Lecture 1: Introduction to Machine Learning
Palestinian Advanced Learning with Machines School, 2022

Daniel Worrall

Intention Setting and Overview

SCIENTISTS FOR PALESTINE

Today's intention: "Basics crash course"

I will show you a lot today.

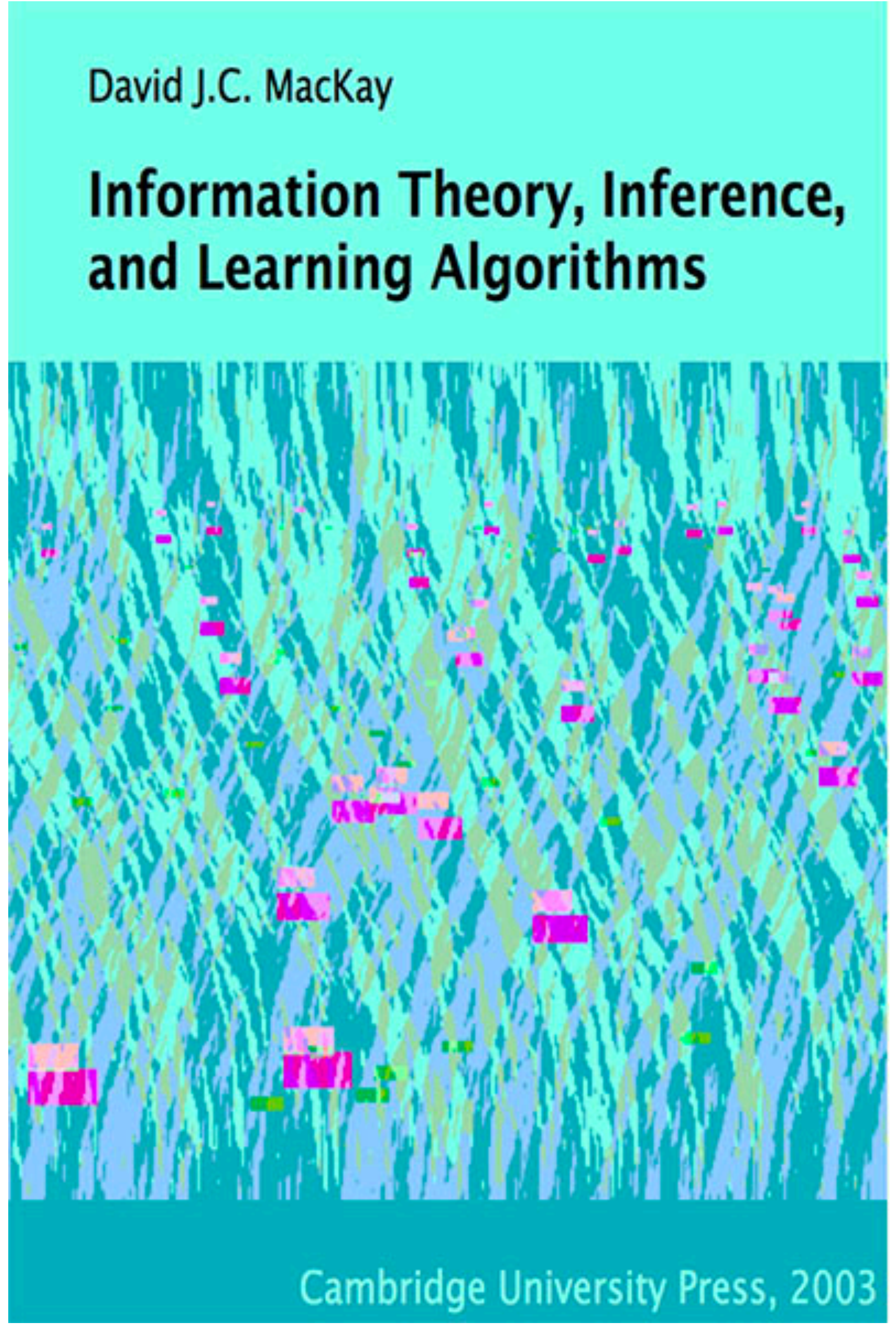
If you cannot follow everything, do not worry!

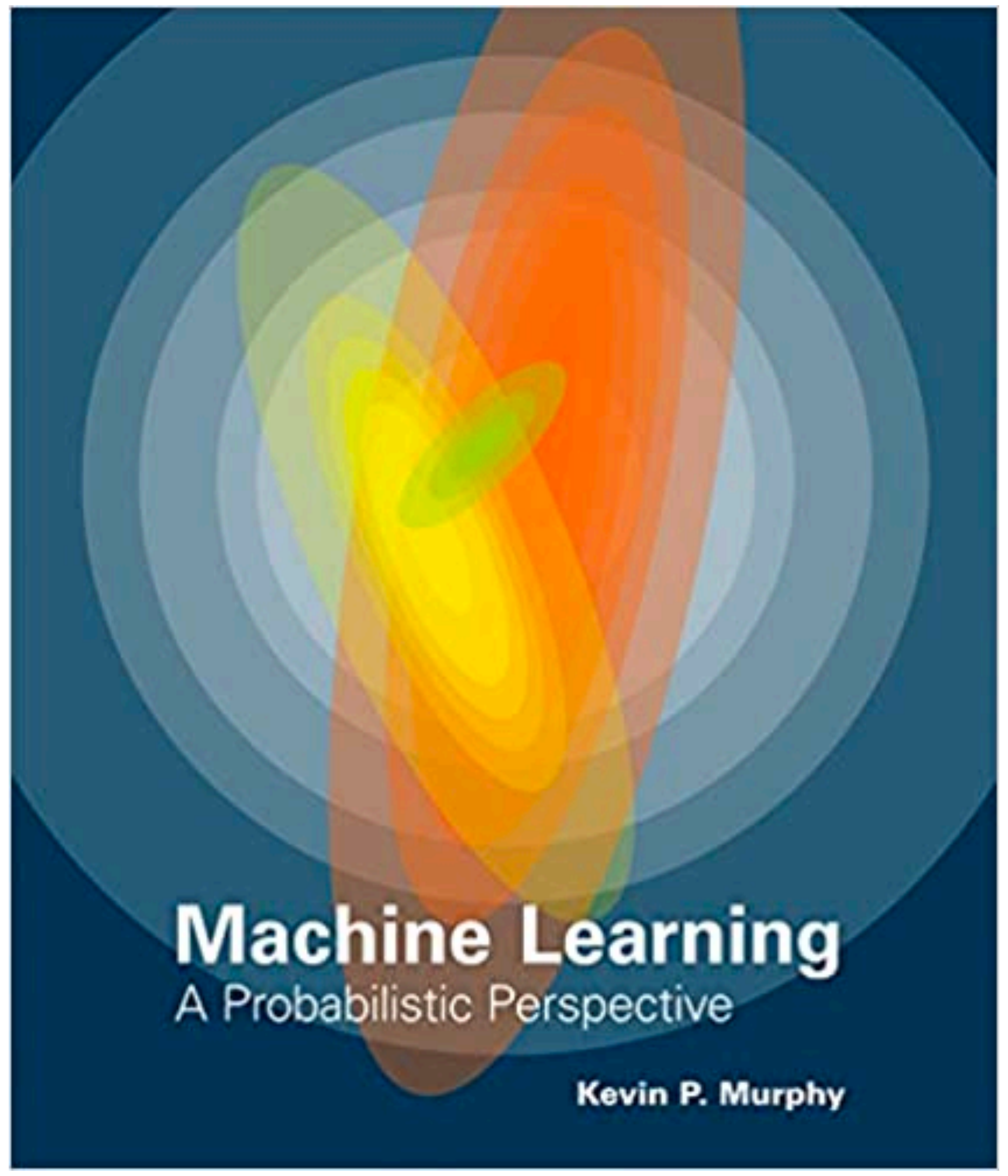
Our goal is exposure, not mastery :)

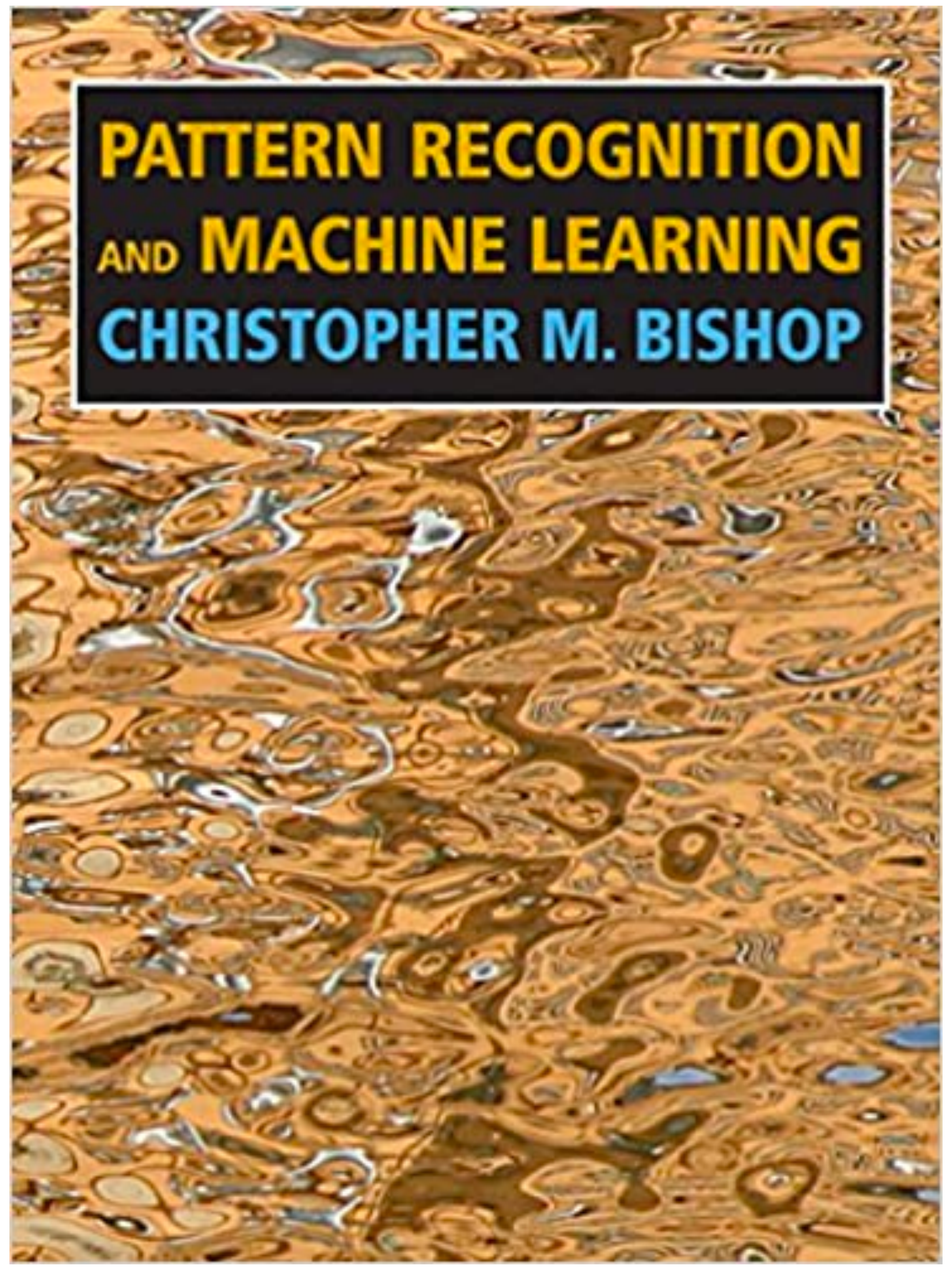
This lecture: Machine Learning Basics

- What is Machine Learning?
- Probability Theory
- Probabilistic models
  - Forward models
  - Independence
- Statistical Inference
  - Maximum Likelihood
  - Bayesian Inference
- Modeling paradigms
  - Prediction
- Model comparison

My recommended books







What is Machine Learning?

