AGENT-BASED MODELS: CONCEPTUAL FOUNDATIONS

A BRIEF INTRODUCTION TO ME

- Assistant Professor at the Paul G. Allen School for Global Public Health at Washington State University
    - Research focus is in computational epidemiology, especially around zoonotic diseases, antimicrobial resistance, and healthcare-associated infections
- Formerly a postdoc at the Network Dynamics and Simulation Science Lab at Virginia Tech
- PhD in Epidemiology from UNC Chapel Hill
- Also a bit of a nerd
- Contact Information:
    - Email: Eric.Lofgren@wsu.edu
    - Twitter: @GermsAndNumbers
    - GitHub: elofgren

- Please feel free to interrupt/ask questions as the presentation is going

A BRIEF INTRODUCTION TO YOU

- How many of you are epidemiologists?
- How many have done any sort of modeling work before?
- Infectious disease vs. Non-infectious disease?
- If you work in a programming language, what language do you use?
  - SAS?
  - R?
  - Python?
  - Other?

GOALS FOR TODAY AND TOMORROW

- Today: Foundations and Theory
  - Give you some idea of what agent-based modeling is, and what it entails
  - Give you some examples of what agent-based modeling looks like in the context of epidemiology
- Tomorrow: Hands-on Work
  - Agent-based modeling is often best learned by getting your hands dirty and writing some code

RESOURCES FOR THIS CLASS

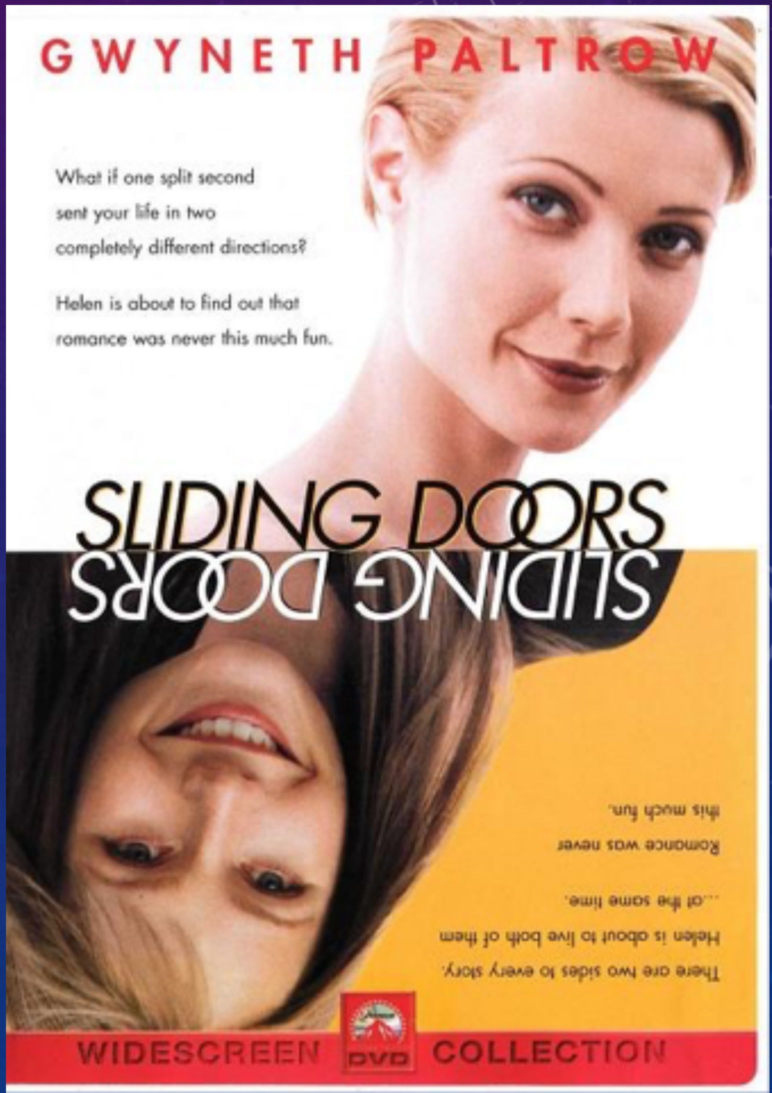
- All the slides (sans animation) and code we use will be available at: https://github.com/elofgren/abmph
- Before tomorrow you should download and install NetLogo from: http://ccl.northwestern.edu/netlogo/

WHAT DO YOU MEAN BY "MODEL"?

- Statistical vs. Mathematical vs. Computational Models
  - Statistical Models: What does our data tell us about the world? *Descriptive*
  - Mathematical Models: How can we use our data to describe how the world works in equations? *Mechanistic*
  - Computational Models: How can we use our data to **simulate** how the world works? *Mechanistic*
- This categorization presents things as having starker divisions than they do in practice, especially for the last two types of models.
- Today focuses on computational models, epidemiology as a field is still heavily dominated by statistical models

WHAT IF? THE FUNDAMENTAL QUESTION OF EPIDEMIOLOGY

- What if a patient had been given Treatment A instead of Treatment B?
- What if someone had never started smoking?
- What if the MMR vaccination rate was 10% higher?
- Counterfactual questions like these are at the core of causal inference, and underlie most medical research
- But they are impossible to answer



OBSERVATIONAL METHODS

- Randomized controlled trials or other randomized experiments are considered the closest means to estimate a causal effect
  - Not without issues – compliance, post-randomization differences between trial arms, etc.
- Other study designs are all methods of attempting to statistically control for differences between groups to isolate an effect
  - Subject to residual confounding, selection bias, etc.
- Limited to within-dataset inference
  - Generalizability must be assumed
  - Indirect or spillover effects are difficult to capture
  - Increasing sample size is expensive
- How do you study large scale policy change? Can you randomize outbreak response? Or policing policy?

WHAT CAN COMPUTATIONAL MODELS DO?

- Dynamics and feedback loops
  - Exposure as a function of current prevalence (“Dependent Happenings”)
- Data Synthesis
  - Inference over multiple data sets, studies, etc.
- Data-free Hypotheticals
  - Preparedness, policy changes, etc.
- Translational Research
  - Apply research findings to a model of a system

WHAT MODELS ARE AND ARE NOT

- Are:
  - A powerful tool for public health planning and research
  - Something every epidemiologist should be passingly familiar with
  - A rigorous, systematic way to try and describe how an entire disease process works
  - Capable of providing truly counterfactual estimates*
- Are Not:

WHAT MODELS ARE AND ARE NOT

- Are:
  - A powerful tool for public health planning and research
  - Something every epidemiologist should be passingly familiar with
  - A rigorous, systematic way to try and describe how an entire disease process works
  - Capable of providing truly counterfactual estimates*
- Are Not:
  - **Magic**

