Last Lecture: Bayesian Statistics and Information Theory

Mar 8th, 2016

Lane McIntosh & Kiah Hardcastle

Math Tools for Neuroscience

Logistics

- Final presentations next week (3/15)
  - 5ish minutes
    - 1-2 min for intro
    - 1-2 min for methods and results
    - 1 min for discussion
- Final written projects due next week (3/15)
  - Guidelines on website
- Any missing homeworks you want us to grade due by 3/15
- Course evaluations

Goals for today

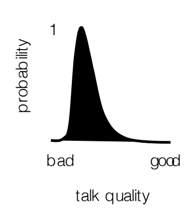
How to account for domain knowledge

How to quantify information transmission

Will the last lecture be any good?

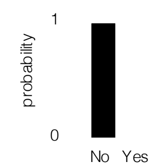
Prior

p(good talk)



Likelihood

p(Lane is talking | good talk)



Today's lecture

Bayesian Statistics

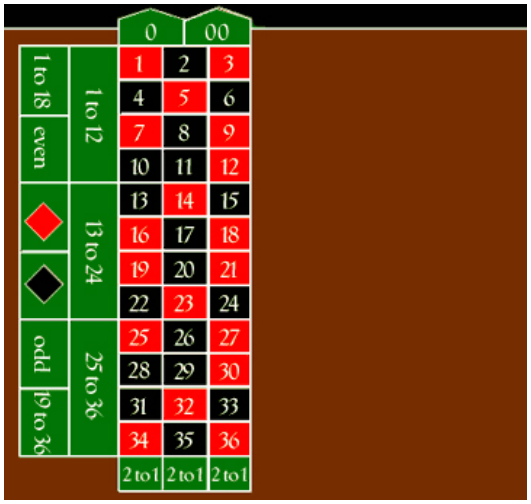
- priors, likelihoods, and posteriors
- Bayes rule
- Bayesian estimators
- uses of Bayesian statistics

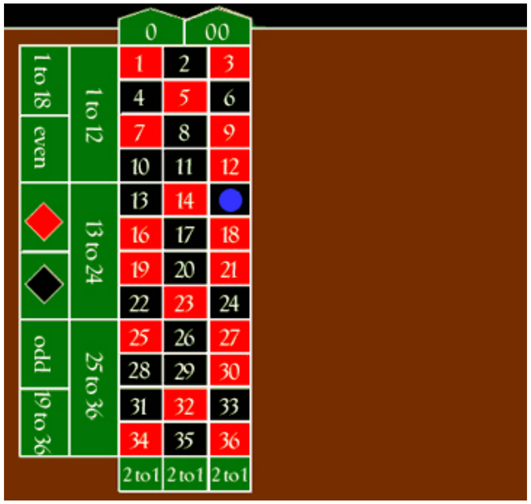
Information Theory

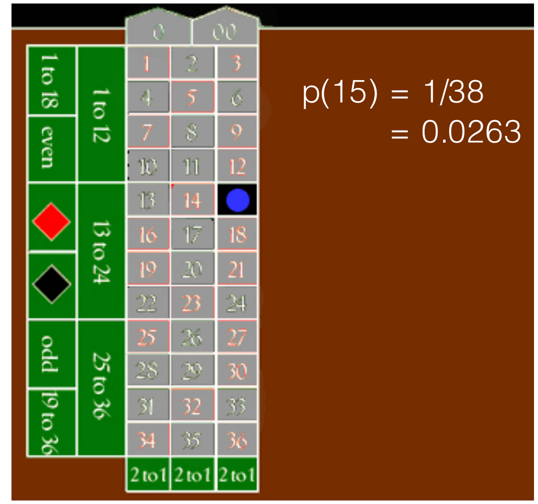
- entropy and mutual information
- how to calculate terms efficiently
- uses of information theory

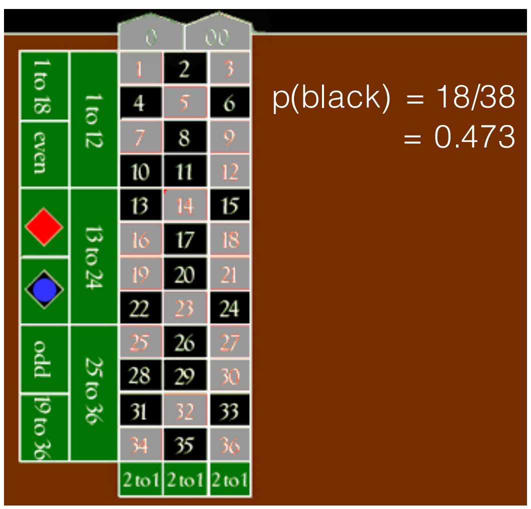
Probability Reminder

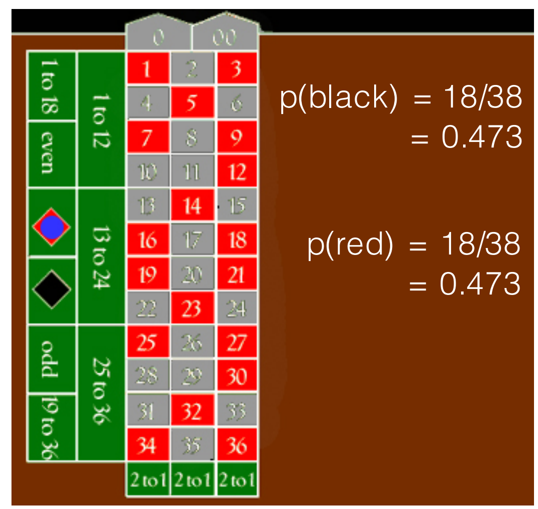
What is a probability distribution?