SCIENTISTS FOR PALESTINE

"Machine learning is the latest in a long line of attempts to distill human knowledge and reasoning into a form that is suitable for constructing machines and engineering automated systems."

—*Mathematics for machine learning*

Machine Learning

Supervised Learning

Reinforcement Learning

Socio-technical challenges

Computer science

Unsupervised Learning

Generative models

Information theory

Self-supervised Learning

Deep learning

AGI safety

Computer vision

Optimisation

Statistics

Computational neuroscience

Natural Language Processing

Causality

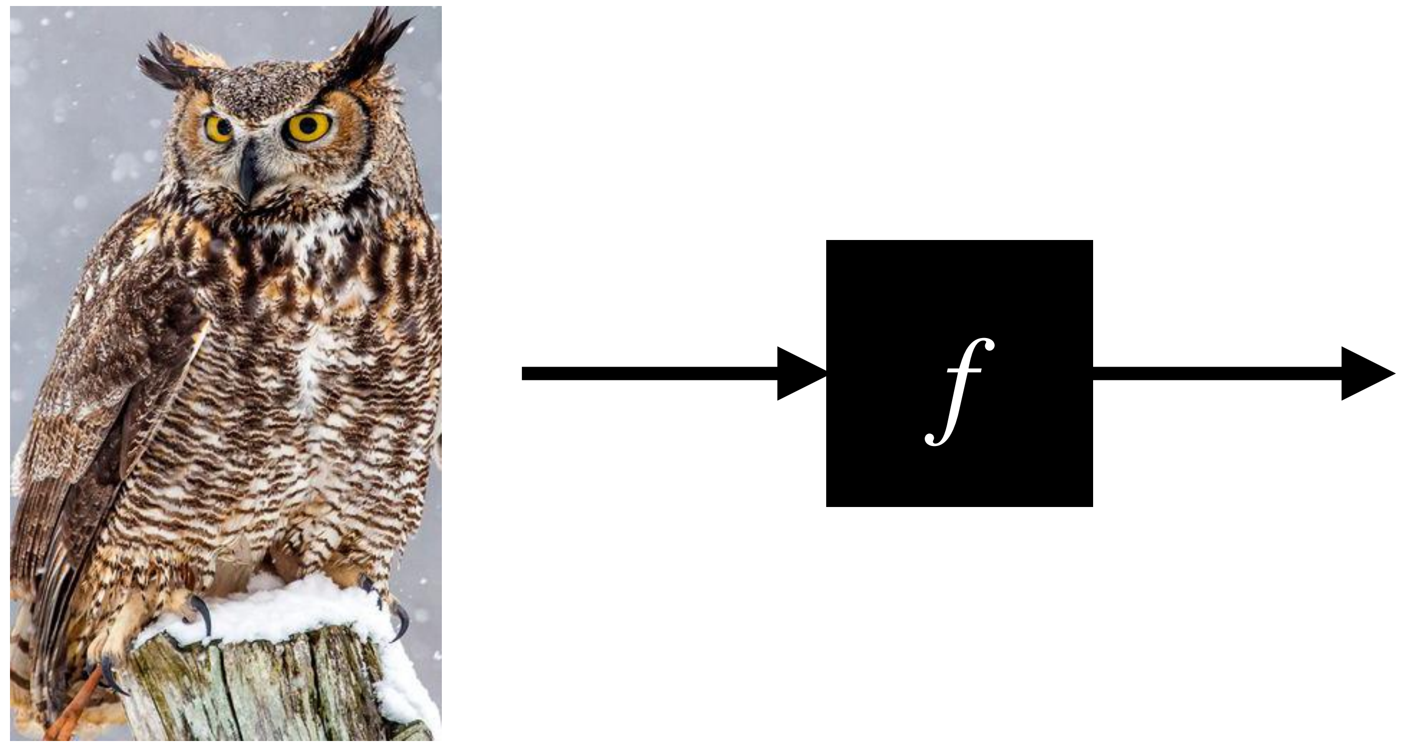
Supervised learning

In these lectures, we focus on *supervised learning*

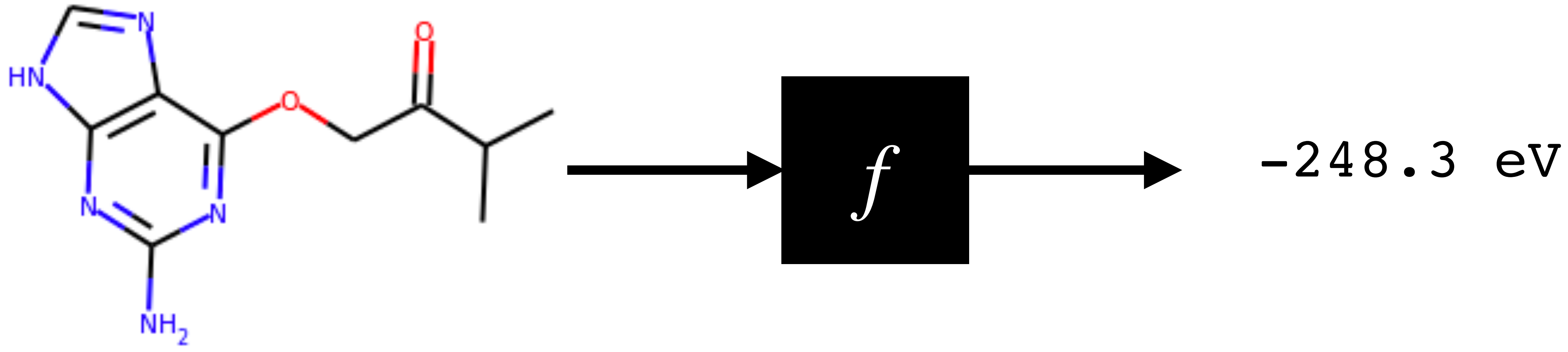
Supervised learning problems have 3 parts

- **Data:** inputs $\mathbf{x}$ and outputs $\mathbf{y}$
- **Model space:** a collection of *models* $\mathscr{M}$ which convert inputs into outputs
- **Algorithm:** a method to choose the best model $m \in \mathscr{M}$ from the model space[1], which best *fits* the data

Classification e.g. image recognition



Regression e.g. molecular property prediction

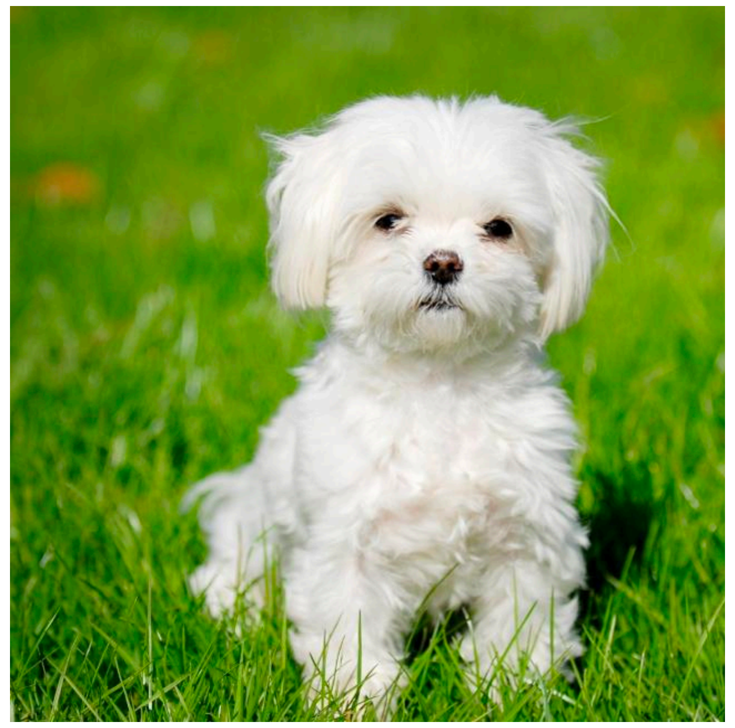


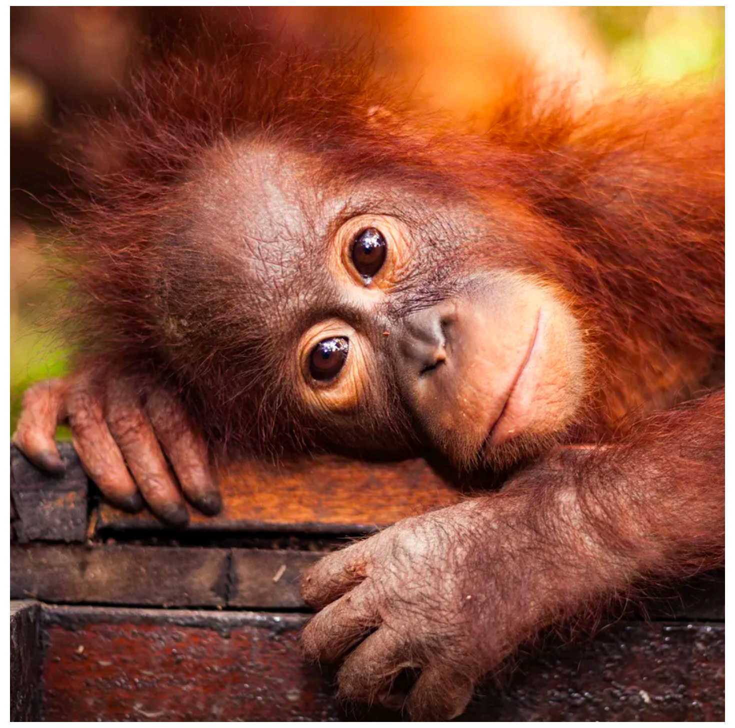
[1] The notation '∈' is pronounced *in*, so $m \in \mathscr{M}$ is spoken 'm in M'

