

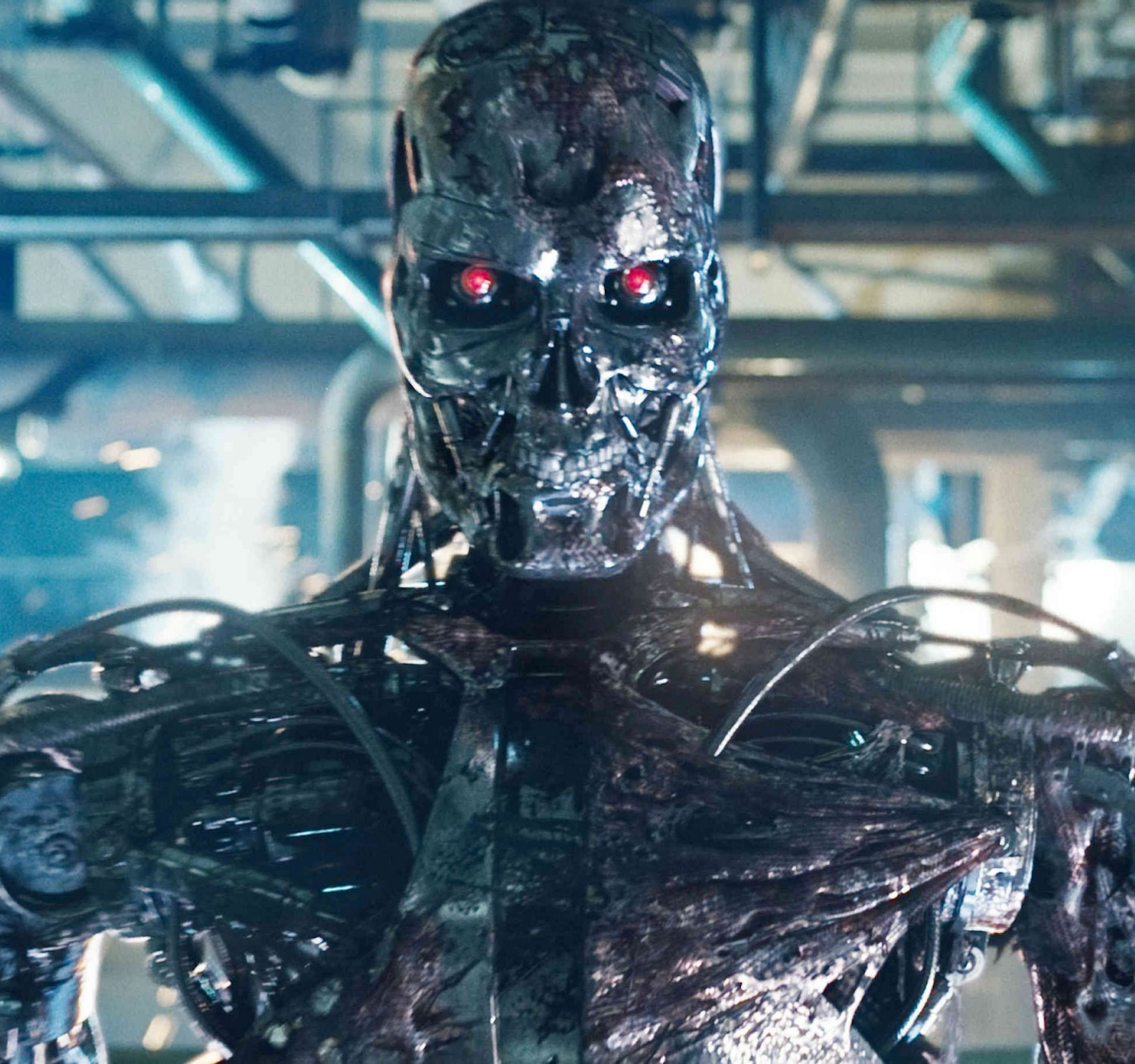


# Data Science and Expertise: **Case of COVID-19**

Rajiv Shah



# What? Who?



**Not AI.**

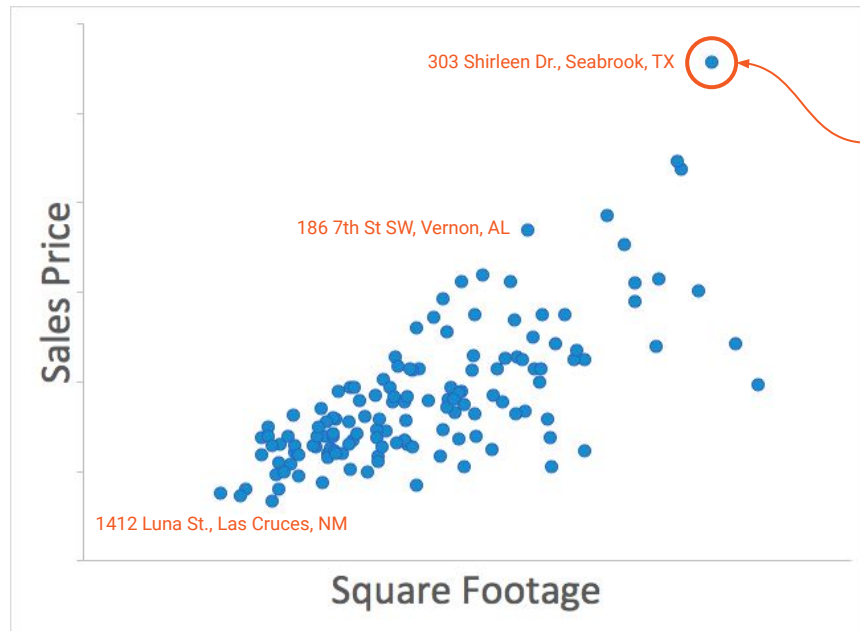


# AI - 200 years



Christian Albrecht Jensen:  
[Portrait of] **Carl Friedrich Gauss**, 1840

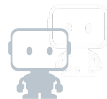
# Machine learning models



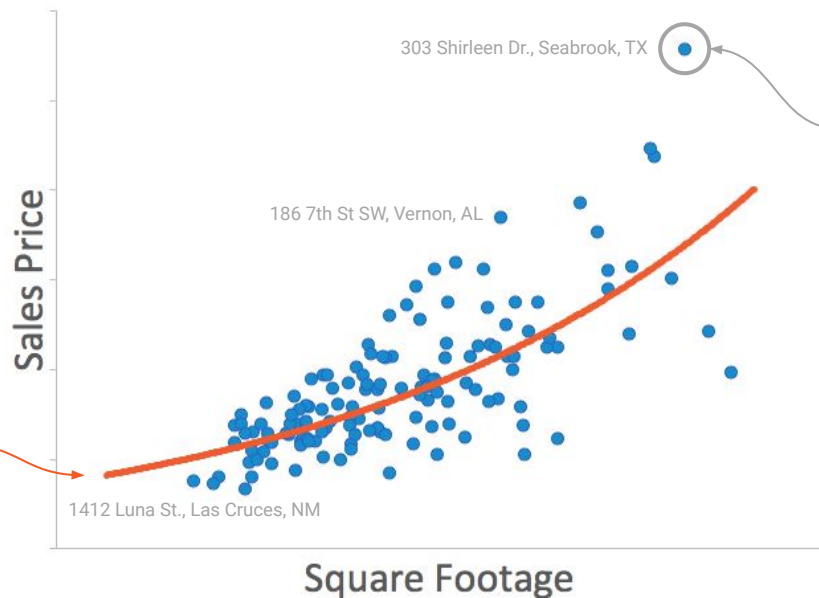
Every point is a  
