
VISITOR
Ames Research Center

1
00:00:01,000 --> 00:00:07,000
[music playing]

2
00:00:16,766 --> 00:00:21,300
- Welcome to the 2016
NASA Ames Summer Series.

3
00:00:22,366 --> 00:00:25,800
If a tree falls in the forest

4
00:00:25,800 --> 00:00:27,800
and no one is around to hear it,

5
00:00:27,800 --> 00:00:30,266
does it make a sound?

6
00:00:30,266 --> 00:00:33,200
Another way to look
at that question:

7
00:00:33,200 --> 00:00:35,533
if a species erase

8
00:00:35,533 --> 00:00:38,900
or a data point disappears

9
00:00:38,900 --> 00:00:41,966
and we do not have
a record of it,

10
00:00:41,966 --> 00:00:44,533
did it exist?

11
00:00:44,533 --> 00:00:47,700
Or does it have any value

12
00:00:47,700 --> 00:00:50,433
in the results of where we seek?

13
00:00:50,433 --> 00:00:52,466
Today's presentation is entitled

14
00:00:52,466 --> 00:00:55,066
"Searching Harsh Environments"

15
00:00:55,066 --> 00:00:59,633
and will be given by
Dr. Ophir Frieder.

16
00:00:59,633 --> 00:01:01,900
He is a professor
of biostatistics,

17
00:01:01,900 --> 00:01:04,900
bioinformatics,
and biomathematics

18
00:01:04,900 --> 00:01:08,500
in the Georgetown University
Medical Center.

19
00:01:08,500 --> 00:01:10,100
He is--he also holds

20
00:01:10,100 --> 00:01:13,533
the Robert and Catherine
McDevitt Chair

21
00:01:13,533 --> 00:01:17,000
in Computer Science
and Information Processing.

22
00:01:19,700 --> 00:01:23,400
Besides that, he also holds
an external position

23
00:01:23,400 --> 00:01:26,966

as the Chief Scientific Officer
of UMBRA Health Corp.

24
00:01:28,533 --> 00:01:30,266
He received
a Bachelor's of Science

25
00:01:30,266 --> 00:01:32,700
in computer science
and communication studies,

26
00:01:32,700 --> 00:01:34,200
a Master's of Science

27
00:01:38,466 --> 00:01:35,800
in computer science
and engineering,

28
00:01:38,466 --> 00:01:41,833
all from
the University of Michigan.

29
00:01:41,833 --> 00:01:44,766
He is a Fellow of the AAAS,

30
00:01:44,766 --> 00:01:48,666
the ACM, IEEE, and NAI.

31
00:01:49,766 --> 00:01:51,066
Please welcome--

32
00:01:51,066 --> 00:01:52,933
please join me in welcoming

33
00:01:52,933 --> 00:01:54,433
Dr. Frieder.

34
00:01:54,433 --> 00:01:57,433
[applause]

35
00:02:00,700 --> 00:02:02,100
- Good morning,

36
00:02:02,100 --> 00:02:04,100
which is, for me,
a little odd to say

37
00:02:04,100 --> 00:02:05,866
considering it's really
good afternoon.

38
00:02:05,866 --> 00:02:08,866
But for you it's good morning.

39
00:02:08,866 --> 00:02:10,700
Thank you for being here.

40
00:02:10,700 --> 00:02:13,200
I see the lights
are nice and shiny.

41
00:02:13,200 --> 00:02:15,466
I'm not used to bright lights,

42
00:02:15,466 --> 00:02:17,300
but we'll do what we can do.

43
00:02:17,300 --> 00:02:21,300
So today's talk is about
searching in harsh environments.

44
00:02:21,300 --> 00:02:23,033
And before I came here,

45
00:02:23,033 --> 00:02:26,300
I looked at the variety of
talks that are gonna be given

46
00:02:26,300 --> 00:02:27,966
through the summer series.

47
00:02:27,966 --> 00:02:31,733
And this one is not quite
in the same spirit

48
00:02:31,733 --> 00:02:33,133
as some of the other ones.

49
00:02:33,133 --> 00:02:35,400
This will be a little different.

50
00:02:35,400 --> 00:02:38,200
This--but we'll see if,
hopefully,

51
00:02:38,200 --> 00:02:39,566
it'll pique your interest

52
00:02:39,566 --> 00:02:41,800
or at least something.

53
00:02:41,800 --> 00:02:44,066
All right, so first of all,
I have to correct you

54
00:02:44,066 --> 00:02:45,433
a well-known myth.

55
00:02:45,433 --> 00:02:47,600
How many of you use Google?

56
00:02:47,600 --> 00:02:49,633
Okay, considering I'm like
a stone's throw away,

57
00:02:49,633 --> 00:02:51,000

it would be pretty sad.

58
00:02:51,000 --> 00:02:53,133
How many of you basically
use Google today?

59
00:02:54,733 --> 00:02:56,733
Okay.
No surprise.

60
00:02:56,733 --> 00:02:58,266
Well, if you actually
give a talk

61
00:02:58,266 --> 00:02:59,766
in any search community,

62
00:02:59,766 --> 00:03:03,133
what you'll hear is that
Google solves it all.

63
00:03:03,133 --> 00:03:05,700
And for the most part,

64
00:03:05,700 --> 00:03:08,200
they actually have solved
quite a bit of it.

65
00:03:08,200 --> 00:03:11,100
But it isn't exactly the truth,

66
00:03:11,100 --> 00:03:14,800
because Google really solved
computerized data.

67
00:03:14,800 --> 00:03:17,400
And a lot of data
isn't computerized.

68

00:03:17,400 --> 00:03:20,200
It's digitized, and I'll show
you what I mean in a second.

69
00:03:20,200 --> 00:03:22,000
And more so,

70
00:03:22,000 --> 00:03:25,566
Google is hardly
a social media player.

71
00:03:25,566 --> 00:03:27,900
How many of you've actually used
Google Plus?

72
00:03:29,066 --> 00:03:30,500
Wow, there's actually
three of you.

73
00:03:30,500 --> 00:03:32,433
I'm impressed.

74
00:03:32,433 --> 00:03:35,166
I've given these talks
in large audiences

75
00:03:35,166 --> 00:03:37,466
and not even a single hand,

76
00:03:37,466 --> 00:03:40,600
so obviously, Google Plus
didn't quite make the splash

77
00:03:40,600 --> 00:03:42,200
they were hoping it to make,

78
00:03:42,200 --> 00:03:43,633
and obviously, it's--

79

00:03:43,633 --> 00:03:45,433
social media's here
all over the place.

80
00:03:45,433 --> 00:03:48,933
So Google is hardly
a social media player.

81
00:03:48,933 --> 00:03:51,600
So what am I gonna
talk to you about today?

82
00:03:51,600 --> 00:03:55,300
As you can see, I'm gonna talk
about a gamut of things.

83
00:03:55,300 --> 00:03:58,100
And the way I usually talk--
give a presentation is,

84
00:03:58,100 --> 00:04:01,433
I usually pick something that
I consider as more engineering,

85
00:04:01,433 --> 00:04:05,200
meaning that the solution
to a real problem exists--

86
00:04:05,200 --> 00:04:07,800
I put it together--

87
00:04:07,800 --> 00:04:11,133
by "me," I say my students
actually put it together--

88
00:04:11,133 --> 00:04:14,700
and basically is in real use.

89
00:04:14,700 --> 00:04:16,466
To the other extreme

is research,

90
00:04:16,466 --> 00:04:18,500
which is something that we
hope to do in the future;

91
00:04:18,500 --> 00:04:20,666
we hope to get it to go
into applied practice.

92
00:04:20,666 --> 00:04:23,666
And I'll talk to you about
searching and mining

93
00:04:23,666 --> 00:04:25,166
of social media.

94
00:04:25,166 --> 00:04:27,300
But first, I'm gonna start
talking to you about

95
00:04:27,300 --> 00:04:30,600
what is complex document
information processing.

96
00:04:30,600 --> 00:04:32,233
And by "complex information,"

97
00:04:32,233 --> 00:04:34,633
I try to find something
that would relate to NASA

98
00:04:34,633 --> 00:04:35,966
along the way.

99
00:04:35,966 --> 00:04:37,900
Now, I got to tell you
that I'm not an expert

100

00:04:37,900 --> 00:04:39,300
in anything that NASA does.

101
00:04:39,300 --> 00:04:40,833
In fact, I'm not even a novice.

102
00:04:40,833 --> 00:04:44,000
But I do realize that when you
actually look at the brochures

103
00:04:44,000 --> 00:04:45,600
and you actually look
at some of the stuff

104
00:04:45,600 --> 00:04:46,866
that you find online,

105
00:04:46,866 --> 00:04:51,166
you talk about
the "Mission Assurance System."

106
00:04:51,166 --> 00:04:54,066
And they talk about having
timeliness of information--

107
00:04:54,066 --> 00:04:56,133
that it will get you
complete information

108
00:04:56,133 --> 00:04:57,933
as quickly as you can.

109
00:04:57,933 --> 00:04:59,966
Now, complete information
really means that you need

110
00:04:59,966 --> 00:05:01,333
to get it from all sources,

111

00:05:01,333 --> 00:05:03,466
not just the standard
computerized as you know it.

112
00:05:03,466 --> 00:05:05,233
So I'm gonna talk about that,

113
00:05:05,233 --> 00:05:07,966
and obviously, I'm gonna talk
about the research direction

114
00:05:07,966 --> 00:05:10,000
in this area, so let me--

115
00:05:10,000 --> 00:05:13,133
by the way, feel free to raise
your hand and stop me.

116
00:05:14,566 --> 00:05:17,100
All right,
so this is a complex document.

117
00:05:17,100 --> 00:05:19,300
When I talk about what
is a complex document,

118
00:05:19,300 --> 00:05:20,900
I talk about
documents like this.

119
00:05:20,900 --> 00:05:23,966
And as you can see, this is not
your standard web document

120
00:05:23,966 --> 00:05:26,033
or isn't your standard
Microsoft Word document.

121
00:05:26,033 --> 00:05:29,133
It's got stamps,

it's got signatures,

122
00:05:29,133 --> 00:05:30,566
it's got logos--

123
00:05:30,566 --> 00:05:32,433
it's got a whole
variety of things,

124
00:05:32,433 --> 00:05:34,400
and that's pretty complex.

125
00:05:34,400 --> 00:05:38,066
But this one's
a little more complex.

126
00:05:38,066 --> 00:05:39,566
So this was
a government-funded project

127
00:05:39,566 --> 00:05:40,633
that I was working on,

128
00:05:40,633 --> 00:05:43,266
And this too has logos,

129
00:05:43,266 --> 00:05:45,066
and it too has a signature,

130
00:05:45,066 --> 00:05:48,400
and it also has
different type components,

131
00:05:48,400 --> 00:05:50,300
including tables
and handwriting,

132
00:05:50,300 --> 00:05:53,600
and by the way,

anybody can read this?

133
00:05:53,600 --> 00:05:55,100
No one can read this.

134
00:05:55,100 --> 00:05:58,833
Okay, anybody can read it
from left to right?

135
00:05:58,833 --> 00:06:00,333
I'm glad nobody said yes,

136
00:06:00,333 --> 00:06:02,333
because in Arabic,
you read from right to left.

137
00:06:02,333 --> 00:06:04,833
But anyway,
the fact of the matter is

138
00:06:04,833 --> 00:06:07,066
that this is
a complex document.

139
00:06:07,066 --> 00:06:08,933
And you need to search that.

140
00:06:08,933 --> 00:06:13,100
Now, fortunately,
we know how to search it.

141
00:06:13,100 --> 00:06:14,766
We know how to play with it.

142
00:06:14,766 --> 00:06:17,533
Like most of the software
development today,

143
00:06:17,533 --> 00:06:19,533

your target is to be able
to have software

144
00:06:19,533 --> 00:06:22,433
that basically you can utilize
other people's software

145
00:06:22,433 --> 00:06:24,533
or freely-available
commodity software,

146
00:06:24,533 --> 00:06:27,000
'cause if you actually tried
to redevelop your software,

147
00:06:27,000 --> 00:06:28,266
what you're gonna end up doing

148
00:06:28,266 --> 00:06:30,233
is having lots of errors,

149
00:06:30,233 --> 00:06:33,066
if you ever succeed at all
to have anything done.

150
00:06:33,066 --> 00:06:35,900
So what we do is, we capitalize
on other people's technology--

151
00:06:35,900 --> 00:06:39,300
freely available software,
freely available component,

152
00:06:39,300 --> 00:06:41,333
some more robust than others--

153
00:06:41,333 --> 00:06:43,133
and we try to get an answer.

