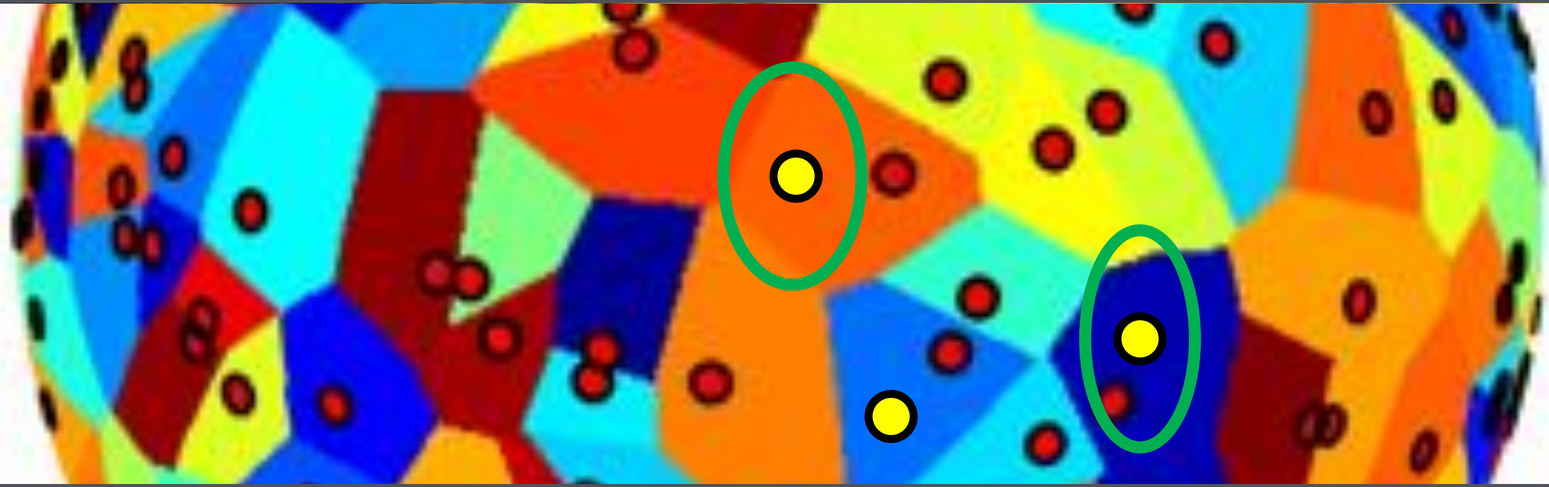
School of Psychology and Clinical Language Sciences

University of Reading

$\pi(a|s,g)$



Andrew Glennerster

University of Reading

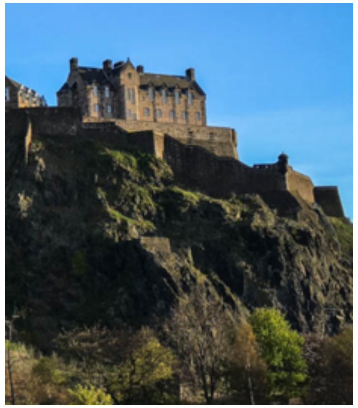
Marialena Stefanou

Alex Muryy

Luise Gootjes-Dreesbach

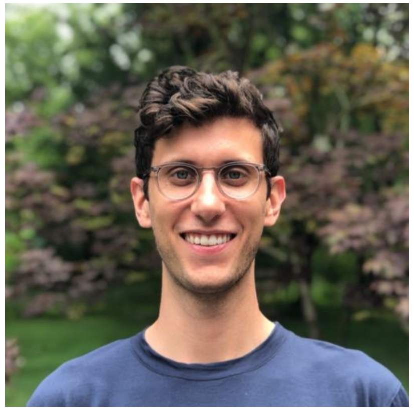
Experiments

Phil Torr



N. Siddharth

Nantas Nardelli



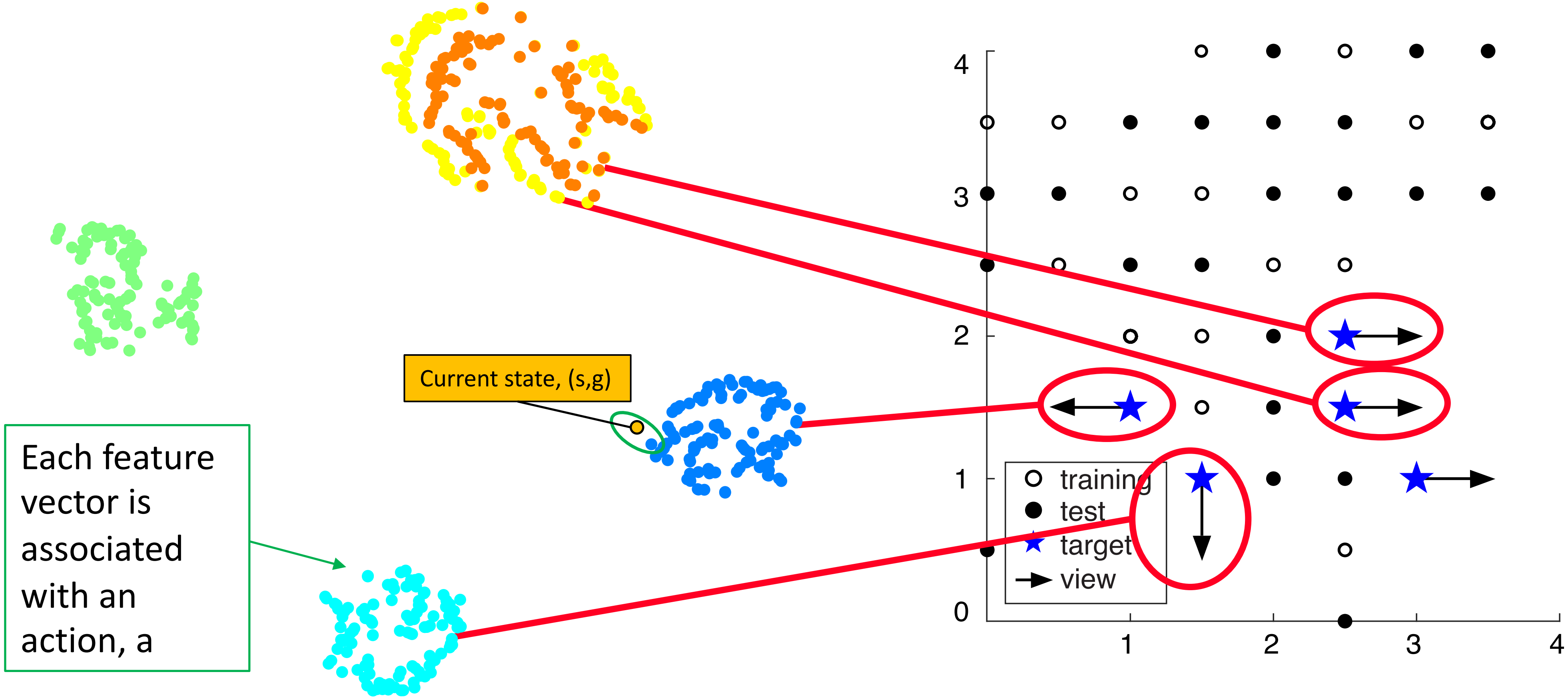
Mark Edmonds

Theory

EPSRC

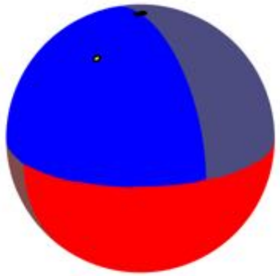
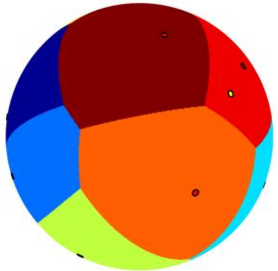
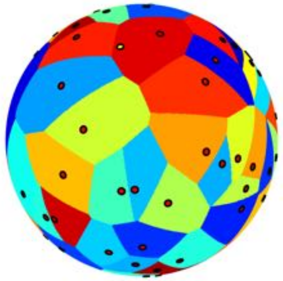
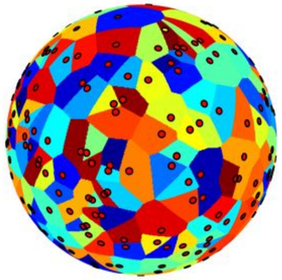
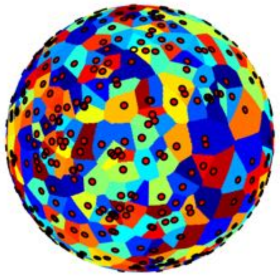
[dstl]

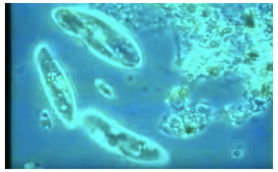
π(a|s,g) for navigation

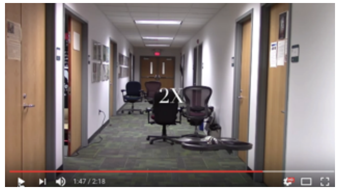


- Zhu et al (2016)
- tSNE shows projection of feature vectors (s,g)
- g dominates

Evolution of $\pi(a|s,g)$





?

