Computer Science

IB Data Science

Lecturer

Dr Damon Wischik

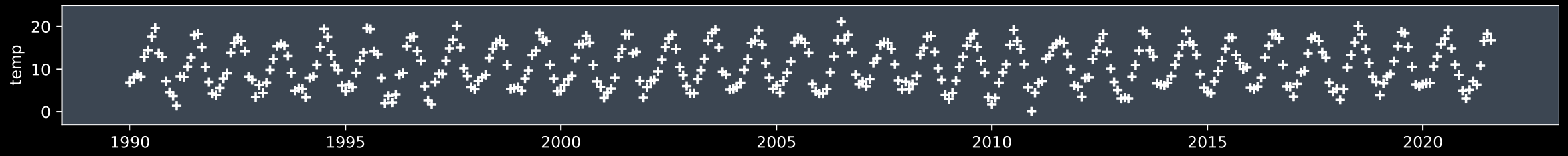
Met Office climate dataset

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

Monthly readings from 37 weather stations around the country. Let's look at Cambridge, from 1990.

station	yyyy	mm	t	af	rain	sun	tmin	tmax	temp
Cambridge	1990	1	1990.00	0	43.8	64.7	4.0	9.8	6.90
Cambridge	1990	2	1990.08	1	71.1	102.0	4.7	11.4	8.05
Cambridge	1990	3	1990.16	3	23.2	153.2	4.7	12.9	8.80

⋮

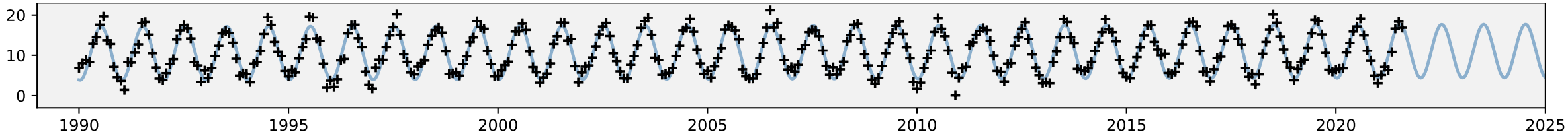


What model / formula would you suggest to fit this dataset?

```
def temp_model(t, …):
    return …
```

A SCIENTIST'S DETERMINISTIC MODEL

```
def temp_model(t, α, ϕ, c, γ):
    return c + α * np.sin(2*π*(t+ϕ)) + γ*t
```

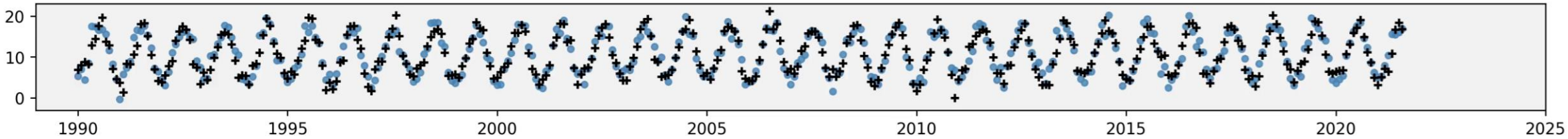


why?

- to describe the data in front of my eyes!
- to tell my fitting routine how much attention to pay to errors/outliers
- to reason about confidence, e.g. "$\gamma = 1.4°C$ (only $\pm$ 0.7°C"

A DATA SCIENTIST'S PROBABILITY MODEL

```
def rtemp(t, α, ϕ, c, γ, σ):
    pred = c + α * np.sin(2*π*(t+ϕ)) + γ*t
    return np.random.normal(loc=pred, scale=σ)
```



All of machine learning is based on a single idea:

1. Write out a probability model
2. Fit the model from data

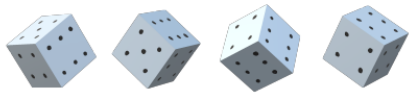
This is behind

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

LECTURE NOTES

IB Data Science:
modelling and machine learning with probability

Damon Wischik, Computer Laboratory, Cambridge University



Contents

- ABRIDGED NOTES
  (contain all examinable material + extras)
- EXTENDED NOTES
  (goes to masters-level, available soon)

- The handout has more wordy explanations and more examples than lectures
- Use the handout like a textbook and take your own notes during lectures

