

Computer Science

IB Data Science

Lecturer

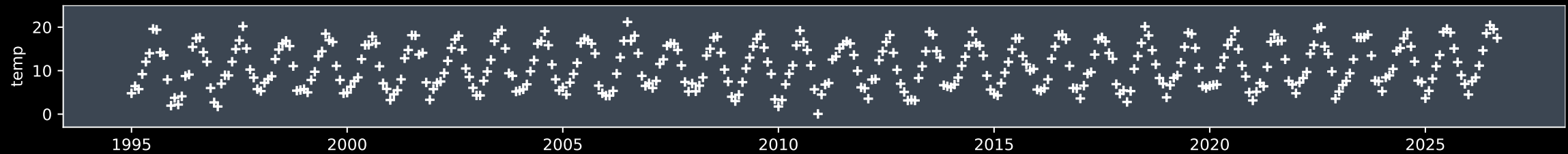
Dr Damon Wischik

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
⋮									

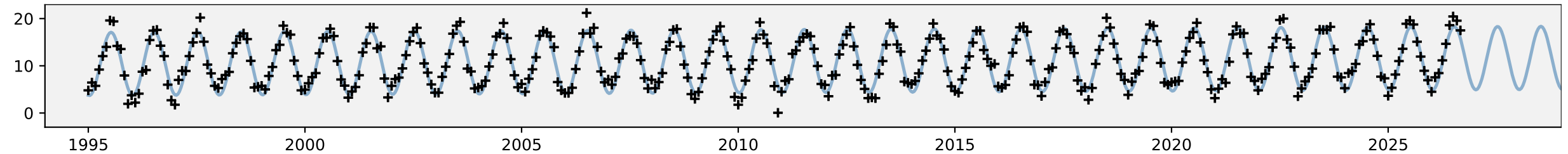


What model / formula would you suggest, to fit this dataset?

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

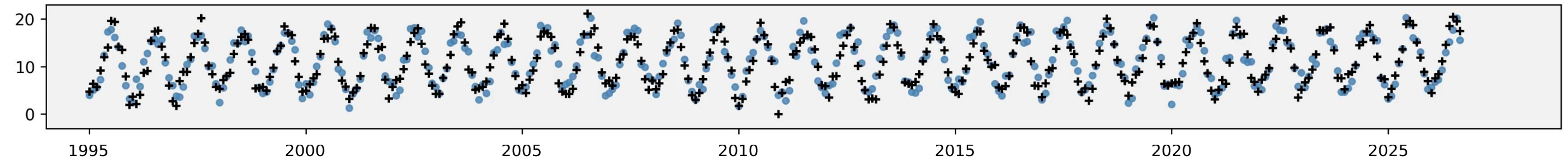
A PHYSICAL SCIENTIST'S DETERMINISTIC MODEL

```
def temp_model(t,  $\alpha$ ,  $\phi$ ,  $c$ ,  $\gamma$ ):  
    return  $c + \alpha * \sin(2*\pi*(t+\phi)) + \gamma*t$ 
```



A DATA SCIENTIST'S PROBABILITY MODEL

```
def rtemp(t,  $\alpha$ ,  $\phi$ ,  $c$ ,  $\gamma$ ,  $\sigma$ ):  
    pred =  $c + \alpha * \sin(2*\pi*(t+\phi)) + \gamma*t$   
    return random.normal(loc=pred, scale= $\sigma$ )
```



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

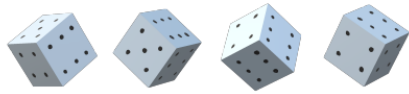
- A-level statistics formulae
- our climate model
- ChatGPT training

