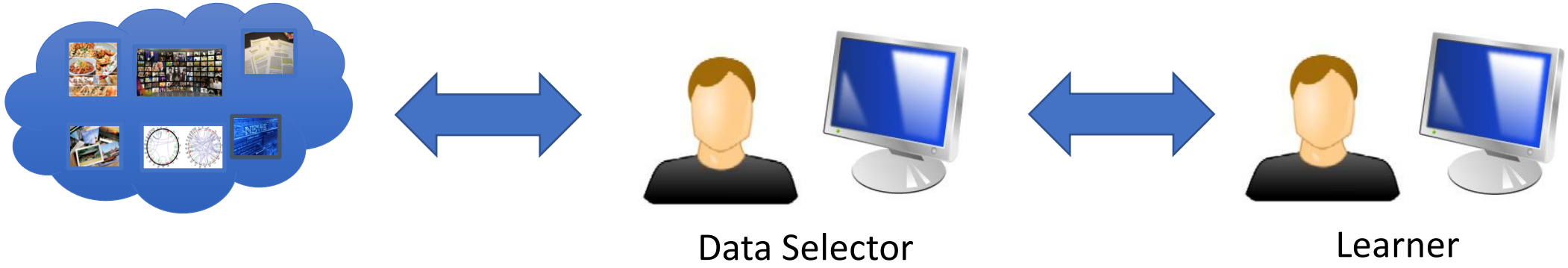
Machine Teaching for Human Learners

Yisong Yue

Caltech

Machine Teaching vs Machine Learning



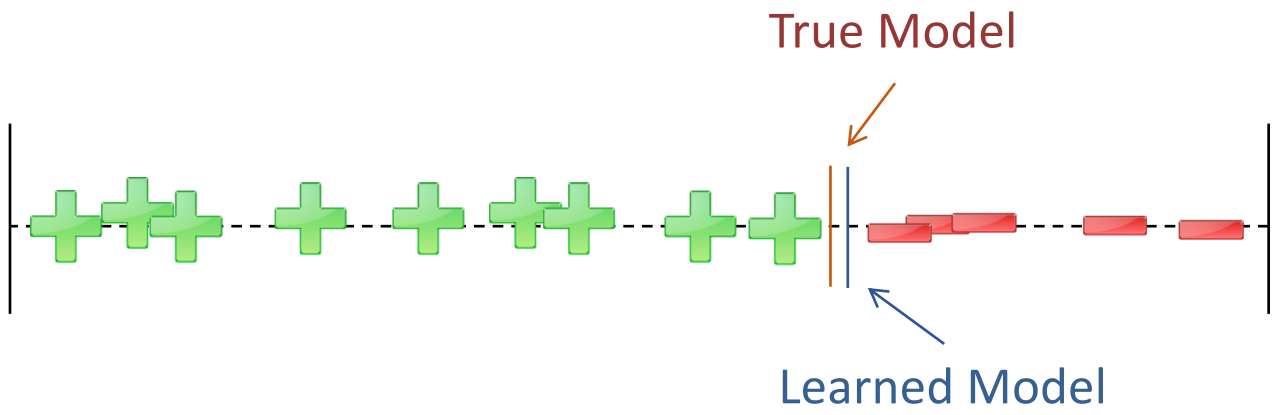
	Learner Type	Data Selector	Know True Model?
Passive Learning	We Control	IID Sampling	No
Active Learning	We Control (sometimes)	We Control	No
Teaching	Given to Us	We Control	Yes

An Overview of Machine Teaching: https://arxiv.org/abs/1801.05927

Simple Example

- 1 feature
- Noise free & realizable
- Learn threshold function

Passive Learning
Sample IID from distribution

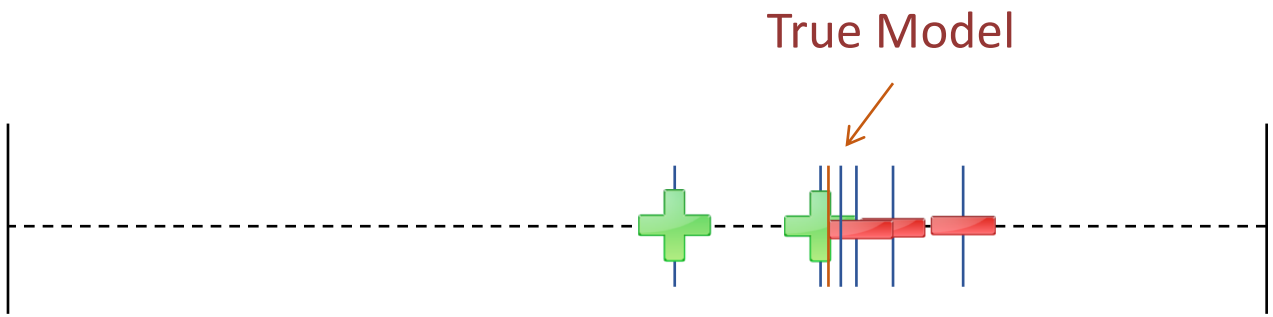


Simple Example

- 1 feature
- Noise free & realizable
- Learn threshold function

Active Learning

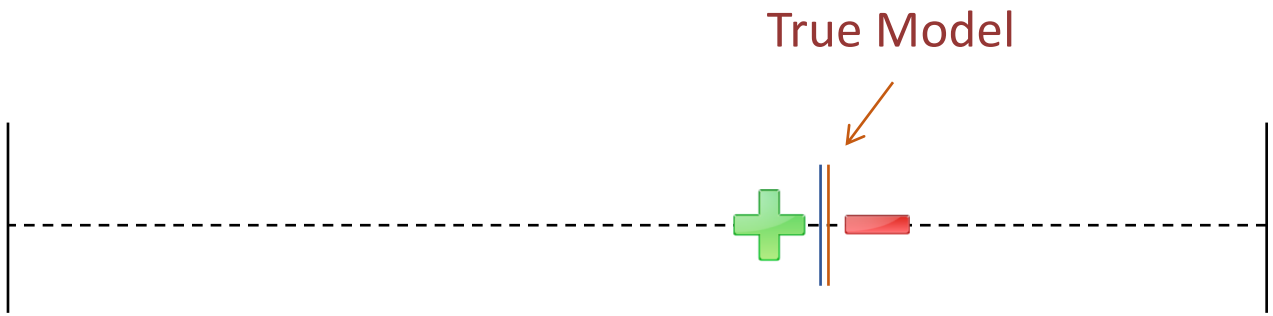
Binary Search to Reduce Uncertainty



Simple Example

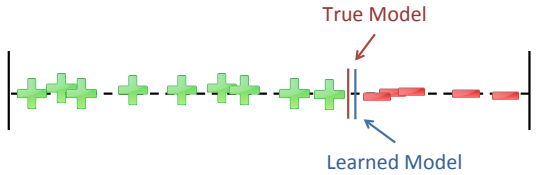
- 1 feature
- Noise free & realizable
- Learn threshold function

