Models of Salience and Eye Movements

Nicholas Butko,
Javier Movellan, Garrison Cottrell,
Terry Sejnowski, Matt Tong,
Chris Kanan, Lingyun Zhang,
Leanne Chukoskie, Mike Mozer,
Mike Arnold

Mathematical Formulation of Information

- "Information" has two (related) mathematical meanings:
  - 1) How much did you expect something you experience ("it is going to rain")?

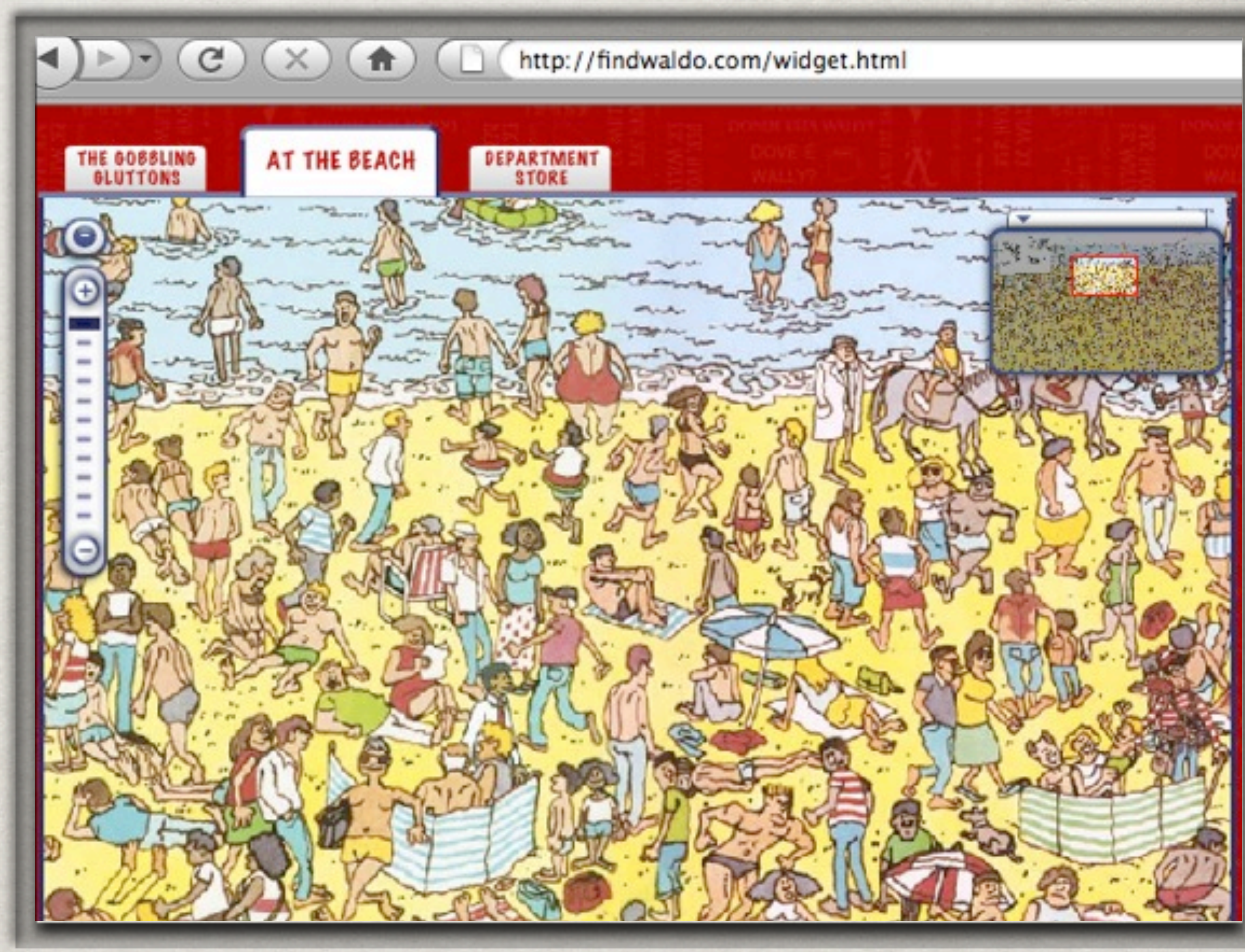
$$-\log p(x)$$

  - 2) How unsure are you about some aspect of the world (e.g. "is it going to rain?")?

$$-\sum_{Possibilities} p(x) \log p(x)$$

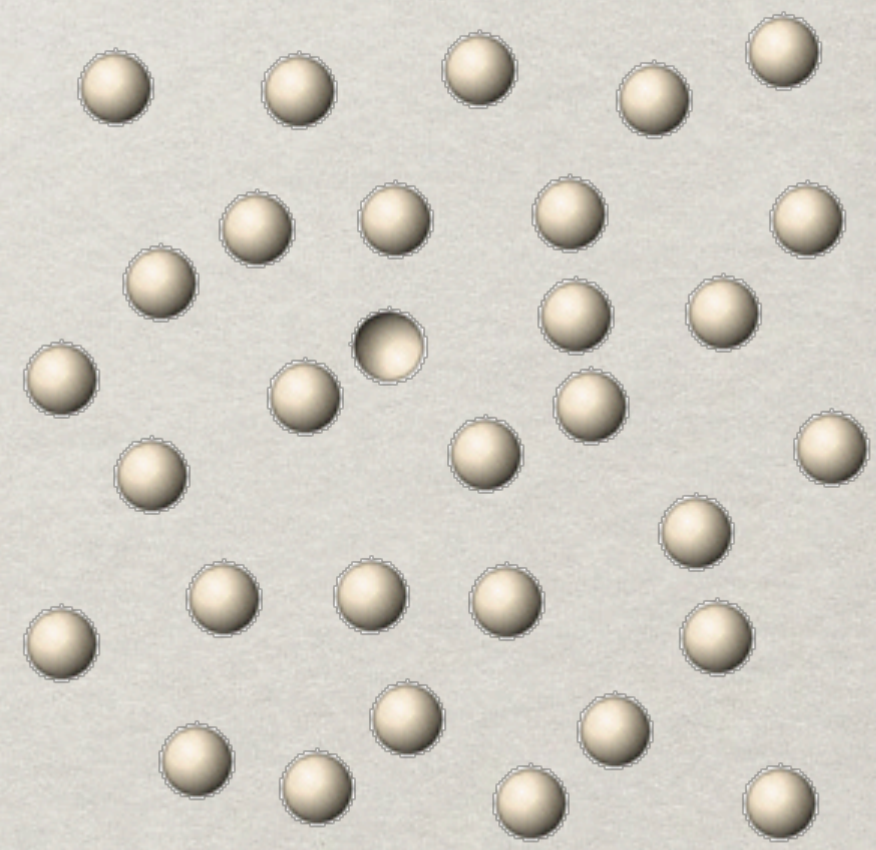
How Humans Search Scenes

- People don't closely examine every inch of the world.
- Eye-movements are tuned to optimally gather information (Sense 1 and Sense 2).
- These different notions of information led to very different models:
  - 1)Visual Saliency
  - 2) Digital Retina



Visual Saliency

- Salient objects "pop out" of visual scenes.
  - Simple preprocessing step directs computational resources.
  - Rare (improbable) image features are more salient than common (probable ones)
  - Improbable events carry more *information* (Sense 1).



- We developed an efficient way to model the statistics of a video stream, and analyze it for salient "pop out".

*Zhang, Cottrell

A Promising Framework

- A common framework is shared by several authors.
- Claim: The **goal** of eye-movements is to **find visual targets**.
- Approach: Attend to regions x of the visual plane which **contain visual targets with high probability**.
  - In open-ended tasks, drop class-specific terms.

$$
\begin{aligned}
\text{Salience}(x) &= \log[p(C_x = 1 | ImgFeats_x)] \\
&= \cancel{\log[p(ImgFeats_x | C_x = 1)]} \cancel{+ \log[p(C_x = 1)]} - \log[p(ImgFeats_x)]
\end{aligned}
$$

Estimate using current image histogram (Torralba et al.)

