دوره‌ی آموزشی «علم داده»
Data Science Course

جلسه سی و یکم

نکاتی تکمیلی در خصوص پیش‌پردازش

مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

PREPROCESSING

هرگونه دست‌کاری روی دیتاست قبل از دادن اون به مدل رو پیش‌پردازش می‌گن

Any manipulation of the dataset before running it through the model

اما چرا این‌کار رو انجام می‌دیم؟

PREPROCESSING

motivation

- Compatibility

PREPROCESSING

motivation

- Compatibility

Exchange rate	Volume
~1	~100,000

- Orders of magnitude

Mathematically 1 is negligible w.r.t. 100,000

(~100,000) + (~1) = ~100,000

PREPROCESSING

motivation

- Compatibility
- Orders of magnitude
- Generalization

Same model, different issue

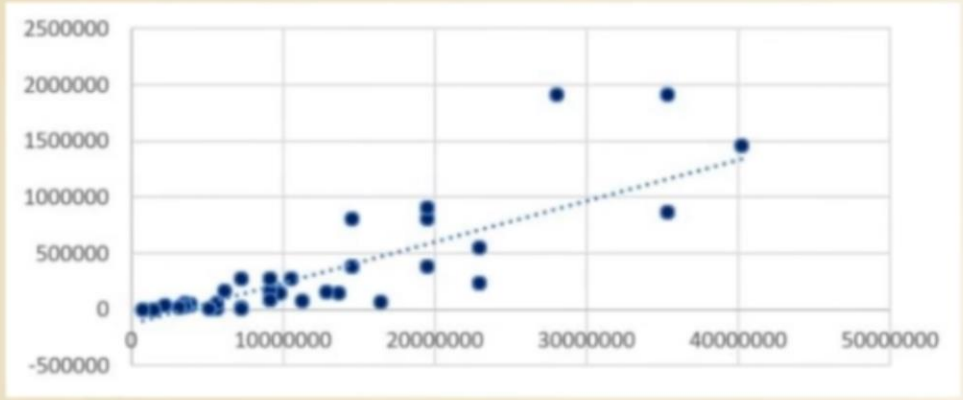
خب. با این مقدمه بریم سراغ موضوع و بحث اصلی‌مون

Relative metrics are especially useful when we have time-series data

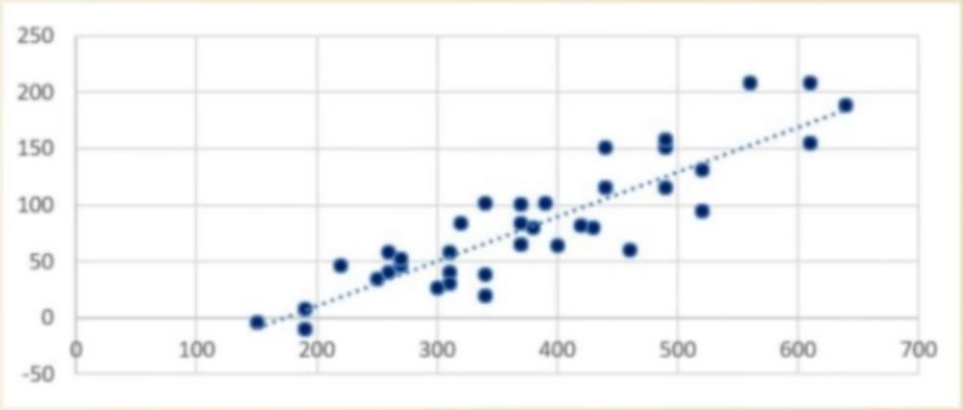
در دنیای بورس و بازارهای مالی، دوتا مقدار وجود داره. مقدار مطلق و مقدار نسبی (Relative Value). گاهاً بدلایل مختلف از مقدار نسبی استفاده می‌کنند که این باعث میشه دیتاست در ظاهر پراکندگی‌های زیادی در بخش‌هایش بوجود بیاد. اما خیلی ساده از یک تبدیل log (اسلاید بعد) استفاده می‌کنیم و این میزان زیادی از پراکندگی‌ها رو کاهش میده. پس دوتا متر اینجا وجود داشت. متر نسبی و متر مطلق.

