Learning by Investigation: A Context for Integrating Statistics, Mathematics, and Computation

Jeff Knisley

Mosaic Kickoff – June 30, 2010

East Tennessee State University

The Institute for Mathematics and its Applications

Introduction – The Symbiosis Project

- The Symbiosis Project is an HHMI-funded project to
  - Integrate a freshman Biology lab sequence with Statistics and Calculus
  - Present Biology within a conceptual framework (rather than encyclopaedic)
  - Introduce topics from Informatics, Computational Science, and Discrete Math as needed
- The Symbiosis Project addresses many of the BIO2010 recommendations
- The Symbiosis curriculum addresses many of the AAMC/HHMI competencies expected of medical students once the redesigned MCAT is implemented in 2013 (or 2014)

Three Huge Challenges

1. Maintaining individual integrity of the science, math, stats within the integration
2. Developing essential skills in stats, calculus, and computational science (coding, data types, etcetera)
3. Assessing performance *Without significant grade inflation!*
   - We can't afford (in Symbiosis) for students to enter Calc II, Differential equations, genetics, etcetera where technology-based assignments allowed them to pass without obtaining math/stat/bio skills and concepts.
   - We can't afford (in Symbiosis) for technology-based assignments to be so small a factor grading-wise that students don't take them seriously

Three Huge Challenges

1. Maintaining individual integrity of the science, math, stats within the integration
2. Developing essential skills in stats, calculus, and computational science (coding, data types, etcetera)
3. Assessing performance *Without significant grade inflation!*
   - We can't afford (in Symbiosis) for students to enter Calc II, Differential equations, genetics, etcetera where technology-based assignments allowed them to pass without obtaining math/stat/bio skills and concepts.
   - We can't afford (in Symbiosis) for technology-based assignments to be so small a factor grading-wise that students don't take them seriously

Three Huge Challenges

1. Maintaining individual integrity of the science, math, stats within the integration
2. Developing essential skills in stats, calculus, and computational science (coding, data types, etcetera)
3. Assessing performance *Without significant grade inflation!*
   - We can't afford (in Symbiosis) for students to enter Calc II, Differential equations, genetics, etcetera where technology-based assignments allowed them to pass without obtaining math/stat/bio skills and concepts.
   - We can't afford (in Symbiosis) for technology-based assignments to be so small a factor grading-wise that students don't take them seriously

Three Huge Challenges

1. Maintaining individual integrity of the science, math, stats within the integration
2. Developing essential skills in stats, calculus, and computational science (coding, data types, etcetera)
3. Assessing performance *Without significant grade inflation!*
   - We can't afford (in Symbiosis) for students to enter Calc II, Differential equations, genetics, etcetera where technology-based assignments allowed them to pass without obtaining math/stat/bio skills and concepts.
   - We can't afford (in Symbiosis) for technology-based assignments to be so small a factor grading-wise that students don't take them seriously

Three Huge Challenges

1. Maintaining individual integrity of the science, math, stats within the integration
2. Developing essential skills in stats, calculus, and computational science (coding, data types, etcetera)
3. Assessing performance *Without significant grade inflation!*
   - We can't afford (in Symbiosis) for students to enter Calc II, Differential equations, genetics, etcetera where technology-based assignments allowed them to pass without obtaining math/stat/bio skills and concepts.
   - We can't afford (in Symbiosis) for technology-based assignments to be so small a factor grading-wise that students don't take them seriously

Investigative Case-Based Learning (ICBL)

- From BioQuest: http://bioquest.org/icbl/
  - Investigative Case-Based Learning ... encourages students to develop questions that can be explored further by reasonable investigative approaches.
  - Students employ a variety of methods and resources, including ... software simulations and models, data sets, internet-based tools and information retrieval methods.
- From Stanley, Waterman, 2005: (http://serc.carleton.edu/introgeo/icbl/)
  - Cases serve as springboards to student-designed investigations.
  - Students structure their own learning using the "story" of the case as a problem space.
  - Cases engage students and faculty in collaborative ... problem solving.

ICBL's transformed

- Problem with Sciences oriented ICBL's w.r.t Math/Stats/Computation
  - Cases may not be good contexts for math, stats, or computation
  - Open-ended questions to develop Calculus skills???
- Goal: Implement desirable features of ICBL's into a Math/Stat/Computation format
  - Students like the "story"
  - Collaboration, "ownership" of problems, etcetera
- *Computation itself can serve as a context for both Math and Stats!*

ICBL's transformed

- Problem with Sciences oriented ICBL's w.r.t Math/Stats/Computation
  - Cases may not be good contexts for math, stats, or computation
    - Open-ended questions to develop Calculus skills???
- Goal: Implement desirable features of ICBL's into a Math/Stat/Computation format
  - Students like the "story"
  - Collaboration, "ownership" of problems, etcetera
- *Computation itself can serve as a context for both Math and Stats!*

ICBL's transformed

- Problem with Sciences oriented ICBL's w.r.t Math/Stats/Computation
  - Cases may not be good contexts for math, stats, or computation
  - Open-ended questions to develop Calculus skills???
- Goal: Implement desirable features of ICBL's into a Math/Stat/Computation format
  - Students like the "story"
  - Collaboration, "ownership" of problems, etcetera
- *Computation itself can serve as a context for both Math and Stats!*

ICBL's transformed

- Problem with Sciences oriented ICBL's w.r.t Math/Stats/Computation
  - Cases may not be good contexts for math, stats, or computation
  - Open-ended questions to develop Calculus skills???
- Goal: Implement desirable features of ICBL's into a Math/Stat/Computation format
  - Students like the "story"
  - Collaboration, "ownership" of problems, etcetera
- *Computation itself can serve as a context for both Math and Stats!*

ICBL's transformed

- Problem with Sciences oriented ICBL's w.r.t Math/Stats/Computation
  - Cases may not be good contexts for math, stats, or computation
  - Open-ended questions to develop Calculus skills???
- Goal: Implement desirable features of ICBL's into a Math/Stat/Computation format
  - Students like the "story"
  - Collaboration, "ownership" of problems, etcetera
- *Computation itself can serve as a context for both Math and Stats!*

Investigations

An investigation is an ICBL whose context is computational that is, choosing "cases" that occur naturally as simulations, approximations, bootstraps, resampling, etcetera. The computational context subsequently allows/requires both mathematical and statistical exploration, providing the integration with each other and the science.

(Possible) Steps to Creating an Investigation

Step 1: A (scientific) story that motivates a computational approach

Step 2: An Overarching Question that utilizes the computational context

Step 3: Models built mathematically within/related to computational context

Step 4: Analysis via statistical tests, methods of the (simulated) data

Example: You Be Walter Reed!

We use the following to start the Symbiosis Project:
The goal is to motivate/discuss the importance of the Scientific method and evidence-based science via data analysis.

At the end of the 19th century, yellow fever was known as "the American Plague." Thousands died in epidemics in Memphis, New Orleans, and Philadelphia, to name a few. During the Spanish-American war, thirteen times more soldiers died from yellow fever than from combat, and yellow fever eventually drove the US out of Cuba – 75% of the soldiers were unfit to serve by the time of the withdrawal. Although they never succeeded in stopping any epidemics, policy-makers in 1900 were convinced – and had "evidence" – that the key to preventing yellow fever was sanitation, sterilization, and quarantine.

Example: You Be Walter Reed!

We use the following to start the Symbiosis Project:
The goal is to motivate/discuss the importance of the Scientific method and evidence-based science via data analysis.