Example I - Image Classification

SCIENTISTS FOR PALESTINE

Image classification





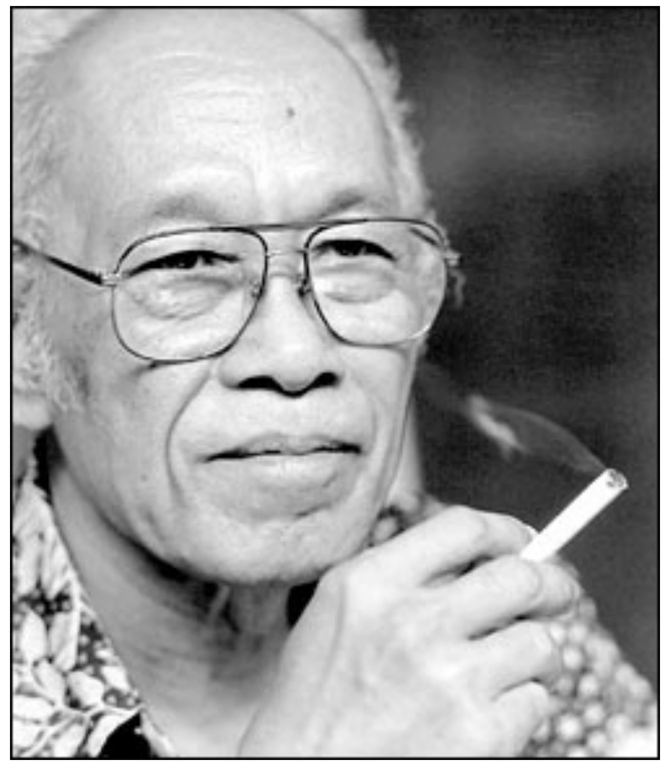
- **Inputs**: 256x256 pixel RGB images
- **Outputs**: labels in {‘dog’, ‘cat’, ‘orangutan’}

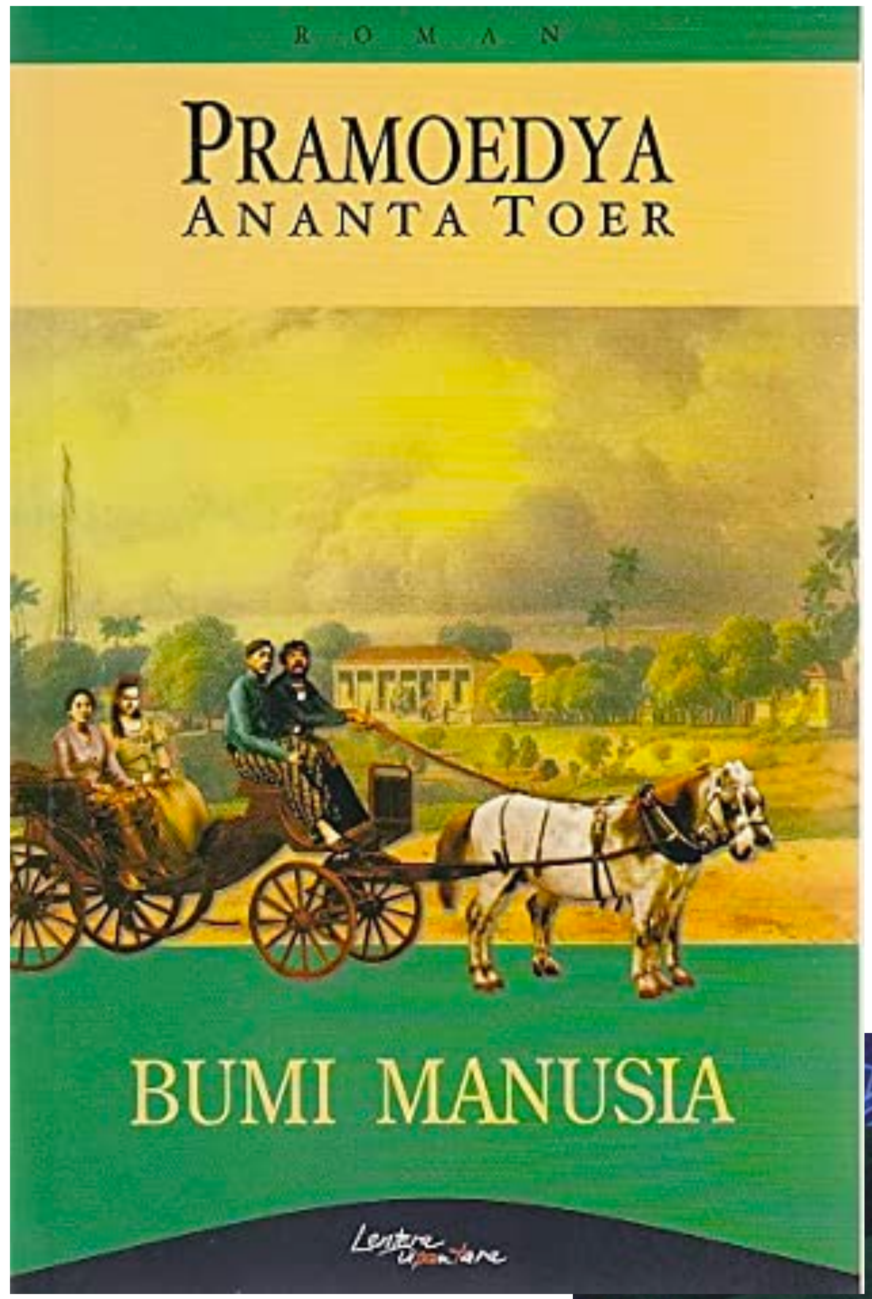
- **Model**: ?
    - **Suggestion:** Collect examples of images $\{x_1, x_2, ...\}$ with labels $\{y_1, y_2, ...\}$ and build a *lookup table*. If new input $\mathbf{x}_* = \mathbf{x}_j$, where $\mathbf{x}_j \in \{x_1, x_2, ...\}$ then its *predicted label* is $\mathbf{y}_* = \mathbf{y}_j$.
    - **Problem 1**: What if we have never seen $\mathbf{x}_*$ before?
    - **Problem 2:** What if we feed in something which is not a dog/cat/orangutan?
    - **Problem 3**: Are there ways to quantify how good/bad our model is?

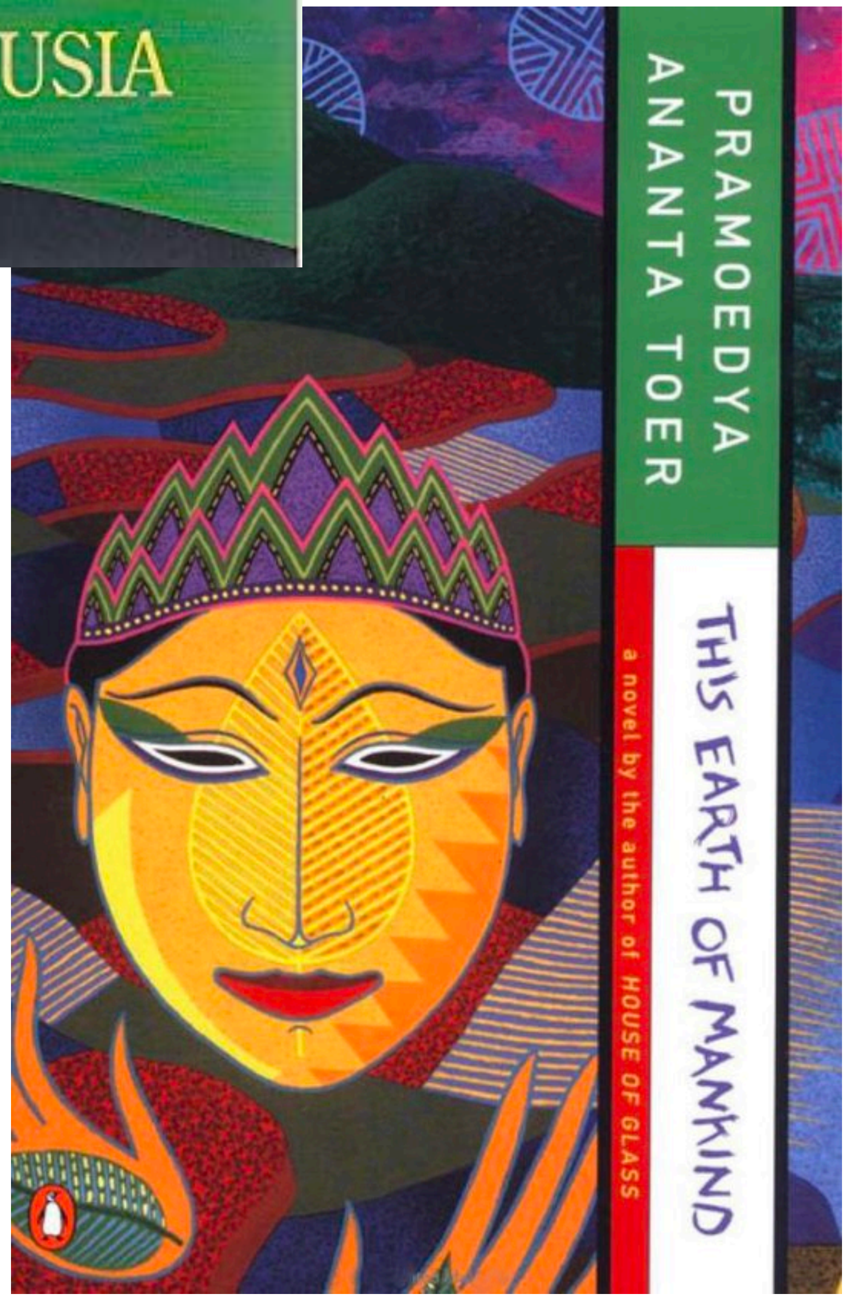
Example II - Machine Translation

‘Orang memanggil aku: Minke. Namaku sendiri ... sementara ini tak perlu kusebutkan. Bukan karena gila misteri. Telah aku timbang: belum perlu benar tampilkan diri dihadapan mata orang lain.’

