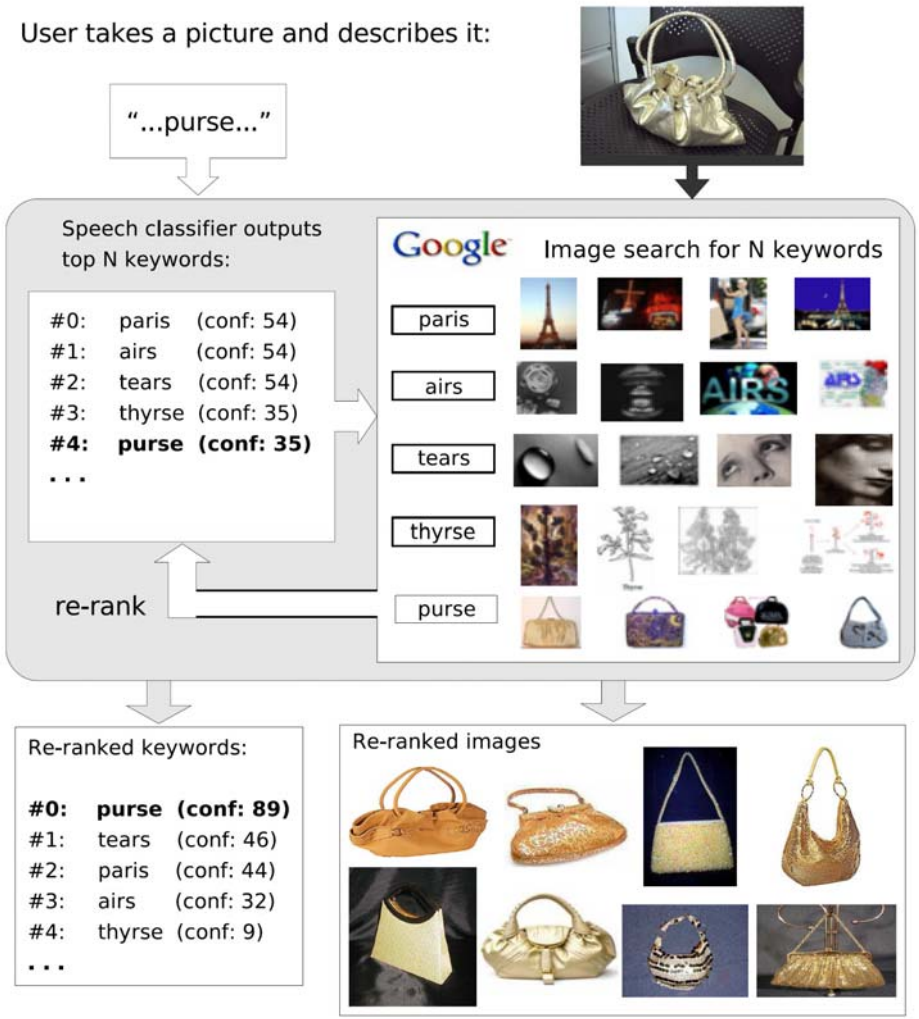
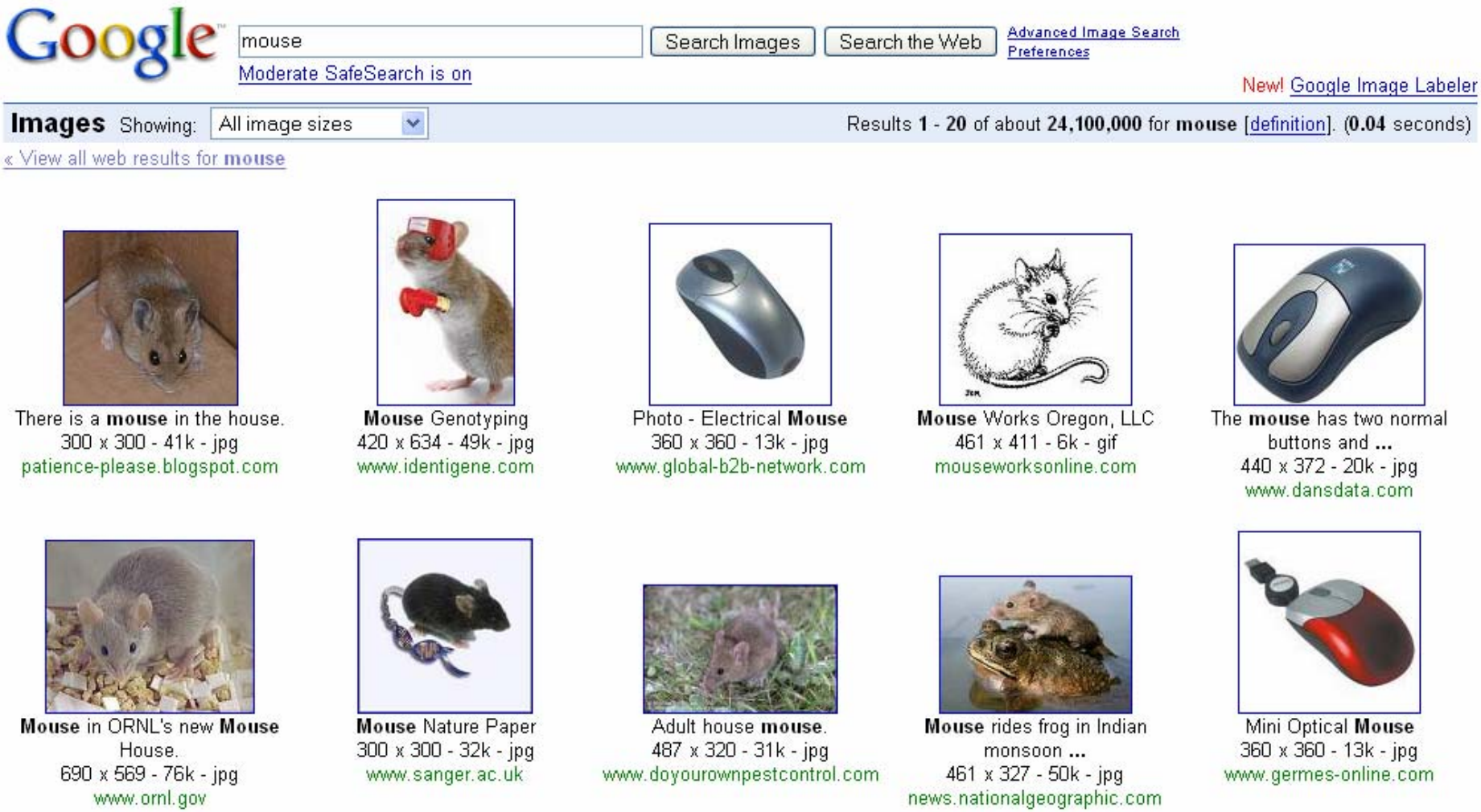
Dictionary-based Visual Sense Models

