```
00:00:01.230 --> 00:00:03.780
<v ->And I'm gonna turn everything over to Dr. Wright.</v>

2
00:00:07.730 --> 00:00:09.770
<v ->Good morning and welcome to the third session</v>

3
00:00:09.770 --> 00:00:12.940
of our four-part Data Science Career Seminar Series:

4
00:00:12.940 --> 00:00:15.350
Bringing Data Science to Addiction Research.

5
00:00:15.350 --> 00:00:17.040
My name is Susan Wright and I'm

6
00:00:17.040 --> 00:00:19.940
from the Division of Neuroscience and Behavior or D&amp;B

7
00:00:19.940 --> 00:00:20.920
and I'm the Program Director

8
00:00:20.920 --> 00:00:22.930
for Big Data and Computational Science.

9
00:00:22.930 --> 00:00:25.750
I'm leading our data science efforts here at NIDA.

10
00:00:25.750 --> 00:00:28.180
Training in data science is also a priority for NIDA

11
00:00:28.180 --> 00:00:29.240
and it's supported by our new

12
00:00:29.240 --> 00:00:32.320
Office of Research Training, Diversity, and Disparities

13
00:00:32.320 --> 00:00:34.550
```

or ORTDD.

14
00:00:34.550 --> 00:00:36.130
We've organized a seminar series

15
00:00:36.130 --> 00:00:38.760
with the full support of our NIDA Director, Dr. Nora Volkow,

16
00:00:38.760 --> 00:00:40.420
and the organizers include members

17
00:00:40.420 --> 00:00:42.720
of the Division of Neuroscience and Behavior and the

18
00:00:42.720 --> 00:00:45.543
Office of Research Training, Diversity, and Disparities.

19
00:00:47.400 --> 00:00:50.460
The organizers include myself, Roger Little,

20
00:00:50.460 --> 00:00:51.540
who is the Deputy Director

21
00:00:51.540 --> 00:00:54.090
of the Division of Neuroscience and Behavior.

22
00:00:54.090 --> 00:00:56.810
Dr. Wilson Compton, NIDA Deputy Director

23
00:00:56.810 --> 00:00:58.080
and Acting Director of the

24
00:00:58.080 --> 00:01:02.140
Office of Research Training, Diversity, and Disparities.

25
00:01:02.140 --> 00:01:04.620
Dr. Albert Avila, the deputy director of the

```
26
00:01:04.620 --> 00:01:07.430
Office of Research Training, Diversity, and Disparities,

27
00:01:07.430 --> 00:01:08.263
and the Director

28
00:01:08.263 --> 00:01:10.720
of the Office of Disparities and Health Disparities.

29
00:01:10.720 --> 00:01:11.970
And Dr. Lindsey Friend,

30
00:01:11.970 --> 00:01:14.600
the Research and Career Development Program Officer in the

31
00:01:14.600 --> 00:01:18.120
Office of Research Training, Diversity, and Disparities.

32
00:01:18.120 --> 00:01:20.870
I want to thank Roger, Wilson, Albert, and Lindsey

33
00:01:20.870 --> 00:01:23.200
for their help with organizing the seminar series.

34
00:01:23.200 --> 00:01:24.341
And I also want to thank the team

35
00:01:24.341 --> 00:01:26.370
that has been helping with the technical details

36
00:01:26.370 --> 00:01:29.410
and that includes Usha Charya, Susan Holbrook,

37
00:01:29.410 --> 00:01:31.790
Caitlin Dudevoir, and David Metzger.

38
00:01:32.650 --> 00:01:35.860
```

For this session, we're highlighting women in data science.

39
00:01:35.860 --> 00:01:39.200
First, we'll have an interview with Dr. Kristian Lum,

40
00:01:39.200 --> 00:01:41.350
then a presentation by Dr. Brenda Curtis,

41
00:01:41.350 --> 00:01:43.260
followed by a joint Q&amp;A session

42
00:01:43.260 --> 00:01:45.210
where we'll take questions from the audience.

43
00:01:45.210 --> 00:01:47.580
Please use the chat box to submit your questions

44
00:01:47.580 --> 00:01:49.730
and we'll get to as many of them as we can.

45
00:01:51.520 --> 00:01:54.610
Dr. Kristian Lum is an Assistant Research Professor

46
00:01:54.610 --> 00:01:56.900
in the Department of Computer and Information Science

47
00:01:56.900 --> 00:01:58.920
at the University of Pennsylvania.

48
00:01:58.920 --> 00:02:00.490
She studies and develops statistical

49
00:02:00.490 --> 00:02:01.520
and machine learning models

50
00:02:01.520 --> 00:02:03.770
to tackle problems with social impact.

```
51
00:02:03.770 --> 00:02:06.580
This includes statistical population estimation models

52
00:02:06.580 --> 00:02:08.610
to estimate the number of undocumented victims

53
00:02:08.610 --> 00:02:11.460
of human rights violations, fair algorithms for use

54
00:02:11.460 --> 00:02:14.590
in high stakes decision making and epidemiological models

55
00:02:14.590 --> 00:02:16.170
to study disease spread among

56
00:02:16.170 --> 00:02:20.050
and between marginalized populations in a better community.

57
00:02:20.050 --> 00:02:21.610
Dr. Lum is particularly interested

58
00:02:21.610 --> 00:02:23.780
in applications in criminal justice.

59
00:02:23.780 --> 00:02:25.770
She enjoys using the tools of statistics

60
00:02:25.770 --> 00:02:27.040
and machine learning to shine a light

61
00:02:27.040 --> 00:02:29.370
on alternative interpretations of data.

62
00:02:29.370 --> 00:02:32.340
She is often just as interested

63
00:02:32.340 --> 00:02:34.981
```

in what is missing from a dataset and what is in it.

64
00:02:34.981 --> 00:02:38.110
So please join me in welcoming Dr. Kristian Lum

65
00:02:38.110 --> 00:02:39.660
for the interview this morning.

66
00:02:40.500 --> 00:02:42.123
Thanks for joining us, Kristian.

67
00:02:43.040 --> 00:02:44.190
<v ->Thanks for having me.</v>

68
00:02:45.250 --> 00:02:48.480
<v ->So you're currently an Assistant Research Professor</v>

69
00:02:48.480 --> 00:02:50.340
in the Department of Computer and Information Science

70
00:02:50.340 --> 00:02:52.240
at the University of Pennsylvania,

71
00:02:52.240 --> 00:02:53.577
can you tell us about your career path

72
00:02:53.577 --> 00:02:55.990
and how it led you to your current position?

73
00:02:55.990 --> 00:02:56.823
<v ->Yeah, sure.</v>

74
00:02:56.823 --> 00:03:00.430
So it definitely has not been a linear path.

75
00:03:00.430 --> 00:03:03.300
So I started out in statistics.

76
00:03:03.300 --> 00:03:04.900
This is about,

77
00:03:04.900 --> 00:03:06.077
Gosh, like 10 years ago or so.

78
00:03:06.077 --> 00:03:09.077
It was when I finished my PhD in that, so it's been a while.

79
00:03:10.200 --> 00:03:12.290
After that, I actually took some time off,

80
00:03:12.290 --> 00:03:15.160
which I, by the way, I think is a great thing to do

81
00:03:15.160 --> 00:03:17.010
if you need a little bit of a refresh,

82
00:03:17.010 --> 00:03:18.470
and I realize that not everybody can,

83
00:03:18.470 --> 00:03:20.200
but for me, it was the right call.

84
00:03:20.200 --> 00:03:22.370
And when I came back, I started

85
00:03:22.370 --> 00:03:25.383
as an Assistant Research Professor at Virginia Tech,

86
00:03:26.250 --> 00:03:27.083
where I was working

87
00:03:27.083 --> 00:03:29.630
in the Network Dynamics and Simulation Science Laboratory.

88
00:03:29.630 --> 00:03:32.440

It's been renamed since, but it was essentially working

89
00:03:32.440 --> 00:03:34.450
on things like age based models,

90
00:03:34.450 --> 00:03:37.770
some computational epidemiology on these age based models.

91
00:03:37.770 --> 00:03:38.820
And mostly what I focused

92
00:03:38.820 --> 00:03:43.010
on there was developing synthetic populations

93
00:03:44.170 --> 00:03:46.270
of the the United States and other countries.

94
00:03:46.270 --> 00:03:49.350
So basically simulated representations of individual people

95
00:03:49.350 --> 00:03:51.630
that were demographically representative and like layering

96
00:03:51.630 --> 00:03:53.850
on top all sorts of other information in a way

97
00:03:53.850 --> 00:03:57.130
that was consistent with data sources that were available.

98
00:03:57.130 --> 00:04:01.440
From there, I ended up taking a different job.

99
00:04:01.440 --> 00:04:04.210
So I worked briefly in at Datapad,

100
00:04:04.210 --> 00:04:06.770
which was sort of Silicon Valley

```
101
00:04:06.770 --> 00:04:08.510
like data science nonprofit,

102
00:04:08.510 --> 00:04:09.820
sorry, start up, not nonprofit,

103
00:04:09.820 --> 00:04:12.000
I'll get to that in a second, it was a startup

104
00:04:12.000 --> 00:04:14.263
and that was acquired soon after I started.

105
00:04:15.650 --> 00:04:17.930
And so I will take a quick step back,

106
00:04:17.930 --> 00:04:20.010
which is gonna explain the next six years

107
00:04:20.010 --> 00:04:23.900
of my career after this is that when I was in grad school,

108
00:04:23.900 --> 00:04:27.760
I sort of like cold emailed a person,

109
00:04:27.760 --> 00:04:28.730
sort of an organization

110
00:04:28.730 --> 00:04:30.280
whose work I was really interested in.

111
00:04:30.280 --> 00:04:33.730
So Patrick Ball was the leader of the organization

112
00:04:33.730 --> 00:04:36.300
and he sort of was really involved

113
00:04:36.300 --> 00:04:37.870
```

```
in this casualty estimation stuff.

114
00:04:37.870 --> 00:04:39.900
So estimating, like I think you said in my intro,

115
00:04:39.900 --> 00:04:43.160
the number of people who've been killed in various conflicts

116
00:04:43.160 --> 00:04:44.380
around the world when you don't believe

117
00:04:44.380 --> 00:04:46.540
that everybody has been recorded.

118
00:04:46.540 --> 00:04:48.320
And so I reached out to him

119
00:04:48.320 --> 00:04:50.370
when I was at midway through grad school

120
00:04:50.370 --> 00:04:52.730
and I ended up interning there for the summer.

121
00:04:52.730 --> 00:04:54.390
And in all of these sort of intervening years,

122
00:04:54.390 --> 00:04:56.730
I had sort of stayed in the periphery of the organization.

123
00:04:56.730 --> 00:04:58.800
It was called the Human Rights Data Analysis Group,

124
00:04:58.800 --> 00:05:02.400
still is called that in fact, doing some consulting,

125
00:05:02.400 --> 00:05:04.030
just sort of staying involved and interested
```

```
126
00:05:04.030 --> 00:05:06.740
'cause I was really interested in what they did,

127
00:05:06.740 --> 00:05:07.990
thought they do some really high impact

128
00:05:07.990 --> 00:05:09.520
and interesting projects.

129
00:05:09.520 --> 00:05:13.430
And so after Datapad was acquired,

130
00:05:13.430 --> 00:05:16.820
I ended up doing part-time consulting for them again

131
00:05:16.820 --> 00:05:19.010
and doing part-time data science consulting.

132
00:05:19.010 --> 00:05:20.040
So through Datapad

133
00:05:20.040 --> 00:05:22.250
I'd met quite a few people who were

134
00:05:22.250 --> 00:05:25.230
sort of in the startup world and in sort of tech basically

135
00:05:25.230 --> 00:05:27.800
and did data science consulting for them.

136
00:05:27.800 --> 00:05:31.510
And I think it was around 2015 or so

137
00:05:31.510 --> 00:05:33.340
and the years are sort of blending together

138
00:05:33.340 --> 00:05:37.350
```

```
at this point for me, but they had a budget

139
00:05:37.350 --> 00:05:38.677
to hire me full-time at HRDAG,

140
00:05:38.677 --> 00:05:41.250
and so I jumped on board and that was where I spent

141
00:05:42.360 --> 00:05:44.040
from whenever that point was again,

142
00:05:44.040 --> 00:05:45.620
'cause I was sort of consulting

143
00:05:45.620 --> 00:05:46.880
at an increasing rate over time,

144
00:05:46.880 --> 00:05:48.320
so it's hard to remember the exact date

145
00:05:48.320 --> 00:05:50.250
where I jumped on full-time.

146
00:05:50.250 --> 00:05:52.700
But from whenever that was roughly 2015

147
00:05:52.700 --> 00:05:55.320
up until March of last year,

148
00:05:55.320 --> 00:05:56.810
I was at HRDAG and there,

149
00:05:56.810 --> 00:06:00.160
I led project on the United States.

150
00:06:00.160 --> 00:06:02.640
So in the human rights world, there's often this criticism
```

```
151
00:06:02.640 --> 00:06:06.460
that sort of human rights organizations are mostly based

152
00:06:06.460 --> 00:06:08.160
in like the United States and Western Europe

153
00:06:08.160 --> 00:06:10.260
and have a tendency to be looking outward

154
00:06:10.260 --> 00:06:11.680
and sort of pointing all over the world

155
00:06:11.680 --> 00:06:13.930
and saying, you people in developing countries

156
00:06:13.930 --> 00:06:17.093
like these are the things that, sorry, my dog,

157
00:06:18.700 --> 00:06:21.810
these are the abuses that you that you're all perpetrating.

158
00:06:21.810 --> 00:06:23.020
And we were thinking,

159
00:06:23.020 --> 00:06:24.550
it was time for us to look inwards,

160
00:06:24.550 --> 00:06:25.847
the United States has its own problems.

161
00:06:25.847 --> 00:06:28.970
And so we thought, what projects could we do

162
00:06:29.850 --> 00:06:31.970
that would speak to issues in the United States?

163
00:06:31.970 --> 00:06:34.960
```

So usually, we look for projects where our skills

164
00:06:34.960 --> 00:06:38.130
as statisticians and data scientists

165
00:06:38.130 --> 00:06:40.380
actually give us sort of like unique,

166
00:06:40.380 --> 00:06:41.213
I don't want to say advantage

167
00:06:41.213 --> 00:06:43.890
like a unique perspective on a problem, right?

168
00:06:43.890 --> 00:06:45.560
And so we ended up working

169
00:06:45.560 --> 00:06:47.890
on the criminal justice system in the United States.

170
00:06:47.890 --> 00:06:49.350
So there, I started studying things

171
00:06:49.350 --> 00:06:53.160
like predictive policing systems, risk assessment models

172
00:06:53.160 --> 00:06:56.210
that evaluate someone's likelihood of rearrest

173
00:06:56.210 --> 00:06:59.820
if that'll you use to then say, make recommendations

174
00:06:59.820 --> 00:07:02.830
about pretrial release, things like that.

175
00:07:02.830 --> 00:07:06.330
And so I did that for about five, maybe six years,

```
176
00:07:06.330 --> 00:07:08.830
somewhere in that range.

177
00:07:08.830 --> 00:07:11.850
I also became very involved in the algorithmic fairness,

178
00:07:11.850 --> 00:07:14.093
accountability and transparency community,

179
00:07:15.380 --> 00:07:17.570
which I think we might talk about a little bit later.

180
00:07:17.570 --> 00:07:19.960
And then last year on March 2nd,

181
00:07:19.960 --> 00:07:22.910
I began my new job at the University of Pennsylvania.

182
00:07:22.910 --> 00:07:25.930
And I'll note that's an interesting time to start a new job.

183
00:07:25.930 --> 00:07:28.880
It was I think maybe one week exactly before the university

184
00:07:28.880 --> 00:07:29.713
sort of shut down due to COVID.

185
00:07:29.713 --> 00:07:33.310
So yeah, interesting times, but sorry,

186
00:07:33.310 --> 00:07:36.150
maybe that was longer than you expected, but my career path

187
00:07:36.150 --> 00:07:40.310
to get here has been a little bit of an adventure.

188
00:07:40.310 --> 00:07:41.210
```

```
It's definitely non-linear

189
00:07:41.210 --> 00:07:43.987
and so it was kind of thought I'd go through it all.

190
00:07:45.816 --> 00:07:47.630
<v ->That was great to hear about, thank you.</v>

191
00:07:47.630 --> 00:07:49.430
So we always hear about how women make

192
00:07:49.430 --> 00:07:51.860
up less than the workforce in STEM careers.

193
00:07:51.860 --> 00:07:54.070
Were you always interested in pursuing a career

194
00:07:54.070 --> 00:07:56.140
in STEM and if so, were there any points

195
00:07:56.140 --> 00:07:58.140
in either your education or your career

196
00:07:58.140 --> 00:08:01.480
where you felt especially supported or not supported?

197
00:08:01.480 --> 00:08:04.190
<v ->Yeah, I think I was always interested in STEM</v>

198
00:08:04.190 --> 00:08:06.790
'cause as long as I can remember, even going back

199
00:08:06.790 --> 00:08:09.967
to being a little kid, I always really liked math,

200
00:08:09.967 --> 00:08:14.530
and I can remember being in high school and taking calculus
```

```
201
00:08:14.530 --> 00:08:16.530
at the local community college over the summer

202
00:08:16.530 --> 00:08:17.880
just cause I kind of wanted to get ahead.

203
00:08:17.880 --> 00:08:19.900
I thought it'd be fun, I guess I was a big nerd,

204
00:08:19.900 --> 00:08:20.733
I don't know.

205
00:08:21.760 --> 00:08:23.187
And it was the first time where I was like,

206
00:08:23.187 --> 00:08:25.940
"Wow, this is awesome."

207
00:08:25.940 --> 00:08:28.540
Because it wasn't just here's the algorithm,

208
00:08:28.540 --> 00:08:31.690
you execute as a human to say like do long division.

209
00:08:31.690 --> 00:08:34.340
I felt like it was actually explaining why things happened

210
00:08:34.340 --> 00:08:36.380
and helping you to set up problems,

211
00:08:36.380 --> 00:08:38.490
so you could address, you have set up the math

212
00:08:38.490 --> 00:08:40.360
so you could address a real world problem.

213
00:08:40.360 --> 00:08:41.837
```

And I remember just thinking like,

214
00:08:41.837 --> 00:08:43.770
"Wow, this is what I wanna do."

215
00:08:43.770 --> 00:08:46.140
And so that was, I think maybe my junior year of high school

216
00:08:46.140 --> 00:08:49.160
and basically from there, it was a pretty straight shot.

217
00:08:49.160 --> 00:08:51.940
I went into college knowing I wanted to do math

218
00:08:51.940 --> 00:08:53.370
about halfway through.

219
00:08:53.370 --> 00:08:55.800
I took a statistics course and I sort of had another one

220
00:08:55.800 --> 00:08:57.217
of those moments where I was like,

221
00:08:57.217 --> 00:08:59.860
"Yup, this is the thing, I really like this."

222
00:08:59.860 --> 00:09:01.750
I liked math as well.

223
00:09:01.750 --> 00:09:04.340
I really was drawn to sort of statistics

224
00:09:04.340 --> 00:09:07.930
because of the grounding in real world problems a little bit

225
00:09:07.930 --> 00:09:11.450
more than math when you get up to some of the higher levels.

```
226
00:09:11.450 --> 00:09:13.240
And of course, people do applied math, it's very grounded,

227
00:09:13.240 --> 00:09:16.380
but in my experience, a statistics was sort of the path

228
00:09:16.380 --> 00:09:17.700
that made sense for me.

229
00:09:17.700 --> 00:09:19.323
So yeah, it's always been,

230
00:09:20.210 --> 00:09:22.930
I think that was definitely always something I knew

231
00:09:22.930 --> 00:09:24.270
I wanted to do.

232
00:09:24.270 --> 00:09:25.803
In terms of feeling supported,

233
00:09:27.843 --> 00:09:31.820
I think it's a little bit of both, to be honest.

234
00:09:31.820 --> 00:09:35.140
I think I've always felt quite a bit of support from say

235
00:09:35.140 --> 00:09:38.230
like my family and most of my professors,

236
00:09:38.230 --> 00:09:41.070
and I felt that there are paths

237
00:09:41.070 --> 00:09:43.600
to having a career in this area.

238
00:09:43.600 --> 00:09:46.000
```

```
But there have also been times where I felt like

239
00:09:47.990 --> 00:09:49.280
I don't even know how to say this exactly,

240
00:09:49.280 --> 00:09:53.000
but like very unwelcome in the sense of

241
00:09:54.920 --> 00:09:56.233
feeling like,

242
00:09:58.510 --> 00:09:59.880
I'm trying to word this delicately.

243
00:09:59.880 --> 00:10:01.840
So I guess I'll be blunt and say that

244
00:10:01.840 --> 00:10:03.727
in 2017 I wrote an article, it was called,

245
00:10:03.727 --> 00:10:05.460
"Statistics We have a Problem."

246
00:10:05.460 --> 00:10:06.970
And it was about my experience

247
00:10:06.970 --> 00:10:09.390
with sexual harassment in statistics.

248
00:10:09.390 --> 00:10:11.900
And if anyone wants to check that out,

249
00:10:11.900 --> 00:10:14.010
you should feel more than welcome to do so,

250
00:10:14.010 --> 00:10:15.430
it's still on the Internet.
```

251
00:10:15.430 --> 00:10:17.530
It made somewhat of a splash in the field,

252
00:10:17.530 --> 00:10:22.050
I don't want to rehash all of the details of that here,

253
00:10:22.050 --> 00:10:24.863
I don't think this is exactly the time and place for that,

254
00:10:25.700 --> 00:10:27.920
but it's not because I'm shy about talking about it.

