Human-Machine Cooperative Classification

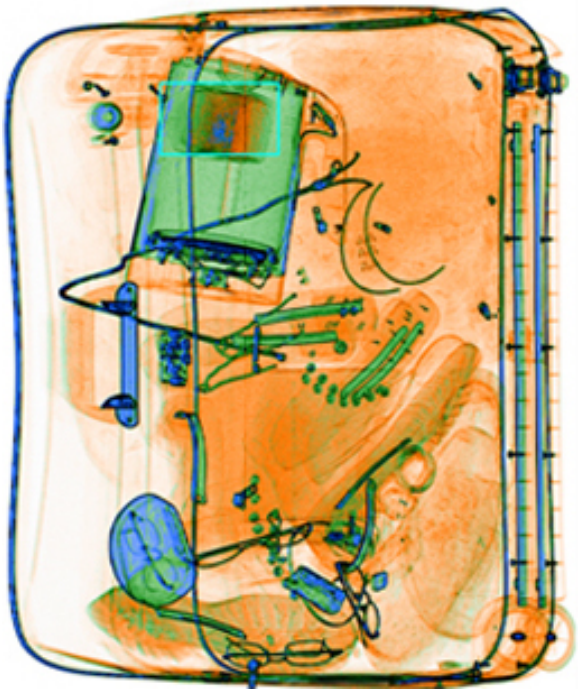
Ron Kneusel

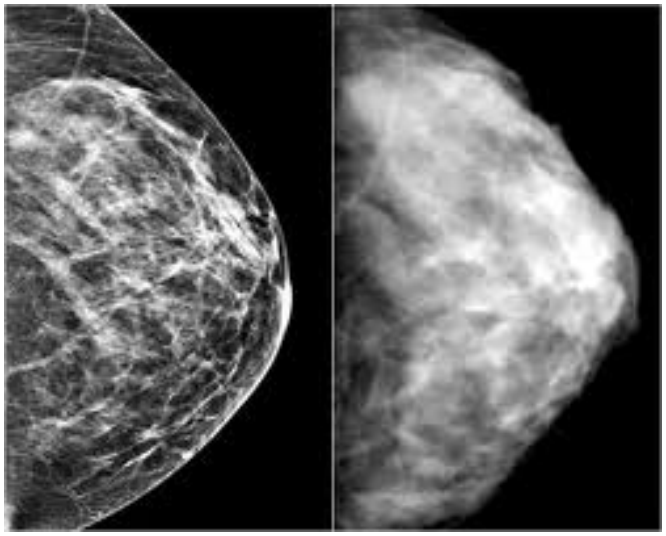
Brett D. Roads

Michael C. Mozer

University of Colorado, Boulder

Challenging Problems For People And Machines In Visual Search and Classification



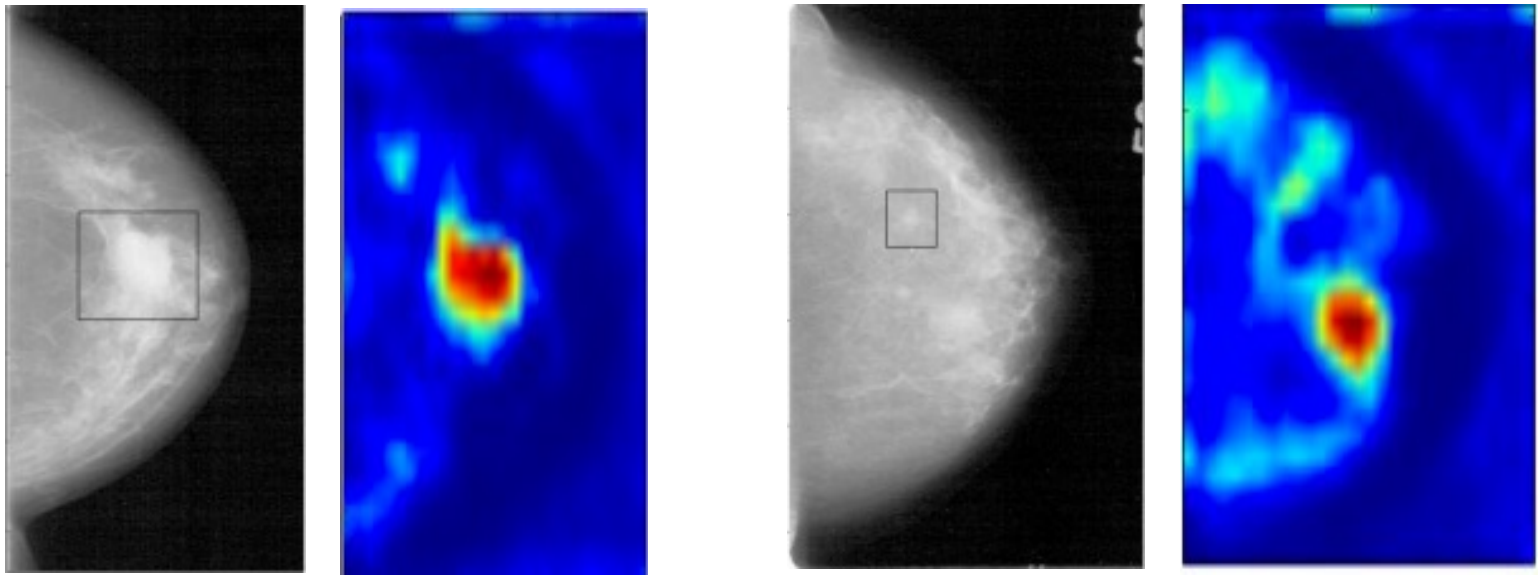


Two Projects

Can we beat the performance of *either* human experts and ML classifiers with a system that allows *both* to cooperate?

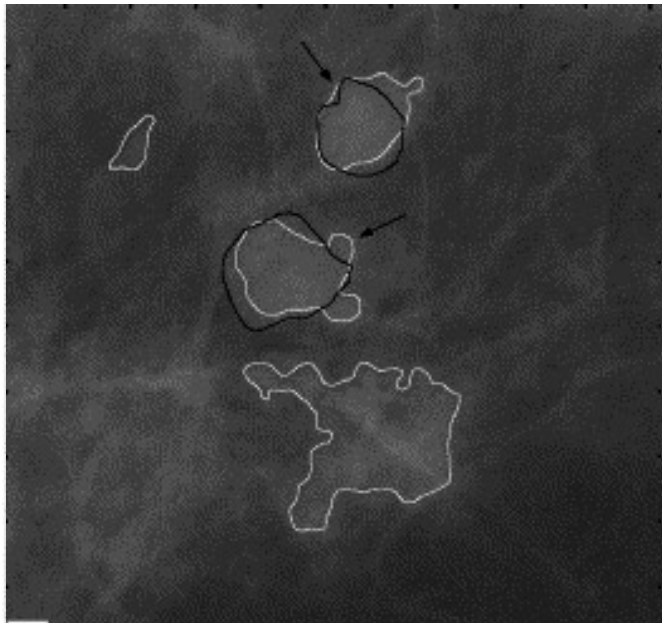
Can we develop methods of querying novices that allow them to classify like an expert?

The Promise of Machine Learning



Images from Ertosun & Rubin (2015)

Can Classifiers Be Used To Boost Human Performance?



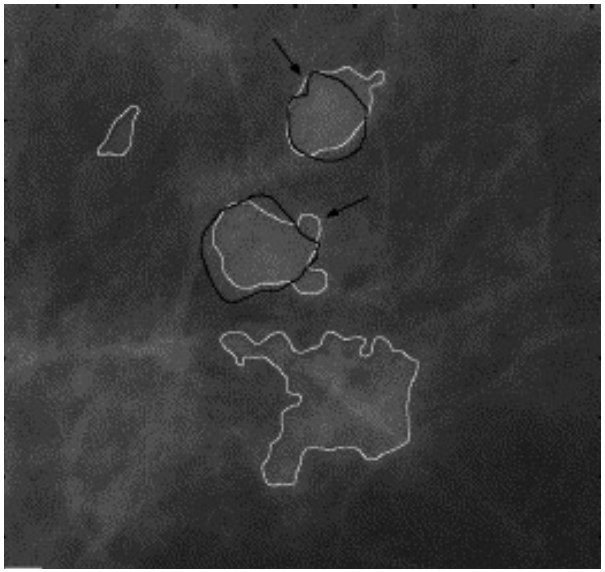
Human-machine cooperative visual search

- Highlights placed at locations identified as possible targets by classifier

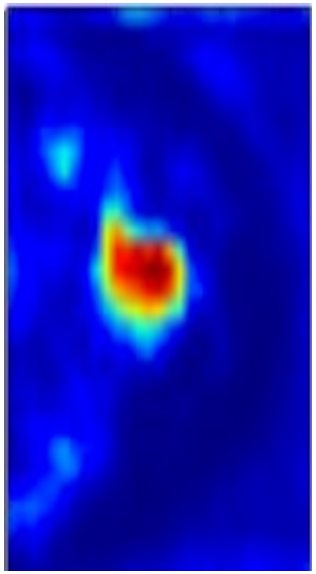
Serious challenges (Alberdi et al., 2004; Drew et al., 2012; Krupinski et al., 1993; Zheng et al., 2001, 2004)

- Highlighted locations are often obvious targets or clearly nontargets
- Nonhighlighted locations missed more often

Focus Of Our Research



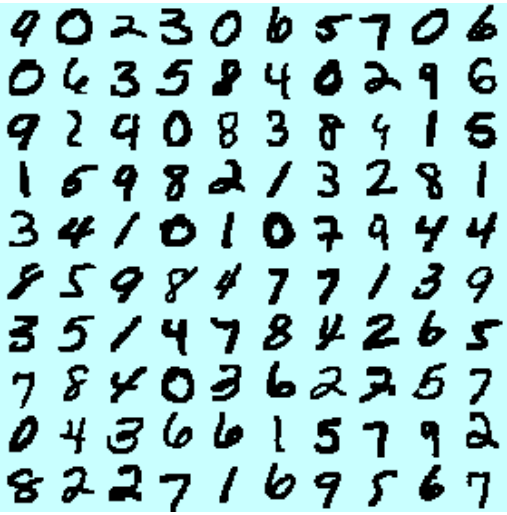
Highlights are all-or-none



Classifiers typically output confidence levels or probabilities

Can we leverage graded information from classifiers to improve human performance?

Experiments 1-3



hard highlighting

soft highlighting

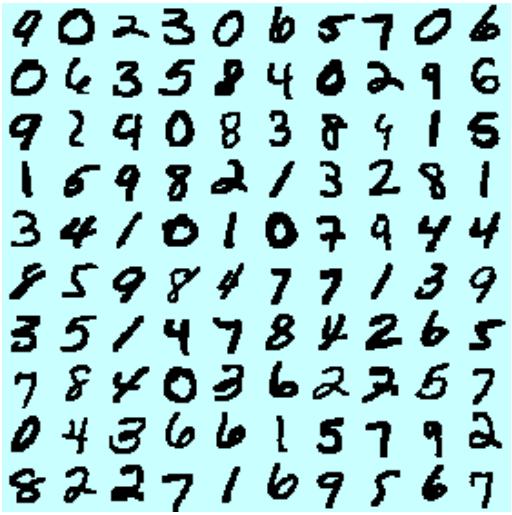
Search digit array for instances of target “2”

Click on target

- target replaced by background color

Multiple targets

Motivation For Soft Highlighting



soft highlighting

Soft highlighting

- provides additional information if classifier is well calibrated
- can leverage bottom-up attentional guidance

  assuming saliency ~ classifier confidence

Experiment 1

10 targets per display

Mechanical turk study

- n=25, 42 trials total

We manipulate quality of classifier

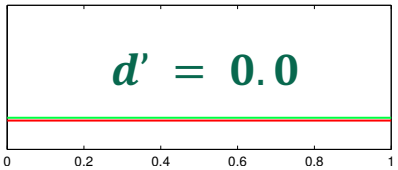
Strong classifier

Weak classifier

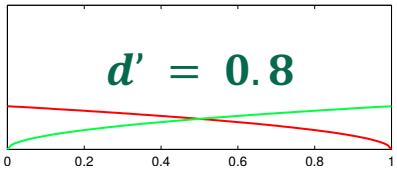
Seven Conditions

control

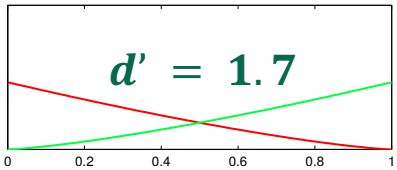
Beta(1,1) vs. Beta(1,1)



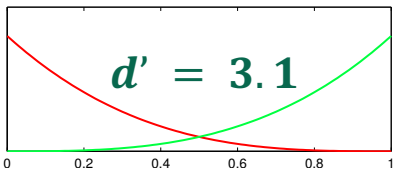
Beta(1,1.5) vs. Beta(1.5,1)



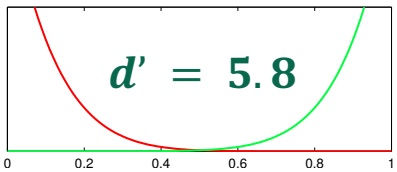
Beta(1,2.33) vs. Beta(2.33,1)



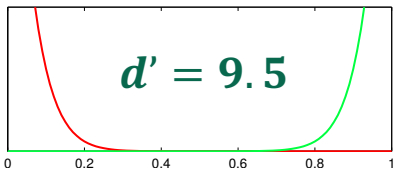
Beta(1,4) vs. Beta(4,1)



