دوره‌ی آموزشی «علم داده»
Data Science Course

جلسه سی و سوم (بخش اول)

پروژه‌ی پیش‌بینی خرید مجدد مشتریان از یک پلت‌فرم فروش کتاب صوتی

(آشنایی با داده‌ها)

مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

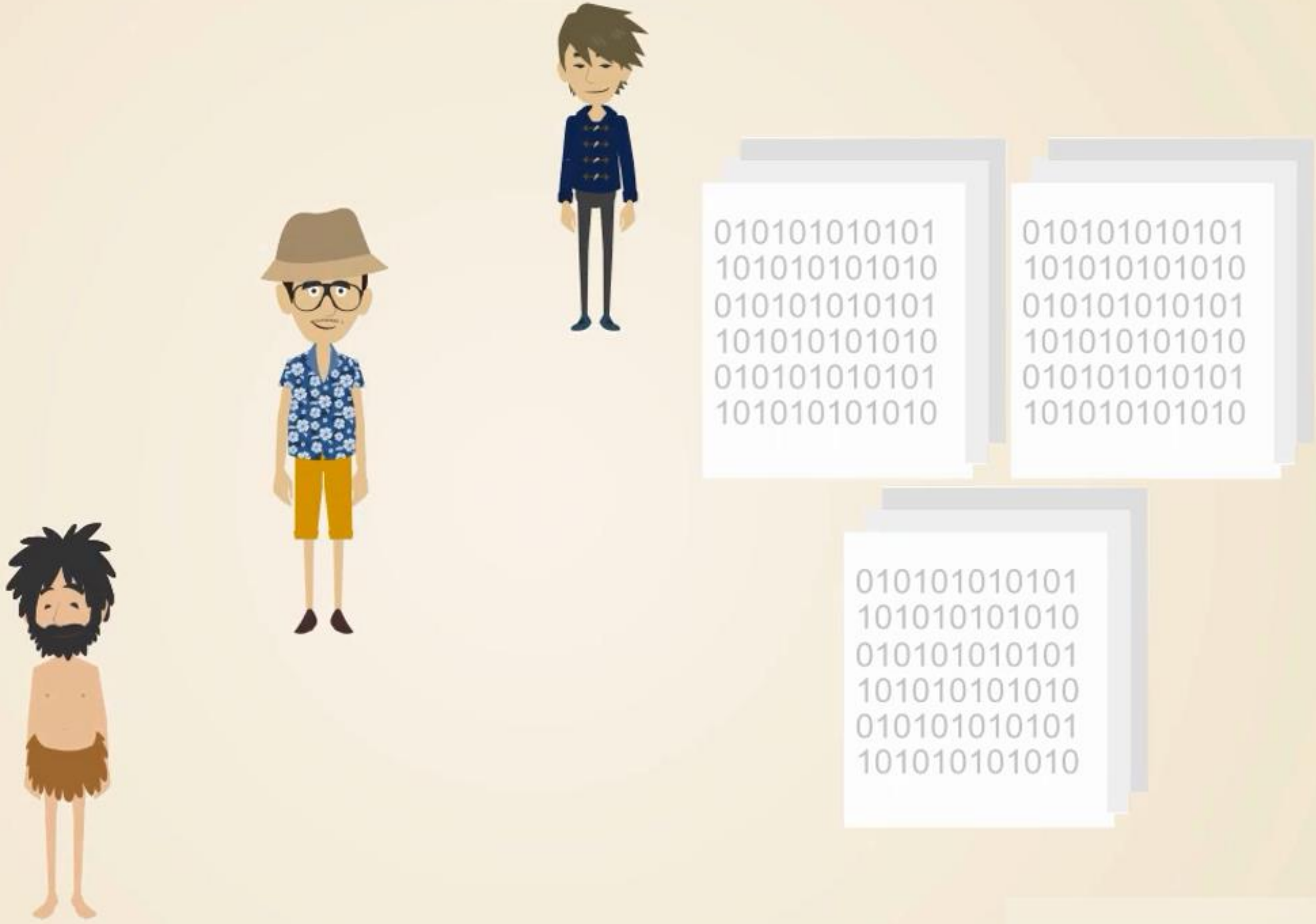
خب! رسیدیم
به آخرای دوره

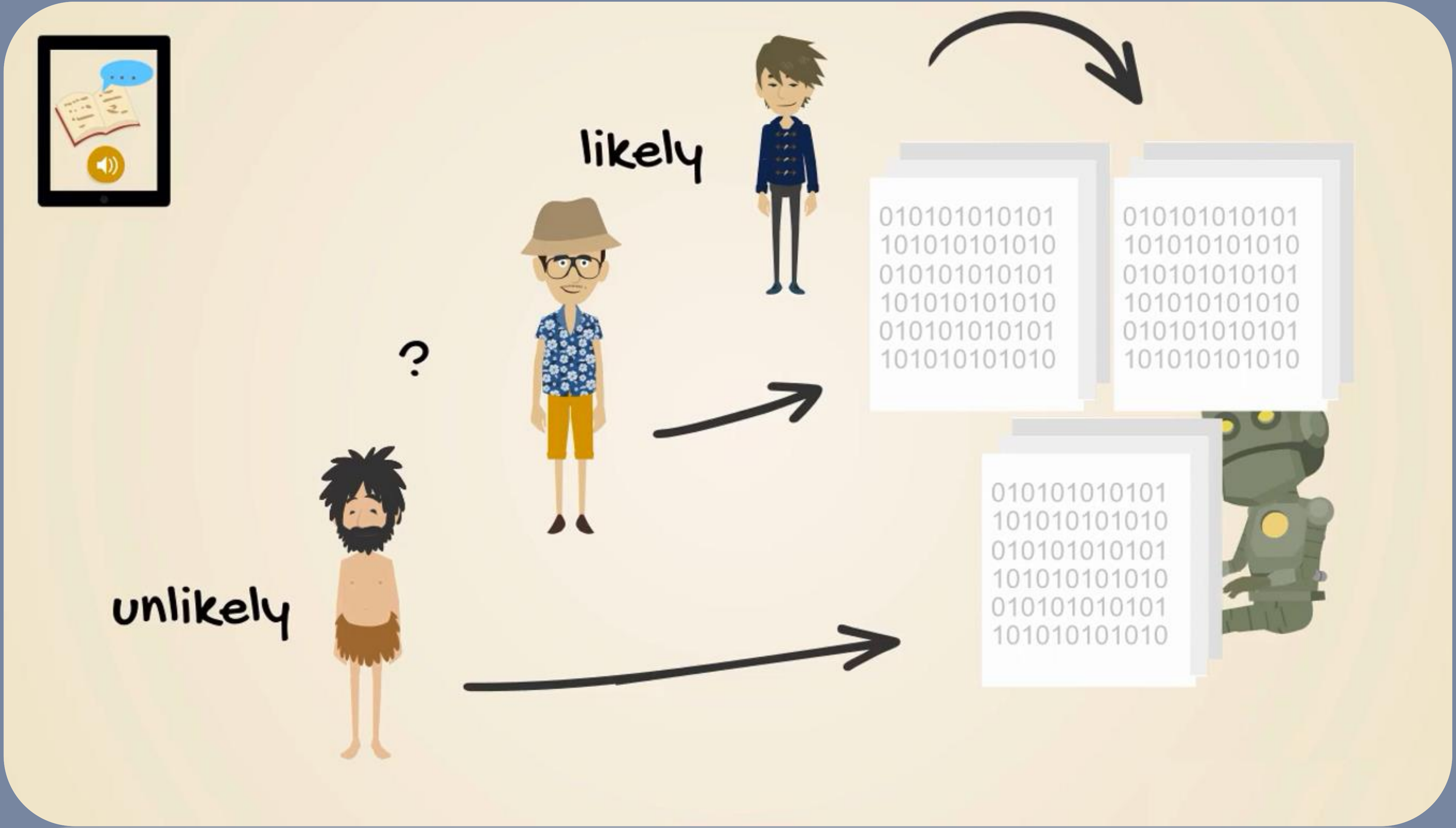
But it is simply the application of everything you already know

