BCM0504

Natureza da Informação

Teoria da Informação e Entropia

Prof. Alexandre Donizeti Alves

Universidade Federal do ABC

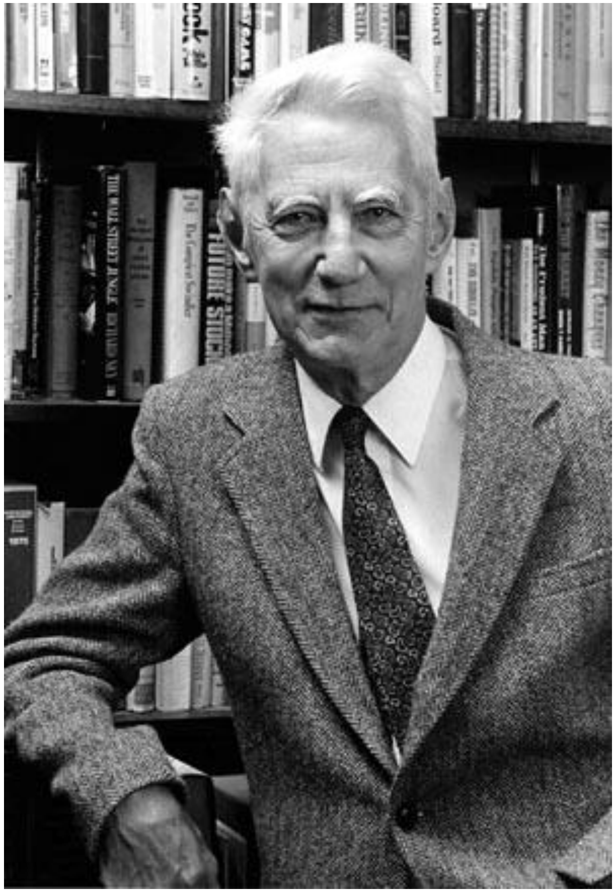
Bacharelado em Ciência e Tecnologia

Bacharelado em Ciências e Humanidades

Terceiro Quadrimestre - 2018

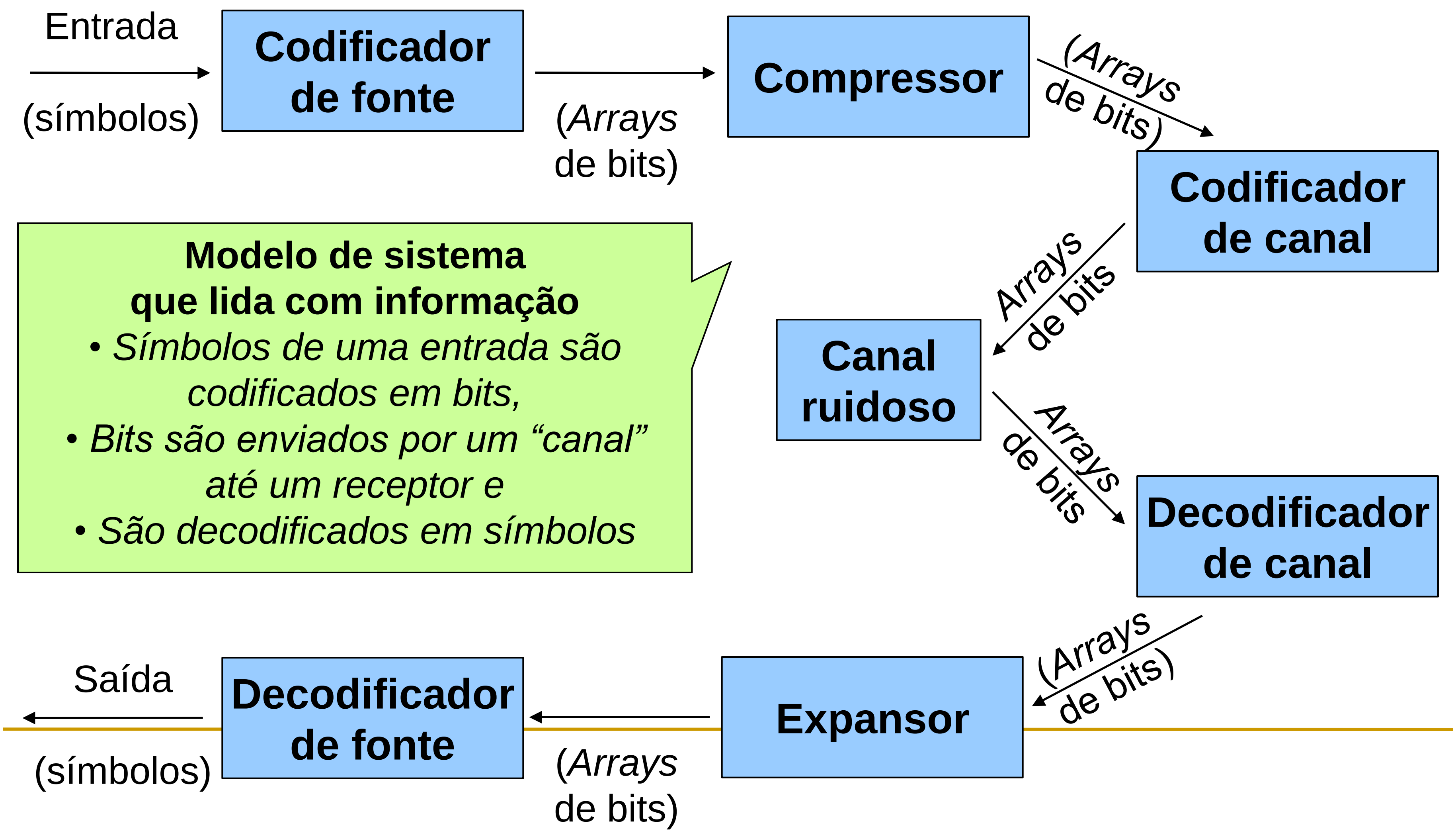
Claude Elwood Shannon (1916-2001)