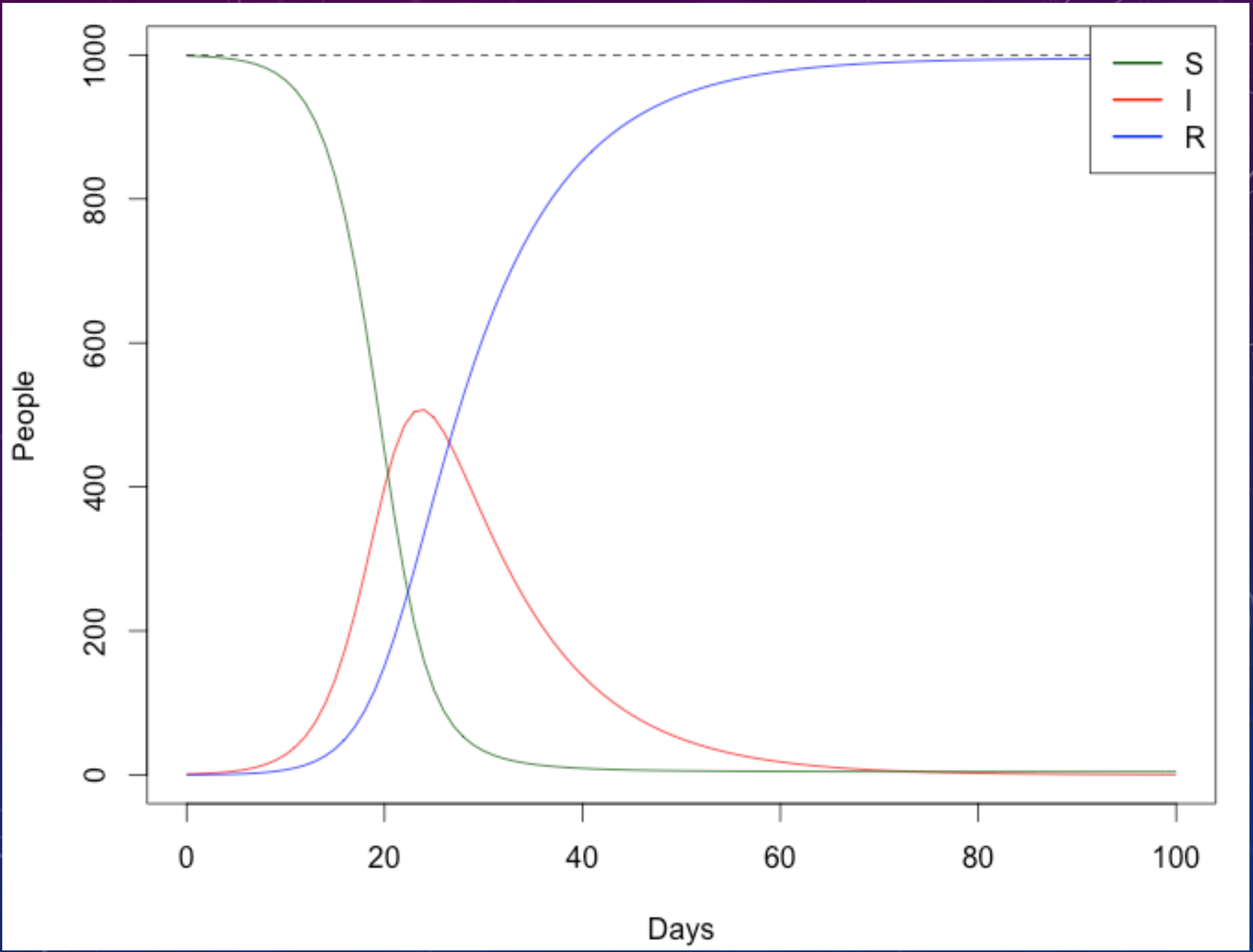
COMPARTMENTAL MODELS

- In order to understand agent-based models, it is helpful to briefly touch on what came before
- Compartmental models are frequently used, especially in infectious disease research
- Patients are divided up into a number of disease states
    - The classic disease model has *S* (Susceptible), *I* (Infected) and *R* (Removed) classes
- Movement between the compartments governed (usually) by a system of ordinary differential equations

$$\frac{dS}{dt} = -\beta S \frac{I}{N}$$

$$\frac{dI}{dt} = \beta S \frac{I}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$



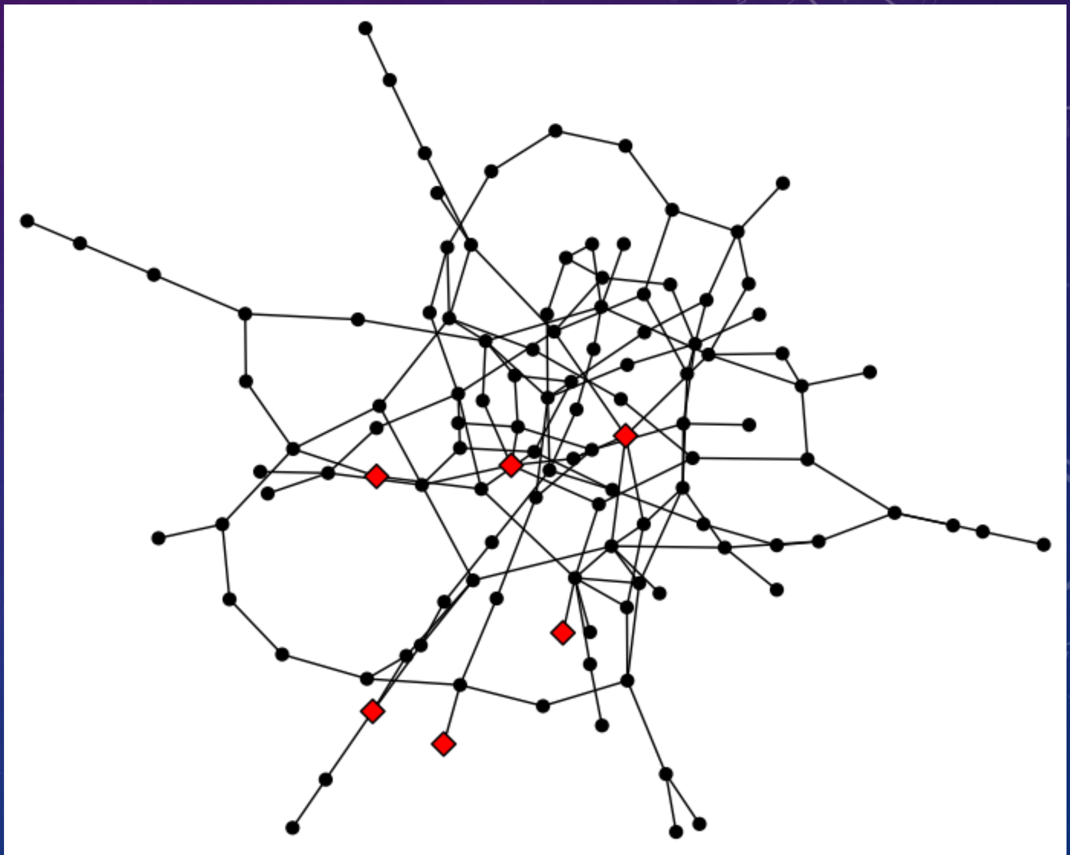
1000
800
600
400
200
0
People
0
20
40
60
80
100
Days
S
I
R

ASSUMPTIONS

- Random Mixing
- Deterministic*
- Populations not individuals
- Population-level heterogeneity is cumbersome to implement

