§11.3 – §11.6
Behaviour of Markov chains

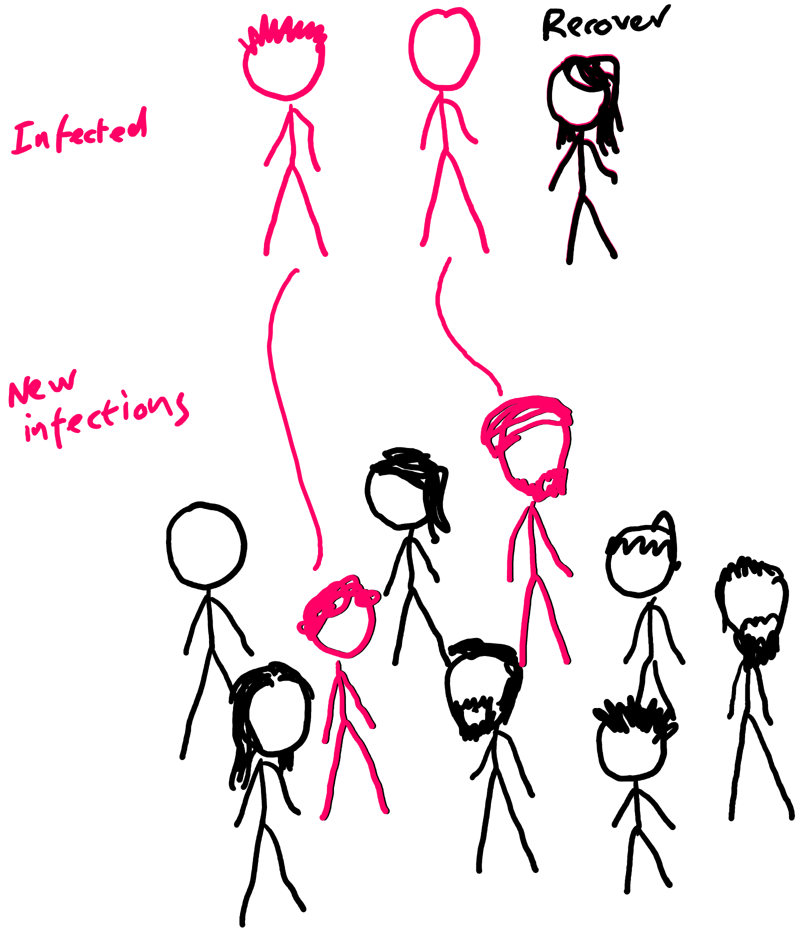
Example 11.1.2: epidemic model

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

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

(We'll let the distributions of $\mathrm{Recoveries}_n$ and $\mathrm{Infections}_n$ depend only on X_n, making this a Markov model.)

Day 2: #infected = 3 + 2 - 1 = 4



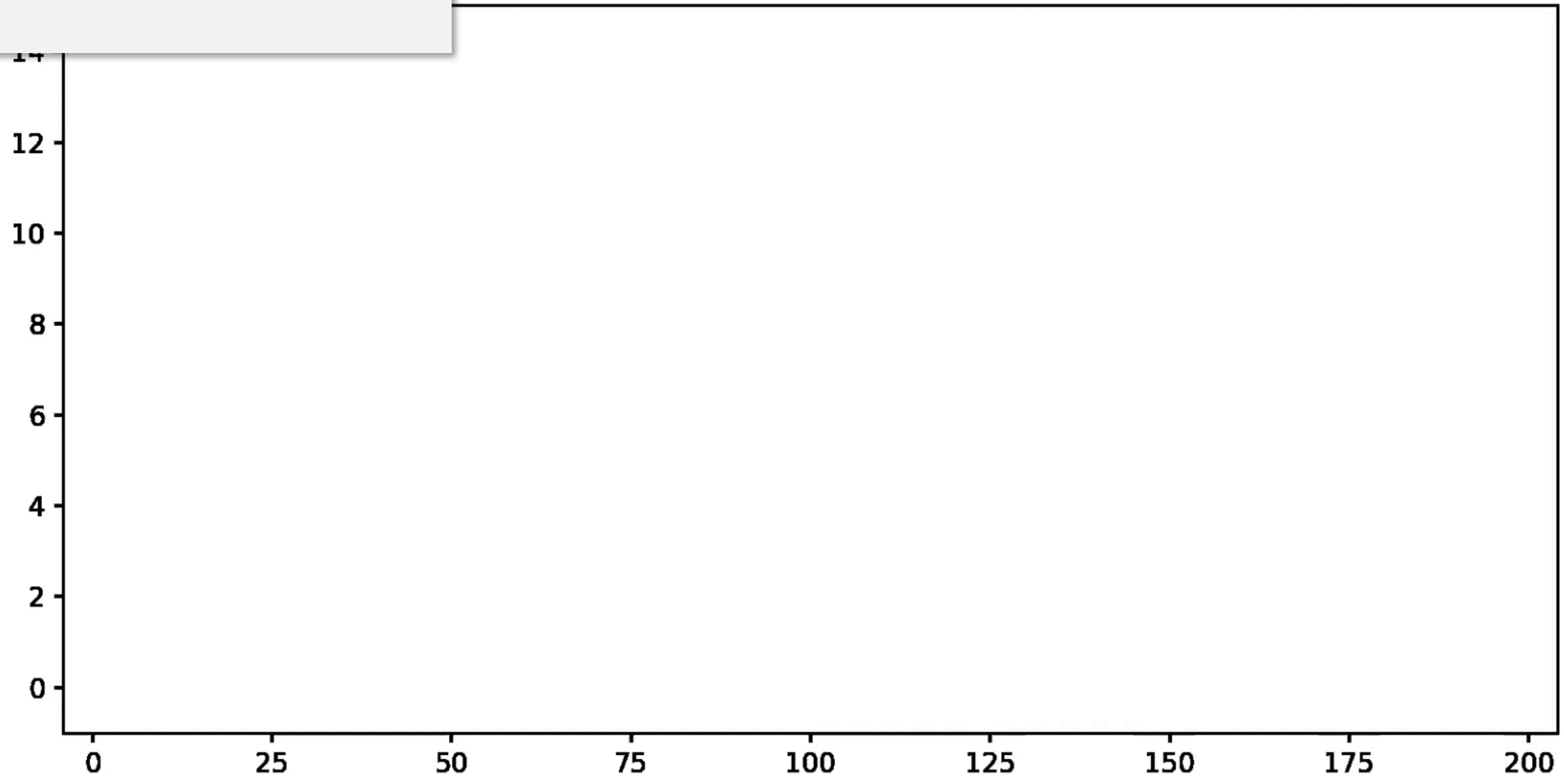
Example 11.1.2: epidemic model

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

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

(We'll let the distributions of $\mathrm{Recoveries}_n$ and $\mathrm{Infections}_n$ depend only on X_n, making this a Markov model.)

