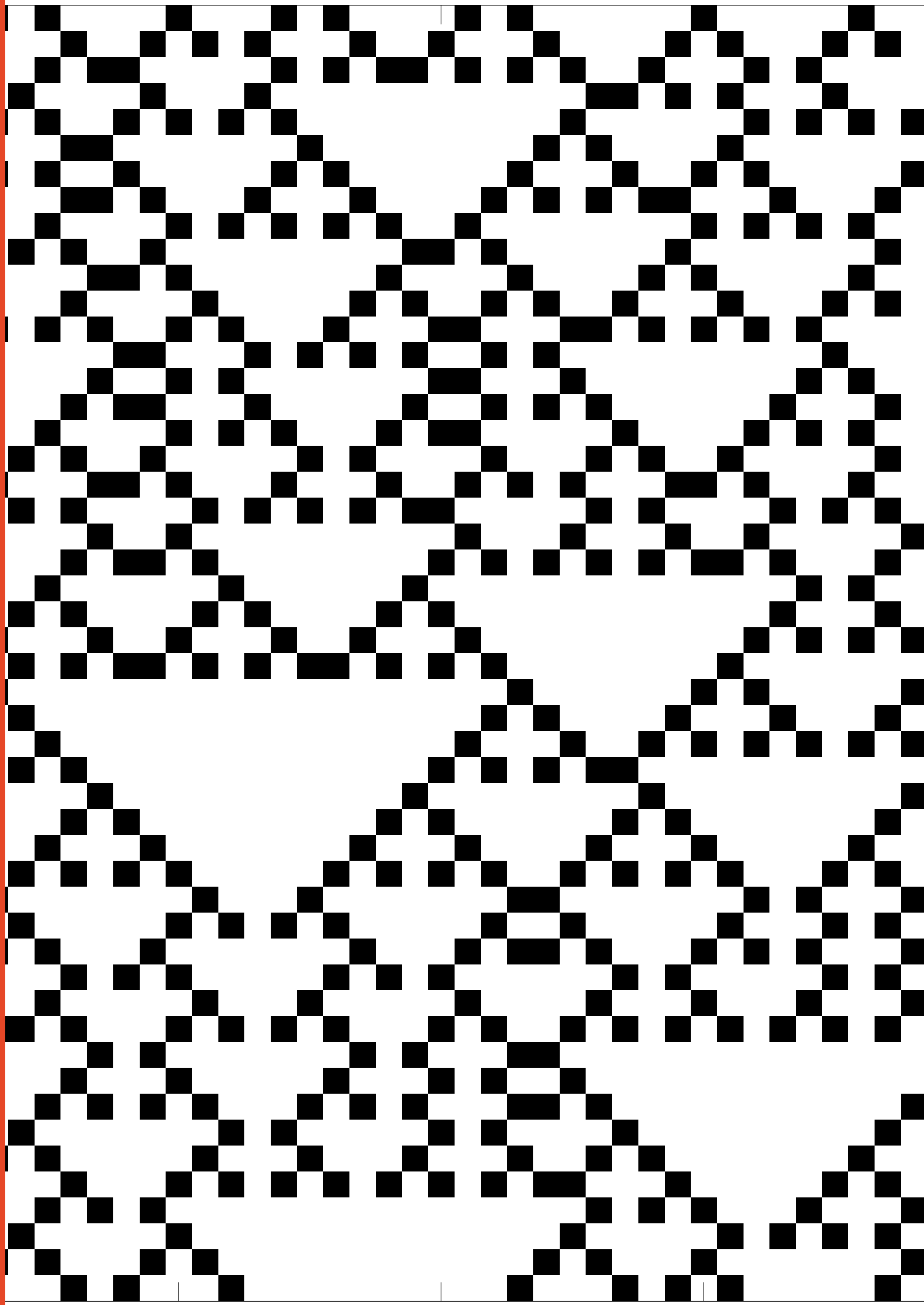
# Introduction to Information Theory

Dr. Joseph Lizier

THE UNIVERSITY OF SYDNEY



# Information, order and randomness

- Recall where we have come across concepts of information:
  - Information processing in biological systems

- Recall where we have come across concepts of order and randomness
  - Concept of “Edge of chaos”
  - Self-organisation as an increase in order over time (without external control) – Sayama p.6
  - Emergence as increase in order over scale – Sayama p.6

H. Sayama, “*Introduction to the Modeling and Analysis of Complex Systems*”, Geneseo, NY: Open SUNY Textbooks, 2015; chapter 1