- bacterium
- 3D state space
- 2 actions
- few stored contexts for action (s,g)

- plants
- more dimensions for state space
- (more actions)
- more complex π

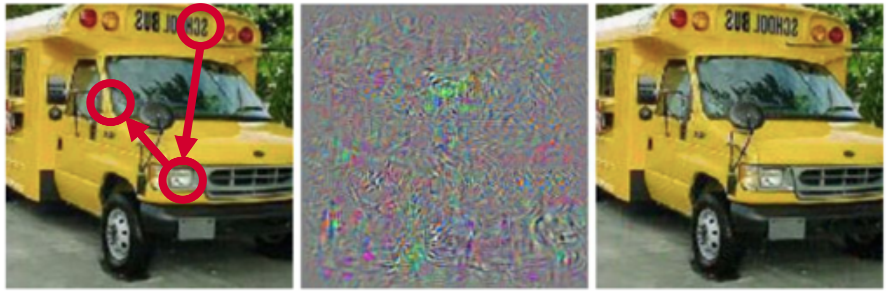
- RL
- e.g. 4096 dimensions for state space
- more complex π

- humans
- more dimensions for state space
- (including ‘virtual’ actions)
- more complex π

higher intelligence
- more dimensions
- more complex π
- paths through state space are rewarded

Behzad's question:

- What are computer vision/reinforcement learning researchers doing wrong?
    - Using the wrong basis set
    - Often trying to do one-shot recognition
    - Task should be integral to recognition



'Bus' 'Ostrich'

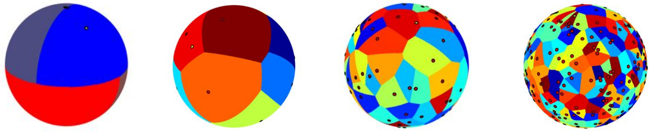
$$\pi(a|s,g)$$

saccades

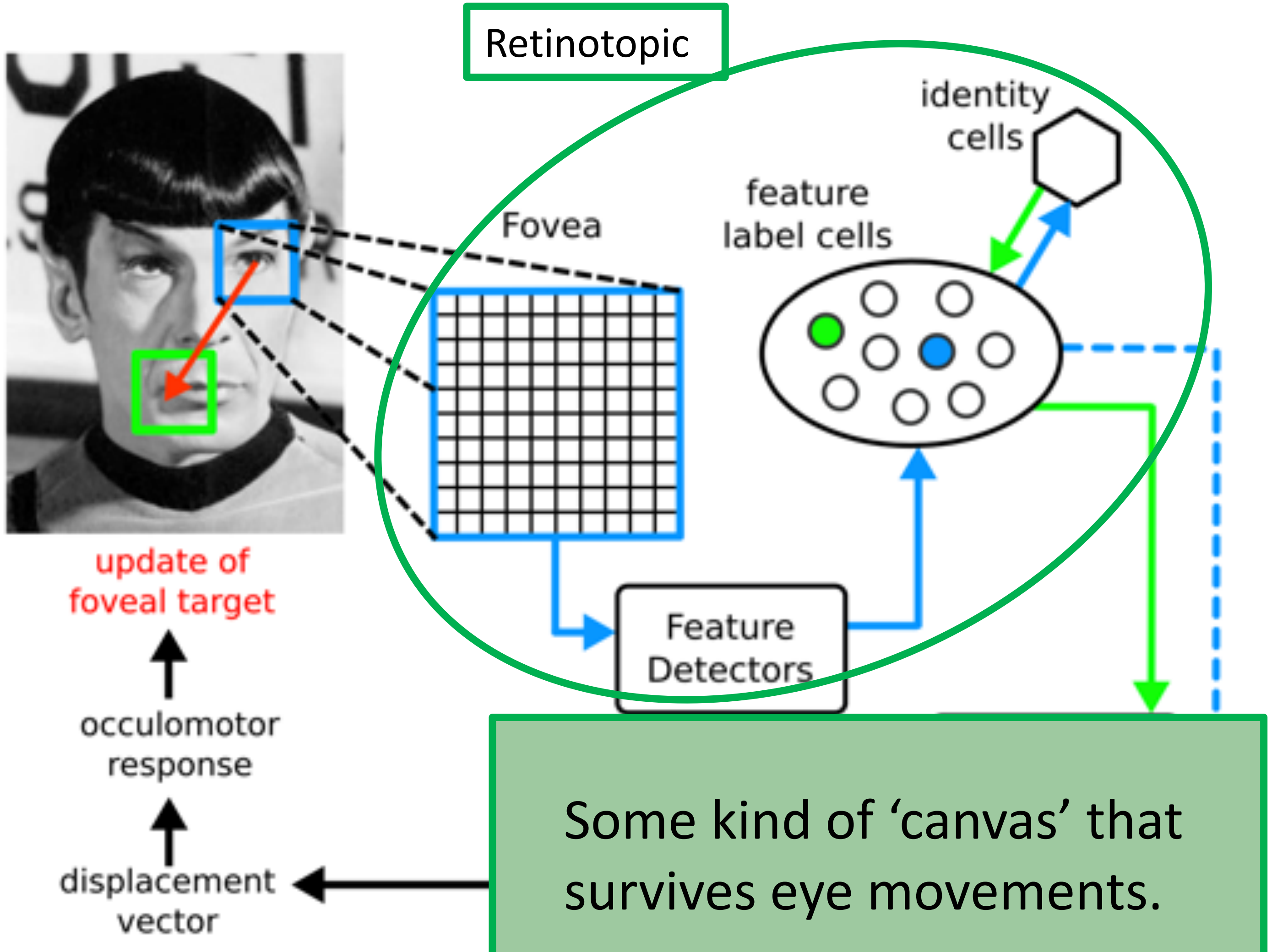
parts of image

seeking confirmation

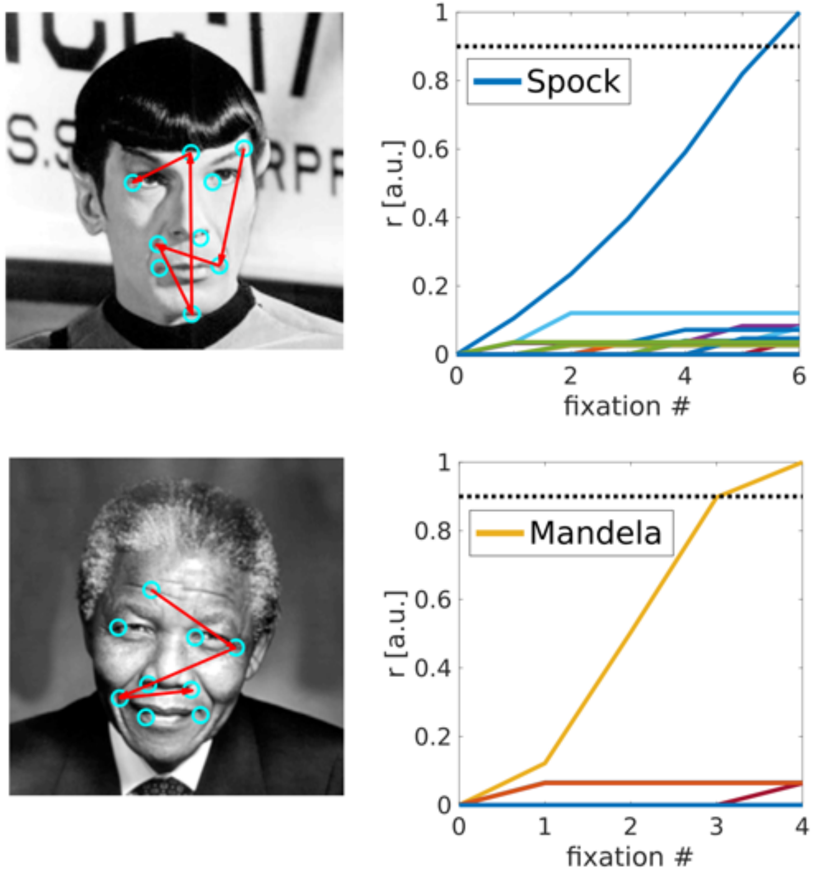
- 3D vision as an example
    - stereoacuity with a moving eye
    - up to navigation
    - avoiding 3D coordinate frames
        - experiment in VR



π(a|s,g) for face recognition



Some kind of ‘canvas’ that survives eye movements.



- The ‘canvas’ can be a policy network
- Bicanski and Burgess (2019)
- Bill Triggs: ’necessary’

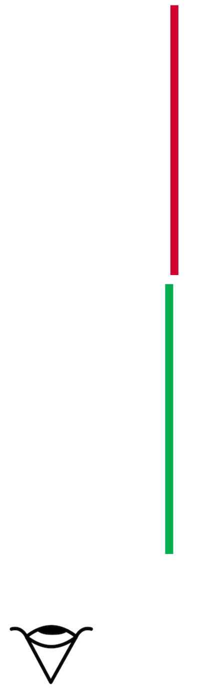
A ‘canvas’

