Lexical Semantics

Intro to NLP, CS585, Fall 2015

http://people.cs.umass.edu/~brenocon/inlp2015/

Brendan O'Connor

- Word-based features in supervised ML: fundamentally, too much data sparsity
- Idea: we want a database of word meanings to help analyze texts. How to represent word meanings?
    - Today: word senses and taxonomies
    - Next week: context vectors (distributional semantics)

Word senses (concepts)

- A single word (word form) can have different *senses* or word meanings.
  - Coarse ambiguity: POS or proper-vs-common noun
  - Finer-grained ambiguity:
    - I went to my <u>bank</u> today.
    - I saw the <u>bank</u> of a river.

Computing with word senses

- If we could disambiguate word senses
  - Could link to a knowledge base of entities or concepts (e.g. "Clinton")
  - Could do better synonym/hypernym analysis (e.g. "find documents talking about female politicians")
- Questions
  - 1. Where do the word senses come from? (Today: pre-specified knowledge base)
  - 2. What is this good for?
  - 3. How to disambiguate?

Word sense relations

Word sense relations

- Binary relations between word senses explain how to generalize or relate their meanings.

Word sense relations

- Binary relations between word senses explain how to generalize or relate their meanings.
- **Synonymy**: mean the same thing
  - I drank cocoa <==> I drank hot chocolate
    - *set equivalence; bidirectional entailment under substitution*

Word sense relations

- Binary relations between word senses explain how to generalize or relate their meanings.
- **Synonymy**: mean the same thing
  - I drank cocoa <==> I drank hot chocolate
    - *set equivalence; bidirectional entailment under substitution*
- **Hypernymy** (IS-A): more general (supersets)
  - financial institution is a hypernym of bank
  - bank is a hyponym of financial institution
  - I went to my bank ==> I went to my financial institution
    - *supersets; directional entailment under substitution*