# Main implications for complex systems (so far)

- Ordered systems (fewer outcomes at system level) have less uncertainty, less information
- Disordered or random systems (more outcomes at system level) have more uncertainty, more information

- Lower probability states are more surprising, carry more information

- We've talked on and off about how complex systems process information, but we don't yet know how to measure that…

# Information, order and randomness

– "*Although they (complex adaptive systems) differ widely in their physical attributes, they resemble one another in the way they handle information. That common feature is perhaps the best starting point for exploring how they operate.*"

  Murray Gell-Mann

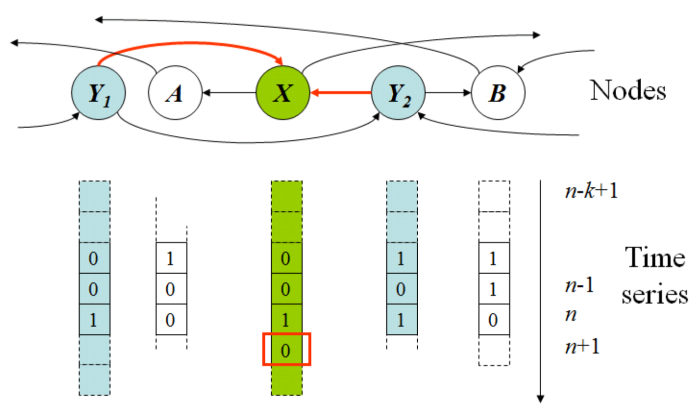
– In order to quantity these key concepts, we turn to information theory

M. Gell-Mann, The Quark and the Jaguar. New York: W.H. Freeman, 1994
in
M. Mitchell, "*Complexity: A guided tour*", New York: Oxford University Press, 2009

# Learning outcomes

1. **Understand** basic information-theoretic measures, and advanced measures for time-series, and how to use these to **analyse** and **dissect** the nature, structure, function and evolution of complex systems.
2. Develop scientific programming skills which can be **applied** in complex system analysis and design.
3. To be able to **understand** the design of and to extend the **design** of a piece of software using techniques from class and your own readings.
4. Ability to **apply** and make informed decisions in selecting and using information-theoretic measures and software tools to analyse complex systems.
5. Ability to **create** information-theoretic analyses of real-world data sets, in particular in a student's domain area of expertise.
6. Capacity to **critically evaluate** investigations of self-organisation and relationships in complex systems using information theory, and the insights provided.

$$\mathrm{H}(X) = -\sum_{x \in A_x} p(x) \log_2 p(x)$$




JIDT

# Information theory: what will cover

## Lectures/activities

1. Introduction to information theory and entropy
2. What is information?
3. Introduction to JIDT
4. Information-theoretic estimators and JIDT
5. Statistical significance and undersampling
6. Information theory and self-organisation
7. Information processing in complex systems
8. Information storage
9. Information transfer
10. Effective network inference

## Resources:

– Texts: Cover and Thomas, Mackay, Bossomaier et al., Lizier (JIDT)
– Software: JIDT JIDT

# Introduction to Information Theory: session outcomes

- Ability to express ideas about uncertainty and information.
- Understand fundamental measures of information theory including: entropy, joint entropy, conditional entropy.
- Ability to partially construct Matlab code to compute such measures, and apply that code to examples.

- Primary references:
  - Cover and Thomas, "Elements of Information Theory", Hoboken, New Jersey: John Wiley and Sons, Inc., 2006 (2nd ed.); chapter 2 (up to and including section 2.2 only)
  - Mackay, "Information Theory, Inference, and Learning Algorithms", Cambridge: Cambridge University Press, 2003; sections 2.4-2.5, 8.1 (up to first mention of mutual information with equation 8.8).
  - Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; Chapter 3, up to and including section 3.2.1.
  - Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics and AI, 1:11, 2014; Appendix A.1 and A.3 (up to first mention of mutual information in both)

# What is information?

– You tell me …

# A game about information: Guess Who? (Hasbro)

1. How does it work?
   a. Game board / rules
   b. Play yourself (character sheets, e.g. sports)
2. Who wants to play?

3. What did we learn from this game?
   a. What are the best/worst questions to ask or strategies?
   b. What types of information did we encounter?

# What is information theory?

– An approach to quantitatively capture the notion of information.
– Traditionally, information theory provides answers to two fundamental questions (Cover and Thomas, 1991):
  1. What is the ultimate data compression?
     - *How small can I zip up a file?*
  2. What is the ultimate transmission rate of communication?
     - *What is my max download speed at home?*

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991.
D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003

# What is information theory

– It's also about far more than these traditional areas:

How do complex systems process information?

