Bayesian Brain Cosmo 2014

Paul R Schrater,

Departments of Psychology and Computer Science, University of Minnesota

Bayesian Brain?

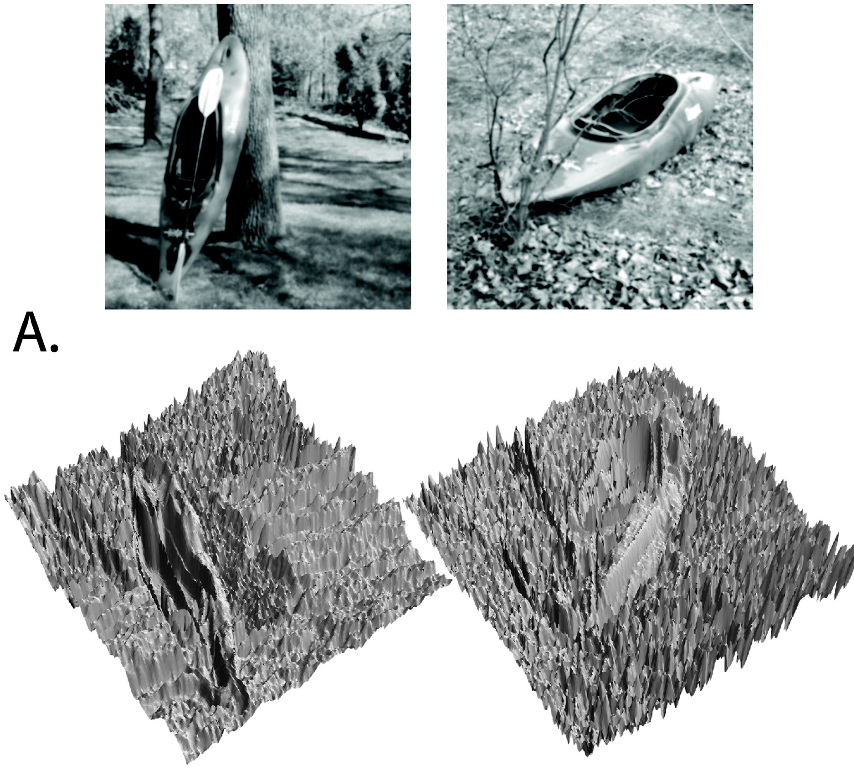
- Ubiquity of sensory uncertainty: - e.g. mapping of 3D objects to 2D image
  - sensory information is impoverished relative to problems human solve
  - intrinsic limitations of the sensory systems (e.g. number and quality of receptors in the retina)
  - neural noise

→ multiple interpretations about the world are possible;

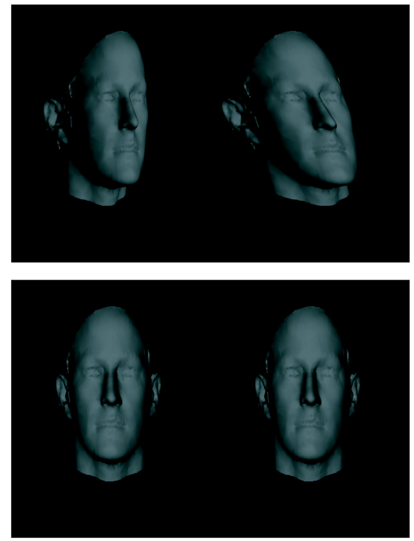
- The brain must represent and process uncertainty to guide actions, allocate time and resources (e.g. attention, computation, sensory processing)

Complex Perceptual Problems are ambiguous

A.

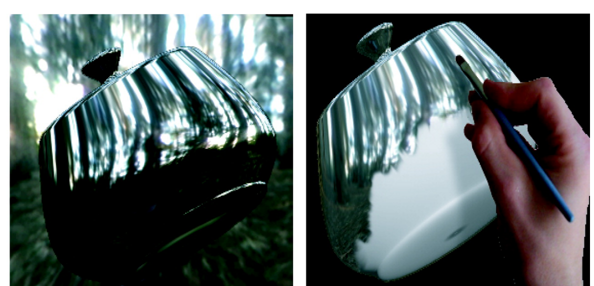


B.



C.

D.



Recognition, Shape, Material

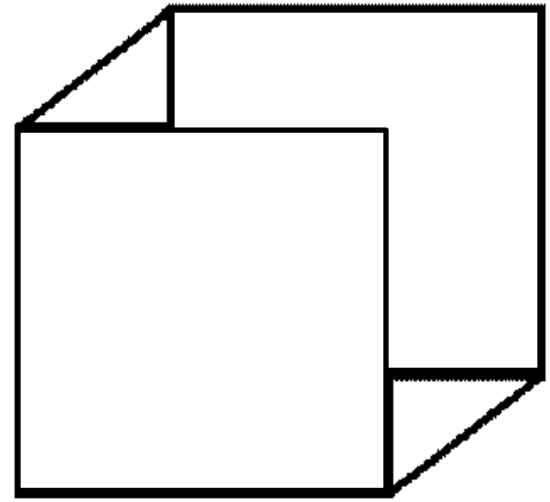
A) Invariance to Pose, lighting, and shading.

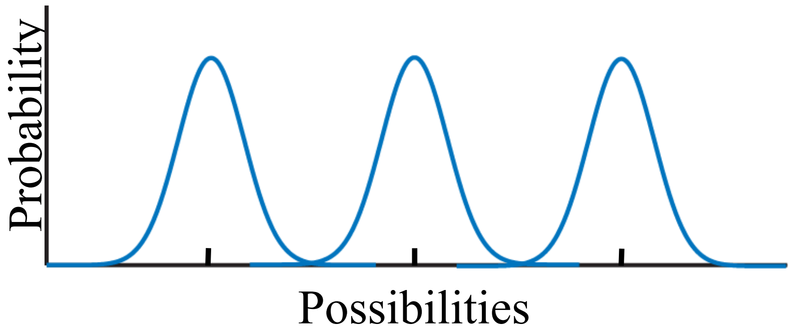
B) Single image ambiguity: Bas relief transform of shape lighting (viewpoint

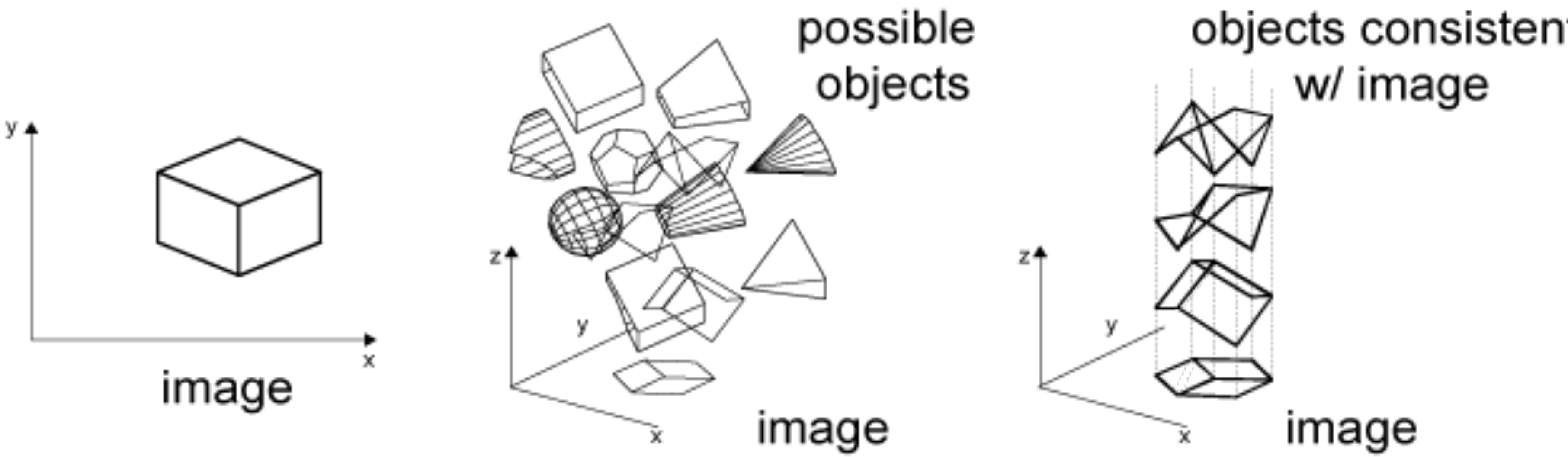
C,D) Reflectivity vs. paint

Ambiguity:

can be characterized by a **probability distribution** for which multiple possibilities have equal/similar probability.

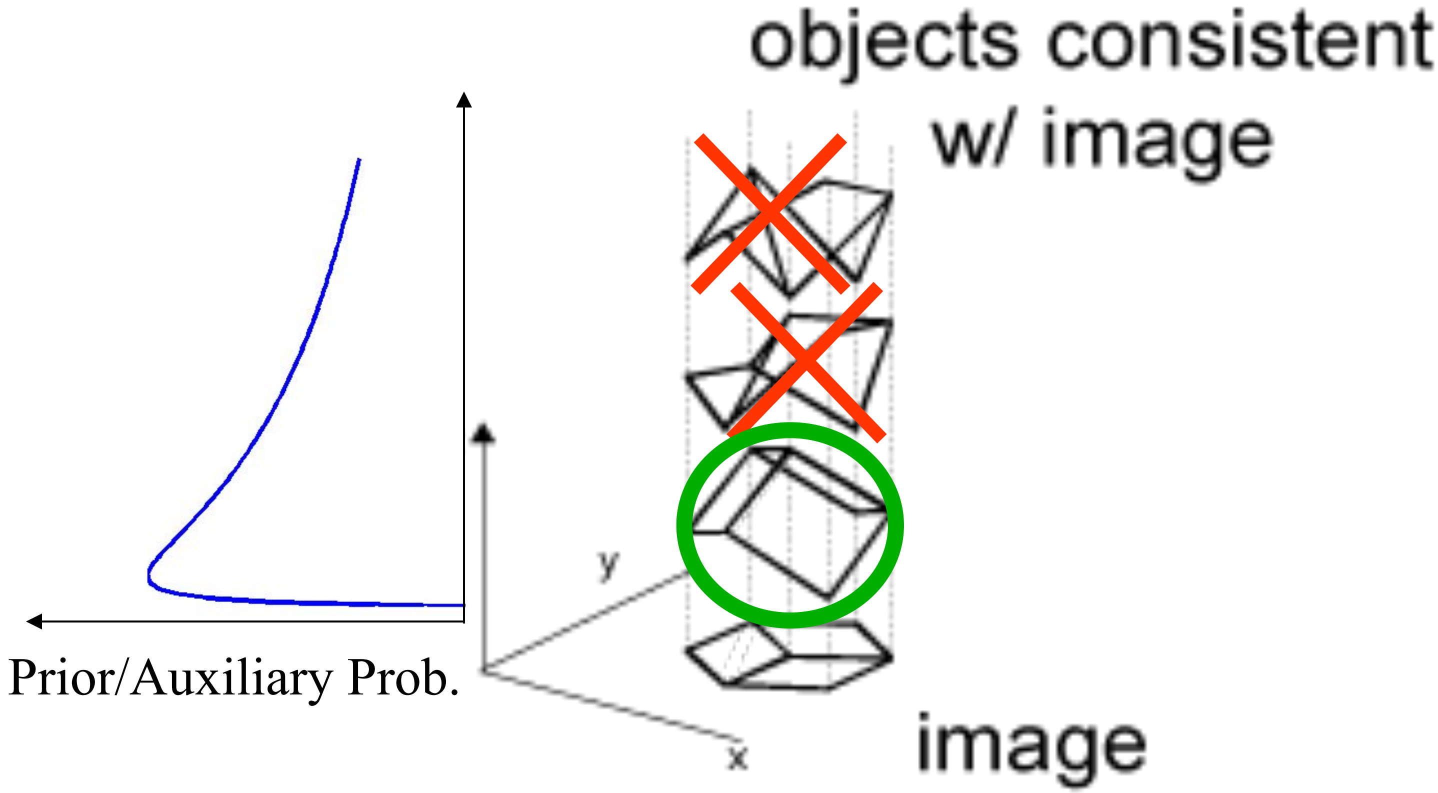






Overcoming ambiguity requires applying additional ***knowledge***

Prior knowledge and *auxiliary information* can further disambiguate candidate scene interpretations



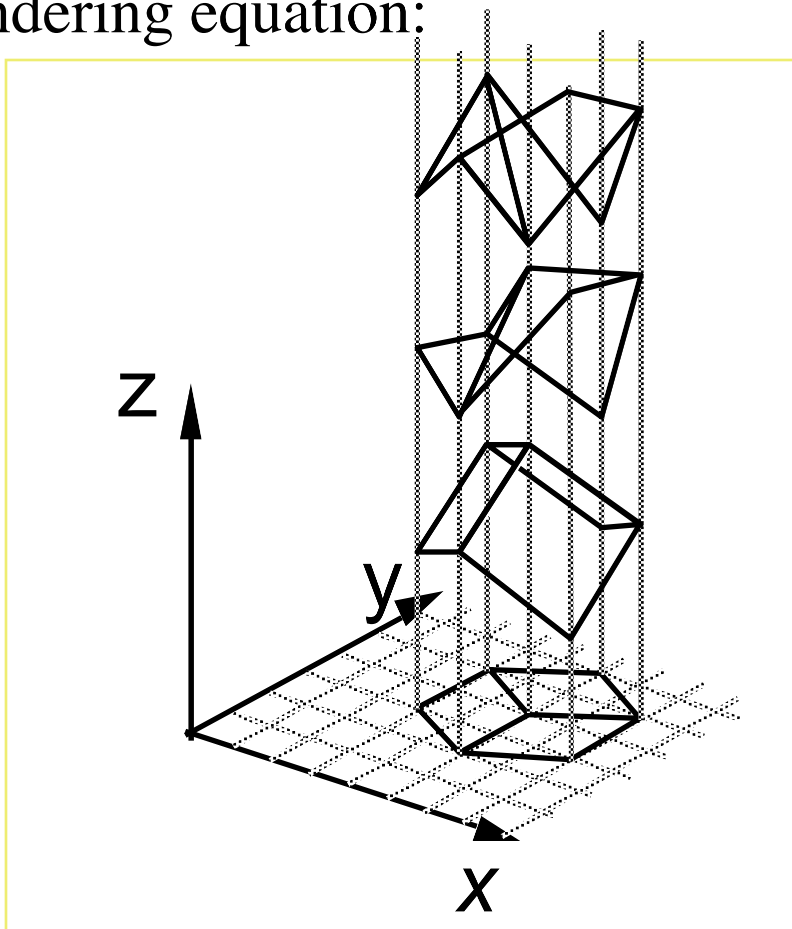
Outline

- How do we specify/describe what we mean by generative knowledge
- What kinds of generative knowledge do people use?
- How do we test for its use?

Forward models for perception:

Built in knowledge of image formation

Images are produced by physical processes that can be modeled via a rendering equation:



$I = f(A, L, V) = f(scene)$

$A =$ object attributes

$L =$ description of the scene lighting

$V =$ viewpoint and imaging parameters (e.g. focus)

Modeling rendering probabilistically:

Likelihood: $p(I \mid scene)$

e.g. for no rendering noise

$$p(I \mid scene) = \delta\big(I - f(scene)\big)$$

How do we describe the other kinds of generative knowledge?

Example

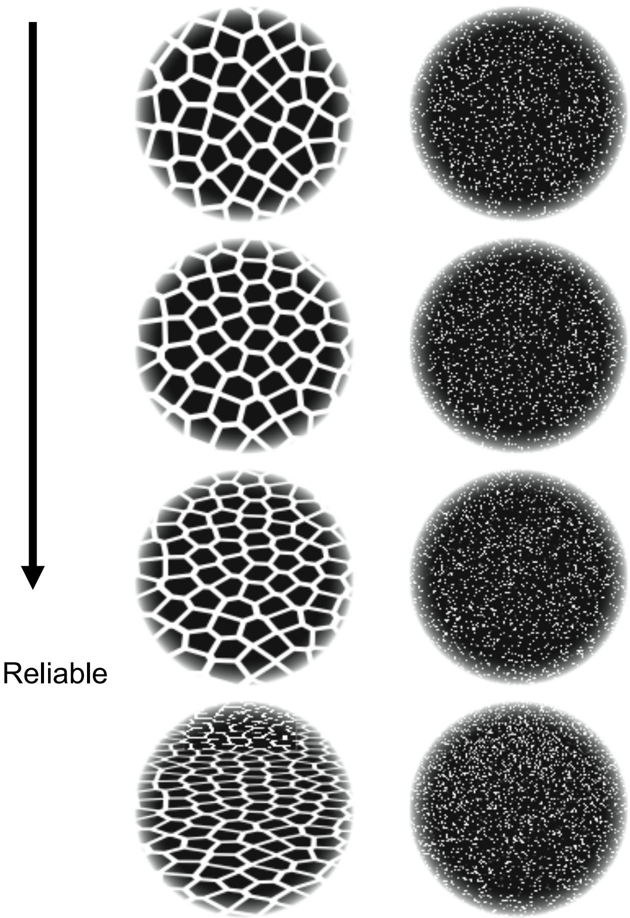
Texture information

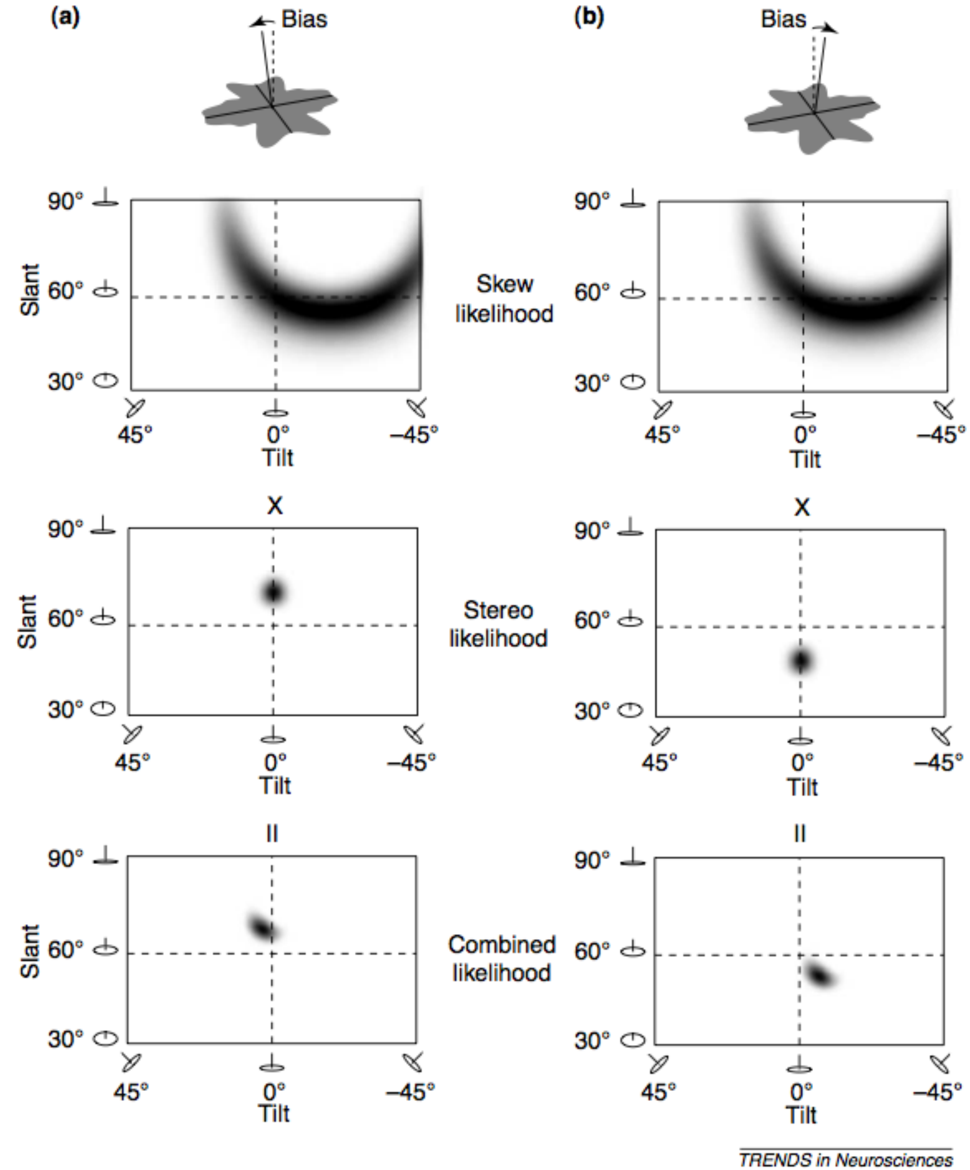
Least Reliable

Most Reliable

Binocular information

Equally reliable





Bayesian Networks: Modeling complex inferences

This model represents the decomposition:

$$P(X_1, X_2, X_3, X_4) = P(X_4 \mid X_2)\, P(X_3 \mid X_1, X_2)\, P(X_1) P(X_2)$$

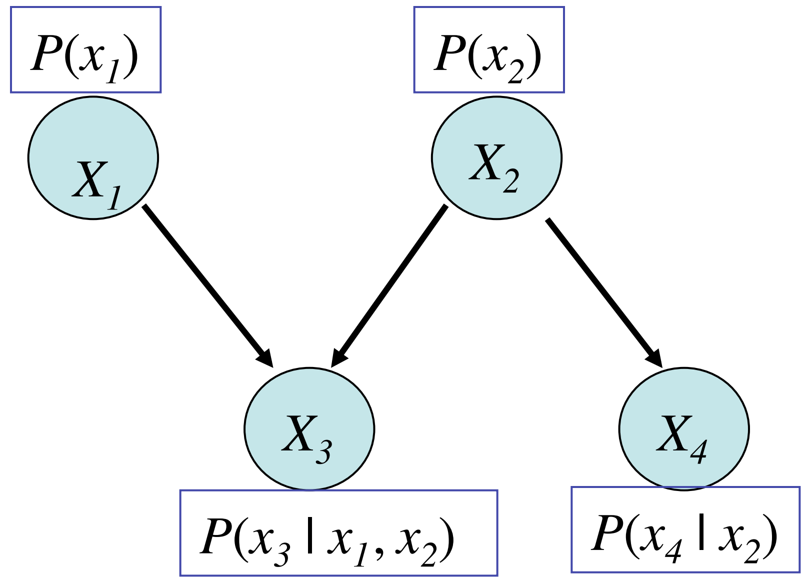
Nodes: random variables

$$X_1, \ldots, X_4$$

Each node has a conditional probability distribution

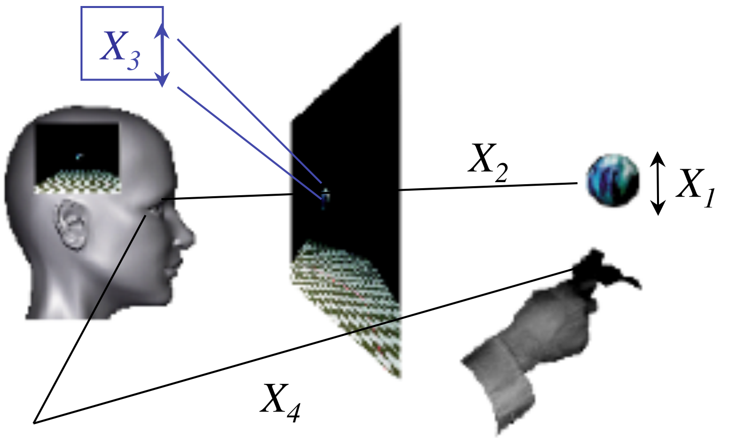
Links: direct dependencies

Data: observations of X_3 and X_4

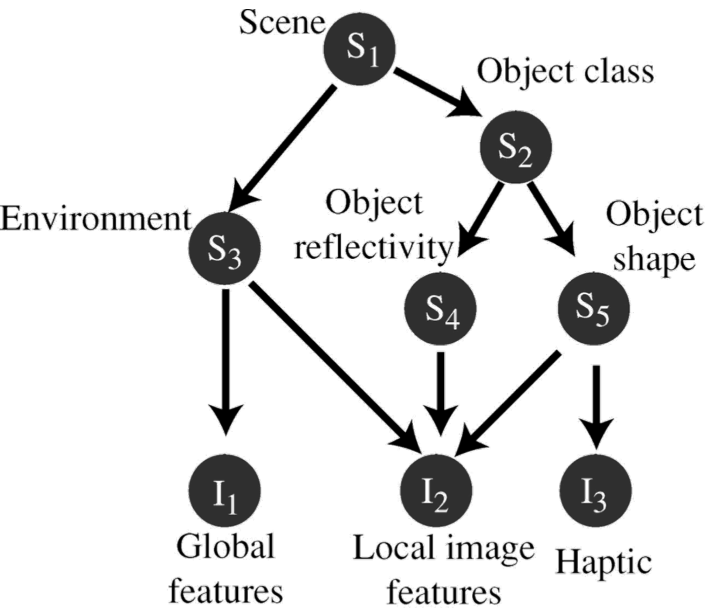


EXAMPLE

X_1 object size
X_2 object distance
X_3 image size
X_4 "felt" distance

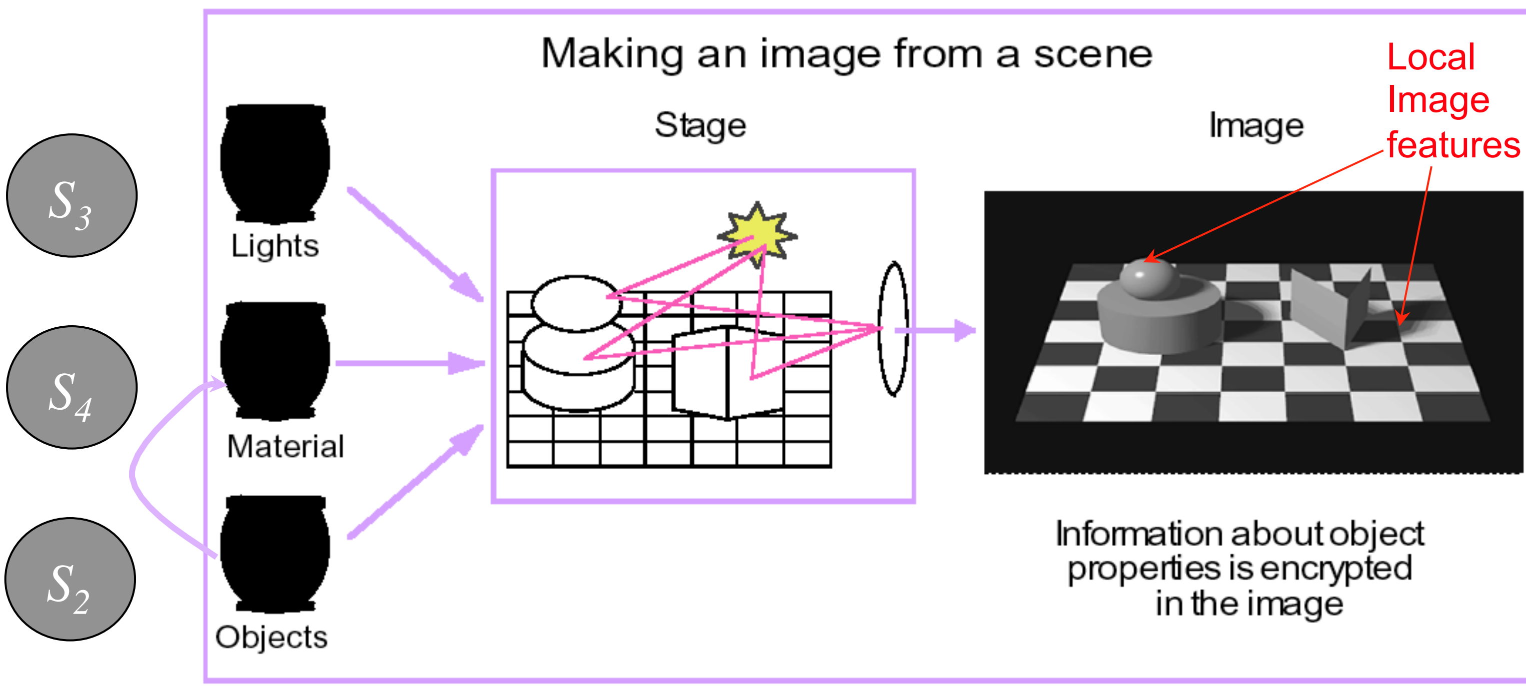


Generative Knowledge



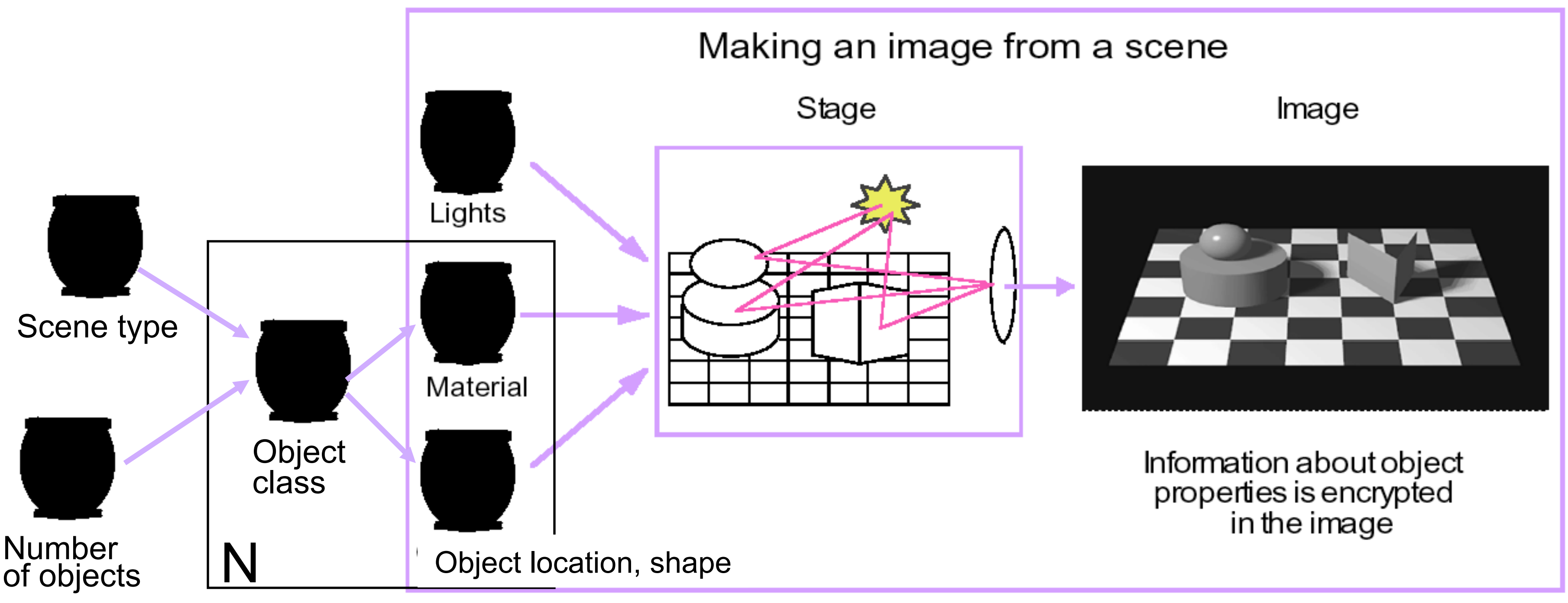
Knowledge about the dependencies between variables can be represented by a graphical model ***in two ways***

1) ***As a connective graph (right)***
2) ***As an inferential graph (explained next)***

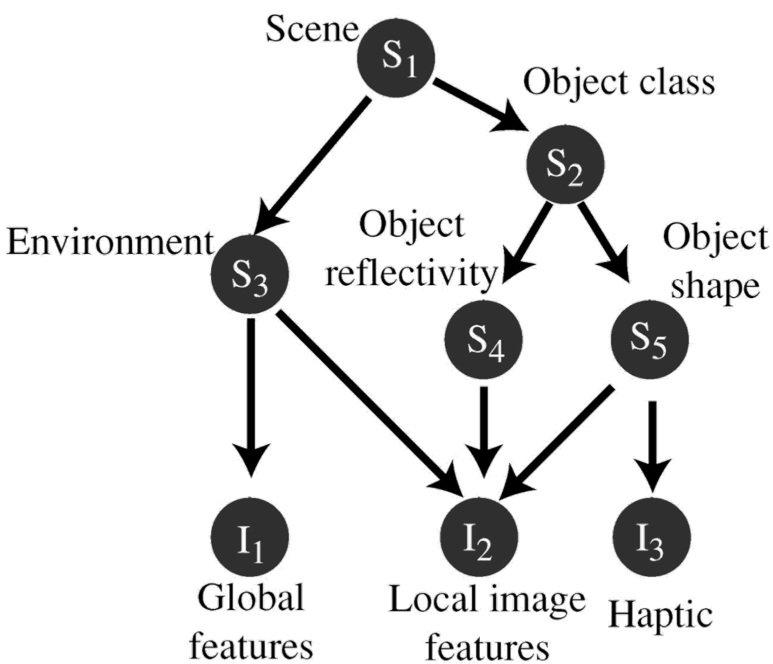


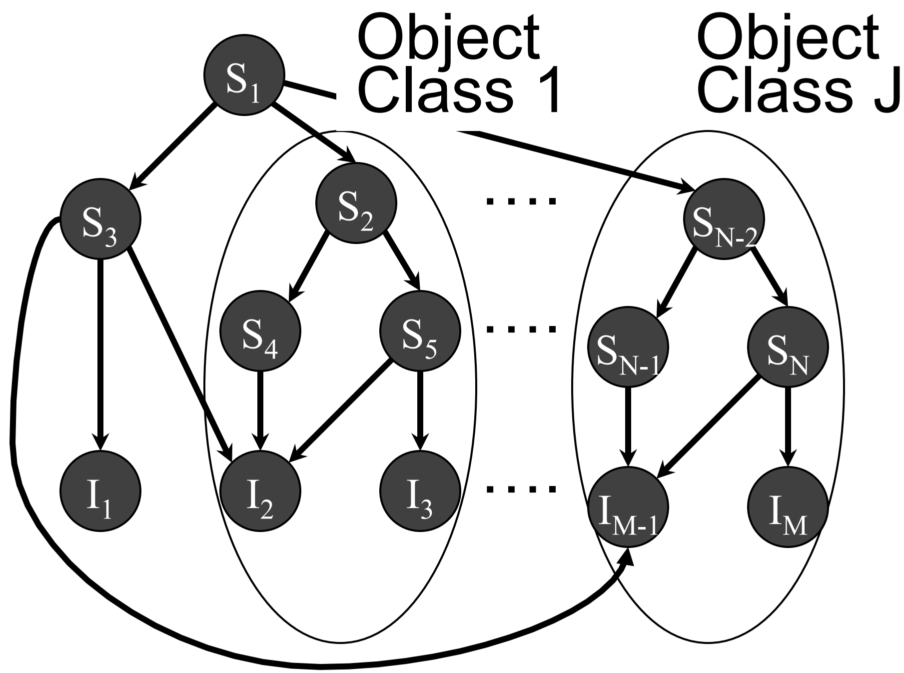
Forward Graphics Analogy

- Sample a scene type
- Sample N object classes
- Sample Objects from each class (locations and attributes for each object)
- Sample rendering variables (lights, viewpoint)
- Sample image features from rendered scene



The graphical model for scene ***inference*** requires different structure for each scene

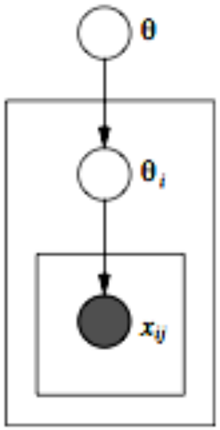




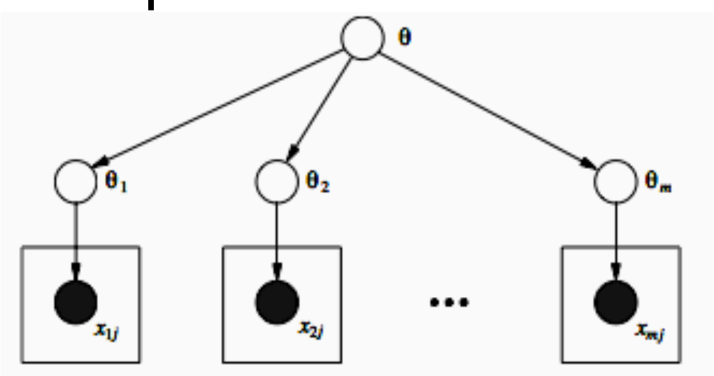
However, this structure is part of what we INFER in scene perception!

