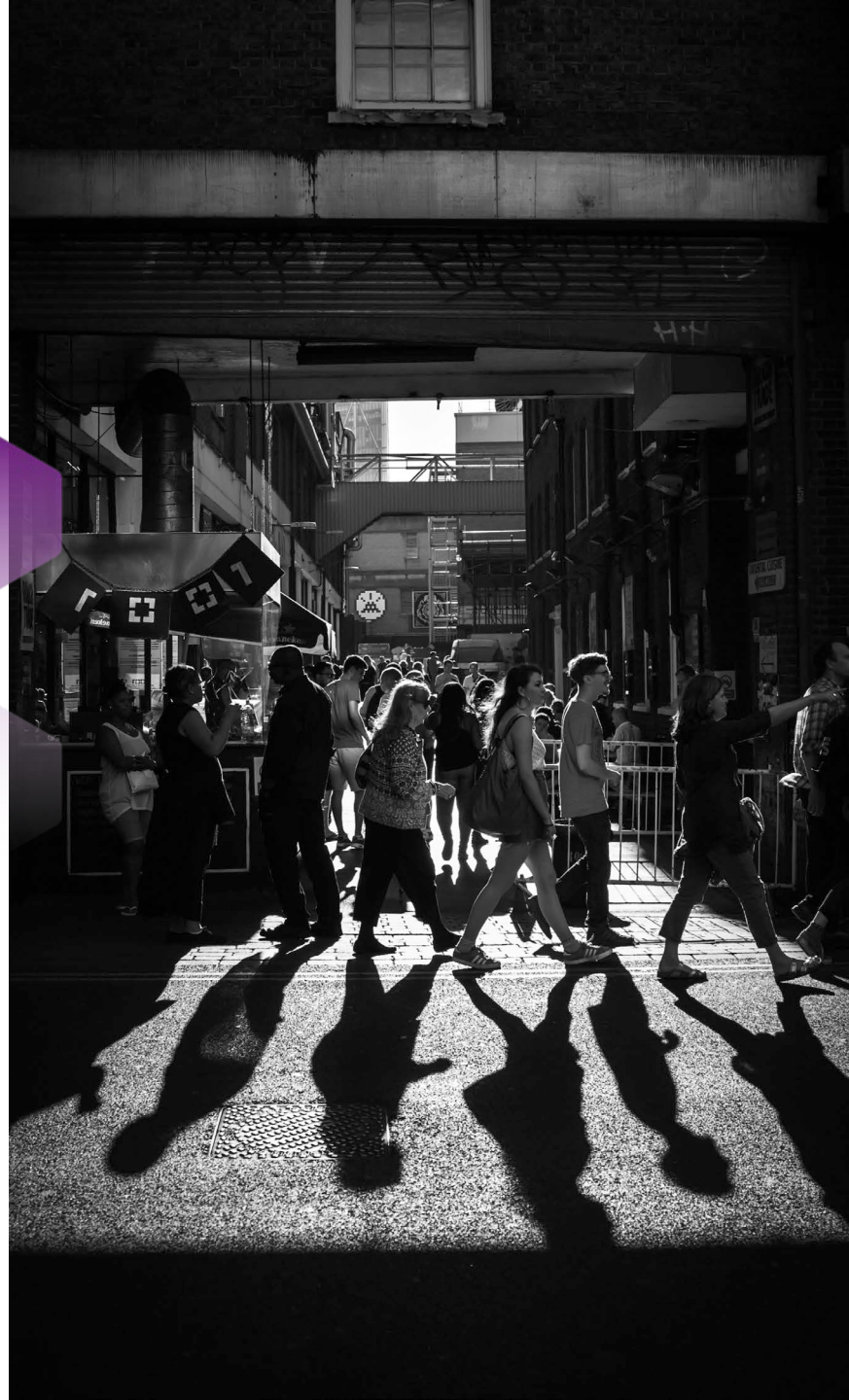
UK Data Service

Agent-based models for generating synthetic data

Dr. J. Kasmire
Computational Social Science Training Lead
UK Data Service
2026



UKRI Economic and Social Research Council

Table of Contents

What is synthetic data?

Any data that is **generated** rather than **observed**

What is fidelity?