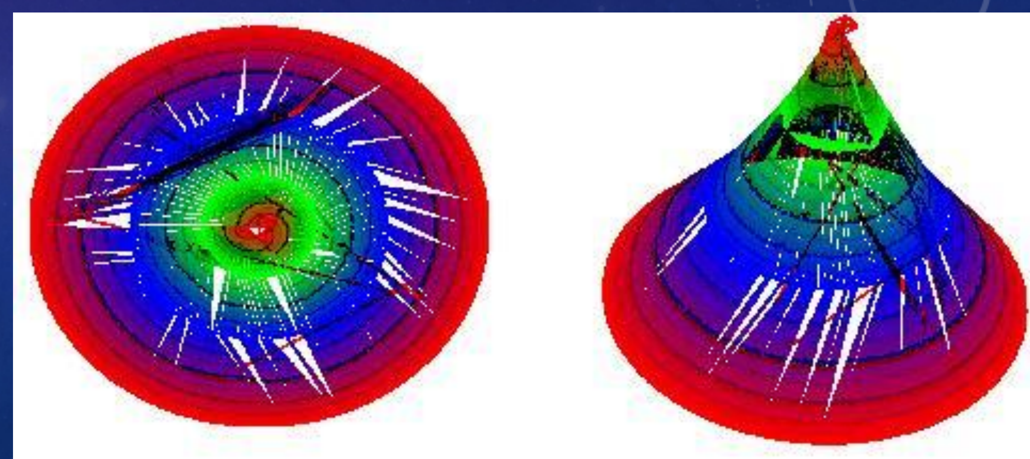
NETWORK MODELS

- Stochastic
- Track individuals (nodes)
- Non-random mixing
- Can incorporate heterogeneity
- But…



HOW DO YOU GET THE NETWORK?

- For large populations, networks are very, very hard to sample
- Ethics, population connectivity, and economics are all barriers
- Estimated networks from sensors, social media, etc. don't necessarily capture meaningful contact
- Mixing isn't random, but *is* assigned



ENTER THE AGENT-BASED MODEL

- Use a computer simulation to model lots of individuals in the same environment
- Stochastic
- Tracks Individuals
- Population is modeled as a set of autonomous “agents” with relatively simple rules
- Contact is driven by behavior
    - How we interact instead of who we interact with


RULE #1.
CARDIO


RULE #31
CHECK THE
BACK SEAT


RULE #32
ENJOY THE
LITTLE
THINGS


RULE #18
LIMBER
UP


RULE #17
BE A HERO

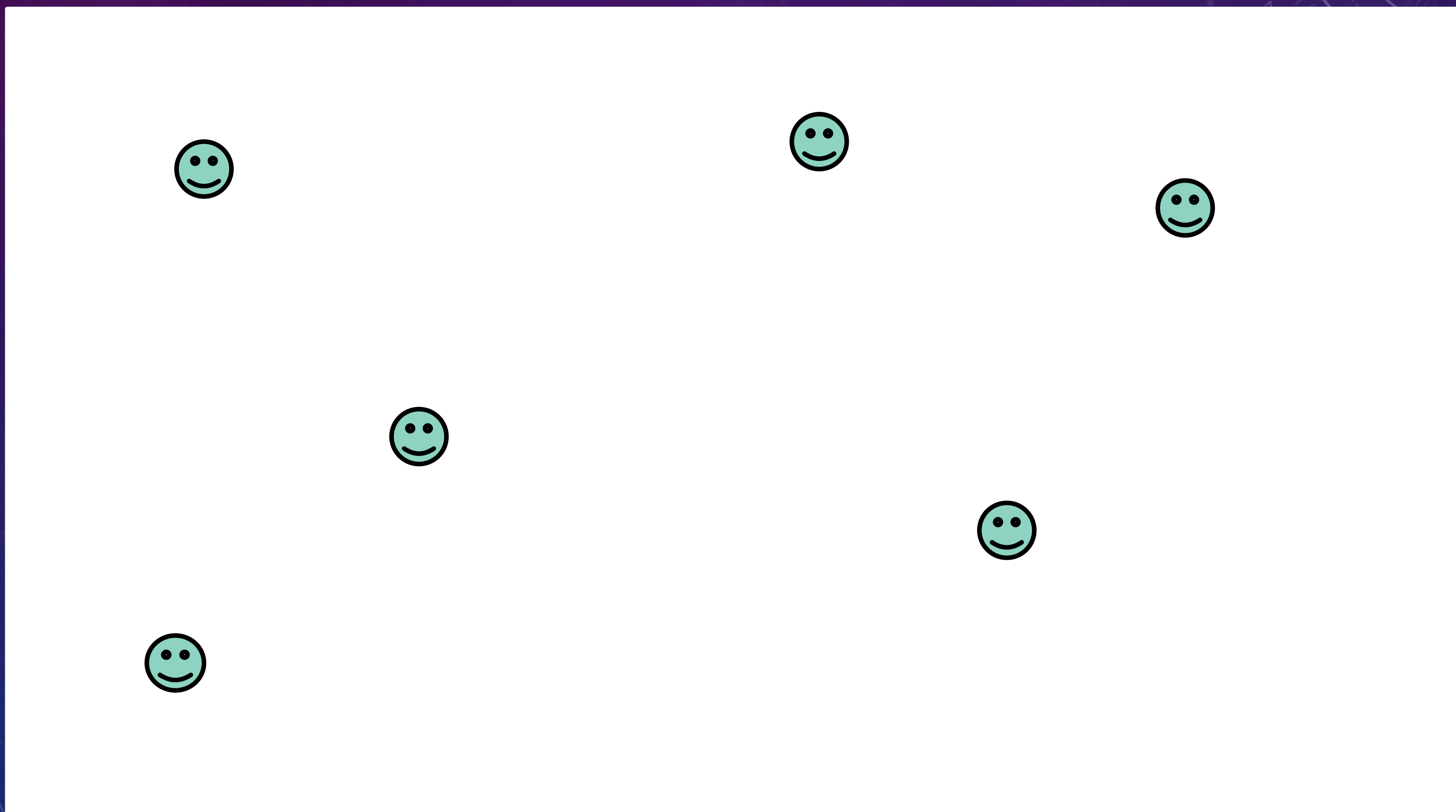
WHY IS THIS INTERESTING

- Very flexible approach
- New kinds of randomness (behaviors can be drawn individually from a distribution)
    - Complex results can arise from simple, low level interactions
    - Can help discover patterns other models later describe