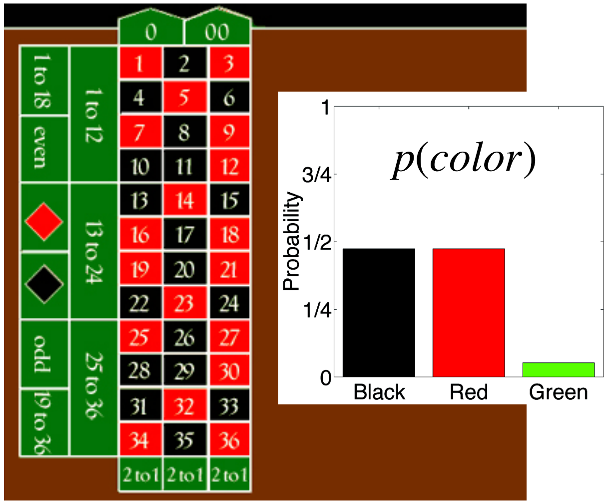




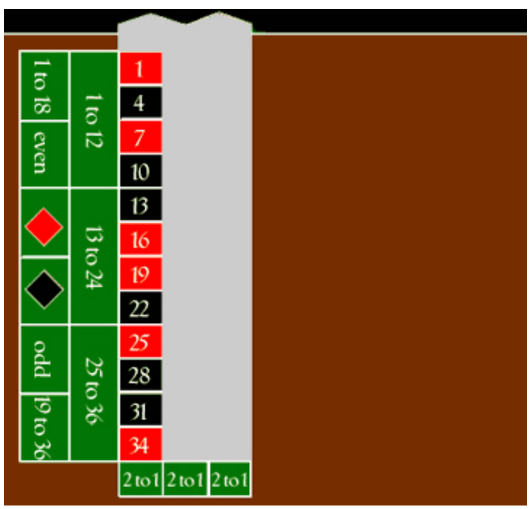

p(15) = 1/38
= 0.0263

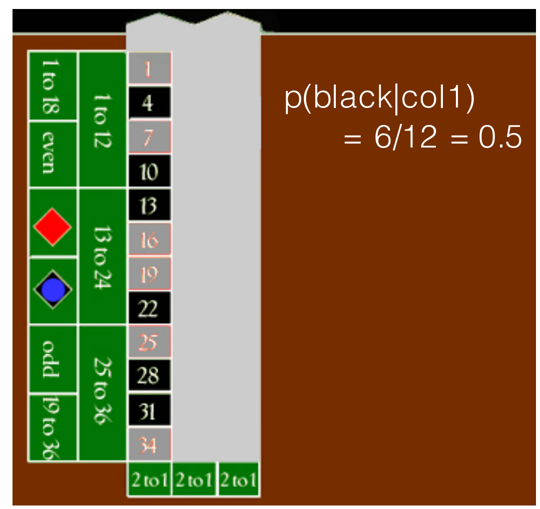

p(black) = 18/38
= 0.473

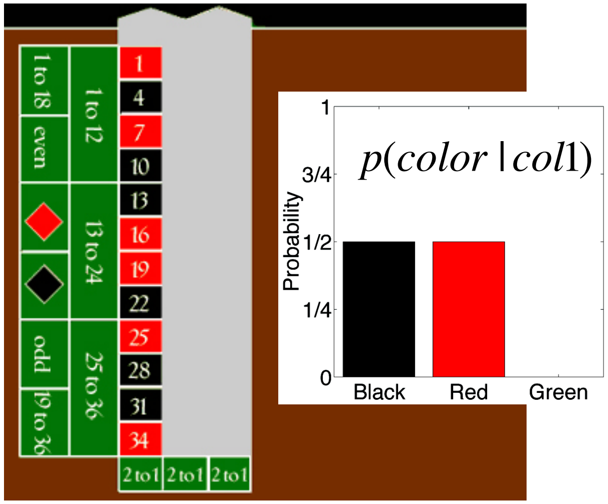

p(black) = 18/38
= 0.473
p(red) = 18/38
= 0.473

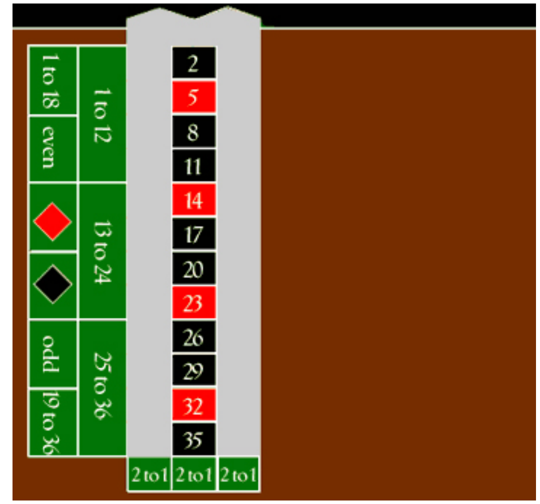


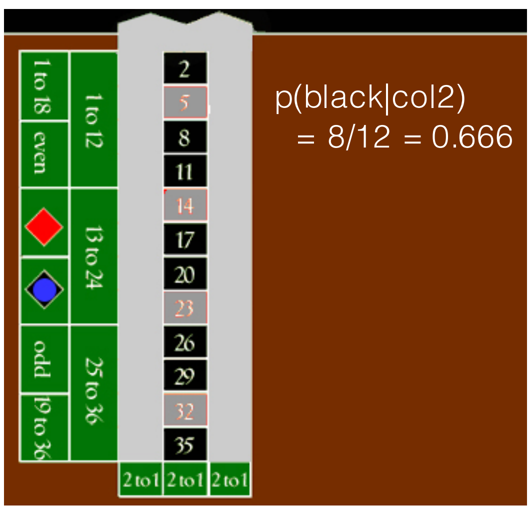
What is a conditional probability?

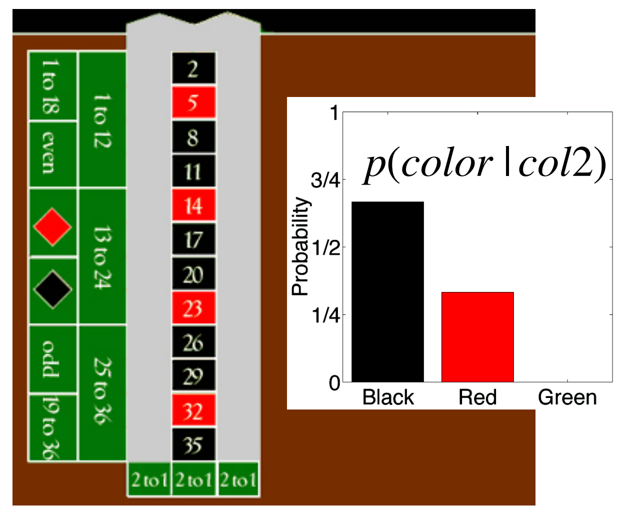





1 to 18
even
odd
19 to 36
1 to 12
13 to 24
25 to 36
1
4
7
10
13
16
19
22
25
28
31
34
2 to 1
2 to 1
2 to 1
$p(color \mid col1)$
Probability
1
3/4
1/2
1/4
0
Black
Red
Green


1 to 18
even
odd
19 to 36
1 to 12
13 to 24
25 to 36
2
5
8
11
14
17
20
23
26
29
32
35
2 to 1
2 to 1
2 to 1


1 to 18
even
odd
19 to 36
1 to 12
13 to 24
25 to 36
2
5
8
11
14
17
20
23
26
29
32
35
2 to 1
2 to 1
2 to 1
p(black|col2)
= 8/12 = 0.666


1 to 18
even
odd
19 to 36
1 to 12
13 to 24
25 to 36
2
5
8
11
14
17
20
23
26
29
32
35
2 to 1
2 to 1
2 to 1
$p(color \mid col2)$
Probability
1
3/4
1/2
1/4
0
Black
Red
Green

