Achieving Reliable Communications and Data Storage Using Error-Correcting Codes

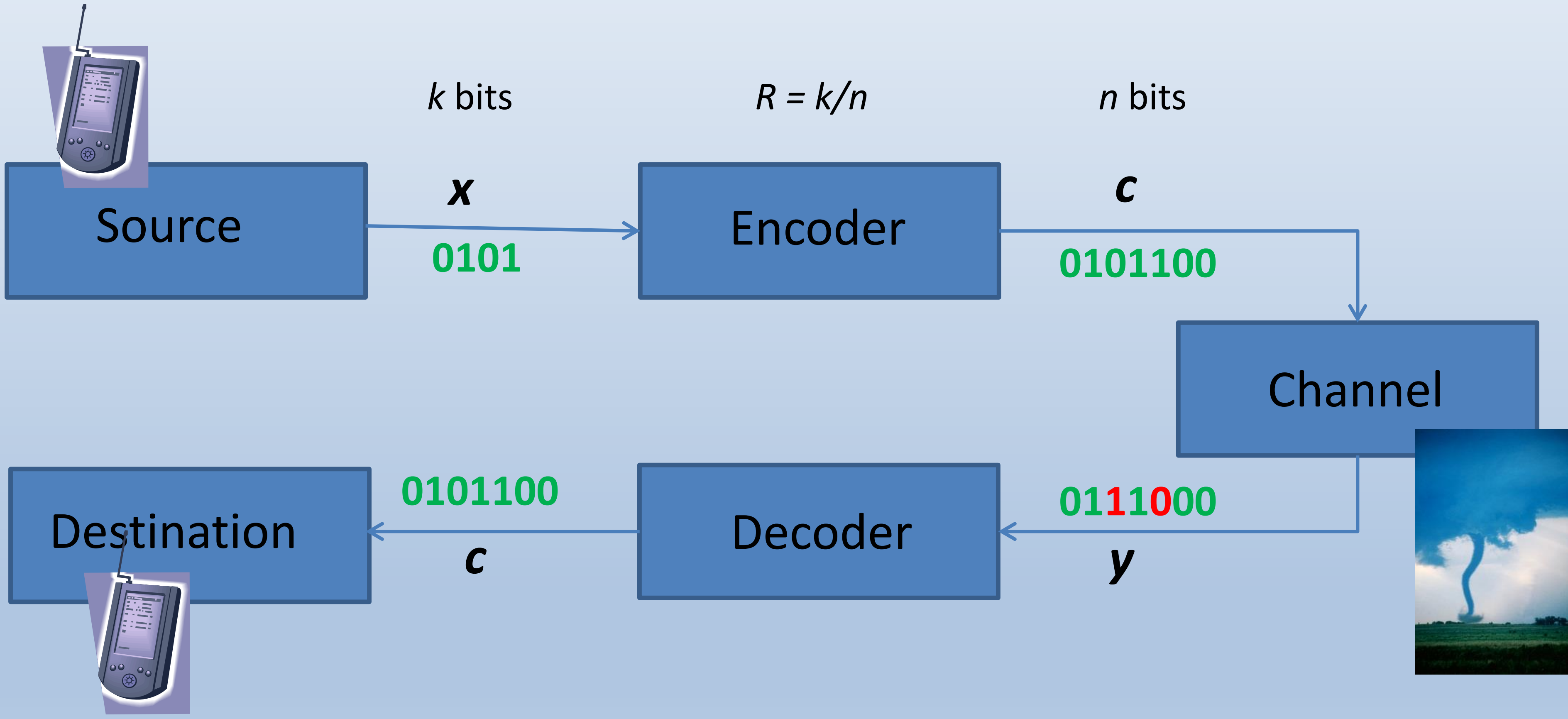
Vitaly Skachek

University of Tartu

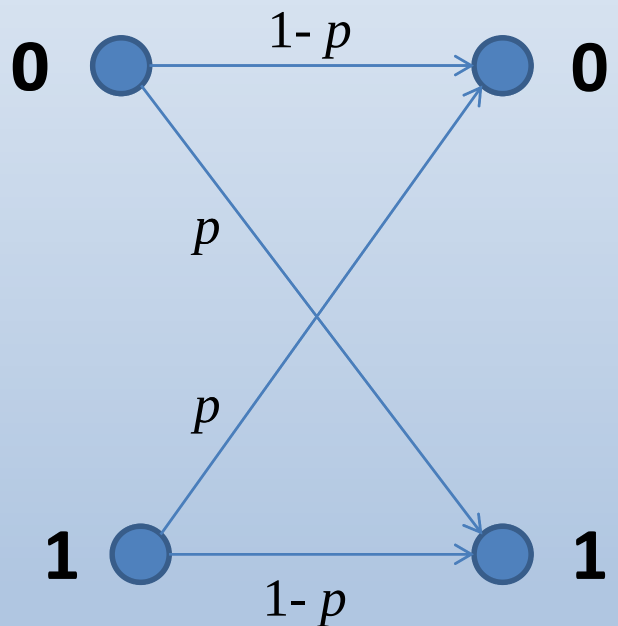
April 22nd, 2025

Some used images are courtesy of Wikipedia/Wikimedia Commons

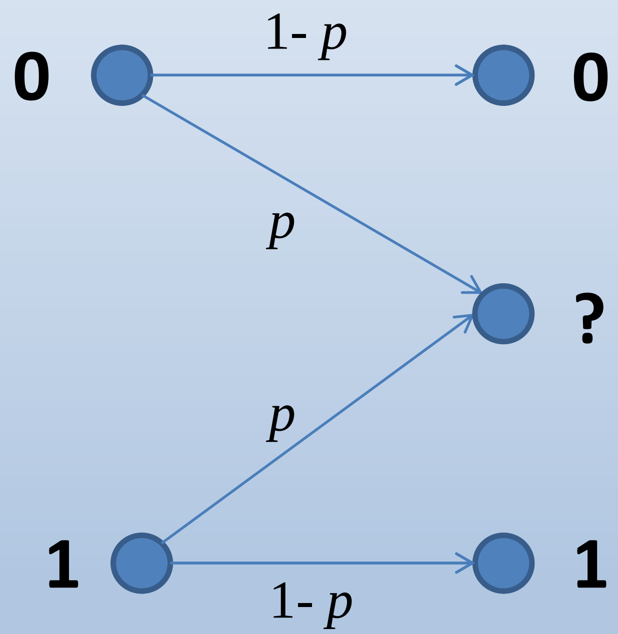
Shannon-Weaver Communications Model



Communications Channels



Binary Symmetric Channel



Binary Erasure Channel

Shannon's Channel Coding Theorems

- A code is a mapping from the set of all vectors of length *k* to a set of vectors of length *n* (over alphabet Σ)

- Given a channel *S*, there is a quantity *C*(*S*) called channel capacity

Claude Shannon
(1916-2001)

Shannon's Channel Coding Theorems

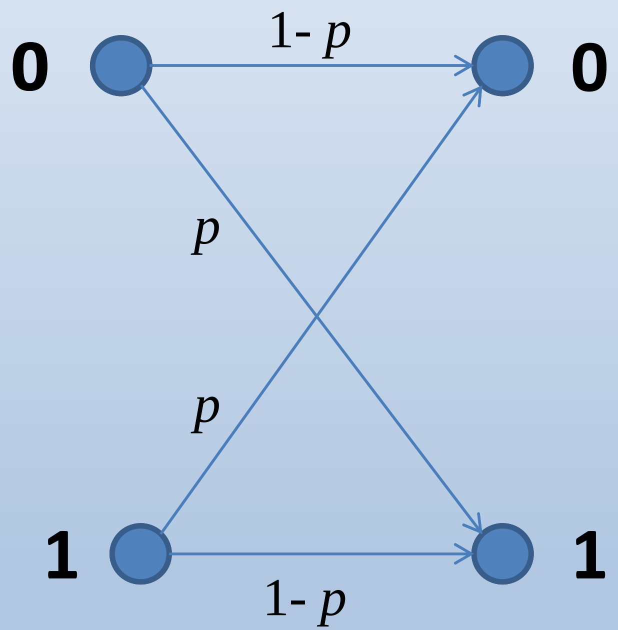
For any rate $R < C(S)$, there exists an infinite sequence of block codes C_i of growing lengths n_i such that $\frac{k_i}{n_i} \geq R$, and there exists a coding scheme for those codes such that the decoding error probability approaches 0 as $i \rightarrow \infty$.

Shannon's Channel Coding Theorems

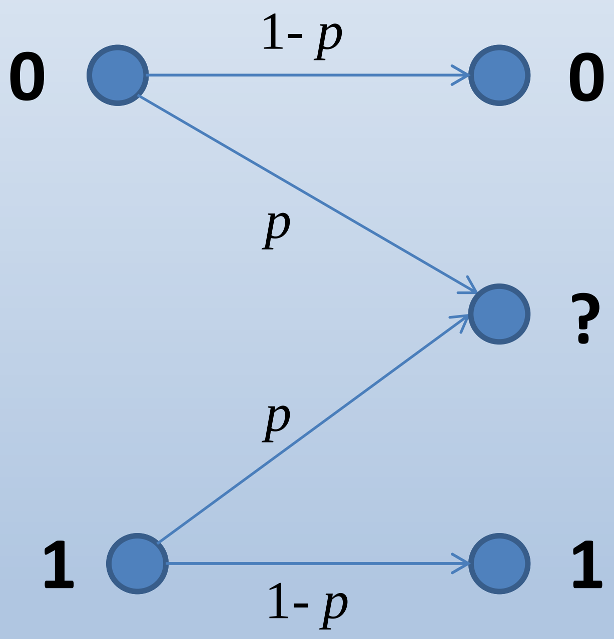
For any rate *R* < *C*(*S*), there exists an infinite sequence of block codes C_i of growing lengths n_i such that $\frac{k_i}{n_i} \geq R$, and there exists a coding scheme for those codes such that the decoding error probability approaches 0 as $i \rightarrow \infty$.

Let *R* > *C*(*S*). For any infinite sequence of block codes C_i of growing lengths n_i such that $\frac{k_i}{n_i} \geq R$, and for any coding scheme for those codes, the decoding error probability is bounded away from 0 as $i \rightarrow \infty$.