Estimate using local region histogram (Bruce & Tsotsos)

Estimate using natural image histograms (Zhang et al.)*

*Best suited to real-time implementation

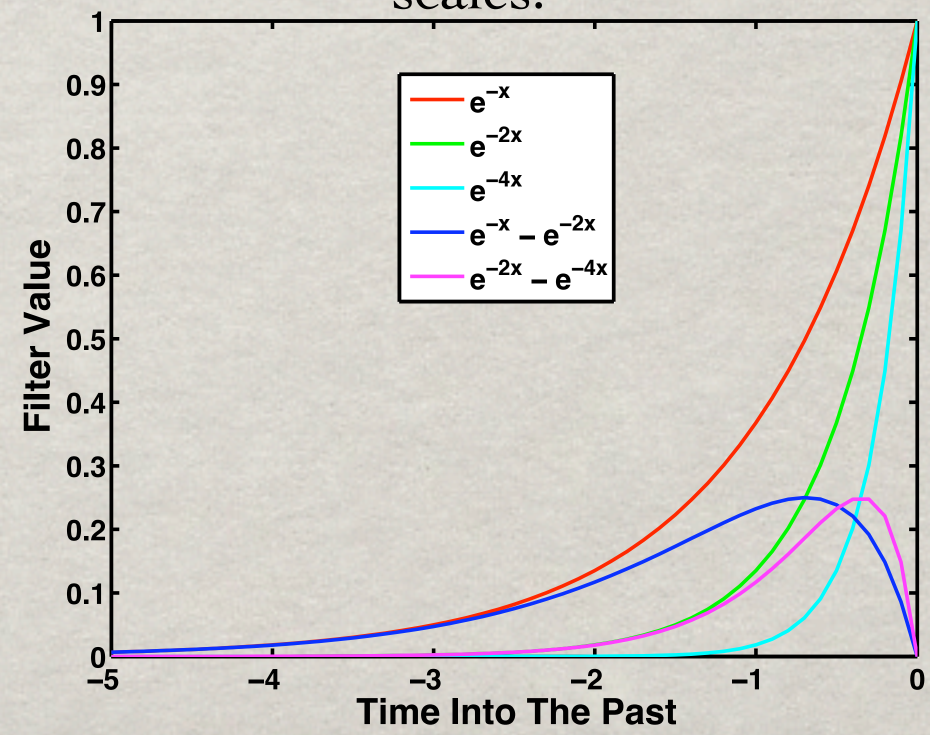
Zhang's SUN Model

- Zhang et al. created the "Saliency Using Natural-statistics" model of visual saliency.

$$\text{Salience}(x) = -\log[p(ImgFeats_x)]$$

Space: Difference Of Gaussians filters at increasing spatial scales.

Time: Difference Of Exponentials filters at increasing temporal scales.



Probability: Generalized Gaussian

$$p(ImgFeats_x) \quad = \quad \prod_i C_i \exp(|ImgFeats_x^i/\sigma_i|^{\theta_i})$$

*Parameters σ and θ are estimated from Image Features in natural images.

SUN Algorithm Sketch

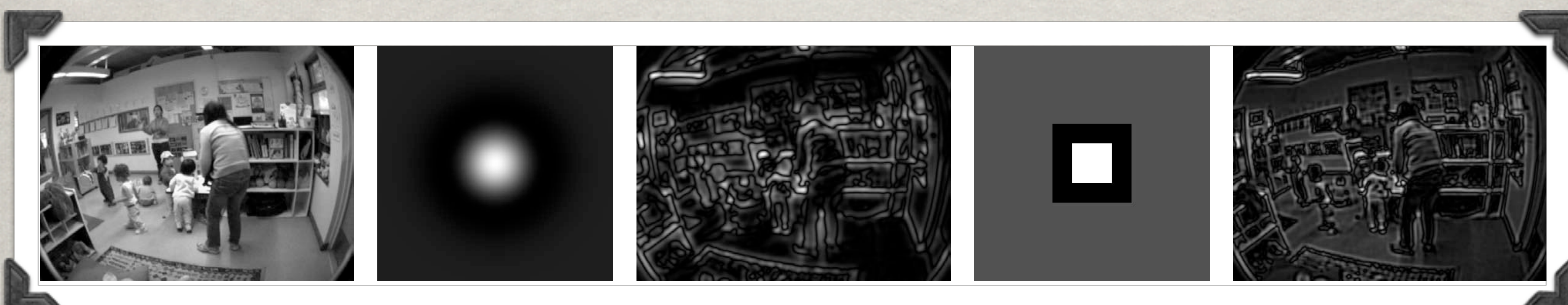
1. Grab a new video frame
2. Filter frame with N Difference-of-Gaussian filters at increasing spatial scales.
3. Integrate each DoG filter with M+1 previous Exponential filters at increasing time-scales.
   $[\tau_j/(1+\tau_j)\ DoG_k + 1/(1+\tau_j)\ OldExponential_{kj}]$
4. Compute NM Difference-of-Exponential temporal filters.
5. Compute -log $p(DoE)$ for each pixel x of each DoE filter i:
   -log $p(DoE_i) = |DoE_i / \sigma_i|^{\theta_i}$; for θ_i and σ_i fit to spatiotemporal scale in natural images.
6. Sum all NM -log $p(DoE)$ to get salience for each pixel x.

Efficient Approximation

- Our goal is a much-faster-than-real-time algorithm.
- We achieve this with two approximations.

1. Instead of Difference-of-Gaussian spatial filters, we use Difference-of-Box Haar-style filters. [Efficient convolution]

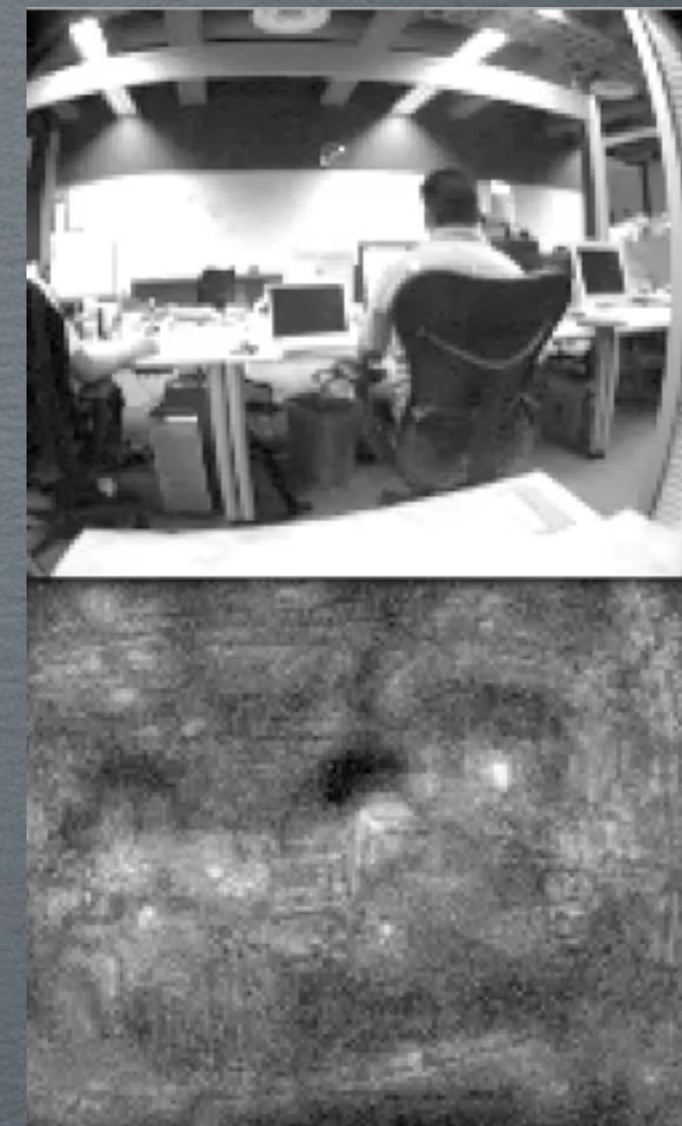
2. Instead of Generalized Gaussian probability model, we use Laplacian probability model. [Efficient inference]



