EARTH-LIFE
SCIENCE
INSTITUTE

Nicholas Guttenberg

Understanding intelligence by trying to build it

1
00:00:15,259 --> 00:00:13,310
the subject of my talk was changed my

2
00:00:16,429 --> 00:00:15,269
mark just now because I actually wasn't

3
00:00:19,159 --> 00:00:16,439
going to talk about computational

4
00:00:21,769 --> 00:00:19,169
chemistry but actually it's a good way

5
00:00:23,390 --> 00:00:21,779
to introduce I was here at LC for a

6
00:00:25,760 --> 00:00:23,400
couple years as a research scientist and

7
00:00:27,169 --> 00:00:25,770
now I'm at a private company in Tokyo

8
00:00:29,359 --> 00:00:27,179
working on machine learning and

9
00:00:31,490 --> 00:00:29,369
artificial intelligence and one of the

10
00:00:34,369 --> 00:00:31,500
things I found to my surprise I still

11
00:00:36,410 --> 00:00:34,379
come to LC twice a week and all of these

12
00:00:38,450 --> 00:00:36,420
projects have opened up because of the

13
00:00:40,820 --> 00:00:38,460

stuff that I'm doing now in AI it's

14
00:00:43,750 --> 00:00:40,830
actually quite applicable to analyzing

15
00:00:45,889 --> 00:00:43,760
the kind of data that for example

16
00:00:50,360 --> 00:00:45,899
Elizabeth was talking about in her talk

17
00:00:52,040 --> 00:00:50,370
so we have a few limited observations of

18
00:00:54,380 --> 00:00:52,050
exoplanets and we want to know more

19
00:00:56,750 --> 00:00:54,390
about them and it turns out some of the

20
00:00:58,010 --> 00:00:56,760
most the recently developed techniques

21
00:01:01,540 --> 00:00:58,020
and artificial intelligence can actually

22
00:01:05,180 --> 00:01:01,550
help answer that it's everywhere now um

23
00:01:07,130 --> 00:01:05,190
so we had this sort of dead period in

24
00:01:09,529 --> 00:01:07,140
artificial intelligence where you didn't

25
00:01:12,020 --> 00:01:09,539
really hear much about it anymore

26
00:01:14,419 --> 00:01:12,030
and now in the last three years or so

27
00:01:17,690 --> 00:01:14,429
it's been completely expanding it's

28
00:01:19,640 --> 00:01:17,700
everywhere the news covers it companies

29
00:01:23,030 --> 00:01:19,650
are now using it for business purposes

30
00:01:24,980 --> 00:01:23,040
it's used in industry it's it has a lot

31
00:01:26,749 --> 00:01:24,990
of implications for things like privacy

32
00:01:28,010 --> 00:01:26,759
and society because governments apply it

33
00:01:31,130 --> 00:01:28,020
as well

34
00:01:33,260 --> 00:01:31,140
but the kind of artificial intelligence

35
00:01:35,660 --> 00:01:33,270
that people are working on now that

36
00:01:37,490 --> 00:01:35,670
people are developing now this sort of

37
00:01:40,010 --> 00:01:37,500
new artificial intelligence is very

38
00:01:42,679 --> 00:01:40,020

different than the sort of old style

39
00:01:44,270 --> 00:01:42,689
which is more like what you probably

40
00:01:49,850 --> 00:01:44,280
would have read about in science fiction

41
00:01:52,429 --> 00:01:49,860
novels so the mental image of machine

42
00:01:55,190 --> 00:01:52,439
intelligence had the sort of idea that

43
00:01:58,340 --> 00:01:55,200
it's logical and that it works by a

44
00:02:00,020 --> 00:01:58,350
deduction machines can't understand

45
00:02:02,780 --> 00:02:00,030
emotion but they can't understand how

46
00:02:06,940 --> 00:02:02,790
truth kind of applies and they can make

47
00:02:09,979 --> 00:02:06,950
very very wide-ranging sort of plans and

48
00:02:13,130 --> 00:02:09,989
actually ironically people tried to

49
00:02:15,319 --> 00:02:13,140
build that in the 60s and it didn't

50
00:02:18,259 --> 00:02:15,329
really work and we're kind of in a local

51
00:02:21,110 --> 00:02:18,269
way but it was very fragile so when you

52
00:02:22,970 --> 00:02:21,120
try to have it deal with raw data with

53
00:02:24,300 --> 00:02:22,980
images from a camera rather than

54
00:02:26,580 --> 00:02:24,310
something constructed in our

55
00:02:31,020 --> 00:02:26,590
put into shape by a human practitioner

56
00:02:32,730 --> 00:02:31,030
it just would suddenly break as a result

57
00:02:37,050 --> 00:02:32,740
of this people stopped funding AI

58
00:02:38,880 --> 00:02:37,060
research by and large in the 70s and at

59
00:02:40,710 --> 00:02:38,890
the same time there were techniques

60
00:02:43,830 --> 00:02:40,720
still being used in industry and

61
00:02:46,860 --> 00:02:43,840
business that did what we would just

62
00:02:49,590 --> 00:02:46,870
