Obtaining Features in an Unstructured Setting

School Connect: Intro to DS & AI

Vivek Sivaramakrishnan
Indian Institute of Technology, Madras

Data

- Structured Data (Tabular)

Player	Team	Ave	SR
Ibrahim Zadran	Afghanistan	28.87	107.44
Rahmanullah Gurbaz	Afghanistan	35.12	124.33
JC Buttler	England	42.8	158.51
PD Salt	England	37.6	159.32

- *Unstructured Data* - text, images, videos, ...
- How to use such data?
- *Somehow* convert them to (a collection of) numbers

Text

- How is text stored in a computer?

```
Input text here...
```

```
ASCII Output here
```

- The above is a (possible) way of representing text as numbers (features).

The Bag-of-Words approach

- Count the number of words in each sentence/document.

```
quick brown fox jumps over lazy dog
the lazy dog slept in the sun
the quick brown fox and the lazy dog
jumped over the lazy dog quickly
```

	quick	brown	fox	jumps	over	lazy	dog	the	slept	in	sun	and	jumped	quickly
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
2	0	0	0	0	0	1	1	2	1	1	1	0	0	0
3	1	1	1	0	0	1	1	2	0	0	0	1	0	0
4	0	0	0	0	1	1	1	1	0	0	0	0	1	1

Sentence Similarity

- Consider the following sentences:

```
1  quick brown fox jumps over the lazy dog
2  pen is mightier than the sword
3  lazy dog slept in the sun
```

- Observe that sentences 1 and 3 are more *similar* to each other, than with 2.
- We must concretize what we mean by *similarity*
- A possible definition:

> **Sentence Similarity**
>
> Given 2 sentences, the *similarity* between them can be defined as the number of words common to each other.

Sentence Similarity Example

quick brown fox jumps over the lazy dog

lazy dog slept in the sun

quick brown fox jumps over the lazy dog

pen is mightier than the sword

Sentence Similarity Example

quick brown fox jumps over the lazy dog

lazy dog slept in the sun

quick brown fox jumps over the lazy dog

pen is mightier than the sword

- The first and second pairs of sentences have a similarity score of 3 and 1 respectively.

Enter `Bag-of-Words`

- Now suppose the `Bag-of-Words` features are given instead of the sentences.
- Can the similarity score (between `1-3` and `2-3`) still be computed?

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

- Yes! Consider the dot product.

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) =
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) =
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) =  1 + 1 + 1
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) = 3
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) = 3
similarity(2, 3) =
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) = 3

similarity(2, 3) =
```

✨ Dot-Product ✨

	quick	brown	fox	jumps	over	the	lazy	dog	slept	in	sun	pen	is	mightier	than	sword
1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
3	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0