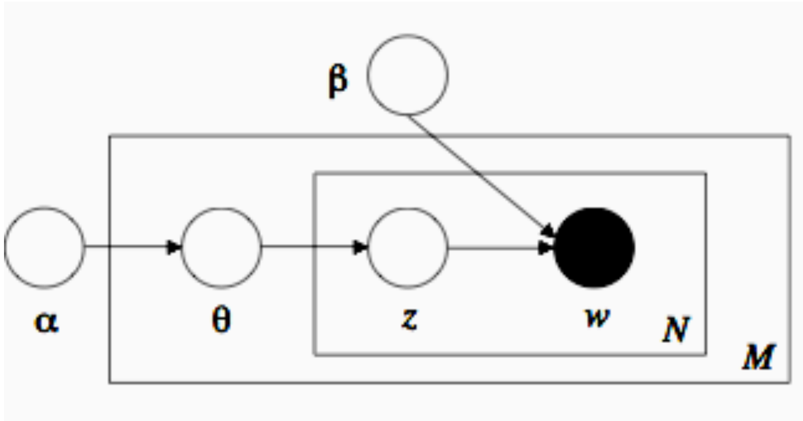
Non-parametric Bayes

Plate notation:



Is equivalent to:

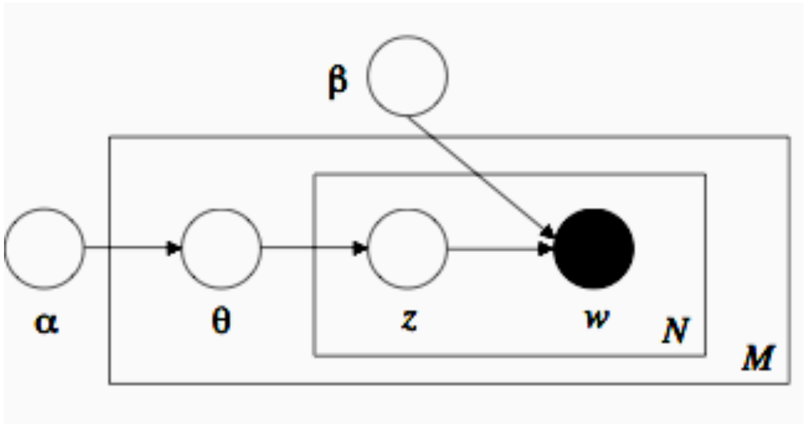




- *Random variables for document clustering*
  - *A word is a multinomial random variable w*
  - *A topic is a multinomial random variable z*
  - *A document is a Dirichlet random variable θ*

Treats number of words and topics as *random variables*

An analogy

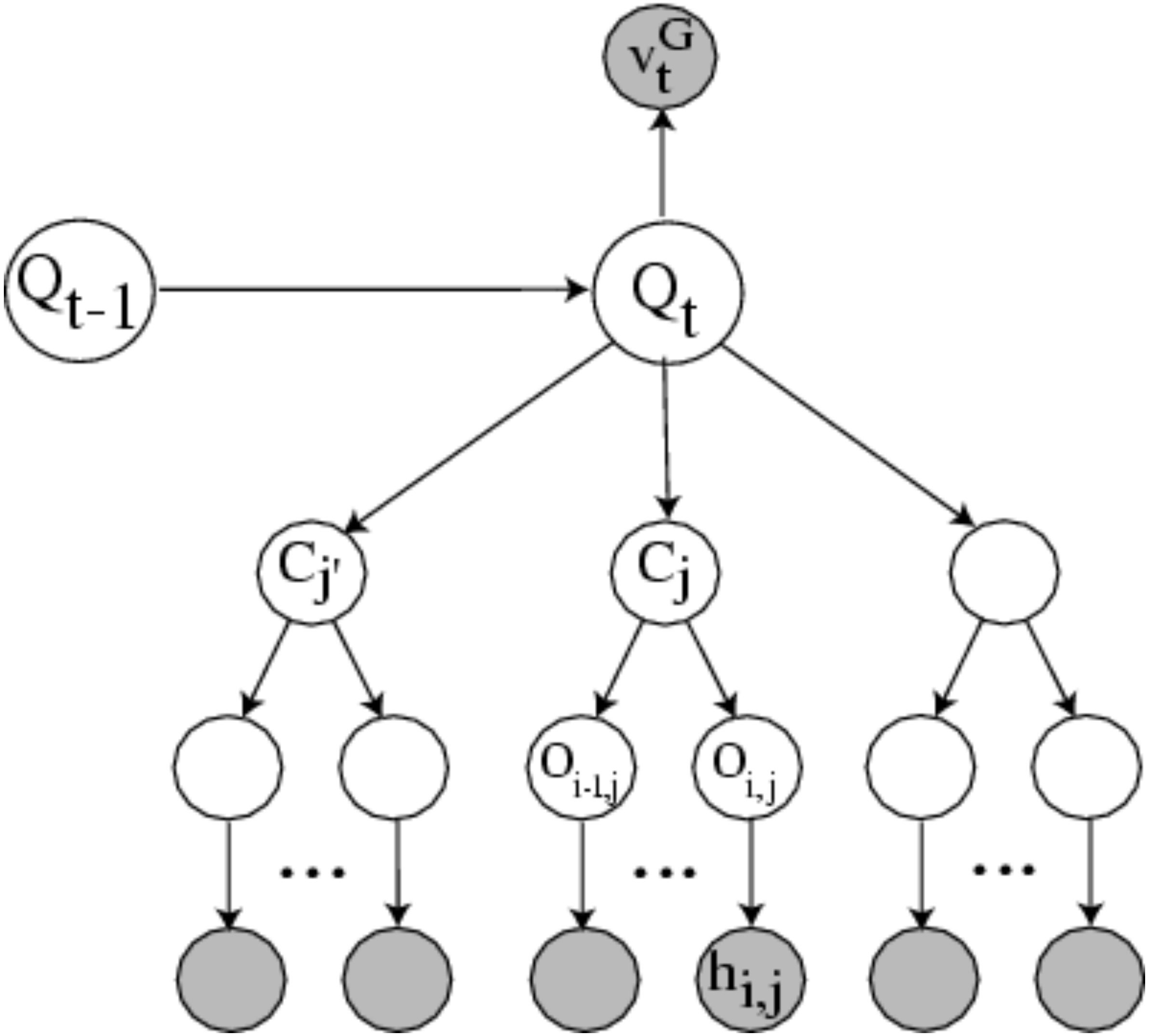


- *Random variables for document clustering*
  - *A word is a multinomial random variable w*
  - *A topic is a multinomial random variable z*
  - *A document is a Dirichlet random variable θ*
- *Random variables for scene inference*
  - *An object class is a multinomial random variable w*
  - *A subscene is a multinomial random variable z*
  - *A scene is a Dirichlet random variable θ*

Non-Parametric Bayes Model

- Parametric vs. non-parametric Bayes
  - Parametric: Fixed parameterization of the prior
    - Needs prior on space of all possible scenes
    - Difficult to learn models (curse of dimensionality)
    - Has generated skepticism of Bayes for vision
  - Non-parametric:
    - Developed in response to limitations of parametric approach
    - ***Only generates scene graph during inference***
    - Needs prior on scene construction (not scenes)
    - Parameters naturally coupled, reducing dimensionality
    - Increasingly used for “hard problems” in machine learning
    - Examples: Latent Dirichlet allocation, Chinese restaurant process, Indian buffet process, etc.

Computer vision architecture (Sudderth et al, 2006)



Visual "gist" observations

Scene category

kitchen, office, lab, conference room, open area, corridor, elevator and street.

Object class

Particular objects

Local image features

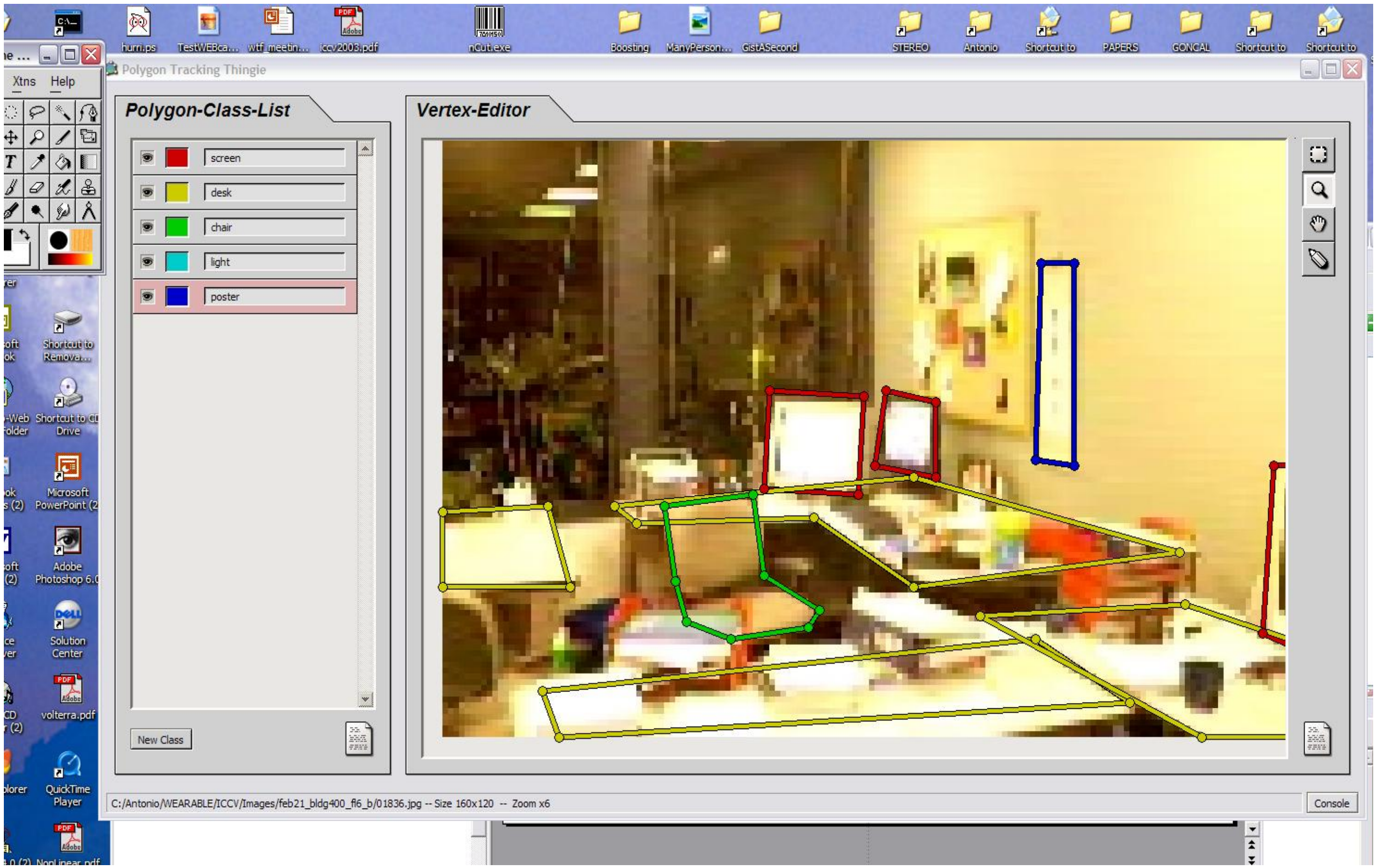
“Top-down” information: a representation for image context

Images

80-dimensional representation

Credit: Antonio Torralba

"Bottom-up" information: labeled training data for object recognition.

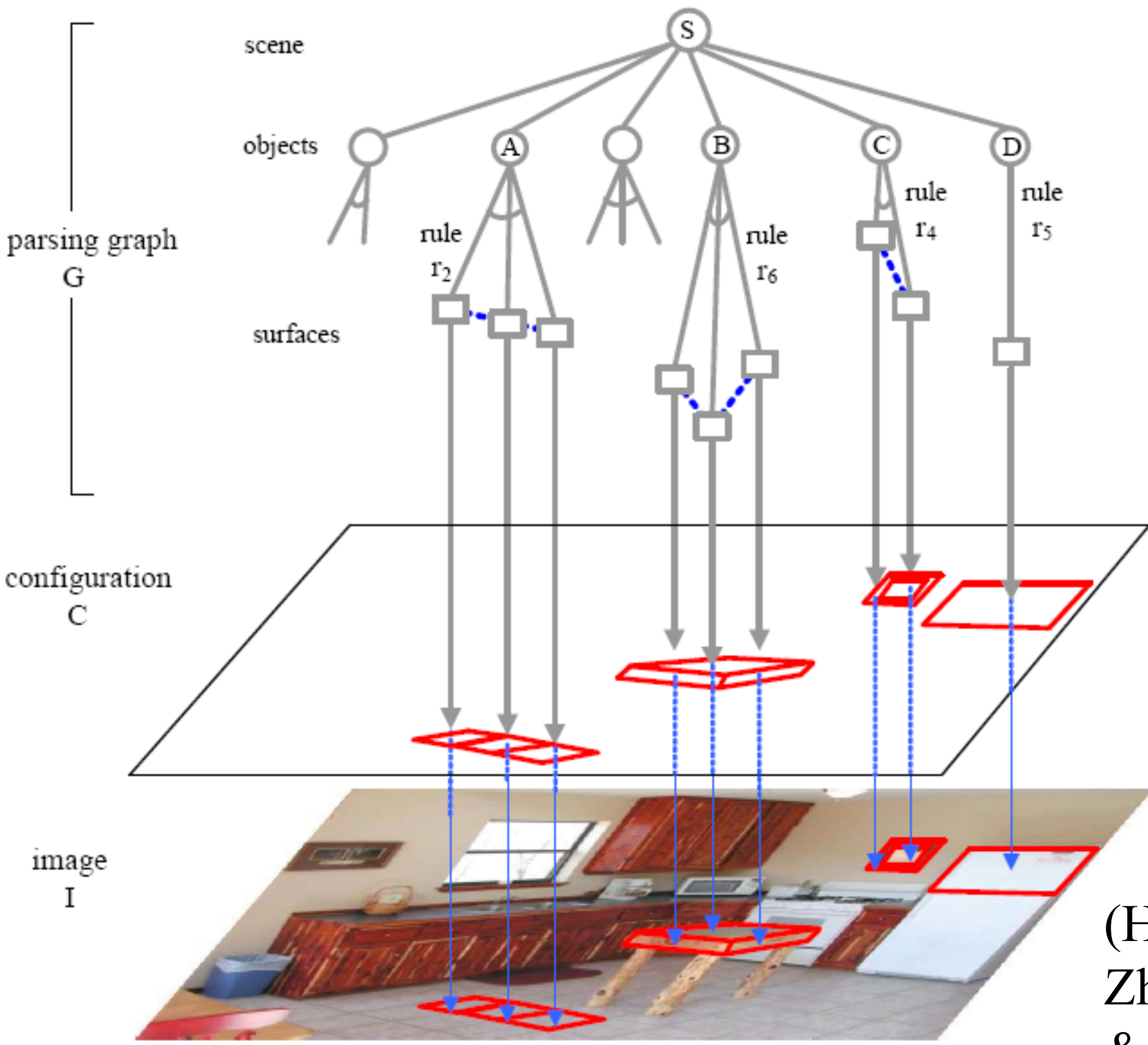


•Hand-annotated 1200 frames of video from a wearable webcam

•Trained detectors for 9 types of objects: bookshelf, desk, screen (frontal) , steps, building facade, etc.

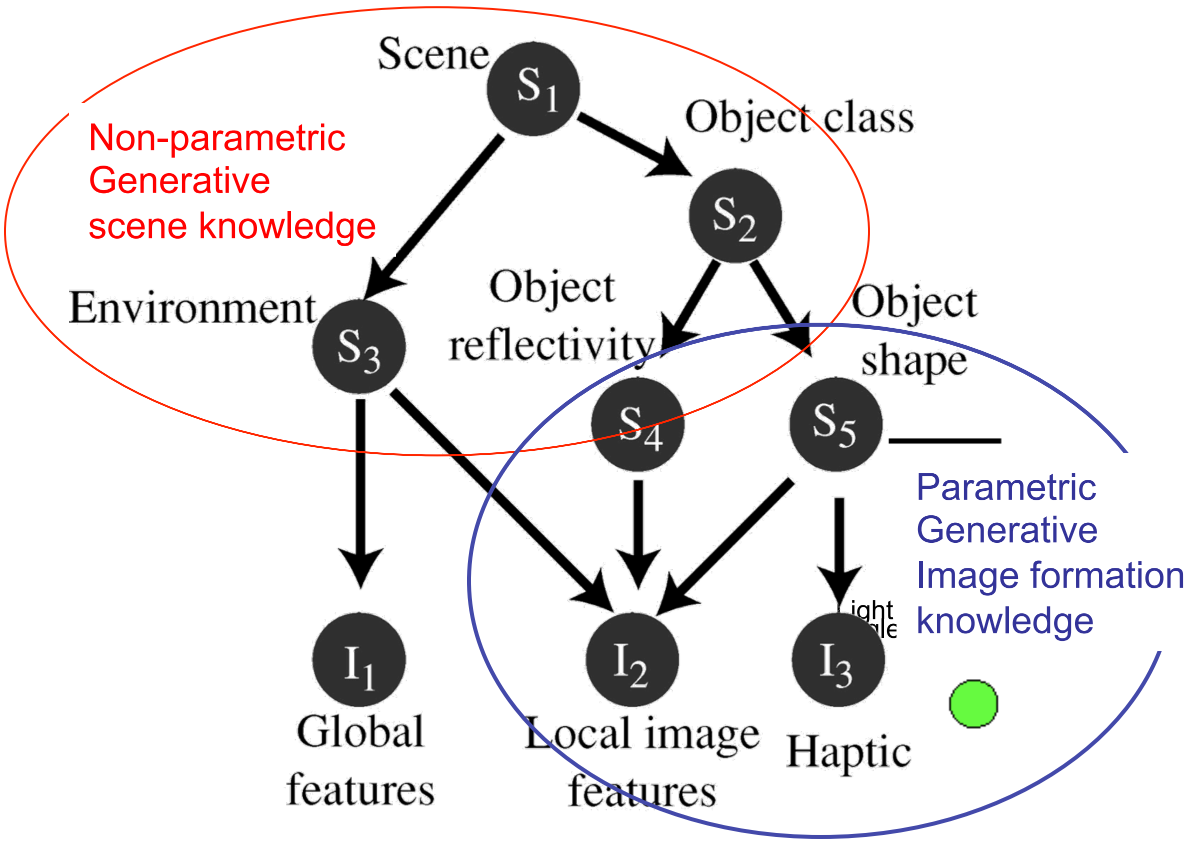
•100-200 positive patches, > 10,000 negative patches

Vision as probabilistic parsing



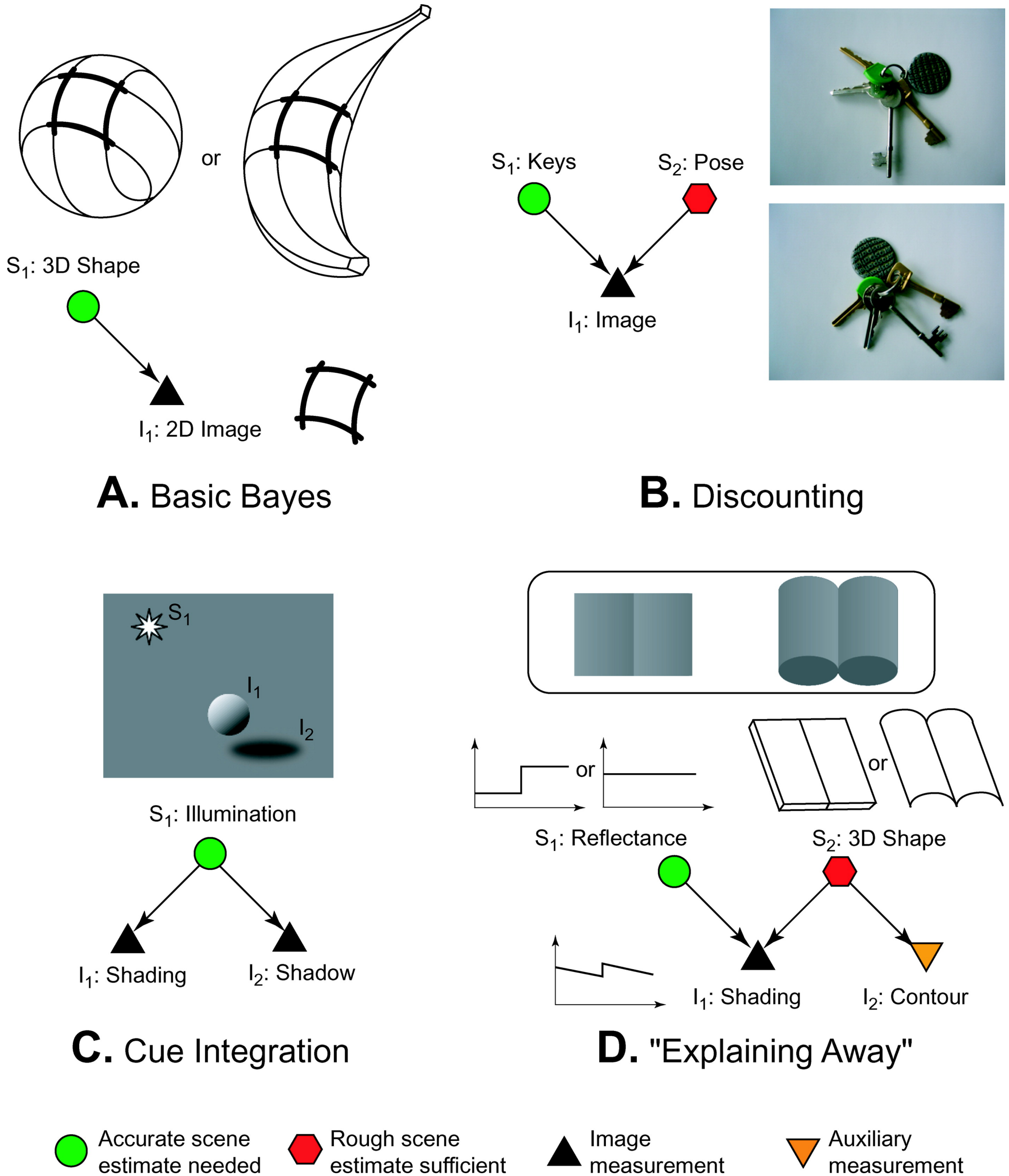
(Han & Zhu, 2006; c.f., Zhu, Yuanhao & Yuille NIPS 06)

Kinds of Generative Model



- **Scene**: *type* puts distributions on constituents, layout, lighting, etc
- **Object class**: puts distribution on object attributes
- **Image formation**: puts distribution on image measurements given objects
- **Dynamics model:** *transformations*

Image formation generative knowledge

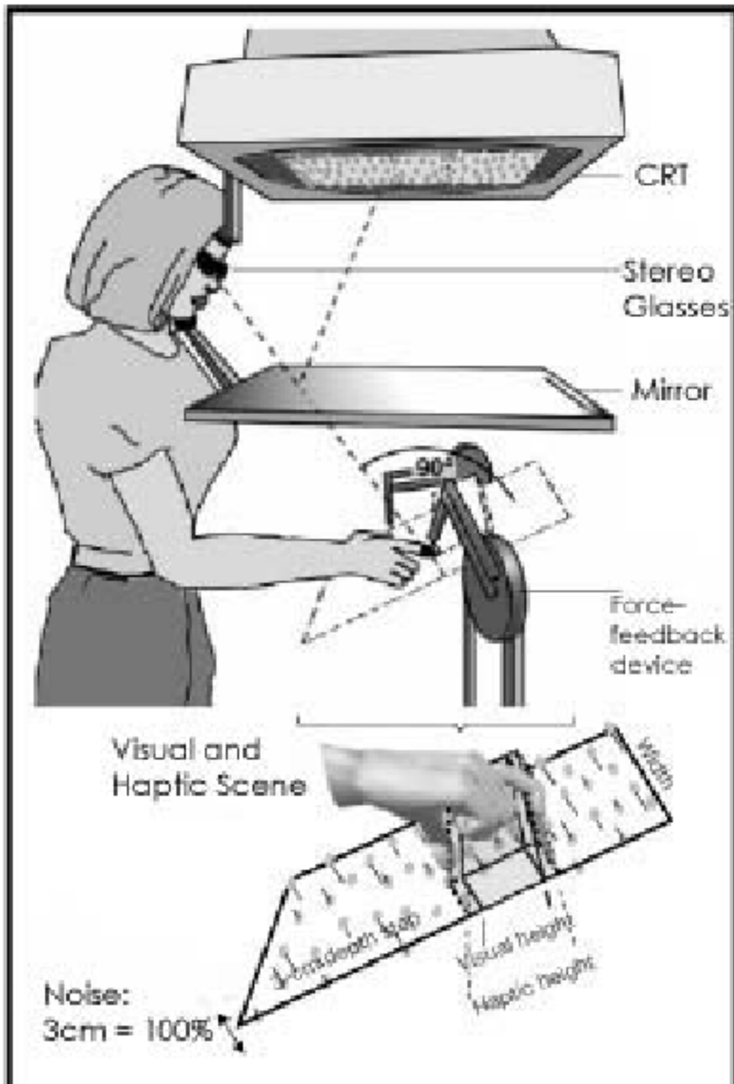


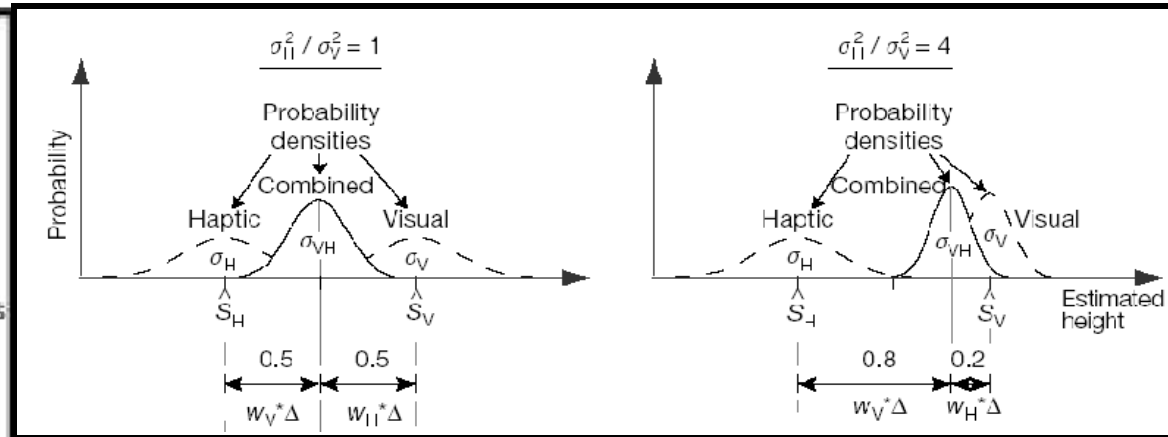
- Different relationships between image measurements and object attributes lead to different inference problems.
- Object property inference frequently requires knowing aspects of the scene (how many objects are present, illumination, object layout and pose, etc)

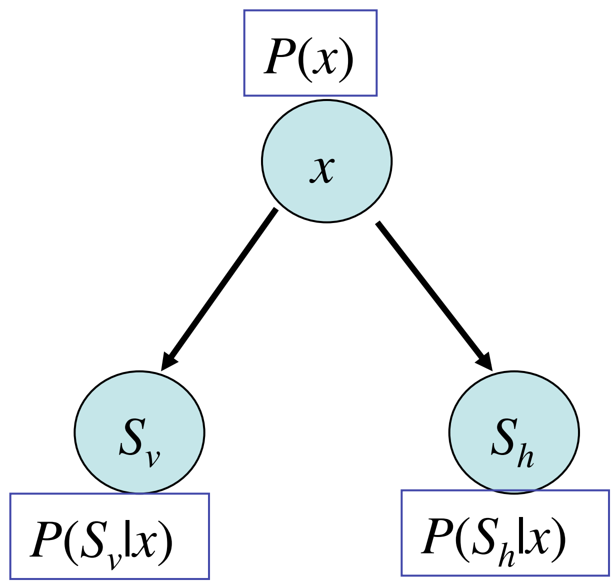
Testing Image generative knowledge

- How do we test whether people understand the relationship between object attributes and image measurements?
- Difficulty: Experimental design must eliminate *ambiguity in scene perception (number of objects, lighting, etc).*
  - (otherwise not studying image formation generative knowledge at all)
- Case studies:
  - Cue integration (quantitative)
  - Explaining away (previously qualitative)

Cue Integration

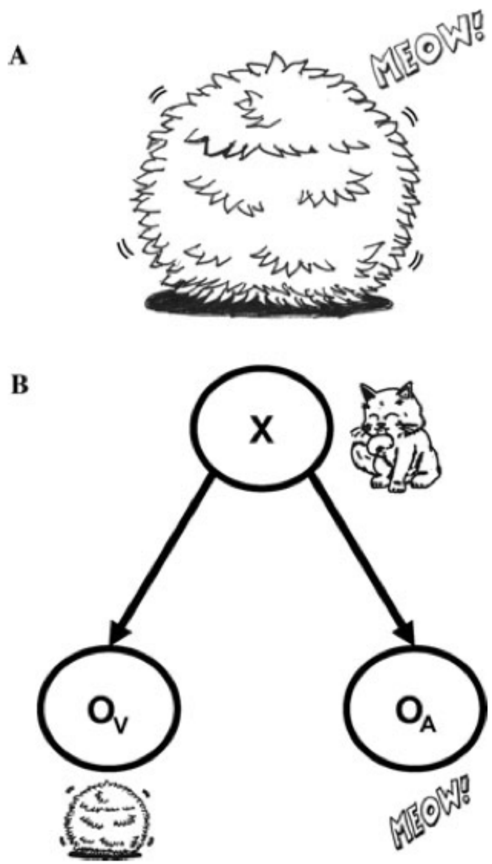




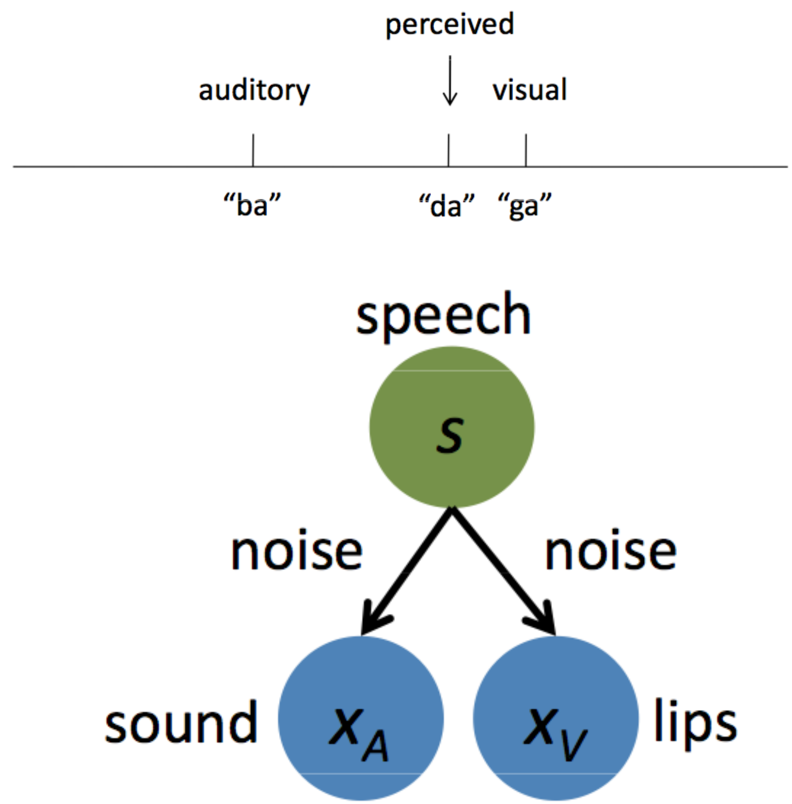


Examples

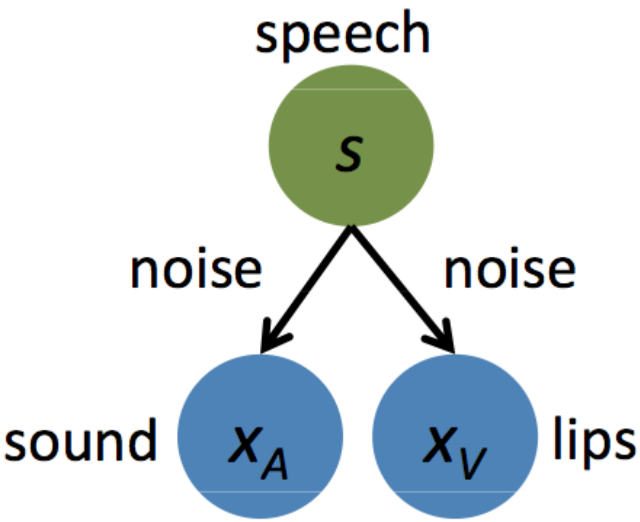
Audio-visual localization



McGurk effect



McGurk Math

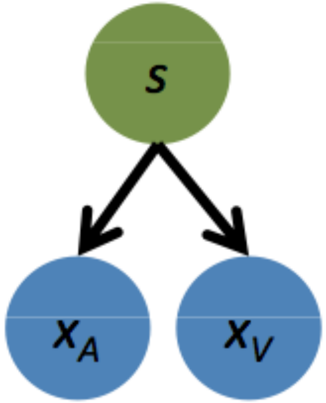


- **Hypotheses**:
  "ba", "ga", "da", other syllables
- (A) auditory evidence for "ba"
- (V) visual evidence for "ga"
- The brain computes
  p(syllable | A,V)

What is the posterior over *s*, given this generative model?

$$p(s \mid x_A, x_V) \propto p(x_A, x_V \mid s)\, p(s)$$
$$= p(x_A \mid s)\, p(x_V \mid s)\, p(s)$$

Conditional independence → *multiplying* likelihood functions



$$p(s \mid x_A, x_V) \propto p(x_A, x_V \mid s)\, p(s)$$
$$= p(x_A \mid s)\, p(x_V \mid s)\, p(s)$$

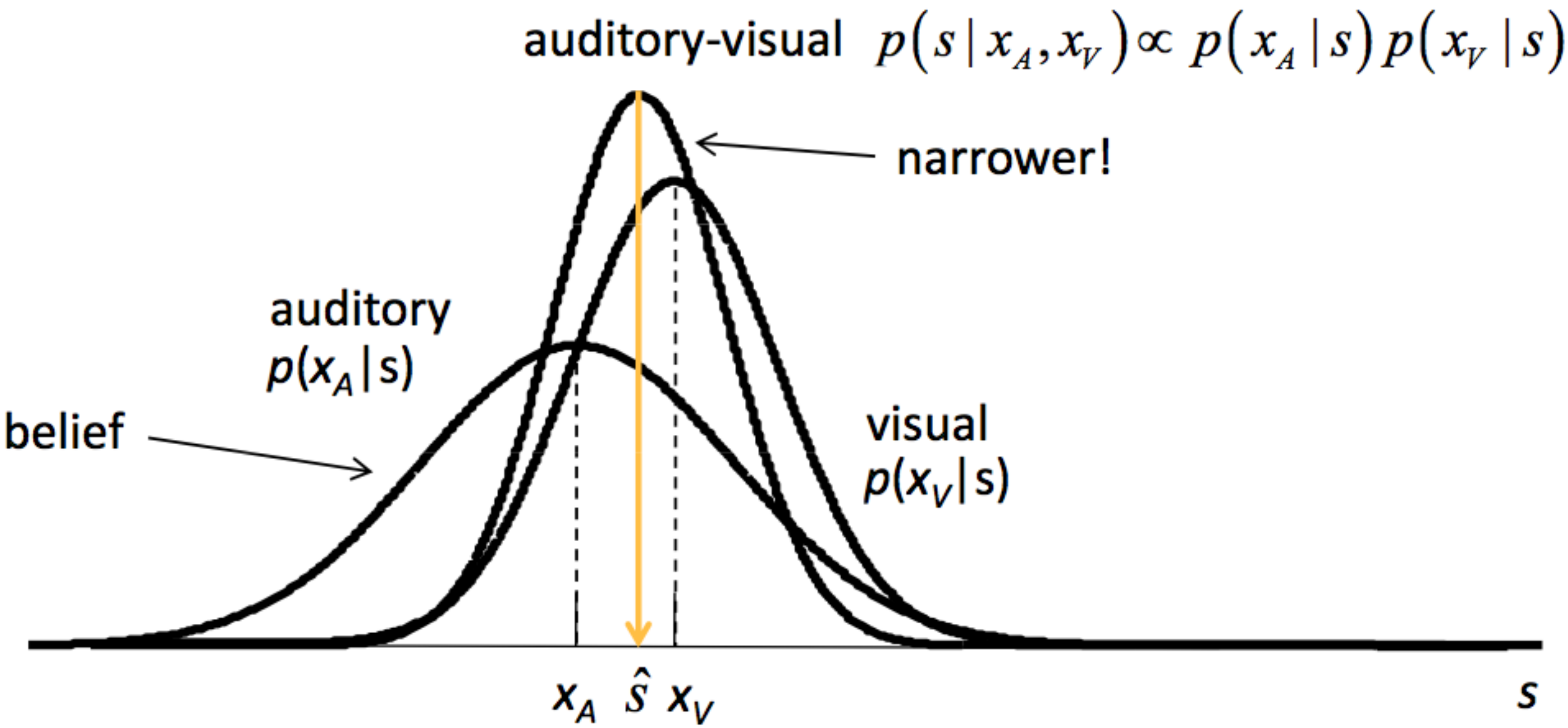
Assumptions about these distributions:

$$p(x_A \mid s) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{(x_A - s)^2}{2\sigma_A^2}}$$

$$p(x_V \mid s) = \frac{1}{\sqrt{2\pi\sigma_V^2}} e^{-\frac{(x_V - s)^2}{2\sigma_V^2}}$$

$$p(s) = \text{constant}$$

Multiplying likelihoods



Classic Cue Combination

Given $p\left(s \mid x_A, x_V\right) \propto p\left(x_A \mid s\right) p\left(x_V \mid s\right)$

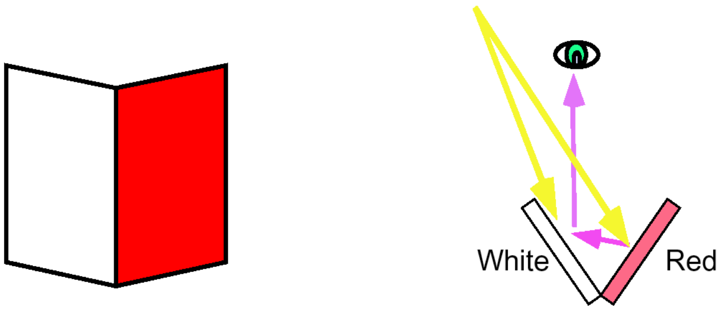
$$p\left(x_A \mid s\right) = \frac{1}{\sqrt{2\pi\sigma_A^2}} e^{-\frac{\left(x_A - s\right)^2}{2\sigma_A^2}} \qquad p\left(x_V \mid s\right) = \frac{1}{\sqrt{2\pi\sigma_V^2}} e^{-\frac{\left(x_V - s\right)^2}{2\sigma_V^2}}$$

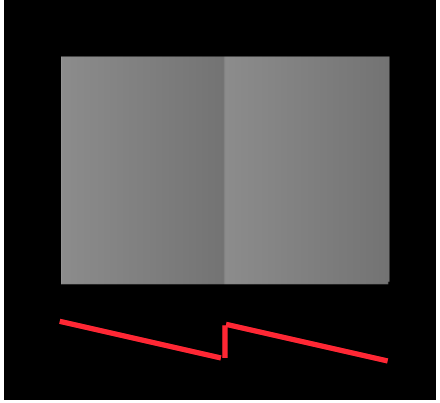
show that $p(s \mid x_A, x_V)$ is a normal distribution over s, with mean $\hat{s} = \frac{w_A x_A + w_V x_V}{w_A + w_V}$

where $w_A = \frac{1}{\sigma_A^2} \quad w_V = \frac{1}{\sigma_V^2}$

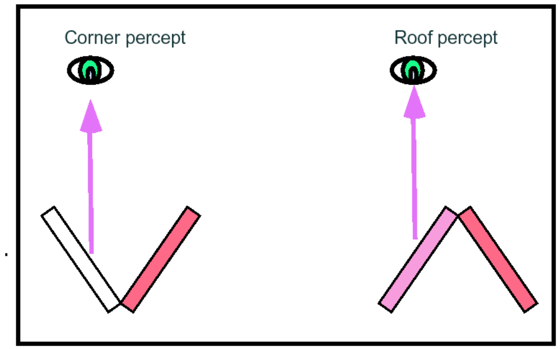
Explaining away

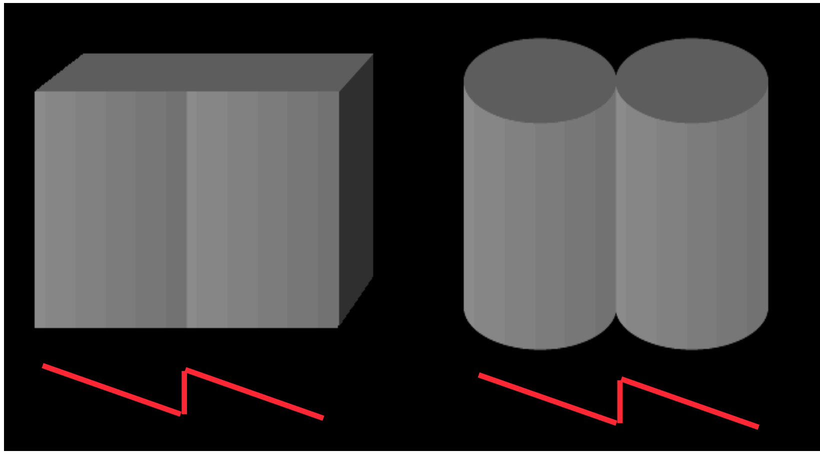
A.

