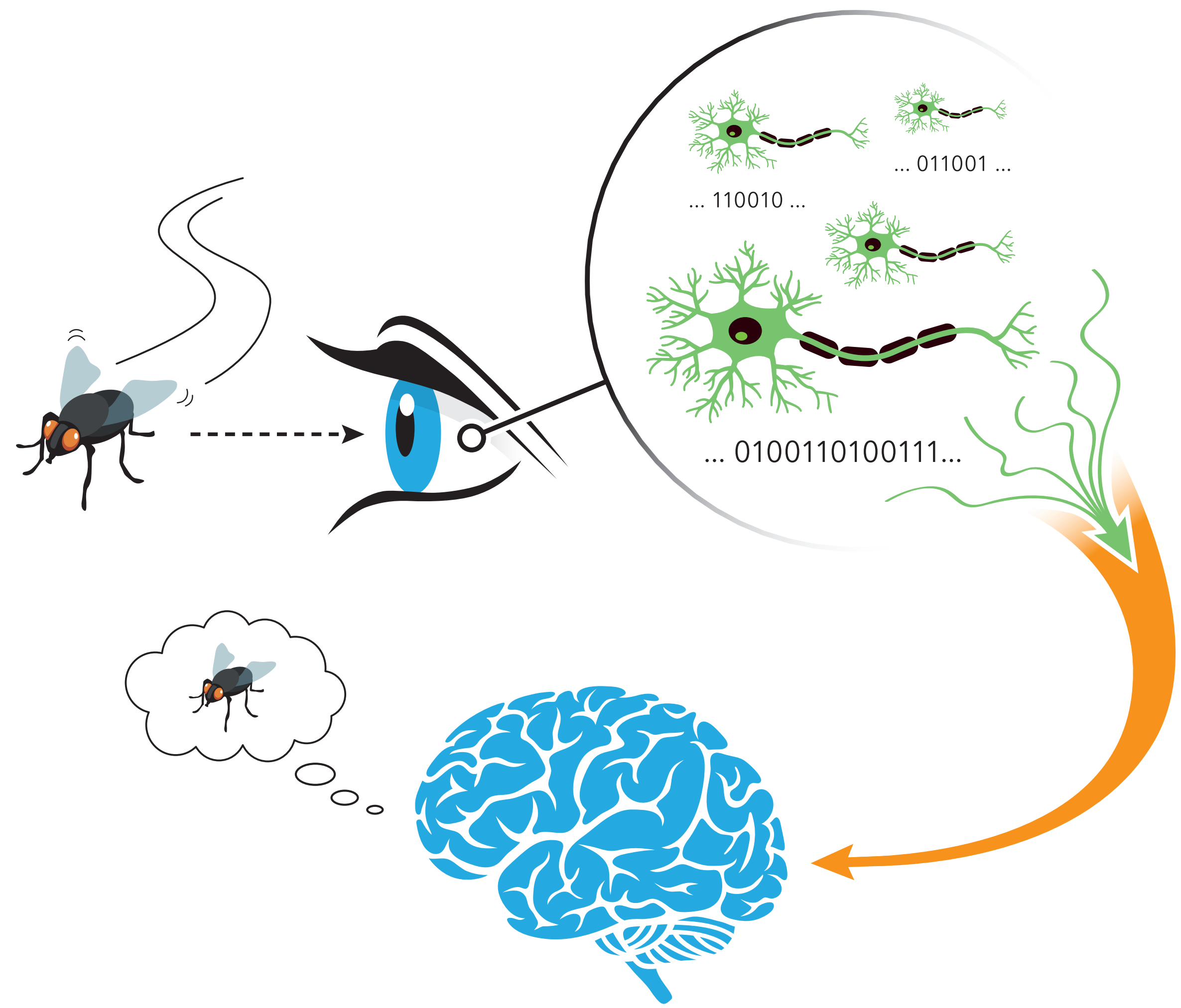
Normative models of visual neural coding

Matthew Chalk

Institut de la Vision
Sorbonne Université
Paris, France

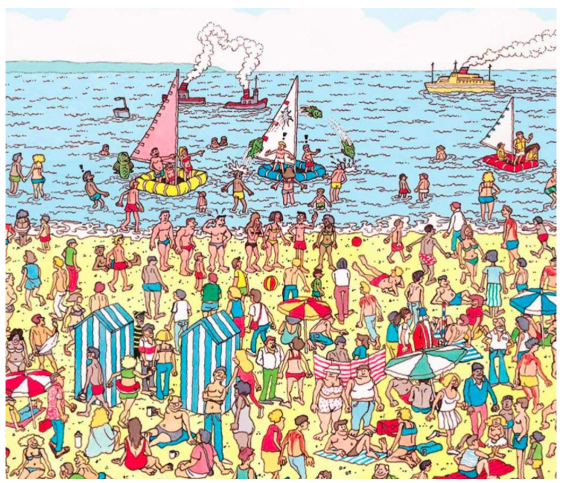


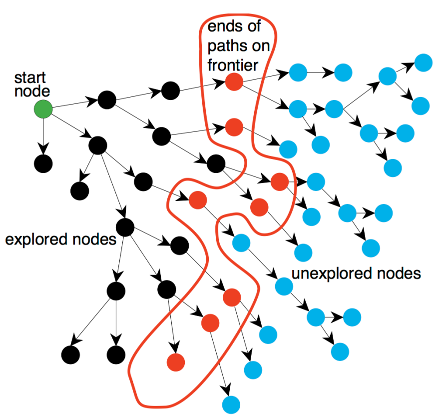
Marre, Vision, 1982

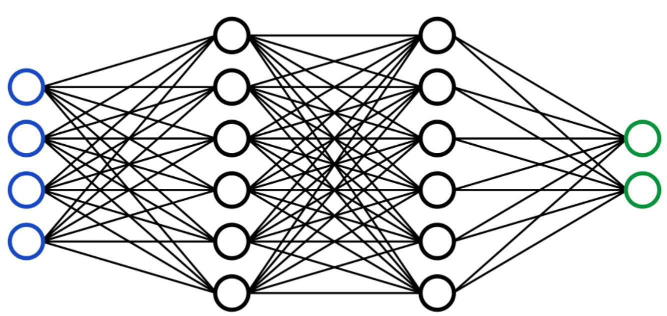
What are ‘normative’ or ‘top-down’ models?

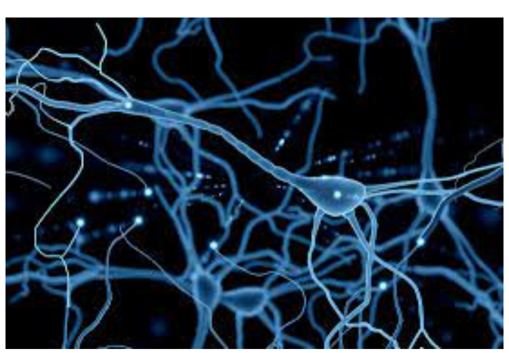
Marre’s 3 levels

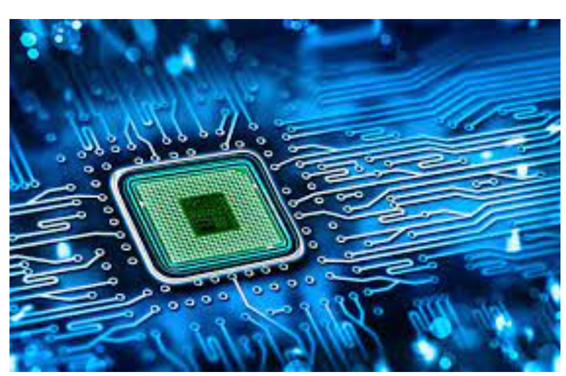
où est Charlie ?











bottom-up

computational goal

algorithms

implementation

top-down

Recap: Efficient coding

Horace Barlow (1961):

Neurons have evolved to encode maximal information about natural stimuli, given limited resources

Mutual information: $I(X;R) = H(R) - \cancel{H(R|X)}$

low noise limit

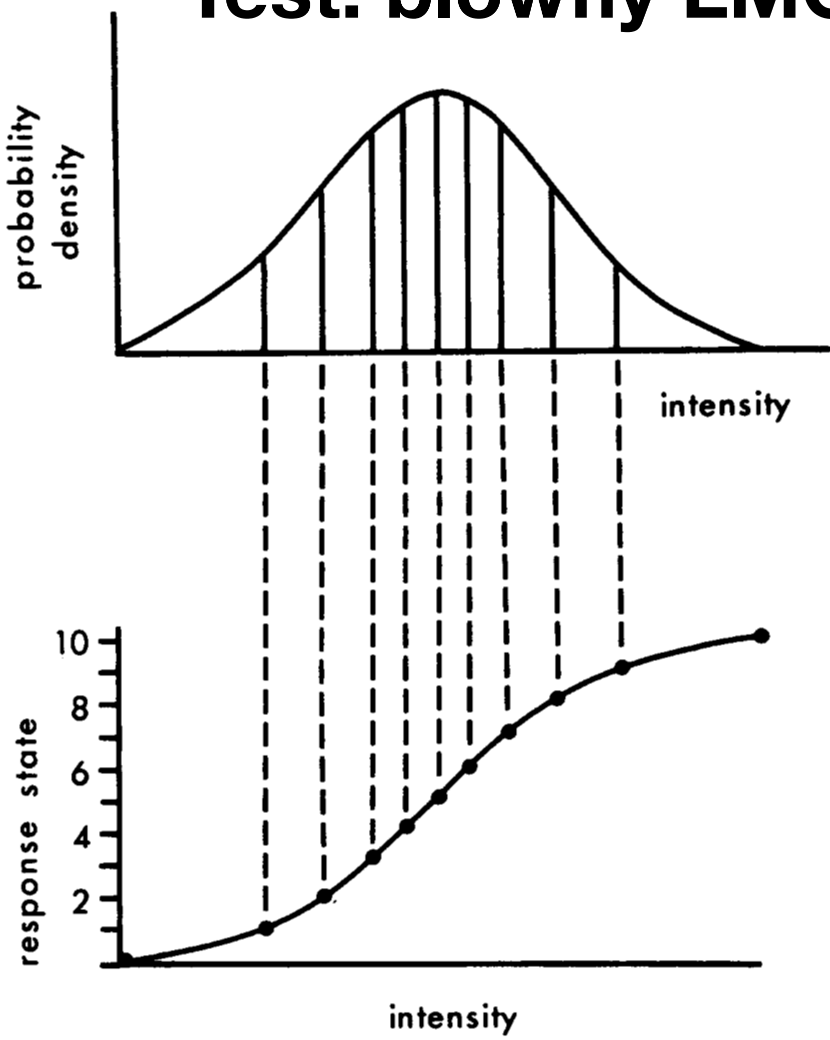
Single neurons, with constrained firing rate: $0 \leq r \leq r_{max}$

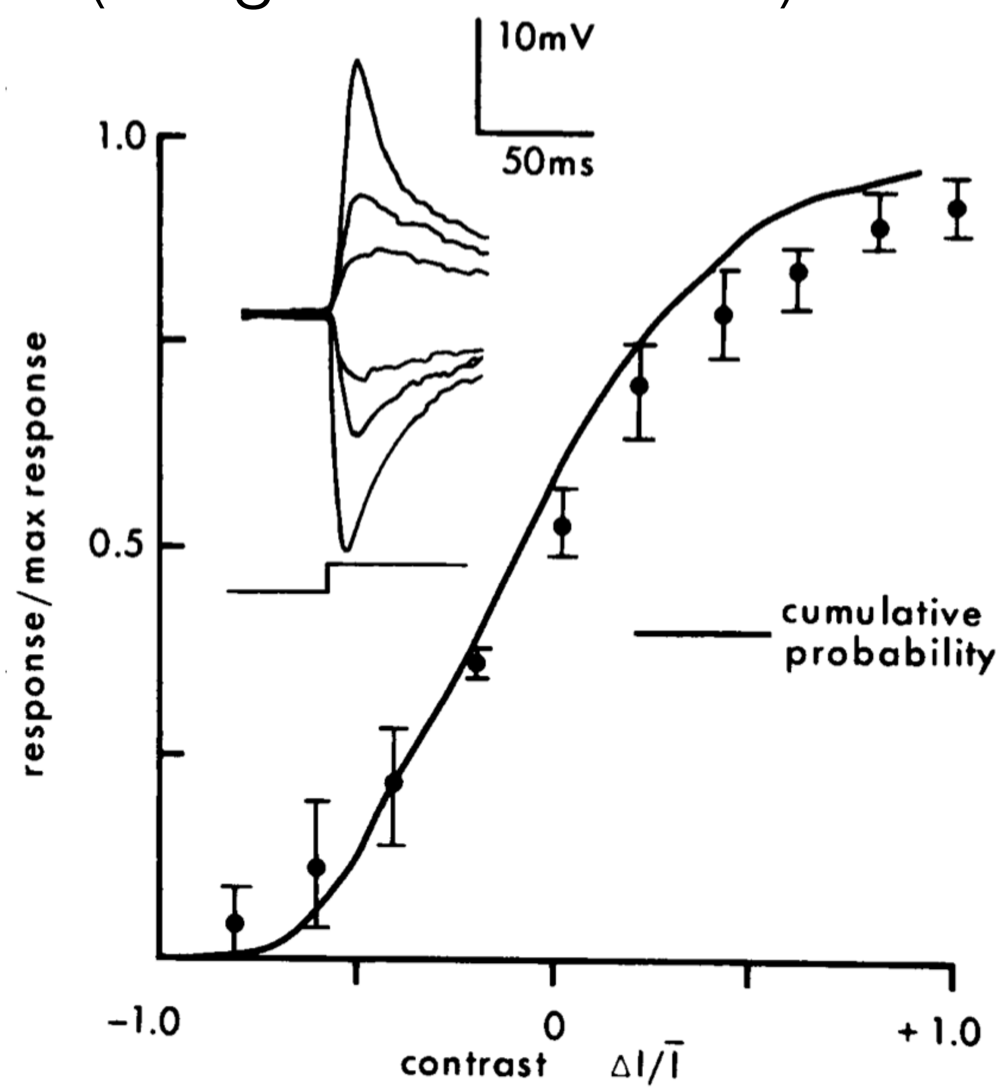
Optimal solution: $r(x) = \int_{x'=x_{\min}}^{x} p(x)dx'$

Recap: Efficient coding

Optimal solution: $$r(x) = \int_{x'=x_{\min}}^{x} p(x)dx'$$

Test: blowfly LMC neuron (Laughlin et al. 1981)





Efficient coding by neural populations

Neurons encode information in a population

Low-noise limit: $$H = -\int p(\boldsymbol{r}) \log p(\boldsymbol{r}) d\boldsymbol{r}$$

Entropy of one neuron: $$H_i = -\int p(\boldsymbol{r}) \log p(r_i)$$

Difference between full entropy, and sum of individual entropies:

$$\sum_i H_i - H = \int p(\boldsymbol{r}) \log \left(\frac{p(\boldsymbol{r})}{\prod_i p(r_i)} \right)$$

$$= KL\left(p(\boldsymbol{r}) \middle| \prod p(r_i) \right) \geq 0$$

Efficient coding by neural populations

Redundancy: $$\sum_i H_i - H \geq 0$$

To maximise inf (at low noise):

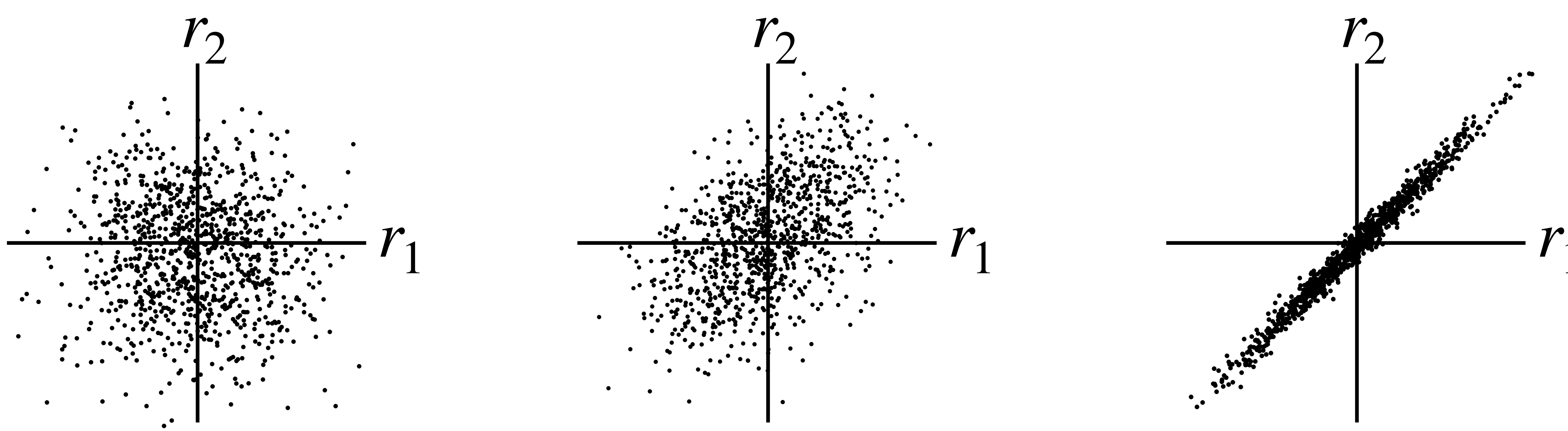
1. maximise H_i
2. minimise redundancy: $p(\boldsymbol{r}) = \prod_i p(r_i)$

Full independence hard to achieve. Consider 2nd order statistics only for now:

$$C_{ij} = \left\langle r_i r_j \right\rangle - \left\langle r_i \right\rangle \left\langle r_j \right\rangle = \sigma^2 I$$

Efficient coding by neural populations

$$C_{ij} = \left\langle r_i r_j \right\rangle - \langle r_i \rangle \left\langle r_j \right\rangle = \sigma^2 I \qquad H_{gauss} = \frac{1}{2}\log(2\pi e\,|\,\Sigma\,|\,)$$



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$$

Can you compute the entropy of each distribution?