Teaching
Choose examples knowing true model

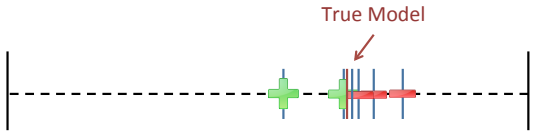


Comparison (Simple Example)

- # samples to be within ε of true model
- Passive Learning: $O\left(\frac{1}{\varepsilon}\right)$

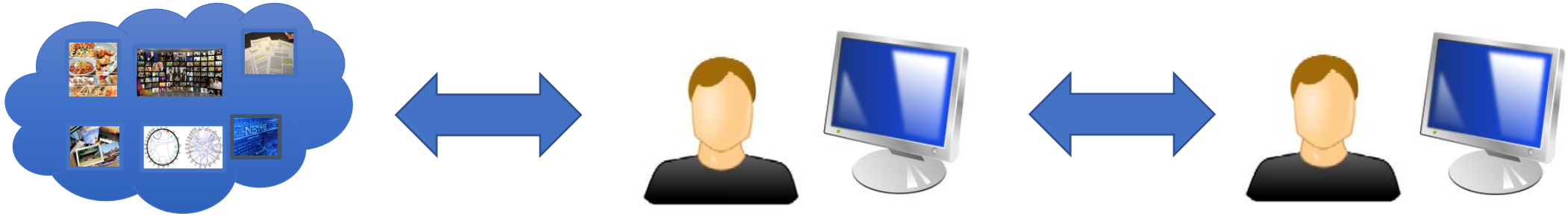


- Active Learning: $O\left(log\frac{1}{\varepsilon}\right)$



- Teaching: $O(1)$

Machine Teaching vs Machine Learning



Data/Model | Data Selector | Learner

	Learner Type	Data Selector	Know True Model?
Passive Learning	We Control	IID Sampling	No
Active Learning	We Control (sometimes)	We Control	No
Teaching	Given to Us	We Control	Yes

An Overview of Machine Teaching: https://arxiv.org/abs/1801.05927

Why is Teaching Important?

- Understanding data poisoning
- Teaching non-experts
- Onboarding at a new organization
- Debugging & testing
- Etc…

Teaching Human Learners

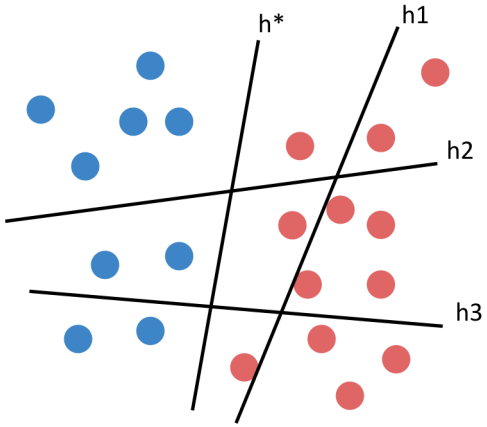
Data/Model

Teaching Algorithm

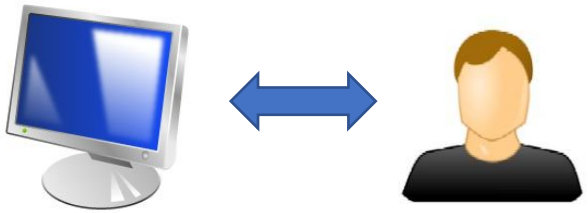
Human Learner

- Tailored to humans
  - Explanations
  - Forgetfulness
- Algorithmic questions
  - Structured prediction
  - Interactive algorithms

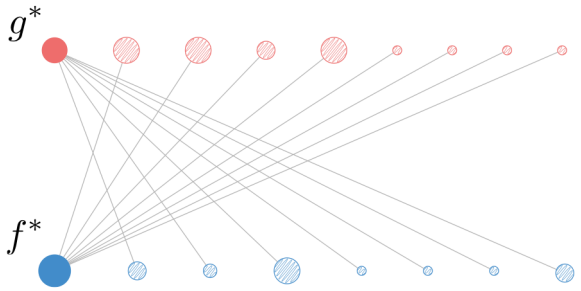
Design Choices



Learner Model



Interaction Protocol



Teaching Algorithm

Teaching with Explanations

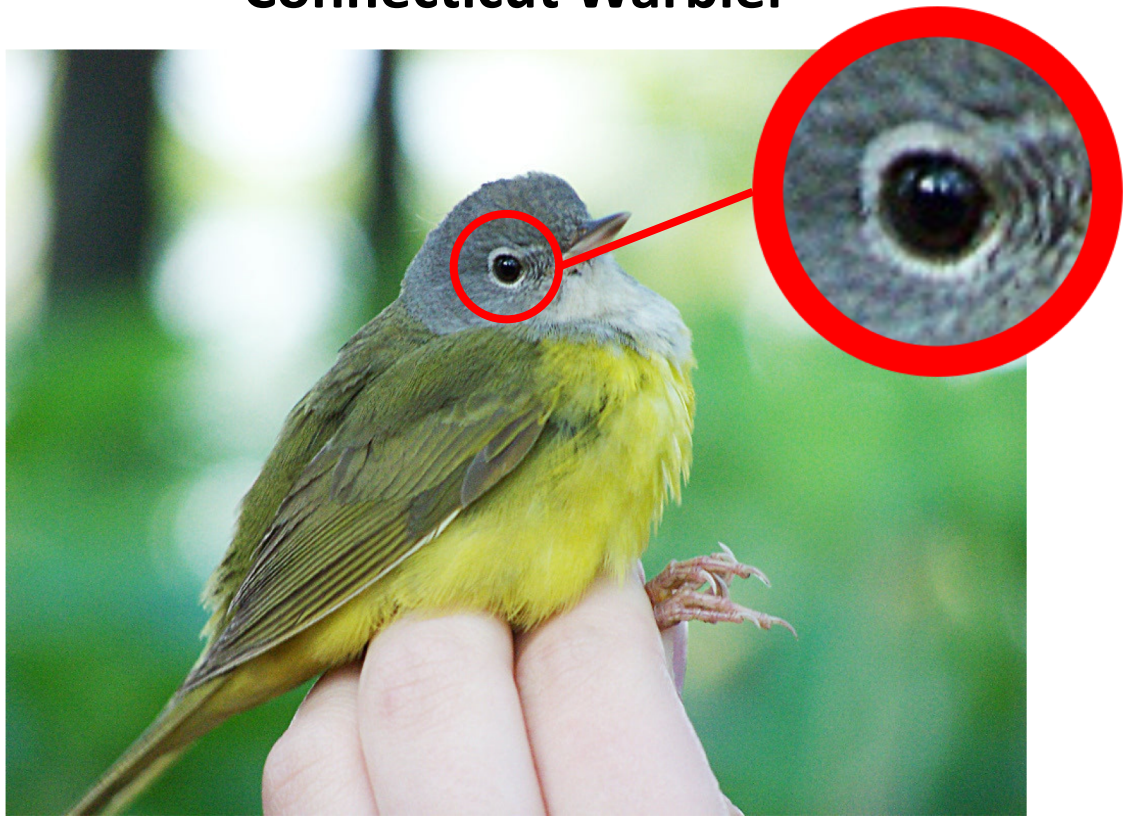
Near-Optimal Machine Teaching via Explanatory Teaching Sets
Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, Yisong Yue, AISTATS 2018