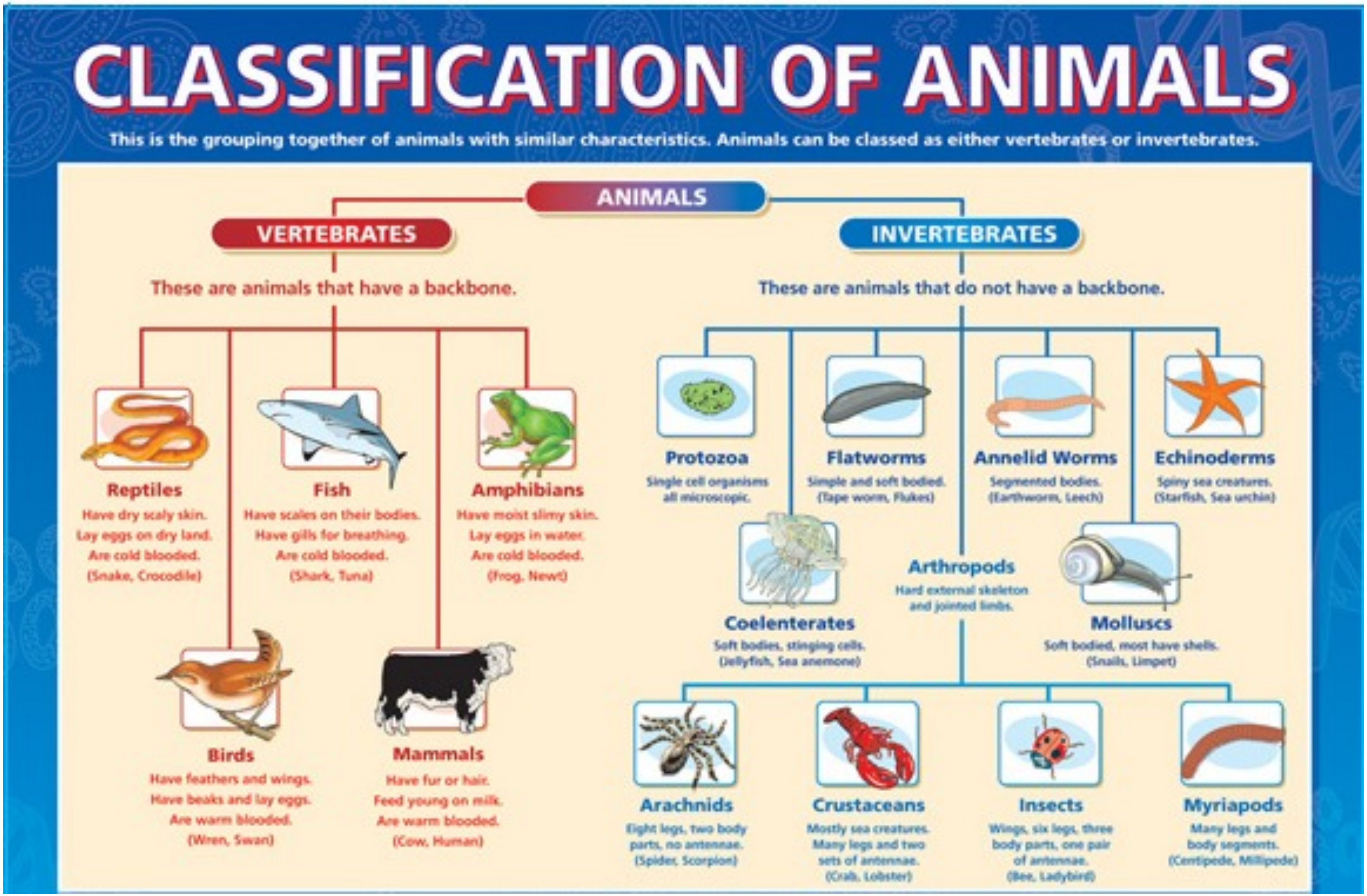
Word sense relations

- Binary relations between word senses explain how to generalize or relate their meanings.
- **Synonymy**: mean the same thing
  - I drank cocoa <==> I drank hot chocolate
    - *set equivalence; bidirectional entailment under substitution*
- **Hypernymy** (IS-A): more general (supersets)
  - financial institution is a hypernym of bank
  - bank is a hyponym of financial institution
  - I went to my bank ==> I went to my financial institution
    - *supersets; directional entailment under substitution*
- These relations hold between *senses* (not words)
  - I saw the river bank =/=> I saw the river financial institution
  - I saw bill today =/=> I saw check

Word sense relations

- Binary relations between word senses explain how to generalize or relate their meanings.
- **Synonymy**: mean the same thing
    - I drank <u>cocoa</u> <==> I drank <u>hot chocolate</u>
        - *set equivalence; bidirectional entailment under substitution*
- **Hypernymy** (IS-A): more general (supersets)
    - <u>financial institution</u> is a hypernym of <u>bank</u>
    - <u>bank</u> is a hyponym of <u>financial institution</u>
    - I went to my <u>bank</u> ==> I went to my <u>financial institution</u>
        - *supersets; directional entailment under substitution*
- These relations hold between *senses* (not words)
    - I saw the river <u>bank</u> =/=> I saw the river <u>financial institution</u>
    - I saw <u>bill</u> today =/=> I saw <u>check</u>
- *<u>Taxonomy</u>*: a hypernym graph

Taxonomies



CLASSIFICATION OF ANIMALS

This is the grouping together of animals with similar characteristics. Animals can be classed as either vertebrates or invertebrates.

ANIMALS

VERTEBRATES

These are animals that have a backbone.

Reptiles
Have dry scaly skin.
Lay eggs on dry land.
Are cold blooded.
(Snake, Crocodile)

Fish
Have scales on their bodies.
Have gills for breathing.
Are cold blooded.
(Shark, Tuna)

Amphibians
Have moist slimy skin.
Lay eggs in water.
Are cold blooded.
(Frog, Newt)

Birds
Have feathers and wings.
Have beaks and lay eggs.
Are warm blooded.
(Wren, Swan)

Mammals
Have fur or hair.
Feed young on milk.
Are warm blooded.
(Cow, Human)

INVERTEBRATES

These are animals that do not have a backbone.

Protozoa
Single cell organisms
all microscopic.

