1
00:00:00,000 --> 00:00:09,766

2
00:00:09,766 --> 00:00:12,333
Hi everyone we're gonna go ahead and get

3
00:00:12,333 --> 00:00:15,000
started. Welcome to - can everybody hear me?

4
00:00:15,000 --> 00:00:16,833
Okay to get started? Okay great!

5
00:00:16,833 --> 00:00:18,733
Welcome to Data Management for Medical

6
00:00:18,733 --> 00:00:21,233
Researchers. My name is Sarah Katz, I am

7
00:00:21,233 --> 00:00:22,866
the Health Science Librarian here at the

8
00:00:22,866 --> 00:00:25,000
University of Delaware, and I just want

9
00:00:25,000 --> 00:00:25,966
to give you a little bit of background

10
00:00:25,966 --> 00:00:28,566
about the workshop. Since August the

11
00:00:28,566 --> 00:00:30,566
University of Delaware Library has been

12
00:00:30,566 --> 00:00:33,066
participating in a research data

13
00:00:33,066 --> 00:00:35,000
management pilot project that's been

14
00:00:35,000 --> 00:00:36,866
funded by the National Network of

15
00:00:36,866 --> 00:00:38,733
Libraries of Medicine the Mid-Atlantic

16
00:00:38,733 --> 00:00:41,200
Region and this pilot's been hosted by

17
00:00:41,200 --> 00:00:44,266
the NYU Health Science Library. This

18
00:00:44,266 --> 00:00:47,233
project provides a holistic

19
00:00:47,233 --> 00:00:49,666
approach to developing data services

20
00:00:49,666 --> 00:00:52,100
that focuses on building a required

21
00:00:52,100 --> 00:00:54,033
knowledge base, understanding

22
00:00:54,033 --> 00:00:56,433
and connecting with researchers,

23
00:00:56,433 --> 00:00:58,533
and promoting effective outreach

24
00:00:58,533 --> 00:01:00,733
strategies, and integrating a broader

25
00:01:00,733 --> 00:01:02,700
institutional data community.

26
00:01:02,700 --> 00:01:05,400
Throughout this project we've conducted

27
00:01:05,400 --> 00:01:07,500
interviews with many faculty here at the

28
00:01:07,500 --> 00:01:09,700
University of Delaware College of Health

29
00:01:09,700 --> 00:01:13,900
Sciences and basically to kind of

30
00:01:13,900 --> 00:01:16,500
understand what we're doing here already

31
00:01:16,500 --> 00:01:19,800
about data services. In addition, this

32
00:01:19,800 --> 00:01:22,666
workshop is another aspect of this pilot.

33
00:01:22,666 --> 00:01:25,566
Ultimately we hope to improve data

34
00:01:25,566 --> 00:01:27,800
services on campus throughout the

35
00:01:27,800 --> 00:01:31,566
participation in this project. For

36
00:01:31,566 --> 00:01:34,433
some other information about what we've

37
00:01:34,433 --> 00:01:36,500
already been doing with regards to

38
00:01:36,500 --> 00:01:38,533
research data management I'm going to

39
00:01:38,533 --> 00:01:40,566
turn it over briefly to my colleague Tom

40
00:01:40,566 --> 00:01:42,733
Melvin. My name is Tom Melvin. Why

41
00:01:42,733 --> 00:01:44,266
am I here? Well, I'm the engineering

42
00:01:44,266 --> 00:01:46,400
liaison to most of the engineering

43
00:01:46,400 --> 00:01:48,166
departments. I do not do chemistry or

44
00:01:48,166 --> 00:01:52,700
biochemical engineering, but since 1989 I

45
00:01:52,700 --> 00:01:57,500
have been here and I represent

46
00:01:57,500 --> 00:01:59,200
(I've been in bands I know how to do

47
00:01:59,200 --> 00:02:02,633
this). I liaison to civil and

48
00:02:02,633 --> 00:02:04,133
environmental, material science,

49
00:02:04,133 --> 00:02:06,566
mechanical, electrical, computer,

50
00:02:06,566 --> 00:02:09,133
engineering. In the last four years I

51
00:02:09,133 --> 00:02:12,233
have also started liaisoning to geology

52
00:02:12,233 --> 00:02:14,566
geography and marine studies. Kind of

53
00:02:14,566 --> 00:02:16,366
somebody retired they asked "could you do

54
00:02:16,366 --> 00:02:18,366
this temporarily?" And I said, "yeah". Temporary

55
00:02:18,366 --> 00:02:21,166
became permanent, so. I work with

56
00:02:21,166 --> 00:02:22,500
faculty a lot on campus.

57
00:02:22,500 --> 00:02:25,200
In this aspect I have, for the last about

58
00:02:25,200 --> 00:02:27,866
five or six years, I've kind of been the

59
00:02:27,866 --> 00:02:29,566
contact person in our department

60
00:02:29,566 --> 00:02:32,766
(Reference and Instructional Services) for any

61
00:02:32,766 --> 00:02:35,533
data management issue. I'm wondering if

62
00:02:35,533 --> 00:02:37,533
anybody actually uses research guides

63
00:02:37,533 --> 00:02:40,166
off the library webpage? So there is a

64
00:02:40,166 --> 00:02:42,266
research guide which is a webpage on

65
00:02:42,266 --> 00:02:44,233
data management. If you just click on the

66
00:02:44,233 --> 00:02:47,366
research icon, the research guide icon on

67
00:02:47,366 --> 00:02:48,933
the library homepage, and type the word

68
00:02:48,933 --> 00:02:51,766
data you will get the link to it it. It has

69
00:02:51,766 --> 00:02:53,766
a lot of general information which is

70
00:02:53,766 --> 00:02:55,800
the nature of these guides. It will have

71
00:02:55,800 --> 00:02:57,900
links to some of the data repositories

72
00:02:57,900 --> 00:02:59,266
we're going to be talking about, some of

73
00:02:59,266 --> 00:03:01,400
the metadata directories that we're

74
00:03:01,400 --> 00:03:03,600
going to be talking about. We've

75
00:03:03,600 --> 00:03:06,166
filmed one of the writing and data

76
00:03:06,166 --> 00:03:08,500
management plan workshops. You can watch

77
00:03:08,500 --> 00:03:10,900
that video from there. We are filming the

78
00:03:10,900 --> 00:03:13,766
video today, and this will be uploaded

79
00:03:13,766 --> 00:03:15,333
there, as well as the slideshow.