Kate Saenko and Trevor Darrell

Goal: Open Object Vocabulary



Problem: visual polysemy



Sources of visual polysemy

Hurricane, tornado watch

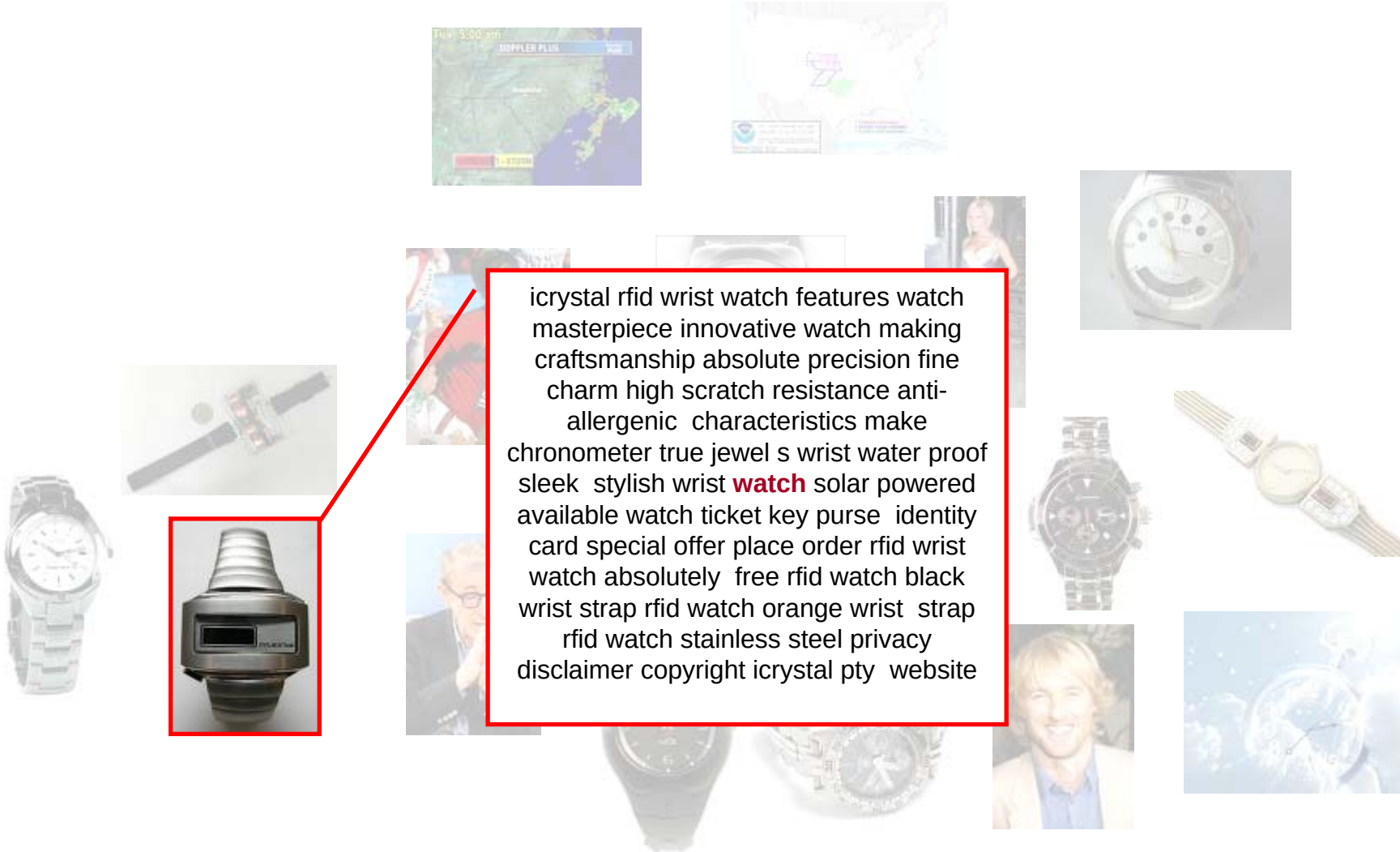
Celebrity watch

Watch out!

Would rather watch…

Suicide watch

Text context provides sense cues



icrystal rfid wrist watch features watch masterpiece innovative watch making craftsmanship absolute precision fine charm high scratch resistance anti-allergenic characteristics make chronometer true jewel s wrist water proof sleek stylish wrist **watch** solar powered available watch ticket key purse identity card special offer place order rfid wrist watch absolutely free rfid watch black wrist strap rfid watch orange wrist strap rfid watch stainless steel privacy disclaimer copyright icrystal pty website

Comparison to previous approaches to web-based object categorization

- Some learn only from image features
  - Fei-Fei Li et. al. bootstrap from labeled images
  - Fergus et. al. rely on first page of results as validation set to select correct cluster
- Some also incorporate text features
  - Schroff et. al. use a category-independent text classifier to re-rank images
  - Berg et. al. discover text topics and ask user to sort them into positive and negative
- **None address polysemy directly**
- **All rely on labeled images of correct sense**

Solution: use online dictionary

WORDNET: Noun

- S: (n) **watch**, ticker (a small portable timepiece)
- S: (n) **watch** (a period of time (4 or 2 hours) during which some of a ship's crew are on duty)
- S: (n) **watch**, vigil (a purposeful surveillance to guard or observe)
- S: (n) **watch** (the period during which someone (especially a guard) is on duty)
- S: (n) lookout, lookout man, sentinel, sentry, **watch**, spotter, scout, picket (a person employed to keep watch for some anticipated event)
- S: (n) vigil, **watch** (the rite of staying awake for devotional purposes (especially on the eve of a religious festival))

