Analyzing Entropy in Biosurveillance

Initiated as a DHS MSI Funded Summer Program
Continuing under DyDAn Support

Student Researchers:
Nakeya Williams, Morgan State University
Devroy McFarlane, Howard University
Ashley Crump, Howard University
Anthony Ogbuka, Morgan State University

Faculty:
Dr. Abdul-Aziz Yakubu, Howard University
Dr. Asamoah Nkwanta, Morgan State Univ.

DIMACS/DyDAn Project Mentors:
Dr. Nina Fefferman
Dr. Fred Roberts

HOWARD UNIVERSITY

MORGAN

DIMACS
Center for Discrete Mathematics & Theoretical Computer Science
Founded as a National Science Foundation Science and Technology Center

DyDAn
THE HOMELAND SECURITY CENTER FOR DYNAMIC DATA ANALYSIS

Early Detection Of Disease Outbreaks Is Crucial For Public Safety

Failure of biosurveillance increases disease the incidence and mortality

True with all infectious diseases including

- natural exposure from zoonotic infections
- purposeful acts of bioterrorism

Image from <www.alpharubicon.com/basicnbc/basicnbc.htm>

- Smallpox
- Avian influenza
- Rift valley fever
- Brucellosis
- Tularemia
- Anthrax

Image from <microbes.historique.net/anthracis.html>

“We continue to have a great deal of difficulty in determining when outbreaks of infection occur in animals and in humans overseas. Just to be brutally honest, we have a lot of trouble determining when we have an outbreak of infectious disease in a community here in the United States.”

Dr. Rajeev Venkayya

Special Assistant to the President for Biodefense

“Almost every problem that you come across is befuddled with all kinds of extraneous data of one sort or another; and if you can bring this problem down into the main issues, you can see more clearly what you are trying to do and perhaps find a solution.”

C. Shannon

A New Possibility For Biosurveillance: Entropy

Entropy quantifies the amount of information communicated within a signal

Signal strength may change when an outbreak starts

We are hoping to detect changes in signal strength **early** into the onset of an outbreak

Our Ultimate Goal: Effective Biosurveillance

Use Entropy to statistically quantify the differences in the strength of disease incidence signal

Exploit differences for earlier detection of a disease outbreak

Current Methods Of Outbreak Detection Are Hit Or Miss

A frequently used method: CuSum

Compares current cumulative summed incidence to average

– needs a lot of historical “non-outbreak” data
(bad for newly emerging threats)

– has to be manually “reset”

(it’s also bad if new sensitivity of detection is available)

Other methods have similar problems

We need a better method

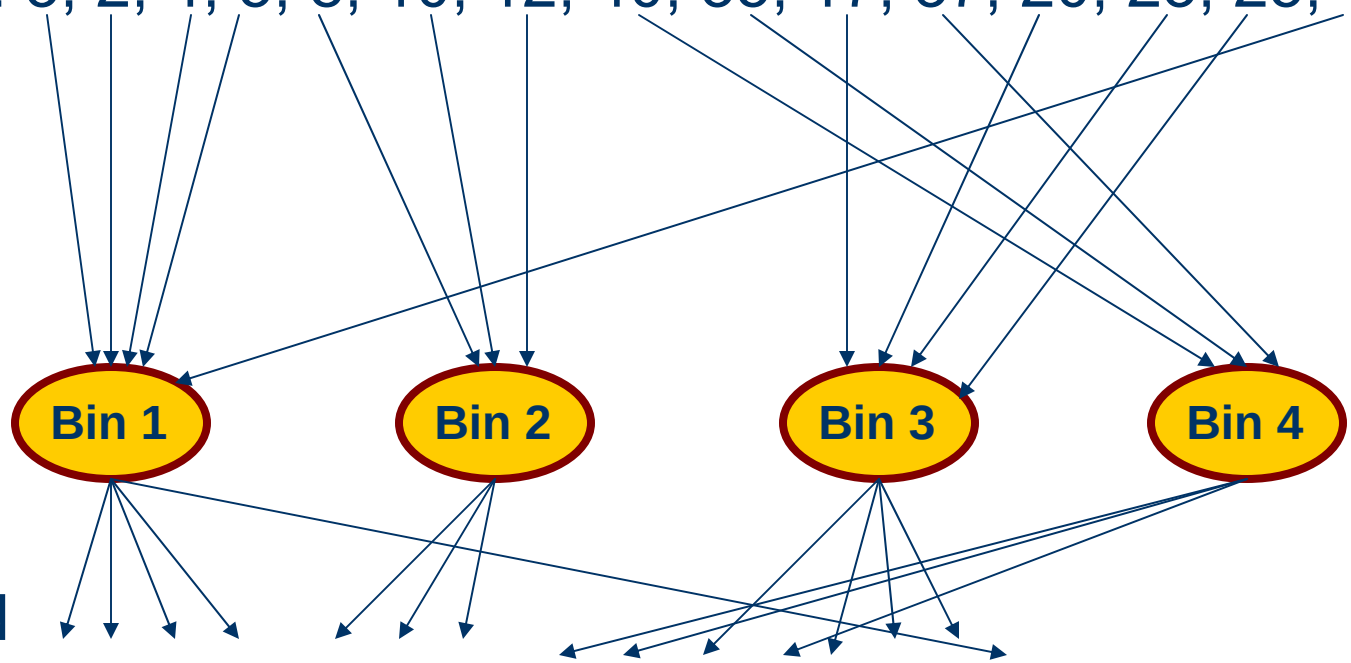
Our Entropy Method Involved the Development of 3 Pre-computational Steps

