§1.5 Better notation for likelihood

All of machine learning is based on a single idea:

1. Write out a probability model

2. Fit the model from data
   by maximizing the likelihood

This is behind

- A-level statistics formulae
- our climate model
- ChatGPT training

Step 1.5. Find an expression for the likelihood

$\Pr_X(x)$ is "the likelihood for X of x"

The *likelihood function* for a random variable X is written $\Pr_X(x)$ and defined as

$$\Pr_X(x) = \mathbb{P}(X = x) \quad \text{in the case where } X \text{ is discrete}$$

and as

$$\Pr_X(x) = \mathrm{pdf}(x) \quad \text{in the case where } X \text{ is continuous with prob. density function } \mathrm{pdf}(x)$$

For parameterized random variables, write

$$\Pr_X(x\,;\theta) \quad \text{or} \quad \Pr_X(x \mid \theta) \quad \text{or} \quad \Pr_X(x)$$

The likelihood notation has two parts:
a random variable i.e. a RNG, and a value.
e.g. $\Pr_{X+Y}(0.2)$

I have a RNG called "X+Y", which calls the X RNG and the Y RNG and adds their outputs together. What's the chance that "X+Y" gives output 0.2?

Pairs of random variables:
$\Pr_{X,Y}(x,y)$ is called the *joint likelihood* of X and Y

Independent random variables:
$\Pr_{X,Y}(x,y) = \Pr_X(x)\Pr_Y(y)$

Independent identically-distributed (IID) sample from X:
$\Pr(x_1,\ldots,x_n) = \Pr_X(x_1)\times\cdots\times\Pr_X(x_n)$

Sequential generation of X, then Y based on X:
$\Pr_{X,Y}(x,y) = \Pr_X(x)\Pr_Y(y\,;x)$

This is how to think of it.

TECHNICAL ASIDE

For discrete random variables,

$$\mathbb{P}(X=x, Y=y) = \Pr_{X,Y}(x,y)$$

$$\mathbb{P}(X\in A, Y\in B) = \sum_{x\in A, y\in B}\Pr_{X,Y}(x,y)$$

For continuous random variables,

$$\mathbb{P}(X\in A, Y\in B) = \int_{x\in A, y\in B}\Pr_{X,Y}(x,y)\,dxdy$$

See IA Probability lecture 6

Maximum Likelihood Estimation, again

If we've seen an outcome x, and we've proposed a probability model X, and if its distribution involves some unknown parameters θ,

the *maximum likelihood estimator* for θ is

$$\hat{\theta} = \arg\max_{\theta} \Pr_X(x\,;\theta)$$

- x could be discrete or continuous
- x could be a single observation or a dataset with many observations

The point of the likelihood notation is so that we can write down a single equation and have it cover all these cases.

Rules of Probability

Understand what is meant by *sample space*, written Ω, and know that $\mathbb{P}(\Omega) = 1$. Be able to reason about probabilities of events with Venn diagrams. Know the core definitions and laws ...

Conditional probability, or equivalently the chain rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0$$

$$\mathbb{P}(B, A) = \mathbb{P}(B)\,\mathbb{P}(A \mid B) \qquad \text{(chain rule)}$$

If A and B are independent: here, A and B are events

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

$$\mathbb{P}(A \mid B) = \mathbb{P}(A)$$

Sum rule, and the law of total probability, for events $\{B_1, B_2, \ldots\}$ that partition the sample space (i.e. for events that are mutually exclusive and where $\bigcup_i B_i = \Omega$):

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, B_i)$$

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i) \qquad \text{(law of total probability)}$$

Bayes's rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0.$$

The fundamental rules of probability still hold, they're just written different in likelihood notation.

here, X and Y are random variables

$$\Pr_{X,Y}(x, y) = \Pr_X(x)\,\Pr_Y(y)$$

$$\left\{ \begin{aligned} \Pr_X(x) &= \sum_y \Pr_{X,Y}(x, y) \\ \Pr_X(x) &= \int_y \Pr_{X,Y}(x, y)\,dy \end{aligned} \right.$$

$\Pr_X(\cdot)$ is a function,
$\Pr_X : \Omega \to \mathbb{R}_{\geq 0}$ where Ω is the sample space of X.

$\Pr_X$ is a deterministic function, call it $f : \Omega \to \mathbb{R}$.

When I apply a deterministic function to a random value, my answer is random.

e.g. $Y \sim N(0,1)$
$\Rightarrow$ $\sin(Y)$ is a random variable

Thus, $f(X)$ is a random variable.

Brain teaser

For this random variable X, what is $\Pr_X(X)$?

$$X = \begin{cases} \text{cat} & \text{with prob. } 1/6 \\ \text{stoat} & \text{with prob. } 3/6 \\ \text{dog} & \text{with prob. } 2/6 \end{cases}$$

What is the deterministic function f?