Application to the retina

Assumption: receptive fields in the retina are linear & translationally invariant

Response

$$L_s(\boldsymbol{a}) = \int D(\boldsymbol{x} - \boldsymbol{a}) s(\boldsymbol{x}) d\boldsymbol{x}$$

linear response

input filter

stimulus

Correlation

$$C(\boldsymbol{a}, \boldsymbol{b}) = \int d\boldsymbol{x} d\boldsymbol{y} D(\boldsymbol{x} - \boldsymbol{a}) D(\boldsymbol{y} - \boldsymbol{b}) C_{ss}(\boldsymbol{x} - \boldsymbol{y}) = \sigma^2 \delta(\boldsymbol{a} - \boldsymbol{b})$$

independent

Application to the retina

$$C(\boldsymbol{a},\boldsymbol{b}) = \int d\boldsymbol{x}d\boldsymbol{y}D(\boldsymbol{x}-\boldsymbol{a})D(\boldsymbol{y}-\boldsymbol{b})C_{ss}(\boldsymbol{x}-\boldsymbol{y}) = \sigma^2\delta(\boldsymbol{a}-\boldsymbol{b})$$

In Fourier space, convolution -> multiplication

$$|\tilde{D}(\boldsymbol{\omega})|^2\tilde{Q}_{ss}(\boldsymbol{\omega}) = \sigma^2$$

Therefore: $$|\tilde{D}(\boldsymbol{\omega})| = \frac{\sigma}{\sqrt{\tilde{Q}_{ss}(\boldsymbol{\omega})}}$$

Measured Q: $$Q_{ss}(\boldsymbol{\omega}) \propto \frac{e^{-\alpha|k|}}{|k|^2 + k_0^2}$$

$\longleftarrow$ optics of eye

$\longleftarrow$ natural scenes

So $|D(k)| \to \infty$ as $|k| \to \infty$ …? What is missing?

TD

- Apply a whitening filter, W, to image patches, x
- Compute the correlation, $\left\langle xx^T \right\rangle - \langle x \rangle \langle x \rangle^T$
- Convolve the image with the filter: $w \star x$
- What happens when you add noise to the image?

Application to the retina

Need to consider noise! $I(X;R) = H(R) - H(R\,|\,X)$

Input includes signal and noise: $s(\boldsymbol{x}) + \eta(\boldsymbol{x})$

Assume filter takes form: $\tilde{D}(\omega) = \tilde{D}_s(\omega)\tilde{D}_\eta(\omega)$

whitening filter

remove noise

Minimise mean squared error: $$\int d\omega\,|\,\tilde{D}_\eta(\tilde{s}(\omega) + \tilde{\eta}(\omega)) - \tilde{s}(\omega)\,|^2$$

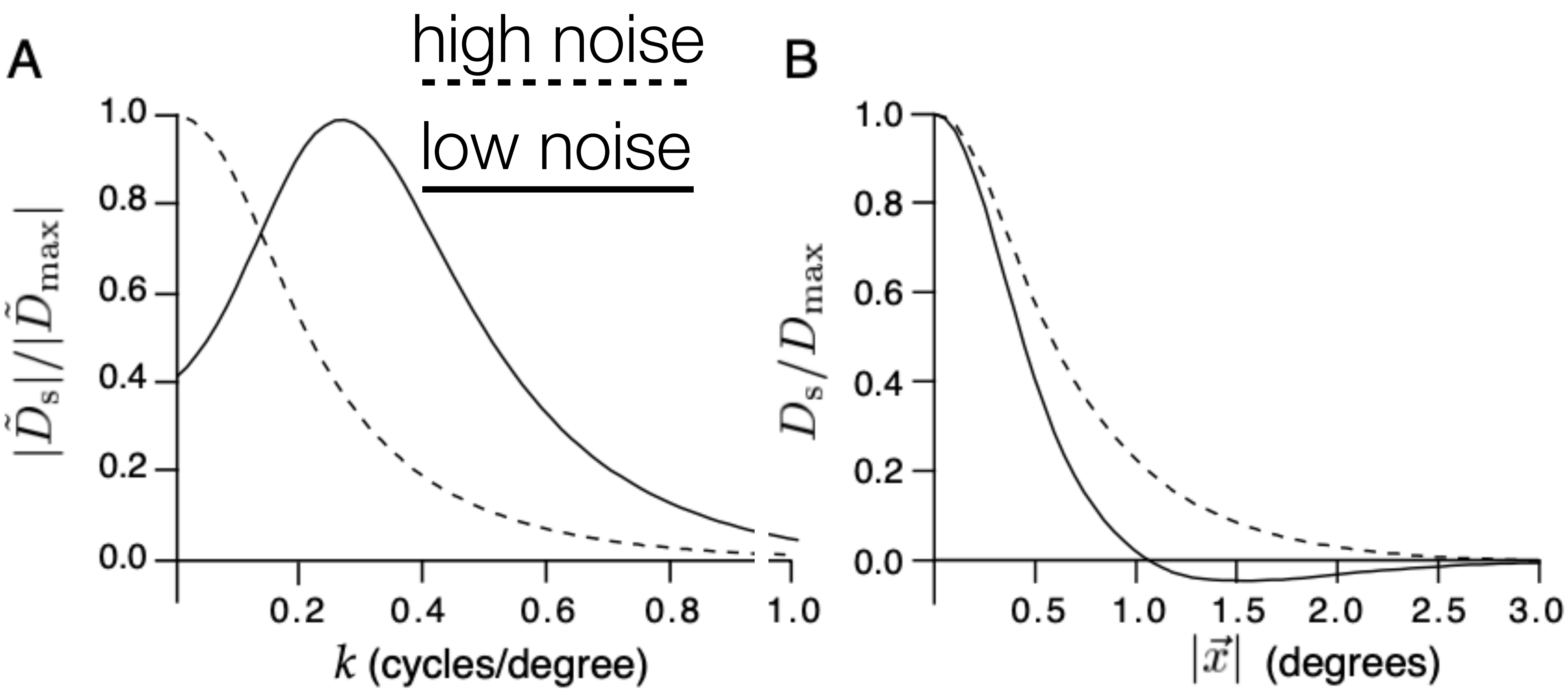
Solution: $$\tilde{D}(\omega) \propto \frac{\sqrt{Q_{ss}(\omega)}}{Q_{ss}(\omega) + Q_{\eta\eta}(\omega)}$$

Atick & Redlich 1992

Application to the retina

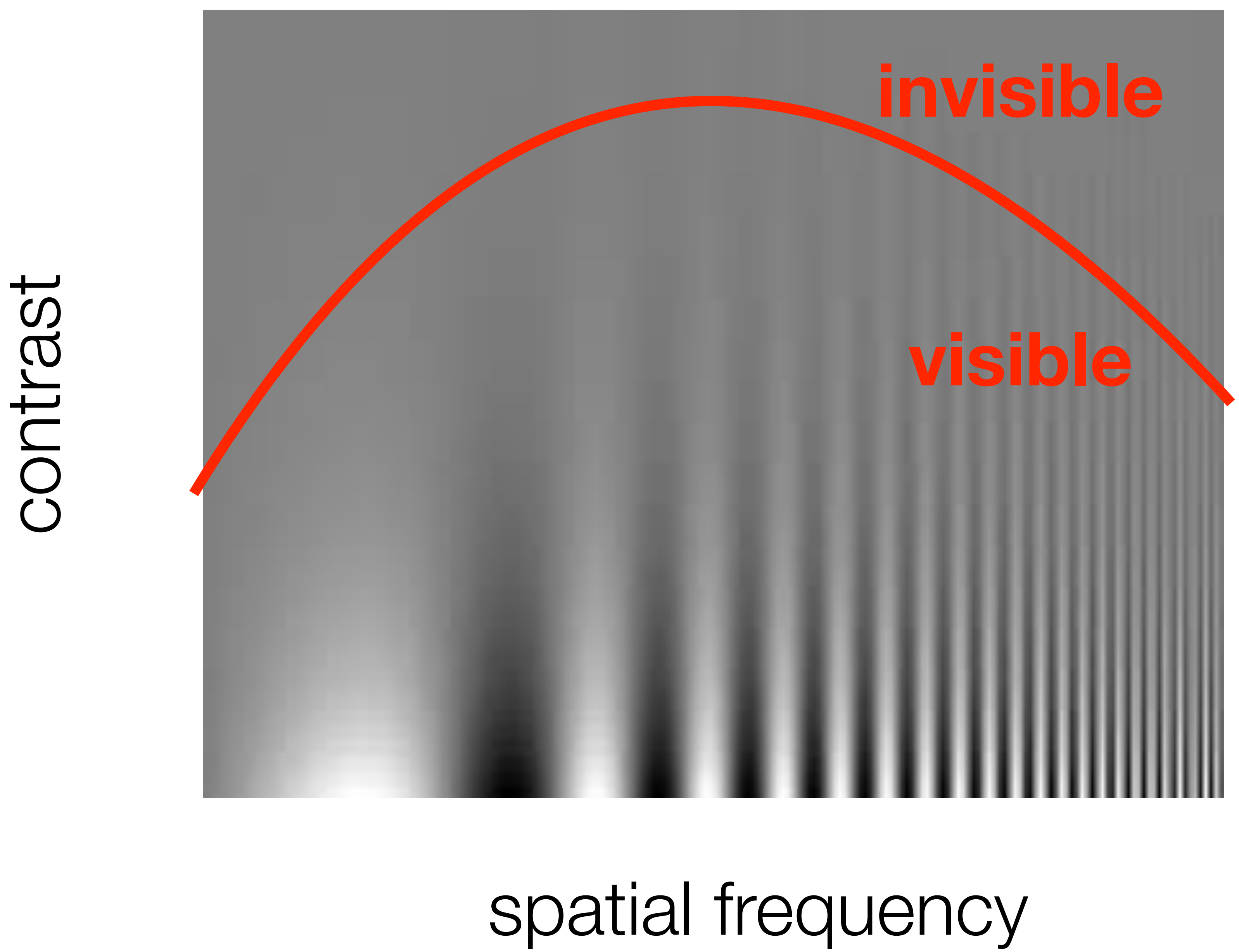
Solution:

$$\tilde{D}(\omega) \propto \frac{\sqrt{Q_{ss}(\omega)}}{Q_{ss}(\omega) + Q_{\eta\eta}(\omega)}$$



Atick & Redlich 1992

Perception

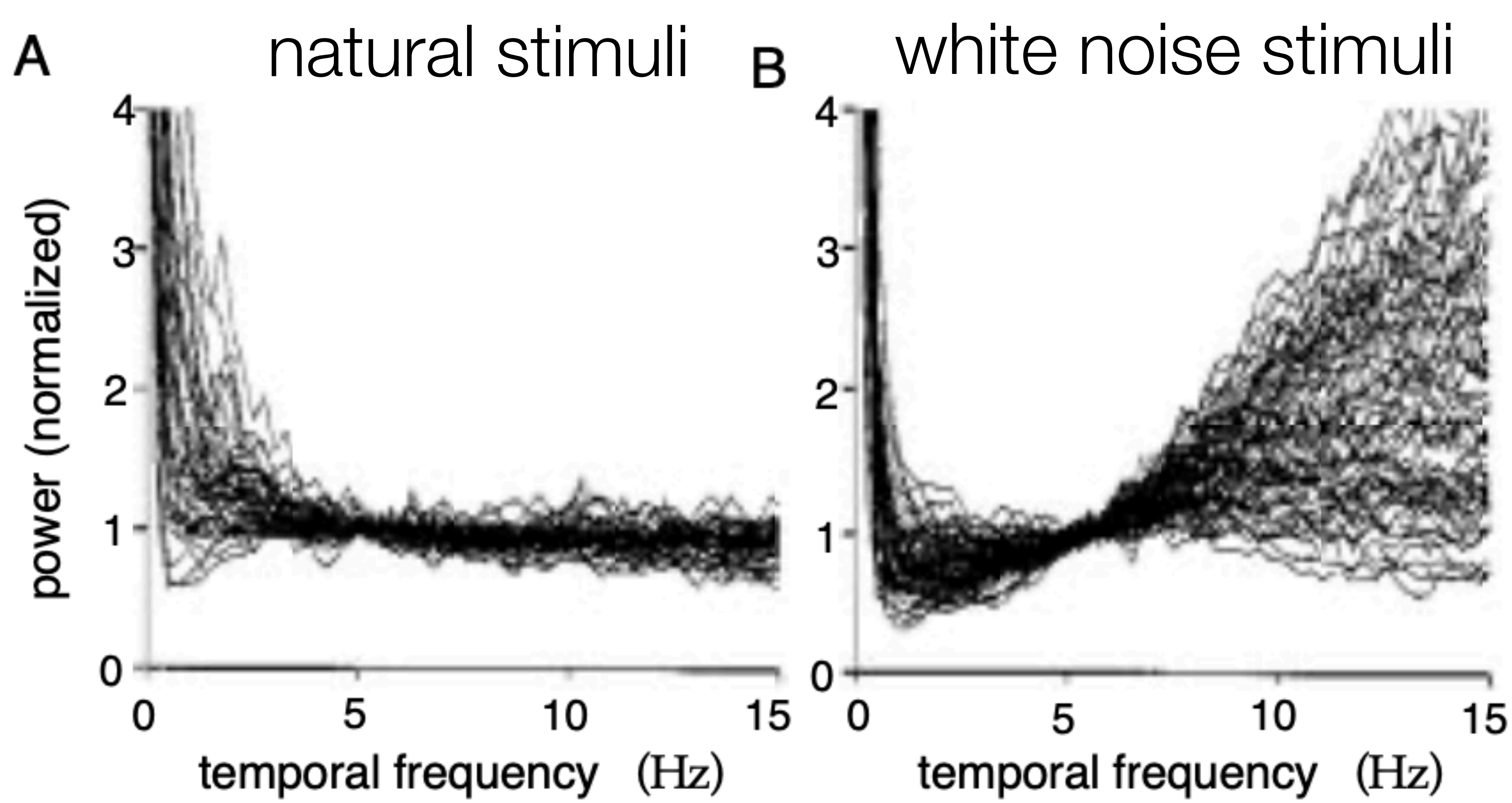


Dan, Atick & Reid 1996

Temporal processing in the LGN

Hypothesis: Neurons amplify high-frequencies to whiten natural signals

Prediction: Stimuli that are already ‘white’ should have high frequencies amplified

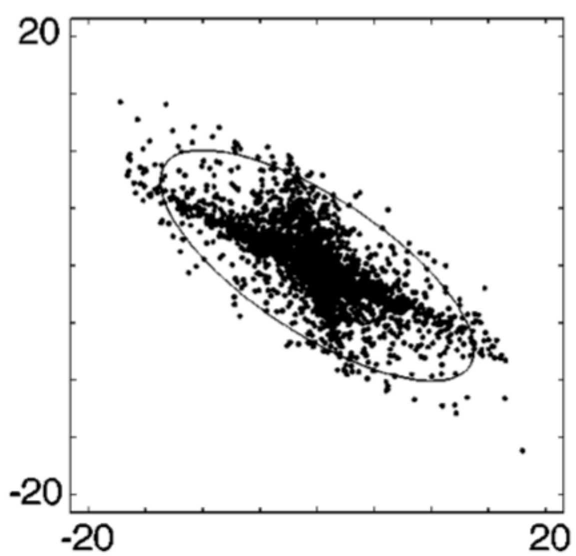


TD

- Can you derive an optimal linear denoising filter?
  - Input: $x_n = x + \eta$
  - Minimise mean squared error: $||W_\eta(x + \eta) - x||^2$
  - Solution: $W_n = \left\langle (x + \eta)(x + \eta)^T \right\rangle^{-1} \left\langle (x + \eta)x^T \right\rangle$
  - (To compute solution, what is $\left\langle xx^T \right\rangle$ and $\left\langle \eta x^T \right\rangle$ and $\left\langle \eta\eta^T \right\rangle$?)
- What does the combined filter, $W = W_{decorr}W_n$ look like?
- What does the resulting neural image $w \star x$ look like?
  - How does this change when we add noise to the image?

