Computer Science

IB Data Science

Lecturer

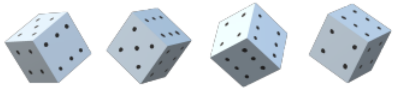
Dr Andrea Marinoni

am2920@cam.ac.uk

HANDOUT

Likelihood:
modelling and machine learning with probability

Damon Wischik, Computer Laboratory, Cambridge University



Contents

- ABRIDGED NOTES
(contain all examinable material)
- EXTENDED NOTES
(contain all examinable material + extras)

- For more printouts, ask student admin

- The handout has more wordy explanations and more examples than lectures

- Use the handout like a textbook and take your own notes during lectures

TIMETABLE

1 October 2024

DAY-BY-DAY COMPUTER SCIENCE TIMETABLE 2024–25

PARTS IA, IB AND II

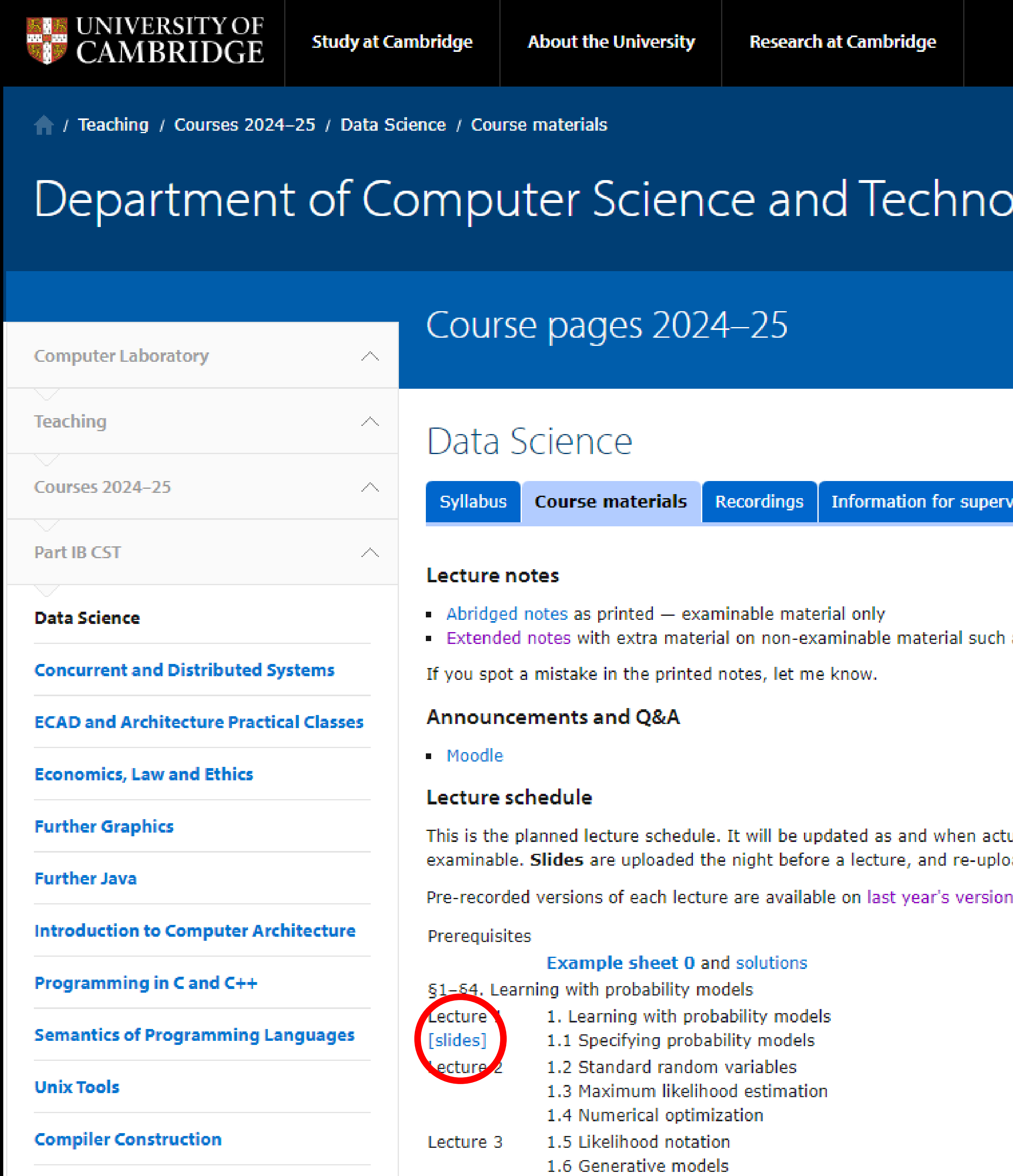
MICHAELMAS TERM 2024

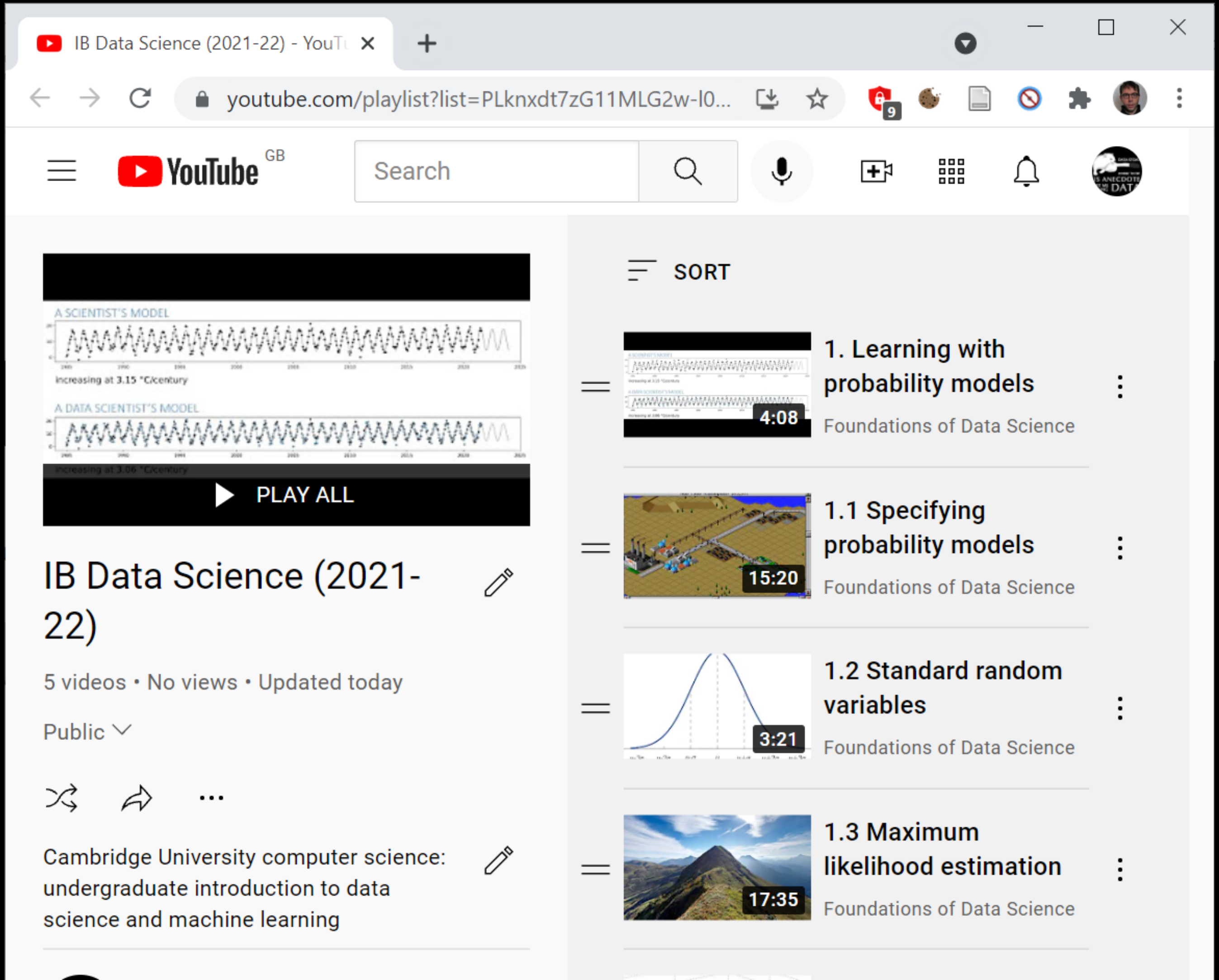
```
          --14 Oct --21 Oct --28 Oct --4 Nov  --11 Nov--18 Nov--25 Nov--2 Dec
09:00-10:00
 P   V   --      --      --      --      --      --K K K--K K K--K K MGK   UnixTools