- Trabalhava nos Laboratórios Bell
- Quanta informação pode passar por uma linha de telefone?
- Inventou o termo bit
- Toda informação pode ser representada por uma cadeia de bits



Modelo de sistema de comunicação

- Modelo estendido



Componentes

- Voltando à fonte (entrada)
  - Modelar em termos de distribuições de probabilidade
  - Função de fonte: prover um símbolo ou uma sequência de símbolos
    - Selecionados de um conjunto de símbolos
    - Seleção a partir, por exemplo, de:
      - Experimento
        - Ex.: jogar moeda ou dado
      - Observação de ações
      - Representação de um objeto
        - Ex.: caracteres de texto, pixels de imagem

Fonte

- Consideraremos número finito de símbolos
  - E mutuamente exclusivos
    - Só um pode ser escolhido a cada instante

- Cada escolha = um “resultado”
  - **Objetivo**: rastrear a sequência de resultados
    - E informação que as acompanha, da entrada à saída do sistema de comunicação
    - Necessário saber qual é o resultado e propriedades dele
      - *Propriedades*: coisas que se aplicam ou não a cada símbolo

Fonte

- Sabendo o resultado, como denotá-lo?
  - Fornecendo sua denominação
- E se não sabemos ainda o resultado, ou estamos incertos sobre ele?
  - Como expressar conhecimento sobre ele se há incerteza?
    - Usar probabilidade

