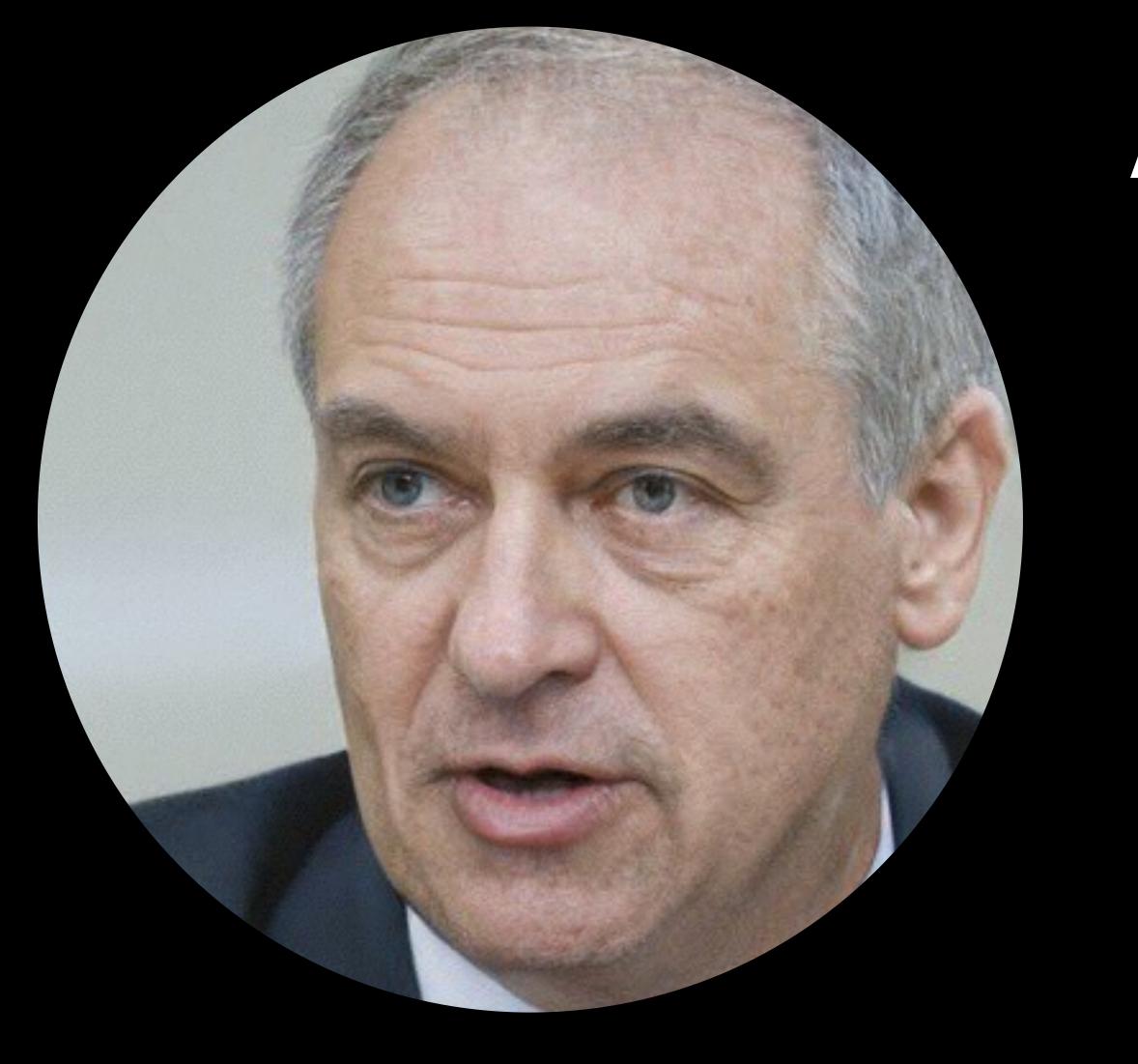
# Adversarial Examples for Robust Detection of Synthetic Media

# or, why you shouldn't trust machine learning

# or, why you shouldn't trust machine learning .... ever





Congressional Candidate



Congressional Candidate

Verified by Twitter



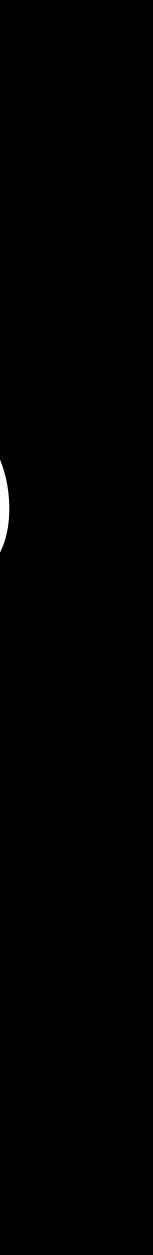
- Congressional Candidate
- Verified by Twitter
- Not a real person