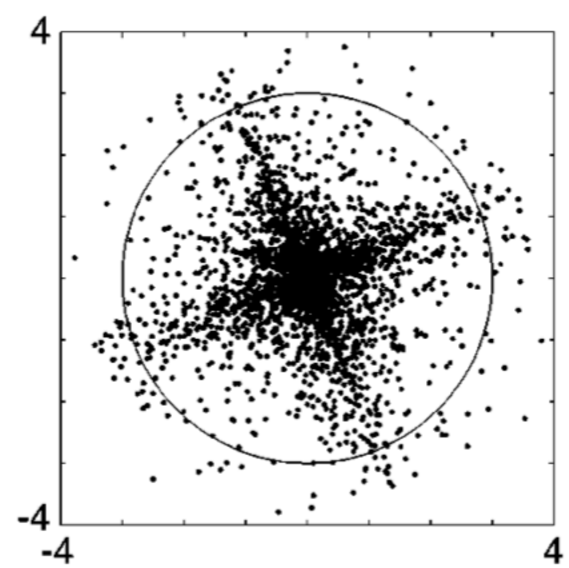
Beyond second order statistics

$$r_i r_j \neq \sigma^2 \delta_{ij}$$
$$p(r_1, r_2) \neq p(r_1)p(r_2)$$



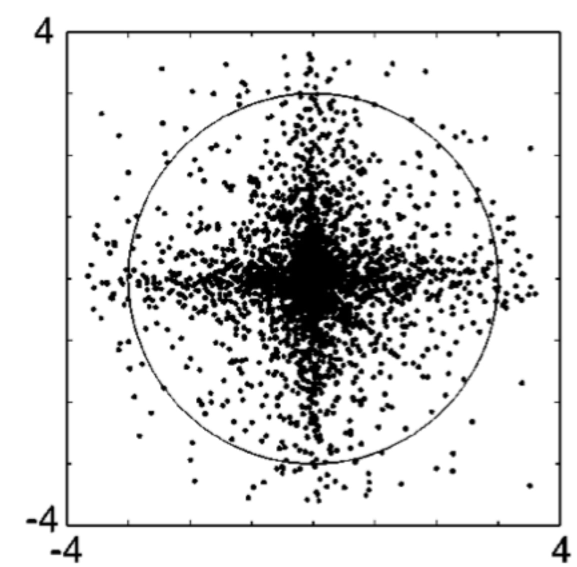
- Correlated, dependent

$$r_i r_j = \sigma^2 \delta_{ij}$$
$$p(r_1, r_2) \neq p(r_1)p(r_2)$$



- Uncorrelated, dependent

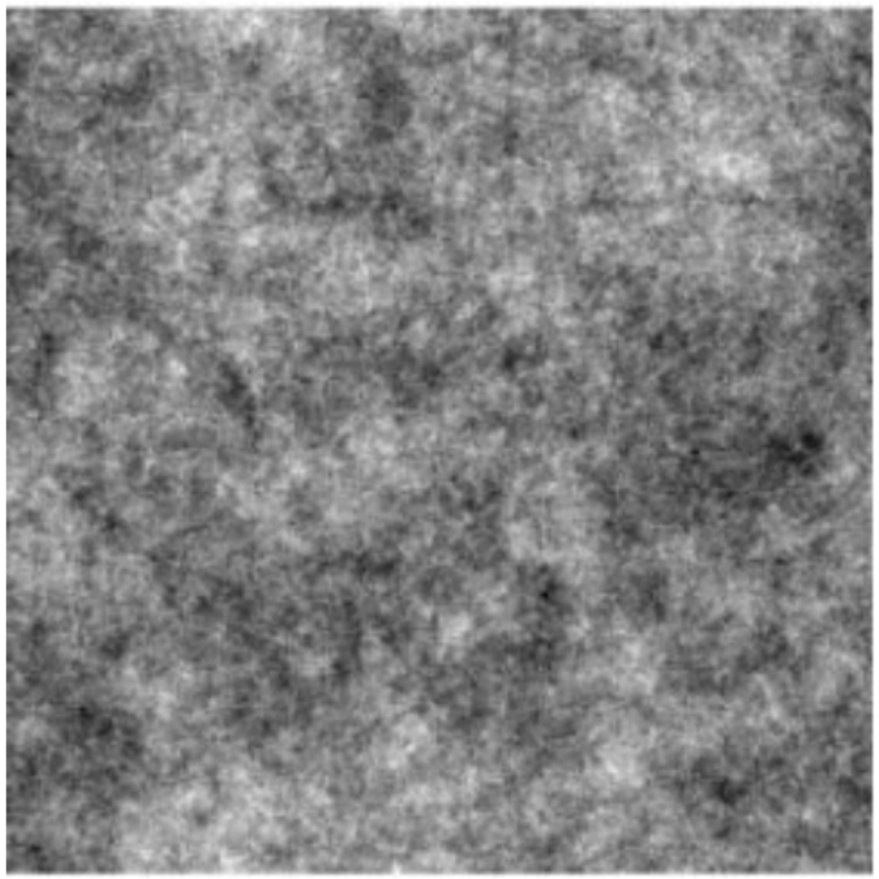
$$r_i r_j = \sigma^2 \delta_{ij}$$
$$p(r_1, r_2) = p(r_1)p(r_2)$$



- Uncorrelated, independent

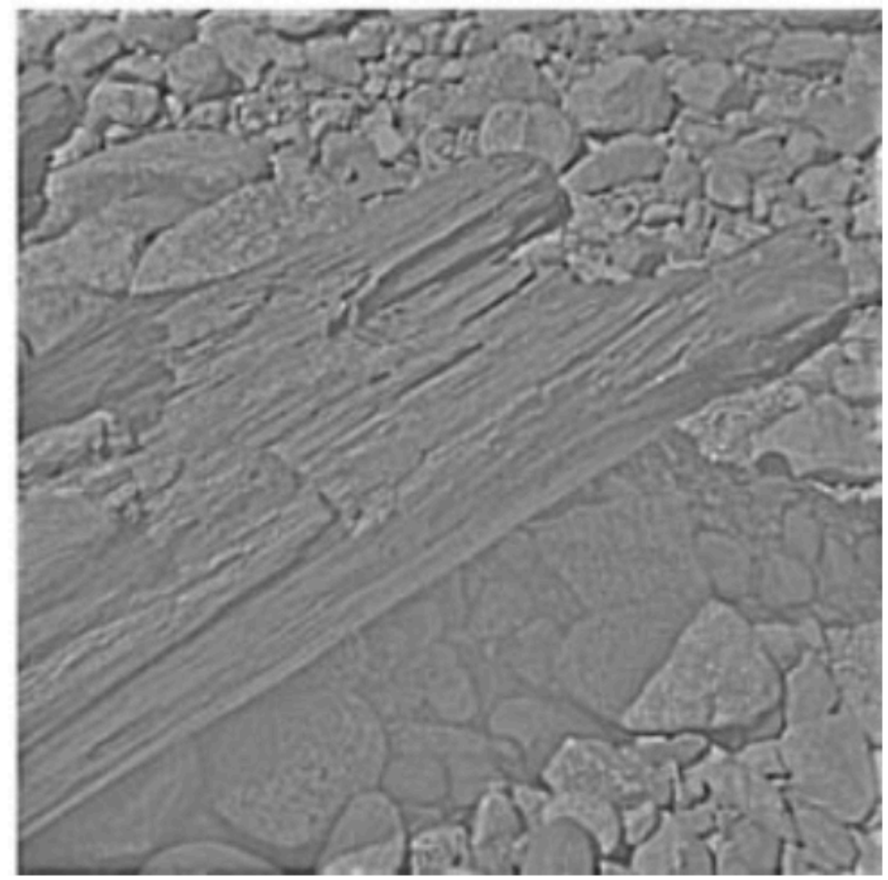
Beyond second order statistics

a.



$$C^{a}_{xy} = \langle s(x)s(y) \rangle$$

b.



$$C^{b}_{xy} = \langle s(x)s(y) \rangle$$

Covariance of ***a*** and ***b*** is the same….
Must be more than 2nd order statistics!

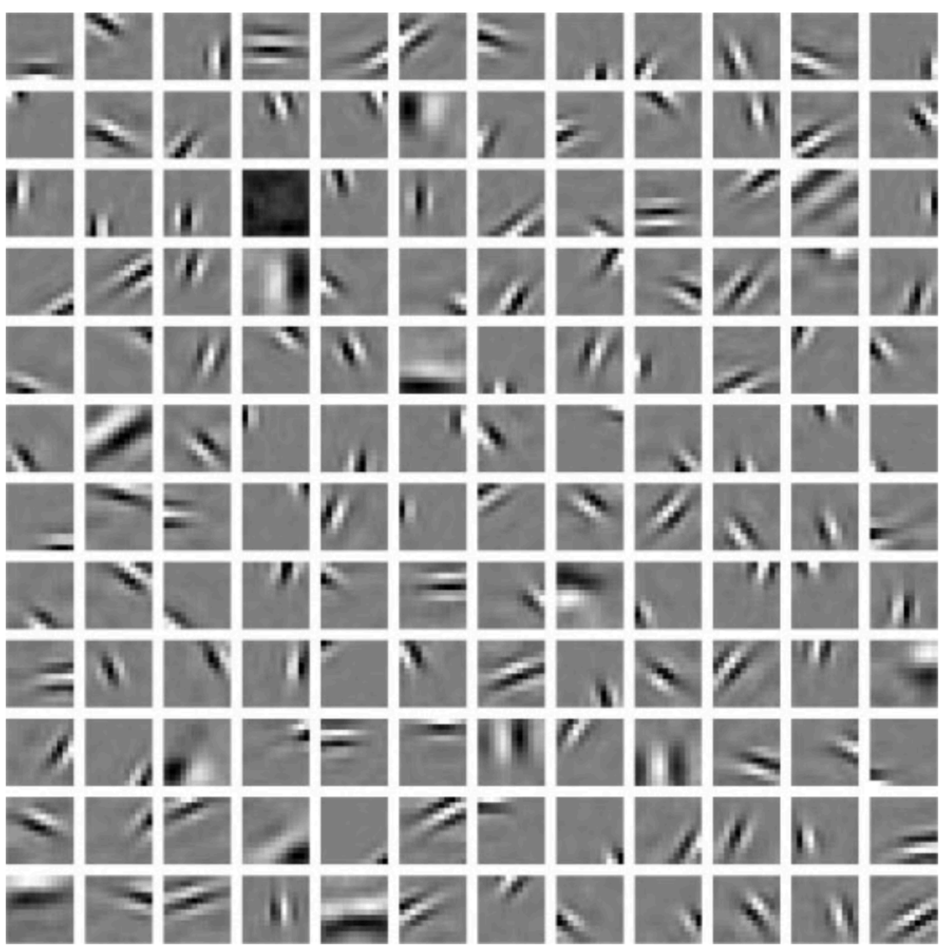
Bell & Sejnowski 1996, Olshausen & Field 1996

Independent component analysis

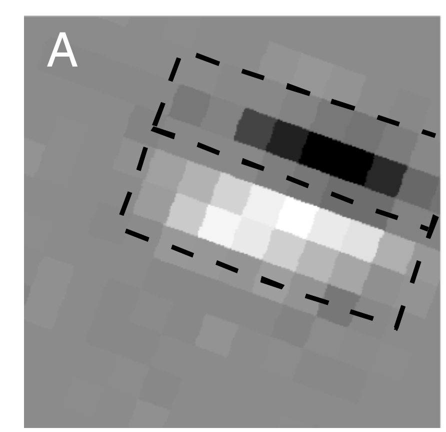
$$r = \sum_i \boldsymbol{w}_i x_i \qquad\qquad p(\boldsymbol{r}) \approx \prod p(r_i)$$

ICA (independent component analysis): maximises $H(\boldsymbol{r})$

Learned filters:



Primary visual cortex:



Carandini et al. 2005

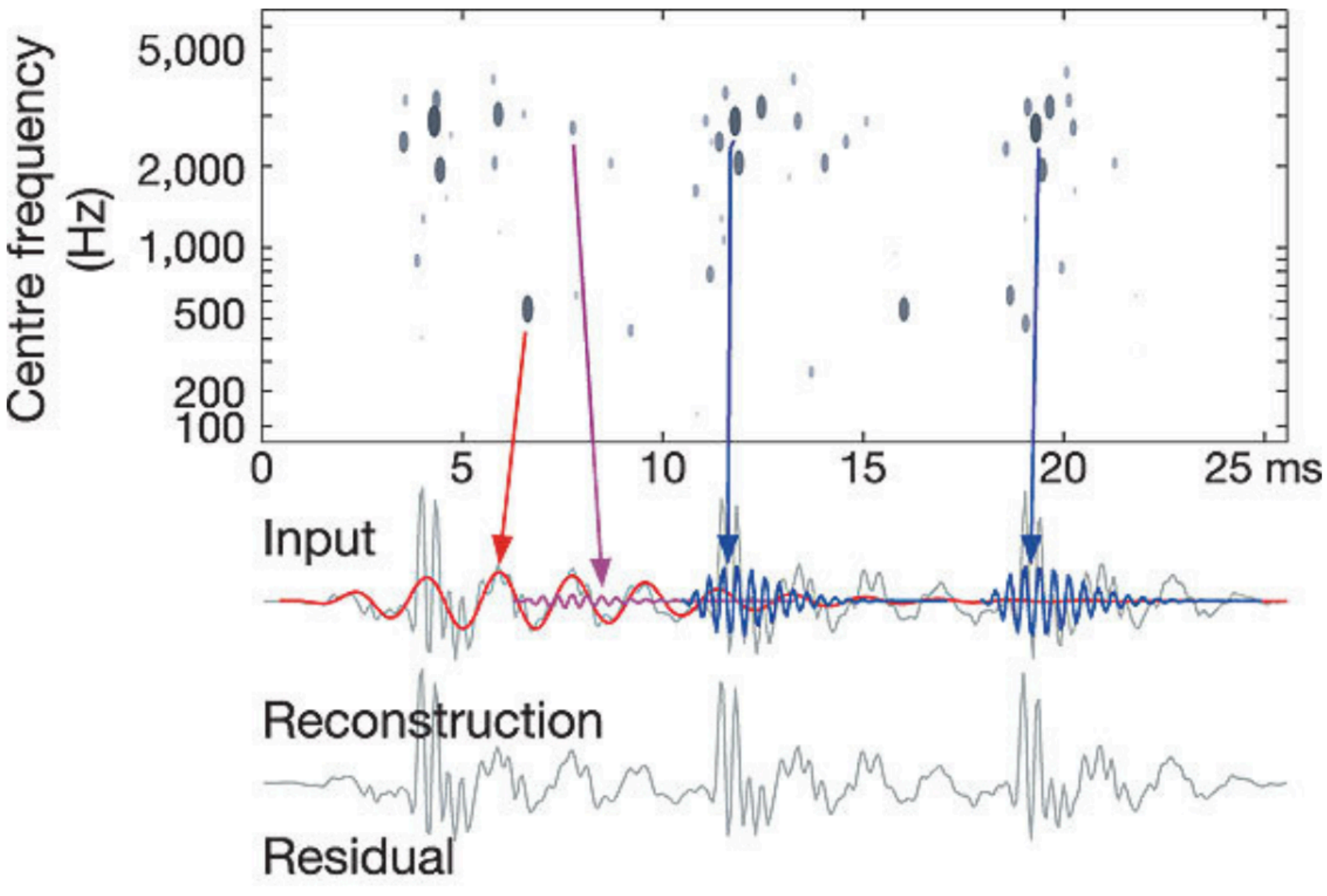
Independent component analysis

TD: plot the histogram of $z_{ICA} = W_{ICA}x$, compare to $z = Wx$

What is the difference? Can you comment ?