Flatworms
Simple and soft bodied.
(Tape worm, Flukes)

Annelid Worms
Segmented bodies.
(Earthworm, Leech)

Echinoderms
Spiny sea creatures.
(Starfish, Sea urchin)

Coelenterates
Soft bodies, stinging cells.
(Jellyfish, Sea anemone)

Arthropods
Hard external skeleton
and jointed limbs.

Molluscs
Soft bodied, most have shells.
(Snails, Limpet)

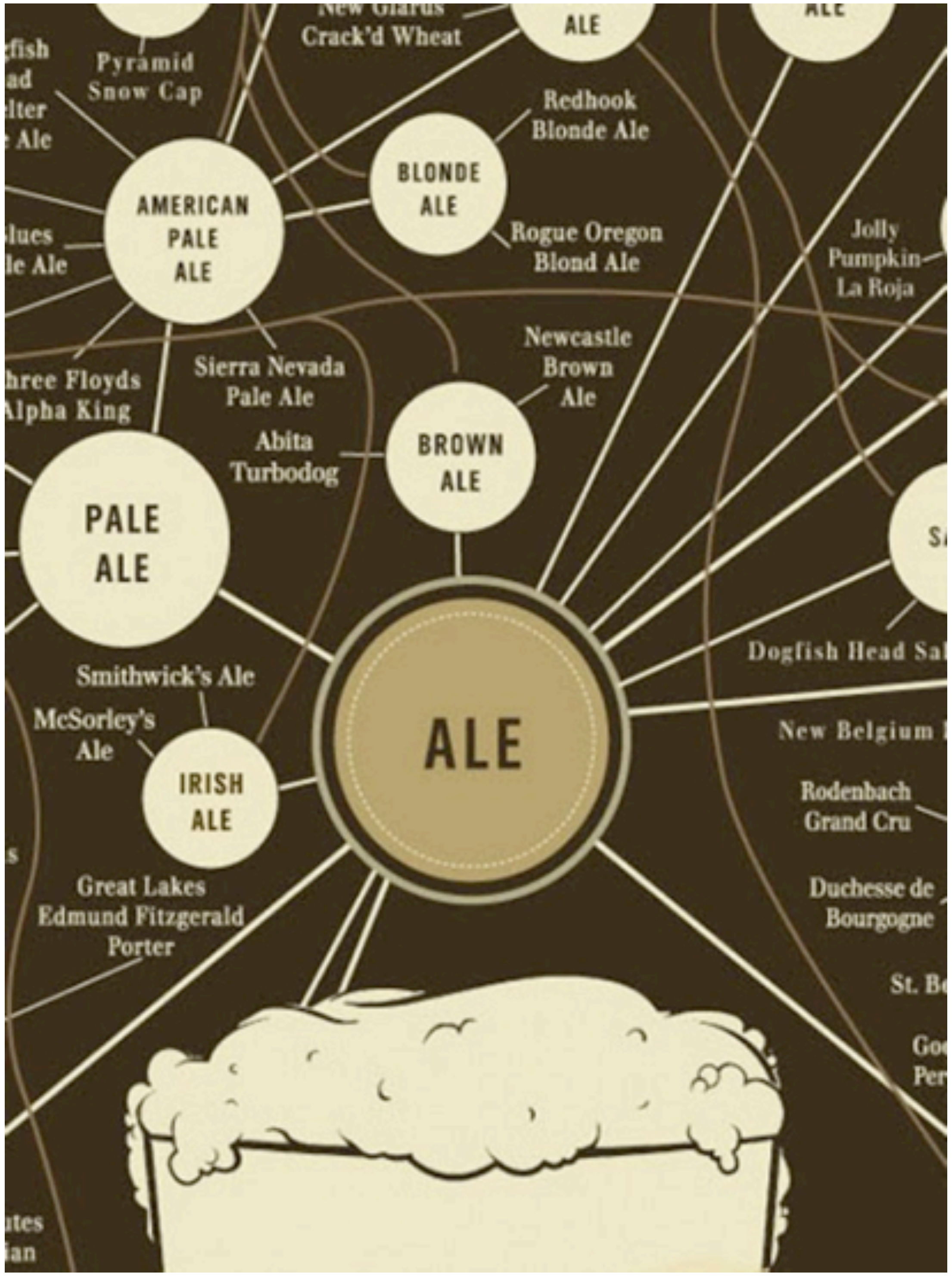
Arachnids
Eight legs, two body
parts, no antennae.
(Spider, Scorpion)