80
00:03:15,333 --> 00:03:17,400
So I wanted to show you the page - we're

81
00:03:17,400 --> 00:03:18,766
having technical difficulties so that

82
00:03:18,766 --> 00:03:21,500
won't be possible. Maybe I'll be able to

83
00:03:21,500 --> 00:03:23,800
get to it later. As always if you have

84
00:03:23,800 --> 00:03:26,800
any questions please contact us. We're

85
00:03:26,800 --> 00:03:29,100
easy to find up at the library, and I'm

86
00:03:29,100 --> 00:03:33,766
gonna turn it back over to Sarah.

87
00:03:33,766 --> 00:03:36,033
We're gonna send out all the links to

88
00:03:36,033 --> 00:03:37,833
anybody who's registered to the workshop

89
00:03:37,833 --> 00:03:40,866
so we'll have a link to the data

90
00:03:40,866 --> 00:03:42,633
management research guide as well as

91
00:03:42,633 --> 00:03:44,366
everything we've talked about today.

92
00:03:44,366 --> 00:03:47,100
If we can't get to these pages later we

93
00:03:47,100 --> 00:03:49,466
will be sure to link to them so that you

94
00:03:49,466 --> 00:03:53,400
can see them. Alright so a little bit of

95
00:03:53,400 --> 00:03:55,700
audience-participation. We won't

96
00:03:55,700 --> 00:03:57,466
have too much of that, but a little bit.

97
00:03:57,466 --> 00:04:00,333
When you think of data management what

98
00:04:00,333 --> 00:04:03,900
exactly do you think of? Is there

99
00:04:03,900 --> 00:04:05,566
anything in particular that first comes

100
00:04:05,566 --> 00:04:07,366
to mind?

101
00:04:07,366 --> 00:04:12,666
Spreadsheets. What else? Yes -

102
00:04:12,666 --> 00:04:24,700
Security. File cabinets. Red cap. How to

103
00:04:24,700 --> 00:04:30,233
organize. Great. Version control. All great

104
00:04:30,233 --> 00:04:32,400
things exactly. These are all different

105
00:04:32,400 --> 00:04:33,733
types of things that we should be

106
00:04:33,733 --> 00:04:35,933
keeping in mind or that come in

107
00:04:35,933 --> 00:04:39,933
mind. Data management - all different

108
00:04:39,933 --> 00:04:41,733
things that we need to keep in mind, so

109
00:04:41,733 --> 00:04:44,266
data management leads to greater

110
00:04:44,266 --> 00:04:46,566
organization of data and workflows.

111
00:04:46,566 --> 00:04:49,700
Organized data is more comprehensible

112
00:04:49,700 --> 00:04:52,300
either to others or to researchers when

113
00:04:52,300 --> 00:04:54,533
they look back at it later. Organize

114
00:04:54,533 --> 00:04:56,633
workflows lead to more efficient

115
00:04:56,633 --> 00:04:59,700
research process. Data management ensures

116
00:04:59,700 --> 00:05:01,866
access to data in the future either to

117
00:05:01,866 --> 00:05:04,300
others or for the researcher. So all of

118
00:05:04,300 --> 00:05:06,000
these things are very very important

119
00:05:06,000 --> 00:05:08,433
when we think of data management. Both

120
00:05:08,433 --> 00:05:10,533
for you and for others looking at your

121
00:05:10,533 --> 00:05:15,033
data. So just a quick little thing a

122
00:05:15,033 --> 00:05:17,233
little overview of what we plan to cover

123
00:05:17,233 --> 00:05:20,866
today. Our learning objectives. So you

124
00:05:20,866 --> 00:05:22,200
could read them here but I'm just gonna

125
00:05:22,200 --> 00:05:24,100
quickly go over this just to give you an

126
00:05:24,100 --> 00:05:26,733
idea of what we plan to go over today.

127
00:05:26,733 --> 00:05:28,966
Recognize the current and forthcoming

128
00:05:28,966 --> 00:05:31,400
requirements that mandate the management

129
00:05:31,400 --> 00:05:34,533
and sharing of research data. Identify

130
00:05:34,533 --> 00:05:36,566
current requirements and issues around

131
00:05:36,566 --> 00:05:39,566
rigor and reproducibility. Apply best

132
00:05:39,566 --> 00:05:42,566
practices for creating and documenting

133
00:05:42,566 --> 00:05:45,633
file names, variables, and workflows.

134
00:05:45,633 --> 00:05:48,533
Identify appropriate options for storing

135
00:05:48,533 --> 00:05:51,433
and preserving research data. Locate

136
00:05:51,433 --> 00:05:54,633
appropriate standards, if any, and

137
00:05:54,633 --> 00:05:57,300
recognize their value for research.

138
00:05:57,300 --> 00:06:00,033
Evaluate repositories and determine the

139
00:06:00,033 --> 00:06:03,066
best sharing options for data. When we

140
00:06:03,066 --> 00:06:04,500
talk about the research data management

141
00:06:04,500 --> 00:06:07,700
climate what we're discussing here is

142
00:06:07,700 --> 00:06:12,033
why is this a big issue now? I know in

143
00:06:12,033 --> 00:06:13,633
the library world the last 10 years this

144
00:06:13,633 --> 00:06:15,400
keeps popping up a lot. That's why we did

145
00:06:15,400 --> 00:06:17,166
the courses we did, the workshops, the

146
00:06:17,166 --> 00:06:20,166
training. And one of the reasons we did

147
00:06:20,166 --> 00:06:22,066
the previous workshops we did was kind

148
00:06:22,066 --> 00:06:23,433
of just to get a feel of what's going on

149
00:06:23,433 --> 00:06:25,400
on campus. How many people are thinking

150
00:06:25,400 --> 00:06:26,233
about this how

151
00:06:26,233 --> 00:06:28,633
many people are aware of it. And as you

152
00:06:28,633 --> 00:06:30,233
can imagine what we discovered is it's

153
00:06:30,233 --> 00:06:32,100
everything's but he's kind of in their

154
00:06:32,100 --> 00:06:33,733
own place. There isn't much of a

155
00:06:33,733 --> 00:06:36,100
coordinated effort now, which is one of

156
00:06:36,100 --> 00:06:37,600
the things we're trying to become part

157
00:06:37,600 --> 00:06:40,166
of is moving in a more coordinate and

158
00:06:40,166 --> 00:06:42,666
coordinated campus-wide effort because

159
00:06:42,666 --> 00:06:44,000
it's becoming more of an issue

160
00:06:44,000 --> 00:06:46,566
specifically from grant funders and

161
00:06:46,566 --> 00:06:48,966
publishers both. Which we will both talk

162
00:06:48,966 --> 00:06:53,066
about. So, basically what does data

163
00:06:53,066 --> 00:06:55,600
management mean for your future? Why do

164
00:06:55,600 --> 00:06:58,600
you have to be aware of these things?

