MACHINE LEARNING
DEPARTMENT

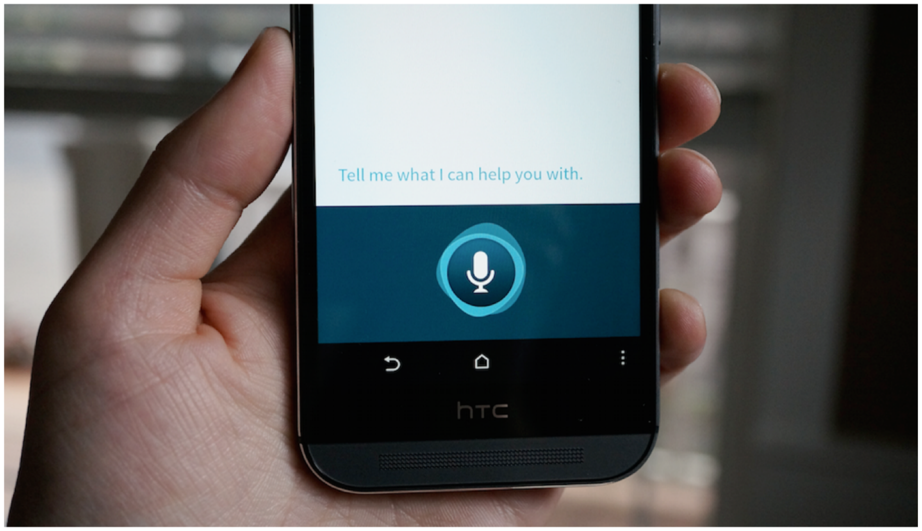
Learning to Classify from Natural Language Explanations

Shashank Srivastava

Joint work with Igor Labutov, Tom Mitchell

CARNEGIE MELLON UNIVERSITY
PITTSBURGH PENNSYLVANIA 1900

Is this email important?



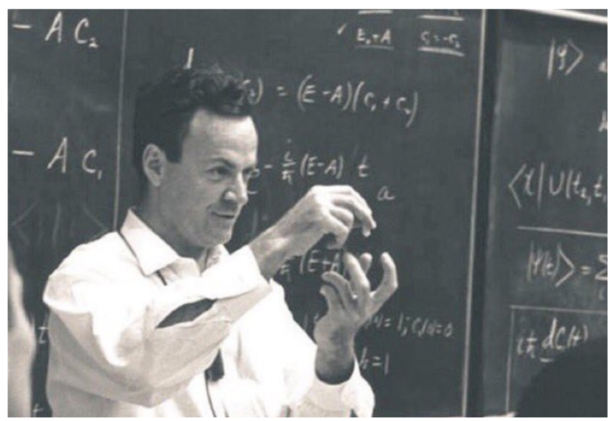
- *'Emails from my boss are usually important'*
- *'Such emails mention a deadline or a meeting'*
- *'The subject might say urgent ...'*

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

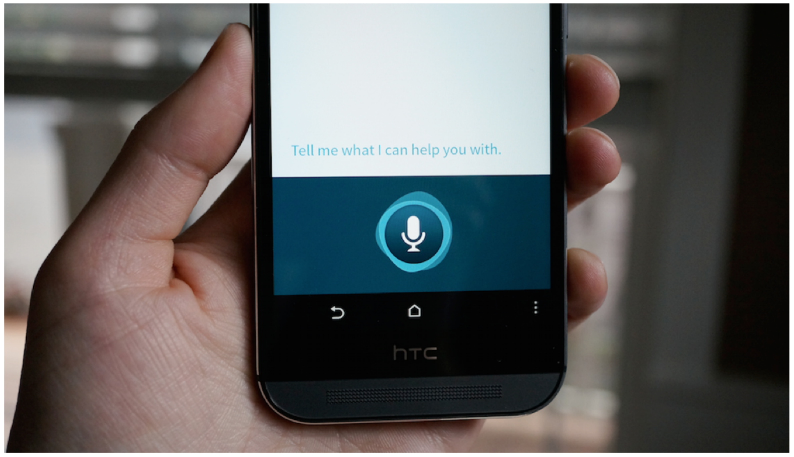
ACL 2018

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)	Logical form (l)	**Evaluate** in a context ($z = [l]_x$)
'three less than twenty times six'	minus(prod(20, 6), 3)	117
'What is the longest river that flows through Pittsburgh?'	argmax(river(x) $\wedge$ traverse(x,y) $\wedge$ const(y, Pittsburgh), length)	Ohio
'Phishing emails often mention prices'	findSemanticCategory(MONEY, field:body)	Yes/No

How to interpret explanations?

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

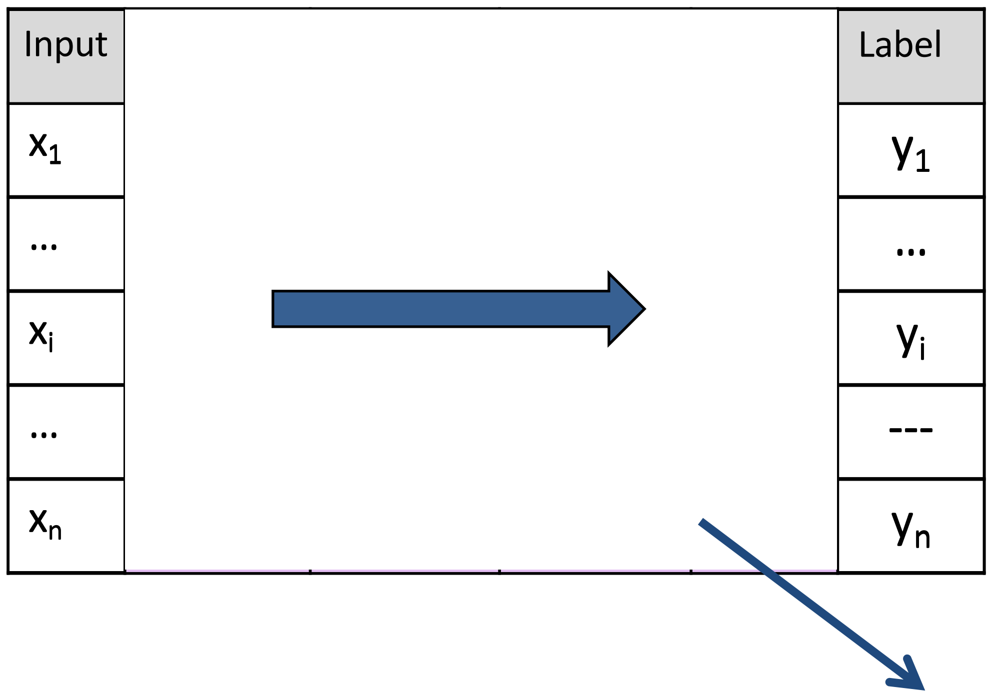
NL Explanation: *'Phishing emails often mention prices'*

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

Jointly learn a classifier and a semantic parser!