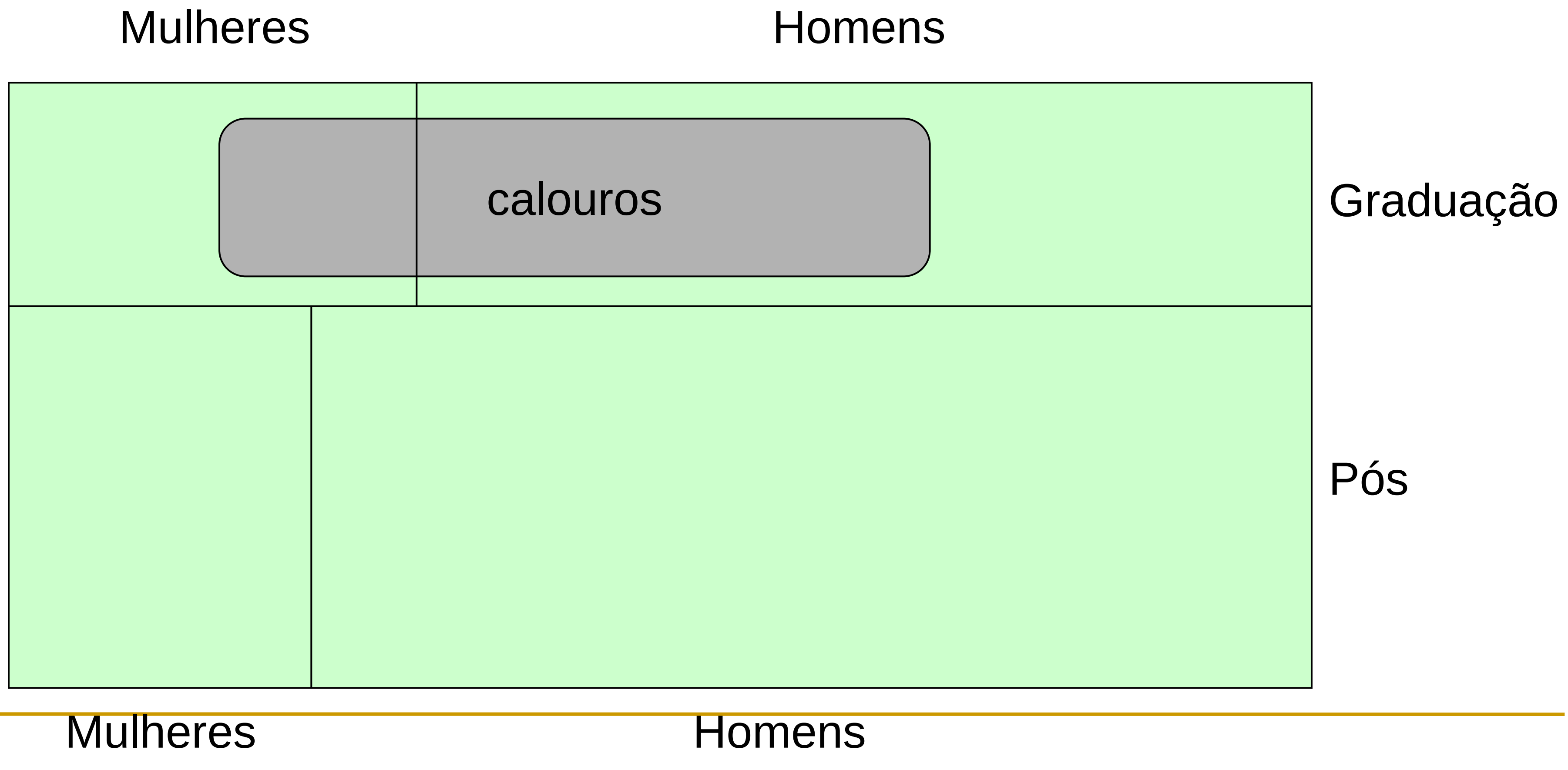
Probabilidades

- **Exemplo**: características dos estudantes MIT - 2007

Tipo/número	**Mulheres**	**Homens**	**Total**
Calouros	482	596	1078
Graduação	1916	2316	4232
Pós-graduação	1916	4236	6152
Total estudantes	3832	6552	10384

Probabilidades

- Ex.: características dos estudantes MIT



Probabilidades

- Ex.: características dos estudantes MIT

  - Supor que um calouro é selecionado
    - Símbolo = um estudante individual
    - Conjunto de possíveis símbolos = 1078
  - Não sabendo qual foi, é homem ou mulher?
    - Como você caracteriza o seu conhecimento?
      - Qual é a probabilidade de uma mulher ter sido selecionada?

Probabilidades

- Ex.: características dos estudantes MIT
  - Supor que um calouro é selecionado
    - Qual é a probabilidade de uma mulher ter sido selecionada?
      - 45% dos calouros são mulheres (482 / 1078): estatística
      - Se todos têm mesma probabilidade de serem escolhidos, probabilidade de selecionar mulher é 45%
      - E se seleção é feita no corredor de um dormitório feminino?
        - Probabilidade será maior que 45%

Eventos

- Resultado: algo que segue como consequência
  - Símbolo selecionado, conhecido ou não para nós
- Evento: subconjunto dos possíveis resultados de um experimento
  - Quando seleção é feita, há vários eventos
    - Um é o próprio resultado: *evento fundamental*
    - Outros: seleção de símbolo com propriedade particular
    - Por simplicidade, as seleções serão chamadas eventos

Eventos

- Ex.: um calouro do MIT é selecionado
  - Resultado é a pessoa em específico selecionada
    - Evento fundamental
  - Outros eventos:
    - Seleção de uma mulher
    - Seleção de alguém da California
    - Seleção de alguém maior de 18 anos
    - Seleção de mulher to Texas
    - Seleção de qualquer pessoa
      - Evento universal
    - Seleção de nenhum símbolo
      - Evento nulo
    - etc.

Eventos

- Diferentes eventos podem ou não se sobrepor
  - Ocorrer para o mesmo resultado

Eventos que não se sobrepõem: mutuamente exclusivos
Ex.: aluno selecionado ser homem ou mulher

- Conjunto de eventos exaustivo: ao menos um deles ocorre quando um símbolo é escolhido
  - Ex.: aluno escolhido tem:
    - *Evento 1*: menos que 25 anos
    - *Evento 2*: mais que 17 anos
      - São exaustivos, mas não são mutuamente exclusivos

Eventos

- Partição: conjunto de eventos mutuamente exclusivos e exaustivos
  - Partição fundamental: contém todos os eventos fundamentais
    - Ex.: Eventos selecionar mulher e selecionar homem formam uma partição
    - Ex.: Eventos fundamentais associados a cada uma das 1078 pessoas formam partição fundamental

Resultados conhecidos

- Sabendo um resultado, é fácil denotá-lo
  - Especificando o símbolo que foi selecionado
  - Sabe então que eventos ocorreram
  - Mas deve conhecer o resultado
    - Enquanto não é conhecido, não é possível expressar dessa forma