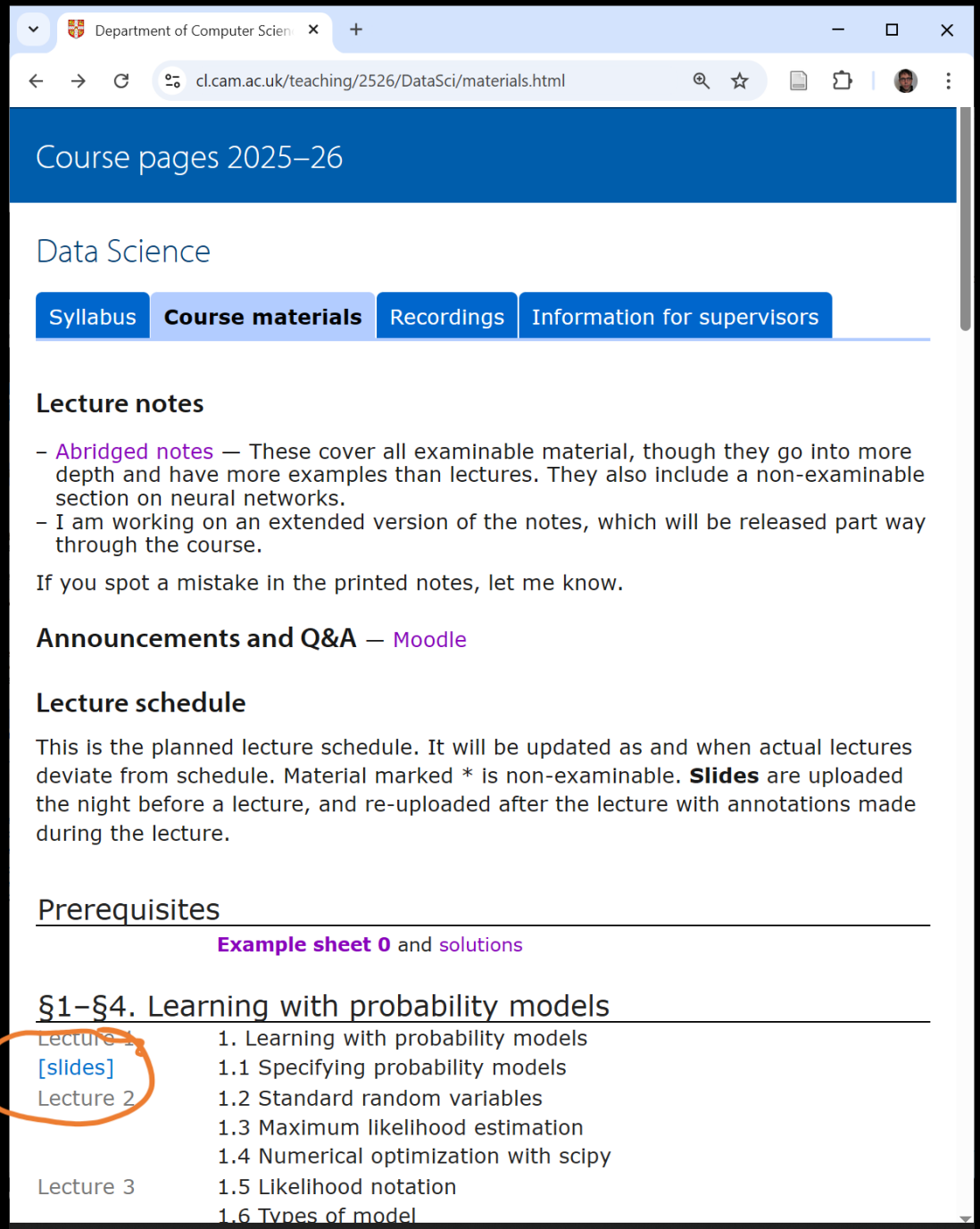
Do you want printouts?

$\frac{1}{4}$ of you: yes.