LOGARITHMS

Financial data

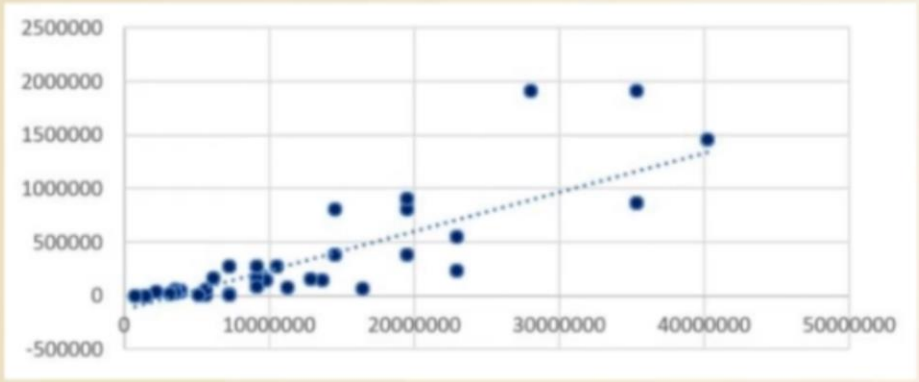


Log transformed financial data

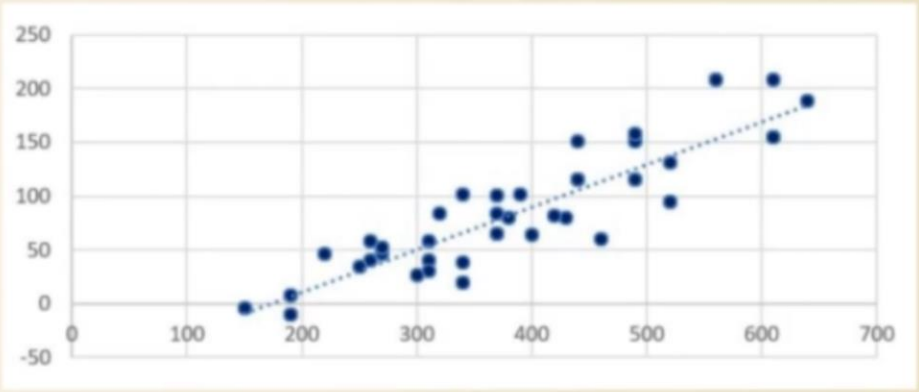


LOGARITHMS

Financial data



Log transformed financial data



Advantages:

Faster computation

Lower order of magnitude

Clearer relationships

Homogeneous variance

STANDARDIZATION

FEATURE SCALING

NORMALIZATION

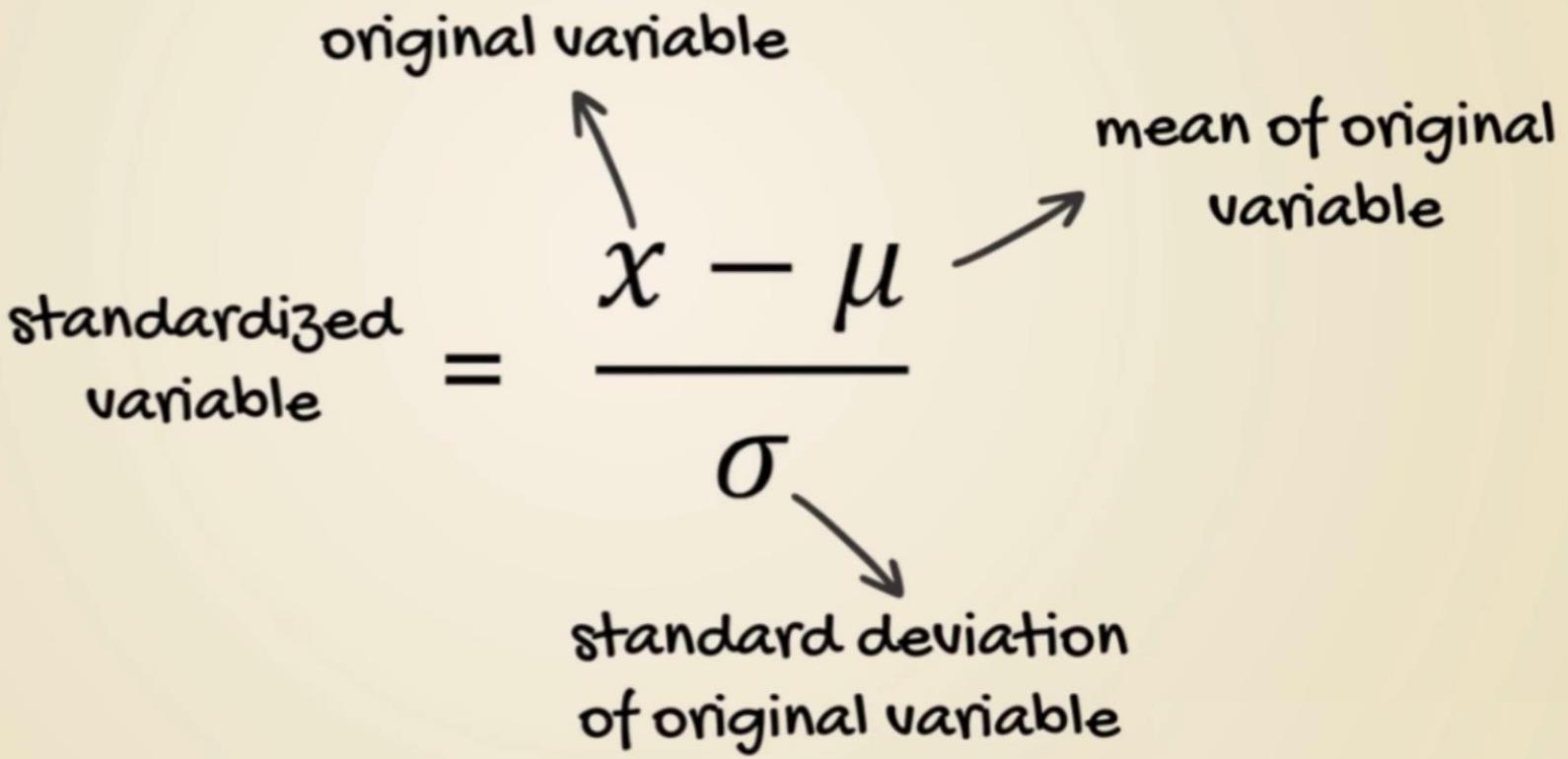
STANDARDIZATION

FEATURE SCALING

The process of transforming data into a standard scale

STANDARDIZATION

FEATURE SCALING



STANDARDIZATION

FEATURE SCALING

day	Exchange rate	Daily trading volume
1	1.3	110,000

STANDARDIZATION

FEATURE SCALING

day	Exchange rate	Daily trading volume
1	1.3	110,000
2	1.34	98,700

STANDARDIZATION

FEATURE SCALING

day	Exchange rate	Daily trading volume
1	1.3	110,000
2	1.34	98,700
3	1.25	135,000

STANDARDIZATION

FEATURE SCALING

day	Exchange rate	Daily trading volume
1	1.3	110,000
2	1.34	98,700
3	1.25	135,000

mean: 1.3

std: 0.045

STANDARDIZATION

FEATURE SCALING

$$\frac{x - \mu}{\sigma}$$

day	Exchange rate		Daily trading volume
1	0.07	(1.3)	110,000
2	0.96	(1.34)	98,700
3	-1.03	(1.25)	135,000

STANDARDIZATION

FEATURE SCALING

$$\frac{x - \mu}{\sigma}$$

day	Exchange rate		Daily trading volume	
1	0.07	(1.3)	-0.25	(110,000)
2	0.96	(1.34)	-0.85	(98,700)
3	-1.03	(1.25)	1.1	(135,000)