**historical example**  
that the machine  
learns from



# Machine learning models

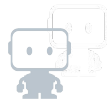


The line is a model. If you tell the model a square footage, it will make a prediction.

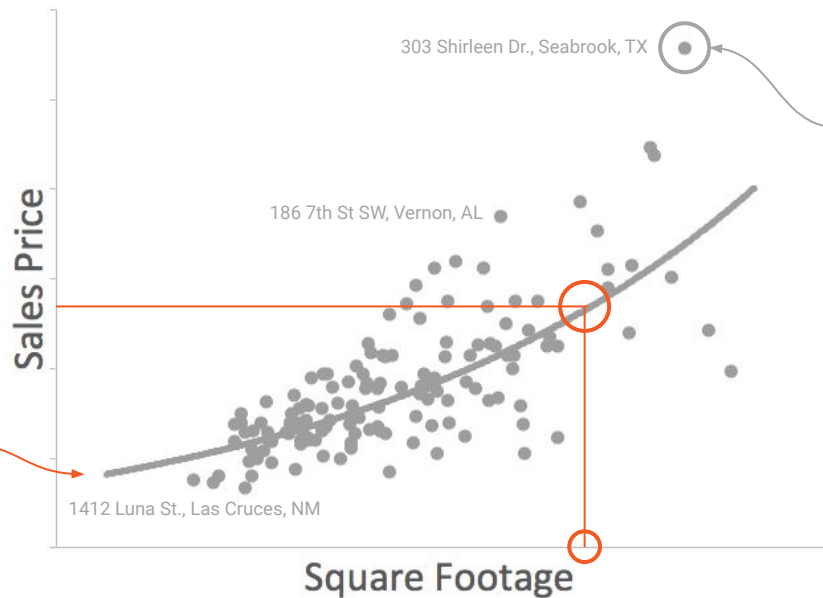


Every point is a **historical example** that the machine learns from

# Machine learning models



The line is a model. If you tell the model a square footage, it will make a prediction.