§x

- Slides for each lecture are on the website
  and most slides say which section they're for

- What's examinable?
  Everything in the lecture schedule,
  except for sections marked ∗



Consent to recordings of live lectures with Panopto

https://www.educationalpolicy.admin.cam.ac.uk/policy-index/recording

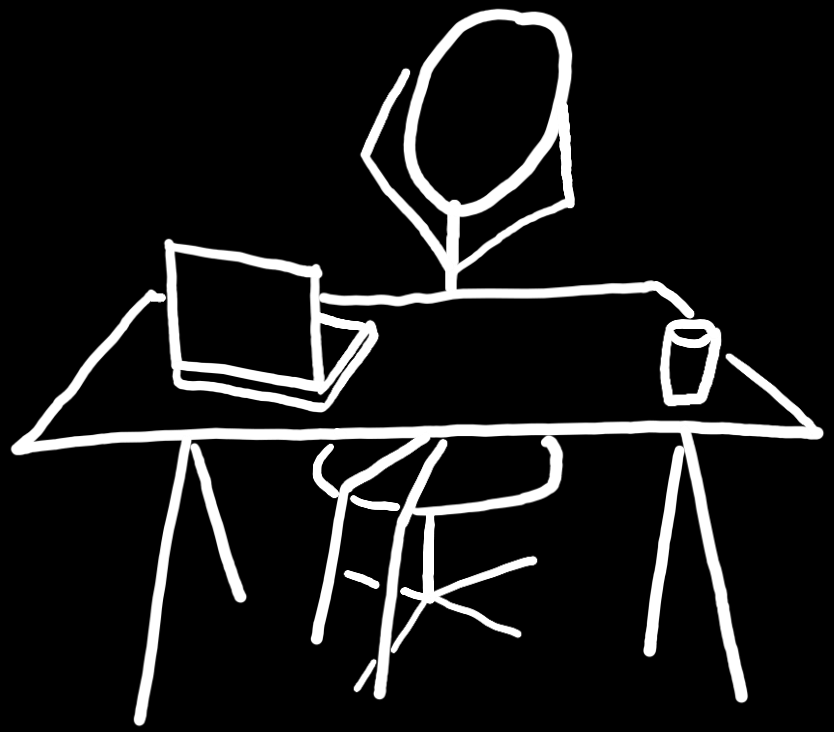
For any teaching session where your contribution is mandatory or expected, we must seek your consent to be recorded.

You are not obliged to give this consent, and you have the right to withdraw your consent after it has been given.

Do you give your consent to recordings?

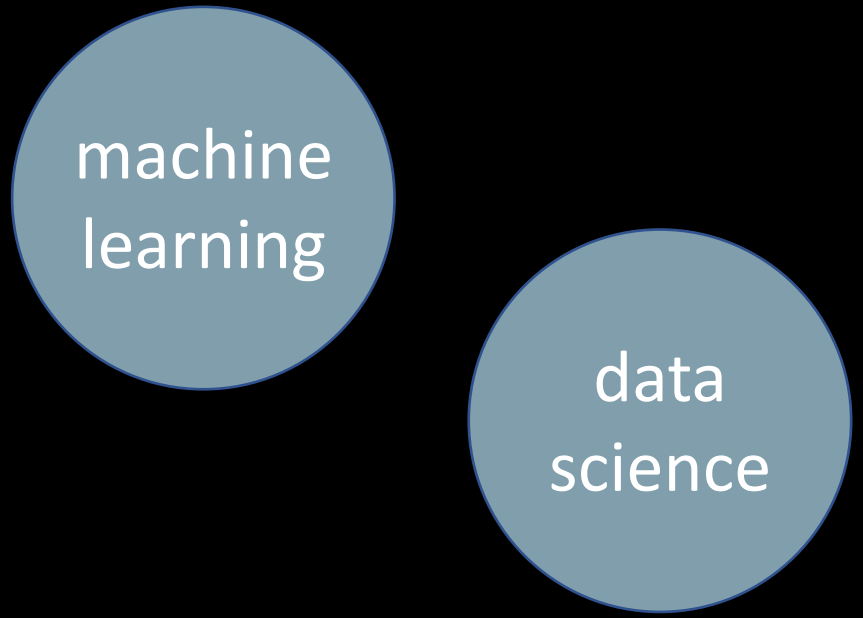
Yes.

- I’ll be in the cafe area 1–1.30pm today for office hours.



