:07:06.843 to that journal paper, I guess.

01:07:08.290 --> 01:07:12.490 But so what happened, I think, is around late January,

01:07:12.490 --> 01:07:17.150 early February, all of people have tried to estimate

01:07:17.150 --> 01:07:21.240 the R-naught and the doubling time of the epidemic,

01:07:21.240 --> 01:07:23.020 and what I've found interesting was

01:07:23.020 --> 01:07:24.900 there were kind of two modes.

01:07:24.900 --> 01:07:28.530 There's several papers estimated that the doubling time

01:07:28.530 --> 01:07:31.293 was about six to seven days, and there were several papers

01:07:31.293 --> 01:07:35.683 that estimated doubling times of about two to four days.

01:07:37.410 --> 01:07:41.040 And I think, ah, 01:07:41.040 --> 01:07:45.280 at least I have shown that the Lancet paper,

01:07:45.280 --> 01:07:48.983 that their whole method seems to be very flawed.

01:07:50.040 --> 01:07:53.820 But whether this means that our estimate is very close

01:07:53.820 --> 01:07:57.790 to the truth, it doesn't necessarily mean so.

01:07:57.790 --> 01:08:00.873 Because we also have a lot of limitations.

01:08:02.490 --> 01:08:03.323 - Okay, thanks.

01:08:09.410 --> 01:08:11.003 Any more question for Qingyuan?

01:08:13.650 --> 01:08:15.510 Okay, thanks Qing.

01:08:15.510 --> 01:08:19.510 I guess that's all for today, and it's a great talk.

01:08:19.510 --> 01:08:21.090 If you have any more questions for Qing,

01:08:21.090 --> 01:08:25.100 you can send him an email, and you can find his email

01:08:25.100 --> 01:08:28.730 on his website, okay?

01:08:28.730 --> 01:08:30.105 - Okay.

01:08:30.105 --> 01:08:32.453 (muttering)

01:08:32.453 --> 01:08:34.860 All right, okay, thank you everyone.

01:08:34.860 --> 01:08:37.875 - Thank you, oh, we got a new message?

01:08:37.875 --> 01:08:39.880 (muttering)

01:08:39.880 --> 01:08:42.912 - It's just a, Keyong said thank you.

01:08:42.912 --> 01:08:44.660 - Okay, okay, bye!

01:08:44.660 --> 01:08:46.060 - [Qingyuan] All right, bye.