154

00:06:43,133 --> 00:06:45,866
And we do know how to deal
with handwriting recognition,

155
00:06:45,866 --> 00:06:48,766
and we do know how to deal
with structured extraction,

156
00:06:48,766 --> 00:06:50,800
and we do know how
to deal with OCR,

157
00:06:50,800 --> 00:06:53,533
and we do know how to deal
with all of that.

158
00:06:53,533 --> 00:06:56,200
And the way we deal
with all that is simply,

159
00:06:56,200 --> 00:06:57,733
we take the document--

160
00:06:57,733 --> 00:07:00,500
we basically--

161
00:07:00,500 --> 00:07:02,366
so there's, like, a document.

162
00:07:02,366 --> 00:07:04,533
We then basically
enhance the document,

163
00:07:04,533 --> 00:07:07,033
because you can see
it's kind of bad.

164
00:07:07,033 --> 00:07:08,266
We layer it.

165
00:07:08,266 --> 00:07:10,700
We OCR the various components.

166
00:07:10,700 --> 00:07:12,966
We take all
the various components

167
00:07:12,966 --> 00:07:15,133
and we run it through all
the various software routines

168
00:07:15,133 --> 00:07:16,500
that we know.

169
00:07:16,500 --> 00:07:18,433
We put it together

170
00:07:18,433 --> 00:07:20,233
and then we try
to either search it

171
00:07:20,233 --> 00:07:21,266
or mine it.

172
00:07:21,266 --> 00:07:22,900
Now, obviously here,

173
00:07:22,900 --> 00:07:24,133
I can tell you

174
00:07:24,133 --> 00:07:26,433
that there's a need

175
00:07:26,433 --> 00:07:28,566
to "enhance" the document.

176
00:07:28,566 --> 00:07:29,800
I have to prove it to you.

177
00:07:29,800 --> 00:07:31,433
Obviously, taking the documents

178
00:07:31,433 --> 00:07:33,333
you'll give me is granted,

179
00:07:33,333 --> 00:07:36,066
and extracting stuff from it
you'll give me is granted.

180
00:07:36,066 --> 00:07:37,833
But you're--what I still owe you

181
00:07:37,833 --> 00:07:40,700
is an explanation of why
I have to enhance it

182
00:07:40,700 --> 00:07:43,866
and why does integration
make any difference at all?

183
00:07:43,866 --> 00:07:46,300
So I owe you that explanation.

184
00:07:46,300 --> 00:07:48,400
So let me start with this.

185
00:07:48,400 --> 00:07:50,633
You can guess
which side is enhanced

186
00:07:50,633 --> 00:07:52,433
and which side is not.

187
00:07:52,433 --> 00:07:53,900
Right?

188

00:07:53,900 --> 00:07:55,400
I'm looking at people
who are, like, wondering--

189
00:07:55,400 --> 00:07:56,400
okay, you do know.

190
00:07:56,400 --> 00:07:57,733
Very good.

191
00:07:57,733 --> 00:07:58,966
So you can tell
which side's enhanced

192
00:07:58,966 --> 00:08:00,466
and which side's not,
and one--

193
00:08:00,466 --> 00:08:02,200
and you can tell
it's a handwritten document.

194
00:08:02,200 --> 00:08:04,033
What this is
is documents taken

195
00:08:04,033 --> 00:08:07,233
from a diary--an old diary

196
00:08:07,233 --> 00:08:08,833
that's gone through,
let us say,

197
00:08:08,833 --> 00:08:10,100
adverse conditions

198
00:08:10,100 --> 00:08:11,700
during World War II.

199

00:08:11,700 --> 00:08:14,333
And basically,
it's stored in a museum.

200
00:08:14,333 --> 00:08:16,133
So what we have here
is basically

201
00:08:16,133 --> 00:08:18,966
a scanned image of it,

202
00:08:18,966 --> 00:08:20,966
and as you can tell,
you have to scratch it,

203
00:08:20,966 --> 00:08:23,000
and you--and you can't
really read it.

204
00:08:23,000 --> 00:08:25,866
And even if you did try
to do OCR, forget it.

205
00:08:25,866 --> 00:08:28,366
You can't get anywhere with it.

206
00:08:28,366 --> 00:08:29,733
But how about this one?

207
00:08:31,566 --> 00:08:34,566
Now you can get somewhere on it.

208
00:08:34,566 --> 00:08:36,100
This one's even a little harder,

209
00:08:36,100 --> 00:08:38,500
because it's got a composition
of a handwritten

210

00:08:38,500 --> 00:08:42,500
and typed.

211
00:08:42,500 --> 00:08:44,266
And this one shows you
that if you enhance it,

212
00:08:44,266 --> 00:08:45,466
you really can get anywhere.

213
00:08:45,466 --> 00:08:46,800
What you can look at is--

214
00:08:46,800 --> 00:08:51,200
if you look at the black
square--rectangle,

215
00:08:51,200 --> 00:08:53,466
you can't really read anything
in the unenhanced,

216
00:08:53,466 --> 00:08:56,033
and you can easily read it
in the other hand.

217
00:08:56,033 --> 00:08:58,166
Anybody want to try
to translate it?

218
00:09:01,900 --> 00:09:03,666
I could make it bigger.

219
00:09:03,666 --> 00:09:05,300
You speak fluent German?

220
00:09:05,300 --> 00:09:06,633
Well,
if you speak fluent German,

221

00:09:06,633 --> 00:09:08,900
then you can translate it.

222
00:09:08,900 --> 00:09:10,533
Ah, fair enough.

223
00:09:10,533 --> 00:09:12,566
I can translate it
for you if you'd like,

224
00:09:12,566 --> 00:09:14,633
but it's a little hard
even for me to read,

225
00:09:14,633 --> 00:09:16,233
and I'm a little closer.

226
00:09:16,233 --> 00:09:18,433
Anyway, so you'll buy the fact

227
00:09:18,433 --> 00:09:21,000
that you need to enhance,
right?

228
00:09:21,000 --> 00:09:24,266
You--you should enhance,
because if you don't enhance

229
00:09:24,266 --> 00:09:26,100
things like that,
you won't be able to process--

230
00:09:26,100 --> 00:09:28,566
certainly not with OCR.

231
00:09:28,566 --> 00:09:31,000
But I also owe you the proof
that you actually--

232

00:09:31,000 --> 00:09:33,400
that it--integration actually
makes a difference--

233
00:09:33,400 --> 00:09:36,266
that you really do want
to break the pieces into

234
00:09:36,266 --> 00:09:37,633
the whole and the sum
of the pieces

235
00:09:37,633 --> 00:09:38,933
and sum it together.

236
00:09:38,933 --> 00:09:41,466
And the best way
to do it is this.

237
00:09:41,466 --> 00:09:42,733
This is me.

238
00:09:42,733 --> 00:09:44,733
Kind of like a business card.

239
00:09:44,733 --> 00:09:46,433
I don't really use
business cards anymore,

240
00:09:46,433 --> 00:09:48,966
but it's an old
historical artifact.

241
00:09:48,966 --> 00:09:52,400
And I have to ask you
the question of,

242
00:09:52,400 --> 00:09:54,600
what positions do I hold?

243
00:09:54,600 --> 00:09:55,933
Now, if you actually
look at it,

244
00:09:55,933 --> 00:09:57,900
and you did not deal
with the text--

245
00:09:57,900 --> 00:09:59,466
if you basically only deal

246
00:09:59,466 --> 00:10:01,800
with the components
that are not text,

247
00:10:01,800 --> 00:10:03,466
you cannot answer that.

248
00:10:03,466 --> 00:10:04,933
Well, that's a given.

249
00:10:04,933 --> 00:10:06,533
But the real question is,

250
00:10:06,533 --> 00:10:09,133
at which institution am I at?

251
00:10:09,133 --> 00:10:10,466
And as you can see,

252
00:10:10,466 --> 00:10:12,333
without processing the logo,

253
00:10:12,333 --> 00:10:13,500
you couldn't answer that.

254
00:10:13,500 --> 00:10:15,000
So without integration,

255
00:10:15,000 --> 00:10:17,733
you couldn't answer either
of these two questions.

256
00:10:17,733 --> 00:10:20,766
So let me show you a little bit
of what we did.

257
00:10:20,766 --> 00:10:22,733
But before we do that,

258
00:10:22,733 --> 00:10:24,766
and I look at the age
of the audience,

259
00:10:24,766 --> 00:10:26,866
and this--it makes it
very clear for me--

260
00:10:26,866 --> 00:10:29,900
I have to address the notion
of technology.

261
00:10:29,900 --> 00:10:32,533
We built this prototype,
and by the time we were done,

262
00:10:32,533 --> 00:10:34,666
no one would use it,

263
00:10:34,666 --> 00:10:37,566
because technology moves
so quickly and so rapidly

264
00:10:37,566 --> 00:10:40,166
that it becomes out of date
fairly quickly.

265

00:10:40,166 --> 00:10:42,500
But unfortunately
or fortunately,

266
00:10:42,500 --> 00:10:44,900
what doesn't come out of date

267
00:10:44,900 --> 00:10:46,733
are benchmarks.

268
00:10:46,733 --> 00:10:48,566
Now I'll show you how sad--

269
00:10:48,566 --> 00:10:50,766
or inspiring,

270
00:10:50,766 --> 00:10:52,166
whichever way
you want to view it--

271
00:10:52,166 --> 00:10:54,366
benchmarks are.

272
00:10:54,366 --> 00:10:57,366
So, in early-on days,

273
00:10:57,366 --> 00:10:59,133
there's some competition
called TREC.

274
00:10:59,133 --> 00:11:00,966
Anybody ever heard of TREC?

275
00:11:00,966 --> 00:11:02,566
Nobody?
Ah.

276
00:11:02,566 --> 00:11:05,566
More than the number of people

that use Google Plus.

277
00:11:07,600 --> 00:11:11,500
TREC is an international
competition for text.

278
00:11:11,500 --> 00:11:13,900
It's basically not
a competition.

279
00:11:13,900 --> 00:11:15,166
It's sponsored by NIST.

280
00:11:15,166 --> 00:11:17,433
It's a bake-off where everybody
gets together.

281
00:11:17,433 --> 00:11:19,666
They submit--they give you
a set of queries,

282
00:11:19,666 --> 00:11:21,733
they give you a set of data,

283
00:11:21,733 --> 00:11:23,466
and they tell you,
"Go run your system on it.

284
00:11:23,466 --> 00:11:24,600
Submit the results."

285
00:11:24,600 --> 00:11:26,133
You send it to NIST,

286
00:11:26,133 --> 00:11:28,033
and NIST basically does
the evaluation for you.

287
00:11:28,033 --> 00:11:30,600

It's been running on
for a long, long time.

288
00:11:30,600 --> 00:11:32,366
In 1993--

289
00:11:32,366 --> 00:11:35,100
which based on some
of the audience age here

290
00:11:35,100 --> 00:11:37,800
was probably before
some of you were born--

291
00:11:37,800 --> 00:11:40,933
in 1993,

292
00:11:40,933 --> 00:11:45,033
they basically had a very,
very large collection--

293
00:11:45,033 --> 00:11:46,566
a collection that was so large,

294
00:11:46,566 --> 00:11:48,200
they basically felt
that the academics

295
00:11:48,200 --> 00:11:50,633
can't really do a lot with it,

296
00:11:50,633 --> 00:11:53,600
and what they did was that
they actually had a subset of it

297
00:11:53,600 --> 00:11:55,866
they also evaluated on.

298
00:11:55,866 --> 00:12:00,366

So anybody want to guess what is
a "large" collection in 1993--

299
00:12:00,366 --> 00:12:03,333
a very large collection that
academics couldn't handle?

300
00:12:04,500 --> 00:12:06,300
Order of magnitude.

301
00:12:06,300 --> 00:12:09,600
Was it--thinking terabytes?

302
00:12:09,600 --> 00:12:11,533
No, how about hundreds of tera--

303
00:12:11,533 --> 00:12:13,633
how about hundreds of gigabytes?

304
00:12:16,333 --> 00:12:19,833
It was actually two gigabytes.

305
00:12:19,833 --> 00:12:21,300
People couldn't store it.

306
00:12:21,300 --> 00:12:25,266
Forget--forget actually--
store it.

307
00:12:25,266 --> 00:12:27,800
And they had to rely on
a small subset collection

308
00:12:27,800 --> 00:12:30,000
which was 500 megabytes.

