Simulating Language
1: Introduction

Kenny Smith
kenny.smith@ed.ac.uk

Why simulate language

Some important questions...

What is the goal of linguistics?

To answer the question:
Why is language the way it is?

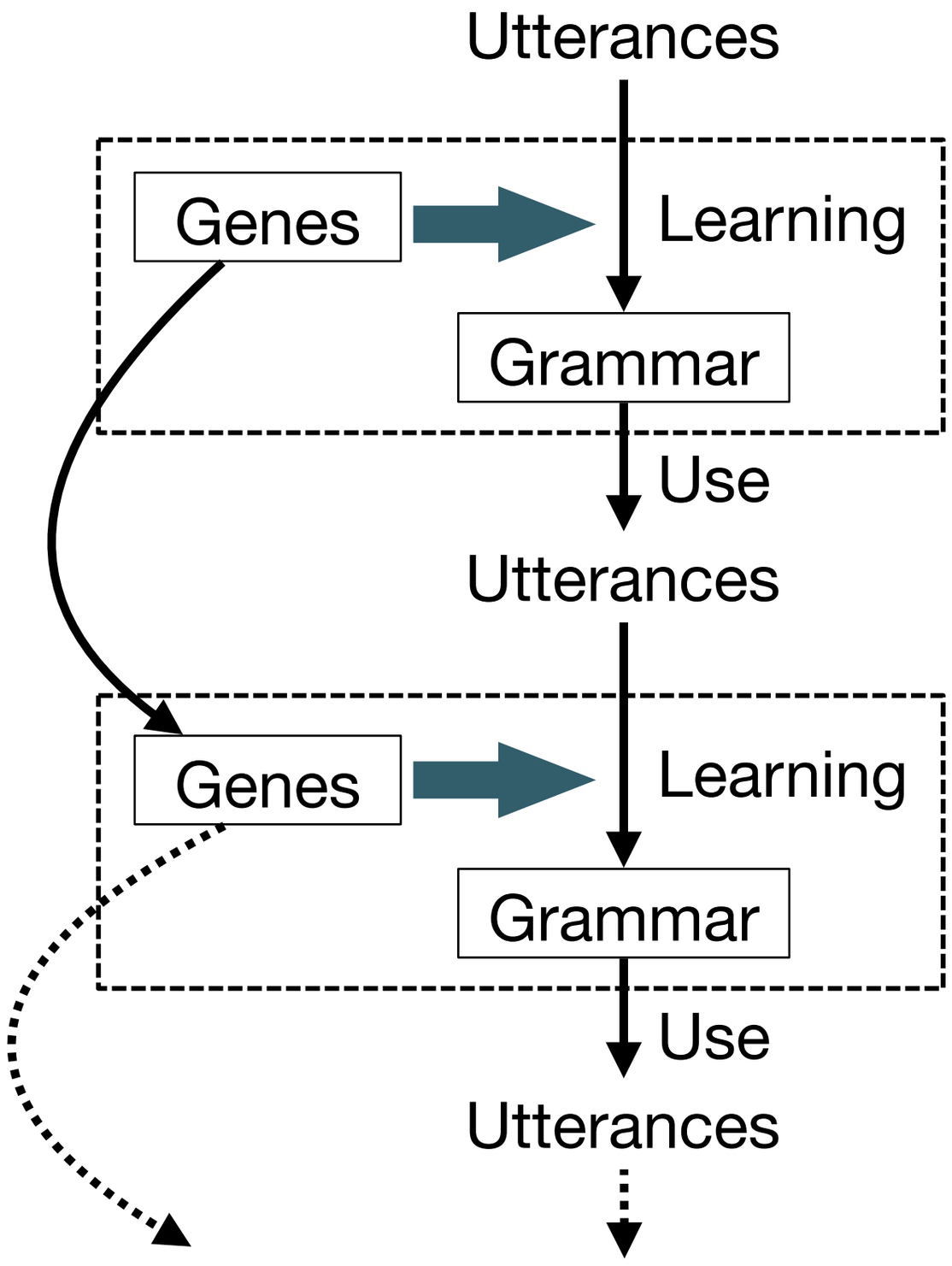
OK... but what should we do to approach this question?

The evolutionary approach to language

We answer the **why question** by posing a **how question.**

We can figure out **why** is language the way it is, by understanding **how** it came to be that way.