num. infected X_n
(5 simulation runs)



time n

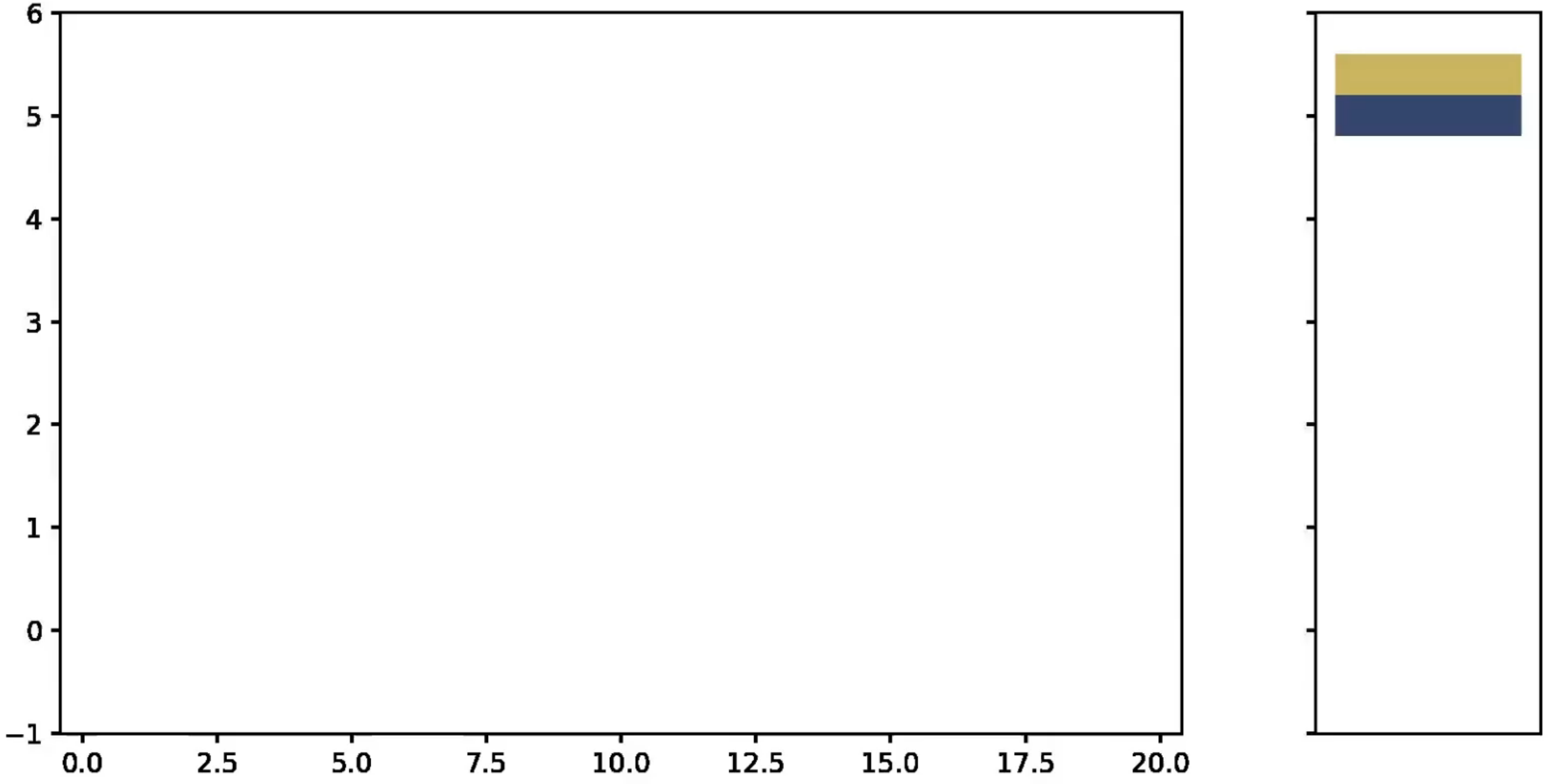
Example 11.1.3 (active users)

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

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

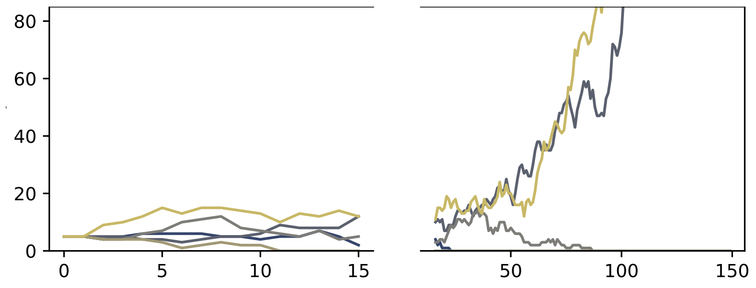
(We'll let the distributions of $\mathrm{Newusers}_n$ and $\mathrm{Departures}_n$ depend only on X_n, making this a Markov model.)

num. users X_n (2 simulation runs)



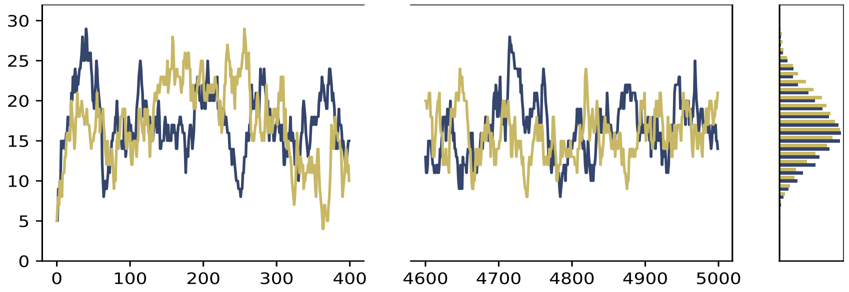
time n

EPIDEMIC MODEL



- How likely is it that the epidemic dies out? §11.3 “Hitting probabilities” (* non-examinable)
- If it doesn’t die out, what’s the growth rate? §11.6 “Drift models” (* non-examinable)

ACTIVE USERS MODEL



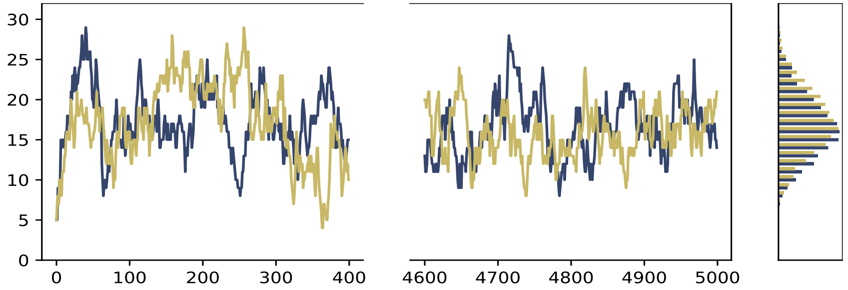
- What’s the average number of active users?