Don't need annotated logical forms

Problem setting



Latent variables

$s_i \rightarrow l_i$ (parsing)

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

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

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)

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

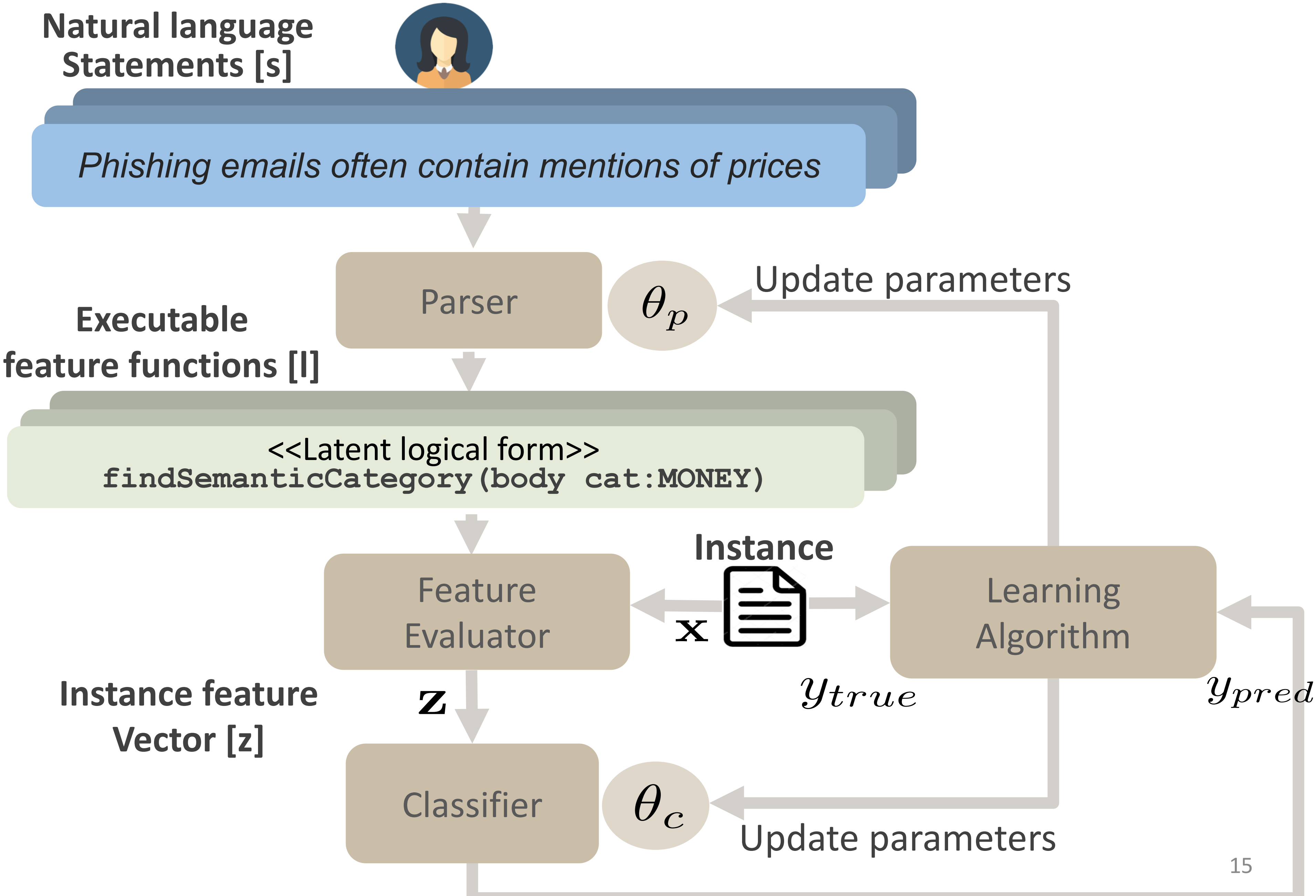
Prefer values that are discriminative

Prefer interpretations supportec 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

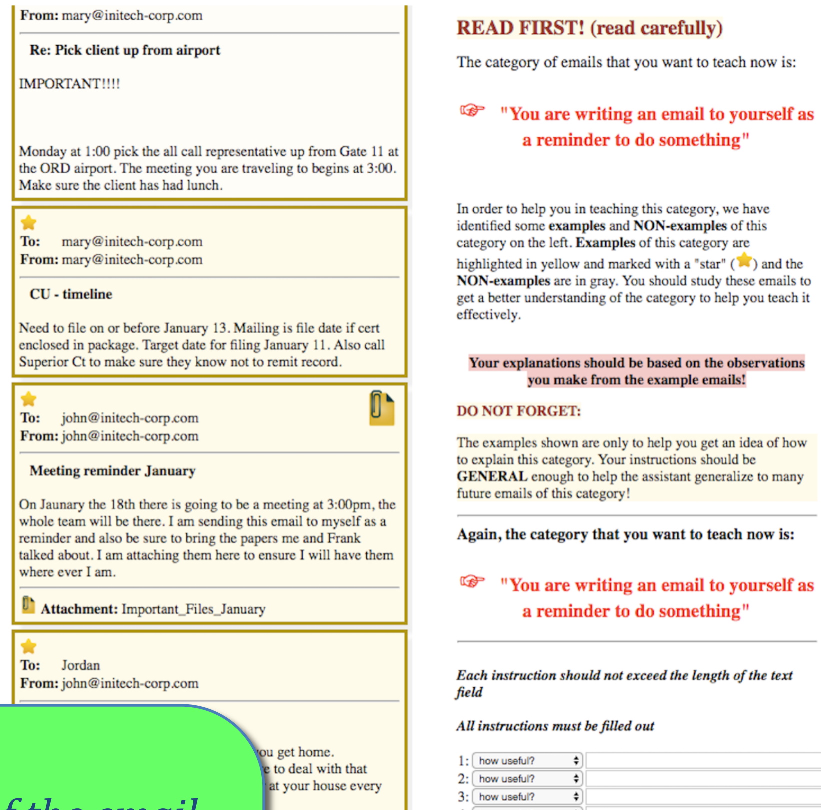
- Emails representing common email categories through AMT
  - Reminders, meeting invitations, requests from boss, internet humor, going out with friends, policy announcements, etc.
  - 1100 emails, 7 types

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

Subject: Note to self - Move the Bodies
From: john@initech-corp.com
To: john@initech-corp.com
Body: 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.
Attachment: none

Data: NL Explanations

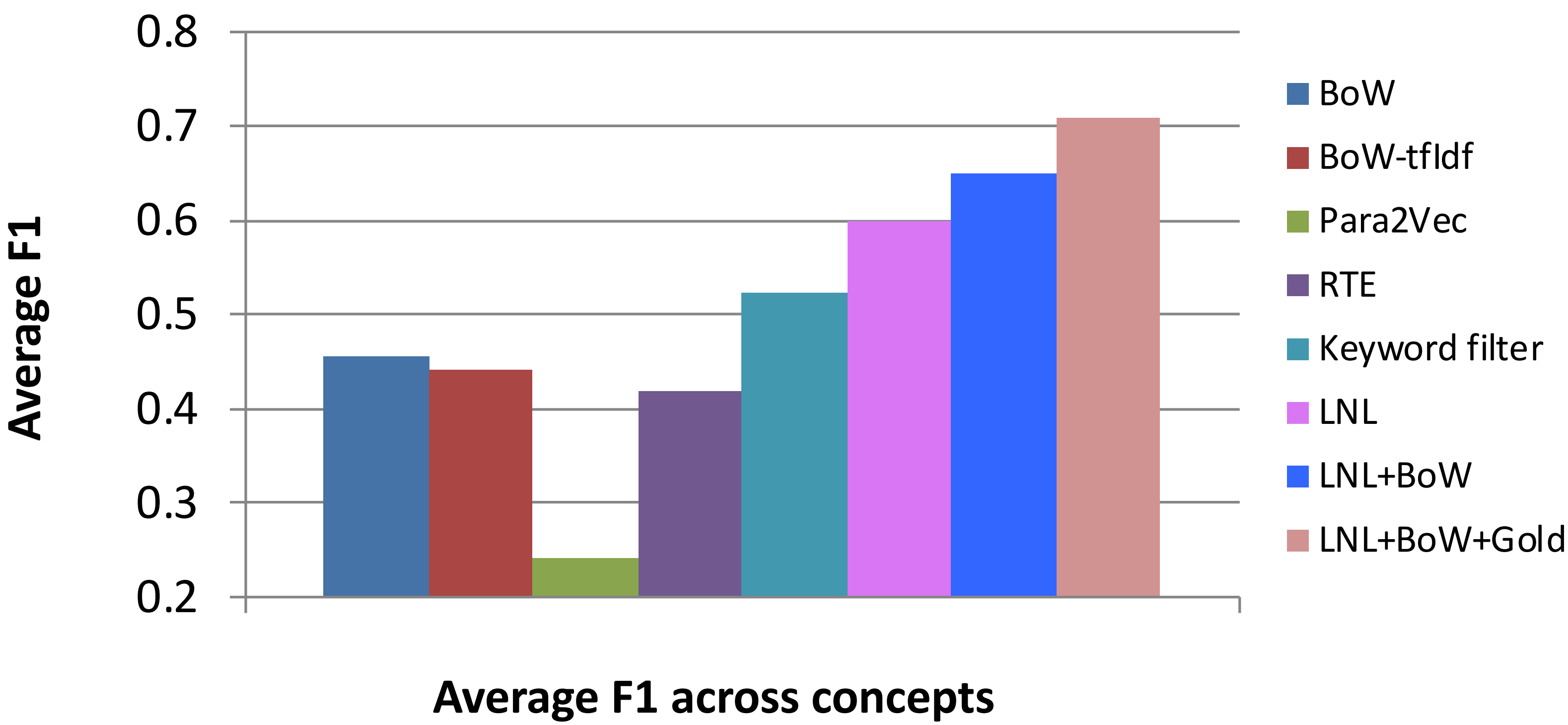
- Dataset of statements explaining each concept
- Turkers describe emails from each category
- 30 statements for each category



