

# ***Group Testing – A Classical Search Method***

簡介群試理論

Hung-Lin Fu 傅 恒 霖

交通大學應用數學系

# Warming Up

- 猜數字遊戲：我有一個介於1和63的整數，你們要如何用最少個問題來猜出它？（如何問？）



# At most 6 Queries

- (1) Is the number larger than 32?  
Yes or No?
- (2) If the answer is “Yes”, then ask if the number is larger than 48 or not. On the other hand, if the answer is “No”, then ask if the number is larger than 16 or not.
- Continued ...
- 6 (Binary Splitting Algorithm)
- Do you have a better idea?

# Another idea

- We can also ask several questions simultaneously and then make a decision!



# Magic cards

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 |
| 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 |
| 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 |

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 |
| 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 |

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 2  | 3  | 6  | 7  | 10 | 11 | 14 | 15 |
| 18 | 19 | 22 | 23 | 26 | 27 | 30 | 31 |
| 34 | 35 | 38 | 39 | 42 | 43 | 46 | 47 |
| 50 | 51 | 54 | 55 | 58 | 59 | 62 | 63 |

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 |
| 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 |
| 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 |

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 4  | 5  | 6  | 7  | 12 | 13 | 14 | 15 |
| 20 | 21 | 22 | 23 | 28 | 29 | 30 | 31 |
| 36 | 37 | 38 | 39 | 44 | 45 | 46 | 47 |
| 52 | 53 | 54 | 55 | 60 | 61 | 62 | 63 |

|    |    |    |    |    |    |    |    |
|----|----|----|----|----|----|----|----|
| 1  | 3  | 5  | 7  | 9  | 11 | 13 | 15 |
| 17 | 19 | 21 | 23 | 25 | 27 | 29 | 31 |
| 33 | 35 | 37 | 39 | 41 | 43 | 45 | 47 |
| 49 | 51 | 53 | 55 | 57 | 59 | 61 | 63 |

Guess ?!

|   |   |   |   |
|---|---|---|---|
|   |   |   |   |
| 4 | 5 | 6 | 7 |

|   |   |   |   |
|---|---|---|---|
|   |   |   |   |
| 2 | 3 | 6 | 7 |

|   |   |   |   |
|---|---|---|---|
|   |   |   |   |
| 1 | 3 | 5 | 7 |

# A coincidence

- It is about Christmas trees.
- How to check “the light bulbs are good”?



# Information to Know

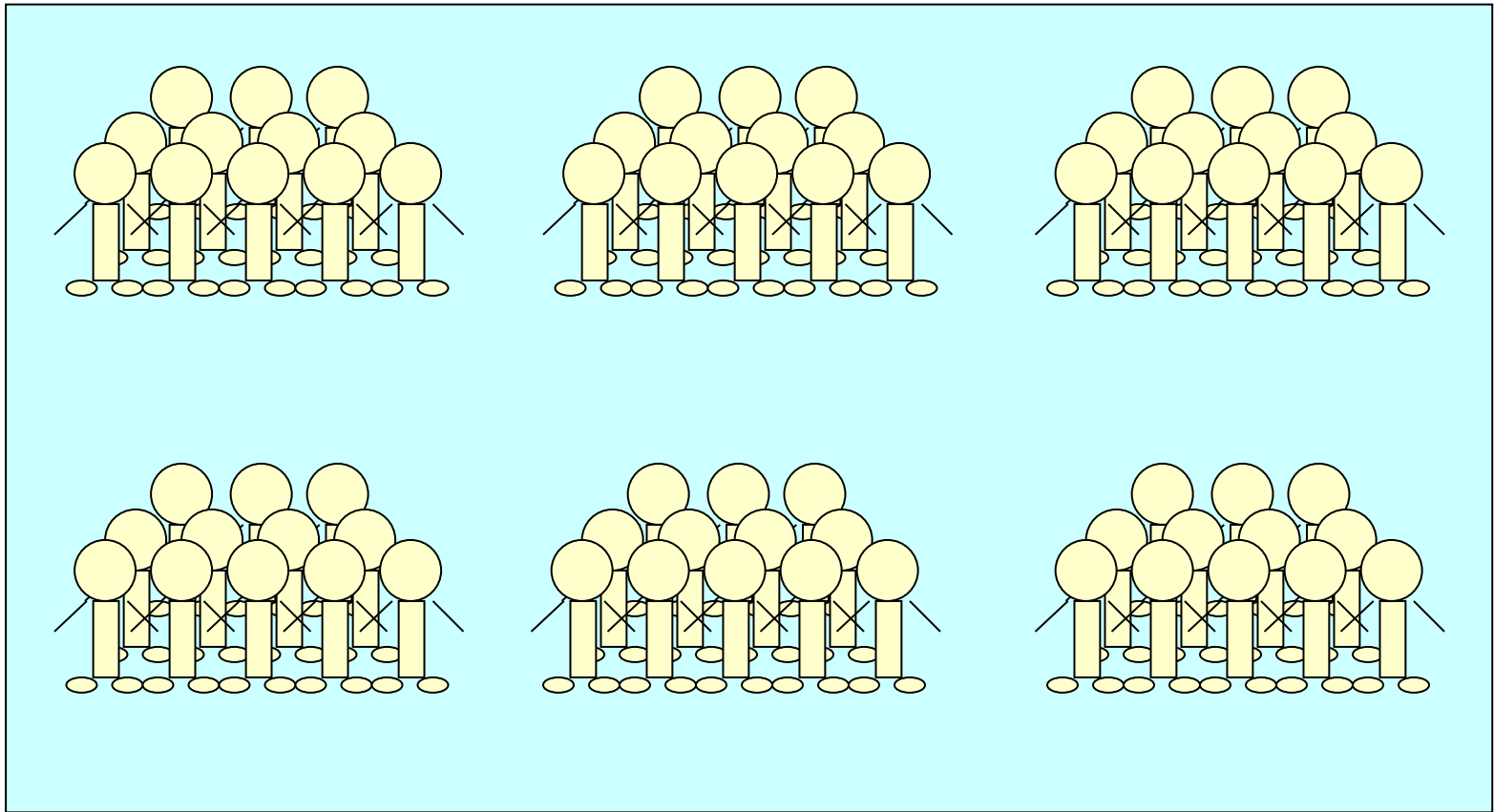
- Hsin Chu played the most important role in the production of Christmas light bulbs. (Before the construction of Science Based park, around 1950 - 1960).
- People lived along 光復路 make the key contribution!
- They knew how to apply **Group Testing** already.
- Currently, more than 80% such light bulbs are manufactured by a **Taiwanese company** located in mainland China.