$f(\text{cat}) = \Pr_X(\text{cat}) = \mathbb{P}(X = \text{cat}) = 1/6$

$f(\text{stoat}) = \frac{1}{2}$

$f(\text{dog}) = 1/3.$

Thus:

$$f(X) = \begin{cases} 1/6 & \text{w. prob. } 1/6 \\ 1/2 & \text{w. prob } 1/2 \\ 1/3 & \text{w.p. } 1/3 \end{cases}$$

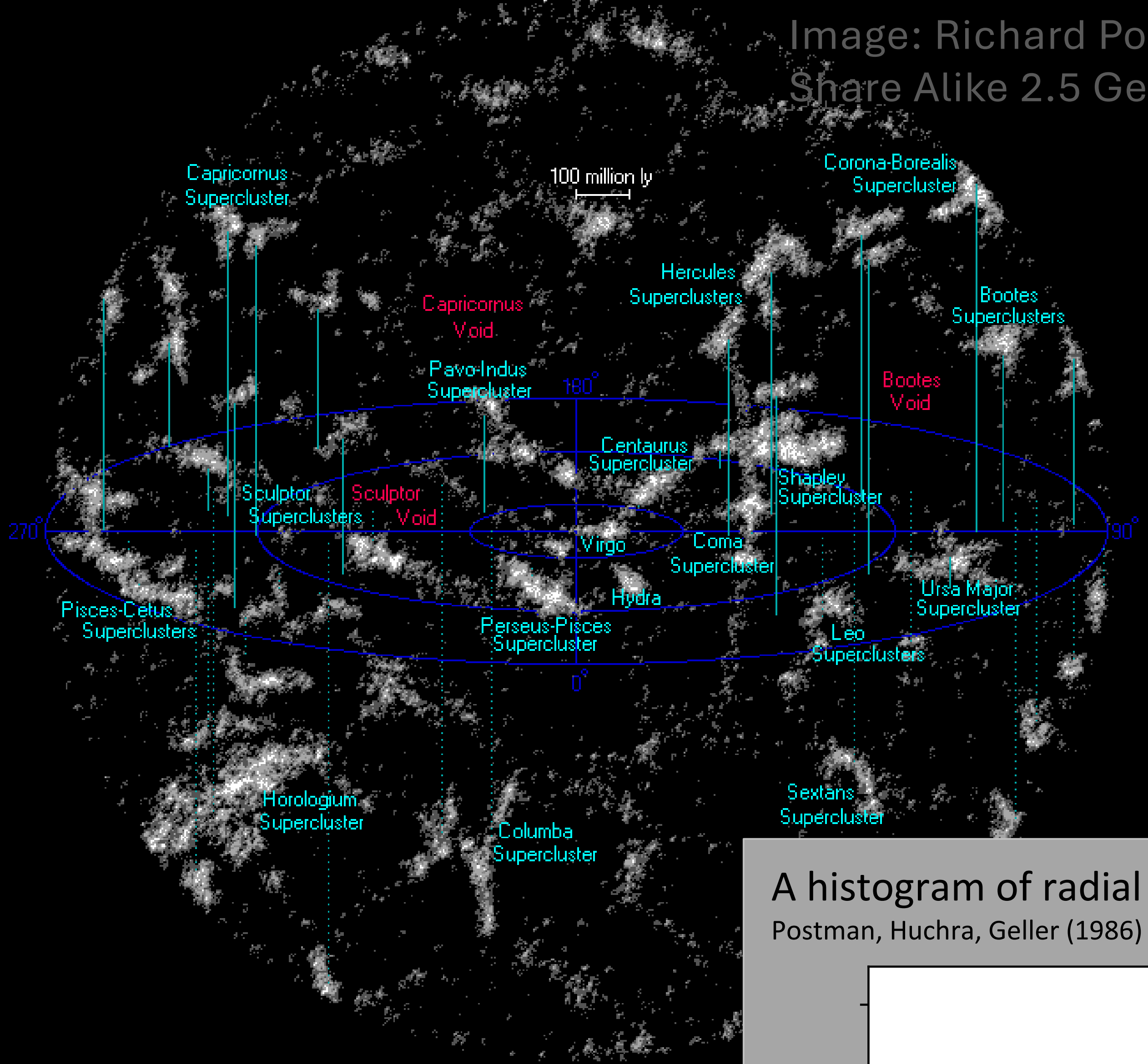
There are standard numerical random variables that you should know:

DISCRETE RANDOM VARIABLES

Binomial $X \sim \text{Bin}(n, p)$	$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ $x \in \{0,1,\ldots,n\}$	For count data, e.g. number of heads in n coin tosses
Poisson $X \sim \text{Pois}(\lambda)$	$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $x \in \{0,1,\ldots\}$	For count data, e.g. number of buses passing a spot
Categorical $X \sim \text{Cat}([p_1, \ldots, p_k])$	$\mathbb{P}(X = x) = p_x$ $x \in \{1,\ldots,k\}$	For picking one of a fixed number of choices

