UNIVERSITY OF CAMBRIDGE

How heavy is 1 kg of information?

A very short introduction to information theory

Dr. Jossy Sayir, Dept of Engineering and EBI

About the speaker…

- Affiliated lecturer at the Department of Engineering
- Fellow of Robinson College and Director of Studies at Newnham College
- Currently teaching:
  - 2nd year probability
  - 3rd year information theory
  - 4th year coding theory
- Research in information theory, coding and communications
- Research fellow at the European Bioinformatics Institute

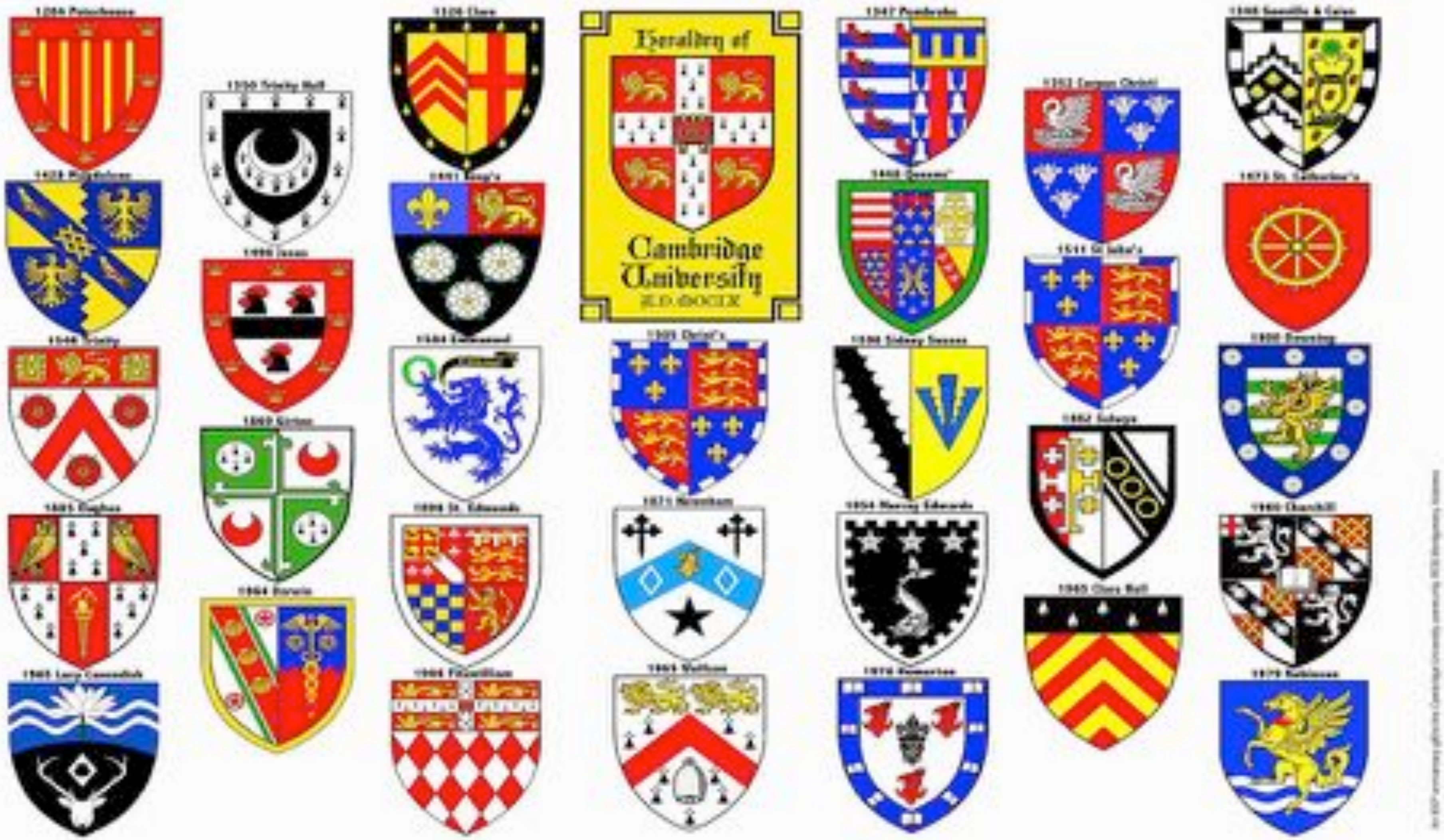
About the talk…

- Material from 3rd year course on information theory (without the maths)
- Claude Shannon's "Mathematical Theory of Communications" (1948)
- Big Bang of the information age
- Modern basis for data communication, storage, processing
- You use information theory in your mobile phone, when you skype, when you use the internet, when you listen to music, etc.