Communications Channels



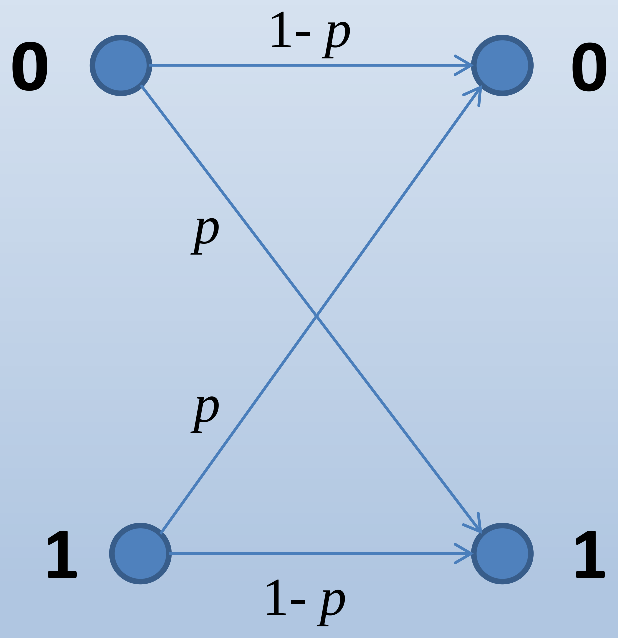
Binary Symmetric Channel



Binary Erasure Channel

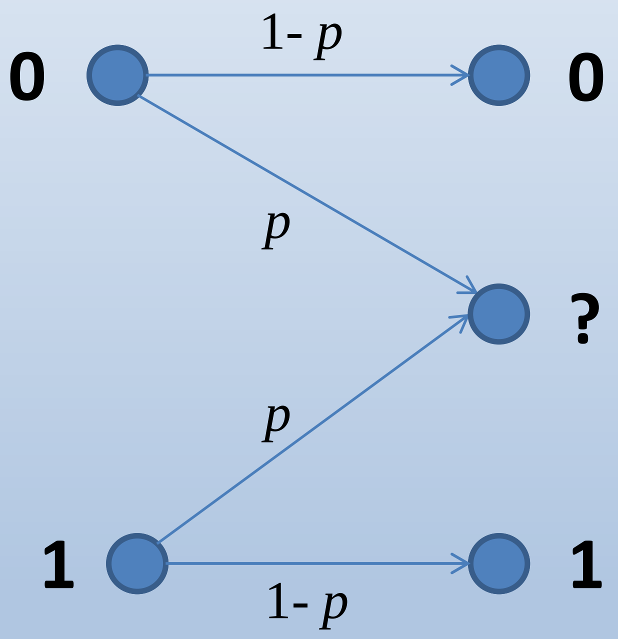
Communications Channels

$C(S)=1-h_2(p)$



Binary Symmetric Channel

$C(S)=1-p$

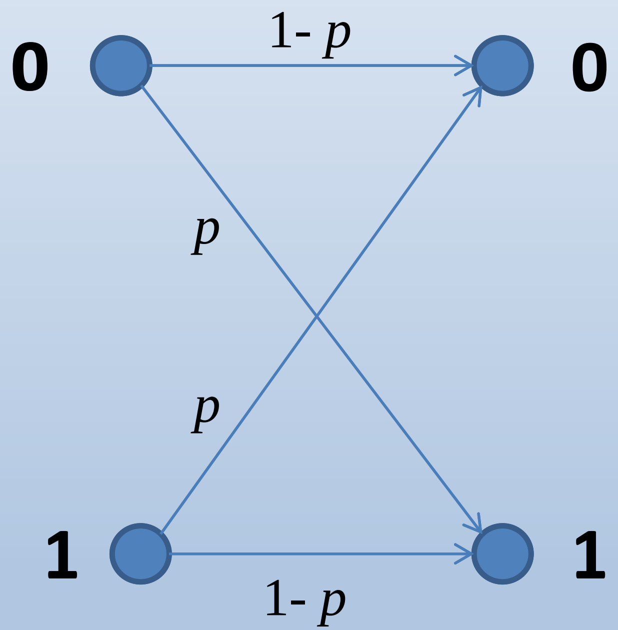


Binary Erasure Channel

Communications Channels

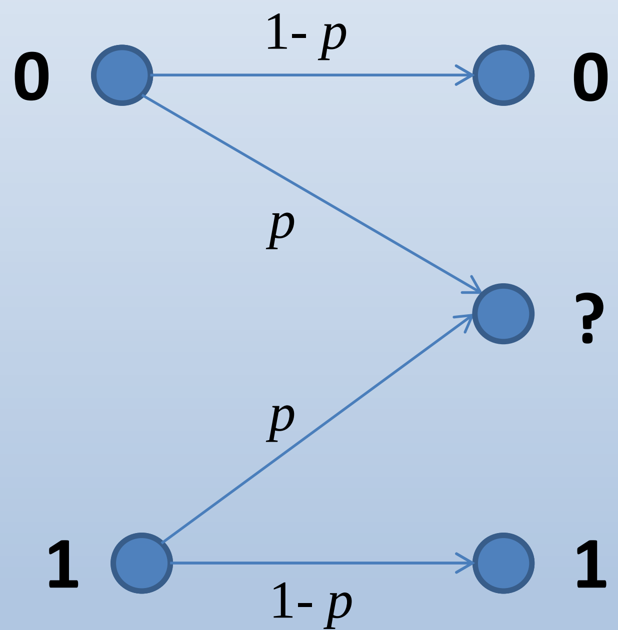
$C(S)$=1-$h_2(p)$

$h_2(x) = -x\ \mathrm{log}x - (1-x)log(1-x)$



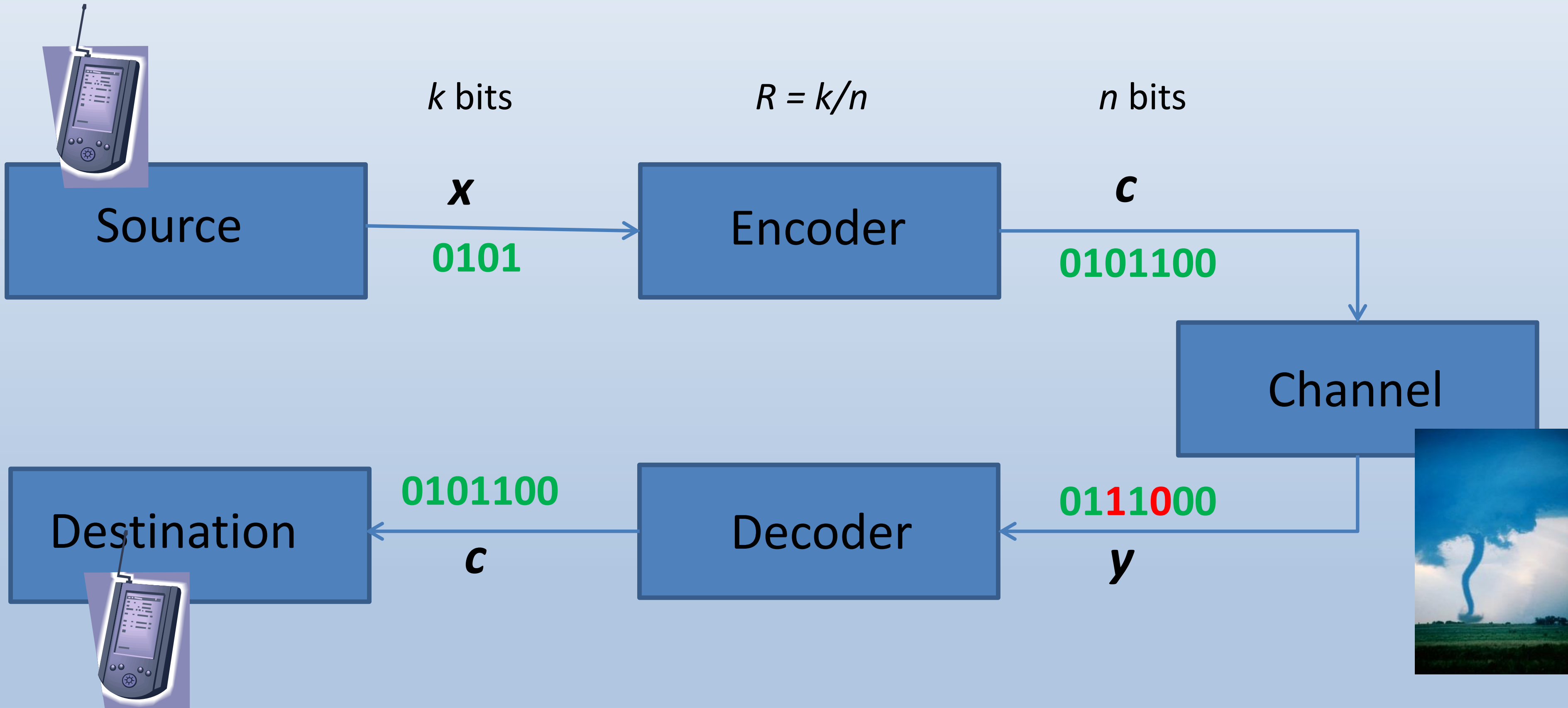
Binary Symmetric Channel

$C(S)$=1-p



Binary Erasure Channel

Communications Model



Parameters to Consider

Other parameters to consider:

- Speed of convergence Pr (err) $\to 0$ as $n \to \infty$. Low error probability for short lengths is needed!
- Time complexity of encoding and decoding algorithms. Structured codes are needed!

Minimum Distance

- The Hamming distance between $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$, $d(x, y)$, is the number of pairs of symbols (x_i, y_i), such that $x_i \neq y_i$.

- The minimum distance of a code C is

$$d = \min_{\{x, y \in C, x \neq y\}} d(x, y)$$

Linear Codes

- A code C over field *F* is a linear [*n*, *k*, *d*] code if there exists a matrix H with *n* columns and rank *n* – *k* such that
$$H \cdot c^T = 0^T \iff c \in C.$$

$k-1$

- The matrix *H* is called a parity-check matrix.
- The value *k* is called the dimension of the code C.
- The ratio *R* = *k*/*n* is called the rate of the code C.
- All vectors (codewords) of C are exactly all linear combinations of rows of a *k* × *n* generator matrix *G*.

Challenge

- Large values of $R = k/n$ correspond to high efficiency of transmission.
- Large values of d correspond to high error resilience.

$\Rightarrow$ We want to make k and d as large as possible at the same time.

Bounds

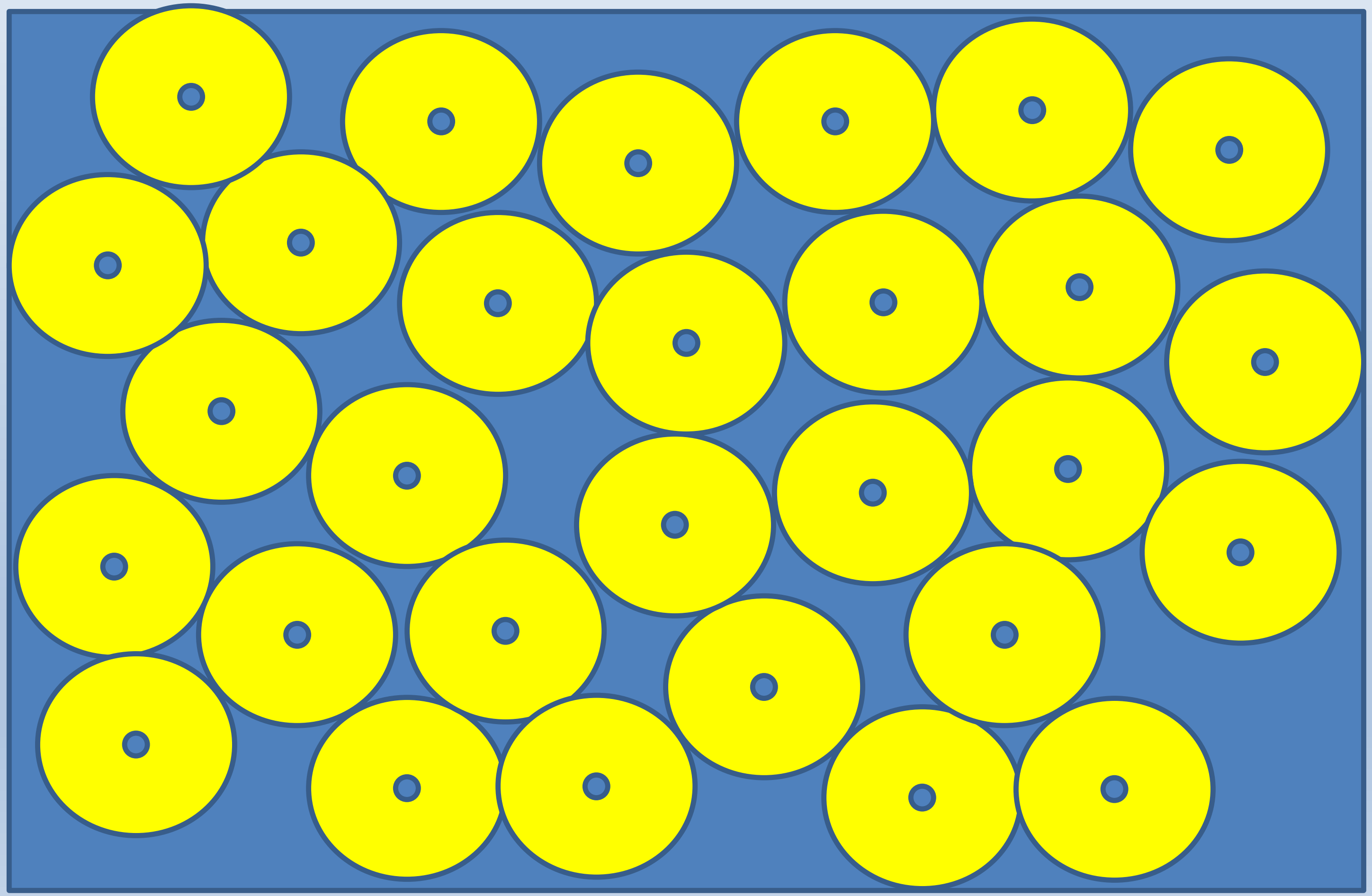
- Sphere-packing (Hamming) bound:

$$\sum_{i=0}^{\lfloor \frac{d-1}{2} \rfloor} \binom{n}{i} (q-1)^i \leq q^{n-k}$$

- Singleton bound:

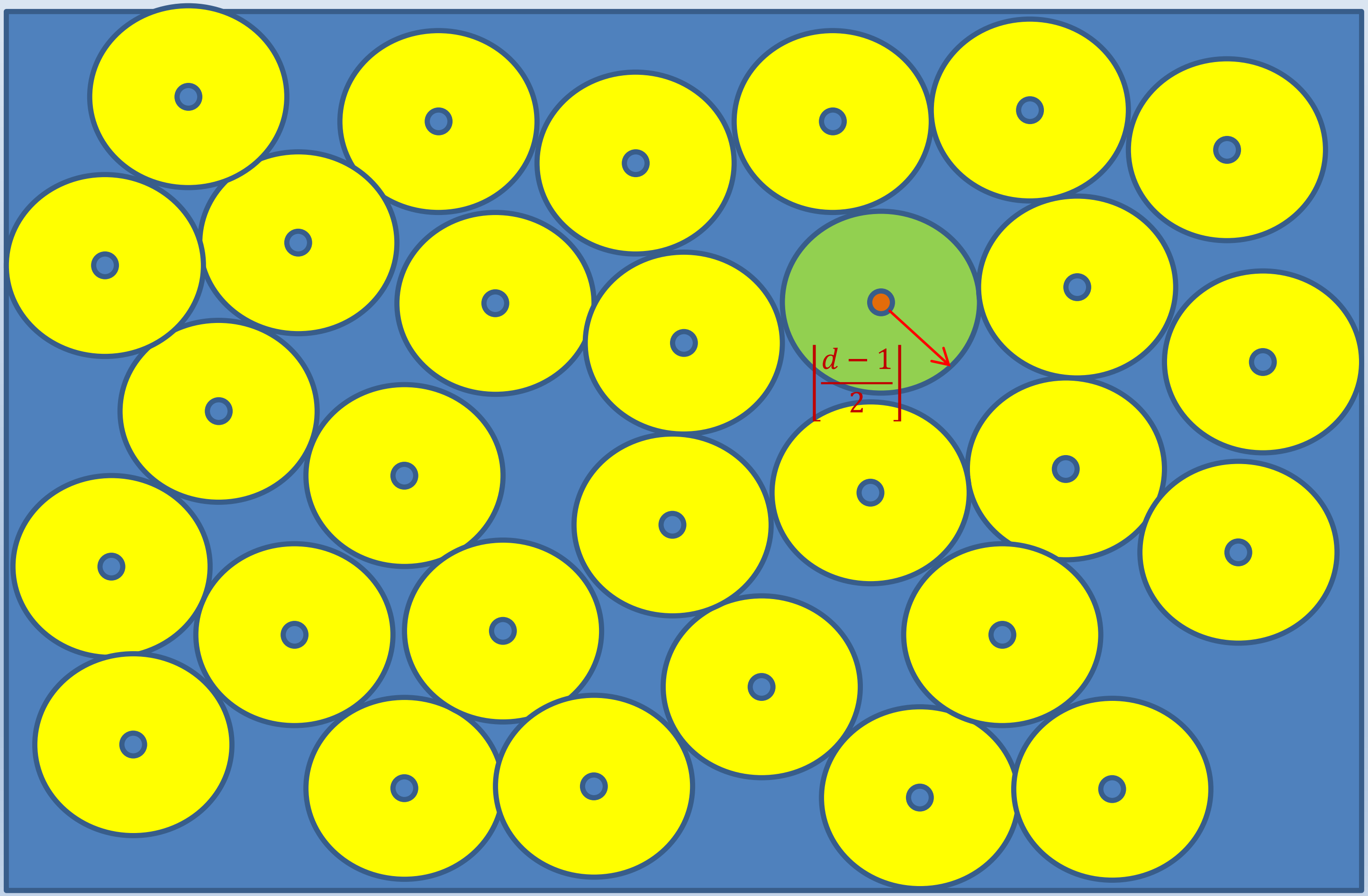
$$d + k - 1 \leq n.$$

Sphere-packing idea

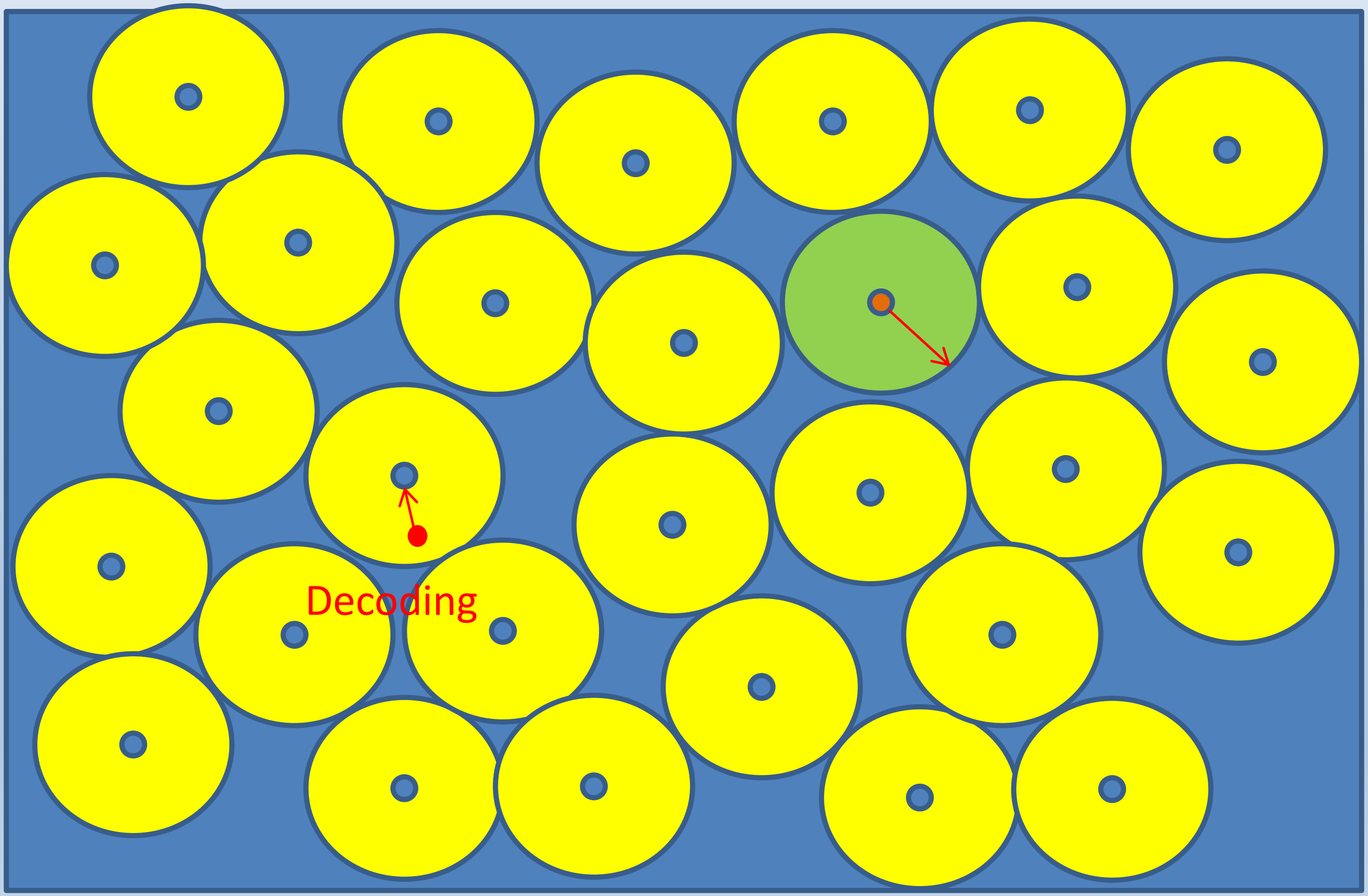


Sphere-packing idea

Sphere-packing idea



Sphere-packing idea



Hamming Codes

- The *m* x *n* binary parity-check matrix:

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Here: $n = 2^m - 1, k = 2^m - m - 1, d = 3$.

- Attains the sphere-packing bound.
  - Optimal trade-off between the parameters
- Used in DRAM memory chips and satellite communications.

Reed-Solomon Codes

- Let $\alpha_1, \alpha_2, \dots, \alpha_n \in F$ be *n* distinct nonzero elements of the final field F.
- The generator matrix:

$$G = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \dots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \dots & \alpha_n^{k-1} \end{bmatrix}$$

- Attains the Singleton bound: $n = d + k - 1$
  - Optimal trade-off between the parameters

Reed-Solomon Codes (cont.)

- Encoding:

$$[x_0 x_1 \ldots x_{k-1}] \cdot$$

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \ldots & \alpha_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \ldots & \alpha_n^{k-1} \end{bmatrix}$$

Polynomial Interpolation Viewpoint

- Input vector $[x_0 x_1 \ldots x_{k-1}]$ is associated with the polynomial
$$P(z) = x_{k-1}z^{k-1} + x_{k-2}z^{k-2} + x_1 z + x_0$$

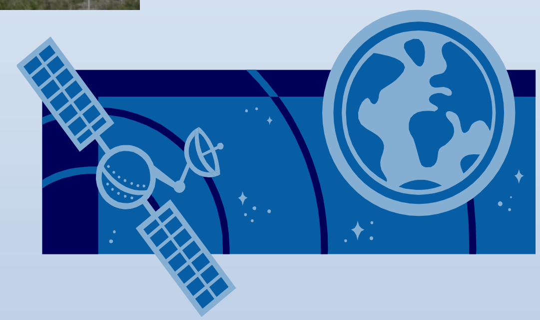
- Encoding is an evaluation:
$$\left(P(\alpha_1), P(\alpha_2), \ldots, P(\alpha_n)\right)$$

- Decoding is an interpolation of the polynomial of degree $\leq k-1$

Reed-Solomon Codes are Used in:

- Wired and wireless communications

- Satellite communications

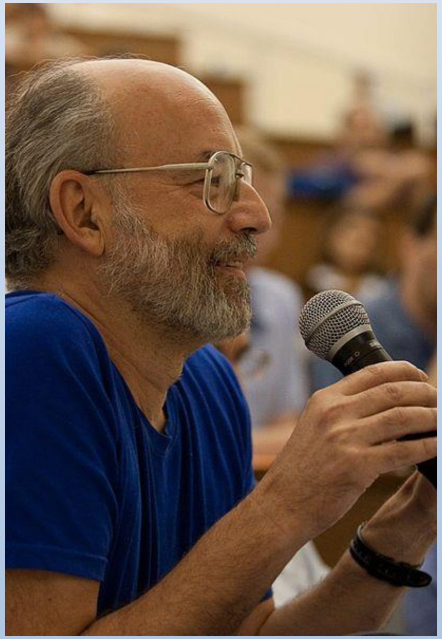


- Hard drives and compact disks

- Flash memory devices

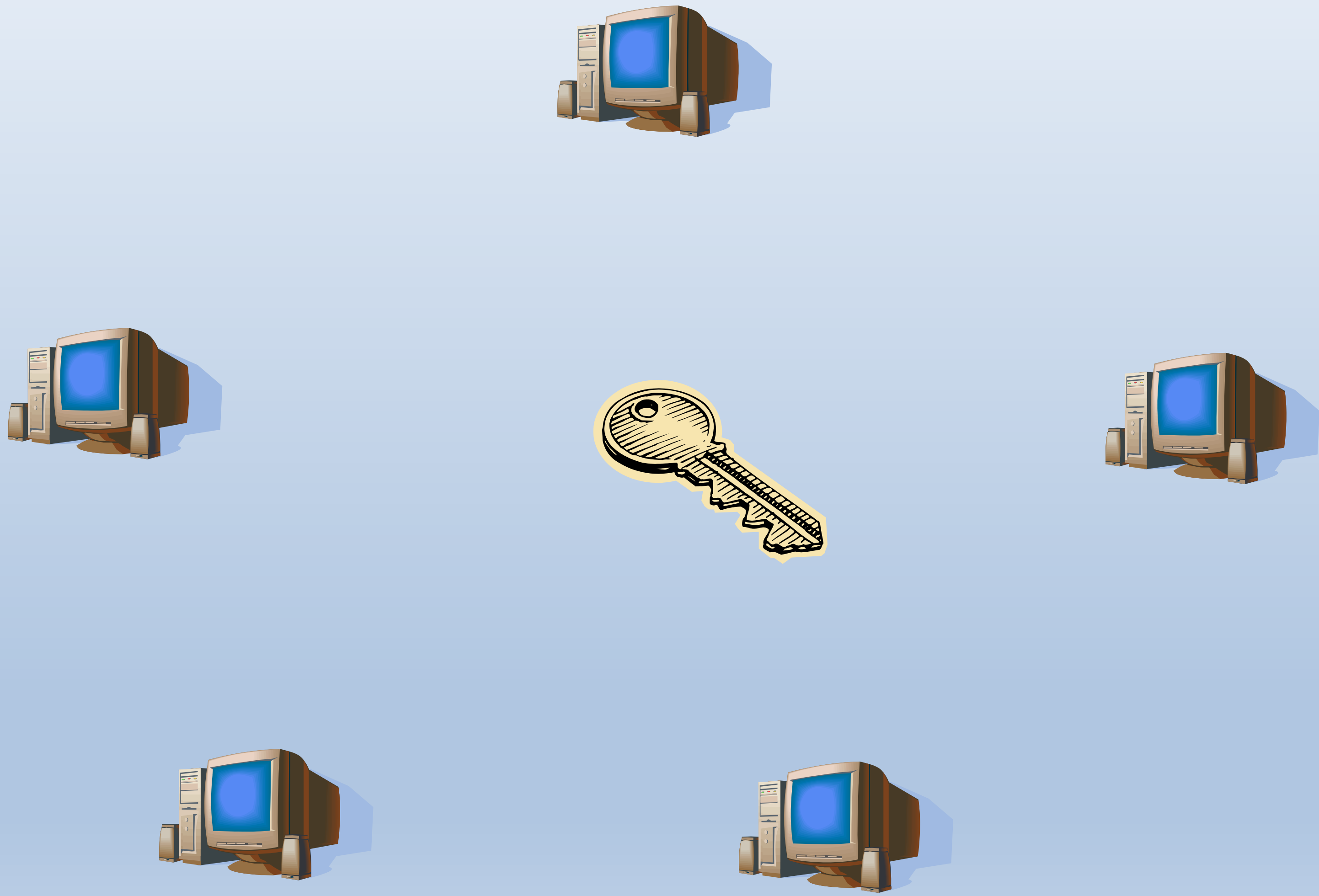
Application of Reed-Solomon Codes

- Shamir's Secret-Sharing Scheme '79
- n users
- 1 key (element of F)
- Any coalition of $< t$ users does not have any information about the key
- Any coalition of $\geq t$ users can recover the key

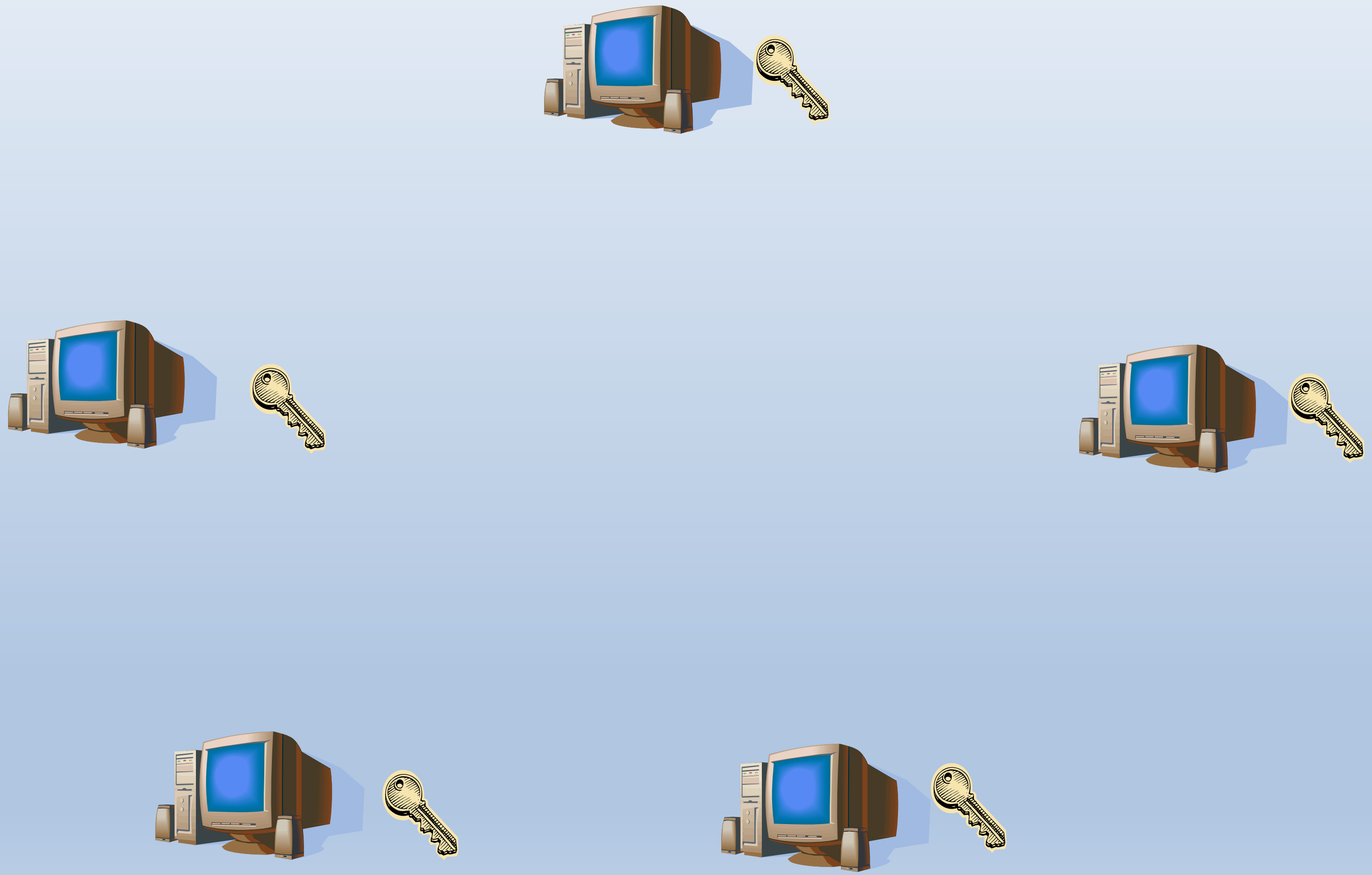


Adi Shamir

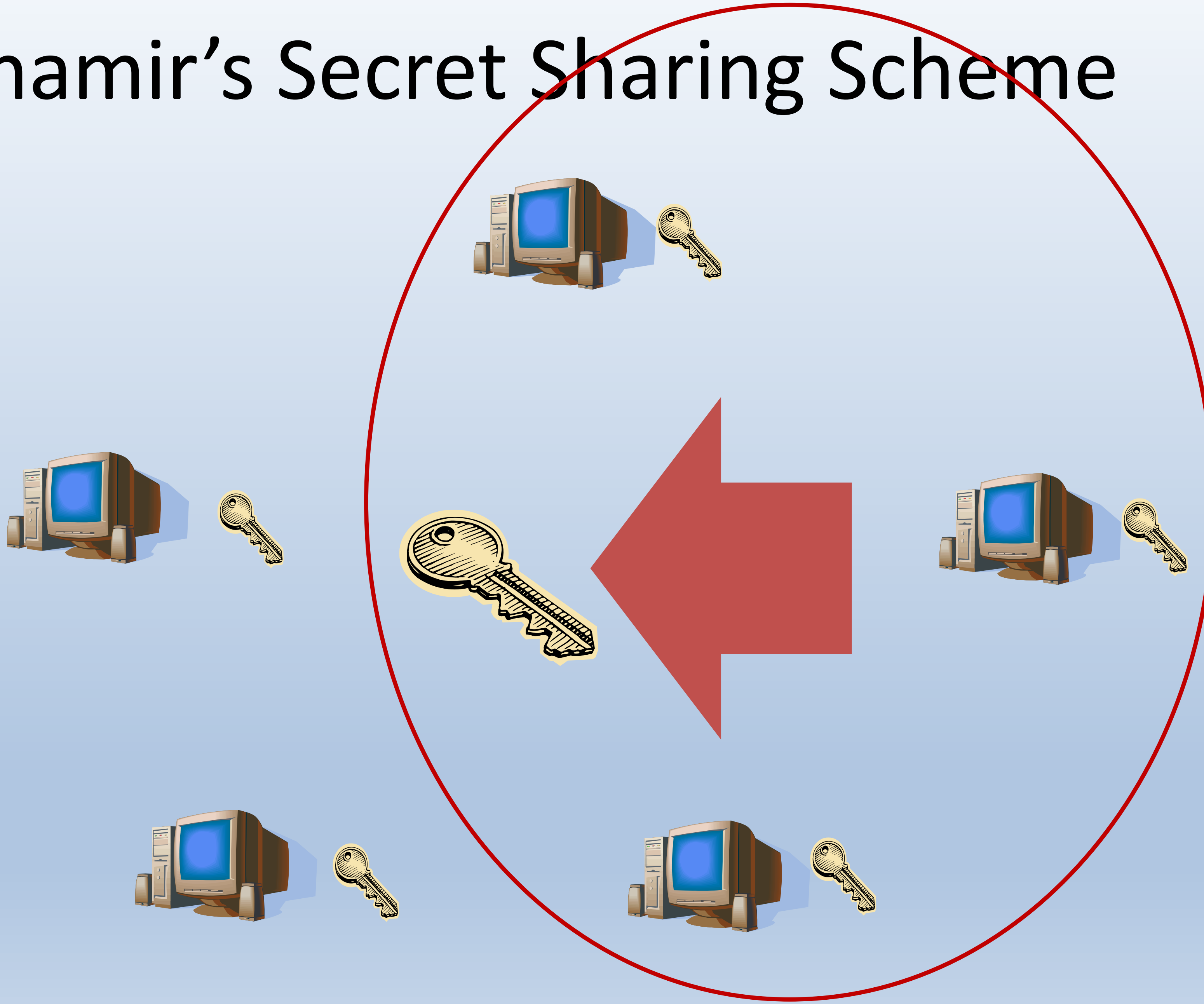
Shamir's Secret Sharing Scheme



Shamir's Secret Sharing Scheme



Shamir's Secret Sharing Scheme

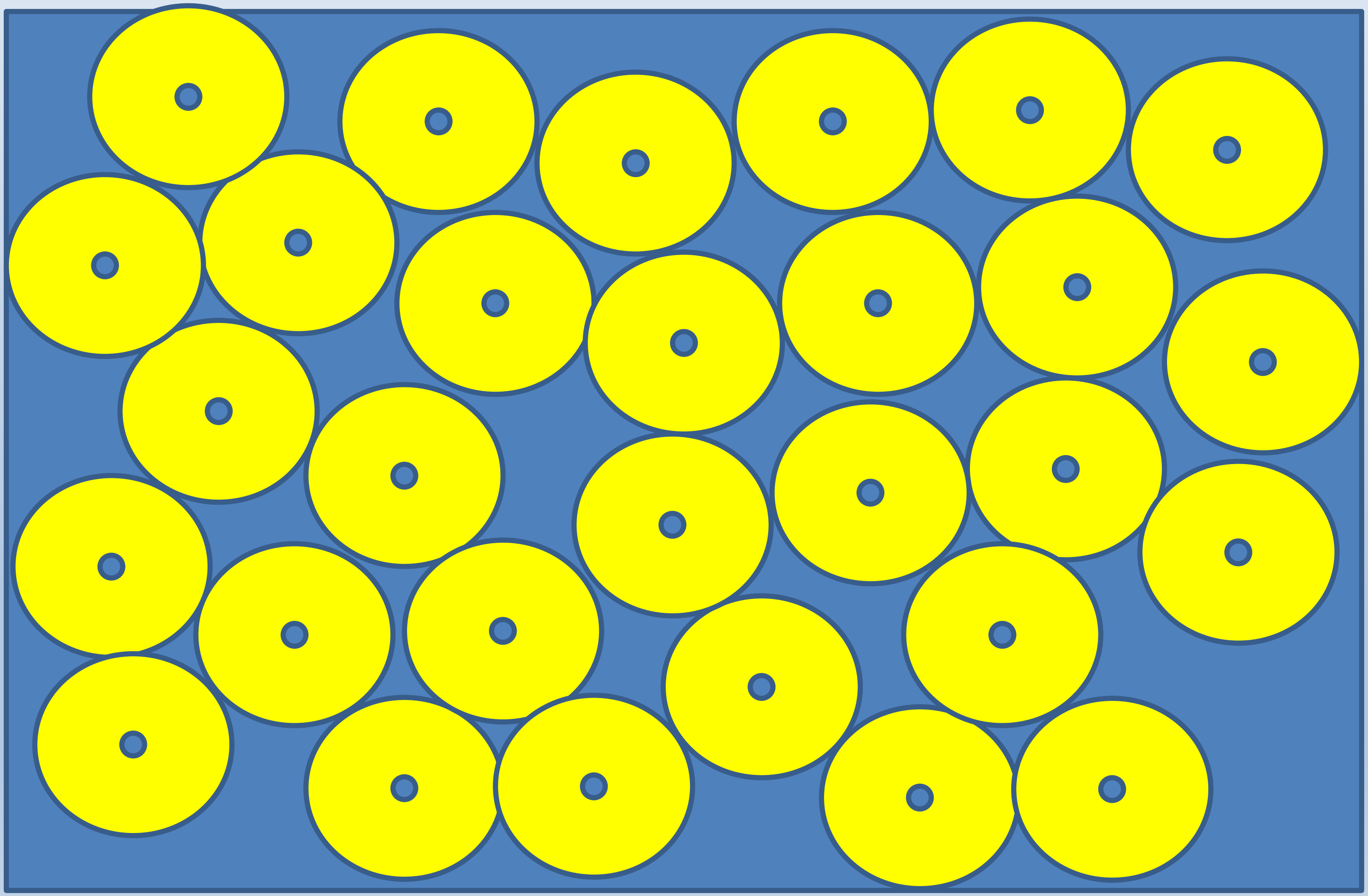


Shamir's Secret Sharing Scheme (cont.)

