LAWS OF PROBABILITY

Chain rule of probability:

$$\Pr_{A,B,C}(a,b,c) = \Pr_A(a)\,\Pr_B(b|A=a)\,\Pr_C(c|A=a,B=b)$$

shorthand: p(a,b,c) = p(a) p(b|a) p(c|a,b)

Law of total probability:

$$\Pr_A(a) = \sum_b \Pr_{A,B}(a,b) = \sum_b \Pr_A(a|B=b)\,\Pr_B(b)$$

$$\Pr_A(a|C=c) = \sum_b \Pr_A(a|B=b,C=c)\,\Pr_B(b|C=c)$$

called "LoTP with baggage $\{C=c\}$"

All the laws of probability "just work" with baggage...

$p(a, \overline{b,c}) = p(a)\, b(b,c|a)$

by defn. of cond prob.

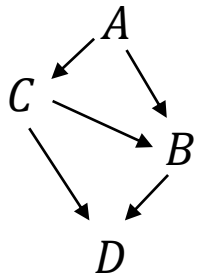
$= p(a)\ p(b|a)\ p(c|b,a)$

by cond. prob. with baggage a.

Thus, the chain rule is derived from "conditional probability, with and without baggage."

CAUSAL DIAGRAMS

The causal diagram tells us how a joint distribution factorizes: e.g.



$$p(a,b,c,d) = p(a)\,p(c|a)\,p(b|a,c)\,p(d|b,c)$$

STRATEGY

- Try to get the variables in their causal order
  so we can simplify by using the structure of the causal diagram

Example (extension of Bayes's rule).
In the causal diagram $A \to B \to C$, calculate $p(b|a,c)$.

$$p(b|a,c) = \frac{p(b,a,c)}{p(a,c)} \quad \text{by defn. of cond. prob}$$

$$= \frac{p(a)\,p(b|a)\,p(c|b)}{\sum_{b'} p(a,b',c)} \quad \text{by causal diag}$$

$$\text{by LoTP / sum rule}$$

$$= \frac{\cancel{p(a)}\;p(b|a)\,p(c|b)}{\sum_{b'} \cancel{p(a)}\;p(b'|a)\,p(c|b')}$$

cf Bayes Rule

$B \to C$

$$p(b|c) = \frac{p(b)\,p(c|b)}{\sum_{b'} p(b')\,p(c|b')}$$

Example (Multi-step transition probabilities).

Consider a Markov chain $X_0 \to X_1 \to \cdots$ with transition matrix P. Calculate $\mathbb{P}(X_2 = x_2 | X_0 = x_0)$.

This means: P is a matrix, and $\mathbb{P}(X_{n+1} = y | X_n = x) = P_{xy}$

STRATEGY

- Try to get the variables in their causal order
  so we can simplify by using the structure of the causal diagram

$$\mathbb{P}(X_2 = x_2 \mid X_0 = x_0) = \sum_{x_1} \mathbb{P}(X_2 = x_2 \mid X_1 = x_1, X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)$$

by LoTP with baggage $\{X_0 = x_0\}$

$$= \sum_{x_1} \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)$$

by causal diagram / memorylessness

$$= \sum_{x_1} P_{x_1 x_2} P_{x_0 x_1}$$

by defn. of P

$$= [P^2]_{x_0 x_2}$$

The Memorylessness Theorem

For a Markov chain $X_0 \to X_1 \to X_2 \to \cdots$,
conditional on the present, the future is independent of the past.

future present past

$$\mathbb{P}(X_3 = x_3 \mid X_2 = x_2, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_2 = x_2)$$

future present past

$$\mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_1 = x_1)$$

future present past

$$\mathbb{P}(X_3 = x_3 \mid X_2 = x_2, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_2 = x_2)$$

Proof: by induction, using LoTP with baggage.

Hidden Markov model



$H_1 \longrightarrow H_2 \longrightarrow H_3 \longrightarrow \cdots$ *hidden 'true' state*

$\downarrow \quad \downarrow \quad \downarrow$

$X_1 \quad X_2 \quad X_3 \quad \cdots$ *noisy observed sequence*

The full likelihood is

$$p(h_1, x_1, \ldots, h_n, x_n) = p(h_1) \left(\prod_{i=2}^{n} p(h_i | h_{i-1}) \right) \left(\prod_{i=1}^{n} p(x_i | h_i) \right)$$

by reading off the causal diagram

The marginal likelihood of an observed sequence is

$$p(x_1, x_2, \ldots, x_n) = \sum_{h_1, \ldots, h_n} p(h_1, x_1, \ldots, h_n, x_n)$$

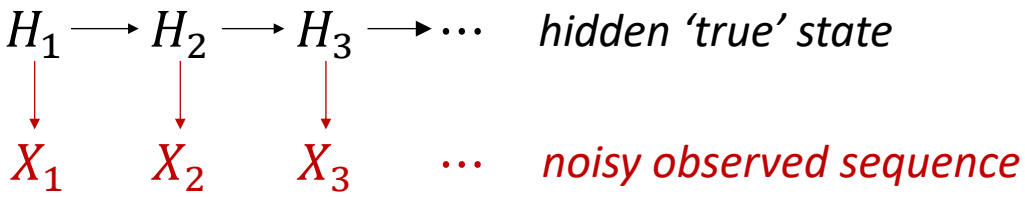
by the law of total probability

If we have a dataset of observed sequences, we could (in principle) estimate all the transition probabilities, using the marginal likelihood formula

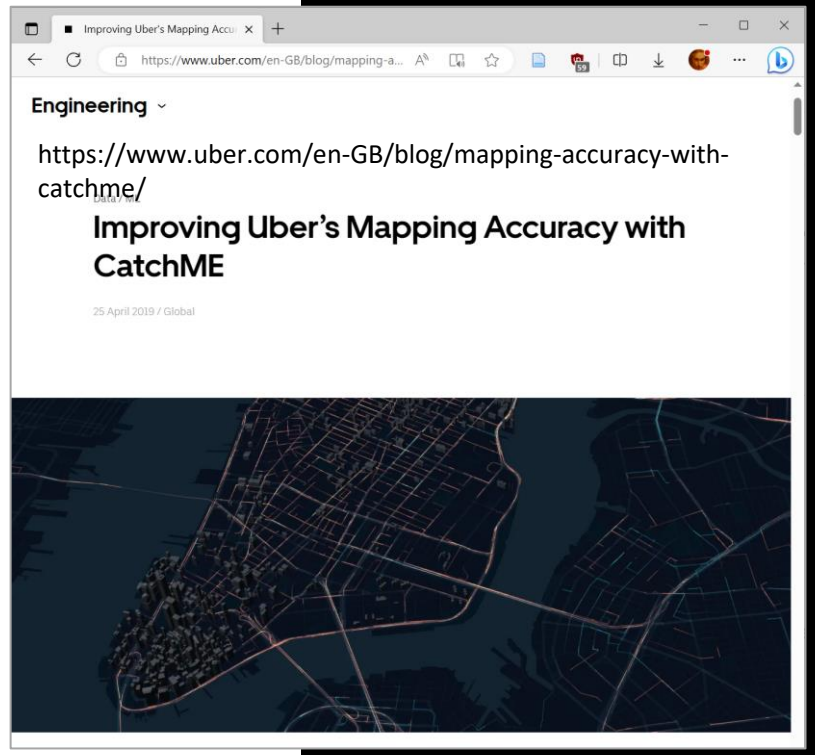
using maximum likelihood estimation

but the likelihood formula has a yucky sum, so we don't want to do this

Hidden Markov model







The practical use of HMMs is:

1. Gather a dataset of hidden+observed sequences,
   and fit the transition probabilities using the full likelihood formula
2. For a new unit with only observed data $x_1, \ldots, x_n$, we want $(H_n | x_1, \ldots, x_n)$
   - Viterbi: cunning trick to find the most likely value of H_n
     [only works for discrete state spaces]
   - Probabilistic approach: calculate the full posterior distribution of H_n
     [see Example Sheet 4]

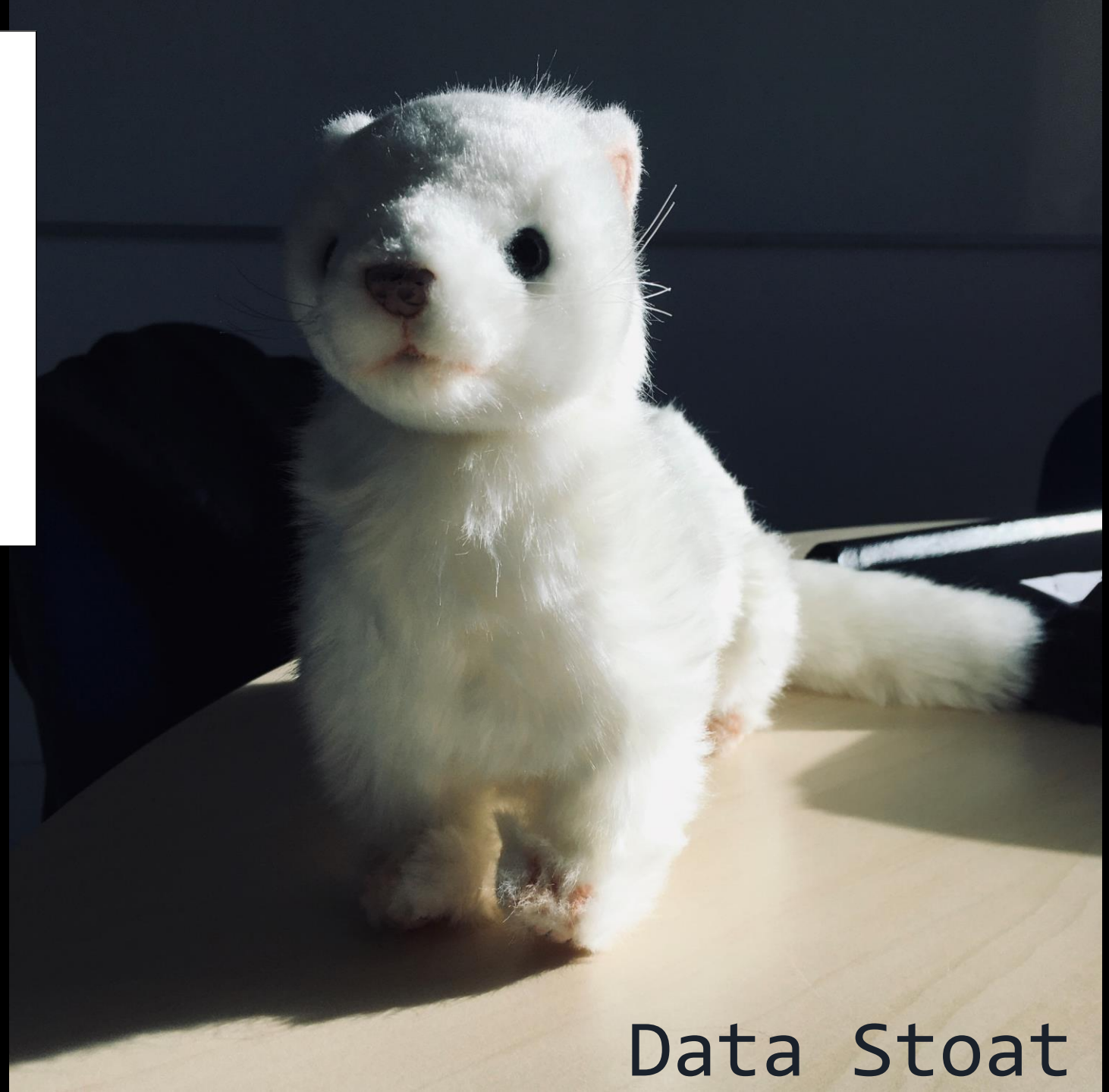
Challenge.

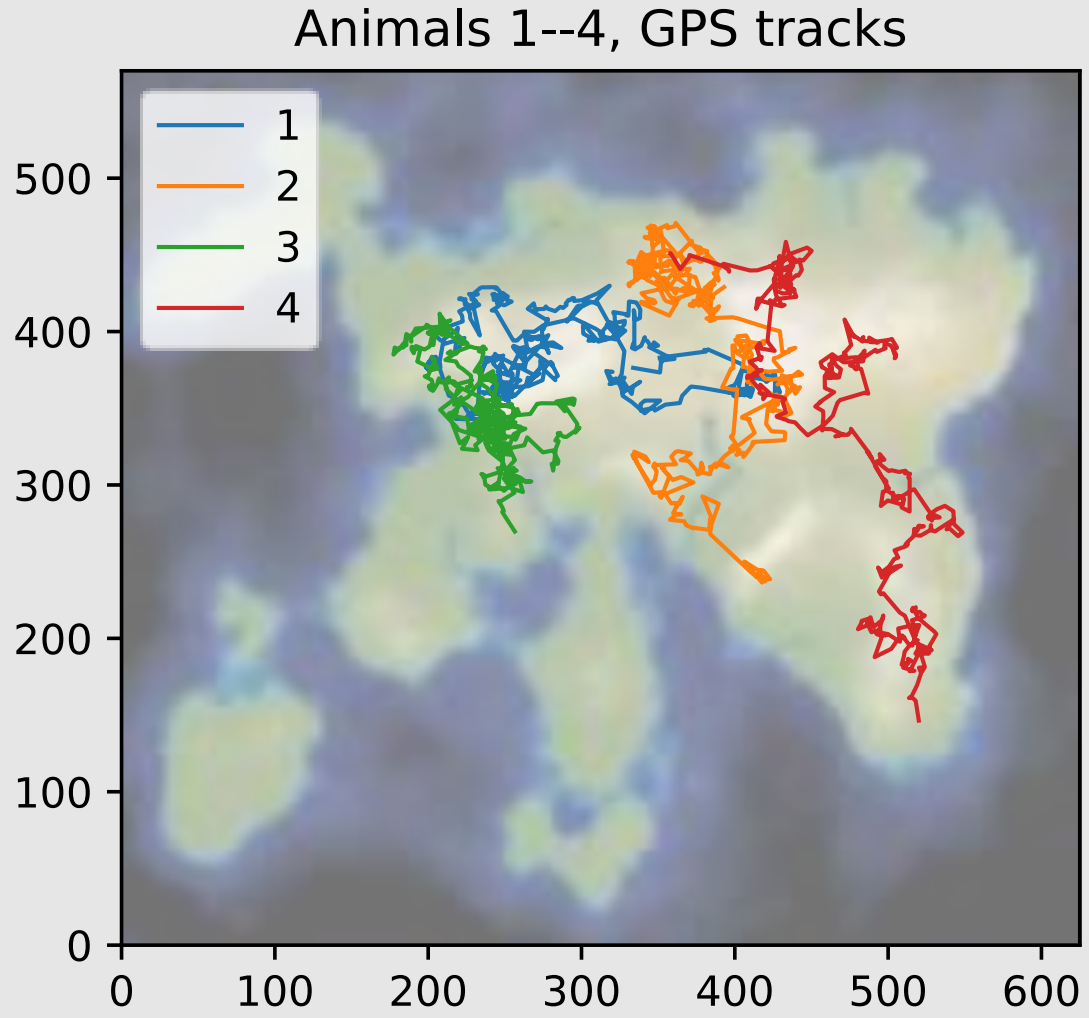
Data Stoat has gone missing!