# Fake Coins



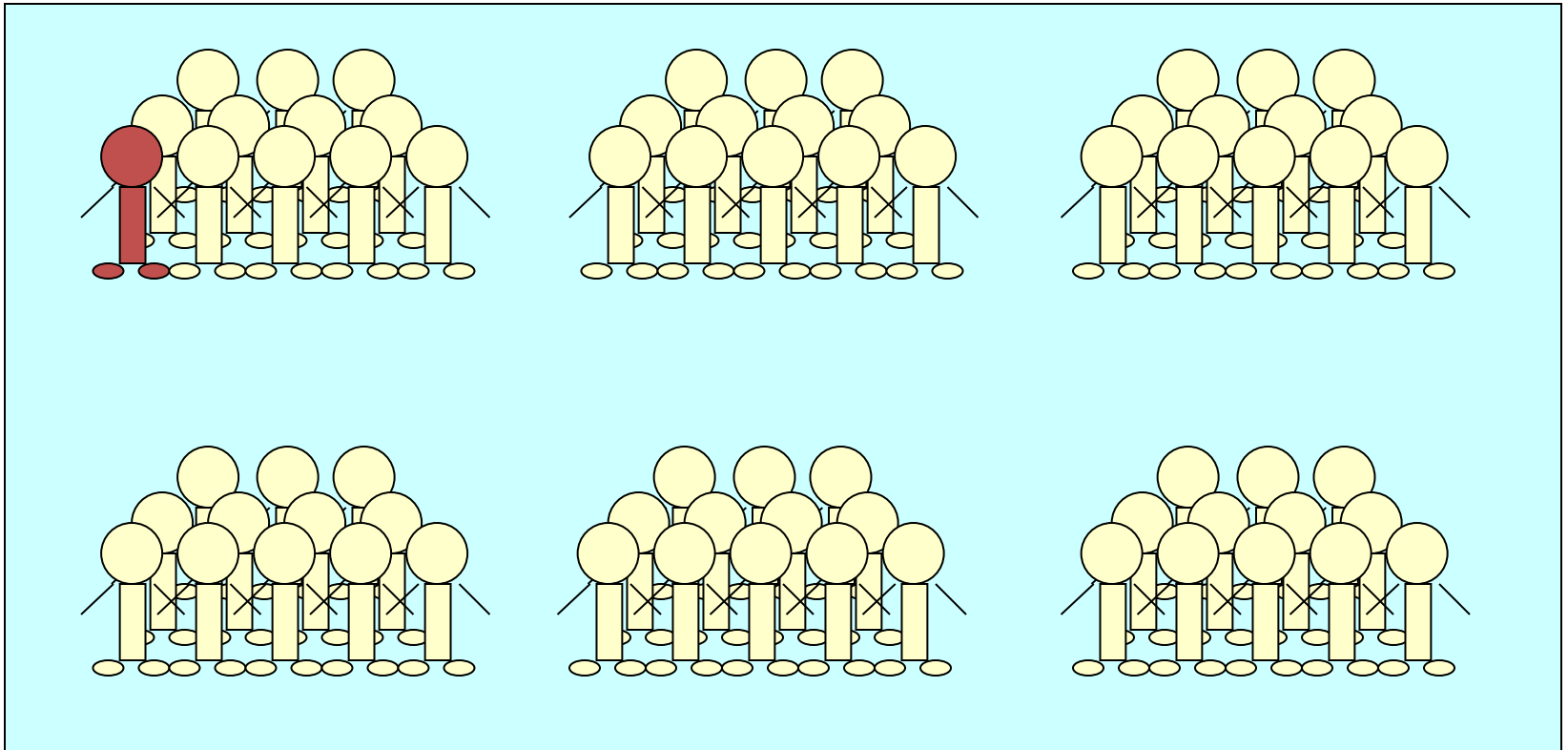
# One More Example

*Syphilis (梅毒) Tests (World war II, 1943)*

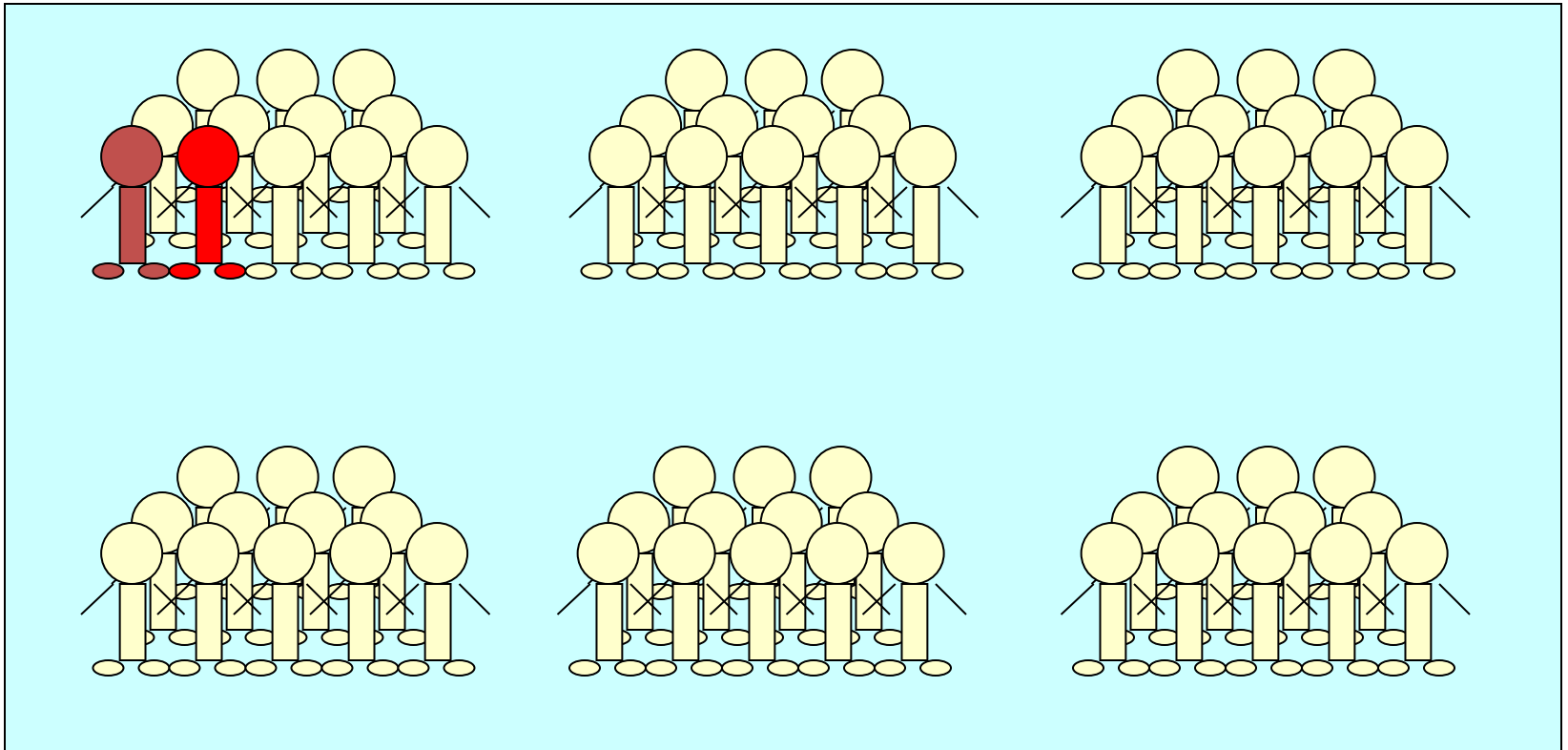


# Individual Test

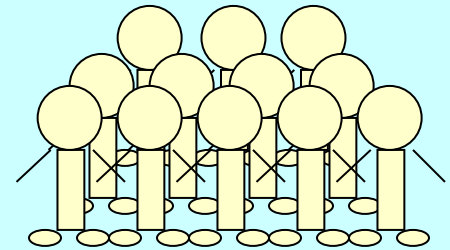
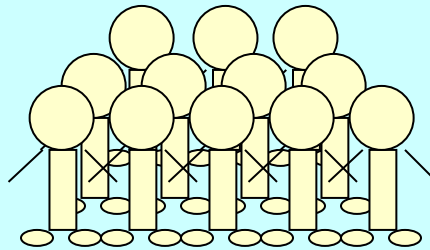
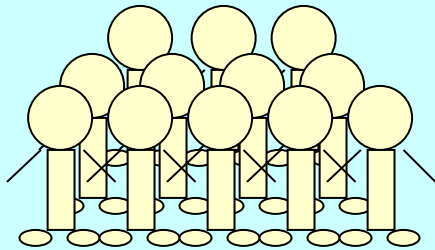
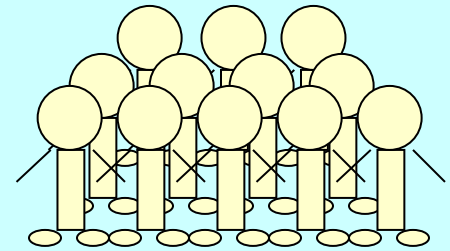
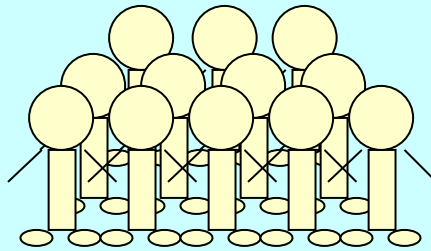
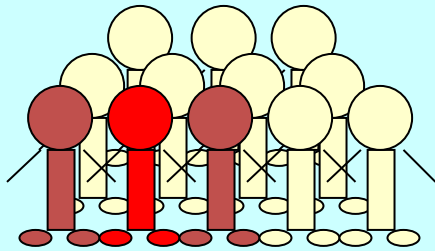
72 tests!