```
similarity(1, 3) = 3
similarity(2, 3) = 1
```

The *Bag-of-Words* Advantage

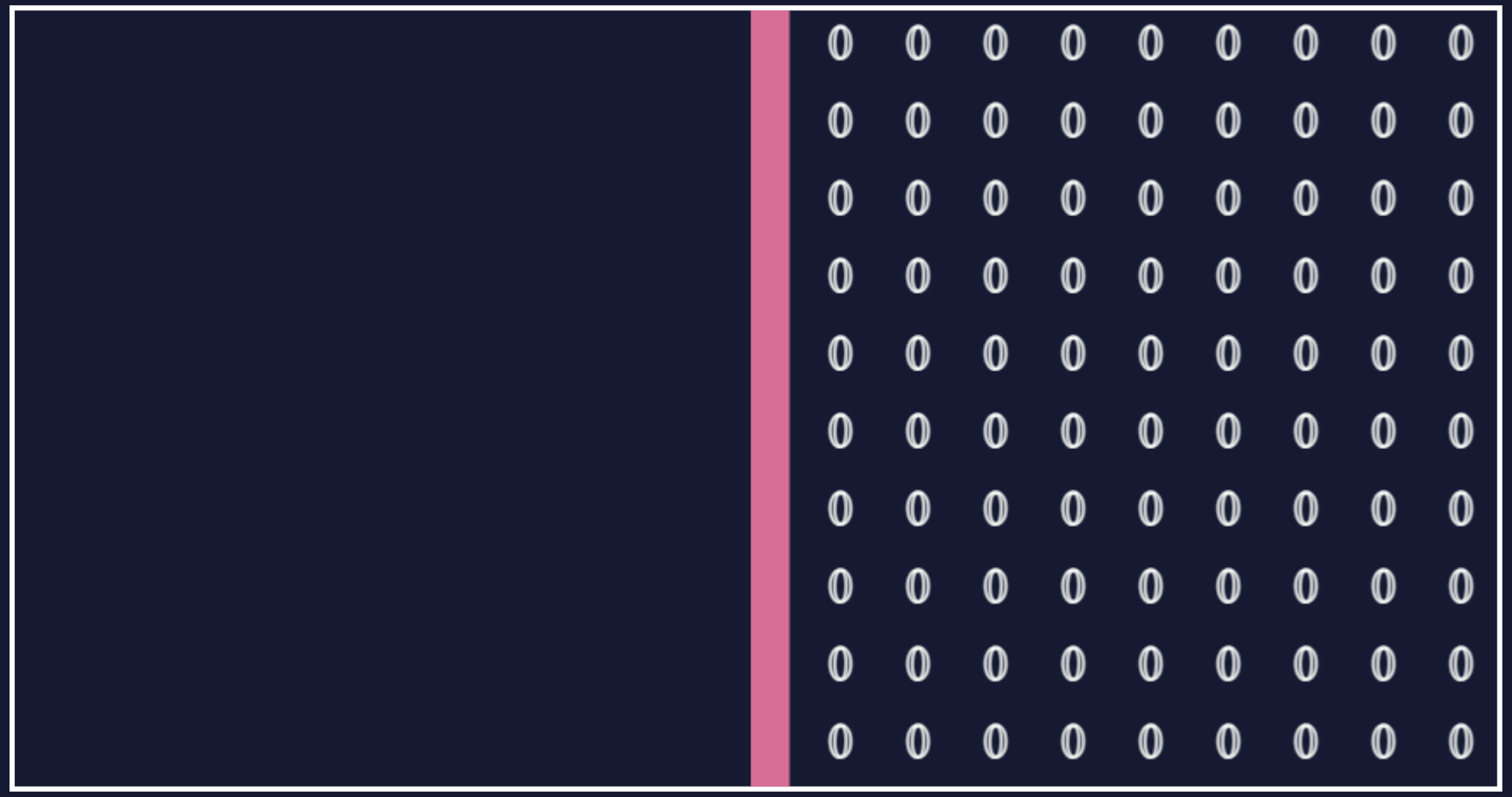
- Observe that by construction of the bag-of-words features, the calculation for sentence-similarity is more straightforward.
- All the resulting features are of an equal *length*. This is generally a desirable property.
- The ASCII representation will incur more steps in the calculation of sentence similarity.