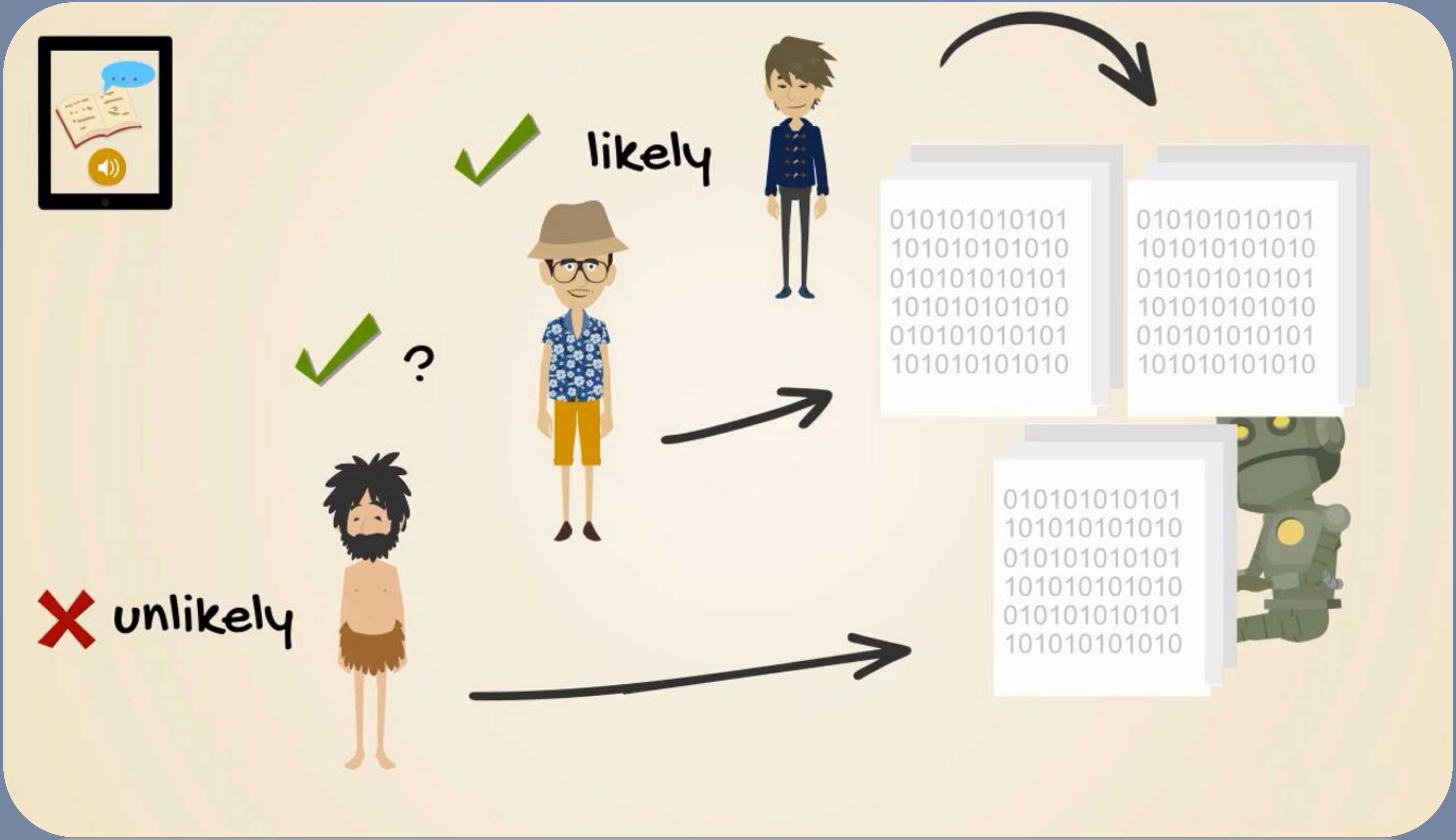
اما پروژه‌ی پیش‌رو، چیزی
نیست، جز بکارگیری مطالبی
که تا الان یاد گرفتیم

مسئله چیه؟

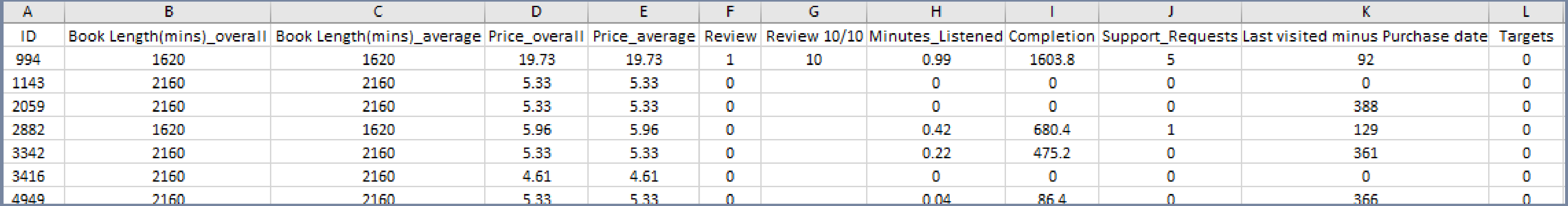
What we should deal with?


010101010101
101010101010
010101010101
101010101010
010101010101
101010101010


likely
?
unlikely
010101010101
101010101010
010101010101
101010101010
010101010101
101010101010


✓ likely
✓ ?
✗ unlikely
010101010101
101010101010
010101010101
101010101010
010101010101
101010101010

Important metrics for conversion



A	B	C	D	E	F	G	H	I	J	K	L
ID	Book Length(mins)_overall	Book Length(mins)_average	Price_overall	Price_average	Review	Review 10/10	Minutes_Listened	Completion	Support_Requests	Last visited minus Purchase date	Targets
994	1620	1620	19.73	19.73	1	10	0.99	1603.8	5	92	0
1143	2160	2160	5.33	5.33	0		0	0	0	0	0
2059	2160	2160	5.33	5.33	0		0	0	0	388	0
2882	1620	1620	5.96	5.96	0		0.42	680.4	1	129	0
3342	2160	2160	5.33	5.33	0		0.22	475.2	0	361	0
3416	2160	2160	4.61	4.61	0		0	0	0	0	0
4949	2160	2160	5.33	5.33	0		0.04	86.4	0	366	0

1. ID
2. Book length(mins)_overall
3. Book length(mins)_average
4. Price_overall
5. Price_average
6. Review
7. Review 10/10
8. Minutes_listened
9. Completion
10. Support_requests
11. Last visited minus purchase date
12. Targets

ID is like a name

The overall book length is the sum of the lengths of purchases

The average book length is the sum divided by the number of purchases

The # purchases = overall length / average length

The price variable is almost always a good predictor!

Review 10/10

- It measures the review of a customer from 1 to 10
- For our ML algorithm, 8.91 = status quo
a review > 8.91indicates above average "feelings"
a review < 8.91indicates below average "feelings"
- The average review indicates the average feelings (towards content / platform / medium)

Total minutes listened is a measure of engagement

Completion is the total minutes listened / book length_overall

Support requests shows the total number of support requests (forgotten password to assistance)