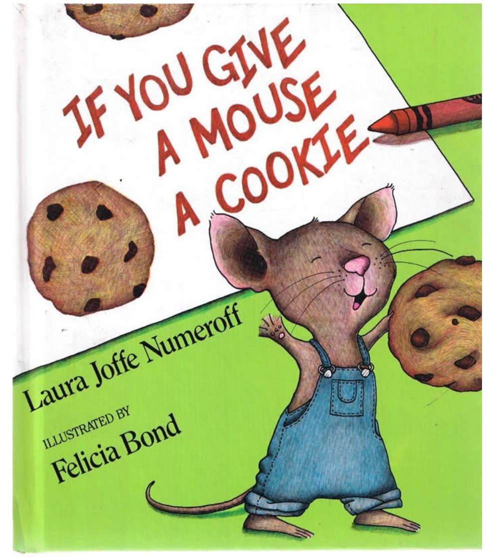
Probability rules

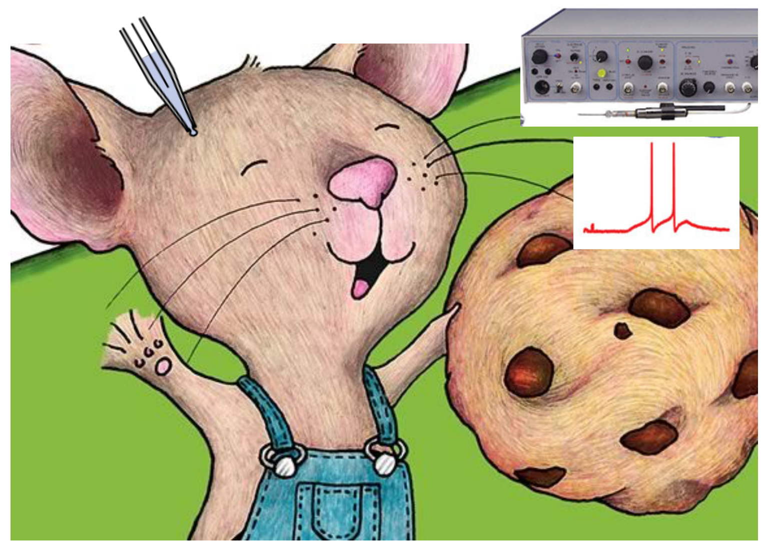
$$P(X,Y) = P(Y,X)$$

$$P(X,Y) = P(X|Y)P(Y)$$

$$\sum_X P(X) = 1$$

$$\sum_X P(X|Y) = 1$$





Probability rules

$$P(X,Y) = P(Y,X)$$

$$P(X,Y) = P(X|Y)P(Y)$$

$$\sum_X P(X) = 1$$

$$\sum_X P(X|Y) = 1$$

Probability rules

$$P(\text{spike}, \text{gave a mouse a cookie}) = P(\text{gave a mouse a cookie}, \text{spike})$$

$$P(X,Y) = P(X|Y)P(Y)$$

$$\sum_X P(X) = 1$$

$$\sum_X P(X|Y) = 1$$

Probability rules

$$P(\text{spike}, \text{gave a mouse a cookie}) = P(\text{gave a mouse a cookie}, \text{spike})$$

$$= P(\text{gave a mouse a cookie}|\text{spike})P(\text{spike})$$

$$\sum_X P(X) = 1$$

$$\sum_X P(X|Y) = 1$$

Probability rules

$$P(\text{spike}, \text{gave a mouse a cookie}) = P(\text{gave a mouse a cookie}, \text{spike})$$

$$= P(\text{gave a mouse a cookie}|\text{spike})P(\text{spike})$$

$$\sum_{\text{cookie states}} P(\text{gave a mouse a cookie}) = 1$$

$$\sum_X P(X|Y) = 1$$

Probability rules

$$P(\text{spike, gave a mouse a cookie}) = P(\text{gave a mouse a cookie, spike})$$

$$= P(\text{gave a mouse a cookie}|\text{spike})P(\text{spike})$$

$$\sum_{\text{cookie states}} P(\text{gave a mouse a cookie}) = 1$$

$$\sum_{\text{cookie states}} P(\text{gave a mouse a cookie}|\text{spike}) = 1$$

Bayesian Inference

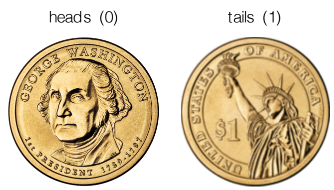
When do people use Bayesian Inference?

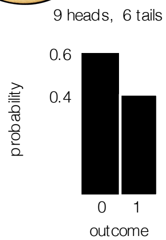
- You have probability distribution over the state of a variable of interest, *x*
- You learn something new, for example that some other random variable, *y,* has a particular value
- You'd like to update your beliefs about *x*, to incorporate this new evidence

What do you need to know to use it?

- You need to be able to express your prior beliefs about *x* as a probability distribution, $p(x)$
- You must able to relate your new evidence to your variable of interest in terms of it's likelihood, $p(y|x)$
- You must be able to multiply.

Frequentists vs Bayesians

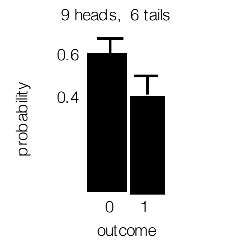




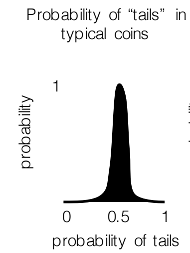
Frequentists vs Bayesians

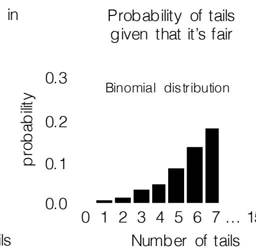
Frequentists
"After many repeated experiments we will arrive at the true frequency of heads/tails"

Bayesians
"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."



Estimate of the true probability

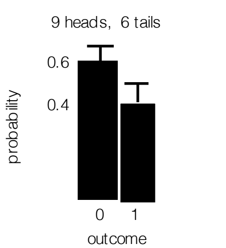




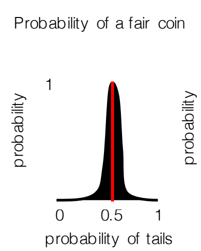
Frequentists vs Bayesians

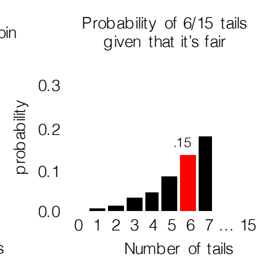
Frequentists
"After many repeated experiments we will arrive at the true frequency of heads/tails"

Bayesians
"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."



Estimate of the true probability





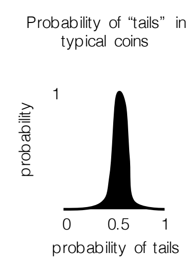
Probability that the coin is fair given past experience

Priors

Bayesians
"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(A)$$

Probability of getting "tails" in your experience with coins



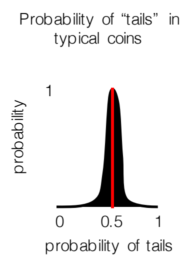
Priors

Bayesians

"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(A)$$

Probability of a fair coin in your experience with coins



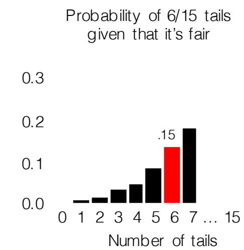
Likelihood

Bayesians

"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(B|A)$$

Probability of 6/15 tails given that it's a fair coin



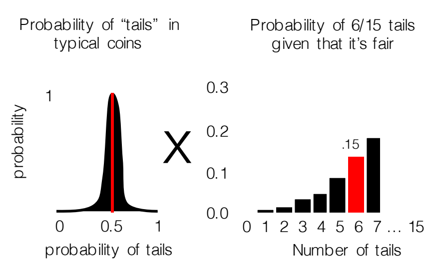
Posterior

Bayesians

"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(A|B)$$

Probability it's a fair coin given 6/15 tails



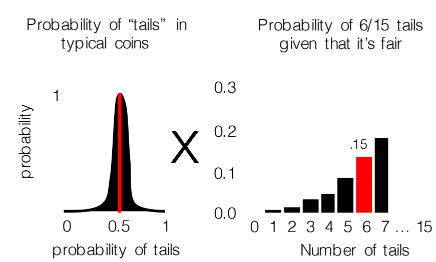
Bayes' Rule

Bayesians

"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}.$$

Probability it's a fair coin given 6/15 tails



$P(A)$ $\qquad$ $P(B|A)$

Bayes' Rule

$$p(x, y) = p(y, x)$$

$$p(x|y)p(y) = p(y|x)p(x)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

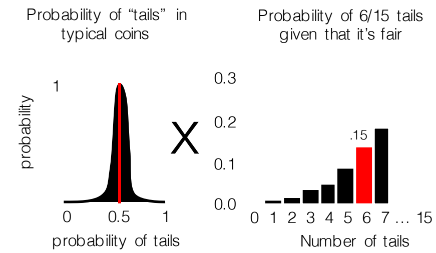
Bayes' Rule

Bayesians

"Probabilities reflect uncertainties in the world, and should take into account all of our knowledge about coins."