Every point is a **historical example** that the machine learns from

# Reality is more complex than one variable



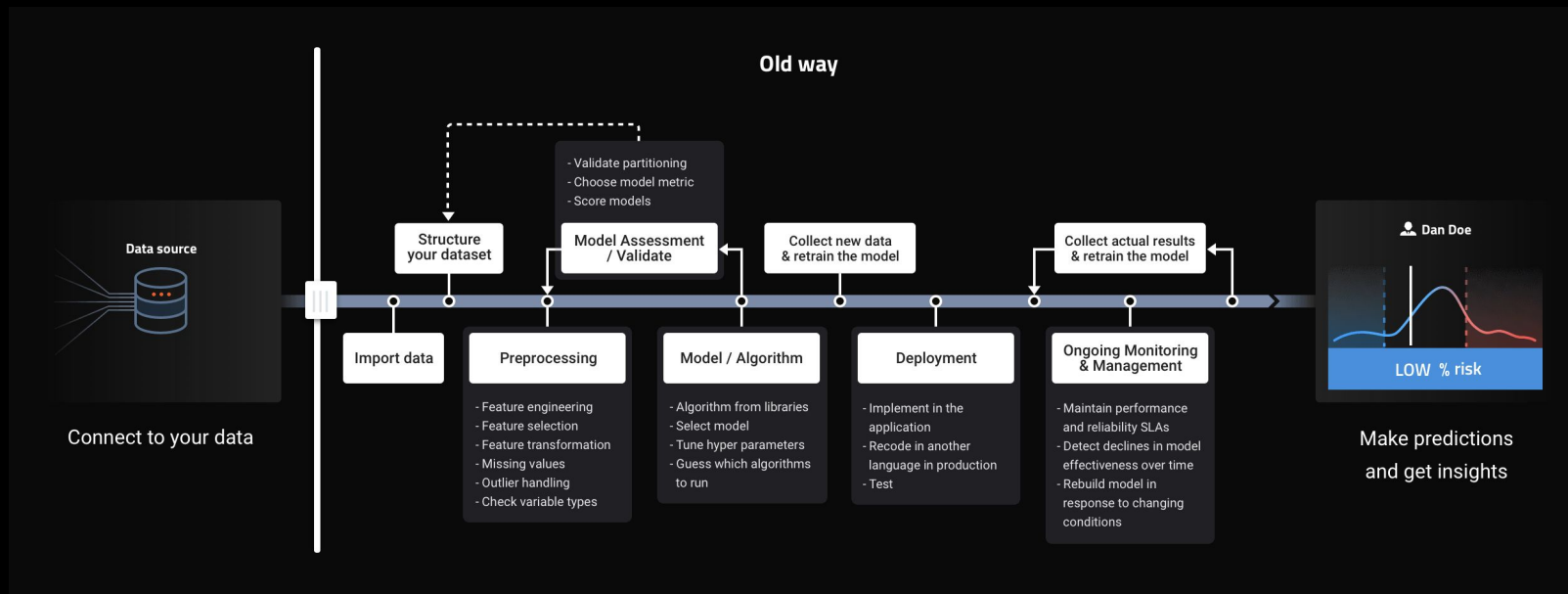
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	LotArea	LandSlope	BldgType	HouseStyle	YearBuilt	YearRemodA	RoofStyle	Exterior1st	Foundation	TotalBsmtSF	Heating	GrLivArea	FullBath	HalfBath	YrSold	SalePrice
2	8450	Gtl		2Story	2003	2003			PConc	856		1710	2	1	2008	
3	9600	Gtl	1Fam	1Story	1976	1976	Gable	MetalSd	CBlock	1262	GasA	1262	2	0	2007	181500
4	11250	Gtl	1Fam	2Story	2001		Gable	VinylSd	PConc	920	GasA	1786	2	1	2008	223500
5	9550	Gtl	1Fam	2Story	1915	1970	Gable	Wd Sdng	BrkTil	756	GasA	1717	1	0		140000
6	14260	Gtl	1Fam	2Story	2000	2000	Gable	VinylSd	PConc	1145	GasA	2198	2	1	2008	250000
7	14115	Gtl	1Fam	1.5Fin		1995	Gable	VinylSd	Wood		GasA		1	1	2009	143000
8	10084	Gtl		1Story		2005				1686	GasA		2	0	2007	307000
9	10382	Gtl	1Fam	2Story		1973	Gable		CBlock	1107	GasA		2	1	2009	200000
10	6120	Gtl	1Fam	1.5Fin	1931	1950	Gable	BrkFace	BrkTil	952	GasA	1774	2	0	2008	129900
11	7420	Gtl	2fmCon	1.5Unf	1939	1950	Gable	MetalSd	BrkTil	991	GasA	1077	1	0	2008	118000
12	11200	Gtl	1Fam	1Story	1965	1965	Hip	HdBoard	CBlock	1040	GasA	1040	1	0	2008	
13	11924	Gtl	1Fam	2Story	2005	2006	Hip	WdShing	PConc	1175	GasA	2324	3	0	2006	345000
14	12968	Gtl	1Fam	1Story	1962	1962		HdBoard	CBlock	912	GasA	912	1	0	2008	144000
15	10652	Gtl	1Fam	1Story	2006	2007	Gable	VinylSd	PConc	1494	GasA	1494	2	0	2007	279500
16	10920	Gtl	1Fam	1Story	1960	1960	Hip	MetalSd	CBlock	1253	GasA	1253	1	1	2008	157000
17	6120	Gtl	1Fam	1.5Unf	1929	2001	Gable	Wd Sdng	BrkTil	832	GasA	854	1	0	2007	132000
18	11241	Gtl	1Fam	1Story	1970		Gable	Wd Sdng	CBlock	1004	GasA	1004	1	0	2010	149000
19	10791	Gtl	Duplex	1Story	1967	1967	Gable	MetalSd	Slab	0	GasA	1296	2	0	2006	90000
20	13695	Gtl	1Fam		2004	2004	Gable	VinylSd	PConc	1114	GasA	1114	1	1	2008	159000
21	7560	Gtl	1Fam	1Story	1958	1965	Hip	BrkFace	CBlock	1029	GasA	1339	1	0	2009	139000
22	14215	Gtl		2Story	2005	2006	Gable	VinylSd	PConc	1158	GasA	2376	3	1	2006	325300
23	7449	Gtl		1.5Unf	1930	1950	Gable	Wd Sdng	PConc	637	GasA	1108	1	0	2007	

