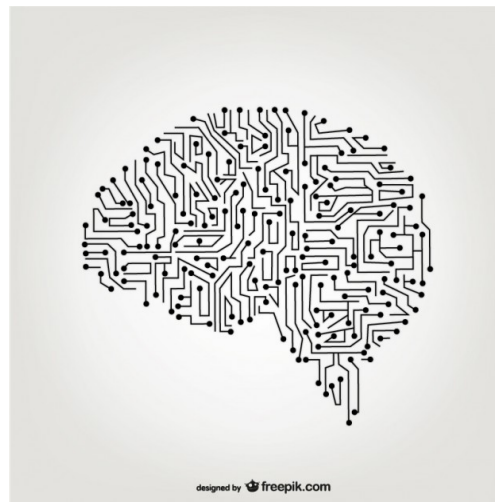

Efficient Coding

Odelia Schwartz

2020

Levels of modeling

- Descriptive (what)
- Mechanistic (how)
- Interpretive (why)



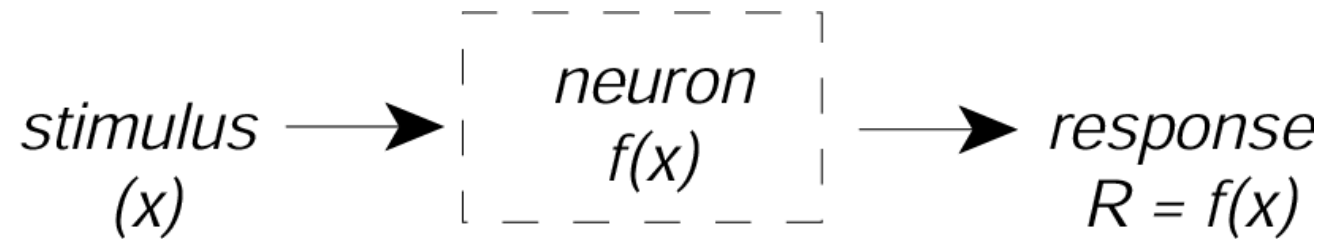
Levels of modeling

- Fitting a receptive field model to experimental data (e.g., using spike-triggered stimuli; you've seen)

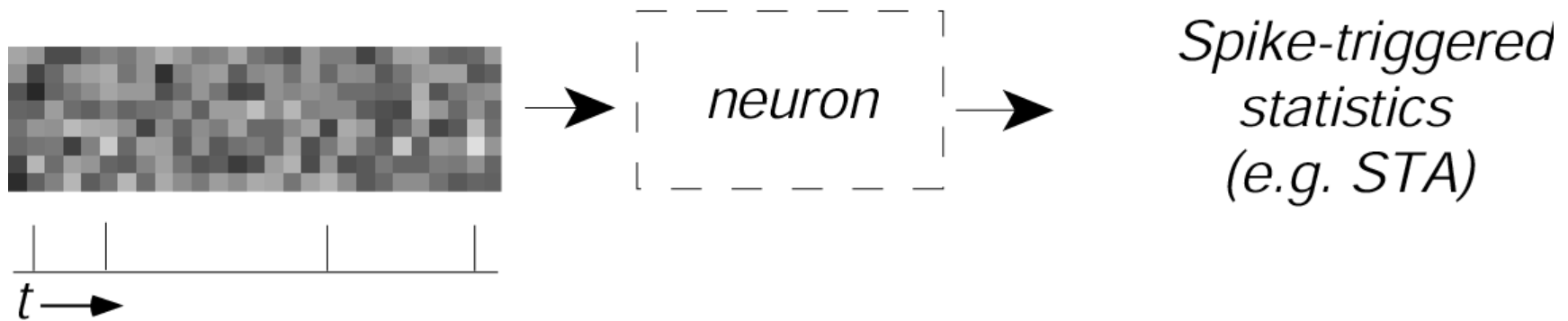
Versus

- Deriving receptive field model based on theoretical principles (e.g., statistical structure of scenes)

Fitting a model to data



Fitting a model to data

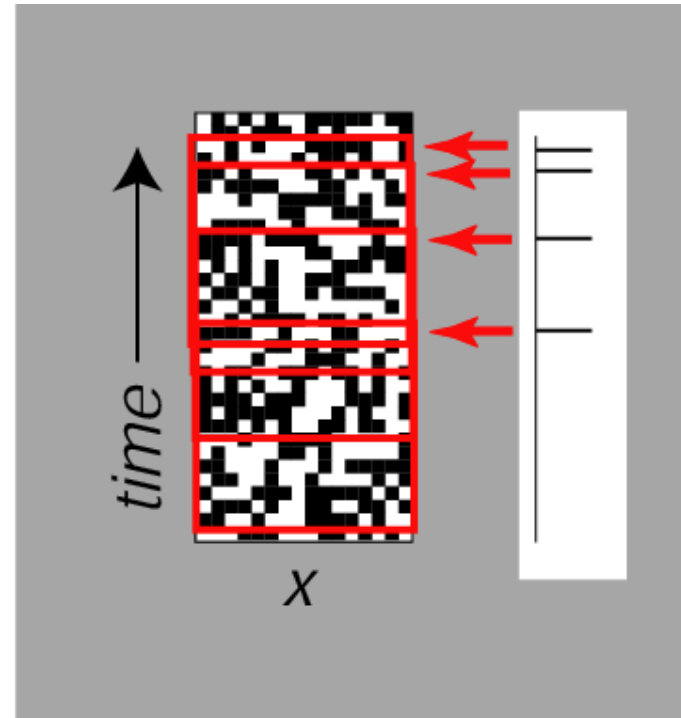


Primary visual cortex

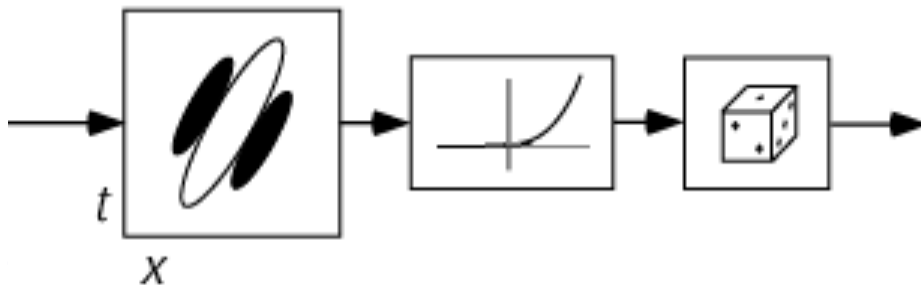
STA



Average of
spike-triggered
stimuli



Model:



Primary visual cortex

Hubel and Wiesel, 1959



Stimuli

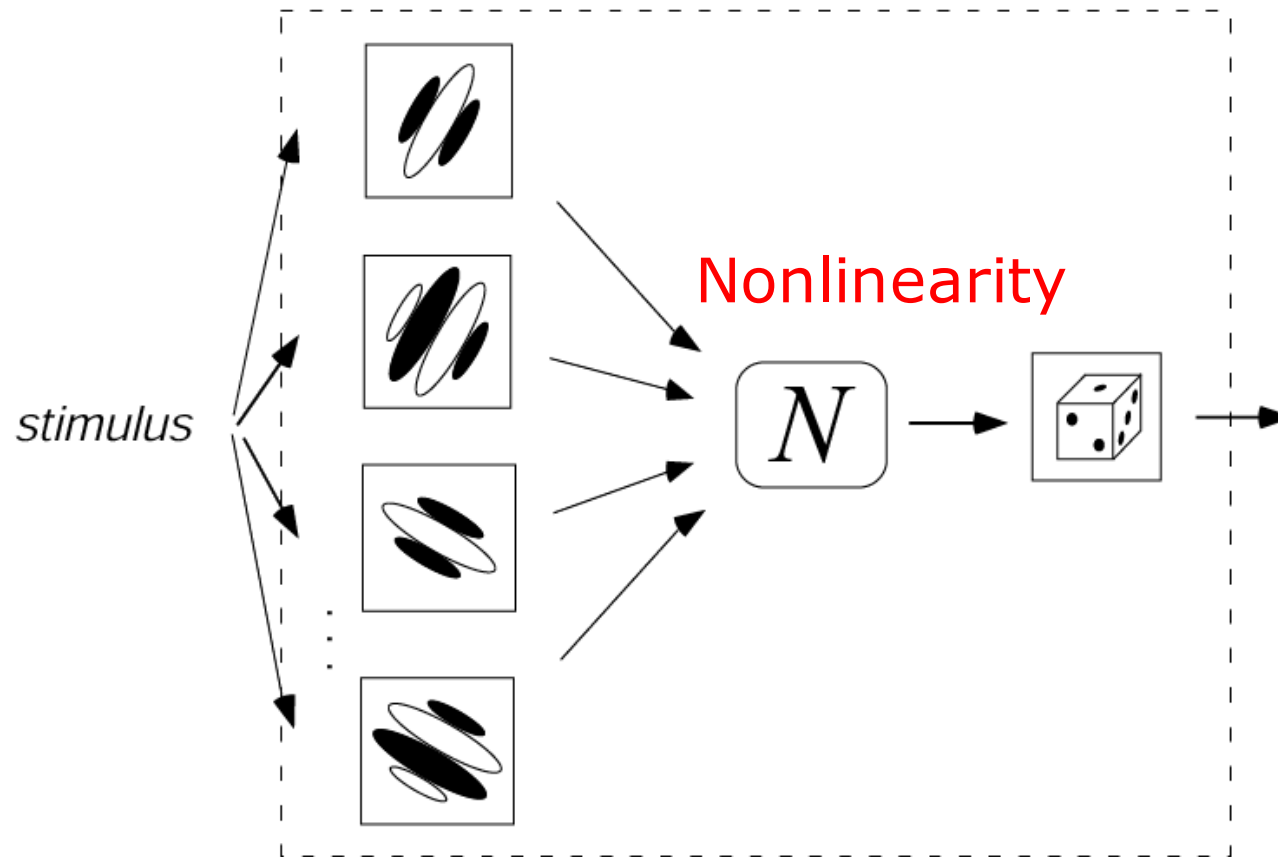
Spikes



Receptive field (filter)



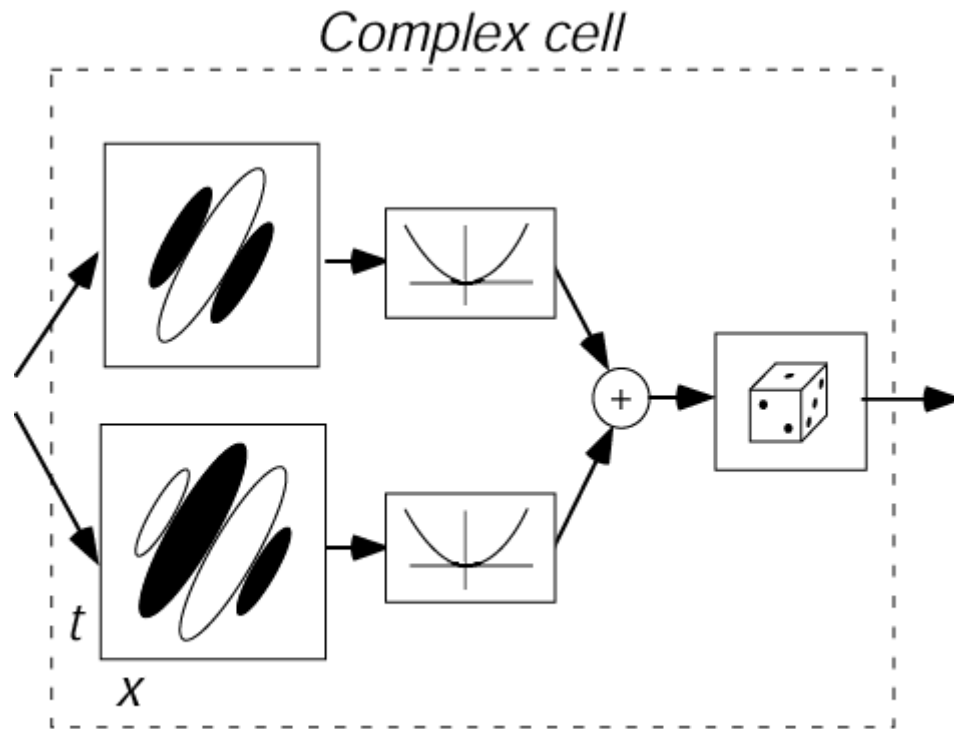
Fitting a model to spike data



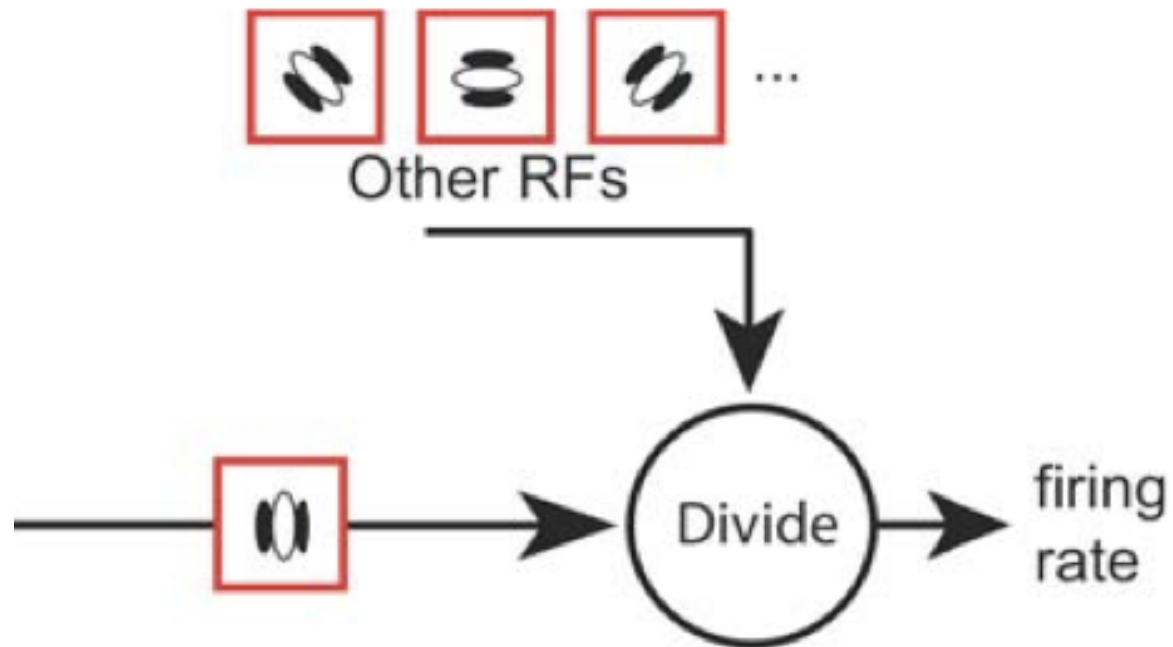
Fitting a model to spike data

What kind of nonlinearities?