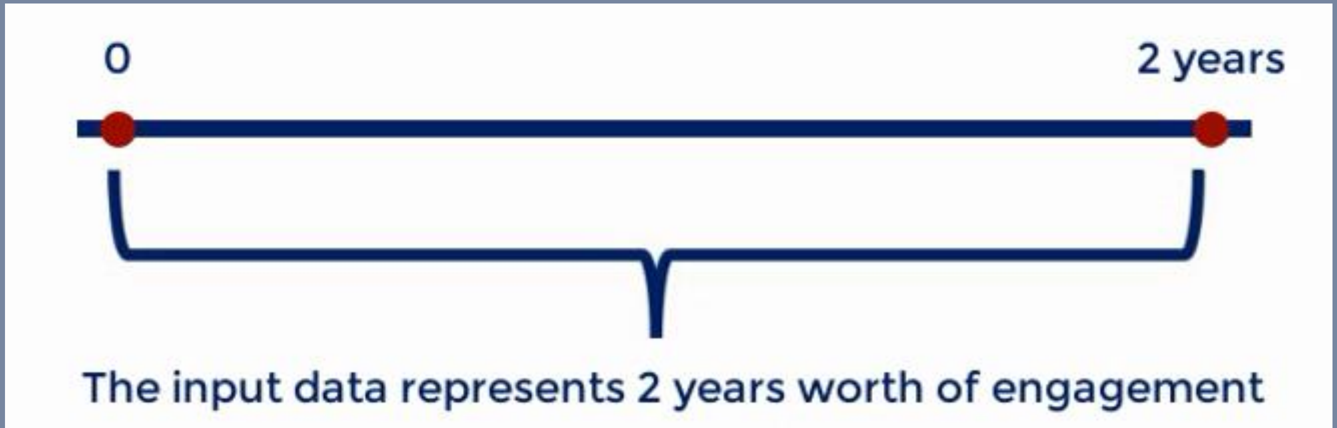
Last visited minus purchase date → The bigger the difference, the bigger the engagement

↓

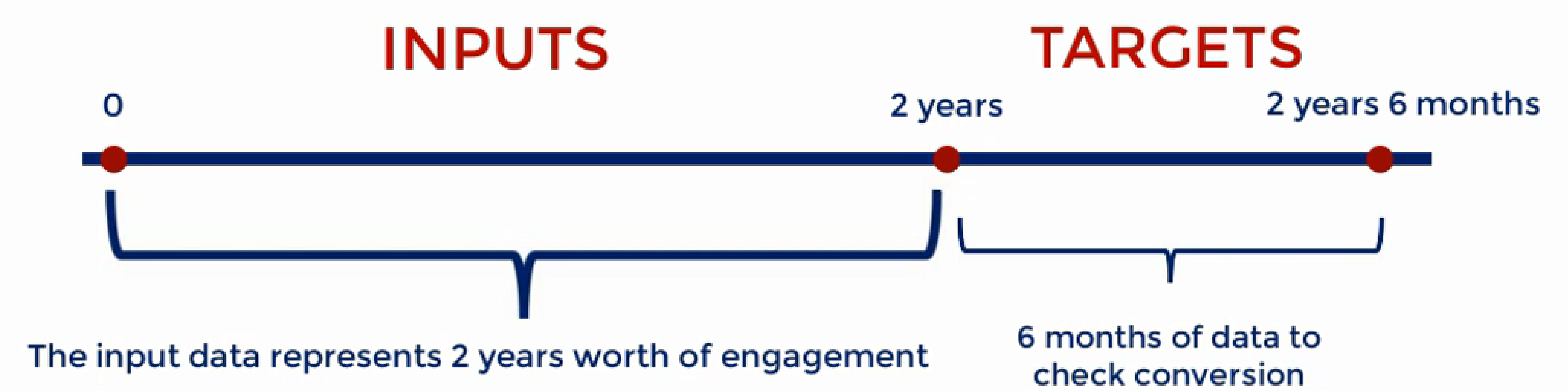
If the value is 0, we are sure the customer has never accessed what he/she has bought

Targets: 1 if a customer bought again in the last 6 months of data,
0 if a customer did not buy again

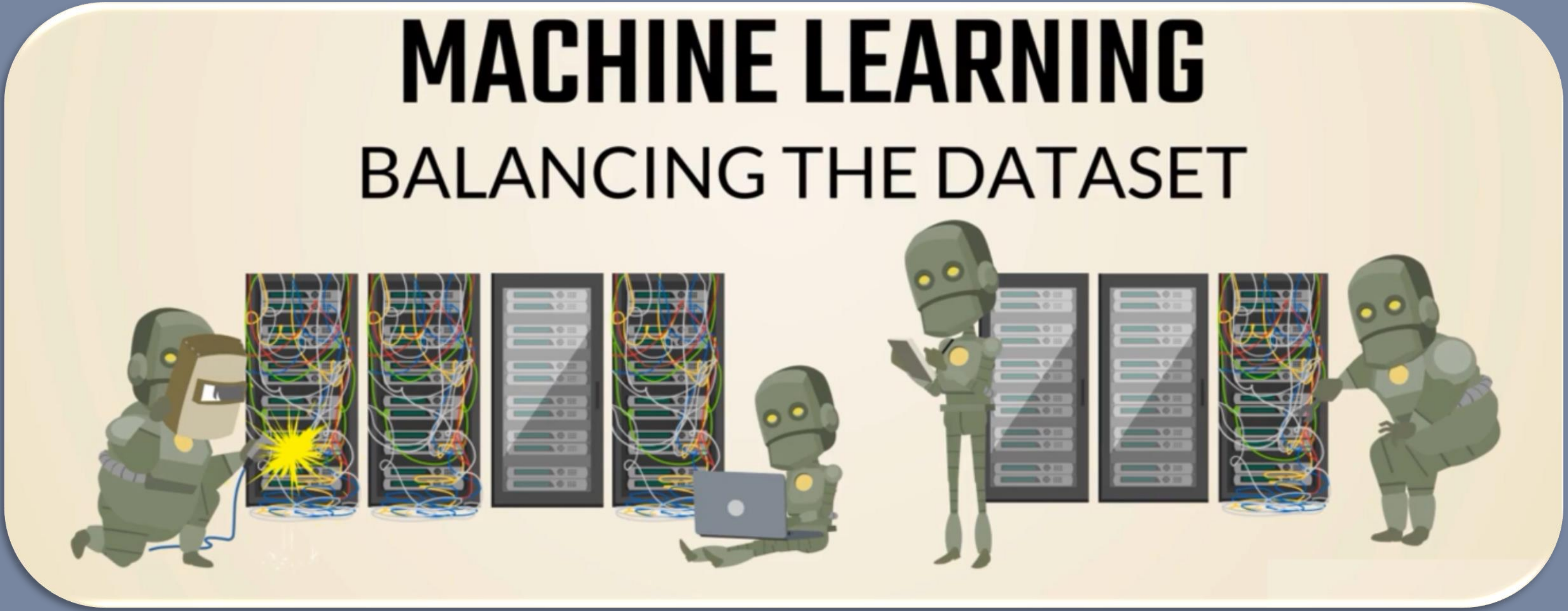
It is always necessary to ask how the data was gathered!