309
00:12:31,900 --> 00:12:34,500
If you take your iPhone
in your pocket

310
00:12:34,500 --> 00:12:36,233
or whatever device you have

311
00:12:36,233 --> 00:12:38,633
and think how much you've
got storage on it,

312
00:12:38,633 --> 00:12:41,500
and think of how many songs
you have on it.

313
00:12:41,500 --> 00:12:45,266
Well, anyway, that benchmark
that they still use--

314
00:12:45,266 --> 00:12:48,766
still is used
periodically today--

315
00:12:48,766 --> 00:12:51,066
long time ago,
still used today.

316
00:12:51,066 --> 00:12:53,966
So we felt that we--if we're
gonna have any benchmarks--

317
00:12:53,966 --> 00:12:55,833
that we were gonna
build any benchmarks--

318
00:12:55,833 --> 00:12:57,766
we actually needed
to build a benchmark

319
00:12:57,766 --> 00:12:59,733
that would stand
the test of time.

320
00:12:59,733 --> 00:13:02,233
But more the fact,

321
00:13:02,233 --> 00:13:03,633
we knew we had to build
a benchmark,

322
00:13:03,633 --> 00:13:06,100
because no such systems
as we described were available,

323
00:13:06,100 --> 00:13:09,066
and no--no way of
evaluating them was possible.

324
00:13:09,066 --> 00:13:10,233
So we had to spend
a lot of time

325
00:13:10,233 --> 00:13:11,600
creating our own benchmarks.

326
00:13:11,600 --> 00:13:13,100
And you know what
the cardinal rule of

327
00:13:13,100 --> 00:13:14,766
creating your own benchmark is?

328
00:13:16,266 --> 00:13:19,233
You're--you're guaranteed
to be the best.

329
00:13:20,366 --> 00:13:21,933
You're also guaranteed
to be the worst,

330
00:13:21,933 --> 00:13:23,066
but you never say that.

331
00:13:24,766 --> 00:13:26,066
So we built a benchmark,

332
00:13:26,066 --> 00:13:27,700
and what are the benchmarks?

333
00:13:27,700 --> 00:13:28,800
In order to have a benchmark,

334
00:13:28,800 --> 00:13:30,200
we built a set
of characteristics.

335
00:13:30,200 --> 00:13:32,466
What are the characteristics
that we felt important?

336
00:13:32,466 --> 00:13:34,200
It had to vary in inputs.

337
00:13:34,200 --> 00:13:35,833
It had to vary in fonts.

338
00:13:35,833 --> 00:13:37,300
It had to vary in
graphical elements.

339
00:13:37,300 --> 00:13:39,066
It had to vary in everything--

340
00:13:39,066 --> 00:13:40,833
had to vary in the whole things

341
00:13:40,833 --> 00:13:43,766
and key to it

342
00:13:43,766 --> 00:13:46,600

is it had to be...

343
00:13:46,600 --> 00:13:49,500
free.

344
00:13:49,500 --> 00:13:50,966
How many times have you bought?

345
00:13:50,966 --> 00:13:52,433
So when I was growing up

346
00:13:52,433 --> 00:13:55,666
and we wanted to buy music,

347
00:13:55,666 --> 00:13:57,266
we did something
that's very archaic,

348
00:13:57,266 --> 00:13:59,766
called "we bought records."

349
00:13:59,766 --> 00:14:01,233
You've seen records?

350
00:14:01,233 --> 00:14:03,200
Okay.

351
00:14:03,200 --> 00:14:05,833
So when we--
we would buy records.

352
00:14:05,833 --> 00:14:08,566
Later on, we would buy CDs.

353
00:14:08,566 --> 00:14:10,533
You've seen CDs?

354
00:14:10,533 --> 00:14:14,666

Now, let us say,
you "borrow" music

355
00:14:14,666 --> 00:14:18,033
from various different sources.

356
00:14:18,033 --> 00:14:20,566
You never pay for data.

357
00:14:20,566 --> 00:14:22,866
You don't pay to
read the newspaper.

358
00:14:22,866 --> 00:14:25,533
You don't pay to
search any media.

359
00:14:25,533 --> 00:14:27,400
It has to be free.

360
00:14:27,400 --> 00:14:29,200
So for anybody
to actually use it,

361
00:14:29,200 --> 00:14:32,400
you had to worry about
the copyright problems

362
00:14:32,400 --> 00:14:34,966
and you had to make sure that
if you solve the copyright,

363
00:14:34,966 --> 00:14:37,333
you solved it in such a way
that it was free.

364
00:14:37,333 --> 00:14:39,533
So we actually built a system--

365

00:14:39,533 --> 00:14:41,166
we built a benchmark
that's free.

366
00:14:41,166 --> 00:14:43,266
Make a long story short,

367
00:14:43,266 --> 00:14:45,233
we built this benchmark.

368
00:14:45,233 --> 00:14:46,233
It was used.

369
00:14:46,233 --> 00:14:47,500
It's about 7 million documents.

370
00:14:47,500 --> 00:14:49,600
42 million TIFF images.

371
00:14:49,600 --> 00:14:51,166
It's about 100--

372
00:14:51,166 --> 00:14:54,166
it's about 100 gigabytes of OCR

373
00:14:54,166 --> 00:14:55,533
after you OCR'd it.

374
00:14:55,533 --> 00:14:58,333
It's 1.5 terabytes in size.

375
00:14:58,333 --> 00:15:00,133
It's still credible, and it's--

376
00:15:00,133 --> 00:15:01,700
most importantly,
it's still used;

377

00:15:01,700 --> 00:15:02,933
it's still there.

378
00:15:02,933 --> 00:15:04,666
If you want it,
you can go to NIST,

379
00:15:04,666 --> 00:15:05,866
and you can get it.

380
00:15:08,000 --> 00:15:09,366
Basically, what we did was,

381
00:15:09,366 --> 00:15:10,766
we actually came up with a way

382
00:15:10,766 --> 00:15:13,266
of actually searching
this collection, and it--

383
00:15:13,266 --> 00:15:14,800
it's been used and liked,

384
00:15:14,800 --> 00:15:16,266
but suffice to say,

385
00:15:16,266 --> 00:15:18,600
is we built a simple
prototype system.

386
00:15:18,600 --> 00:15:20,766
It had very simple integration.

387
00:15:20,766 --> 00:15:22,166
These are all
the components to it.

388
00:15:22,166 --> 00:15:25,633
You just plug in

component together,

389
00:15:25,633 --> 00:15:27,233
and you basically ran it.

390
00:15:27,233 --> 00:15:31,133
Now,
without going into details,

391
00:15:31,133 --> 00:15:33,633
let me show you why
this system is--

392
00:15:33,633 --> 00:15:35,233
should be of your interest.

393
00:15:35,233 --> 00:15:37,233
So here's a query for you.

394
00:15:38,900 --> 00:15:41,133
The query says you want a logo

395
00:15:41,133 --> 00:15:43,600
of the American Tobacco
Company.

396
00:15:43,600 --> 00:15:45,566
It has to talk
about the documents--

397
00:15:45,566 --> 00:15:47,300
to talk about income forecast.

398
00:15:47,300 --> 00:15:48,433
And it has to be of--

399
00:15:48,433 --> 00:15:49,633
talking about income forecast

400
00:15:49,633 --> 00:15:52,666
of greater than 500,000.

401
00:15:52,666 --> 00:15:53,966
Now, if you look at it,

402
00:15:53,966 --> 00:15:56,733
the logo is pretty clear.

403
00:15:56,733 --> 00:15:58,433
And if you actually look at it,

404
00:15:58,433 --> 00:15:59,833
the word in the red box

405
00:15:59,833 --> 00:16:01,566
that you probably can't see

406
00:16:01,566 --> 00:16:04,433
says "income forecast."

407
00:16:04,433 --> 00:16:06,100
So far, pretty straightforward.

408
00:16:06,100 --> 00:16:08,533
But in the black box,

409
00:16:08,533 --> 00:16:10,766
it says "800,000."

410
00:16:12,533 --> 00:16:15,366
Now, the reason I use
that as an example is,

411
00:16:15,366 --> 00:16:18,100
if you actually
did a text search,

412
00:16:18,100 --> 00:16:20,166
you wouldn't find
this document,

413
00:16:20,166 --> 00:16:22,066
even if you found the logo,

414
00:16:22,066 --> 00:16:23,766
because a text document--

415
00:16:23,766 --> 00:16:26,300
when you basically
type in the number 500,000,

416
00:16:26,300 --> 00:16:28,600
what are you actually doing?

417
00:16:28,600 --> 00:16:31,433
You're looking for a string
that says "500,000."

418
00:16:31,433 --> 00:16:33,333
But in a database search,

419
00:16:33,333 --> 00:16:36,700
when you say "500,000"
or "greater than 500,000,"

420
00:16:36,700 --> 00:16:38,266
say an SQL,

421
00:16:38,266 --> 00:16:41,500
you actually get a result
that's greater than 500,000.

422
00:16:41,500 --> 00:16:43,333
So this document fits.

423

00:16:46,000 --> 00:16:48,233
Here's another example.

424
00:16:48,233 --> 00:16:52,666
RJR logo with the filtration
efficiency and a signature.

425
00:16:52,666 --> 00:16:55,900
Why do you care
about signatures?

426
00:16:55,900 --> 00:16:58,900
When do you sign?

427
00:16:58,900 --> 00:17:02,566
Typically, if you are
in management particularly,

428
00:17:02,566 --> 00:17:06,133
all you do all day long
is like this,

429
00:17:06,133 --> 00:17:08,266
because that's
an authorization.

430
00:17:08,266 --> 00:17:11,233
So this is a way
of identifying

431
00:17:11,233 --> 00:17:13,066
something that has a logo.

432
00:17:13,066 --> 00:17:14,466
It talks about filtration,

433
00:17:14,466 --> 00:17:16,033
and somebody has
an authorization

434
00:17:16,033 --> 00:17:17,800
that did something about it,

435
00:17:17,800 --> 00:17:20,766
hence this document.

436
00:17:20,766 --> 00:17:23,766
Another example is five--

437
00:17:23,766 --> 00:17:26,666
find the five
highest signatures

438
00:17:26,666 --> 00:17:30,033
of dollar allocations
by people.

439
00:17:30,033 --> 00:17:32,133
Now, you can imagine
there are some organizations

440
00:17:32,133 --> 00:17:33,933
that care about payments--

441
00:17:33,933 --> 00:17:37,300
that pay two people by certain
authorizing people...

442
00:17:37,300 --> 00:17:38,633
I guess.

443
00:17:38,633 --> 00:17:41,800
Too many "peoples"
in that commentary.

444
00:17:41,800 --> 00:17:45,966
But to--the goal--
in order to solve that,

445
00:17:45,966 --> 00:17:47,500
you have to identify
all the documents

446
00:17:47,500 --> 00:17:49,066
signed by the people.

447
00:17:49,066 --> 00:17:51,800
You have to "sum total" any
money that they talk about,

448
00:17:51,800 --> 00:17:53,833
and then you have
to sort order the ones

449
00:17:53,833 --> 00:17:56,933
to be most highest pay.

450
00:17:56,933 --> 00:17:59,933
But that's only relatively
easy part of the challenge.

451
00:18:01,866 --> 00:18:06,400
How many have--how many people
have signed ten things in a row?

452
00:18:08,233 --> 00:18:11,333
More of you than that, I'm sure.

453
00:18:11,333 --> 00:18:14,700
The problem is, when you sign
ten times in a row,

454
00:18:14,700 --> 00:18:17,900
how similar is your signature?

455
00:18:17,900 --> 00:18:19,966
Hundred times in a row?

456
00:18:19,966 --> 00:18:21,933
For that matter,
every time I sign something,

457
00:18:21,933 --> 00:18:23,433
it looks different.

458
00:18:25,033 --> 00:18:26,333
So the fact of the matter is,

459
00:18:26,333 --> 00:18:28,633
you have to do
signature matching.

460
00:18:28,633 --> 00:18:31,033
And unlike simple,
straightforward things,

461
00:18:31,033 --> 00:18:33,900
signatures are
very hard to match.

462
00:18:33,900 --> 00:18:36,266
But we cheat.

463
00:18:36,266 --> 00:18:38,966
How do you think we cheat?

464
00:18:38,966 --> 00:18:41,133
Well, we take different
extraction procedures,

465
00:18:41,133 --> 00:18:42,933
and we're integrating
everything together,

466
00:18:42,933 --> 00:18:45,100
and some byline is
gonna have a name,

467
00:18:45,100 --> 00:18:47,366
or some header's
gonna have a name.

468
00:18:47,366 --> 00:18:49,766
And that's the way
we match it.