Fidelity = faithfulness or similarity between synthetic data and real-world data

More about fidelity

Can never be 100%

Doesn’t always apply

Is not binary or even a simple low to high

Higher is not always better

Needs good documentation

Uses and generation methods for synthetic data

Preview

Code development

Proof of concept

Presentation

Availability

Remote work

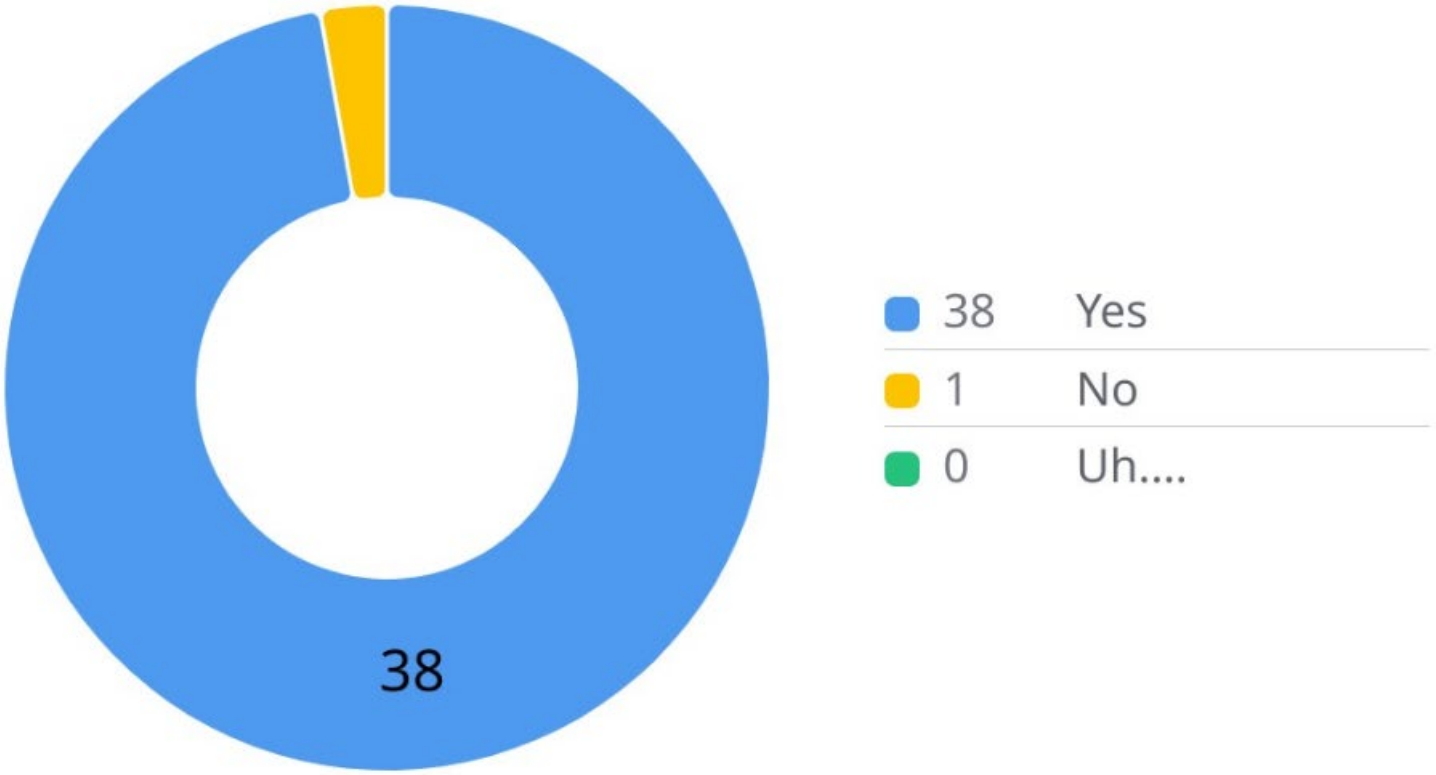
Handmade

Random/nonsense

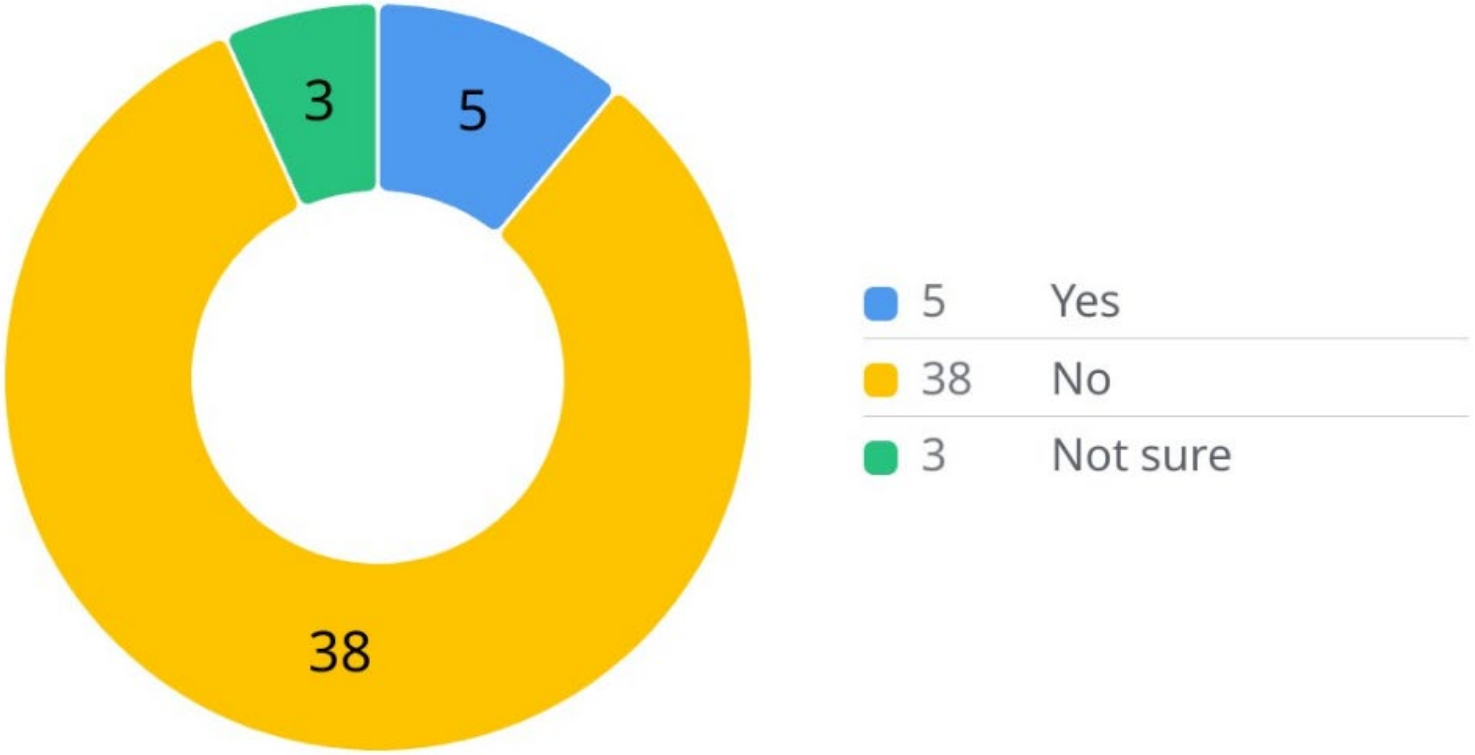
Machine learning

Simulation

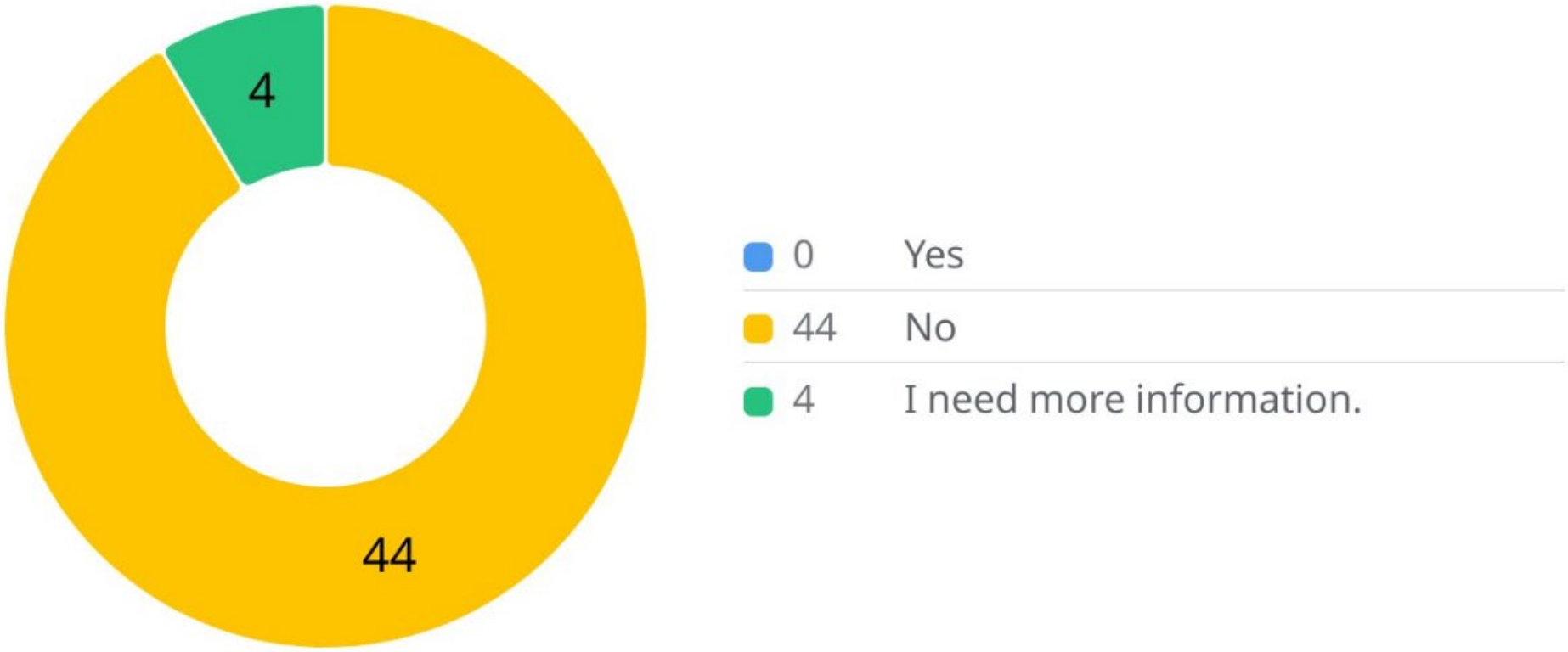
Are predictions synthetic data?