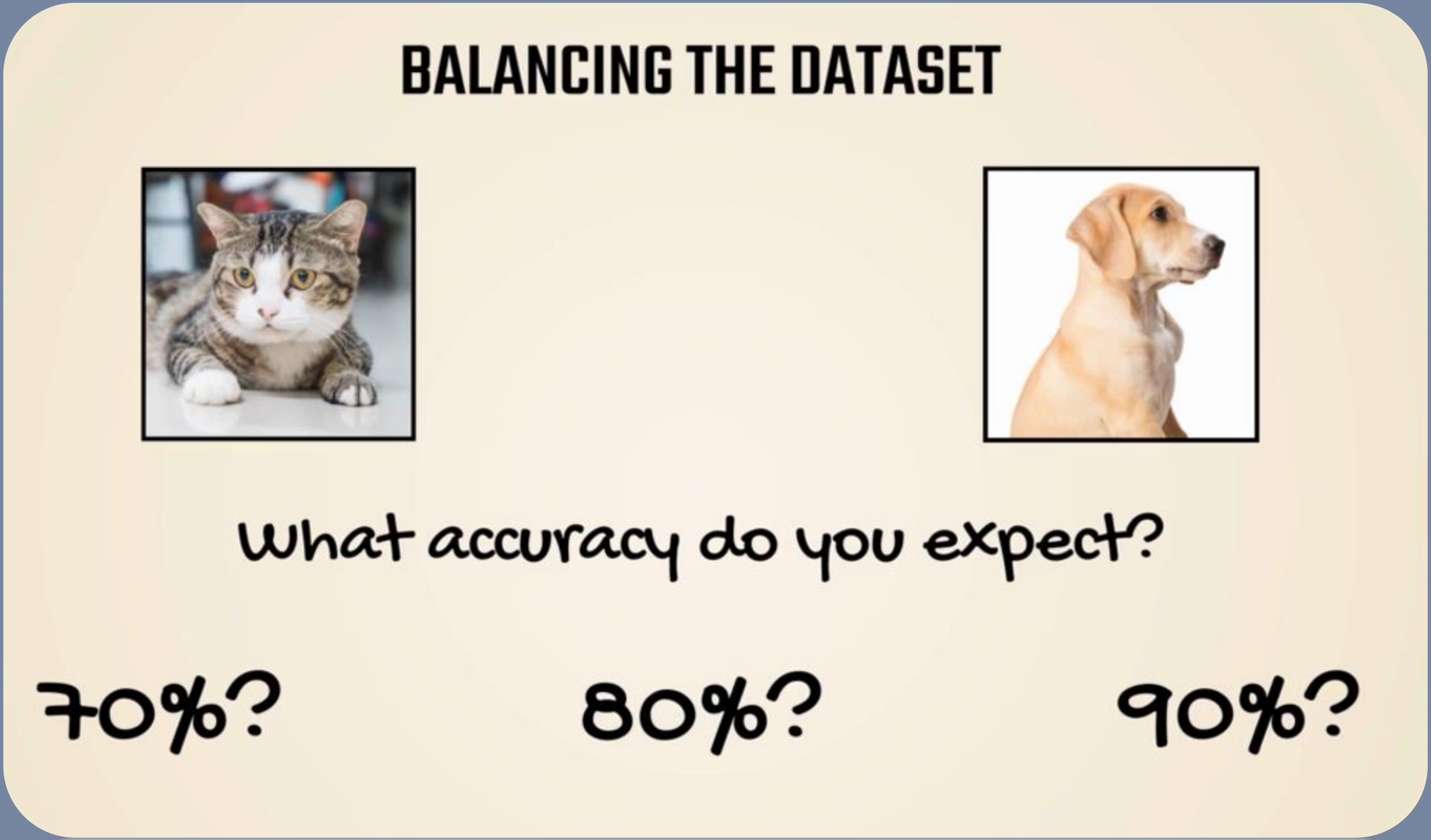


What does it mean to convert?


INPUTS
TARGETS
0
2 years
2 years 6 months
The input data represents 2 years worth of engagement
6 months of data to check conversion

بالانس کردن دیتاست!
یکی از مهمترین مراحل

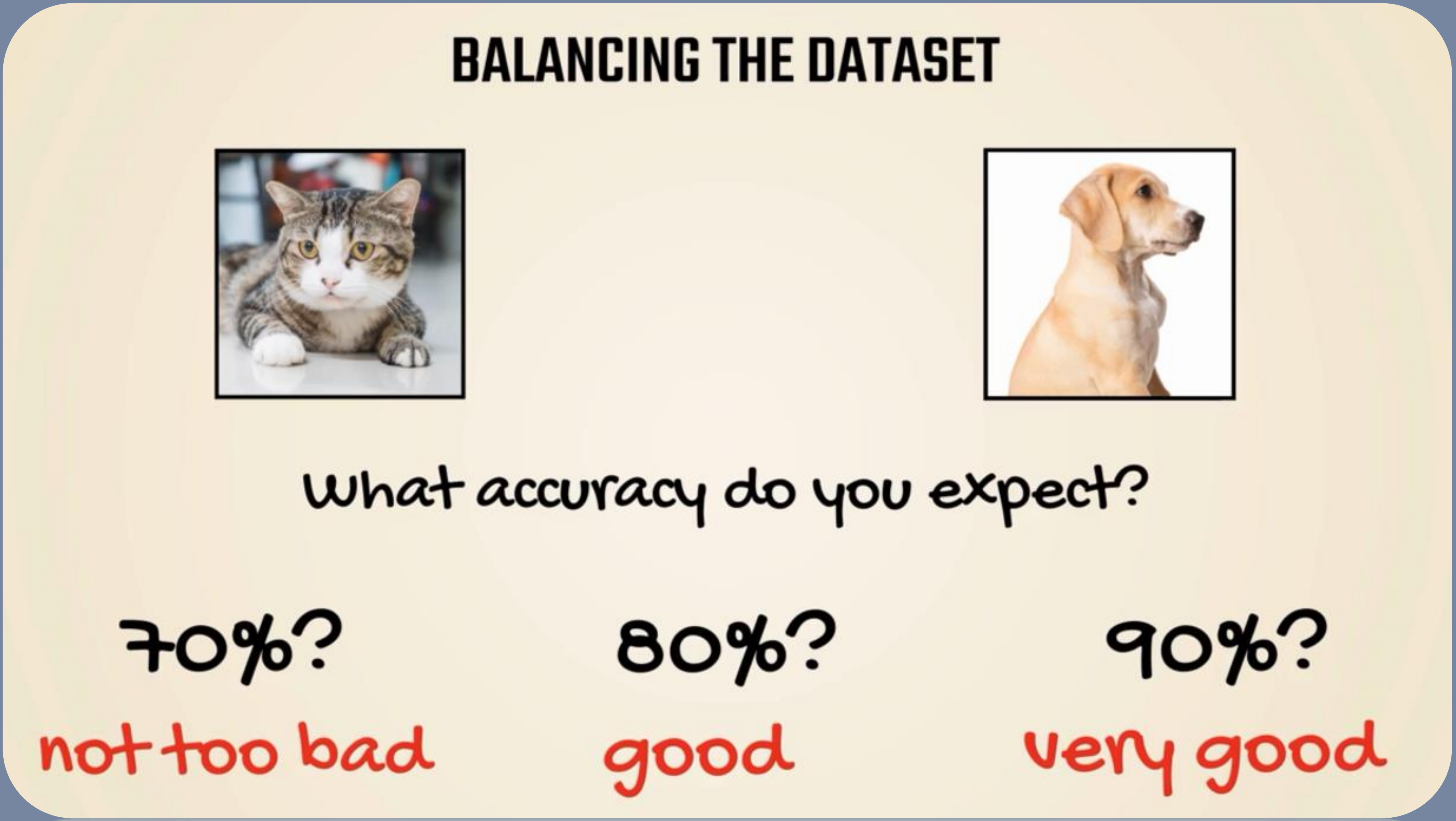




BALANCING THE DATASET

What accuracy do you expect?

70%? 80%? 90%?