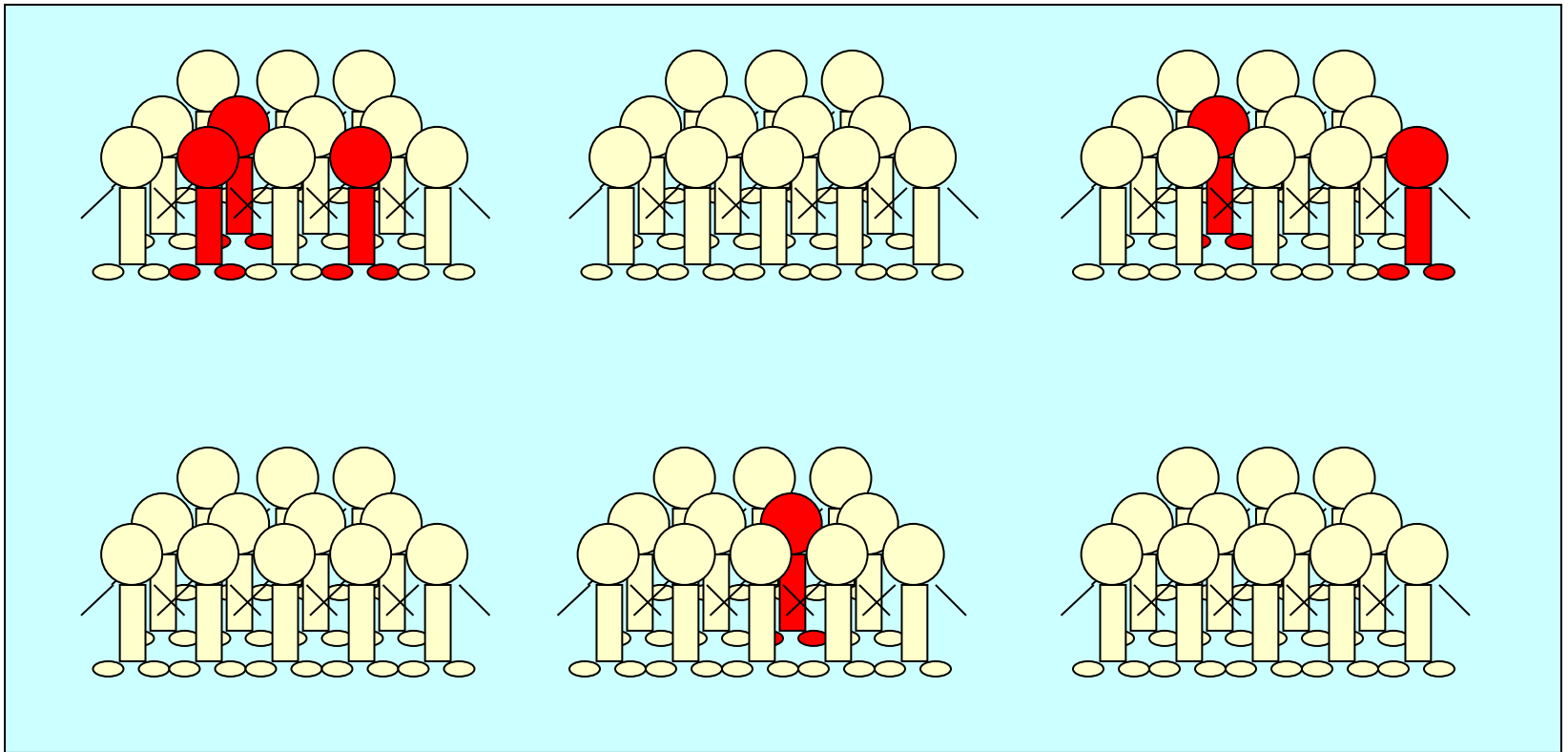
**Red** *is positive and* **Brown** *is negative*