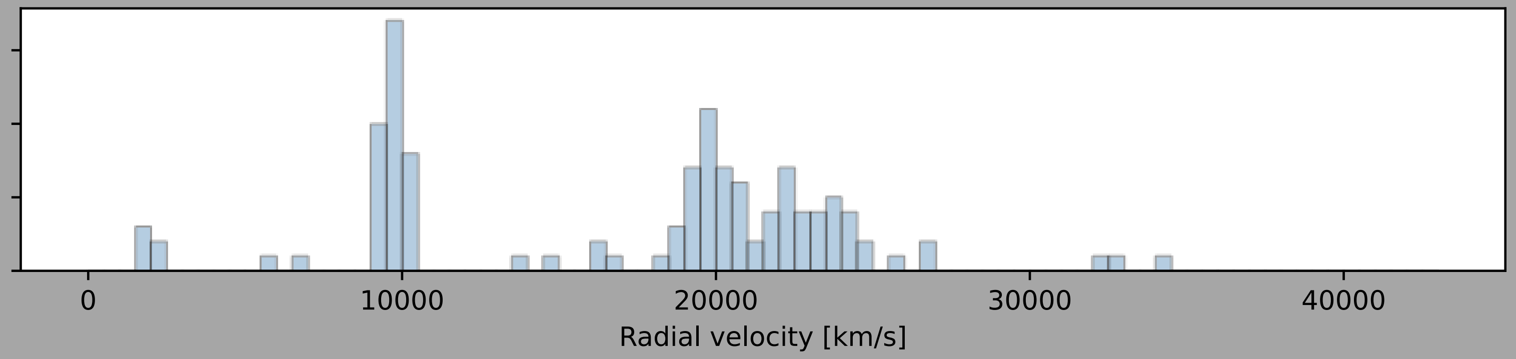
CONTINUOUS RANDOM VARIABLES

Uniform $X \sim U[a, b]$	$\text{pdf}(x) = \frac{1}{b-a}$ $x \in [a, b]$	A uniformly-distributed floating point value
Normal / Gaussian $X \sim N(\mu, \sigma^2)$	$\text{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ $x \in \mathbb{R}$	For data about magnitudes, e.g. temperature or height
Pareto $X \sim \text{Pareto}(\alpha)$	$\text{pdf}(x) = \alpha\, x^{-(\alpha+1)}$ $x \geq 1$	For data about “cascade” magnitudes, e.g. forest fires
Exponential $X \sim \text{Exp}(\lambda)$	$\text{pdf}(x) = \lambda\, e^{-\lambda x}$ $x > 0$	For waiting times, e.g. time until next bus
Beta $X \sim \text{Beta}(a, b)$	$\text{pdf}(x) \propto x^{a-1}(1-x)^{b-1}$ $x \in (0,1)$	Arises in Bayesian inference

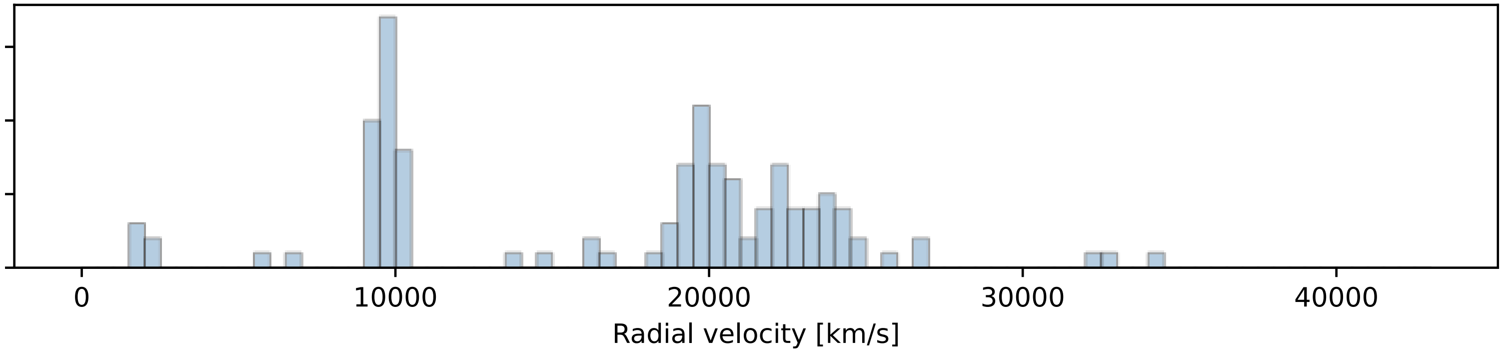


A histogram of radial velocities of 120 galaxies in the Corona Borealis region

Postman, Huchra, Geller (1986)