Auditory coding in the cochlear

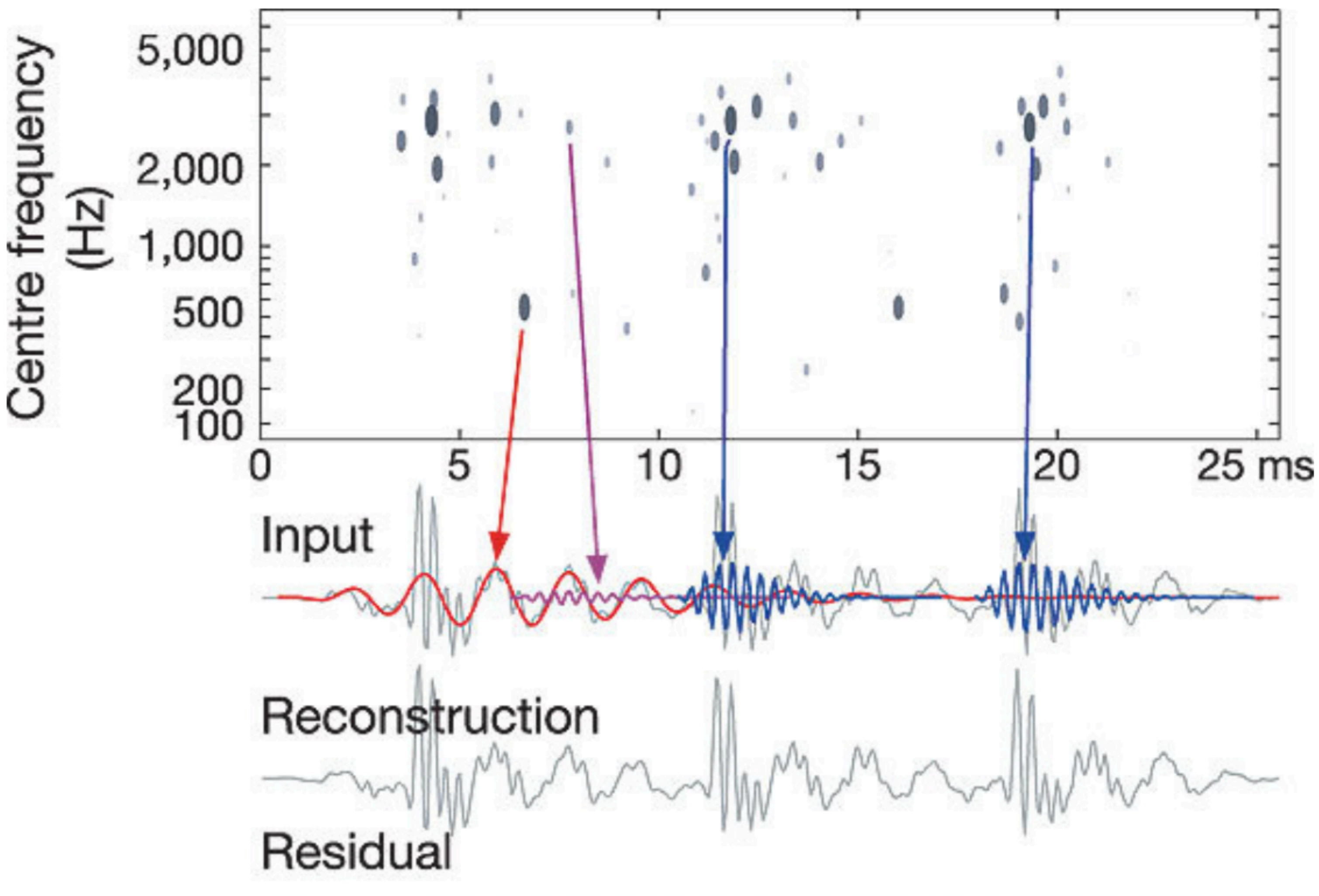
Efficient coding model of auditory nerve fibres

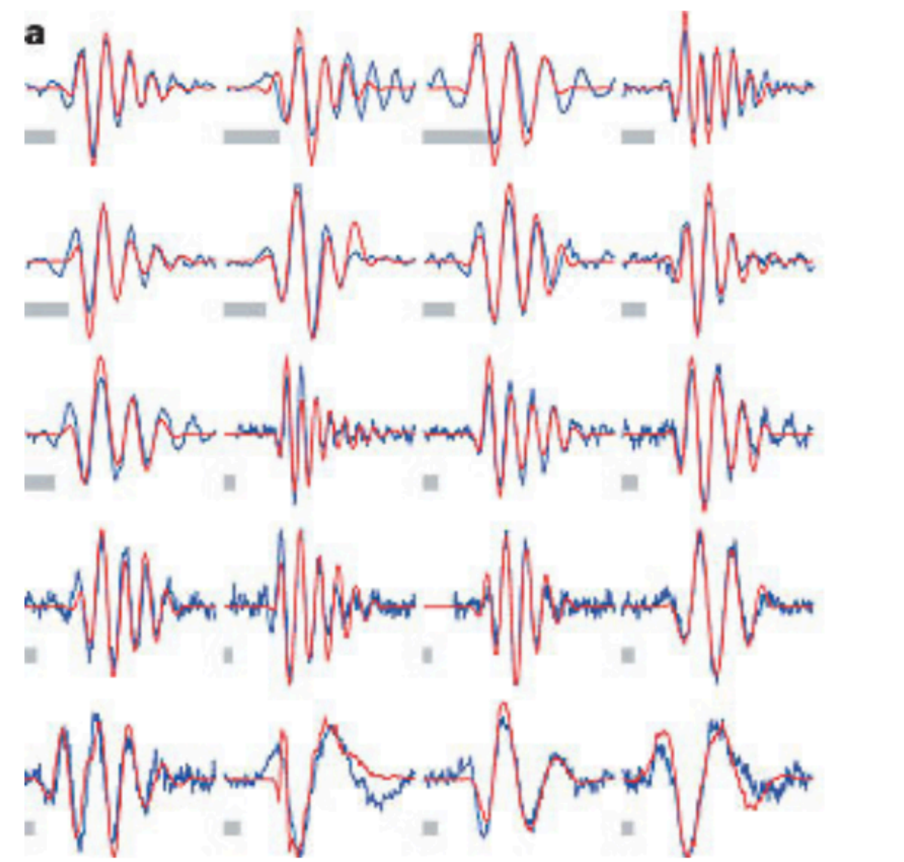


Smith & Lewicki 2006

Auditory coding in the cochlear

Efficient coding model of auditory nerve fibres





Beyond independent component analysis

TD2: plot the conditional histograms of two units, $p(z_1 \,|\, z_2)$

Hint: I approximated $p(z_1, z_2)$ from data (using histogram)

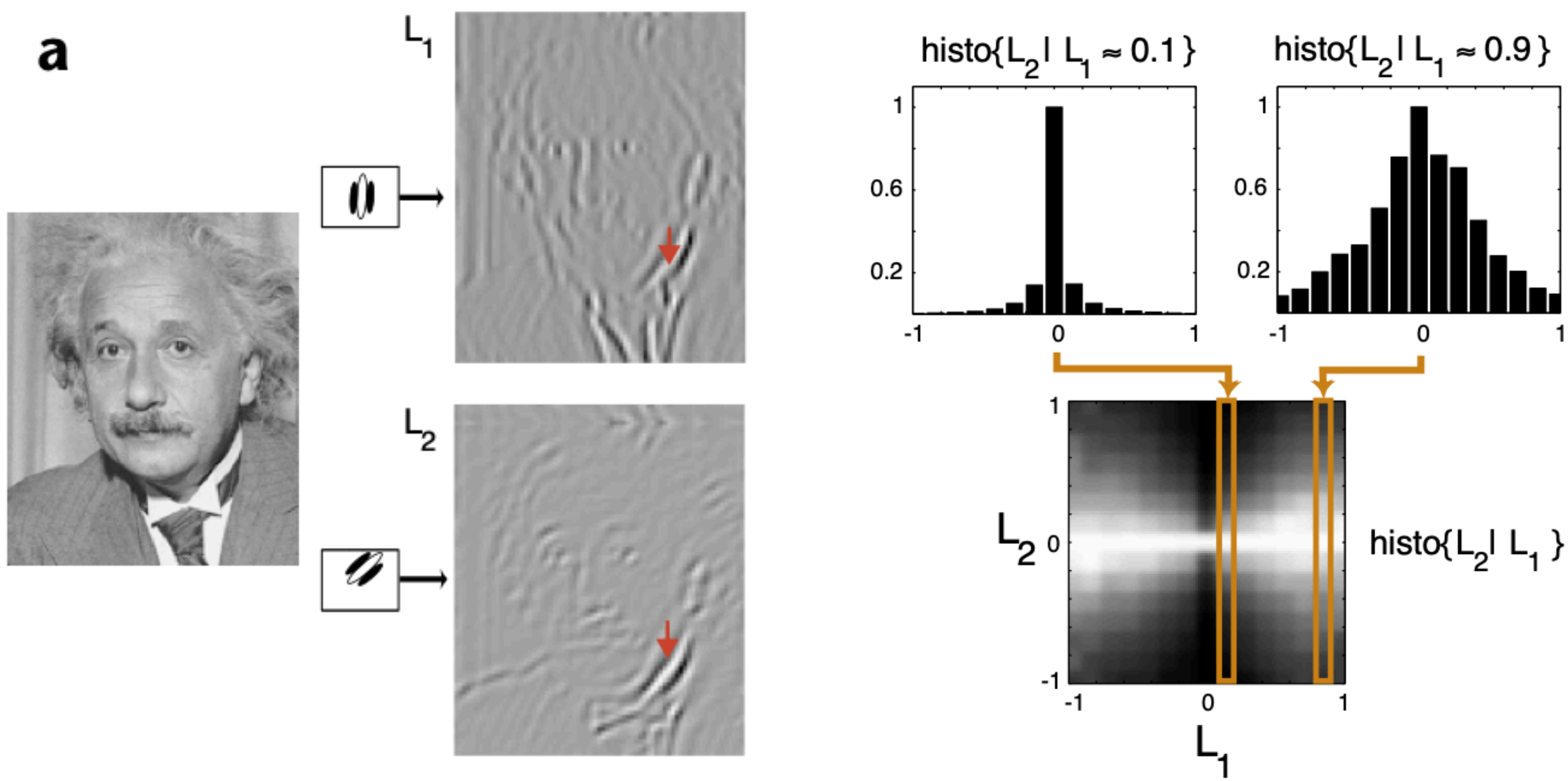
Compute $p(z_1)$

Remember, from Bayes' theorem: $p(z_1 \,|\, z_2) = \dfrac{p(z_1, z_2)}{p(z_2)}$

Are the units independent of each other?
What does this mean?

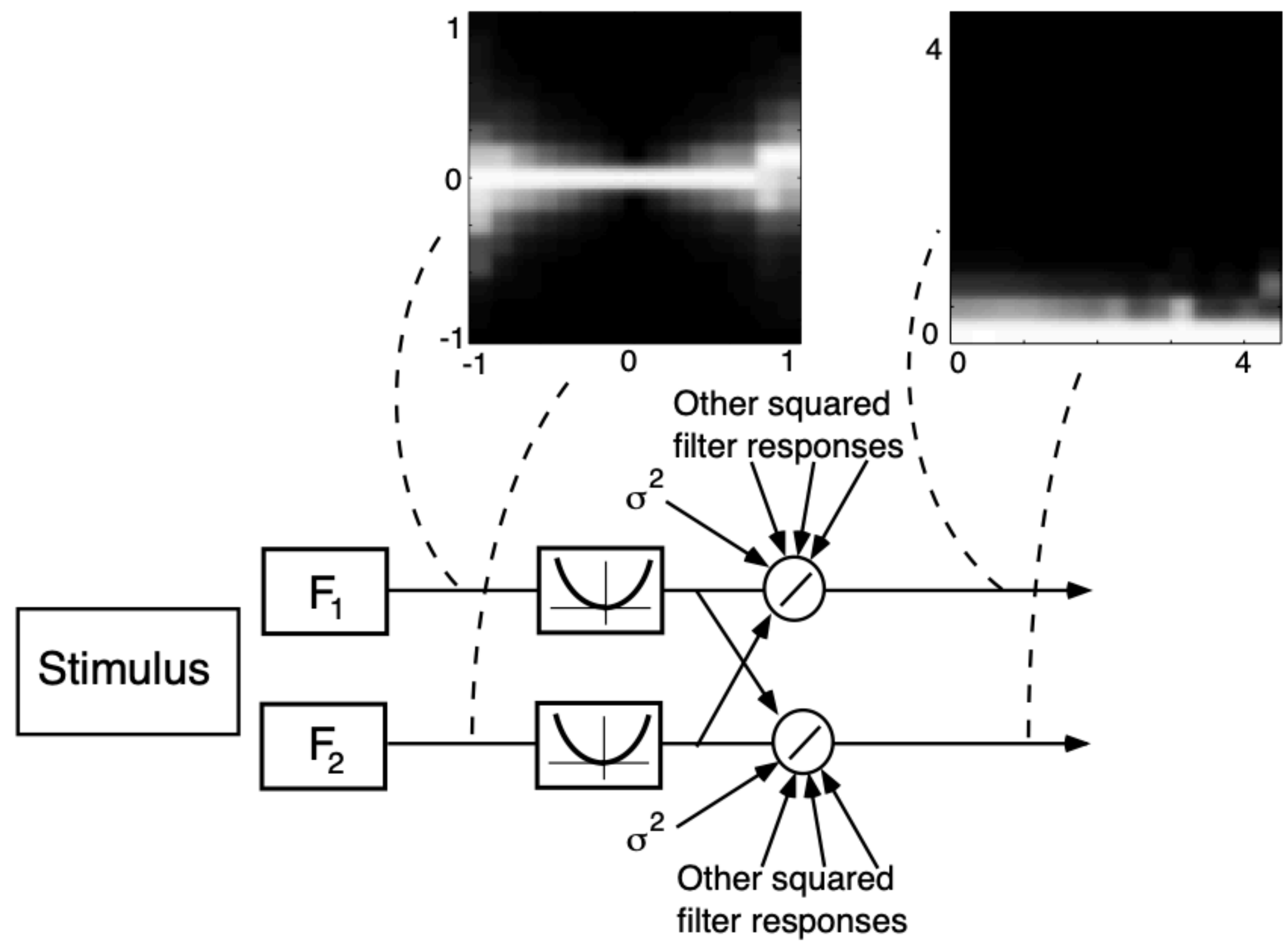
Beyond linear coding

Best linear filters ***not*** independent!



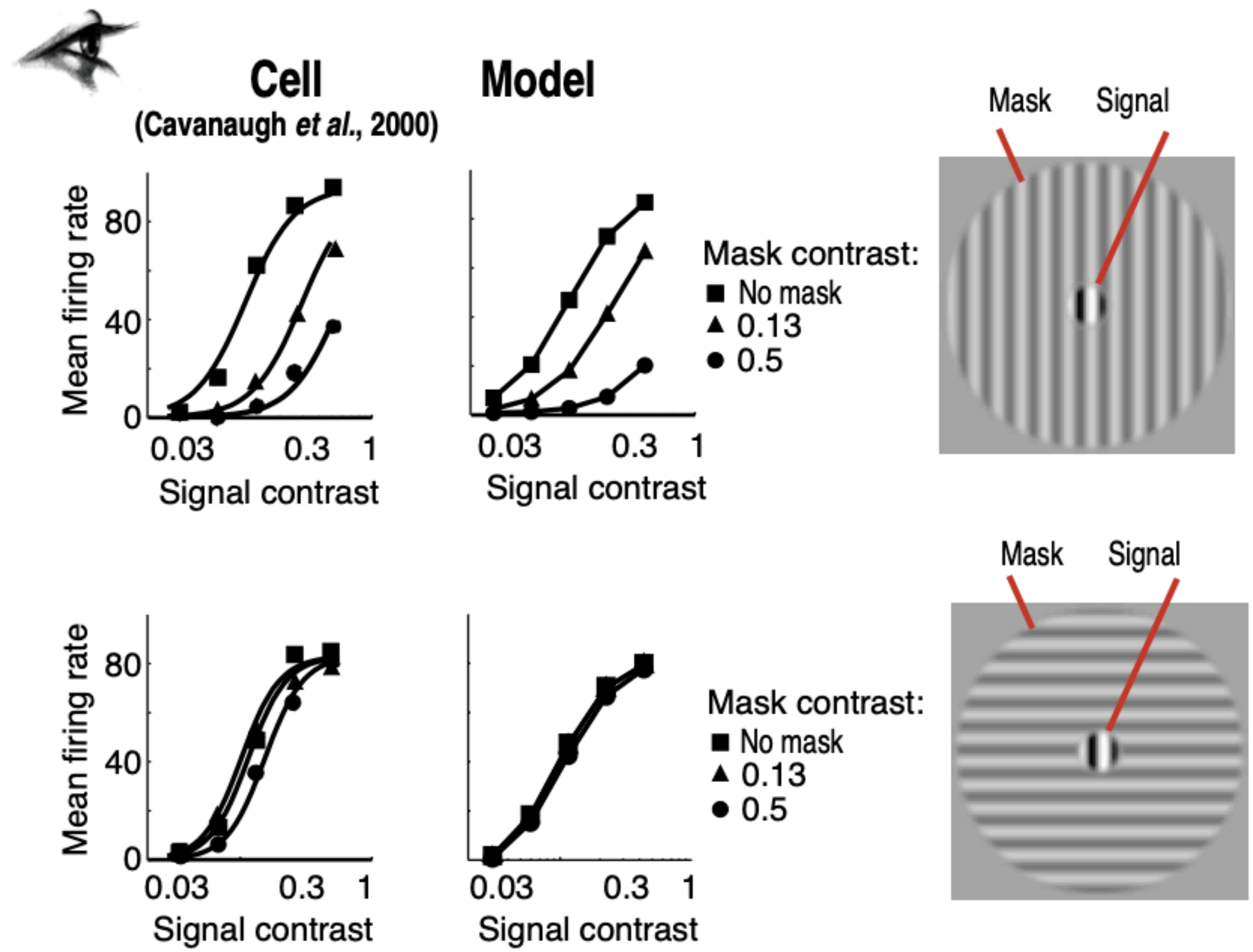
Beyond linear coding

$$R_i = \frac{L_i^2}{\sum_j w_{ji} L_j^2 + \sigma^2}$$



Beyond linear coding

$$R_i = \frac{L_i^2}{\sum_j w_{ji} L_j^2 + \sigma^2}$$



Encoding ‘relevant’ information

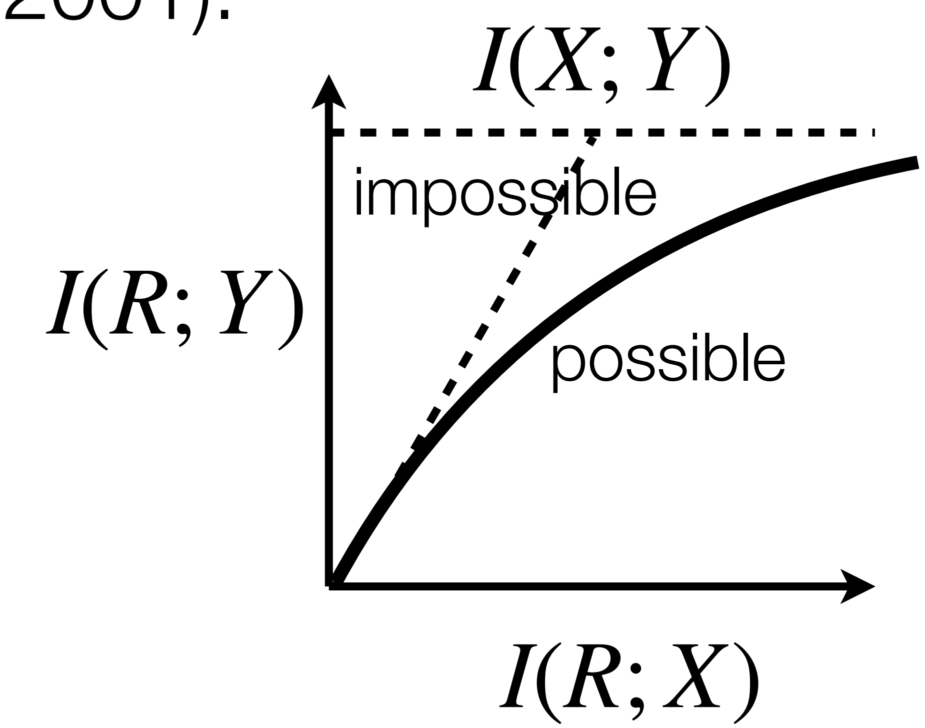
Not all visual information is equal.