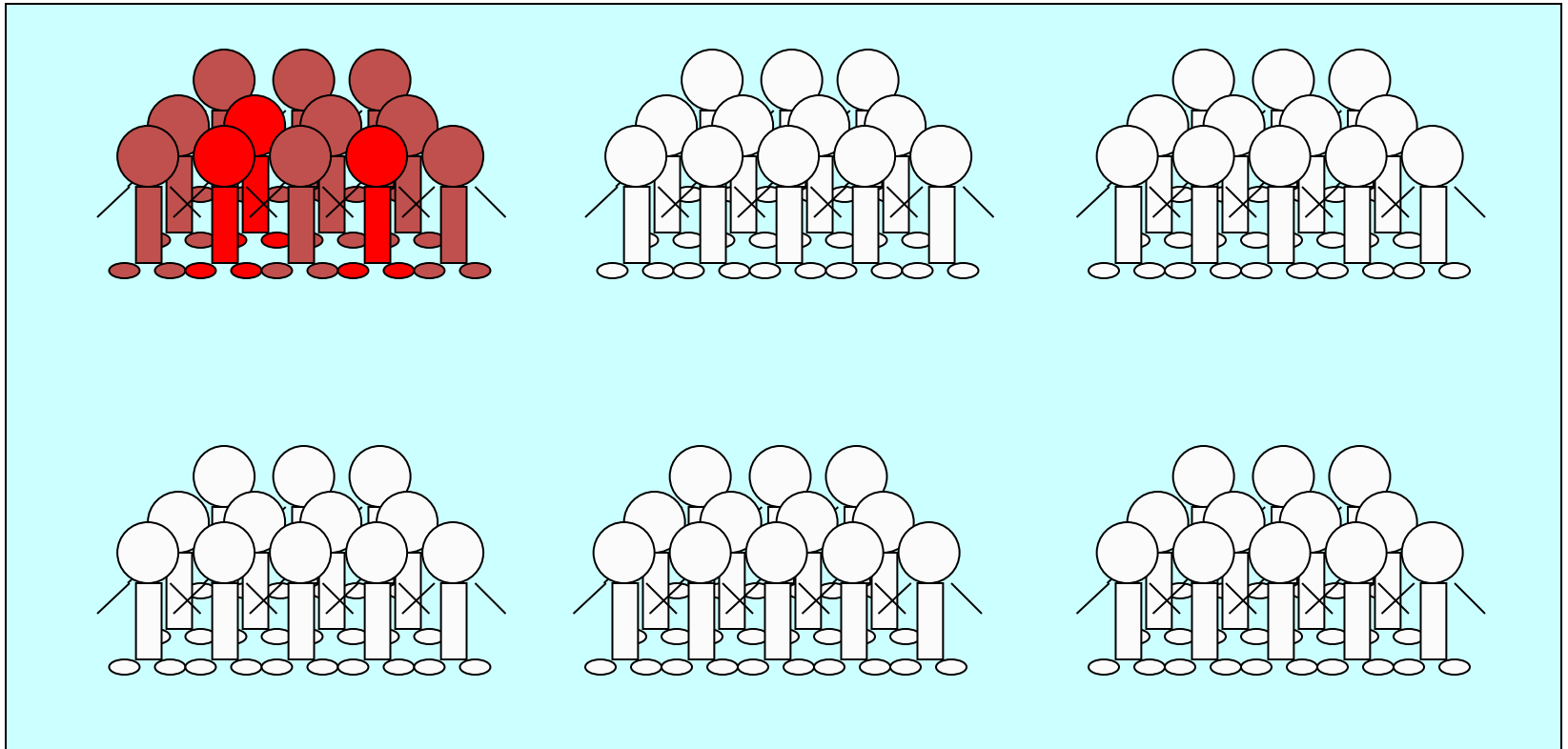
*Keep going*



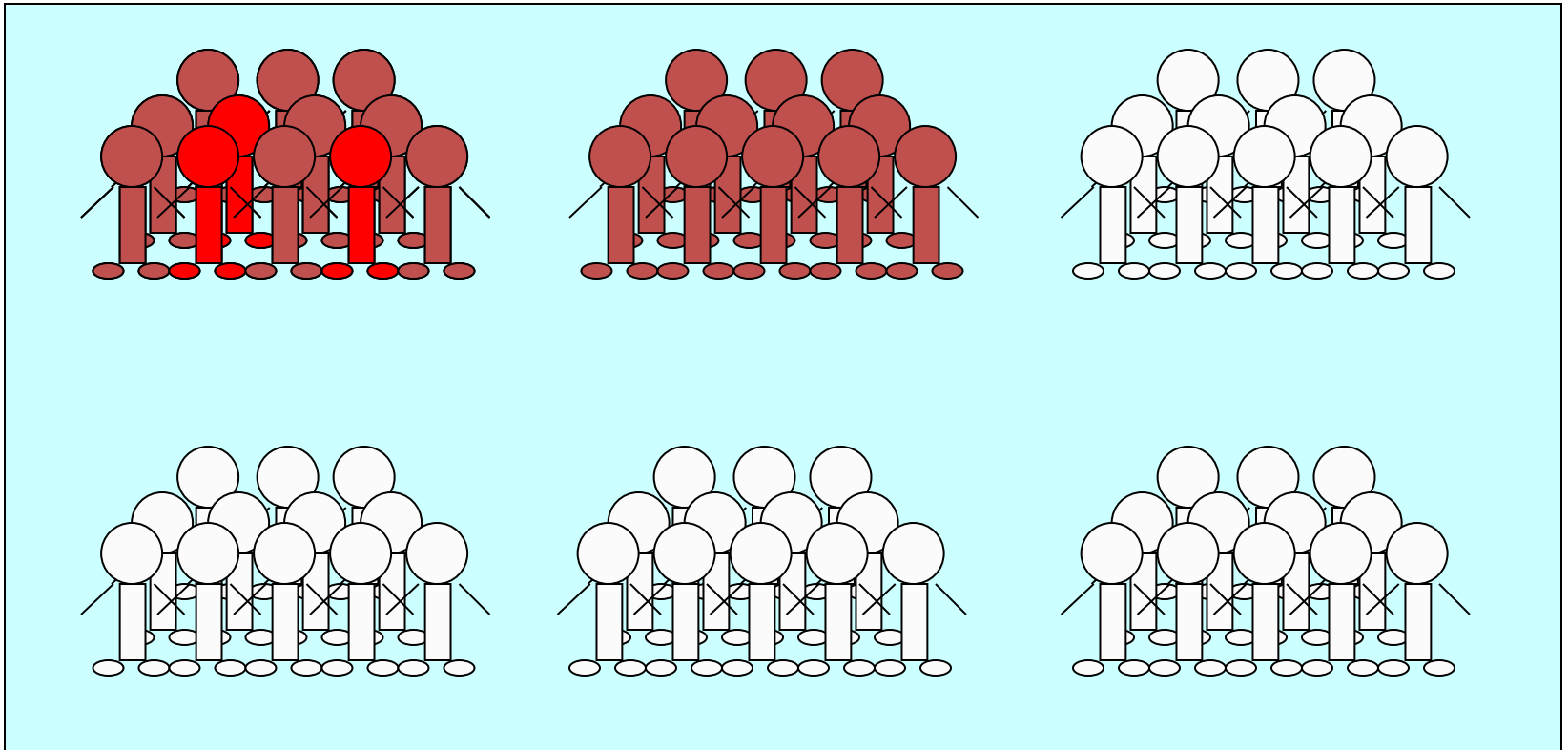
# *More positives (72 tests)*



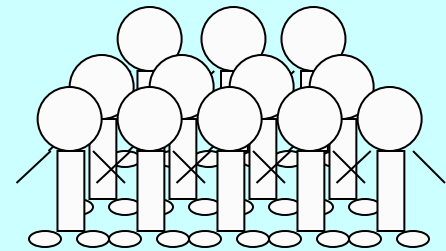
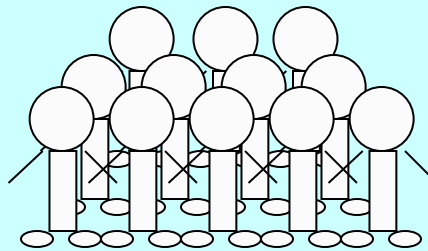
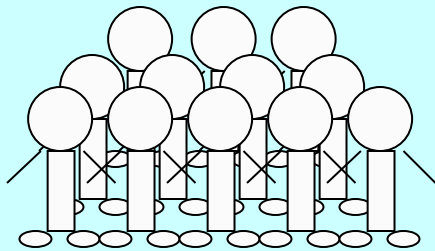
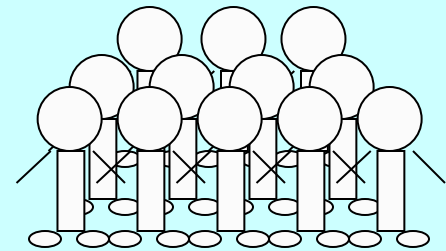
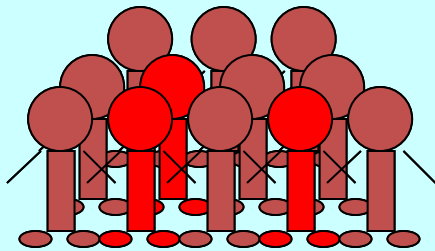
# ***Test a group at one time***



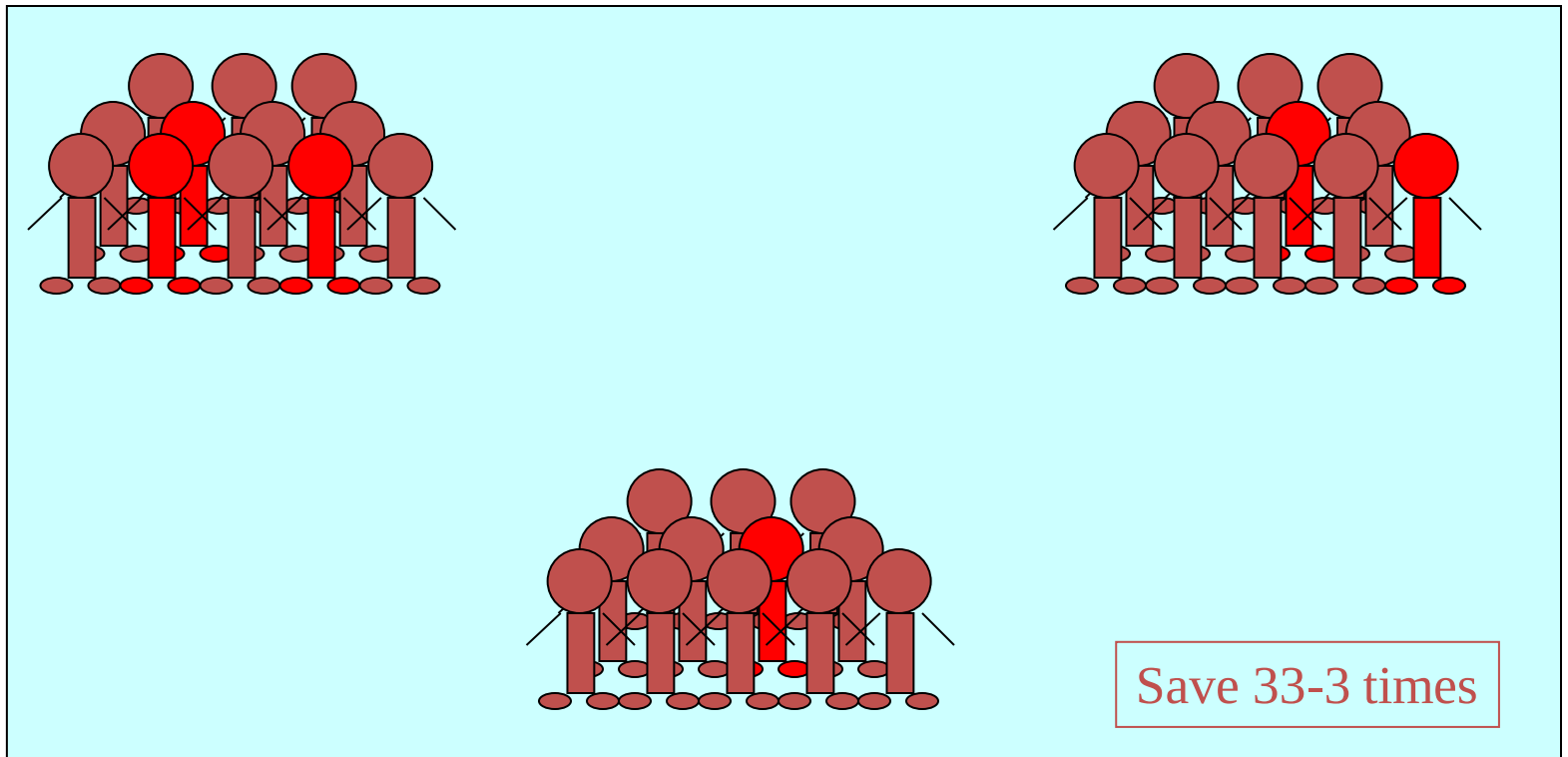
***Some group is negative!***



***They are done!***



# ***Save a lot of Money!***



*What's next?*

# Research Facts

- The idea was proposed by a team member (Rosenblatt) of R. Dorfman and Dorfman went ahead to write a journal paper to emphasize the importance of this idea.
- The detection of defective members of large populations, Ann. Math. Statist. 14 (1943), 436 – 440.
- Upon the report, this idea has never been used then. ***People are not confident about it!***

# Group testing

- Consider a set  $N$  of  $n$  items consisting of at most  $d$  *positive* (used to be called defective) items with the others being *negative* (used to be called good) items.
- A group test, sometimes called a *pool*, can be applied to an arbitrary set  $S$  of items with two possible outcomes; *negative*: all items in  $S$  are negative; *positive*: at least one positive item in  $S$ , not knowing which one or how many.

