Mathematical Modeling Software: Ins and Outs

Eric T. Lofgren, MSPH
Dept. of Epidemiology
UNC Gillings School of Global Public Health

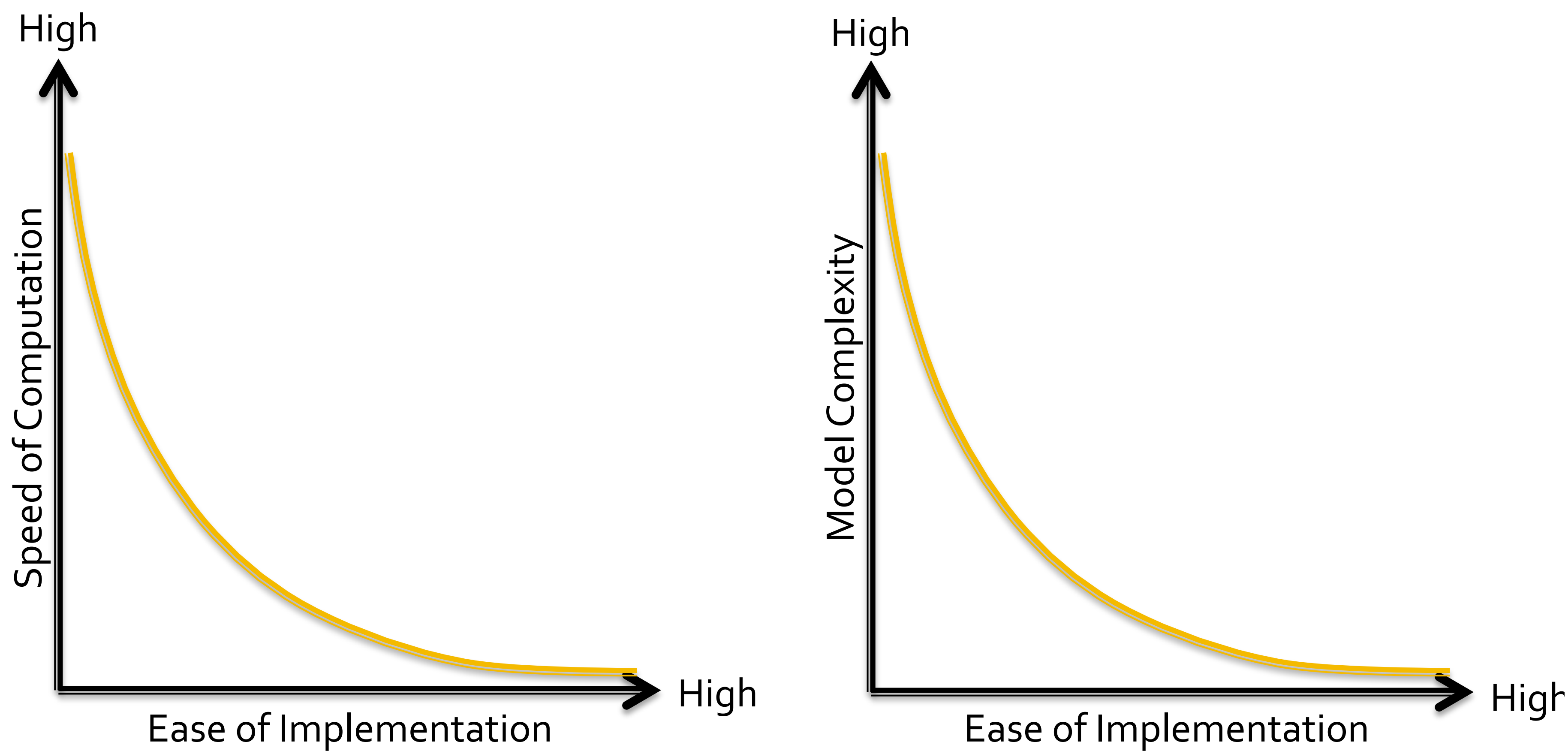
A Disclosure Reminder

- ETL has no conflicts of interest to disclose and has no financial relationship with any commercial software provider discussed in this presentation.