$$X \rightarrow R \rightarrow Y$$

input response relevant variable

Information bottleneck (Tishby et al. 2001):

$$L_{p(R|X)} = I(R;Y) - \gamma I(R;X)$$

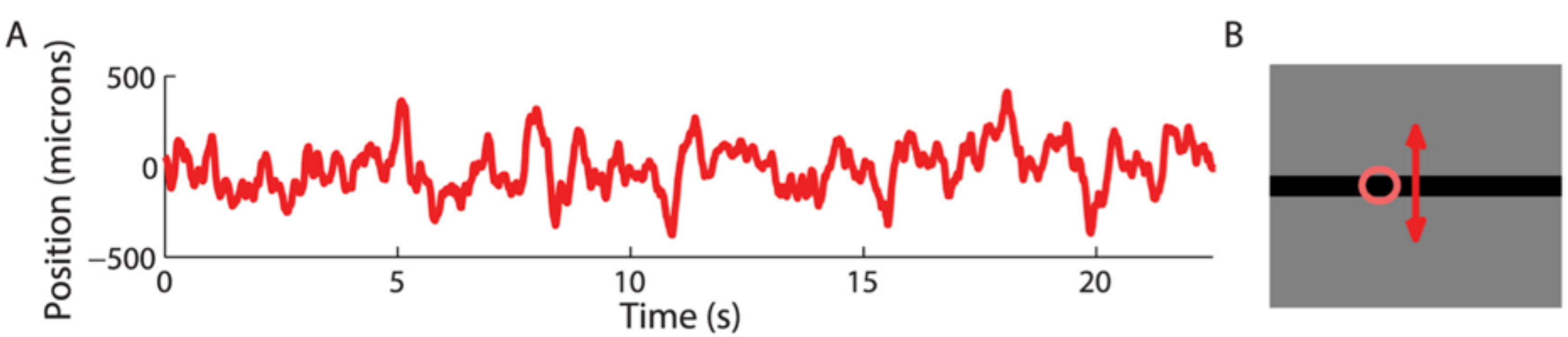


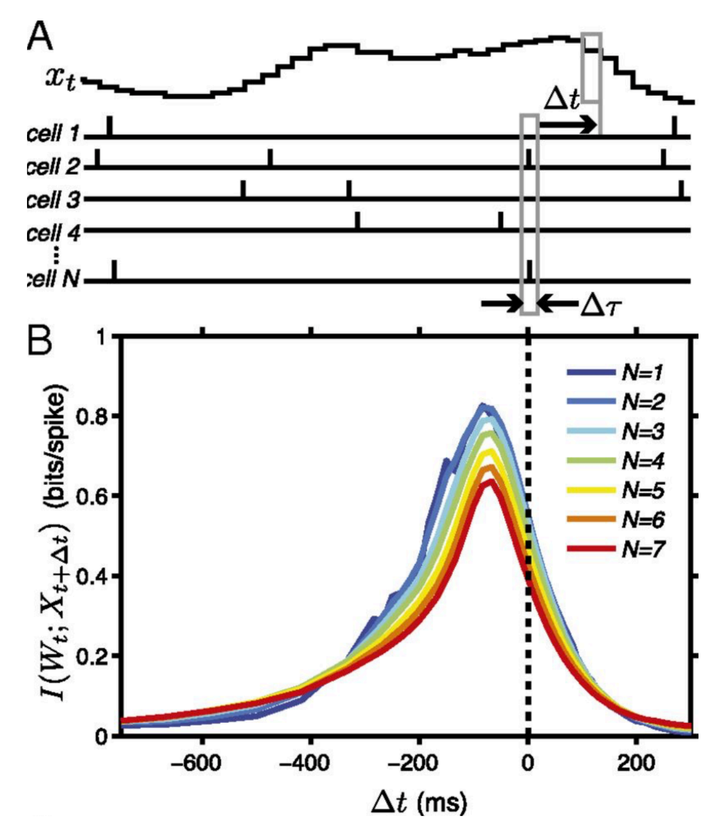
Palmer et al., PNAS, 2015

Encoding predictive information

Predictions are important:

$$L_{p(R_t|X_t)} = I(R_t; X_{>t}) - \gamma I(R_t; X_t)$$



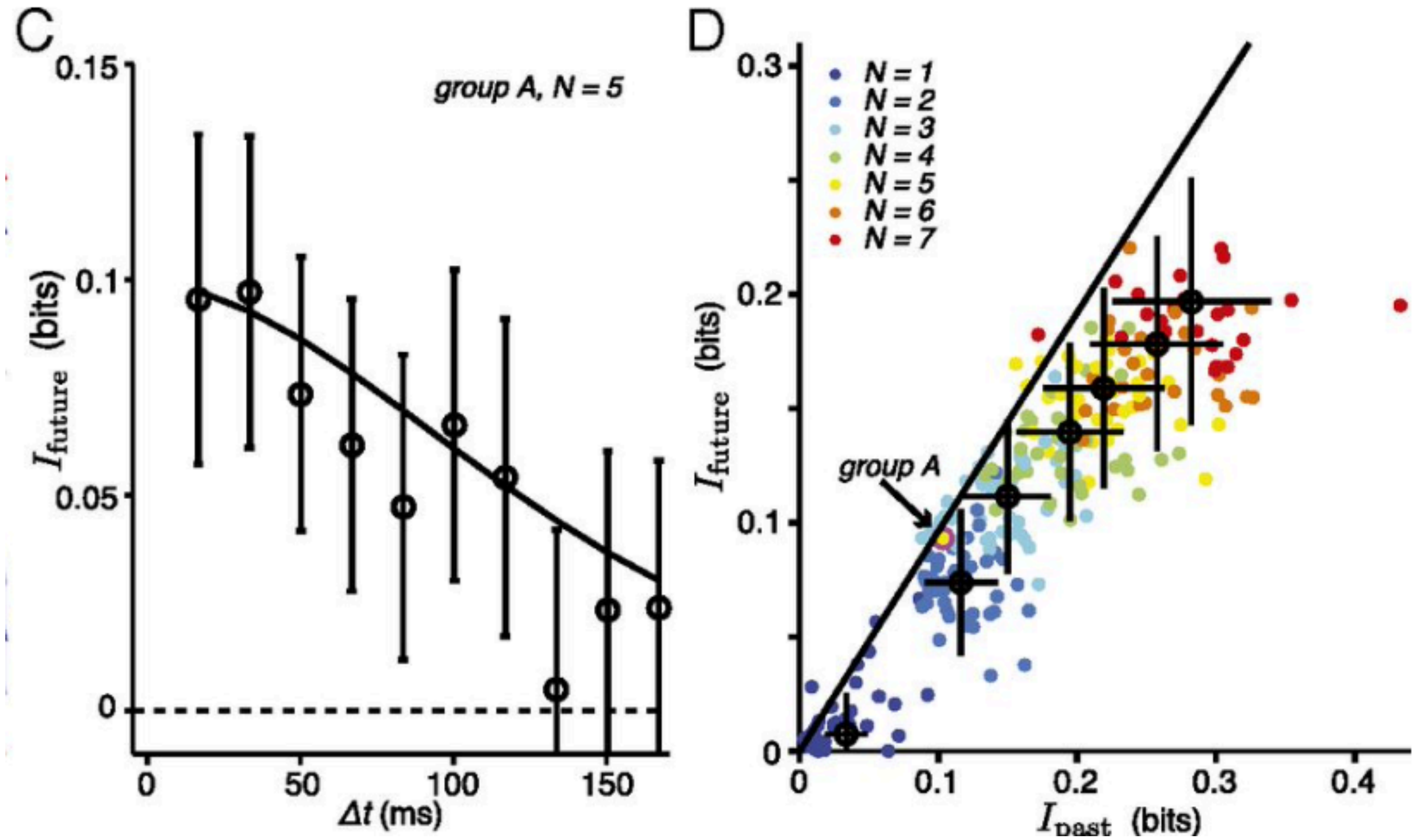


Palmer et al., PNAS, 2015

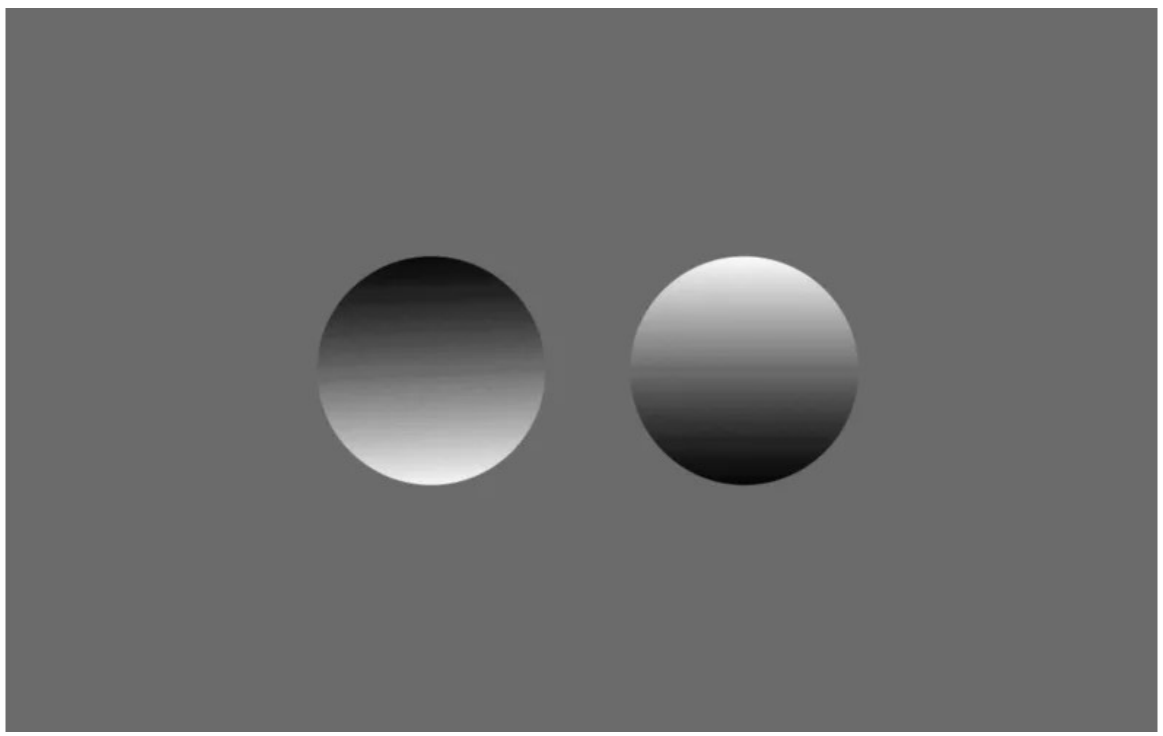
Encoding relevant predictive information

We need to predict the future:

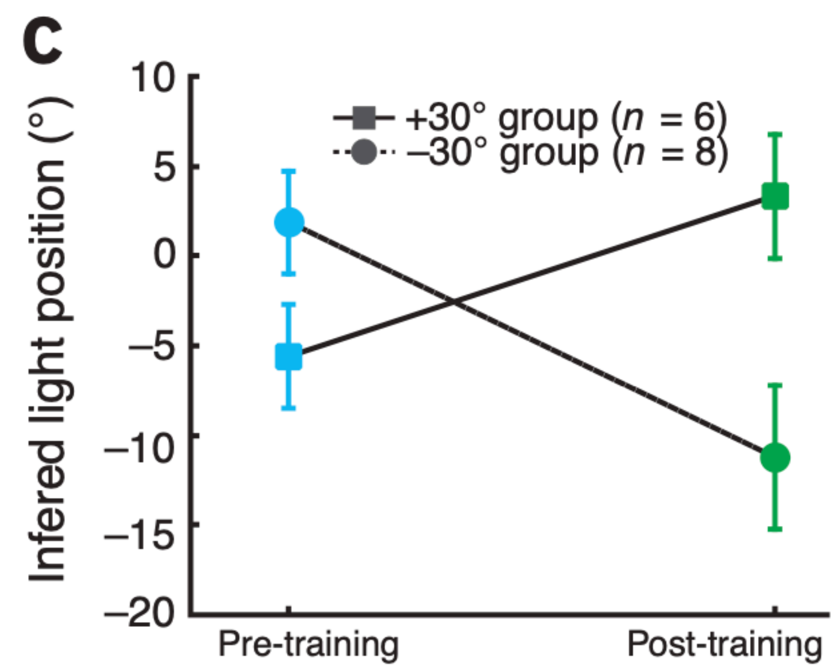
$$L_{p(R_t|X_t)} = I(R_t; X_{>t}) - \gamma I(R_t; X_t)$$



The Bayesian brain



Can be reversed with experience!