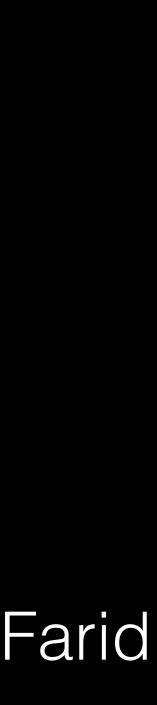






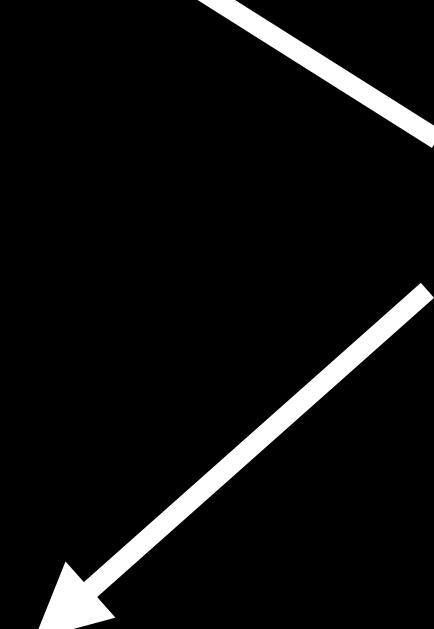
# Four trivial attacks

Joint work with Hany Farid



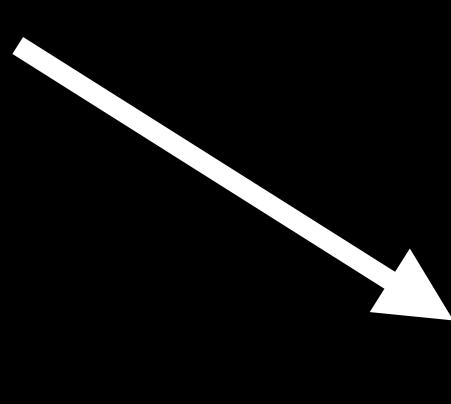












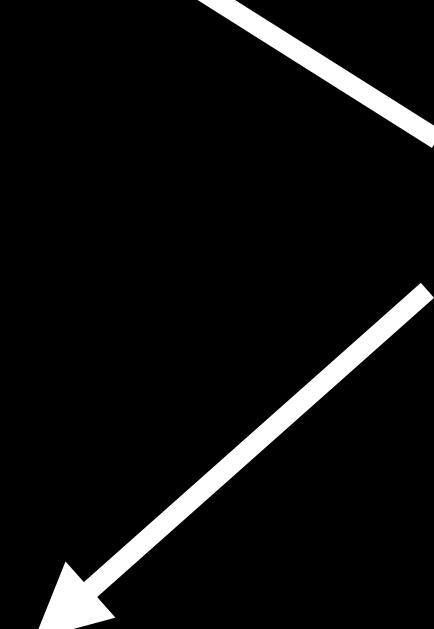






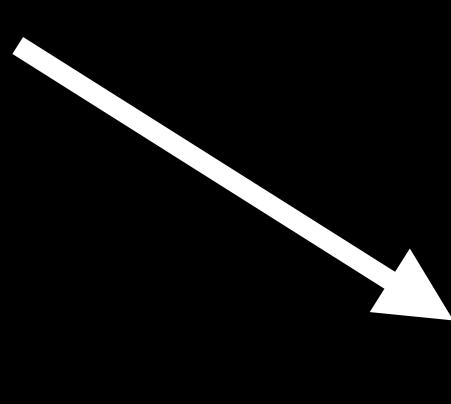










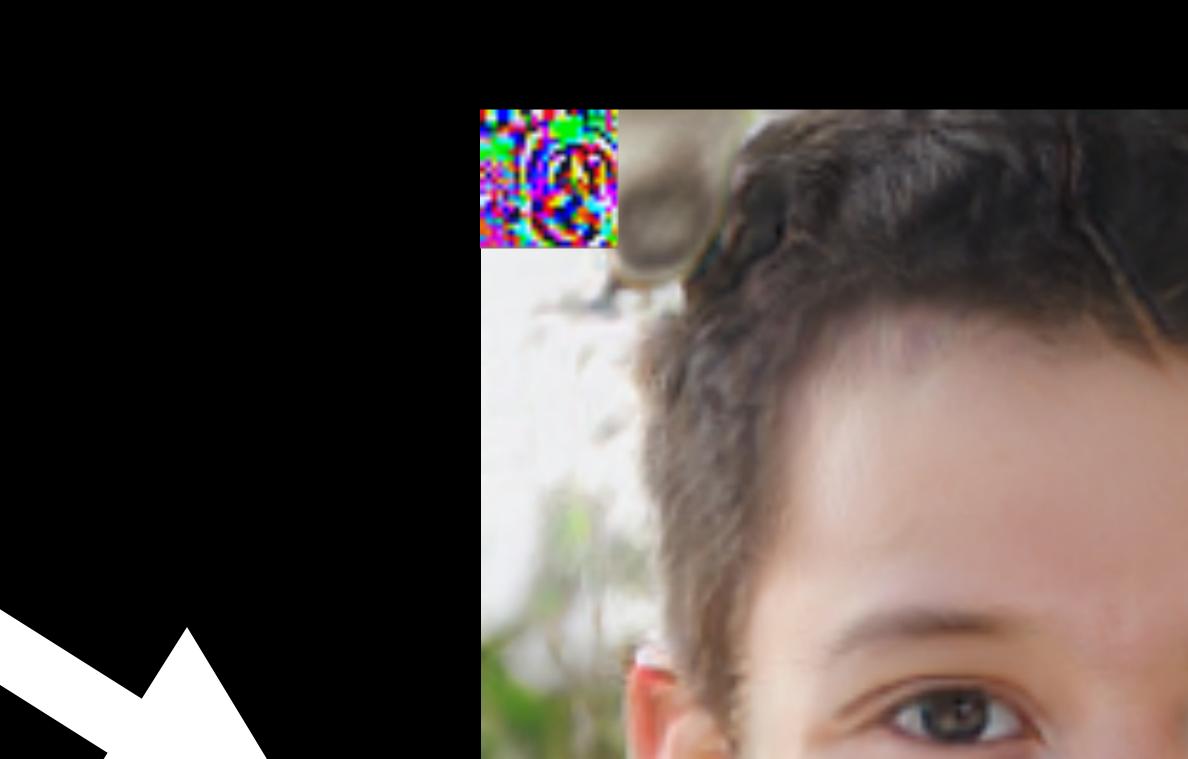






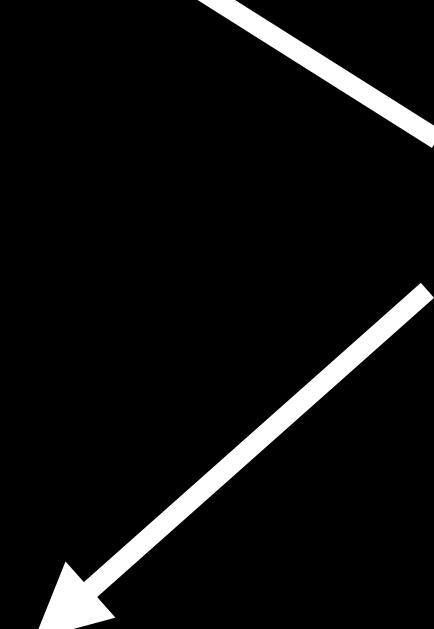






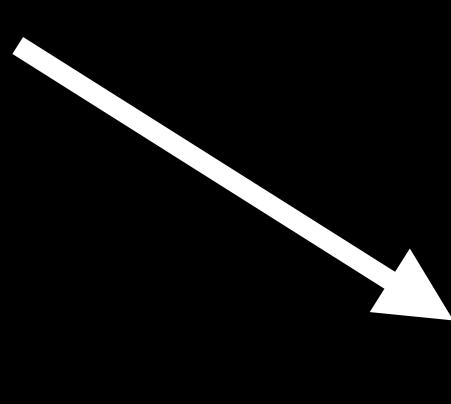




































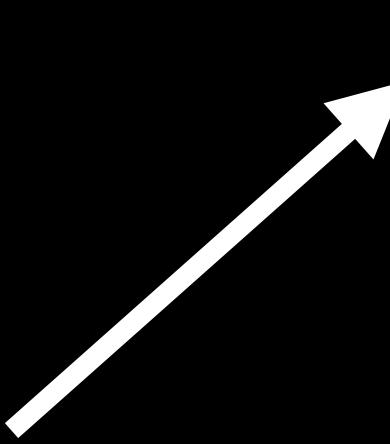


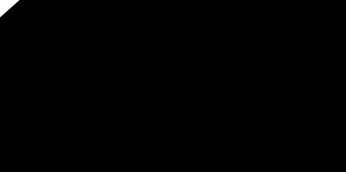


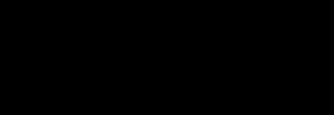


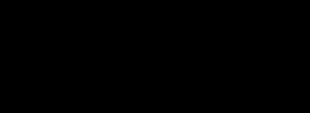


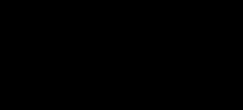


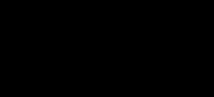






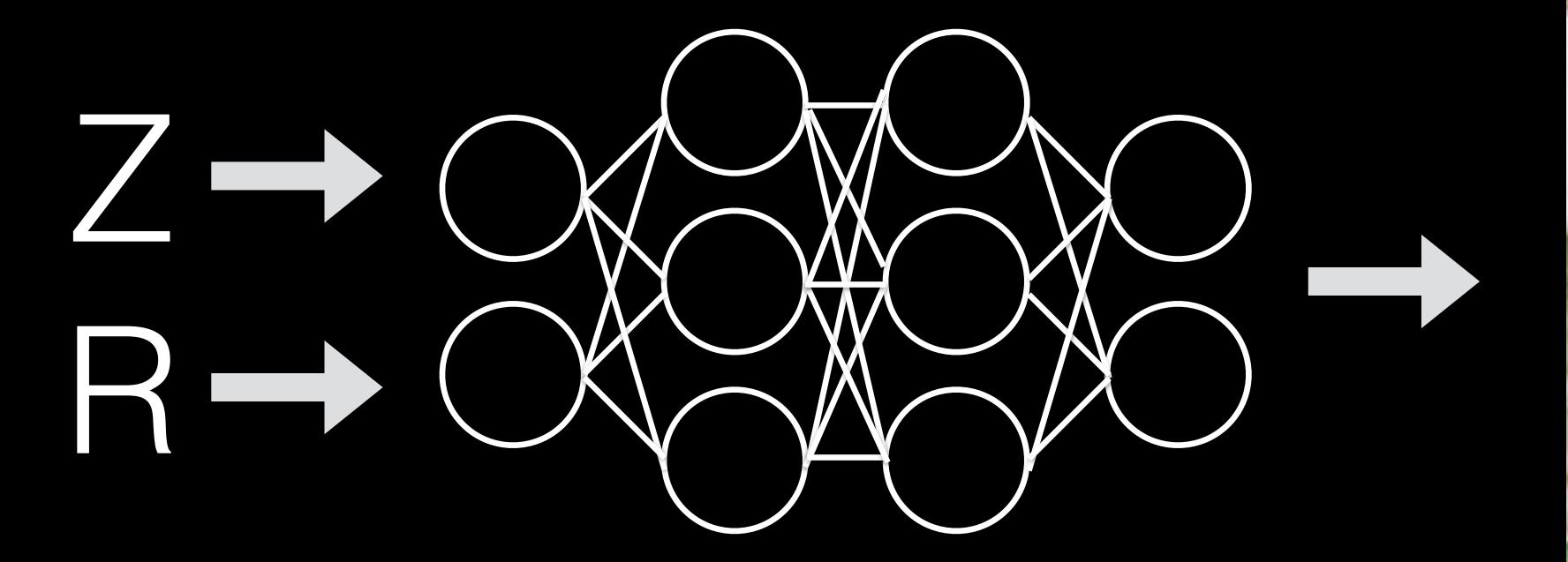
















# G(Z, R)



## 

# G(Z, Radv)



## REAL

# $G(Z_1, R_1)$ $G(Z_2, R_2)$ $G(Z_3, R_3)$











# $G(Z_1, R_{adv}) G(Z_2, R_{adv}) G(Z_3, R_{adv})$















# $G(Z_1, R_1)$ $G(Z_2, R_2)$ $G(Z_3, R_3)$











# $G(Z_1, R_1)$ $G(Z_2, R_2)$ $G(Z_3, R_3)$











# Adversarial Distribution Shifts

# or, why you shouldn't trust machine learning .... ever

# Natural

# Distribution Shifts

Joint work with Rohan Taori, Achal Dave, Vaishaal Shankar, Benjamin Recht, Ludwig Schmidt



