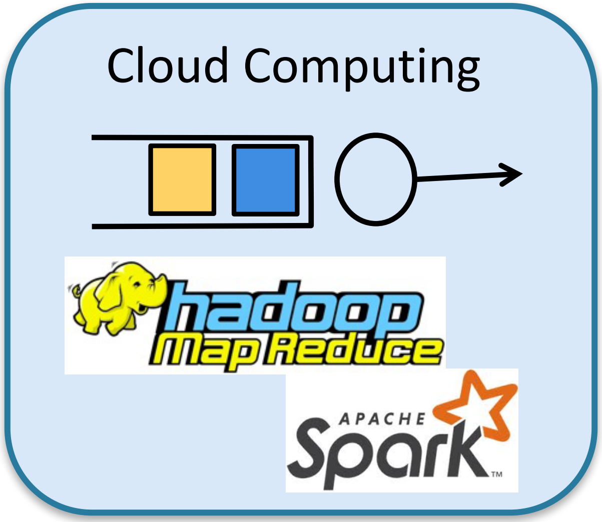
18-847F: Special Topics in Computer Systems

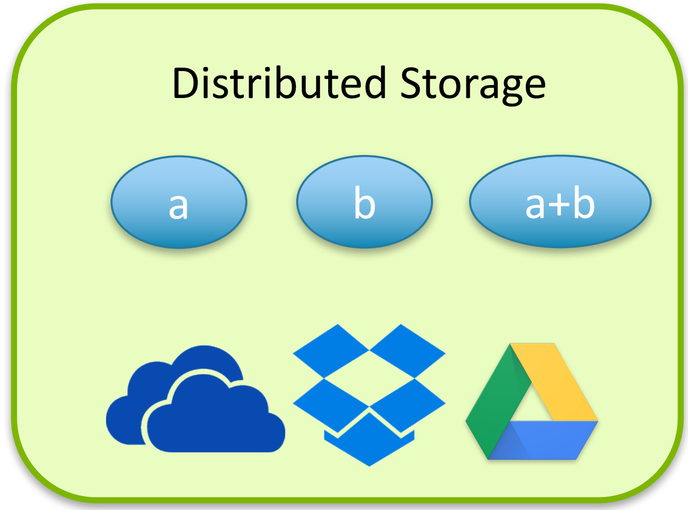
Foundations of Cloud and Machine Learning Infrastructure