Does fidelity apply to all synth data?



Is higher fidelity always better?



Top-down vs. bottom-up systems

Systems can be top-down or bottom-up.

Top-down vs. bottom-up problems

Systems can be top-down or bottom-up.

Problems can also be top-down or bottom-up.

Systems or problems are top-down if they:

- Are ‘whole’, ‘well-understood’ and have central control or structure
- Can be broken into defined parts and interactions
- Work with ‘classic scientific method’ ideas of isolation, prediction, repetition, etc.

Systems or problems are bottom-up if they:

- Are poorly-understood, open/flowing, have partial/decentralised control
- Can’t be reduced to parts
- Have emergent behaviour
- Don’t work well with classic scientific method
- Best understood through simulation or modelling

What about data?

Think about how we might get data from an engine vs. from an ecosystem in the real-world

What about data expectations?

What are the sensors or the measurements?

Is there a “right” measurement or not?

Are these the measurements we want?

Are we checking performance or trying to understand?

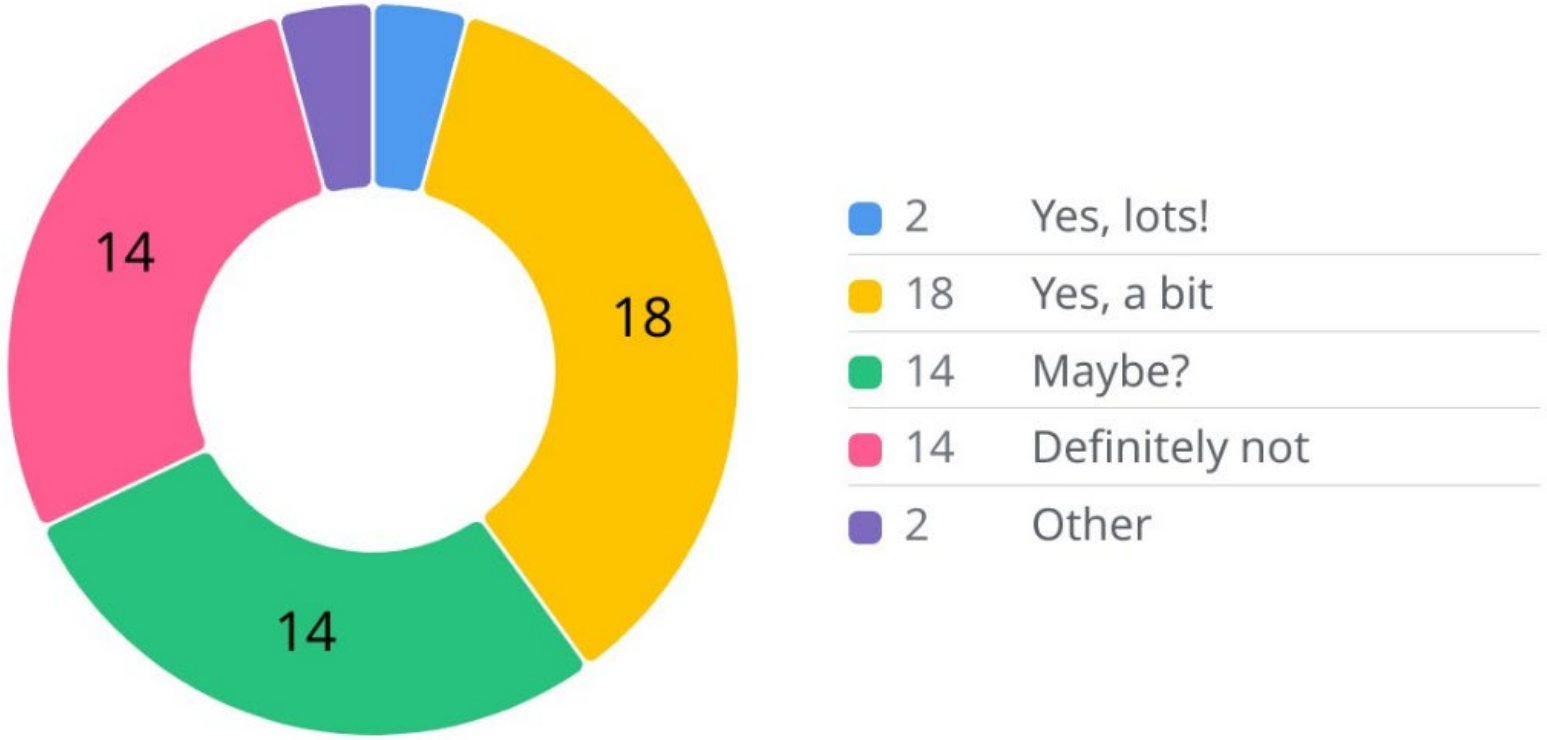
Synth data for top-down or bottom-up

Think about how to generate synthetic data for:

- an engine (top-down) and
- an ecosystem (bottom-up)

Would you use the same generation method for both?

Have you ever considered top-down vs. bottom-up before?



Suggest some top-down systems

Suggest some bottom-up systems

What are agent-based models:

ABM are simulated worlds with:

- States
- Rules
- Objects (optional)
- Agents that:
  - have their own states and rules and
  - take actions/make decisions

The World

- Can represent anything at any scale
- Proceeds through time in discrete ‘time steps’
- Includes things that matter
- Is unique for each agent as it contains everything *else* (including other agents)
- Has states and follows rules (more on this later)

The agents

The defining feature of agent-based models!