 P   S K -- K K -- K K -- K K -- K K -- K K -- K K -- K K -- K  MK    ConcDisSys
  QM F F --    F --    F --    F --    F --    F --    F --    F --    MPF   CAT        2h
  Q  L K -- K    --      --      --      --      --      --      --    MGK   TeX+Julia
  Q  L   --      --      -- L L--  L L--  L L--  L L--  L L--    ML-B  DenotSem
10:00-11:00
N    C  M--M M M--M M M--M M M--M M  --      --      --      --    AVSM  FoundsCS
N    C   --      --      --      --   F--F F F--F F F--F F F--F F MPF   DiscMath
 P   L   --M M M--M M M--M M J--J J J--  J J--J J --      --    RM+   IntComArch
 P   L O --    O --    O --    O --      --      --      --      --    ACO   FGraphics  2h
 P   L   --      --      --      --   I -- I I -- I C -- C B -- B  KI+   EconLaw
 P   S   --      --      --      --     --G      --      --      --    TG    GroupProj
  QM W  S--   S  --  S S--S S  --S S  --      --      --      --    WS+   NLP
  QM T   --S   S--S    --      --      --      --      --      --    WS+   NLP
  QM E   --      --      --   S--      --      --      --      --    WS+   NLP
  QM S   --      --      --   L--L L L--L L L--L L L--      --    NDL+  ADS

11:00-12:00
N    C  W--W W W--W W W--W W W--W W  --      --      --      --    IJW   DigElec
N    C   --      --      --      --   H--H H H--H H H--H H H--H H RKH   OOProg
N    S   -- G    -- G    -- G    -- G    --      --      --      --    DJG   Databases
N    S   --      --      --      --   Z -- Z Z -- Z Z -- Z Z -- Z  EZ    Graphics
 P   S  M--M M M--M M M--M M M--M M  --M M  --M M  --      --    AM    DataSci
  Q  L  K--K K K--K K K--K K K--K K  --      --      --      --    NK    Types
  Q  L   --      --      --      --     --M M  --M M  --M M  --M M SAM   Business
  Q  L C -- C C -- C C -- C C -- C C -- C C -- C C -- C C -- C  JAC   PrincComm
```