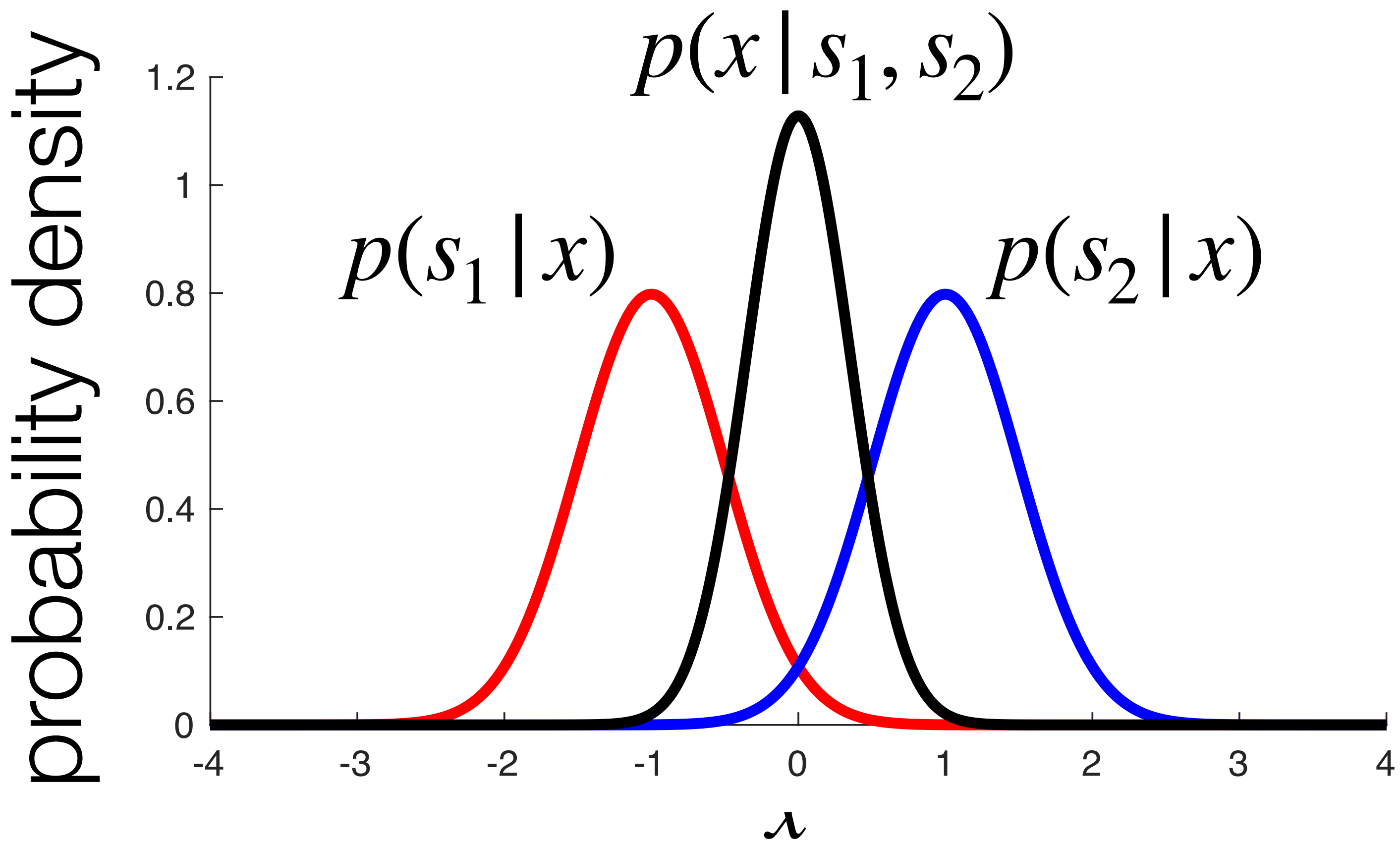
Bayes' theorem:

$$p(\text{world} \mid \text{observation}) \propto p(\text{observation} \mid \text{world}) \times p(\text{world})$$

The Bayesian brain

Bayesian cue combination

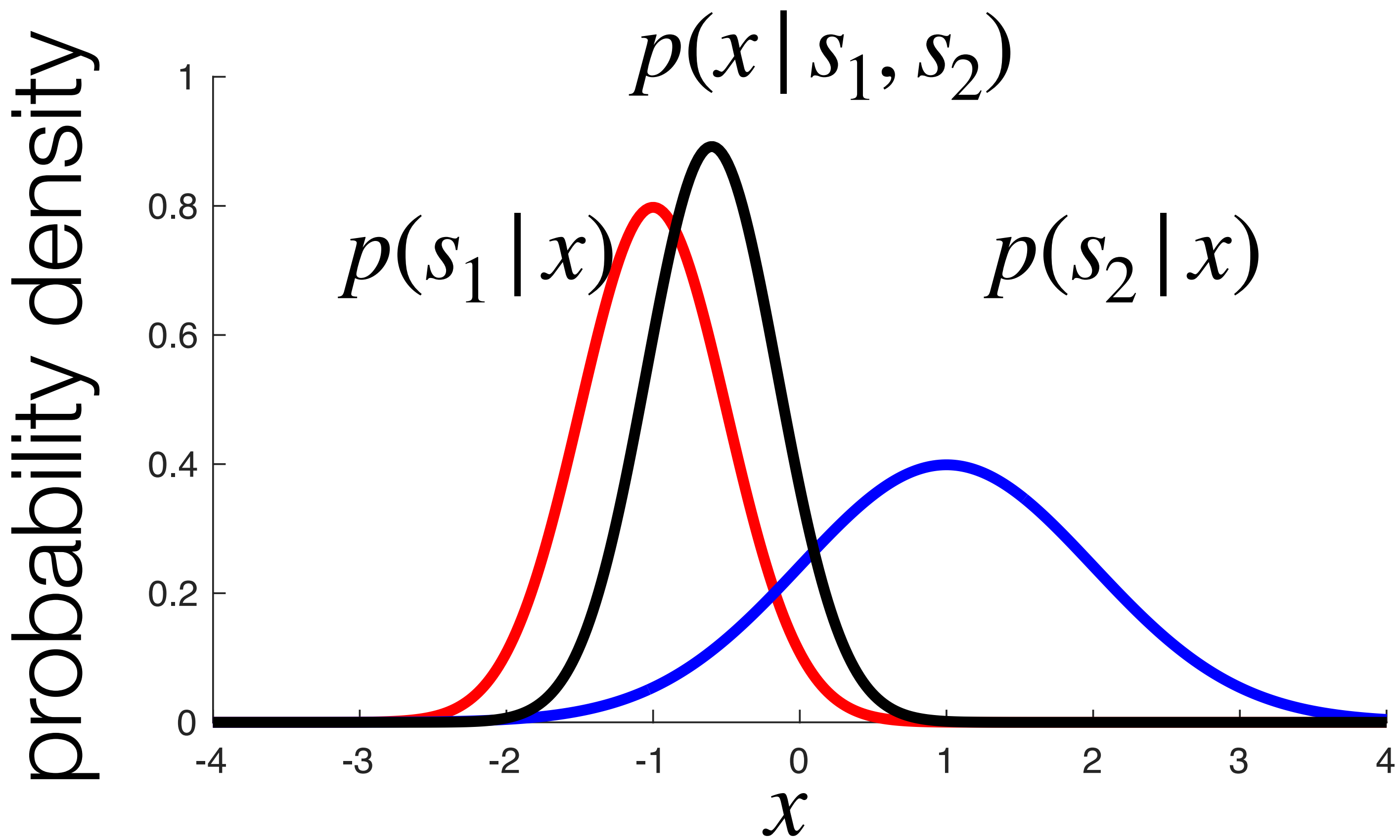
$$p(x \mid s_1, s_2) \propto p(s_1 \mid x)p(s_2 \mid x)p(x)$$



The Bayesian brain

Bayesian cue combination

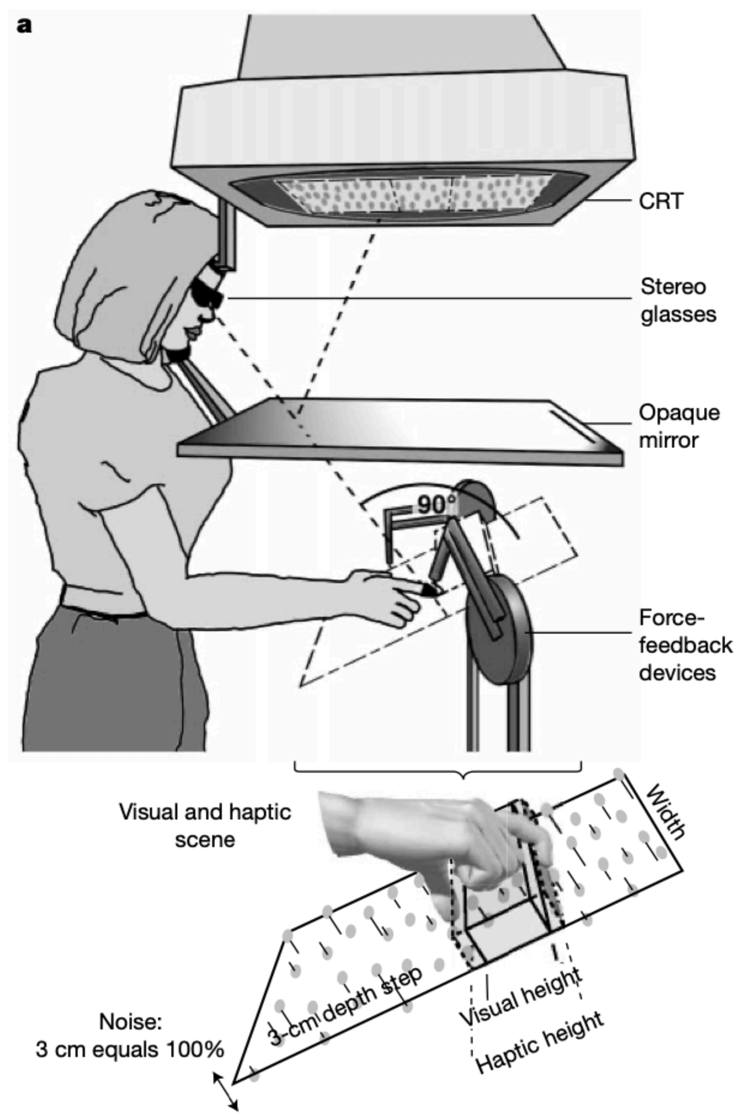
$$p(x \mid s_1, s_2) \propto p(s_1 \mid x)p(s_2 \mid x)p(x)$$

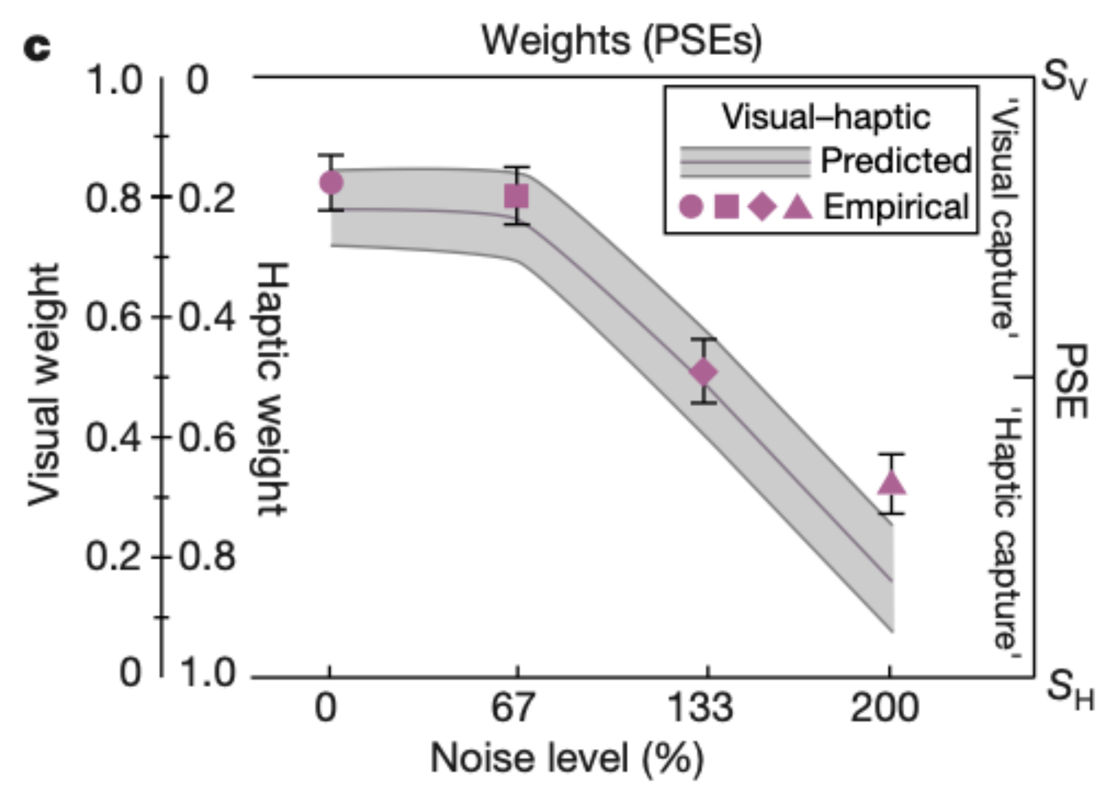


The Bayesian brain

Bayesian cue combination

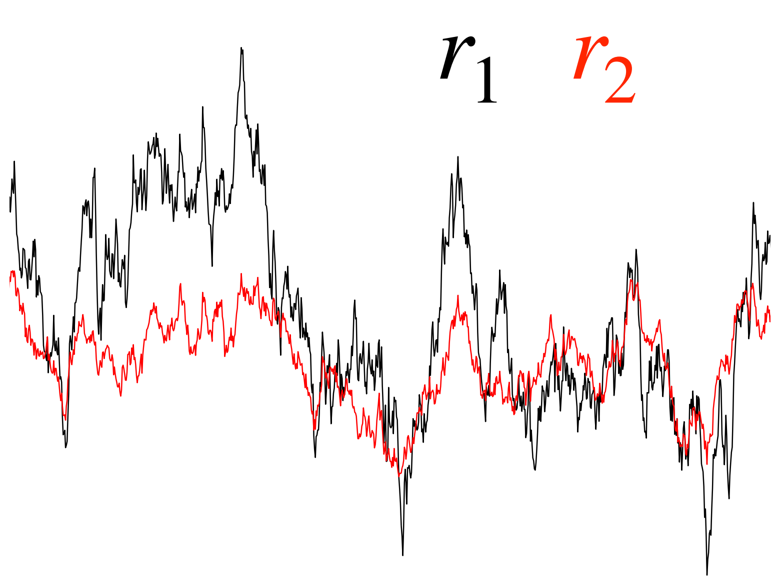
$$p(x \mid s_1, s_2) \propto p(s_1 \mid x)p(s_2 \mid x)p(x)$$

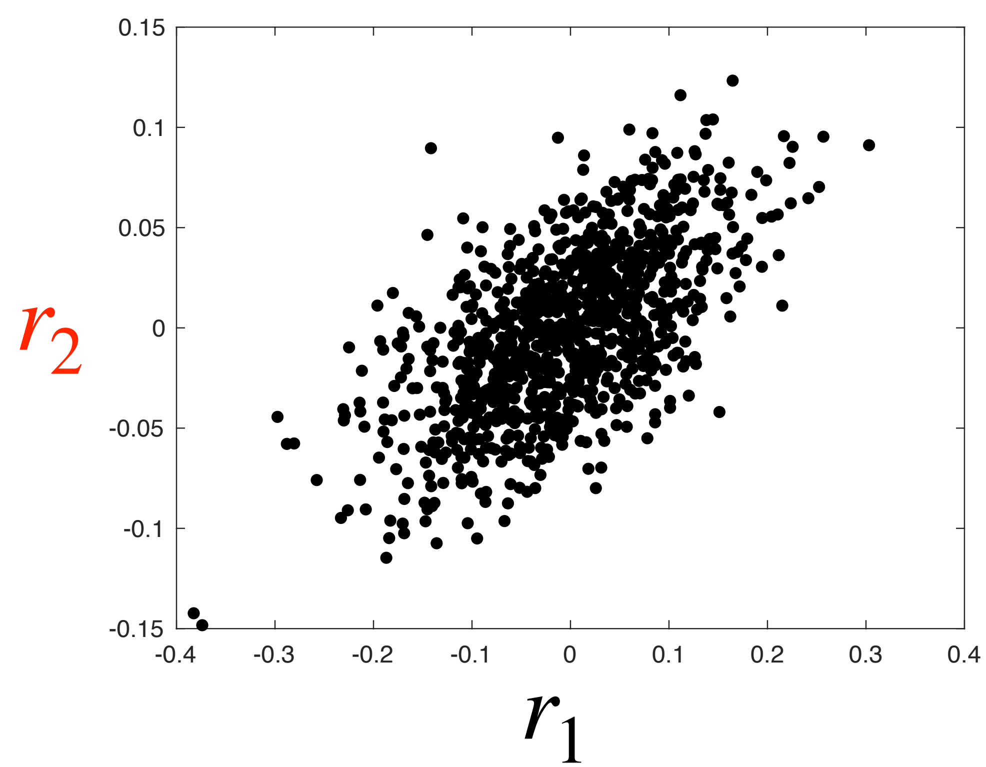




Bayesian neural codes

- Do visual neurons encode uncertainty (and probability distributions)?
  - Parametric codes: encode parameters of distribution (e.g mean & variance)
  - Sampling codes: encode samples from a distribution: noise = uncertainty



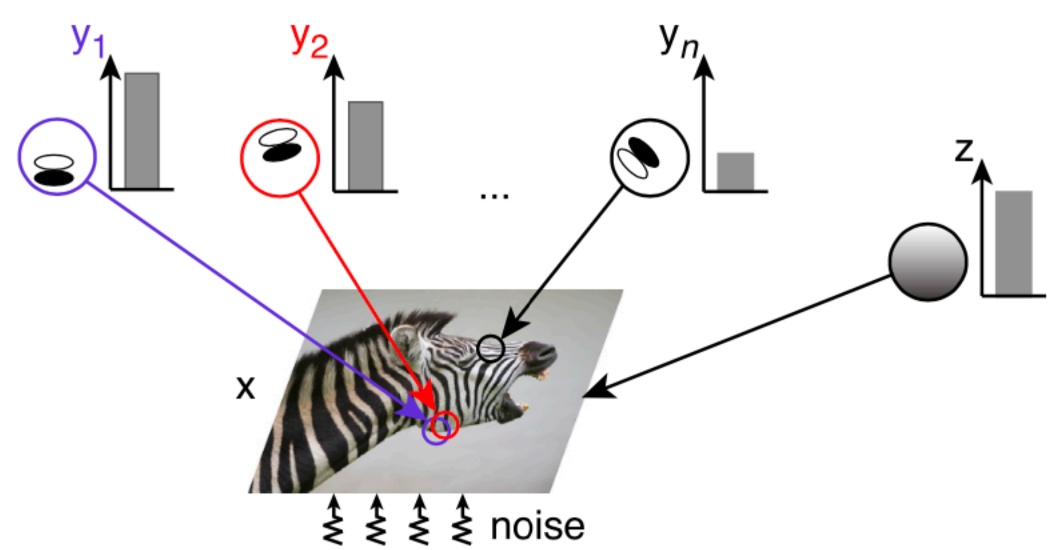


Probabilistic sampling in V1

- **Construct** latent generative model of natural images:

$$p(x \mid y)p(y)$$

image   latent feature

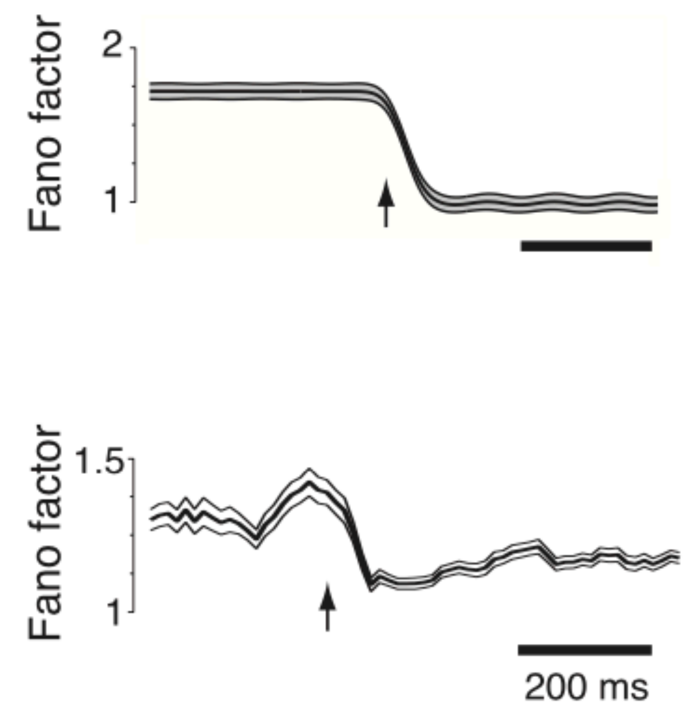


- **Assume** individual neurons sample posterior probability over latent features