- Outra forma de denotar: probabilidades
  - Generalizável a situação em que resultado ainda não é conhecido

Resultados conhecidos

- Seja *i* um índice dentro de uma partição
  - De 0 a *n*-1 (*n* é o número de eventos na partição)

- Para qualquer evento particular A_i
  - $p(A_i) = 1$ se resultado correspondente é selecionado
  - $p(A_i) = 0$ caso contrário
    - Partição $\Rightarrow$ será 1 para exatamente um evento *i* e 0 para demais eventos
      - Ex. *p*(evento universal) = 1 e *p*(evento nulo) = 0

Mesma notação se aplica a eventos *A* quaisquer (não necessariamente em uma partição)

Resultados desconhecidos

- Se símbolo ainda não foi selecionado, você ainda não conhece o resultado
  - Então cada $p(A)$ pode ter um valor entre 0 e 1
    - Valores maiores = maior crença de que o evento vai ocorrer
    - Valores menores = menor crença de que o evento vai ocorrer
    - Se evento é certamente impossível $\Rightarrow p(A) = 0$
    - Quando resultado é aprendido, cada $p(A)$ pode ser ajustado para 0 ou 1

Resultados desconhecidos

- Forma de atribuir os valores
  - Obedecer teoria da probabilidade
    - Valores = probabilidades
      - Conjunto de probabilidades que se aplicam a uma partição = distribuição de probabilidade

Axiomas da probabilidade:

• Para qualquer evento A: $0 \leq p(A) \leq 1$
• Se um evento A ocorre somente em função de outros eventos mutuamente exclusivos A_i (porque, por exemplo, formam uma partição): $p(A) = \Sigma\, p(A_i)$
• Para qualquer partição: $\Sigma\, p(A_i) = 1$
(já que p(evento universal) = 1)

Eventos conjuntos

- Probabilidade de símbolo escolhido ter duas propriedades diferentes
  - Ex.: escolha de caloura (mulher) do Texas
    - $p(M)$ = probabilidade de ser mulher
    - $p(T)$ = probabilidade de ser do Texas
    - $p(M,T)$ = probabilidade de ser mulher do Texas
- Se os eventos são independentes $\Rightarrow$ multiplica probabilidades dos eventos individuais
  - Probabilidade de um não depende do outro ocorrer
    - $p(A,B) = p(A)\, p(B)$

Eventos conjuntos

- Independência não é usual
  - Fórmula mais geral para a probabilidade do evento conjunto (ambos ocorrerem)
    - Probabilidades condicionais: probabilidade de um evento dado que outro ocorreu
      - Ex.: $p(M \mid T)$ = probabilidade condicional de selecionar mulher, dado que o calouro escolhido é do Texas

$$
\begin{aligned}
p(A,B) &= p(B)\, p(A \mid B) \\
&= p(A)\, p(B \mid A)
\end{aligned}
$$

Teorema de *Bayes*

Eventos conjuntos

- Ex.:
  - $p(M, T) = \mathrm{p}(T)\ p(M \mid T)$
    - *Probabilidade de calouro escolhido ser mulher do Texas é probabilidade de estudante ser do Texas vezes a probabilidade de que, sendo texana, a pessoa é mulher*

OU

  - $p(M, T) = \mathrm{p}(M)\ p(T \mid M)$
    - *Probabilidade de calouro escolhido ser mulher do Texas é probabilidade de estudante ser mulher vezes a probabilidade de que a pessoa escolhida, sendo mulher, é texana*

Exemplo

- Considere que um estudante qualquer é selecionado (entre todos) aleatoriamente
  - Igual probabilidade para todos estudantes
  - Partição fundamental: 10384 eventos fundamentais
    - Cada aluno em particular
    - Soma de todas probabilidades = 1
      - Então cada um tem probabilidade 1/10384 = 0,01%

Exemplo

- Considere que um estudante qualquer é selecionado (entre todos) aleatoriamente
  - Qual é a probabilidade de ser um graduando?
    - $p(G) = 4232 / 10384 = 0{,}41$
    - Soma das probabilidades fundamentais dos 4232 eventos associados a estudantes graduandos

Exemplo

- Considere que um estudante qualquer é selecionado (entre todos) aleatoriamente
  - Qual é a probabilidade de ser um homem graduando?
    - $p(G) = 0,41$
    - Probabilidade conjunta $p(H,G)$?
    - Selecionado graduando, qual é a probabilidade condicional dele ser um homem?
      - $p(H \mid G)$?