*Butko, Zhang, Cottrell, Movellan

Two Examples

Online: Camera Control

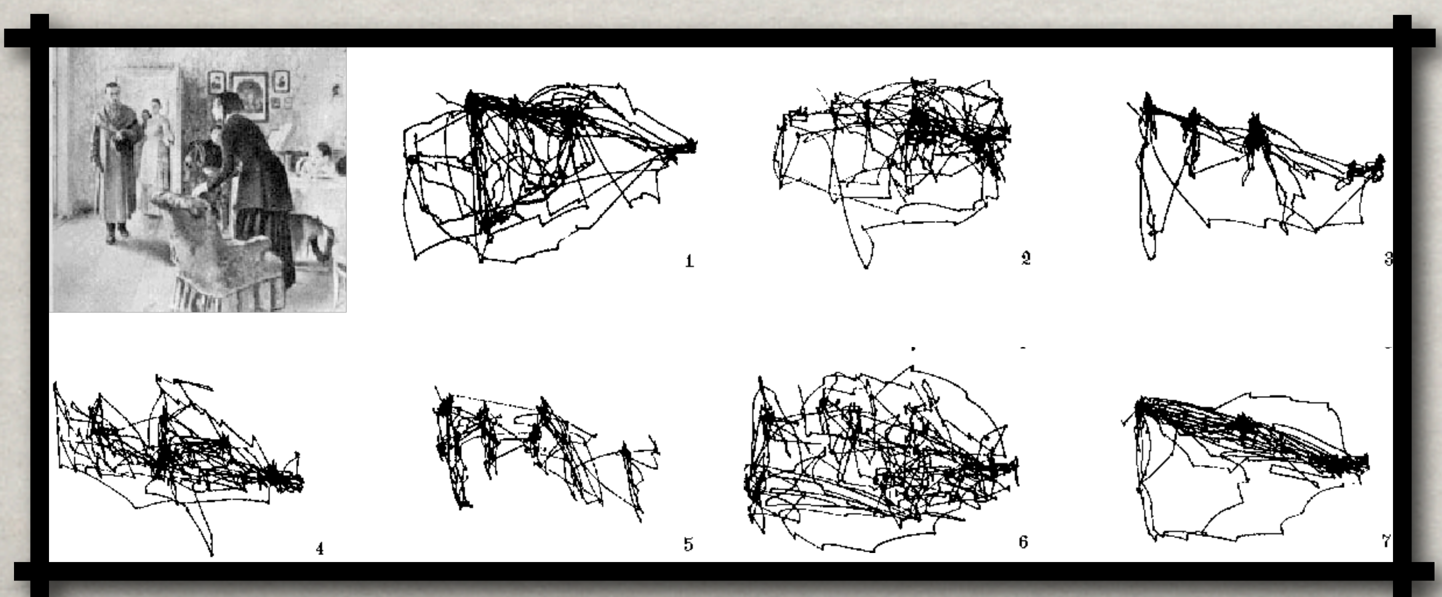


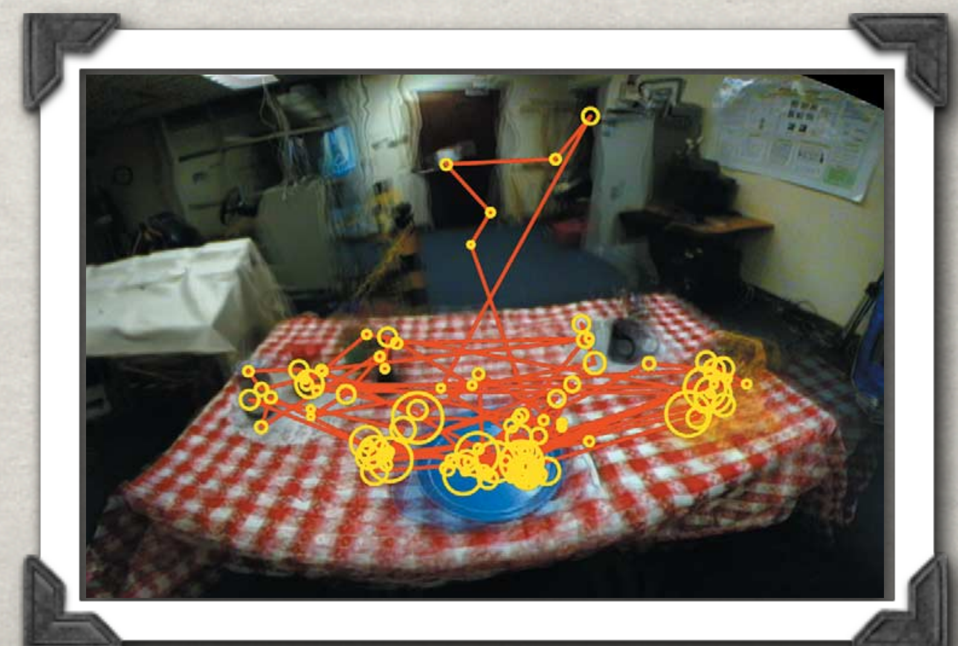
Offline: Video Analysis

“Pop Out” Helps Track People

Task Directed Looking Behavior

- Visual Popout can be useful for robots, and it seems to be important in people, but it can't account for task-specific looking behavior.
- It has long been known that where people look depends on what information they are trying to gather [Yarbus 1967]
- Current studies have difficulty making quantitative claims: "Fixations are tightly linked in time to the evolution task. Very few irrelevant regions are fixated." [Hayhoe & Ballard 2005]





A Promising Framework

- A common framework is shared by several authors.
- Claim: The **goal** of eye-movements is to **find visual targets**.
- Approach: Attend to regions x of the visual plane which **contain visual targets with high probability**.
  - In open-ended tasks, drop class-specific terms.

$$
\begin{aligned}
\text{Salience}(x) &= \log[p(C_x = 1 | ImgFeats_x)] \\
&= \cancel{\log[p(ImgFeats_x | C_x = 1)]} + \cancel{\log[p(C_x = 1)]} - \log[p(ImgFeats_x)]
\end{aligned}
$$

Object Appearance Information

Location Prior

Image Channel Information

"Mutual Information" between object presence and image features.

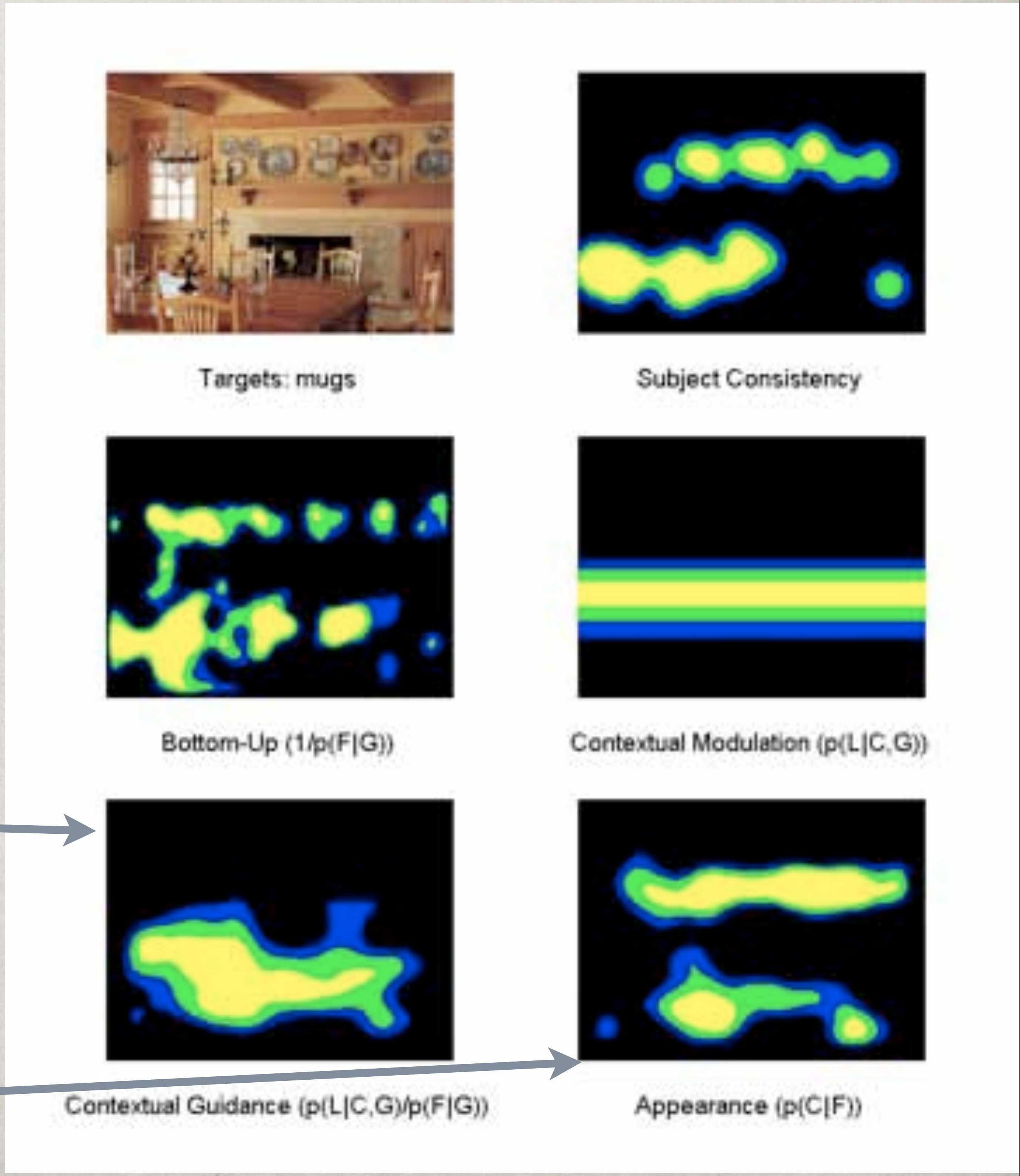
***Tong, Kanan, Cottrell**

Qualitative Results (mug search)