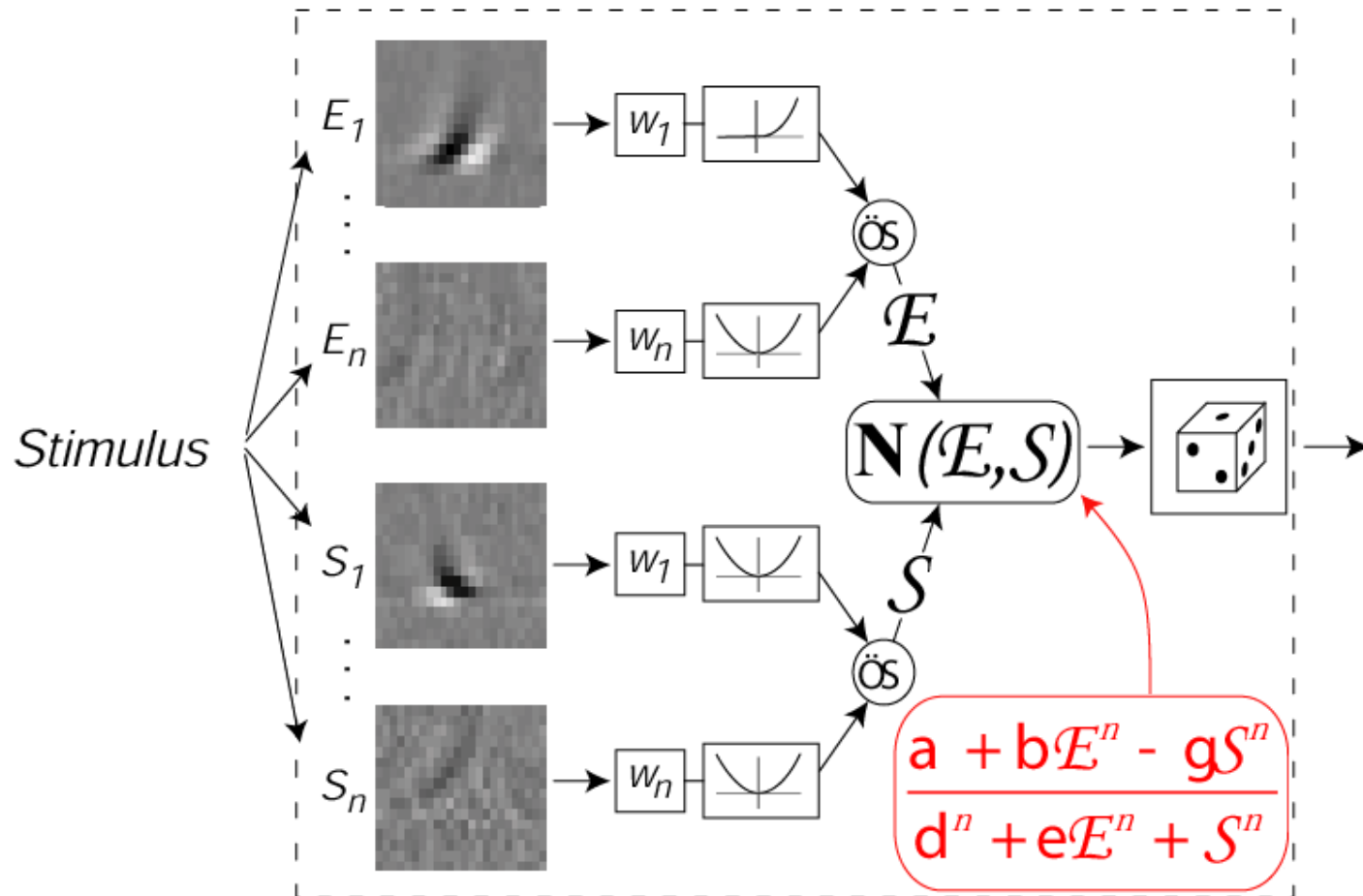
Complex cell



Divisive normalization



Fitting a model to spike data



Levels of modeling

- Fitting a receptive field model to experimental data (e.g., using spike-triggered stimuli)

Versus

- Deriving receptive field model based on theoretical principles (e.g., statistical structure of scenes)—adding an interpretive layer.

Complementary question



Can we derive or constrain a neural model by understanding statistical regularities in scenes?

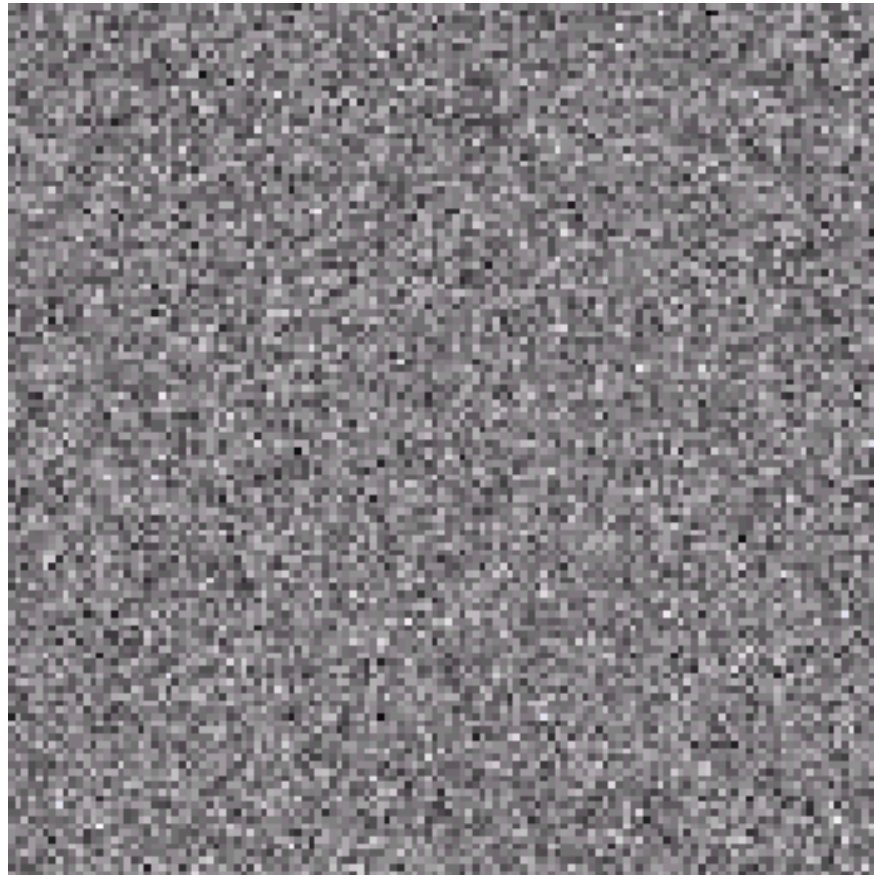
Complementary question



16

Appealing hypothesis: brain evolved to capture probabilistic aspects of the natural environment

Unlike...





18

Also quite different from traditional experimental stimulus in vision experiments...

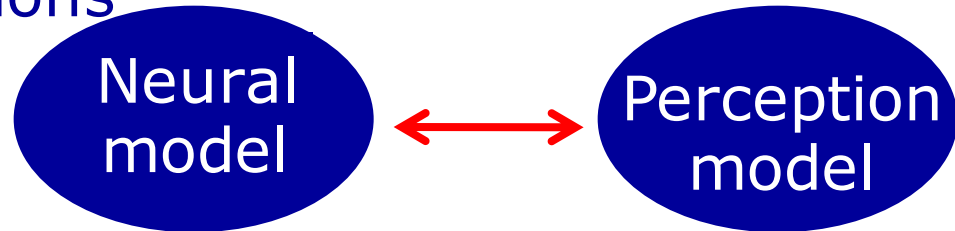
Visual representation



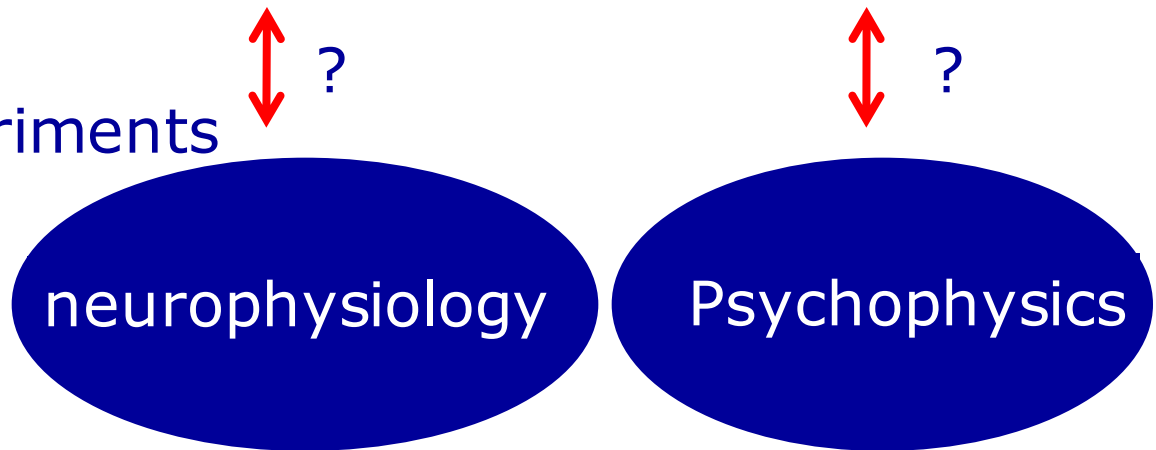
Visual input



Simulations

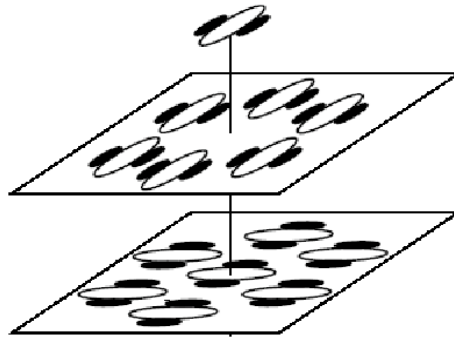


Experiments



- Goals: Principled and predictive understanding

Building model from scene stats



- Cortical computations as interactions of RFs across space, orientation, etc.
- RFs and interaction constrained by scene statistics? Can we derive them?