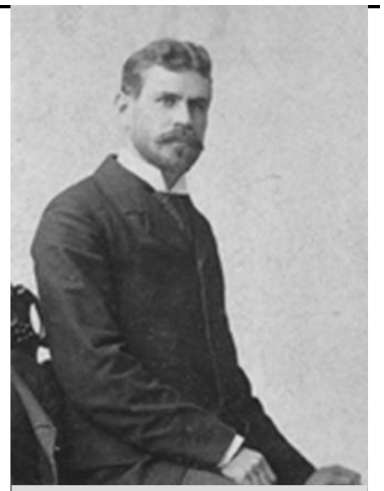
> *At the end of the 19th century, yellow fever was known as "the American Plague." Thousands died in epidemics in Memphis, New Orleans, and Philadelphia, to name a few. During the Spanish-American war, thirteen times more soldiers died from yellow fever than from combat, and yellow fever eventually drove the US out of Cuba – 75% of the soldiers were unfit to serve by the time of the withdrawal. Although they never succeeded in stopping any epidemics, policy-makers in 1900 were convinced – and had "evidence" – that the key to preventing yellow fever was sanitation, sterilization, and quarantine.*

Example: You Be Walter Reed!

In 1881, the Cuban physician Carlos Finlay discovered that yellow fever was spread by Mosquito bite, but no one believed him. In 1901, Walter Reed showed that Finlay was right – although only after volunteer Lazear's death. However, policy makers remained unconvinced. Only after an epidemic in New Orleans in 1905 did Congress and other Authorities finally accept Reed's/Finlay's scientific results (note: Reed died in 1904).

Finlay, seated on left, with Havana public health experts

Major Walter Reed



Jesse Lazear

Example: You Be Walter Reed!

Now, 100 years later, the question surfaces in the mid 1990's, "Can the AIDS virus be spread by mosquito?" How do you answer that question and in so doing convince policy makers that you are right?

- Faculty questions/comments/starter points
  - You can't kill anyone!!
    - Medical Ethics prevents what the Reed commission did.
    - Clara Maas and Jesse Lazear signed the first ever "informed consent" forms
  - *Question: How does a mosquito-borne epidemic differs from a "poor sanitation caused" epidemic*

Example: You Be Walter Reed!

Now, 100 years later, the question surfaces in the mid 1990's, "Can the AIDS virus be spread by mosquito?" How do you answer that question and in so doing convince policy makers that you are right?

- Faculty questions/comments/starter points
  - You can't kill anyone!!
    - Medical Ethics prevents what the Reed commission did.
    - Clara Maas and Jesse Lazear signed the first ever "informed consent" forms
  - *Question: How does a mosquito-borne epidemic differs from a "poor sanitation caused" epidemic*

Example: You Be Walter Reed!

Now, 100 years later, the question surfaces in the mid 1990's, "Can the AIDS virus be spread by mosquito?" How do you answer that question and in so doing convince policy makers that you are right?

- Faculty questions/comments/starter points
  - You can't kill anyone!!
    - Medical Ethics prevents what the Reed commission did.
    - Clara Maas and Jesse Lazear signed the first ever "informed consent" forms
    - *Question: How does a mosquito-borne epidemic differs from a "poor sanitation caused" epidemic*

Example: You Be Walter Reed!

Now, 100 years later, the question surfaces in the mid 1990's, "Can the AIDS virus be spread by mosquito?" How do you answer that question and in so doing convince policy makers that you are right?

- Faculty questions/comments/starter points
  - You can't kill anyone!!
    - Medical Ethics prevents what the Reed commission did.
    - Clara Maas and Jesse Lazear signed the first ever "informed consent" forms
  - *Question: How does a mosquito-borne epidemic differs from a "poor sanitation caused" epidemic*

Example: You Be Walter Reed!

Now, 100 years later, the question surfaces in the mid 1990's, "Can the AIDS virus be spread by mosquito?" How do you answer that question and in so doing convince policy makers that you are right?

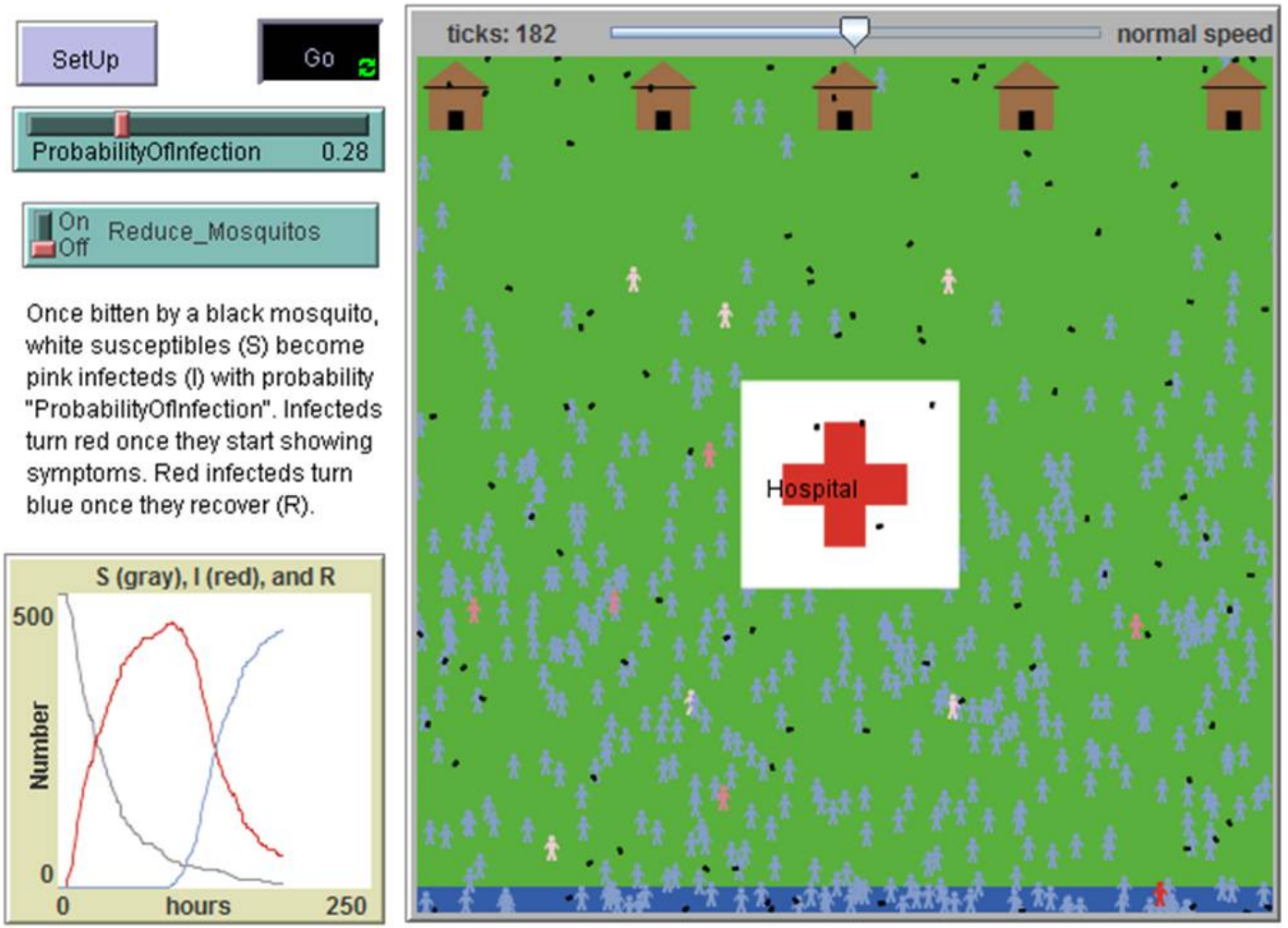
- Faculty questions/comments/starter points
  - You can't kill anyone!!
    - Medical Ethics prevents what the Reed commission did.
    - Clara Maas and Jesse Lazear signed the first ever "informed consent" forms
  - *Question: How does a mosquito-borne epidemic differs from a "poor sanitation caused" epidemic*

Example: You Be Walter Reed!