§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 $*$





- Pre-recorded videos from 2021-22 are on YouTube
- All examinable material is in these videos
- For recordings of lectures …

Consent to recordings of live lectures

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

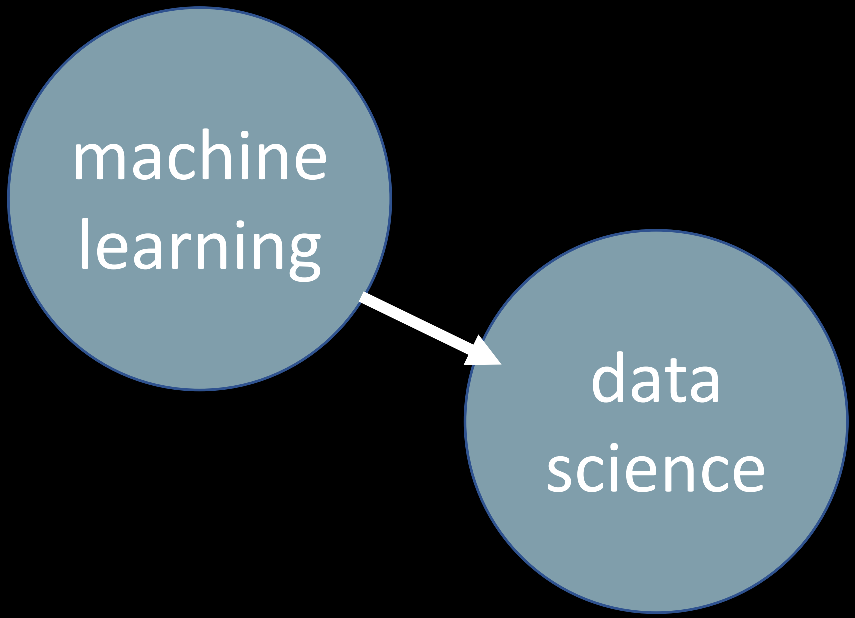
For any teaching session where your contribution is mandatory or expected, we must seek your consent to be recorded.

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?

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

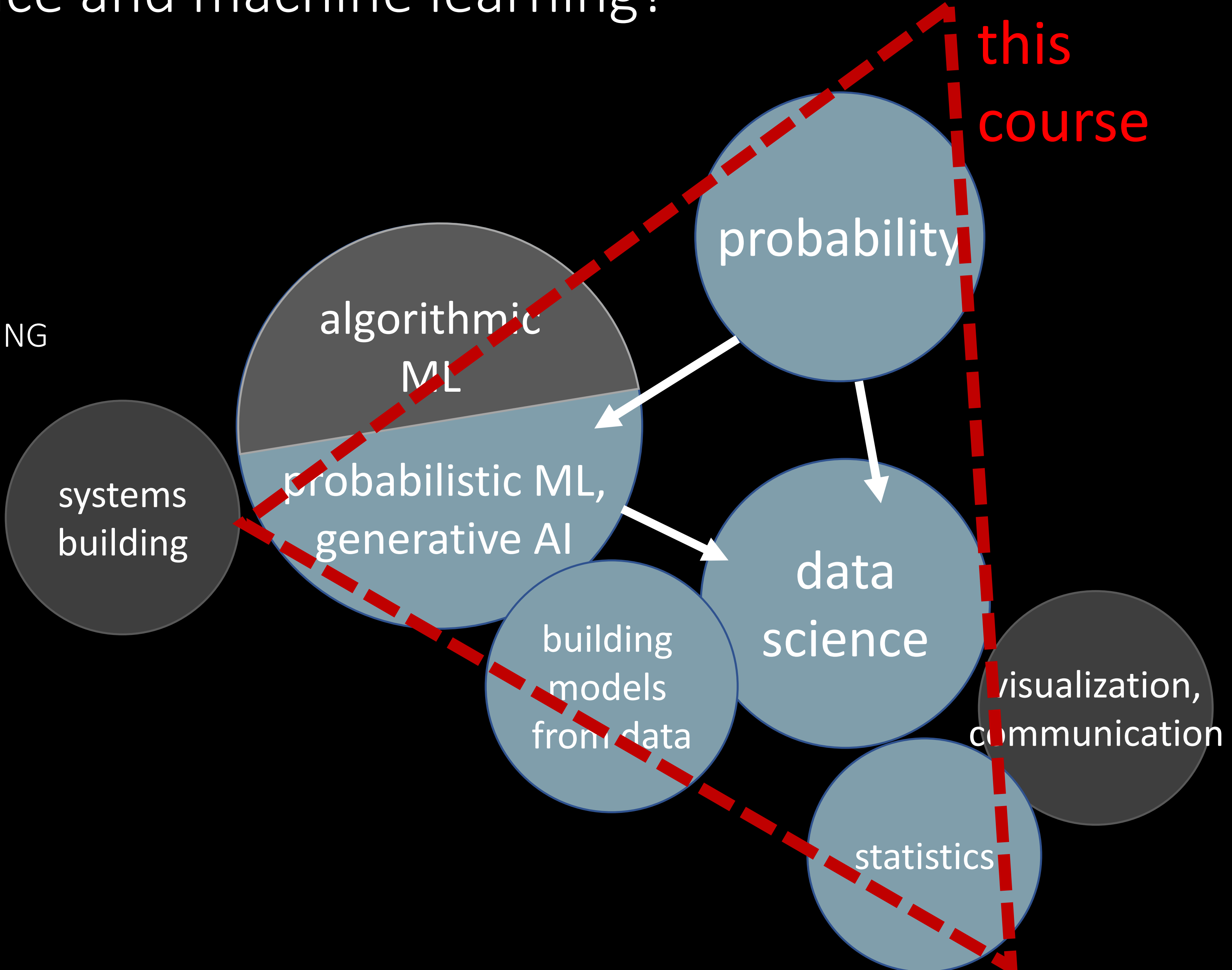
MACHINE LEARNING
Building systems that learn from data.