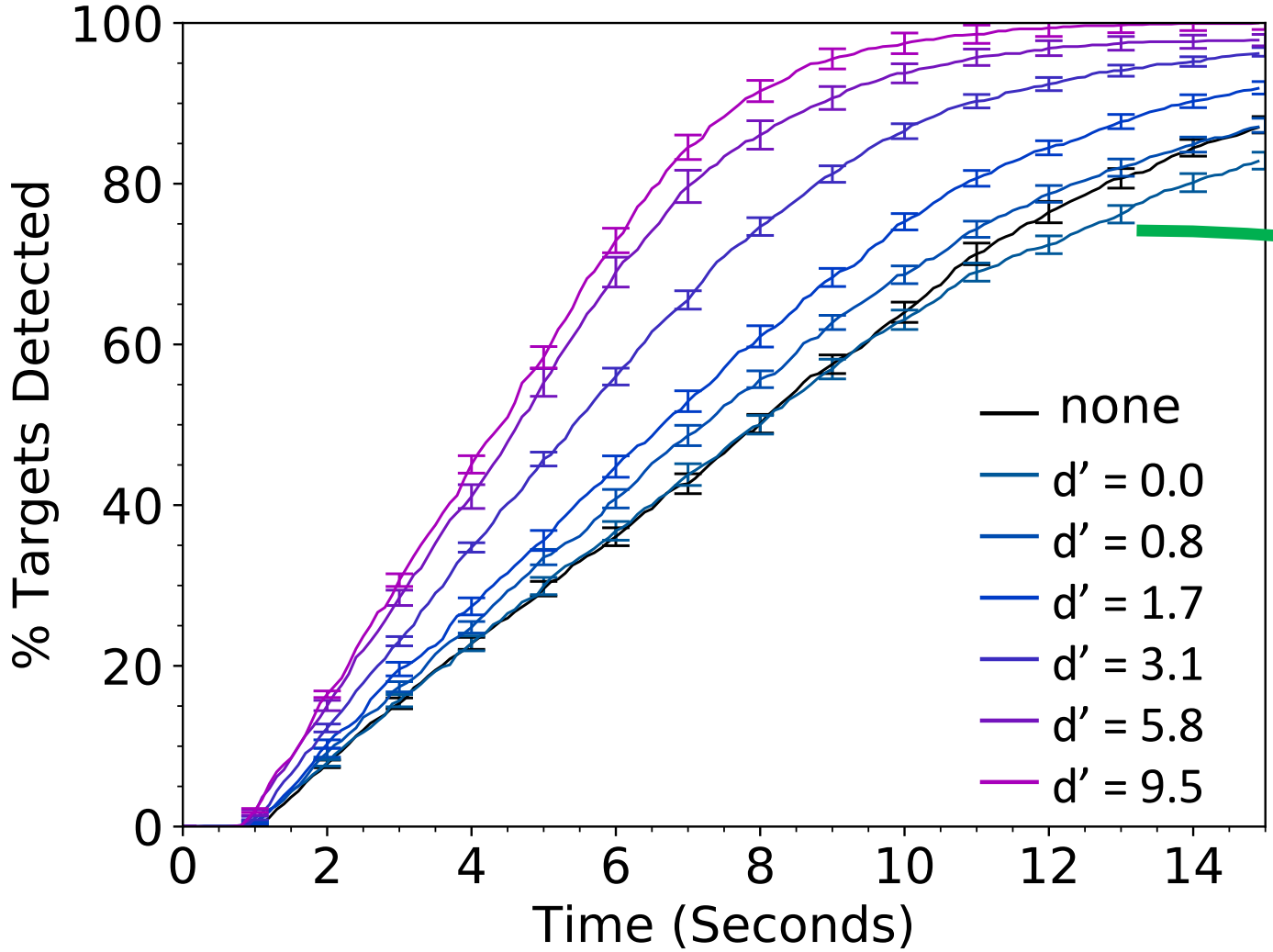
Beta(1,9) vs. Beta(9,1)



Beta(1,19) vs. Beta(19,1)



Experiment 1



1 SEM corrected for between subject variability (Masson & Loftus, 2003)

Even weak classifiers can boost human performance with soft highlighting

Experiment 2

Compare

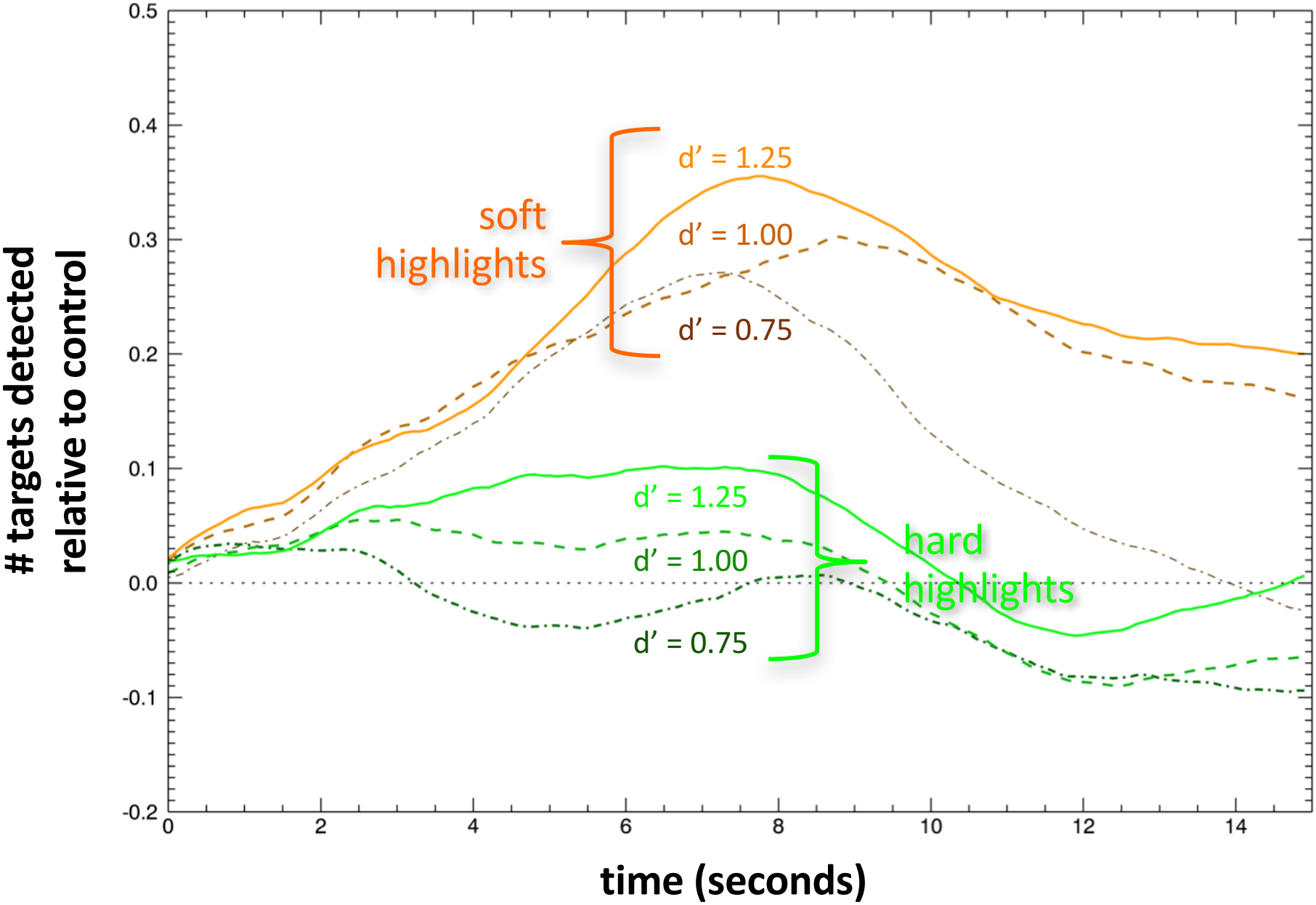
hard highlighting

soft highlighting

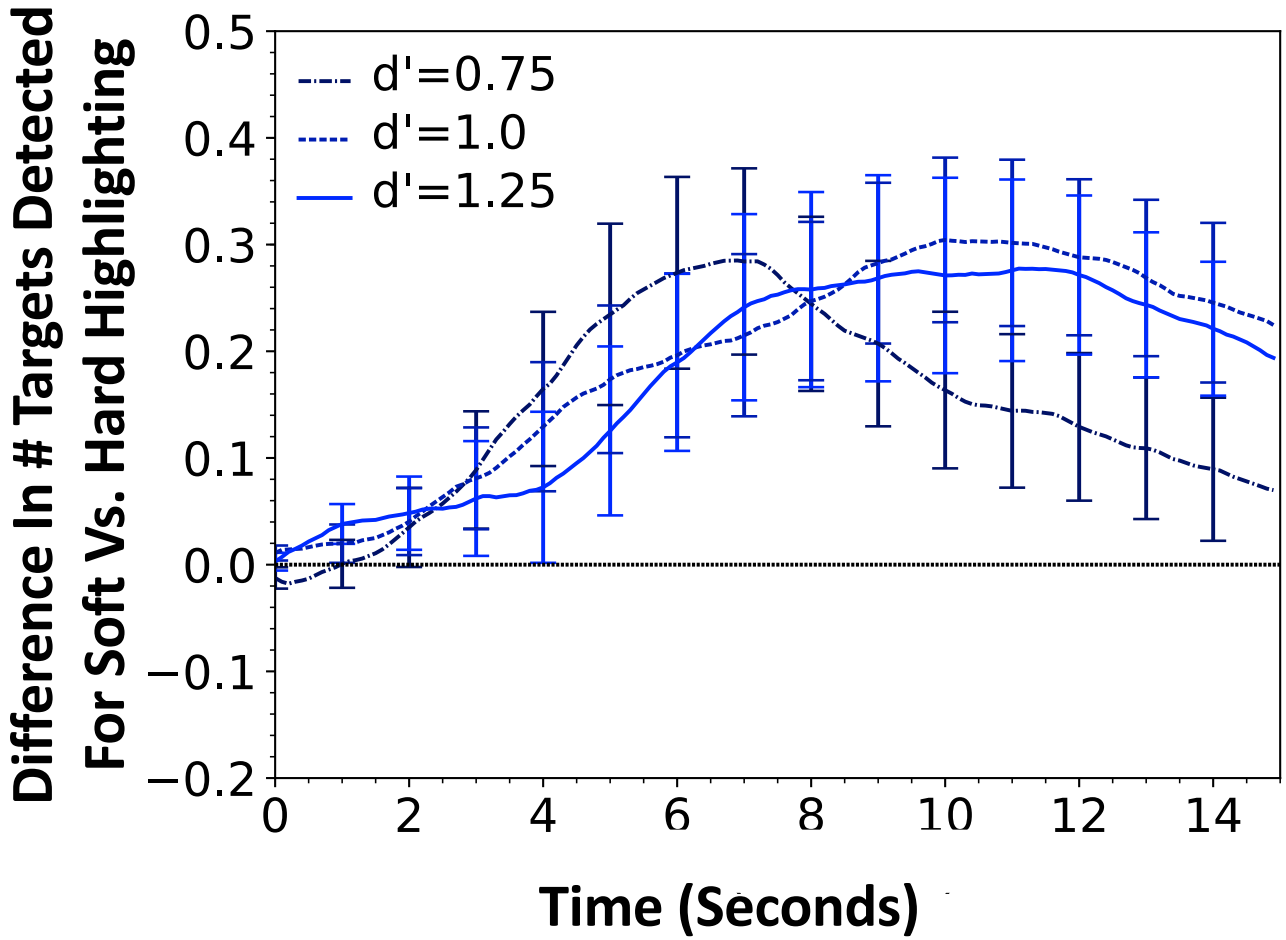
Vary classifier strength

- $d' \in \{0.75, 1.00, 1.25\}$

Experiment 2



Experiment 2



Soft highlighting consistently beats hard highlighting over time course of trial.

(Both beat control condition with no highlighting.)

Experiment 2.5

Does highlighting help by

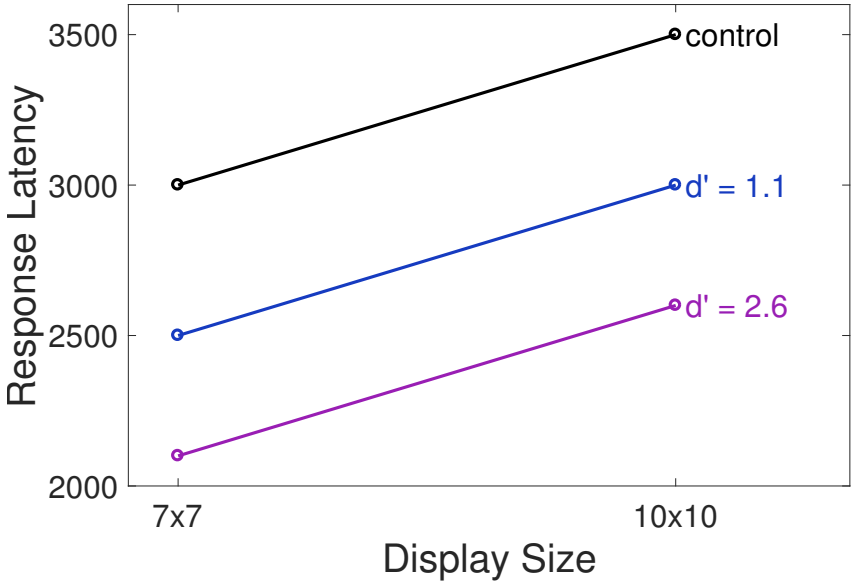
- segmenting and grouping elements in display, or
- making search more efficient?