General Philosophy

- Familiar is better than new
- Flexible is better than specialized
- “Free” is better than expensive
- As much power as is needed, not as much as is available
  - Everything discussed today *could* be done with a pencil and paper

Two General Rules



Microsoft Excel

- Spreadsheet packages like Excel work well for difference equation based compartmental models
  - Almost all of this is true for Numbers (Apple) or OpenOffice/LibreOffice (Open Source)
  - Excellent tool to explore some of the concepts we've discussed today
- Lots of good example SIR models available online
- Best for models that are:
  - Simple – many compartments become difficult to see on a single sheet
  - Short – Excel struggles with large data sets

Supplementing Excel

- Excel's problems with large datasets primarily a data visualization problem
  - Newer versions *very* slow
  - Export results from Excel as a .xls or .csv file and open them in a software package meant for making publications ready graphs
  - Most major statistical software programs will do
- NodeXL (http://nodexl.codeplex.com/) and PopTools PopTools (www.poptools.org) provide useful extensions to the Windows version of Excel for epidemiologic models.
  - Note PopTools may be on the way out due to very old code

Some Problems with Excel

- Visualization
  - Excel's graphs are looked down on for reasons both valid and not
  - As mentioned previously, the solution to this problem is to bring your data into another tool
- Difference equation perception problems
  - Differential equations viewed as more sophisticated
    - Even at times when they aren't
  - Presents a problem in the *publication* of results
  - Won't matter for learning how models work, exploring them for yourself
  - This is changing

SAS

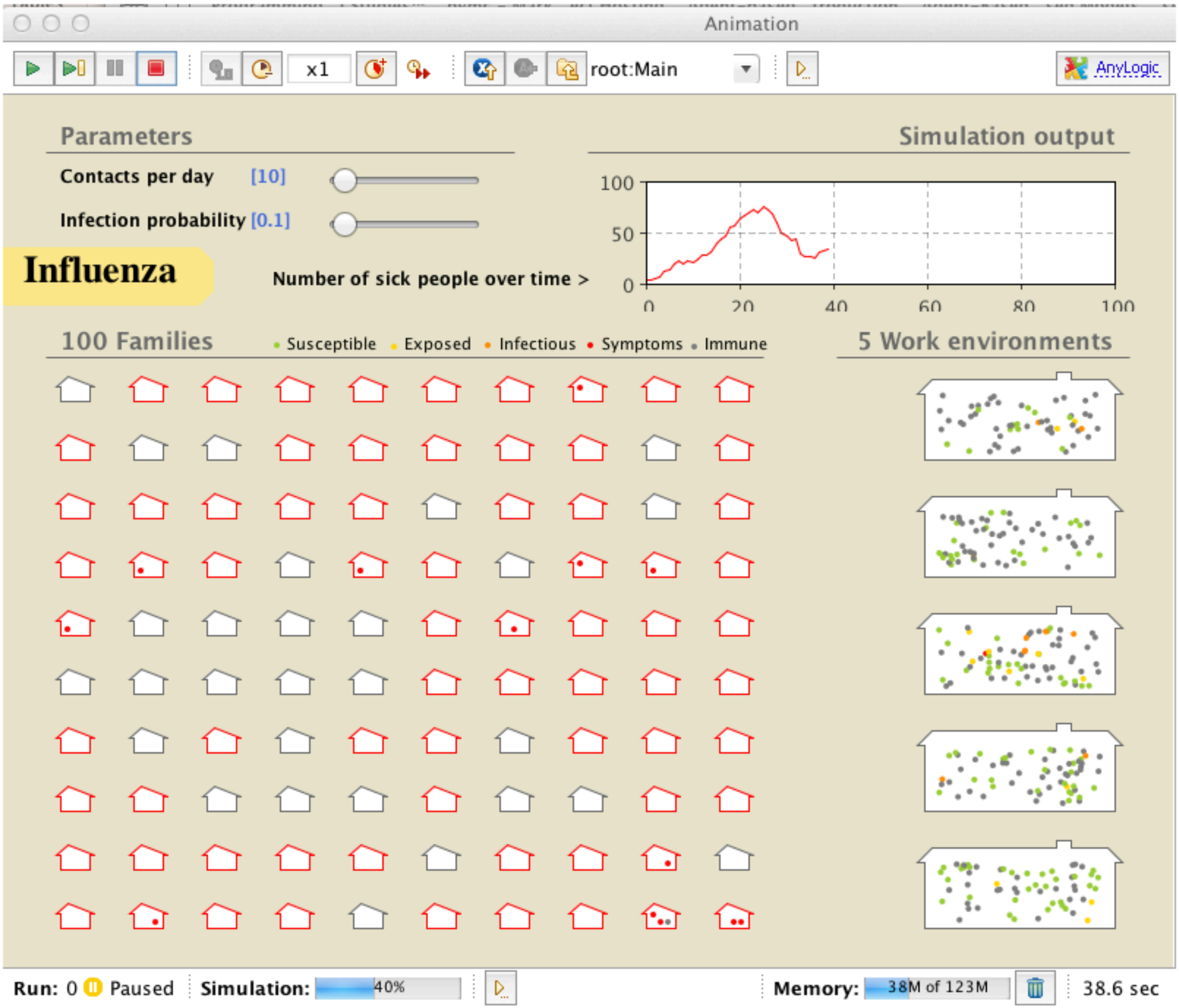
- It is *possible* to do SIR models in SAS
- Not necessarily what it was meant for
- Two methods in SAS:
  - DO Loops, generating the model as data
  - PROC MODEL
    - http://www.nesug.org/proceedings/nesug99/po/po053.PDF
  - May be possible to use IML for network models based on their matrix form, and more sophisticated SIR model analysis not mentioned in this Learning Institute

