WEBVTT

1 00:00:00.080 --> 00:00:02.010 <v ->Today it is my honor to introduce,</v>

2 00:00:02.010 --> 00:00:04.309 Dr. Atul Deshpande.

3 00:00:04.309 --> 00:00:06.790 Dr. Deshpande is a postdoctoral researcher

4 00:00:06.790 --> 00:00:09.080 in the lab of Dr. Elana Fertig

5 00:00:09.080 --> 00:00:10.630 in the department of oncology,

6 00:00:10.630 --> 00:00:12.920 at Johns Hopkins University.

7 00:00:12.920 --> 00:00:14.730 He has a PhD in electrical engineering

8 00:00:14.730 --> 00:00:17.460 from the University of Wisconsin-Madison,

9 00:00:17.460 --> 00:00:18.380 and his interests include

10 00:00:18.380 --> 00:00:20.100 the use of time series analysis

11 00:00:20.100 --> 00:00:21.310 and spatial statistics

12 00:00:21.310 --> 00:00:23.920 for modeling biological processes.

13 00:00:23.920 --> 00:00:26.080 He's currently developing analysis techniques

14 00:00:26.080 --> 00:00:28.027 to use single cell and spacial multigenomics

15 00:00:28.027 --> 00:00:30.110 for the characterization of

16 00:00:30.110 --> 00:00:31.509 the tumor microenvironment

17 00:00:31.509 --> 00:00:34.340 and intracellular signaling networks.

18 00:00:34.340 --> 00:00:37.413 Welcome. (students applause)

19 00:00:40.090 --> 00:00:40.960 <v ->Well, thank you so much.</v>

20 00:00:40.960 --> 00:00:43.160 And once I figure out my...

21 00:00:48.010 --> 00:00:49.380 Where my PowerPoint window is,

22 00:00:49.380 --> 00:00:51.860 we can start in earnest.

23 00:00:51.860 --> 00:00:54.940 Okay, yeah, thank you for the kind introduction.

24 00:00:54.940 --> 00:00:56.830 So, I'm Atul Deshpande,

25 00:00:56.830 --> 00:01:00.850 and today the title of my talk is exploring time

26 00:01:00.850 --> 00:01:03.640 and space for identifying gene interactions

27 00:01:03.640 --> 00:01:05.340 using single cell transcriptomics.

28 00:01:06.630 --> 00:01:10.440 So, what do time and space mean

29 00:01:10.440 --> 00:01:12.649 in the context of this talk?

30 00:01:12.649 --> 00:01:14.820 So, they refer to recent technological advances

31 00:01:14.820 --> 00:01:17.070 and the algorithms, which are the foundation

32 00:01:17.070 --> 00:01:19.120 for the projects I will be talking about.

33 00:01:20.450 --> 00:01:23.540 And the first advance is the ability

34 00:01:23.540 --> 00:01:26.730 to measure gene expression in individual cells.

35 00:01:26.730 --> 00:01:28.790 This in turn inspired development

36 00:01:28.790 --> 00:01:31.740 of algorithms that ordered these cells along

37 00:01:31.740 --> 00:01:33.083 the biological trajectory.

38 00:01:34.060 --> 00:01:37.460 Using these algorithms, we can observe changes

39 00:01:37.460 --> 00:01:39.440 in gene expression in

40 00:01:39.440 --> 00:01:42.850 a pseudo temporal reference for pseudo time,

41 00:01:42.850 --> 00:01:44.870 which is a measure of the progress

42 00:01:44.870 --> 00:01:46.653 of the biological process.

43 00:01:47.880 --> 00:01:50.170 The second is a more recent ability

44 00:01:50.170 --> 00:01:51.500 to measure gene expression

45 00:01:51.500 --> 00:01:54.210 within the spatial context of the tissue.

46 00:01:54.210 --> 00:01:55.770 But this we can analyze changes

47 00:01:55.770 --> 00:01:56.840 in gene expression

48 00:01:58.300 --> 00:02:00.450 as cellular neighborhoods change,

49 00:02:00.450 --> 00:02:02.183 or as the tissue type changes.

50 00:02:06.620 --> 00:02:09.800 So, before single cell transcriptomics,

51 00:02:09.800 --> 00:02:11.550 we would usually get one measurement

52 00:02:11.550 --> 00:02:15.300 of gene expression from a collected sample.

53 00:02:15.300 --> 00:02:18.300 And this is now called

54 00:02:18.300 --> 00:02:22.450 bulk RNA-seq in retroactively.

55 00:02:22.450 --> 00:02:25.510 However, as this measurement would just be

56 00:02:25.510 --> 00:02:27.480 an average of the population of cells

57 00:02:27.480 --> 00:02:30.740 in the sample, and it would obscure information

58 00:02:30.740 --> 00:02:33.550 about the different cell types, or different

59 00:02:33.550 --> 00:02:35.093 cell states in the population.

60 00:02:36.060 --> 00:02:37.490 With single-cell RNA-seq,

61 00:02:37.490 --> 00:02:39.510 we can now measure gene expression

62 00:02:39.510 --> 00:02:41.390 in individual cells.

63 00:02:41.390 --> 00:02:43.180 Depending on technology, this can range

64 00:02:43.180 --> 00:02:46.320 from a few hundred cells up to hundreds

65 00:02:46.320 --> 00:02:48.690 of thousands of cells.

66 00:02:48.690 --> 00:02:51.000 And this allows us to observe

67 00:02:51.000 --> 00:02:54.680 the full heterogeneity of the cell population

68 00:02:56.000 --> 00:02:58.510 represented by gene expression.

69 00:02:58.510 --> 00:03:01.770 And using this high dimensional data

70 00:03:01.770 --> 00:03:03.240 that we now have,

71 00:03:03.240 --> 00:03:05.320 we can characterize different cell types

72 00:03:05.320 --> 00:03:10.320 and cell states as gene expression vectors.

73 00:03:11.670 --> 00:03:13.740 So, one drawback of this technique

74 00:03:13.740 --> 00:03:16.680 is the issue of technical dropouts.

75 00:03:16.680 --> 00:03:20.790 Now, this is characterized by observing,

76 00:03:20.790 --> 00:03:23.040 as in us observing a lot

77 00:03:23.040 --> 00:03:26.300 of false zeroes, or zero inflated measurements,

78 00:03:26.300 --> 00:03:29.050 because we are unable to reliably measure

79 00:03:29.050 --> 00:03:31.213 the low iron accounts in individual cells.

80 00:03:34.696 --> 00:03:36.529 Now, the first project

81 00:03:39.330 --> 00:03:41.520 that I will discuss uses

82 00:03:43.204 --> 00:03:46.430 a single cell RNA-seq technology,

83 00:03:46.430 --> 00:03:48.800 or as it's downstream of that.

84 00:03:48.800 --> 00:03:52.900 And it uses also downstream of algorithms,

85 00:03:52.900 --> 00:03:57.900 which order single cell data into trajectories,

86 00:03:57.970 --> 00:04:00.520 which represent the biology

87 00:04:00.520 --> 00:04:02.220 that they might be studying.

88 00:04:02.220 --> 00:04:04.280 For example, let's say if you are...

89 00:04:04.280 --> 00:04:08.580 You have a dataset, which corresponds

90 00:04:08.580 --> 00:04:10.570 to stem cell differentiation,

91 00:04:10.570 --> 00:04:13.480 there are probably now 70 different

92 00:04:15.366 --> 00:04:17.030 trajectory inference methods depending on what

93 00:04:17.030 --> 00:04:20.900 kind of datasets you are studying,

94 00:04:20.900 --> 00:04:23.090 what biology you want to study,

95 00:04:23.090 --> 00:04:24.920 how big the dataset is,

96 00:04:24.920 --> 00:04:27.520 or what the expected trajectory is

97 00:04:27.520 --> 00:04:30.470 of the biology that you're studying maybe.

98 00:04:30.470 --> 00:04:33.700 And they attempt to order these cells based

99 00:04:33.700 --> 00:04:36.500 on the expression of potentially

100 00:04:36.500 --> 00:04:40.390 a few key marker genes, or how, which genes

101 00:04:40.390 --> 00:04:42.870 are differentially expressed along

102 00:04:42.870 --> 00:04:44.363 the biological process.

103 00:04:45.630 --> 00:04:47.620 So, anytime you collect,

104 00:04:47.620 --> 00:04:51.200 let's say a single cell RNA-seq data,

105 00:04:51.200 --> 00:04:54.460 you would find a mix of cells,

106 00:04:54.460 --> 00:04:55.900 and that was the entire motivation

107 00:04:55.900 --> 00:04:57.180 for doing this.

108 00:04:57.180 --> 00:05:01.340 But that mix of cells would have

109 00:05:01.340 --> 00:05:03.440 a range of cell states,

110 00:05:03.440 --> 00:05:05.960 which could correspond to

111 00:05:06.860 --> 00:05:08.960 from the beginning of the biological process,

112 00:05:08.960 --> 00:05:11.610 to the very end of the biological process.

113 00:05:11.610 --> 00:05:14.480 And what these algorithms are trying to do

114 00:05:14.480 --> 00:05:18.050 is they're trying to fit these cells

115 00:05:18.050 --> 00:05:22.500 in their right place, in the biological process.

116 00:05:22.500 --> 00:05:25.490 And once we do that, we can actually observe

117 00:05:25.490 --> 00:05:30.250 the gene expression along this ordering.

118 00:05:30.250 --> 00:05:34.110 And a lot of these methods also assign

119 00:05:34.110 --> 00:05:35.450 a pseudo time to each cell,

120 00:05:35.450 --> 00:05:38.810 which tells you how far along in the biology

121 00:05:38.810 --> 00:05:43.460 they think, or they hypothesize that the cell is.

122 00:05:43.460 --> 00:05:44.660 And so, the question that we wanted

123 00:05:44.660 --> 00:05:49.660 to ask is given this pseudo temporal ordering

124 00:05:50.290 --> 00:05:52.660 of the cells, which gives us

125 00:05:53.560 --> 00:05:55.140 a gene expression dynamics

126 00:05:55.140 --> 00:05:57.970 in the pseudo temporal reference.

127 00:05:57.970 --> 00:06:01.760 Can we use these dynamics

128 00:06:02.730 --> 00:06:05.803 to infer gene regulatory networks?

129 00:06:06.980 --> 00:06:10.100 Or any directed networks from say,

130 00:06:10.100 --> 00:06:13.500 sets of genes to their targets.

131 00:06:13.500 --> 00:06:14.360 And the second question

132 00:06:14.360 --> 00:06:19.250 was whether the assigned pseudo time values help

133 00:06:19.250 --> 00:06:22.313 us in the network inference task.

134 00:06:25.370 --> 00:06:30.370 So, to make the, I guess,

135 00:06:31.470 --> 00:06:33.750 explanation more approachable,

136 00:06:33.750 --> 00:06:36.523 I will just use an example dataset.

137 00:06:37.380 --> 00:06:41.220 And as I explained, the concepts I've...

138 00:06:41.220 --> 00:06:43.540 We will just see what that means

139 00:06:43.540 --> 00:06:45.400 in terms of this dataset.

140 00:06:45.400 --> 00:06:49.650 So, this is a dataset from Semrau et al,

141 00:06:49.650 --> 00:06:52.530 and this is a single cell data

142 00:06:52.530 --> 00:06:56.730 from retinoic acid, driven differentiation.

143 00:06:56.730 --> 00:07:00.740 And in this mouse, embryonic stem cells

144 00:07:00.740 --> 00:07:02.456 differentiate into neuroectoderm

145 00:07:02.456 --> 00:07:05.570 and extraembryonic endoderm cells.

146 00:07:05.570 --> 00:07:09.790 Now the data as collected had nine samples,

147 00:07:09.790 --> 00:07:12.090 one before the differentiation starts

148 00:07:12.090 --> 00:07:15.180 and one after every six hours.

149 00:07:15.180 --> 00:07:18.650 So, you have data collected over 96 hours

150 00:07:18.650 --> 00:07:22.453 from nine samples, and each sample has 384 cells.

151 00:07:23.960 --> 00:07:26.530 So overall, I believe we have something

152 00:07:26.530 --> 00:07:28.750 like you can do the math.

153 00:07:28.750 --> 00:07:31.793 I guess, 2,600 cells or something like that.

154 00:07:33.230 --> 00:07:37.240 So, we chose to apply

155 00:07:37.240 --> 00:07:39.380 two trajectory inference methods to this.

156 00:07:39.380 --> 00:07:41.630 So, the first one is monocle 2,

157 00:07:41.630 --> 00:07:44.800 which is also called Monocle DDR tree, I believe.

158 00:07:44.800 --> 00:07:46.960 And the second one is PAGA Tree.

159 00:07:46.960 --> 00:07:50.400 So, both of these methods identify

160 00:07:50.400 --> 00:07:53.480 a bifurcating trajectory from these cells.

161 00:07:53.480 --> 00:07:56.170 And so, the first one is to the left

162 00:07:56.170 --> 00:08:00.570 where the embryonic stem cells are actually

163 00:08:00.570 --> 00:08:01.910 on the right of...