call statistics so they just analyzed

63
00:02:51,420 --> 00:02:49,600

data they tried to extract correlations

64
00:02:54,030 --> 00:02:51,430
from data and got more and more

65
00:02:56,610 --> 00:02:54,040
sophisticated with that at a recent

66
00:02:59,250 --> 00:02:56,620
point around 2000 those techniques

67
00:03:00,690 --> 00:02:59,260
really came into their own and with the

68
00:03:02,400 --> 00:03:00,700
increase in computing power some of the

69
00:03:05,340 --> 00:03:02,410
old techniques that were seen as more of

70
00:03:07,380 --> 00:03:05,350
an actual dedicated model of the brain

71
00:03:09,600 --> 00:03:07,390
or artificial intelligence kind of thing

72
00:03:10,770 --> 00:03:09,610
ended up almost being merged with these

73
00:03:11,910 --> 00:03:10,780
business techniques and these

74
00:03:13,740 --> 00:03:11,920
statistical techniques that had been

75
00:03:15,990 --> 00:03:13,750
steadily growing throughout this sort of

76
00:03:17,850 --> 00:03:16,000
dead zone of AI in the 70s and so now we

77
00:03:19,890 --> 00:03:17,860
have something that's very different

78
00:03:24,300 --> 00:03:19,900
than the science-fiction image of a kind

79
00:03:26,910 --> 00:03:24,310
of a logical deductive robot so in the

80
00:03:29,130 --> 00:03:26,920
old style of AI the idea is to

81
00:03:30,650 --> 00:03:29,140
explicitly represent human knowledge and

82
00:03:33,810 --> 00:03:30,660
then look at how human knowledge

83
00:03:35,190 --> 00:03:33,820
interacts so you have a few statements

84
00:03:36,750 --> 00:03:35,200
that you know are true and you have a

85
00:03:38,460 --> 00:03:36,760
few rules of how statements can be

86
00:03:40,259 --> 00:03:38,470
combined with each other and through

87
00:03:41,699 --> 00:03:40,269
that you generate everything that's true

88
00:03:44,220 --> 00:03:41,709

or everything that you could possibly

89
00:03:46,560 --> 00:03:44,230
know is true and that was sort of the

90
00:03:50,009 --> 00:03:46,570
mental image of what intelligence was

91
00:03:52,650 --> 00:03:50,019
like and the problem is this requires a

92
00:03:54,120 --> 00:03:52,660
lot of work from humans to set the thing

93
00:03:56,250 --> 00:03:54,130
up in a way that those rules are going

94
00:03:58,289 --> 00:03:56,260
to let the system go anywhere other than

95
00:04:00,479 --> 00:03:58,299
what was imagined so even though this

96
00:04:03,420 --> 00:04:00,489
could in principle extrapolate very far

97
00:04:05,130 --> 00:04:03,430
in practice it was brittle if you change

98
00:04:06,599 --> 00:04:05,140
things a little bit then suddenly these

99
00:04:09,240 --> 00:04:06,609
rules don't apply it doesn't really help

100
00:04:12,750 --> 00:04:09,250
you anymore the new approach basically

101
00:04:15,210 --> 00:04:12,760
is very hands-off human practitioners of

102
00:04:17,430 --> 00:04:15,220
it tried not to put anything in if at

103
00:04:21,659 --> 00:04:17,440
all possible let the Machine figure out

104
00:04:23,760 --> 00:04:21,669
what it can from raw data and the kind

105
00:04:27,480 --> 00:04:23,770
of message of this is the data itself

106
00:04:29,130 --> 00:04:27,490
contains almost a way to figure out how

107
00:04:30,750 --> 00:04:29,140
to process it if you just have enough

108
00:04:32,279 --> 00:04:30,760
data and you use robust enough

109
00:04:34,050 --> 00:04:32,289
techniques

110
00:04:37,500 --> 00:04:34,060
the result is you have a machine that

111
00:04:38,040 --> 00:04:37,510
almost acts like human intuition it can

112
00:04:40,080 --> 00:04:38,050
make

113
00:04:41,850 --> 00:04:40,090

guesses the guesses are generally

114
00:04:45,300 --> 00:04:41,860
correct or they're more correct than

115
00:04:47,939 --> 00:04:45,310
chance but at the same time it's really

116
00:04:50,460 --> 00:04:47,949
hard for the machine to explain why is

117
00:04:52,110 --> 00:04:50,470
this right why do I think this and that

118
00:04:54,360 --> 00:04:52,120
makes it very challenging for humans to

119
00:04:57,480 --> 00:04:54,370
interact with and to understand when

120
00:04:59,520 --> 00:04:57,490
things are going wrong so I want to kind

121
00:05:02,939 --> 00:04:59,530
of explain what this is

122
00:05:04,710 --> 00:05:02,949
concretely one of the the most common

123
00:05:06,270 --> 00:05:04,720
techniques now is using neural networks

124
00:05:07,680 --> 00:05:06,280
and these are on all sorts of different

125
00:05:10,920 --> 00:05:07,690