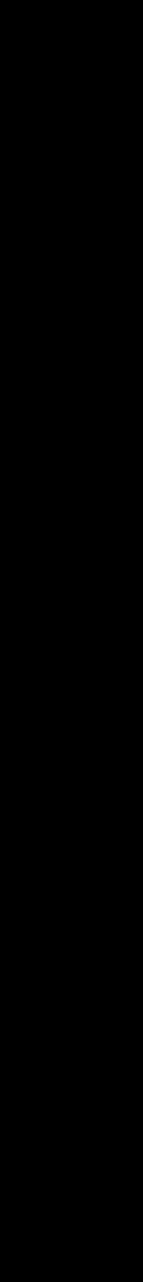
## What we want

1. Someone wants to know what breed of dog they just saw on the street 2. They take out their phone 3. Open up the camera app

- 4. Take a picture, and run a ResNet on the image

## What we have

1. Someone wants to know what breed of dog they just saw on the street 2. They take out their phone 3. Open up the camera app 4. Close the camera app. Open up the browser. Visit <u>http://image-net.org/</u>. Download the ILSVRC2012 test set. Select an image of a dog uniformly at random. Ask the resnet model to classify that random image. Ignore the real dog.



### Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht<sup>\*</sup> Rebecca Roelofs Ludwig Schmidt UC Berkeley UC Berkeley UC Berkeley

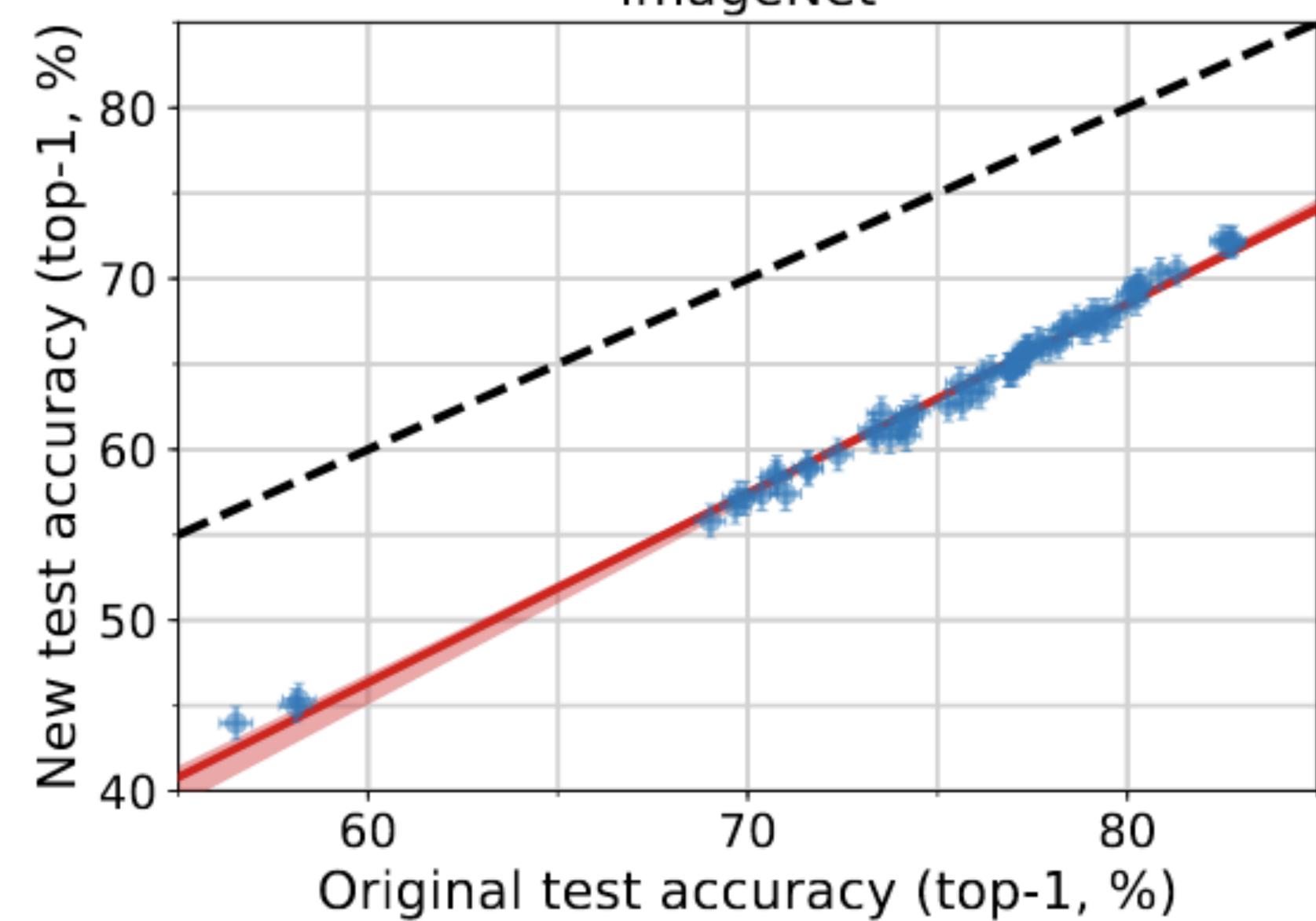
### Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% - 15% on CIFAR-10 and 11% - 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Vaishaal Shankar UC Berkeley