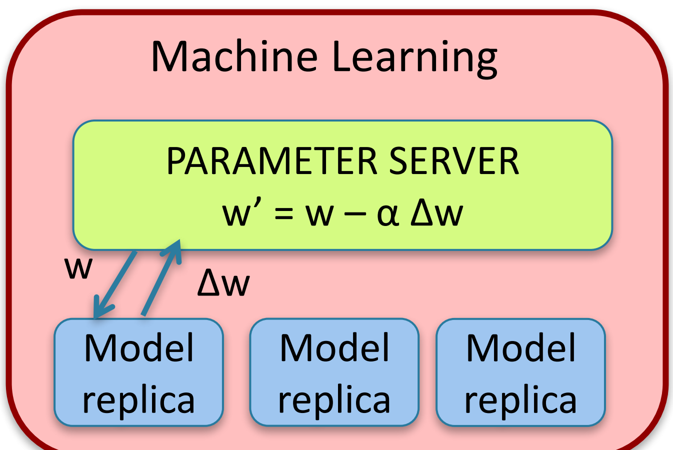
Lecture 8: Intro to Coding Theory

Foundations of Cloud and Machine Learning Infrastructure

Topics Covered

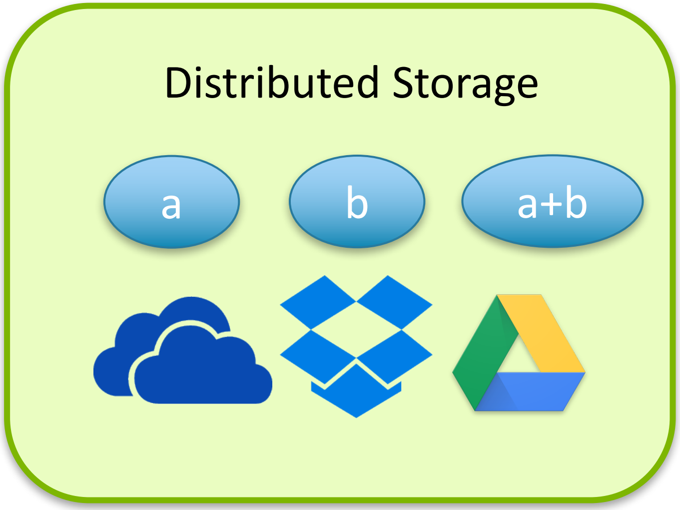






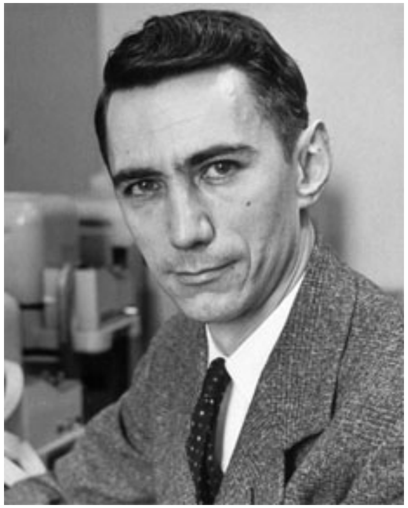
Topics Covered

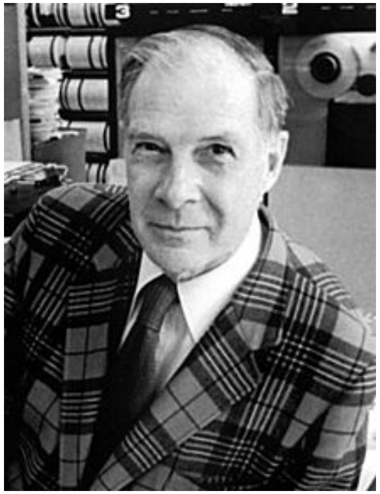
- Coding for locality/repair
- Reducing latency in content download
- Coded Computing



Coding Theory

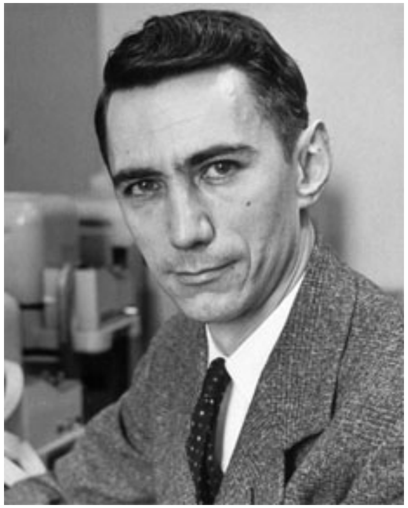
- For reliable communication in presence of noise
- Bell Labs was one of the leaders in 1950's
- Key figures: Claude Shannon and Richard Hamming

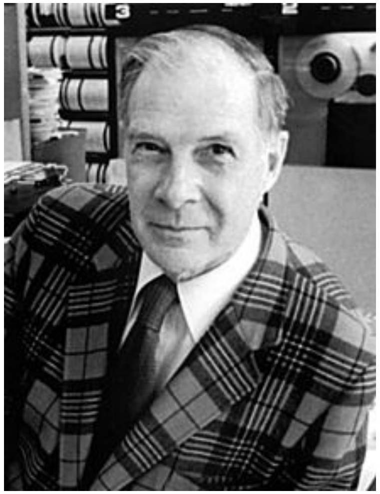




Coding Theory

- Two types of Coding:
  - Source Coding: Data Compression
  - Channel Coding: Error Correction

