GROUNDED FINE-GRAINED CLASSIFICATION

OLIVIA WINN

GOAL

Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition

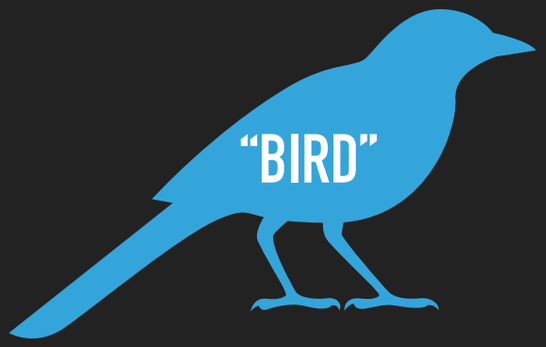


Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition
  - Part identification

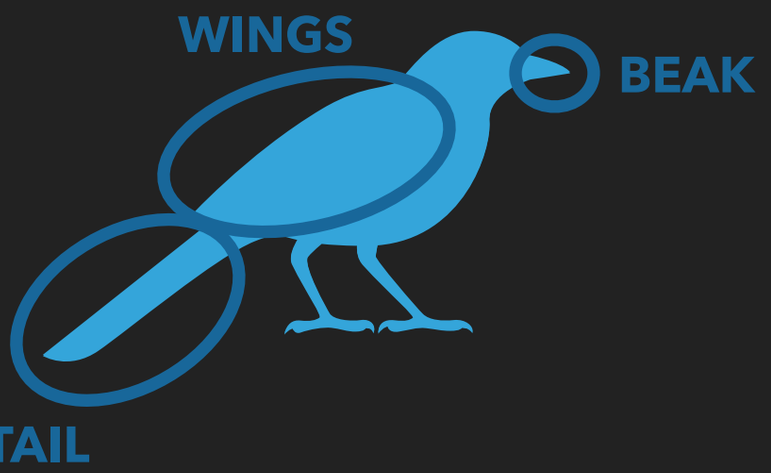
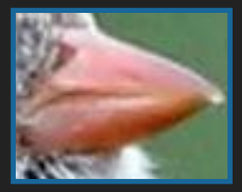


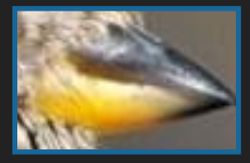
Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition
  - Part identification
- Properties



"orange"



"yellow"

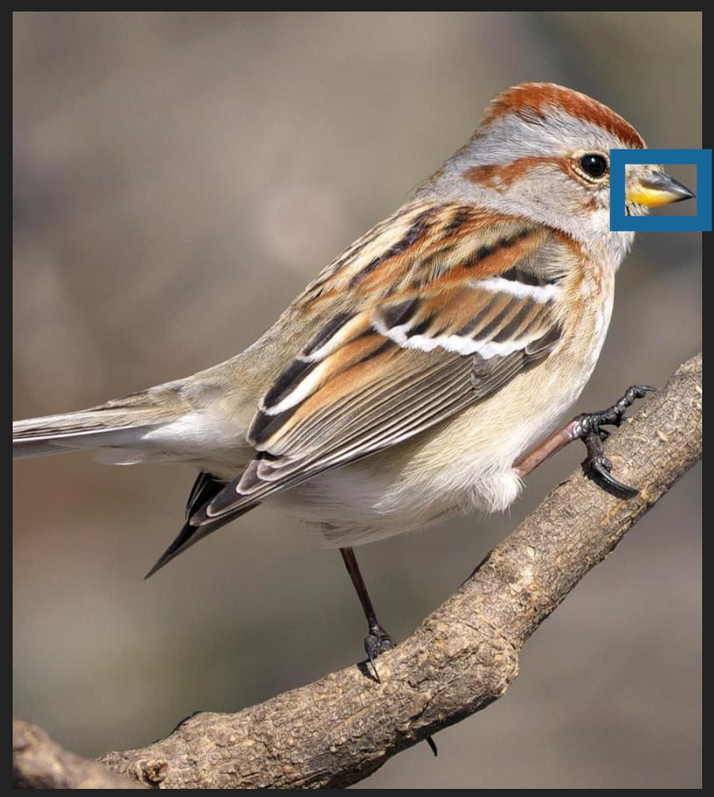
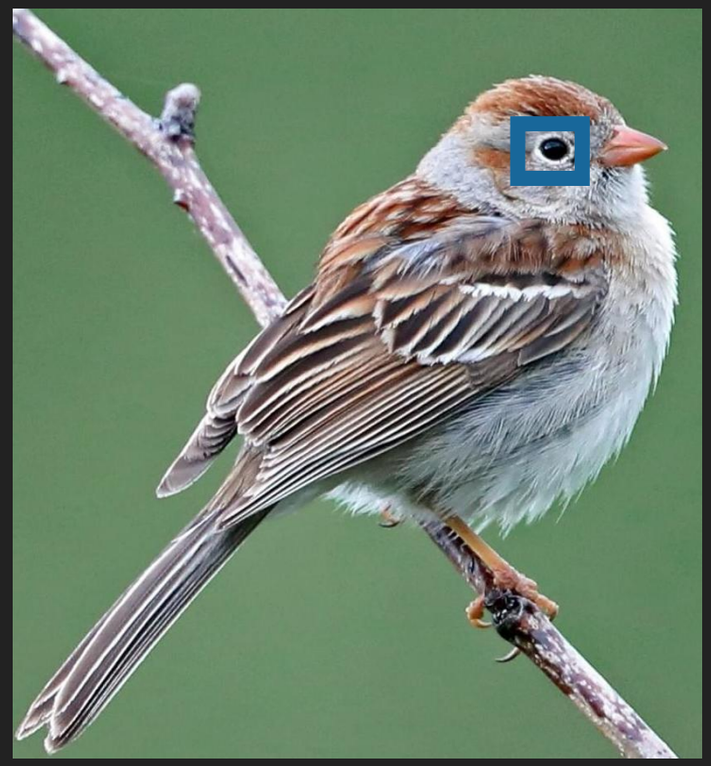


Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition
  - Part identification
- Properties
  - Gradability

"white eye ring"



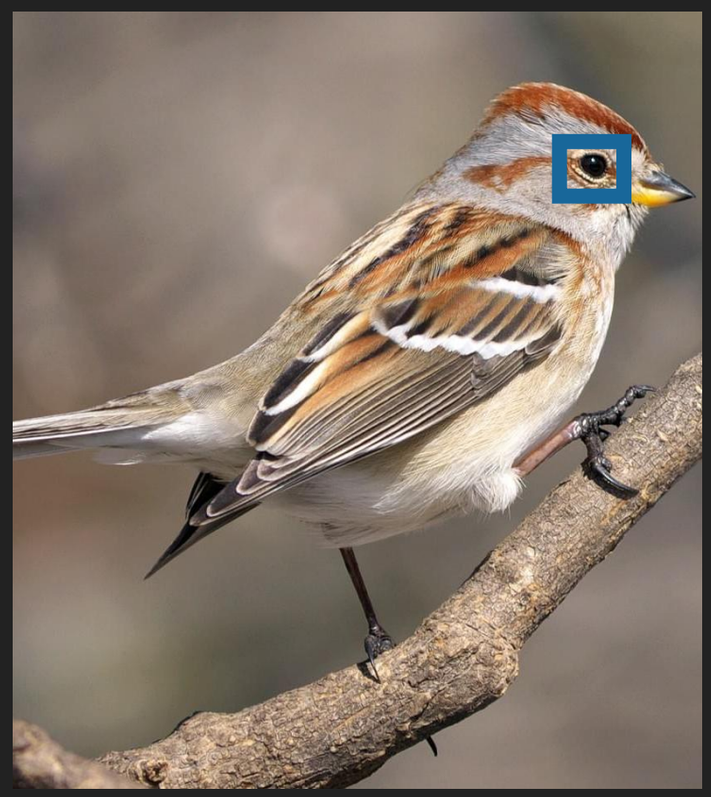
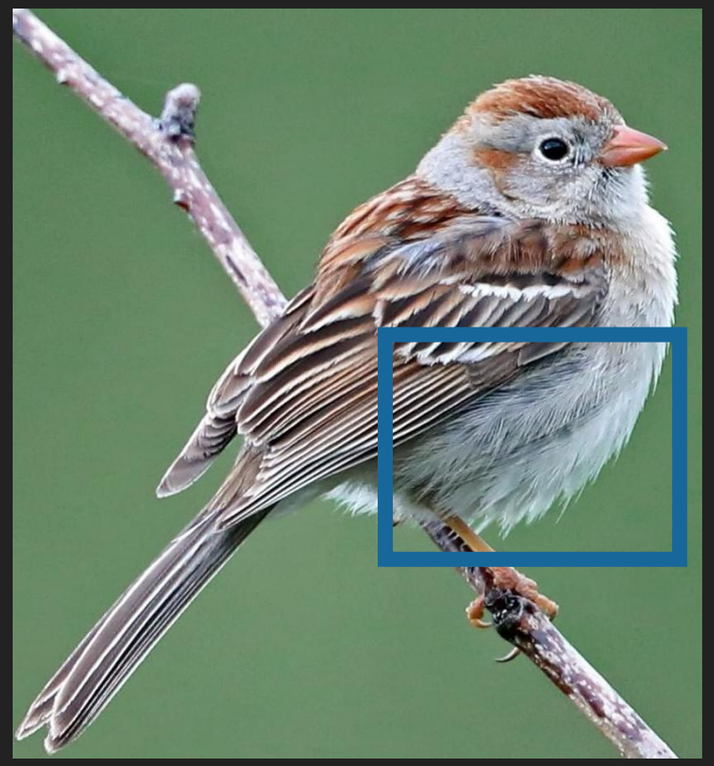


Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition
  - Part identification
- Properties
  - Gradability
  - Vagueness

"mostly white"



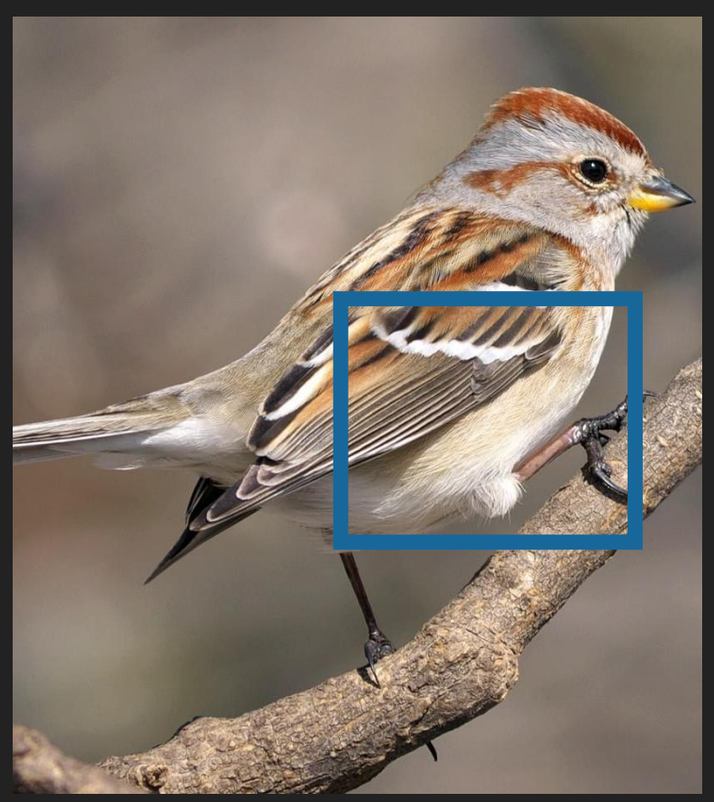


Image Credit: allaboutbirds.org

ASSOCIATING WORDS AND IMAGES

- Object recognition
  - Part identification
  - Properties
    - Gradability
    - Vagueness
- Composition

"white" + "eye" + "ring"

"[American Tree Sparrows] don't have as strong of a white eye ring as Field Sparrows"

Image Credit: allaboutbirds.org

OUTLINE

- Object recognition
  - Part identification
- Properties
  - Gradability
  - Vagueness
- Composition

- Methodology
  - LDA to skipgram
  - Attribute-based learning
- Modifiers
  - Adjective Gradability / Scales
  - Quantifiers / Vagueness
- Composition in Distributional Semantics

OUTLINE

- **Methodology**
  - From LDA to Skipgram
  - Attribute-Based Learning
- Modifiers
- Compositionality in Distributional Semantics

FROM LDA TO SKIPGRAM

- LDA
  - *Document governed by collection of latent topics*
  - *Words have varying probabilities given each topic*

- Skip-Gram
  - *Similar words are used in similar context*
  - *Words represented as points in high-dimensional space where word similarity is measured through cos angle*

FROM LDA TO SKIPGRAM

- LDA
  - *Document governed by collectio*
  - *Words have varying probabilities*
  - How do topics govern multi-modal documents?
- Skip-Gram
  - *Similar words are used in similar c*
  - *Words represented as points in hi where word similarity is measurec*

Feng & Lapata (2010)

Blei & Jordan (2003)

Roller & Walde (2013)

Lazaridou et al (2016)

Wang et al (2017)

Silberer et al (2014)

FROM LDA TO SKIPGRAM

- LDA
  - *Document governed by collectio*
  - *Words have varying probabilities*
  - How do topics govern multi-modal documents?
- Skip-Gram
  - *Similar words are used in similar c*
  - *Words represented as points in hi where word similarity is measurec*
- What text should be used?

Feng & Lapata (2010)

Blei & Jordan (2003)

Roller & Walde (2013)

Lazaridou et al (2016)

Wang et al (2017)

Silberer et al (2014)

FROM LDA TO SKIPGRAM

- LDA
  - *Document governed by collectio*
  - *Words have varying probabilities*
  - How do topics govern multi-modal documents?
- Skip-Gram
  - *Similar words are used in similar c*
  - *Words represented as points in hi where word similarity is measurec*
- What text should be used?
- How can we learn new information?

Feng & Lapata (2010)

Blei & Jordan (2003)

Roller & Walde (2013)

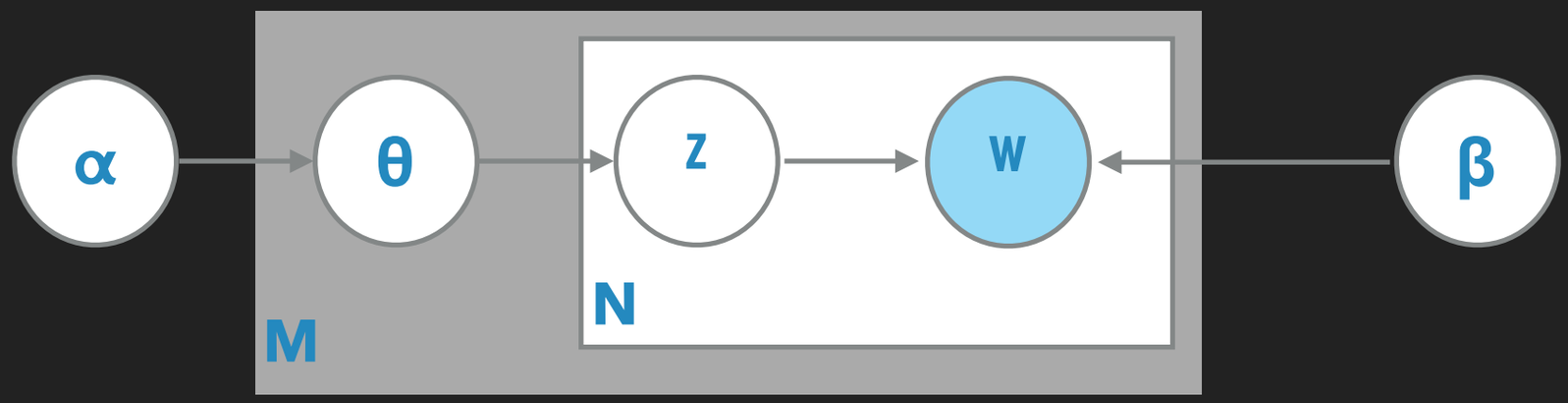
Lazaridou et al (2016)

Wang et al (2017)

Silberer et al (2014)

VISUALLY-INFORMED LDA

Feng & Lapata (2010)

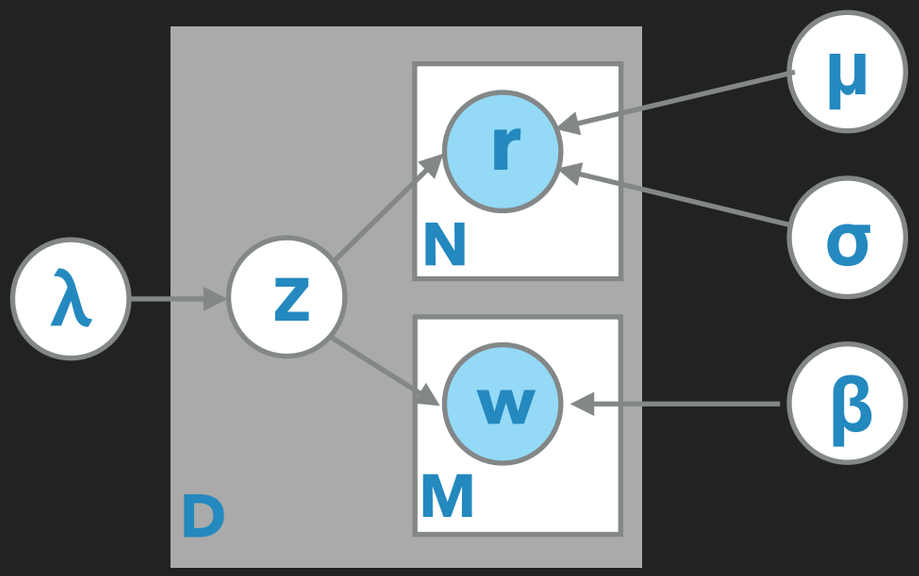


- **Document**: bag of words and image features (BoVW) together
- **Goal**: enhance word meaning through visual information
  - Use β as word representation to measure word similarity and word association
- **Data**: BBC news articles with images
- **Result**: Visual information improvement over pure text model
- **Limitation**: No correlation between words and images

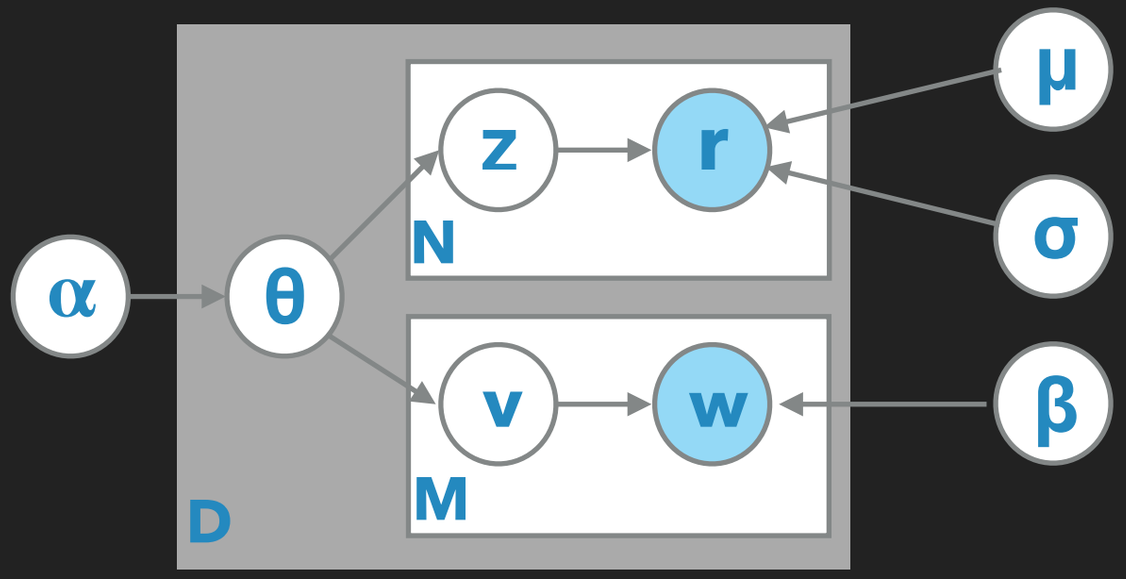
MULTI-VARIATE LDA

Blei & Jordan (2003)

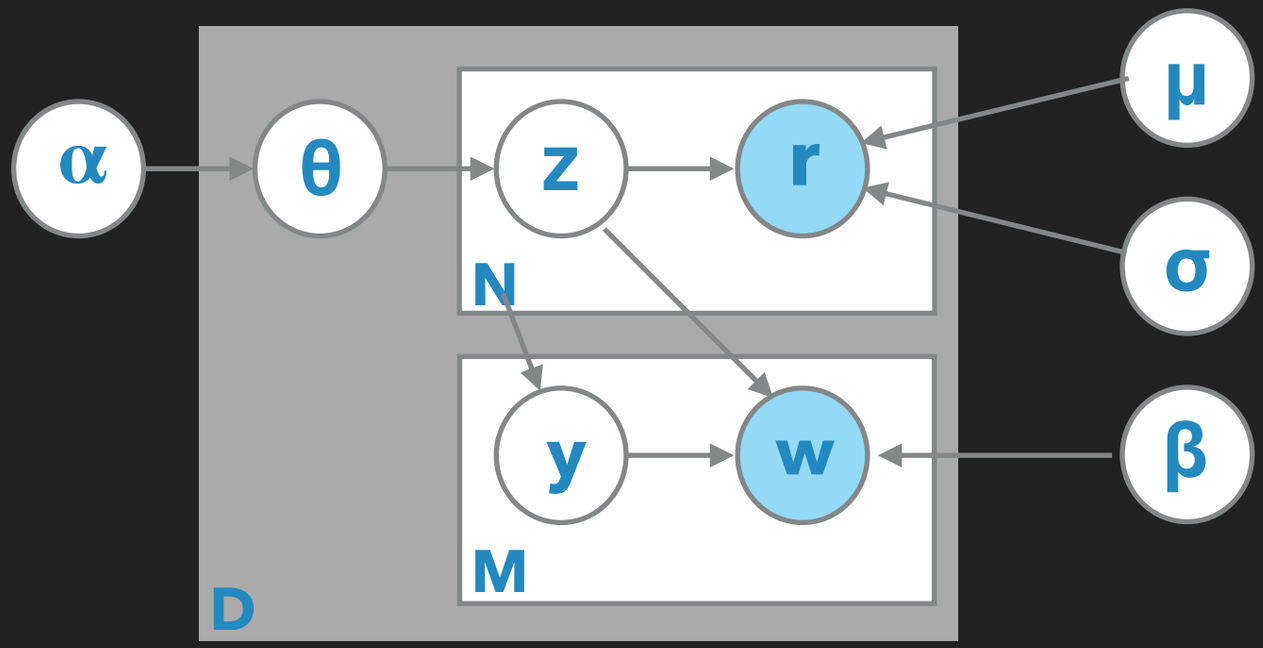
GM-MIXTURE



GM-LDA



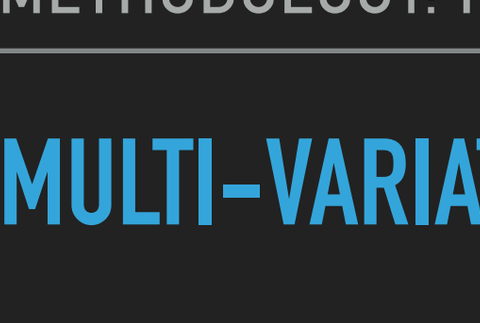
Corr-LDA



MULTI-VARIATE LDA

Blei & Jordan (2003)

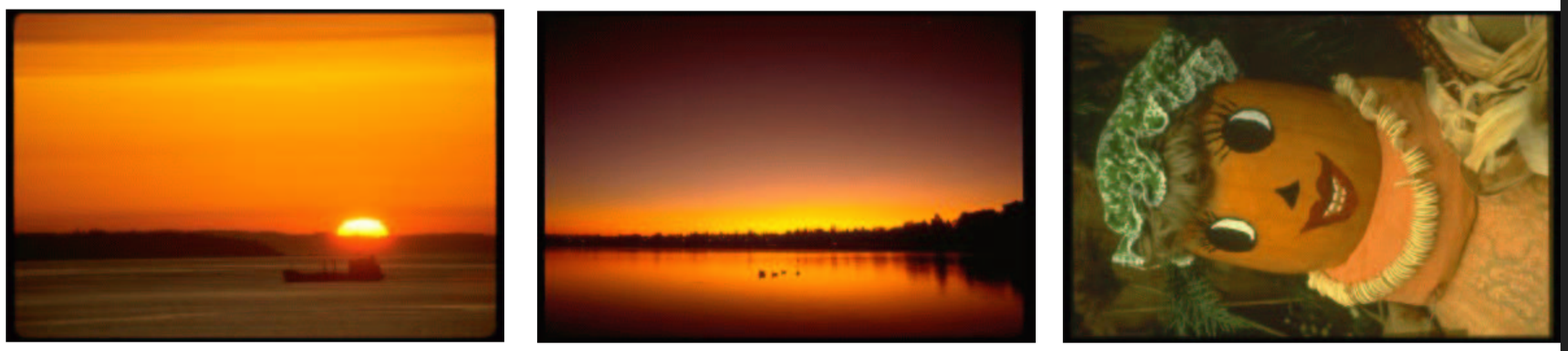
- **Document**: Image w/ words generated by image regions
- **Goal**: Caption generation & text-based image retrieval
- **Data**: 7000 image-caption pairs from Corel database



True caption
birds tree

Corr–LDA
birds nest leaves branch tree

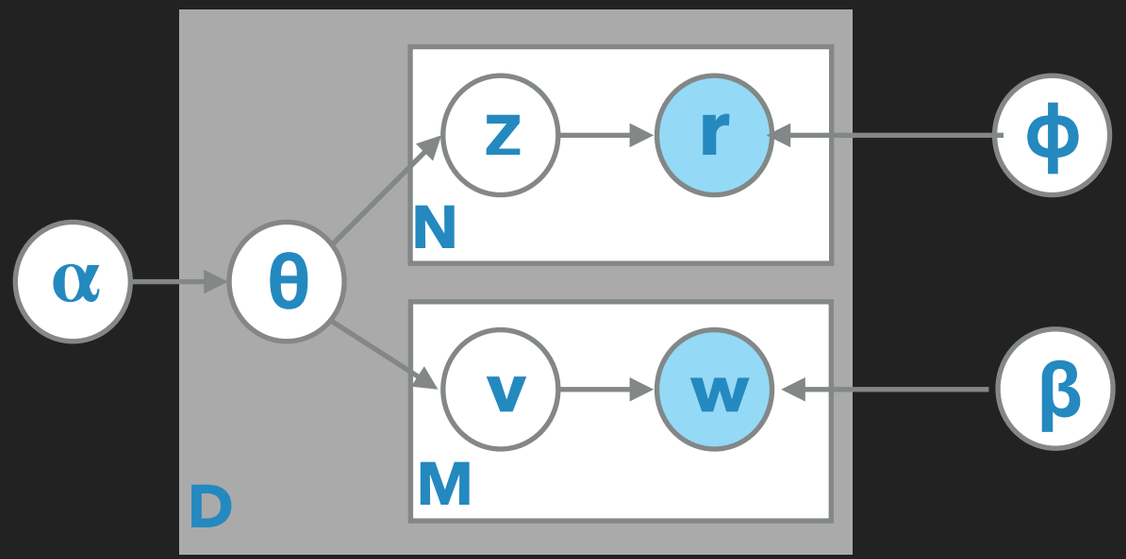
Sunset



- **Limitation**: Context is lost

MULTI-VARIATE LDA

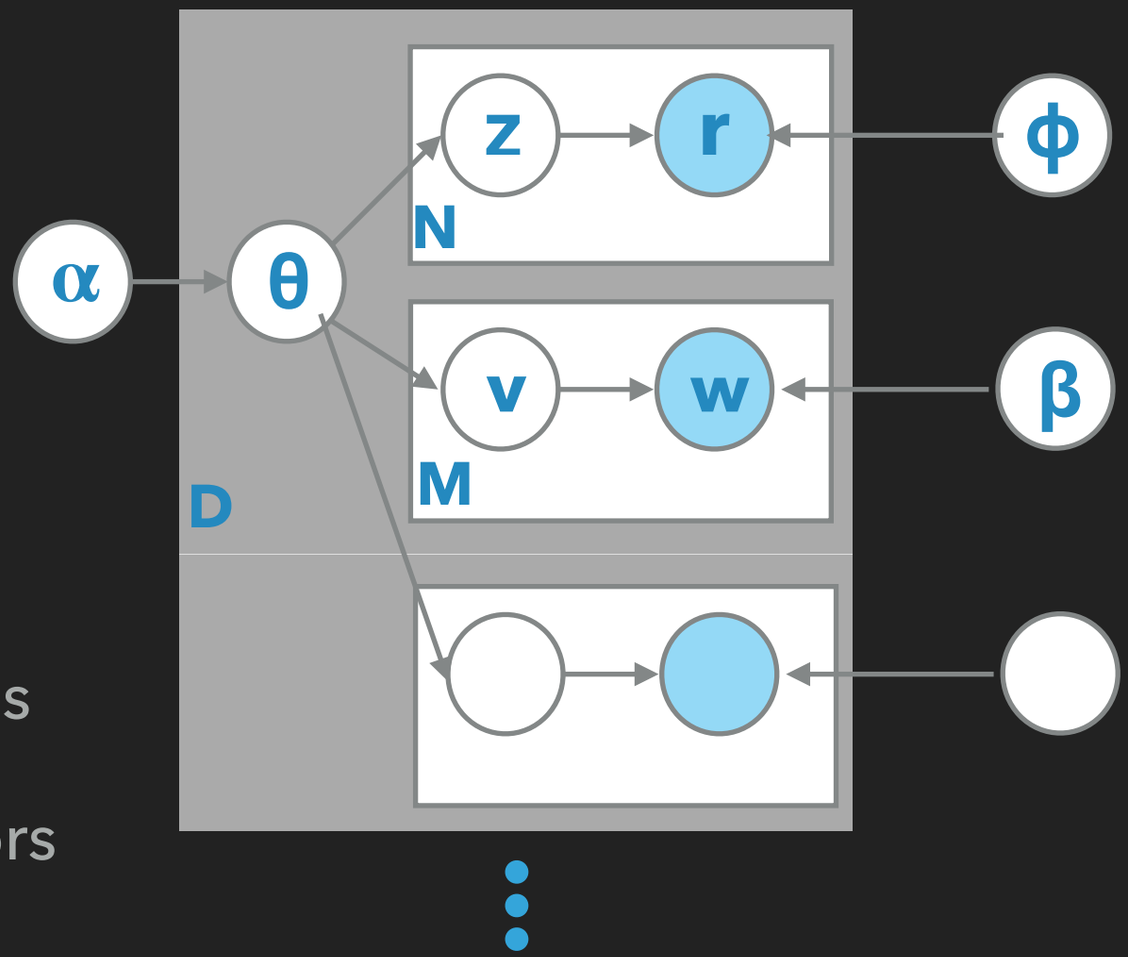
Roller & Walde (2013)



- **Goal**: Introduce context as additional variable

MULTI-VARIATE LDA

Roller & Walde (2013)



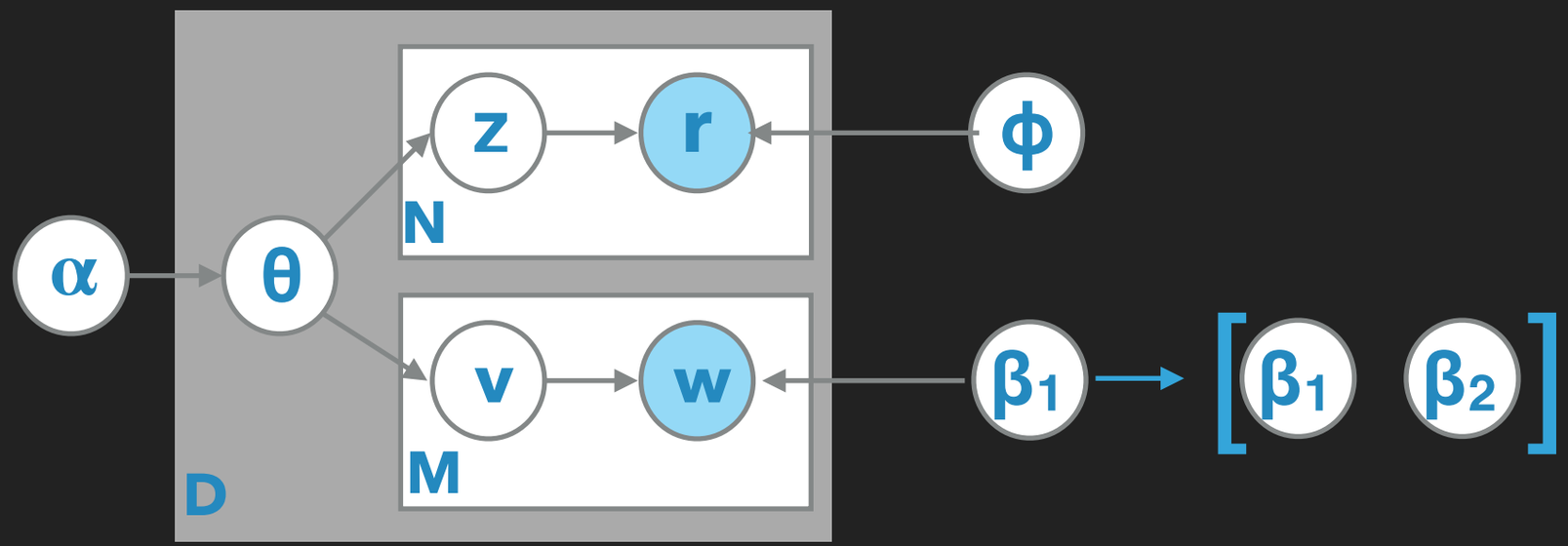
Associated Nouns

Object Descriptors

- **Goal**: Introduce context as additional variable
- **Document**:
  - 3D-LDA

MULTI-VARIATE LDA

Roller & Walde (2013)



- **Goal**: Introduce context as additional variable
- **Document**:
  - 3D-LDA
  - HybridLDA: concatenate β from separately trained models

MULTI-VARIATE LDA

Roller & Walde (2013)

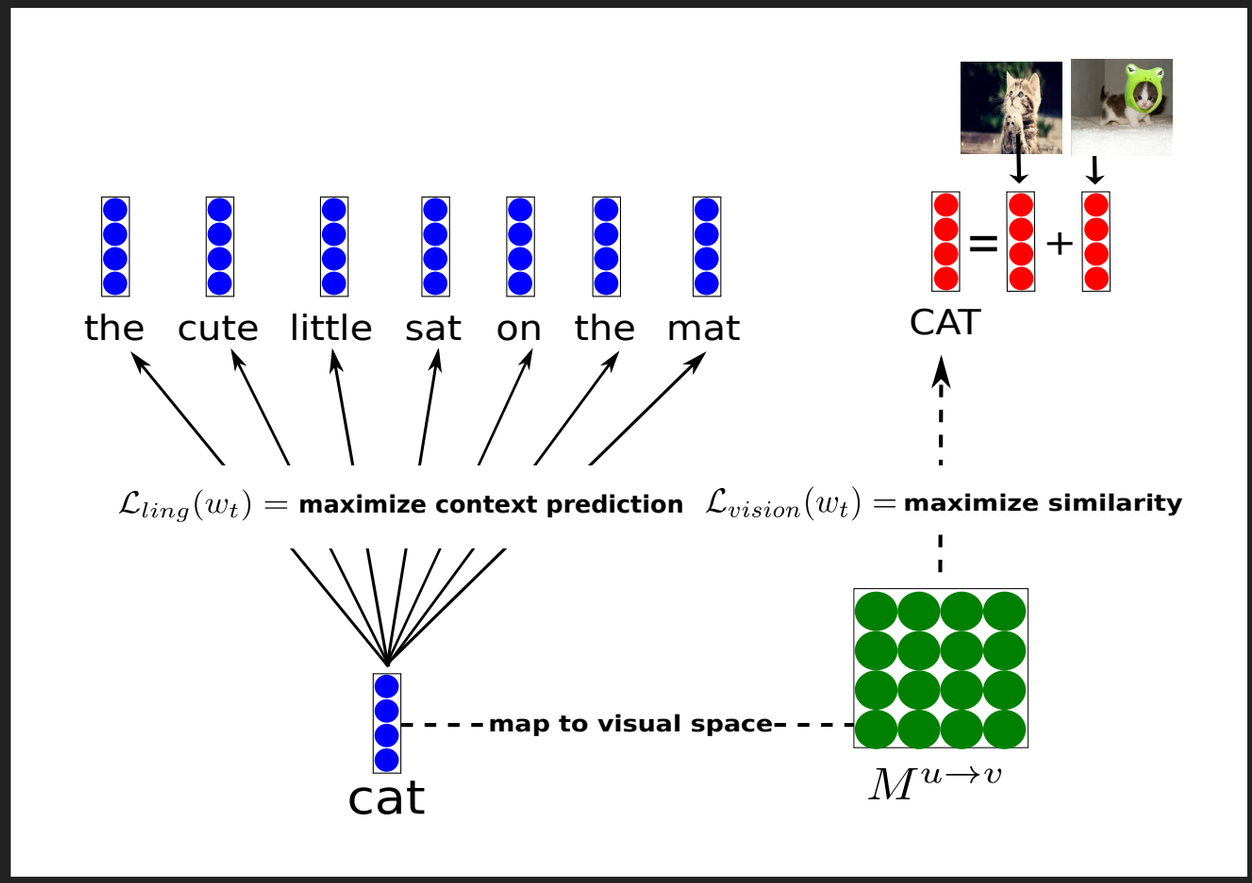
- **Data**: ImageNet, deWaC, association norms, feature norms
- **Result**: Hybrid LDA combining all data is most successful

Top Words	Prototypical Images
water ship lake sea meter river	
(number) clock Friday Sunday Saturday Monday	

- **Limitation**: Images represented by feature clusters; semantically unrelated components can be combined

MULTIMODAL SKIPGRAM

Lazaridou et al (2016)

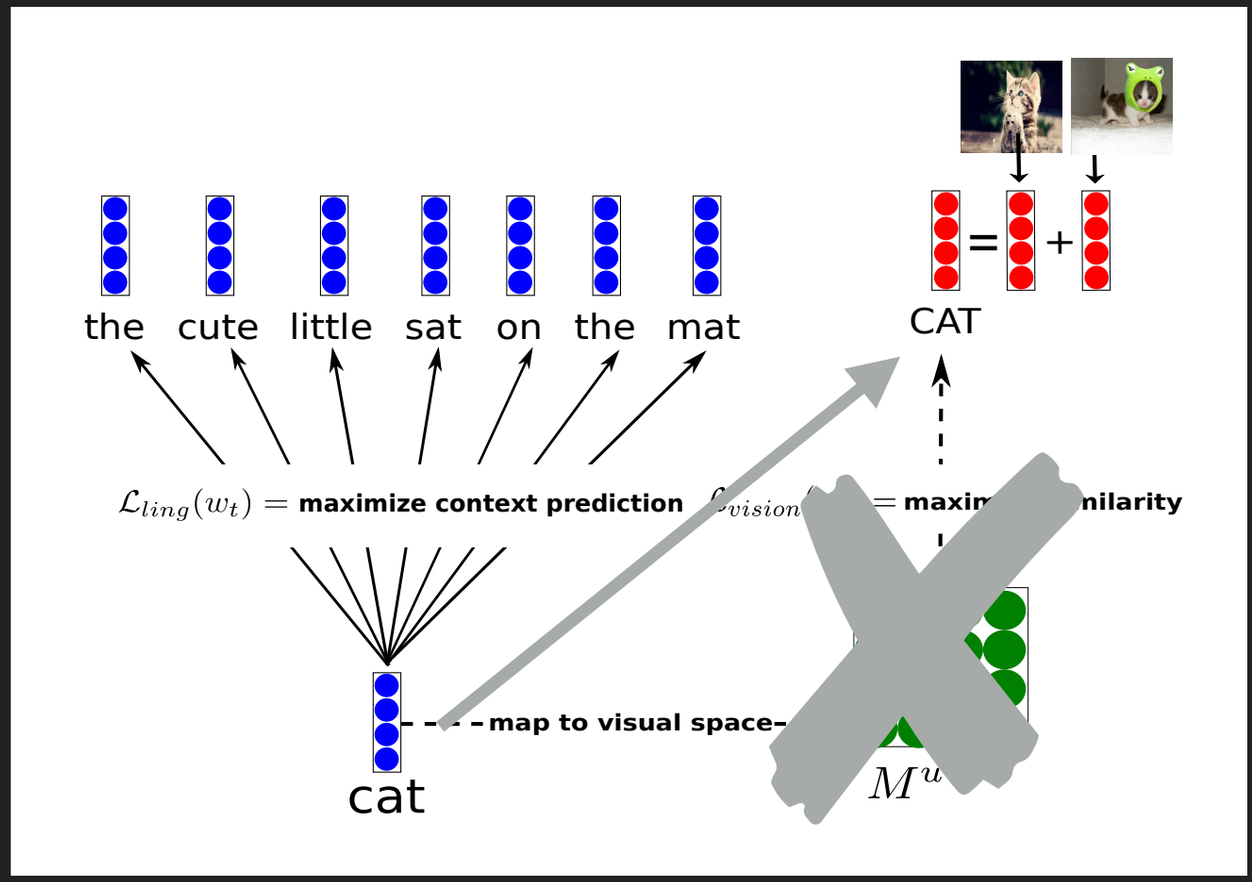


MMSkip-Gram-B

- **Goal**: Include visual information in skip-gram context
- **Data**: Wikipedia & ImageNet

MULTIMODAL SKIPGRAM

Lazaridou et al (2016)



MMSkip-Gram-A

- **Goal**: Include visual information in skip-gram context
- **Data**: Wikipedia & ImageNet

MULTIMODAL SKIPGRAM

Lazaridou et al (2016)

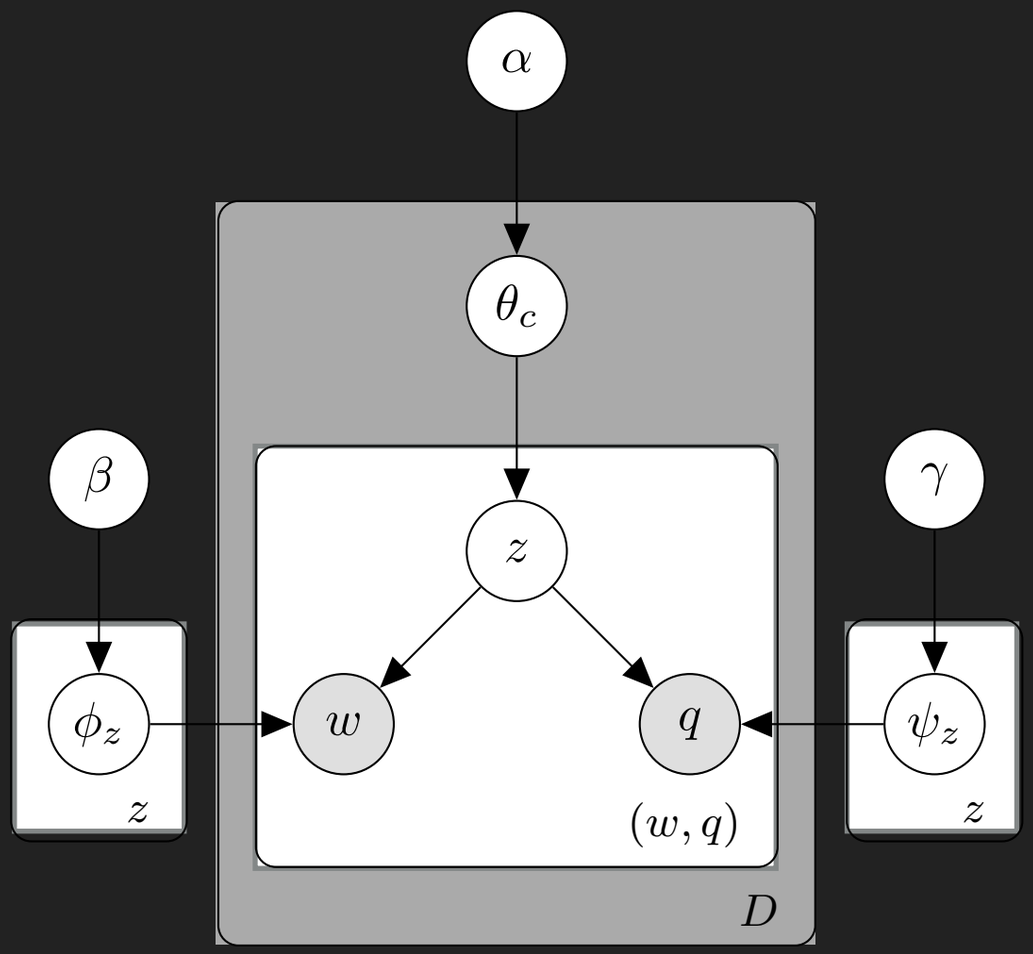
- **Results**:
  - Adding visual information for only some words improves word similarity for all

Target	SKIP-GRAM	MMSKIP-GRAM-A	MMSKIP-GRAM-B
donut	fridge, diner, candy	pizza, sushi, sandwich	pizza, sushi, sandwich
owl	pheasant, woodpecker, squirrel	eagle, woodpecker, falcon	eagle, falcon, hawk
mural	sculpture, painting, portrait	painting, portrait, sculpture	painting, portrait, sculpture
tobacco	coffee, cigarette, corn	cigarette, cigar, corn	cigarette, cigar, smoking
depth	size, bottom, meter	sea, underwater, level	sea, size, underwater
chaos	anarchy, despair, demon	demon, anarchy, destruction	demon, anarchy, shadow

- **Limitations**:
  - Word similarity can vary depending on context
  - What about uncommon words?