Theory

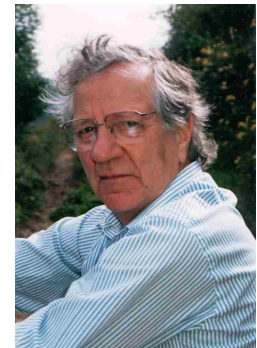
- **Locke:** The mind is a “tabula rasa” and only filled with knowledge after sense experience



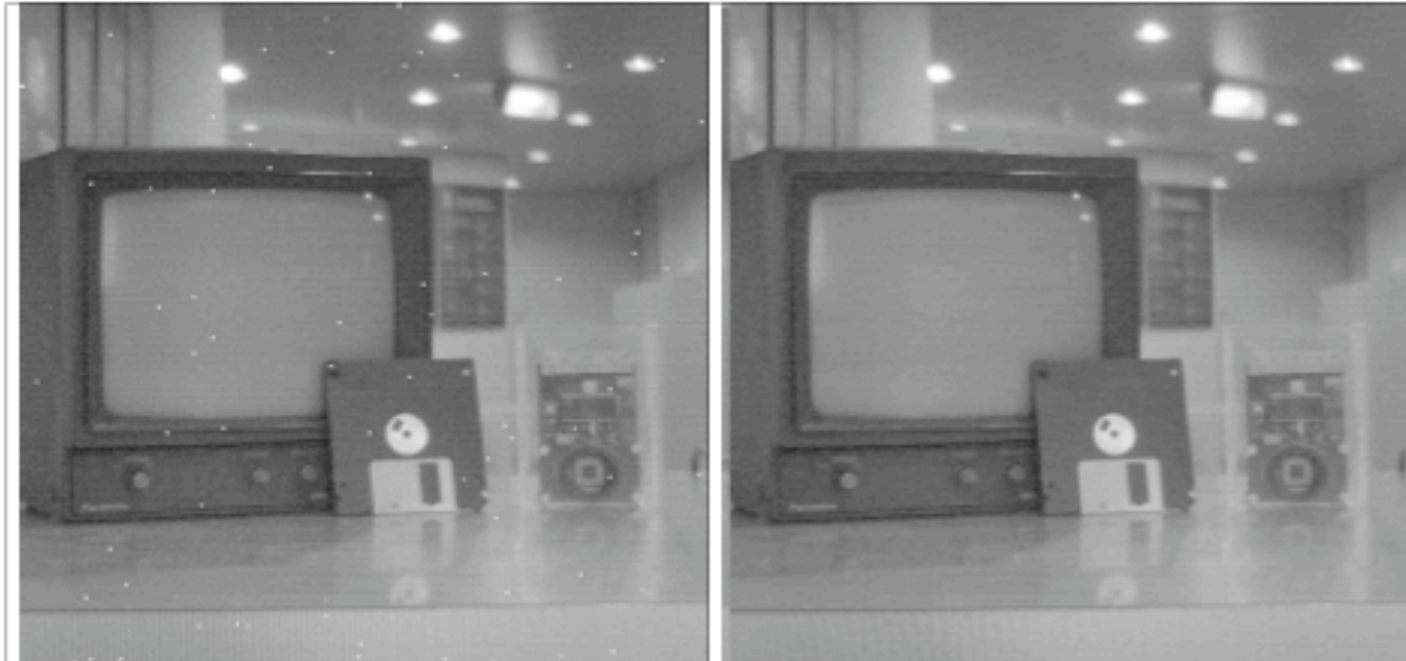
- **Helmholtz:** perception as inference of the properties of sensory stimuli



- **Attneave, Barlow:** Hypothesized in the 1950s that sensory processing matched to statistics of environment (reduce redundancy; increase independence)



Images are spatially redundant

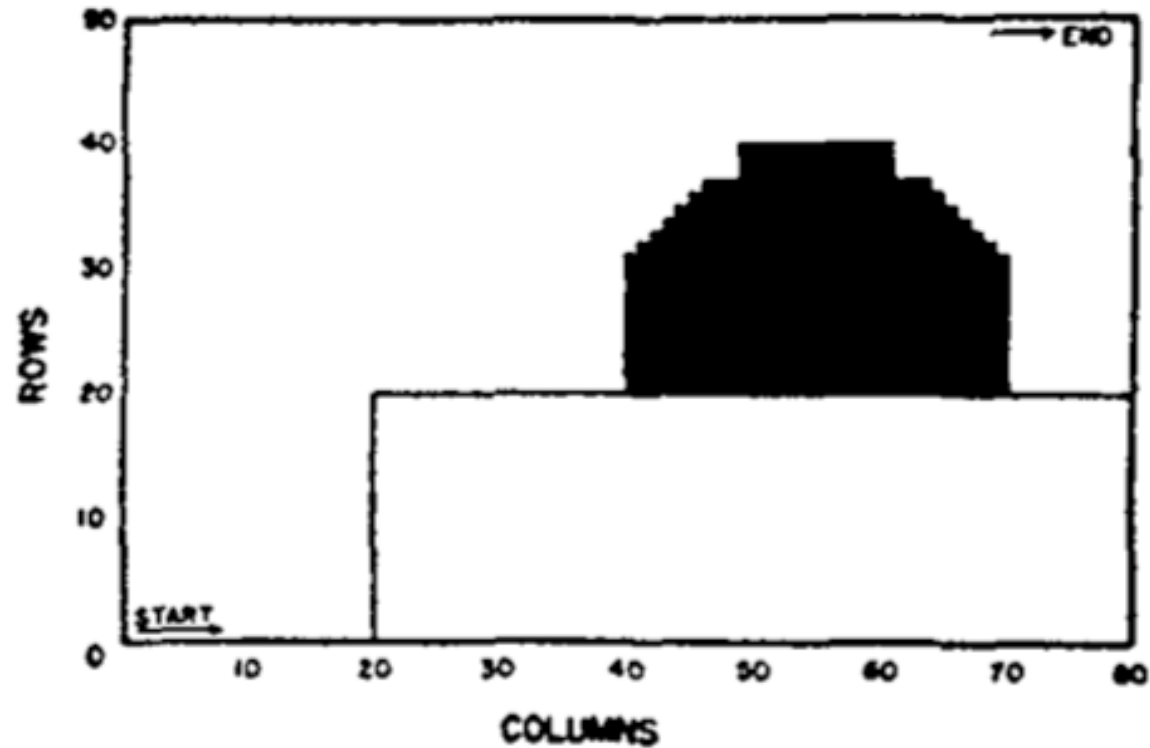


Kersten, 1992 (psychophysics);

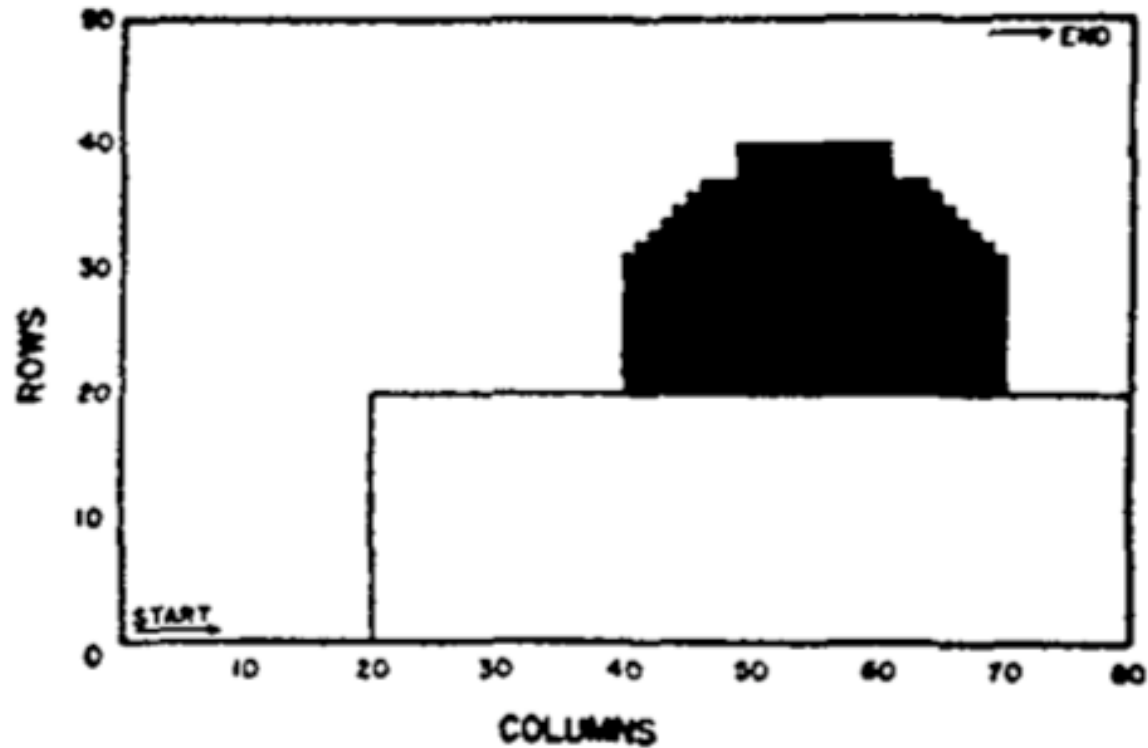
23

Dierickx and Meynants, 1987 (computer)

Images are spatially redundant



Images are spatially redundant

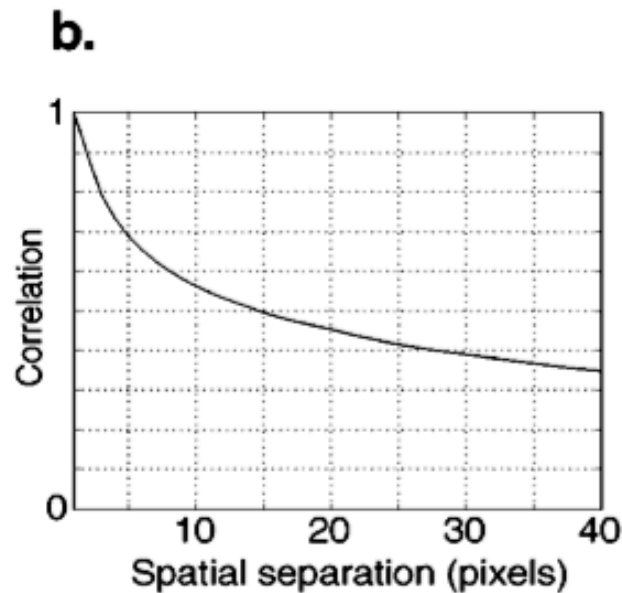
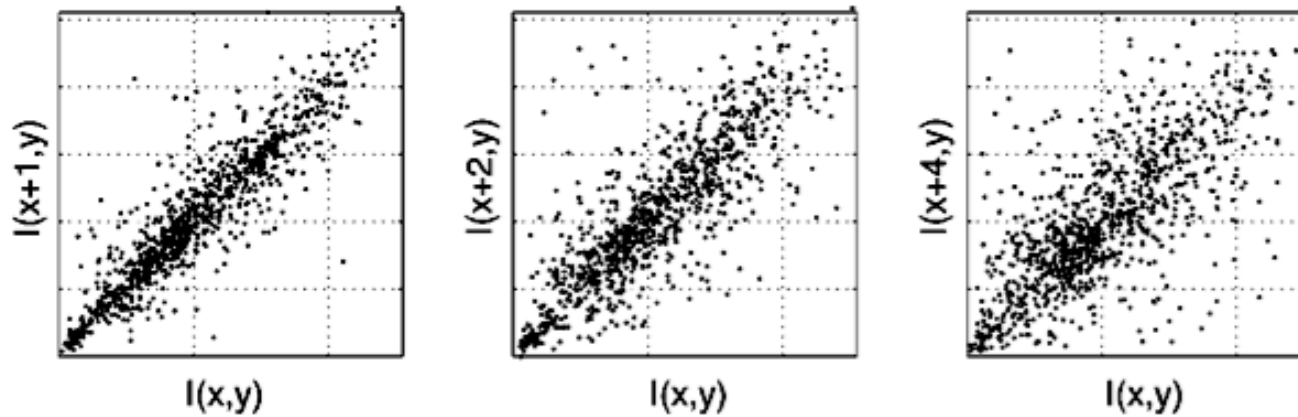


25 Attneave 1951; “ink bottle on the corner of the desk”

Images are spatially redundant



Statistics of images show dependencies



Scene statistics approaches

Two main approaches for studying scene statistics

1. Bottom-up

2. Top-down, generative

Scene statistics approaches

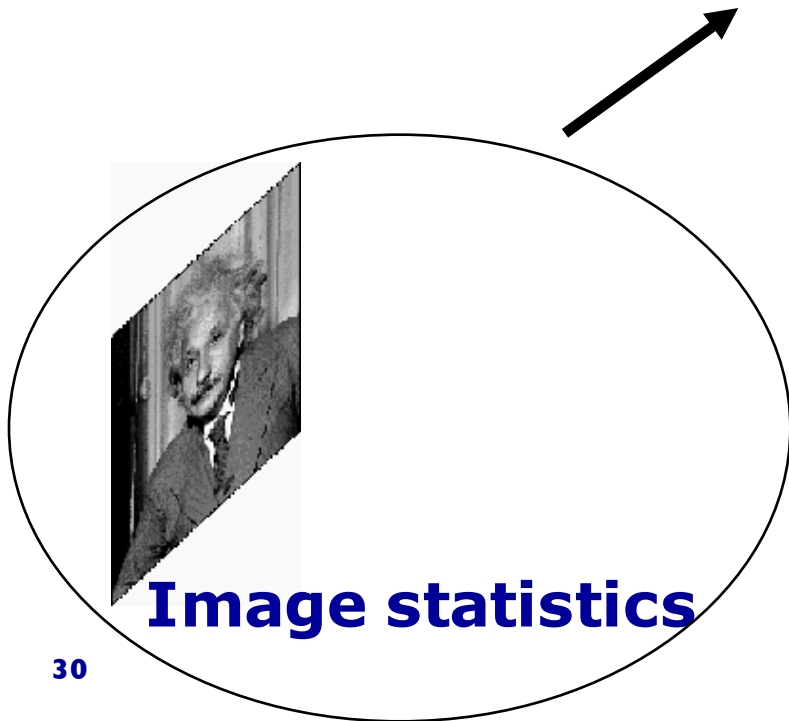
Two main approaches for studying scene statistics

1. Bottom-up (This class!)

2. Top-down, generative

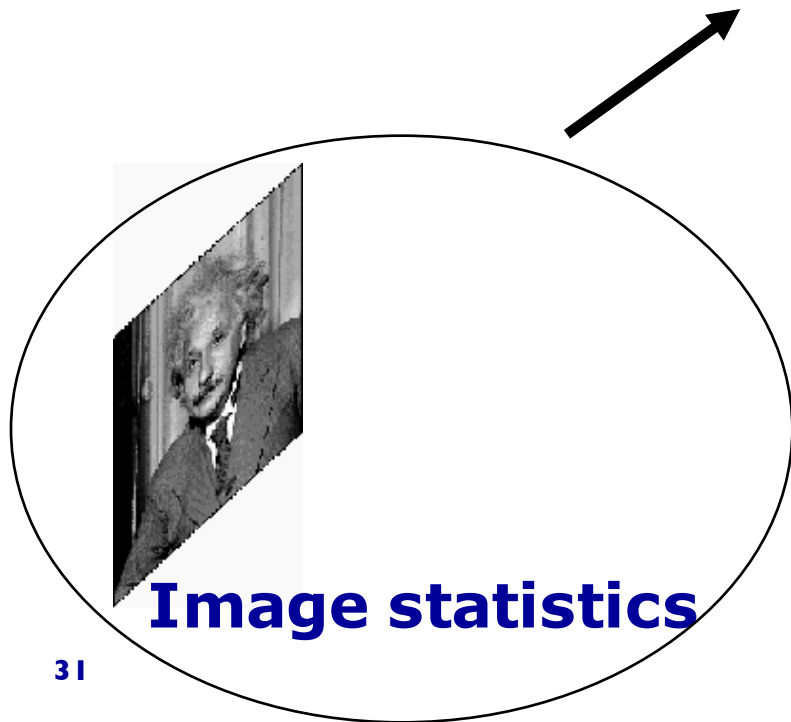
Bottom-up approach

Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics



Bottom-up approach

We'll take a small detour and talk about Information Theory...