Agents

- Can represent almost anything that acts or decides
- Have things or features that matter
- Are unique and behave uniquely
- Have states and follow rules
- Take actions, make decisions per states and rules
- Can exist during the entire simulation or can “spawn” and/or “die” during it

Output

- Synthetic data
- Decided by the modeller
- Typically consists of self-reports from the world and/or the agents about their state or features
- Can be at each time step, just the end, or specified intervals throughout

Rules! Starting simple

If [weather = raining] AND [temp < 5 degrees]

[travel by bus]

Else [travel by bike]

Rules! Still simple

If [weather = raining] AND [temp < 5 degrees]

[travel by bus]

Else [travel by bike]

If [weather = snowing] AND [temp < 3 degrees]

[travel by bus]

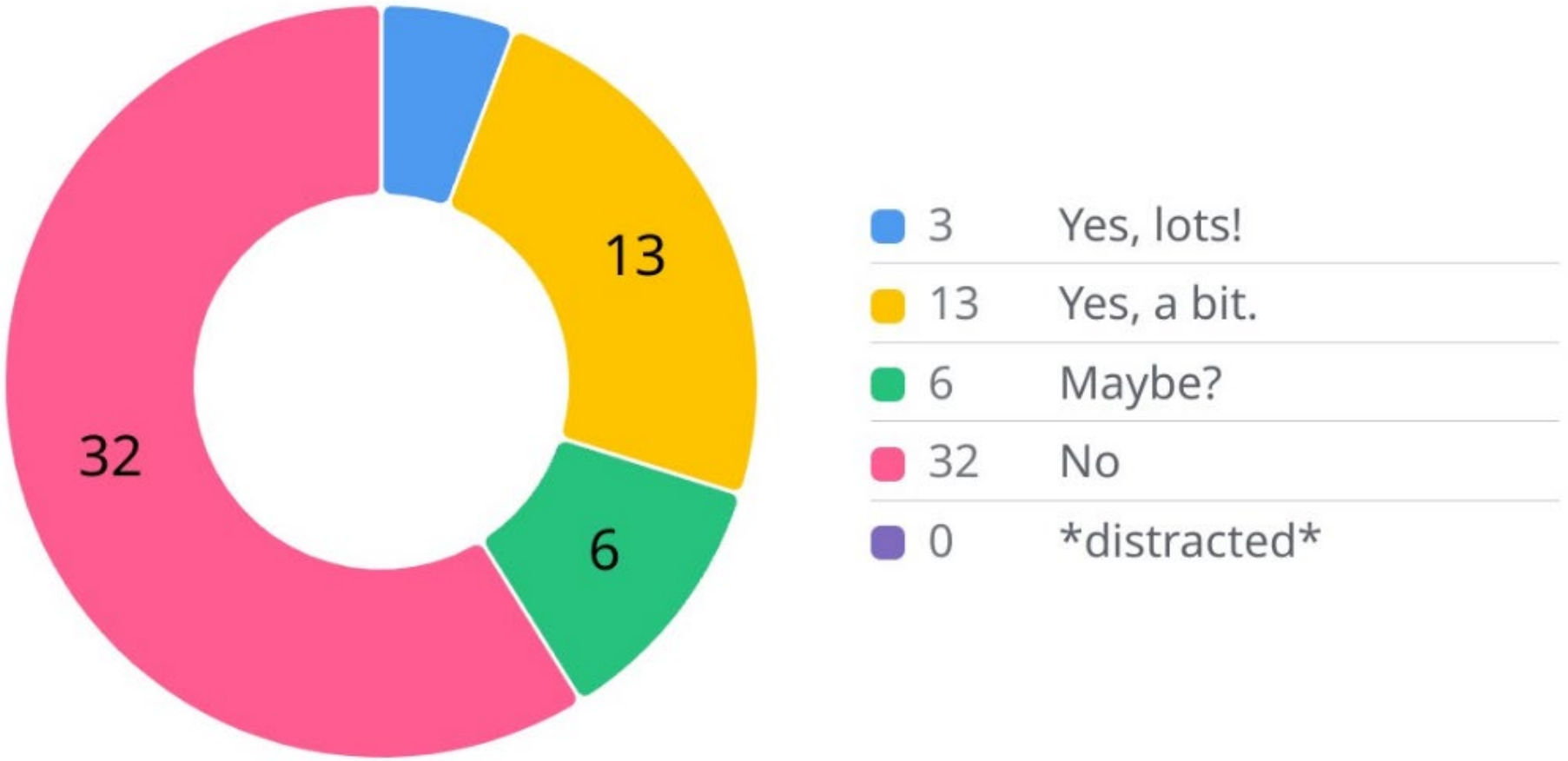
Else [travel by bike]

Rules! Adding feedback

1) If [weather = raining] AND [temp < 5 degrees]
   [travel by bus]
   Else [travel by bike]

2) If [yesterday's transport = success]
   [Follow normal rule in number 1]
   Else [copy travel of nearest agent who had a successful yesterday's transport]

Have you ever used ABM/simulation before?



What ABM might you want to build/use?