The **processes** that shape language



Language learning

Language use

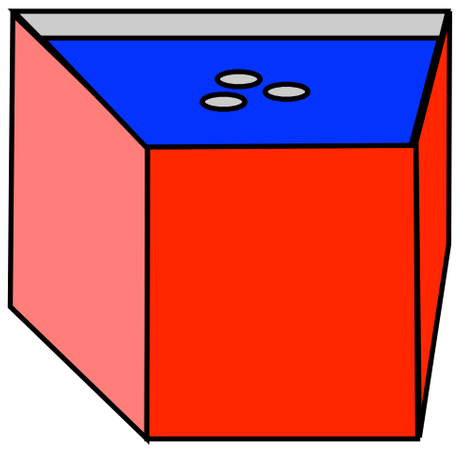
Cultural evolution

Biological evolution

How on earth do we study this?!
We are going to build **models**

Reading recap: What is a model?

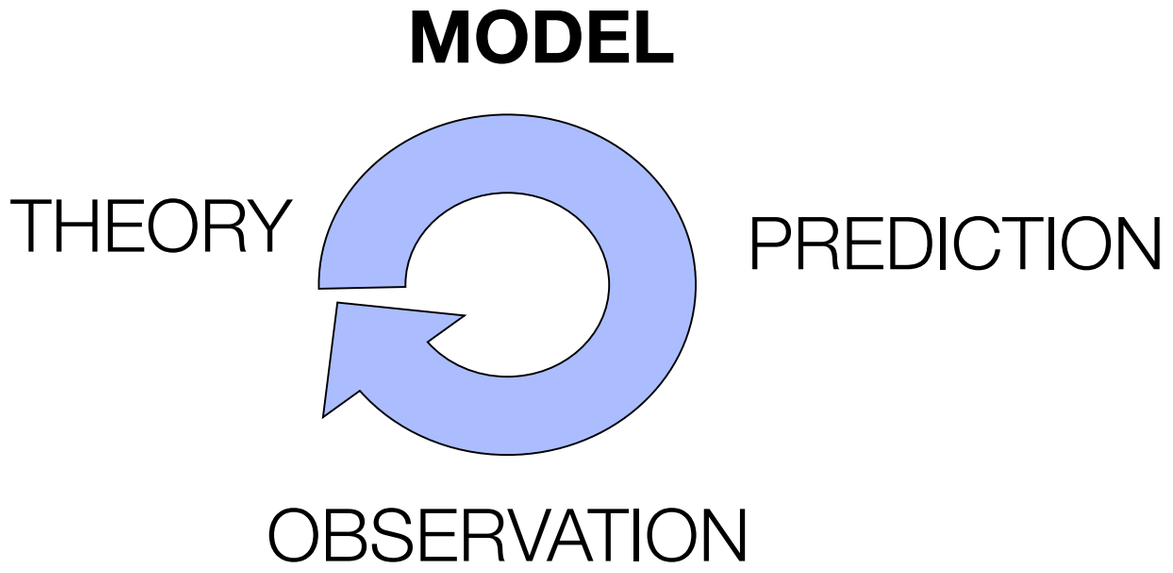
- A miniature version of the system you are interested in
- Gains
  - Simplicity
  - Control
  - Ease of observation
- Losses: ?



Reading recap: What is a model for?

- One possibility: We use models when we can't be sure what our theories predict

- Especially useful when dealing with *complex systems*



- An alternative / complementary approach: models as tools for understanding

  "the *insights* offered by a model are at least as important as its *predictions*: they help in understanding things by playing with them." (Sigmund, 1995, *Games of Life*, p. 4)

Computational modelling is the solution

- Computers are very good for models of many interacting components
- This has proved particularly valuable in allowing us to build a fundamentally *evolutionary* approach to understanding language
- In this course, we will be building and playing with models to tackle questions like:
  - How do data and learners' prior biases interact to shape linguistic knowledge?
  - Why do languages have grammars?
  - What would it mean to say that language is innate?

Course outline

Lectures

Kenny Smith (me)

Simon Kirby

Course organiser: Simon Kirby
Lab director: Matt Spike

Labs

Matt Spike

Annie Holtz

Henry Conklin

Claire Graf

Weekly schedule

- All details on website: https://centre-for-language-evolution.github.io/simlang2021/index.html

- 3 hours per week contact time
  - Tuesday: lecture
    - Preparatory work (readings, quizzes) sometimes required before lectures
  - Thursdays/Friday: **labs**
    - Labs are the primary source of feedback throughout the course, opportunity for one-to-one help and discussion.

This is a **practical** course: lots of time playing with code, working with simulations

---

- You do not need to know how to program, but you do need not to be scared of computers, and willing to try things out. You do not need any fancy maths, but you do need not to be scared of a little arithmetic and seeing the occasional equation (starting today!)