To distinguish these alternatives,

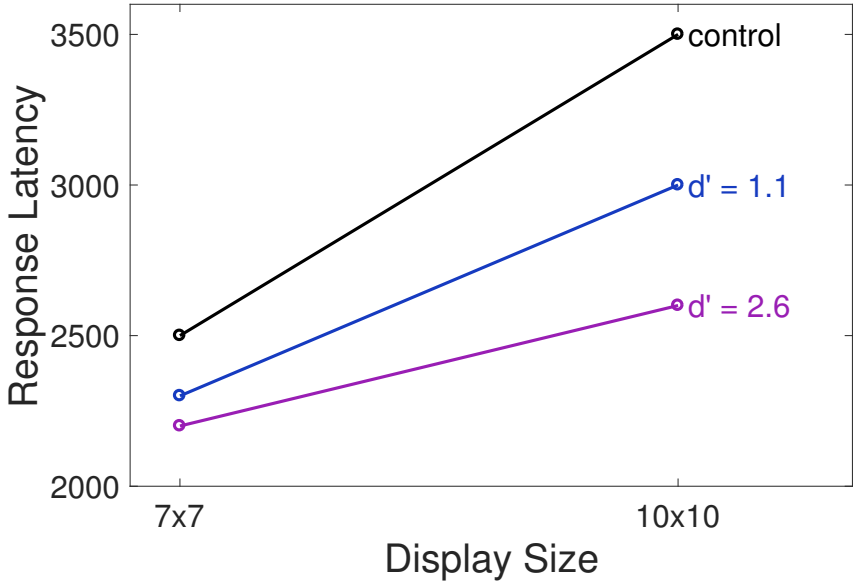
- Two display sizes (7x7 vs. 10x10)
- Two classifier strengths, $d' \in \{1.1, 2.6\}$
- single target "2"
- measure RT

Experiment 2.5

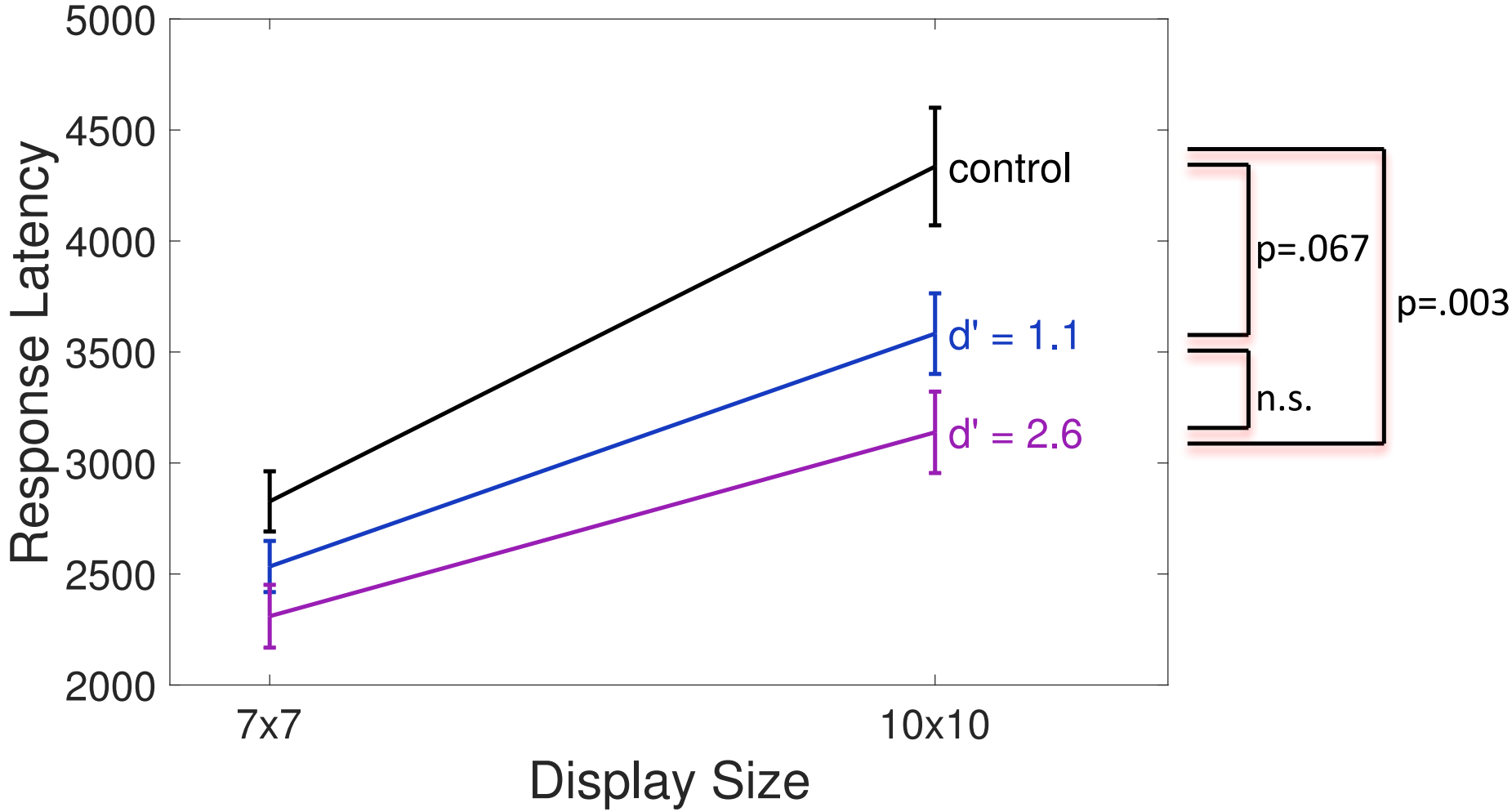
If highlighting helps segment and group display elements



If highlighting makes search more efficient



Experiment 2.5



Highlighting makes search more efficient

In Experiments 1 and 2, targets present on every trial.

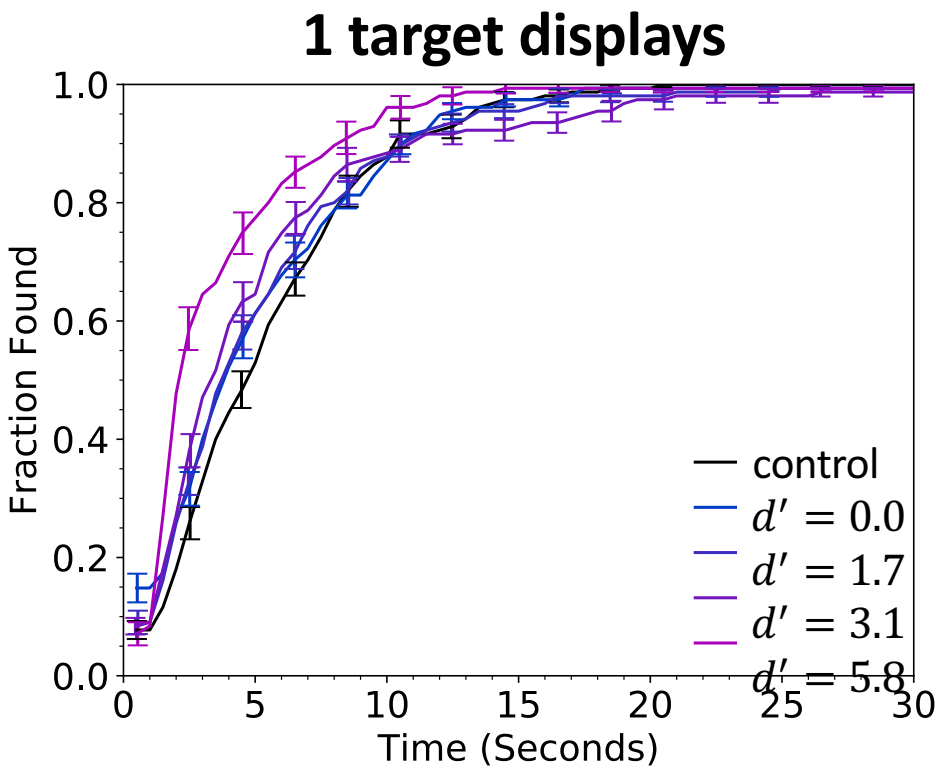
- Although highlighting speeds target detection, it may slow search when no target is present.

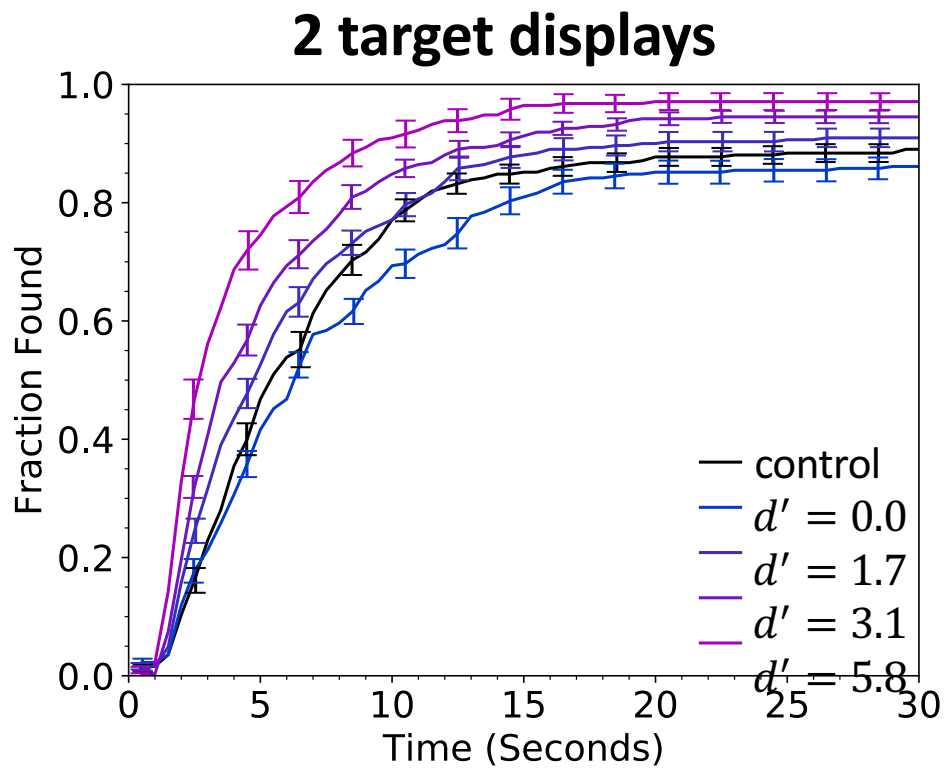
Experiment 3

- 0, 1, or 2 targets
- Four classifier strengths, $d' \in \{0.0, 1.7, 3.1, 5.8\}$
- Subject clicks on targets, terminates trial by pressing 'done' when they feel all have been found.

DONE

Experiment 3



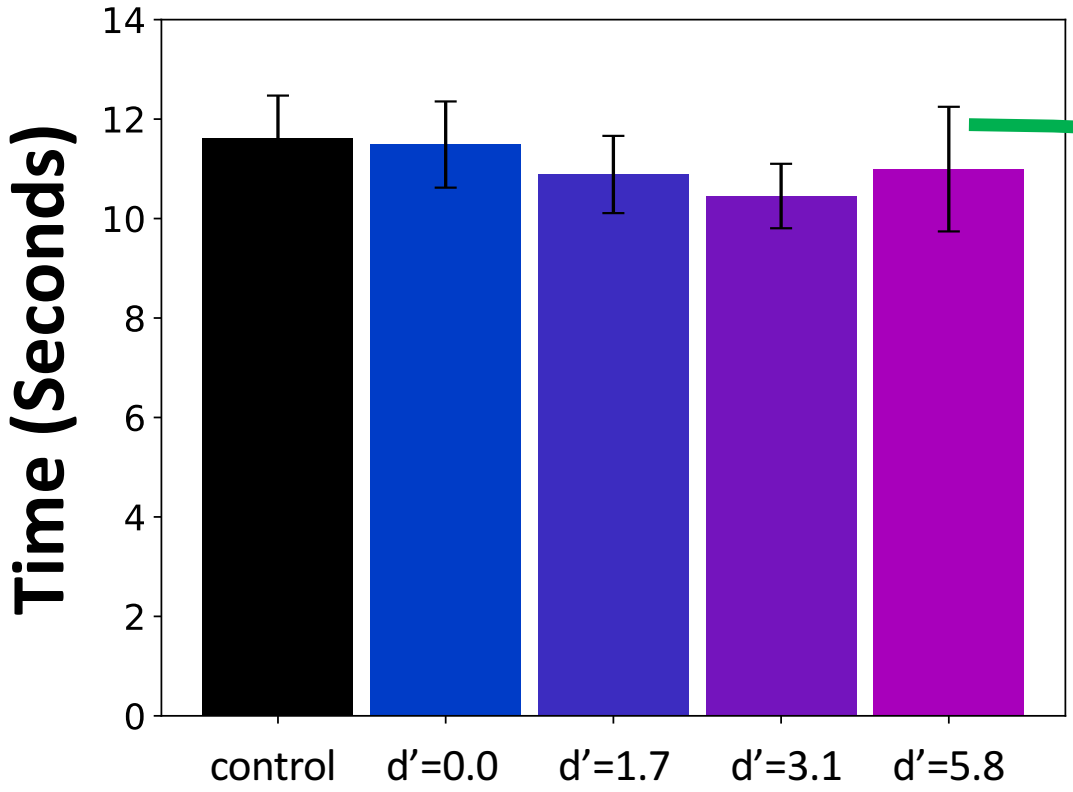


For displays with 1 or 2 targets,

- **highlighting speeds detection**
- **highlighting results in higher asymptotic detection rate**

Weakest classifier ($d' = 1.7$) is not effective, in contrast to previous experiments.

Experiment 3



1 SEM corrected for between subject variability (Masson & Loftus, 2003)

For displays with no targets, highlighting has no harmful effect on trial termination.