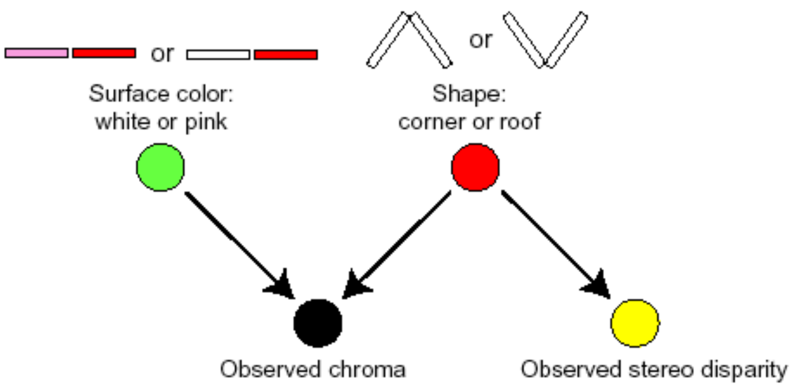


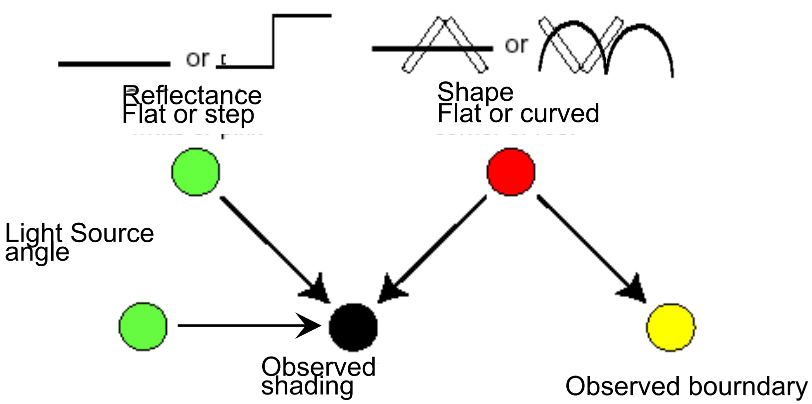


B.







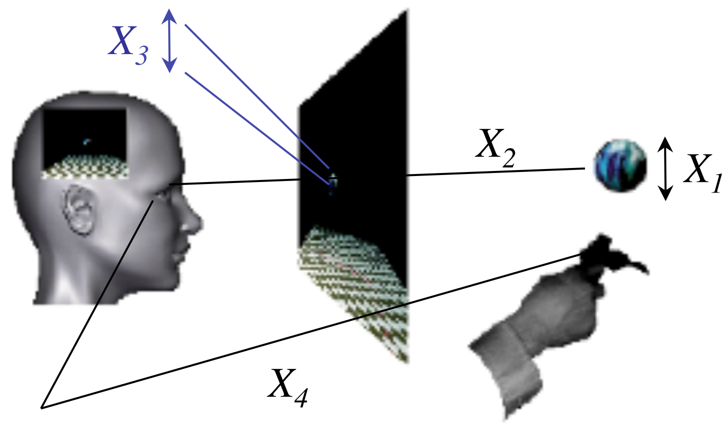


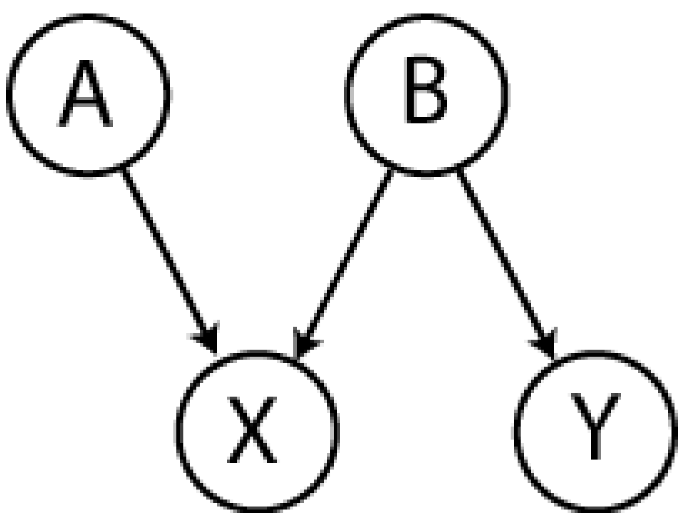
Bloj, M. G., Kersten, D., & Hurlbert, A. C. (1999). Perception of three-dimensional shape influences colour perception via mutual illumination. Nature, 402, 877-879.

Quantitative Predictions for Explaining away?

EXAMPLE

A object size
B object distance
X image size
Y "felt" distance



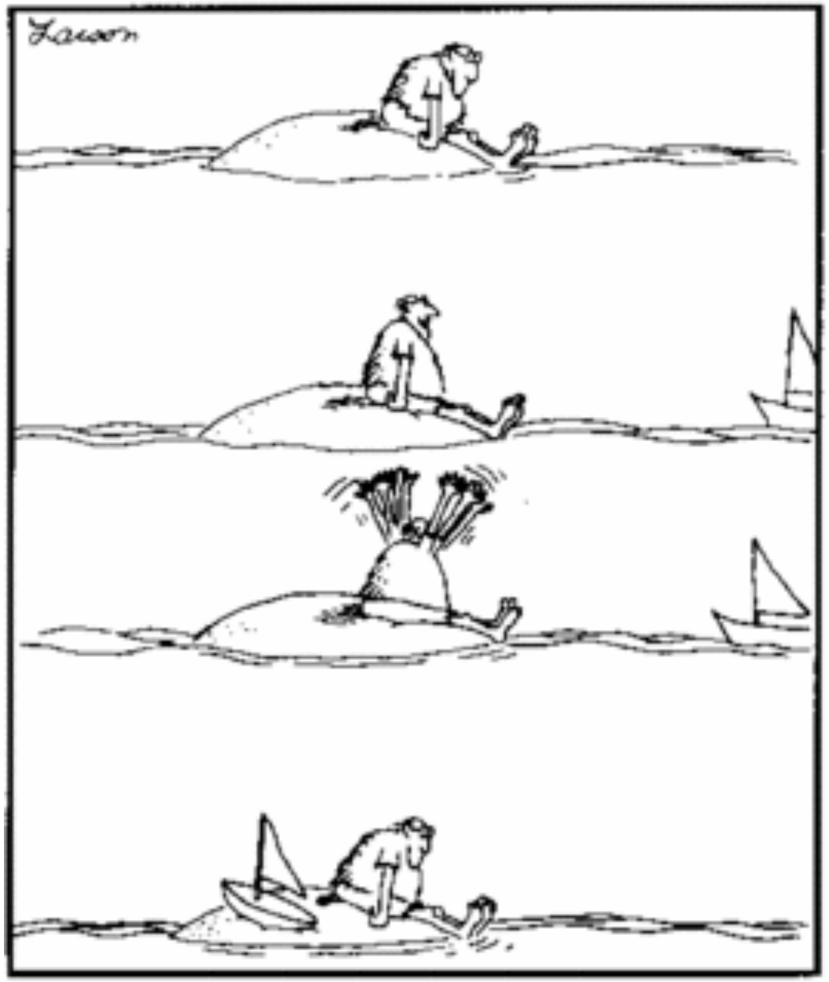


- *Sensory generative knowledge:*
  - constrains possible **size** & **distance** combinations to those consistent with the **image size cue (**Epstein et al., 1961)

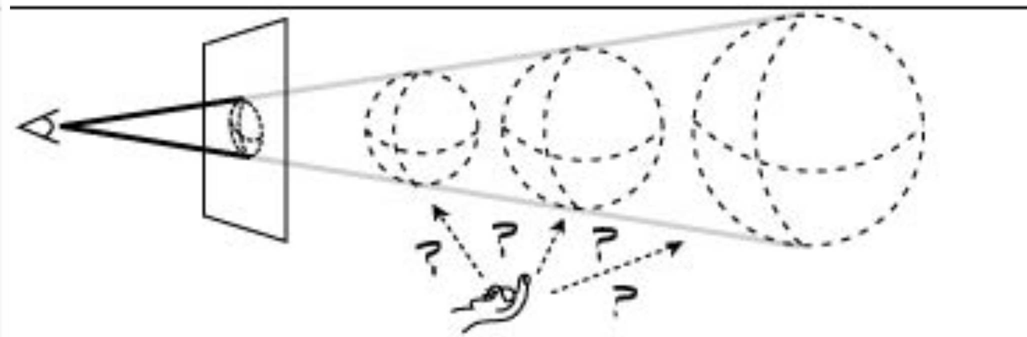
- *Auxiliary* size cue:
  - rules out **size** & **distance** combinations that are inconsistent with auxiliary cue
  - allows unambiguous *inference* of **distance**

- Consistent with feature of Bayesian reasoning: Explaining Away (Pearl, 1988)

The “size / distance” problem



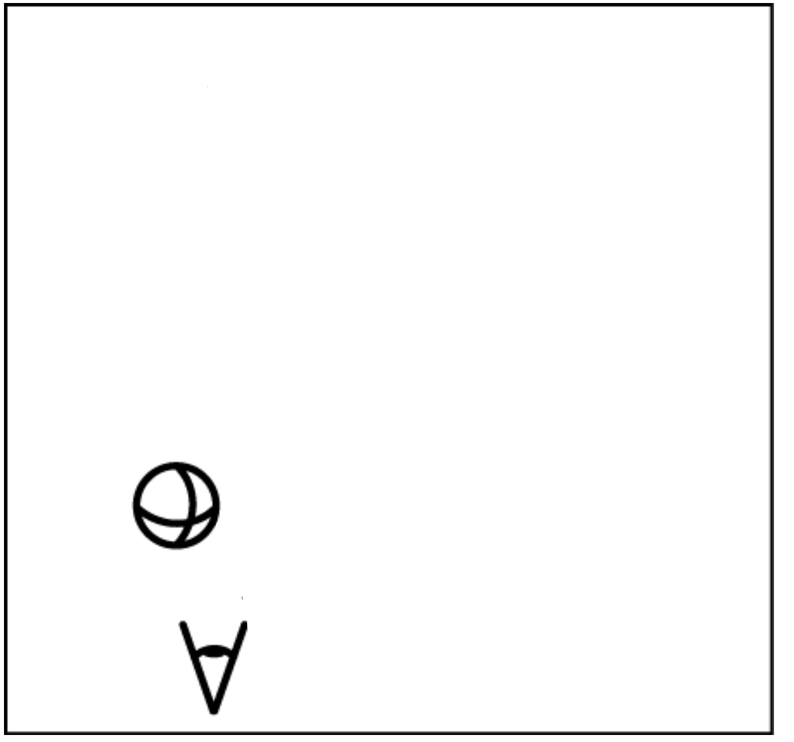
- **Size** and **distance** are *ambiguous* given only a monocular image size cue
  - Emmert’s Law (Boring, 1940; Weintraub & Gardner, 1970)

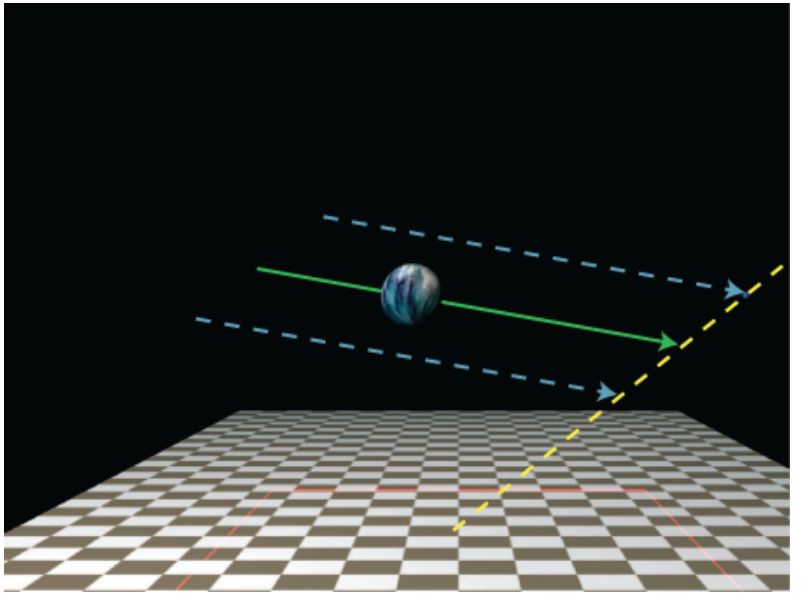


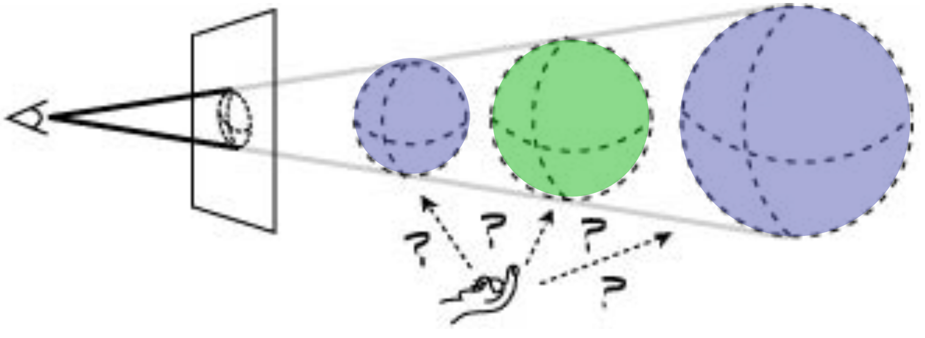
Humans use size cues to improve distance perception

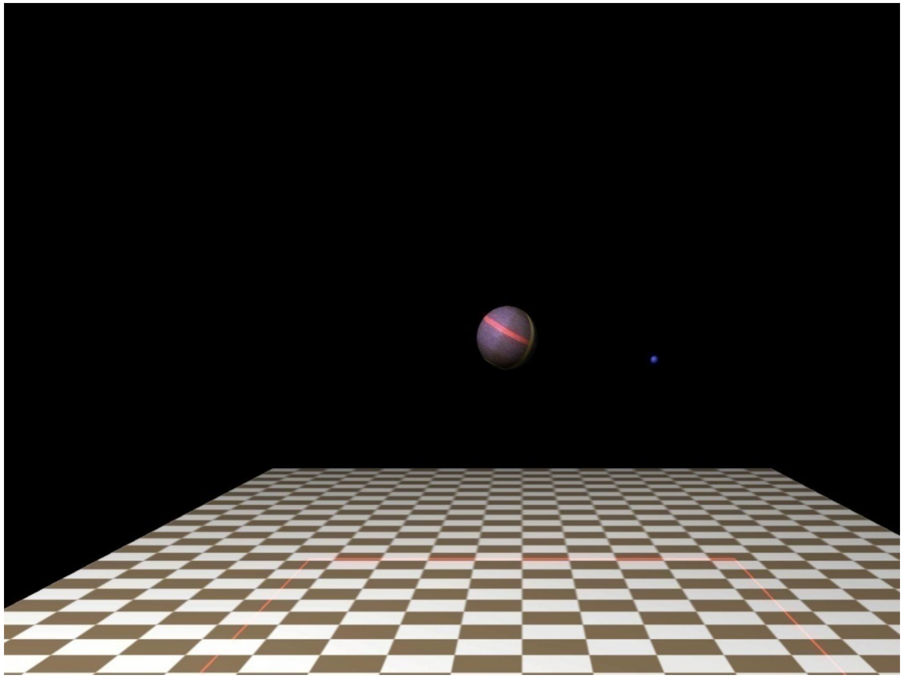
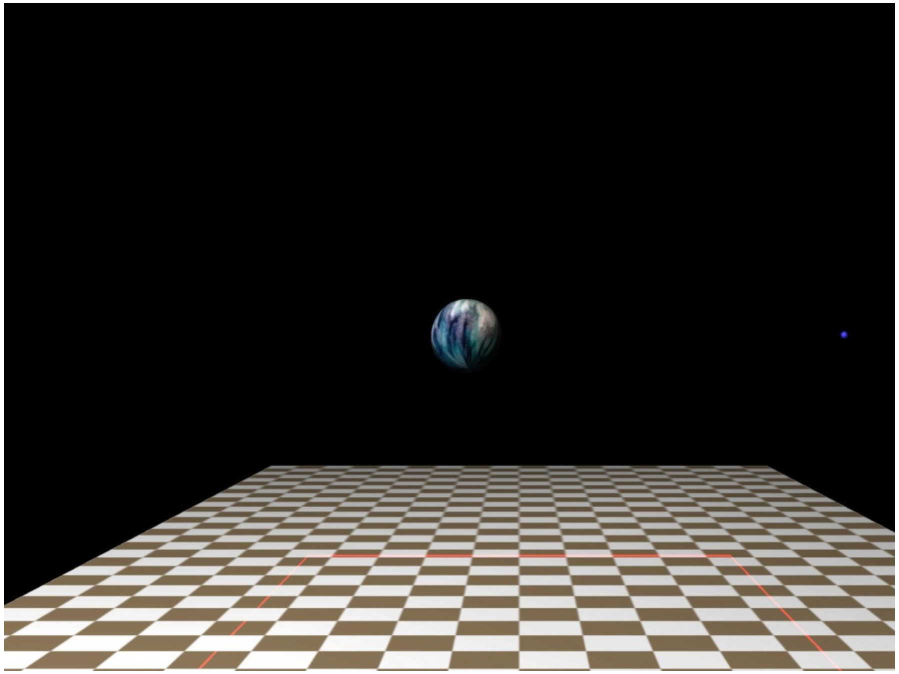
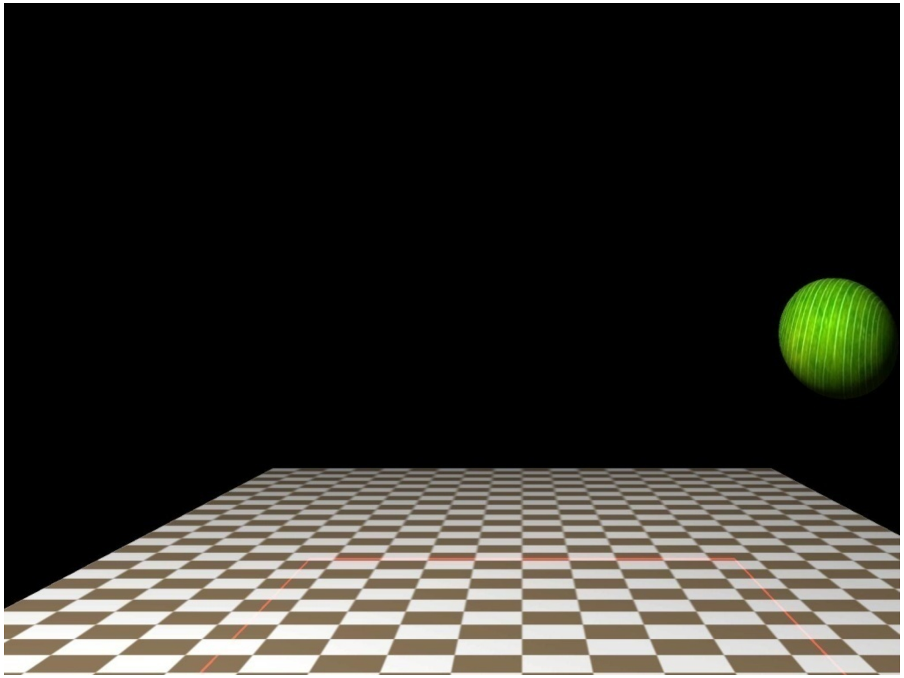
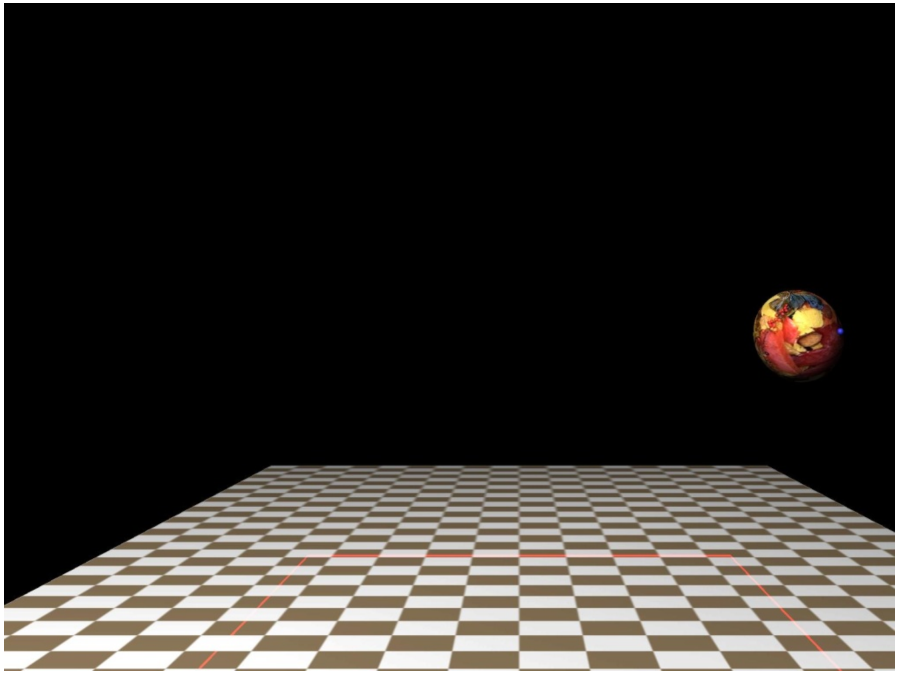
- *Interception phase*:
  - Depress mouse
  - Ball moves to left of scene
  - Begins to approach and move rightward
  - Participant positions fingertip along "constraint line" to intercept

- Computer records:
  - True distance as ***crossing distance***
  - Fingertip position as ***judged distance***



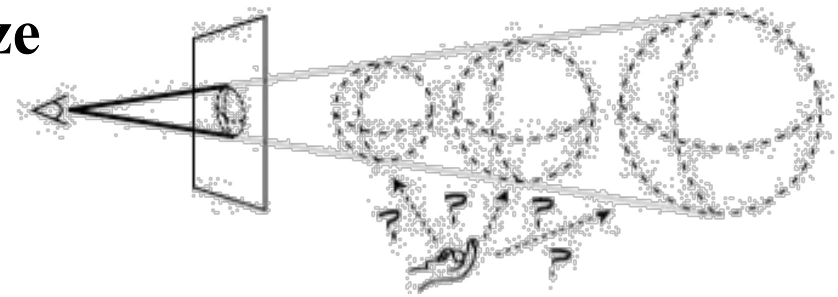




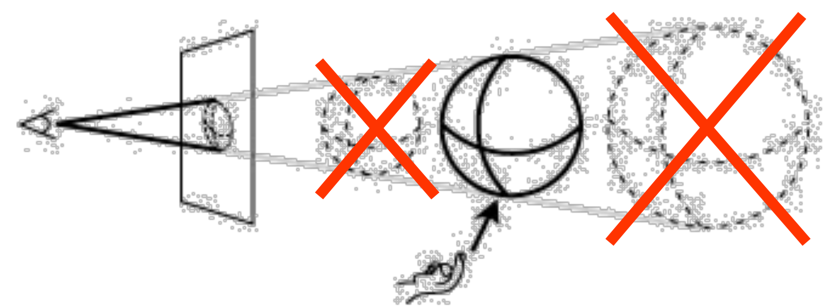





Predictions:

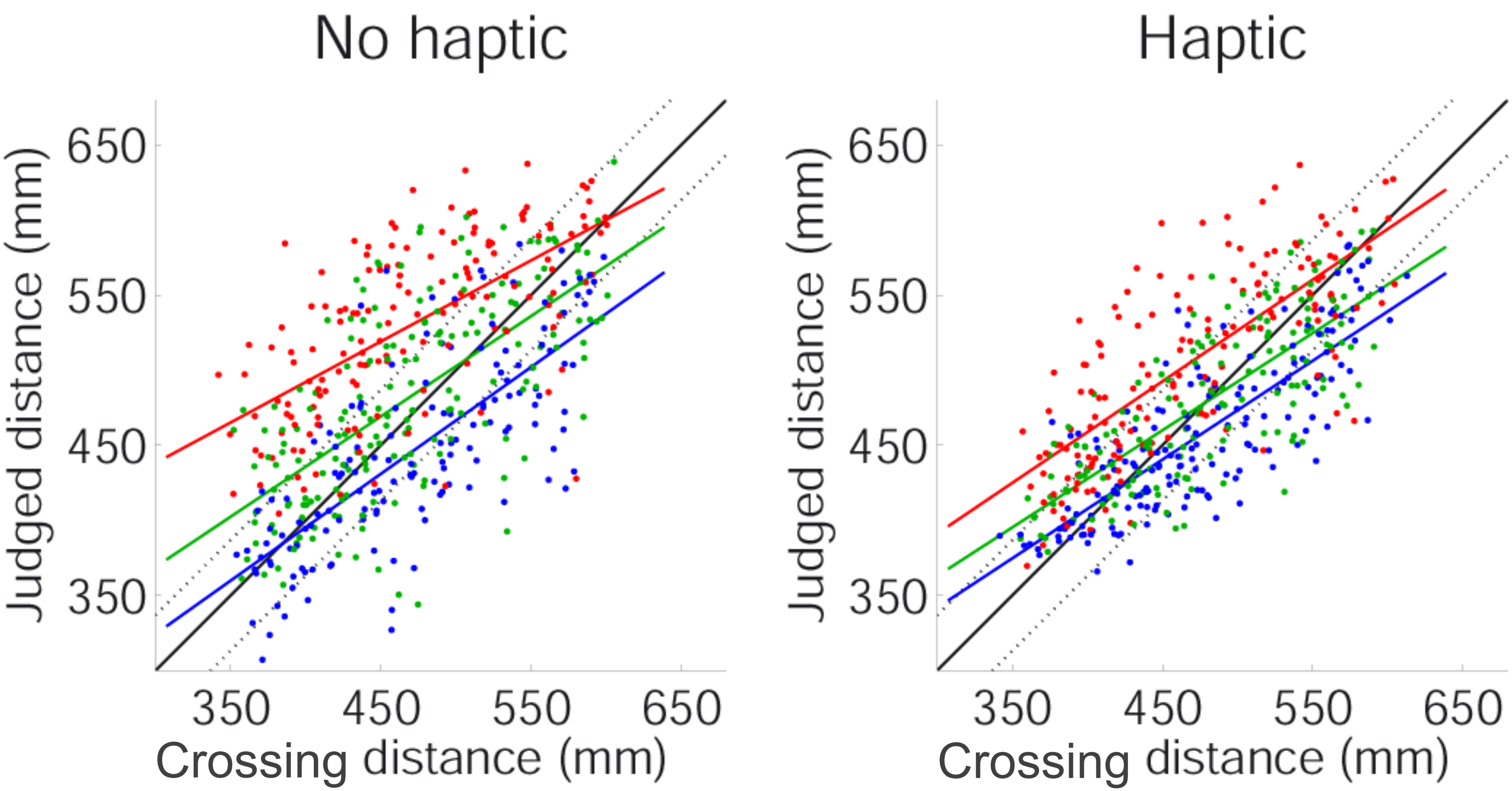
1) **NO-HAPTIC** case:
   - *Judged distances* depend on **ball size**
   - Substantial errors in *judged distances* due to ambiguity



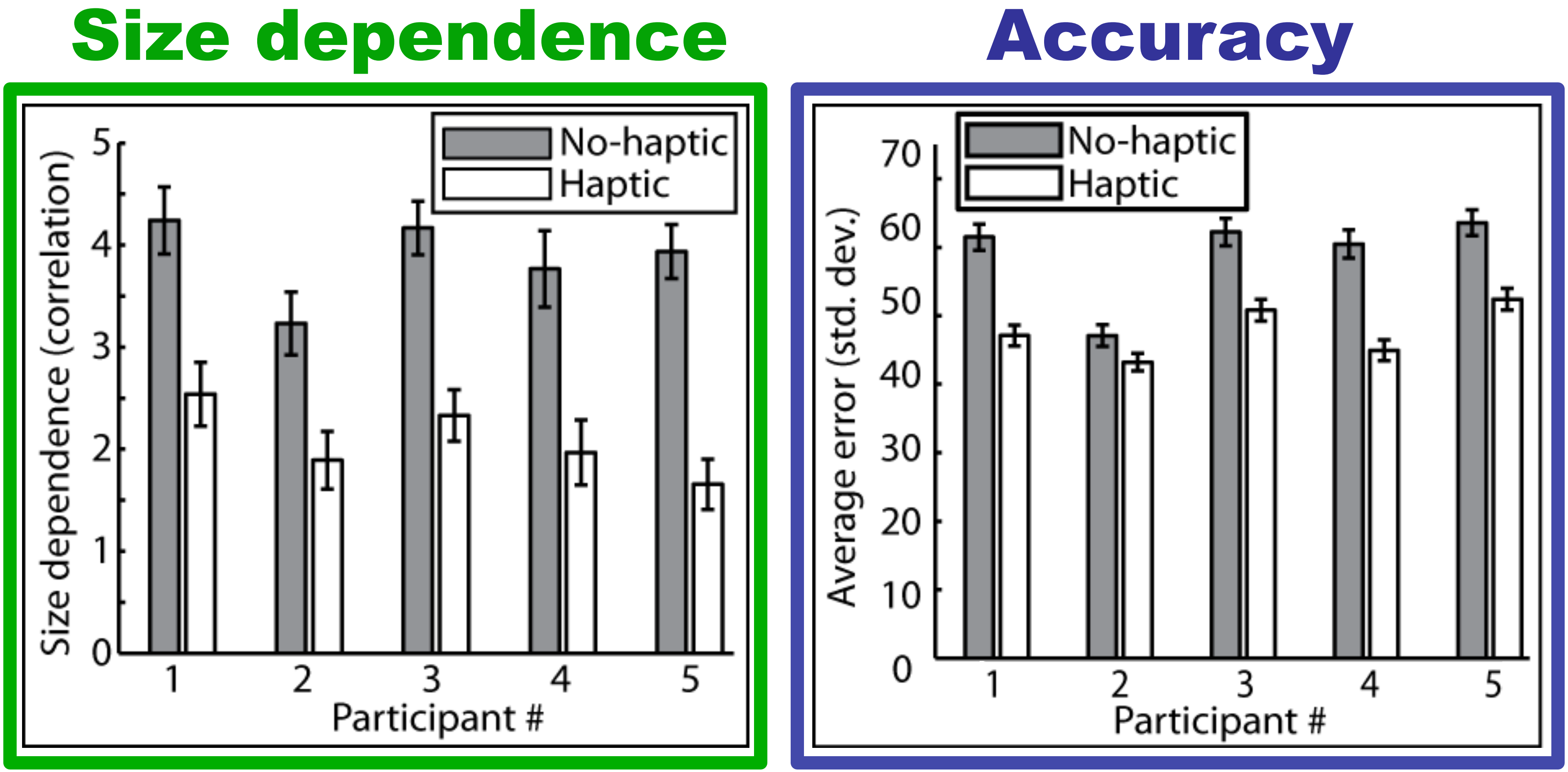
2) **HAPTIC** case:
   - *Judged distances* depend **LESS** on **ball size**
   - Reduced errors due to *explaining away* of inconsistent distances



Judged distances vs. crossing distances (participant 4)



