Dan Jurafsky and James Martin
Speech and Language Processing

Chapter 6:
Vector Semantics

But how to define a concept?

Classical ("Aristotelian") Theory of Concepts

The meaning of a word:

a concept defined by **necessary** and **sufficient** conditions

A **necessary** condition for being an X is a condition C that X must satisfy in order for it to be an X.

- If not C, then not X
- "Having four sides" is necessary to be a square.

A **sufficient** condition for being an X is condition such that if something satisfies condition C, then it must be an X.

- If and only if C, then X
- The following necessary conditions, jointly, are sufficient to be a square
  - x has (exactly) four sides
  - each of x's sides is straight
  - x is a closed figure
  - x lies in a plane
  - each of x's sides is equal in length to each of the others
  - each of x's interior angles is equal to the others (right angles)
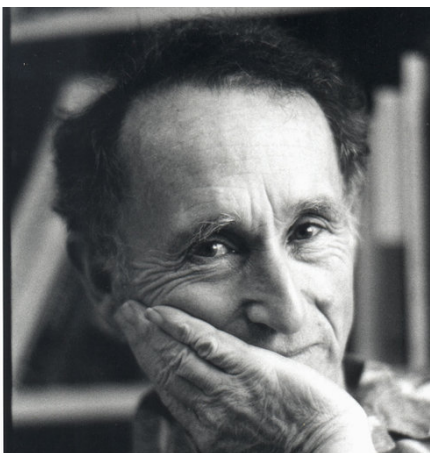 - the sides of x are joined at their ends

Example from Norman Swartz, SFU

Problem 1: The features are complex and may be context-dependent
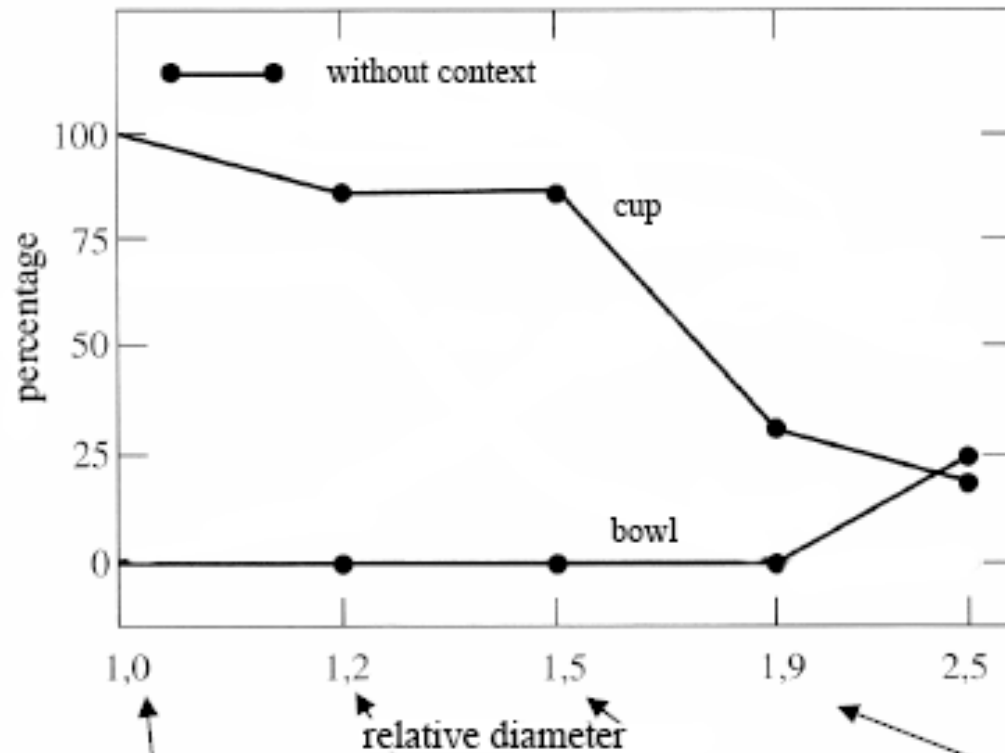
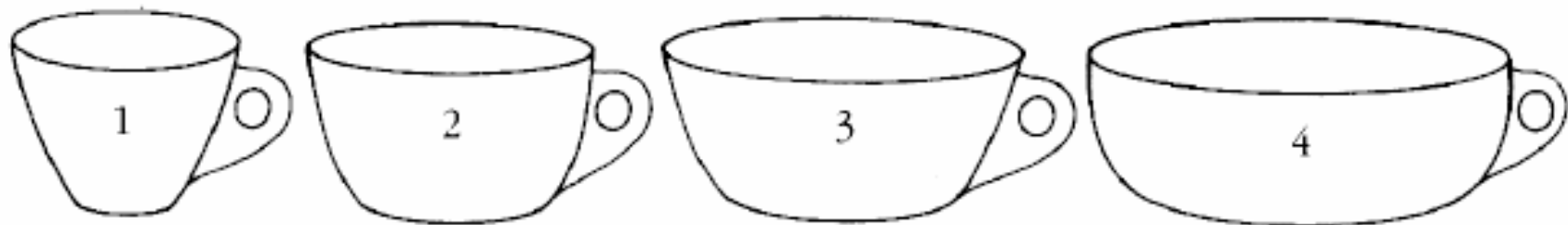William Labov. 1975