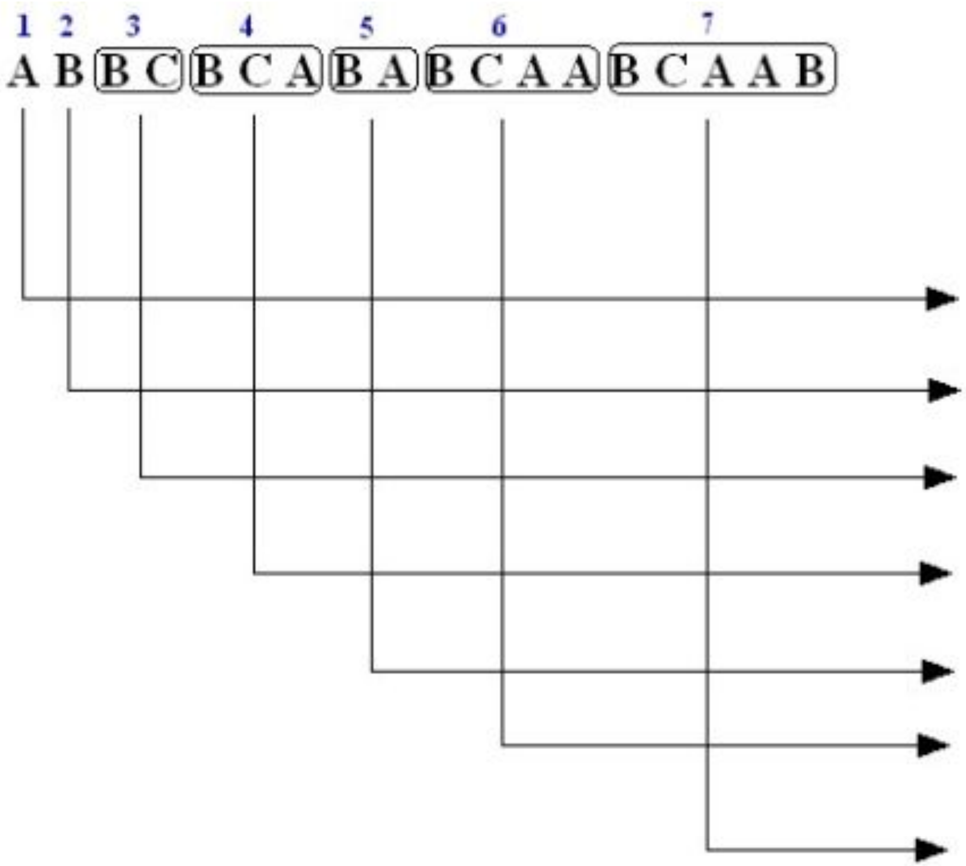




Source Coding

- Huffman Coding
- Zip Data Compression: Lempel-Ziv Coding
- Image/Video Compression: JPEG, MPEG
- Modern applications: Gradient & Model Compression

Source Coding: Lempel-Ziv Coding



output	Dictionary	
	index	string
(0, A)	1	A
(0, B)	2	B
(2, C)	3	BC
(3, A)	4	BCA
(2, A)	5	BA
(4, A)	6	BCAA
(6, B)	7	BCAAB

Simplest Channel Codes

- Repetition Code
  - 0 → 0 0 0 : Rate: 1/3
  - If receive 0 ? ? we can recover from 2 erasures

- (3,2) code: Data bits: a, b Parity bit: (a XOR b)
  - Example: 0 1 1, 1 1 0: Rate 2/3
  - If we receive 0 ? 1 or ? 1 0 we can correct the failed bit
  - 2 bit symbols: (0 1) **?** (1 1)

Simplest Channel Codes

- Repetition Code
  - 0 → 0 0 0 : Rate: 1/3
  - If receive 0 ? ? we can recover from 2 erasures
- (3,2) code: Data bits: a, b Parity bit: (a XOR b)
  - Example: 0 1 1, 1 1 0: Rate 2/3
  - If we receive 0 ? 1 or ? 1 0 we can correct the failed bit
  - 2 bit symbols: (0 1) **(1,0)** (1 1)

Linear Codes: Definition and Notation

Linear Codes

An (n,k) linear code C is a dimension-k subspace of F_q^n, where F_q is a finite field of q elements

Generator Matrix

G is an k x n matrix for code C, if its k rows span C

For an (7,4) binary (q=2) code

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Linear Codes: Definition and Notation

With an (7,4) code, we encode a 4-bit string (a,b,c,d) as

The code is said to be systematic if $G = [I_k \mid A]$

$$\begin{array}{c} a \\ b \\ c \\ d \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

a, b, c, d, a+c+d, a+b+c, b+c+d

Linear Codes: Definition and Notation

Rate of the Code

An (n,k) code has code rate r = k/n

For an (7,4)
binary (q=2) code

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Linear Codes: Definition and Notation

Distance

Minimum Hamming distance between any two codewords. For linear codes, it is the minimum Hamming weight of a non-zero codeword.

