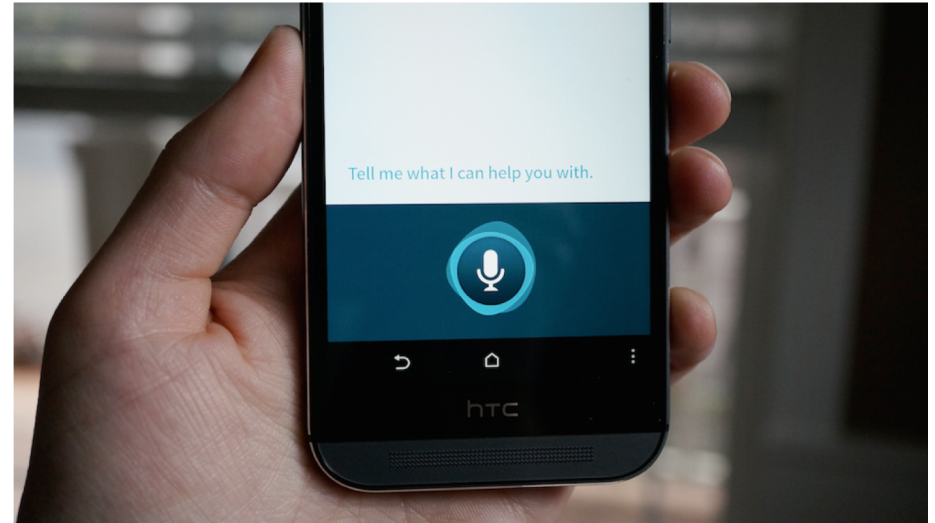


# Learning to Classify from Natural Language Explanations

*Shashank Srivastava*

*Joint work with Igor Labutov, Tom Mitchell*





## ***Is this email important?***

- *'Emails from my boss are usually important'*
- *'Such emails mention a deadline or a meeting'*
- *'The subject might say urgent ...'*

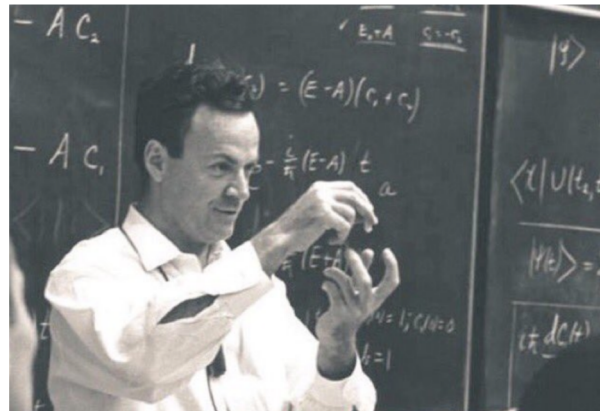
# Towards Conversational ML?

- Traditional dependence on 'big data'
  - Widely successful
  - Infeasible for long tail of learning problems
- Inherent statistical limitations
  - Coarsely,  $n \approx \log(H)$
  - Intractable for representations like ontologies
- Extend ML to richer forms of input
  - Explanations, instructions, clarifications ...



# Learning from Language

- Much of human learning is through language
  - Think books, lectures, student-teacher dialogue





# Why now?

- If there is a new publication relevant to my current project, email it to me
- Whenever it snows at night, wake me up 30 minutes earlier
- If I receive a late night email from my advisor, ring alarm at full blast



Every user can be a programmer

# Core issues

- Learning to Interpret NL

- Parsing of NL statements to formal semantic representations

*'Emails from my boss are usually important'*



*equals( email.sender, getContactEmail("boss") )*

Semantic parsing

- Using Language to Operationalize Learning

- E.g., Learning classification tasks from language



$\{0,1\}$

Binary classification

# How can language operationalize learning?

- ① *By defining expressive features* for learning tasks

*Joint Concept Learning and Semantic Parsing  
from Natural Language Explanations*

EMNLP 2017

- ② *By specifying model constraints* that can supervise training

*Zero-shot Learning of Classifiers from Natural  
Language Quantification*

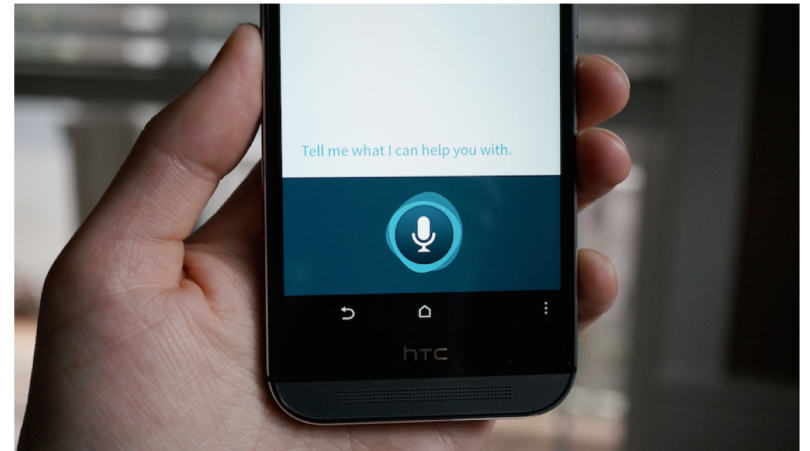
ACL 2018

*Part 1*: Defining features using NL  
explanations

# Defining features using NL

***Is this email important?***

*'Emails from my boss are usually important'*  
*'Such emails mention a deadline or a meeting'*  
*'The subject might say urgent ...'*



NL explanations



Executable feature  
functions



# NL Explanations as feature definitions

Semantic parsing maps NL to formal logical forms

## Natural language statement (s)

*'three less than twenty times six'*

*'What is the longest river that flows through Pittsburgh?'*

*'Phishing emails often mention prices'*

## Logical form (l)

`minus( prod(20, 6), 3 )`

`argmax( river(x)  $\wedge$  traverse(x,y)  $\wedge$  const(y, Pittsburgh), length)`

`findSemanticCategory( MONEY, field:body )`

**Evaluate** in a context ( $z = [l]_x$ )

117

Ohio

Yes/No

# How to interpret explanations?

- Pragmatics of language can guide parsing
  - A teacher's intention would be use discriminative statements

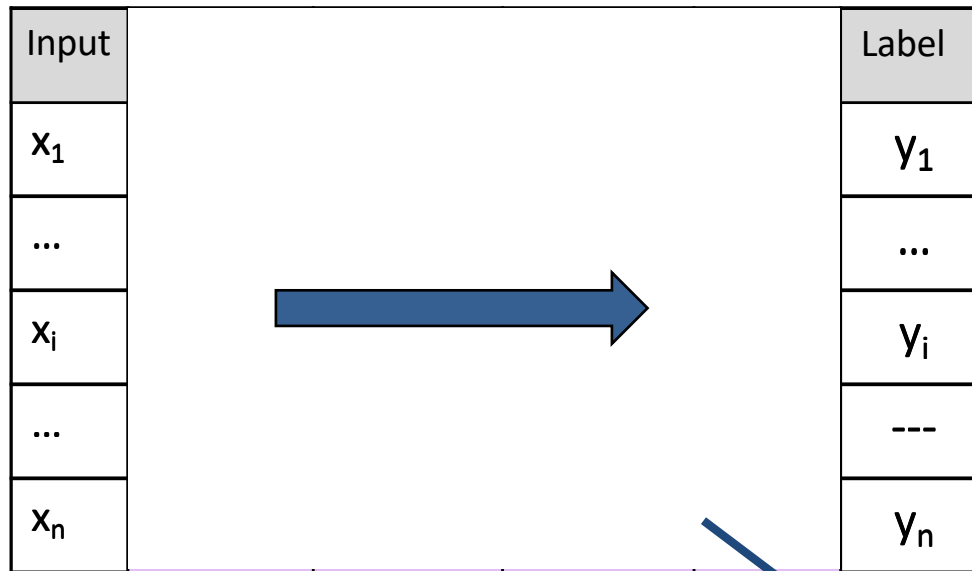
NL Explanation: *'Phishing emails often mention prices'*

Interpretation	Discriminative?
I1: findWord('prices', body)	X
I2: findSemanticCategory(cat:MONEY, body)	✓

Jointly learn a classifier and a semantic parser!