SAS Pros and Cons

- Pros:
  - Familiar language to some
  - Active community to support programming questions
  - Decent graphics with ODS (SAS 9.2 and above)
  - Handles large data sets very well
  - Very strong random number generation
- Cons:
  - *Expensive*
  - Somewhat awkward to use for this purpose
- Alternate uses:
  - Graphing data from Excel
  - Analysis of models from other software using SAS's statistical routines

AnyLogic

- Very flexible tool
  - Can run agent-based, network and SIR-type models
- Can create stand-alone programs to demonstrate your results
- Lots of demos of relevant models online, strong corporate support
- Cons: Expensive, proprietary
- Worth visiting their website anyway to play with some of their demo models or try out the software


Animation
x1
root:Main
AnyLogic
Parameters
Contacts per day [10]
Infection probability [0.1]
Influenza
Number of sick people over time >
Simulation output
100
50
0
0 20 40 60 80 100
100 Families
Susceptible
Exposed
Infectious
Symptoms
Immune
5 Work environments
Run: 0 Paused
Simulation: 40%
Memory: 38M of 123M
38.6 sec

MATLAB and Mathematica

- Dominant mathematics software packages
- If you work with modelers, engineers, applied mathematicians etc. odds are they know one of these two
- Very good for more complicated math in SIR models not covered in this Learning Institute
- Good support and established user communities
- Overkill for most things covered today
- Expensive
- Open source alternatives like SAGE and Octave

Free Software

R

- Open-source programming language intended for use in statistics
- Now has extensions into other areas, can be used to implement SIR and agent models and probably network models
- Very good visualization tools
- Fast growing and popular
- Likely able to also do network models and agent based models; this entails some difficulty
- Somewhat steep learning curve
- Will be what we use for the hands-on section later today

Python

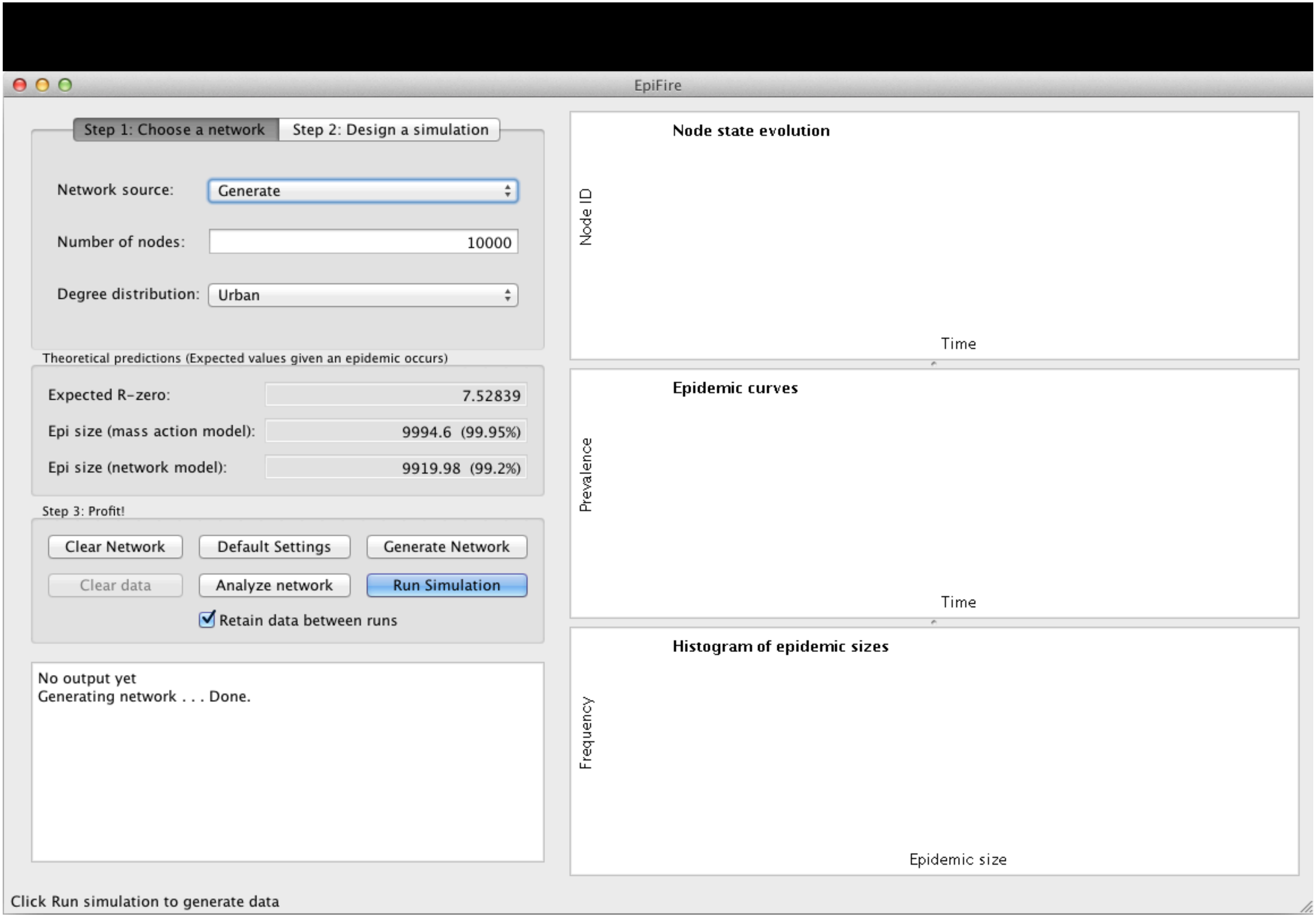
- General purpose programming language
- Can be used to implement all the models discussed today with varying ease
- Slower than some other programming languages (this is a problem for R as well)
- Fast growing and popular
- Slightly easier to learn than R (this is an opinion, not fact)
- Stronger for math and modeling, much weaker for statistics
- All the SIR models seen so far were coded in Python
  - Code is available on GitHub