—Pramoedya Ananta Toer (~1980).







‘People called me Minke. My own name … for the time being I need not tell it. Not because I’m crazy for mystery. I’ve thought about it quite a lot: I don’t yet really need to reveal who I am before the eyes of others.’

—Max Lane (1981)

‘People called me Minke. As for my real name … for now it doesn’t need to be mentioned. Not because I need mystery. I have weighed it up: I needn’t yet reveal myself before the eyes of others. ’

—Daniel Worrall (2020)

‘People call me: Minke. My own name … meanwhile I don't need to mention it. Not because of a mystery mad. I have weighed: do not need to properly present yourself before the eyes of others.’

—Google Translate (2020)

Example II - Machine Translation

SCIENTISTS FOR PALESTINE

Machine translation is an input—output task

- **Inputs**: variable length Indonesian strings
- **Outputs**: variable length English strings

- **Problem 1**: Each word has multiple translations
- **Problem 2**: We cannot possibly collect all input—output pairs
- **Problem 3**: What is a good translation?

Many of the problems we have seen can be addressed (to some extent) by thinking *probabilistically*. To understand what this means though, we need to learn what probability is.

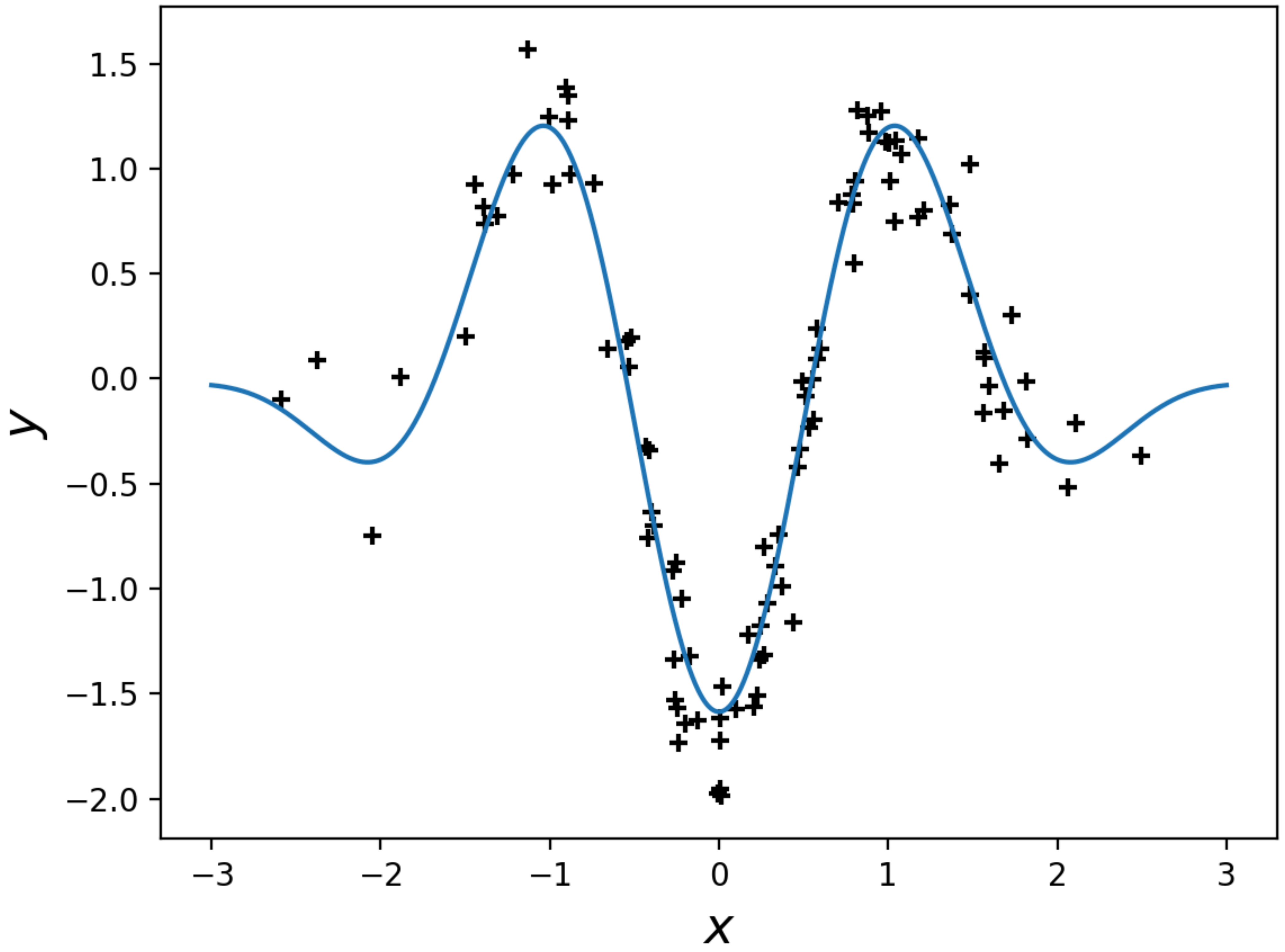
Example III - Regression

Regression is the canonical input—output task

Inputs: real numbers $x_i \in \mathbb{R}$

Outputs: real numbers $y_i \in \mathbb{R}$

- **Problem 1**: How to handle residual error?
- **Problem 2**: Is a linear model the best we can do?
- **Problem 3**: What about higher dimensions?



The previous two examples are just variants to of regression (in a very liberal sense).

Data → Machines

Machine learning is primarily a conceptual discipline.

Machine learning is automated machine building*

Machine learning is…

Computational machines need:

- Computers
- Programming
- Mathematics: Probability, calculus, linear algebra

Mathematics is the language of data

Mathematics is just one aspect of ML

- Socio-technical aspects: Fairness, Decoloniality, …

* Some would argue that it is the automation of the scientific method.