  How can we learn this distribution? §11.4, 11.5 “Stationarity”

  Needed for Part II Machine Learning & Bayesian Inference

- How can we tell which of these two behaviours we’ll see? §11.4.2 “Existence and uniqueness” (* non-examinable)

ACTIVE USERS MODEL



It looks like this distribution is stable i.e. unchanging over time

$X_i \sim \pi$ means:
$\mathbb{P}(X_i = x) = \pi_x$ for all x in the state space

Can we find a stable probability distribution π, i.e. a distribution such that $X_0 \sim \pi \;\Rightarrow\; X_1 \sim \pi$?
(If so, and if $X_0 \sim \pi$, then $X_i \sim \pi$ for all $i > 0$. We then say the chain is *stationary*.)

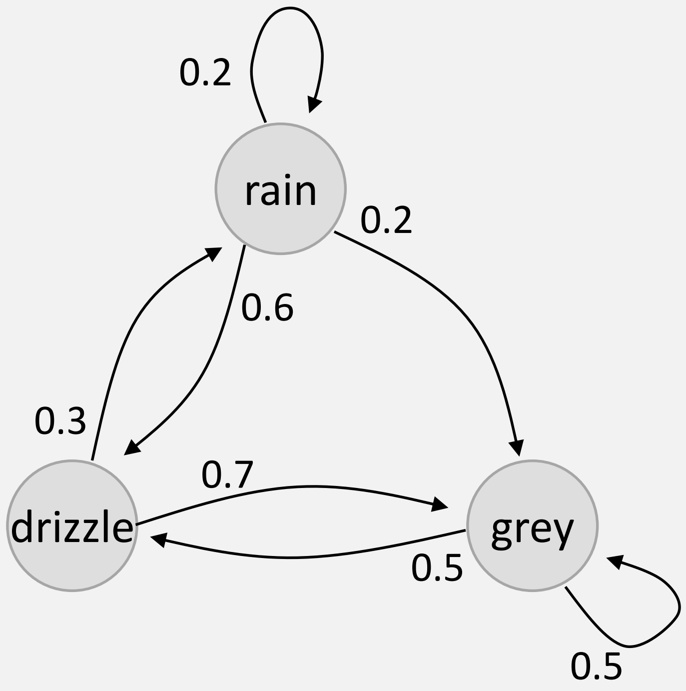
A distribution π over the state space is called a *stationary distribution* or *equilibrium distribution* if

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

§11.5. What does stationarity have to do with the histogram above, which shows time-averages?
If $X_i \sim \pi$ for all i, then the time-averages are given by π, because

$$\mathbb{E}\Big(\overbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{X_i=x}}^{\text{long-run fraction of time spent in state } x}\Big) = \frac{1}{n}\sum_i \mathbb{E}(1_{X_i=x}) = \frac{1}{n}\sum_i \mathbb{P}(X_i = x) \overset{\text{by stationarity}}{=} \frac{1}{n}\sum_i \pi_x = \pi_x$$

Example 11.4.1 (Stationary distribution)
Find the stationary distribution of Cambridge weather.



let's suppose that a stationary dist. exists. call it π.

Suppose $X_0 \sim \pi$.

Then $X_1 \sim \pi$ by definition of stationarity

But also

$$\mathbb{P}(X_1 = x) = \sum_y \mathbb{P}(X_1 = x \mid X_0 = y)\, \mathbb{P}(X_0 = y) \quad \text{by LoTP.}$$

$$\pi_x = \sum_y P_{yx}\, \pi_y \quad = \sum_y \pi_y P_{yx}$$

where P is the transition matrix

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

$x = $ drizzle: $\pi_{drizzle} = 0.6\, \pi_{rain} + 0.5\, \pi_{grey}$

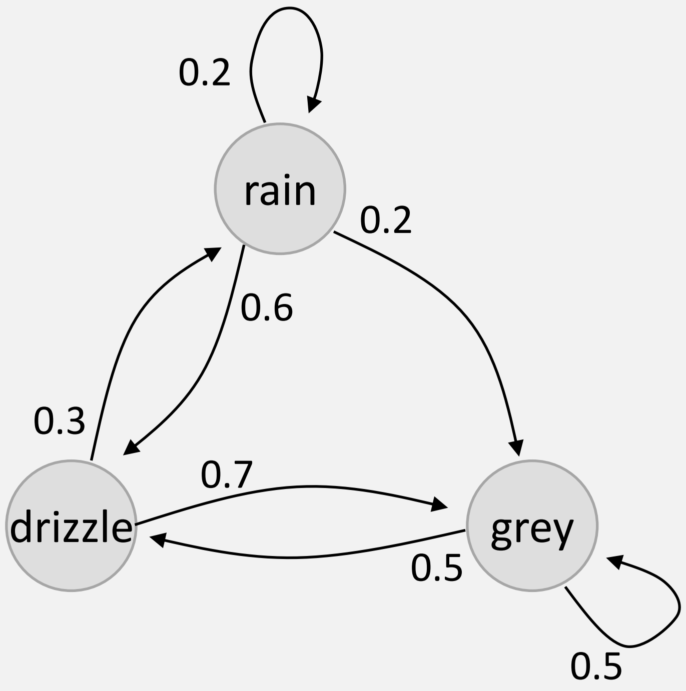
$x = $ grey: $\pi_{grey} = 0.2\, \pi_{rain} + 0.7\, \pi_{drizzle} + 0.5\, \pi_{grey}$

Also: $\pi_{rain} + \pi_{drizzle} + \pi_{grey} = 1.$ because we've assumed π is a dist.

Can solve for π.

Example 11.4.1 (Stationary distribution)

Find the stationary distribution of Cambridge weather.



In matrix notation,

$$\pi = \pi P \quad \text{or equivalently} \quad (P-I)^{\mathsf{T}}\pi = 0$$
$$\pi \cdot 1 = 1$$