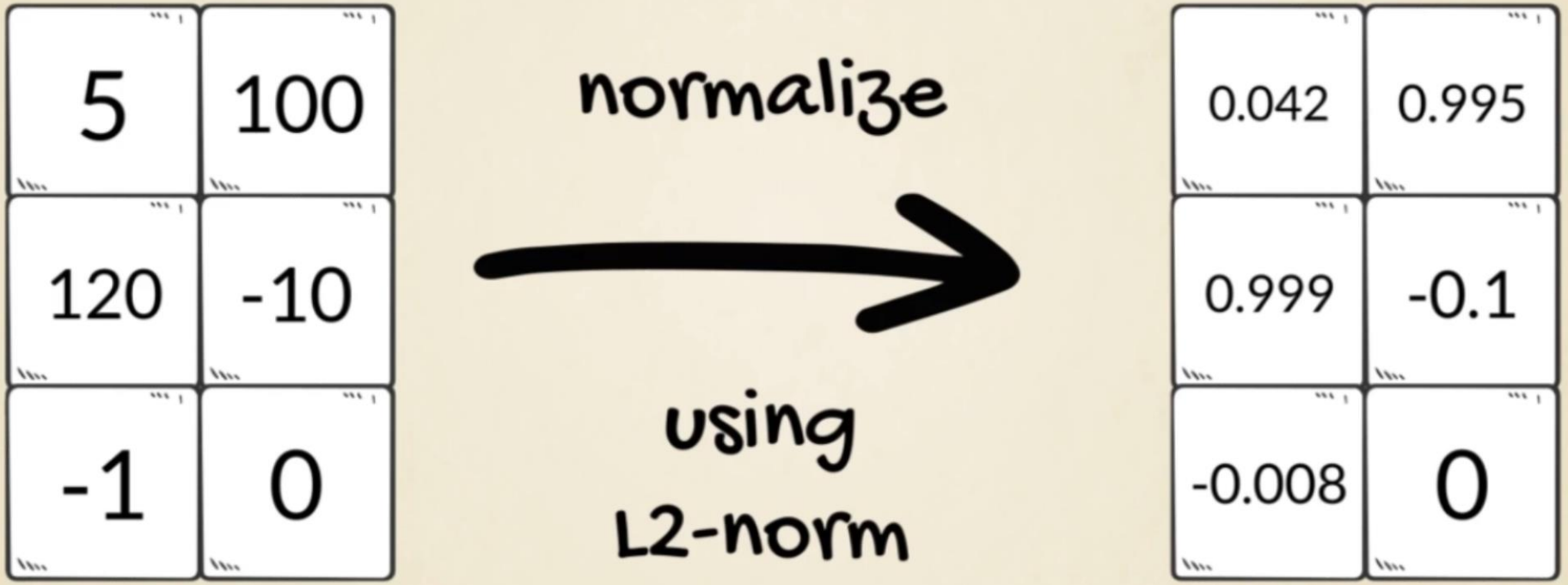
STANDARDIZATION

FEATURE SCALING

day	Exchange rate		Daily trading volume	
1	0.07	(1.3)	-0.25	(110,000)
2	0.96	(1.34)	-0.85	(98,700)
3	-1.03	(1.25)	1.1	(135,000)

We have forced the features to appear similar

NORMALIZATION



$$\begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 2 \end{bmatrix} \rightarrow \left[\begin{array}{c|c|c} -0.70710678 & 0 & 0.70710678 \\ \hline 0 & 0.4472136 & 0.89442719 \end{array}\right]$$

$$L^2 - Norm\,(-1,0,1) = \sqrt{(-1)^2 + 0^2 + 1^2} = \sqrt{2}$$

$$\left(\frac{-1}{\sqrt{2}}, \frac{0}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = (-0.70710678, 0, 0.70710678)$$