164 00:08:03.340 --> 00:08:07.400 I'm not sure if people can see my mouse pointer,

165 00:08:07.400 --> 00:08:09.300 but yeah, they're on the right of the trajectory.

166 00:08:09.300 --> 00:08:13.520 And then, towards the bottom left,

167 00:08:13.520 --> 00:08:15.640 you go into a neuroectoderm state

168 00:08:15.640 --> 00:08:16.660 and towards the...

169 00:08:18.770 --> 00:08:23.770 Right, top left, you go into an endoderm state.

170 00:08:24.390 --> 00:08:27.250 And on the right side, the way PAGA Tree

171 00:08:27.250 --> 00:08:29.880 infers trajectory is you have

172 00:08:29.880 --> 00:08:32.980 the embryonic stem cells on the top left.

173 00:08:32.980 --> 00:08:35.100 And then, it identifies

174 00:08:35.100 --> 00:08:39.100 a few more branches than Monocle does.

175 00:08:39.100 --> 00:08:40.230 But both of these

176 00:08:40.230 --> 00:08:43.030 identify branching trajectories.

177 00:08:43.030 --> 00:08:47.000 And in each case we selected

178 00:08:48.150 --> 00:08:49.210 the two branches,

179 00:08:49.210 --> 00:08:53.530 which corresponded to markers, which were,

180 00:08:53.530 --> 00:08:55.920 which ended up being high for neuroectoderm.

181 00:08:55.920 --> 00:08:59.670 So, the trajectories, the sub trajectories

182 00:08:59.670 --> 00:09:02.810 from each method that we've wanted to study

183 00:09:02.810 --> 00:09:07.420 was the embryonic stem cells to neuroectoderm,

184 00:09:08.480 --> 00:09:10.293 using these two methods.

185 00:09:11.370 --> 00:09:13.540 So, this as in, so we had...

186 00:09:13.540 --> 00:09:16.240 We have these two trajectory inference methods,

187 00:09:16.240 --> 00:09:18.380 which assigned their own pseudo times,

188 00:09:18.380 --> 00:09:22.500 and this is the pseudo temporal expression

189 00:09:22.500 --> 00:09:25.410 dynamics for the same gene.

190 00:09:25.410 --> 00:09:28.670 I did not mark which gene it was, but yeah,

191 00:09:28.670 --> 00:09:30.220 so this was for the same gene.

192 00:09:30.220 --> 00:09:32.760 And you can see that the dynamics

193 00:09:32.760 --> 00:09:35.290 that each of these trajectories gives

194 00:09:35.290 --> 00:09:36.680 us is different.

195 00:09:36.680 --> 00:09:39.690 First of all, the main branch,

196 00:09:39.690 --> 00:09:41.820 or sub part of the trajectory that

197 00:09:41.820 --> 00:09:43.960 we are considering has

198 00:09:43.960 --> 00:09:45.610 a different number of cells.

199 00:09:45.610 --> 00:09:47.940 And these cells may not necessarily be common

200 00:09:47.940 --> 00:09:48.800 to both end.

201 00:09:48.800 --> 00:09:49.810 There will be some which are common

202 00:09:49.810 --> 00:09:51.590 to both of these trajectories,

203 00:09:51.590 --> 00:09:54.240 but some others which are completely different.

204 00:09:54.240 --> 00:09:56.580 But also, that the cell ordering itself

205 00:09:56.580 --> 00:10:01.320 that each method based on whatever mathematics

206 00:10:01.320 --> 00:10:03.420 they use, or whatever algorithms they use,

207 00:10:04.600 --> 00:10:07.780 would differ between these two methods.

208 00:10:07.780 --> 00:10:11.840 So, as you see, Monocle has a higher expression

209 00:10:11.840 --> 00:10:13.620 much earlier in the pseudo time,

210 00:10:13.620 --> 00:10:17.820 as opposed to PAGA Tree, which has much later.

211 00:10:17.820 --> 00:10:20.020 And the pseudo times here,

212 00:10:20.020 --> 00:10:22.010 were not exactly 100, they're just nominalized

213 00:10:22.010 --> 00:10:25.470 to 100 just represent progress from 0%

214 00:10:25.470 --> 00:10:27.863 of the biology to 100% of the biology,

215 00:10:30.620 --> 00:10:32.973 or as inferred by that method.

216 00:10:33.810 --> 00:10:36.980 So, now what are the challenges associated

217 00:10:36.980 --> 00:10:39.400 with order single-cell data?

218 00:10:39.400 --> 00:10:42.840 So, the first one is that unlike say,

219 00:10:42.840 --> 00:10:46.580 stock data, or say weather data,

220 00:10:46.580 --> 00:10:49.130 or something like that, you don't necessarily

221 00:10:49.130 --> 00:10:53.630 have a uniform distribution of cells.

222 00:10:53.630 --> 00:10:56.330 And if you're going to do a time series analysis,

223 00:10:56.330 --> 00:10:57.480 that would mean that you do not

224 00:10:57.480 --> 00:10:59.820 have regularly spaced time series,

225 00:10:59.820 --> 00:11:00.653 but you actually

226 00:11:00.653 --> 00:11:03.020 have irregularly space time series.

227 00:11:03.020 --> 00:11:05.090 On top of that, the pseudo time values

228 00:11:05.090 --> 00:11:07.270 that are assigned to the cells

229 00:11:07.270 --> 00:11:10.133 and ordering stem cells is uncertain.

230 00:11:12.720 --> 00:11:16.720 Now, finally, we recall that we had the issue

231 00:11:16.720 --> 00:11:18.740 of zero inflated measurements,

232 00:11:18.740 --> 00:11:20.870 or false zeroes in the meter

233 00:11:20.870 --> 00:11:22.370 because of technical dropouts.

234 00:11:25.798 --> 00:11:28.720 So, the question is how to overcome all

235 00:11:28.720 --> 00:11:32.370 of these drawbacks

236 00:11:32.370 --> 00:11:33.720 to try and find

237 00:11:35.730 --> 00:11:39.293 networks from this time series data.

238 00:11:40.160 --> 00:11:43.310 So, the project that we had,

239 00:11:43.310 --> 00:11:45.000 it resulted in basically

240 00:11:45.000 --> 00:11:46.270 an algorithm called SINGE,

241 00:11:46.270 --> 00:11:48.050 which is single cell inference

242 00:11:48.050 --> 00:11:50.210 of networks from Granger ensembles.

243 00:11:50.210 --> 00:11:52.730 So, this was done at the Morgridge Institute

244 00:11:52.730 --> 00:11:55.160 for Research in Madison, Wisconsin.

245 00:11:55.160 --> 00:11:58.633 And these are my collaborators on this project.

246 00:12:00.770 --> 00:12:02.860 And let's see, okay.

247 00:12:02.860 --> 00:12:06.390 So, the main concept that we build on

248 00:12:06.390 --> 00:12:08.400 is basically the Granger causality test.

249 00:12:08.400 --> 00:12:11.603 It was introduced by Clive Granger in 1960s.

250 00:12:14.489 --> 00:12:15.550 And to give a very simple example

251 00:12:15.550 --> 00:12:17.330 of what it's trying to say is, let's say

252 00:12:17.330 --> 00:12:21.670 if you have two times series X and Y,

253 00:12:21.670 --> 00:12:23.970 now Granger causality tests, whether

254 00:12:25.940 --> 00:12:28.330 the prediction of current values of Y

255 00:12:28.330 --> 00:12:30.860 improves by using past values of X,

256 00:12:30.860 --> 00:12:32.703 in addition to past values of Y.

257 00:12:34.210 --> 00:12:35.870 And if that happens, then we say

258 00:12:35.870 --> 00:12:37.940 that X Granger causes Y.

259 00:12:37.940 --> 00:12:40.590 So, this is basically a lag regression

260 00:12:40.590 --> 00:12:41.890 between X and Y.

261 00:12:41.890 --> 00:12:43.860 So, this has had applications

262 00:12:43.860 --> 00:12:46.060 in econometrics and finance,

263 00:12:46.060 --> 00:12:47.170 and is also being used

264 00:12:47.170 --> 00:12:50.600 in computational neuroscience and biology,

265 00:12:50.600 --> 00:12:53.643 as noted in these examples here.

266 00:12:55.310 --> 00:12:57.650 Now, the multivariate Granger causality test

267 00:12:57.650 --> 00:13:00.290 can be thought of as setting up and solving

268 00:13:00.290 --> 00:13:02.320 a vector, or regression model,

269 00:13:02.320 --> 00:13:05.430 where you have say, P genes, T time points

270 00:13:05.430 --> 00:13:06.440 and L lags.

271 00:13:06.440 --> 00:13:09.073 Where L lags is telling you how many,

272 00:13:11.010 --> 00:13:13.760 say your relationships with the past expressions

273 00:13:13.760 --> 00:13:15.670 you're trying to model.

274 00:13:15.670 --> 00:13:17.150 And once you have that,

275 00:13:17.150 --> 00:13:21.200 you could think of solving this way,

276 00:13:21.200 --> 00:13:23.740 our model by just minimizing

277 00:13:23.740 --> 00:13:25.253 this objective function here.

278 00:13:26.610 --> 00:13:28.060 And that would give you, I guess,

279 00:13:28.060 --> 00:13:30.700 a few edges between the past values

280 00:13:30.700 --> 00:13:34.160 of all of the genes and your target gene.

281 00:13:34.160 --> 00:13:35.610 Okay, maybe I should have explained

282 00:13:35.610 --> 00:13:36.920 this figure first.

283 00:13:36.920 --> 00:13:39.250 So, you have all the regular,

284 00:13:39.250 --> 00:13:42.050 all the possible regulators of a gene,

285 00:13:42.050 --> 00:13:43.340 and then you have a target gene,

286 00:13:43.340 --> 00:13:44.790 and you're trying to identify

287 00:13:46.470 --> 00:13:48.840 what explains what past values

288 00:13:48.840 --> 00:13:51.400 of any of these genes explains

289 00:13:51.400 --> 00:13:53.263 the current values of the target gene.

290 00:13:54.580 --> 00:13:58.730 And if you wanted to have

291 00:13:58.730 --> 00:14:01.710 a sparse representation of this network,

292 00:14:01.710 --> 00:14:03.300 or have an...

293 00:14:03.300 --> 00:14:05.030 Count only a few of the edges,

294 00:14:05.030 --> 00:14:08.210 you would introduce this by CT parameter,

295 00:14:08.210 --> 00:14:11.960 which would ensure that the edges from say,

296 00:14:11.960 --> 00:14:14.590 all of these genes to your target

297 00:14:14.590 --> 00:14:15.840 are not numerous.

298 00:14:15.840 --> 00:14:18.453 And you can explain the biology in a few edges.

299 00:14:22.378 --> 00:14:25.543 Now, to counter the irregularity

300 00:14:26.590 --> 00:14:29.480 of the time series, we use

301 00:14:30.340 --> 00:14:33.220 an idea called Generalized Lasso Granger.

302 00:14:33.220 --> 00:14:36.452 So, what this does is,

303 00:14:36.452 --> 00:14:39.110 I'm not sure, maybe I have...

304 00:14:39.110 --> 00:14:43.530 Yeah, okay, so just to recall, right?

305 00:14:43.530 --> 00:14:45.630 So, you have a pseudo temporal data,

306 00:14:45.630 --> 00:14:48.140 which has irregular time series,

307 00:14:48.140 --> 00:14:50.680 and you have missing values,

308 00:14:50.680 --> 00:14:54.150 which show up as zeros here, right?

309 00:14:54.150 --> 00:14:59.150 So, we want to adapt the Lasso Granger test

310 00:15:00.280 --> 00:15:01.990 for irregular time series.

311 00:15:01.990 --> 00:15:04.740 So, what was previously,

312 00:15:04.740 --> 00:15:07.320 basically coefficients from older samples

313 00:15:07.320 --> 00:15:08.820 in regular time series,

314 00:15:08.820 --> 00:15:13.720 now becomes coefficients from just timestamps

315 00:15:14.880 --> 00:15:16.010 in the past.

316 00:15:16.010 --> 00:15:17.730 Because you might not necessarily have

317 00:15:17.730 --> 00:15:19.553 a sample at that point.

318 00:15:20.710 --> 00:15:25.357 Furthermore, we can rethink basically,

319 00:15:28.091 --> 00:15:32.780 the object to function as originally,

320 00:15:32.780 --> 00:15:34.890 if it was a dot predict between

321 00:15:34.890 --> 00:15:37.650 the coefficients and the values

322 00:15:37.650 --> 00:15:40.560 of the gene expression,

323 00:15:40.560 --> 00:15:45.200 we rethink that as a weighted dot predict,

324 00:15:45.200 --> 00:15:46.650 where basically we...

325 00:15:47.590 --> 00:15:48.840 And this is the description

326 00:15:48.840 --> 00:15:51.450 of the weighted dot predict, where you use

327 00:15:51.450 --> 00:15:55.720 a Gaussian kernel to weight the inputs

328 00:15:55.720 --> 00:15:58.540 pseudo product based on their proximity

329 00:15:58.540 --> 00:16:01.830 to the timestamps that you...