469
00:18:49,766 --> 00:18:51,666
So that--so these
are the five people

470
00:18:51,666 --> 00:18:53,100
who paid the most
amount of money

471
00:18:53,100 --> 00:18:56,500
in some payments on
tobacco litigation.

472
00:18:56,500 --> 00:18:58,933
Here's another example.

473
00:18:58,933 --> 00:19:03,000
This is a query about
the associations of people:

474
00:19:03,000 --> 00:19:05,700
where they've paid,
who they've paid,

475
00:19:05,700 --> 00:19:08,666
who Dr. Stone has paid.

476
00:19:08,666 --> 00:19:10,866
Again, not your typical search

477
00:19:10,866 --> 00:19:12,866

that is supported.

478
00:19:12,866 --> 00:19:15,866
And, of course, when I
actually say that I built it,

479
00:19:15,866 --> 00:19:17,233
actually I didn't do it very--

480
00:19:17,233 --> 00:19:18,500
I was just running the effort.

481
00:19:18,500 --> 00:19:20,200
These are the people
who actually did it.

482
00:19:20,200 --> 00:19:21,666
And they did it over time,

483
00:19:21,666 --> 00:19:23,766
and over time, they've changed
their affiliations.

484
00:19:23,766 --> 00:19:25,766
Some are students,
some are collaborators,

485
00:19:25,766 --> 00:19:29,633
some are colleagues,
and so on, and so forth.

486
00:19:29,633 --> 00:19:32,566
And since I am an academic,

487
00:19:32,566 --> 00:19:36,900
we have to worry
about publications.

488
00:19:36,900 --> 00:19:39,300

In the new age
of universities today,

489
00:19:39,300 --> 00:19:40,900
a lot of people are
looking at patents,

490
00:19:40,900 --> 00:19:44,566
so there's some patents
about it later on.

491
00:19:44,566 --> 00:19:47,166
So that was kind of basically
in the middle ground.

492
00:19:47,166 --> 00:19:50,933
It was somewhere between
research and practice.

493
00:19:50,933 --> 00:19:52,600
We're trying to solve
a real problem.

494
00:19:52,600 --> 00:19:55,766
We were trying to get something
in the prototype stage.

495
00:19:55,766 --> 00:19:58,833
But it hasn't really
seen major deployment

496
00:19:58,833 --> 00:20:01,300
or any specific deployment.

497
00:20:01,300 --> 00:20:03,100
The next effort
we're talking about

498
00:20:03,100 --> 00:20:05,233
is a little bit more...

499
00:20:05,233 --> 00:20:07,000
to the point.

500
00:20:07,000 --> 00:20:09,433
And it's actually very relevant

501
00:20:09,433 --> 00:20:12,833
to systems that basically
have user reportings.

502
00:20:12,833 --> 00:20:16,600
And user reporting systems have
wildly interesting spellings

503
00:20:16,600 --> 00:20:18,400
and have wildly
interesting phrasing

504
00:20:18,400 --> 00:20:20,866
and have wildly
interesting grammars.

505
00:20:20,866 --> 00:20:24,466
And when you have a real
problem searching those,

506
00:20:24,466 --> 00:20:28,833
like you have in the various
reporting systems like NASA has,

507
00:20:28,833 --> 00:20:31,133
you need to come up with
such a solution.

508
00:20:31,133 --> 00:20:34,733
So, searching in
adverse conditions--

509

00:20:34,733 --> 00:20:38,366
what does it mean?

510
00:20:38,366 --> 00:20:41,366
Spelling is difficult.

511
00:20:41,366 --> 00:20:43,200
It's getting worse
of a problem,

512
00:20:43,200 --> 00:20:44,966
because we've got autocorrect
for everything,

513
00:20:44,966 --> 00:20:47,533
so basically people learn how
to spell less and less well,

514
00:20:47,533 --> 00:20:49,400
and then they basically
try to spell

515
00:20:49,400 --> 00:20:51,566
using the "Twitterese"

516
00:20:51,566 --> 00:20:54,866
or any other
encryption mechanism

517
00:20:54,866 --> 00:20:57,833
that they call spelling.

518
00:20:57,833 --> 00:21:00,533
And it gets very difficult
in the spelling.

519
00:21:00,533 --> 00:21:02,066
As a faculty,
I look at spelling

520
00:21:02,066 --> 00:21:03,333
when people write on exams,

521
00:21:03,333 --> 00:21:04,700
and it's getting
worse and worse.

522
00:21:06,333 --> 00:21:08,866
But it's even worse when
you're starting to spell

523
00:21:08,866 --> 00:21:11,000
in foreign languages,

524
00:21:11,000 --> 00:21:13,433
particularly if you don't know
the foreign language,

525
00:21:13,433 --> 00:21:14,500
but you're gonna do a search

526
00:21:14,500 --> 00:21:17,300
that is related
to the foreign language.

527
00:21:17,300 --> 00:21:19,866
So one such situation exists

528
00:21:19,866 --> 00:21:22,533
in a collection called
Yizkor Books.

529
00:21:23,866 --> 00:21:26,466
And the other issue is
when you start dealing

530
00:21:26,466 --> 00:21:28,466
with medical texts,

531
00:21:28,466 --> 00:21:30,900
and I'll tell you about
both of them independently.

532
00:21:30,900 --> 00:21:32,500
So Yizkor Books--

533
00:21:32,500 --> 00:21:34,933
"yizkor," the word in Hebrew,
means "remember."

534
00:21:34,933 --> 00:21:37,933
It's a collection of books that
are scattered around the world,

535
00:21:37,933 --> 00:21:39,133
including in the Holocaust--

536
00:21:39,133 --> 00:21:41,600
United States Holocaust
Memorial Museum.

537
00:21:41,600 --> 00:21:44,566
And in this collection,
the restriction--

538
00:21:44,566 --> 00:21:46,000
there's restricted access,

539
00:21:46,000 --> 00:21:48,000
and there's also a section

540
00:21:48,000 --> 00:21:50,833
that actually has an archive.

541
00:21:50,833 --> 00:21:53,100
And the archive--
people come in there and say,

542
00:21:53,100 --> 00:21:55,566
"I'd like to find information

543
00:21:55,566 --> 00:21:58,400
about someone."

544
00:21:58,400 --> 00:22:00,633
And so, "Okay, fine, who
are you looking for?"

545
00:22:00,633 --> 00:22:03,400
"Oh, they lived in some city."

546
00:22:03,400 --> 00:22:05,533
"Okay, fine, when is this from?"

547
00:22:05,533 --> 00:22:07,333
"Oh, this is from
World War II era."

548
00:22:07,333 --> 00:22:09,200
"Okay, fine.

549
00:22:09,200 --> 00:22:10,766
Could you be a little bit
more specific?

550
00:22:10,766 --> 00:22:12,233
Where did they live?"

551
00:22:12,233 --> 00:22:14,933
"In Europe."

552
00:22:14,933 --> 00:22:18,433
Not really helpful,
but it's a start.

553
00:22:18,433 --> 00:22:20,266

"Okay, what about--

554
00:22:20,266 --> 00:22:22,466
could you tell me
what language they spoke?"

555
00:22:22,466 --> 00:22:24,633
"Oh, yeah, yeah, yeah.
They spoke German."

556
00:22:24,633 --> 00:22:27,166
Getting a little better.

557
00:22:27,166 --> 00:22:29,333
"What was the name of the city
they lived in?"

558
00:22:30,900 --> 00:22:34,133
"I don't know, but it had a...

559
00:22:34,133 --> 00:22:35,533
'burg' in it."

560
00:22:37,900 --> 00:22:39,500
"Anything else?"

561
00:22:39,500 --> 00:22:40,700
"Um, not quite.

562
00:22:40,700 --> 00:22:41,866
I don't know.

563
00:22:41,866 --> 00:22:44,400
It kind of sounds like..."

564
00:22:44,400 --> 00:22:49,966
So, I said, "Okay, so where
would be an example of this?"

565
00:22:49,966 --> 00:22:51,866
So one example of such city

566
00:22:51,866 --> 00:22:55,733
is called Bratislava.

567
00:22:55,733 --> 00:22:58,500
Anybody heard of Bratislava?

568
00:22:58,500 --> 00:23:01,000
It's in Slovakia.

569
00:23:02,666 --> 00:23:06,766
Does Bratislava have
the "burg" in it?

570
00:23:06,766 --> 00:23:08,566
I don't know how to spell
Bratislava too well,

571
00:23:08,566 --> 00:23:10,333
but it doesn't sound like
it has "burg" in it.

572
00:23:10,333 --> 00:23:12,866
But it does,
because the word "Pressburg"

573
00:23:12,866 --> 00:23:15,833
was another name for Bratislava.

574
00:23:15,833 --> 00:23:19,033
And that one does have it.

575
00:23:19,033 --> 00:23:20,466
So it gets a little difficult--

576
00:23:20,466 --> 00:23:21,933

and they did speak German,

577
00:23:21,933 --> 00:23:24,066
because the Jews of that area
called it Pressburg,

578
00:23:24,066 --> 00:23:26,633
and they actually were
speaking German.

579
00:23:28,200 --> 00:23:31,033
So we wanted to build
a search system

580
00:23:31,033 --> 00:23:33,333
where people didn't quite know
how to pronounce things--

581
00:23:33,333 --> 00:23:38,166
people didn't quite know
how to, location-wise, find it.

582
00:23:38,166 --> 00:23:41,566
People were--but yet
they wanted information.

583
00:23:41,566 --> 00:23:43,933
And we built a system

584
00:23:43,933 --> 00:23:48,133
that basically has
such an interface.

585
00:23:48,133 --> 00:23:49,933
But most importantly,
they give you--

586
00:23:49,933 --> 00:23:52,500
when they give--
they do a search,

587
00:23:52,500 --> 00:23:55,966
they actually can find
a collection from the documents

588
00:23:55,966 --> 00:23:57,800
that actually have
multiple languages

589
00:23:57,800 --> 00:23:59,533
in the same documents,

590
00:23:59,533 --> 00:24:01,266
let alone different documents

591
00:24:01,266 --> 00:24:02,766
of different languages.

592
00:24:02,766 --> 00:24:05,233
And it produces
a simple ranking system.

593
00:24:05,233 --> 00:24:07,133
Now, as you can see,

594
00:24:07,133 --> 00:24:09,300
it's not doing
particularly great.

595
00:24:09,300 --> 00:24:10,733
But it is finding it.

596
00:24:10,733 --> 00:24:11,933
And the way it finds it

597
00:24:11,933 --> 00:24:14,500
is based on
this simple algorithm.

598
00:24:14,500 --> 00:24:17,633
If you look at
the top green box--

599
00:24:17,633 --> 00:24:19,266
segment section--

600
00:24:19,266 --> 00:24:21,433
that is basically
breaking things up

601
00:24:21,433 --> 00:24:23,600
in random pieces.

602
00:24:23,600 --> 00:24:25,133
And if you look at
the bottom green box,

603
00:24:25,133 --> 00:24:27,233
that is a traditional approach
of what people do.

604
00:24:27,233 --> 00:24:29,200
It's called n-grams.

605
00:24:29,200 --> 00:24:31,766
N-grams is sliding
character windows

606
00:24:31,766 --> 00:24:33,366
that go overlap each other

607
00:24:33,366 --> 00:24:35,366
in order to find
what you're looking for.

608
00:24:35,366 --> 00:24:39,800
The problem is--
is it's sliding windows.

609
00:24:39,800 --> 00:24:42,100
And sliding means contiguous.

610
00:24:42,100 --> 00:24:44,800
And if it's contiguous,
you cannot find some things

611
00:24:44,800 --> 00:24:47,300
that are basically chopped
in the middle.

612
00:24:47,300 --> 00:24:49,566
So the top one takes care of it.

613
00:24:49,566 --> 00:24:51,900
And we built this system,

614
00:24:51,900 --> 00:24:56,100
in use in the archive section
of the Holocaust Museum today.

615
00:24:56,100 --> 00:24:58,600
And it uses simple,
simplistic rules

616
00:24:58,600 --> 00:24:59,833
to break things up,

617
00:24:59,833 --> 00:25:01,400
which are kind of
crazy-looking,

618
00:25:01,400 --> 00:25:04,000
but they are to try variety

619
00:25:04,000 --> 00:25:05,733
and add mutation into things.

620
00:25:07,233 --> 00:25:09,566
And here's a standard way
of evaluating it.

621
00:25:09,566 --> 00:25:12,200
So the way you evaluate
these systems

622
00:25:12,200 --> 00:25:13,900
is by taking--

623
00:25:13,900 --> 00:25:16,033
the standard approach is
basically language bases

624
00:25:16,033 --> 00:25:17,500
called D-M Soundex--

625
00:25:17,500 --> 00:25:18,733
it's a Soundex approach;

626
00:25:18,733 --> 00:25:20,233
it's by sound--

627
00:25:20,233 --> 00:25:23,466
which doesn't help if
you cannot pronounce it.