# Adaptive (Sequential) and Non-adaptive

- Can you see the *difference* between the above two ways of finding the number in  $[1,63]$ ?
- The first one (*adaptive or sequential*): You ask the second question (query) after knowing the answer of the first one and continue .... That is, the previous knowledge will be used later.
- The second one (*non-adaptive*): You can ask all the questions (queries) *at the same time*.

# Algorithms

- Adaptive algorithm (*Count on you!*)
- Non-adaptive algorithm (*Construction of Arrays?*)
- k-stage algorithm

The most popular one is a 2-stage algorithm in which we use an algorithm (either one) in the first stage and then in the second stage we test the left “suspected” items from the first stage one by one.

# Fundamental Problem

- Among all possible algorithms let  $M(d, n)$  denote the minimum number of tests we need to use in the worst case.  $M(1, 64) = 6$ .
- **Problem: Determine  $M(d, n)$ .**
- For example, can we determine  $M(2, 1,000)$ ?
- **Fact:**  $d \cdot \{\log_2 n\} \geq M(d, n) \geq \{\log_2 C(n, d)\}$ .  
( $C(n, d)$  denotes  $n$ -choose- $d$ .)
- The inequality of the right hand side is known as the information lower bound. (?)

# Development of GT (1)

- - 1999 Complex model: Molecular biology (Torney)
- - 1998 Mutually obscuring defetives (Damaschke)
- - 1997 Inhibitor model (Farach *et al.*)
- - 1993 Combinatorial group testing (堵丁柱與黃光明)
- - 1984 Multiple-access communication (Berger et al.)
- - 1943 Blood testing (Dorfman & Rosenblatt)

# Development of GT (2)

- 2006 – Present: Our contributions- See references.
- 2006 Pooling designs and nonadaptive group testing (堵丁柱與黃光明).
- Graph reconstruction (contig sequencing)
- Non-unique probe selection problem
- - 2006 Inhibitor complex model (Chang *et al.*)
- - 2001 Drug discovery (Xie *et al.*)
- - 2000 Combinatorial group testing, 2ed (堵丁柱與黃光明)

# Non-adaptive Algorithm

We can use a matrix to describe a non-adaptive algorithm.

*Items* are indexed by columns and the tests are indexed by rows. Therefore, the  $(i, j)$  entry is **1** if the item  $j$  is included in the *pool*  $i$  (for test), and **0** otherwise.

## *An example (12 items and 9 tests)*

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 |   |   | 1 |   |   | 1 |   |   | 1 |   |   |
| 1 |   |   |   | 1 |   |   | 1 |   |   | 1 |   |
| 1 |   |   |   |   | 1 |   |   | 1 |   |   | 1 |
|   | 1 |   | 1 |   |   |   |   | 1 |   | 1 |   |
|   | 1 |   |   | 1 |   | 1 |   |   |   |   | 1 |
|   | 1 |   |   |   | 1 |   | 1 |   | 1 |   |   |
|   |   | 1 | 1 |   |   |   | 1 |   |   |   | 1 |
|   |   | 1 |   | 1 |   |   |   | 1 | 1 |   |   |
|   |   | 1 |   |   | 1 | 1 |   |   |   | 1 |   |

The blanks  
are zeros.

***Items: 010000010000 (2 & 8)***

***Outcome vector: 010111100***

|   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 |   |   | 1 |   |   | 1 |   |   | 1 |   |   |
| 1 |   |   |   | 1 |   |   | 1 |   |   | 1 |   |
| 1 |   |   |   |   | 1 |   |   | 1 |   |   | 1 |
|   | 1 |   | 1 |   |   |   |   | 1 |   | 1 |   |
|   | 1 |   |   | 1 |   | 1 |   |   |   |   | 1 |
|   | 1 |   |   |   | 1 |   | 1 |   | 1 |   |   |
|   |   | 1 | 1 |   |   |   | 1 |   |   |   | 1 |
|   |   | 1 |   | 1 |   |   |   | 1 | 1 |   |   |
|   |   | 1 |   |   | 1 | 1 |   |   |   | 1 |   |

The blanks  
are zeros.

# Decoding Idea

- Let each input be a 12-dim column vector  $\mathbf{x}$  and  $A$  be the above  $9 \times 12$  matrix. Then the outcome  $(0,1)$ -vector  $\mathbf{y}$  is obtained by  $A\mathbf{x}$  such that its  $i^{\text{th}}$  component is 1 if the  $i^{\text{th}}$  component of  $A\mathbf{x}$  is positive and 0 otherwise.
- If the mapping  $A$  is 1-1, then we are able to find the positives, that is to determine  $\mathbf{x}$ , see it?

# Outcome Vectors

- The vector  $\mathbf{y}$  is an *outcome vector* which is corresponding to an input  $\mathbf{x}$ .
- If  $A$  is 1-1, then we can decode  $\mathbf{x}$  long as we know its outcome vector.
- In order to get the job done with lower decoding complexity, extra properties for  $A$  is needed.

# Set notation

- It is easier to apply combinatorial structures to construct pooling designs, therefore, we use sets (or codewords) for columns in non-adaptive algorithms.
- We shall use set notation most of the time.
- The set corresponding to a codeword (binary vector) is called the *support* of the codeword.

# Relation with designs

- Let  $M = [m_{i,j}]$  be a  $(0,1)$   $t \times n$  matrix. Then we can use  $n$  sets (ordered)  $S_i$ 's to represent the columns of  $M$  where  
$$S_k = \{i : m_{i,k} = 1, i = 1, 2, \dots, t\}, k = 1, 2, \dots, n.$$
- The following sets represent the columns the  $(0,1)$ -matrix mentioned above:  
 $\{1,2,3\}, \{4,5,6\}, \{7,8,9\}, \{1,4,7\}, \{2,5,8\}, \{3,6,9\}, \{1,5,9\},$   
 $\{2,6,7\}, \{3,4,8\}, \{1,6,8\}, \{2,4,9\}, \{3,5,7\}.$
- This a collection of subsets of  $[1,9]$ , so a design.
- In fact, this design is the **Affine plane of order 3**.

# Remark

- A *design* is a pair  $(X, B)$  where  $X$  is a v-set and  $B$  is a collection of subsets of  $X$ . So, the matrix mentioned above can be viewed as a *design*.
- A design defined on a v-set can be viewed as a *binary code of length v*.
- A design  $(X, B)$  can also be viewed as a *hypergraph* with vertex set  $X$  and edge set  $B$ .

# Find Positives?

- We can do it, if the number of positives is not too many, say **at most 2**, by running the **9** tests simultaneously corresponding to rows.
- The reason is that the union of (**at most**) 2 columns are all distinct. In a word, the matrix is indeed an injective mapping.

# d-separable and d-disjunct matrices

- A matrix is *d-separable* if  $\cup D \neq \cup D'$  for any two distinct d-sets D and D' (columns), i.e. no two unions of d columns are the same.
- A matrix is *d-disjunct* if no column is contained in the union of any other d columns.
- A *d-disjunct* matrix is *d-separable*, *but the other way around may not be correct. (?)*

# Important Facts

- *A  $d$ -disjunct matrix can be applied to find  $k (\leq d)$  positives.*

*Proof.* The union of  $k (\leq d)$  columns corresponds to distinct *outcome vector*. (This is also true for  $d$ -separable matrices.)

- *$d$ -disjunct matrices have a simple decoding algorithm, namely, a column is positive if and only if it does not appear in a negative row.*

# Various Models

- *Errors* occurred (*in experiments*).
- There are *inhibitors* (*cancel the positive effects*).
- There is a *threshold* (*enough positives*).
- Defective items are sets: *complex* model.
- *Competitive* model: the number of defectives is unknown.
- The defectives are *mutually obscuring*.

# Further Remarks

- We can approach this study via different topics such as *combinatorial design, algebraic combinatorics, coding theory and graph theory*, and of course *algorithms*.
- Group testing does play an important role in applications such as *computational molecular biology, network security, image compression, data patterns, fingerprinting in multimedia*, ..., etc.

# *Human Genome Project*



- The Human Genome Project was a 15-year-long, publicly funded project initiated in 1990 with the objective of determining the DNA sequence of the entire euchromatic human genome within 15 years.<sup>[4]</sup> In May 1985, Robert Sinsheimer organized a workshop to discuss sequencing the human genome,<sup>[5]</sup> but for a number of reasons the NIH was uninterested in pursuing the proposal. The following March, the Santa Fe Workshop was organized by [Charles DeLisi](#) and David Smith of the Department of Energy's Office of Health and Environmental Research (OHER).<sup>[6]</sup> At the same time [Renato Dulbecco](#) proposed whole genome sequencing in an essay in Science.<sup>[7]</sup> James Watson followed two months later with a workshop held at the Cold Spring Harbor Laboratory.