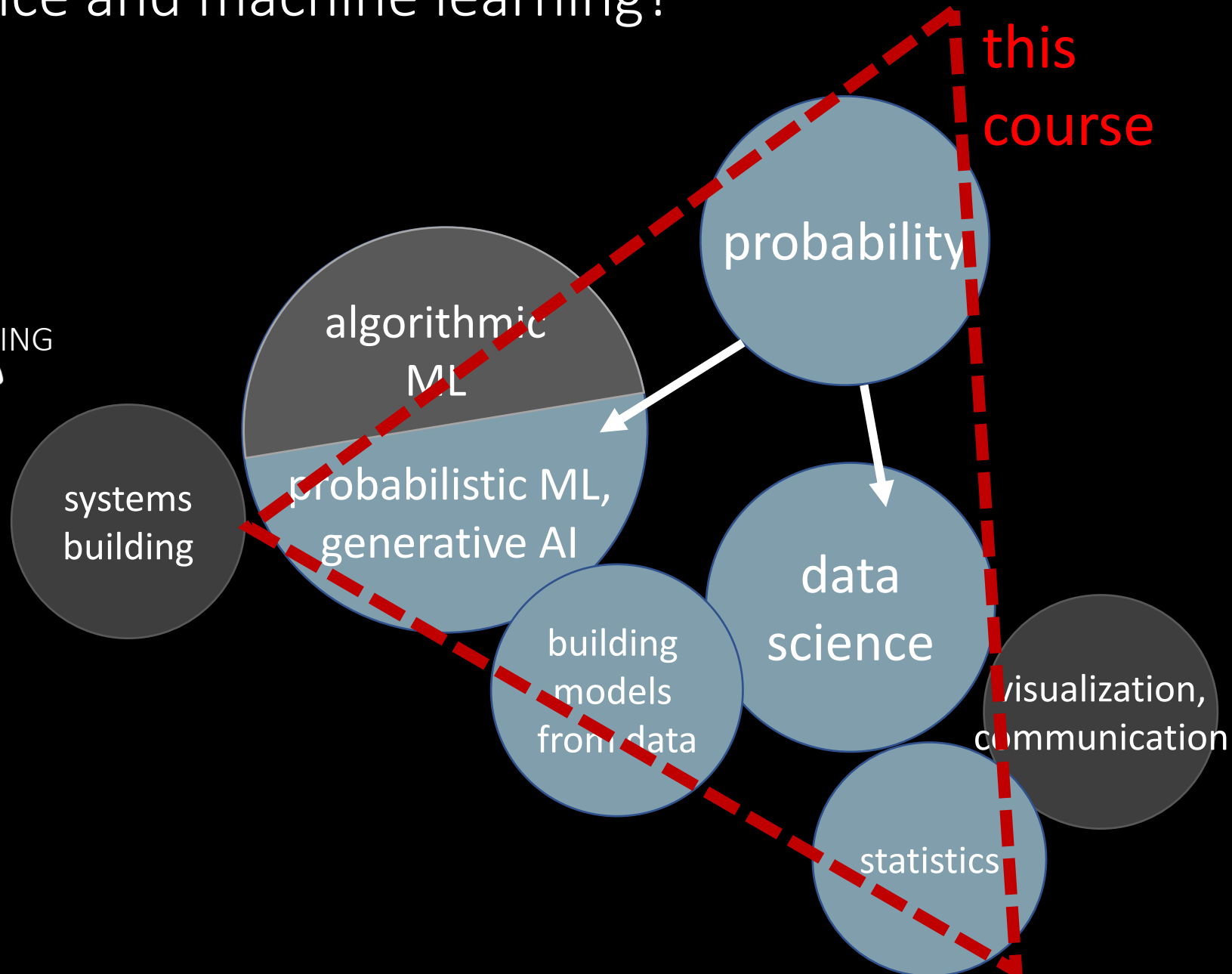
What is data science? What's the difference between data science and machine learning?

MACHINE LEARNING
Building systems
that learn
from data.



DATA SCIENCE
Finding
patterns in a
dataset.

What is data science? What's the difference between data science and machine learning?



MACHINE LEARNING

~~Building systems that learn from data.~~

Building models and applying them.

DATA SCIENCE

~~Finding patterns in a dataset~~

Building models to learn about the dataset.

