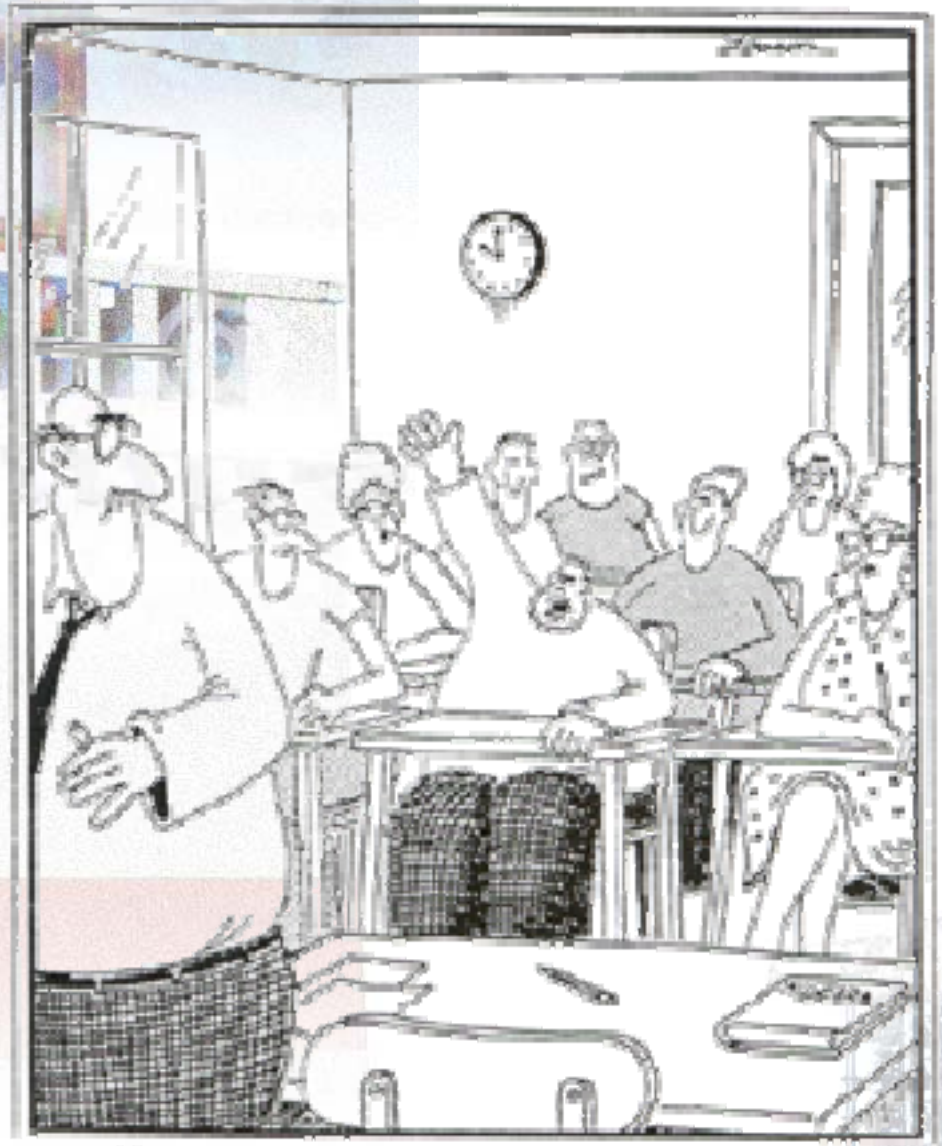
Introduction to Informatics

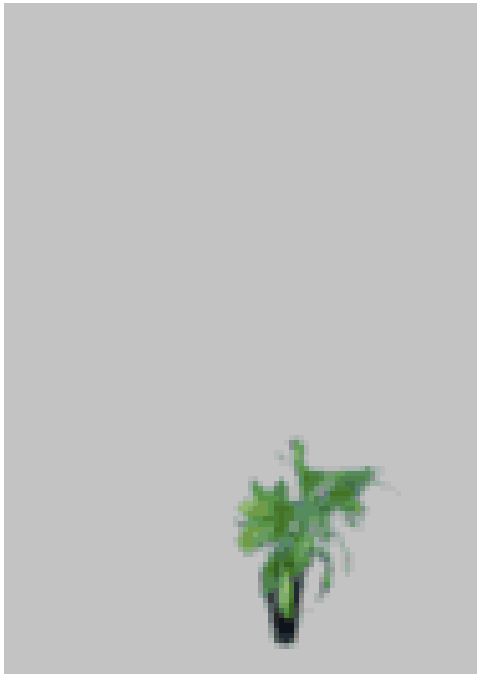


Lecture 16: Inductive Model Building

The Black Box and Text Frequency Analysis

Luis M.Rocha and Santiago Schnell

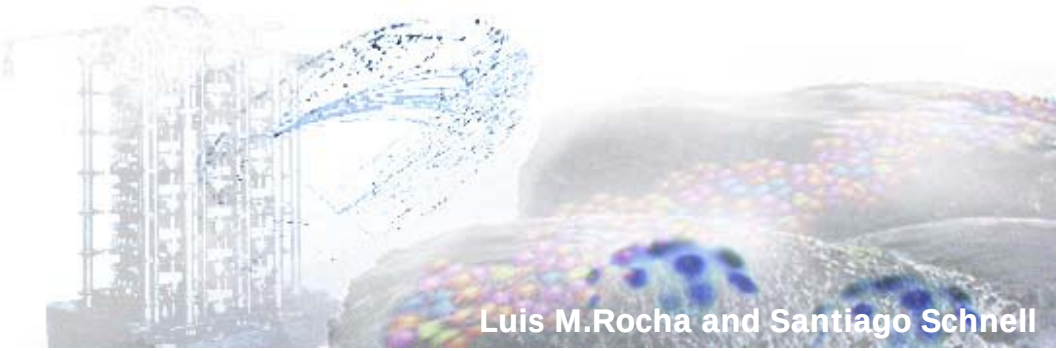
Readings until now



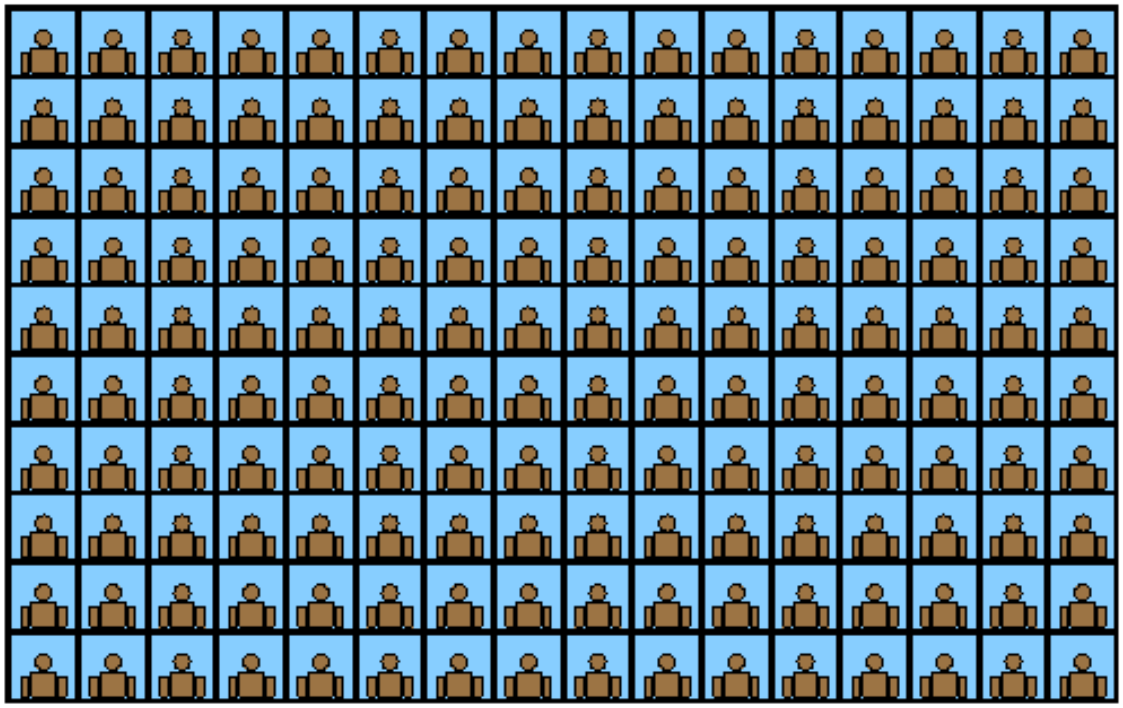
- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - *@ infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
    - From Andy Clark's book "*Natural-Born Cyborgs*"
      - Chapters 2 and 6 (pages 19 - 67)
    - From Irv Englander's book "*The Architecture of Computer Hardware and Systems Software*"
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)

Exam Schedule

- 11595 (T/R Class)
  - Midterm
    - March 1st (Thursday)
      - Being Graded
  - Final Exam
    - Thursday, May 3rd, 7:15-9:15 p.m.

