What to do next?

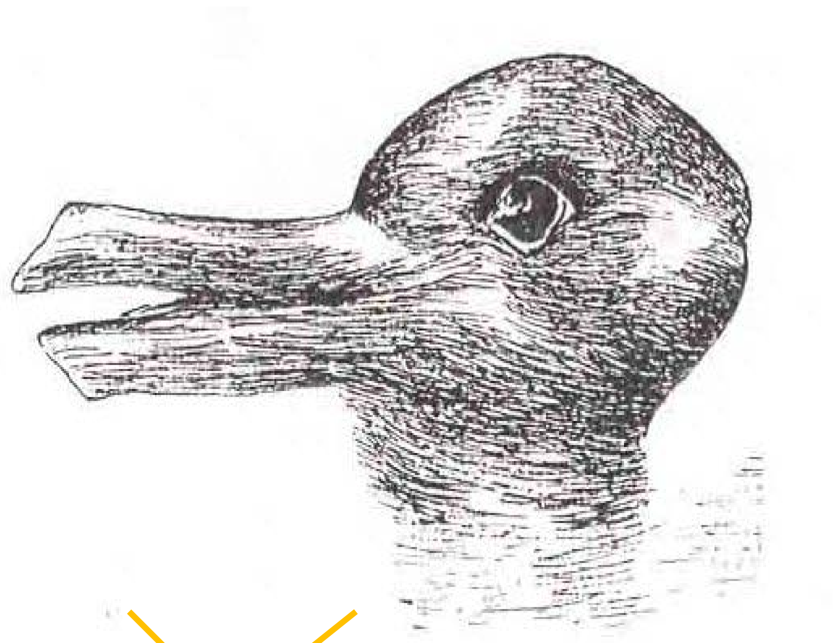
- on top down task driven feature saliency

Lars Kai Hansen,
Seliz Kardogan &
Letizia Marchegiani

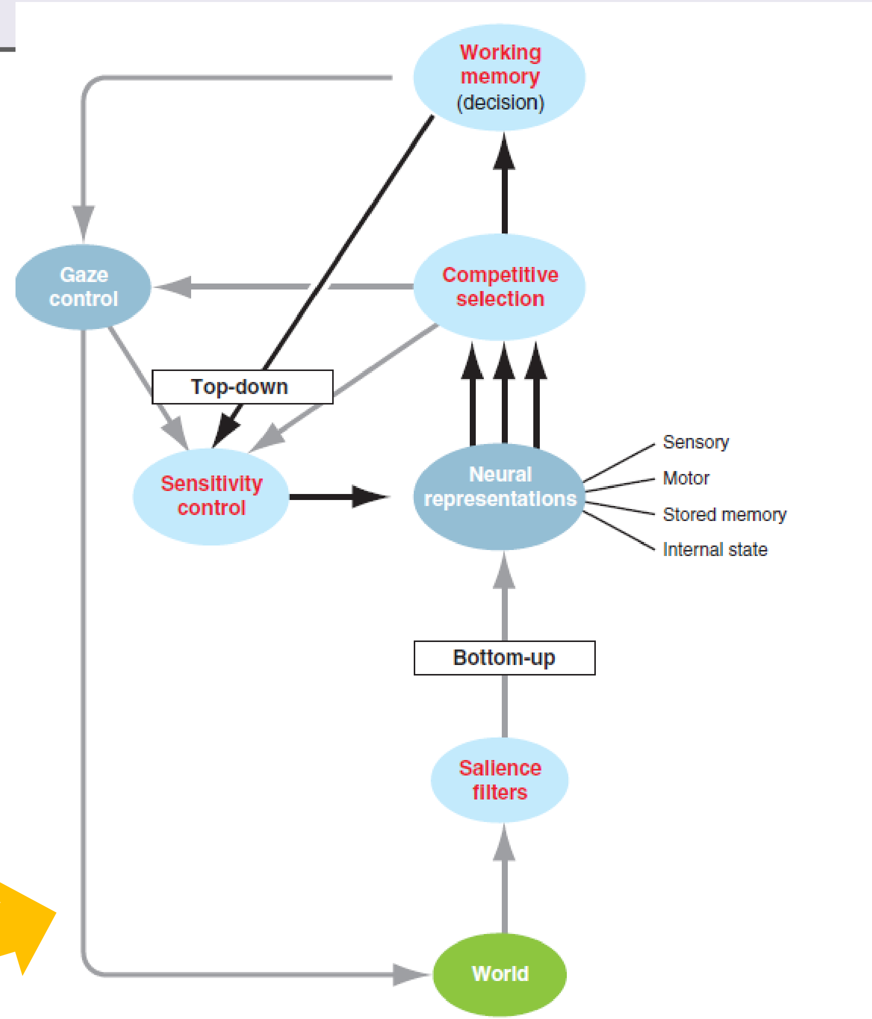
Technical University of Denmark
DTU Informatics
Lyngby, Denmark