- Where we disagree the most with Torralba et al. (2006)

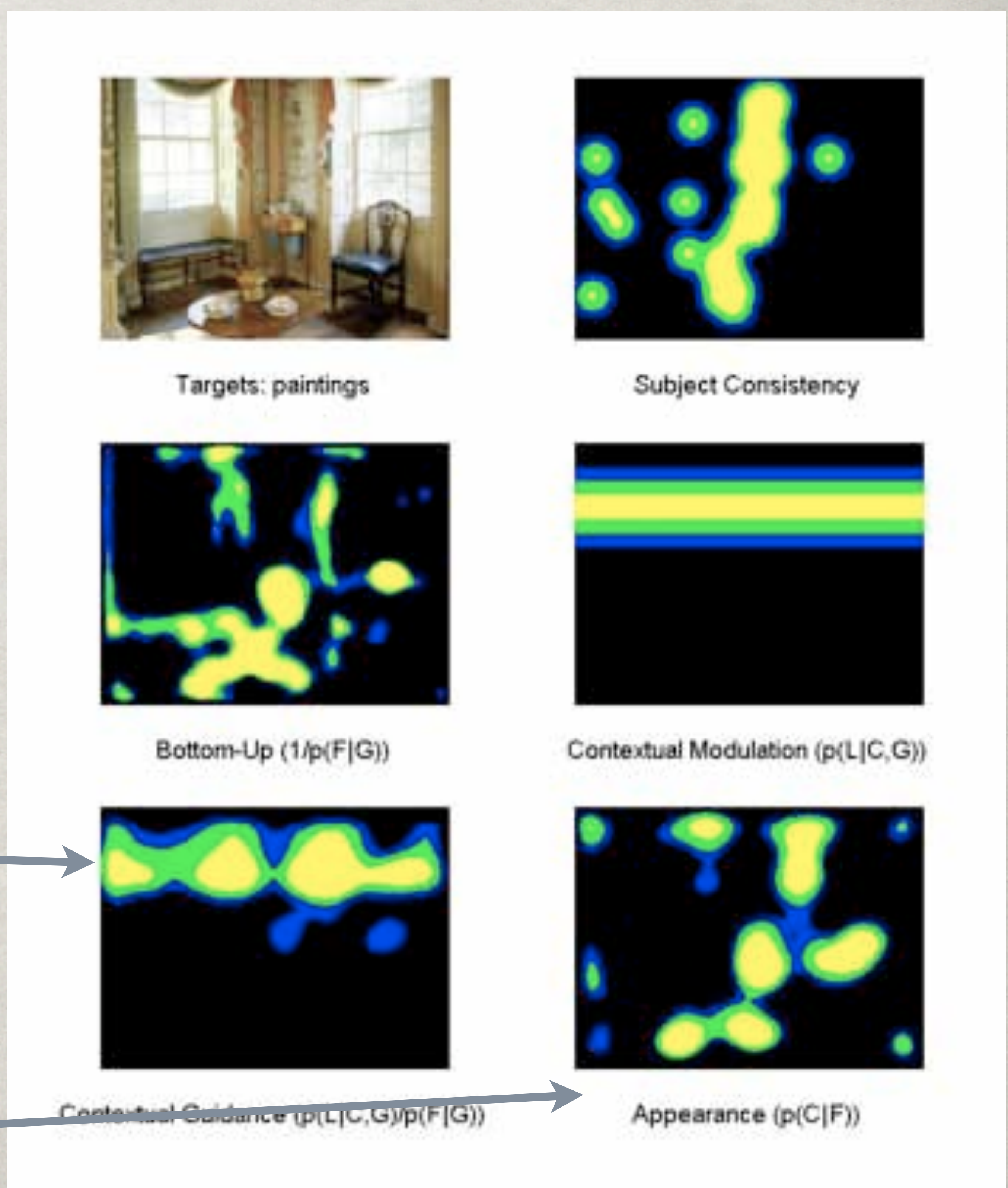
Gist

SUN



Qualitative Results (picture search)

- Where we disagree the most with Torralba et al. (2006)



Gist

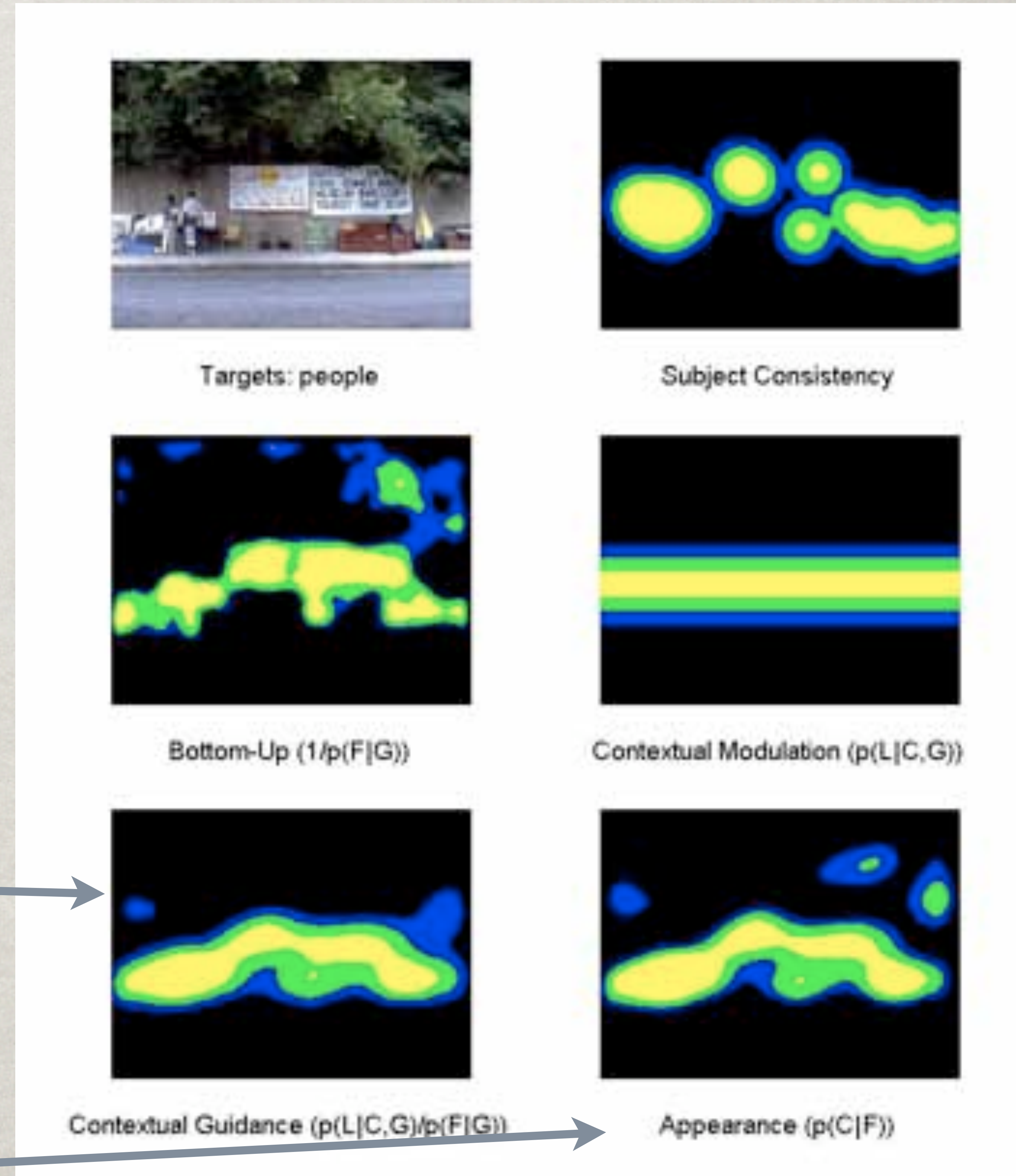
SUN

Qualitative Results (people search)

