1
00:00:02.010 --> 00:00:02.843
<v ->Good morning, everyone,</v>

2
00:00:02.843 --> 00:00:04.910
and welcome to the fourth and final session

3
00:00:04.910 --> 00:00:07.430
of our Data Science Career Seminar Series,

4
00:00:07.430 --> 00:00:09.790
Bringing Data Science to Addiction Research.

5
00:00:09.790 --> 00:00:10.810
My name is Susan Wright,

6
00:00:10.810 --> 00:00:13.420
and I'm from the Division of Neuroscience and Behavior.

7
00:00:13.420 --> 00:00:14.350
I'm the Program Director

8
00:00:14.350 --> 00:00:15.900
for Big Data and Computational Science,

9
00:00:15.900 --> 00:00:18.700
and I'm leading our data science efforts here at NIDA.

10
00:00:18.700 --> 00:00:21.080
Training in data science is a priority for NIDA,

11
00:00:21.080 --> 00:00:23.710
and it's supported by our new Office of Research Training,

12
00:00:23.710 --> 00:00:26.010
Diversity and Disparities.

13
00:00:26.010 --> 00:00:28.240
We've organized a seminar series with the full support

14
00:00:28.240 --> 00:00:30.760
of our NIDA director, Dr. Nora Volkow,

15
00:00:30.760 --> 00:00:32.130
and the organizers include members

16
00:00:32.130 --> 00:00:34.290
of the Division of Neuroscience and Behavior

17
00:00:34.290 --> 00:00:35.860
and the Office of Research, Training,

18
00:00:35.860 --> 00:00:37.940
Diversities, and Disparities.

19
00:00:37.940 --> 00:00:39.620
The organizers include myself,

20
00:00:39.620 --> 00:00:41.720
Dr. Roger Little, Deputy Director

21
00:00:41.720 --> 00:00:43.900
of the Division of Neuroscience and Behavior,

22
00:00:43.900 --> 00:00:46.790
Dr. Wilson Compton, the NIDA Deputy Director,

23
00:00:46.790 --> 00:00:48.650
and the acting Director of the Office of Research,

24
00:00:48.650 --> 00:00:50.800
Training, Diversity, and Disparities,

25
00:00:50.800 --> 00:00:53.060

Dr. Albert Avila, the Deputy Director

26
00:00:53.060 --> 00:00:54.450
of the Office of Research, Training,

27
00:00:54.450 --> 00:00:56.090
Diversity, and Disparities,

28
00:00:56.090 --> 00:00:58.220
and the Director of the Office of Disparities

29
00:00:58.220 --> 00:00:59.560
and Health Disparities,

30
00:00:59.560 --> 00:01:02.210
and Dr. Lindsey Friend, the Research and Career Development

31
00:01:02.210 --> 00:01:04.810
Program Officer and the Office of Research, Training,

32
00:01:04.810 --> 00:01:07.396
Diversity, and Disparities.

33
00:01:07.396 --> 00:01:09.410
I want to thank Roger, Wilson, Albert, and Lindsey,

34
00:01:09.410 --> 00:01:11.900
for their help with organizing this seminar series,

35
00:01:11.900 --> 00:01:13.540
and I also want to thank the team who has been helping

36
00:01:13.540 --> 00:01:14.890
with the technical details,

37
00:01:14.890 --> 00:01:17.680
and that includes Isha Charia, Susan Holbrook,

38
00:01:17.680 --> 00:01:19.803
Caitlin Dutera, and David Maza.

39
00:01:20.643 --> 00:01:21.600
For this session,

40
00:01:21.600 --> 00:01:24.090
we'll first have an interview with Dr. Mike Tamir,

41
00:01:24.090 --> 00:01:26.760
then we'll have a presentation by Dr. Daniel Jacobson,

42
00:01:26.760 --> 00:01:28.030
and then we'll have a joint Q and A session

43
00:01:28.030 --> 00:01:30.470
where we'll take questions from the audience.

44
00:01:30.470 --> 00:01:32.670
Please use the chat box to submit your questions,

45
00:01:32.670 --> 00:01:34.820
and we'll get to as many of them as we can.

46
00:01:36.567 --> 00:01:39.090
So introduction to Dr. Mike Tamir.

47
00:01:39.090 --> 00:01:41.710
Mike serves as the Chief Machine Learning Scientist

48
00:01:41.710 --> 00:01:44.070
and Head of Machine Learning for SIG,

49
00:01:44.070 --> 00:01:46.690
is also UC Berkeley data science faculty,

50
00:01:46.690 --> 00:01:49.300

and the Director of the university's machine learning labs.

51
00:01:49.300 --> 00:01:50.880
He has led teams of data scientists

52
00:01:50.880 --> 00:01:54.890
in the Bay area as Head of Data Science at Uber ATG,

53
00:01:54.890 --> 00:01:57.383
Chief Data Scientist for Intertrust Intact,

54
00:01:58.260 --> 00:02:00.800
the Director of Data Science for MetaScale Sears,

55
00:02:00.800 --> 00:02:02.620
and CSO for Galvanize,

56
00:02:02.620 --> 00:02:06.010
where he founded Galvanize U UNH accredited Masters

57
00:02:06.010 --> 00:02:07.260
in data science degree,

58
00:02:07.260 --> 00:02:09.090
then oversaw the company's transformation

59
00:02:09.090 --> 00:02:12.310
from co-working space to data science organization.

60
00:02:12.310 --> 00:02:14.030
Mike began his career in academia,

61
00:02:14.030 --> 00:02:16.060
serving as mathematics teaching fellow

62
00:02:16.060 --> 00:02:17.470
for Columbia University

63
00:02:17.470 --> 00:02:20.010
before teaching at the University of Pittsburgh.

64
00:02:20.010 --> 00:02:22.960
So welcome Mike, and thank you for joining us this morning.

65
00:02:24.006 --> 00:02:28.730
So since data scientist is a relatively new job title,

66
00:02:28.730 --> 00:02:30.430
sometimes people aren't exactly sure

67
00:02:30.430 --> 00:02:32.340
what a data scientist does.

68
00:02:32.340 --> 00:02:35.040
How do you typically describe what you do, when asked?

69
00:02:39.970 --> 00:02:43.963
<v ->Well, thanks for having me and welcome everyone.</v>

70
00:02:46.190 --> 00:02:51.190
So data scientist as a job title is fairly new.

71
00:02:51.630 --> 00:02:56.630
When I got my first job as a data scientist,

72
00:02:57.300 --> 00:02:59.570
I had to Google it.

73
00:02:59.570 --> 00:03:01.795
I didn't know what it meant.

74
00:03:01.795 --> 00:03:06.795
And so, and you know, so it was not so long ago,

75
00:03:07.170 --> 00:03:08.387

but long enough ago,

76
00:03:08.387 --> 00:03:10.770
there weren't billions of articles,

77
00:03:10.770 --> 00:03:12.940
but or media manual what is a data scientist,

78
00:03:12.940 --> 00:03:15.019
and what data scientists do.

79
00:03:15.019 --> 00:03:19.510
The term has appeared to evolve a little bit

80
00:03:19.510 --> 00:03:21.793
over the past several years,

81
00:03:22.860 --> 00:03:23.693
I think that, you know,

82
00:03:23.693 --> 00:03:27.403
certainly in the early years there wasn't much definition,

83
00:03:28.590 --> 00:03:31.533
whereas on the other end of the tunnel,

84
00:03:32.590 --> 00:03:37.590
now data science more and more is not associated with

85
00:03:38.730 --> 00:03:43.730
analytics but is a little bit more of a mix of the core

86
00:03:44.930 --> 00:03:48.520
machine learning engineering techniques that we use,

87
00:03:48.520 --> 00:03:52.160
or a lot of data scientists use,

88
00:03:52.160 --> 00:03:57.160
and some of the bread and butter data analysis techniques

89
00:03:58.040 --> 00:03:59.570
that you have to do.

90
00:03:59.570 --> 00:04:01.500
So let me break that down a little bit.

91
00:04:01.500 --> 00:04:04.590
First, there is a, you know, whenever you get a data set,

92
00:04:04.590 --> 00:04:05.870
there's a lot that you need to do

93
00:04:05.870 --> 00:04:09.000
in order to just prep the data for modeling.

94
00:04:09.000 --> 00:04:12.170
And certainly when I think about the job

95
00:04:12.170 --> 00:04:16.900
of being a data scientist, it usually involves

96
00:04:16.900 --> 00:04:19.140
some sort of forecasting or creating,

97
00:04:19.140 --> 00:04:20.790
you know, algorithmic estimators.

98
00:04:21.700 --> 00:04:24.010
But more than that, there's what I say,

99
00:04:24.010 --> 00:04:27.260
you know, often, in the job,

100
00:04:27.260 --> 00:04:28.820

we'll build these deep learning models,

101
00:04:28.820 --> 00:04:30.850
and maybe they have to be trained for months,

102
00:04:30.850 --> 00:04:32.380
and you have these results,

103
00:04:32.380 --> 00:04:37.240
and then you get your estimations on a certain phenomenon,

104
00:04:37.240 --> 00:04:39.470
and then you have to say, okay now it's time

105
00:04:39.470 --> 00:04:41.360
for the data science work,

106
00:04:41.360 --> 00:04:45.070
where you actually go through, you look at all

107
00:04:45.070 --> 00:04:47.830
of the residuals of your performance of your data yet,

108
00:04:47.830 --> 00:04:50.220
you dig into where you're making mistakes,

109
00:04:50.220 --> 00:04:51.330
where you're not making mistakes,

110
00:04:51.330 --> 00:04:53.150
you start to produce hypotheses

111
00:04:53.150 --> 00:04:57.850
about why did this particular model do well here,

112
00:04:57.850 --> 00:04:59.790
and not do well there,

113
00:04:59.790 --> 00:05:03.400
you have to involve, it involves a lot of, you know,

114
00:05:03.400 --> 00:05:06.170
bread and butter statistical and experimental techniques,

115
00:05:06.170 --> 00:05:08.020
like making sure you stratify right,

116
00:05:08.020 --> 00:05:09.710
making sure you're using the right metrics,

117
00:05:09.710 --> 00:05:12.270
making sure you have the right sort of experimental design

118
00:05:12.270 --> 00:05:15.680
when you're saying if one particular technique

119
00:05:15.680 --> 00:05:17.200
improves or doesn't improve.

120
00:05:17.200 --> 00:05:22.200
And so there is a lot more science involved

121
00:05:23.780 --> 00:05:27.180
if you think about it, just like computer science, you know,

122
00:05:27.180 --> 00:05:30.130
came from this, you know, looking at natural phenomenons,

123
00:05:30.130 --> 00:05:32.590
and then trying to create hypotheses,

124
00:05:32.590 --> 00:05:35.590
and approach as a scientist, you know,

125
00:05:35.590 --> 00:05:40.590

how do these mechanical objects like computers

126
00:05:40.680 --> 00:05:43.480
approach them and how they run algorithms

127
00:05:44.320 --> 00:05:45.870
from a scientific perspective,

128
00:05:45.870 --> 00:05:50.110
a lot of what I do day-to-day at least now is

129
00:05:50.110 --> 00:05:51.363
for a given set of data,

130
00:05:52.670 --> 00:05:57.230
your experimental design is centered around

131
00:05:57.230 --> 00:06:00.909
what kind of machine learning, primarily, techniques,

132
00:06:00.909 --> 00:06:02.400
deep learning techniques

133
00:06:03.250 --> 00:06:05.210
we'll be able to get at the signal in that.