$$P(A|B) \propto P(B|A)P(A)$$

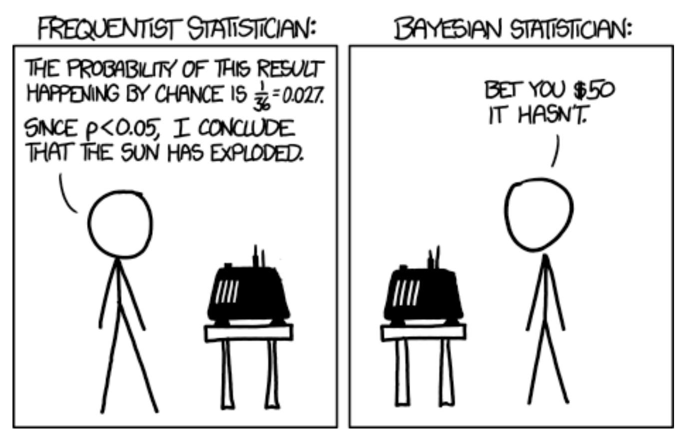
Probability it's a fair coin given 6/15 tails



$P(A)$ $P(B|A)$

Example: Did the sun just explode?

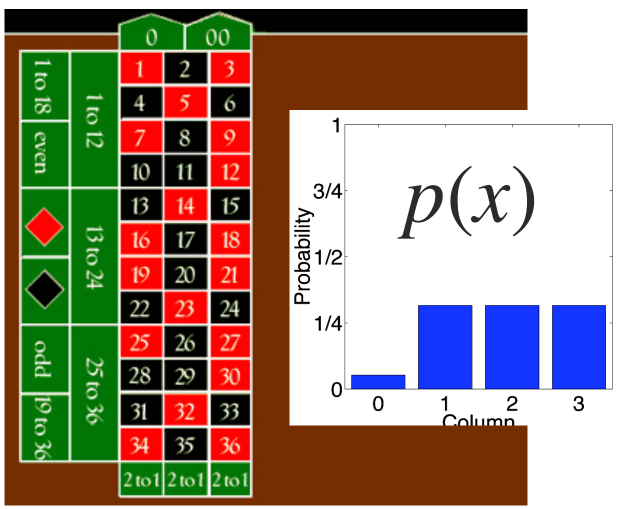
Example: Did the sun just explode?



Bayesian Roulette

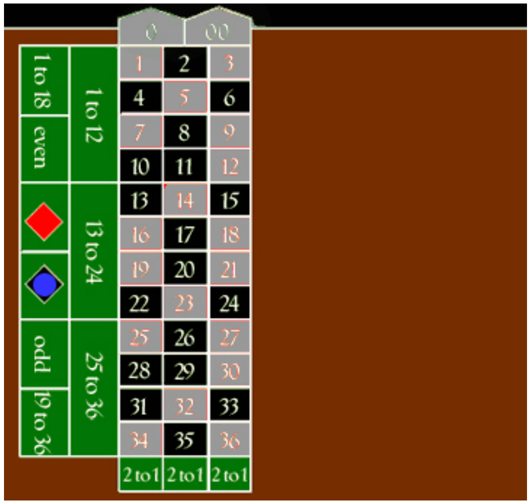
Bayesian Roulette

- We' re interested in which column will win.
- *p*(*column*) is our prior.



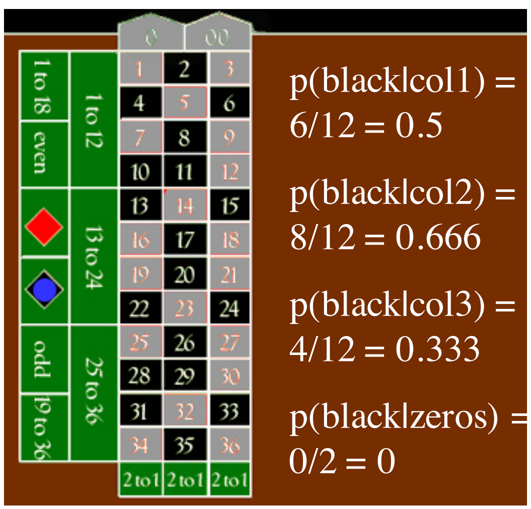
Bayesian Roulette

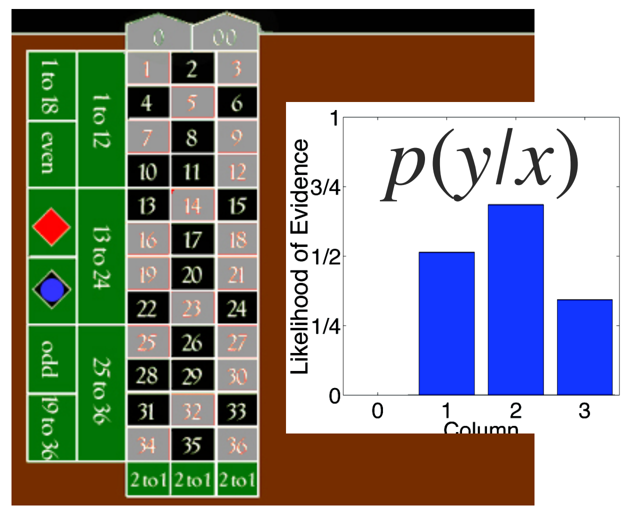
- We' re interested in which column will win.
- *p*(*column*) is our prior.
- We learn *color*=black.



Bayesian Roulette

- We're interested in which column will win.
- *p*(*column*) is our prior.
- We learn *color*=black.
- What is *p*(*color*=black|*column*)?





Bayesian Roulette

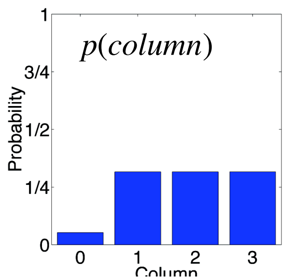
- We're interested in which column will win.
- $p(column)$ is our prior.
- We learn *color*=black.
- What is $p(color\text{=black}|column)$?
- We could calculate $p(color\text{=black})$, but who cares, we'll normalize when we're done.

Bayesian Roulette

- We're interested in which column will win.
- $p(column)$ is our prior.
- We learn *color*=black.
- What is $p(color\text{=black}|column)$?
- We could calculate $p(color\text{=black})$, but who cares, we'll normalize when we're done.

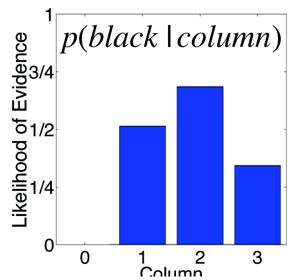
- Go directly to BAYES.

Bayes' Rule



$P(A)$

$\times$

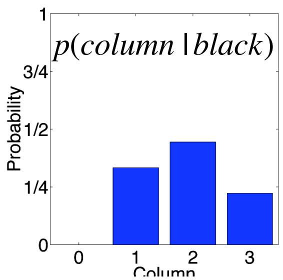


$P(B|A)$

Bayes' Rule

$$P(A|B) \propto P(B|A)P(A)$$

Bayes' Rule



No one would really use Bayesian Inference for Roulette.