330 00:16:01.830 --> 00:16:04.260 That correspond to these coefficients.

331 00:16:04.260 --> 00:16:07.740 So, these ellipses here show kernels,

332 00:16:07.740 --> 00:16:09.600 I guess, they represent kernels.

333 00:16:09.600 --> 00:16:12.310 They don't necessarily stop at these bandwidths,

334 00:16:12.310 --> 00:16:13.210 but they just keep going

335 00:16:13.210 --> 00:16:15.740 because they're ghosting kernels.

336 00:16:15.740 --> 00:16:17.990 But these just represent the kernels,

337 00:16:17.990 --> 00:16:20.260 where basically, if you have

338 00:16:20.260 --> 00:16:22.490 a timestamp corresponding to coefficient

339 00:16:22.490 --> 00:16:25.480 and you have no sample at that timestamp,

340 00:16:25.480 --> 00:16:26.440 that doesn't necessarily mean

341 00:16:26.440 --> 00:16:30.880 that the input to the gene predict it is zero.

342 00:16:30.880 --> 00:16:32.960 So, basically what you would do is

343 00:16:32.960 --> 00:16:36.010 you would just look at a bin around

344 00:16:36.010 --> 00:16:41.010 that timestamp, and weight input from regulators,

345 00:16:42.180 --> 00:16:46.240 depending on their proximity to this timestamp.

346 00:16:46.240 --> 00:16:50.790 So, if the sample is exactly at

347 00:16:50.790 --> 00:16:52.110 the timestamp that you expect,

348 00:16:52.110 --> 00:16:54.400 you would rate it highly based

349 00:16:54.400 --> 00:16:56.440 on discussion kernel, and the farther

350 00:16:56.440 --> 00:16:58.350 you move away from the timestamp,

351 00:16:58.350 --> 00:17:01.170 the weaker the rate of

352 00:17:02.240 --> 00:17:05.360 that particular sample would be.

353 00:17:05.360 --> 00:17:06.900 So, what this helps us do

354 00:17:06.900 --> 00:17:10.160 is if there are say more than one cells

355 00:17:10.160 --> 00:17:13.870 in close proximity, it would take input

356 00:17:13.870 --> 00:17:15.400 from all of them.

357 00:17:15.400 --> 00:17:17.670 If there are no cells in the close proximity

358 00:17:17.670 --> 00:17:19.680 to at least take input from some cells,

359 00:17:19.680 --> 00:17:21.380 which are farther away, and so on.

360 00:17:24.510 --> 00:17:27.340 So, yeah, as in this works

361 00:17:27.340 --> 00:17:28.460 with irregular time series,

362 00:17:28.460 --> 00:17:30.370 because you don't necessarily have

363 00:17:30.370 --> 00:17:33.500 to expect samples in the past at the timestamps

364 00:17:33.500 --> 00:17:34.700 that you wanted them to.

365 00:17:36.480 --> 00:17:39.900 And yeah, I think we already discussed this.

366 00:17:39.900 --> 00:17:44.900 So, now, as in going back to the case for...

367 00:17:45.420 --> 00:17:48.210 So, we had these false zeroes, right?

368 00:17:48.210 --> 00:17:50.310 So now, because of this kernel method,

369 00:17:50.310 --> 00:17:54.010 we have an inherent imputation over missing data.

370 00:17:54.010 --> 00:17:56.423 So, now we get what we could think of as,

371 00:17:57.930 --> 00:18:00.400 instead of taking all of the zeros

372 00:18:00.400 --> 00:18:02.550 as they are at face value,

373 00:18:02.550 --> 00:18:04.290 we can treat them, or some of them

374 00:18:04.290 --> 00:18:09.260 as dropouts, as just missing data.

375 00:18:09.260 --> 00:18:11.230 And we just remove those samples now,

376 00:18:11.230 --> 00:18:12.640 because we can now work

377 00:18:12.640 --> 00:18:14.820 with irregular time series.

378 00:18:14.820 --> 00:18:17.170 And because of this kernel method,

379 00:18:17.170 --> 00:18:19.420 we can actually work with time signature,

380 00:18:19.420 --> 00:18:21.570 all uniquely irregular.

381 00:18:21.570 --> 00:18:22.890 We can work with...

382 00:18:24.420 --> 00:18:26.270 We can remove the zero valued samples

383 00:18:26.270 --> 00:18:29.860 and get a different, differently irregular

384 00:18:29.860 --> 00:18:31.853 time series for each of these genes.

385 00:18:32.790 --> 00:18:36.870 And so, such an action can probably

386 00:18:36.870 --> 00:18:39.750 be informed by imputation techniques like magic,

387 00:18:39.750 --> 00:18:41.910 which help you complete,

388 00:18:41.910 --> 00:18:43.730 or impute zeros in the dataset.

389 00:18:43.730 --> 00:18:45.820 So, instead of imputing the dataset,

390 00:18:45.820 --> 00:18:47.780 as you could just use its output

391 00:18:47.780 --> 00:18:51.110 to decide whether or not to remove the data from,

392 00:18:51.110 --> 00:18:55.943 or remove that zero from this input dataset.

393 00:18:58.140 --> 00:18:59.930 So, this is just an illustration

394 00:18:59.930 --> 00:19:04.330 of a single generalized Lasso Granger test.

395 00:19:04.330 --> 00:19:08.300 So, you have the POU5F1 gene, and it's basically,

396 00:19:08.300 --> 00:19:11.030 you see it's the cells corresponding

397 00:19:11.030 --> 00:19:15.860 to that, or other details expression

398 00:19:15.860 --> 00:19:17.770 along pseudo time.

399 00:19:17.770 --> 00:19:22.010 And what you also see is two trendlines

400 00:19:22.950 --> 00:19:27.000 predicted using a Lambda of 0.1,

401 00:19:27.000 --> 00:19:29.110 which is basically a sparsity constraint of 0.1.

402 00:19:29.110 --> 00:19:31.990 So, it would have fewer edges

403 00:19:31.990 --> 00:19:34.867 between the regulators and POU5F1.

404 00:19:35.940 --> 00:19:40.530 And then a Lambda of 0.02,

405 00:19:40.530 --> 00:19:42.640 which has far more regulators.

406 00:19:42.640 --> 00:19:45.570 And you can see that both of these predict

407 00:19:46.460 --> 00:19:49.350 the trends of POU5F1 when using

408 00:19:49.350 --> 00:19:50.943 the past values quite well.

409 00:19:53.970 --> 00:19:58.093 So, now that was just one GLG test.

410 00:19:59.120 --> 00:20:01.460 Now, what SINGE does, is it performs multiple

411 00:20:01.460 --> 00:20:04.310 such GLG tests where you sub-sample

412 00:20:04.310 --> 00:20:06.840 the time series different ways

413 00:20:06.840 --> 00:20:11.610 to get different irregulars time series again.

414 00:20:11.610 --> 00:20:14.110 And you also use diverse hyper-parameters

415 00:20:14.110 --> 00:20:16.990 to effectively using these two combinations,

416 00:20:16.990 --> 00:20:20.220 slice the cake multiple ways and trying

417 00:20:20.220 --> 00:20:21.720 to look at the data.

418 00:20:21.720 --> 00:20:22.890 So, the type of barometers

419 00:20:22.890 --> 00:20:25.210 that we use are Lambda, which determines

420 00:20:25.210 --> 00:20:28.650 the sparsity of the network that we get,

421 00:20:28.650 --> 00:20:30.830 or get into metrics that we get.

422 00:20:30.830 --> 00:20:35.830 And we have Delta T, which gives us

423 00:20:35.900 --> 00:20:39.530 a time resolution of the lags between say,

424 00:20:39.530 --> 00:20:41.280 the past regulators

425 00:20:41.280 --> 00:20:44.990 and the current target timestamps,

426 00:20:44.990 --> 00:20:47.360 and the number of likes that you have.

427 00:20:47.360 --> 00:20:51.170 So together, they will tell you how far behind

428 00:20:51.170 --> 00:20:53.550 in pseudo time should you be looking to try

429 00:20:53.550 --> 00:20:57.400 to predict the expression of the target.

430 00:20:57.400 --> 00:20:58.680 And finally, the kernel width,

431 00:20:58.680 --> 00:21:03.030 which tells how far, how wide the width should be

432 00:21:03.030 --> 00:21:06.813 around the timestamp that you are considering.

433 00:21:08.456 --> 00:21:09.289 Now, once we get

434 00:21:11.370 --> 00:21:13.340 adjacency matrices from all of these,

435 00:21:13.340 --> 00:21:16.860 we get, we considered them as partial networks,

436 00:21:16.860 --> 00:21:20.210 and we get ranked lists from each of them.

437 00:21:20.210 --> 00:21:22.020 And we aggregate these rank lists

438 00:21:22.020 --> 00:21:24.330 using a modified border count.

439 00:21:24.330 --> 00:21:25.390 So, border count is something

440 00:21:25.390 --> 00:21:28.806 which has been used in election.

441 00:21:28.806 --> 00:21:31.420 It's basically an election, I guess,

442 00:21:31.420 --> 00:21:33.630 result aggregating strategy,

443 00:21:33.630 --> 00:21:35.950 where if you have five candidates,

444 00:21:35.950 --> 00:21:39.190 you rank them from one to five,

445 00:21:39.190 --> 00:21:41.600 and then the person who has, I guess,

446 00:21:41.600 --> 00:21:44.360 the lowest number here over all

447 00:21:44.360 --> 00:21:46.630 of the people that voted,

448 00:21:46.630 --> 00:21:48.930 they would win the vote.

449 00:21:48.930 --> 00:21:51.507 So, the modified border width

450 00:21:51.507 --> 00:21:53.190 is basically the same concept,

451 00:21:53.190 --> 00:21:55.260 but the only change that we did

452 00:21:55.260 --> 00:22:00.260 was we wanted to place more weight

453 00:22:02.870 --> 00:22:05.180 to a ranking, which distinguishes

454 00:22:05.180 --> 00:22:09.580 between say a one, the first interaction

455 00:22:09.580 --> 00:22:12.310 we find with the 10th interaction we find.

456 00:22:12.310 --> 00:22:15.160 As opposed to say, the 10,000th interaction

457 00:22:15.160 --> 00:22:17.620 we find with the 10,010th interaction

458 00:22:17.620 --> 00:22:18.620 that we find.

459 00:22:18.620 --> 00:22:23.320 So, that's why the weighting before adding

460 00:22:23.320 --> 00:22:26.210 these border weights is one over N squared,

461 00:22:26.210 --> 00:22:29.243 as opposed to say, N here.

462 00:22:33.100 --> 00:22:36.460 So, yeah, once we aggregate this,

463 00:22:36.460 --> 00:22:39.310 we get a final rank list.

464 00:22:39.310 --> 00:22:43.200 And so, we had to do in for trajectories,

465 00:22:43.200 --> 00:22:45.770 we got gene dynamics from them,

466 00:22:45.770 --> 00:22:49.070 and now that results in two different networks.

467 00:22:49.070 --> 00:22:52.840 And there's just showing the top 100 edges

468 00:22:52.840 --> 00:22:54.910 from Monocle 2 and PAGA Tree.

469 00:22:54.910 --> 00:22:56.320 Now, you can obviously see

470 00:22:56.320 --> 00:22:58.600 that they look very different.

471 00:22:58.600 --> 00:23:01.690 Some of the edges I think, are common,

472 00:23:01.690 --> 00:23:04.660 but they can be very, very different.

473 00:23:04.660 --> 00:23:07.710 So, now the question is,

474 00:23:07.710 --> 00:23:11.520 which of these is right, or better?

475 00:23:11.520 --> 00:23:14.070 So, for that we would have

476 00:23:14.070 --> 00:23:15.110 to first think of, okay,

477 00:23:15.110 --> 00:23:16.940 how do we evaluate this?

478 00:23:16.940 --> 00:23:19.940 So, one way to evaluate that would be

479 00:23:19.940 --> 00:23:23.570 to do a precision recall evaluation.

480 00:23:23.570 --> 00:23:25.370 So, let's say we have this rank list

481 00:23:25.370 --> 00:23:27.790 of candidate gene interactions that we just got

482 00:23:27.790 --> 00:23:30.680 from SINGE and a gold standard,

483 00:23:30.680 --> 00:23:31.780 which knows the truth.

484 00:23:32.680 --> 00:23:34.390 As we go down this rank list,

485 00:23:34.390 --> 00:23:36.870 the precision metric tells us

486 00:23:36.870 --> 00:23:38.300 what fraction of the prediction

487 00:23:38.300 --> 00:23:40.370 so far have been correct.

488 00:23:40.370 --> 00:23:41.710 And the recall metric tells us

489 00:23:41.710 --> 00:23:44.320 how many of the total interactions

490 00:23:44.320 --> 00:23:46.110 in the gold standard, which were correct

491 00:23:46.110 --> 00:23:47.433 have so far been covered.

492 00:23:48.350 --> 00:23:50.880 So, the figure on the right shows

493 00:23:50.880 --> 00:23:53.940 a precision recall curve for two rank lists.