*Don't need annotated logical forms*

# Problem setting



No annotations of logical forms, supervision is only through concept labels  $\{0,1\}$  for examples

Latent variables

$s_i \rightarrow l_i$  (parsing)

$[l_i]_{x_j} \rightarrow z_{ij}$  (evaluation)

# Coupled parsing and concept classification

Input	$s_1$	$s_2$	$s_j$	$s_m$	Label
$x_1$	$z_{11}$	$z_{12}$	$z_{1j}$	$z_{1m}$	$y_1$
...					...
$x_i$	$z_{i1}$	$z_{i2}$	$z_{ij}$	$z_{im}$	$y_i$
...					---
$x_n$	$z_{n1}$	$z_{n2}$	$z_{nj}$	$z_{nm}$	$y_n$

$$\log P(y_i | x_i, s, \theta) = \log P(y_i | z_{i:}, \theta_{pred}) + \log P(z_{i:} | x_i, s, \theta_{parse})$$

Classifier

How likely are the observed concept labels, taking evaluations of NL statements as given?

Parser =  $\sum_{[l]_{x_i=z_{ij}}} P(l | s_j)$

How probable is a NL statement to apply for a given email (marginalized over all interpretations)?

# Model training

## ➤ Variational EM:

- **E- step:** Calculate estimates of  $z_{ij}$  (evaluations of statements in different contexts)

$$q_j(z_j) \propto \exp \left( \mathbb{E}_{j' \neq j} [\log p_{\theta_c}(y|\mathbf{z}, x)] + \log p_{\theta_p}(z_j|x, s_j) \right)$$

Prefer values that are  
discriminative

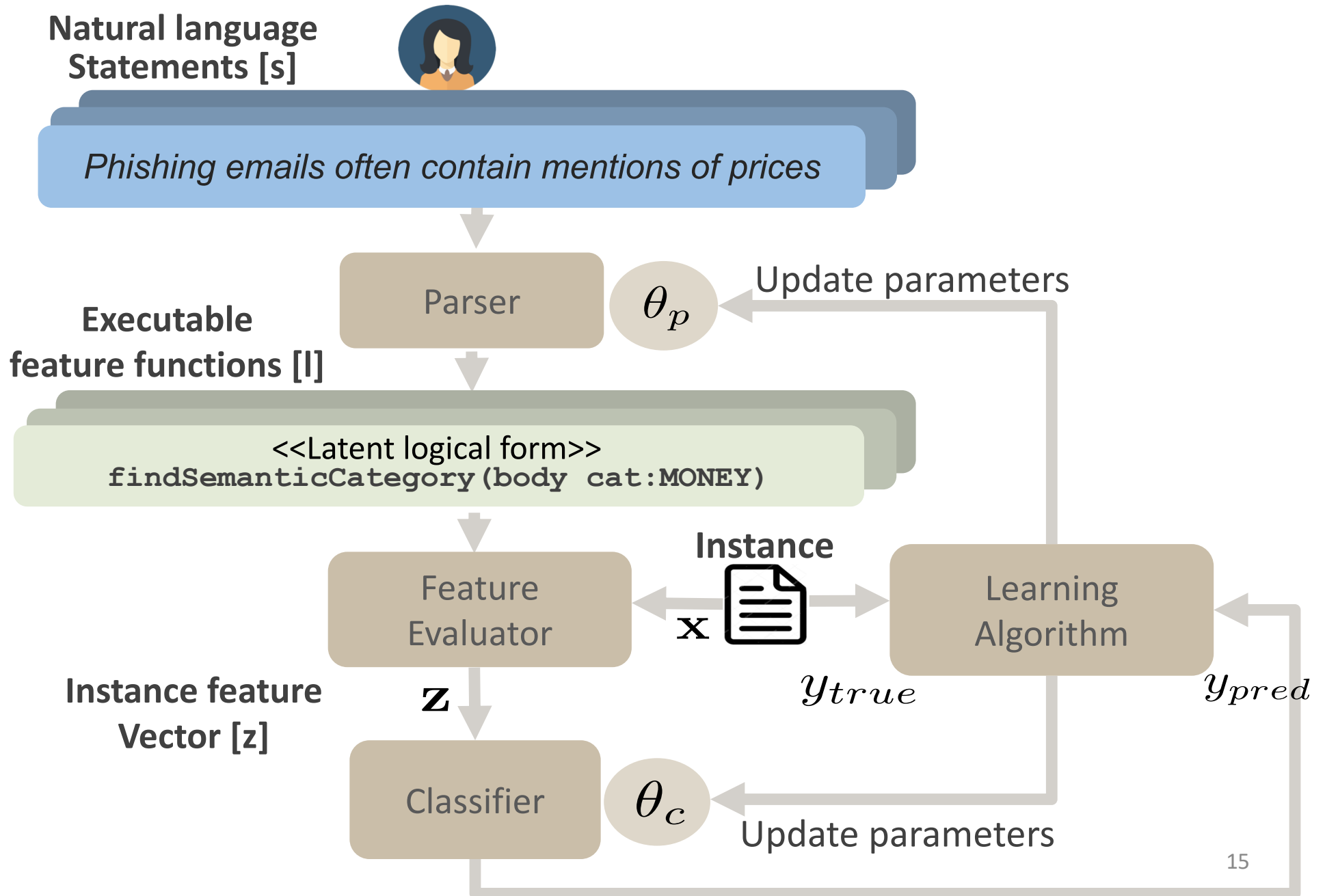
Prefer interpretations supported  
by linguistic evidence

Prefer interpretations of sentences that are both discriminative as well as supported by linguistic evidence

- **M- step:** Updates concept classifier and semantic parsing models taking  $z_{ij}$  's as given.