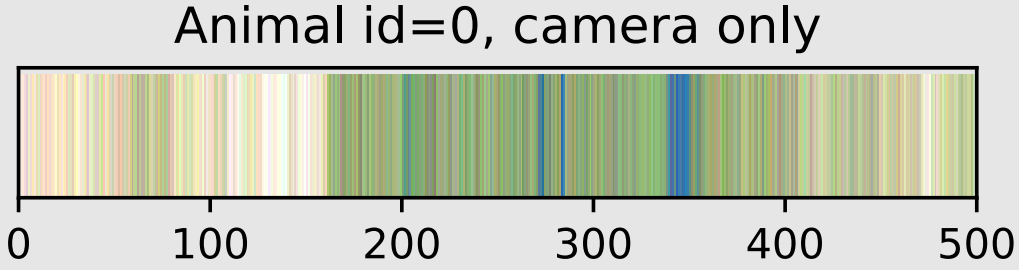
The GPS sensor that he normally carries has stopped working. But he still has a low-res camera with mobile uplink, so we know what sort of environment he's in.

We also have full data from some of Data Stoat's friends, including Correlation Weasel and Bayes Ferret.

Can you help find Data Stoat?




Animals 1--4, GPS tracks
1
2
3
4
0
100
200
300
400
500
600

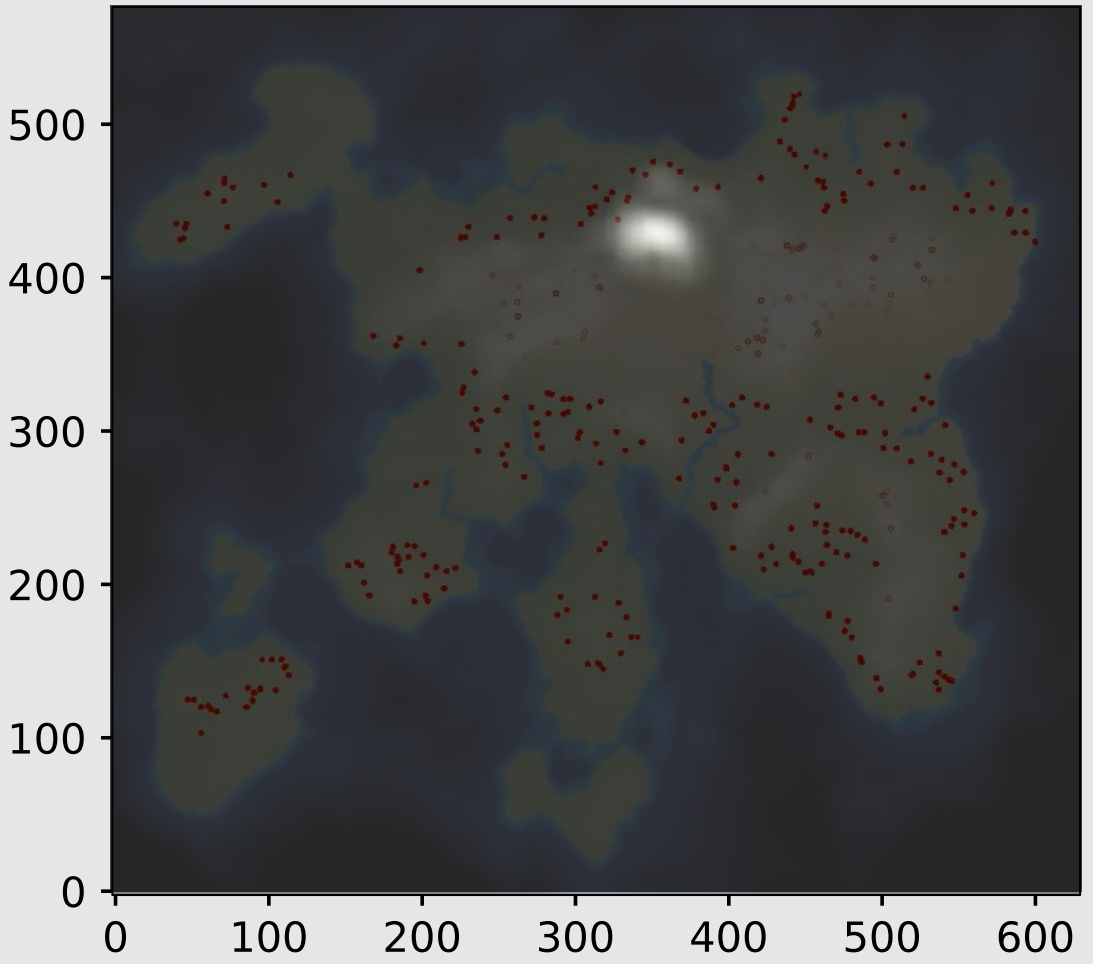

Animal id=0, camera only
0
100
200
300
400
500

Submit your answer as a heatmap.

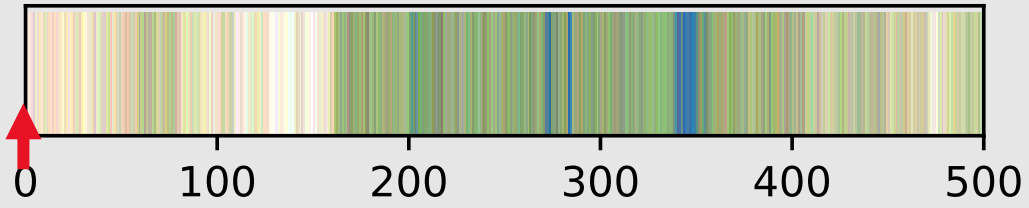
Your score will be the probability that you assign to Data Stoat's true location.

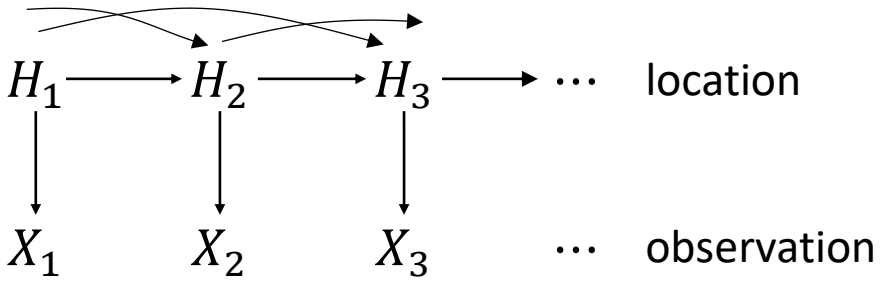
Winner will be announced on Christmas day.

