

Computer Science

# IB Data Science

Lecturer

**Dr Andrea Marinoni**

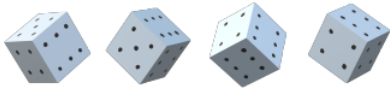
[am2920@cam.ac.uk](mailto:am2920@cam.ac.uk)

# HANDOUT

## Likelihood:

modelling and machine learning with probability

Damon Wischik, Computer Laboratory, Cambridge University



### Contents

Introduction	v
<b>I Learning with probability models</b>	<b>1</b>
<b>1 Specifying and fitting models</b>	<b>3</b>
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	17
1.5 Likelihood notation	19
1.6 Generative models / unsupervised learning	22
1.7 Supervised learning	25
<b>2 Feature spaces / linear regression</b>	<b>29</b>
2.1 Fitting a linear model	30
2.2 Feature design	33
2.2.1 One-hot coding	33
2.2.2 Non-linear response	34
2.2.3 Comparing groups	34
2.2.4 Periodic patterns	35
2.2.5 Secular trend	36
2.3 Diagnosing a linear model	38
2.4 Linear regression and least squares	40
2.5 The geometry of linear models	41
2.6 Interpreting parameters	44
2.7 Gauss's invention of least squares	49
<b>3 Neural networks</b>	<b>51</b>
3.1 Prediction accuracy	53
3.2 Probabilistic deep learning	56

- 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 I<sub>A</sub>, I<sub>B</sub> 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	MK	ConcDisSys		
	QM	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	--		ML-B	DenotSem	

10:00-11:00

N	C	M--M	M M--M	M M--M	M M--M	M M--M	M M--M	M --	--	--	--	--	AVSM	FoundCS
N	C	--	--	--	--	--	--	F--F	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--	J J--	J J--	J J--	J J--	RM+	IntComArch
P	L	O --	O --	O --	O --	O --	O --	I --	I I --	I C --	C B --	B	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	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--L	L L--	--	--	NDL+	ADS

11:00-12:00

N	C	W--W	W W	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	00Prog
N	S	--	G	--	G	--	G	--	G	--	--	--	--	DJG	Databases
N	S	--	--	--	--	--	--	Z--Z	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--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 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 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 \*

Home / Teaching / Courses 2024–25 / Data Science / Course materials

Department of Computer Science and Technology

Course pages 2024–25

Computer Laboratory

Teaching

Courses 2024–25

Part IB CST

Data Science

Concurrent and Distributed Systems

ECAD and Architecture Practical Classes

Economics, Law and Ethics

Further Graphics

Further Java

Introduction to Computer Architecture

Programming in C and C++

Semantics of Programming Languages

Unix Tools

Compiler Construction

Syllabus

Course materials

Recordings

Information for supervisors

Lecture notes

- Abridged notes as printed — examinable material only
- Extended notes with extra material on non-examinable material such as...

If you spot a mistake in the printed notes, let me know.

Announcements and Q&A

- Moodle

Lecture schedule

This is the planned lecture schedule. It will be updated as and when actual dates are confirmed. **Slides** are uploaded the night before a lecture, and re-uploaded if necessary.

Pre-recorded versions of each lecture are available on [last year's version](#) of the course page.

Prerequisites

Example sheet 0 and solutions

§1–64. Learning with probability models

Lecture 1

- 1. Learning with probability models
- 1.1 Specifying probability models

Lecture 2

- 1.2 Standard random variables
- 1.3 Maximum likelihood estimation
- 1.4 Numerical optimization

Lecture 3

- 1.5 Likelihood notation
- 1.6 Generative models

IB Data Science (2021-22) - YouTube x

youtube.com/playlist?list=PLknxdt7zG11MLG2w-l0...