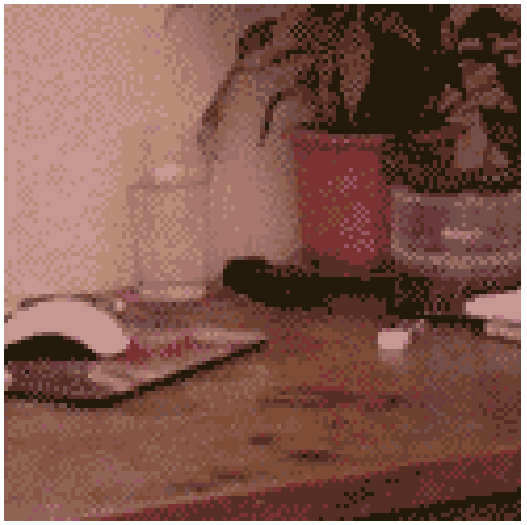


NO LAB THIS WEEK !!!



Luis M.Rocha and Santiago Schnell

Assignment Situation



- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed: due Friday, March 9
  - Next: Lab 8
    - Intro to Statistical Analysis using Excel
      - March 22 & 23, Due Friday, March 30

- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2, Being Graded
    - Third installment
      - Presented on March 8th, Due on March 30th
  - Group
    - First Installment
      - Presented: March 6th, Due March 9th

Get a Group NOW!

Summarizing Data

- Frequency
  - Number of times an item or value occurs in a collection
- Frequency Distribution
  - Given a collection of data items/values, the specification of all the distinctive values in the collection together with the number of times each of these items/values ***occurs*** in the collection
    - Table that organizes data into mutually exclusive classes
    - Shows number of observations from data set that fall into each class

[Chase and Brown, "General Statistics"]

Frequency Distribution (values)

Sorted Data: 30 data values (Carpet Looms)

15.2	15.2	15.3	15.3	15.3	15.3	15.3	15.4	15.4	15.4
15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.4	15.5	15.5
15.5	15.5	15.5	15.5	15.6	15.6	15.6	15.7	15.7	15.7

Frequency Distribution	Class	Tallies	Frequency
	15.2	//	2
	15.3	///	5
	15.4	/// /// /	11
	15.5	/// /	6
	15.6	///	3
	15.7	///	3

Luis M.Rocha and Santiago Schnell

Relative Frequency Distribution (values)

Relative Frequency Distribution	Class	Frequency (1)	Relative Freq. (1) ÷ 30	Cumulative Relative Frequency
	15.2	2	0.07	0.07
	15.3	5	0.16	0.23
	15.4	11	0.37	0.60
	15.5	6	0.20	0.80
	15.6	3	0.10	0.90
	15.7	3	0.10	1.00
		30	1.00	

Number of values

Freq. Distribution Film data (items)

Raw Data: Your favorite films

The Big Lebowski, Kung Fu Hustle, Team America – World Police, Kill Bill 1 + 2, Good Night, and Good Luck, Pulp Fiction,….

Sorted Movie Preferences

dumb and dumber	8
wedding crashers	7
office space	6
the matrix	5
jackass 2	4
old school	4
tommy boy	4
anchorman	3
mission impossible	3
scarface	3
super troopers	3
the departed	3

Class	Freq.	Rel. Freq.	%
…….			
dragon ball z	1	0.004	0.4%
dream catcher	1	0.004	0.4%
dumb & dumber	8	0.036	3.6%
elf	1	0.004	0.4%
enough	1	0.004	0.4%
face off	1	0.004	0.4%
Fast/furious/tokyo	1	0.004	0.4%
Fear/loath/vegas	2	0.009	0.9%
…….			
	223	1.000	100%

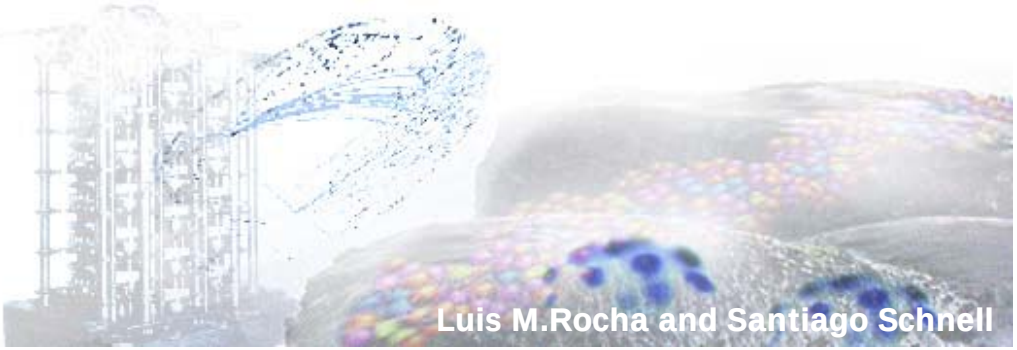
Movies

Votes (# Items)

Luis M.Rocha and Santiago Schnell

Grouped Frequency Distribution

- Further summarizes the data
  - Very important when the number of data values is large
- Procedure
  - Divide the interval describing the range of values in a small number of subintervals: ***classes***
    - E.g. 10, usually of equal width
  - Count how many values fall in each class



Building a Grouped Frequency Distribution

Data

100	**74**	84	95	95	110	99	87	100	108
85	103	99	83	91	91	84	110	113	105
100	98	100	108	100	98	100	107	79	86
123	107	87	105	88	85	99	101	93	99

- R = 123 - 74 = 49
- 49/10 = 4.9
- Tentative Class Interval Size = 5

Constructing a Grouped Frequency Distribution

Grouped Frequency Distribution

Class	Frequency (1)	Relative Freq. (1) ÷ 40	Cumulative Relative Frequency
70 - 74	1	.025	.025
75 - 79	1	.025	.050
80 - 84	3	.075	.125
85 - 89	6	.150	.275
90 - 94	3	.075	.350
95 - 99	8	.200	.550
100 - 104	8	.200	.750
105 - 109	6	.150	.900
110 - 114	3	.075	.975
115 - 119	0	.000	.975
120 - 124	1	.025	1.000

Additional Terms Associated with Grouped Frequency Distributions

Grouped Frequency Distribution

Class	Frequency (1)	Relative Freq. (1) ÷ 30	Cumulative Relative Frequency
70 - 74			
75 - 79			
80 - 84			
85 - 89			
90 - 94			
95 - 99			
100 - 104			
105 - 109			
110 - 114	3	.075	.975
115 - 119	0	.000	.975
120 - 124	1	.025	1.000