Distance = d = 3

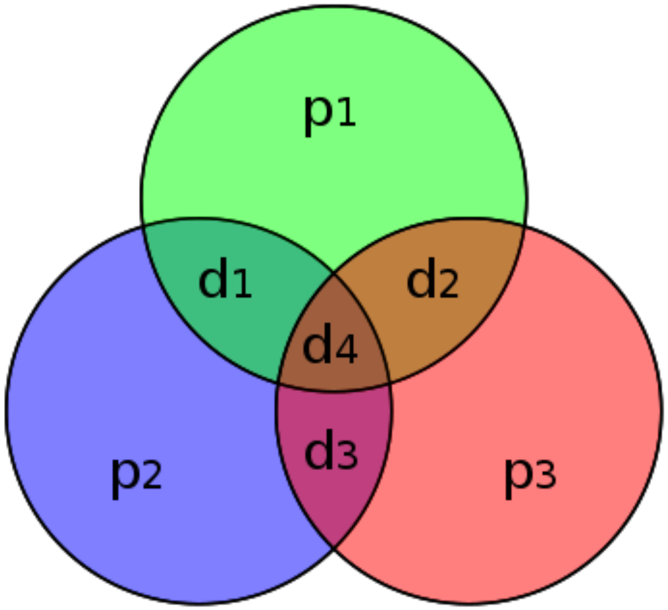
For an (7,4) binary (q=2) code

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

Codes with d = n-k+1 are called maximum-distance separable (MDS) codes

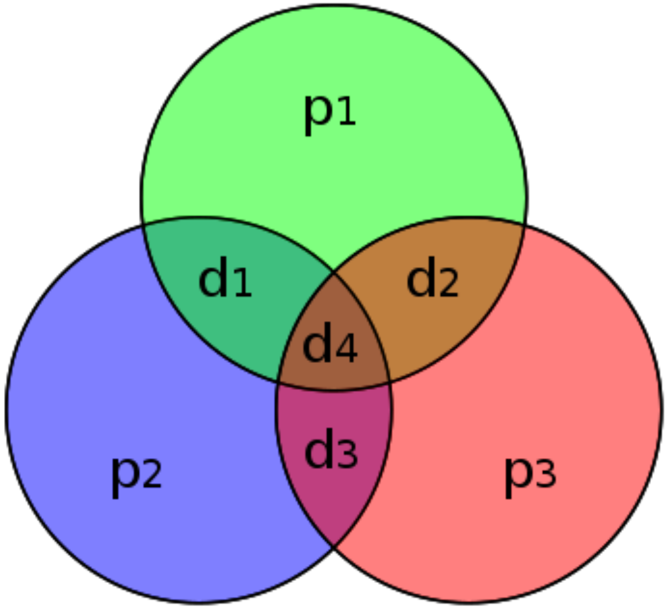
Hamming Codes

- (7,4) Hamming Code: 4 data bits, 3 parity bits
- Parity $p_1 = d_1 \oplus d_2 \oplus d_4$
- Can correct 1-bit errors or 2-bit erasures
- Can detect 1 or 2-bit errors



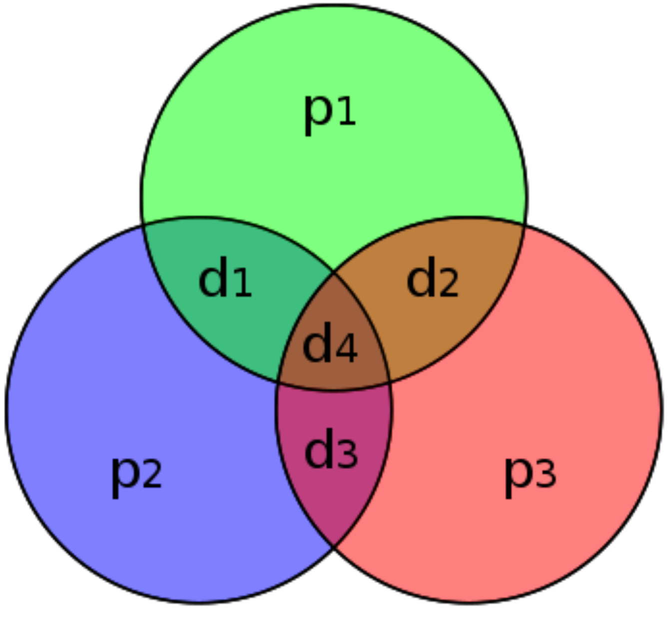
Concept Check: Erasure Codes

- What is the rate and distance of this code?
- Correct the 2 erasures
    - $(d_1, d_2, d_3, d_4, p_1, p_2, p_3) = (0, ?, 1, ?, 1, 0, 0)$



Concept Check: Answer

- What is the rate of the code? $r = 4/7$, $d = 3$
- Correct the 2 erasures
  - $(d_1, d_2, d_3, d_4, p_1, p_2, p_3) = (0, 0, 1, 1, 1, 0, 0)$