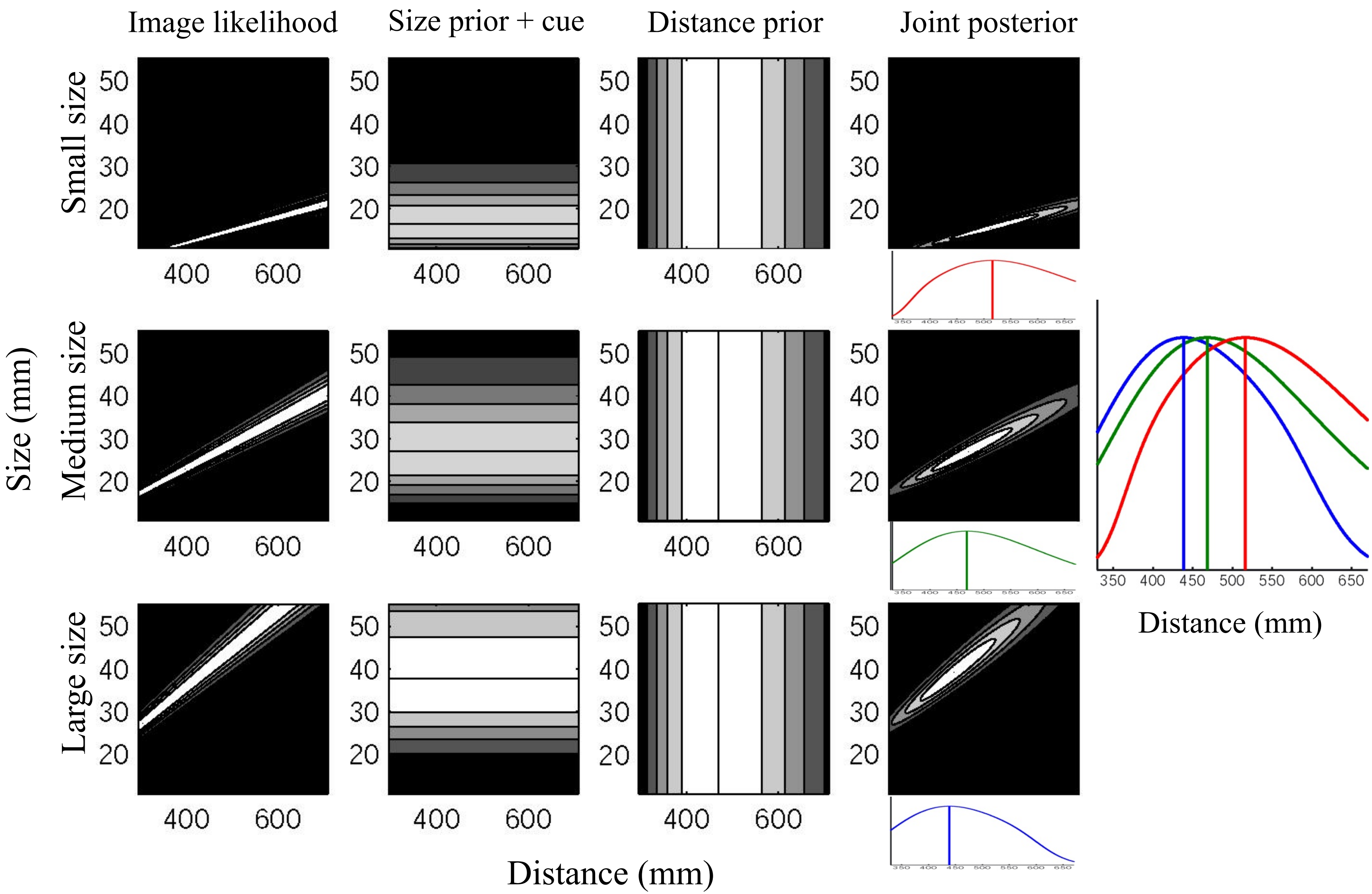
Results:


Size dependence
Accuracy
No-haptic
Haptic
Size dependence (correlation)
Participant #
Average error (std. dev.)
0
1
2
3
4
5
10
20
30
40
50
60
70

ALL SAME DISTANCE

NO-HAPTIC

HAPTIC



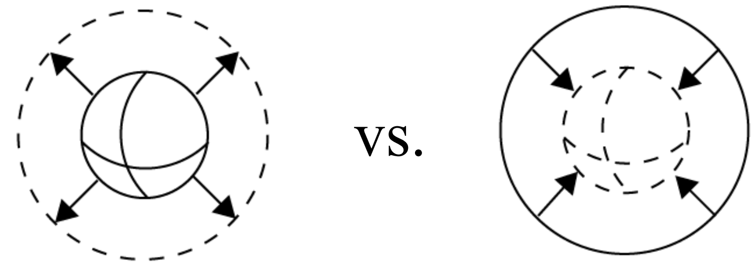
- Bayesian model does a good job of predicting data

- Modeling the participants as "sampling from their posteriors" does better job of predicting data than modeling them as "MAP estimators"

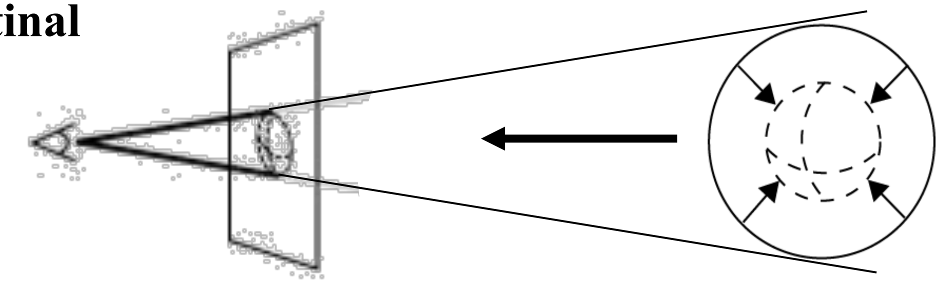
- <u>Reasonable noise estimates:</u>
    - Vis. angle noise std. dev. ~ [6, 30] minutes @ [81, 410]
    - Haptic size noise std. dev. ~ [2, 5] mm @ [14, 42]

Size-change perception

- Extension of *size/distance* problem:
  - **size-change** perception



- Example:
  - Imagine viewing a balloon whose **retinal image size** is *shrinking*
  - The balloon may be *deflating*, OR *inflating* and receding rapidly
  - Knowing the **distance-change** rate can disambiguate the **size-change** rate



- Experimental question:
  - Can auxiliary **distance-change cues** improve **size-change** judgments?
  - Are both HAPTIC and STEREO **distance-change** cues effective?

Psychophysical Methods 1

- 11 human participants in virtual reality workbench (PHANToM & 3D graphics)
  - (1 outlier was removed)
- <u>Stimulus</u>: monocularly-viewed ball that changed in size and distance
- Distance-change cues:
  - **HAPTIC**: 1 fingertip "stuck" in center of ball as it moves
  - **STEREO**: binocular images consistent with real physical projection
- After 1000ms, participant chooses:
  - **INFLATING** or **DEFLATING**

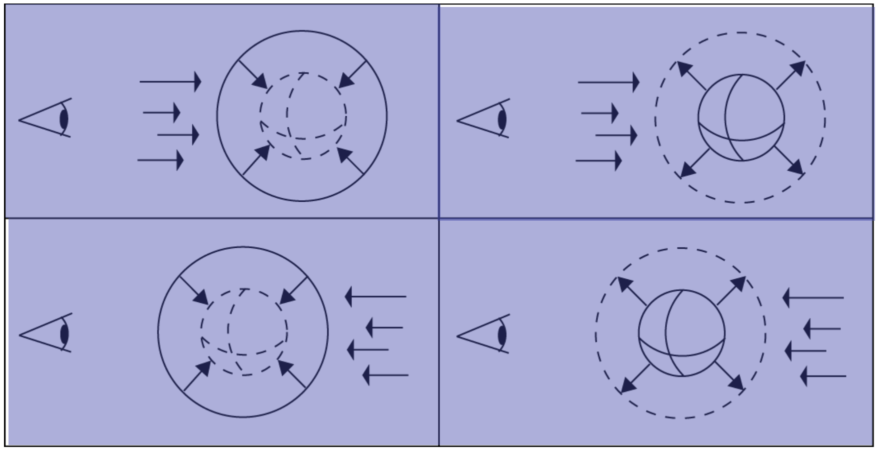
Methods 2:

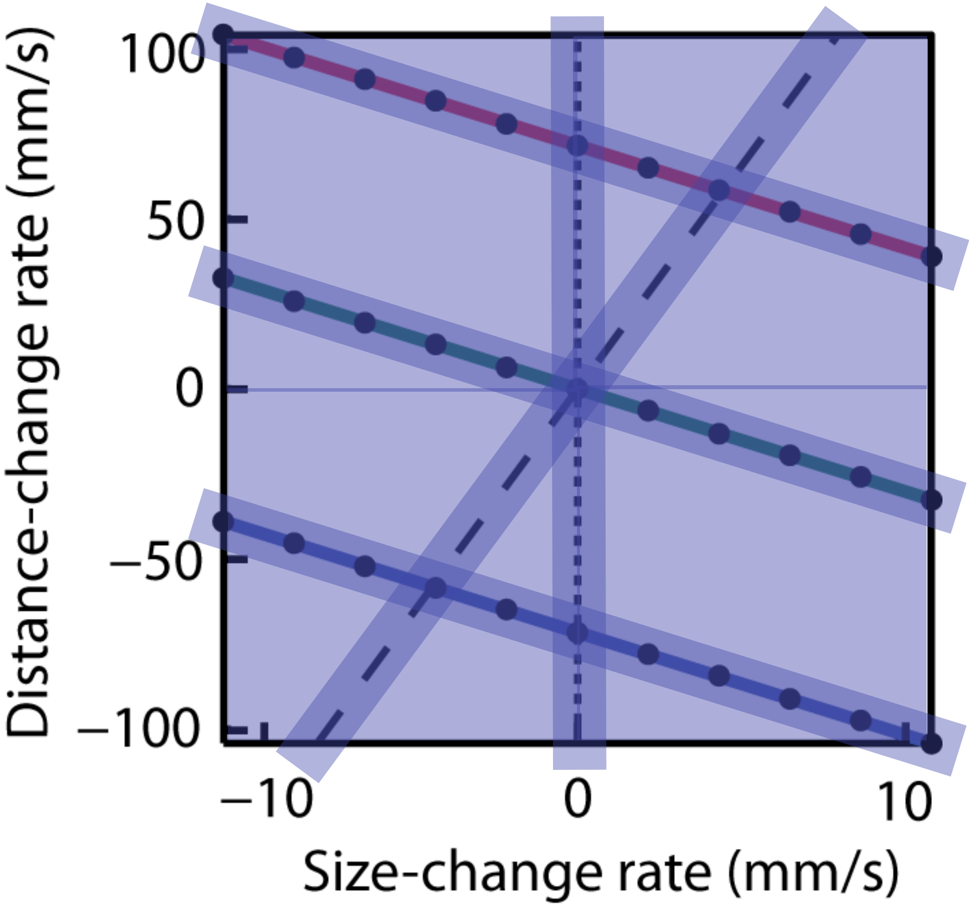
- 330 trials per 4 distance-cue cases:
  1) No Auxiliary cues
  2) Haptic-only
  3) Stereo-only
  4) Haptic & Stereo

- Each case: 3 psychometric functions - 11 points x 10 repetitions per point (black dots) - were measured.

- Diagonal, dashed line: size- & distance-change combinations that yield **ZERO** image size-change.

- Vertical, dotted line: boundary of unbiased discrimination between inflating and deflating sizes.



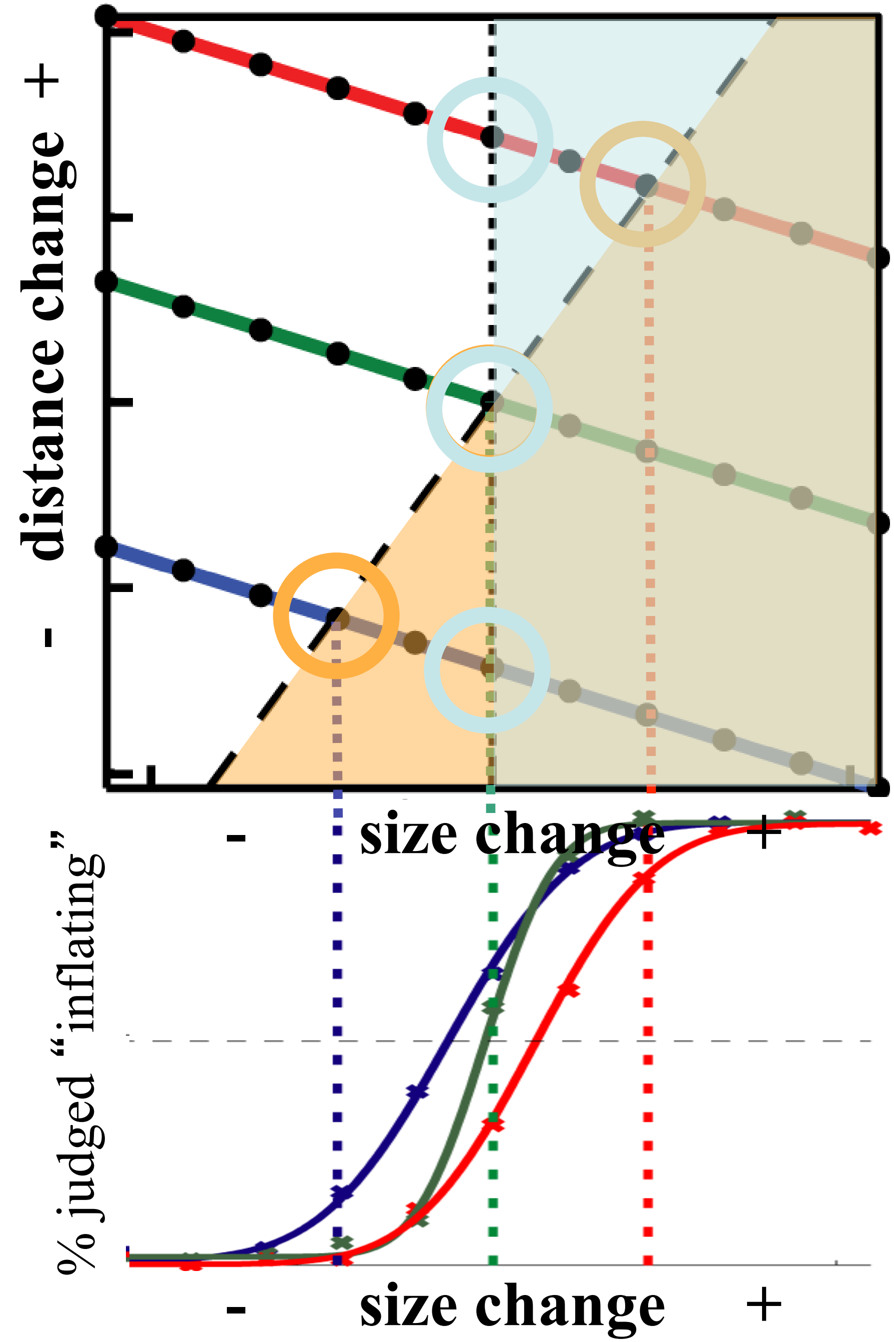


Predictions:

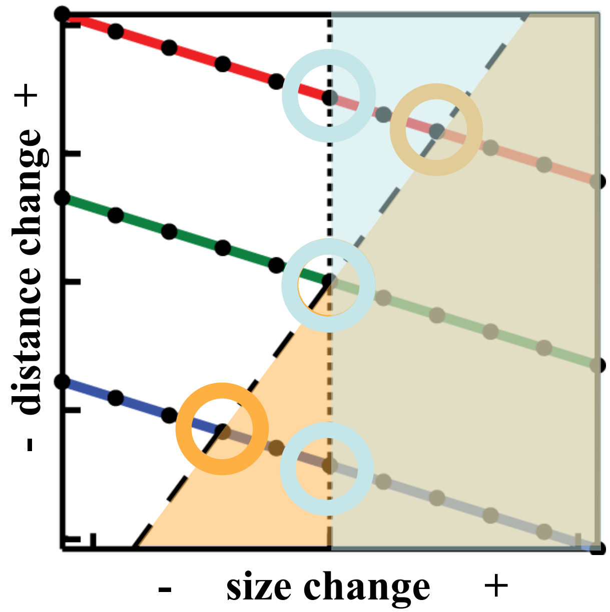
2 predictions for "explaining away" observer:

1. <u>No Auxiliary case</u>: psychometric curves along the diagonal, dotted line

2. <u>Auxiliary cases</u>: psychometric curves along the vertical, dotted line

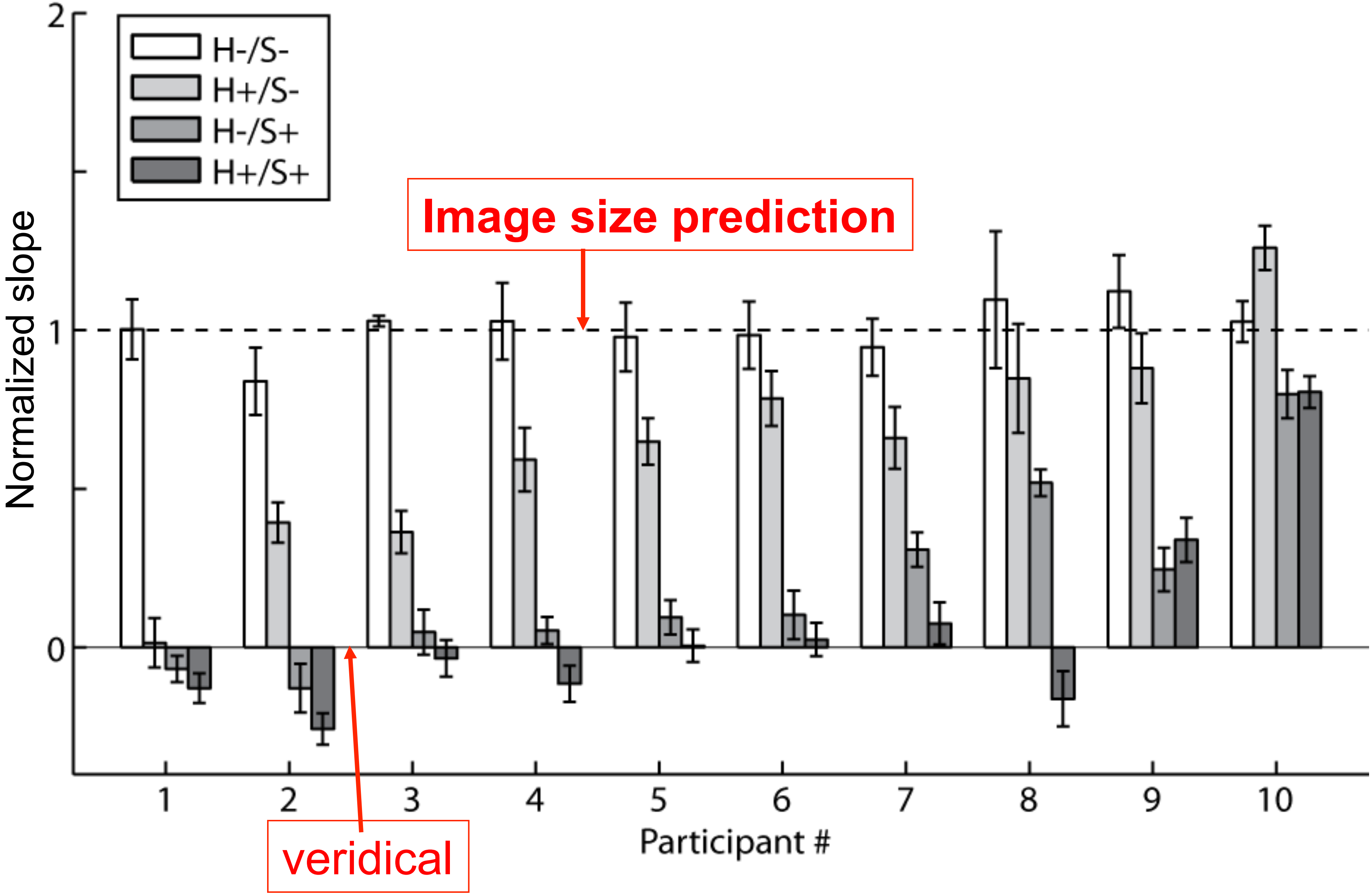


Results



1. No Auxiliary case: the size-change judgments are based on image size-change.
2. Haptic-only, Stereo-only, Haptic & Stereo: increased veridicality, physical size-change is more accurately judged.

Summary of participants' normalized slopes

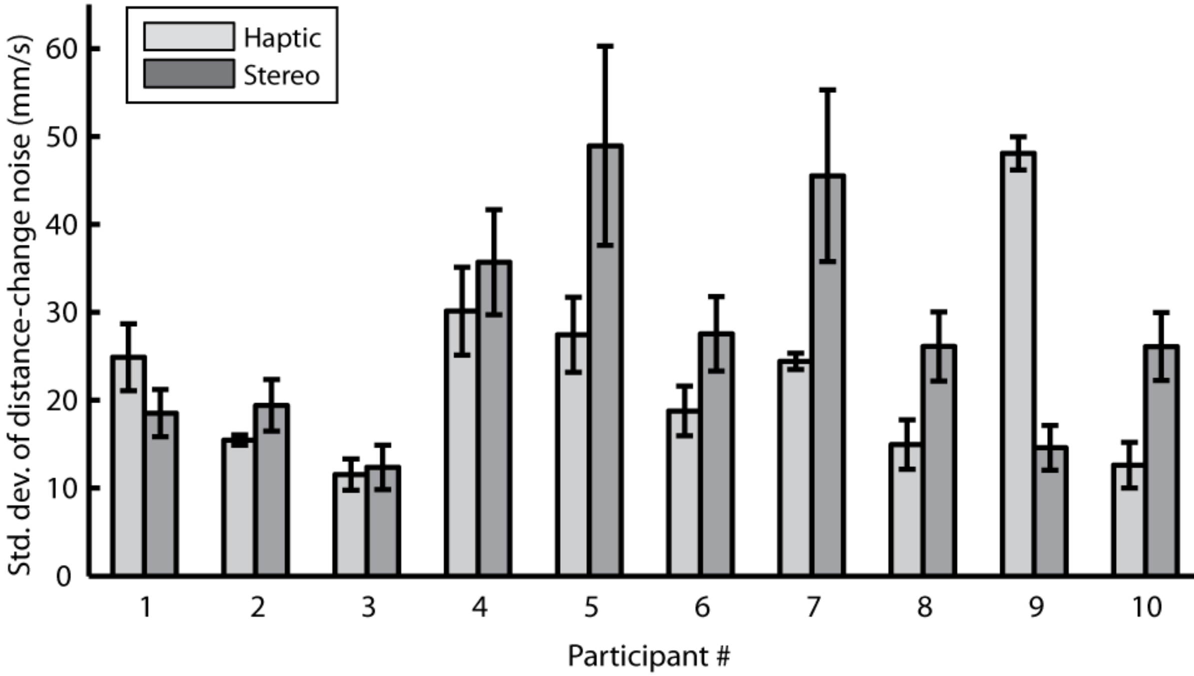


Why is **stereo** > **haptic**?

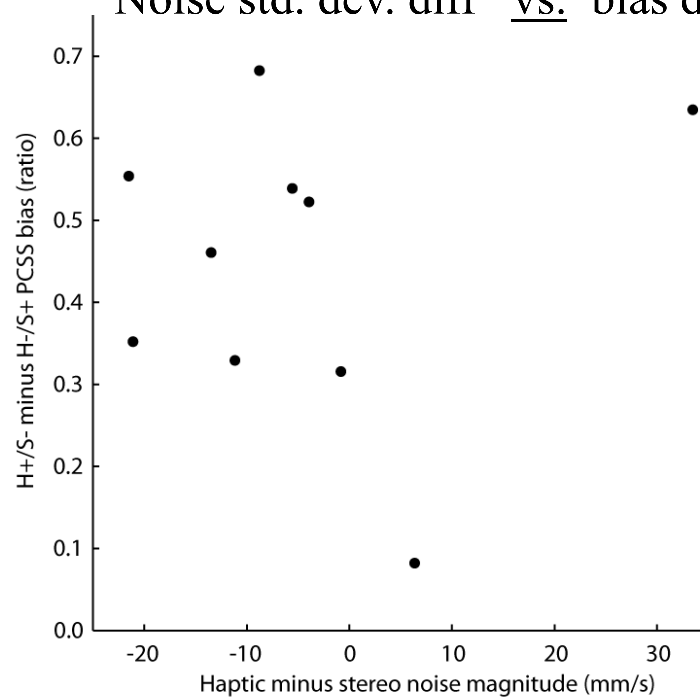
- <u>Follow-up experiment</u>: measured stereo & haptic distance-change cue reliabilities (Ernst, 2005)

- <u>2IFC</u>: "Which interval contained faster ball?"

- Psychometric function (cumulative normal) slope gives us each cue's noise std. dev.



Noise std. dev. diff vs. bias diff



NO CORRELATION → not simply a difference in auxiliary cue quality

Experiment 2: Conclusions

- Participants use **distance-change cues** to improve their **size-change** perception.
- *Stereo* **distance-change cue** is more useful than *haptic*
  - There is a discrepancy between how haptic and stereo distance information are used to improve size-change judgments.
- *Haptic* and *stereo* distance-change cues have similar reliability
  - (perhaps even haptic > stereo)

Possible reasons for stereo/haptic discrepancy:

- Brain is suboptimal - does not exploit haptic cue's full potential
- Brain understands haptic distance cue is less likely to be causally-related to image size cue, thus only integrates it partially (Koerding et al., 2007)

- Next steps:
  - Quantitative Bayesian model
  - Causal model

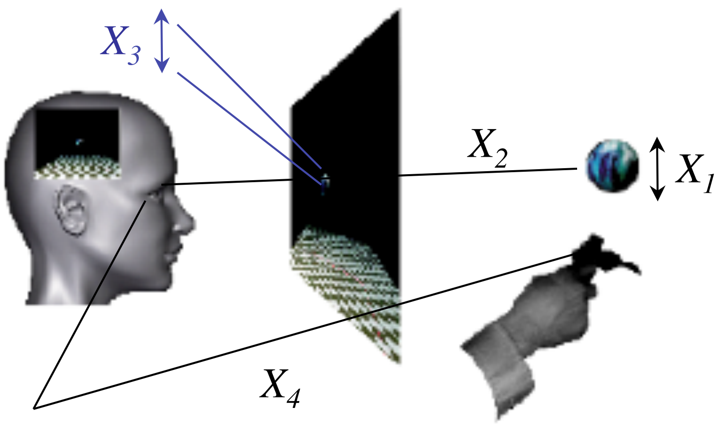
General Conclusions

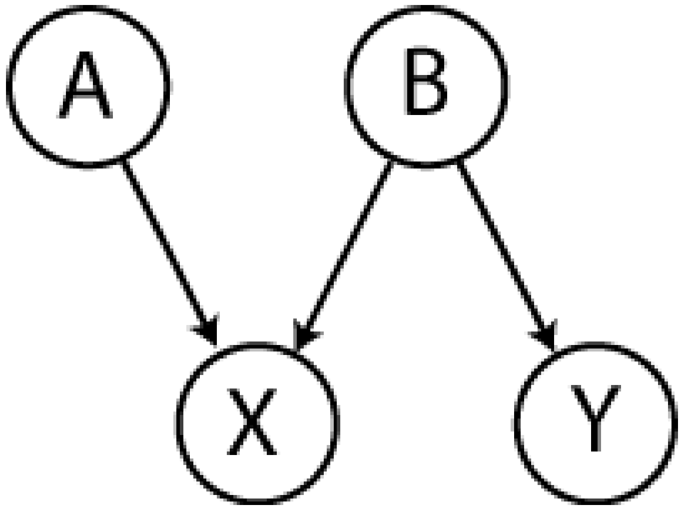
- Uncertainty and ambiguity plague perceptually-guided actions.
- The brain has knowledge of each, and forms percepts and plans actions to overcome their negative consequences.
- Generative knowledge has (potentially) a hierarchical structure
- Non-parametric Bayesian models provide a language to handle the difference between fixed relationships and those that vary from scene to scene, sharing relevant information across scenes.
- Such processing is characteristic of Bayesian reasoning and decision-making.

Quantitative Predictions for Explaining away?

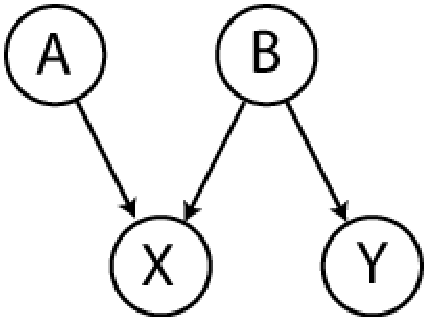
EXAMPLE

- A object size
- B object distance
- X image size
- Y "felt" distance





Making more Complex ***Qualitative*** Predictions

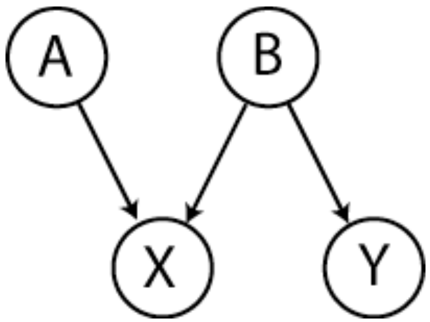


GOAL: Not meant to be a substitute for modeling, but how do you get cute "cue weight formulas" for complex models

- Given a network structure
- Linearize around values of hidden variables to 2nd order (moment matching, taylor, Laplace)

$$\begin{bmatrix} X \\ Y \end{bmatrix} = T \cdot \begin{bmatrix} A \\ B \end{bmatrix} + \begin{bmatrix} \omega_X \\ \omega_Y \end{bmatrix}$$

Making more Complex ***Qualitative*** Predictions



GOAL: Not meant to be a substitute for modeling, but how do you might get cute "cue weight formulas" for complex models

- Linearization

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} \omega_X \\ \omega_Y \end{bmatrix}$$

$$\mathbf{z} = T\mathbf{x} + \mathbf{w}$$

Assume Gaussian Noise

PRIOR

$$P(a)P(b) = P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$$

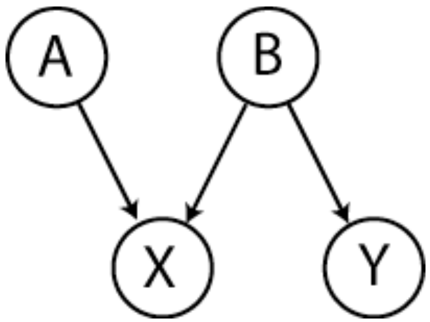
$$\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad C_{prior} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

LIKELIHOOD

$$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$$

$$C_{XY} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}; \qquad T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Making more Complex *Qualitative* Predictions



WANTED

$$P(b \mid \mathbf{z})$$

PRIOR

$$P(a)P(b) = P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$$

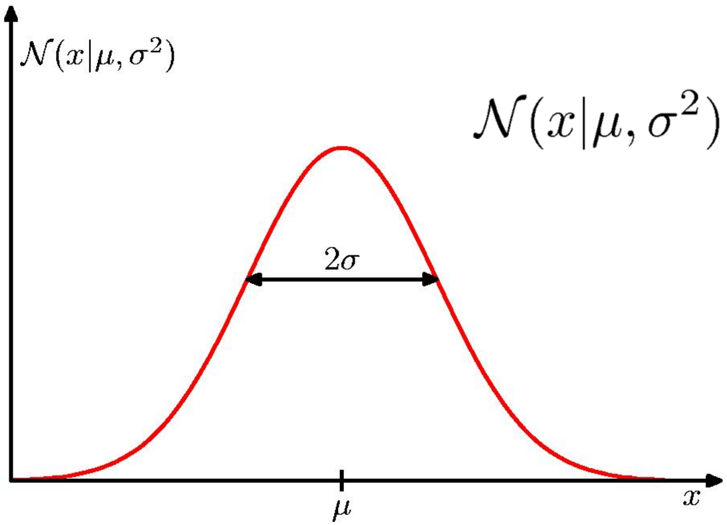
$$\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad C_{prior} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

LIKELIHOOD

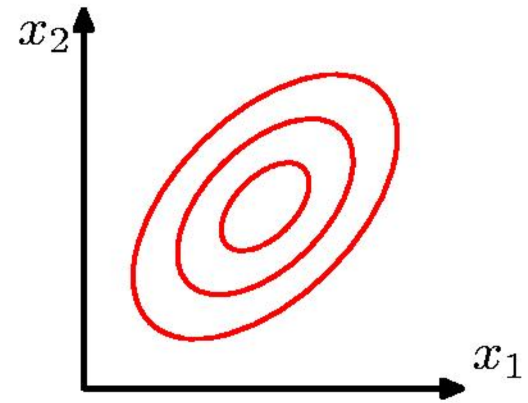
$$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$$

$$C_{XY} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}; \qquad T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

The Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

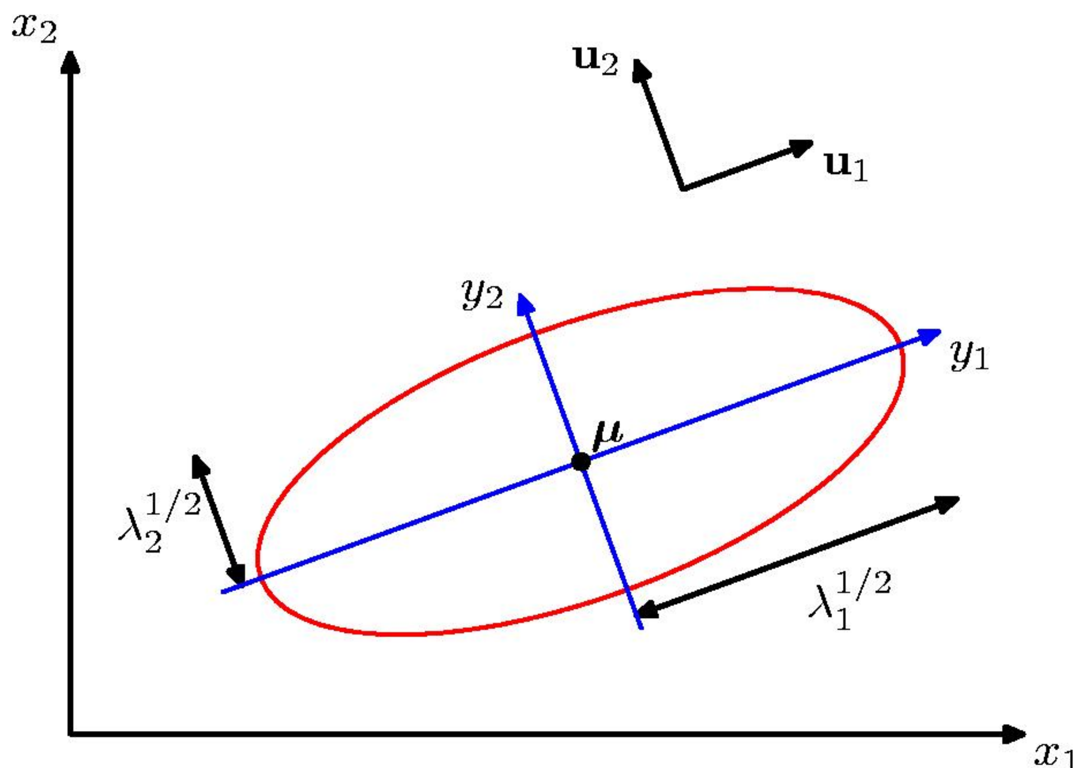
Geometry of the Multivariate Gaussian

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}} (\mathbf{x} - \boldsymbol{\mu})$$



Moments of the Multivariate Gaussian (1)

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\end{aligned}$$

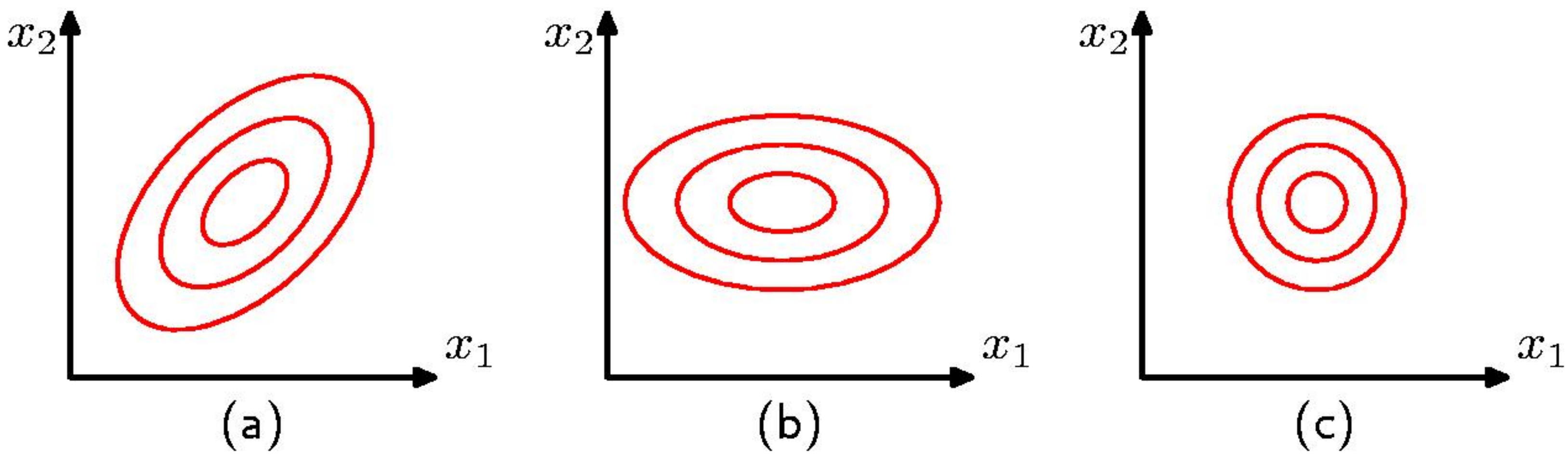
thanks to anti-symmetry of z

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Multivariate Gaussian (2)