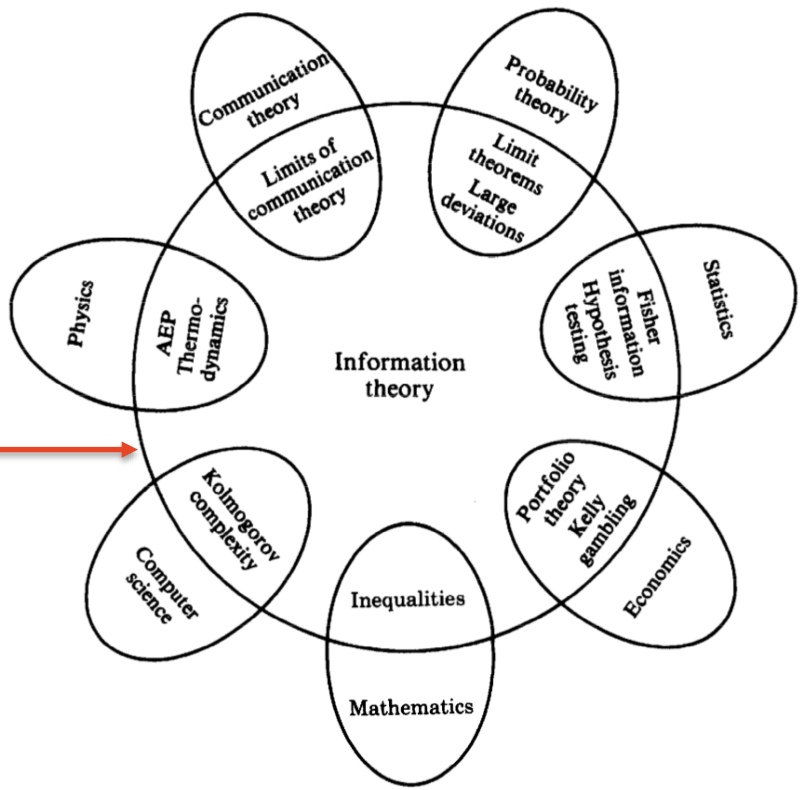



**Figure 1.1.** The relationship of information theory with other fields.

Image from Cover and Thomas (1991)

# Defining information – first pass

- JL: "*Information is all about questions and answers*"

- Information is the amount by which
  - one variable (an answer/signal/measurement)
  - reduces our uncertainty or surprises us
  - about another variable.

- We need to quantify both:
  - Uncertainty (entropy)
  - Uncertainty reduction (information)

- This was quantified by Claude Shannon


C. E. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27(3–4):379–423, 623–656, 1948.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991.

D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003

# Information is measured in bits

– 1 bit is the amount of uncertainty about (*an equiprobable*) yes/no question.

– The answer to that question provides 1 bit of uncertainty reduction or information.

– E.g. *What will my next coin flip be, heads or tails?*

# Quantifying information – preliminaries

- X is a random variable
  - A variable whose value is subject to chance.
  - i.e. an answer/signal/measurement
  - e.g. result of a coin flip, whether it rains today, etc.
- x is a sample or outcome or measurement of X
  - drawn from some discrete alphabet $A_X = \{x_1, x_2, \dots\}$
  - For binary X , $A_X = \{0, 1\}$
  - For a coin toss, $A_X = \{\text{heads}, \text{tails}\}$
  - For hair colour in Guess Who?, $A_X = \{?\}$
- We have probability distribution function (PDF) defined: $p(x) = Pr(X = x),\ x \in A_X$
  - $0 \le p(x) \le 1,\ \forall x \in A_X$
  - $\sum_{x \in A_X} p(x) = 1$

For background, see:
Bossomaier et al., "An introduction to transfer entropy: Information flow in Complex Systems", Springer, Cham, 2016; chapter 2.