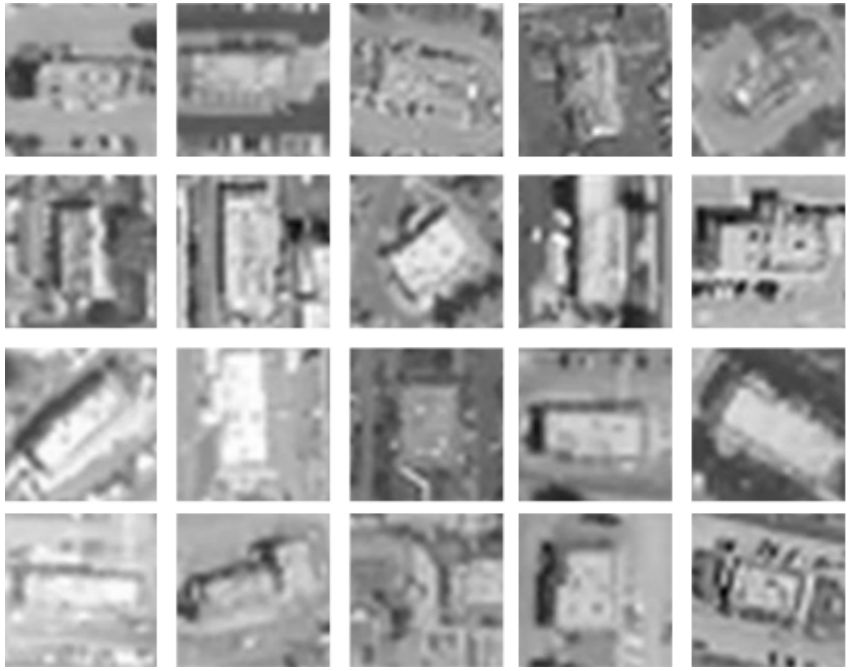
The Story So Far

- Even very noisy classifiers can boost human visual search
- Soft highlighting superior to hard highlighting
- No drawbacks found with soft highlighting

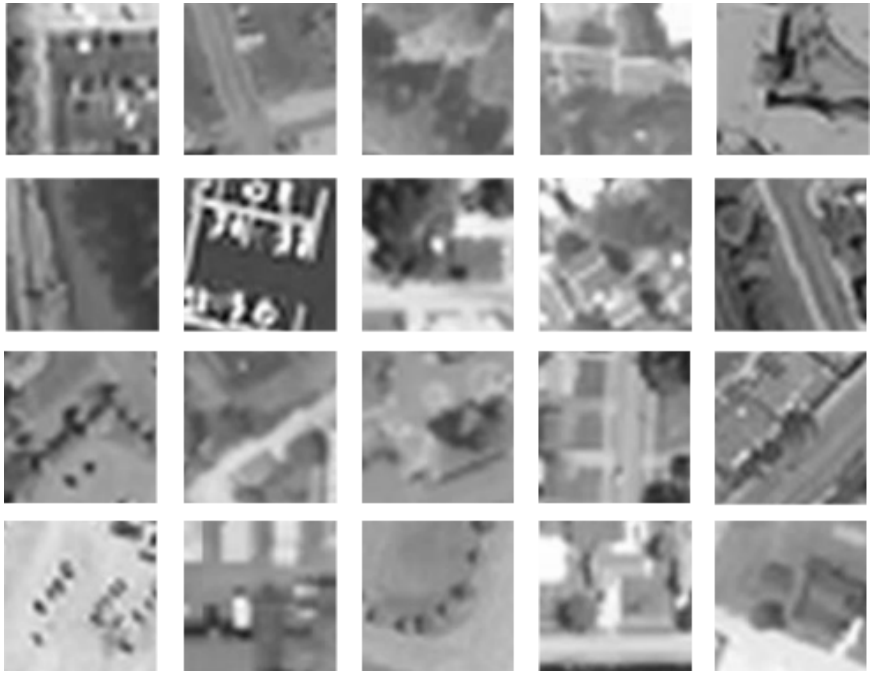
Satellite Imagery Search

Search for fast-food restaurant in suburban scene.

target-present patches



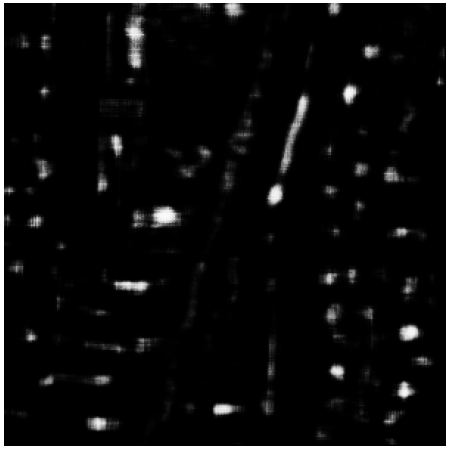
target-absent patches



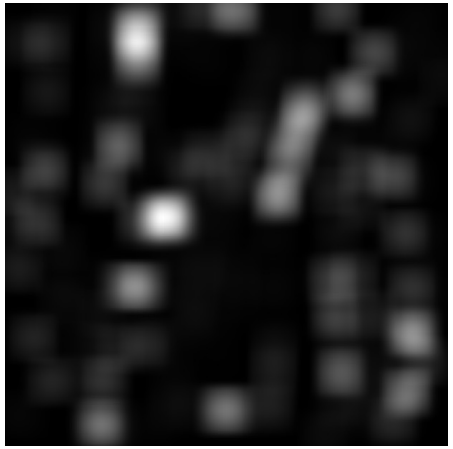
Use patches like these to train a state-of-the-art convolutional neural network classifier

Trained Classifier

original image



pixel-wise classification heat map



Gaussian convolution

saliency manipulation via saturation

Soft Highlighting

Hard Highlighting

Experiment 4

Task

- Click on location of restaurant

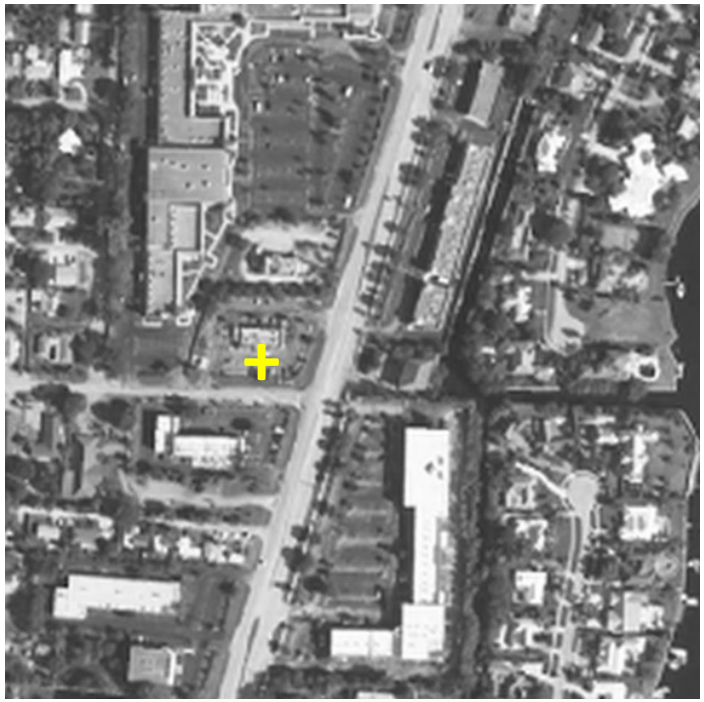
Stimulus set contains

- 2/3 single-target images
- 1/3 target-absent images

If clicked location is not target, buzzer sounds.

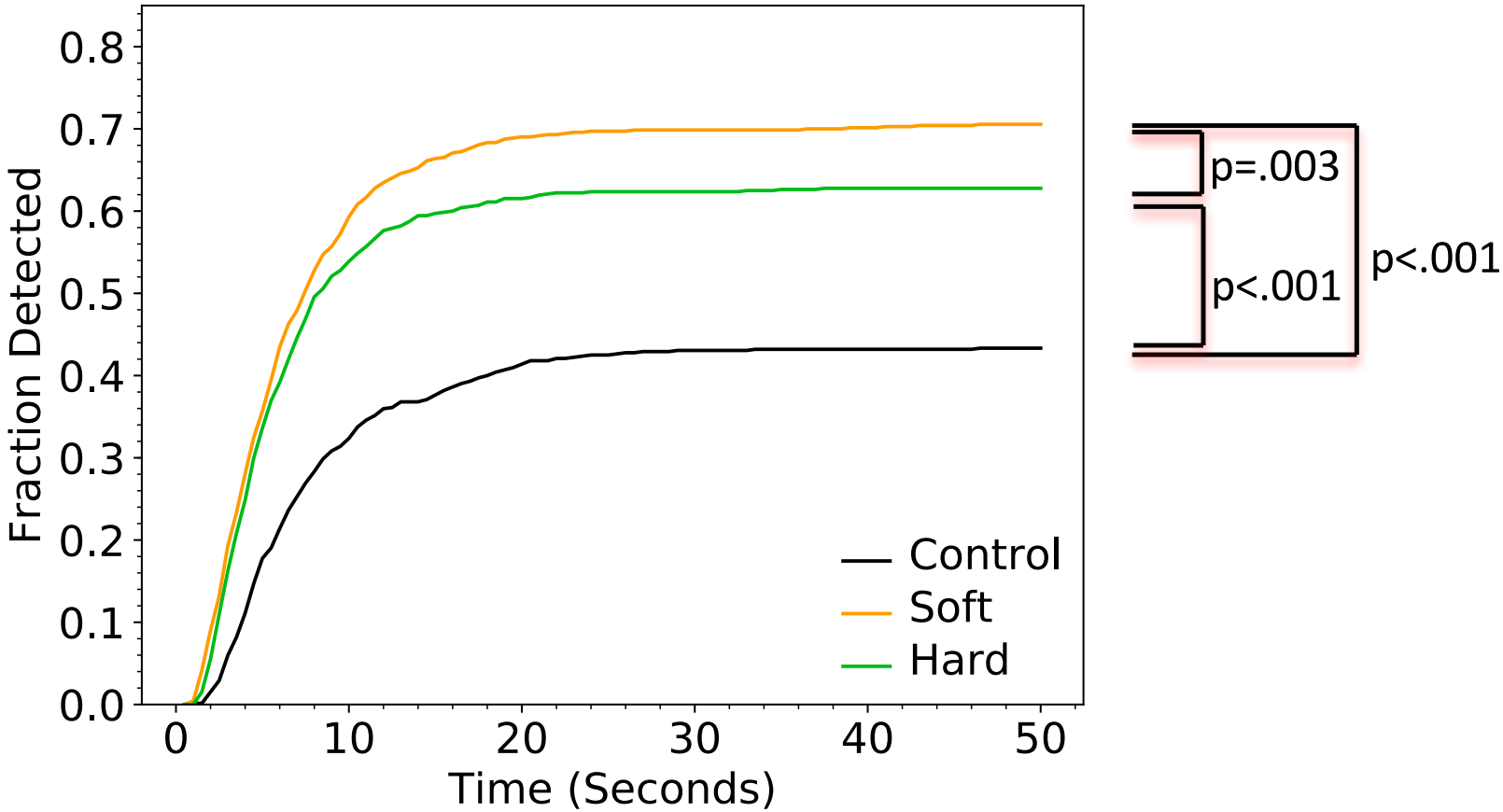
Trial continues until

- target is located
- 'No target present' is pressed



No Target Present

Experiment 4
Target-Present Trials



Detection accuracy higher with soft than hard

- likely because more locations receive some highlighting

Experiment 5

Task: Find target without feedback

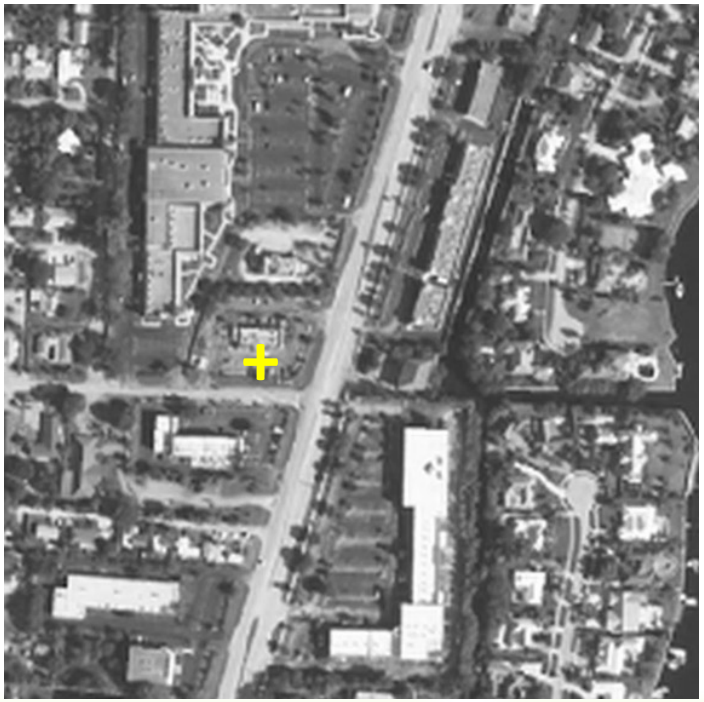
- click on candidate location
- can change selection
- or indicate no target is present
- click 'Next' to commit decision
- feedback provided after trial