Thought Exercise 1

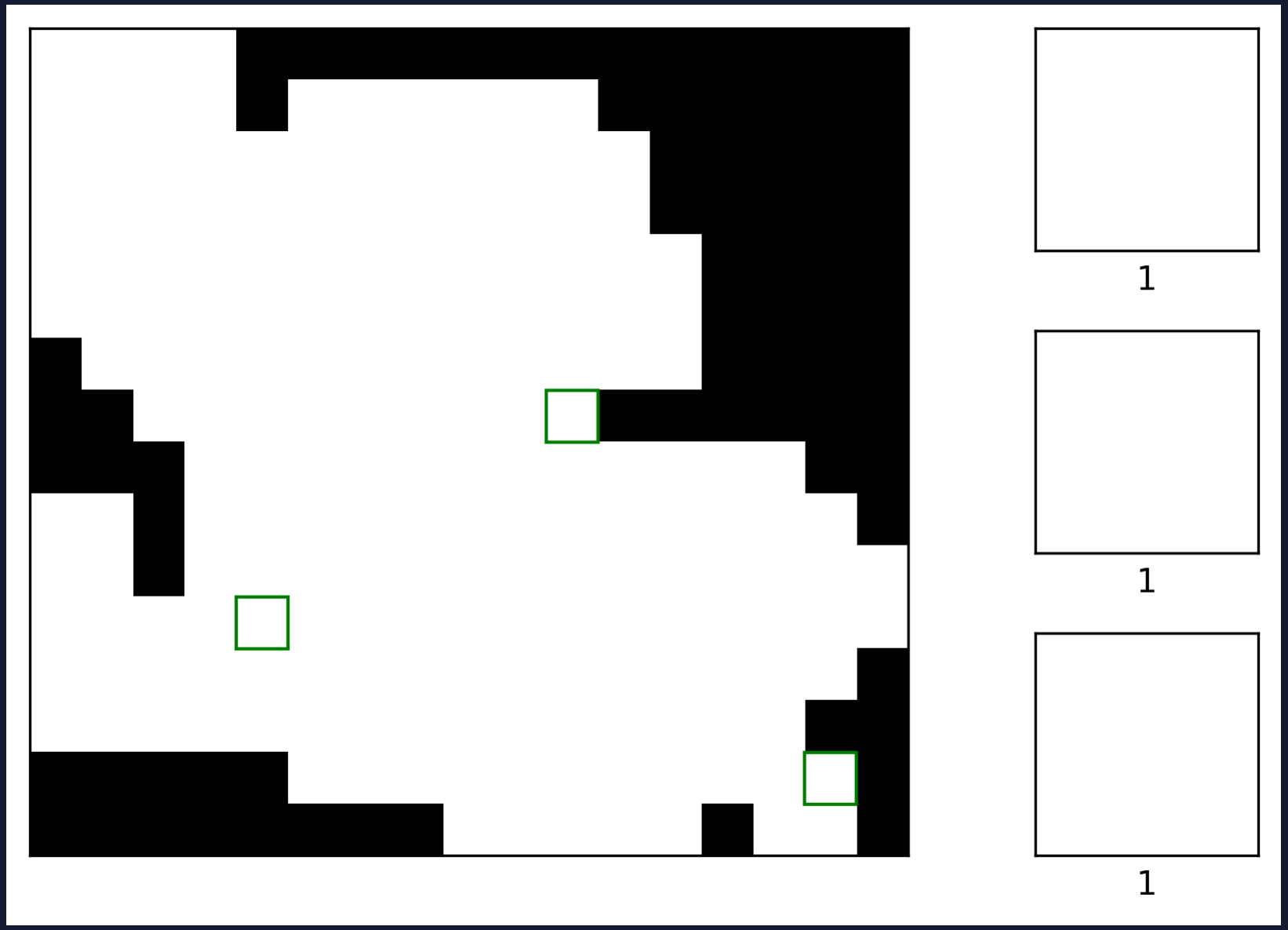
- Given the *Bag-of-Words* features, is it possible to reconstruct the original sentence? Is it possible in the ASCII case?
- Recall that the similarity between sentences `2` and `3` is `1`, because of the word `the`.
- The words `slept` and `sleeping` are classified into two separate columns.
- Words like `sleepy` and `drowsy` that convey more or less the same meanings (synonyms) also are given separate columns