165
00:06:58,600 --> 00:07:00,200
A brief history, and what we're going to

166
00:07:00,200 --> 00:07:02,200
cover basically are the NIH data

167
00:07:02,200 --> 00:07:03,800
management requirements. We're going to

168
00:07:03,800 --> 00:07:05,433
be concentrating on the NIH because

169
00:07:05,433 --> 00:07:07,200
we're assuming that's where most of you

170
00:07:07,200 --> 00:07:09,033
are getting your funding from. It is the

171
00:07:09,033 --> 00:07:10,800
federal agency you deal with the most.

172
00:07:10,800 --> 00:07:13,166
When I did these before the engineering

173
00:07:13,166 --> 00:07:15,933
people lot of it was NSF. The NSF is

174
00:07:15,933 --> 00:07:17,333
actually a little more advanced on this

175
00:07:17,333 --> 00:07:18,933
and we'll talk about that a little bit.

176
00:07:18,933 --> 00:07:20,800
So we're gonna talk about the data

177
00:07:20,800 --> 00:07:22,766
management and sharing requirements that

178
00:07:22,766 --> 00:07:24,633
they have proposed. Nothing is codified

179
00:07:24,633 --> 00:07:26,766
with them yet, but they have very strong

180
00:07:26,766 --> 00:07:30,133
suggestions, as well as publisher data

181
00:07:30,133 --> 00:07:31,366
sharing requirements. You have to be

182
00:07:31,366 --> 00:07:33,900
aware of both. Because you might have a

183
00:07:33,900 --> 00:07:35,800
different requirements coming from who

184
00:07:35,800 --> 00:07:38,233
you get the funding from, as opposed to

185
00:07:38,233 --> 00:07:41,133
who you published with. And Sarah will

186
00:07:41,133 --> 00:07:42,500
talk a little bit about the rigor of

187
00:07:42,500 --> 00:07:44,333
reproducibility of the data which is

188
00:07:44,333 --> 00:07:46,000
really the main issue that we're talking

189
00:07:46,000 --> 00:07:48,433
about. Making sure that other people who

190
00:07:48,433 --> 00:07:51,033
want to access your data later can use

191
00:07:51,033 --> 00:07:53,733
it to get to the conclusions you came to.

192
00:07:53,733 --> 00:07:56,200
So, where did we start?

193
00:07:56,200 --> 00:07:59,500
Well back in 2003 is when this first

194
00:07:59,500 --> 00:08:01,766
became an issue. And the NIH came up on

195
00:08:01,766 --> 00:08:05,000
the policy that any grant application

196
00:08:05,000 --> 00:08:09,166
for over $500,000 had to submit a data

197
00:08:09,166 --> 00:08:11,566
management or data sharing plan with the

198
00:08:11,566 --> 00:08:14,400
grant application. It is my understanding

199
00:08:14,400 --> 00:08:16,966
that there was kind of an out on that

200
00:08:16,966 --> 00:08:18,633
that if you had a good reason why you

201
00:08:18,633 --> 00:08:21,000
didn't feel you could submit a data

202
00:08:21,000 --> 00:08:22,600
management plan you could just write

203
00:08:22,600 --> 00:08:25,600
that in "here's why I can't". I suspect a

204
00:08:25,600 --> 00:08:27,633
lot of people did that. That's just kind

205
00:08:27,633 --> 00:08:29,833
of a guess of mine, but that's a guess

206
00:08:29,833 --> 00:08:31,533
based on a lot of experience of dealing

207
00:08:31,533 --> 00:08:32,733
with people over the last five or six

208
00:08:32,733 --> 00:08:35,200
years. So I don't think they were very

209
00:08:35,200 --> 00:08:38,400
stringent about it at the time. And one

210
00:08:38,400 --> 00:08:39,266
of the things that's important to

211
00:08:39,266 --> 00:08:39,633
remember

212
00:08:39,633 --> 00:08:40,900
which I'm gonna mention now is that when

213
00:08:40,900 --> 00:08:42,000
they talk... How many people have actually

214
00:08:42,000 --> 00:08:43,666
done a data management plan for a grant

215
00:08:43,666 --> 00:08:47,066
application? Well you know a couple. So as

216
00:08:47,066 --> 00:08:48,700
you know they don't have to be long. And

217
00:08:48,700 --> 00:08:50,866
in fact they want them short. So this

218
00:08:50,866 --> 00:08:52,200
doesn't have to be when you right the

219
00:08:52,200 --> 00:08:54,433
DMP it doesn't have to be a long arduous

220
00:08:54,433 --> 00:08:57,233
process. Some of the agencies actually

221
00:08:57,233 --> 00:08:59,966
stipulate no more than two pages. They

222
00:08:59,966 --> 00:09:02,333
want certain elements but they don't

223
00:09:02,333 --> 00:09:04,000
want it to actually be long. So you don't

224
00:09:04,000 --> 00:09:05,333
have to think of this as an arduous

225
00:09:05,333 --> 00:09:09,400
thing. So it first came up in 2003. For

226
00:09:09,400 --> 00:09:11,133
years and years it kind of just stayed

227
00:09:11,133 --> 00:09:14,800
the same. In February of 2015

228
00:09:14,800 --> 00:09:17,300
NIH published a white paper called "The

229
00:09:17,300 --> 00:09:19,633
Plan for Increasing Access to Scientific

230
00:09:19,633 --> 00:09:21,866
Publications and Digital Scientific Data".

231
00:09:21,866 --> 00:09:27,066
From NIH funded scientific research in

232
00:09:27,066 --> 00:09:28,900
the white paper they came up with some

233
00:09:28,900 --> 00:09:31,000
suggestions. They wanted full

234
00:09:31,000 --> 00:09:32,900
descriptions of the data and how the

235
00:09:32,900 --> 00:09:34,933
data was collected. All the different

236
00:09:34,933 --> 00:09:36,466
types of data you expected to collect

237
00:09:36,466 --> 00:09:39,800
and how you collected that data. They

238
00:09:39,800 --> 00:09:41,666
want to know what software tools you use

239
00:09:41,666 --> 00:09:44,466
to analyze or create or produce final

240
00:09:44,466 --> 00:09:46,900
results of the data. That's important

241
00:09:46,900 --> 00:09:48,500
because there's been a lot of experience

242
00:09:48,500 --> 00:09:50,700
of people going back and trying to use

243
00:09:50,700 --> 00:09:53,533
data from 10, 15, 20 years ago and they

244
00:09:53,533 --> 00:09:55,600
cannot locate the software that was used

245
00:09:55,600 --> 00:09:59,333
at the time. So the data gets a little

246
00:09:59,333 --> 00:10:00,933
useless at that point and I've actually

247
00:10:00,933 --> 00:10:03,366
read accounts of people who had to spend

248
00:10:03,366 --> 00:10:04,833
a good deal of their funding research

249
00:10:04,833 --> 00:10:08,200
funds recreating the software that

250
00:10:08,200 --> 00:10:11,433
was used to use the old data again. So

251
00:10:11,433 --> 00:10:13,733
it's saying saving the software along

252
00:10:13,733 --> 00:10:16,666
with the data is an access issue. You can

253
00:10:16,666 --> 00:10:18,300
store it somewhere, you also want to make

254
00:10:18,300 --> 00:10:20,400
sure people can use it. So they want to

255
00:10:20,400 --> 00:10:22,966
know that. What protocols or steps you

256
00:10:22,966 --> 00:10:25,600
have used to create the data. Again this

257
00:10:25,600 --> 00:10:27,266
is so somebody later on down the line

258
00:10:27,266 --> 00:10:30,333
can recreate the research. How you ensure

259
00:10:30,333 --> 00:10:32,100
the long-term preservation of your data.