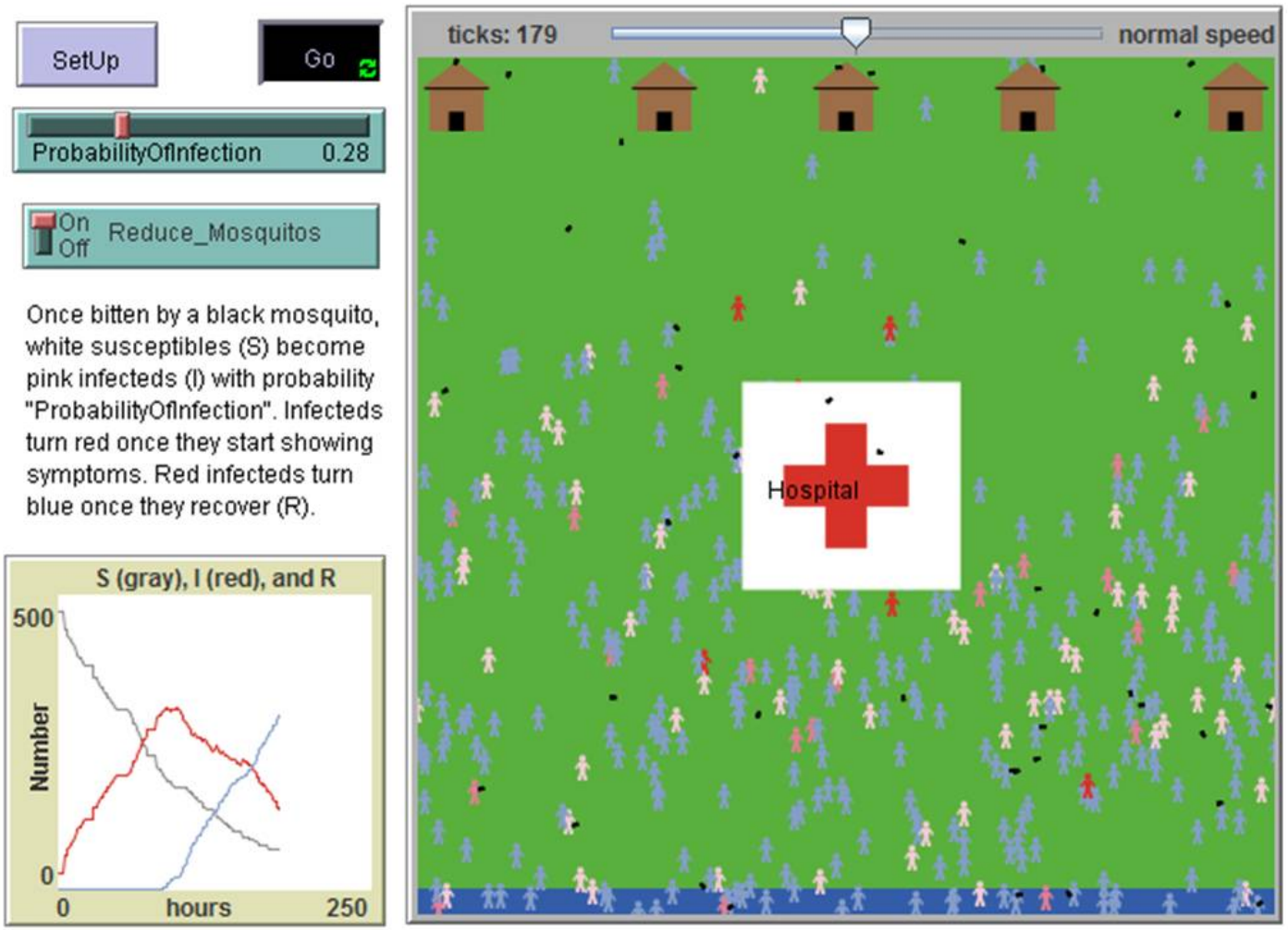
Final Context: Epidemic Simulations

Although the traditional SIR model is a decent model of diseases spread by human-to-human contact, diseases spread by insects, tainted food, poor sanitation, or polluted water all require slightly different models. Can we tell the difference between different infectious agent models using data drawn from various implementations of the epidemics? Can we detect when a particular intervention is having a significant effect independent of knowing the cause of the disease?