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 the behaviour.

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

Example sheet 0

Prerequisites

IB Data Science—DJW—2023/2024

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

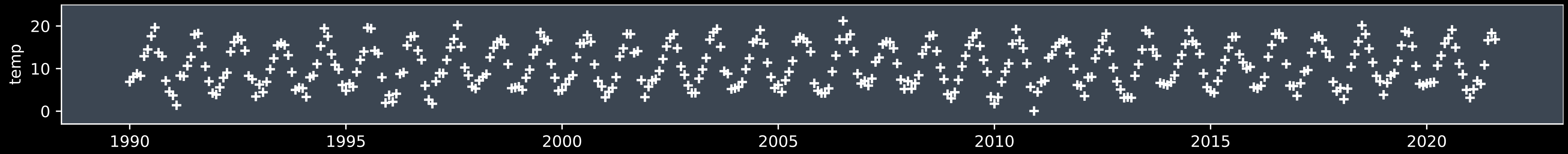
Met Office climate dataset

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

⋮

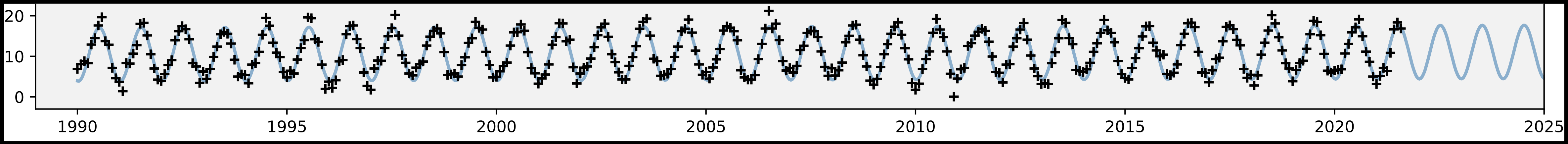


What model / formula would you suggest to fit this dataset?

```
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
```



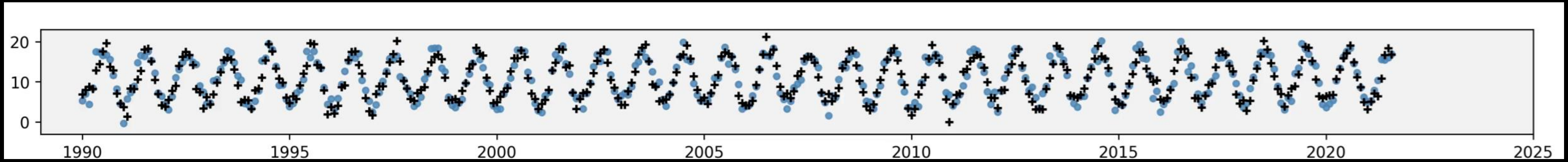
why? To describe the data in front of me!