260
00:10:32,100 --> 00:10:33,966
Where you're going to store it, what

261
00:10:33,966 --> 00:10:35,833
their policies are in terms of backup

262
00:10:35,833 --> 00:10:38,500
how long it will be there. And how you

263
00:10:38,500 --> 00:10:41,566
provide access to the data. Specifically

264
00:10:41,566 --> 00:10:44,100
are you going to limit access for some

265
00:10:44,100 --> 00:10:46,700
reason? Now you're allowed to do that, but

266
00:10:46,700 --> 00:10:48,366
you have to say why you're going to do

267
00:10:48,366 --> 00:10:49,900
it. So for instance if you're going to be

268
00:10:49,900 --> 00:10:51,800
using some of this research to apply for

269
00:10:51,800 --> 00:10:52,466
a patent

270
00:10:52,466 --> 00:10:54,366
you might not want everyone getting into

271
00:10:54,366 --> 00:10:56,700
that research. So that's a legitimate

272
00:10:56,700 --> 00:10:58,233
reason they just want to know what the

273
00:10:58,233 --> 00:11:01,300
access policy is going to be. So this is

274
00:11:01,300 --> 00:11:03,066
basically what they've proposed in the

275
00:11:03,066 --> 00:11:07,033
white paper. Then in 2016 they put out an

276
00:11:07,033 --> 00:11:09,933
RFI for strategies for how they could

277
00:11:09,933 --> 00:11:13,566
implement these ideas into the community.

278
00:11:13,566 --> 00:11:16,533
Basically they were looking to

279
00:11:16,533 --> 00:11:18,166
solicit the following information: How

280
00:11:18,166 --> 00:11:20,233
the data should be managed, how the data

281
00:11:20,233 --> 00:11:22,000
should be made publicly available, and

282
00:11:22,000 --> 00:11:24,166
how to set standards for citing the

283
00:11:24,166 --> 00:11:26,366
shared data and software. We're not going

284
00:11:26,366 --> 00:11:29,066
to talk a lot about citing now but that

285
00:11:29,066 --> 00:11:31,133
is an issue. Like any other publication

286
00:11:31,133 --> 00:11:34,700
you want your data to be locatable.

287
00:11:34,700 --> 00:11:37,066
anybody you used the data Citation Index

288
00:11:37,066 --> 00:11:40,200
yet? Curious. Anybody know... what about... yeah

289
00:11:40,200 --> 00:11:42,966
it's not a lot to use. So there are

290
00:11:42,966 --> 00:11:45,533
two databases now basically. The data

291
00:11:45,533 --> 00:11:47,500
Citation Index is a product of the

292
00:11:47,500 --> 00:11:49,300
people who put out Thomson Reuters the

293
00:11:49,300 --> 00:11:50,166
people who put out the web of science.

294
00:11:50,166 --> 00:11:52,933
And Base is just a database of datasets

295
00:11:52,933 --> 00:11:55,000
and you can go in there and locate.

296
00:11:55,000 --> 00:11:58,400
They do a very good job of trawling over

297
00:11:58,400 --> 00:12:01,766
200 different repositories and uploading

298
00:12:01,766 --> 00:12:04,200
and indexing the information. So

299
00:12:04,200 --> 00:12:07,433
basically citing the data using

300
00:12:07,433 --> 00:12:08,833
metadata, which we'll talk a lot a little

301
00:12:08,833 --> 00:12:10,833
bit to cite the data is important as

302
00:12:10,833 --> 00:12:14,633
well. So these are the issues they wanted

303
00:12:14,633 --> 00:12:17,466
one of the areas they've actually been

304
00:12:17,466 --> 00:12:21,833
pretty serious about and enforcing their

305
00:12:21,833 --> 00:12:25,166
policies on, is genomic data sharing. I

306
00:12:25,166 --> 00:12:27,300
assume and this is just an assumption of

307
00:12:27,300 --> 00:12:29,033
mine I don't I'm not involved in medical

308
00:12:29,033 --> 00:12:31,100
research, but from what I've learned

309
00:12:31,100 --> 00:12:33,766
about data in general, I'm assuming that

310
00:12:33,766 --> 00:12:35,100
they went here first because there's a

311
00:12:35,100 --> 00:12:37,666
lot of it going on. So that it was kind

312
00:12:37,666 --> 00:12:39,633
of an easy place to go first and set

313
00:12:39,633 --> 00:12:43,100
some standards. So right now the NIH

314
00:12:43,100 --> 00:12:45,333
wants any large-scale data dealing

315
00:12:45,333 --> 00:12:47,833
with any of these GWAS, SNPs,

316
00:12:47,833 --> 00:12:50,733
genomic sequencing, metagenomic,

317
00:12:50,733 --> 00:12:53,800
epigenomic gene expressions... They do want

318
00:12:53,800 --> 00:12:56,033
the requirements are that you must

319
00:12:56,033 --> 00:12:58,266
submit a genomic data sharing plan with

320
00:12:58,266 --> 00:13:01,300
your request and that you must agree to

321
00:13:01,300 --> 00:13:03,100
share the data no later than the date of

322
00:13:03,100 --> 00:13:04,133
publication.

323
00:13:04,133 --> 00:13:06,033
Again I don't know how much they're

324
00:13:06,033 --> 00:13:08,166
enforcing this. Has anybody encountered

325
00:13:08,166 --> 00:13:11,566
this I'm kind of curious? Not yet, okay.

326
00:13:11,566 --> 00:13:13,400
So but that's one of the policies

327
00:13:13,400 --> 00:13:14,533
they're actually trying to enforce

328
00:13:14,533 --> 00:13:16,366
fairly strictly right now. And again I

329
00:13:16,366 --> 00:13:17,800
assume they've just gone there because

330
00:13:17,800 --> 00:13:19,166
there's so much of this research going

331
00:13:19,166 --> 00:13:20,933
on it's a good place to start.