# Concept to Learn: Phishing Emails



# Data: Email classification

- Emails representing common email categories through AMT
  - Reminders, meeting invitations, requests from boss, internet humor, going out with friends, policy announcements, etc.
  - 1100 emails, 7 types

*E.g. You are writing an email to yourself as a reminder to do something*

<b>Subject:</b> Note to self - Move the Bodies
<b>From:</b> <a href="mailto:john@initech-corp.com">john@initech-corp.com</a>
<b>To:</b> <a href="mailto:john@initech-corp.com">john@initech-corp.com</a>
<b>Body:</b> Blasted police. I need to pick up lye and move the bodies tonight. Forecast is rain and the swamp's filling up. Need to remember galoshes, too.
<b>Attachment:</b> none

# Data: NL Explanations

- Dataset of statements explaining each concept
- Turkers describe emails from each category
- 30 statements for each category

## Sample explanations:

*Most reminders mention a date and a time in the message of the email*

*The sender of the email is the same as the recipient*

*These emails usually close with a name or title*

*These emails sometimes have jpg attachments*

*The email likely has words like "policy" or "announcement" in the subject*

*Emails from a public domain are not office requests*

From: mary@initech-corp.com

**Re: Pick client up from airport**

IMPORTANT!!!!

Monday at 1:00 pick the all call representative up from Gate 11 at the ORD airport. The meeting you are traveling to begins at 3:00. Make sure the client has had lunch.

---

★

To: mary@initech-corp.com

From: mary@initech-corp.com

**CU - timeline**

Need to file on or before January 13. Mailing is file date if cert enclosed in package. Target date for filing January 11. Also call Superior Ct to make sure they know not to remit record.

---

★

To: john@initech-corp.com

From: john@initech-corp.com

**Meeting reminder January**

On January the 18th there is going to be a meeting at 3:00pm, the whole team will be there. I am sending this email to myself as a reminder and also be sure to bring the papers me and Frank talked about. I am attaching them here to ensure I will have them where ever I am.

Attachment: Important\_Files\_January

---

★

To: Jordan

From: john@initech-corp.com

you get home.  
to deal with that  
at your house every

## READ FIRST! (read carefully)

The category of emails that you want to teach now is:

☞ "You are writing an email to yourself as a reminder to do something"

In order to help you in teaching this category, we have identified some **examples** and **NON-examples** of this category on the left. **Examples** of this category are highlighted in yellow and marked with a "star" (★) and the **NON-examples** are in gray. You should study these emails to get a better understanding of the category to help you teach it effectively.

**Your explanations should be based on the observations you make from the example emails!**

## DO NOT FORGET:

The examples shown are only to help you get an idea of how to explain this category. Your instructions should be **GENERAL** enough to help the assistant generalize to many future emails of this category!

Again, the category that you want to teach now is:

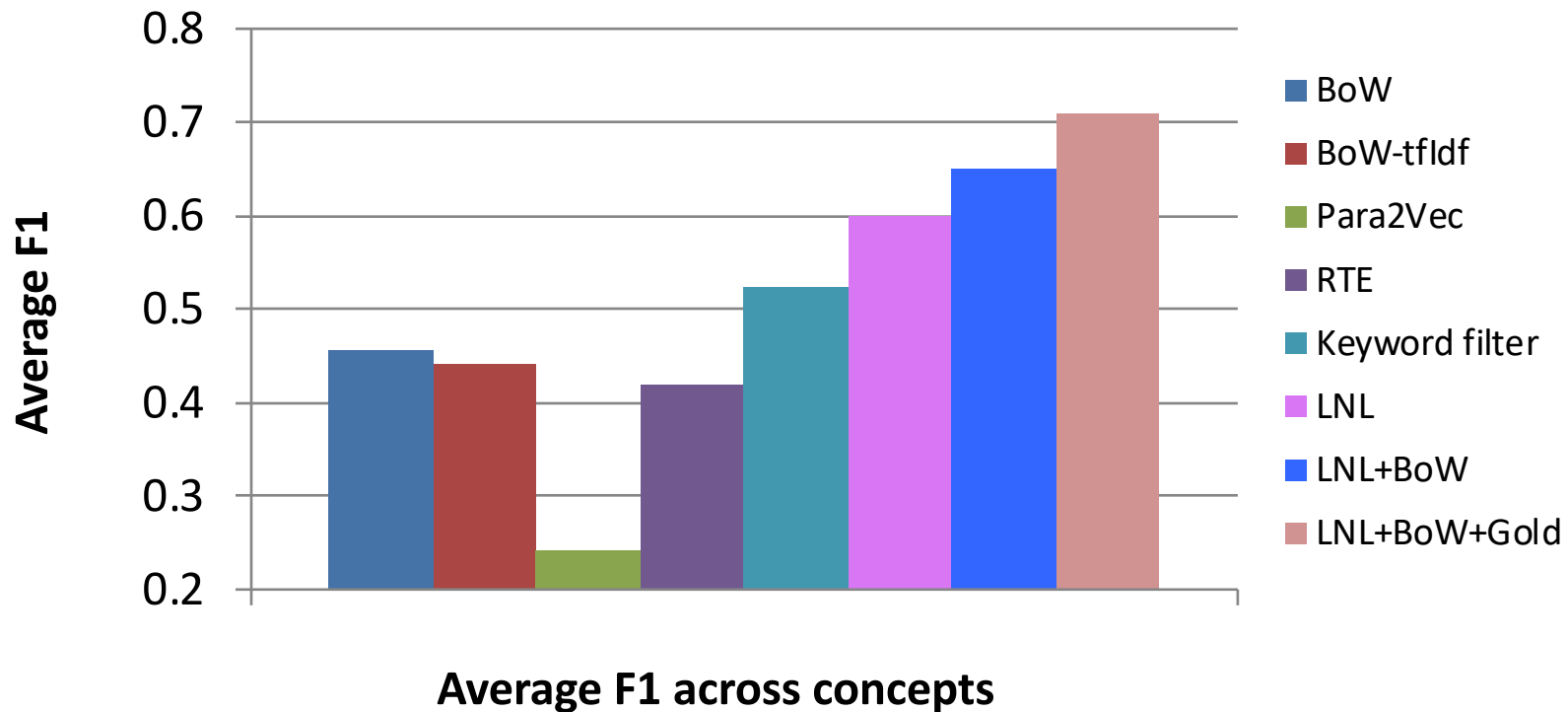
☞ "You are writing an email to yourself as a reminder to do something"