From the Bell Labs Webpage

- Information, like weight or energy, can be measured and quantified

Cambridge College 20 Questions



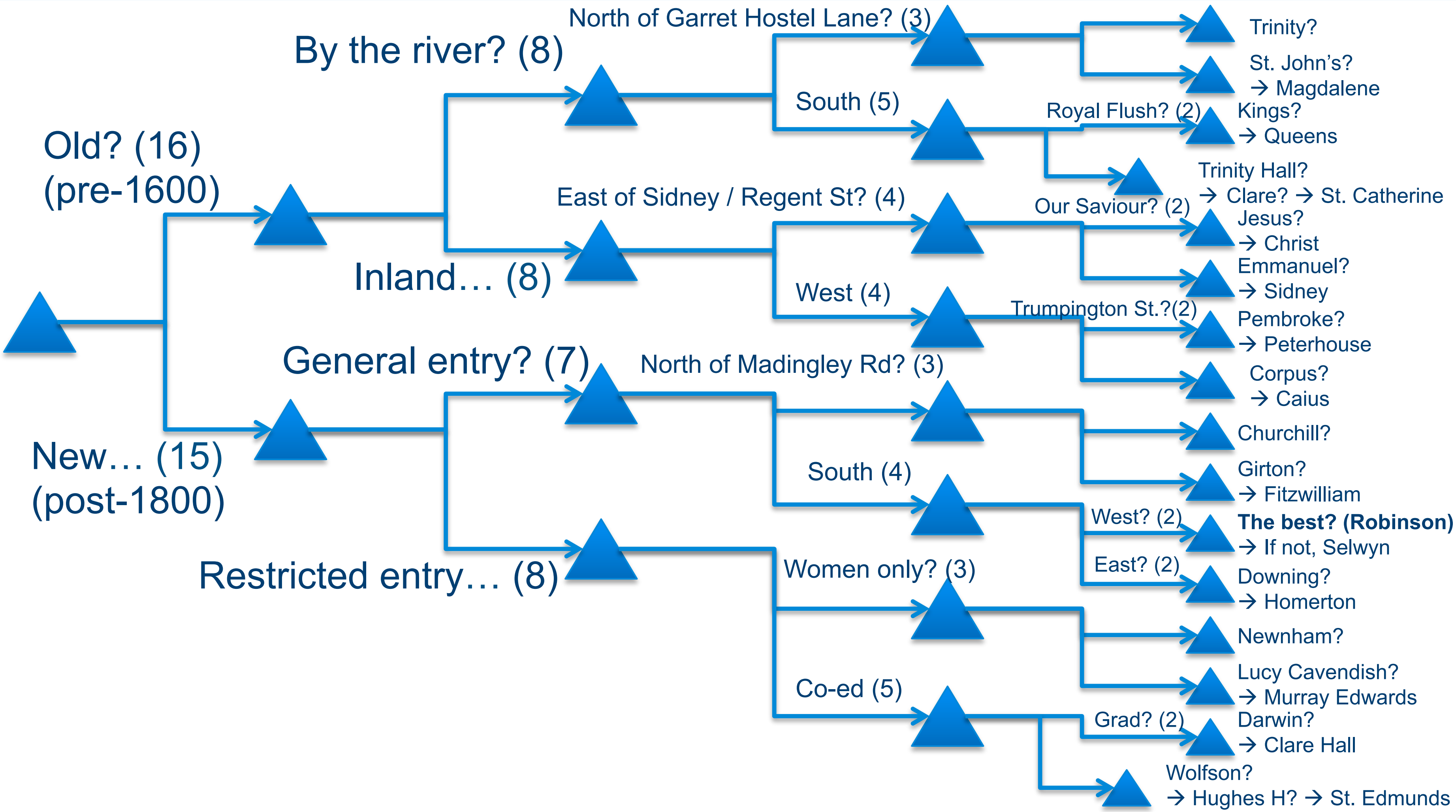
UNIVERSITY OF CAMBRIDGE

Cambridge College 20 Questions

- Guess a college in as few as possible "yes/no" questions
- 31 colleges
- How many questions?