494 00:23:53.940 --> 00:23:56.240 The ideal precision recall curve

495 00:23:56.240 --> 00:23:58.220 would place all the edges in the gold standard

496 00:23:58.220 --> 00:23:59.053 at the top of the list.

497 00:23:59.053 --> 00:24:03.560 So, that's the dotted line that you see here,

498 00:24:03.560 --> 00:24:06.040 and the area under that precision

499 00:24:06.040 --> 00:24:08.530 we call curve (mumbles) blue one.

500 00:24:08.530 --> 00:24:12.970 A random list in expectation would be flat.

501 00:24:12.970 --> 00:24:14.940 So, and it would have a precision

502 00:24:14.940 --> 00:24:17.930 recall curve, and the area under

503 00:24:17.930 --> 00:24:20.260 that curve would be 0.5.

504 00:24:20.260 --> 00:24:24.277 and here, I guess, to make belief orderings.

505 00:24:27.090 --> 00:24:29.310 And in this example, we can see

506 00:24:29.310 --> 00:24:34.063 that the precision we call curve of A,

507 00:24:35.150 --> 00:24:39.750 which I guess, the predictor A is better

508 00:24:39.750 --> 00:24:44.750 because it starts off with having more ones,

509 00:24:44.880 --> 00:24:47.930 or as in a high precision, and then falls

510 00:24:47.930 --> 00:24:49.990 as opposed to B, which rises

511 00:24:49.990 --> 00:24:51.053 from a low precision.

512 00:24:51.053 --> 00:24:54.600 What it means that A gets more hits

513 00:24:54.600 --> 00:24:56.860 in the top of its list as opposed to B,

514 00:24:56.860 --> 00:24:58.030 and so on.

515 00:24:58.030 --> 00:25:01.370 And so, one way to also evaluate

516 00:25:01.370 --> 00:25:03.180 these position we call curves is to just look

517 00:25:03.180 --> 00:25:05.750 at the area under the curve, which is so A here

518 00:25:05.750 --> 00:25:07.490 is 0.7 and B's 0.52.

519 00:25:07.490 --> 00:25:09.910 And that tells us that on an average

520 00:25:09.910 --> 00:25:14.910 A ranks edges better as opposed to B.

521 00:25:16.320 --> 00:25:19.280 Now, we would like to use near this,

522 00:25:19.280 --> 00:25:22.350 and the question is what could we use as

523 00:25:22.350 --> 00:25:23.183 a gold standard?

524 00:25:24.030 --> 00:25:25.500 Now, this is real biological data

525 00:25:25.500 --> 00:25:28.570 that we are using, and for that,

526 00:25:28.570 --> 00:25:31.670 we would also need to look into

527 00:25:31.670 --> 00:25:35.400 the literature to find validation.

528 00:25:35.400 --> 00:25:37.460 So, one good source of information

529 00:25:37.460 --> 00:25:39.310 is the escape database curated

530 00:25:39.310 --> 00:25:41.010 by the Ma'ayan lab.

531 00:25:41.010 --> 00:25:44.150 And this database includes the results

532 00:25:44.150 --> 00:25:46.710 of loss of function and gain of experiments

533 00:25:46.710 --> 00:25:48.640 done on genes, and also

534 00:25:48.640 --> 00:25:50.010 and also ChIP-seq experiments,

535 00:25:50.010 --> 00:25:51.700 which identify binding sites

536 00:25:51.700 --> 00:25:53.193 of transcription factors.

537 00:25:54.330 --> 00:25:57.940 Now, the problem being that even this database

538 00:25:57.940 --> 00:26:00.970 is incomplete because the gaps

539 00:26:00.970 --> 00:26:03.980 in biological knowledge remain and doesn't,

540 00:26:03.980 --> 00:26:05.530 I guess over the time, over time,

541 00:26:05.530 --> 00:26:08.630 it would be completed, filled more and more.

542 00:26:08.630 --> 00:26:12.020 But when we were doing this evaluation,

543 00:26:12.020 --> 00:26:14.330 we had to deal with what was effectively

544 00:26:14.330 --> 00:26:15.500 a partial gold standard,

545 00:26:15.500 --> 00:26:17.760 or an incomplete gold standard.

546 00:26:17.760 --> 00:26:20.290 So, the evaluation that we did was not

547 00:26:20.290 --> 00:26:22.920 for all of the genes in the dataset,

548 00:26:22.920 --> 00:26:26.453 but only a fraction of the genes.

549 00:26:28.210 --> 00:26:32.940 So, we had these two methods

550 00:26:32.940 --> 00:26:36.470 and two pseudo times, which we got from that.

551 00:26:36.470 --> 00:26:37.790 So, what we wanted, what we did

552 00:26:37.790 --> 00:26:41.710 is we compared the performance of SINGE

553 00:26:42.740 --> 00:26:46.200 using say, Monocle 2 and the pseudo time,

554 00:26:46.200 --> 00:26:48.940 as well as Monocle 2 with only the ordering.

555 00:26:48.940 --> 00:26:50.290 And some of the least PAGA Tree

556 00:26:50.290 --> 00:26:51.690 fed the pseudo time and PAGA Tree

557 00:26:51.690 --> 00:26:52.840 with only the ordering.

558 00:26:53.710 --> 00:26:57.650 And so, this is how the precision recall curves

559 00:26:57.650 --> 00:27:01.060 of these four methods look.

560 00:27:01.060 --> 00:27:04.420 So, we look at the average precision,

561 00:27:04.420 --> 00:27:06.240 which is the same thing as the area under

562 00:27:06.240 --> 00:27:07.890 the precision recall curve.

563 00:27:07.890 --> 00:27:09.783 And we also look at the average precision

564 00:27:09.783 --> 00:27:13.640 in the early part of the precision recall curve.

565 00:27:13.640 --> 00:27:16.313 And the point for that being that,

566 00:27:17.960 --> 00:27:20.650 in say, a usual workflow,

567 00:27:20.650 --> 00:27:23.850 you would have a combination method,

568 00:27:23.850 --> 00:27:27.713 which would point to some important edges,

569 00:27:28.620 --> 00:27:31.490 and then, you would potentially tell

570 00:27:31.490 --> 00:27:33.680 a collaborator to try

571 00:27:33.680 --> 00:27:35.820 and experimentally validate that.

572 00:27:35.820 --> 00:27:38.190 And in that sense, you would be giving

573 00:27:38.190 --> 00:27:40.110 them results from the top of your list,

574 00:27:40.110 --> 00:27:43.100 as opposed to trying to tell how well

575 00:27:43.100 --> 00:27:45.300 the 10,000th edge in the list

576 00:27:45.300 --> 00:27:46.673 is placed in the rankings.

577 00:27:47.580 --> 00:27:49.820 So, with that in mind, we also look

578 00:27:49.820 --> 00:27:52.780 at what's the average early precision

579 00:27:52.780 --> 00:27:54.093 of these curves.

580 00:27:55.030 --> 00:27:58.540 And for that, we basically say what happened,

581 00:27:58.540 --> 00:28:03.057 as to what extent is the precision maintained

582 00:28:04.150 --> 00:28:06.380 until 10% of the genes

583 00:28:06.380 --> 00:28:08.340 and the gold standard are...

584 00:28:08.340 --> 00:28:09.700 Or interactions with the gold standard

585 00:28:09.700 --> 00:28:14.550 are regarded in the list that we have.

586 00:28:14.550 --> 00:28:17.760 So, the figure to the right shows

587 00:28:17.760 --> 00:28:19.900 a scatterplot of these, the average precision

588 00:28:19.900 --> 00:28:21.080 and the average early precision

589 00:28:21.080 --> 00:28:24.670 for these four methods, for these four options.

590 00:28:24.670 --> 00:28:27.420 And what we see is that the...

591 00:28:27.420 --> 00:28:29.380 The best performing combination

592 00:28:29.380 --> 00:28:31.170 is using Monocle's ordering,

593 00:28:31.170 --> 00:28:35.540 but not its pseudo time, and Monocle applying

594 00:28:35.540 --> 00:28:37.610 the pseudo time that it order,

595 00:28:37.610 --> 00:28:41.050 that it assigns to the cells,

596 00:28:41.050 --> 00:28:44.213 actually degrades the performance quite a bit.

597 00:28:45.670 --> 00:28:48.950 And both of the PAGA Tree options

598 00:28:48.950 --> 00:28:50.600 with, or without pseudo time,

599 00:28:50.600 --> 00:28:52.350 are in between these.

600 00:28:52.350 --> 00:28:55.680 So, now why would this happen?

601 00:28:55.680 --> 00:28:57.100 For example, and let's take

602 00:28:57.100 --> 00:28:58.420 an extreme case, right?

603 00:28:58.420 --> 00:29:01.630 And okay, before that, there's not necessarily

604 00:29:04.060 --> 00:29:05.680 something that's wrong with Monocle,

605 00:29:05.680 --> 00:29:08.960 but it's basically that for this dataset,

606 00:29:08.960 --> 00:29:11.530 in this instance, the pseudo time values

607 00:29:11.530 --> 00:29:14.500 did not necessarily make a lot of sense.

608 00:29:14.500 --> 00:29:17.170 So, let's say you have perfectly ordered cells.

609 00:29:17.170 --> 00:29:18.890 And for the first half of the cells,

610 00:29:18.890 --> 00:29:22.110 you just assign a value very close

611 00:29:22.110 --> 00:29:23.180 to zero and the second half,

612 00:29:23.180 --> 00:29:25.570 you assign a value very close to one.

613 00:29:25.570 --> 00:29:27.470 So, even though the ordering of the cells

614 00:29:27.470 --> 00:29:31.170 was quite nice and reliable, just because

615 00:29:31.170 --> 00:29:33.810 we ended up assigning a value

616 00:29:33.810 --> 00:29:35.750 to the pseudo times, often times,

617 00:29:35.750 --> 00:29:38.090 which is completely unrealistic.

618 00:29:38.090 --> 00:29:40.970 We might end up losing

619 00:29:40.970 --> 00:29:41.950 a lot of information

620 00:29:41.950 --> 00:29:44.320 that we otherwise had in the dataset,

621 00:29:44.320 --> 00:29:45.270 or in the ordering.

622 00:29:48.920 --> 00:29:52.880 So, yeah, as an extended,

623 00:29:52.880 --> 00:29:55.520 the ideas from this particular figure, right?

624 00:29:55.520 --> 00:29:57.240 So, you have two methods,

625 00:29:57.240 --> 00:29:59.030 they're giving you two different...

626 00:30:00.010 --> 00:30:01.323 Okay, two methods with their orderings

627 00:30:01.323 --> 00:30:04.760 and pseudo times, so basically four cases,

628 00:30:04.760 --> 00:30:07.590 and they all give you different rankings,

629 00:30:07.590 --> 00:30:12.390 which have different performances

630 00:30:12.390 --> 00:30:14.410 in terms of network evaluation.

631 00:30:14.410 --> 00:30:16.670 And in a sense, you could say

632 00:30:18.711 --> 00:30:21.960 that each of these PAGA Tree inference methods

633 00:30:21.960 --> 00:30:24.870 itself with all their inefficiencies

634 00:30:24.870 --> 00:30:27.990 and efficiencies are only partially looking

635 00:30:27.990 --> 00:30:29.990 at the biological data.

636 00:30:29.990 --> 00:30:33.730 So, from that perspective, each

637 00:30:33.730 --> 00:30:37.150 of these orderings and pseudo time values

638 00:30:37.150 --> 00:30:38.560 can be considered as sources

639 00:30:38.560 --> 00:30:39.980 of noisy information,

640 00:30:39.980 --> 00:30:41.580 or noisy sources of information.

641 00:30:42.430 --> 00:30:46.357 So, instead of trying to just infer

642 00:30:48.860 --> 00:30:52.100 one pseudo time trajectory from

643 00:30:52.100 --> 00:30:54.810 the dataset and finding the network,

644 00:30:54.810 --> 00:30:55.940 or say another, and finding

645 00:30:55.940 --> 00:30:58.480 the network from that, we could think

646 00:30:58.480 --> 00:31:01.380 of the trajectory inference method itself

647 00:31:01.380 --> 00:31:03.140 as an additional hyper parameter

648 00:31:03.140 --> 00:31:06.310 on top of the sparsity, and kernel bits,

649 00:31:06.310 --> 00:31:07.520 and so on.

650 00:31:07.520 --> 00:31:10.290 So, instead of aggregating at this point

651 00:31:10.290 --> 00:31:11.900 after just one trajectory inference method,

652 00:31:11.900 --> 00:31:14.040 we could just say that maybe

653 00:31:14.040 --> 00:31:16.110 we have four trajectory inference methods

654 00:31:18.950 --> 00:31:20.200 in the beginning.

655 00:31:20.200 --> 00:31:22.760 And after that, we do all

656 00:31:22.760 --> 00:31:24.980 of these sub sampling and application

657 00:31:24.980 --> 00:31:27.520 of hyper-parameters, and multiple tests.

658 00:31:27.520 --> 00:31:29.370 And then, we aggregate over all

659 00:31:29.370 --> 00:31:30.910 of these results across

660 00:31:30.910 --> 00:31:32.820 trajectory inference methods.