Exemplo

- Considere que um estudante qualquer é selecionado (entre todos) aleatoriamente
  - $p(H \mid G)$
    - Nova partição fundamental = 4232 possíveis graduandos
      - 2316 desses são homens
      - Cada um tem igual probabilidade de ser selecionado, ou seja, 1/4232
      - Evento selecionar um homem é relacionado a 4236 desses eventos fundamentais
        - $p(H \mid G) = 2316/4232 = 0,55$

Exemplo

- Considere que um estudante qualquer é selecionado (entre todos) aleatoriamente
  - $p(H, G)$
    - Teorema de Bayes
      - $p(H, G) = \mathrm{p}(G)\,\mathrm{p}(H \mid G)$

$$= \frac{4232}{10384} \times \frac{2316}{4232} = \frac{2316}{10384} = \mathbf{22{,}3\%}$$

Informação

- Queremos expressar a informação (ou falta dela) a respeito da escolha de um símbolo
  - Conhecida a resposta, não há incerteza sobre o símbolo escolhido ou suas propriedades
    - E que eventos ocorrem como consequência dessa seleção
  - E antes da seleção ser feita? Ou de sabermos a resposta?
    - Temos incerteza
      - Quanta?

Um método simples para medir a informação:

Quantas perguntas preciso fazer para saber qual número você pensou dentre este conjunto de números?

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

- Um dos alunos pensa um número
- O professor faz as perguntas, enquanto que outro aluno escreve um 1 no quadro branco se a resposta for sim e zero se a resposta for não

Um método simples para medir a informação:

Supor que o aluno pensou cinco

- O número é maior do que 8? Não 0
- O número é maior que 4? Sim   01
- O número é maior que 6? Não   010
- O número é maior que 5? Não   0100

Então é o 5

Se as questões estão corretamente formuladas, é possível identificar o número somente com $\log_2(16)=4$ questões ou 4 bits

Informação

- Sabendo a resposta, a informação pode ser contada a outro especificando o símbolo escolhido
  - Se há dois símbolos possíveis $\Rightarrow$ um bit pode ser usado para tal
    - Ex.: cara ou coroa em jogada de moeda
    - 1 bit é a quantidade de informação necessária para tomar uma decisão perante duas opções igualmente prováveis
  - Se há quatro possíveis eventos $\Rightarrow$ resposta pode ser expressa por dois bits
    - Ex.: naipes de carta pega de um baralho
  - Se há n possíveis respostas $\Rightarrow$ são necessários $\log_2 n$ bits

O Bit

- Grau de imprevisibilidade
- Bit é a quantidade de informação necessária para tomar uma decisão perante duas opções igualmente prováveis
- Calcular grau de imprevisibilidade (em bits) segundo a fórmula de Boltzmann
  - S = k log(W)
  - W são possíveis configurações que toma um determinado arranjo de particulas
  - k = 1

A informação é medida em bits

$$S = \log_2(n)$$

$$S = \log_2(1) = 0\ bits$$

$$S = \log_2(2) = 1\ bit$$

$$S = \log_2(6) = 2{,}58\ bits$$

Informação

- Quantidade de informação aprendida ao conhecer o resultado é o número mínimo de bits que seriam usados para especificar o símbolo

Exemplo

- Classe de 32 alunos: 2 mulheres e 30 homens
  - Um aluno é escolhido
    - Objetivo é saber qual
  - Incerteza inicial é de 5 bits
    - Necessário para especificar o resultado
  - Escolha aleatória $\Rightarrow$ probabilidade de cada um ser selecionado é 1/32
    - Mulher: $p(M) = 2/32$
    - Homem: $p(H) = 30/32$

Exemplo

- Classe de 32 alunos: 2 mulheres e 30 homens
  - Quanta informação ganhamos sabendo que a escolha é de uma mulher, sem saber qual?
    - Incerteza é diminuída de 5 bits para 1 bit
      - Necessário para especificar qual das duas mulheres
      - Ganhamos 4 bits de informação!
  - E se for homem?
    - Reduz incerteza de 5 a 4,91 bits ($\log_2 30$)
    - Aprendemos 0,09 bits de informação

Informação

- E uma partição em que os eventos têm probabilidades diferentes?
  - Aprendemos diferentes quantidades de informação
    - Se resultado era provável, aprendemos menos do que se ele era improvável
    - Informação ganha por uma resposta *i* é $\log_2(1 / p(A_i))$
      - $= - \log_2(p(A_i))$
      - Ex.: Informação ganha sabendo que é mulher
        - $p(M) = 2/32$
        - $I(M) = \log_2(32/2) = \log_2(16) = 4$

Uma interpretação da fórmula: informação = -log(probabilidade)

- Numa seqüência binária, se todos são um, não há informação
  - Ex: 1111111111

- Mas se o número 1 aparece 10% das vezes, ele possui $-\log_2(1/10)$ de informação ou:

  $$-\log_2(p) = -\log_2(0{,}1) = \log_2(10) = 3{,}3219 \text{ bits}$$

- Enquanto que o número 0 possui:

  $$-\log_2(p) = -\log_2(0{,}9) = \log_2(10/9) = 0{,}152 \text{ bits}$$

- Então, quanta informação temos nesta mensagem?