255
00:10:27.920 --> 00:10:31.010
But I guess what I would say, I think like big picture,

256
00:10:31.010 --> 00:10:32.340
I felt like supported,

257
00:10:32.340 --> 00:10:34.980
but then there are all these things that happen

258
00:10:34.980 --> 00:10:37.700
along the way that make you maybe feel

259
00:10:37.700 --> 00:10:41.590
like you might have to compromise some of your values

260
00:10:41.590 --> 00:10:43.630
or integrity or put up with things

261
00:10:43.630 --> 00:10:47.970
that you just don't wanna put up with to make it

262
00:10:48.925 --> 00:10:49.870
in this career.

263
00:10:49.870 --> 00:10:52.420

```
And so in the article,

264
00:10:52.420 --> 00:10:53.450
I do talk about this a little bit,

265
00:10:53.450 --> 00:10:54.630
that was one of the precipitate,

266
00:10:54.630 --> 00:10:55.980
there were some precipitating events

267
00:10:55.980 --> 00:10:58.270
involving sexual harassment where I did decide,

268
00:10:58.270 --> 00:11:02.510
okay, statistics or academic statistics is not going

269
00:11:02.510 --> 00:11:05.420
to be a place where I can thrive as a researcher

270
00:11:05.420 --> 00:11:09.070
and as a human who has to live among these colleagues,

271
00:11:09.070 --> 00:11:12.303
and so that was part of the reason that I left at one point.

272
00:11:17.300 --> 00:11:18.580
<v ->So what do you think can be done</v>

273
00:11:18.580 --> 00:11:20.710
to help women feel more supported in STEM careers

274
00:11:20.710 --> 00:11:23.760
in general, but particularly data science?

275
00:11:23.760 --> 00:11:25.540
<v ->What do I think can help?</v>
```

276
00:11:25.540 --> 00:11:30.540
I think if you see that people are being treated poorly,

277
00:11:30.800 --> 00:11:32.230
even sort of minor things,

278
00:11:32.230 --> 00:11:34.730
I think it helps to say something.

279
00:11:34.730 --> 00:11:37.170
I mean, I know it can be embarrassing for the person

280
00:11:37.170 --> 00:11:39.240
or for you to to step in and say something,

281
00:11:39.240 --> 00:11:42.300
but I think, simple things like, "Hey, I saw that,"

282
00:11:42.300 --> 00:11:44.270
or "You doing okay?"

283
00:11:44.270 --> 00:11:45.390
And these are sort of the minor things,

284
00:11:45.390 --> 00:11:47.790
I'm talking about the sort of harassment type of things

285
00:11:47.790 --> 00:11:51.610
or even talking to folks who are behaving inappropriately

286
00:11:51.610 --> 00:11:55.390
and being like, "What I saw wasn't okay."

287
00:11:55.390 --> 00:11:57.170
I actually think that makes a big difference

288
00:11:57.170 --> 00:12:00.470

```
because I think what people who maybe don't experience

289
00:12:00.470 --> 00:12:02.600
these things themselves don't understand,

290
00:12:02.600 --> 00:12:04.540
it's not just the one thing that you saw,

291
00:12:04.540 --> 00:12:05.940
it's like all of these other things

292
00:12:05.940 --> 00:12:07.950
that sort of accumulate over time

293
00:12:07.950 --> 00:12:11.390
and really just wear someone down and make them feel...

294
00:12:11.390 --> 00:12:12.480
And it's worse than unwelcome,

295
00:12:12.480 --> 00:12:15.420
make them feel like this isn't a place

296
00:12:15.420 --> 00:12:17.330
where they can thrive.

297
00:12:17.330 --> 00:12:18.910
And so I think that's one thing

298
00:12:18.910 --> 00:12:21.100
and I think there are bigger picture things too.

299
00:12:21.100 --> 00:12:24.570
Like the things where I feel like we're doing pretty well,

300
00:12:24.570 --> 00:12:27.910
just panels and things like to encourage women
```

```
301
00:12:27.910 --> 00:12:30.130
that this is a place where you can use your skills,

302
00:12:30.130 --> 00:12:34.476
you will be valued, having opportunities

303
00:12:34.476 --> 00:12:36.880
for people to learn more about the field.

304
00:12:36.880 --> 00:12:39.590
I think it's a little bit of everything, to be honest

305
00:12:41.870 --> 00:12:44.290
<v ->Definitely will be checking out your article.</v>

306
00:12:44.290 --> 00:12:46.030
So your work on the development

307
00:12:46.030 --> 00:12:47.740
of statistical and machine learning models

308
00:12:47.740 --> 00:12:49.680
to tackle problems with social impact,

309
00:12:49.680 --> 00:12:51.260
and I know that you're particularly interested

310
00:12:51.260 --> 00:12:53.270
in applications to criminal justice.

311
00:12:53.270 --> 00:12:55.340
Can you tell us about some of the specific projects

312
00:12:55.340 --> 00:12:56.380
you've worked on where you felt

313
00:12:56.380 --> 00:12:58.470
```

```
like you've made the biggest difference?

314
00:12:58.470 --> 00:13:00.260
<v ->Yeah, so I think probably the project</v>

315
00:13:00.260 --> 00:13:02.010
that made the biggest difference came out

316
00:13:02.010 --> 00:13:03.370
quite a few years ago at this point.

317
00:13:03.370 --> 00:13:05.020
So it was in, I think 2016

318
00:13:06.162 --> 00:13:08.450
and this is the project on predictive policing.

319
00:13:08.450 --> 00:13:11.170
So predictive policing is essentially the idea

320
00:13:11.170 --> 00:13:13.350
that you could use data

321
00:13:13.350 --> 00:13:15.750
and it's usually things like police records of crime

322
00:13:15.750 --> 00:13:19.410
to make predictions about who will commit a crime

323
00:13:19.410 --> 00:13:21.350
in the future, sometimes who will be a victim

324
00:13:21.350 --> 00:13:23.230
of a crime in the future, lumped together.

325
00:13:23.230 --> 00:13:25.680
It's kind of a strange thing to do in my opinion
```

```
326
00:13:25.680 --> 00:13:28.360
or where crime will occur in the future.

327
00:13:28.360 --> 00:13:29.850
And these things are actually pretty popular,

328
00:13:29.850 --> 00:13:32.370
these sorts of models, some police departments make

329
00:13:32.370 --> 00:13:33.880
their own if they're larger departments

330
00:13:33.880 --> 00:13:36.750
but a lot of them buy tools

331
00:13:36.750 --> 00:13:40.440
or software from companies that sell these sorts of things.

332
00:13:40.440 --> 00:13:42.440
And so I wrote a paper

333
00:13:42.440 --> 00:13:45.730
where we reproduced a predictive policing algorithm

334
00:13:45.730 --> 00:13:47.280
that was published in Jaza,

335
00:13:47.280 --> 00:13:49.450
so one of the big statistics journals,

336
00:13:49.450 --> 00:13:53.980
and we applied it to data in from Oakland, California

337
00:13:53.980 --> 00:13:57.580
to see what would happen if this algorithm had been

338
00:13:57.580 --> 00:13:59.450
```

in the past applied there.

339
00:13:59.450 --> 00:14:02.540
And so we actually, we did a few comparison.

340
00:14:02.540 --> 00:14:04.890
So the idea is if police records

341
00:14:04.890 --> 00:14:07.370
are not a representative sample of crime and in particular,

342
00:14:07.370 --> 00:14:10.400
if they over represent say communities of color,

343
00:14:10.400 --> 00:14:13.130
then the algorithms will learn those disparate patterns

344
00:14:13.130 --> 00:14:17.340
and be used to concentrate and reallocate policing back

345
00:14:17.340 --> 00:14:19.650
to the locations where there were overrepresented

346
00:14:19.650 --> 00:14:22.160
in the past and in particular, they could amplify,

347
00:14:22.160 --> 00:14:25.020
or reproduce, or perpetuate historical racial bias

348
00:14:25.020 --> 00:14:27.670
in policing and that was sort of the idea.

349
00:14:27.670 --> 00:14:30.980
And so we needed to compare to like, where do we think?

350
00:14:30.980 --> 00:14:33.920
So we used a dataset on drug crimes in Oakland,

```
351
00:14:33.920 --> 00:14:35.360
I needed a data set to say like, okay,

352
00:14:35.360 --> 00:14:37.930
where do we think drug use is actually happening,

353
00:14:37.930 --> 00:14:38.890
seems like a reasonable thing to think about.

354
00:14:38.890 --> 00:14:41.930
And so we used some public health data compared

355
00:14:41.930 --> 00:14:46.030
it to census data, just to get a sense that if you look

356
00:14:46.030 --> 00:14:47.380
at this public health data,

357
00:14:48.720 --> 00:14:52.210
it seems like drug use is probably all over the place.

358
00:14:52.210 --> 00:14:54.670
But then when you look at the historical police records

359
00:14:54.670 --> 00:14:57.200
of drug crime, it's really highly concentrated

360
00:14:57.200 --> 00:14:59.880
in black communities and Hispanic communities.

361
00:14:59.880 --> 00:15:02.500
And so when we ran the algorithm on that historical data

362
00:15:02.500 --> 00:15:05.340
just like would be done if the algorithm had been applied

363
00:15:05.340 --> 00:15:08.340
```

to this data and I'll give a caveat in a second there,

364
00:15:08.340 --> 00:15:11.330
we saw that in fact, it would reproduce the historical bias

365
00:15:11.330 --> 00:15:14.180
in policing and could be used to sort of perpetuate that.

366
00:15:15.070 --> 00:15:18.640
In practice, the algorithm that we applied,

367
00:15:18.640 --> 00:15:19.550
they say they don't apply it

368
00:15:19.550 --> 00:15:20.590
to drug crime data,

369
00:15:20.590 --> 00:15:24.740
but I think the sort of broad overarching point remains.

370
00:15:24.740 --> 00:15:28.860
And so I think this particular project had a

371
00:15:28.860 --> 00:15:29.693
pretty big impact.

372
00:15:29.693 --> 00:15:33.440
This was something that advocacy groups,

373
00:15:33.440 --> 00:15:35.100
people who work in this space

374
00:15:35.100 --> 00:15:38.510
in like less quantitative ways had been saying for awhile.

375
00:15:38.510 --> 00:15:39.970
And this is one of the things that I think is

376
00:15:39.970 --> 00:15:43.450
really important if you want to work in, I don't know,

377
00:15:43.450 --> 00:15:44.380
I don't actually love this term,

378
00:15:44.380 --> 00:15:45.430
but like data for social good

379
00:15:45.430 --> 00:15:47.520
or whatever that sort

380
00:15:47.520 --> 00:15:52.520
of field is called is that people who work in those areas

381
00:15:52.520 --> 00:15:54.620
in non-quantitative ways often, already kind of know

382
00:15:54.620 --> 00:15:56.300
what's going on, and it can be useful

383
00:15:56.300 --> 00:16:00.410
to give a sort of quantitative voice or rigor

384
00:16:00.410 --> 00:16:01.840
to some of those same concerns.

385
00:16:01.840 --> 00:16:02.917
So like look into it and be like,

386
00:16:02.917 --> 00:16:04.920
"Okay, this is what you think is going on."

387
00:16:04.920 --> 00:16:06.770
Let's investigate this using data,

388
00:16:06.770 --> 00:16:08.880

```
using the tools of data science, because this is

389
00:16:08.880 --> 00:16:12.120
what people in power, people who could make decisions listen

390
00:16:12.120 --> 00:16:13.533
to in a lot of ways.

391
00:16:15.226 --> 00:16:19.220
Let's see if what anecdotally you think is happening

392
00:16:19.220 --> 00:16:21.170
or might happen is born out by the data

393
00:16:25.930 --> 00:16:27.760
<v ->Just to follow up on that last part,</v>

394
00:16:27.760 --> 00:16:29.830
you mentioned that term, you said you don't like,

395
00:16:29.830 --> 00:16:33.110
could you clarify why or what should be used instead

396
00:16:33.110 --> 00:16:35.630
<v ->Just because I feel like it's too broad.</v>

397
00:16:35.630 --> 00:16:37.800
I feel like when you're saying social good,

398
00:16:37.800 --> 00:16:38.803
Who's social good?

399
00:16:39.895 --> 00:16:40.880
One thing that I've learned working

400
00:16:40.880 --> 00:16:45.120
in the criminal justice area is that there are
```

```
401
00:16:45.120 --> 00:16:46.930
many people who work in this area

402
00:16:46.930 --> 00:16:48.830
who think what they're doing is for social good

403
00:16:48.830 --> 00:16:53.050
and it sort of a very broad spectrum of approaches

404
00:16:53.050 --> 00:16:55.210
to dealing with social issues.

405
00:16:55.210 --> 00:16:57.750
And in a lot of cases, people on both sides

406
00:16:57.750 --> 00:17:00.580
of the spectrum think they're basically doing data science

407
00:17:00.580 --> 00:17:02.360
for social good and have very different opinions

408
00:17:02.360 --> 00:17:04.520
about the world they'd like to create

409
00:17:04.520 --> 00:17:07.440
or how they'd like to get there, right?

410
00:17:07.440 --> 00:17:10.110
And so I just like to be kind of specific

411
00:17:10.110 --> 00:17:13.150
about the sorts of projects that I do

412
00:17:13.150 --> 00:17:16.140
and the types of lenses I take to looking at data

413
00:17:17.870 --> 00:17:20.860
```

```
just because when it doesn't say a whole lot,

414
00:17:20.860 --> 00:17:22.710
I guess when you say for social good.

415
00:17:25.370 --> 00:17:28.660
<v ->Makes sense, so in machine learning algorithms are said</v>

416
00:17:28.660 --> 00:17:30.790
to be fair if the results are independent

417
00:17:30.790 --> 00:17:34.080
of given variables, typically those considered sensitive,

418
00:17:34.080 --> 00:17:36.240
can you talk a little bit about algorithmic bias

419
00:17:36.240 --> 00:17:38.680
and the impacts it can have on society?

420
00:17:38.680 --> 00:17:41.770
<v ->Yeah, sure, so what you listed is is one example</v>

421
00:17:41.770 --> 00:17:42.870
of a definition of fairness

422
00:17:42.870 --> 00:17:45.410
and that's one that I've written about in the past.

423
00:17:45.410 --> 00:17:47.960
But it's one of many definitions.

424
00:17:47.960 --> 00:17:50.460
So there's independence, there also equality

425
00:17:50.460 --> 00:17:53.670
of say false positive rates, equality of say accuracy,
```

426
00:17:53.670 --> 00:17:56.350
there's any number of ways you can dice it.

427
00:17:56.350 --> 00:17:58.530
And so when I was writing about this,

428
00:17:58.530 --> 00:18:00.510
this is back probably 2016, 2017.

429
00:18:00.510 --> 00:18:03.190
Also, this is really what sort of like the state

430
00:18:03.190 --> 00:18:05.500
of the field was, and I think it's useful to be thinking

431
00:18:05.500 --> 00:18:07.660
about different ways to measure inequality

432
00:18:07.660 --> 00:18:09.470
and what fairness might look like.

433
00:18:09.470 --> 00:18:10.630
But those really only take

434
00:18:10.630 --> 00:18:12.900
into account the model in isolation.

435
00:18:12.900 --> 00:18:17.010
And so I think since then, this field has expanded the scope

436
00:18:17.010 --> 00:18:21.870
of what it means to talk about a model behaving fairly,

437
00:18:21.870 --> 00:18:24.110
to think about sort of like upstream

438
00:18:24.110 --> 00:18:25.850

```
and downstream consequences as well.

439
00:18:25.850 --> 00:18:28.330
So will the model be used fairly, right?

440
00:18:28.330 --> 00:18:29.500
If say it's a model that's used

441
00:18:29.500 --> 00:18:33.250
to inform human decision makers, does that actually lead

442
00:18:33.250 --> 00:18:37.350
to humans making decisions that are better or more fair

443
00:18:37.350 --> 00:18:40.960
or along any dimension that you might want like,

444
00:18:40.960 --> 00:18:43.250
is this moving us towards a world that we might like,

445
00:18:43.250 --> 00:18:47.280
is it moving us towards say in the criminal justice context,

446
00:18:47.280 --> 00:18:50.320
often people want to use predictive models

447
00:18:50.320 --> 00:18:51.920
as part of a bail reform,

448
00:18:51.920 --> 00:18:54.520
type of movement to reduce pretrial detention.

449
00:18:54.520 --> 00:18:56.620
So is the introduction of model actually moving

450
00:18:56.620 --> 00:18:58.030
us towards that goal of having
```

451
00:18:58.030 --> 00:19:00.150
fewer people incarcerated pre-trial?

452
00:19:00.150 --> 00:19:02.770
So sort of expanding the scope is I think one

453
00:19:02.770 --> 00:19:05.370
of the important things when we think about talking

454
00:19:05.370 --> 00:19:06.450
about fair models, right?

455
00:19:06.450 --> 00:19:09.380
So I think doing these sorts of evaluations of the model

456
00:19:09.380 --> 00:19:12.140
in isolation, it's an important component

457
00:19:12.140 --> 00:19:16.830
because I think it's very unlikely that as you start

458
00:19:16.830 --> 00:19:18.430
sort of telescoping out if you have

459
00:19:18.430 --> 00:19:20.950
some sort of extreme unfairness say at one step

460
00:19:21.970 --> 00:19:23.730
that it's not going to get worse as time goes on,

461
00:19:23.730 --> 00:19:28.450
but it doesn't guarantee that the system is going

462
00:19:28.450 --> 00:19:29.750
to be working in a way that's moving

463
00:19:29.750 --> 00:19:31.440

you towards say larger social goals.

464
00:19:31.440 --> 00:19:35.530
Like for example, reduced number of people in jail.

465
00:19:35.530 --> 00:19:39.020
And also often when people think about the model

466
00:19:39.020 --> 00:19:40.650
in isolation, don't think a whole lot

467
00:19:40.650 --> 00:19:42.820
about the data generating process.

468
00:19:42.820 --> 00:19:45.320
So for example, what are the inputs to the model?

469
00:19:45.320 --> 00:19:46.180
Where did those come from?

470
00:19:46.180 --> 00:19:50.350
Are those measured in a way that is fair, right?

471
00:19:50.350 --> 00:19:52.720
So one of the examples I can think of is a project

472
00:19:52.720 --> 00:19:55.380
I did looking at the role of overbooking

473
00:19:55.380 --> 00:19:57.950
on these recidivism prediction models.

474
00:19:57.950 --> 00:19:59.880
So basically saying, okay, they take in a bunch

475
00:19:59.880 --> 00:20:02.820
of inputs including person's criminal history

476
00:20:02.820 --> 00:20:05.700
and what the police say, like the booking charges,

477
00:20:05.700 --> 00:20:06.930
what the police say the person did

478
00:20:06.930 --> 00:20:08.610
before they've ever been convicted,

479
00:20:08.610 --> 00:20:09.980
before it's ever been tried,

480
00:20:09.980 --> 00:20:12.960
even for prosecutors really looked at those charges, right?

481
00:20:12.960 --> 00:20:16.430
How often are charges that are ultimately unsubstantiated

482
00:20:16.430 --> 00:20:19.023
by the court system used to,

483
00:20:20.420 --> 00:20:23.120
they are pushed through a model to result

484
00:20:23.120 --> 00:20:25.630
in a higher recommended level of supervision.

485
00:20:25.630 --> 00:20:27.700
So a higher level of being safe supervised

486
00:20:27.700 --> 00:20:31.370
by pretrial services, may be detention based only on charges

487
00:20:31.370 --> 00:20:32.780
that are ultimately unsubstantiated.

488
00:20:32.780 --> 00:20:34.840

So how often the charges the person isn't convicted

489
00:20:34.840 --> 00:20:37.770
or caused them to be recommended for higher,

490
00:20:37.770 --> 00:20:42.710
more sort of serious or punitive conditions.

491
00:20:42.710 --> 00:20:44.170
And the answer was actually pretty high,

492
00:20:44.170 --> 00:20:45.300
it was close to 30%.

493
00:20:45.300 --> 00:20:47.910
And so thinking about all these sorts of things.

494
00:20:47.910 --> 00:20:49.830
but where does the data come from?

495
00:20:49.830 --> 00:20:50.920
Could it be gamed, right?

496
00:20:50.920 --> 00:20:51.753
Could somebody be like,

497
00:20:51.753 --> 00:20:55.170
"Hey, we think this person's needs to be off the street."

498
00:20:55.170 --> 00:20:57.060
We could just sort of tag on some extra charges

499
00:20:57.060 --> 00:20:59.610
and that will then induce them to be recommended

500
00:20:59.610 --> 00:21:02.430
to say not be released, all these sorts of things,

501
00:21:02.430 --> 00:21:06.200
sort of considering both the human component and the model

502
00:21:06.200 --> 00:21:10.440
and sort of effects you wanna see in the real world,

503
00:21:10.440 --> 00:21:12.380
all together, it's complicated.

504
00:21:12.380 --> 00:21:13.760
There are a lot of moving pieces.

505
00:21:13.760 --> 00:21:16.910
I think each component it's important

506
00:21:16.910 --> 00:21:18.730
to evaluate say the model in isolation,

507
00:21:18.730 --> 00:21:20.923
but that's not the end of the story.

