Introduction to Information Theory

Lecture 7 I400/I590

Artificial Life as an approach to Artificial Intelligence

Larry Yaeger

Professor of Informatics, Indiana University

It Starts with Probability

- Which starts with gambling
- In 1550 Cardan wrote a manuscript outlining the probabilities of dice rolls, points, and gave a rough definition of probability
  - However, the document was lost and not discovered until 1576, and not printed until 1663
  - So credit is normally given to someone else
    - What are the odds?
- In 1654 the Chevalier De Mere asked Blaise Pascal why he lost more frequently on one bet, rolling dice, than he did on another bet, when it seemed to him the chance of success in the two bets should be equal
  - He also asked how to correctly distribute the stakes when a dice game was incomplete

Roll the Dice

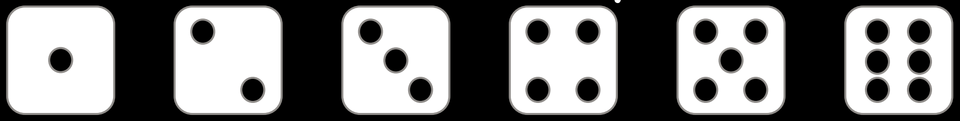
- Blaise Pascal exchanged a series of five letters with Pierre de Fermat, regarding the dice and points problems, in which they outlined the fundamentals of probability theory
- de Mere's dice-roll question was about the odds in two different bets:
  - That he would roll at least one six in four rolls of a single die
  - That he would roll at least one pair of sixes in 24 rolls of a pair of dice
- He reasoned (incorrectly) that the odds should be the same:
  - 4 * (1/6) = 2/3
  - 24 * (1/36) = 2/3

Basic Definitions

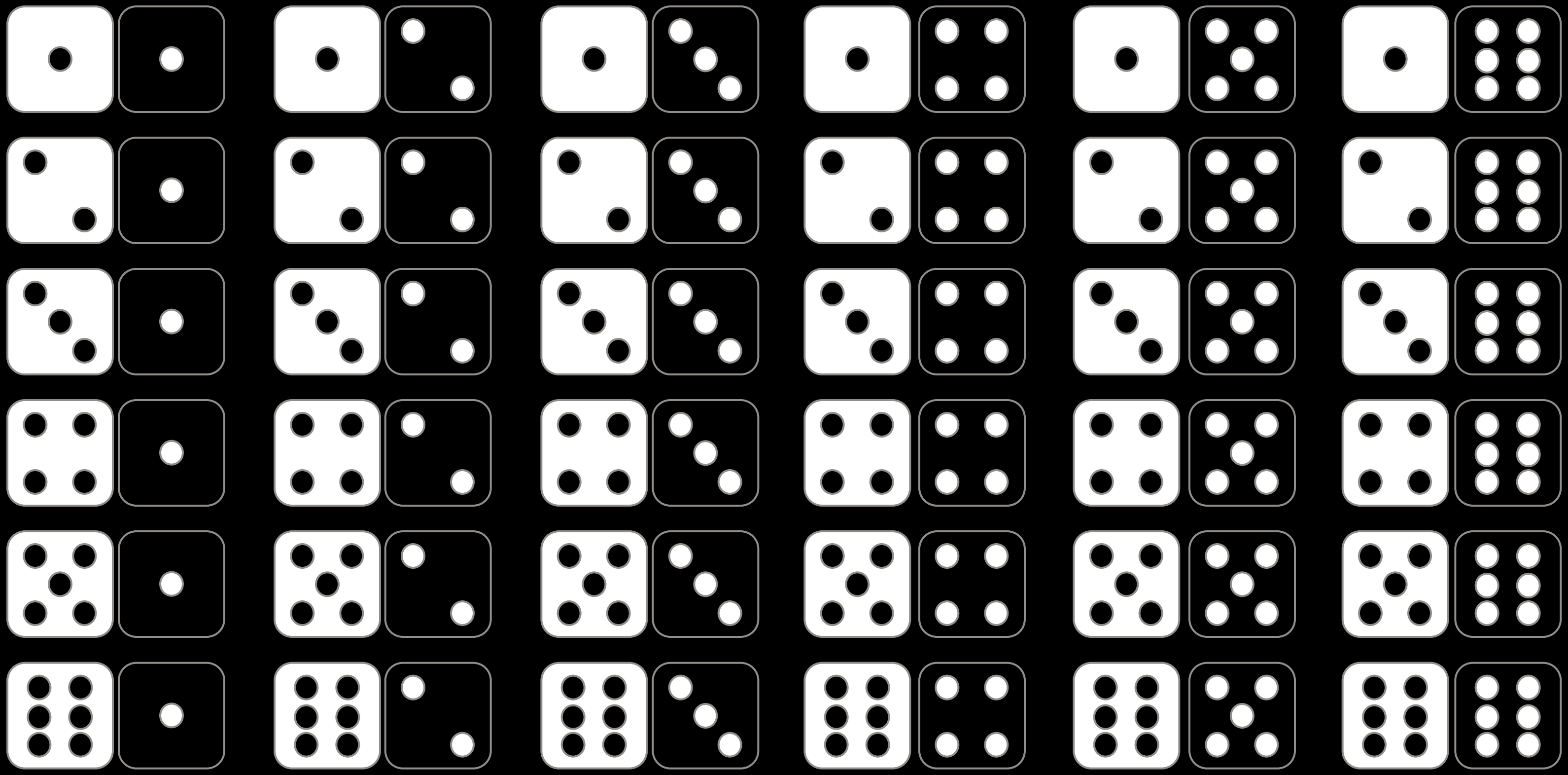
- *Random experiment* — The process of observing the outcome of a chance event
- *Elementary outcomes* — The possible results of a random experiment
- *Sample space* — The set of all elementary outcomes
- So if the event is the toss of a coin, then
    - Random experiment = recording the outcome of a single toss
    - Elementary outcomes = Heads or Tails
    - Sample space = {H,T}

Sample Space for Dice

- Single die has six elementary outcomes:



- Two dice have 36 elementary outcomes:



Probability = Idealized Frequency

- Let’s toss a coin 10 times and record a 1 for every heads and a 0 for every tails:

0
0
1
0
1
1
1
1
0
1

$$N = 10$$
$$N_H = 6$$
$$N_T = 4$$

$$F_H = N_H/N = 6/10 = 0.6$$
$$F_T = N_T/N = 4/10 = 0.4$$

$$P(H) \approx F_H = 0.6$$
$$P(T) \approx F_T = 0.4$$

- Why not 0.5?
  - Statistical stability (one kind of sampling error)

Probability = Idealized Frequency

- Now toss the coin 100 times, still recording a 1 for every heads and a 0 for every tails:

```
0 0 0 1 0 1 0 0 1 0
0 0 0 1 0 1 1 0 0 1
1 0 0 1 0 0 1 1 0 0
1 1 0 1 0 1 0 0 0 0
0 1 0 1 1 1 1 0 1 1
0 1 1 1 0 0 1 0 1 1
0 1 1 0 1 1 1 1 1 1
1 0 1 1 0 1 0 0 0 1
0 1 1 1 0 0 1 0 1 1
0 1 0 1 0 1 0 0 0 1
```

$N = 100$

$N_H = 51$

$N_T = 49$

$F_H = N_H/N = 51/100 = 0.51$

$F_T = N_T/N = 49/100 = 0.49$

$P(H) \approx F_H = 0.51$

$P(T) \approx F_T = 0.49$

- *Law of large numbers* gives us convergence to the actual probability as the number of samples goes to infinity

M Out of N

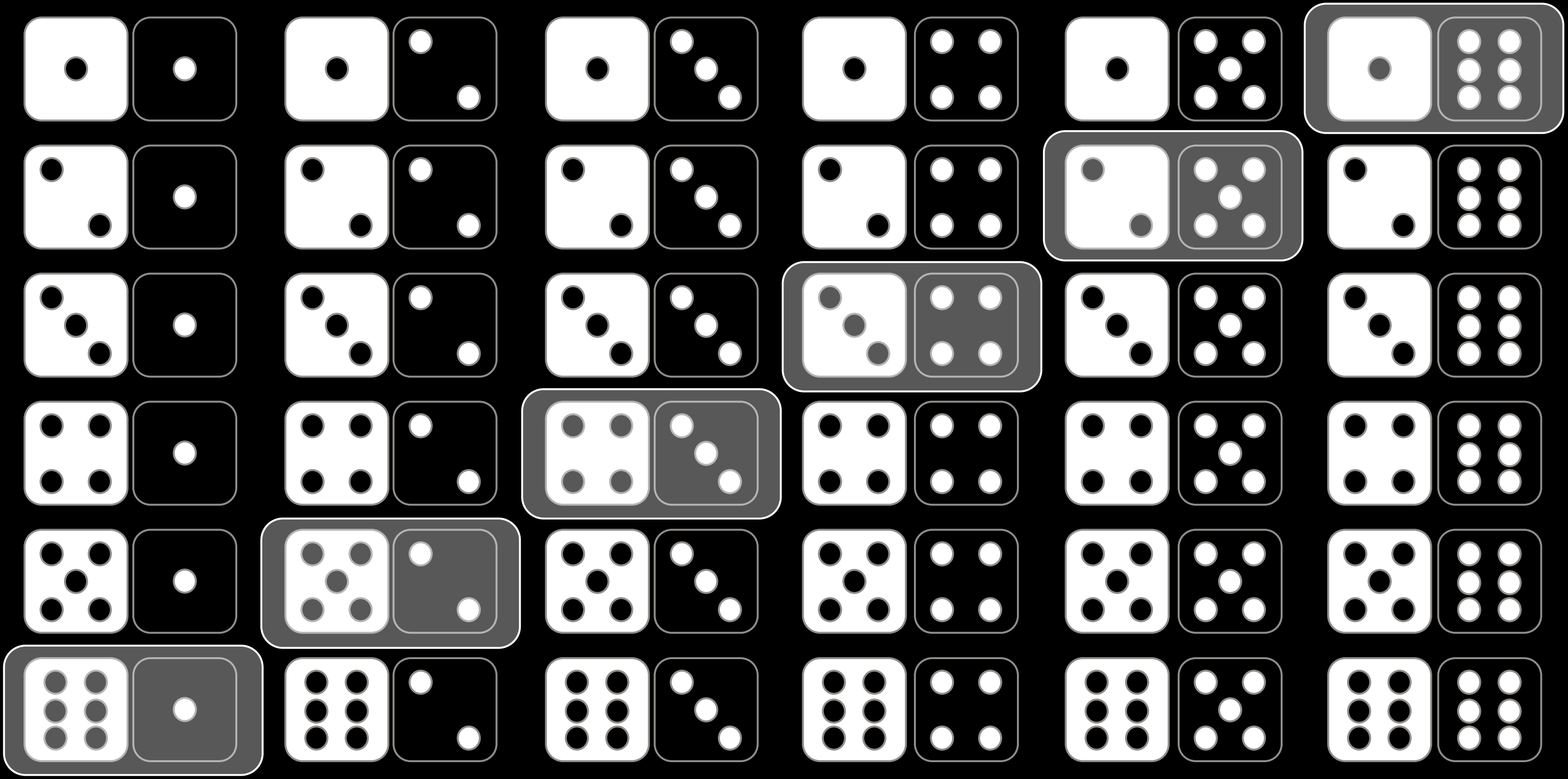
- When the random experiments (the samples) are independent of each other, then you can simply look at the number of possible ways to obtain a particular outcome relative to the total number of possible outcomes
- Tossing a (fair) coin will always produce either Heads or Tails, independent of previous experiments, so

  P(H) = 1/2 = 0.5
- Rolling a particular combination of dice, such as (White 5 and Black 2) represents one possible outcome out of 36 possible outcomes, so

  P(W5 and B2) = 1/36 = 0.02777…

Come On, Seven!