# Shannon information content

- The *fundamental* quantity of information theory
- Shannon information content of a sample or outcome x:

$$h(x) = \log_2\left(\frac{1}{p(x)}\right)$$

- Units are bits for log in base 2.
- Is a measure of surprise at the value of this sample or outcome x given p(x):
  - $h(x) \geq 0$
  - *No surprise* if there is only ever one outcome $p(x) = 1$;
  - There is always some level of surprise if there exists more than one outcome with $p(x) > 0$
  - Our *surprise increases* as x becomes less likely;

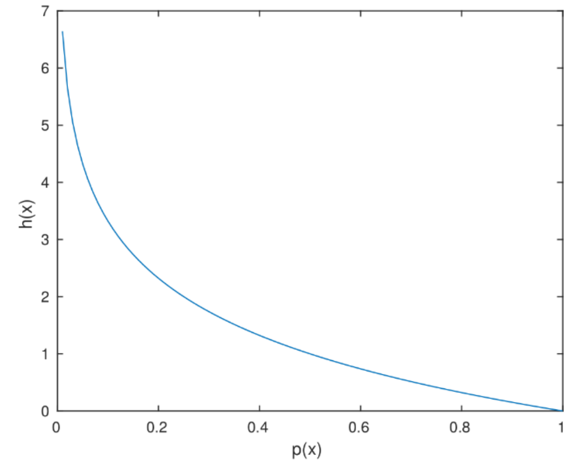
D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003; section 2.4

T. Bossomaier et al., “An Introduction to Transfer Entropy: Information flow in Complex Systems”, Springer, Cham, 2016; chapter 3

# Shannon information content

- Shannon information content of a sample or outcome x:

$$h(x) = \log_2\left(\frac{1}{p(x)}\right)$$
$$h(x) = -\log_2 p(x)$$




- Exercise:
  1. Edit the Matlab function `infocontent(p)` to return the Shannon information content for an outcome x with probability p(x)
  2. Compute using your function:
     a. h(heads) for a fair coin?
     b. h(1) for a 6-sided die? h(not 1) for a 6-sided die?
     c. h(1) for a 20-sided die? h(not 1) for a 20-sided die?
  3. Reproduce the above plot of h(x) versus p(x) (hint: input p as a vector across the range 0.01:0.01:1)

# Shannon information content

- Shannon information content of a sample or outcome x:

$$h(x) = \log_2\left(\frac{1}{p(x)}\right)$$
$$h(x) = -\log_2 p(x)$$

- Examples – Guess Who? (original version):
  - h(`alex`)? $\log_2\left(\frac{1}{1/24}\right)$=4.585 bits
  - h(`female`)? $\log_2\left(\frac{1}{5/24}\right)$=2.263 bits
  - h(`male`)? $\log_2\left(\frac{1}{19/24}\right)$=0.337 bits
  - Is "*female*?" a good question to ask first?
  - Is "*alex*?" a good question to ask first?

# (Shannon) entropy

- Shannon entropy of a random variable X:

$$\mathrm{H}(X) = \sum_{x \in A_x} p(x) \log_2 \frac{1}{p(x)}$$

$$\mathrm{H}(X) = \sum_{x \in A_x} -p(x) \log_2 p(x)$$

$$\mathrm{H}(X) = \langle h(x) \rangle$$

- ***Expectation value*** of Shannon information content. *H(X)* $\geq$ *0*
- Measures our uncertainty of the answer to our question
- $p \log p \rightarrow 0$ in limit as $p \rightarrow 0$
- Examples:
  - If $\exists x, p(x) = 1 \ \rightarrow H(X) = 0$
  - For binary X, $p(0) = 0.5, p(1) = 0.5 \ \rightarrow H(X) = 1$ bit
  - $p(x) = \frac{1}{|A_X|}, \forall x \rightarrow H(X) = \log_2 |A_X|$ bits

# (Shannon) entropy

– Shannon entropy of a random variable X:

$$\mathrm{H}(X) = \sum_{x \in A_x} -p(x) \log_2 p(x)$$

– Exercise: Let's code it!

1. For a binary X with $p_1 = P(X = 1)$:
   - $\mathrm{H(X)} = -p_1 \log_2 p_1 - (1 - p_1) \log_2 (1 - p_1)$
2. Edit the Matlab function `entropy(p)` to return the Shannon entropy for the given distribution over outcomes *x* of *X*
   a. Input is p as a vector
   b. How to sum over *x*?
   c. What are some possible error conditions here?
   d. Test: entropy([0.5, 0.5]), entropy([1, 0])
   e. Plot H(X) as a function of P(X=1) for binary *X*.

# (Shannon) entropy

– Shannon entropy of a random variable X:

$$\mathrm{H}(X) = \sum_{x \in A_X} -p(x) \log_2 p(x)$$

– Examples: *Guess Who?* (verify with your code)

1. H(`who`)? 4.585 bits
2. H(`sex`)? p(`male`) x h(`male`)+p(`female`) x h(`female`) = 0.738 bits
3. Value of questions:
   - Is "*female?*" a good question to ask first?
   - Is "*alex?*" a good question to ask first?
   - What is the best question to ask first, and why?