508
00:21:24.827 --> 00:21:26.940
<v ->And that's very interesting.</v>

509
00:21:26.940 --> 00:21:29.800
So many commercial companies prefer to keep the details

510
00:21:29.800 --> 00:21:32.280
of the algorithms they use confidential,

511
00:21:32.280 --> 00:21:33.510
there only seem to be a lot of standards

512
00:21:33.510 --> 00:21:35.320
for their construction and operation,

513
00:21:35.320 --> 00:21:37.220

```
do you see this changing anytime soon?

514
00:21:38.720 --> 00:21:39.885
<v ->No.</v>

515
00:21:39.885 --> 00:21:42.552
(laughs loudly)

516
00:21:43.495 --> 00:21:46.870
I think that's accurate and likely to continue.

517
00:21:52.570 --> 00:21:53.960
<v ->So I know that you're also interested</v>

518
00:21:53.960 --> 00:21:56.350
in some alternative interpretations of data.

519
00:21:56.350 --> 00:21:57.690
Can you tell us a bit about your work

520
00:21:57.690 --> 00:21:59.470
in this area and some of the guidelines

521
00:21:59.470 --> 00:22:01.793
around exploring alternative interpretations?

522
00:22:04.200 --> 00:22:05.130
<v ->'Cause I can sort of take this back</v>

523
00:22:05.130 --> 00:22:06.880
to the predictive policing context, right?

524
00:22:06.880 --> 00:22:09.760
People usually think of police records of crime is sort

525
00:22:09.760 --> 00:22:11.520
of a ground truth measure
```

526
00:22:11.520 --> 00:22:13.700
of where crime is occurring, right.

527
00:22:13.700 --> 00:22:15.160
And that's sort of the premise

528
00:22:15.160 --> 00:22:17.850
of using them to make predictions about future crime, right?

529
00:22:17.850 --> 00:22:19.690
And so I think one of the alternative lenses is

530
00:22:19.690 --> 00:22:21.710
that these are records

531
00:22:21.710 --> 00:22:23.640
of where police are enforcing crimes

532
00:22:23.640 --> 00:22:25.100
and sort of leaving it at that, right?

533
00:22:25.100 --> 00:22:28.110
Like this doesn't necessarily tell us a whole lot

534
00:22:28.110 --> 00:22:31.780
about where crimes are occurring or where the problems are.

535
00:22:31.780 --> 00:22:34.240
And so I think that's one example.

536
00:22:34.240 --> 00:22:38.380
I think often the approach I take is to use

537
00:22:38.380 --> 00:22:41.030
some of the same tools, say of data science

538
00:22:41.030 --> 00:22:43.440

or sort of the type of tools that are used

539
00:22:43.440 --> 00:22:44.890
in these areas and sort of trying to turn

540
00:22:44.890 --> 00:22:46.040
it around in that way, like, okay,

541
00:22:46.040 --> 00:22:49.870
say what would happen if we interpret this data differently?

542
00:22:49.870 --> 00:22:53.010
Or how can we use the same sort of mathematical tools

543
00:22:53.010 --> 00:22:56.440
to like highlight issues with some population

544
00:22:56.440 --> 00:22:58.690
who's often not represented?

545
00:22:58.690 --> 00:23:01.920
So another example of this is a recent project I did

546
00:23:01.920 --> 00:23:04.740
in collaboration with some researchers at the ACLU

547
00:23:04.740 --> 00:23:09.740
and Eric Lofgren at Washington State University

548
00:23:09.880 --> 00:23:11.530
and Nina Fefferman at University of Tennessee,

549
00:23:11.530 --> 00:23:14.850
both epidemiologists to model the spread of COVID

550
00:23:14.850 --> 00:23:17.350
within in between jails and communities.

551
00:23:17.350 --> 00:23:18.750
The idea being that there are all these sorts

552
00:23:18.750 --> 00:23:20.900
of epidemiological models out there going

553
00:23:20.900 --> 00:23:23.990
on to make predictions and drive policy

554
00:23:23.990 --> 00:23:25.036
about what we should be doing

555
00:23:25.036 --> 00:23:29.930
to deal with what's going on with COVID

556
00:23:29.930 --> 00:23:32.230
or this is especially true

557
00:23:32.230 --> 00:23:34.550
during the beginning of the pandemic.

558
00:23:34.550 --> 00:23:38.180
And although epidemiologists and public health folks were

559
00:23:38.180 --> 00:23:39.718
sort of already ringing the alarm bells

560
00:23:39.718 --> 00:23:43.300
that jails are often hotbeds of infection.

561
00:23:43.300 --> 00:23:45.540
In fact, because of very poorest places,

562
00:23:45.540 --> 00:23:48.980
people will travel back and forth between them quite a bit,

563
00:23:48.980 --> 00:23:50.660

```
that can drive infection in the community.

564
00:23:50.660 --> 00:23:53.200
We wanted to be able to use some of those tools

565
00:23:53.200 --> 00:23:57.840
of computational epi to sort of highlight how big

566
00:23:57.840 --> 00:23:59.690
of a problem this could be.

567
00:23:59.690 --> 00:24:01.160
And so that sort of another,

568
00:24:01.160 --> 00:24:03.120
I don't know if it's never another interpretation on data,

569
00:24:03.120 --> 00:24:06.620
but it's sort of another, using those types of tools

570
00:24:06.620 --> 00:24:08.780
like data-based or computational tools

571
00:24:08.780 --> 00:24:11.333
to highlight different problems in the world.

572
00:24:14.854 --> 00:24:17.510
<v ->Great, so data missing from a data set is</v>

573
00:24:17.510 --> 00:24:19.610
something that data scientists frequently do

574
00:24:19.610 --> 00:24:20.990
and something that you've noted

575
00:24:20.990 --> 00:24:24.200
that you were often just as interested in if not,
```

576
00:24:24.200 --> 00:24:25.060
can you tell us a little bit

577
00:24:25.060 --> 00:24:27.120
about how you approach this challenge?

578
00:24:27.120 --> 00:24:28.260
<v ->Yeah, so I can tell you a little bit</v>

579
00:24:28.260 --> 00:24:31.000
about the organization I worked for,

580
00:24:31.000 --> 00:24:32.890
the Human Rights Data Analysis Group,

581
00:24:32.890 --> 00:24:34.740
usually abbreviated as HRDAG.

582
00:24:34.740 --> 00:24:38.040
And while I was there how I approached this

583
00:24:38.040 --> 00:24:40.160
because this has been like a large part of my career

584
00:24:40.160 --> 00:24:43.160
at this point, I think it's actually pretty interesting.

585
00:24:43.160 --> 00:24:47.080
So the idea there was that people in times

586
00:24:47.080 --> 00:24:50.010
of say civil conflicts often it's really hard

587
00:24:50.010 --> 00:24:52.010
to get good data on the number

588
00:24:52.010 --> 00:24:54.573

of people who've been killed or disappeared.

589
00:24:55.910 --> 00:24:56.980
The reasons are numerous.

590
00:24:56.980 --> 00:24:58.860
It could be things like perpetrators

591
00:24:58.860 --> 00:25:02.390
intentionally hiding bodies, just lack of infrastructure

592
00:25:02.390 --> 00:25:06.530
to record and just things can be kind of hectic

593
00:25:07.840 --> 00:25:09.163
during a conflict, right.

594
00:25:10.240 --> 00:25:11.550
And the reason I think

595
00:25:11.550 --> 00:25:14.490
that it's important to have a good sense on what's missing

596
00:25:14.490 --> 00:25:17.680
from the data is that if you don't account for that,

597
00:25:17.680 --> 00:25:20.170
you can tell very different stories.

598
00:25:20.170 --> 00:25:22.060
The data can tell you very different

599
00:25:22.060 --> 00:25:23.220
and misleading stories, right?

600
00:25:23.220 --> 00:25:24.053
So for example,

601
00:25:24.053 --> 00:25:28.870
if we have say one perpetrator organization hiding bodies,

602
00:25:28.870 --> 00:25:31.150
for example, then you were to do some analysis

603
00:25:31.150 --> 00:25:33.430
on how many people were killed and by whom,

604
00:25:33.430 --> 00:25:34.740
you might find that the people who are

605
00:25:34.740 --> 00:25:36.560
actually killing way more people,

606
00:25:36.560 --> 00:25:39.240
so committing many more human rights abuses,

607
00:25:39.240 --> 00:25:42.160
had fewer recorded killings.

608
00:25:42.160 --> 00:25:43.370
And then you might conclude that they were

609
00:25:43.370 --> 00:25:44.963
actually responsible for fewer.

610
00:25:46.020 --> 00:25:47.440
It's also, I think important

611
00:25:47.440 --> 00:25:49.610
for say retrospective policy analysis.

612
00:25:49.610 --> 00:25:51.870
So if you have, for example,

613
00:25:51.870 --> 00:25:54.930

```
one region where there just hasn't been a lot

614
00:25:54.930 --> 00:25:56.790
of documentation, perhaps it's more rural,

615
00:25:56.790 --> 00:25:57.970
that's something that we see a lot,

616
00:25:57.970 --> 00:25:59.983
this sort of urban, rural bias.

617
00:26:01.000 --> 00:26:02.477
Then if you wanted to look at,

618
00:26:02.477 --> 00:26:05.100
"Hey, how did this policy change if it was say implemented

619
00:26:05.100 --> 00:26:06.910
across time and space differently,

620
00:26:06.910 --> 00:26:08.840
the number of killings that were occurring."

621
00:26:08.840 --> 00:26:13.170
You might also make untrue or erroneous conclusions

622
00:26:13.170 --> 00:26:15.920
about what was effective because what you might be picking

623
00:26:15.920 --> 00:26:18.760
up on is some combination

624
00:26:18.760 --> 00:26:21.440
of people just aren't being recorded as much over here.

625
00:26:21.440 --> 00:26:24.240
And maybe the impact was just different over time.
```

```
626
00:26:24.240 --> 00:26:25.202
And these are all things we've seen.

627
00:26:25.202 --> 00:26:28.720
We've also seen things like, for example,

628
00:26:28.720 --> 00:26:31.360
a project in Columbia that we were working on

629
00:26:31.360 --> 00:26:33.400
and we got one data set on the number of killings,

630
00:26:33.400 --> 00:26:35.720
and it was really pretty flat

631
00:26:35.720 --> 00:26:37.800
over a very long stretch of time.

632
00:26:37.800 --> 00:26:39.520
And we were like, "That doesn't make a lot of sense."

633
00:26:39.520 --> 00:26:42.090
from what we're hearing from and what we understand

634
00:26:42.090 --> 00:26:44.870
of the conflict and what our partners who live

635
00:26:44.870 --> 00:26:47.070
there have said it, that's not exactly true.

636
00:26:47.070 --> 00:26:49.320
There have been sort of spikes

637
00:26:49.320 --> 00:26:50.840
in violence during this conflict.

638
00:26:50.840 --> 00:26:51.673
```

```
What is this?

639
00:26:51.673 --> 00:26:53.010
And so as we dug into it more

640
00:26:53.010 --> 00:26:55.780
and we spoke to the people in charge of the data,

641
00:26:55.780 --> 00:26:57.077
they were basically like,

642
00:26:57.077 --> 00:26:58.620
"Yeah, I mean, that makes perfect sense."

643
00:26:58.620 --> 00:26:59.810
This is an organization

644
00:26:59.810 --> 00:27:02.840
that actually investigates the killings.

645
00:27:02.840 --> 00:27:06.060
And so they just have an organizational capacity

646
00:27:06.060 --> 00:27:09.180
of X and not exactly X, right.

647
00:27:09.180 --> 00:27:12.320
But you know about X and so if it exceeds that capacity,

648
00:27:12.320 --> 00:27:14.180
well, those things don't make it into the database.

649
00:27:14.180 --> 00:27:16.677
And so you look at it across time and you could conclude,

650
00:27:16.677 --> 00:27:20.040
"Okay, this is like really a pretty steady conflict.
```

651
00:27:20.040 --> 00:27:21.460
There weren't really spikes and violence."

652
00:27:21.460 --> 00:27:22.810
But then if you try to correct for that

653
00:27:22.810 --> 00:27:25.690
and understand what's missing,

654
00:27:25.690 --> 00:27:29.820
you would instead find the sort of dynamics

655
00:27:29.820 --> 00:27:31.890
of the conflict were very different then

656
00:27:31.890 --> 00:27:32.723
than what you thought.

657
00:27:32.723 --> 00:27:35.507
And so the approach that we took at HRDAG

658
00:27:35.507 --> 00:27:36.950
and they still take

659
00:27:36.950 --> 00:27:38.300
and I still interested in this area though

660
00:27:38.300 --> 00:27:41.470
I haven't worked on it for about the past year is

661
00:27:41.470 --> 00:27:43.600
to basically do population estimation.

662
00:27:43.600 --> 00:27:45.820
Sometimes this is called capture recapture,

663
00:27:45.820 --> 00:27:48.080

though because of the context area,

664
00:27:48.080 --> 00:27:50.640
that term can be a little bit misleading

665
00:27:50.640 --> 00:27:53.580
because people are actually captured and we're not

666
00:27:53.580 --> 00:27:56.700
we aren't like capturing humans and releasing them.

667
00:27:56.700 --> 00:27:59.040
But if you're familiar with that terminology,

668
00:27:59.040 --> 00:28:00.160
that's essentially the same,

669
00:28:00.160 --> 00:28:02.640
it's the same statistical methodology.

670
00:28:02.640 --> 00:28:05.570
And essentially, the idea is if you get multiple lists

671
00:28:05.570 --> 00:28:09.050
of people who've been killed, you can look at the overlaps

672
00:28:09.050 --> 00:28:13.010
among them to infer the number that were not recorded

673
00:28:13.010 --> 00:28:14.020
on any of those lists.

674
00:28:14.020 --> 00:28:15.640
So just to give some intuition for that.

675
00:28:15.640 --> 00:28:16.800
So you had two lists

676
00:28:16.800 --> 00:28:19.180
and both of them overlapped a whole lot.

677
00:28:19.180 --> 00:28:20.260
Well, then you might conclude,

678
00:28:20.260 --> 00:28:22.310
that's probably close to the whole universe

679
00:28:22.310 --> 00:28:23.700
of people who've been killed.

680
00:28:23.700 --> 00:28:26.040
Whereas if you have two lists and they barely overlap,

681
00:28:26.040 --> 00:28:27.030
you'd probably conclude

682
00:28:27.030 --> 00:28:29.070
that there are a whole lot more people

683
00:28:29.070 --> 00:28:31.100
in that population that you haven't recorded.

684
00:28:31.100 --> 00:28:34.570
And of course, that sort of logic is formalized

685
00:28:34.570 --> 00:28:35.600
in statistical models.

686
00:28:35.600 --> 00:28:37.010
It's not just us looking at overlaps

687
00:28:37.010 --> 00:28:39.450
and being like, "1,400," right?

688
00:28:39.450 --> 00:28:42.020

Like it's actually comes out of statistical models

689
00:28:42.020 --> 00:28:45.020
and those of course have their own assumptions

690
00:28:46.923 --> 00:28:48.920
that you need to deal with.

691
00:28:48.920 --> 00:28:52.460
But in any case that's the general idea

692
00:28:52.460 --> 00:28:54.460
to get a better sort of statistical handle

693
00:28:54.460 --> 00:28:56.593
on what's missing from your data.

694
00:28:59.548 --> 00:29:01.430
<v ->So you have worked on a number of very important</v>

695
00:29:01.430 --> 00:29:03.180
and interesting projects.

696
00:29:03.180 --> 00:29:05.190
So what are some of the most important lessons

697
00:29:05.190 --> 00:29:07.340
you've learned so far in your career?

698
00:29:07.340 --> 00:29:08.610
And if you had to do things over,

699
00:29:08.610 --> 00:29:10.110
what would you do differently?

700
00:29:11.372 --> 00:29:13.660
<v ->What are some of the most, let me see,</v>

701
00:29:13.660 --> 00:29:15.560
what are the most important lessons

702
00:29:15.560 --> 00:29:17.270
I've learned in my career?

703
00:29:17.270 --> 00:29:19.500
I think, especially in the areas that I work

704
00:29:19.500 --> 00:29:20.650
on when we're talking about data

705
00:29:20.650 --> 00:29:22.453
that represents real people,

706
00:29:23.580 --> 00:29:26.880
I think treating the data with the respect it deserves

707
00:29:26.880 --> 00:29:29.750
like not sort of, I think it's often easy

708
00:29:29.750 --> 00:29:32.410
as data scientists to just get lost in like the X's

709
00:29:32.410 --> 00:29:36.630
and Y's in sort of it's just a database of numbers

710
00:29:36.630 --> 00:29:39.750
sort of abstracting away because to some extent,

711
00:29:39.750 --> 00:29:42.100
I think to do that sort work day in and day out,

712
00:29:42.100 --> 00:29:44.300
you kind of have to, but at the end of the day,

713
00:29:44.300 --> 00:29:46.390

I think remembering that each,

714
00:29:46.390 --> 00:29:48.337
say row in your dataset represents a real person

715
00:29:48.337 --> 00:29:50.600
and a real person's story

716
00:29:50.600 --> 00:29:53.560
and sort of treating it with the care it deserves

717
00:29:53.560 --> 00:29:57.640
when you say describe the data, when you talk about results,

718
00:29:57.640 --> 00:30:01.110
when you make recommendations for what say to do

719
00:30:01.110 --> 00:30:02.033
in the real world.

720
00:30:03.126 --> 00:30:05.550
I think that really goes a long way

721
00:30:05.550 --> 00:30:07.930
and that's something I've come to appreciate

722
00:30:07.930 --> 00:30:10.593
more and more over time is that,

723
00:30:12.050 --> 00:30:14.590
yes, when you say like write down a regression model

724
00:30:14.590 --> 00:30:16.040
or something, there's something very abstract

725
00:30:16.040 --> 00:30:19.370
about that, and that can help you to understand the world,

```
726
00:30:19.370 --> 00:30:20.850
but at the end of the day, like the data is

727
00:30:20.850 --> 00:30:22.140
representing real people.

728
00:30:22.140 --> 00:30:24.030
And so having some care about that,

729
00:30:24.030 --> 00:30:25.530
I think goes a really long way

730
00:30:28.510 --> 00:30:30.620
<v ->That's very important to keep in mind.</v>

731
00:30:30.620 --> 00:30:32.780
So what is the best and worst career advice

732
00:30:32.780 --> 00:30:33.830
you've ever received?

733
00:30:37.269 --> 00:30:38.943
<v ->I don't think I've received a lot of career advice.</v>

734
00:30:38.943 --> 00:30:40.530
And I'm gonna be honest,

735
00:30:40.530 --> 00:30:43.080
maybe I should be asking for it more, I don't know.

736
00:30:47.780 --> 00:30:49.700
I guess I'll say I can tell you what's been good

737
00:30:49.700 --> 00:30:52.520
and bad about my own career strategy.

738
00:30:52.520 --> 00:30:53.620
```

And I don't know if anyone's advice,

739
00:30:53.620 --> 00:30:55.630
maybe this is extremely unadvisable.

740
00:30:55.630 --> 00:30:56.463
I don't know,

741
00:30:56.463 --> 00:31:01.463
but I don't think anyone's really told me not to do this.

742
00:31:02.010 --> 00:31:04.360
And it's basically that I've kind of just followed

743
00:31:04.360 --> 00:31:05.290
what I think is interesting.

744
00:31:05.290 --> 00:31:06.123
Like you said, I've worked

745
00:31:06.123 --> 00:31:07.970
on some pretty interesting projects.

746
00:31:07.970 --> 00:31:11.050
I've been kind of less concerned

747
00:31:11.050 --> 00:31:14.280
about say like career advancement

748
00:31:14.280 --> 00:31:18.133
or clawing my way to the top of something.

749
00:31:20.030 --> 00:31:21.780
I honestly think that served me really well.

750
00:31:21.780 --> 00:31:26.650
Like I have really enjoyed the journey.

751
00:31:26.650 --> 00:31:29.000
I really enjoy all the projects that I work on.

752
00:31:30.450 --> 00:31:31.770
And so I think I have,

753
00:31:31.770 --> 00:31:33.520
actually I have been encouraged

754
00:31:33.520 --> 00:31:36.033
to do that in the past, maybe not explicitly.

755
00:31:37.407 --> 00:31:38.280
Is that good advice?

756
00:31:38.280 --> 00:31:39.480
Yeah, I don't know, right?

757
00:31:39.480 --> 00:31:42.550
Like I think also as you start looking around it,

758
00:31:42.550 --> 00:31:45.050
your peers, you might find if you take that approach

759
00:31:45.050 --> 00:31:46.850
that people are say getting tenure

760
00:31:46.850 --> 00:31:49.600
and you're kind of still drifting around.

761
00:31:49.600 --> 00:31:51.350
And so if that's something that's important,

762
00:31:51.350 --> 00:31:54.670
maybe a more linear path is a better choice,

763
00:31:54.670 --> 00:31:56.050

```
but what I can say is I'm pretty happy

764
00:31:56.050 --> 00:31:57.650
with how I've been doing things.

765
00:32:00.990 --> 00:32:04.103
<v ->Well, it sounds like you've had a great career so far.</v>

766
00:32:04.103 --> 00:32:05.030
<v ->Thanks.</v>

767
00:32:05.030 --> 00:32:07.990
<v ->Based on your past projects and experiences,</v>

768
00:32:07.990 --> 00:32:09.420
where do you see data science

769
00:32:09.420 --> 00:32:11.920
and addiction research intersecting in the future?

770
00:32:18.530 --> 00:32:20.510
<v ->I think data science really has the power</v>

771
00:32:20.510 --> 00:32:25.255
to sort of touch all facets of any sort of science, right?

772
00:32:25.255 --> 00:32:28.817
Any sort of social impact sort of areas where there's data

773
00:32:28.817 --> 00:32:31.760
and there's clearly data in this area,

774
00:32:31.760 --> 00:32:32.970
I guess just sort of going back to one

775
00:32:32.970 --> 00:32:35.090
of the things I was saying earlier,
```

776
00:32:35.090 --> 00:32:37.330
the way that drug addiction

777
00:32:37.330 --> 00:32:38.780
and drug use has been criminalized.