The column **SalePrice** is the target to predict

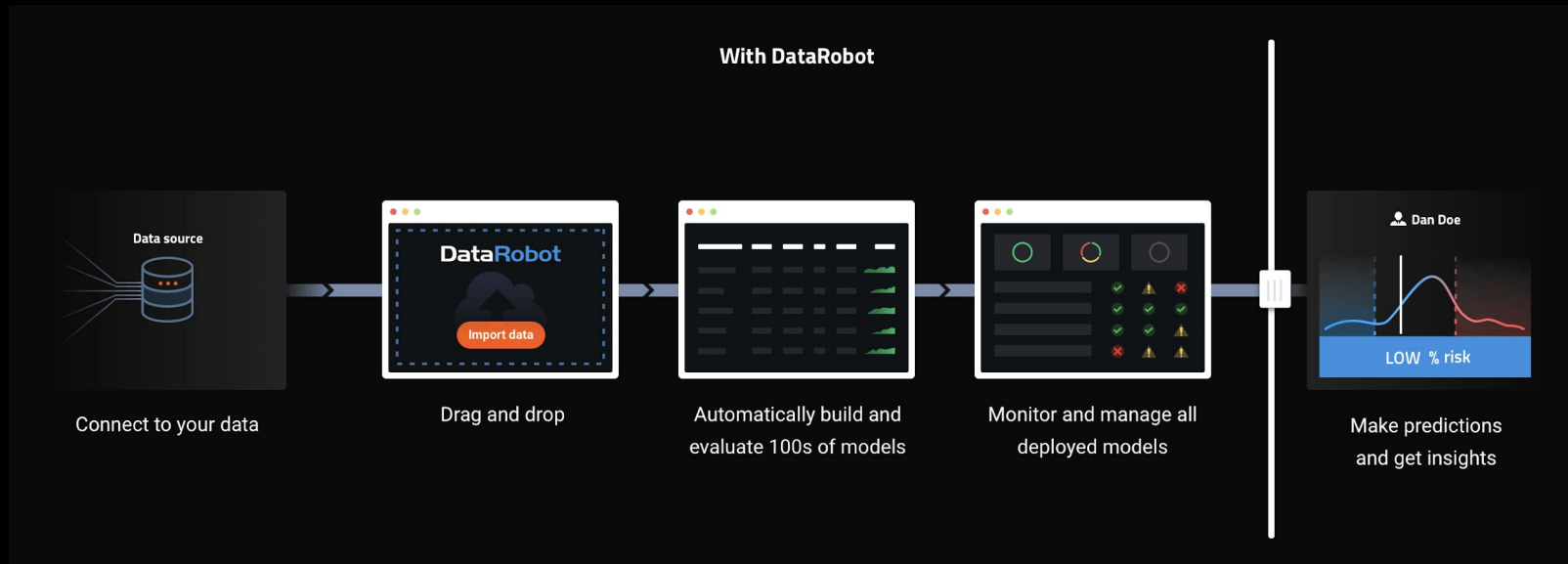




# Lowering the barrier to AI - AutoML



# Lowering the barrier to AI - AutoML



# Expertise in Real Estate



## House Hunters - Full Episodes

Watch full episodes of HGTV's long-running series *House Hunters* — right here at HGTV.com.



CONTINUE:

**Choosy in Chesapeake**



WISH LIST:

Him: studio space



8:34





# COVID-19

# Data Science and COVID-19

A screenshot of a web browser displaying the DataRobot COVID-19 Response Effort page. The browser's address bar shows the URL 'datarobot.com/lp/covid-19-response-effort/'. The page features a background image of healthcare workers in blue scrubs and masks. The DataRobot logo is in the top left, and the tagline 'Enabling the AI-Driven Enterprise' is in the top right. The main heading is 'DataRobot COVID-19 Response Effort'. Below it, there are three paragraphs of text.

DataRobot

Enabling the AI-Driven Enterprise

## DataRobot COVID-19 Response Effort

In unprecedented times like these, we all must do what we can to help. That's why DataRobot is providing its services pro bono to help with the response to the COVID-19 virus.

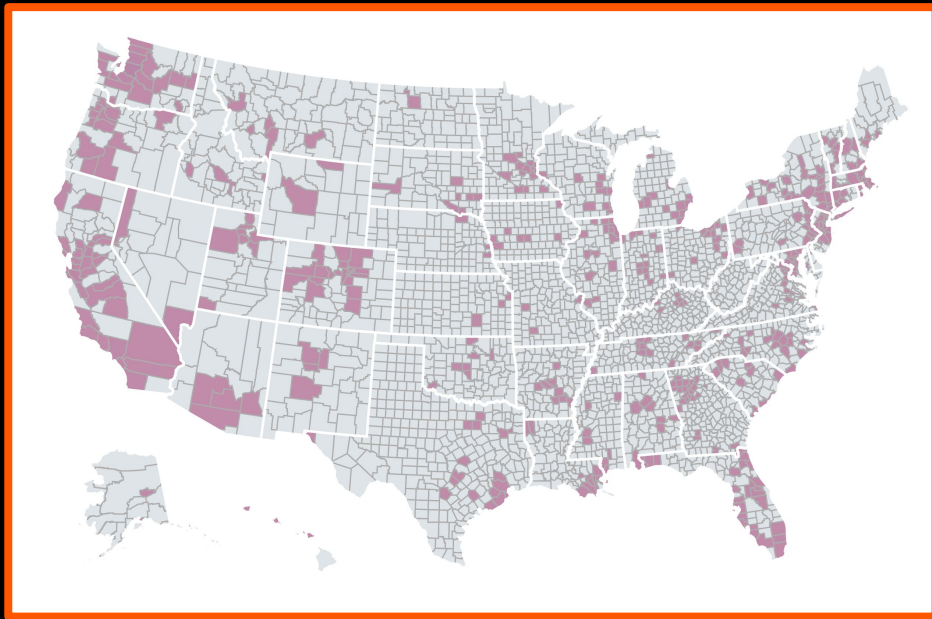
Inspired by the passion of our employees, partners, and customers, we believe that the data science and machine learning community can turn expertise into powerful predictions that can benefit the lives of many.

Choose your path below to start using one of DataRobot's products for the greater good.

# COVID-19 in the United States



**Research Question:**  
**Which counties in the**  
**United States have COVID?**



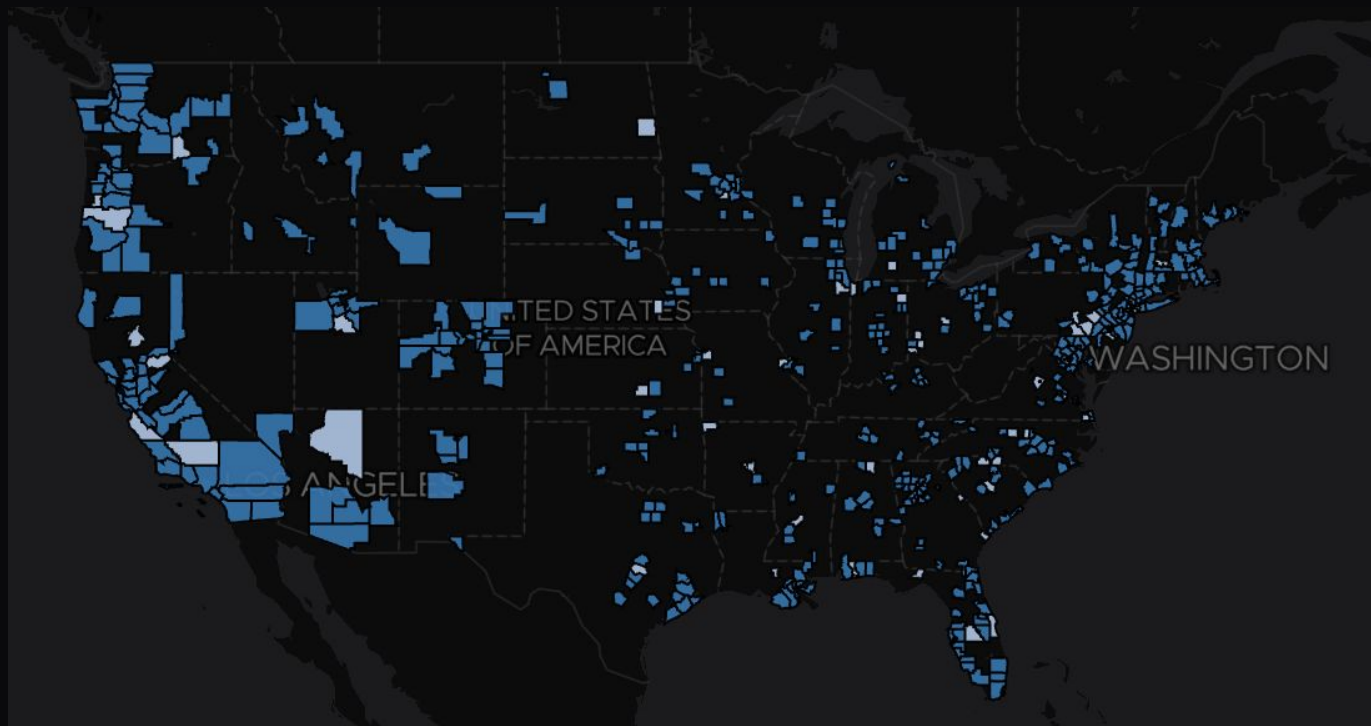
<https://blog.datarobot.com/predicting-the-covid-19-on-the-u.s.-county-level>