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$



Partitioned Conditionals and Marginals

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

Conditionals

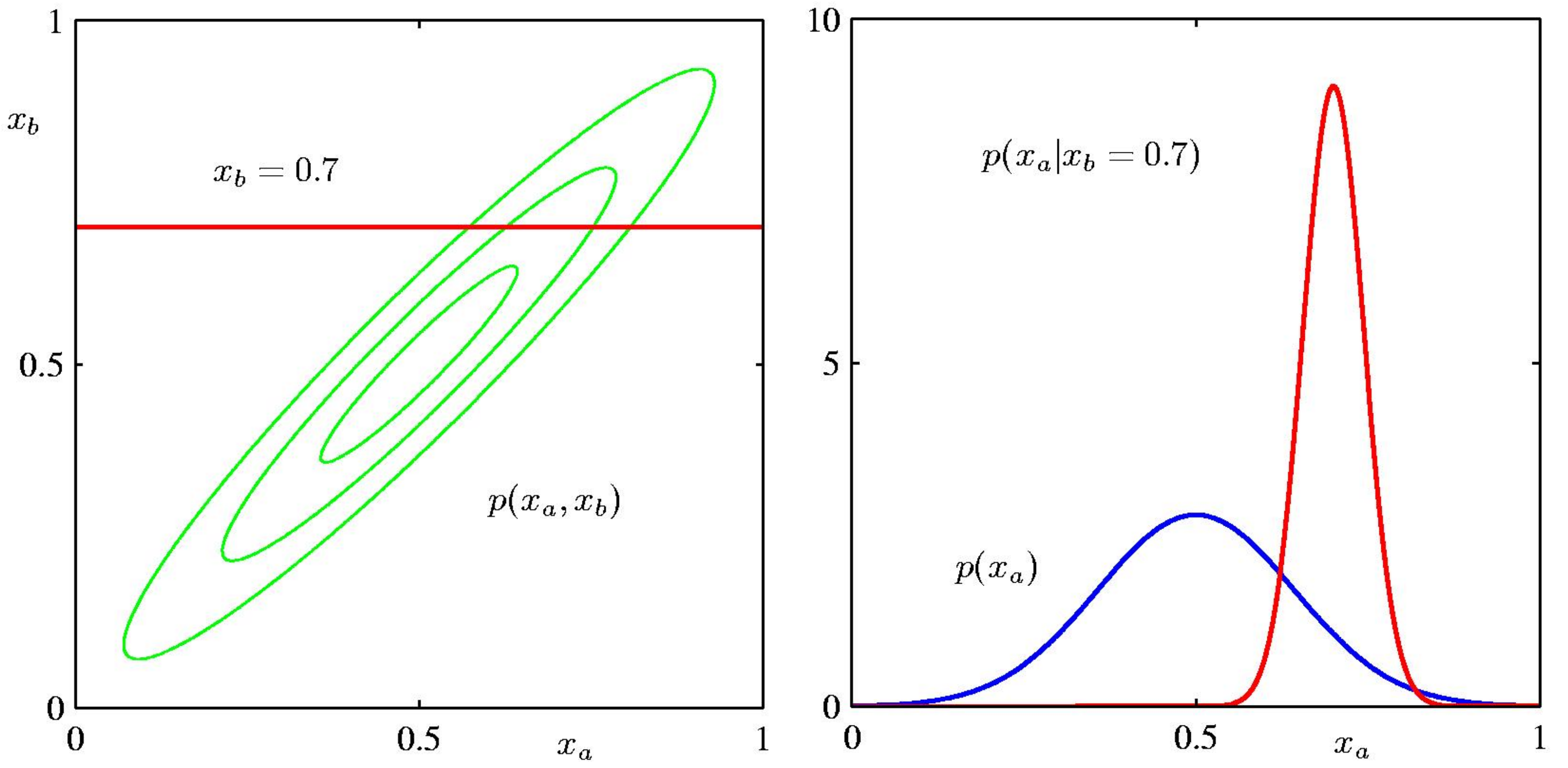
$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\right\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

Marginals

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\, \mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

Partitioned Conditionals and Marginals



Bayes' Theorem for Gaussian Variables

- Given

- we have

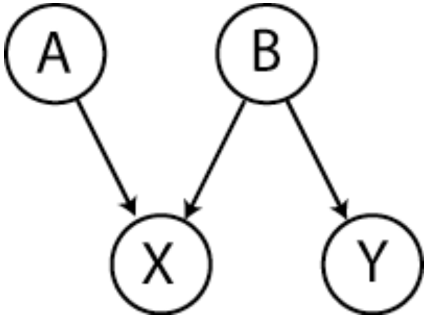
$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

- where

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})^{-1}
\end{aligned}
$$

$$
\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}
$$

Making more Complex *Qualitative* Predictions



WANTED: $P(b \mid \mathbf{z})$

1) Bayes: *Given*

$P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$

$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$

$$P(\mathbf{x} \mid \mathbf{z}) = N(\mathbf{x} \mid \mu_{post}, C_{post})$$

$$\mu_{post,} = C_{post}^{-1}\left(T^T C_{XY}^{-1}\mathbf{z} + C_{prior}^{-1}\mu_{prior}\right)$$

$$C_{post} = \left(C_{prior}^{-1} + T^T C_{XY}^{-1} T\right)^{-1}$$

PRIOR

$$P(a)P(b) = P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$$

$$\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad C_{prior} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

LIKELIHOOD

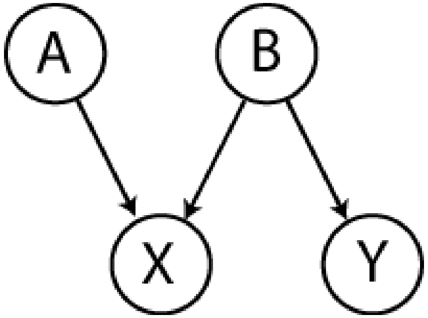
$$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$$

$$C_{XY} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}; \qquad T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

2) Marginalize *a*:

$$P(b \mid \mathbf{z}) = N(b \mid \mu_{post,}^{b}, C_{post}^{bb})$$

Making more Complex Quantitative Predictions



EXAMPLE FOR: ***P(a|z)***

$$\bar{\mu}_{Post} = \bar{\mu}_{prior} + C_{prior}^{T} \cdot T^{T} \cdot \left(T \cdot C_{prior} \cdot T^{T} + C_{XY}\right)^{-1} \cdot \left(\mathbf{z} - T \cdot \bar{\mu}_{prior}\right)$$

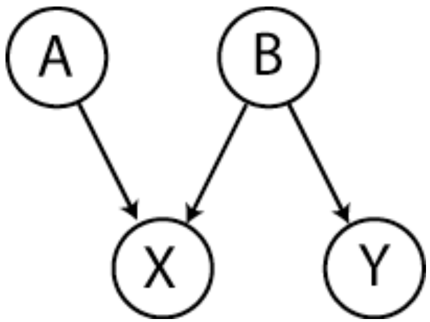
Different properties than cue combination!

$$T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \qquad C_{prior} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} \qquad C_{XY} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix};$$

$$A^{*} = \frac{\alpha}{\sigma_X^2\left(\beta + \sigma_Y^2\right) + \alpha\left(\beta + \sigma_Y^2 + \sigma_X^2\right)} \left\{ \left(\beta + \sigma_Y^2\right) X + \sigma_X^2\, Y + \left(\beta + \sigma_X^2 + \sigma_Y^2\right) \bar{\mu}_{prior}^{A} + \left(\beta + \sigma_Y^2\right) \bar{\mu}_{prior}^{B} \right\}$$

Cue weights don’t sum to one, both priors matter, etc.

Making more Complex ***Qualitative*** Predictions



WANTED: $P(b \mid \mathbf{z})$

1) Bayes: *Given*

$$P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$$
$$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$$

$$P(\mathbf{x} \mid \mathbf{z}) = N(\mathbf{x} \mid \mu_{post}, C_{post})$$
$$\mu_{post,} = C_{post}^{-1}\left(T^T C_{XY}^{-1}\mathbf{z} + C_{prior}^{-1}\mu_{prior}\right)$$
$$C_{post} = \left(C_{prior}^{-1} + T^T C_{XY}^{-1} T\right)^{-1}$$

2) Marginalize *a*:

$$P(b \mid \mathbf{z}) = N(b \mid \mu_{post,}^{b}, C_{post}^{bb})$$

PRIOR

$$P(a)P(b) = P(\mathbf{x}) = N(\mathbf{x} \mid \mu_{prior}, C_{prior})$$

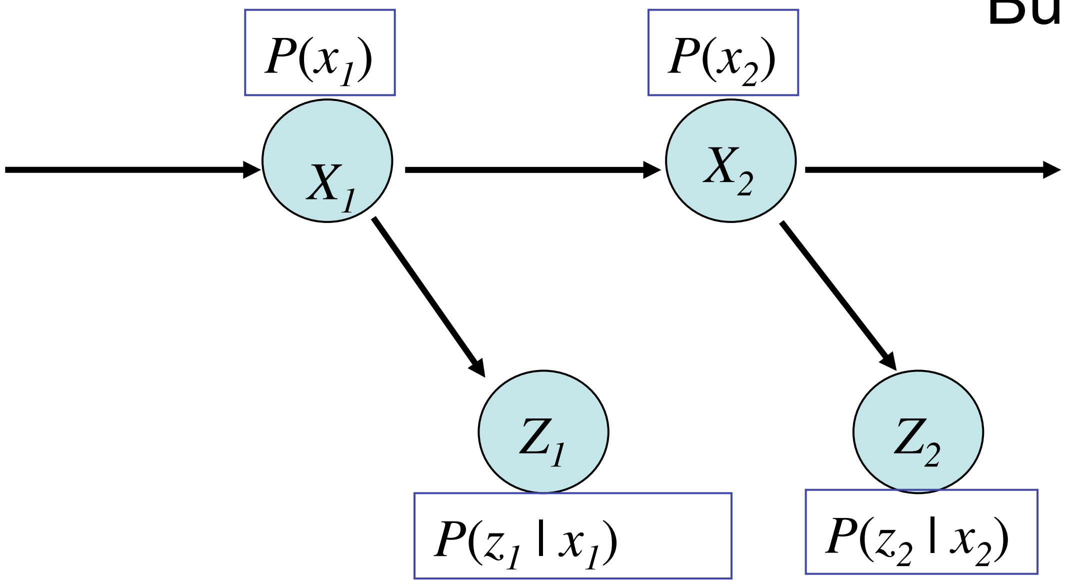
$$\mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix} \qquad C_{prior} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

LIKELIHOOD

$$P(\mathbf{z} \mid \mathbf{x}) = N(\mathbf{z} \mid T\mathbf{x}, C_{XY})$$

$$C_{XY} = \begin{pmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{pmatrix}; \qquad T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Bayesian Networks: Modeling temporal dependence

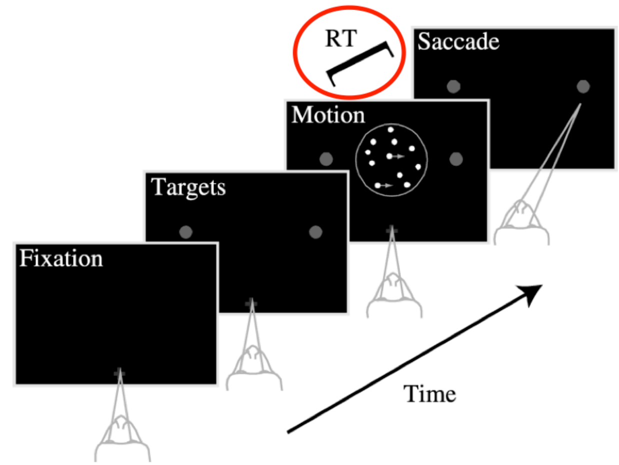


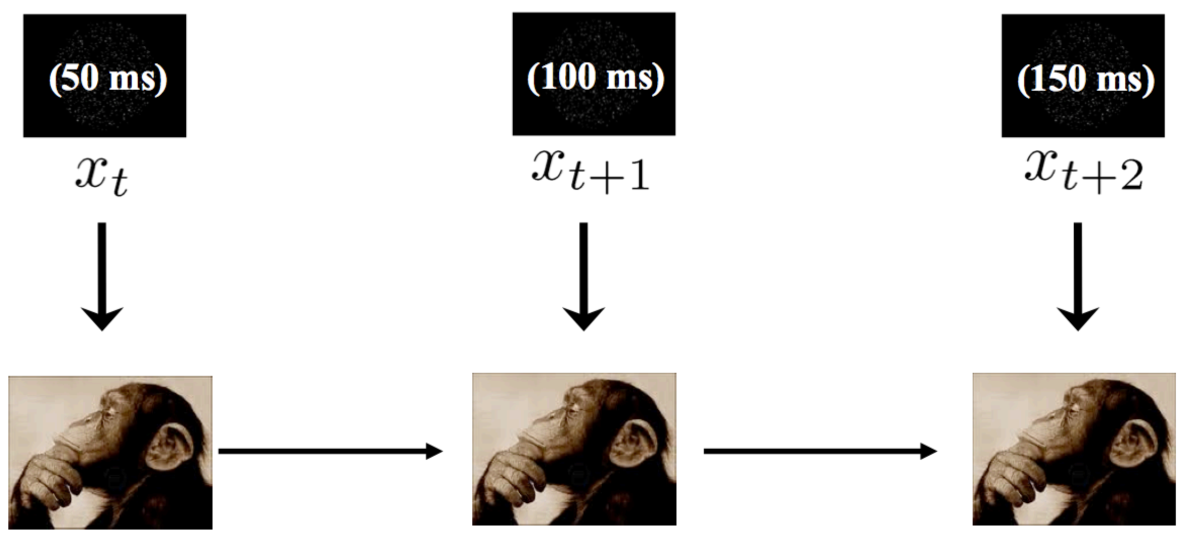
This is just cue combination
But with a more complex prior.

EXAMPLES

Sensori-motor integration
Calibration
Learning
Trajectory Perception

Bayesian Networks: Temporal inference





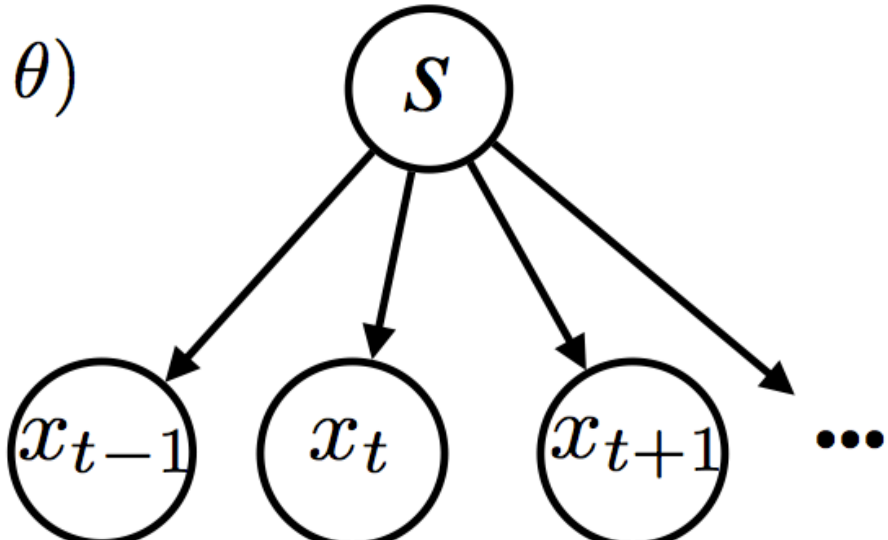
Left or Right?

Bayesian Inference: Review

Generative Model: statistical assumptions about the world

hidden variable: L or R

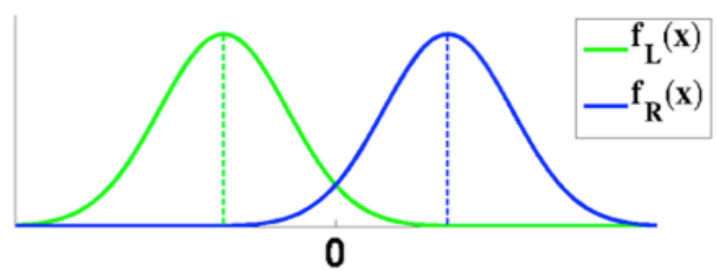
prior $p(s;\theta)$



likelihood

$p(x|s;\phi)$

observed data

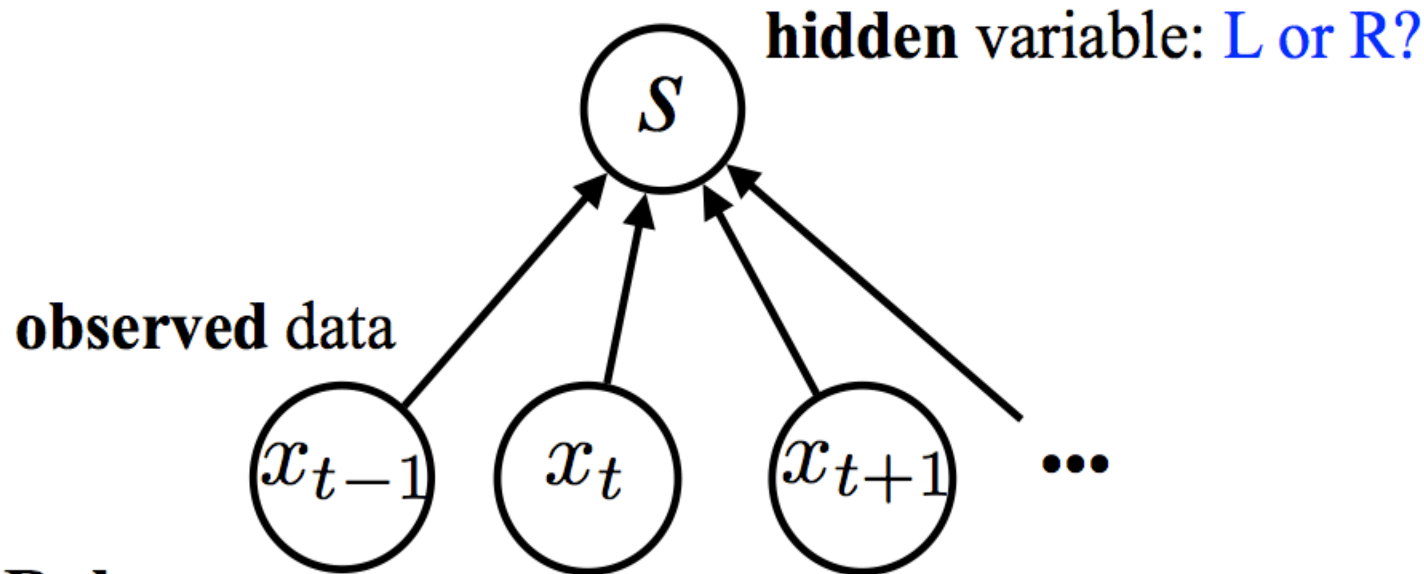


***iid* noise**

$$p(\mathbf{x}_t|s;\phi) = \prod_{i=1}^{t} p(x_i|s;\phi)$$

$$\mathbf{x}_t := (x_1, \ldots, x_t)$$

Bayesian Inference



Bayes' Rule

(batch)

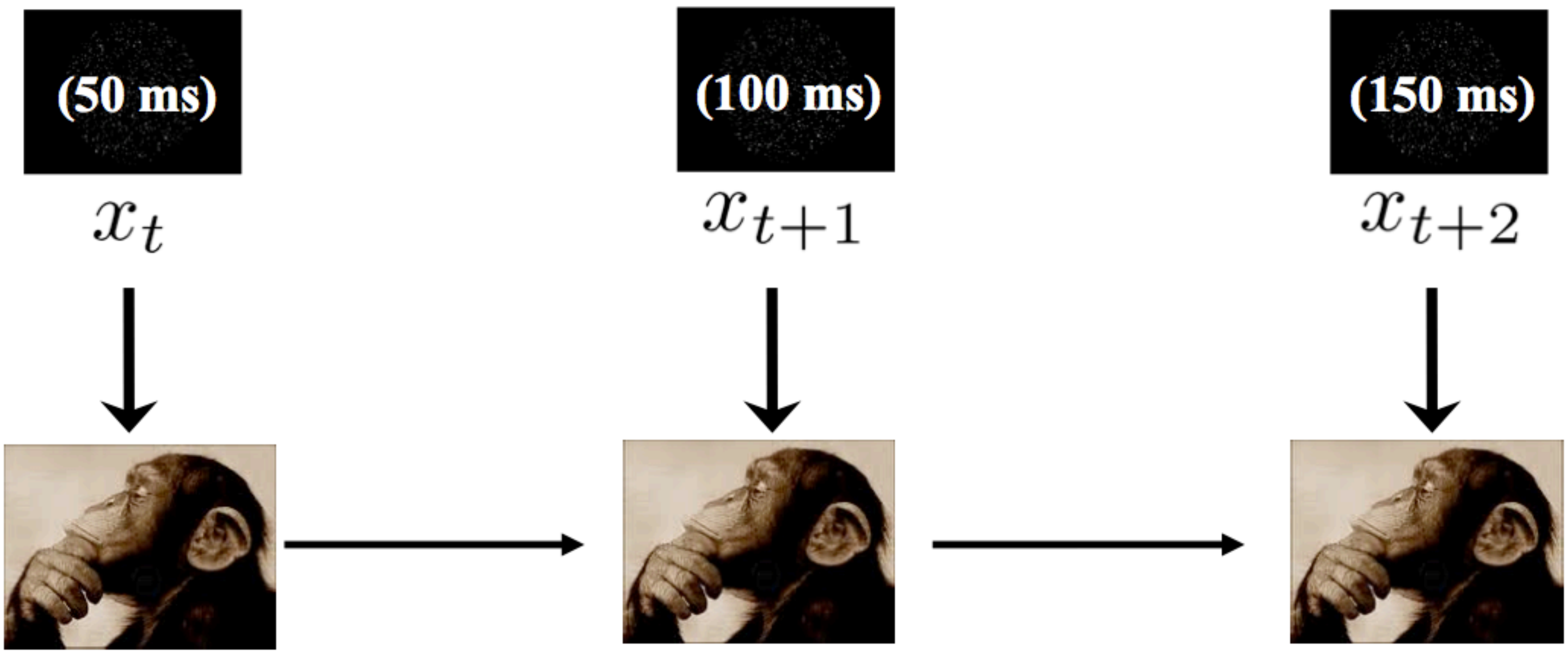
$$p(s|\mathbf{x}_t) = \frac{p(\mathbf{x}_t|s)p(s)}{p(\mathbf{x}_t)} = \frac{p(\mathbf{x}_t|s)p(s)}{\cancel{\int p(\mathbf{x}_t|s')p(s')ds'} \rightarrow Z} \propto p(\mathbf{x}_t|s)p(s) = p(s)\prod_{i=1}^{t} p(x_i|s)$$

$$= \frac{p(x_t|s, \cancel{\mathbf{x}_{t-1}})p(s|\mathbf{x}_{t-1})}{p(x_t|\mathbf{x}_{t-1})} = \frac{p(x_t|s)p(s|\mathbf{x}_{t-1})}{\cancel{\int p(x_t|s')p(s'|\mathbf{x}_{t-1})ds'} \rightarrow Z'} \propto p(x_t|s)p(s|\mathbf{x}_{t-1})$$

(online)

Sequential Update

A Running Example



$$P_0 = p(s=1) \qquad P_t = \frac{f_1(x_t)P_{t-1}}{Z_t} \qquad P_{t+1} = \frac{f_1(x_{t+1})P_t}{Z_{t+1}} \qquad P_{t+2} = \frac{f_1(x_{t+2})P_{t+1}}{Z_{t+2}}$$

Sequential Estimation, temporal independence

Contribution of the N^{th} data point, x_N

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}\right)
\end{aligned}
$$

correction given x_N

correction weight

old estimate

Learning as Inference: Kalman

- Basic Idea:

Make prediction based on previous data

↓

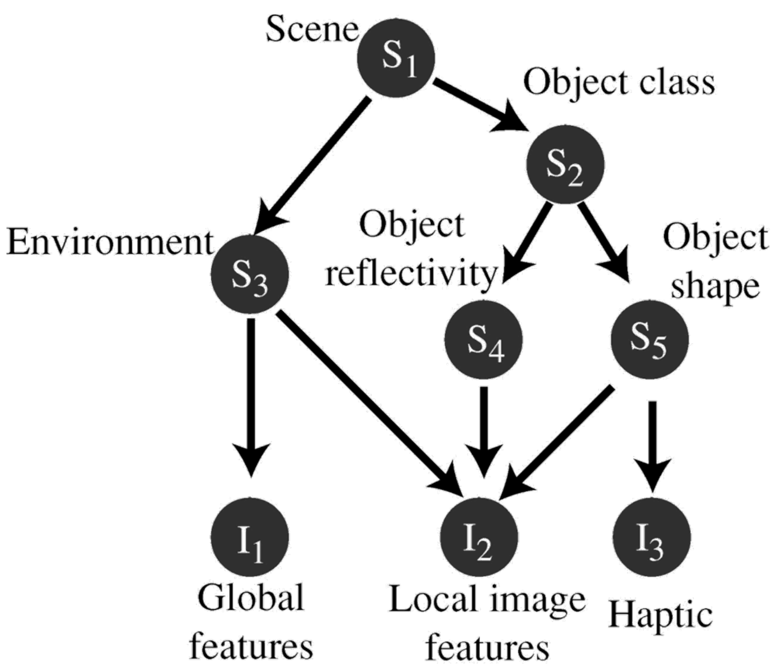
Take measurement

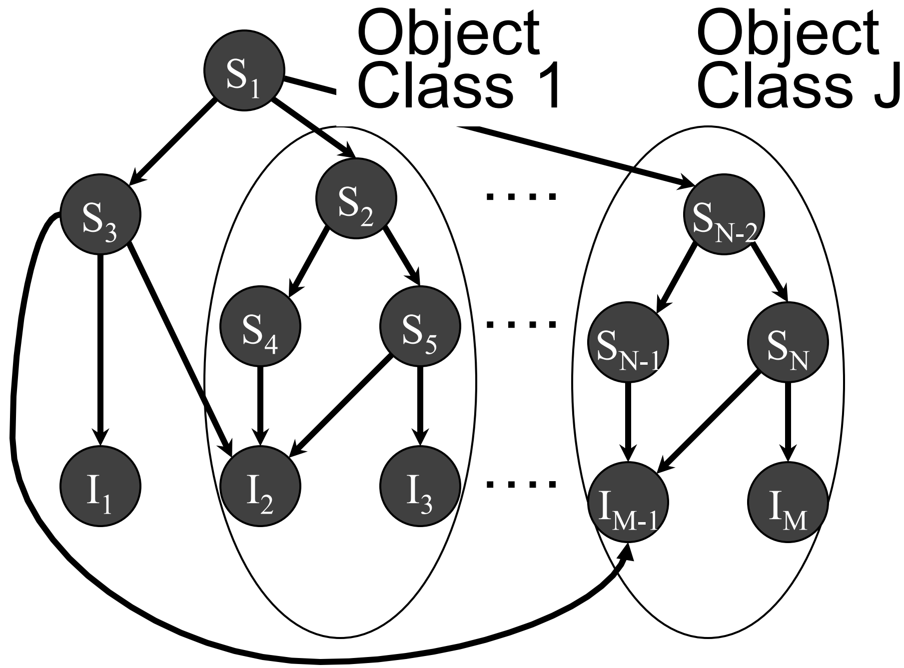
↓

Optimal estimate (ŷ) =
Prediction + (Kalman Gain) * (Measurement - Prediction)

Variance of estimate =
Variance of prediction * (1 – Kalman Gain)

Structure Learning: Inferring variable relations

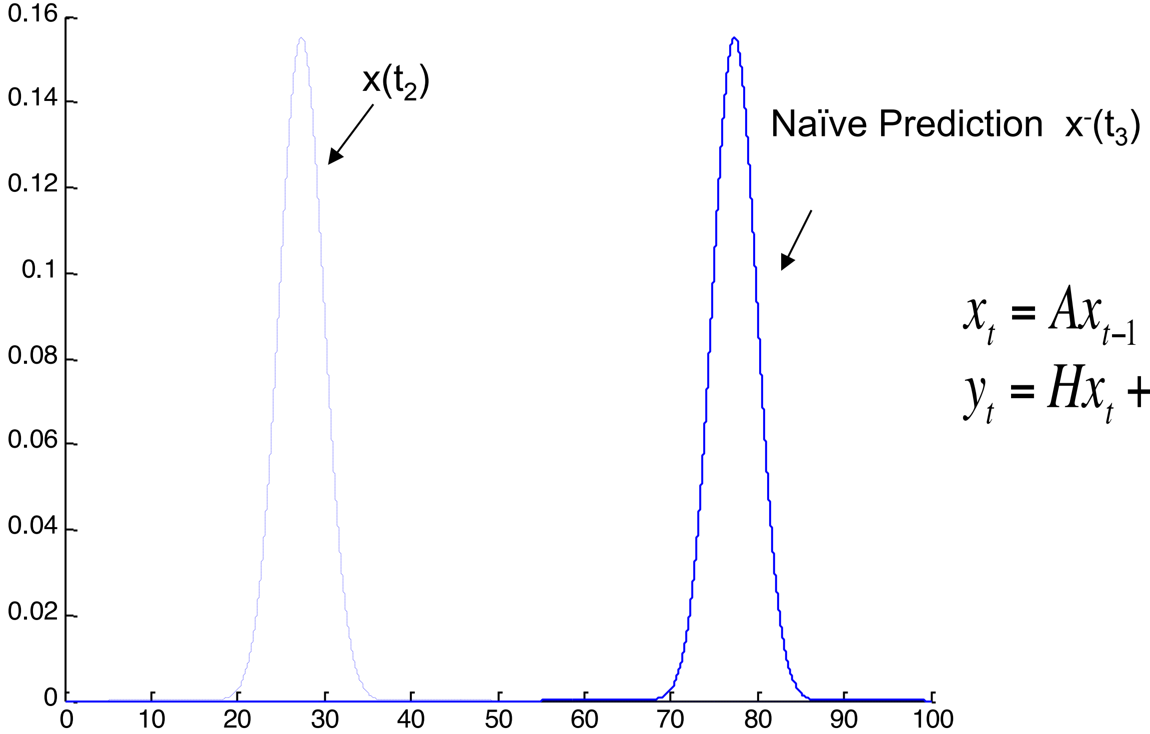




Learning the graph is often MORE important!!