*Each instruction should not exceed the length of the text field*

*All instructions must be filled out*

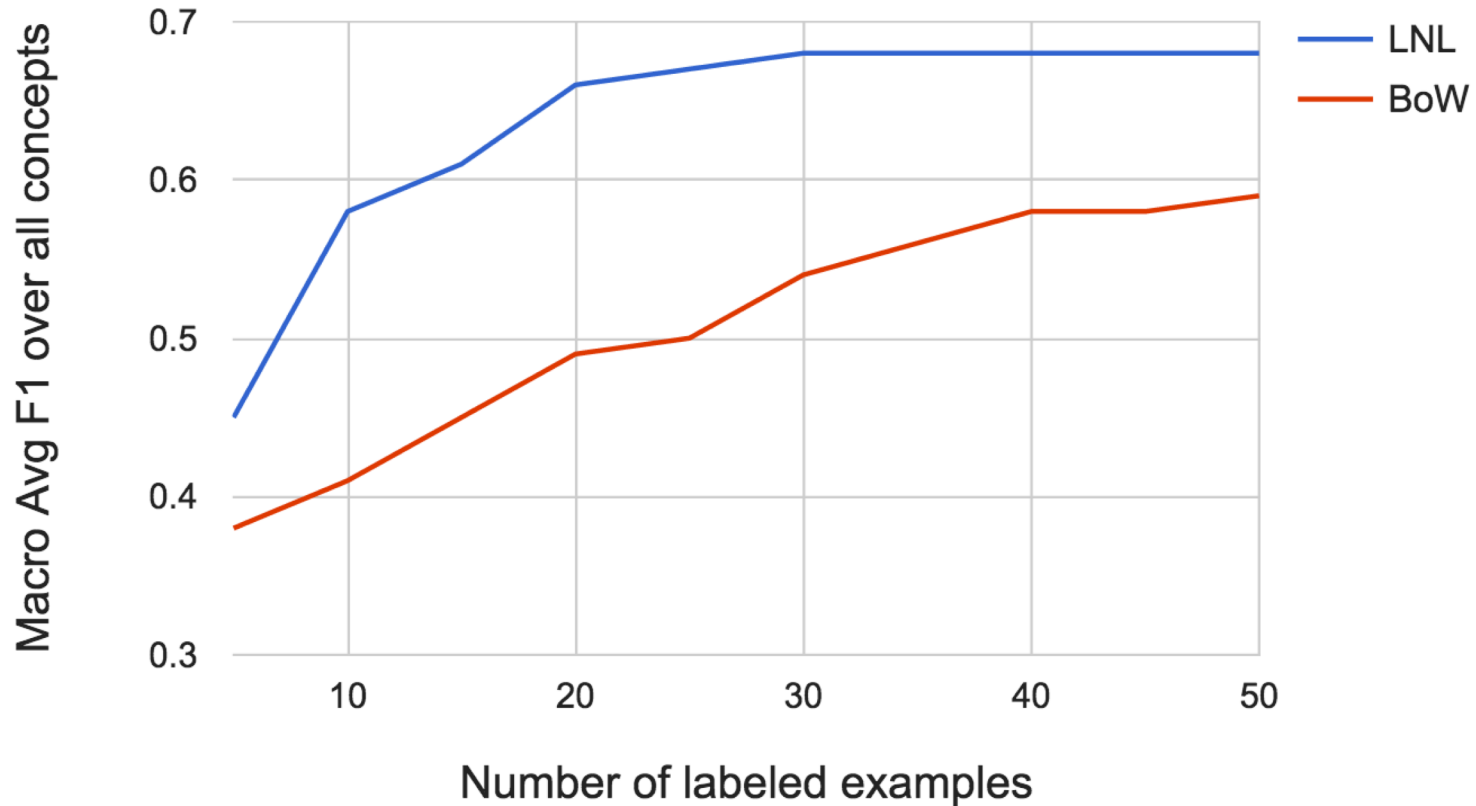
1:	how useful?	↓
2:	how useful?	↓
3:	how useful?	↓

# Results: Email classification



➤ Significantly better than best baseline for 6 of 7 categories

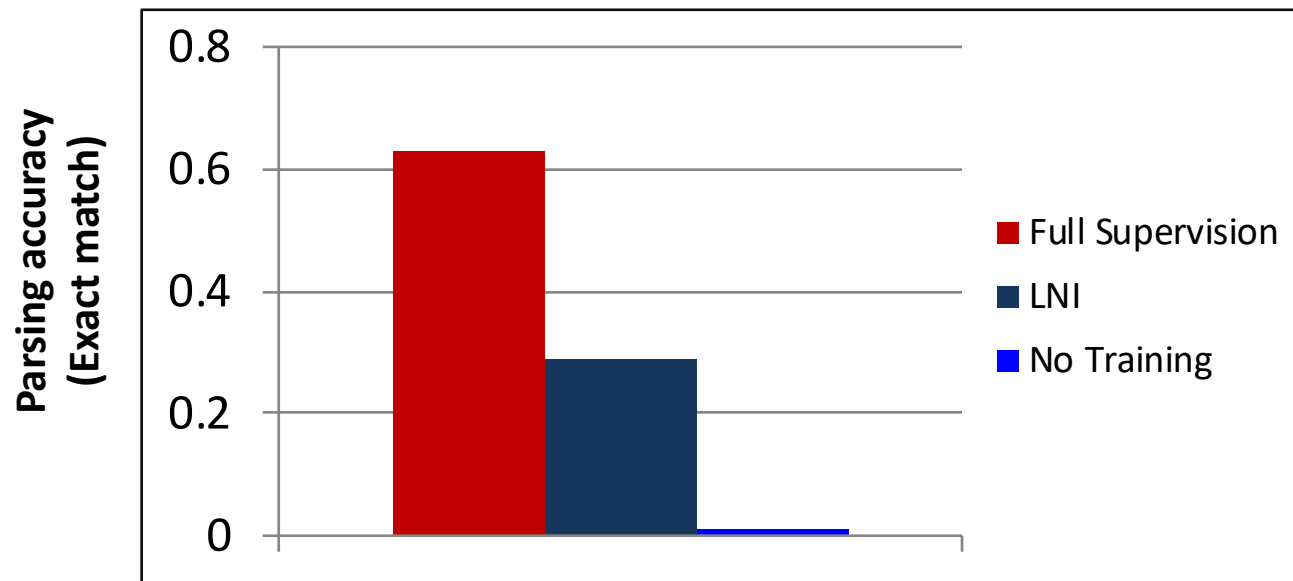
# Learning from fewer examples



- LNL consistently outperforms BoW, especially with fewer examples



# Results: Semantic Parsing



- Baseline (red): traditional supervised model trained on statements paired with logical forms

Predicted logical forms are often highly correlated

```
getPhraseMention( email, stringVal('meeting'))
```

```
getPhraseMention( body, stringVal('meeting'))
```

# Summary

- *NL explanations can define executable feature functions that improve concept learning performance*
- *Pragmatic context can guide learning of semantic parsers even with very weak supervision (class-labels only)*
- *Each domain requires specifying a DSL (one-time effort)*
  - *Reusable across long tail of categories*

*Part 2*: Incorporating model  
constraints from NL

# NL advice as defining model constraints

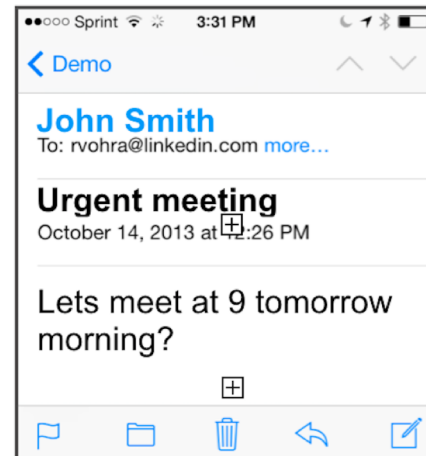


If the subject says 'urgent', it is almost **certainly** important.

**Most** emails from John are important.

Emails that I reply to are **usually** important.

Unimportant emails are **often** sent to a list



➔ Important email!

↓

sender:	John Smith
subject:	Urgent meeting ...
Fwd:	NO
Addressed to:	.....