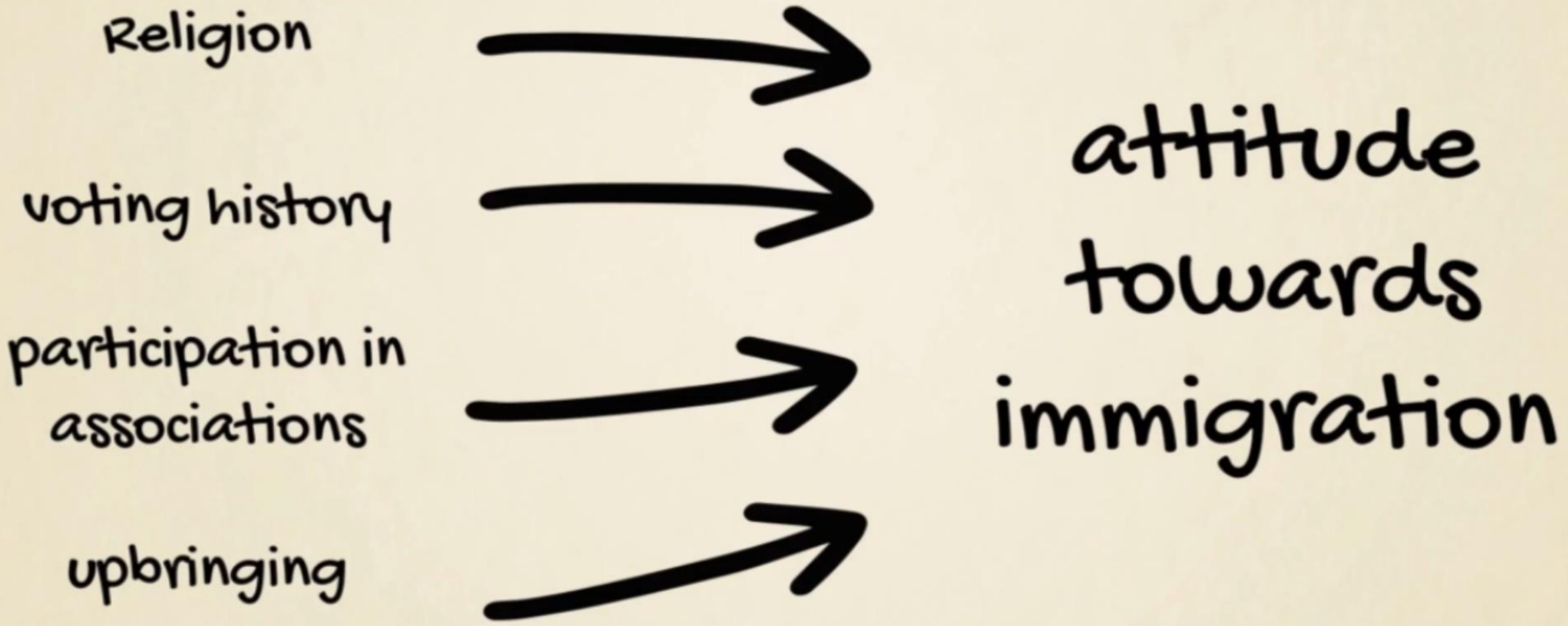
PCA

/principal components analysis/

Dimension reduction technique used to combine several variables into a bigger (latent) variable