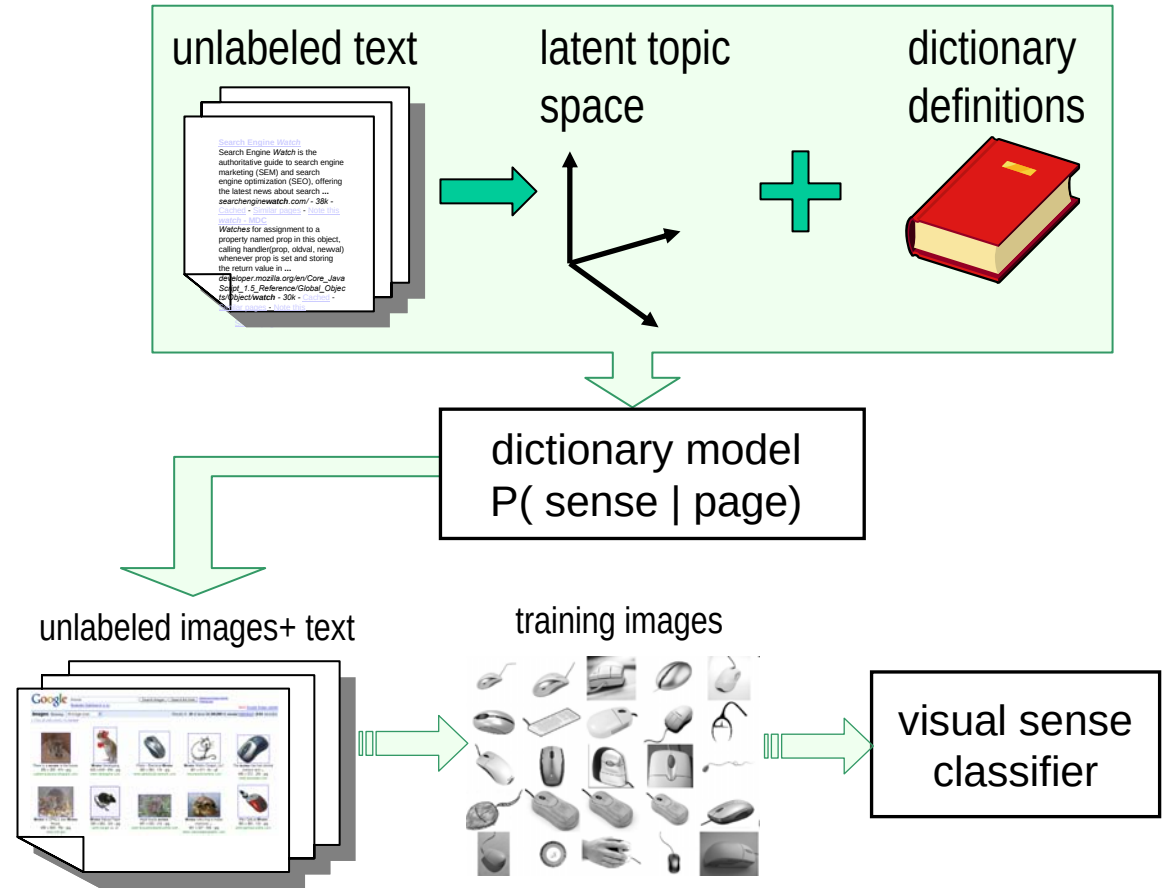
WIKIPEDIA: Watch may also refer to:

- Watch system, a period of work duty
- Tropical cyclone warnings and watches, alerts issued to coastal areas threatened by severe storms
- Watch (Unix), a Unix command
- Watch (TV channel) a TV station launching in Autumn 2008
- Watch (computer programming)
- Help:Watching pages on Wikipedia
- Watch (dog), name of the pet dog in the the Boxcar Children

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. ACL, 1995.

Overview of approach

- Given a word, learn a probabilistic model of each sense as defined by dictionary
- Use text model to construct sense-specific image classifiers
- Assumptions: only noun entries, one sense per image



Dictionary model

- Use sense entry text to learn a probability distribution over words for that sense
- Problem: entries contain very little text
  - Expand by adding synonyms, hyponyms, 1st-level hypernyms
  - Still, very few words!

•S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
•*direct hyponym* / *full hyponym*
  •S: (n) house mouse, Mus musculus (brownish-grey Old World mouse now a common household pest worldwide)
  •S: (n) harvest mouse, Micromyx minutus (small reddish-brown Eurasian mouse inhabiting e.g. cornfields)
  •S: (n) field mouse, fieldmouse (any nocturnal Old World mouse of the genus Apodemus inhabiting woods and fields and gardens)
  •S: (n) nude mouse (a mouse with a genetic defect that prevents them from growing hair and also prevents them from immunologically rejecting human cells and tissues; widely used in preclinical trials)
  •S: (n) wood mouse (any of various New World woodland mice)
•***direct hypernym*** / *inherited hypernym* / *sister term*
  •S: (n) rodent, gnawer (relatively small placental mammals having a single pair of constantly growing incisor teeth specialized for gnawing)

Dictionary model

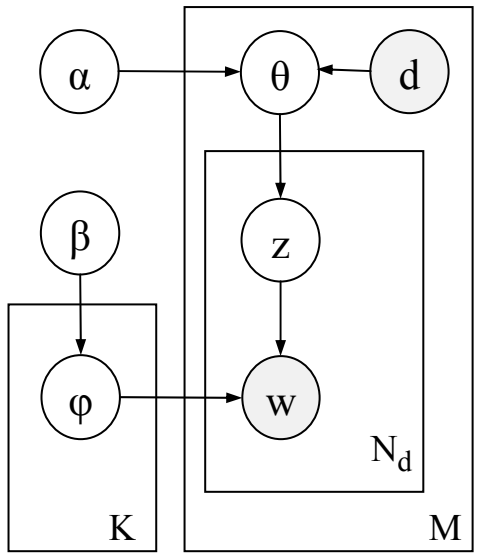
- Use sense entry text to learn a probability distribution over words for that sense
- Problem: entries contain very little text
  - Expand by adding synonyms, hyponyms, 1st-level hypernyms
  - Still, very few words!