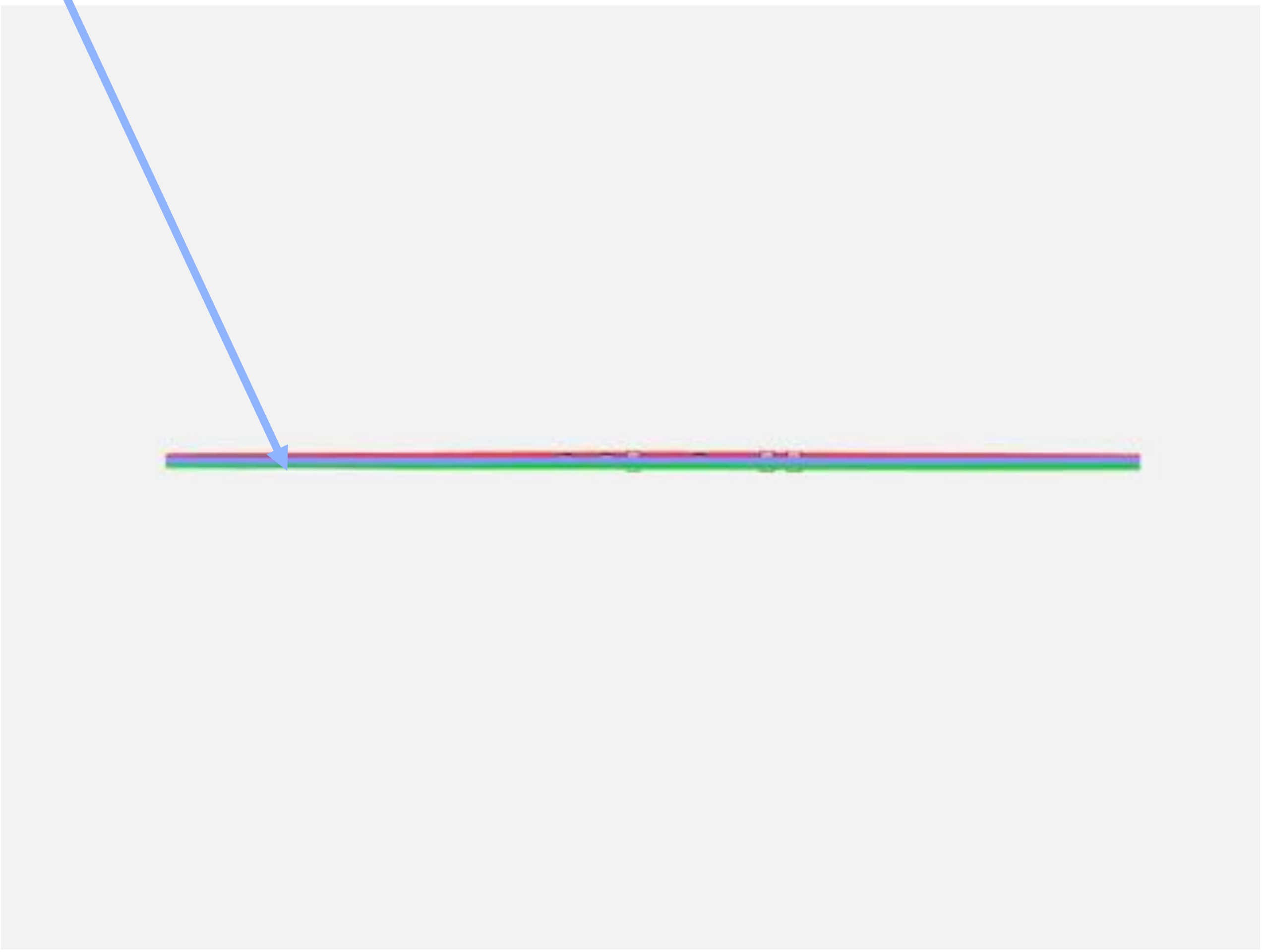
What is this?

Brain fits a smooth curve and reports the red and green centroids

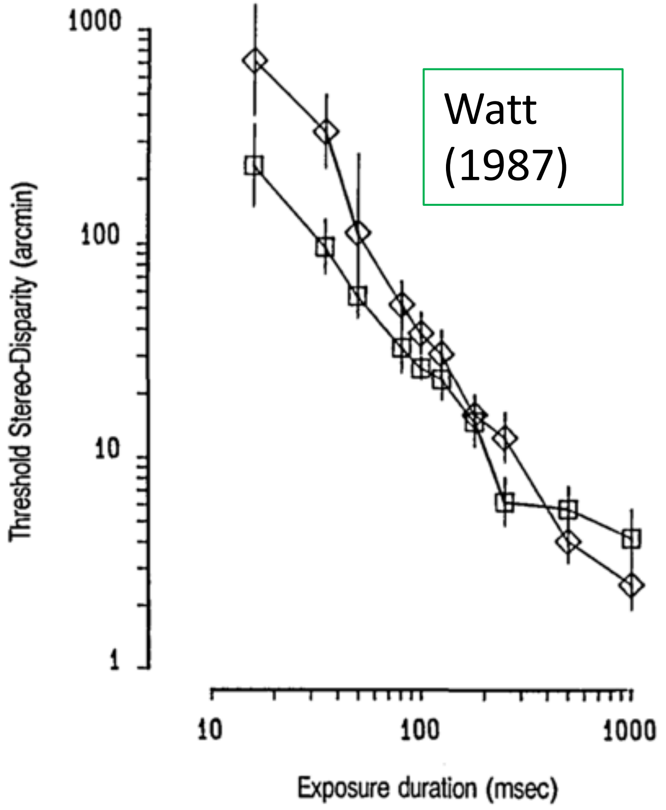
To do this, it must average spikes across time and space

Vernier task: red line left or right of green?





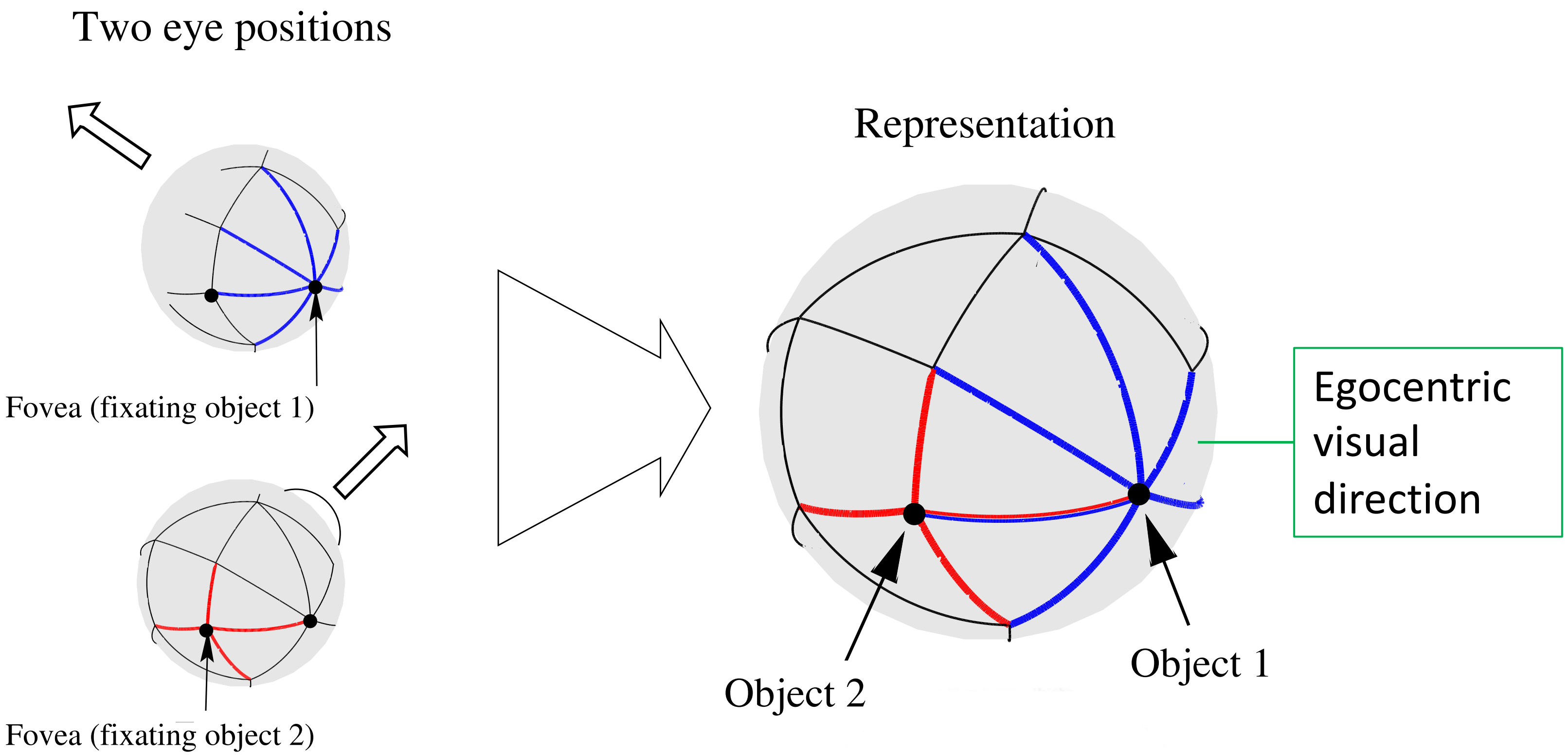
Brain’s task: integrate across different eye positions



Watt (1987)

- Melcher and Morrone (2003)
- Coordinate frame for accumulating evidence is not retinotopic
- Product of experts: Hinton (1999)

A ‘canvas’ surrounding the camera



- Extend the ‘canvas’ all the way around the camera.
- An egocentric representation of direction

$$\pi(a|s,g)$$

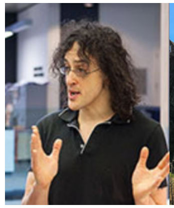
saccades

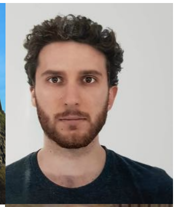
next fixation target

part of the optic array

- Glennerster, Hansard and Fitzgibbon (2001, 2009), optic flow in fixating observers