(n,k) Reed-Solomon Codes: 1960

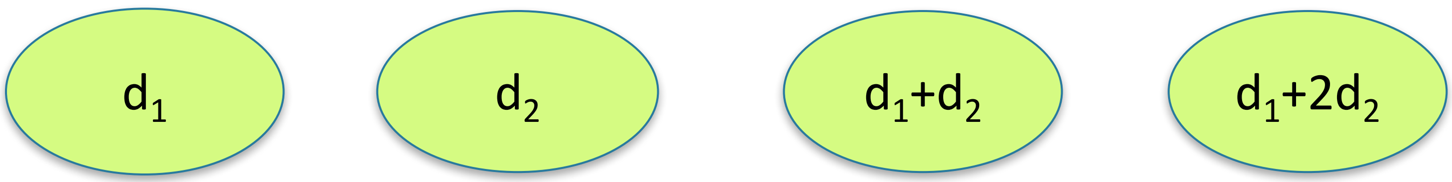
- Data: $d_1, d_2, d_3, \ldots d_k$
- Polynomial: $d_1 + d_2 x + d_3 x^2 + \ldots d_k x^{k-1}$
- Parity bits: Evaluate at n-k points:

x=1:	$d_1 + d_2 + d_3 + d_4$
x=2:	$d_1 + 2 d_2 + 4 d_3 + 8 d_4$
x=3 :	….
x=4:	….
x=n:	…