To tell my fitting procedure how much attention to pay to outliers

To be able to say "This model can't really fit the data"

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)
```



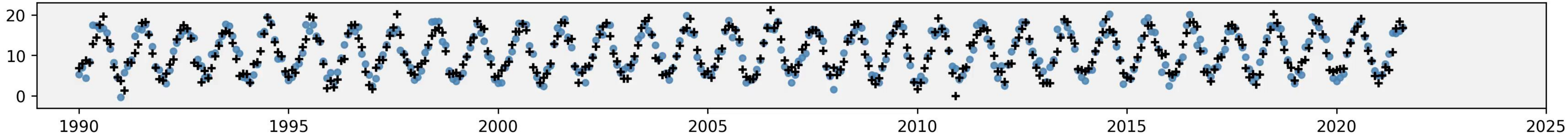
All of machine learning is based on a single idea:

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

This is behind

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

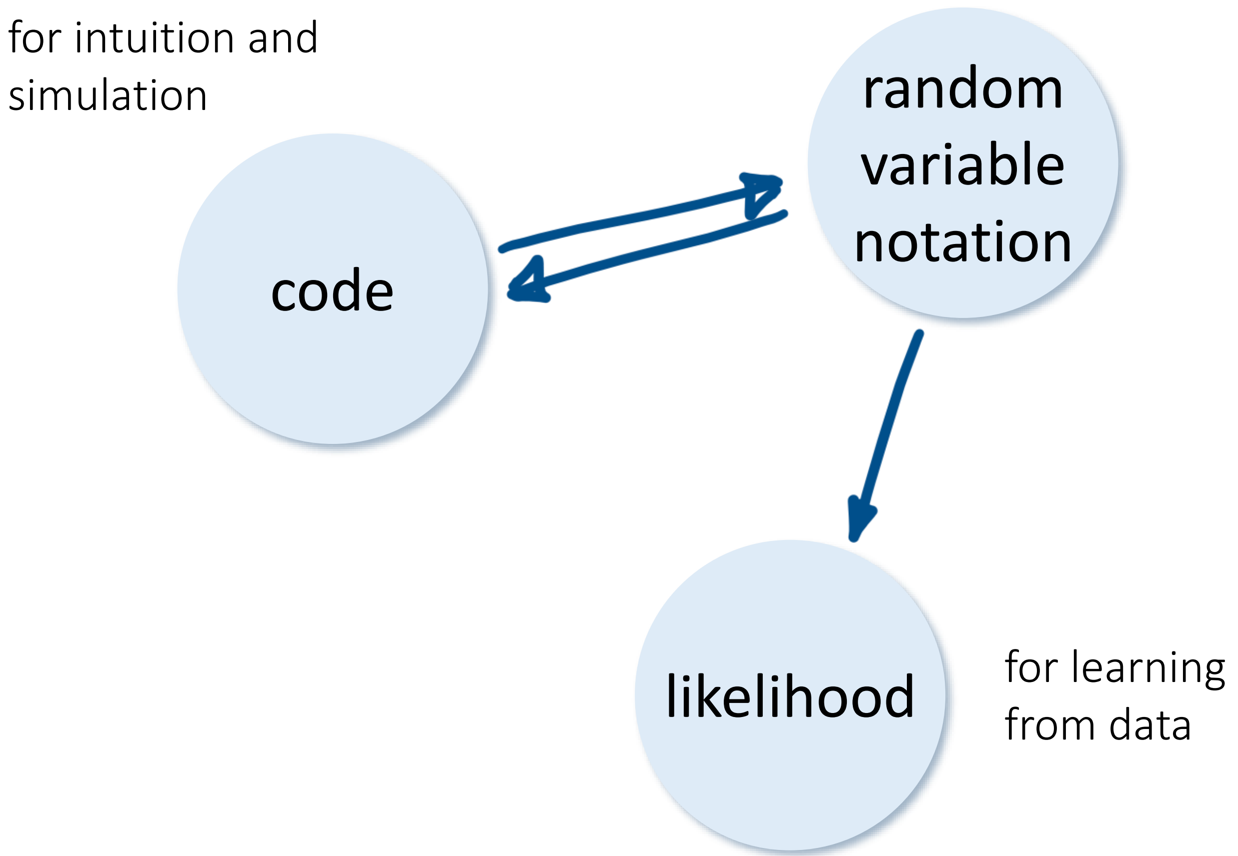
How to specify a probability model