Idea: use large collection of unlabeled text to learn hidden topics which align with different senses/uses of the word

Latent Dirichlet allocation (Blei et. al. '03)

- One of several techniques for discovering latent dimensions in bag-of-words data

- Given:
  - M documents, N_d words each, and parameters α and β
- Generate each document as follows:
  - For each topic $j=1,\ldots,K$, sample φ^j
  - For each document d, sample θ_d
  - For each word i, sample topic z_i from θ_d and then choose the word w_i from φ^{z_i}



Example: latent text space

- Example of 8 discovered latent topics for “watch” from Yahoo web search data (15 most likely stemmed words are shown)

1. price band winder box jewelri ship leather order strap item sale shop store gift case
2. time movement dial clock mechan hand quartz seiko design chronograph timepiec collect
3. pocket gold diamond antiqu silver fine ladi steel circa sold vintag dial ring jewelri face
4. rolex servic repair batteri omega replica men tag heuer breitl swiss replac gucci button test
5. new view ad comment video http episod var opera movi game download onlin music dai
6. updat new design post site video home plai inform axcent health link center relat cola
7. new world media right sai said hous april obama islam march bush war american time
8. us dai make time just look like need year great want good run case know

Dictionary model using latent space

- Given query word with sense ***s*** in set ***{1, … ,S},*** and a text document $\boldsymbol{d^t}$, compute probability of each sense as

$$P(s|d^t) = \sum_{j=1}^{K} P(s|z=j)P(z=j|d^t).$$

Dictionary model using latent space

- Given query word with sense ***s*** in set ***{1, … ,S},*** and a text document $\boldsymbol{d^t}$, compute probability of each sense as

$$P(s|d^t) = \sum_{j=1}^{K} P(s|z=j)P(z=j|d^t).$$

- Define the likelihood of sense ***s*** with entry $\mathbf{e_s = w_1, \ldots, w_{Es}}$ given topic ***j*** as

$$P(s|z=j) \equiv \frac{1}{E_s} \sum_{i=1}^{E_s} P(w_i|z=j),$$

Dictionary model using latent space

- Given query word with sense ***s*** in set ***{1, … ,S},*** and a text document d^t, compute probability of each sense as

$$P(s|d^t) = \sum_{j=1}^{K} P(s|z=j)P(z=j|d^t).$$

- Define the likelihood of sense ***s*** with entry $\mathbf{e_s = w_1, \ldots, w_{Es}}$ given topic ***j*** as

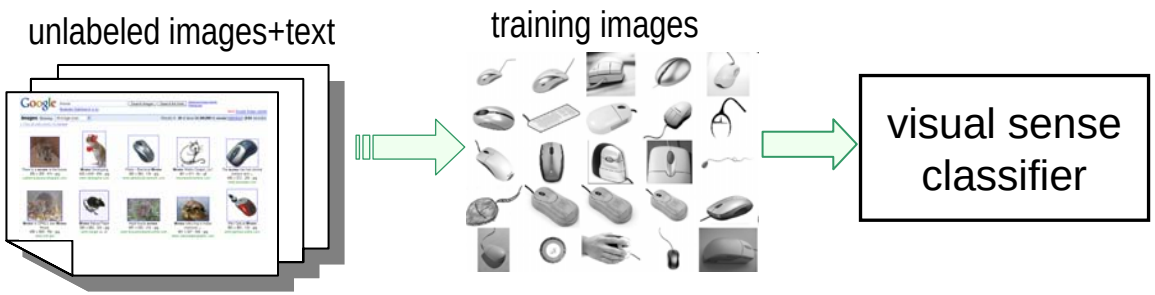
$$P(s|z=j) \equiv \frac{1}{E_s} \sum_{i=1}^{E_s} P(w_i|z=j),$$

- Marginalize across words to compute distribution of topics in d^t :

$$P(z=j|d^t) = \sum_{i=1}^{D} P(z=j|w_i) = \sum_{i=1}^{D} \frac{P(w_i|z=j)P(z=j)}{P(w_i)}$$

Visual sense classifier

- Use dictionary model to generate labeled training data for visual classifier
- Choice of classifier not crucial (we use SVM as it has yielded state-of-the-art image classification results)
- For each sense,
  - Use $P(s|d^t)$ to select N highest-ranked image-text pairs, use the images as positive training data
  - Negative data is drawn from background class



Baseline approach: sense-specific search term generation

- Automatically generate sense-specific keywords from Wordnet entries
  - Limit queries to 3 terms
  - Append word to synonyms and direct hypernyms
  - E.g. mouse + computer mouse, mouse + electronic device
- Result:

bass: bass+percoid, bass+percoid+fish, bass+percoidean, freshwater+bass
face: face+external+body+part, human+face
mouse: computer+mouse, mouse+electronic+device
speaker: loudspeaker, loudspeaker+system, speaker+electro-acoustic+transducer, speaker+unit
watch: ticker, watch+horologe, watch+timekeeper, watch+timepiece