# Now we have a new dataset. Identical in every way to the original. How do models do on this new dataset?





### ImageNet

# Possible explanations

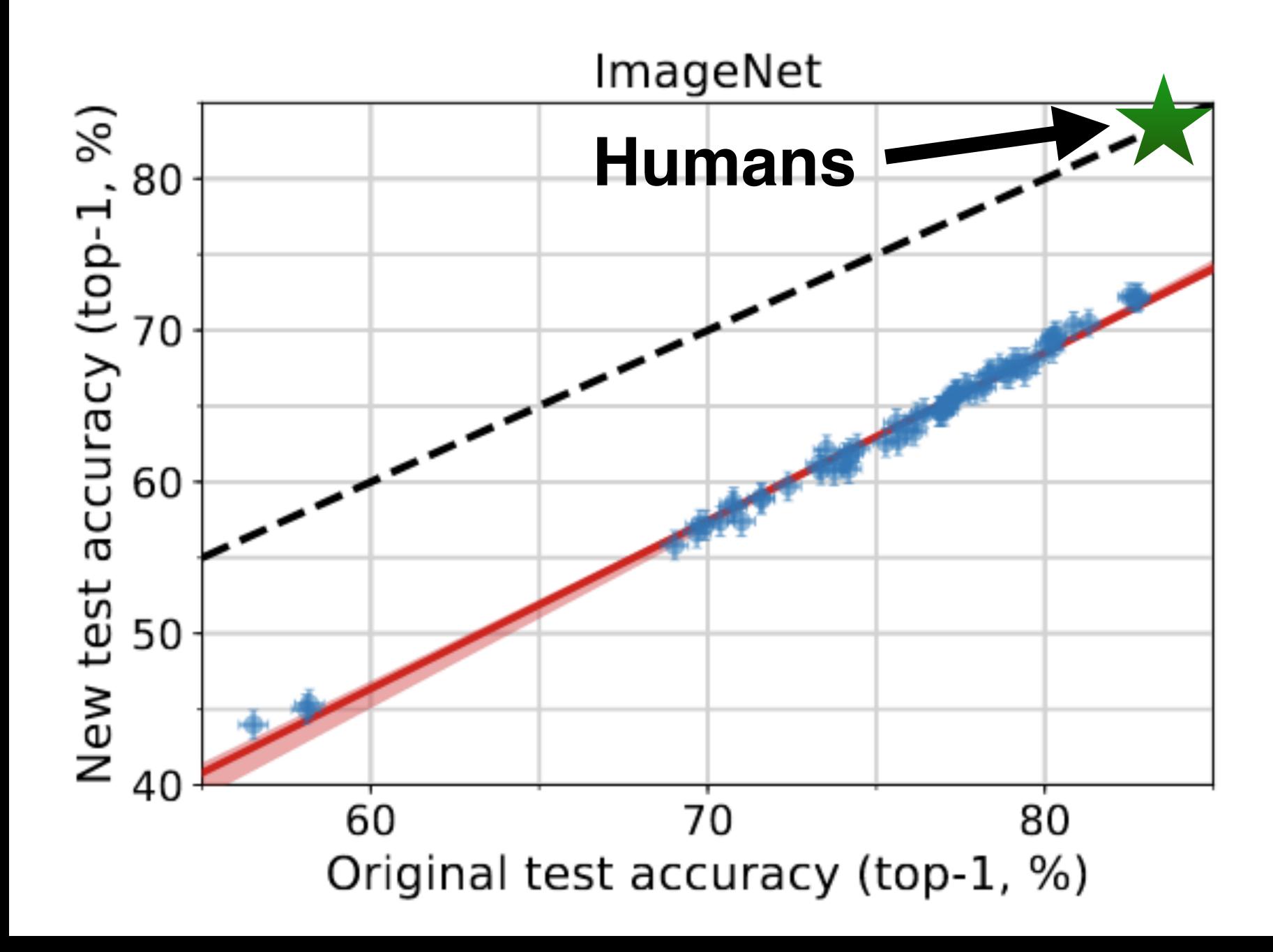
# 1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift



# Possible explanations

# 1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift





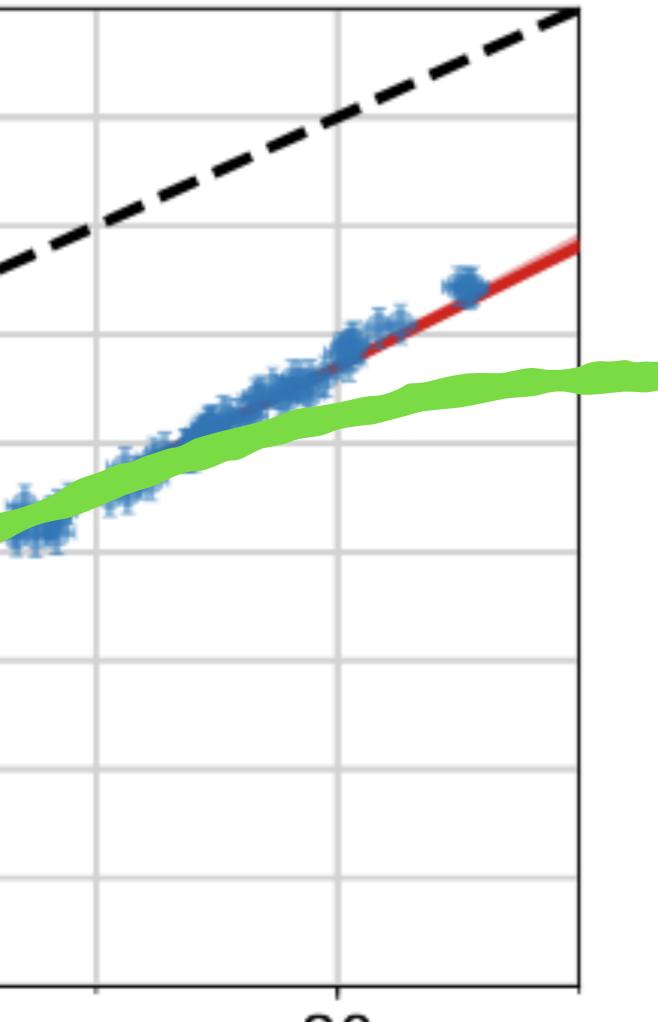
# Possible explanations

# 1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift



# ImageNet

New test accuracy (top-1, %) 40 60 70 Original test accuracy (top-1, %)