Crustaceans
Mostly sea creatures.
Many legs and two
sets of antennae.
(Crab, Lobster)

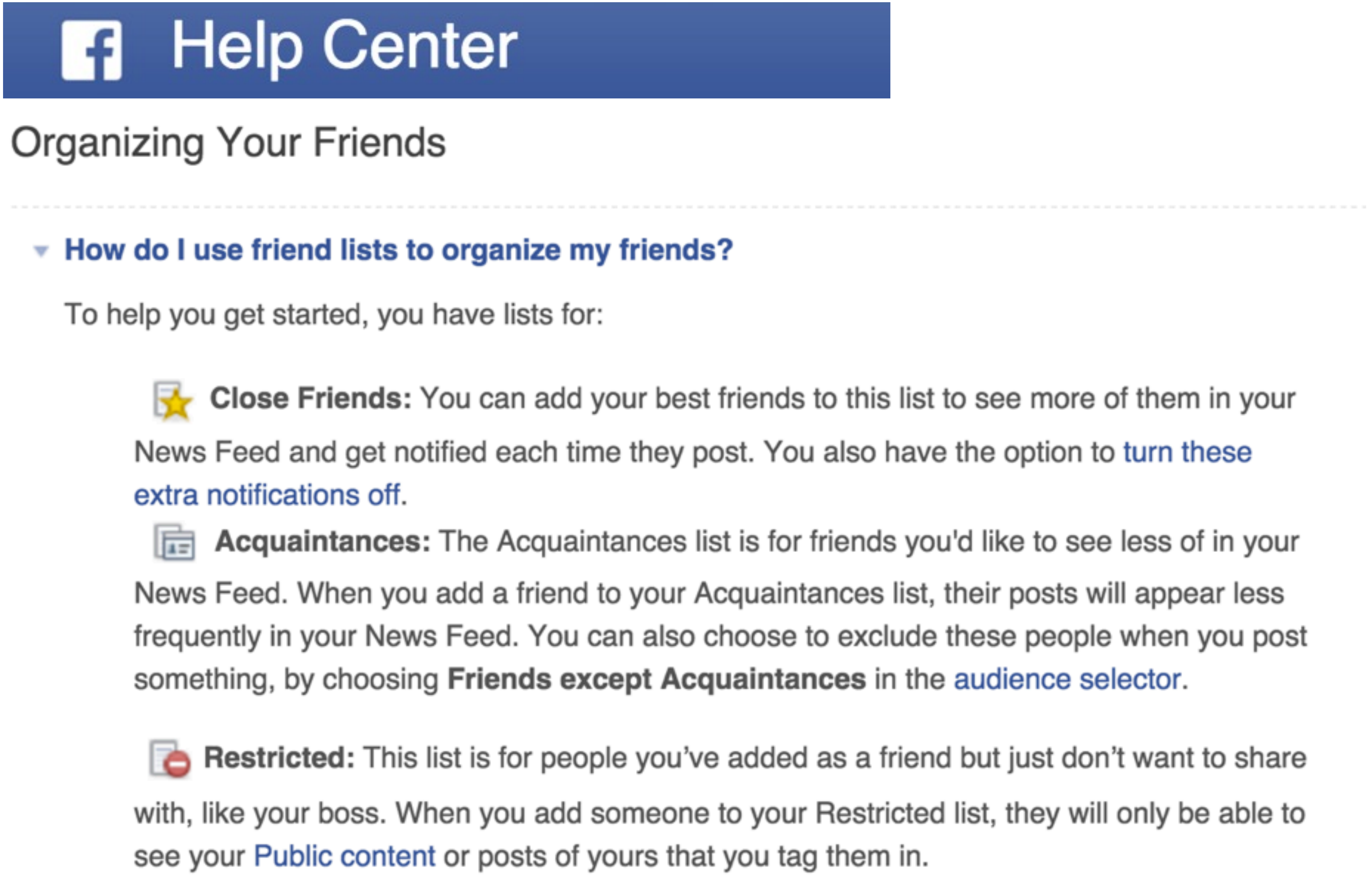
Insects
Wings, six legs, three
body parts, one pair
of antennae.
(Bee, Ladybird)