778
00:32:38.780 --> 00:32:41.900
I think it's reasonable to think that the data

779
00:32:41.900 --> 00:32:43.520
that sort of traditional perspectives

780
00:32:43.520 --> 00:32:45.340
on the data maybe aren't the full story.

781
00:32:45.340 --> 00:32:49.680
And so using that data or using other perspectives

782
00:32:49.680 --> 00:32:52.530
to look at it will be an important direction to go.

783
00:32:52.530 --> 00:32:54.920
It's also very likely missing a whole lot of things.

784
00:32:54.920 --> 00:32:56.750
So thinking about what is missing,

785
00:32:56.750 --> 00:32:58.100
what's systematically missing,

786
00:32:58.100 --> 00:32:59.430
how might that change the story

787
00:32:59.430 --> 00:33:03.350
that you're pulling out of the data will be

788
00:33:03.350 --> 00:33:05.483

```
an important avenue going forward.

789
00:33:08.790 --> 00:33:09.733
<v ->Are there any areas you would</v>

790
00:33:09.733 --> 00:33:11.773
like to see this research expanded?

791
00:33:12.870 --> 00:33:15.660
<v ->Data science or addiction research?</v>

792
00:33:15.660 --> 00:33:17.560
<v ->Data science for addiction research.</v>

793
00:33:18.970 --> 00:33:21.310
<v ->Honestly, this is not really my area of expertise,</v>

794
00:33:21.310 --> 00:33:22.143
so I will leave

795
00:33:22.143 --> 00:33:24.430
that to the folks who know what they're talking about.

796
00:33:24.430 --> 00:33:28.930
I think that's always a good idea to not talk

797
00:33:28.930 --> 00:33:30.620
when you don't know what you're talking about.

798
00:33:30.620 --> 00:33:32.400
So I'll exercise that

799
00:33:33.387 --> 00:33:36.100
<v ->Are there any areas though for, I guess, data science</v>

800
00:33:36.100 --> 00:33:38.420
in general, that you'd like to see expanded then?
```

```
801
00:33:39.850 --> 00:33:41.610
<v ->I mean, I really like this area</v>

802
00:33:41.610 --> 00:33:44.930
of data science where you kind of...

803
00:33:44.930 --> 00:33:49.930
Data science and the advocacy world, data science to sort of

804
00:33:50.180 --> 00:33:53.170
like more of more, this is a noxious word,

805
00:33:53.170 --> 00:33:55.630
so be prepared but like bespoke type of analysis

806
00:33:55.630 --> 00:33:57.230
where you're like, "Hey, let's take

807
00:33:57.230 --> 00:33:58.440
on this like very kind of niche

808
00:33:58.440 --> 00:34:00.780
and small question and like throw like the

809
00:34:00.780 --> 00:34:02.770
most powerful tools at it to see what we can get."

810
00:34:02.770 --> 00:34:05.110
I think for a while we've been sort of focused

811
00:34:05.110 --> 00:34:10.110
on scaling things and small batch, artisanal data science,

812
00:34:10.500 --> 00:34:12.003
I think is pretty cool.

813
00:34:15.930 --> 00:34:17.800
```

```
<v ->And lastly, are there any areas</v>

814
00:34:17.800 --> 00:34:20.283
you'd like to see NIH doing in data science?

815
00:34:23.740 --> 00:34:25.280
<v ->I mean, everything, right?</v>

816
00:34:25.280 --> 00:34:27.480
I don't know if I can narrow it down really.

817
00:34:30.700 --> 00:34:32.990
<v ->That's fine, it's been great talking to you this morning.</v>

818
00:34:32.990 --> 00:34:34.287
We really enjoyed hearing about your career

819
00:34:34.287 --> 00:34:37.070
and your experiences and your advice and everything.

820
00:34:37.070 --> 00:34:37.903
So thank you so much,

821
00:34:37.903 --> 00:34:39.600
I wanna have everyone clap for you,

822
00:34:39.600 --> 00:34:41.210
I know it's hard to do in this virtual environment.

823
00:34:41.210 --> 00:34:42.664
<v ->Thank you.</v>

824
00:34:42.664 --> 00:34:45.000
<v ->And next, we're gonna be having a presentation</v>

825
00:34:45.000 --> 00:34:48.250
from Dr. Brenda Curtis from the NIDA IRP.
```

```
826
00:34:48.250 --> 00:34:50.130
So Dr. Brenda Curtis is the chief

827
00:34:50.130 --> 00:34:52.620
of technology and translational research unit

828
00:34:52.620 --> 00:34:54.950
of the NIDA intramural research program.

829
00:34:54.950 --> 00:34:57.030
She earned both a bachelor's degree in biology

830
00:34:57.030 --> 00:34:58.233
and a master's degree in public health

831
00:34:58.233 --> 00:35:00.050
from the University of Illinois

832
00:35:00.050 --> 00:35:01.150
and then obtained her doctorate

833
00:35:01.150 --> 00:35:03.560
in communications from the University of Pennsylvania

834
00:35:03.560 --> 00:35:05.230
where she most recently held the appointment

835
00:35:05.230 --> 00:35:07.520
of assistant professor of psychology

836
00:35:07.520 --> 00:35:08.810
and psychiatry addictions

837
00:35:08.810 --> 00:35:10.730
at the Perelman School of Medicine.

838
00:35:10.730 --> 00:35:12.750
```

```
Dr. Curtis also completed a fellowship

839
00:35:12.750 --> 00:35:13.870
in the Fordham University,

840
00:35:13.870 --> 00:35:16.120
HIV and Drug Abuse Prevention Research

841
00:35:16.120 --> 00:35:17.740
Ethics Training Institute.

842
00:35:17.740 --> 00:35:18.990
Her training in public health

843
00:35:18.990 --> 00:35:20.790
and health communication allows her

844
00:35:20.790 --> 00:35:22.470
to employ a public health approach

845
00:35:22.470 --> 00:35:24.220
by using effective communication techniques

846
00:35:24.220 --> 00:35:26.650
to ensure recruitment and retention rates are achieved.

847
00:35:26.650 --> 00:35:29.900
Her research focuses translational leveraging social media

848
00:35:29.900 --> 00:35:32.390
and big data methodology to form the development,

849
00:35:32.390 --> 00:35:34.300
evaluation, and implementation

850
00:35:34.300 --> 00:35:36.750
of technology-based tools that address substance use
```

```
851
00:35:36.750 --> 00:35:39.530
and related conditions such as HIV&amp;AIDS.

852
00:35:39.530 --> 00:35:41.580
Dr. Curtis employees multiple methodologies

853
00:35:41.580 --> 00:35:43.970
to facilitate the flow of scientific discovery

854
00:35:43.970 --> 00:35:45.780
to practical application allowing

855
00:35:45.780 --> 00:35:48.200
her to not only reach underserved populations,

856
00:35:48.200 --> 00:35:49.590
but to design health monitoring

857
00:35:49.590 --> 00:35:51.800
and behavioral change interventions

858
00:35:51.800 --> 00:35:54.610
that are user-centered, inclusive, and evidence-based.

859
00:35:54.610 --> 00:35:57.860
So please join me in welcoming Dr. Brenda Curtis,

860
00:35:57.860 --> 00:35:59.090
virtual applause here,

861
00:35:59.090 --> 00:36:01.480
and you can go ahead and share your sides now.

862
00:36:01.480 --> 00:36:02.313
<v ->Okay.</v>

863
00:36:08.320 --> 00:36:10.640
```

```
So thank you for inviting me

864
00:36:10.640 --> 00:36:14.330
and I'm gonna walk you through kind of start off

865
00:36:14.330 --> 00:36:18.223
with the why I do this and kind of how I got here.

866
00:36:20.070 --> 00:36:21.323
Let me move.

867
00:36:22.490 --> 00:36:23.770
Okay, slides moving?

868
00:36:23.770 --> 00:36:25.120
Yes, okay.

869
00:36:25.120 --> 00:36:29.790
So I am from a small town in East St. Louis, Illinois.

870
00:36:29.790 --> 00:36:31.160
It's across from Illinois.

871
00:36:31.160 --> 00:36:34.540
So sorry, across from Missouri, St. Louis.

872
00:36:34.540 --> 00:36:35.373
And it's

873
00:36:43.630 --> 00:36:44.893
just got a warning,

874
00:36:49.220 --> 00:36:51.673
had arise,

875
00:36:53.424 --> 00:36:54.757
I'm getting RMS.
```

```
876
00:36:58.420 --> 00:37:00.570
<v ->It sounds like your audio is cutting out.</v>

877
00:37:04.660 --> 00:37:05.493
<v ->Interesting.</v>

878
00:37:13.970 --> 00:37:15.213
<v Susan>Your Internet, girl.</v>

879
00:37:20.780 --> 00:37:22.380
Hello.

880
00:37:22.380 --> 00:37:23.220
<v ->We can still hear you,</v>

881
00:37:23.220 --> 00:37:25.847
it sounds like audio is cutting out little bit.

882
00:37:25.847 --> 00:37:27.851
<v ->Yeah, I lost Internet connection, I will do it.</v>

883
00:37:27.851 --> 00:37:28.690
<v Susan>Okay.</v>

884
00:37:28.690 --> 00:37:30.670
<v ->Let me try again.</v>

885
00:37:30.670 --> 00:37:31.503
<v Susan>Yeah.</v>

886
00:37:31.503 --> 00:37:33.303
<v ->And hopefully, this time it will work.</v>

887
00:37:36.090 --> 00:37:37.890
So I'm from a small town

888
00:37:37.890 --> 00:37:41.520
```

and in a town that was having really high poverty

889
00:37:41.520 --> 00:37:43.440
and really high crime.

890
00:37:43.440 --> 00:37:47.600
And when I started off in it,

891
00:37:47.600 --> 00:37:50.640
I left to go to undergrad and my undergrad degree was

892
00:37:50.640 --> 00:37:54.293
in biology and I planned on being a doctor.

893
00:37:56.310 --> 00:37:58.140
I kind of didn't really have

894
00:37:58.140 --> 00:38:00.930
or know of many like career paths.

895
00:38:00.930 --> 00:38:04.110
I knew doctor, lawyer, accountant, things like that.

896
00:38:04.110 --> 00:38:06.560
And so I went to the University of Illinois

897
00:38:06.560 --> 00:38:10.623
in Urbana-Champagne and wanted to go pre-med.

898
00:38:11.620 --> 00:38:14.690
I had always my whole life been interested in science,

899
00:38:14.690 --> 00:38:16.660
I can not remember a name to save my life

900
00:38:16.660 --> 00:38:19.400
and I cannot remember a date to save my life.

901
00:38:19.400 --> 00:38:23.210
But math was this universal language to me,

902
00:38:23.210 --> 00:38:27.220
and my mom, she had three children,

903
00:38:27.220 --> 00:38:29.280
she had her first child

904
00:38:29.280 --> 00:38:31.120
while she was in high school

905
00:38:31.120 --> 00:38:33.730
and had to drop out and then had to go back

906
00:38:33.730 --> 00:38:34.913
to get to finish.

907
00:38:37.108 --> 00:38:41.700
I'm the youngest and she was trying to go back to school

908
00:38:41.700 --> 00:38:44.150
so that she could get a promotion at work.

909
00:38:44.150 --> 00:38:48.700
And so had three children, managing a full-time job,

910
00:38:48.700 --> 00:38:52.670
a house and everything, she would leave her homework out

911
00:38:52.670 --> 00:38:55.810
and as to be a great daughter, be supportive,

912
00:38:55.810 --> 00:38:58.590
I would sit there and sometimes I would just do

913
00:38:58.590 --> 00:38:59.860

```
her homework for her.

914
00:38:59.860 --> 00:39:00.747
And she'd come and be like,

915
00:39:00.747 --> 00:39:02.040
"You're not supposed to do my homework,"

916
00:39:02.040 --> 00:39:04.070
but then of course she'd be extremely happy,

917
00:39:04.070 --> 00:39:06.770
but so I would always do her homework for her.

918
00:39:06.770 --> 00:39:08.880
And I just kind of had a knack for math,

919
00:39:08.880 --> 00:39:10.760
and I used to do like math competitions

920
00:39:10.760 --> 00:39:12.270
and different things.

921
00:39:12.270 --> 00:39:14.660
So when I went away for undergrad

922
00:39:14.660 --> 00:39:17.680
and I mean, I grew up in poverty and I grew up poor,

923
00:39:17.680 --> 00:39:19.443
but a lot of people do.

924
00:39:21.000 --> 00:39:24.070
When I came back, I saw a community

925
00:39:24.070 --> 00:39:26.320
that was being devastated.
```

```
926
00:39:26.320 --> 00:39:28.640
We've always had poverty that was there,

927
00:39:28.640 --> 00:39:32.250
but crack cocaine, heroin took over

928
00:39:32.250 --> 00:39:34.580
and we had always had alcohol problems.

929
00:39:34.580 --> 00:39:38.470
It really took over and the drugs fit into the crime.

930
00:39:38.470 --> 00:39:42.910
And I initially, when I came out, I was doing work

931
00:39:42.910 --> 00:39:45.340
at Washington University in St. Louis

932
00:39:45.340 --> 00:39:47.430
in one other transplant units.

933
00:39:47.430 --> 00:39:52.220
And I would hear from family and friends

934
00:39:52.220 --> 00:39:57.050
about a gang shootout or some other problems.

935
00:39:57.050 --> 00:39:58.193
And then I would be called

936
00:39:58.193 --> 00:40:01.630
into the hospital to receive organs.

937
00:40:01.630 --> 00:40:04.890
And it kind of was a connection

938
00:40:04.890 --> 00:40:07.560
```

that I was having a hard time making.

939
00:40:07.560 --> 00:40:12.560
And I decided to leave doing research at Wash U

940
00:40:13.640 --> 00:40:18.640
and to work at a drug treatment center doing HIV counseling

941
00:40:19.090 --> 00:40:20.763
as well as drug treatment work.

942
00:40:21.880 --> 00:40:24.180
I wanted to give back to my community.

943
00:40:24.180 --> 00:40:27.250
And so my history kinda always starts from that,

944
00:40:27.250 --> 00:40:31.670
figuring out a way that I could give back to my community.

945
00:40:31.670 --> 00:40:35.170
One of the biggest problems that I saw is

946
00:40:35.170 --> 00:40:37.750
that no one was coming to help us

947
00:40:37.750 --> 00:40:40.970
like East St. Louis still today as you saw

948
00:40:40.970 --> 00:40:43.750
in the other slide, we have the highest crime rate

949
00:40:43.750 --> 00:40:46.840
and the community is just devastated

950
00:40:46.840 --> 00:40:48.900
and there wasn't any one coming to solve those,

```
951
00:40:48.900 --> 00:40:51.970
we weren't getting Gates money or some other foundation.

952
00:40:51.970 --> 00:40:55.250
And definitely the state of Illinois was not coming.

953
00:40:55.250 --> 00:40:56.990
And there's lots of communities

954
00:40:56.990 --> 00:40:58.600
like East St. Louis across the US.

955
00:40:58.600 --> 00:41:01.160
And so I want it to figure out ways

956
00:41:01.160 --> 00:41:03.250
to get people the information

957
00:41:03.250 --> 00:41:05.740
and specifically around prevention and treatment.

958
00:41:05.740 --> 00:41:09.160
And I felt like and I still do today feel like the Internet

959
00:41:09.160 --> 00:41:13.623
and smartphones are ways to track, deliver,

960
00:41:15.013 --> 00:41:17.650
and do interventions as well as prevention.

961
00:41:17.650 --> 00:41:19.670
And one of the good things about smart phones is

962
00:41:19.670 --> 00:41:22.330
that everyone kind of has them.

963
00:41:22.330 --> 00:41:25.380
```

And so when you look at national numbers,

964
00:41:25.380 --> 00:41:27.562
you see that there's like 96%

965
00:41:27.562 --> 00:41:29.890
of the us population has a cell phone,

966
00:41:29.890 --> 00:41:32.103
about 81% have a smartphone.

967
00:41:33.010 --> 00:41:35.580
And we started looking across men and women

968
00:41:35.580 --> 00:41:38.160
that those numbers are similar.

969
00:41:38.160 --> 00:41:40.820
Yes, more younger people have a smartphone,

970
00:41:40.820 --> 00:41:43.040
but even as you get older

971
00:41:43.040 --> 00:41:44.903
we have greater populations,

972
00:41:44.903 --> 00:41:47.360
a good percentage have a smartphone.

973
00:41:47.360 --> 00:41:49.620
The age that I'm typically working in

974
00:41:49.620 --> 00:41:50.770
'cause I'm working in the clinic,

975
00:41:50.770 --> 00:41:53.810
it's kind of that 30 to 49 age range.

976
00:41:53.810 --> 00:41:55.780
And a sizable percentage

977
00:41:55.780 --> 00:41:58.143
of them around 92% have a smartphone.

978
00:41:59.240 --> 00:42:01.137
And when we look at, this is the Pew data,

979
00:42:01.137 --> 00:42:03.520
and these are the three ethnicities

980
00:42:03.520 --> 00:42:05.930
that they have represented.

981
00:42:05.930 --> 00:42:08.460
We get pretty good coverage across

982
00:42:10.080 --> 00:42:11.170
having a smartphone.

983
00:42:11.170 --> 00:42:15.810
Where we see lots of disparities is when we start dealing

984
00:42:15.810 --> 00:42:17.210
with the education level.

985
00:42:17.210 --> 00:42:19.790
And that's kind of where we start falling off

986
00:42:19.790 --> 00:42:22.263
and seeing less smartphone ownership.

987
00:42:23.834 --> 00:42:28.270
The smartphones today are super computers of yesterday.

988
00:42:28.270 --> 00:42:29.763

```
And we can do a lot on there.

989
00:42:29.763 --> 00:42:31.650
We can not only track,

990
00:42:31.650 --> 00:42:34.940
but we can deliver and we can monitor.

991
00:42:34.940 --> 00:42:39.080
And so that was kind of how I started was like,

992
00:42:39.080 --> 00:42:41.240
okay I want to get...

993
00:42:41.240 --> 00:42:42.470
No one's coming to help me,

994
00:42:42.470 --> 00:42:44.480
I kind of felt like I'll help my communities.

995
00:42:44.480 --> 00:42:48.387
And so how can I be a person be it to get treatment

996
00:42:48.387 --> 00:42:50.533
and information out there.

997
00:42:51.800 --> 00:42:55.240
And so one of the questions that I started

998
00:42:55.240 --> 00:42:57.393
with was kind of online surveillance.

999
00:42:58.340 --> 00:43:00.700
I initially started with looking at how drugs emerge,

1000
00:43:00.700 --> 00:43:02.440
like synthetic marijuana
```

1001
00:43:02.440 --> 00:43:04.720
that was taking a hold in communities.

1002
00:43:04.720 --> 00:43:08.160
And then I moved to like the online surveillance of alcohol.

1003
00:43:08.160 --> 00:43:10.910
And alcohol, it was when I got my first RO1 grant

1004
00:43:10.910 --> 00:43:13.270
from NIDA, I am a NIDA baby starting

1005
00:43:13.270 --> 00:43:16.970
with a diversity supplement that I met Albert at that time.

1006
00:43:16.970 --> 00:43:19.357
But when I got my first RO1,

1007
00:43:21.154 --> 00:43:23.943
the question was looking at social media language.

1008
00:43:25.300 --> 00:43:27.250
I was doing a clinic study,

1009
00:43:27.250 --> 00:43:30.170
but we wanted to test some of these questions out.

1010
00:43:30.170 --> 00:43:34.580
And so we decided to look at doing a sample online.

1011
00:43:34.580 --> 00:43:39.000
Twitter language is free, it's public.

1012
00:43:39.000 --> 00:43:40.823
And we just had a simple question,

1013
00:43:43.133 --> 00:43:47.682

can Twitter reuse to predict excessive alcohol consumption

1014
00:43:47.682 --> 00:43:49.410
at a county level?

1015
00:43:49.410 --> 00:43:52.670
So instead of doing an individual, we are looking

1016
00:43:52.670 --> 00:43:54.760
at the county level.

1017
00:43:54.760 --> 00:43:58.143
And we did a simple study.

1018
00:43:59.934 --> 00:44:01.060
We did the random sample,

1019
00:44:01.060 --> 00:44:03.690
we collected on 1% of Twitter posts

1020
00:44:03.690 --> 00:44:06.233
and then we collected national survey data.

1021
00:44:08.460 --> 00:44:11.120
One of the reasons I'm asking, can we use Twitter is

1022
00:44:11.120 --> 00:44:12.780
because it's cheaper.

1023
00:44:12.780 --> 00:44:15.550
National surveys are expensive.