Class Interval Midpoint
(70 + 74)/2 = 72

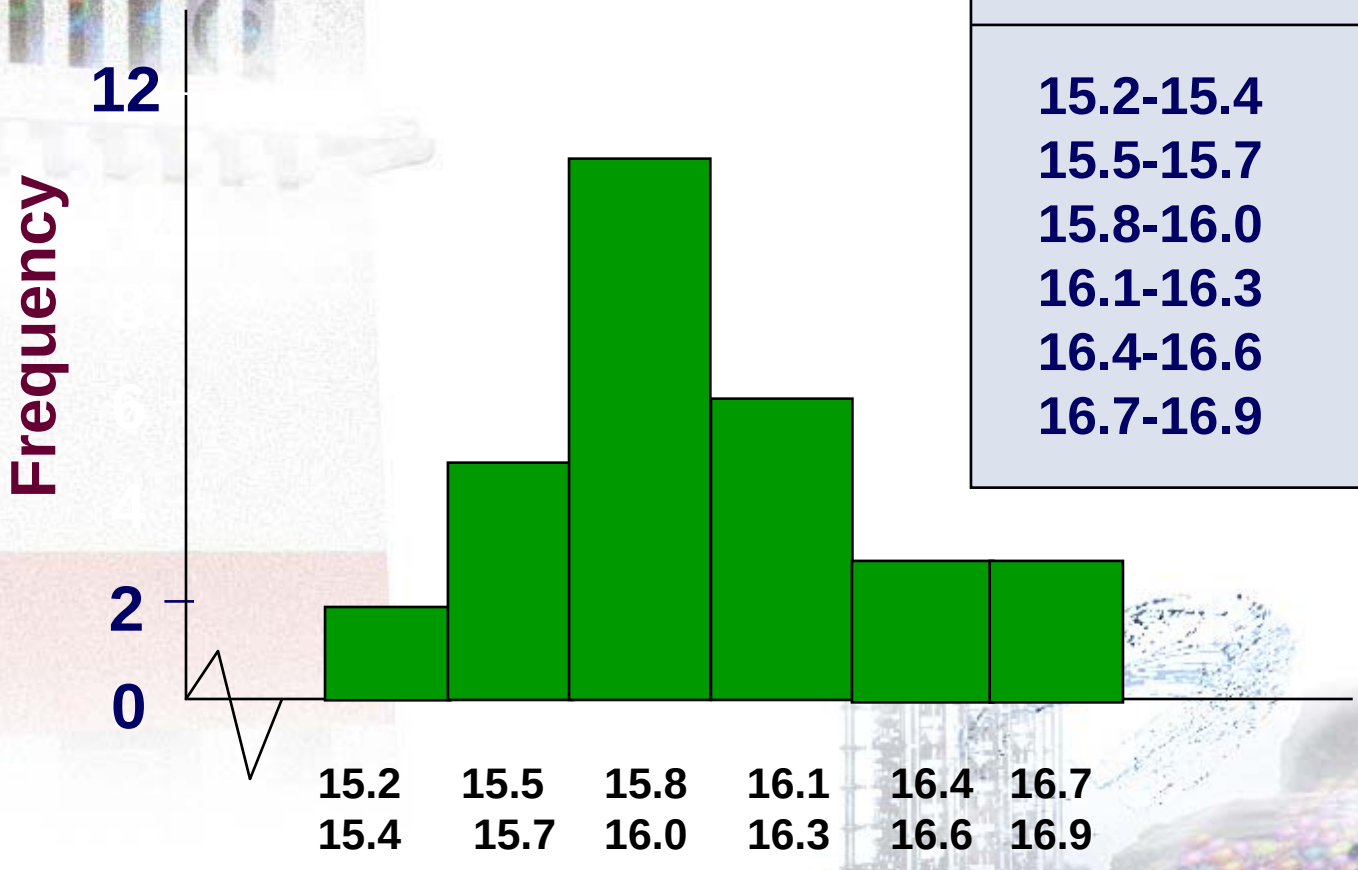
Upper Limit
84

Lower Limit
80

Frequency Histogram

Grouped Frequency Distribution for Carpet Loom Example (Yards of carpet)

Class	Frequency
15.2-15.4	2
15.5-15.7	5
15.8-16.0	11
16.1-16.3	6
16.4-16.6	3
16.7-16.9	3

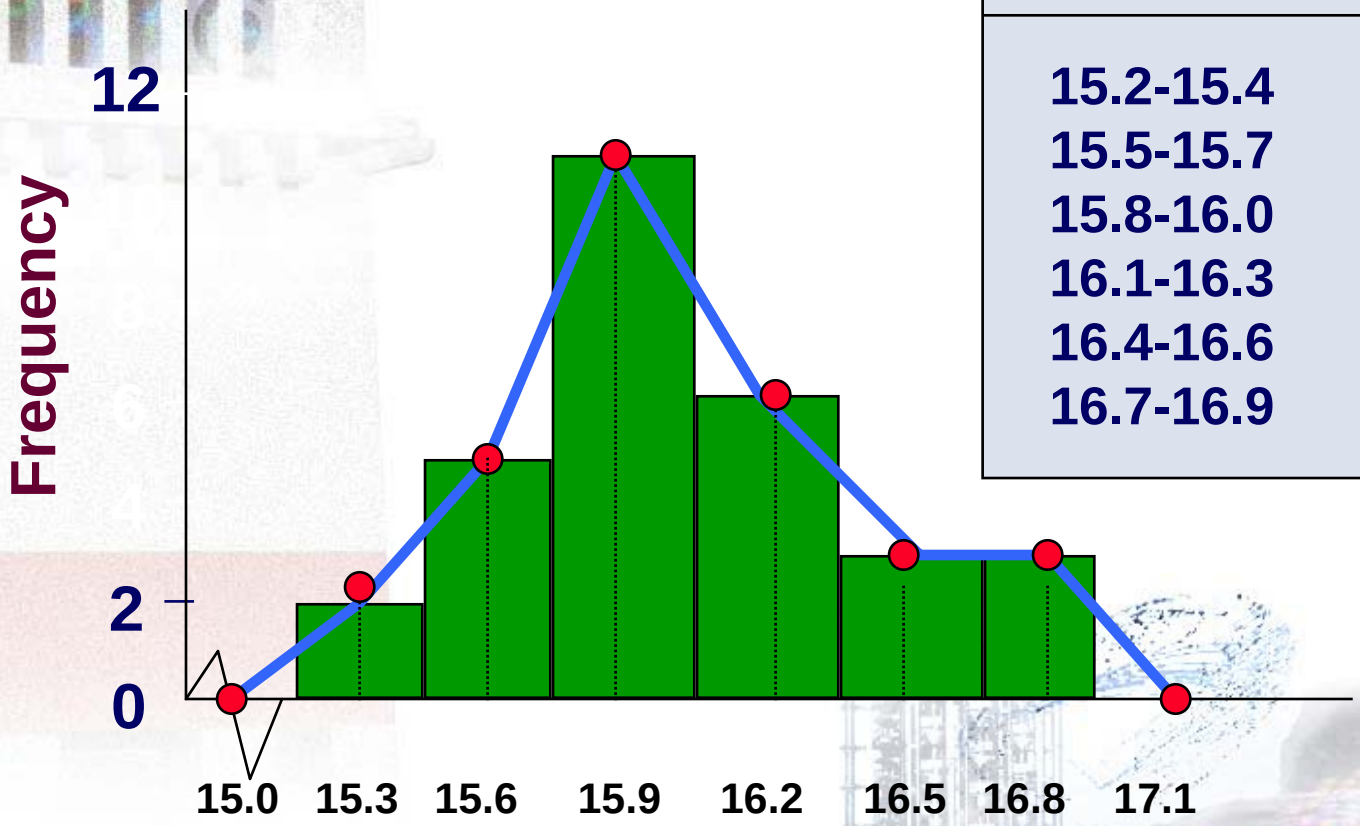


Luis M.Rocha and Santiago Schnell

Frequency Polygon

Grouped Frequency Distribution for Carpet Loom Example (Yards of carpet)

Class	Frequency
15.2-15.4	2
15.5-15.7	5
15.8-16.0	11
16.1-16.3	6
16.4-16.6	3
16.7-16.9	3