Particle Filter to estimate Animal 0's location

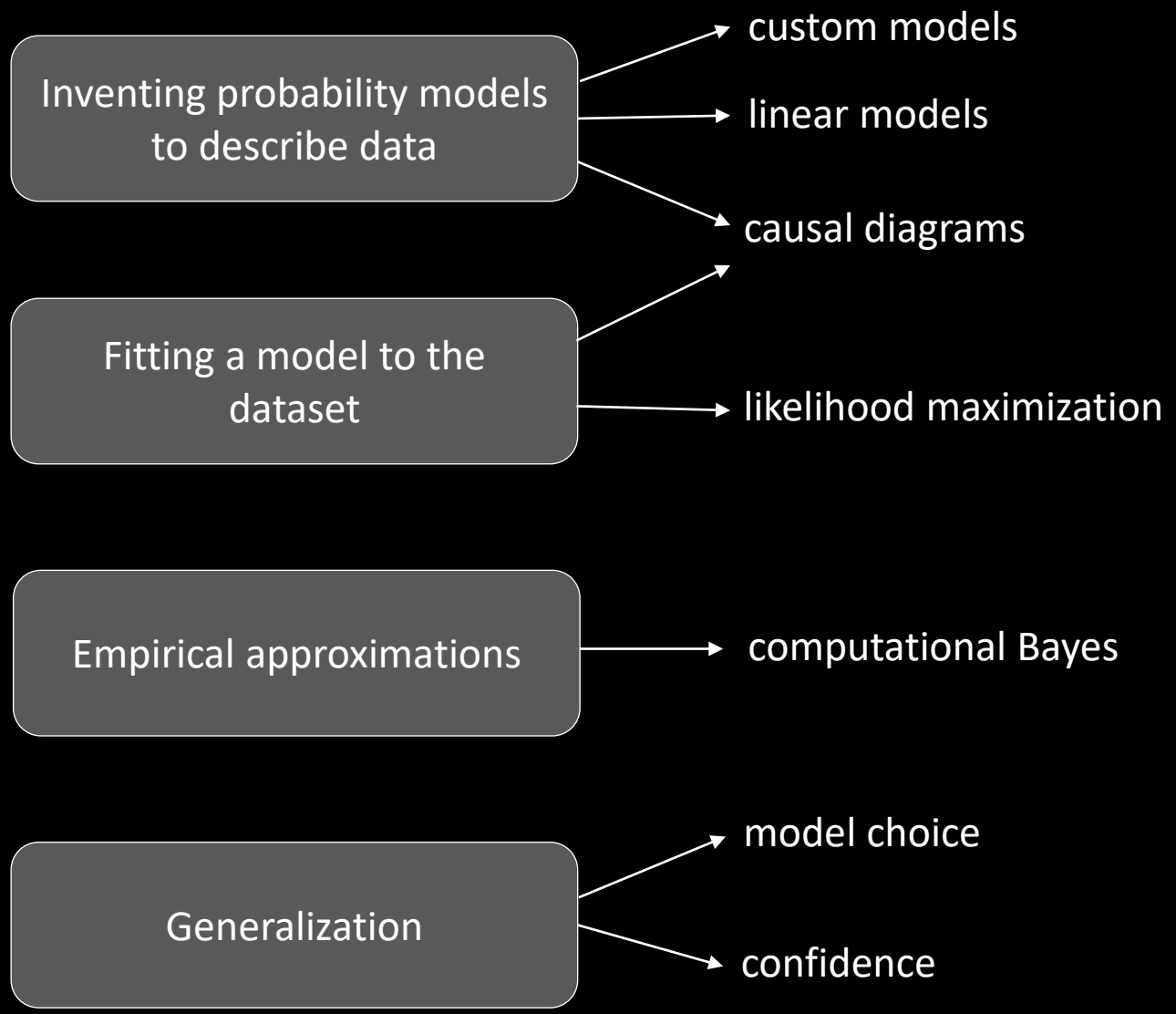


Animal id=0, camera only





- invent models for $\underline{H}$ and $(X_i|H_i)$
- fit them
- computational Bayes (particle filter)
- choose between models
- produce a heatmap based on confidence



PART IV

§4 Long-run behaviour of Markov chains

& preview of some topics for next year

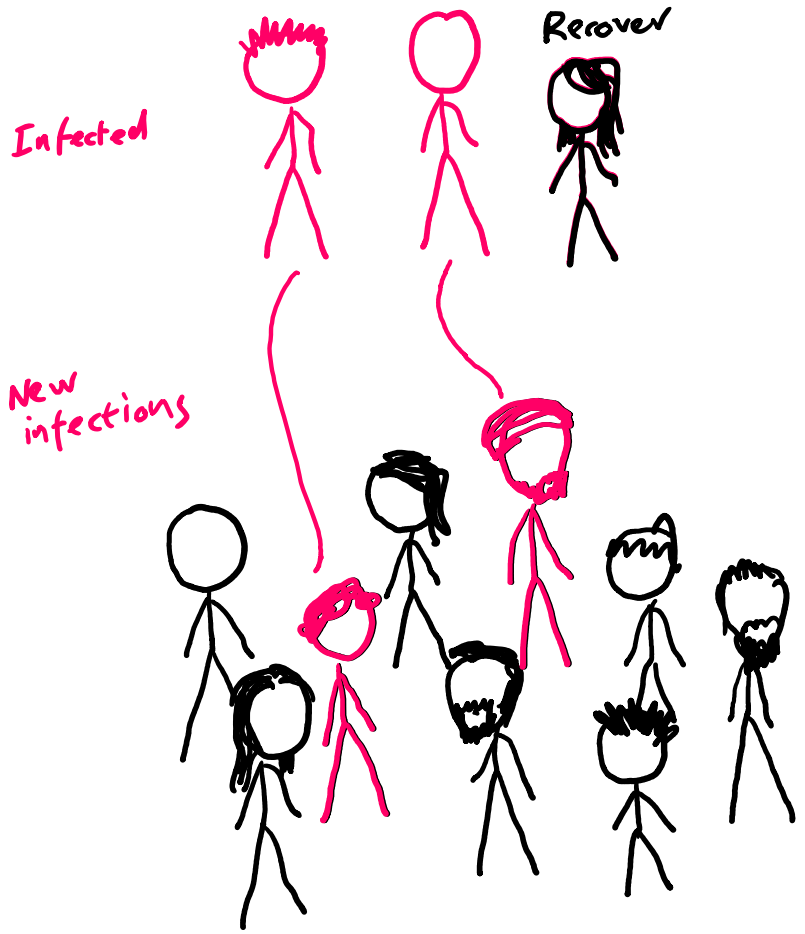
Example: dynamical system model of an epidemic

Let $X_n \in \mathbb{N}$ be the number of infected people on day n, and let it evolve according to

$$X_{n+1} = X_n - \text{Recoveries}_n + \text{Infections}_n$$

(We'll let the distributions of Recoveries_n and Infections_n depend only on X_n, making this a Markov model.)