- Where we agree the most with Torralba et al. (2006)

Gist

SUN



Qualitative Results (painting search)

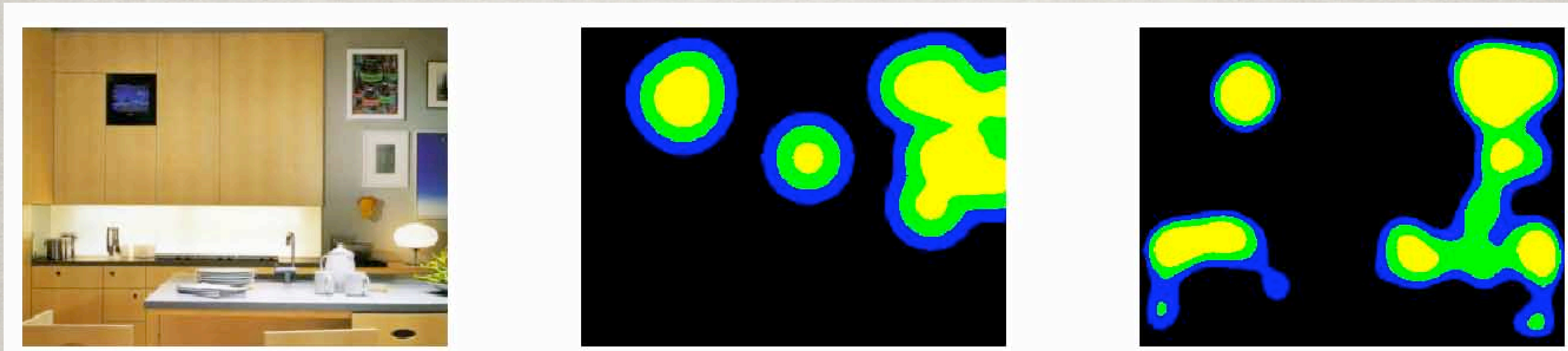


Image Humans SUN

- This is an example where SUN and humans make the same mistake due to the similar appearance of TV's and pictures (the black square in the upper left is a TV!).

*Tong, Kanan, Cottrell

Mathematical Formulation of Information

- "Information" has two (related) mathematical meanings:
  - 1) How much did you expect something you experience ("it is going to rain")?

$$-\log p(x)$$

  - 2) How unsure are you about some aspect of the world (e.g. "is it going to rain?")?