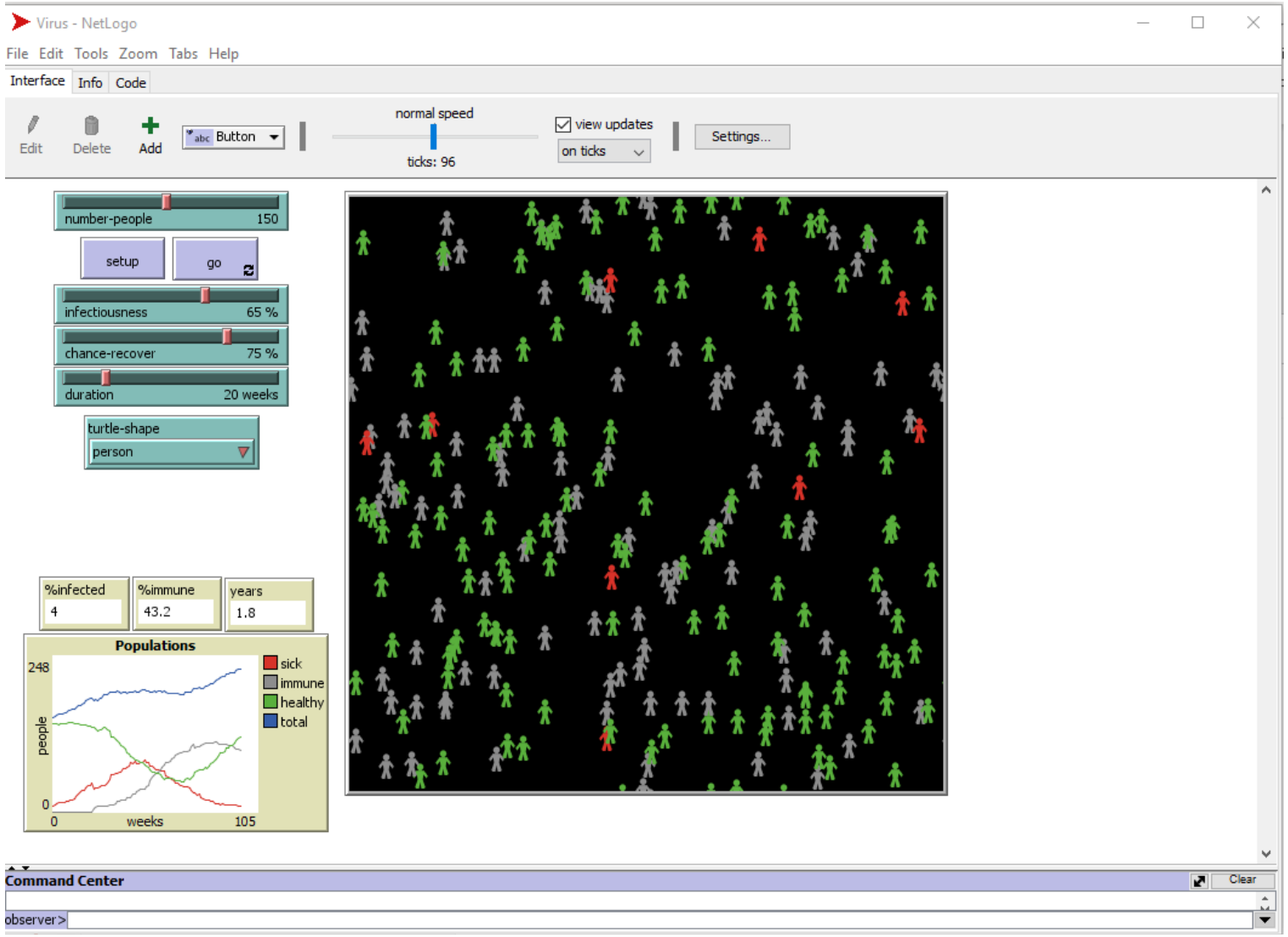
Audience responses included:

- Policy response, urbanisation, geopolitics, public sentiment shifts, tax benefits
- Queuing, emotional responses to service failures, public toilets
- Adaptive leadership, change and transformation, idea generation
- Crime prediction, crime prevention,
- Healthcare use, disease transmission, food use, smoking, childcare provision
- Student behaviour, fatigue, social networks
- Home working, dynamic energy use, active travel, data/telecoms use
- Consumer behaviour, household decision making, product purchases
- Human and animal populations, migration, protected areas, conservation
- Kung fu robots and world domination

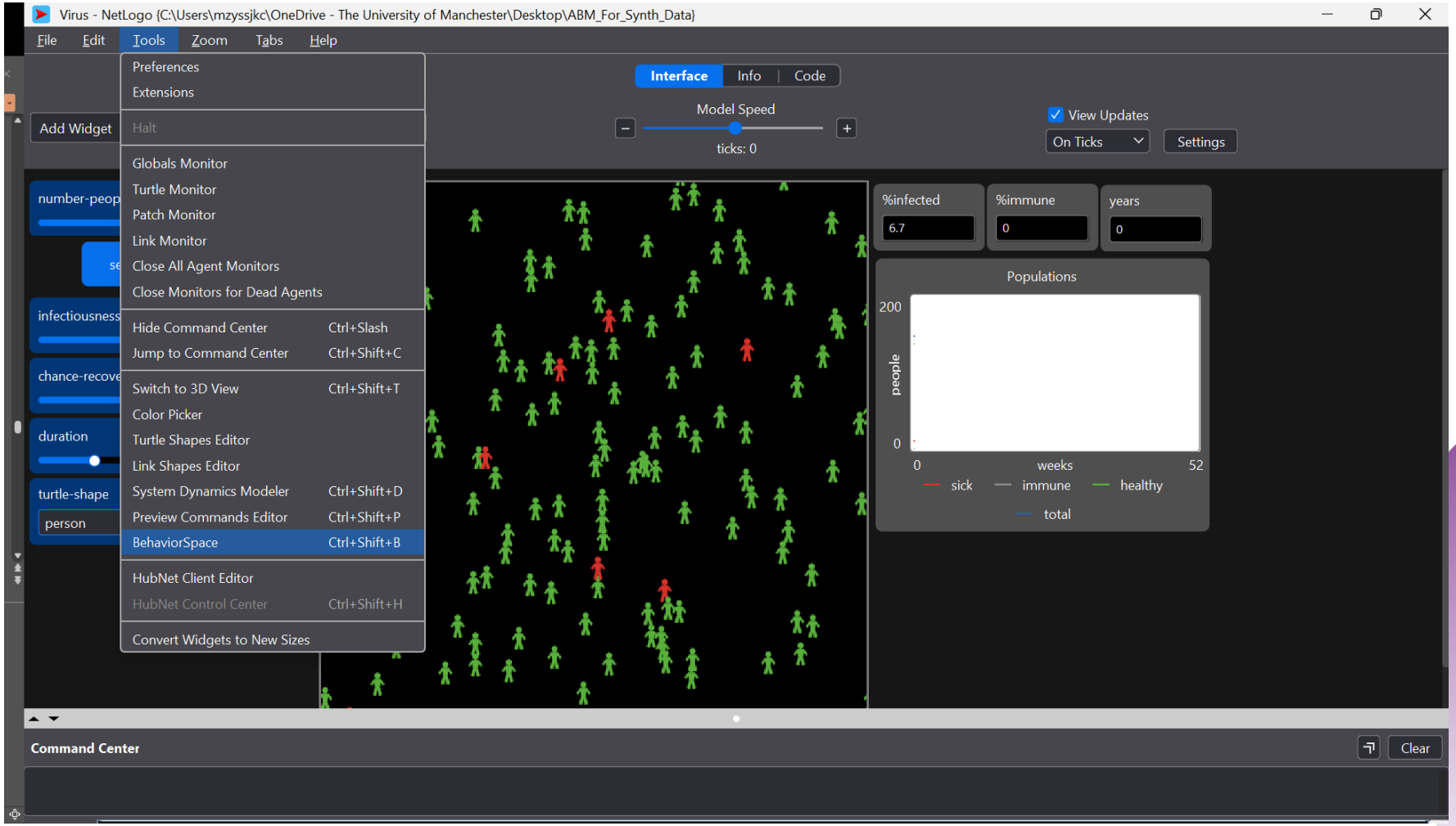
Let’s look at some examples!