Stimulus set contains

- 2/3 target present images
- 1/3 target-absent images

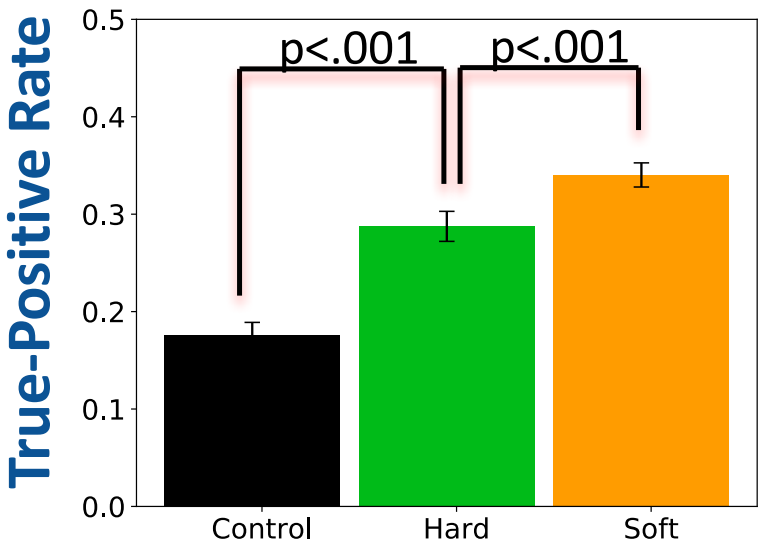
Mechanical turk study

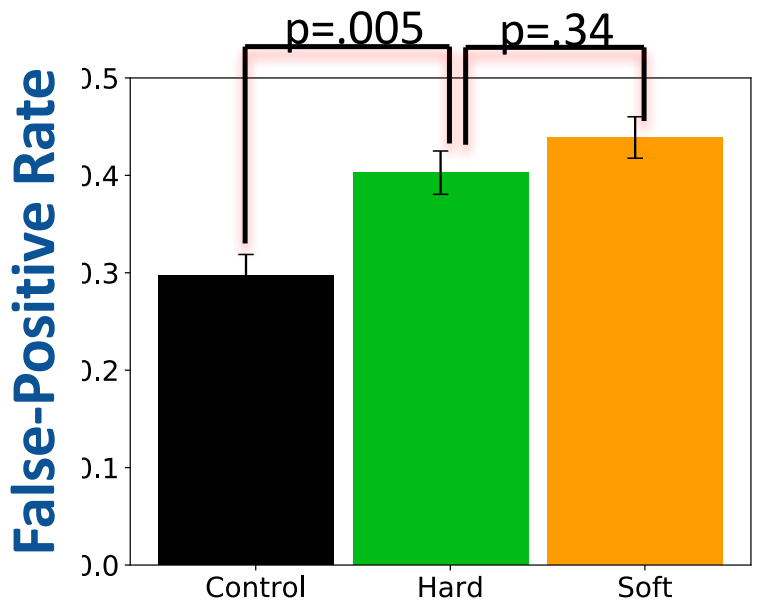
- n=90, 36 trials per subject



No Target Present | Next

Experiment 5





- Soft > hard > control
- Stronger result than Experiment 4, because no feedback is provided during trial.

- Because each image has many potential target locations but subjects can choose only one, we have to estimate a patchwise rate.
- Soft and hard can't be distinguished by statistical test

Experiment 5

d' Analysis

population	condition	d'
human	control	2.23
classifier		2.28
human + classifier	hard	2.64
human + classifier	soft	2.80

Cohen's d = 0.39

Summary

Even poor-quality classifiers can boost human visual search

- **Seems to succeed because visual system is good at integrating bottom-up saliency signals from highlights with top-down task-related guidance**

Soft highlighting seems superior to hard highlighting

- **Caveat: difference is modest in final experiment**

Topic 2

Can we develop methods of querying novices that allow them to classify like an expert?

Acta Derm Venereol. 2011 May ; 91(3): 279–283. doi:10.2340/00015555-1049.

Utility of Non-rule-based Visual Matching as a Strategy to Allow Novices to Achieve Skin Lesion Diagnosis

R. Benjamin ALDRIDGE[1], **Dominik GLODZIK**[2], **Lucia BALLERINI**[2], **Robert B. FISHER**[2], and **Jonathan L. REES**[1]

[1] Department of Dermatology, University of Edinburgh, Edinburgh, UK

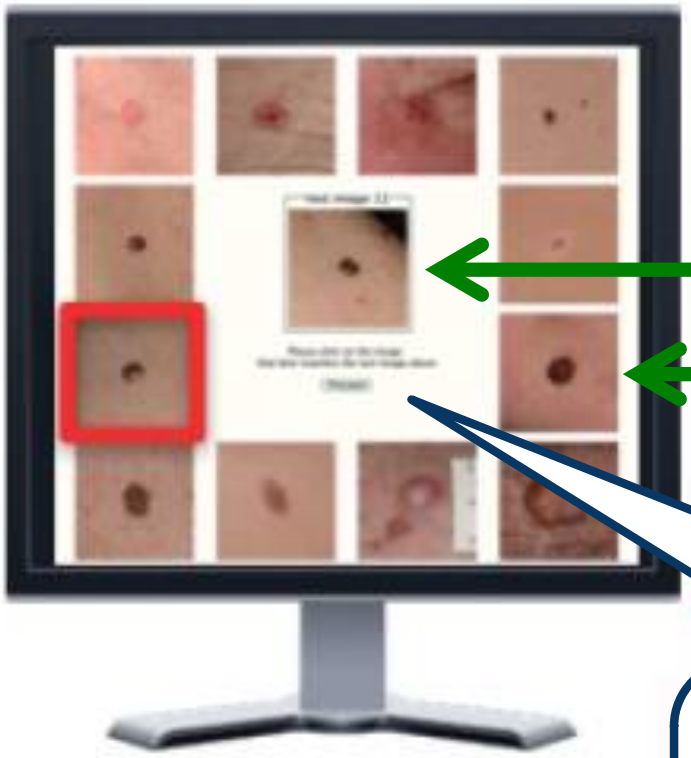
[2] School of Informatics, University of Edinburgh, Edinburgh, UK

Abstract

Non-analytical reasoning is thought to play a key role in dermatology diagnosis. Considering its potential importance, surprisingly little work has been done to research whether similar identification processes can be supported in non-experts. We describe here a prototype diagnostic

Could novices classify skin lesions as accurately as experts, given appropriate support?

Lesion Classification Procedure
(Aldridge et al., 2011)

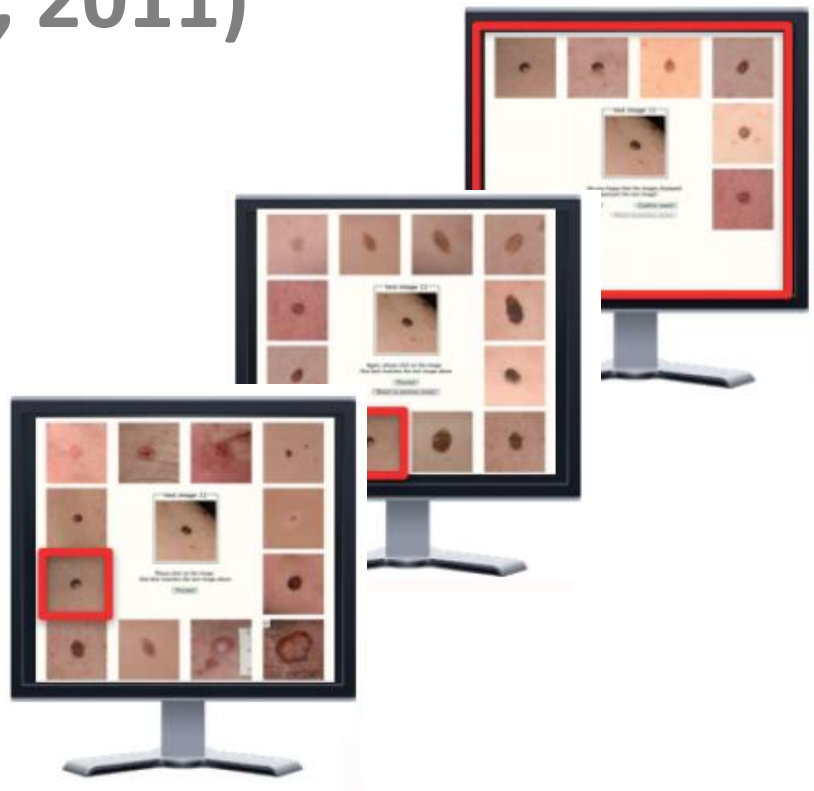


query image

reference image

Please select the reference image that best matches the query image.

Lesion Classification Procedure
(Aldridge et al., 2011)



Subjects perform a sequence of judgments in which they select the best-matching reference exemplar.

Reference exemplars have a known classification.

Similarity-based matching ➜ implicit classification of query

Result

- **Novices perform as well on a 5-way classification task as dermatology students who had completed 8 intro lectures and a 2-week clinical attachment**

Return on Investment
Reduce Diagnostic Error
Engage and Educate Patients
Current Therapy and Management
In The Workflow
Recognize Drug Reactions
Fulfills Meaningful Use

The benefits of using VisualDx

Faster, More Accurate Diagnosis

VisualDx is a diagnostic clinical decision support system that leverages the innate human ability to recognize visual patterns to assist healthcare providers in making faster, more accurate diagnoses. VisualDx combines high-quality, peer-reviewed medical images and concise, actionable information to support today's busy physicians in the accurate recognition and management of disease. Health care professionals can input visual clues, symptoms, and patient history to help make the correct diagnosis and avoid costly and dangerous errors at the point of care.

Reduce Diagnostic Errors and Improve Patient Outcomes

10% to 20% of all diagnoses are inaccurate[1] and may result in patient dissatisfaction, harm, or serious injury including death. Physicians and patients know that proper patient management and therapy depends on diagnostic accuracy. Moreover, diagnostic error often involves common clinical scenarios and is not limited to rare diagnoses or unusual presentations. VisualDx is proven to reduce diagnostic error. Read more about how images improve diagnostic accuracy.

Recognize Drug Reactions

Diagnose medication reactions early and accurately with VisualDx. The skin and mucosae are often the first sites of an adverse reaction, including life-threatening disease reactions. An essential safety tool for recognizing drug-induced conditions, VisualDx links to literature evidence documenting drug-disease associations. Learn more about diagnosing adverse medication reactions.

Reduce Risk for Hospitals

Hospital physicians may have only minutes to assess and diagnose their patients – often with incomplete data. Read how VisualDx helps hospitals mitigate risk.

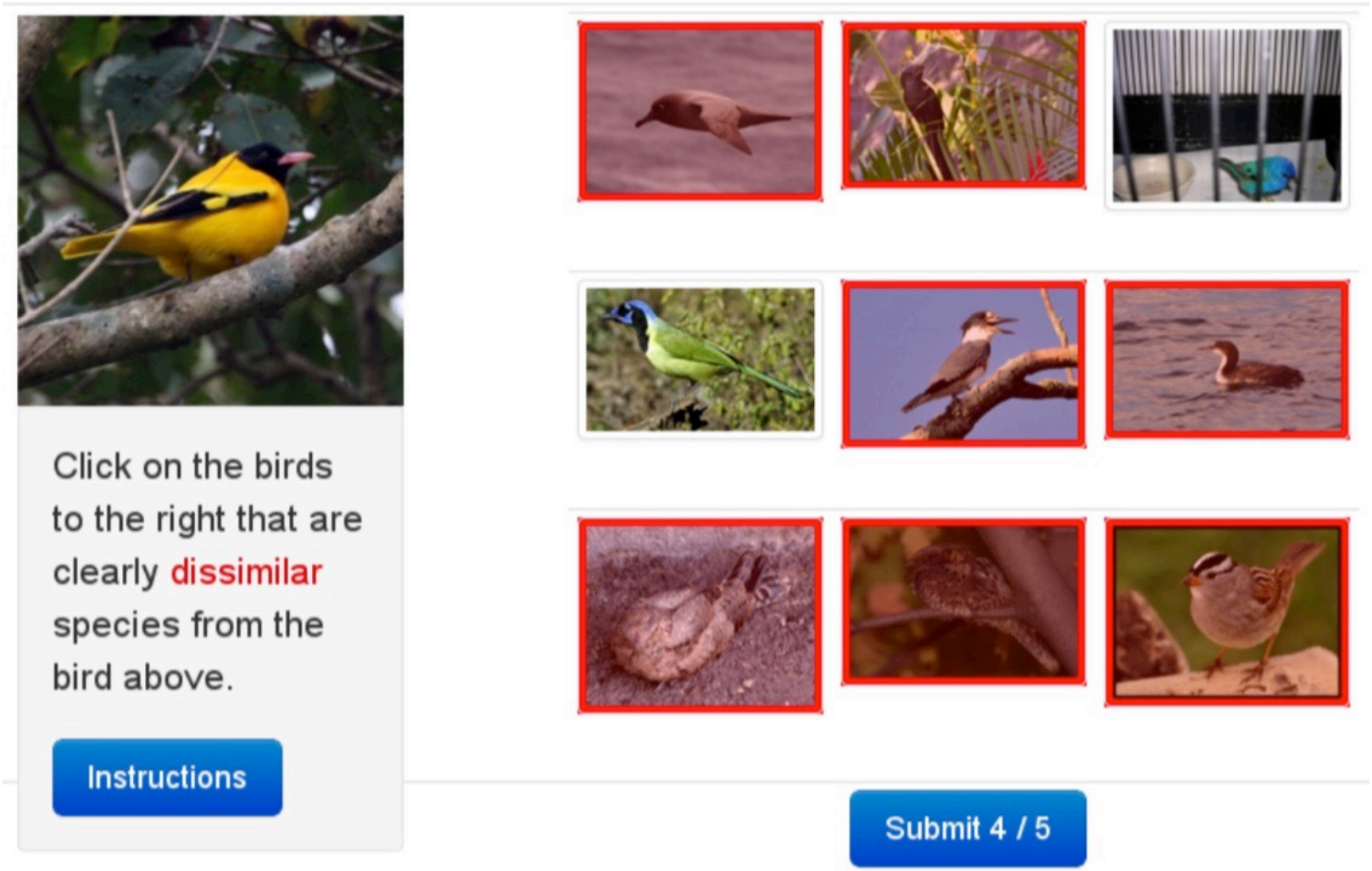
Empower Primary Care Physicians

Primary care physicians are under tremendous pressure to evaluate more patients in less time. Learn how VisualDx empowers the primary care physician.