1024
00:44:15.550 --> 00:44:17.490
And so one of the things we did was

1025
00:44:17.490 --> 00:44:19.820
we took some national survey data.

```
1026
00:44:19.820 --> 00:44:23.350
We used the behavioral risk factor surveillance system

1027
00:44:23.350 --> 00:44:26.740
across the US including,

1028
00:44:26.740 --> 00:44:31.040
Puerto Rico, DC, and the Virgin, Islands and womp.

1029
00:44:31.040 --> 00:44:33.570
We also collected demographic data

1030
00:44:33.570 --> 00:44:35.760
from the US Census Bureau,

1031
00:44:35.760 --> 00:44:39.583
and we use data from the American Community Survey.

1032
00:44:40.490 --> 00:44:42.240
So one of the questions we wanted to know

1033
00:44:42.240 --> 00:44:44.840
at the basic level is can we out predict

1034
00:44:46.550 --> 00:44:50.000
kind of the the survey data that the information

1035
00:44:50.000 --> 00:44:55.000
that we have already about who is reporting substance use?

1036
00:44:56.040 --> 00:44:57.620
And so we wanted to know,

1037
00:44:57.620 --> 00:45:01.240
could we use social media language to do that?

1038
00:45:01.240 --> 00:45:05.190
```

And the answer was, yes, we could out predict

1039
00:45:06.320 --> 00:45:08.920
just kind of the social demographic variables

1040
00:45:09.780 --> 00:45:11.520
that the survey data.

1041
00:45:11.520 --> 00:45:14.480
So we did some fancy things where you have all the data,

1042
00:45:14.480 --> 00:45:17.317
you train the model, and then you test it, you hold out

1043
00:45:17.317 --> 00:45:20.173
and you test it on another set of the model.

1044
00:45:21.810 --> 00:45:24.540
I mean, I'm not so interested in, can we,

1045
00:45:24.540 --> 00:45:27.020
that's the first question, always, can you?

1046
00:45:27.020 --> 00:45:30.230
But I'm always more interested in what do I gain about it?

1047
00:45:30.230 --> 00:45:34.140
Because my background, my undergrad was in biology.

1048
00:45:34.140 --> 00:45:37.350
My master's is in public health

1049
00:45:37.350 --> 00:45:40.150
and my PhD is in health communication.

1050
00:45:40.150 --> 00:45:43.050
So I'm kind of always looking from this lens

1051
00:45:43.050 --> 00:45:47.340
of from a public health and a health communication person,

1052
00:45:47.340 --> 00:45:50.180
how they can use this information

1053
00:45:50.180 --> 00:45:55.180
to inform treatment, prevention, and message delivery,

1054
00:45:55.780 --> 00:45:58.193
and message persuasiveness to individuals,

1055
00:45:59.330 --> 00:46:02.810
looking to how to change behavior that way.

1056
00:46:02.810 --> 00:46:05.250
So one of the questions is like counties

1057
00:46:05.250 --> 00:46:07.380
that had higher levels

1058
00:46:07.380 --> 00:46:10.903
of excessive drinking compared to counties with lower.

1059
00:46:12.200 --> 00:46:14.040
And one of the things that we see is

1060
00:46:14.040 --> 00:46:16.740
that when we look at the language that's used,

1061
00:46:16.740 --> 00:46:19.760
in counties that have higher levels of excessive drinking,

1062
00:46:19.760 --> 00:46:23.370
they're talking about sports, things like hockey,

1063
00:46:23.370 --> 00:46:25.400

they're talking about going to festivals

1064
00:46:25.400 --> 00:46:28.260
and art films, and folk films.

1065
00:46:28.260 --> 00:46:29.623
Of course, they talk about drinking,

1066
00:46:29.623 --> 00:46:33.640
they're talking about higher education type things,

1067
00:46:33.640 --> 00:46:36.470
research papers, projects, assignments.

1068
00:46:36.470 --> 00:46:39.960
And of course, they're talking about going out on weekends.

1069
00:46:39.960 --> 00:46:43.110
Now, if anyone's wondering, a lot of times in society,

1070
00:46:43.110 --> 00:46:46.480
we have this view of people who,

1071
00:46:46.480 --> 00:46:49.030
certain populations who have a higher levels

1072
00:46:49.030 --> 00:46:49.960
of excessive drinking.

1073
00:46:49.960 --> 00:46:51.410
And hopefully, I think you're saying

1074
00:46:51.410 --> 00:46:54.360
that if you had that view that they were black

1075
00:46:54.360 --> 00:46:57.010
and brown populations, the answer's, no.

1076
00:46:57.010 --> 00:46:59.220
We have higher rates of drinking are

1077
00:46:59.220 --> 00:47:03.330
in higher income, higher educated communities.

1078
00:47:03.330 --> 00:47:04.737
And so that's always interesting

1079
00:47:04.737 --> 00:47:07.170
'cause people sometimes have preconceived notion

1080
00:47:08.474 --> 00:47:12.320
about what we see because of stigma and stereotypes.

1081
00:47:12.320 --> 00:47:15.930
But in here, the language kind of speaks for itself.

1082
00:47:15.930 --> 00:47:19.320
In areas where we have less excessive drinking,

1083
00:47:19.320 --> 00:47:23.360
people are talking about religion, prayer, God,

1084
00:47:23.360 --> 00:47:25.470
doing things with family.

1085
00:47:25.470 --> 00:47:29.310
So one person could say, well, okay, so how is this useful?

1086
00:47:29.310 --> 00:47:32.660
Well, if you're in a community, you could be looking

1087
00:47:32.660 --> 00:47:36.150
at Twitter or you could know what's going on with festivals,

1088
00:47:36.150 --> 00:47:38.540

what sports activities, and you could send out

1089
00:47:38.540 --> 00:47:41.670
as a public health specialist or communicator,

1090
00:47:41.670 --> 00:47:45.770
you could send out harm reduction information on Twitter.

1091
00:47:45.770 --> 00:47:48.763
So that's one type of thing that you can do.

1092
00:47:50.390 --> 00:47:52.490
There's over 3000 counties in the US,

1093
00:47:52.490 --> 00:47:55.220
and we didn't think that counties was necessarily

1094
00:47:55.220 --> 00:47:58.800
the best way to think of this.

1095
00:47:58.800 --> 00:48:01.680
And so in the lab told us

1096
00:48:01.680 --> 00:48:04.820
about something called the American Community Project,

1097
00:48:04.820 --> 00:48:06.320
which was developed by a group

1098
00:48:06.320 --> 00:48:09.170
of researchers at George Washington University.

1099
00:48:09.170 --> 00:48:13.000
And what it is is it groups US counties

1100
00:48:13.000 --> 00:48:15.340
into 15 community types.

1101
00:48:15.340 --> 00:48:19.270
And these community types are based upon 36 demographic,

1102
00:48:19.270 --> 00:48:21.870
social economic, and cultural indicators

1103
00:48:21.870 --> 00:48:24.170
including population density, income, race,

1104
00:48:24.170 --> 00:48:25.510
things like that.

1105
00:48:25.510 --> 00:48:29.910
And so, for example, like you would see Philadelphia, LA,

1106
00:48:29.910 --> 00:48:31.930
New York group together in big city.

1107
00:48:31.930 --> 00:48:35.070
So it doesn't depend upon the actual space

1108
00:48:35.070 --> 00:48:37.690
that some were our proximity, that's not where it is,

1109
00:48:37.690 --> 00:48:39.460
is looking at the features.

1110
00:48:39.460 --> 00:48:42.540
And so we argued that clustering this way would give

1111
00:48:42.540 --> 00:48:47.320
us more information culturally, and kind of to be able

1112
00:48:47.320 --> 00:48:51.160
to start looking at excessive alcohol consumption

1113
00:48:51.160 --> 00:48:53.513

```
and more specifically, how to target that.

1114
00:48:54.480 --> 00:48:58.000
And here's just a sample of some of the results.

1115
00:48:58.000 --> 00:49:00.880
Using the American Communities Project,

1116
00:49:00.880 --> 00:49:04.610
we can say how these four communities talk

1117
00:49:04.610 --> 00:49:08.243
about alcohol use and drinking differently.

1118
00:49:09.360 --> 00:49:11.850
African-American communities are discussing drinking

1119
00:49:11.850 --> 00:49:16.043
in the context of night clubs and dance places.

1120
00:49:17.090 --> 00:49:19.230
College towns, not too surprising,

1121
00:49:19.230 --> 00:49:21.170
they're talking about drinking on weekends,

1122
00:49:21.170 --> 00:49:22.470
they're doing drinking humor,

1123
00:49:22.470 --> 00:49:24.540
and they do a lot of sex related topics

1124
00:49:24.540 --> 00:49:26.073
in relationship to drinking.

1125
00:49:27.310 --> 00:49:30.690
Hispanics centers were talking more about drinking
```

```
1126
00:49:30.690 --> 00:49:33.880
with family and family was related to drinking.

1127
00:49:33.880 --> 00:49:36.470
Whereas evangelical hubs were talking

1128
00:49:36.470 --> 00:49:39.430
more about their referencing sobriety

1129
00:49:39.430 --> 00:49:44.430
and different types dealing with family and responsibility,

1130
00:49:44.630 --> 00:49:46.250
things like that.

1131
00:49:46.250 --> 00:49:49.440
So it's just interesting because we can use this type

1132
00:49:49.440 --> 00:49:53.600
of information, not only to predict higher and lower levels

1133
00:49:53.600 --> 00:49:55.477
of excessive drinking in the county level,

1134
00:49:55.477 --> 00:49:58.070
but more importantly, it gives us the insight

1135
00:49:58.070 --> 00:50:01.340
into how we make able to develop messages

1136
00:50:01.340 --> 00:50:05.060
and develop interventions and treatment information.

1137
00:50:05.060 --> 00:50:08.150
So that was like the first level of that grant.

1138
00:50:08.150 --> 00:50:11.370
```

The second and the reason why I think NIDA gave

1139
00:50:11.370 --> 00:50:14.160
me the grant, the more important reason are not important,

1140
00:50:14.160 --> 00:50:16.930
but the reason that they were kind of concerned

1141
00:50:16.930 --> 00:50:21.380
with was we wanted to know from a clinical level

1142
00:50:21.380 --> 00:50:24.320
how we could use big data and machine learning

1143
00:50:24.320 --> 00:50:29.320
in order to figure out ways of predicting treatment outcomes

1144
00:50:29.560 --> 00:50:32.310
and excessive drinking, not excessive drinking,

1145
00:50:32.310 --> 00:50:35.760
but substance use and substance use outcomes.

1146
00:50:35.760 --> 00:50:39.230
So in here, we had a simple question,

1147
00:50:39.230 --> 00:50:43.290
can we use social media language,

1148
00:50:43.290 --> 00:50:45.350
we collect social media language literally

1149
00:50:45.350 --> 00:50:46.890
by a click of a button.

1150
00:50:46.890 --> 00:50:49.680
We wanted to know if we can predict treatment outcomes

```
1151
00:50:49.680 --> 00:50:51.200
of people who were attending

1152
00:50:51.200 --> 00:50:53.910
intensive outpatient drug treatment.

1153
00:50:53.910 --> 00:50:55.650
So these are people who are going

1154
00:50:55.650 --> 00:51:00.170
to a clinic based treatment center four times a week

1155
00:51:00.170 --> 00:51:04.960
at least, we collected this in Philadelphia,

1156
00:51:04.960 --> 00:51:08.270
the area where we're at where kind of near a lot

1157
00:51:08.270 --> 00:51:11.830
of the open area drug markets

1158
00:51:11.830 --> 00:51:14.700
that I believe have been close.

1159
00:51:14.700 --> 00:51:17.120
But so we were kind of in this community

1160
00:51:17.120 --> 00:51:19.803
where we had high poverty,

1161
00:51:21.160 --> 00:51:24.930
but yet we were able to kind of,

1162
00:51:24.930 --> 00:51:28.720
there was a lot of community groups

1163
00:51:28.720 --> 00:51:31.740
```

that were providing services to participants.

1164
00:51:31.740 --> 00:51:34.440
And so we were able to kind of partner with a lot

1165
00:51:34.440 --> 00:51:38.500
of them in order to come in and do this research study.

1166
00:51:38.500 --> 00:51:41.300
So in this study we had the question about,

1167
00:51:41.300 --> 00:51:43.550
can we use social media language

1168
00:51:43.550 --> 00:51:45.240
to predict treatment outcomes?

1169
00:51:45.240 --> 00:51:46.700
And so at the top, you'll see

1170
00:51:46.700 --> 00:51:49.070
this is kind of how the study went,

1171
00:51:49.070 --> 00:51:53.140
a person at baseline, which was at treatment intake.

1172
00:51:53.140 --> 00:51:55.350
So a treatment intake, a person they come

1173
00:51:55.350 --> 00:51:59.160
to the treatment center and they say, "I need treatment."

1174
00:51:59.160 --> 00:52:03.200
And while they're doing the intake and they're waiting,

1175
00:52:03.200 --> 00:52:07.403
we would come up to them and say, "Do you have social media?

1176
00:52:08.260 --> 00:52:09.307
do you have a social media account?"

1177
00:52:09.307 --> 00:52:11.250
Any social media in which we interested

1178
00:52:11.250 --> 00:52:13.070
in our research study.

1179
00:52:13.070 --> 00:52:14.940
So we were at the treatment sites

1180
00:52:14.940 --> 00:52:18.910
and I was doing while as at the University of Pennsylvania.

1181
00:52:18.910 --> 00:52:22.450
And so we would get their social media language.

1182
00:52:22.450 --> 00:52:25.990
We also had a trained staff who would do

1183
00:52:25.990 --> 00:52:29.900
something called the ASI, the Addiction Severity Index.

1184
00:52:29.900 --> 00:52:32.279
And this right here is like a gold standard

1185
00:52:32.279 --> 00:52:35.190
of assessing addiction severity.

1186
00:52:35.190 --> 00:52:37.270
So we would do like an hour long interview

1187
00:52:37.270 --> 00:52:39.220
for the addiction severity index

1188
00:52:39.220 --> 00:52:41.250

and kind of like that click of a button

1189
00:52:41.250 --> 00:52:44.050
for your social media link.

1190
00:52:44.050 --> 00:52:45.960
We collected the social media language

1191
00:52:45.960 --> 00:52:48.320
for two years before treatment.

1192
00:52:48.320 --> 00:52:50.260
And we're using the social media language

1193
00:52:50.260 --> 00:52:54.650
and the Addiction Severity Index to predict out 90 days,

1194
00:52:54.650 --> 00:52:57.100
we did some fancy stuff of initially predicting

1195
00:52:57.100 --> 00:52:57.960
kind of three outcomes,

1196
00:52:57.960 --> 00:53:00.070
but we settled with one, two outcomes,

1197
00:53:00.070 --> 00:53:02.910
either dropping out of treatment or staying in treatment.

1198
00:53:02.910 --> 00:53:05.720
And the reason why we kind of threw out relapse is

1199
00:53:05.720 --> 00:53:09.470
because if you can relapse, but stay in treatment,

1200
00:53:09.470 --> 00:53:10.880
when you are out of treatment,

1201
00:53:10.880 --> 00:53:12.980
you're gone and we can't deliver interventions

1202
00:53:12.980 --> 00:53:14.710
to you anymore or treatment to you.

1203
00:53:14.710 --> 00:53:16.370
So the key is to keep people in treatment

1204
00:53:16.370 --> 00:53:18.820
and we felt that that was the best indicator.

1205
00:53:18.820 --> 00:53:20.880
So what we wanna do is we wanna predict that.

1206
00:53:20.880 --> 00:53:24.650
and our AUCs, which are area under the curve

1207
00:53:24.650 --> 00:53:27.860
And to give you an idea, 50% would indicate chance

1208
00:53:27.860 --> 00:53:30.140
and 100% indicated perfection.

1209
00:53:30.140 --> 00:53:33.960
We were able to predict 90-day treatment outcomes

1210
00:53:33.960 --> 00:53:38.580
at about 79%, which is huge, very, very good.

1211
00:53:38.580 --> 00:53:40.610
Our social behavior work is

1212
00:53:40.610 --> 00:53:44.250
typically in the 68, 70ish, early '70s.

1213
00:53:44.250 --> 00:53:49.140

So getting to a 78.9 or 79% was huge for us,

1214
00:53:49.140 --> 00:53:50.843
and so this was a win.

1215
00:53:52.150 --> 00:53:56.080
So we can predict treatment outcomes, that's great, right?

1216
00:53:56.080 --> 00:54:00.560
But I need to figure out how can we apply this?

1217
00:54:00.560 --> 00:54:02.840
And so that was kind of one of our big questions is

1218
00:54:02.840 --> 00:54:04.150
how to apply this.

1219
00:54:04.150 --> 00:54:08.763
And what we decided was to do was to create risk scores.

1220
00:54:10.120 --> 00:54:13.770
So we did kind of, we simulated a clinical application.

1221
00:54:13.770 --> 00:54:17.460
So what we did was we took only social media language

1222
00:54:17.460 --> 00:54:19.460
we have at baseline.

1223
00:54:19.460 --> 00:54:23.090
And we took the ASI because that's traditionally done

1224
00:54:23.090 --> 00:54:25.980
at baseline, treatment centers are not gonna get rid

1225
00:54:25.980 --> 00:54:28.630
of the ASI, so we said, let's simulate this.

1226
00:54:28.630 --> 00:54:30.410
So we have demographic information

1227
00:54:30.410 --> 00:54:32.420
which you get from the intake form,

1228
00:54:32.420 --> 00:54:35.930
we have social media language at baseline,

1229
00:54:35.930 --> 00:54:37.600
and we have the ASI.

1230
00:54:37.600 --> 00:54:39.540
And what we did was we use that language

1231
00:54:39.540 --> 00:54:42.530
to put people into four risk categories,

1232
00:54:42.530 --> 00:54:46.230
and then predict out where we think they should be

1233
00:54:46.230 --> 00:54:48.880
at 90 days or they are at 90 days.

1234
00:54:48.880 --> 00:54:51.130
So we're able to label each participant

1235
00:54:51.130 --> 00:54:55.250
with a risk value of how from highest risk, lowest risk.

1236
00:54:55.250 --> 00:54:59.530
And as you can see, those first 30 days are crucial.

1237
00:54:59.530 --> 00:55:02.600
And while you still get dropout from 30 to 90,

1238
00:55:02.600 --> 00:55:04.630

the first 30 days are crucial.

1239
00:55:04.630 --> 00:55:08.110
And so we can use this risk score at baseline,

1240
00:55:08.110 --> 00:55:10.670
the day a person walk into the treatment center

1241
00:55:10.670 --> 00:55:14.160
to be like, "Hey, we really need to."

1242
00:55:14.160 --> 00:55:17.330
hopefully we're investing time and energy into everyone,

1243
00:55:17.330 --> 00:55:20.310
but this group that's in the high risk category

1244
00:55:20.310 --> 00:55:21.910
or high risk categories,

1245
00:55:21.910 --> 00:55:25.610
we should start doing some more intensive treatment,

1246
00:55:25.610 --> 00:55:29.870
some wraparound services, they're gonna need more support.

1247
00:55:29.870 --> 00:55:31.890
So in a situation, unfortunately,

1248
00:55:31.890 --> 00:55:34.270
where we have limited resources,

1249
00:55:34.270 --> 00:55:38.290
this allows us to triage those resources

1250
00:55:38.290 --> 00:55:42.240
to people at most in need and early on because remember,

1251
00:55:42.240 --> 00:55:44.850
when we lose them they are gone,

1252
00:55:44.850 --> 00:55:48.883
so if we can keep them in treatment longer, we can succeed.

1253
00:55:50.750 --> 00:55:54.540
So that was using social media language.

1254
00:55:54.540 --> 00:55:57.040
And we we've been analyzing the data

1255
00:55:57.040 --> 00:55:59.460
and working with it and figuring out ways

1256
00:55:59.460 --> 00:56:02.850
of we're moving to using, we also collected language

1257
00:56:02.850 --> 00:56:05.020
while in treatment, we're looking at that.

1258
00:56:05.020 --> 00:56:06.900
And we also are looking at what is

1259
00:56:06.900 --> 00:56:09.163
it specifically about the treatment.

1260
00:56:10.870 --> 00:56:13.300
Spoiler alert, it is looking very similar

1261
00:56:13.300 --> 00:56:15.990
to the county level excessive drinking

1262
00:56:15.990 --> 00:56:17.870
where we're finding language

1263
00:56:17.870 --> 00:56:21.720

about very religious language is showing,

1264
00:56:21.720 --> 00:56:24.282
indicating better outcomes and treatment

1265
00:56:24.282 --> 00:56:25.600
or at least in this treatment.

1266
00:56:25.600 --> 00:56:26.640
But I would like to point out

1267
00:56:26.640 --> 00:56:28.730
that those were community-based treatment centers

1268
00:56:28.730 --> 00:56:31.780
and a lot of them have kind of a 12-step orientation.

1269
00:56:31.780 --> 00:56:33.890
So maybe it's just treatment matching

1270
00:56:33.890 --> 00:56:35.620
because that's where people kind of,

1271
00:56:35.620 --> 00:56:37.170
those treatment centers talk a lot

1272
00:56:37.170 --> 00:56:40.030
about high risk, higher power.

1273
00:56:40.030 --> 00:56:44.020
So COVID hit, and we have a situation,

1274
00:56:44.020 --> 00:56:48.480
we're at NIDA clinic and Baltimore shut down,

1275
00:56:48.480 --> 00:56:52.130
and we decided my team and some collaborators

1276
00:56:52.130 --> 00:56:55.490
to make lemonade out of lemons and felt

1277
00:56:55.490 --> 00:57:00.450
that who else is best suited to do a national online study

1278
00:57:00.450 --> 00:57:02.270
than someone who uses social media

1279
00:57:02.270 --> 00:57:04.130
and all of these digital tools.

1280
00:57:04.130 --> 00:57:06.950
So we got into the lab, not physically,

1281
00:57:06.950 --> 00:57:10.023
but virtually and designed the study.