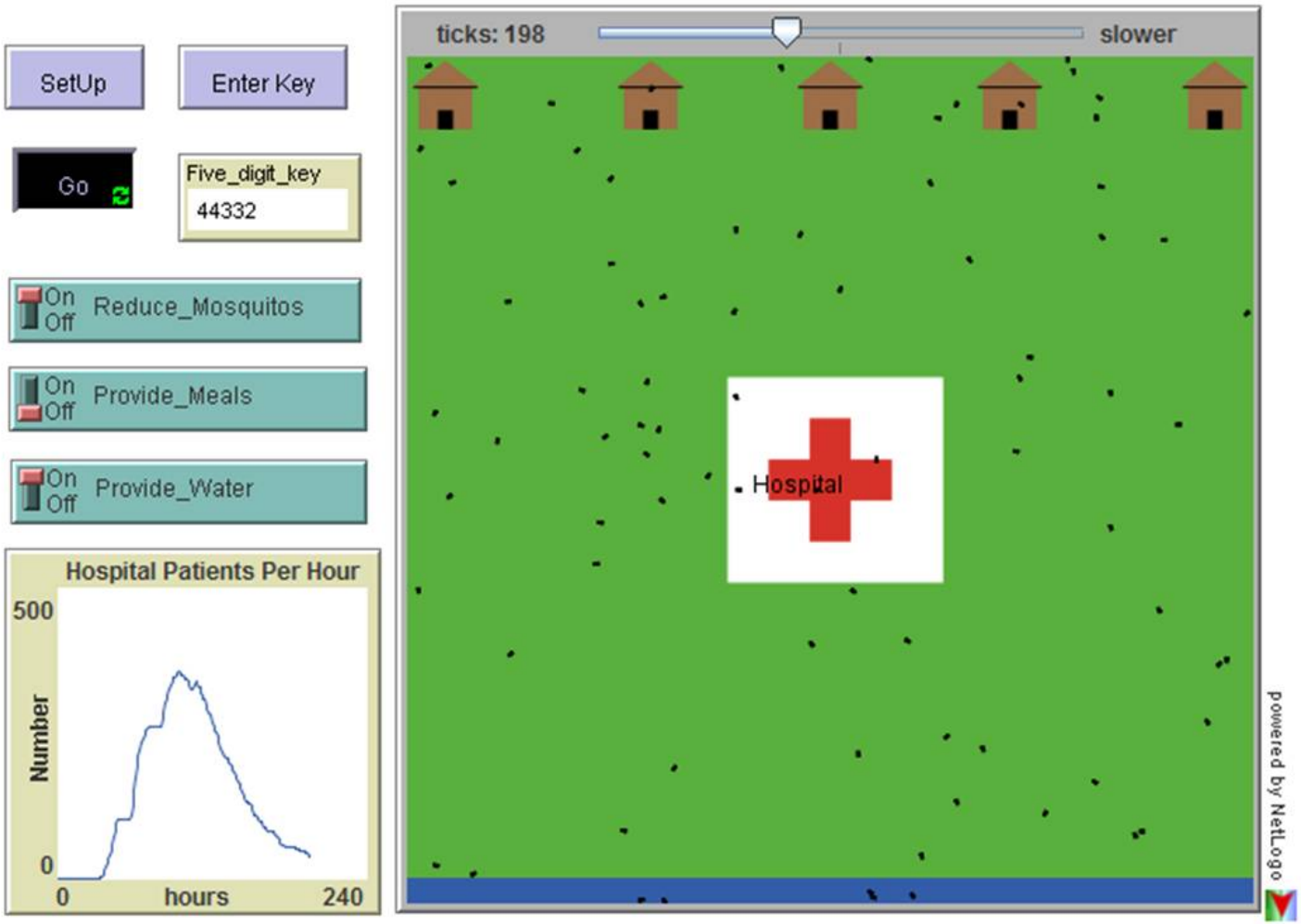
Mosquito-Borne Disease Epidemic



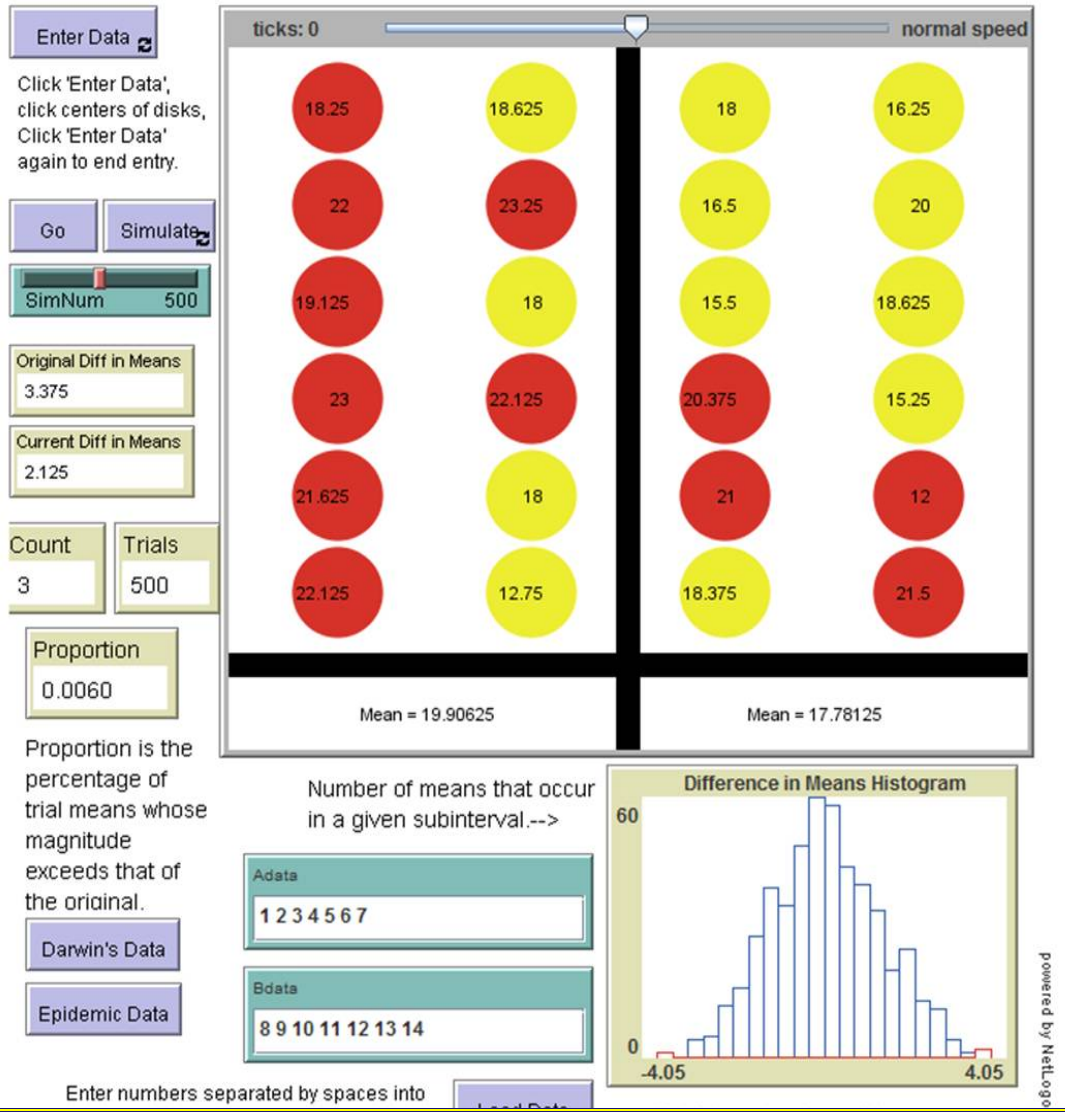
Mosquito-Borne Disease Epidemic



A Simple Epidemic Investigation



The Randomization Test



ICBL versus Investigation

- What Example looks like as an ICBL
  - Students tend to debate and persuade
  - They don't have any real "experiments" they can do
- But as an "Investigation"
  - Case + Data via applet/CAS/StatTool
  - We noticed early on that our Case study + Technology efforts were doing as much for the math and stats as for the biology.
  - Technology replaces the "word problem approach" in getting students to connect math to "the real world"

ICBL versus Investigation

- What Example looks like as an ICBL
  - Students tend to debate and persuade
  - They don't have any real "experiments" they can do
- But as an "Investigation"
  - Case + Data via applet/CAS/StatTool
  - We noticed early on that our Case study + Technology efforts were doing as much for the math and stats as for the biology.
  - Technology replaces the "word problem approach" in getting students to connect math to "the real world"

Investigations

- Motivation came from the Biologists
  - Exploration and Experiment – i.e., "crunching the numbers" – is both a traditional and an effective means that Scientists use to learn mathematical concepts and how to apply them
    - Biologists like the Lotka-Volterra predator-prey model even though it is not "biologically realistic"
    - Cellular Automata – e.g., the Game of Life – are considered important scientific tools
  - But assessment is still an issue!!
- Technology (computation) is a natural context for integrating mathematics and statistics

Investigations

- Motivation came from the Biologists
  - Exploration and Experiment – i.e., "crunching the numbers" – is both a traditional and an effective means that Scientists use to learn mathematical concepts and how to apply them
    - Biologists like the Lotka-Volterra predator-prey model even though it is not "biologically realistic"
    - Cellular Automata – e.g., the Game of Life – are considered important scientific tools
  - But assessment is still an issue!!
- Technology (computation) is a natural context for integrating mathematics and statistics

Investigations

- Motivation came from the Biologists
  - Exploration and Experiment – i.e., "crunching the numbers" – is both a traditional and an effective means that Scientists use to learn mathematical concepts and how to apply them
    - Biologists like the Lotka-Volterra predator-prey model even though it is not "biologically realistic"
    - Cellular Automata – e.g., the Game of Life – are considered important scientific tools
  - But assessment is still an issue!!
- Technology (computation) is a natural context for integrating mathematics and statistics

Investigations as "Word Problems"