Intro Information Theory

Redundancy:

- Marginal distribution (eg., in English “a” more often than “q”)
- Joint distribution (eg, ”sh” more often than “sd”)
- Analogous to images marginal and joint... (later)

Intro Information Theory

Redundancy and relation to coding in bits:

BABABABADABACAABAACABDAAAAABAAAAAAAADBCA

A → 00

B → 01 0100010001000100110001001000000100001000

C → 10 0111000000000001000000000000000011011000

D → 11

A → 0

B → 10 1001001001001110100110001000110010111000

C → 110 001000000000111101100

D → 111

Intro Information Theory

Entropy:

$$H(y) = -\sum_y p(y) \log_2 p(y)$$

- Measure of uncertainty or how interesting
- Always positive and equal to zero iff outcome is certain
- Log base 2 – expressed in bits
- Relates to minimal coding length

Intro Information Theory

Entropy:

$$H(y) = -\sum_y p(y) \log_2 p(y)$$

- Example: $P(A)=1/2$; $P(B)=1/4$; $P(C)=1/8$; $P(D)=1/8$
- Entropy = 1.75 bits (compared to 2 bits if all equal)

Intro Information Theory

Entropy:

$$H(y) = - \sum_y p(y) \log_2 p(y)$$

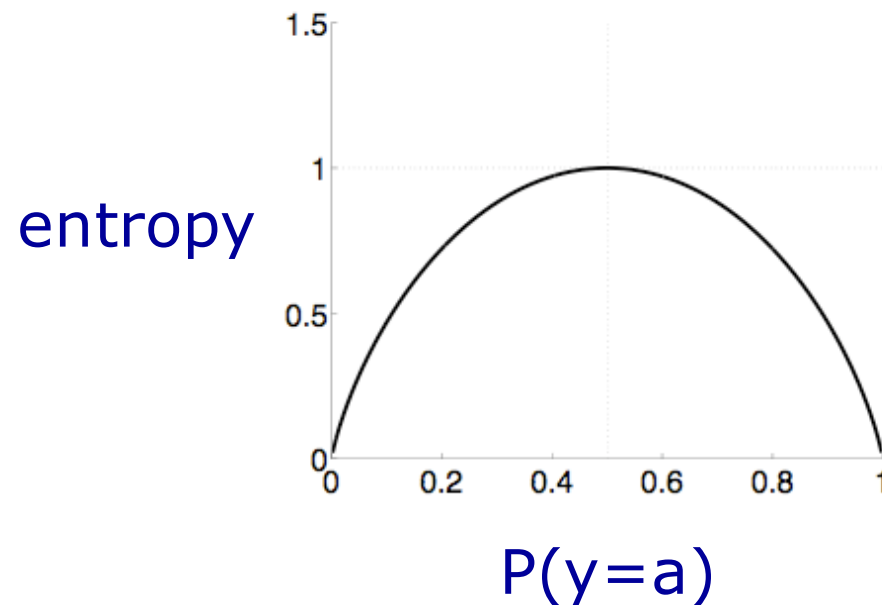
- If there are 2 possible outcomes with probability p and $1-p$, **when is the entropy maximal?**

Intro Information Theory

Entropy:

$$H(y) = -\sum_y p(y) \log_2 p(y)$$

Example: possible outcomes: a, b

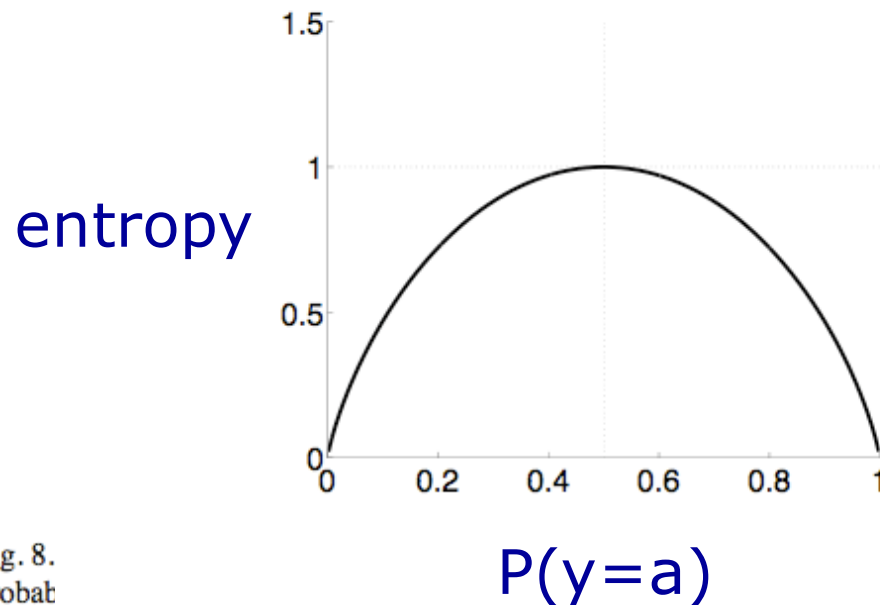


Intro Information Theory

Entropy:

$$H(y) = -\sum_y p(y) \log_2 p(y)$$

Example: possible outcomes: a, b



Maximum entropy
when most random
(0.5), or more generally
for uniform distribution

Fig. 8.
probat

Intro Information Theory

We've thus far looked at marginal distributions through one channel; we would like to also look at joint...

Conditional Entropy:

$$H(y | x) = - \sum_x p(x) \sum_y p(y | x) \log_2 p(y | x)$$

How much entropy left
in y when we know x

Intro Information Theory

We've thus far looked at marginal distributions through one channel; we would like to also look at joint...

Conditional Entropy:

$$H(y | x) = - \sum_x p(x) \sum_y p(y | x) \log_2 p(y | x)$$

averaged
over all x

How much entropy left
in y when we know x

Intro Information Theory

We've thus far looked at marginal distributions through one channel; we would like to also look at joint...

Conditional Entropy:

$$H(y | x) = - \sum_x p(x) \sum_y p(y | x) \log_2 p(y | x)$$

averaged over all x	How much entropy left in y when we know x
------------------------	--

- What happens when x and y are independent?
Dependent? Equal?

Intro Information Theory

Conditional Entropy:

$$H(y | x) = - \sum_x p(x) \sum_y p(y | x) \log_2 p(y | x)$$

- How much entropy left in y when we know x , averaged over all x
- What happens when x and y are independent? Dependent?

Independent: $H(y | x) = H(y)$

Dependent: $H(y | x) < H(y)$

Equal: $H(y | x) = 0$

Intro Information Theory

Mutual information:

$$I(x, y) = H(y) - H(y | x)$$

- What is the mutual information if x and y are independent?

Intro Information Theory

Mutual information:

$$I(x, y) = h(y) - h(y | x) = \dots$$

$$\sum_{x,y} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

- What is the mutual information if x and y are independent?
- Also Kullback-Leibler between...

Intro Information Theory

Marginal and joint entropy:

$$H(y_1, y_2, \dots, y_n) \leq H(y_1) + H(y_2) + \dots + H(y_n)$$

Equality iff independent: $p(y_1, y_2, \dots, y_n) = p(y_1)p(y_2)\dots p(y_n)$

Intro Information Theory

Marginal and joint entropy:

$$H(y_1, y_2, \dots, y_n) \leq H(y_1) + H(y_2) + \dots + H(y_n)$$

Equality iff independent: $p(y_1, y_2, \dots, y_n) = p(y_1)p(y_2)\dots p(y_n)$

Maximal entropy when:

- Outputs through a single channel as random as possible (but subject to constraints on channel)
- Independent. In general, hard to achieve. Restrict to, eg, linear.

Intro Information Theory

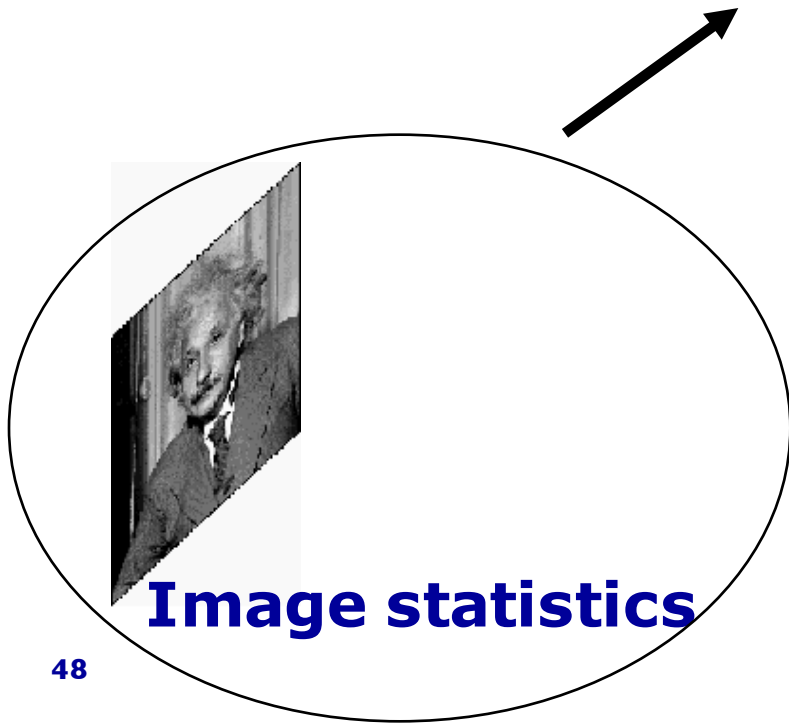
Redundancy; we can optimize...

- Marginal distribution (eg., in English “a” more often than “q”)
- Joint distribution (eg, ”sh” more often than “sd”)

What about images...?

Bottom-up approach

Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics



We'll go through past
examples in the field,
building up to more
recent approaches...