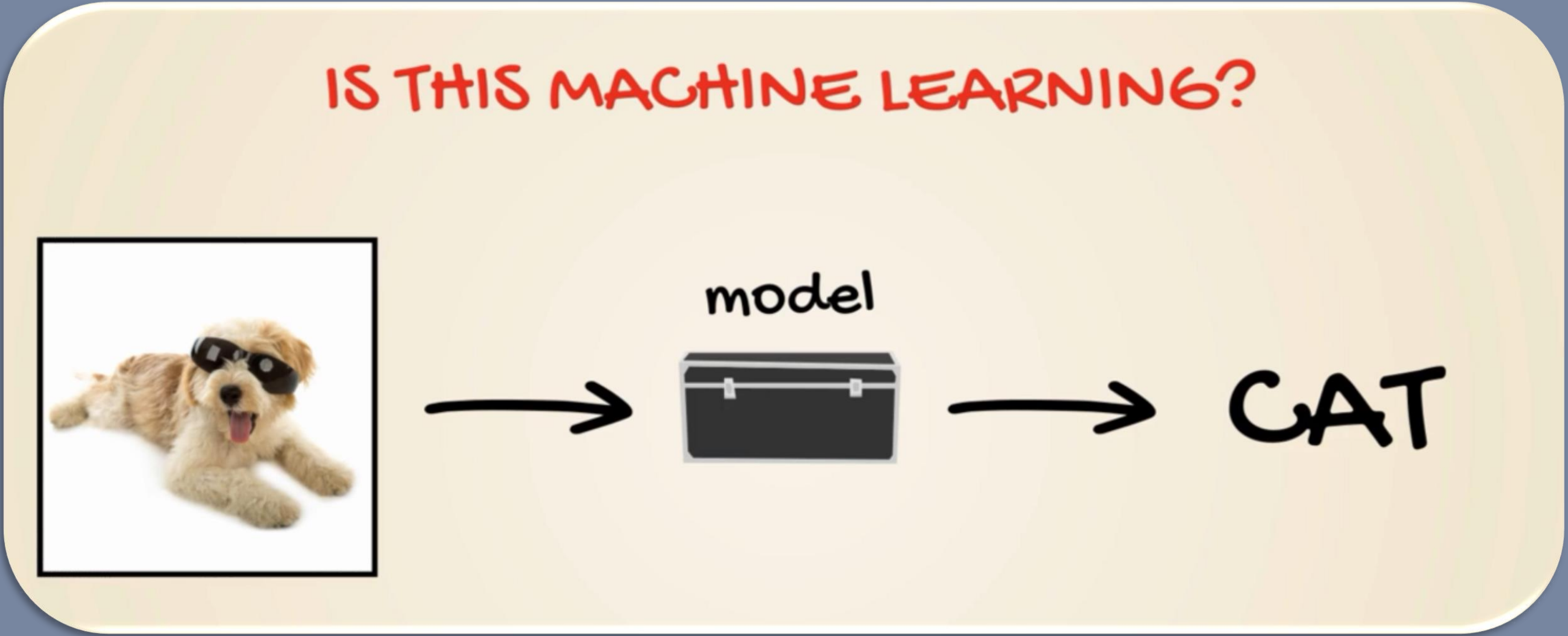
BALANCING THE DATASET

What accuracy do you expect?