- We will be working in a simplified subset of **Python.** We will supply the code for the practicals, but you will need to modify it to carry out the tasks on the interactive notebooks. You can do this from anywhere you have access to a web-browser.

- This isn’t a programming course: we aren’t going to teach you how to program, but we will teach you just enough to understand and use some simple models we provide. You will have to meet us half way: you’ll get on much better if you get your hands dirty and try things out.

Lab classes

- Labs take place on Gather.Town
- You will be allocated to a timeslot and a tutor today
- Info on how to access gather.town will appear on Teams; go to your lab class at the allocated time and get help with the programming practicals
- Two ways to use the labs
  - as drop-ins
  - **a dedicated time to work on the practicals with help/company at hand**

Assessment

- All aspects of the course could be assessed (i.e. every lecture, every lab)
- Two assignments, each consisting of short answer questions.
- Schedule of assessment release, deadlines, feedback dates will appear on the course webpage
- For UG students, each worth 50%. For PG students, second worth 70%.

Getting started: introducing Bayes

A medical quiz

- Your friend complains of a headache. Is this headache caused by:

A. A brain tumour

B. Dehydration

C. Athlete’s foot

- Resolving this question requires you to draw on two probabilities:
  - How likely is it that someone with the medical condition in question would exhibit that symptom?
  - How common is each medical condition?

Likelihood of symptoms given underlying condition

Brain tumour: headache is very likely, if you have a brain tumour

Dehydration: headache is very likely, if you are dehydrated

Athlete's foot: a headache is very very unlikely to be caused by athlete's foot

- If all we care about are the likelihood of the symptoms given each underlying condition, we would conclude that your friend either has a brain tumour or is dehydrated

Probability of underlying conditions

Brain tumours: very rare

Dehydration: very common

Athlete's foot: very common (let's say)

- If all we care about is the prevalence of each underlying condition, we would conclude that your friend is either dehydrated or suffering from athlete's foot

- But you didn't conclude this: you brought these two quantities together in a smart way. How did you do it?

The Bayesian approach

- What you're trying to figure out is the probability that your friend has a particular medical condition or illness, given the symptoms they are exhibiting. We call this quantity "the probability of the illness given the symptoms", or: $$P(illness|symptoms)$$
- We are trying to work this out based on two quantities which we know (roughly):
  - The **likelihood** of exhibiting a particular symptom given that you have a certain illness $$P(symptoms|illness)$$
  - The **prior** probability of each illness $$P(illness)$$

Bayes' rule

- Bayes' rule provides a convenient way of expressing the quantity we want to know in terms of the quantities we already know:

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- Or, in full:

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

Breaking it down

$$P(illness|symptoms) = \frac{P(symptoms|illness)P(illness)}{P(symptoms)}$$

$P(illness|symptoms)$

- The thing we want to know is called the **posterior**

$P(symptoms|illness)$

- The probability of a particular set of symptoms given that you have a specific illness is called the **likelihood**

$P(illness)$

- The probability that you have a particular illness, before I have any evidence from your symptoms, is called the **prior**

$P(symptoms)$

- The term on the bottom (the probability of the symptoms independent of illness) is actually not very interesting to us, since it is the same for all illnesses.

It makes intuitive sense...

$$P(illness|symptoms) \propto P(symptoms|illness)P(illness)$$

- If the likelihood of symptoms given a certain illness is high, this will increase the posterior probability of that illness
- If the prior probability of a certain illness is high, this will increase the posterior probability of that illness
- If a particular illness has low prior probability, we need some really convincing evidence to make us believe it to be true

Errr... hello... isn't this a course about language?

- In the headache example, we were trying to use evidence provided by symptoms to infer what your friend's underlying problem was

- What if you aren't an amateur medic, but a child hearing utterances from a parent and learning a language? You are trying to use evidence provided by utterances to infer what grammar your parent has in their head

  illnesses = languages
  symptoms = utterances
  prior for each illness = **prior for particular (types of) language**

- An ideal language learner will estimate the posterior probability of each possible language given the utterances heard