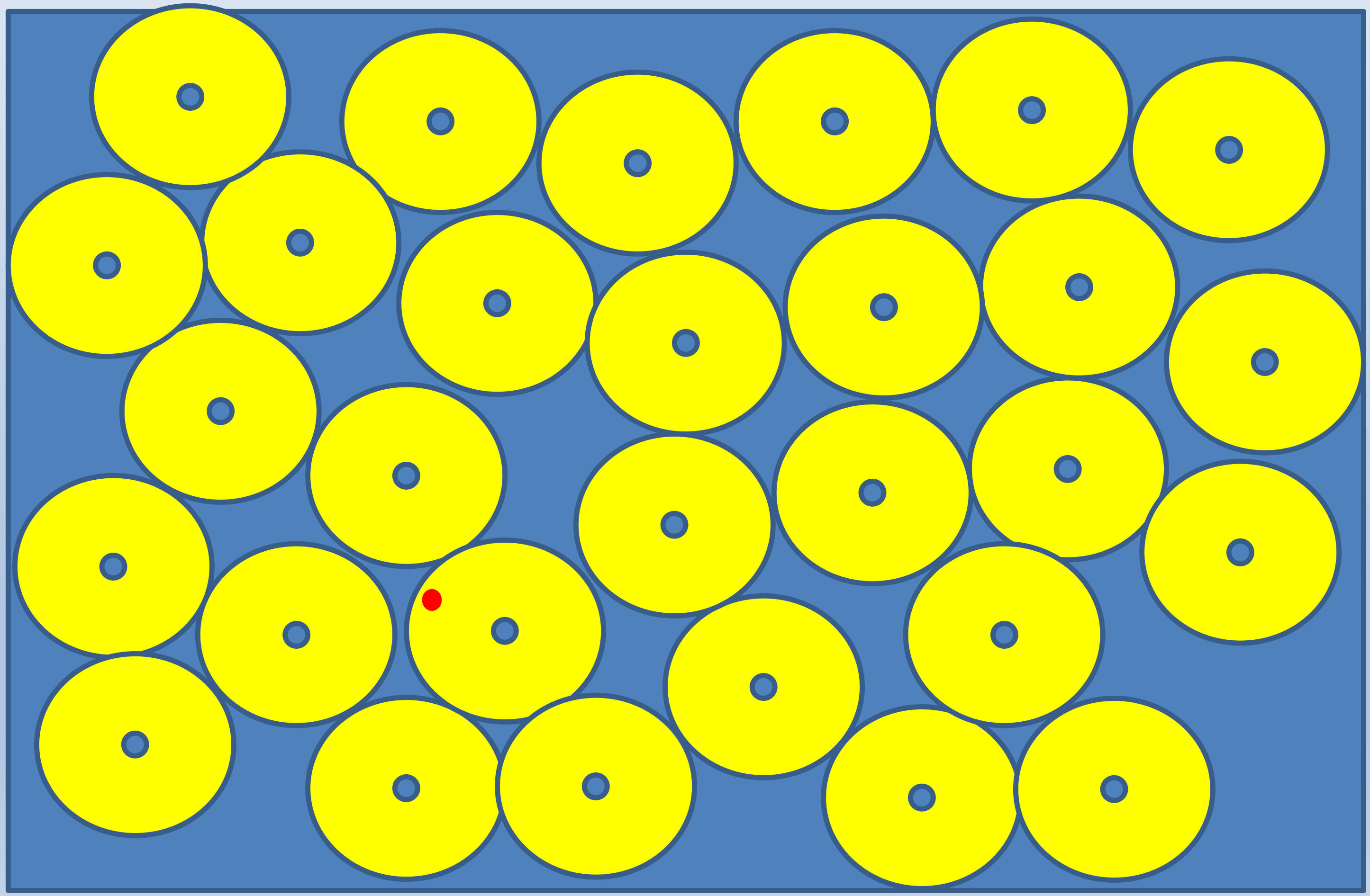
- Select randomly $x_1, x_2, \dots, x_{k-1}$. Let x_0 be a secret key. Construct a polynomial
$$P(z) = x_{k-1}z^{k-1} + x_{k-2}z^{k-2} + x_1 z + x_0$$
- Give $(\alpha_i, P(\alpha_i))$ to user i

- Large coalition has enough points to interpolate the polynomial
- Small coalition has no information about the polynomial

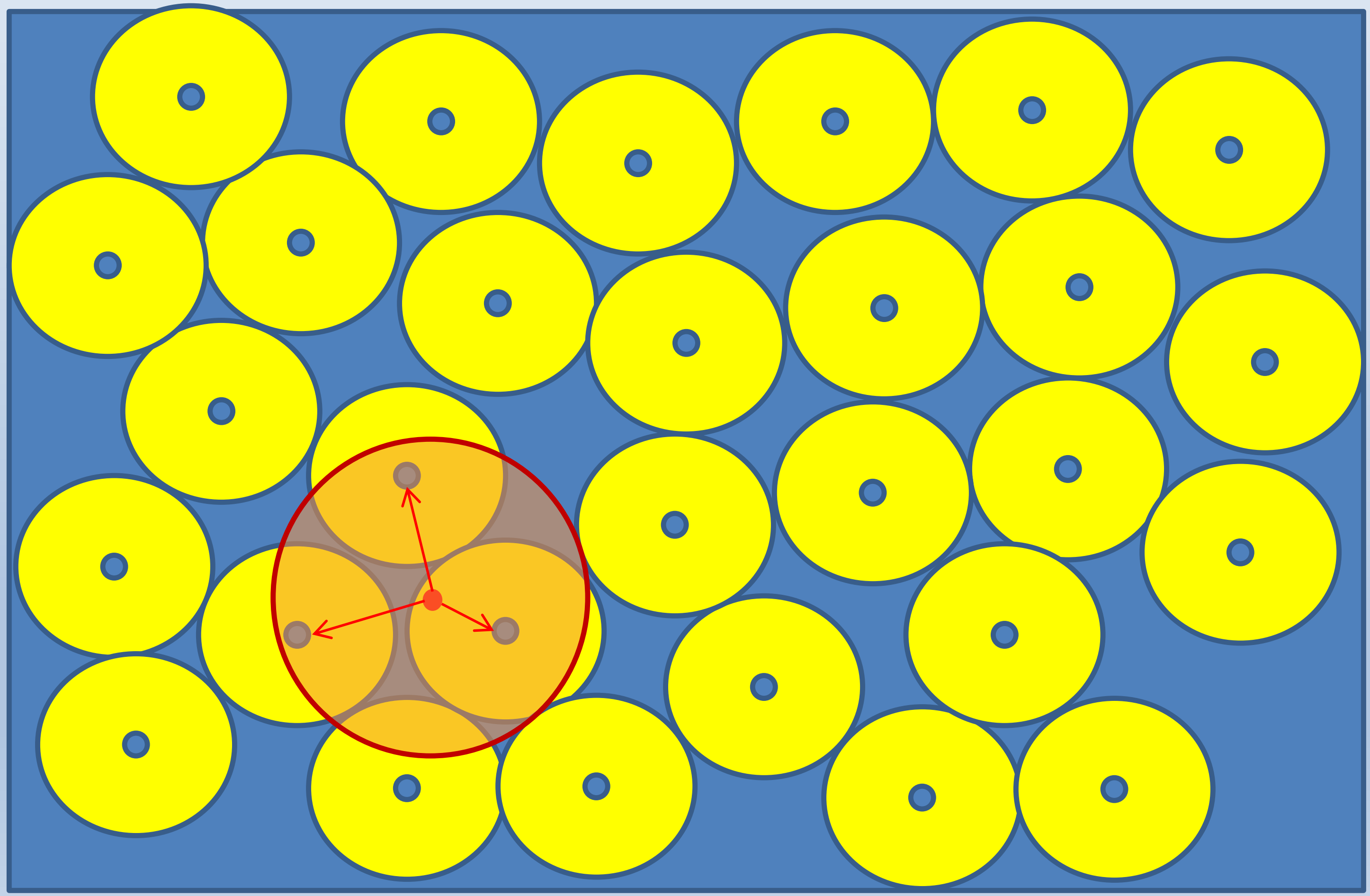
List-decoding of Reed-Solomon Codes



List-decoding of Reed-Solomon Codes

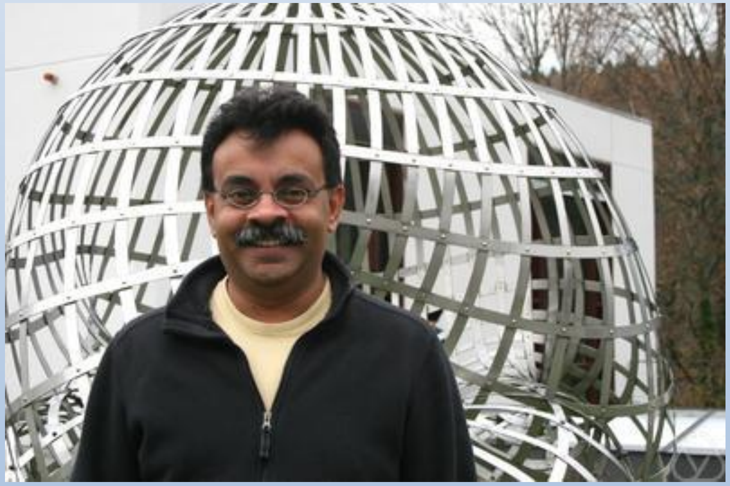


List-decoding of Reed-Solomon Codes



List-decoding of Reed-Solomon Codes

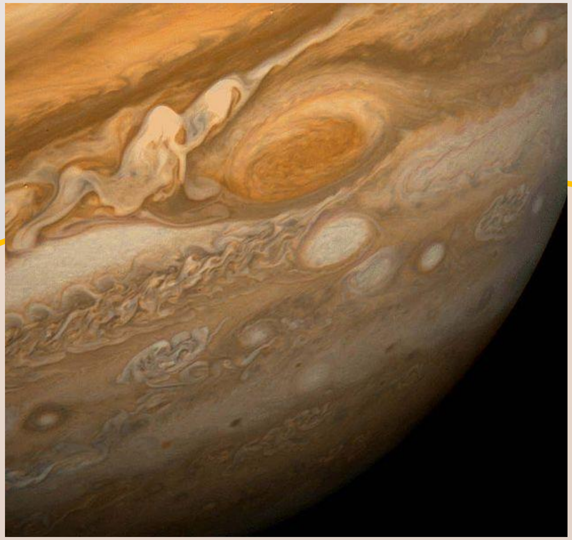
- Sudan ‘97, Guruswami ‘99, Vardy-Parvaresh ‘05, Guruswami-Rudra ‘06

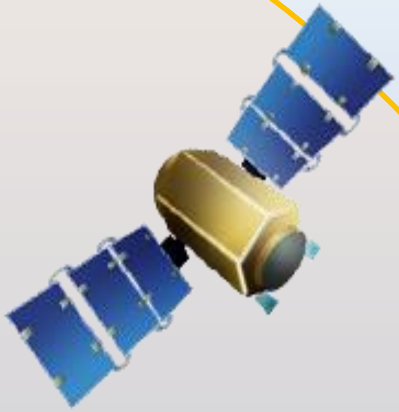


Madhu Sudan

Venkatesan Guruswami

List Decoding of RS Codes



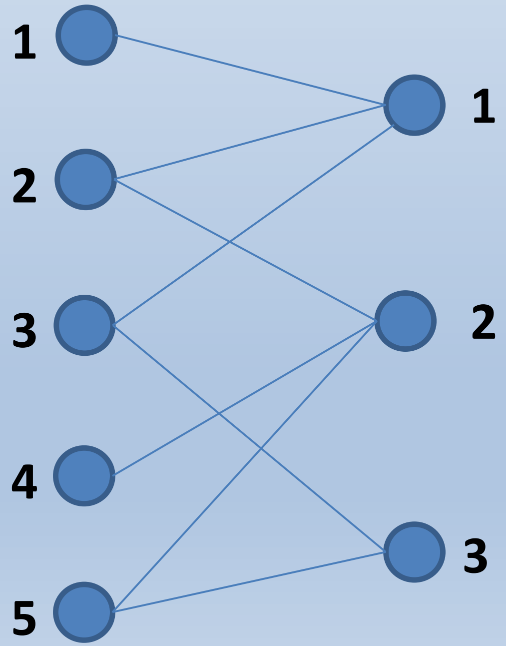


Voyager 1 – the first manmade object to leave the Solar System. Launched in 1977.