forms so this is just a sort of a

126
00:05:13,230 --> 00:05:10,930
prototype a toy toy example and there's

127
00:05:15,290 --> 00:05:13,240
you have some input to the network in

128
00:05:19,350 --> 00:05:15,300
this case an image of my cat

129
00:05:24,089 --> 00:05:19,360
that input is encoded in numbers so

130
00:05:25,830 --> 00:05:24,099
pixel values things like that and those

131
00:05:29,520 --> 00:05:25,840
numbers are then pushed through a

132
00:05:32,100 --> 00:05:29,530
network of successive computations so I

133
00:05:34,529 --> 00:05:32,110
start with these inputs in the next

134
00:05:36,839 --> 00:05:34,539
layer they become some new numbers in

135
00:05:39,270 --> 00:05:36,849
the next layer they become some new

136
00:05:41,610 --> 00:05:39,280
numbers and so on each of these layers

137
00:05:43,439 --> 00:05:41,620
is related to each other by a

138
00:05:46,499 --> 00:05:43,449

mathematical operation which I've

139
00:05:48,360 --> 00:05:46,509
defined and I put some kind of unknown

140
00:05:52,050 --> 00:05:48,370
values into that operation in this case

141
00:05:53,640 --> 00:05:52,060
these are these these weights so the way

142
00:05:56,730 --> 00:05:53,650
that the red layer goes to the green

143
00:06:00,600 --> 00:05:56,740
layer is specified by these weights and

144
00:06:02,850 --> 00:06:00,610
those weights will change in order to

145
00:06:04,860 --> 00:06:02,860
make the predictions better so right now

146
00:06:09,450 --> 00:06:04,870
this network is very bad it thinks this

147
00:06:12,360 --> 00:06:09,460
is a dog 70% chance but since I know how

148
00:06:14,640 --> 00:06:12,370
I got those numbers I can change the

149
00:06:16,740 --> 00:06:14,650
network to make it so that when shown

150
00:06:19,290 --> 00:06:16,750
the same thing now it's going to say cat

151
00:06:20,909 --> 00:06:19,300
and I can figure out by going back

152
00:06:22,950 --> 00:06:20,919
through the network by propagating those

153
00:06:24,990 --> 00:06:22,960
errors backwards through the same

154
00:06:26,399 --> 00:06:25,000
calculation I've just done I can figure

155
00:06:28,320 --> 00:06:26,409
out how each of those weights should

156
00:06:30,420 --> 00:06:28,330
change just a little bit to make it a

157
00:06:32,909 --> 00:06:30,430
little bit less mistaken on this one

158
00:06:34,800 --> 00:06:32,919
case and then in the future this is

159
00:06:38,369 --> 00:06:34,810
exaggerated but the next time I run it

160
00:06:40,260 --> 00:06:38,379
now it thinks probably a cat in practice

161
00:06:43,379 --> 00:06:40,270
if I do this on one image and if I just

162
00:06:46,230 --> 00:06:43,389
do this once the next cat may be a black

163
00:06:48,180 --> 00:06:46,240

cat it didn't help but if I do this on a

164
00:06:50,639 --> 00:06:48,190
hundred million cats and I do it over

165
00:06:51,820 --> 00:06:50,649
and over and over I gradually remove all

166
00:06:54,220 --> 00:06:51,830
of the ways that the network

167
00:06:56,650 --> 00:06:54,230
makes mistakes at least as covered by

168
00:06:58,480 --> 00:06:56,660
the data that I've given it but at the

169
00:07:03,220 --> 00:06:58,490
end of this it can't tell me why it

170
00:07:04,750 --> 00:07:03,230
thinks this is a cat so another property

171
00:07:08,920 --> 00:07:04,760
of these is that they're very good at

172
00:07:10,960 --> 00:07:08,930
filling in gaps in between what it's

173
00:07:13,120 --> 00:07:10,970
shown so if I show it a black cat and a

174
00:07:14,980 --> 00:07:13,130
gray cat and an orange cat and so on it

175
00:07:16,900 --> 00:07:14,990
can kind of figure out what's between

176
00:07:19,300 --> 00:07:16,910
them but if I show it a very different

177
00:07:21,510 --> 00:07:19,310
kind of cat instead of showing a you

178
00:07:24,940 --> 00:07:21,520
know a house cat I show it a cheetah

179
00:07:26,410 --> 00:07:24,950
then it's hopeless it's never seen

180
00:07:28,690 --> 00:07:26,420
anything like it it doesn't know how to

181
00:07:31,240 --> 00:07:28,700
fix the errors that it has built into it

182
00:07:34,060 --> 00:07:31,250
about how to process a cheetah so it's

183
00:07:35,980 --> 00:07:34,070
very good at this sort of inside of what

184
00:07:38,170 --> 00:07:35,990
it sees but outside of what it sees like

185
00:07:42,430 --> 00:07:38,180
here I've taken one of these things and

186
00:07:44,500 --> 00:07:42,440
I've just trained it on the blue data in