ONE-SHOT

Wang et al (2017)



- **Document**: made up of words and properties that appeared as children of the same <*word-dependency relation*> pair
- **Goal**: Learn properties from single exposure to object in a context
- **Data**: QMR & AD (quantified attr. datasets) Text-only approach
- **Result**:

Top 5 properties for 'gown'; context *undo-dobj*	
bi-TM one-shot	`clothing`, `is_long`, `made_of_material`, `different_colours`, `has_sleeves`

- **Limitations**: Visual properties score lowest

zero-shot:

$$P(q|d_u) = \sum_z P(z|\theta_u)P(q|\psi_z)$$

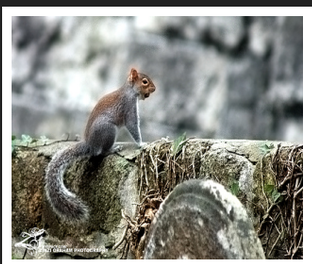
one-shot:

$$P(q|w) = \sum_z P(z|w)P(q|\psi_z)$$

VISUAL INFORMATION

Silberer et al (2014)

- **Document**: collection of objects & visual representations which share the same property
- **Goal**: Use visual information to corroborate properties
- **Data**: McRae attributes, Wikipedia extracted word-attribute pairs
- **Results**:
  - Physically grounding text adds meaning



climbs, climbs_trees, crawls, hops, jumps, eats, eats_nuts, is_small, has_bushy_tail
has_4_legs, has_head, has_neck, has_nose, has_snout, has_tail, has_claws
has_eyes, has_feet, has_toes,

- **Limitations**:
  - No correlation between image features and individual attributes

FROM LDA TO SKIPGRAM

	Document	Contribution	Limitation
Feng & Lapata (2010)	Unordered collection of text and image words	Images **enhance** word distributions in **LDA**	**No correlation** between text and images
Blei & Jordan (2003)	Image region –> caption word	Can name **multiple objects** in image	Linguistic and visual **context lost**
Roller & Walde (2013)	Image, name, association norms, feature norms	Images and textual context best governed by **separate latent factors**	**Correlations** in the spaces **do not map**
Lazaridou et al (2016)	**Skip-Gram**: full text + some images	Visual information **informs entire space**	**Assume single meaning**
Wang et al (2017)	Dependency parse relation (text only)	Object properties can be **learned from context**	Without images, **visual properties hardest**
Silberer et al (2014)	Property (object + image)	Images **inform property understanding**	**No mapping** between visual properties and property words

FROM LDA TO SKIPGRAM

	Document	Contribution	Limitation
Feng & Lapata (2010)	Unordered collection of text and image words	Images **enhance** word distributions in **LDA**	**No correlation** between text and images
Blei & Jordan (2003)	Image region –> caption word	Can name **multiple objects**	Linguistic and visual **context lost**
Roller & Walde (2013)	Image, name, association norms, feature norms	ontext **parate latent factors**	**Correlations** in the spaces **do not map**
Lazaridou et al (2016)	**Skip-Gram**: full text + some images	Visual information **informs entire space**	**Assume single meaning**
Wang et al (2017)	Dependency parse relation (text only)	Object properties can be **learned from context**	Without images, **visual properties hardest**
Silberer et al (2014)	Property (object + image)	Images **inform property understanding**	**No mapping** between visual properties and property words

- Name: ??
- Properties:
  - Dog-like
  - Striped
  - Black tail

Image Credit: arkive.org

ATTRIBUTE-BASED LEARNING

- **Attribute**: ‘human-nameable mid-level semantic property’
- Object: co-occurring correlated bundles of attributes

  - Dog-like
  - Striped
  - Black tail

Image Credits: arkive.org, wikipedia.org, wildlife-photographs.blogspot.com

ATTRIBUTE-BASED LEARNING

- **Attribute**: ‘human-nameable mid-level semantic property’
- Object: co-occurring correlated bundles of attributes

- Dog-like
- Striped
- Black tail

Lazaridou et al (2014)

Hwang & Sigal (2014)

Chen et al (2017)

Vedantam et al (2017)

Image Credits: arkive.org, wikipedia.org, wildlife-photographs.blogspot.com

ATTRIBUTE-BASED LEARNING

- **Attribute**: ‘human-nameable mid-level semantic property’
- Object: co-occurring correlated bundles of attributes

- Dog-like
- Striped
- Black tail
- Has ears

Lazaridou et al (2014)

Hwang & Sigal (2014)

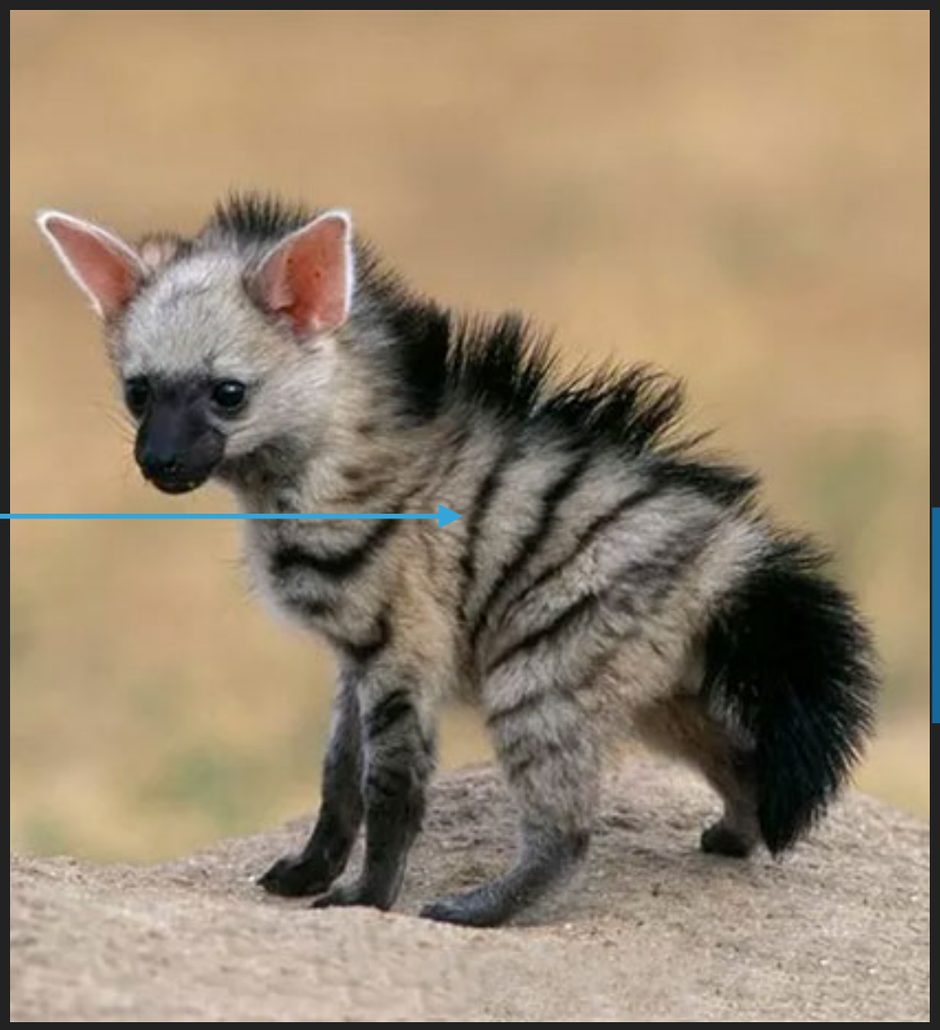
Chen et al (2017)

Vedantam et al (2017)

Image Credits: arkive.org, wikipedia.org, wildlife-photographs.blogspot.com

ATTRIBUTE-BASED LEARNING

- **Attribute**: ‘human-nameable mid-level semantic property’
- Object: co-occurring correlated bundles of attributes



- Dog-like
- Striped
- Black tail

Lazaridou et al (2014)

Hwang & Sigal (2014)

Chen et al (2017)

Vedantam et al (2017)

Image Credits: arkive.org, wikipedia.org, wildlife-photographs.blogspot.com

ATTRIBUTE-BASED LEARNING

- **Attribute**: ‘human-nameable mid-level semantic property’
- Object: co-occurring correlated bundles of attributes

- Dog-like
- Striped
- Black tail

Lazaridou et al (2014)

Hwang & Sigal (2014)

Chen et al (2017)

Vedantam et al (2017)

Image Credits: arkive.org, wikipedia.org, wildlife-photographs.blogspot.com

IMPLICIT ATTRIBUTES

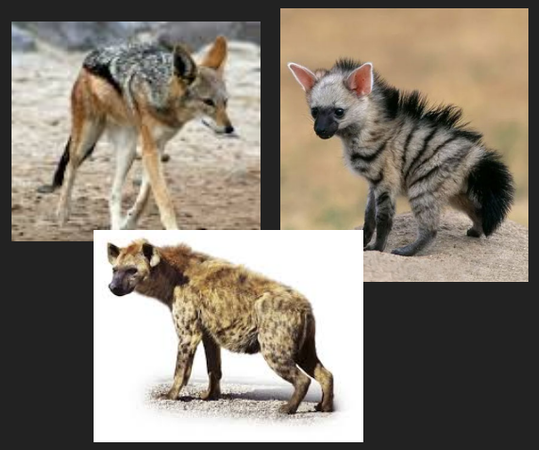
Lazaridou et al (2014)

- **Fast-mapping:** people immediately learn new object from limited info

→ 'dog-like'

'Aardwolf' → 

- "Aardwolf cubs often share the den with their mother" → mammal, fur
- Distributional representations bring words of similar context together
- Objects with the same properties have similar image features



Cross-Modal Mapping

'Aardwolf'

'Mongoose'

'Hyena'

IMPLICIT ATTRIBUTES

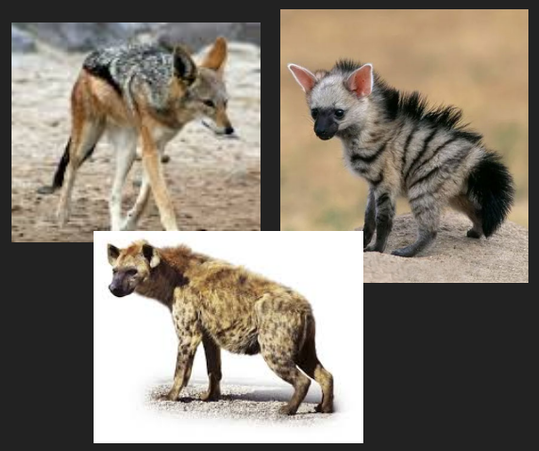
Lazaridou et al (2014)

- **Fast-mapping:** people immediately learn new object from limited info

→ 'dog-like'

'Aardwolf' → ?

- "Aardwolf cubs often share the den with their mother" → mammal, fur
- Distributional representations bring words of similar context together
- Objects with the same properties have similar image features



$f_{proj_{v \to w}} = \Theta_{v \to w}$

Neural Network

$f_{proj_{w \to v}} = \Theta_{w \to v}$

'Aardwolf'

'Mongoose'

'Hyena'

- **Data**: Wikipedia articles, CIFAR-10 & ESP images

CROSS-MODAL MAPPING

Lazaridou et al (2014)

- **Results**: Categorization induced by hidden layer of neural network

	Seen Concepts	Unseen Concept	Rank of Correct Unseen Concept	CIFAR-100 Category
Unit 1	**sunflower**, **tulip**, pear	butterfly	2 (**rose**)	flowers
Unit 2	**cattle**, **camel**, **bear**	squirrel	2 (**elephant**)	large omnivores and herbivores
Unit 3	**castle**, **bridge**, **house**	bus	4 (**skyscraper**)	large man-made outdoor things

- Neighbors of mapped vectors reveal information

→ spoke, wheel, brake, tyre, motorcycle

- **Limitation**: Similarity in the spaces do not always correspond

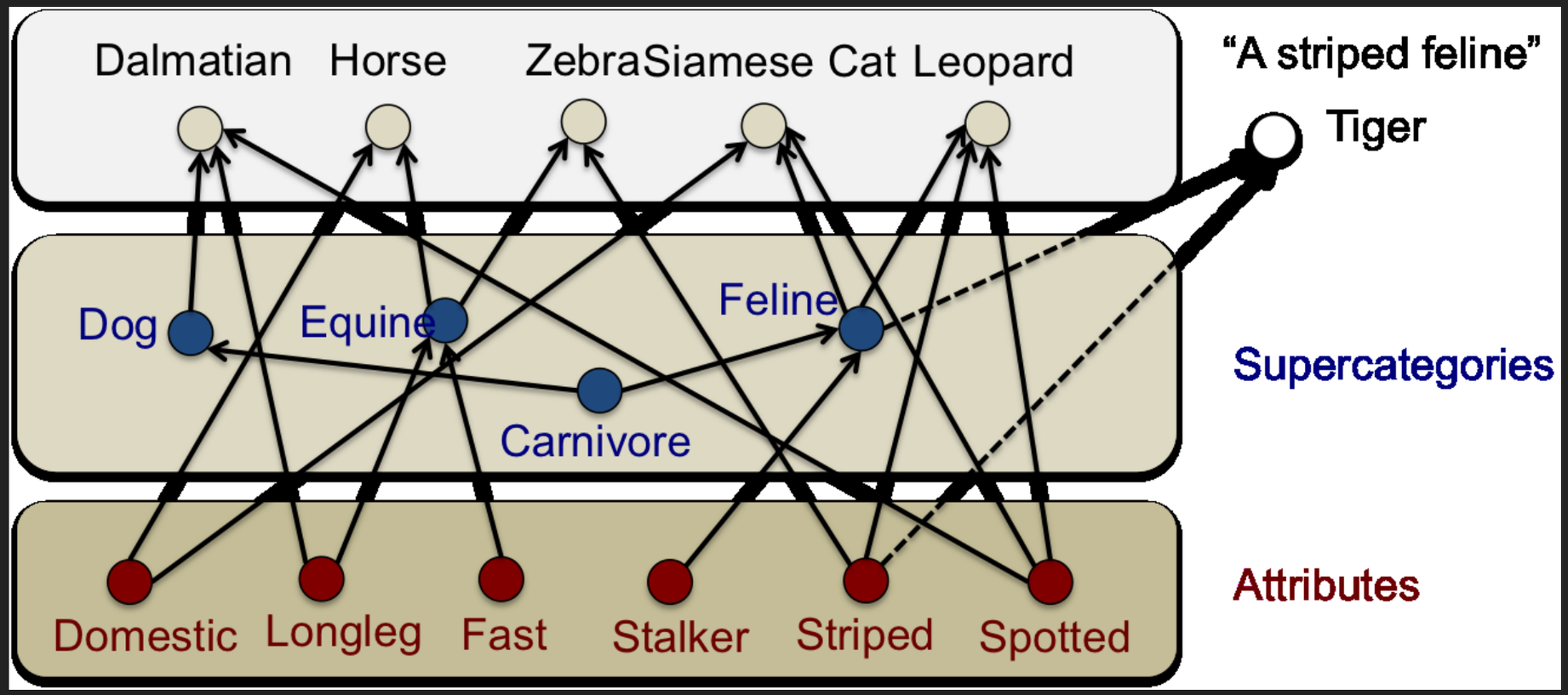
dishwasher →   → potato

UNIFIED SEMANTIC EMBEDDING

Hwang & Sigal (2014)

- ‘Dog-like’
  - Inherent, indescribable properties → posture, head shape
  - Attributes → four legs, tail, ears, snout…
- Class = ‘super-class’ + unique attributes

- Dog-like
- Striped
- Black tail



UNIFIED SEMANTIC EMBEDDING

Hwang & Sigal (2014)

- Mapping to joint space:
  1. Image representation close to class vector (and farther from others)
  2. Class vector closer to its super-class than to the other classes
  3. Attribute vectors maximize correlation with respective images
- 'Relationship regularization' ***in*** joint space:
  4. Class = superclass + attributes

$$\mathcal{R}(U, B) = \sum_{c}^{C} ||u_c - u_p - U^A \beta_c||_2^2 + \gamma_2 ||\beta_c + \beta_o||_2^2$$

UNIFIED SEMANTIC EMBEDDING

Hwang & Sigal (2014)

- Mapping to joint space:
  1. Image representation close to class vector (and farther from others)
  2. Class vector closer to its super-class than to the other classes
  3. Attribute vectors maximize correlation with respective images

- 'Relationship regularization' ***in*** joint space:
  4. Class = superclass + attributes

$$\mathcal{R}(U,B) = \sum_{c}^{C} ||u_c - u_p - U^A \beta_c||_2^2 + \gamma_2 ||\beta_c + \beta_o||_2^2$$

  5. $0 \preceq \beta_c$ Describe objects with the attributes they have

UNIFIED SEMANTIC EMBEDDING

Hwang & Sigal (2014)

- Mapping to joint space:
  1. Image representation close to class vector (and farther from others)
  2. Class vector closer to its super-class than to the other classes
  3. Attribute vectors maximize correlation with respective images
- 'Relationship regularization' ***in*** joint space:
  4. Class = superclass + attributes

$$\mathcal{R}(U,B) = \sum_{c}^{C} ||u_c - u_p - U^A \beta_c||_2^2 + \gamma_2 ||\beta_c + \beta_o||_2^2$$

  5. $0 \preceq \beta_c$ Describe objects with the attributes they have
  6. 'exclusive' regularization - ensures unique decomposition per class

UNIFIED SEMANTIC EMBEDDING

Hwang & Sigal (2014)

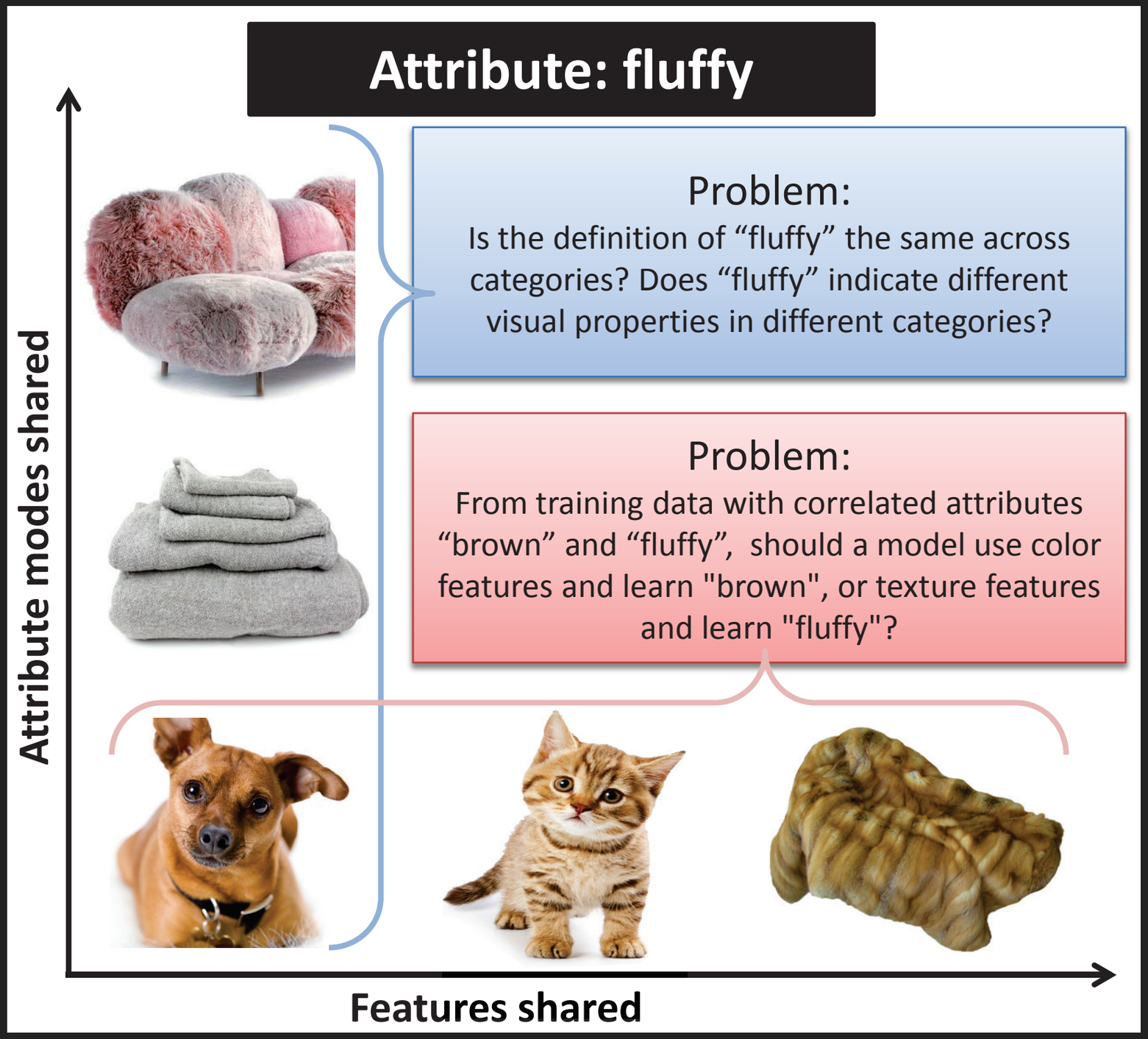
- **Data**: Animals with Attributes, super-classes from WordNet hierarchy

Category	Ground-truth attributes	
 Otter	An animal that swims, fish, water, new world, small, flippers, furry, black, brown, tail, . . .	
	Supercategory + learned attributes	
	A musteline mammal that is quadrapedal, flippers, furry, ocean	
Primate	N/A	An animal, that has hands and bipedal

- **Limitations**:
  - Strict hierarchy not applicable to all domains
  - Cannot handle recognition through *lack* of attribute

MULTI-TASK ATTRIBUTE LEARNING

Chen et al (2017)

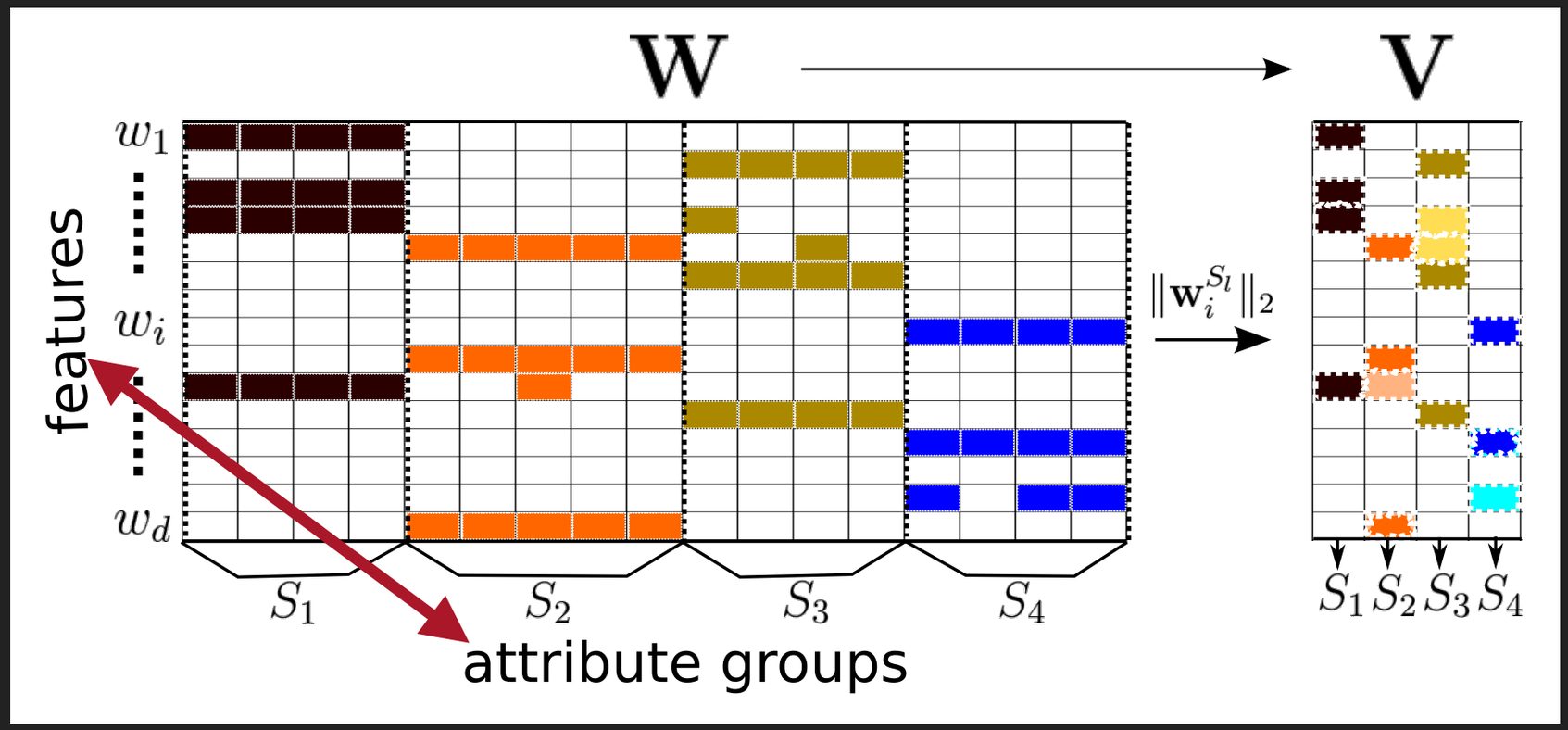


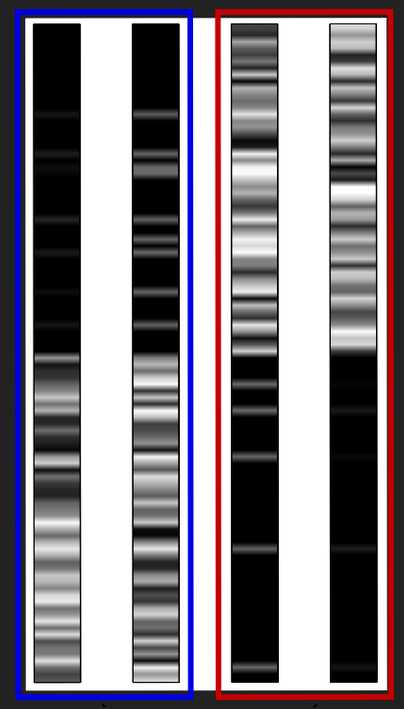
Data: CUB, AWA, aPascal/aYahoo images w/ attributes

MULTI-TASK ATTRIBUTE LEARNING

Chen et al (2017)

- **Selective Sharing**



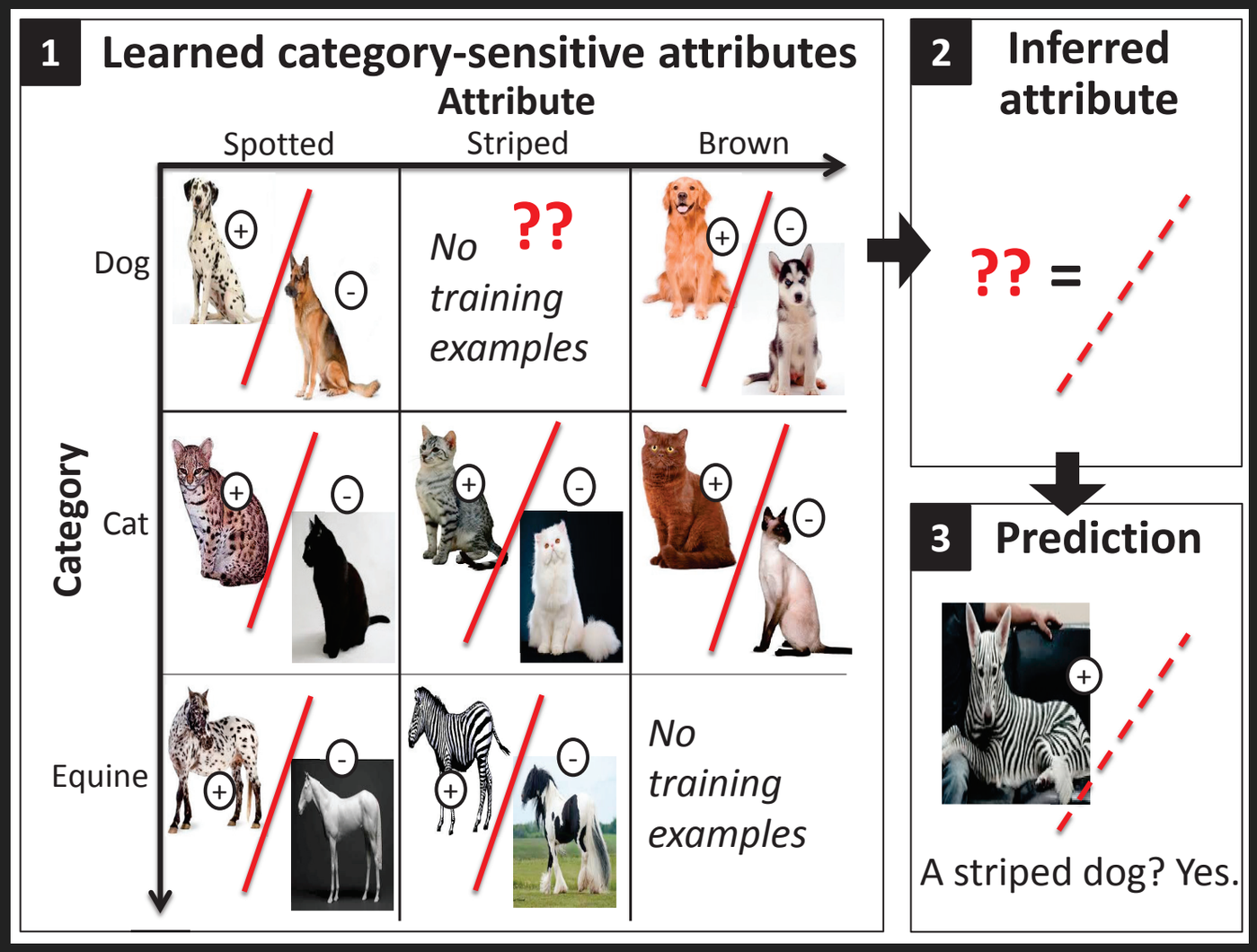


- **Limitation**: Features must match attribute groups

MULTI-TASK ATTRIBUTE LEARNING

Chen et al (2017)

- **Category-Sensitive Attributes**

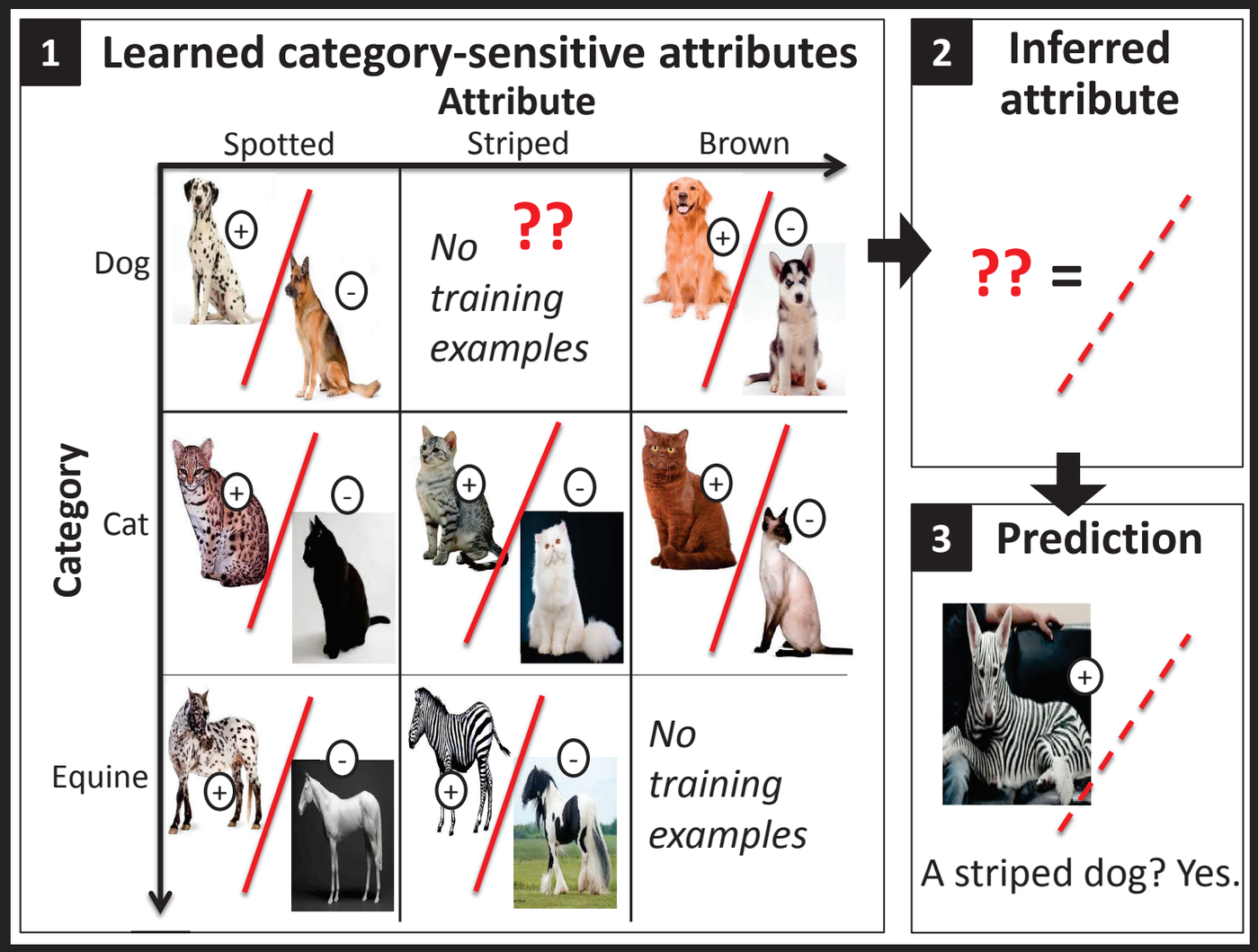


- Train SVM for each class-specific attribute
  - Use all attribute instances
  - In-class penalty is higher
- Represent models as tensor
- Use tensor completion to ‘hypothesize’ missing classifiers

MULTI-TASK ATTRIBUTE LEARNING

Chen et al (2017)

- **Category-Sensitive Attributes**



- **Limitation**: Correlation of attributes can be useful

VISUALLY GROUNDED IMAGINATION

Vedantam et al (2017)

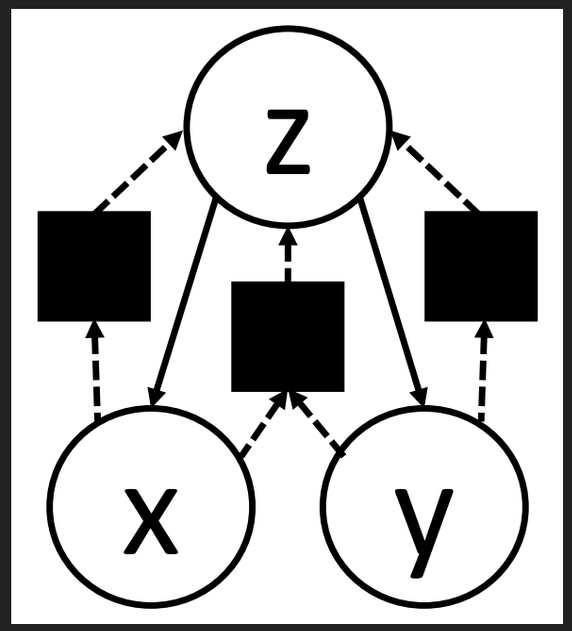
- Attribute: Interaction between adjective and noun
- Three aspects: Coverage, Correctness, Compositionality
- **Goal**: How do we handle cases of missing information?