- The fact that the Santa Fe workshop was motivated and supported by a Federal Agency opened a path, albeit a difficult and tortuous one,<sup>[8]</sup> for converting the idea into public policy. In a memo to the Assistant Secretary for Energy Research (Alvin Trivelpiece), Charles DeLisi, who was then Director of the OHER, outlined a broad plan for the project.<sup>[9]</sup>

# Computational Molecular Biology

- DNA解序後，需分析訊息，但資料量30億對。
  - 需要現代科技輔助。

- 主要課題

- 序列組合
- 序列分析
- 基因認定
- 生物資訊資料庫、種族樹建構、蛋白質三維結構推測…。

# 序列比對探究

- 說明
  - 找序列中「相似」及「相異」的部份。
  - 為何要比對序列？
  - 為何要使用電腦比對？
  - 如何使用電腦比對？
  - 困難點：序列型態多樣、需建構不同的資料結構及**演算法** (*Algorithm*)。

# Example of Group Testing

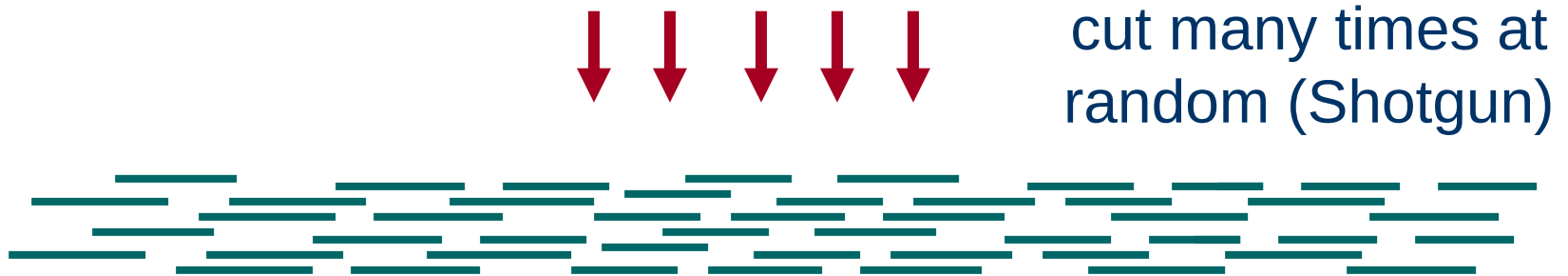
- In screening clone library the goal is to determine which clones in the clone library hybridize with a given probe in an efficient fashion. A clone is said to be **positive** if it hybridize with the **probe(探針)**, and **negative** otherwise.

# Shotgun Sequencing

- Shotgun sequencing is a throughput technique resulting in the sequencing of a large number of bacterial genomes, mouse genomes and the celebrated human genomes. In all such projects, we are left with a collection of **contigs** that for special reasons cannot be assembled with general assembly algorithms. **Continued ...**

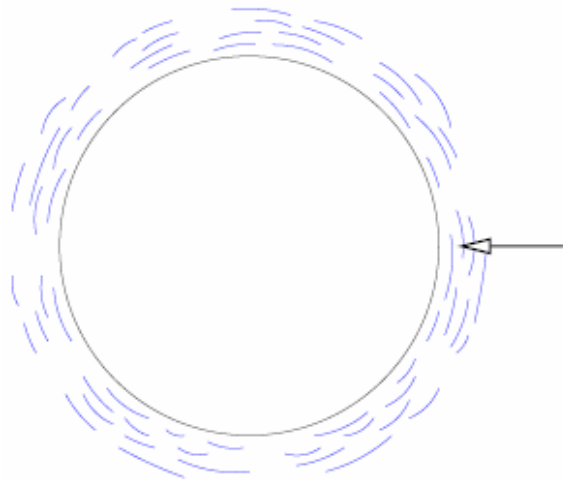
# Shotgun approach

genomic segment



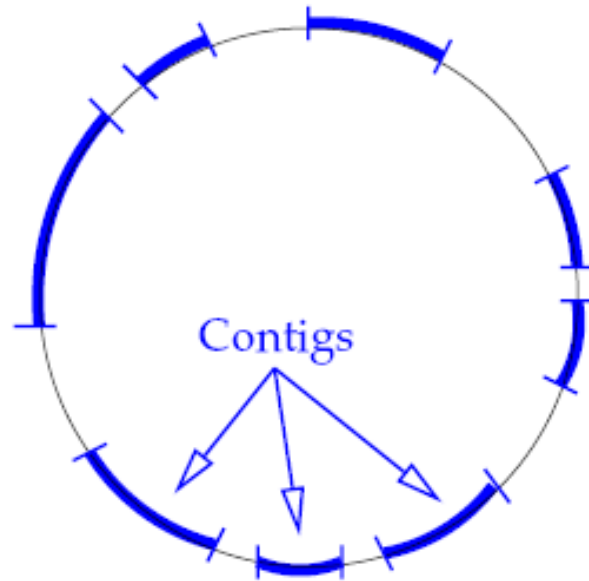
# Whole-genome shotgun sequencing

- Short reads are obtained and covering the genome with redundancy and possible gaps.

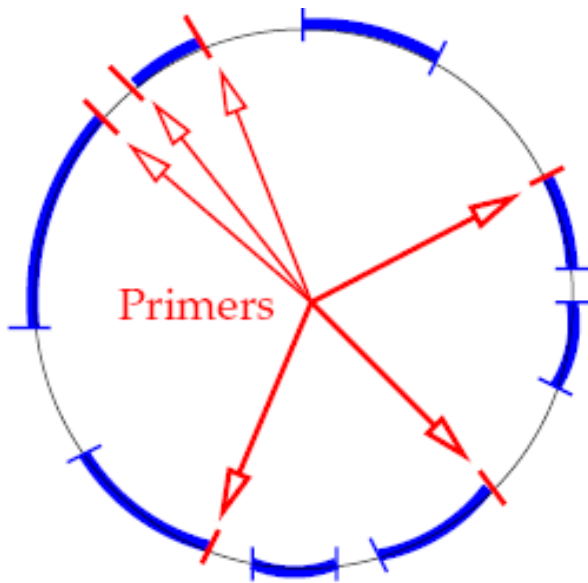


Circular genome

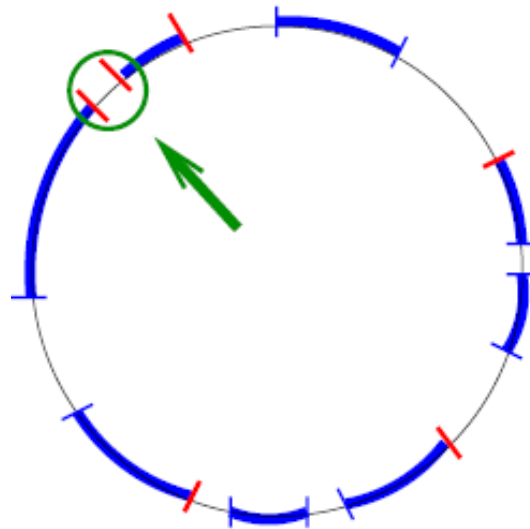
- Reads are assembled into contigs with unknown relative placement.

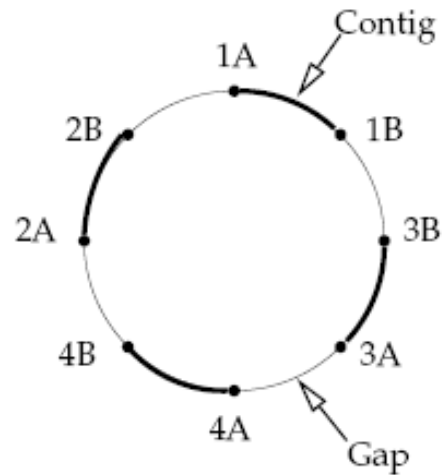


– **Primers** : (short) fragments of DNA characterizing ends of contigs.



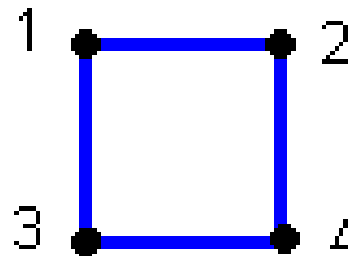
- A **PCR (Polymerase Chain Reaction)** reaction reveals if two primers are proximate (adjacent to the same gap).
- **Multiplex PCR** can treat multiple primers simultaneously and outputs if there is a pair of adjacent primers in the input set and even sometimes the number of such pairs.