Teaching Categories to Human Learners with Visual Explanations
Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, Yisong Yue, CVPR 2018

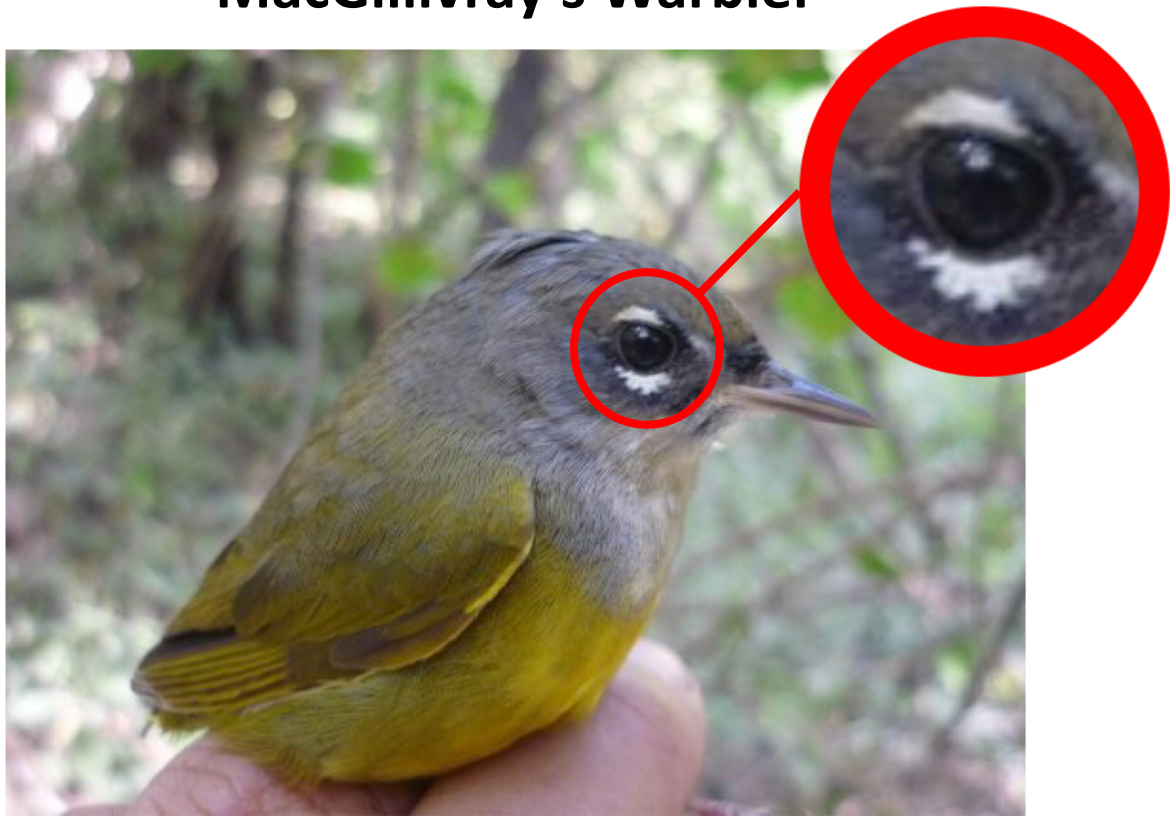
Connecticut Warbler
or ~~MacGillivray's Warbler~~

https://www.inaturalist.org/observations/9869215

Connecticut Warbler



MacGillivray's Warbler



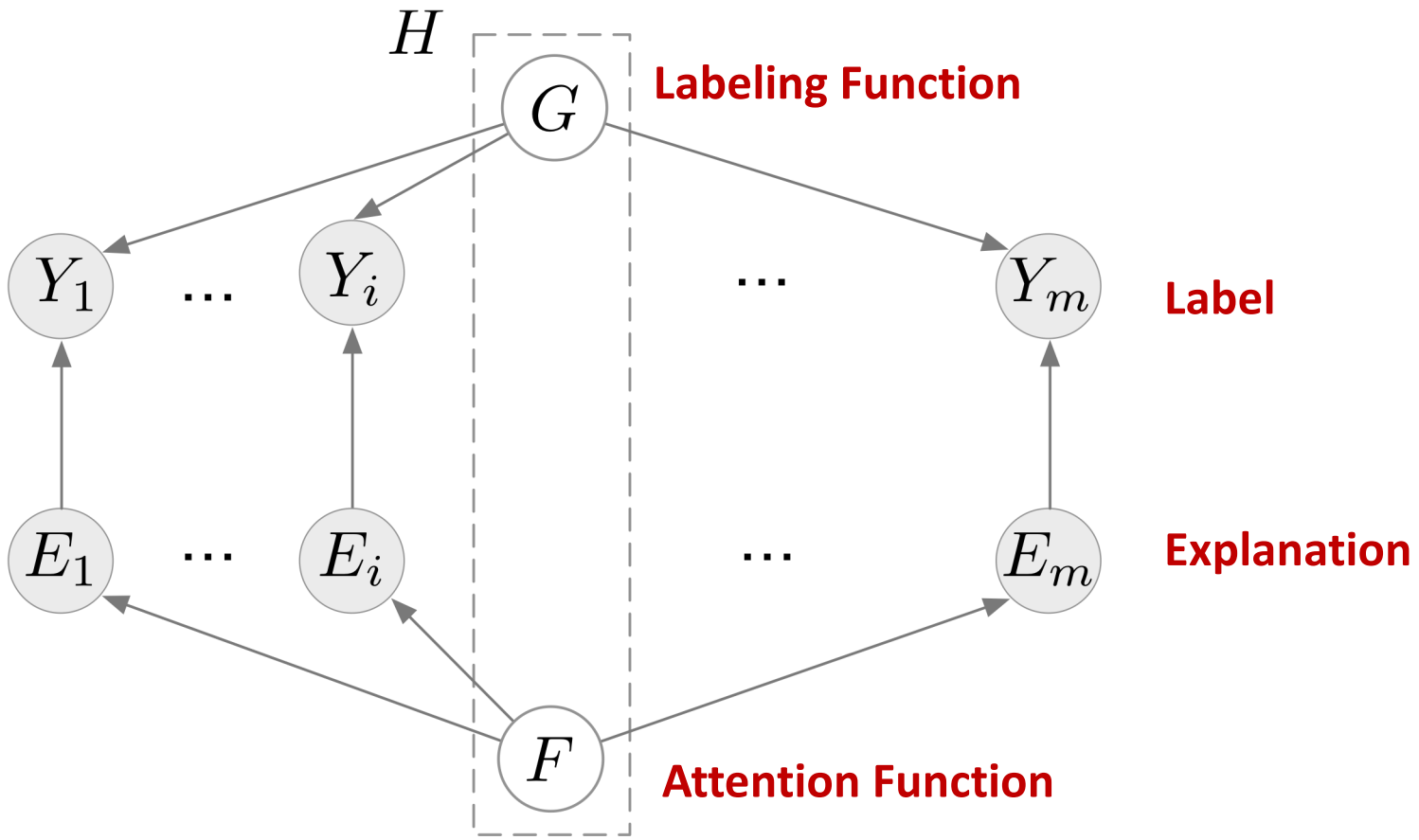
https://www.inaturalist.org/observations/9869215
https://www.inaturalist.org/observations/3949369