"striped" ✓ "dog-like" ✓ → "striped" + "dog-like" = ?

→ **striped ?**
(independent of other aspects)

VISUALLY GROUNDED IMAGINATION

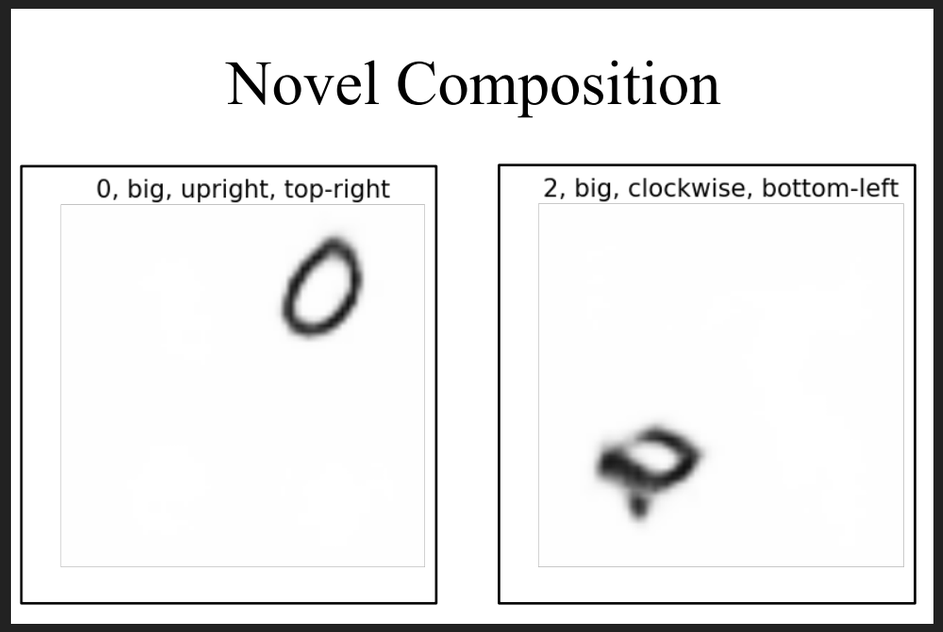
Vedantam et al (2017)

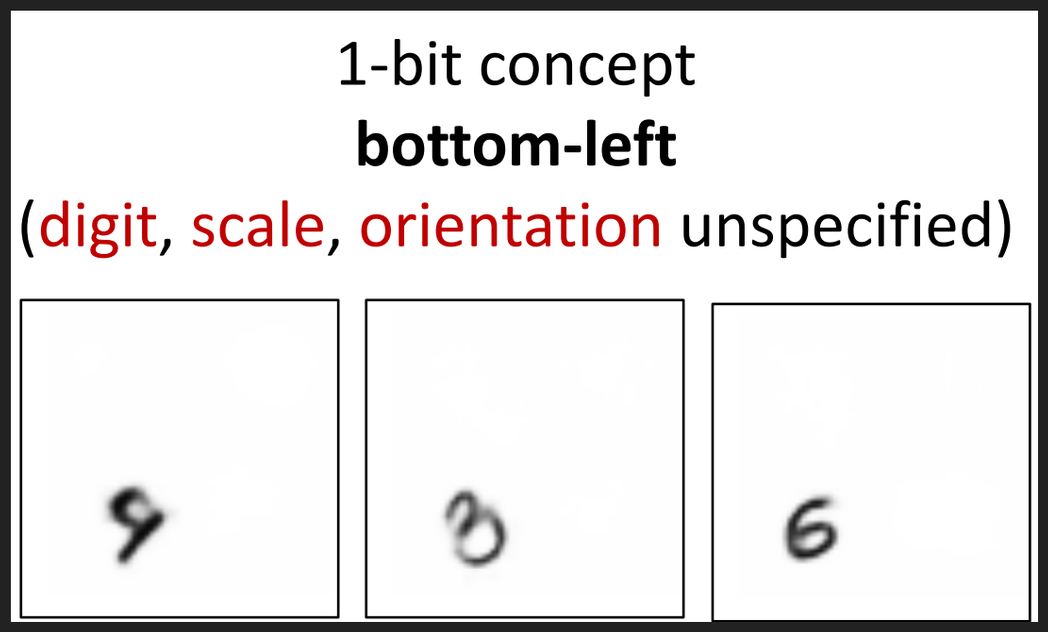


- Partial attributes: "**striped** **any color** **bird**"

$$q(z|y_O) \propto p(z) \prod_{k \in O} q(z|y_k)$$

- **Product of Experts**: Latent space only adjusted by attributes when specified

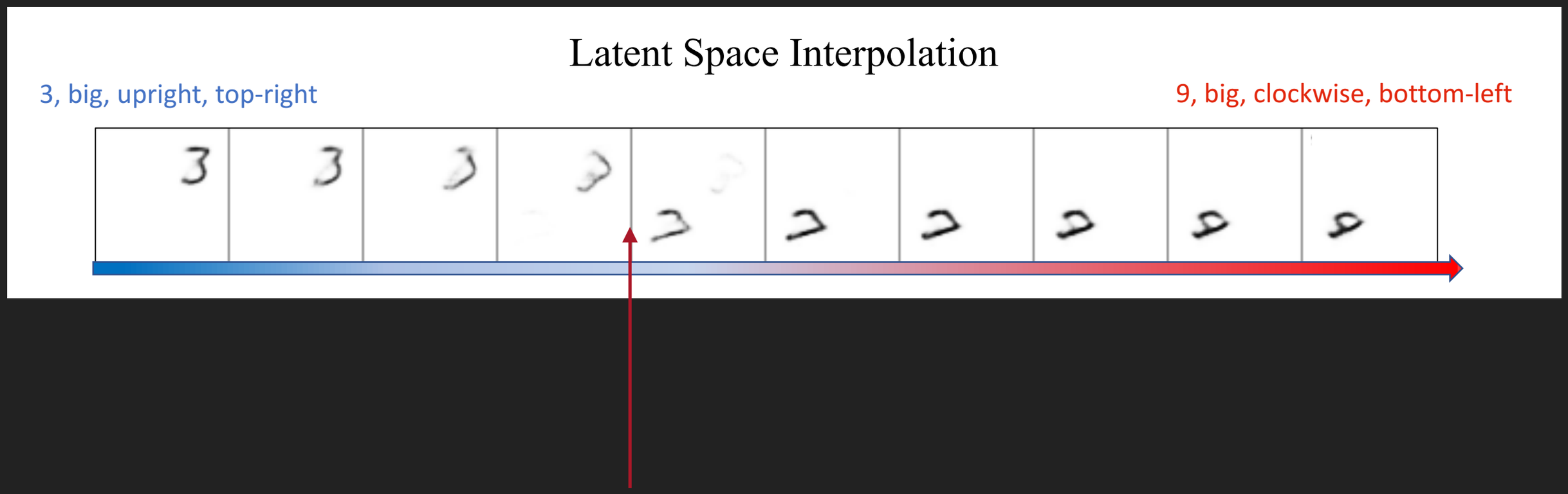




- **Data**: MNist images and attributes

VISUALLY GROUNDED IMAGINATION

Vedantam et al (2017)

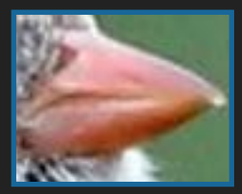


- **Limitation**: Not truly learning proper interpolation of attributes
  Not representing the attributes themselves

ATTRIBUTE-BASED REPRESENTATION

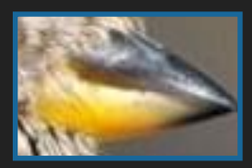
	Approach	Result	Limitation
Lazaridou et al (2014)	Cross-modal map	**Context** provides information	Context is not always appropriate
Hwang & Sigal (2014)	Joint hierarchical map	Context can be **structured**	But not too structured
Chen et al (2017)	Feature & class specific classifiers	Dependent on **components**	Have classifier but not representation
Vedantam et al (2017)	Joint latent space	**Continuous** space between attributes	Cannot model space between attributes

- Attributes are the result of **adjectives** *modifying* a **noun**
  - Nouns are abstract: contain all objects which fit under the label
  - Adjectives provide concrete picture or example
- How do we model that modification in feature space?
- First, examine the linguistics of modification



"orange"

"beak"



"yellow"

OUTLINE

- Methodology
- **Modifiers**
  - Adjectives
    - Gradability / Scales
    - Comparison
  - Quantifiers, Vagueness
- Compositionality in Distributional Semantics

GRADABILITY

- Modification can occur at varying intensities

tiny small large huge

de Melo & Bansal (2013)

Qing & Franke (2014)

Lassiter & Goodman (2015)

GRADABILITY

- Modification can occur at varying intensities
  - Can we automatically learn adjective intensity?

tiny small large huge

de Melo & Bansal (2013)

Qing & Franke (2014)

Lassiter & Goodman (2015)

GRADABILITY

- Modification can occur at varying intensities
  - Can we automatically learn adjective intensity?
- Individual adjectives have ranges they can apply to
  - How do we determine these ranges and their cutoffs?

tiny small large huge

de Melo & Bansal (2013)

Qing & Franke (2014)

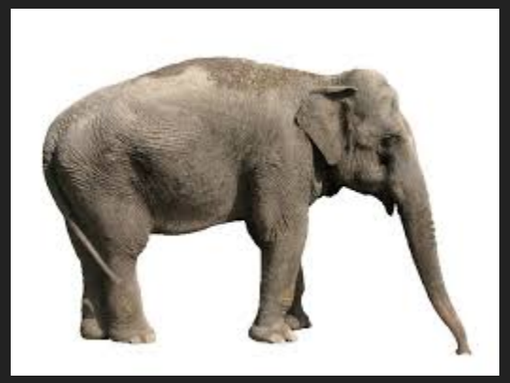
Lassiter & Goodman (2015)

GRADABILITY

- Modification can occur at varying intensities
  - Can we automatically learn adjective intensity?
- Individual adjectives have ranges they can apply to
  - How do we determine these ranges and their cutoffs?

~~huge~~

large



de Melo & Bansal (2013)

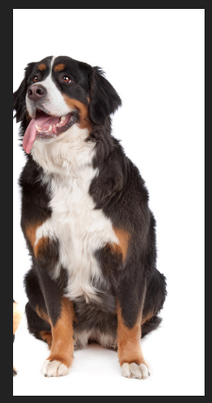
Qing & Franke (2014)

Lassiter & Goodman (2015)

GRADABILITY

- Modification can occur at varying intensities
  - Can we automatically learn adjective intensity?
- Individual adjectives have ranges they can apply to
  - How do we determine these ranges and their cutoffs?

- Information –> word choice



huge

large

?

de Melo & Bansal (2013)

Qing & Franke (2014)

Lassiter & Goodman (2015)

GRADABILITY

- Modification can occur at varying intensities
  - Can we automatically learn adjective intensity?
- Individual adjectives have ranges they can apply to
  - How do we determine these ranges and their cutoffs?

- Information –> word choice
- Word –> interpretation

de Melo & Bansal (2013)

Qing & Franke (2014)

Lassiter & Goodman (2015)

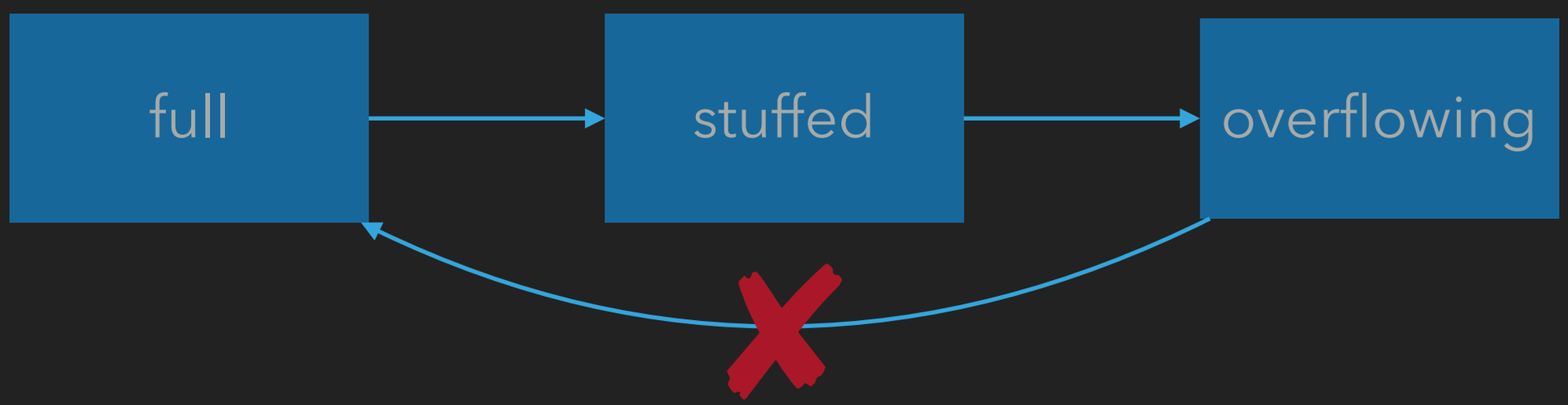
INFERRING SEMANTIC INTENSITIES

de Melo & Bansal (2013)

- **Goal**: Automatically learn adjective scales
- Use known syntactic patterns to collect word pairs

e.g. '★ (,) but not ★' ; 'not ★ (,) though still ★'

- Generate weak-strong scores for each word pair based on pattern counts



INFERRING SEMANTIC INTENSITIES

de Melo & Bansal (2013)

- **Data**: WordNet & Web Scraping
- **Results**:

Method	Pairwise Accuracy	Avg. τ	Avg. $\lvert\tau\rvert$	Avg. ρ	Avg. $\lvert\rho\rvert$
Web Baseline	48.2%	N/A	N/A	N/A	N/A
Divide-and-Conquer	50.6%	0.45	0.53	0.52	0.62
Sheinman and Tokunaga (2009)	55.5%	N/A	N/A	N/A	N/A
MILP	69.6%	0.57	0.65	0.64	0.73
MILP with synonymy	78.2%	0.57	0.66	0.67	0.80
Inter-Annotator Agreement	78.0%	0.67	0.76	0.75	0.86

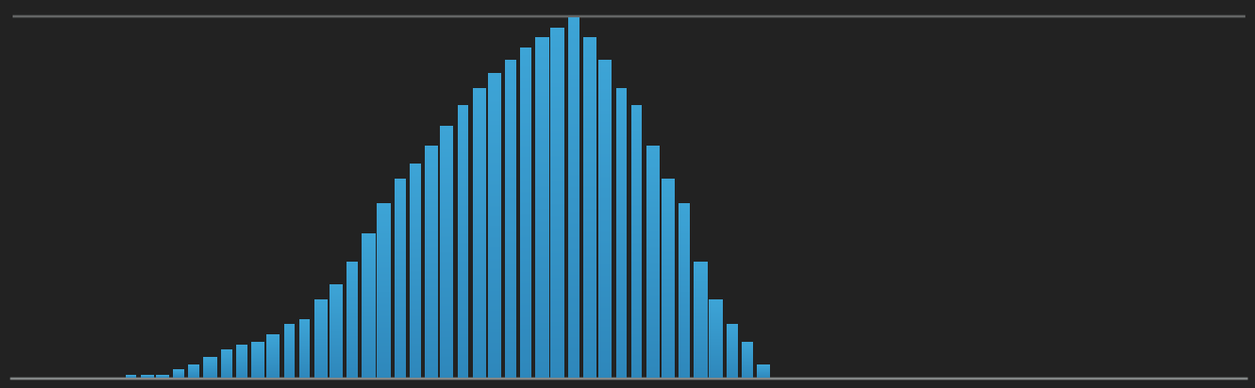
tiny → small → large → huge

- **Limitations**: No sense as to scope of individual words

SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- An adjective is used when the property described exceeds a threshold
  - Ex: A cookie is 'large' if its diameter is more than 4 inches
- Depends on 'comparison class': large cookie vs. large tree
- **Vagueness:** threshold is uncertain, *even with perfect knowledge*



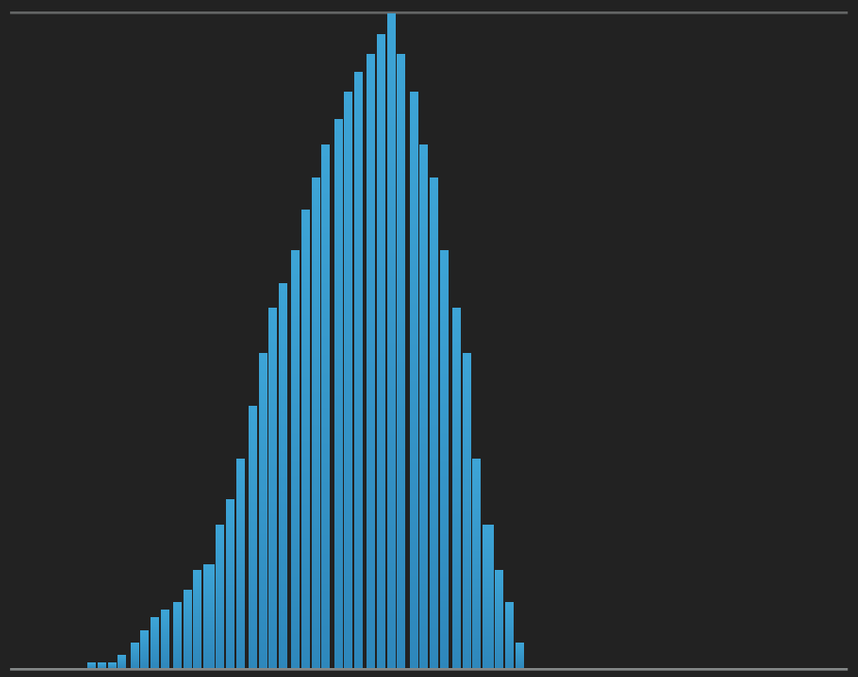
Sizes of cookies I've baked

- **Goal**: Model word usage as probability –> understand vagueness

SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- "I made a **large** cookie"
- Word use is [ideally] efficient: minimal effort accurate statement
- Speaker model: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta) \mathrm{d}\theta$



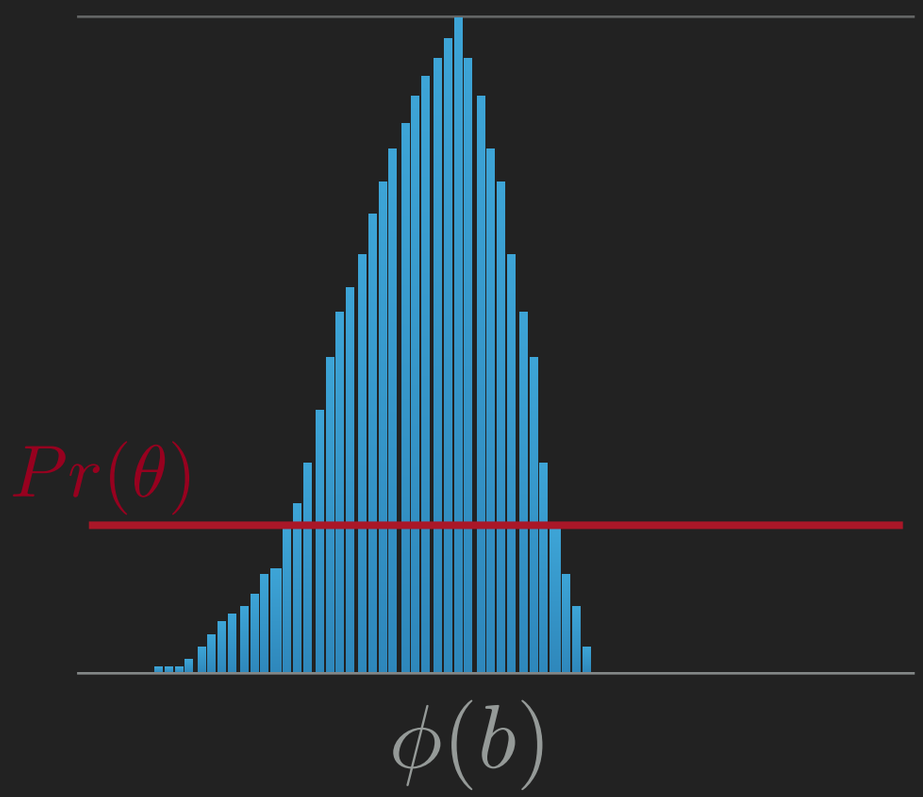
Sizes of cookies I've baked

SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- "I made a **large** cookie"
- Word use is [ideally] efficient: minimal effort accurate statement
- Speaker model: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta) \mathrm{d}\theta$

LITERAL LISTENER L_0

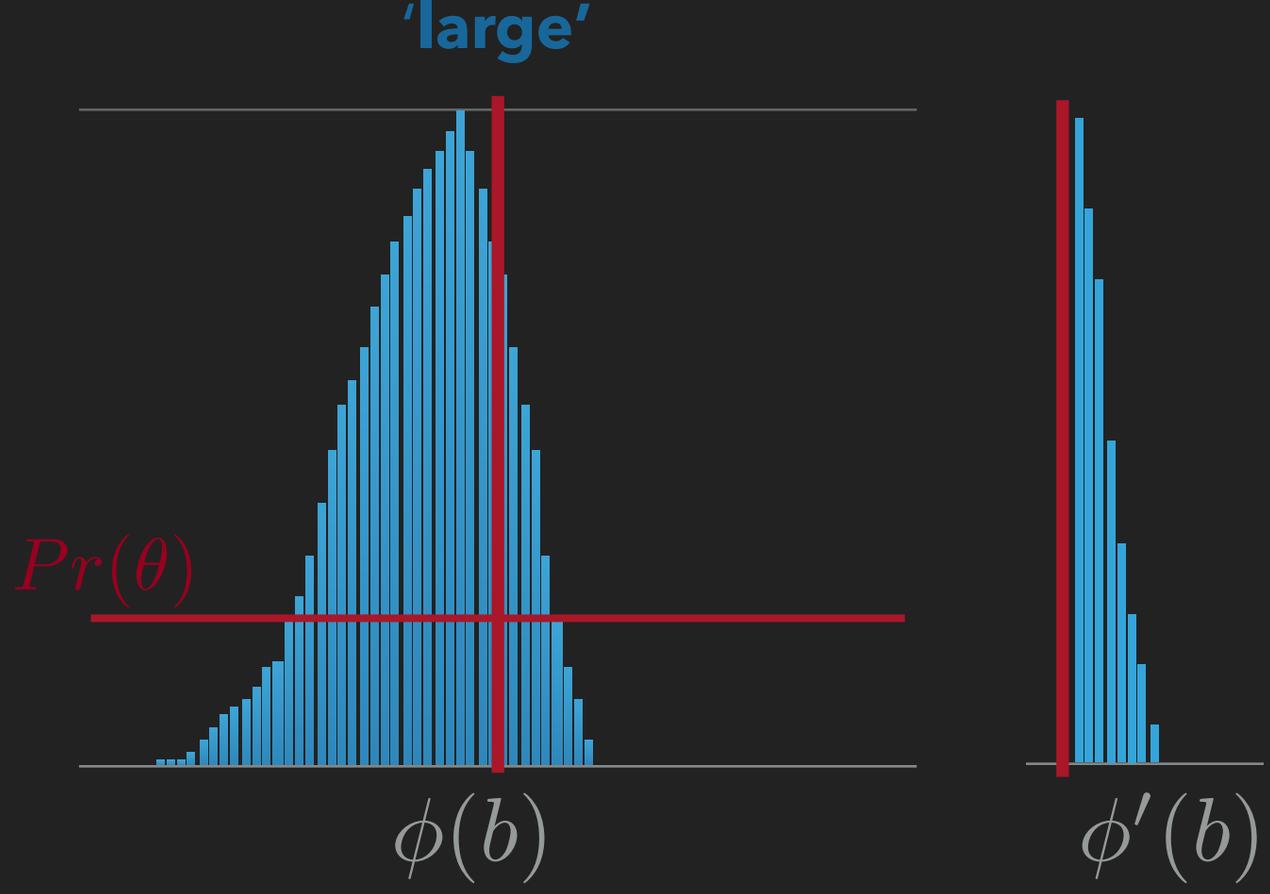


SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- "I made a **large** cookie"
- Word use is [ideally] efficient: minimal effort accurate statement
- Speaker model: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta) \mathrm{d}\theta$

LITERAL LISTENER L_0



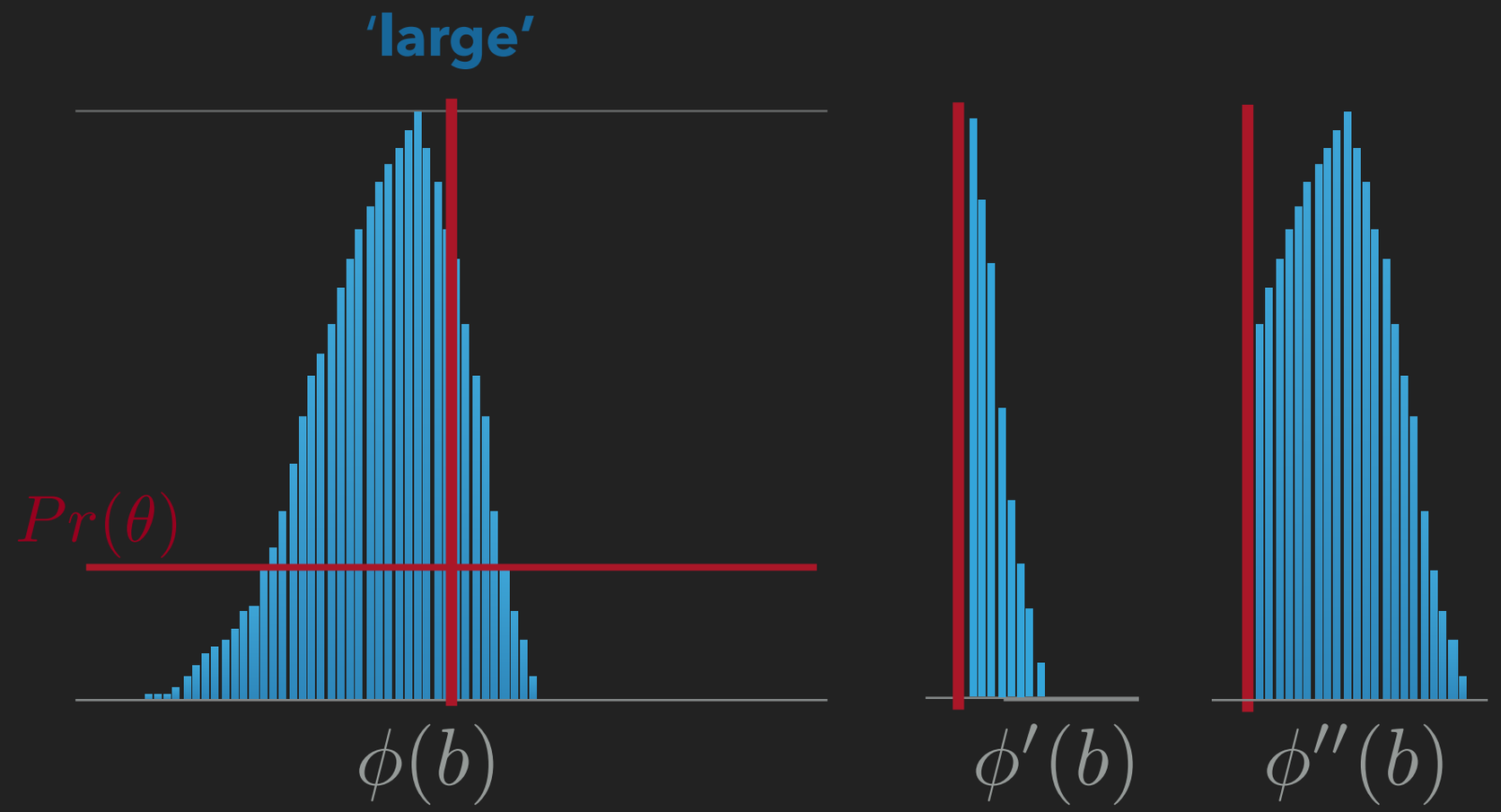
SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- "I made a **large** cookie"
- Word use is [ideally] efficient: minimal effort accurate statement
- Speaker model: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta) \mathrm{d}\theta$

PRAGMATIC SPEAKER S_1

LITERAL LISTENER L_0



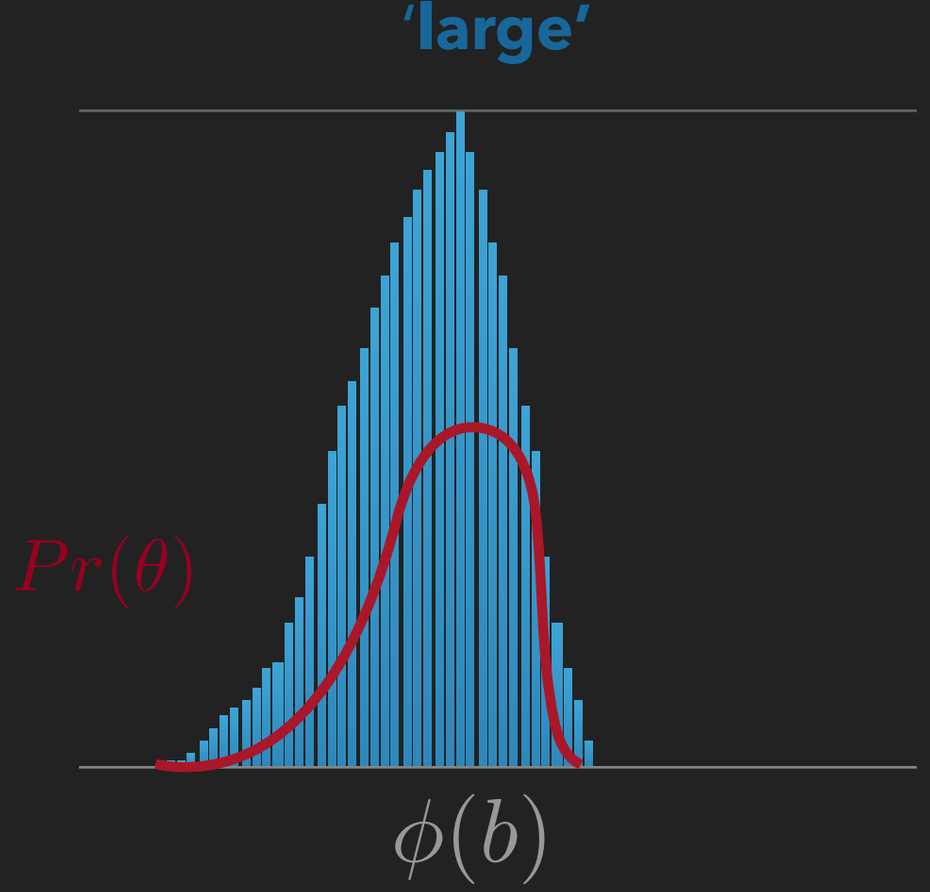
SPEAKER-ORIENTED MODEL

Qing & Franke (2014)

- "I made a **large** cookie"
- Word use is [ideally] efficient: minimal effort accurate statement
- Speaker model: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta) \mathrm{d}\theta$

PRAGMATIC SPEAKER S_1

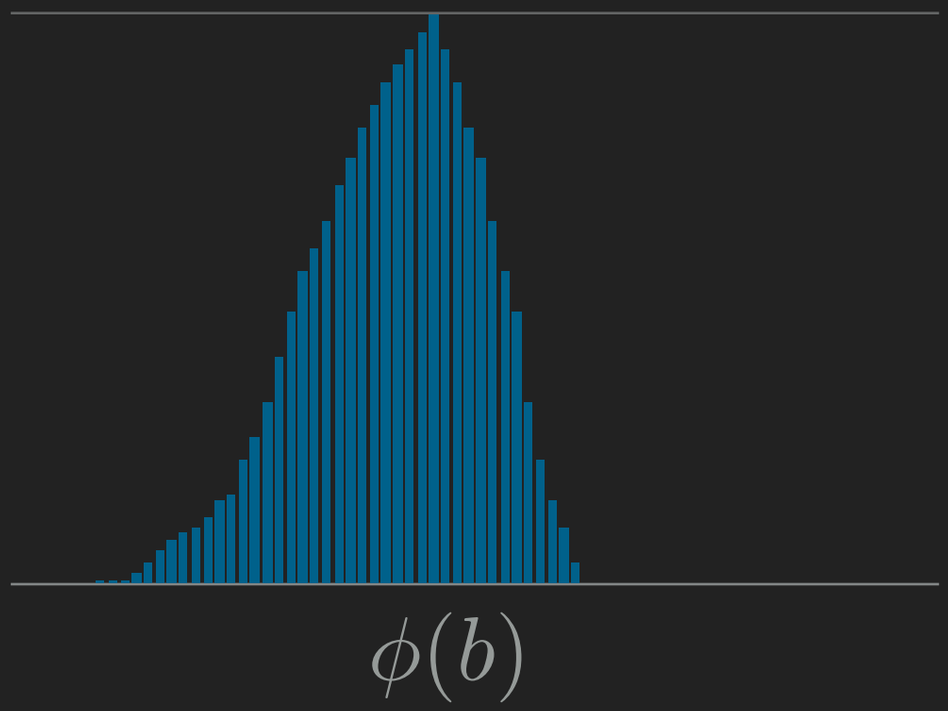
LITERAL LISTENER L_0



SPEAKER-LISTENER INTERACTION

Qing & Franke (2014)

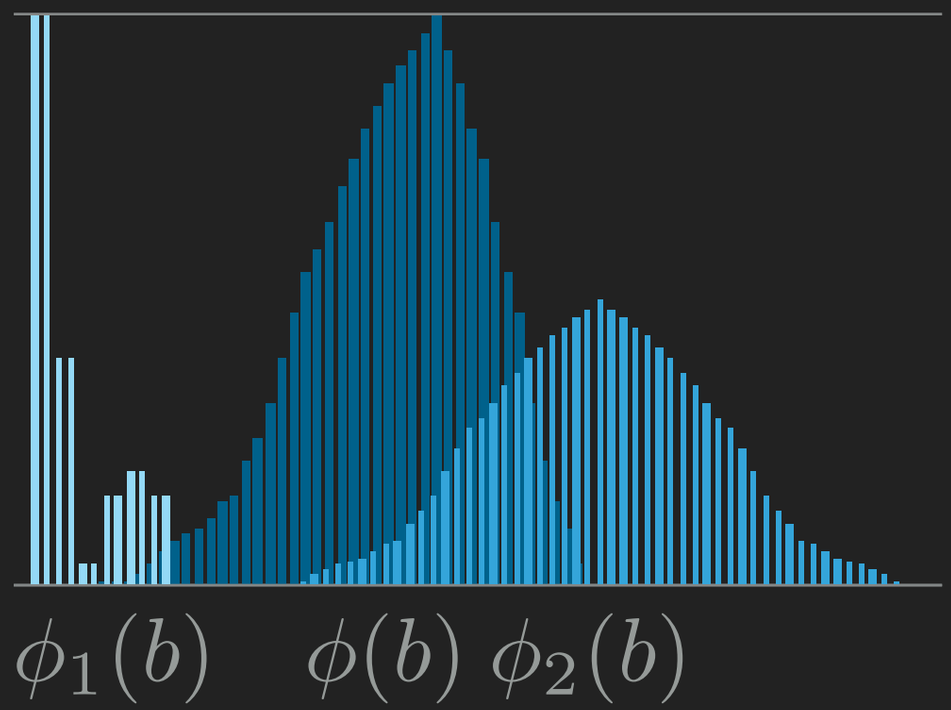
- "I made a **large** cookie"



SPEAKER-LISTENER INTERACTION

Qing & Franke (2014)

- "I made a **large** cookie"

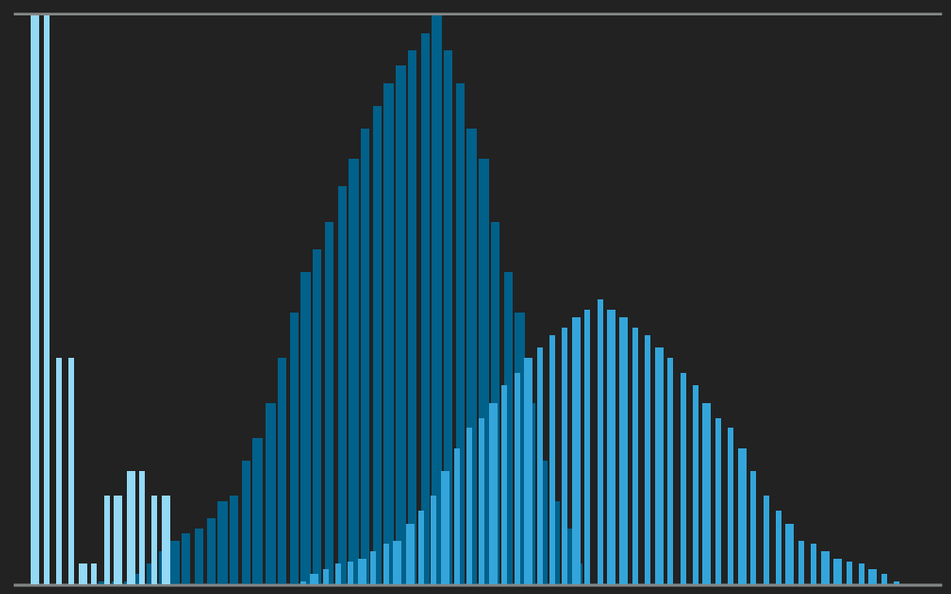


- **Limitation**: What happens when the priors are different?