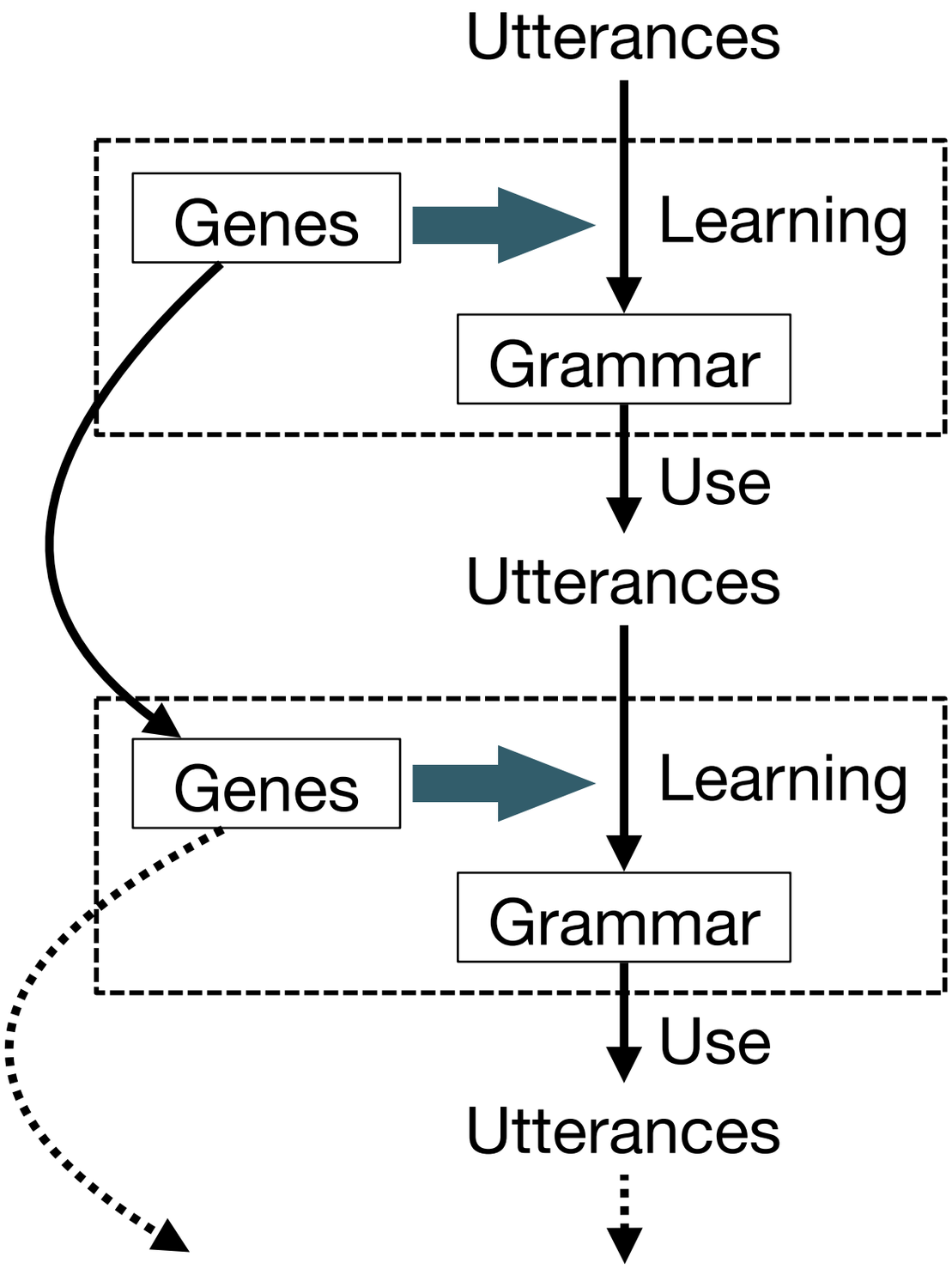
- Children probably don't calculate sums in their head while learning, but if their learning process is sensible, we can characterise it this way

Bayesian language learning

- Evaluate hypotheses about language given some prior bias (perhaps provided by your biology?) and the data that you've heard

$$P(h|d) \propto P(d|h)P(h)$$

- h = hypothesis about the language
- d = linguistic data


Utterances
Genes
Learning
Grammar
Use
Utterances
Genes
Learning
Grammar
Use
Utterances

Bayesian language learning $P(h|d) \propto P(d|h)P(h)$

- This modelling approach provides several advantages for our purposes
  - Quantitative (we can put numbers on stuff)
  - Simple (just multiplying and dividing)
  - Transparent (nice clean representation of the role of **prior knowledge**)
  - Surprisingly powerful (as we'll see)

Coming up next!

- **Thursday and Friday:** labs, introduction to python, noteable, etc
- Next week: a Bayesian model of word (concept) learning
    - Reading and a pre-lecture quiz required before next week's lecture!