Low-Density Parity-Check Codes

- Gallager ’62
- Urbanke, Richardson and Shokrollahi ’01
- Parity-check matrix *H* is sparse
- Performance close to channel capacity
- Decoding complexity linear in *n*

Tanner graph:

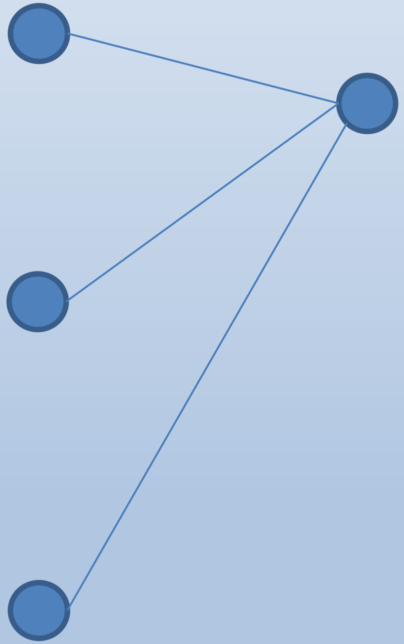


$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

Low-Density Parity-Check Codes

- Belief-propagation decoding algorithm (message-passing algorithm)

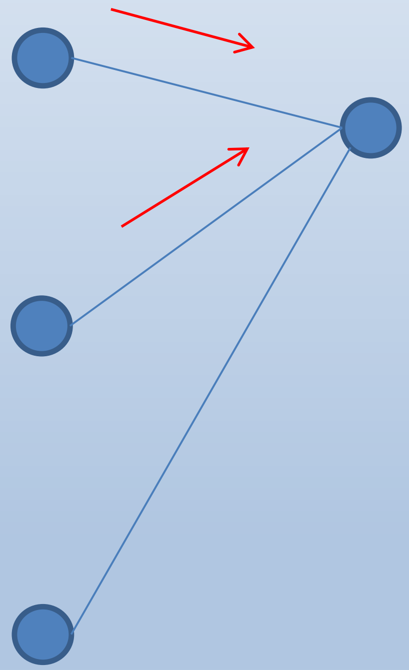
(Pr(0),Pr(1))



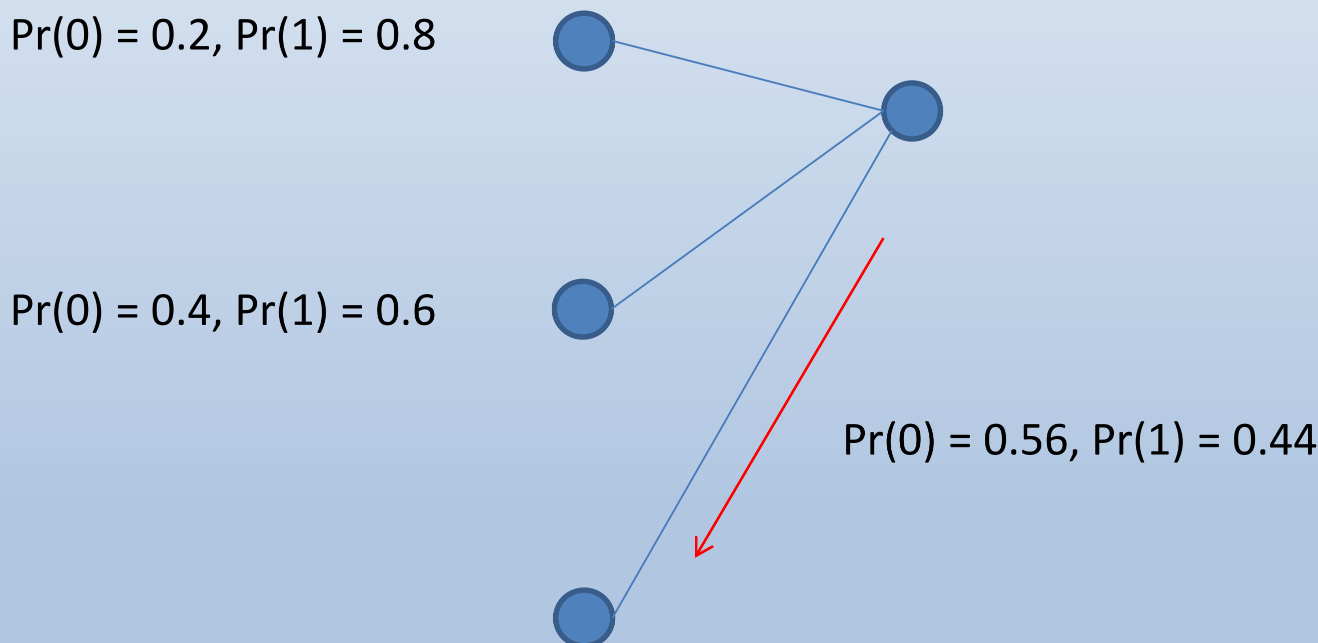
Belief-Propagation Algorithm

Pr(0) = 0.2, Pr(1) = 0.8

Pr(0) = 0.4, Pr(1) = 0.6



Low-Density Parity-Check Codes



Reed-Solomon Codes are Used in:

- Wired and wireless communications

- Satellite communications

- Hard drives and compact disks

- Flash memory devices

LDPC Codes are Used in:

- Wired and wireless communications

- Satellite communications

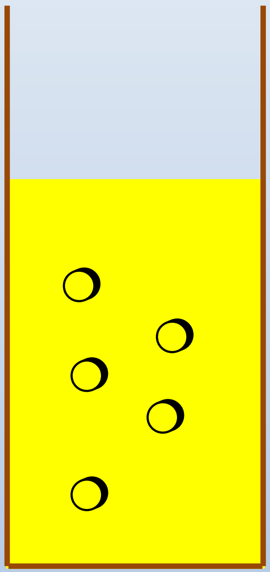
- Hard drives and compact disks

- Flash memory devices

Emerging Applications of Coding Theory

Flash memories

- Easy to add electric charge, hard to remove
- The charge “leaks” with the time
- Neighboring cells influence each other



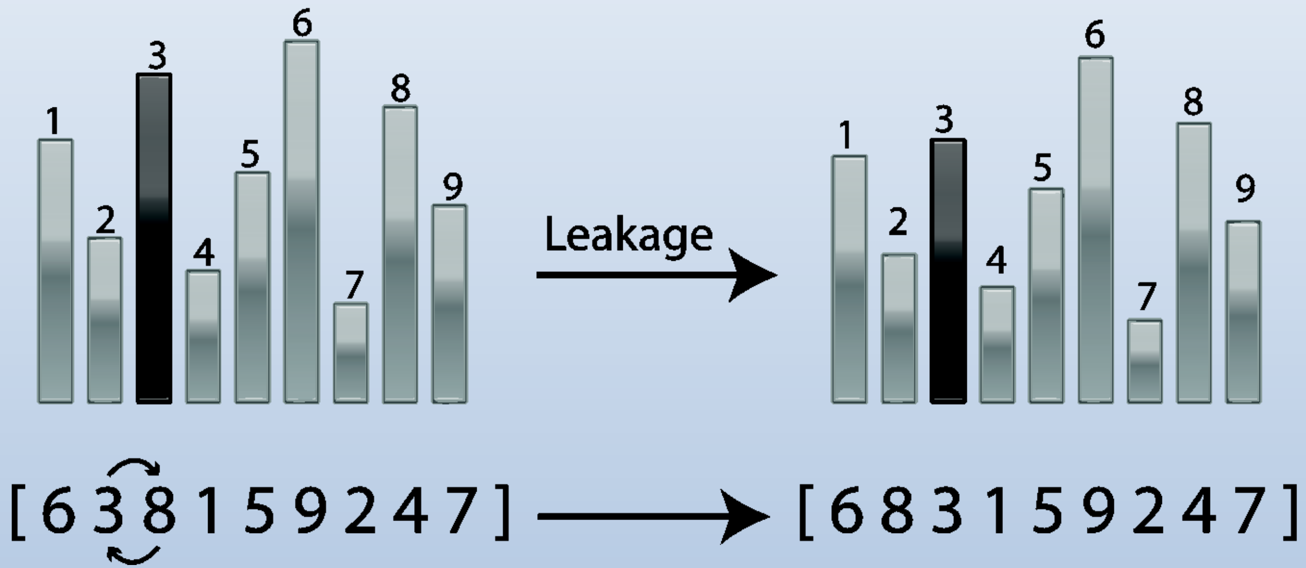
Flash memory cell

Flash memories

- Rank modulation
- The information is represented using relative levels of charge, invariant to leakage
- Coding over permutations

Jiang, Mateescu, Schwartz, Bruck ‘2006

Flash memories



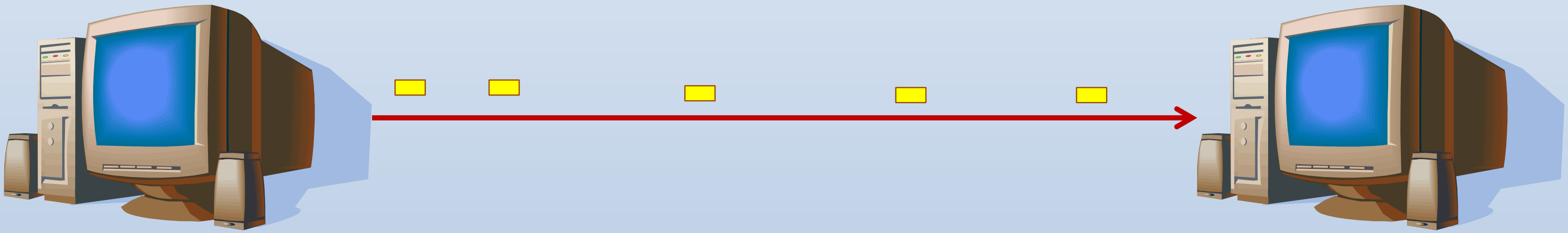
Networking

- Rateless Codes

- A. Shokrollahi '2004
- Used in DVB-H standard for IP datacast for handheld devices