- Potentially enable learning without labeled examples?
- Leverage quantifier expressions in language

# Sequential Approach

*Emails that I reply to are usually important*

Mapping language to quantitative constraints

Semantic Parser

$\mathbf{x} \rightarrow (\text{email.replied} == \text{true})$   
 $\mathbf{y} \rightarrow \text{important:true}$   
 $\mathbb{E}_{y|x}[\phi(x, y)] = b_{\text{usually}}$

Incorporating constraints in model training

Posterior Regularization

$\theta$

Classifier  
 $f : x \rightarrow y$

Unlabeled data



# Sequential Approach

*Emails that I reply to are usually important*

Mapping language to quantitative constraints

Semantic Parser

$\mathbf{x} \rightarrow (\text{email.replied} == \text{true})$   
 $\mathbf{y} \rightarrow \text{important:true}$   
 $\mathbb{E}_{y|x}[\phi(x, y)] = b_{\text{usually}}$

Posterior Regularization

$\theta$

Classifier  
 $f : x \rightarrow y$

Unlabeled data

# Training classifiers from declarative NL

- Explanations encode multiple properties that can aid statistical learning

*'Emails that I reply to are usually important'*

1. Features important for a learning problem
  - ✓  $\mathbf{x}$  : repliedTo:true
2. Class labels
  - ✓  $y$  : Important
3. Type of Relationship b/w features and labels
  - ✓  $P(y | \mathbf{x})$
4. Strength of Relationship
  - ✓ Specified by quantifier?

# Semantic parsing

## ➤ Constraint types:

- i. *About a third of the emails that I get are important* :  $P(y)$
- ii. *Emails that I reply to are usually important* :  $P(y | x)$
- iii. *I almost always reply to important emails* :  $P(x | y)$

## ➤ Novelty largely in identifying the type of the assertion

### ➤ Primarily depends on syntactic features

- ✓ Features based on dependency paths
- ✓ Presence/absence of negation words
- ✓ Identifying active/passive voice
- ✓ Order of occurrence of triggers for x and y

*'Emails that I reply to are usually important'*

$P(\text{important} | \text{replied:true}) \approx p_{\text{usually}}$

# Semantic parsing

- Leverage semantics of linguistic quantifiers
  - Associate point probability estimates for frequency adverbs and determiners

Frequency quantifier	Probability value
always , certainly , definitely , all	0.95
usually , normally , generally , likely	0.70
most , majority	0.60
often , half	0.50
many	0.40
sometimes , frequently , some	0.30
few , occasionally	0.20
rarely , seldom	0.10
never	0.05

- Purely subjective beliefs, not calibrated on any data

# Sequential Approach

*Emails that I reply to are usually  
important*

Semantic  
Parser

$\mathbf{x} \rightarrow (\text{email.replied} == \text{true})$   
 $\mathbf{y} \rightarrow \text{important:true}$   
 $\mathbb{E}_{y|x}[\phi(x, y)] = b_{\text{usually}}$

Posterior Regularization

Incorporating constraints  
in model training

$\theta$

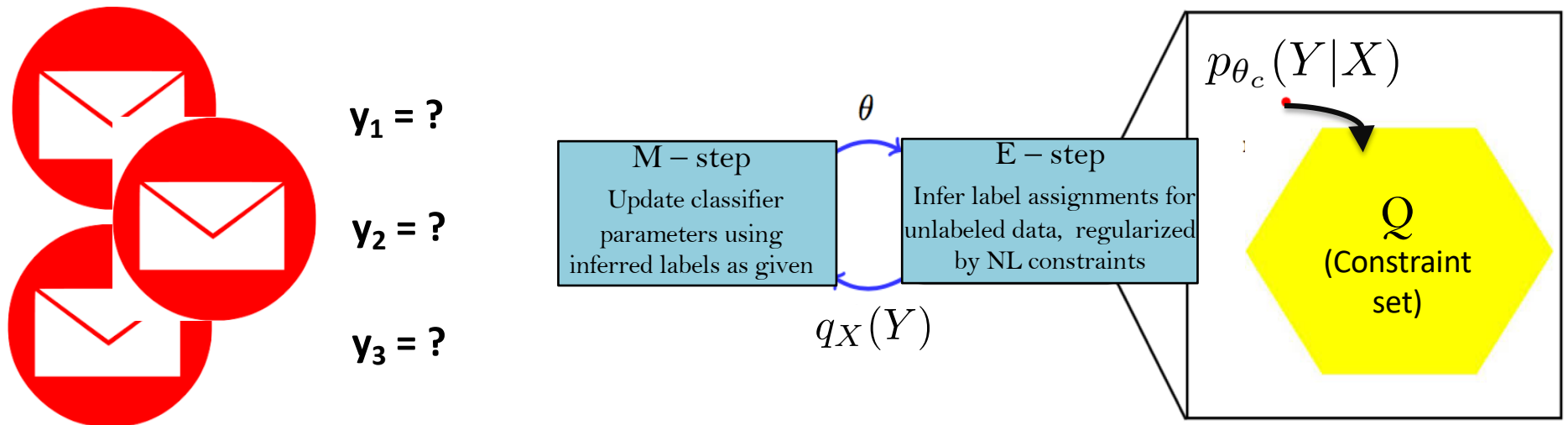
Classifier  
 $f : x \rightarrow y$



Unlabeled data

# Posterior Regularization

- Use the posterior regularization (PR) principle to imbue human-provided advice in learned models
  - Unobserved class labels as latent variables
- PR optimizes a latent variable model subject to a set of constraints on the posterior distribution  $p_{\theta}(\mathbf{y} \mid \mathbf{x})$



# Probability Assertions as PR Constraints

- PR can handle linear constraints over distributions of latent variables

$$Q := \{q_{\mathbf{x}}(\mathbf{y}) : \mathbb{E}_q[\phi(\mathbf{x}, \mathbf{y})] \leq \mathbf{b}\}$$

Linear bounds on expected values of features under  $q$

- Can convert each constraint type to this form:

Type	Example	
$P(y x)$	<i>Emails that I reply to are usually important</i>	$\mathbb{E}[\mathbb{I}_{y=important,reply(x):true}] - p_{usually} \times \mathbb{E}[\mathbb{I}_{reply(x):true}] = 0$
$P(x y)$	<i>I almost always reply to important emails</i>	$\mathbb{E}[\mathbb{I}_{y=important,reply(x):true}] - p_{always} \times \mathbb{E}[\mathbb{I}_{y=important}] = 0$
$P(y)$	<i>About a third of all emails I get are important</i>	Same as $P(y   x)$ , when $x$ is a constant feature

# Posterior Regularization

- Each constraint from the semantic parser can be expressed in the form compatible with PR
  - Conjunction of all such constraints specifies  $Q$
- Train with modified EM to maximize PR objective:

$$J_Q(\theta) = \mathcal{L}(\theta) - \min_{q \in Q} KL(q | p_\theta(Y|X))$$

Improve data likelihood

Emulate human advice



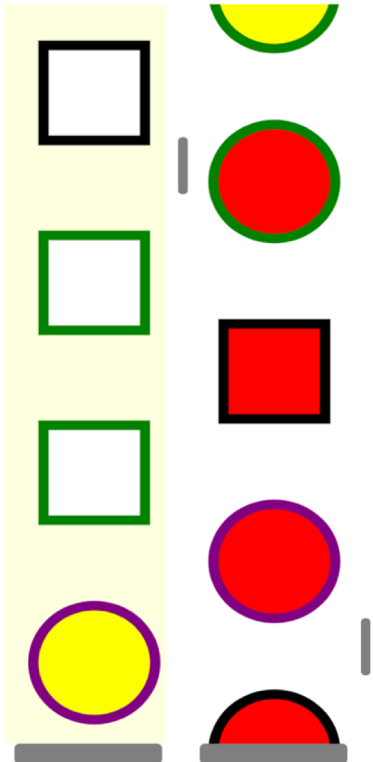
# Synthetic shape classification

- Turkers observe samples of shapes from synthetically generated datasets, and describe them through statements.