A histogram of radial velocities of 120 galaxies



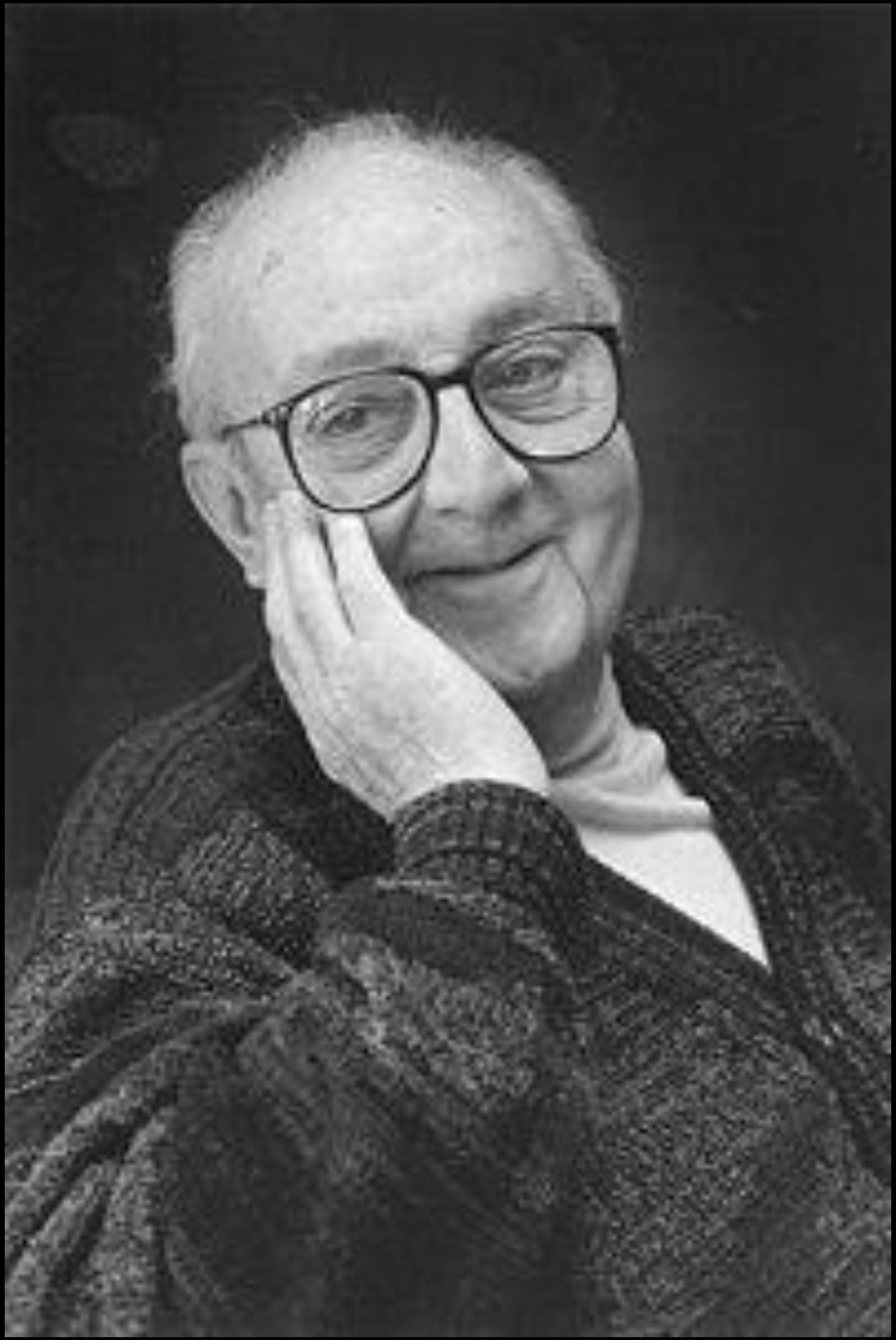
How might you complete this code?

```
def rgalaxy(...):
    # TODO: return a single random galaxy speed

def rgalaxies(size):
    return [rgalaxy(…) for _ in range(size)]
```

$$K = \begin{cases} 1 & \text{w.p. } p_1 \\ 2 & \text{w.p. } p_2 \\ 3 & \text{w.p. } p_3 \end{cases}$$