Or, putting these two together,

$$A\pi = b$$

$$\underbrace{\begin{bmatrix} (P-I)^{\mathsf{T}} \\ \text{---}\,1\,\text{---} \end{bmatrix}}_{\text{The } A \text{ matrix}} \begin{bmatrix} \vert \\ \pi \\ \vert \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ \vert \\ 0 \\ 1 \end{bmatrix}}_{\text{The } b \text{ vector}}$$

```
# 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.

But does this help us to find a stationary distribution? Can these equations even be solved?

- What if there's no solution?
- What if there are multiple solutions?

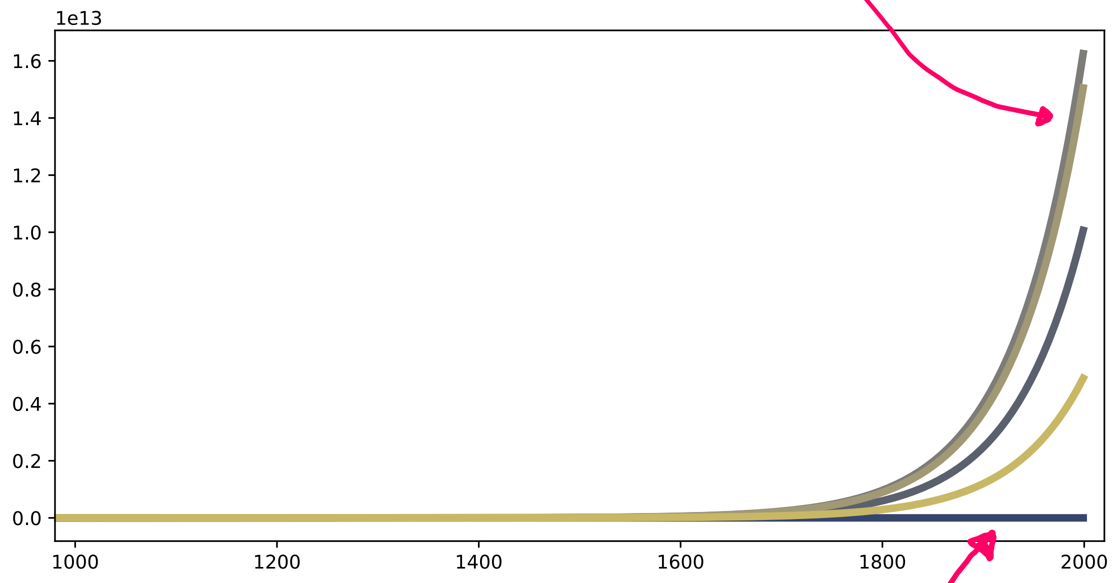
§11.4.2 "Existence and uniqueness" (* non-examinable)

Existence and uniqueness

Suppose (1) the state space is finite, and (2) the state space is *irreducible* i.e. there's a path from any state to any other

Then there is a unique stationary distribution, and it specifies the long-run time-average distribution.

If the state space is infinite, the Markov chain might 'explode'



If there are 'absorbing' states, the Markov chain might get stuck

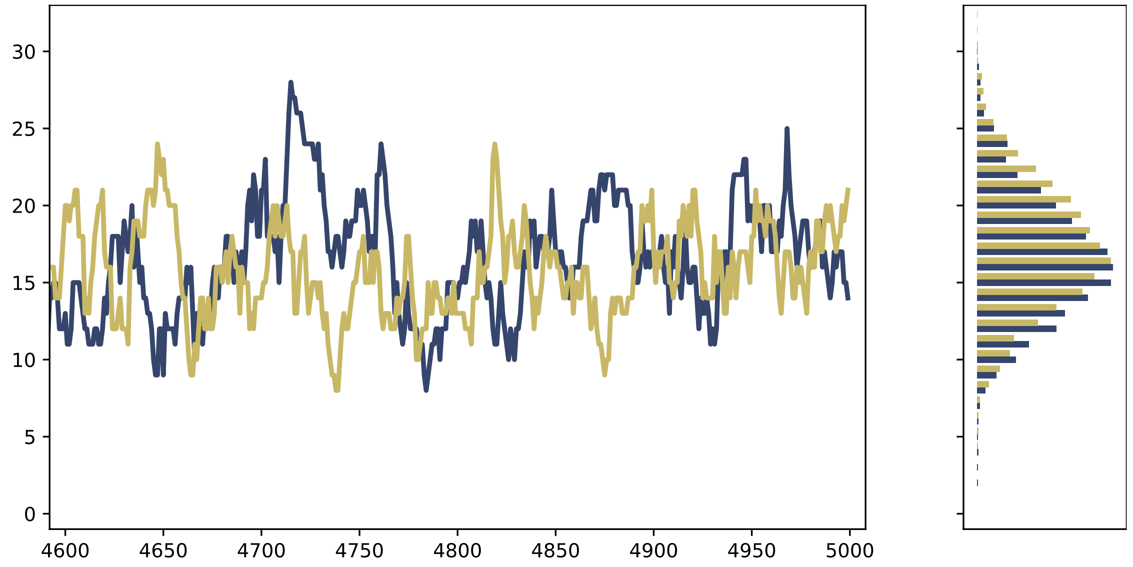
This epidemic model *does* have a unique stationary distribution (namely the "stuck at zero" distribution), but the epidemic may nonetheless explode in which case long-run time-averages aren't equal to stationary.

Existence and uniqueness

Suppose (1) the state space is finite, and (2) the state space is *irreducible* i.e. there's a path from any state to any other

Then there is a unique stationary distribution, and it specifies the long-run time-average distribution.

Even if the state space is infinite, there might still be a stationary distribution



In practice, just go ahead and solve $\pi = \pi P$. (This can always be solved.)

- If there's a unique solution **and** it can be normalized to sum to 1, then it's the unique stationary distribution.
- Otherwise, we have to work harder to classify the Markov chain's behaviour.

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.

Detailed balance equations

Lemma. If π is a vector that satisfies

$$\pi_x P_{xy} = \pi_y P_{yx} \text{ for all } x, y$$

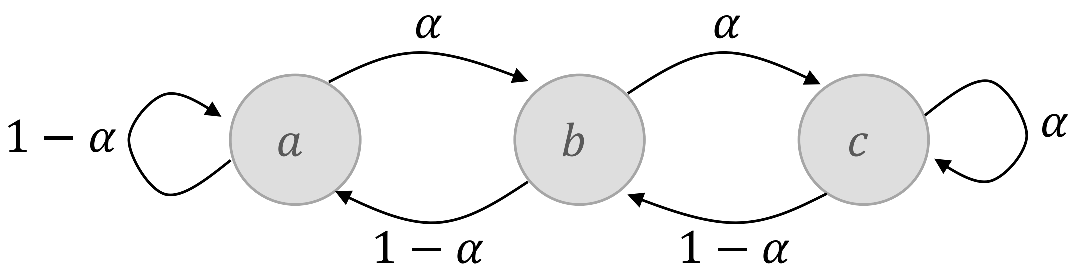
then π solves $\pi = \pi P$.

It doesn't hurt to try to solve detailed balance!

- If we're lucky, it tells us the stationary distribution
- If not, we just have to slog through solving $\pi = \pi P$

Example 11.4.4
(Stationary distribution via detailed balance)

Find the stationary distribution of the Markov chain



let's see if we can solve the detailed balance equations:

$$\underline{\forall x, y} \quad \pi_x P_{xy} = \pi_y P_{yx}.$$

$x = a,\ y = b.$ $\quad \pi_a \alpha = \pi_b (1-\alpha)$

$x = b,\ y = a.$ $\quad$ same.

$x = a,\ y = c.$ and vice versa $\quad \pi_a 0 = \pi_c 0$ – redundant

$x = b,\ y = c$ and vice versa $\quad \pi_b \alpha = \pi_c (1-\alpha)$

$x = a,\ y = a$ etc. $\quad \pi_a (1-\alpha) = \pi_a (1-\alpha)$ – redundant

$$\Rightarrow \quad \pi_b = \pi_a \frac{\alpha}{1-\alpha}$$

$$\pi_c = \pi_b \frac{\alpha}{1-\alpha} = \pi_a \left(\frac{\alpha}{1-\alpha}\right)^2$$

Add in the constraint

$$\pi_a + \pi_b + \pi_c = 1$$

and we get a stat. dist.

Also, since this chain is finite and irreducible, then there is a unique stationary distribution.
The π we have just found must be it.

IB Data Science syllabus

Using a probability model to describe data

- *Deriving the likelihood*
  Supervised and generative models built from standard distributions
- Models that depend on linear combinations of features → Parameter interpretation and identifiability
  Linear mathematics
- Sequence models
  Stationarity
  Probability calculations

Fitting a model's unknown parameters using MLE

- Least squares optimization
- Brute force optimization

Reasoning about parameter uncertainty

- *Bayes's rule*
  Bayesianist formulation
  Monte Carlo
- Confidence intervals and hypothesis testing
  Empirical distributions
- Prediction confidence (example sheets 2&3)

Please give feedback via the Qualtrix link
you should have received in your email.
(Or in person!)

- Café office hours today 1–2pm
- Café office hours Friday 11am–12noon

- I recommend the talk by Simon Peyton Jones, today at 3pm

BIG IDEA 1

Probability modelling is a great way to approach machine learning

Why don’t more people adopt it?
Because it’s unnatural!

> If you don’t get this elementary, but mildly unnatural, mathematics of elementary probability into your repertoire, then you go through a long life like a one-legged man in an ass kicking contest.
>
> Charles Munger,
> business partner of Warren Buffett

~~ALGORITHMIC~~ MACHINE LEARNING

Supervised Learning

Data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Labels: $y_1, y_2, \ldots, y_n$

Task: ~~Predict the label~~ Model $\Pr_Y(y_i\,;\,x_i)$

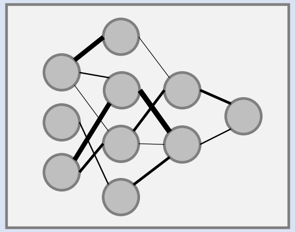
$\cancel{y_i \approx f_\theta(x_i)}$

Holdout evaluation: ~~Invent a prediction loss function e.g.~~

$\cancel{L(y, f_\theta(x)) = |y - f_\theta(x)|^2}$

~~and measure the prediction loss on a holdout dataset~~

holdout log lik

$x \longrightarrow$  $\longrightarrow f_\theta(x) \longrightarrow Y$