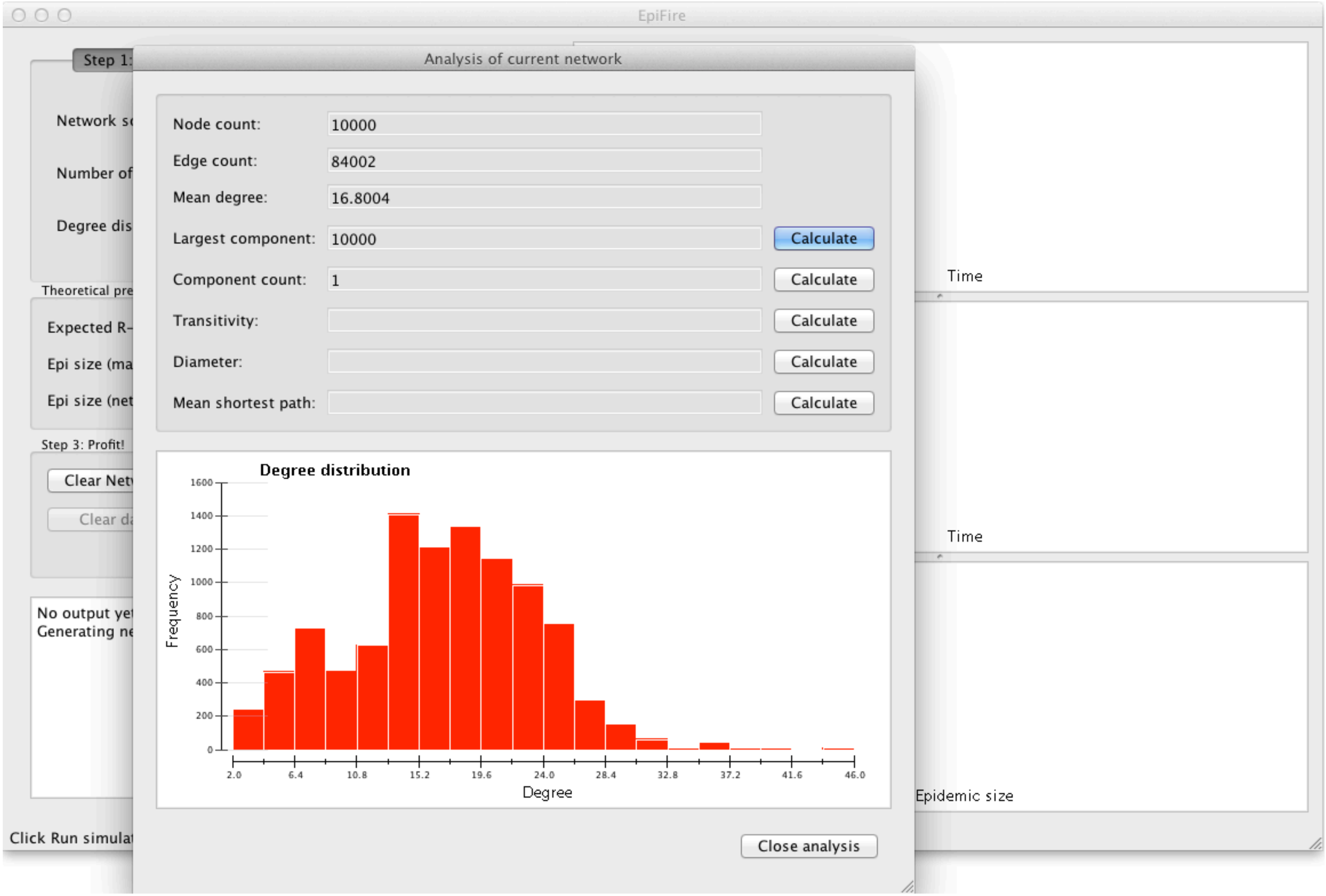
Benefits of a Programming Language

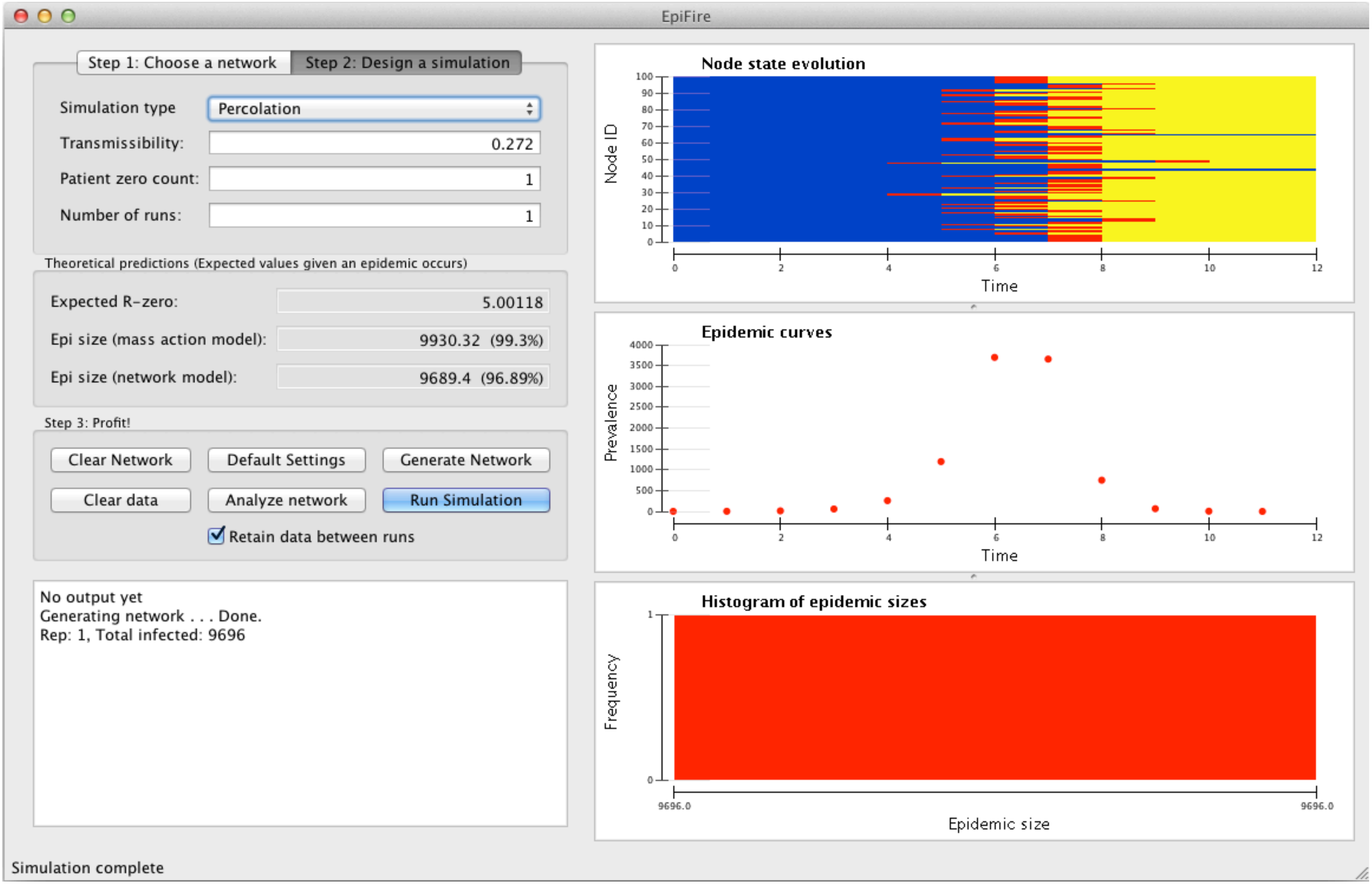
- Maximum flexibility over your model, freedom to answer your question
  - "Mad Libs is a game where key words in a short story have been replaced with blanks. Players fill in the blanks with designated parts of speech ("noun", "adverb") or types of words ("body part", "type of liquid"), without seeing the rest of the story. Occasionally, hilarity ensues, but no one really believes that this is an effective method for generating great literature." – Amanda Cox, New York Times
- The only cost is your time
- But…
  - It does cost your time
  - It isn't easy
  - Only as polished as you're willing to make it
  - Isn't always necessary if you want to learn and explore the concepts of modeling