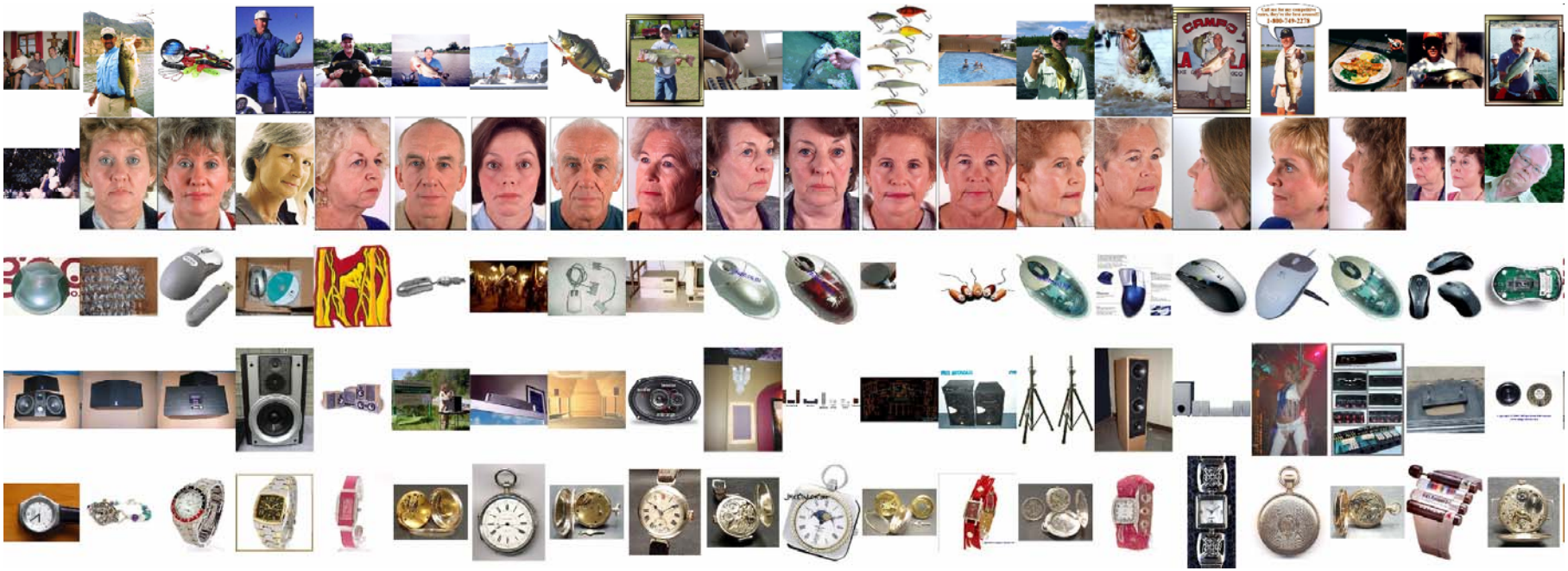
Dataset

- Five polysemous words: *bass, face, mouse, speaker, watch*
- For each word, collected three unlabeled datasets:
  - Yahoo image search using given word
  - Yahoo image search using sense-specific terms
  - Yahoo text search using given word
- For evaluation, human annotator labeled the presense/absense of the following *target* senses:
  - Bass-fish, face-human, mouse-computer, speaker-device, watch-timepiece

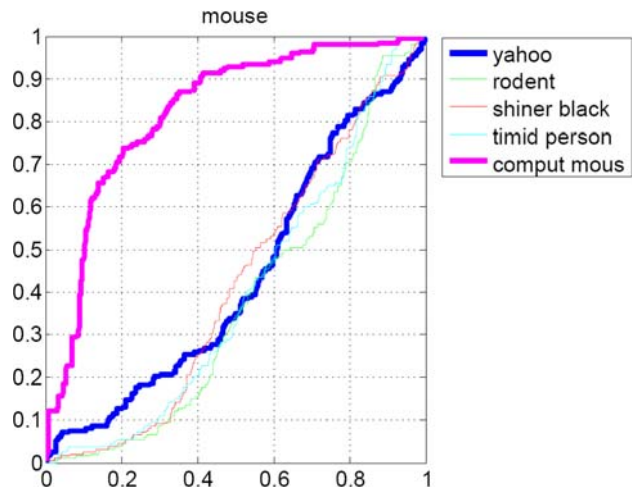
category	size of datasets			distribution of labels in the keyword dataset	
	text-only	sense term	keyword	positive (good)	negative (partial, unrelated)
Bass	984	357	678	146	532
Face	961	798	756	130	626
Mouse	987	726	768	198	570
Speaker	984	2270	660	235	425
Watch	936	2373	777	512	265

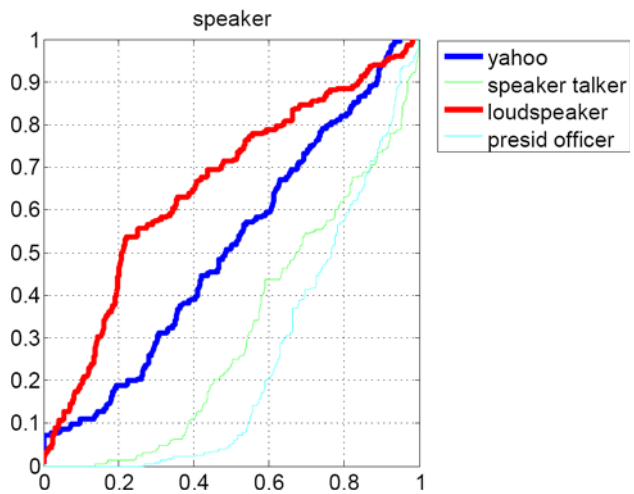
Experiments: re-ranking

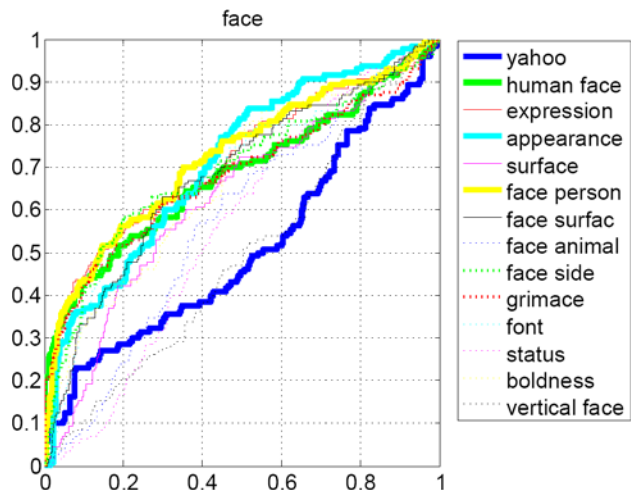
- Unlabeled images re-ranked using dictionary model (top 20)
- Choose one target sense per word