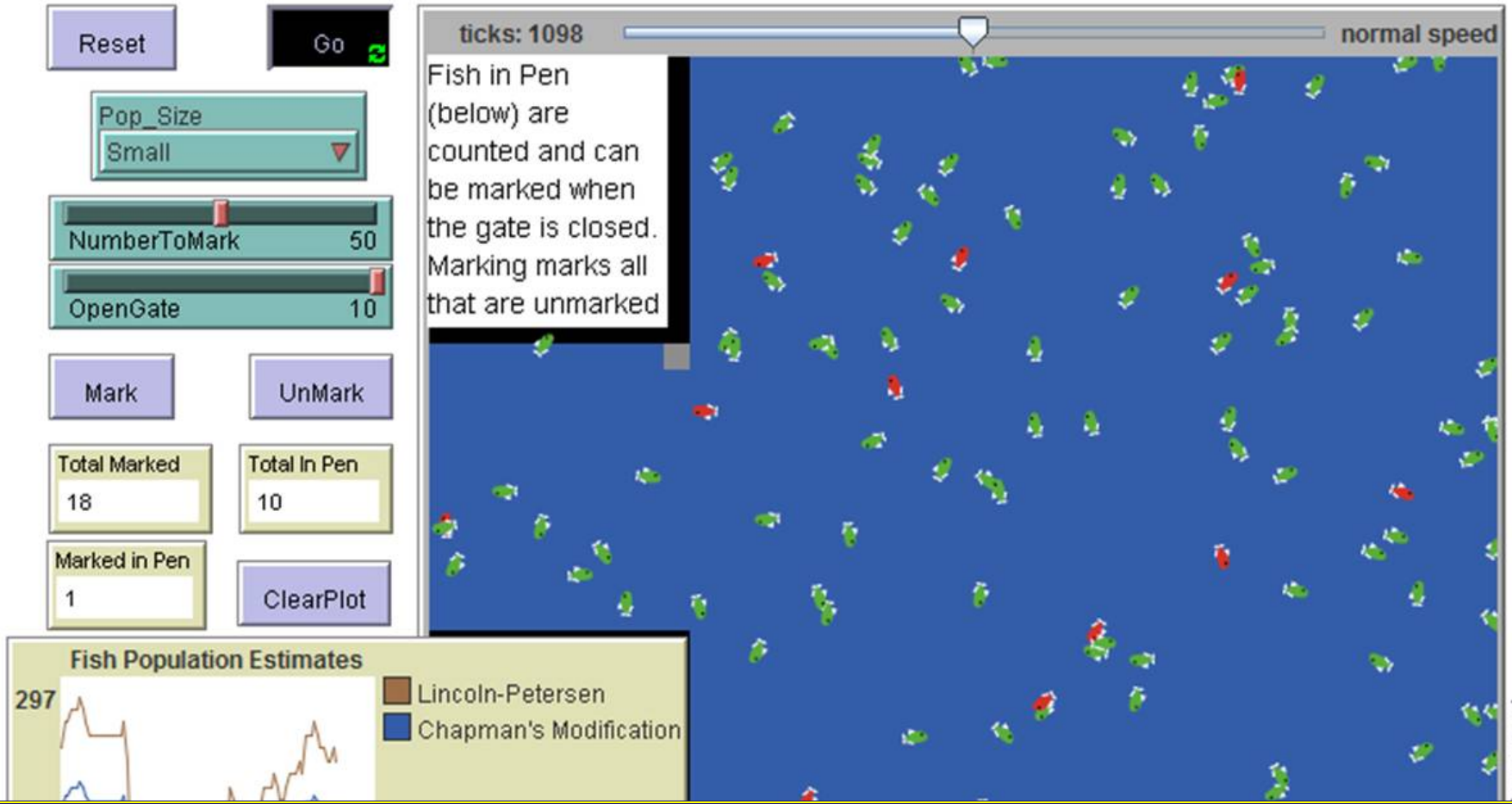
Key: Creating an investigation is analogous to writing a word problem.

1. Begin with a story that requires mathematics/modeling
2. Tell the story within a context of data / simulations / other computations
3. Implement the story outcome in a computational context – data exploration, simulations, etcetera
4. The technology context will thus require the mathematical skill/concept that motivated the investigation initially.

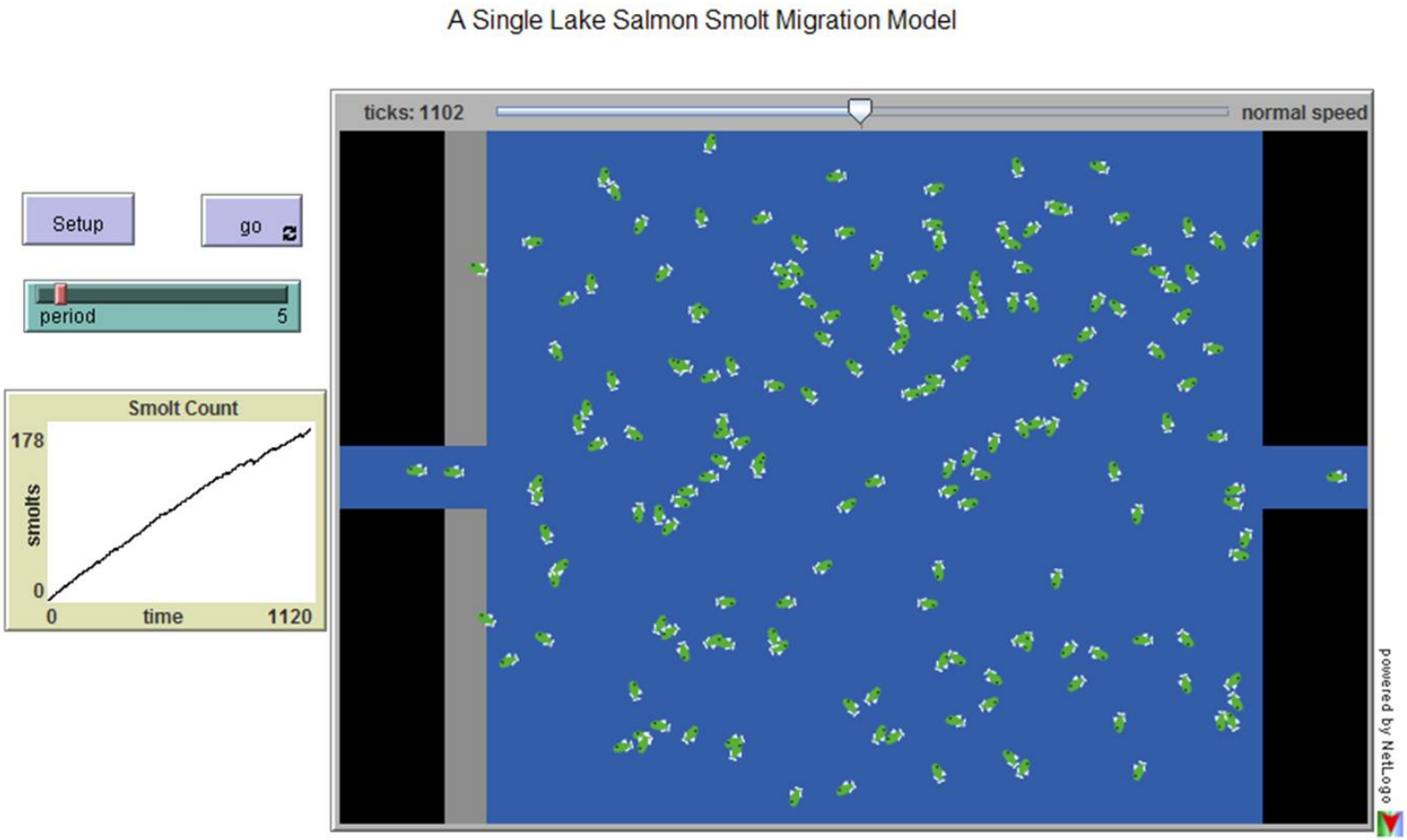
Types of Investigations

- Investigations can be as small as a "test problem"
  - Example: Mark Recapture
  - Example: t-test
- Or as large as a semester-long project
  - Salmon smolt migration that uniformly randomly fills a lake is not "biologically realistic"
    - Salmon don't stay in the lake that long
    - They learn to move through a lake sensing flow toward the lake exit
  - Analyzing a game to determine if first player has an advantage