```
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)
```

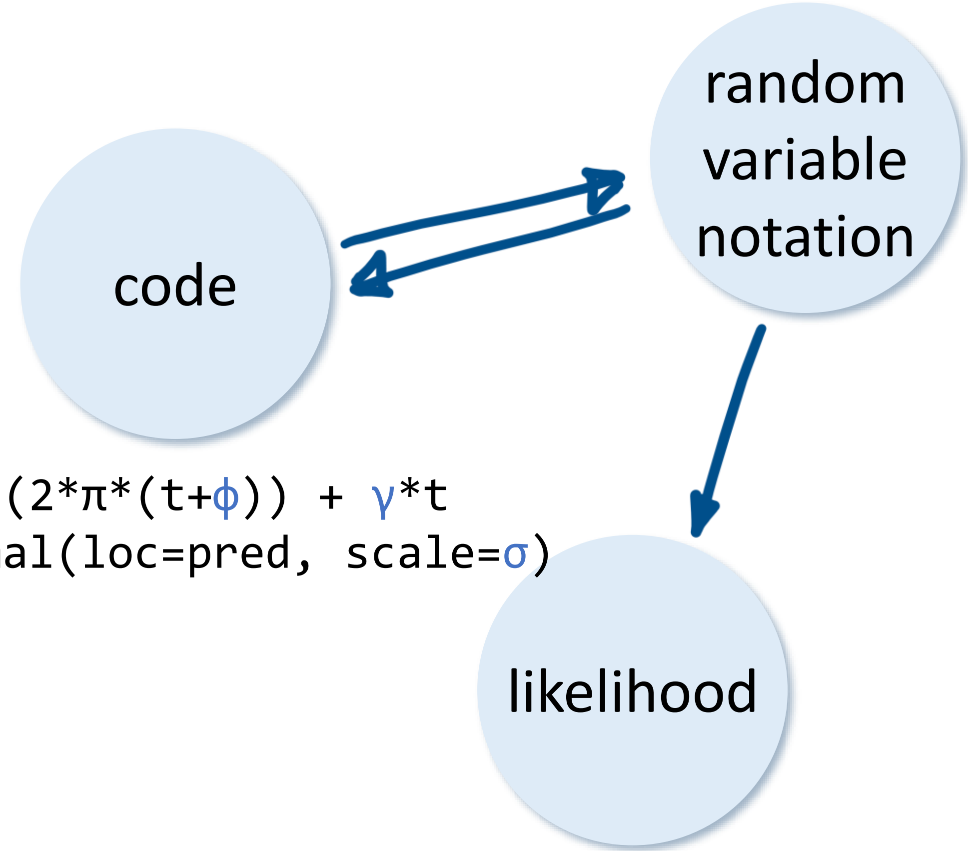
A probability model is a piece of code
where the output is random.

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)
```

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

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

Generate X from the Uniform distribution.

```python
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: random variables
Lower case: parameters, constants.

```python
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 are generated independently"
— knowing the value of one tells us nothing about the value of the other.

In random variable notation, assume independence unless specified otherwise (like this)

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

$(Y, Z) \sim \text{Myrandpair}$

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

λ is lower-case, so it refers to a fixed value

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

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

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

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

$\sim$: "has the same distribution"
" has the same histogram "