Cambridge College 20 Questions

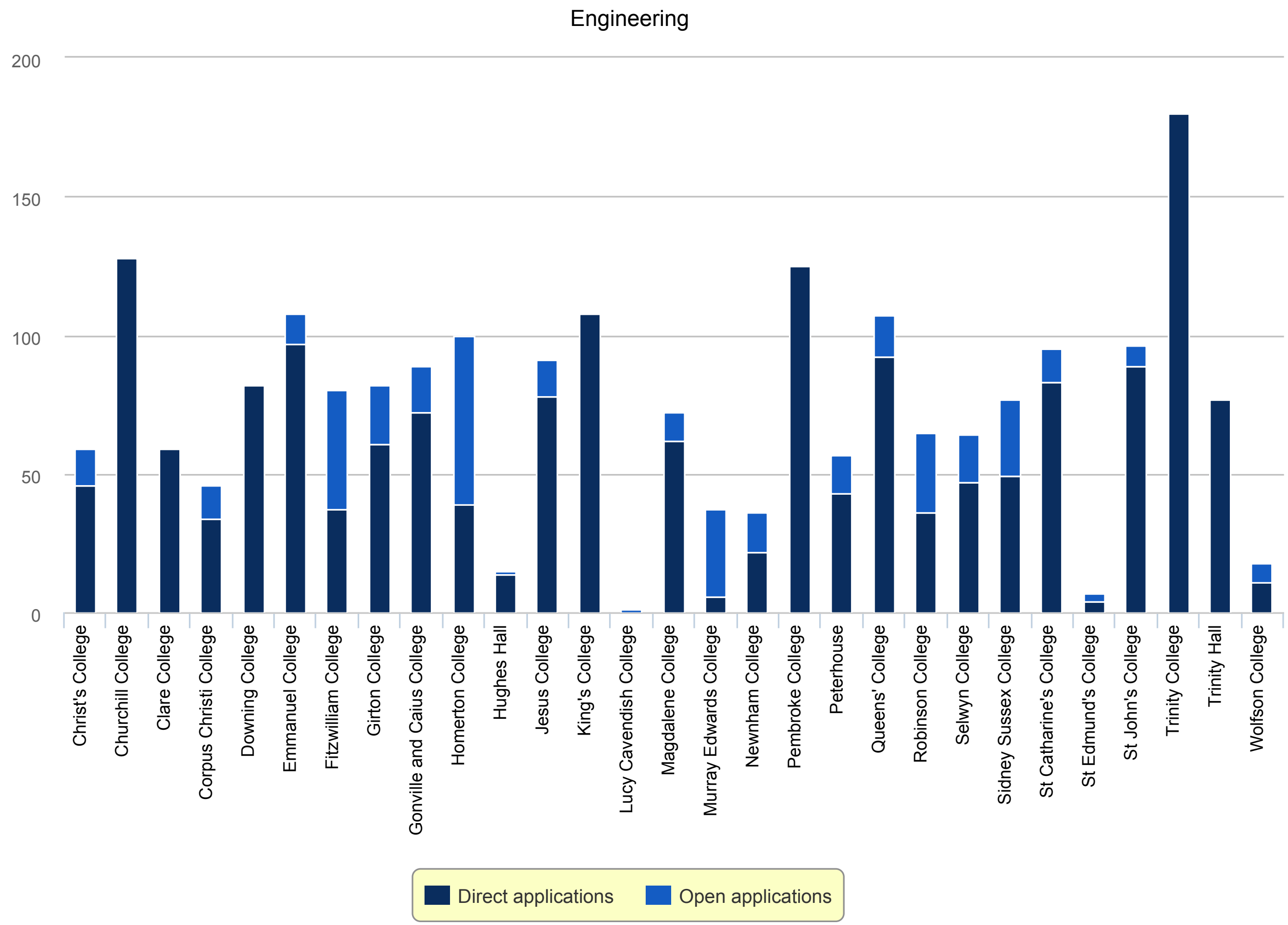
Guessing tree



20 Questions - analysis

- This tree could be improved to get to 5 questions
- Can you think of how you could ask all question at once?
- What college do you think I would pick?