- Is this desirable? Can we come up with more meaningful features, such that the similarity score also becomes meaningful?
- The *Natural Language Processing* field delves into the above questions deeply.

Pictures

- How are pictures stored in a computer? Using `Pixels`

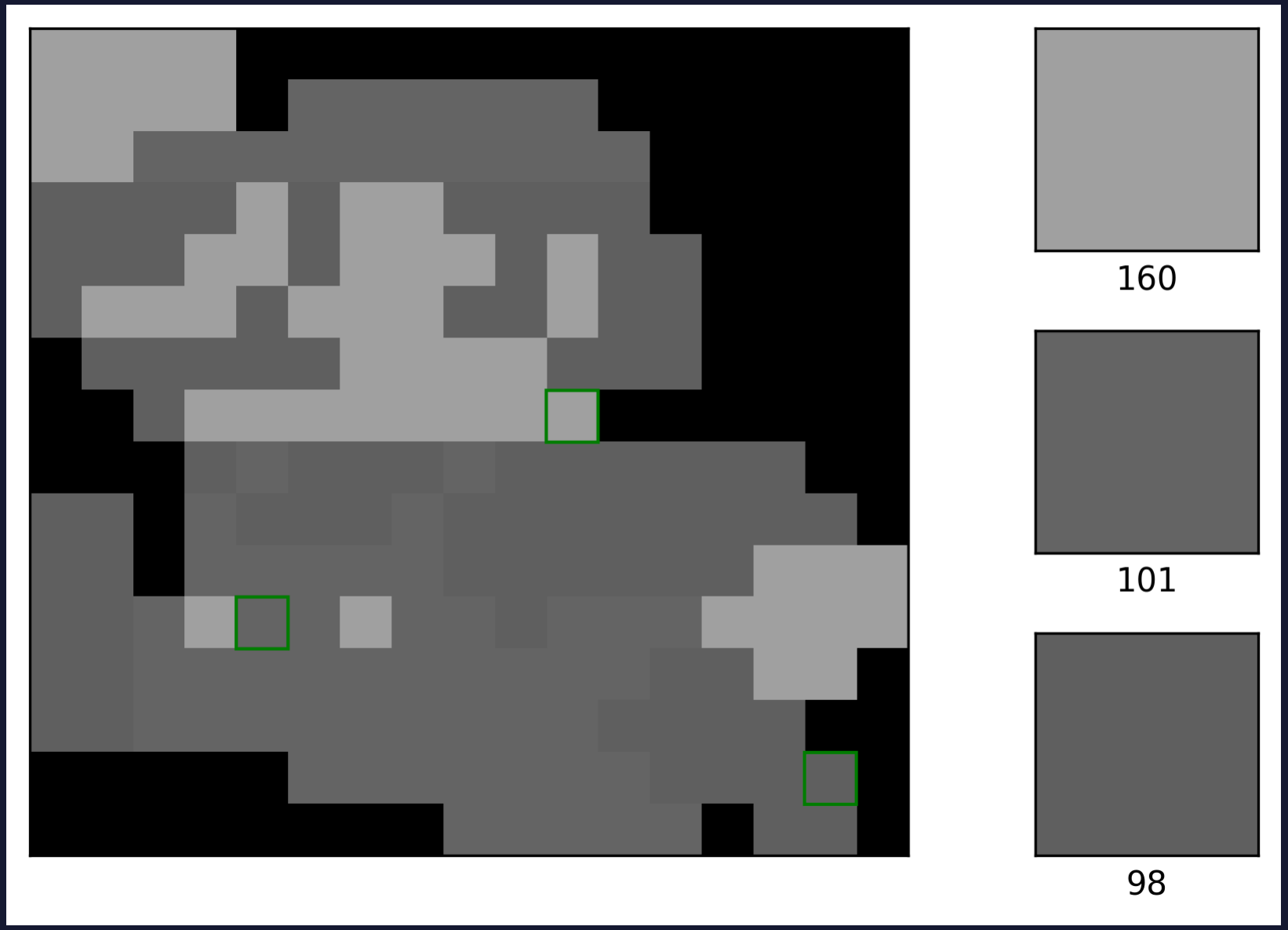


The more the colours, the more possible values a pixel can take.