# (Shannon) entropy – empirical data

– Shannon entropy of a random variable X:

$$\mathrm{H}(X) = \sum_{x \in A_X} -p(x)\log_2 p(x)$$

– Exercise: Let's code it assuming we don't have *P*(*X*) given, but are given empirical data to compute *P*(*X*) from:

1. Edit the Matlab function `entropyempirical(xn)` to return the Shannon entropy for the given samples $x_n$ of $X$ (n is sample index)
   a. Input is $x_n$ as a vector of samples – work out $A_X$ from the $x_n$
   b. How to compute $p(x)$ for each outcome $x$?
   c. Test: `entropyempirical([0,0,1,1])` (should be 1)
   d. Test: create a vector of inputs from 10 coin tosses that you do
   e. Test: create vectors of inputs from random data, e.g. `randi(2, 1, 10)`

# Meaning of entropy, and traditional usage

- Using an optimal compression or encoding scheme given $p(x)$:
  - $h(x)$ is the number of bits for a symbol to communicate x
  - $H(X)$ is the number of bits to communicate the x *on average*.
- Or (in bits): how *few* yes/no questions would I need to ask (on average) to determine the value of $x$?

- Think about *Guess who?* as a *decoding* task

# Meaning of entropy, and traditional usage

- Using an optimal compression or encoding scheme given $p(x)$:
  - $h(x)$ is the number of bits for a symbol to communicate x
  - $H(X)$ is the number of bits to communicate the x *on average.*
- Or (in bits): how *few* yes/no questions would I need to ask (on average) to determine the value of x?

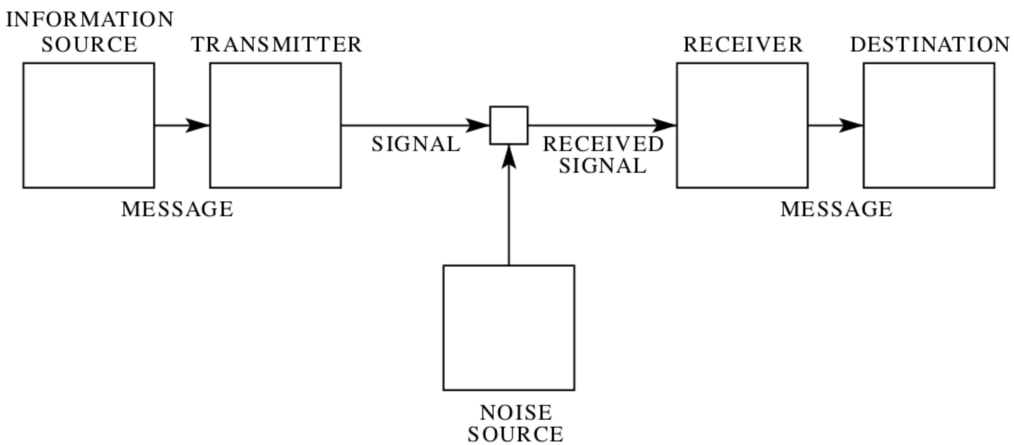


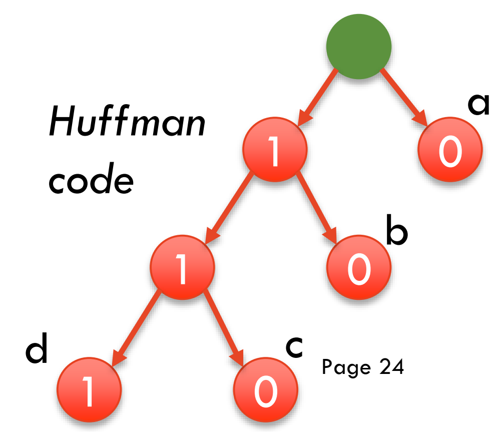

Fig. 1—Schematic diagram of a general communication system.