- However, rolling a seven can be done in any of six ways:



- P(Seven) = 6 / 36 = 1 / 6 = 0.1666…

Simple Rules of Probability

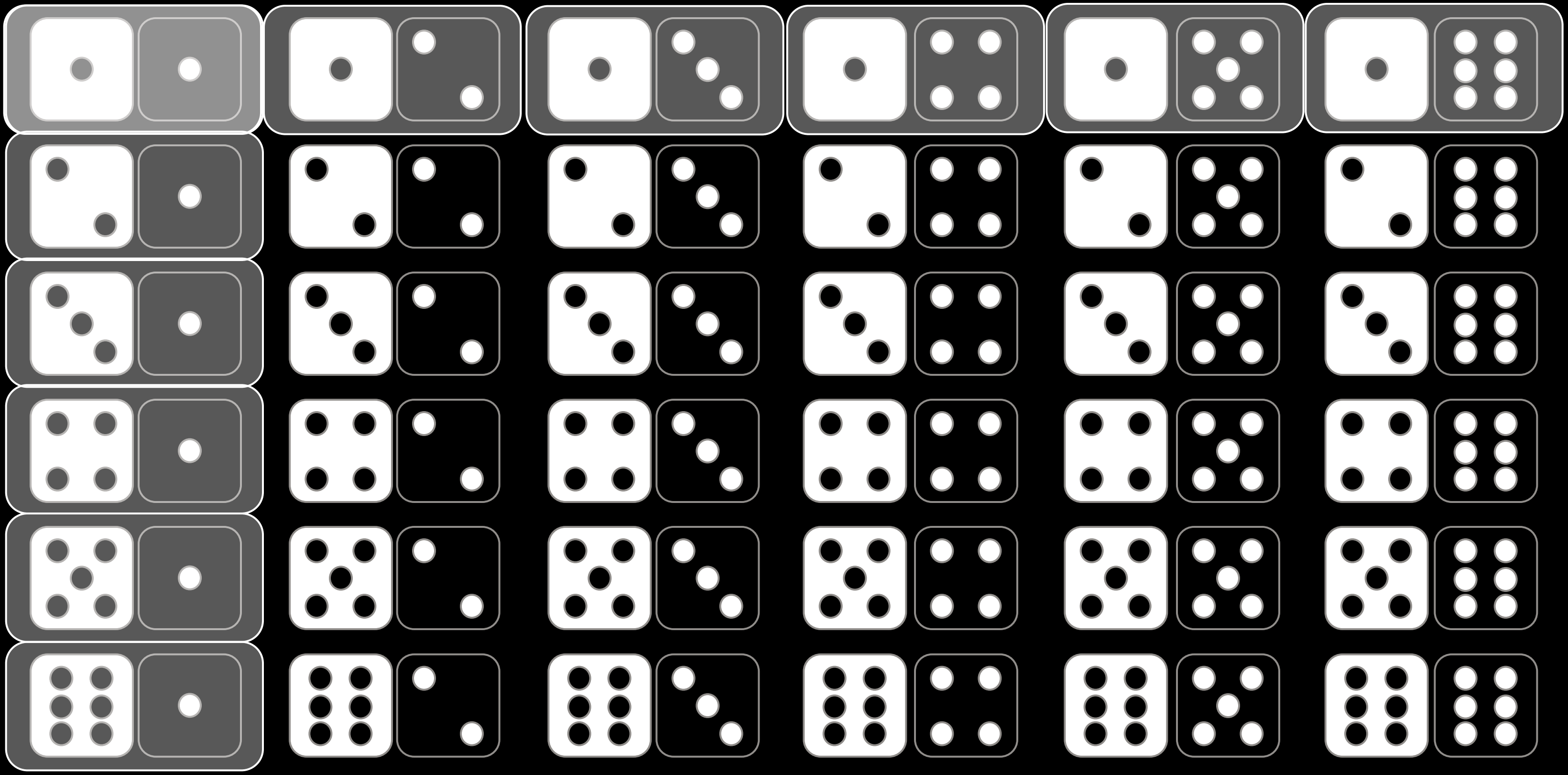
- Define $P(o_i)$ = probability of observing outcome o_i
- $0.0 \leq P(o_i) \leq 1.0$
- $\sum_i P(o_i) = 1.0$
- $P(\text{NOT } o_i) = 1.0 - P(o_i)$
  - Known as the *Subtraction Rule*

Events

- *Event* — a set of elementary outcomes
  - E.g., rolling a seven
    - Remember, you could get a seven with any of six different elementary outcomes
- Define P(x) = probability of observing event x
  - $P(x_i)$ = probability of observing i^{th} possible event
- $P(x) = \sum_j P(o_j)$
  - o_j are the elementary outcomes that produce event x
  - E.g., six ways of rolling seven yields 6 * (1/36) = 1/6
- $0.0 \le P(x_i) \le 1.0$
- $\sum_i P(x_i) = 1.0$
- $P(\text{NOT } x_i) = 1.0 - P(x_i)$

The Addition Rule

- Now throw a pair of black & white dice, and ask: What is the probability of throwing at least one one?
  - Let event a = the white die will show a one
  - Let event b = the black die will show a one



The Addition Rule

- Probability of throwing at least one one is P(a OR b), also written as $P(a \cup b)$
- Note that the elementary outcome when both dice are one (snake eyes) is counted twice if you just sum P(a) and P(b), so P(a AND b) must be subtracted, yielding
- P(a OR b) = P(a) + P(b) - P(a AND b)