Admissions Numbers Engineering 2014 cycle



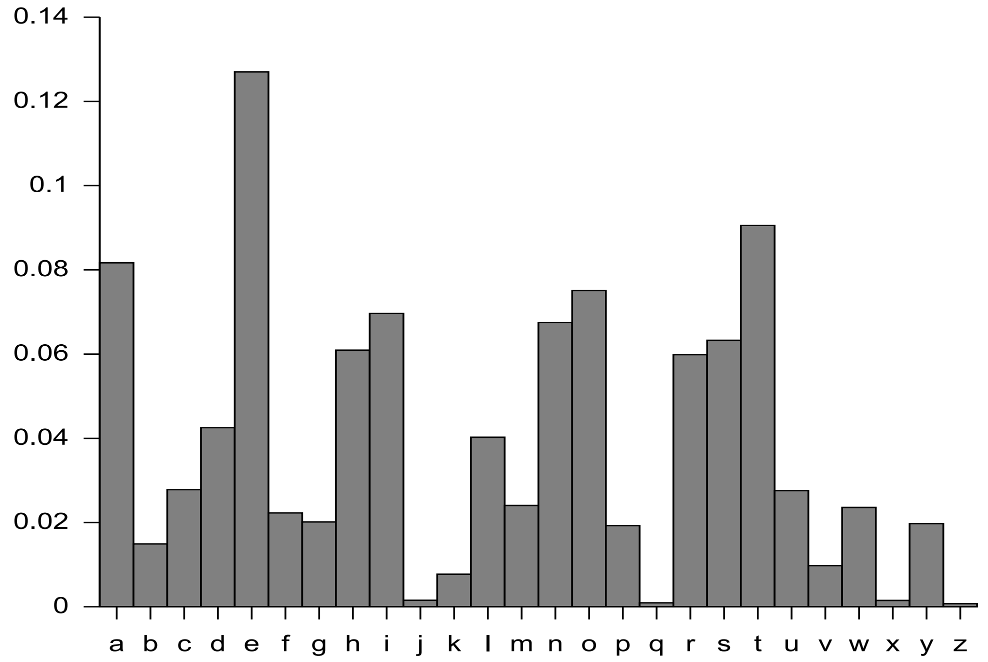
Total
2,161

Information measures…

- **Hartley's information:** if a question has N possible answers, the information content in its answer is log N
- **Shannon's information:** if an answer has probability p, it's as if it were one of $N_p=1/p$ equally likely answers and hence its information content is $\log(1/p) = -\log p$
- **Shannon's "entropy" formula:** $H = -\Sigma_n p_n \log_b p_n$
- What base is the log?

English text

- Entropy of English is H=4.17 bits
- Better than 5, but do we really need 4-5 yes/no questions to guess the next letter in English text?