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

People who try to do ML without probability

Example sheet 0

Prerequisites

IB Data Science—DJW—2025/2026

This course assumes that you know how to handle basic probability problems and that you know about random variables, as taught in IA *Introduction to Probability*. It also assumes that you know how to find the maximum or minimum of a function, using calculus, as taught in IA *Maths for NST*. The code snippets in the course are in Python and numpy, and you should be familiar with numpy's way of writing vectorized computations.

This example sheet reviews the material that you need to know. Please look through, and make sure you remember how to answer these questions! Solutions are provided on the course website. *For supervisors: this example sheet is not intended for supervision.*

Rules of probability (IA Probability lecture 1)

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 ...

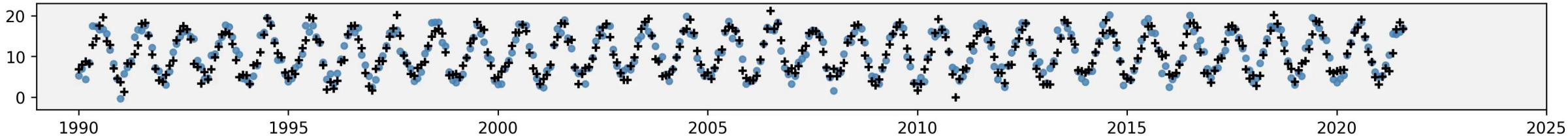
Conditional probability, or equivalently the chain rule:

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

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

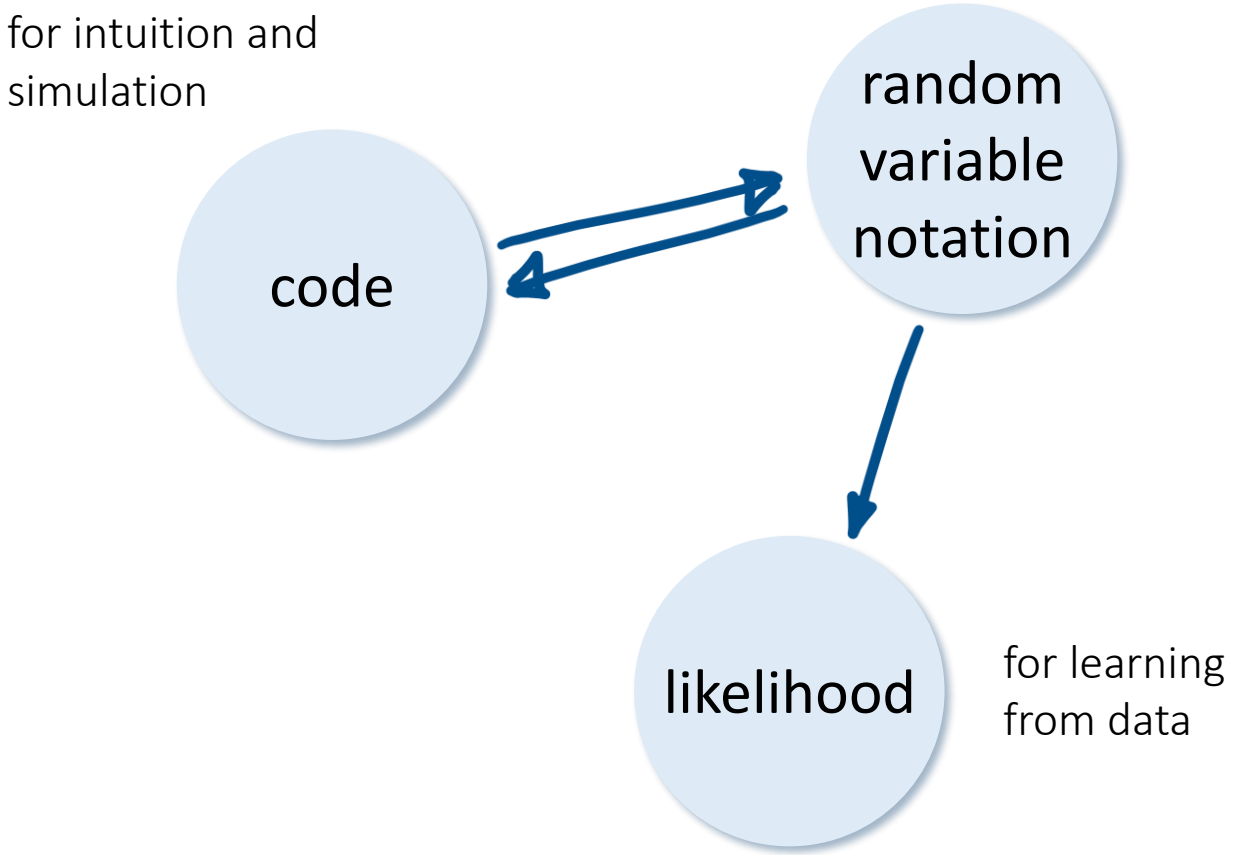
- Example sheet 0 is to remind you about IA Probability, Maths for NST, and Scientific Computing
- It's not for supervision; solutions are provided