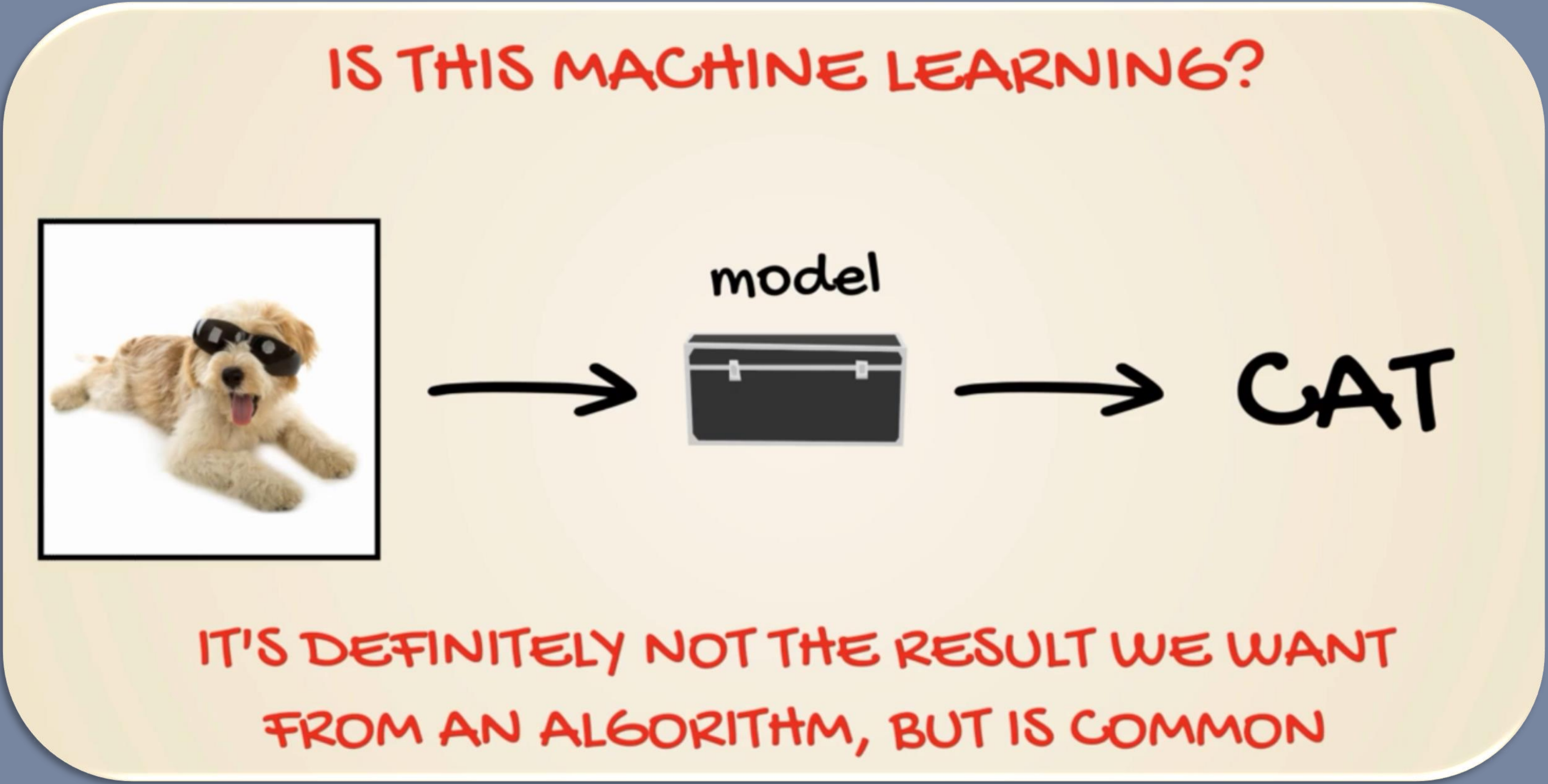
70%? not too bad

80%? good

90%? very good

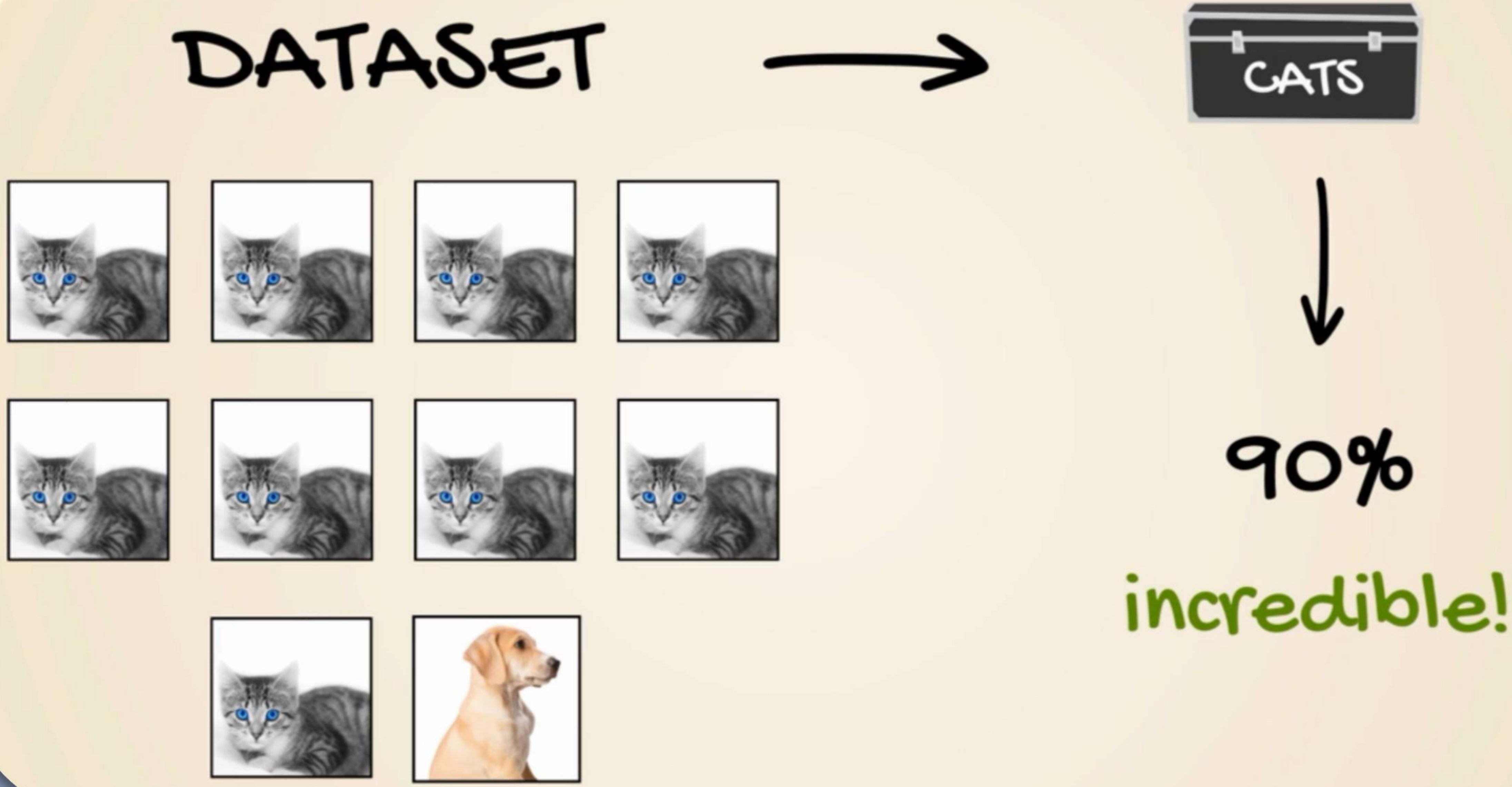


IS THIS MACHINE LEARNING?



IS THIS MACHINE LEARNING?