Myriapods
Many legs and
body segments.
(Centipede, Millipede)

Taxonomies


New Glarus
Crack'd Wheat
ALE
ALE
fish
ad
lter
Ale
Pyramid
Snow Cap
AMERICAN
PALE
ALE
Redhook
Blonde Ale
BLONDE
ALE
Rogue Oregon
Blond Ale
Jolly
Pumpkin
La Roja
lues
le Ale
Newcastle
Brown
Ale
hree Floyds
Alpha King
Sierra Nevada
Pale Ale
Abita
Turbodog
BROWN
ALE
PALE
ALE
S
Dogfish Head Sa
Smithwick's Ale
McSorley's
Ale
IRISH
ALE
ALE
New Belgium
Rodenbach
Grand Cru
Great Lakes
Edmund Fitzgerald
Porter
Duchesse de
Bourgogne
St. B
Go
Per
tes
ian

Taxonomies



Help Center

Organizing Your Friends

How do I use friend lists to organize my friends?

To help you get started, you have lists for:

Close Friends: You can add your best friends to this list to see more of them in your News Feed and get notified each time they post. You also have the option to turn these extra notifications off.

Acquaintances: The Acquaintances list is for friends you'd like to see less of in your News Feed. When you add a friend to your Acquaintances list, their posts will appear less frequently in your News Feed. You can also choose to exclude these people when you post something, by choosing **Friends except Acquaintances** in the audience selector.

Restricted: This list is for people you've added as a friend but just don't want to share with, like your boss. When you add someone to your Restricted list, they will only be able to see your Public content or posts of yours that you tag them in.

Lexical knowledge bases

Lexical knowledge bases

- “KB”=knowledge base. Database of <u>concepts</u>. Each has
  - Words associated with it
  - (possibly) Relationships to other concepts, e.g. taxonomy

Lexical knowledge bases

- "KB"=knowledge base. Database of concepts. Each has
  - Words associated with it
  - (possibly) Relationships to other concepts, e.g. taxonomy
- Very simple KB: an entity list
  - list of world countries, with multiple possible names for each ("USA", "United States", "United States of America")
  - list of all baseball teams ... baseball players ...
  - others?

Lexical knowledge bases

Call my main squeeze.

In cars, Google's Android Auto can remember your pet names for your significant other.

http://time.com/google-now/

Personalized KBs? e.g. your phone's contacts

Lexical knowledge bases

Lexical knowledge bases

- Two popular general-domain, manually/semiauto. curated KBs:
  - Freebase: KB of entities (mostly), heavily derived from Wikipedia https://www.freebase.com/
    - (google is now killing freebase. see also DBPedia)
  - Wordnet: (common) nouns, adjective, adverbs http://wordnetweb.princeton.edu/perl/webwn

Lexical knowledge bases

- Two popular general-domain, manually/semiauto. curated KBs:
  - Freebase: KB of entities (mostly), heavily derived from Wikipedia https://www.freebase.com/
    - (google is now killing freebase. see also DBPedia)
  - Wordnet: (common) nouns, adjective, adverbs http://wordnetweb.princeton.edu/perl/webwn
- Zillions of similar proprietary databases