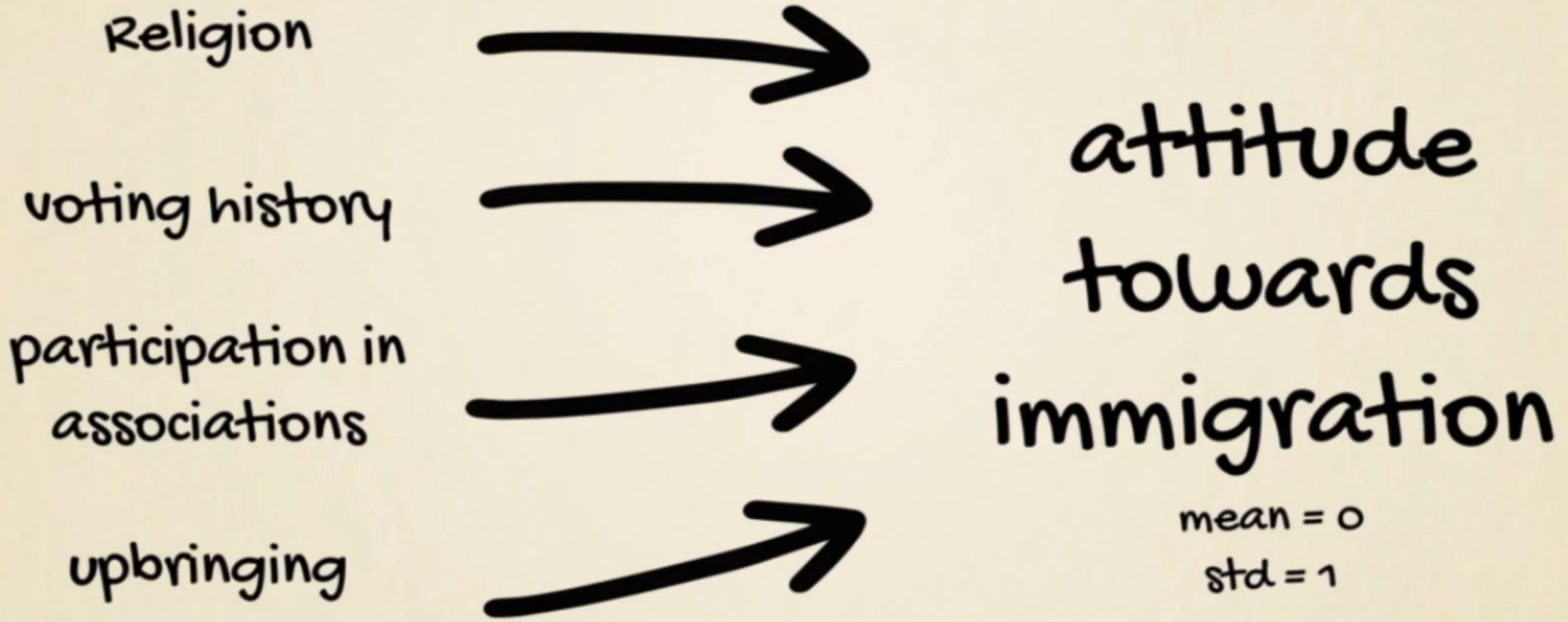
PCA

/principal components analysis/



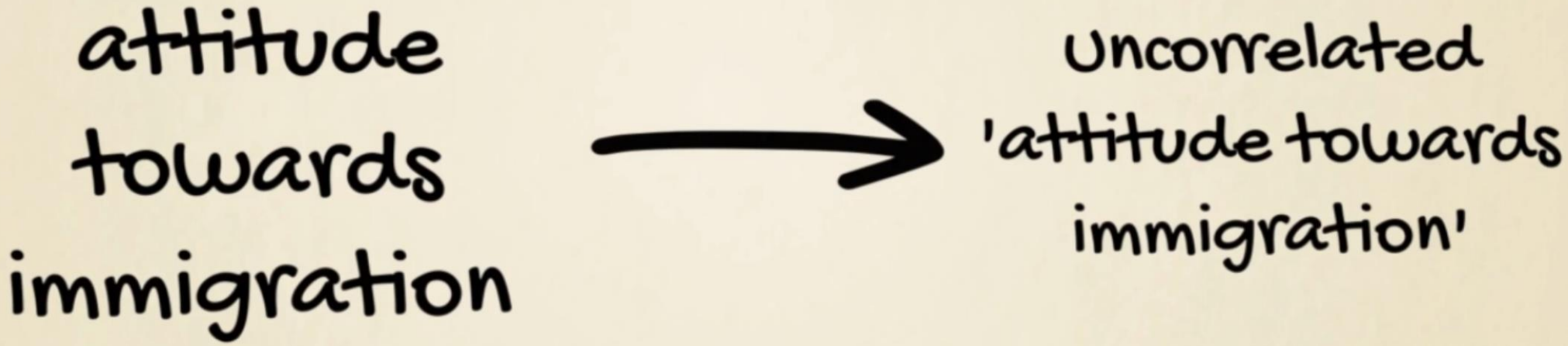
PCA

/principal components analysis/



WHITENING

معمولاً بعد از PCA و به منظور حذف
داده‌های وابسته بکار گرفته می‌شود



A whitening transformation or sphering transformation is a linear transformation that transforms a vector of random variables with a known covariance matrix into a set of new variables whose covariance is the identity matrix, **meaning that they are uncorrelated and each have variance 1.**

توضیح دقیق تمامی این موارد بسیار زمان‌بر است و عموماً هر کدام از اینها به مسئله‌ی در دست وابسته هست

اما مهمترین موضوع برای ما در ادامه دوره، بیشتر

Standardization

می‌باشد

CATEGORICAL DATA

Categories or groups (non-numerical data)

CATEGORICAL DATA

Categories or groups (non-numerical data)

numbers →  → numbers

CATEGORICAL DATA

Categories or groups (non-numerical data)

numbers →  → numbers

CATEGORICAL DATA

Categories or groups (non-numerical data)

tensors →  → numbers

CATEGORICAL DATA

CATEGORICAL DATA

1 2 3

CATEGORICAL DATA

We have assumed order where there isn't

1 < 2 < 3

CATEGORICAL DATA

How to encode categories in a way useful for ML

CATEGORICAL DATA

How to encode categories in a way useful for ML

One-hot encoding

Binary encoding

BINARY ENCODING

	ordinal	binary
	1	0 1
	2	1 0
	3	1 1

BINARY ENCODING

	ordinal	var 1	var 2
	1	0	1
	2	1	0
	3	1	1

BINARY ENCODING

	var 1	var 2
	0	1
	1	0
	1	1

By splitting the numbers in two variables we have removed the order

...but...