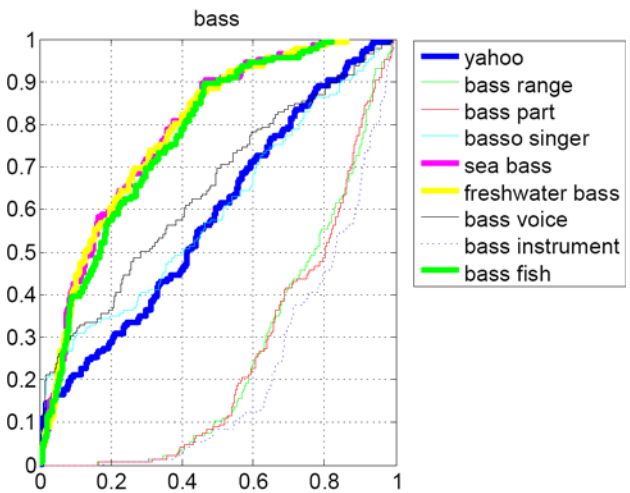


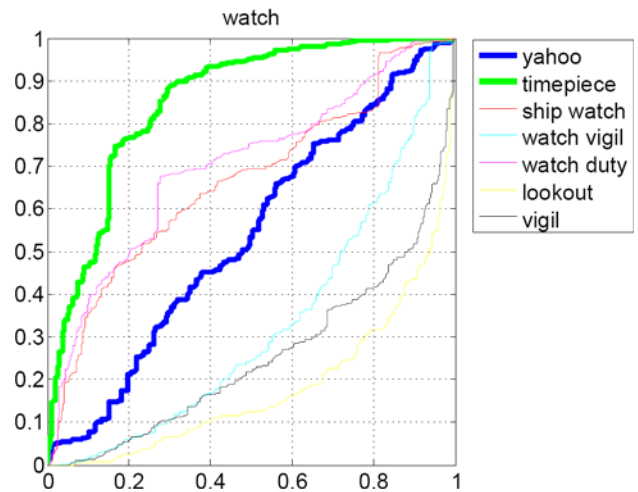
Results: retrieval of target sense



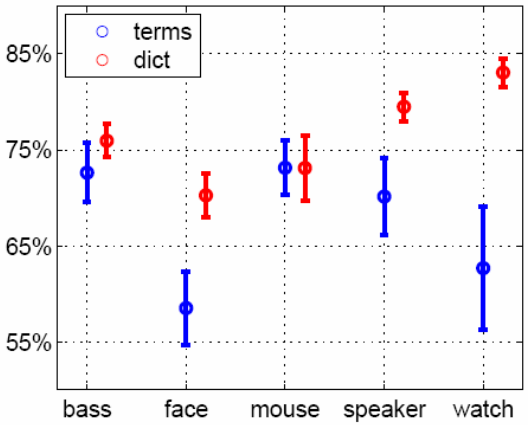




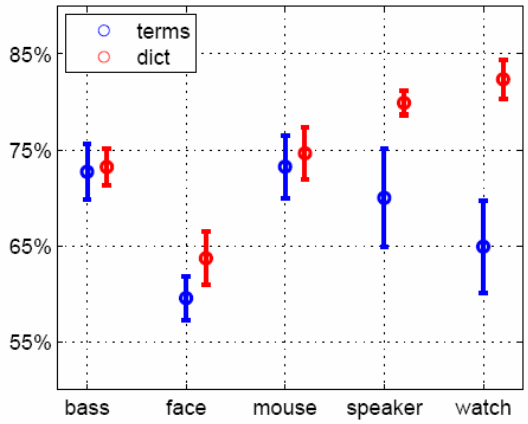




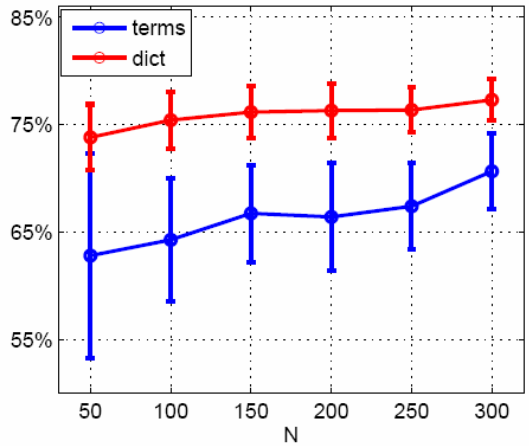
Results: classifying unseen images



(a) 1-SENSE test set



(b) MIX-SENSE test set



(c) 1-SENSE average vs. N

- 1-SENSE test set: negative class consists of ground-truth senses of other objects
- MIX-SENSE test set: negative class also includes the other senses of the word

Adding an Image Term to the Dictionary Model

- The text-only dictionary model defines $P(s|d^t)$, but does not take into account the unlabeled images d^i
- To estimate $P(s|d^i)$, we first fit an LDA model to the unlabeled images in the dataset.

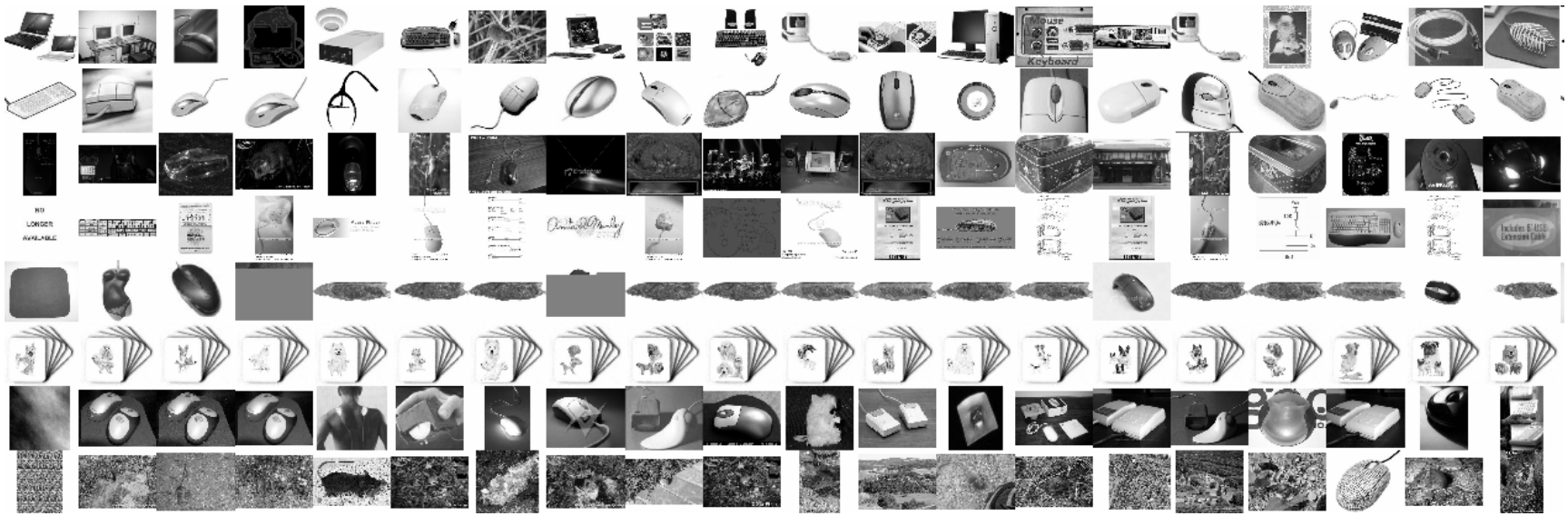


Figure: Most likely images for each image LDA topic for “mouse”