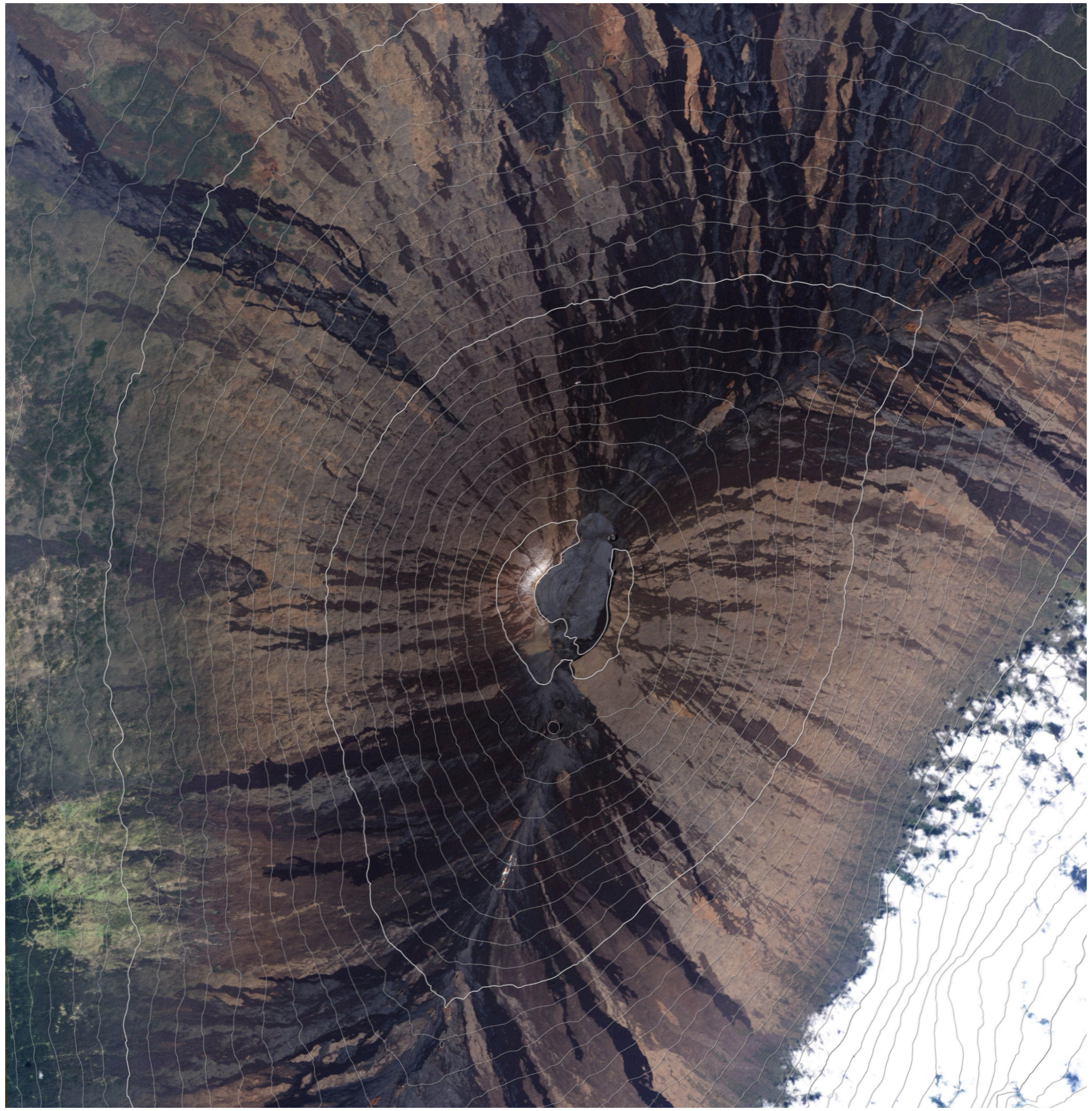
Probability and Statistics

SCIENTISTS FOR PALESTINE

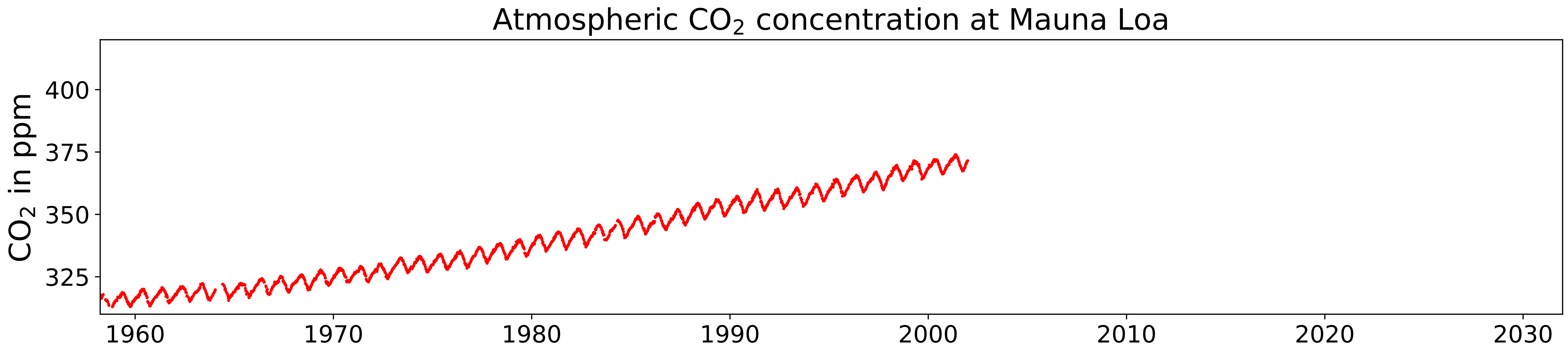
Mauna Loa

SCIENTISTS FOR PALESTINE

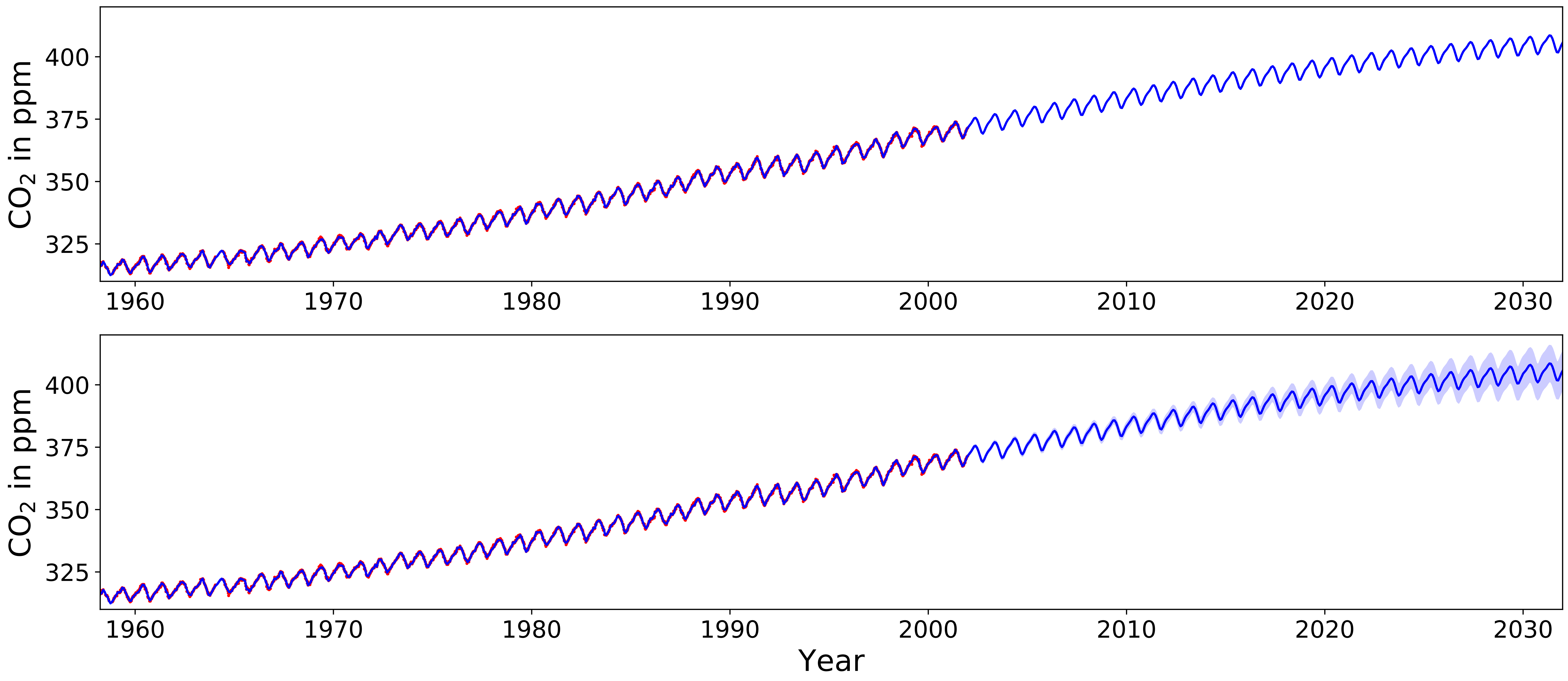
Mauna Loa is one of 5 volcanoes forming the Island of Hawaii in the U.S. state of Hawaii in the Pacific Ocean. The largest subaerial volcano in both mass and volume, Mauna Loa, has historically been considered the largest volcano on Earth, dwarfed only by Tamu Massif.



Mauna Loa



In machine learning we use past data to make predictions about the future



Mauna Loa

How would we model the Mauna Loa CO_2 levels as a function of time? Perhaps

$$y = w_1 x + w_2$$

Well this is a straight line, we need an extra periodic component, so how about

$$y = w_1 x + w_2 \cos(2\pi x + w_3) + w_4$$

But this is no good, because y is negative for certain values of x...

The reality of modeling

"All models are wrong, but some are useful."

—George Box, 1976

- How do we choose a model?
- Does it matter that we will no doubt make mistakes?
- How do I tell if one model is better than another?

SCIENTISTS FOR PALESTINE

probability: **a mathematical formalism describing uncertain events**

statistics: **the science of collecting and analyzing data**

—Carl Rasmussen

Bayesian statistics is a branch of statistics loved by machine learners for its computational nature.

- Why is it useful?
- What can we say about uncertain events?
- What be measured?

Intuitive Probability

SCIENTISTS FOR PALESTINE

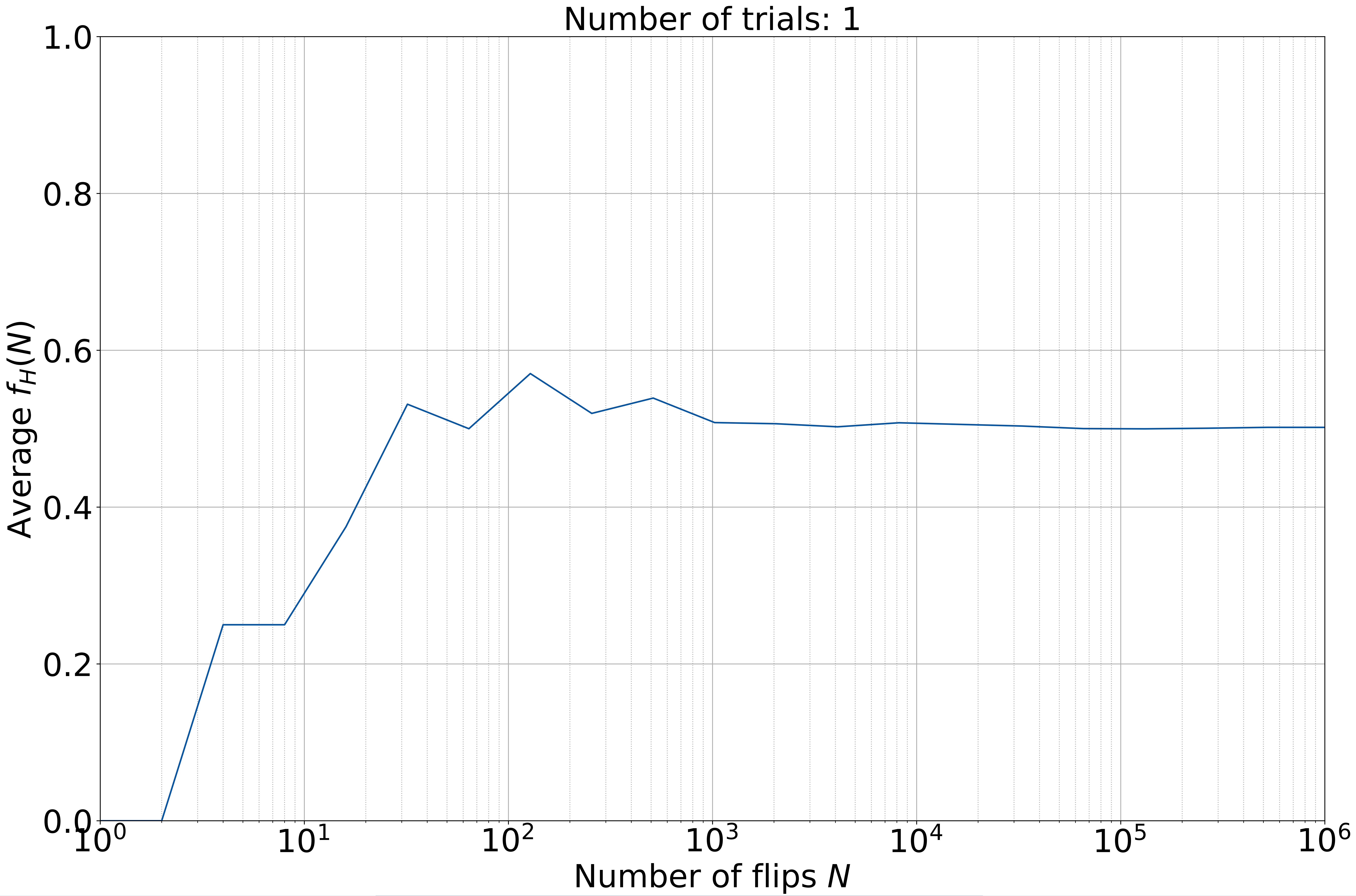
Take a coin. Label heads with 1 and tails with 0. Now flip the coin N times and take the average. Now do this again multiple times.