Instance-Level Labels
(conventional teaching)

Teaching with Explanations

Bayesian Learner Model

Index over entire dataset
Suppressing X for brevity



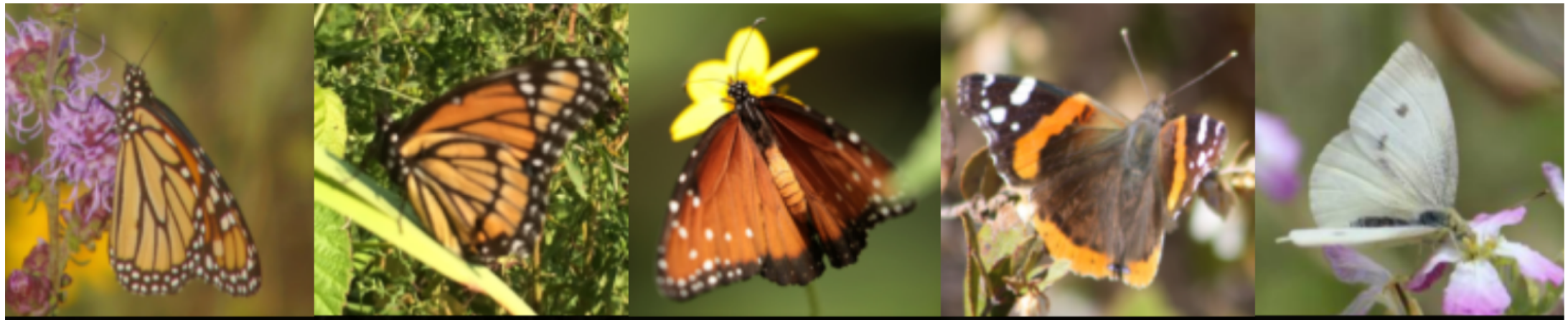
Butterflies

2,224 images, 5 classes

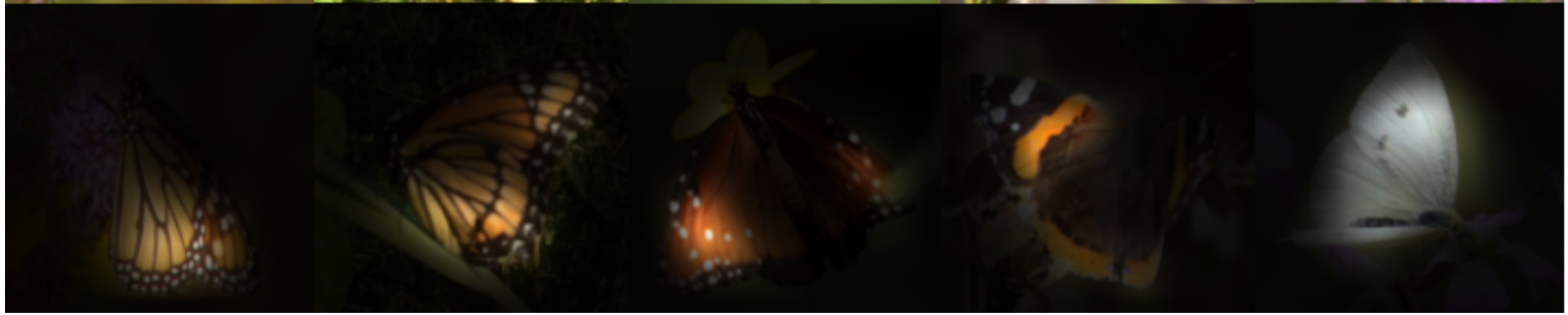
Y

Monarch | Viceroy | Queen | Red Admiral | Cabbage White

X



E



OCT Retina Images

1,125 images, 3 classes

Y

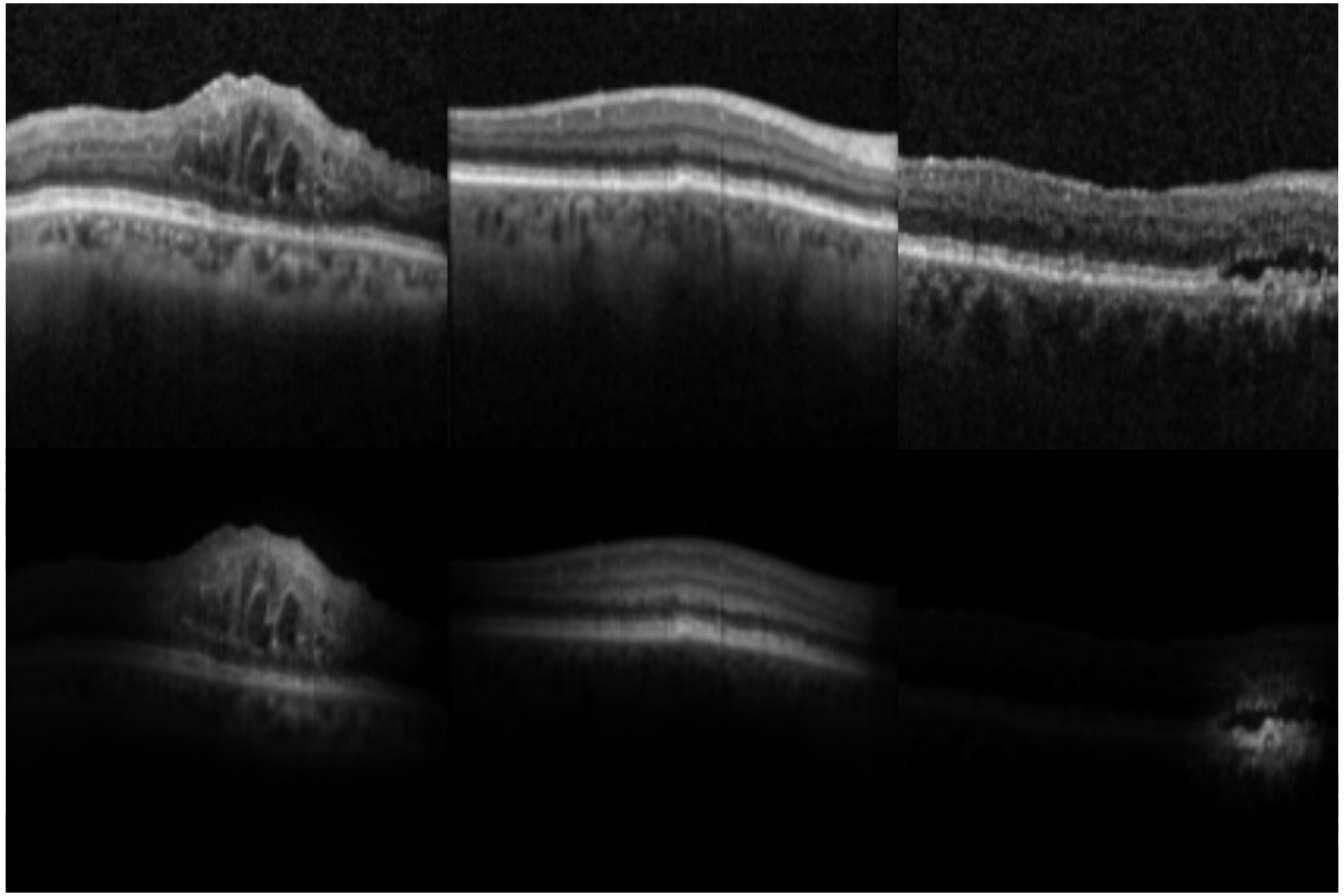
X

E

Macular Edema

Normal

Subretinal Fluid