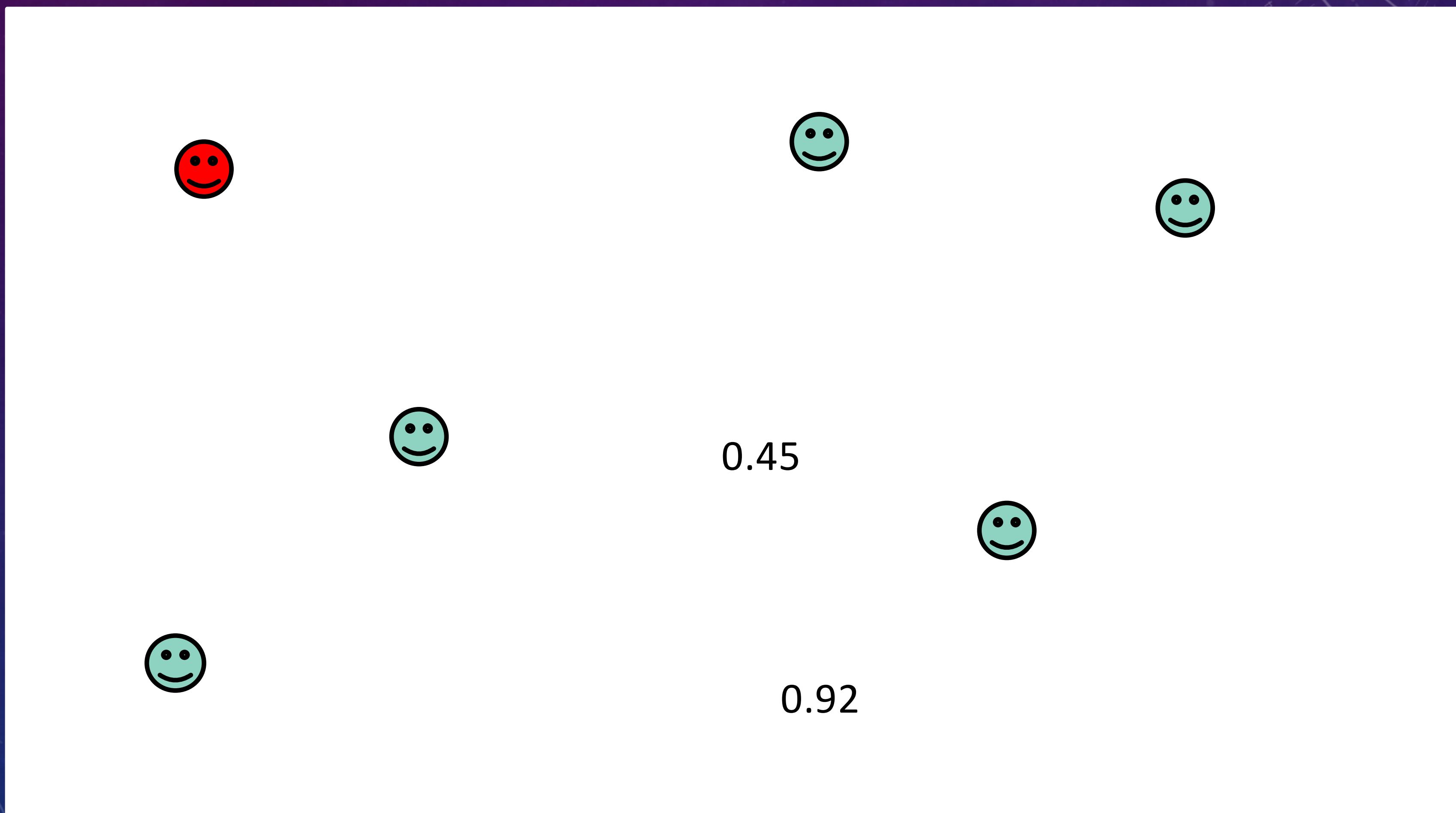
THE FUZZY GREY AREA

- Compartmental, Network and Agent-based models are often considered to be discrete entities.
- They really aren't
  - What if the behavior of an agent is "mix randomly"?
  - What if we make a compartment for every person?
  - What if nodes in a network add and remove links to one another based on rules?
  - What if we use an agent-based model to estimate the formation of a network?
    - NDSSL does this last one