Conceptual Overview

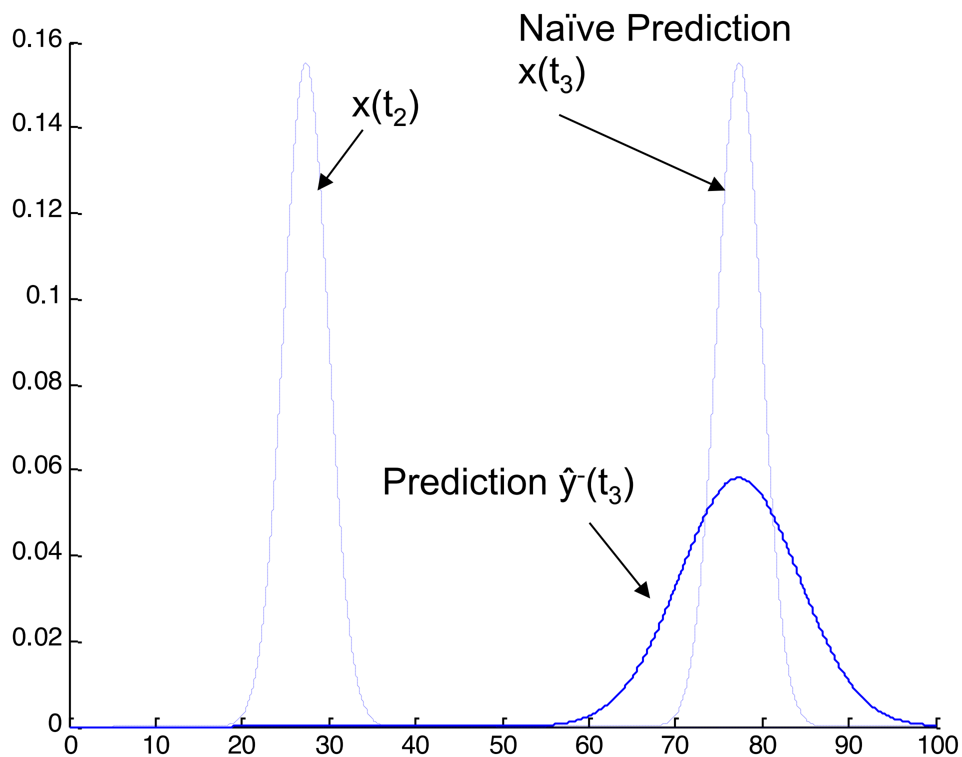
Predict new location if an observer was moving?



$$x_t = Ax_{t-1} + Bx + \omega_{walk}$$
$$y_t = Hx_t + \omega_{sensory}$$

- At time t_3, observer moves with velocity dy/dt=u
- Naïve approach: Shift probability to the right to predict
- This would work if we knew the velocity exactly (perfect model)

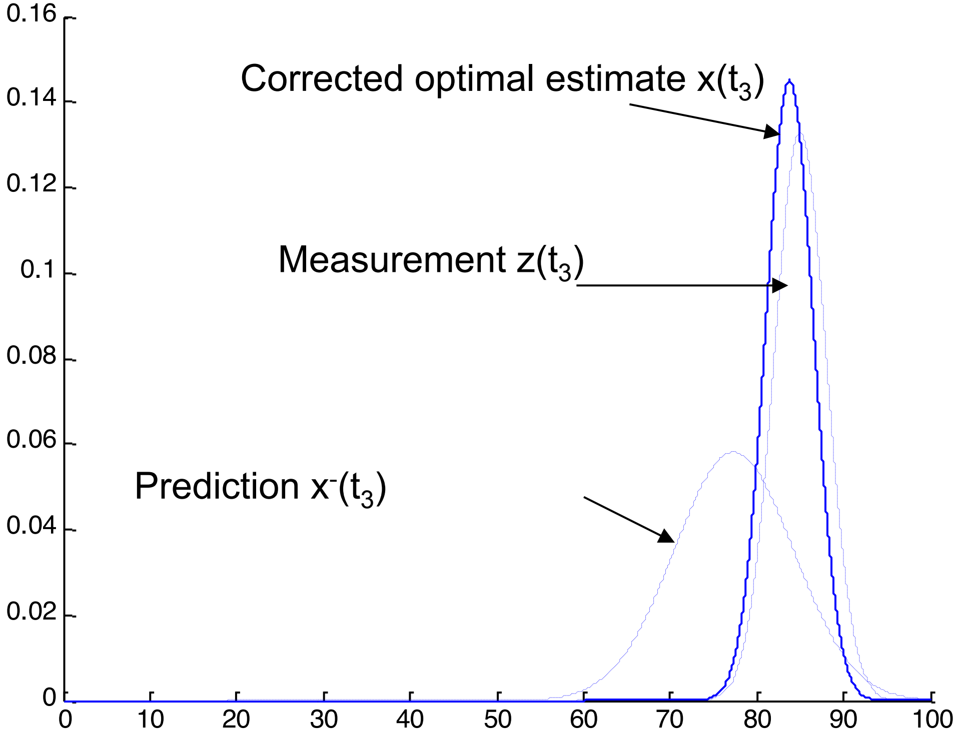
Conceptual Overview



But you may not be so sure about the exact velocity

- Better to assume imperfect model by adding Gaussian noise
- $dy/dt = u + w$
- Distribution for prediction moves and spreads out

Conceptual Overview



- Now we take a measurement at t_3
- Need to once again correct the prediction
- Same as before

Conceptual Overview

- Initial conditions (x_{k-1} and σ_{k-1})
- Prediction (x^-_k , σ^-_k)
  - Use initial conditions and model (eg. constant velocity) to make prediction
- Measurement (z_k)
  - Take measurement
- Correction (x_k , σ_k)
  - Use measurement to correct prediction by 'blending' prediction and residual – always a case of merging only two Gaussians
  - Optimal estimate with smaller variance

Blending Factor

- If we are sure about measurements:
  - Measurement error covariance (R) decreases to zero
  - K decreases and weights residual more heavily than prediction
- If we are sure about prediction
  - Prediction error covariance P^{-}_{k} decreases to zero
  - K increases and weights prediction more heavily than residual

The set of Kalman Filtering Equations in Detail

Prediction (Time Update)

(1) Project the state ahead

$$\hat{y}^-_k = Ay_{k-1} + Bu_k$$

(2) Project the error covariance ahead

$$P^-_k = AP_{k-1}A^T + Q$$

Correction (Measurement Update)

(1) Compute the Kalman Gain

$$K = P^-_k H^T (HP^-_k H^T + R)^{-1}$$

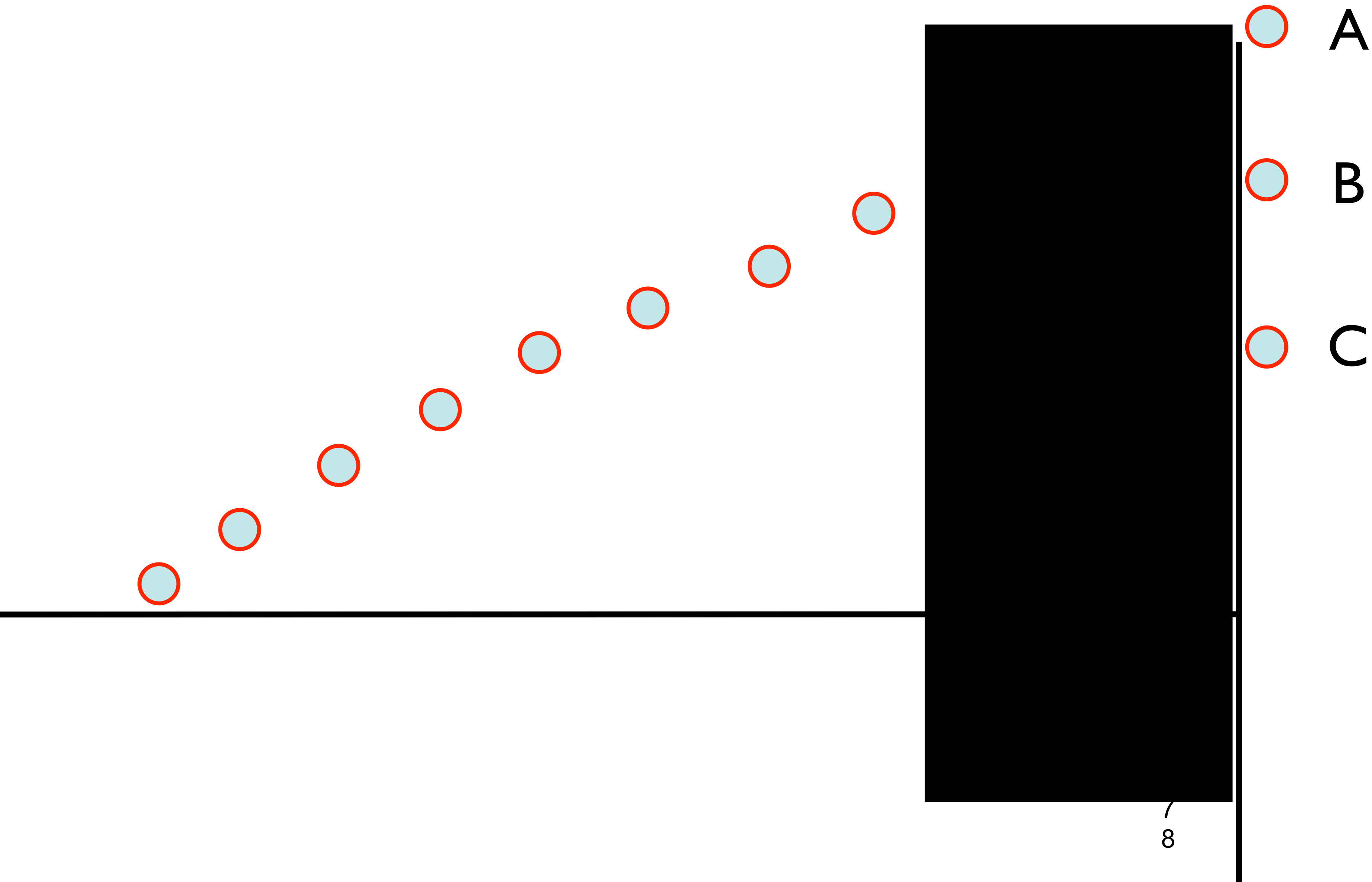
(2) Update estimate with measurement z_k

$$\hat{y}_k = \hat{y}^-_k + K(z_k - H\,\hat{y}^-_k)$$

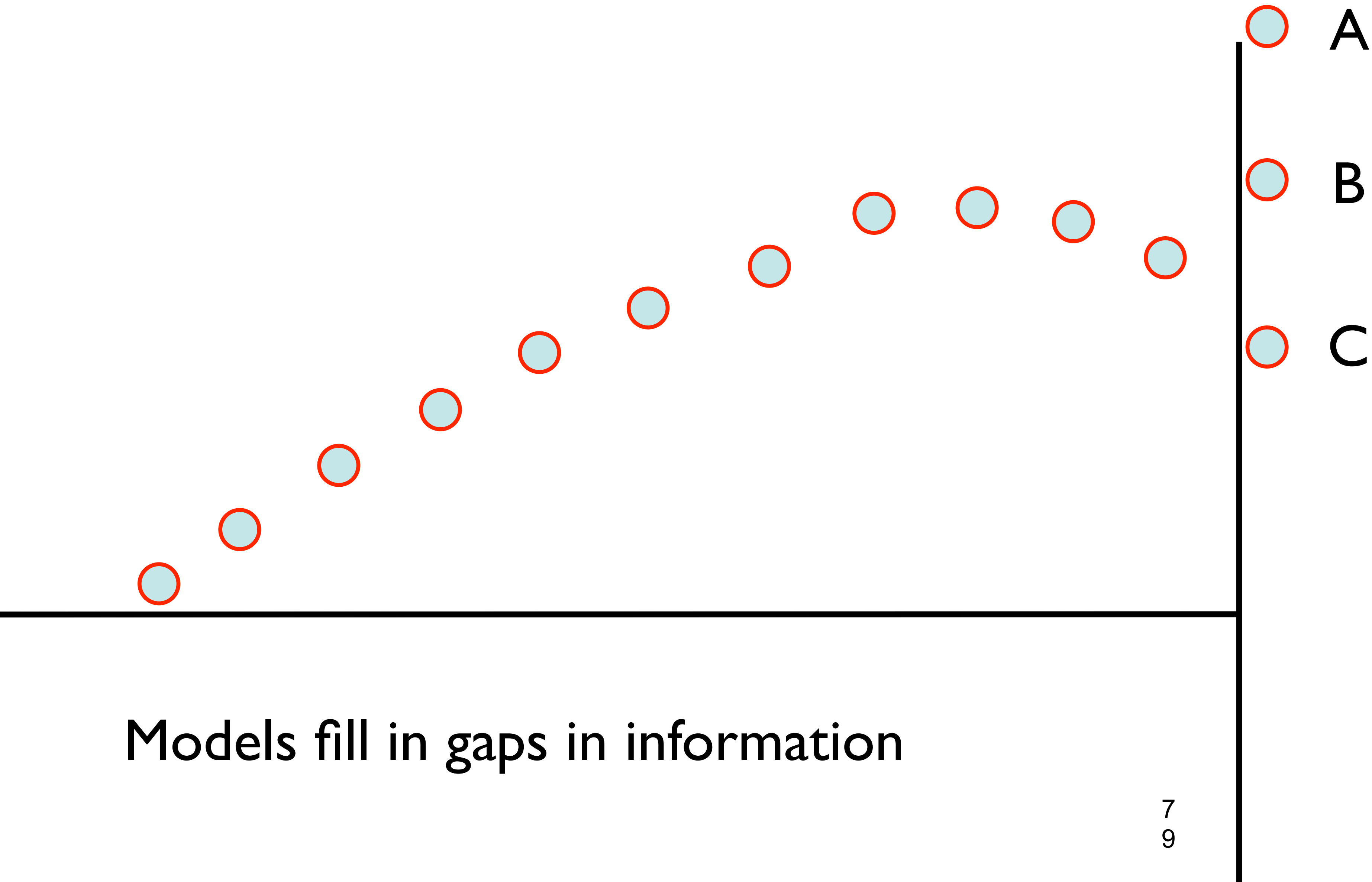
(3) Update Error Covariance

$$P_k = (I - KH)P^-_k$$

Model example

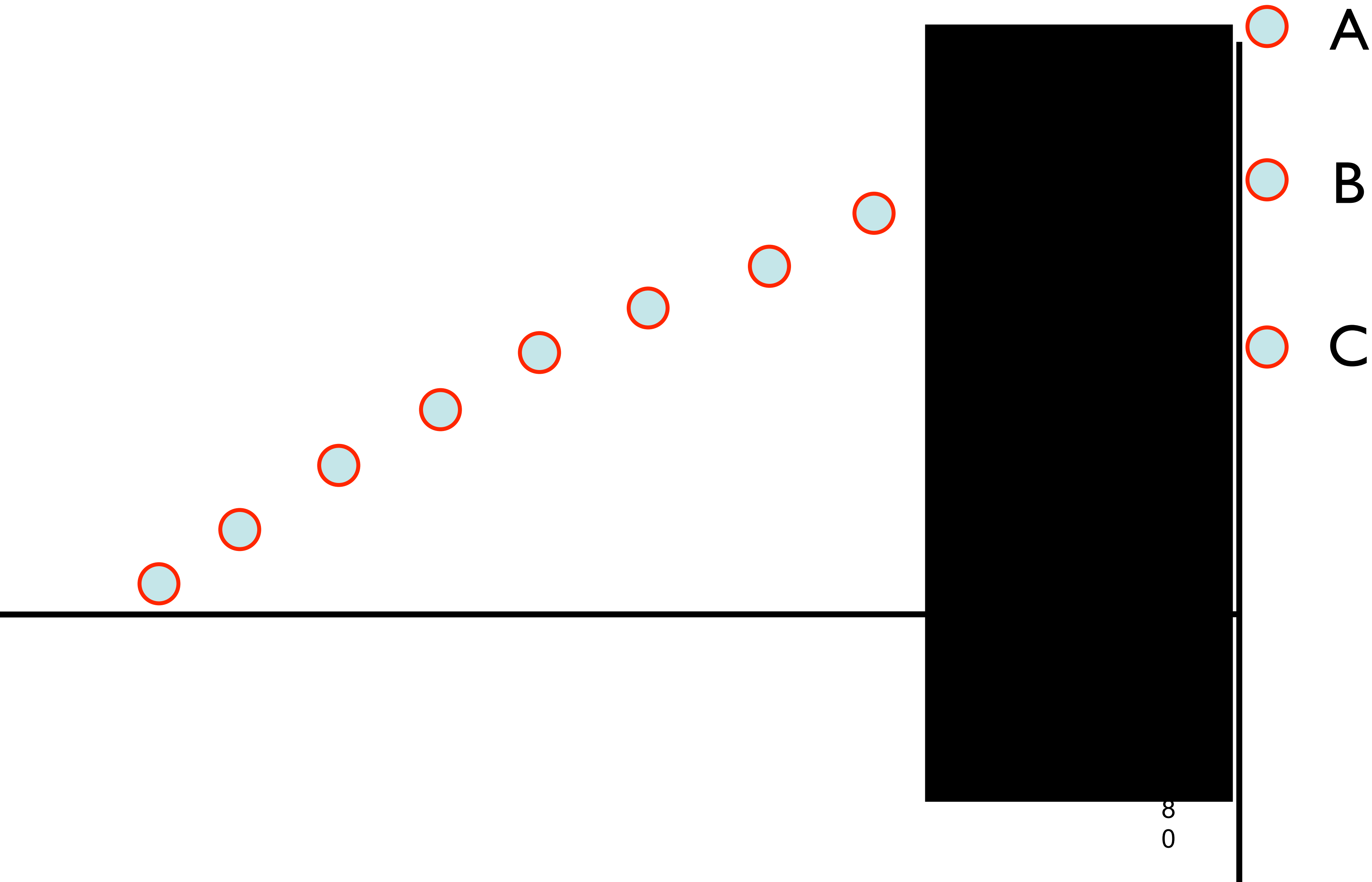


Model Example

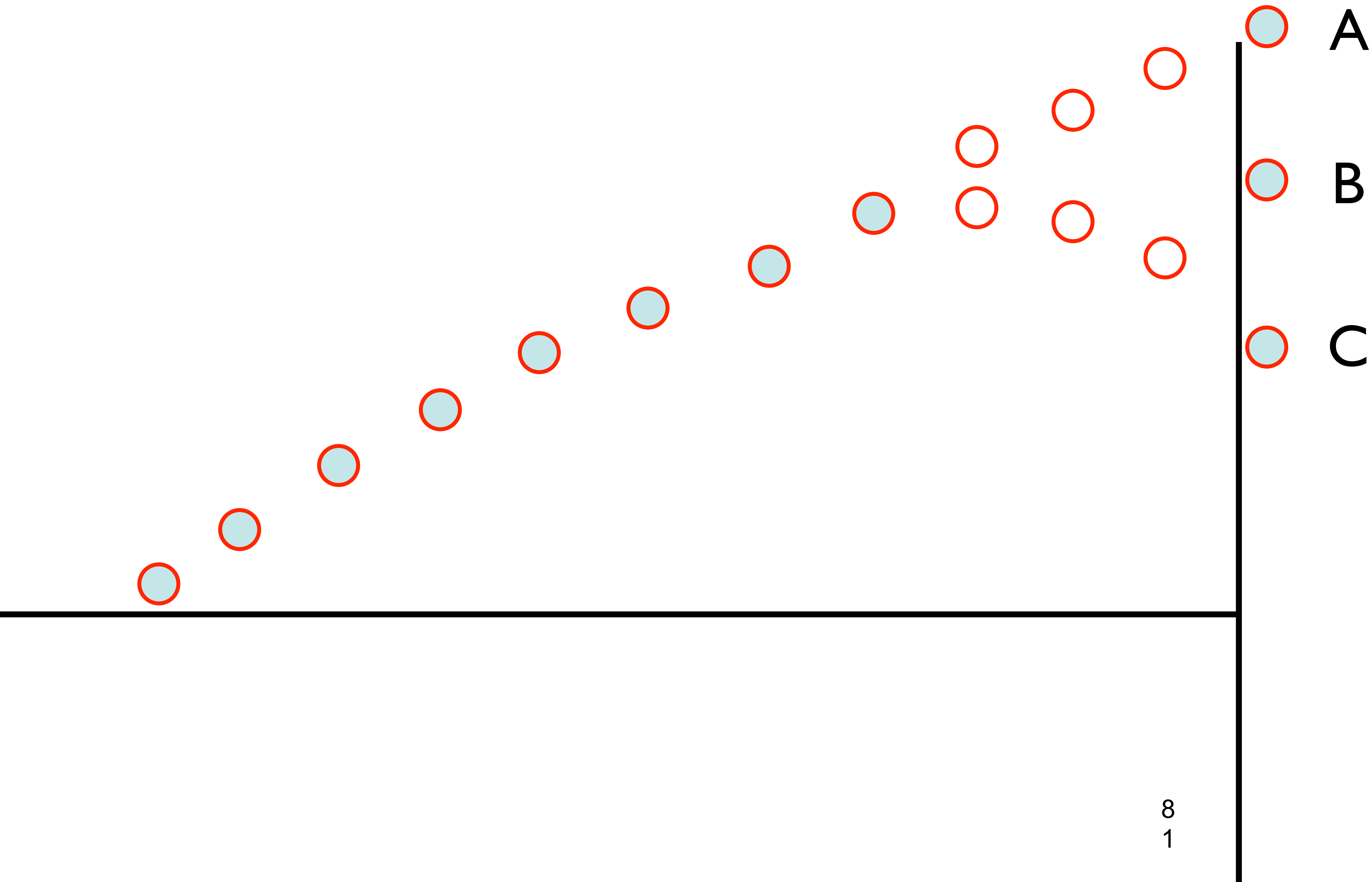


Models fill in gaps in information

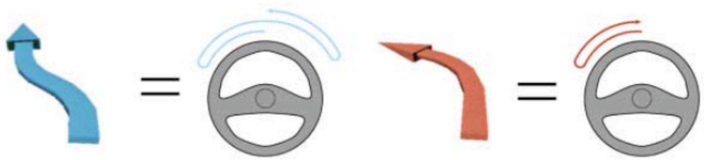
Model example

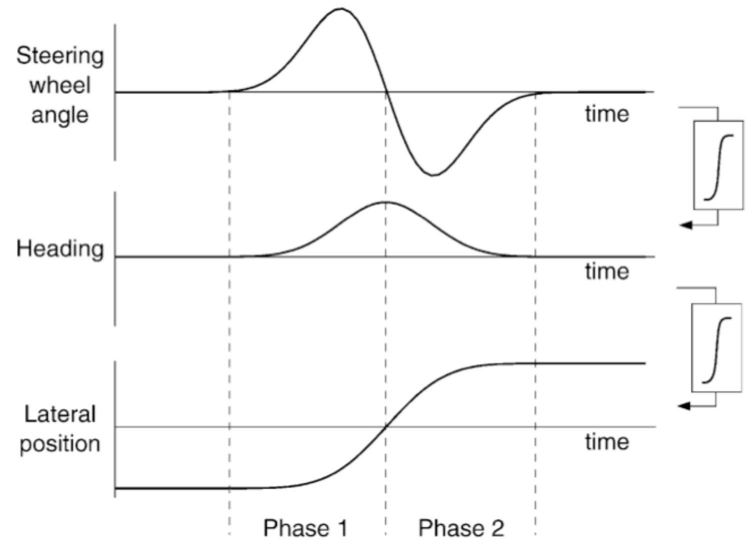


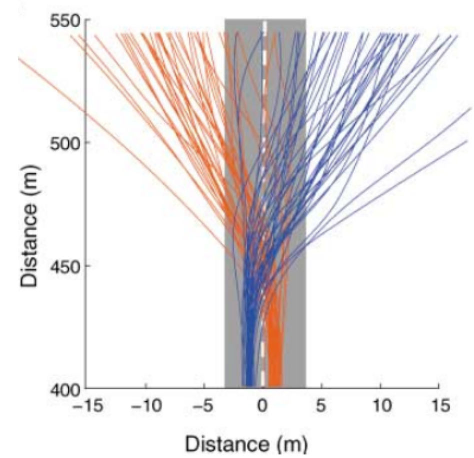
Extrapolation depends on model



Do we have internal models for everything? NO!



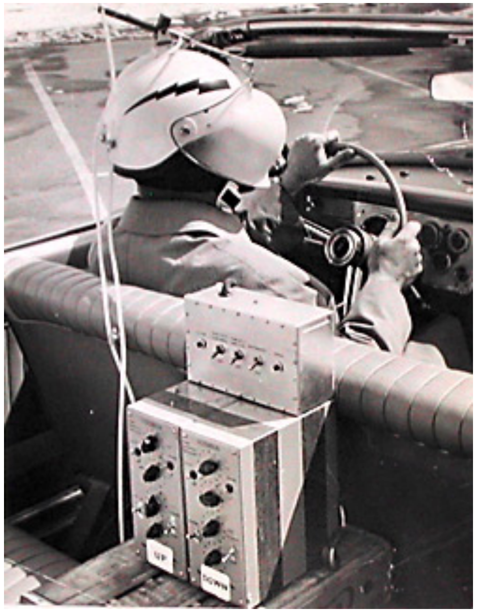




Classic example
of a failure to learn
Internal model

Wallis, G.M., A. Chatziastros and H.H. Bülthoff: An unexpected role for visual feedback in vehicle steering control. Current Biology 12, 295-299 (2002)

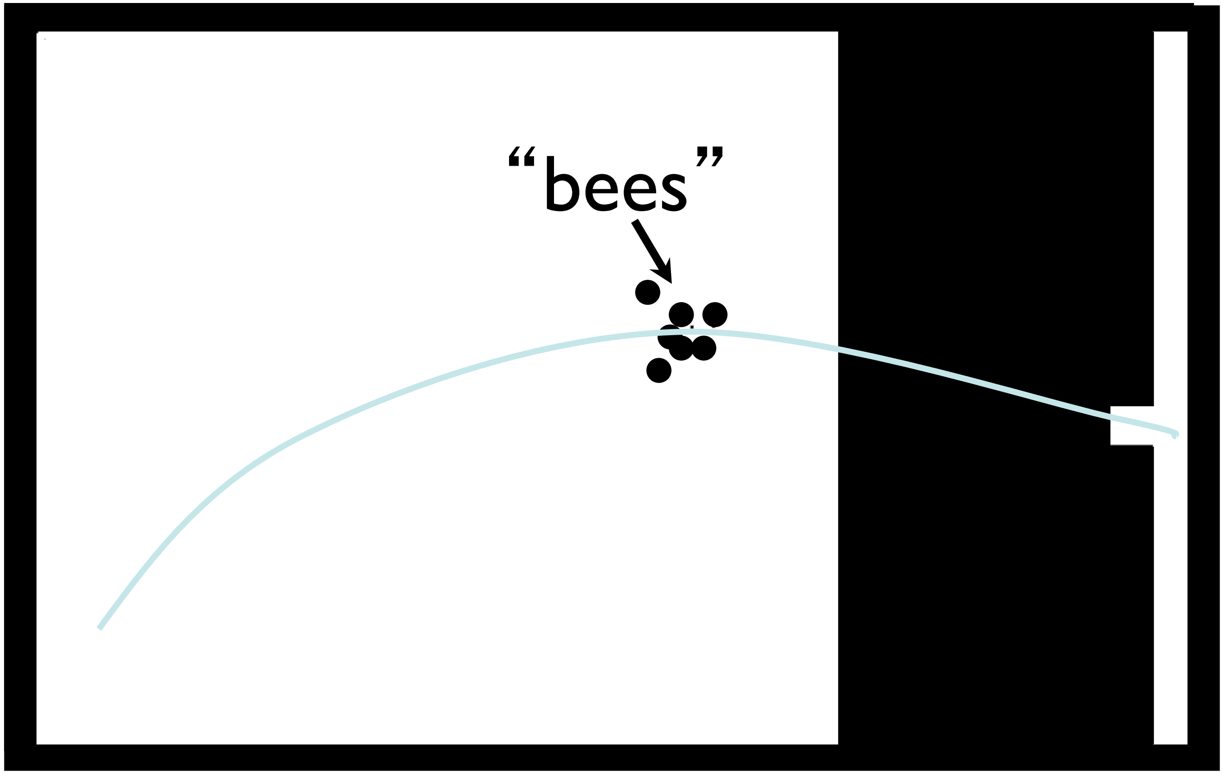
Prediction - the reason for models



http://www.youtube.com/watch?v=kOgusISPpqo

Moving Dot task

- Prediction task
- Watch the dots move
- Position "bucket" to catch the emerging dots



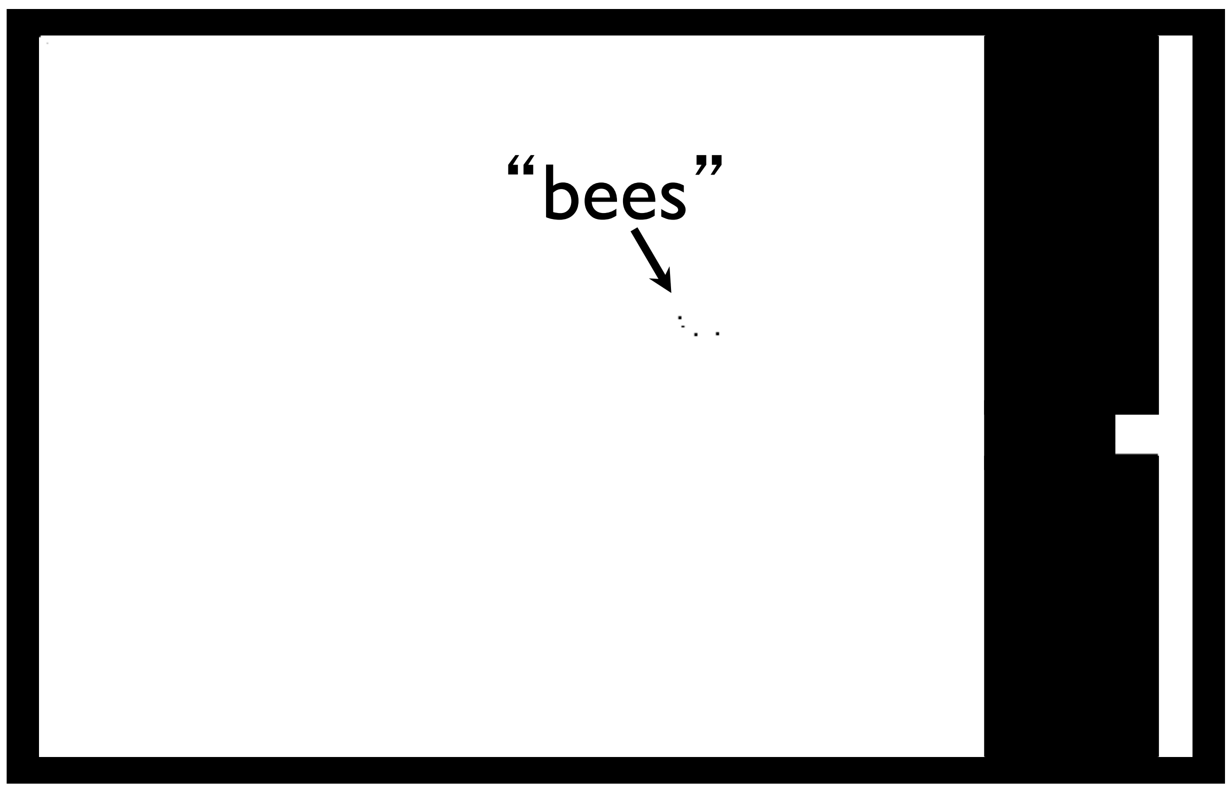
Moving Dot task

- Prediction task
- Watch the dots move
- Position "bucket" to catch the emerging dots

Stimuli designed to be optimal for matched Kalman filter

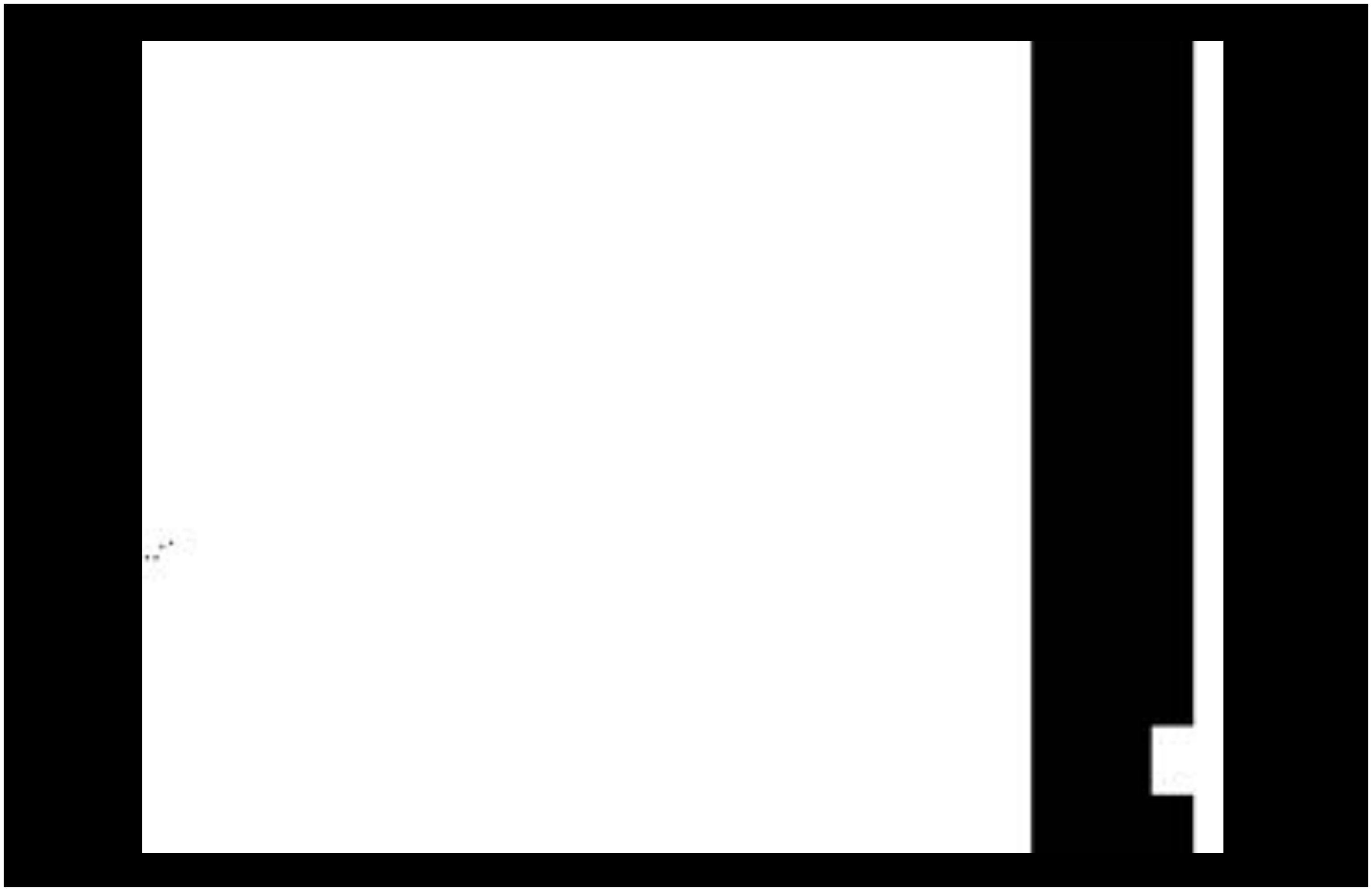
Moving Dot task

- Capture the "bees"

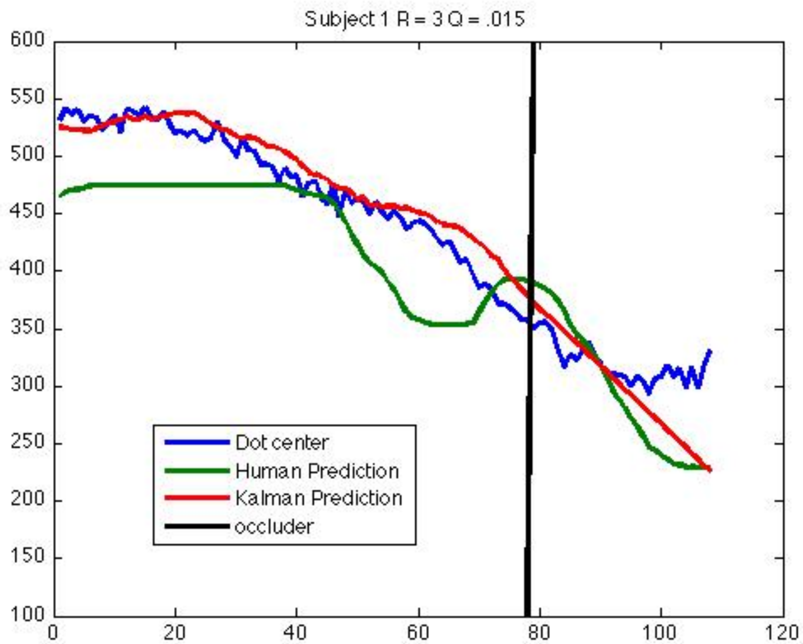


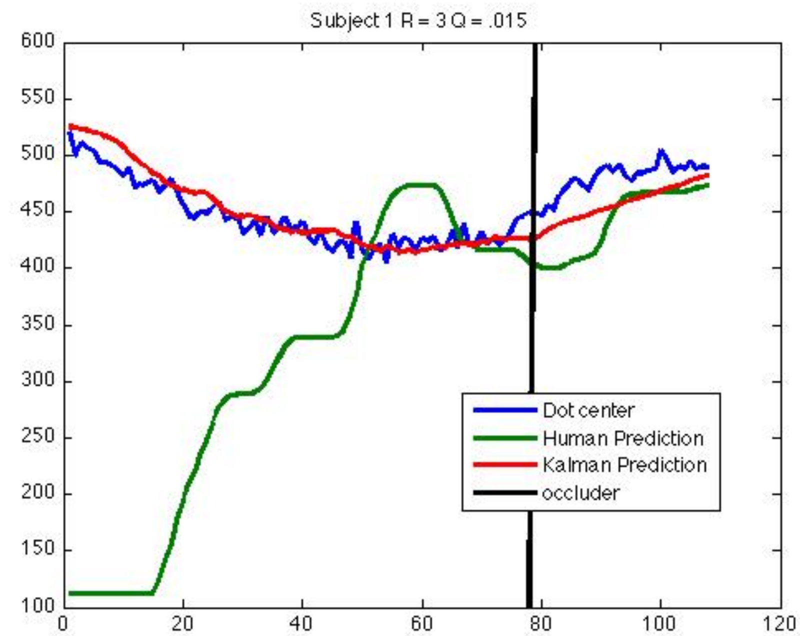
Trajectory = ~random walk

movie demo

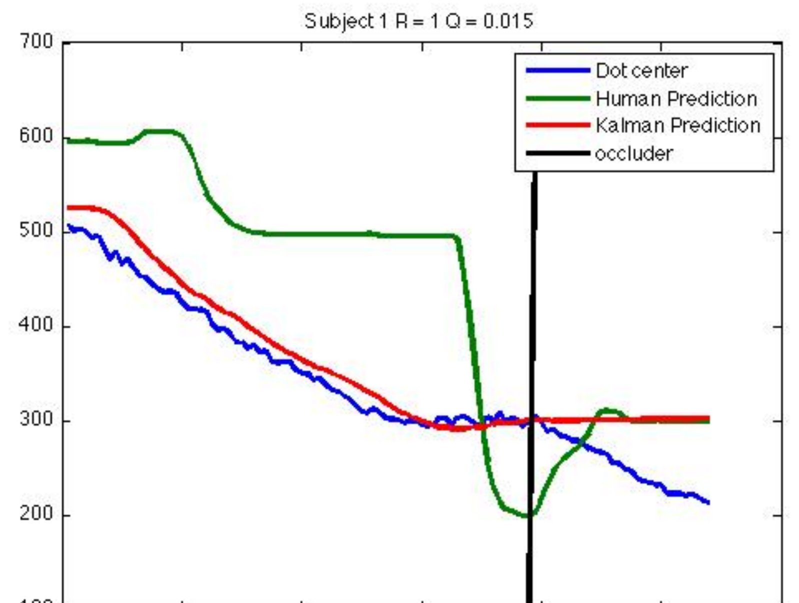


Humans vs. Kalman Filter

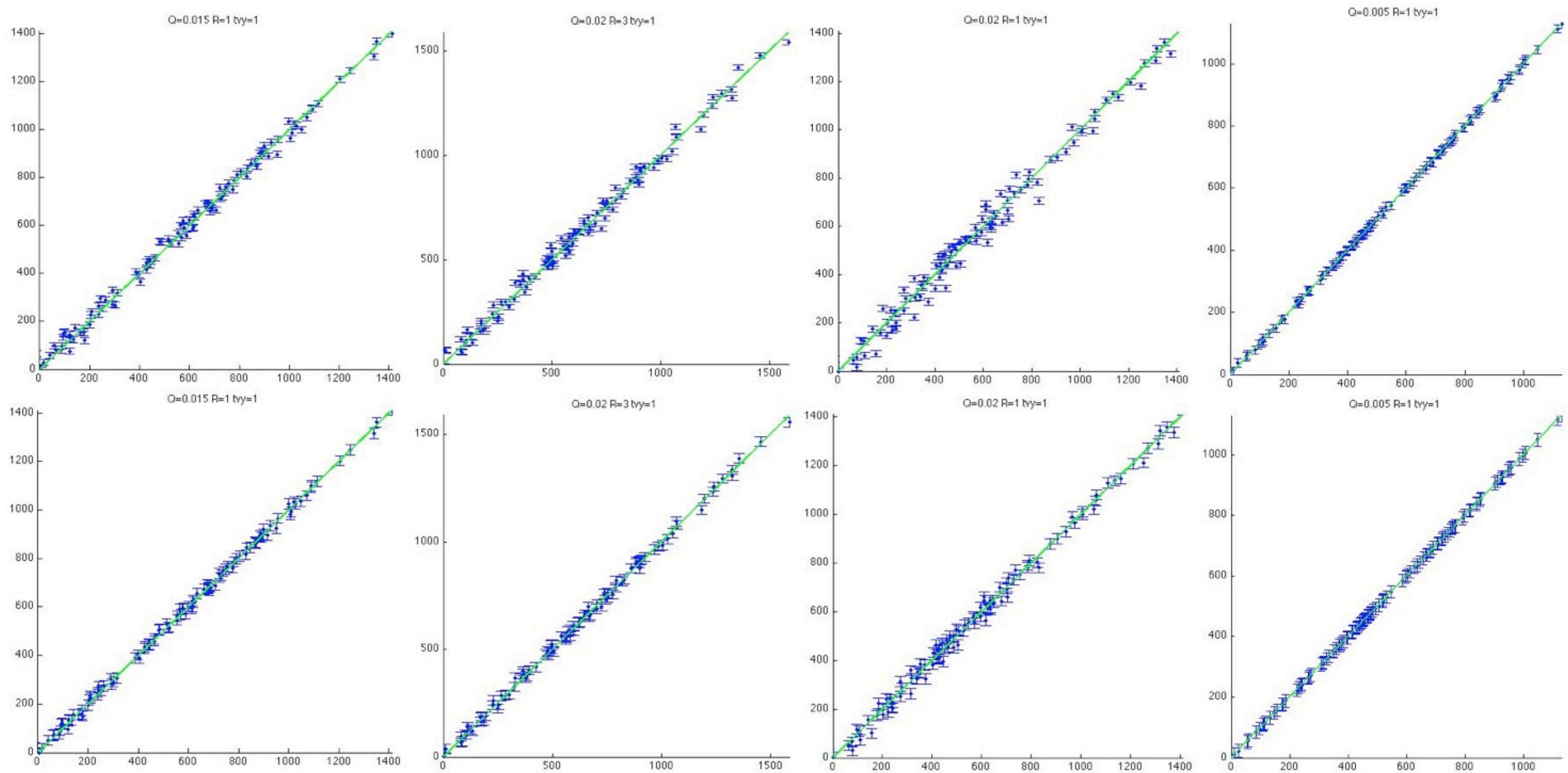




- Demonstration of the task, human vs. filter performance
- Kalman filter predicts human behavior well



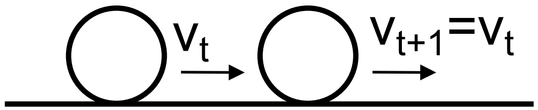
Matched Kalman excellent predictor



What are Human default Motion Models?