- Can solve for the coefficients from any k coded symbols

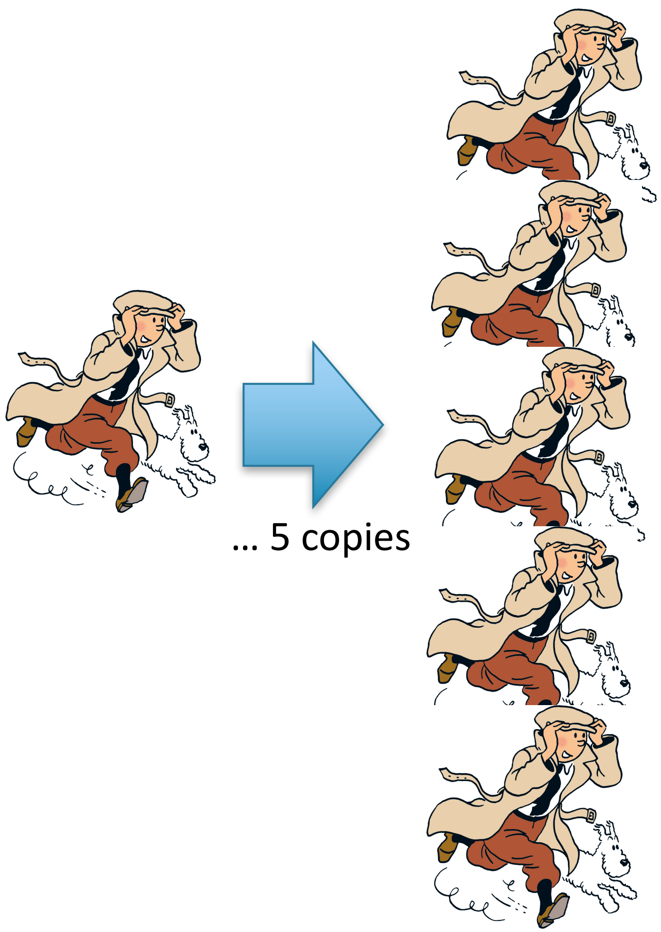
Example: (4,2) Reed-Solomon Code

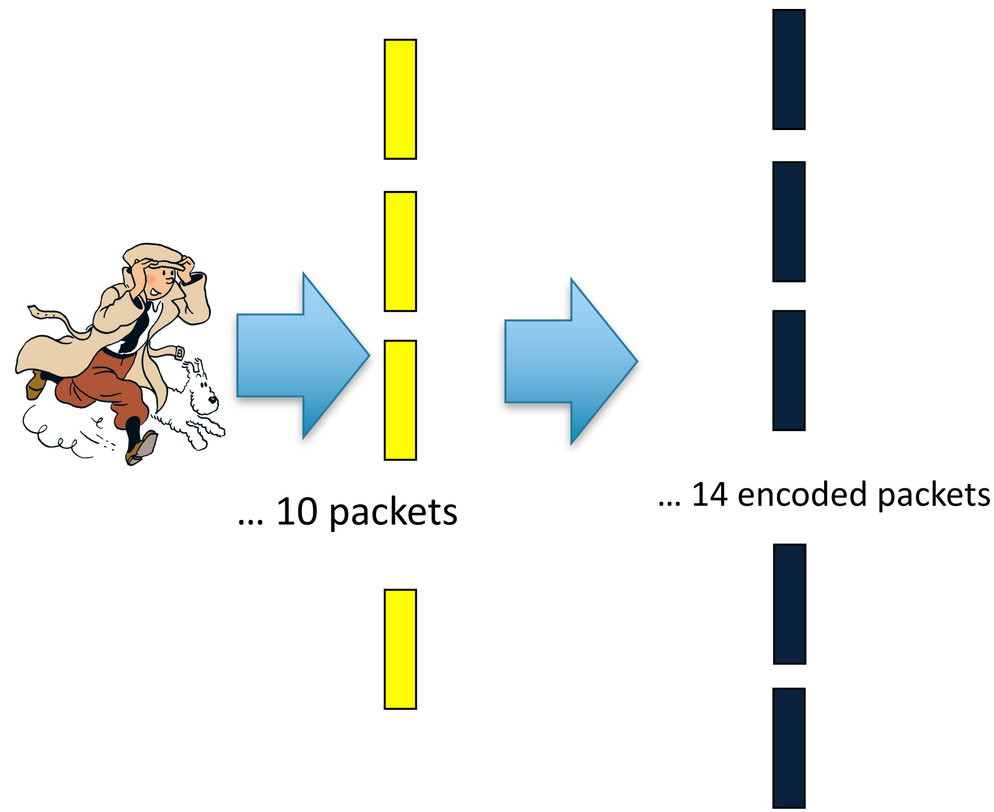
- Data: $d_1, d_2 \rightarrow$ Polynomial: $d_1 + d_2 x + d_3 x^{2} + \dots d_k x^{k-1}$



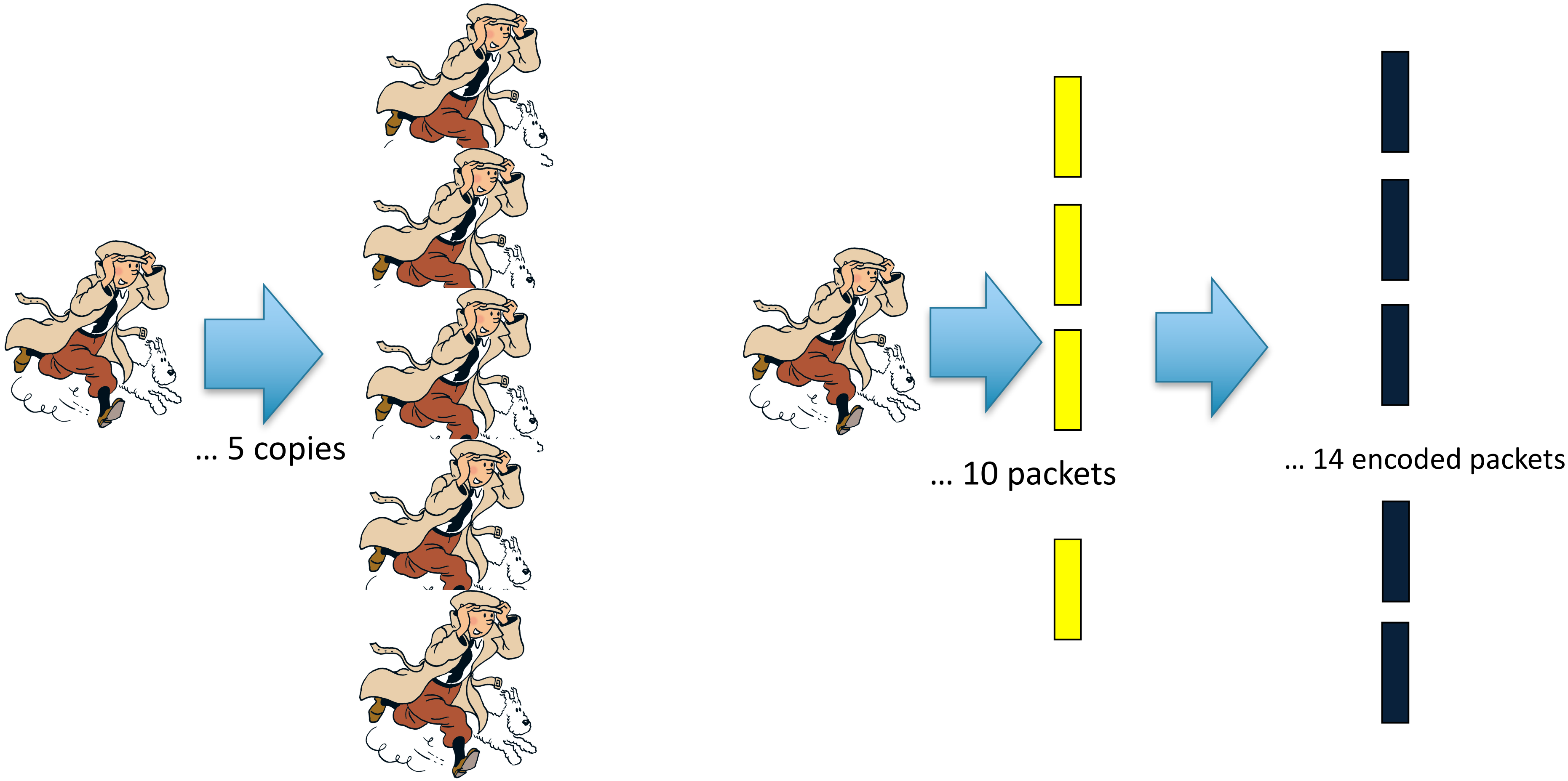
- Can solve for the coefficients from any k coded symbols
- Microsoft uses (7, 4) code
- Facebook uses (14,10) code