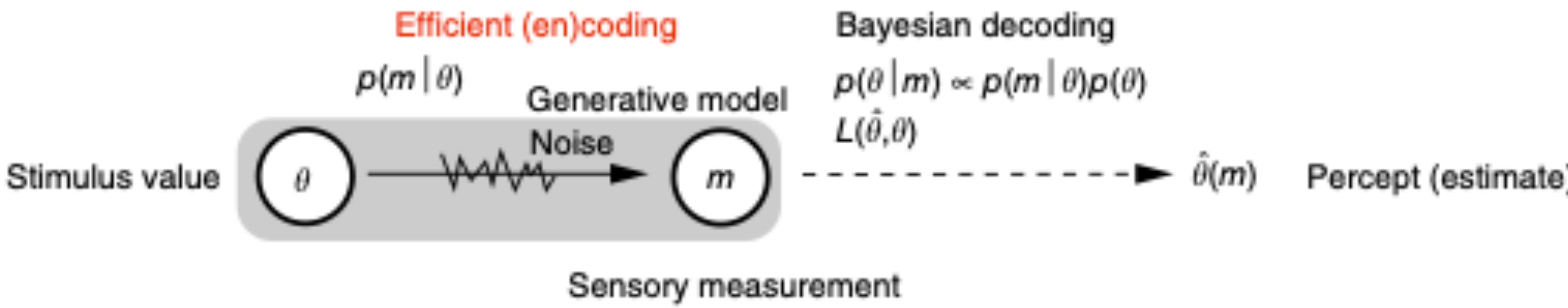
$$r \sim p(y \mid x)$$

- **Prediction:** neural variability should reflect uncertainty

- Can predict various aspects of V1 neurons responses. (e.g. quenching of noise at stimulus onset)



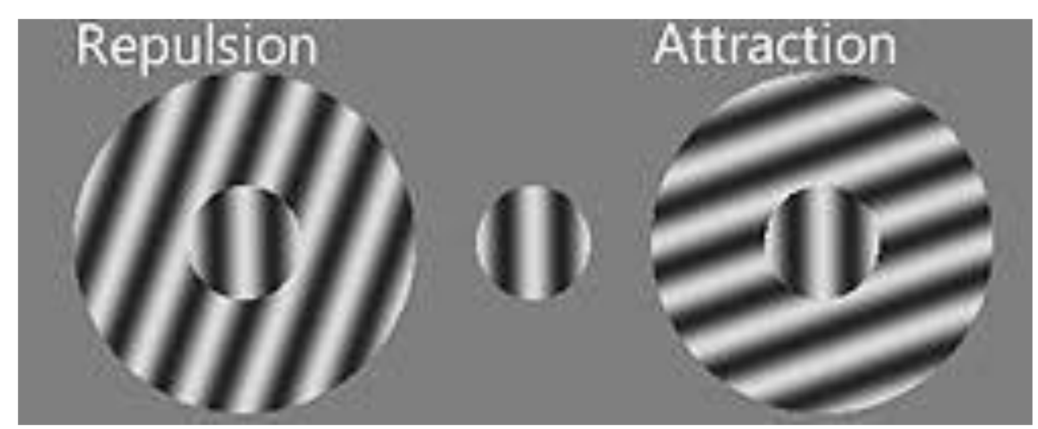
Bayesian inference meets efficient coding



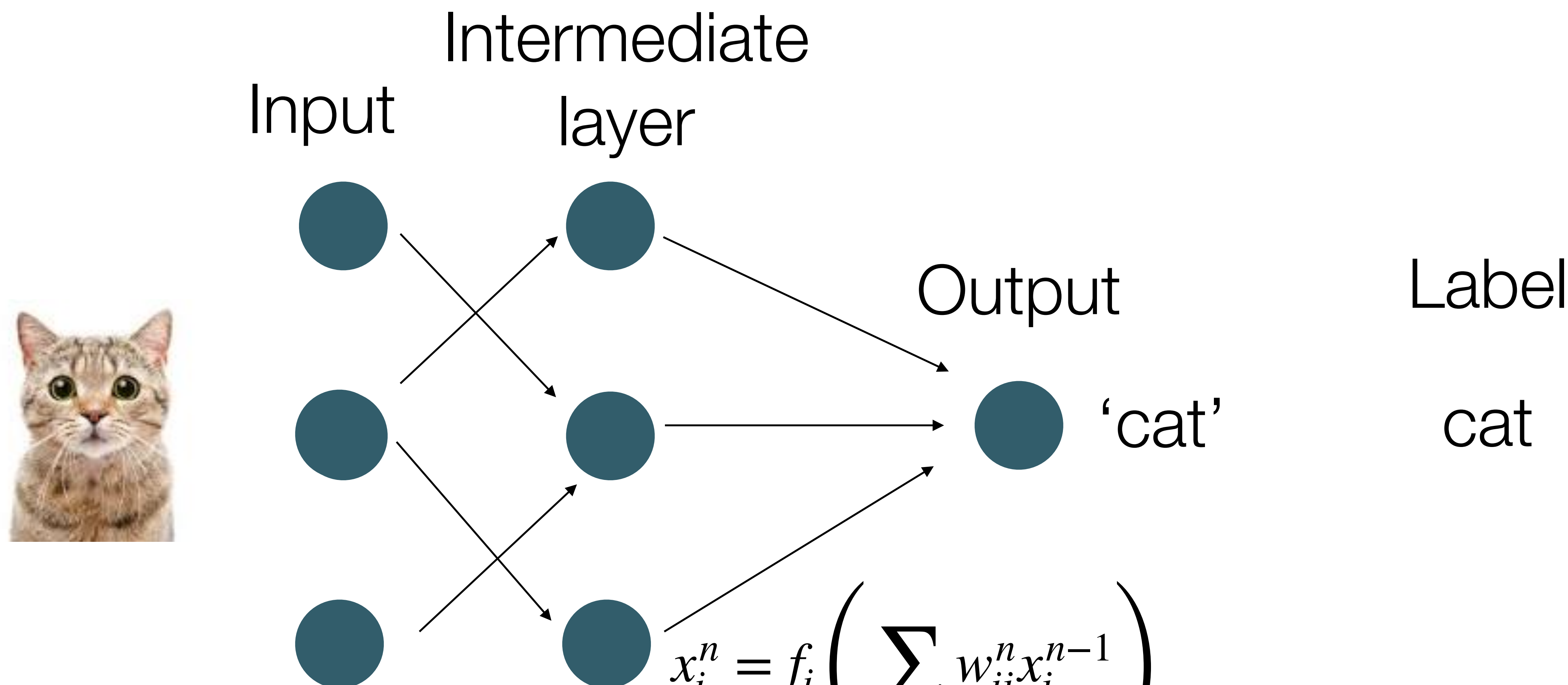
Efficient coding: sensory neurons dedicate resources into encoding most frequent stimuli

Bayesian decoding: downstream neurons 'decode' using Bayesian inference

Can explain repulsive/attractive perceptual biases (Wei & Stocker 2015 Nature Neuro)



Artificial neural networks (ANNs)



$$x_i^n = f_i\left(\sum_i w_{ij}^n x_j^{n-1}\right)$$

Define **loss function**: $L(o_w(x), o^{target})$

Learn weights using **gradient descent:**

$$w \rightarrow w - \eta \nabla_w L$$

Artificial neural networks (ANNs)

Define loss function: $L(o_w(x), o^{target})$

Learn weights using **gradient descent** on training data-set:

$$w \rightarrow w - \eta \nabla_w L$$

Back-propogation: efficiently compute gradients (by propogating through network)

Stochastic gradient descent: approximate gradient using small subset of data: allows for *large* ($>10^6$) data-sets

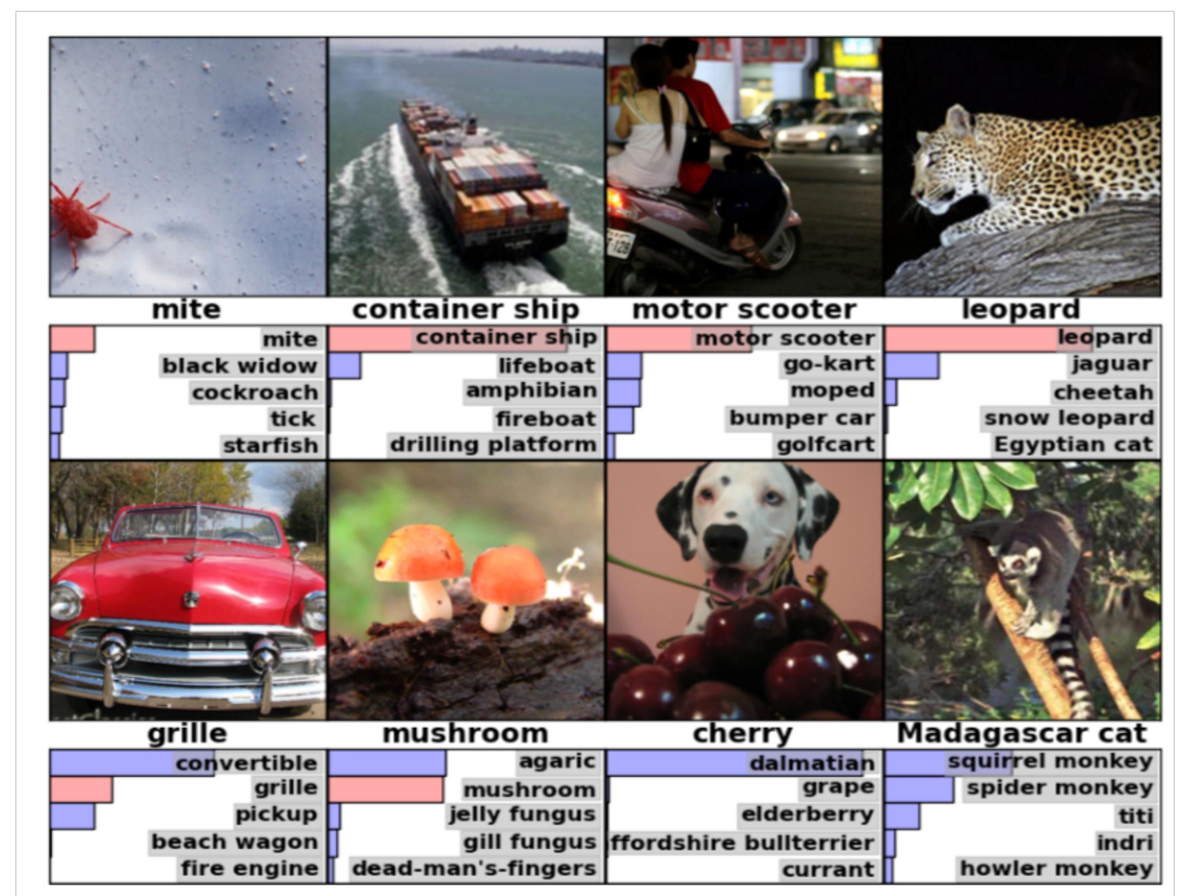
Convolution: each 'channel' in the network performs spatial convolution of previous layer.

Imagenet challenge

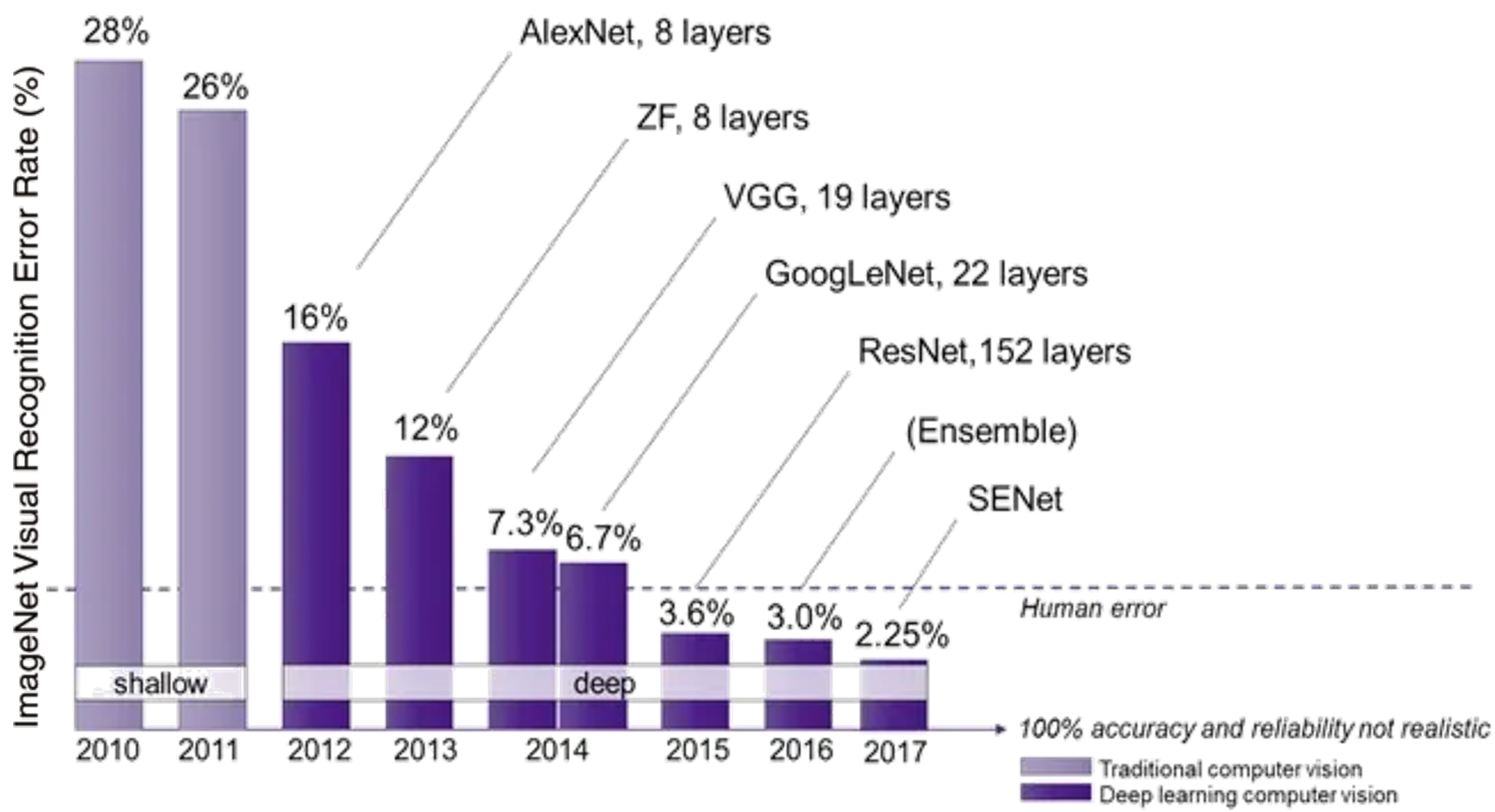
Imagenet large scale visual recognition challenge

14M images, 1000 classes

Top 1 and top 5 accuracy (metric)