332
00:13:20,933 --> 00:13:24,966
So in addition to any requirements that

333
00:13:24,966 --> 00:13:28,366
might be issued by the granting agency

334
00:13:28,366 --> 00:13:29,733
and one of the things we'll do later on

335
00:13:29,733 --> 00:13:31,433
to show you how to get that information

336
00:13:31,433 --> 00:13:34,433
it's pretty easy to get. But we're again

337
00:13:34,433 --> 00:13:35,800
we're concentrating on the NIH here just

338
00:13:35,800 --> 00:13:37,600
because we assume that's the big fish in

339
00:13:37,600 --> 00:13:40,600
the pond here. Many publishers are now

340
00:13:40,600 --> 00:13:43,000
also coming up with policies. They

341
00:13:43,000 --> 00:13:45,600
usually will not conflict

342
00:13:45,600 --> 00:13:47,166
with any policies put out by the

343
00:13:47,166 --> 00:13:49,433
granting agency. You just simply want to

344
00:13:49,433 --> 00:13:50,800
be aware where you're going to be

345
00:13:50,800 --> 00:13:52,700
publishing and if they have any kind of

346
00:13:52,700 --> 00:13:55,233
data sharing policy associated with

347
00:13:55,233 --> 00:13:58,033
their publication practices. One of the

348
00:13:58,033 --> 00:13:59,400
first ones actually come up with a

349
00:13:59,400 --> 00:14:01,133
strict one was PLOS - the Public Library

350
00:14:01,133 --> 00:14:03,333
of Online Science - which is an open

351
00:14:03,333 --> 00:14:06,533
access platform where a lot of science

352
00:14:06,533 --> 00:14:09,766
stuff gets published. And they basically

353
00:14:09,766 --> 00:14:11,766
have said refusal to share the data in

354
00:14:11,766 --> 00:14:13,500
accordance with their policy will be

355
00:14:13,500 --> 00:14:16,500
grounds for rejection. Their policy

356
00:14:16,500 --> 00:14:19,000
is that the authors must show proof that

357
00:14:19,000 --> 00:14:21,733
they have shared their data somewhere by

358
00:14:21,733 --> 00:14:23,466
providing the journal with a unique

359
00:14:23,466 --> 00:14:26,300
identifier. Basically a DOI. If the proof

360
00:14:26,300 --> 00:14:28,233
is not provided PLOS can reject the

361
00:14:28,233 --> 00:14:30,900
paper outright or retract it if it's

362
00:14:30,900 --> 00:14:32,366
already been published and an author

363
00:14:32,366 --> 00:14:35,100
removes the data from public view. So you

364
00:14:35,100 --> 00:14:37,233
must specify that the data are deposited

365
00:14:37,233 --> 00:14:39,766
publicly and list the name or names of

366
00:14:39,766 --> 00:14:42,133
the repositories along with the DOI or

367
00:14:42,133 --> 00:14:44,233
an accession number. And they're pretty

368
00:14:44,233 --> 00:14:46,733
strict about that. Which makes sense

369
00:14:46,733 --> 00:14:48,400
they're open access, they're going to be

370
00:14:48,400 --> 00:14:50,200
one of the first people to go here. But

371
00:14:50,200 --> 00:14:52,666
other publishers are coming on board

372
00:14:52,666 --> 00:14:55,266
with this. Nature Science, Cell have

373
00:14:55,266 --> 00:14:56,900
language written into their policies

374
00:14:56,900 --> 00:14:59,533
that state that data necessary to

375
00:14:59,533 --> 00:15:01,966
understand, assess, and extend the

376
00:15:01,966 --> 00:15:04,100
conclusions of a manuscript must be

377
00:15:04,100 --> 00:15:06,533
shared and proof of this must be

378
00:15:06,533 --> 00:15:08,933
required specifically for genomic data.

379
00:15:08,933 --> 00:15:11,133
All other kinds of data they don't

380
00:15:11,133 --> 00:15:13,533
require proof yet like PLOS does. PLOS

381
00:15:13,533 --> 00:15:16,266
wants it for everything. Nature, Cellpress

382
00:15:16,266 --> 00:15:17,666
Science are just asking for

383
00:15:17,666 --> 00:15:19,300
but now but they're probably going to be

384
00:15:19,300 --> 00:15:23,266
moving more in the area of... there's

385
00:15:23,266 --> 00:15:25,000
basically broad discussions going on in

386
00:15:25,000 --> 00:15:26,866
the publishing community that are going

387
00:15:26,866 --> 00:15:28,366
to point to greater enforcement of this

388
00:15:28,366 --> 00:15:33,200
for all data sets. A good example of

389
00:15:33,200 --> 00:15:35,100
publisher requirements that are coming

390
00:15:35,100 --> 00:15:36,700
on board now that you need to be aware

391
00:15:36,700 --> 00:15:38,700
of, there's one that came from the

392
00:15:38,700 --> 00:15:40,100
International Committee of Medical

393
00:15:40,100 --> 00:15:42,466
Journal Editors that released a proposal

394
00:15:42,466 --> 00:15:45,333
in January 2016 to require that

395
00:15:45,333 --> 00:15:47,933
de-identify patient data from clinical

396
00:15:47,933 --> 00:15:50,100
trials that underly a journal article

397
00:15:50,100 --> 00:15:52,933
must be made public within six months of

398
00:15:52,933 --> 00:15:55,800
publication. So this group consists of

399
00:15:55,800 --> 00:15:57,333
JAMA, The Lancet, and the New England

400
00:15:57,333 --> 00:15:59,100
Journal of Medicine, these are all

401
00:15:59,100 --> 00:16:01,666
prominent journals. This is just a good

402
00:16:01,666 --> 00:16:03,433
example of the policies that are coming

403
00:16:03,433 --> 00:16:05,033
on board that you're going to need to be

404
00:16:05,033 --> 00:16:06,966
aware of if you want to publish in these

405
00:16:06,966 --> 00:16:10,633
journals. So a quick quiz. Just shout out

406
00:16:10,633 --> 00:16:12,466
the right answer. According to multiple

407
00:16:12,466 --> 00:16:15,266
publisher requirements, all data

408
00:16:15,266 --> 00:16:18,333
necessary to which of these must be

409
00:16:18,333 --> 00:16:20,933
available? A) Understand, B) Assess, C)

410
00:16:20,933 --> 00:16:25,933
Extend, or D) All the above. D - see you paid

411
00:16:25,933 --> 00:16:28,833
attention that's great! So the point here

412
00:16:28,833 --> 00:16:31,333
is that your funding agencies are going

413
00:16:31,333 --> 00:16:33,066
to have requirements increasingly so in

414
00:16:33,066 --> 00:16:35,500
the future so will the publishers. My

415
00:16:35,500 --> 00:16:36,600
experience is that they don't normally

416
00:16:36,600 --> 00:16:39,300
conflict. They're pretty much going to be

417
00:16:39,300 --> 00:16:40,866
the same. You just have to be aware that

418
00:16:40,866 --> 00:16:42,600
you have to deal with this on each end.