Stanford University Hospital

p(hepititus | fever, hematuria, pale stool, abdominal pain, jaundice)

NASA

p(hull breach | pressure loss, tremor, attitude sensor failure)

Stanford Bioinformatics Group

p(transmembrane protein | genetic sequence)

iPhone Texting

p(you are writing "have" | last word, last 5 keystrokes)

Exercise: believing a paper

Is this result real, chance, or falsified?

p(result truth | p = 0.04, published in Science)

p(published in Science) = 0.07

p(p-value = 0.04, chance) = 0.04

p(chance, p-value=0.04 | published in Science) = 0.001

p(falsification) = 0.02

p(chance) = p(real) = 0.49

p(true, p-value=0.04 | published in Science) = 0.01

p(falsified, p-value=0.04 | published in Science) = 0.0005

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

$$p(x,y) = p(x|y)p(y)$$

$$p(x,y|z) = p(x|y,z)p(y|z)$$

Exercise: believing a paper

Is this result real, chance, or falsified?

p(result truth | p = 0.04, published in Science)

$\propto$ p(result truth) * p(p = 0.04, published in Science | result truth)

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Exercise: believing a paper

Is this result chance?

p(chance | p = 0.04, published in Science)

$\propto$ p(chance) * p(p = 0.04, published in Science | chance)

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Exercise: believing a paper

Is this result chance?

p(chance | p = 0.04, published in Science)

$\propto$ p(chance) * p(p = 0.04, published in Science | chance)

$=$ p(chance) * $\dfrac{\text{p(p = 0.04, published in Science, chance)}}{\text{p(chance)}}$

$$p(x,y) = p(x|y)p(y) \qquad p(x|y) = \frac{p(x,y)}{p(y)}$$

Exercise: believing a paper

Is this result chance?

p(chance | p = 0.04, published in Science)

$\propto$ p(chance) * p(p = 0.04, published in Science | chance)

$=$ p(chance) * $\dfrac{\text{p(p = 0.04, published in Science, chance)}}{\text{p(chance)}}$

$=$ p(chance) * $\dfrac{\text{p(published in Science) * p(0.04, chance | published in Science)}}{\text{p(chance)}}$

Exercise: believing a paper

Is this result chance?

p(chance | p = 0.04, published in Science)

$\propto$ p(chance) * p(p = 0.04, published in Science | chance)

$=$ p(chance) * $\dfrac{\text{p(p = 0.04, published in Science, chance)}}{\text{p(chance)}}$

$=$ p(chance) * $\dfrac{\text{p(published in Science) * p(0.04, chance | published in Science)}}{\text{p(chance)}}$

$=$ p(published in Science) * p(0.04, chance | published in Science)

$=$ 0.07 * 0.001 $=$ 0.00007

Normalization

p(true, p-value=0.04 | published in Science) = 0.01

p(falsified, p-value=0.04 | published in Science) = 0.0005

Is this result real?

p(true | p = 0.04, published in Science)

$\propto$

p(published in Science) * p(0.04, true | published in Science)

$= 0.07 * 0.01 = 0.0007$

Is this result falsified?

p(true | p = 0.04, published in Science)

$\propto$

p(published in Science) * p(0.04, falsified | published in Science)

$= 0.07 * 0.0005 = 0.000035$

Normalization

Is this result real, chance, or falsified?

p(result | p = 0.04, published in Science)

Is this result real?

p(true | p=0.04, published in Science)

$$\frac{0.0007}{0.0007 + 0.0007 + 0.000035} = 0.87$$

Is this result chance?

p(chance | p=0.04, published in Sci.)

$$\frac{0.00007}{0.0007 + 0.0007 + 0.000035} = 0.09$$

Is this result falsified?

p(falsified | p=0.04, published in Science)

$$\frac{0.000035}{0.0007 + 0.0007 + 0.000035} = 0.05$$

But does stuff actually have applications in Neuroscience?

Smarter data analysis

Bayesian models as models of behavior or the brain.

Smarter data analysis

The null distribution of stochastic search gene suggestion: a Bayesian approach to gene mapping.

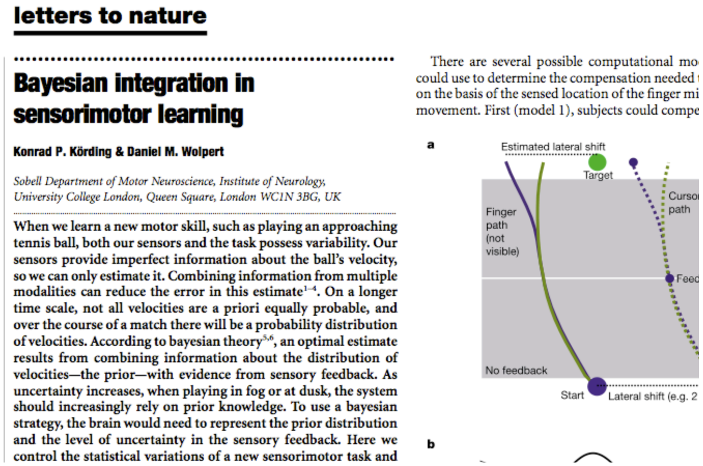
Bayesian hierarchical model for estimating gene expression intensity using multiple scanned microarrays.

A Bayesian genome-wide linkage analysis of quantitative traits for rheumatoid arthritis via perfect sampling.

Bayesian hierarchical modeling of means and covariances of gene expression data within families.

Nonlinear predictive modeling of MHC class II-peptide binding using Bayesian neural networks.

WWBD?



letters to nature

Bayesian integration in sensorimotor learning

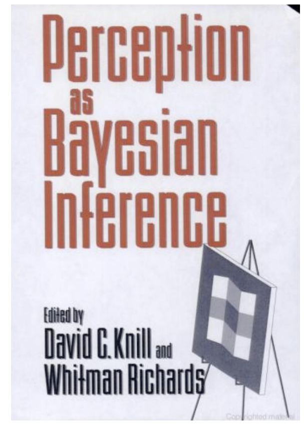
Konrad P. Körding & Daniel M. Wolpert

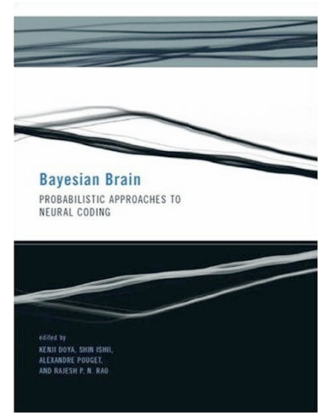
Sobell Department of Motor Neuroscience, Institute of Neurology, University College London, Queen Square, London WC1N 3BG, UK

When we learn a new motor skill, such as playing an approaching tennis ball, both our sensors and the task possess variability. Our sensors provide imperfect information about the ball's velocity, so we can only estimate it. Combining information from multiple modalities can reduce the error in this estimate[1–4]. On a longer time scale, not all velocities are a priori equally probable, and over the course of a match there will be a probability distribution of velocities. According to bayesian theory[5,6], an optimal estimate results from combining information about the distribution of velocities—the prior—with evidence from sensory feedback. As uncertainty increases, when playing in fog or at dusk, the system should increasingly rely on prior knowledge. To use a bayesian strategy, the brain would need to represent the prior distribution and the level of uncertainty in the sensory feedback. Here we control the statistical variations of a new sensorimotor task and

There are several possible computational mo could use to determine the compensation needed on the basis of the sensed location of the finger mi movement. First (model 1), subjects could compe