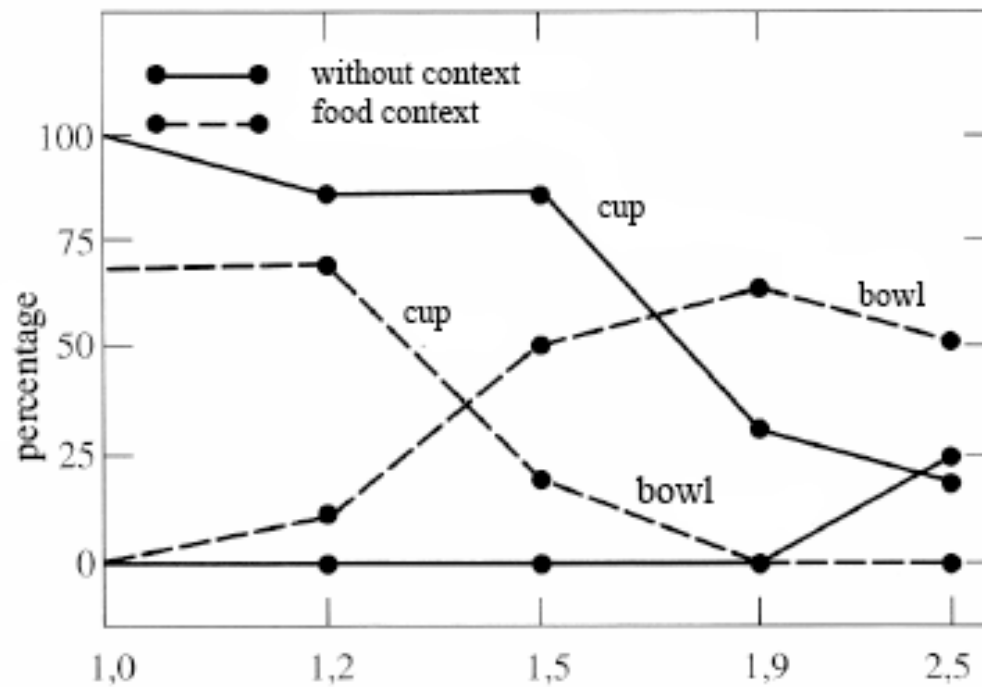What are these?

Cup or bowl?