LECTURE NOTES

IB Data Science:

modelling and machine learning with probability

Damon Wischik, Computer Laboratory, Cambridge University



Contents

Introduction	v
Prerequisites	viii
I Learning with probability models	1
1 Specifying and fitting models	3
1.1 Specifying a probability model	3
1.2 Standard random variables	8
1.3 Maximum likelihood estimation	11
1.4 Numerical optimization with scipy	18
1.5 Likelihood notation	20
1.6 Types of model	23
1.7 Supervised and unsupervised learning	25
2 Feature spaces / linear regression	31
2.1 Fitting a linear model	32
2.2 Feature design	35
2.2.1 One-hot coding	35
2.2.2 Non-linear response	36
2.2.3 Comparing groups	36
2.2.4 Periodic patterns	37
2.2.5 Secular trend	38
2.3 Diagnosing a linear model	40
2.4 Linear regression and least squares	42

- The handout has more wordy explanations and more examples than lectures
- Ask teaching admin if you want a printout

§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 *

Department of Computer Science

cl.cam.ac.uk/teaching/2627/DataSci/materials.html

Course pages 2026–27

Data Science

[Syllabus](#) **Course materials** [Recordings](#) [Information for supervisors](#)

Lecture notes — [\[datasci2026.pdf\]](#). If you spot a mistake in these notes, please let me know.

Announcements, timetable — [Moodle](#)

Lecture schedule

This is the planned lecture schedule. It will be updated as and when actual lectures deviate from schedule. Material marked * is non-examinable. **Slides** are uploaded the night before a lecture, and re-uploaded after the lecture with annotations made during the lecture.

Prerequisites

[Example sheet 0](#) and [solutions](#)
[NumPy tutorial](#)

§1–§4. Learning with probability models

Lecture 1	1. Learning with probability models
[slides]	1.1 Specifying probability models
Lecture 2	1.2 Standard random variables
	1.3 Maximum likelihood estimation
	1.4 Numerical optimization with scipy
Lecture 3	1.5 Likelihood notation
	1.6 Types of model

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



Example sheet 0

Prerequisites
IB Data Science—DJW—2026/2027

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 it will help to 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:

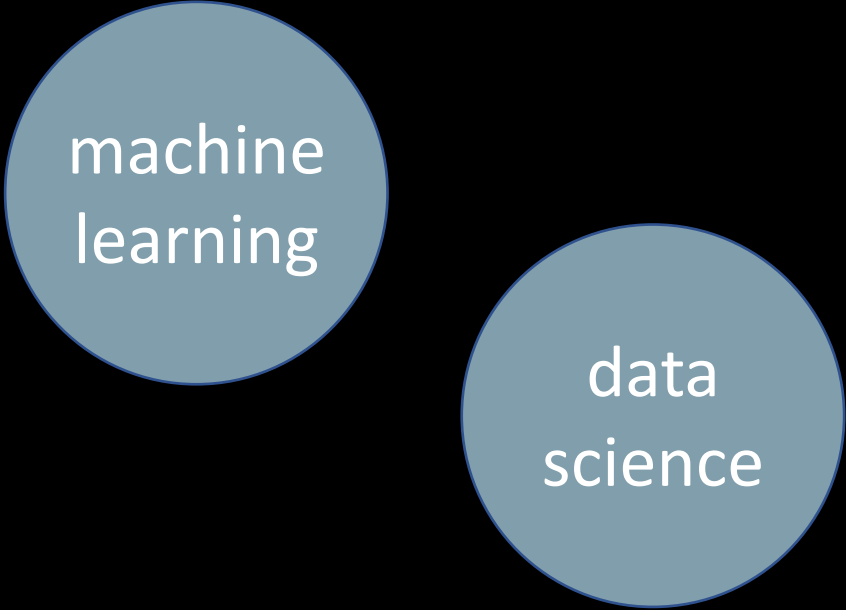
$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \text{ if } \mathbb{P}(B) > 0$$

$$\mathbb{P}(B, A) = \mathbb{P}(B) \mathbb{P}(A | B) \quad (\text{chain rule})$$

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

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

MACHINE LEARNING
Building systems
that learn
from data.