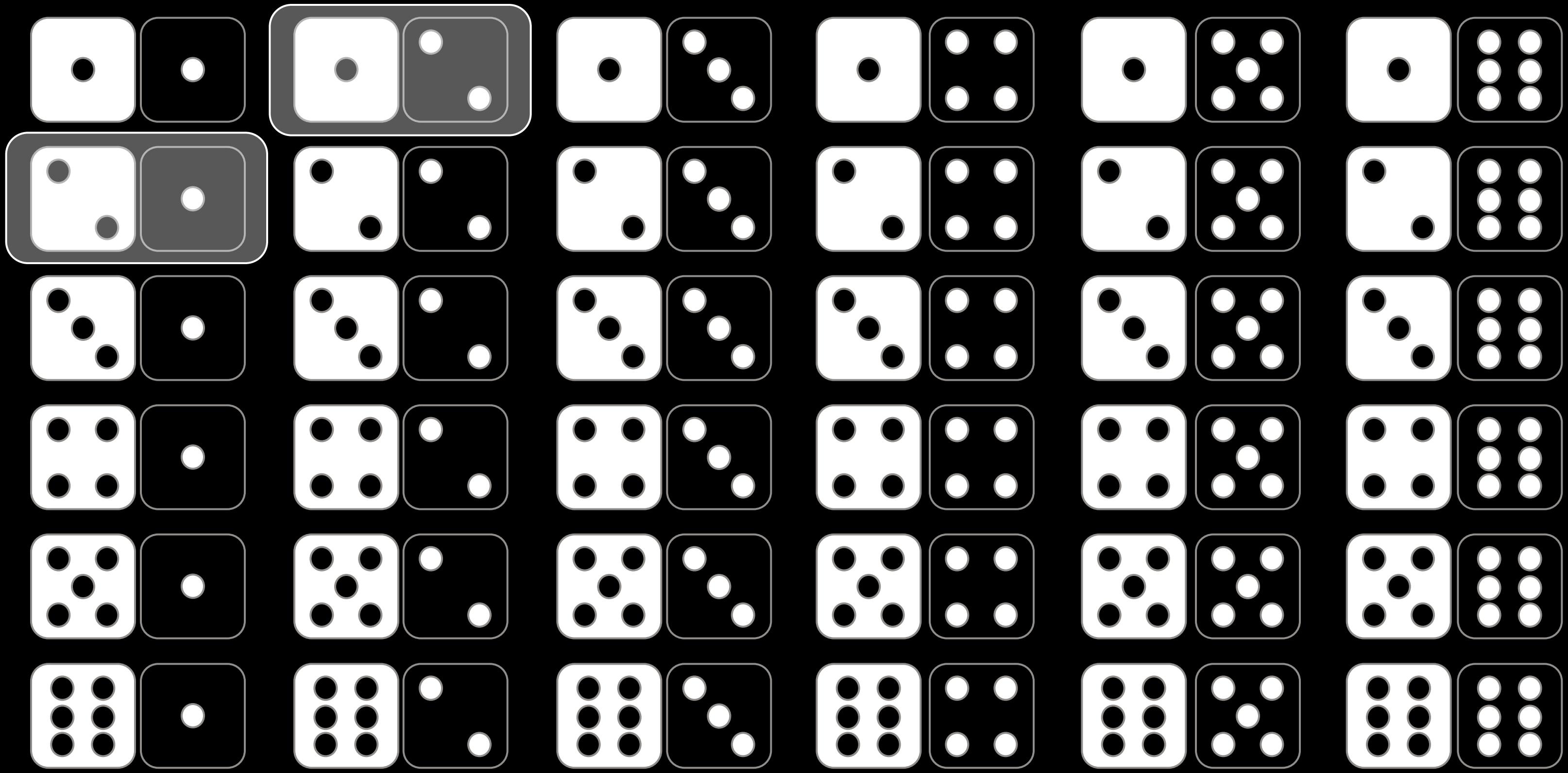
  = 1/6 + 1/6 - 1/36 = 11/36 = 0.30555…
- P(x OR y) = P(x) + P(y) - P(x AND y)
  - Known as the *Addition Rule*
- If and only if the two events are *mutually exclusive*, which is just another way of saying P(x AND y) = 0.0, then we get the special case
  - P(x OR y) = P(x) + P(y)

Joint Probability

- The *joint probability* of two events is just the probability of both events occurring (at the same time)
  - It's the thing that is zero when the events are mutually exclusive
- P(x,y) = P(x AND y)
  - Also written as $P(x \cap y)$
- In our example (a = white one, b = black one), then
  - P(a,b) = P(a AND b) = 1/36 = 0.02777…
    - Snake eyes!

Conditional Probability

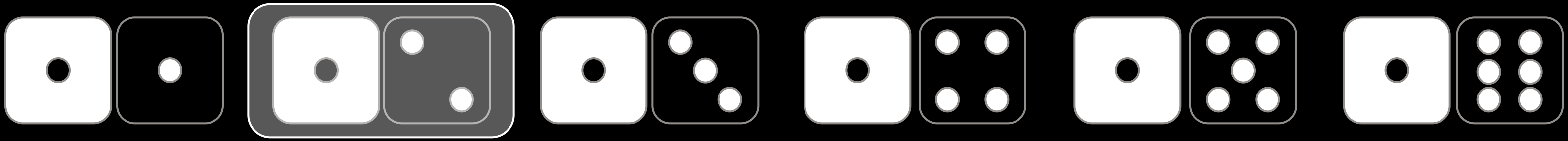
- Let event c = the dice sum to three



- P(c) = 2/36 = 1/18 = 0.0555…

Conditional Probability

- Suppose we change the rules and throw the dice one at a time, first white, then black
- This obviously makes no difference before any dice are rolled
- However, suppose we have now rolled the first die and it has come up one (event a), then our possible elementary outcomes are reduced to:



- P(c|a) = 1/6 = 0.1666…

Conditional Probability

- $P(x|y) = P(x \text{ AND } y) / P(y)$
  - In our example, $P(c|a) = P(c \text{ AND } a) / P(a)$

    $= (1/36) / (1/6)$

    $= 1 / 6$
- $P(x \text{ AND } y) = P(x|y)\,P(y)$
  - Known as the *Multiplication Rule*
- $P(x \text{ AND } y) \equiv P(y \text{ AND } x)$
  - $P(x|y)\,P(y) = P(y|x)\,P(x)$

Statistical Independence

- If x and y are statistically *independent*, then
  - P(x AND y) = P(x) P(y)
    - P(x|y) = P(x)
    - P(y|x) = P(y)
- From our first dice example, of rolling two ones
  - P(a) = P(b) = 1/6
  - P(a AND b) = 1/36
  - P(a|b) = P(a AND b) / P(b)
    = (1/36) / (1/6) = 1/6 = P(a)
  - The two events, a and b, are independent
- From our sequential dice example, rolling a three
  - P(c) = 1/18
  - P(c|a) = 1/6 ≠ P(c)
  - The two events, a and c, are not independent

Marginal Probability

- Marginal probability is the probability of one event, ignoring any information about other events
  - The marginal probability of event x is just P(x)
  - The marginal probability of event y is just P(y)
- If knowledge is specified in terms of conditional probabilities or joint probabilities, then marginal probabilities may be computed by summing over the ignored event(s)

$$P(x) = \sum_j p(x,y_j) = \sum_j p(x|y_j)\, p(y_j)$$

Enough Probability, But What of de Mere?