419
00:16:42,600 --> 00:16:45,166
It's fairly easy to find out that

420
00:16:45,166 --> 00:16:47,333
information. We learn one of the things

421
00:16:47,333 --> 00:16:49,200
we can do if you are having trouble

422
00:16:49,200 --> 00:16:51,100
finding any information about that

423
00:16:51,100 --> 00:16:53,200
contact us we can help you get to the

424
00:16:53,200 --> 00:16:58,200
right place.

425
00:16:58,200 --> 00:17:02,433
In January of 2016 the NIH added the

426
00:17:02,433 --> 00:17:05,333
requirement that grant applications must

427
00:17:05,333 --> 00:17:07,766
address a number of issues related to

428
00:17:07,766 --> 00:17:09,566
the rigor and reproducibility of

429
00:17:09,566 --> 00:17:13,566
research. So we've got these new

430
00:17:13,566 --> 00:17:16,300
guidelines that were started in January

431
00:17:16,300 --> 00:17:19,866
of 2016. And the most important thing is

432
00:17:19,866 --> 00:17:23,233
to start paying attention here to some

433
00:17:23,233 --> 00:17:24,733
of the percentages that we're going to

434
00:17:24,733 --> 00:17:28,300
talk about here, and the numbers that

435
00:17:28,300 --> 00:17:31,700
we're addressing. These requirements grew

436
00:17:31,700 --> 00:17:34,666
out of a growing concern about a lack of

437
00:17:34,666 --> 00:17:37,800
reproducibility in many areas of science

438
00:17:37,800 --> 00:17:41,466
particularly biomedical research. So the

439
00:17:41,466 --> 00:17:44,266
first example here is when Bayer sought

440
00:17:44,266 --> 00:17:47,133
to validate published results on

441
00:17:47,133 --> 00:17:49,266
potential drug targets they

442
00:17:49,266 --> 00:17:51,633
determined that the data collected was

443
00:17:51,633 --> 00:17:53,500
consistent with the published literature

444
00:17:53,500 --> 00:17:58,866
only 21% of the time. So our first

445
00:17:58,866 --> 00:18:06,400
example here is 21%.

446
00:18:06,400 --> 00:18:11,333
Our second example here is Amgen.

447
00:18:11,333 --> 00:18:14,266
So Amgen, another drug company, tried to

448
00:18:14,266 --> 00:18:17,466
reproduce findings in 53 landmark

449
00:18:17,466 --> 00:18:19,900
preclinical cancer studies and they were

450
00:18:19,900 --> 00:18:23,533
only able to confirm the findings of 11%

451
00:18:23,533 --> 00:18:28,000
of those studies. So first we had 21% now

452
00:18:28,000 --> 00:18:32,300
we have 11%. And these are not just in

453
00:18:32,300 --> 00:18:35,133
drug trials. The open science

454
00:18:35,133 --> 00:18:37,300
collaboration group together to

455
00:18:37,300 --> 00:18:39,266
duplicate a hundred experiments

456
00:18:39,266 --> 00:18:42,200
published in 2008 in three high-ranking

457
00:18:42,200 --> 00:18:44,900
psychology journals. They found that they

458
00:18:44,900 --> 00:18:47,200
were able to reproduce the exact same

459
00:18:47,200 --> 00:18:52,600
results in only 39 cases. Of those cases

460
00:18:52,600 --> 00:18:53,833
that didn't match, some were

461
00:18:53,833 --> 00:19:00,600
significantly off. And 15 showed that

462
00:19:00,600 --> 00:19:02,400
the results were not similar at all.

463
00:19:02,400 --> 00:19:05,000
These studies and many others raised

464
00:19:05,000 --> 00:19:08,133
concerns at the NIH and plans emerged to

465
00:19:08,133 --> 00:19:09,966
adjust the NIH grant submission

466
00:19:09,966 --> 00:19:11,933
guidelines to enhance reproducibility.

467
00:19:11,933 --> 00:19:13,500
Whether you believe there's a

468
00:19:13,500 --> 00:19:15,600
reproducibility crisis or not you will

469
00:19:15,600 --> 00:19:18,100
have to adhere to NIH's new guidelines

470
00:19:18,100 --> 00:19:21,433
regarding rigor and reproducibility. From

471
00:19:21,433 --> 00:19:23,433
a data management perspective it's a

472
00:19:23,433 --> 00:19:25,266
guideline for scientific rigor that's

473
00:19:25,266 --> 00:19:27,766
relevant. In NIH's description of what's

474
00:19:27,766 --> 00:19:30,266
needed to ensure scientific rigor they

475
00:19:30,266 --> 00:19:33,033
state that it includes full transparency

476
00:19:33,033 --> 00:19:35,500
and report in reporting experimental

477
00:19:35,500 --> 00:19:41,766
details. What this means is that saving

478
00:19:41,766 --> 00:19:44,266
the data produced by experiment is not

479
00:19:44,266 --> 00:19:47,000
sufficient. The data will not allow

480
00:19:47,000 --> 00:19:49,666
others to reproduce your work. In

481
00:19:49,666 --> 00:19:52,600
thinking about data management it's now

482
00:19:52,600 --> 00:19:56,700
necessary to take a broader perspective.

483
00:19:56,700 --> 00:19:59,966
So we have to preserve the lab notebook.

484
00:19:59,966 --> 00:20:03,533
And the experimental workflow must also

485
00:20:03,533 --> 00:20:08,333
be documented and preserved. And now also

486
00:20:08,333 --> 00:20:10,366
the computer programs must be saved

487
00:20:10,366 --> 00:20:12,933
including the information about

488
00:20:12,933 --> 00:20:16,133
parameters, versions, operating systems,

489
00:20:16,133 --> 00:20:18,300
and anything else that's necessary to

490
00:20:18,300 --> 00:20:23,400
reproduce the results. So for a clinical

491
00:20:23,400 --> 00:20:25,733
trial the protocol must be saved

492
00:20:25,733 --> 00:20:27,800
including the study design, the

493
00:20:27,800 --> 00:20:30,200
intervention, the inclusion, and

494
00:20:30,200 --> 00:20:33,800
exclusion criteria. So, in short, the data

495
00:20:33,800 --> 00:20:36,600
management encompasses ensuring that

496
00:20:36,600 --> 00:20:38,666
everything needed to reproduce a study

497
00:20:38,666 --> 00:20:43,733
has been preserved. So there's now some

498
00:20:43,733 --> 00:20:47,600
data danger zones we have to talk about.