## 80

## Adaptive Overfitting



# Possible explanations

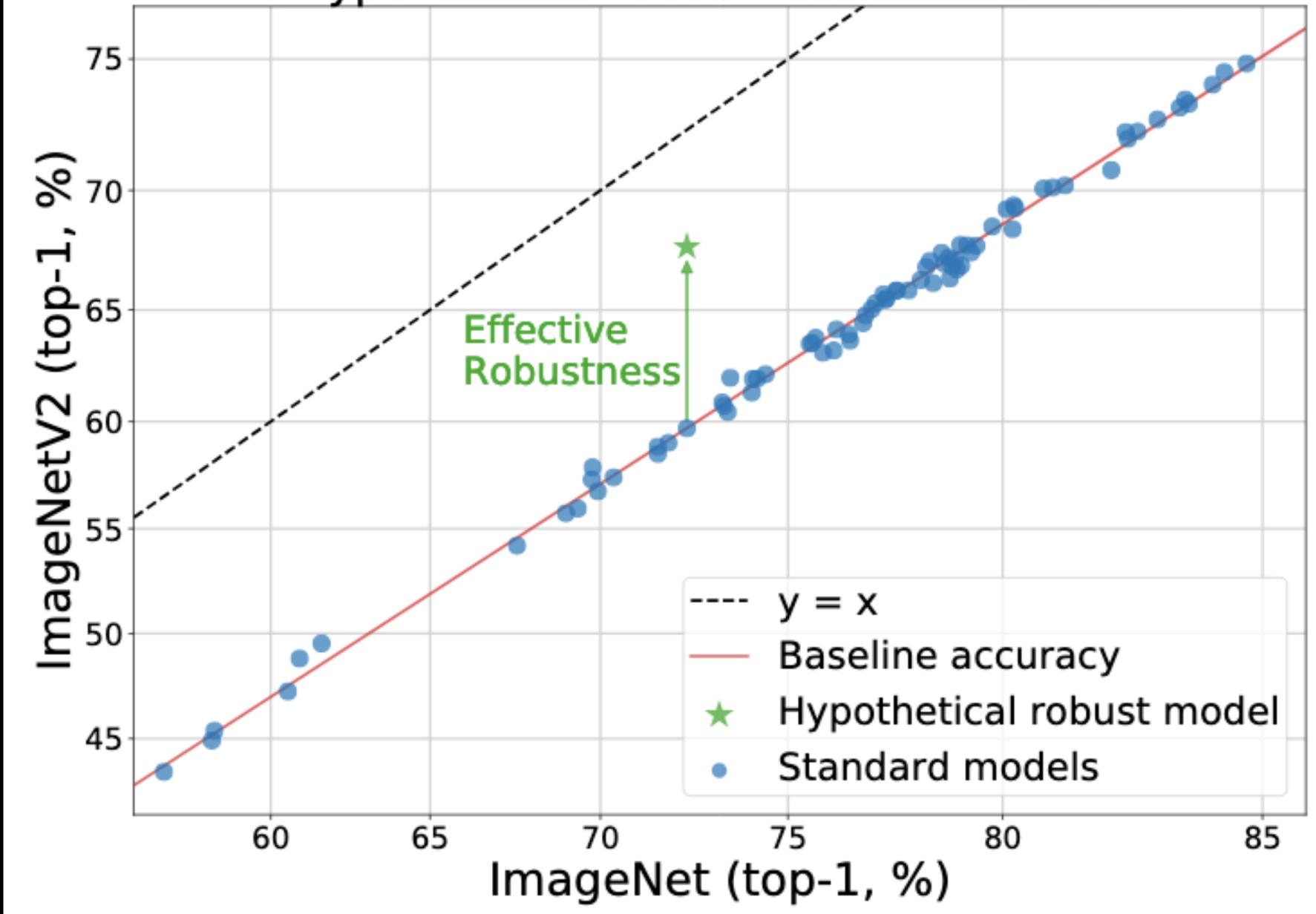
## 1. It's just a harder dataset 2. Adaptive overfitting 3. Distribution shift