YouTube GB

Search

PLAY ALL

IB Data Science (2021-22)

5 videos • No views • Updated today

Public

Cambridge University computer science: undergraduate introduction to data science and machine learning

SORT

1. Learning with probability models  
Foundations of Data Science  
4:08

1.1 Specifying probability models  
Foundations of Data Science  
15:20

1.2 Standard random variables  
Foundations of Data Science  
3:21

1.3 Maximum likelihood estimation  
Foundations of Data Science  
17:35

- 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>

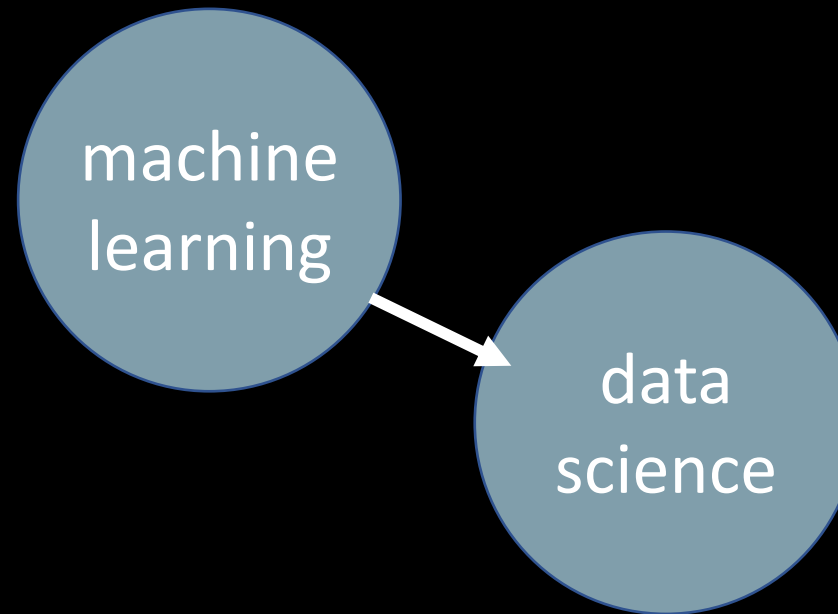
For any teaching session where your contribution is mandatory or expected, we must seek your consent to be recorded.

**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?

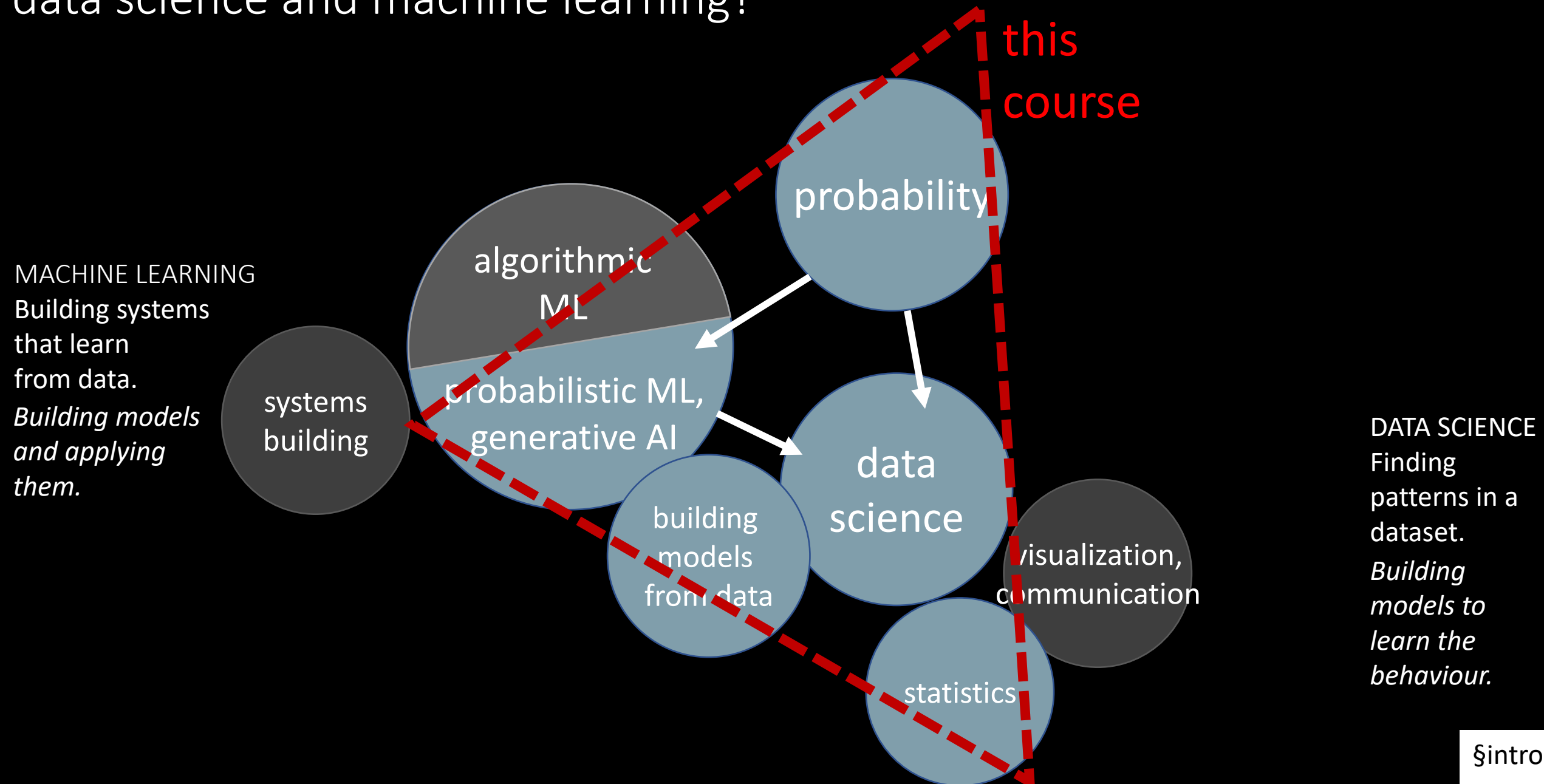
# What is data science? What's the difference between data science and machine learning?