628
00:25:23,466 --> 00:25:26,266
Anybody speak Czech here?

629
00:25:26,266 --> 00:25:29,666
Ah, well then, you can
invalidate my statement.

630
00:25:29,666 --> 00:25:31,433
I don't speak Czech.

631

00:25:31,433 --> 00:25:33,933
But there are words
that are yea long

632
00:25:33,933 --> 00:25:36,666
that just forget about vowels.

633
00:25:36,666 --> 00:25:40,100
They just don't think
it's necessary.

634
00:25:40,100 --> 00:25:42,433
And for somebody like me
to try to pronounce it

635
00:25:42,433 --> 00:25:44,166
is a new experience.

636
00:25:46,600 --> 00:25:48,366
Make a long story short,

637
00:25:48,366 --> 00:25:50,900
doesn't help
the phonetics of it.

638
00:25:50,900 --> 00:25:52,866
So the way that, generally--

639
00:25:52,866 --> 00:25:54,800
other approaches
is called n-grams,

640
00:25:54,800 --> 00:25:57,033
and ours is the one
at the bottom.

641
00:25:57,033 --> 00:25:58,566
And what you're looking at

642

00:25:58,566 --> 00:26:00,600
is basically a standard way
of evaluating

643
00:26:00,600 --> 00:26:03,200
search techniques for letters.

644
00:26:03,200 --> 00:26:05,833
It's basically--
you add a letter,

645
00:26:05,833 --> 00:26:07,733
you drop a letter,
you replace a letter,

646
00:26:07,733 --> 00:26:09,733
or you swap a random pair.

647
00:26:09,733 --> 00:26:11,100
And what you're looking at--

648
00:26:11,100 --> 00:26:13,133
the ones in red
are the ones that--

649
00:26:13,133 --> 00:26:14,333
the best score.

650
00:26:14,333 --> 00:26:15,666
And what you're looking at,

651
00:26:15,666 --> 00:26:16,966
and when you talk
about the rank,

652
00:26:16,966 --> 00:26:19,066
is where in the list
did you find that--

653

00:26:19,066 --> 00:26:20,366
what was the average rank

654
00:26:20,366 --> 00:26:23,066
for all the collection
that you tried?

655
00:26:23,066 --> 00:26:25,100
And the reason
there's two numbers

656
00:26:25,100 --> 00:26:28,500
is for the bottom number--

657
00:26:28,500 --> 00:26:30,666
sorry, the top--
the bottom number

658
00:26:30,666 --> 00:26:33,800
is what did you find that
n-grams also found?

659
00:26:33,800 --> 00:26:38,333
So the USHMM search
finds everything

660
00:26:38,333 --> 00:26:41,466
the n-grams search does,
plus some other ones.

661
00:26:41,466 --> 00:26:44,266
So if you only look
at what the USHMM did,

662
00:26:44,266 --> 00:26:46,633
and what the n-gram--
the n-gram also found,

663
00:26:46,633 --> 00:26:47,766
it's the bottom number.

664
00:26:47,766 --> 00:26:48,900
And as you can tell,

665
00:26:48,900 --> 00:26:50,200
these are ugly numbers,

666
00:26:50,200 --> 00:26:52,166
but what you want to see
is where the trend is.

667
00:26:53,933 --> 00:26:57,600
And this is if you wanted
to get really scary.

668
00:26:57,600 --> 00:27:01,100
Add two characters,
add three characters,

669
00:27:01,100 --> 00:27:02,633
add four random characters--

670
00:27:02,633 --> 00:27:05,600
remove two, three, and four--

671
00:27:05,600 --> 00:27:07,800
replace two, three, and four--

672
00:27:07,800 --> 00:27:09,666
and swap two, three, and four.

673
00:27:09,666 --> 00:27:11,066
And whenever you--
what you see is,

674
00:27:11,066 --> 00:27:12,833
the trends are the same.

675

00:27:12,833 --> 00:27:14,633
So these are the algorithms
that's used today

676
00:27:14,633 --> 00:27:15,900
in the archive section

677
00:27:15,900 --> 00:27:17,533
to find names.

678
00:27:19,800 --> 00:27:21,200
But I'm in a medical center--

679
00:27:21,200 --> 00:27:23,500
one of my appointments
is the medical center.

680
00:27:23,500 --> 00:27:25,500
And if you're in
the medical center,

681
00:27:25,500 --> 00:27:27,666
you basically want to show
that you're doing something

682
00:27:27,666 --> 00:27:29,533
that's useful for medicine.

683
00:27:29,533 --> 00:27:32,766
So we actually tried
to deal with transcriptions

684
00:27:32,766 --> 00:27:35,333
and how much errors occur.

685
00:27:35,333 --> 00:27:37,833
Now, by transcription error--

686
00:27:37,833 --> 00:27:41,800

how many of you have
been to a hospital?

687
00:27:41,800 --> 00:27:44,266
How many of you have been
treated in a hospital?

688
00:27:44,266 --> 00:27:45,633
Quite a few of you.

689
00:27:45,633 --> 00:27:47,733
How many of you would wish
they would not be there

690
00:27:47,733 --> 00:27:49,700
during the time shifts
between--

691
00:27:49,700 --> 00:27:53,400
when nurses go from 3:00 to 3:30

692
00:27:53,400 --> 00:27:55,633
or from 7:00 to 7:30?

693
00:27:55,633 --> 00:27:58,400
The nurses overlap.

694
00:27:58,400 --> 00:28:00,866
And the reason they overlap is,
they hand off the information

695
00:28:00,866 --> 00:28:02,466
from one shift to the next.

696
00:28:02,466 --> 00:28:05,600
But a lot of times,
they kind of write things down.

697
00:28:05,600 --> 00:28:10,000

And so--and how many of you have
seen doctors' handwritings?

698
00:28:10,000 --> 00:28:13,500
They don't write things down;
they kind of scribble.

699
00:28:13,500 --> 00:28:16,400
The bottom line is that
transcription errors

700
00:28:16,400 --> 00:28:18,333
and the names of medications

701
00:28:18,333 --> 00:28:19,500
and so and so--

702
00:28:19,500 --> 00:28:21,333
forget about it.

703
00:28:21,333 --> 00:28:24,200
So a lot of these errors occur,

704
00:28:24,200 --> 00:28:26,633
and they account for quite a bit
of the possibilities.

705
00:28:26,633 --> 00:28:28,300
Now, not all--
not all, necessarily,

706
00:28:28,300 --> 00:28:29,666
are transcription errors.

707
00:28:29,666 --> 00:28:31,200
But some of them are.

708
00:28:33,766 --> 00:28:37,100
So we ran it

on a medical dictionary.

709
00:28:37,100 --> 00:28:39,300
And this is a standard
characteristics,

710
00:28:39,300 --> 00:28:41,600
and we were actually
quite happy.

711
00:28:41,600 --> 00:28:44,000
We did okay.

712
00:28:44,000 --> 00:28:45,966
And again, we did fairly well.

713
00:28:45,966 --> 00:28:47,700
In fact, we found
almost everything,

714
00:28:47,700 --> 00:28:49,600
even when you basically tried

715
00:28:49,600 --> 00:28:52,700
a complete mess of four swaps,

716
00:28:52,700 --> 00:28:54,333
and so on, and so forth.

717
00:28:56,433 --> 00:28:58,133
So that just shows us
a little bit more.

718
00:28:58,133 --> 00:28:59,833
And this is, again--
this is basically

719
00:28:59,833 --> 00:29:01,733
a straightforward combination

720
00:29:01,733 --> 00:29:04,300
of modifying n-gram solutions

721
00:29:04,300 --> 00:29:05,466
that are well-known,

722
00:29:05,466 --> 00:29:07,633
adding a little noise into it,

723
00:29:07,633 --> 00:29:08,933
because noise helps--

724
00:29:08,933 --> 00:29:09,966
by the way, how many of you

725
00:29:09,966 --> 00:29:13,733
deal with optimization
algorithms here?

726
00:29:13,733 --> 00:29:15,066
Some of you.

727
00:29:15,066 --> 00:29:17,233
There's famous
optimization algorithms.

728
00:29:17,233 --> 00:29:20,633
There's things like
simulated annealing

729
00:29:20,633 --> 00:29:22,033
in genetic algorithms.

730
00:29:22,033 --> 00:29:24,833
And you know what
a key of their success is?

731

00:29:24,833 --> 00:29:27,400
Add some noise:

732
00:29:27,400 --> 00:29:29,166
mutations,

733
00:29:29,166 --> 00:29:31,766
random heatings.

734
00:29:31,766 --> 00:29:33,666
Right?
So we did the same thing here.

735
00:29:33,666 --> 00:29:36,566
And it worked here too.

736
00:29:36,566 --> 00:29:38,800
Again,
these are the people involved.

737
00:29:38,800 --> 00:29:41,100
And by--when I say "I,"

738
00:29:41,100 --> 00:29:42,333
I didn't do much.

739
00:29:42,333 --> 00:29:44,900
It's obviously
my collaborators.

740
00:29:44,900 --> 00:29:46,700
And we thank them for--

741
00:29:46,700 --> 00:29:50,133
massive users of the system,
for their comments.

742
00:29:50,133 --> 00:29:52,500
And again, being an academic,

743
00:29:52,500 --> 00:29:54,633
and, of course, my students
very much appreciated

744
00:29:54,633 --> 00:29:57,800
the last academic stint

745
00:29:57,800 --> 00:30:00,066
to Santorini in April.

746
00:30:00,066 --> 00:30:01,433
Very nice place.

747
00:30:03,333 --> 00:30:07,166
And now I kind of
want to leave you with it,

748
00:30:07,166 --> 00:30:08,933
to show you that really,

749
00:30:08,933 --> 00:30:12,166
we do some things that are,
hopefully, for the future.

750
00:30:12,166 --> 00:30:14,200
And I'm going to talk about
searching social media,

751
00:30:14,200 --> 00:30:15,666
and I don't need
to motivate that.

752
00:30:15,666 --> 00:30:18,266
I don't need to give you
a NASA program that does that.

753
00:30:18,266 --> 00:30:19,566
It's everywhere.

754
00:30:19,566 --> 00:30:21,566
Everything they do
is social media.

755
00:30:23,233 --> 00:30:25,233
So I'm gonna talk about

756
00:30:25,233 --> 00:30:27,366
public health surveillance.

757
00:30:27,366 --> 00:30:30,633
And the way that
it's usually done is,

758
00:30:30,633 --> 00:30:32,433
it takes a long time,

759
00:30:32,433 --> 00:30:36,800
because it takes a huge amount
of human effort.

760
00:30:36,800 --> 00:30:39,533
It basically involves when
somebody comes around

761
00:30:39,533 --> 00:30:40,900
and looks for you

762
00:30:40,900 --> 00:30:42,800
and goes to the doctor
feeling sick.

763
00:30:42,800 --> 00:30:45,366
And enough times different
people go to the doctor,

764
00:30:45,366 --> 00:30:46,933
eventually the doctor

reports it.

765
00:30:46,933 --> 00:30:48,266
If enough doctors report it,

766
00:30:48,266 --> 00:30:50,133
eventually it's caught.

767
00:30:50,133 --> 00:30:51,900
But it takes a long time,

768
00:30:51,900 --> 00:30:54,633
and therefore,

769
00:30:54,633 --> 00:30:56,066
you need to expedite it.

770
00:30:56,066 --> 00:30:58,800
And the way you expedite
almost everything nowadays

771
00:30:58,800 --> 00:31:01,200
is social media.

772
00:31:02,966 --> 00:31:05,733
So, how do you deal
with social media?

773
00:31:05,733 --> 00:31:06,833
Well, we're not the first.

774
00:31:06,833 --> 00:31:08,233
In fact, many people
have done it.

775
00:31:08,233 --> 00:31:11,666
Typically, they talk about
a known basic problem.

776
00:31:11,666 --> 00:31:13,200
So they're gonna say,

777
00:31:13,200 --> 00:31:15,133
"I'm gonna look for influenza."

778
00:31:15,133 --> 00:31:17,833
And they're gonna
find influenza.

779
00:31:17,833 --> 00:31:18,966
And they're gonna look
through it,

780
00:31:18,966 --> 00:31:20,066
because they know what they're--

781
00:31:20,066 --> 00:31:21,833
exactly what
they're looking for.