Bayesian Behavior & Bayesian Brain

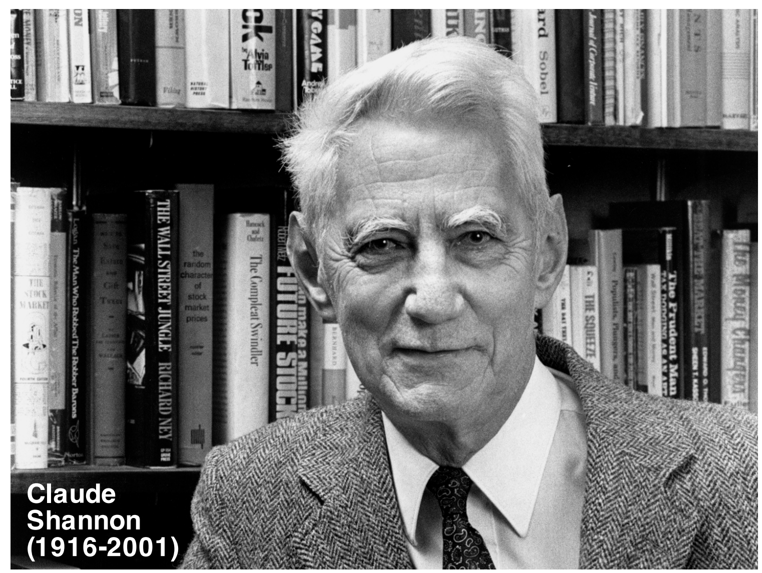




Today's lecture

Bayesian Statistics
- priors, likelihoods, and posteriors
- Bayes rule
- Bayesian estimators
- uses of Bayesian statistics

Information Theory
- entropy and mutual information
- how to calculate terms efficiently
- uses of information theory



Crux of information theory

Defined for any probability distributions

Entropy
"uncertainty or complexity"

$$H(X) = -\sum_X p(X) \log p(X)$$

Information
"decrease in uncertainty"

$$I(X,Y) = H(X) - H(X|Y)$$

Entropy

A measure of uncertainty

A measure of disorder

A measure of complexity

Entropy

A measure of uncertainty

A measure of disorder

A measure of complexity

The (average) number of yes/no questions needed to completely specify the state of a system

The (average) number of yes/no questions needed to completely specify the state of a system

What if there were two coins?

What if there were two coins?

What if there were two coins?

What if there were two coins?

number of states $= 2^{\text{number of yes-no questions}}$

2 states.	1 question.
4 states.	2 questions.
8 states.	3 questions.
16 states.	4 questions.

number of states $= 2^{\text{number of yes-no questions}}$

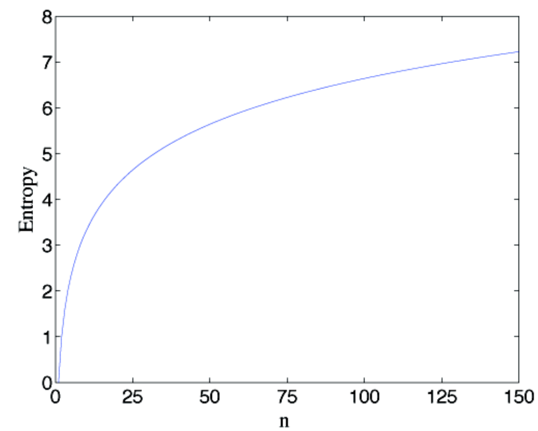
$\log_2$(number of states) = number of yes-no questions

H(X) is entropy, the number of yes-no questions required to specify the state of *X*

n is the number of states of the system, assumed (for now) to be equally likely

$$H(X) = \log_2 n$$

$$H(X) = \log_2 n$$



Consider Dice

The Six Sided Die

$H = \log_2(6) = 2.585$ bits

The Four Sided Die

$H = \log_2(4) = 2.000$ bits

The Twenty Sided Die

$H = \log_2(20) = 4.322$ bits

What about all three dice?

$H = \log_2(4 \times 6 \times 20)$

What about all three dice?

$H = \log_2(4) + \log_2(6) + \log_2(20)$

What about all three dice?

H = 8.907 bits

What about all three dice?

Entropy, from independent elements of a system, adds

Let's rewrite this a bit...

$$H(X) = \log_2 n$$

Trivial Fact 1:
$\log_2(x) = -\log_2(1/x)$

$$H(X) = \log_2 n$$

Trivial Fact 1:
$\log_2(x) = -\log_2(1/x)$

$$H(X) = \log_2 n$$

Trivial Fact 1:
$\log_2(x) = -\log_2(1/x)$

$$H(X) = -\log_2 \frac{1}{n}$$

Trivial Fact 2:

if there are *n* equally likely possibilities,
p = 1/n

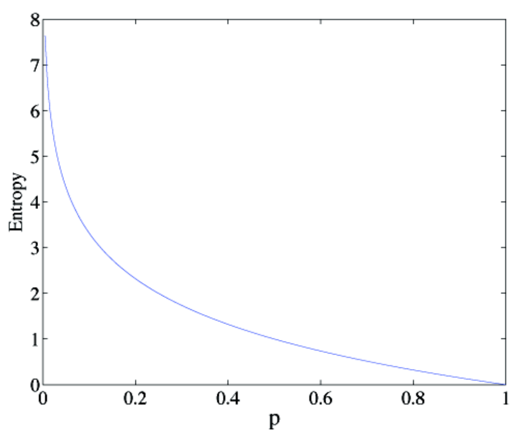
$$H(X) = -\log_2 \frac{1}{n}$$

Trivial Fact 2:

if there are *n* equally likely possibilities,
p = 1/n

$$H(X) = -\log_2 p$$

$$H(X) = -\log_2 p$$



$$H(X) = -\log_2 p$$

What if the n states
are not equally probable?

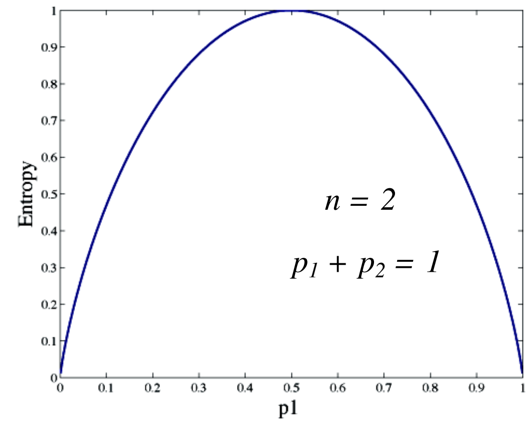
Maybe we should use the
expected value of the entropies,
a weighted average by probability

$$H(X) = -\sum_X p(X) \log p(X)$$

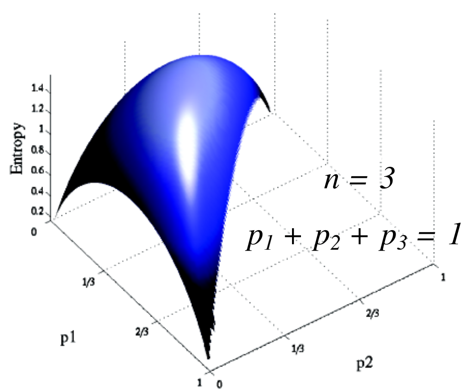
Let's do a simple example:

n = 2, how does H change as we vary $p(X_1)$ and $p(X_2)$?