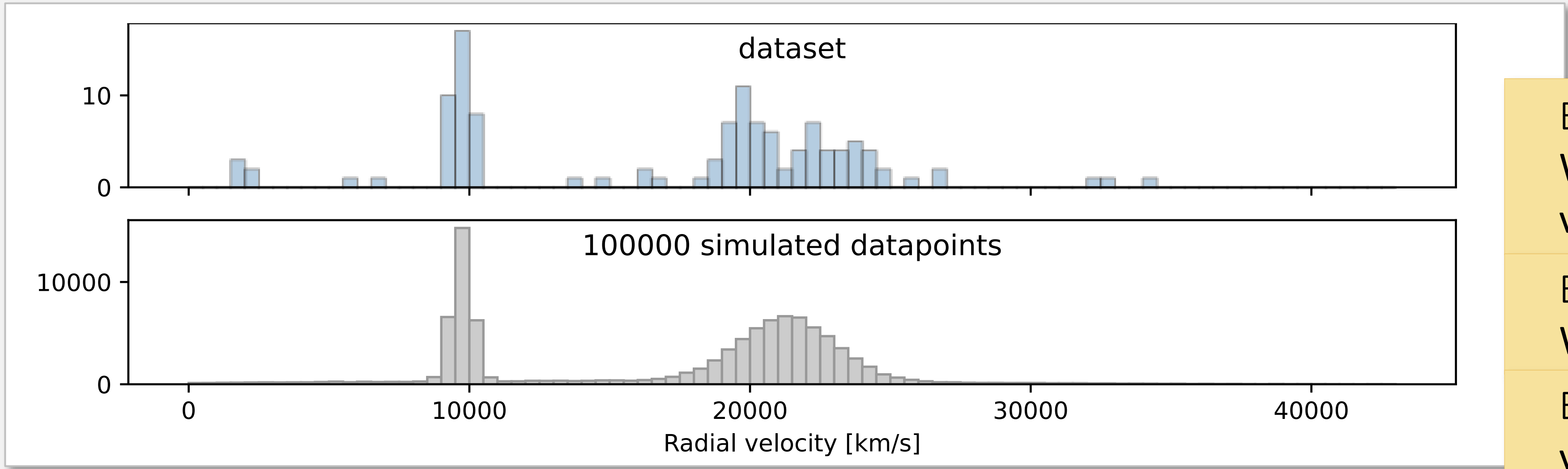
$$X \sim N(\mu_K, \sigma_K^2)$$



George Box
1919–2013

"All models are wrong,
but some are useful"

There's no "right" model. Don't be shy, just go ahead and invent a model!



EXERCISE 1.1.3.
Write this model in random variable notation.

EXERCISE 1.7.4.
What's the likelihood?

EXERCISE 1.7.4. How would you fit this model?

```
def rgalaxy(p,μ,σ):
    k = np.random.choice([0,1,2], p=p)
    x = np.random.normal(loc=μ[k], scale=σ[k])
    return x

def rgalaxies(size, p,μ,σ):
    return [rgalaxy(p,μ,σ) for _ in range(size)]


p = [0.28, 0.54, 0.18]
μ = [9740, 21300, 15000]
σ = [340, 1700, 10600]
```

$K \sim \mathrm{Cat}(p)$ $\quad p = (p_0, p_1, p_2) \in \mathbb{R}^3$

$X \sim N(\mu_K, \sigma_K^2)$.