SPEAKER–LISTENER INTERACTION

Lassiter & Goodman (2017)

- "I made a **large** cookie"



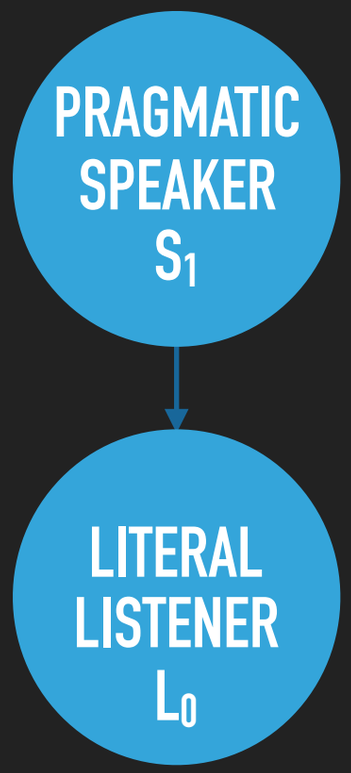
- **Limitation**: What happens when the priors are different?
- **Goal**: How do we interpret the use of an adjective?

$$P_S(u|w) \propto P_L(w|u) \cdot P_S(u) \qquad P_L(w|u) \propto P_S(u|w) \cdot P_L(w)$$

LISTENER MODEL

Lassiter & Goodman (2017)

- “I made a **large** cookie”



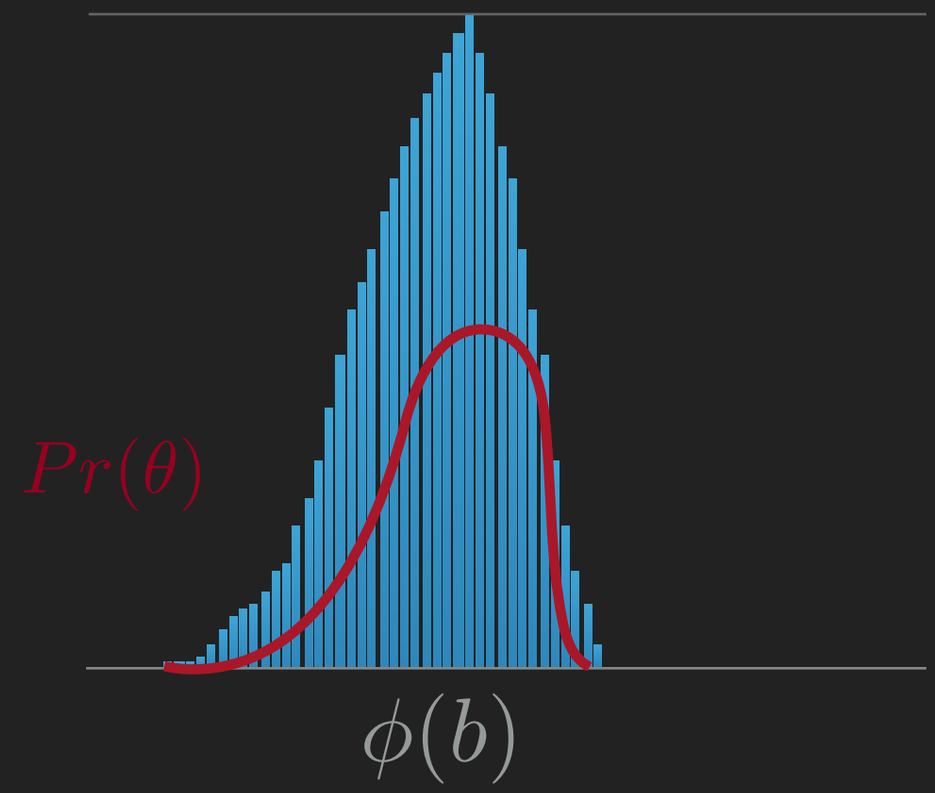
{∅, ‘small’, ‘large’}

$$P_{L_0}(A|u,V) = P_{L_0}(A[[u]]^V = 1)$$

LISTENER MODEL

Lassiter & Goodman (2017)

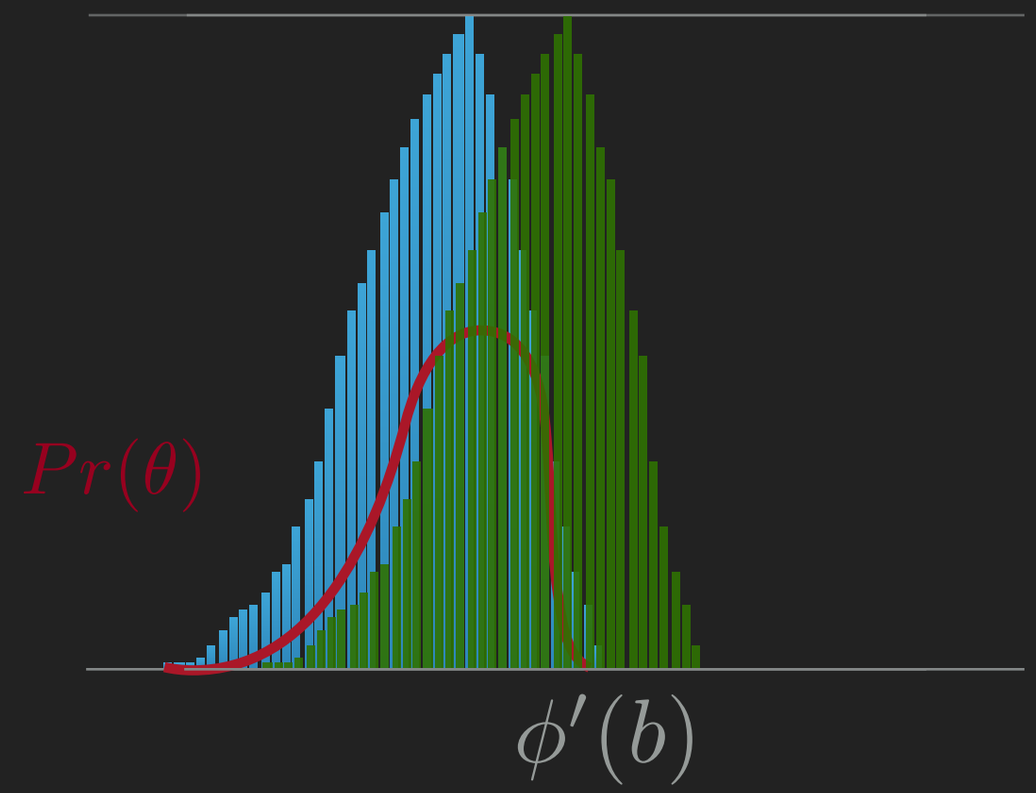
- "I made a **large** cookie"



LISTENER MODEL

Lassiter & Goodman (2017)

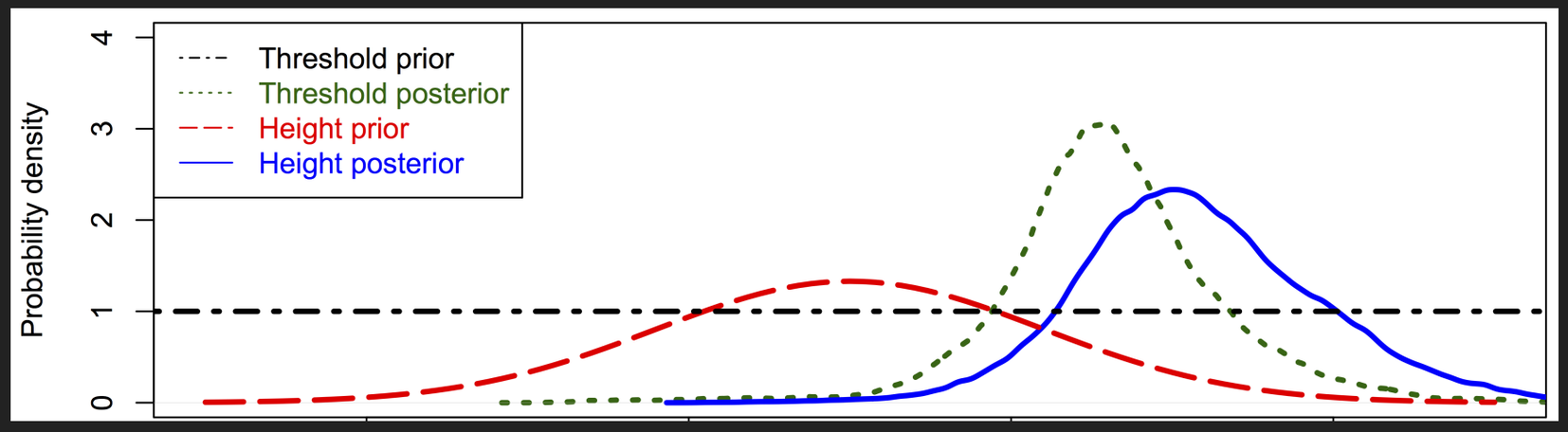
- "I made a **large** cookie"



LISTENER MODEL

Lassiter & Goodman (2017)

- "I made a **large** cookie"



Limitation: Model is theoretic

GRADABILITY TO COMPARATIVES

	Purpose	Result	Limitation
de Melo & Bansal (2013)	Ordering by intensity	**Automatic** from syntax	Do not know **range** of individual words
Qing & Franke (2014)	Modeling word use	Model **word usage**	Assumes basic listener with **same prior**
Lassiter & Goodman (2017)	Modeling interpretation	Model **word interpretation**	Model is **theoretic**

- How do we **grade** words in the context of visual information?
  - Individual words can have a range of interpretations, i.e. their *groundings* are variable
  - Multiple words can refer to the same visual feature

GRADABILITY TO COMPARATIVES

	Purpose	Result	Limitation
de Melo & Bansal (2013)	Ordering by intensity	Automatic from syntax	Do not know **range** of individual words
Qing & Franke (2014)	Modeling word use	Model	Assumes basic listener with **same prior**
Lassiter & Goodman (2017)	Modeling interpretation	Model	Model is **theoretic**

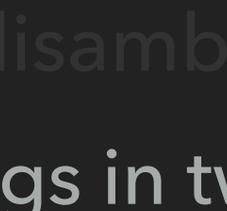
"white eye ring"

"[American Tree Sparrows] don't have as strong of a white eye ring as Field Sparrows"

- How do we **grade** words in the context of visual information?
  - Individual words can have a range of interpretations, i.e. their *groundings* are variable
  - Multiple words can refer to the same visual feature
- Need **context** to disambiguate, i.e. **compare**

GRADABILITY TO COMPARATIVES

	Purpose	Result	Limitation
de Melo & Bansal (2013)	Ord intensity	Automatic from syntax	Do not know **range** of individual words
Qing & Franke (2014)	Mo ord use	Model	Assumes basic listener with **same prior**
Lassiter & Goodman (2017)	Modeling interpretation	Model	Model is **theoretic**



"white eye ring"

"[American Tree Sparrows] don't have as strong of a white eye ring as Field Sparrows"

- How do we **grade** words in the context of visua
  - Individual words can have a range of interp their *groundings* are variable
  - Multiple words can refer to the same visual
- Need **context** to disambiguate, i.e. **compare**
- Example groundings in two common properties: color and size

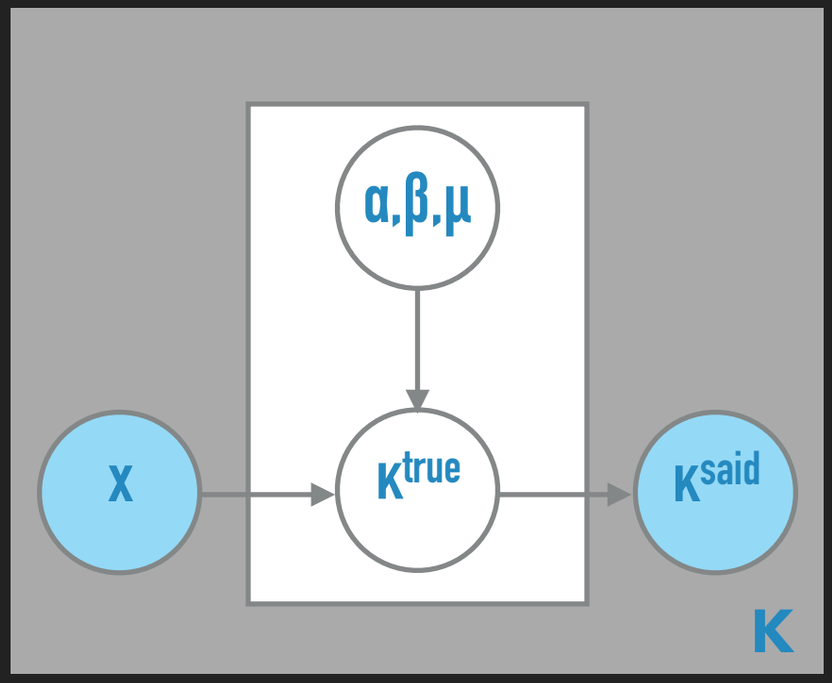
McMahan & Stone (2015)

Monroe et al (2017)

Bagherinezhad et al (2016)

GROUNDED COLOR SEMANTICS

McMahan & Stone (2015)



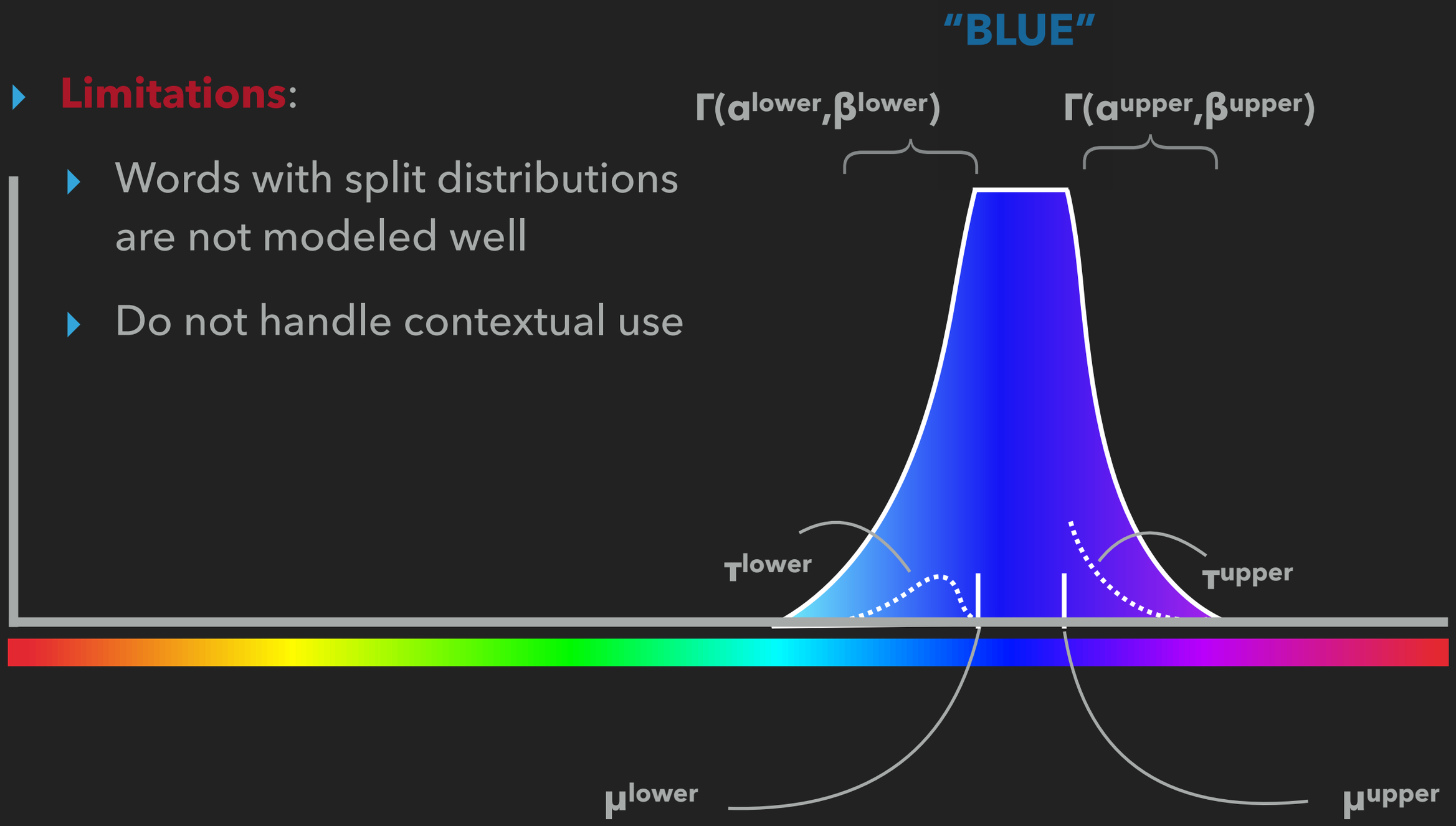
Applicability: $P(k^{true} \mid x)$

Availability: $P(k^{said} \mid k^{true})$

GROUNDED COLOR SEMANTICS

McMahan & Stone (2015)

- **Data**: XKCD online color survey
- **Limitations**:
  - Words with split distributions are not modeled well
  - Do not handle contextual use



GROUNDED COLORS IN CONTEXT

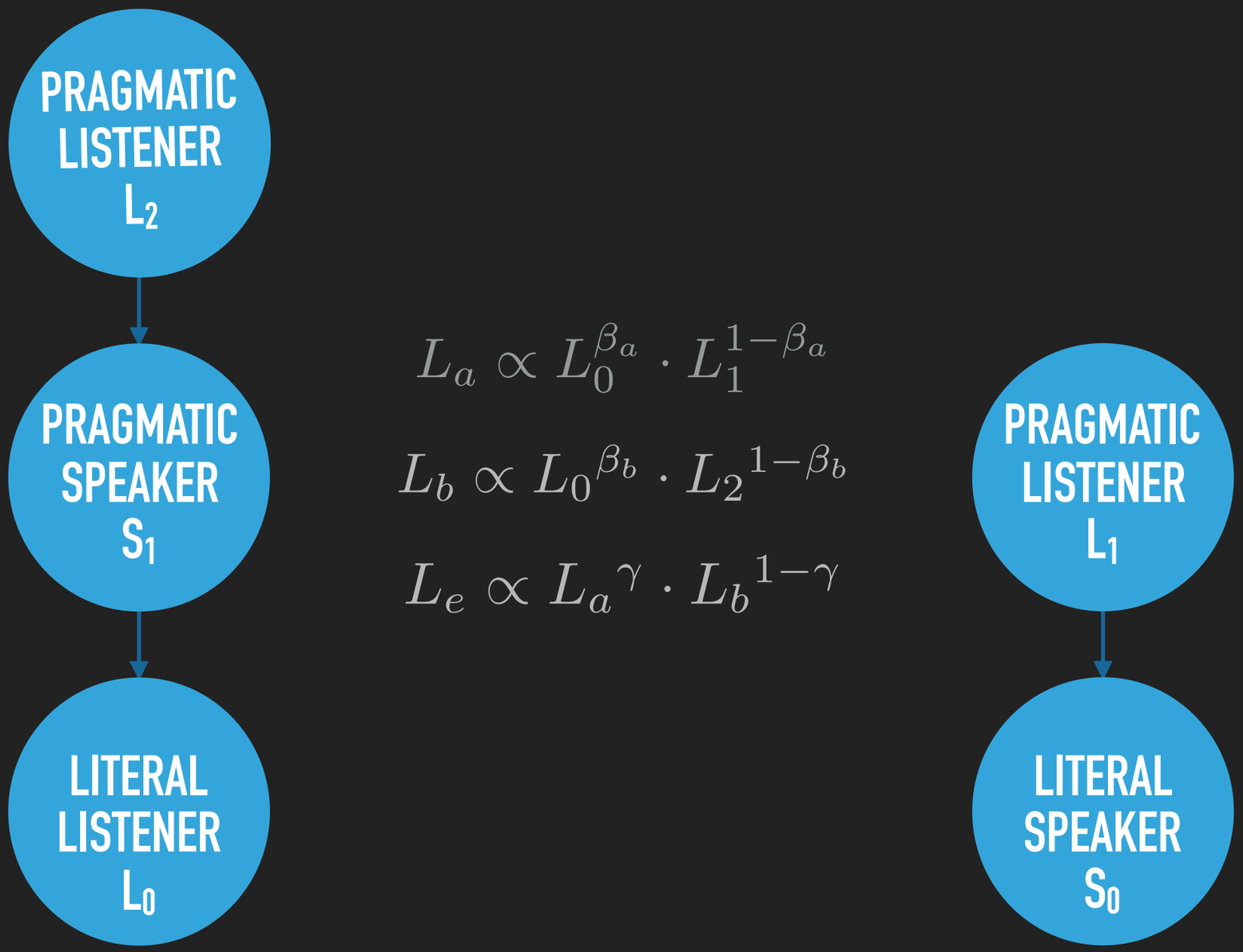
Monroe et al (2017)

	Context	Utterance
1.		darker blue
2.		Purple
3.		blue
4.		blue

- **Goal**: How is color label use affected by other colors present?
- **Data**:
  - Task: describe **target** color in context of **2 distractors**
  - Distractors could be **close**, **split**, or **far**
- Model: speaker/listener approach
  - Threshold now governed by contextual information

GROUNDED COLORS IN CONTEXT

Monroe et al (2017)



GROUNDED COLORS IN CONTEXT

Monroe et al (2017)

- **Results**:
  - Comparative terms used most often when one distractor is similar to the target

L_2			
blue	**68**	32	<1
S_0	*5.71*	***7.63***	*0.01*
L_1	43	**57**	<1
L_a	**50**	**50**	<1
L_b	**68**	32	<1
L_e	**59**	41	<1

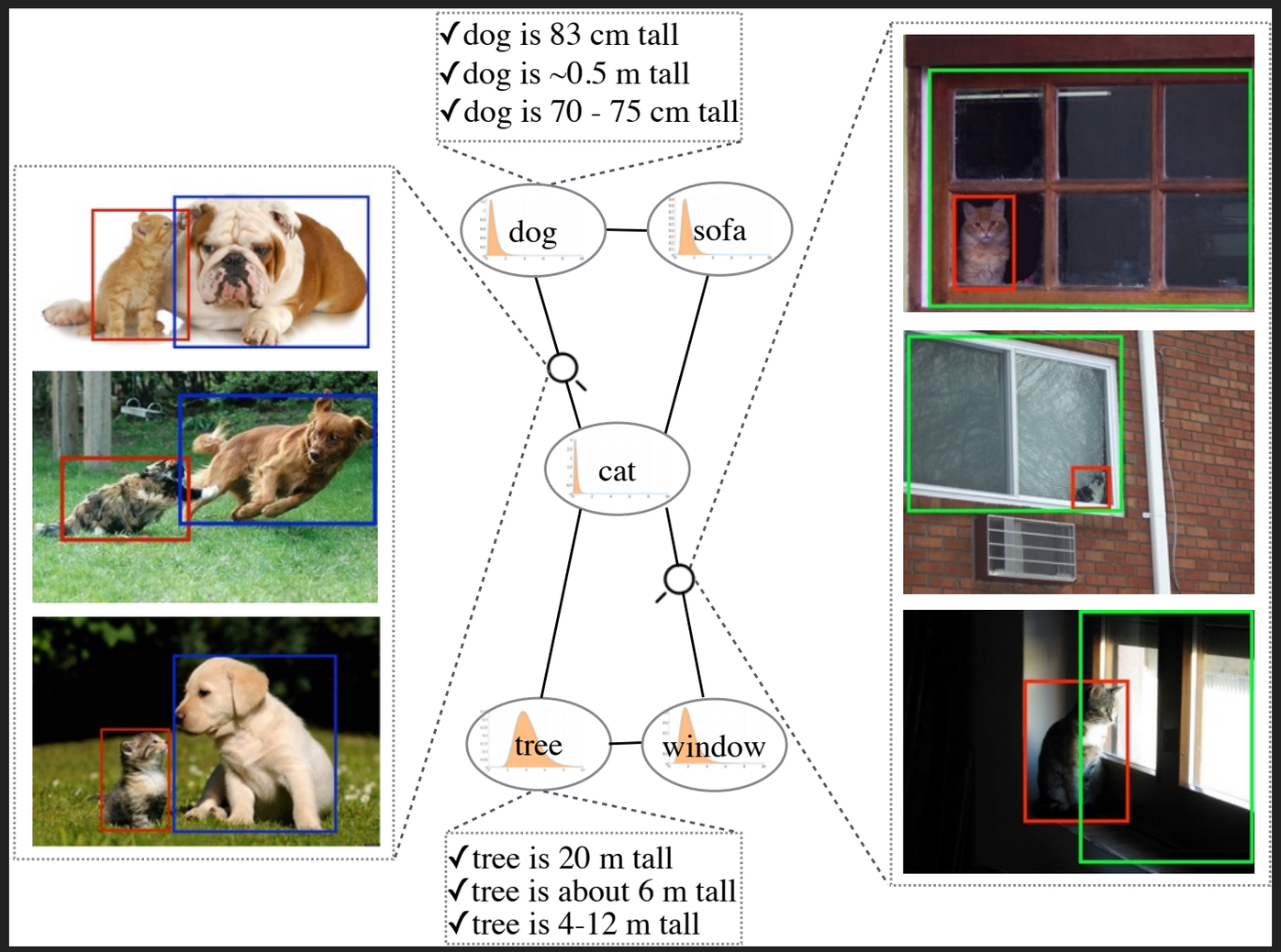
L_2			
drab green not the bluer one	5	<1	**95**
S_0 ($\times 10^{-9}$)	***5.85***	*0.38*	*<0.01*
L_1	**94**	6	<1
L_a	**92**	6	2
L_b	8	1	**91**
L_e	**63**	6	32

- **Limitation**:
  - Do not have representation of comparatives

COMPARATIVE SIZES

Bagherinezhad et al (2016)

- **Goal**: Use images to learn about object sizes
- Text absolute but incomplete; image information only relative

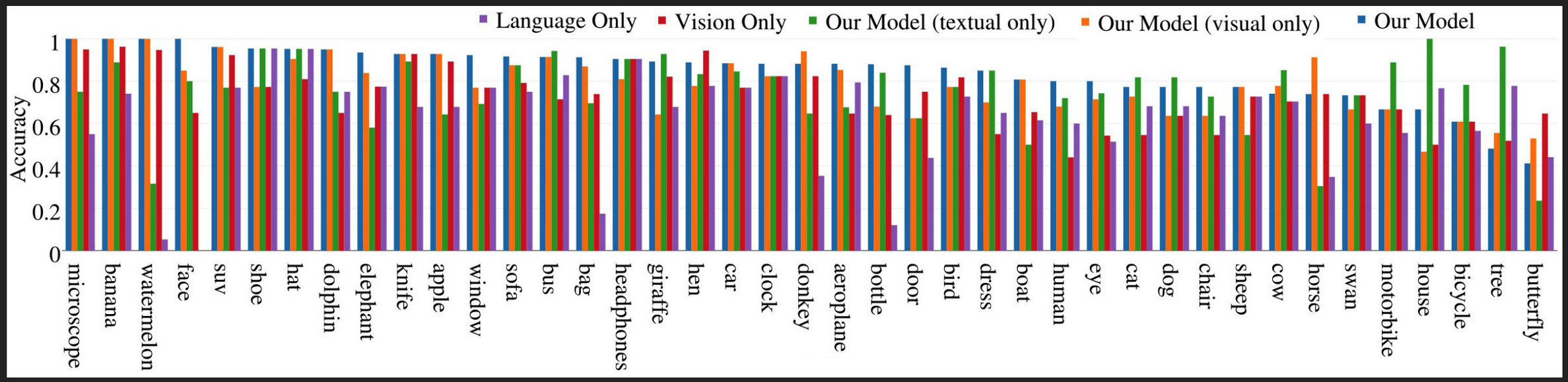


- **Data**: 41 objects, 486 object pairs ,100 Flickr images per pair

COMPARATIVE SIZES

Bagherinezhad et al (2016)

- **Results**: Minimal size information required for high accuracy



- Transitivity: size of chairs mostly affected by the size of cats
- **Limitation**: Difficult to handle objects with highly variable sizes
  Do not use comparative textual information

COMPARATIVES

	Purpose	Result	Limitation
McMahan & Stone (2015)	Vagueness of color	**Probabilistic color labels**	**Assume single cluster of word**
Monroe et al (2017)	Contextual color use	**Contextually based color understaning**	**No explicit comparisons**
Bagherinezhad (2016)	Automatic size understanding	**Learn absolute sizes from relative**	**Do not use comparatives**

- Both useful and limiting that each approach focused on a single property, using property-specific representations of the data
- A combination of both global (absolute) information and local (relative) details are necessary to properly contextualize descriptions
- As of yet, are not handling comparative adjectives themselves, only comparing and contextualizing

QUANTIFIERS

- Adjectives don't always apply to all instances of an object
  - '**Most** field sparrows have a white eye ring'
  - '**Some** dogs are large'

1. Can quantifiers be incorporated into representations?

Herbelot & Vecchi (2015) Sorodoc et al (2016) Pezzelle et al (2017)

QUANTIFIERS

- Adjectives don’t always apply to all instances of an object
  - ‘**Most** field sparrows have a white eye ring’
  - ‘**Some** dogs are large’

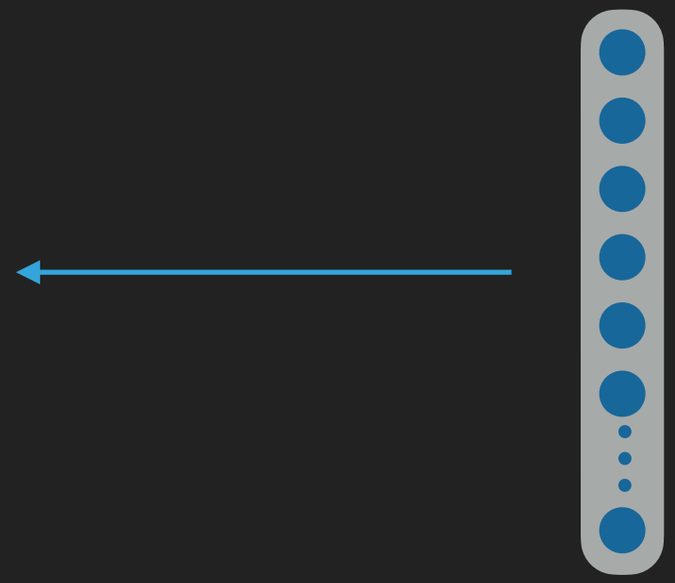
1. Can quantifiers be incorporated into representations?
2. Can quantifiers be grounded?

Herbelot & Vecchi (2015) | Sorodoc et al (2016) | Pezzelle et al (2017)

QUANTIFIERS IN MODEL-THEORETIC SPACE Herbelot & Vecchi (2015)

- **Model Theoretic Space:** Objects are vectors where each dimension equals the proportion of attribute possession

HORSE	
a_mammal	1
has_four_legs	0.95
is_brown	0.35
is_scaly	0



- **Goal**: Learn quantifiers through linear map from existing distributional spaces to model-theoretic space

QUANTIFIERS IN MODEL-THEORETIC SPACE Herbelot & Vecchi (2015)

- **Data**: QMR & AD (quantifiers), Wikipedia & Google News
- **Results**:
  - Training and testing on animals yields best mapping

Instance	Mapped	Gold
raven a_bird	most	all
pigeon has_hair	few	no
elephant has_eyes	most	all
crab is_blind	few	few
snail a_predator	no	no

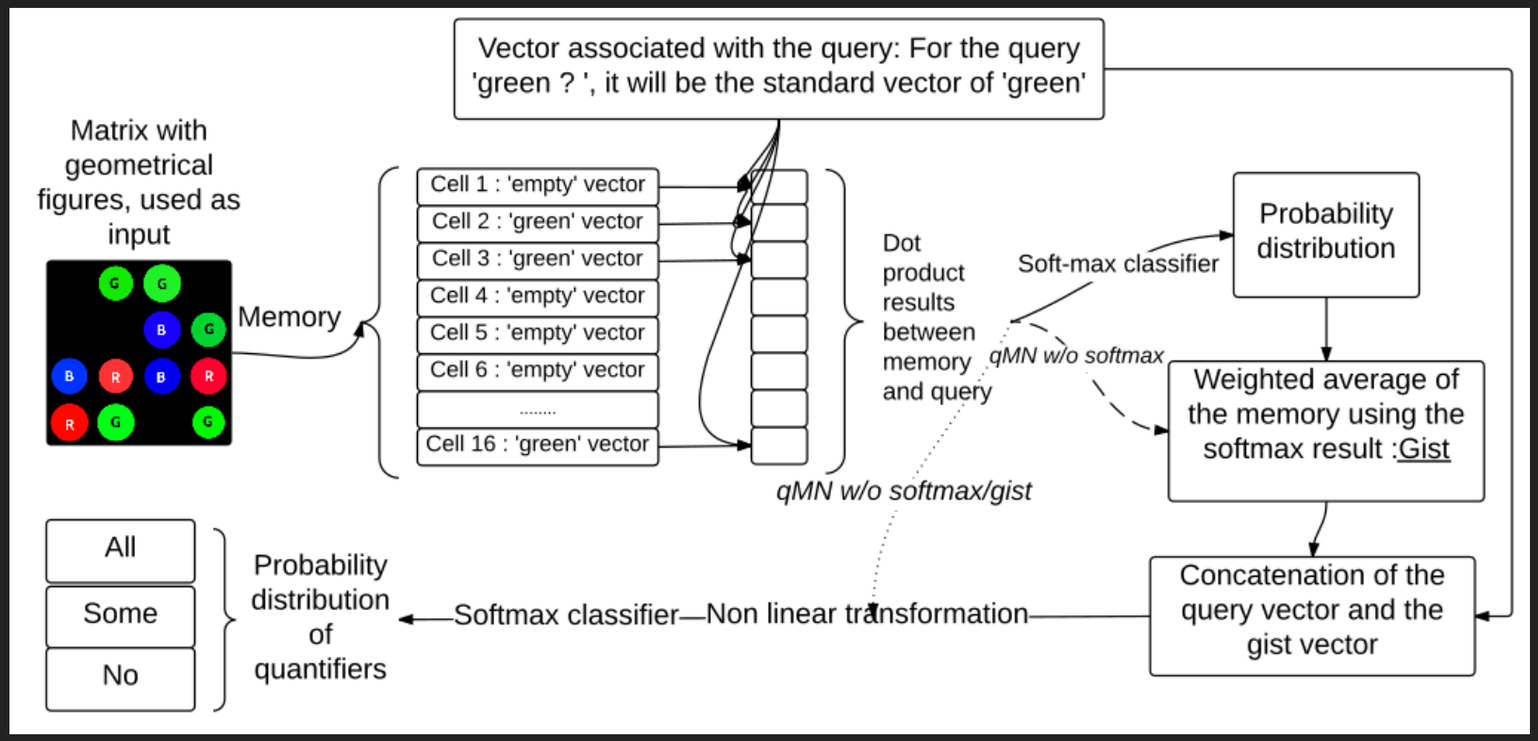
plum	*cottage*
a_fruit	has_a_roof
grows_on_trees	used_for_shelter*
tastes_sweet	has_doors*
is_citrus*	worn_on_feet*

axe	*hatchet*
a tool	a tool
is sharp	is sharp
has a handle	has a handle
used for cutting	used for cutting
has a metal blade	made of metal
a weapon	an axe
has a head	is small
used for chopping	–
has a blade	–
is dangerous	–
is heavy	–
used by lumberjacks	–
used for killing	–

- **Limitation**: Missing data negatively affects mapping
  No contextual dependency - *when/where* do differences occur

QUANTIFIERS IN IMAGES

Sorodoc et al (2016)



- **Goal**: Grounded quantification
- **Data**: Generated images of colored circles
- **Results**: Proportion-based method outperforms count-based method
- **Limitations**: Highly controlled images and limited queries

QUANTIFIERS IN IMAGES

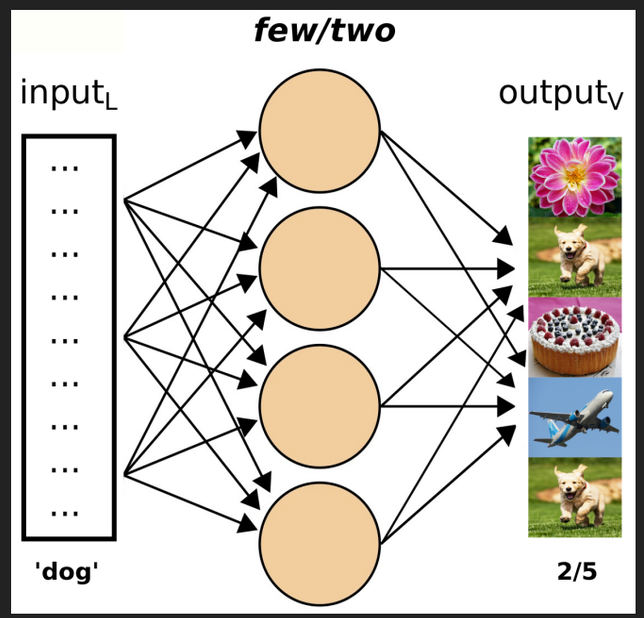
Pezzelle et al (2017)

How many are dogs? *Three/most*

- A person recognizes both small numbers and proportions
- **Goal**: Map from text to image learn quantifiers from varying proportions

QUANTIFIERS IN IMAGES

Pezzelle et al (2017)



- A person recognizes both small numbers and proportions
- **Goal**: Map from text to image learn quantifiers from varying proportions
- **Data**: ImageNet images constructed into collages
- **Result**: Quantifiers and cardinals require different similarity measures (cos similarity and dot product, respectively)
- **Limitation**: Restricted learning space

QUANTIFIERS

	Purpose	Result	Limitation
Herbelot & Vecchi (2015)	Use MT to learn quantified attributes	**Partial attributes** can be inferred	Assumes global truth (**not contextual**)
Sorodoc et al (2016)	Ground quantifiers in image data	**Proportional** approach	**Count-based** representation
Pezzelle et al (2017)	Ground quantifiers and cardinals	**Separate metric** from counting	**Restricted data**

- Quantifiers in images:
  - Correspond to proportions
  - Can be learned alongside cardinals
- Future work: Applying quantifiers to grounded classification methods

‘**mostly** white’

QUANTIFIERS

	Purpose	Result	Limitation
Herbelot & Vecchi (2015)	Use MT to learn quantified attributes	**ibutes** can be inferred	Assumes global truth (**not contextual**)
Sorodoc et al (2016)	Ground quantifiers in image data	**onal** approach	**Count-based** representation
Pezzelle et al (2017)	Ground quantifiers and cardinals	**metric** from counting	**Restricted data**

“white” + “eye” + “ring”

- Quantifiers in images:
  - Correspond to proportions
  - Can be learned alongside cardinals
- Future work: Applying quantifiers to grounded classification methods

‘mostly white’

- How do we put everything together?