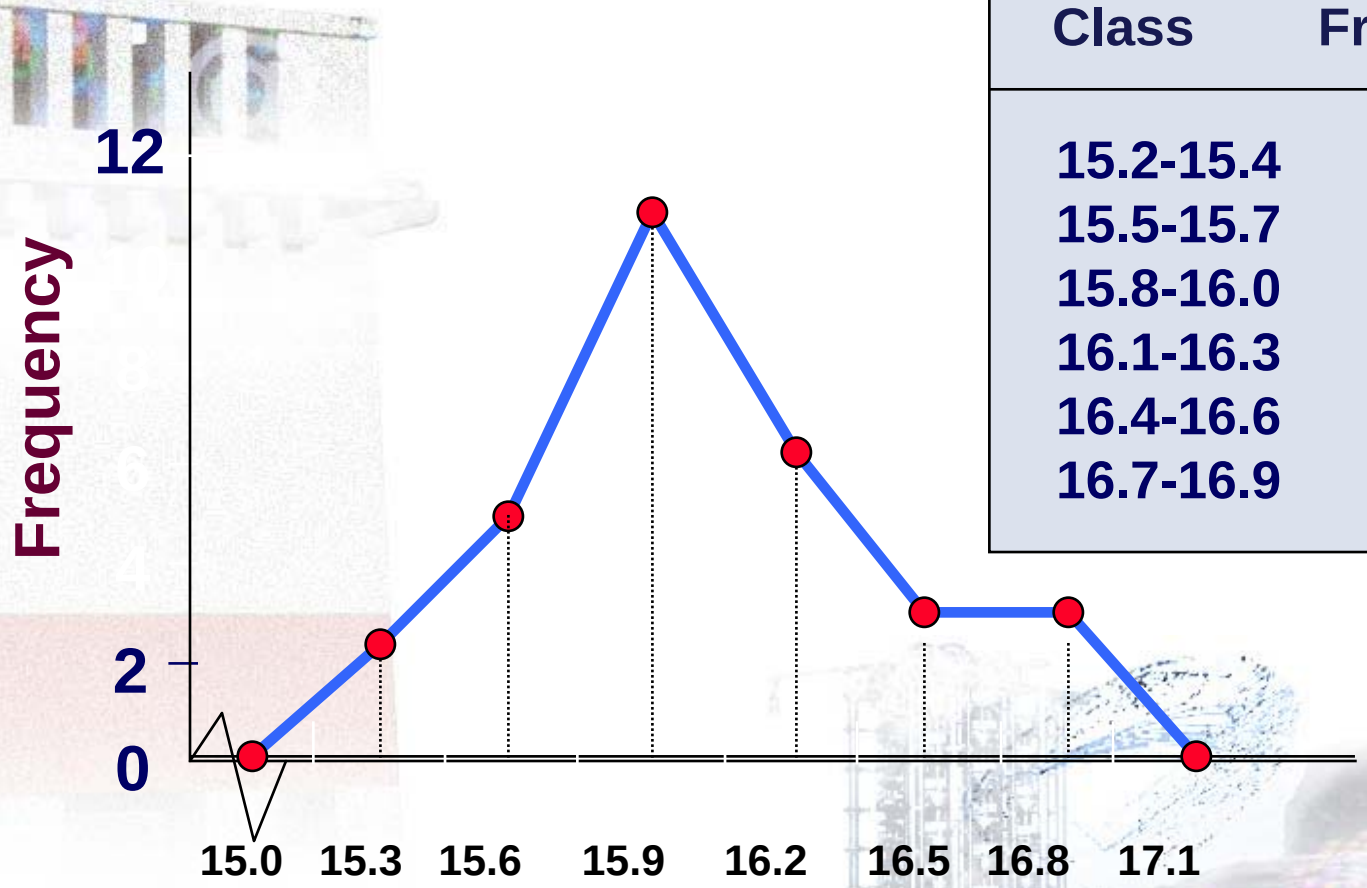


Luis M.Rocha and Santiago Schnell

Frequency Polygon

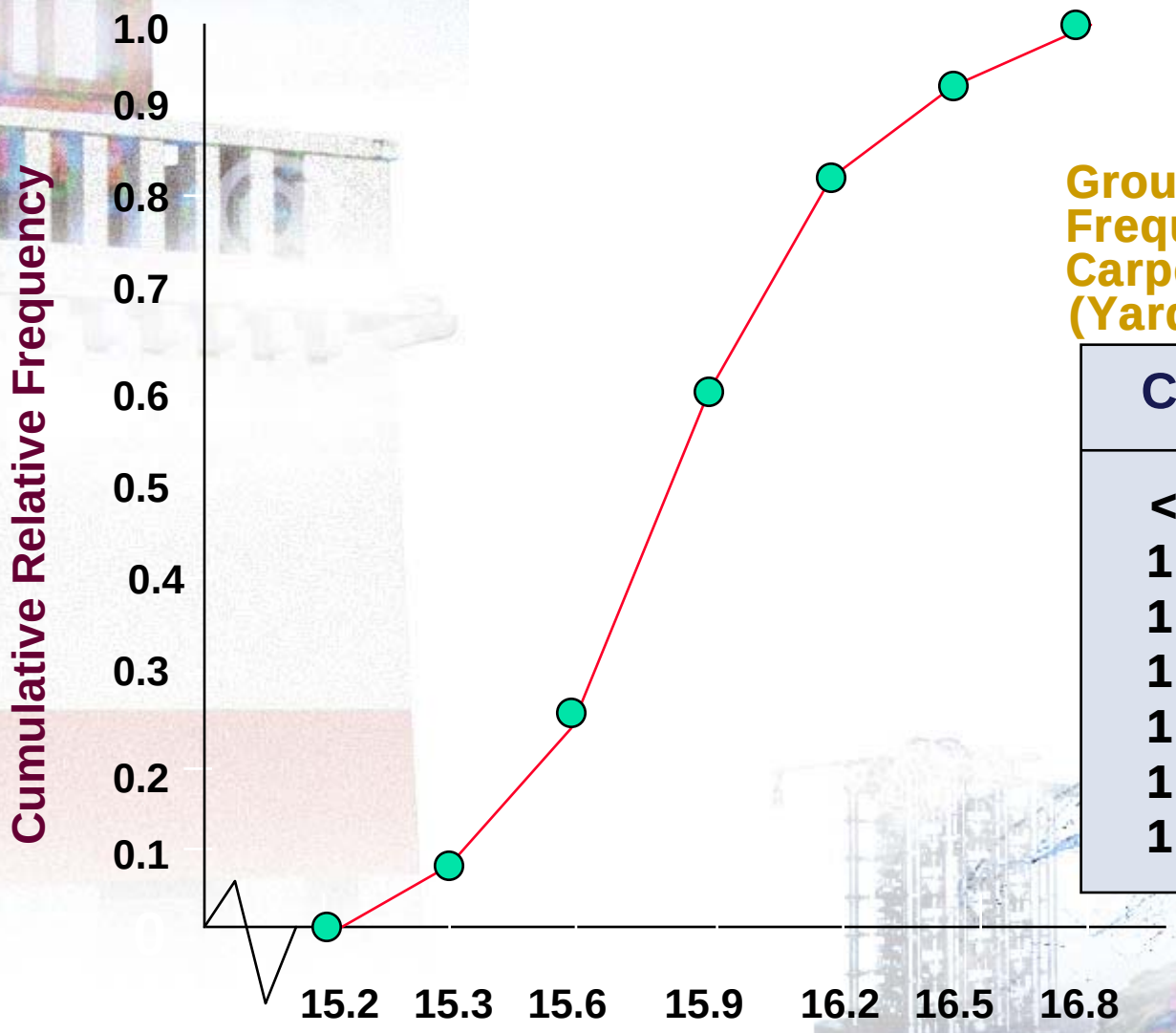
Grouped Frequency Distribution for Carpet Loom Example (Yards of carpet)

Class	Frequency
15.2-15.4	2
15.5-15.7	5
15.8-16.0	11
16.1-16.3	6
16.4-16.6	3
16.7-16.9	3



Luis M.Rocha and Santiago Schnell

Cumulative Relative Frequency Polygon: Ogive



Grouped Cumulative Frequency Distribution for Carpet Loom Example (Yards of carpet)