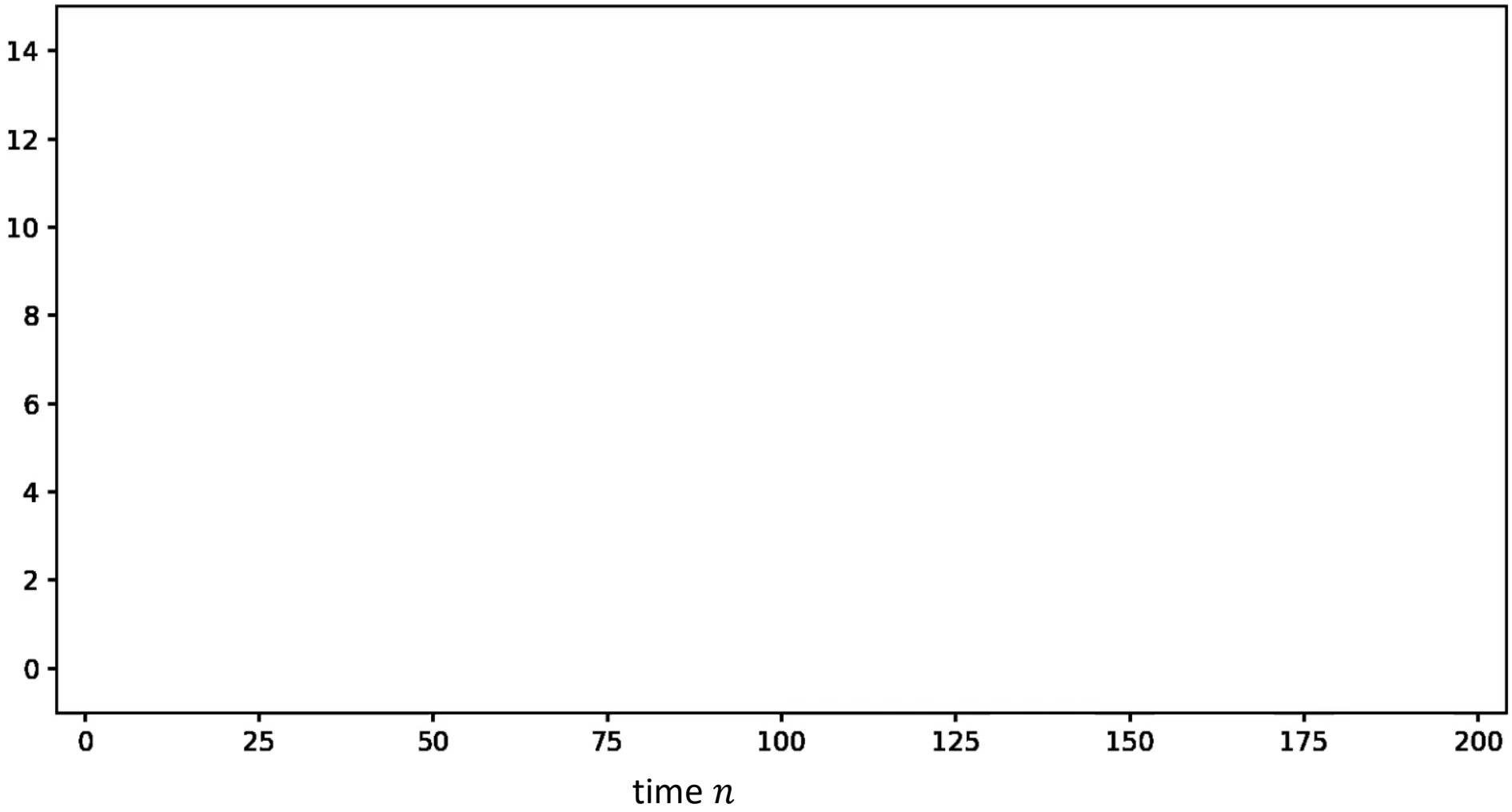
Day 1: ##infected $= 33 + 2 - 1 = 4$



Example: dynamical system model of an epidemic

Let $X_n \in \mathbb{N}$ be the number of infected people on day n

num. infected X_n
(5 simulation runs)



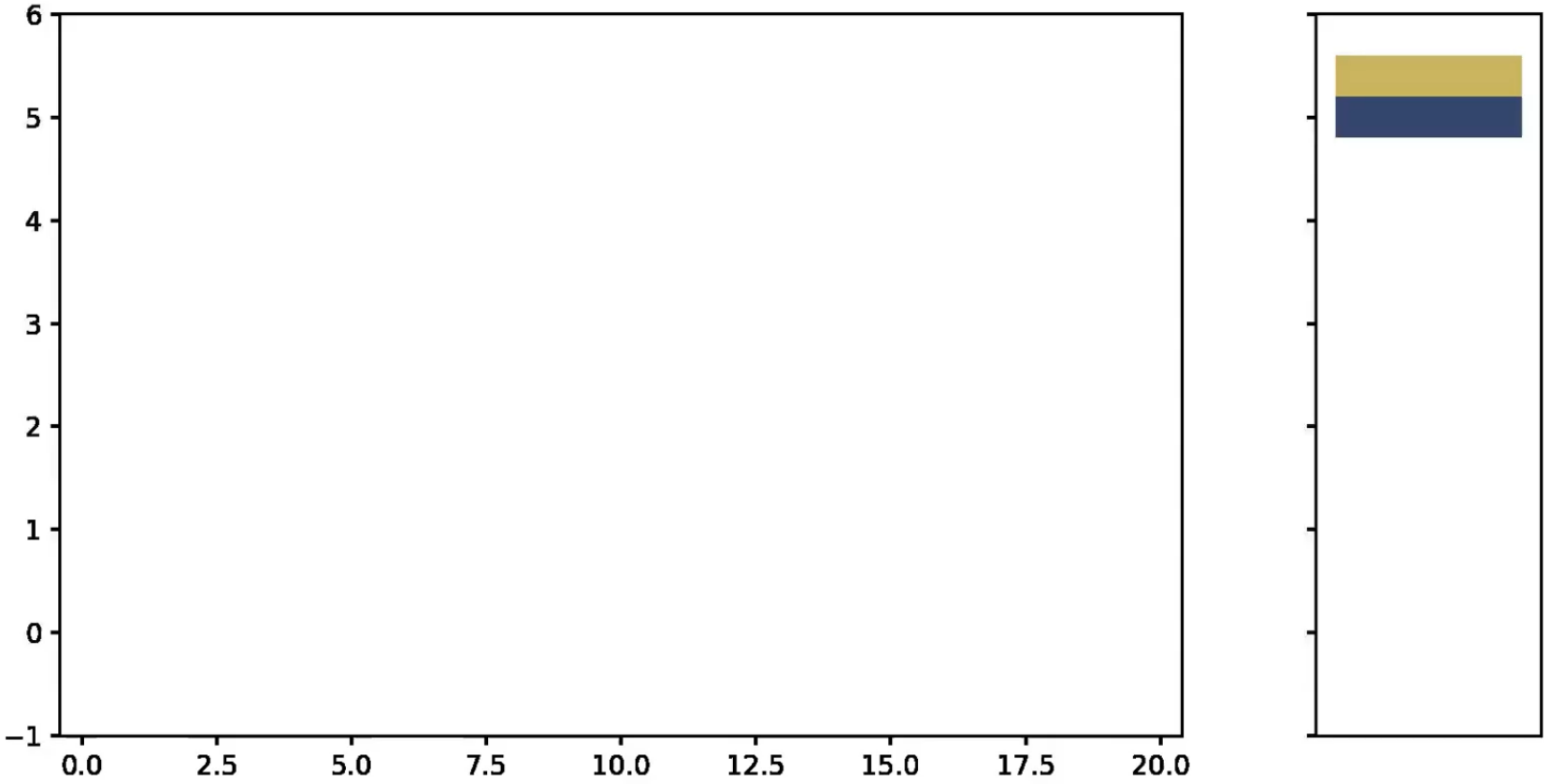
Example: dynamical system model of an online platform

Let $X_n \in \mathbb{N}$ be the number of users currently using an online platform at timestep n, and let it evolve according to

$$X_{n+1} = X_n + \text{Newusers}_n - \text{Departures}_n$$

(We'll let the distributions of Newusers_n and Departures_n depend only on X_n, making this a Markov model.)

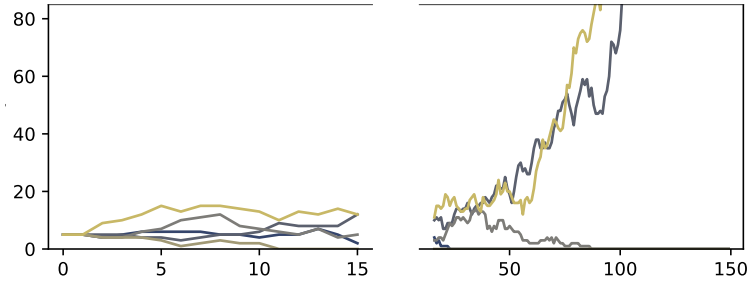
num. users X_n (2 simulation runs)



time n

Markov models can have very different behaviours, depending on the distribution of $(X_{n+1}|X_n)$.