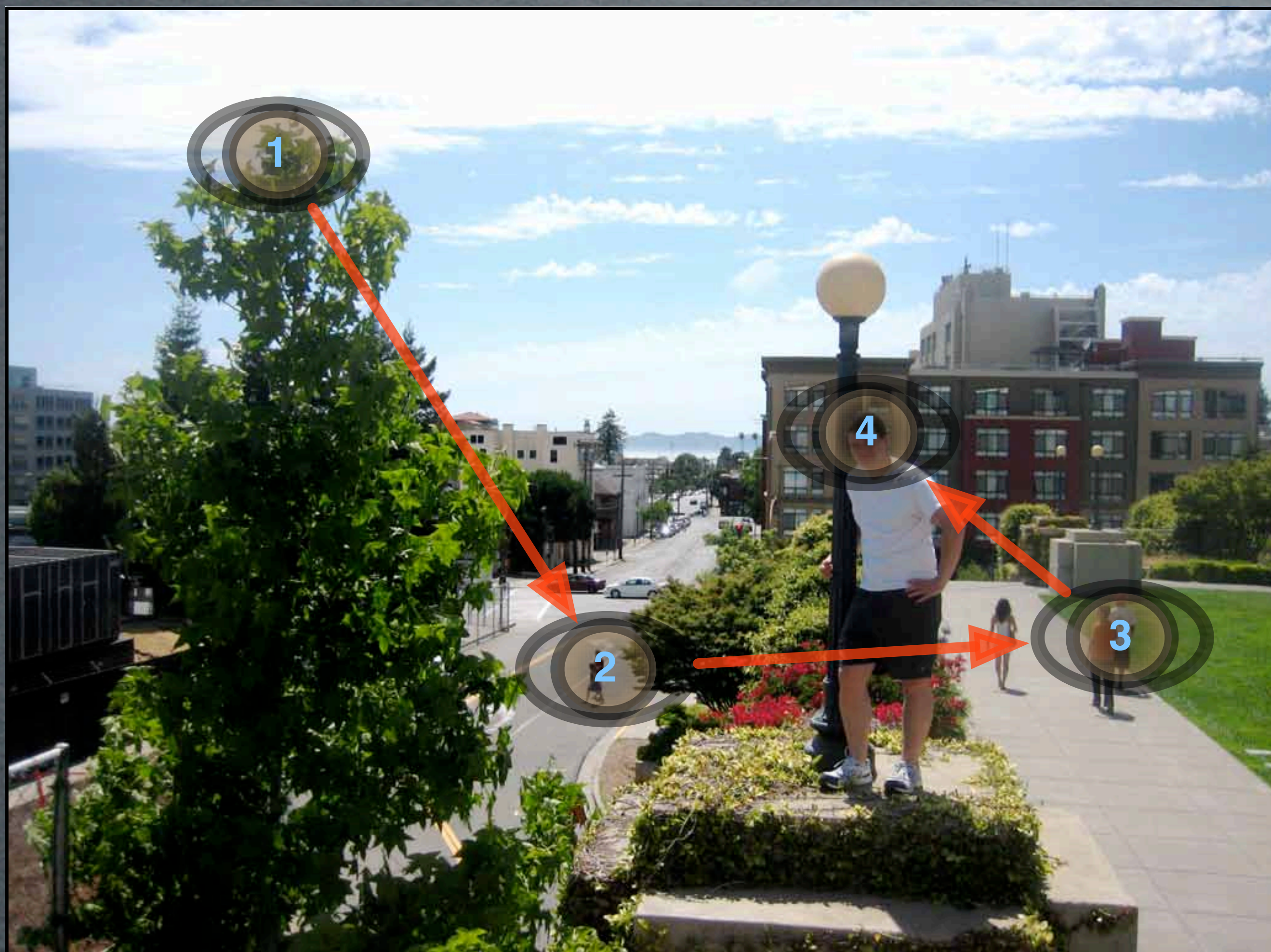
$$-\sum_{Possibilities} p(x)\log p(x)$$

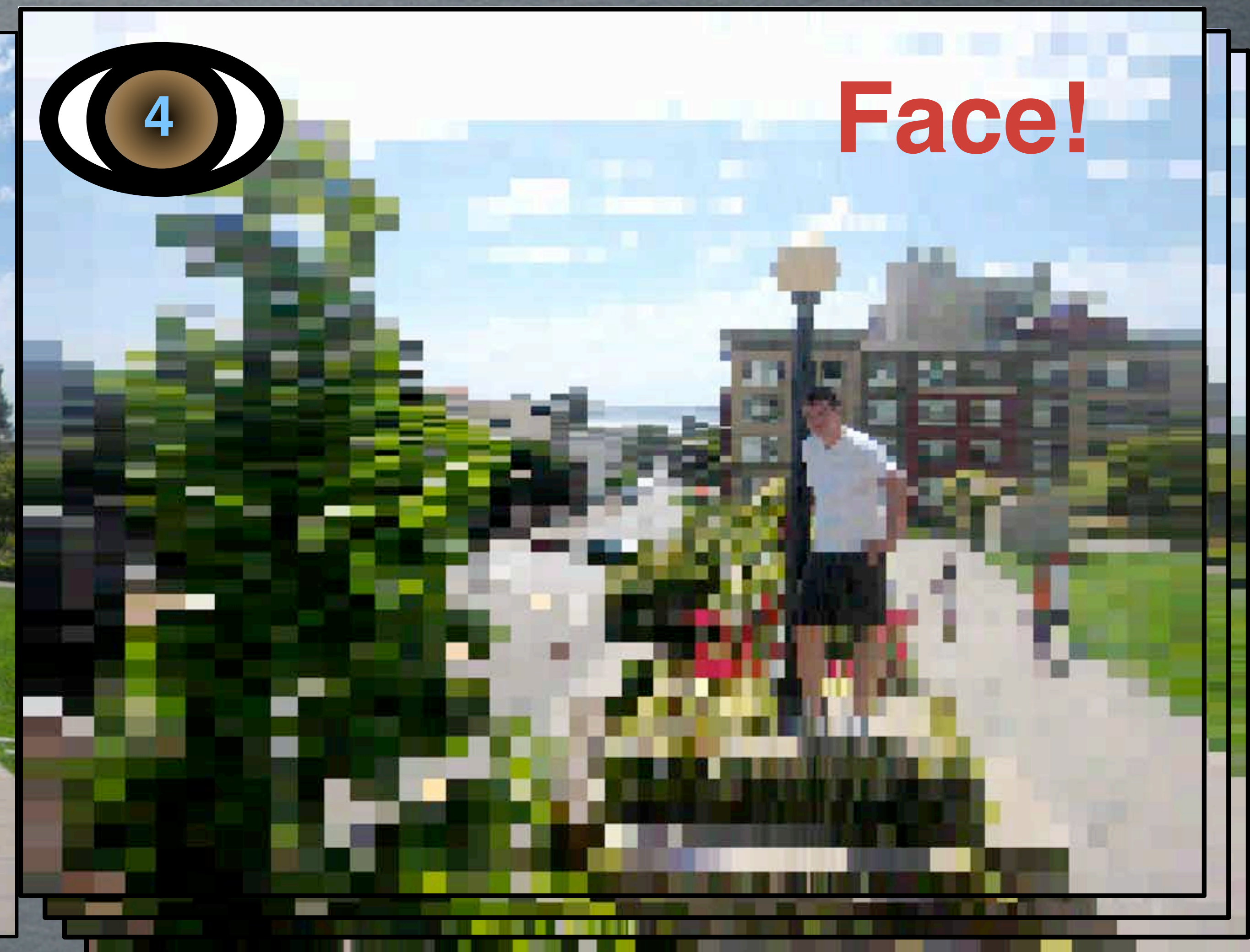
Gathering Information

$$-\sum_{Possibilities} p(x)\log p(x)$$

- Question:
  - "Where is a face?"
- Possibilities:
  - Top-left, Middle, Bottom-right, etc...
  - Or, nowhere.
- Information (above) says how much information we have left to gather about the face location.
  - Once we have gathered the maximum amount of information, we will know where the face is.

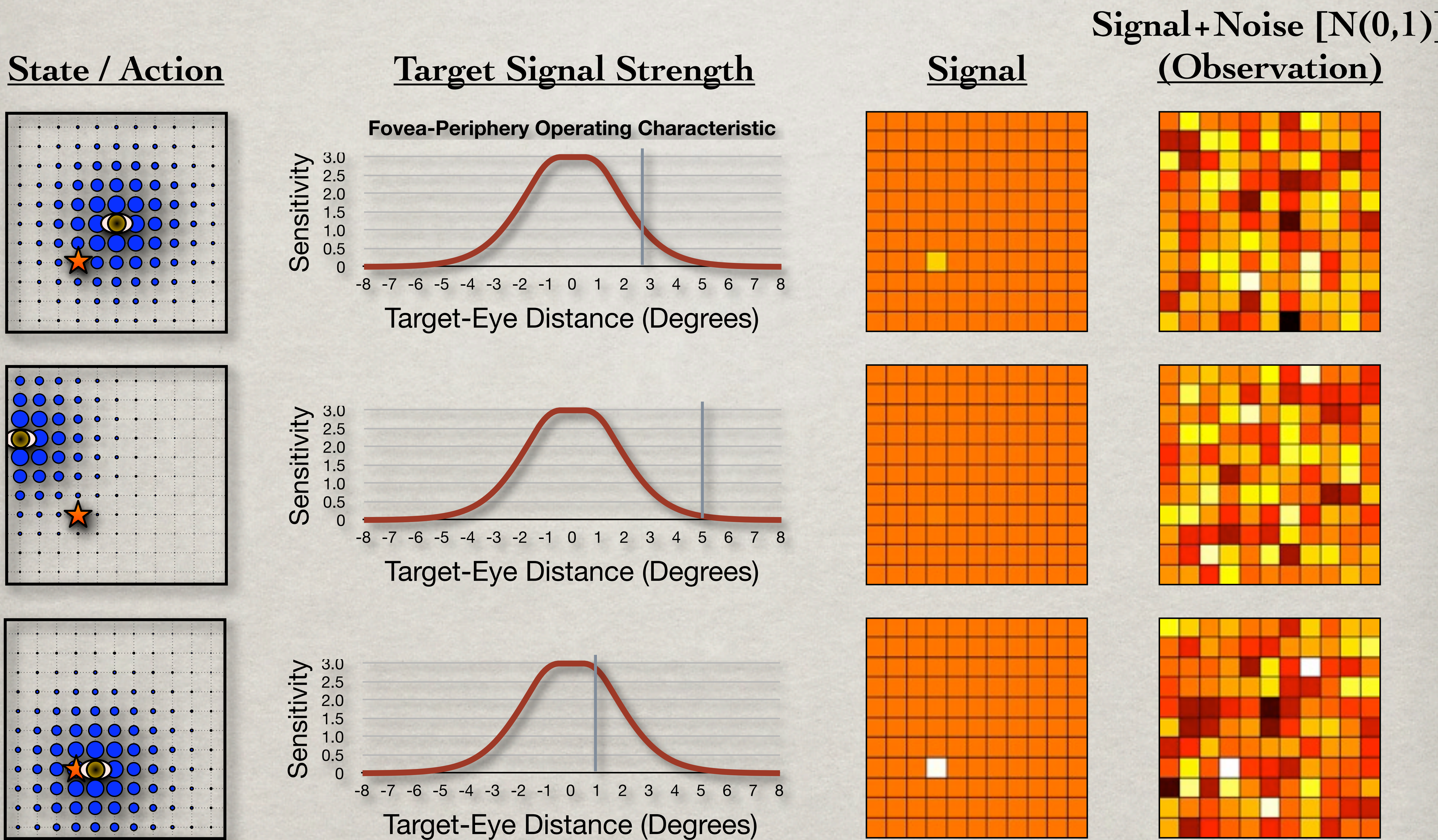
***Butko, Movellan**





Searching for Faces

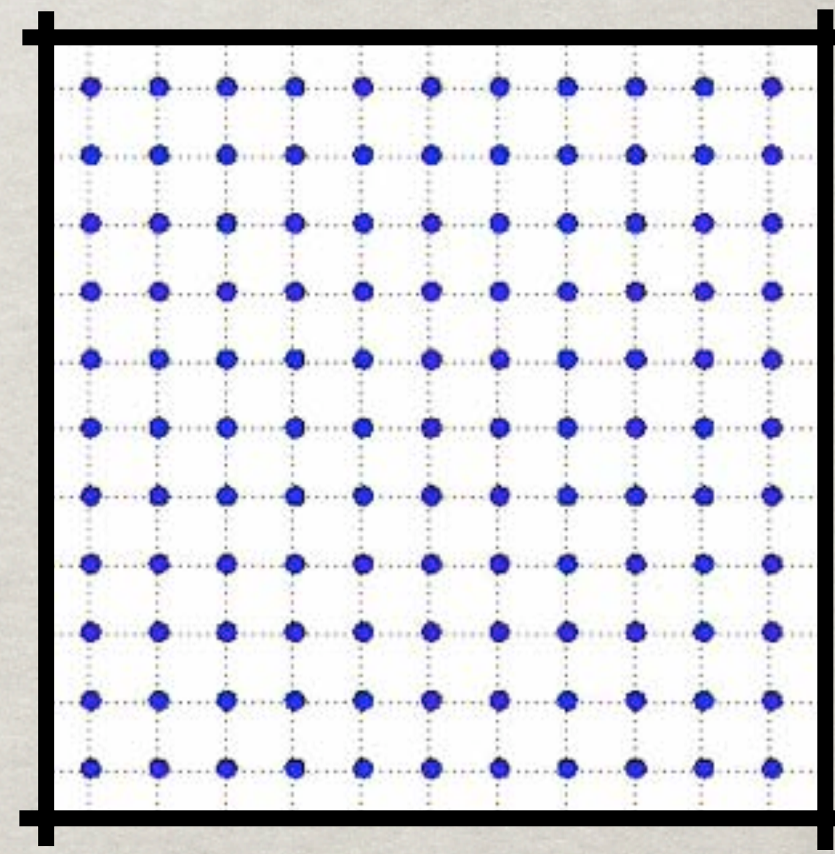
Modeling the Retina
[Adapted from Najemnik & Geisler 2005]



*Apply Infomax principle to learn optimal eye-movement behavior!