- What has information theory ever done for me? zip files, mp3s, encoding mobile telecoms / ADSL *etc.*

# Meaning of entropy, and traditional usage

- Using an optimal compression or encoding scheme given $p(x)$:
  - $h(x)$ is the number of bits for a symbol to communicate x
  - $H(X)$ is the number of bits to communicate the x *on average*.
- Example: say we want to communicate the result of a horse race with four horses $\{a, b, c, d\}$:
  - How many bits to encode each outcome?
    - Assume $p(x) = 0.25, \forall x$
      to give 2 bits. (max. entropy assumption)
  - Or: if $p(a) = 0.5, p(b) = 0.25, p(c) = p(d) = 0.125$?
    - $h(x)$ tells us to use 1 bit for a (say "0"), 2 bits for b (say "10") and 3 bits for c and d (say "110" and "111");
    - $H(X) = 1.75$ bits.
    - Using the actual $p(x)$ leads to more efficient coding
  - Information is not about meaning

*Huffman code*



# Entropy of text and compression

- Think about coding letters in English language text
- Can we get any insights into how many bits to use for each letter?[1]
- Look at entropy of alphabet in MacKay (2003)
- Meaning of a non-integer number of bits:
  - Encoding one sample at a time can only be done with an integer number of bits
  - To reach the lower limits suggested by information theory, we would need to use block coding (i.e. encoding multiple samples together)
  - But – we can do better still if we look at the entropy of these blocks …

| $i$ | $a_i$ | $p_i$ | $h(p_i)$ |
|---|---|---|---|
| 1 | a | .0575 | 4.1 |
| 2 | b | .0128 | 6.3 |
| 3 | c | .0263 | 5.2 |
| 4 | d | .0285 | 5.1 |
| 5 | e | .0913 | 3.5 |
| 6 | f | .0173 | 5.9 |
| 7 | g | .0133 | 6.2 |
| 8 | h | .0313 | 5.0 |
| 9 | i | .0599 | 4.1 |
| 10 | j | .0006 | 10.7 |
| 11 | k | .0084 | 6.9 |
| 12 | l | .0335 | 4.9 |
| 13 | m | .0235 | 5.4 |
| 14 | n | .0596 | 4.1 |
| 15 | o | .0689 | 3.9 |
| 16 | p | .0192 | 5.7 |
| 17 | q | .0008 | 10.3 |
| 18 | r | .0508 | 4.3 |
| 19 | s | .0567 | 4.1 |
| 20 | t | .0706 | 3.8 |
| 21 | u | .0334 | 4.9 |
| 22 | v | .0069 | 7.2 |
| 23 | w | .0119 | 6.4 |
| 24 | x | .0073 | 7.1 |
| 25 | y | .0164 | 5.9 |
| 26 | z | .0007 | 10.4 |
| 27 | - | .1928 | 2.4 |
| | $\sum_i p_i \log_2 \frac{1}{p_i}$ | | 4.1 |

Table 2.9. Shannon information contents of the outcomes a–z.

1. How to determine the coding to use is out of scope…

# Joint Entropy

- We can consider joint entropy of a multivariate, e.g. $\{X, Y\}$:

$$\mathrm{H}(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

$$\mathrm{H}(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} -p(x,y) \log_2 p(x,y)$$

$$\mathrm{H}(X,Y) = \langle h(x,y) \rangle$$

- Surprise $h(x,y)$ / Uncertainty $H(X,Y)$ for the joint sample $\{x,y\}$
- Same properties as for $H(X)$, only now $X$ is multivariate!
- $H(X,Y) \geq H(X)$
- We refer to H(X) or H(Y) as *marginal* entropies to distinguish them
- Is $H(X, Y) = H(X) + H(Y)$?
  - Only for *independent* variables where p(x, y) = p(x)p(y) !


T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.2

D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003. Section 8.1

# Joint Entropy

– We can consider joint entropy of a multivariate, e.g. $\{X, Y\}$:

$$\mathrm{H}(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} p(x,y) \log_2 \frac{1}{p(x,y)}$$

$$\mathrm{H}(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} -p(x,y) \log_2 p(x,y)$$

$$\mathrm{H}(X,Y) = \langle h(x,y) \rangle$$

– Exercise: How to code *H(X,Y)*?

1. Edit the Matlab function `jointentropy(p)` to return the joint Shannon entropy for the joint probability p.
   a. You can assume p is 2D (`p(x,y)`) for now.
   b. Can you simply alter your `entropy(p)` code?
   c. Try some test cases that you come up with yourself.
   d. *Challenge*: try dropping the assumption that p is 2D. Does your code change?