134
00:06:05.210 --> 00:06:10.210
And you're really, when we have term papers for students,

135
00:06:11.420 --> 00:06:15.100
this is how I describe how to think about their term papers.

136
00:06:15.100 --> 00:06:19.340
You have a baseline, that's your control,

137
00:06:19.340 --> 00:06:21.210
and then you have a treatment model,

138
00:06:21.210 --> 00:06:23.610
I'm gonna try this tweak in my architecture,

139
00:06:23.610 --> 00:06:26.230
I'm gonna try this tweak in my training algorithm.

140
00:06:26.230 --> 00:06:28.540
And that's the way you think about it,

141
00:06:28.540 --> 00:06:33.540
and, you know, returning to that scientific approach,

142
00:06:34.200 --> 00:06:35.610
really helps to make sure

143
00:06:35.610 --> 00:06:38.980
that you're not just kind of randomly pressing buttons

144
00:06:38.980 --> 00:06:43.980
or trying to get lucky with a certain neural architecture.

145
00:06:50.100 --> 00:06:51.487
<v ->Yeah, definitely very interesting to hear</v>

146
00:06:51.487 --> 00:06:53.440
how the term has evolved.

147
00:06:53.440 --> 00:06:54.710
So along those lines,

148
00:06:54.710 --> 00:06:57.680
a lot of people have been rebranding as data scientists.

149
00:06:57.680 --> 00:07:00.870
Do you see a need to define data science roles more clearly?

150
00:07:00.870 --> 00:07:03.070

And do you see a big difference in the skills needed

151
00:07:03.070 --> 00:07:04.520
for academia versus industry?

152
00:07:06.900 --> 00:07:11.900
<v ->So I don't think that the rebranding possibly, you know,</v>

153
00:07:16.752 --> 00:07:20.760
the cliche is that if you are,

154
00:07:20.760 --> 00:07:22.250
at least for the past decade has been,

155
00:07:22.250 --> 00:07:24.890
if you change your job title on LinkedIn,

156
00:07:24.890 --> 00:07:29.493
you get a 20% salary bump and that's a good strategy, right?

157
00:07:31.250 --> 00:07:35.540
That being said, we don't actually have, you know,

158
00:07:35.540 --> 00:07:40.255
like open JDs for scientists at SIG right now,

159
00:07:40.255 --> 00:07:41.680
for data scientists.

160
00:07:41.680 --> 00:07:44.400
We have them for machine learning research scientists.

161
00:07:44.400 --> 00:07:48.593
We have them for machine learning engineers.

162
00:07:49.750 --> 00:07:53.053
Honestly, I would think that a lot of that is nominal,

163
00:07:53.970 --> 00:07:57.440
the engineers and the research scientists have to do

164
00:07:57.440 --> 00:08:01.470
what I usually refer to as the last question is that

165
00:08:01.470 --> 00:08:03.100
the bread and butter data science work,

166
00:08:03.100 --> 00:08:06.370
which really looks at, you know,

167
00:08:06.370 --> 00:08:08.230
when I say the data science work,

168
00:08:08.230 --> 00:08:11.590
I mean the exploring the data, getting manually

169
00:08:11.590 --> 00:08:16.410
into feature engineering or data prep

170
00:08:16.410 --> 00:08:19.160
for that particular model or tactic

171
00:08:19.160 --> 00:08:20.430
that you're going to be using,

172
00:08:20.430 --> 00:08:24.210
and the residual analysis and making sure that you do

173
00:08:24.210 --> 00:08:28.040
careful exploration of the residual analysis,

174
00:08:28.040 --> 00:08:30.190
in order to generate your next hypothesis.

175
00:08:30.190 --> 00:08:34.950

That's the scientific, the data science practice

176
00:08:37.400 --> 00:08:39.450
what gets called a data scientist,

177
00:08:39.450 --> 00:08:44.450
or that is maybe a little bit vaguer.

178
00:08:44.810 --> 00:08:47.040
And I've noticed in some companies,

179
00:08:47.040 --> 00:08:49.060
some of the larger companies in particular,

180
00:08:49.060 --> 00:08:51.960
sometimes data science is more like a product role.

181
00:08:51.960 --> 00:08:54.250
Sometimes it's indistinguishable

182
00:08:54.250 --> 00:08:55.850
from machine learning engineer,

183
00:08:55.850 --> 00:08:59.880
and so it really, what I would do if you're not sure

184
00:08:59.880 --> 00:09:01.920
about what you're signing up for,

185
00:09:01.920 --> 00:09:04.223
I would make sure to ask the hiring manager and ask

186
00:09:04.223 --> 00:09:06.440
the other people at that role, what their day-to-day is,

187
00:09:06.440 --> 00:09:09.230
'cause that's probably gonna be your best insight

188
00:09:09.230 --> 00:09:11.230
into what your job is actually gonna be.

189
00:09:17.960 --> 00:09:19.970
<v ->So you have a couple of different roles right now.</v>

190
00:09:19.970 --> 00:09:22.733
What would you say attracted you to your current roles?

191
00:09:26.810 --> 00:09:31.770
<v ->So at Susquehanna in particular, you know,</v>

192
00:09:31.770 --> 00:09:35.643
I've been, or I had been before joining,

193
00:09:36.570 --> 00:09:40.090
wandering for quite a long time, you know,

194
00:09:40.090 --> 00:09:44.330
if it was even possible to separate signal from noise,

195
00:09:44.330 --> 00:09:49.330
in such a narrow signal to noise ratio context,

196
00:09:53.490 --> 00:09:56.010
this is the sort of situation where, you know,

197
00:09:56.010 --> 00:09:58.200
you get a lot of data.

198
00:09:58.200 --> 00:10:00.380
You know, usually our problem is

199
00:10:00.380 --> 00:10:02.150
not that we don't have enough data volume

200
00:10:02.150 --> 00:10:07.150

to support saturation in full training in convergence

201
00:10:08.290 --> 00:10:11.630
with as large of a neural network, as you want,

202
00:10:11.630 --> 00:10:13.120
that's not the issue.

203
00:10:13.120 --> 00:10:16.880
The two main issues are with frequency,

204
00:10:16.880 --> 00:10:18.200
seeing if you can then, you know,

205
00:10:18.200 --> 00:10:19.890
once you are able to separate the signal,

206
00:10:19.890 --> 00:10:24.040
can I do it, and serve a result in the, you know,

207
00:10:24.040 --> 00:10:26.500
milliseconds or nanoseconds, whatever it is,

208
00:10:26.500 --> 00:10:28.570
that you need for certain applications.

209
00:10:28.570 --> 00:10:29.910
And there's always a trade off there, right?

210
00:10:29.910 --> 00:10:32.820
The more, the quicker the frequency,

211
00:10:32.820 --> 00:10:35.780
the lower the latency window that you have,

212
00:10:35.780 --> 00:10:36.670
the more data you have.

213
00:10:36.670 --> 00:10:39.220
Great, that means you can train a neural network,

214
00:10:39.220 --> 00:10:40.750
and you have all the data you want, right?

215
00:10:40.750 --> 00:10:43.693
You've got this wealth the embarrassment of riches, right?

216
00:10:45.470 --> 00:10:49.100
The, unfortunately, coupled with that is the fact

217
00:10:49.100 --> 00:10:50.240
that if you have low latency,

218
00:10:50.240 --> 00:10:53.290
then that also means if you create a large enough model

219
00:10:53.290 --> 00:10:55.210
that takes note, it has enough calculations,

220
00:10:55.210 --> 00:10:58.060
you may not have enough time to actually use that model

221
00:10:58.060 --> 00:11:00.900
because you need to think about all the ways of,

222
00:11:00.900 --> 00:11:03.310
for serving to execute,

223
00:11:03.310 --> 00:11:05.100
just generating that forecast quick enough,

224
00:11:05.100 --> 00:11:07.210
in order to use it once you get the data.

225
00:11:07.210 --> 00:11:09.800

And so that's, you know, you might think,

226
00:11:09.800 --> 00:11:11.210
oh, one day I'll work at a company,

227
00:11:11.210 --> 00:11:12.830
which has all the data in the world,

228
00:11:12.830 --> 00:11:15.220
and then I'll be able to do whatever I want,

229
00:11:15.220 --> 00:11:16.850
but there's gonna be this natural trade off,

230
00:11:16.850 --> 00:11:19.530
it's almost like a, you know, bias variance trade off

231
00:11:19.530 --> 00:11:21.610
where it kind of, there's a balance

232
00:11:21.610 --> 00:11:25.490
in the middle of having enough frequencies to get the data

233
00:11:25.490 --> 00:11:28.860
and being able to actually use it.

234
00:11:28.860 --> 00:11:31.210
The other thing is just, you know,

235
00:11:31.210 --> 00:11:33.853
I mentioned the signal noise ratio,

236
00:11:33.853 --> 00:11:38.610
that is a huge area where it just means

237
00:11:38.610 --> 00:11:41.980
you have to focus on a lot more of that

238
00:11:41.980 --> 00:11:44.360
really careful residual analysis,

239
00:11:44.360 --> 00:11:49.090
any sort of information leakage

240
00:11:49.090 --> 00:11:54.090
or error in your metrics where you get like phantom boost

241
00:11:56.700 --> 00:11:59.690
to your performance is, could be very risky, right?

242
00:11:59.690 --> 00:12:04.500
Because the signal is so weak that, or compared to the noise

243
00:12:04.500 --> 00:12:06.470
that these small margins matter.

244
00:12:06.470 --> 00:12:09.240
And these small margins are in the neighborhood

245
00:12:09.240 --> 00:12:11.490
of what can happen with just, you know,

246
00:12:11.490 --> 00:12:14.670
small missteps in terms of scientific practice.

247
00:12:14.670 --> 00:12:18.040
So that's something else that is an added challenge

248
00:12:18.040 --> 00:12:20.033
that attracted me to Susquehanna.

249
00:12:24.170 --> 00:12:26.560
<v ->Sounds like a very interesting job.</v>

250
00:12:26.560 --> 00:12:28.440

I know that you do a lot of machine learning.

251
00:12:28.440 --> 00:12:30.157
So what are the most interesting (microphone cuts out)

252
00:12:30.157 --> 00:12:32.307
for machine learning that you've worked on?

253
00:12:35.180 --> 00:12:38.053
<v ->So I think you cut out there for a moment</v>

254
00:12:38.053 --> 00:12:40.903
when you said what were the most interesting applications

255
00:12:42.139 --> 00:12:43.164
for machine learning?

256
00:12:43.164 --> 00:12:45.363
<v ->For machine learning that you've worked on?</v>

257
00:12:47.120 --> 00:12:52.120
<v ->Well, so my natural place that I like to do research,</v>

258
00:12:54.580 --> 00:12:57.990
or that I like to to focus continues to be

259
00:12:59.030 --> 00:13:00.723
in natural language processing.

260
00:13:01.850 --> 00:13:04.830
There's something very interesting about teaching computers

261
00:13:04.830 --> 00:13:08.347
how to read and understand language.

262
00:13:08.347 --> 00:13:10.697
And as I understand with an asterisk of course,

263
00:13:11.550 --> 00:13:16.550
you know, the last two years give or take since 2018

264
00:13:18.640 --> 00:13:23.600
or really near 2017 which is at the very end of 2017

265
00:13:23.600 --> 00:13:28.430
with transformers has, you know,

266
00:13:28.430 --> 00:13:33.430
it has induced a new wave of making us think about

267
00:13:33.570 --> 00:13:37.700
what's possible in terms of natural language processing.