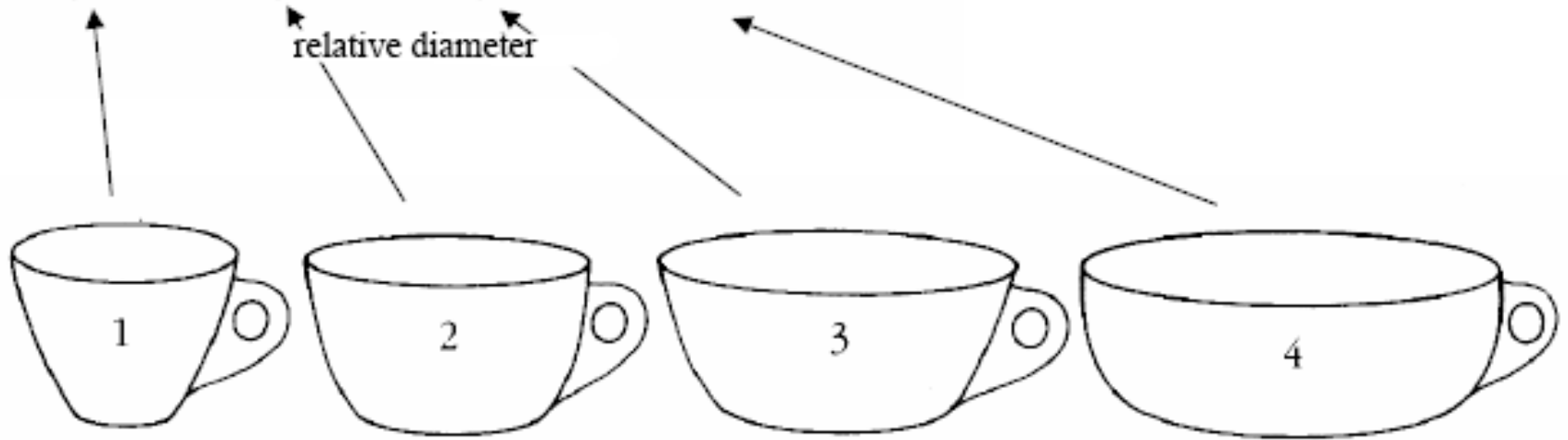# The category depends on complex features of the object (diameter, etc)

Where does the category „cup" end?

The category depends on the context! (If there is food in it, it's a bowl)

Boundaries between cups and bowls are context sensitive

Labov's definition of cup

The term *cup* is used to denote round containers with a ratio of depth to width of $1\pm r$ where $r \leq r_b$, and $r_b = \alpha_1 + \alpha_2 + \ldots \alpha_0$ and α_1 is a positive quality when the feature i is present and 0 otherwise.

feature 1 = with one handle
        2 = made of opaque vitreous material
        3 = used for consumption of food
        4 = used for the consumption of liquid food
        5 = used for consumption of hot liquid food
        6 = with a saucer
        7 = tapering
        8 = circular in cross-section

Cup is used variably to denote such containers with ratios width to depth $1\pm r$ where $r_b \leq r \leq r_1$ with a probability of $r_1 - r/r_t - r_b$. The quantity $1\pm r_b$ expresses the distance from the modal value of width to height.

Ludwig Wittgenstein (1889-1951)

Philosopher of language

In his late years, a proponent of studying "ordinary language"
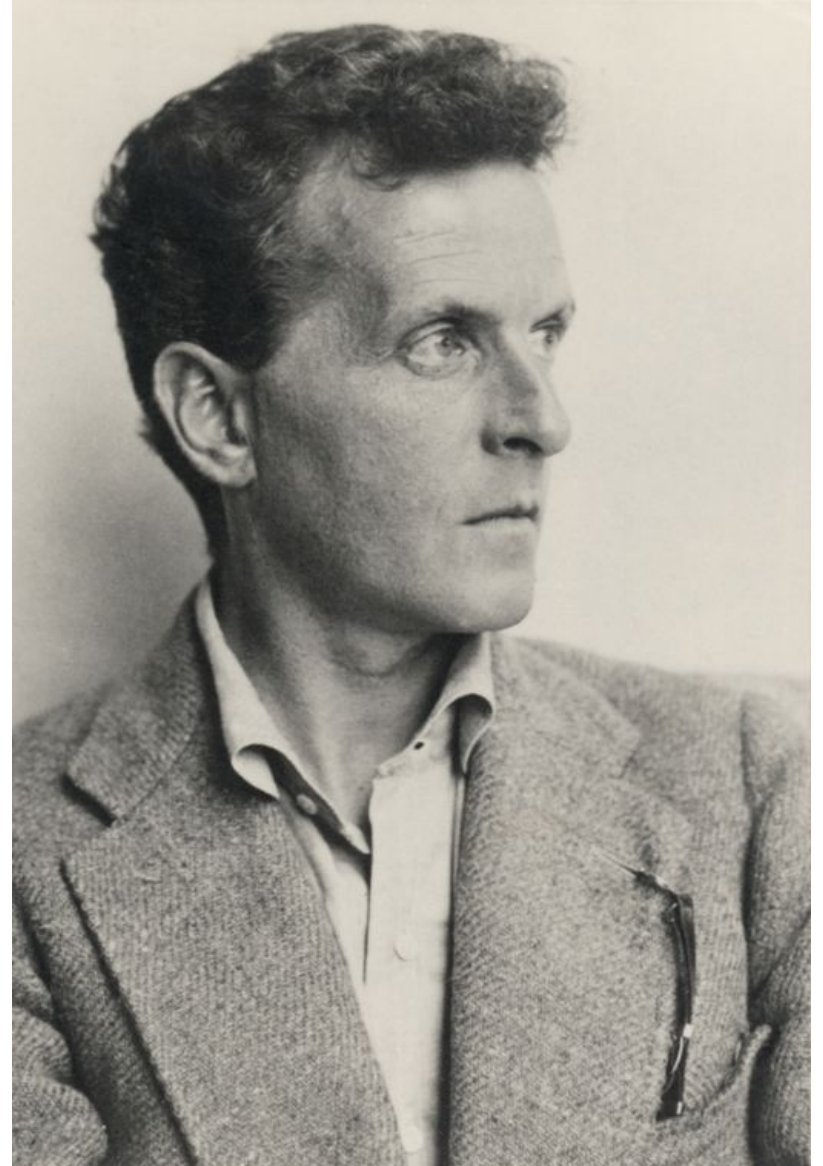
Wittgenstein (1945) *Philosophical Investigations.* Paragraphs 66,67

66. Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don't say: "There *must* be something common, or they would not be called 'games'"—but *look and see* whether there is anything common to all.—For if you look at them you will not see something that is common to *all*, but similarities, relationships, and a whole series of them at that. To repeat: don't think, but look!—Look for example at board-games, with their multifarious relationships. Now pass to card-games; here you find many correspondences with the first group, but many common features drop out, and others appear. When we pass next to ball-games, much that is common is retained, but much is lost.—Are they all 'amusing'? Compare chess with noughts and crosses. Or is there always winning and losing, or competition between players? Think of patience. In ball games there is winning and losing; but when a child throws his ball at the wall and catches it again, this feature has disappeared. Look at the parts played by skill and luck; and at the difference between skill in chess and skill in tennis. Think now of games like ring-a-ring-a-roses; here is the element of amusement, but how many other characteristic features have disappeared! And we can go through the many, many other groups of games in the same way; can see how similarities crop up and disappear.

And the result of this examination is: we see a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail.

67. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family: build, features, colour of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way.—And I shall say: 'games' form a family.

And for instance the kinds of number form a family in the same way. Why do we call something a "number"? Well, perhaps because it has a—direct—relationship with several things that have hitherto been called number; and this can be said to give it an indirect relationship to other things we call the same name. And we extend our concept of number as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres.

But if someone wished to say: "There is something common to all these constructions—namely the disjunction of all their common properties"—I should reply: Now you are only playing with words. One might as well say: "Something runs through the whole thread—namely the continuous overlapping of those fibres".

What is a game?

Wittgenstein's thought experiment on "What is a game":

PI #66:

"Don't say "there must be something common, or they would not be called `games'"—but *look and see* whether there is anything common to all"

Is it amusing?

Is there competition?

Is there long-term strategy?

Is skill required?

Must luck play a role?

Are there cards?

Is there a ball?

Family Resemblance

Game 1	Game 2	Game 3	Game 4
ABC	BCD	ACD	ABD

"each item has at least one, and probably several, elements in common with one or more items, but no, or few, elements are common to all items"   Rosch and Mervis

How about a radically different approach?

Ludwig Wittgenstein

PI #43:
  "The meaning of a word is its use in the language"

Let's define words by their usages

In particular, words are defined by their environments (the words around them)

Zellig Harris (1954): **If A and B have almost identical environments we say that they are synonyms**.

What does ongchoi mean?