Coding vs Replication



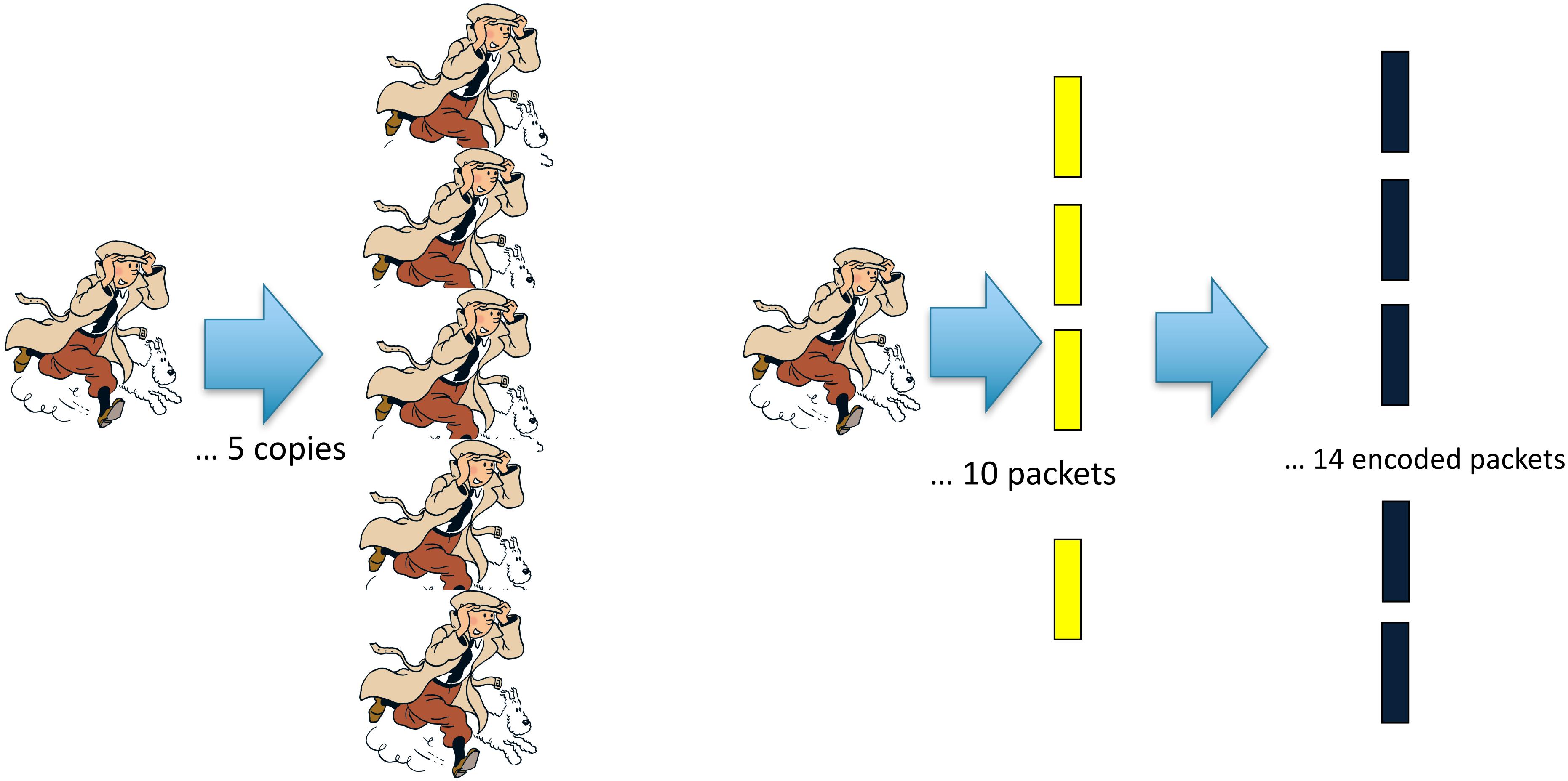


Concept Check: Coding vs Replication



- How many node-failures can each system tolerate?
- What is the code rate of each system?