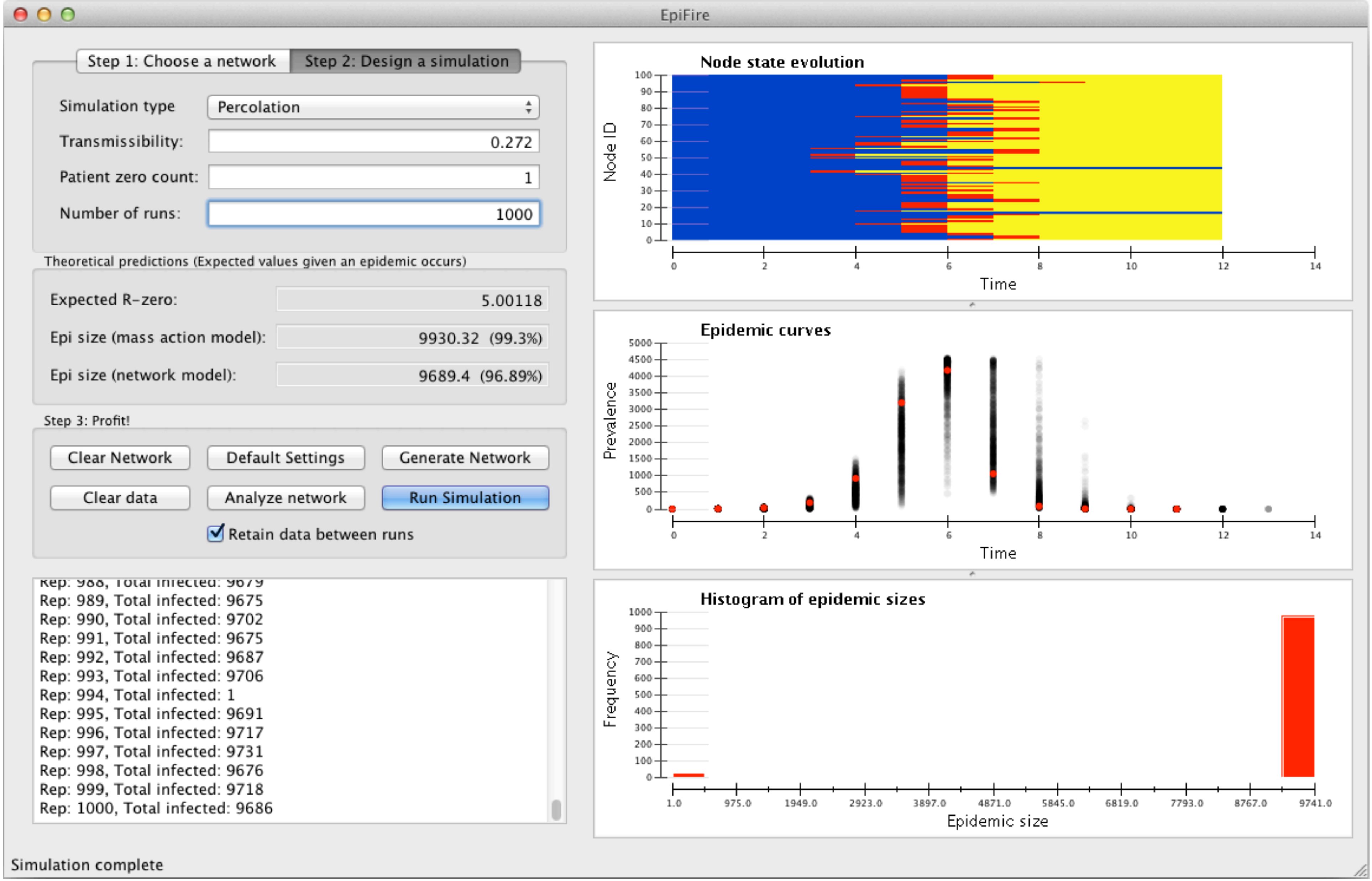
EpiFire

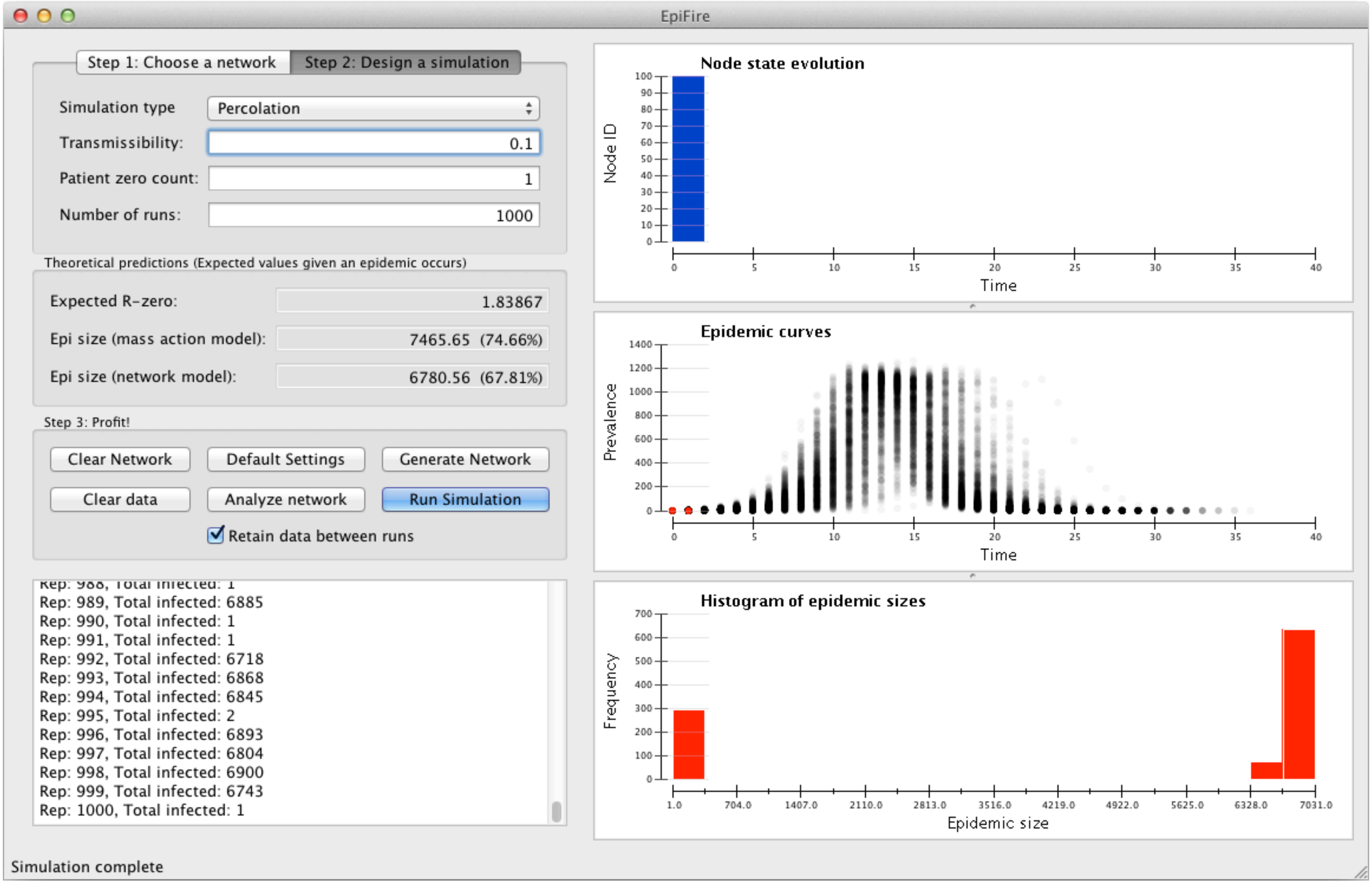
- Well-designed, free software for network epidemic models
- http://epifire.sourceforge.net/


EpiFire
Step 1: Choose a network
Step 2: Design a simulation
Network source:
Generate
Number of nodes:
10000
Degree distribution:
Urban
Theoretical predictions (Expected values given an epidemic occurs)
Expected R-zero:
7.52839
Epi size (mass action model):
9994.6 (99.95%)
Epi size (network model):
9919.98 (99.2%)
Step 3: Profit!
Clear Network
Default Settings
Generate Network
Clear data
Analyze network
Run Simulation
Retain data between runs
No output yet
Generating network . . . Done.
Click Run simulation to generate data
Node state evolution
Node ID
Time
Epidemic curves
Prevalence
Time
Histogram of epidemic sizes
Frequency
Epidemic size