Sample explanations:

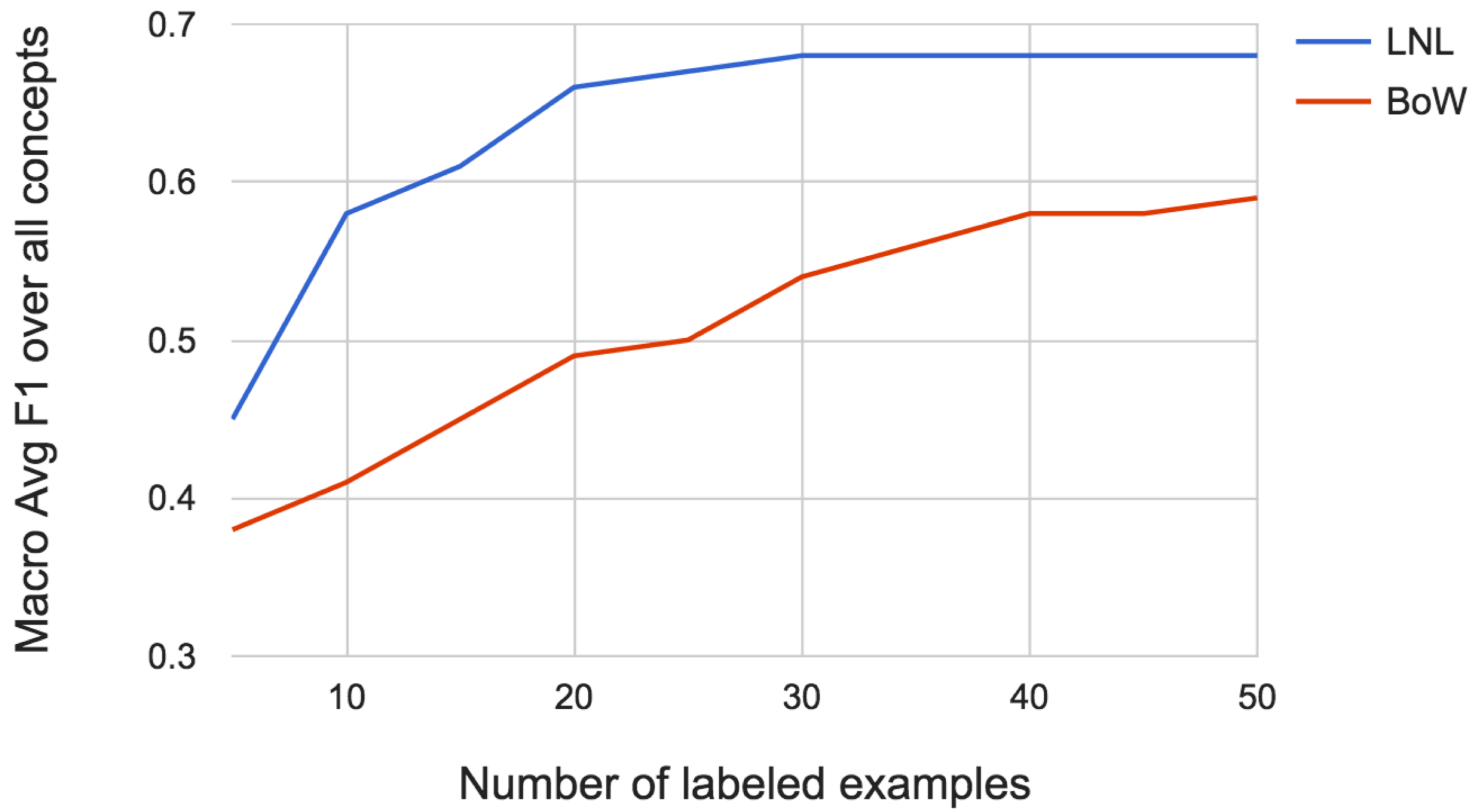
Most reminders mention a date and a time in the message of the email
The sender of the email is the same as the recipient
These emails usually close with a name or title
These emails sometimes have jpg attachments
The email likely has words like "policy" or "announcement" in the subject
Emails from a public domain are not office requests

Results: Email classification



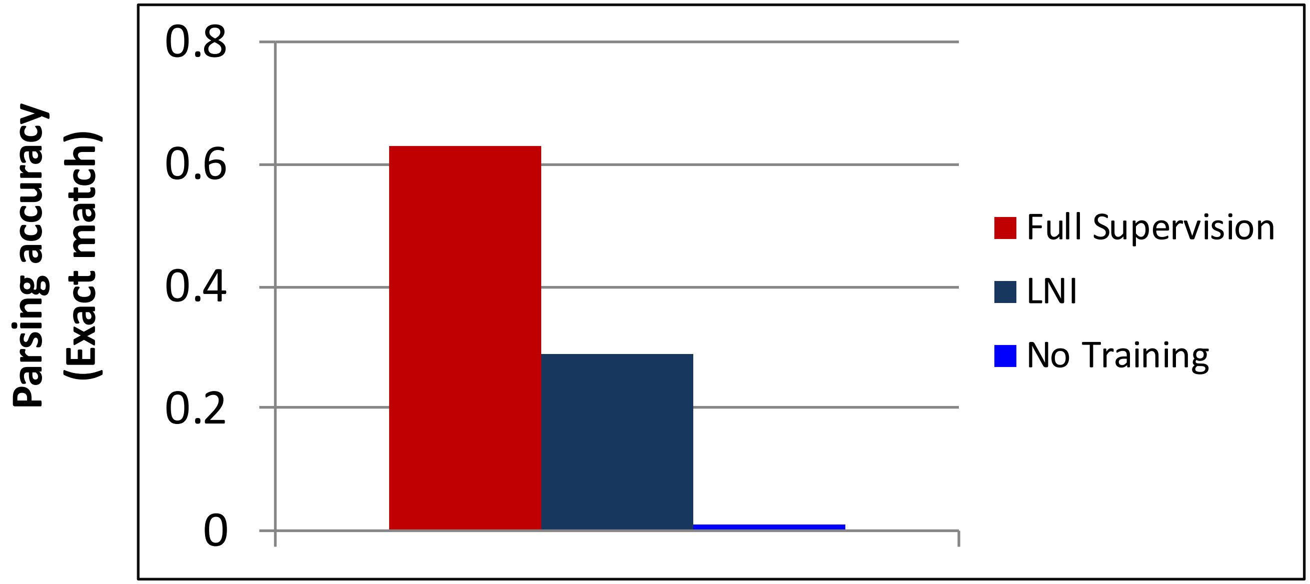
- Significantly better than best baseline for 6 of 7 categories