0000010000

Poderia pensar que como temos 10 dígitos, teríamos 10 bits de informação, mas na verdade cada 0 vale 0,152bits porque tem pouca incerteza (imprevisibilidade ou surpresa), enquanto que o único 1 tem muita informação (3,321 bits)

A **entropia** ou informação total da mensagem seria:

H_{total}=9x(informação do zero)+1x(informação do 1)=

=9x0,152+1x3,321= 4,6890 bits ao invés de 10 bits

Informação

- Se queremos quantificar nossa incerteza antes de saber uma resposta
  - Média sobre todos os possíveis resultados
    - Todos eventos na partição com probabilidade não nula
  - Informação média:
    - Soma da multiplicação da informação de cada evento A_i por $p(A_i)$

$$H = - \Sigma \; p(A_i) \log_2(p(A_i))$$

Entropia de uma fonte

Fundamental para caracterizar informações de fontes

Informação

- Atenção: cuidado quando probabilidade de um evento é 0
  - Considerar que na realidade é valor bem pequeno, mas não nulo
  - O produto então aproxima o valor 0

Informação

- Ex.: Informação contida na jogada de uma moeda alterada para cair 60% das vezes em cara e 40% em coroa

$$H = -0{,}6\log_2(0{,}6) - (0{,}4)\log_2(0{,}4) = 0{,}97\,bits$$

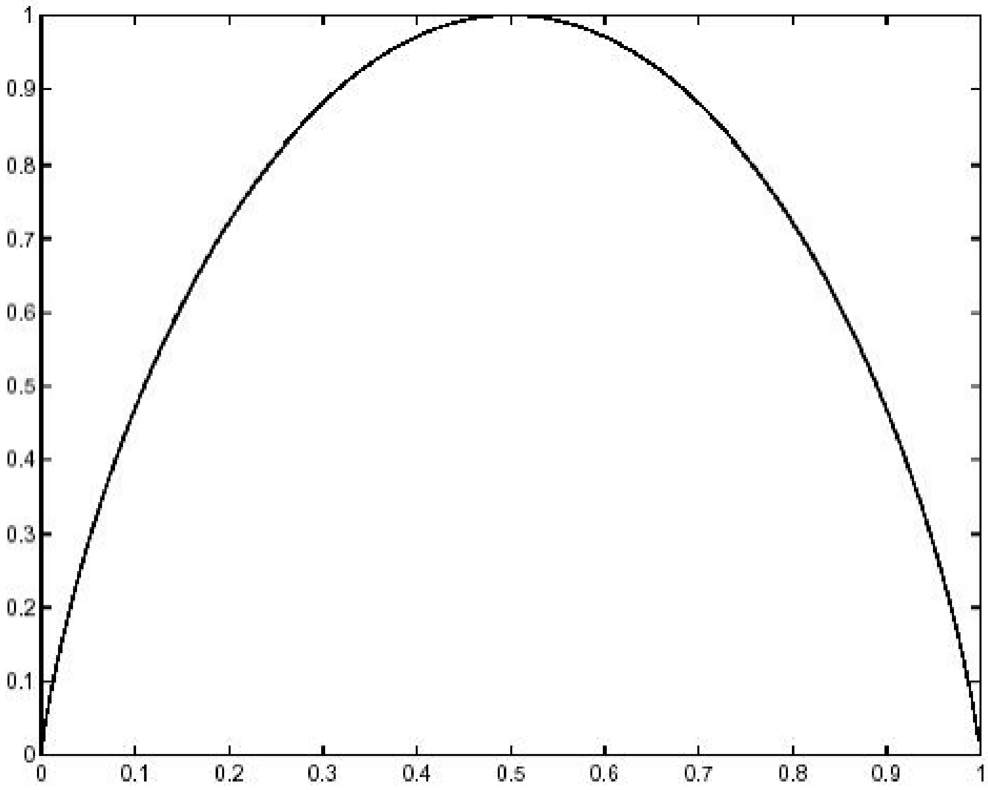
  - Observar que a soma das duas probabilidades é 1 (0,6+0,4=1)

Propriedades de Informação

- É conveniente pensar em informação como quantidade física com dimensões
  - Ex. como velocidade: tamanho/tempo (m/s)
  - Menos natural, uma vez que probabilidades não possuem dimensão
  - Mas fórmula usa $\log_2$
    - Informação em log de base 2 é expressa em bits
      - Poderia usar outras bases
        - $\log_k(x) = \log_2(x) / \log_2(k)$

Propriedades da informação

- Se há dois eventos na partição com probabilidades p e (1-p), a informação por símbolo é:
  - $H = - p \log_2(p) - (1-p) \log_2(1-p)$



Entropia de uma fonte com 2 símbolos como função de p

Entropia (Shannon):

É maior (1 bit) quando p = 0,5 (probabilidade dos dois eventos é igual)