Integrating Observations

- Characterized the noise properties of the sensory system.
- The POMDP framework specifies how to *infer* the likelihood that the target is at each location:



$$
\begin{aligned}
p(S=i|A_{1:t},\vec{O}_{1:t}) &\propto p(\vec{O}_t=\vec{o}|S=i,A_t)p(S=i|\vec{O}_{1:t-1},A_{1:t-1}) \\
p(\vec{O}_t=\vec{o}|S=i,A_t=k) &= \prod_{j=1}^{N} p(o_j|S=i,A_t=k) \\
&= 1/\sqrt{2\pi}\exp((o_i-d_{i,k})^2/2)\prod_{j\neq i}1/\sqrt{2\pi}\exp((o_j)^2/2) \\
&= \frac{\exp((o_i-d_{i,k})^2/2)}{\exp((o_i)^2/2)}Z \\
&= \exp(\alpha_{i,k}d_{i,k})Z; \quad \alpha\equiv(o_i-d_{i,k})/2 \\
B_t^i &\propto \exp(\alpha_{i,k}d_{i,k})B_{t-1}
\end{aligned}
$$

- I-POMDP Bayesian analysis:
  - Online Learning
  - Local update rule

***Butko, Movellan**

Digital Retina in Action

Infomax Approach Improves State of the Art AI

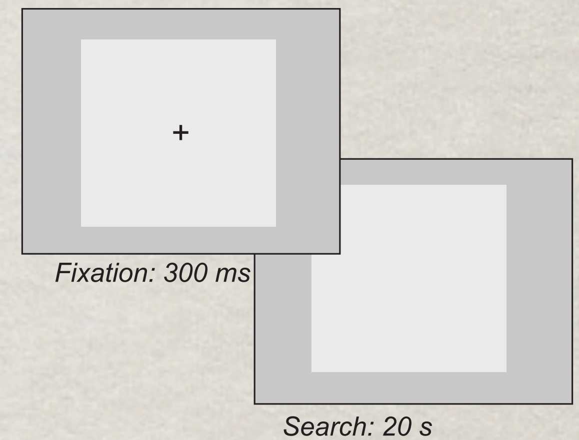
- Apply digital retina sequentially to a static image Vs. search for faces using a standard face detector.
  - Achieve two-fold speed increase with minimal loss in accuracy.
- Optimal Information (Sense 2) gathering.

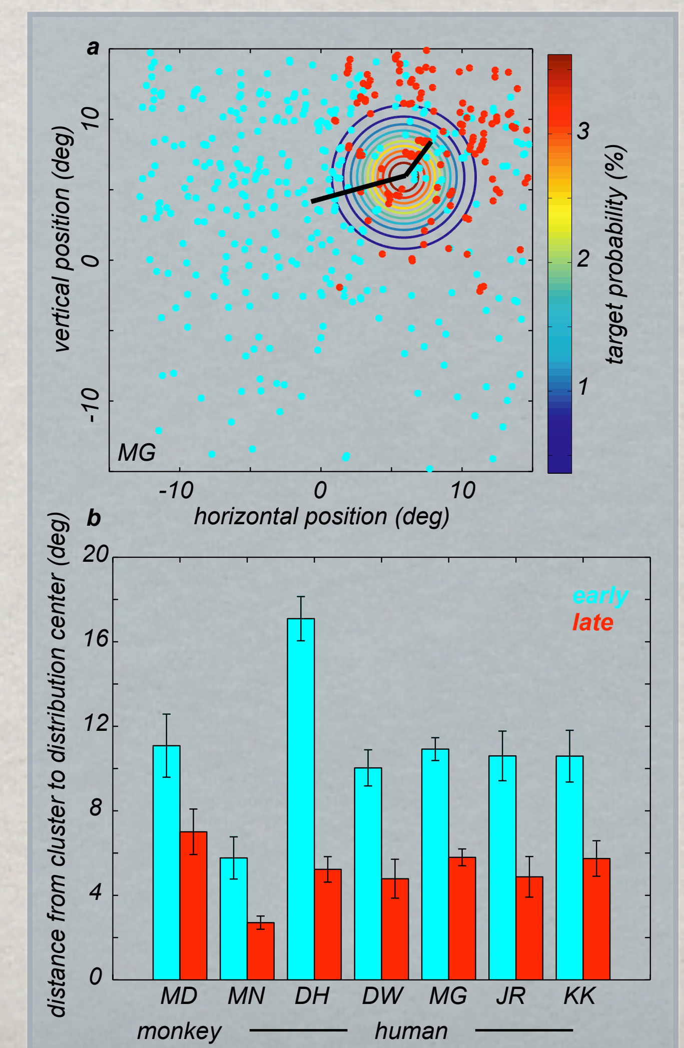
	Digital Retina	Full Image
Runtime (ms/1000px)	**0.57**	1.25
Displacement (% Width)	7.6%	**6%**

***Butko, Movellan**

Is Information The Only model for Eye-Movements?

- Given a set of eye-movement data, how should we model it?
- Experiments in “top-down” effects of eye-movement:
  - “Hidden” target, have to look at something invisible to end trial.
    - No “bottom up” visual information to aid eye-movement.
  - After many trials, learn where target is likely to be, move eyes in absence of visual cues.
- How can we model this study?

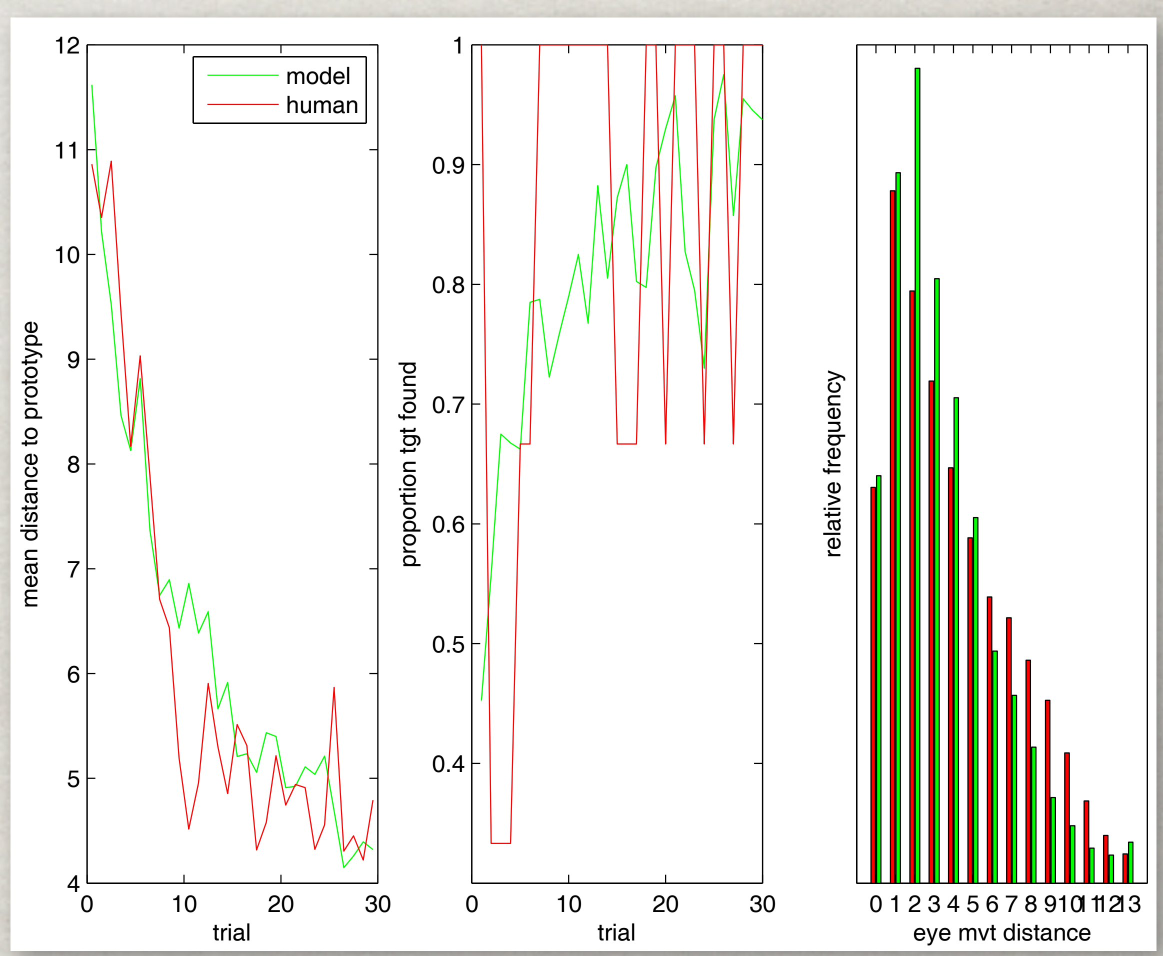




*Chukoskie, Sejnowski

Model 1: Spatial Q-Learning (RL)