661 00:31:32.820 --> 00:31:34.080 So, hopefully what that would do

662 00:31:34.080 --> 00:31:39.080 is that would account for all the inefficiencies,

663 00:31:39.290 --> 00:31:40.470 or counter then inefficiencies

664 00:31:40.470 --> 00:31:42.910 of individual trajectory inference methods,

665 00:31:42.910 --> 00:31:47.383 and give us a more robust network at the end.

666 00:31:49.280 --> 00:31:52.110 And I have not, I guess, shown

667 00:31:52.110 --> 00:31:54.010 our comparisons for the other methods,

668 00:31:55.730 --> 00:31:57.640 which obviously isn't in our paper.

669 00:31:57.640 --> 00:31:59.760 We are doing better than them.

670 00:31:59.760 --> 00:32:01.800 So, but you can have a look at

671 00:32:02.680 --> 00:32:05.030 that in the paper if you're interested,

672 00:32:05.030 --> 00:32:07.670 because I just wanted to conceptually focus

673 00:32:07.670 --> 00:32:09.723 on these ideas a little bit more.

674 00:32:11.010 --> 00:32:13.900 So, I guess, one problem with trying

675 00:32:13.900 --> 00:32:16.570 to run four different, or five different

676 00:32:16.570 --> 00:32:19.060 trajectory inference methods is depending on

677 00:32:19.060 --> 00:32:20.237 what kind of data set you have

678 00:32:20.237 --> 00:32:22.393 and what kind of biology you are studying,

679 00:32:23.370 --> 00:32:27.330 you might not necessarily have

680 00:32:27.330 --> 00:32:28.850 to try only four methods.

681 00:32:28.850 --> 00:32:29.683 You will probably have

682 00:32:29.683 --> 00:32:32.080 to try multiple methods before,

683 00:32:32.080 --> 00:32:34.210 which let's say, if you know

684 00:32:34.210 --> 00:32:35.310 it's a branching trajectory,

685 00:32:35.310 --> 00:32:38.150 you end up seeing a branching trajectory.

686 00:32:38.150 --> 00:32:41.200 And each of these methods would have

687 00:32:41.200 --> 00:32:44.320 their own input data format,

688 00:32:44.320 --> 00:32:46.503 up data formats, visualizations,

689 00:32:49.267 --> 00:32:52.300 and all of these other intricacies.

690 00:32:52.300 --> 00:32:55.170 And that's where the dynverse project comes

691 00:32:55.170 --> 00:32:56.430 to our rescue.

692 00:32:56.430 --> 00:32:59.790 So, if anyone is looking to do

693 00:32:59.790 --> 00:33:01.000 a lot of trajectory inference methods,

694 00:33:01.000 --> 00:33:03.900 I would strongly encourage you to look at that.

695 00:33:03.900 --> 00:33:06.000 So, these in this project,

696 00:33:06.000 --> 00:33:09.850 they have streamlined the use of, I think,

697 00:33:09.850 --> 00:33:11.670 55 trajectory inference methods.

698 00:33:11.670 --> 00:33:14.060 So, you don't necessarily need to install

699 00:33:14.060 --> 00:33:14.893 each one of them.

700 00:33:14.893 --> 00:33:16.300 You just install this project

701 00:33:16.300 --> 00:33:18.240 and they run each

702 00:33:18.240 --> 00:33:20.700 of these methods using a docker.

703 00:33:20.700 --> 00:33:23.470 And so, what it also helps you do

704 00:33:23.470 --> 00:33:26.160 is it helps you visualize

705 00:33:26.160 --> 00:33:31.160 all of these trajectories and evaluate them using

706 00:33:31.420 --> 00:33:34.590 the same, I guess, support scripts

707 00:33:34.590 --> 00:33:37.900 and support functions, which they also provide.

708 00:33:37.900 --> 00:33:41.720 And in all this, this would make

709 00:33:41.720 --> 00:33:43.920 your lives quite easy.

710 00:33:43.920 --> 00:33:46.850 And they also have basically a user,

711 00:33:46.850 --> 00:33:48.020 a graphical user interface,

712 00:33:48.020 --> 00:33:51.740 which helps you prioritize

713 00:33:51.740 --> 00:33:55.060 what trajectory inference method to use,

714 00:33:55.060 --> 00:33:59.902 depending on what biology you want to study.

715 00:33:59.902 --> 00:34:02.340 How many cells you have, what compute power

716 00:34:02.340 --> 00:34:05.973 you might have access to, and so on.

717 00:34:12.272 --> 00:34:16.650 So, okay just some final comments on the use

718 00:34:16.650 --> 00:34:18.980 of, I guess, the utility of trajectory inference

719 00:34:18.980 --> 00:34:22.250 and pseudo times for further analysis.

720 00:34:22.250 --> 00:34:24.980 And so, first of all, as in trajectories

721 00:34:24.980 --> 00:34:27.990 look really nice, they visually,

722 00:34:27.990 --> 00:34:30.510 they give us a lot of information.

723 00:34:30.510 --> 00:34:33.270 And so, based on what we saw,

724 00:34:33.270 --> 00:34:35.773 we did see that there's some,

725 00:34:38.646 --> 00:34:40.510 the ordering information

726 00:34:40.510 --> 00:34:43.040 and the pseudo time values can help

727 00:34:43.040 --> 00:34:44.113 in network inference.

728 00:34:45.090 --> 00:34:48.520 The good pseudo times can help a little bit,

729 00:34:48.520 --> 00:34:51.480 but if you have exceptionally bad pseudo times,

730 00:34:51.480 --> 00:34:54.033 it can hurt a lot as opposed to ordering.

731 00:34:54.960 --> 00:34:58.710 And not every dataset is really suitable

732 00:34:58.710 --> 00:34:59.560 for trajectory inference.

733 00:34:59.560 --> 00:35:00.820 What do I mean by that?

734 00:35:00.820 --> 00:35:04.160 So, the dataset that I chose,

735 00:35:04.160 --> 00:35:07.430 and I guess a lot of what is...

736 00:35:08.400 --> 00:35:09.630 What particular inference methods

737 00:35:09.630 --> 00:35:11.190 are built around, as say,

738 00:35:11.190 --> 00:35:14.310 stem cell differentiation in general,

739 00:35:14.310 --> 00:35:19.220 where it's as in the biology is quite neat

740 00:35:19.220 --> 00:35:20.053 to begin with.

741 00:35:20.053 --> 00:35:23.260 As in you start off from a single cell type,

742 00:35:23.260 --> 00:35:27.180 and a lot of the biology is already known.

743 00:35:27.180 --> 00:35:30.010 So, you don't have to worry, you know

744 00:35:30.010 --> 00:35:31.820 that it's going to be a branching,

745 00:35:31.820 --> 00:35:36.460 or bifurcating, or multi furcating trajectory.

746 00:35:36.460 --> 00:35:38.130 So, you know that the quality of the biology,

747 00:35:38.130 --> 00:35:42.580 you know what cell states to exist, to expect,

748 00:35:42.580 --> 00:35:43.770 and so on, and so forth.

749 00:35:43.770 --> 00:35:45.930 You know the markers of each of those.

750 00:35:45.930 --> 00:35:49.360 And so, studying something like that

751 00:35:49.360 --> 00:35:53.040 is much more easier using trajectory inference,

752 00:35:53.040 --> 00:35:54.460 or pseudo time.

753 00:35:54.460 --> 00:35:55.970 On the other hand, let's say,

754 00:35:55.970 --> 00:35:59.330 if you had a sample from a cancer tumor

755 00:35:59.330 --> 00:36:02.380 in that you would find cancer cells,

756 00:36:02.380 --> 00:36:06.140 normal cells, a bunch of immune cells,

757 00:36:06.140 --> 00:36:10.350 probably 10 to 20 kinds of immune cells,

758 00:36:10.350 --> 00:36:11.580 and so on.

759 00:36:11.580 --> 00:36:13.620 So, the trajectory inference method

760 00:36:14.680 --> 00:36:18.040 usually tracks, or predicts places,

761 00:36:18.040 --> 00:36:19.820 cell states and context.

762 00:36:19.820 --> 00:36:22.770 Not cell types themselves.

763 00:36:22.770 --> 00:36:25.270 So, you wouldn't necessarily be able

764 00:36:25.270 --> 00:36:28.560 to reliably run a trajectory inference method

765 00:36:28.560 --> 00:36:33.020 across as in using a mix of different cell types,

766 00:36:33.020 --> 00:36:34.990 as opposed to cell states.

767 00:36:34.990 --> 00:36:38.220 Now, with the stem cell differentiation,

768 00:36:38.220 --> 00:36:40.780 the good thing is that the cell states

769 00:36:40.780 --> 00:36:43.460 themselves after a point, transition

770 00:36:43.460 --> 00:36:45.200 into different cell types,

771 00:36:45.200 --> 00:36:47.030 because it's the same cell,

772 00:36:47.030 --> 00:36:50.170 or same cell type which transitions

773 00:36:50.170 --> 00:36:51.803 through multiple cell types,

774 00:36:52.800 --> 00:36:55.560 through these cell states.

775 00:36:55.560 --> 00:36:58.170 But that's not the case with cancer biology,

776 00:36:58.170 --> 00:37:00.640 where you already start off

777 00:37:00.640 --> 00:37:05.640 with a mix of cell types and trajectory inference

778 00:37:05.800 --> 00:37:08.430 would not make sense for that mix.

779 00:37:08.430 --> 00:37:10.610 What people have tried is isolate,

780 00:37:10.610 --> 00:37:15.610 just say a T-cell type, and then try

781 00:37:15.910 --> 00:37:18.920 to order, or find the trajectory only

782 00:37:18.920 --> 00:37:21.230 for those T-cells.

783 00:37:21.230 --> 00:37:23.350 And there has been some success in that.

784 00:37:23.350 --> 00:37:27.230 So, you could run trajectory inference

785 00:37:27.230 --> 00:37:29.920 for a subset of the dataset, but not necessarily

786 00:37:29.920 --> 00:37:30.870 the entire dataset.

787 00:37:32.230 --> 00:37:37.230 And so, depending on what biological processes

788 00:37:37.910 --> 00:37:38.853 you want to study,

789 00:37:40.820 --> 00:37:42.620 there are trajectory inference methods,

790 00:37:42.620 --> 00:37:44.600 which may or may not be suitable for it.

791 00:37:44.600 --> 00:37:47.350 For example, a number of methods

792 00:37:47.350 --> 00:37:51.140 like Monocle and PAGA Tree,

793 00:37:51.140 --> 00:37:56.140 they try to find tree-like structures

794 00:37:56.300 --> 00:37:58.180 in the trajectories,

795 00:37:58.180 --> 00:37:59.230 so they would not be suitable

796 00:37:59.230 --> 00:38:03.440 for a cyclic biological process

797 00:38:03.440 --> 00:38:06.043 like just maintenance processes in cells.

798 00:38:06.900 --> 00:38:08.010 And then, there are other methods

799 00:38:08.010 --> 00:38:10.800 which actually try to find cell cycles,

800 00:38:10.800 --> 00:38:12.210 and they would not be appropriate

801 00:38:12.210 --> 00:38:13.653 for branching processes.

802 00:38:15.550 --> 00:38:18.630 And I guess, as a no single

803 00:38:18.630 --> 00:38:20.030 trajectory inference method,

804 00:38:22.550 --> 00:38:25.200 accurately represents the biology.

805 00:38:25.200 --> 00:38:26.720 So, it's all basically

806 00:38:26.720 --> 00:38:28.780 some mathematical abstraction

807 00:38:28.780 --> 00:38:31.133 of what might be happening in the cells.

808 00:38:35.216 --> 00:38:36.049 And yeah, as an if...

809 00:38:36.049 --> 00:38:37.170 If at the outset, you know

810 00:38:37.170 --> 00:38:41.090 what kind of trajectory to expect, then it helps

811 00:38:41.090 --> 00:38:42.240 in trying to

812 00:38:44.770 --> 00:38:45.910 at least first really,

813 00:38:45.910 --> 00:38:49.790 say whether the trajectory that you're getting

814 00:38:49.790 --> 00:38:52.060 and the pseudo times that you get

815 00:38:52.060 --> 00:38:55.030 is of any worth.

816 00:38:55.030 --> 00:38:57.920 So, just to give you an example.

817 00:38:57.920 --> 00:39:00.290 So, we started off with Monocle 2

818 00:39:00.290 --> 00:39:02.560 as one of our examples in our paper,

819 00:39:02.560 --> 00:39:05.330 and then we wanted to have another method

820 00:39:05.330 --> 00:39:07.413 to compare the effects of different

821 00:39:07.413 --> 00:39:09.890 trajectory inference methods.

822 00:39:09.890 --> 00:39:12.850 And PAGA Tree was not necessarily the first one.

823 00:39:12.850 --> 00:39:14.290 We tried a number of other ones,

824 00:39:14.290 --> 00:39:16.080 which did not.

825 00:39:16.080 --> 00:39:18.380 And we knew what to expect here.