$$H(X) = -\sum_X p(X) \log p(X)$$



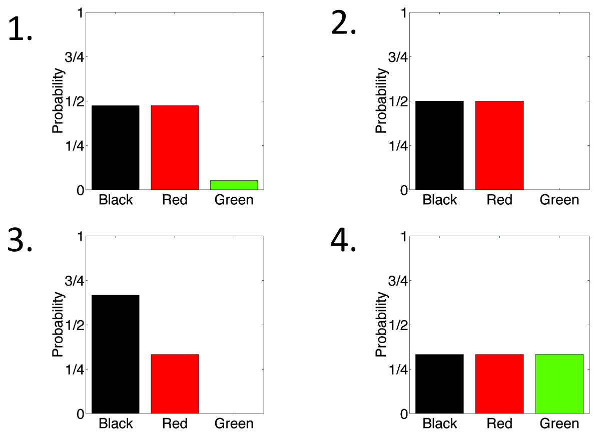
$$H(X) = -\sum_X p(X) \log p(X)$$



The bottom line intuitions for Entropy:

- Entropy is a statistic for describing a probability distribution.
- Probabilities distributions which are flat, broad, dense, etc. have HIGH entropy.
- Probability distributions which are peaked, sharp, narrow, sparse, etc. have LOW entropy.
- Entropy adds for independent elements of a system, thus entropy grows with the dimensionality of the probability distribution.
- Entropy is zero IFF the system is in a definite state, i.e. p = 1 somewhere and 0 everywhere else.

Pop Quiz: rank from high to low entropy



Uses of entropy

The (average) number of yes/no questions needed
to completely specify the state of a system

Redundancy of the English language (1,025,109.8 words)
If every word were used equally,
H(English word) = $\log_2$(1,025,109.8) = 19.97 bits

Uses of entropy

The (average) number of yes/no questions needed
to completely specify the state of a system

Redundancy of the English language (1,025,109.8 words)
If every word were used equally,
H(English word) = $\log_2$(1,025,109.8) = 19.97 bits

but…

H(English) = 11.82

Uses of entropy

The (average) number of yes/no questions needed
to completely specify the state of a system

Redundancy of the English language (1,025,109.8 words)
If every word were used equally,
H(English word) = $\log_2$(1,025,109.8) = 19.97 bits

but…

H(English) = 11.82

Redundancy = $1 - H_{max}/H_{true}$ = 1 – 19.97/11.82 = 0.41

Uses of entropy

Amount of information transmitted by a neuron

Complexity of an experimental stimulus

Redundancy of experimental variables

Reliability of a neuron

Crux of information theory

Defined for any probability distributions

Entropy
"uncertainty or complexity"

$$H(X) = -\sum_X p(X) \log p(X)$$

Information
"decrease in uncertainty"

$$I(X,Y) = H(X) - H(X|Y)$$

Information

$$I(X,Y) = H(X) - H(X|Y)$$

reduction in uncertainty

information X gives about Y

newsworthiness

a change in what you don't know

Information as a measure of correlation

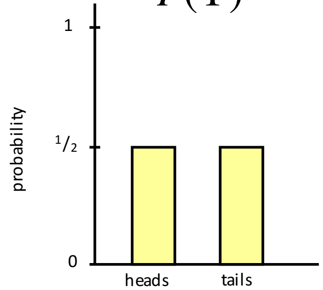
x

y

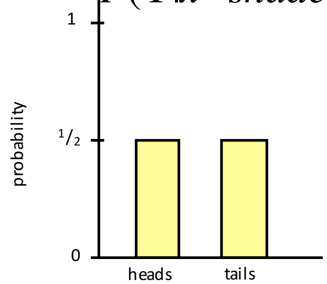
$$I\,(X;Y) = H(Y) - H(Y|X) = 0 \text{ bits}$$

$H(Y) = 1$ $\qquad$ $H(Y\,|\,x{=}shades) = 1$

$P(Y)$



$P(Y|x{=}shades\,)$



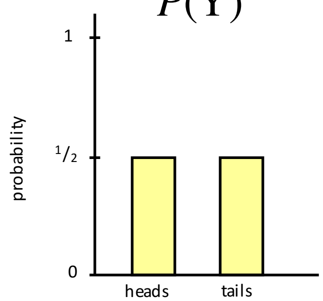
Information as a measure of correlation

X $\qquad$ Y

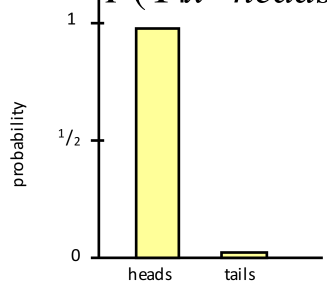
$$I\,(X;Y) = H(Y) - H(Y|X) \sim 1 \text{ bit}$$

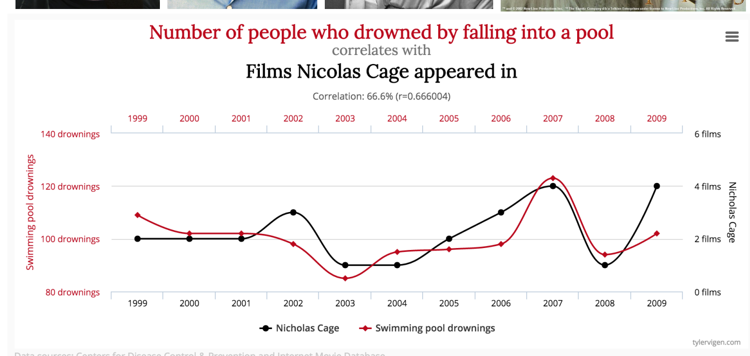
$H(Y) = 1$ $\qquad$ $H(Y\,|\,x{=}heads) \sim 0$

$P(Y)$



$P(Y|x{=}heads\,)$





Information theory in Neuroscience

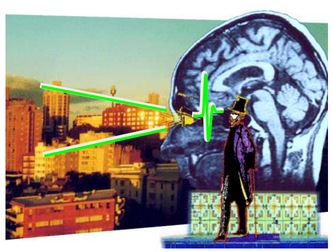
x **y**

Information is Mutual

$$I(X;Y) = I(Y;X)$$

$$H(Y)-H(Y|X) = H(X)-H(X|Y)$$

Information is Mutual



$$I(\text{Stimulus};\text{Spike}) = I(\text{Spike};\text{Stimulus})$$

What a spike tells the Brain about the stimulus,
is the same as what our stimulus choice tells us about
the likelihood of a spike.

Uses of mutual information

How similar are two neurons' representations?

How much information does a visual neuron transmit about a visual stimulus?

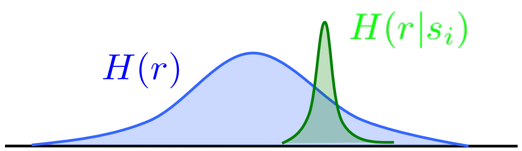
How much is a time series recording influenced by an experimental perturbation?

What experimental variable is most affected by an experimental perturbation?

Computing mutual information

How much information does a visual neuron transmit about a visual stimulus?

$$I(X,Y) = H(X) - H(X|Y)$$



How to screw it up

Choose stimuli which are not representative.

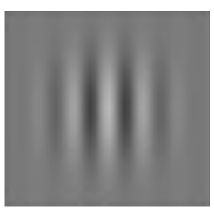
Measure the "wrong" aspect of the response.

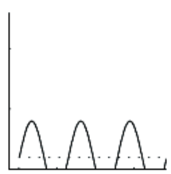
Don't take enough data to estimate P( ) well.

Use a crappy method of computing H( ).