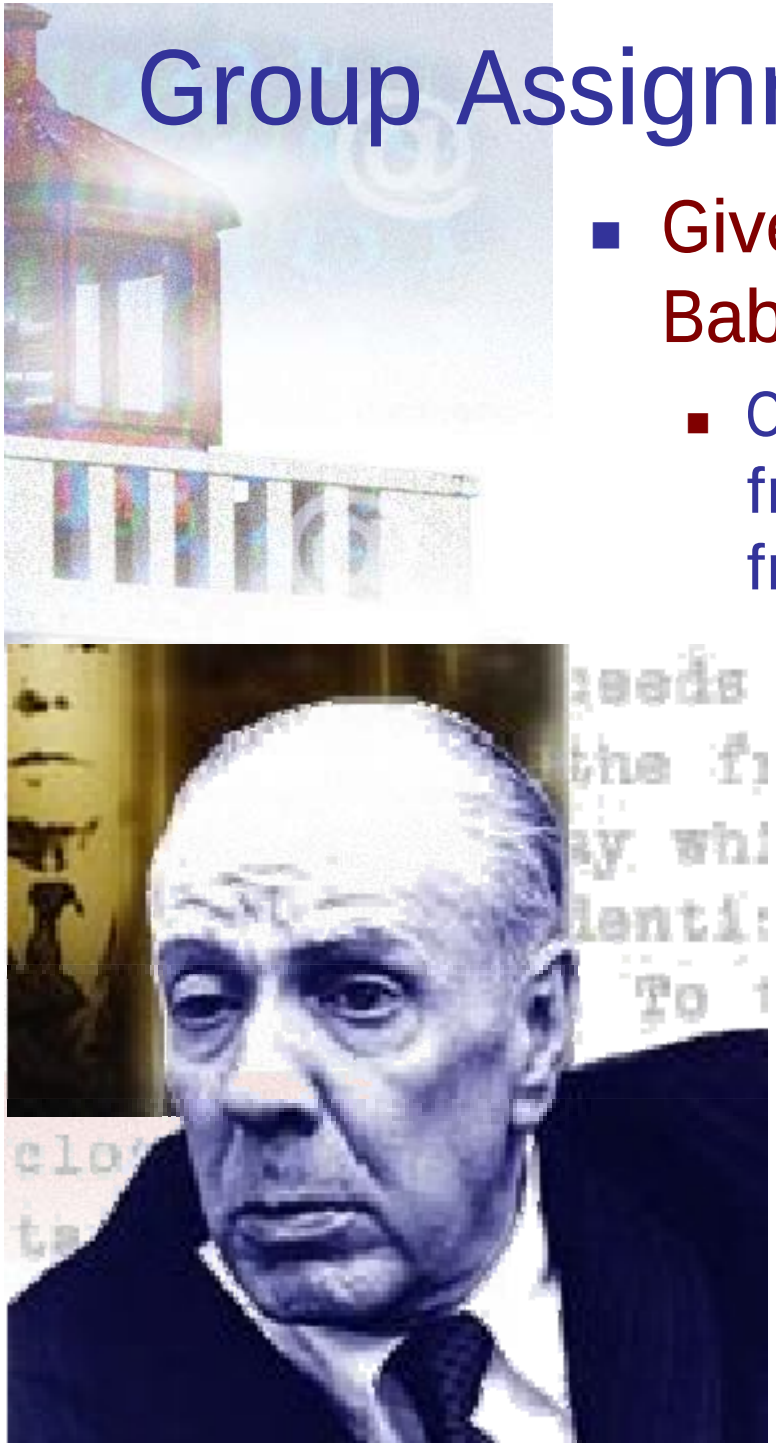
Class	C.R.Freq.
< 15.2	0.00
15.2-15.4	0.07
15.5-15.7	0.23
15.8-16.0	0.60
16.1-16.3	0.80
16.4-16.6	0.90
16.7-16.9	1.00

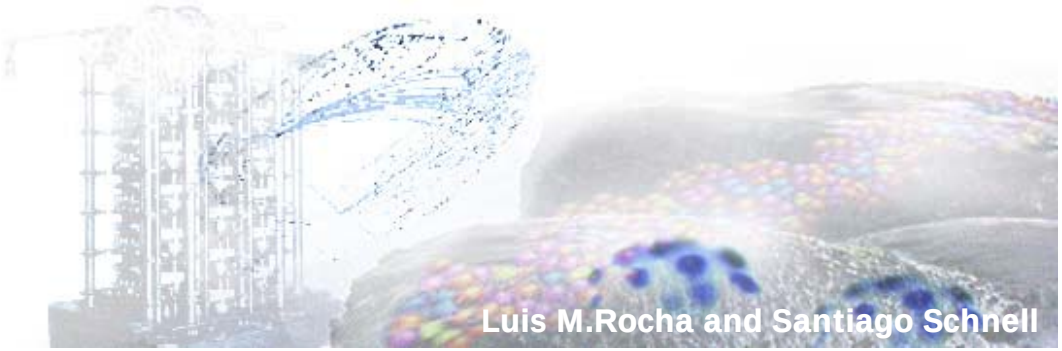
Frequency Analysis and Cryptography

- Cryptography
  - Derived from the Greek word *Kryptos*: hidden
- See Simon Singh's The Code Book CD-ROM
- Check out Cybersecurity group @Informatics

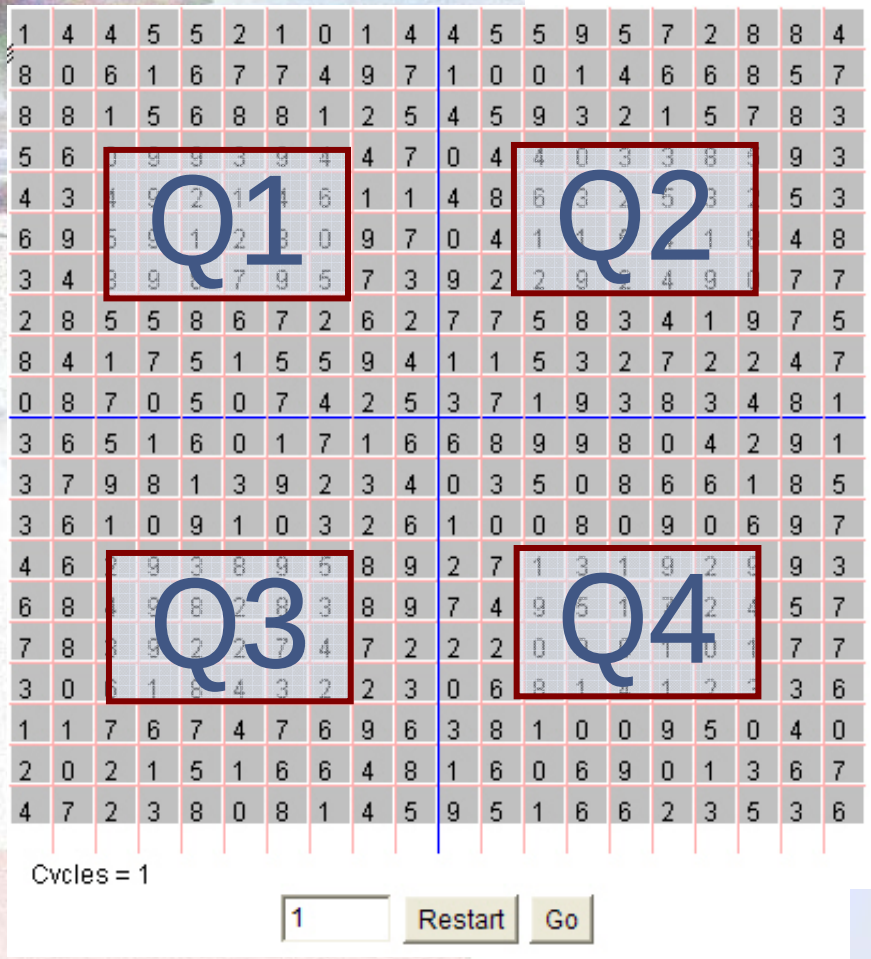
Group Assignment: First Installment

- Given the text of “Lottery of Babylon” by Jorge Luis Borges
  - Compute the frequency, relative frequency, and cumulative relative frequency distribution of letters
    - In the Spanish and the English Text
    - Upload to Oncourse
      - Note: in the Spanish version, lookout for ñ, á, é, í, ó, ú