499
00:20:47,600 --> 00:20:51,333
So let's see if I can get this to work

500
00:20:51,333 --> 00:20:55,966
our next portion. So I have a little

501
00:20:55,966 --> 00:21:03,633
video that hopefully will play...

502
00:21:03,633 --> 00:21:07,833
Hello my name is Dr. Judy Benign. I'm an

503
00:21:07,833 --> 00:21:10,566
oncologist at NYU School of Medicine.

504
00:21:10,566 --> 00:21:13,266
Hello Dr. Judy Benign. I read your

505
00:21:13,266 --> 00:21:16,266
article on b-cell function. I think that

506
00:21:16,266 --> 00:21:18,100
I could use the data for my work on

507
00:21:18,100 --> 00:21:22,200
pancreatic cancer. I am NOT an oncologist.

508
00:21:22,200 --> 00:21:24,433
I know but I think I could use the data

509
00:21:24,433 --> 00:21:27,666
for my work on pancreatic cancer. Do you

510
00:21:27,666 --> 00:21:29,800
have the data? Everything you need to

511
00:21:29,800 --> 00:21:32,966
know is in the article. No. What I need is

512
00:21:32,966 --> 00:21:36,633
the data. Will you share your data? I am

513
00:21:36,633 --> 00:21:39,166
not sure that will be possible. But your

514
00:21:39,166 --> 00:21:41,366
work is in PubMed Central and was funded

515
00:21:41,366 --> 00:21:44,700
by NIH. That is true. And it was published

516
00:21:44,700 --> 00:21:47,266
in Science which requires that you share

517
00:21:47,266 --> 00:21:50,100
your data. I did publish in Science. Then

518
00:21:50,100 --> 00:21:52,700
I am requesting your data.

519
00:21:52,700 --> 00:21:58,033
Can I have a copy of your data?

520
00:21:58,033 --> 00:22:00,633
I am not sure where my data is. But

521
00:22:00,633 --> 00:22:03,900
surely you saved your data! I did. I saved

522
00:22:03,900 --> 00:22:08,466
it on a USB Drive. Where is the USB Drive?

523
00:22:08,466 --> 00:22:12,900
It is in a box.

524
00:22:12,900 --> 00:22:16,100
It is in a box at home. I just moved.

525
00:22:16,100 --> 00:22:18,233
But can I use your data?

526
00:22:18,233 --> 00:22:24,533
There are many boxes...so many boxes!

527
00:22:24,533 --> 00:22:32,700
I forgot to label the boxes!

528
00:22:32,700 --> 00:22:35,200

529
00:22:35,200 --> 00:22:38,266
Hello again. Thank you for sending me a

530
00:22:38,266 --> 00:22:40,900
copy of your data on a USB Drive.

531
00:22:40,900 --> 00:22:44,033
I received the envelope yesterday. You

532
00:22:44,033 --> 00:22:46,233
were welcome. But I will need that back

533
00:22:46,233 --> 00:22:48,766
when you are finished. That is my only

534
00:22:48,766 --> 00:22:50,200
copy.

535
00:22:50,200 --> 00:22:52,600
I did have a question. What is your

536
00:22:52,600 --> 00:22:55,800
question? You might find the answer in my

537
00:22:55,800 --> 00:22:59,266
article. No. I received the data but when

538
00:22:59,266 --> 00:23:02,466
I opened it up it was in hexadecimal. Yes

539
00:23:02,466 --> 00:23:04,266
that is right.

540
00:23:04,266 --> 00:23:07,100
I cannot read hexadecimal. You asked for

541
00:23:07,100 --> 00:23:10,100
my data and I gave it to you. I have done

542
00:23:10,100 --> 00:23:12,200
what you asked.

543
00:23:12,200 --> 00:23:13,733
But is there a way to read the

544
00:23:13,733 --> 00:23:16,066
hexadecimal? You will need the program

545
00:23:16,066 --> 00:23:19,300
that created the hexadecimal file. Yes. I

546
00:23:19,300 --> 00:23:21,900
will. What is the name of the program?

547
00:23:21,900 --> 00:23:23,266
Saito synth.

548
00:23:23,266 --> 00:23:26,100
I do not know this program. It was a very

549
00:23:26,100 --> 00:23:29,033
good program. The company that made the

550
00:23:29,033 --> 00:23:32,400
program went bankrupt in 2007. Do you

551
00:23:32,400 --> 00:23:34,866
have a copy of the program? I do not use

552
00:23:34,866 --> 00:23:37,133
this program anymore because the company

553
00:23:37,133 --> 00:23:40,600
that made it went bankrupt. Maybe you can

554
00:23:40,600 --> 00:23:49,133
buy a copy on eBay?

555
00:23:49,133 --> 00:23:52,100
I have good news.

556
00:23:52,100 --> 00:23:55,400
You again? I talked to my colleague. She

557
00:23:55,400 --> 00:23:57,033
knew a person with a copy of the

558
00:23:57,033 --> 00:23:59,400
software. Then why do you need me!?

559
00:23:59,400 --> 00:24:01,633
Everything you need to know about the

560
00:24:01,633 --> 00:24:04,000
data is in the article. I opened the data

561
00:24:04,000 --> 00:24:06,433
and I could not understand it. If you

562
00:24:06,433 --> 00:24:08,266
have the program you will find it as

563
00:24:08,266 --> 00:24:11,233
clear. Well, I noticed that you called

564
00:24:11,233 --> 00:24:13,966
your data fields Sam. Is that an

565
00:24:13,966 --> 00:24:16,800
abbreviation? Yes! It is an abbreviation

566
00:24:16,800 --> 00:24:20,300
of my co-authors name. His name is Samuel

567
00:24:20,300 --> 00:24:24,600
Li. We call him Sam. I see. And what is

568
00:24:24,600 --> 00:24:26,500
the content of the field called?

569
00:24:26,500 --> 00:24:29,966
Sam1? Ah, yes! Sam1 is the level of

570
00:24:29,966 --> 00:24:33,100
CXCR4 expression. And what is the content

571
00:24:33,100 --> 00:24:36,033
of the field called Sam2? That is

572
00:24:36,033 --> 00:24:39,100
logical if you think about it. What is

573
00:24:39,100 --> 00:24:45,066
the content of the field called Sam2?

574
00:24:45,066 --> 00:24:51,366
I don't remember. What about Sam3?

575
00:24:51,366 --> 00:24:53,700
Is there a guide to the data anywhere?

576
00:24:53,700 --> 00:24:57,200
Yes of course. It is the article that is

577
00:24:57,200 --> 00:25:00,700
published in Science. The article does

578
00:25:00,700 --> 00:25:03,000
not tell me what the field names mean. Is

579
00:25:03,000 --> 00:25:05,200
there any record of what these field

580
00:25:05,200 --> 00:25:06,900
names mean? Yes.