- What is the probability of rolling a six in four rolls of a single die (call this event S)?
  - Let event d_i = a die shows a six on the ith roll
  - $P(S) = P(d_1 \text{ OR } d_2 \text{ OR } d_3 \text{ OR } d_4)$
    $= P((d_1 \text{ OR } d_2) \text{ OR } (d_3 \text{ OR } d_4))$
  - $P(d_1 \text{ OR } d_2) = P(d_1) + P(d_2) - P(d_1 \text{ AND } d_2)$
    $= 1/6 + 1/6 - 1/36 = 11/36$
    $= P(d_3 \text{ OR } d_4))$
  - $P(S) = 11/36 + 11/36 - (11/36)^2$
    $= 0.517747$
  - Let event e_i = NOT d_i (a six does not show)
  - $P(\text{NOT } S) = P(e_1 \text{ AND } e_2 \text{ AND } e_3 \text{ AND } e_4)$
    $= (5/6)^4 = 0.482253$ (statistically independent)
  - $P(S) = 1.0 - P(\text{NOT } S) = 0.517747$

Wanna Bet?

- What is the probability of rolling a pair of sixes in twenty-four rolls of a pair of dice (call this event T)?
  - Let event f_i = dice show a pair of sixes on the i^{th} roll
  - $P(T) = P(f_1 \text{ OR } f_2 \dots \text{ OR } f_{24})$
    $= P((f_1 \text{ OR } f_2) \text{ OR } (f_3 \text{ OR } f_4) \dots \text{ OR } (f_{23} \text{ OR } f_{24}))$
  - … could do it, but entirely too painful …
  - Let event g_i = NOT f_i (a pair of sixes does not show)
  - $P(\text{NOT } T) = P(g_1 \text{ AND } g_2 \dots \text{ AND } g_{24})$
    $= (35/36)^{24} = 0.508596$
  - $P(T) = 1.0 - P(\text{NOT } T) = 0.491404$ $P(S) = 0.517747$
- So P(S) > P(T), and de Mere's observation that he lost more often when he bet on double sixes than when he bet on single sixes was remarkably astute

Information Theory (finally)

- Claude E. Shannon also called it "communication theory"
- The theory was developed and published as "The Mathematical Theory of Communication" in the July and October 1948 issues of the *Bell System Technical Journal*
- Shannon's concerns were clearly rooted in the communication of signals and symbols in a telephony system, but his formalization was so rigorous and general that it has since found many applications
- He was aware of similarities and concerned about differences with thermodynamic entropy, but was encouraged to adopt the term by Von Neumann, who said, "Don't worry. No one knows what entropy is, so in a debate you will always have the advantage."

Entropy

- Physicist Edwin T. Jaynes identified a direct connection between Shannon entropy and physical entropy in 1957
- Ludwig Boltzmann's grave is embossed with his equation:

  S = k log W

  Entropy = Boltzmann's-constant

  * log(function of # of possible micro-states)
- Shannon's measure of information (or uncertainty or entropy) can be written:

  I = K log Ω

  Entropy = constant (usually dropped)

  * log(function of # of possible micro-states)

Energy -> Information -> Life

- John Avery (*Information Theory and Evolution*) relates physical entropy to informational entropy as

  1 electron volt / kelvin = 16,743 bits
- So converting one electron-volt of energy into heat, at room temperature will produce an entropy change of

  1 electron volt / 298.15 kelvin = 56.157 bits
- Thus energy, such as that which washes over the Earth from the Sun, can be seen as providing a constant flow of not just “free energy”, but free information
- Living systems take advantage of, and encode this information, temporarily and locally reducing the conversion of energy into entropy
  - Brains encode rapidly changing information in neural structures
  - Genes encode slowly changing information in DNA

History, As Always

- Samuel F.B. Morse worried about letter frequencies when designing (both versions of) the Morse code (1838)
  - Made the most common letters use the shortest codes
  - Obtained his estimate of letter frequency by counting the pieces of type in a printer's type box
  - Observed transmission problems with buried cables
- William Thompson, aka Lord Kelvin, Henri Poincaré, Oliver Heaviside, Michael Pupin, and G.A. Campbell all helped formalize the mathematics of signal transmission, based on the methods of Joseph Fourier (mid to late 1800's)
- Harry Nyquist published the Nyquist Theorem in 1928
- R.V.L. Hartley published "Transmission of Information" in 1928, containing a definition of information that is the same as Shannon's for equiprobable, independent symbols

History

- During WWII, A.N. Kolmogoroff, in Russia, and Norbert Weiner, in the U.S., devised formal analyses of the problem of extracting signals from noise (aircraft trajectories from noisy radar data)
- In 1946 Dennis Gabor published “Theory of Communication”, which addressed related themes, but ignored noise
- In 1948 Norbert Wiener published *Cybernetics*, dealing with communication and control
- In 1948 Shannon published his work
- In 1949 W.G. Tuller published “Theoretical Limits on the Rate of Transmission of Information” that parallels Shannon’s work on channel capacity

Stochastic Signal Sources

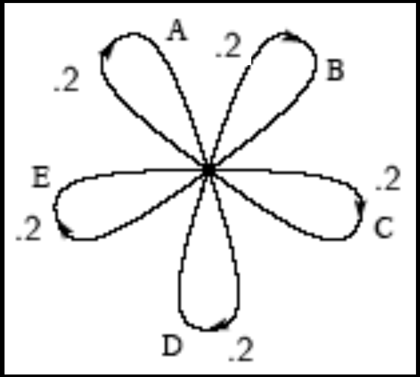
- Suppose we have a set of 5 symbols—the English letters A, B, C, D, and E
- If symbols from this set are chosen with equal probability (0.2), you would get something like:

B D C B C E C C C A D C B D D A A E C E E

A A B B D A E E C A C E E B A E E C B C E

A D

- This source may be represented as follows

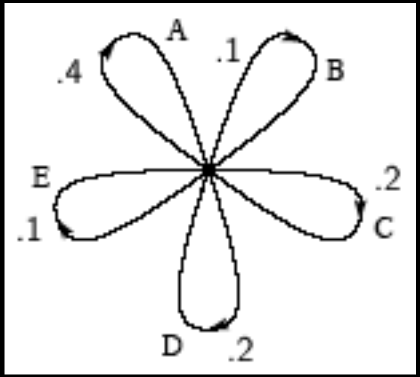


Stochastic Signal Sources

- If the same symbols (A, B, C, D, E) are chosen with uneven probabilities 0.4, 0.1, 0.2, 0.2, 0.1, respectively, one obtains:

```
A A A C D C B D C E A A D A D A C E D A E
A D C A B E D A D D C E C A A A A A D
```