782
00:31:21,833 --> 00:31:23,300
This is a problem

783
00:31:23,300 --> 00:31:25,500
if you don't know what
you're looking for.

784
00:31:25,500 --> 00:31:28,466
Often,
they use complex solutions.

785
00:31:28,466 --> 00:31:32,166
I firmly believe
in the KISS principle.

786
00:31:32,166 --> 00:31:33,900
Anybody know what
the KISS principle is?

787
00:31:33,900 --> 00:31:35,200
I'm sure you do.

788
00:31:35,200 --> 00:31:36,733
Keep it...

789
00:31:36,733 --> 00:31:38,433
[unintelligible chatter]

790
00:31:38,433 --> 00:31:41,433
I'll leave to you
to complete the rest of it.

791
00:31:42,866 --> 00:31:44,900
So the--and those don't
usually work,

792
00:31:44,900 --> 00:31:46,566
because complex solutions
take forever,

793
00:31:46,566 --> 00:31:49,966
and they aren't really
heavily adopted.

794
00:31:49,966 --> 00:31:54,466
Or they use their own resources
that you can't get access to.

795
00:31:54,466 --> 00:31:56,466
For example, query logs.

796
00:31:56,466 --> 00:31:59,033
Query logs are heavily used.

797
00:31:59,033 --> 00:32:01,966
Do you have an access
to the query log?

798
00:32:01,966 --> 00:32:04,266
Not really, so you can't do it.

799
00:32:05,966 --> 00:32:07,766
So what we wanted to do was,

800
00:32:07,766 --> 00:32:10,233
we wanted to change
the old way,

801
00:32:10,233 --> 00:32:12,400
which was saying,
"Is there a flu?

802
00:32:12,400 --> 00:32:14,366
Are you looking
for something specific?"

803
00:32:14,366 --> 00:32:16,366
To general things like,

804
00:32:16,366 --> 00:32:19,233
is there something occurring?

805
00:32:19,233 --> 00:32:21,033
Is there something new?

806
00:32:21,033 --> 00:32:23,166
If so, yes,
there is something occurring.

807
00:32:23,166 --> 00:32:25,633
It's the flu.

808
00:32:25,633 --> 00:32:28,633
So what we have is, we have
a collection of tweets.

809
00:32:28,633 --> 00:32:30,133
It's 2 billion tweets,

810
00:32:30,133 --> 00:32:33,600
which were given to us
by Johns Hopkins.

811
00:32:33,600 --> 00:32:36,200
We basically took those tweets--

812
00:32:36,200 --> 00:32:38,733
we partitioned them by time.

813
00:32:38,733 --> 00:32:42,733
We then checked for them
being trending.

814
00:32:42,733 --> 00:32:44,966
And once we identified
what is trending,

815
00:32:44,966 --> 00:32:47,500
we saw if it's something
that should alert us.

816
00:32:47,500 --> 00:32:51,033
That's the nutshell,
but I'll go in more specifics.

817
00:32:51,033 --> 00:32:52,533
So here's a tweet collection--

818
00:32:52,533 --> 00:32:53,966
2 billion.

819
00:32:53,966 --> 00:32:56,066
We cleaned out the ones
that are not health-related.

820
00:32:56,066 --> 00:32:58,466
We left 1.6 million.

821
00:32:58,466 --> 00:33:01,733
That, we--
it's partitioned over time.

822
00:33:01,733 --> 00:33:03,800
So then we partitioned it
over the time.

823
00:33:05,733 --> 00:33:07,500
We cleaned it up some more--

824
00:33:07,500 --> 00:33:09,133
got rid of some punctuations,

825
00:33:09,133 --> 00:33:11,633
stop-words, and so on,
and so forth.

826
00:33:11,633 --> 00:33:14,333
Then we found out what
is actually associating

827
00:33:14,333 --> 00:33:16,900
with one another.

828
00:33:16,900 --> 00:33:18,700
And we saw that if
you took the tweets

829
00:33:18,700 --> 00:33:21,300
"pounding head"--
"pounding headache,"

830
00:33:21,300 --> 00:33:24,300
"sore throat,"
and "low-grade flu and fever"

831
00:33:24,300 --> 00:33:25,966
versus the other ones,

832
00:33:25,966 --> 00:33:27,866
you saw that certain things
had a certain

833
00:33:27,866 --> 00:33:29,266
sufficient amount of support

834
00:33:29,266 --> 00:33:30,366
of what's going on.

835
00:33:30,366 --> 00:33:32,133
So "flu, sore throat"

836
00:33:35,533 --> 00:33:34,033
had support of 3,

837
00:33:35,533 --> 00:33:37,000
"cough" had support of 2.

838
00:33:37,000 --> 00:33:39,133
And basically,
we decided to clean out

839
00:33:39,133 --> 00:33:41,566
and keep only things
above a certain threshold.

840
00:33:41,566 --> 00:33:43,066
So that's an example.

841
00:33:44,766 --> 00:33:49,500
Then we decided to see if
that information is trending.

842

00:33:49,500 --> 00:33:51,300
And what we did was,
we looked for--

843
00:33:51,300 --> 00:33:55,300
oh, we looked for slope

844
00:33:55,300 --> 00:33:56,833
and a side.

845
00:33:58,733 --> 00:34:00,533
How many of you are
still think--

846
00:34:00,533 --> 00:34:02,933
writing your dissertations?

847
00:34:02,933 --> 00:34:05,133
Or thinking about
writing dissertations?

848
00:34:05,133 --> 00:34:07,866
Thinking about writing
research papers?

849
00:34:07,866 --> 00:34:09,533
Okay.

850
00:34:09,533 --> 00:34:13,166
I said we looked
for the change in slope.

851
00:34:13,166 --> 00:34:15,066
Bad word.

852
00:34:15,066 --> 00:34:16,433
I'll tell you a better word.

853
00:34:16,433 --> 00:34:18,466

We changed--we looked for the--

854
00:34:18,466 --> 00:34:20,566
for a derivative,

855
00:34:20,566 --> 00:34:25,100
because slope is delta-y
over delta-x...

856
00:34:26,266 --> 00:34:27,966
Not to be confused
with derivative,

857
00:34:27,966 --> 00:34:31,600
which is delta-y over delta-x.

858
00:34:31,600 --> 00:34:33,600
Sound the same?
It is the same.

859
00:34:33,600 --> 00:34:36,100
But--but--

860
00:34:36,100 --> 00:34:38,233
one is easier to publish with,

861
00:34:38,233 --> 00:34:40,933
'cause it sounds very
sophisticated: "derivative."

862
00:34:40,933 --> 00:34:44,566
And one is much--sounds like
you're in junior high,

863
00:34:44,566 --> 00:34:45,733
you're talking about "slope."

864
00:34:45,733 --> 00:34:47,666
So we did derivatives.

865
00:34:47,666 --> 00:34:49,066
We didn't do slope.

866
00:34:49,066 --> 00:34:50,466
Forget the slide.

867
00:34:53,166 --> 00:34:55,466
So here's an example of things
which we looked for.

868
00:34:55,466 --> 00:34:57,600
It's derivative.

869
00:34:57,600 --> 00:34:59,566
So here is a trending decision.

870
00:34:59,566 --> 00:35:02,233
This one occurs
very frequently, right?

871
00:35:02,233 --> 00:35:04,866
So the term--so "feel sick"
occurred very frequently

872
00:35:04,866 --> 00:35:06,300
over time.

873
00:35:06,300 --> 00:35:08,500
This one did not
occur so frequently,

874
00:35:08,500 --> 00:35:09,766
but this one trends.

875
00:35:09,766 --> 00:35:11,100
The first one does not trend,

876

00:35:11,100 --> 00:35:12,400
because although it's high,

877
00:35:12,400 --> 00:35:15,133
there's not much change to it.

878
00:35:15,133 --> 00:35:17,833
We took those.

879
00:35:17,833 --> 00:35:20,833
We used Wikipedia
to map things to like--

880
00:35:20,833 --> 00:35:23,100
to like sections.

881
00:35:23,100 --> 00:35:24,300
And why Wikipedia?

882
00:35:24,300 --> 00:35:26,233
Because it's layman's terms.

883
00:35:26,233 --> 00:35:28,266
What's tweets written in?

884
00:35:28,266 --> 00:35:30,600
Do you see--
do you see "neoplasm"

885
00:35:30,600 --> 00:35:33,633
or "sarcoma"

886
00:35:33,633 --> 00:35:35,000
listed in tweets?

887
00:35:35,000 --> 00:35:37,866
No, you see words like "cancer."

888

00:35:37,866 --> 00:35:40,366
So we wanted a clean--
use clean English.

889
00:35:42,066 --> 00:35:45,633
We looked for where in Wikipedia

890
00:35:45,633 --> 00:35:48,100
it mapped onto a concept.

891
00:35:48,100 --> 00:35:49,833
So here's the words
that mapped on.

892
00:35:49,833 --> 00:35:52,466
"Sore throat" we knew is
actually a medical thing.

893
00:35:52,466 --> 00:35:54,800
How do we know that?

894
00:35:54,800 --> 00:35:56,000
We cheat.

895
00:35:58,066 --> 00:35:59,500
See the red circle?

896
00:35:59,500 --> 00:36:02,500
It says ICD-9 and ICD-10 codes.

897
00:36:02,500 --> 00:36:05,766
What are ICD-9
and ICD-10 codes?

898
00:36:05,766 --> 00:36:08,600
Billing codes--

899
00:36:08,600 --> 00:36:10,766
billing codes for medicine.

900
00:36:10,766 --> 00:36:12,166
Now it's ICD-10.

901
00:36:12,166 --> 00:36:15,000
Previously it was ICD-9.

902
00:36:15,000 --> 00:36:16,800
If it has a billing code,

903
00:36:16,800 --> 00:36:18,433
it's a medical condition...

904
00:36:20,200 --> 00:36:21,533
Mostly likely.

905
00:36:23,566 --> 00:36:27,600
And we compared reaction to flu.

906
00:36:27,600 --> 00:36:29,566
And as you can see,

907
00:36:29,566 --> 00:36:33,400
the corrected Google,
and the CDC, and ours--

908
00:36:33,400 --> 00:36:36,300
you do not--you shouldn't
compare the scale.

909
00:36:36,300 --> 00:36:37,666
But you should compare that--

910
00:36:37,666 --> 00:36:40,633
you see we detected the trends
at the right time.

911
00:36:40,633 --> 00:36:42,300

So we were a little optimistic

912
00:36:42,300 --> 00:36:44,866
and said,
"Maybe there's a hope for it."

913
00:36:44,866 --> 00:36:46,800
But I have to give you
a word of warning.

914
00:36:46,800 --> 00:36:50,900
Using tweets, while
they are helpful in times,

915
00:36:50,900 --> 00:36:53,800
they're not quite as helpful
in others.

916
00:36:53,800 --> 00:36:56,166
So it is true that
the landing on the Hudson

917
00:36:56,166 --> 00:36:57,566
and the Mumbai terror attacks

918
00:36:57,566 --> 00:37:00,433
were detected quickly
in Twitter.

919
00:37:00,433 --> 00:37:02,300
It is also true--

920
00:37:02,300 --> 00:37:03,500
slightly less accurate

921
00:37:03,500 --> 00:37:05,066
is the flu tweet detection.

922
00:37:06,866 --> 00:37:09,300

Hurricane Sandy
had a bunch of misses.

923
00:37:09,300 --> 00:37:10,866
In fact, some cases were--

924
00:37:10,866 --> 00:37:13,800
the misses were indicating
the wrong locations.

925
00:37:13,800 --> 00:37:15,466
And some locations
for the meetups--

926
00:37:15,466 --> 00:37:19,200
the wrong locations occurred in

927
00:37:19,200 --> 00:37:21,766
off to the east of New Jersey,

928
00:37:21,766 --> 00:37:23,066
which, if you know geography,

929
00:37:23,066 --> 00:37:26,066
is not exactly a very
useful place to meet.

930
00:37:27,666 --> 00:37:30,566
But most wonderful
was my favorite

931
00:37:30,566 --> 00:37:32,466
of the celebrity deaths,

932
00:37:32,466 --> 00:37:34,233
all 'cause of "Colbert Report."

933
00:37:34,233 --> 00:37:36,900
How many of you used to see

the "Colbert Report"?

934
00:37:39,000 --> 00:37:42,833
He had a show of speaking
to Jeff Goldblum from the dead,

935
00:37:42,833 --> 00:37:45,700
because on Twitter,
Jeff Goldblum was killed.

936
00:37:45,700 --> 00:37:47,033
Only problem?

