# Data Representation

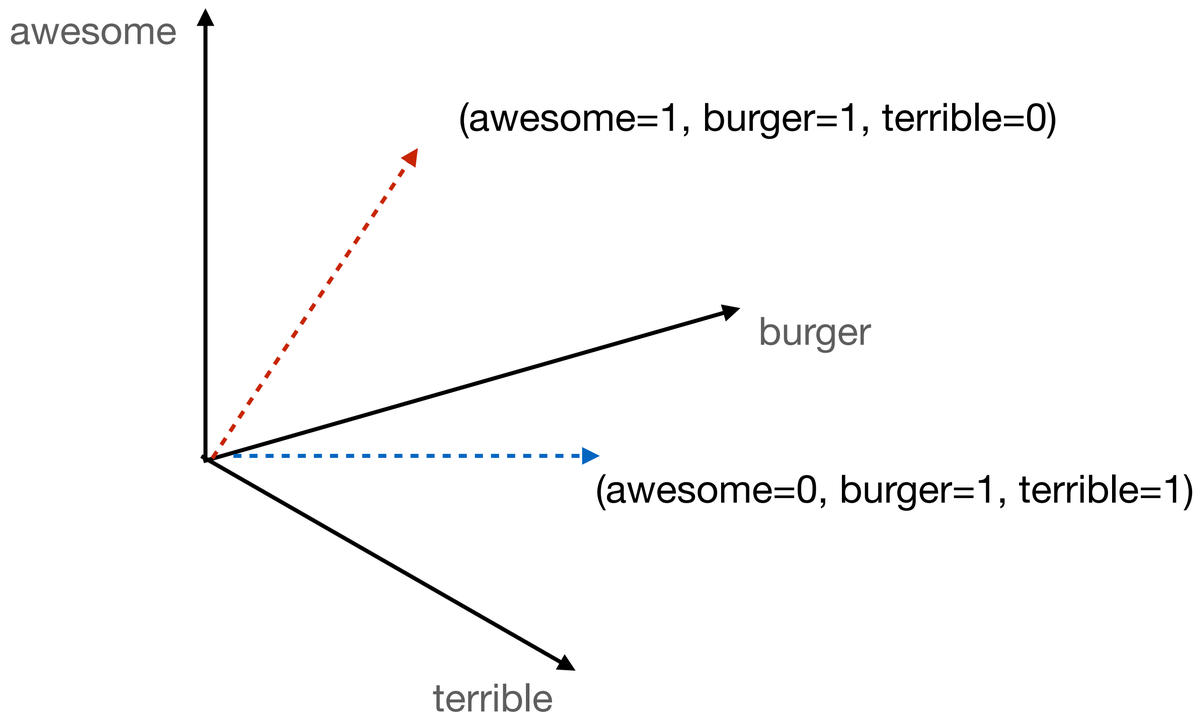
COMP337/COMP527 - Data Mining and Visualisation

**Procheta Sen**

UNIVERSITY OF LIVERPOOL

# Representing categorical data

If you must represent categorical data, then you could do so by representing each category as a separate dimension of a  vector.



# Data representation

- Data representation is one of the first things we must do in data mining
- What we can mine is largely determined by our data representation
- There is no one best data representation method for all data mining tasks.

# Representing text data: example

In what ways can we represent the following sentence?

The burger I ate was an awesome burger!

**Method 1:** By a list of words?

[“the”, “burger”, “i”, “ate”, “was”, “an”, “awesome”, “burger”]

**Method 2:** By the set of words?

{“the”, “burger”, “i”, “ate”, “was”, “an”, “awesome”}

**Method 3:** By a vector of word frequency?

(“the”:1, “burger”:2, “i”:1, “ate”:1, “was”:1, “an”:1, “awesome”:1)

**Method 4:** By a vector of letter frequency?

{‘a’: 4, ' ': 7, 'b': 2, 'e': 6, 'g': 2, 'i': 1, 'h': 1, 'm': 1, 'o': 1, 'n': 1, 's': 2, 'r': 4, 'u': 2, 't': 2, 'w': 2}