EPIDEMIC MODEL

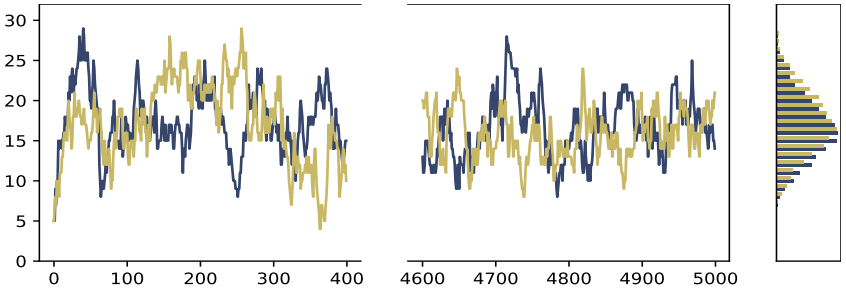


It might collapse or explode.

- How likely is it that the epidemic dies out?
- If it doesn't die out, what's the growth rate?

no longer on the syllabus

ACTIVE USERS MODEL



It might settle down to a stable **stationary distribution**.

- How can we calculate this distribution?

A distribution π over the state space is called a **stationary distribution** if

$$X_0 \sim \pi \quad \Rightarrow \quad X_1 \sim \pi$$

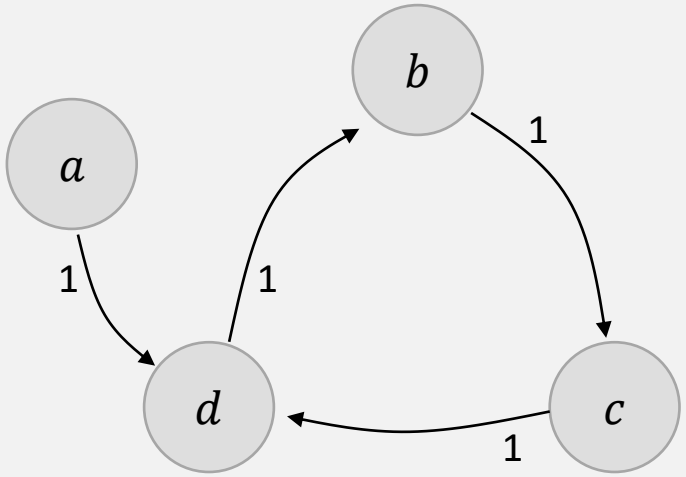
If π is a stationary distribution, and if $X_0 \sim \pi$, then by induction $X_n \sim \pi$ for all $n > 0$.

For Markov chains with a finite state space,

- $\pi \in \mathbb{R}^n_{\geq 0}$ where n is the number of states
- "$X_i \sim \pi$" means $\mathbb{P}(X_i = x) = \pi_x$ for all states x

If the state space is continuous, let π be a likelihood function instead.

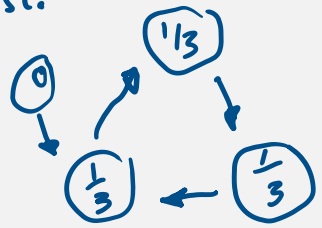
QUESTION. What's the stationary distribution of this Markov chain?



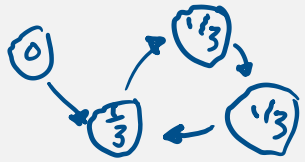
The stationary dist. is

$(\pi_a, \pi_b, \pi_c, \pi_d) = (0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

To see this, suppose X_0 has this dist.

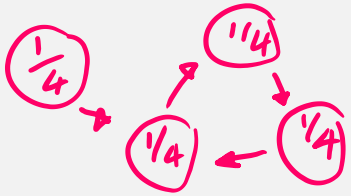


Then what is X_1?

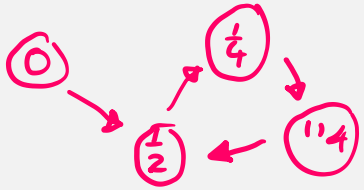


So this π is stationary.

Consider choosing $X_0 \sim \pi$ where $(\pi_a, \pi_b, \pi_c, \pi_d) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$



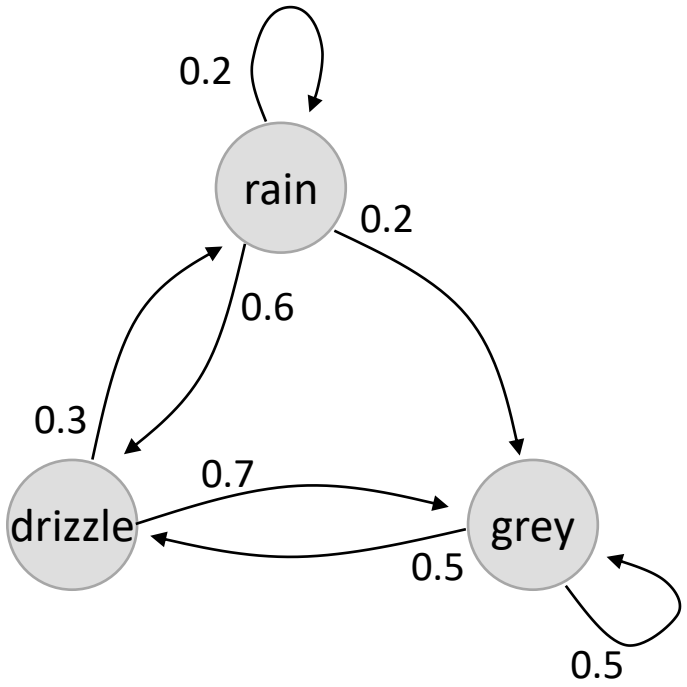
Then, after one timestep, X_1 has distribution



Thus this π is <u>not</u> a stationary distribution.

Example (Stationary distribution)

Find the stationary distribution of Cambridge weather, generated from this Markov chain:



Suppose a stationary distribution exists, call it π.

Suppose $X_0 \sim \pi$.

Then $X_1 \sim \pi$ by defn. of stationarity

But also $\mathbb{P}(X_1 = x) = \sum_y \mathbb{P}(X_1 = x | X_0 = y)\mathbb{P}(X_0 = y)$ by LoTP

π_x where P is the transition matrix P_{yx} π_y

Thus $\pi_x = \sum_y \pi_y P_{yx}$ (and x was arbitrary so the equation holds for all x)

x = rain: $\pi_{rain} = 0.2\pi_{rain} + 0.3\pi_{drizzle} + 0.5\pi_{grey}$

x = drizzle: $\pi_{drizzle} = 0.6\pi_{rain}$

x = grey: $\pi_{grey} = 0.2\pi_{rain} + 0.7\pi_{drizzle} + 0.2\pi_{grey}$

$\pi_{rain} + \pi_{drizzle} + \pi_{grey} = 1$ because π is a dist.

So we can solve for π.

Describing a Markov chain's transitions

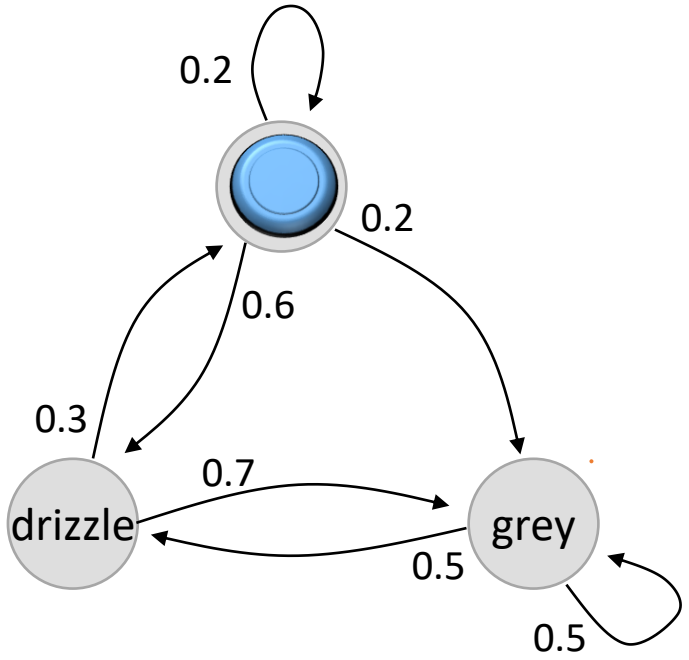
For a Markov chain $X_0 \to X_1 \to X_2 \to \cdots$ it's useful to have concise notation for describing the transition probabilities.