Individual Assignment – Part II



Analysis and Observations

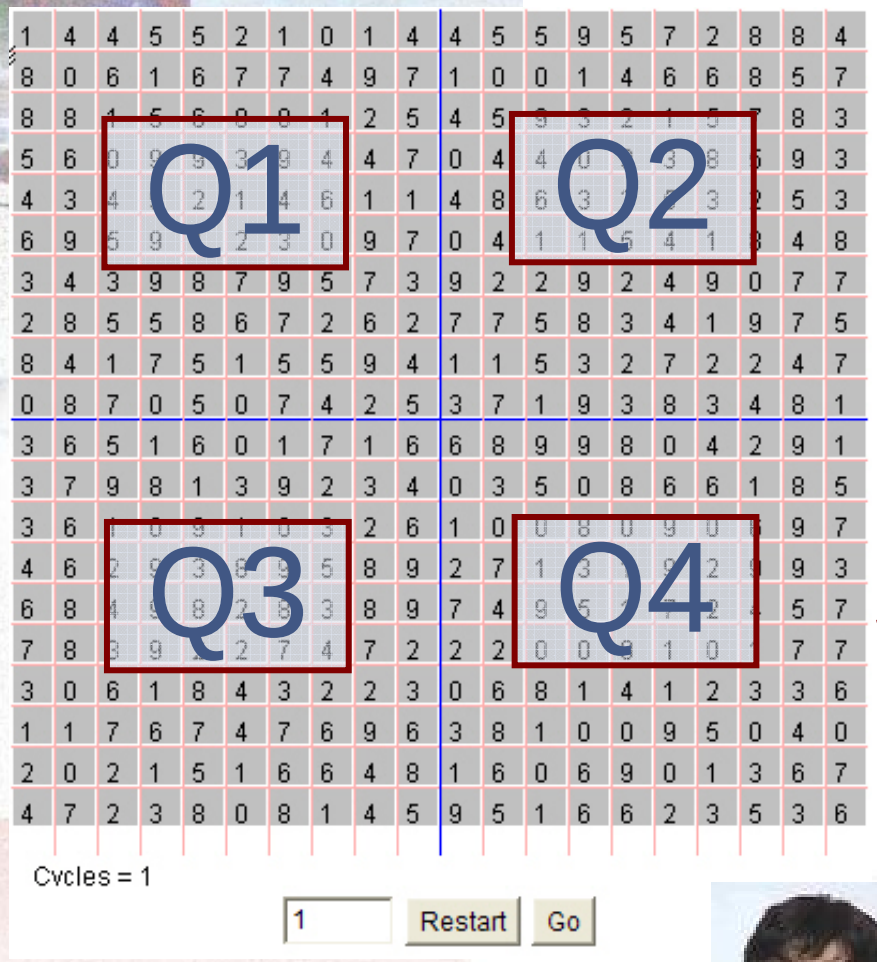
- Jeffrey Randall Cooley
  - Q4
    - looked at the number **4** one iteration at a time.
      - 4 -> 1 -> 9 -> 4 -> 1
    - All numbers keep appearing
  - Q2
    - Can go to all 0
    - Can go to all 0 and evens (2, 4, 6, and 8)
  - Q1
    - Groups numbers together
  - Q3
    - Always goes to all zero (black)
- Andrew James Dempsey
  - Concentrated on Q3
    - Dies between 400-800 cycles
    - One or none changes per cycle
    - 5 always changes to 0
    - Every number goes to zero
    - 1,2,4,6,7,9 all go to four different numbers
    - Only 3 -> 1
    - 3,4,6,8 -> 2 ;
    - 9 -> 3;
    - No numbers go to 7 or 9!

Test these hypotheses further!

Luis M.Rocha and Santiago Schnell

Individual Assignment – Part II

- Andrew Y M Kim



- Looked at 20th cell by 19th row
  - Out of 100 trials, 40 started even and ended odd
    - Odd-to-Odd: 24%
    - Did not change: 16%
    - Even-to-Even: 12%
    - Odd-to-Even: 8%
  - Proposed that the cell has rules as a logic gate does for each respective number

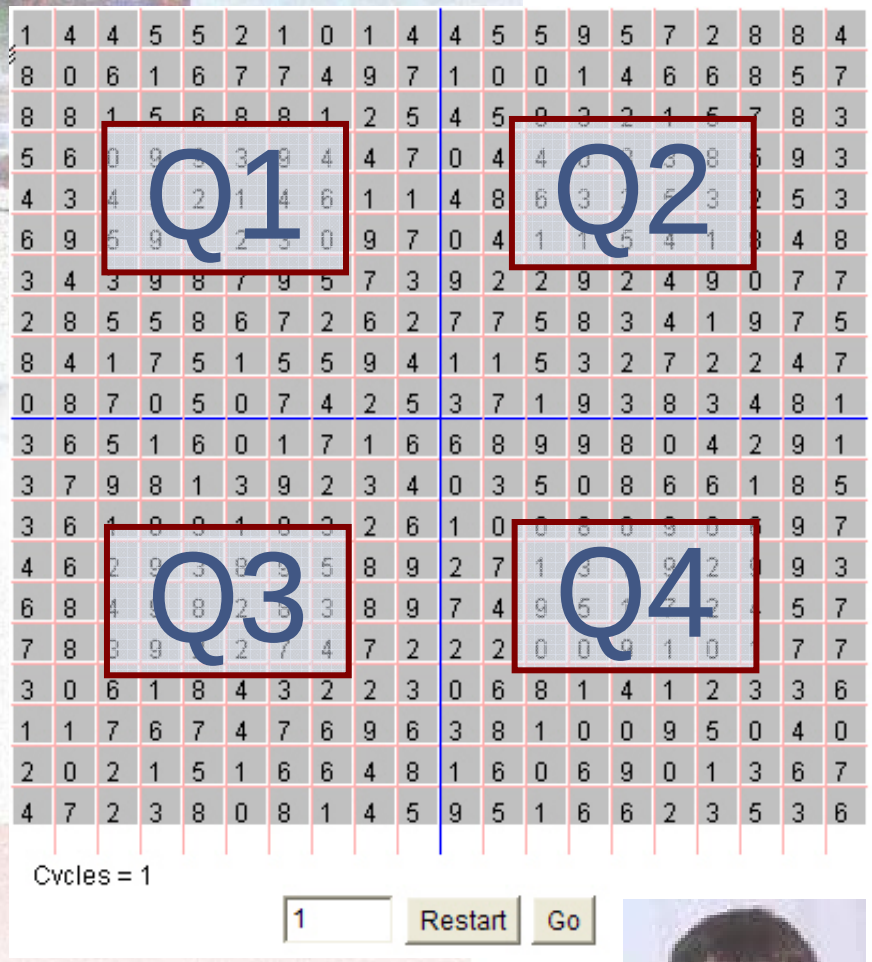
John Oglesby

- Q3
  - Observed when specific digits disappear
    - 1 goes first: 250-255 cycles
    - 9 next: 265 and 270 cycles.....
    - 2 last to die
    - Etc.
  - **Conclusions**
    - Even numbers are more likely to stay
  - But how many tries!!?