OUTLINE

- Methodology
- Modifiers
- **Compositionality in Distributional Semantics**
  - Language
  - Language & Vision

COMPOSITIONALITY IN DISTRIBUTIONAL SEMANTICS

Mitchell & Lapata (2010)

Baroni & Zamparelli (2010)

Hartung et al (2017)

Boleda et al (2013)

Vecchi et al (2013)

Dunlop et al (2010)

- Linguistic aspects:
  - Composition methods

“black” ? **“tail”** = **“black tail”**

COMPOSITIONALITY IN DISTRIBUTIONAL SEMANTICS

Mitchell & Lapata (2010)

Baroni & Zamparelli (2010)

Hartung et al (2017)

Boleda et al (2013)

Vecchi et al (2013)

Dunlop et al (2010)

- Linguistic aspects:
  - Composition methods — **“black”** ? **“tail”** = **“black tail”**
  - Property being described — **“black”** –> color, emotion, legality…

COMPOSITIONALITY IN DISTRIBUTIONAL SEMANTICS

Mitchell & Lapata (2010)

Baroni & Zamparelli (2010)

Hartung et al (2017)

Boleda et al (2013)

Vecchi et al (2013)

Dunlop et al (2010)

- Linguistic aspects:
  - Composition methods — **“black”** ? **“tail”** = **“black tail”**
  - Property being described — **“black”** –> color, emotion, legality…
  - Intensionality — **“alleged murderer”** ≠ **“alleged”** & **“murderer”**

COMPOSITIONALITY IN DISTRIBUTIONAL SEMANTICS

Mitchell & Lapata (2010)

Baroni & Zamparelli (2010)

Hartung et al (2017)

Boleda et al (2013)

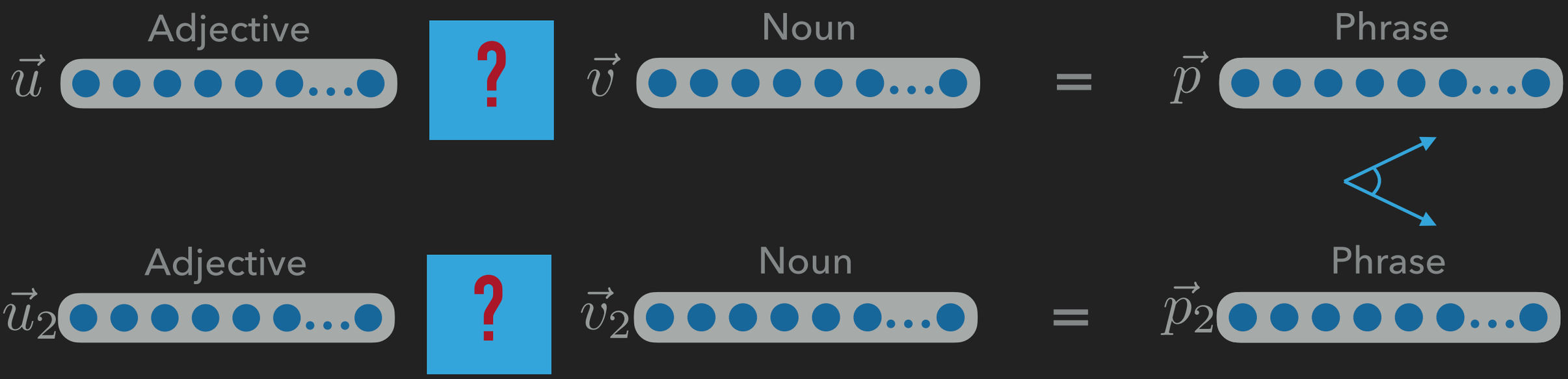
Vecchi et al (2013)

Dunlop et al (2010)

- Linguistic aspects:
  - Composition methods — **“black”** ? **“tail”** = **“black tail”**
  - Property being described — **“black”** –> color, emotion, legality…
  - Intensionality — **“alleged murderer”** ≠ **“alleged”** & **“murderer”**
  - Ordering — **“light grey bag”** vs **“grey light bag”**

ADJECTIVE–NOUN PHRASES

Mitchell & Lapata (2010)



- **Goal**: Find composition function that optimizes similarity between composed vectors, depending on representation

ADJECTIVE–NOUN PHRASES

Mitchell & Lapata (2010)

- 2 semantic spaces:
  1. context co-occurence

$$v_i(t) = \frac{p(c_t|t)}{p(c_i)}$$

  2. LDA topic proportions

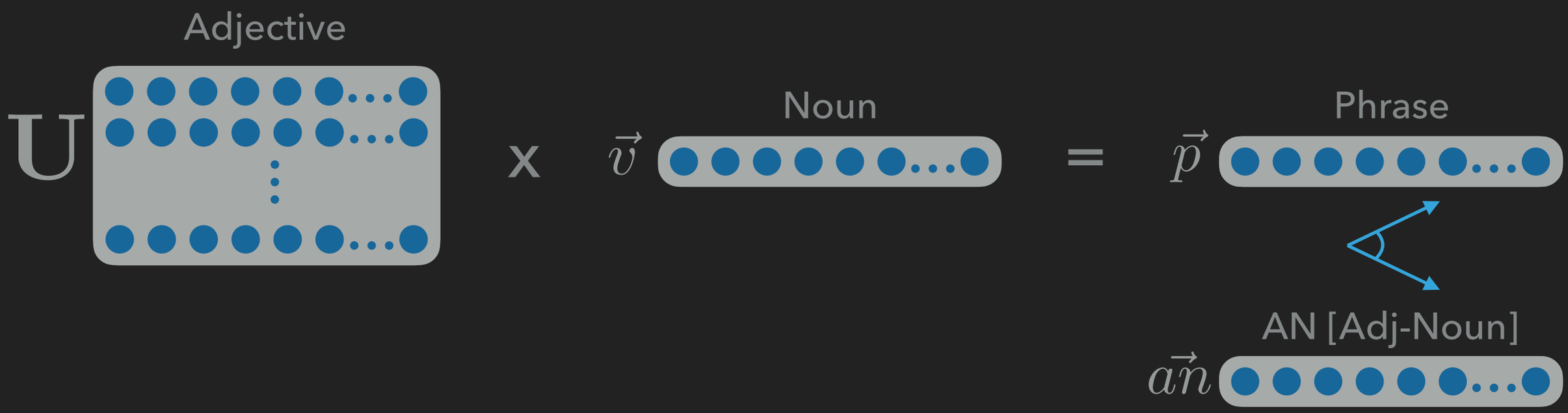
$$\beta_{ij} = p(w_i|z_j)$$

Spearman's ρ	Context-based	LDA
Additive	.36	.37
Kintsch	.32	.30
Multiplicative	.46	.25
Tensor product	.41	.39
Convolution	.09	.15
Weighted additive	.44	.38
Dilation	.44	.38
Target unit	.43	
Head only	.43	.35
Humans	.52	.52

- **Data**: BNC corpus
- **Results**: Multiplication was best for context-based vectors, but additive functions are best overall
- **Limitation**: Only measuring similarity between the constructed vectors

ADJECTIVE–NOUN PHRASES

Baroni & Zamparelli (2010)



- Adjectives transform noun to noun-phrase
- Noun-phrases are corpus-generated vectors
- **Goal**: Learn adjectives as functions over nouns

ADJECTIVE–NOUN PHRASES

Baroni & Zamparelli (2010)

- **Data**: Wikipedia + BNC
- **Results**:
  - Composed vectors are semantically related to corpus-derived phrase vectors

    'young' * 'man' –> 'small son', 'small daughter', 'mistress'

  - Adjectives cluster well based on property described
- **Limitations**: Vector space derived from dimensionality reduction using only most common words

ADJECTIVE–NOUN CLASSES

Hartung et al (2017)

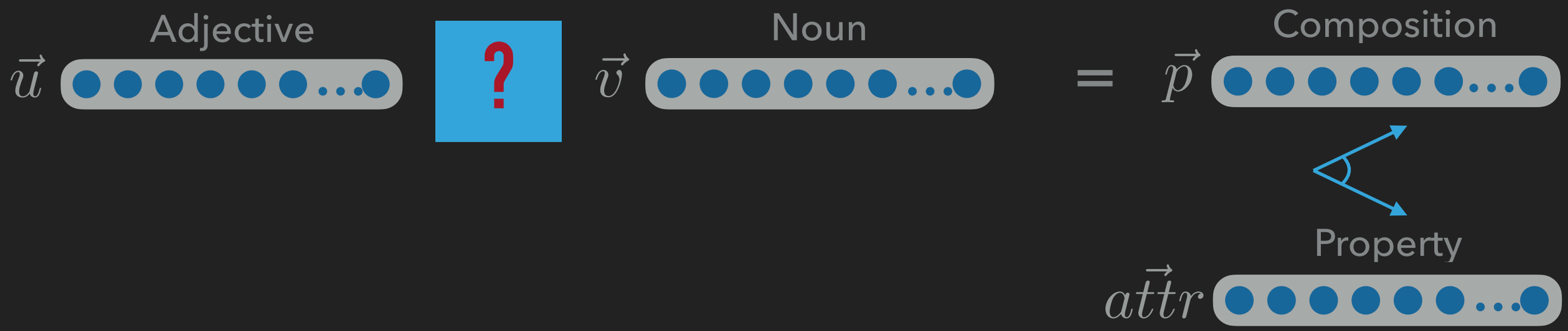
"hot summer"

temperature

"hot debate"

emotion

- Property being described depends on noun in phrase
- Properties have names: these are also nouns
- Find correct property through composition



ADJECTIVE–NOUN CLASSES

Hartung et al (2017)

- **Data**: HeiPLAS adj-property-noun triples (Hartung 2015)
  Google News word2vec
- **Results**:
  - Weighted addition is best
- **Limitations**:
  - Probability-based spaces do not work
  - Property can be context-dependent

	Compositional Model	P@1	P@5
predict models	Adjective	0.33	0.50
	Noun	0.03	0.10
	Vector Addition ($\oplus$)	0.24	0.45
	Weighted Vector Addition	0.33	0.51
	Vector Multiplication ($\odot$)	0.00	0.02
	Adj. Dilation ($\lambda = 2$)	0.06	0.18
	Noun Dilation ($\lambda = 2$)	0.33	0.51
	Full Add. Weighted Noun	0.33	0.54
	Full Add. Weighted Adjective	0.46	0.71
	Full Add. Weighted Adj. and Noun	**0.56**	**0.75**
	Trained Tensor Product ($\otimes$)	0.44	0.57
count	C-LDA (Hartung, 2015)	0.09	n/a
	L-LDA (Hartung, 2015)	0.16	n/a

INTENSIONALITY

Boleda et al (2013)

- Intersective:
  - A "white towel" **is** both white and a towel
- Subsective:
  - A "skilled surgeon" is **not necessarily** skilled in general
- Intensional:
  - An "alleged murderer" **is not** a murderer (nor 'alleged')
- **Goal**: Determine if compositional methods are affected by intensionality of adjective

INTENSIONALITY

Boleda et al (2013)

Data: Wikipedia + BNC

I	alleged	former	future	hypothetical	impossible	likely	mere	mock
N	loose	wide	white	naive	severe	hard	intelligent	ripe
I	necessary	past	possible	potential	presumed	probable	putative	theoretical
N	modern	black	free	safe	vile	nasty	meagre	stable

Model	Global	Intensional	Non-intensional	NN=A	NN=N
observed	-	-	-	*8.2*	*3.3*
lexical function	**0.60**±0.11	**0.60**±0.10	**0.60**±0.10	0.9	0.6
full additive	0.52±0.13	0.52±0.13	0.51±0.12	10.0	4.8
weighted additive	0.48±0.14	0.48±0.14	0.48±0.14	23.2	13.3
dilation	0.42±0.18	0.42±0.17	0.42±0.17	31.0	11.6
multiplicative	0.32±0.21	0.32±0.20	0.32±0.20	29.9	16.6
noun only	*0.40*±*0.18*	*0.40*±*0.17*	*0.40*±*0.17*	-	-

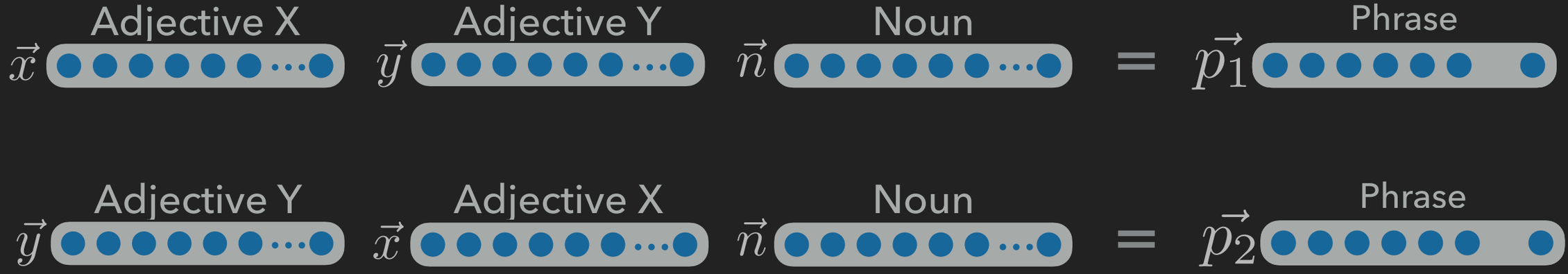
Predicted-to-observed vector

Limitation: Do not compose multiple adjectives

ADJECTIVE ORDERING

Vecchi et al (2013)

- Syntax makes adjective ordering easy to learn (Dunlop 2010)
- **Goal**: Understand adjective ordering in distributional space as function of adjective modification strength



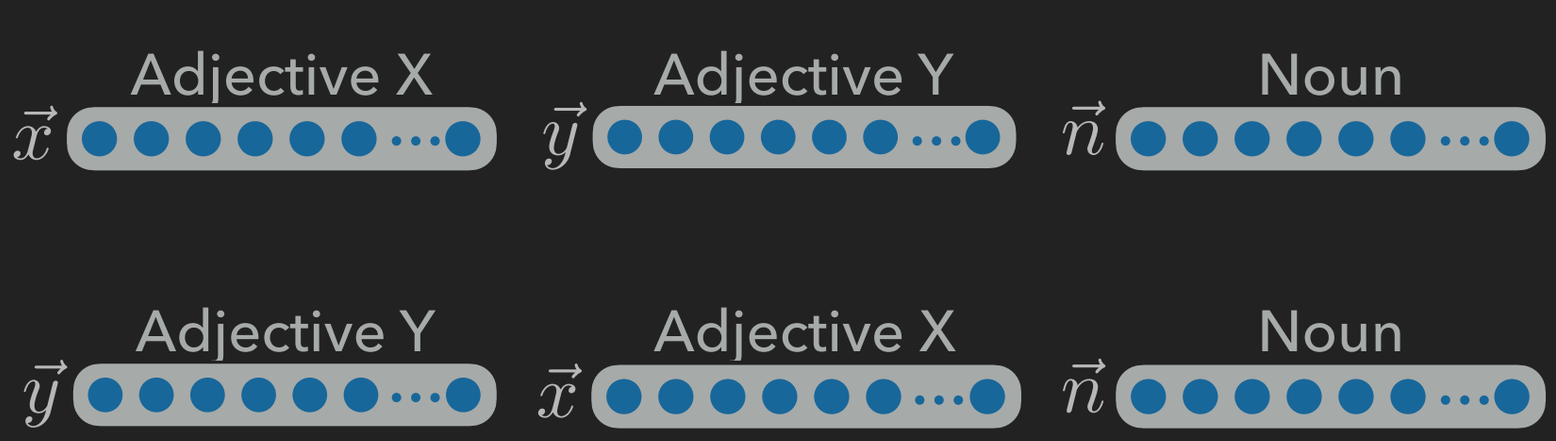
$$\cos\angle : \vec{x}, \vec{y}, \vec{n}, \vec{x} \cdot \vec{n}, \vec{y} \cdot \vec{n}$$

- **Data**: Wikipedia + BNC with dimension reduction
- **Limitations**: Treat flexible ordering as equivalent meaning

ADJECTIVE ORDERING

Vecchi et al (2013)

- Syntax makes adjective ordering easy to learn (Dunlop 2010)
- **Goal**: Understand adjective ordering in distributional space as function of adjective modification strength



$$\cos\angle : \vec{x}, \vec{y}, \vec{n}, \vec{x} \cdot \vec{n}, \vec{y} \cdot \vec{n}$$

Model	ρ	*M&L*
CORP	0.41	0.43
W.ADD	0.41	0.44
F.ADD	0.40	–
MULT	0.33	0.46
LFM	0.40	–

	Gold	*FO*	*RO*
W.ADD	0.565	0.572	0.558
F.ADD	0.618	0.622	0.614
MULT	0.424	0.468	0.384
LFM	0.655	0.675	0.637

- **Data**: Wikipedia + BNC with dimension reduction
- **Limitations**: Treat flexible ordering as equivalent meaning

LINGUISTIC ASPECTS OF COMPOSITIONALITY

	Initial Data	Goal	[Best] Method
Mitchell & Lapata (2010)	Adjective, noun vectors	Examine composition methods	Multiplication
Baroni & Zamparelli (2010)	Noun, noun phrase vectors	Represent adjective as matrix	Linear mapping
Hartung et al (2017)	Adjective, noun, property name vectors	Learn adjective property from composition	Weighted addition
Boleda et al (2013)	Adjective, noun, phrase vectors	Effect of intensionality on composition function	Lexical Function (linear mapping)
Vecchi et al (2013)	Adjective, noun	Learn adjective ordering	Weighted Addition

- Composition depends on representation and purpose

GROUNDED COMPOSITION

How is composition affected by the addition of visual information?

- Abstract vs. concrete phrases
- Images as visual phrases
- What information each modality provides

Shutova et al (2016)

Lazaridou et al (2014)

Collell & Moens (2016)

ABSTRACT COMPOSITION

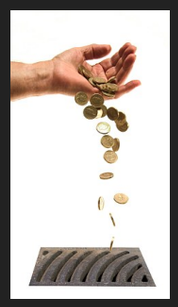
Shutova et al (2016)

- Metaphor comes from combining imagery of different domains

liquid ← **'pour money'** → ***finance***

- **Goal**: Use visual information to separate abstract vs. concrete
- **Data**: Wikipedia, ImageNet; MOH and TSV for testing

Text: $\vec{x}$ **'pour'** $\cos\angle$ $\vec{y}$ **'money'**

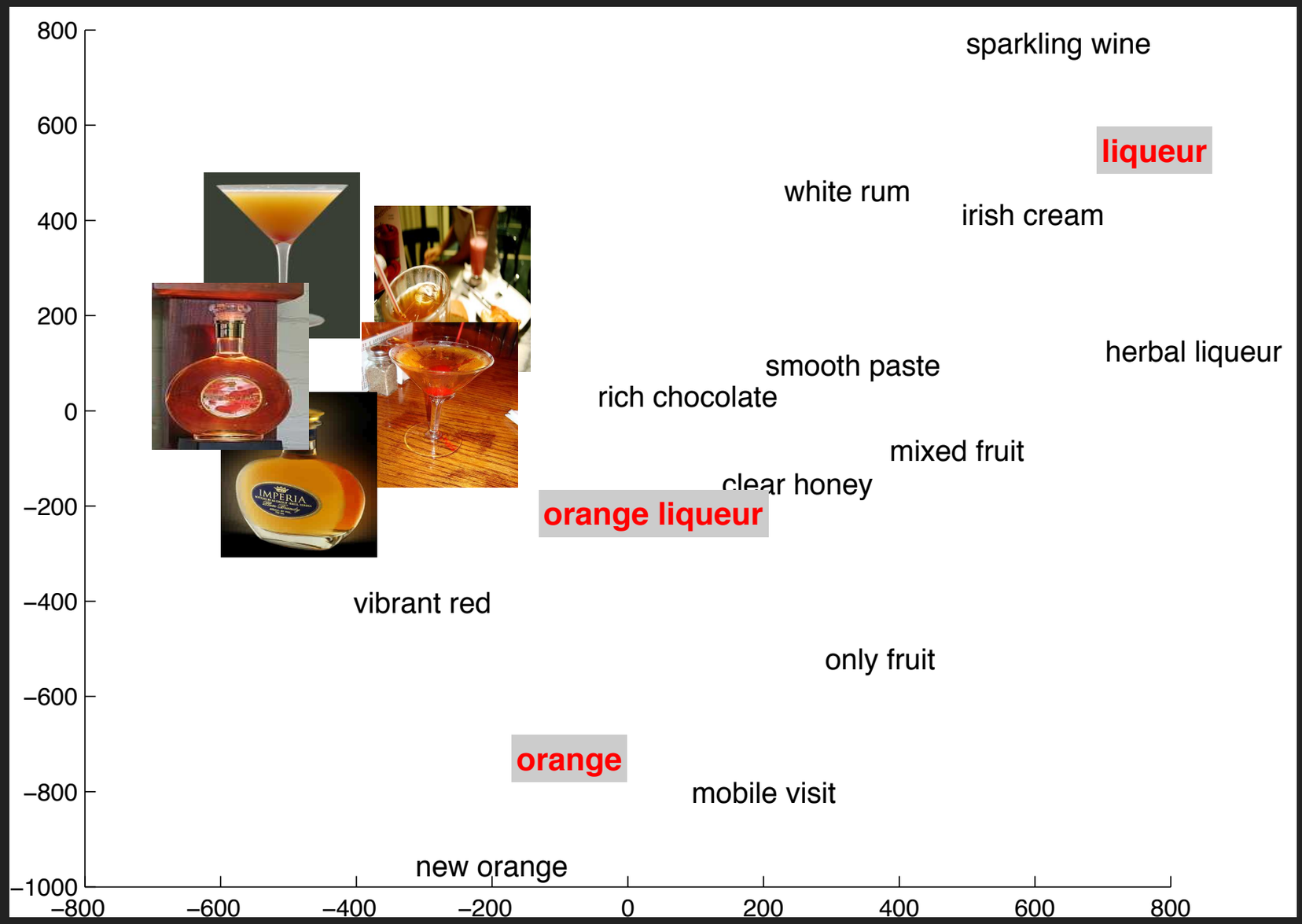
Image:  - $\vec{y}$ **'money'** = $\vec{u}$ **~pour~** $\cos\angle$ $\vec{x}$ **'pour'**

- **Limitations**:
  - Dependent on visual representation being incoherent for metaphors
  - Assume composition is addition

ATTRIBUTE COMPOSITION

Lazaridou et al (2015)

- Treat visual features as the phrase representations

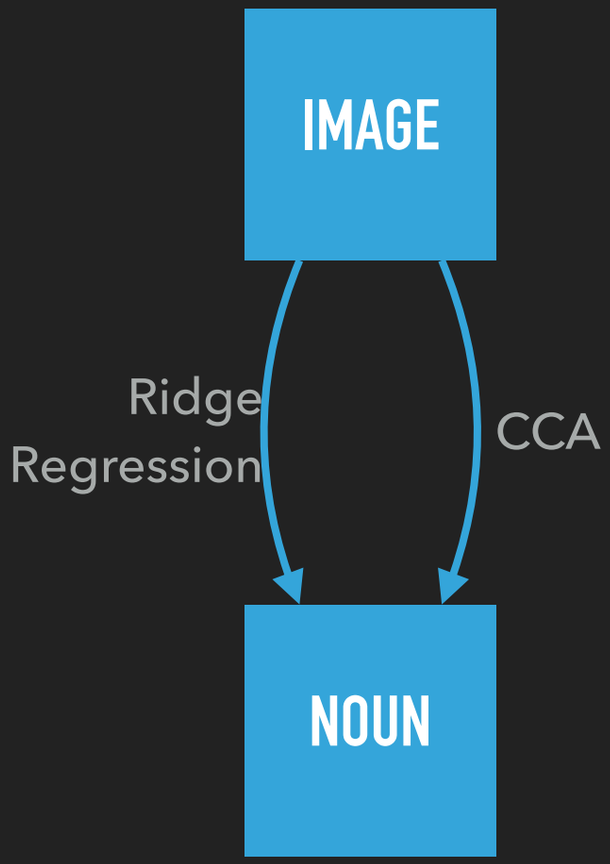


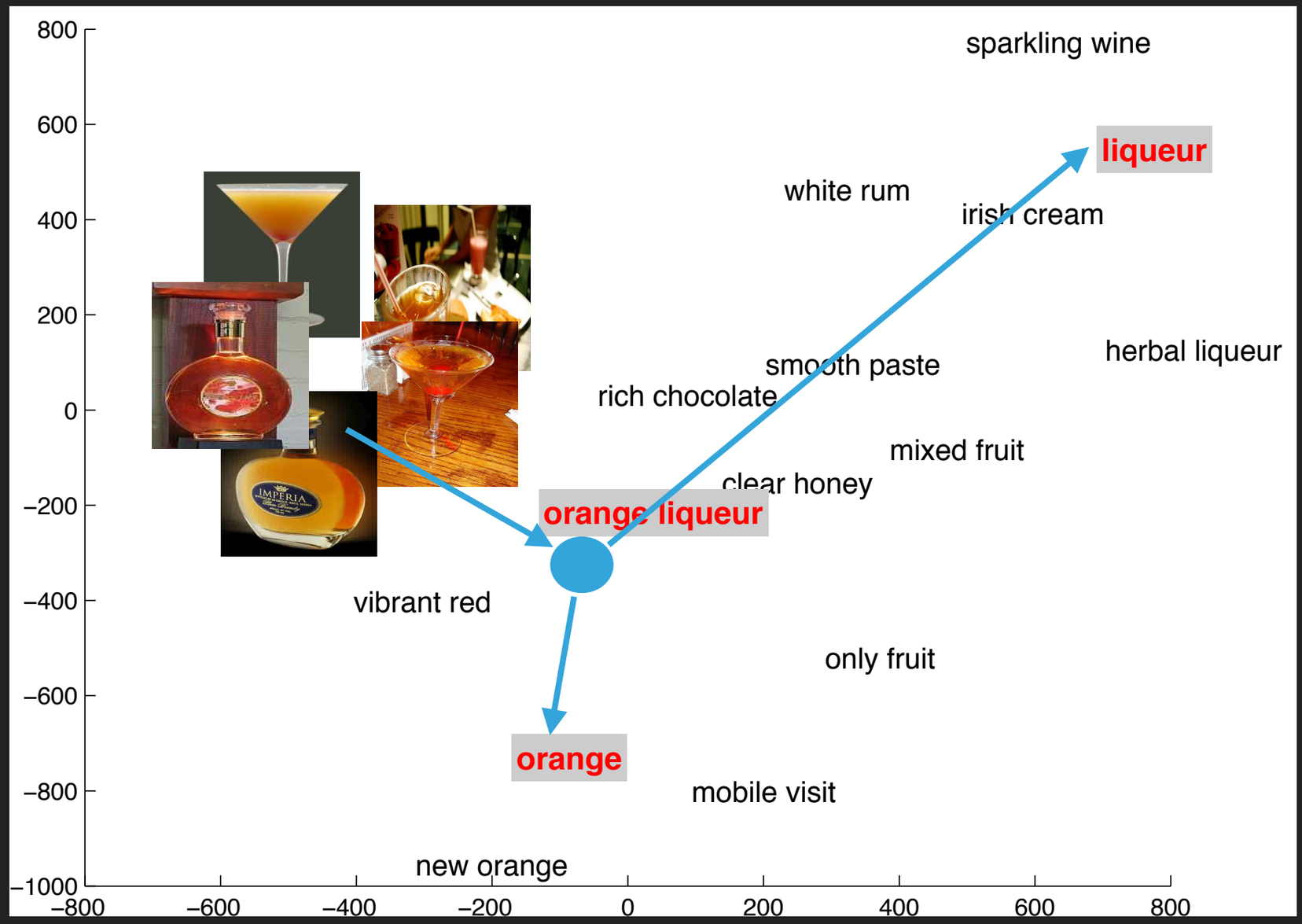
- **Data**: Wikipedia CBOW, 384 WordNet/ImageNet synsets

ATTRIBUTE COMPOSITION

Lazaridou et al (2015)

- **Goal**: Zero-shot attribute learning



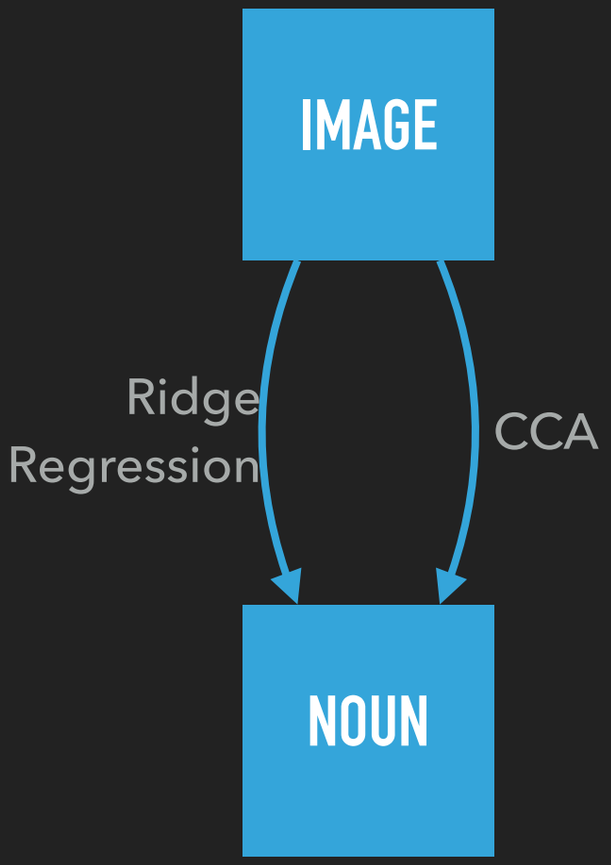


- **Data**: Wikipedia CBOW, 384 WordNet/ImageNet synsets

ATTRIBUTE COMPOSITION

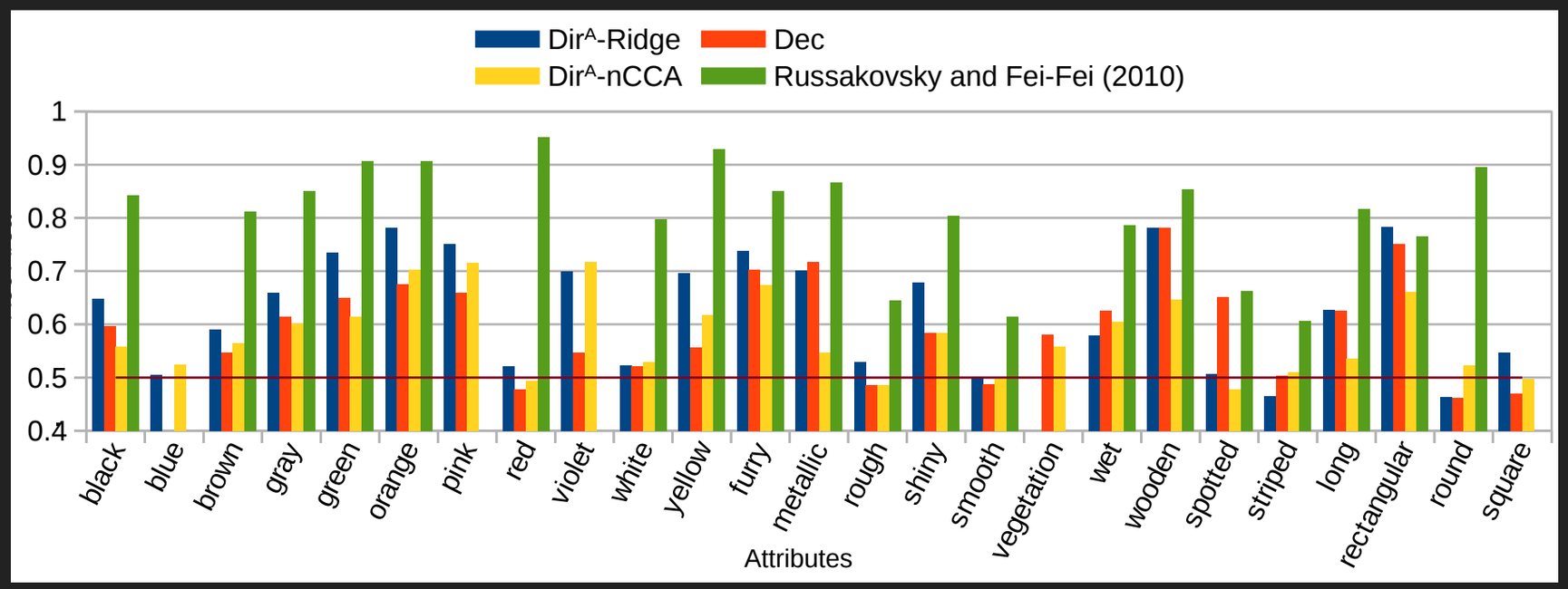
Lazaridou et al (2015)

- **Goal**: Zero-shot attribute learning



Dir^A: Attributes outside training

Dec: Leave-one-out training



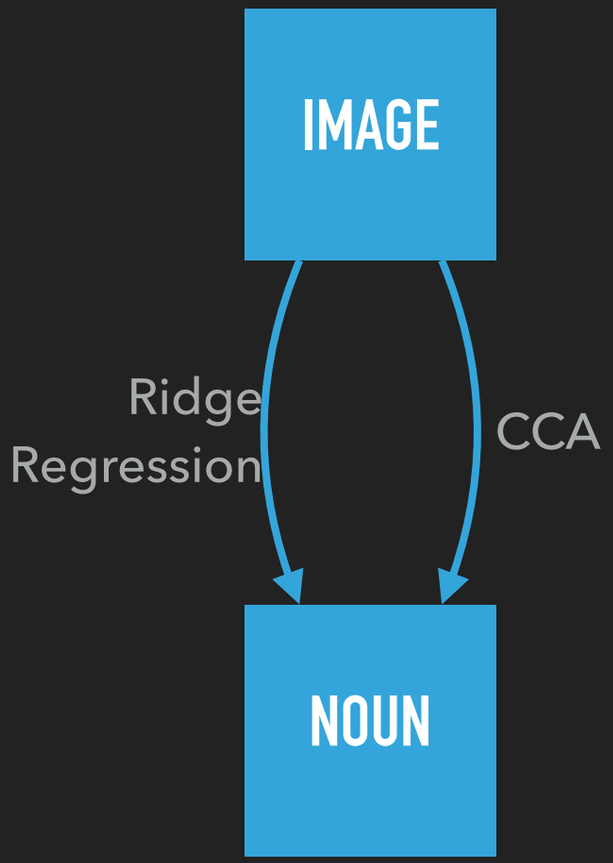
- **Limitation**: Decomposition only in linguistic space

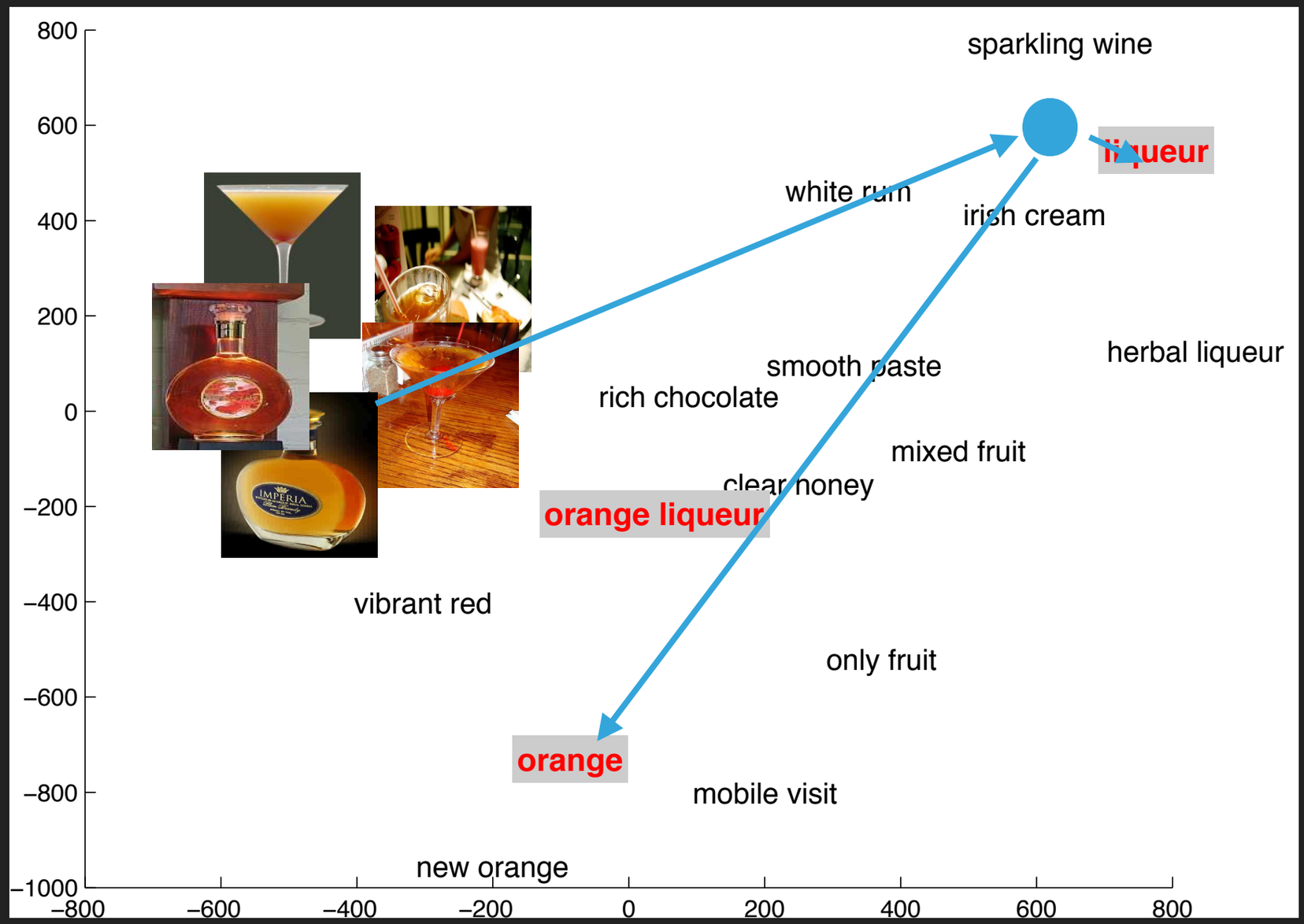
"Sunflowers are on average yellow (mean rank 2.3), fields are green (4.4), cabinets are wooden (4), and vans metallic (6.6) (strawberries are, suspiciously, blue, 2.7.)"