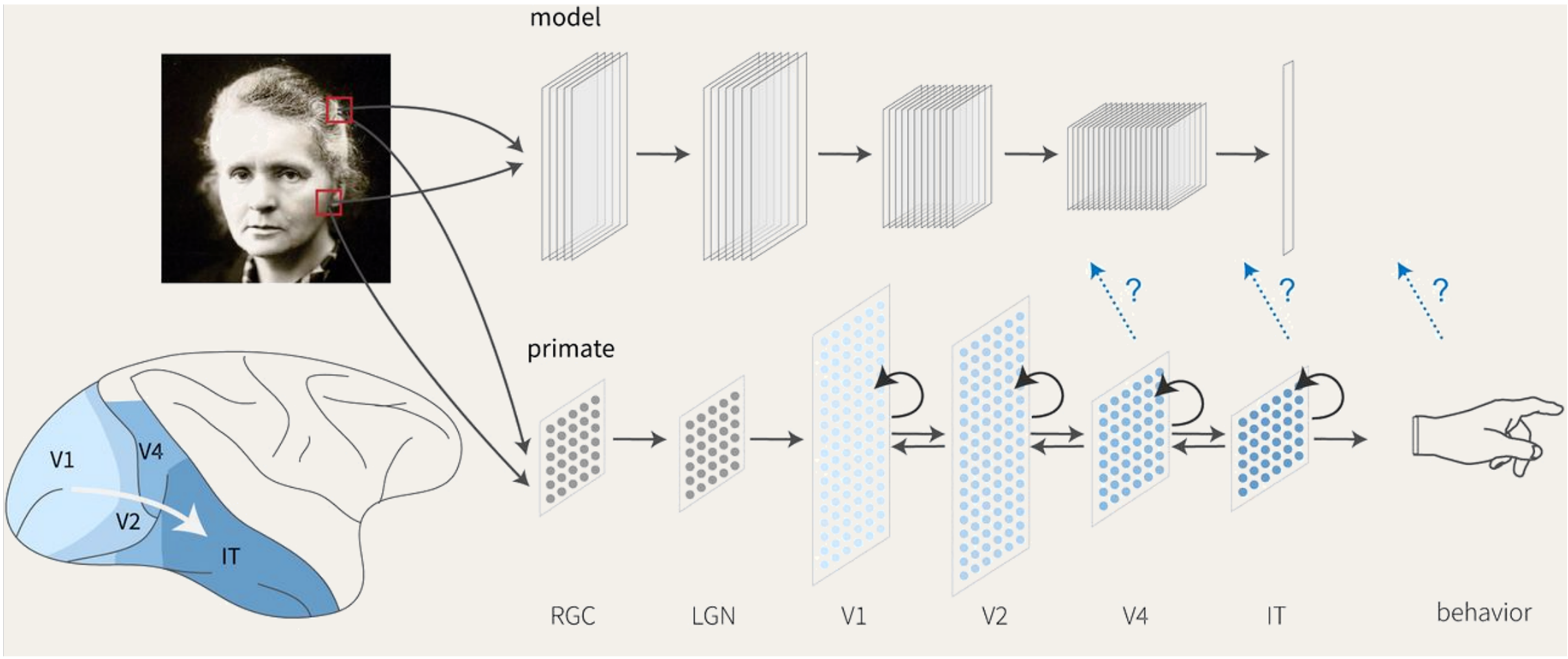


Deep learning revolution



But are deep networks ‘brain like’?

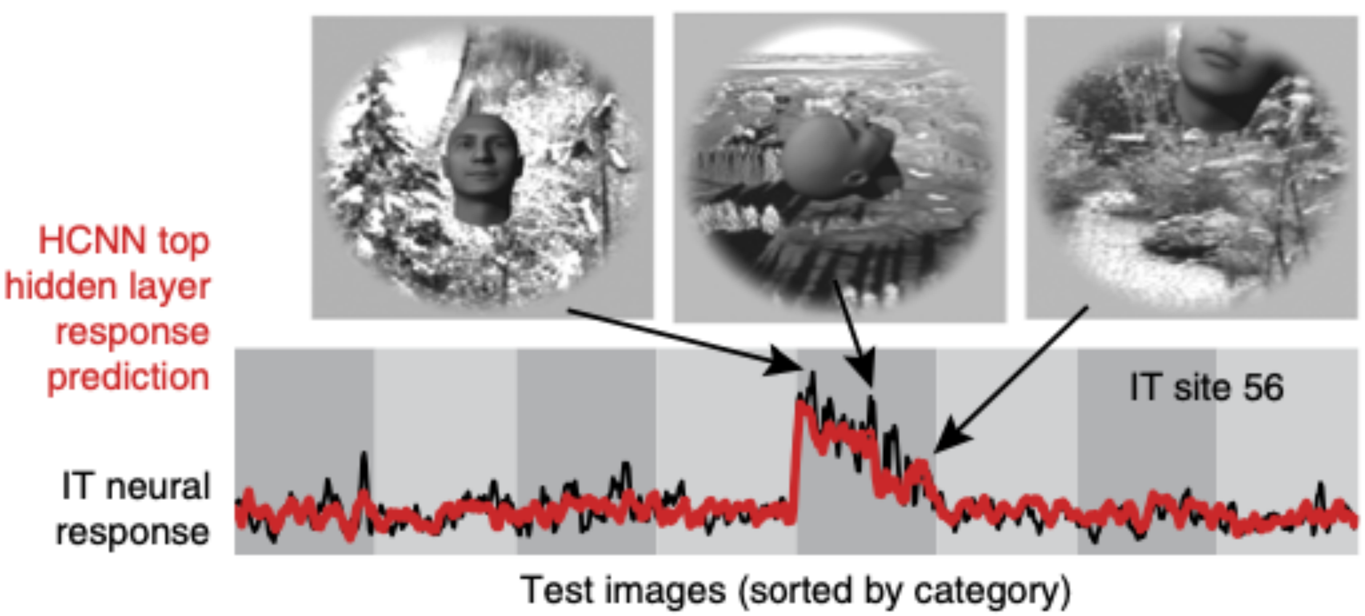
How do ANNs compare to brain?



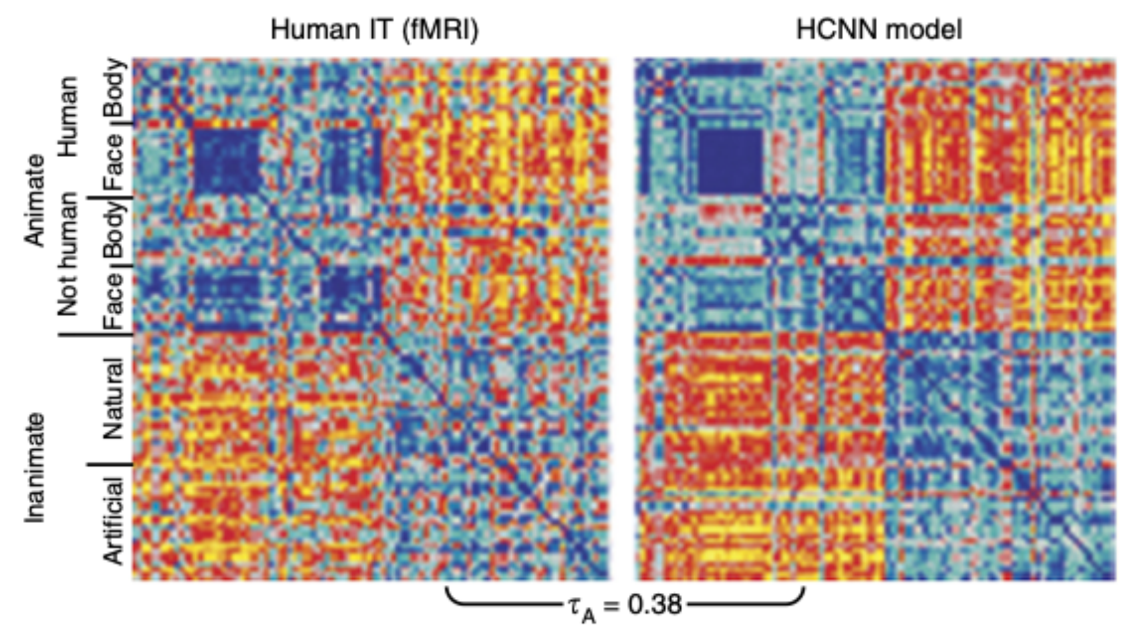
Can compre different layers in ANN, and different layers in visual processing

Comparing ANNS with neural activity

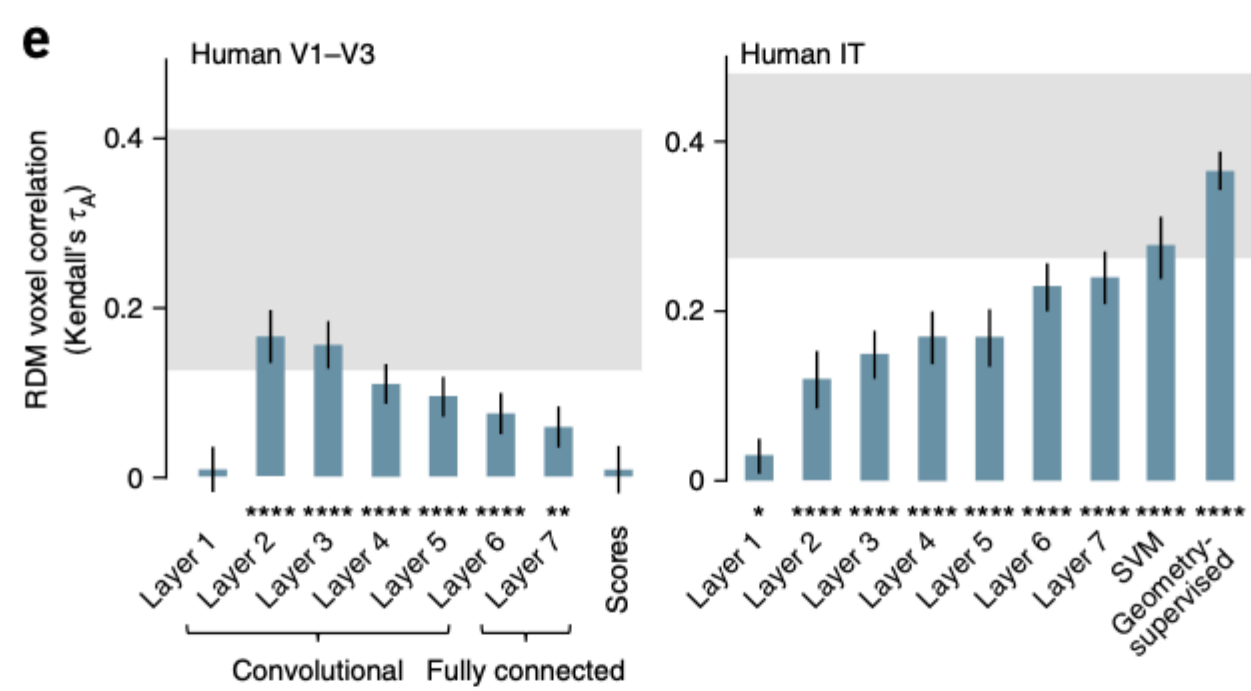
Single neurons



Internal representations

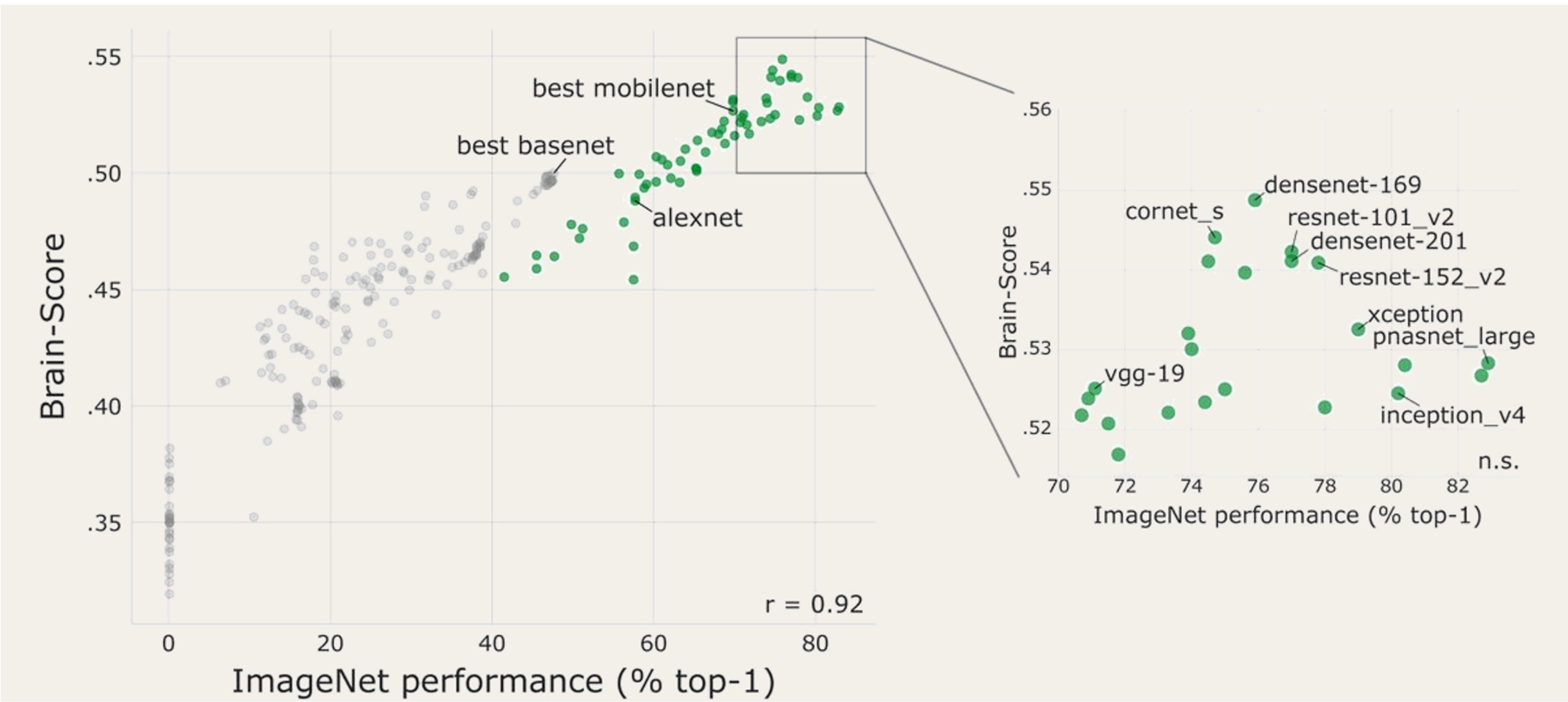


Comparison between different layers



Brain score

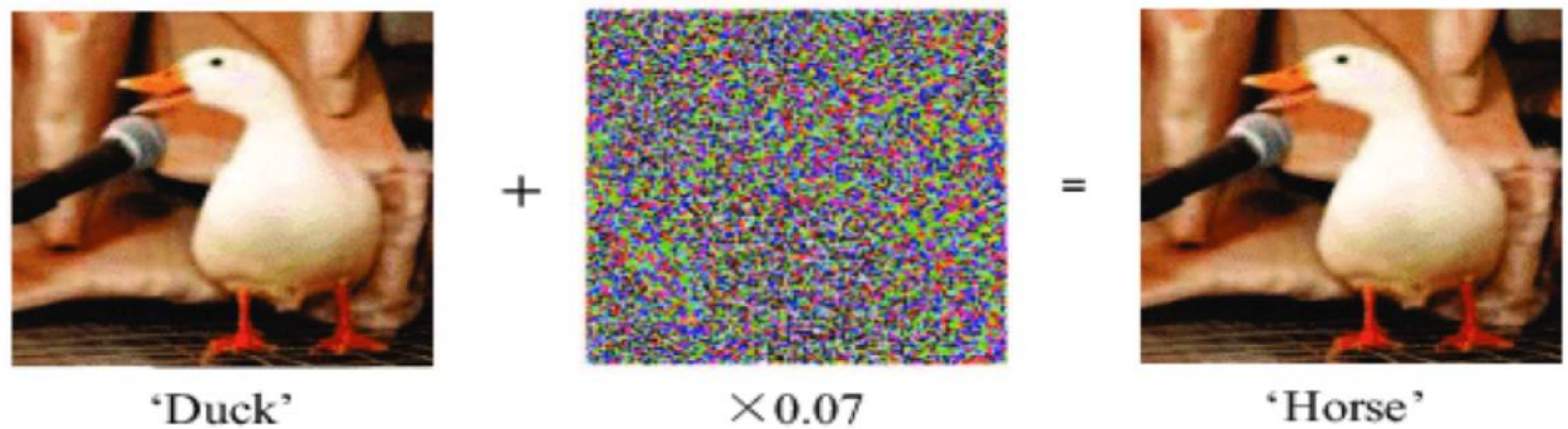
- An overall benchmark for comparing ANNs and neurons (Yamins 2014)



- brain-score performance with ImageNet performance

Still some way to go…

- Adversarial attacks



- Texture versus object

- Robustness to different types of noise
- Invariances (e.g. object rotations)
- More importantly: what is inside the black box? Can we extract 'principles' of neural coding?

Summary

- **Efficient coding**
  - Redundancy reduction by neural populations
  - Efficient coding in the retina: redundancy reduction vs noise cancellation
  - Independent/sparse coding and primary visual cortex
  - Divisive normalisation
- **Information bottleneck** and predictive coding
- **Bayesian and probabilistic models**
- **Deep learning**