187
00:07:46,510 --> 00:07:44,510
between that's fine outside it doesn't

188
00:07:48,130 --> 00:07:46,520

extrapolate at all it doesn't understand

189
00:07:51,400 --> 00:07:48,140
this pattern should be repeated over and

190
00:07:53,550 --> 00:07:51,410
over and over which means that in the

191
00:07:55,930 --> 00:07:53,560
end of the day this kind of AI is very

192
00:07:58,150 --> 00:07:55,940
strictly limited by the quality of the

193
00:08:00,490 --> 00:07:58,160
data that it can be provided it can't

194
00:08:02,590 --> 00:08:00,500
become infinitely good just sitting in a

195
00:08:04,690 --> 00:08:02,600
room it has to constantly be receiving

196
00:08:06,670 --> 00:08:04,700
some kind of feedback and information

197
00:08:10,810 --> 00:08:06,680
from its environment and that controls

198
00:08:12,610 --> 00:08:10,820
what happens google recently did a study

199
00:08:15,580 --> 00:08:12,620
where they took all of the images from

200
00:08:17,770 --> 00:08:15,590
YouTube and all these photo sharing

201
00:08:19,960 --> 00:08:17,780
sites and things like that made a data

202
00:08:21,970 --> 00:08:19,970
set with 300 million images and that

203
00:08:23,920 --> 00:08:21,980
still performed noticeably better than

204
00:08:25,540 --> 00:08:23,930
the best AI people could generate with a

205
00:08:27,850 --> 00:08:25,550
hundred million images with ten million

206
00:08:30,070 --> 00:08:27,860
images and there's just a simple

207
00:08:32,620 --> 00:08:30,080
logarithmic scaling of performance with

208
00:08:34,540 --> 00:08:32,630
the amount of data so these things are

209
00:08:36,490 --> 00:08:34,550
very data bald necked

210
00:08:39,130 --> 00:08:36,500
that determines how good the thing you

211
00:08:41,980 --> 00:08:39,140
get is going to be all right what's

212
00:08:43,180 --> 00:08:41,990
going on inside of the network they're

213
00:08:45,640 --> 00:08:43,190

hard to interpret but they're not

214
00:08:47,620 --> 00:08:45,650
impossible the nice thing about having

215
00:08:49,450 --> 00:08:47,630
something on a computer is we can cut it

216
00:08:50,890 --> 00:08:49,460
open and see what's going on exactly and

217
00:08:54,220 --> 00:08:50,900
we can change something a little bit and

218
00:08:57,040 --> 00:08:54,230
see how that changes and the sort of

219
00:08:59,350 --> 00:08:57,050
image that has emerged from studying

220
00:09:00,730 --> 00:08:59,360
these things empirically is that what

221
00:09:03,220 --> 00:09:00,740
they really do is they filter out

222
00:09:05,660 --> 00:09:03,230
irrelevant information so in that first

223
00:09:08,180 --> 00:09:05,670
layer they keep

224
00:09:08,900 --> 00:09:08,190
as much as they can but then the next

225
00:09:10,700 --> 00:09:08,910
layer down

226
00:09:13,010 --> 00:09:10,710
they've discarded a few things that

227
00:09:14,180 --> 00:09:13,020
ended up not being very constructive to

228
00:09:15,590 --> 00:09:14,190
the question that was being asked and

229
00:09:18,020 --> 00:09:15,600
the next layer down they discard a bit

230
00:09:20,270 --> 00:09:18,030
more and a bit more and at the end the

231
00:09:22,400 --> 00:09:20,280
only information that exists in the

232
00:09:24,050 --> 00:09:22,410
network at the very last layer is what

233
00:09:26,480 --> 00:09:24,060
the network needs to answer the

234
00:09:28,400 --> 00:09:26,490
questions it's been asked but

235
00:09:30,500 --> 00:09:28,410
interestingly at the top at the very

236
00:09:34,160 --> 00:09:30,510
first layer it throws out things that

237
00:09:36,590 --> 00:09:34,170
are just generally not useful so in this

238
00:09:38,540 --> 00:09:36,600

case this is a network that's trained to

239
00:09:41,750 --> 00:09:38,550
identify the age and gender of a face

240
00:09:44,390 --> 00:09:41,760
and at the top layer you can see it can

241
00:09:46,430 --> 00:09:44,400
still see the rims of my glasses it can

242
00:09:48,680 --> 00:09:46,440
see my mouth all of these different

243
00:09:50,300 --> 00:09:48,690
things are the activations of one neuron

244
00:09:52,430 --> 00:09:50,310
in the network and I've just picked

245
00:09:54,950 --> 00:09:52,440
seven or seven neurons for that layer to

246
00:09:56,630 --> 00:09:54,960
visualize but there's 128 of them I

247
00:09:58,490 --> 00:09:56,640
think the way that it's actually

248
00:10:00,470 --> 00:09:58,500
detecting is if I actually look at it it

249