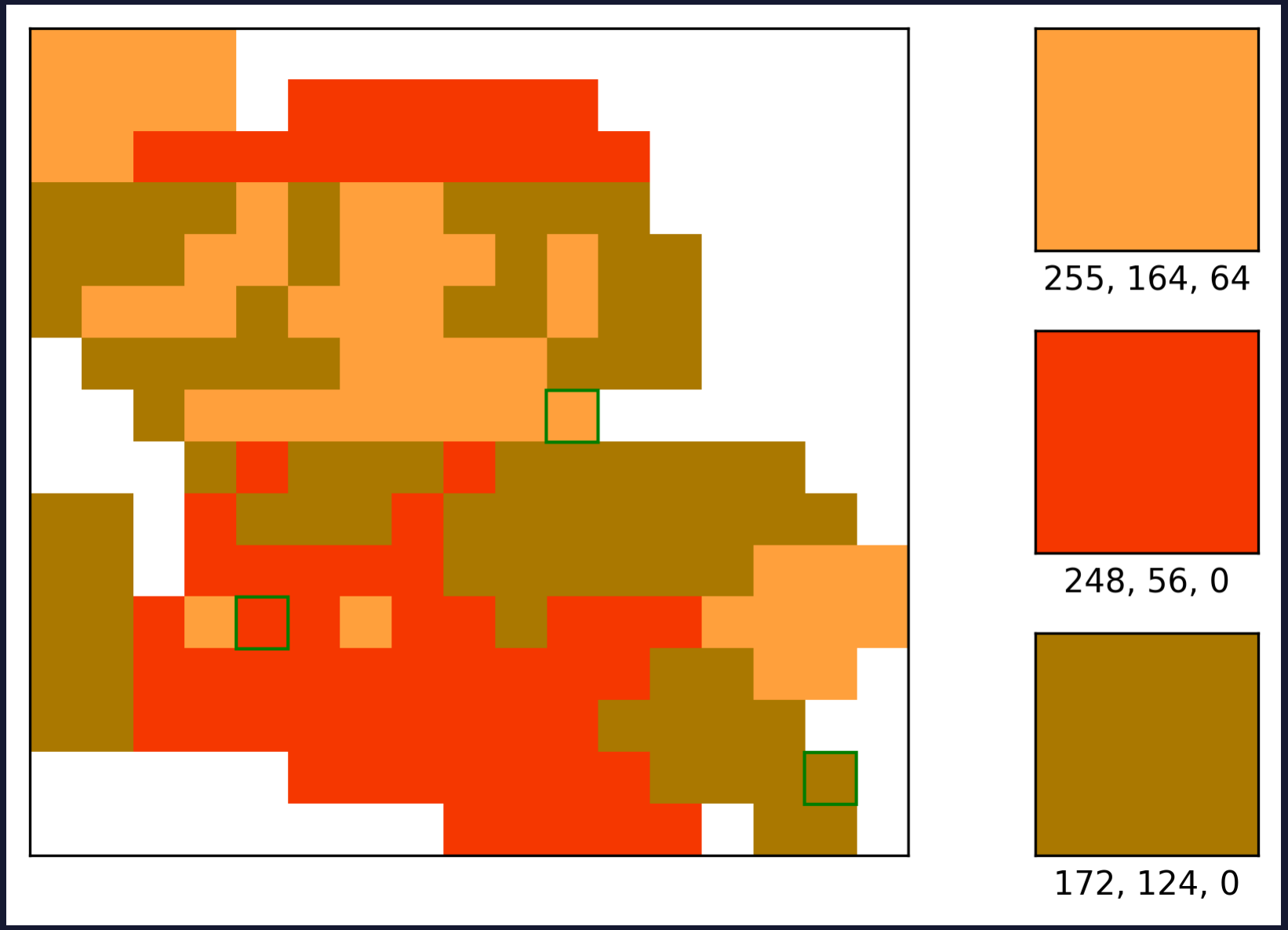
- Number of Colors = 2

The more the colours, the more possible values a pixel can take.



- Number of Colors = 256

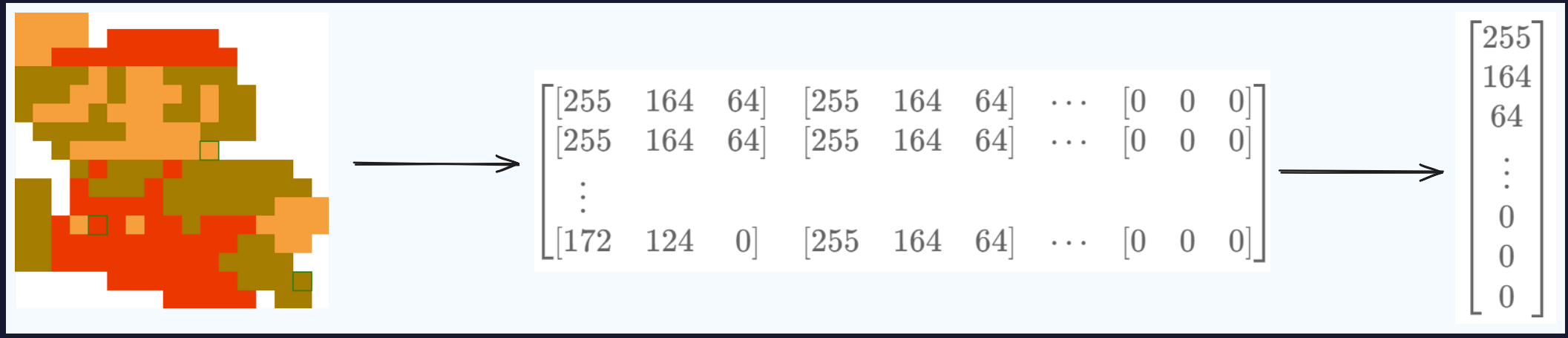
The more the colours, the more possible values a pixel can take.



- Number of Colors = $256^3 = 16777216$

Picture Features

The raw pixel values of the image can be used as features for algorithms.



Mario To Pixels