Bottom-up approach

Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics



Optimizing marginal
statistics

Efficient coding: single neuron fly vision

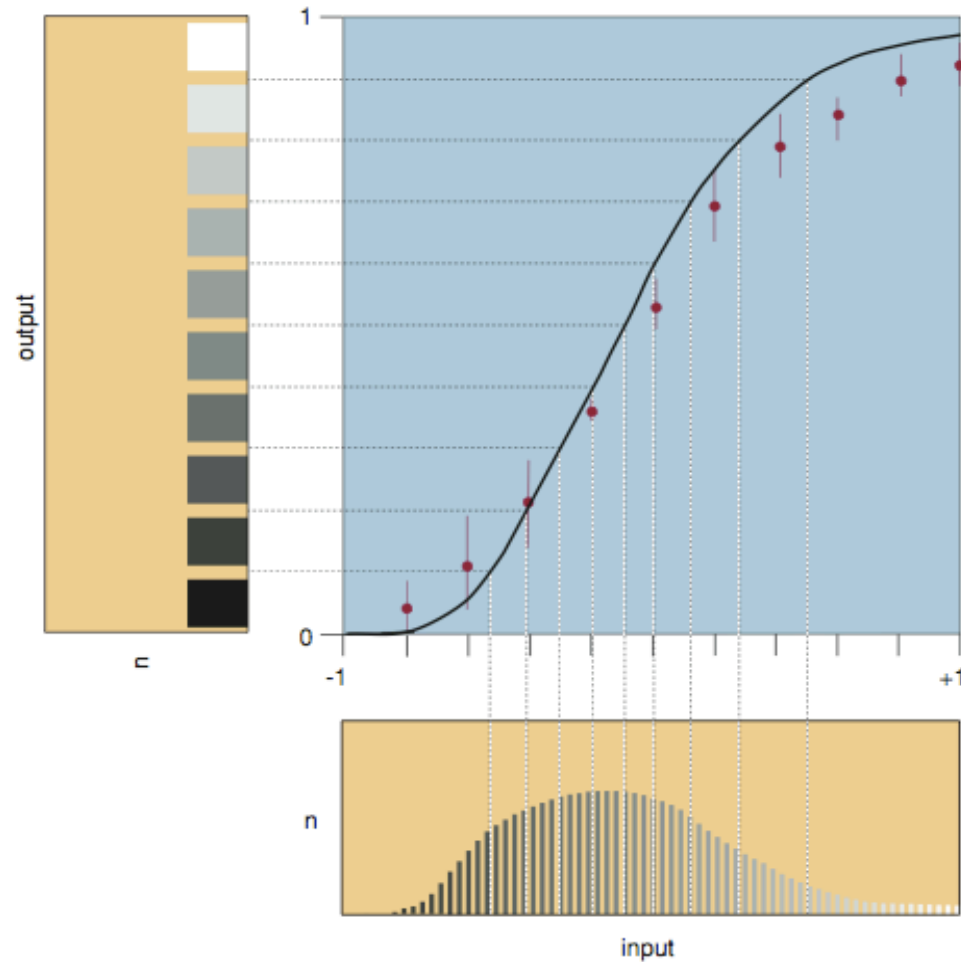
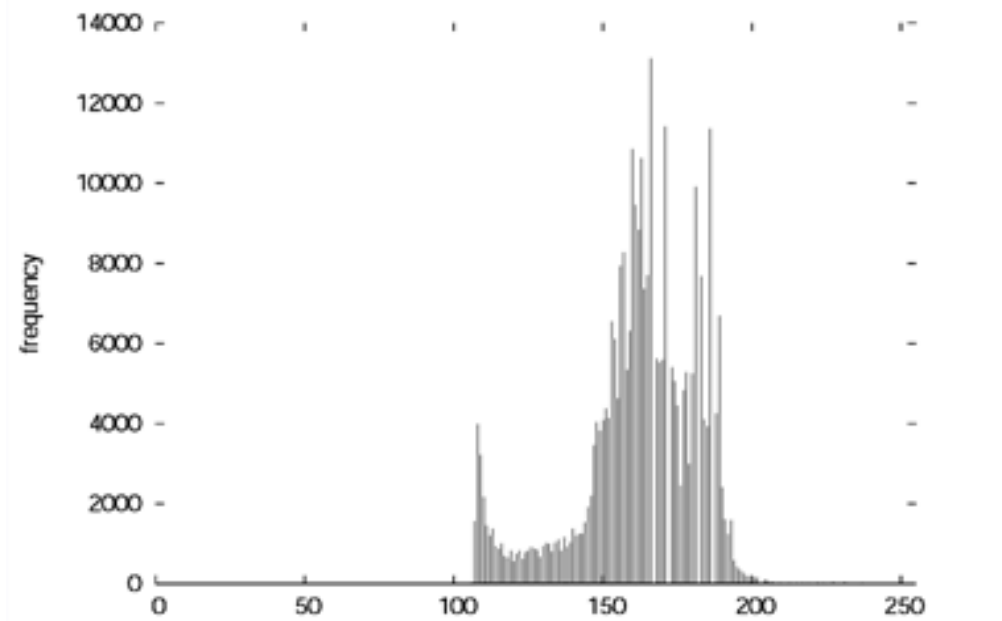
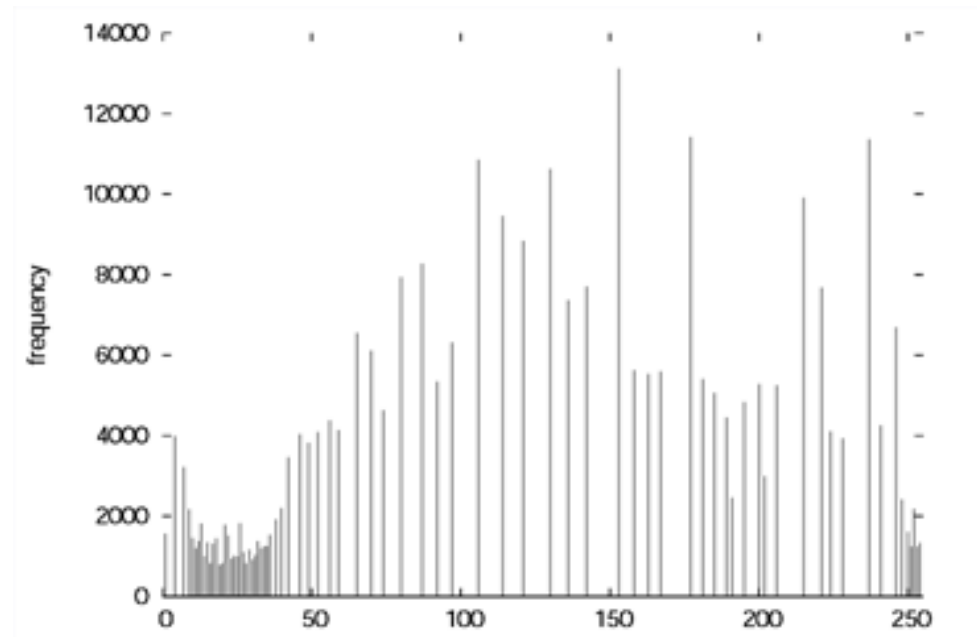


Figure from Olshausen & Field 2000; adapted from Laughlin 1981; Measured contrasts in natural scenes and showed that the membrane potential of fly visual neurons approximately transforms to uniform

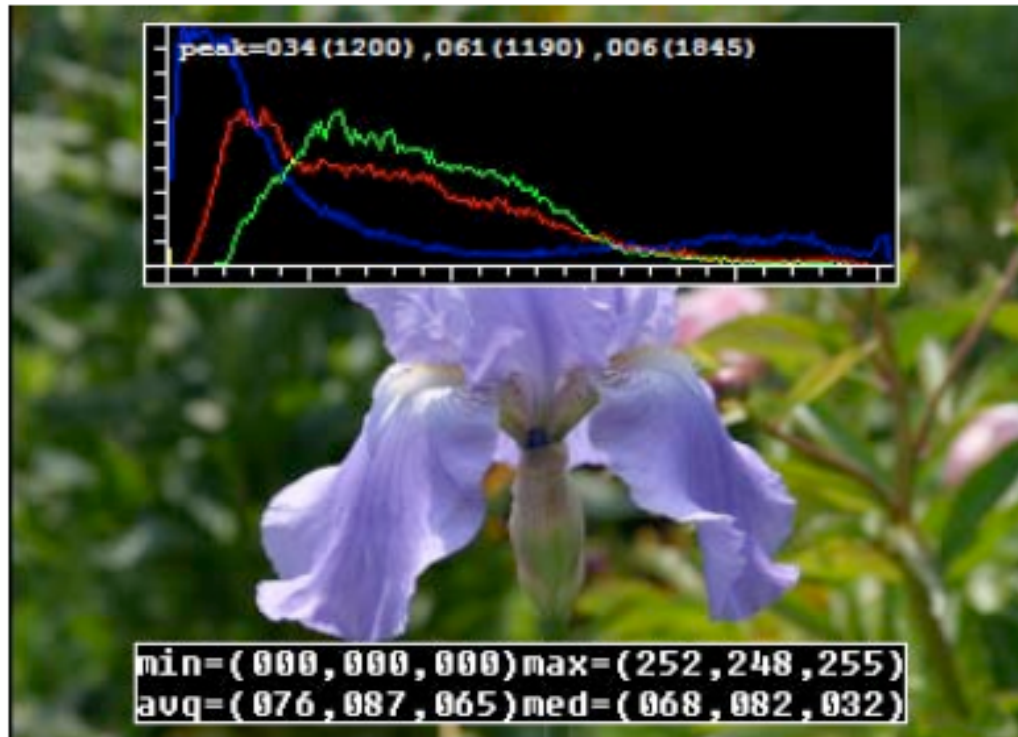
*Similar idea in image processing...
Histogram equalization*



*Similar idea in image processing...
Histogram equalization*

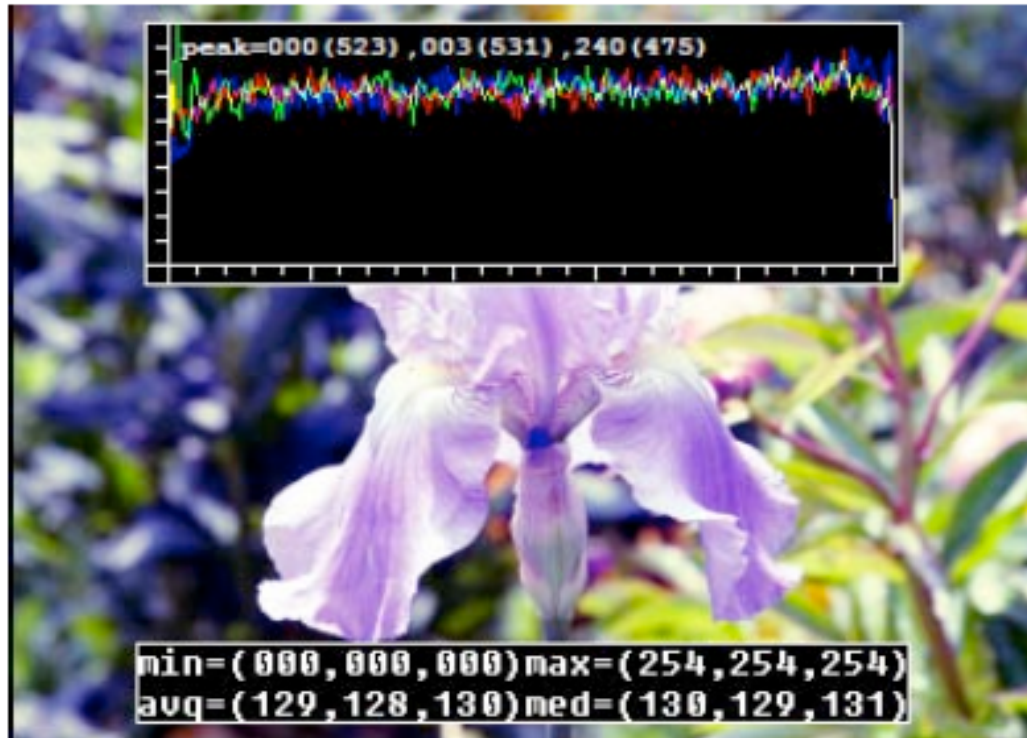


*Similar idea in image processing...
Histogram equalization*



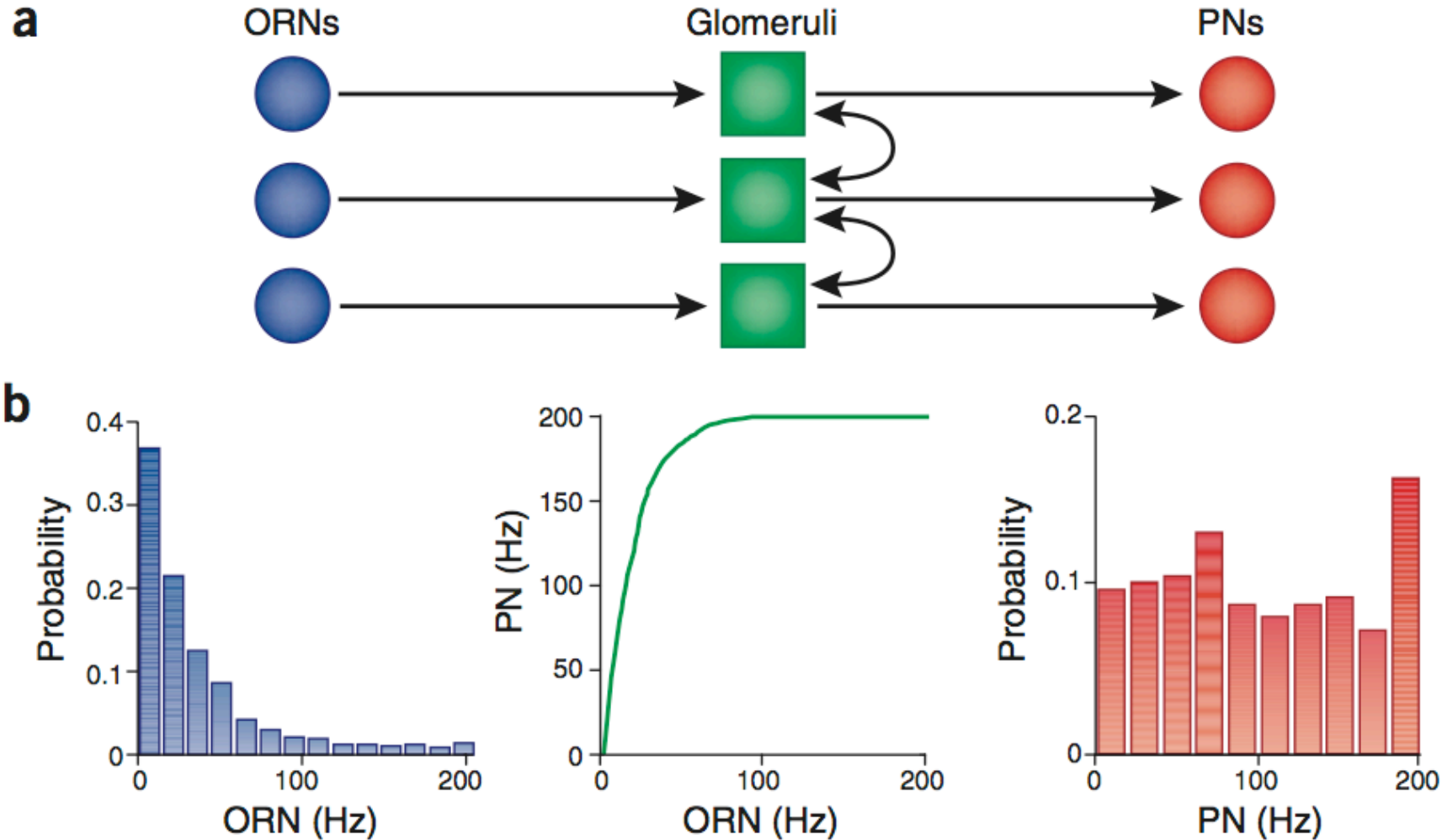
Richard Szeliski, Computer Vision Book 2010