Wordnet

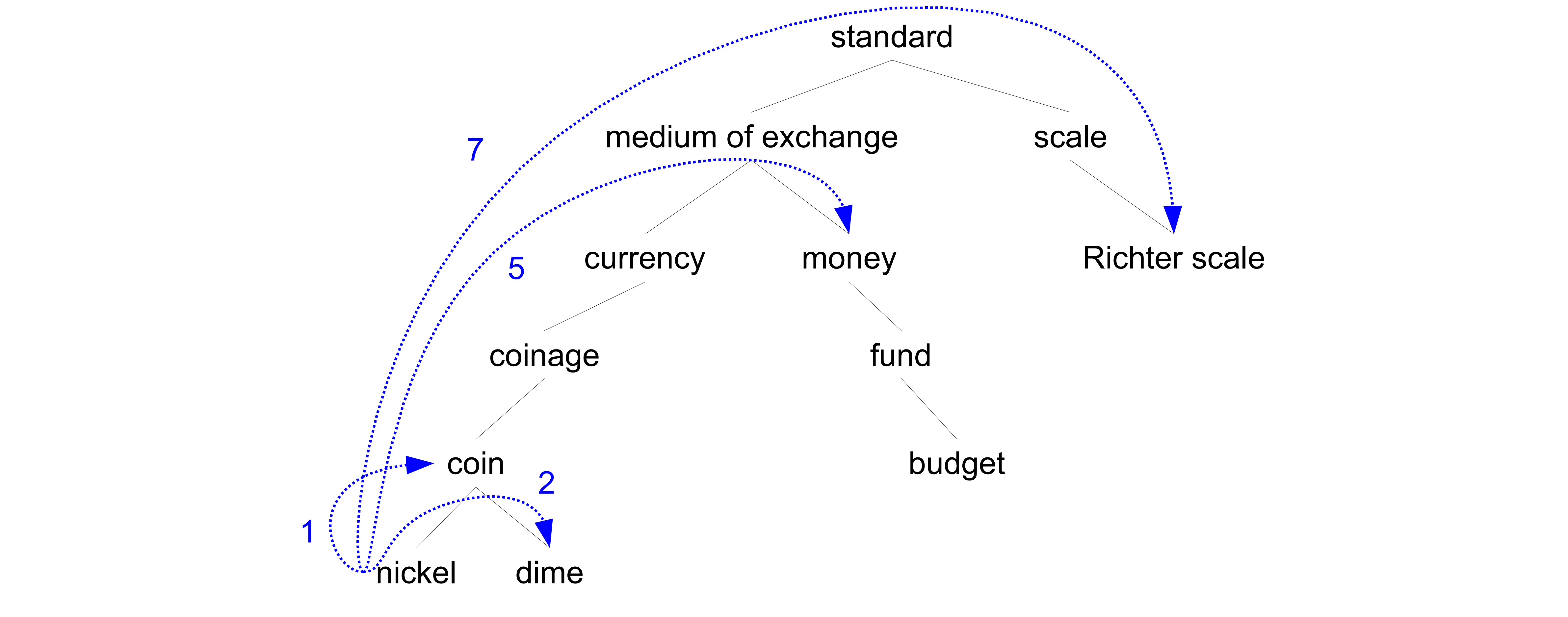


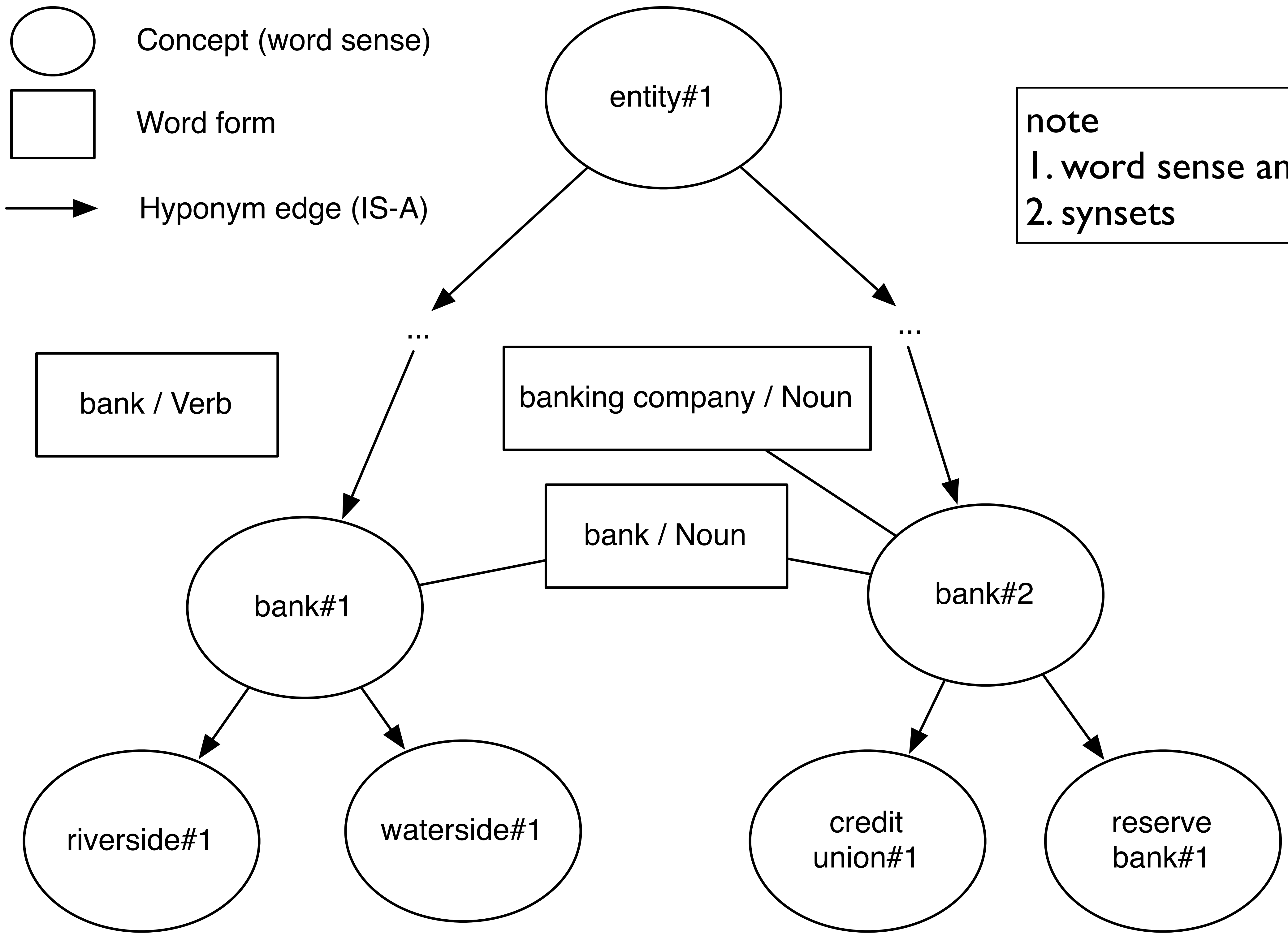
Figure 20.6 A fragment of the WordNet hypernym hierarchy, showing path lengths from *nickel* to *coin* (1), *dime* (2), *money* (5), and *Richter scale* (7).

[Jurafsky & Martin]

Wordnet http://wordnetweb.princeton.edu/perl/webwn

- Hand-curated database of word senses for English
  - nouns, adjective, adverbs
    (has verbs but problematic)
- Each concept (“synset”) has
  - A set of words it corresponds to.
    Each is: (lemma,POS) pair
  - Hypernym relations to other concepts
  - Synonymy among words sharing a synset.