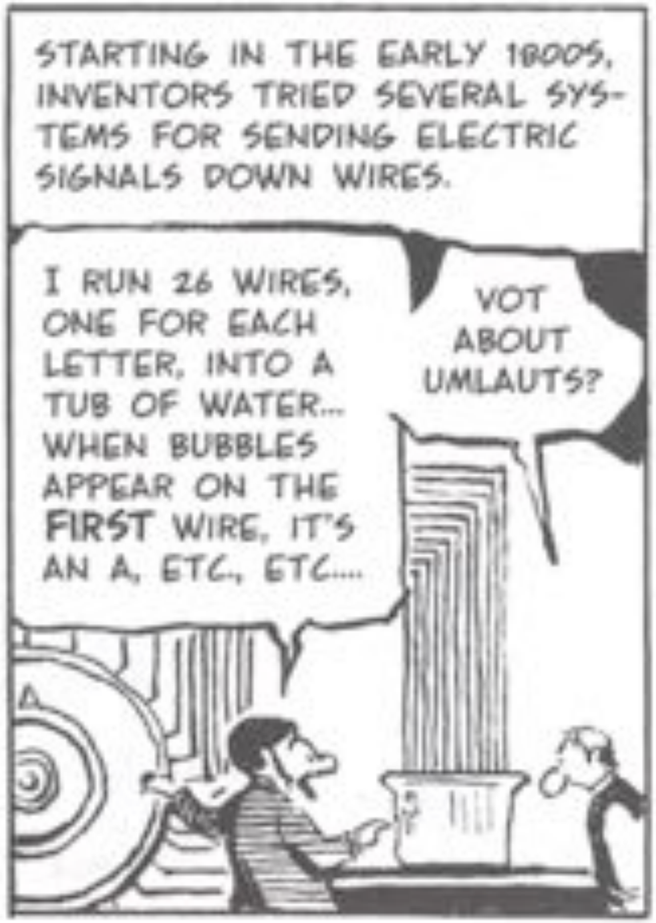


Wikipedia: "The frequency of letters in text has been studied for use in cryptanalysis, and frequency analysis in particular, dating back to the Iraqi mat…"

Source Coding

- English text can be compressed well below 2 bits per letter by modern data compression algorithms
- All sources (images, sound, video, files) are compressed before transmission (lossy or lossless)
- **Data compression removes all redundancy so that the result is perfectly unpredictable**
- Can you compress the 6 numbers between 1 and 59 resulting from a lottery draw?
- "Compressing Sets and Multisets of Sequences" Christian Steinruecken

Interlude


STARTING IN THE EARLY 1800S, INVENTORS TRIED SEVERAL SYS-TEMS FOR SENDING ELECTRIC SIGNALS DOWN WIRES.
I RUN 26 WIRES, ONE FOR EACH LETTER, INTO A TUB OF WATER... WHEN BUBBLES APPEAR ON THE FIRST WIRE, IT'S AN A, ETC., ETC....
VOT ABOUT UMLAUTS?

Reed Solomon Coding

SUDOKU

5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

SUDOKU

5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9

SUDOKU

5	3	4	5	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	1	5	6	7
9	5	9	7	6	1	4	2	3
4	2	7	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	8	2	8	4
2	8	7	4	1	9	6	2	5
3	4	5	2	8	6	1	7	9

1-error correction

0	1	1	0	
0	1	0	1	
1	0	1	1	
0	1	0	1	

1-error correction

0	1	1	0	0
0	1	0	1	0
1	0	1	1	1
0	1	0	1	0
1	1	0	1	1

1-error correction

Can we still correct erasures?