- A virus model including behaviour space, experiments, and output
- 2 to 4 more models (depending on time available)

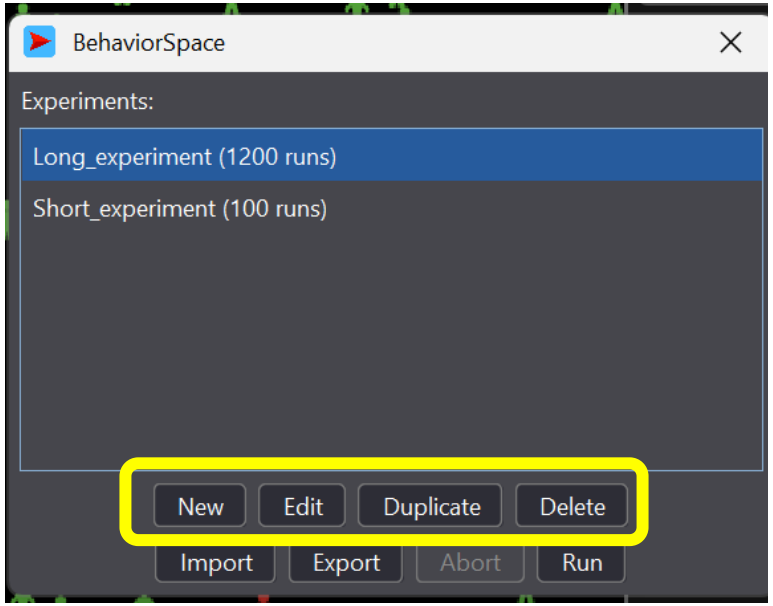
A simple example - Virus

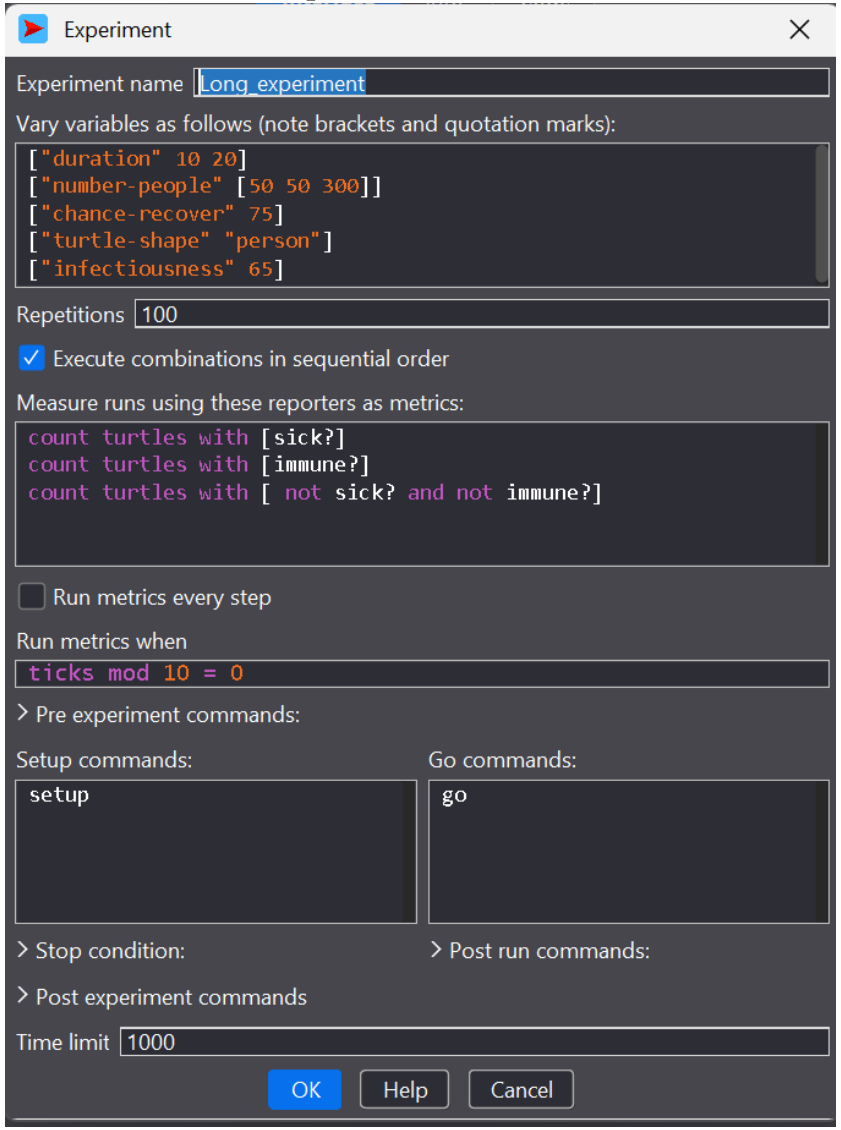


Virus – open behaviour space

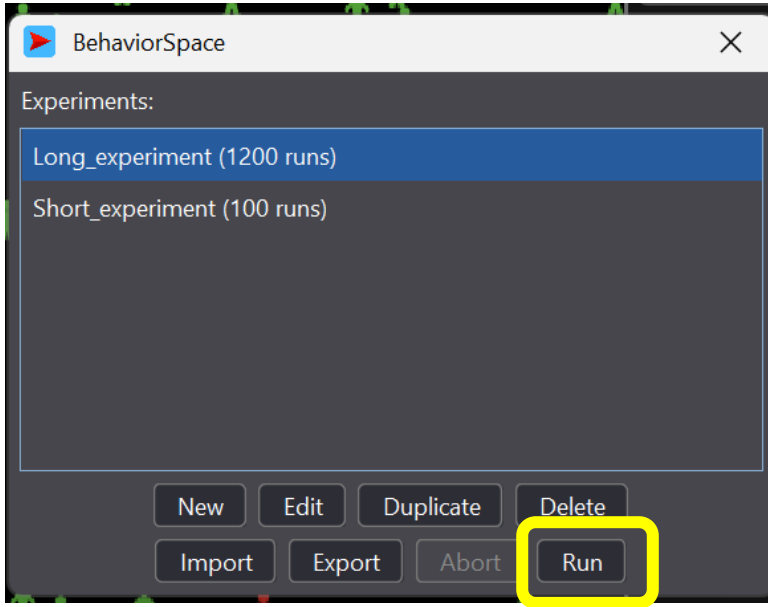


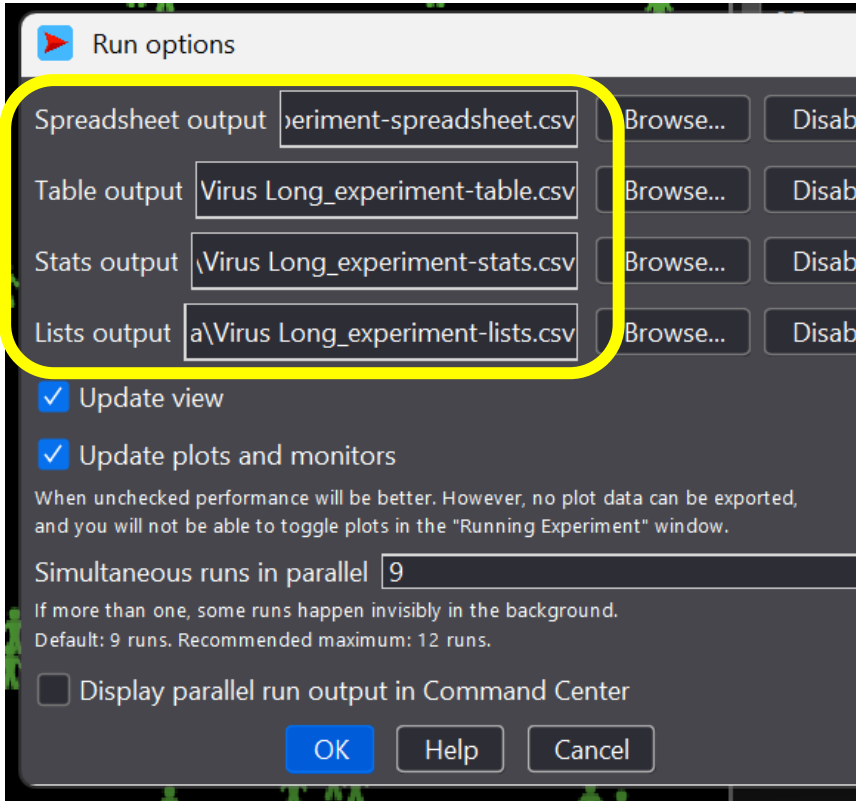
Virus – create / edit experiments



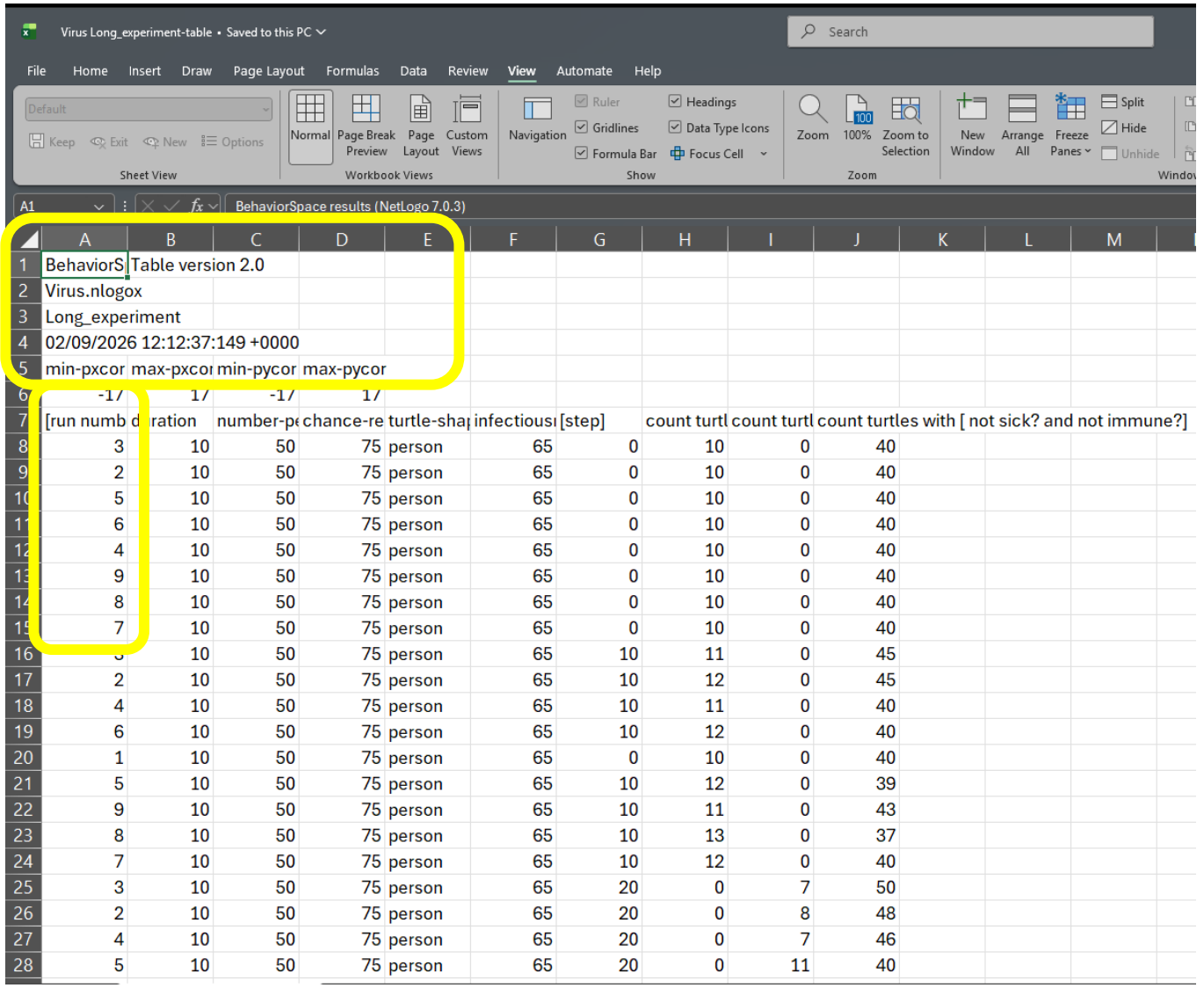


Virus – run experiments



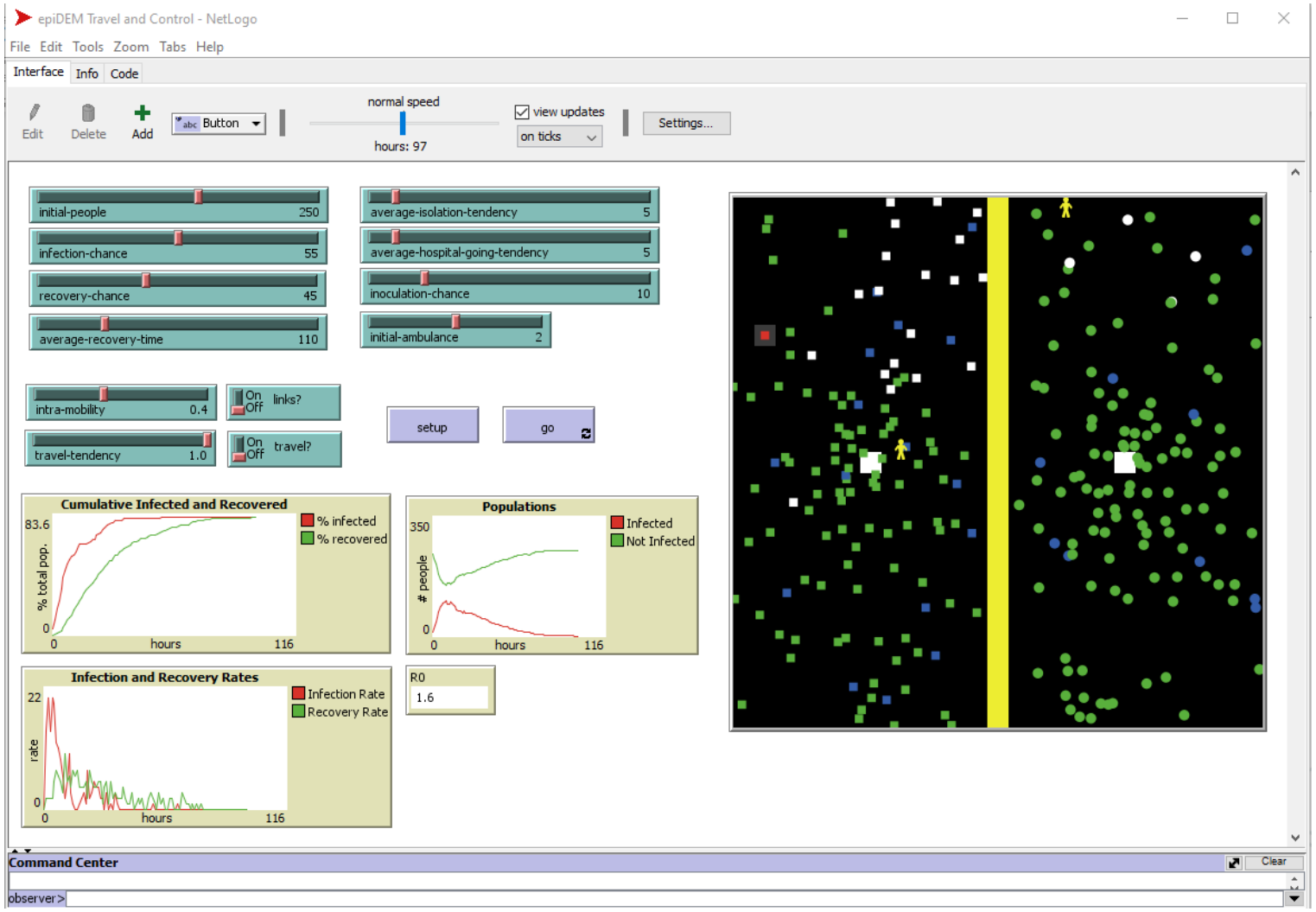


Virus – experiments output



	A	B	C	D	E	F	G	H	I	J
1	BehaviorS	Table version 2.0								
2	Virus.nlogox									
3	Long_experiment									
4	02/09/2026 12:12:37:149 +0000									
5	min-pxcor	max-pxcor	min-pycor	max-pycor						
6	-17	17	-17	17						
7	[run numb	duration	number-pe	chance-re	turtle-sha	infectious	[step]	count turtl	count turtl	count turtles with [not sick? and not immune?]
8	3	10	50	75	person	65	0	10	0	40
9	2	10	50	75	person	65	0	10	0	40
10	5	10	50	75	person	65	0	10	0	40
11	6	10	50	75	person	65	0	10	0	40
12	4	10	50	75	person	65	0	10	0	40
13	9	10	50	75	person	65	0	10	0	40
14	8	10	50	75	person	65	0	10	0	40
15	7	10	50	75	person	65	0	10	0	40
16		10	50	75	person	65	10	11	0	45
17	2	10	50	75	person	65	10	12	0	45
18	4	10	50	75	person	65	10	11	0	40
19	6	10	50	75	person	65	10	12	0	40
20	1	10	50	75	person	65	0	10	0	40