1) **Binning** the Incidence Data
2) Analyzing within a Temporal **Window**
3) Moving the temporal window according to different **Step Sizes**

Step 1 – Binning

Weekly Disease Incidence

Data: 3, 2, 4, 5, 8, 10, 12, 40, 35, 17, 37, 20, 23, 25, 4,…



Binned

Data: 1, 1, 1, 1, 2, 2, 2, 4, 4, 3, 4, 3, 3, 3, 1

Step 2 & 3 – Window & Step Size

Window Size = 7

Step Size = 1

Incidence Data:

3, 2, 4, 5, 8, 10, 12, 40, 35, 17, 37, 20, 23, 25, 4,…

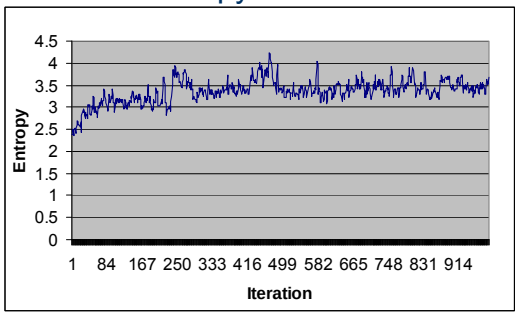
Calculate Entropy

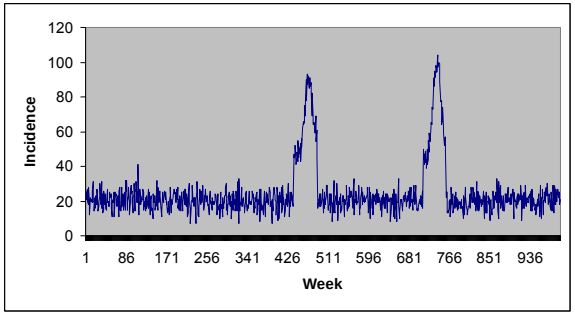
Algorithm

1) Calculate frequency of symbols

2) Calculate entropy

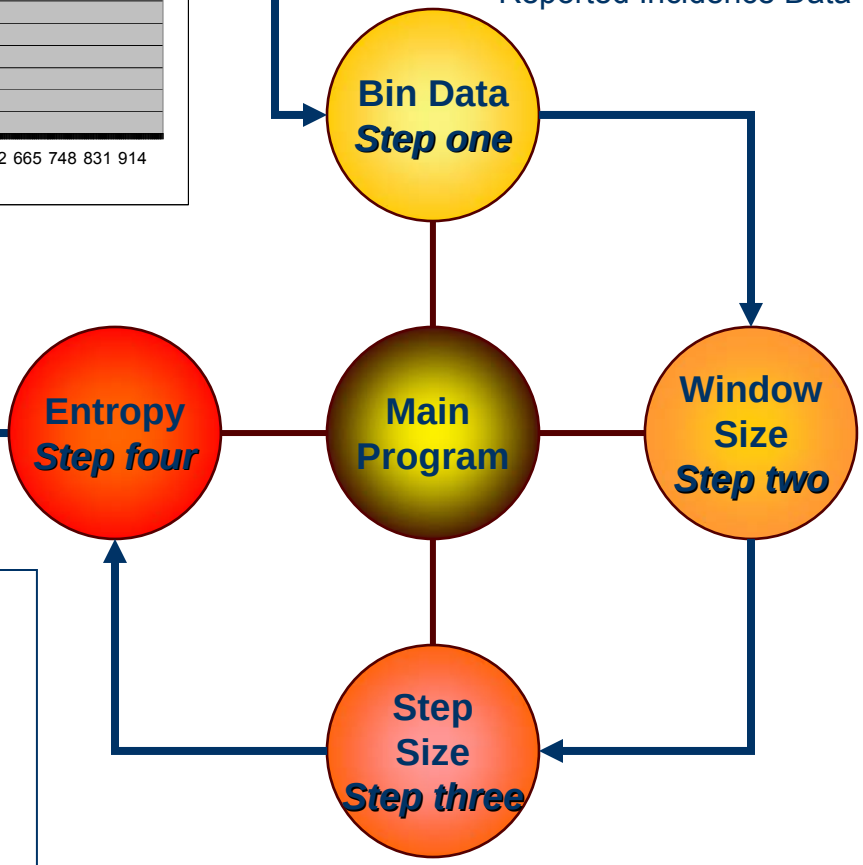
Entropy Outcome





Reported Incidence Data

Repeats Step 4 until data is used up



Entropy Formula:

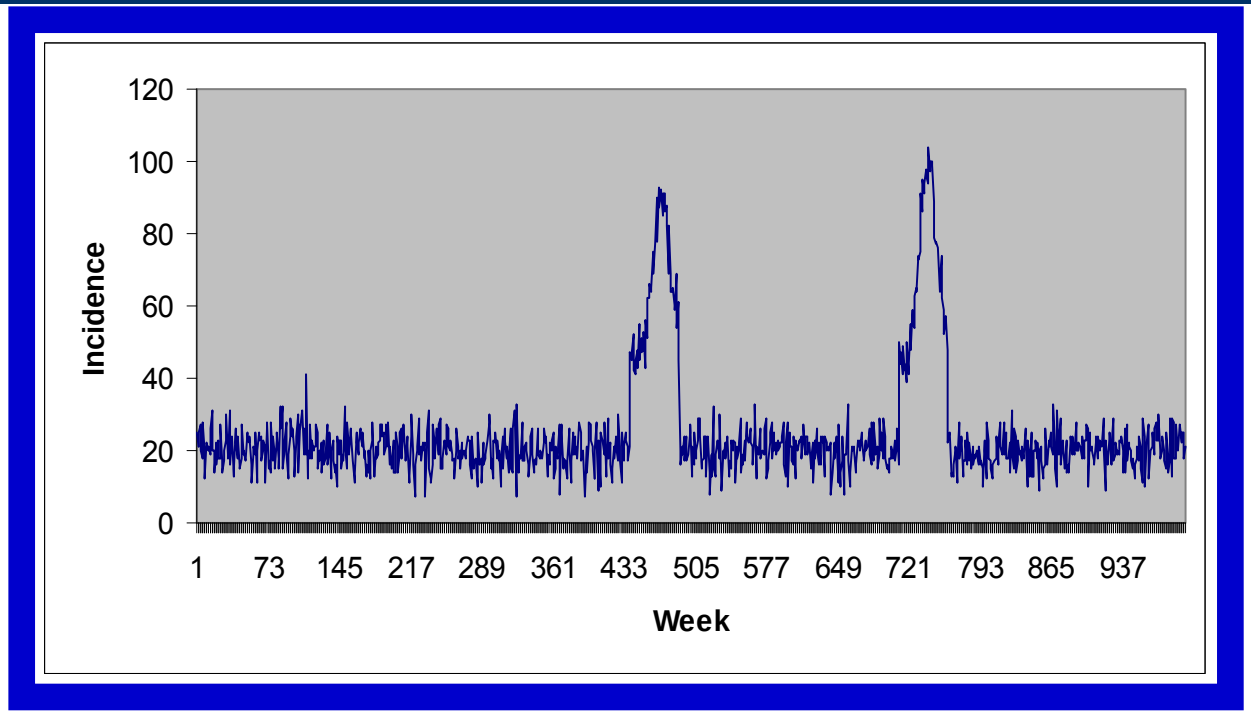
$$H(X) = E(I(X)) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