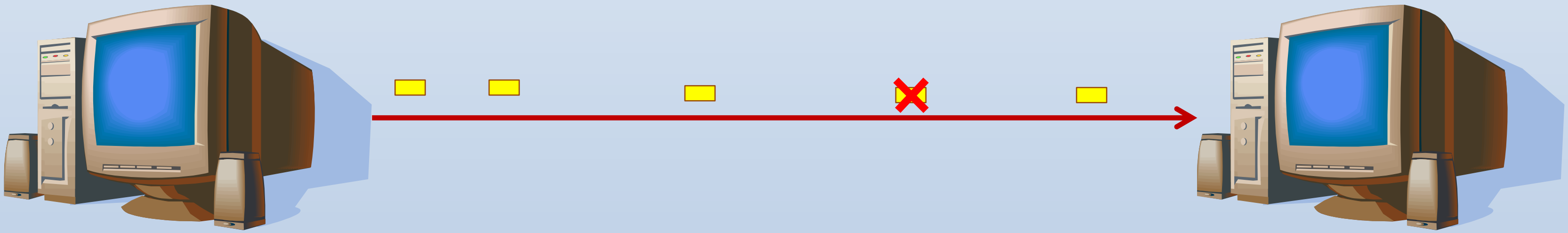
Networking

- Rateless Codes



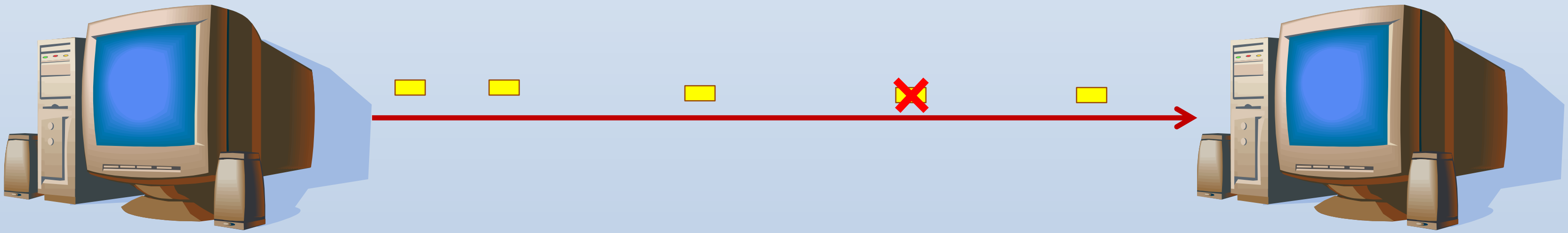
Networking

- Rateless Codes



Networking

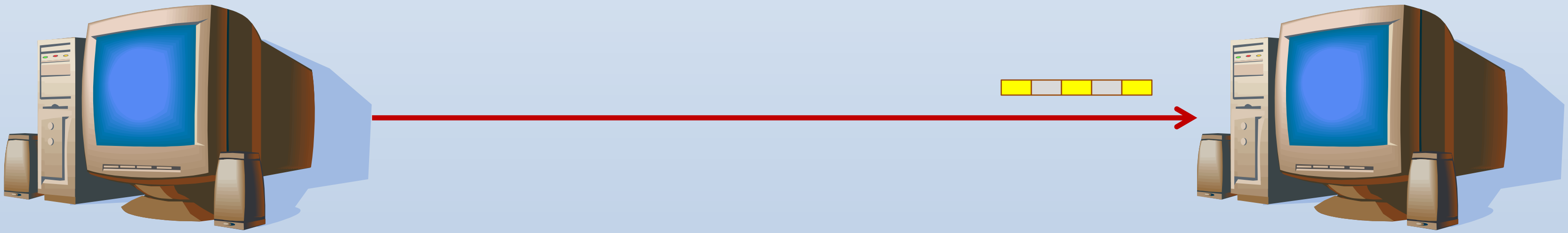
- Rateless Codes



- Possible solution: ARQs (retransmissions) – slow!
- Alternative: large error-correcting code

Networking

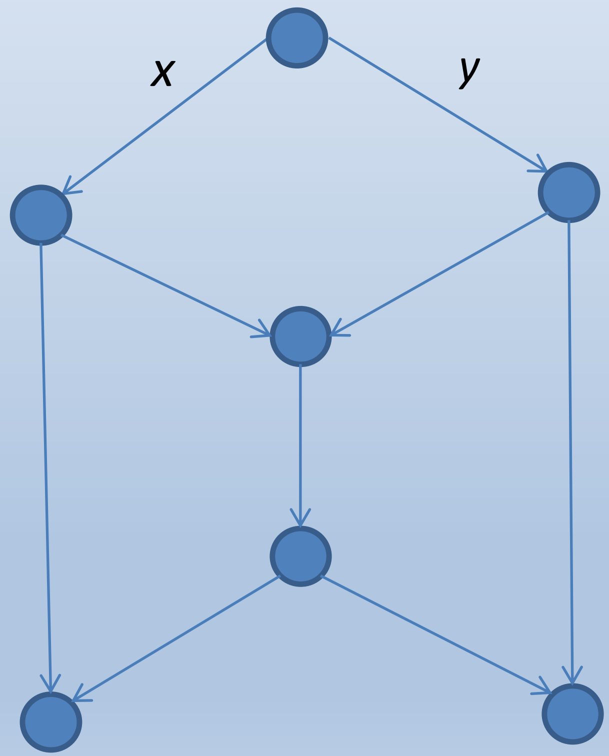
- Rateless Codes



Network coding

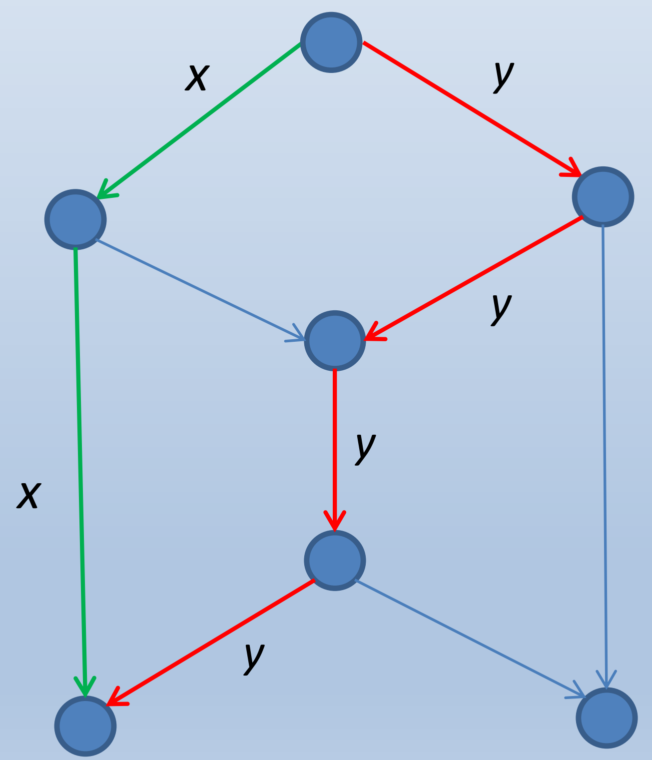
- Butterfly network

Ahlswede, Cai, Li and Yeung, 2000



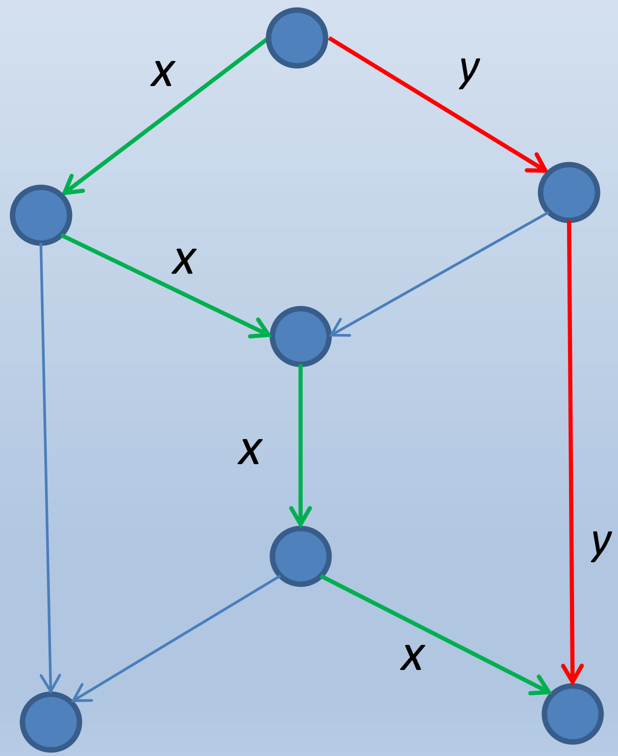
Network coding

- Butterfly network



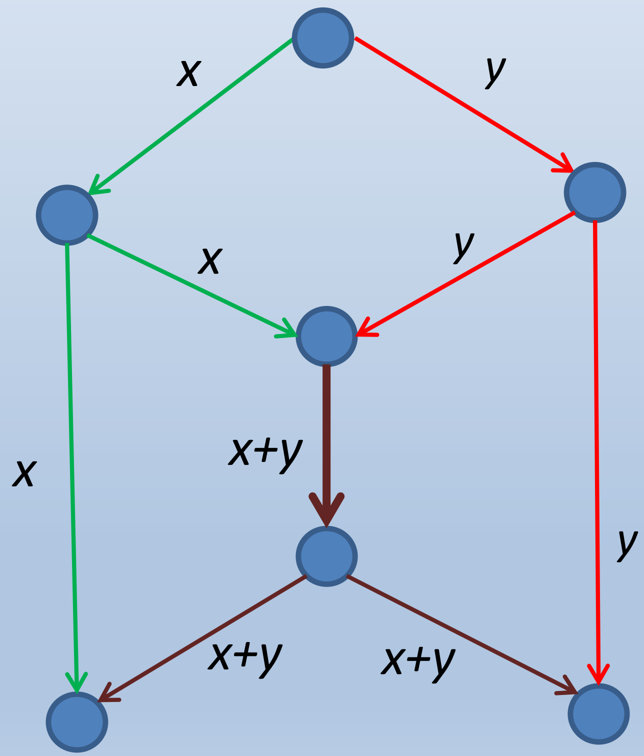
Network coding

- Butterfly network



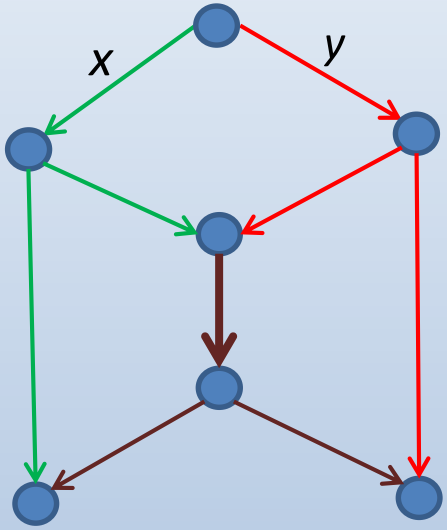
Network coding

- Butterfly network



Network coding

- The number of bits deliverable to each destination is equal to min-cut between source and each of destinations
- Avalanche P2P Network (Microsoft, 2005)
- Experiments for use in mobile communications



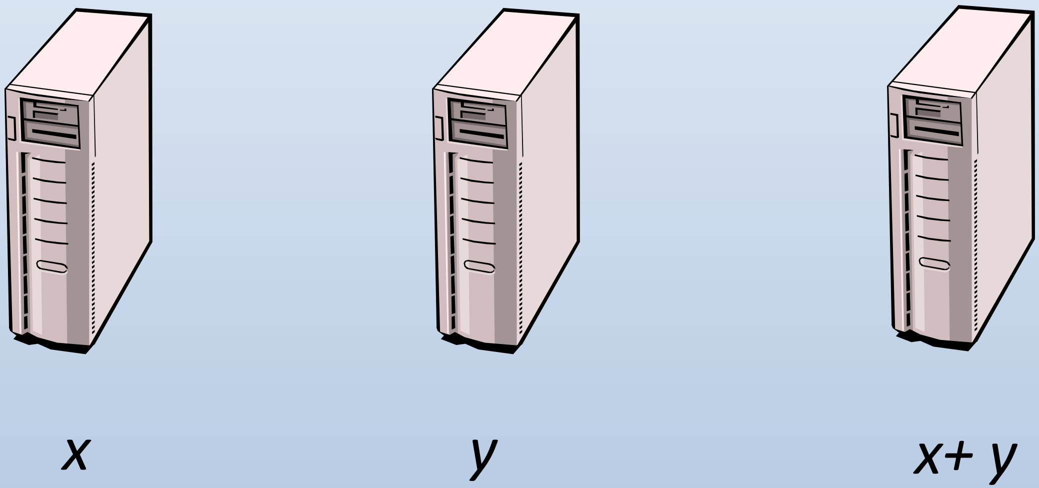
Distributed Storage

- Huge amounts of data stored by big data companies (Google, Amazon, Facebook, Dropbox)