BASIC EXAMPLE



WITH INFECTION

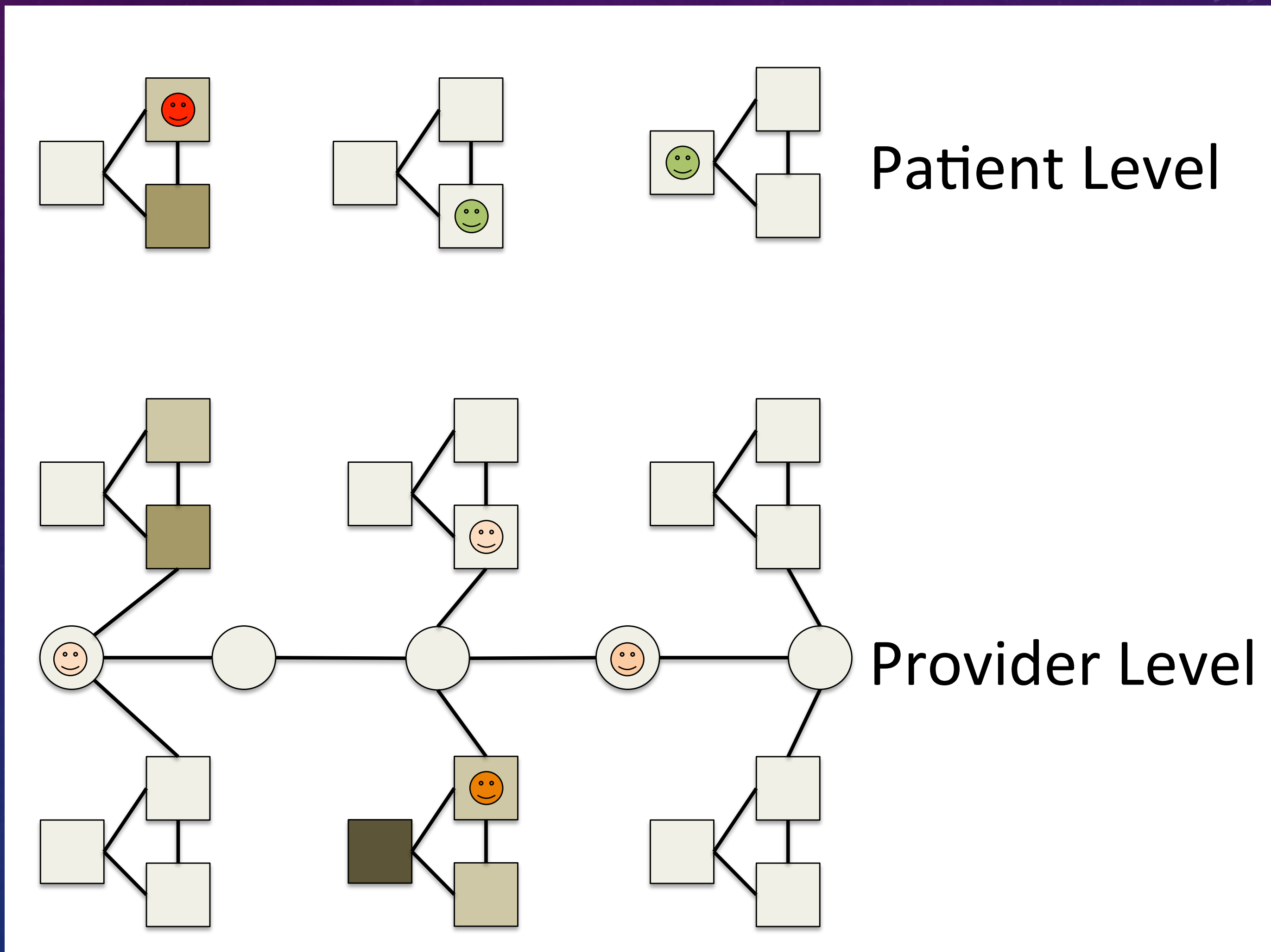


Transmission probability = 0.65

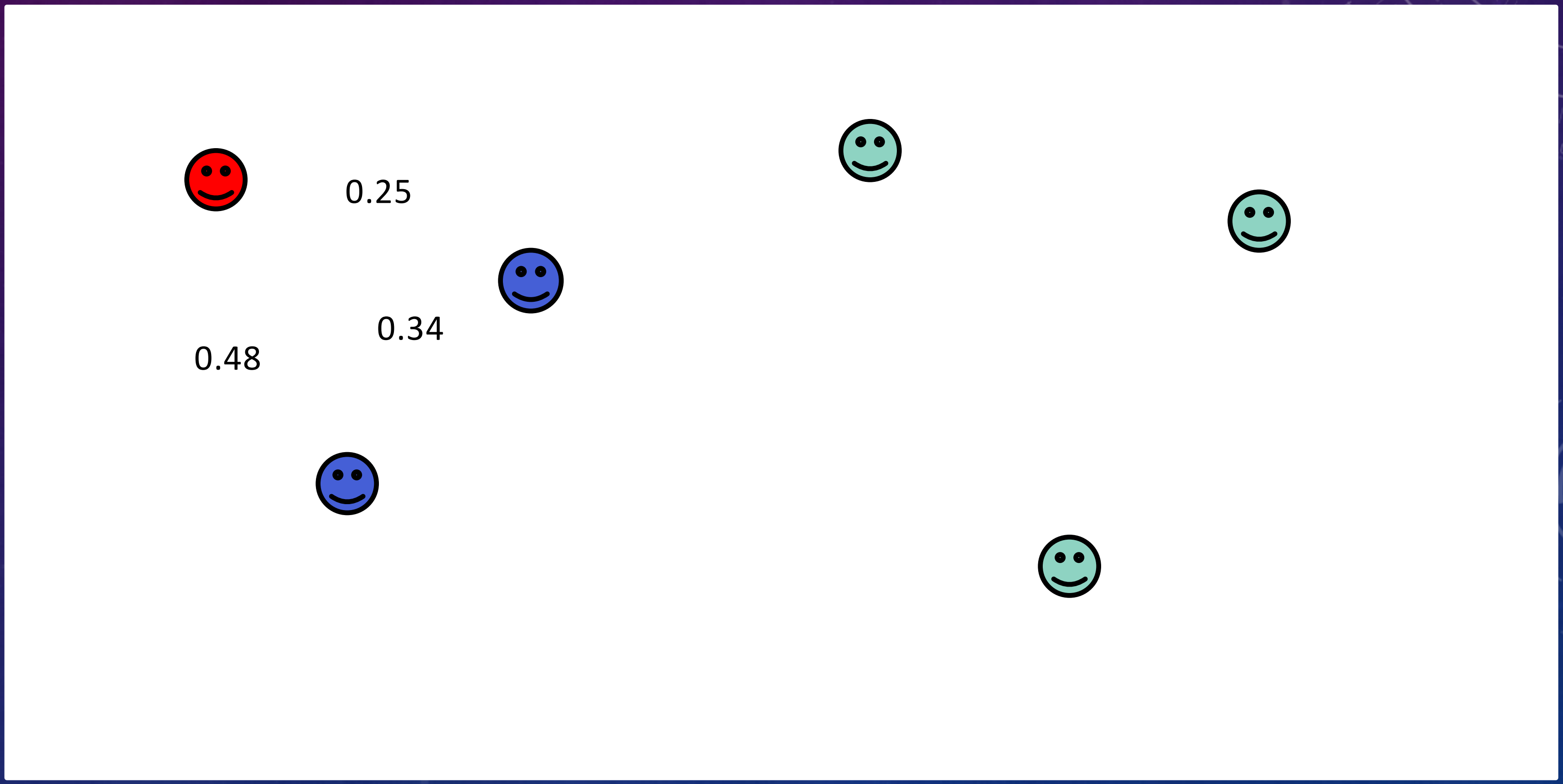
WHERE ABMS ARE PARTICULARLY STRONG

- Adding a type of stochasticity not present in other models
  - Random but rule-based mixing
  - Interactions with the environment
  - Positions, states and information about other agents
- Modeling different classes of individuals more easily
  - Draw parameters from a distribution, rather than a fixed value
  - Easily create a new type of agent by changing behavior rules

ENVIRONMENTAL AND STATE SENSING



DIFFERENT CLASSES



Transmission probability (Civilian) = 0.65
Transmission probability (HCW) = 0.30
Treatment probability = 0.80

MORE COMPLEXITY

- What if p(Infection|HCW) was a distribution, representing experienced and inexperienced first responders?
- What if p(Infection|HCW) changed with time, representing fatigue?
- What if infected individuals move randomly *until* they see a HCW?
- What if they try to *avoid* HCWs?
  - This was the case for some Ebola patients
- How about adding terrain?

A NOTE OF CAUTION

- Clearly, ABMs are a *very* powerful tool, and lend themselves well to sophisticated and complex models
  - Grouping and behavior processes, interaction with the environment, huge numbers of agents (a human body, an entire hospital, an entire healthcare system, an entire city…)
- It is easy to add complexity, it is hard to *implement* it
  - More complex models mean slower models
  - Parameter choices are difficult to find
- Easy to get carried away
  - Focus shifts to modeling the system, not the research question
- Randomness means you have to simulate the system *many* times

OTHER TRADEOFFS

- Few analytical solutions
  - Simulation results instead of proofs
  - Those that do exist are *hard*
- Difficult to describe
  - Consider the figures in this presentation
  - Can use SIR-like flow charts, but harder to represent the whole population
    - No equations
    - Reproducibility is difficult
- Programming expertise

GETTING STARTED

BASIC QUESTIONS

- What is the question or system you want to model?
- Why does it need to be modeled?
- What kind of model does it need?
- How fast do you need an answer?

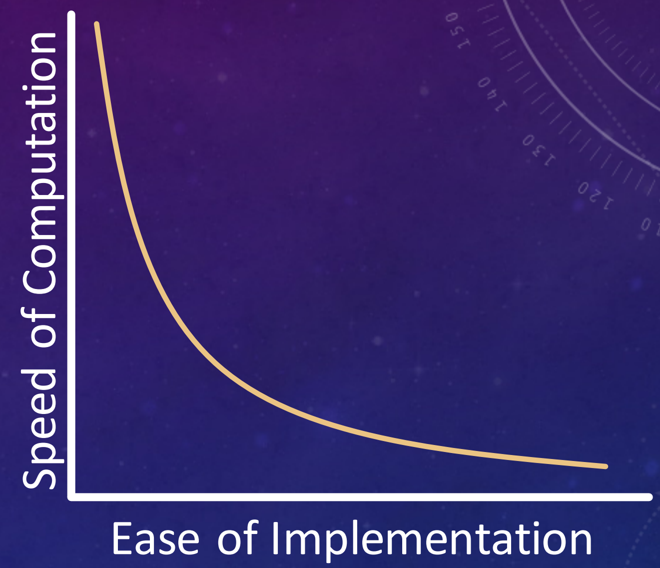
- ”I want to study the effect of incarceration policy on neighborhood resilience. I think there is a lot of indirect effects and feedback loops that exist. I’d like to model fairly sophisticated behavior, and people’s interactions with the environment, so I think I need an agent-based approach.”
- “I want to make an agent-based model of tuberculosis.”