Learning from fewer examples



➢ LNL consistently outperforms BoW, especially with fewer examples

Results: Semantic Parsing



- <u>Baseline (red)</u>: 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

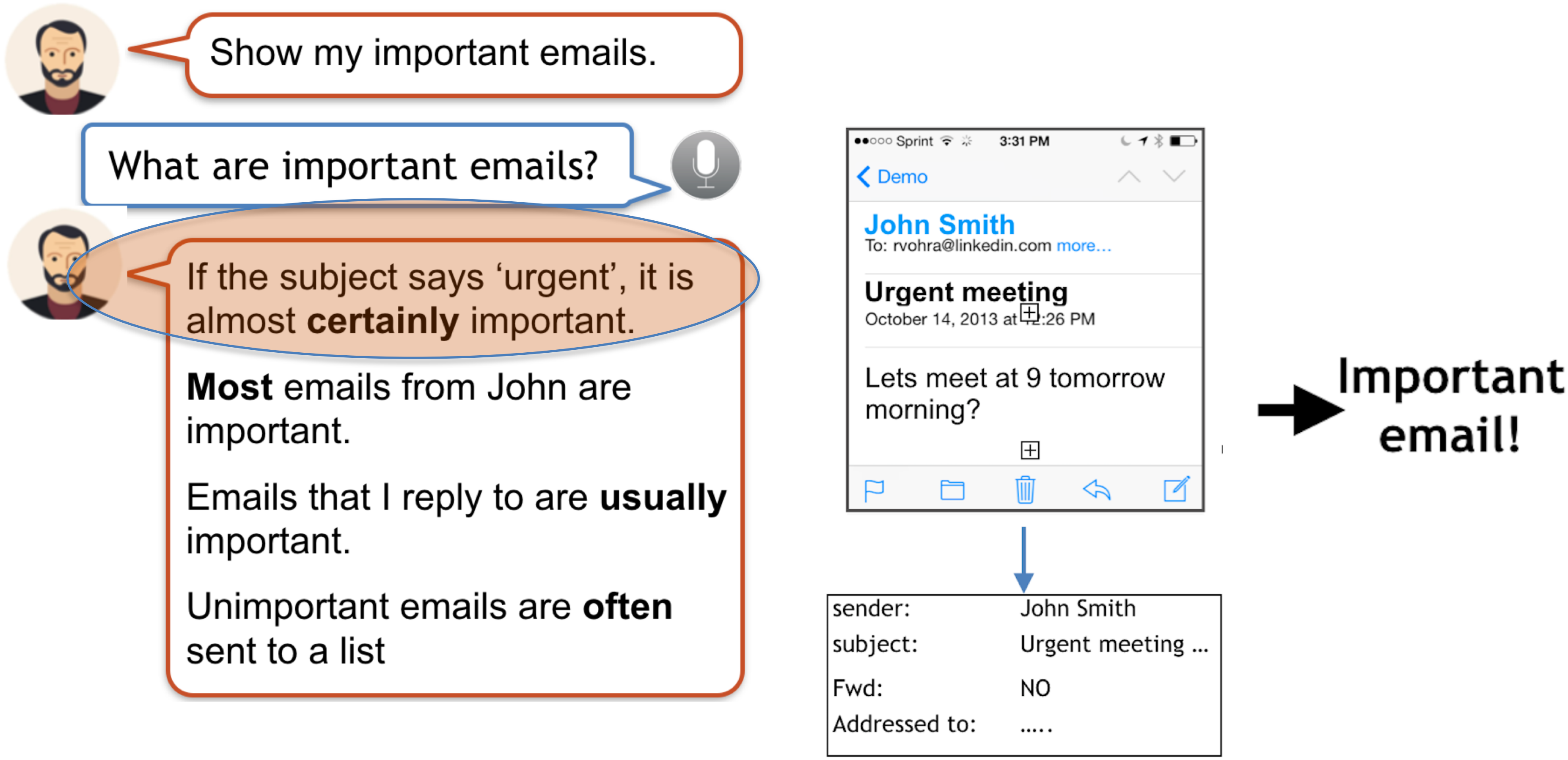
- *NL explanations can define executable feature functions that improve concept learning performance*

- *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

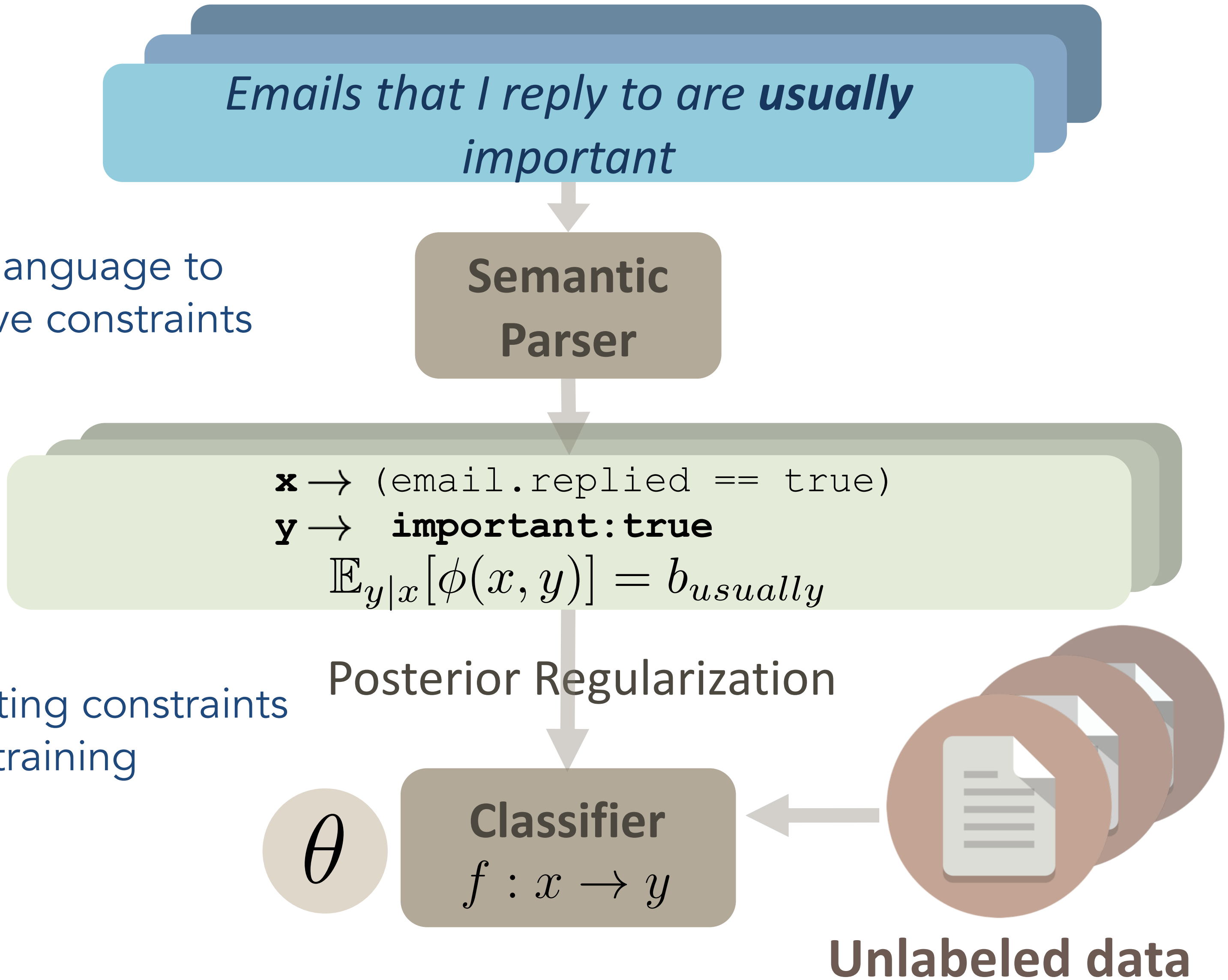


- ➢ Potentially enable learning without labeled examples?
- ➢ Leverage quantifier expressions in language