ATTRIBUTE COMPOSITION

Lazaridou et al (2015)

- **Goal**: Zero-shot attribute learning





ATTRIBUTE COMPOSITION

Lazaridou et al (2015)

- **Results**:

Image	Model	Top item	Top hit (Rank)
 A: white, brown N: dog	DEC	A: white N: dog	white (1) dog (1)
	DIR^{O}	A: animal N: goat	white (27) dog (25)
	LM	A: stray	brown (74)
	vLM	A: pet	brown (17)

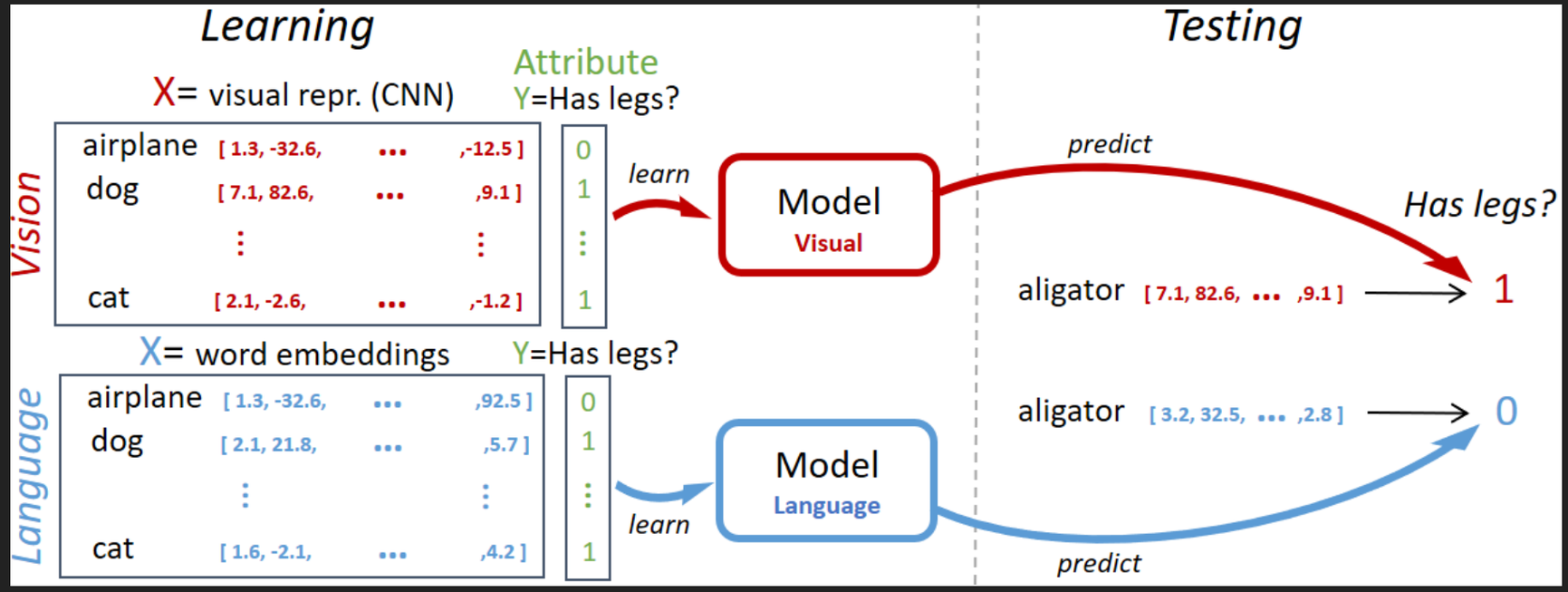
Image	Object	Predicted Attributes
	aeroplane	thick, wet, dry, cylindrical, motionless, translucent
	dog	cuddly, wild, cute, furry, white, coloured

- **Limitations**: Single object per image; have no correspondence between adjective closeness and relevance

VISUAL & LINGUISTIC INFORMATION

Collell & Moens (2016)

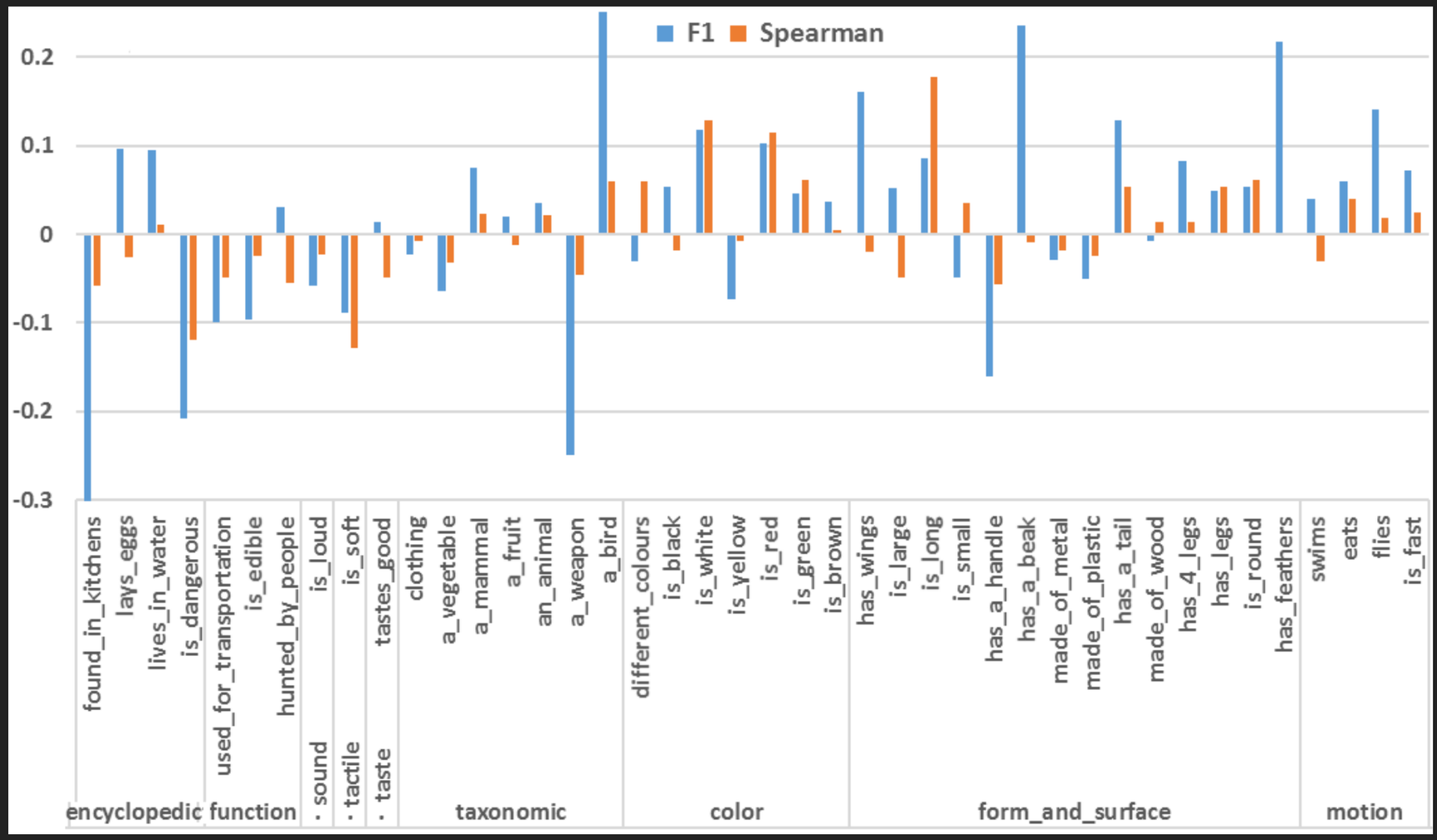
- **Goal**: Understand the information each modality captures
- **Data**: ImageNet, GloVe, McRae



VISUAL & LINGUISTIC REPRESENTATION

Collell & Moens (2016)

- **Results**: Positive is visual contribution, negative is text



- **Limitation**: Examined one textual and one visual representation

GROUNDED COMPOSITION

	Hypothesis	Result	Limitation
Shutova et al (2016)	Metaphor based on visual dichotomy	**Phrase composition reveals metaphoricity**	**Assumed addition-based composition**
Lazaridou et al (2015)	Image represents adjective-noun composition	**Zero-shot attribute learning**	**Decomposition only in linguistic space**
Collell & Moens (2016)	Text and images provide different information	**Images inform visual properties, text abstract ones**	**Limited representation analysis**

- Different representations have different strengths and weaknesses, as do the different modalities
- The more different the semantics of the linguistic space and visual space are, the more difficult it is to map between them

CONCLUSION

- What we can do:
  - Recognize objects and their component parts
  - Handle vagueness of adjective modification
  - Quantify attribute exhibition in a class
  - Compose adjectives and nouns

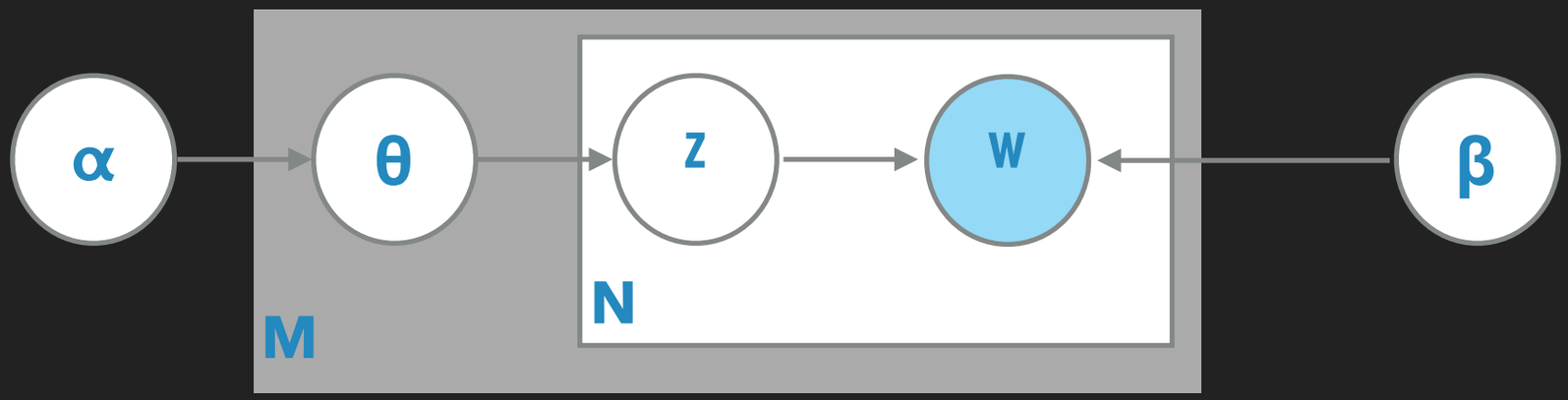
- Future considerations:
  - Applying quantifiers to properties of a single object (not just proportions of countable features)
  - Grounding comparative terms
  - Using attribute absence as a property of an object
  - Understanding human limits of differentiation

THANK YOU

QUESTIONS?

BACKUP SLIDES

FENG & LAPATA (2010)



$$\prod_{d=1}^{M} \int P(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} P(z_{dn}|\theta_d) P(w_{dn}|z_{dn}, \beta) \right) d\theta_d$$

$$(\gamma^*, \phi^*) = \underset{\gamma,\phi}{\operatorname{argmin}}\, d(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$

- KL divergence $D(p,q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j}$
- Jensen Shannon Divergence $JS(p,q) = \frac{1}{2}\left[D\left(p, \frac{(p+q)}{2}\right) + D\left(q, \frac{(p+q)}{2}\right)\right]$
- Conditional distribution $P(w_2|w_1) = \sum_{z=1}^{K} P(w_2|z)P(z|w_1)$

$$P(z|w_1) \propto P(w_1|z)P(z)$$

- Measure correlation with human similarity metrics using Pearson's *r*

FENG & LAPATA (2010)

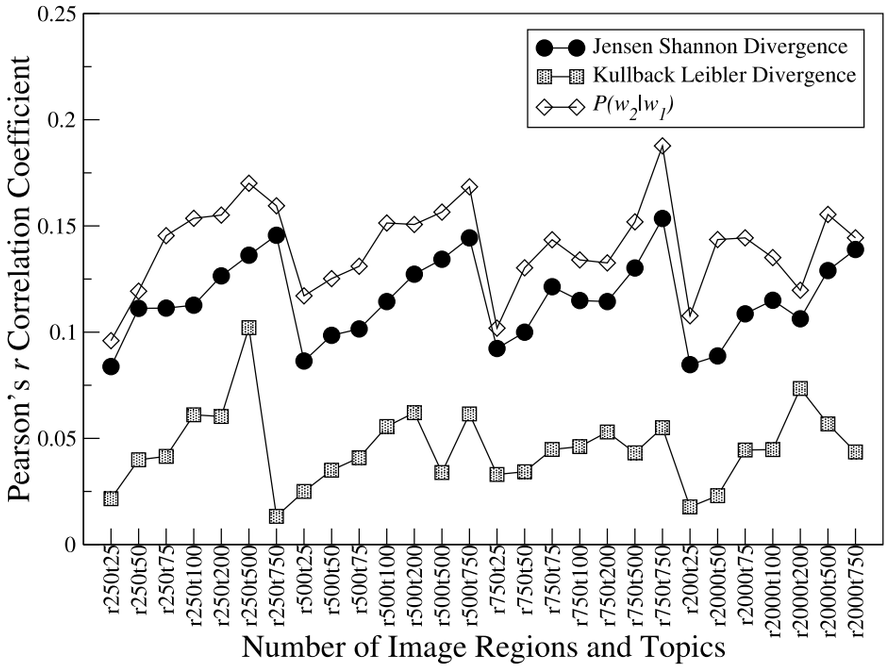


Figure 2: Performance of multimodal topic model on predicting word association under varying topics and visual terms (development set).

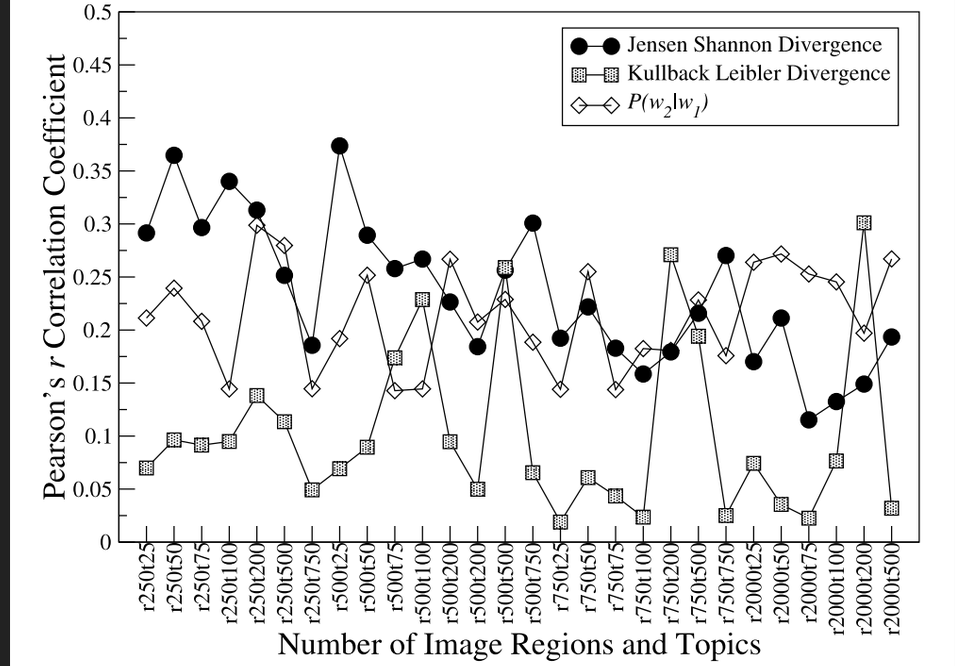


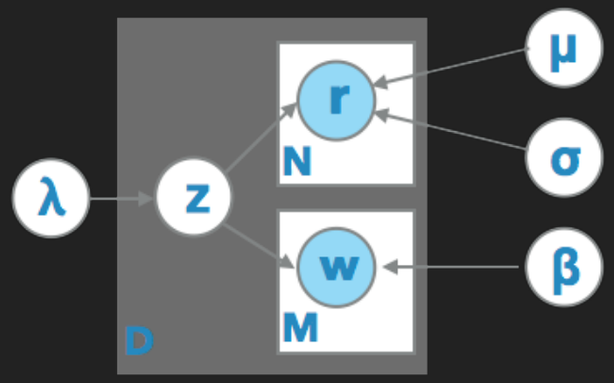
Figure 3: Performance of multimodal topic model on predicting word similarity under varying topics and visual terms (development set).

Model	Word Association	Word Similarity
UpperBnd	0.400	0.545
MixLDA	0.123	0.318
TxtLDA	0.077	0.247

Table 2: Model performance on word association and similarity (test set).

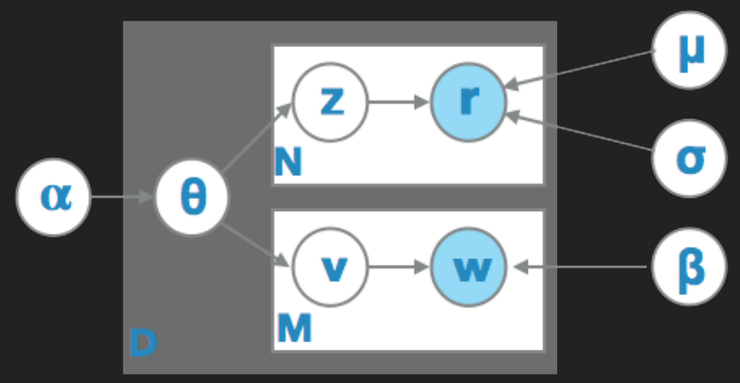
BLEI & JORDAN (2003)

GM-MIXTURE



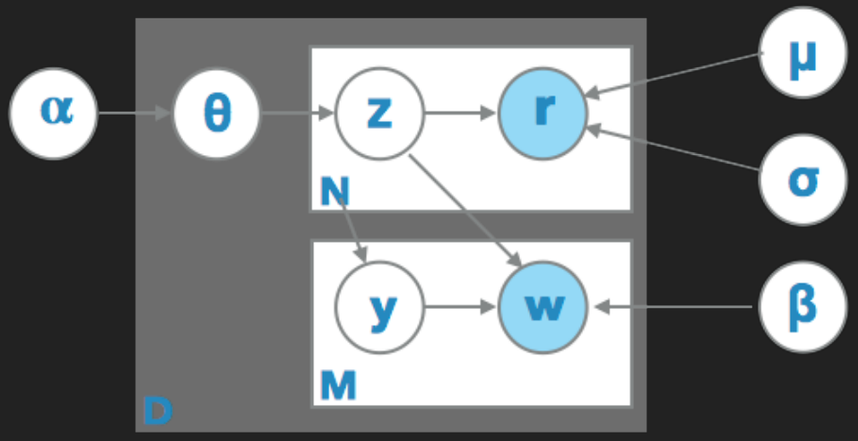
$$p(z, \mathbf{r}, \mathbf{w}) = p(z|\lambda)\Big(\prod_{n=1}^{N} p(r_n|z, \mu, \sigma)\Big)\Big(\prod_{m=1}^{M} p(w_m|z, \beta)\Big)$$

GM-LDA



$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{v}) = p(\theta|\alpha)\Big(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n, \mu, \sigma)\Big)\Big(\prod_{m=1}^{M} p(v_m|\theta)p(w_m|v_m, \beta)\Big)$$

Corr-LDA



$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{v}) = p(\theta|\alpha)\Big(\prod_{n=1}^{N} p(z_n|\theta)p(r_n|z_n, \mu, \sigma)\Big)\Big(\prod_{m=1}^{M} p(y_m|N)p(w_m|y_m, z, \beta)\Big)$$

BLEI & JORDAN (2003)

Variational Inference:

$$q(\theta, \mathbf{z}, \mathbf{y}) = q(\theta|\gamma)\Big(\prod_{n=1}^{N} q(z_n|\phi_n)\Big)\Big(\prod_{m=1}^{M} q(y_m|\lambda_m)\Big)$$

Update posterior Dirichlet

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}$$

Update posterior for each image region

$$\phi_{ni} \propto p(r_n|z_n = i, \mu, \sigma) \exp\{\mathrm{E}_q[\log \theta_i|\gamma]\}$$
$$\cdot \exp\left\{\sum_{m=1}^{M} \lambda_{mn} \log p(w_m|y_m = n, z_m = i, \beta)\right\}$$
$$\mathrm{E}_q[\log \theta_i|\gamma] = \Phi(\gamma_i) - \Phi(\textstyle\sum \gamma_j)$$

Update posterior for each word

$$\lambda_{mn} \propto \exp\left\{\sum_{i=1}^{K} \phi_{ni} \log p(w_m|y_m = n, z_n = i, \beta)\right\}$$

Approximate word dist:

$$p(w|\mathbf{r}) \approx \sum_{n=1}^{N} \sum_{z_n} q(z_n|\phi_n) p(w|z_n, \beta)$$

$$p(w|\mathbf{r}, r_n) \approx \sum_{z_n} q(z_n|\phi_n) p(w|z_n, \beta)$$

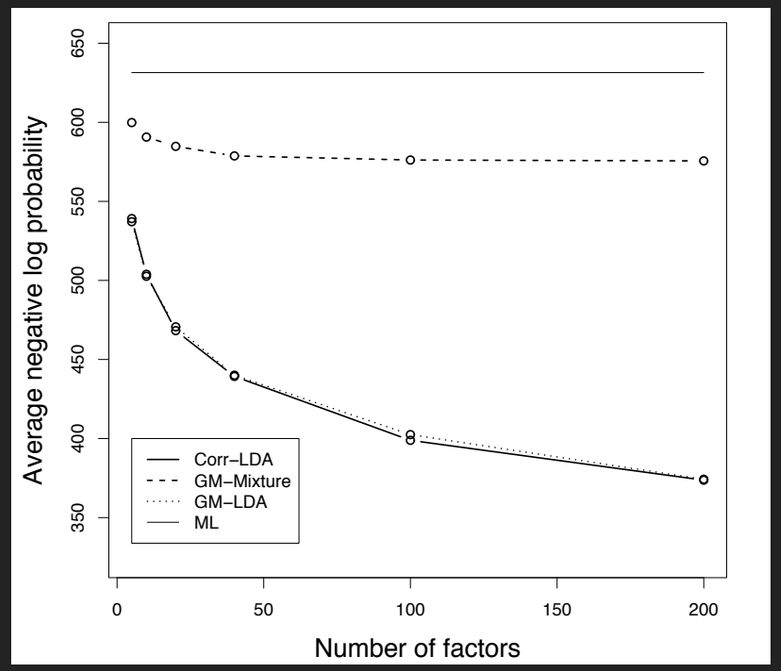
Smoothing: add prior dist. to β

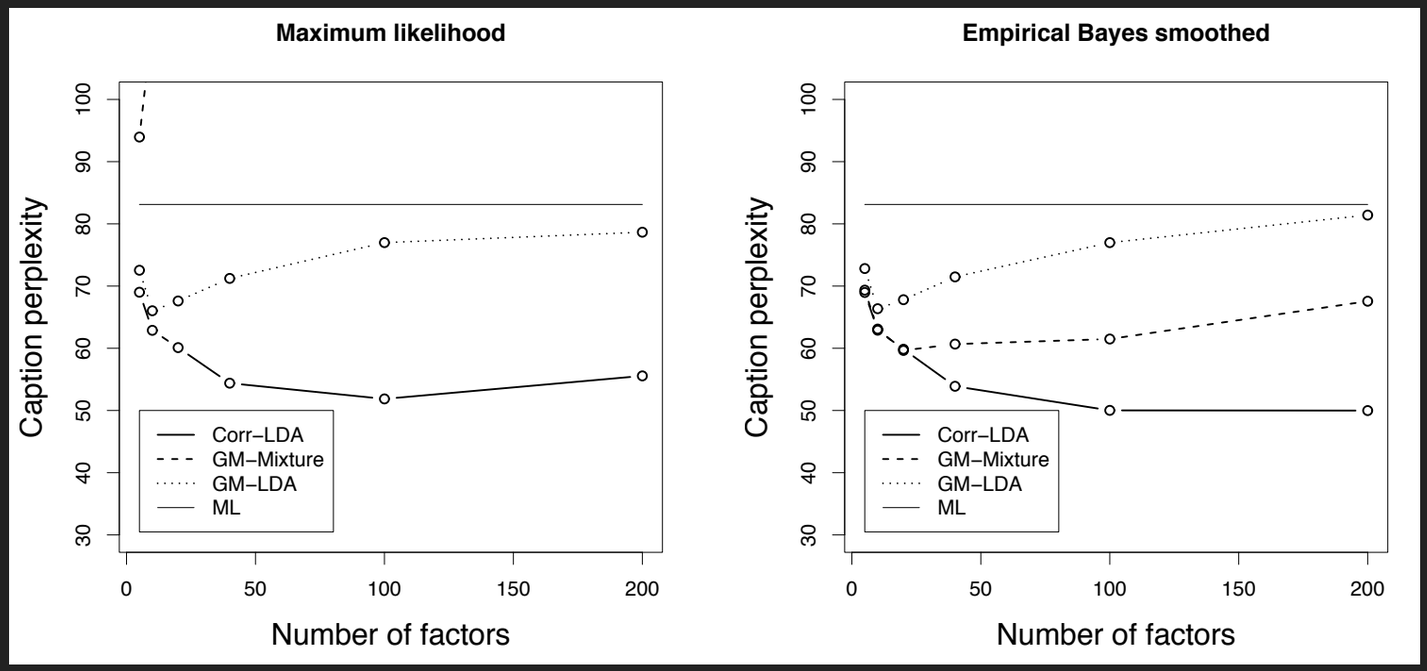
$$\beta_i \sim \mathrm{Dir}(\eta, \eta, ..., \eta)$$

$$\rho_{ij} = \eta + \textstyle\sum_{d=1}^{D} \sum_{m=1}^{M} 1(w_d m = j) \sum_{n=1}^{N} \phi_{ni} \lambda_{mn}$$

$$\beta \to \exp\{\mathrm{E}[\log \beta|\rho]\}$$

BLEI & JORDAN (2003)





$$\text{annotation perplexity} = \exp\{-\sum_{d=1}^{D}\sum_{m=1}^{M_d}\log p(w_m|r_d)/\sum_{d=1}^{D}M_d\}$$

inverse of geometric mean per-word likelihood

ROLLER & WALDE (2014)

- Text: DeWaC 1,038,883 documents consisting of 75,678 word types and 466M word tokens
- Association Norms: 95,214 cue-response pairs for 1,012 nouns and 5,716 response types
- Feature Norms: 11,714 cue-response pairs for 569 nouns and 2,589 response types
- Images: BidlerNetle 2022 word-synset mappings for just 309 words

Modality	K	Assoc.
Text Only		
Text Only (LDA)	200	.679
Bimodal mLDA		
Text + Feature Norms	150	.676
Text + SURF	50	.789 ***
Text + GIST	100	.739 ***
Text + SURF Clusters	200	.618 ***
Text + GIST Clusters	150	.690
3D mLDA		
Text + FN + SURF	100	.722 ***
Text + FN + GC	200	.601 ***
Hybrid mLDA		
FN, SURF	150+50	.800 ***
FN, GC	150+150	.742 ***
FN, GC, SURF	150+150+50	**.804** ***

Table 2: Average predicted rank similarity between cue words and their associates. Stars indicate statistical significance compared to the text-only modality, with gray stars indicating the model is statistically worse than the text model. The Hybrid models are the concatenation of the corresponding Bimodal mLDA models.

- Images: SURF & GIST ft
- Clusters: k-means 500 clusters of BoVW, images are membership in clusters

Modality	K	ρ
Text Only		
Text Only (LDA)	200	.204
Bimodal mLDA		
Text + Feature Norms	150	.310 ***
Text + Assoc. Norms	200	.328 **
Text + SURF	50	.251
Text + GIST	100	.204
Text + SURF Clusters	200	.159
Text + GIST Clusters	150	.233
3D mLDA		
Text + FN + AN	250	.259
Text + FN + SURF	100	.286 *
Text + FN + GC	200	.261 *
Hybrid mLDA		
FN, AN	150+200	.390 ***
FN, SURF	150+50	.350 ***
FN, GC	150+150	.340 ***
FN, AN, GC	150+200+150	.395 ***
FN, AN, SURF	150+200+50	.404 ***
FN, AN, SURF, GC	150+200+50+150	**.406** ***

Table 1: Average rank correlations between $-sKL(w_{compound}, w_{constituent})$ and our Compositionality gold standard. The Hybrid models are the concatenation of the corresponding Bimodal mLDA models. Stars indicate statistical significance compared to the text-only setting at the .05, .01 and .001 levels using a two-tailed t-test.

LAZARIDOU ET AL (2015)

- Skip-Gram:

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\underbrace{\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)}_{\mathcal{L}_{ling}(w_t)}\Big) \qquad p(w_{t+j}|w_t) = \frac{e^{{u'_{w_{t+j}}}^T u_{w_t}}}{\sum_{w'=1}^{W} e^{{u'_{w'}}^T u_{w_t}}}$$

- Visual knowledge: $\frac{1}{T}\sum_{t=1}^{T}(\mathcal{L}_{ling}(w_t) + \mathcal{L}_{vision}(w_t))$
- MM-Skipgram-A: $\mathcal{L}_{vision}(w_t) = -\sum_{w' \; P_n(w)} \max(0, \gamma - cos(u_{w_t}, v_{w_t}) + cos(u_{w_t}, v_{w'}))$
- MM-Skipgram-B: $u_{w_t}\text{replaced by } z_{w_t} = M^{u\to v}u_{w_t}$

LAZARIDOU ET AL (2015)

MEN: General relatedness
Simlex: Taxonomic sim.
SemSim: Semantic sim.
VisSim: Visual sim.

Pickle –> hamburger
PIckle –> cucumber
Pickle –> onion
Pickle –> zucchini

Model	*MEN*		*Simlex-999*		*SemSim*		*VisSim*	
	100%	42%	100%	29%	100%	85%	100%	85%
KIELA AND BOTTOU	-	0.74	-	0.33	-	0.60	-	0.50
BRUNI ET AL.	-	0.77	-	0.44	-	0.69	-	0.56
SILBERER AND LAPATA	-	-	-	-	0.70	-	0.64	-
CNN FEATURES	-	0.62	-	0.54	-	0.55	-	0.56
SKIP-GRAM	0.70	0.68	0.33	0.29	0.62	0.62	0.48	0.48
CONCATENATION	-	0.74	-	0.46	-	0.68	-	0.60
SVD	0.61	0.74	0.28	0.46	0.65	0.68	0.58	0.60
MMSKIP-GRAM-A	0.75	0.74	0.37	0.50	0.72	0.72	0.63	0.63
MMSKIP-GRAM-B	0.74	0.76	0.40	0.53	0.66	0.68	0.60	0.60

Human preference for nearest neighbor vs random image

	global	*|words|*	*unseen*	*|words|*
all	48%	198	30%	127
concrete	73%	99	53%	30
abstract	23%	99	23%	97

	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.5	2.6	14.2	23.5	36.1
MMSKIP-GRAM-A	2.1	3.7	16.7	24.6	37.6
MMSKIP-GRAM-B	2.2	5.1	20.2	28.5	43.5

Table 3: Percentage precision@k results in the zero-shot image labeling task.

	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.9	3.3	11.5	18.5	30.4
MMSKIP-GRAM-A	1.9	3.2	13.9	20.2	33.6
MMSKIP-GRAM-B	1.9	3.8	13.2	22.5	38.3

Table 4: Percentage precision@k results in the zero-shot image retrieval task.

WANG ET AL (2017)

Count-based comparison models

Independent Bernoulli

$$\mathbf{w}^{\alpha} = \mathbf{w}^{\alpha} + \underline{\mathbf{c}}$$

$$\mathbf{w}^{\beta} = \mathbf{w}^{\beta} + (1 - \underline{\mathbf{c}})$$

$$\mathbf{w}_{Ind} = \frac{\mathbf{w}^{\alpha}}{\mathbf{w}^{\alpha} + \mathbf{w}^{\beta}}$$

- α and β are parameters of Beta distribution
- Represents uncertainty about probability of property through Beta distribution over Bernoulli probabilities

Multinomial (properties compete)

- $\mathbf{c}_{\text{Mult}}$ is multinomial over properties
- *w* is Dirichlet with Q parameters

bi-TM

- $\mathbf{c}_{\text{Mult}}$ is Bernoulli mixture instead of independent properties (can represent co-occurrences)
- no competition between properties

WANG ET AL (2017)

Models		QMR BOW5	QMR Syn	Animal Syn
Baseline		0.12	0.16	0.63
PLS		**0.24**	0.35	0.71
Count	Mult.	0.13	0.25	0.64
	Ind.	0.11	0.23	0.64
	BernMix H1	0.11	0.17	0.65
	BernMix H2	0.10	0.18	0.63
bi-TM	plain	0.23	**0.36**	0.80
	BernMix H2	0.20	0.34	**0.81**

Table 1: MAP scores, multi-shot learning on the QMR and Animal datasets

	Models		all	oracle top20	AvgCos top20
QMR	Count	Mult.	0.16	0.37	0.28
		BernMix H1	0.14	0.33	0.21
		BernMix H2	0.15	0.31	0.22
	bi-TM	plain	**0.21**	**0.47**	**0.35**
		BernMix H2	0.18	0.45	0.34
Animal	Count	Mult.	0.58	0.77	0.61
		BernMix H1	0.60	0.80	0.57
		BernMix H2	0.59	0.81	0.59
	bi-TM	plain	0.64	0.88	0.63
		BernMix H2	**0.65**	**0.89**	**0.66**

Table 2: MAP scores, one-shot learning on the QMR and Animal datasets

MAP: Mean Average Precision
Measure what extent the model ranks definitional properties in the correct order

$$AP = \frac{1}{\sum_{i=1}^{n} I(i)} \sum_{i=1}^{n} \text{Prec}_i \cdot I(i)$$

Type	MAP
Function	0.45
Taxonomic	**0.62**
Visual	0.34
Encyclopaedic	0.35
Perc	0.40

Table 6: QMR, bi-TM, one-shot: MAP by property type over (oracle) top 20 context items

SILBERER ET AL (2014)

- Attribute classifiers
  - SVM trained on 4 features: color, texture, visual words, edges
- Image representation: normalized vector of attribute classification scores

$$\mathrm{p}_w = \frac{(sum_{i_w \in I_w} \mathrm{score}_a(i_w))_{a=1,\ldots,F}}{\sum_{a=1}^{F} \sum_{i_w \in I_w} \mathrm{score}_a(i_w)}$$

- Comparison Models
  - Concatenation
  - CCA

SILBERER ET AL (2014)

	Nelson	Concat	CCA	TopicAttr	TextAttr
Concat	0.24				
CCA	0.30	0.72			
TopicAttr	0.26	0.55	0.28		
TextAttr	0.21	0.80	0.83	0.34	
VisAttr	0.23	0.65	0.52	0.40	0.39

Table 5: Correlation matrix for seen Nelson et al. (1998) cue-associate pairs and five distributional models. All correlation coefficients are statistically significant ($p < 0.01, N = 435$).

	Nelson	Concat	CCA	TopicAttr	TextAttr
Concat	0.11				
CCA	0.15	0.66			
TopicAttr	0.17	0.69	0.48		
TextAttr	0.11	0.65	0.25	0.39	
VisAttr	0.13	0.57	0.87	0.57	0.34

Table 6: Correlation matrix for unseen Nelson et al. (1998) cue-associate pairs and five distributional models. All correlation coefficients are statistically significant ($p < 0.01, N = 1,716$).

Models	Seen
All Attributes	0.28
Text Attributes	0.20
Visual Attributes	0.25

Table 7: Model performance on seen Nelson et al. (1998) cue-associate pairs; models are based on gold human generated attributes (McRae et al., 2005). All correlation coefficients are statistically significant ($p < 0.01, N = 435$).

Models	Seen	Unseen
Concat	0.22	0.10
CCA	0.26	0.15
TopicAttr	0.23	0.19
TextAttr	0.20	0.08
VisAttr	0.21	0.13
MixLDA	0.16	0.11

Table 8: Model performance on a subset of Nelson et al. (1998) cue-associate pairs. Seen are concepts known to the attribute classifiers and covered by MixLDA ($N = 85$). Unseen are concepts covered by LDA but unknown to the attribute classifiers ($N = 388$). All correlation coefficients are statistically significant ($p < 0.05$).

LAZARIDOU ET AL (2014)

- 4 zero-shot approaches
  - Linear Projection: $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = (\mathbf{V}_s^T\mathbf{V}s)^{-1}\mathbf{V}_s^T\mathbf{W}_s$
  - CCA: $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = \mathbf{C}_V{\mathbf{C}_W}^{-1}$
  - SVD: $\mathrm{f}_{\mathrm{proj}_{\mathrm{v}\rightarrow\mathrm{w}}} = \mathbf{Z}_k\mathbf{Z}_k^T$

    $[\hat{\mathbf{V}}_s\hat{\mathbf{W}}_s] = \mathbf{U}_k\Sigma_k\mathbf{Z}_k^T$
  - Neural Network (used hyperbolic tangent)

Model \ k	1	2	3	5	10	20
Chance	1.1	2.2	3.3	5.5	11.0	22.0
SVD	1.9	5.0	8.1	14.5	29.0	48.6
CCA	3.0	6.9	10.7	17.9	31.7	51.7
lin	2.4	6.4	10.5	18.7	33.0	55.0
NN	3.9	6.6	10.6	21.9	37.9	58.2

Table 2: Percentage accuracy among top k nearest neighbors on CIFAR-100.

Context \ Mapping	v → w	w → v
Chance	17	17
context 1	12.6	14.5
context 5	8.08	13.29
context 10	7.29	13.44
context 20	6.02	12.17
context full	5.52	5.88

Table 7: Mean rank results averaged across 34 concepts when mapping an image-based vector and retrieving its linguistic neighbors (v → w) as well as when mapping a text-based vector and retrieving its visual neighbors (w → v). Lower numbers cue better performance.

HWANG & SIGAL (2014)

$$\mathcal{L}_C(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{x_i}, y_i) = \sum_c [1 + ||\boldsymbol{W}\boldsymbol{x_i} - \boldsymbol{u}_{y_i}||_2^2 - ||\boldsymbol{W}\boldsymbol{x_i} - \boldsymbol{u}_c||_2^2]_+, \ \forall c \neq y_i \quad (1)$$

$$\min_{\boldsymbol{W},\boldsymbol{U}} \sum_i^N \mathcal{L}_C(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{x_i}, y_i) + \lambda||\boldsymbol{W}||_F^2 + \lambda||\boldsymbol{U}||_F^2, y_i \in \{1, ..., m\} \quad (2)$$

$$\mathcal{L}_S(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{x_i}, y_i) = \sum_{s \in \mathcal{P}_{\dagger\rangle}} \sum_{c \in \mathcal{S}_s} [1 + ||\boldsymbol{W}\boldsymbol{x_i} - \boldsymbol{u}_s||_2^2 - ||\boldsymbol{W}\boldsymbol{x_i} - \boldsymbol{u}_c||_2^2]_+ \quad (3)$$

$$\mathcal{L}_A(\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{x_i}, y_i) = 1 - \sum_a (\boldsymbol{W}\boldsymbol{x_i})^T y_i^a \boldsymbol{u}_a, ||\boldsymbol{u}_a||^2 \leq 1, y_i^a \in \{0, 1\}, \forall a \in \mathcal{A}_{y_i} \quad (4)$$

$$\boldsymbol{u}_c = \boldsymbol{u}_p + \boldsymbol{U}^A \beta c, c \in \mathcal{C}_p, ||\beta_c||_0 \preceq \gamma_1, \beta_c \succeq 0, \forall c \in \{1, ..., C\} \quad (5)$$

$$\mathcal{R}(\boldsymbol{U}, \boldsymbol{B}) = \sum_c^C ||\boldsymbol{u}_c - \boldsymbol{u}_p - \boldsymbol{U}^A \beta_c||_2^2 + \gamma_2 ||\beta_c + \beta_o||_2^2$$

$$c \in \mathcal{C}_p, o \in \mathcal{P}_c \cup \mathcal{S}_c, 0 \preceq \beta_c \preceq \gamma_1, \forall c \in \{1, ..., C\} \quad (6)$$

CHEN ET AL

Selective Sharing

Multi-task lasso: all-competing **λ** balances sparcity against classification loss

$$\mathbf{W}^* \underset{\mathbf{W}}{\operatorname{argmin}} L(\mathbf{X}, \mathbf{Y}; \mathbf{W}) + \lambda \sum_m ||\mathbf{w}^m||_1$$

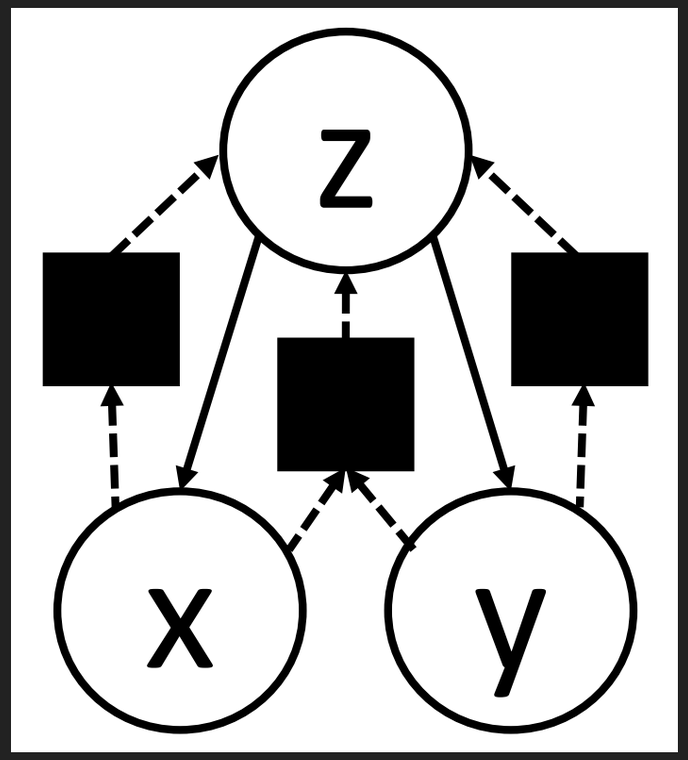
$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \sum_{m,n} \log(1 + \exp((1 - 2y_n^m)\boldsymbol{x}_n^T \boldsymbol{w}^m)) + \lambda \sum_{d=1}^{D} \underbrace{\sum_{l=1}^{L} ||\boldsymbol{w}_d^{S_l}||_2}_{||V||_1}$$

Category-Specific Attributes

$$\text{minimize } \left(\frac{1}{2}||\mathbf{w}||^2 + C_S \sum_i \xi_i + C_O \sum_j \gamma_j\right)$$

$$\begin{aligned} s.t. \quad & y_i \boldsymbol{w}^T \boldsymbol{x}_i \geq 1 - \xi_i; \forall i \in \mathscr{I} \\ & y_j \boldsymbol{w}^T \boldsymbol{x}_j \geq 1 - \gamma_j; \forall j \in \mathscr{O} \\ & \xi_i \geq 0; \gamma_j \geq 0 \end{aligned}$$

VEDANTAM ET AL (2017)



$$p(x, y, z) = p(z)p(x|z)p(y|z)$$

$$q(z|x, y), q(z|x), q(z|y)$$

$$p(y|x) = \int p(y|z)q(z|x)dz$$
$$p(x|y) = \int p(x|z)q(z|y)dz$$

- **Attribute Independence:**

$$p(y|z) = \prod_{k \in \mathcal{A}} p(y_k|z)$$

- **Product of Experts:**

$$q(z|y_{\mathcal{O}}) \propto p(z) \prod_{k \in O} q(z|y_k)$$

$$\text{correctness}(\mathcal{S}, y_{\mathcal{O}}) = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \frac{1}{|\mathcal{O}|} \sum_{k \in \mathcal{O}} \mathbb{I}(\hat{y}(x)_k = y_k)$$

Frac. of attr. that match desc.

$$\text{coverage}(\mathcal{S}, y_{\mathcal{O}}) = \frac{1}{|\mathcal{M}|} \sum_{k \in \mathcal{M}} (1 - JS(p_k, q_k))$$

Meas. diversity of underspec attr.