21	5	10	50	75	person	65	10	12	0	39
22	9	10	50	75	person	65	10	11	0	43
23	8	10	50	75	person	65	10	13	0	37
24	7	10	50	75	person	65	10	12	0	40
25	3	10	50	75	person	65	20	0	7	50
26	2	10	50	75	person	65	20	0	8	48
27	4	10	50	75	person	65	20	0	7	46
28	5	10	50	75	person	65	20	0	11	40

A complex example - Virus



To the simulation-ator!

1 Virus

Simple https://tinyurl.com/5cd7hfa8

Travel and control https://tinyurl.com/4zxywzd5

2 Flocking https://tinyurl.com/2euny6a5

3 Giant component https://tinyurl.com/tps24rhc

4 Rebellion https://tinyurl.com/466n65wf

If you want to generate synth data via ABM

- Check ABM makes sense for this system/problem
- Consider the pros and cons
- Identify the world, states, agents, objects, rules
- Plan the data output and parameters

Pros of ABM

Pros

- No intuition needed
- Allows mental model inspection
- Relatively cheap/easy/fast
- ‘Impossible’ data
- Can be natural to explain

Pros and cons of ABM

Pros

- No intuition needed
- Allows mental model inspection
- Relatively cheap/easy/fast
- ‘Impossible’ data
- Can be natural to explain

Cons

- Tricky to reflect critically
- Capturing abstract concepts is difficult
- Treated as “truthiness” or Garbage in/garbage out
- Hard to verify
- Hard to use chaotic/complex results

Sharing ABM data

- Output is synthetic