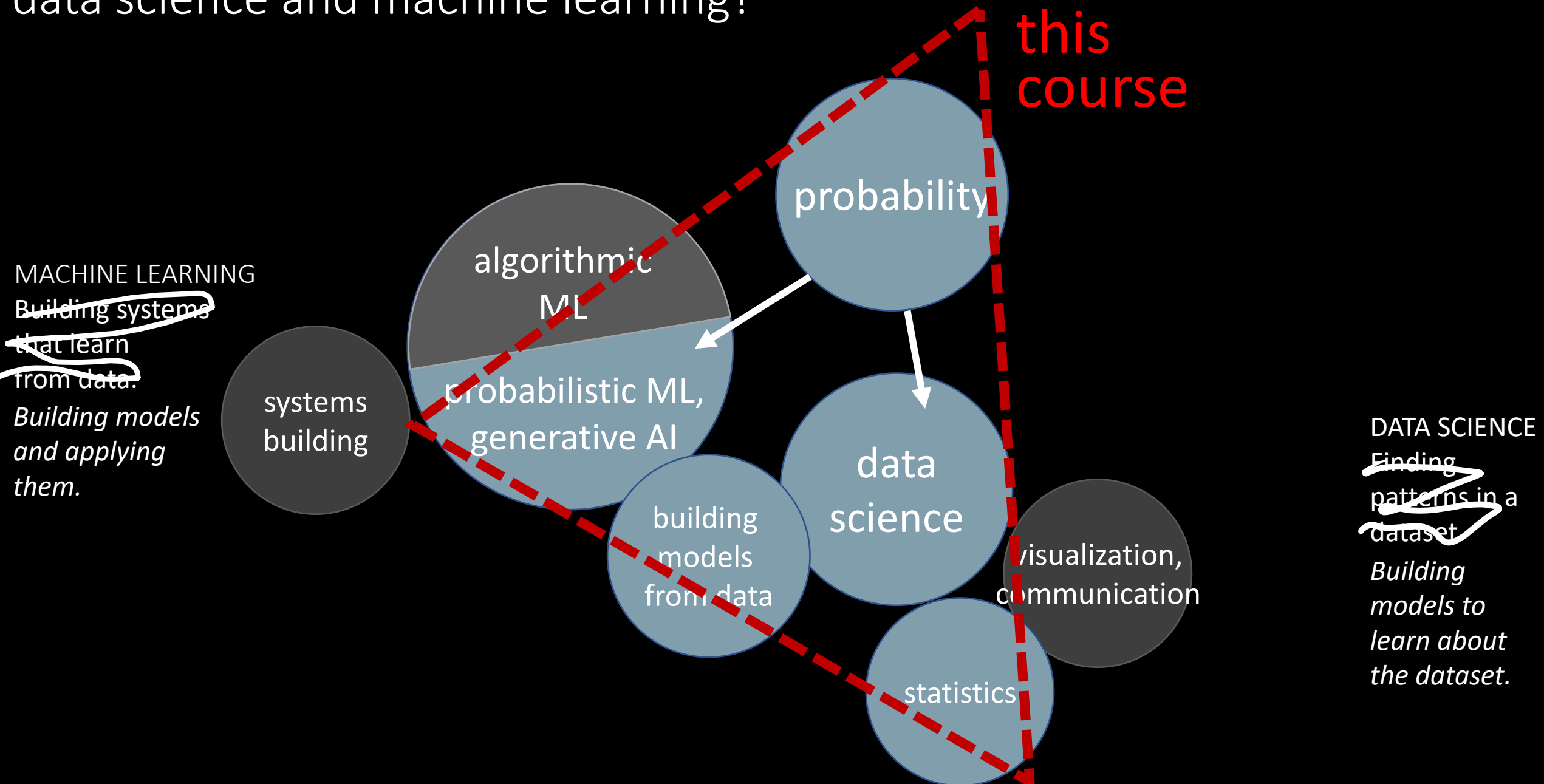


machine
learning

data
science

DATA SCIENCE
Finding
patterns in a
dataset.