268
00:13:37.700 --> 00:13:39.100
And that's been very exciting,

269
00:13:39.100 --> 00:13:43.000
because you can now use these new tactics,

270
00:13:43.000 --> 00:13:46.390
you know, the attention mechanism has kind of turned into

271
00:13:47.290 --> 00:13:52.230
with the convolutional layer has been for image processing

272
00:13:53.700 --> 00:13:55.480
and for computer vision.

273
00:13:55.480 --> 00:13:58.450
Now, not only is it a lot easier

274
00:13:58.450 --> 00:14:03.240
to solve natural language-based problems

275
00:14:03.240 --> 00:14:06.740

with these techniques,

276
00:14:06.740 --> 00:14:09.190
these attention mechanism-based techniques

277
00:14:09.190 --> 00:14:10.703
that transformer techniques,

278
00:14:12.120 --> 00:14:15.720
but also they're just getting so huge.

279
00:14:15.720 --> 00:14:17.880
You know, every couple months you hear

280
00:14:17.880 --> 00:14:18.940
about Google just came out

281
00:14:18.940 --> 00:14:22.310
with one that's on the order of trillions, you know,

282
00:14:22.310 --> 00:14:25.380
that's for the record, it's give or take

283
00:14:25.380 --> 00:14:29.630
one and a half trillion connections in your brain.

284
00:14:29.630 --> 00:14:32.310
So we're getting into the right order of magnitude

285
00:14:32.310 --> 00:14:36.840
of how many parameters in a neural network

286
00:14:36.840 --> 00:14:38.890
to neurons in your brain,

287
00:14:38.890 --> 00:14:42.360
and in fact, the way that Google did it was

288
00:14:43.900 --> 00:14:46.800
probably a lot truer to the way, you know,

289
00:14:46.800 --> 00:14:48.940
the connections in your brain are made

290
00:14:48.940 --> 00:14:51.920
which is that, you know, you have these little submodules,

291
00:14:51.920 --> 00:14:53.700
this mixture of experts of different parts

292
00:14:53.700 --> 00:14:55.700
of the neural network are responsible for different kinds

293
00:14:55.700 --> 00:14:58.350
of questions and get activated in different contexts.

294
00:14:59.630 --> 00:15:01.750
And so, you know, some of the work that I've done in that,

295
00:15:01.750 --> 00:15:04.714
that more recently is, you know, thinking about

296
00:15:04.714 --> 00:15:09.714
how do you deal with things like the fake news problem,

297
00:15:10.050 --> 00:15:14.810
how do you try to sort fact from fiction

298
00:15:14.810 --> 00:15:18.010
or more attractively,

299
00:15:18.010 --> 00:15:21.580
how do you sort a manipulation from, you know,

300
00:15:21.580 --> 00:15:24.990

honest and non-manipulative information exchange,

301
00:15:24.990 --> 00:15:27.240
and is there something that you can actually detect

302
00:15:27.240 --> 00:15:31.840
in the way people use languages, the way people use language

303
00:15:31.840 --> 00:15:33.100
when they write journal articles,

304
00:15:33.100 --> 00:15:34.660
the way they use language when they write blogs,

305
00:15:34.660 --> 00:15:37.090
or the way they use language when they're doing, you know,

306
00:15:37.090 --> 00:15:41.410
it's more punditry or influence peddling.

307
00:15:41.410 --> 00:15:45.807
And it seems like, you know, even for sources like RT,

308
00:15:46.950 --> 00:15:50.500
which is a Russian propaganda site,

309
00:15:50.500 --> 00:15:55.500
you're able to, at least detect some of that signal

310
00:15:55.600 --> 00:15:59.268
that are telltale signs that you know,

311
00:15:59.268 --> 00:16:04.110
this is not a journalism.

312
00:16:04.110 --> 00:16:05.950
This is not a scientific paper.

313
00:16:05.950 --> 00:16:09.410
This is something that's intending to do more,

314
00:16:09.410 --> 00:16:12.900
and that more has to do with manipulating

315
00:16:12.900 --> 00:16:15.900
our human emotions when we read stuff

316
00:16:15.900 --> 00:16:19.820
rather than our logical assessment,

317
00:16:19.820 --> 00:16:21.417
when we think about things.

318
00:16:24.827 --> 00:16:26.890
<v ->Yeah, it's very exciting to hear about these advancements,</v>

319
00:16:26.890 --> 00:16:28.090
but it's definitely very important

320
00:16:28.090 --> 00:16:31.040
to consider all these critical issues.

321
00:16:31.040 --> 00:16:33.960
So some of the long-term implications about transitioning

322
00:16:33.960 --> 00:16:36.710
from ultra big data models to Turing questions,

323
00:16:36.710 --> 00:16:39.160
can you tell us more about your thoughts on this?

324
00:16:41.180 --> 00:16:42.495
<v ->Sure.</v>

325
00:16:42.495 --> 00:16:44.450

There's, you know,

326
00:16:44.450 --> 00:16:48.950
so Turing question being, you know,

327
00:16:48.950 --> 00:16:52.520
can we create machines that are indistinguishable

328
00:16:53.480 --> 00:16:58.070
from an experience perspective of talking with them,

329
00:16:58.070 --> 00:17:00.330
or at least interacting with words

330
00:17:00.330 --> 00:17:02.260
with a machine versus a human.

331
00:17:02.260 --> 00:17:05.517
And so Turing test being, you know,

332
00:17:07.840 --> 00:17:11.150
it may or may not have been Turing's original intent,

333
00:17:11.150 --> 00:17:15.780
but it's too late now, you know, a machine passes the test

334
00:17:15.780 --> 00:17:19.240
if you can't tell the difference, right?

335
00:17:19.240 --> 00:17:21.920
And so if as many people guessed

336
00:17:21.920 --> 00:17:26.920
that the machine is a human as they would for humans,

337
00:17:27.600 --> 00:17:29.710
just naturally making a mistake that a human is a machine,

338
00:17:29.710 --> 00:17:33.050
or vice versa, then it passes the test,

339
00:17:33.050 --> 00:17:37.270
and Turing machines, or, sorry, not Turing machines,

340
00:17:37.270 --> 00:17:42.270
modern ultra big models, these language models,

341
00:17:42.880 --> 00:17:47.880
so a language model is just this construct of

342
00:17:48.000 --> 00:17:50.880
I give words, I input words,

343
00:17:50.880 --> 00:17:53.830
and then it produces a, you know

344
00:17:53.830 --> 00:17:57.520
it continues that string of words in an appropriate way,

345
00:17:57.520 --> 00:17:59.700
in a way that sounds fluid and appropriate,

346
00:17:59.700 --> 00:18:02.700
and in context for whatever words prompted it,

347
00:18:02.700 --> 00:18:03.800
that's a language model.

348
00:18:03.800 --> 00:18:07.652
And so we experience language models all the time,

349
00:18:07.652 --> 00:18:09.900
or, you know, at least, well now they're actually

350
00:18:09.900 --> 00:18:10.733

transformer-based too,

351
00:18:10.733 --> 00:18:13.470
but when you do autocomplete in your browser,

352
00:18:13.470 --> 00:18:15.930
that is a language model,

353
00:18:15.930 --> 00:18:18.870
when it tries to tell you what word

354
00:18:18.870 --> 00:18:20.600
you're probably going to say next,

355
00:18:20.600 --> 00:18:23.053
and it gives you an option for that search.

356
00:18:24.460 --> 00:18:28.380
Now, actually, as of, I think, 18 months ago,

357
00:18:28.380 --> 00:18:30.460
Google has introduced transformers

358
00:18:30.460 --> 00:18:33.550
even to that process of select search

359
00:18:33.550 --> 00:18:35.680
will use transformer models.

360
00:18:35.680 --> 00:18:39.150
These very big models like GPT, first it was GPT-1,

361
00:18:39.150 --> 00:18:41.810
and then GPT-2, and then GPT-3,

362
00:18:41.810 --> 00:18:44.540
which had the crown for the most parameters

363
00:18:44.540 --> 00:18:45.380
as of a year ago.

364
00:18:45.380 --> 00:18:48.890
And now there's one, you know, Microsoft and Google

365
00:18:48.890 --> 00:18:51.930
have come up with a slightly bigger ones,

366
00:18:51.930 --> 00:18:55.470
all of these are focused on just pouring on

367
00:18:55.470 --> 00:18:57.730
more and more and more and more parameters

368
00:18:57.730 --> 00:19:00.570
and more neurons, more connections between those neurons,

369
00:19:00.570 --> 00:19:05.080
each connection has a parameter in order to get better

370
00:19:05.080 --> 00:19:07.920
at finding the response, the appropriate response,

371
00:19:07.920 --> 00:19:09.730
but these are still fundamentally language models.

372
00:19:09.730 --> 00:19:12.380
And so they do fascinating things,

373
00:19:12.380 --> 00:19:15.980
like you can ask factual questions of GPT-3,

374
00:19:15.980 --> 00:19:19.030
like who was the President in 1876,

375
00:19:19.030 --> 00:19:20.723

and it'll probably get it right.

376
00:19:22.660 --> 00:19:23.710
But it's just the language model,

377
00:19:23.710 --> 00:19:24.750
it can just find

378
00:19:24.750 --> 00:19:27.130
what's the most appropriate response given the prompting.

379
00:19:27.130 --> 00:19:30.060
So you can also ask it

380
00:19:30.060 --> 00:19:33.470
who was the President of the United States in 1776,

381
00:19:33.470 --> 00:19:37.560
and, or, sorry, not 1776, 1676.

382
00:19:37.560 --> 00:19:42.560
And it'll respond with a perfect Dunning Kruger confidence.

383
00:19:43.760 --> 00:19:46.250
Well, the President was William Penn, or, you know,

384
00:19:46.250 --> 00:19:49.240
someone who was important and was a relevant figure

385
00:19:49.240 --> 00:19:51.570
of the time and can get that context right,

386
00:19:51.570 --> 00:19:55.100
but it doesn't have any sort of self-auditing of,

387
00:19:55.100 --> 00:19:56.810
you know, does this make sense?

388
00:19:56.810 --> 00:19:57.960
Was there a, you know,

389
00:19:57.960 --> 00:19:59.800
were these additional facts

390
00:19:59.800 --> 00:20:03.620
like United States was not even a twinkle in the eye

391
00:20:03.620 --> 00:20:07.390
of Americans yet, you know,

392
00:20:07.390 --> 00:20:08.690
it won't be able to tell that.

393
00:20:08.690 --> 00:20:13.690
And so I forget the, it was, you know,

394
00:20:14.130 --> 00:20:19.130
LeCun or Bengio, I'm gonna roughly paraphrase.

395
00:20:20.240 --> 00:20:23.010
There's a lot of criticism that these language models

396
00:20:23.010 --> 00:20:26.310
are not, you know, fully robust general AI,

397
00:20:26.310 --> 00:20:31.060
and the comparison was that he made is, you know,

398
00:20:31.060 --> 00:20:34.580
comparing a language model to a generalized AI,

399
00:20:34.580 --> 00:20:37.430
it's like comparing a high altitude aircraft

400
00:20:37.430 --> 00:20:39.820

to a rocket to the Moon.