Suppose you see these sentences:

- Ong choi is delicious **sautéed with garlic**.
- Ong choi is superb **over rice**
- Ong choi **leaves** with salty sauces

And you've also seen these:

- …spinach **sautéed with garlic over rice**
- Chard stems and **leaves** are **delicious**
- Collard greens and other **salty** leafy greens

Conclusion:

- ◦ Ongchoi is a leafy green like spinach, chard, or collard greens

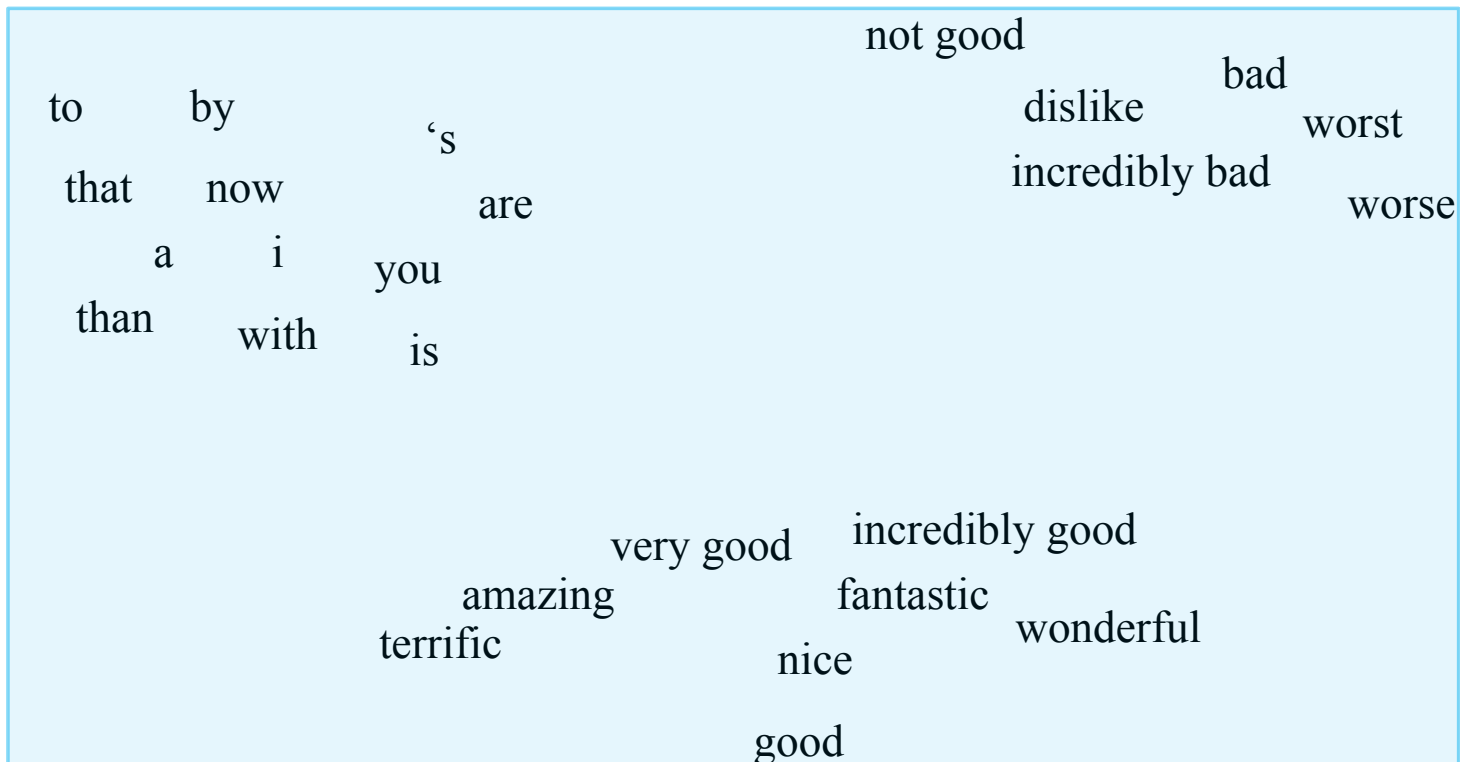# Ong choi: *Ipomoea aquatica "Water Spinach"*

We'll build a new model of meaning focusing on similarity

Each word = a vector
- Not just "word" or word45.

Similar words are "nearby in space"

not good

to   by

that   now   's

dislike   bad

worst

incredibly bad

worse

a   i   are

than   with   you

is

very good   incredibly good

amazing   fantastic

terrific   wonderful

nice

good

We define a word as a vector

Called an "embedding" because it's embedded into a space

The standard way to represent meaning in NLP

Fine-grained model of meaning for similarity
◦ NLP tasks like sentiment analysis
  ◦ With words, requires **same** word to be in training and test
  ◦ With embeddings: ok if **similar** words occurred!!!
◦ Question answering, conversational agents, etc