– $I(X)$ is the information content of X

– $p(x_i) = \Pr(X = x_i)$ is the probability mass function of X

Test Data Set

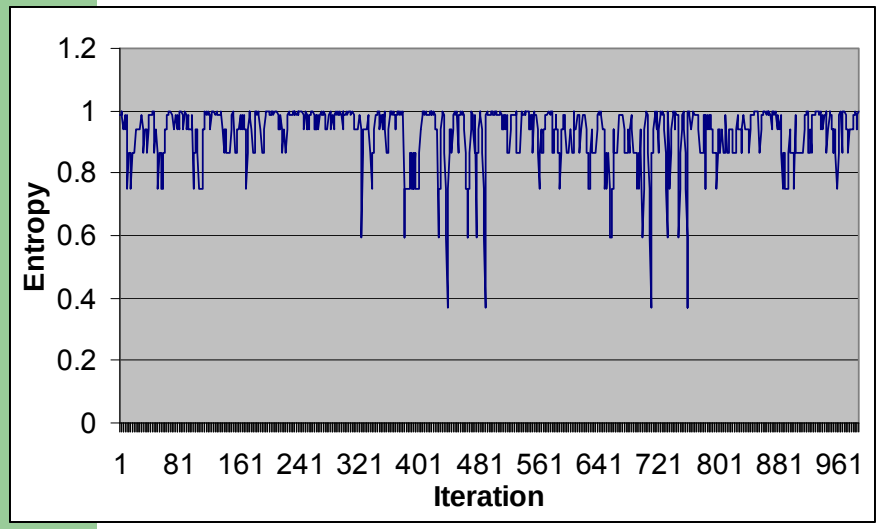
Influenza data from the Blood Research Institute of the BloodCenter of Wisconsin (an NIH funded research center)

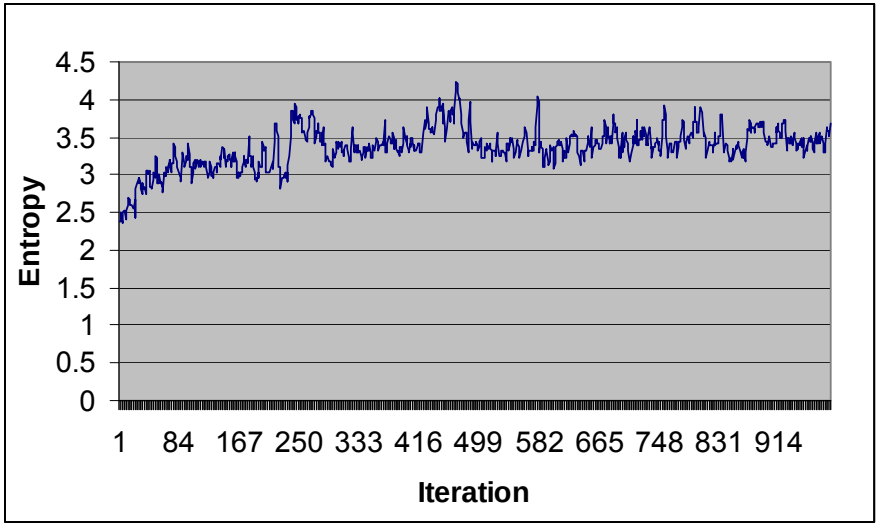


Modified by artificially elongating the non-outbreak intervals to allow for development and testing of entropy algorithm

Step 1: Binning

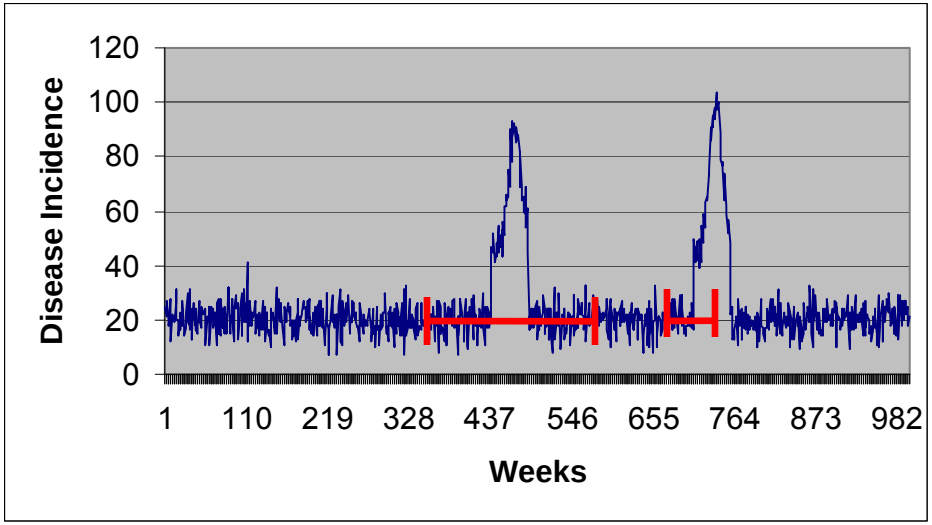
- Bin disease incidence data to minimize small fluctuations
- Pick random number of bins, evenly distribute
- Divide data into intervals then use random number of bins
- Bin dynamically to simulate incoming data