401
00:20:39.820 --> 00:20:41.980
They're just still trying to do different things,

402
00:20:41.980 --> 00:20:42.990
and so we're not there yet,

403
00:20:42.990 --> 00:20:45.930
although the technology is certainly

404
00:20:45.930 --> 00:20:47.430
a step in the right direction.

405
00:20:51.560 --> 00:20:53.890
<v ->So to talk a little bit more about the Turing test.</v>

406
00:20:53.890 --> 00:20:55.690
It was one of the most important milestones

407
00:20:55.690 --> 00:20:57.630
in the research and development of AI,

408
00:20:57.630 --> 00:20:59.750
but now the focus has turned more

409
00:20:59.750 --> 00:21:03.000
to making human-computer interactions as smooth as possible.

410
00:21:03.000 --> 00:21:03.833
What are your thoughts on this,

411
00:21:03.833 --> 00:21:05.973
and where do you see this going in the future?

412
00:21:09.150 --> 00:21:14.150
<v ->Well, you know, it's not there yet in terms of, you know,</v>

413
00:21:16.960 --> 00:21:21.960
when you're waiting on hold or when you sign into your,

414
00:21:23.100 --> 00:21:24.520
I don't know, your bank or something,

415
00:21:24.520 --> 00:21:26.720
and they have their own version of Siri,

416
00:21:26.720 --> 00:21:29.983
and it tries to answer a question and can't, right?

417
00:21:32.400 --> 00:21:34.790
There are still a lot to do

418
00:21:34.790 --> 00:21:39.790
in terms of practical applications of chat bots

419
00:21:40.250 --> 00:21:42.690
and using these language models,

420
00:21:42.690 --> 00:21:46.660
much of that probably has to do with this mix of

421
00:21:49.590 --> 00:21:51.520
being able to respond fluently,

422
00:21:51.520 --> 00:21:54.430
and even contentfully by pure language model,

423
00:21:54.430 --> 00:21:56.270
brute force language model means,

424
00:21:56.270 --> 00:22:01.270
versus being able to understand, you know,

425
00:22:01.790 --> 00:22:04.180

what is the restricted context of

426
00:22:04.180 --> 00:22:06.580
what this person is looking for,

427
00:22:06.580 --> 00:22:10.610
you know, Siri and Alexa,

428
00:22:10.610 --> 00:22:15.610
and a lot of these digital assistants really lean, you know,

429
00:22:15.640 --> 00:22:20.640
as wonderful and amazing as neural architectures have been

430
00:22:21.639 --> 00:22:23.903
in increasing fluidity,

431
00:22:25.764 --> 00:22:28.670
the digital assistants mostly rely on dropdown menus

432
00:22:28.670 --> 00:22:32.070
and trying to classify what's the question of a list

433
00:22:32.070 --> 00:22:34.700
of questions that I'm going to be asked,

434
00:22:34.700 --> 00:22:39.700
and everything else is, I don't deal with that.

435
00:22:39.900 --> 00:22:42.060
My three-year-old daughter will sometimes

436
00:22:42.060 --> 00:22:45.870
get a hold of the phone and, you know, especially, you know,

437
00:22:45.870 --> 00:22:50.070
since we can't take her to play dates in the age of COVID

438
00:22:50.070 --> 00:22:51.500
has made very good friends with, you know,

439
00:22:51.500 --> 00:22:55.387
having long conversations with Siri (laughs)

440
00:22:55.387 --> 00:22:57.890
but they're very rogue responses,

441
00:22:57.890 --> 00:23:01.913
and you can start to see the patterns of those, nowadays,

442
00:23:02.890 --> 00:23:05.240
or you know, in this day and age, I should say.

443
00:23:10.100 --> 00:23:11.850
<v ->So artificial intelligence has made</v>

444
00:23:11.850 --> 00:23:15.400
a lot of things possible that previously seemed impossible.

445
00:23:15.400 --> 00:23:16.970
What things seem impossible now

446
00:23:16.970 --> 00:23:19.610
do you think might become a reality in the near future

447
00:23:19.610 --> 00:23:21.260
or just in the future in general?

448
00:23:30.760 --> 00:23:31.910
<v ->It's a good question.</v>

449
00:23:36.510 --> 00:23:40.397
One thing that I keep telling, I keep coming back

450
00:23:47.090 --> 00:23:50.100

to adding this, you know, what's the gap

451
00:23:50.100 --> 00:23:55.100
between language models and generalized AI,

452
00:23:55.770 --> 00:23:57.340
that's a big one.

453
00:23:57.340 --> 00:24:02.340
There are also gaps in, you know, every time we have

454
00:24:04.300 --> 00:24:09.100
a big breakthrough in reinforcement learning or in vision,

455
00:24:09.100 --> 00:24:12.950
or in, what I'm thinking about is image captioning,

456
00:24:12.950 --> 00:24:15.170
which really is a combination

457
00:24:15.170 --> 00:24:18.880
of language model or natural language processing,

458
00:24:18.880 --> 00:24:23.613
deep natural language processing and image processing.

459
00:24:24.660 --> 00:24:26.173
How do we connect these?

460
00:24:27.430 --> 00:24:31.973
So there's a paper by Bender that came out last year,

461
00:24:35.310 --> 00:24:38.770
the title was something like towards NLU,

462
00:24:38.770 --> 00:24:40.350
natural language understanding,

463
00:24:40.350 --> 00:24:43.250
or the mountain of NLU, something like that.

464
00:24:43.250 --> 00:24:47.690
And it was written from a linguistics background,

465
00:24:47.690 --> 00:24:50.550
or the perspective of a linguist I should say

466
00:24:50.550 --> 00:24:55.550
of how do we know when someone has

467
00:24:56.830 --> 00:24:58.660
an understanding of a term,

468
00:24:58.660 --> 00:24:59.890
which is a very different question

469
00:24:59.890 --> 00:25:01.900
from someone can use a term or someone,

470
00:25:01.900 --> 00:25:03.313
or maybe is a different question now,

471
00:25:03.313 --> 00:25:05.423
that's what's interesting,

472
00:25:07.790 --> 00:25:10.820
of a term that they've never, you know,

473
00:25:10.820 --> 00:25:15.820
you said a cat or a lunchbox, right?

474
00:25:16.470 --> 00:25:17.970
How do you know that, you know,

475
00:25:17.970 --> 00:25:19.330

it's completely ungrounded,

476
00:25:19.330 --> 00:25:21.430
this machine is you know, sitting on a desk,

477
00:25:21.430 --> 00:25:26.430
or a server center somewhere has never interacted

478
00:25:26.980 --> 00:25:31.160
with a lunchbox, but can talk about lunchboxes,

479
00:25:31.160 --> 00:25:33.360
and give the right context,

480
00:25:33.360 --> 00:25:37.580
for a lunchbox, could tell you what goes into a lunch box,

481
00:25:37.580 --> 00:25:38.880
you know, sandwiches,

482
00:25:38.880 --> 00:25:41.343
but not dump trucks, that sort of thing.

483
00:25:42.950 --> 00:25:47.380
But in a sense, it's ungrounded usage,

484
00:25:47.380 --> 00:25:48.620
even though it's perfectly contentful,

485
00:25:48.620 --> 00:25:51.520
and perfectly appropriate usage.

486
00:25:51.520 --> 00:25:56.520
And so one part of grounding that kind of machine learning

487
00:26:00.289 --> 00:26:02.100
is by adding to it,

488
00:26:02.100 --> 00:26:04.580
not just the context of using the words,

489
00:26:04.580 --> 00:26:06.180
and then when you use the words,

490
00:26:07.460 --> 00:26:11.120
in context of what other words, but also direct,

491
00:26:11.120 --> 00:26:12.200
what you might think of

492
00:26:12.200 --> 00:26:13.870
as the analog of direct experience, right?

493
00:26:13.870 --> 00:26:17.510
So being able to see a lunchbox,

494
00:26:17.510 --> 00:26:19.420
and identify images of lunchboxes,

495
00:26:19.420 --> 00:26:22.560
and create text like text captioning,

496
00:26:22.560 --> 00:26:26.960
which while something that, you know, where we are

497
00:26:26.960 --> 00:26:29.620
with text captioning at this point is,

498
00:26:29.620 --> 00:26:32.290
or captioning at this point

499
00:26:32.290 --> 00:26:37.290
is far beyond where eight years ago I thought we would be

500
00:26:37.290 --> 00:26:39.653

in turn of the new decade,

501
00:26:40.950 --> 00:26:45.950
it's still not the magnificent of you know, 99%,

502
00:26:46.090 --> 00:26:48.070
oh, this is always appropriate,

503
00:26:48.070 --> 00:26:51.623
at best, we are hitting up against the wall of,

504
00:26:52.534 --> 00:26:53.870
you know, well there's a lot of different ways

505
00:26:53.870 --> 00:26:55.794
you can caption something,

506
00:26:55.794 --> 00:26:57.400
and when you're working with,

507
00:26:57.400 --> 00:27:01.800
you know, machine translation, summarization, or text,

508
00:27:01.800 --> 00:27:03.130
anytime there's free form,

509
00:27:03.130 --> 00:27:04.720
knowing what the right answer is,

510
00:27:04.720 --> 00:27:07.410
when there's multiple ways of describing something,

511
00:27:07.410 --> 00:27:11.500
is an intrinsic barrier for machine learning,

512
00:27:11.500 --> 00:27:15.000
because you need to figure out how to update

513
00:27:15.000 --> 00:27:16.550
these models very quickly, and how to, you know,

514
00:27:16.550 --> 00:27:18.515
propagate the errors, you know,

515
00:27:18.515 --> 00:27:23.190
the grading correction through each of the parameters.

516
00:27:23.190 --> 00:27:25.160
And so having the right metrics,

517
00:27:25.160 --> 00:27:27.170
there's a lot of noise there, especially if you only have

518
00:27:27.170 --> 00:27:31.960
one or two labels of what the right text is.

519
00:27:31.960 --> 00:27:34.870
And so I think that's really where

520
00:27:34.870 --> 00:27:38.090
it would be interesting to see us move next is

521
00:27:38.090 --> 00:27:39.240
number one, you know,

522
00:27:39.240 --> 00:27:41.630
theoretically it seems it's very difficult

523
00:27:41.630 --> 00:27:44.340
to solve these free text problems.

524
00:27:44.340 --> 00:27:46.957
So that's one area I'd love to see improvement in,

525
00:27:46.957 --> 00:27:50.510

but also just seeing improvement in, you know,

526
00:27:50.510 --> 00:27:55.210
some of these hybrid contexts, where you're using, you know,

527
00:27:55.210 --> 00:27:59.530
vision and if not sound then at least text.

528
00:27:59.530 --> 00:28:01.790
And, you know, combining that with the use,

529
00:28:01.790 --> 00:28:03.940
you've seen that a little bit when in reinforcement learning

530
00:28:03.940 --> 00:28:08.930
where, you know, a lot of these word model techniques

531
00:28:08.930 --> 00:28:11.010
are focused on, you know, giving

532
00:28:11.010 --> 00:28:14.170
a reinforcement learning agent a model, a visual model,

533
00:28:14.170 --> 00:28:17.670
where they can observe the environment around them,

534
00:28:17.670 --> 00:28:20.320
and then compress that into a much more manageable space,

535
00:28:20.320 --> 00:28:23.040
and then start doing these high speed predictions

536
00:28:23.040 --> 00:28:24.187
about what's likely to happen,

537
00:28:24.187 --> 00:28:25.310
and what are you forecasting,

538
00:28:25.310 --> 00:28:27.120
if I take this action, then what will happen,

539
00:28:27.120 --> 00:28:30.170
and do this, what's called a technical concept

540
00:28:30.170 --> 00:28:34.840
of machine planning, but we haven't seen that,

541
00:28:34.840 --> 00:28:37.377
you know, we haven't seen all of these together yet,

542
00:28:37.377 --> 00:28:39.220
and that's really where, you know,

543
00:28:39.220 --> 00:28:42.210
I think it'd be interesting to see the industry grow

544
00:28:42.210 --> 00:28:43.043
in the future.