VEDANTAM ET AL (2017)

Name	Ref	Model	Objective
VAE	(Kingma et al., 2014)	$p(z)p(x\|z)$	$\text{elbo}(x\|z;\ z\|x)$
triple ELBO	This	$p(z)p(x\|z)p(y\|z)$	$\text{elbo}(x,y\|z;\ z\|x,y)$ $+\text{elbo}(x\|z;\ z\|x) + \text{elbo}(y\|z;\ z\|y)$
JMVAE	(Suzuki et al., 2017)	$p(z)p(x\|z)p(y\|z)$	$\text{elbo}(x,y\|z;\ z\|x,y)$ $-\alpha\text{KL}(q(z\|x,y),q(z\|x))$ $-\alpha\text{KL}(q(z\|x,y),q(z\|y))$
bi-VCCA	(Wang et al., 2016)	$p(z)p(x\|z)p(y\|z)$	$\mu\ \text{elbo}(x,y\|z;\ z\|x)$ $+(1-\mu)\text{elbo}(x,y\|z;\ z\|y)$
JVAE-Pu	(Pu et al., 2016)	$p(z)p(x\|z)p(y\|z)$	$\text{elbo}(x,y\|z;\ z\|x) + \text{elbo}(x\|z;\ z\|x)$
JVAE-Kingma	(Kingma et al., 2014)	$p(z)p(y)p(x\|z,y)$	$\text{elbo}(x\|y,z;\ z\|x,y) + \log p(y)$
CVAE-Yan	(Yan et al., 2016)	$p(z)p(x\|y,z)$	$\text{elbo}(x\|y,z;\ z\|x,y)$
CVAE-Sohn	(Sohn et al., 2015)	$p(z\|x)p(y\|x,z)$	$\text{elbo}(y\|x,z;\ z\|x,y;\ z\|x)$
CMMA	(Pandey et al., 2017)	$p(z\|y)p(x\|z)$	See text.

Table 2: Comparison of different approaches on MNIST-a test set. Higher numbers are better. Error bars (in parentheses) are standard error of the mean. For concrete concepts (where all 4 attributes are specified), we do not use a PoE inference network, and we do not report coverage. Hyperparameter settings for each result are discussed in the supplementary material.

Method	#Attributes	Coverage (%)	Correctness (%)	PoE?	Training set
triple ELBO	4	-	**90.76 (0.11)**	N	`iid`
JMVAE	4	-	86.38 (0.14)	N	`iid`
bi-VCCA	4	-	80.57 (0.26)	N	`iid`
triple ELBO	3	**90.76 (0.21)**	77.79 (0.30)	Y	`iid`
JMVAE	3	89.99 (0.20)	**79.30 (0.26)**	Y	`iid`
bi-VCCA	3	85.60 (0.34)	75.52 (0.43)	Y	`iid`
triple ELBO	2	**90.58 (0.17)**	**80.10 (0.47)**	Y	`iid`
JMVAE	2	89.55 (0.30)	77.32 (0.44)	Y	`iid`
bi-VCCA	2	85.75 (0.32)	75.98 (0.78)	Y	`iid`
triple ELBO	1	**91.55 (0.05)**	**81.90 (0.48)**	Y	`iid`
JMVAE	1	89.50 (0.09)	81.06 (0.23)	Y	`iid`
bi-VCCA	1	87.77 (0.10)	76.33 (0.67)	Y	`iid`
triple ELBO	4	-	**83.10 (0.07)**	N	`comp`
JMVAE	4	-	79.34 (0.52)	N	`comp`
bi-VCCA	4	-	75.18 (0.51)	N	`comp`

VEDANTAM ET AL (2017)

Label space partitioning: for each $y^{(2)} \in Y_2$ there is no $y^{(1)} \in Y_1$ that is identical but there is some $y^{(1)}$ which shares at least one attribute with $y^{(2)}$

Example:

$Y = \{(b,r),(b,w),(s,r),(s,w)\}$ Cartesian product of (big, small) with (red, white)

$Y_1 = \{(b,r),(s,w)\}$

$Y_2 = \{(b,w),(s,r)\}$

$R_1(x,y) = p_{true}(x,y|y \in Y_1)$

$R_2(x,y) = p_{true}(x,y|y \in Y_2)$

} distribution over images and concepts for each split

$D_{xy1}^{train} = \{(x,y) \sim p_1\}$

$D_{xy2}^{test} = \{(x,y) \sim p_2\}$

Evaluation

$S(y_0) = \{x^{(n)} \sim p(x|y_0) : n = 1:N\}$: x (images) generated from y_0 (observed descriptions) N times

$\hat{y}(x)$ predicted attribute vector

VEDANTAM ET AL (2017)

VAE-based models:

$p_\theta(x,z) = p_\theta(z)p_\theta(x|z)$

- $p_\theta(x|z)$ → likelihood ("decoder") – usually done with a neural network
- $p_\theta(z)$ → prior – assume Gaussian ($p(z) = \mathcal{N}(z|0,I)$)

Posterior inference: fit $q_\phi(z|x)$ ("encoder") to maximize Evidence Lower BOund

$elbo(x) = E_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x), p(z))$

↓ to TRIPLE

$p(x,y,z) = p(z)p(x|z)\prod_{k\in A} p(y_k|z)$

- $\prod_{k\in A} p(y_k|z)$ → MLP for k^{th} attribute – maps z to softmax dist. over possible values for y_k
- $p(x|z)$ → image decoder (used DCGAN architecture)

Single modality:

$\sum_y p(x,y|z) = \sum_y [p(x|z)p(y|z)] = p(x|z)\left[\sum_y p(y|z)\right] = p(x|z)$ → inference: $q(z|x)$

assumes independence

↳ thus $\sum_x p(x,y|z) = p(y|z)$ → inference: $q(z|y)$

VEDANTAM ET AL (2017)

TRIPLE ELBO:

$$\mathcal{L}(x,y) = E_{q(z|x,y)}[\lambda_x^{xy} \log p(x|z)] + \lambda_y^{yx} \log p(y|z)] - KL(q(z|x,y), p(z))$$
$$+ E_{q(z|x)}[\lambda_x^{x} \log p(x|z)] - KL(q(z|x), p(z))$$
$$+ E_{q(z|y)}[\lambda_y^{y} \log p(y|z)] - KL(q(z|y), p(z))$$

$\lambda_x^x, \lambda_y^y, \lambda_x^{xy}, \lambda_y^{yx}$: scaling terms to balance likelihood contributions of each data type

Justification:

$$q(z|D_{xy}) = \frac{1}{|D_{xy}|} \sum_{x \in D_{xy}} q(z|x,y)$$ → want latent space of attributes to match latent space of corresponding image-attribute pairs

★ Want $q(z|y)$ to match $q(z|D_{xy})$ to 'cover' all examples, but need some room to handle novel examples (i.e. be 'broad', 'entropic', as close to $\mathcal{N}(0,I)$ prior as possible

★ Parameters of $p(y|z)$ only trained in $elbo(x,y)$ [not $elbo(y)$] } ELBO + FL
★ Optimizing the $elbo(y)$ term maps into precomputed embedding space } "frozen likelihood"

→ Enforces the model to use the same embedding space whether or not the data is paired

VEDANTAM ET AL (2017)

Scaling terms: Decoders $(p(x|z), p(y|z))$ are Gaussians

$$\lambda_x^x \log p(x|z) = -\frac{\lambda_x^x}{2\sigma^2} \cdot (x-\mu(z))^T \Sigma(z)^{-1} (x-\mu(z))^2 + \text{const.}$$

→ outputs of $p(x|z)$ given input z

(unimodal)

$$L = \lambda_x^x E_{q(z|x)}[\log p(x|z)] - KL(q(z|x), p(z))$$

$$= E_{q(z|x)}[\log p(x|z)] - \beta_x KL(q(z|x), p(z)) \qquad \beta_x = 1/\lambda_x^x$$

(multimodal)

$\lambda_y^y > 1$, $\lambda_x^x \leq 1$, $\lambda_y^{yx}/\lambda_x^{xy} > 1$ → images more informative than attributes
→ ensures latent space shared between images & attributes

DE MELO & BANSAL (2013)

- Word scoring

$$W_1 = \frac{1}{P_1} \sum_{p_1 \in P_{ws}} cnt(p_1(a_1, a_2)) \qquad S_1 = \frac{1}{P_2} \sum_{p_2 \in P_{sw}} cnt(p_2(a_1, a_2))$$

$$W_2 = \frac{1}{P_1} \sum_{p_1 \in P_{ws}} cnt(p_1(a_2, a_1)) \qquad S_2 = \frac{1}{P_2} \sum_{p_2 \in P_{sw}} cnt(p_2(a_2, a_1))$$

$$P_1 = \sum_{p_1 \in P_{ws}} cnt(p_1) \qquad P_2 = \sum_{p_2 \in P_{sw}} cnt(p_2)$$

$$score(a_1, a_2) = \frac{(W_1 - S_1) - (W_2 - S_2)}{cnt(a_1) \cdot cnt(a_2)}$$

DE MELO

- MILP

maximize

$$\sum_{(i,j)\notin E}(w_{ij}-s_{ij})\dot{s}core(a_i,a_j)-\sum_{(i,j)\in E}(w_{ij}+s_{ij})C$$

subject to

$$d_{ij}=x_j-x_i \qquad \forall i,j\in\{1,...,N\}$$
$$d_{ij}-w_{ij}C\leq 0 \qquad \forall i,j\in\{1,...,N\}$$
$$d_{ij}+(1-w_{ij})C>0 \qquad \forall i,j\in\{1,...,N\}$$
$$d_{ij}+s_{ij}C\geq 0 \qquad \forall i,j\in\{1,...,N\}$$
$$d_{ij}-(1-s_{ij})C<0 \qquad \forall i,j\in\{1,...,N\}$$
$$x_i\in[0,1] \qquad \forall i\in\{1,...,N\}$$
$$w_{ij}\in\{0,1\} \qquad \forall i,j\in\{1,...,N\}$$
$$s_{ij}\in\{0,1\} \qquad \forall i,j\in\{1,...,N\}$$

DE MELO & BANSAL (2013)

Method	Pairwise Accuracy	Avg. τ	Avg. $\lvert\tau\rvert$	Avg. ρ	Avg. $\lvert\rho\rvert$
Web Baseline	48.2%	N/A	N/A	N/A	N/A
Divide-and-Conquer	50.6%	0.45	0.53	0.52	0.62
Sheinman and Tokunaga (2009)	55.5%	N/A	N/A	N/A	N/A
MILP	69.6%	0.57	0.65	0.64	0.73
MILP with synonymy	78.2%	0.57	0.66	0.67	0.80
Inter-Annotator Agreement	78.0%	0.67	0.76	0.75	0.86

Table 3: Main test results

		Predicted Class		
		Weaker	Tie	Stronger
True Class	Weaker	117	127	15
	Tie	5	42	15
	Stronger	11	122	115

Table 4: Confusion matrix (Web baseline)

		Predicted Class		
		Weaker	Tie	Stronger
True Class	Weaker	177	29	53
	Tie	9	24	29
	Stronger	15	38	195

Table 5: Confusion matrix (MILP)

QING & FRANKE (2014)

Speaker: $\sigma(u_1|b_0, \Pr) = p(\theta \leq b_0) = \int_{-\infty}^{b_0} \Pr(\theta)\mathrm{d}\theta$

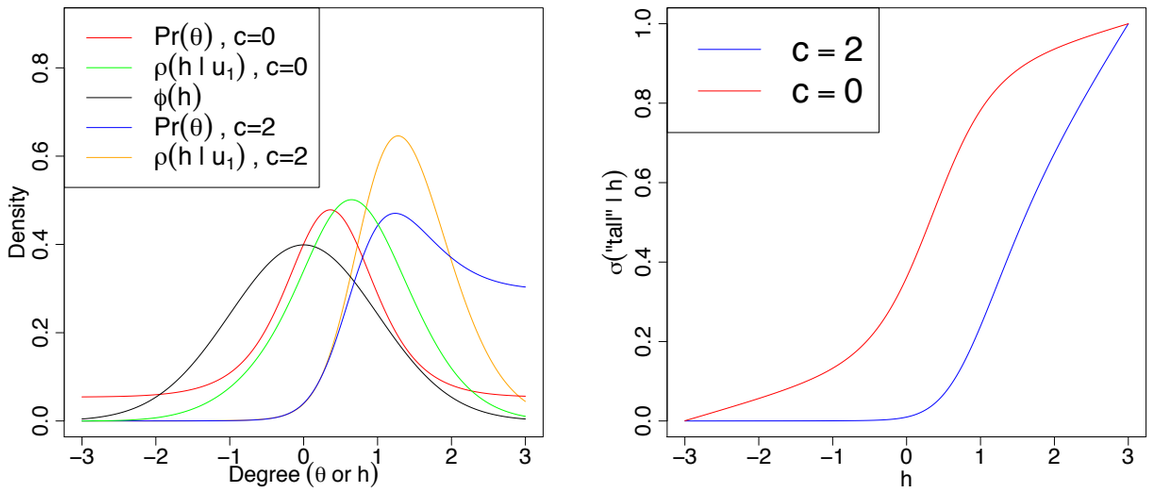
Listener believes correct height: $\frac{\phi(b_0)}{1-\int_{-\infty}^{\theta} \phi(b)\mathrm{d}b}$

Expected success: $ES(\theta) = \int_{-\infty}^{\theta} \phi(b)\phi(b|u_0, \theta)\mathrm{d}b + \int_{\theta}^{\infty} \phi(b)\phi(b|u_1, \theta)\mathrm{d}b$

Utility: $U(\theta) = ES(\theta) - \int_{\theta}^{\infty} \phi(b) \cdot c\mathrm{d}b$

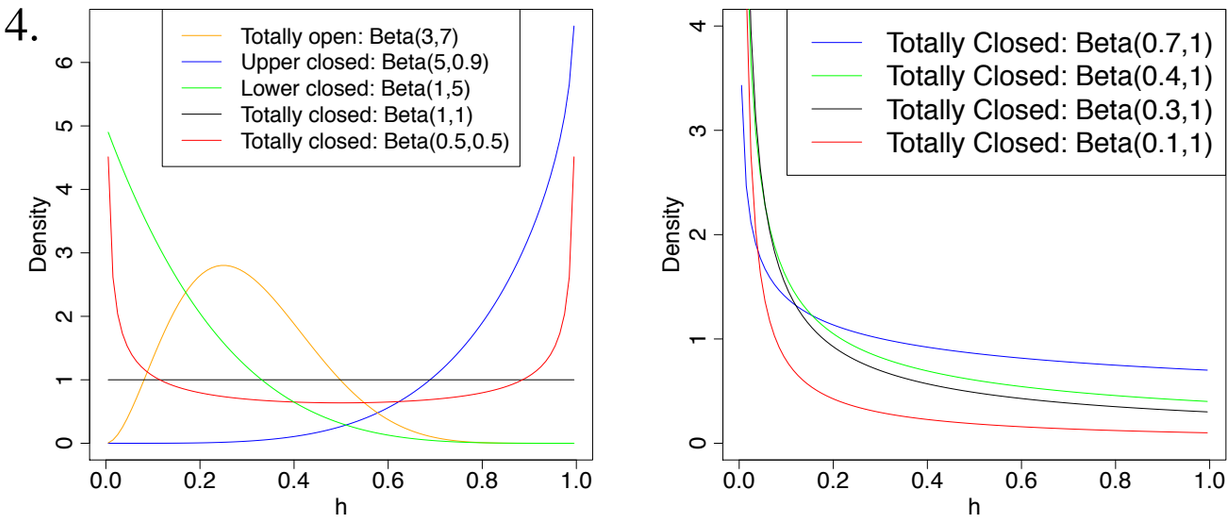
Threshold distribution: $\Pr(\theta) \propto (\exp(\lambda \cdot U(\theta))$

QING & FRANKE (2014)



(a) $\Pr(\theta)$ and $\rho(h \mid u_1)$ (b) $\sigma(u_1 \mid h)$

SOM predictions for Gaussian prior N(0,1), with $\lambda = 4$.



(a) Beta(α, β) (b) Beta$(\alpha, 1)$

Correspondence between beta distributions and scale structures.

LASSITER & GOODMAN (2015)

- Pragmatic Listener

$$P_{L_1}(A, V|u) \propto P_{S_1}(u|A, V) \cdot P_{L_1}(A) \cdot P_{L_1}(V)$$

$$P_{S_1}(u|A, V) \propto \exp(\lambda \cdot \ln[P_{L_0}(A|u, V) - C(u)])$$

{∅, ‘small’, ‘large’}

$$P_{L_0}(A|u, V) = P_{L_0}(A[[u]]^V = 1)$$

LASSITER & GOODMAN (2015)

- "I ate **some** cookies"

PRAGMATIC LISTENER L_1

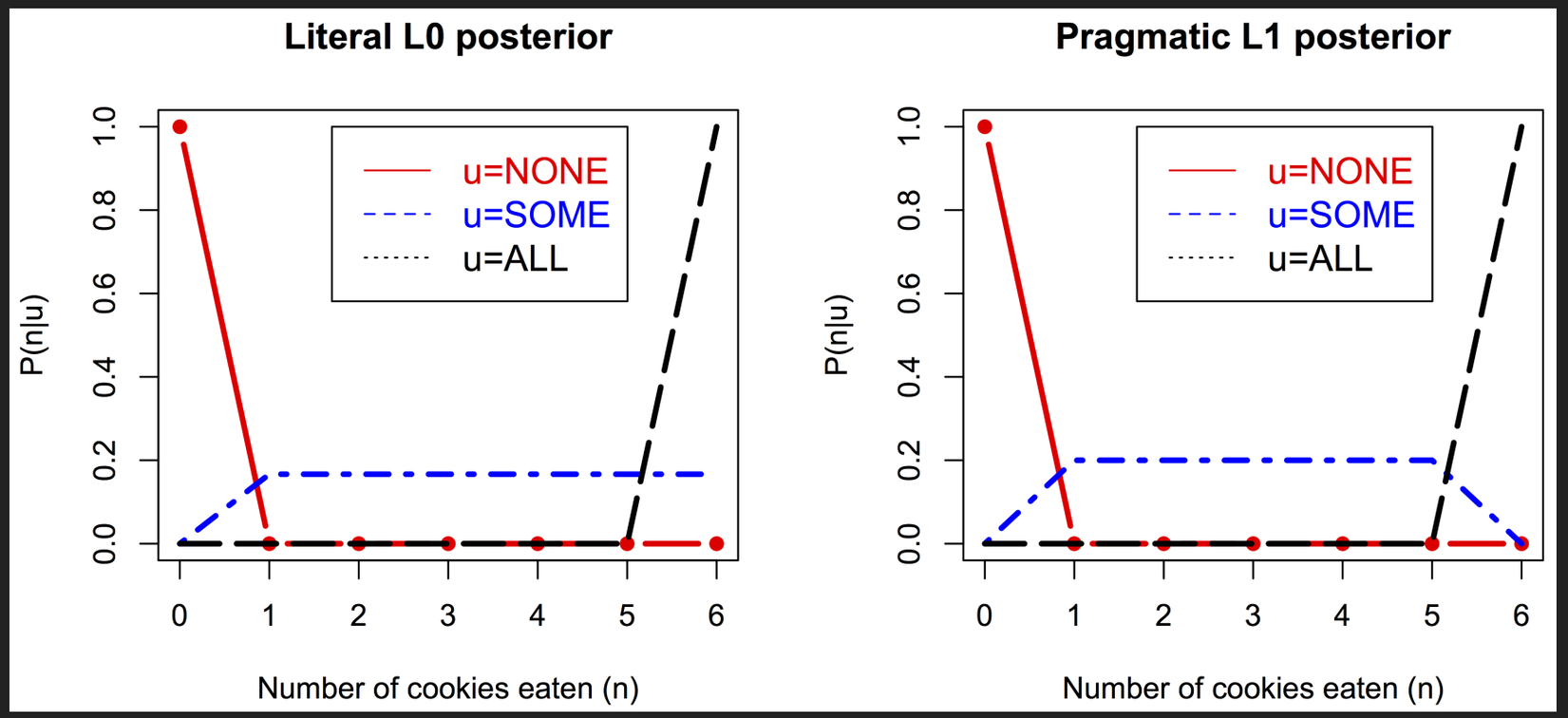
PRAGMATIC SPEAKER S_1

LITERAL LISTENER L_0

$$P_{L_1}(A|u) = \frac{P_{S_1}(u|A) \cdot P_{L_1}(A)}{\sum_{A'} P_{S_1}(u|A') \cdot P_{L_1}(A')}$$

$$P_{S_1}(u|A) = \frac{\exp(\lambda \cdot U_{S_1}(u; A))}{\sum_{u' \in \mathbf{ALT}} \exp(\lambda \cdot U_{S_1}(u'; A))}$$

{∅, 'some', 'all'}



MCMAHAN & STONE (2015)

Thresholds:

$$\tau_k^{Lower,d} \sim \mu_k^{Lower,d} - \Gamma(\alpha_k^{Lower,d}, \beta_k^{Lower,d})$$
$$\tau_k^{Upper,d} \sim \mu_k^{Upper,d} + \Gamma(\alpha_k^{Upper,d}, \beta_k^{Upper,d})$$

Probability of *x* falling in category *k*:

$$P(\tau_k^{Lower,H} < x^H < \tau_k^{Upper,H}) \times$$
$$P(\tau_k^{Lower,S} < x^S < \tau_k^{Upper,S}) \times$$
$$P(\tau_k^{Lower,V} < x^V < \tau_k^{Upper,V})$$
$$= \prod_d P(\tau_k^{L,d} < x_i^d < \tau_k^{U,d})$$

Availability & Applicability

$$P(k^{said}, k^{true}|x) = P(k^{said}|k^{true})P(k^{true}|x)$$

$$P(k^{said}, k^{true}|\mathbf{x}) = \alpha_k \prod_d \phi_k^d(x^d)$$

$$\phi_k^d(x^d) = \begin{cases} P(x^d > \tau_k^{L,d}), & x^d \leq \mu_k^{L,d} \\ P(x^d < \tau_k^{U,d}), & x^d \geq \mu_k^{U,d} \\ 1, & otherwise \end{cases}$$

$$\alpha_k = \frac{P(k^{said}, k^{true})}{P(k^{true})}$$
$$= \frac{count(k)/N}{\int_x P(k^{true}|x)P(x)}$$

MCMAHAN & STONE (2015)

HM: Histogram model (bins colorspace and counts frequency)

GM: $P(x|k^{true})$ Gaussian model w/ diagonal covariance

$$P(k^{said}, k^{true}|x) \propto P(x|k^{true})P(k^{said}, k^{true})$$

	TOP^1	TOP^5	TOP^{10}
LUX	39.55%	69.80%	80.46%
HM	39.40%	71.89%	82.53%
GM	39.05%	69.25%	79.99%

Table 1: Decision-based results. The percentage of correct responses of 544,764 test-set data points are shown.

	$-LL$	$-LLV$	AIC	Perp
LUX	$1.13*10^7$	$2.05*10^6$	$4.13*10^6$	13.61
HM	$1.13*10^7$	$2.09*10^6$	$4.82*10^6$	14.41
GM	$1.34*10^7$	$2.08*10^6$	$4.17*10^6$	14.14

Table 2: Likelihood-based evaluation results: negative log likelihood of the data, negative log likelihood of labels given points, number of parameters, Akaike Information Criterion and perplexity of labels given color values. Parameter counts for AIC are 15751 for LUX, 315669 for HM and 5803 for GM.

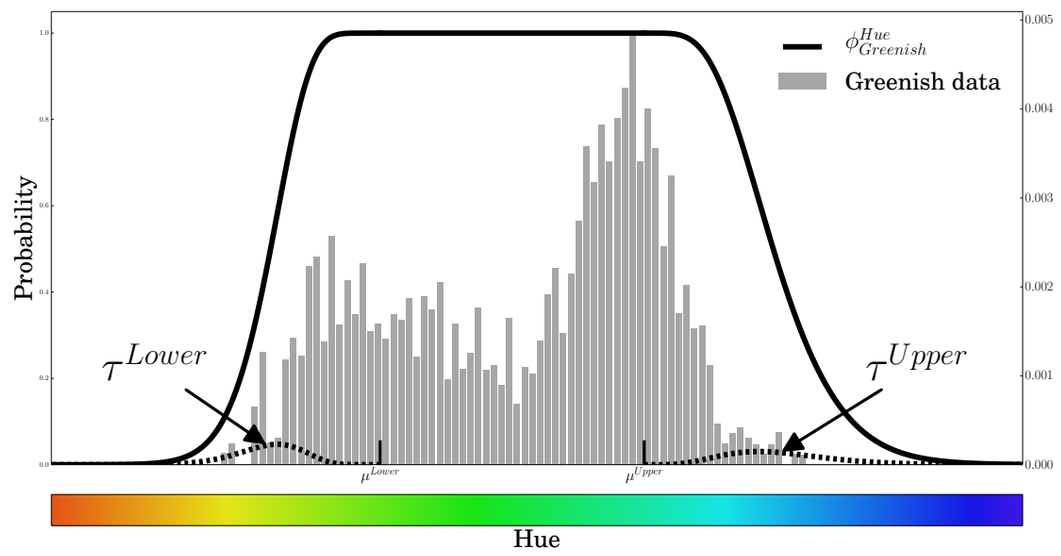
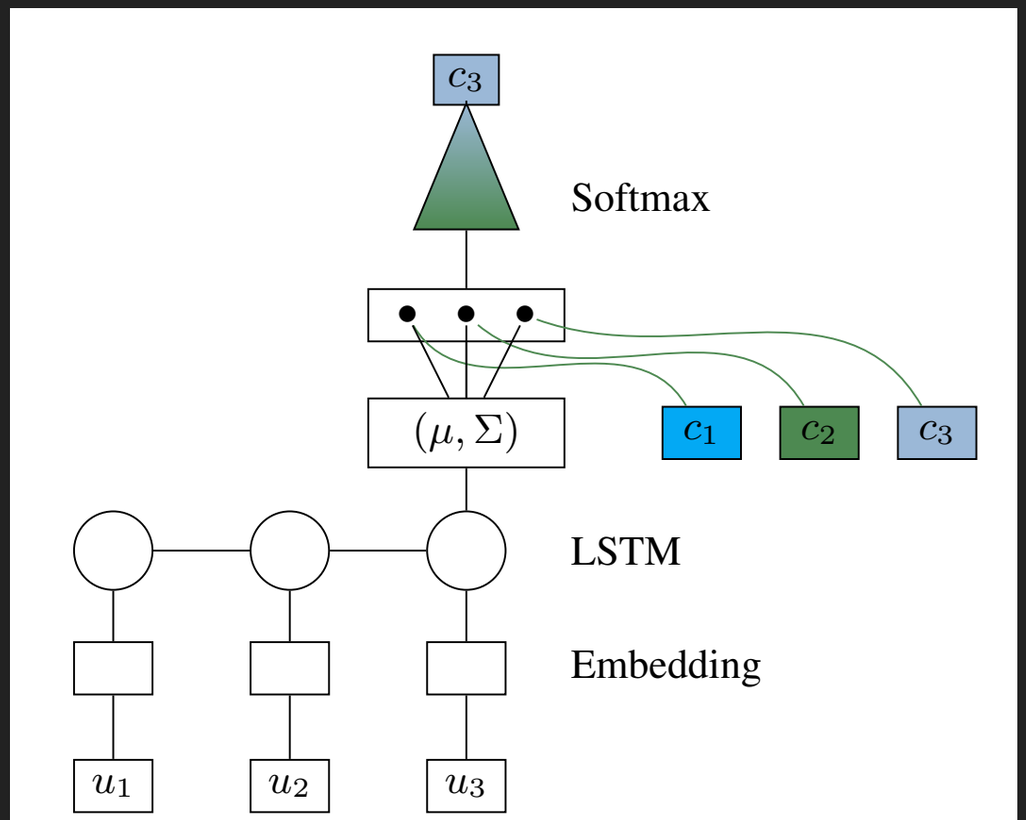
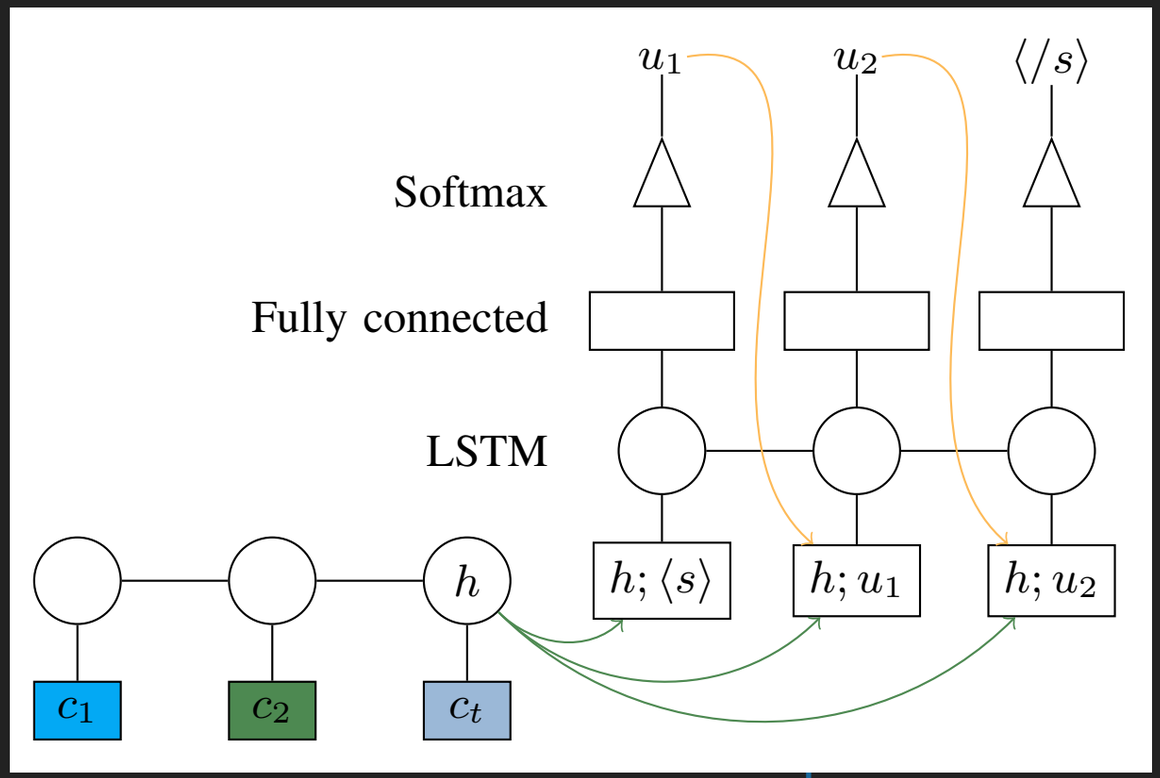


Figure 6: For the Hue dimension, the data for "greenish" is plotted against the LUX model's ϕ curve.

MONROE ET AL (2017)





LITERAL LISTENER L_0

$L_0(t \mid u', C; \theta)$

$S_0(u \mid t, C; \phi)$

LITERAL SPEAKER S_0

MONROE ET AL (2017)

PRAGMATIC LISTENER L_2

$$L_2(t \mid u, C; \theta) = \frac{S_1(u|t,C;\theta)}{\sum_{t'} S_1(u|t',C;\theta)}$$

PRAGMATIC SPEAKER S_1

$$S_1(u \mid t, C; \theta) = \frac{L_0(t|u,C;\theta)}{\sum_{u'} L_0(t|u',C;\theta)}$$

LITERAL LISTENER L_0

$$L_0(t \mid u', C; \theta)$$

$$L_a \propto L_0^{\beta_a} \cdot L_1^{1-\beta_a}$$

$$L_b \propto L_0{}^{\beta_b} \cdot L_2{}^{1-\beta_b}$$

$$L_e \propto L_a{}^{\gamma} \cdot L_b{}^{1-\gamma}$$

$$L_1(t \mid u, C; \theta) = \frac{S_0(u|t,C;\theta)}{\sum_{t'} S_0(u|t',C;\theta)}$$

PRAGMATIC LISTENER L_1

LITERAL SPEAKER S_0

$$S_0(u \mid t, C; \phi)$$

MONROE ET AL (2017)

L_0			
drab green not the bluer one	<1	<1	>**99**
gray	**96**	4	<1
blue dull green	24	**76**	<1
blue	<1	>**99**	<1
bluish	<1	>**99**	<1
green	4	1	**95**
yellow	<1	<1	>**99**
S_1			
drab green not the bluer one	1	<1	**34**
gray	**58**	5	<1
blue dull green	27	28	<1
blue	2	**32**	<1
bluish	1	**32**	<1
green	10	3	33
yellow	<1	<1	**34**
L_2			
drab green not the bluer one	5	<1	**95**
S_0 ($\times 10^{-9}$)	***5.85***	*0.38*	*<0.01*
L_1	**94**	6	<1
L_a	**92**	6	2
L_b	8	1	**91**
L_e	**63**	6	32

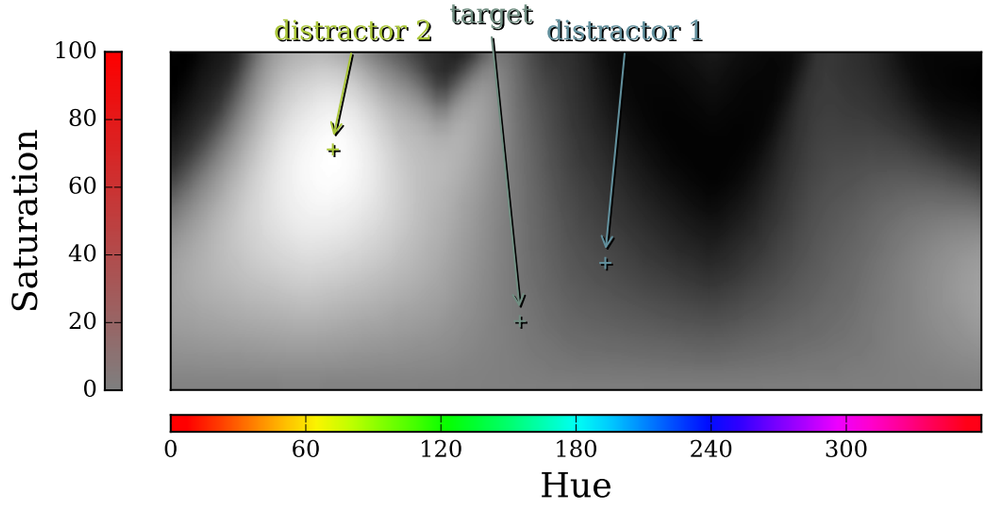
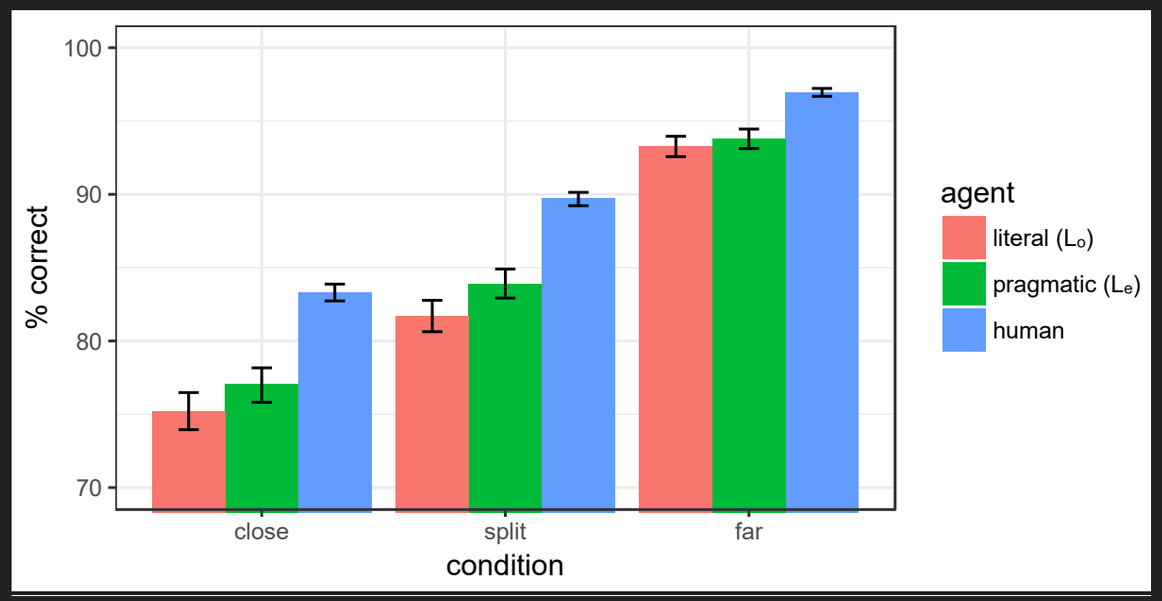


Figure 6: L_0's log marginal probability density, marginalizing over V (value) in HSV space, of color conditioned on the utterance *drab green not the bluer one*. White regions have higher probability. Labeled colors are the three colors from the right column of Figure 5.



MONROE ET AL (2017)

	human			S_0			S_1		
	far	split	close	far	split	close	far	split	close
# Chars	7.8	12.3	14.9	9.0	12.8	16.6	9.0	12.8	16.4
# Words	1.7	2.7	3.3	2.0	2.8	3.7	2.0	2.8	3.7
% Comparatives	1.7	14.2	12.8	3.6	8.8	13.1	4.2	9.0	13.7
% High Specificity	7.0	7.6	7.4	6.4	8.4	7.6	6.8	7.9	7.5
% Negatives	2.8	10.0	12.9	4.8	8.9	13.3	4.4	8.5	14.1
% Superlatives	2.2	6.1	16.7	4.7	9.7	17.2	4.8	10.3	16.6

Table 2: Corpus statistics and statistics of samples from artificial speakers (rates per utterance). S_0: RNN speaker; S_1: pragmatic speaker derived from RNN listener (see Section 4.3). The human and artificial speakers show many of the same correlations between language use and context type.

model	accuracy (%)	perplexity
L_0	83.30	1.73
$L_1 = L(S_0)$	80.51	1.59
$L_2 = L(S(L_0))$	83.95	1.51
$L_a = L_0 \cdot L_1$	84.72	1.47
$L_b = L_0 \cdot L_2$	83.98	1.50
$L_e = L_a \cdot L_b$	**84.84**	**1.45**
human	*90.40*	
L_0	85.08	1.62
L_e	**86.98**	**1.39**
human	*91.08*	

HERBELOT & VECCHI (2015)

Space	*# train vec.*	*# test vec.*	*# dims*	*# test inst.*
MT_{QMR}	400	141	2172	1570
MT_{AD}	60	12	54	648
MT_{QMR+AD}	410	145	2193	1595

Model-Theoretic		*Distributional*		
train	*test*	DS_{cooc}	$DS_{Mikolov}$	*human*
MT_{QMR}	MT_{QMR}	0.350	0.346	0.624
MT_{AD}	MT_{AD}	**0.641**	0.634	–
MT_{QMR+AD}	MT_{QMR+AD}	0.569	0.523	–
MT_{QMR+AD}	$\text{MT}_{animals}$	**0.663**	0.612	–
MT_{QMR+AD}	$\text{MT}_{no\text{-}animals}$	0.353	0.341	–
MT_{QMR}	$\text{MT}_{QMR^{animals}}$	0.419	0.405	–
MT_{QMR+AD}	$\text{MT}_{QMR^{animals}}$	**0.666**	0.600	0.663

	% of gold in...
top 5 neighbours	19% (27/145)
top 10 neighbours	29% (42/145)
top 20 neighbours	46% (67/145)

Table 4: Percentage of gold vectors found in the top neighbours of the mapped concepts, shown for the $\text{DS}_{cooc} \rightarrow \text{MT}_{QMR+AD}$ transformation.

		Gold				
		no	few	some	most	all
Mapped	no	0	-0.05	-0.35	-0.95	-1
	few	-0.05	0	0.2	0.9	0.95
	some	-0.35	-0.2	0	0.6	0.65
	most	-0.95	-0.9	-0.6	0	0.05
	all	-1	-0.95	-0.65	-0.05	0

Table 7: Distance matrix for the evaluation of the natural language quantifiers generation step.

		Gold				
		no	few	some	most	all
Mapped	no	238	66	20	4	2
	few	53	45	30	19	12
	some	6	1	2	3	2
	most	4	6	4	16	56
	all	0	0	0	2	3

Table 8: Confusion matrix for the results of the natural language quantifiers generation.