937
00:37:47,033 --> 00:37:48,666
Not problem--only good thing is

938
00:37:48,666 --> 00:37:50,333
that Jeff Goldblum isn't dead.

939
00:37:51,833 --> 00:37:56,900
So, Colbert talked to
Jeff Goldblum from the dead,

940
00:37:56,900 --> 00:37:58,766
because obviously,
Twitter had killed him.

941
00:37:58,766 --> 00:38:02,166
So don't necessarily bank on it
being the case.

942
00:38:03,666 --> 00:38:07,933
So, we actually looked to see
what it can do.

943
00:38:07,933 --> 00:38:10,900
And we trended on "sinuses,"

944
00:38:10,900 --> 00:38:13,100

and we noticed
that there's a problem

945
00:38:13,100 --> 00:38:16,833
around the April
and May timeframe.

946
00:38:16,833 --> 00:38:18,566
We then look
"allergic response,"

947
00:38:18,566 --> 00:38:21,466
and we saw that that trended
along the April and May.

948
00:38:21,466 --> 00:38:23,500
And in fact,
so did food allergies

949
00:38:23,500 --> 00:38:25,333
along that same timeframe.

950
00:38:25,333 --> 00:38:28,166
So we said, "Maybe there's
something to it."

951
00:38:30,466 --> 00:38:33,466
But the truth is

952
00:38:33,466 --> 00:38:38,733
that what we used is not
to validate a situation,

953
00:38:38,733 --> 00:38:41,300
because if you validated
that Jeff Goldblum is dead,

954
00:38:41,300 --> 00:38:43,033
you would've validated it true.

955
00:38:43,033 --> 00:38:44,900
And it's false.

956
00:38:44,900 --> 00:38:47,666
Social media should not be
necessarily the guide

957
00:38:47,666 --> 00:38:49,466
for validation of a concept.

958
00:38:49,466 --> 00:38:51,466
But it should be a guide,
potentially,

959
00:38:51,466 --> 00:38:54,033
for telling you if something
potentially has changed.

960
00:38:54,033 --> 00:38:58,800
Now, it may be a false alarm,

961
00:38:58,800 --> 00:39:01,100
or it may reality.

962
00:39:01,100 --> 00:39:02,466
And if it's reality,

963
00:39:02,466 --> 00:39:04,700
you will detect it
by other sources.

964
00:39:04,700 --> 00:39:07,033
It at least will give you
something to do.

965
00:39:07,033 --> 00:39:08,700
It'll give you something
to explore.

966
00:39:08,700 --> 00:39:11,900
So it is very useful as
a hypothesis generator,

967
00:39:11,900 --> 00:39:15,433
not so useful as
a truth indicator.

968
00:39:15,433 --> 00:39:17,100
And that's what we did.

969
00:39:18,666 --> 00:39:20,833
And by "we,"

970
00:39:20,833 --> 00:39:22,300
I mean all these people,

971
00:39:22,300 --> 00:39:23,833
including Alek Kolcz

972
00:39:23,833 --> 00:39:25,100
when he was at Twitter

973
00:39:25,100 --> 00:39:27,266
and so on, and so forth

974
00:39:27,266 --> 00:39:29,333
along the way.

975
00:39:29,333 --> 00:39:32,366
And here are some publications.

976
00:39:32,366 --> 00:39:34,133
Again, student appreciated.

977
00:39:34,133 --> 00:39:36,166
He went not only to Cologne,

978
00:39:36,166 --> 00:39:37,766
but he continued
to go to Slovenia

979
00:39:37,766 --> 00:39:39,533
and had a two-and-a-half-week
vacation

980
00:39:39,533 --> 00:39:42,300
courtesy of yours truly.

981
00:39:44,566 --> 00:39:46,566
So, what do we do?

982
00:39:46,566 --> 00:39:49,033
I showed you that the whole

983
00:39:49,033 --> 00:39:50,600
is greater than
the sum of the parts.

984
00:39:50,600 --> 00:39:53,100
When you integrate things,
you can actually get somewhere.

985
00:39:53,100 --> 00:39:56,533
You can solve problems
you couldn't solve otherwise.

986
00:39:56,533 --> 00:39:57,866
I showed you that searching,

987
00:39:57,866 --> 00:39:59,366
which we all know
is very easy,

988
00:39:59,366 --> 00:40:01,466
is not so easy if you can't
really do the search

989
00:40:01,466 --> 00:40:04,133
And if you've got very,
very adverse spelling

990
00:40:04,133 --> 00:40:07,900
and grammar and conditions
along the way.

991
00:40:07,900 --> 00:40:11,000
And I showed you
that social media

992
00:40:11,000 --> 00:40:13,700
should be used as
a warning mechanism

993
00:40:13,700 --> 00:40:15,366
or an alarm indicator.

994
00:40:15,366 --> 00:40:17,366
But it may not be
the ideal situation

995
00:40:17,366 --> 00:40:19,166
for you to go solve

996
00:40:19,166 --> 00:40:22,000
via truth indication.

997
00:40:23,133 --> 00:40:25,900
So let me conclude with one...

998
00:40:25,900 --> 00:40:27,700
the way I always conclude.

999
00:40:30,000 --> 00:40:32,500
I always conclude
with a bunch of statements.

1000
00:40:32,500 --> 00:40:35,900
One, I--
I have three cardinal rules.

1001
00:40:35,900 --> 00:40:38,700
Rule number one

1002
00:40:38,700 --> 00:40:41,000
is always finish on time.

1003
00:40:41,000 --> 00:40:43,033
The reason it's important
to finish on time is,

1004
00:40:43,033 --> 00:40:45,000
if you don't finish on time,

1005
00:40:45,000 --> 00:40:47,566
people start to wonder

1006
00:40:47,566 --> 00:40:49,733
if you didn't know how
to organize your talk

1007
00:40:49,733 --> 00:40:52,100
or if you didn't organize it.

1008
00:40:52,100 --> 00:40:55,066
So rule number one,
finish on time.

1009
00:40:55,066 --> 00:40:57,933
Rule number two...

1010
00:40:57,933 --> 00:41:02,400
rule number two is
always leave room for questions.

1011
00:41:02,400 --> 00:41:04,700
And it's very important
to leave room for questions,

1012
00:41:04,700 --> 00:41:07,166
'cause if you don't leave room--
time for questions,

1013
00:41:07,166 --> 00:41:08,733
people will start to say,

1014
00:41:08,733 --> 00:41:11,133
"Okay, this was a canned speech.

1015
00:41:11,133 --> 00:41:14,300
Person had it rehearsed.

1016
00:41:14,300 --> 00:41:17,266
Was not--did not want
to answer any questions,

1017
00:41:17,266 --> 00:41:18,500
because they were afraid

1018
00:41:18,500 --> 00:41:19,933
that they'll be asked questions

1019
00:41:19,933 --> 00:41:21,433
that they don't know
how to answer."

1020
00:41:21,433 --> 00:41:23,500
So rule number two is,

1021
00:41:23,500 --> 00:41:25,333
always leave room for questions.

1022
00:41:25,333 --> 00:41:29,566

But not any rule is as important
as rule number three.

1023
00:41:29,566 --> 00:41:31,800
Rule number three is by far
the most important.

1024
00:41:31,800 --> 00:41:34,100
Rule number three is,
never leave room

1025
00:41:34,100 --> 00:41:35,766
for too many questions,

1026
00:41:35,766 --> 00:41:37,466
because if you do,

1027
00:41:37,466 --> 00:41:39,333
they will realize

1028
00:41:39,333 --> 00:41:41,266
that you didn't know what
you were talking about,

1029
00:41:41,266 --> 00:41:44,866
and they will ask the questions
that you cannot answer.

1030
00:41:44,866 --> 00:41:47,533
So I was told that
I would have a total

1031
00:41:47,533 --> 00:41:50,300
of about 45 minutes.

1032
00:41:50,300 --> 00:41:53,733
I was told I should
leave room for questions.

1033

00:41:53,733 --> 00:41:56,333
And it is now exactly
43 minute--

1034
00:41:56,333 --> 00:41:59,933
42 minutes and 25 seconds
into the talk.

1035
00:41:59,933 --> 00:42:04,100
So I very much
thank you for coming.

1036
00:42:04,100 --> 00:42:06,233
And now I open the floor
for questions.

1037
00:42:06,233 --> 00:42:08,700
[applause]

1038
00:42:08,700 --> 00:42:10,266
- Very good, very good.

1039
00:42:10,266 --> 00:42:13,600
[applause]

1040
00:42:13,600 --> 00:42:15,100
So great--great rules,

1041
00:42:15,100 --> 00:42:17,166
and so we have time
for questions

1042
00:42:17,166 --> 00:42:19,000
and maybe some deep questions

1043
00:42:19,000 --> 00:42:20,700
that will challenge
the speaker.

1044

00:42:20,700 --> 00:42:22,566
If you have a question,
please raise your hand

1045
00:42:22,566 --> 00:42:23,833
and wait for the microphone

1046
00:42:23,833 --> 00:42:25,300
and ask a question.

1047
00:42:32,800 --> 00:42:35,000
- In one of your
tables of results,

1048
00:42:35,000 --> 00:42:37,666
you had drop a character,
add a character,

1049
00:42:37,666 --> 00:42:40,533
drop multiple characters,
add multiple characters...

1050
00:42:42,466 --> 00:42:44,666
- Those are different tables,

1051
00:42:44,666 --> 00:42:46,500
but yeah.

1052
00:42:46,500 --> 00:42:49,100
- So it seems like
for misspellings,

1053
00:42:49,100 --> 00:42:51,100
you'd need combinations
of those.

1054
00:42:51,100 --> 00:42:55,000
Did you also include
dropping and adding, and--

1055
00:42:55,000 --> 00:42:57,466
or maybe that was
replacing, or...

1056
00:42:57,466 --> 00:42:59,500
- So we tried--we added--

1057
00:42:59,500 --> 00:43:02,366
we evaluated using
adding a character,

1058
00:43:02,366 --> 00:43:03,666
dropping a character,

1059
00:43:03,666 --> 00:43:06,233
replacing a character,
and random swaps of characters.

1060
00:43:06,233 --> 00:43:09,100
And we did it up to
a combination of four.

1061
00:43:09,100 --> 00:43:13,366
So we basically--if you actually
looked at what we searched with

1062
00:43:13,366 --> 00:43:17,000
and you compared it to what--

1063
00:43:17,000 --> 00:43:20,966
to actually what we really
meant to search with,

1064
00:43:20,966 --> 00:43:22,266
it was completely different.

1065
00:43:22,266 --> 00:43:23,933
You wouldn't recognize
some of the terms.

1066
00:43:23,933 --> 00:43:26,766
And we also used actual,
real logs of--

1067
00:43:26,766 --> 00:43:27,833
of the users

1068
00:43:27,833 --> 00:43:29,100
to compare as well.

1069
00:43:33,866 --> 00:43:37,400
- Right now, the FBI and CIA

1070
00:43:37,400 --> 00:43:39,700
are doing huge numbers
of searches

1071
00:43:39,700 --> 00:43:42,300
to try to figure out
if there are people

1072
00:43:42,300 --> 00:43:45,233
stalking us at this very moment,
et cetera.

1073
00:43:45,233 --> 00:43:46,766
Are you working on any of that,

1074
00:43:46,766 --> 00:43:48,333
and does some of the--

1075
00:43:48,333 --> 00:43:50,600
some of the approaches
that you're using

1076
00:43:50,600 --> 00:43:51,966
help to root out

1077
00:43:51,966 --> 00:43:55,233
this evil part of our society?

1078
00:43:55,233 --> 00:43:58,466
- Am I working on...

1079
00:43:58,466 --> 00:43:59,866
stalking people?

1080
00:43:59,866 --> 00:44:01,100
[laughter]

1081
00:44:01,100 --> 00:44:02,933
I try not to stalk people.

1082
00:44:02,933 --> 00:44:05,033
Right.

1083
00:44:05,033 --> 00:44:06,800
I work on technology.

1084
00:44:06,800 --> 00:44:10,133
Technology is used in
various different ways.

1085
00:44:10,133 --> 00:44:13,200
And I've built various
search systems along the way.

1086
00:44:13,200 --> 00:44:16,233
So I don't actually know

1087
00:44:16,233 --> 00:44:17,866
what people are using
the systems

1088
00:44:17,866 --> 00:44:20,300
or the algorithms or the--

that I've built.

1089
00:44:20,300 --> 00:44:21,666
So I really can't answer that.

1090
00:44:21,666 --> 00:44:23,533
But I try not to stalk people,

1091
00:44:23,533 --> 00:44:27,000
if that's the question--
answer to your question.