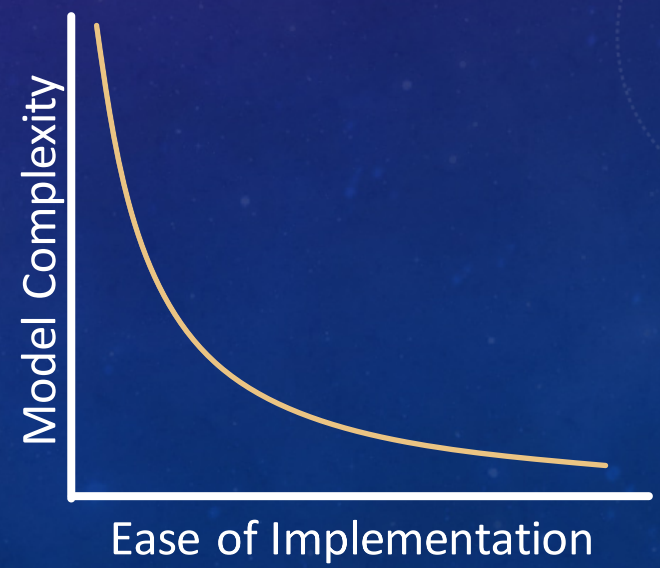
ASSEMBLE A TEAM

- Modeling is inherently a team science endeavor
- Look for potential collaborators:
  - Clinical colleagues
  - Science of behavior/decision-making
    - Psychology, Anthropology, Economics
  - Biology/Ecology
    - Many of these models are also heavily used in those fields
  - Computer Science
  - Mathematics

"WHAT SOFTWARE SHOULD I USE?"

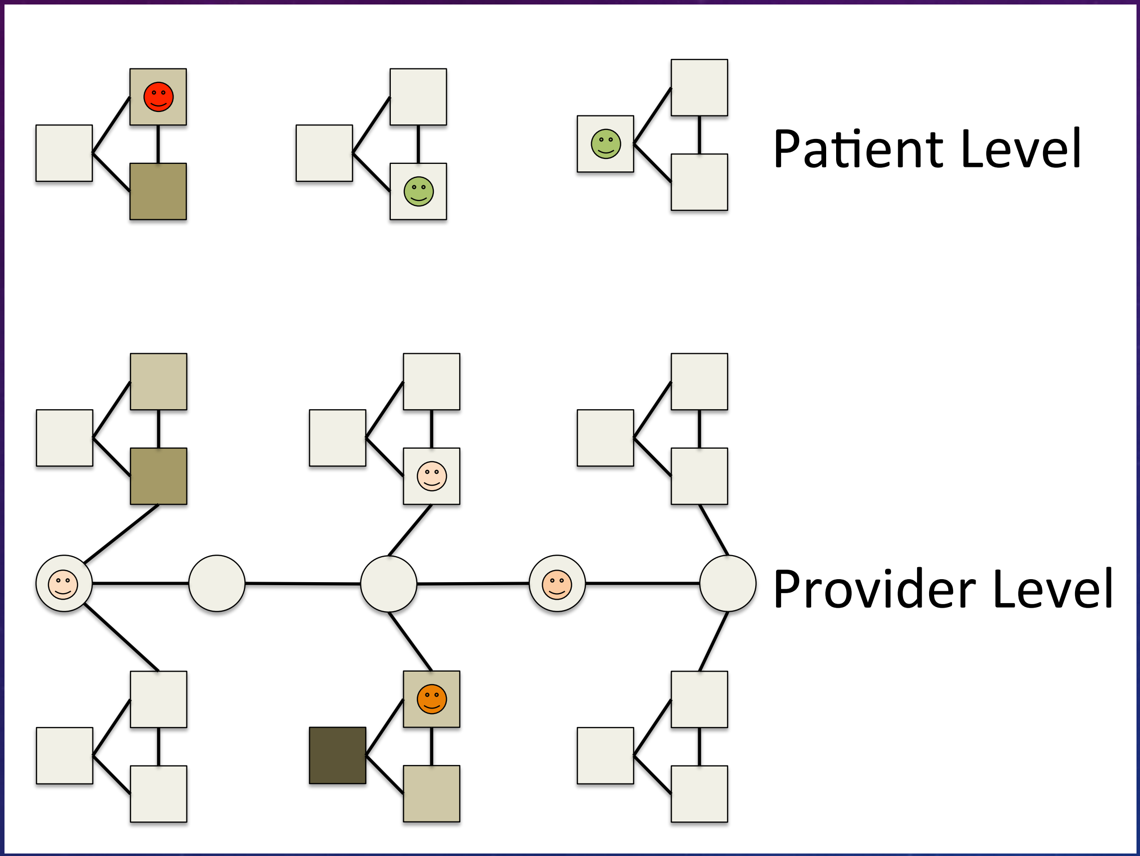
- Very common question
- Lots of possible options
- Open source, proprietary, graphical, etc.
- Could always break down and write your own
  - Lots of flexibility, lots of work
  - Isn't necessary just for learning
- Use what your colleagues/collaborators use