# Daily Predictions to Identify Counties to Have Covid-19



This map shows the 449 counties that are currently infected in dark blue and the predicted 50 high risk counties in light blue.



# Daily Predictions for Covid-19



*[UPDATE] We are releasing new U.S. county predictions based on the data available today  
(03/25/2020):*

MO	Platte	SD	Lincoln	VA	Roanoke
KS	Shawnee	TX	Jefferson	ND	Grand Forks
TX	Randall	OK	Rogers	TX	Rockwall
AZ	Mohave	GA	Jackson	CA.	Kings
VA	Montgomery	VA	Fauquier	ID	Bonneville
RI	Kent	IL	Tazewell	MO	Cape Girardeau
TX	Guadalupe	IN	Kosciusko		

# Start with Socio-Economic Data



- Easily available data
- Use this to identify the relationships between features and counties having COVID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	State	County_x	2013 Rural-ur	2013 Urban l	Less than a hi	High school d	Some college	Bachelor's de	Percent of ad	Percent of ad	Percent of ad	Percent of ad	Economic ty	have_confirm	POP_ESTI
2	AL	Autauga	2	2	4204	12119	10552	10291	11.3	32.6	28.4	27.7	0	FALSE	556
3	AL	Baldwin	3	2	14310	40579	46025	46075	9.7	27.6	31.3	31.3	5	TRUE	2180
4	AL	Barbour	6	6	4901	6486	4566	2220	27	35.7	25.1	12.2	3	FALSE	248
5	AL	Bibb	1	1	2650	7471	3846	1813	16.8	47.3	24.4	11.5	0	FALSE	224
6	AL	Blount	1	1	7861	13489	13267	5010	19.8	34	33.5	12.6	0	FALSE	578
7	AL	Bullock	6	6	1760	2817	1582	945	24.8	39.7	22.3	13.3	3	FALSE	101
8	AL	Butler	6	6	2141	6091	3421	2235	15.4	43.9	24.6	16.1	0	FALSE	196
9	AL	Calhoun	3	2	12620	25653	26643	14219	15.9	32.4	33.7	18	4	TRUE	1142
10	AL	Chambers	6	5	4383	9060	7003	3118	18.6	38.4	29.7	13.2	0	FALSE	336
11	AL	Cherokee	6	6	3692	7157	5417	2407	19.8	38.3	29	12.9	0	FALSE	260
12	AL	Chilton	1	1	5302	13125	7147	4219	17.8	44.1	24	14.2	0	FALSE	441
13	AL	Choctaw	9	10	1754	3606	2851	1225	18.6	38.2	30.2	13	3	FALSE	128
14	AL	Clarke	7	11	3175	7372	4205	2118	18.8	43.7	24.9	12.6	3	FALSE	239
15	AL	Clay	9	10	2275	3366	2700	944	24.5	36.3	29.1	10.2	3	FALSE	132
16	AL	Cleburne	8	4	2364	4135	2518	1428	22.6	39.6	24.1	13.7	0	FALSE	149
17	AL	Coffee	4	5	4833	9723	12181	8175	13.8	27.9	34.9	23.4	0	FALSE	519
18	AL	Colbert	3	2	6016	13927	11459	7236	15.6	36	29.7	18.7	0	FALSE	547
19	AL	Conecuh	7	11	1703	4315	1881	949	19.2	48.8	21.3	10.7	0	FALSE	122
20	AL	Coosa	8	3	1675	3178	2424	953	20.4	38.6	29.5	11.6	3	FALSE	107
21	AL	Covington	6	6	4530	9368	8588	3952	17.1	35.4	32.5	14.9	0	FALSE	369
22	AL	Crenshaw	8	6	2091	3753	2326	1505	21.6	38.8	24	15.6	3	FALSE	138
23	AL	Cullman	4	3	10405	20382	18559	7866	18.2	35.6	32.4	13.7	0	FALSE	834
24	AL	Dale	4	5	4732	10924	12036	5690	14.2	32.7	36.1	17	3	FALSE	489
25	AL	Dallas	4	5	5135	9368	8076	4021	19.3	35.2	30.4	15.1	3	FALSE	383
26	AL	DeKalb	6	6	12394	15956	13441	6085	25.9	33.3	28.1	12.7	3	FALSE	713
27	AL	Elmore	2	2	7594	18304	16197	13664	13.6	32.8	29	24.5	0	TRUE	818
28	AL	Escambia	6	6	4792	11546	6340	3287	18.5	44.5	24.4	12.7	0	FALSE	367
29	AL	Etowah	3	2	11147	23665	24416	12767	15.5	32.9	33.9	17.7	0	FALSE	1025