Adding an Image Term to the Dictionary Model

- For each visual topic ***v=1, … ,L***, compute ***P(v|d^i),*** as before,
- Then estimate the conditional probability ***P(s|v)***

$$P(s|v) = \frac{\sum_{k=1}^{M} P(s|d_k^t)P(v|d_k^i)}{P(v)}$$

- Given a test image ***d$^i_*$,*** we can estimate

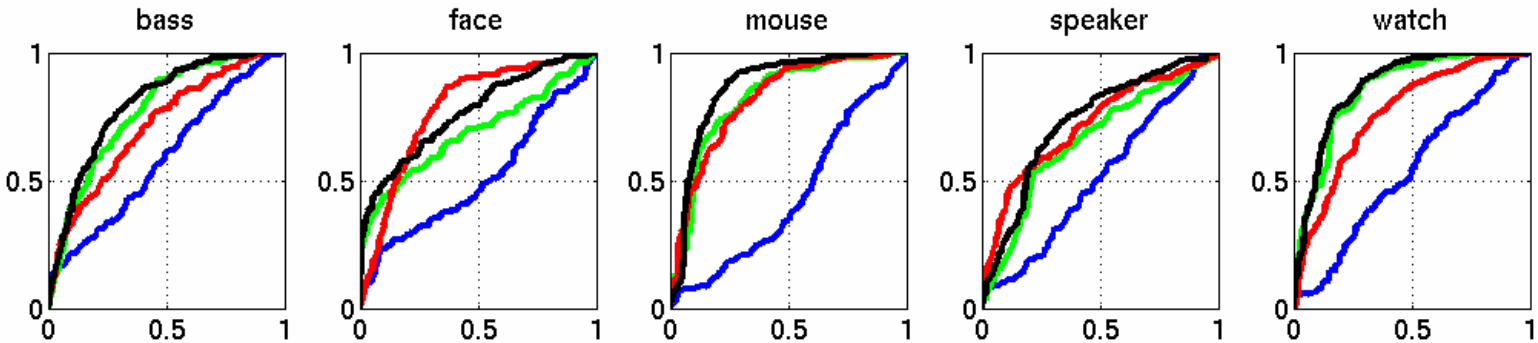
$$P(s|d_*^i) = \sum_{j=1}^{L} P(s|v=j)P(v=j|d_*^i)$$

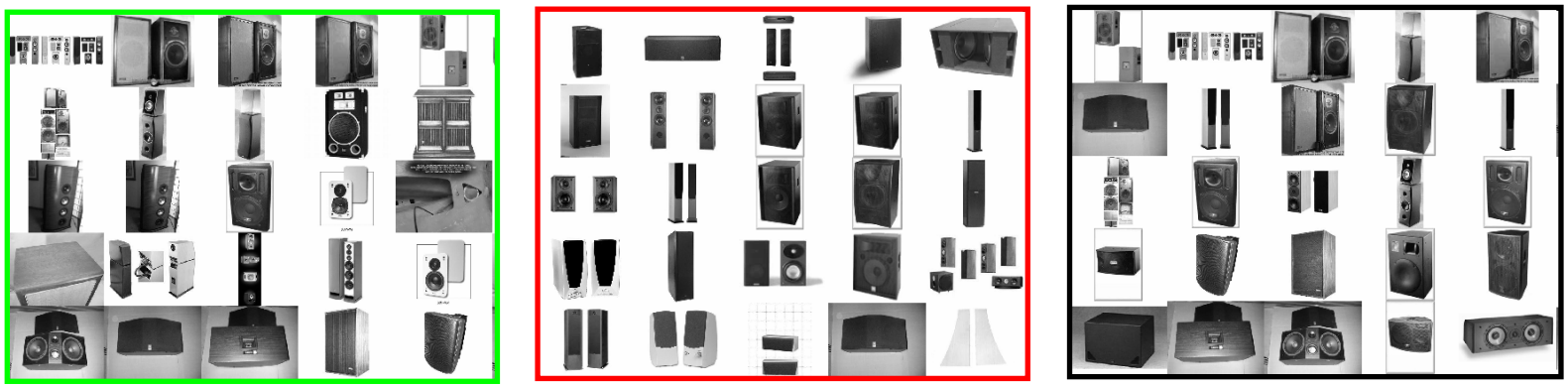
- The multimodal model is defined as

$$P(s|d^i, d^t) = P(s|d^i) + P(s|d^t)$$

Experiments: Image term

- We evaluated retrieval by the image-only and multimodal model and compared it to the previous text-only results
- Evaluated different weights for model combination





Concrete vs. abstract nouns

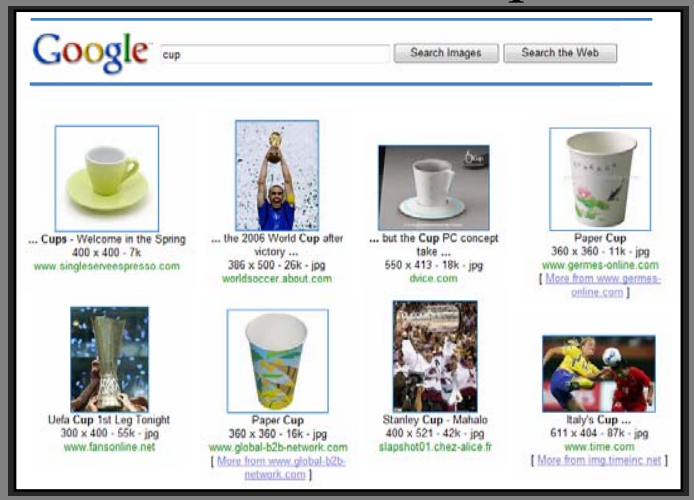
- images associated with an abstract word sense should be excluded when training a visual classifier to learn a model of a physical object.