0		1		0
	1		1	
1	0		1	1
	1	0		0
1	1	0		1

1-error correction

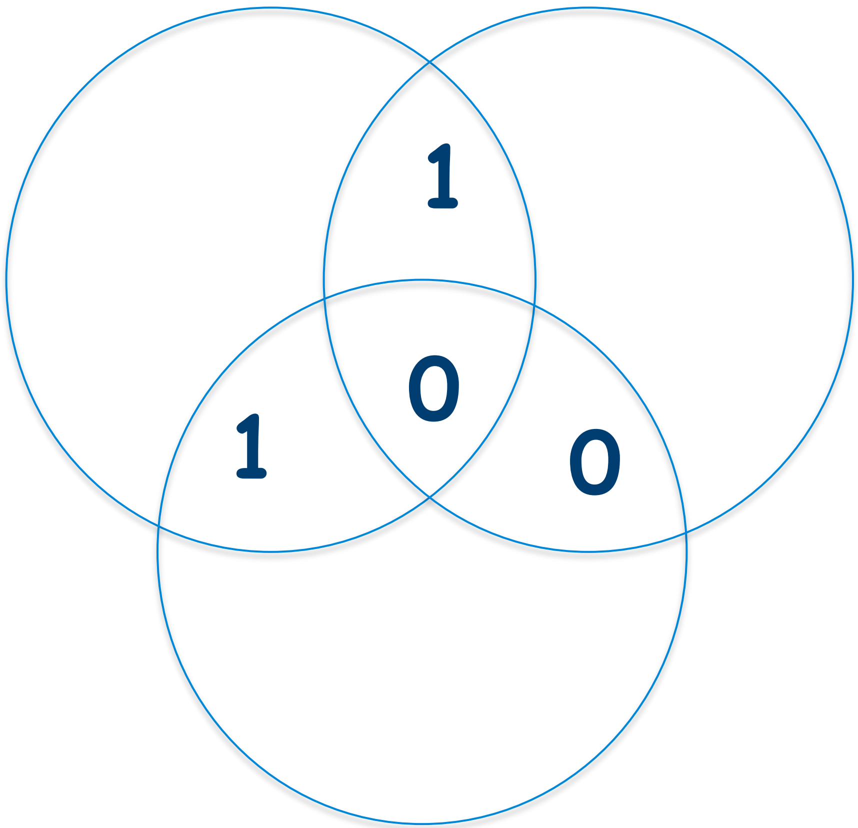
0	1	1	0	0
0	1	0	1	0
1	1	1	1	1
0	1	0	1	0
1	1	0	1	1

Dimensions and rate

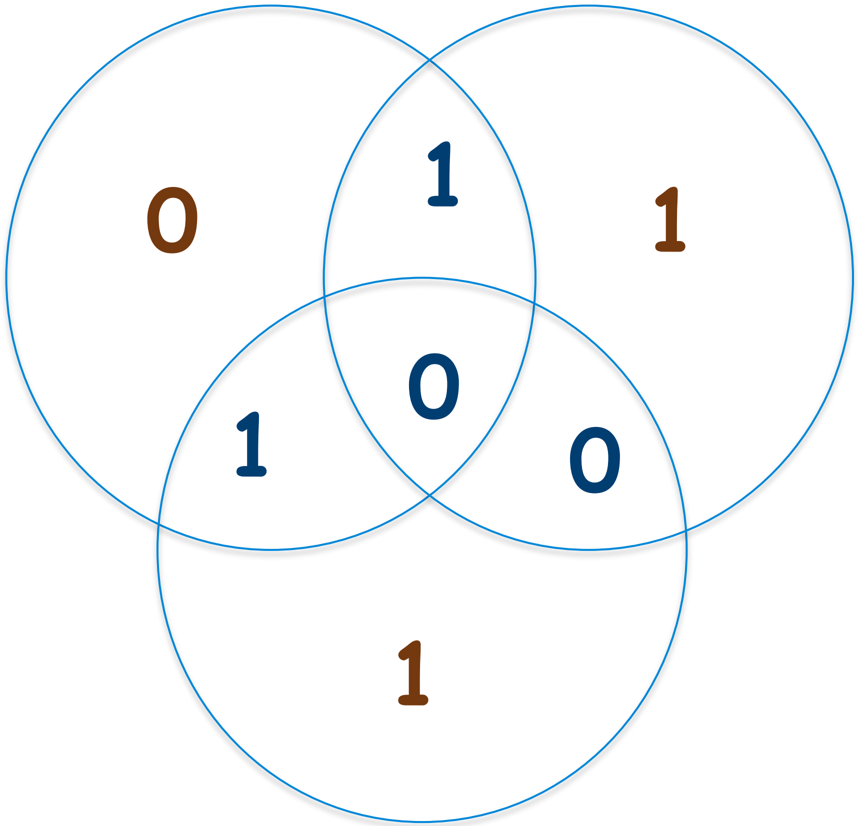
- Load a K=fxf grid of data
- Add 2f+1 redundancy bits
- Total length: $N = f^2+2f+1 = (f+1)^2$
- Information rate: $R = K/N = f^2/(f+1)^2$
- For example, for f=2 we encode K=4 information digits, add 5 redundancy digits for an information rate of 4/9 and correct 1 error
- Can we do better?