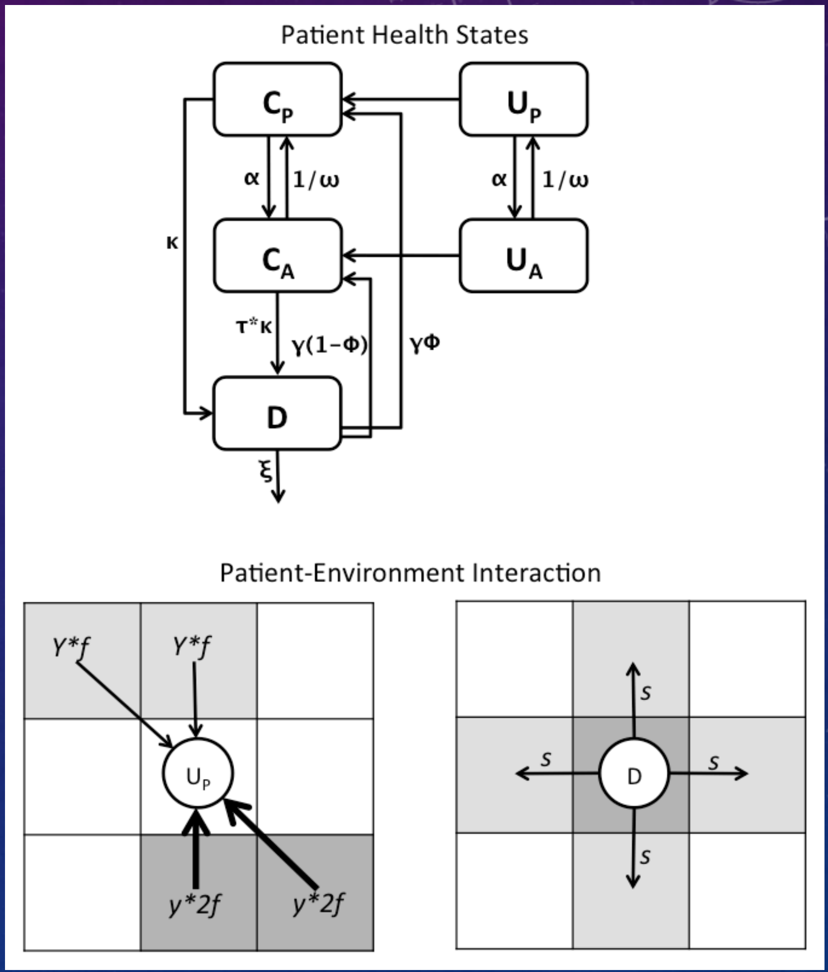




DESIGNING YOUR MODEL

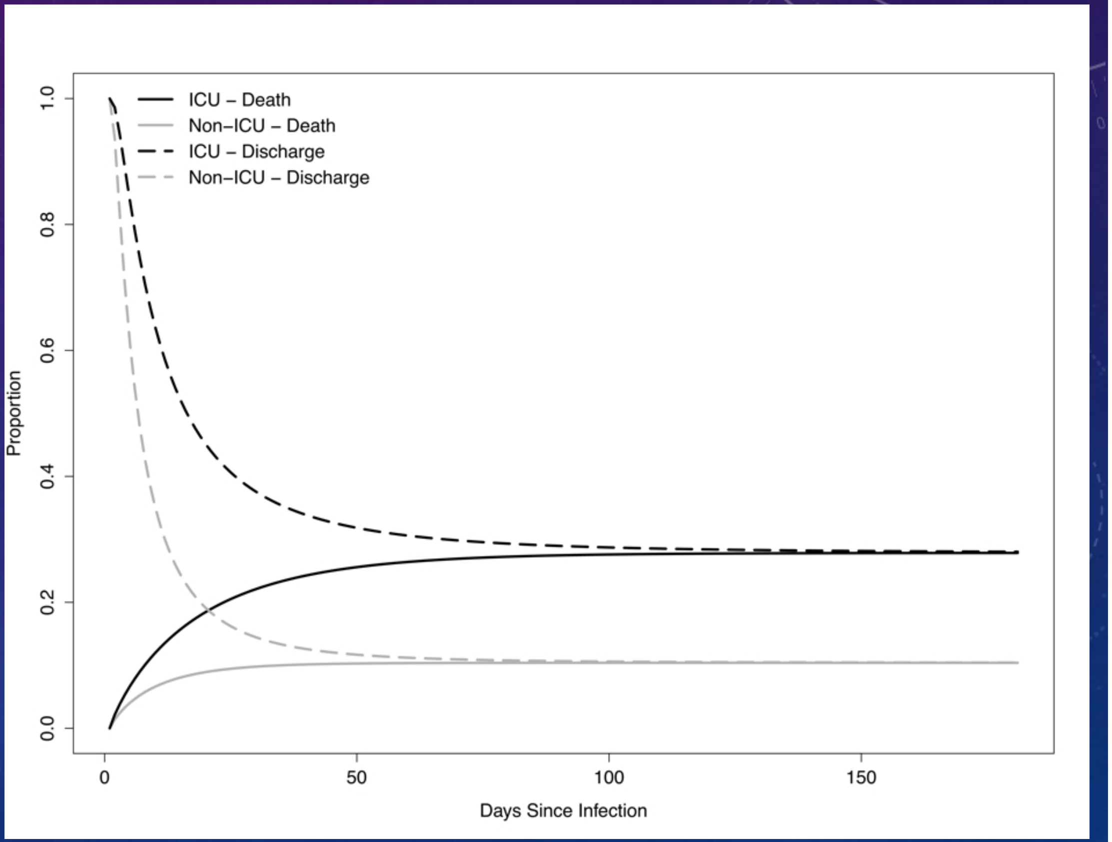
- Sit down, talk the problem over with your team, do a literature search, etc. and try to come up with a working picture of how you think your system works.
- Write that down/diagram it
- Write down agent behavior as a flow chart, identify everywhere you need a parameter
  - Can help to write “pseudocode”
- Start thinking about how you want your results formatted (summary estimates, individual results, etc.)


Patient Level
Provider Level


Patient Health States
C_P
U_P
C_A
U_A
D
α
1/ω
κ
τ*κ
γ(1−Φ)
γΦ
ξ
Patient-Environment Interaction
Y*f
y*2f
U_P
s
D

WHERE DO PARAMETERS COME FROM?

- The Literature
  - Other models
  - Effect estimates, RTC results, etc.
- Perform your own study
  - Epidemiologists are experts in parameter estimation
- Collect data and fit the model to it



FITTING MODELS TO DATA

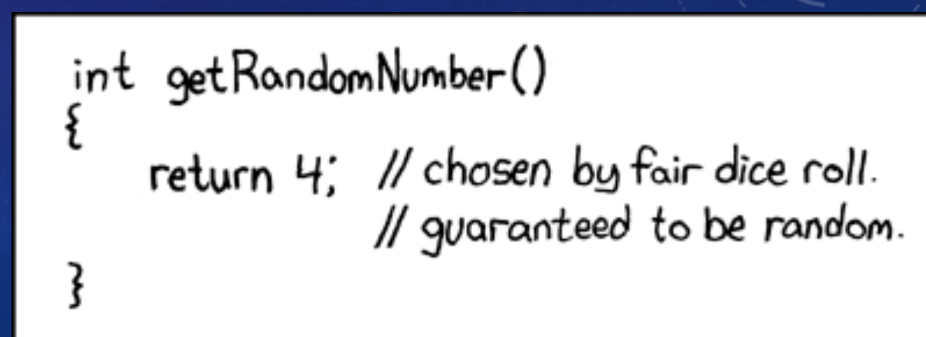
- A whole multi-day workshop in its own right
- Relatively straightforward with compartmental models
- Much less straightforward with agent-based models
  - Multiple dimensions to try and fit
  - Stochasticity – does one run not fitting mean a bad fit, or randomness?
  - Approximate Bayesian Computation, particle filtering, pattern-oriented approaches ("calibrating to experience"), and many, many others
- Verification: My model is giving the correct answer given the inputs I provided (the model is behaving as you expect it to behave. 2 + 2 = 4)
- Validation: My model is giving an answer that corresponds to reality
  - Conventional English use of the term "valid" implies a model is correct if it is successfully validated. **This is not the case.**
  - The model is only "not wrong"

IMPLEMENT!

- Implementation is a major part of the modeling enterprise
- Good coding practice
  - Software Carpentry
- Use modular code, test early and often
- Documentation is critical
  - Nothing you do will make sense 6 months from now/when a reviewer asks for revisions
  - `Model(parameters) # FIX ME` is next to useless
- **Version Control**
  - I like GitHub but there are many others
  - There’s a free plan for academics
  - Access to code is *a* form of reproducability
  - DOIs

AN ASIDE ABOUT RANDOM NUMBERS

- Agent-based models use tremendous amounts of random numbers
- How do random number generators work?
  - Random numbers can be generated from an arbitrary distribution
- What is a “seed” and why do I care?
- V&V using random numbers
- Random numbers in reproducibility and experimental design