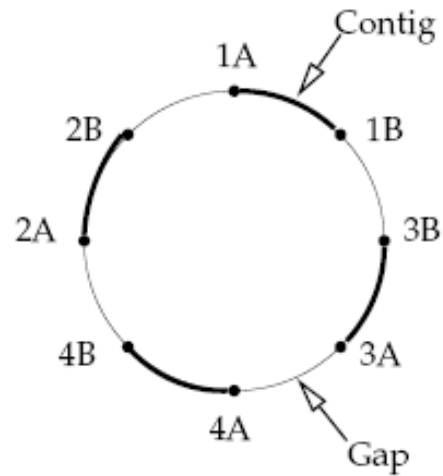




- Two primers of each contig are “mixed together”
  - Find a Hamiltonian cycle by PCRs!

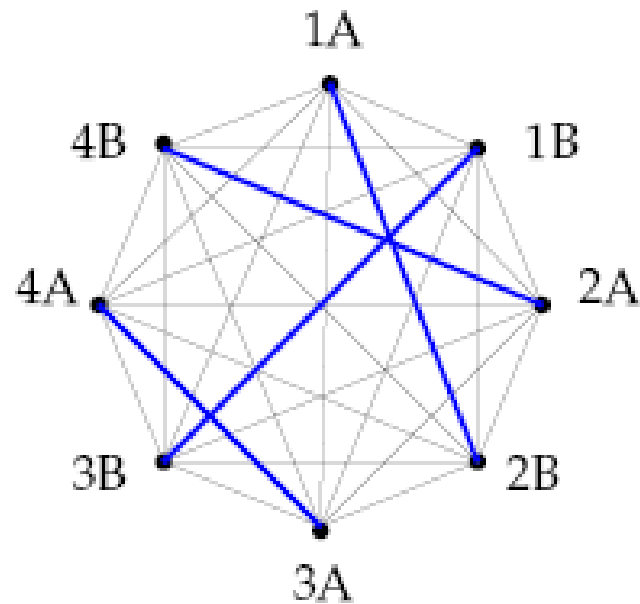
1A ——— 1B  
 2A ——— 2B  
 3A ——— 3B  
 4A ——— 4B





1A ——— 1B  
 2A ——— 2B  
 3A ——— 3B  
 4A ——— 4B

- Primers are treated independently.
  - Find a perfect matching by PCRs.



# Goal

- Our goal is to provide an experimental protocol that identifies all pairs of adjacent primers with as few PCRs (queries) (or multiplex PCRs respectively) as possible.

# Mathematical Models

## *Hidden Graphs or Hypergraphs (Reconstructed)*

- Topology-known graphs, e.g. Hamiltonian cycle, matching, star, clique, bipartite graph, ..., etc.
- Graphs of bounded degree
- Hypergraphs
- Graphs of known number of edges

# Compressed Sensing

- In compressed sensing (**CS**), we are given an  $n$ -dimensional sparse signal with support size  $k$ .
- In general  $k$  is small comparing to  $n$ .
- **Random (?)** projections of the sparse signal are obtained. (We notice the projection if it is not orthogonal to the signal, i.e. the inner product of the two vectors is not equal to zero.)
- ***The goal is to identify the support set while minimizing the number of projections.***

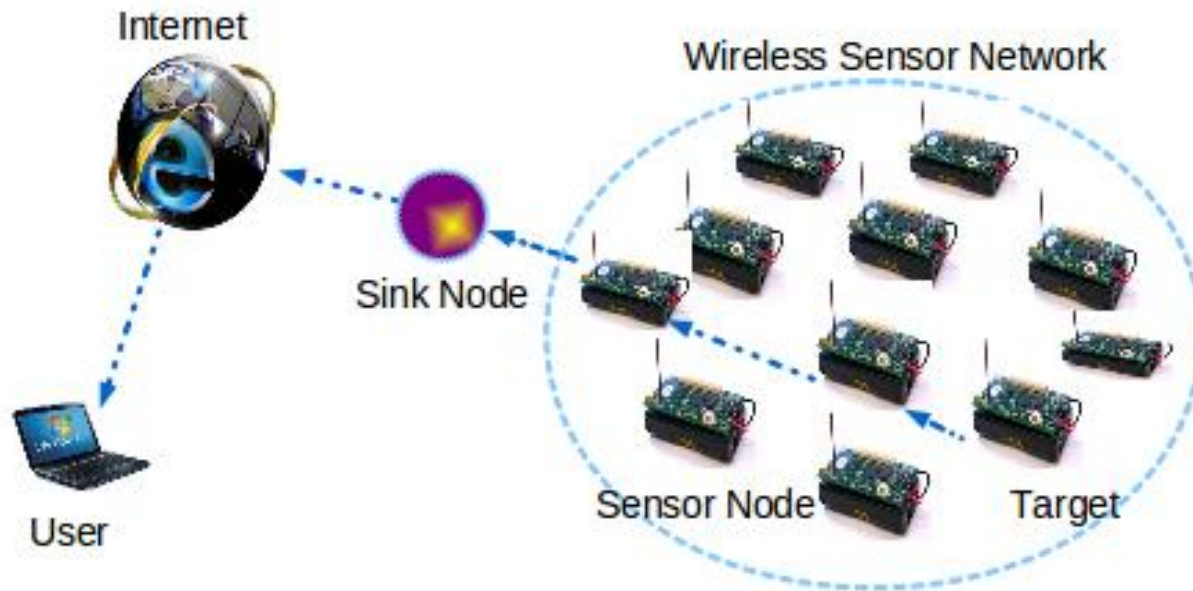
## *Have you seen “group testing”?*

- We shall find a measurement  $t \times n$  matrix  $A$  with as less rows (minimizing  $t$ ) as possible such that after  $t$  projections the support set of the sparse signal can be obtained.
- Here we only consider the Boolean version ( $0, 1$ ) of  $CS$ .
- *Noticed that we use non-adaptive group testing algorithms for compressed sensing.*