1282
00:57:11.592 --> 00:57:14.093
We're doing a national study at NIDA,

1283
00:57:14.093 --> 00:57:17.470
2,500 people with and without substance use disorders,

1284
00:57:17.470 --> 00:57:19.570
including alcohol use disorder

1285
00:57:19.570 --> 00:57:22.640
in their natural environments throughout the US.

1286
00:57:22.640 --> 00:57:24.460
We have already collected data,

1287
00:57:24.460 --> 00:57:27.140
we have already met our recruitment goals.

1288
00:57:27.140 --> 00:57:31.860

For baseline, we have minimum of 2,500 people in the study.

1289
00:57:31.860 --> 00:57:34.190
We come and collect their social media language.

1290
00:57:34.190 --> 00:57:39.190
We do tons of surveys, not tons, it's minimum.

1291
00:57:39.570 --> 00:57:44.570
We collect depression, anxiety, food insecurity,

1292
00:57:44.610 --> 00:57:49.330
economic indicator, loss of jobs, treatment indicators,

1293
00:57:49.330 --> 00:57:52.910
drug use market, COVID risk type scores.

1294
00:57:52.910 --> 00:57:56.370
We did kind of pack that 'cause we weren't sure

1295
00:57:56.370 --> 00:57:58.840
at the beginning of COVID what was important

1296
00:57:58.840 --> 00:58:00.040
and we had some hypothesis

1297
00:58:00.040 --> 00:58:02.460
about social isolation and loneliness

1298
00:58:02.460 --> 00:58:04.693
and access to harm reduction and treatment.

1299
00:58:05.940 --> 00:58:08.020
So we have that at baseline,

1300
00:58:08.020 --> 00:58:12.270
we do a short 30-day initial study

```
1301
00:58:12.270 --> 00:58:15.830
where we use an EMA, Ecological Momentary Assessment,

1302
00:58:15.830 --> 00:58:19.320
sending out a smartphone message, a little survey.

1303
00:58:19.320 --> 00:58:21.120
We do that for 30 days.

1304
00:58:21.120 --> 00:58:23.630
And then we follow them for like the next six months

1305
00:58:23.630 --> 00:58:26.800
where we do more surveys

1306
00:58:26.800 --> 00:58:28.420
and those are a little bit tailored,

1307
00:58:28.420 --> 00:58:32.600
kind of we're ending with vaccine hesitancy

1308
00:58:32.600 --> 00:58:34.510
and things like that.

1309
00:58:34.510 --> 00:58:37.573
So as COVID changed, we were able to change our studies.

1310
00:58:39.042 --> 00:58:39.875
But the key part,

1311
00:58:39.875 --> 00:58:41.590
one of the key things we're gonna show you is

1312
00:58:41.590 --> 00:58:45.710
that of 300 of those, we're doing a smartphone sensor study

1313
00:58:45.710 --> 00:58:48.530
```

```
where they put a sensor on their smartphone

1314
00:58:48.530 --> 00:58:49.890
and we're capturing data

1315
00:58:49.890 --> 00:58:54.560
from over 15 sensors on the smartphone,

1316
00:58:54.560 --> 00:58:57.500
including social media language, as well as keystrokes.

1317
00:58:57.500 --> 00:59:00.790
So things that we can gather from that smartphone,

1318
00:59:00.790 --> 00:59:03.900
the sensor data, when you're online

1319
00:59:03.900 --> 00:59:06.200
and you have your smartphone that you carry everywhere,

1320
00:59:06.200 --> 00:59:09.350
that you text and type and do almost everything with,

1321
00:59:09.350 --> 00:59:11.490
it captures a lot of information about you.

1322
00:59:11.490 --> 00:59:14.363
It probably know better than most people that you knows.

1323
00:59:15.230 --> 00:59:17.510
It knows, we get your accelerometer data,

1324
00:59:17.510 --> 00:59:19.820
we can do things like get your physical movement

1325
00:59:19.820 --> 00:59:21.620
and daily activities.
```

1326
00:59:21.620 --> 00:59:24.180
Bluetooth are kind of telling us about social interactions,

1327
00:59:24.180 --> 00:59:25.490
how close you are to people,

1328
00:59:25.490 --> 00:59:29.000
and how many people you come in contact with.

1329
00:59:29.000 --> 00:59:30.610
GPS, we know where you went,

1330
00:59:30.610 --> 00:59:32.560
we know how long you were there.

1331
00:59:32.560 --> 00:59:37.200
We know if you rode, drove, walked, or ran.

1332
00:59:37.200 --> 00:59:40.030
We have the light sensors, we have WiFi scans.

1333
00:59:40.030 --> 00:59:44.220
We have your calls and texts, smartphone text logs.

1334
00:59:44.220 --> 00:59:46.820
So we can look at things like how,

1335
00:59:46.820 --> 00:59:49.650
are we getting changes in how many people you talk to

1336
00:59:49.650 --> 00:59:50.483
or you texts?

1337
00:59:50.483 --> 00:59:53.620
And we can look at things like what were your mood

1338
00:59:53.620 --> 00:59:56.880

and behaviors before then, after, during,

1339
00:59:56.880 --> 00:59:58.030
are we getting changes?

1340
00:59:58.030 --> 01:00:00.060
We also know what apps you're using

1341
01:00:00.060 --> 01:00:02.600
and we know your smartphone, your language

1342
01:00:02.600 --> 01:00:07.210
and we can predict social interactions and daily activities.

1343
01:00:07.210 --> 01:00:09.970
So to give you an idea of something that we're doing

1344
01:00:09.970 --> 01:00:14.630
with the keystroke data, we can look at, for example,

1345
01:00:14.630 --> 01:00:16.740
we can do like a just account

1346
01:00:16.740 --> 01:00:20.280
of having a list of lexicon list of words.

1347
01:00:20.280 --> 01:00:22.610
We can look at how many times you're talking

1348
01:00:22.610 --> 01:00:25.140
about vaccines or COVID.

1349
01:00:25.140 --> 01:00:28.970
We can also look at people talking about mood and stress,

1350
01:00:28.970 --> 01:00:31.980
and you know how they're doing, how they're coping.

```
1351
01:00:31.980 --> 01:00:33.800
Are they socially isolated?

1352
01:00:33.800 --> 01:00:35.740
Are they getting treatment and things?

1353
01:00:35.740 --> 01:00:38.620
So that's the study that we have in the field right now

1354
01:00:38.620 --> 01:00:42.280
that I'm really proud of my research lab.

1355
01:00:42.280 --> 01:00:44.520
I wouldn't be able to do this without my team.

1356
01:00:44.520 --> 01:00:46.550
We have a great group of people who are

1357
01:00:46.550 --> 01:00:49.613
from computer science, psychology, public health,

1358
01:00:50.660 --> 01:00:54.790
and we all work together to answer some of these questions.

1359
01:00:54.790 --> 01:00:57.200
But one of the key things about my lab is

1360
01:01:00.658 --> 01:01:03.260
I do a lot of ethics work.

1361
01:01:03.260 --> 01:01:04.860
I'm trying to turn off my phone,

1362
01:01:06.200 --> 01:01:09.120
sorry, yes, I have a strange ring through.

1363
01:01:09.120 --> 01:01:10.060
```

```
We have a lot of people

1364
01:01:10.060 --> 01:01:13.470
from different diverse backgrounds that we ask questions

1365
01:01:13.470 --> 01:01:15.080
all the time, especially when we're talking

1366
01:01:15.080 --> 01:01:16.790
about like releasing of data.

1367
01:01:16.790 --> 01:01:18.140
What should be released?

1368
01:01:18.140 --> 01:01:19.630
How has she been released?

1369
01:01:19.630 --> 01:01:22.750
and what are our responsibilities?

1370
01:01:22.750 --> 01:01:24.440
I like to thank my collaborators,

1371
01:01:24.440 --> 01:01:28.810
have lots of collaborators here at NIH and outside.

1372
01:01:28.810 --> 01:01:30.870
And of course, I like to think Nora

1373
01:01:30.870 --> 01:01:35.840
and Amy who have helped me from when I was at Penn

1374
01:01:35.840 --> 01:01:39.193
to being here at NIDA IRP, so thank you.

1375
01:01:43.580 --> 01:01:45.160
<v ->Thank you, Brenda, that was great.</v>
```

```
1376
01:01:45.160 --> 01:01:47.260
It was wonderful to hear about what you're working on

1377
01:01:47.260 --> 01:01:48.093
and about your career

1378
01:01:48.093 --> 01:01:50.880
as well what motivated you in to this area.

1379
01:01:50.880 --> 01:01:54.800
So thanks again to both Brenda and Kristian.

1380
01:01:54.800 --> 01:01:57.790
Next, we're gonna have questions from the audience,

1381
01:01:57.790 --> 01:01:59.430
and I'm gonna turn it over to Dr. Lindsey

1382
01:01:59.430 --> 01:02:01.170
who's gonna moderate that session.

1383
01:02:01.170 --> 01:02:04.023
So please put your questions into the chat box.

1384
01:02:06.380 --> 01:02:07.213
<v ->Awesome, thanks.</v>

1385
01:02:07.213 --> 01:02:08.700
Yeah, there's a little question and answer box

1386
01:02:08.700 --> 01:02:12.340
at the bottom and list the name of the person if you want

1387
01:02:12.340 --> 01:02:14.360
one person specifically.

1388
01:02:14.360 --> 01:02:16.660
```

So first, one of the purposes

1389
01:02:16.660 --> 01:02:18.400
of the seminar series is to reach out

1390
01:02:18.400 --> 01:02:21.390
to younger who are thinking of a career in data science.

1391
01:02:21.390 --> 01:02:24.380
So I'll ask both of our speakers today,

1392
01:02:24.380 --> 01:02:27.030
what advice you have for younger scientists

1393
01:02:27.030 --> 01:02:30.010
that are thinking of starting a career in data science

1394
01:02:34.290 --> 01:02:35.510
Either of you can jump.

1395
01:02:35.510 --> 01:02:38.150
<v ->Okay, well, for me,</v>

1396
01:02:38.150 --> 01:02:40.350
I kind of came in through the back door.

1397
01:02:40.350 --> 01:02:43.030
I'm not a trained data scientist,

1398
01:02:43.030 --> 01:02:45.123
but I think just like public health,

1399
01:02:46.500 --> 01:02:48.780
I felt like public health is a very diverse group

1400
01:02:48.780 --> 01:02:50.560
of kind of who gets there.

```
1401
01:02:50.560 --> 01:02:53.273
And I work with data scientists to help me do,

1402
01:02:54.320 --> 01:02:56.120
using their tools and methodologies

1403
01:02:56.120 --> 01:02:58.783
in order to apply it in this domain.

1404
01:03:00.640 --> 01:03:03.723
I always spoke the language of math and science,

1405
01:03:06.989 --> 01:03:09.780
the community I came with that wasn't discouraged.

1406
01:03:09.780 --> 01:03:13.760
So I had lots of great teachers and support systems

1407
01:03:13.760 --> 01:03:15.900
and so they allowed me to be me.

1408
01:03:15.900 --> 01:03:18.670
And honestly, I felt like it was a language

1409
01:03:18.670 --> 01:03:20.943
that I could reference and understand.

1410
01:03:20.943 --> 01:03:24.093
And so I just kind of leaned into that.

1411
01:03:25.280 --> 01:03:27.580
And so that's what I would say is like don't put limits

1412
01:03:27.580 --> 01:03:31.133
on people and to just go forward in your interest.

1413
01:03:32.920 --> 01:03:35.760
```

```
<v ->Yeah, I think my advice would be</v>

1414
01:03:35.760 --> 01:03:38.450
to get some subject area expertise

1415
01:03:38.450 --> 01:03:40.550
in the areas that you want to work in as well.

1416
01:03:40.550 --> 01:03:45.550
I think the tools of data science are really powerful

1417
01:03:46.040 --> 01:03:50.070
and can be applied sort of blindly.

1418
01:03:50.070 --> 01:03:52.330
If you want to, you can just throw whatever model

1419
01:03:52.330 --> 01:03:56.810
you want at whatever data and get an answer.

1420
01:03:56.810 --> 01:03:59.700
But the answer is kind of meaningless if you don't

1421
01:03:59.700 --> 01:04:02.270
really have a good, deep understanding

1422
01:04:02.270 --> 01:04:04.930
of the subject area, the data comes from,

1423
01:04:04.930 --> 01:04:08.800
like why you would want to be applying that method or model,

1424
01:04:08.800 --> 01:04:12.250
like what are relevant questions in that area, et cetera.

1425
01:04:12.250 --> 01:04:14.820
And so I think even if you are like a very mathy person
```

```
1426
01:04:14.820 --> 01:04:17.670
like I am, and it seems like Dr. Curtis as well,

1427
01:04:17.670 --> 01:04:22.500
I think , it's always worth the time to also dive deep

1428
01:04:22.500 --> 01:04:26.053
on the non-mathematical aspects of the field.

1429
01:04:33.240 --> 01:04:34.690
<v ->Lindsey, we can't hear you.</v>

1430
01:04:35.749 --> 01:04:39.850
<v ->Sorry, so I'm gonna read one question from the attendees,</v>

1431
01:04:39.850 --> 01:04:41.377
thank you for your answers.

1432
01:04:41.377 --> 01:04:44.430
"Dr. Curtis, very interesting application of data science

1433
01:04:44.430 --> 01:04:46.750
to drug use in the phone center project.

1434
01:04:46.750 --> 01:04:47.963
How might predictive algorithms

1435
01:04:47.963 --> 01:04:51.530
on the phone sensors be used in a real world setting?

1436
01:04:51.530 --> 01:04:54.470
And how do you see this being the ultimate end user?

1437
01:04:54.470 --> 01:04:57.620
For example, clinicians, hospitals, healthcare,

1438
01:04:57.620 --> 01:04:59.270
```

healthcare providers, et cetera."

1439
01:05:00.580 --> 01:05:02.640
<v ->My approach because probably I started</v>

1440
01:05:02.640 --> 01:05:06.560
in treatment is how we can help treatment

1441
01:05:06.560 --> 01:05:10.330
and also strongly support peer support and family support.

1442
01:05:10.330 --> 01:05:13.780
And so I'm looking at coming from the lens

1443
01:05:13.780 --> 01:05:15.900
of not necessarily providing intervention

1444
01:05:15.900 --> 01:05:18.440
to the person per se the smartphone,

1445
01:05:18.440 --> 01:05:19.850
but how we can use that information

1446
01:05:19.850 --> 01:05:22.090
that can inform clinical treatment.

1447
01:05:22.090 --> 01:05:25.310
And so, for example, things I've heard from families is

1448
01:05:25.310 --> 01:05:27.170
like, they just want to have an indicator

1449
01:05:27.170 --> 01:05:29.993
that their family member is doing okay.

1450
01:05:30.830 --> 01:05:32.620
Are they similar to they were yesterday

```
1451
01:05:32.620 --> 01:05:35.130
or two the last week, has there been any changes?

1452
01:05:35.130 --> 01:05:37.560
So I can see an application there that can kind

1453
01:05:37.560 --> 01:05:40.400
of be this early detection, early warning tool,

1454
01:05:40.400 --> 01:05:44.920
which can also be applied to treatment centers.

1455
01:05:44.920 --> 01:05:48.320
Case managers and counselors have a huge number

1456
01:05:48.320 --> 01:05:51.960
of participants, patients per clinician.

1457
01:05:51.960 --> 01:05:54.750
And so if we can provide them with just like a

1458
01:05:54.750 --> 01:05:56.810
daily dashboard or an idea

1459
01:05:56.810 --> 01:05:58.930
of how their patients are doing

1460
01:05:58.930 --> 01:06:02.450
and who they may wanna check in with earlier,

1461
01:06:02.450 --> 01:06:04.190
I think that that would be great.

1462
01:06:04.190 --> 01:06:07.090
So I'm using these tools for,

1463
01:06:07.090 --> 01:06:09.060
```

we don't know if there's a certain level

1464
01:06:09.060 --> 01:06:10.880
that indicate risk, right?

1465
01:06:10.880 --> 01:06:12.630
Or is it a change of behavior?

1466
01:06:12.630 --> 01:06:14.977
Is it that a person comes from treatment,

1467
01:06:14.977 --> 01:06:17.260
you know from their treatment records

1468
01:06:17.260 --> 01:06:18.410
and you've been talking to them.

1469
01:06:18.410 --> 01:06:20.910
If these three weeks they've been doing great

1470
01:06:20.910 --> 01:06:22.300
and now. we see a change,

1471
01:06:22.300 --> 01:06:25.130
we don't know if that changes a good change or a bad change,

1472
01:06:25.130 --> 01:06:26.850
but that would allow the counselor to then

1473
01:06:26.850 --> 01:06:29.370
or the treatment provider or peer support to check

1474
01:06:29.370 --> 01:06:31.960
in with that person to see what's going on

1475
01:06:31.960 --> 01:06:35.480
and then to start initiating if higher level

```
1476
01:06:35.480 --> 01:06:36.950
of intervention is needed.

1477
01:06:36.950 --> 01:06:39.510
We can also use this with medication adherence

1478
01:06:39.510 --> 01:06:41.830
and monitoring to see how people are doing

1479
01:06:41.830 --> 01:06:43.703
on medication assisted treatments.

1480
01:06:45.410 --> 01:06:47.860
<v ->Awesome, thank you for that answer, very great.</v>

1481
01:06:49.230 --> 01:06:51.373
Another question that we had was regarding COVID,

1482
01:06:51.373 --> 01:06:52.970
I mean, both of you have mentioned

1483
01:06:52.970 --> 01:06:54.650
how your projects have changed

1484
01:06:54.650 --> 01:06:57.050
because of the pandemic, but what kind

1485
01:06:57.050 --> 01:06:58.530
of advice or outreach would you have

1486
01:06:58.530 --> 01:07:02.730
for younger scientists and this kind of pandemic field

1487
01:07:03.580 --> 01:07:04.530
that they find themselves

1488
01:07:04.530 --> 01:07:06.683
```

```
kind of growing up in education wise?

1489
01:07:09.610 --> 01:07:13.330
<v ->I'll go first, I have a lot of like small things,</v>

1490
01:07:13.330 --> 01:07:16.220
I think one of them, and I have sort of fallen victim

1491
01:07:16.220 --> 01:07:18.010
to not using this advice is

1492
01:07:18.010 --> 01:07:20.480
that you don't have to chase every,

1493
01:07:20.480 --> 01:07:22.120
I don't wanna call it COVID like a news cycle,

1494
01:07:22.120 --> 01:07:24.490
but just because something's the hot area,

1495
01:07:24.490 --> 01:07:25.400
you can still keep working

1496
01:07:25.400 --> 01:07:26.510
on the thing that you're working on.

1497
01:07:26.510 --> 01:07:30.390
You don't have to necessarily pivot to COVID.

1498
01:07:30.390 --> 01:07:32.350
Yes, it's an important topic, and yes,

1499
01:07:32.350 --> 01:07:34.120
we obviously need a lot of research on it,

1500
01:07:34.120 --> 01:07:35.750
but also we still need to keep doing research
```

```
1501
01:07:35.750 --> 01:07:36.730
on other things as well.

1502
01:07:36.730 --> 01:07:39.330
And I think for a lot of people,

1503
01:07:39.330 --> 01:07:42.367
myself obviously included, it's tempting to be like,

1504
01:07:42.367 --> 01:07:43.670
"Okay, this is the thing right now."

1505
01:07:43.670 --> 01:07:46.993
So like I should pivot all my resources towards that,

1506
01:07:48.280 --> 01:07:52.000
but other research is valuable as well.

1507
01:07:52.000 --> 01:07:54.360
I also think one of the really great things

1508
01:07:54.360 --> 01:07:59.360
about the sort of online or like virtual type

1509
01:07:59.780 --> 01:08:02.240
of research that's been happening since COVID,

1510
01:08:02.240 --> 01:08:03.540
I don't want to say great thing about COVID

1511
01:08:03.540 --> 01:08:05.770
'cause I think there's nothing great about that,

1512
01:08:05.770 --> 01:08:08.650
but this sort of virtual world is that I can be here

1513
01:08:08.650 --> 01:08:12.700
```

with you all, people can tune in from all over the place.

1514
01:08:12.700 --> 01:08:15.260
Recently, I was involved in organizing a conference

1515
01:08:15.260 --> 01:08:16.520
that was fully remote.

1516
01:08:16.520 --> 01:08:18.300
It was much cheaper than normal,

1517
01:08:18.300 --> 01:08:20.630
people from all over the world could attend.

1518
01:08:20.630 --> 01:08:23.900
We actually do try to make it accessible even in-person,

1519
01:08:23.900 --> 01:08:25.670
but it was much more accessible this year.

1520
01:08:25.670 --> 01:08:27.690
And so I think sort of taking advantage

1521
01:08:27.690 --> 01:08:29.850
of this situation where everything is online,

1522
01:08:29.850 --> 01:08:31.710
there's so many resources,

1523
01:08:31.710 --> 01:08:34.970
you have access to people who maybe normally,

1524
01:08:34.970 --> 01:08:37.453
it would be hard to talk to.