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



In Data Science, we fit probability models.

Whereas in IA MLRD, you learnt to fit deterministic models:

1. Invent a predictor / classifier, which gives a deterministic answer
2. Fit it, to get high predictive accuracy

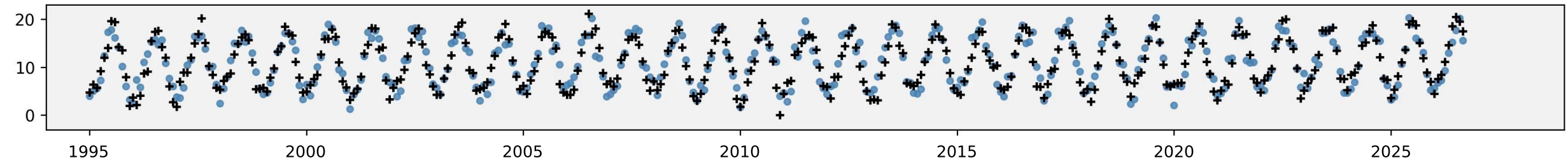
*If you don't get this elementary,
but mildly unnatural, mathematics
of elementary probability into your
repertoire, then you go through a
long life like a one-legged man in
an ass kicking contest.*

Charles Munger, business partner of Warren Buffett

What is a probability model?

ANSWER 1

Any piece of code whose output is random is a probability model.

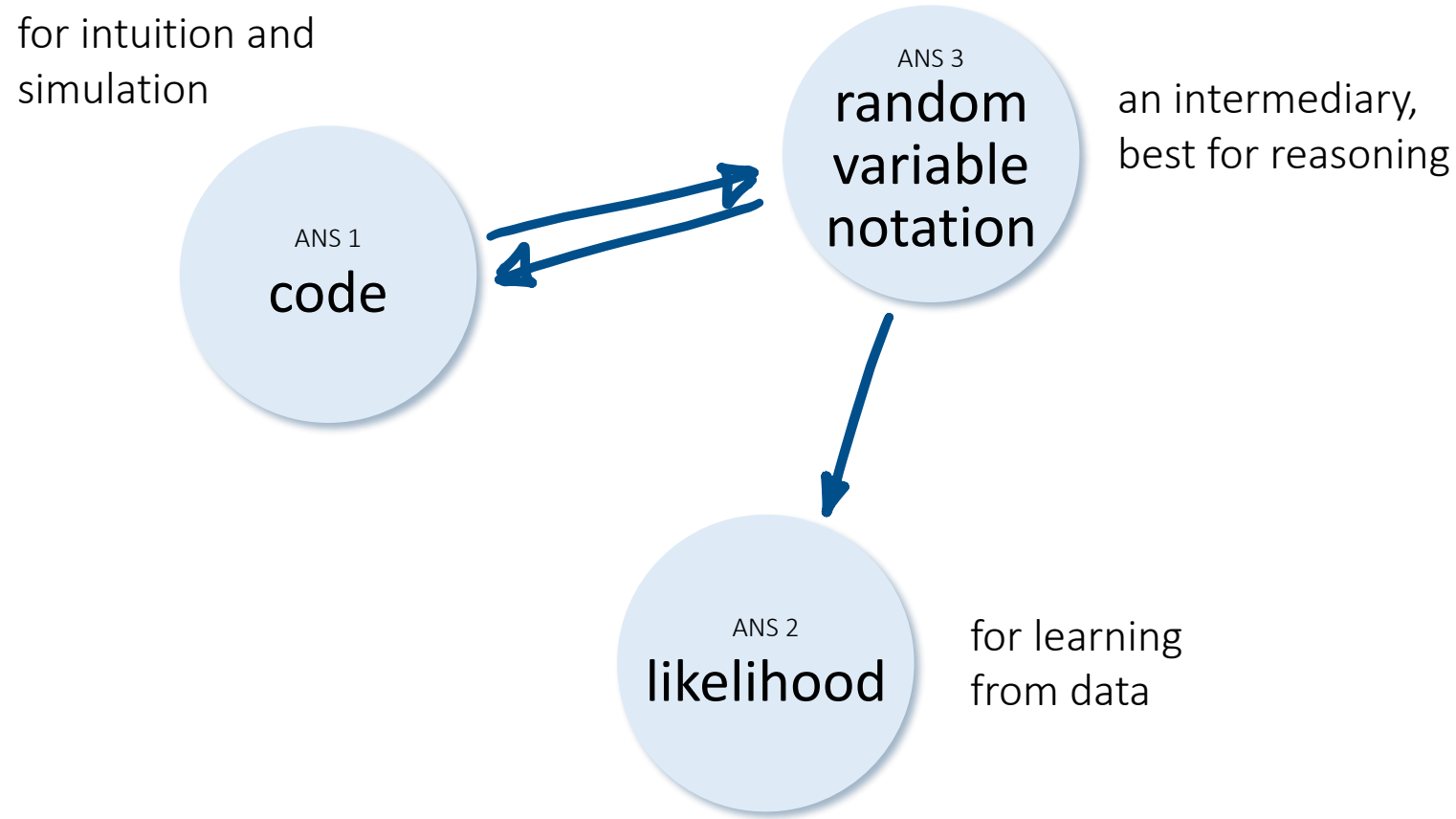


```
def rtemp(t,  $\alpha$ =10,  $\phi$ =-0.25, c=11,  $\gamma$ =0.035,  $\sigma$ =2):  
    pred = c +  $\alpha$  * np.sin(2*np.pi*(t+ $\phi$ )) +  $\gamma$ *t  
    return np.random.normal(loc=pred, scale= $\sigma$ )
```