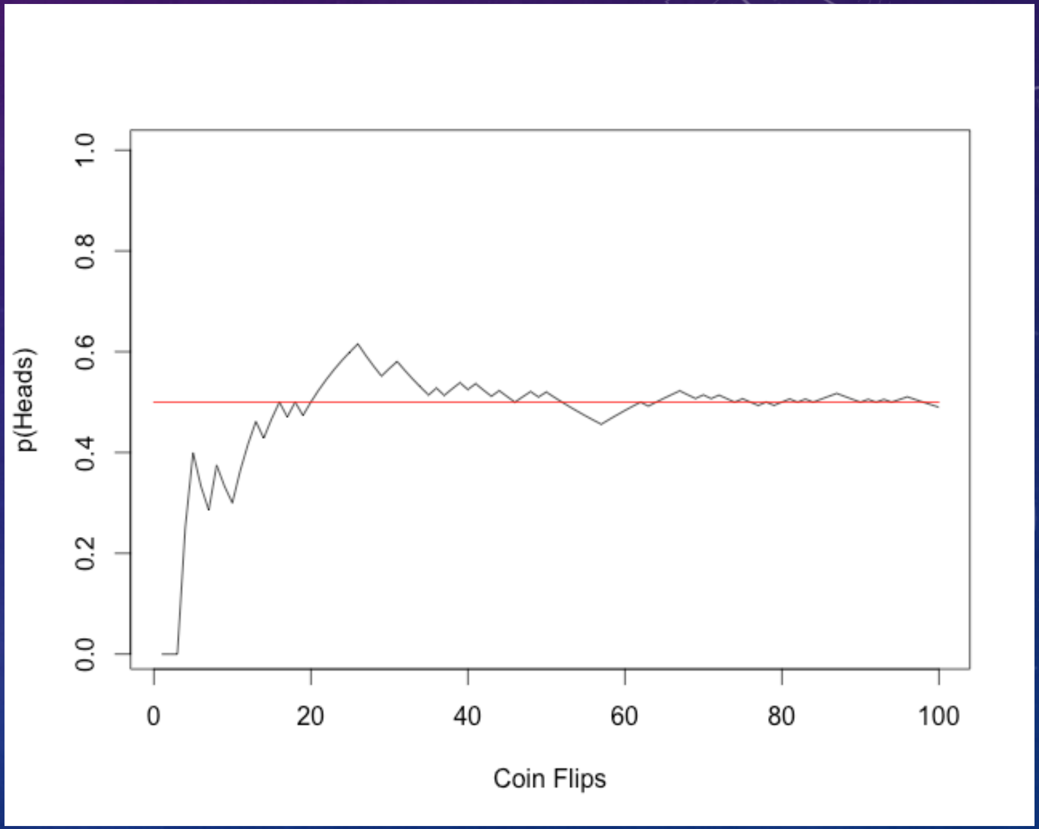


ANALYSIS

- Every time you change a parameter, you create a new counterfactual scenario
- Many/most ABMs are very amenable to basic statistical analysis – t-tests, ANOVA, etc.
- If you design your output correctly, you can analyze agent-based models as virtual cohort studies, or simulate other studies inside of them
- Caveats
  - p-values do not mean what you think they mean
  - Plot all your data at least once – multimodal, non-normal, etc. distributions are quite common

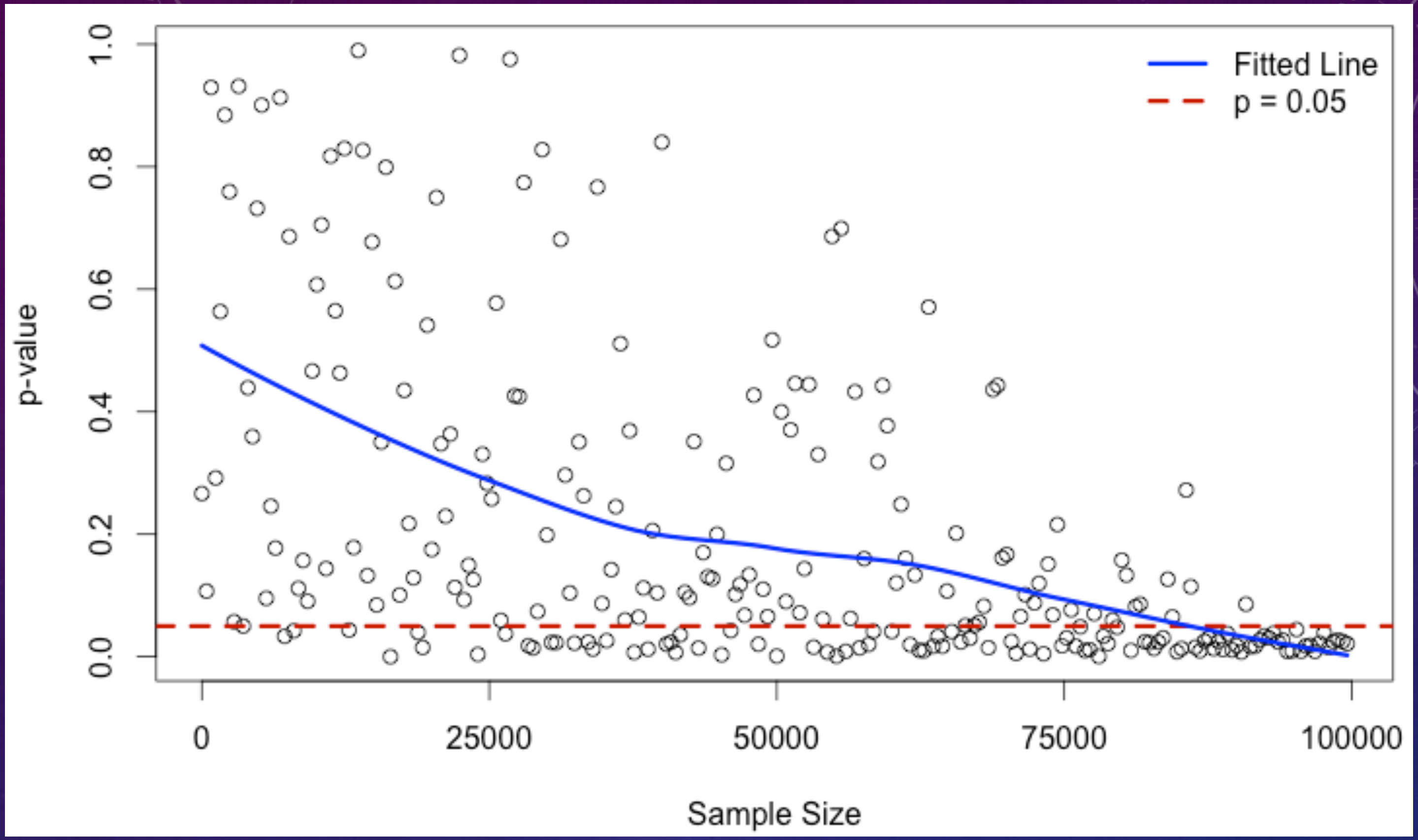
WHY SIMULATIONS WORRY ABOUT SAMPLE SIZE

- Law of Large Numbers
- ***Not*** statistical power
- Goal is to converge on an answer and minimize the impact of extreme random numbers



ON P-VALUES

- Observational Study:
  - *f*(Sample Size, Effect Size, Test,α)
  - All but sample size essentially fixed
  - Sample size is hard to increase – limited source population, recruitment is hard and expensive
- Simulation Study:
  - All those factors
  - But what determines simulation sample size?
    - *f*(Computing Power, Patience)
  - Power is now something trivially modified by the researcher
    - Clusters, cloud computing, three-day weekends



- True difference: RR = 3.37 vs. 3.3701
- All it took was 100,000 runs of the model
- Average 3.37 seconds / run
- 4 processor cores = 25,000 runs / core
- ~ 7 hours of wall time
- All of that overnight

ABMS AND CAUSAL INFERENCE

- Causal inference and Agent-based models sometimes feel at odds with one another
- Different heritage, different nomenclature, etc.
- Opinion:
  - They aren't
  - More strongly: Causal inference models *are* agent-based models with a series of constraints and assumptions imposed on them
  - ABMs provide *indisputably counterfactual scenarios*
  - But those scenarios may be about a fictional universe
  - How willing are you to step outside your data?

- “A model is a lie that helps you see the truth.” – Howard Skipper
- “Data do not speak for themselves - they need context, and they need skeptical evaluation.” – Allen Wilcox
- “It is inappropriate to be concerned about mice when there are tigers abroad.” – George Box