Luis M.Rocha and Santiago Schnell

Individual Assignment – Part II

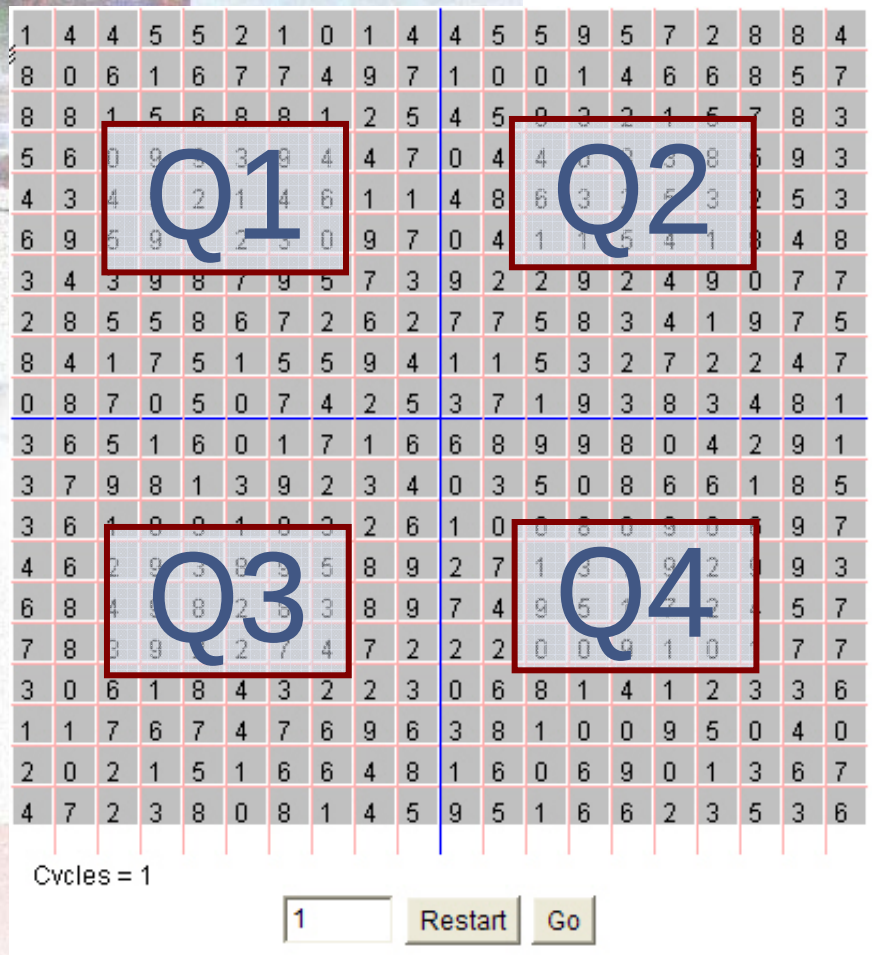
- Donald Peek

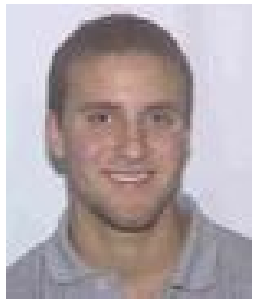


- Found a correlation between colors and numbers
  - 0 = Black ; 1 =GREY;
  - 2 = CYAN; 3 = BLUE;
  - 4 =GREEN/PINK; 5 =GREEN/PINK;
  - 6 =WHITE; 7 = YELLOW/RED;
  - 8 =YELLOW/RED; 9 = LIGHT PINK;
- Q1
  - TEST: ran the box 15 times at 1200 cycles each and recorded which numbers remain in the box.
    - NUMBER TIMES COUNTED RANK
    - 1----011111110010110-----10---2
    - 2----100110001111101------9----3
    - 3----011000110001110------7----4
    - 4----001111000000001------5----5/6
    - 5----000010011011100------5----5/6
    - 6----110000000011000------4----7
    - 7----100001000001000------3----8/9
    - 8----000010000101000------3----8/9
    - 9----100000000000000------1----10
    - 0----011101111101111-----12----1
  - Lower digits more frequent?
    - Test
- Quadrants are independent?
- How to study so many changes?

Luis M.Rocha and Santiago Schnell

Individual Assignment – Part II





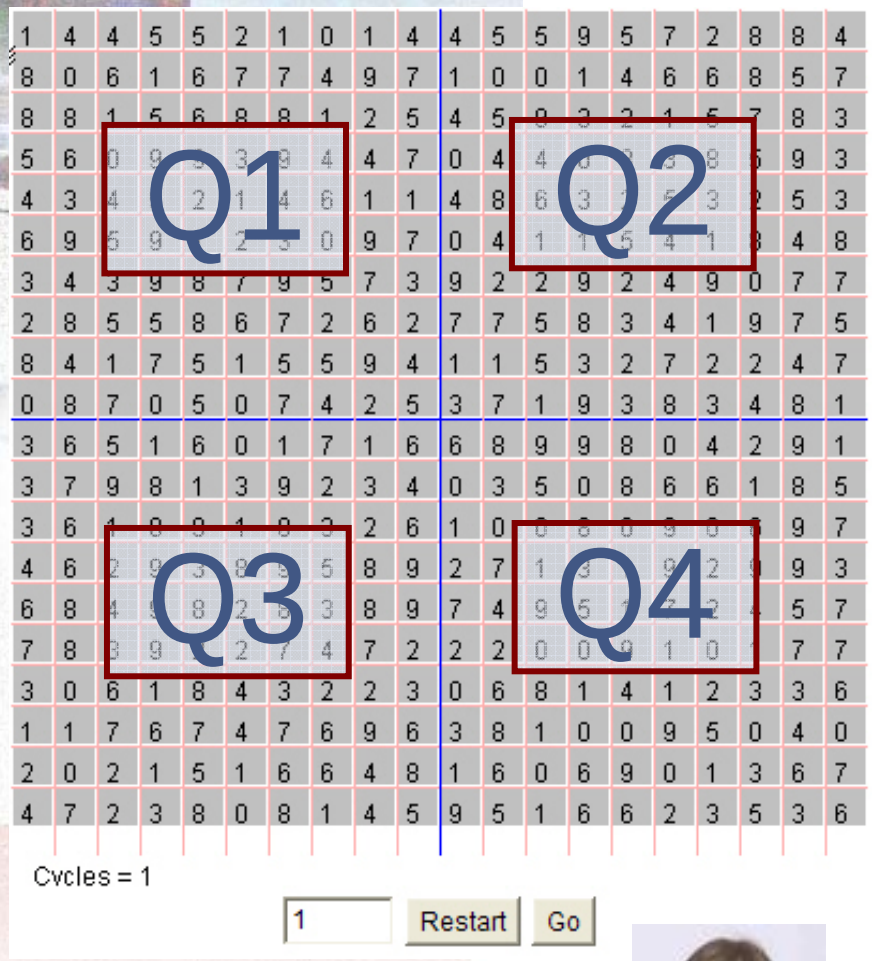
Cliff Taylor

- Boids
  - Quadrants represent one of the 3 behaviors
    - separation, alignment, and cohesion.
    - Q1 contains the results
      - Grouping/flocking
    - Now, the top left quadrant represents all three of
  - Also proposed by Donald Peek for Q1 and Q3
  - Sarah Kepa, J.T.Waugh, Marcus Bigbee and Andrew Philbrick observed grouping behavior in the left quadrants.




Luis M.Rocha and Santiago Schnell

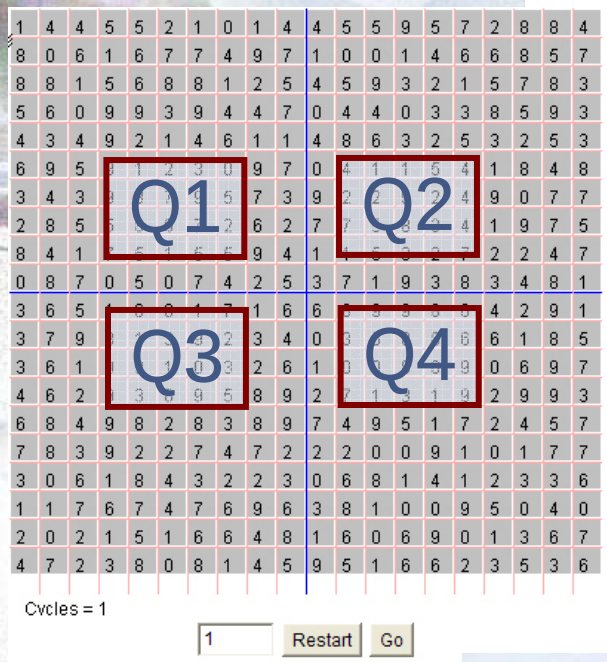
Individual Assignment – Part II



- Jacob Levi Marsh
  - **First Try: 1000 Cycles**
    - Q1: Four numbers...1,2,3,4.
    - Q2: Two numbers...5,0
    - Q3: Only 0
    - Q4: All numbers...0-9
  - **Second Try: 1000 Cycles**
    - Q1 : Contains four numbers...0,1,4,8
    - Q2 : Contains only 0
    - Q3 : Contains only 0
    - Q4 : Contains all numbers...0-9
  - **Third Try: 1000 Cycles**
    - Q1 : Contains six numbers...0,1,2,3,6,9
    - Q2 : Contains all numbers 1-9
    - Q3 : Contains only 0
    - Q4 : Contains all numbers...0-9
  - **Fourth Try: 1000 Cycles**
    - Q1 : Contains five numbers...0,1,2,3,4
    - Q2 : Contains five numbers...0,2,4,6,8
    - Q3 : Contains only 0
    - Q4 : Contains all numbers...0-9
  - **Findings:**
    - Q1 can have 1, 3, 4, 5, 6, or 7 different numbers.
    - Q2 can have 1, 2, 5, or 10 different numbers.
    - Q3 only has one number which is 0.
    - Q4 always has 10 different numbers.
  - Look at transition details!