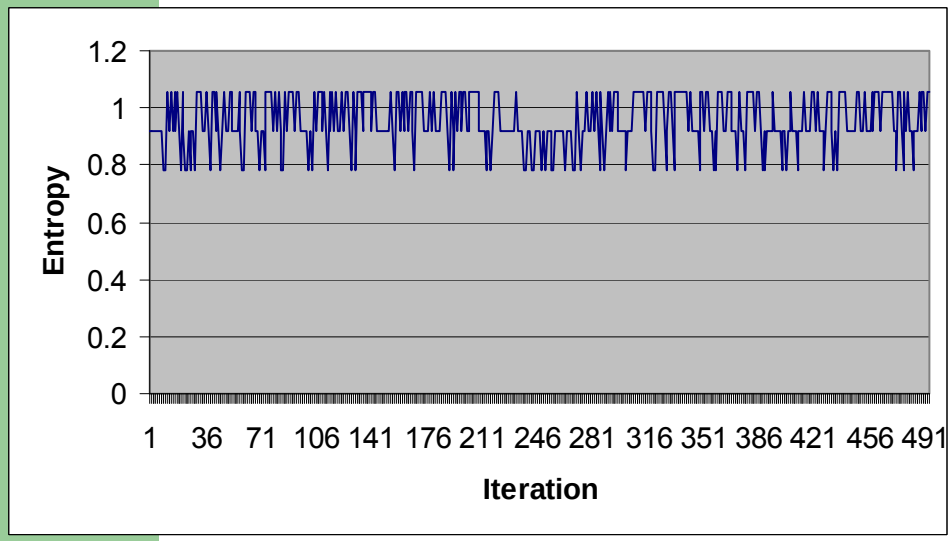


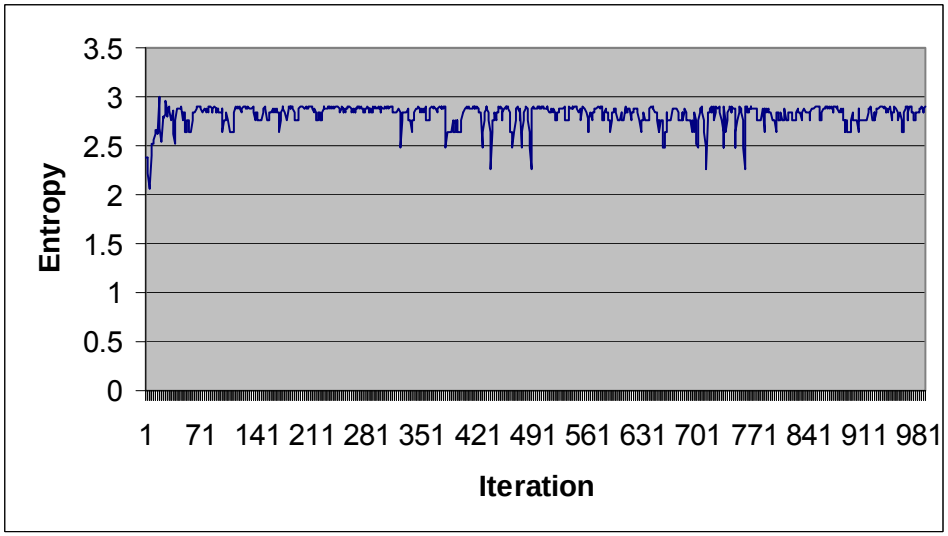


Step 2: Window Size

- The window should be small enough to detect the phase shift from the endemic level to the outbreak.