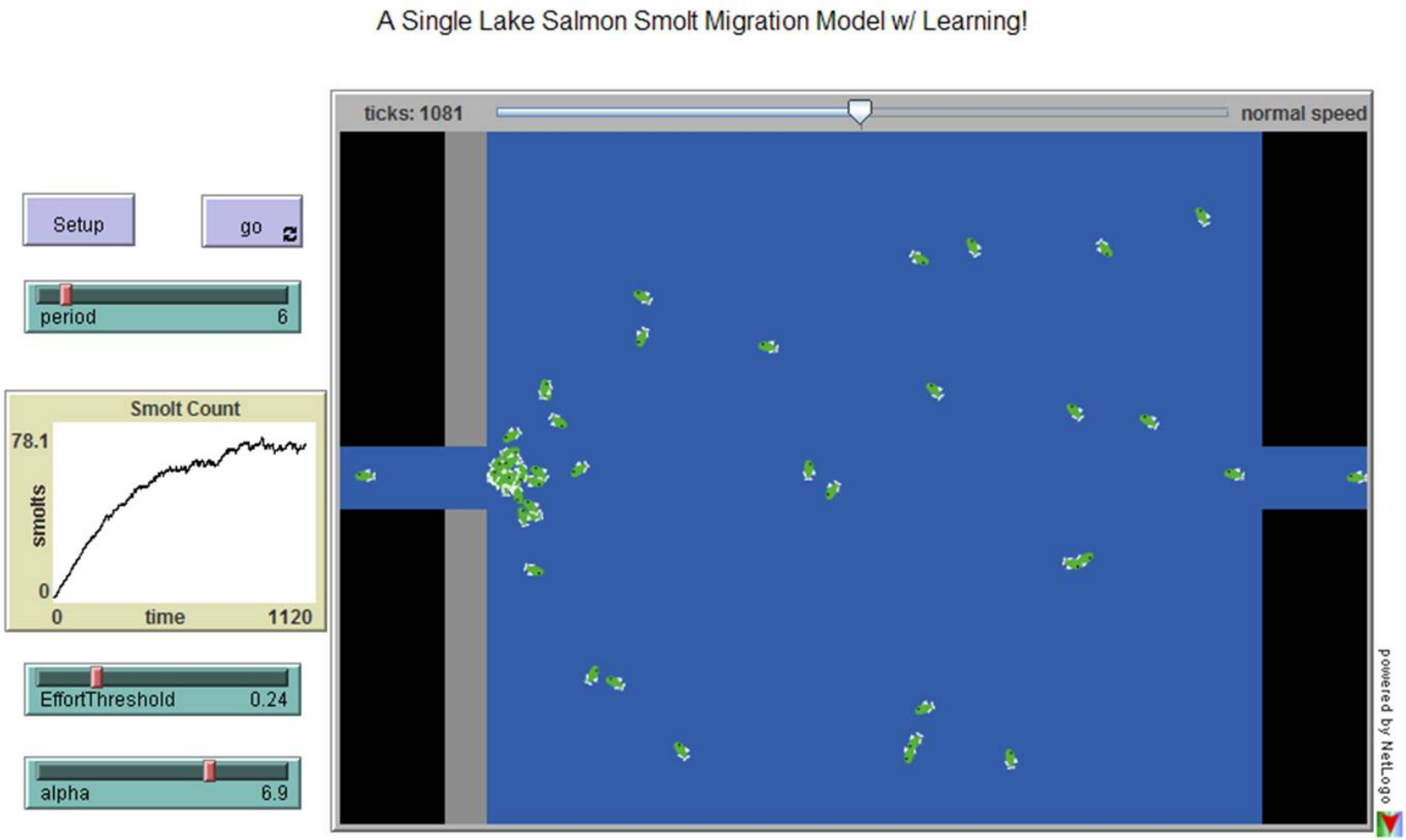
Mark Recapture



Smolts with Uniformly Random Migration



Smolts who learn to find their way



Motivating Stories + Questions

Overarching Question: How would we show that people tend to get shorter as they get older?
Computational Context: Use Data to answer the question "How do humans armspans tend to scale, on average, with their height?"

Overarching Question: How would we use drug dosing for animal models to predict appropriate doses for humans?
Computational Context: Exploring data in hopes of predicting power law relationships

Example: Designing an Investigation

Concept to be Assessed: Linear Regression, including the concept of residuals, correlation coefficient, minimizing total squared error, and regression toward the mean.

The Word Problem: Two parts of an animals body are said to *grow isometrically* if their respectives lengths remain in proportion to one another as the organism increases in size. Describe the isometric relationship between shoe size and height if there is a 1 size increase for every two inches in increased height.

As An Investigation: Give the student several data sets containing length-length data over a population sample that features many different sizes (i.e., stages of development). Illustrate how linear regression can be used to predict an isometry and can provide evidence for the validity of the regression model. Then they must do the same!

More Examples

Intuition tells us that 100 flips of a fair coin should produce about 50 "heads."

Yet suppose the coin is bent slightly (figure 4.0.1). Is the coin still fair? How would we find out? Why should we care?

And what if the coin is "spun" instead of flipped. Or if stacks of coins are used instead?

More Examples

Intuition tells us that 100 flips of a fair coin should produce about 50 "heads."

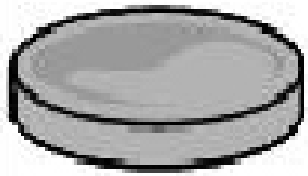
Yet suppose the coin is bent slightly (figure 4.0.1). Is the coin still fair? How would we find out? Why should we care?

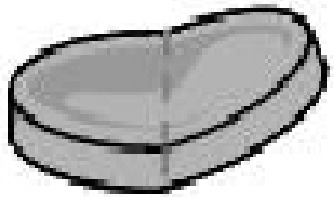
And what if the coin is "spun" instead of flipped. Or if stacks of coins are used instead?

More Examples

Intuition tells us that 100 flips of a fair coin should produce about 50 "heads."

Yet suppose the coin is bent slightly (figure 4.0.1). Is the coin still fair? How would we find out? Why should we care?





And what if the coin is "spun" instead of flipped. Or if stacks of coins are used instead?

More Examples

Intuition tells us that 100 flips of a fair coin should produce about 50 "heads."

Yet suppose the coin is bent slightly (figure 4.0.1). Is the coin still fair? How would we find out? Why should we care?

And what if the coin is "spun" instead of flipped. Or if stacks of coins are used instead?