there are some implied correlation between them

BINARY ENCODING

	var 1	var 2
	0	1
	1	0
	1	1

Bread is the opposite of yogurt

Whatever is bread is not yogurt and vice versa

BINARY ENCODING

	var 1	var 2
	0	1
	1	0
	1	1

Muffins is the opposite of yogurt

Whatever is muffins is not yogurt and vice versa

ONE-HOT ENCODING

	bread	yogurt	muffins
	✓		
		✓	
			✓

ONE-HOT ENCODING

	bread	yogurt	muffins
	1	0	0
	0	1	0
	0	0	1

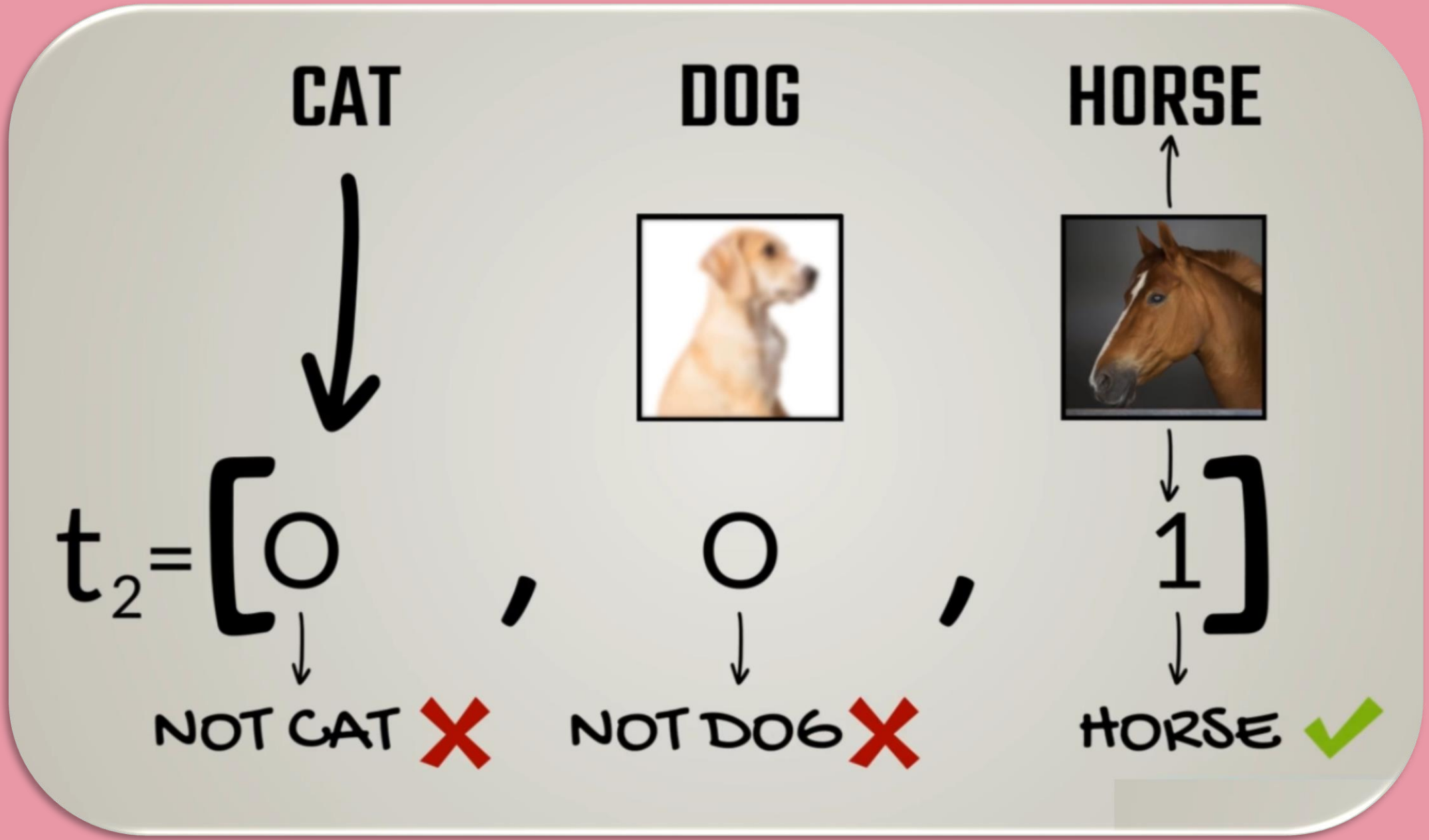
ONE-HOT ENCODING

	bread	yogurt	muffins
	1	0	0
	0	1	0
	0	0	1

uncorrelated

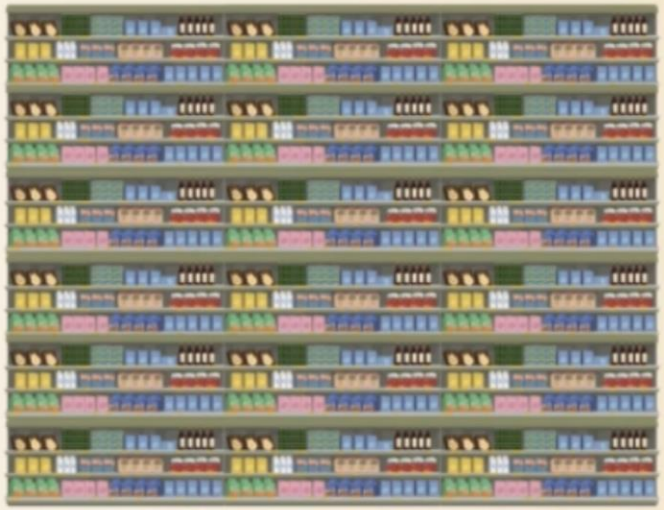
unequivocal

unambiguous


CAT
DOG
HORSE
$t_2 = [0, 0, 1]$
NOT CAT
NOT DOG
HORSE

ONE-HOT ENCODING

problem?