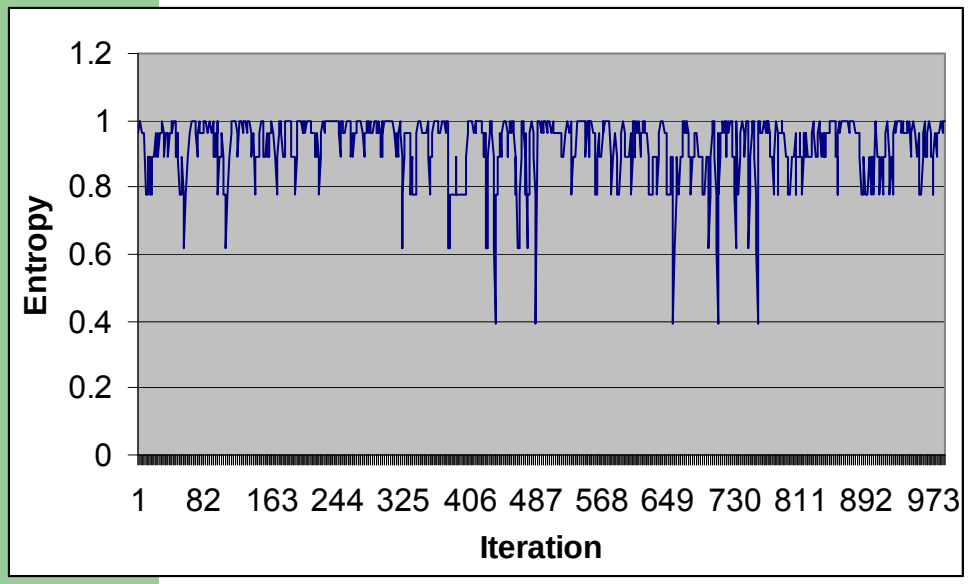


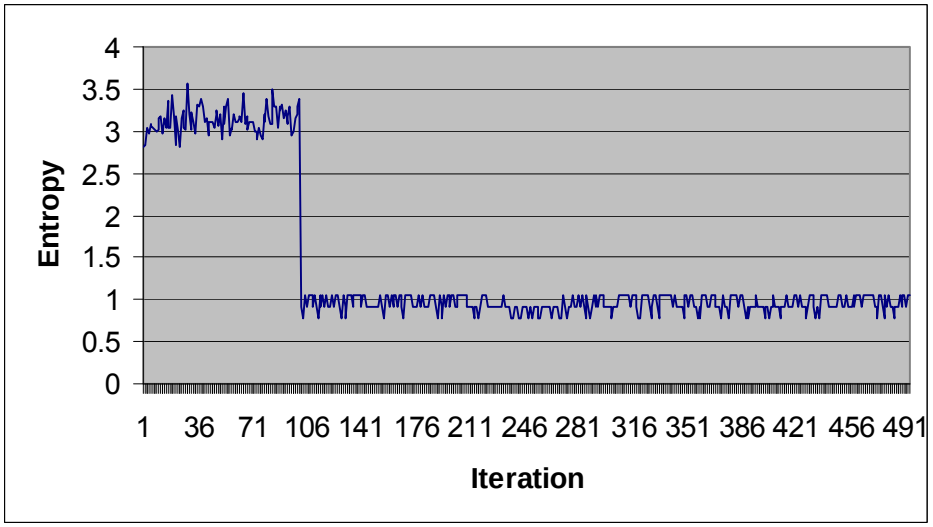




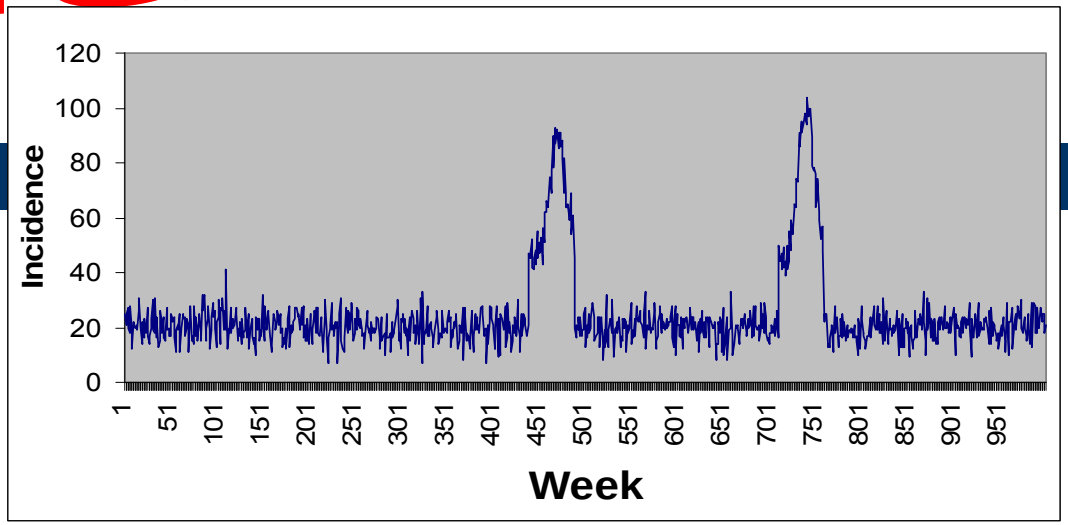
Step 3: Step Size

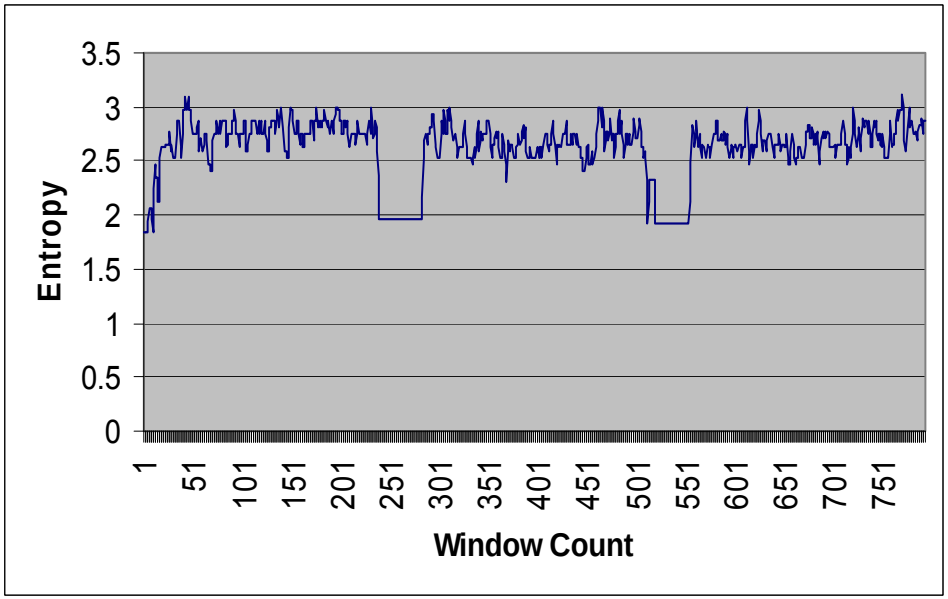
- Adjusting the step size helps eliminate variations in incidence data caused by things like weekends and holidays in daily datasets.





Eureka!!!


Incidence
120
100
80
60
40
20
0