- This source may be represented as follows

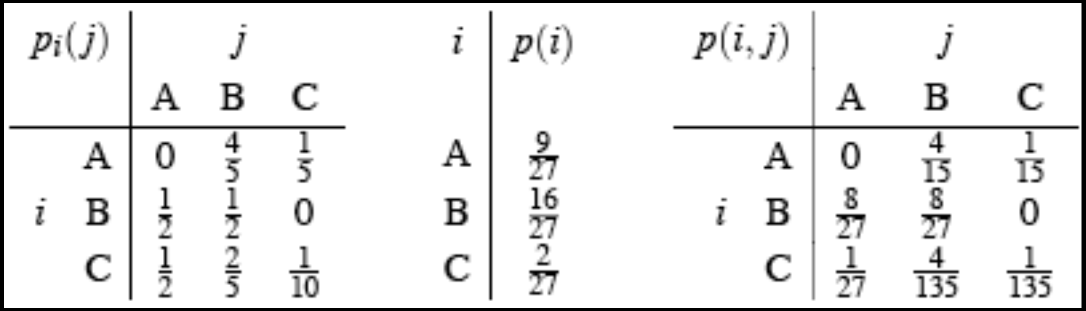


Stochastic Signal Sources

- More complicated models are possible if we base the probability of the current symbol on the preceding symbol, invoking conditional and joint probabilities
- E.g., if we confine ourselves to three symbols, A, B, and C, with the following probability tables

Transition (Conditional) Probabilities

P(j|i)



$p_i(j)$		j		
		A	B	C
	A	0	$\frac{4}{5}$	$\frac{1}{5}$
i	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

i	$p(i)$
A	$\frac{9}{27}$
B	$\frac{16}{27}$
C	$\frac{2}{27}$

$p(i,j)$		j		
		A	B	C
	A	0	$\frac{4}{15}$	$\frac{1}{15}$
i	B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

Digram (Bigram, Joint) Probabilities

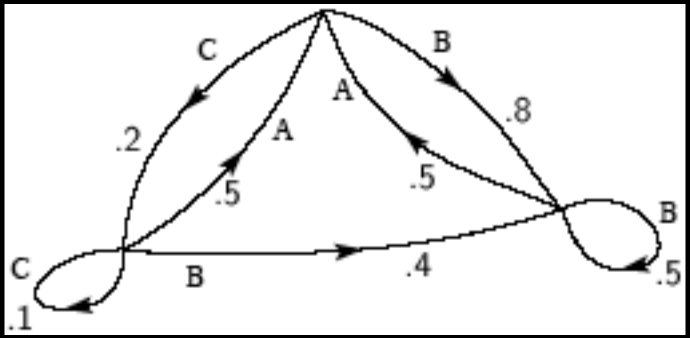
P(i AND j)

one might obtain

A B B A B A B A B A B A B A B B B A B B B
B B A B A B A B A B A B B B A C A C A B B
A B B B B A B B A B A C B B B A B A

Stochastic Signal Sources

- This source may be represented as follows



- Simple *bigrams* (or *digrams*) may, of course, be replaced with *trigrams* or arbitrary depth *n-grams*, if we choose to make the next symbol dependent on more and more history

Stochastic Signal Sources

- Symbols can also be words, not just letters
- Suppose that, based on our five letters, A, B, C, D, and E, we have a vocabulary of 16 “words” with associated probabilities:

.10	A	.16	BEBE	.11	CABED	.04	DEB
.04	ADEB	.04	BED	.05	CEED	.15	DEED
.05	ADEE	.02	BEED	.08	DAB	.01	EAB
.01	BADD	.05	CA	.04	DAD	.05	EE

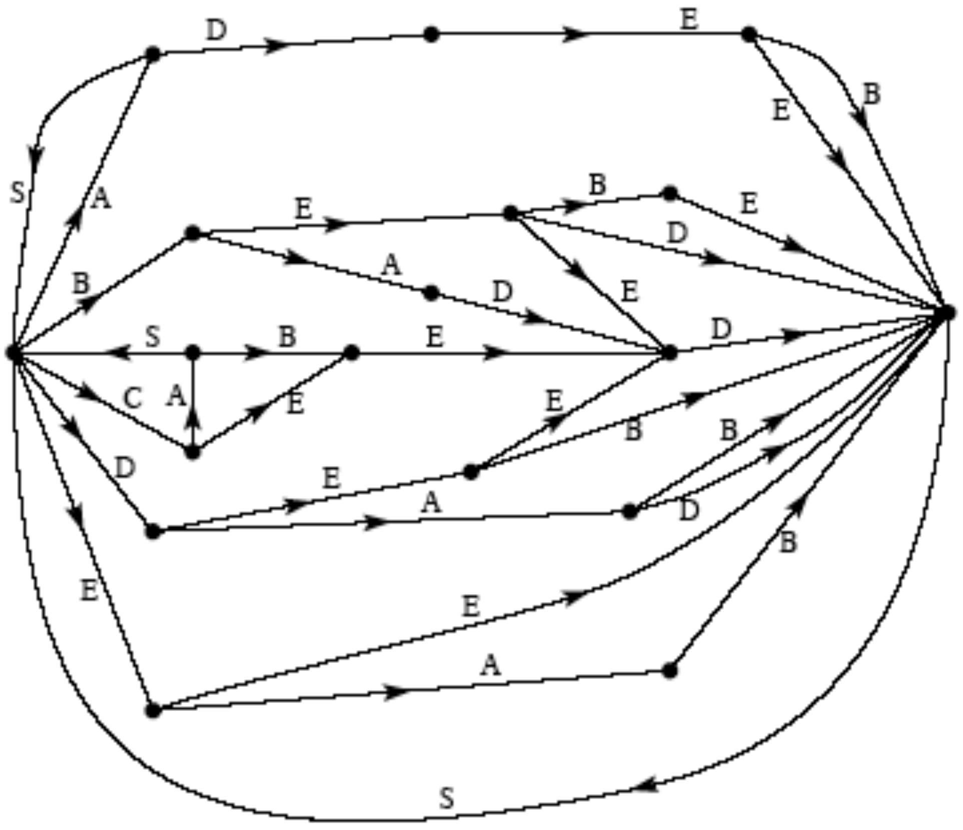
- If successive words are chosen independently and separated by a space, one might obtain:

```
DAB EE A BEBE DEED DEB ADEE ADEE EE DEB
BEBE BEBE BEBE ADEE BED DEED DEED CEED
ADEE A DEED DEED BEBE CABED BEBE BED DAB
DEED ADEB
```