1	1	0
0	1	1
1	0	1

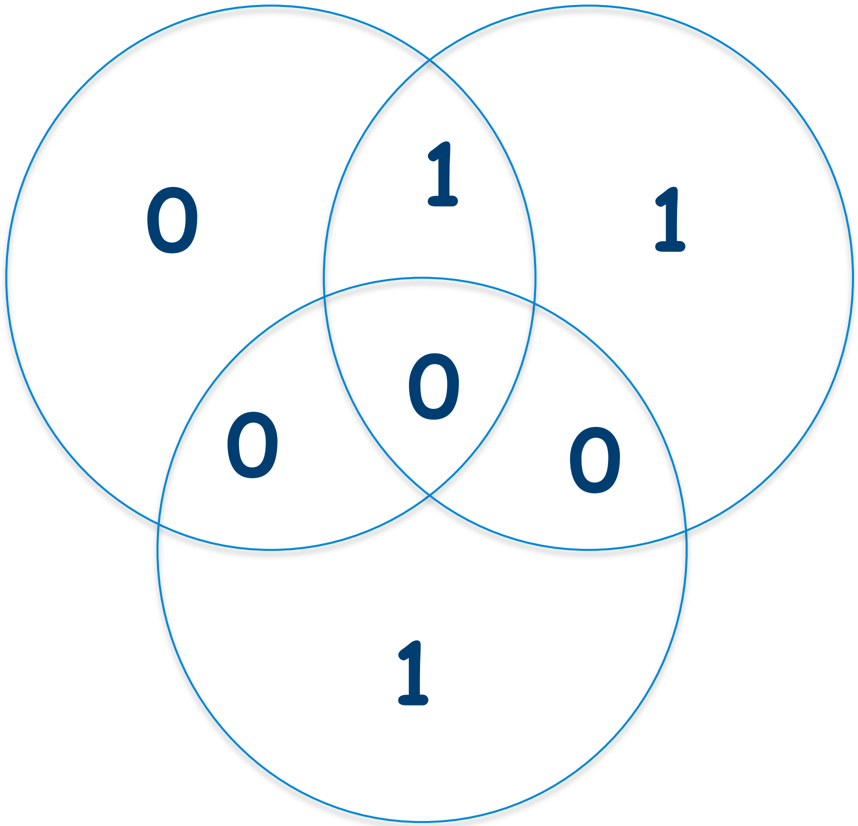
(7,4) Hamming Code



(7,4) Hamming Code



(7,4) Hamming Code



(7,4) Hamming Code Dimensions

- Decoding rule: flip the bit in the intersection of the circles that have the wrong parity
- We encode K=4 information digits and add 3 redundancy digits to transmit N=7 digits
- We can always correct 1 error, at an information rate 4/7 (much better than 4/9)

The best card trick ever…

Analysis

- The "guesser" needs to guess one of 52-4=48 possible cards
- The "guesser" needs to receive $\log_2 48 = 5.58$ bits of information
- The "helper" has a choice among 5 cards to return to the member of the public, followed by a choice of 4x3x2 orderings of the remaining 4 cards, totaling 5! = 120 possibilities
- The "channel" between the helper and the guesser has a capacity to transmit $\log_2 120 = 6.91$ bits of information
- There is ample capacity to comfortably transmit the information the guesser needs
- All you need is a clever code that the "helper" and the "guesser" can work out easily in their heads

An unusual storage channel…

The New York Times: ***Double Helix Serves Double Duty***

Nick Goldman, a molecular biologist at the European Bioinformatics Institute in Hinxton, England, used a technique with error-correction software to store and retrieve data in synthetic DNA molecules.

Channel Coding

- Every communication or storage channel has a capacity that can measured and computed
- Clever coding can achieve any desired error probability for rates below capacity
- Above channel capacity, there is a minimum error probability that cannot be beaten

What we've learned…

- Shannon's legacy:

Source

Remove Bad Redundancy

Add Good Redundancy

From the Bell Labs Webpage

Channel: Communications, Storage

Recover Source

- Information is measureable, like weight and energy
- How much is 1 kg of information?
- On DNA, 2 Petabyte per gram (1 Petabyte is 1000 Terabytes)
- 1kg ≈ the internet (1200 petabytes)