	Trial 1	Trial 2	Trial3	Trial4	Trial 5
$N = 10$	0.5000	0.8000	0.6000	0.6000	0.2000
$N = 100$	0.4800	0.4800	0.4800	0.5400	0.5400
$N = 1000$	0.4950	0.5130	0.5080	0.5080	0.4850
$N = 10000$	0.4967	0.5031	0.4980	0.4988	0.4934

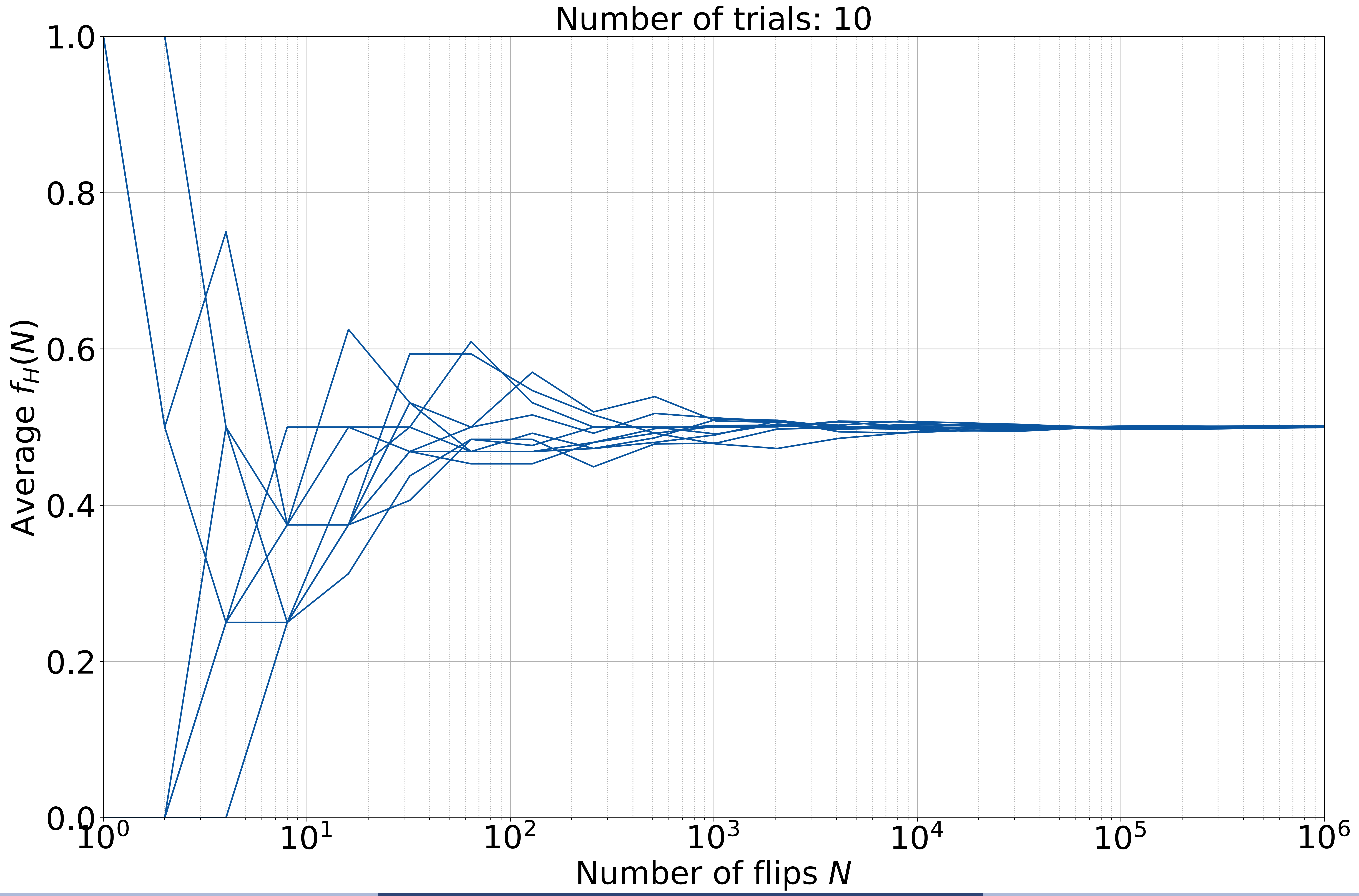
Despite the fact that in each trial we get a different result, there is a trend!

As $N \to \infty$, what do you think will happen?

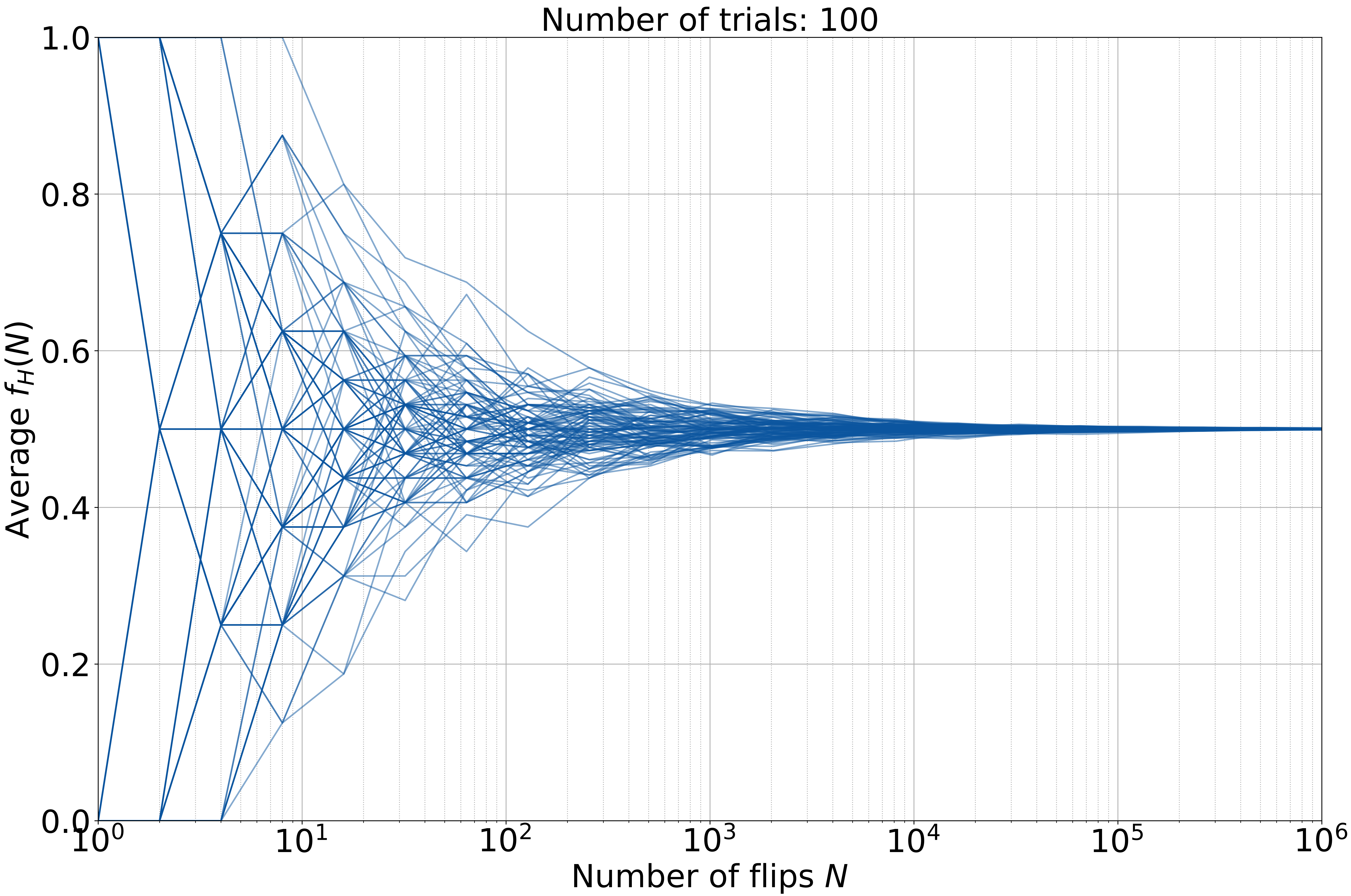
Frequentist Probability



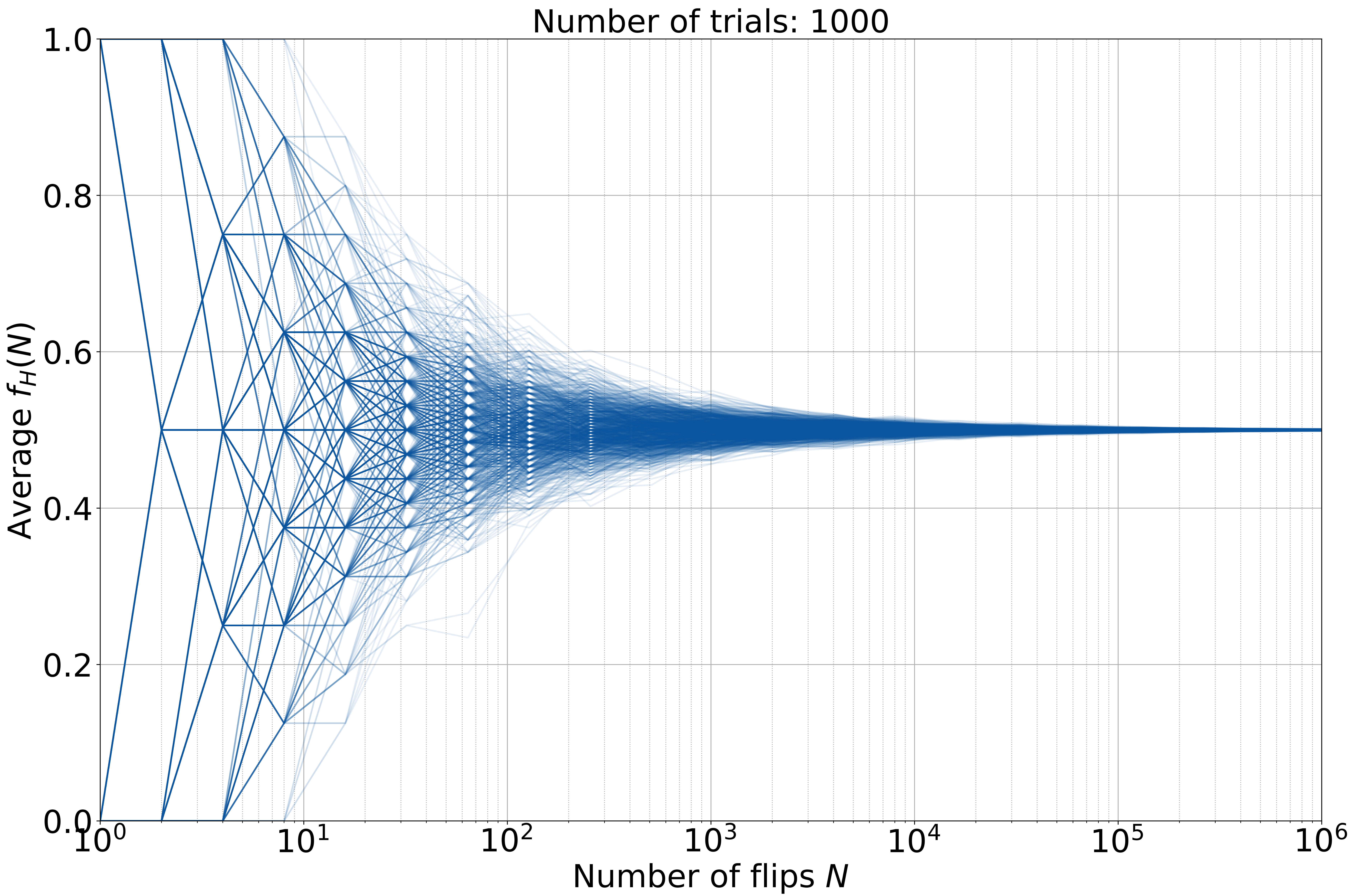
Frequentist Probability



Frequentist Probability



Frequentist Probability



Frequentist Probability

SCIENTISTS FOR PALESTINE

Probabilities can represent *frequencies.*
e.g. Flip a coin N times, define $N(H)$ to be the number of times it lands heads. The *relative frequency* $f_H(N)$ of landing heads is

$$f_H(N) = \frac{N(H)}{N}$$

The probability of a head, written $p(H)$ is

$$p(H) := \lim_{N \to \infty} f_H(N)$$

The symbol $\lim_{N \to \infty}$ is called a *limit*. It is the formal way of saying "when N gets big".
Frequentists: *event probability* = long run frequency in a repeatable experiment.