# Joint Entropy

- We can consider joint entropy of a multivariate, e.g. $\{X, Y\}$:

$$\mathrm{H}(X,Y) = \sum_{x \in A_x} \sum_{y \in A_y} -p(x,y) \log_2 p(x,y)$$

- Exercise: Let's code it assuming we don't have $P(X,Y)$ given, but are given empirical data to compute $P(X,Y)$ from:
  1. Edit the Matlab function `jointentropyempirical(xn)` to return the Shannon entropy for the given multivariate samples $x_n$ of $X$ (n is sample index)
     a. Input is $x_n$ as a matrix of samples, where each row is a vector of samples of each variable (e.g. `[0,1]`)
     b. Trick: can we use our existing `entropyempirical()` by combining {x,y} into a joint variable?

# Aside: Shannon entropy – derivation

– Shannon entropy of a random variable $X$:

$$\mathrm{H}(X) = \sum_{x \in A_X} -p(x)\log_2 p(x)$$

– Is a unique form that satisfies three axioms (Ash, 1965; Shannon, 1948):
  – Continuity w.r.t. $p(x)$
  – Monotony – $H(X)\uparrow$ as $|A_X|\uparrow$, for $p(x) = 1/|A_X|$
  – Grouping – For independent variables $X$ and $Y$, $H(X, Y) = H(X) + H(Y)$

C. E. Shannon. A mathematical theory of communication. Bell System Technical Journal, 27(3–4):379–423, 623–656, 1948.
R. B. Ash. Information Theory. Dover Publications Inc., New York, 1965. p. 5-12.

# Conditional entropy

- What if we already know something that may pertain to X? Does this change our surprise/uncertainty?
- Conditional entropy: (average) surprise remaining about sample *x* of *X* if we already know the sample *y* of *Y*

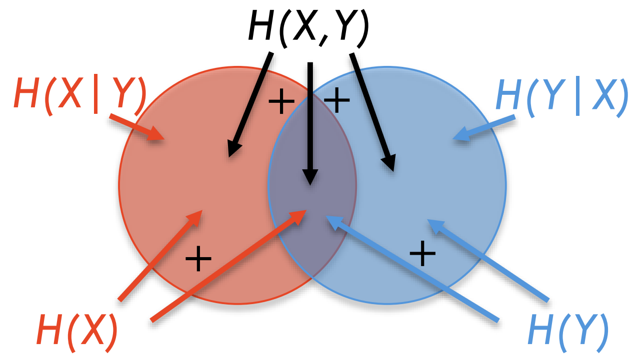
$$h(x|y) = h(x,y) - h(y)$$
$$h(x|y) = -\log_2 p(x|y)$$

$$\mathrm{H}(X|y) = \sum_{x \in A_x} p(x|y) \log_2 \frac{1}{p(x|y)}$$

$$\mathrm{H}(X|Y) = \sum_{y \in A_y} p(y) \mathrm{H}(X|y)$$

$$\mathrm{H}(X|Y) = \sum_{x \in A_x} \sum_{y \in A_y} p(x,y) \log_2 \frac{1}{p(x|y)}$$

$$\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - H(Y)$$




Venn diagram

**Properties:**

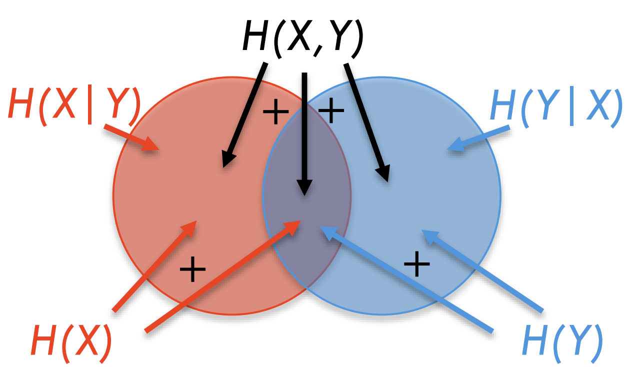
- $0 \leq H(X|Y) \leq H(X)$
- $H(X|Y) = H(X)$ iff $X$ and $Y$ are independent
- $H(X|Y) = 0$ means there is no surprise left in $X$ once we know $Y$.

# Conditional entropy

– Conditional entropy: (average) surprise remaining about sample *x* of *X* if we already know the sample *y* of *Y*

$$h(x|y) = h(x,y) - h(y)$$

$$\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - H(Y)$$




Venn diagram

Coding interpretation of *H*(*X* | *Y*)?