826 00:39:18.380 --> 00:39:21.160 We knew that there was stem cell

827 00:39:21.160 --> 00:39:26.160 to ectoderm trajectory and endoderm trajectory,

828 00:39:26.430 --> 00:39:27.980 or a branch of that.

829 00:39:27.980 --> 00:39:32.980 And using basically, just the first,

830 00:39:35.040 --> 00:39:37.560 I think we tried four methods

831 00:39:37.560 --> 00:39:39.400 and PAGA Tree was basically the fourth method,

832 00:39:39.400 --> 00:39:41.690 which gave us that kind of branching trajectory,

833 00:39:41.690 --> 00:39:44.870 or branching topology for the biology.

834 00:39:44.870 --> 00:39:49.250 And so, none of the methods you try

835 00:39:49.250 --> 00:39:51.803 might necessarily mean anything,

836 00:39:53.010 --> 00:39:55.503 unless you have some way of validating that.

837 00:39:56.520 --> 00:39:59.010 So, at this point, I'm gonna switch

838 00:39:59.010 --> 00:40:04.010 to spatial expression,

839 00:40:04.070 --> 00:40:05.820 or a spatial data and special analysis.

840 00:40:05.820 --> 00:40:08.230 So, if you have any questions

841 00:40:08.230 --> 00:40:12.380 about the pseudo time analysis,

842 00:40:12.380 --> 00:40:13.813 should we take it now, or?

843 00:40:19.010 --> 00:40:20.270 <v Lecturer>Does anybody have any questions</v>

844 00:40:20.270 --> 00:40:22.753 on the first half of the presentation here?

845 00:40:26.265 --> 00:40:27.290 <v Dr. Deshpande>Oh, we can continue on,</v>

846 00:40:27.290 --> 00:40:29.707 then we can come back later.

847 00:40:34.439 --> 00:40:35.689 Shall we go on?

848 00:40:40.670 --> 00:40:42.148 <v Lecturer>Sounds good.</v>

849 00:40:42.148 --> 00:40:44.148 <v Dr. Deshpande>Okay.</v>

850 00:40:47.730 --> 00:40:49.653 Okay, so that was all about,

851 00:40:51.940 --> 00:40:56.940 say how pseudo time is used in our analysis.

852 00:40:56.940 --> 00:41:01.617 And so, the other end of,

853 00:41:03.170 --> 00:41:04.470 I guess, not necessarily end,

854 00:41:04.470 --> 00:41:05.310 the other perspective

855 00:41:05.310 --> 00:41:09.600 is how is space important and how,

856 00:41:09.600 --> 00:41:11.050 what kind of data do we have,

857 00:41:12.540 --> 00:41:14.843 which give us information about space?

858 00:41:15.760 --> 00:41:18.450 So, the spatial context of cells

859 00:41:18.450 --> 00:41:21.570 is very important in many biological processes.

860 00:41:21.570 --> 00:41:23.690 For example, when immune cells respond

861 00:41:23.690 --> 00:41:26.670 to an infection, or a wound, they need

862 00:41:26.670 --> 00:41:28.920 to be in physical proximity of their targets.

863 00:41:31.010 --> 00:41:33.840 Similarly with, I guess, cancer tumor growth,

864 00:41:33.840 --> 00:41:37.560 and the immune response to cancer

865 00:41:37.560 --> 00:41:39.720 happen through intracellular signaling.

866 00:41:39.720 --> 00:41:42.390 Either through cytokine secretion,

867 00:41:42.390 --> 00:41:45.563 or through surface receptors on adjacent cells.

868 00:41:48.410 --> 00:41:50.170 Just knowing the relative location

869 00:41:50.170 --> 00:41:51.520 of different cell types can also

870 00:41:51.520 --> 00:41:52.820 be very informative.

871 00:41:52.820 --> 00:41:55.393 For example, in the figure here,

872 00:41:56.630 --> 00:41:57.960 the information about the presence

873 00:41:57.960 --> 00:42:01.920 of various immune cell types nearest tumor,

874 00:42:01.920 --> 00:42:03.720 and the extent of immune deficient

875 00:42:03.720 --> 00:42:06.163 in the tumor are essential prognostic markers.

876 00:42:07.930 --> 00:42:10.850 And so, single cell RNA-seq,

877 00:42:12.550 --> 00:42:15.270 as good as it is, it associates a cell

878 00:42:15.270 --> 00:42:17.900 from its tissue, due to which

879 00:42:17.900 --> 00:42:20.820 we lose the spatial context of the cell states.

880 00:42:20.820 --> 00:42:23.580 But in recent years, we have been able

881 00:42:23.580 --> 00:42:28.170 to develop both

882 00:42:28.170 --> 00:42:29.760 as in spatial proteomics,

883 00:42:29.760 --> 00:42:34.760 which help you to image protein

884 00:42:35.080 --> 00:42:39.800 and densities of say, up to 30 markers

885 00:42:39.800 --> 00:42:41.883 at single cell resolution in the tissue.

886 00:42:43.340 --> 00:42:45.910 As well as spatial transcriptomics,

887 00:42:45.910 --> 00:42:50.540 which can measure 20,000 genes at spots

888 00:42:50.540 --> 00:42:52.970 in the tissue.

889 00:42:52.970 --> 00:42:56.610 And this was named method of the year last year

890 00:42:56.610 --> 00:42:58.883 in 2020, yeah, that was last year.

891 00:43:01.071 --> 00:43:04.580 So, here's just a workflow

892 00:43:04.580 --> 00:43:06.450 of the next Visium technology,

893 00:43:06.450 --> 00:43:07.283 which is one of these

894 00:43:07.283 --> 00:43:09.950 spatial transcriptomics technologies.

895 00:43:09.950 --> 00:43:14.853 So, this includes 5,000 barcoded spots on slide.

896 00:43:15.800 --> 00:43:20.800 And these are added to the cells in the...

897 00:43:21.250 --> 00:43:24.110 Which are located in those spots.

898 00:43:24.110 --> 00:43:26.120 And this helps preserve the spatial context

899 00:43:26.120 --> 00:43:28.533 of the cells to the actual sequencing.

900 00:43:29.600 --> 00:43:33.240 Now, this technology is not exactly single cell.

901 00:43:33.240 --> 00:43:37.363 It still provides a lot of useful spacial detail.

902 00:43:41.080 --> 00:43:46.080 So yeah, for explaining this project,

903 00:43:46.860 --> 00:43:50.770 I will use the 10x Visium sample,

904 00:43:50.770 --> 00:43:52.140 provided by 10x genomics

905 00:43:53.040 --> 00:43:54.950 of a breast cancer tissue.

906 00:43:54.950 --> 00:43:56.700 So, the figure on the left

907 00:43:56.700 --> 00:43:59.690 is an H and E slide, it's hematoxylin

908 00:44:00.785 --> 00:44:02.580 and eosin stain slide,

909 00:44:02.580 --> 00:44:07.580 which helps pathologists annotate

910 00:44:07.700 --> 00:44:12.700 the sample for tumor, and lesions, and so on.

911 00:44:13.680 --> 00:44:18.500 And the second image is that slide annotated

912 00:44:18.500 --> 00:44:21.500 by a pathologist, and you can see

913 00:44:21.500 --> 00:44:24.640 that there are different biology's

914 00:44:24.640 --> 00:44:27.130 in this one slide.

915 00:44:27.130 --> 00:44:29.090 And for example, the lesion on top

916 00:44:29.090 --> 00:44:31.050 is an invasive cancer lesion, which means

917 00:44:31.050 --> 00:44:33.330 that it can spread beyond the breast tissue,

918 00:44:33.330 --> 00:44:34.900 but the other lesions correspond

919 00:44:34.900 --> 00:44:35.970 to DCAs lesions,

920 00:44:35.970 --> 00:44:38.840 which are not yet classified as invasive,

921 00:44:38.840 --> 00:44:41.700 they could in the future be invasive.

922 00:44:41.700 --> 00:44:43.390 Other important annotations are those

923 00:44:43.390 --> 00:44:46.600 of immune cells and the stromal cells

924 00:44:46.600 --> 00:44:48.013 in between these lesions.

925 00:44:49.680 --> 00:44:51.800 For a good clinical outcome, you would hope

926 00:44:51.800 --> 00:44:54.683 that immune cells can infiltrate these lesions.

927 00:44:55.590 --> 00:44:59.100 And so the figure on the right shows

928 00:44:59.100 --> 00:45:00.360 the same H and E slide

929 00:45:02.000 --> 00:45:04.660 with overlaid Visium spots.

930 00:45:04.660 --> 00:45:06.870 So, each of these spots correspond

931 00:45:06.870 --> 00:45:08.133 to one measurement.

932 00:45:09.510 --> 00:45:14.400 So, this slide shows a couple of examples

933 00:45:14.400 --> 00:45:16.900 of spacial gene expression.

934 00:45:16.900 --> 00:45:18.550 So, the figure to the left

935 00:45:18.550 --> 00:45:21.230 is the same annotated H and E slide

936 00:45:21.230 --> 00:45:23.480 that will help us keep track

937 00:45:23.480 --> 00:45:26.870 of the biology in the slide.

938 00:45:26.870 --> 00:45:29.900 And so, the first figure, the middle figure,

939 00:45:29.900 --> 00:45:32.770 basically it shows the expression of CD8A,

940 00:45:32.770 --> 00:45:35.600 which is a marker of cytotoxic T-cells.

941 00:45:35.600 --> 00:45:37.020 Now, we see this gene expressed

942 00:45:37.020 --> 00:45:42.020 in the blood near the invasive and DCAs lesions,

943 00:45:42.480 --> 00:45:43.810 which means that the immune cells

944 00:45:43.810 --> 00:45:44.900 are responding to a tumor.

945 00:45:44.900 --> 00:45:46.550 However, we see that

946 00:45:46.550 --> 00:45:48.740 there's not much infiltration of these cells

947 00:45:48.740 --> 00:45:49.743 within the lesions.

948 00:45:50.920 --> 00:45:53.990 The second marker is CD14, which is found

949 00:45:53.990 --> 00:45:55.740 in macrophages and dendritic cells,

950 00:45:56.630 --> 00:45:58.310 and its expression is much higher

951 00:45:58.310 --> 00:45:59.730 inside the lesions, which could point

952 00:45:59.730 --> 00:46:03.670 to successful infiltration of these cell types.

953 00:46:03.670 --> 00:46:08.240 Now, just a reminder, these the measurements

954 00:46:08.240 --> 00:46:09.870 that we get from 10x Visium

955 00:46:09.870 --> 00:46:13.000 are not exactly single cell, but they're near,

956 00:46:13.000 --> 00:46:14.660 near single cell.

957 00:46:14.660 --> 00:46:16.490 In a sense that each of these spots

958 00:46:16.490 --> 00:46:18.433 is 55 micro meters wide.

959 00:46:19.360 --> 00:46:21.690 And depending on what cell type

960 00:46:21.690 --> 00:46:22.910 you might have in that spot,

961 00:46:22.910 --> 00:46:26.920 it could have anywhere from one to 10 cells.

962 00:46:26.920 --> 00:46:28.140 And immune cells are much smaller,

963 00:46:28.140 --> 00:46:30.300 so there could be up to 10 immune cells in it,

964 00:46:30.300 --> 00:46:32.280 but maybe only one cancer,

965 00:46:32.280 --> 00:46:34.750 or epithelial cell in that spot.

966 00:46:34.750 --> 00:46:36.410 So, as a result of gene expression

967 00:46:36.410 --> 00:46:38.460 of that spot is the average

968 00:46:38.460 --> 00:46:39.610 of the cells inside it.

969 00:46:41.910 --> 00:46:46.754 Now, our lab has a method called CoGAPS,

970 00:46:46.754 --> 00:46:49.640 oesophageal CoGAPS, which is a Bayesian

971 00:46:49.640 --> 00:46:51.480 Markov chain Monte Carlo method

972 00:46:51.480 --> 00:46:53.330 for nonnegative matrix factorization.

973 00:46:54.370 --> 00:46:58.470 And so, as a result of say,

974 00:46:58.470 --> 00:47:00.940 the 10x Visium measurement,

975 00:47:00.940 --> 00:47:04.210 we now have a high dimensional matrix

976 00:47:04.210 --> 00:47:08.020 with 20,000 genes and around 5,000 spots.

977 00:47:08.020 --> 00:47:11.610 And what CoGAPS does is it helps

978 00:47:11.610 --> 00:47:14.730 to factorize this matrix

979 00:47:14.730 --> 00:47:17.750 into two low rank matrices,

980 00:47:17.750 --> 00:47:19.363 both of which are non-negative,

981 00:47:20.860 --> 00:47:25.450 which correspond to latent patterns in the data.

982 00:47:25.450 --> 00:47:26.790 And in the past, we have seen

983 00:47:26.790 --> 00:47:30.150 that these two correspond to biology's

984 00:47:31.240 --> 00:47:33.033 based on the pattern markers.

985 00:47:34.220 --> 00:47:37.780 So, the two matrices that CoGAPS factorizes

986 00:47:37.780 --> 00:47:40.500 the dataset into are the amplitude matrix,

987 00:47:40.500 --> 00:47:44.220 which has say, 20,000 rows for 20,000 genes

988 00:47:44.220 --> 00:47:46.573 and N columns for the end patterns.