How to specify a probability model



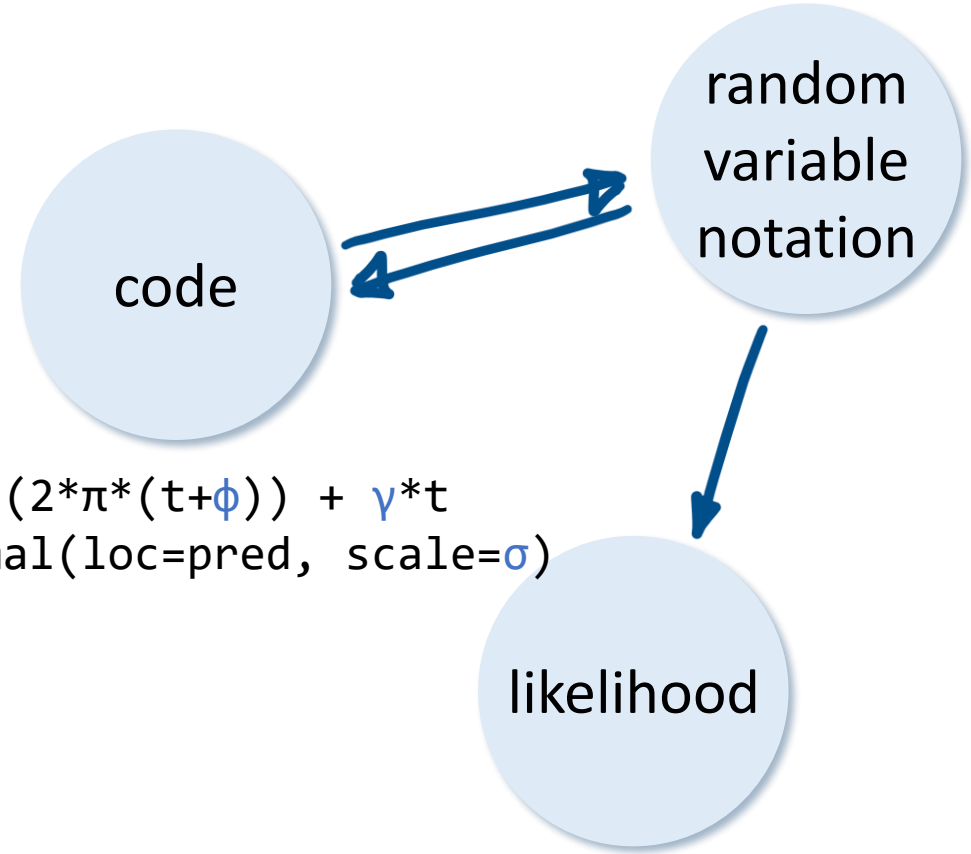

```
def rtemp(t, α=10, ϕ=-0.25, c=11, γ=0.035, σ=2):
    pred = c + α * np.sin(2*π*(t+ϕ)) + γ*t
    return np.random.normal(loc=pred, scale=σ)
```

Three views of a probability model



Three views of a probability model

$$\text{Temp}_i \sim \alpha \sin\big(2\pi(t_i + \varphi)\big) + c + \gamma t_i + \text{Normal}(0, \sigma^2),$$
$$i \in \{1, \ldots, n\}$$




```
def rtemp(t, α,ϕ,c,γ,σ):
    pred = c + α * np.sin(2*π*(t+ϕ)) + γ*t
    return np.random.normal(loc=pred, scale=σ)
```

```
def ry():
    x = random.random()
    y = x ** 2
    return y
```