00:10:03,230 --> 00:10:00,480
finds very general features of images

250
00:10:05,570 --> 00:10:03,240
that are robust like edges circles

251
00:10:08,480 --> 00:10:05,580
corners things like that but it's not

252
00:10:11,390 --> 00:10:08,490
very responsive to say a noise pattern

253
00:10:14,150 --> 00:10:11,400
and that's because even at the most

254
00:10:15,890 --> 00:10:14,160
simple level ignoring noise is just a

255
00:10:17,180 --> 00:10:15,900
generally good strategy to being able to

256
00:10:18,860 --> 00:10:17,190
answer the question deeper in the

257
00:10:20,900 --> 00:10:18,870
network but as I go further and further

258
00:10:22,160 --> 00:10:20,910
down I stop being able to actually

259
00:10:24,620 --> 00:10:22,170
reconstruct some of these features

260
00:10:26,870 --> 00:10:24,630
anymore I have almost lost my glasses

261
00:10:28,700 --> 00:10:26,880
here here they're pretty much gone here

262
00:10:31,250 --> 00:10:28,710
I can't even see my face anymore there's

263
00:10:32,870 --> 00:10:31,260

just some statistical correlations that

264
00:10:34,370 --> 00:10:32,880
the network's held on to that are

265
00:10:37,040 --> 00:10:34,380
informative about age and gender

266
00:10:39,440 --> 00:10:37,050
so networks are generalized that is they

267
00:10:41,540 --> 00:10:39,450
they can work on things that they

268
00:10:43,280 --> 00:10:41,550
haven't seen because they're throwing

269
00:10:44,780 --> 00:10:43,290
away all of the distracting factors that

270
00:10:46,430 --> 00:10:44,790
actually make those things different by

271
00:10:48,230 --> 00:10:46,440
the end of the network all of these

272
00:10:49,430 --> 00:10:48,240
faces are basically the same face as

273
00:10:53,510 --> 00:10:49,440
long as they have the same age and

274
00:10:55,790 --> 00:10:53,520
gender so you guys what do these things

275
00:10:57,650 --> 00:10:55,800
actually know if a network can classify

276
00:10:59,870 --> 00:10:57,660
something does that mean it understands

277
00:11:03,710 --> 00:10:59,880
what it is there's the technique for

278
00:11:06,020 --> 00:11:03,720
actually asking the network or asking it

279
00:11:07,730 --> 00:11:06,030
to reconstruct an input that would

280
00:11:10,970 --> 00:11:07,740
convince it that this is this kind of

281
00:11:13,400 --> 00:11:10,980
class so in this case for example this

282
00:11:15,380 --> 00:11:13,410
is the network that's freely available

283
00:11:16,970 --> 00:11:15,390
people trained to have a very large data

284
00:11:19,069 --> 00:11:16,980
set called imagenet

285
00:11:21,379 --> 00:11:19,079
and it has a thousand different objects

286
00:11:23,389 --> 00:11:21,389
that it knows how to recognize I picked

287
00:11:25,310 --> 00:11:23,399
three that I thought would be very

288
00:11:28,610 --> 00:11:25,320

visually distinctive and I generated

289
00:11:30,259 --> 00:11:28,620
these three reconstructions this is what

290
00:11:33,079 --> 00:11:30,269
it thinks an image that would convince

291
00:11:35,689 --> 00:11:33,089
it it's a Pekingese dog and you can see

292
00:11:39,519 --> 00:11:35,699
there's some eyes and maybe some of the

293
00:11:44,509 --> 00:11:42,620
this is a flamingo and you can see the

294
00:11:46,129 --> 00:11:44,519
kind of the bird shape and some wings

295
00:11:48,500 --> 00:11:46,139
and the color patterns of a flamingo but

296
00:11:49,970 --> 00:11:48,510
again it's not like a flamingo situated

297
00:11:52,879 --> 00:11:49,980
in a realistic background it's not a

298
00:11:54,769 --> 00:11:52,889
full image it's just bits of the idea of

299
00:11:58,370 --> 00:11:54,779
a flamingo that are relevant to it that

300
00:12:00,350 --> 00:11:58,380
it thinks are indicative and the same

301
00:12:02,090 --> 00:12:00,360
for a snake where you can see scales you

302
00:12:04,720 --> 00:12:02,100
can see the kind of looks like the belly

303
00:12:07,819 --> 00:12:04,730
scales and an eye and maybe the face

304
00:12:09,740 --> 00:12:07,829
alright so in the in the case of that

305
00:12:12,019 --> 00:12:09,750
reconstruction that's what the network

306
00:12:15,319 --> 00:12:12,029
knows but it doesn't know that it knows

307
00:12:17,090 --> 00:12:15,329
the network has no sort of introspective

308
00:12:20,210 --> 00:12:17,100
process where it queries itself it just

309
00:12:21,769 --> 00:12:20,220
does things it's like when you when you

310
00:12:23,000 --> 00:12:21,779