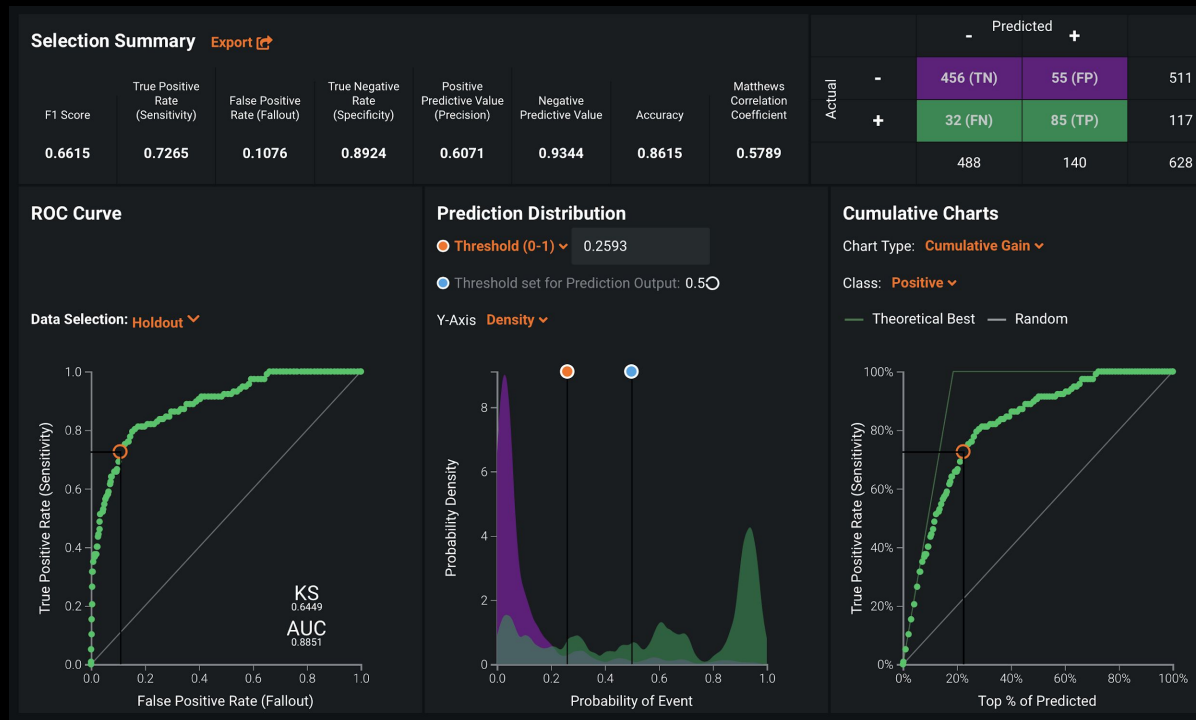
# Evaluating the Model



If we look at the top 50 highest predictions out of the model, 48 of those counties had a COVID case within the next ten 10 days.