Concept Check: Coding vs Replication



- How many node-failures can each system tolerate?: 4
- What is the code rate of each system? 1/5 and 10/14
- Replication uses 357% more storage for the same reliability!

RAID: Redundant Array of Independent Disks (1987)

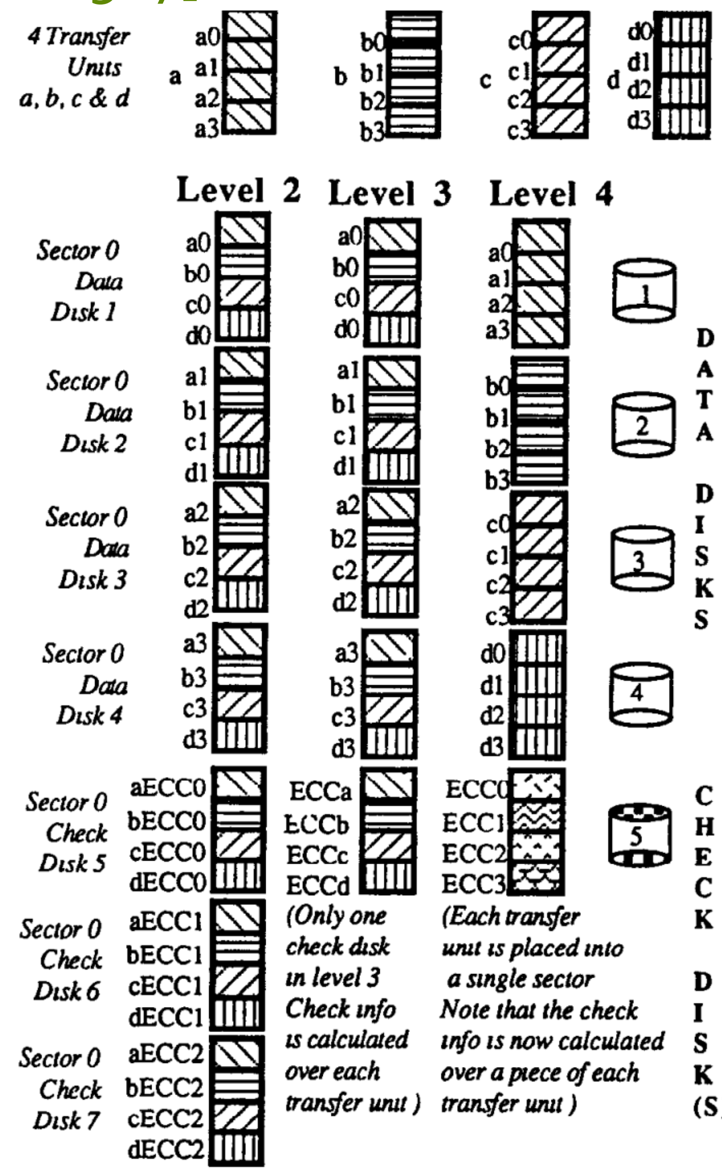
- Levels RAID 0, RAID 1, … : design for different goals such as reliability, availability, capacity etc.

- One of the inventors, Garth Gibson was here at CMU

RAID: Redundant Array of Independent Disks

[Patterson et al 1987]

- RAID 1: Replication
- RAID 2: The (7,4) Hamming code
  Detect 2 errors, correct 1
- RAID 3: Only parity check disk, used for error correction
- RAID 4: Bit interleaving to allow parallel reads/writes
- RAID 5: Spread check and data bits across all disks



RAID: Redundant Array of Independent Disks

[Patterson et al 1987]

- RAID 1: Replication
- RAID 2: The (7,4) Hamming code
  Detect 2 errors, correct 1
- RAID 3: Only parity check disk, used for error correction
- RAID 4: Bit interleaving to allow parallel reads/writes
- RAID 5: Spread check and data bits across all disks