989 00:47:47.560 --> 00:47:49.710 And this helps us identify groups

990 00:47:49.710 --> 00:47:51.640 of co-expressed genes,

991 00:47:51.640 --> 00:47:53.580 which correspond to the patterns.

992 00:47:53.580 --> 00:47:56.520 And the pattern matrix has N rows

993 00:47:56.520 --> 00:48:01.180 and 5,000 columns, and they associate the spots

994 00:48:01.180 --> 00:48:03.820 on the sample with patterns.

995 00:48:03.820 --> 00:48:07.744 So, because of the nature of the CoGAPS,

996 00:48:07.744 --> 00:48:10.850 factorization, and these, the columns

997 00:48:10.850 --> 00:48:13.810 of the matrices here, or the rows of the matrices

998 00:48:13.810 --> 00:48:15.390 here are not really orthogonal.

999 00:48:15.390 --> 00:48:17.410 They are independent, but not orthogonal.

1000 00:48:17.410 --> 00:48:20.240 So, they could co-exist in spots,

1001 00:48:20.240 --> 00:48:24.590 or a gene could be present in multiple processes,

1002 00:48:24.590 --> 00:48:25.423 and multiple patterns,

1003 00:48:25.423 --> 00:48:27.133 which correspond to processes.

1004 00:48:28.720 --> 00:48:33.720 So, when we apply CoGAPS to the Visium data,

1005 00:48:36.950 --> 00:48:39.050 so the first try was basically

1006 00:48:39.050 --> 00:48:42.620 just five patterns, and when we apply it

1007 00:48:42.620 --> 00:48:45.000 to try and find five patterns

1008 00:48:47.390 --> 00:48:50.060 after a factorization, we see that

1009 00:48:50.060 --> 00:48:51.420 a number of them correspond

1010 00:48:51.420 --> 00:48:53.990 to the pathology annotations

1011 00:48:53.990 --> 00:48:56.860 that we see on the figure on the left.

1012 00:48:56.860 --> 00:48:58.870 So, we find a pattern which corresponds

1013 00:48:58.870 --> 00:49:01.220 to the immune cells.

1014 00:49:01.220 --> 00:49:03.710 We find a pattern which corresponds

1015 00:49:03.710 --> 00:49:06.900 to invasive carcinoma on the top left here.

1016 00:49:06.900 --> 00:49:08.640 And we also find a pattern which corresponds

1017 00:49:08.640 --> 00:49:10.193 to the DCAs lesions.

1018 00:49:12.260 --> 00:49:14.860 And as we increase the dimensionality

1019 00:49:14.860 --> 00:49:17.670 of CoGAPS factorization, we start seeing more

1020 00:49:17.670 --> 00:49:19.520 and more tissue heterogeneity.

1021 00:49:19.520 --> 00:49:23.320 For example, we now see three patterns

1022 00:49:23.320 --> 00:49:25.580 which are associated with the mesial carcinoma,

1023 00:49:25.580 --> 00:49:27.160 and we can see that they correspond

1024 00:49:27.160 --> 00:49:31.440 to different regions in that lesion.

1025 00:49:31.440 --> 00:49:33.310 And this for example is completely internal,

1026 00:49:33.310 --> 00:49:35.823 which has no interaction with immune cells.

1027 00:49:36.680 --> 00:49:38.830 We have a pattern which corresponds

1028 00:49:38.830 --> 00:49:40.620 to immune cells, we have a pattern

1029 00:49:40.620 --> 00:49:43.350 which corresponds to the stromal cells.

1030 00:49:43.350 --> 00:49:46.870 And we also have different patterns

1031 00:49:46.870 --> 00:49:50.640 which highlight individual DCAs lesions.

1032 00:49:50.640 --> 00:49:53.300 So, one could say that potentially it's trying,

1033 00:49:53.300 --> 00:49:56.880 it is finding biology's,

1034 00:49:56.880 --> 00:49:59.273 which are unique to these DCAs lesions.

1035 00:50:03.938 --> 00:50:06.780 So, we can analyze the A matrix

1036 00:50:06.780 --> 00:50:08.610 to identify groups of genes associated

1037 00:50:08.610 --> 00:50:10.590 with each pattern, and we call these

1038 00:50:10.590 --> 00:50:12.030 the pattern markers.

1039 00:50:12.030 --> 00:50:14.590 And these help us identify pathways

1040 00:50:14.590 --> 00:50:17.819 that are likely expressed in these patterns,

1041 00:50:17.819 --> 00:50:20.350 or because now, especially in this sample,

1042 00:50:20.350 --> 00:50:22.640 we see a one to one association

1043 00:50:22.640 --> 00:50:25.290 between the pattern and the biology,

1044 00:50:25.290 --> 00:50:27.003 also in the biology, basically.

1045 00:50:28.630 --> 00:50:31.733 So, let's see, how long do we have.

1046 00:50:34.649 --> 00:50:37.570 I think we're close to...

1047 00:50:37.570 --> 00:50:39.850 I'll quickly rush through these.

1048 00:50:39.850 --> 00:50:44.850 So, the other analysis that we can do is given,

1049 00:50:45.480 --> 00:50:48.910 let's say two of these patterns,

1050 00:50:48.910 --> 00:50:52.390 we can try to see how these patterns interact.

1051 00:50:52.390 --> 00:50:53.850 So, you can see that these patterns

1052 00:50:53.850 --> 00:50:57.610 have a lot of spatial structure to it,

1053 00:50:57.610 --> 00:50:59.100 which CoGAPS was not told about.

1054 00:50:59.100 --> 00:51:00.840 CoGAPS, the parameters that Co-GAPS uses

1055 00:51:00.840 --> 00:51:02.790 have no special information,

1056 00:51:02.790 --> 00:51:05.730 and it's still found these spatial structures.

1057 00:51:05.730 --> 00:51:08.150 So, and we also see that these patterns

1058 00:51:08.150 --> 00:51:09.510 are adjacent to each other and we want

1059 00:51:09.510 --> 00:51:11.450 to see how they interact.

1060 00:51:11.450 --> 00:51:13.033 So, what we do is we find,

1061 00:51:14.880 --> 00:51:18.430 basically we estimate the kernel density

1062 00:51:18.430 --> 00:51:21.910 of each of these patterns, which is a function

1063 00:51:21.910 --> 00:51:24.630 of both the pattern intensity at a spot,

1064 00:51:24.630 --> 00:51:28.040 as well as the spatial clustering

1065 00:51:28.040 --> 00:51:30.400 of hyper intensities.

1066 00:51:30.400 --> 00:51:31.740 And we compare that against

1067 00:51:31.740 --> 00:51:34.820 another distribution obtained by

1068 00:51:34.820 --> 00:51:37.120 the density estimation after randomizing

1069 00:51:37.120 --> 00:51:39.353 the locations of these pattern densities.

1070 00:51:40.330 --> 00:51:43.490 So, the intensities which are beyond

1071 00:51:43.490 --> 00:51:46.570 distal distribution are the ones that we...

1072 00:51:48.100 --> 00:51:49.090 Are the spots which correspond

1073 00:51:49.090 --> 00:51:50.930 to these outliers are the ones

1074 00:51:50.930 --> 00:51:55.400 that we count as hotspots of pattern activity.

1075 00:51:55.400 --> 00:51:57.130 Similarly, we can find the hotspots

1076 00:51:57.130 --> 00:51:59.380 of immune response.

1077 00:51:59.380 --> 00:52:02.110 And when we combine both of them,

1078 00:52:02.110 --> 00:52:07.110 we find regions where cancer is active,

1079 00:52:08.720 --> 00:52:10.810 regions where immune cells are active,

1080 00:52:10.810 --> 00:52:13.970 and regions where both of them are active.

1081 00:52:13.970 --> 00:52:16.090 And this is the interaction region.

1082 00:52:16.090 --> 00:52:17.490 And in this region, we are trying

1083 00:52:17.490 --> 00:52:19.380 to find genes which correspond

1084 00:52:19.380 --> 00:52:24.380 to this interaction between cancer and immune,

1085 00:52:25.100 --> 00:52:27.220 and which are not necessarily markers of...

1086 00:52:27.220 --> 00:52:29.290 And regular markers of cancer and immune.

1087 00:52:29.290 --> 00:52:31.950 So, genes which are specifically related

1088 00:52:31.950 --> 00:52:33.910 to the non-linear interactions

1089 00:52:33.910 --> 00:52:36.913 between these patterns.

1090 00:52:37.810 --> 00:52:40.220 And to that end, basically we hypothesize

1091 00:52:40.220 --> 00:52:43.320 that since CoGAPS is already

1092 00:52:44.780 --> 00:52:47.300 an approximation of the dataset

1093 00:52:47.300 --> 00:52:49.720 with a linear combination of the patterns,

1094 00:52:49.720 --> 00:52:51.350 the residuals of CoGAPS,

1095 00:52:51.350 --> 00:52:53.300 of the CoGAPS estimate from the dataset

1096 00:52:54.810 --> 00:52:57.750 could point us to the non-linear interactions

1097 00:52:57.750 --> 00:53:01.090 between the patterns.

1098 00:53:01.090 --> 00:53:04.950 And we are only looking at the region

1099 00:53:06.410 --> 00:53:09.240 where both of the patterns are active

1100 00:53:09.240 --> 00:53:11.753 and comparing the residuals of CoGAPS

1101 00:53:12.870 --> 00:53:15.330 in that region to the residuals

1102 00:53:15.330 --> 00:53:16.480 in only the cancer region,

1103 00:53:16.480 --> 00:53:18.350 and only the immune region.

1104 00:53:18.350 --> 00:53:21.293 And now, this can be done for each of these,

1105 00:53:23.660 --> 00:53:24.840 I guess, pattern combinations,

1106 00:53:24.840 --> 00:53:28.810 and we can find what corresponds

1107 00:53:28.810 --> 00:53:30.010 to pattern interaction

1108 00:53:30.010 --> 00:53:32.340 between these pairs of patterns.

1109 00:53:32.340 --> 00:53:35.290 So, for future work, as part

1110 00:53:35.290 --> 00:53:37.280 of the data collection in clinical trials,

1111 00:53:37.280 --> 00:53:42.280 we're already collecting both spacial

1112 00:53:42.380 --> 00:53:44.440 and single cell transcriptomics

1113 00:53:44.440 --> 00:53:47.260 and proteomics from patients.

1114 00:53:47.260 --> 00:53:49.620 So, we are trying to integrate all

1115 00:53:49.620 --> 00:53:53.960 of this into one big dataset,

1116 00:53:53.960 --> 00:53:56.673 which would represent the tumor microenvironment,

1117 00:53:58.590 --> 00:54:00.550 which would help us characterize

1118 00:54:02.940 --> 00:54:05.030 the patient sample as a whole.

1119 00:54:05.030 --> 00:54:07.490 And we would also like

1120 00:54:07.490 --> 00:54:10.470 to infer intracellular signaling networks

1121 00:54:10.470 --> 00:54:11.670 the same way as we were trying to do

1122 00:54:11.670 --> 00:54:14.210 it using time, but now using space

1123 00:54:14.210 --> 00:54:18.130 where intracellular signaling is a function

1124 00:54:18.130 --> 00:54:19.950 of the distance between the cells

1125 00:54:19.950 --> 00:54:21.410 and the types of neighboring cells

1126 00:54:21.410 --> 00:54:22.463 for a target cell.

1127 00:54:25.549 --> 00:54:27.130 And the learnings from these projects

1128 00:54:27.130 --> 00:54:30.650 would go into a spatial temporal model

1129 00:54:30.650 --> 00:54:33.420 of tumor growth and response to therapy,

1130 00:54:33.420 --> 00:54:35.880 which can be used into building

1131 00:54:35.880 --> 00:54:39.370 a digital patient or digital clone,

1132 00:54:39.370 --> 00:54:43.050 where we can try to test what therapies

1133 00:54:43.050 --> 00:54:47.093 might work on what patients.

1134 00:54:48.430 --> 00:54:50.350 So, these are the people who have been,

1135 00:54:50.350 --> 00:54:51.540 and of course, 10x Genomics,

1136 00:54:51.540 --> 00:54:53.950 who were kind enough to give us the sample

1137 00:54:53.950 --> 00:54:58.510 for studying, as well as my collaborators

1138 00:54:58.510 --> 00:54:59.593 on this project.

1139 00:55:00.750 --> 00:55:01.583 Thank you so much.

1140 00:55:01.583 --> 00:55:03.010 And I can take questions now,

1141 00:55:04.350 --> 00:55:05.950 sorry for the overshooting time.

1142 00:55:09.776 --> 00:55:10.609 <v Lecturer>Thank you so much.</v>

1143 00:55:10.609 --> 00:55:15.320 Do we have any questions to look at?

1144 00:55:22.275 --> 00:55:24.561 People on Zoom? Yeah, question (mumbles).

1145 00:55:24.561 --> 00:55:25.410 <v Female Student>Going back</v>

1146 00:55:25.410 --> 00:55:27.676 to the time series slides.