*Similar idea in image processing...
Histogram equalization*



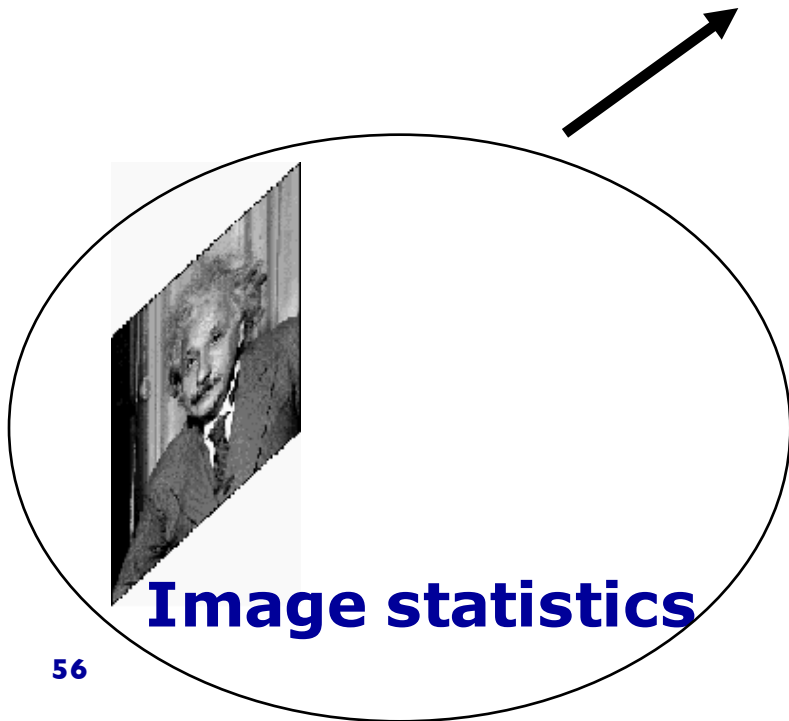
Richard Szeliski, Computer Vision Book 2010

Efficient coding: fly olfaction



Bottom-up approach

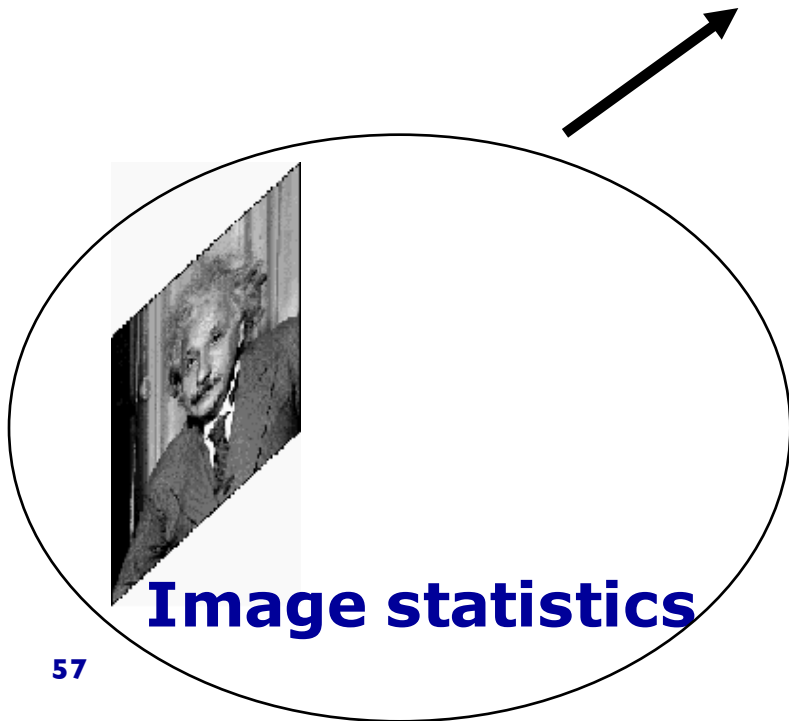
Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics



Assuming a **linear**
system and optimizing
joint statistics:
decorrelation

Bottom-up approach

Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics

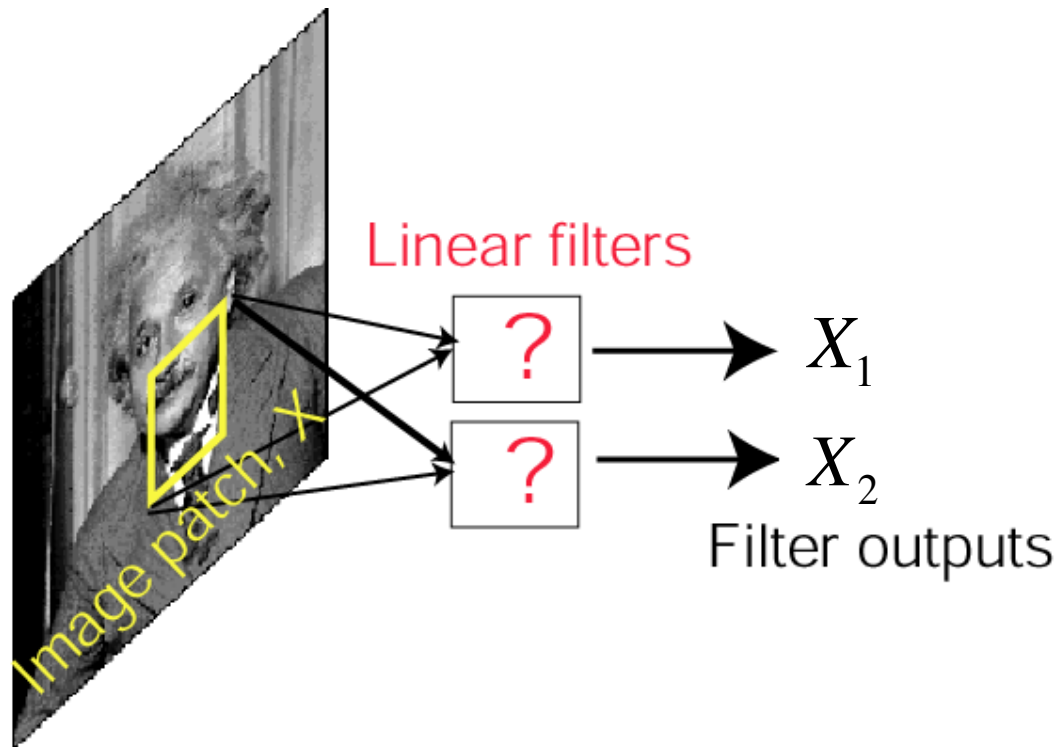


decorrelation

$$E[y_i y_j] = 0; i \neq j$$

**Does this guarantee
independence?**

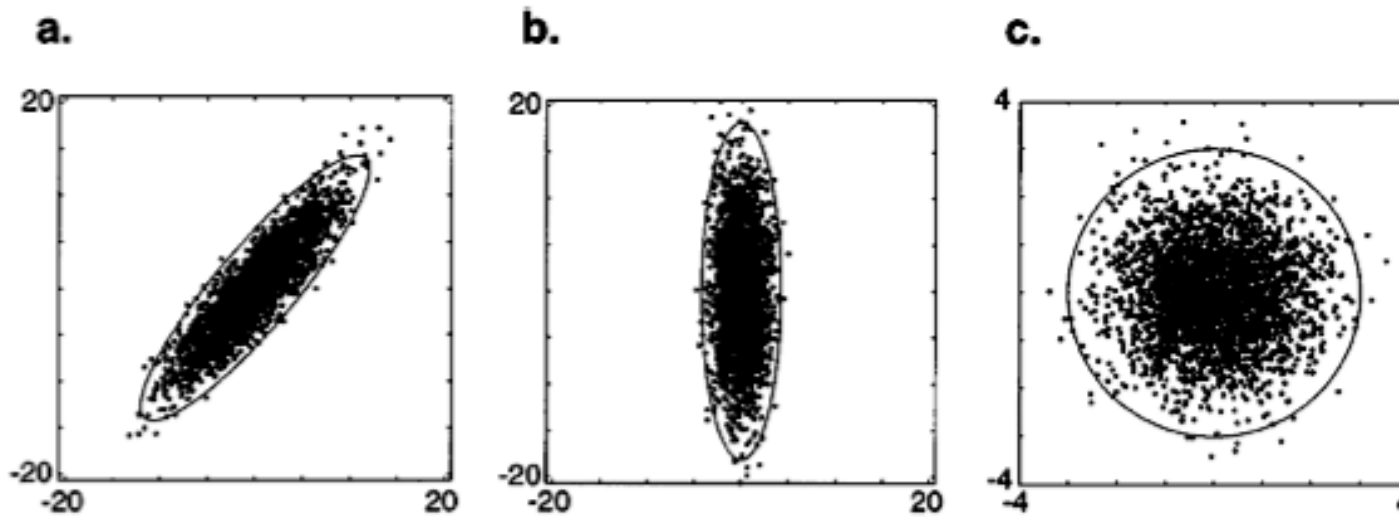
Linear Model: Theory

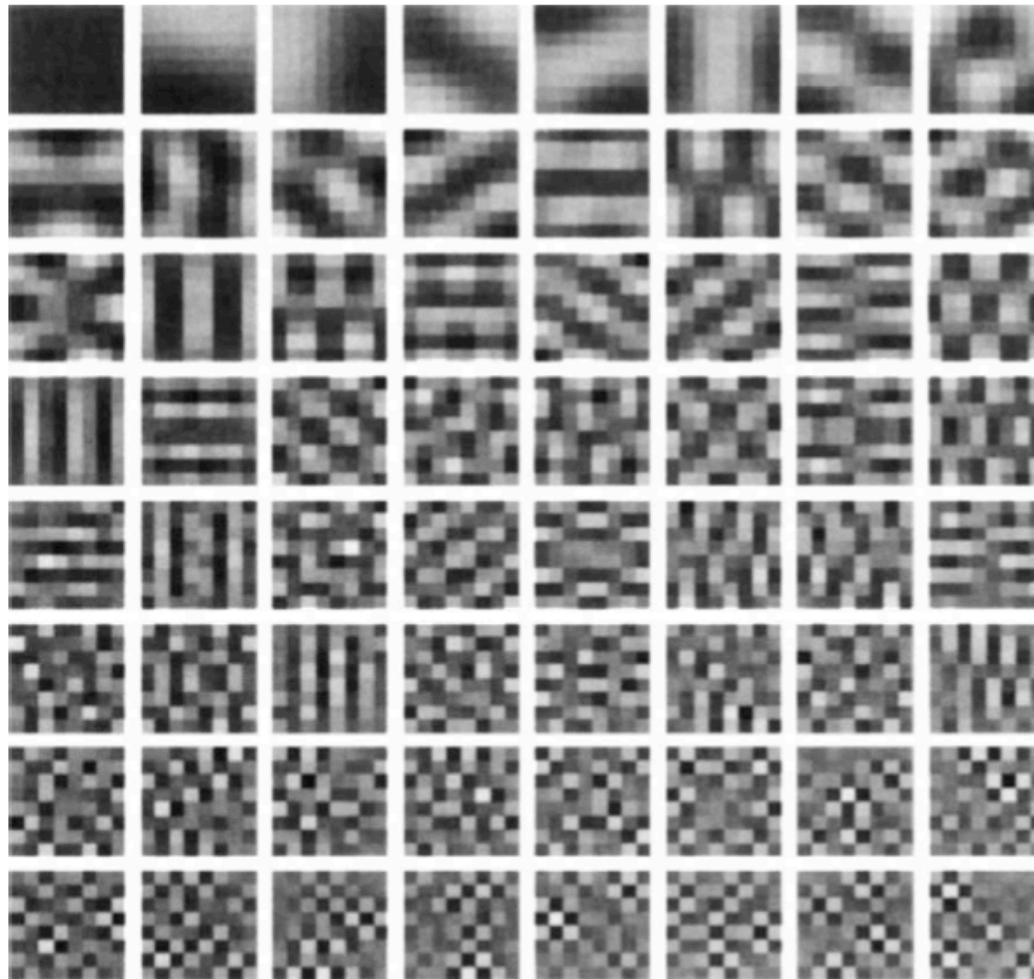


Find **linear filters** that **decorrelate** filter outputs to natural images

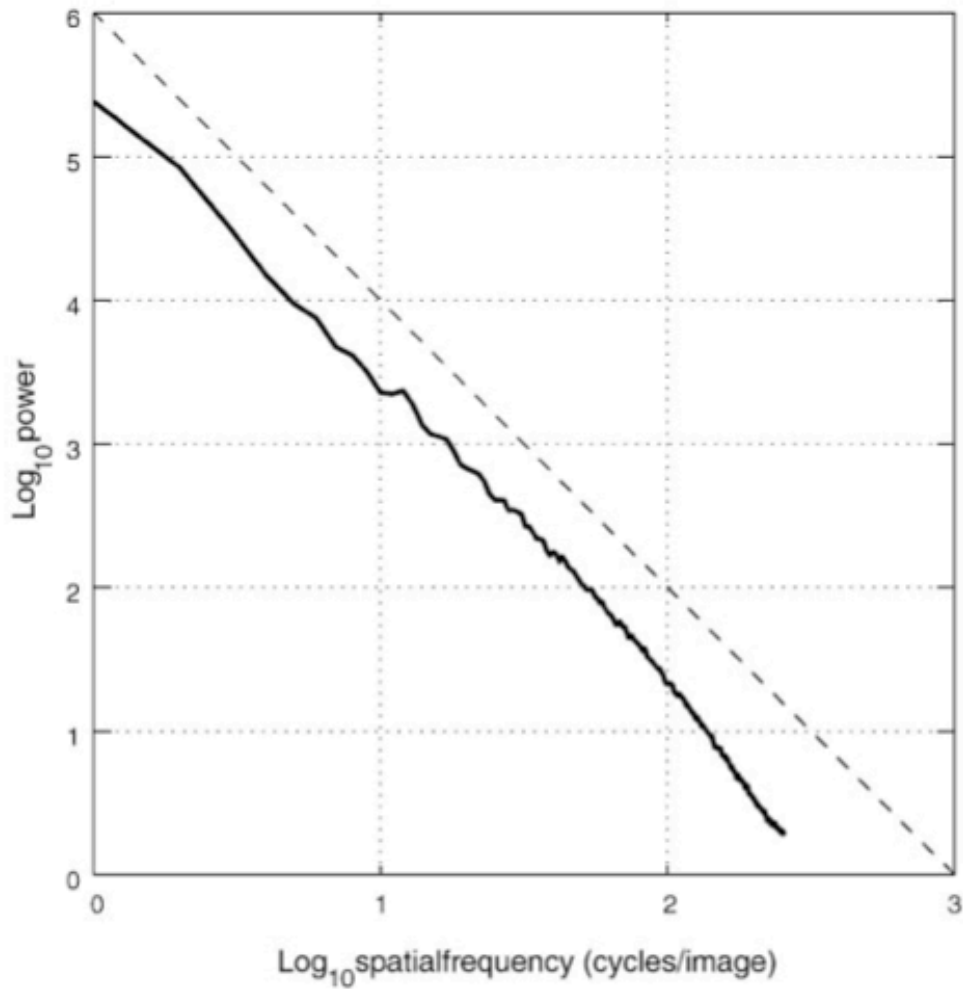
Geometric view (PCA)