Calculate I( ) and report it without comparing it to anything...

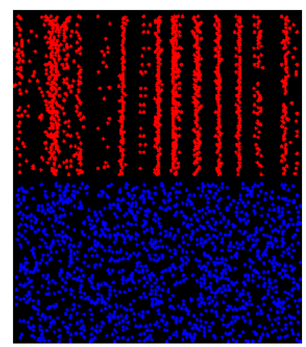
Here's an example of Information Theory applied appropriately





Temporal Coding of Visual Information in the Thalamus
Pamela Reinagel and R. Clay Reid
J. Neurosci. 20(14):5392-5400. (2000)

LGN responses are very reliable.



Is there information in the temporal pattern of spikes?

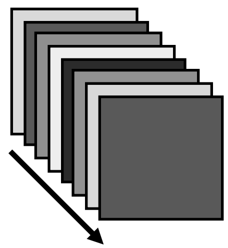
Patterns of Spikes in the LGN

x

spikes

…….0……..
…….1……..

y



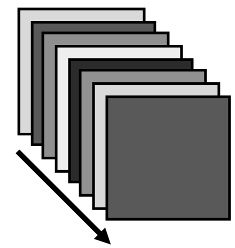
Patterns of Spikes in the LGN

x

spikes

…….00……..
…….10……..
…….01……..
…….11……..

y



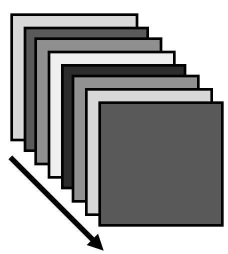
Patterns of Spikes in the LGN

x

spikes

…….000……..
…….101……..
…….011……..
…….100……..

y



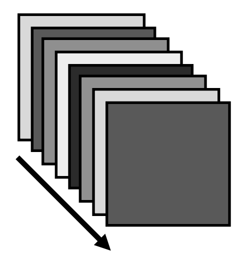
Patterns of Spikes in the LGN

x

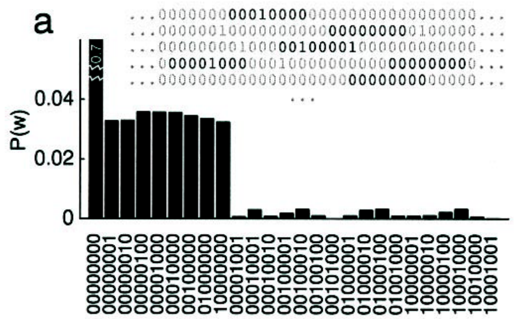
spikes

…….000100……..
…….101101……..
…….011110……..
…….010001……..

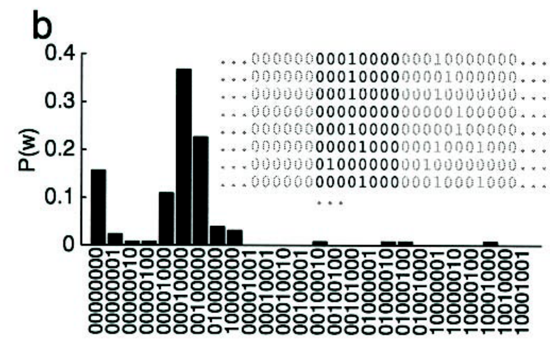
y



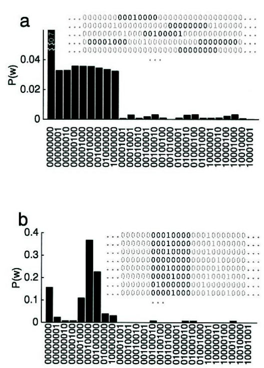
P(spike pattern)

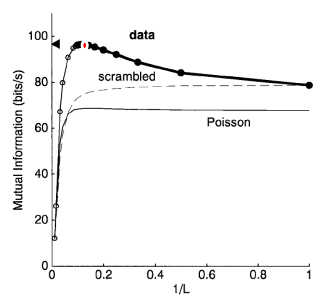


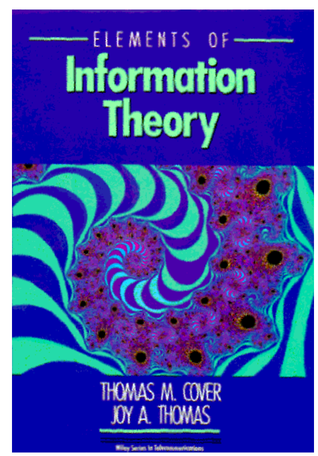
P(spike pattern | stimulus)

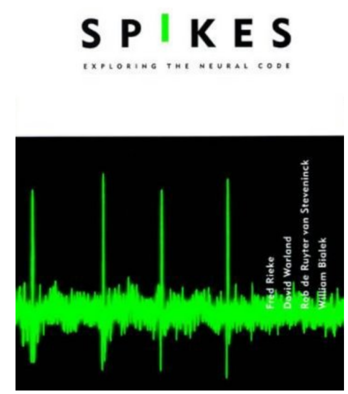


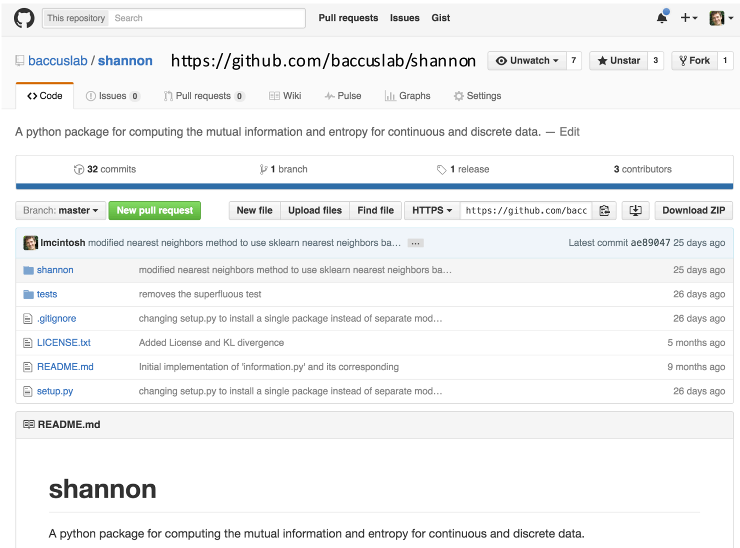
There is some extra Information in Temporal Patterns of spikes.










This repository
Search
Pull requests
Issues
Gist
baccuslab / shannon
https://github.com/baccuslab/shannon
Unwatch 7
Unstar 3
Fork 1
Code
Issues 0
Pull requests 0
Wiki
Pulse
Graphs
Settings
A python package for computing the mutual information and entropy for continuous and discrete data. — Edit
32 commits
1 branch
1 release
3 contributors
Branch: master
New pull request
New file
Upload files
Find file
HTTPS
https://github.com/bacc
Download ZIP
lmcintosh modified nearest neighbors method to use sklearn nearest neighbors ba…
Latest commit ae89047 25 days ago
shannon
modified nearest neighbors method to use sklearn nearest neighbors ba…
25 days ago
tests
removes the superfluous test
26 days ago
.gitignore
changing setup.py to install a single package instead of separate mod…
26 days ago
LICENSE.txt
Added License and KL divergence
5 months ago
README.md
Initial implementation of 'information.py' and its corresponding
9 months ago
setup.py
changing setup.py to install a single package instead of separate mod…
26 days ago
README.md
shannon
A python package for computing the mutual information and entropy for continuous and discrete data.