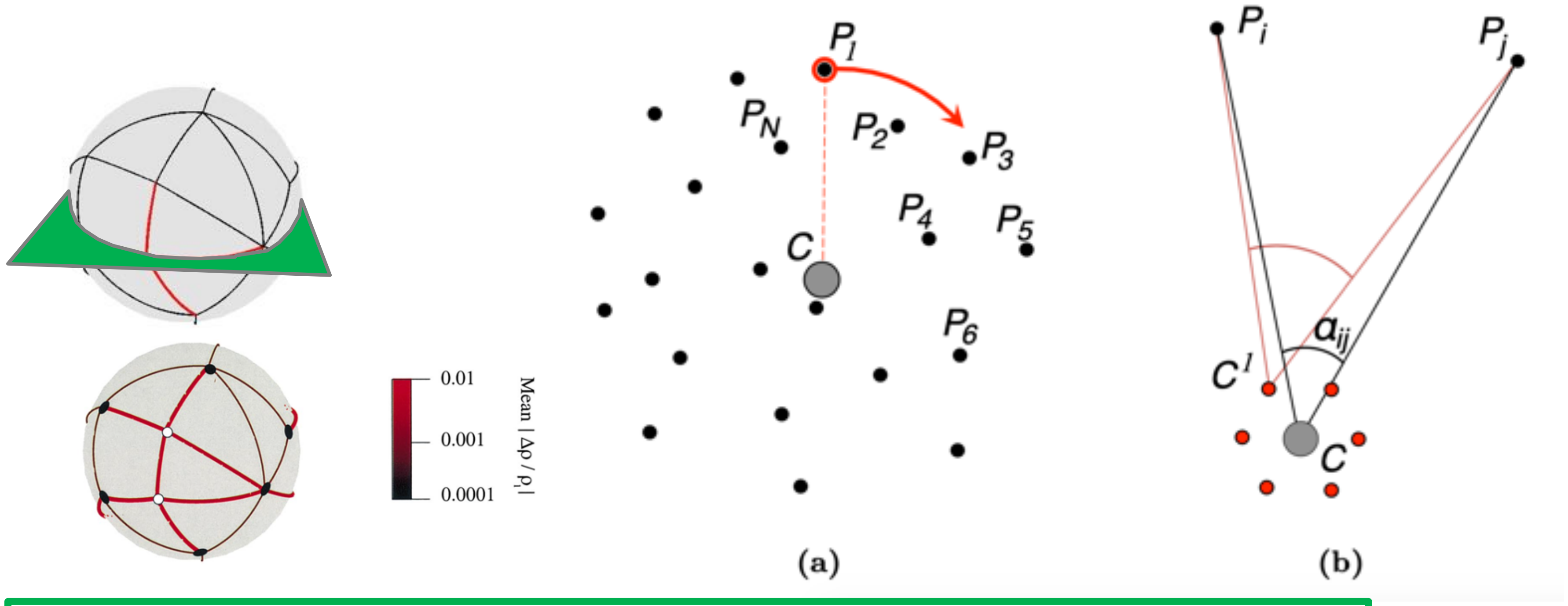
Hierarchical, compositional encoding of location





Muryy,
Siddharth,
Nardelli,
Glennerster
and Torr
(2020)

Hierarchical, compositional encoding of location

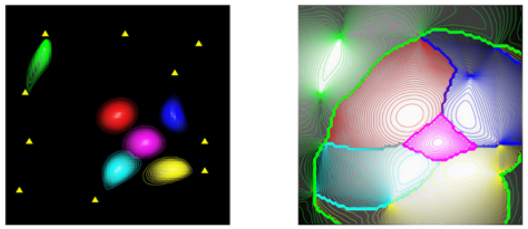


$$\boldsymbol{\varepsilon} = \{\alpha_{ij} : i = 1, \ldots, N, j = (i+1), \ldots, N, 1, \ldots, (i-1)\}. \tag{1}$$

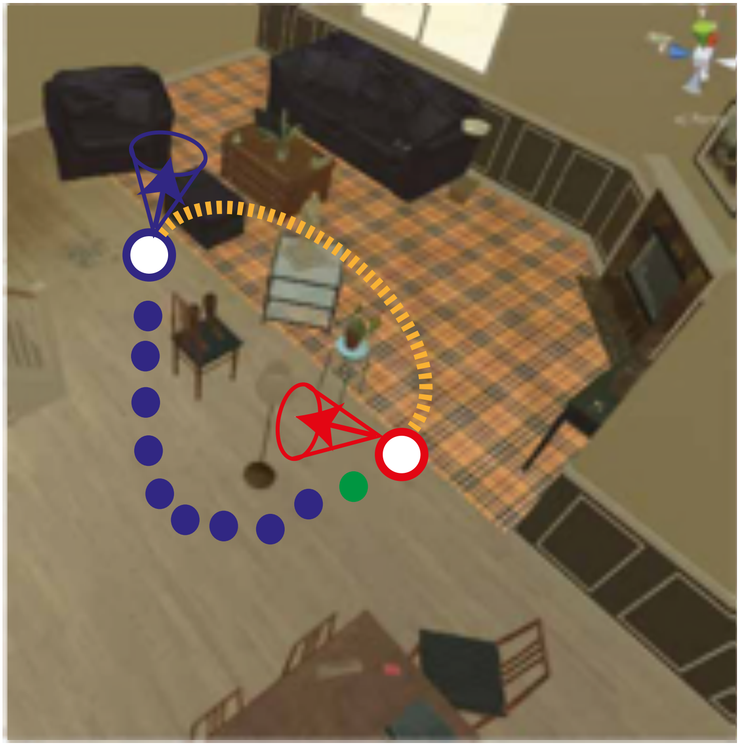
$$\boldsymbol{\psi} = \{\psi_n\}_{n=1}^{N} = \frac{1}{n_C} \sum_{k=1}^{n_C} \frac{\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_{C^k}}{\boldsymbol{\varepsilon}} \tag{2}$$

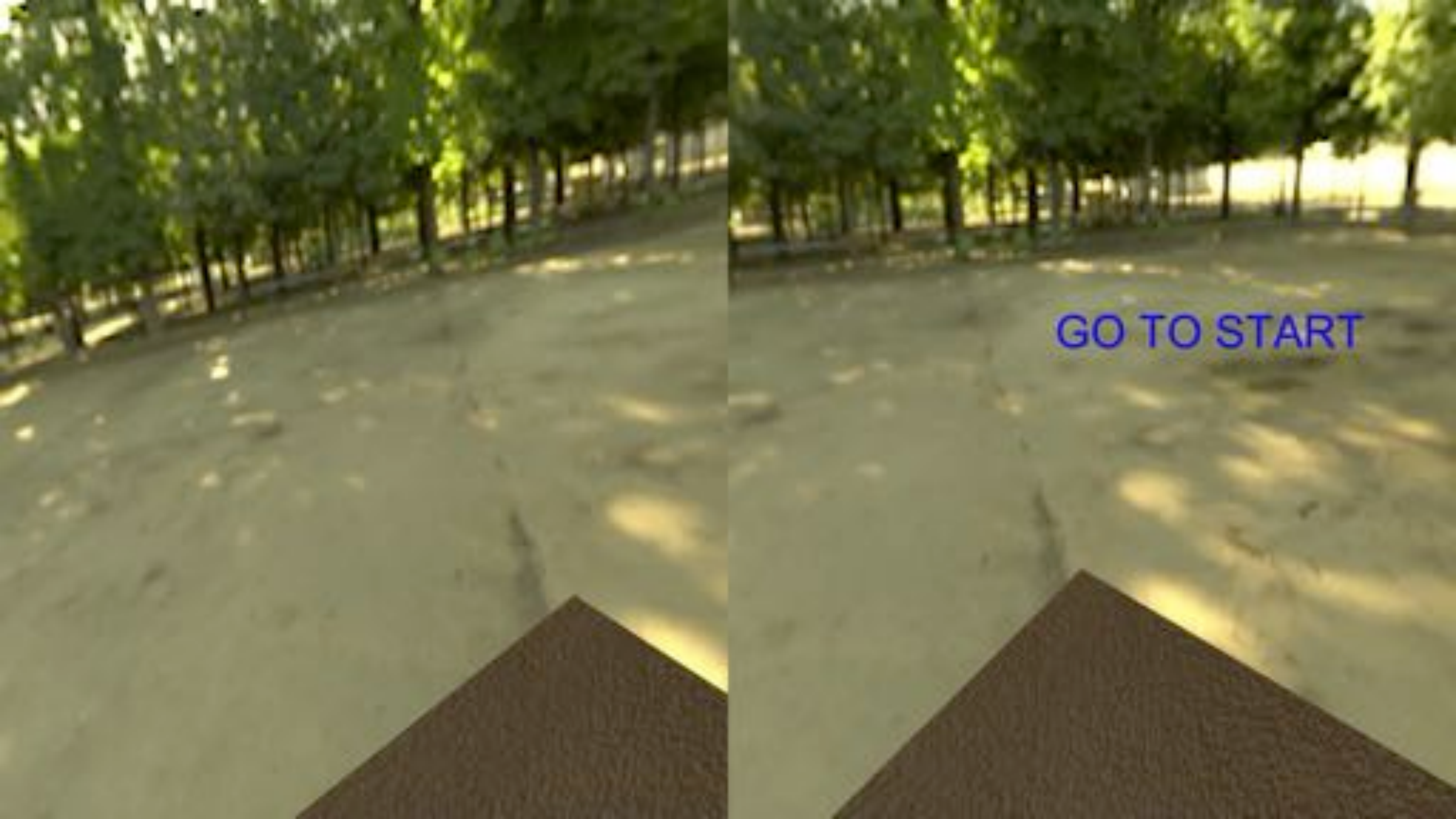
Since $\boldsymbol{\psi}$ has the same ordering of elements as $\boldsymbol{\varepsilon}$, each element of $\boldsymbol{\psi}$ contains a parallax-related measure referring to that particular pair of points.