Chinese Characters

717 images, 3 classes

Y Grass Mound Stem

X

E

Greedy Teaching Algorithm (non-interactive)

- **Prior:** $P(H = h)$

  Our belief over the learner's hypothesis

- **Teaching set:** S

  Set of (y,e) tuples (x implicit)

- **Posterior:** $P(H = h|Y_S, E_S)$

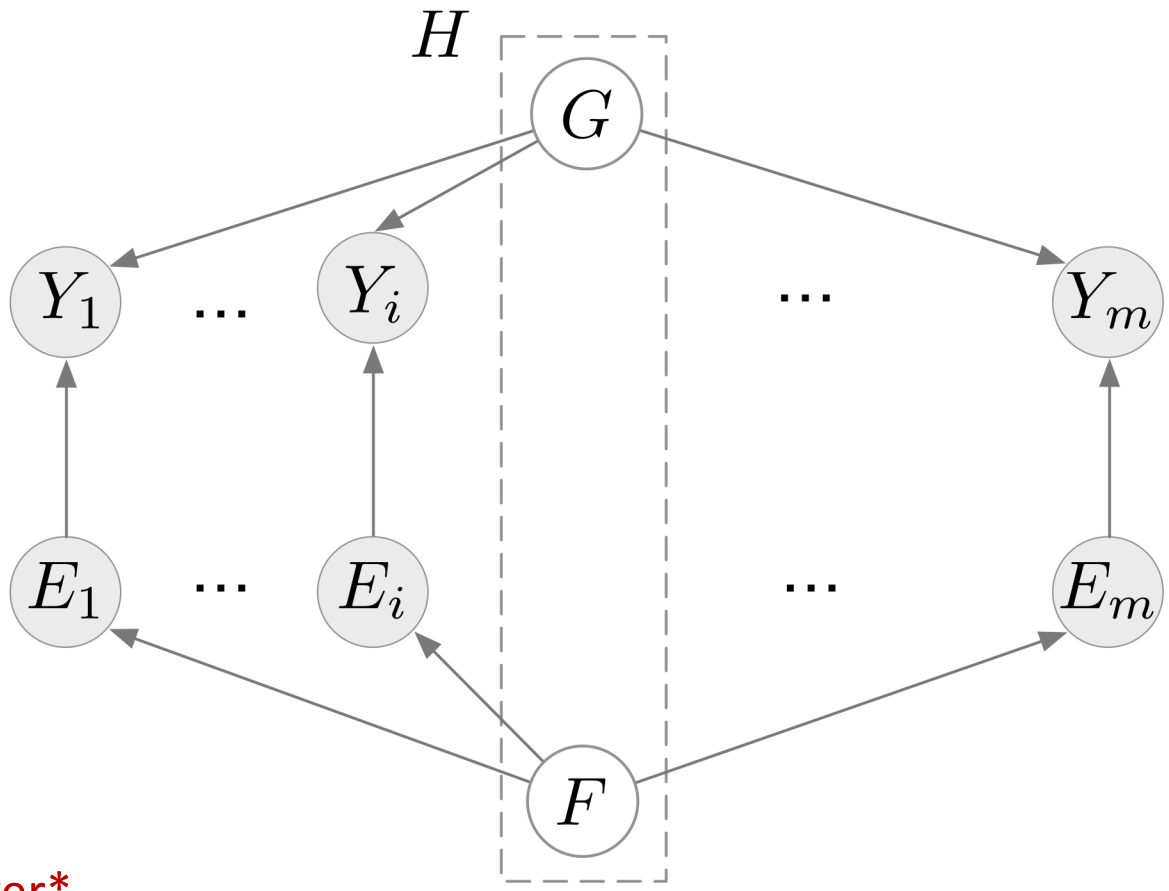
  Our belief over learner's hypothesis after teaching her with S

- **Goal:** Choose S

  (to minimize posterior error)

  Greedy algorithm!