- And again one could introduce transition probabilities

Stochastic Signal Sources

- This source may be represented as follows



Approximations of English (Letters)

- Assume we have a set of 27 symbols—the English alphabet plus a space
- A zero-order model of the English language might then be an equiprobable, independent sequence of these symbols:

```
XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD
```

- A first-order approximation, with independent symbols, but using letter frequencies of English text might yield:

```
OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH
EEI ALHENHTTPA OOBTTVA NAH BRL
```

Approximations of English (Letters)

- A second-order approximation using bigram probabilities from English text might yield:

```
ON IE ANTSOUTINYS ARE T INCTORE ST BE S
DEAMY ACHIN D ILONASIVE TUCOOWE AT
TEASONARE FUSO TIZIN ANDY TOBE SEACE
CTISBE
```

- A third-order approximation using trigram probabilities from English text might yield:

```
IN NO IST LAT WHEY CRATICT FROURE BIRS
GROCID PONDENOME OF DEMONSTURES OF THE
REPTAGIN IS REGOACTIONA OF CRE
```

Approximations of English (Words)

- A first-order word approximation, choosing words independently, but with their appropriate frequencies:

  REPRESENTING AND SPEEDILY IS AN GOOD APT
  OR COME CAN DIFFERENT NATURAL HERE HE THE
  A IN CAME THE TOOF TO EXPERT GRAY COME TO
  FURNISHES THE LINE MESSAGE HAD BE THESE

- A second-order word approximation, using bigram word transition probabilities (but no other grammatical structure):

  THE HEAD AND IN FRONTAL ATTACK ON AN
  ENGLISH WRITER THAT THE CHARACTER OF THIS
  POINT IS THEREFORE ANOTHER METHOD FOR THE
  LETTERS THAT THE TIME OF WHO EVER TOLD
  THE PROBLEM FOR AN UNEXPECTED

- Note that there is reasonably good structure out to about twice the range that is used in construction

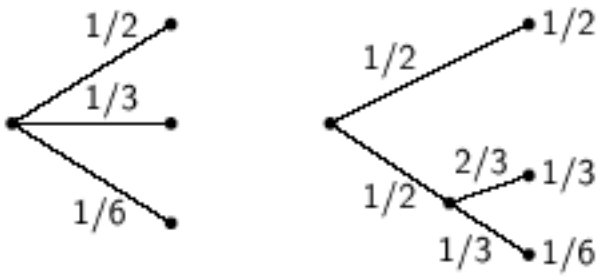
Information in Markov Processes

- The language models just discussed and many other symbol sources can be described as Markov processes (stochastic processes in which future states depend solely on the current state, and not on how the current state was arrived at)
- Can we define a quantity that measures the information produced by, or the information rate of, such a process?
- Let's say that *the information produced by a given symbol is exactly the amount by which we reduce our uncertainty about that symbol when we observe it*
- We therefore now seek a measure of uncertainty

Uncertainty

- Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$
- Say these probabilities are known, but that is all we know concerning which event will occur next
- What properties would a measure of our uncertainty, $H(p_1, p_2, \ldots, p_n)$, about the next symbol require:
  1) H should be continuous in the p_i
  2) If all the p_i are equal ($p_i = 1/n$), then H should be a monotonic increasing function of n
     - With equally likely events, there is more choice, or uncertainty, when there are more possible events
  3) If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H

Uncertainty



- On the left, we have three possibilities:
  $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/6$
- On the right, we first choose between two possibilities:
  $p_1 = 1/2$, $p_2 = 1/2$
  and then on one path choose between two more:
  $p_3 = 2/3$, $p_4 = 1/3$
- Since the final probabilities are the same, we require:
  $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2\ H(2/3, 1/3)$

Entropy

- In a proof that explicitly depends on this decomposibility and on monotonicity, Shannon establishes
  - *Theorem 2: The only H satisfying the three above assumptions is of the form:*

    $$H = - K \sum_{i=1}^{n} p_i \log p_i$$

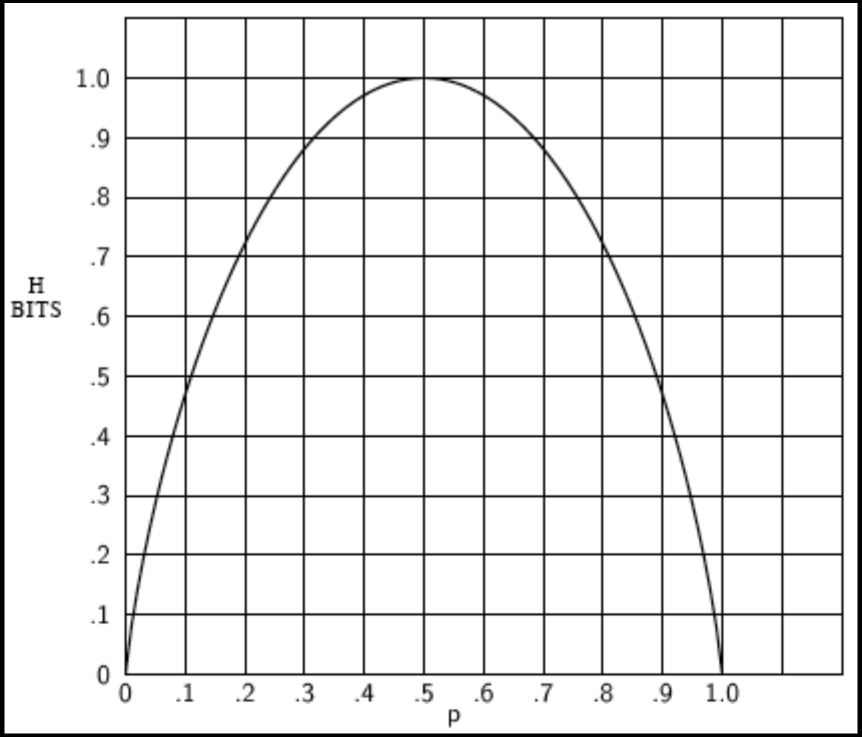
    where K is a positive constant
- Observing the similarity in form to entropy as defined in statistical mechanics, Shannon dubbed H the entropy of the set of probabilities $p_1, p_2, \ldots, p_n$
- Generally, the constant K is dropped; Shannon explains it merely amounts to a choice of unit of measure