edge weights θ

Generative Modelling

Data: $\{x_1, x_2, \ldots, x_n\}$

Task: ~~learn to synthesize new values similar (but not identical) to those in the dataset, …~~ Model $\Pr_X(x_i)$

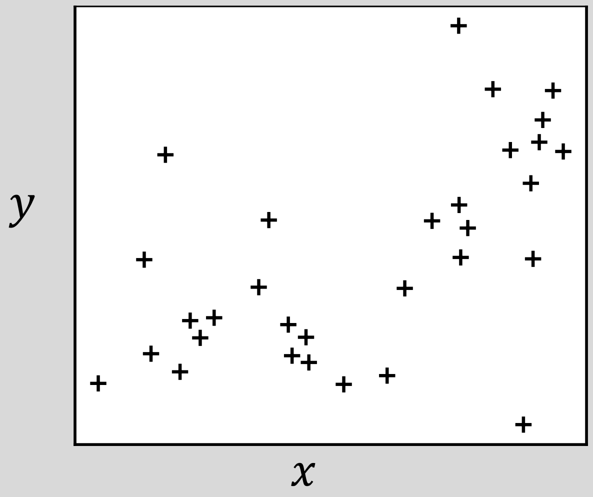
Holdout evaluation: ~~??? there is no prediction, so we can't measure prediction loss ???~~

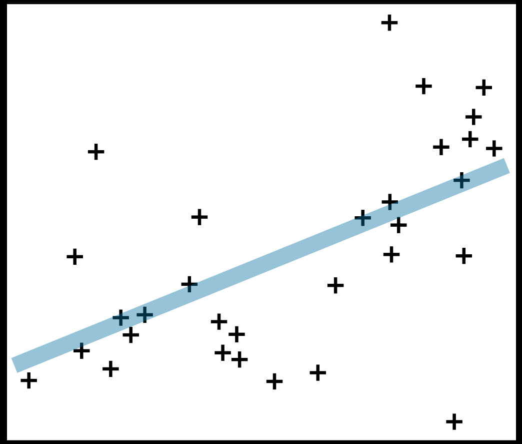
holdout log lik

The real strength of probability modelling is for the generative case, where the algorithmic approach just doesn't have the tools we need.

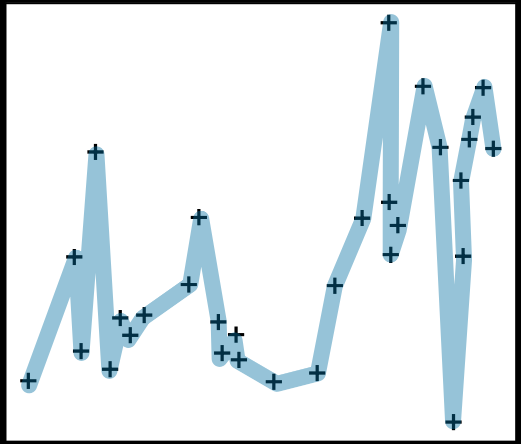
Exercise

I have a labelled dataset of (x, y) pairs and I want to predict y given x. Which of these three models is best?

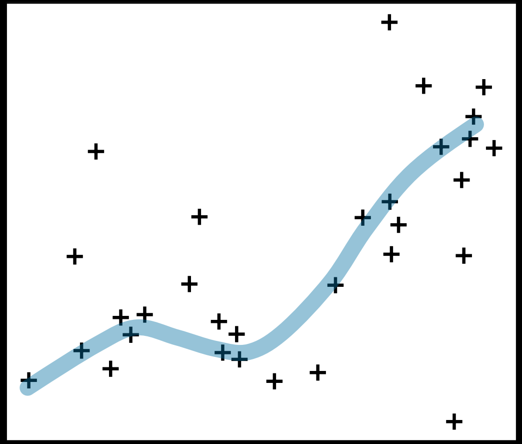




prediction loss: bad



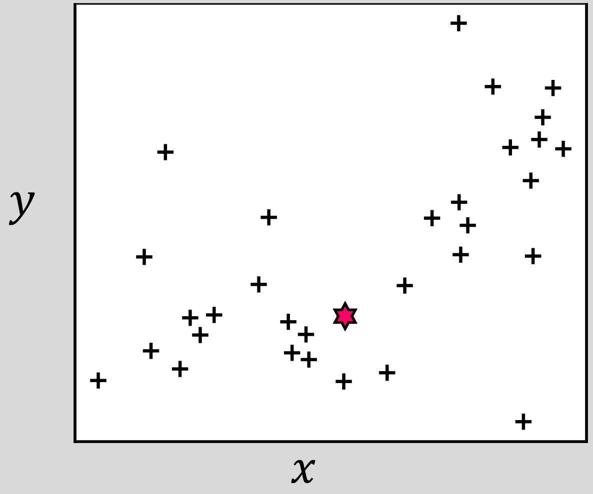
prediction loss: great



prediction loss: ok

Exercise

I have a labelled dataset of (x, y) pairs and I want to model y given x. Which of these three probability models is best?



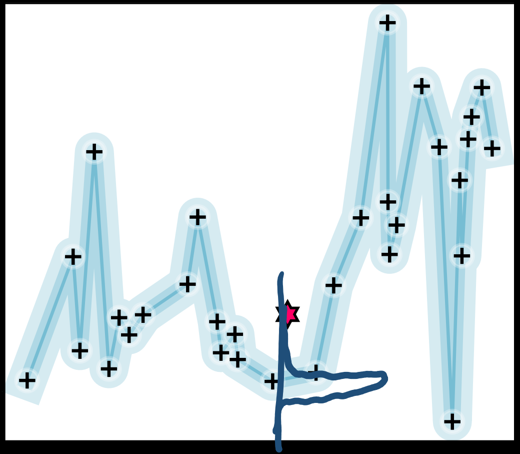
\+ training data

★ holdout data

log lik: bad

holdout log lik: bad

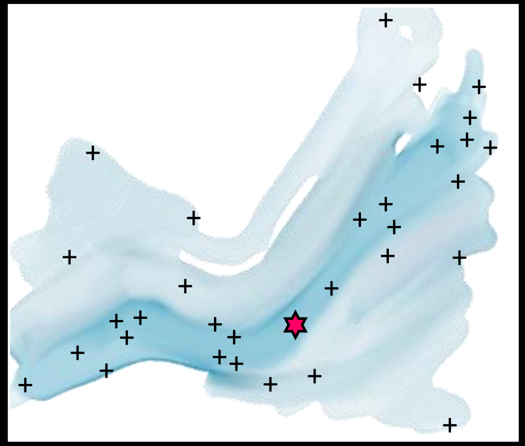
UNDERFIT



log lik: great

holdout log lik: terrible

OVERFIT



log lik: ok

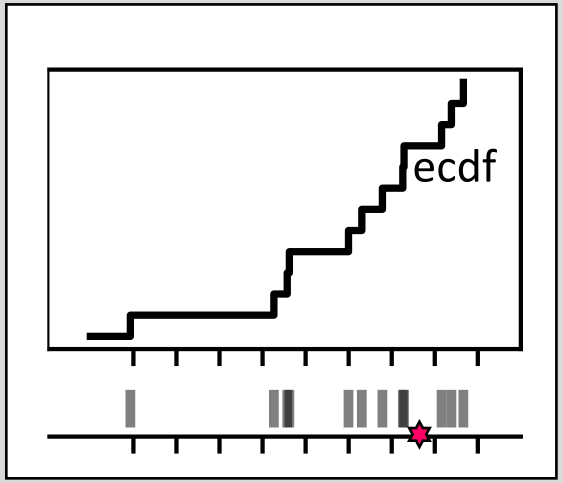
holdout log lik: ok

GOLDILOCKS FIT

Holdout log likelihood is a sensible way to evaluate a probability model. It's the natural way to generalize holdout prediction loss.

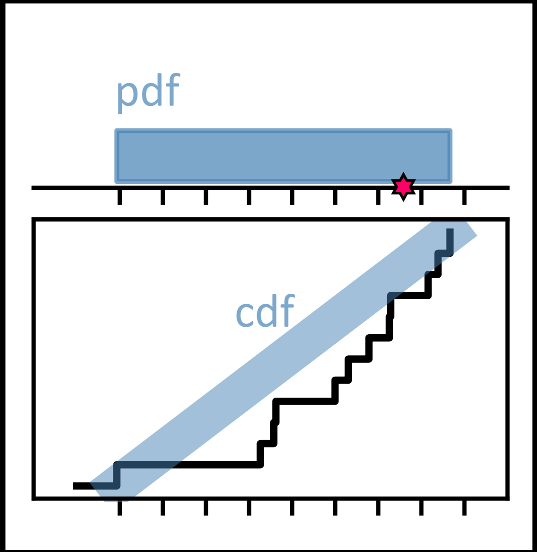
Exercise

I have an unlabelled dataset $\{x_1, x_2, \ldots, x_n\}$ and I want to fit a generative model. Which model is best?



| training data

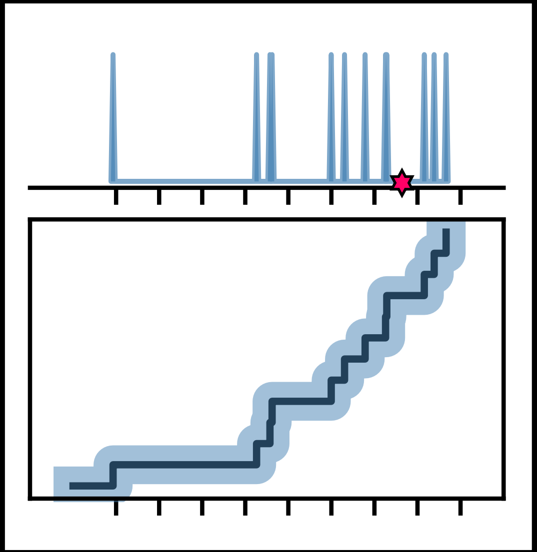
★ holdout data



fit to data: bad

holdout log lik: bad

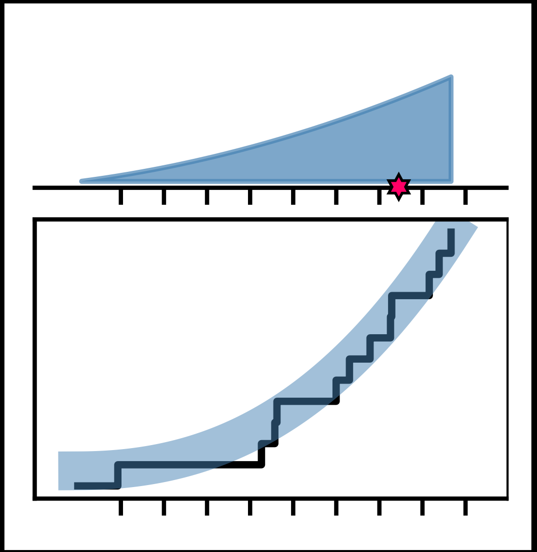
UNDERFIT



fit to data: great

holdout log lik: terrible

OVERFIT



fit to data: ok

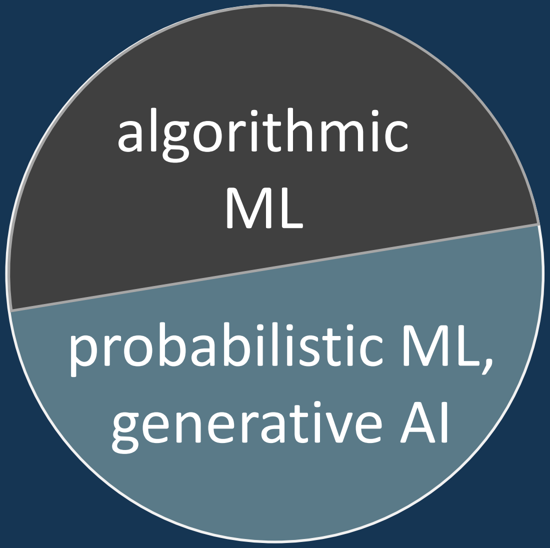
holdout log lik: ok

GOLDILOCKS FIT

Holdout log likelihood is a perfect way to evaluate generative models.

“The job of my code is to make predictions.

“Evaluate my code by how close its predictions are to the ground truth.”



There are many different ways to measure prediction accuracy, for different types of data:

for regression

R^2

mean square error

hinge loss

Cohen’s κ

margin ranking loss

mean absolute error

classification accuracy

area under curve

F1 score

for classification

and inside every sane algorithmic ML procedure there’s a probability model struggling to get out!

“The job of my code is to propose a probability model.