Bayesian Probability

Probabilities can represent *beliefs*

e.g. Given the results of a blood test, the probability that Aisha has a nasty disease is $p\,\%$.

e.g. The probability that it will rain in Ramallah on 31st July 2022 is $q\,\%$.

Such claims cannot be verified through *repeated* experimentation. This subjective or *Bayesian* interpretation expresses *degree of belief*.

Frequentist and Bayesian probability treated with same theory

Revd. Thomas Bayes

N.B. other interpretations exist: propensity, logical probability, mechanistic, etc.

*I've heard a rumor that the gentleman pictured above may not actually be the Reverend Thomas Bayes.

The Probability Axioms

1) Probability of event x is a non-negative real number

$$p(x) \geq 0 \quad \text{for all} \quad x \subseteq \Omega$$

2) Certain events have unit probability

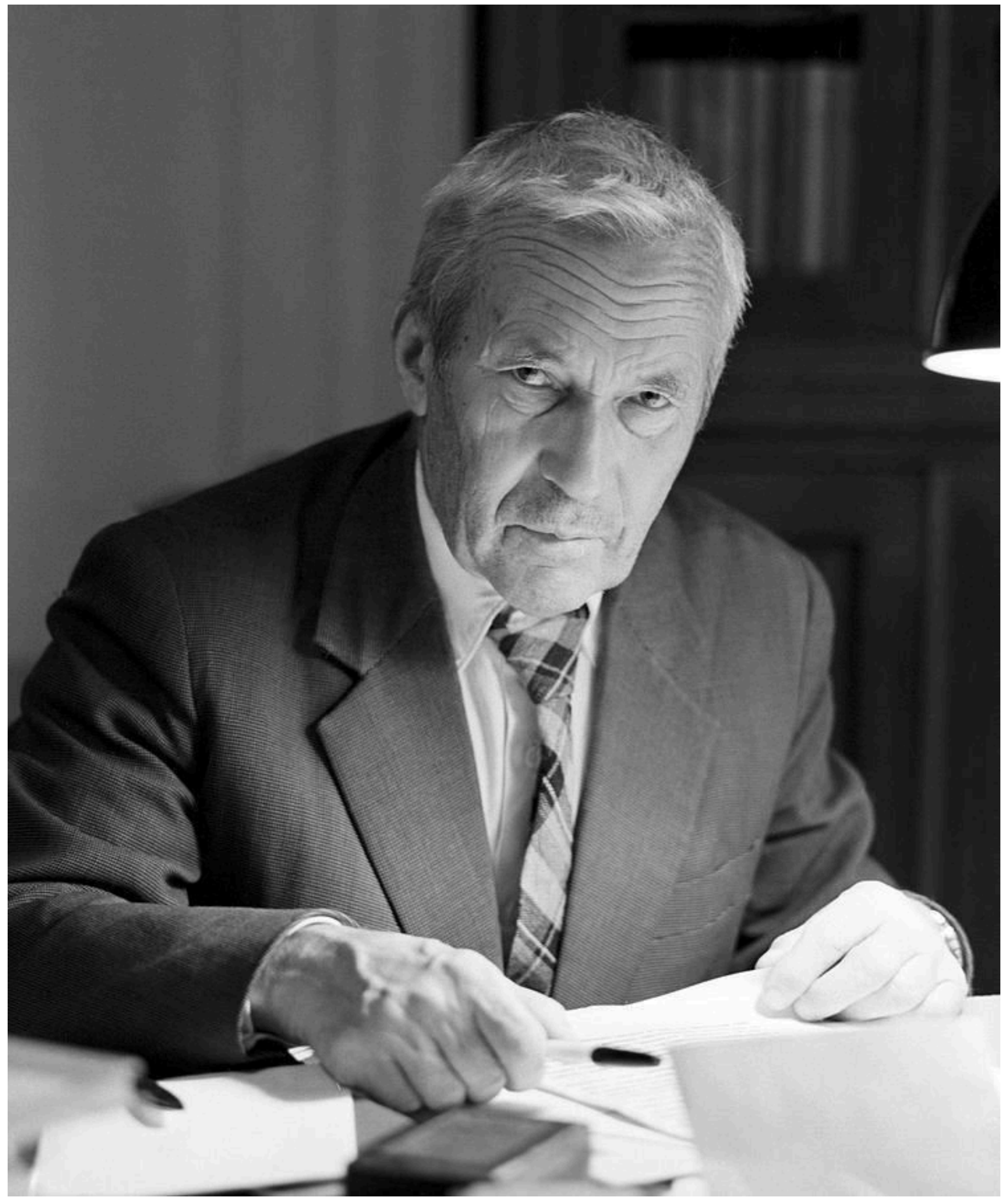
$$p(\Omega) = 1$$

3) Countable additivity: for disjoint events $x_1, x_2, \ldots, x_N$

$$p(x_1 \mathbin{\dot\cup} x_2 \mathbin{\dot\cup} \ldots \mathbin{\dot\cup} x_N) = p(x_1) + p(x_2) + \ldots + p(x_N)$$

Other rules:

- **Complement rule:** $p(\Omega \setminus x) = 1 - p(x)$
- **Impossible events:** $p(\emptyset) = 0$
- **Subsets:** $x_1 \subseteq x_2 \implies p(x_1) \leq p(x_2)$
- **Union rule:** $p(x_1 \cup x_2) = p(x_1) + p(x_2) - p(x_1 \cap x_2)$



Andrey Kolmogorov

Revision: PMFs

Sample space

A *sample space* Ω is the *set* of possible outcomes of an experiment. Outcomes are called *samples.*

Events

An *event* E is a subset of a sample space $E \subseteq \Omega$

Event space

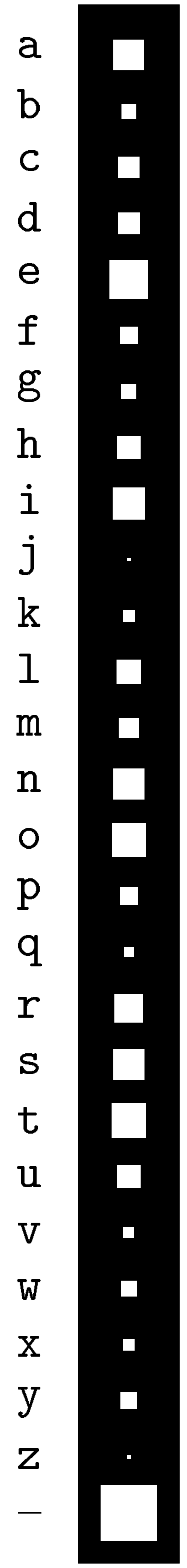
An *event space*[1] Σ is the space of all events $E \subseteq \Omega$

Probability Mass Function (PMF)

A *probability mass function*[2] p assigns a number in $[0,1]$ to every event in the event space.

- p(A) = 1 means that $A \in \Sigma$ is certain
- p(A) = 0 means that $A \in \Sigma$ will never happen
- If p(A) > p(B), then A is more likely than B

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

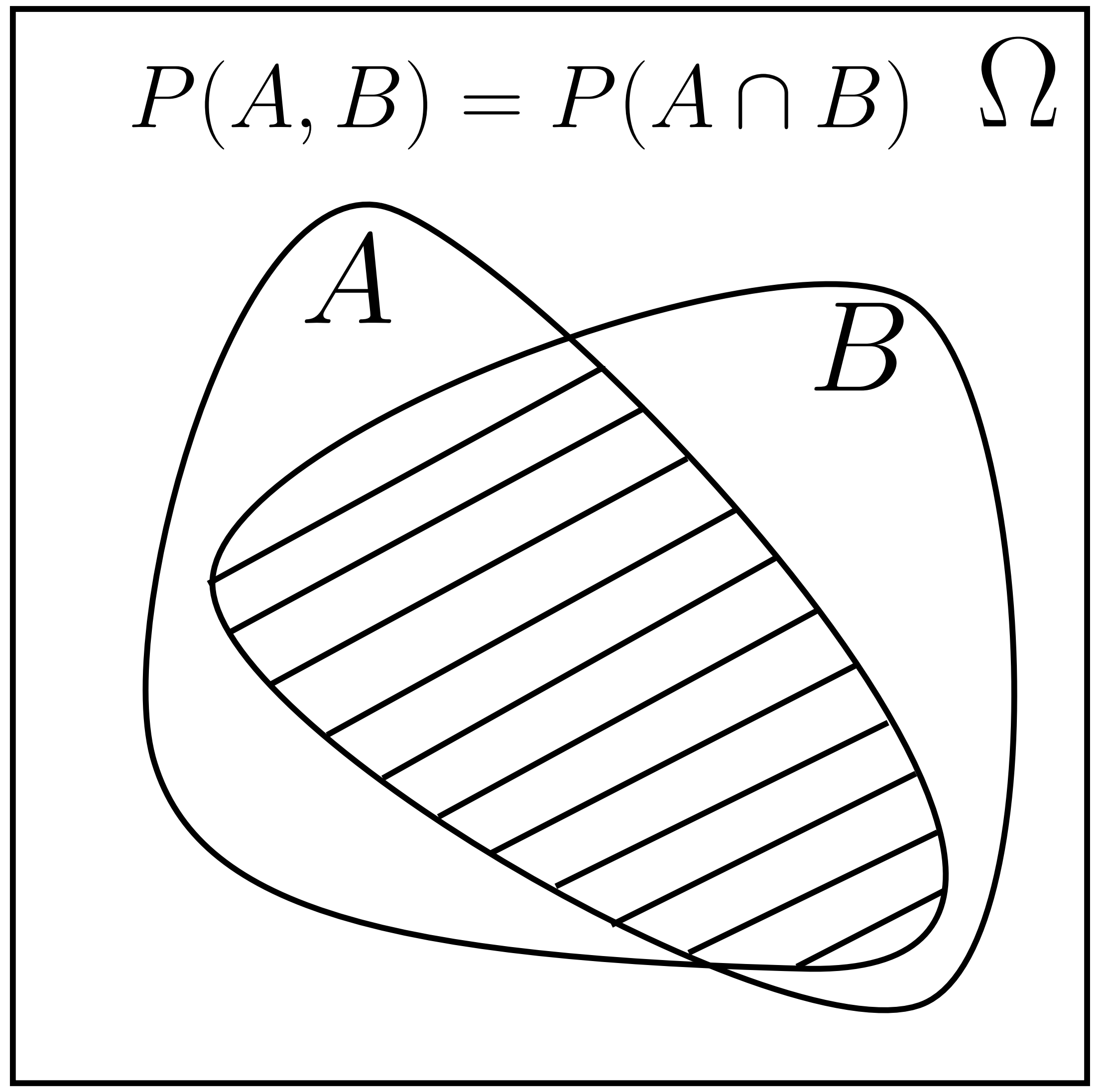


PMF over letters in English

[1] In the continuous setting the definition is a bit fiddly
[2] Note in statistics a capital P is used, but in machine learning we are lazy and just use a small p