This is a *vectorized* piece of code:
we can pass in a vector for t , and the computation
is automatically applied per element.

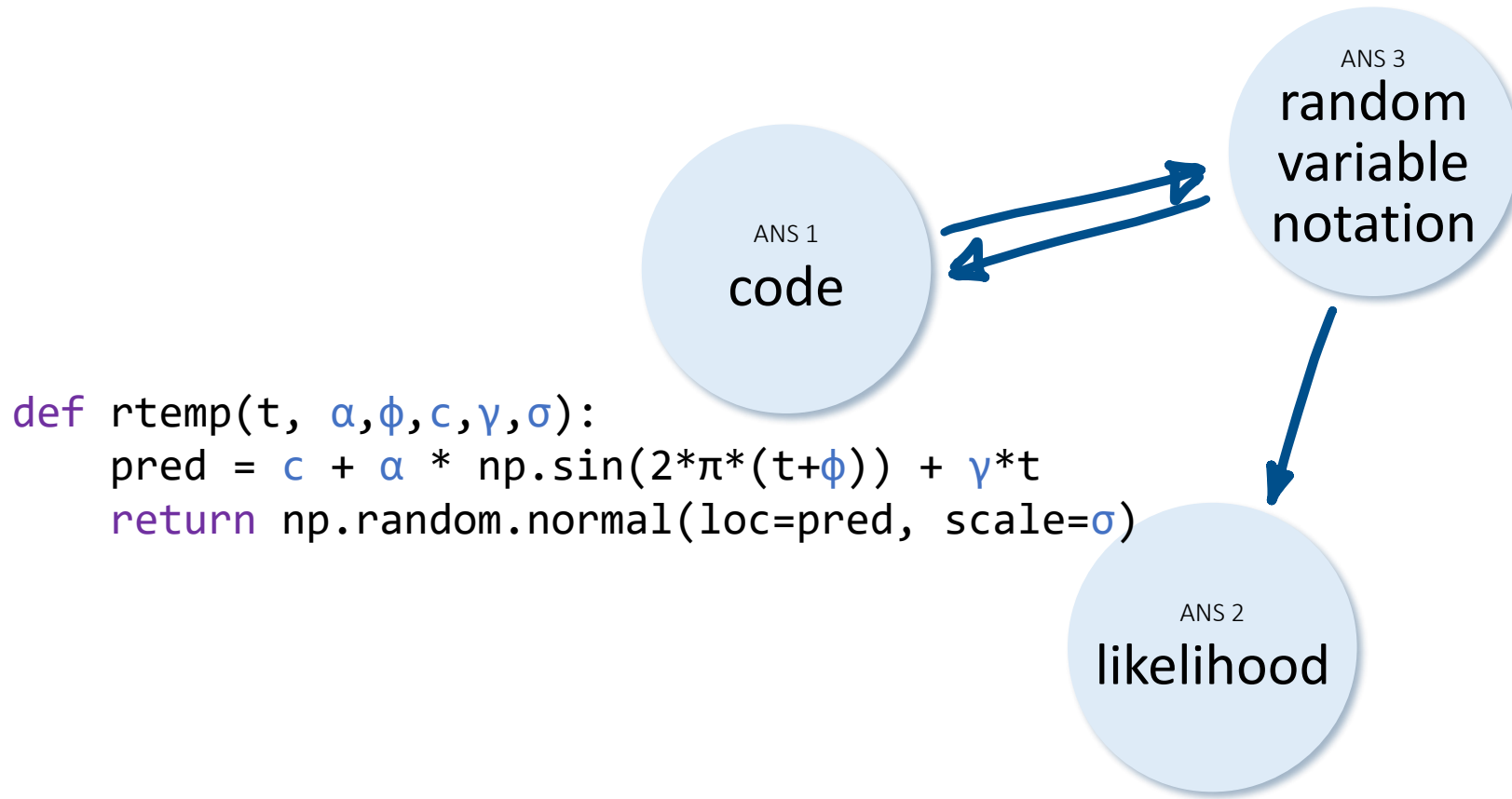
If you want to do further ML, learn the NumPy
library for vectorized computation. There's a tutorial
on the course webpage.

Three views of a probability model



Three views of a probability model

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



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

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

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

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

```
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)
```

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

```
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)
```