More Examples

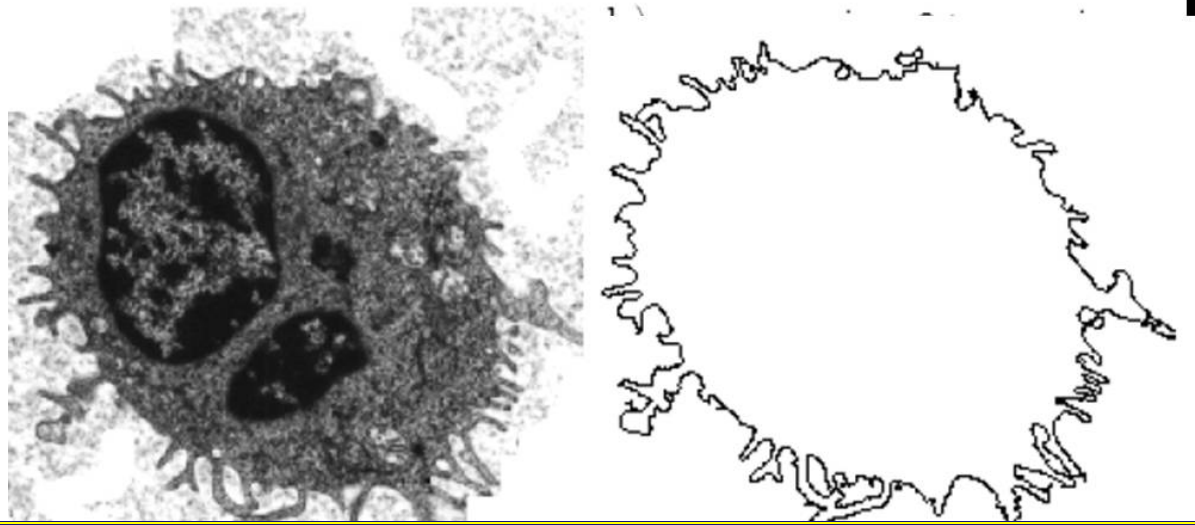
Intuition tells us that 100 flips of a fair coin should produce about 50 "heads."

Yet suppose the coin is bent slightly (figure 4.0.1). Is the coin still fair? How would we find out? Why should we care?

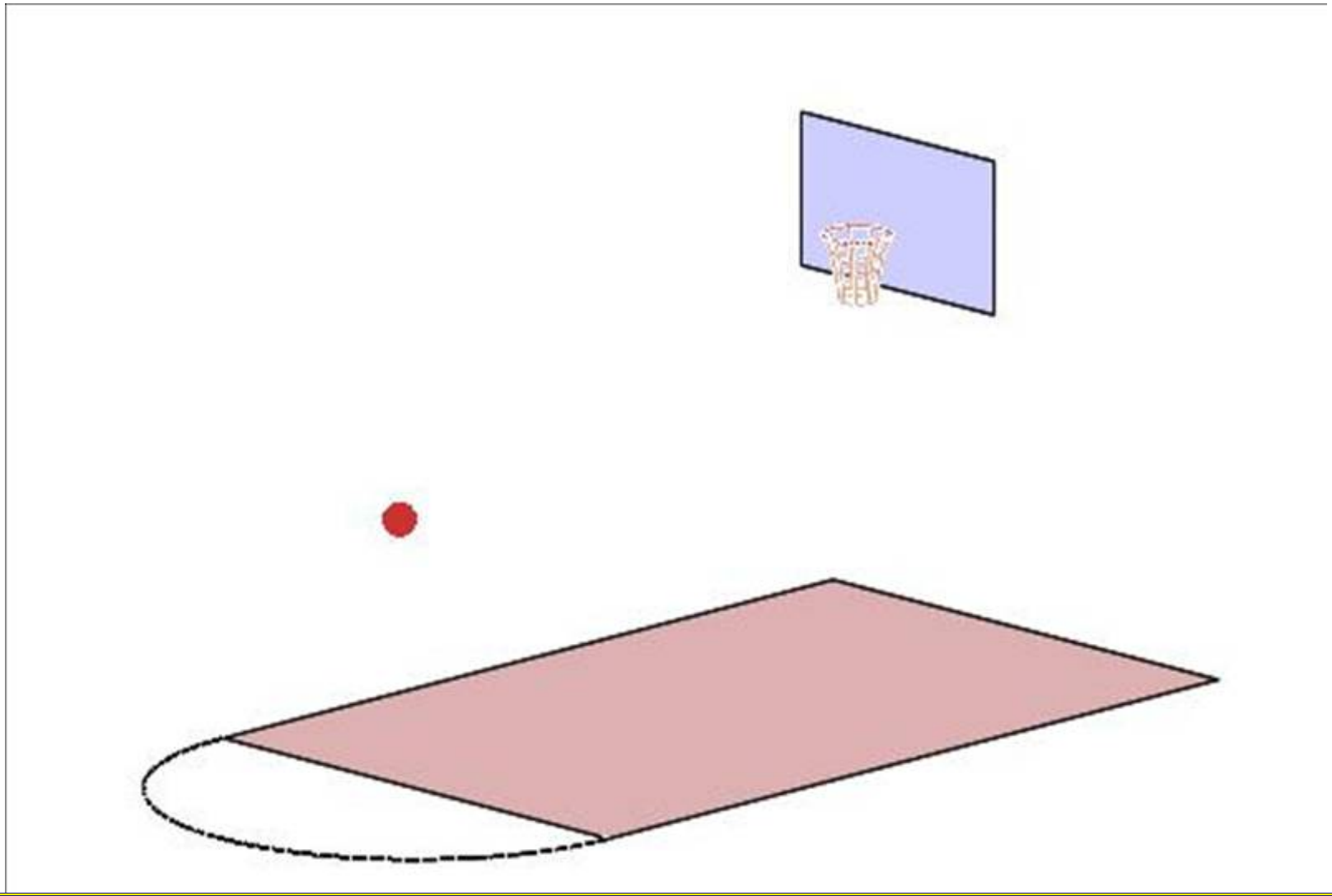
And what if the coin is "spun" instead of flipped. Or if stacks of coins are used instead?

More Examples

"We utilize the fractal dimension of the perimeter surface of cell sections .... to distinguish cancerous from healthy cells .... As a first application we show that it is possible to perform this distinction between patients with hairy-cell lymphocytic leukemia and those with normal blood lymphocytes." Bauer and MacKenzie,