**SELECTED SHAPES**   **OTHER SHAPES**

scroll to see more

scroll to see more



**DO NOT PRESS THE BACK BUTTON, THIS WILL CAUSE THE HIT TO BREAK**

**READ FIRST! (read carefully)**

Please describe the shapes in the **SELECTED** column, in a way that can help other people identify these shapes.

Each sentence should focus on **ONE FEATURE** at a time.  
For example, only focusing on shape, fill or border color.

Please **SELECT FEATURE** in the dropdown box which you are describing in your sentence.

**DO NOT** combine multiple features into a single sentence

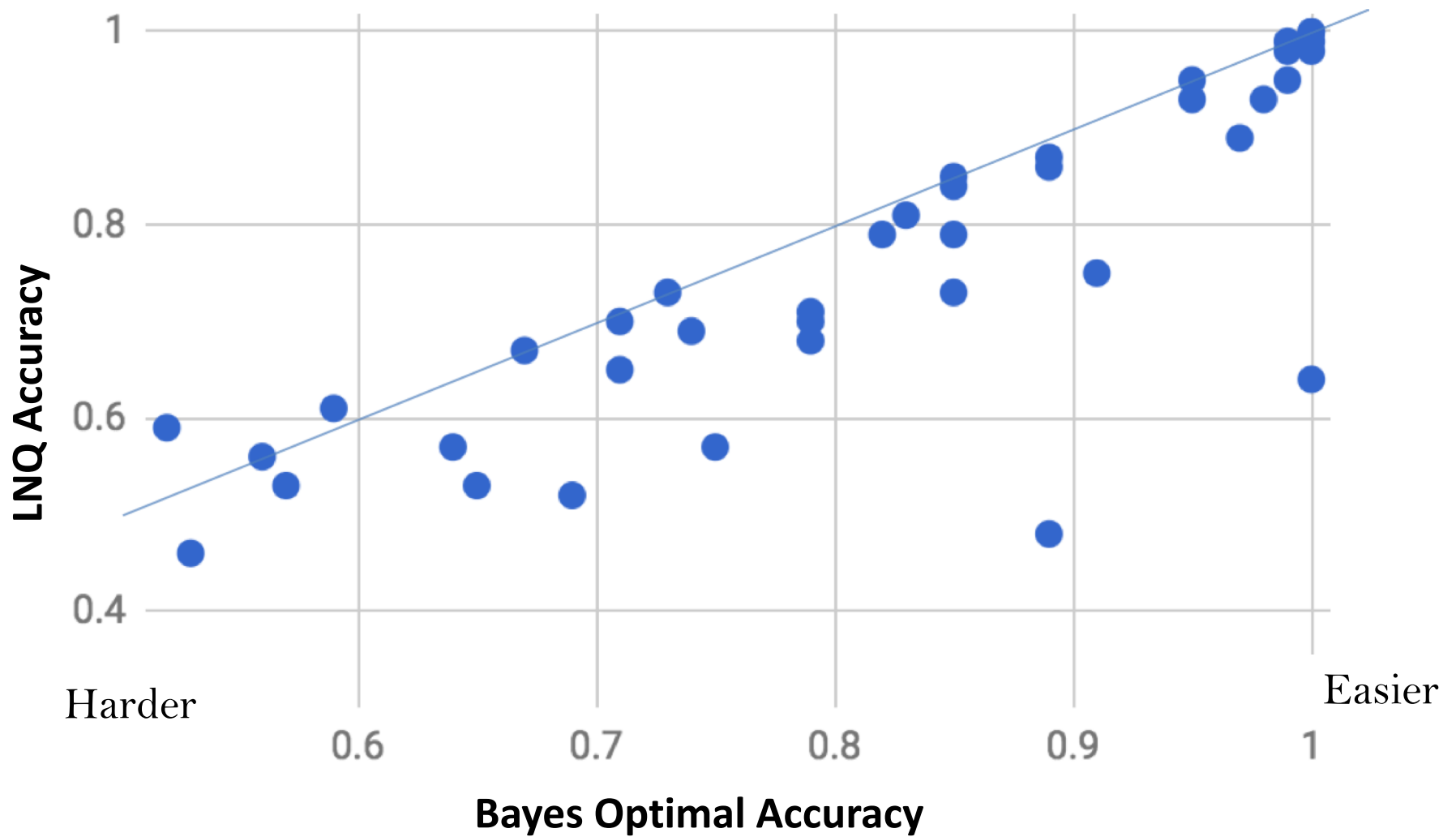
Add another statement

1	Shape	selected shapes are almost always a square
2	Border color	other shapes rarely have a blue border
3	Fill color	if the shape has a red fill color, it's most likely not a selected shape

Finished!

- ✓ 50 datasets
- ✓ 30 workers
- ✓ 4.3 statements per task on average

1. Selected shapes are almost always a square
2. Other shapes rarely have a blue border
3. If a shape has a red fill, it is most likely not a selected shape ...



Each dot represents a dataset (and corresponding classification task) generated from a known distribution

# Average Classification Accuracy (Shapes data)

Approach	Avg Accuracy	Access to labels	Access to statements
LNQ	0.751	no	yes
Bayes Optimal	0.831	--	--
Logistic Regression	0.737	yes	no
Random	0.524	--	--

# Average Classification Accuracy (Shapes data)

Approach	Avg Accuracy	Access to labels	Access to statements
LNQ	0.751	no	yes
Bayes Optimal	0.831	--	--
Logistic Regression	0.737	yes	no
Random	0.524	--	--
LNQ (no quantification)	0.545	no	yes
LNQ (coarse quantification)	0.679	no	yes

# Average Classification Accuracy (Shapes data)

Approach	Avg Accuracy	Access to labels	Access to statements
LNQ	0.751	no	yes
Bayes Optimal	0.831	--	--
Logistic Regression	0.737	yes	no
Random	0.524	--	--
LNQ (no quantification)	0.545	no	yes
LNQ (coarse quantification)	0.679	no	yes
Human teacher	0.802	yes	yes (writes descriptions)
Human learner	0.734	no	yes

# Real classification tasks

## SELECTED BIRDS

## OTHER BIRDS

scroll to see more

scroll to see more



## READ FIRST! (read carefully)

Please describe the birds in the **SELECTED** column, in a way that can help other people identify these shapes.