- The 'visual image' (ε) is a vector listing all the angles (N^2) between all N visible points.
- For each angle, we also calculate a measure of parallax. If P_i and P_j are both distant, this parallax will be small.
- Muryy et al (2020) and Glennerster, Hansard and Fitzgibbon (2001)



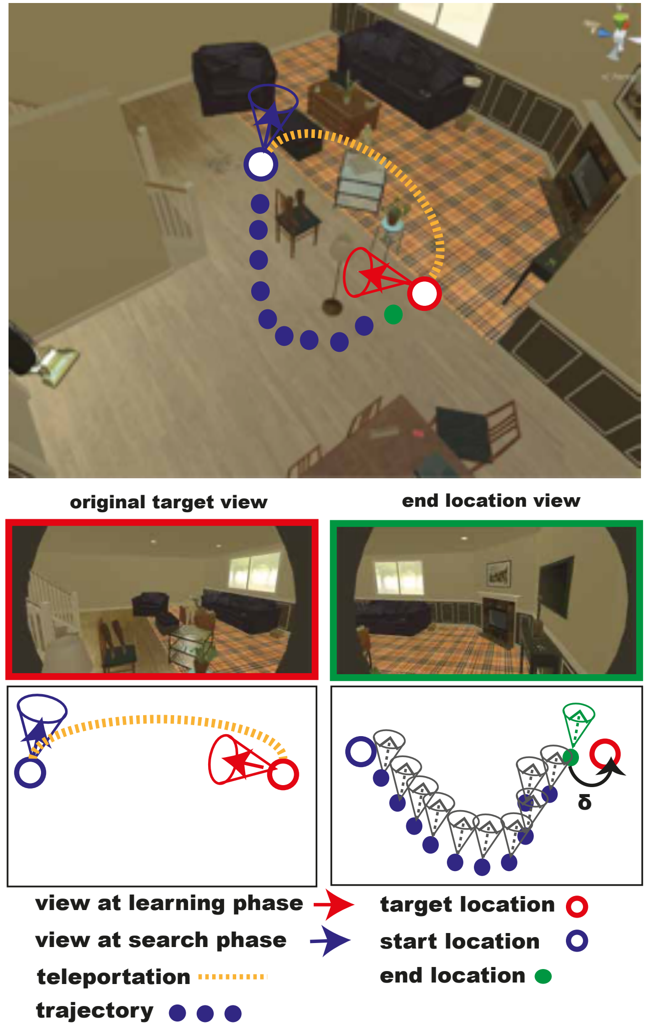
Homing without image matching

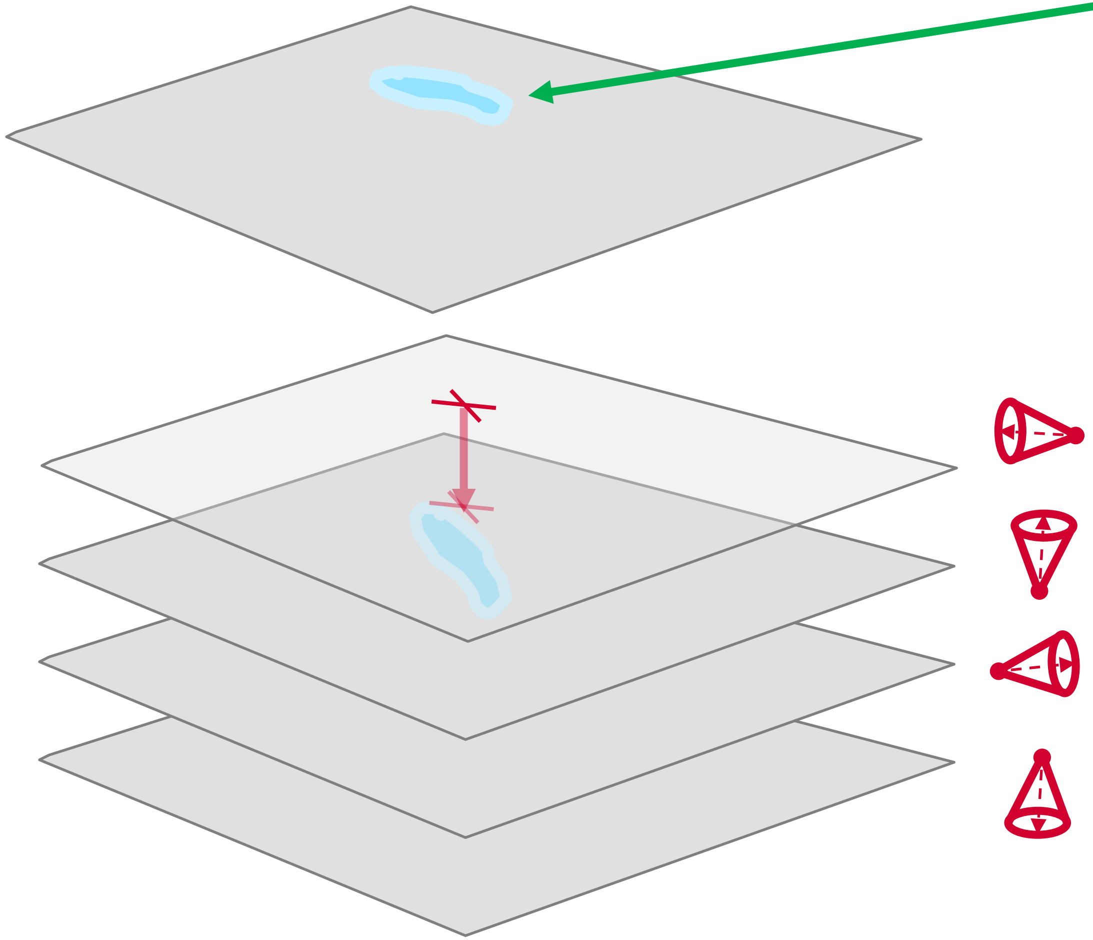


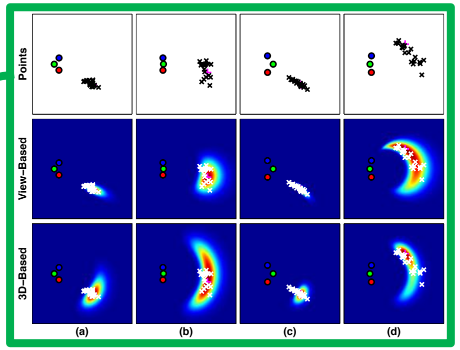

GO TO START

Homing without image matching

University of Reading

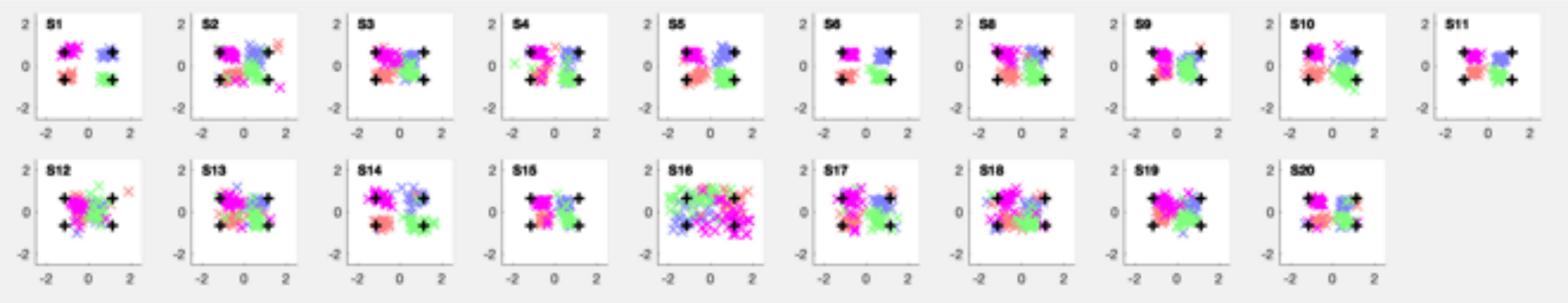


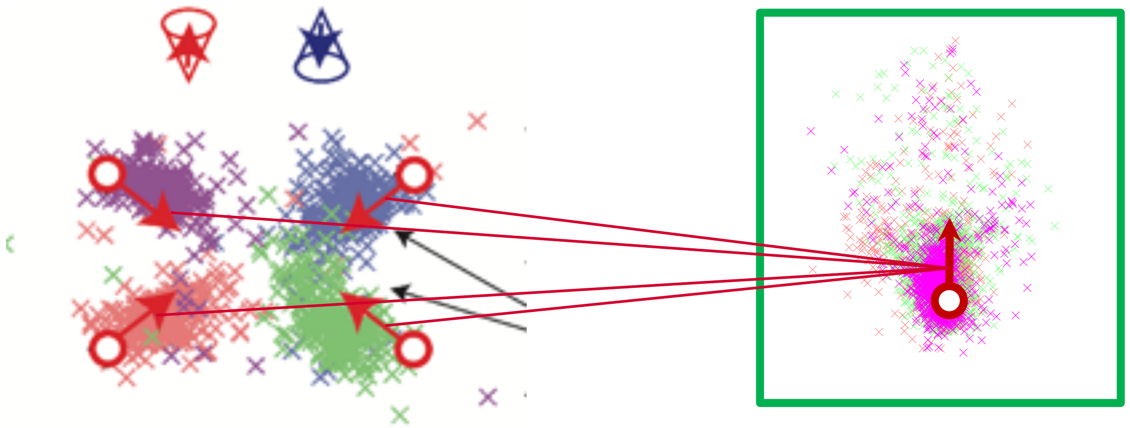




Gootjes-Dreesbach et al (2017)