“Evaluate my code by the likelihood it assigns to the ground truth.”

$$\log \Pr(\text{dataset})$$

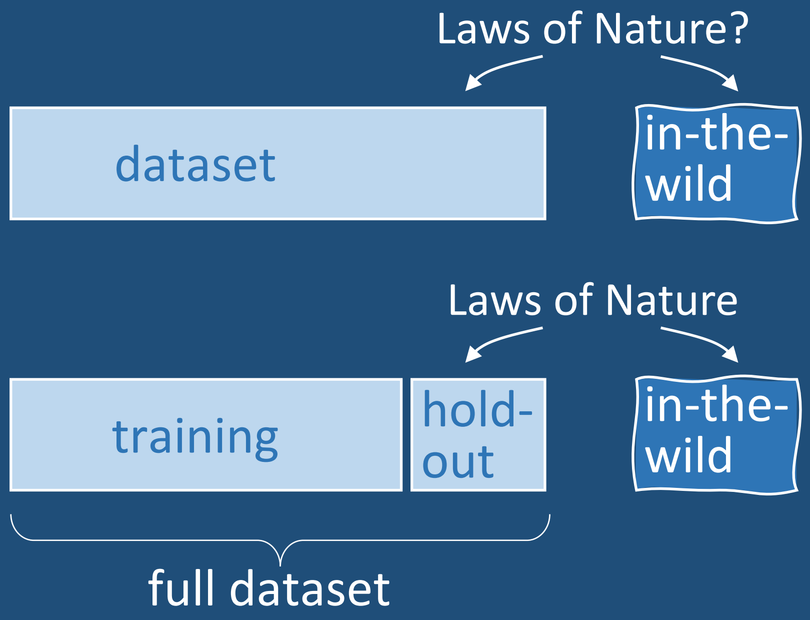
though it’s known by different names in different fields:

- *perplexity* in NLP
- *KL divergence* in ML
- *likelihood ratio* in statistics
- *ignorance score* in sports betting

BIG IDEA 2

We model because we want to make inductive claims

How will my model perform in the wild?



Every genuine scientific theory must be falsifiable.

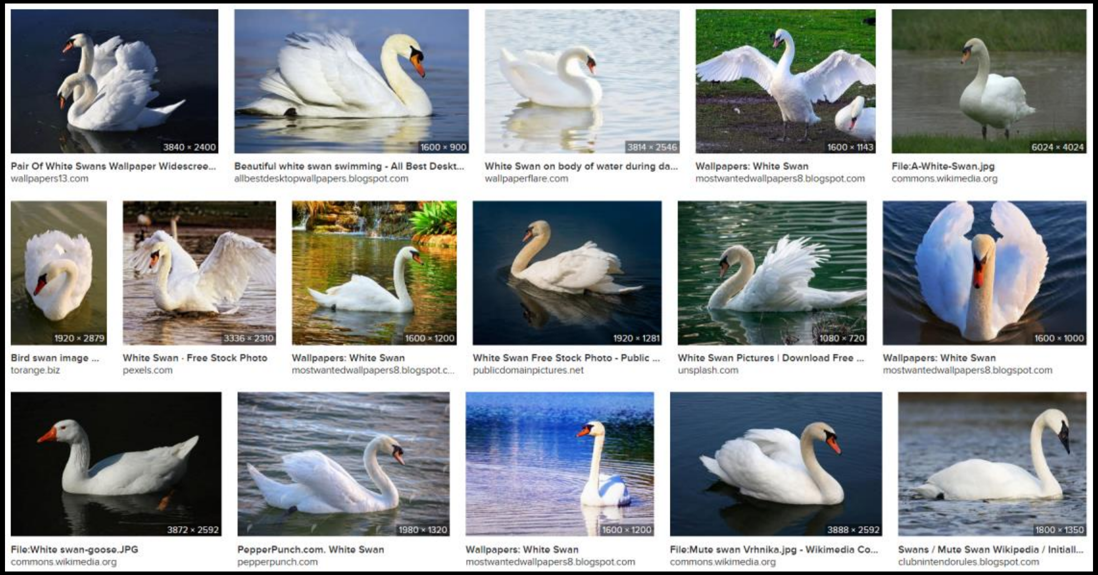
It is easy to obtain evidence in support of virtually any theory; the evidence only counts if it is the positive result of a genuinely risky prediction.

Karl Popper (1902–1994)

Why does Popper not believe in supporting evidence?

HYPOTHESIS
All swans are white, i.e.

$$\forall x \ \ \mathrm{IsSwan}(x) \Rightarrow \mathrm{IsWhite}(x)$$



ANALYSIS

The hypothesis is logically equivalent to

$$\forall x \ \ \neg\mathrm{IsWhite}(x) \Rightarrow \neg\mathrm{IsSwan}(x)$$

SUPPORTING EVIDENCE
My pot plant isn't white, and it isn't a swan.

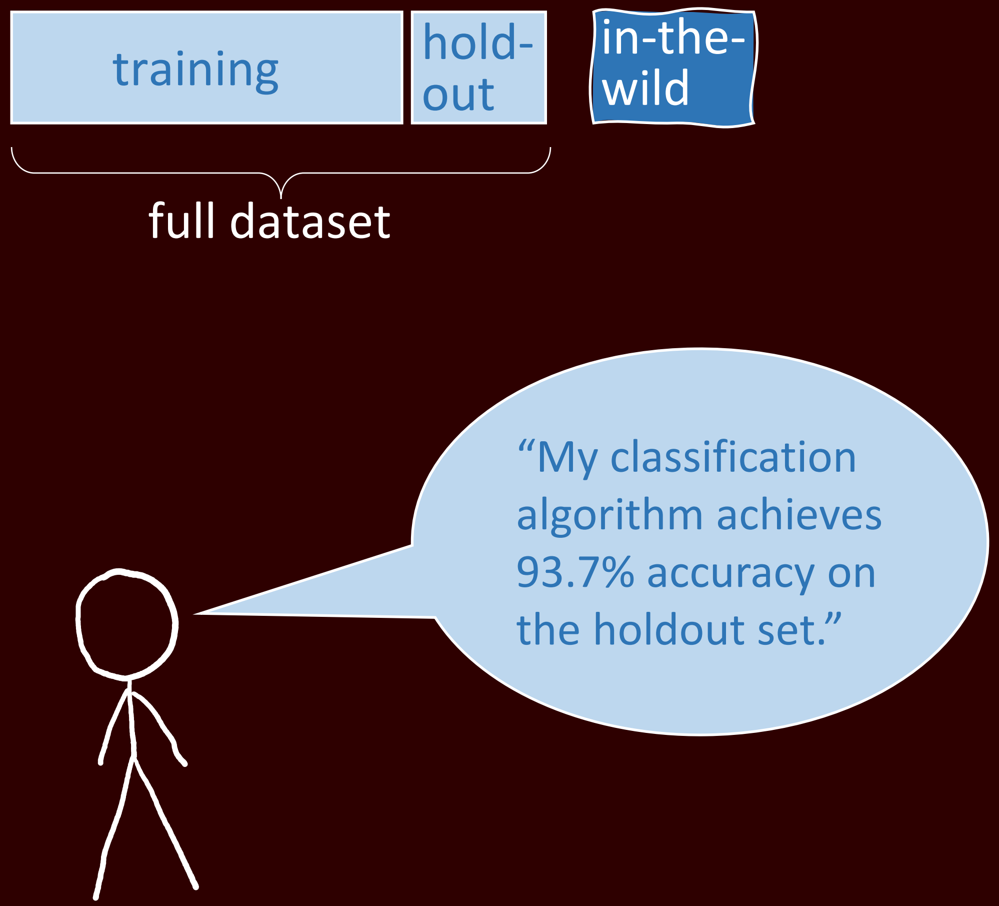
It's hard to make out the inductive claim from the "results" section of a typical ML paper.

Table 2: Results on HotpotQA distractor (dev). (+hyperlink) means usage of extra hyperlink data in Wikipedia. Models beginning with "–" are ablation studies without the corresponding design.

Model	Ans EM	Ans F_1	Sup EM	Sup F_1	Joint EM	Joint F_1
Baseline [53]	45.60	59.02	20.32	64.49	10.83	40.16
DecompRC [29]	55.20	69.63	N/A	N/A	N/A	N/A
QFE [30]	53.86	68.06	57.75	84.49	34.63	59.61
DFGN [36]	56.31	69.69	51.50	81.62	33.62	59.82
SAE [45]	60.36	73.58	56.93	84.63	38.81	64.96
SAE-large	66.92	79.62	61.53	86.86	45.36	**71.45**
HGN [14] (+hyperlink)	66.07	79.36	60.33	87.33	43.57	71.03
HGN-large (+hyperlink)	69.22	82.19	62.76	88.47	47.11	**74.21**
BERT (sliding window) variants						
BERT Plus	55.84	69.76	42.88	80.74	27.13	58.23
LQR-net + BERT	57.20	70.66	50.20	82.42	31.18	59.99
GRN + BERT	55.12	68.98	52.55	84.06	32.88	60.31
EPS + BERT	60.13	73.31	52.55	83.20	35.40	63.41
LQR-net 2 + BERT	60.20	73.78	56.21	84.09	36.56	63.68
P-BERT	61.18	74.16	51.38	82.76	35.42	63.79
EPS + BERT(large)	63.29	76.36	58.25	85.60	41.39	67.92
CogLTX	65.09	78.72	56.15	85.78	39.12	69.21
– multi-step reasoning	62.00	75.39	51.74	83.10	35.85	65.35
– rehearsal & decay	61.44	74.99	7.74	47.37	5.36	37.74
– train-test matching	63.20	77.21	52.57	84.21	36.11	66.90