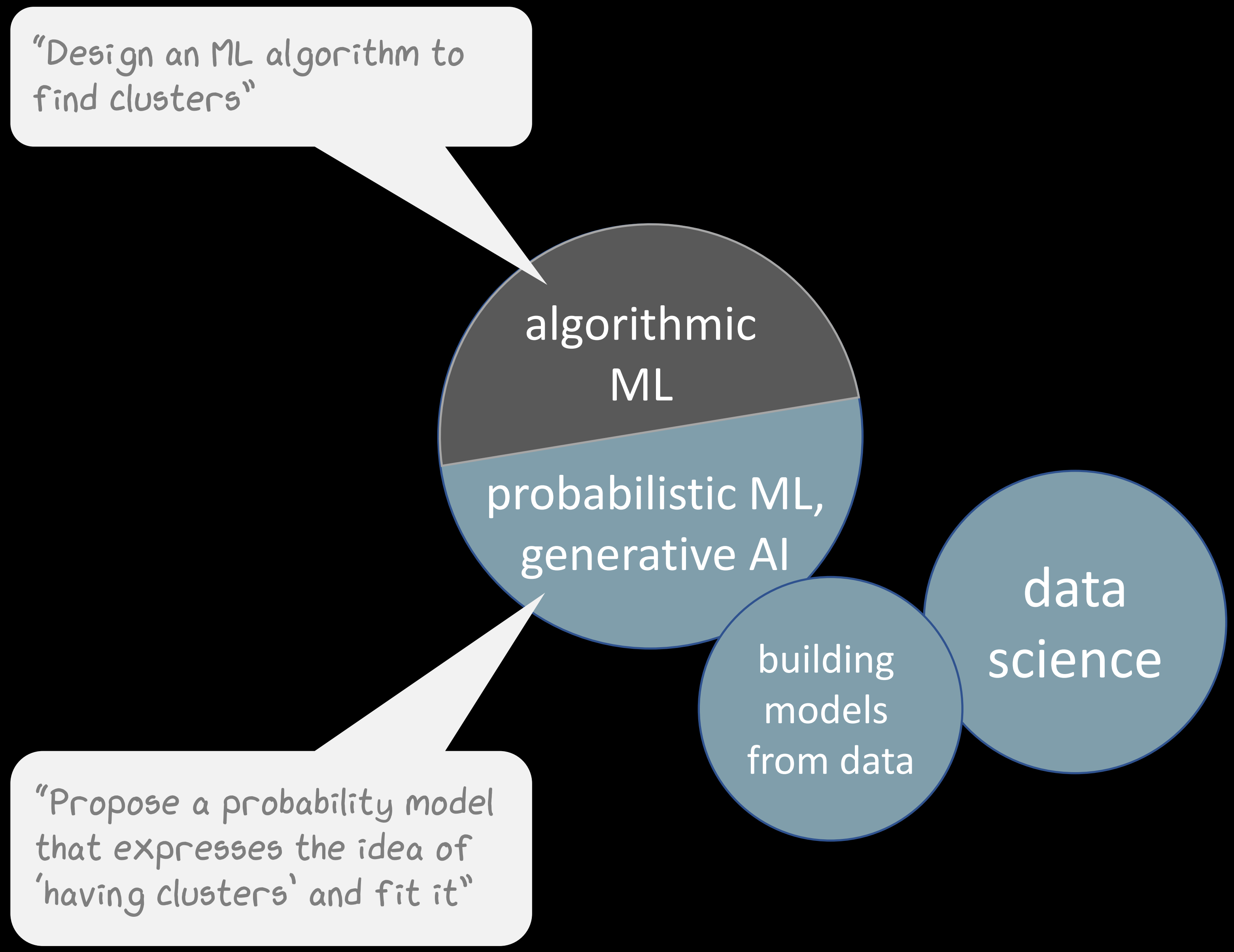
$$\Pr_{K,X}(k,x) = \Pr_K(k)\ \Pr_X(x\,;\,k)$$ [by the rule for sequential generation]

$$= p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(x-\mu_k)^2/2\sigma_k^2}$$ [looking up the likelihood on the r.v. reference sheet]

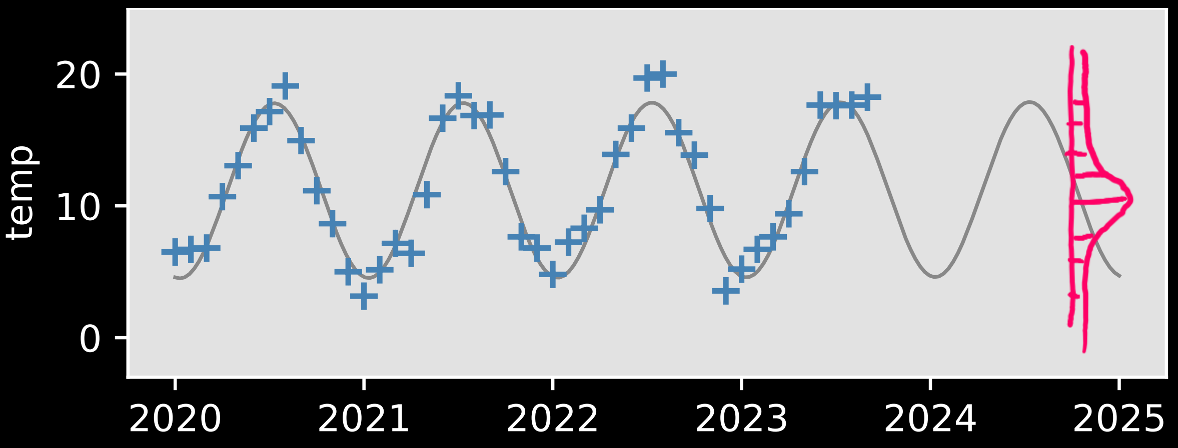
$$\Pr_X(x) = \sum_k p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(x-\mu_k)^2/2\sigma_k^2}$$ [by the law of total probability]

$$\Pr(x_1\cdots x_n\,;\,p,\mu,\sigma) = \prod_{i=1}^{n} \Pr_X(x_i\,;\,p,\mu,\sigma)$$ [since our rgalaxies function makes the datapoints independent]