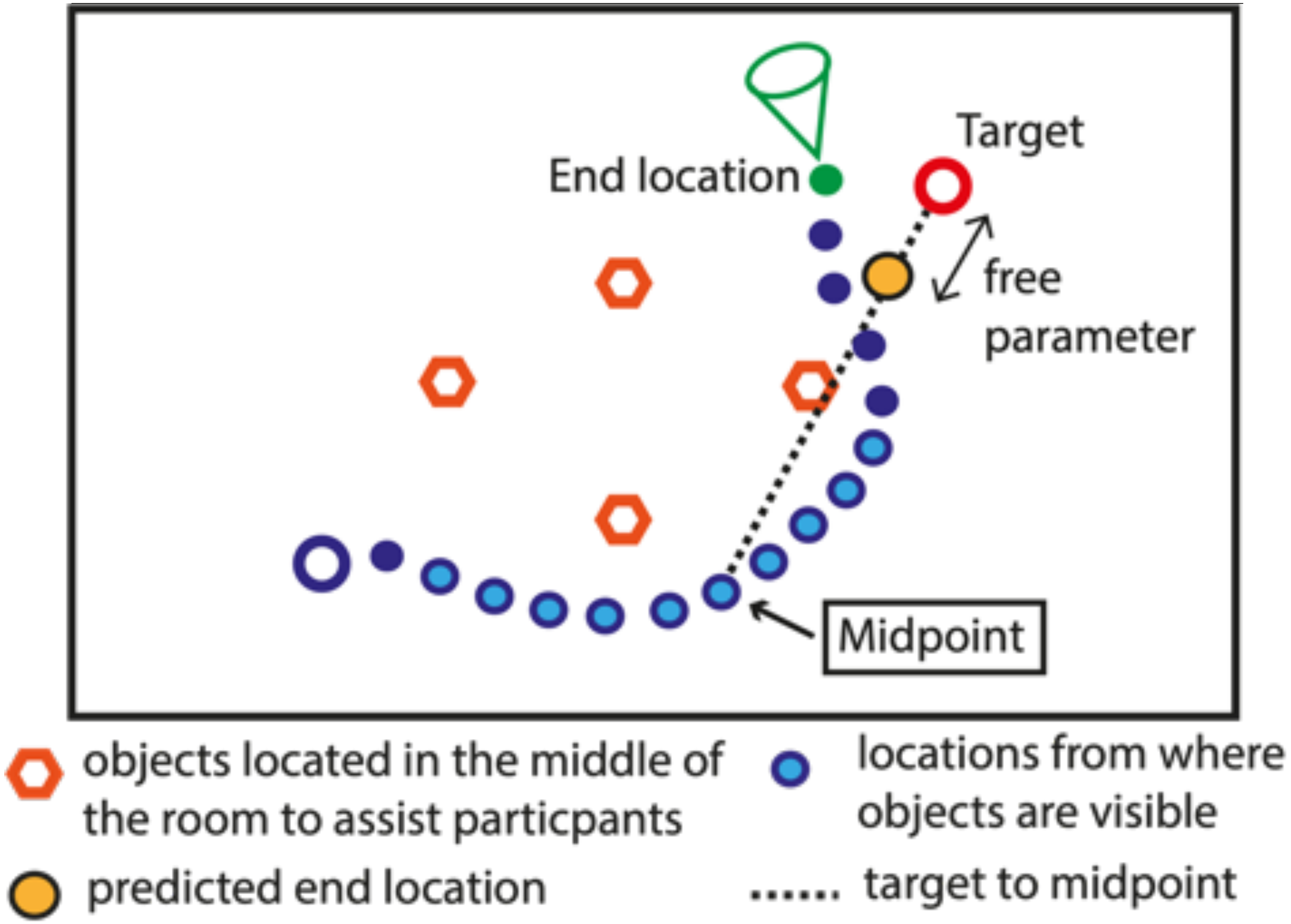
Homing without image matching



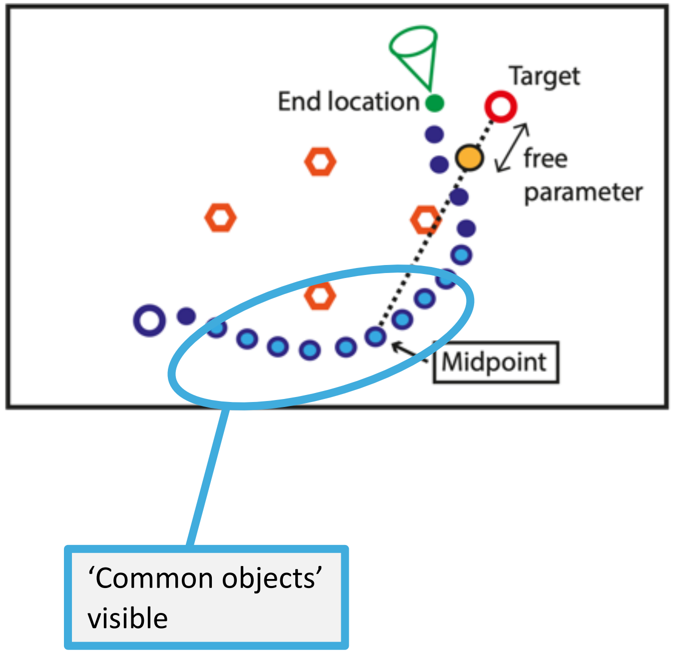


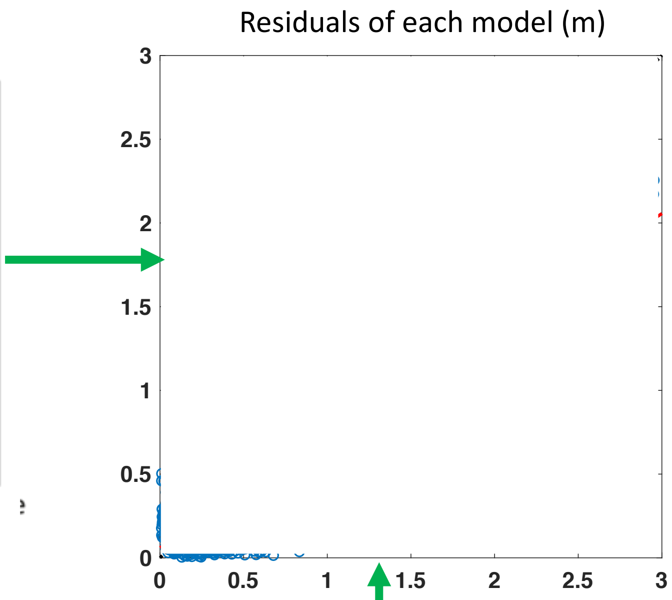
- All participants show a consistent bias relative to the true ‘home’ location

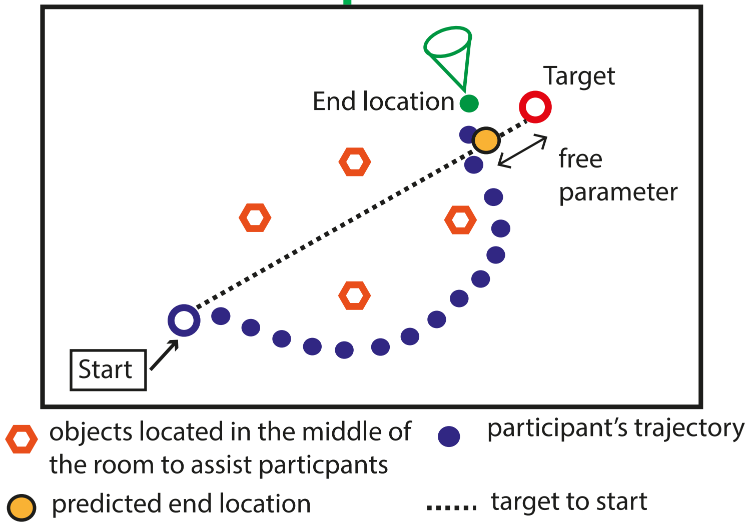
Homing without image matching

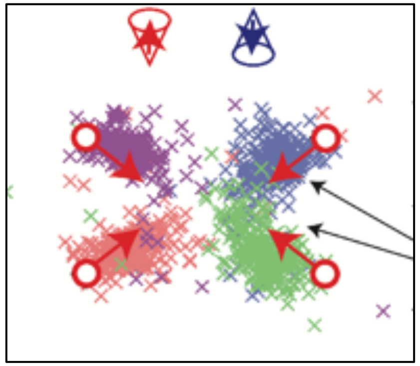


Homing without image matching



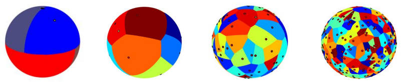






- Participants take a circuitous route to reach the target (to help orient themselves)
- A model that takes this into account does better at explaining the data than one based on the 'Start' location
- Neither a 3D reconstruction model nor a neural network model predict these biases in any obvious way
- We are hoping to model this behaviour in collaboration with Phil Torr's group and DeepMind, i.e. using Generative Query Networks (Eslami et al, 2018)

What are the big problems ahead?

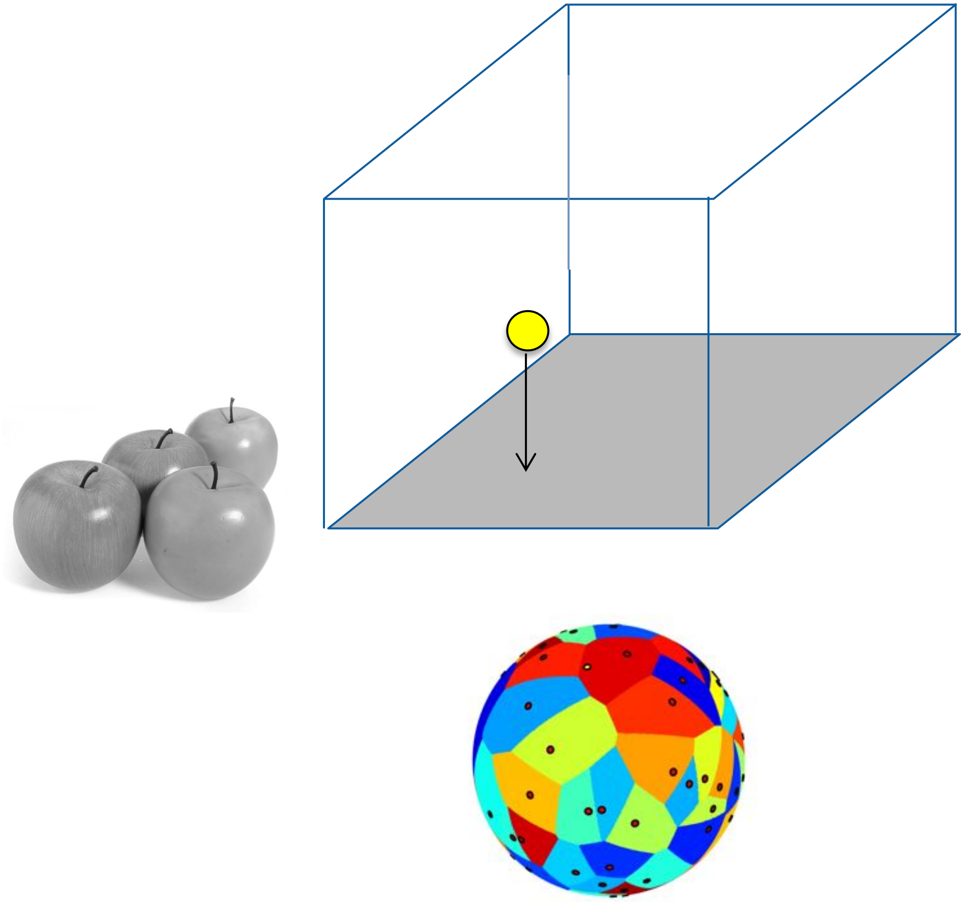


$\pi(a|s,g)$

3D, SLAM

	Straightforward or assumed	Challenging or costly
Model-free inference ('Skinner-like')		
Model-based inference ('Marr-like')		