$X \sim U[0,1]$
$Y = X^2$

Generate X from the Uniform dist.

```
def ri(a,b):
    x = random.random()
    i = math.floor(a*x+b)
    return i
```

$X \sim U[0,1]$
$I = \lfloor aX + b \rfloor$

Upper case for random variables
Lower case for parameters, constants, datapoints.

```
x = random.random()
y = x **2
```

$X \sim U[0,1]$
$Y = X^2$

```
def rz():
    x1 = random.random()
    x2 = random.random()
    return x1 * math.log(x2)
```

$X_1, X_2 \sim U[0,1]$
$Z = X_1 \log X_2$

X_1 and X_2 are generated independently.

"learning the value of one tells us nothing about the other".

unless specified otherwise

```
def rmyrandpair():
    x1 = random.random()
    x2 = random.random()
    y,z = (x1+x2, x1*x2)
    return (y,z)
```

$(Y, Z) \sim$ Myrandpair

λ is lower-case (not Λ) so it refers to a fixed value.

When we say "X_1 and X_2 are independent" we mean "given all the parameters".

```
λ = 3
x1 = random.uniform(0,λ)
x2 = random.uniform(0,λ)
```

$X_1, X_2 \sim U[0, \lambda]$

```
x = random.random()
y = 1 - x
```

$X \sim U[0,1]$
$Y = 1 - X$

~ means "has the same distribution"

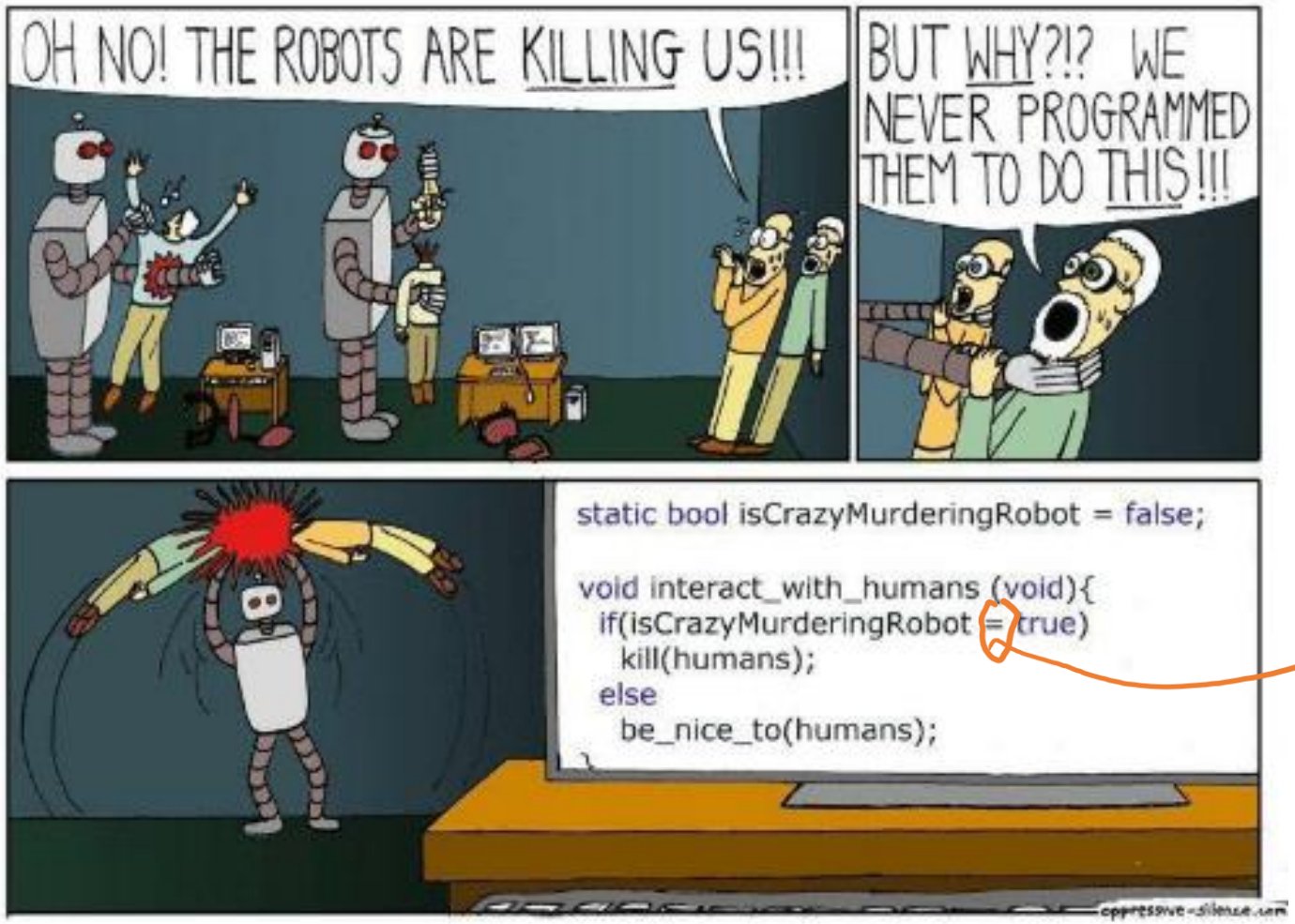
$X \sim U[0,1]$
$Y \sim U[0,1]$
$X \sim Y$
$X \sim 1-X$