É 0 para p = 0 e p =1 (nestes casos, resposta é certa e nenhuma informação é ganha conhecendo-a)

Propriedades da informação

- Ex.: moeda

p(cara) + p(coroa) = 1

$$H = -p(cara)\log_2(p(cara)) - p(coroa)\log_2(p(coroa)) =$$
$$-p(cara)\log_2(p(cara)) - ((1 - p(cara))\log_2(1 - p(cara))$$

Quando ambas possibilidades têm a mesma probabilidade de acontecer, p(cara) = p(coroa) = 0,5 e a entropia ou imprevisibilidade é máxima, e igual a 1 bit

Propriedades da informação

- Para partições com mais de dois eventos, a informação por símbolo pode ser maior
  - Se há n possíveis eventos, a informação por símbolo situa-se entre 0 e $\log_2 n$ bits
    - Valor máximo $H = \log_2(n)$ quando todas as probabilidades são iguais

$$p_i = 1/n, \forall i$$

$$H = -\sum p_i \log_2 p_i = -n\left(\frac{1}{n}\right)\log_2\left(\frac{1}{n}\right) = \log_2(n)$$

Exemplo

- Ele diz bom dia todo dia
  - Há apenas um estado possível
  - 0 bit, não há informação
  - Entropia = 0
  - Não precisamos transmitir nada
  - Já sabemos que ele diz bom dia

Um emissor que fornece sempre a mesma mensagem, fornece 0 bits de informação
(enquanto que conteúdo informativo de uma mensagem pouco previsível é grande)

$$H = -\sum p_i \log_2 p_i$$

$$H = -1 \cdot \log_2(1) = 0\, bits$$

Um único evento com probabilidade =1

Exemplo

- 1 bit – dois estados igualmente prováveis
- Precisamos transmitir um bit para informar sobre o estado da moeda
- Mas sabemos que só pode ser cara ou coroa
  - Um bit resolve

Dois eventos (cara e coroa), cada um deles com probabilidade 0,5

$$H = -\sum p_i \log_2 p_i = -\sum_{i=1}^{2} p_i \log_2 p_i =$$

$$= -p_{cara} \log_2 p_{cara} - p_{coroa} \log_2 p_{coroa} =$$

$$= -0{,}5 \log_2 0{,}5 - 0{,}5 \log_2 0{,}5 =$$

$$= -0{,}5 \log_2 (1/2) - 0{,}5 \log_2 (1/2)) =$$

$$= 0{,}5 \log_2 2 + 0{,}5 \log_2 2 = 0{,}5 + 0{,}5 = 1 bit$$

Exemplo

- Dado: 6 estados
- 2 bits não são suficientes
- 3 bits: “sobram” 2 estados

Seis eventos 1,2,3,4,5,6
cada um deles com
probabilidade 1/6

$$H = -\sum p_i \log_2 p_i = -\sum_{i=1}^{6} p_i \log_2 p_i =$$

$$= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - p_3 \log_2 p_3 -$$

$$- p_4 \log_2 p_4 - p_5 \log_2 p_5 - p_6 \log_2 p_6 =$$

$$= -\left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) - \left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) - \left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right)$$

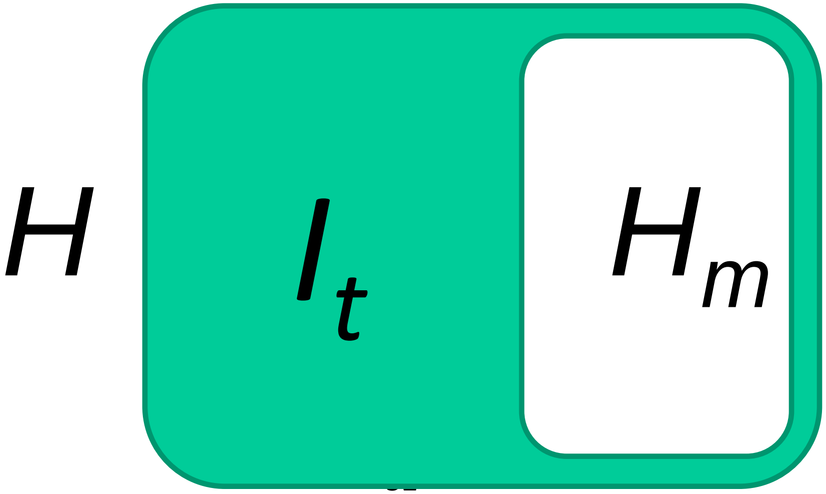
$$- \left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) - \left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) - \left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) =$$

$$= -6\left(\tfrac{1}{6}\right)\log_2\left(\tfrac{1}{6}\right) = -\log_2\left(\tfrac{1}{6}\right) = \log_2 6 = 3{,}32 \log_{10} 6 =$$