88% precision rate for a 5 day forecast

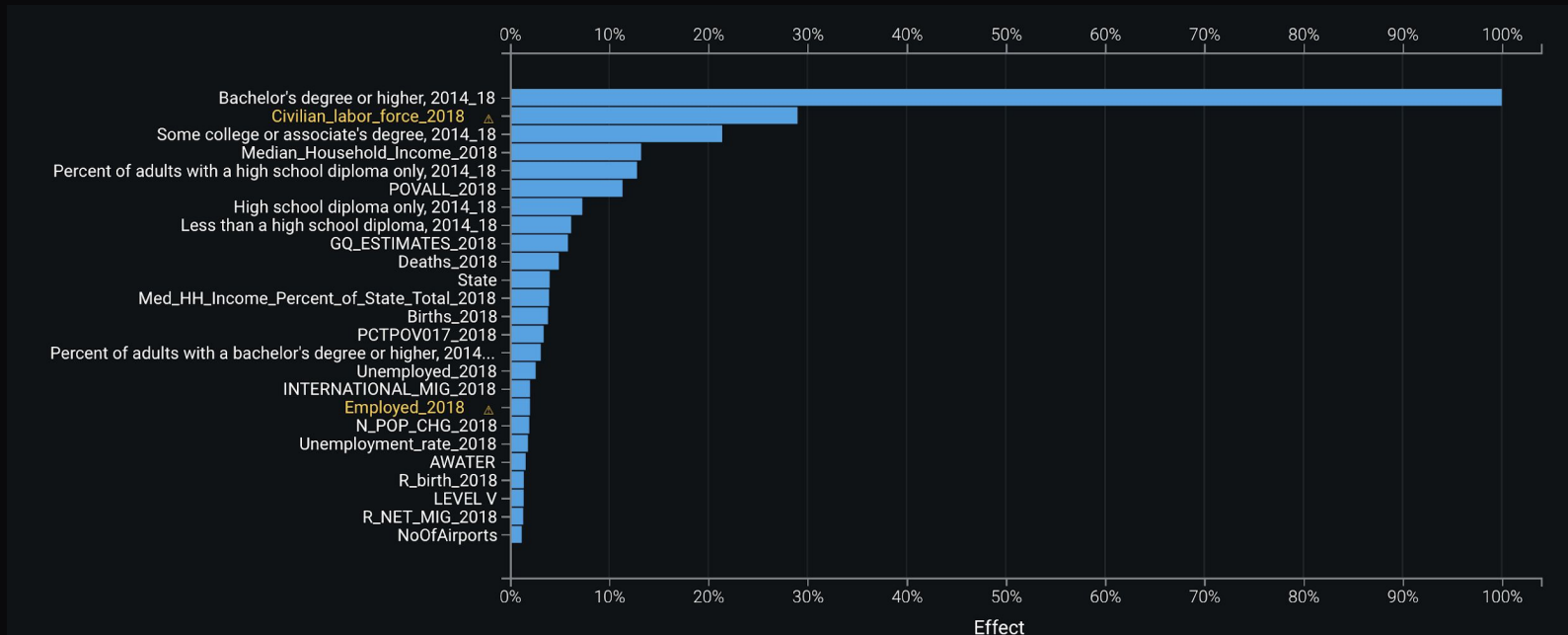
96% precision over a 10 day



# Explaining the Model



See that income and education play a strong role



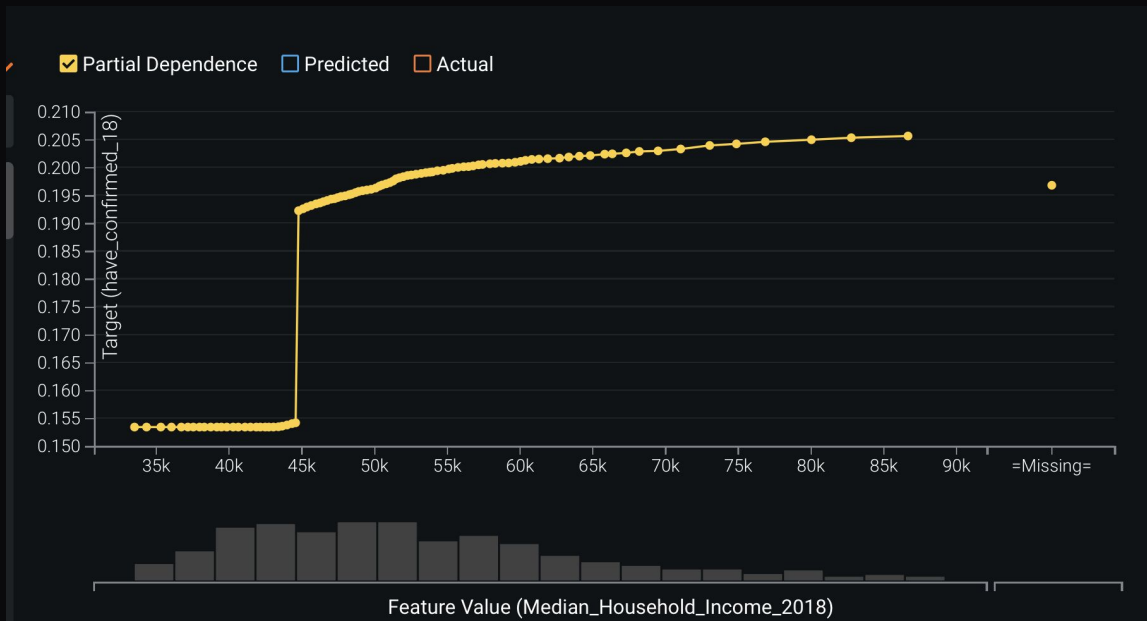
# Explaining the Model



Lower income counties are  
less likely to have COVID-19

...

Is that really all there is to it?





# Next Steps -- Second Generation



**Add data about airports, hospitals, and geo-spatial data**

New Patterns:

- State name has a larger effect
- Number of hospitals is getting more important. - It could indicate two things: infection spread is local now (thus the importance of state as a proxy for proximity) and efficiency of detection which definitely depends on state and number of hospitals.

# What data was used?



- **Johns Hopkins University's dataset of confirmed cases**
- **Existing U.S. county-level socioeconomic data**
- **U.S. county geo-coordinates**
- **News Break Coronavirus Realtime Updates**
- **Claritas demographics data**
- **County data from USAFacts**

# COVID-19 Chest X-rays




Detect which chest  
x-rays indicate COVID


Menu Search Feature List: All Features View Raw Data + Create Feature List

<input type="checkbox"/> Feature Name	Index	Importance <span>▼</span>	Var Type	Unique	Missing
<input type="checkbox"/> class	1	Target	Categorical	2	0
<input type="checkbox"/> image	2	<div></div>	Image	150	0

Image Preview Duplicates



non-covid  
120 rows



covid  
68 rows

<https://blog.datarobot.com/identifying-leakage-in-computer-vision-on-medical-images>

# Results




The model does very well, it predicts everything correctly, either as a True Negative or True Positive

It made one mistake and treated one x-ray as a False Positive

		Predicted		
		-	+	
Actual	-	94 (TN)	1 (FP)	95
	+	0 (FN)	55 (TP)	55
		94	56	150


# Results

Lots of amateurs building out “very accurate” models for detecting COVID-19 from chest x-rays

Ryan Gotesman Follow  
Mar 25 · 4 min read ★

① Anyone can publish on Medium per our Policies, but we don't fact-check every story. For more info about the coronavirus, see [cdc.gov](#).