545
00:28:45.460 --> 00:28:46.714
<v ->It's definitely interesting to think about</v>

546
00:28:46.714 --> 00:28:48.397
what's gonna happen in the future.

547
00:28:48.397 --> 00:28:49.790
And along those same lines,

548
00:28:49.790 --> 00:28:51.830
there has been some concern and hype

549
00:28:51.830 --> 00:28:55.240
about AI capabilities and how it will impact humans.

550
00:28:55.240 --> 00:28:56.884

What are your thoughts on this?

551
00:28:56.884 --> 00:29:00.113
<v ->(chuckles) So when I was at Uber, you know,</v>

552
00:29:05.520 --> 00:29:06.960
we were working on self-driving,

553
00:29:06.960 --> 00:29:11.680
and versus, so that division, but, you know,

554
00:29:11.680 --> 00:29:16.120
I was sitting in the car with another Uber employee,

555
00:29:16.120 --> 00:29:19.650
I was driving, and we were chatting

556
00:29:19.650 --> 00:29:22.150
about what that's gonna mean.

557
00:29:22.150 --> 00:29:23.640
You know, he certainly wasn't worried,

558
00:29:23.640 --> 00:29:26.970
and I'm not too worried for, you know,

559
00:29:26.970 --> 00:29:31.263
robots taking the job of Uber drivers anytime soon.

560
00:29:33.130 --> 00:29:34.733
But, you know, at some point,

561
00:29:37.390 --> 00:29:42.120
he said the phrase, well, you know,

562
00:29:42.120 --> 00:29:45.090
one day the only job that's gonna be left is

563
00:29:46.080 --> 00:29:49.500
programming the robots that do all of our jobs.

564
00:29:49.500 --> 00:29:54.000
And that's certainly something that an approach I take

565
00:29:54.000 --> 00:29:57.950
when I think about my kids' education, you know,

566
00:29:57.950 --> 00:30:00.400
trying to, luckily there's Minecraft,

567
00:30:00.400 --> 00:30:03.120
so it's very easy to teach kids to get into coding,

568
00:30:03.120 --> 00:30:07.163
if they like to play video games,

569
00:30:08.910 --> 00:30:13.910
I don't know that it's going to be as big of an impact

570
00:30:16.280 --> 00:30:19.530
to like, you know, a lot of that impact is that

571
00:30:19.530 --> 00:30:22.560
we've already seen with, you know, manufacturing,

572
00:30:22.560 --> 00:30:25.610
you know, there are certain repetitive,

573
00:30:25.610 --> 00:30:30.610
but more, you know, more skilled and more complicated tasks

574
00:30:32.610 --> 00:30:35.440
like driving that may be replaced,

575
00:30:35.440 --> 00:30:38.290

but then there's going to be a whole host of techniques

576
00:30:38.290 --> 00:30:41.710
or of needs that are gonna be behind that

577
00:30:41.710 --> 00:30:45.090
involved with maintaining the technology

578
00:30:45.090 --> 00:30:46.470
that supports these.

579
00:30:46.470 --> 00:30:49.020
So certainly anyone taking this class

580
00:30:49.020 --> 00:30:52.880
is on the right side of things,

581
00:30:52.880 --> 00:30:55.835
getting the technical training that you need.

582
00:30:55.835 --> 00:31:00.835
I would not shy away from any program

583
00:31:01.820 --> 00:31:06.290
that makes sure that any education program

584
00:31:06.290 --> 00:31:08.810
that makes sure that coding is part of the standard part

585
00:31:08.810 --> 00:31:10.810
of the curriculum at some point,

586
00:31:10.810 --> 00:31:13.297
and certainly that's the future.

587
00:31:13.297 --> 00:31:16.900
You know, that was what Galvanize was really focused on,

588
00:31:16.900 --> 00:31:19.403
was in filling those gaps after the fact,

589
00:31:20.560 --> 00:31:22.050
you know, I don't know that bootcamps

590
00:31:22.050 --> 00:31:24.390
is really the right approach for that sort of thing,

591
00:31:24.390 --> 00:31:29.390
but Master's programs and undergraduate education,

592
00:31:29.560 --> 00:31:32.220
you know, Berkeley, it's become part of the

593
00:31:32.220 --> 00:31:34.397
standard core curriculum to do data science now,

594
00:31:34.397 --> 00:31:37.363
and I think that's incredibly far-sighted.

595
00:31:41.940 --> 00:31:43.190
<v ->Thanks, Mike, it was really great</v>

596
00:31:43.190 --> 00:31:44.023
to hear about your career,

597
00:31:44.023 --> 00:31:46.910
and your perspective on a lot of these questions.

598
00:31:46.910 --> 00:31:48.220
Just as a reminder to the audience,

599
00:31:48.220 --> 00:31:49.910
we will take some additional questions

600
00:31:49.910 --> 00:31:54.100

for both Mike and Dan at the end of the seminar.

601
00:31:54.100 --> 00:31:55.480
So next, we're gonna have a presentation

602
00:31:55.480 --> 00:31:56.853
by Dr. Dan Jacobson,

603
00:31:58.620 --> 00:32:00.620
and first we'll have some virtual applause for Mike,

604
00:32:00.620 --> 00:32:02.330
I know it was very hard in this virtual environment,

605
00:32:02.330 --> 00:32:03.950
but I know audience is clapping,

606
00:32:03.950 --> 00:32:07.620
and happy that you were able to be here for our interview.

607
00:32:07.620 --> 00:32:09.750
<v ->Thank you, thanks for having me.</v>

608
00:32:12.840 --> 00:32:16.206
<v ->So now I'd like to introduce Dr. Dan Jacobson,</v>

609
00:32:16.206 --> 00:32:17.039
he is the Chief Scientist

610
00:32:17.039 --> 00:32:18.950
for the computational systems biology

611
00:32:18.950 --> 00:32:20.670
at Oak Ridge National Laboratory,

612
00:32:20.670 --> 00:32:23.530
which is home to some of the world's largest supercomputers.

613
00:32:23.530 --> 00:32:24.760
Dan's research focuses on

614
00:32:24.760 --> 00:32:26.610
understanding the complex sets of interactions

615
00:32:26.610 --> 00:32:29.670
of molecules of all types across all omics layers,

616
00:32:29.670 --> 00:32:31.500
and cells that lead the phenotypes, traits,

617
00:32:31.500 --> 00:32:33.540
and disease states in organisms,

618
00:32:33.540 --> 00:32:34.790
and how all of that is conditional

619
00:32:34.790 --> 00:32:36.570
on the surrounding environment.

620
00:32:36.570 --> 00:32:38.390
His research team applies these approaches

621
00:32:38.390 --> 00:32:40.390
to grand challenges in bioenergy,

622
00:32:40.390 --> 00:32:43.450
sustainable agriculture, ecosystems and human health,

623
00:32:43.450 --> 00:32:45.640
and the intersections among these areas.

624
00:32:45.640 --> 00:32:47.330
Dan's lab was doing a range of research

625
00:32:47.330 --> 00:32:49.380

to address the COVID-19 pandemic,

626
00:32:49.380 --> 00:32:51.540
including studies of the molecular evolution

627
00:32:51.540 --> 00:32:53.750
of pathogenic elements of coronavirus,

628
00:32:53.750 --> 00:32:56.530
molecular mechanisms, ecogenesis,

629
00:32:56.530 --> 00:32:58.970
and identification of potential new therapies,

630
00:32:58.970 --> 00:33:00.940
environmental variables that affect

631
00:33:00.940 --> 00:33:02.410
COVID-19 disease outcomes,

632
00:33:02.410 --> 00:33:05.520
and the prediction and prevention of future zoonotic

633
00:33:05.520 --> 00:33:07.210
spillovers and pandemics.

634
00:33:07.210 --> 00:33:09.410
For this work, Dan has been awarded the 2021

635
00:33:09.410 --> 00:33:11.580
Secretary of Energy's Achievement Award,

636
00:33:11.580 --> 00:33:14.256
which is the highest award given by the USDOE,

637
00:33:14.256 --> 00:33:18.670
and the 2020 HPCwire top HPC enabled science award.

638
00:33:18.670 --> 00:33:22.320
Dan's team was the first group to break the exascale barrier

639
00:33:22.320 --> 00:33:24.970
and is happy to have done so for a biology project.

640
00:33:24.970 --> 00:33:26.880
At present, this calculation is the fastest

641
00:33:26.880 --> 00:33:29.630
scientific calculation ever done anywhere in the world.

642
00:33:30.740 --> 00:33:32.740
This project led to his team being awarded

643
00:33:32.740 --> 00:33:34.870
the 2018 Gordon Bell Prize,

644
00:33:34.870 --> 00:33:37.040
the first ever presented to biology.

645
00:33:37.040 --> 00:33:39.350
Dan's career as a computational systems biologist

646
00:33:39.350 --> 00:33:40.540
has included a leadership role

647
00:33:40.540 --> 00:33:44.650
in academic corporate NGO and national lab settings.

648
00:33:44.650 --> 00:33:46.060
His lab focuses on the development

649
00:33:46.060 --> 00:33:48.790
and subsequent application of mathematical, statistical,

650
00:33:48.790 --> 00:33:51.440

and computational methods to biological datasets,

651
00:33:51.440 --> 00:33:53.040
in order to yield new insights

652
00:33:53.040 --> 00:33:55.510
in the complex biological systems.

653
00:33:55.510 --> 00:33:59.993
As labs approaches include, sorry.

654
00:34:01.640 --> 00:34:03.590
The use of network theory and topology discovery

655
00:34:03.590 --> 00:34:05.410
clustering wavelet theory,

656
00:34:05.410 --> 00:34:08.230
AI explainable AI together with the traditional

657
00:34:08.230 --> 00:34:11.210
and more advanced supercomputing architectures.

658
00:34:11.210 --> 00:34:13.330
Areas of statistics of particular interest to his lab

659
00:34:13.330 --> 00:34:16.610
include the use of both frequent and invasion models,

660
00:34:16.610 --> 00:34:17.560
as well as the development

661
00:34:17.560 --> 00:34:20.430
of new methods of genome-wide epistasis studies.

662
00:34:20.430 --> 00:34:22.550
These mathematical and statistical models are applied

663
00:34:22.550 --> 00:34:23.720
to various populations,

664
00:34:23.720 --> 00:34:26.440
and metamorphic omics datasets individually,

665
00:34:26.440 --> 00:34:28.120
as well as in combination in an attempt

666
00:34:28.120 --> 00:34:30.380
to better understand the functional relationships,

667
00:34:30.380 --> 00:34:32.860
as well as biosynthesis signaling, transcriptional

668
00:34:32.860 --> 00:34:35.780
translational degradation, and kinetic regulatory networks

669
00:34:35.780 --> 00:34:38.960
at play in biological organisms and communities.

670
00:34:38.960 --> 00:34:41.480
His group takes a broad view of biological complexity

671
00:34:41.480 --> 00:34:42.610
and evolution that stretches

672
00:34:42.610 --> 00:34:45.790
from viruses to microbes, to plants, to humans.

