# Basic Word Representations

Natalie Parde
UIC CS 421

# How, then, should we represent the meaning of a word?

- Two classic strategies:
  - **Bag of words representations:** A word is a string of letters, or an index in a vocabulary list
  - **Logical representation:** A word defines its own meaning (“dog” = DOG)

# How, then, should we represent the meaning of a word?

- Two classic strategies:
  - **Bag of words representations:** A word is a string of letters, or an index in a vocabulary list
  - **Logical representation:** A word defines its own meaning ("dog" = DOG)

# Back to our discussion of vector semantics!

- Under the distributional hypothesis, we define a word by its **environment** or its **distribution** in language use
- This corresponds to the set of **contexts** in which the word occurs
  - **Context:** Neighboring words or grammatical environments
- **Two words with very similar sets of contexts (i.e., similar distributions) are assumed to have very similar meanings**

# We do this to infer meaning in the real world all the time.

- Pretend you don't know what the word *ongchoi* means
- However, you read the following sentences:
  - Ongchoi is delicious sautéed with garlic.
  - Ongchoi is superb over rice.
  - …ongchoi leaves with salty sauces…
- You've seen many of the other context words in these sentences previously:
  - …spinach sautéed with garlic over rice…
  - …chard stems and leaves are delicious…
  - …collard greens and other salty leafy greens…
- Your (correct!) conclusion?
  - Ongchoi is probably a leafy green similar to spinach, chard, or collard greens

**Our goal in NLP is to do the same thing computationally.**

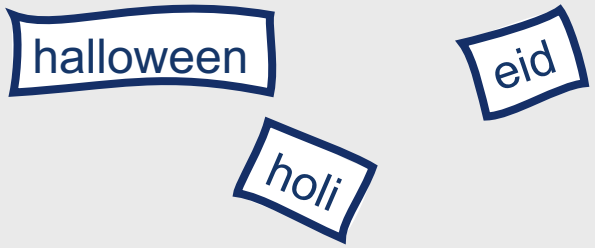
- How would we do this in the sample case from the previous slide?
  - Count the words in the context of *ongchoi*
  - See what other words occur in those same contexts

# We can represent a word's context using vectors.

- Define a word as a single vector point in an *n*-dimensional space
  - For bag of words representations, *n* = vocabulary size
- Represent the presence or absence of words in its surrounding context using numeric values
  - For bag of words representations, the value stored in a dimension *n* corresponds to the presence of a context word *c* in close proximity to the target word *w*

**The goal is for the values in these vector representations to correspond with dimensions of meaning.**

- Assuming this is the case, we should be able to:
  - Cluster vectors into semantic groups
  - Perform operations that are semantically intuitive

**The goal is for the values in these vector representations to correspond with dimensions of meaning.**

- Assuming this is the case, we should be able to:
  - Cluster vectors into semantic groups
  - Perform operations that are semantically intuitive