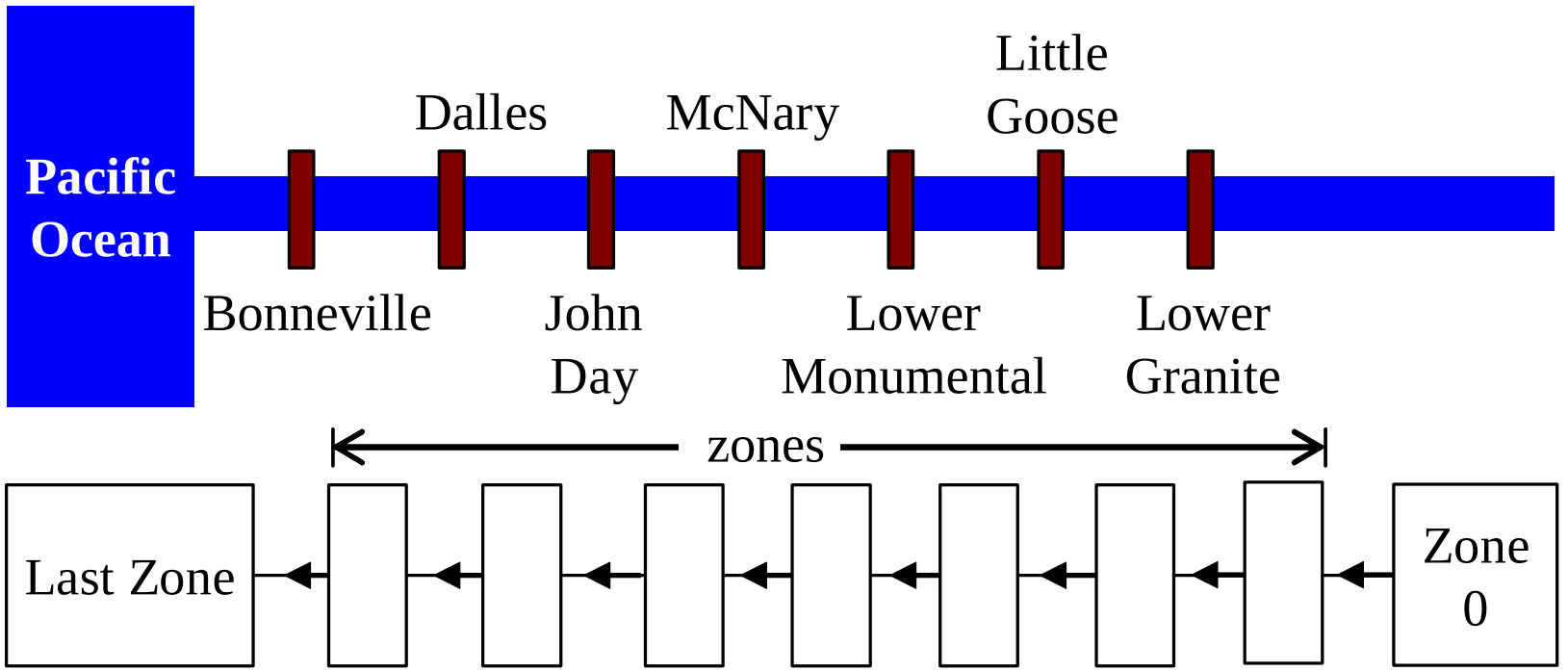
The Perfect Freethrow



Smolt Migration

It takes on average about 27 days for the smolt migration to be completed, which corresponds to a per capita migration rate of $m = 0.568$. Supposing that 100,000 smolts are initially in zone 0, what does the smolt migration system of solutions look like for the Columbia-Snake-Tucannon river system.

Smolt Migration



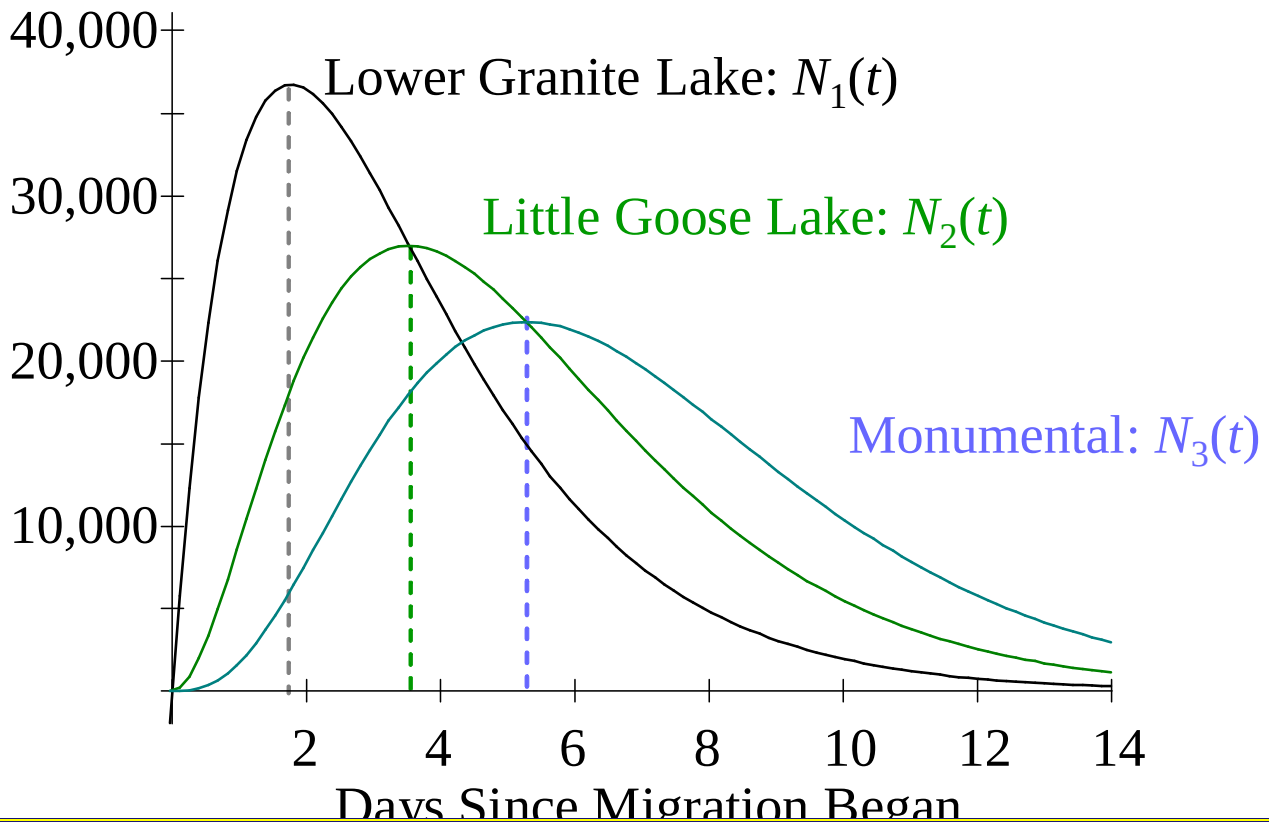
Solution

If $m = 0.568$, then $N_0(t) = 100,000e^{-0.568t}$. However, because "zone 0" is abstractly defined, the greater significance of $N_0(t)$ is in its use above in finding solutions $N_1(t), \ldots, N_8(t)$. The first few of these solutions are

$$\begin{aligned} N_1(t) &= 100,000(0.568t)e^{-0.568t} = 56,800te^{-0.568t} \\ N_2(t) &= \frac{100,000(0.568)^2 t^2}{2}e^{-0.568t} = 16,100t^2e^{-0.568t} \\ N_3(t) &= \frac{100,000(0.568)^3 t^3}{6}e^{-0.568t} = 3,054t^3e^{-0.568t} \end{aligned}$$

Considering Multiple Lakes in Series

These solutions predict that once the smolt population peaks in the lake behind Lower Granite Dam, the populations in subsequent lakes will peak about 2 days after the previous.



Remaining Solutions

Similarly, the remaining solutions are

$$N_4(t) = 100,000\frac{(0.568)^4 t^4}{4!}e^{-0.568t} = 433t^4e^{-0.568t}$$

$$N_5(t) = 100,000\frac{(0.568)^5 t^5}{5!}e^{-0.568t} = 49t^5e^{-0.568t}$$

$$N_6(t) = 100,000\frac{(0.568)^6 t^6}{6!}e^{-0.568t} = 4.7t^6e^{-0.568t}$$

and $N_7(t) = 0.38t^7e^{-0.568t}$.

This is a Poisson Distribution!!

Remaining Solutions

Similarly, the remaining solutions are

$$N_4(t) = 100,000\frac{(0.568)^4 t^4}{4!}e^{-0.568t} = 433t^4e^{-0.568t}$$

$$N_5(t) = 100,000\frac{(0.568)^5 t^5}{5!}e^{-0.568t} = 49t^5e^{-0.568t}$$

$$N_6(t) = 100,000\frac{(0.568)^6 t^6}{6!}e^{-0.568t} = 4.7t^6e^{-0.568t}$$

and $N_7(t) = 0.38t^7e^{-0.568t}$.
This is a Poisson Distribution!!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Final Thoughts

- It will be necessary to know at least one technology
  - More than 1 is not really necessary
  - Computer Algebra Systems like Maple, Sage, Mathematica are sufficient
  - Programming languages R or Python
  - Java, C++, Csharp, etc. might be good
- Possible Goal: Libraries of computation-based investigations
  - Biologists already have large repositories – Bioquest, for example
  - Issue: Finding/creating an appropriate computational context
    - Sage is a possibility...
    - But perhaps we need a version of Sage just for this purpose!

Thank you!

Any Questions