- Ashlee Nicole Sweeden
  - Similar conclusions

Individual Assignment –Part II

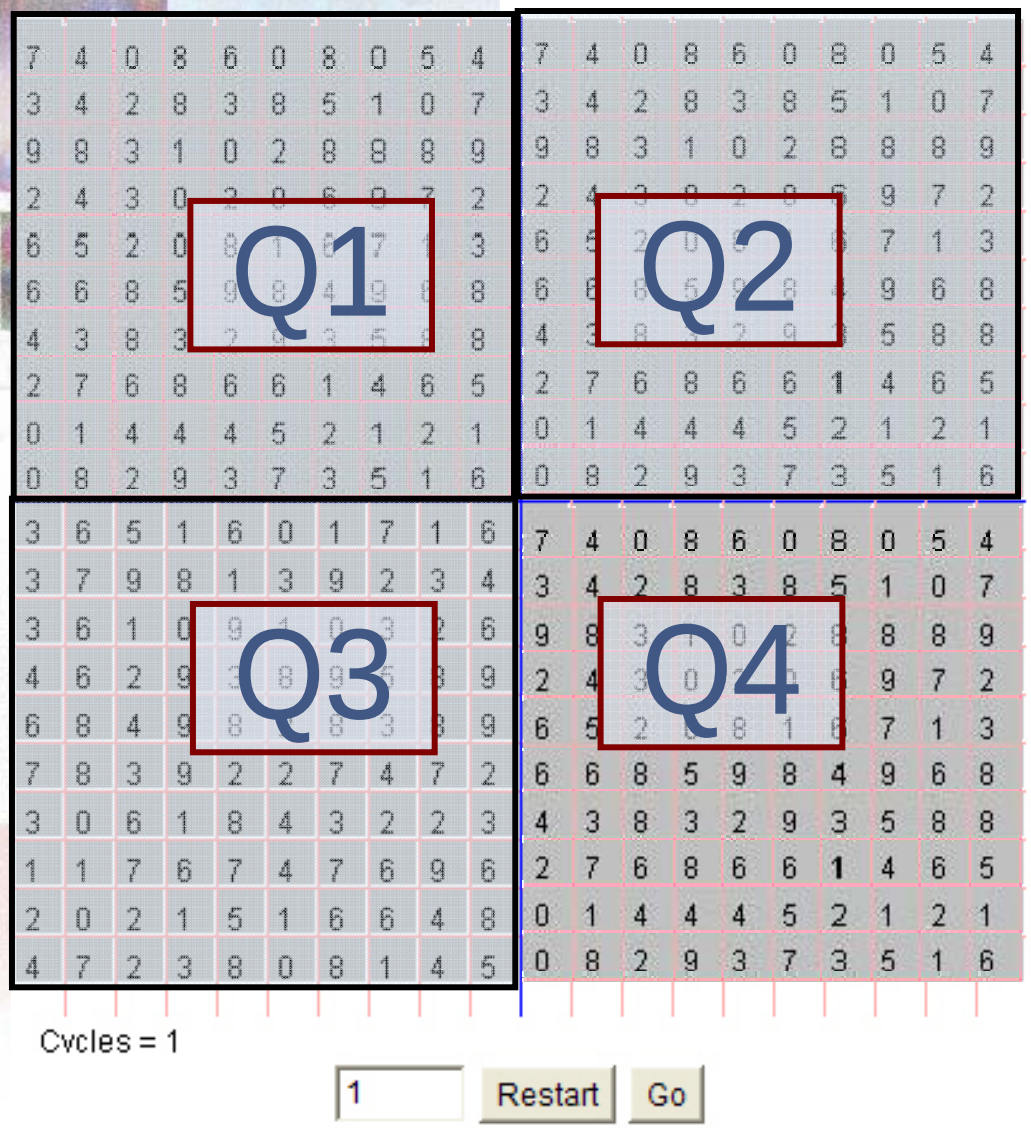


- Samuel Abraham Ritter
  - Ran 1,000 cycles and it appeared at first that the number in each of the corners was represented by the number that was depicted most in each of the 4 individual squares.
    - After counting, my theory was not correct since one of the inner squares did not follow this .....it was only the bottom right inner square that did not follow this theory

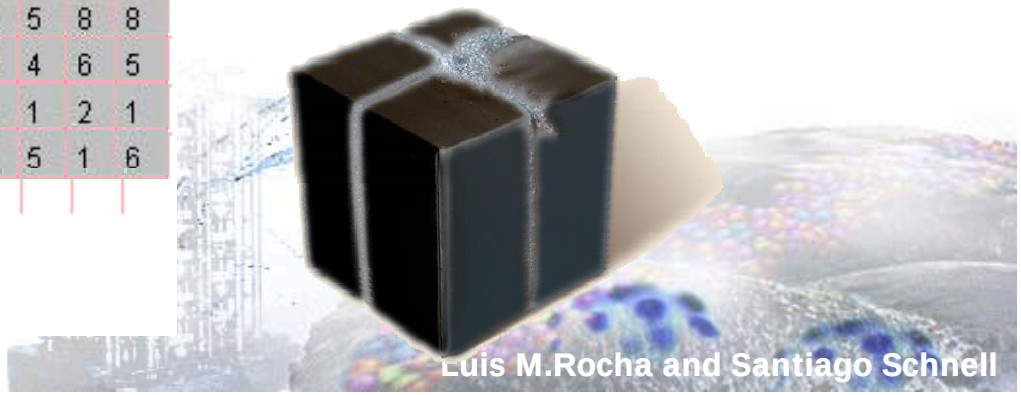
- Q2
  - Nathaniel Wishart
    - an odd number can either be replaced by another odd number or an even number, but an even number can only be replaced by another even number.
  - Sarah Kepa
    - Unless they go to zero, the odd numbers seem to change to odd numbers, and the even number change to even numbers.
  - Andrew Glenn Philbrick
    - From what I noticed even numbers are only able to be replaced by a new even number for a while, but once a great number of cycles are ran the pattern seems to disappear (?)

- Which one is True?

Individual Assignment – Part III

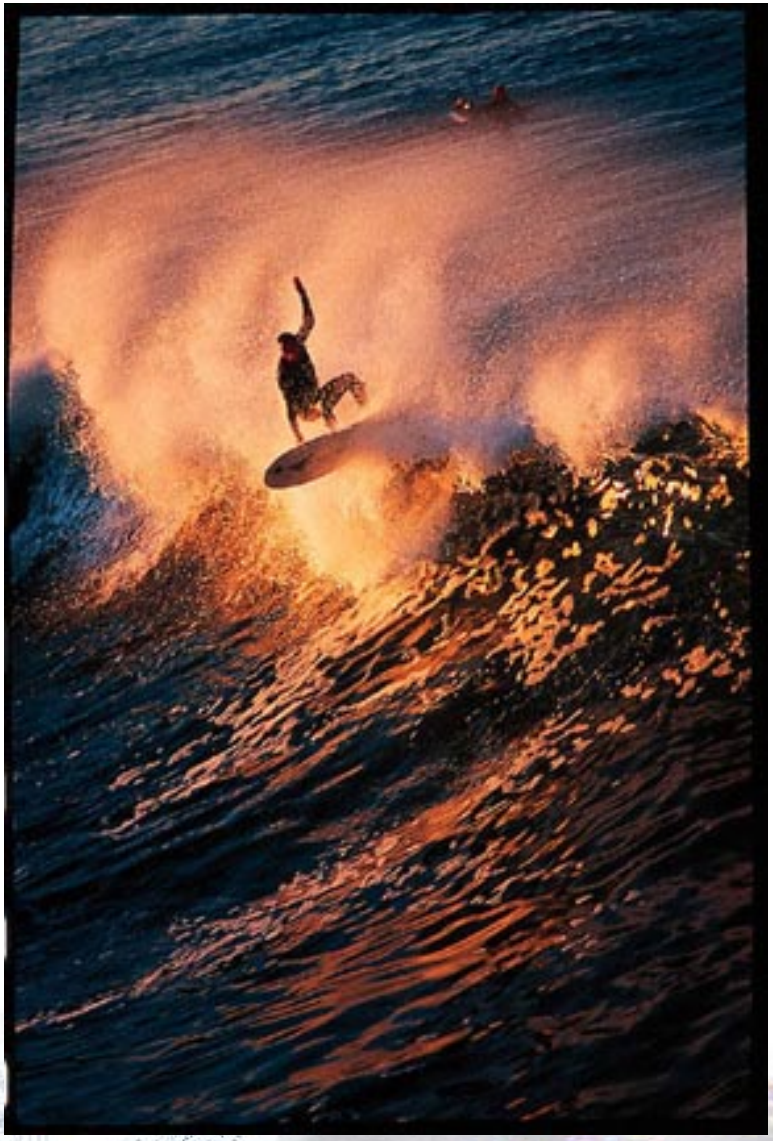


- Step by step analysis of "dying" squares
  - 3rd Installment
    - Presented: March 8th
    - Due: March 30th
  - 4th Installment
    - Presented: April 5th
    - Due: April 20th
- Use descriptive statistics
  - To uncover rules inductively
    - E.g. the behavior of evens and odds, individual numbers, or ranges of cycles, etc.



Next class

- Spring Break



Next Class!

- Topics
  - More Inductive Reasoning Modeling
    - Measures of Dispersion and Position
    - Regression
- Readings for Next week
  - @ *infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 109-134)
- Lab 8
  - Data analysis with Excel (linear regression)
  - NO LAB THIS WEEK!!!!