catch a ball you don't have time to

311
00:12:24,379 --> 00:12:23,010
think about what are you going to do you

312
00:12:26,420 --> 00:12:24,389
just do it and then you can look at it

313
00:12:28,460 --> 00:12:26,430

after the fact and say this is what it

314
00:12:29,900 --> 00:12:28,470
felt like to catch a ball the network I

315
00:12:32,750 --> 00:12:29,910
described didn't have that mechanism

316
00:12:34,540 --> 00:12:32,760
it's just a it's just like your visual

317
00:12:37,639 --> 00:12:34,550
system or something very reflexive

318
00:12:39,319 --> 00:12:37,649
however we can intentionally design

319
00:12:40,939 --> 00:12:39,329
networks that have this kind of

320
00:12:43,879 --> 00:12:40,949
reflective mechanism or this kind of

321
00:12:47,150 --> 00:12:43,889
introspective mechanism one of the

322
00:12:49,280 --> 00:12:47,160
things that lets us do this we have a

323
00:12:50,689 --> 00:12:49,290
large amount of data available to us in

324
00:12:52,610 --> 00:12:50,699
our senses all the time but we can only

325
00:12:55,519 --> 00:12:52,620
pay attention to a little bit and we

326
00:12:58,100 --> 00:12:55,529
direct that attention to pay attention

327
00:12:59,960 --> 00:12:58,110
to what's to look at or to only take the

328
00:13:03,590 --> 00:12:59,970
data that's relevant to what we're

329
00:13:05,480 --> 00:13:03,600
trying to do the network that's static

330
00:13:08,150 --> 00:13:05,490
that I described before all of that

331
00:13:09,860 --> 00:13:08,160
stuff is frozen in it says I'm going to

332
00:13:12,470 --> 00:13:09,870
discard this information at this layer

333
00:13:14,840 --> 00:13:12,480
that's how I am but you can make it

334
00:13:16,819 --> 00:13:14,850
dynamic instead you can ask the network

335
00:13:19,430 --> 00:13:16,829
to direct what it discards based on the

336
00:13:22,220 --> 00:13:19,440
current circumstances and this kind of

337
00:13:24,880 --> 00:13:22,230
attentional model is actually really

338
00:13:28,579 --> 00:13:24,890

good for processing human language so

339
00:13:30,269 --> 00:13:28,589
the the newest Google state-of-the-art

340
00:13:32,369 --> 00:13:30,279
machine translation uses

341
00:13:34,259 --> 00:13:32,379
attention to model the relationship

342
00:13:36,449 --> 00:13:34,269
between words in the sentence and you

343
00:13:39,179 --> 00:13:36,459
can train it on about a 10 thousand

344
00:13:40,590 --> 00:13:39,189
times less in computer cycles than if

345
00:13:43,050 --> 00:13:40,600
you do one that doesn't have this kind

346
00:13:45,090 --> 00:13:43,060
of attention mechanism so this makes a

347
00:13:46,739 --> 00:13:45,100
really big difference to the performance

348
00:13:49,980 --> 00:13:46,749
of this kind of of this kind of

349
00:13:52,110 --> 00:13:49,990
technology all of what I've been talking

350
00:13:53,850 --> 00:13:52,120
about is networks that essentially get

351
00:13:56,100 --> 00:13:53,860
rid of information they discard what

352
00:13:58,139 --> 00:13:56,110
they don't care about you might ask well

353
00:14:00,389 --> 00:13:58,149
if I want to make something that creates

354
00:14:02,819 --> 00:14:00,399
new information if I want a creative AI

355
00:14:04,439 --> 00:14:02,829
could I do that and there's a technique

356
00:14:07,499 --> 00:14:04,449
that came out a few years ago called

357
00:14:09,059 --> 00:14:07,509
generative adversarial networks the

358
00:14:10,619 --> 00:14:09,069
technique here is that you have two

359
00:14:12,840 --> 00:14:10,629
networks that interact with each other

360
00:14:15,269 --> 00:14:12,850
and they essentially fight to try to

361
00:14:17,369 --> 00:14:15,279
fool each other one network produces a

362
00:14:18,840 --> 00:14:17,379
fake image and the other Network tries

363
00:14:21,689 --> 00:14:18,850

to learn how to tell the difference

364
00:14:24,929 --> 00:14:21,699
between reality and fake images and as a

365
00:14:27,119 --> 00:14:24,939
result they sort of explore the space of

366
00:14:30,030 --> 00:14:27,129
images to figure out what makes an image

367
00:14:32,610 --> 00:14:30,040
realistic and you can get very nice

368
00:14:34,920 --> 00:14:32,620
results even with a very small amount of

369
00:14:37,319 --> 00:14:34,930
data for instance this these are all

370
00:14:39,689 --> 00:14:37,329
fake all of these butterflies are

371
00:14:42,780 --> 00:14:39,699
generated by the generative part of the

372
00:14:44,610 --> 00:14:42,790