Facebook data center in Oregon

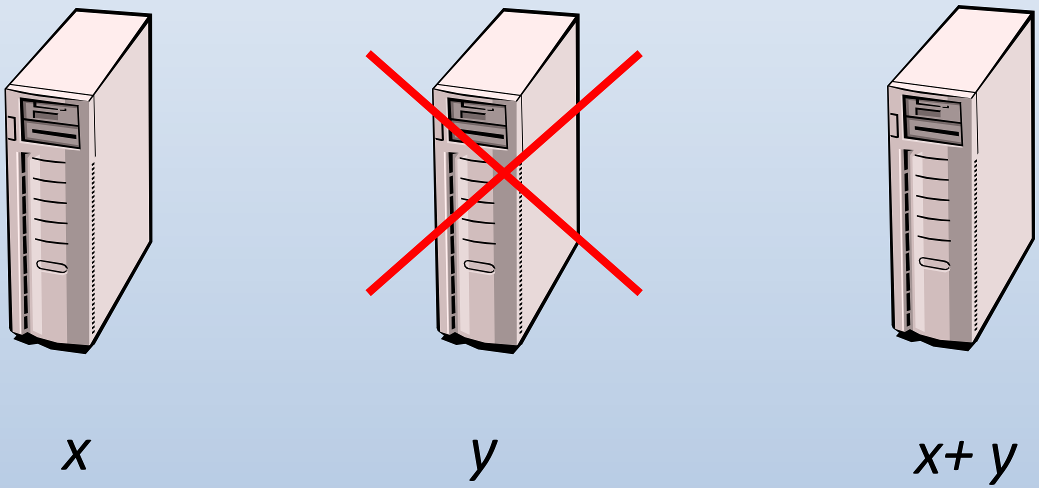
Server room at
Wikipedia data center

Distributed data storage

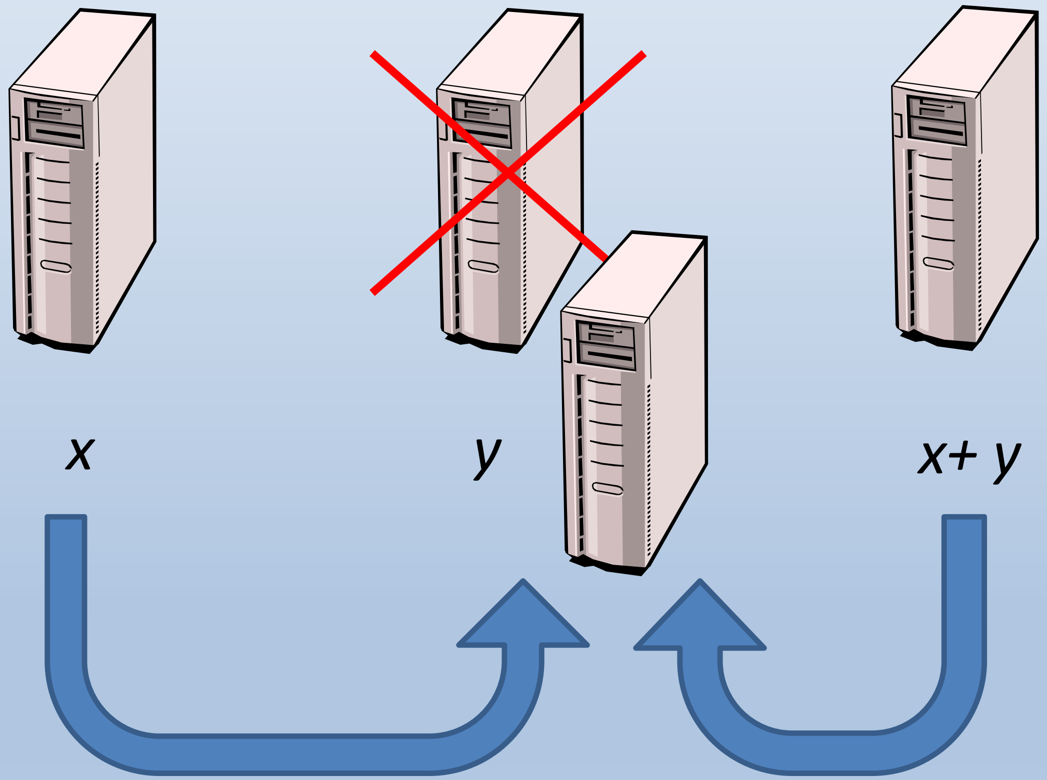


Dimakis, Godfrey, Wu, Wainwright, Ramchandran ‘2008

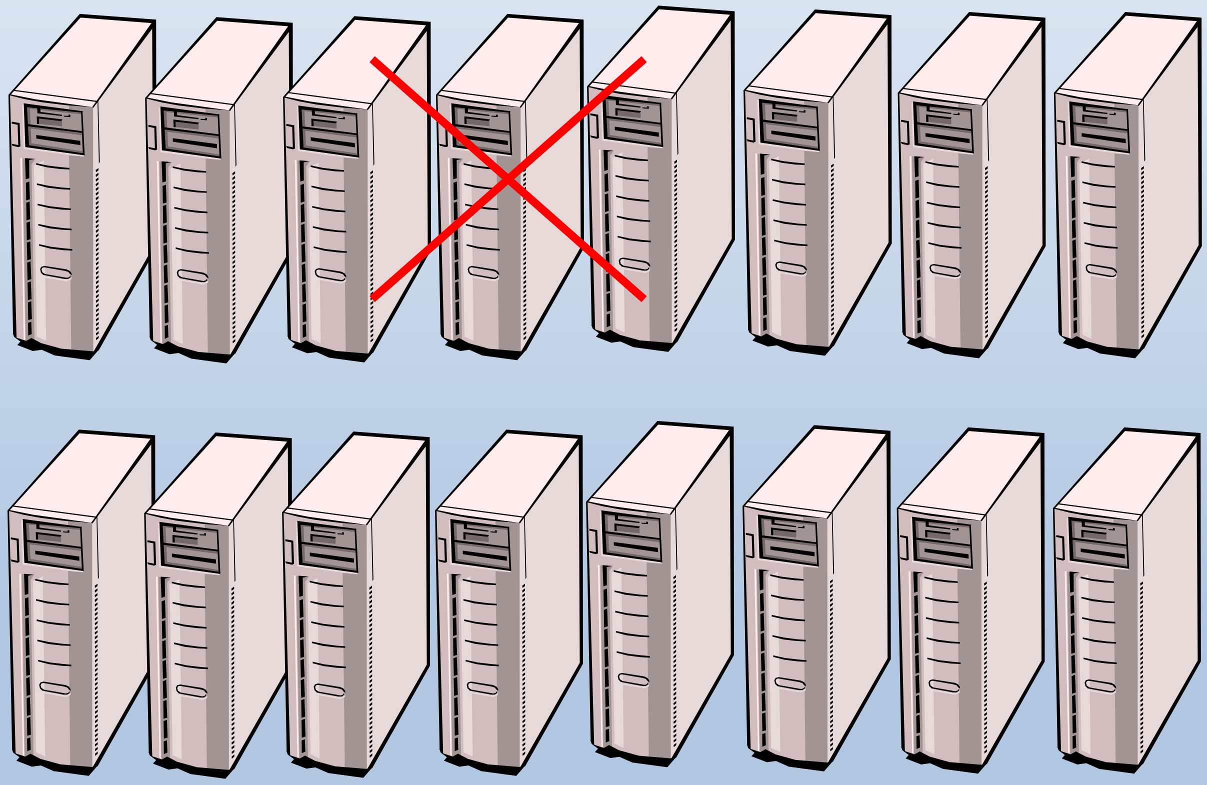
Distributed data storage



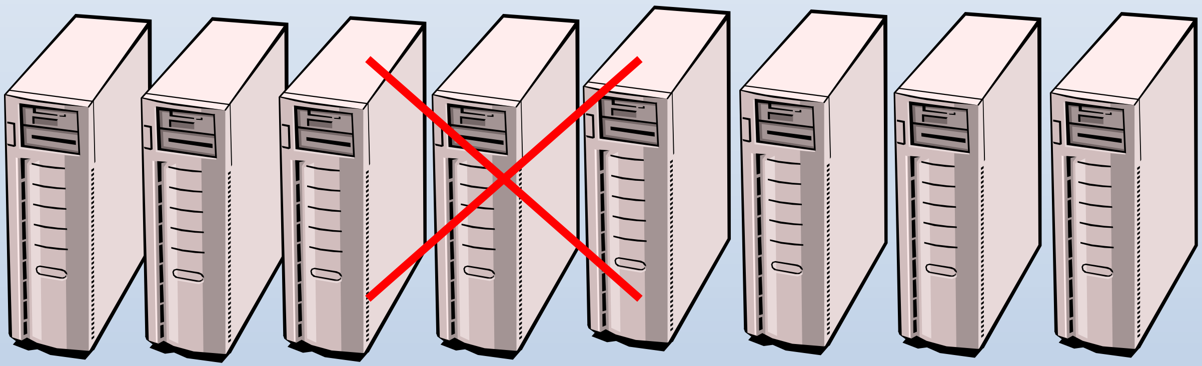
Distributed data storage



Distributed data storage

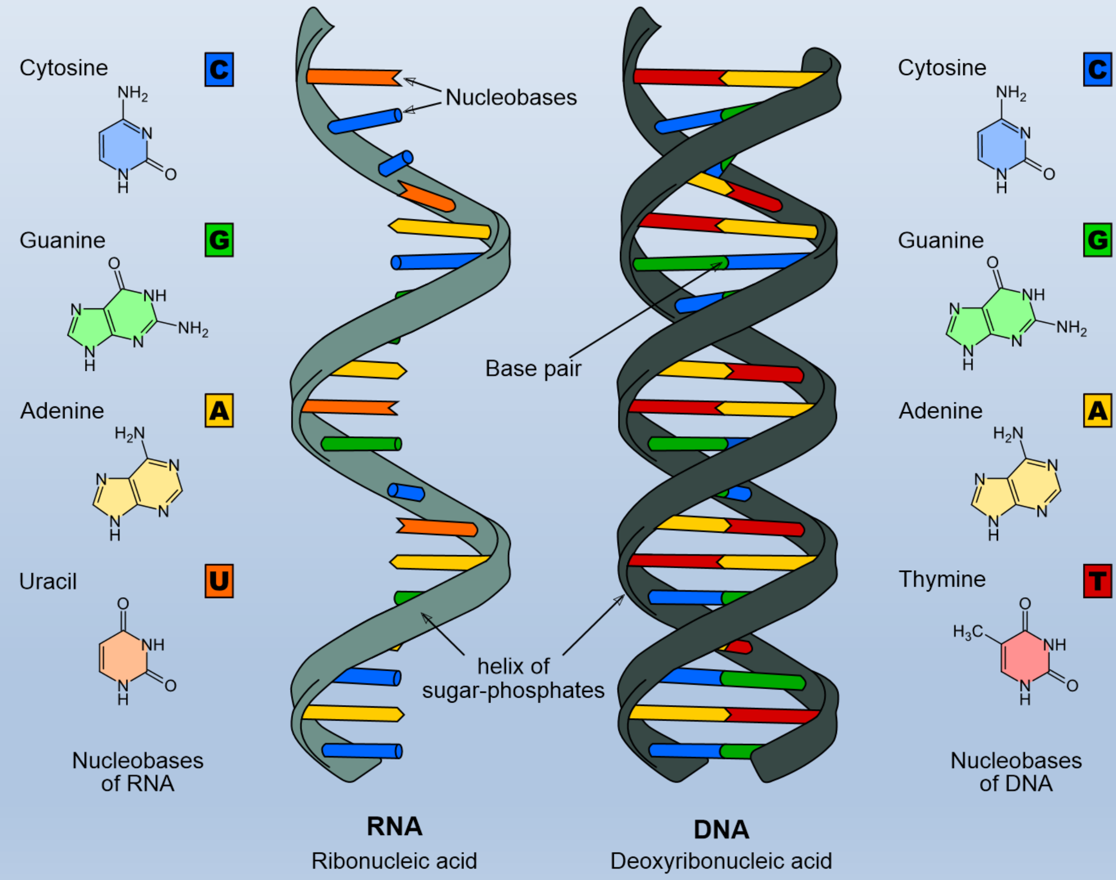


Distributed data storage



- Classical error-correcting codes can be employed
- Local correction is needed (using few other servers) to facilitate the correction

DNA-based Data Storage



Error Correction in DNA Storage

- Four amino acids: A, T, G, C
- Mechanisms for error correction are required

March 2018: University of Washington and Microsoft demonstrated storage and retrieval of 200MB of data.

Thanks!