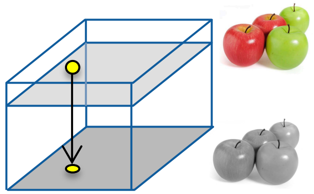
Adding dimensions as tasks evolve



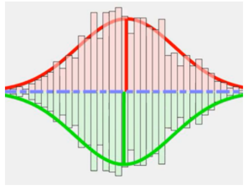
- … or losing dimensions at dusk…
- contexts for action are always compositional

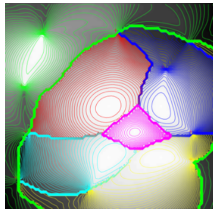
In summary, to answer Behzad's question:

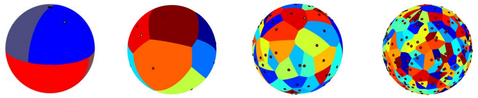
- What are computer vision/reinforcement learning researchers doing wrong?
  - Using the wrong basis set
  - Task should be integral to recognition
  - Often trying to do one-shot recognition



- 3D vision as an example
  - stereoacuity with a moving eye
  - up to navigation
  - could avoid 3D coordinate frames (testable)





-  $\equiv$ $\pi(a|s,g)$ … a good way to think about the brain

Collaborations, COVID, future plans

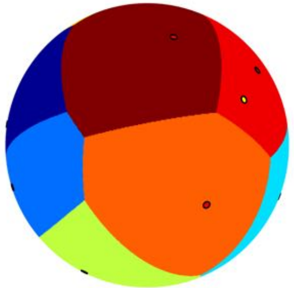
Collaborations and future plans

- Oxford
  - One joint paper (Muryy et al, 2020)
  - Grant to EPSRC about to be submitted
    - Reading
    - Oxford
    - Edinburgh (Siddharth)
    - Leeds
  - Will explore GQN modelling of human navigation and inductive biases inspired by human navigation results
- UCLA
  - Review paper with Mark Edmonds

COVID strategy for experiments

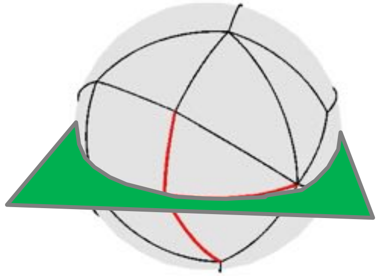
- One participant uses the head set until all their data is collected. Then clean and rest HMD.
- Ventilated lab, PPE, distancing and no touching of items in common

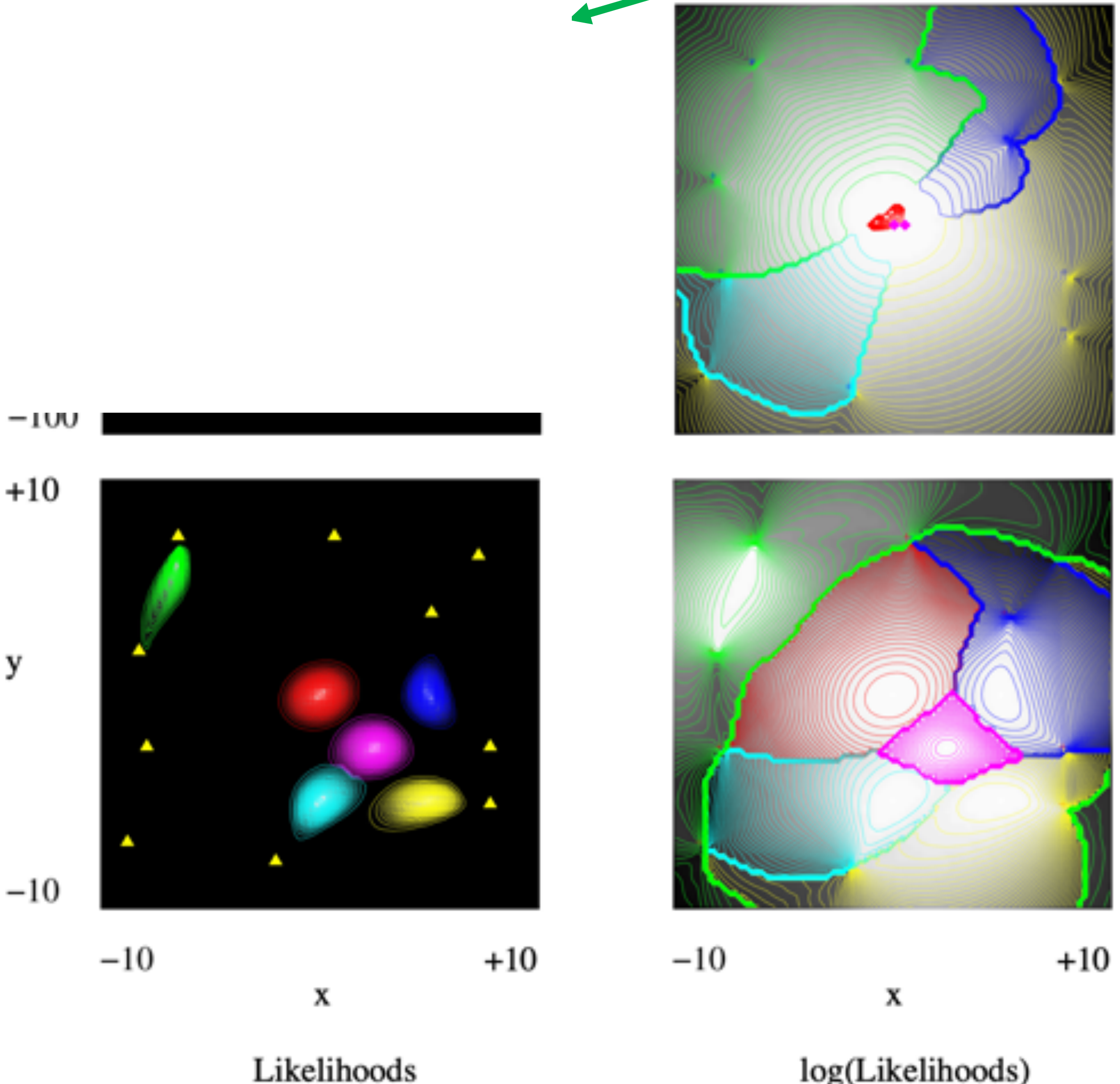
Glossary



Terms used here	Terms used in DNN papers
Current feature vector (yellow dot)	"feature activations in the last fully connected layer" (Krizhevsky et al, 2012)
Stored feature vectors (red dots)	Rows in W where $f(x_i, W) = W x_i$ (Karpathy tutorial)
Voronoi cells (assumes feature vectors are same length, e.g. unit vectors)	"Computing similarity by using Euclidean distance between two 4096-dimensional, real-valued vectors" (Krizhevsky et al, 2012)
Dimensionality of state space Basis vector	As above ($\mathbb{R}^{4096}$) One of the 4096 dimensions
Number of stored feature vectors	"For multi-class classification, we have $y_n \in [K] := \{1, 2, \ldots, K\}$ and K is the number of classes" (Li et al 2019, https://arxiv.org/pdf/1808.05385.pdf)
Policy	$\Pi(a\|s,g)$
Number of actions can be very small	e.g. "binary labels $y_n \in \{-1, 1\}$"