1525
01:08:38.480 --> 01:08:40.390
There's a lot less overhead and a lot less cost,

```
1526
01:08:40.390 --> 01:08:43.080
I think too for people to just do a quick Zoom

1527
01:08:43.080 --> 01:08:45.010
or tune into something like this.

1528
01:08:45.010 --> 01:08:47.580
And so that's been one of the positives

1529
01:08:47.580 --> 01:08:51.060
of this increasingly virtual research world

1530
01:08:51.060 --> 01:08:52.110
that we're living in.

1531
01:08:53.030 --> 01:08:56.220
<v ->For us, it was a kind of an interesting point</v>

1532
01:08:56.220 --> 01:08:59.240
because we were like, as our center was closing down,

1533
01:08:59.240 --> 01:09:02.050
we realized that a lot of treatment centers

1534
01:09:02.050 --> 01:09:04.620
across the US where people remember they go four

1535
01:09:04.620 --> 01:09:06.630
or five days a week was closing down.

1536
01:09:06.630 --> 01:09:09.540
And so we were really afraid of that,

1537
01:09:09.540 --> 01:09:13.330
we also know that during the the opiate epidemic

1538
01:09:13.330 --> 01:09:16.050
```

that our fear was that places

1539
01:09:16.050 --> 01:09:18.974
that were providing harm reduction like Naloxone

1540
01:09:18.974 --> 01:09:22.270
and other things were going to disappear.

1541
01:09:22.270 --> 01:09:25.820
And so we were going to have higher levels of overdoses.

1542
01:09:25.820 --> 01:09:29.550
So we were afraid of that because harm reduction,

1543
01:09:29.550 --> 01:09:32.330
treatment sites closing their doors

1544
01:09:32.330 --> 01:09:34.040
and having social distance,

1545
01:09:34.040 --> 01:09:37.260
we were also afraid because substance use is

1546
01:09:37.260 --> 01:09:41.890
already a social isolating type of situation,

1547
01:09:41.890 --> 01:09:44.760
and people who are in treatment are highly stigmatized

1548
01:09:44.760 --> 01:09:46.390
and already kind of like an outcast

1549
01:09:46.390 --> 01:09:48.030
unfortunately many times.

1550
01:09:48.030 --> 01:09:49.010
And so we were afraid

```
1551
01:09:49.010 --> 01:09:51.110
that that isolation was gonna continue.

1552
01:09:51.110 --> 01:09:53.710
And even when they come into treatment centers,

1553
01:09:53.710 --> 01:09:55.860
if they have COVID that they were going

1554
01:09:55.860 --> 01:09:57.420
to get discriminated against.

1555
01:09:57.420 --> 01:09:59.840
And so we had these kinds of questions that were happening

1556
01:09:59.840 --> 01:10:04.050
to this population and that we were concerned about.

1557
01:10:04.050 --> 01:10:07.990
And so I was like on the phone, and texting, and typing

1558
01:10:07.990 --> 01:10:12.047
with different people, my collaborator and saying like,

1559
01:10:12.047 --> 01:10:13.847
"This is a problem, we should do it,

1560
01:10:14.740 --> 01:10:16.430
this could be horrible."

1561
01:10:16.430 --> 01:10:18.730
And more importantly, we were really afraid

1562
01:10:18.730 --> 01:10:21.000
that this population wasn't gonna get heard,

1563
01:10:21.000 --> 01:10:22.340
```

that this was gonna happen

1564
01:10:22.340 --> 01:10:23.900
and we weren't gonna know about it,

1565
01:10:23.900 --> 01:10:26.340
the impact of COVID to this population.

1566
01:10:26.340 --> 01:10:29.740
So that was that, and then at the same time,

1567
01:10:29.740 --> 01:10:34.170
I'm doing this, Nora puts out this statement article

1568
01:10:34.170 --> 01:10:35.417
about the harms and I was like,

1569
01:10:35.417 --> 01:10:36.640
"Oh, my God," now I was teasing,

1570
01:10:36.640 --> 01:10:38.530
I was like, "Does she have access to my share drive?"

1571
01:10:38.530 --> 01:10:39.730
No, I know she doesn't,

1572
01:10:39.730 --> 01:10:42.020
but it was like, this timing was perfect

1573
01:10:42.020 --> 01:10:44.280
and it gave me kind of this,

1574
01:10:44.280 --> 01:10:48.170
I felt like, okay, I'm one track and I can push through.

1575
01:10:48.170 --> 01:10:50.280
Getting something like this type of study done

1576
01:10:50.280 --> 01:10:53.190
at NIDA, inside the government was kind of hard.

1577
01:10:53.190 --> 01:10:55.520
I'm talking about a lot of stuff

1578
01:10:55.520 --> 01:10:58.340
that normally governments don't collect.

1579
01:10:58.340 --> 01:11:00.500
So we had to partner with Penn

1580
01:11:00.500 --> 01:11:02.000
and we're able to kind of get this done.

1581
01:11:02.000 --> 01:11:05.530
Everyone was supportive, Nora Collins, everyone.

1582
01:11:05.530 --> 01:11:08.613
And so we got it and Amy and we got it done.

1583
01:11:09.850 --> 01:11:12.380
So COVID in that sense really helped me push

1584
01:11:12.380 --> 01:11:15.130
my research more national.

1585
01:11:15.130 --> 01:11:18.460
And also, I think hopefully, give voice to a population

1586
01:11:18.460 --> 01:11:22.260
that were not going to get heard,

1587
01:11:22.260 --> 01:11:24.378
and I was really, we were really kind of scared

1588
01:11:24.378 --> 01:11:26.003

of what was gonna be going on.

1589
01:11:27.560 --> 01:11:29.000
<v ->Awesome, thanks, it's good that you're able</v>

1590
01:11:29.000 --> 01:11:31.853
to find some positives at a time of so many negatives.

1591
01:11:32.910 --> 01:11:35.370
I know Wilson has a question, so I'll turn it over to Wilson

1592
01:11:35.370 --> 01:11:37.653
to ask his question to the panelists.

1593
01:11:38.960 --> 01:11:42.680
<v ->Thanks very much, this has been a wonderful session today,</v>

1594
01:11:42.680 --> 01:11:46.760
and I really appreciate both the career path

1595
01:11:46.760 --> 01:11:49.440
that Dr. Lum you described and Dr. Curtis,

1596
01:11:49.440 --> 01:11:53.030
your wonderful research, as well as some new information

1597
01:11:53.030 --> 01:11:54.620
for me about your career path,

1598
01:11:54.620 --> 01:11:56.190
how you started and how you got launched,

1599
01:11:56.190 --> 01:11:58.870
it was really important.

1600
01:11:58.870 --> 01:12:00.950
I have a couple of questions,

```
1601
01:12:00.950 --> 01:12:04.160
First, I think applies mostly to you, Dr. Lum,

1602
01:12:04.160 --> 01:12:08.000
look at a question about you've moved

1603
01:12:08.000 --> 01:12:12.220
between academic and NGO and start up

1604
01:12:12.220 --> 01:12:17.210
and shifting back to academic is not

1605
01:12:17.210 --> 01:12:19.540
always so easy after you been out

1606
01:12:19.540 --> 01:12:20.840
of an academic environment,

1607
01:12:20.840 --> 01:12:22.450
what has it been like coming back

1608
01:12:22.450 --> 01:12:26.300
and any advice about how to make those shifts?

1609
01:12:26.300 --> 01:12:29.520
<v ->Yeah, I mean, I think what you said is totally true.</v>

1610
01:12:29.520 --> 01:12:32.050
It's not always, I guess I'd say

1611
01:12:32.050 --> 01:12:33.730
something sort of that it's not ever easy

1612
01:12:33.730 --> 01:12:38.463
to completely shift sectors like that.

1613
01:12:39.350 --> 01:12:42.080
```

```
How it's been for me, it's been okay.

1614
01:12:42.080 --> 01:12:46.840
I think for the most part continued doing the same type

1615
01:12:46.840 --> 01:12:48.630
of research that I was doing before.

1616
01:12:48.630 --> 01:12:51.240
I feel like some of the differences are

1617
01:12:51.240 --> 01:12:53.530
when I was at the NGO, we were doing research.

1618
01:12:53.530 --> 01:12:56.700
And so I think working in a career

1619
01:12:56.700 --> 01:12:58.710
where you're still doing research

1620
01:12:58.710 --> 01:13:00.760
even if your affiliation is an academic,

1621
01:13:00.760 --> 01:13:04.020
that does leave doors open to go back to academia

1622
01:13:04.020 --> 01:13:05.490
'cause you're still publishing,

1623
01:13:05.490 --> 01:13:07.360
you're sort of still sort of like keeping a presence

1624
01:13:07.360 --> 01:13:08.370
in the research world.

1625
01:13:08.370 --> 01:13:12.023
So I think if somebody wanted to keep those doors open,
```

1626
01:13:13.970 --> 01:13:17.380
picking a career path where there are still opportunities

1627
01:13:17.380 --> 01:13:19.960
to continue publishing would would sort of help keep

1628
01:13:19.960 --> 01:13:20.793
those doors open.

1629
01:13:20.793 --> 01:13:21.870
I think in terms of differences,

1630
01:13:21.870 --> 01:13:26.280
one thing I really appreciated about working in research,

1631
01:13:26.280 --> 01:13:29.980
but in an NGO was we measured our impact,

1632
01:13:29.980 --> 01:13:32.300
I think a little bit differently than people in academia do.

1633
01:13:32.300 --> 01:13:36.050
So it wasn't, I don't wanna say this is true in academia,

1634
01:13:36.050 --> 01:13:37.130
but I think a caricature is

1635
01:13:37.130 --> 01:13:40.060
that it's purely about citation counts

1636
01:13:40.060 --> 01:13:43.020
and impact factors and all of those things

1637
01:13:43.020 --> 01:13:47.340
whereas when I was at the NGO, it was, yeah,

1638
01:13:47.340 --> 01:13:49.110

we were publishing 'cause we were researchers,

1639
01:13:49.110 --> 01:13:51.993
but also part of how we measured our impact was

1640
01:13:51.993 --> 01:13:54.800
were we helping advocacy groups understand the issues,

1641
01:13:54.800 --> 01:13:56.150
were we talking to journalists

1642
01:13:56.150 --> 01:13:58.760
and helping them understand the important issues

1643
01:13:58.760 --> 01:14:03.760
to be covering, were we working with other say NGOs

1644
01:14:05.690 --> 01:14:08.450
to help them say, scope their research

1645
01:14:08.450 --> 01:14:10.120
in say the AI space, right?

1646
01:14:10.120 --> 01:14:14.870
So I think we had a sort of more flexible definition

1647
01:14:14.870 --> 01:14:18.750
of success and that's something that I really appreciated.

1648
01:14:18.750 --> 01:14:21.500
So coming back to academia has been a little bit interesting

1649
01:14:21.500 --> 01:14:25.400
in the sense that I feel like I need to shift

1650
01:14:25.400 --> 01:14:26.690
my attention more to publishing

```
1651
01:14:26.690 --> 01:14:28.820
rather than these other aspects

1652
01:14:28.820 --> 01:14:29.900
that I've focused on for so long.

1653
01:14:29.900 --> 01:14:30.733
And it can be difficult,

1654
01:14:30.733 --> 01:14:33.450
I think to reramp up on the like publish, publish,

1655
01:14:33.450 --> 01:14:37.530
publish side of things, but it's been enjoyable.

1656
01:14:37.530 --> 01:14:39.680
And I like my current job, that sounds negative,

1657
01:14:39.680 --> 01:14:41.320
I don't mean it to be negative about academia,

1658
01:14:41.320 --> 01:14:43.890
but I think one of the difficulties is

1659
01:14:43.890 --> 01:14:46.550
kind of reworking how you think

1660
01:14:46.550 --> 01:14:48.640
about your work and sort of revamping

1661
01:14:48.640 --> 01:14:51.730
where you put your priorities to sort of set

1662
01:14:51.730 --> 01:14:53.810
yourself up for success in areas

1663
01:14:53.810 --> 01:14:55.313
```

```
where the incentives are different.

1664
01:14:56.370 --> 01:14:58.060
<v ->Certainly strikes me that both you</v>

1665
01:14:58.060 --> 01:15:00.330
and Dr. Curtis really approach your work

1666
01:15:00.330 --> 01:15:04.963
from a passion for improving lives of groups

1667
01:15:05.830 --> 01:15:09.290
that have been traditionally ignored by so much of society.

1668
01:15:09.290 --> 01:15:12.780
And I'm curious about how you balance

1669
01:15:12.780 --> 01:15:17.500
sort of the academic nature of your work with this,

1670
01:15:17.500 --> 01:15:21.470
the real world, changing the world that we live in

1671
01:15:21.470 --> 01:15:22.980
and making lives better,

1672
01:15:22.980 --> 01:15:25.150
which clearly motivates both of you very much,

1673
01:15:25.150 --> 01:15:26.750
it's just wonderful to see.

1674
01:15:26.750 --> 01:15:28.100
But how do you balance some of those

1675
01:15:28.100 --> 01:15:30.230
'cause because there's a need for the rigorous
```

```
1676
01:15:30.230 --> 01:15:32.280
in terms of the math and the science

1677
01:15:32.280 --> 01:15:35.763
as well as a focus on the long term goals?

1678
01:15:37.530 --> 01:15:39.230
<v ->Well, I actually think one of the things</v>

1679
01:15:39.230 --> 01:15:43.400
that Kristian talked about, which is very key is

1680
01:15:43.400 --> 01:15:47.650
that while we are doing our research and we are publishing

1681
01:15:47.650 --> 01:15:51.540
and getting information out to an academic institution,

1682
01:15:51.540 --> 01:15:55.310
it is if not more important to make sure

1683
01:15:55.310 --> 01:15:58.360
that that information is getting to people on the ground,

1684
01:15:58.360 --> 01:16:00.690
that's going to be using it.

1685
01:16:00.690 --> 01:16:03.380
So when, for example, I published an article,

1686
01:16:03.380 --> 01:16:05.890
we may make an infographic or something that we can put

1687
01:16:05.890 --> 01:16:09.237
on social media that can boil down this information

1688
01:16:09.237 --> 01:16:13.133
```

and make it accessible to lots of different people.

1689
01:16:15.550 --> 01:16:17.040
So that's one of the things I do

1690
01:16:17.040 --> 01:16:19.430
as well as the organizations and groups

1691
01:16:19.430 --> 01:16:22.720
that I've participated in the type of community outreach

1692
01:16:22.720 --> 01:16:25.840
that I give and making myself accessible

1693
01:16:25.840 --> 01:16:30.840
to high school students and students all along the pipeline

1694
01:16:31.730 --> 01:16:35.630
and my trainees in order to kind of, to do this.

1695
01:16:35.630 --> 01:16:39.770
So I think that we have an obligation and well,

1696
01:16:39.770 --> 01:16:41.030
I can't, at this point,

1697
01:16:41.030 --> 01:16:43.900
I used to be with the Public Policy Center at Annenberg,

1698
01:16:43.900 --> 01:16:46.300
but while at this point I don't do policy work,

1699
01:16:46.300 --> 01:16:49.290
I can make sure that my publications

1700
01:16:49.290 --> 01:16:52.500
as well as information that I put out could easily be used

```
1701
01:16:52.500 --> 01:16:54.900
and accessed by policy makers

1702
01:16:54.900 --> 01:16:57.290
so that they can have empirical evidence

1703
01:16:57.290 --> 01:16:59.113
in order to support decisions.

1704
01:17:01.780 --> 01:17:03.580
<v ->Thank you, I think that's a wonderful idea.</v>

1705
01:17:03.580 --> 01:17:06.800
And one I would personally strive for and encourage people

1706
01:17:06.800 --> 01:17:11.000
to emulate to make sure that work has multiple levels

1707
01:17:11.000 --> 01:17:13.463
that it can be used.

1708
01:17:15.970 --> 01:17:17.720
<v ->Awesome, thank you.</v>

1709
01:17:17.720 --> 01:17:20.490
I wanted to ask about one phrase I heard both of you use

1710
01:17:20.490 --> 01:17:23.500
and that was the lens that I look at the data through.

1711
01:17:23.500 --> 01:17:25.430
You both kind of mentioned like your individual lens,

1712
01:17:25.430 --> 01:17:26.670
so I was wondering if you could like speak

1713
01:17:26.670 --> 01:17:29.370
```

```
to that a little bit more, kind of your unique perspective

1714
01:17:29.370 --> 01:17:30.940
on your tackling these problems

1715
01:17:30.940 --> 01:17:33.163
and how that maybe has evolved over time.

1716
01:17:34.640 --> 01:17:36.513
<v ->So I guess, I can start.</v>

1717
01:17:40.483 --> 01:17:42.080
I think the the way I think about data

1718
01:17:42.080 --> 01:17:44.820
or the lens through which I look at data is very much shaped

1719
01:17:44.820 --> 01:17:48.405
by my time working at the NGO at HRDAG

1720
01:17:48.405 --> 01:17:52.280
where we really were sort of trying to think

1721
01:17:52.280 --> 01:17:53.590
of data from a different perspective.

1722
01:17:53.590 --> 01:17:55.770
I think often that lens is shaped

1723
01:17:55.770 --> 01:17:58.270
by one of the other things I was talking about earlier,

1724
01:17:58.270 --> 01:18:01.180
sort of diving deep into a substantive area,

1725
01:18:01.180 --> 01:18:04.700
trying to understand what the issues are there,
```

1726
01:18:04.700 --> 01:18:06.930
thinking about how the data was generated

1727
01:18:07.770 --> 01:18:11.433
and taking seriously concerns from people,

1728
01:18:11.433 --> 01:18:13.470
like when we're talking about, for example,

1729
01:18:13.470 --> 01:18:15.930
data that's going to be used to make recommendations

1730
01:18:15.930 --> 01:18:19.450
about real humans, taking seriously the current concerns

1731
01:18:19.450 --> 01:18:22.240
of the people about whom those recommendations will be made.

1732
01:18:22.240 --> 01:18:25.120
Right, so looking at the data, trying at least look

1733
01:18:25.120 --> 01:18:27.160
at the data from their perspective, often I'm not the person

1734
01:18:27.160 --> 01:18:30.580
at the sharp end of these models, in fact, almost never.

1735
01:18:30.580 --> 01:18:32.590
And so I can't say that I can fully adopt

1736
01:18:32.590 --> 01:18:34.650
someone's perspective, I don't know what it's like,

1737
01:18:34.650 --> 01:18:39.650
for example, to be a person evaluated for recidivism risk,