Each sentence should focus on **ONE FEATURE** at a time. For example, only focusing on **crown color**, **primary color** or **wing pattern**

Please **SELECT FEATURE** in the dropdown box which you are describing in your sentence and use the table below to help you identify names for these features

**DO NOT** combine multiple features into a single sentence

- **Bill shape**  
curved, dagger, hooked, hooked (seabird), all-purpose, cone
- **Size**  
very large, large, medium, small, very small
- **Shape**  
long-legged-like | duck-like | gull-like | hummingbird-like | pigeon-like | tree-clinging-like | hawk-like | sandpiper-like | swallow-like | perching-like
- **Tail pattern**  
solid | spotted | striped | multi-colored
- **Primary color**  
blue | brown | grey | yellow | olive | green | black | white | red | buff
- **Crown color**  
blue | brown | grey | yellow | olive | green | black | white | red | buff
- **Wing pattern**  
solid, spotted, striped, multi-colored

Add another statement

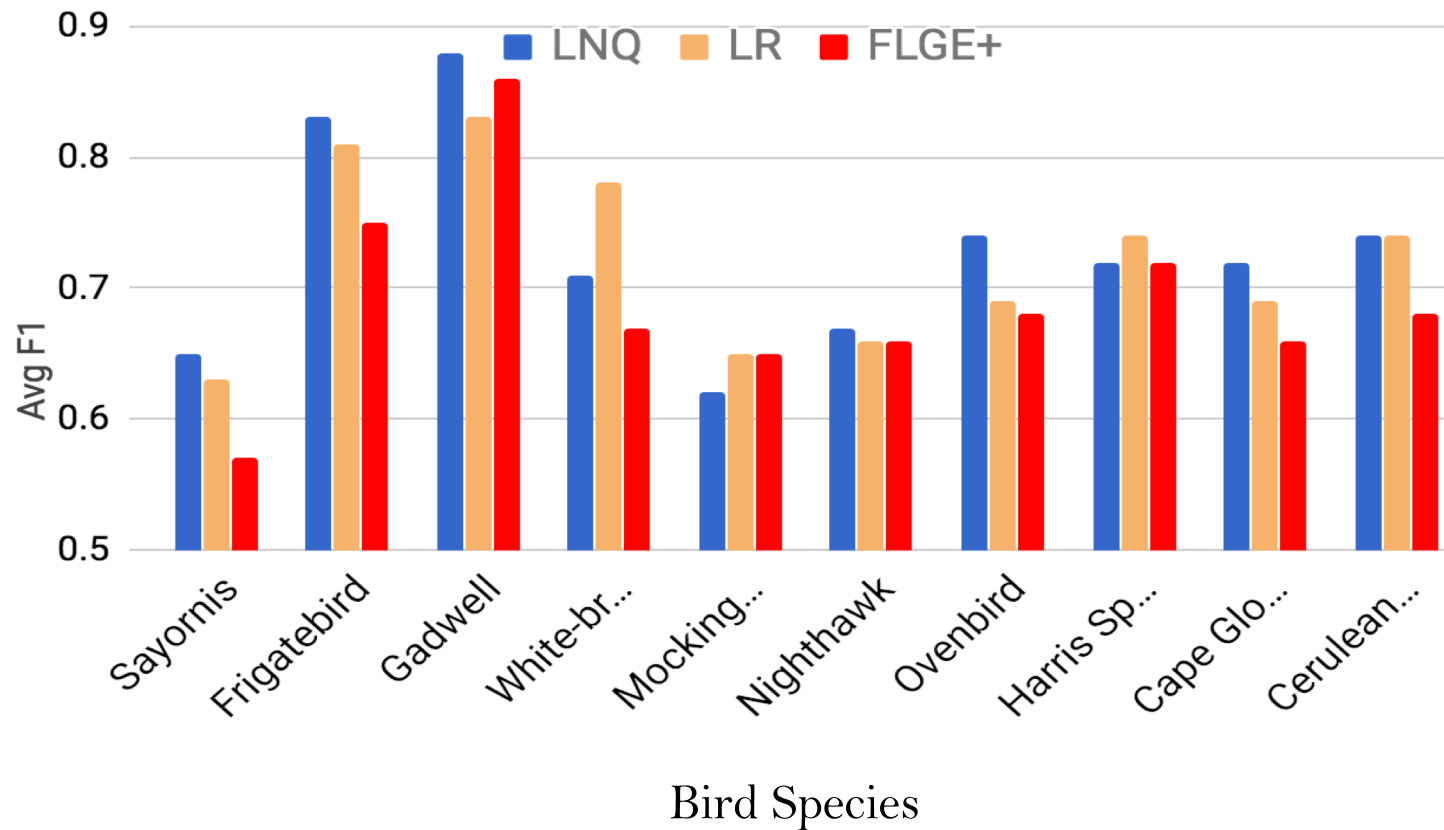
- 1 Primary color all selected birds have a brown primary color
- 2 - what feature? -
- 3 - what feature? -
- 4 - what feature? -

- ✓ 10 species from CUB-200 dataset
- ✓ 60 examples per species
- ✓ 53 pre-specified attributes
- ✓ 6.1 statements per task on average

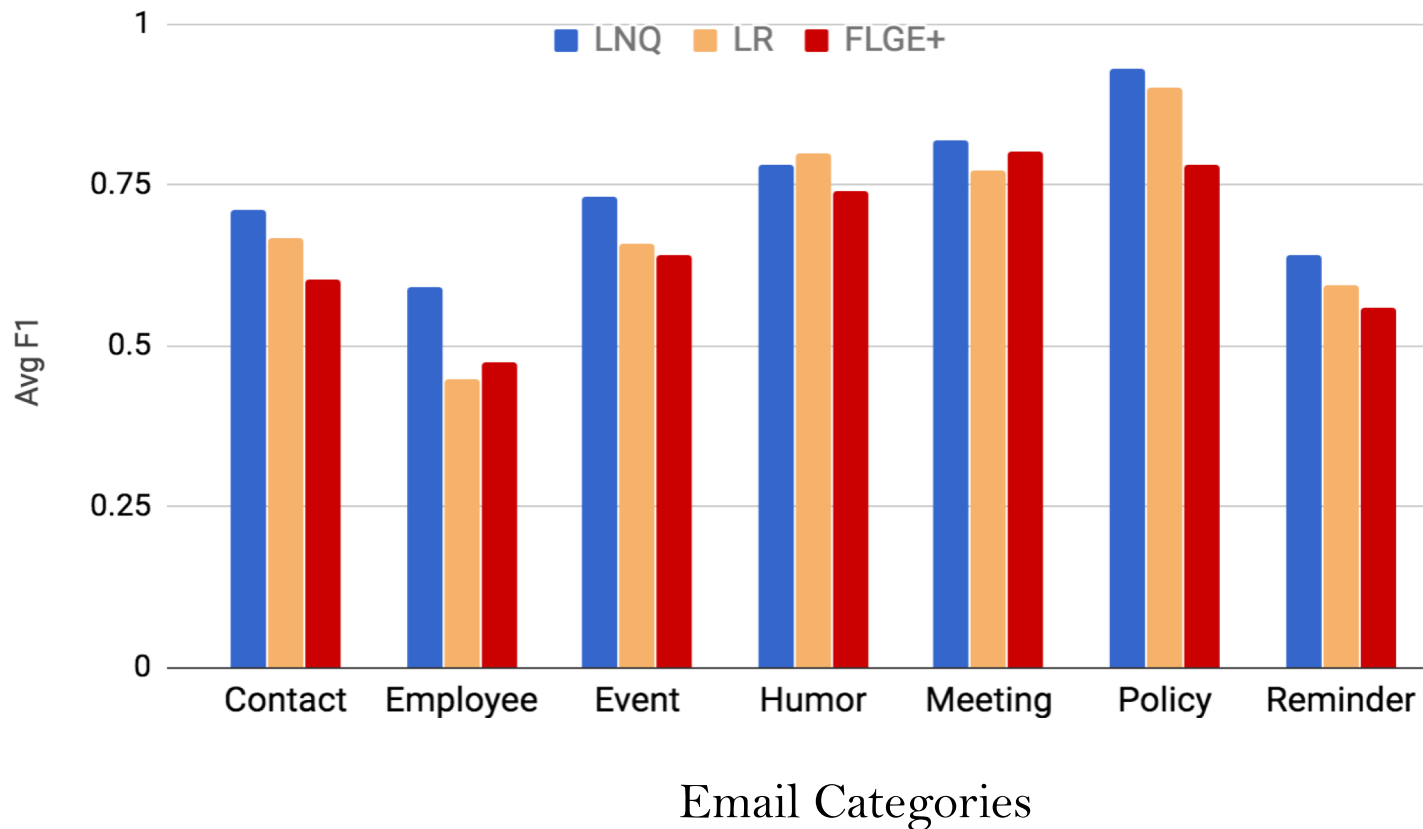
## Example statements:

- A specimen that has a striped crown is likely to be a selected bird
- Birds in the other category rarely ever have dagger-shaped beaks

# Results: Bird Species Identification



# Results: Emails Categorization

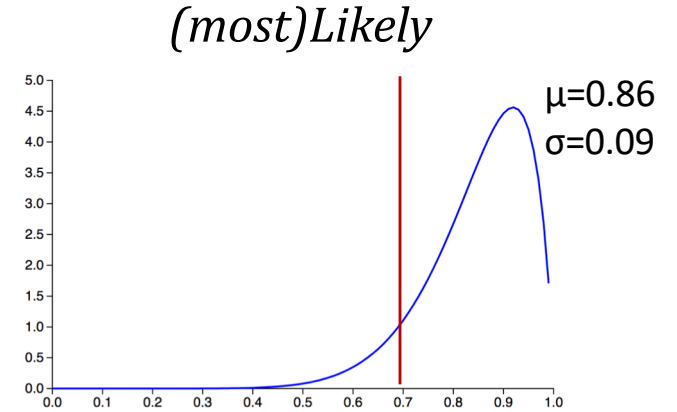
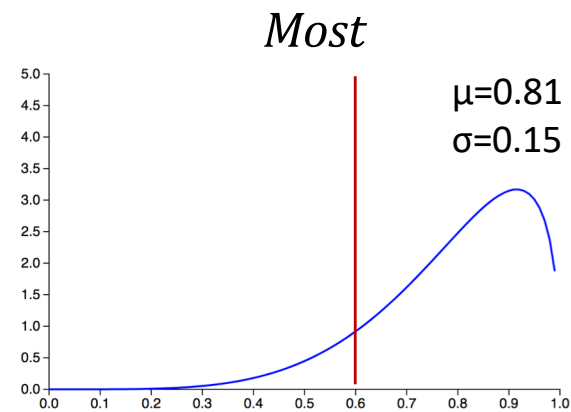
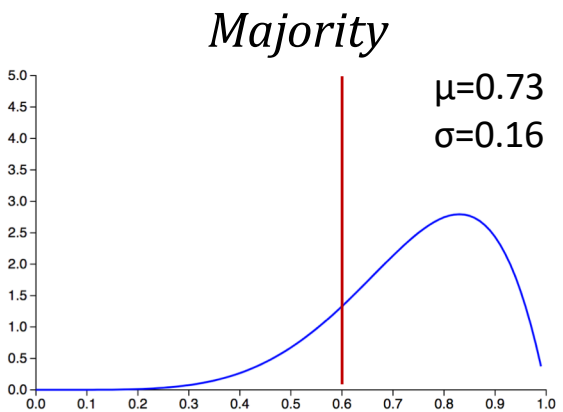
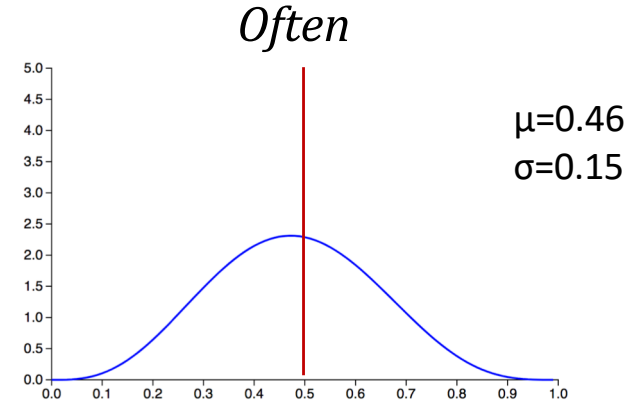
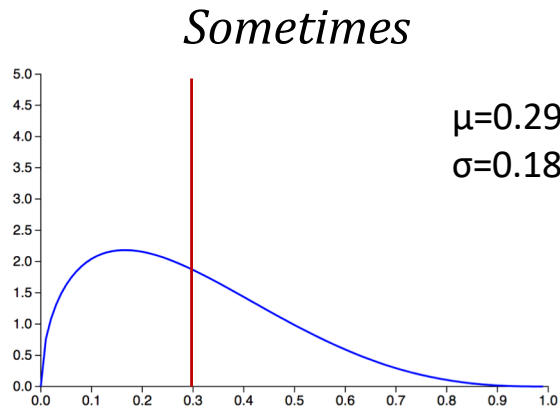
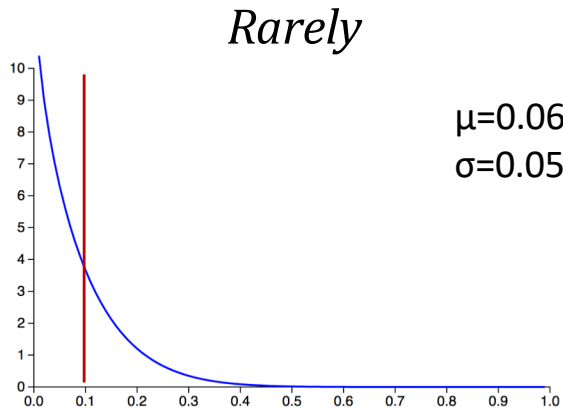


Performance by training from both quantification and labels

- About a third of statements used quantifiers



# Empirical distributions of probability values



# Summary

- *Declarative NL can supervise learning in limited data settings*
- *Differential associative strengths of linguistic quantifiers can be effective towards zero-shot concept learning*
- *Possible to learn through a blend of strategies*

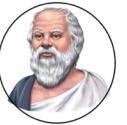
# Other directions

- Learning with mixed initiative dialog
  - Allow the learner to ask questions?



Can you label this email as important/not important?

Sure. This is actually an important email



Thanks. Can you give me an explanation of the concept?

Emails from CMU are usually important



- Learning from multiple teachers
  - How to learn from contradictory advice?



- Pairing explanations with demonstrations, curricular learning,...

Questions?