$$= 2{,}58 bits$$

Informação transmitida, I_t

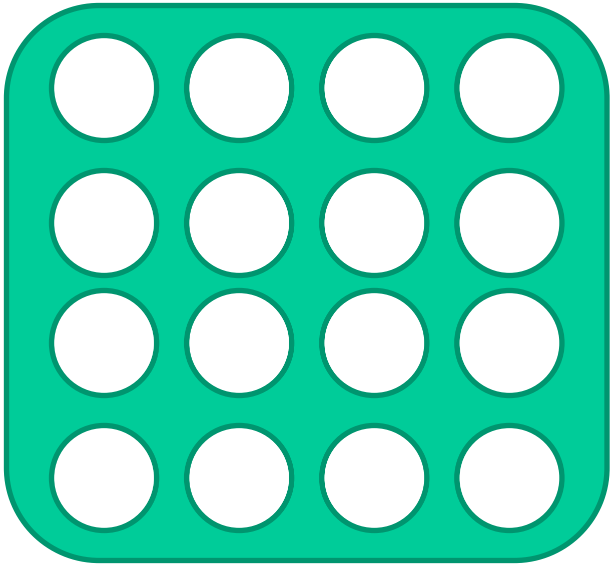
- É a diferença entre o grau de entropia ou imprevisibilidade inicial (*H*) e a imprevisibilidade final H_m



H, I_t e H_m são entropias

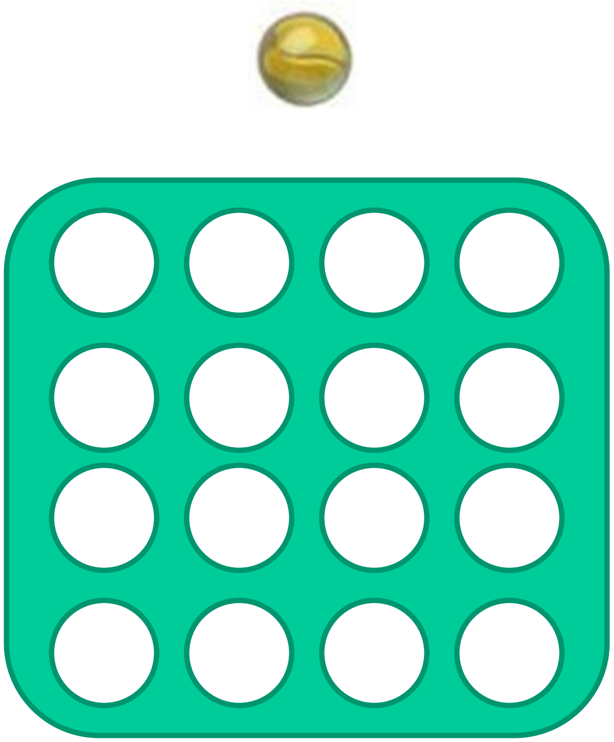
- A entropia H é chamada também entropia da fonte de informação
  - Representa a capacidade potencial de informação que pode ser fornecida por um determinado arranjo ou dispositivo no instante inicial
- I_t é a chamada Informação transmitida ou entropia do canal
- H_m é a entropia final, depois da mensagem "m" ter sido transmitida

Exemplo:



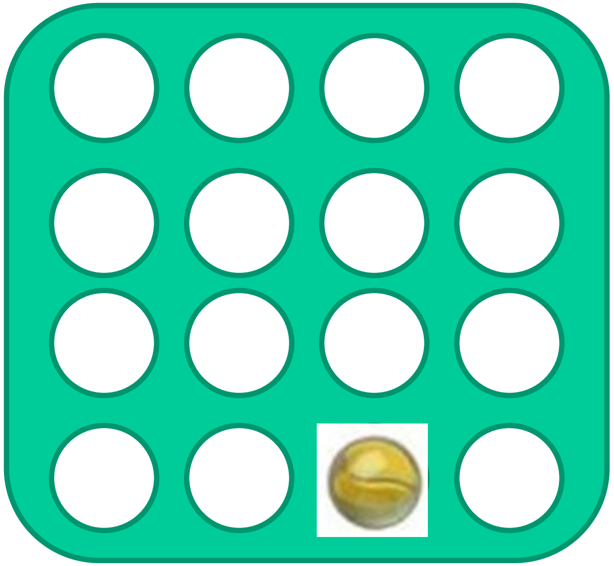
- Jogar uma bola em uma matriz com 16 posições possíveis
  - Qual seria a entropia inicial e qual a entropia final?
  - Qual seria a quantidade de informação fornecida pela bola?

Exemplo:



- Entropia inicial: $H = \log_2(16) = 4$

Exemplo:



- Entropia final: $H = \log_2(15)$
- Informação: diferença de entropias:
- $I = \text{Entropia}_{\text{inicial}} - \text{Entropia}_{\text{final}}$
- $I = \log_2(16) - \log_2(15) =$

$= \log_2(16/15) = 0{,}09$ bits