When the state space is discrete, we'll draw the *state space diagram* and give the *transition probability matrix*:

STATE SPACE DIAGRAM

state space = {rain, drizzle, grey}

X_n = state on day n



TRANSITION PROBABILITY MATRIX

$$P = \begin{array}{c} \\ \text{rain} \\ \text{drizzle} \\ \text{grey} \end{array} \begin{array}{c} \begin{array}{ccc} \text{rain} & \text{drizzle} & \text{grey} \end{array} \\ \begin{bmatrix} .2 & .6 & .2 \\ .3 & 0 & .7 \\ 0 & .5 & .5 \end{bmatrix} \end{array}$$

useful sanity check: rows sum to 1

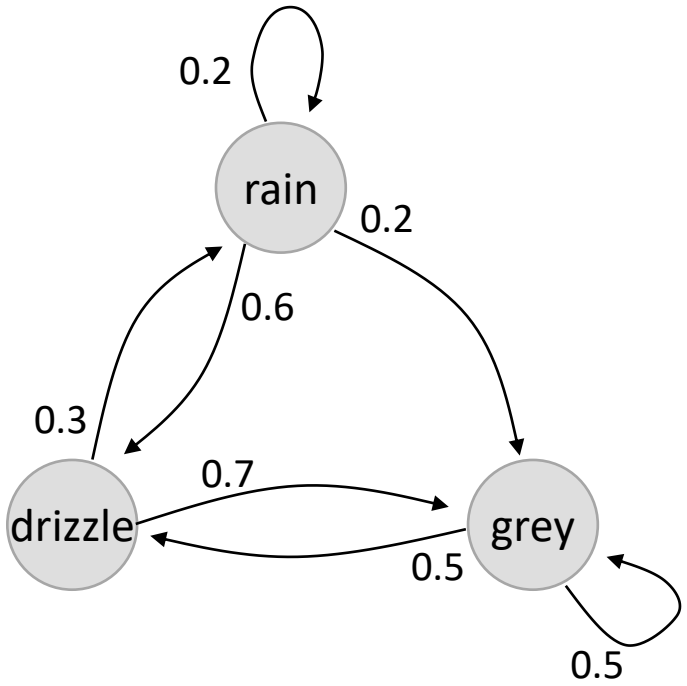
$$\mathbb{P}(X_{n+1} = j \,|X_n = i) = P_{i,j}$$

$$X_{n+1} \sim \text{Cat}(P_{X_n \bullet})$$

row X_n of P

Example (Stationary distribution)

Find the stationary distribution of Cambridge weather, generated from this Markov chain:



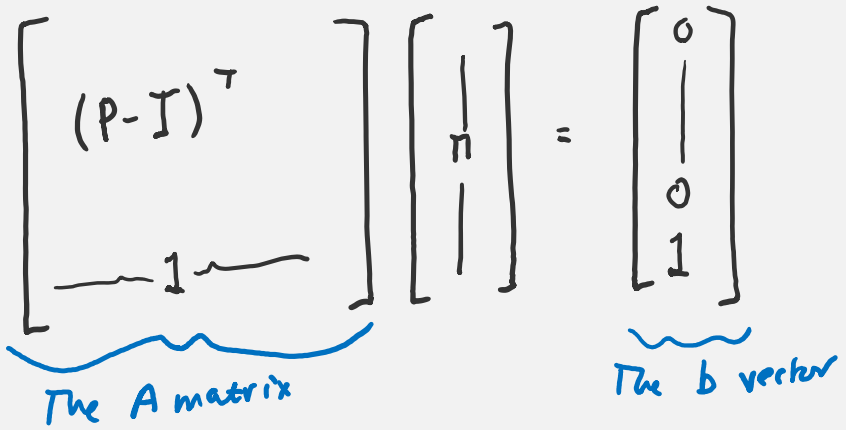
In matrix notation,

$\pi = \pi P$ or equivalently $\pi(P - I) = 0$ or $(P - I)^{\mathsf{T}} \pi = 0$

$\pi \cdot 1 = 1$

Or, putting these two together,

$A\pi = b$




```
# let states be rain=0, drizzle=1, grey=2
P = np.array([[.2,.6,.2], [.3,0,.7], [0,.5,.5]])
A = np.concatenate([(P-numpy.eye(3)).T, [[1,1,1]]])
π = np.linalg.lstsq(A, [0,0,0,1])[0]
```

- `np.linalg.lstsq(A,b)` seeks $\min_x |Ax - b|^2$. If $Ax = b$ can be solved, it will find a solution. It doesn't care about redundant equations.
- `np.linalg.solve(A,b)` solves $Ax = b$. It requires an exact system of equations, i.e. A square with no redundant equations.

Stationarity equations

If π is a stationary distribution, then it solves

$$\pi = \pi P, \qquad \pi \cdot 1 = 1$$

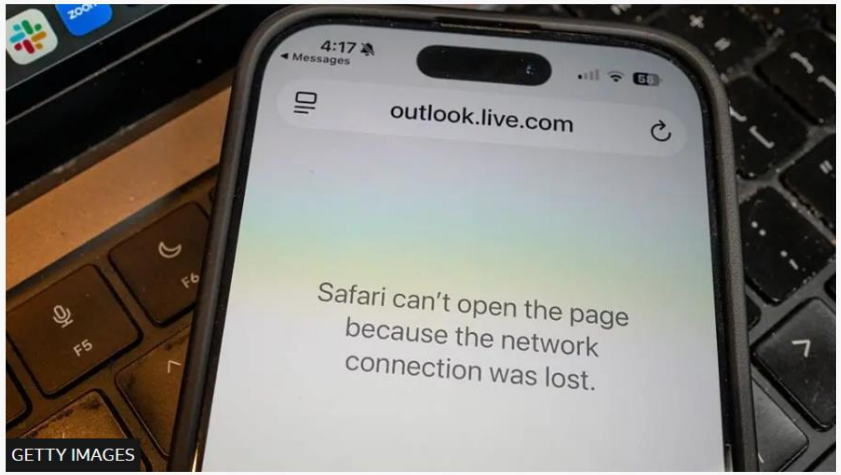
Conversely, if π is a distribution that solves $\pi = \pi P$ then π is a stationary distribution.