12,000 products → 12,000 columns

12,000 products

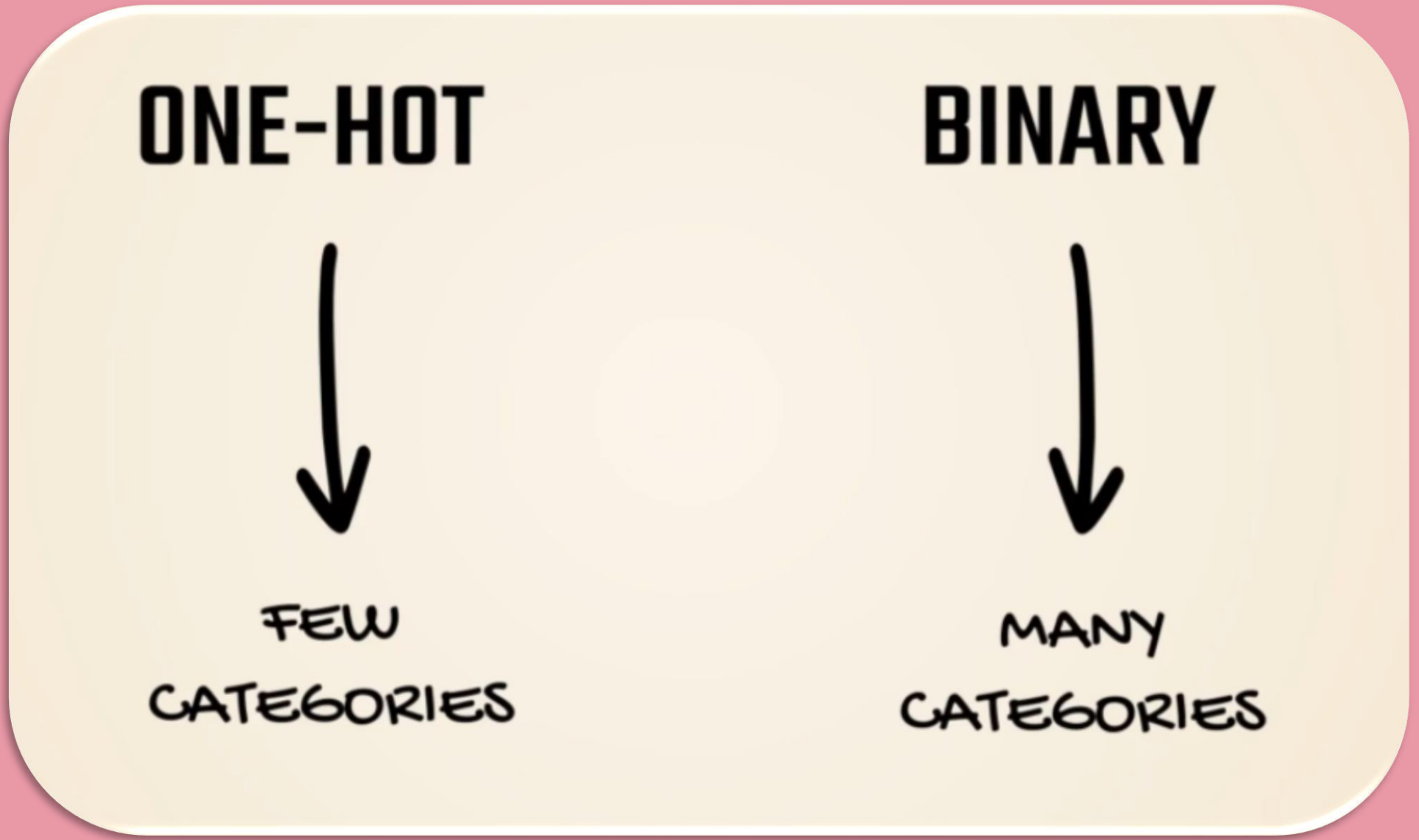
	ONE-HOT	BINARY
NEW COLUMNS	12,000	16

12,000 IN BINARY: 10111011100000

12,000 products

	ONE-HOT	BINARY
NEW COLUMNS	12,000	16 BETTER!!!

12,000 IN BINARY: 10111011100000


ONE-HOT
FEW CATEGORIES
BINARY
MANY CATEGORIES

در ویدیوی بعدی راجع به دسته‌بندی دیتاست
MNIST
صحبت خواهیم کرد

Stay Tuned and have FUN