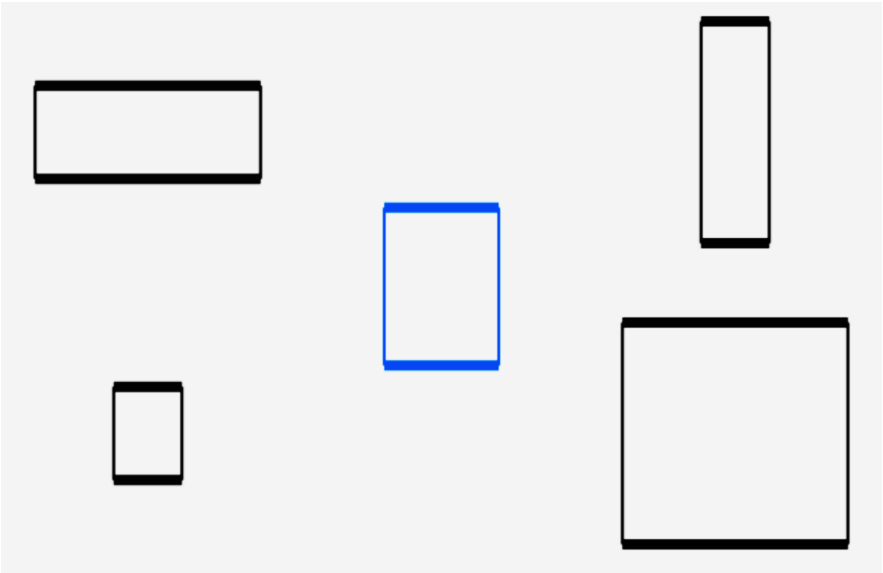
Watch the Larry Weed 1971 Grand Rounds Video

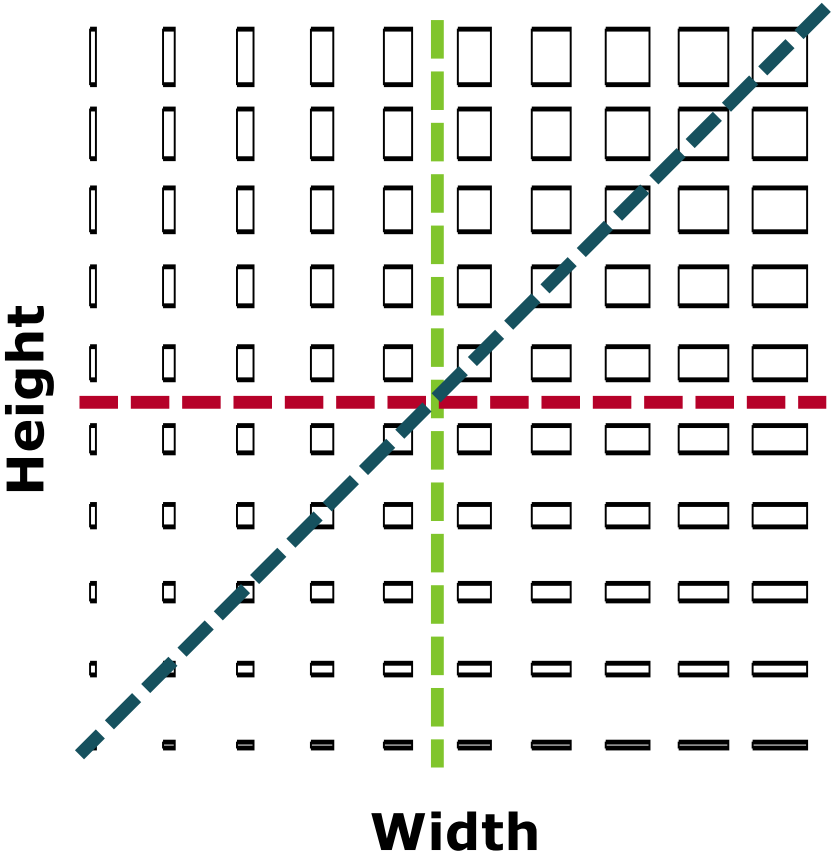
Similarity Comparisons for Interactive Fine-Grained Categorization

Catherine Wah[†] Grant Van Horn[†] Steve Branson[‡] Subhransu Maji[*] Pietro Perona[‡] Serge Belongie[♯]
[†]UC-San Diego [‡]Caltech [*]TTI at Chicago [♯]Cornell Tech
vision.ucsd.edu vision.caltech.edu ttic.edu tech.cornell.edu



Rectangle Experiment





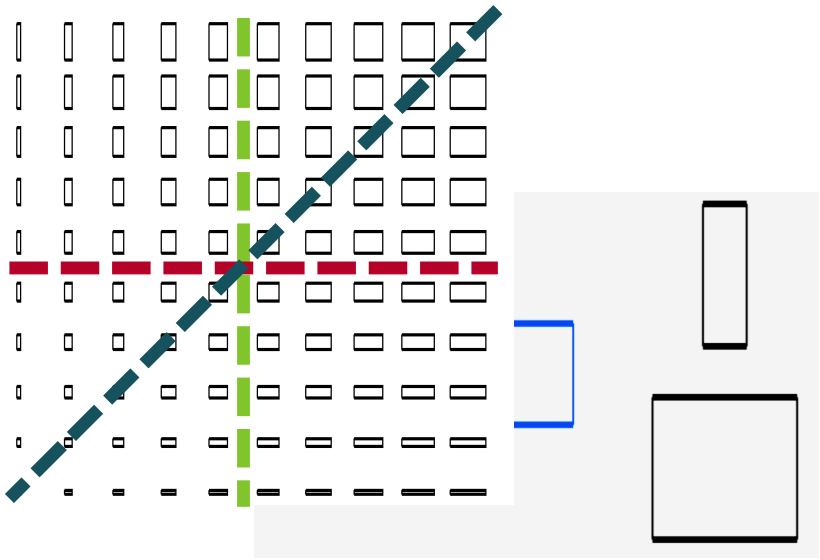
Experiment Setup

Three binary classification tasks

- intermixed trial to trial

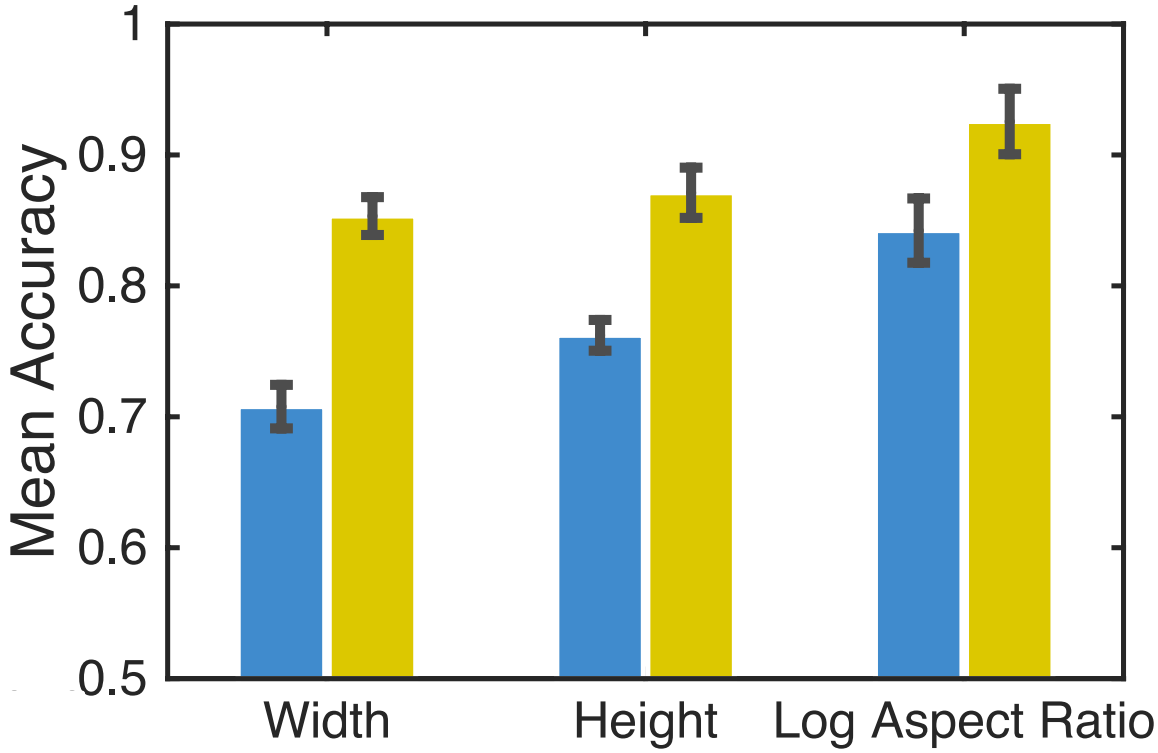
On each trial, the query and referen rectangles are chosen at random

- constrained to have 2 references of each category

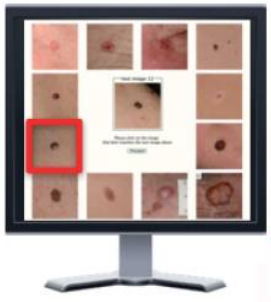


Subject instructed to find the reference that best matches the query

Classification scored as correct if the chosen reference has the same category as the query.


Mean Accuracy
1
0.9
0.8
0.7
0.6
0.5
Width
Height
Log Aspect Ratio

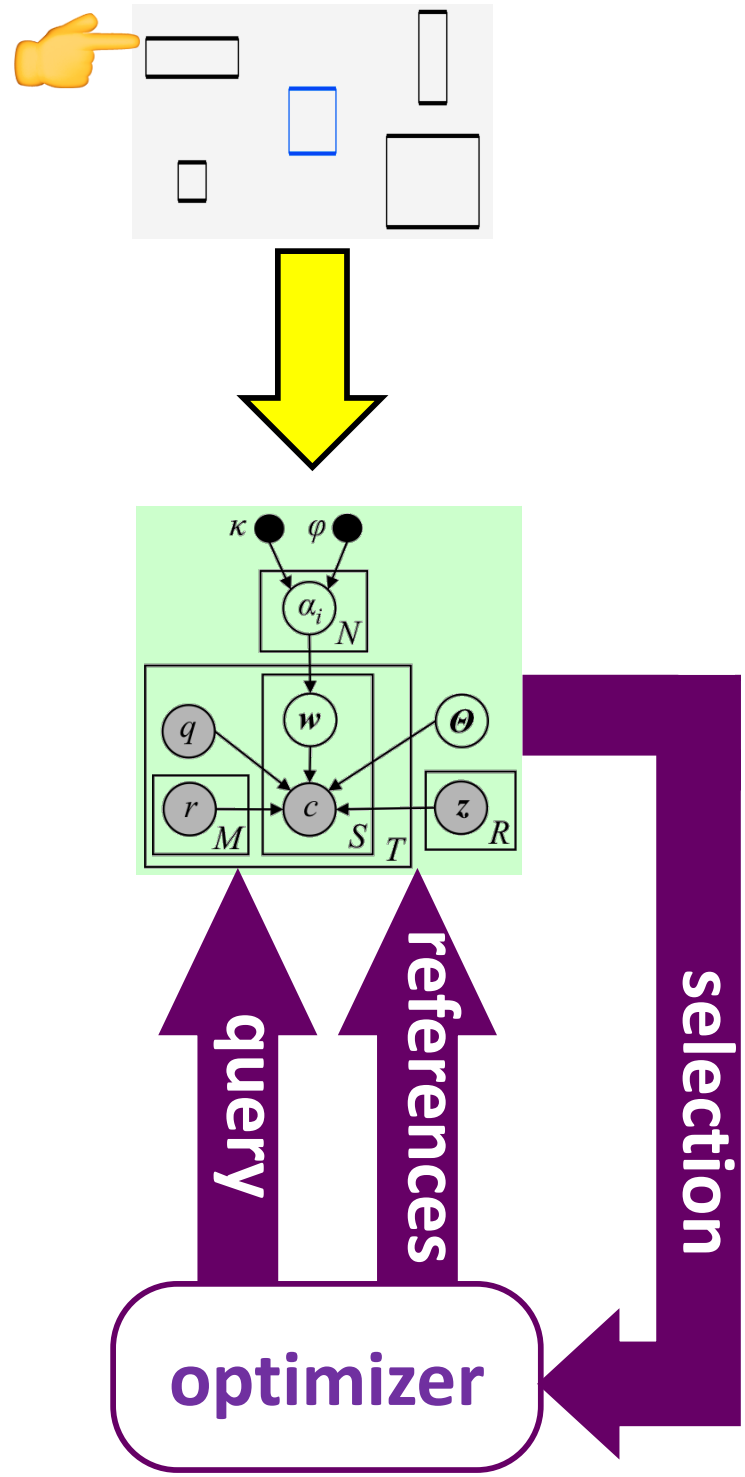
Aldridge et al. (2011) used intuition to choose reference exemplars.



Can we determine the *optimal* set of references for a given categorization task?