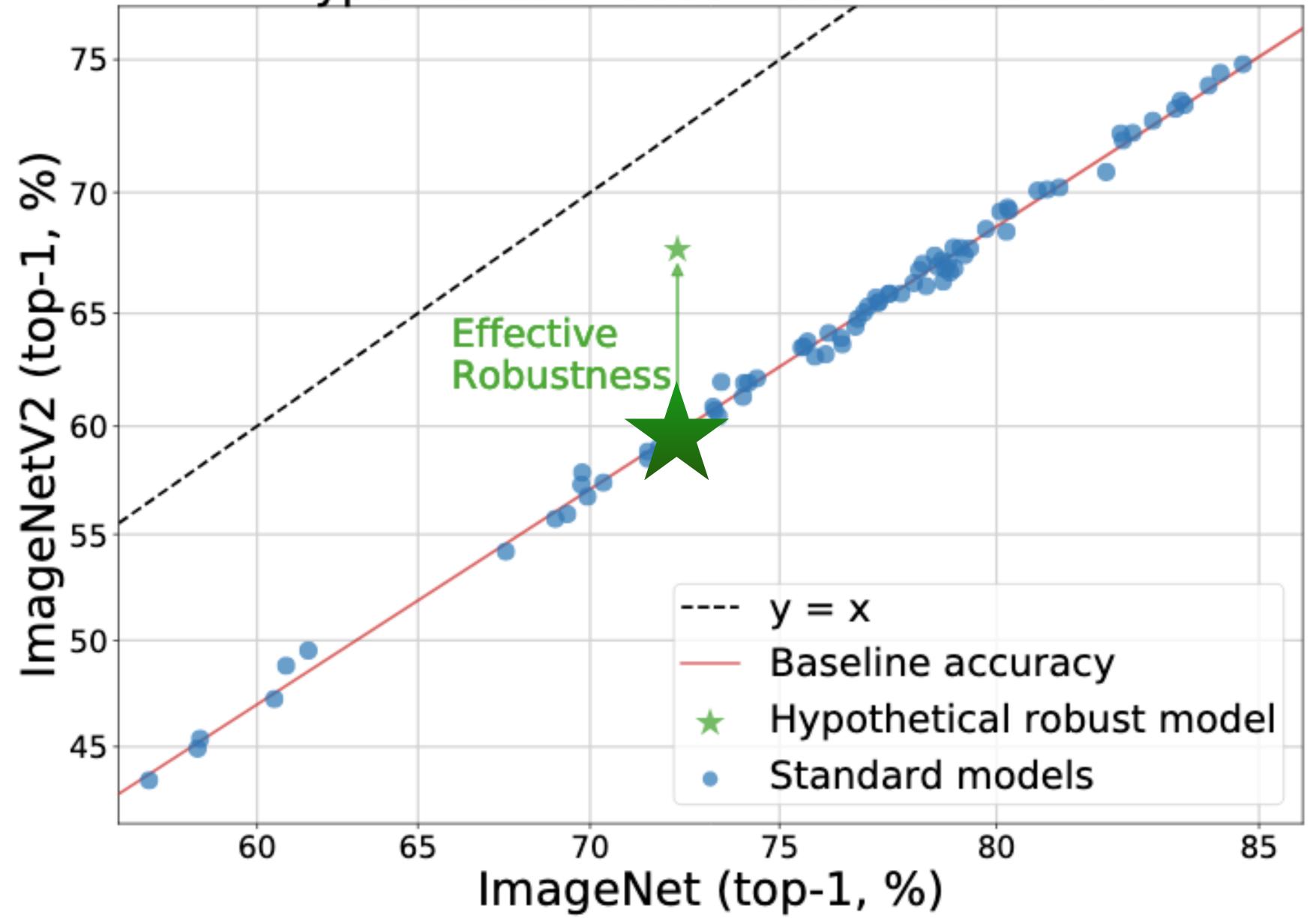


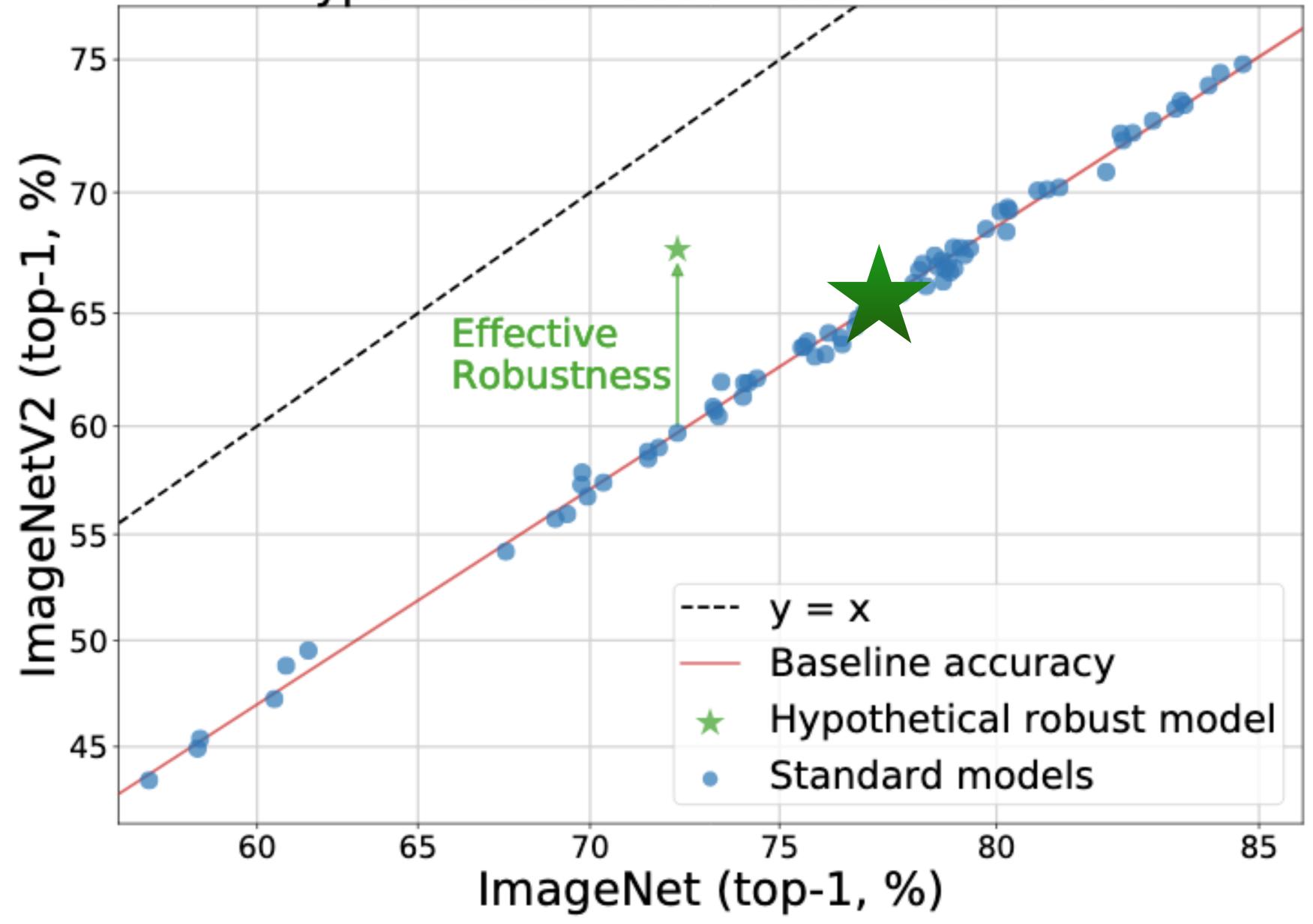
																																																			22 40 26 28 60 6
																																																			23 40 26 29 59 6
44 2.3		11	7.4	49	9.5	12	59	64	38	26	36	50		30	33	20	9.7	40	31	11		57	18	27	42	24	36	58	2	8 23	3	<b>15</b> :	36	27	28	20	19	16	40	27	32	52	2 18	8 .	z	3	3 40	3 40 25	3 40 25 29	3 40 25 29 59	3 40 25 29 59 6
																																																			43 31 33 59 6
42 1.8	e	3.9	7.1	46	8.1	17	59	63	37	29	41	44		27	35	21	15	49	35	27	-	57	19	25	46	25	35	58	2	18 28	4	ю	33	26	32	23	21	18	48	30	36	53	3 2	1.	23		43	43 28	43 28 29	43 28 29 59	43 28 29 59 6
46 2.3		10	8.2	41	9.4	19	58	63	59	36	38	74	1	40			20	47	35	37		53	23	32	44	28	36	58	4	14 30	3	8	53	37	33	22	23	15	44	31	39	48	3 21	0	30		42	42 26	42 26 30	42 26 30 59	42 26 30 59 6
44 1.9		10	7	54	9	17	59	65	42	32	41	46	3	31	38	28	2	48	35	4		23	26		47	34		D	з	12			37	30	33		22	20	46	31	43	54	1 2	4	25		43	43 32	43 32 31	43 32 31 59	43 32 31 59 6
43 2.6		12	7.1	52	9.6	12	58	63	37	26	36	47		27	32	20	0	41	30	15	Ę	96	17		41			58	2	27 23	3			26	27	20	18	15	40	26	31	51	11	8	23		40	40 24	40 24 28	40 24 28 59	40 24 28 59 6
43 2.1	7	7.7	6.2	40	6.9	11	60	64	39	28	36	47		30			4.3	475	53	2	ę	57	16			22		58	2	18		15 :	34		28	20	15	14	41	28	30	51	11	8 .	22		40	40 25	40 25 29	40 25 29 59	40 25 29 59 6
44 1.6	S	9.7	7.2	53	9.1	17	59	64	43	33	40	47		32	39	27	22	47	36	4	ŧ	58	25	29	46	33	38	60	3	13 30	3	9 3	37	30	34	26	22	21	46	32	44	54	4 2	4	26		42	42 32	42 32 31	42 32 31 59	43 28 29 59 6   42 26 30 59 6   43 32 31 59 6   40 24 28 59 6   40 24 28 59 6   40 25 29 59 6   42 32 31 59 6
																																																			43 32 33 59 6
47 1.3		13	8.9	54	12	16	56	62	37	32	41	43	1	32	39	28	26	51	35	4		56	27	23	43	33	36	58	2	19 3(	4	ю	33	28	33	23	24	17	47	31	41	51	2	1.	23		41	41 28	41 28 30	41 28 30 58	41 28 30 58 6
43 1.7	e	3.9						62	з		36	45						43	32	з	) (	66	16	Г					2	18 25	3	и :	34			18	19						17	7	22				30	30 57	30 57 6
52 2	1	70		70	70	20		61	2		51	7			43	37	3.		38	6	: 6	57	37			46	42			30	4	9	11			34	38	31	62		60	53	3 3	3.	29					58	58 6
41 1.8		11		50	9.4	10		60	3		35			5	30	17	7.5	36	28	18		54	14			20	34	56		21	3	13					16	12	36		28	48	3 19	5	18			21	21	2) 57	21 57 6
45 1.7		14						65	5		41			35	35	t.			33	5		33	33			37	32	61		25	3	9		33	25		22	22	47		51	58	3 21	8				33	33	33 58	33 58 6
41 1.1		10		52	9.1	17		62	3		41			9	38	25	18		34	4	ę	6	23			30	36	Ę		28	4	17					23	18	46		42	53	3 2	1							
37 2.3	9	9.6		38	9.7	17		60	3		44				35	14	,		33	3	) =	54	13			19	34			26				26	30			12	39		36	50	0 1				38	38 23	38 23 2	38 23 24	38 23 24 6
41 1.6		14					-5	59	3		37	39							28	21		i2	15						2	13 21			29	23	25	1)		11	38		29	47					36	36 22	36 22 27	36 22 27	36 22 27
49 2.8		11	8.1	46	10	14	50	58	38	45	51	41		32	48	23	21	53	38	7:	2	33	23	21	41	30	40	62	3	14 41	5	51 3	38	34	44	24	32	19	50	35	68	55	5 Z	3.	27		46	46 31	46 31 38	46 31 38 57	46 31 38 57 6
46		69	66	26	4.8	28	57	65	19	30	56	17		49	34	39	27	65	39	54		31	38	51	56	45	42	60	1	0 23	4	9 6	5.8	42	26	38	32	29	59	31	55	53	3 30	6 .	42		48	48 43	48 43 34	48 43 34 55	48 43 34 55 6
48 2.4		68	48	68	68	19	50	59	18	33	50	17		39	40	43	41	65	38	61		i3	42	28	51	51	42	56	1	1 30	4	18 7	7.8	35	33	41	37	39	61	35	59	49	9 40	0	28		46	46 49	46 49 37	46 49 37 55	46 49 37 55 6
37 1.9		10	8.1	42	9	13	52	57	27	22	39	34		21	27	14	9.5	41	30	24		50	13	18	39	20	31	52	2	2 2	3	37 2	27	21	26	15	18	11	40	27	31	46	5 14	4	18		37	37 20	37 20 25	37 20 25 54	37 20 25 54 6
48 25	e	13	78	35	89	16	48	58	37	49	49	42		25	49.	21	17	46	72			8	19	25	40	25	53	60	3	11 37	5		17	30	39	21	29	15	47	63	47	52	> 1	9	29		44	44 25	44 25 46	44 25 46 56	44 25 46 56 6
42		or 05	63	26	3.2	97	54	62	19	29	53	18		48	21	30	22	62	36											.5 20										29	53	E.			40			48 41	48 41 21	48 41 21 53	48 41 21 53 8
7° 47 20				20	5.2	40		54																						.7 26												46	, <b>,</b>	-	7V 70		-*0	45 E2	45 50 50	45 53 36 54	40 41 51 55 0
4/ 22		04	53	04	04	18	40	94	10	31	50	15	2	31	30	40	45	63	31	5		19	40	29	50	94	41	52	· ·	. 2	4	Hr -	Ð	33	31	45	30	43	60	34		40	> 4:	-	20		40	40 52	45 52 36	45 52 36 51	45 52 36 51 6
2.3	1	11	6.1	47	8.5	15	67	51	30	18	25	35	į	27	26	31	28	46	23	3:		54	29	18	35	33	23	43	2	1	2	24 2	23	21	19	21	11	16	41	18	30	49	9 19	9	18		32	32 23	32 23 15	32 23 15 51	32 23 15 51 6