673
00:34:45.790 --> 00:34:46.640
So let's welcome Dan,

674
00:34:46.640 --> 00:34:48.420
and I see that you have your slides ready,

675
00:34:48.420 --> 00:34:49.520

so take it away, Dan.

676
00:34:49.520 --> 00:34:51.120
I think it was fascinating to hear about your research,

677
00:34:51.120 --> 00:34:52.790
and congratulations on the new computation,

678
00:34:52.790 --> 00:34:54.280
I can't wait to hear details about that

679
00:34:54.280 --> 00:34:55.330
when they're available.

680
00:34:55.330 --> 00:34:57.650
So a round of applause for Dan,

681
00:34:57.650 --> 00:34:59.280
and so now we're gonna turn it over

682
00:34:59.280 --> 00:35:00.800
to Dr. Lindsey Friend,

683
00:35:00.800 --> 00:35:03.040
who's gonna be moderating questions from the chat.

684
00:35:03.040 --> 00:35:04.120
So if you have any questions,

685
00:35:04.120 --> 00:35:06.220
feel free to put them in the chat box,

686
00:35:06.220 --> 00:35:08.173
and I'll turn it over to Lindsey.

687
00:35:14.290 --> 00:35:15.253
<v ->Who is muted?</v>

688
00:35:16.659 --> 00:35:18.742
(laughs)

689
00:35:26.160 --> 00:35:28.977
It's not a Zoom session unless somebody is muted.

690
00:35:28.977 --> 00:35:30.340
<v ->Can you hear me now?</v>

691
00:35:30.340 --> 00:35:32.490
<v ->Yes.</v>
<v ->Oh, okay, thank you.</v>

692
00:35:32.490 --> 00:35:33.393
Sorry about that.

693
00:35:34.450 --> 00:35:35.460
Okay, great.

694
00:35:35.460 --> 00:35:37.810
Thank you again for your talks.

695
00:35:37.810 --> 00:35:39.560
Sorry for the technical problems.

696
00:35:39.560 --> 00:35:41.930
The first question I think is directed at Mike.

697
00:35:41.930 --> 00:35:44.280
He discussed gaps and being grounded,

698
00:35:44.280 --> 00:35:47.223
when you were discussing the language analysis portion.

699
00:35:48.270 --> 00:35:50.190
So the question is, can you give another kind of

700

00:35:50.190 --> 00:35:52.150
real-world example using those terms,

701
00:35:52.150 --> 00:35:55.063
and how they would apply to another scenario?

702
00:35:58.190 --> 00:36:02.970
<v ->For when something is grounded versus ungrounded</v>

703
00:36:02.970 --> 00:36:04.633
in machine usage?

704
00:36:07.580 --> 00:36:12.580
Well, so maybe I'll try to think of two cases.

705
00:36:12.900 --> 00:36:16.100
One is text in general, right?

706
00:36:16.100 --> 00:36:20.090
So I mentioned large language models,

707
00:36:20.090 --> 00:36:23.790
and how they may not interact with the things to which

708
00:36:23.790 --> 00:36:28.790
the text tokens, ostensibly refer, you know,

709
00:36:30.942 --> 00:36:34.540
that's been the case for not just

710
00:36:38.070 --> 00:36:39.490
more recent transformer models

711
00:36:39.490 --> 00:36:43.330
but also historical, just a pure word embeddings,

712
00:36:43.330 --> 00:36:48.330
certainly it's something that we've had to be very,

713
00:36:50.030 --> 00:36:52.020
you know, mindful of, and make sure

714
00:36:52.020 --> 00:36:56.560
that we are ruling out as appropriate is

715
00:36:56.560 --> 00:37:01.180
there are a lot of these models learn from texts

716
00:37:02.380 --> 00:37:07.380
that humans use and humans are filled with implicit

717
00:37:07.560 --> 00:37:11.110
and unfortunately too frequently

718
00:37:12.100 --> 00:37:15.690
just as often explicit biases.

719
00:37:15.690 --> 00:37:20.690
And this means that when a machine learns from written texts

720
00:37:21.570 --> 00:37:23.740
that it's filled with those biases,

721
00:37:23.740 --> 00:37:25.070
it also learns those biases.

722
00:37:25.070 --> 00:37:26.730
There are some famous examples of, you know,

723
00:37:26.730 --> 00:37:29.920
Twitter bots that were created by AIs,

724
00:37:29.920 --> 00:37:33.220
that became inexplicably racist.

725

00:37:33.220 --> 00:37:37.880
And that's something that you have to be very careful

726
00:37:37.880 --> 00:37:42.880
that you're not teaching a machine to follow these patterns.

727
00:37:43.230 --> 00:37:47.240
Now there's been some great work in how to screen out

728
00:37:47.240 --> 00:37:51.423
certain biases, but this is by far not a solved problem.

729
00:37:52.325 --> 00:37:57.180
And so there's this balance between,

730
00:37:57.180 --> 00:37:59.830
you know, this idea that meaning is use.

731
00:37:59.830 --> 00:38:02.900
And so if people use terms in a certain way,

732
00:38:02.900 --> 00:38:06.080
then that's what they mean by those terms.

733
00:38:06.080 --> 00:38:09.230
And meaning is intention.

734
00:38:09.230 --> 00:38:11.830
And finding a way to balance

735
00:38:11.830 --> 00:38:16.643
between those two poles becomes, especially a domain,

736
00:38:18.640 --> 00:38:22.070
when your you realize that the usage that people have

737
00:38:22.070 --> 00:38:25.630
is not always either the appropriate usage

738
00:38:25.630 --> 00:38:27.550
or the intended usage.

739
00:38:27.550 --> 00:38:30.967
And so that's one of these two reviews

740
00:38:32.940 --> 00:38:35.773
of what meaning is kind of peel apart.

741
00:38:37.610 --> 00:38:39.490
So that's maybe not another example,

742
00:38:39.490 --> 00:38:43.000
but another aspect of both grounding and usage

743
00:38:43.000 --> 00:38:45.430
of terms versus the actual meaning of those terms,

744
00:38:45.430 --> 00:38:47.903
and what those terms are representing.

745
00:38:49.572 --> 00:38:52.822
Ungrounded in terms of another context.

746
00:38:57.220 --> 00:39:02.220
So usually the concept of groundedness is tied to text.

747
00:39:03.660 --> 00:39:06.975
I'm tempted to say that, you know,

748
00:39:06.975 --> 00:39:11.975
in image classification, when you start to detect objects

749
00:39:16.900 --> 00:39:19.570
or misclassify, get false positives for objects

750

00:39:19.570 --> 00:39:23.100
based on spurious features,

751
00:39:23.100 --> 00:39:25.210
so in particular situations like

752
00:39:25.210 --> 00:39:30.210
where you have adversarial features,

753
00:39:30.580 --> 00:39:34.395
or adversarially trained examples,

754
00:39:34.395 --> 00:39:38.090
that can trick an estimator into thinking

755
00:39:38.090 --> 00:39:40.160
that something is say a cat when it's not a cat,

756
00:39:40.160 --> 00:39:41.883
or whatever the case may be.

757
00:39:43.276 --> 00:39:48.276
While it's not the traditional use of ungrounded language,

758
00:39:49.810 --> 00:39:53.433
it's certainly related to this idea that, you know,

759
00:39:54.820 --> 00:39:59.260
there's a difference between being able to recognize

760
00:39:59.260 --> 00:40:02.800
why something is worthy of a high estimation

761
00:40:02.800 --> 00:40:07.270
of a certain classification, as say a cat versus not.

762
00:40:07.270 --> 00:40:10.740
So those are both those answers are not other examples,

763
00:40:10.740 --> 00:40:13.109
they're just related examples,

764
00:40:13.109 --> 00:40:14.830
but hopefully it gets in the ballpark

765
00:40:14.830 --> 00:40:16.198
of answering the question.

766
00:40:16.198 --> 00:40:17.090
(electronic chime)

767
00:40:17.090 --> 00:40:19.030
<v ->Great, thank you very much.</v>

768
00:40:19.030 --> 00:40:22.130
Another question, I think that's for Dan,

769
00:40:22.130 --> 00:40:23.790
how can data science contribute

770
00:40:23.790 --> 00:40:26.650
to the enhancement of treatment specificity

771
00:40:26.650 --> 00:40:29.620
through individualized psychosocial interventions?

772
00:40:29.620 --> 00:40:31.850
So an intervention focus, Dan?

773
00:40:31.850 --> 00:40:34.460
<v ->And that's a great question that of course</v>

774
00:40:34.460 --> 00:40:36.250
is really what we're striving towards,

775

00:40:36.250 --> 00:40:40.640
and trying to understand the underlying

776
00:40:40.640 --> 00:40:44.060
genetic architectures, as well as influencing environment,

777
00:40:44.060 --> 00:40:47.170
to get us towards thinking about precision medicine,

778
00:40:47.170 --> 00:40:49.090
personalized medicine,

779
00:40:49.090 --> 00:40:51.630
finding the right intervention for the right person

780
00:40:51.630 --> 00:40:52.530
at the right time.

781
00:40:53.470 --> 00:40:56.220
Right now is we're, I'm sure we're all aware

782
00:40:56.220 --> 00:41:00.530
and prescriptions are really a trial and error game often

783
00:41:00.530 --> 00:41:03.040
trying different medic medications on people,

784
00:41:03.040 --> 00:41:04.990
and hoping that they're in the part of the population

785
00:41:04.990 --> 00:41:05.890
that will work on.

786
00:41:07.410 --> 00:41:10.080
The long-term goal is to really better understand

787
00:41:10.080 --> 00:41:12.530
the systems biology and increasingly

788
00:41:12.530 --> 00:41:16.150
the psychosocial elements of environment,

789
00:41:16.150 --> 00:41:20.060
and current and fire stress and exposure

790
00:41:20.060 --> 00:41:22.380
that's gonna lead to that phenotypic outcome.

791
00:41:22.380 --> 00:41:25.780
And then to be able to deliver the right treatment

792
00:41:25.780 --> 00:41:27.150
whether it's pharmaceutical

793
00:41:28.210 --> 00:41:33.210
or other other psychological interventions

794
00:41:33.330 --> 00:41:35.670
that's gonna help that patient.

795
00:41:35.670 --> 00:41:38.320
That's a long-term goal.

796
00:41:38.320 --> 00:41:42.130
I think there's a long ways to go, but trying to gain

797
00:41:42.130 --> 00:41:44.850
this fundamental mechanistic understanding,

798
00:41:44.850 --> 00:41:48.060
and the heterogeneity of that across the population,

799
00:41:48.060 --> 00:41:51.120
there's not one size fits all that you get the same,

800

00:41:51.120 --> 00:41:53.752
you can get the same phenotypic output

801
00:41:53.752 --> 00:41:57.800
from a range of different underlying architectures.

802
00:41:57.800 --> 00:42:00.060
We have a genetic omic or an environment,

803
00:42:00.060 --> 00:42:04.350
that it's that combination of different alleles,

804
00:42:04.350 --> 00:42:06.180
combination of different environments

805
00:42:06.180 --> 00:42:08.380
with those alleles that leads to an outcome.

806
00:42:09.250 --> 00:42:12.150
We tend to classify disease

807
00:42:12.150 --> 00:42:14.120
in these sort of very broad categories,

808
00:42:14.120 --> 00:42:17.280
I mean, in ICD-9 or ICD-10 code

809
00:42:17.280 --> 00:42:19.950
often is not representing the true underlying biology.