Example 1:

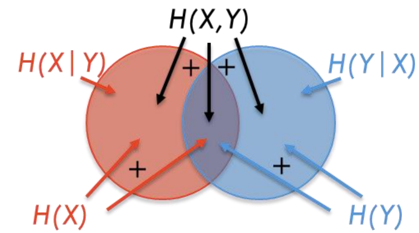
– Coding characters in English text –what variable *Y* would drop *H*(*X*) to some *H*(*X* | *Y*) and therefore the code length for a conditional encoding of incoming character *X*?

– Context of previous character(s) *Y* changes the probability of the next character *X* – Markov chains

| $i$ | $a_i$ | $p_i$ | $h(p_i)$ |
|---|---|---|---|
| 1 | a | .0575 | 4.1 |
| 2 | b | .0128 | 6.3 |
| 3 | c | .0263 | 5.2 |
| 4 | d | .0285 | 5.1 |
| 5 | e | .0913 | 3.5 |
| 6 | f | .0173 | 5.9 |
| 7 | g | .0133 | 6.2 |
| 8 | h | .0313 | 5.0 |
| 9 | i | .0599 | 4.1 |
| 10 | j | .0006 | 10.7 |
| 11 | k | .0084 | 6.9 |
| 12 | l | .0335 | 4.9 |
| 13 | m | .0235 | 5.4 |
| 14 | n | .0596 | 4.1 |
| 15 | o | .0689 | 3.9 |
| 16 | p | .0192 | 5.7 |
| 17 | q | .0008 | 10.3 |
| 18 | r | .0508 | 4.3 |
| 19 | s | .0567 | 4.1 |
| 20 | t | .0706 | 3.8 |
| 21 | u | .0334 | 4.9 |
| 22 | v | .0069 | 7.2 |
| 23 | w | .0119 | 6.4 |
| 24 | x | .0073 | 7.1 |
| 25 | y | .0164 | 5.9 |
| 26 | z | .0007 | 10.4 |
| 27 | - | .1928 | 2.4 |
| $\sum_i p_i \log_2 \frac{1}{p_i}$ | | | 4.1 |

Table 2.9. Shannon information contents of the outcomes a–z.

# Conditional entropy




– Conditional entropy: (average) surprise remaining about sample *x* of *X* if we already know the sample *y* of *Y*

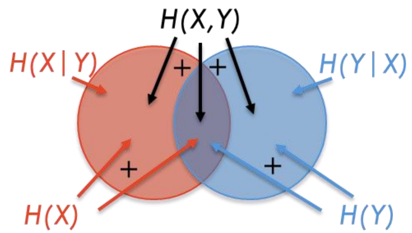
$$h(x|y) = h(x,y) - h(y)$$

$$\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - H(Y)$$

– Exercise: Let's code it!

1. Edit the Matlab function `conditionalentropy(p)` to return the conditional entropy for *X given Y* for the joint probability *p*.
   a. You can assume *p* is 2D (`p(x,y)`); this is the input.
   b. Trick: can we use our existing `entropy()` and `jointentropy()`?
   c. Test: `conditionalentropy([0.5,0;0,0.5])`=0
   d. Test: `conditionalentropy([0.25,0.25;0.25,0.25])` = 1
   e. Guess Who? H(`sex|earings`)? Construct p(`sex,earings`) first. Is H(`earings|sex`) the same?

# Conditional entropy




– Conditional entropy: (average) surprise remaining about sample *x* of *X* if we already know the sample *y* of *Y*

$$h(x|y) = h(x,y) - h(y)$$

$$\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - H(Y)$$

– Exercise: Let's code it!

1. Edit the Matlab function `conditionalentropyempirical(xn,yn)` to return the conditional entropy for *X given Y* from empirical samples $x_n,y_n$:
   a. Input is samples $x_n,y_n$.
   b. Trick: can we use our existing `entropyempirical()` and `jointentropyempirical()`?
   c. Test: `conditionalentropyempirical([0,0,1,1], [0,1,0,1])` =1
   d. Test: `conditionalentropyempirical([0,0,1,1], [0,0,1,1])` =0

# Chain rule for entropy and information content

- Chain rule for entropy:
  - $H(X,Y) = H(X) + H(Y|X)$
  - $H(X,Y) = H(Y) + H(X|Y)$
  - $H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i | X_1 .. X_{i-1})$
  - Same applies for $h(x,y)$, $H(X,Y|Z)$ and $h(x,y|z)$.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.5
D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003. Section 8.1

# Introduction to Information Theory: summary

- We've been introduced to the ideas of uncertainty and surprise.
- Understand the meaning of entropy as -*p log p*
- Know how to calculate entropy

- Coming up: Move onto measuring information

# Questions