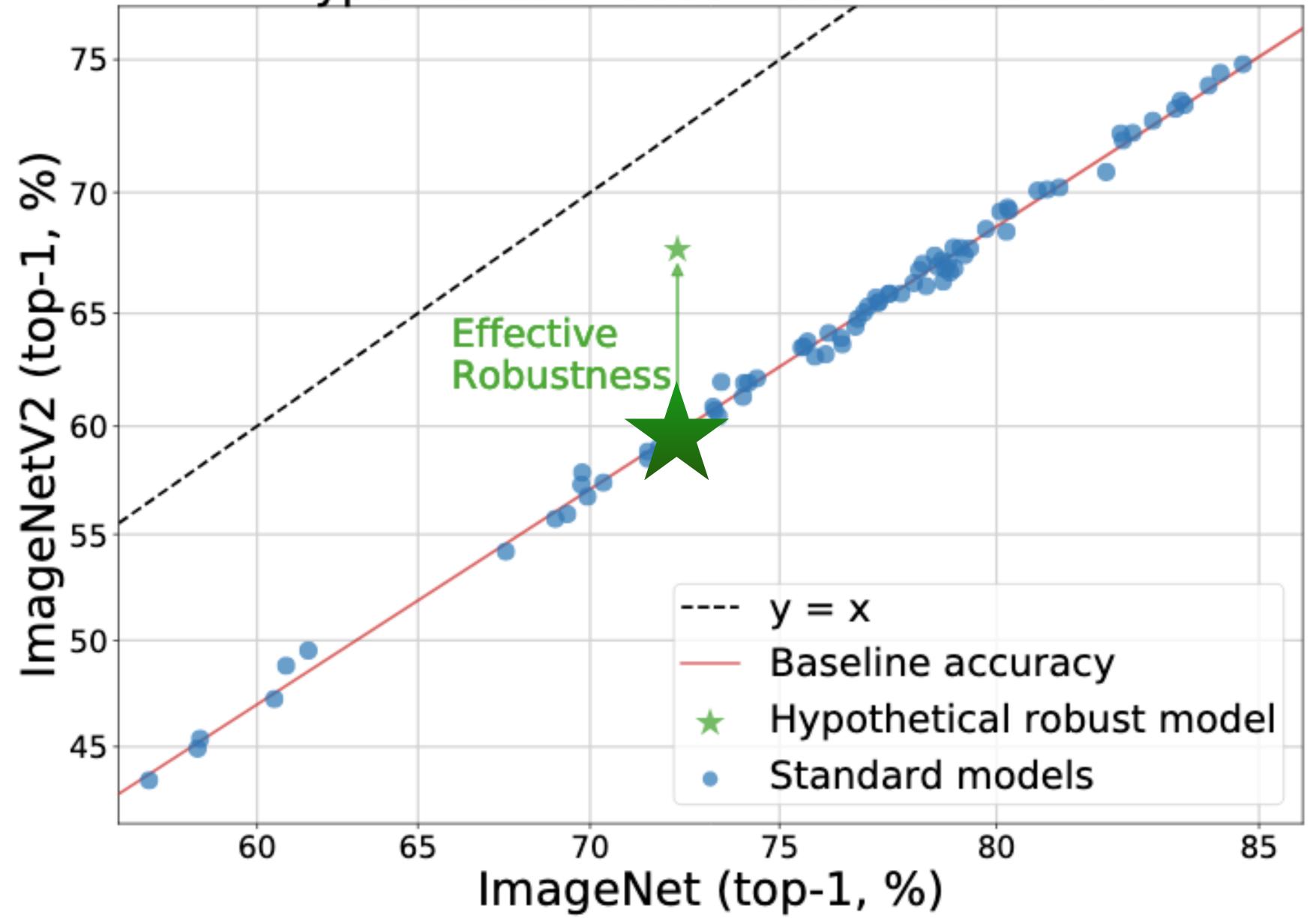
# Formalization:

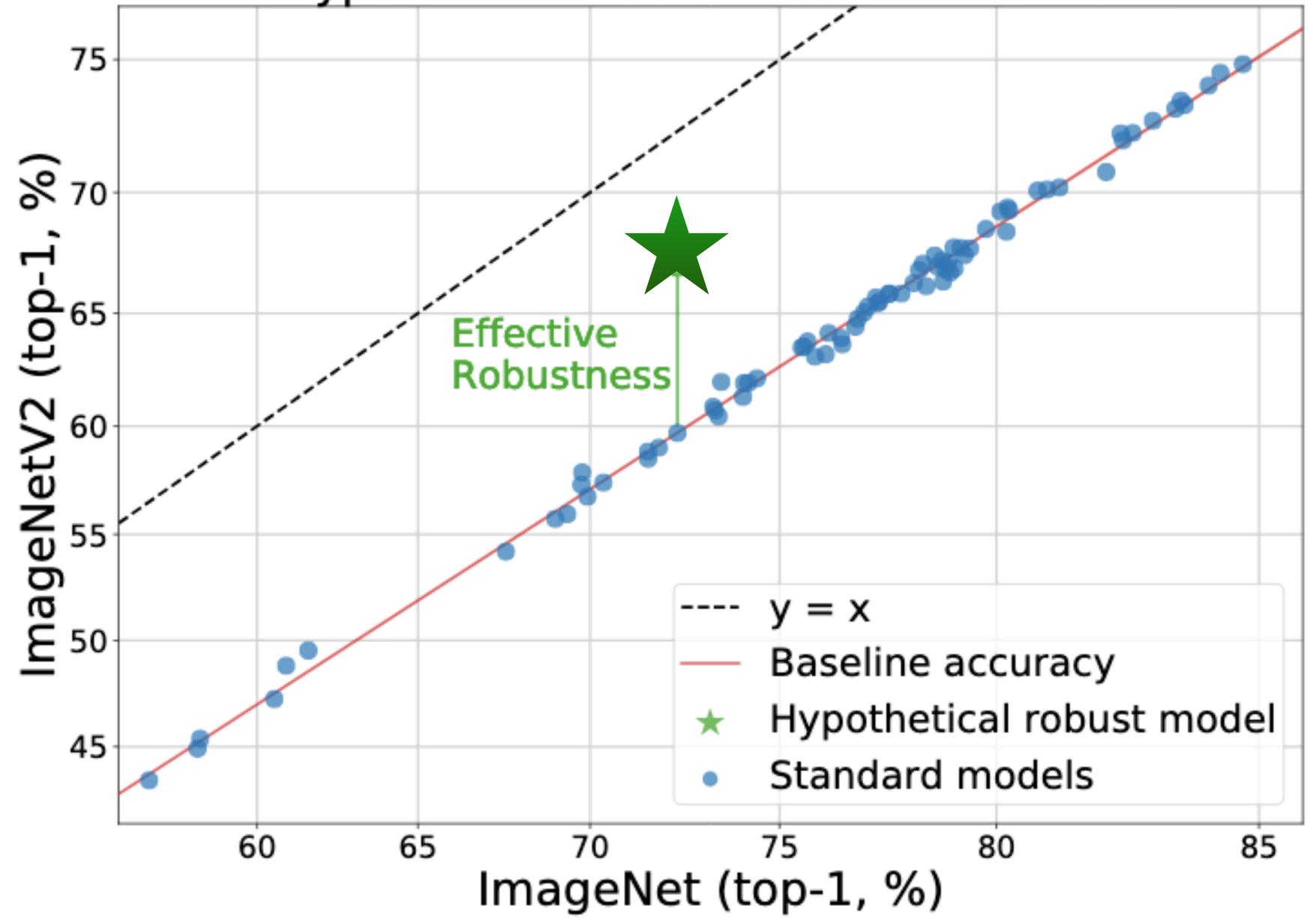
# Effective Robustness



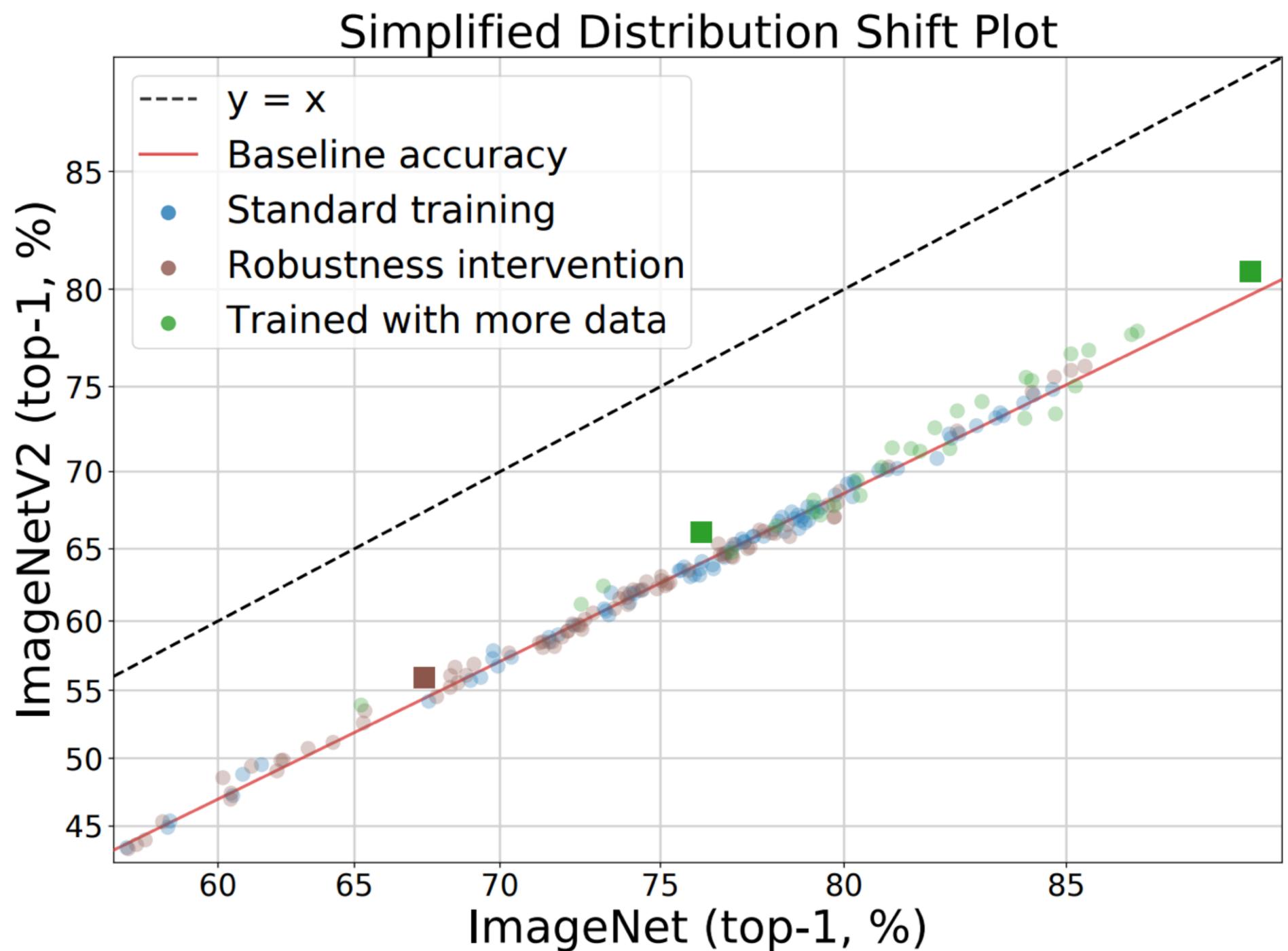








# So what helps?



## 



# If you use machine learning, you're vulnerable

**Don't** go and try to solve adversarially robust forensic detection

Do consider evaluating the robustness of your Classifiers

nicholas@carlini.com

Machine learning is not robust, in neither acversarial nor natural data settings

https://nicholas.carlini.com