**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?





*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  $\Omega$ , 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:

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

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

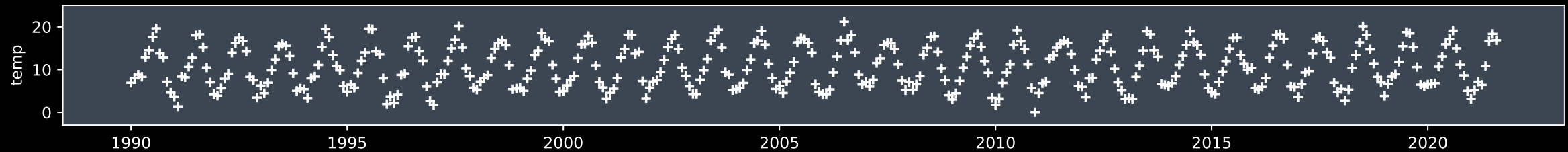
- Example sheet 0 is to remind you about IA Probability, Maths for NST, and Scientific Computing
- It's not for supervision; solutions are provided

# 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
⋮									

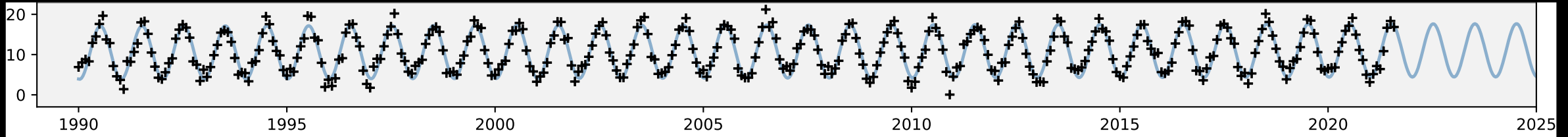


What model / formula would you suggest to fit this dataset?

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

## A SCIENTIST'S DETERMINISTIC MODEL

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



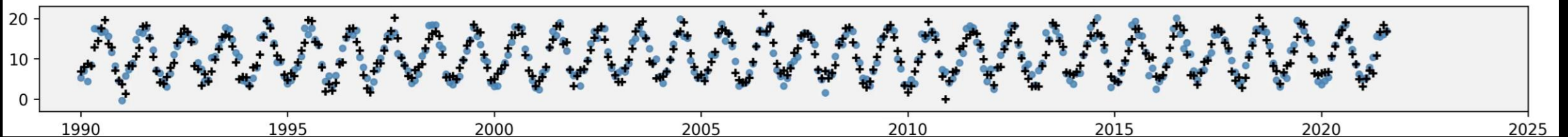
why? To describe the data in front of me!