1147 00:55:27.676 --> 00:55:28.509 <v ->Mm-hmm.</v>

1148 00:55:28.509 --> 00:55:29.342 <v Female Student>Can you talk</v>

1149 00:55:29.342 --> 00:55:30.520 about how you know if you have good,

1150 00:55:30.520 --> 00:55:32.240 or bad pseudo times?

1151 00:55:32.240 --> 00:55:35.040 And is there a way to fix bad pseudo times?

1152 00:55:35.040 --> 00:55:38.820 <v ->So, yeah, as in what I've not shared on here</v>

1153 00:55:38.820 --> 00:55:42.850 is so, in our experiments,

1154 00:55:42.850 --> 00:55:45.560 we also, we knew for example,

1155 00:55:45.560 --> 00:55:46.580 that we were studying...

1156 00:55:46.580 --> 00:55:48.710 We wanted to study a trajectory which goes

1157 00:55:48.710 --> 00:55:53.710 from stem cells to neuroectoderm,

1158 00:55:55.710 --> 00:55:57.120 and we had markers.

1159 00:55:57.120 --> 00:55:59.713 And I think, some (mumbles) themselves.

1160 00:56:00.860 --> 00:56:03.170 They have identified markers

1161 00:56:03.170 --> 00:56:08.170 of stem cells neuroectoderms and endoderm cells.

1162 00:56:08.660 --> 00:56:10.270 So, if we're looking at the trajectories

1163 00:56:10.270 --> 00:56:13.040 of the markers along the pseudo time

1164 00:56:13.040 --> 00:56:14.493 to see if those make sense.

1165 00:56:15.360 --> 00:56:17.800 For example, a marker which is supposed

1166 00:56:17.800 --> 00:56:21.250 to be high in stem cells would,

1167 00:56:21.250 --> 00:56:23.460 should be tapering down to zero

1168 00:56:23.460 --> 00:56:26.220 along pseudo time, and a marker,

1169 00:56:26.220 --> 00:56:29.610 which is supposed to be high in neuroectoderm

1170 00:56:29.610 --> 00:56:34.170 should be increasing with pseudo time.

1171 00:56:34.170 --> 00:56:37.600 So, we had, I think six oral markers

1172 00:56:37.600 --> 00:56:40.610 to each of stem cells, neuroectoderm

1173 00:56:40.610 --> 00:56:45.500 and endoderm cells.

1174 00:56:45.500 --> 00:56:49.040 And we were trying to confirm the combination

1175 00:56:49.040 --> 00:56:51.790 that neuroectoderm markers increase

1176 00:56:51.790 --> 00:56:53.750 with pseudo time, but the other two decrease,

1177 00:56:53.750 --> 00:56:58.130 or the endoderm shouldn't decrease necessarily,

1178 00:56:58.130 --> 00:57:00.300 but it shouldn't have

1179 00:57:00.300 --> 00:57:05.300 a monotonic increase like the neuroectoderm one.

1180 00:57:08.280 --> 00:57:10.903 And it should not be present in the initial.

1181 00:57:12.380 --> 00:57:13.213 Does that...

1182 00:57:14.310 --> 00:57:16.360 So, that was one way to do it, basically.

1183 00:57:21.120 --> 00:57:22.116 <v Lecturer>Thank you.</v>

1184 00:57:22.116 --> 00:57:23.783 Any other questions?

1185 00:57:39.510 --> 00:57:41.423 So, with the combination of many cells,

1186 00:57:41.423 --> 00:57:43.630 and the spatial stuff, is there any hope

1187 00:57:43.630 --> 00:57:45.740 of getting a temporal signal out of any of that,

1188 00:57:45.740 --> 00:57:46.940 or is that (indistinct)?

1189 00:57:49.897 --> 00:57:52.060 <v ->In spatial did you mean?</v>

1190 00:57:52.060 --> 00:57:53.350 <v Lecturer>Yeah.</v>

1191 00:57:53.350 --> 00:57:58.350 <v Dr. Deshpande>So, I think,</v>

1192 00:57:58.840 --> 00:58:00.060 the issue would be, I guess,

1193 00:58:00.060 --> 00:58:01.653 not in clinical, I suppose.

1194 00:58:03.740 --> 00:58:05.740 In a sense that, okay, are you thinking

1195 00:58:05.740 --> 00:58:08.045 about pseudo temporal, or just clinical?

1196 00:58:08.045 --> 00:58:09.700 <v Lecturer>Yeah.</v>

1197 00:58:09.700 --> 00:58:11.330 <v ->Pseudo temporal, I think there might</v>

1198 00:58:11.330 --> 00:58:12.290 be some possibility,

1199 00:58:12.290 --> 00:58:13.570 and I've been thinking of

1200 00:58:17.940 --> 00:58:19.500 as in, we would still have to isolate,

1201 00:58:19.500 --> 00:58:21.610 I guess, cell types, for example.

1202 00:58:21.610 --> 00:58:24.130 So, one of the problems with that

1203 00:58:24.130 --> 00:58:26.640 is that as I mentioned,

1204 00:58:26.640 --> 00:58:30.550 the spots are not exactly single cell, right?

1205 00:58:30.550 --> 00:58:32.800 So, especially, let's say if you're trying

1206 00:58:32.800 --> 00:58:35.500 to do a pseudo temporal ordering

1207 00:58:35.500 --> 00:58:38.283 of CD8 T-cells,

1208 00:58:39.380 --> 00:58:40.500 they are more,

1209 00:58:40.500 --> 00:58:44.460 more likely than not, co-localized

1210 00:58:44.460 --> 00:58:48.830 with other cell types, which would also,

1211 00:58:48.830 --> 00:58:51.250 I guess, corrupt the expression

1212 00:58:51.250 --> 00:58:53.030 that you are seeing.

1213 00:58:53.030 --> 00:58:57.180 So, that would make it slightly different.

1214 00:58:57.180 --> 00:59:01.300 We could think of ordering the spot

1215 00:59:01.300 --> 00:59:02.944 as a whole, basically.

1216 00:59:02.944 --> 00:59:03.777 And my...

1217 00:59:04.910 --> 00:59:07.120 I belong to a school of thought that basically,

1218 00:59:07.120 --> 00:59:08.446 if you have a...

1219 00:59:08.446 --> 00:59:09.600 And then, so what people try to do

1220 00:59:09.600 --> 00:59:13.000 with say, this kind of data,

1221 00:59:13.000 --> 00:59:14.560 this spacial Visium data,

1222 00:59:14.560 --> 00:59:16.830 where you have say, up to 10 cells,

1223 00:59:16.830 --> 00:59:21.830 they try to resolve this into cell types.

1224 00:59:22.730 --> 00:59:25.010 So, they would compare that to, there is I think,

1225 00:59:25.010 --> 00:59:27.703 one paper called RTCD, or RCTD.

1226 00:59:29.940 --> 00:59:32.223 RCTD robust cell type decomposition.

1227 00:59:33.470 --> 00:59:35.510 So, what they do is basically,

1228 00:59:35.510 --> 00:59:37.870 they take the spatial data,

1229 00:59:37.870 --> 00:59:41.060 they have a reference single cell data,

1230 00:59:41.060 --> 00:59:46.060 and they try to assign each spot,

1231 00:59:46.990 --> 00:59:51.030 or a resolve each spot into a mixture

1232 00:59:51.030 --> 00:59:53.930 of the cell types that might exist

1233 00:59:53.930 --> 00:59:55.543 in the single cell data.

1234 00:59:56.750 --> 01:00:01.010 And that could help you to say,

1235 01:00:01.010 --> 01:00:04.380 identify what the mixture in general is.

1236 01:00:04.380 --> 01:00:08.960 But my as in my thought is that we could

1237 01:00:08.960 --> 01:00:12.970 just think of each spot as some representation

1238 01:00:15.200 --> 01:00:16.820 of the biology in that neighborhood.

1239 01:00:16.820 --> 01:00:19.710 So, each spot could just represent

1240 01:00:19.710 --> 01:00:22.360 a neighborhood, as opposed to trying to find

1241 01:00:22.360 --> 01:00:23.893 what the individual cells are.

1242 01:00:24.990 --> 01:00:28.840 And that would basically abstract out

1243 01:00:30.460 --> 01:00:33.340 the representation and the biology to that

1244 01:00:33.340 --> 01:00:34.900 of the spots.

1245 01:00:34.900 --> 01:00:36.790 And we'll have to think about how to do that,

1246 01:00:36.790 --> 01:00:40.350 but I think there could be some ordering to that,

1247 01:00:40.350 --> 01:00:45.130 but we'll need to see what makes sense.

1248 01:00:45.130 --> 01:00:49.410 And then, for a lot of cells, cell states,

1249 01:00:49.410 --> 01:00:50.650 they are quite well-characterized.

1250 01:00:50.650 --> 01:00:52.670 For example, if you say that a T-cell

1251 01:00:52.670 --> 01:00:55.210 is activated, or a T-cell as naive,

1252 01:00:55.210 --> 01:00:59.150 or exhausted, you know what markers to expect.

1253 01:00:59.150 --> 01:01:01.400 But what would you be able to say

1254 01:01:02.340 --> 01:01:04.043 for spots instead?

1255 01:01:05.420 --> 01:01:09.900 The other thing to think of is,

1256 01:01:09.900 --> 01:01:12.320 especially with say, the proteomics as well,

1257 01:01:12.320 --> 01:01:14.370 where you can get actual single cell

1258 01:01:18.000 --> 01:01:22.380 and distributions, and neighborhood characterization.

1259 01:01:22.380 --> 01:01:25.033 You could think of it as can you,

1260 01:01:26.810 --> 01:01:28.347 so the same thing that...

1261 01:01:28.347 --> 01:01:30.620 The same ideas that were used

1262 01:01:30.620 --> 01:01:32.743 for pseudo temporal ordering of cells,

1263 01:01:33.590 --> 01:01:35.720 can they be used for pseudo temporal

1264 01:01:35.720 --> 01:01:38.680 ordering of neighborhoods?

1265 01:01:38.680 --> 01:01:40.730 For example, if you have a cell neighborhood,

1266 01:01:40.730 --> 01:01:44.868 which as they're presented as whatever,

1267 01:01:44.868 --> 01:01:47.763 the central cell, and it's five neighbors.

1268 01:01:48.710 --> 01:01:51.540 Now, depending on, are they all tumor?

1269 01:01:51.540 --> 01:01:52.860 Then maybe they have...

1270 01:01:52.860 --> 01:01:54.310 They're basically deep in the cancer,

1271 01:01:54.310 --> 01:01:57.183 which has never been visited by an immune cell,

1272 01:01:58.140 --> 01:01:59.380 is that a mix of tumor

1273 01:01:59.380 --> 01:02:01.570 and activated immune cells?

1274 01:02:01.570 --> 01:02:03.930 So, that is basically an active tumor

1275 01:02:03.930 --> 01:02:06.150 immune interaction that's happening.

1276 01:02:06.150 --> 01:02:10.220 Is that exhausted T-cells and tumor,

1277 01:02:10.220 --> 01:02:11.100 where basically the tumor

1278 01:02:11.100 --> 01:02:15.690 has fought back and tried to suppress the...

1279 01:02:15.690 --> 01:02:16.930 Or it's basically sent signals

1280 01:02:16.930 --> 01:02:21.130 to suppress the immune response, and so on.

1281 01:02:21.130 --> 01:02:22.433 So, perhaps there could be

1282 01:02:22.433 --> 01:02:24.730 a trajectory of neighborhoods,

1283 01:02:24.730 --> 01:02:28.810 where you could say that depending on all

1284 01:02:28.810 --> 01:02:31.130 the possible combinations that you expect

1285 01:02:31.130 --> 01:02:33.453 in cellular neighborhoods,

1286 01:02:35.330 --> 01:02:39.960 this current neighborhood is this far along

1287 01:02:39.960 --> 01:02:42.680 that process, or that branch of a process.

1288 01:02:44.393 --> 01:02:46.621 That was a long and winding answer.

1289 01:02:46.621 --> 01:02:47.740 (chuckles) I don't know if

1290 01:02:48.680 --> 01:02:51.690 that necessarily answered it. <v Lecturer>Thank you.</v>

1291 01:02:51.690 --> 01:02:54.490 Thank you, any last questions?

1292 01:02:54.490 --> 01:02:55.770 I wanna be mindful of time.

1293 01:02:55.770 --> 01:02:58.333 Any questions that come to you, or?

1294 01:03:06.287 --> 01:03:09.277 All right, well if not, thank you again.

1295 01:03:09.277 --> 01:03:11.168 (students applaud) We really appreciate that.

1296 01:03:11.168 --> 01:03:14.752 <v Dr. Deshpande>Thank you a lot.</v>

1297 01:03:14.752 --> 01:03:15.977 <v Lecturer>You have a wonderful (indistinct).</v>

1298 01:03:15.977 --> 01:03:16.810 <v ->Mm-hmm.</v>

1299 01:03:20.394 --> 01:03:24.394 (lecturer mumbles indistinctly)

1300 01:03:26.984 --> 01:03:31.067 (students chatter indistinctly)