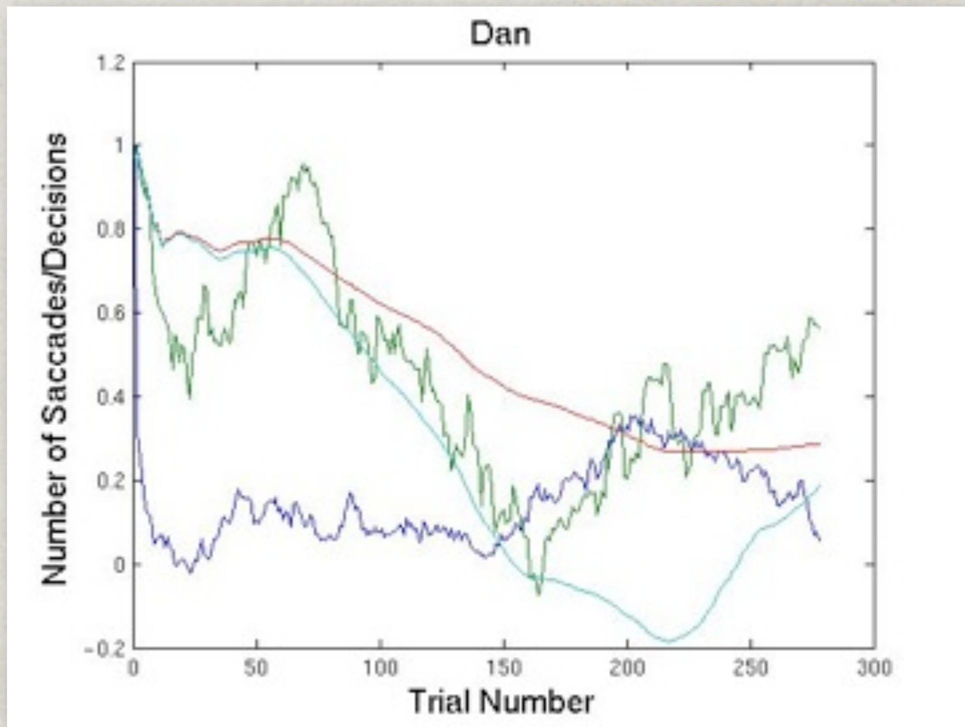
- A subject doing this task is "rewarded" for finding the search target, "penalized" for moving their eyes too much (wasting energy).
  - Can leverage reinforcement learning.
    - Actions that are close (in retrospect, L1 distance) to the target are rewarded more than far away ones.
    - Movements that are far (L1 distance) from the last fixation are penalized.
  - Learn reward structure.
- Choose an eye-movement as a soft-max over the reward of each state.

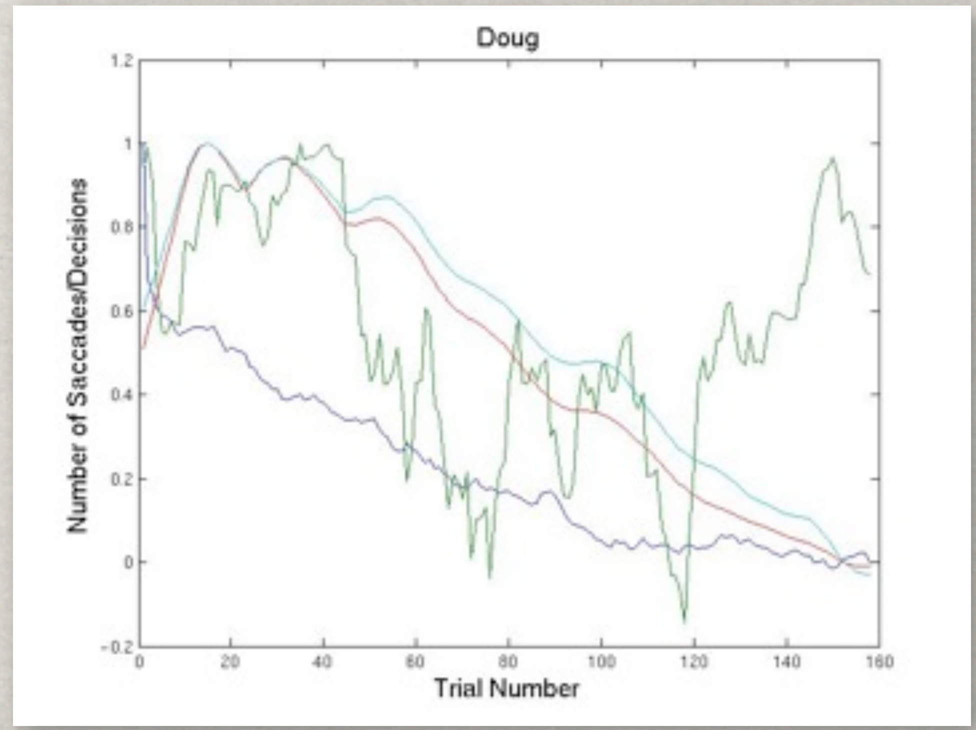


*Mozer

Model 2: Bayesian Optimal Observer

- Use Bayesian parameter estimation to estimate target location distribution.
- After each found-target, update estimates of [x-mean, x-variance, y-mean, y-variance].
- For each trial, sample eye-movements from location distribution.
  - If not-found, set probability for that location to zero.
  - Renormalize and resample to generate next fixation.
- Very similar to Infomax Approach (earlier)





Subject Performance

Model Performance

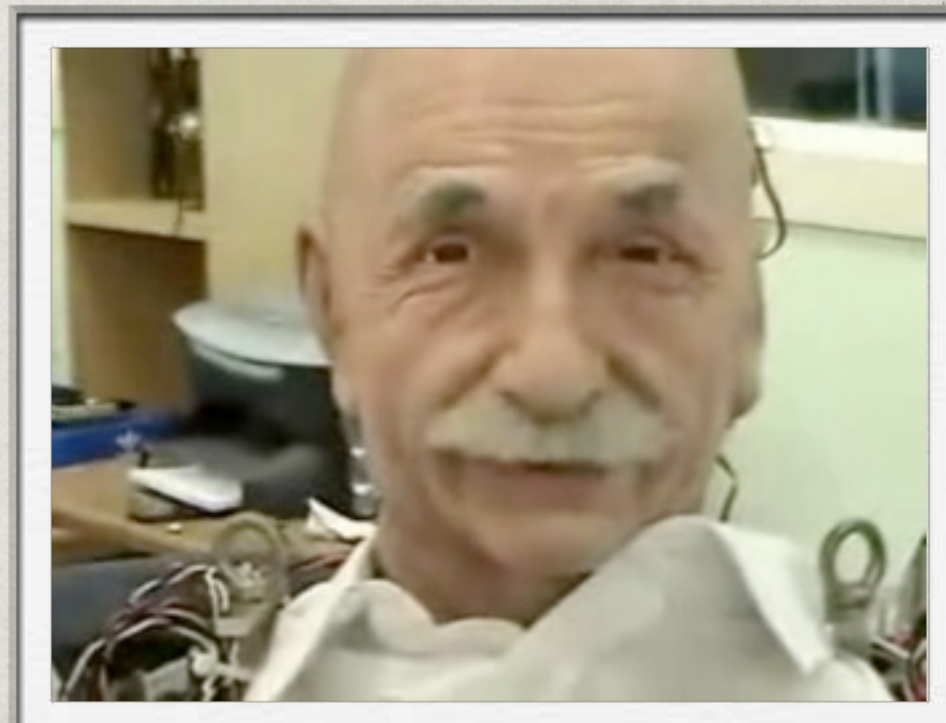
KL Distance Data Distr. to Target Distr.

KL Distance Recent Data to Target Distr.

*Arnold

Next Steps

- Einstein Tutor
  - In order to teach, you need to effectively gather information about the mental state of your pupil.
  - Attentive? Confused? Bored?



- Project One
  - Robotic Platform to simulate developmental processes and learning during the first year of life.