ENCODING AND TRANSFORMING DATA

LECTURE #8
SECTION 1
JUNE 15

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

UNIVERSITY of
RWANDA

CATEGORICAL DATA

CATEGORICAL DATA:

So far our covariates have been numerical data, data that are recorded as real numbers and these values can be ordered, e.g. Age = 50 is greater that Age = 20.

But many important real-life measurements are not numerical, e.g. if nationality = "USA", "Rwanda", ... Why can't we simply encode the values as numbers?

"USA" = 0, "Rwanda" = 1, ...

Encoding categories as numbers is misleading, since numbers imply ordering:

"Rwanda" = 1 > 0 = "USA"...

Instead, we encode these categorical variables as binary vectors:

One-hot Encoding

Nationality	"USA"	"Rwanda"	"China"	"Mexico"
"USA" →	[1 ,	0 ,	0 ,	0]
"Rwanda" →	[0 ,	1 ,	0 ,	0]
"China" →	[0 ,	0 ,	1 ,	0]
"Mexico" →	[0 ,	0 ,	0 ,	1]

TEXT DATA

REPRESENTING TEXT DATA:

Sometimes our covariates contain entire blocks of text, e.g.

diagostic_notes_1 = "The Patient has fever and chill and has not eaten."
diagostic_notes_2 = "The patient has no fever and had no food."

We can encode each word using one-hot encoding and add-up all the onehot vectors. This simply counts the number of times each word occures in the text. This is called the count vector:

Count Vector Representations

	the	Patient	has	fever	and	chills	not	eaten	no	food
diagostic_notes_1	1	1	2	1	2	1	1	1	0	0
diagostic_notes_2	1	1	1	1	1	0	0	0	1	1

We need to worry about:

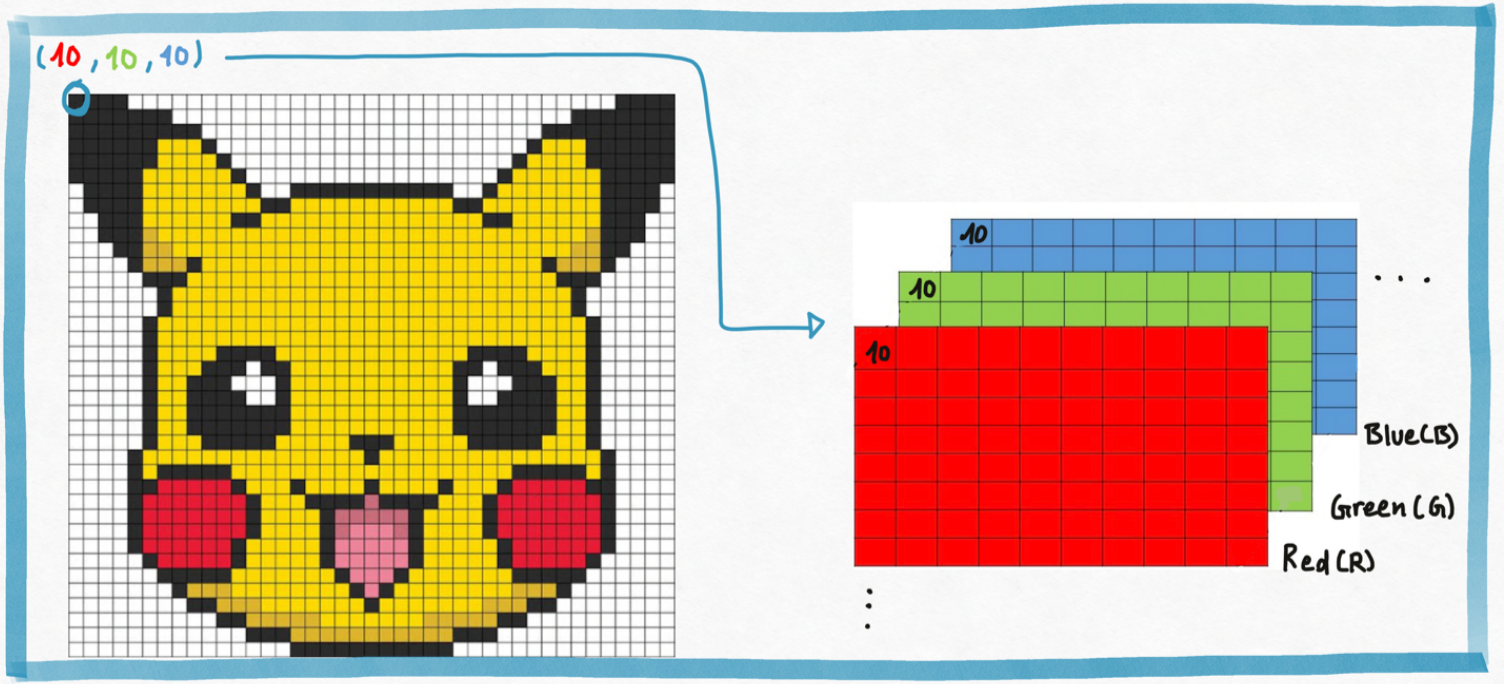
1. Punctuation
2. conjugation
3. words with little meaning: "the", "an", "and",...
4. ambiguity: "the river bank", "the bank account"

IMAGE DATA

REPRESENTING IMAGE DATA:

Images are represented as a grid of numbers. Each cell in the grid is called a pixel and consists of three numbers: (R, G, B) for how much red, green and blue color to display for the pixel.

So each image with LxH pixels is represented by 3 LxH matrices, one for each colour channel.



FLATTENED IMAGE VECTORS:

The representation of images as three color matrices is inconvenient to work with, all of our models take vectors as inputs.

In order to apply our models to image data, we flatten the matrices into a single vector.