Wordnet: concepts vs words



note

1. word sense ambiguity
2. synsets

Wordnet

POS	**#**
Noun	117,097
Adjective	22,141
Verb	11,488
Adverb	4,601

http://wordnetweb.princeton.edu/perl/webwn

Hyponyms of “person” in WN

7588 total -- with MFS restriction. *[from Michael Heilman]*

- vintager
- matrisib
- horseback rider
- ceo
- seeker
- fieldhand
- radiologist
- captain
- moujik
- research director
- damsel
- nibbler
- nailer
- nude person
- seismologist
- oddball
- prankster
- radiotherapist
- nebraskan
- cupbearer
- psychic
- accompanist
- plagiariser
- timberman
- photographer's model
- lombard
- debaser
- courtier
- dutch uncle
- schlemiel
- dizygotic twin
- mental case
- matriarch
- vocalist
- internist
- transplanter
- techie
- sniffler
- marrano
- first baseman
- government man
- child prodigy
- athenian
- hospital chaplain
- dominatrix
- bibliopole
- hombre
- east indian
- ballet master
- bad person
- rock 'n' roll musician
- flack catcher
- telephoner
- dominus
- cheater
- groveler
- accomplice
- herb doctor
- schoolfriend
- preteen
- gastronome
- concierge
- shogun
- flutist
- bottom dog
- imperialist
- emir
- libeler
- manichaean
- abnegator
- cousin-german
- masorite
- trouble maker
- villainess
- rajpoot
- calapooya
- overlord
- bank guard
- tumbler
- polycarp
- radiographer
- slave owner
- stick-in-the-mud
- audile
- deadbeat
- maltman
- jeweler

Is Wordnet useful?

Is Wordnet useful?

- Going beyond individual words to more general meanings (combat data sparsity)
  - Synonym expansion
  - Derive sets of terms for specific senses
  - Word similarity as path distance (Resnik similarity)

Is Wordnet useful?

- Going beyond individual words to more general meanings (combat data sparsity)
  - Synonym expansion
  - Derive sets of terms for specific senses
  - Word similarity as path distance (Resnik similarity)
- The platonic ideal of WN is right. What about WN itself?
  - WN's sense inventory is too ambitious / fine-grained
  - Coverage is often problematic
  - General-domain knowledge bases are very hard to design and make! (Knowledge/ontology engineering is a whole discipline.
    e.g. library scientists are trained in taxonomy design.)

Is Wordnet useful?

- Going beyond individual words to more general meanings (combat data sparsity)
    - Synonym expansion
    - Derive sets of terms for specific senses
    - Word similarity as path distance (Resnik similarity)
- The platonic ideal of WN is right. What about WN itself?
    - WN's sense inventory is too ambitious / fine-grained
    - Coverage is often problematic
    - General-domain knowledge bases are very hard to design and make! (Knowledge/ontology engineering is a whole discipline. e.g. library scientists are trained in taxonomy design.)
- If your task has lots of training data, WN typically is not helpful (e.g. has failed to help MT), though clearly helps in low-data cases

Is Wordnet useful?

- Going beyond individual words to more general meanings (combat data sparsity)
  - Synonym expansion
  - Derive sets of terms for specific senses
  - Word similarity as path distance (Resnik similarity)
- The platonic ideal of WN is right. What about WN itself?
  - WN's sense inventory is too ambitious / fine-grained
  - Coverage is often problematic
  - General-domain knowledge bases are very hard to design and make! (Knowledge/ontology engineering is a whole discipline. e.g. library scientists are trained in taxonomy design.)
- If your task has lots of training data, WN typically is not helpful (e.g. has failed to help MT), though clearly helps in low-data cases
- Entity-centric databases came next -- less epistemically ambitious but more immediately useful (Freebase, DBpedia)

Word sense disambiguation

- Given KB and text:
  Want to tag spans in text with concept IDs
- Disambiguation problem
  - "I saw the bank" => bank#1 or bank#2?
  - "Michael Jordan" => ?

Word sense disambiguation

- Given KB and text:
  Want to tag spans in text with concept IDs
- Disambiguation problem
  - "I saw the bank" => bank#1 or bank#2?
  - "Michael Jordan" => ?

- Many terms for this: concept tagging, entity linking, "wikification", WSD

Word sense disambiguation

- Supervised setting: ground-truth concept IDs for words in text
- Given candidate concepts for a word: similar problem as POS disambiguation, named entity type disambiguation, etc.
- Contextual features
  - Word immediately to left ... to right ...
  - Word within 10 word window (20 word window? entire document?)
- Features from matching a concept description, if your KB has one
  - *Michael Jeffrey Jordan (born February 17, 1963), also known by his initials, MJ,[1] is an American former professional basketball player. He is also a businessman, and principal owner and chairman of the Charlotte Hornets. Jordan played 15 seasons in the National Basketball Association (NBA) for theChicago Bulls and Washington Wizards.*
- Most frequent sense baseline
  - For WN, hard to beat (?!)
- Contrast to <u>distributional semantics</u>:
  unsupervised learning of word meanings