EpiFire
Analysis of current network
Step 1:
Network s
Number of
Degree dis
Theoretical pre
Expected R-
Epi size (ma
Epi size (net
Step 3: Profit!
Clear Net
Clear d
No output ye
Generating ne
Click Run simula
Node count:
10000
Edge count:
84002
Mean degree:
16.8004
Largest component:
10000
Calculate
Component count:
1
Calculate
Transitivity:
Calculate
Diameter:
Calculate
Mean shortest path:
Calculate
Degree distribution
Frequency
Degree
Time
Time
Epidemic size
Close analysis


EpiFire
Step 1: Choose a network
Step 2: Design a simulation
Simulation type
Percolation
Transmissibility:
0.272
Patient zero count:
1
Number of runs:
1
Theoretical predictions (Expected values given an epidemic occurs)
Expected R-zero:
5.00118
Epi size (mass action model):
9930.32 (99.3%)
Epi size (network model):
9689.4 (96.89%)
Step 3: Profit!
Clear Network
Default Settings
Generate Network
Clear data
Analyze network
Run Simulation
Retain data between runs
No output yet
Generating network . . . Done.
Rep: 1, Total infected: 9696
Simulation complete
Node state evolution
Node ID
Time
Epidemic curves
Prevalence
Time
Histogram of epidemic sizes
Frequency
Epidemic size
9696.0
9696.0


EpiFire
Step 1: Choose a network
Step 2: Design a simulation
Simulation type
Percolation
Transmissibility:
0.272
Patient zero count:
1
Number of runs:
1000
Theoretical predictions (Expected values given an epidemic occurs)
Expected R-zero:
5.00118
Epi size (mass action model):
9930.32 (99.3%)
Epi size (network model):
9689.4 (96.89%)
Step 3: Profit!
Clear Network
Default Settings
Generate Network
Clear data
Analyze network
Run Simulation
Retain data between runs
Rep: 989, Total infected: 9675
Rep: 990, Total infected: 9702
Rep: 991, Total infected: 9675
Rep: 992, Total infected: 9687
Rep: 993, Total infected: 9706
Rep: 994, Total infected: 1
Rep: 995, Total infected: 9691
Rep: 996, Total infected: 9717
Rep: 997, Total infected: 9731
Rep: 998, Total infected: 9676
Rep: 999, Total infected: 9718
Rep: 1000, Total infected: 9686
Simulation complete
Node state evolution
Node ID
Time
Epidemic curves
Prevalence
Time
Histogram of epidemic sizes
Frequency
Epidemic size
1.0
975.0
1949.0
2923.0
3897.0
4871.0
5845.0
6819.0
7793.0
8767.0
9741.0


EpiFire
Step 1: Choose a network
Step 2: Design a simulation
Simulation type
Percolation
Transmissibility:
0.1
Patient zero count:
1
Number of runs:
1000
Theoretical predictions (Expected values given an epidemic occurs)
Expected R-zero:
1.83867
Epi size (mass action model):
7465.65 (74.66%)
Epi size (network model):
6780.56 (67.81%)
Step 3: Profit!
Clear Network
Default Settings
Generate Network
Clear data
Analyze network
Run Simulation
Retain data between runs
Rep: 989, Total infected: 6885
Rep: 990, Total infected: 1
Rep: 991, Total infected: 1
Rep: 992, Total infected: 6718
Rep: 993, Total infected: 6868
Rep: 994, Total infected: 6845
Rep: 995, Total infected: 2
Rep: 996, Total infected: 6893
Rep: 997, Total infected: 6804
Rep: 998, Total infected: 6900
Rep: 999, Total infected: 6743
Rep: 1000, Total infected: 1
Simulation complete
Node state evolution
Node ID
Time
Epidemic curves
Prevalence
Time
Histogram of epidemic sizes
Frequency
Epidemic size
1.0
704.0
1407.0
2110.0
2813.0
3516.0
4219.0
4922.0
5625.0
6328.0
7031.0

Software in Summary

- Excel and EpiFire will get you 90% of the way there
  - Good tools for learning, and serious models can be and are done in Excel
- Modeling is associated with many fields: Epidemiology, engineering, finance, applied math, etc.
- As a consequence there are tons of possible options
  - None of them a clear winner
- Least expensive and most flexible is to learn a programming language
- **Use what your colleagues/collaborators use**