- I.e., set that will maximize implicit-classification accuracy regardless of query image

Strategy



Propose computational model of similarity-based matching

Use experimental data to fit model parameters

Plug model into an automatic procedure to optimize the set of references for each classification task

- Optimal = maximizes probability that subject's choice will correctly categorize any query image

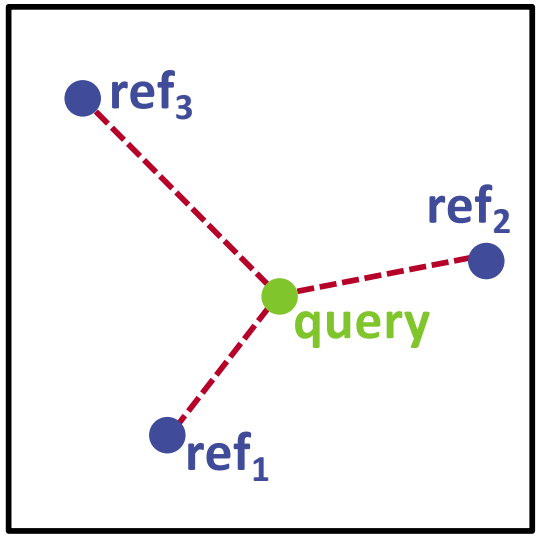
Theory Of Similarity Judgment

Builds on rich literature

- Nosofsky (1986)
- Shepard (1987)
- Tenenbaum (1999)
- Jones, Maddox, and Love (2005, 2006)
- Jones, Love, and Maddox (2006)
- Sinha and Russell (2011)

Theory Of Similarity Judgment

1. Similarity is based on an internal representation (*psychological embedding*)



2. Distance is computed between query and each reference

$$d(z_q, z_r) = \left(\sum_{i=1}^{N} w_i \left| z_{qi} - z_{ri} \right|^{\rho} \right)^{1/\rho}$$

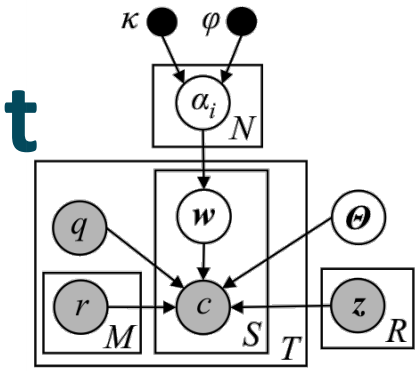
3. Similarity is based on distance

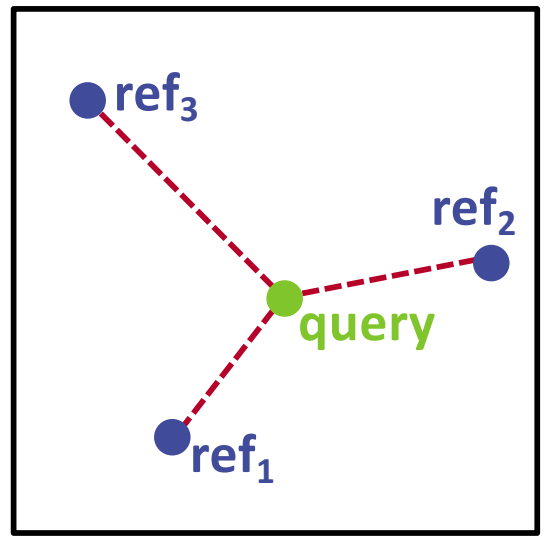
$$g(d) = \gamma + \exp(-\beta d^{\tau})$$

4. Probability of selecting a reference is proportional to similarity

$$P(r|q) \propto g(d(z_q, z_r))$$

Theory Of Similarity Judgment





- **Free parameters**
- **Psychological dimensions**

  {height, width}
  {log height, log width}
  {area, aspect ratio}
  {log area, log aspect ratio}

$$d(z_q, z_r) = \left(\sum_{i=1}^{N} w_i \left| z_{qi} - z_{ri} \right|^{\rho} \right)^{1/\rho}$$

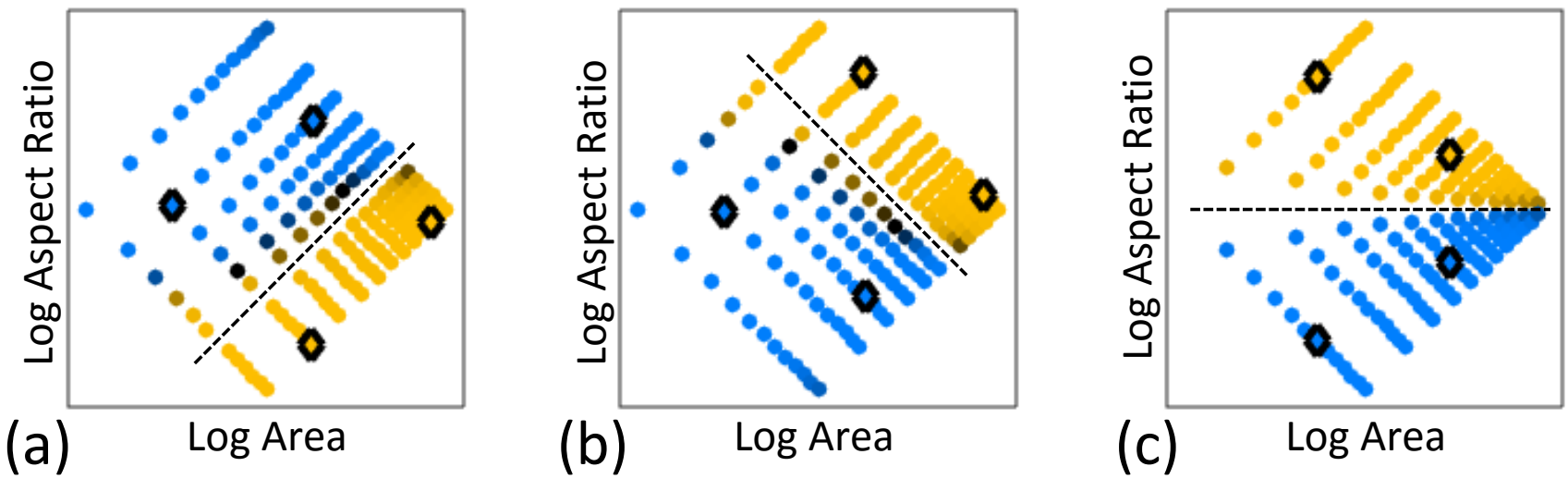
- **Attentional weights**

  Assumption: weights fluctuate randomly due to context effects (Tversky, 1977)

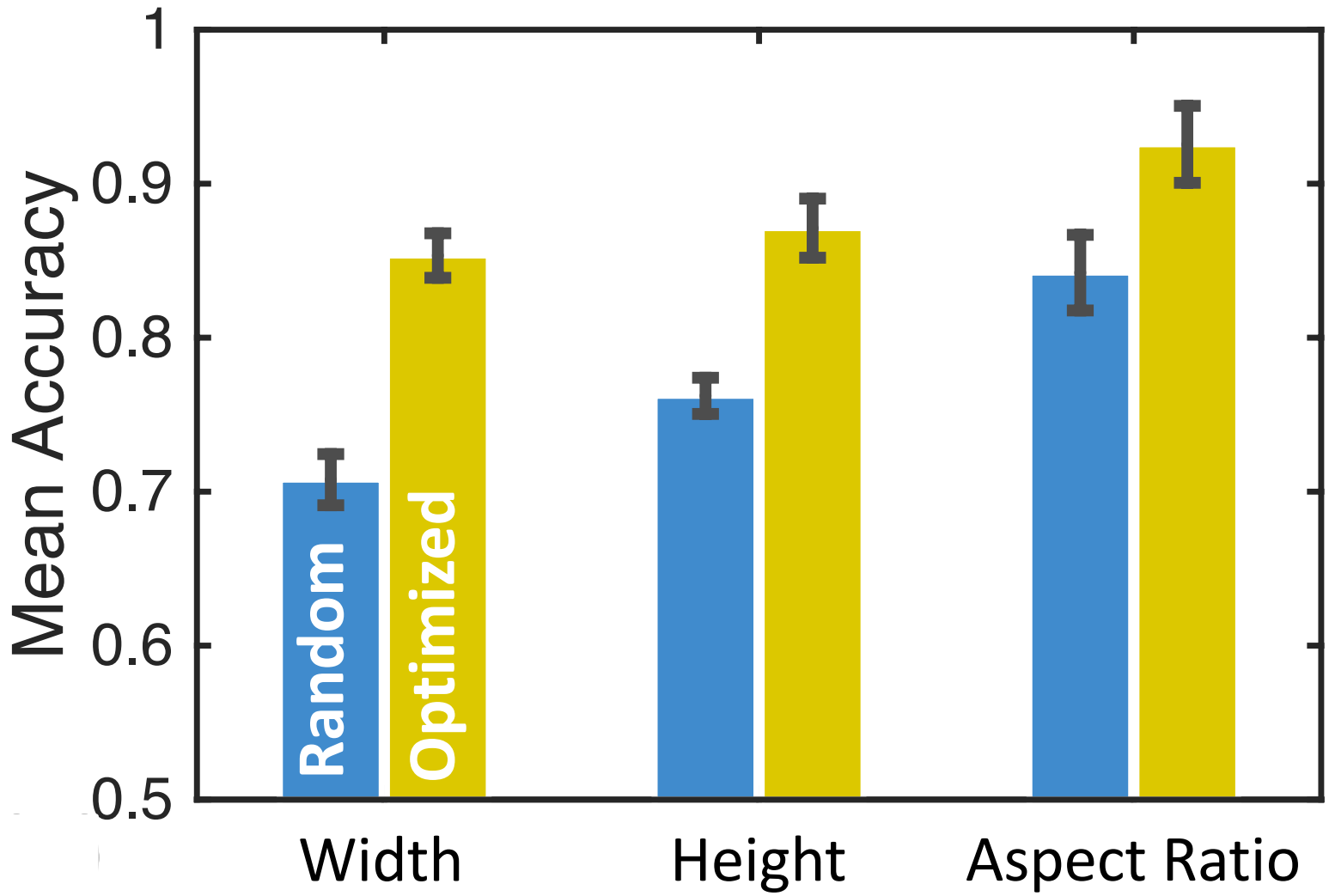
$$g(d) = \gamma + \exp(-\beta d^{\tau})$$