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



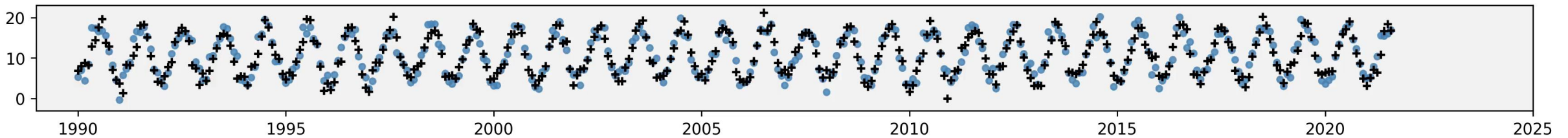
All of machine learning is based on a single idea:

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

This is behind

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

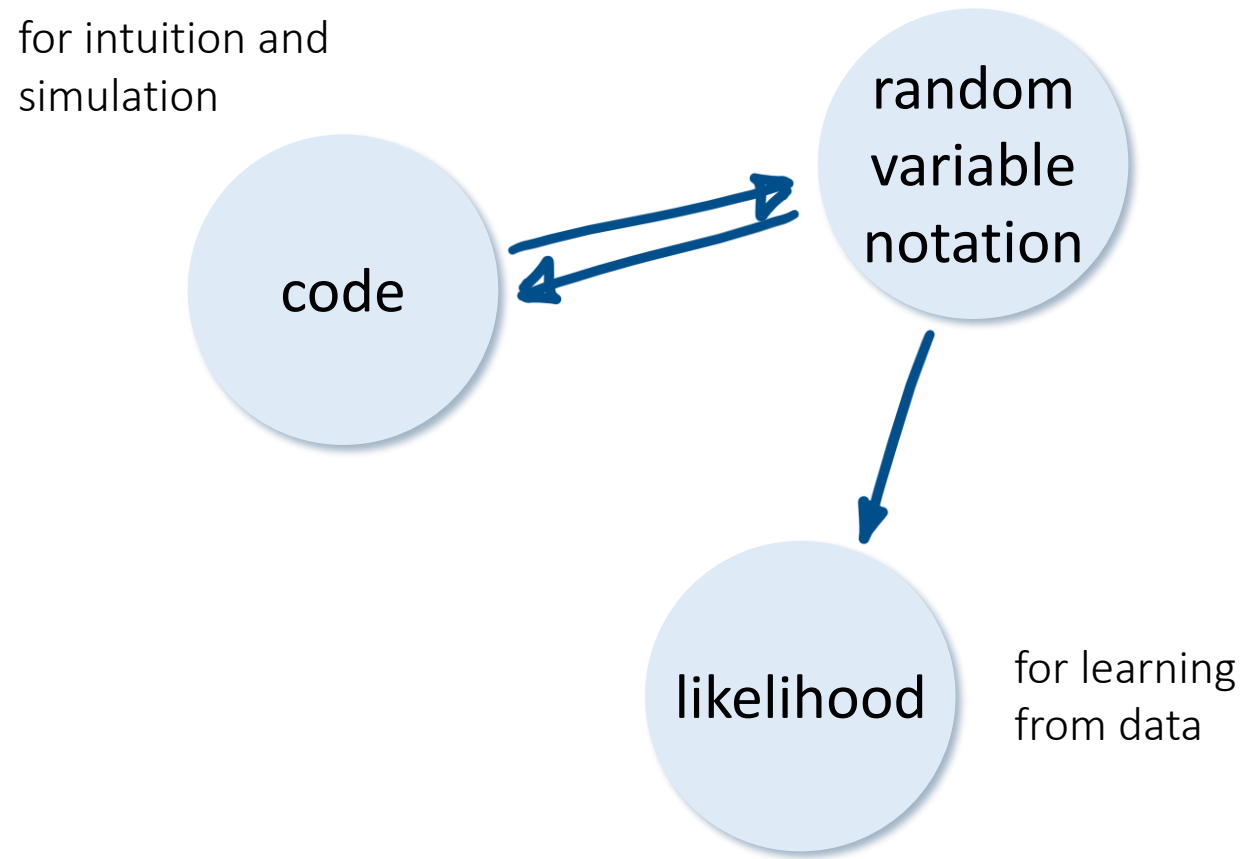
# How to specify a probability model



```
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= $\sigma$ )
```