Center for integrated molecular brain imaging

Motivating example



~~BIRD~~ RABBIT



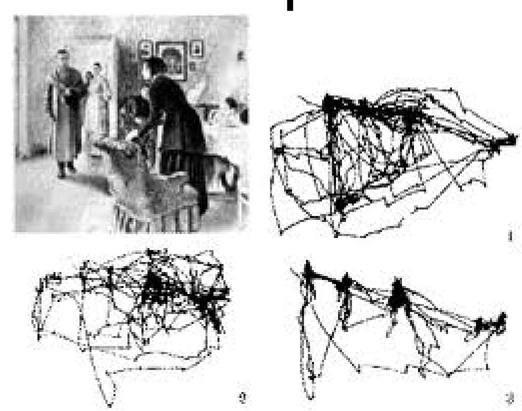
E.I. Knudsen: Fundamental Components of Attention, Annu. Rev. Neurosci. 2007. 30:57–78.

Lars Kai Hansen
IMM, Technical University of Denmark

Why pursue top down attention mechanisms?

- Predictive modeling of human behavior

- Inverse reconstruct human intention /sphere of interest from actions



V. Navalpakkam, L. Itti, Modeling the influence of task on attention, *Vision Research,* Vol. **45**, No. 2, pp. 205-231, Jan 2005.

Top down vs bottom up attention

Bottom up

Attention determined by feature of the input

Audio

Cocktail party effect
(Cherry, 64)

Visual

Classical spatial
novelty saliency
(Itti+Koch, 04)

Top down

Attention determined by state of the observer

Audio

Cocktail party problem (Cherry, 64)

Visual

ambiguous pictures

eye tracking

See e.g. J.M. Wolfe et al. ”How fast can you change your mind? The speed of top-down guidance in visual search” Vision Research 44 (2004) 1411–1426

A model of top down attention

1. Task is implemented as decision problem

2. Attention is represented as choice of feature

Standard probabilistic classifier

Model of posterior probability

$$p(c|x,z)$$

Two sets of features

i) Features setting the context **'the gist (x)**

(Friedman,79; Torralba et al., 04)

ii) **Potential features (z)** considered by the attention mechanism

Friedman A. Framing pictures: the role of knowledge in automatized encoding and memory of gist. Journal of Experimental Psychology: General 1979;108:316–355.

Mathematical model

The **GIST**

We are interested in a partial observation **x** under a decision task: Choose among ”C” actions

$$
\begin{aligned}
p(c|\boldsymbol{x}) &= \int p(c,\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \\
&= \frac{\int p(c,\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z}}{\sum_{c=1}^{C}\int p(c,\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z}}
\end{aligned}
$$

versus getting additional information though z_j

$$
\begin{aligned}
p(c|\boldsymbol{x},z_j) &= \sum_{c=1}^{C}\int p(c,\boldsymbol{z}|\boldsymbol{x})\prod_{i\neq j} dz_i \\
&= \frac{\int p(c,\boldsymbol{x},\boldsymbol{z})\prod_{i\neq j} dz_i}{\sum_{c=1}^{C}\int p(c,\boldsymbol{x},\boldsymbol{z})\prod_{i\neq j} dz_i}
\end{aligned}
$$

Measure the information gain

First used by Lindley (1956) for experimental design..