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
801
851
901
951
Week


Entropy
3.5
3
2.5
2
1.5
1
0.5
0
1
51
101
151
201
251
301
351
401
451
501
551
601
651
701
751
Window Count

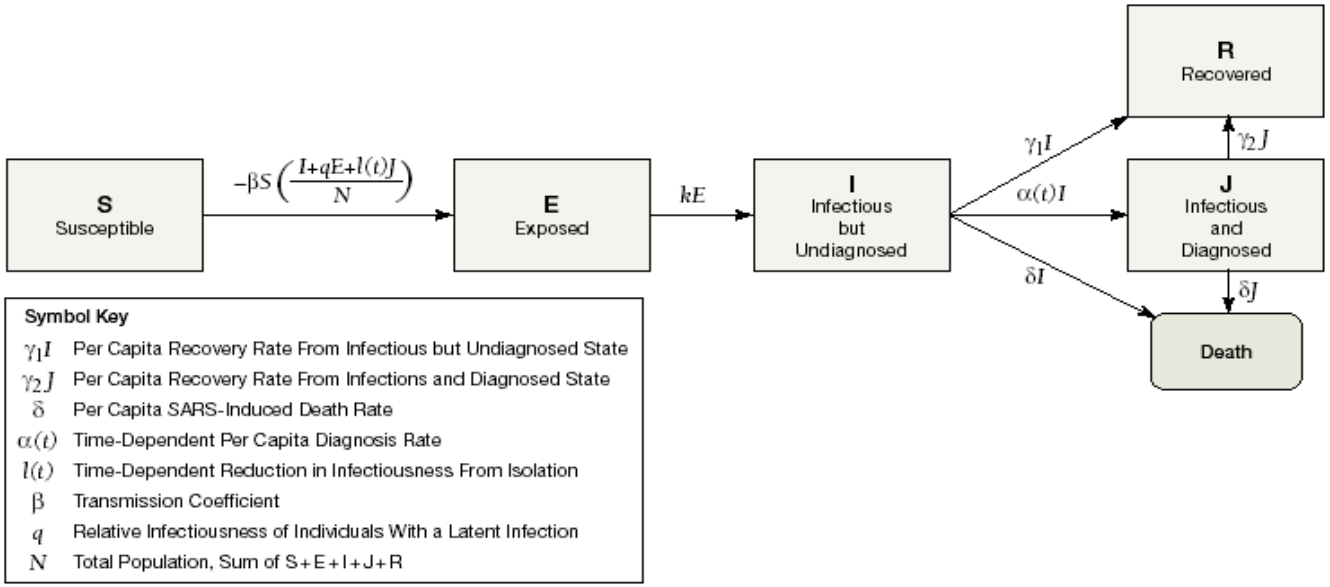
Next steps: Refinements & Extensions

- Impact of binning, window and step sizes on entropy technique
  - Experimental and theoretical work needed
- Surveillance data versus “actual” disease process