To fit, we maximize this likelihood over the unknown parameters p, μ, σ


"Design an ML algorithm to find clusters"
algorithmic ML
probabilistic ML, generative AI
building models from data
data science
"Propose a probability model that expresses the idea of 'having clusters' and fit it"

supervised learning

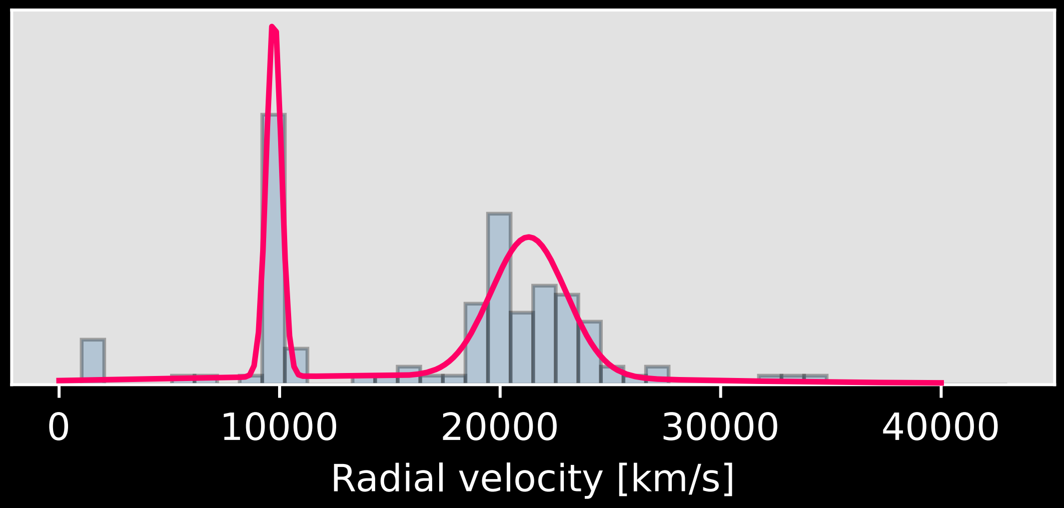


Given a dataset of (t_i, temp_i) pairs, $i \in \{1, \ldots, n\}$,

~~I'd like to predict temp as a function of t.~~

I'd like to fit a probability model for Temp, where the parameters of the distribution depend on t.

unsupervised learning



Given a dataset $[x_1, \ldots, x_n]$ of galaxy speeds,

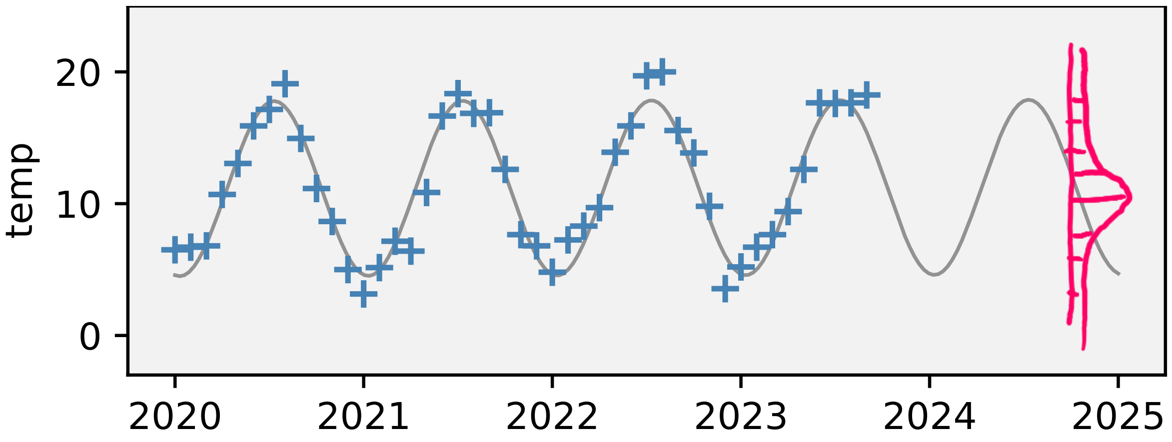
~~I'd like to identify the clusters.~~

I'd like to fit a probability model for speed X, using a model that has clusters.

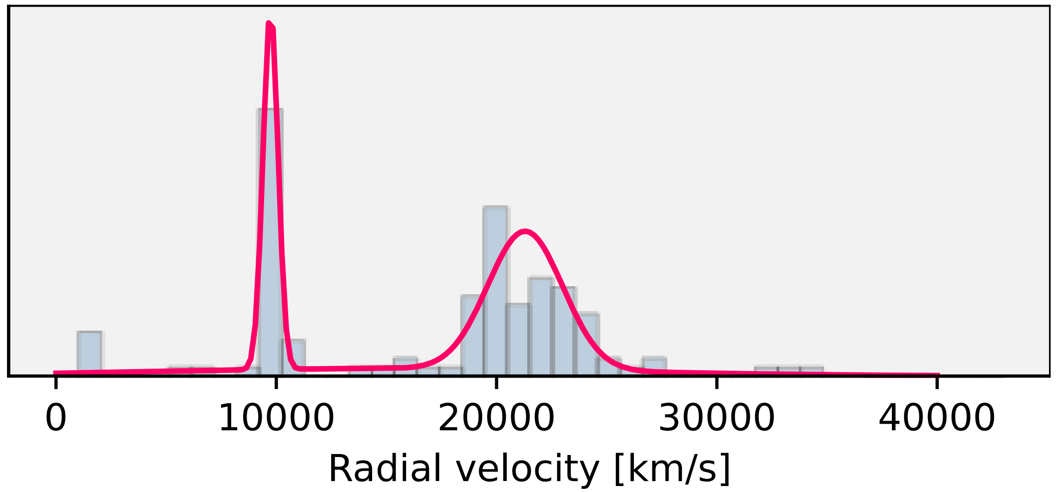
Some terminology:

~~prediction, description, generation~~ probabilistic modelling

supervised learning



unsupervised learning



Given a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ where y_i is the label in record i and x_i is the predictor variable or variables:

Given a dataset $x_1, \ldots, x_n$:

0. Propose a probability model
   $Y \sim \cdots$ or $X \sim \cdots$

1. Write out the likelihood for a single datapoint:
   $\Pr_Y(y; x, \theta)$ or $\Pr_X(x; \theta)$

2. Model the dataset as independent observations and write out the likelihood for the full dataset:
   $\Pr(\text{dataset}) = \prod_{i=1}^{n} \Pr_Y(y_i; x_i, \theta)$ or $\prod_{i=1}^{n} \Pr_X(x_i; \theta)$

3. Learn θ using maximum likelihood estimation

Terminology for supervised learning

station	yyyy	mm	t	af	rain	sun	tmin	tmax	temp
Cambridge	1985	1	1985.00	23	37.3	40.7	-2.2	3.4	0.6
Cambridge	1985	2	1985.08	13	14.6	79	-1.9	4.9	1.5
Cambridge	1985	3	1985.16	10	45.8	97.8	1.1	8.7	4.9
⋮									

called the PREDICTOR variable,
or the FEATURE,
or the COVARIATE

called the RESPONSE,
or the LABEL variable,
or the GROUND TRUTH.

- Here the response is real-valued, so we call it REGRESSION.
- If the response were categorical, we'd call it CLASSIFICATION.

https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data

Exercise 1.6.1 (types of model).

The MNIST database of handwritten images consists of records (x_i, y_i) where $x_i \in \mathbb{R}^{28\times 28}$ is a greyscale image with 28×28 pixels, and $y_i \in \{0, \ldots, 9\}$ is the digit.

Give examples of predictive and generative modelling tasks based on this data.

image	digit
	2
	1
	3
	1
	4

Data from http://yann.lecun.com/exdb/mnist/

- Simple prediction (**image classifier**): given an image x, output the digit y
- Generative prediction (**probabilistic image classifier**): given an image x, output a distribution over digits Y
- Generative prediction (**handwriting generator**): given a digit y, generate a random image X
- Generative prediction (**image infill**): given a partial image, generate a complete image X
- Pure generation: generate a random image X, of any digit
- Pure generation: generate a random pair (X, Y) with joint likelihood $\Pr_{X,Y}(x,y)$.

Then, the marginal distribution $\mathbb{P}(Y=y \mid X=x)$ corresponds to the probabilistic image classifier

$\Pr_X(x \mid Y=y)$ corresponds to the handwriting generator.

On Friday we'll do a mock exam question.
Have a look at it beforehand!

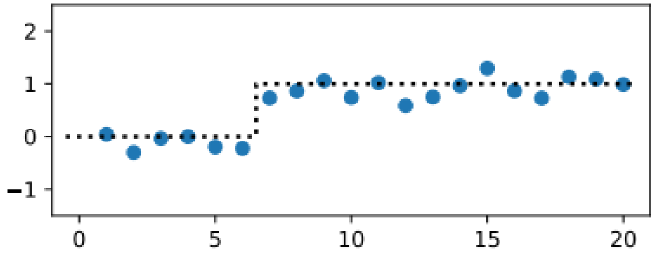
COMPUTER SCIENCE TRIPOS Part IB – mock – Paper 6

1 Foundations of Data Science (DJW)

(*a*) A 0/1 signal is being transmitted. The transmitted signal at timeslot $i \in \{1,\ldots,n\}$ is $x_i \in \{0,1\}$, and we have been told that this signal starts at 0 and then flips to 1, i.e. there is a parameter $\theta \in \{1,\ldots,n-1\}$ such that $x_i = 1_{i>\theta}$. The value of this parameter is unknown. The channel is noisy, and the received signal in timeslot i is

$$Y_i \sim x_i + \mathrm{Normal}(0, \varepsilon^2)$$

where ε is known.



(*i*) Given received signals $(y_1,\ldots,y_n)$, find an expression for the log likelihood, $\log \Pr(y_1,\ldots,y_n\,;\,\theta)$. Explain your working. [5 marks]