581
00:25:06,900 --> 00:25:09,533
My co-author knows what the content of

582
00:25:09,533 --> 00:25:15,866
Sam2 is, and Sam3.

583
00:25:15,866 --> 00:25:20,500
And Sam4. Can I talk to your co-author?

584
00:25:20,500 --> 00:25:23,000
I'm not sure. I would very much like to

585
00:25:23,000 --> 00:25:25,866
talk to your co-author. Well he was a

586
00:25:25,866 --> 00:25:28,600
graduate student. He went back to China

587
00:25:28,600 --> 00:25:32,233
two years ago. Can I have his contact

588
00:25:32,233 --> 00:25:34,333
information?

589
00:25:34,333 --> 00:25:38,866
He is in China. His name is Sam Li.

590
00:25:38,866 --> 00:25:41,466
I think I cannot use your data. You could

591
00:25:41,466 --> 00:25:43,233
check the article to see if what you

592
00:25:43,233 --> 00:25:45,866
need is there? Please stop talking now.

593
00:25:45,866 --> 00:25:48,400
All right. But it almost showed a little

594
00:25:48,400 --> 00:25:50,200
bit without having even less information

595
00:25:50,200 --> 00:25:53,366
the confusion that sometimes by not

596
00:25:53,366 --> 00:25:56,800
recording, you can end up with. But I will

597
00:25:56,800 --> 00:26:00,200
be sure to - when we send out our email

598
00:26:00,200 --> 00:26:01,633
with our links and everything - to include

599
00:26:01,633 --> 00:26:03,366
the full video so that you can not only

600
00:26:03,366 --> 00:26:06,966
see some of those links but you can see

601
00:26:06,966 --> 00:26:09,833
the full video as well. But that does

602
00:26:09,833 --> 00:26:11,566
illustrate some of the common challenges

603
00:26:11,566 --> 00:26:14,466
that you can face by not properly

604
00:26:14,466 --> 00:26:17,333
managing all of your data. And some of

605
00:26:17,333 --> 00:26:19,833
the snafus that can occur after the

606
00:26:19,833 --> 00:26:24,100
fact when data is not properly managed.

607
00:26:24,100 --> 00:26:26,333
Alright so when you think about all of

608
00:26:26,333 --> 00:26:27,866
the steps you take when creating,

609
00:26:27,866 --> 00:26:30,566
processing, and analyzing your data you

610
00:26:30,566 --> 00:26:32,600
want to avoid a scenario where all your

611
00:26:32,600 --> 00:26:35,633
work is tied up in a USB Drive that only

612
00:26:35,633 --> 00:26:37,733
you and sometimes not even you can

613
00:26:37,733 --> 00:26:40,200
understand. You don't want your data to

614
00:26:40,200 --> 00:26:44,366
end up with Sam Li in China. In

615
00:26:44,366 --> 00:26:48,400
particular. And so now what I'd like you

616
00:26:48,400 --> 00:26:49,966
all to do is think about your current

617
00:26:49,966 --> 00:26:51,666
workflow and ask yourself the following

618
00:26:51,666 --> 00:26:56,033
questions. Can you easily locate your raw

619
00:26:56,033 --> 00:27:00,700
data? Can you understand it? Can you

620
00:27:00,700 --> 00:27:02,900
understand the processes that you took

621
00:27:02,900 --> 00:27:05,233
to get the raw data to the processed

622
00:27:05,233 --> 00:27:08,266
data? And what tools or versions of

623
00:27:08,266 --> 00:27:10,200
software did you use to get it to those

624
00:27:10,200 --> 00:27:17,233
stages? Can you connect the different

625
00:27:17,233 --> 00:27:19,733
types of related data you collected? If

626
00:27:19,733 --> 00:27:22,333
those collected forms of data and

627
00:27:22,333 --> 00:27:25,033
imaging data that you that are related

628
00:27:25,033 --> 00:27:27,166
is it all easy to track down?

629
00:27:27,166 --> 00:27:29,833
So can you connect your steps? Could you

630
00:27:29,833 --> 00:27:31,766
figure out how you got things from one

631
00:27:31,766 --> 00:27:33,433
step to the next?

632
00:27:33,433 --> 00:27:36,800
And are your naming conventions

633
00:27:36,800 --> 00:27:39,466
consistent with others on your team? And

634
00:27:39,466 --> 00:27:41,833
finally do you name your subjects,

635
00:27:41,833 --> 00:27:43,766
animals, and specimens, in a consistent

636
00:27:43,766 --> 00:27:46,033
way just as everyone else on your team

637
00:27:46,033 --> 00:27:48,500
does? Is it a documentation that would

638
00:27:48,500 --> 00:27:49,933
allow you to share your data with

639
00:27:49,933 --> 00:27:52,100
someone else, and would they know exactly

640
00:27:52,100 --> 00:27:55,600
how you named something? So could you

641
00:27:55,600 --> 00:27:56,600
share it and have somebody actually

642
00:27:56,600 --> 00:27:58,733
understand it not just know

643
00:27:58,733 --> 00:28:01,100
what Sam1 is. You want to ask

644
00:28:01,100 --> 00:28:02,600
yourself these questions because you

645
00:28:02,600 --> 00:28:04,333
don't want a simple data management

646
00:28:04,333 --> 00:28:06,033
error to be a cause of something really

647
00:28:06,033 --> 00:28:08,366
bad to happen. So take this

648
00:28:08,366 --> 00:28:10,000
article from the New England Journal of

649
00:28:10,000 --> 00:28:12,033
Medicine - the editors found multiple

650
00:28:12,033 --> 00:28:16,900
errors within a table in the paper. Even

651
00:28:16,900 --> 00:28:18,566
though the errors didn't alter the

652
00:28:18,566 --> 00:28:20,833
conclusion of the article, the authors

653
00:28:20,833 --> 00:28:22,733
couldn't find the primary data they

654
00:28:22,733 --> 00:28:24,700
collected. Because they couldn't locate

655
00:28:24,700 --> 00:28:26,933
the primary data, their paper was

656
00:28:26,933 --> 00:28:30,133
ultimately retracted. And this is a

657
00:28:30,133 --> 00:28:32,300
perfect example of where poor data

658
00:28:32,300 --> 00:28:35,700
management had really significant

659
00:28:35,700 --> 00:28:45,333
consequences. [Title slide indicating that this is the end of Part 1 of this series. Please continue with Part 2 of 3.]

660
00:28:45,333 --> 00:28:47,399
[Title slide indicating that this is the end of Part 1 of this series. Please continue with Part 2 of 3.]