## Prototyping a Neural Network to diagnose Covid-19 from Chest X-ray



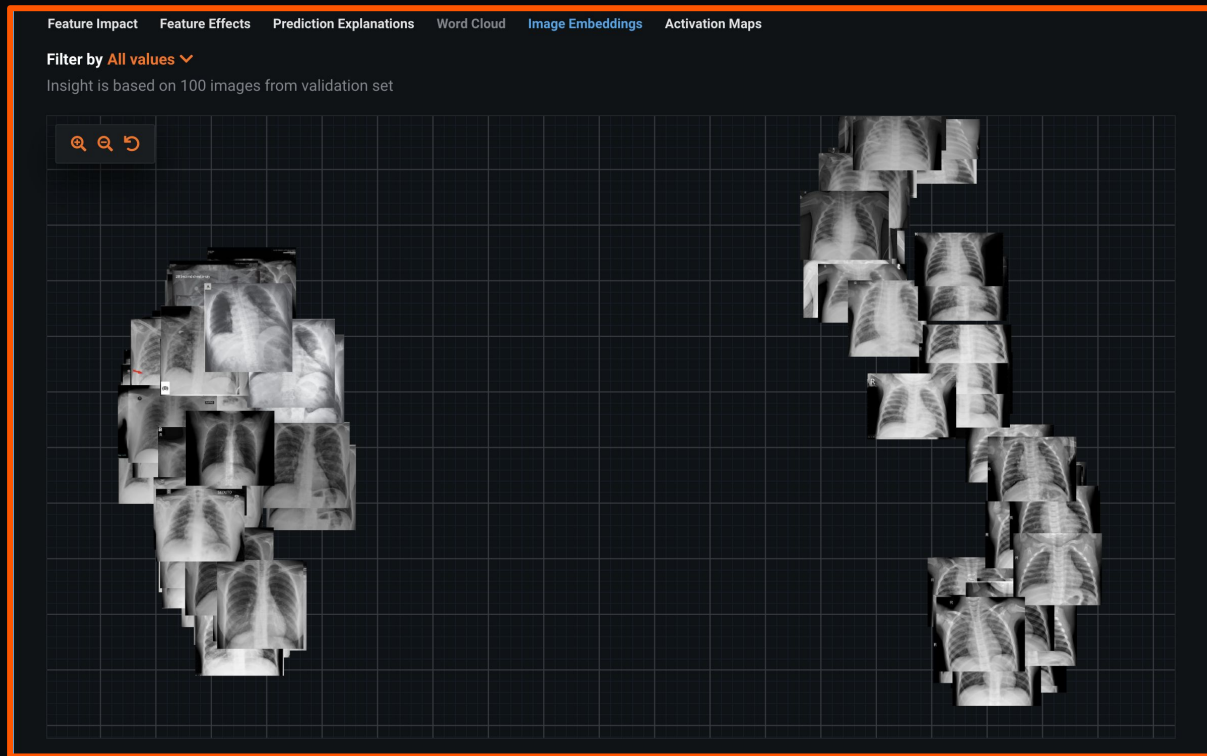
Chest x-ray showing pneumonia

Imagine a patient walks into a emergency room with severe pneumonia. You (nurse, doctor) worry it is Covid-19. You take a nasal swab and send it off for testing. The results will come back in about 12–24 hours based on where your hospital is located. During that long wait, out of caution, the patient is quarantined, using valuable space, and every interaction necessitates...

# COVID Embeddings



Organizes images by how similar they are to each other





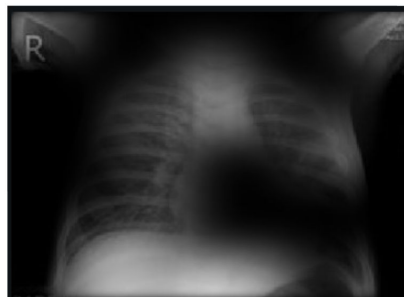
# Explaining the Model

The computer recognizes  
they come from two different  
datasets:

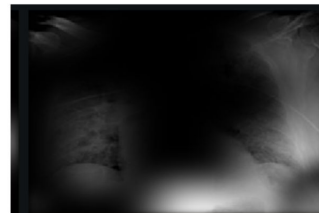
written labels

body structures

COVID-19 negative



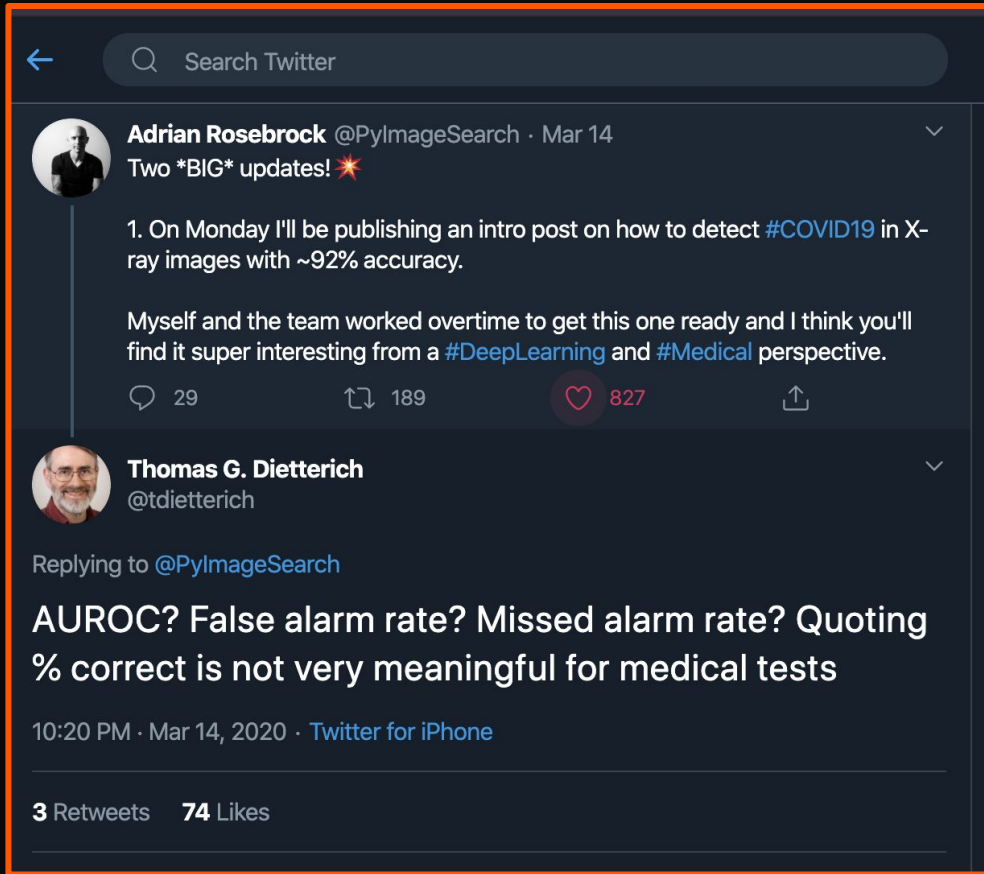
COVID-19 positive



# Expertise

Anyone can find patterns in data

Subject matter experts are the ones that can link it to existing bodies of knowledge





**because you can  
doesn't mean you should**



# Discussion and Q/A



Rajiv Shah

**@rajcs4**

[linkedin.com/in/rcshah/](https://www.linkedin.com/in/rcshah/)