  Analysis: reduction to submodular set cover*



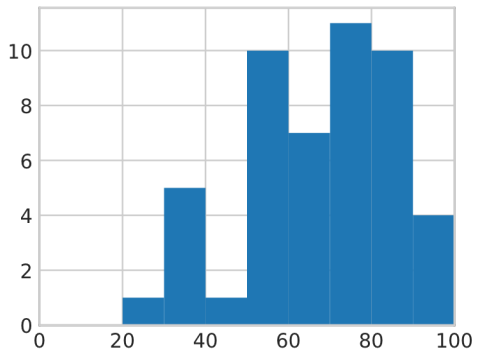
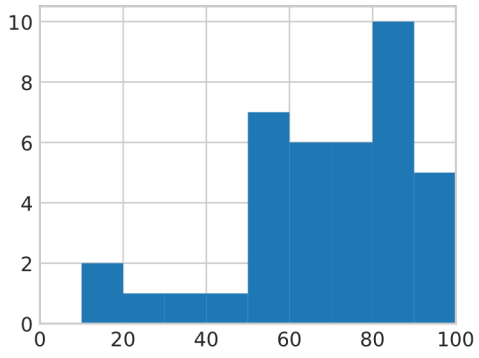
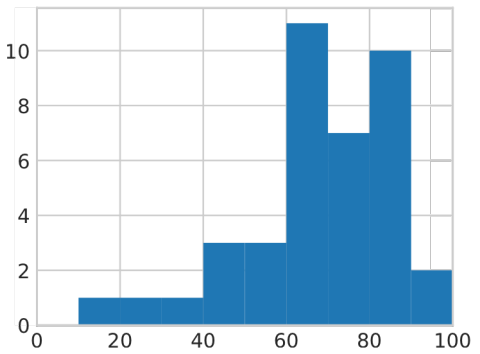
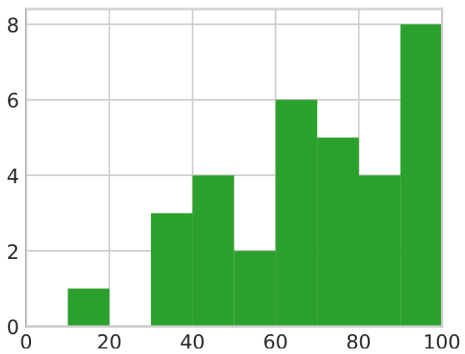
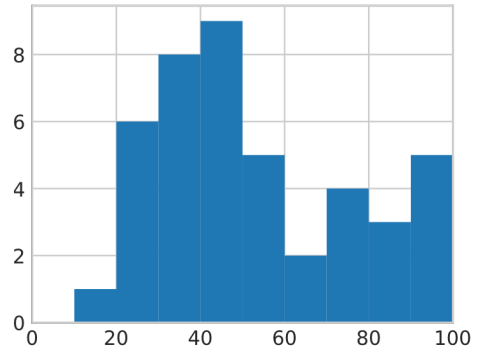
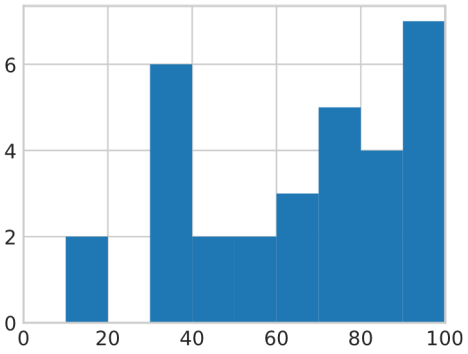
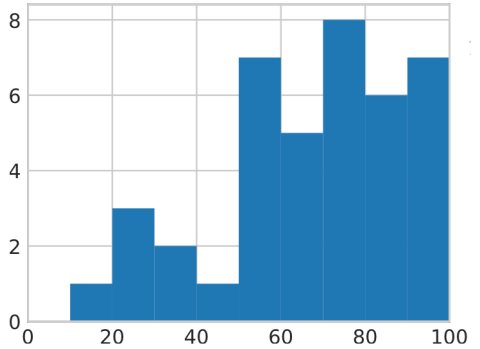
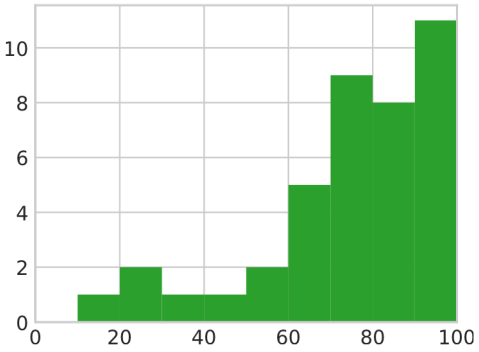
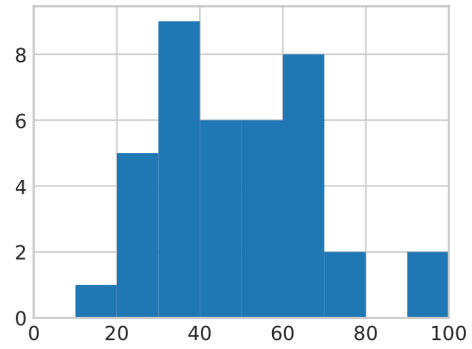
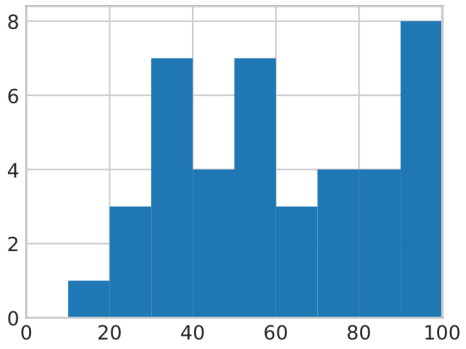
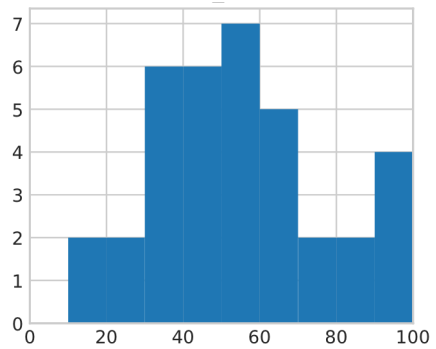
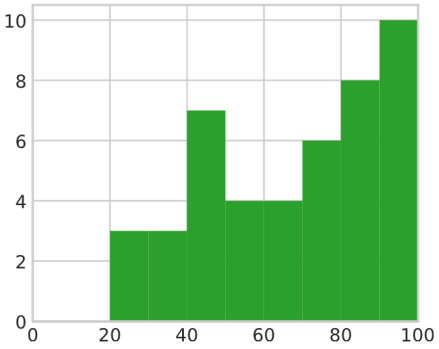
*Generalizes previous work on submodular teaching [Singla et al., ICML 2014]

User Study

Familiarize participants with interface

Teach for 20 iterations

Test for 20 iterations (to measure performance)

	Random Label Only	Random w/ Explain	Greedy Label Only	Our Approach
Butterflies				
OCT Eyes				
Chinese				

Teaching w/ Explanations

- Redefine teacher/learner interface
  - Humans don’t learn from only labels
- First formalization of explainable teaching
- Rigorously developed approach
- Good empirical results

Teaching Forgetful Learners

Teaching Multiple Concepts to Forgetful Learners
Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, Adish Singla, *(in submission)*

Modeling Forgetfulness

Half-life Regression (HRL) model [Settles & Meeder, ACL 2016]

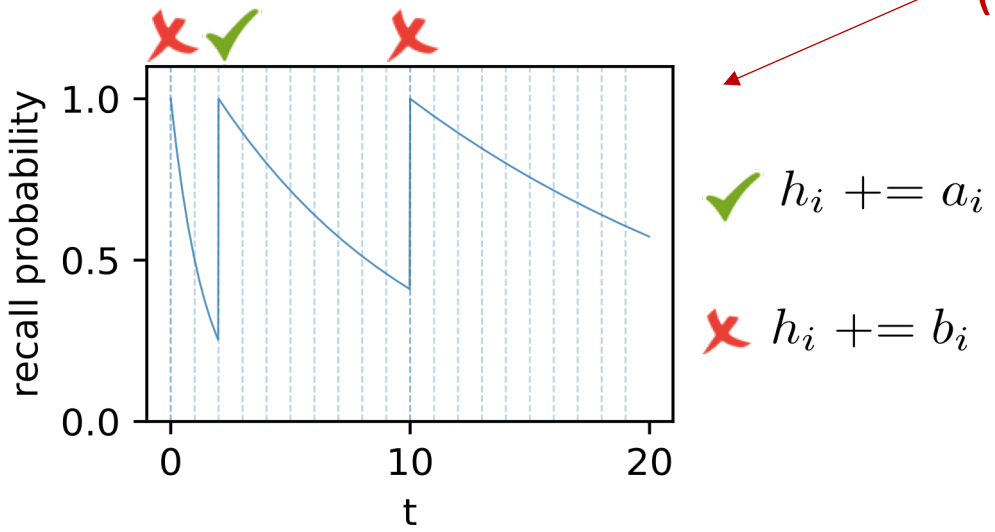
Time since last teaching concept

Time t

Recall Probability of Concept i:

$$p_i(t \mid \text{history}) = 2^{-\frac{\Delta t_i}{h_i}}$$

Half-life estimate (depends on feedback)