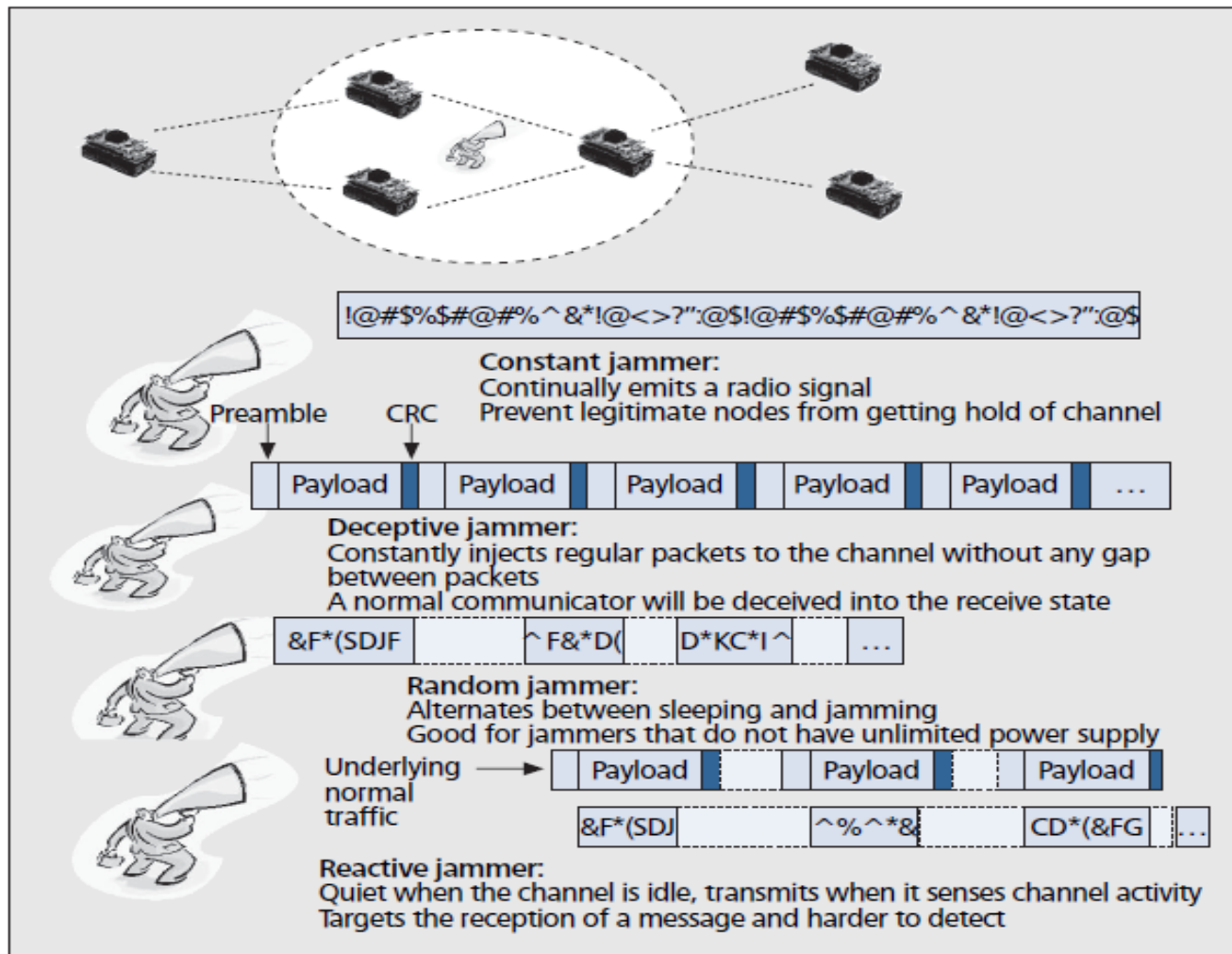
# *Randomized Constructions*

- We may construct the measurement matrix by selecting its rows randomly. (Apply Lovász Local Lemma.)
- If the matrix has  $t$  rows, then we obtain an outcome vector which is a  $t \times 1$  column vector.
- Noticed that the “0” components of the outcome vector plays the most important role.
- If the channel is noise-free, then we can find a bunch of items which are not defective. (See it?)

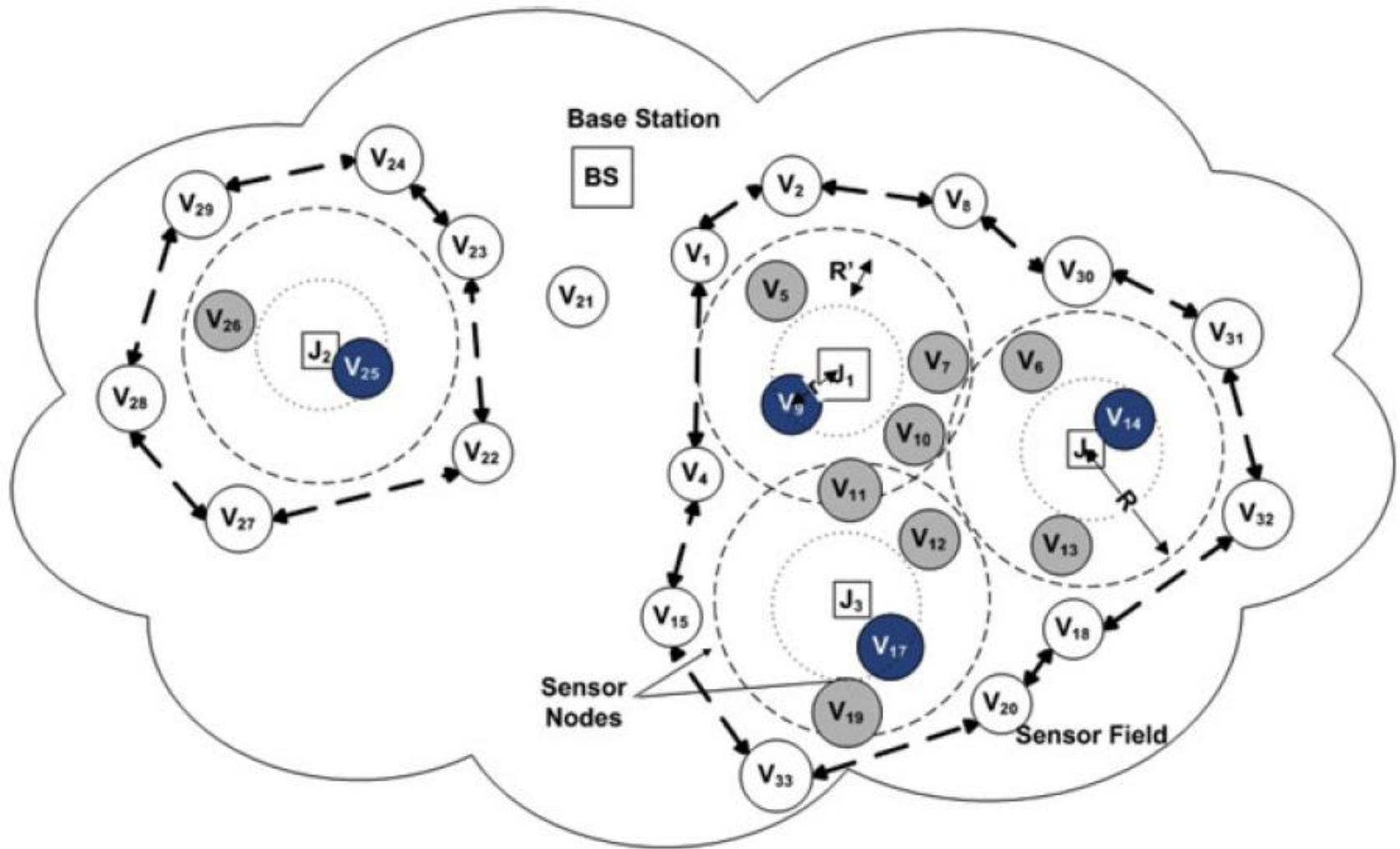
# Wireless Sensor Network



# Jammer



# Basic Model



- *Group testing?*
- *Please refer to the book  
“Group Testing Theory in  
Network Security” written by  
My T. Thai.*

# Fingerprinting for Multimedia

- *Digital fingerprinting* is a technique for identifying users who use multimedia content for unintended purposes, such as redistribution.
- These fingerprints are typically *embedded* into the content using *watermarking* techniques that are designed to be robust to a variety of attacks.

# Somewhere in India



# Anti-collusion

- A cost-effective attack against such digital fingerprints is *collusion*, where several differently marked copies of the same content are combined to disrupt the underlying fingerprints.
- 集体犯案
- How can we design a “code” to catch them if it does happen? *Again, group testing works!*

# References

- A new construction of 3-bar-*separable* matrices via improved decoding of Macula construction (with F. K. Hwang), **Discrete Optim.**, 5(2008), 700-704.
- An upper bound of the number of tests in pooling designs for the *error-tolerant complex model* (with Hong-Bin. Chen and F. K. Hwang), **Optim. Lett.**, 2(2008), no. 3, 425-431.
- The minimum number of e-vertex-cover among *hypergraphs* with e edges of given rank (with F. H. Chang, F. K. Hwang and B. C. Lin), **Discrete Applied Math.**, 157(2009), 164-169.
- Non-adaptive algorithms for *threshold* group testing (with Hong-Bin Chen), **Discrete Applied Math.**, 157(2009), 1581-1585.

# Continued

- Identification and classification problem on pooling designs for *inhibitor* models (with Huilan Chang and Hong-Bin Chen), **J. Computational Biology**, 17(2010), No. 7, 927-941.
- Reconstruction of *hidden graphs and threshold* group testing (with Huilan Chang, Hong-Bin Chen and Chi-Huai Shih), **J. Combin. Optimization**, 22(2011), no. 2, 270 – 281.
- Group testing with multiple *mutually-obscuring positives* (with Hong-Bin Chen), **Lecture Notes Computer Science**, 7777(2013), 557 – 568.
- *Threshold* group testing on *inhibitor* model (Huilan Chang and Chie-Huai Shih), **J. Computational Biology**, Vol. 20, No. 6, 2013, 1 – 7.

# Continued

- Learning a *hidden graph* (with Huilan Chang and Chie-Huai Shih), **Optim. Lett.**, 8(2014), no. 8, 2341 – 2348.
- New bound on *2-bar-separable* codes of length 2 (with Minquan Cheng, J. Jiang, Yuan-Hsun Lo and Ying Miao), **Des. Code Cryptography**, 74(2015), no. 1, 31 – 40.
- Codes with the identifiable parent property for *multimedia fingerprinting* (with Minquan Cheng, J. Jiang, Yuan-Hsun Lo and Ying Miao), **Des. Code Cryptography**, 83(2017), no. 1, 71 – 82.
- Learning a hidden *uniform hypergraph* (with Huilan Chang and Chi-Huai Shih), **Optim. Lett.**, 12(2018), 55 - 62.

***Keep Moving Forward!***