810
00:42:19.950 --> 00:42:21.550
It's a clinical pigeonhole

811
00:42:21.550 --> 00:42:25.280
that's convenient for billing, convenient for record keeping

812
00:42:25.280 --> 00:42:28.570
but it's not trying to capture all the underlying biology.

813
00:42:28.570 --> 00:42:31.040
So we're trying to sort of shine the light

814
00:42:31.040 --> 00:42:33.640
and find all the different types of biology

815
00:42:33.640 --> 00:42:35.690
that can lead to that diagnosis,

816
00:42:35.690 --> 00:42:38.483
and ideally then in the long-term,

817
00:42:40.610 --> 00:42:43.240
the vision of course is to help clinicians then

818
00:42:43.240 --> 00:42:44.390
choose the right therapy

819
00:42:44.390 --> 00:42:46.880
for the right person under the right conditions.

820
00:42:46.880 --> 00:42:49.457
<v ->I actually had a related question to that, Dan,</v>

821
00:42:49.457 --> 00:42:51.370
and that is,

822
00:42:51.370 --> 00:42:56.370
if you had the data available for drugs' side effects,

823
00:42:56.950 --> 00:42:58.950
how effective might your approaches be

824
00:42:58.950 --> 00:43:02.823
to understanding the unknown biology underlying those?

825

00:43:03.970 --> 00:43:08.970
<v ->Funny you should ask that, we're very much engaged</v>

826
00:43:09.210 --> 00:43:13.420
in polypharmacy research, looking at drug interactions,

827
00:43:13.420 --> 00:43:16.780
both leveraging the known interactions,

828
00:43:16.780 --> 00:43:19.080
as well as discovering new interactions,

829
00:43:19.080 --> 00:43:21.500
as part of that sort of questions,

830
00:43:21.500 --> 00:43:25.160
concomitant with that comes with the unintended consequences

831
00:43:25.160 --> 00:43:28.220
the off-target effects of pharmaceuticals.

832
00:43:28.220 --> 00:43:32.357
And so we're looking at in building data sets, again,

833
00:43:34.130 --> 00:43:38.010
ranging across the clinical layers of information

834
00:43:38.010 --> 00:43:40.020
and co-morbidity information,

835
00:43:40.020 --> 00:43:43.870
and of course, prescription pharmacy fill

836
00:43:43.870 --> 00:43:46.900
and consumption information along with the molecular details

837
00:43:46.900 --> 00:43:49.170
of looking at molecular profiles of,

838
00:43:49.170 --> 00:43:53.070
omics profiles of how cells are responding

839
00:43:53.070 --> 00:43:54.083
to different drugs,

840
00:43:55.210 --> 00:43:57.160
and trying to integrate lots of different

841
00:43:57.160 --> 00:43:59.900
types of essays together to get exactly that question,

842
00:43:59.900 --> 00:44:03.650
what can we tell is

843
00:44:03.650 --> 00:44:06.880
from known and unknown interaction space,

844
00:44:06.880 --> 00:44:08.840
what is from un-targeted information,

845
00:44:08.840 --> 00:44:13.150
how can we build in the genomics and environment

846
00:44:13.150 --> 00:44:16.090
of the underlying patients to tease all that apart?

847
00:44:16.090 --> 00:44:18.630
And how can we bring in the structural biology component

848
00:44:18.630 --> 00:44:21.990
as well to once we have candidates for off-target effects,

849
00:44:21.990 --> 00:44:24.550
can we show them what those interactions are,

850

00:44:24.550 --> 00:44:27.260
and long-term goal again, is to avoid those in the future

851
00:44:27.260 --> 00:44:30.220
that you can make your therapies more and more specific

852
00:44:30.220 --> 00:44:31.860
and contextually specific

853
00:44:31.860 --> 00:44:34.123
to minimize the sort of off-target effects.

854
00:44:36.410 --> 00:44:38.843
<v ->Great, thank you for that perspective.</v>

855
00:44:38.843 --> 00:44:40.550
I think Wilson has a couple of questions,

856
00:44:40.550 --> 00:44:41.623
turn it over to him.

857
00:44:43.090 --> 00:44:44.343
<v ->Sure, thanks very much.</v>

858
00:44:45.340 --> 00:44:47.860
First off, I couldn't count the number of zeros

859
00:44:47.860 --> 00:44:52.860
in a Zetta up, it's larger than a quadrillion,

860
00:44:53.210 --> 00:44:55.930
and that's as much as I sort of stopped being able

861
00:44:55.930 --> 00:44:59.730
to even imagine, that will be wonderful

862
00:44:59.730 --> 00:45:02.760
to read the report about that.

863
00:45:02.760 --> 00:45:06.190
I was struck as a clinician by your foray

864
00:45:06.190 --> 00:45:10.440
into electronic health records, personalized medicine,

865
00:45:10.440 --> 00:45:13.420
and the morass that is represented

866
00:45:13.420 --> 00:45:17.480
by electronic health records, where your data may come from.

867
00:45:17.480 --> 00:45:20.830
I will say that in natural language processing,

868
00:45:20.830 --> 00:45:24.710
the fact that there's misinterpretation by a machine,

869
00:45:24.710 --> 00:45:27.460
isn't always that different from what humans do,

870
00:45:27.460 --> 00:45:29.500
people misinterpret written language

871
00:45:29.500 --> 00:45:31.660
or verbal language all the time.

872
00:45:31.660 --> 00:45:35.753
Try explaining sarcasm to someone who doesn't quite get it.

873
00:45:36.720 --> 00:45:38.450
That's an example of how something

874
00:45:38.450 --> 00:45:41.330
that can be amusing to one in one context

875

00:45:41.330 --> 00:45:44.033
can be quite mean and cutting in another context.

876
00:45:45.150 --> 00:45:46.190
Just as one.

877
00:45:46.190 --> 00:45:48.080
When we think about medical records, though,

878
00:45:48.080 --> 00:45:50.870
this is a huge gap for the addictions field,

879
00:45:50.870 --> 00:45:54.410
where substances are not routinely recorded

880
00:45:54.410 --> 00:45:56.320
in medical records.

881
00:45:56.320 --> 00:45:59.360
We're hoping to take advantage of sort of the free text

882
00:45:59.360 --> 00:46:02.040
fields to build some of this.

883
00:46:02.040 --> 00:46:03.440
But I'm curious,

884
00:46:03.440 --> 00:46:07.600
not just about sort of that patchy nature of the data,

885
00:46:07.600 --> 00:46:09.920
but the way that sometimes these things are missing

886
00:46:09.920 --> 00:46:12.297
in a systematic or biased way.

887
00:46:13.290 --> 00:46:15.460
Any thoughts about how we can address that,

888
00:46:15.460 --> 00:46:19.040
and how we can use AI and big data approaches

889
00:46:19.040 --> 00:46:23.423
to address these inherent limitations of the data systems?

890
00:46:24.838 --> 00:46:26.297
<v ->I couldn't agree more.</v>

891
00:46:26.297 --> 00:46:29.763
I mean, HR records are messy, right?

892
00:46:30.770 --> 00:46:34.700
And most of the structured data and the unstructured,

893
00:46:34.700 --> 00:46:38.150
the text data bring big challenges.

894
00:46:38.150 --> 00:46:41.200
We're fortunate to have the clinical records

895
00:46:41.200 --> 00:46:45.240
for 23 million patients going back about 20 years

896
00:46:45.240 --> 00:46:48.120
here at Oak Ridge as part of the VA collaboration.

897
00:46:48.120 --> 00:46:51.830
And so that gives us plenty large corpuses to,

898
00:46:51.830 --> 00:46:52.680
or is that corpi?

899
00:46:54.440 --> 00:46:58.805
I really have a purpose to learn on,

900

00:46:58.805 --> 00:47:01.420
and to learn on all the different layers of information.

901
00:47:01.420 --> 00:47:04.330
And so what we're finding is,

902
00:47:04.330 --> 00:47:07.070
if you look at each layer independently,

903
00:47:07.070 --> 00:47:10.230
yes there's all sorts of challenges with each layer,

904
00:47:10.230 --> 00:47:12.083
but as you start to combine them,

905
00:47:13.200 --> 00:47:17.330
and use these AI and its wonderful AI approaches together,

906
00:47:17.330 --> 00:47:20.490
they start to support each other so that when, are ICD-9

907
00:47:20.490 --> 00:47:22.550
and 10 codes useful?

908
00:47:22.550 --> 00:47:24.110
Yes, they're useful.

909
00:47:24.110 --> 00:47:24.943
Are they perfect?

910
00:47:24.943 --> 00:47:26.790
No, they're not.

911
00:47:26.790 --> 00:47:30.700
But if you combine them with lab values,

912
00:47:30.700 --> 00:47:34.370
if you combine them with prescription information,

913
00:47:34.370 --> 00:47:37.970
if you combine them with outpatient information,

914
00:47:37.970 --> 00:47:40.330
as you build all these layers together,

915
00:47:40.330 --> 00:47:41.980
they start to support each other.

916
00:47:41.980 --> 00:47:43.000
Is it perfect?

917
00:47:43.000 --> 00:47:45.110
Sure, no, it's not perfect.

918
00:47:45.110 --> 00:47:47.140
But we're starting to get better and better

919
00:47:47.140 --> 00:47:50.820
at defining phenotypes, not just from diagnosis codes,

920
00:47:50.820 --> 00:47:52.667
but from the whole body of information.

921
00:47:52.667 --> 00:47:55.030
And that's one of the hopes in substance abuse

922
00:47:56.220 --> 00:48:00.760
is that from a standpoint of a phenotype,

923
00:48:00.760 --> 00:48:02.950
can we learn about what's really predictive

924
00:48:02.950 --> 00:48:05.820
of that phenotype that's been sort of rigorously done

925

00:48:05.820 --> 00:48:07.720
by surveys and chart reviews.

926
00:48:07.720 --> 00:48:09.270
Can we learn the clinical information

927
00:48:09.270 --> 00:48:10.940
that's predictive of that,

928
00:48:10.940 --> 00:48:13.255
and build multi-variant proxy phenotypes

929
00:48:13.255 --> 00:48:17.710
to find those missing cases, and you're completely right.

930
00:48:17.710 --> 00:48:19.150
And it's very challenging,

931
00:48:19.150 --> 00:48:22.383
and substance abuse in your cases and your controls,

932
00:48:23.690 --> 00:48:25.660
making sure your controls aren't contaminated,

933
00:48:25.660 --> 00:48:28.520
and making sure your cases are truthful,

934
00:48:28.520 --> 00:48:31.140
are your controls exposed, controls or not,

935
00:48:31.140 --> 00:48:33.560
we can have all sorts of debates about all of those,

936
00:48:33.560 --> 00:48:35.400
but by taking the totality of the record,

937
00:48:35.400 --> 00:48:40.400
as well as PRO, patient reported outcomes,

938
00:48:40.530 --> 00:48:42.600
patient reported information,

939
00:48:42.600 --> 00:48:44.090
you start to fill in those gaps.

940
00:48:44.090 --> 00:48:46.410
And so we're taking those sorts of approaches

941
00:48:46.410 --> 00:48:48.960
and a range of different projects,

942
00:48:48.960 --> 00:48:51.480
and showing the benefit of doing that.