✓ $h_i \mathrel{+}= a_i$

✗ $h_i \mathrel{+}= b_i$

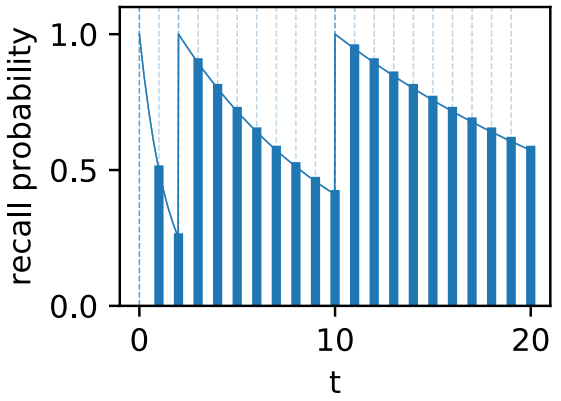
Interactive Teaching Protocol

- For t = 1…T
  - Teacher chooses concept $i \epsilon \{1, \ldots, m\}$ (e.g., a flashcard)
  - Learner tries to recall concept (success or fail)
  - Teacher reveals answer (e.g., “Spielzug”)

- Goal: maximize

$$f(\text{history}) = \frac{1}{m}\frac{1}{T}\sum_{i=1}^{m}\sum_{t=1}^{T} p_i(t|\text{history}_{1:t-1})$$

“Area Under Curve”



Naive Approaches

- Round Robin
  - Doesn’t adapt to new estimates of learner recall probabilities
  - Over-teaches easy concepts
  - Under-teaches hard concepts

- Lowest Recall Probability
  - Doesn’t consider change to recall probability

Greedy Teaching Algorithm (interactive)

- Choose concept i to maximize

$$\Delta(i|\text{history}) = E_{y_t}\left[f\left(\text{history}\oplus(i, y_t)\right) - f\left(\text{history}\right)\right]$$

y_t: success or failure of recall at time t

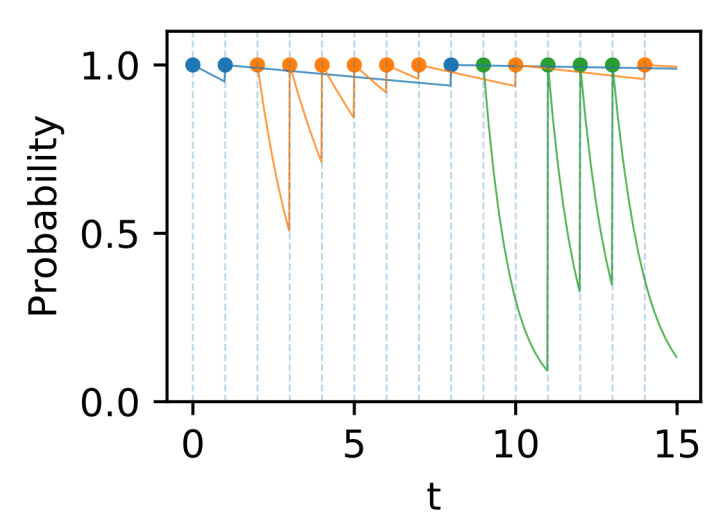
(randomness over model estimate)

$$p_i(t \mid \text{history}) = 2^{-\frac{\Delta t_i}{h_i}}$$ (h_i updated after observing y_t)

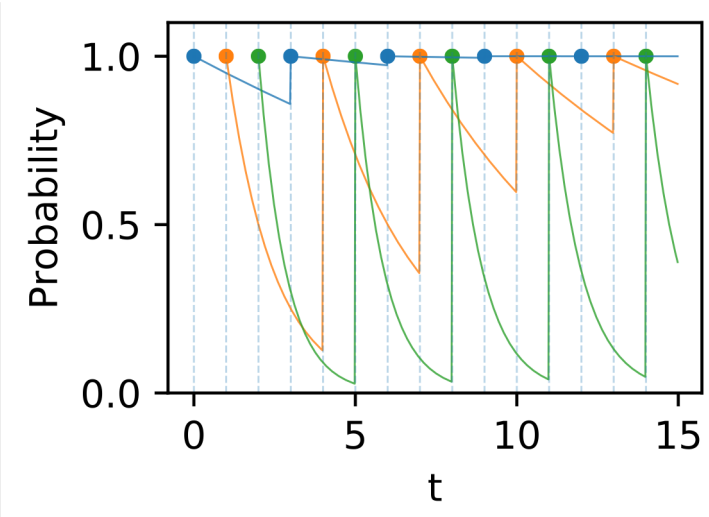
*Theoretical analysis in paper based on submodular ratio