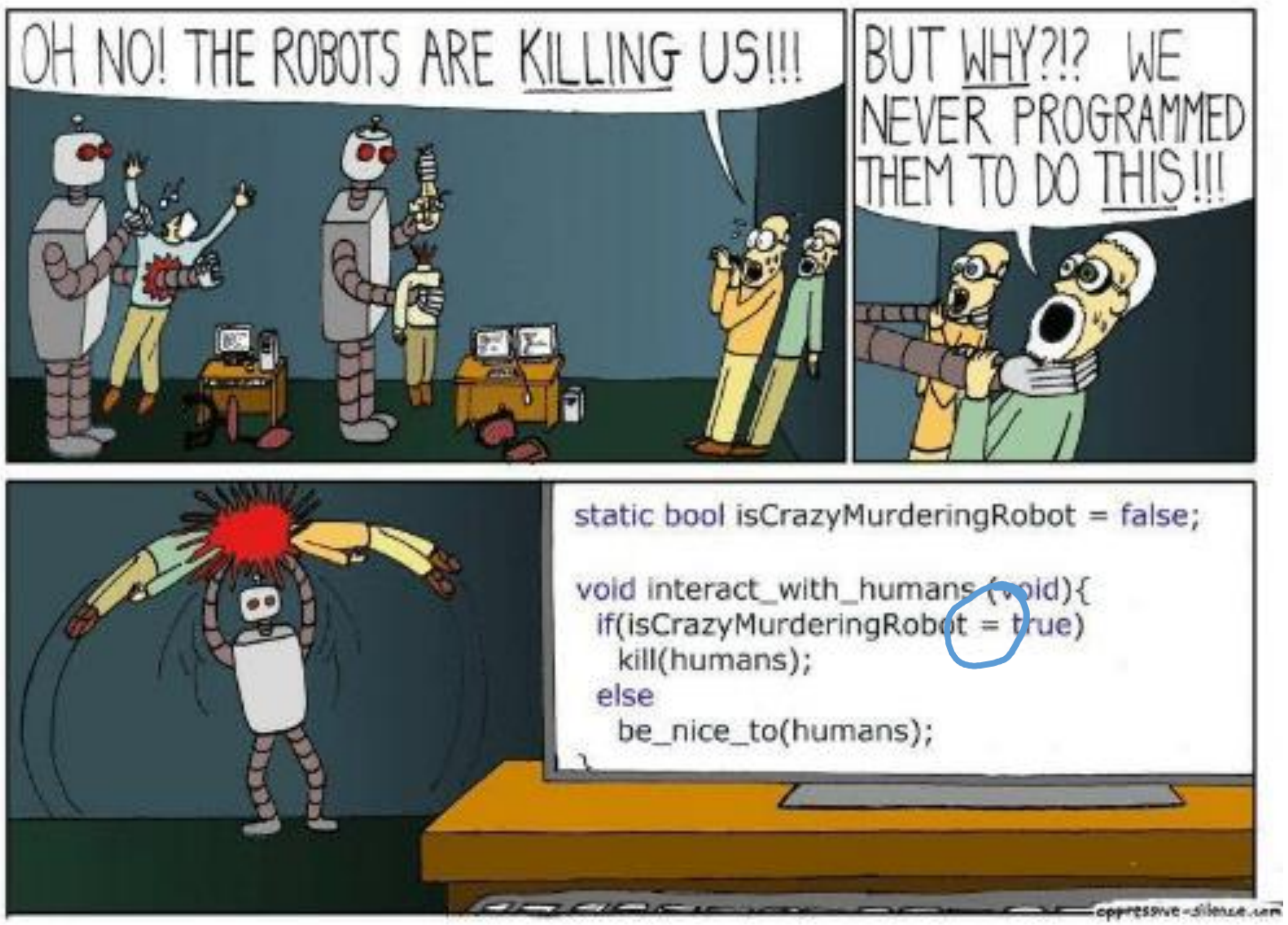
$X \sim U[0,1]$
$Y \sim U[0,1]$

$X \sim Y$
$X \sim 1-Y$

$=$: "always has the same value every time I run the code"

$Y = 1 - X$

$X + Y = 1$



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

$X \sim U[0,1]$

$Y \sim N(X, 0.1^2)$

"first generate x then use it to generate y"

```
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)
```

df.t is a vector of length 380

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

This expresses 380 separate equations.
Each of these eqns has an independent $N(0,\sigma^2)$.
(That's what the np.random.normal call generates)

Or, equivalently,

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

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

§1

A core skill is being able to design probability models. This course is for you to learn this skill, through examples.