Results. Table 2 shows that CogLTX outperforms most of previous methods and all **7** BERT variants solutions on the leaderboard.[4] These solutions basically follow the framework of aggregating the results from sliding windows by extra neural networks, leading to bounded performances attributed to insufficient interaction across paragraphs.

Limitations of current ML: it makes weak inductive claims



What your readers think (and want):

For a new in-the-wild datapoint x,

$$\mathbb{P}(\text{classify } x \text{ correctly}) = 93.7\%$$

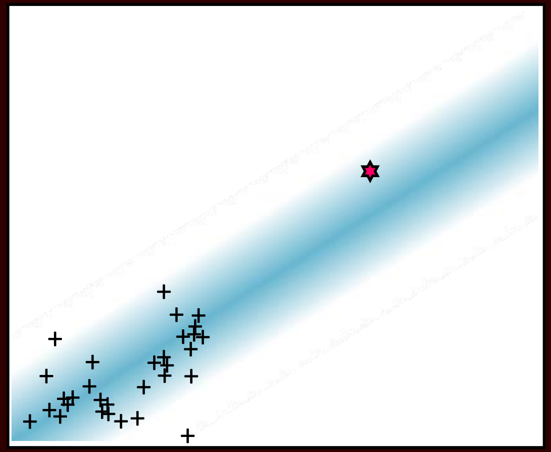
What you actually meant:

When I take a big bunch of new in-the-wild datapoints* then, averaged across this bunch,

$$\text{fraction classified correctly} = 93.7\%$$

* assuming that this bunch matches the composition of my holdout set

Limitations of current ML: it has trouble with generalization



Machine learning approach

“The job of a model is to **generalize to new data.** I’ll split my data into training + holdout, and measure how accurate it is on the holdout set.”

Scientist’s approach

“The job of a model is to **generalize to novel situations.** Any model that’s not based on well-grounded scientific concepts will not align with the Laws of Nature, so it will probably make bad predictions.”

Everything should be made as simple as possible, but not simpler.

Albert Einstein

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

the Herbert Spencer Lecture,
Oxford, 10 June 1933

“It is more important to have beauty in one’s equations than to have them fit experiment.”

Paul Dirac

Limitations of current ML: logical learning

DW **You**

Bleen is an adjective, meaning "if the year is before 2000 then blue, otherwise green". I have a bleen car, and the year is 2023. What colour is my car?

ChatGPT

Your car, described as "bleen" and given the year 2023, would be green. The definition of "bleen" you provided changes the color based on the year: blue for years before 2000 and green for years from 2000 onwards. Since 2023 is after 2000, the color of your car would be green.

Human:

Did you know that there is no country in Africa that starts with the letter "k"?

ChatGPT:

Yes, that's correct! While there are 54 recognized countries in Africa, none of them begin with the letter "K". The closest is Kenya, which starts with a "K" sound, but is actually spelled with a "K" sound. It's always interesting to learn new trivia facts like this.

Current AI does some amazing things with one-shot learning, but it's still missing the picture.

When you study IB Artificial Intelligence, think hard about why old-school AI can do things that modern AI can't.