Hierarchical, compositional encoding of location





ws

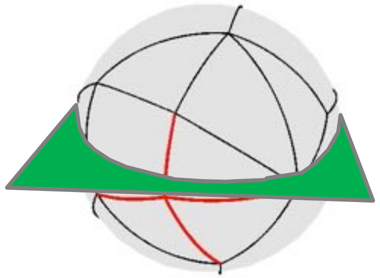
n the
earth.
the
tion

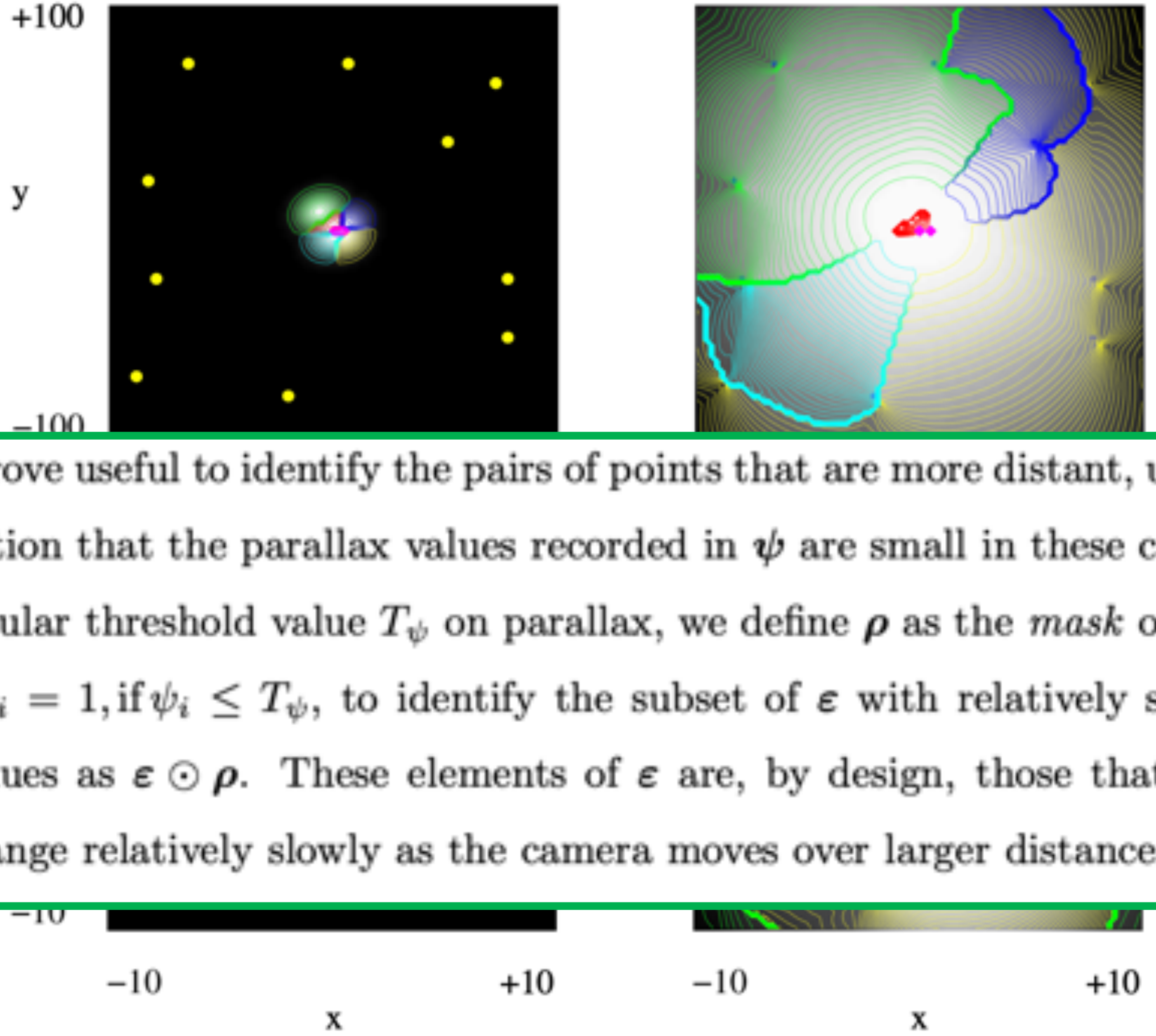
re

nal
-
ns of
)

Hansard

Hierarchical, compositional encoding of location





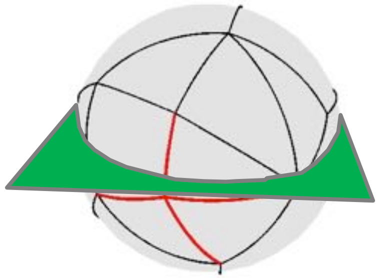
Likelihoods

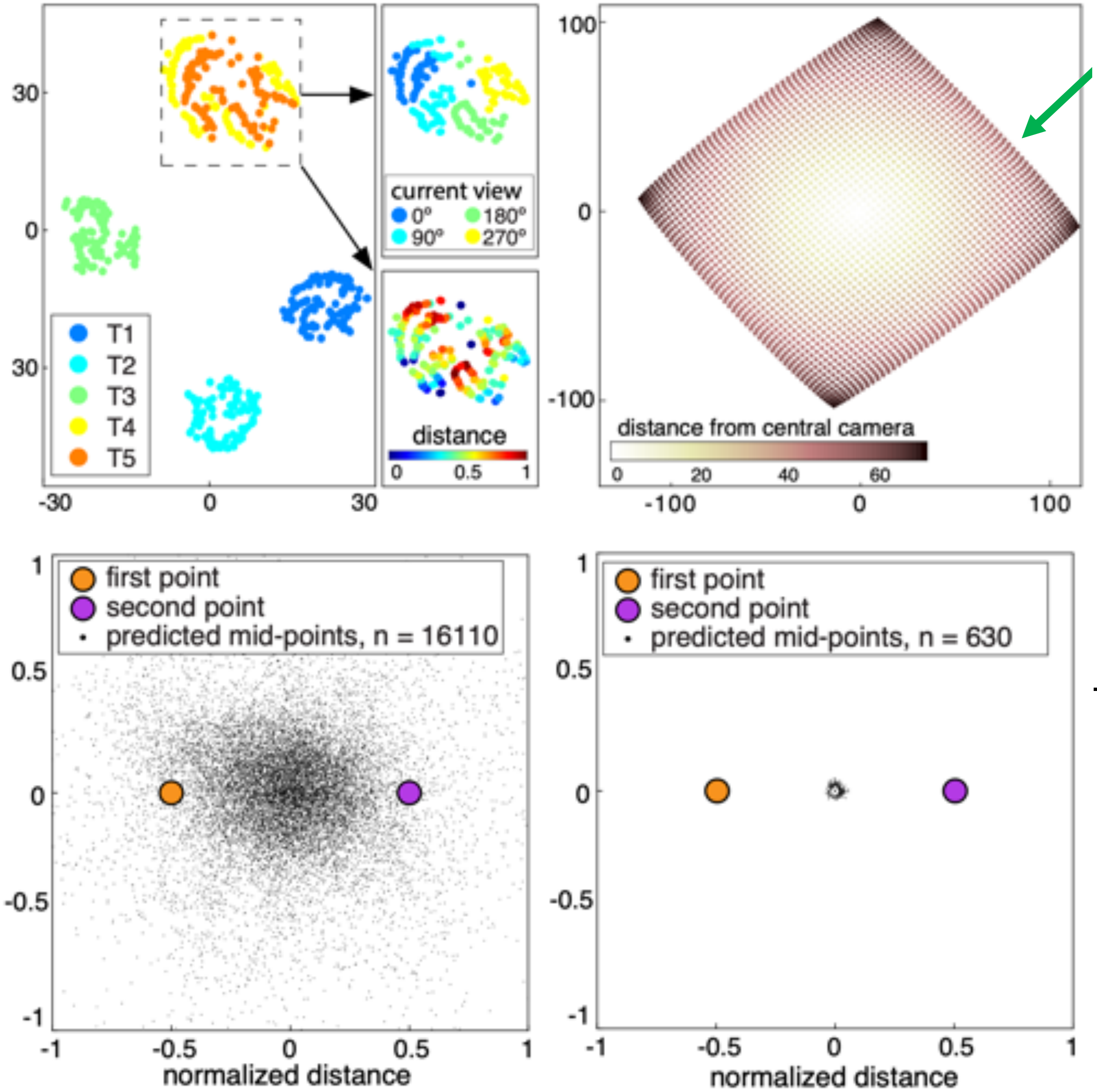
log(Likelihoods)

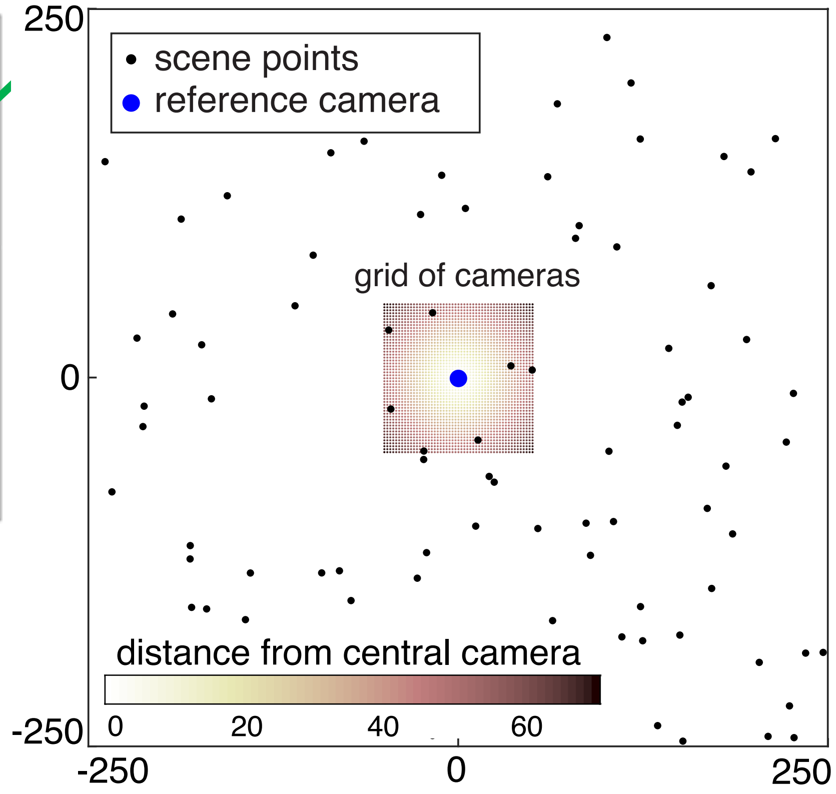
It will prove useful to identify the pairs of points that are more distant, using the observation that the parallax values recorded in $\boldsymbol{\psi}$ are small in these cases. For a particular threshold value T_ψ on parallax, we define $\boldsymbol{\rho}$ as the *mask* on $\boldsymbol{\psi}$, such that $\rho_i = 1, \text{if}\, \psi_i \le T_\psi$, to identify the subset of $\boldsymbol{\varepsilon}$ with relatively small parallax values as $\boldsymbol{\varepsilon} \odot \boldsymbol{\rho}$. These elements of $\boldsymbol{\varepsilon}$ are, by design, those that are likely to change relatively slowly as the camera moves over larger distances.

- The 'visual image' (ε) is a vector listing all the angles (N^2) between all N visible points.
- For each angle, we also calculate a measure of parallax. If P_i and P_j are both distant, this parallax will be small.

Hierarchical, compositional encoding of location







compositional (hypotheses 'split' into sub-categories as the dimensions of the representation increase)

- Muryy et al (2020) and Glennerster, Hansard and Fitzgibbon (2001)