ON A MEASURE OF THE INFORMATION PROVIDED BY AN EXPERIMENT[1, 2]

By D. V. Lindley

University of Cambridge and University of Chicago

1. Summary. A measure is introduced of the information provided by an experiment. The measure is derived from the work of Shannon [10] and involves the knowledge prior to performing the experiment, expressed through a prior probability distribution over the parameter space. The measure is used to compare some pairs of experiments without reference to prior distributions; this method of comparison is contrasted with the methods discussed by Black-[illegible]. Finally, the measure is applied to provide a solution to some problems of experimental design, where the object of experimentation is not to reach decisions but rather to gain knowledge about the world.

D. V. Lindley, "On a measure of the information provided by an experiment," Annals Mathematical Statistics, vol. 4, pp. 986–1005, 1956.

Information theoretical model

$$
\begin{aligned}
\Delta S_j(\boldsymbol{x}, z_j) &= - \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\
&\quad + \sum_{c=1}^{C} \log p(c|\boldsymbol{x}, z_j) p(c|\boldsymbol{x}, z_j)
\end{aligned}
$$

$$
\begin{aligned}
G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\
&= \sum_{c=1}^{C} \int \log p(c|\boldsymbol{x}, z_j) p(c, z_j|\boldsymbol{x}) dz_j \\
&\quad - \sum_{c=1}^{C} \int \log \cancel{p(c, \boldsymbol{z}|\boldsymbol{x})} p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}.
\end{aligned}
$$

Lars Kai Hansen
IMM, Technical University of Denmark

Gaussian-Discrete distribution

... allows closed form marginalization and conditionals

$$p(c, \boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, \boldsymbol{z}|k)$$

Component probabilities

Class emission probabilities

Joints Gaussian component
density of of ”gist” and features

Information gain by requesting of j'th feature

1D integrals over normal distribution pdf

$$
\begin{aligned}
G_j(\boldsymbol{x}) &= \sum_{c=1}^{C}\sum_{k=1}^{K} p(c|k)p(k|\boldsymbol{x}) \times \\
&\qquad \int \log\left[p(c,\boldsymbol{x},z_j)\right] p(z_j|\boldsymbol{x},k)dz_j \\
&- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log\left[p(\boldsymbol{x},z_j)\right] p(z_j|\boldsymbol{x},k)dz_j \\
&+ \text{const.}
\end{aligned}
$$

Experimental design

Consider a conventional classification problem

Train the Gaussian-discrete joint input (gist/feature) distribution on training data

Test with split gist+feature input: Give the gist (**x**) and determine by the information gain which feature ”j” is best

Compare: train, features, only gist, gist + best, gist + random

Lars Kai Hansen
IMM, Technical University of Denmark

Results:
Pima diabetes problem (UCI)