Sequential Approach

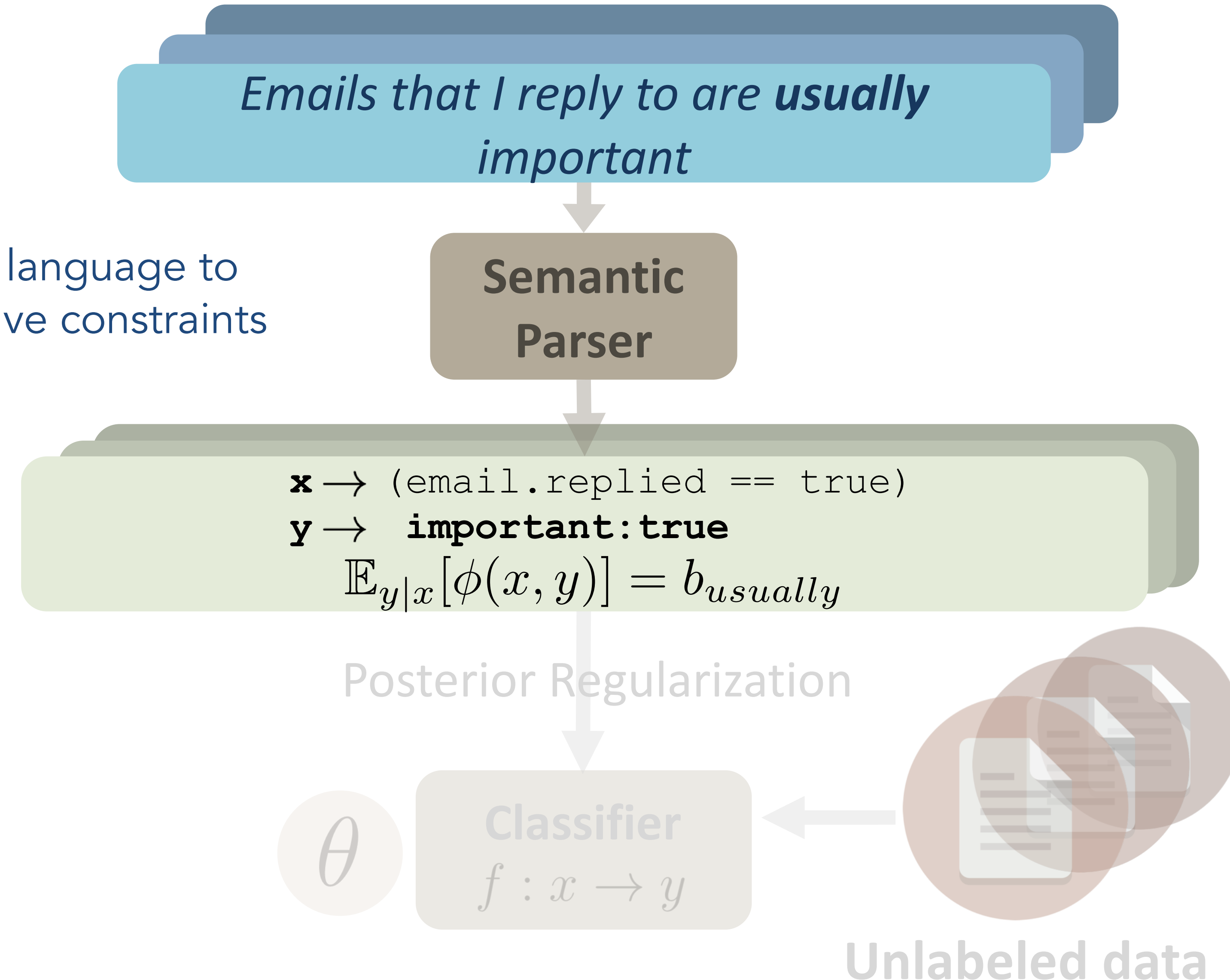
Mapping language to quantitative constraints

Incorporating constraints in model training



Sequential Approach

Mapping language to quantitative constraints



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
   - ✓ **x** : repliedTo:true
2. Class labels
   - ✓ **y** : Important
3. Type of Relationship b/w features and labels
   - ✓ $P(y|x)$
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 :* $P(y \mid x)$
  - iii. *I almost always reply to important emails :* $P(x \mid y)$

- 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'

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

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

$\mathbf{x} \longrightarrow$ `(email.replied == true)`

$\mathbf{y} \longrightarrow$ **`important:true`**

$$\mathbb{E}_{y|x}[\phi(x,y)] = b_{usually}$$

Posterior Regularization

Incorporating constraints in model training

θ

Classifier

$f : x \rightarrow y$

Unlabeled data

Posterior Regularization

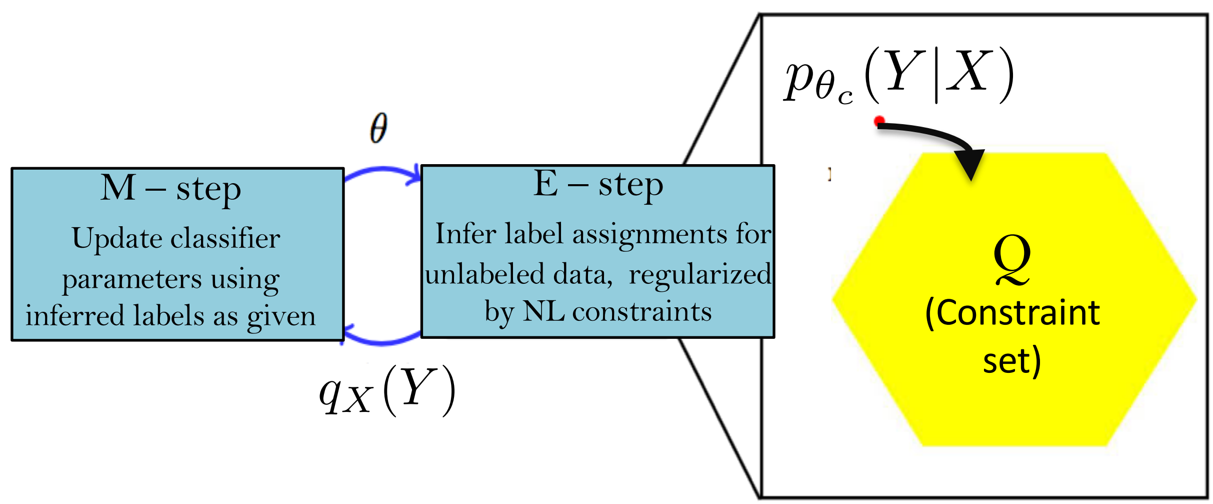
- Use the posterior regularization (PR) principle to imbue human-provided advice in learned models
  - Unobserved class labels as latent variables

- PR optimizes a latent variable model subject to a set of constraints on the posterior distribution $p_\theta(\mathbf{y} \mid \mathbf{x})$

y_1 = ?

y_2 = ?

y_3 = ?