Simulation Results

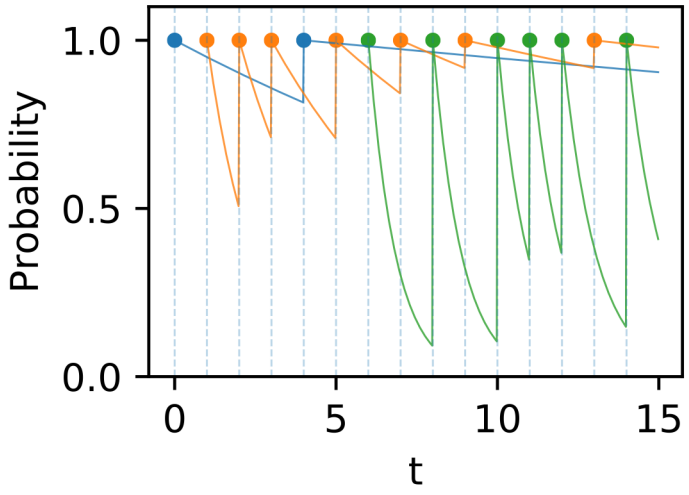
Greedy

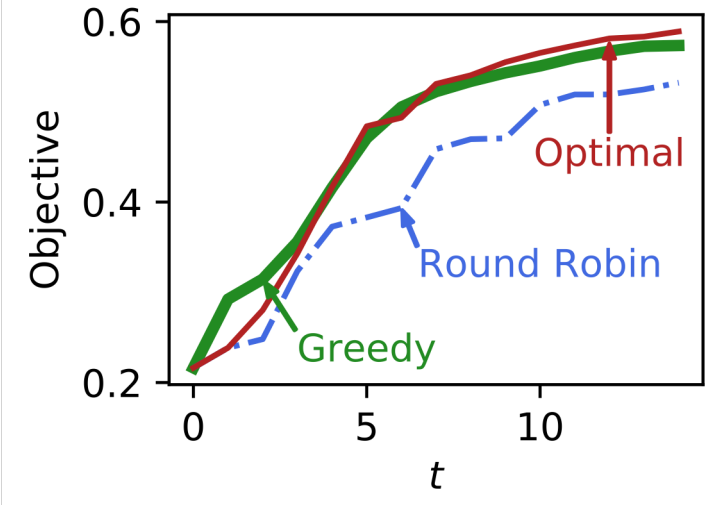


Round Robin

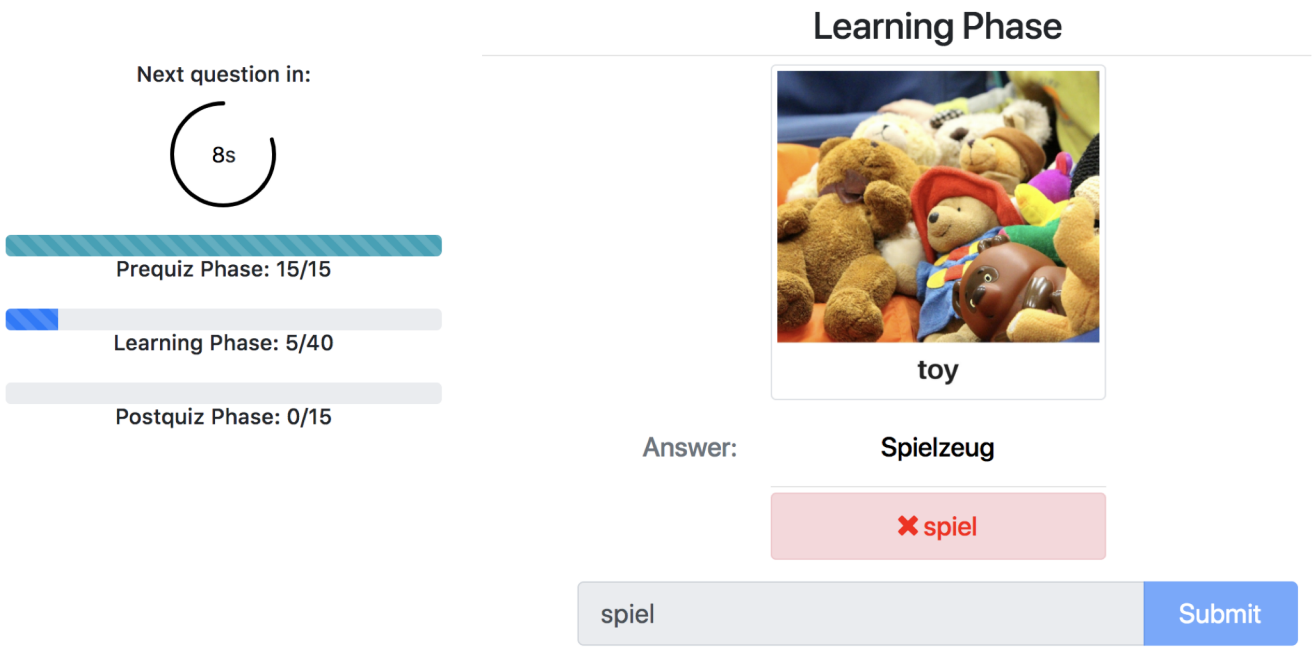


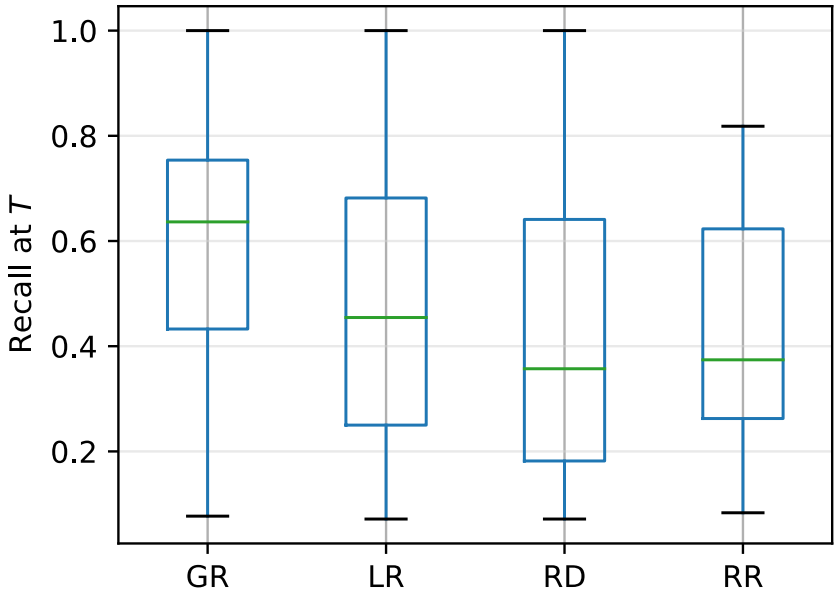
Optimal





User Study





- 150 participants from Mechanical Turk platform
- T=40, m=15, total study time is about 25 mins

Teaching Human Learners

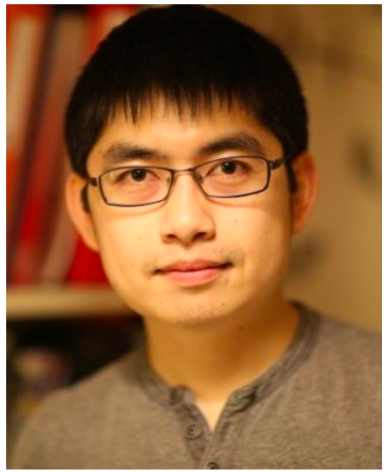
Data/Model

Teaching
Algorithm

Human
Learner

- Tailored to humans
  - Explanations
  - Forgetfulness
- Algorithmic questions
  - Structured prediction
  - Interactive algorithms

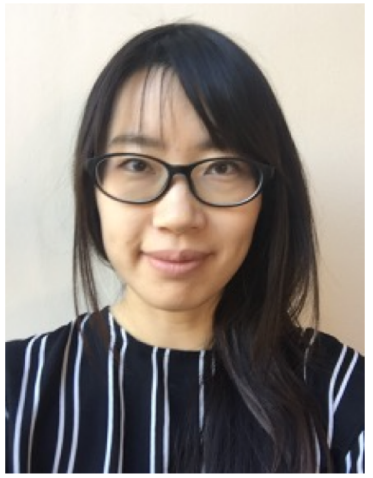
Online learning platforms: German vocabulary: https://www.teaching-german.cc/
Species names: https://www.teaching-biodiversity.cc/



Yuxin Chen

Oisin Mac Aodha

Anette Hunziker



Shihan Su

Adish Singla

Pietro Perona

Manuel Gomez Rodriguez

Andreas Krause

References

Near-Optimal Machine Teaching via Explanatory Teaching Sets
Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, Yisong Yue, AISTATS 2018

Teaching Categories to Human Learners with Visual Explanations
Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, Yisong Yue, CVPR 2018

Teaching Multiple Concepts to Forgetful Learners
Anette Hunziker, Yuxin Chen, Oisin Mac Aodha, Manuel Gomez Rodriguez, Andreas Krause, Pietro Perona, Yisong Yue, Adish Singla, *(in submission)*

Online Platforms:

German vocabulary: https://www.teaching-german.cc/

Species names: https://www.teaching-biodiversity.cc/