Wikipedia: Concrete nouns and abstract nouns

Concrete nouns refer to physical bodies that can be observed by at least one of the senses. For instance, "chair", "apple", or "Janet". *Abstract nouns* on the other hand refer to abstract objects, that is ideas or concepts, such as "justice" or "hate". While this distinction is sometimes useful, the boundary between the two of them is not always clear; consider, for example, the noun "art". In English, many abstract nouns are formed by adding noun-forming suffixes ("-ness", "-ity", "-tion") to adjectives or verbs. Examples are "happiness", "circulation" and "serenity".

Filtering out abstract senses

Search Word: "cup"



Online Dictionary

Word to search for: cup

Search Dictionary

Noun

- cup (a small open container usually used for drinking; usually has a handle) *"he put the cup back in the saucer"; "the handle of the cup was missing"*
- cup, loving cup (a large metal vessel with two handles that is awarded as a trophy to the winner of a competition) *"the school kept the cups is a special glass case"*

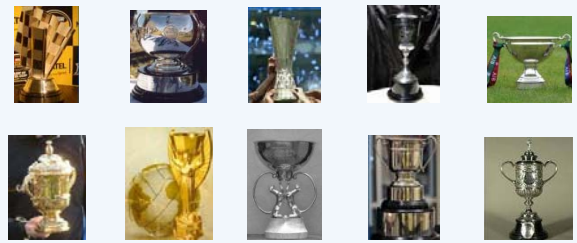
⋮

- a major sporting event or competition *"the world cup", "the Stanley cup"*

Clustering + Word Sense Model

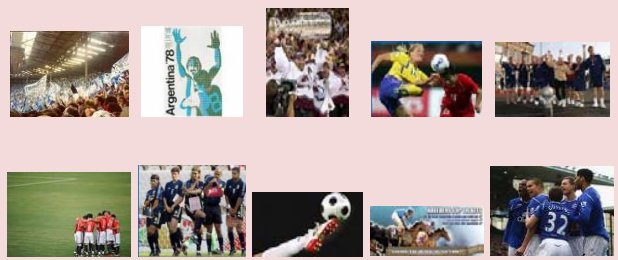
Object Sense: *drinking container*

Object Sense: *loving cup* (*trophy*)



⋮

Abstract Sense: *sporting event*



Filtering abstract senses

- How can we determine if a sense is concrete or abstract?
  - Use a natural language processing method to learn classifier
  - Use existing dictionary information: e.g. WordNet's lexical file tags

Mouse: Noun

•<noun.animal>S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)

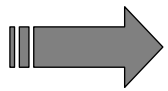
•<noun.state>S: (n) shiner, black eye, **mouse** (a swollen bruise caused by a blow to the eye)

•<noun.person>S: (n) **mouse** (person who is quiet or timid)

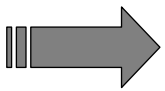
•<noun.artifact>S: (n) **mouse**, computer mouse (a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) *"a mouse takes much more room than a trackball"*

Results: re-ranking search results

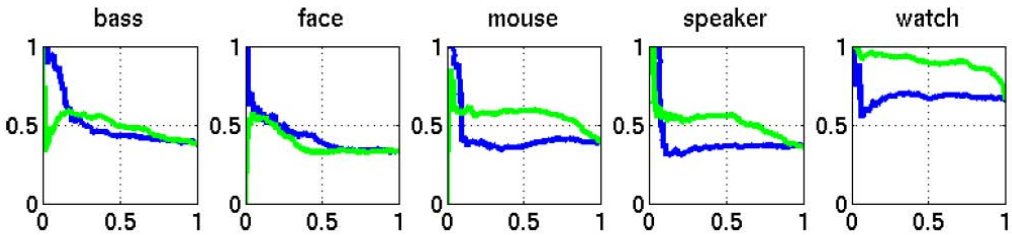
crane



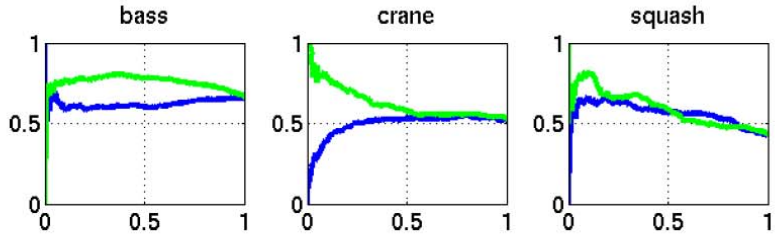
watch



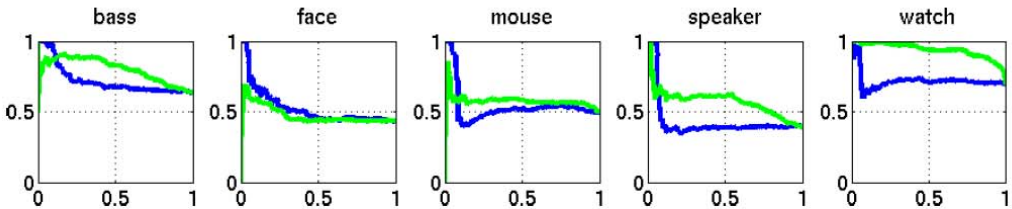
RPC of all concrete senses



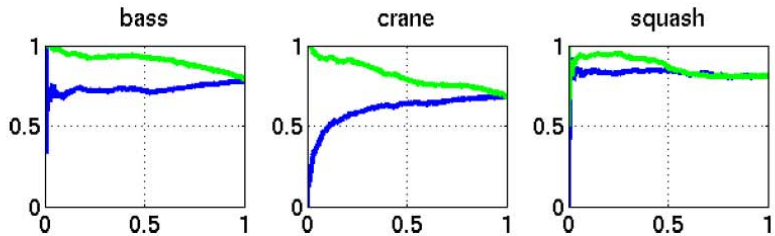
(a) Core Senses, MIT



(b) Core Senses, UIUC



(c) Core+Related Senses, UIUC



(d) Core+Related Senses, UIUC

Figure 4. Retrieval of concrete senses on MIT and UIUC data.

Conclusions

- Proposed an unsupervised method to learn sense-specific object models from web text and image data

- Extended proposed method to filter out non-physical word senses