Why do we care? Because looking for solutions to these equations can tell us how the system is likely to behave:

unique solution	system behaves predictably
no solution (for ∞ state spaces)	system explodes
multiple solutions	multiple modes of behaviour, needs care when testing

There's a nifty shortcut called *detailed balance* that sometimes helps solve these equations. [See handout on course webpage]

Global Microsoft outage affecting tens of users



Adam Goldsmith, **Jessica Lamb** and **Osmond Chia**

12 November 2026

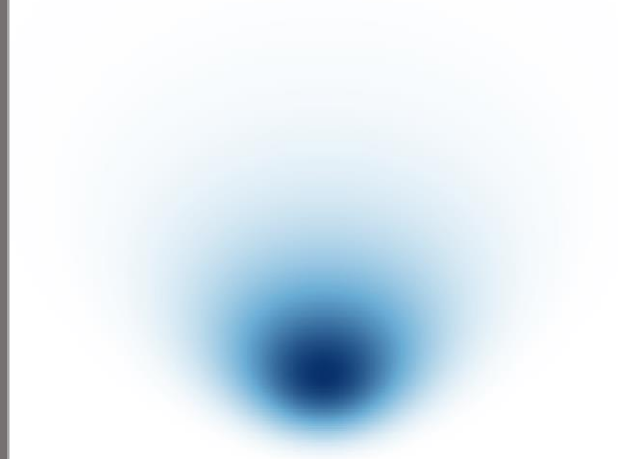
Updated 13 November 2026

Microsoft is struggling to fix a major global outage of its service that is reportedly affecting tens of thousands of users.

Part II Computer Systems Modelling



MATHEMATICAL BAYES

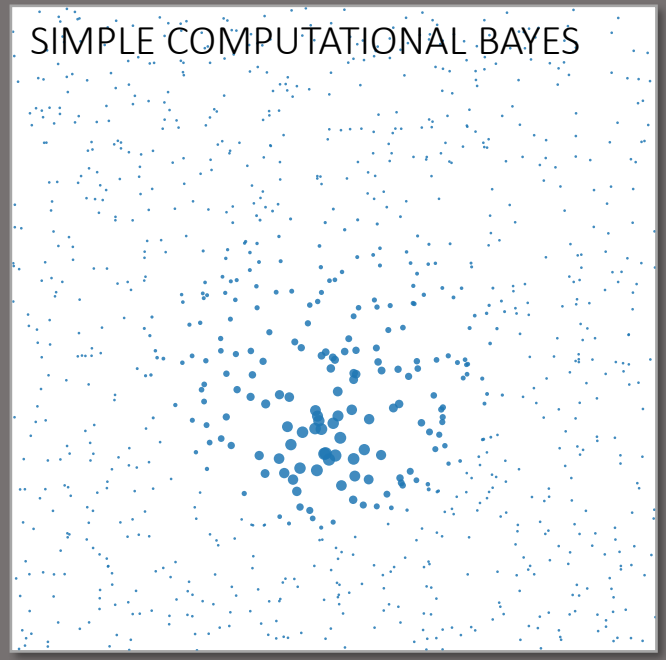


space of all possible θ

$\Pr_\Theta(\theta|\text{data}) \propto \Pr_\Theta(\theta)\ \Pr(\text{data}|\theta)$

Calculate posterior probabilities by integration

SIMPLE COMPUTATIONAL BAYES



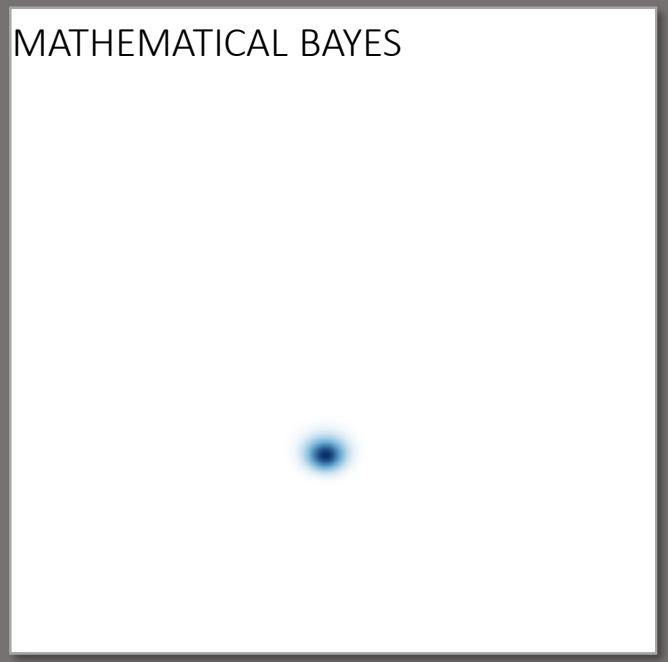
Sample points θ_j from $\Pr_\Theta$, $j = 1, \ldots, m$
Weight them by $w_j = \Pr(\text{data}|\theta_j)$

Estimate posterior probabilities using these weighted samples

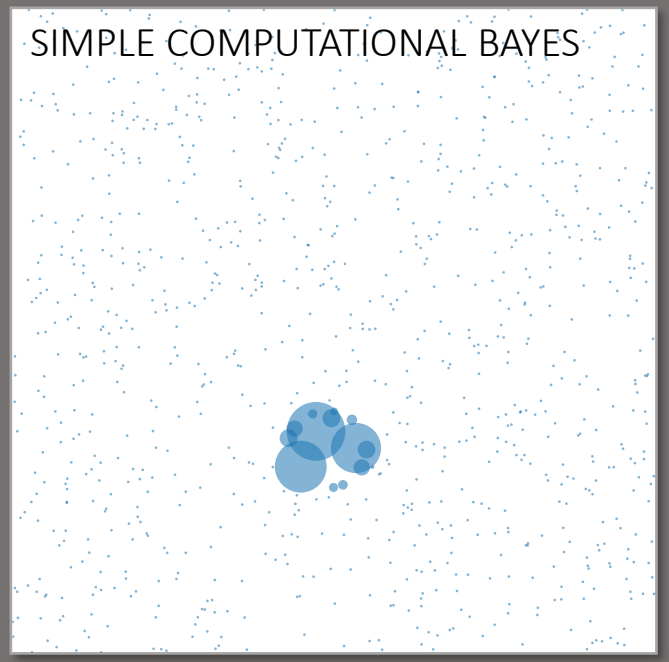


Part II Machine Learning and Bayesian Inference

MATHEMATICAL BAYES

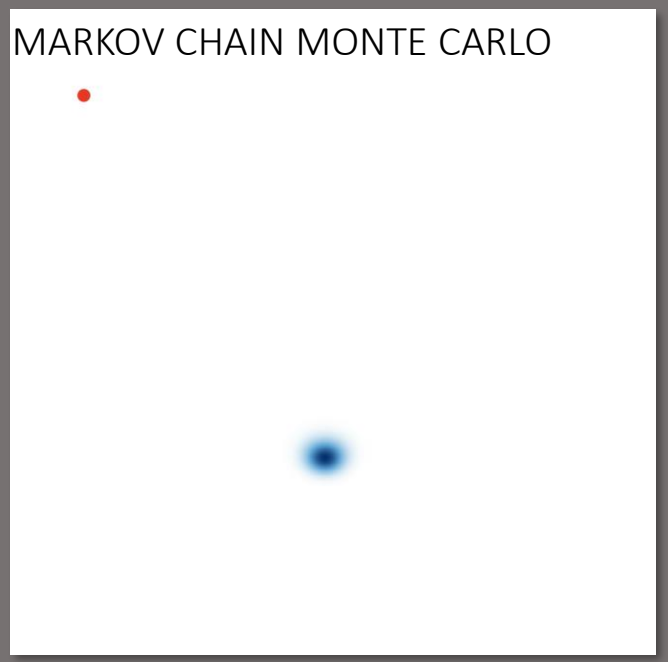


SIMPLE COMPUTATIONAL BAYES



For large datasets, our simple computational procedure becomes useless.

MARKOV CHAIN MONTE CARLO



Design a Markov chain whose stationary distribution is $\Pr_\Theta(\theta|\text{data})$

To get samples of the posterior, just use samples from this chain.

THANK YOU!



Pick up a Data Stoat sticker

[available in the teaching office,
next to reception]

Office hours today

12–1pm in the cafe area