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

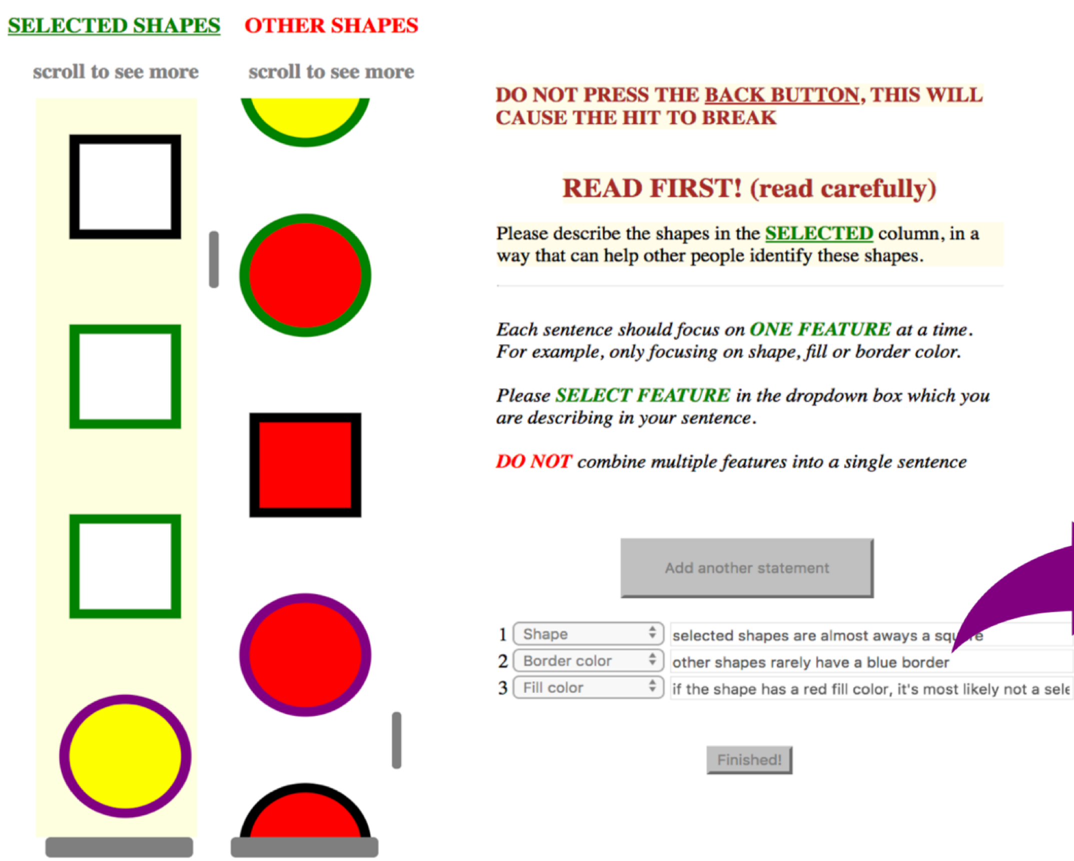
Posterior Regularization

- 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:

$$J_Q(\theta) = \underbrace{\mathcal{L}(\theta)}_{\text{Improve data likelihood}} - \underbrace{\min_{q \in Q} KL(q \mid p_\theta(Y|X))}_{\text{Emulate human advice}}$$