} these are all true statements.

= means "always have the same value whenever I run the code". Not assignment!

$Y = 1-X$
$X+Y = 1$
$X = 1-Y$

} all true statements



In Java, = is for assignment.

```
x = random.random()
y = np.random.normal(
        loc=x, scale=0.1)
```

$$X \sim U[0,1]$$
$$Y \sim N(X, 0.1^2)$$

This is how we indicate that Y is generated based on X.

It's good to accompany this with a "causal diagram" $X \to Y$.

(We'll see lots of these at the end of the course.)

```
def rtemp(t, α=10, ϕ=-0.25, c=11, γ=0.035, σ=2):
    pred = α*np.sin(2*π*(t+ϕ)) + c + γ*t
    return np.random.normal(loc=pred, scale=σ)

df = pandas.read_csv(...)  # data frame, n=380 rows
Temp = rtemp(df.t)
```

numpy vectorized code: generates 380 independent Normal random variables

380 equations, one for each i, each with its own Normal. They're all independent, by our notation convention

$$\text{Temp}_i \sim \alpha \sin\big(2\pi(t_i + \varphi)\big) + c + \gamma t_i + \text{Normal}(0, \sigma^2), \qquad i \in \{1, \ldots, n\}$$

$$\text{Temp}_i = \alpha \sin\big(2\pi(t_i + \varphi)\big) + c + \gamma t_i + E_i, \qquad E_i \sim \text{Normal}(0, \sigma^2), \qquad i \in \{1, \ldots, n\}$$

What are the r.v.? Temp_i, Normal for eqn i

What are the non-r.v.? parameters: $\alpha, \varphi, c, \gamma, \sigma$

others: t_i

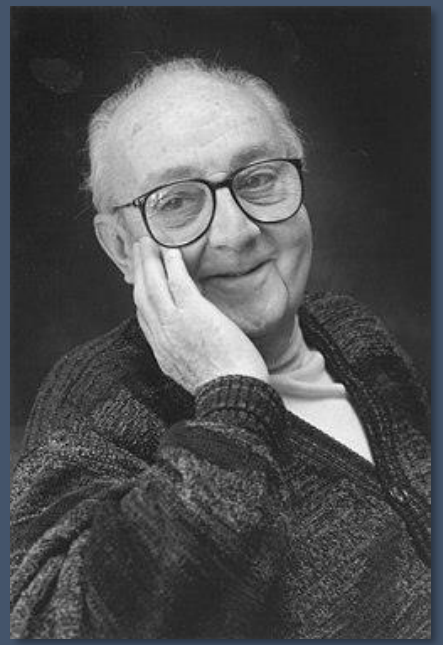
All of machine learning is based on a single idea:

1. Write out a probability model ← today
2. Fit the model from data ← next lecture

This is behind
- A-level statistics formulae
- our climate model
- ChatGPT training

§1

A core skill is being able to build useful probability models. During this course you will pick up this skill, through examples.



"All models are wrong but some are useful"
George Box

... so don't be shy when building your own. There *is* no "correct model".