1092
00:44:28,466 --> 00:44:30,033
If I didn't answer
your question,

1093
00:44:30,033 --> 00:44:32,100
ask it again.

1094
00:44:32,100 --> 00:44:33,433
- I was thinking
more in terms of

1095
00:44:33,433 --> 00:44:35,833
utilizing the technology
that you're using

1096
00:44:35,833 --> 00:44:37,700
to sort of read
between the lines

1097
00:44:37,700 --> 00:44:41,033
in some of these
communication systems

1098
00:44:41,033 --> 00:44:43,833
that people are using.

1099
00:44:43,833 --> 00:44:45,666

Did you hear what I said?

1100
00:44:45,666 --> 00:44:47,500
- I heard some of it.
Could you please repeat it?

1101
00:44:47,500 --> 00:44:50,833
- Oh, I said, I was thinking
more of using your technology

1102
00:44:50,833 --> 00:44:52,400
to read between the lines

1103
00:44:52,400 --> 00:44:56,700
in some of these--some of these
communication systems--

1104
00:44:56,700 --> 00:44:59,300
you know, organizations like
ISIS is supposedly using

1105
00:44:59,300 --> 00:45:01,933
over the Internet
and whether or not--

1106
00:45:01,933 --> 00:45:03,800
I mean, you're not
directly involved,

1107
00:45:03,800 --> 00:45:06,133
is what you answered me
in saying, but--

1108
00:45:06,133 --> 00:45:09,533
so you're not working
with the FBI and the CIA,

1109
00:45:09,533 --> 00:45:10,933
as far as you--
as far as you know,

1110
00:45:10,933 --> 00:45:12,700
your technology is not
part of that?

1111
00:45:12,700 --> 00:45:13,933
[laughter]

1112
00:45:13,933 --> 00:45:15,100
- As far as I know--

1113
00:45:15,100 --> 00:45:16,433
- If it were,
you wouldn't tell us.

1114
00:45:16,433 --> 00:45:17,733
- As far as I know, no,

1115
00:45:17,733 --> 00:45:20,866
but I can tell you
that you can search...

1116
00:45:20,866 --> 00:45:23,566
you can search basic
different languages

1117
00:45:23,566 --> 00:45:25,500
along the way

1118
00:45:25,500 --> 00:45:28,033
using different algorithms
and different approaches,

1119
00:45:28,033 --> 00:45:30,633
some of which I've actually
used and developed.

1120
00:45:37,566 --> 00:45:39,366
- Thank you.

1121
00:45:39,366 --> 00:45:41,400
So what you are doing is great,

1122
00:45:41,400 --> 00:45:44,266
but before you can do this,

1123
00:45:44,266 --> 00:45:48,066
you have to have these
documents digitized

1124
00:45:48,066 --> 00:45:50,366
in this electronic form.

1125
00:45:50,366 --> 00:45:52,500
And that's a big barrier,

1126
00:45:52,500 --> 00:45:56,366
both technically
and, like, legally.

1127
00:45:56,366 --> 00:45:59,333
So not all documents

1128
00:45:59,333 --> 00:46:01,866
are available for this,
actually.

1129
00:46:01,866 --> 00:46:06,200
Maybe too few documents
are available for this.

1130
00:46:06,200 --> 00:46:08,300
So how about this barrier?

1131
00:46:08,300 --> 00:46:11,233
- So the reason that
we had the first part

1132
00:46:11,233 --> 00:46:14,200
that I talked about
is in order to try to get

1133
00:46:14,200 --> 00:46:16,700
"scanned documents"

1134
00:46:16,700 --> 00:46:19,666
into a form that you can
actually OCR some of it--

1135
00:46:19,666 --> 00:46:21,733
so you can actually do
some of this interpretation.

1136
00:46:21,733 --> 00:46:24,566
The goal is that
if you can do the OCR--

1137
00:46:24,566 --> 00:46:25,933
and it's a big "if"--

1138
00:46:25,933 --> 00:46:27,400
if you can do the OCR,

1139
00:46:27,400 --> 00:46:28,900
then you can use some of these

1140
00:46:28,900 --> 00:46:31,033
foreign language
search techniques,

1141
00:46:31,033 --> 00:46:33,800
or foreign garbled
search techniques,

1142
00:46:33,800 --> 00:46:36,566
to try to correct some
of the OCR errors on top of it,

1143
00:46:36,566 --> 00:46:38,533
or at least try to
search the documents

1144
00:46:38,533 --> 00:46:41,666
that have them, even if they're
poor OCR corrections.

1145
00:46:41,666 --> 00:46:45,366
But yes,
digitization is a process,

1146
00:46:45,366 --> 00:46:47,366
and OCR doesn't always work.

1147
00:46:47,366 --> 00:46:49,833
In fact, it works--
it often fails.

1148
00:46:49,833 --> 00:46:53,733
And some documents
from various collections,

1149
00:46:53,733 --> 00:46:56,833
we've actually tried to help
people actually OCR,

1150
00:46:56,833 --> 00:46:59,666
and we basically
put our hands up and said,

1151
00:46:59,666 --> 00:47:01,133
"Never gonna happen,"

1152
00:47:01,133 --> 00:47:03,433
at least never as far
as computer science,

1153

00:47:03,433 --> 00:47:05,433
which means five years.

1154
00:47:09,500 --> 00:47:11,700
- Are there any other
lessons you learned

1155
00:47:11,700 --> 00:47:14,033
from your work that you can
apply to other areas

1156
00:47:14,033 --> 00:47:15,766
of your life?

1157
00:47:15,766 --> 00:47:18,833
- Have I learned any lessons?

1158
00:47:18,833 --> 00:47:21,233
I learned a lot of lessons.

1159
00:47:21,233 --> 00:47:24,433
First and foremost,
when you do search technology,

1160
00:47:24,433 --> 00:47:26,966
you should actually
get your head--

1161
00:47:26,966 --> 00:47:29,366
or do any evaluation
of search development--

1162
00:47:29,366 --> 00:47:31,400
you better have a solid--

1163
00:47:31,400 --> 00:47:34,500
a benchmark to actually
evaluate your systems with.

1164

00:47:34,500 --> 00:47:36,866
Gold standard is--

1165
00:47:36,866 --> 00:47:38,633
there's no replacement
for gold standards.

1166
00:47:38,633 --> 00:47:42,200
Every approximation
that you have

1167
00:47:42,200 --> 00:47:44,366
pales to actually having
the reality of it.

1168
00:47:44,366 --> 00:47:47,500
I'm not saying that you
necessarily will always have it,

1169
00:47:47,500 --> 00:47:49,666
but you should try
to aspire it.

1170
00:47:49,666 --> 00:47:53,000
Another thing is, get really
solid graduate students.

1171
00:47:54,566 --> 00:47:56,566
Being an academic,

1172
00:47:56,566 --> 00:48:00,933
I made my...career,

1173
00:48:00,933 --> 00:48:02,700
or whatever it is,

1174
00:48:02,700 --> 00:48:05,700
based on the hard work

1175

00:48:05,700 --> 00:48:07,733
and ingenuity of
my graduate students,

1176
00:48:07,733 --> 00:48:08,900
which I'm indebted for--

1177
00:48:08,900 --> 00:48:10,833
and also my colleagues.

1178
00:48:10,833 --> 00:48:12,466
But really, the fundamental work

1179
00:48:12,466 --> 00:48:13,800
is done by
the graduate students,

1180
00:48:13,800 --> 00:48:15,333
so I guess--resources.

1181
00:48:15,333 --> 00:48:16,833
Get the resources you need
is probably

1182
00:48:16,833 --> 00:48:19,433
the best way of grouping
everything together.

1183
00:48:19,433 --> 00:48:21,433
So that's the short answer
for you.

1184
00:48:23,333 --> 00:48:25,366
- So, in your Twitter example,

1185
00:48:25,366 --> 00:48:28,433
you got a very large database
from Johns Hopkins.

1186

00:48:28,433 --> 00:48:31,466
Is it possibly to actually
apply your model

1187
00:48:31,466 --> 00:48:33,166
to live data,

1188
00:48:33,166 --> 00:48:36,266
and is it something that
anyone's planning to do

1189
00:48:36,266 --> 00:48:37,800
in the future?

1190
00:48:37,800 --> 00:48:39,000
- So it's interesting you ask.

1191
00:48:39,000 --> 00:48:40,933
The initial goal was
to try to do it

1192
00:48:40,933 --> 00:48:42,266
on live data.

1193
00:48:42,266 --> 00:48:43,700
In fact, we had an intern--

1194
00:48:43,700 --> 00:48:47,666
collaborator plus a researcher

1195
00:48:47,666 --> 00:48:50,533
at Twitter at the time

1196
00:48:50,533 --> 00:48:52,266
that was gonna try
to do it live.

1197
00:48:52,266 --> 00:48:53,800
We wanted to do--

what we wanted to do

1198
00:48:53,800 --> 00:48:56,300
was a real, live feed,
and basically be able to--

1199
00:48:56,300 --> 00:48:58,533
to parse any trend situation

1200
00:48:58,533 --> 00:48:59,866
for different domains.

1201
00:48:59,866 --> 00:49:01,933
You need a domain,
'cause we have to identify

1202
00:49:01,933 --> 00:49:03,433
the vocabulary and the like

1203
00:49:03,433 --> 00:49:04,933
of what you're trying to track.

1204
00:49:04,933 --> 00:49:08,900
But...that was the intent.

1205
00:49:08,900 --> 00:49:10,633
Hasn't gone very far.

1206
00:49:10,633 --> 00:49:14,433
We're trying--we're trying to do
some of that now.

1207
00:49:14,433 --> 00:49:16,033
But it's still in an infancy,

1208
00:49:16,033 --> 00:49:19,033
so I can't really tell you if
it actually will work or not.

1209
00:49:19,033 --> 00:49:22,433
That's why when I pieced
this talk together,

1210
00:49:22,433 --> 00:49:26,366
it was what is kind of ripe
for reality--

1211
00:49:26,366 --> 00:49:29,600
what is--what is in use,

1212
00:49:29,600 --> 00:49:31,633
and versus what is--
you hope will be eventually

1213
00:49:31,633 --> 00:49:33,000
ripe for reality.

1214
00:49:33,000 --> 00:49:34,500
- What are the major setbacks?

1215
00:49:36,333 --> 00:49:38,233
[inaudible]

1216
00:49:38,233 --> 00:49:40,966
- So vocabulary was one.

1217
00:49:40,966 --> 00:49:43,466
Two was noise.

1218
00:49:43,466 --> 00:49:45,166
You need
a strong enough signal--

1219
00:49:45,166 --> 00:49:48,100
you need a base,
enough general vocabulary,

1220

00:49:48,100 --> 00:49:50,800
so that you'll be able to--

1221
00:49:50,800 --> 00:49:52,800
be able to track
enough information.

1222
00:49:52,800 --> 00:49:54,133
You need it specific enough

1223
00:49:54,133 --> 00:49:57,133
so that basically you can
isolate it correctly.

1224
00:49:57,133 --> 00:49:59,433
And the problem comes in that
you need also enough noise

1225
00:49:59,433 --> 00:50:01,466
to be able to detect
in the live stream

1226
00:50:01,466 --> 00:50:04,433
that when it spikes
and when it goes down.

1227
00:50:04,433 --> 00:50:07,533
So isolating noise and non--

1228
00:50:07,533 --> 00:50:11,633
non-major topics was
the biggest problem.

1229
00:50:11,633 --> 00:50:13,733
If you have a major topic
like influenza,

1230
00:50:13,733 --> 00:50:15,600
no problem.

1231
00:50:15,600 --> 00:50:17,300
If you have a topic that
basically deals with

1232
00:50:17,300 --> 00:50:20,100
food-borne illnesses,
we didn't quite get as far,

1233
00:50:20,100 --> 00:50:22,566
because geographical coding
of food-borne illnesses

1234
00:50:22,566 --> 00:50:24,366
and people tweeting on it
was not sufficient

1235
00:50:24,366 --> 00:50:25,433
for us to catch it,

1236
00:50:25,433 --> 00:50:27,500
as of yet.

1237
00:50:27,500 --> 00:50:30,200
I'm hopeful

1238
00:50:30,200 --> 00:50:32,333
but cannot guarantee it for you.

1239
00:50:34,700 --> 00:50:36,366
Wow, now we're way over.

1240
00:50:36,366 --> 00:50:40,233
- So with that, please join me
in thanking Dr. Frieder.

1241
00:50:40,233 --> 00:50:41,733
[applause]
Thank you very much.

1242
00:50:41,733 --> 00:50:44,100
[applause]