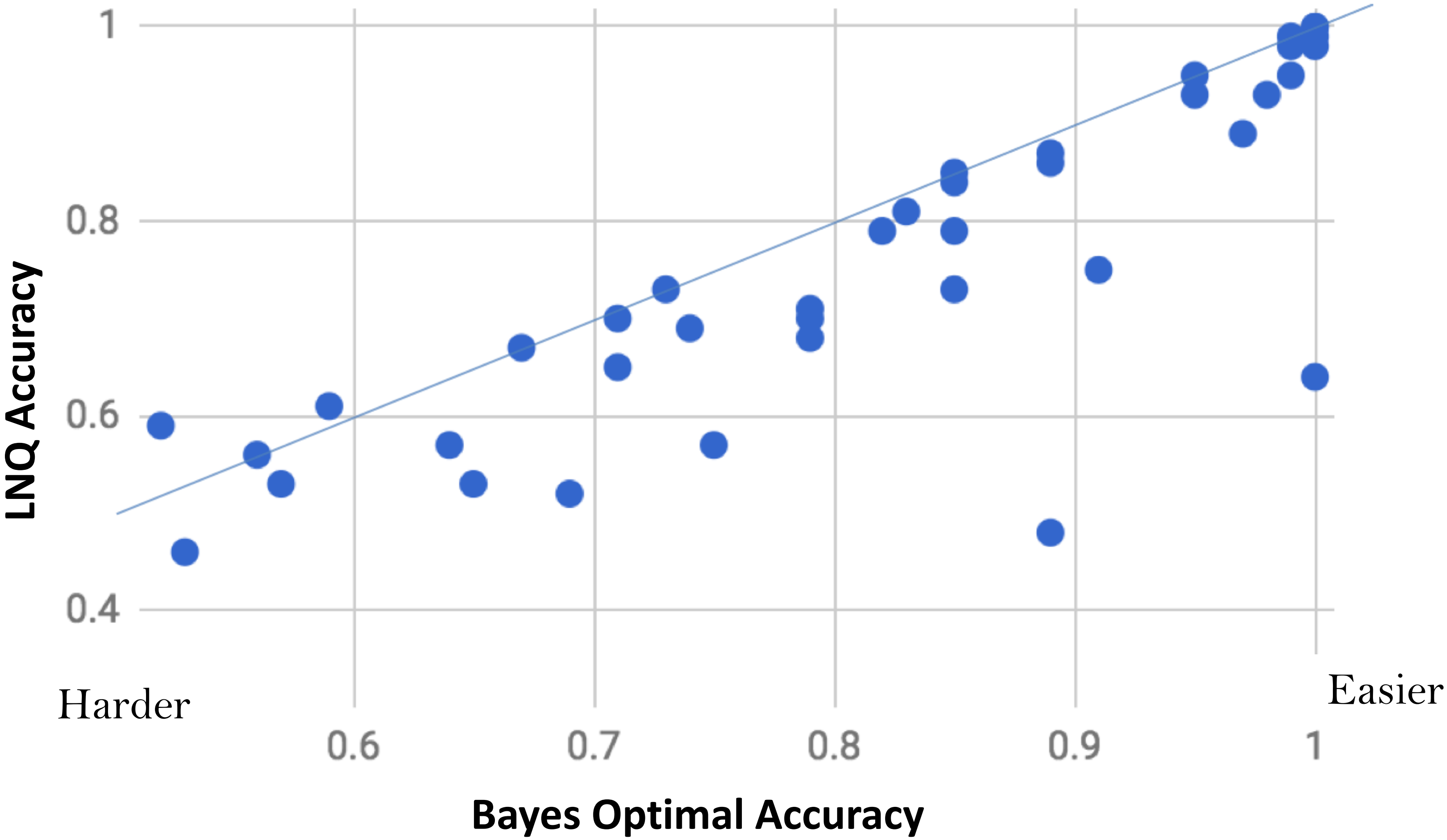
Synthetic shape classification

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



- ✓ 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	--	--

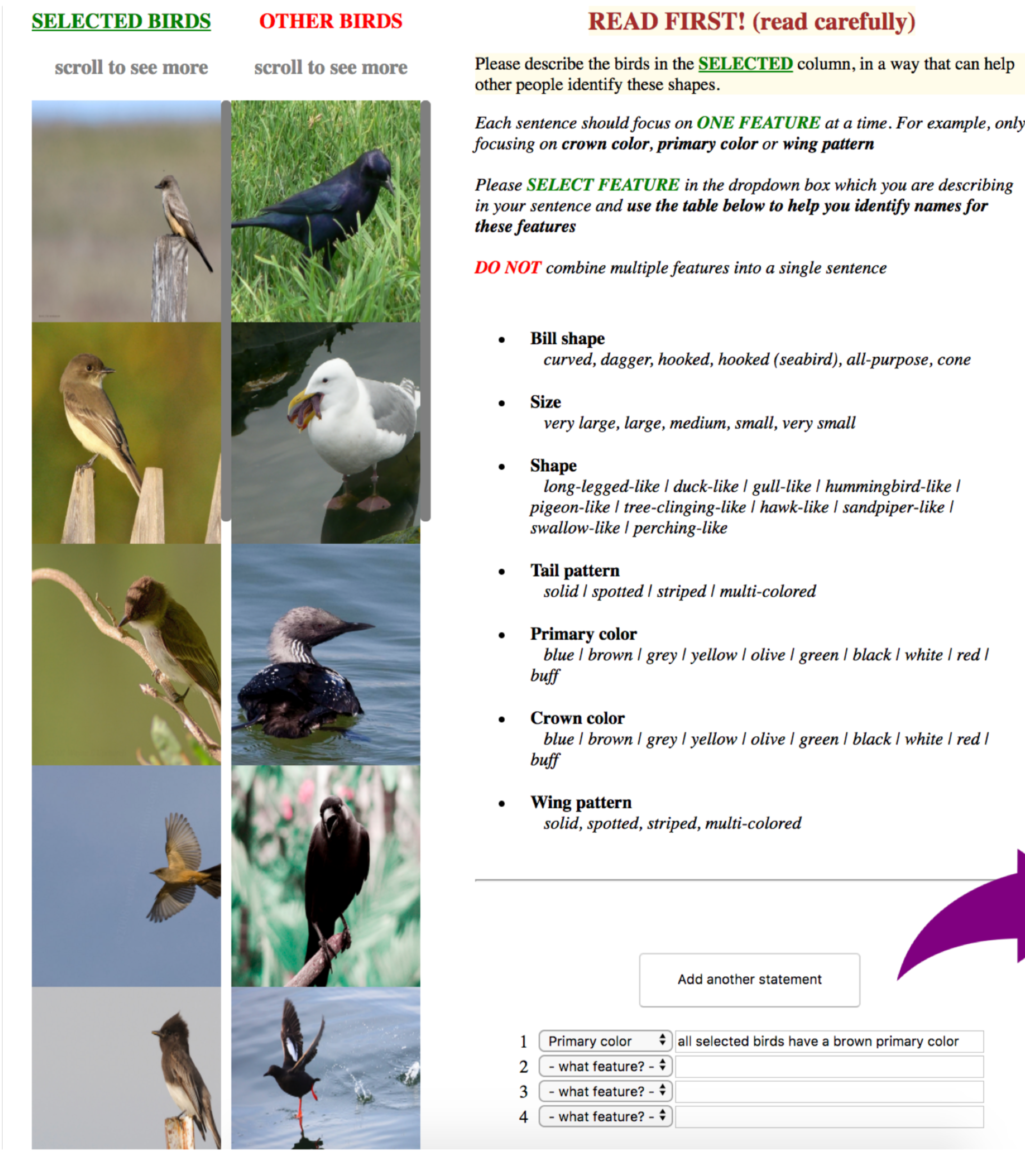
Average Classification Accuracy (Shapes data)

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

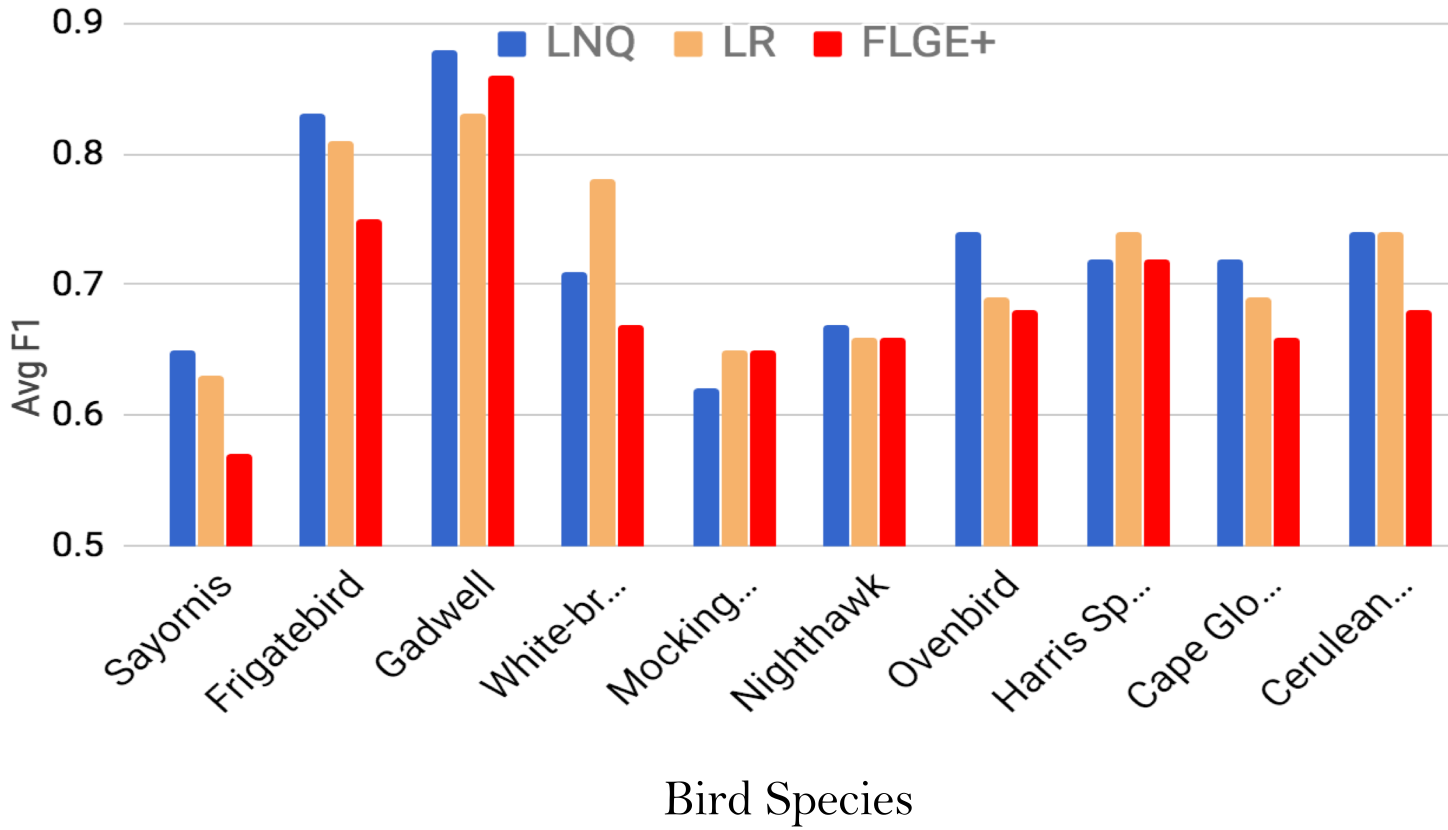


- ✓ 10 species from CUB-200 dataset
- ✓ 60 examples per species
- ✓ 53 pre-specified attributes
- ✓ 6.1 statements per task on average