Gaussian distribution



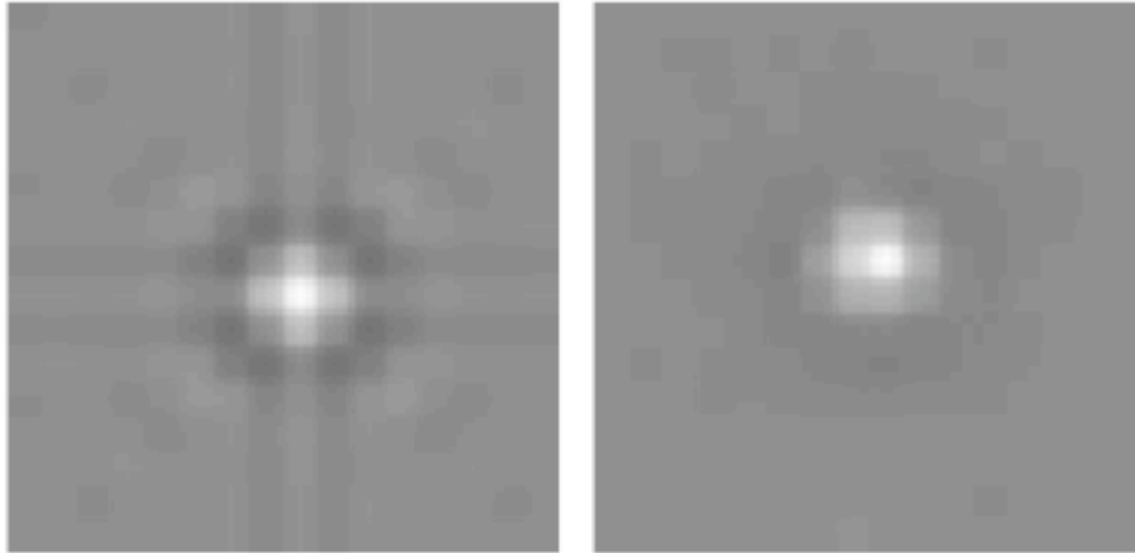


- Principal Component Analysis on image patches



- Power spectrum of natural images (from Simoncelli & Olshausen review)

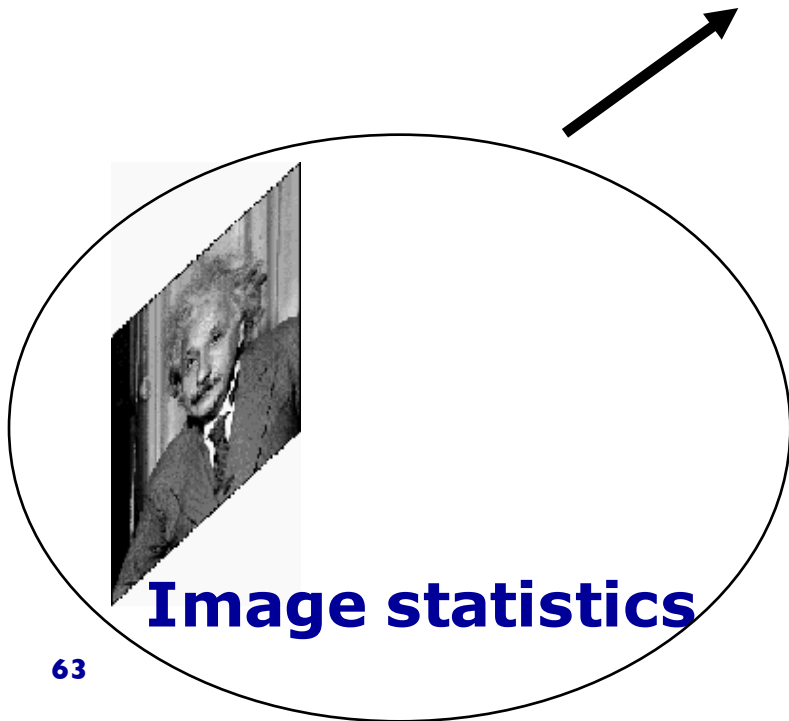
Bottom-up approach



- PCA and imposing extra constraints such as Spatially localized filters (from Hyvarinen book; see Atick & Redlich 1992; Zhaoping 2006)
- 62 • Remember decorrelated does not mean independent

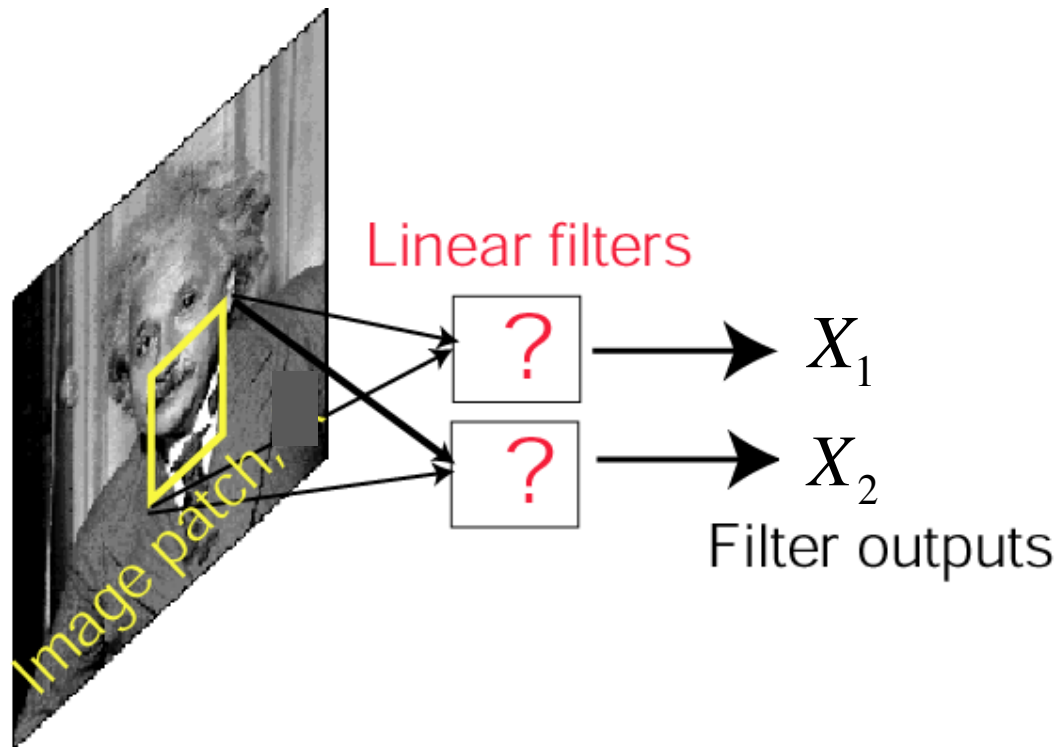
Bottom-up approach

Choose and manipulate projections,
to optimize probabilistic and
information-theoretic metrics



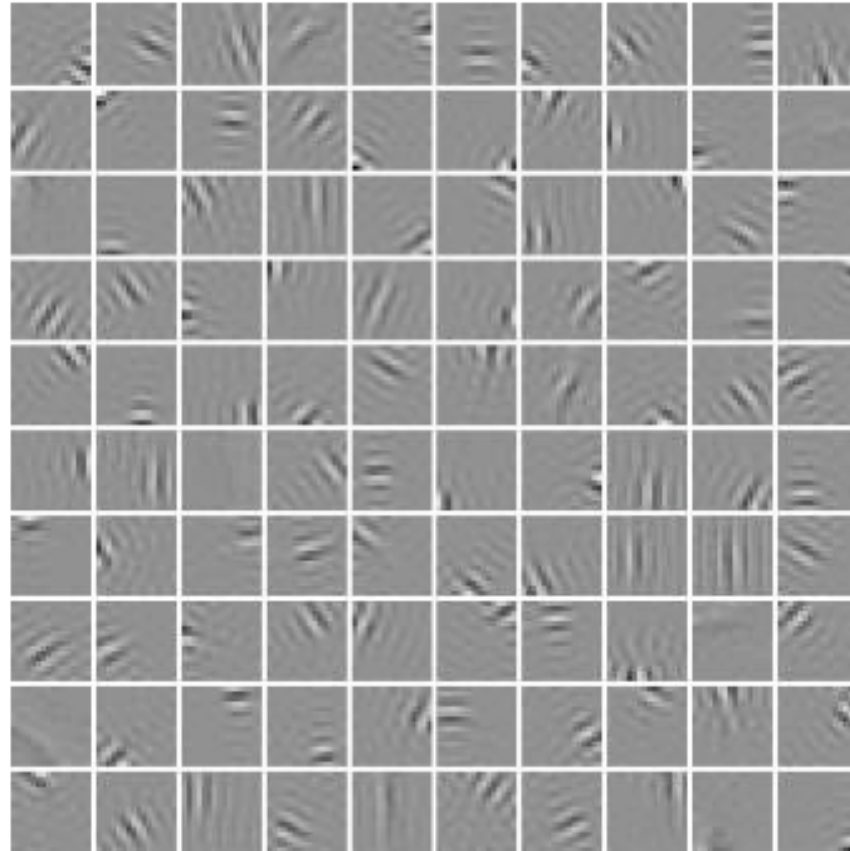
Assuming a **linear**
system and optimizing
joint statistics:
independence

Linear Model: Theory



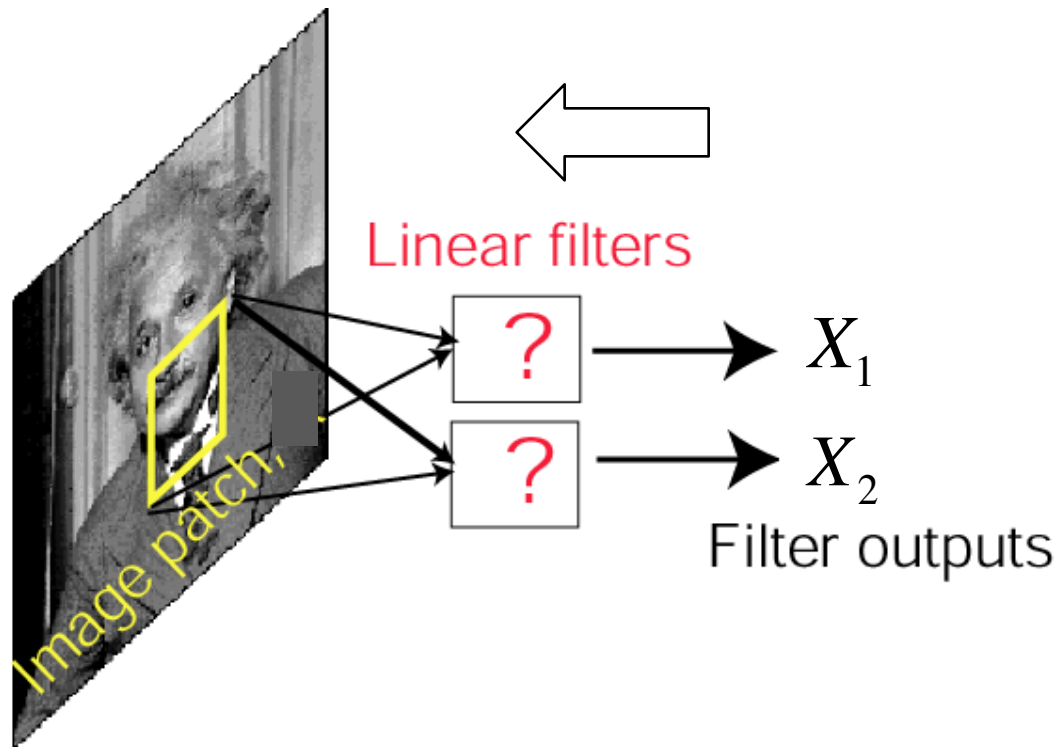
Find **linear filters** that maximize measure of statistical independence (or sparseness) between filter outputs to natural images (e.g., *Olshausen & Field, 1996*; *Bell & Sejnowski 1997*)