943
00:48:51.480 --> 00:48:56.113
So our phenotypes are getting more sophisticated,

944
00:48:56.113 --> 00:48:58.340
then as the phenotypes get more sophisticated,

945
00:48:58.340 --> 00:49:00.390
and closer to biology, the systems biology,

946
00:49:00.390 --> 00:49:02.300
we can do on them, gets better,

947
00:49:02.300 --> 00:49:06.540
but you're spot on it's a really challenging issue,

948
00:49:06.540 --> 00:49:10.150
but it's this taking the totality of the data together

949
00:49:10.150 --> 00:49:14.010
is our strategy of solving as much of that as we can.

950

00:49:14.010 --> 00:49:16.310
And then also filling in by what we can get

951
00:49:16.310 --> 00:49:17.260
from patients now,

952
00:49:17.260 --> 00:49:21.320
and with apps and increasingly interest in patients

953
00:49:21.320 --> 00:49:24.270
in their own treatment and research,

954
00:49:24.270 --> 00:49:26.510
that's more and more doable.

955
00:49:26.510 --> 00:49:29.420
And other projects we're talking about,

956
00:49:29.420 --> 00:49:32.350
for a neuro-psychological condition, potentially,

957
00:49:32.350 --> 00:49:35.480
now cell phenotyping PRO-based for, you know,

958
00:49:35.480 --> 00:49:37.023
maybe millions of people.

959
00:49:37.930 --> 00:49:41.840
That's a huge distributed app-driven cohort,

960
00:49:41.840 --> 00:49:44.480
and then you filter down from that

961
00:49:44.480 --> 00:49:47.740
to folks that you wanna do really deep phenotyping on.

962
00:49:47.740 --> 00:49:51.070
And then you combine with that information

963
00:49:51.070 --> 00:49:55.130
from animal models to help you explore that biology as well,

964
00:49:55.130 --> 00:49:57.590
and in ways that you can't do in humans.

965
00:49:57.590 --> 00:50:01.440
But yeah, the clinical records are super challenging,

966
00:50:01.440 --> 00:50:03.470
but it's the holistic view is

967
00:50:03.470 --> 00:50:04.800
where we're starting to get there,

968
00:50:04.800 --> 00:50:06.500
and seeing some really cool stuff.

969
00:50:07.610 --> 00:50:08.443
<v ->Thank you.</v>

970
00:50:09.590 --> 00:50:10.890
Mike, I had a question for you

971
00:50:10.890 --> 00:50:12.450
about natural language processing

972
00:50:12.450 --> 00:50:15.180
in terms of, from what I'm familiar with,

973
00:50:15.180 --> 00:50:17.120
most of it's done with English language.

974
00:50:17.120 --> 00:50:20.810
Any advantages of using other languages for some of this?

975

00:50:20.810 --> 00:50:23.540
Is there more, I don't know enough about linguistics

976
00:50:23.540 --> 00:50:25.240
to know whether there might be some languages

977
00:50:25.240 --> 00:50:27.023
where it's more precise.

978
00:50:29.870 --> 00:50:32.140
<v ->So, I don't know about more precise.</v>

979
00:50:32.140 --> 00:50:34.900
Certainly there's a lot of efforts

980
00:50:34.900 --> 00:50:39.394
in working with other languages, as well.

981
00:50:39.394 --> 00:50:44.393
There are major, I think it's corpora

982
00:50:45.774 --> 00:50:50.363
in different languages,

983
00:50:51.910 --> 00:50:54.891
you know, some of this has because of machine translation

984
00:50:54.891 --> 00:50:59.830
is a very important, you know, application

985
00:50:59.830 --> 00:51:01.728
for natural language processing.

986
00:51:01.728 --> 00:51:03.950
I'm not super familiar,

987
00:51:03.950 --> 00:51:07.850
but I am aware that Facebook and Google

988
00:51:07.850 --> 00:51:11.130
are doing a lot of work in trying to come up

989
00:51:11.130 --> 00:51:15.330
with embedding models that are almost this like

990
00:51:15.330 --> 00:51:20.223
intro lingua, you know, particular language independence,

991
00:51:21.130 --> 00:51:23.800
so that they can expand to different countries,

992
00:51:23.800 --> 00:51:28.050
and they are very pragmatically going from

993
00:51:28.050 --> 00:51:31.400
most used to least used in that order,

994
00:51:31.400 --> 00:51:32.650
in terms of integrating them,

995
00:51:32.650 --> 00:51:37.320
and now, these are more focused on being able to embed text

996
00:51:38.200 --> 00:51:43.200
you know, in English and in Spanish and in Mandarin

997
00:51:45.040 --> 00:51:48.500
versus, and then seeing how all of those embeddings

998
00:51:48.500 --> 00:51:49.793
can be realigned.

999
00:51:51.060 --> 00:51:55.060
I did some some research where several years ago now

1000

00:51:55.060 --> 00:51:59.380
on what can be done with just pure, you know,

1001
00:51:59.380 --> 00:52:00.213
single word embeddings.

1002
00:52:00.213 --> 00:52:02.040
And if you think about what happens

1003
00:52:02.040 --> 00:52:04.160
when you embed a language, you know,

1004
00:52:04.160 --> 00:52:06.437
each word embedding is sort of fuzzy in the sense that

1005
00:52:06.437 --> 00:52:09.313
you know, there are multiple meanings of the term bank.

1006
00:52:10.330 --> 00:52:13.100
And so you end up kind of mode averaging, you know,

1007
00:52:13.100 --> 00:52:14.250
the thing where you put in money,

1008
00:52:14.250 --> 00:52:16.760
and the thing that you do when you go around a turn,

1009
00:52:16.760 --> 00:52:18.660
and the thing that you do,

1010
00:52:18.660 --> 00:52:20.653
or that you see at the edge of water.

1011
00:52:22.640 --> 00:52:23.750
But for the most part,

1012
00:52:23.750 --> 00:52:28.520
these language embeddings are kind of, you can realign them.

1013
00:52:28.520 --> 00:52:31.230
So you might take German and English,

1014
00:52:31.230 --> 00:52:34.910
and embed the entire constellation of vectors

1015
00:52:34.910 --> 00:52:36.130
that correspond to the words,

1016
00:52:36.130 --> 00:52:38.160
and then you can do with a rotation.

1017
00:52:38.160 --> 00:52:39.930
So we're talking about transformation,

1018
00:52:39.930 --> 00:52:43.370
you can align the bases of those embeddings,

1019
00:52:43.370 --> 00:52:46.110
and then measure how similar two languages are

1020
00:52:46.110 --> 00:52:48.350
to one another on a vocab level,

1021
00:52:48.350 --> 00:52:51.390
with all of that noisiness of multiple senses.

1022
00:52:51.390 --> 00:52:54.020
And you can even do

1023
00:52:54.020 --> 00:52:56.920
a little bit of a single value decomposition

1024
00:52:56.920 --> 00:52:58.950
of that linear transformation between the two,

1025

00:52:58.950 --> 00:53:00.810
'cause you know,

1026
00:53:00.810 --> 00:53:02.900
the linear transformation is gonna be a square matrix,

1027
00:53:02.900 --> 00:53:05.540
and square matrices can always be separated out

1028
00:53:05.540 --> 00:53:06.850
into a rotation,

1029
00:53:06.850 --> 00:53:09.460
a scale rotation with singular value decomposition.

1030
00:53:09.460 --> 00:53:13.590
So by analyzing the diagonals on that scaling matrix,

1031
00:53:13.590 --> 00:53:15.600
you can actually see how close it is to identity,

1032
00:53:15.600 --> 00:53:17.400
which would be just a pure rotation.

1033
00:53:19.720 --> 00:53:20.553
<v ->Right.</v>

1034
00:53:20.553 --> 00:53:22.820
Thank you both for your thoughtful answers.

1035
00:53:22.820 --> 00:53:23.653
I see that we're at time,

1036
00:53:23.653 --> 00:53:25.843
so I'm gonna turn the mic back over to Susan.

1037
00:53:27.620 --> 00:53:28.830
<v ->Thank you, Lindsey.</v>

1038
00:53:28.830 --> 00:53:30.870
So I just wanna thank both Dr. Tamir

1039
00:53:30.870 --> 00:53:32.390
and Dr. Jacobson again,

1040
00:53:32.390 --> 00:53:34.420
with a big round of virtual applause.

1041
00:53:34.420 --> 00:53:36.010
So it's been really wonderful to hear about

1042
00:53:36.010 --> 00:53:37.860
both of their careers and their perspectives,

1043
00:53:37.860 --> 00:53:39.160
and to hear the advice they've given

1044
00:53:39.160 --> 00:53:41.090
to those considering this career path.

1045
00:53:41.090 --> 00:53:43.340
I also wanna thank the speakers throughout the series,

1046
00:53:43.340 --> 00:53:45.830
the organizers, the technical team, and the audience.

1047
00:53:45.830 --> 00:53:48.620
Thank you all for making this seminar a series of success.

1048
00:53:48.620 --> 00:53:51.210
We'd love to hear your feedback about the seminar series,

1049
00:53:51.210 --> 00:53:53.010
and whether there are any other types of seminars

1050

00:53:53.010 --> 00:53:55.150
you'd like to see about data science careers,

1051
00:53:55.150 --> 00:53:56.930
or just data science in general.

1052
00:53:56.930 --> 00:53:58.850
It's possible that we may continue this series

1053
00:53:58.850 --> 00:54:00.750
in the fall or next year.

1054
00:54:00.750 --> 00:54:02.140
And lastly, I just wanna acknowledge

1055
00:54:02.140 --> 00:54:04.530
that there are several fellowships and job opportunities

1056
00:54:04.530 --> 00:54:07.020
for data science at NIH that are available

1057
00:54:07.020 --> 00:54:09.160
to both students and professionals.

1058
00:54:09.160 --> 00:54:10.700
These opportunities are coordinated

1059
00:54:10.700 --> 00:54:13.150
by the NIH Office of Data Science Strategy,

1060
00:54:13.150 --> 00:54:14.860
which is doing a tremendous job of coordinating

1061
00:54:14.860 --> 00:54:17.870
and inspiring data science activities across NIH.

1062
00:54:17.870 --> 00:54:20.060
One program is the data scholars program,

1063
00:54:20.060 --> 00:54:22.610
which is geared towards experienced data scientists,

1064
00:54:22.610 --> 00:54:24.410
and I believe they're still taking applications

1065
00:54:24.410 --> 00:54:25.243
until April 9th,

1066
00:54:25.243 --> 00:54:28.170
so there's still some time if you're interested in applying.

1067
00:54:28.170 --> 00:54:30.330
There's also the Civic Digital Fellowship program,

1068
00:54:30.330 --> 00:54:33.530
which is geared towards undergraduate and graduate students.

1069
00:54:33.530 --> 00:54:35.930
And there's also a Graduate Data Science Program

1070
00:54:35.930 --> 00:54:37.580
for Master's level students.

1071
00:54:37.580 --> 00:54:38.413
So if you're interested,

1072
00:54:38.413 --> 00:54:39.790
I would urge you to check out the website

1073
00:54:39.790 --> 00:54:42.060
for additional details online,

1074
00:54:42.060 --> 00:54:43.380
and I just wanna thank everyone again,

1075

00:54:43.380 --> 00:54:45.000
and please feel free to contact me,

1076
00:54:45.000 --> 00:54:46.080
or any of the other organizers

1077
00:54:46.080 --> 00:54:47.970
with any questions or feedback.

1078
00:54:47.970 --> 00:54:49.123
So thanks everyone.