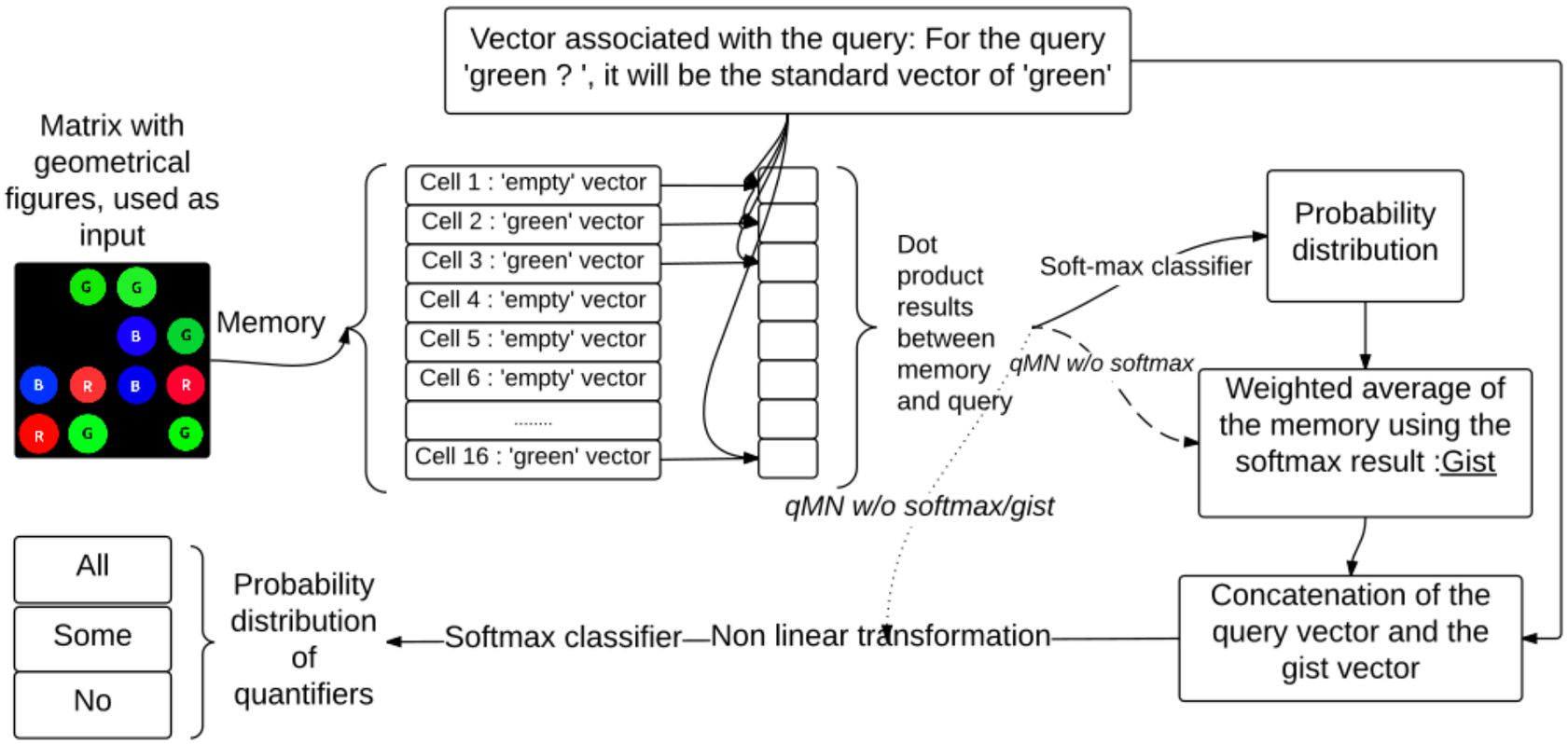
HERBELOT & VECCHI (2015)

Instance	Mapped	Gold
raven a_bird	most	all
pigeon has_hair	few	no
elephant has_eyes	most	all
crab is_blind	few	few
snail a_predator	no	no
octopus is_stout	no	few
turtle roosts	no	few
moose is_yellow	no	no
cobra hunted_by_people	some	some
snail forages	few	no
chicken is_nocturnal	few	no
moose has_a_heart	most	all
pigeon hunted_by_people	no	few
cobra bites	few	most

Table 9: Examples of mapped concept-predicate pairs

plum	*cottage*
a_fruit	has_a_roof
grows_on_trees	used_for_shelter*
tastes_sweet	has_doors*
is_edible	a_house
is_round	has_windows
is_small	is_small
has_skin	a_building*
is_juicy	used_for_living_in
tastes_good	made_of_wood*
has_seeds*	made_by_humans*
is_green*	worn_on_feet*
has_peel*	has_rooms*
is_orange*	used_for_storing_farm_equipment*
is_citrus*	found_on_farms*
is_yellow*	found_in_the_country
has_vitamin_C*	an_appliance*
has_leaves*	has_tenants*
has_a_pit	has_a_bathroom*
has_a_stem*	requires_rent*
grows_in_warm_climates*	requires_a_landlord*

SORODOC ET AL (2016)



Models	familiar	unseen quantities	unseen colors
RNN	65.7	62.0	49.7
Counting	86.5	78.4	32.8
qMN	**88.8**	**97.0**	**54.9**
-softmax	85.9	66.6	54.4
-softmax/gist	51.4	51.8	44.4

Table 1: Model accuracies (in %).

Counting:
Image is 16-D vector (one dimension per color plus empty cell)
Value is freq. of color scaled by color similarity (colors in image governed by small Gaussian to add noise)

PEZZELLE ET AL (2016)

Train-q				Train-c			
no	few	most	all	one	two	three	four
0/1	1/6	2/3	1/1	1/1	2/2	3/3	4/4
0/2	2/5	3/4	2/2	1/3	2/3	3/4	4/5
0/3	2/7	3/5	3/3	1/4	2/5	3/5	4/6
0/4	3/8	4/5	4/4	1/6	2/7	3/8	4/7
Test-q				**Test-c**			
no	few	most	all	one	two	three	four
0/5	1/7	4/6	5/5	1/2	2/4	3/7	4/8
0/8	4/9	6/8	9/9	1/7	2/9	3/9	4/9

Table 1: Combinations in Train and Test.

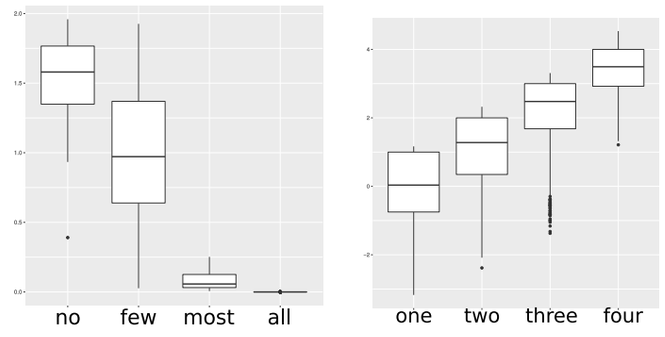


Figure 2: Left: quantifiers against cosine distance. Right: cardinals against dot product.

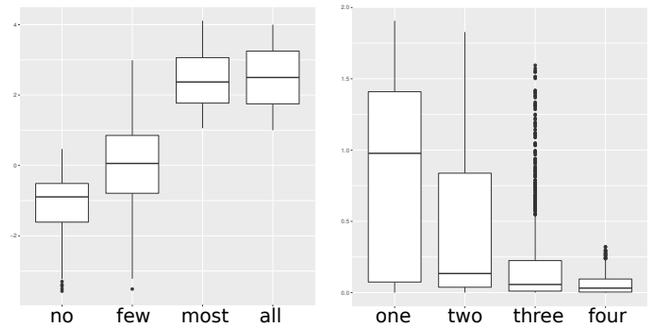


Figure 3: Left: quantifiers against dot product. Right: cardinals against cosine distance.

	lin		nn-cos		nn-dot	
	mAP	*P2*	*mAP*	*P2*	*mAP*	*P2*
no	0.78	0.65	**0.87**	0.77	0.54	0.37
few	0.59	0.39	**0.68**	0.51	0.59	0.43
most	0.61	0.36	0.60	0.29	**0.62**	0.45
all	0.75	0.66	**1**	1	0.33	0.12
one	0.44	0.30	0.38	0.21	**0.61**	0.45
two	0.35	0.15	0.38	0.21	**0.57**	0.43
three	0.38	0.16	0.36	0.13	**0.56**	0.40
four	0.65	0.47	0.75	0.60	**0.76**	0.61

Table 2: R-target. *mAP* and *P2* for each model.

	no	few	most	all
no	**288**	88	0	0
few	141	**191**	38	6
most	0	0	111	265
all	0	0	0	**376**
	one	two	three	four
one	**168**	113	54	41
two	64	**136**	124	52
three	23	80	130	145
four	10	24	72	**272**

Table 3: Top: Q nn-cos, number of cases retrieved in top-2 positions. Bottom: same for C nn-dot.

MITCHELL & LAPATA (2010)

Model	Function
Additive	$p_i = u_i + v_i$
Kintsch	$p_i = u_i + v_i + n_i$
Multiplicative	$p_i = u_i \cdot v_i$
Tensor product	$p_{i,j} = u_i \cdot v_j$
Circular convolution	$p_i = \sum_j u_j \cdot v_{i-j}$
Weighted additive	$p_i = \alpha v_i + \beta u_i$
Dilation	$p_i = v_i \sum_j u_j u_j + (\lambda - 1) u_i \sum_j u_j v_j$
Head only	$p_i = v_i$
Target unit	$p_i = v_i(t_1 t_2)$

- Training set creation:
  - Phrase similarity: sum of similarity of constituents

MITCHEL & LAPATA (2010)

Table 6
Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ) using a simple semantic space

Model	Adjective–Noun	Noun–Noun	Verb–Object
Additive	.36	.39	.30
Kintsch	.32	.22	.29
Multiplicative	.46	.49	.37
Tensor product	.41	.36	.33
Convolution	.09	.05	.10
Weighted additive	.44	.41	.34
Dilation	.44	.41	.38
Target unit	.43	.34	.29
Head only	.43	.17	.24
Humans	.52	.49	.55

Table 7
Correlation coefficients of model predictions with subject similarity ratings (Spearman's ρ) using the LDA topic model

Model	Adjective–Noun	Noun–Noun	Verb–Object
Additive	.37	.45	.40
Kintsch	.30	.28	.33
Multiplicative	.25	.45	.34
Tensor product	.39	.43	.33
Convolution	.15	.17	.12
Weighted additive	.38	.46	.40
Dilation	.38	.45	.41
Head only	.35	.27	.17
Humans	.52	.49	.55

BARONI & ZAMPARELLI (2010)

American N	*black N*	*easy N*
Am. representative	black face	easy start
Am. territory	black hand	quick
Am. source	black (n)	little cost
green N	*historical N*	*mental N*
green (n)	historical	mental activity
red road	hist. event	mental experience
green colour	hist. content	mental energy
necessary N	*nice N*	*young N*
necessary	nice	youthful
necessary degree	good bit	young doctor
sufficient	nice break	young staff

Nearest neighbors of centroid ANs

bad luck	*electronic communication*	*historical map*
bad	elec. storage	topographical
bad weekend	elec. transmission	atlas
good spirit	purpose	hist. material
important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

Nearest neighbors of specific ANs

method	*25%*	*median*	*75%*
alm	17	170	$\geq$1K
add	27	257	$\geq$1K
noun	72	448	$\geq$1K
mult	279	$\geq$1K	$\geq$1K
slm	629	$\geq$1K	$\geq$1K
adj	$\geq$1K	$\geq$1K	$\geq$1K

Quartile ranks of observed ANs in cosine-ranked list of predicted AN vectors

BARONI & ZAMPARELLI (2010)

input	*purity*
matrix	73.7 (68.4-94.7)
centroid	73.7 (63.2-94.7)
vector	68.4 (63.2-89.5)
random	45.9 (36.8-57.9)

- Matrix: learned adjective matrix
- Centroid: center of all ANs containing the adjective
- Vector: traditional co-occurrence
- Random: constraint that no cluster is left empty

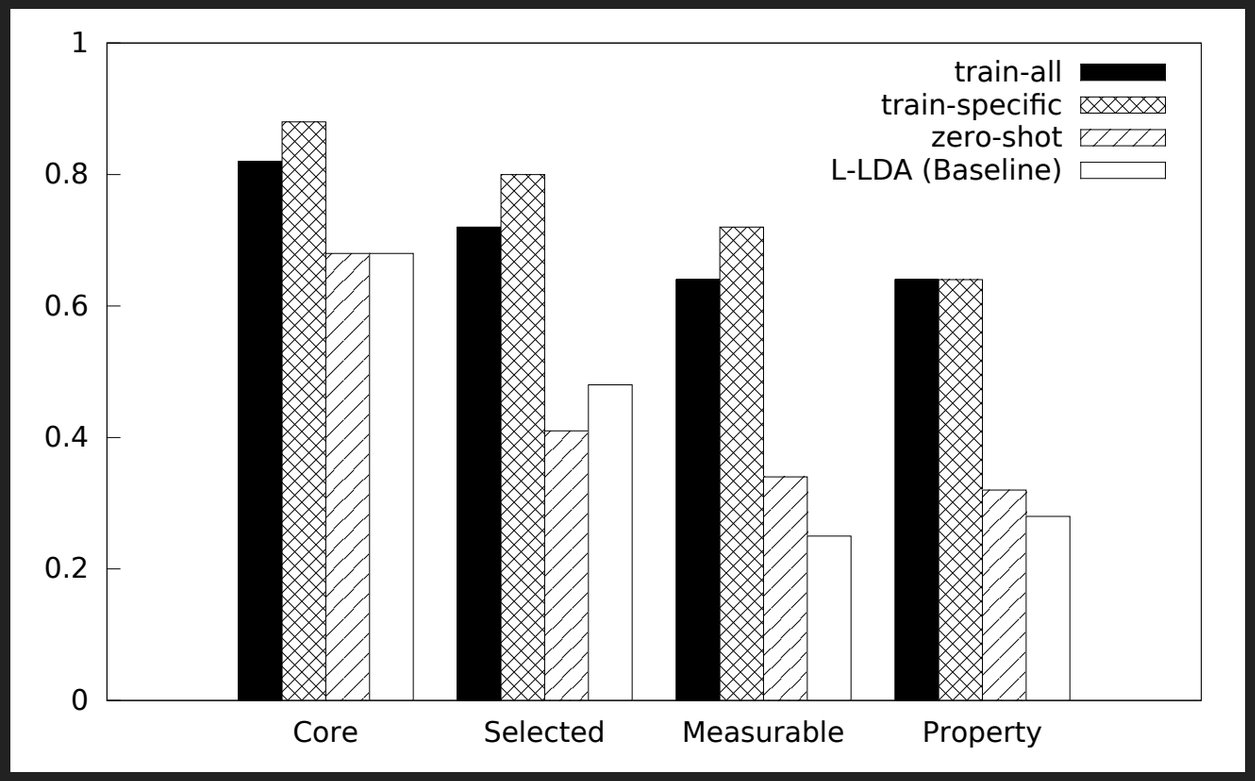
BARONI & ZAMPARELLI (2010)

- 36 adjectives:
  - size (big, great, huge, large, major, small, little),
  - denominal (American, European, national, mental, historical, electronic)
  - colors (white, black, red, green)
  - positive evaluation (nice, excellent, important, appropriate)
  - temporal (old, recent, new, young, current), modal (necessary, possible)
  - common abstract antonymous pairs (difficult, easy, good, bad, special, general, different, common)
- 1420 nouns (occurring ≥ 300 times w/ adjective)
- Semantic space
  - LMI scores of co-occurrence counts w/ 10k most common words
  - SVD to 300D

HARTUNG ET AL (2017)

Subset	Num. Attributes	Num. Train. Triples	Example Phrases
Core	10	72	*silvery hair* (COLOR), *huge wave* (SIZE), *longstanding conflict* (DURATION)
Selected	23	153	*sufficient food* (QUANTITY), *grave decision* (IMPORTANCE), *broad river* (WIDTH)
Measurable	65	261	*heavy load* (WEIGHT), *short hair* (LENGTH), *slow walker* (SPEED)
Property	73	300	*young people* (AGE), *high mountain* (HEIGHT), *straight line* (SHAPE)
All	254	869	*dry paint* (WETNESS), *scentless wisp* (SMELL), *vehement defense* (STRENGTH)

Table 1: Overview of subsets of attributes contained in HeiPLAS data, together with example phrases



	Compositional Model	P@1	P@5
predict models	Adjective	0.33	0.50
	Noun	0.03	0.10
	Vector Addition ($\oplus$)	0.24	0.45
	Weighted Vector Addition	0.33	0.51
	Vector Multiplication ($\odot$)	0.00	0.02
	Adj. Dilation ($\lambda = 2$)	0.06	0.18
	Noun Dilation ($\lambda = 2$)	0.33	0.51
	Full Add. Weighted Noun	0.33	0.54
	Full Add. Weighted Adjective	0.46	0.71
	Full Add. Weighted Adj. and Noun	**0.56**	**0.75**
	Trained Tensor Product ($\otimes$)	0.44	0.57
count	C-LDA (Hartung, 2015)	0.09	n/a
	L-LDA (Hartung, 2015)	0.16	n/a

HARTUNG ET AL (2017)

Underlying Word Representation	$\odot$	$\oplus$	Weighted Addition	Full Additive
word2vec	0.36	0.48	0.42	**0.50**
M&L-BoW	0.46	0.36	0.44	n/a
M&L-Topic	0.25	0.37	0.38	n/a
C-LDA	0.28	0.19	n/a	n/a

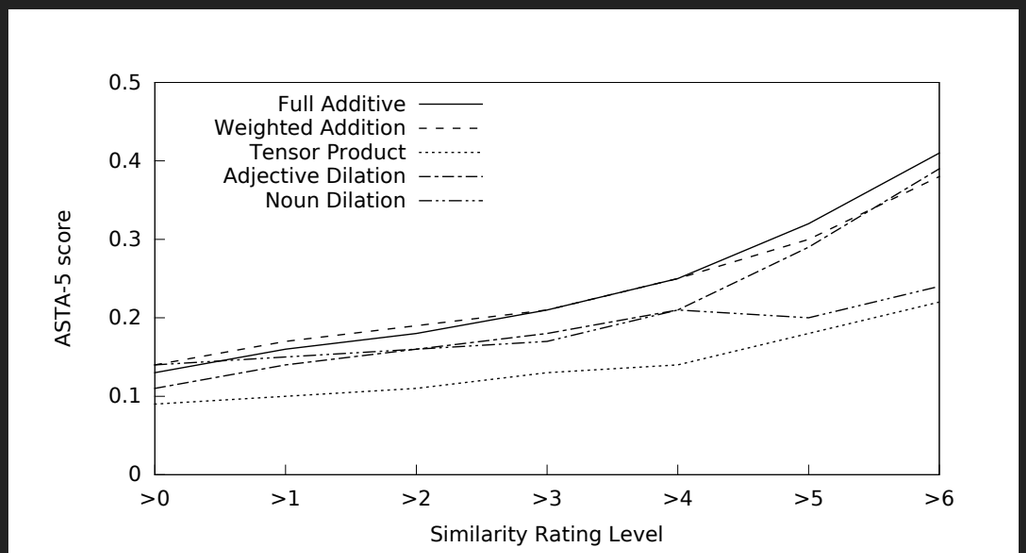


Figure 2: ASTA-5 scores over different levels of human similarity ratings (cf. Experiment 4)

BOLEDA ET AL (2013)

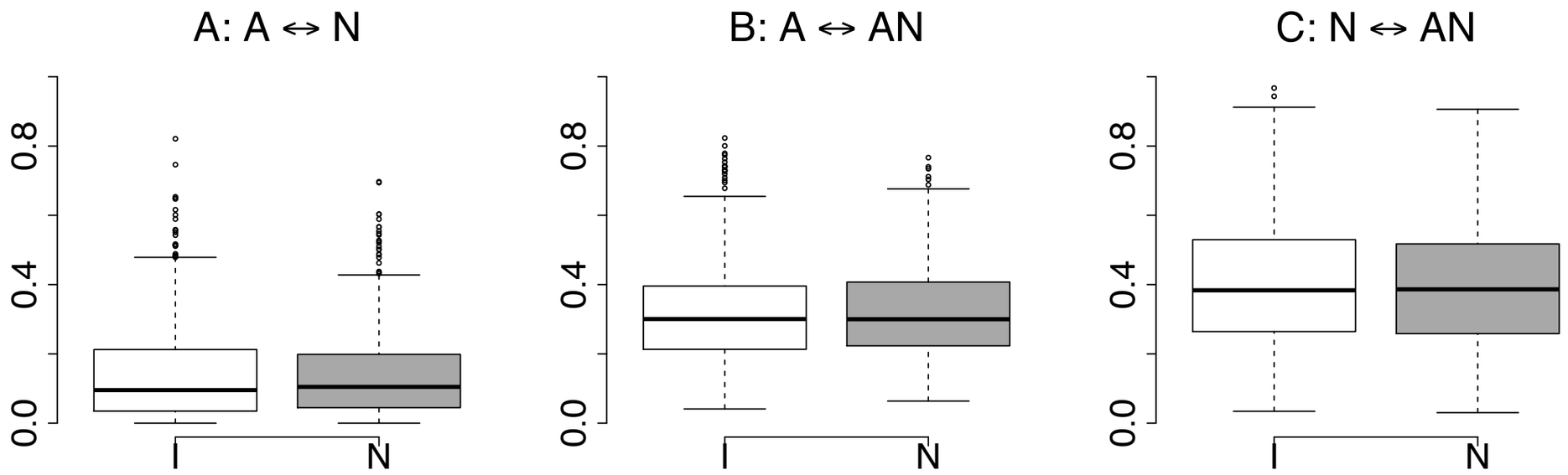


Figure 1: Distribution of cosines for observed vectors, by adjective type (intensional, I, or non-intensional, N). From left to right, adjective vs. noun, adjective vs. phrase, and noun vs. phrase cosines.

	Monosemous	**Polysemous**
I	*alleged accomplice*, *former surname*, *necessary competence*	*mock charge*, *putative point*, *past range*
N	*modern aircraft*, *severe hypertension*, *wide disparity*	*nasty review*, *ripe shock*, *meagre part*
	Typical	**Nontypical**
I	*former mayor*, *likely threat*, *alleged killer*	*former retreat*, *likely base*, *alleged fact*
N	*severe pain*, *free download*, *wide perspective*	*severe budget*, *free attention*, *wide detail*

VECCHI ET AL (2013)

	Measure	*t*	*sig.*	
CORP	cosA$_x$	2.478		
	cosA$_y$	-4.348	*	RO>FO
	cosN	4.656	*	FO>RO
	cosA$_x$N	5.913	*	FO>RO
	cosA$_y$N	1.970		
W.ADD	cosA$_x$	4.805	*	FO>RO
	cosA$_y$	-1.109		
	cosN	1.140		
	cosA$_x$N	1.059		
	cosA$_y$N	0.584		
F.ADD	cosA$_x$	2.050		
	cosA$_y$	-1.451		
	cosN	4.493	*	FO>RO
	cosA$_x$N	-0.445		
	cosA$_y$N	2.300		
MULT	cosA$_x$	3.830	*	FO>RO
	cosA$_y$	-0.503		
	cosN	5.090	*	FO>RO
	cosA$_x$N	4.435	*	FO>RO
	cosA$_y$N	3.900	*	FO>RO
LFM	cosA$_x$	-1.649		
	cosA$_y$	-1.272		
	cosN	5.539	*	FO>RO
	cosA$_x$N	3.336	*	FO>RO
	cosA$_y$N	4.215	*	FO>RO
ΔPMI		8.701	*	FO>RO

Table 4: **Flexible vs. Rigid Order AANs.** t-normalized differences between flexible order (FO) and rigid order (FO) mean cosines (or mean ΔPMI values) for corpus-extracted and model-generated vectors. For significant differences (p<0.05 after Bonferroni correction), the last column reports whether mean cosine (or ΔPMI) is larger for flexible order (FO) or rigid order (RO) class.

	Measure	*t*	*sig.*	
W.ADD	cosA$_x$	-7.840	*	U>A
	cosA$_y$	7.924	*	A>U
	cosN	2.394		
	cosA$_x$N	-5.462	*	U>A
	cosA$_y$N	3.627	*	A>U
F.ADD	cosA$_x$	-8.418	*	U>A
	cosA$_y$	6.534	*	A>U
	cosN	-1.927		
	cosA$_x$N	-3.583	*	U>A
	cosA$_y$N	-2.185		
MULT	cosA$_x$	-5.100	*	U>A
	cosA$_y$	5.100	*	A>U
	cosN	0.000		
	cosA$_x$N	-0.598		
	cosA$_y$N	0.598		
LFM	cosA$_x$	-7.498	*	U>A
	cosA$_y$	7.227	*	A>U
	cosN	-2.172		
	cosA$_x$N	-5.792	*	U>A
	cosA$_y$N	0.774		
ΔPMI		-11.448	*	U>A

Table 5: **Attested- vs. unattested-order rigid order AANs.** t-normalized mean paired cosine (or ΔPMI) differences between attested (A) and unattested (U) AANs with their components. For significant differences (paired t-test p<0.05 after Bonferroni correction), last column reports whether cosines (or ΔPMI) are on average larger for A or U.

national daily newspaper	*new regional government*
national newspaper	regional government
major newspaper	local reform
daily newspaper	regional council
daily national newspaper	*fresh organic vegetable*
national daily newspaper	organic vegetable
well-known journalist	organic fruit
weekly column	organic product

SHUTOVA ET AL (2016)

Features	Method	P	R	$F1$
Linguistic	WORDCOS	0.67	0.76	**0.71**
	PHRASCOS1	0.38	0.94	0.54
Visual	WORDCOS	0.49	0.97	0.65
	PHRASCOS1	0.56	0.79	0.66
Multimodal	WORDMID	0.56	0.86	0.68
	PHRASMID	0.44	0.93	0.59
	WORDLATE	0.49	0.96	0.65
	PHRASLATE	0.41	0.92	0.57
	MIXLATE	0.65	0.87	**0.75**

Table 1: System performance on Mohammad et al. dataset (MOH) in terms of precision (P), recall (R) and F-score ($F1$)

Features	Method	P	R	$F1$
Linguistic	WORDCOS	0.73	0.80	**0.76**
	PHRASCOS1	0.43	0.96	0.57
Visual	WORDCOS	0.50	0.95	0.66
	PHRASCOS1	0.60	0.91	**0.73**
Multimodal	WORDMID	0.59	0.85	0.70
	PHRASMID	0.54	0.93	0.68
	WORDLATE	0.69	0.72	0.70
	PHRASLATE	0.50	1.00	0.67
	MIXLATE	0.67	0.96	**0.79**

Table 2: System performance on Tsvetkov et al. test set (TSV-TEST) in terms of precision (P), recall (R) and F-score ($F1$)

LAZARIDOU ET AL (2016)

Ridge Regression:

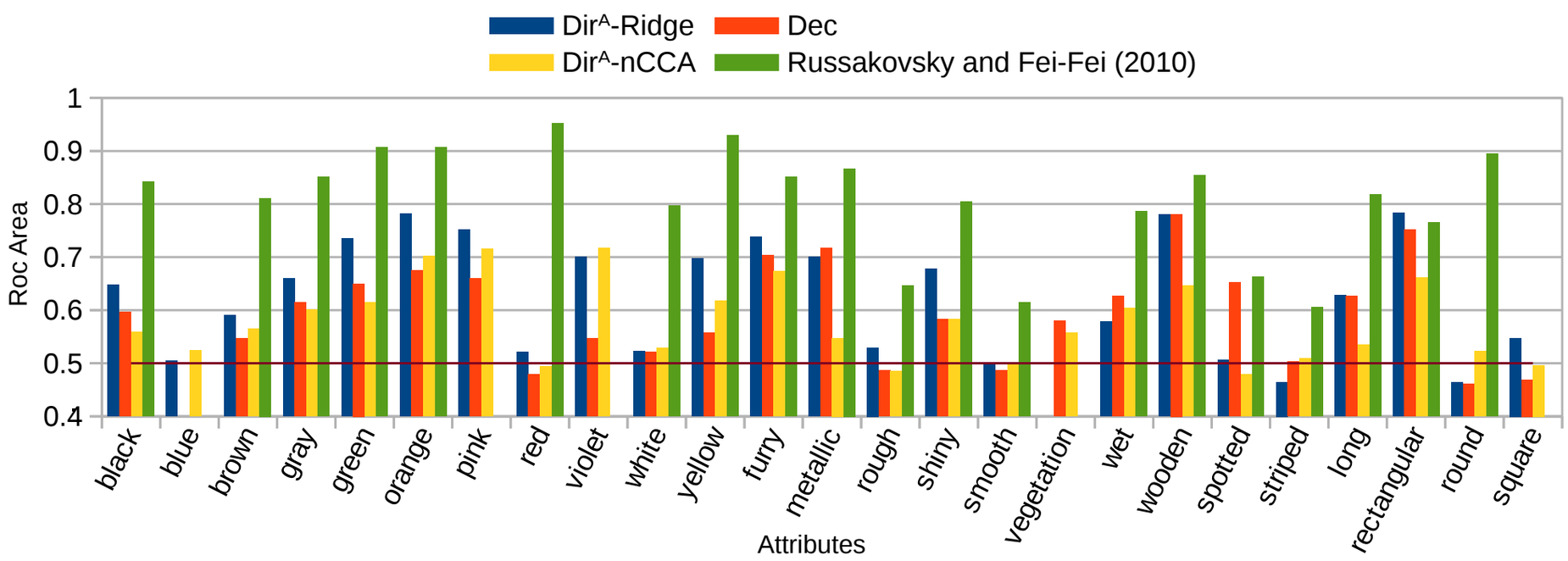
$$||W^{Tr} - F_{proj}V^{Tr}||_2^2 - ||\lambda F_{proj}||_2^2$$

Decomposition:

$$||[W_{adj}^{Tr}; W_{noun}^{Tr}] - F_{dec}W_{AN}^{Tr}||_2^2 - ||\lambda F_{dec}||_2^2$$

	Training			Evaluation		
	#im.	#attr.	#obj.	#im.	#attr.	#obj.
Exp. 1	10,749	97	-	*leave-one-attribute-out*		
Exp. 2	23,000	-	750	8,449	25	203

Table 3: Summary of training and evaluation sets.



LAZARIDOU ET AL (2016)

	LM	SP	vLM	DIR^{O}	DEC	DIR^{A}
@1	2	0	5	1	10	7
@5	5	7	16	4	31	23
@10	8	9	29	9	44	37
@20	18	17	50	19	59	51
@50	33	32	72	43	81	68
@100	56	55	82	67	89	77

Table 4: Percentage hit@k attribute retrieval scores.

	LM	SP	vLM	DIR^{O}	DEC
@1	1	0	2	0	4
@5	2	3	7	2	15
@10	3	5	15	4	23
@20	9	10	30	9	35
@50	20	20	49	22	59
@100	35	34	61	44	70

Table 5: Percentage recall@k attribute retrieval scores.

	DIR^{O}	DEC	DIR^{A}
@1	1	2	0
@5	3	10	0
@10	5	14	1
@20	9	20	2
@50	20	29	6
@100	33	41	12

Table 6: Percentage hit@k noun retrieval scores.

Concreteness: Average concreteness of the nouns the adjective modifies in the corpus

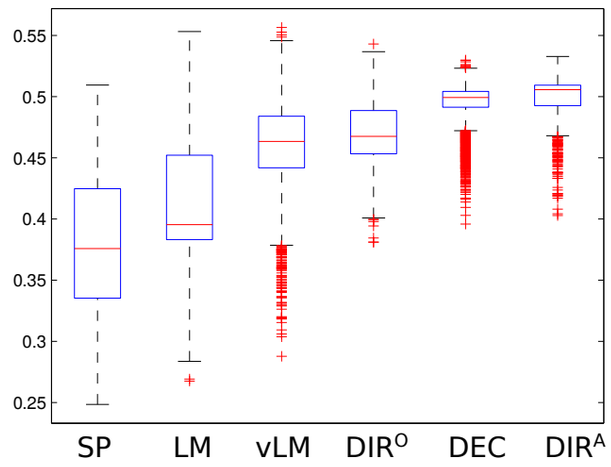


Figure 4: Distributions of (per-image) concreteness scores across different models. Red line marks median values, box edges correspond to 1st and 3rd quartiles, the wiskers extend to the most extreme data points and outliers are plotted individually.

LAZARIDOU ET AL (2016)

Attribute-based classification: object-trained method is improvement over standard BoVW features

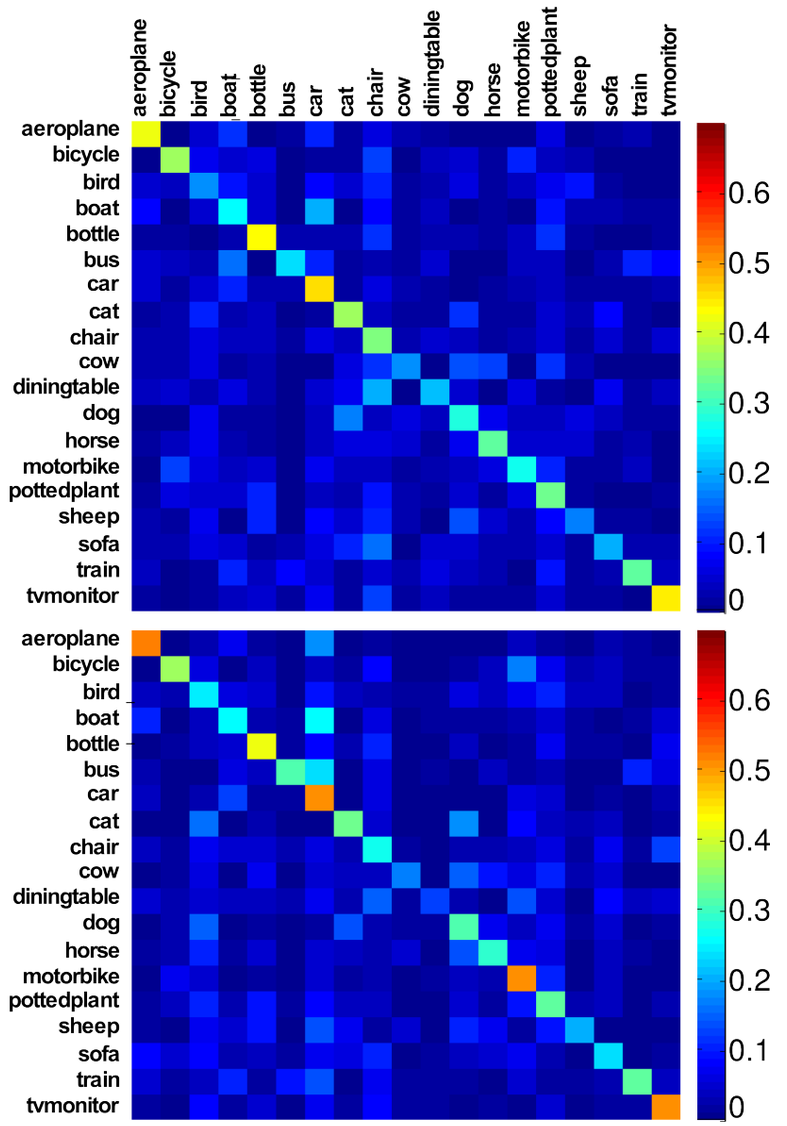


Figure 5: Confusion matrices for PHOW (**top**) and DEC (**bottom**). Warmer-color cells correspond to higher proportions of images with gold row label tagged by an algorithm with the column label (e.g., the first cells show that DEC tags a larger proportion of aeroplanes correctly).

COLLELL & MOENS (2016)

- Image representations:
  - (i) Averaging: Component-wise average of the CNN feature vectors of individual images. (i.e. cluster center of individual representations)
  - (ii) Maxpool: Computes the component-wise maximum of the CNN feature vectors of individual image (i.e. vector components “visual properties.”)

COLLELL & MOENS (2016)

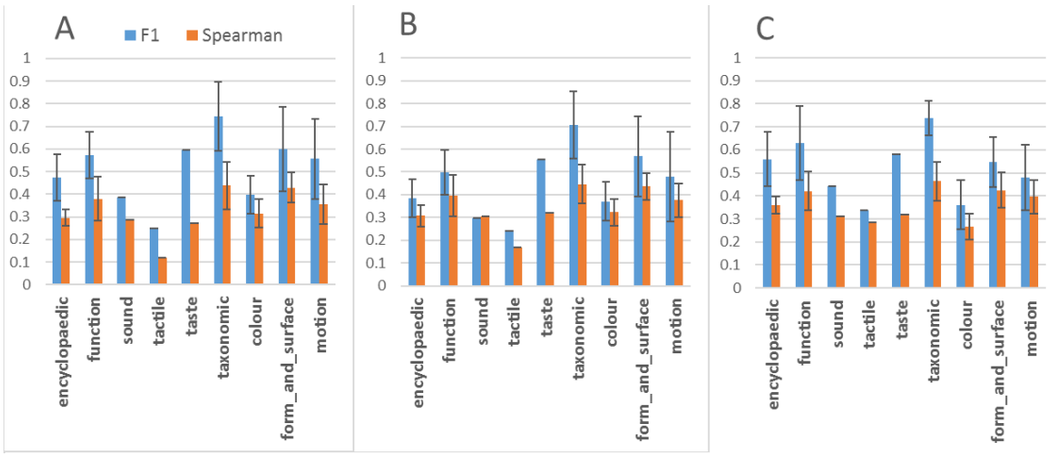


Figure 2: Averages of F1 (classification) and Spearman (regression) measures per attribute type (i.e., averaging individual attributes) for VIS_{avg} (A), VIS_{max} (B) and GloVe (C). Error bars show standard error.

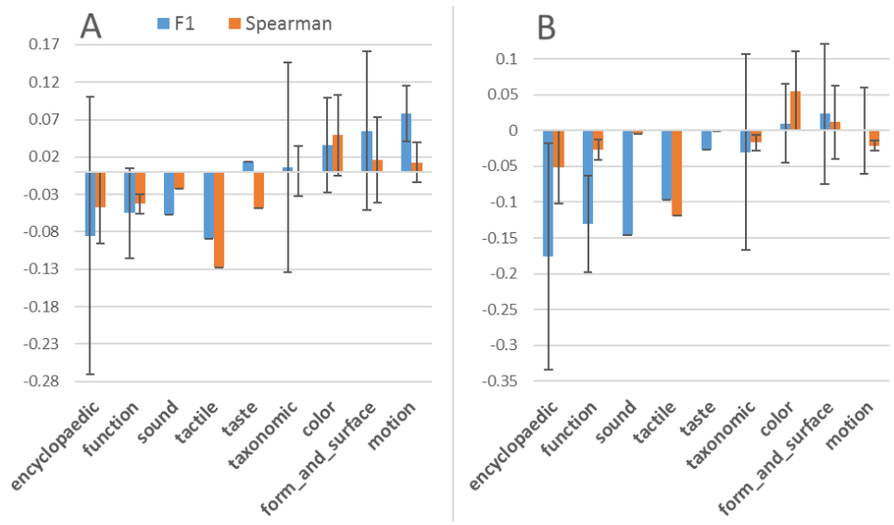


Figure 4: Averages of performance difference per attribute type. For each attribute type (e.g., taxonomic, taste, etc.), the bar indicates the average performance difference of its set of attributes. Plot A shows performance difference between VIS_{avg} and GloVe and B between VIS_{max} and GloVe. As in Fig. 3, positive bars indicate better performance of visual embeddings and negative bars otherwise. Error bars show standard error.

RSA – RATIONAL SPEECH ACT

$$s_0(u, \mid t, \mathcal{L}) \propto \mathcal{L}(u, t) e^{-\kappa(u)}$$

$$l_1(t \mid u, \mathcal{L}) \propto s_0(u, \mid t, \mathcal{L}) P(t)$$

- Pragmatic listener can start from a literal speaker or a pragmatic speaker
- Set of utterances U and $\mathcal{L}$ usually specified by hand

$$l_0(t \mid u, \mathcal{L}) \propto \mathcal{L}(u, t) P(t)$$

$$s_1(u \mid t, \mathcal{L}) \propto e^{\alpha \log(l_0(t, \mid u, \mathcal{L})) - \kappa(U)}$$

$$l_2(t \mid u, \mathcal{L}) \propto s_1(u, \mid t, \mathcal{L}) P(t)$$

- If U not finite, cannot