- Model code and/or output can be shared if no licensed data used as input or within the model
- ALWAYS check the terms of use for any licensed data that you use
- Consider sharing model code without input data or with dummy input data
- https://ukdataservice.ac.uk/app/uploads/cd137-enduserlicence.pdf

Platforms, languages, etc.

Free, open-source, small download size, tutorials and info available on web

- Mason – easy to learn, not well recognised
- Repast – hard to learn, more powerful
- NetLogo – built in visualisation, not so powerful
- EMLab-Agentspring – modular, not well recognised
- Object-oriented software (e.g. Python, Julia) – hard to learn, powerful

Resources for learning more

Wikipedia articles

https://en.wikipedia.org/wiki/Comparison_of_agent-based_modeling_software

https://en.wikipedia.org/wiki/Agent-based_model

Online ABM repository https://www.comses.net/

NetLogo Resources

https://ccl.northwestern.edu/netlogo/docs/dictionary.html

https://ccl.northwestern.edu/netlogo/docs/tutorial1.html

Agent-based models for socio-technical systems https://link.springer.com/book/10.1007/978-94-007-4933-7

Agent-Based Modelling and Geographical Information Systems https://www.abmgis.org/

Thanks!

Dr Jools Kasmire

j.kasmire@manchester.ac.uk

@JKasmireComplex@mastodonapp.uk

@jools-cyborg.bsky.social

UK Data Service