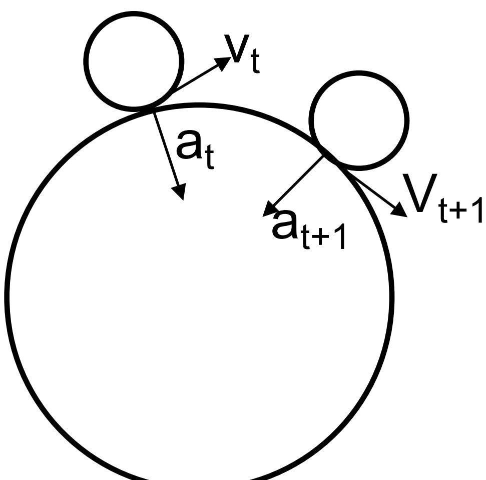
<u>Object velocity:</u> speed 5 m/s
direction south

1. Constant velocity (**CV**)



-maintain speed and direction

2. Constant acceleration (**CA**)



-constant change in speed and/or direction

Motion extrapolation task

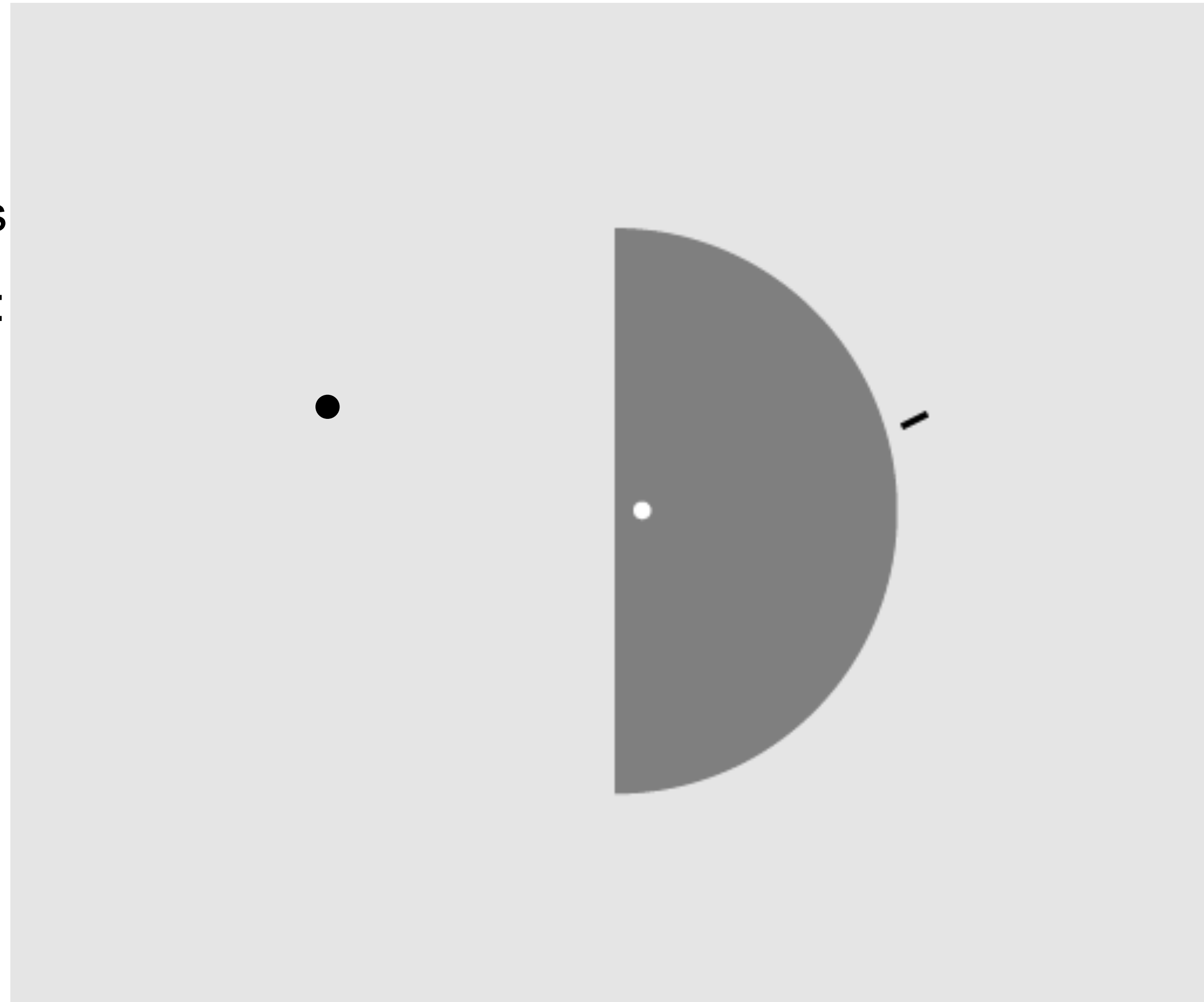
-Fixation

-After 500ms dot travels

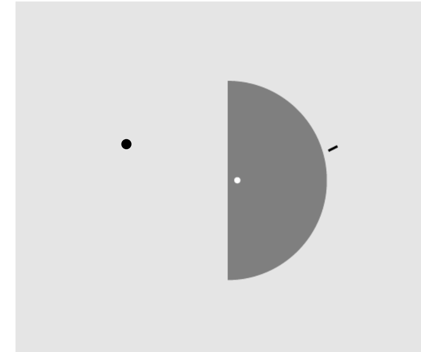
-Extrapolation judgment:
“above” or “below”

-No reemergence;
no feedback

-Determine the PSE
based on staircase
procedure



Motion extrapolation: Kalman filters for simple motions



Parameters of dot motion:

$$x_k = [x, y, vx, vy, ax, ay]_k^T$$

position velocity acceleration

Process:

True state: $x_k = A_k x_{k-1} + w_{k-1}$

$$\begin{pmatrix} x_k \\ y_k \\ vx_k \\ vy_k \\ ax_k \\ ay_k \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta & 0 & 0 & 0 \\ 0 & 1 & 0 & \Delta & 0 & 0 \\ 0 & 0 & 1 & 0 & \Delta & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k-1} \\ y_{k-1} \\ vx_{k-1} \\ vy_{k-1} \\ ax_{k-1} \\ ay_{k-1} \end{pmatrix} + \begin{pmatrix} w_{x_{k-1}} \\ w_{y_{k-1}} \\ w_{vx_{k-1}} \\ w_{vy_{k}-1} \\ w_{ax_{k-1}} \\ w_{ay_{k-1}} \end{pmatrix}$$

"w" ~N(0,Q),
"Q" = covariance; reflects trust in prior ("A")

Q = 0 → complete trust

"A" represents the prior model in the absence of data

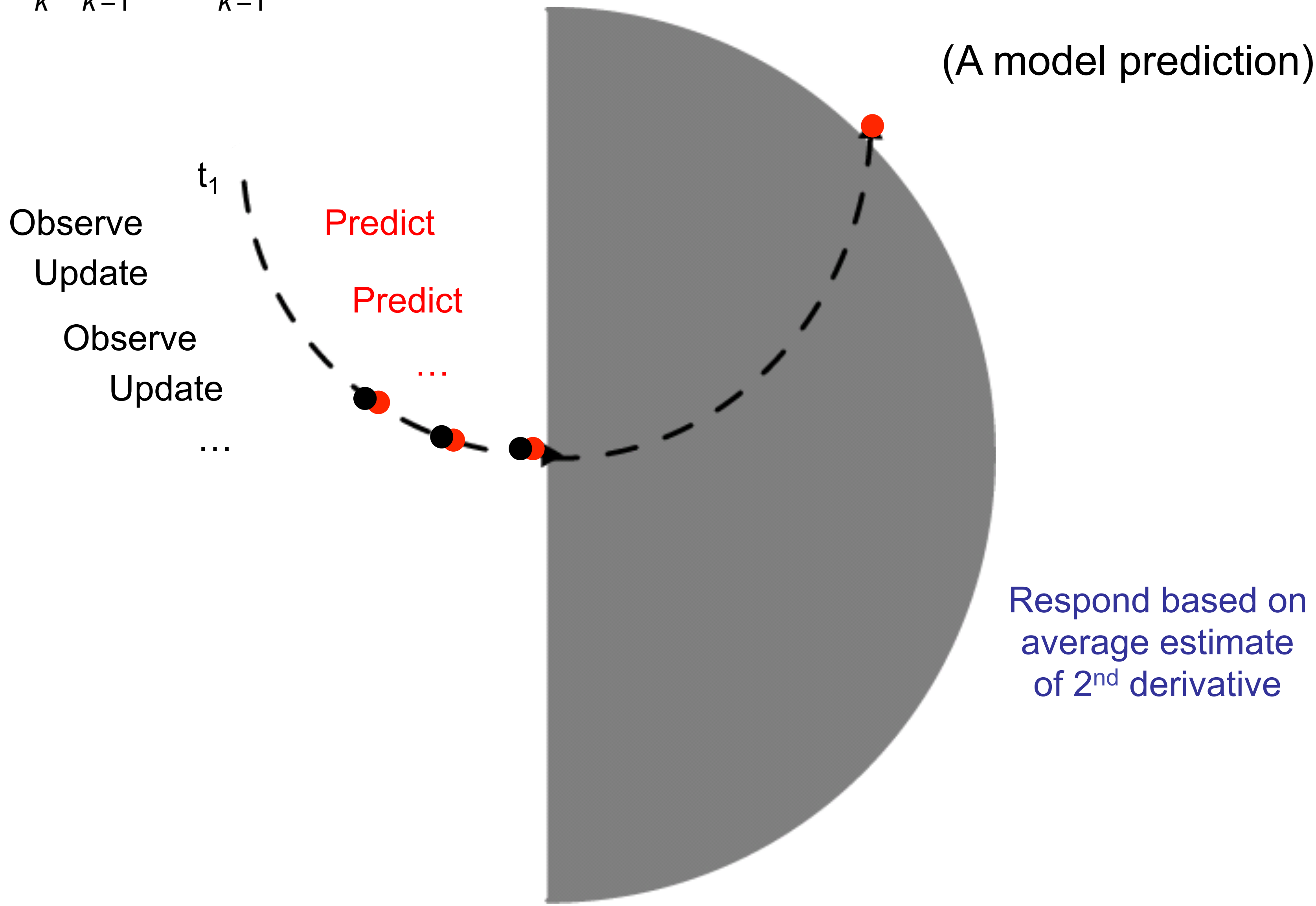
→ **CV:** constant speed & direction: Linear motion prior

→ **CA:** constant change in direction: Circular motion prior

Motion extrapolation: Model behavior

CA prediction using a Kalman filter

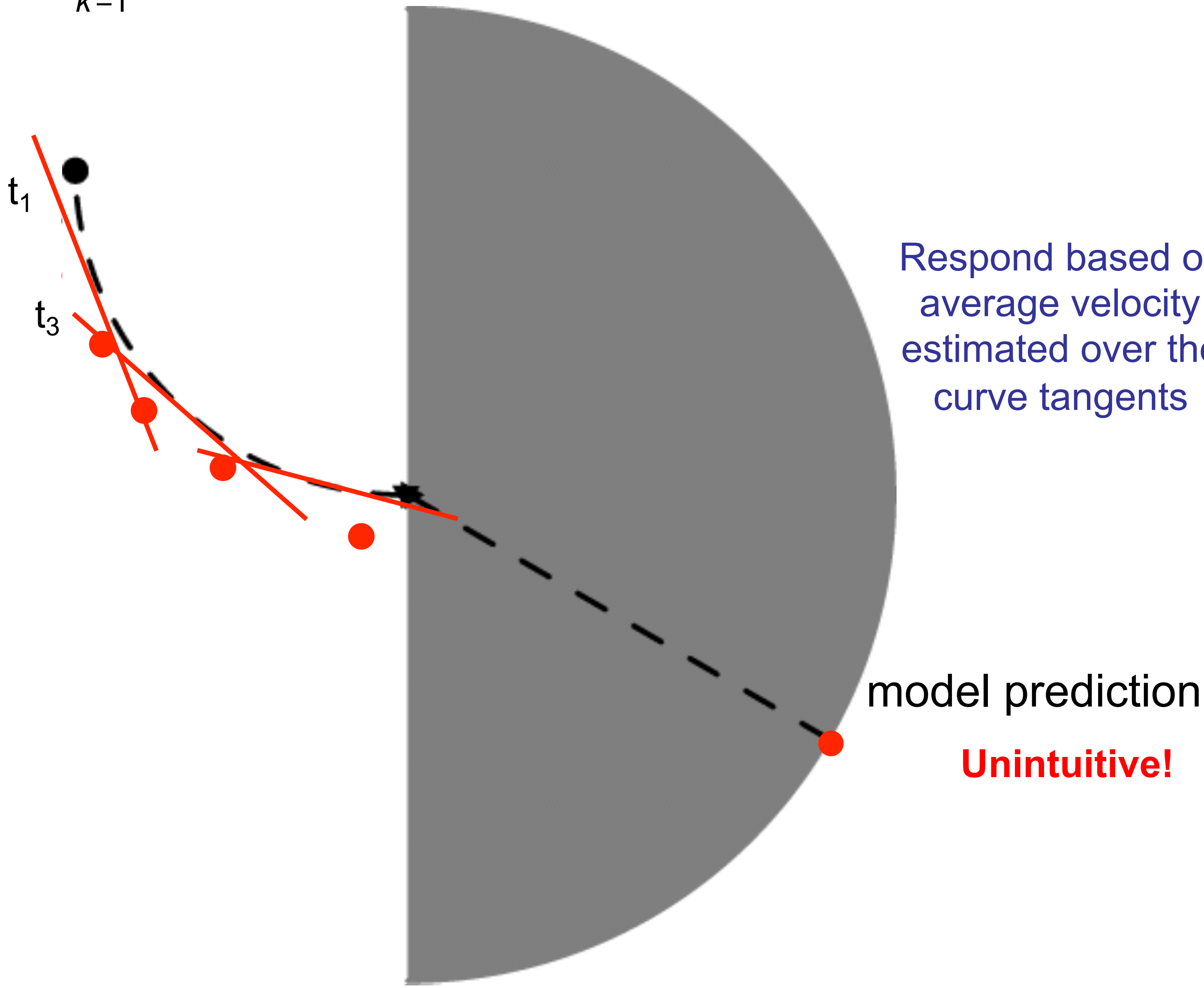
$$x_k = A_k x_{k-1} + w_{k-1}$$



Motion extrapolation: Model behavior

CV prediction using a Kalman filter

$$x_k = A_k x_{k-1} + w_{k-1}$$

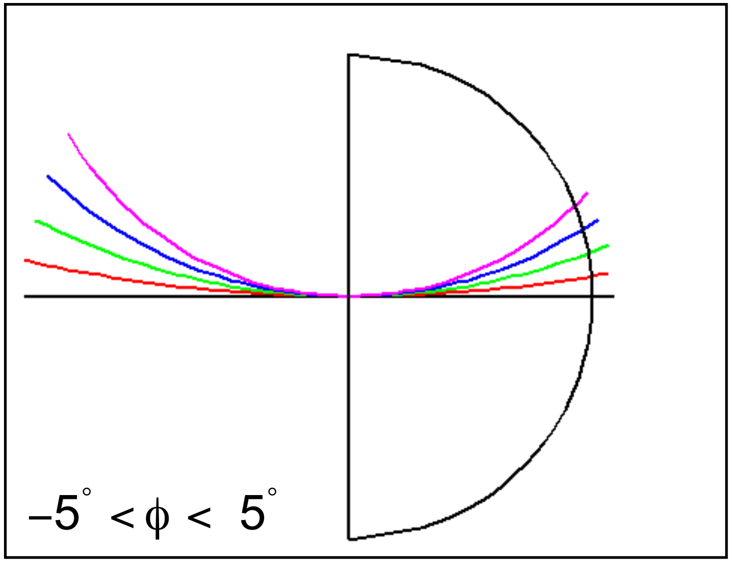


Respond based on average velocity estimated over the curve tangents

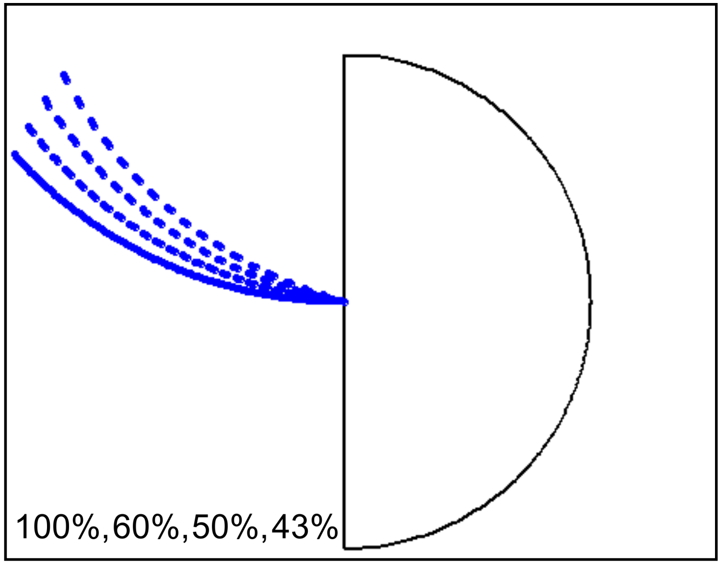
Unintuitive!

Stimulus manipulations:

Path curvature (5)



Motion sampling (4)



-Dot speed: 5 deg/s (constant)

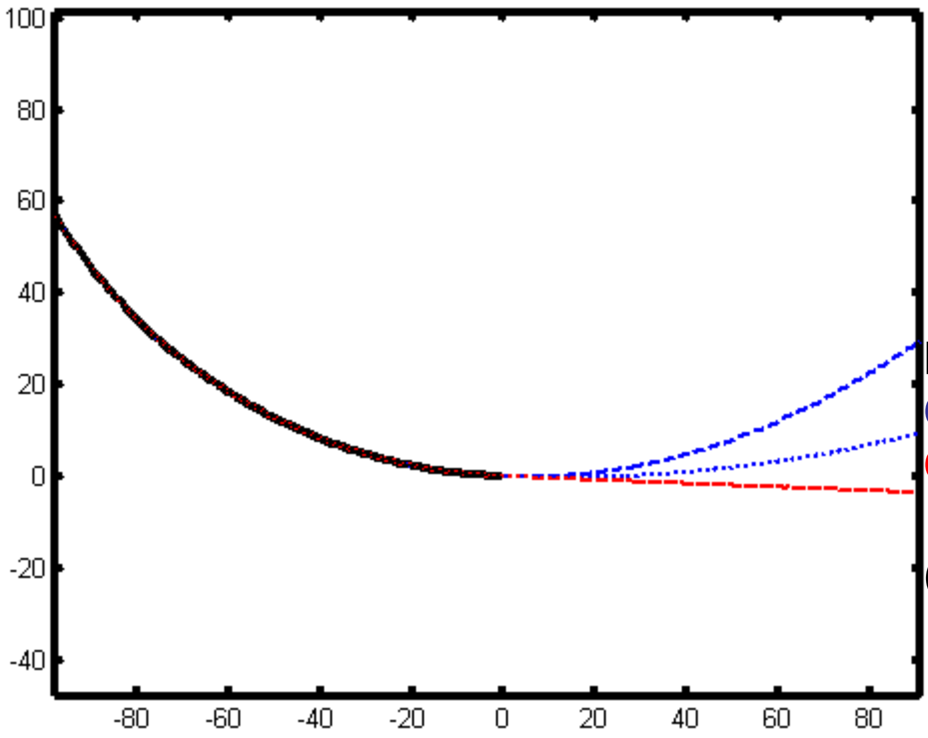
-2 staircases (i.e. 1U-2D, 2U-1D) per condition (curvature x sampling)
-100 trials per staircase

-10 participants unaware of the purpose of the experiment

Motion extrapolation: Model behavior

The simple linear process predicts a wide range of behaviors by varying:

i. The specific internal model (CA, CV)
ii. Trust in model predictions vs. measurements



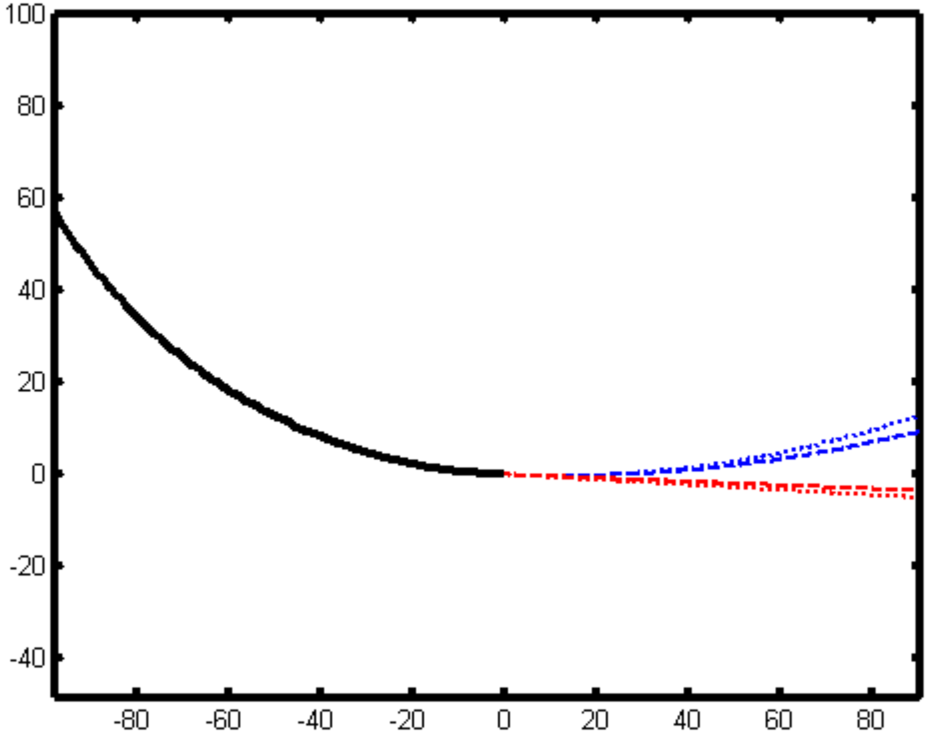
Decreased trust ("Q"):
CA – flatter extrapolation
CV – no change

(Less of the curve is used)

Motion extrapolation: Model behavior

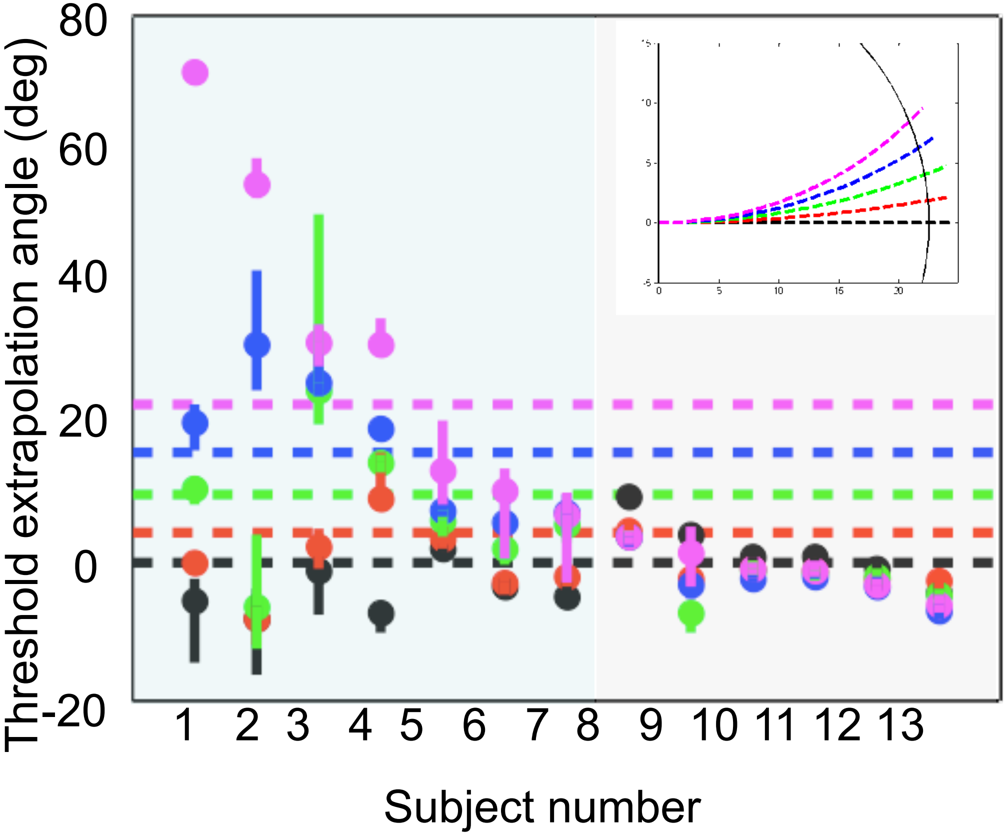
The model predicts a wide range of behaviors by varying:

i. The specific internal model (CA, CV)
ii. Trust in the model
iii. Motion sampling

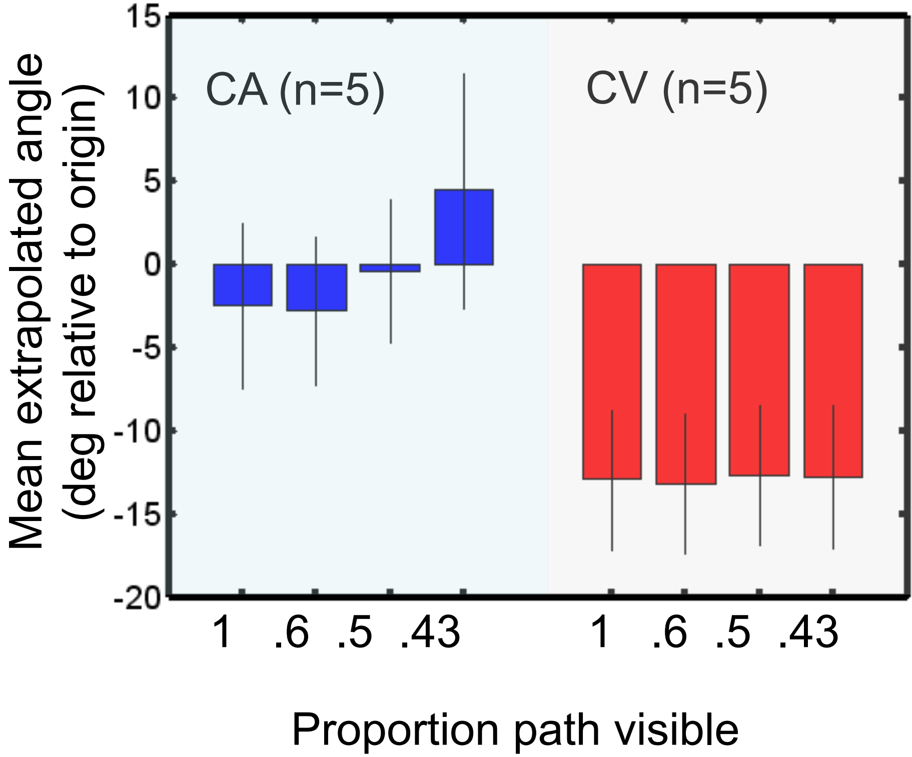


Sparser sampling
CA – slightly more curved
CV – little change

Results

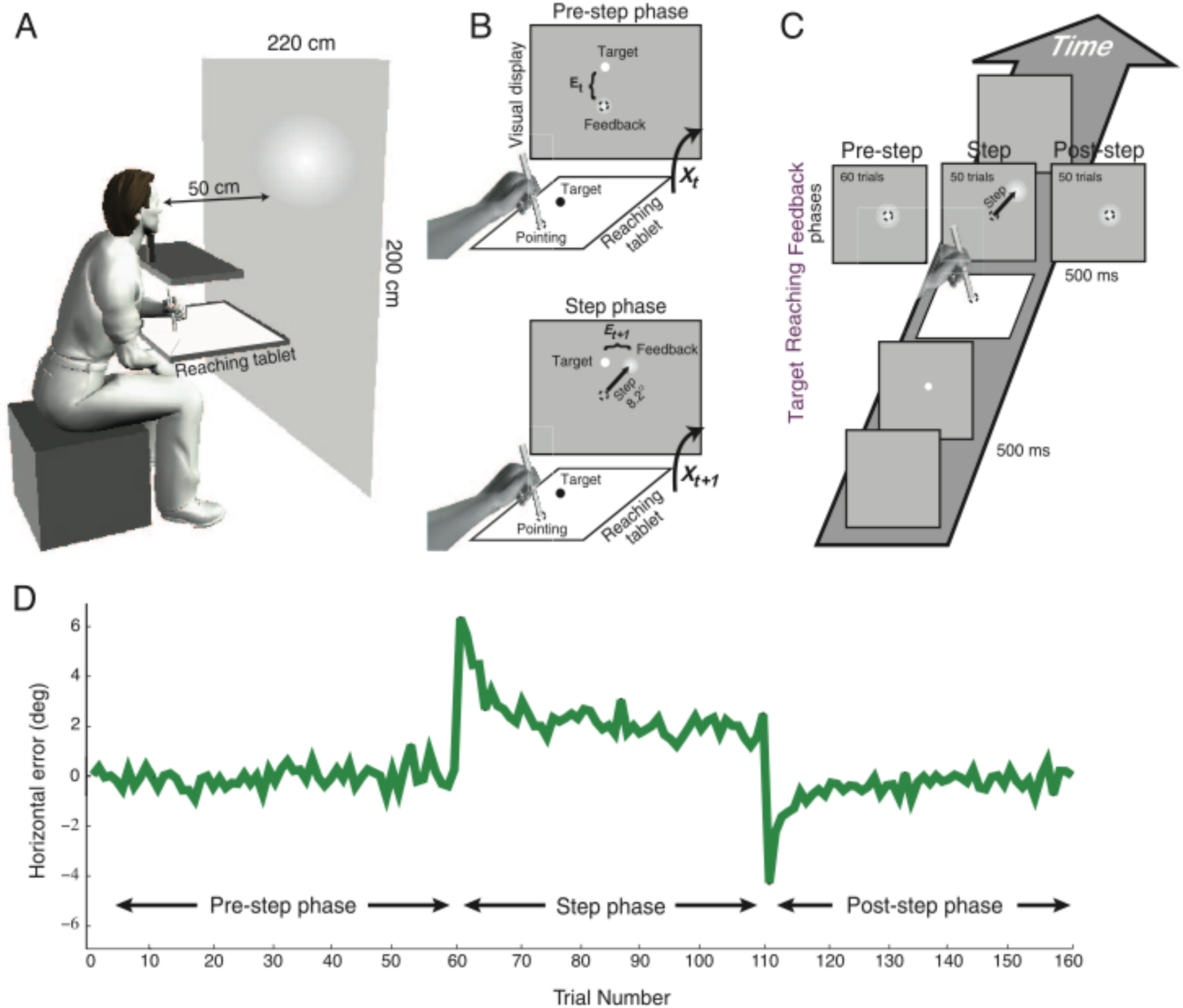


80
60
40
20
0
-20
Threshold extrapolation angle (deg)
1 2 3 4 5 6 7 8 9 10 11 12 13
Subject number

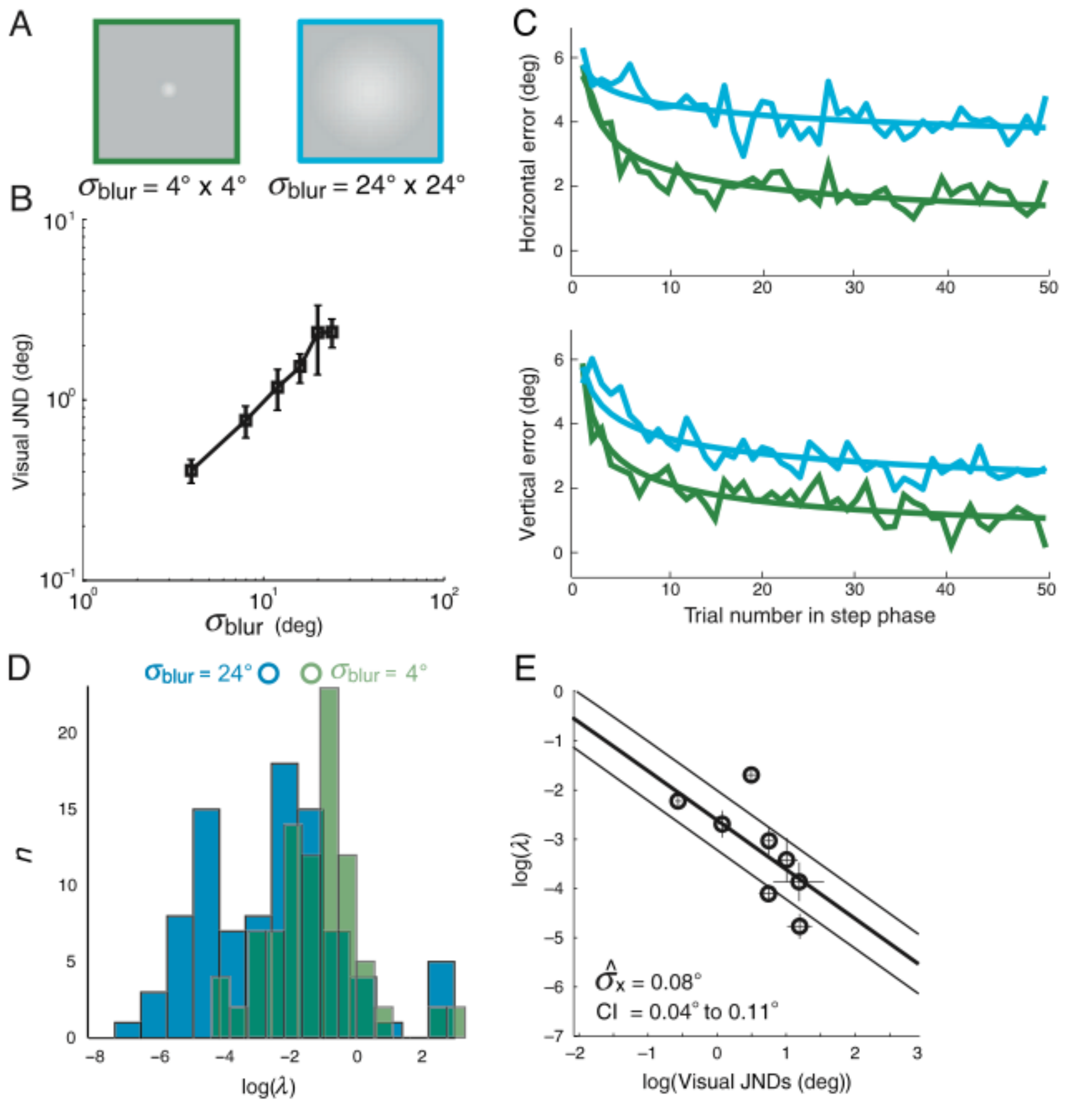


CA (n=5)
CV (n=5)
15
10
5
0
-5
-10
-15
-20
Mean extrapolated angle
(deg relative to origin)
1 .6 .5 .43 1 .6 .5 .43
Proportion path visible