$$P(r|q) \propto g(d(z_q, z_r))$$

Optimized References



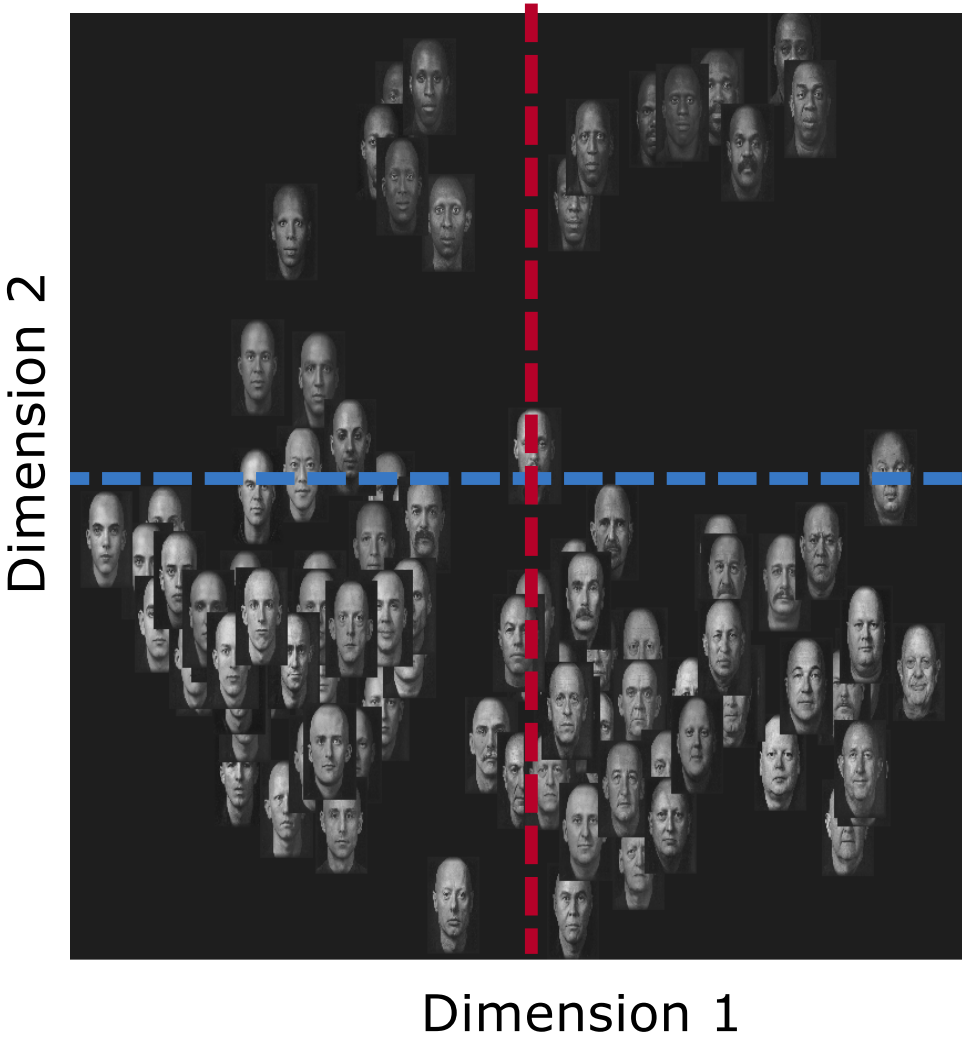
Results

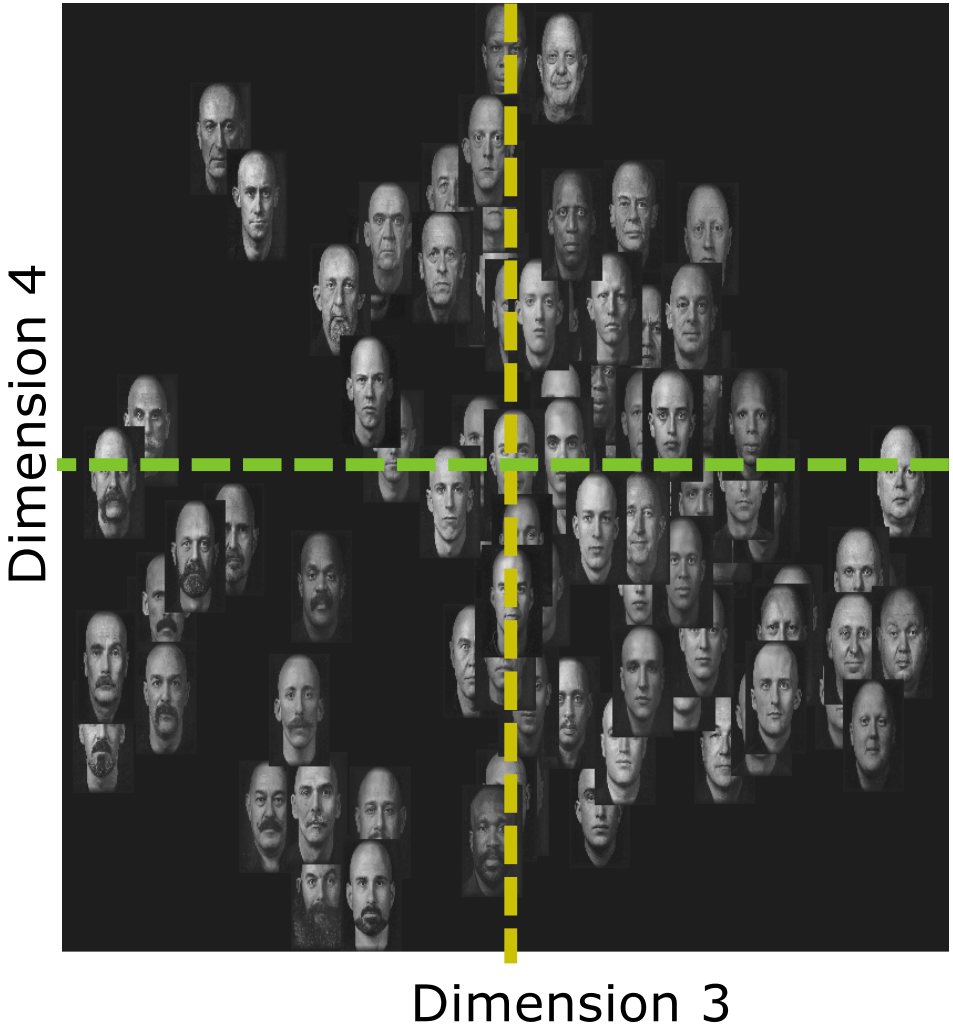

Mean Accuracy
1
0.9
0.8
0.7
0.6
0.5
Random
Optimized
Width
Height
Aspect Ratio

Face Experiment

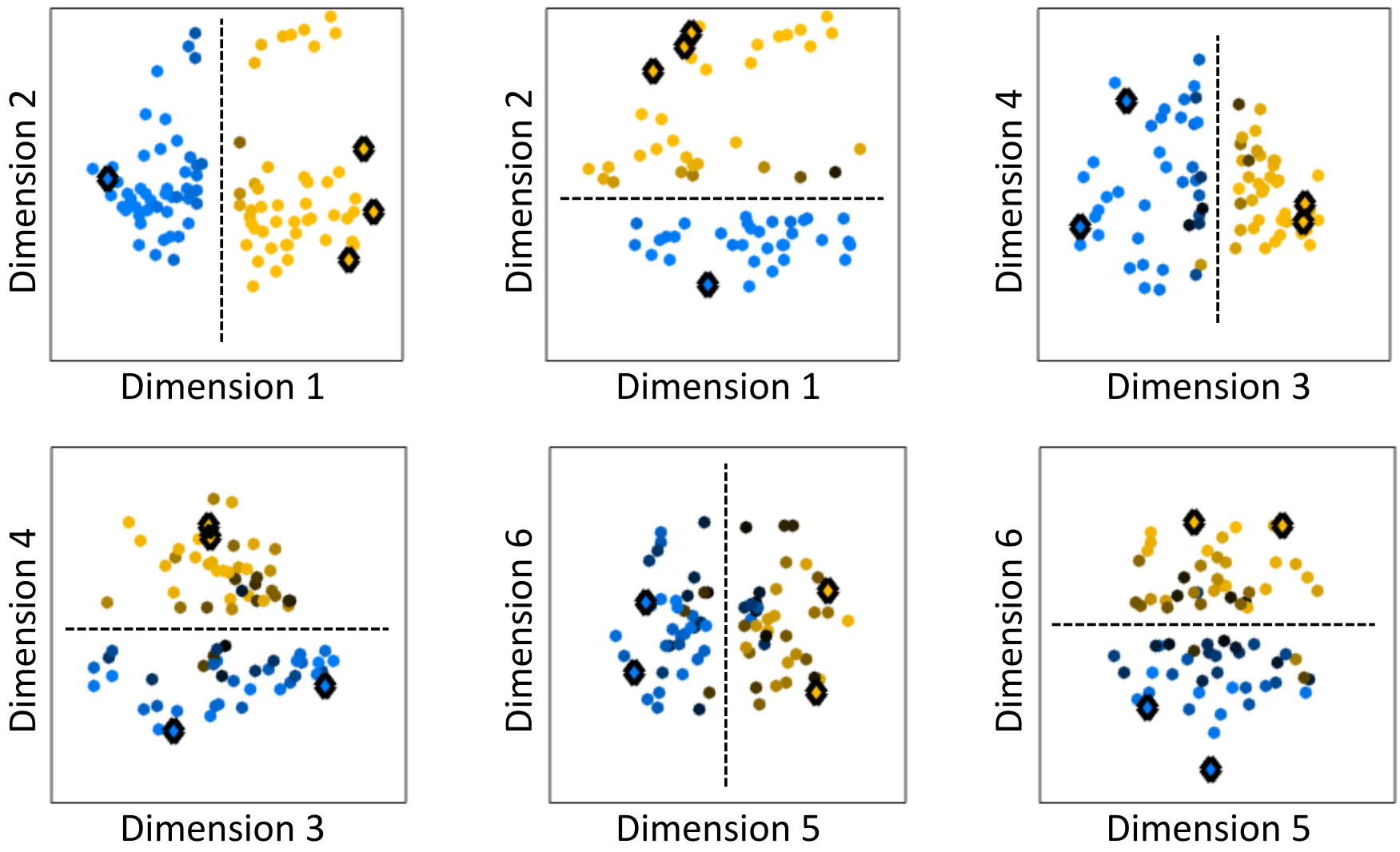
104 grayscale bald male faces

Six-dimensional psychological embedding
(Jones and Goldstone, 2013)

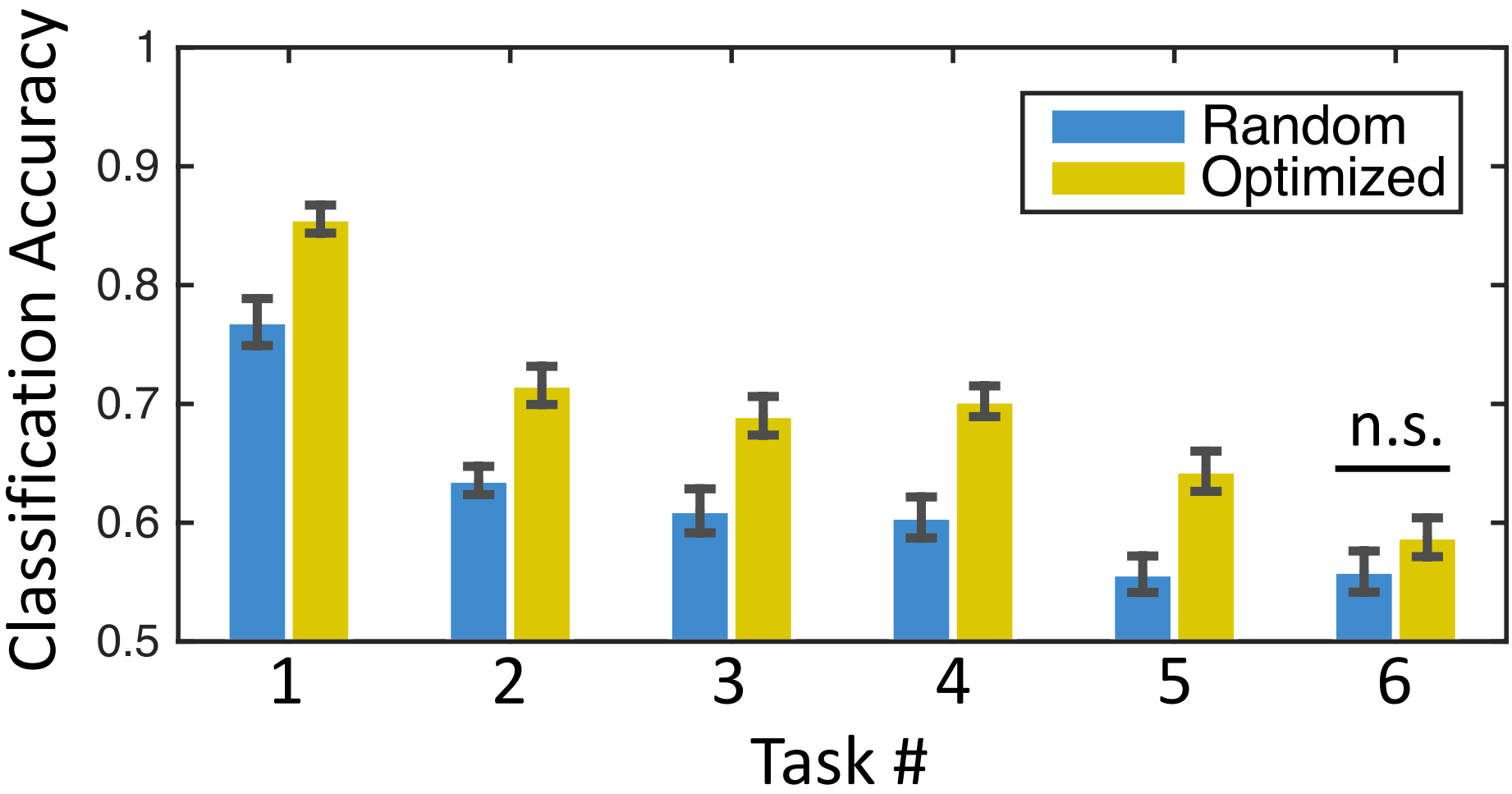

Dimension 2
Dimension 1


Dimension 4
Dimension 3

Optimized References For Face Tasks

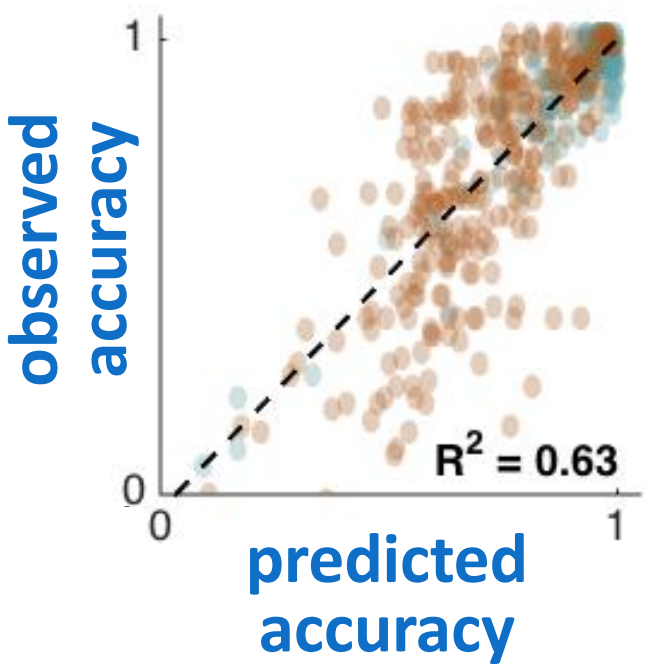


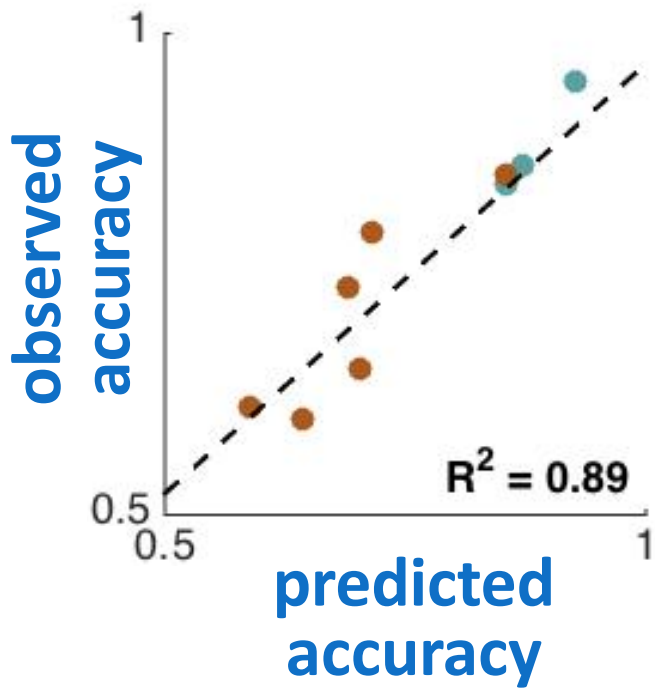
Categorization Accuracy On Six Face Tasks



Once we fit model parameters to judgments collected with random references…

We can make fully-constrained predictions of subject choices for the optimized references.





Contributions

We developed a predictive model of similarity-based choice.

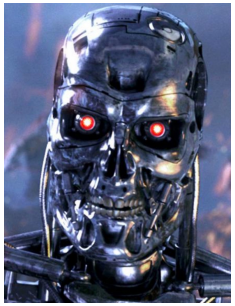
We use the model as a proxy for human subjects and optimize a set of reference exemplars for a given implicit categorization task.

We show that untrained novices can perform implicit classification pretty well via similarity judgments.

Questions we're interested in:

- **Can we scale up this approach to realistic problems?**
- **Can this paradigm be improved with a bit of instruction?**

Human-ML Interaction



ML approaches require an understanding of psychological mechanisms

- e.g., attentional salience
- e.g., utility functions (time vs. accuracy)
- e.g., naïve representations of domain images
- e.g., similarity judgment