network from a noise pattern however the

373
00:14:47,579 --> 00:14:44,620
generator and discriminator were trained

374
00:14:50,519 --> 00:14:47,589
on a data set of I think it's three

375
00:14:53,220 --> 00:14:50,529
thousand three thousand butterfly images

376
00:14:56,220 --> 00:14:53,230
that look very similar to this but are

377
00:14:58,470 --> 00:14:56,230
different in detail as a result of this

378
00:15:01,889 --> 00:14:58,480
the model that the generator learns

379
00:15:03,900 --> 00:15:01,899
allows us to even generate intervening

380
00:15:06,389 --> 00:15:03,910
butterflies we can generate a butterfly

381
00:15:08,309 --> 00:15:06,399
halfway between two species or halfway

382
00:15:10,379 --> 00:15:08,319
between two genders about a fly halfway

383
00:15:12,329 --> 00:15:10,389
between front and back very strange

384
00:15:14,400 --> 00:15:12,339
things like that that don't exist in the

385
00:15:16,650 --> 00:15:14,410
world but that the understanding that

386
00:15:21,509 --> 00:15:16,660
the machine has developed allows it to

387
00:15:23,040 --> 00:15:21,519
imagine another interesting thing about

388
00:15:25,470 --> 00:15:23,050

this is once you have these creative

389
00:15:28,079 --> 00:15:25,480
machines well what's the relationship

390
00:15:30,239 --> 00:15:28,089
between that and human artists so you

391
00:15:32,189 --> 00:15:30,249
could just try to let the machine do

392
00:15:35,879 --> 00:15:32,199
everything but a more sort of

393
00:15:38,189 --> 00:15:35,889
interesting or a useful way to go about

394
00:15:40,829 --> 00:15:38,199
this is to ask what would a human artist

395
00:15:42,269 --> 00:15:40,839
do when given access to the machine the

396
00:15:45,100 --> 00:15:42,279
nice thing about this kind of technique

397
00:15:48,110 --> 00:15:45,110
is you can give it cute

398
00:15:49,910 --> 00:15:48,120
you can say I want you to draw something

399
00:15:52,189 --> 00:15:49,920
that fits within a certain outline so

400
00:15:55,309 --> 00:15:52,199
this is something called pics - pics and

401
00:15:57,139 --> 00:15:55,319
I drew this very bad cat and it

402
00:16:00,019 --> 00:15:57,149
generated this it filled in the details

403
00:16:02,420 --> 00:16:00,029
or if I give it something that's not a

404
00:16:04,069 --> 00:16:02,430
cat it still tries to make it into

405
00:16:06,079 --> 00:16:04,079
something like you cat it takes my cue

406
00:16:09,350 --> 00:16:06,089
and then it's just another tool I can

407
00:16:10,999 --> 00:16:09,360
use like a paintbrush this is an example

408
00:16:13,819 --> 00:16:11,009
of a piece of software you can download

409
00:16:16,660 --> 00:16:13,829
this it involves installing quite a lot

410
00:16:18,829 --> 00:16:16,670
of stuff unfortunately but it

411
00:16:20,780 --> 00:16:18,839
understands mountains at least to some

412
00:16:23,930 --> 00:16:20,790
degree and it can take cues of the color

413
00:16:26,300 --> 00:16:23,940

of the sky or the grass things like that

414
00:16:30,019 --> 00:16:26,310
and generate a mountain responding to

415
00:16:32,180 --> 00:16:30,029
your to your input all of this stuff is

416
00:16:34,759 --> 00:16:32,190
kind of data processing still even the

417
00:16:36,050 --> 00:16:34,769
generative stuff another place that

418
00:16:38,600 --> 00:16:36,060
people try to use artificial

419
00:16:40,819 --> 00:16:38,610
intelligence is to drive behaviors so a

420
00:16:43,699 --> 00:16:40,829
big example of this is self-driving cars

421
00:16:45,740 --> 00:16:43,709
you want the machine to control the

422
00:16:47,420 --> 00:16:45,750
motion of the vehicle it has to choose

423
00:16:49,850 --> 00:16:47,430
where to go it has to choose how to

424
00:16:51,800 --> 00:16:49,860
respond to a situation it has to

425
00:16:55,670 --> 00:16:51,810
generate actions not just correctly

426
00:16:58,370 --> 00:16:55,680
classify a perceptual input and this

427
00:17:00,439 --> 00:16:58,380
creates a really big problem when I want

428
00:17:02,120 --> 00:17:00,449
to classify perceptual input I can be

429
00:17:04,100 --> 00:17:02,130
like Google and collect 300 million

430
00:17:07,039 --> 00:17:04,110
images independent of any artificial

431
00:17:08,899 --> 00:17:07,049
intelligence development I can just

432
00:17:11,449 --> 00:17:08,909
gather as much data as I want and I'll

433
00:17:13,039 --> 00:17:11,459
get a very good AI out of that if I want