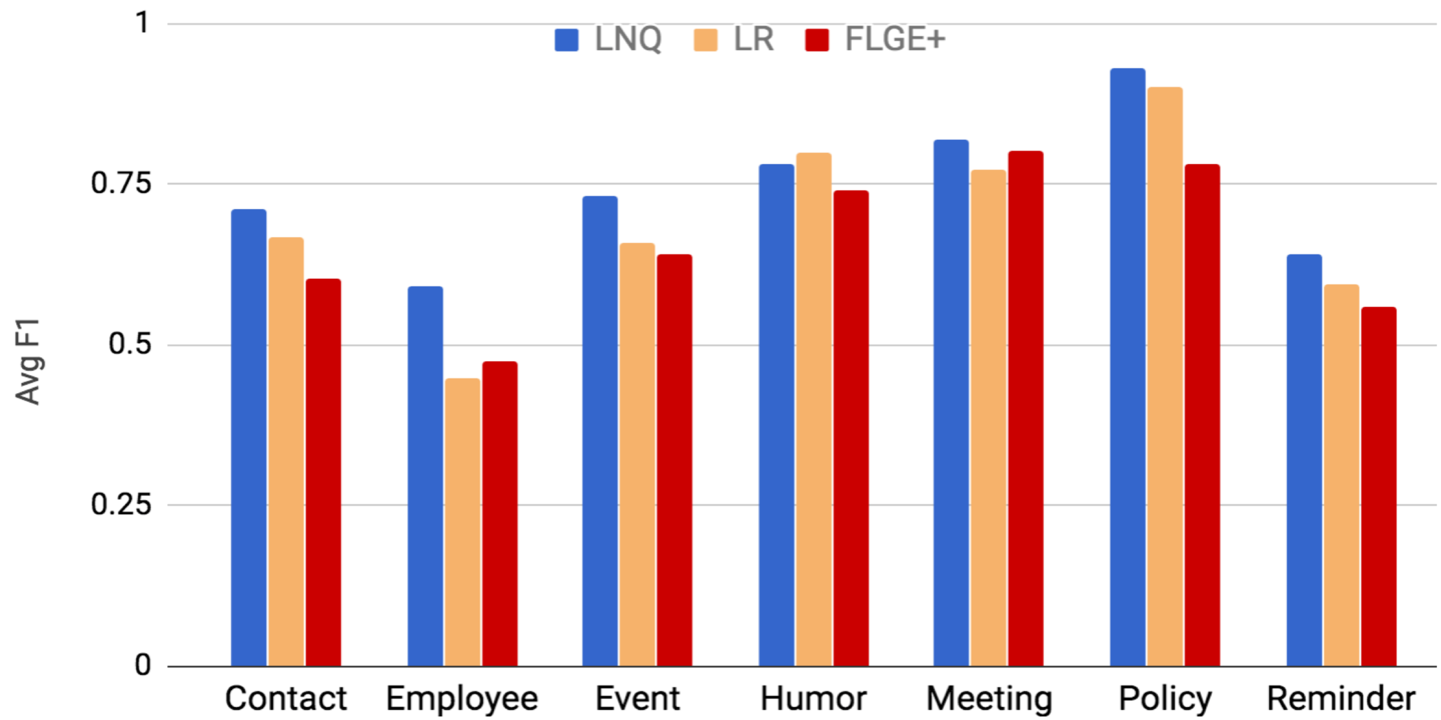
<u>*Example statements:*</u>

- *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



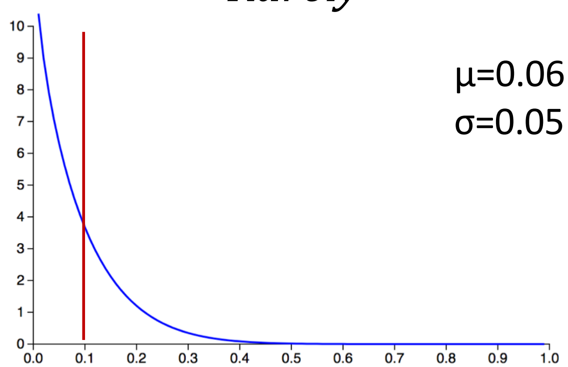
Email Categories

Performance by training from both quantification and labels

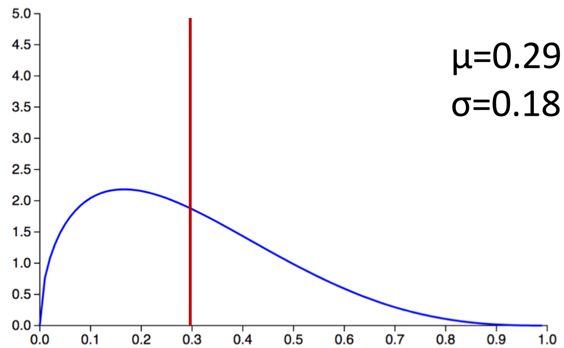
- ➢ About a third of statements used quantifiers

Empirical distributions of probability values

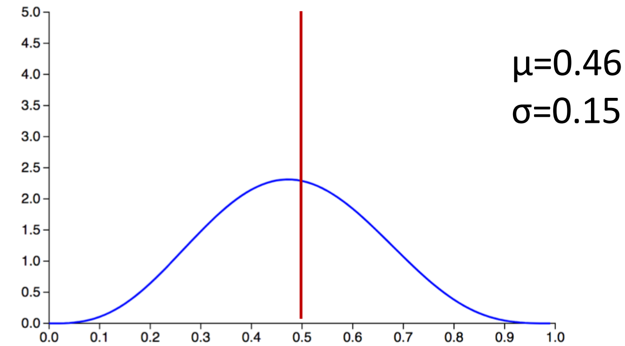
Rarely



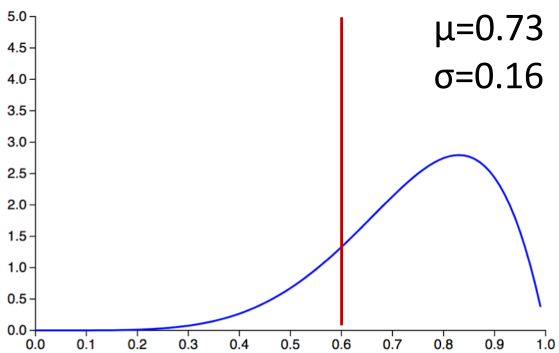
Sometimes



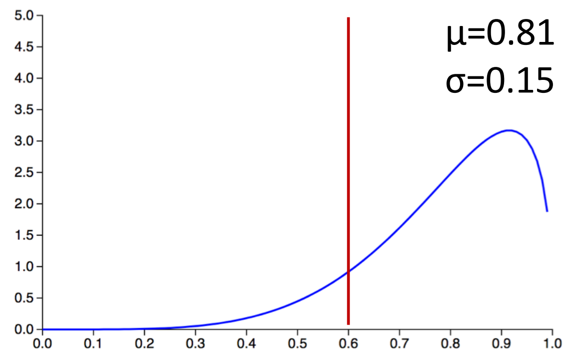
Often



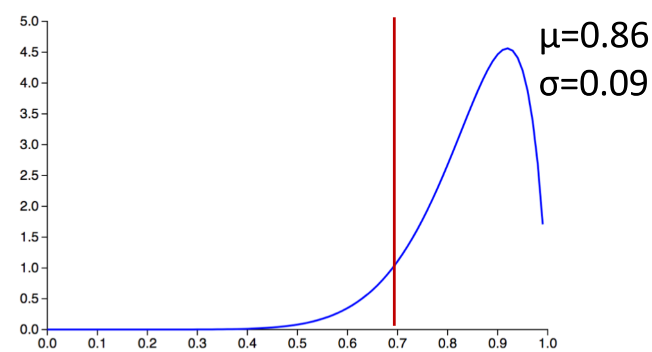
Majority



Most



(most)Likely



Summary

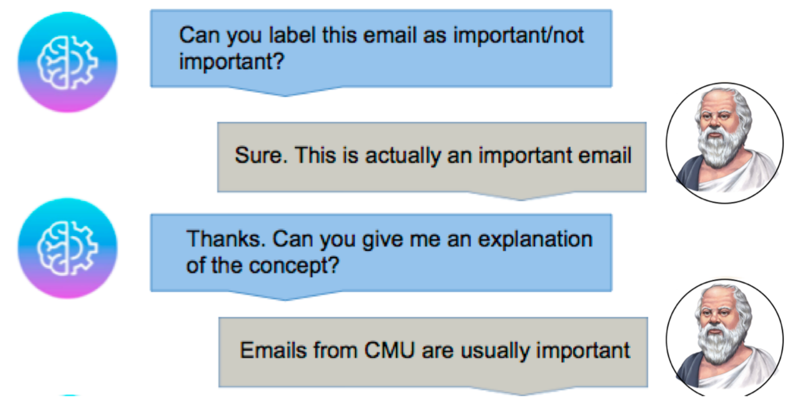
- *Declarative NL can supervise learning in limited data settings*

- *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?



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

- Pairing explanations with demonstrations, curricular learning,…

Questions?