IT'S DEFINITELY NOT THE RESULT WE WANT FROM AN ALGORITHM, BUT IS COMMON



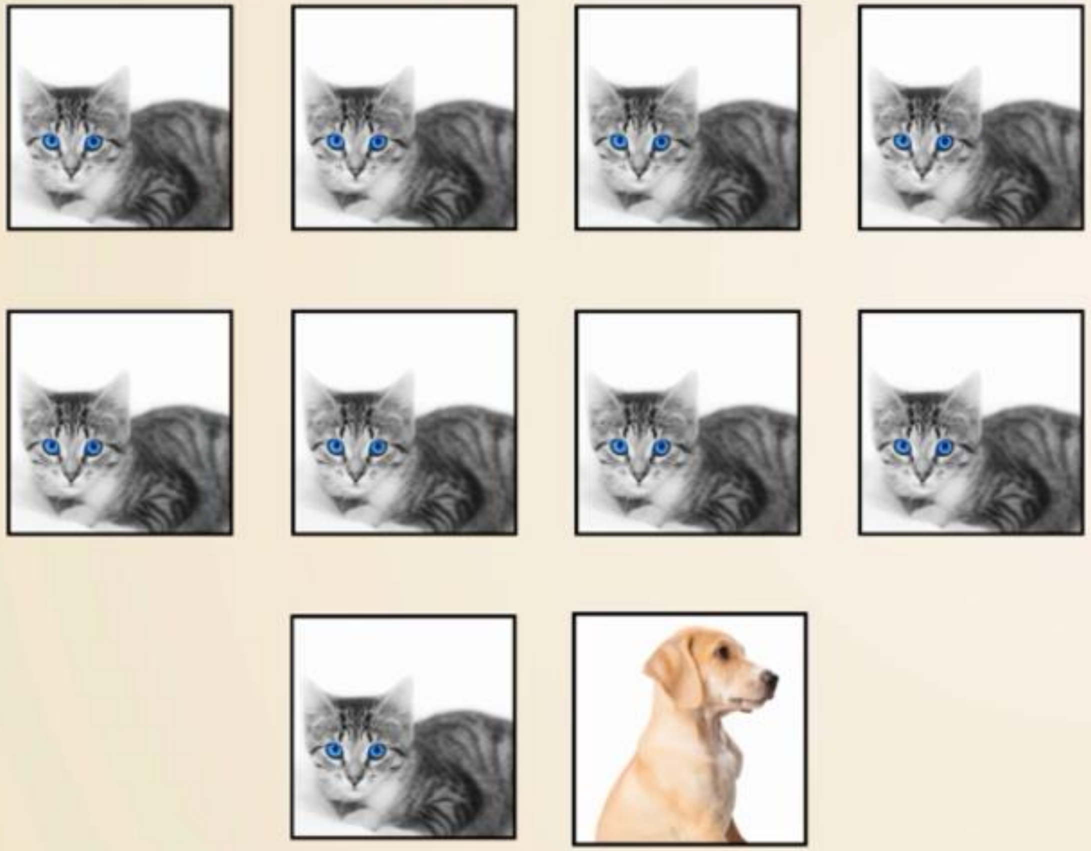
DATASET

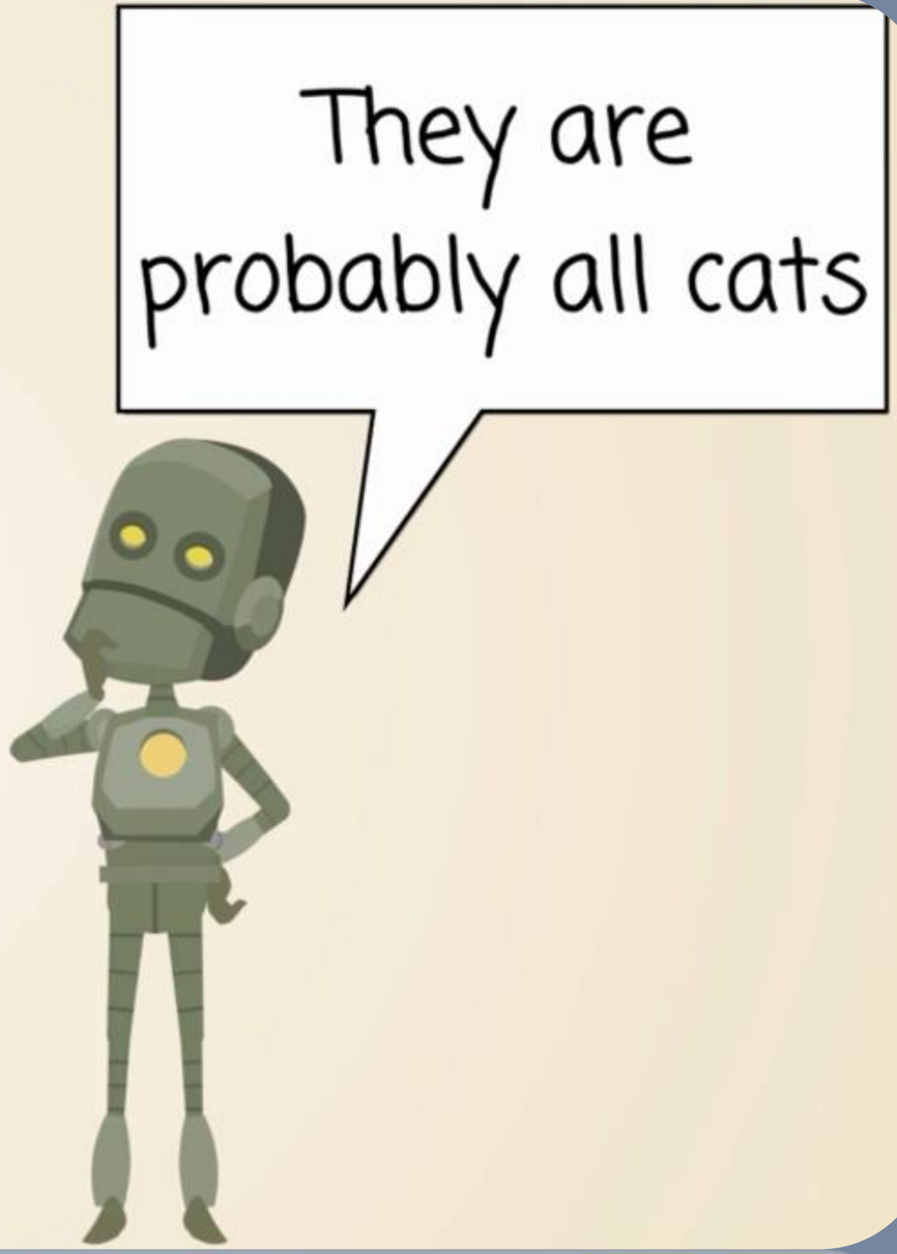
CATS

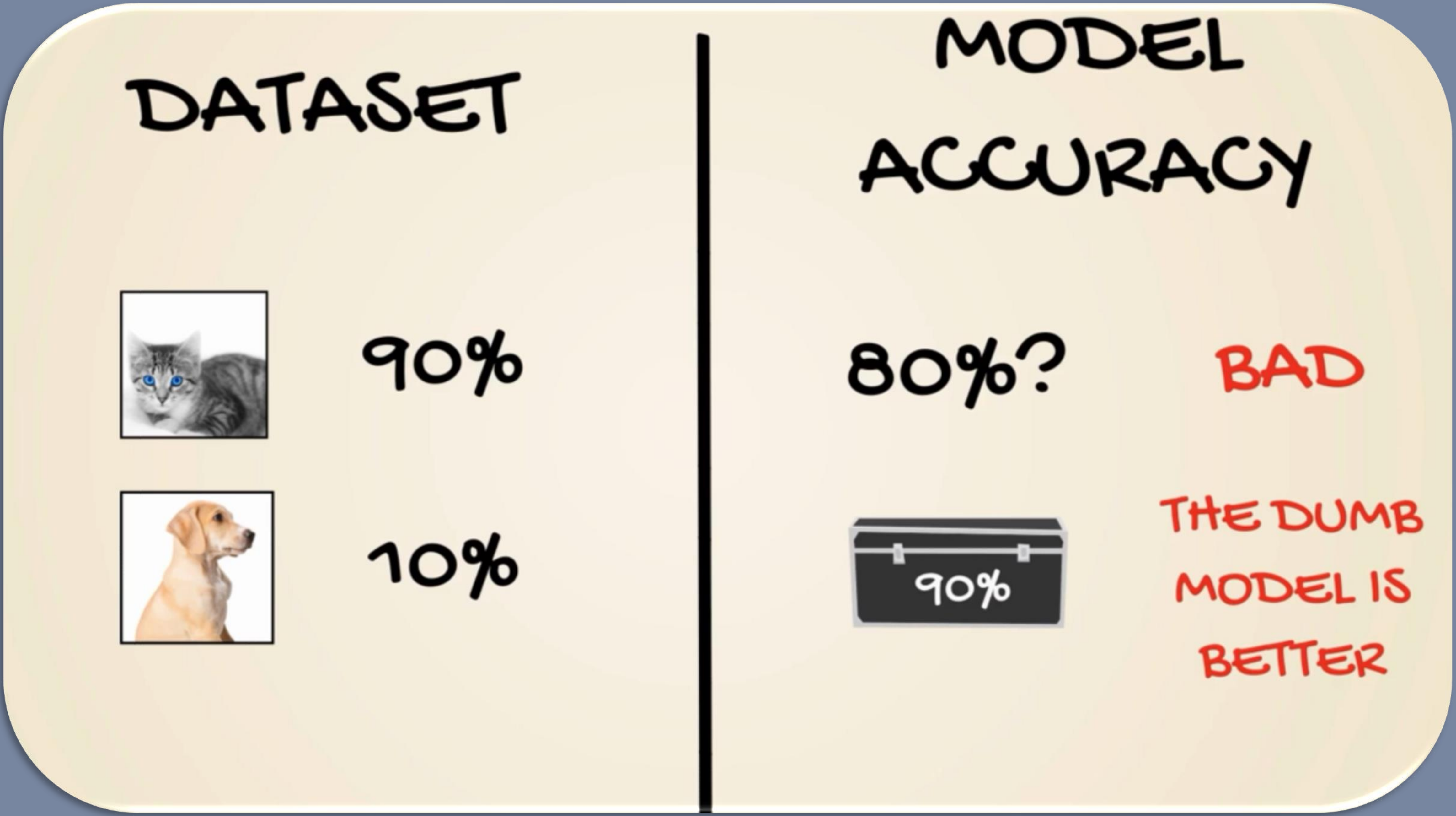
90%

incredible!