1738
01:18:40.000 --> 01:18:41.450

```
that's just not my experience,

1739
01:18:41.450 --> 01:18:43.660
but listening is as well as you can

1740
01:18:43.660 --> 01:18:45.620
and taking seriously the concerns

1741
01:18:45.620 --> 01:18:46.870
and incorporating those concerns

1742
01:18:46.870 --> 01:18:49.603
or that perspective into the data analysis.

1743
01:18:50.780 --> 01:18:53.473
<v ->For me, my lens has been coming</v>

1744
01:18:57.567 --> 01:18:59.880
from two lenses, maybe three,

1745
01:18:59.880 --> 01:19:02.240
one lens is coming from the communities

1746
01:19:02.240 --> 01:19:04.530
that are highly stigmatized

1747
01:19:04.530 --> 01:19:07.790
and highly stereotyped.

1748
01:19:07.790 --> 01:19:12.680
And knowing that I hear constantly those types

1749
01:19:12.680 --> 01:19:15.800
of stereotypes being, and data coming out

1750
01:19:17.306 --> 01:19:18.500
of what I should be
```

```
1751
01:19:18.500 --> 01:19:22.003
or if we go by the averages, what I would look like.

1752
01:19:23.140 --> 01:19:27.573
Also having a lens of being educated from an Ivy league,

1753
01:19:28.477 --> 01:19:31.180
what that means as well as you're not a researcher.

1754
01:19:31.180 --> 01:19:35.270
So trying to look at my data from various perspectives

1755
01:19:35.270 --> 01:19:36.451
as well as the gender,

1756
01:19:36.451 --> 01:19:40.023
the gender is an interesting lens to use.

1757
01:19:41.644 --> 01:19:42.530
And having, making sure

1758
01:19:42.530 --> 01:19:44.750
that that's why I like to have a diverse lab rule,

1759
01:19:44.750 --> 01:19:47.280
all different areas of peoples so that we can try

1760
01:19:47.280 --> 01:19:48.670
to get as many voices.

1761
01:19:48.670 --> 01:19:53.180
But more importantly, going back to participants.

1762
01:19:53.180 --> 01:19:56.180
I do a lot of ethics works where I ask participants

1763
01:19:56.180 --> 01:19:59.430
```

```
about the data, about the information we're collecting,

1764
01:19:59.430 --> 01:20:00.780
features they would like to have

1765
01:20:00.780 --> 01:20:02.980
to ensure privacy, confidentiality,

1766
01:20:02.980 --> 01:20:05.550
who they would like to have access to the data.

1767
01:20:05.550 --> 01:20:08.130
What types of apps are features

1768
01:20:08.130 --> 01:20:09.930
they would like built in of who,

1769
01:20:09.930 --> 01:20:13.060
when to turn on access and turn off access.

1770
01:20:13.060 --> 01:20:14.800
So I think it's key

1771
01:20:14.800 --> 01:20:17.700
to ask people who you are collecting data

1772
01:20:17.700 --> 01:20:20.890
from what they feel about it and what they want used,

1773
01:20:20.890 --> 01:20:24.580
and then to incorporate that into your research.

1774
01:20:24.580 --> 01:20:29.370
But also to let IRBs and regulators know this information

1775
01:20:29.370 --> 01:20:30.540
'cause they could be making demands
```

```
1776
01:20:30.540 --> 01:20:32.140
and you could say, no, no, no, my population,

1777
01:20:32.140 --> 01:20:34.540
they literally said they don't want that.

1778
01:20:34.540 --> 01:20:36.170
And so applying it that way.

1779
01:20:36.170 --> 01:20:39.890
I think it's key to include our participants

1780
01:20:39.890 --> 01:20:44.000
and people in our community and community advisory boards

1781
01:20:44.000 --> 01:20:46.470
in our research so that they can have a voice

1782
01:20:46.470 --> 01:20:47.303
and not necessarily like,

1783
01:20:47.303 --> 01:20:48.820
"We're just gonna give you a voice,"

1784
01:20:48.820 --> 01:20:51.830
but they need to have a real legitimate voice

1785
01:20:51.830 --> 01:20:55.270
and stake into our research and into what we're developing

1786
01:20:55.270 --> 01:20:56.570
and how we disseminate it.

1787
01:20:58.380 --> 01:20:59.930
<v ->Awesome, thank you very much.</v>

1788
01:21:00.870 --> 01:21:03.057
```

```
Another question was for you, Dr. Curtis specifically,

1789
01:21:03.057 --> 01:21:04.570
"Are you active on Twitter?

1790
01:21:04.570 --> 01:21:06.220
Does your lab have a Twitter presence?

1791
01:21:06.220 --> 01:21:10.287
How does that social media outreach come from you, guys?"

1792
01:21:11.920 --> 01:21:16.653
<v ->My lab currently here, we are not active on Twitter.</v>

1793
01:21:19.130 --> 01:21:22.067
Being part of NIH and NIDA, I just came in 2019,

1794
01:21:22.067 --> 01:21:24.620
and I'm still trying to figure out the whole communication

1795
01:21:24.620 --> 01:21:26.690
of what we can distribute versus not.

1796
01:21:26.690 --> 01:21:30.710
So what I typically do is we make things,

1797
01:21:30.710 --> 01:21:34.147
and then we will let our partners and our collaborators,

1798
01:21:34.147 --> 01:21:36.960
and a lot of times they put stuff out.

1799
01:21:36.960 --> 01:21:39.390
So we have not done it.

1800
01:21:39.390 --> 01:21:41.030
I don't even know if we're allowed to have
```

```
1801
01:21:41.030 --> 01:21:43.390
our own lab Twitter page.

1802
01:21:43.390 --> 01:21:45.068
I actually think the answer is no.

1803
01:21:45.068 --> 01:21:45.901
(chuckle loudly)

1804
01:21:45.901 --> 01:21:47.170
I'm not sure Wilson will say,

1805
01:21:47.170 --> 01:21:48.461
yeah, the answer is no.

1806
01:21:48.461 --> 01:21:50.670
I will let you know though, but yeah,

1807
01:21:50.670 --> 01:21:52.490
I haven't read their bioethics.

1808
01:21:52.490 --> 01:21:54.500
<v ->I'm not totally sure, I'll find out.</v>

1809
01:21:54.500 --> 01:21:55.452
It's a great question.

1810
01:21:55.452 --> 01:21:56.285
<v ->I think the answer is...</v>

1811
01:21:56.285 --> 01:21:58.470
<v ->Intramural is a little different than extramural,</v>

1812
01:21:58.470 --> 01:21:59.940
but we'll find out.

1813
01:21:59.940 --> 01:22:01.550
```

```
<v Brenda>Yeah.</v>

1814
01:22:01.550 --> 01:22:04.103
<v ->Great, thanks, I think Albert has a question.</v>

1815
01:22:05.977 --> 01:22:06.927
Yup, there you are.

1816
01:22:07.920 --> 01:22:12.920
<v ->Thanks to both of you guys for your talk</v>

1817
01:22:13.710 --> 01:22:15.290
in your career journey.

1818
01:22:15.290 --> 01:22:18.300
One of the things that I would like you, guys,

1819
01:22:18.300 --> 01:22:21.270
to both address so that we could share

1820
01:22:21.270 --> 01:22:25.660
this with investigators down the road is

1821
01:22:25.660 --> 01:22:28.200
you've touched upon you have to be passionate,

1822
01:22:28.200 --> 01:22:31.710
and you have to really be interested in your field,

1823
01:22:31.710 --> 01:22:33.760
but you guys came from two different places

1824
01:22:33.760 --> 01:22:34.810
and your journeys have been

1825
01:22:34.810 --> 01:22:36.773
really different and interesting.
```

```
1826
01:22:37.980 --> 01:22:40.040
I would like to hear if you guys could share

1827
01:22:40.040 --> 01:22:41.580
actually like tangible

1828
01:22:41.580 --> 01:22:44.880
things that someone right now who's an undergrad

1829
01:22:44.880 --> 01:22:47.140
who may be kind of wondering if data science

1830
01:22:47.140 --> 01:22:49.520
or computer science is something for them,

1831
01:22:49.520 --> 01:22:51.040
what are tangibles that they could do

1832
01:22:51.040 --> 01:22:53.580
besides following your passion and your gut

1833
01:22:53.580 --> 01:22:54.460
and things like that,

1834
01:22:54.460 --> 01:22:56.270
what are the things that they could take?

1835
01:22:56.270 --> 01:22:58.290
People who may not have the same sort

1836
01:22:58.290 --> 01:23:01.620
of opportunities that others may have, what could they do

1837
01:23:04.700 --> 01:23:06.380
<v ->Should I start?</v>

1838
01:23:06.380 --> 01:23:09.640
```

I think one thing that people could do to sort of test

1839
01:23:09.640 --> 01:23:12.918
this out and this I'm not entirely obeying the question

1840
01:23:12.918 --> 01:23:15.780
by saying, by veering away from like find

1841
01:23:15.780 --> 01:23:18.280
something you're passionate about, but learn to code.

1842
01:23:18.280 --> 01:23:20.650
Like I think that is obviously a big part

1843
01:23:20.650 --> 01:23:22.870
of this career path.

1844
01:23:22.870 --> 01:23:25.610
And I personally find it really miserable trying

1845
01:23:25.610 --> 01:23:28.120
to learn a new language when you're just like walking

1846
01:23:28.120 --> 01:23:31.230
through a book and like copying in the commands

1847
01:23:31.230 --> 01:23:32.970
and then just fiddling with them.

1848
01:23:32.970 --> 01:23:36.380
I find the best way to do it is to find a question

1849
01:23:36.380 --> 01:23:37.290
you really wanna answer

1850
01:23:37.290 --> 01:23:39.230
and some data that can help you answer it

```
1851
01:23:39.230 --> 01:23:41.750
or maybe something you really wanna build

1852
01:23:41.750 --> 01:23:44.880
and maybe requires data, maybe it doesn't.

1853
01:23:44.880 --> 01:23:46.000
But if it does require data,

1854
01:23:46.000 --> 01:23:47.050
the data you'd need to build it.

1855
01:23:47.050 --> 01:23:48.150
And actually just try,

1856
01:23:49.250 --> 01:23:50.750
even if you don't have a lot of resources

1857
01:23:50.750 --> 01:23:52.670
there are a lot of free resources on the Internet.

1858
01:23:52.670 --> 01:23:54.960
I'm not even gonna like, just Google it.

1859
01:23:54.960 --> 01:23:56.580
There's tons of stuff out there.

1860
01:23:56.580 --> 01:23:59.170
There are a lot of free programming languages,

1861
01:23:59.170 --> 01:24:01.910
I think in the past when people were all using SAS

1862
01:24:01.910 --> 01:24:03.650
and everything and that's expensive

1863
01:24:03.650 --> 01:24:06.670
```

and you needed university access if you were a student.

1864
01:24:06.670 --> 01:24:09.660
Now, you can just use R, now there's Python.

1865
01:24:09.660 --> 01:24:12.160
There's all sorts of tools out there that are free

1866
01:24:13.479 --> 01:24:15.880
and there are great resources to learn as well.

1867
01:24:15.880 --> 01:24:19.500
So I think if you really want to figure out

1868
01:24:19.500 --> 01:24:22.340
if data science is a good career path for you,

1869
01:24:22.340 --> 01:24:25.570
I would just start by like, yeah, figuring out a project

1870
01:24:25.570 --> 01:24:29.850
you wanna do and just give it a shot, just go around browse.

1871
01:24:29.850 --> 01:24:31.990
People will explain things differently depending

1872
01:24:31.990 --> 01:24:32.970
on the resource you find,

1873
01:24:32.970 --> 01:24:36.020
find somebody who's explanations work for you,

1874
01:24:36.020 --> 01:24:36.853
work through it

1875
01:24:36.853 --> 01:24:38.500
and see it'll be really frustrating at first.

1876
01:24:38.500 --> 01:24:40.790
I remember when I first came to programming

1877
01:24:40.790 --> 01:24:43.950
I was in a class and it was like, i equals i plus one

1878
01:24:43.950 --> 01:24:48.950
and I was like false, no, unless it's infinity, no,

1879
01:24:48.950 --> 01:24:51.610
that's not how math works and having to like work through,

1880
01:24:51.610 --> 01:24:54.000
sort of rewire your brain to think,

1881
01:24:54.000 --> 01:24:56.490
okay, no, I'm actually assigning it.

1882
01:24:56.490 --> 01:24:57.580
Yeah. It'll be frustrating.

1883
01:24:57.580 --> 01:24:59.780
It'll be like banging your head against a wall,

1884
01:24:59.780 --> 01:25:00.630
I think at times.

1885
01:25:00.630 --> 01:25:04.070
But if you enjoy that kind of like head banging,

1886
01:25:04.070 --> 01:25:07.940
brain twisting sort of work, this is probably a good place.

1887
01:25:07.940 --> 01:25:08.780
<v ->Thanks Dr. Lum.</v>

1888
01:25:08.780 --> 01:25:09.920

So give it a try and see

1889
01:25:09.920 --> 01:25:11.930
if that's something that you're pulled to,

1890
01:25:11.930 --> 01:25:13.670
gravitate towards, that sounds great.

1891
01:25:13.670 --> 01:25:14.970
What about you, Dr. Curtis?

1892
01:25:14.970 --> 01:25:17.350
<v ->Well, actually, one of the things I didn't include is</v>

1893
01:25:17.350 --> 01:25:20.750
when I was in high school, my dad saw

1894
01:25:20.750 --> 01:25:22.570
or he heard some people at work talking

1895
01:25:22.570 --> 01:25:24.420
about a computer programming class

1896
01:25:24.420 --> 01:25:27.100
that the college was doing for high school students.

1897
01:25:27.100 --> 01:25:30.330
So of course, that's how I spent my Saturday afternoons.

1898
01:25:30.330 --> 01:25:34.240
So I did learn initially to do some basic coding

1899
01:25:34.240 --> 01:25:37.720
'cause my dad was like computers are gonna are the future.

1900
01:25:37.720 --> 01:25:41.930
And so I did that, and then when I got my job

1901
01:25:41.930 --> 01:25:43.480
at the drug treatment center

1902
01:25:43.480 --> 01:25:45.290
where I was doing HIV counseling,

1903
01:25:45.290 --> 01:25:49.050
the first thing I developed, (chuckles) do you, guys,

1904
01:25:49.050 --> 01:25:50.460
know those like cosmic quiz

1905
01:25:50.460 --> 01:25:52.440
like what type of person are you?

1906
01:25:52.440 --> 01:25:54.763
I did one where it was like,

1907
01:25:56.780 --> 01:25:59.670
it was about your sexual activity and your risk.

1908
01:25:59.670 --> 01:26:02.500
So I was asking questions like what type of sex do like,

1909
01:26:02.500 --> 01:26:05.520
how do you it, all the different flavors.

1910
01:26:05.520 --> 01:26:07.370
And then it was like a risk score

1911
01:26:07.370 --> 01:26:09.730
and depending on their different behaviors,

1912
01:26:09.730 --> 01:26:11.080
it would assign risks.

1913
01:26:11.080 --> 01:26:13.280

And so I would do that instead of doing,

1914
01:26:13.280 --> 01:26:15.530
we had this boring regular questionnaire.

1915
01:26:15.530 --> 01:26:17.018
So I would still have them do the questionnaire

1916
01:26:17.018 --> 01:26:18.600
because I have to have it for documentation.

1917
01:26:18.600 --> 01:26:21.780
But then I put people in a computer program, they do it.

1918
01:26:21.780 --> 01:26:24.990
And then we would talk about this in harm reduction.

1919
01:26:24.990 --> 01:26:27.643
So for me, while I'm not a computer programmer

1920
01:26:27.643 --> 01:26:29.540
or a data scientist, in that sense,

1921
01:26:29.540 --> 01:26:31.280
you had to use some of those tools.

1922
01:26:31.280 --> 01:26:34.850
It was more like how to get people engaged

1923
01:26:34.850 --> 01:26:38.370
in the content to assess risk, to assess things,

1924
01:26:38.370 --> 01:26:40.580
and then to deliver something that was cool

1925
01:26:40.580 --> 01:26:42.350
and innovative that would keep them engaged

```
1926
01:26:42.350 --> 01:26:44.143
so that we could work on treatment.

1927
01:26:47.110 --> 01:26:50.270
<v ->So then the advice would be to those young...</v>

1928
01:26:50.270 --> 01:26:53.060
<v ->My advice would be to have fun with it.</v>

1929
01:26:53.060 --> 01:26:55.820
You're not gonna wanna do something that's boring.

1930
01:26:55.820 --> 01:26:58.270
You do need to learn the skills, which is sometime boring,

1931
01:26:58.270 --> 01:27:00.120
but then apply it to what you wanna do.

1932
01:27:00.120 --> 01:27:00.953
I mean, I don't know,

1933
01:27:00.953 --> 01:27:04.050
even you wanna play with what Pokemon character are you

1934
01:27:04.050 --> 01:27:05.210
or whatever it is,

1935
01:27:05.210 --> 01:27:07.640
apply it to something that you're interested in

1936
01:27:07.640 --> 01:27:10.280
so that you can just keep getting the skills.

1937
01:27:10.280 --> 01:27:13.970
You do need the skills, but fortunately those skills,

1938
01:27:13.970 --> 01:27:15.860
```

```
a lot of times, they're not reliant

1939
01:27:15.860 --> 01:27:17.630
of one of your social economic class.

1940
01:27:17.630 --> 01:27:18.830
As you plan out,

1941
01:27:18.830 --> 01:27:21.130
there are lots of free things that you can do.

1942
01:27:22.002 --> 01:27:24.020
Well, I'm not gonna list a couple of the programs either,

1943
01:27:24.020 --> 01:27:27.060
but there's lots of stuff that have broken

1944
01:27:27.060 --> 01:27:28.740
those barriers of income.

1945
01:27:28.740 --> 01:27:30.150
And if you have a smart phone, some of them,

1946
01:27:30.150 --> 01:27:32.360
you can even just use your smart phone and code on.

1947
01:27:32.360 --> 01:27:34.810
So go after it, have fun,

1948
01:27:34.810 --> 01:27:37.750
apply it to something you're interested in.

1949
01:27:37.750 --> 01:27:38.741
<v ->Thanks to you both,</v>

1950
01:27:38.741 --> 01:27:41.093
I think that's very, very useful info.
```

```
1951
01:27:44.410 --> 01:27:45.243
<v ->Great, thank you.</v>

1952
01:27:45.243 --> 01:27:46.740
So we had one question come in

1953
01:27:46.740 --> 01:27:50.580
and it was about with all of the different ways

1954
01:27:50.580 --> 01:27:52.463
to gather data and to survey,

1955
01:27:53.480 --> 01:27:56.320
how do you decide what is a high priority

1956
01:27:56.320 --> 01:27:58.010
for you to spend your time on?

1957
01:27:58.010 --> 01:27:59.950
How do you decide what kind

1958
01:27:59.950 --> 01:28:00.980
of questions do you wanna tackle?

1959
01:28:00.980 --> 01:28:02.423
What's a high priority?

1960
01:28:03.360 --> 01:28:04.550
Where does that drive come from

1961
01:28:04.550 --> 01:28:06.000
or how do you narrow it down?

1962
01:28:09.020 --> 01:28:10.730
<v ->Okay, I'll start.</v>

1963
01:28:10.730 --> 01:28:14.030
```

```
I go for clinical relevance.

1964
01:28:14.030 --> 01:28:18.430
What I think is going to have a high impact in the clinic

1965
01:28:18.430 --> 01:28:23.430
as well as what could have a high level

1966
01:28:24.240 --> 01:28:27.510
nationally kind of surveillance tool

1967
01:28:27.510 --> 01:28:30.820
and it's quick and easy and something that can apply to them

1968
01:28:30.820 --> 01:28:33.463
at a level from the community to the nation.

1969
01:28:34.550 --> 01:28:39.550
Also has some indicators that could not be applied well,

1970
01:28:40.000 --> 01:28:41.650
but so I kind of go from there,

1971
01:28:41.650 --> 01:28:45.173
more of an application use case scenario.

1972
01:28:47.050 --> 01:28:49.030
<v ->Yeah, I think I borrow some of my thinking</v>

1973
01:28:49.030 --> 01:28:52.930
on this from my time at HRDAG where there are a suite

1974
01:28:52.930 --> 01:28:54.860
of questions that we asked ourselves there.

1975
01:28:54.860 --> 01:28:56.290
I don't even honestly remember
```

1976
01:28:56.290 --> 01:28:59.540
all of them off the top of my head, but one of them is

1977
01:28:59.540 --> 01:29:02.610
about how can my skills actually contribute to this?

1978
01:29:02.610 --> 01:29:04.450
Do I have something unique to say in this area?

1979
01:29:04.450 --> 01:29:06.380
So that's one of the things I ask myself.

1980
01:29:06.380 --> 01:29:10.080
And the other one is in this case say we do a great job

1981
01:29:10.080 --> 01:29:13.440
and we get to some sort of truth, does the truth matter?

1982
01:29:13.440 --> 01:29:15.890
And in some sense, the truth always matters,

1983
01:29:15.890 --> 01:29:17.750
just sort of as a value I hold,

1984
01:29:17.750 --> 01:29:19.700
the truth is important just in general,

1985
01:29:20.750 --> 01:29:21.583
but also does it matter?

1986
01:29:21.583 --> 01:29:22.680
Is it policy relevant?

1987
01:29:22.680 --> 01:29:25.100
Is this something where people know the truth,

1988
01:29:25.100 --> 01:29:27.940

```
they can act on it, they can maybe just sort of speaking

1989
01:29:27.940 --> 01:29:29.010
and generally do

1990
01:29:29.010 --> 01:29:31.140
something about it to make the world better?

1991
01:29:31.140 --> 01:29:32.670
Will this be relevant to policy makers,

1992
01:29:32.670 --> 01:29:34.600
will this actually move things forward?

1993
01:29:34.600 --> 01:29:37.840
And so those are sort of the two main questions

1994
01:29:37.840 --> 01:29:39.580
that I think I've, again, borrowed from my time

1995
01:29:39.580 --> 01:29:43.830
at HRDAG that helped me sort of whittle down all the

1996
01:29:43.830 --> 01:29:45.860
sort of infinite possibilities of projects

1997
01:29:45.860 --> 01:29:47.860
that one could work on at this point to try

1998
01:29:47.860 --> 01:29:52.793
to find one that I think is a good use of time and skill.

1999
01:29:54.050 --> 01:29:54.883
<v ->Thank you very much.</v>

2000
01:29:54.883 --> 01:29:56.600
Thank you for answering all of the questions
```

```
2001
01:29:56.600 --> 01:29:59.050
and for your presentations.

2002
01:29:59.050 --> 01:29:59.980
I see that we're out of time,

2003
01:29:59.980 --> 01:30:01.963
so I will turn it back over to Susan.

2004
01:30:04.350 --> 01:30:05.183
<v ->Thank you, Lindsey.</v>

2005
01:30:05.183 --> 01:30:08.363
I just wanna thank you again, Dr. Lum and Dr. Curtis.

2006
01:30:08.363 --> 01:30:10.200
It's been great hearing about your careers

2007
01:30:10.200 --> 01:30:11.120
and what you're working on.

2008
01:30:11.120 --> 01:30:13.720
And I think you've given a lot of great advice

2009
01:30:13.720 --> 01:30:15.870
to the the next generation of data scientists.

2010
01:30:15.870 --> 01:30:17.460
It's been really enjoyable this morning hearing

2011
01:30:17.460 --> 01:30:18.547
from both of you.

2012
01:30:18.547 --> 01:30:20.650
And I also just wanna announce that next week,

2013
01:30:20.650 --> 01:30:22.880
```

next Monday at 9:00 a.m. is our final session

2014
01:30:22.880 --> 01:30:24.890
of this four-part series.

2015
01:30:24.890 --> 01:30:26.500
We'll be hearing from Dan Jacobson

2016
01:30:26.500 --> 01:30:28.410
from Oak Ridge National Laboratory

2017
01:30:28.410 --> 01:30:30.120
and Mike Tamir from SIG

2018
01:30:30.120 --> 01:30:32.500
and he's also a faculty at UC Berkeley.

2019
01:30:32.500 --> 01:30:33.610
So please, go ahead and register.

2020
01:30:33.610 --> 01:30:36.020
I think the registration link is in the chat

2021
01:30:36.020 --> 01:30:37.610
and hopefully, we'll see you all again next week.

2022
01:30:37.610 --> 01:30:39.320
And thanks again to our speakers

2023
01:30:39.320 --> 01:30:40.406
and thanks again to the organizers

2024
01:30:40.406 --> 01:30:41.840
and everyone that's on the team.

2025
01:30:41.840 --> 01:30:43.630
And I know it's hard in the virtual environment,

2026
01:30:43.630 --> 01:30:44.560
but some virtual applause

2027
01:30:44.560 --> 01:30:45.940
for everybody for being a part

2028
01:30:45.940 --> 01:30:47.580
of this this morning.

2029
01:30:47.580 --> 01:30:48.430
Thanks, everyone.