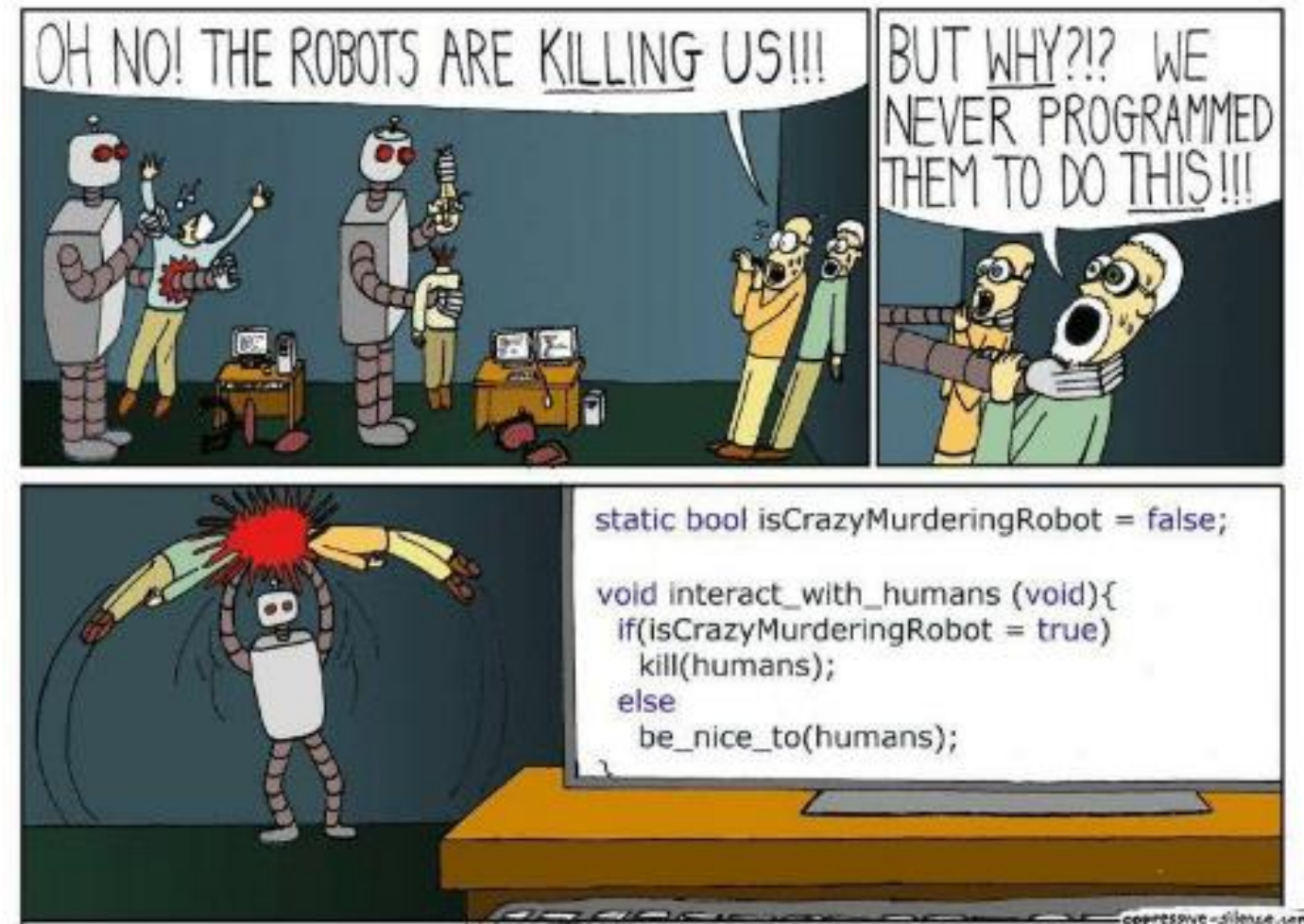
$$X_1, X_2 \sim U[0, \lambda]$$

QUESTION.
Are X_1 and X_2 independent?

When we say “ X_1 and X_2 are independent”, we implicitly mean “assuming the parameters are known”.

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

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



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

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

```
def rtemp(t,  $\alpha$ =10,  $\phi$ =-0.25,  $c$ =11,  $\gamma$ =0.035,  $\sigma$ =2):
    pred =  $\alpha$ *np.sin(2*np.pi*(t+ $\phi$ )) +  $c$  +  $\gamma$ *t
    return np.random.normal(loc=pred, scale= $\sigma$ )

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

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

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

QUESTION.

What are the random variables?

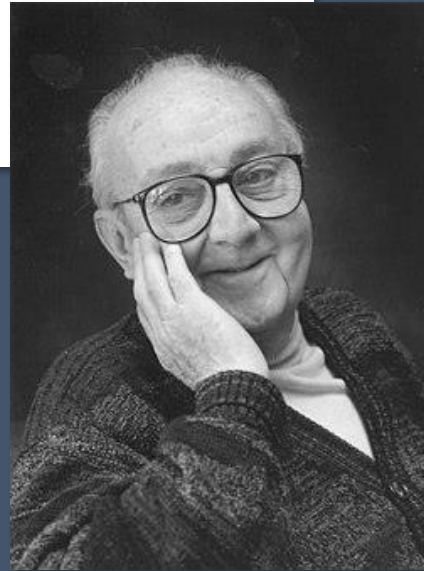
What are the parameters?

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



Next lecture: fitting models, i.e. “learning”

But we can only fit once we’ve got the right model. How do we get the right model?

*“All models are wrong
but some are useful”*

George Box

... so don’t be shy when building your own.
There *is* no “correct model”.

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