434
00:17:16,880 --> 00:17:13,049
something that's actually taking action

435
00:17:19,039 --> 00:17:16,890
it needs to know what its actions would

436
00:17:21,500 --> 00:17:19,049
do so if I want to understand how to

437
00:17:23,990 --> 00:17:21,510
make a safe self-driving car that

438
00:17:26,120 --> 00:17:24,000

doesn't or that can say recover from a

439
00:17:28,100 --> 00:17:26,130
near crash situation like if it starts

440
00:17:29,899 --> 00:17:28,110
to swerve off the road I want it to be

441
00:17:32,200 --> 00:17:29,909
able to recover I need to collect data

442
00:17:33,860 --> 00:17:32,210
of that car swerving off the road

443
00:17:36,350 --> 00:17:33,870
because these things can only

444
00:17:37,789 --> 00:17:36,360
interpolate so they need to have some

445
00:17:39,890 --> 00:17:37,799
experience of the situation's they're

446
00:17:42,470 --> 00:17:39,900
like please deal with or that you want

447
00:17:45,020 --> 00:17:42,480
them to behave competently in and that

448
00:17:46,940 --> 00:17:45,030
means for this kind of AI right now it's

449
00:17:49,970 --> 00:17:46,950
very far behind the sort of perceptual

450
00:17:52,010 --> 00:17:49,980
AI because of this data limit because

451
00:17:55,520 --> 00:17:52,020
you actually need to engage the AI in

452
00:17:57,080 --> 00:17:55,530
those real world control situations in

453
00:17:58,269 --> 00:17:57,090
order for it to actually have what it

454
00:18:01,879 --> 00:17:58,279
needs to

455
00:18:03,590 --> 00:18:01,889
alright so in conclusion the the main

456
00:18:05,299 --> 00:18:03,600
thing about modern AI compared to the

457
00:18:07,159 --> 00:18:05,309
sort of science-fiction images it's all

458
00:18:10,099 --> 00:18:07,169
experience driven it's all very

459
00:18:12,169 --> 00:18:10,109
intuitive it's essentially learning from

460
00:18:14,869 --> 00:18:12,179
doing a massive statistical analysis of

461
00:18:18,229 --> 00:18:14,879
lifetime or many human lifetimes worth

462
00:18:20,479 --> 00:18:18,239
of data that it's given but it doesn't

463
00:18:22,789 --> 00:18:20,489

actually work by deduction it doesn't

464
00:18:25,310 --> 00:18:22,799
work by following into some kind of

465
00:18:27,979 --> 00:18:25,320
extrapolate Ori story or what we would

466
00:18:30,589 --> 00:18:27,989
generally call understanding the nice

467
00:18:32,659 --> 00:18:30,599
thing about this setup is that it means

468
00:18:34,190 --> 00:18:32,669
you don't actually have to understand

469
00:18:35,330 --> 00:18:34,200
how to do the task you want to teach it

470
00:18:38,149 --> 00:18:35,340
you just have to have a lot of examples

471
00:18:40,460 --> 00:18:38,159
and there's a lot of flexibility in that

472
00:18:42,560 --> 00:18:40,470
you can set up the network with very

473
00:18:45,859 --> 00:18:42,570
complicated structures and let it take

474
00:18:48,289 --> 00:18:45,869
care of itself given the data and given

475
00:18:49,940 --> 00:18:48,299
a network the optimization process of

476
00:18:51,979 --> 00:18:49,950
fixing all the small errors bit by bit

477
00:18:53,419 --> 00:18:51,989
will take it to some kind of functional

478
00:18:55,099 --> 00:18:53,429
state and you don't actually have to

479
00:18:56,810 --> 00:18:55,109
understand how it's doing it at the end

480
00:18:58,700 --> 00:18:56,820
you just have to set up a circumstance

481
00:19:00,279 --> 00:18:58,710
where if it works you know it and you

482
00:19:03,320 --> 00:19:00,289
get what you want

483
00:19:05,810 --> 00:19:03,330
however because it can't extrapolate

484
00:19:07,969 --> 00:19:05,820
this puts a really strong limit on where

485
00:19:10,249 --> 00:19:07,979
we can easily use it versus where we

486
00:19:11,989 --> 00:19:10,259
might like to use it but it's not really

487
00:19:14,659 --> 00:19:11,999
going to give us what we want if we want

488
00:19:16,820 --> 00:19:14,669

to do things in very new situations

489
00:19:18,259 --> 00:19:16,830
that's a really big problem right now

490
00:19:20,359 --> 00:19:18,269
and learning how to make an actual

491
00:19:22,249 --> 00:19:20,369
functional extrapolate of AI I think

492
00:19:25,039 --> 00:19:22,259
right now is is one of the key problems

493
00:19:26,479 --> 00:19:25,049
and the data requirements pose certain

494
00:19:28,879 --> 00:19:26,489
limits on where and when you can

495
00:19:33,940 --> 00:19:28,889
actually use these technologies so thank