Behavior of the Entropy Function

- In the simple case of two possibilities with probability p and q = 1 - p, entropy takes the form

  $$H = -(p \log p + q \log q)$$

  and is plotted here as a function of p:



Behavior of the Entropy Function

- In general, H = 0 if and only if all the p_i are zero, except one which has a value of one
- For a given n, H is a maximum (and equal to log n) when all p_i are equal (1/n)
  - Intuitively, this is the most uncertain situation
- Any change toward equalization of the probabilities p_1, p_2, …, p_n increases H
  - If $p_i \neq p_j$, adjusting p_i and p_j so they are more nearly equal increases H
  - Any “averaging” operation on the p_i increases H

Joint Entropy

- For two events, x and y, with m possible states for x and n possible states for y, the entropy of the joint event may be written in terms of the joint probabilities

  $$H(x,y) = -\sum_{i,j} p(x_i,y_j) \log p(x_i,y_j)$$

  while

  $$H(x) = -\sum_{i,j} p(x_i,y_j) \log \sum_{j} p(x_i,y_j)$$

  $$H(y) = -\sum_{i,j} p(x_i,y_j) \log \sum_{i} p(x_i,y_j)$$

- It is "easily" shown that

  $$H(x,y) \leq H(x) + H(y)$$

  - Uncertainty of a joint event is less than or equal to the sum of the individual uncertainties
  - Only equal if the events are independent
    - $p(x,y) = p(x)\,p(y)$

Conditional Entropy

- Suppose there are two chance events, x and y, not necessarily independent. For any particular value x_i that x may take, there is a conditional probability that y will have the value y_j, which may be written

$$p(y_j|x_i) = p(x_i,y_j) \;/\; \sum_j p(x_i,y_j) = p(x_i,y_j) / p(x_i)$$

- Define the *conditional entropy* of y given x, H(y|x), as the average of the entropy of y given each value of x, weighted according to the probability of getting that particular x

$$H(y|x) = - \sum_{i,j} p(x_i)\, p(y_j|x_i) \log p(y_j|x_i)$$

$$H(y|x) = - \sum_{i,j} p(x_i,y_j) \log p(y_j|x_i)$$

  - This quantity measures, on the average, how uncertain we are about y when we know x

Joint, Conditional, & Marginal Entropy

- Substituting for $p(y_j|x_i)$, simplifying, and rearranging yields

  $$H(x,y) = H(x) + H(y|x)$$

  - The uncertainty, or entropy, of the joint event x, y is the sum of the uncertainty of x plus the uncertainty of y when x is known

- Since $H(x,y) \leq H(x) + H(y)$, and given the above, then

  $$H(y) \geq H(y|x)$$

  - The uncertainty of y is never increased by knowledge of x
    - It will be decreased unless x and y are independent, in which case it will remain unchanged

Maximum and Normalized Entropy

- *Maximum entropy*, when all probabilities are equal is

  $H_{Max} = \log n$

- Normalized entropy is the ratio of entropy to maximum entropy

  $H_o(x) = H(x) / H_{Max}$

- Since entropy varies with the number of states, n, normalized entropy can be a better way of comparing across systems
- Shannon called this *relative entropy*
- (Some cardiologists and physiologists call entropy divided by total signal power normalized entropy)

Mutual Information

- Define *Mutual Information* (aka *Shannon Information Rate*) as
  $$I(x,y) = \sum_{i,j} p(x_i,y_j) \log [\ p(x_i,y_j) \ / \ p(x_i)p(y_j)\ ]$$
- When x and y are independent $p(x_i,y_j) = p(x_i)p(y_j)$, so $I(x,y)$ is zero
- When x and y are the same, the mutual information of x,y is the same as the information conveyed by x (or y) alone, which is just $H(x)$
- Mutual information can also be expressed as
  $$I(x,y) = H(x) - H(x|y) = H(y) - H(y|x)$$
- Mutual information is nonnegative
- Mutual information is symmetric; i.e., $I(x,y) = I(y,x)$

Probability and Uncertainty

- Marginal

  $p(x)$ $\qquad H(x) = -\sum_{i} p(x_i) \log p(x_i)$

- Joint

  $p(x,y)$ $\qquad H(x,y) = -\sum_{i,j} p(x_i,y_j) \log p(x_i,y_j)$

- Conditional

  $p(y|x)$ $\qquad H(y|x) = -\sum_{i,j} p(x_i,y_j) \log p(y_j|x_i)$

- Mutual

  $$I(x,y) = \sum_{i,j} p(x_i,y_j) \log [\, p(x_i,y_j) / p(x_i)p(y_j) \,]$$

Credits

- Die photo on slide 3 from budgetstockphoto.com
- Penny photos on slide 4 from usmint.gov
- Some organization and examples of basic probability theory taken from Larry Gonick's excellent *The Cartoon Guide to Statistics* http://www.amazon.com/exec/obidos/ASIN/0062731025/
- Some historical notes are from John R. Pierce's *An Introduction to Information Theory* http://www.amazon.com/exec/obidos/ASIN/0486240614/
- Physical entropy relation to Shannon entropy and energy-to-information material derived from John Avery's Information Theory and Evolution http://www.amazon.com/exec/obidos/ASIN/9812384006/
- Information theory examples from Claude E. Shannon's *The Mathematical Theory of Communication* http://www.amazon.com/exec/obidos/ASIN/0252725484/

References

- http://www.umiacs.umd.edu/users/resnik/nlstat_tutorial_summer1998/Lab_ngrams.html
- http://www-2.cs.cmu.edu/~dst/Tutorials/Info-Theory/
- http://szabo.best.vwh.net/kolmogorov.html
- http://www.saliu.com/theory-of-probability.html
- http://www.wikipedia.org/