Linear Model: Theory



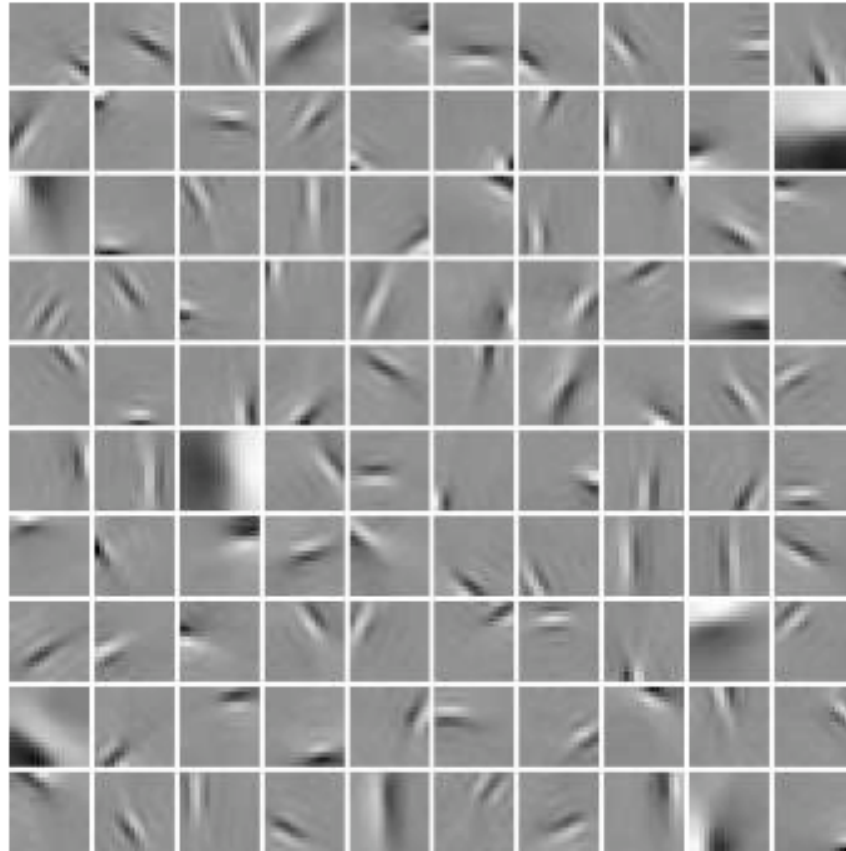
- ICA filters plotted from Hoyer images (e.g., *Olshausen & Field, 1996; Bell & Sejnowski 1997— here at Redwood!*)
- Qualitatively related to V1 RFs

Linear Model: Theory



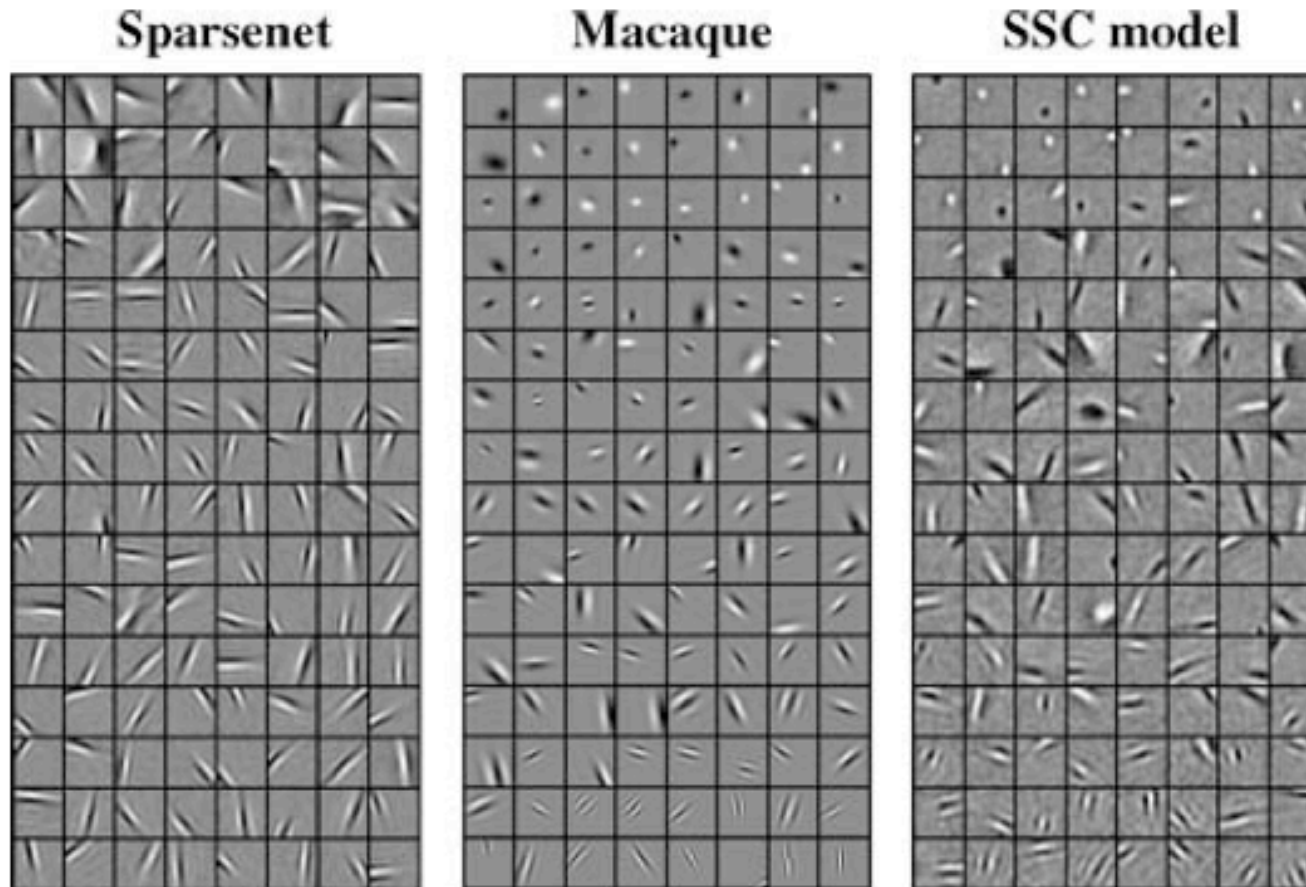
Linear transform, so from filter outputs can also go back to the image...

Linear Model: Theory



- ICA basis functions; from Hoyer
- Olshausen & Field, 1996; Bell & Sejnowski 1997

Linear Model: Theory

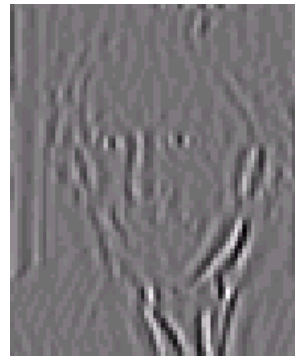
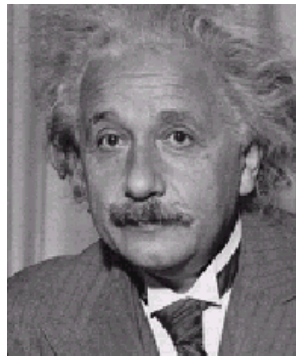


- Note also more recent work explaining neural diversity
- Rehn and Sommer, 2007 (data: Ringach)

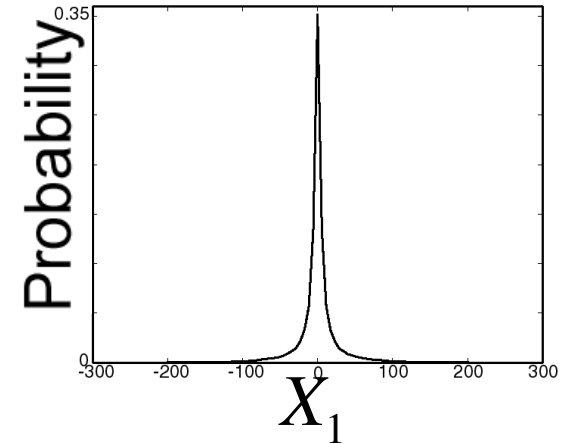
Bottom-up approach

What about sparse?
(e.g., Olshausen & Field)

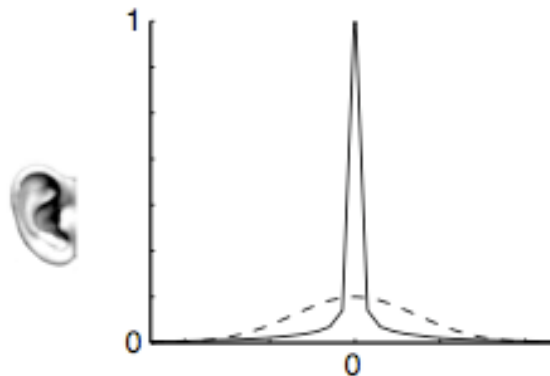
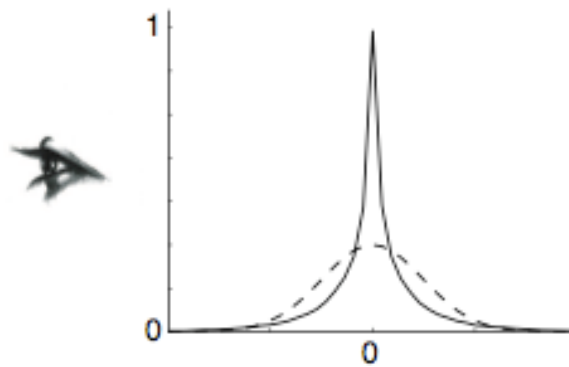
Bottom-up Statistics



X_1

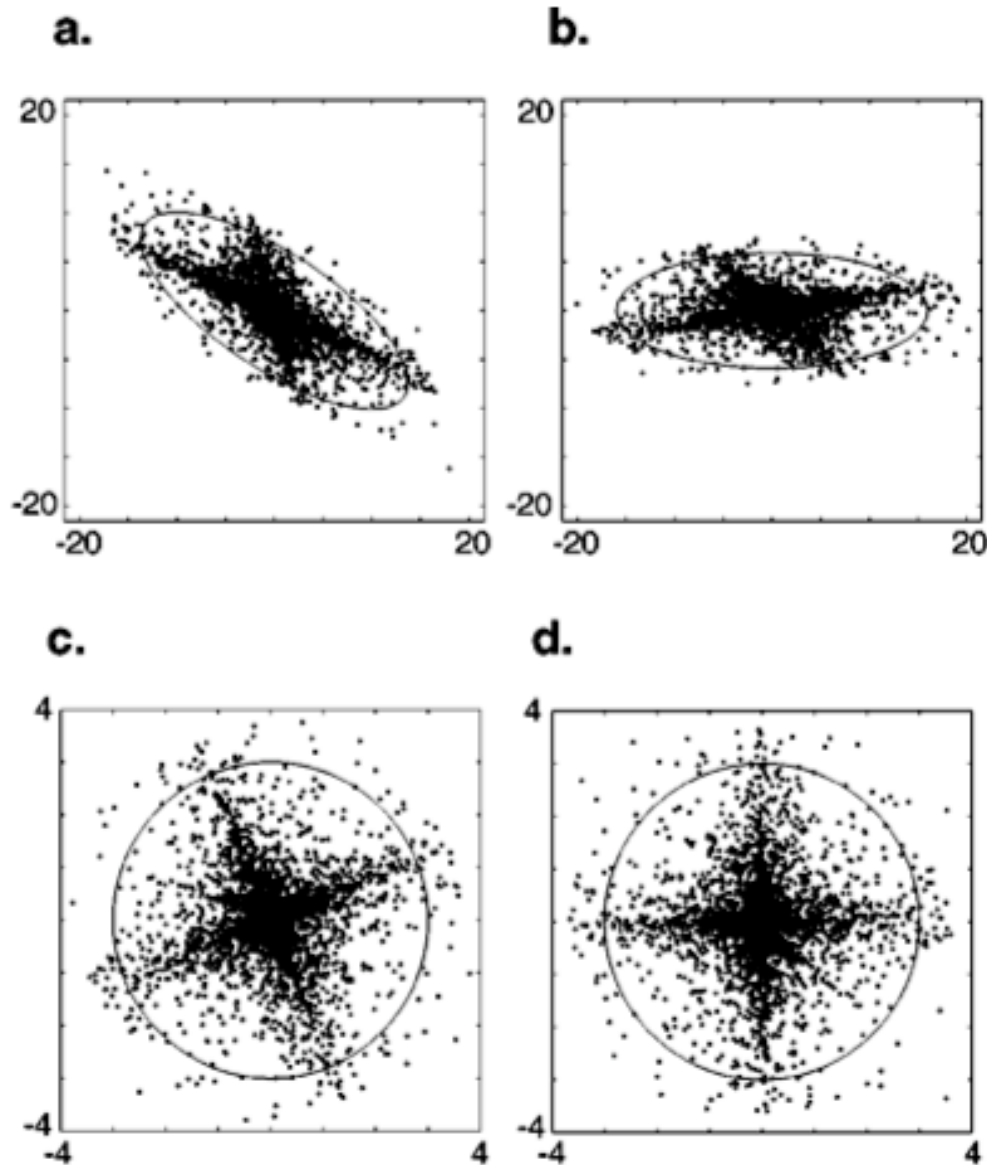


Bottom-up Statistics



- Well described by, eg, generalized Gaussian distribution

Geometric view (ICA)



Non Gaussian
(sparse)
distribution

Simoncelli & Olshausen
review, 2001

ICA and sparse coding

- In ICA maximizing independence assuming a linear transform (e.g., by maximizing joint entropy of the output).
- But should also assume that the outputs have a sparse distribution...

Summary

- Different levels of modeling...
- We've considered bottom-up scene statistics, efficient coding, and relation of linear transforms to visual filters
- Efficient coding through one channel and multiple channels
- Can we propagate statistical principles (such as efficient coding?) and how far?
- Next class: nonlinearities