Features

1*) number of times pregnant

2) plasma glucose concentration a 2 hours in an oral glucose tolerance test

3) diastolic blood pressure

4) triceps skin fold thickness

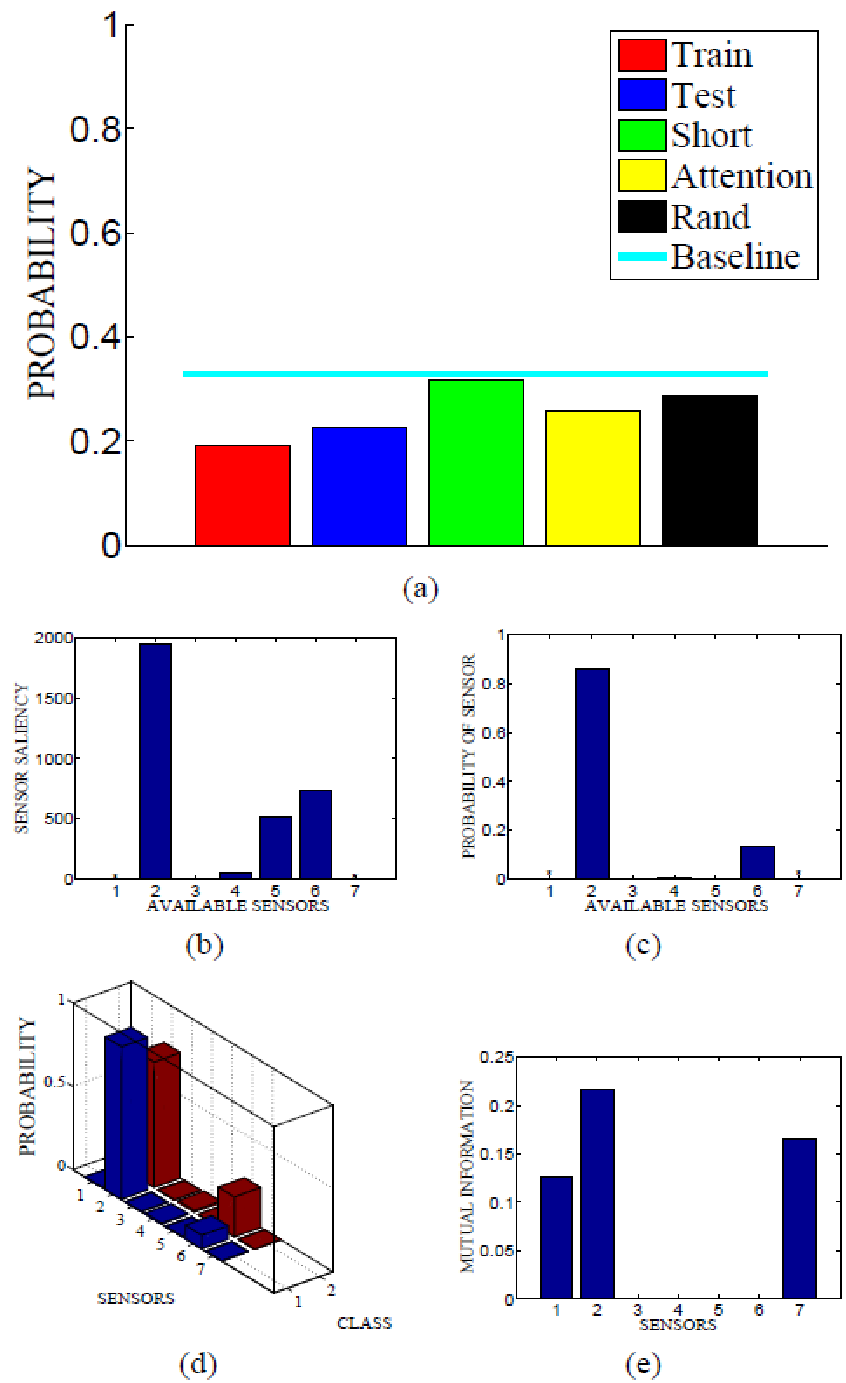
5) body mass index

6) diabetes pedigree function

7*) age (years).

$N_{train} = 200$

$N_{test} = 332$



Results:
Abelone age classification problem

Features:

1*) gender (M, F),

2*) length, longest shell measurement,

3) diameter, perpendicular to length,

4) height, with meat in shell,

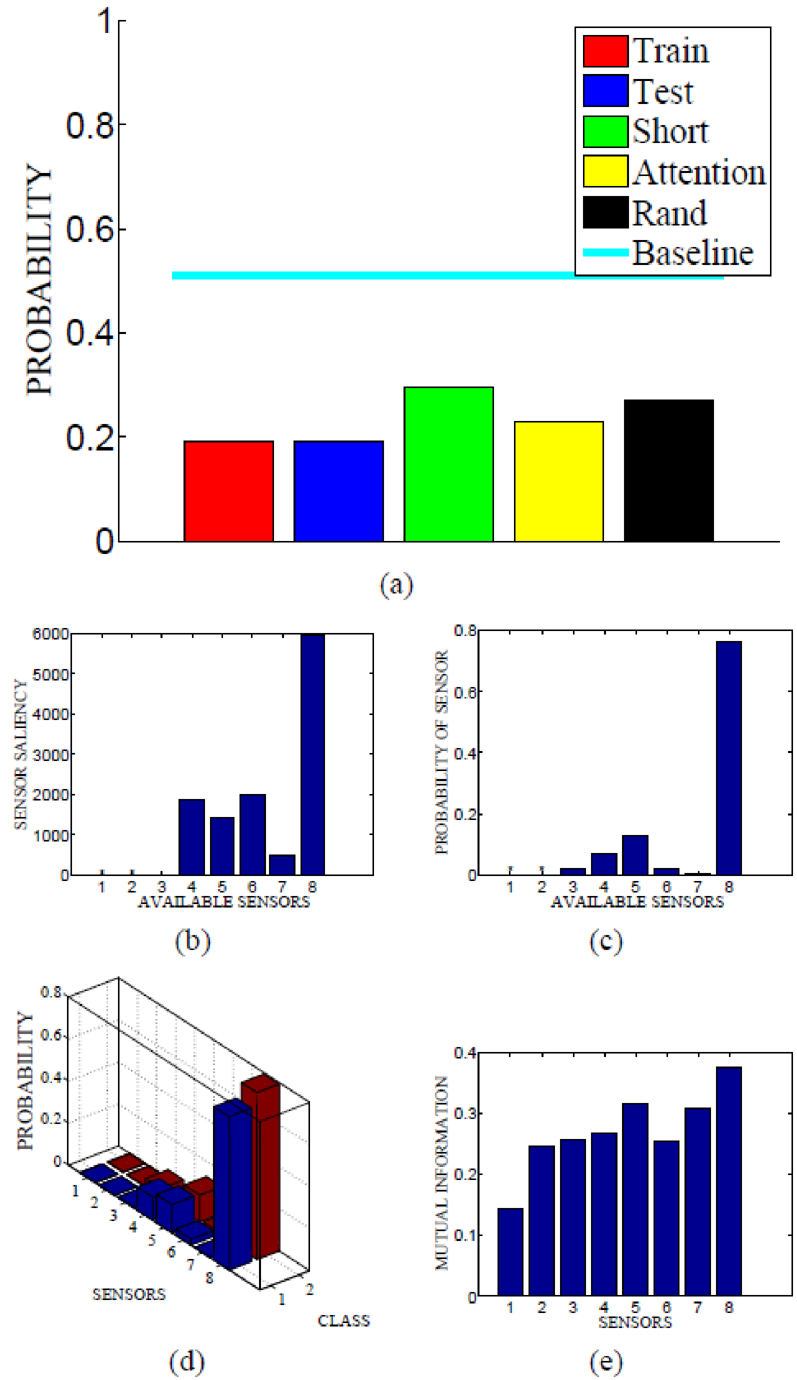
5) whole weight, whole abalone,

6) shucked weight, weight of meat,

7) viscera weight, gut weight (after bleeding),

8) shell weight, after being dried.

$N_{train} = 3500 \quad N_{test} = 677$



Lars Kai Hansen
IMM, Technical University of Denmark

Conclusion

- Attention mechanisms combine properties of the input field and the state/goal/task of the beholder
- A simple information optimizing mechanism can **use task information to determine what to do next and improves decision making**
- Perspectives
  - Engineering Q: Fast evaluation of a proxy for the entropy
  - Scientific Q: Are human observers optimal?
  - Cognitive Systems Q: Top-down attention can be used to infer the state of the beholder

Research supported by Lundbeck Fonden - www.cimbi.dk