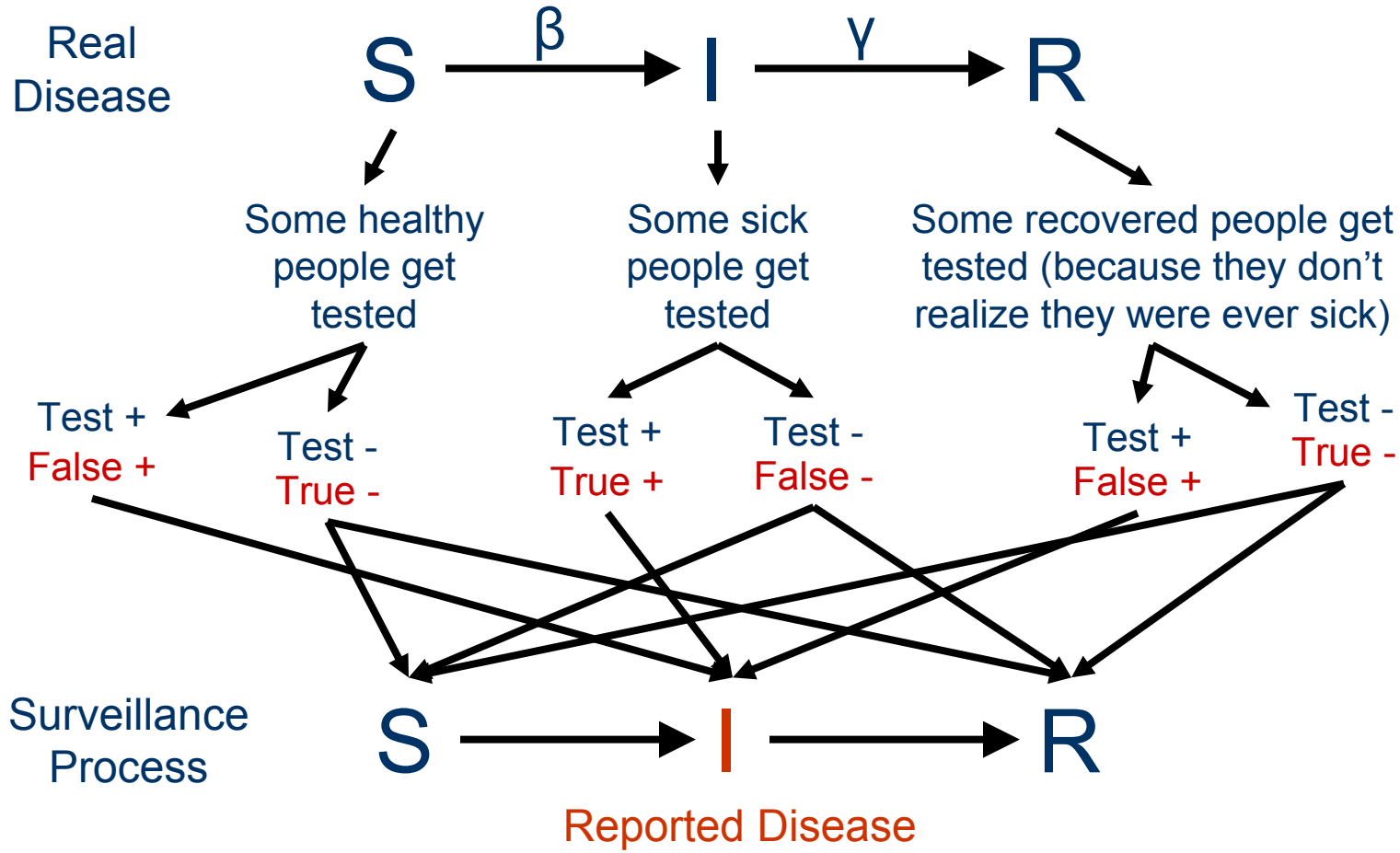
Mathematical Epidemiology

A Mathematical model of SARS showed how Isolation and Quarantine measures could reduce the size of a SARS outbreak by a factor of 1000. The mathematical results agreed with actual observations in the greater Toronto area (JAMA, 2003)

Figure. Box Diagram Illustrating a Mathematical Model of Outcomes of a SARS Epidemic



Surveillance vs. Disease



Related Research Topics:

- Application of entropy to biosurveillance and bioterrorism data.
- Algorithms for calculating and monitoring changes in entropy.
- Entropy aided detection of beginnings of outbreak scenarios.
- Tying in Infectious Disease Models