DATASET







DATASET

90%

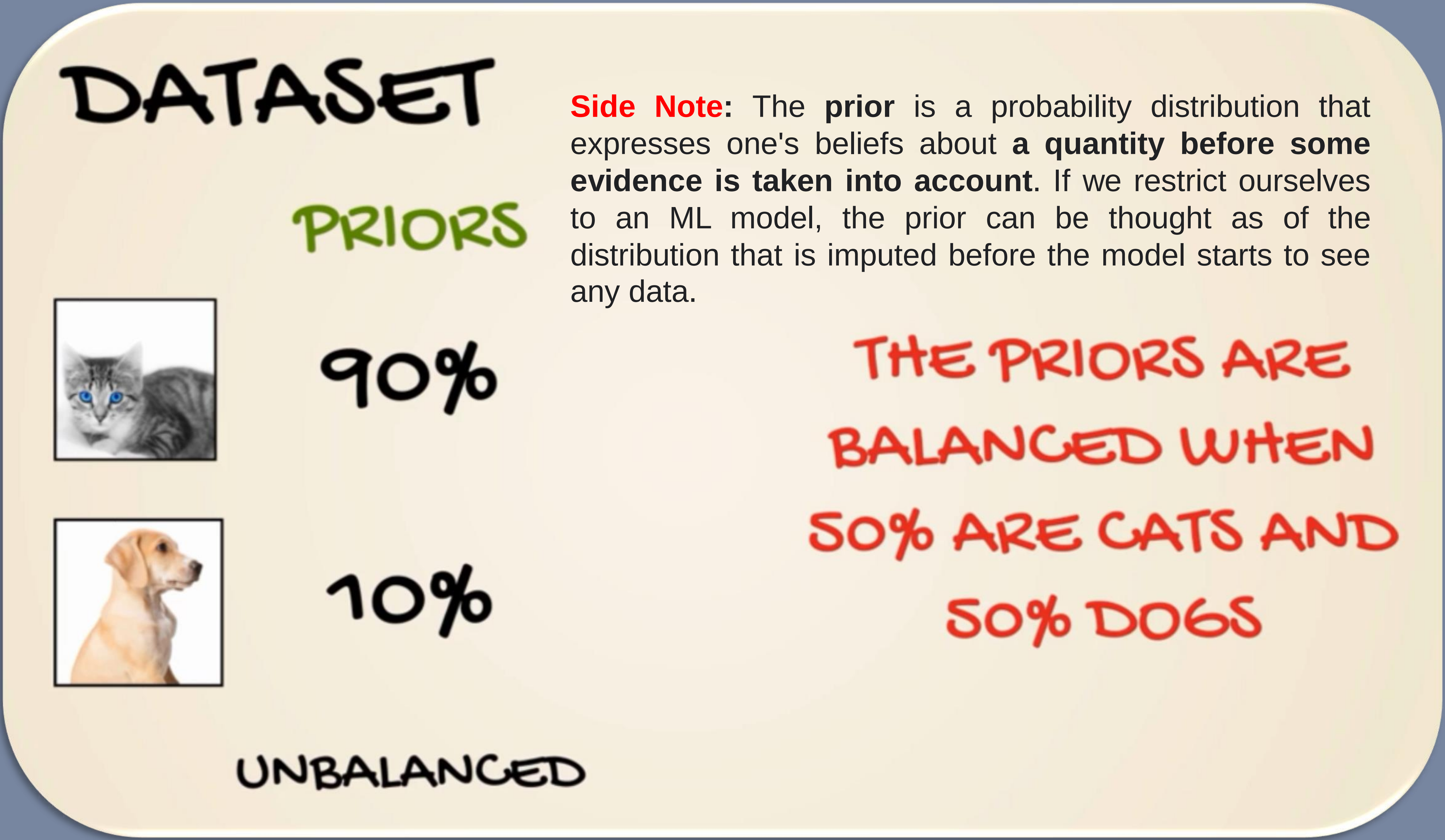
10%

MODEL ACCURACY

80%? BAD

90%

THE DUMB MODEL IS BETTER

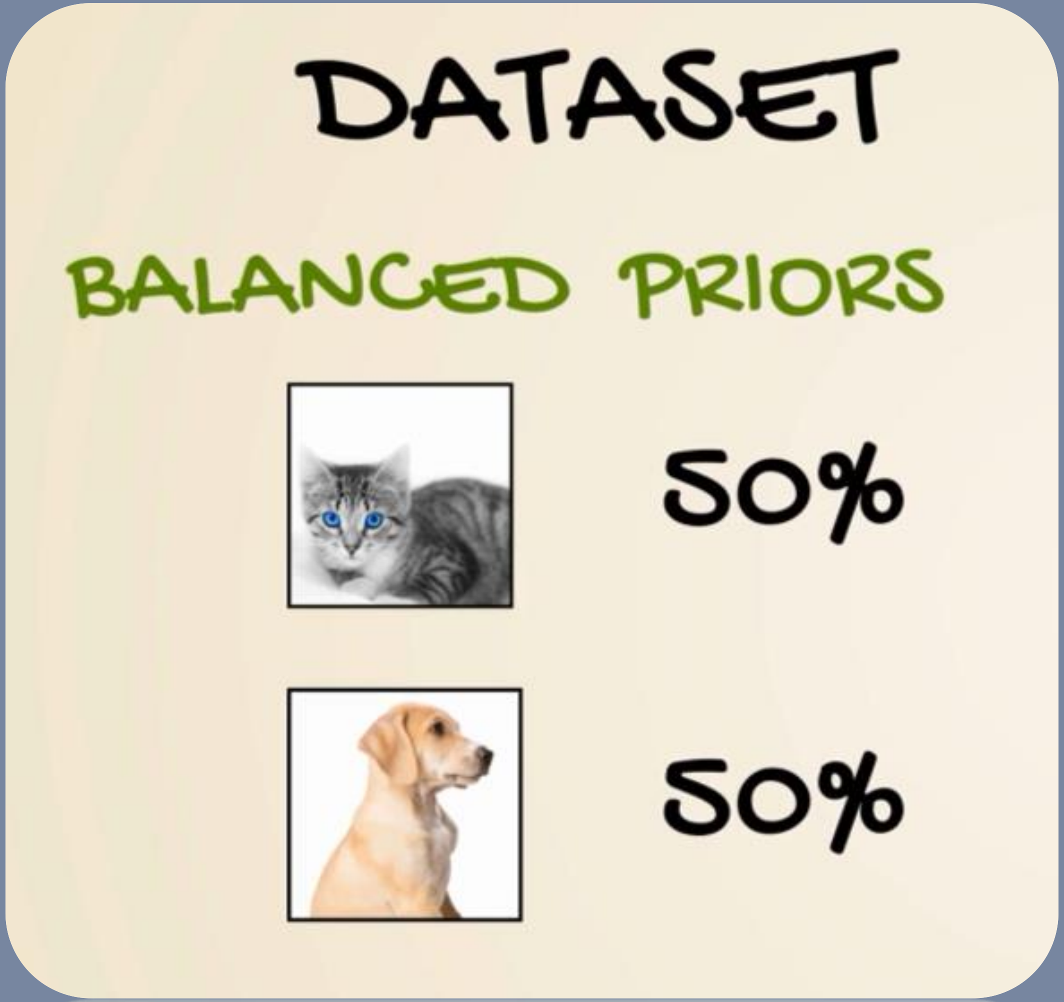


Side Note: The **prior** is a probability distribution that expresses one's beliefs about **a quantity before some evidence is taken into account**. If we restrict ourselves to an ML model, the prior can be thought as of the distribution that is imputed before the model starts to see any data.

DATASET

PRIORS

	90%	70%	60%
	10%	30%	40%
	UNBALANCED	UNBALANCED	UNBALANCED

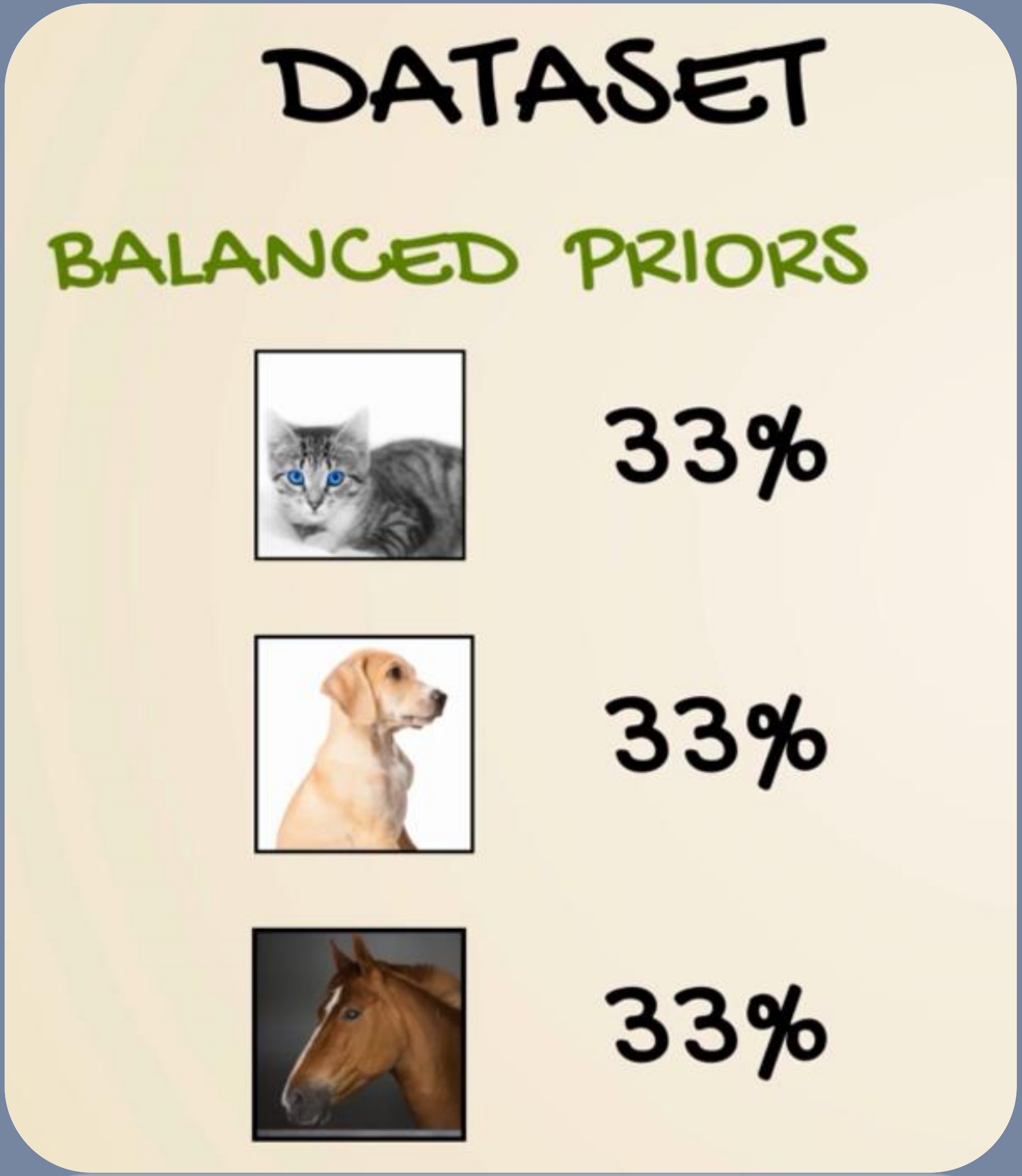


DATASET

BALANCED PRIORS

50%

50%



DATASET

BALANCED PRIORS

33%

33%

33%

Now, let’s take a look at the dataset in Excel