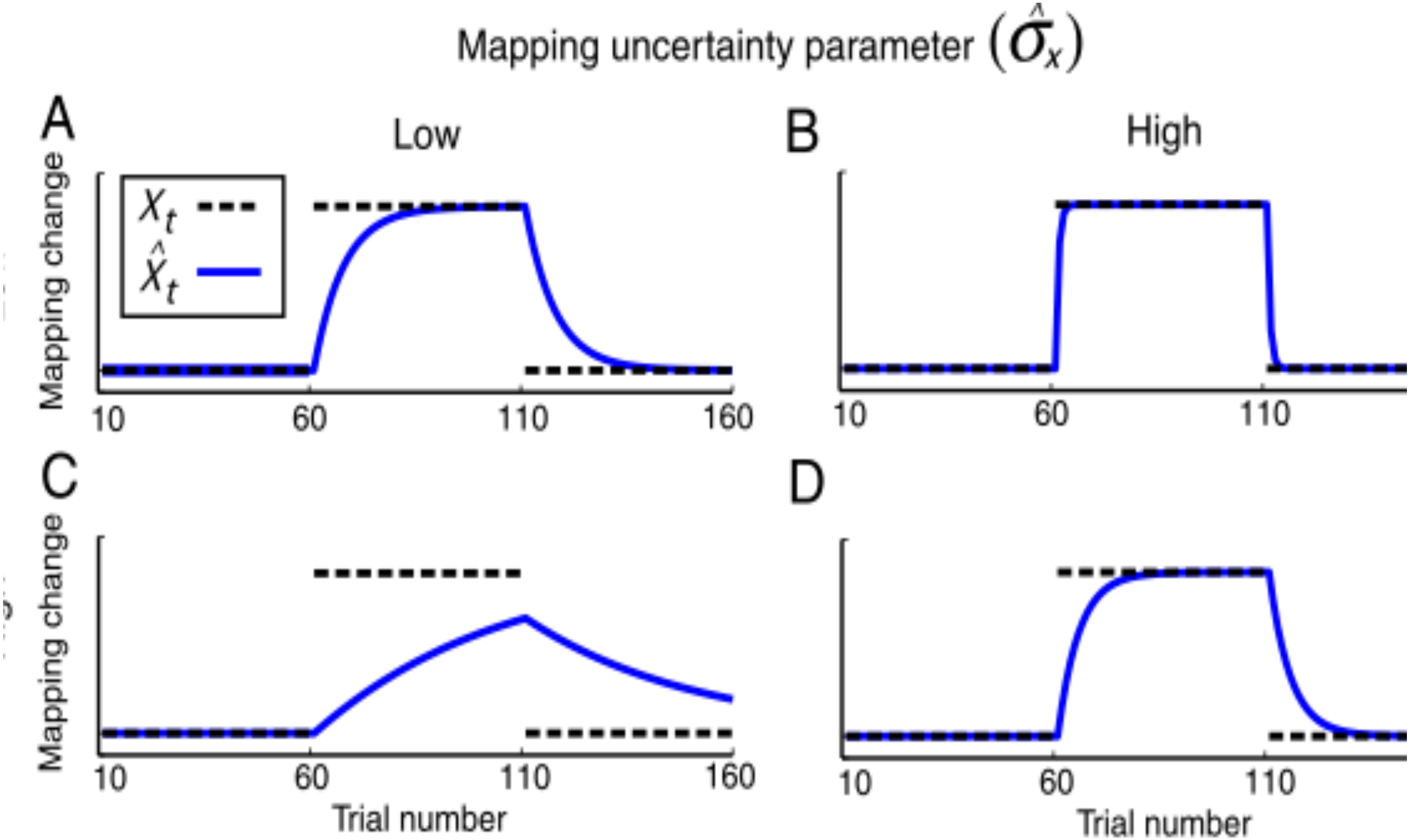
Temporal dependence in cue weighting



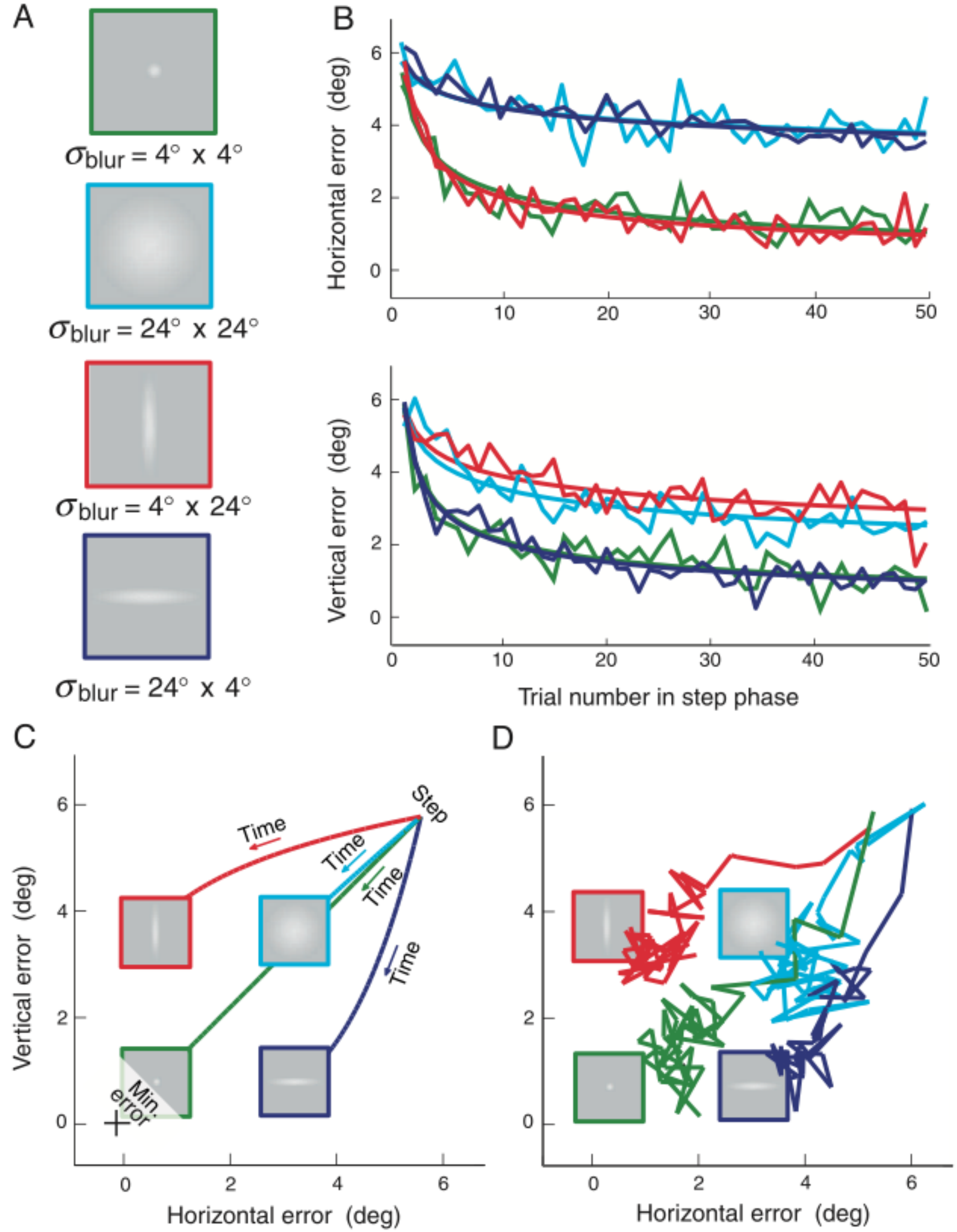
Position uncertainty and blur



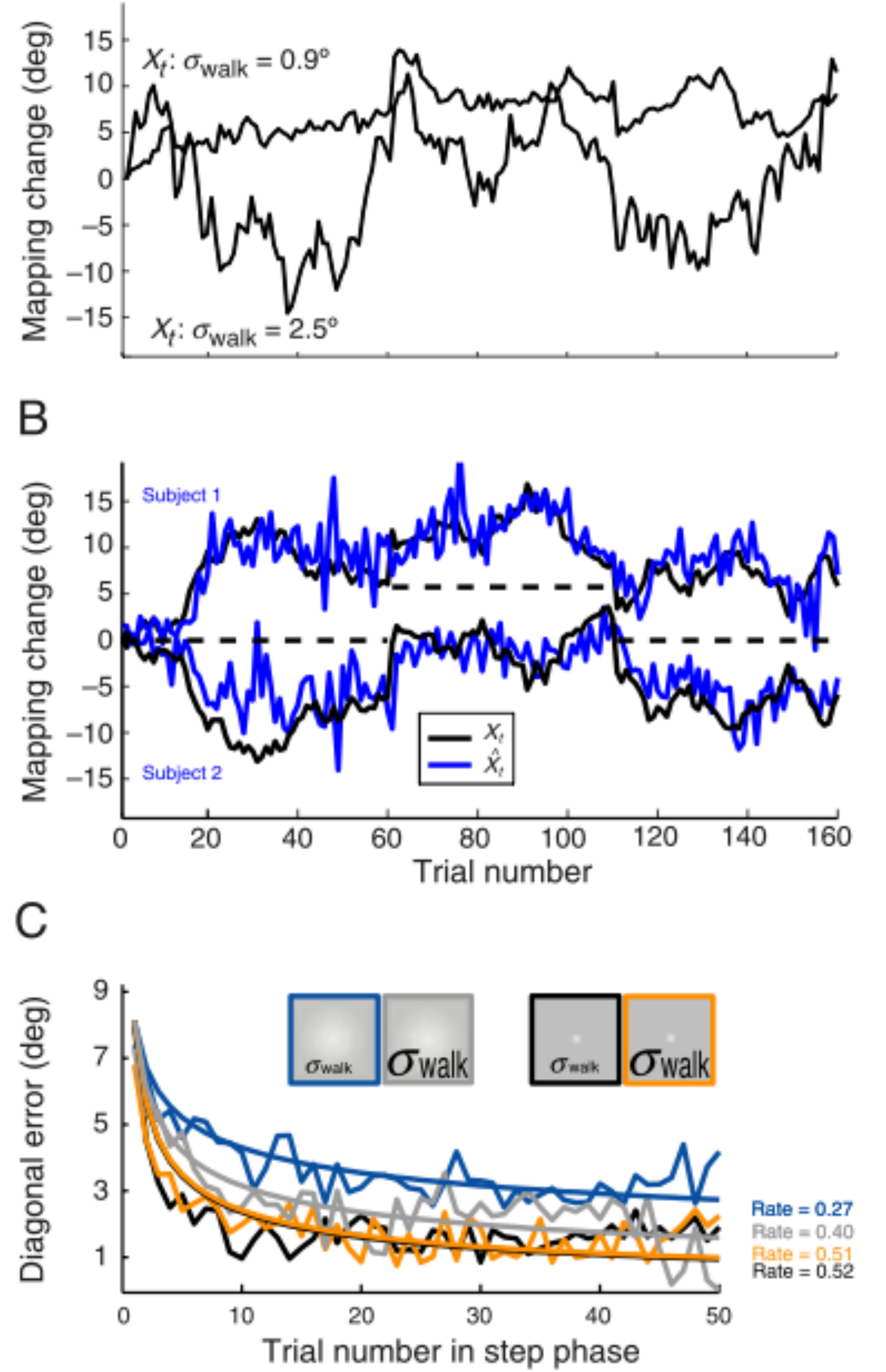
Predictions



Directional Blur

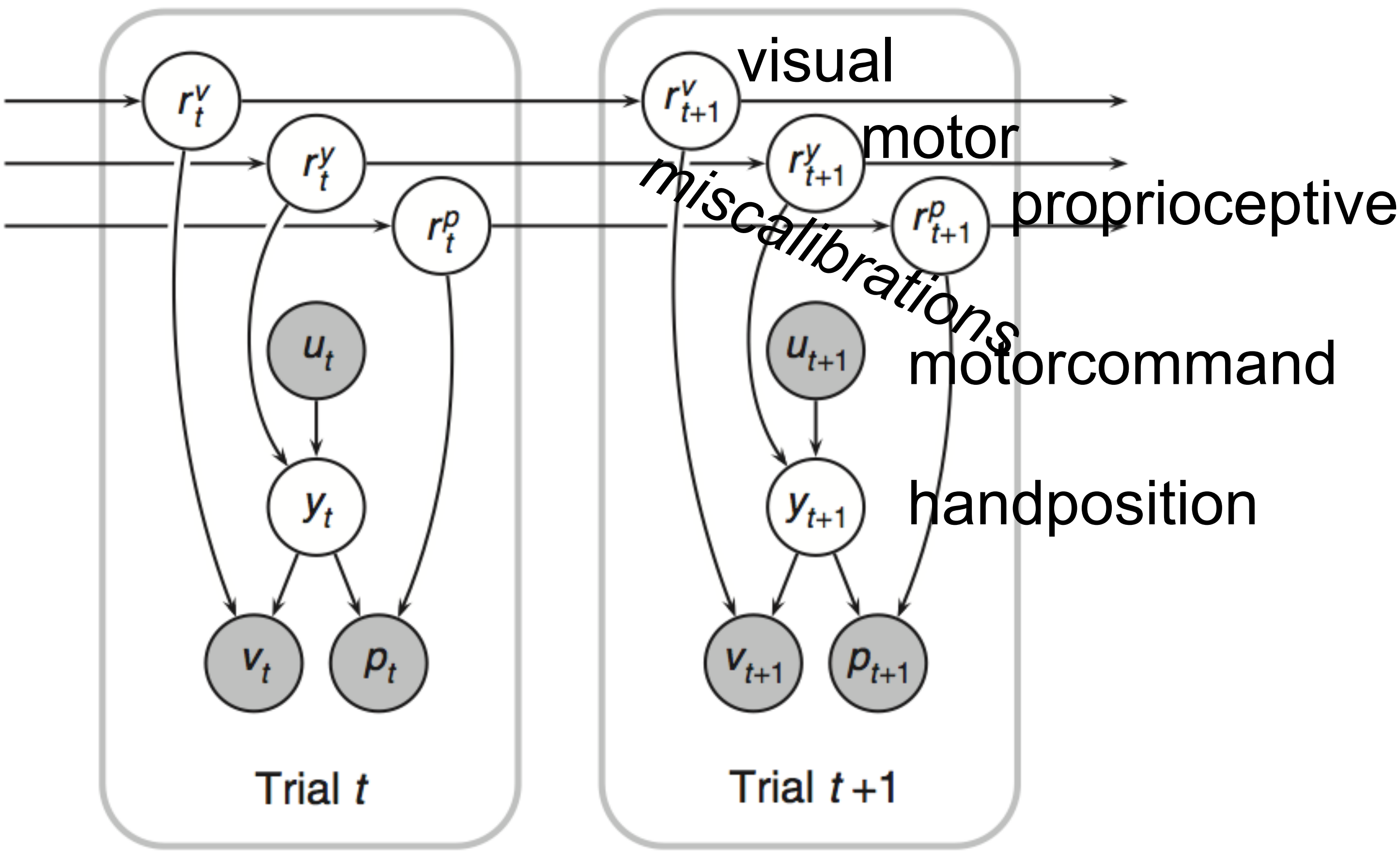


Random walk increases adaptation rate



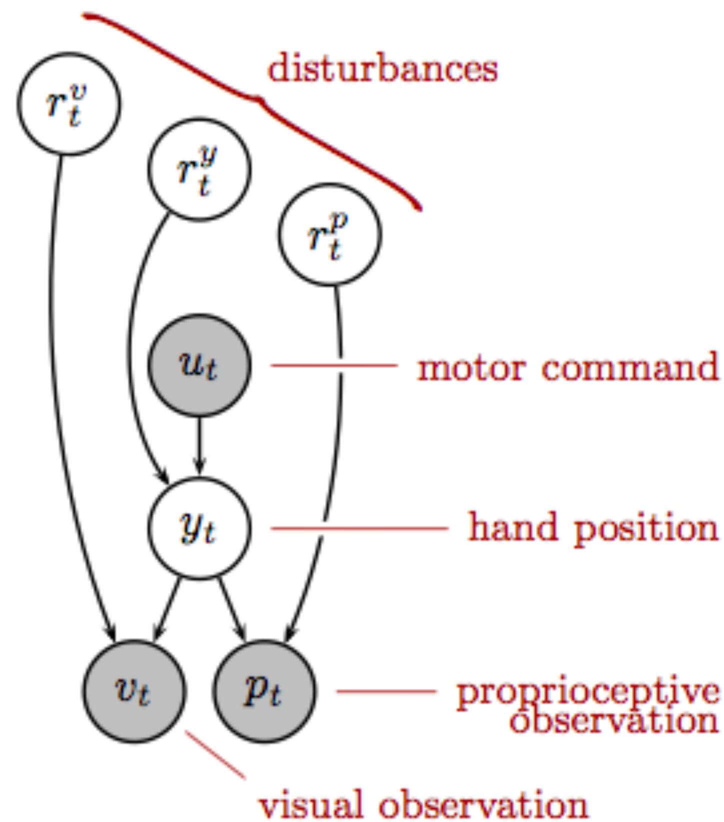
Bayesian sensory- and motor-adaptation model.

Shaded circles represent observed random variables
Unshaded circles represent unobserved random variables



Rewrite as Kalman

$$v_t = y_t + r_t^v + \varepsilon_t^v$$
$$p_t = y_t + r_t^p + \varepsilon_t^p$$
$$y_t = u_t + r_t^y + \varepsilon_t^y$$



Problem: This mixes observable and unobserved variables

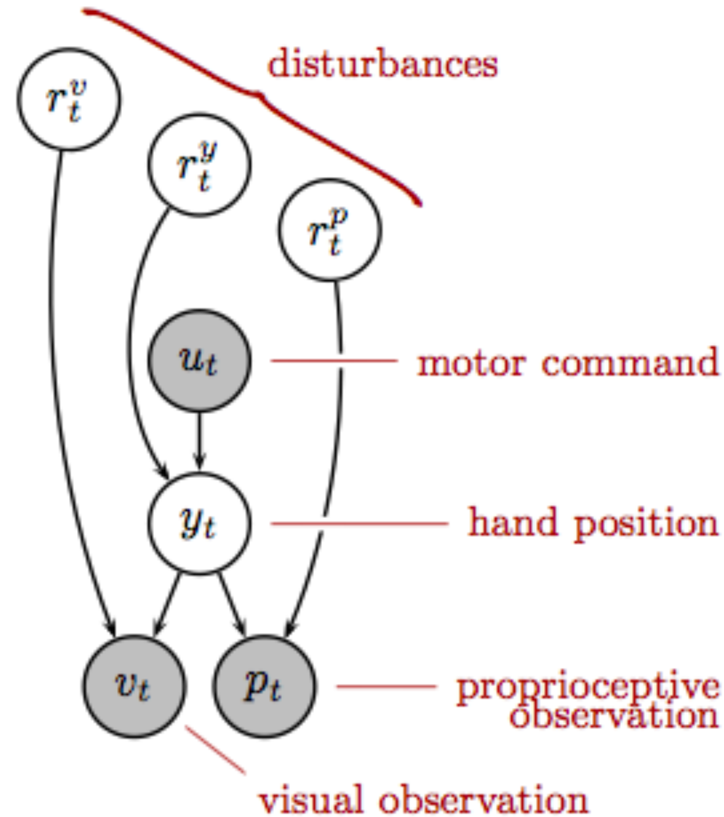
Rewrite as Kalman

Because Linear and Gaussian, we can rewrite:

$$v_t = y_t + r_t^v + \varepsilon_t^v$$

$$p_t = y_t + r_t^p + \varepsilon_t^p$$

$$u_t = y_t - r_t^y - \varepsilon_t^y$$

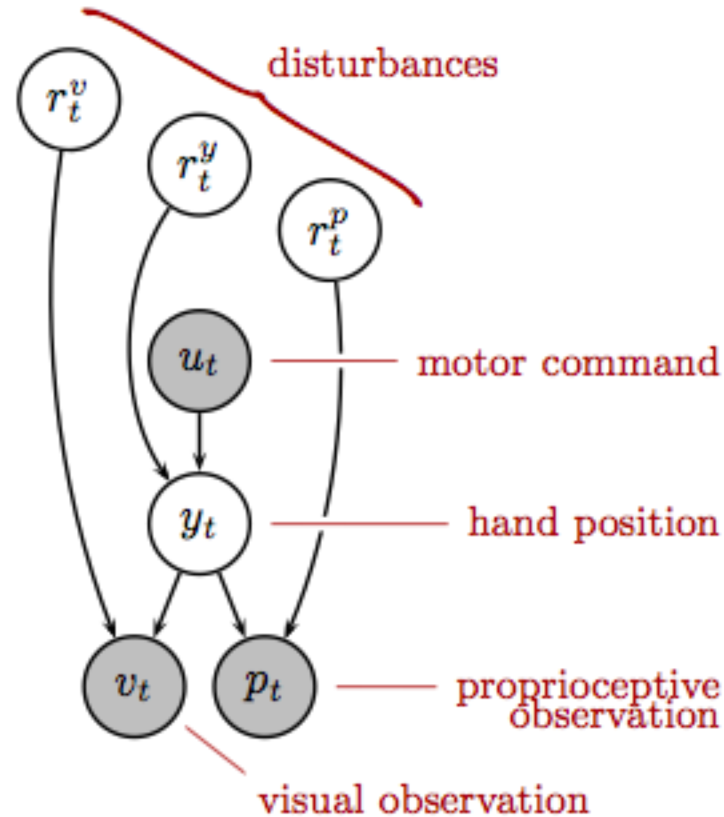


Rewrite as Kalman

$$v_t = y_t + r_t^v + \varepsilon_t^v$$

$$p_t = y_t + r_t^p + \varepsilon_t^p$$

$$u_t = y_t - r_t^y - \varepsilon_t^y$$

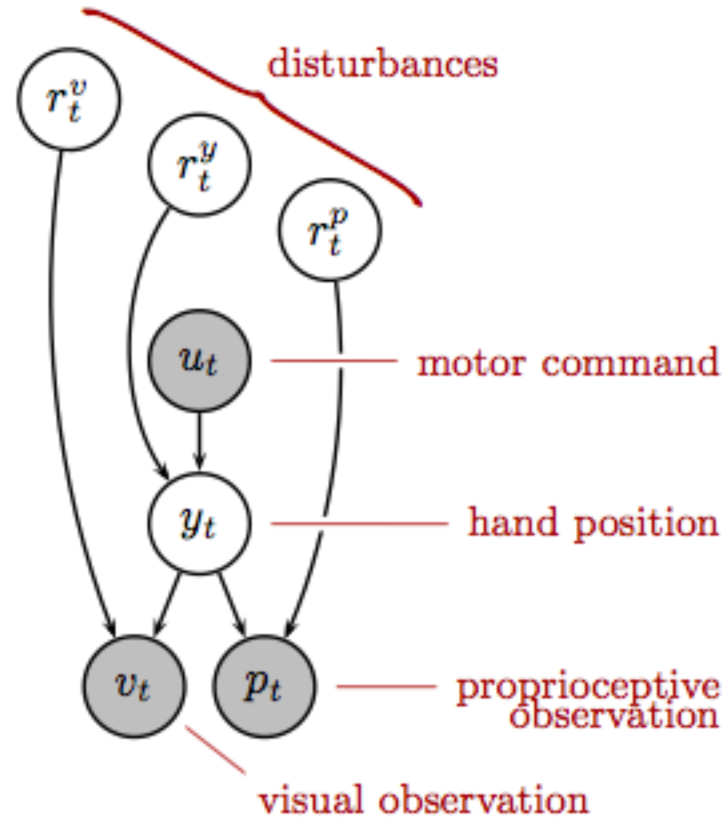


$$\begin{bmatrix} v_t \\ p_t \\ u_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} r_t^v \\ r_t^p \\ r_t^y \\ y_t \end{bmatrix} + \begin{bmatrix} \varepsilon_t^v \\ \varepsilon_t^p \\ -\varepsilon_t^y \end{bmatrix}$$

Rewrite as Kalman

$$\begin{bmatrix} v_t \\ p_t \\ u_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} r_t^v \\ r_t^p \\ r_t^y \\ y_t \end{bmatrix} + \begin{bmatrix} \varepsilon_t^v \\ \varepsilon_t^p \\ -\varepsilon_t^y \end{bmatrix}$$

THEY DIDN'T DO THIS, BUT COULD HAVE



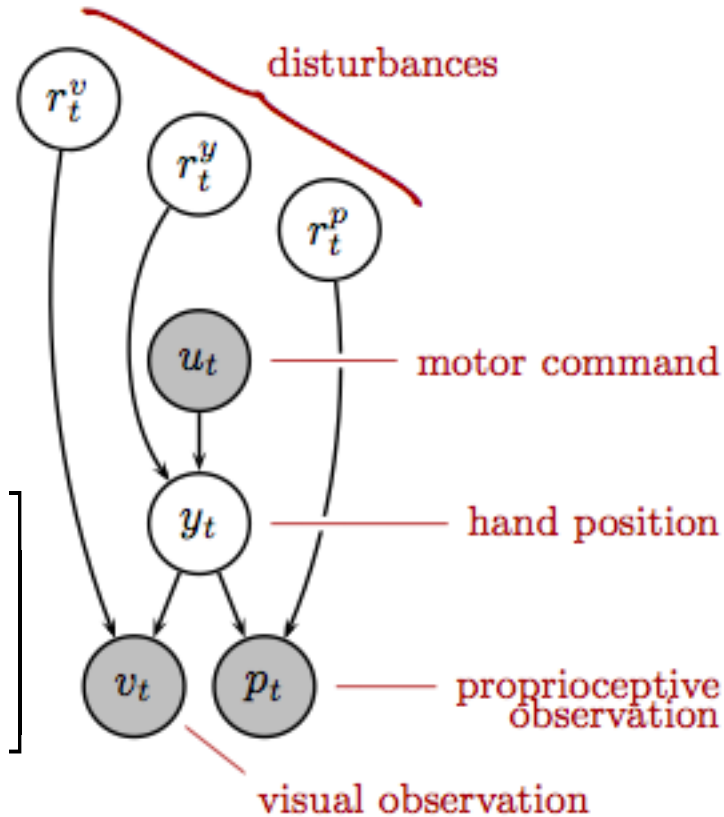
Rewrite as Kalman

$$y_t = u_t + r_t^y + \varepsilon_t^y$$

$$v_t = (u_t + r_t^y + \varepsilon_t^y) + r_t^v + \varepsilon_t^v$$
$$p_t = \left(u_t + r_t^y + \varepsilon_t^y\right) + r_t^p + \varepsilon_t^p$$

$$\begin{bmatrix} v_t - u_t \\ p_t - u_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} r_t^v \\ r_t^p \\ r_t^y \end{bmatrix} + \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_t^v \\ \varepsilon_t^p \\ \varepsilon_t^y \end{bmatrix}$$

$$\mathbf{z}_t = H\mathbf{r}_t + H\varepsilon_t$$



Simple Kalman Filter

Dynamics Model

$$\mathbf{r}_{t+1} = A\mathbf{r}_t + \eta_t$$

$$\mathbf{z}_t = H\mathbf{r}_t + H\varepsilon_t$$

$$A = \begin{bmatrix} a^v & 0 & 0 \\ 0 & a^p & 0 \\ 0 & 0 & a^y \end{bmatrix}$$

$$H = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

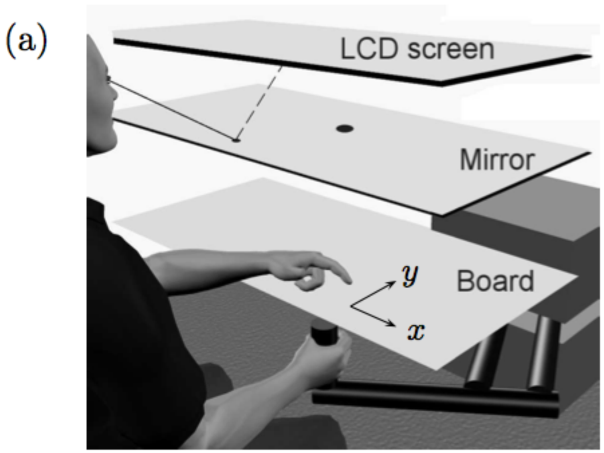
$$\eta_t \sim N(0,Q)$$

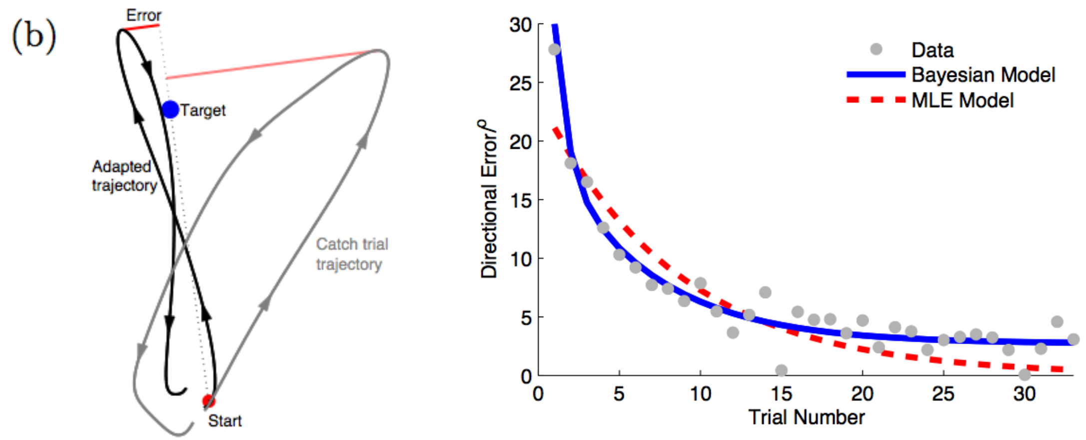
$$\eta_t \sim N(0,R)$$

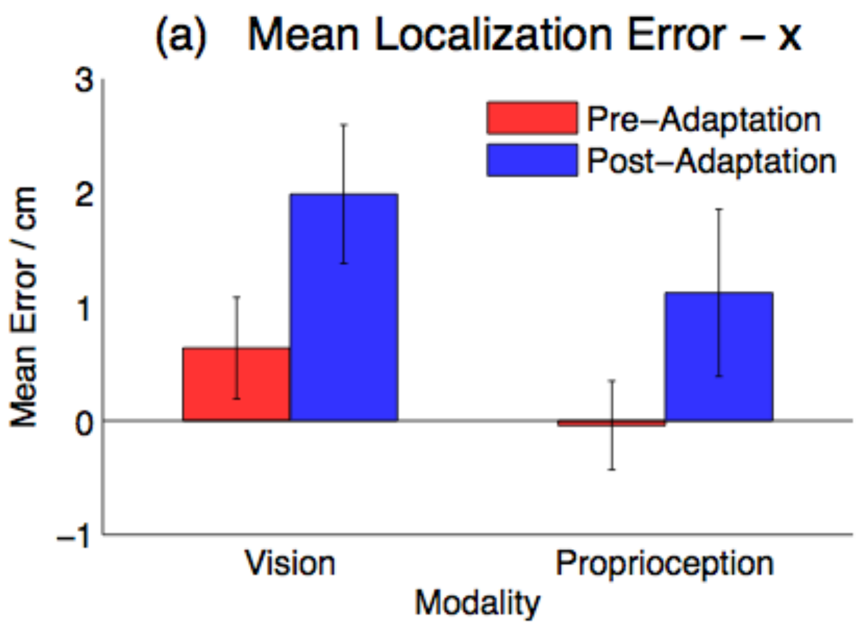
$$Q = \begin{bmatrix} q^v & 0 & 0 \\ 0 & q^p & 0 \\ 0 & 0 & q^y \end{bmatrix}$$

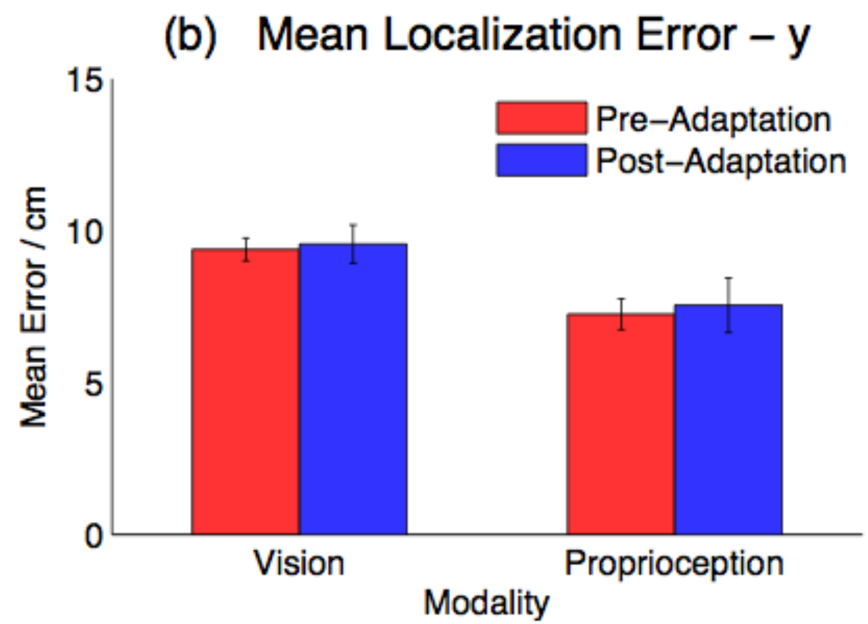
$$R = E\left[(H\boldsymbol{\epsilon}_t)(H\boldsymbol{\epsilon}_t)^T\right] = \begin{pmatrix} \sigma_v^2 + \sigma_u^2 & \sigma_u^2 \\ \sigma_u^2 & \sigma_p^2 + \sigma_u^2 \end{pmatrix}$$

Experimental results



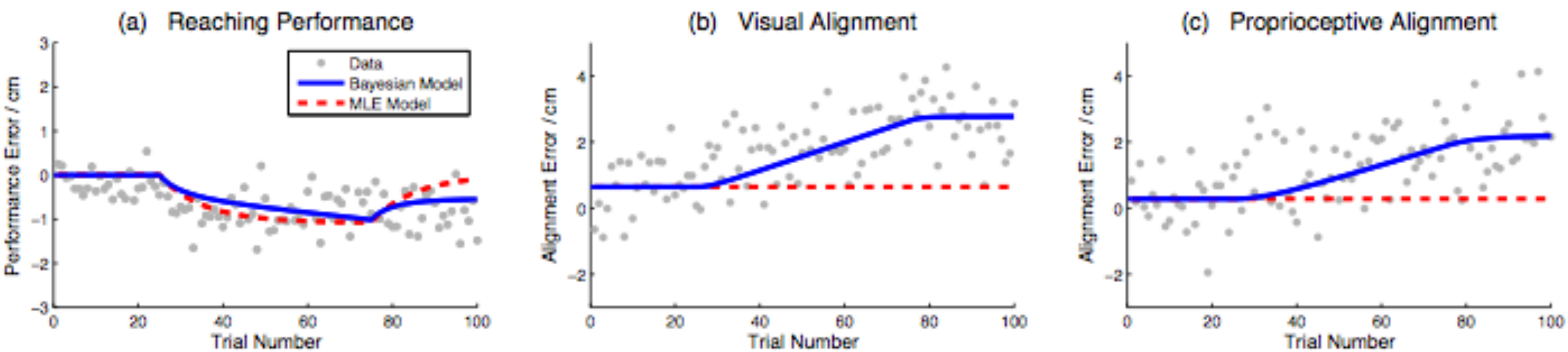






Results contd

Three tasks: Reach to target (right hand),
left hand to visual
left hand to right hand's location



Summing Up so far

- Bayesian models provide a principled language to describe uncertainty, information fusion under uncertainty, and make non-trivially verified predictions about perceptual processing.

- The brain needs to represent priors and likelihoods –
- Not always the case we are Bayesian….