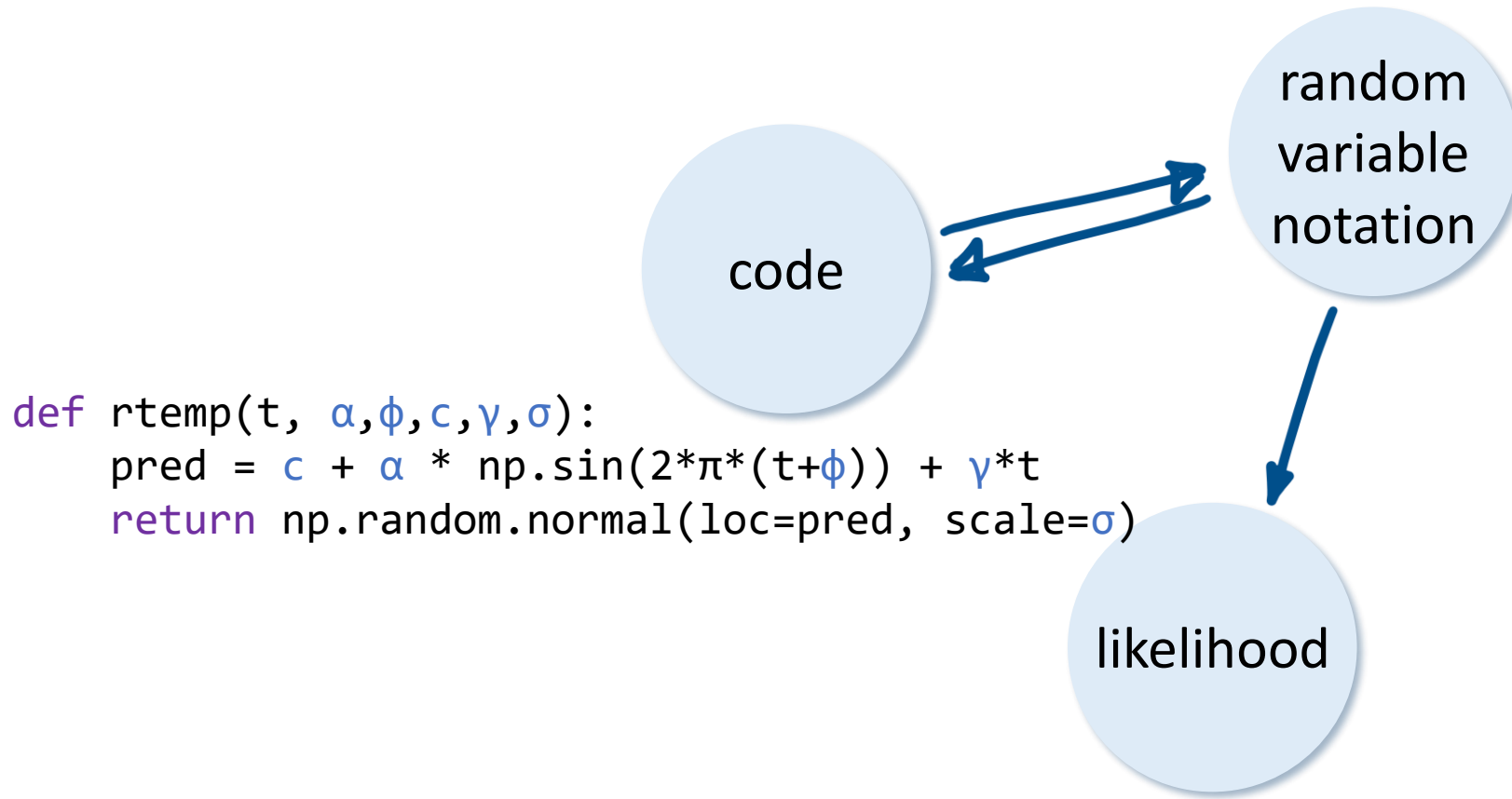
A probability model is a piece of code  
where the output is random.

# Three views of a probability model



# Three views of a probability model

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





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

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

Generate  $X$  from the  
Uniform distribution.

```
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 variable  
Lower case: parameters, constants.

```
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"  
— knowing the value of one tells us  
nothing about the value of the other.

In random variable notation,  
assume independence

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

$(Y, Z) \sim \text{Myrandpair}$  unless specified otherwise  
(like this)



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

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

$\lambda$  is lower-case, so it refers to a fixed value,

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
```

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

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,

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

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.