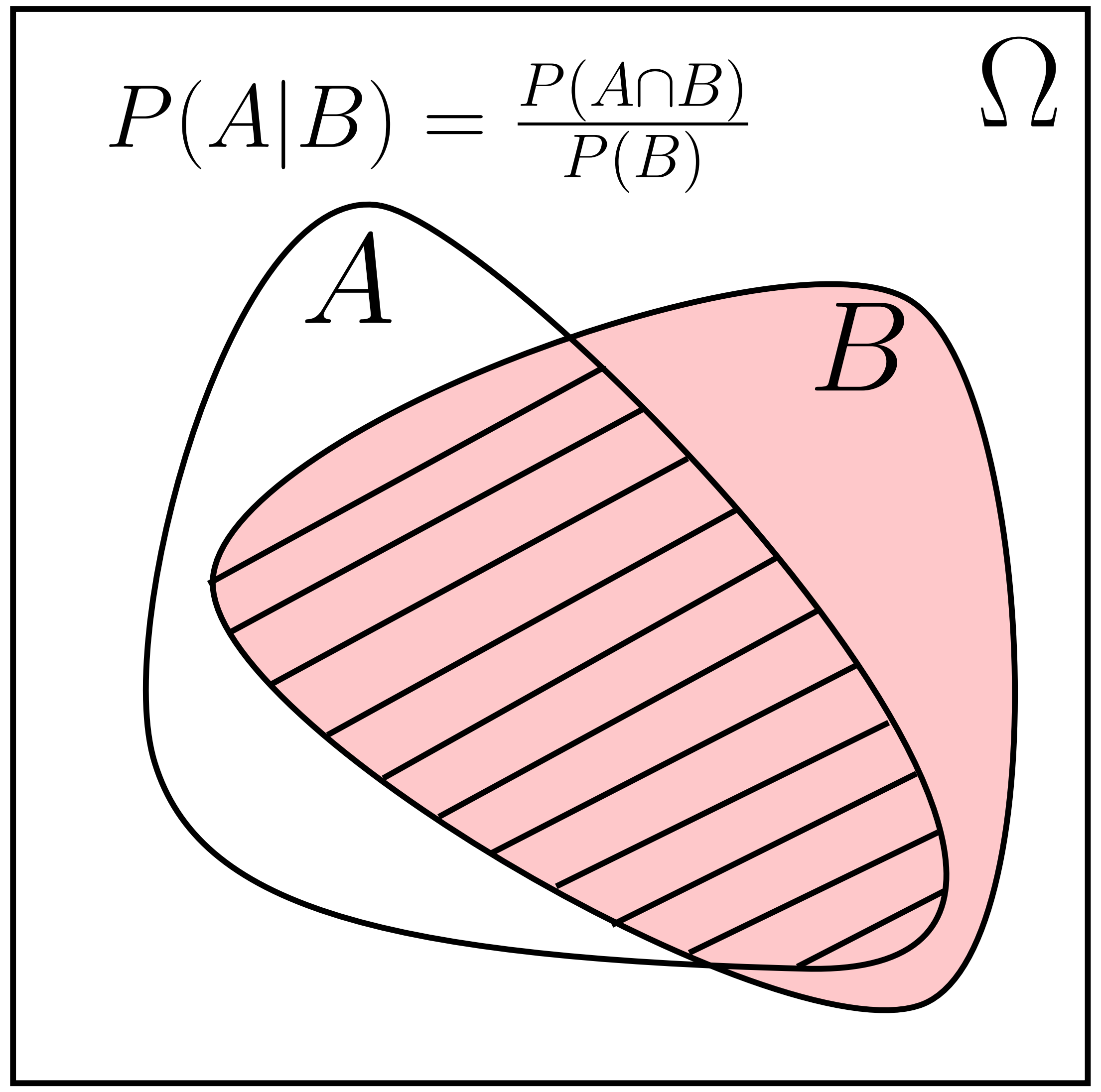
Conditional and Joint Probability

Joint probability: A and B co-occur



$$p(A, B) = p(B, A)$$

Conditional probability: B then A



$$p(A|B) \neq p(B|A)$$

Product rule $p(A, B) = p(A|B)p(B)$

Sum rule $p_A(A) = \sum_B p_{A,B}(A, B)$ **or** $p_A(A) = \int_B p_{A,B}(A, B) \, \mathrm{d}B$

Sum rule proof $\sum_B p(A, B) = \sum_B p(B|A)p(A) = p(A) \underbrace{\sum_B p(B|A)}_{=1} = p(A)$

Note sometimes we write $p(x)$ and other times we will write $p_X(x)$ depending on context

Some definitions (Google them later if you don't know them)

SCIENTISTS FOR PALESTINE

Bayes' Rule

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Given y has happened, what is prob. of x?

Proof

$$p(y|x)p(x) = p(x, y) = p(x|y)p(y)$$

Expectations

$$\mathbb{E}_{p(\mathbf{x})}[f(\mathbf{x})] = \int f(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x} \quad \text{or} \quad \sum_{\mathbf{x}} f(\mathbf{x})p(\mathbf{x})$$

Mean

$$\bar{\mathbf{x}} = \mathbb{E}_{p(\mathbf{x})}[\mathbf{x}]$$

Covariance

$$\boldsymbol{\Sigma} = \mathbb{E}_{p(\mathbf{x})}\left[(\mathbf{x} - \bar{\boldsymbol{x}})(\mathbf{x} - \bar{\boldsymbol{x}})^\top\right]$$

Change of variables formula

$$p_X(x) = p_Z(z)\left|\frac{\partial z}{\partial x}\right| \qquad x = f(z)$$

Jacobian tracks volume change

Proof

$$p_X(x) = \int_Z p(x|z)p_Z(z)\,\mathrm{d}z$$

$$\overset{\text{e.g.}}{=} \int_Z \delta(x - f(z))p_Z(z)\,\mathrm{d}z$$

$$= \int_U \delta(x - u)p_Z(f^{-1}(u))\left|\frac{\partial z}{\partial u}\right|\mathrm{d}u$$

$$= p_Z(z)\left|\frac{\partial z}{\partial x}\right|$$

Information Theory

Surprisal

$$I(\mathbf{x}) = -\log p(\mathbf{x})$$

Entropy = average surprise

$$H(\mathbf{x}) = \mathbb{E}_{p(\mathbf{x})}[-\log p(\mathbf{x})]$$

Kullback-Leibler divergence

$$D_{KL}(p(\mathbf{x})||q(\mathbf{x})) = \int p(\mathbf{x})\log\frac{p(\mathbf{x})}{q(\mathbf{x})}\,\mathrm{d}\mathbf{x}$$

Mutual information

$$\mathbb{I}[\mathbf{x}; \mathbf{y}] = D_{KL}(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y}))$$

Models

Probabilistic models

We are mostly concerned with models* which look like

$$p(\mathbf{x} \mid \boldsymbol{\theta})$$

In many cases x refers to an *observation* and refers to a set of *parameters*.

Probability

$$\boldsymbol{\theta}$$

$$\downarrow$$

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$$

*Sometimes we refer to $\{p(\mathbf{x}|\theta)\}_{\theta \in \Theta}$ as a model, other times we refer to $p(\mathbf{x}|\theta)$ for a single θ as the model

You build a model of radioactive decay.

Model 1: Particles decay x cm from the source, following an exponential distribution:

$$p(x|\lambda) = \frac{1}{Z} e^{-\lambda x}$$

Variable — Parameter — Normalization constant

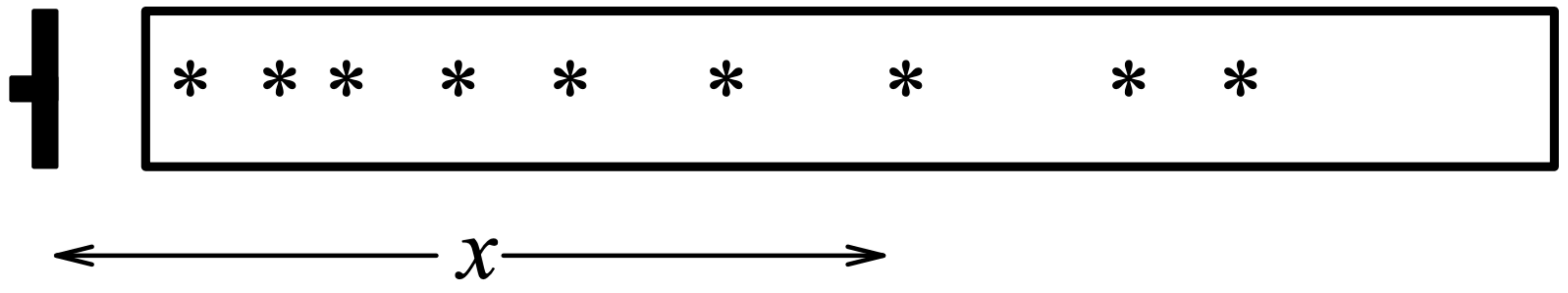
$$\int_0^\infty \frac{1}{Z} e^{-\lambda x}\, \mathrm{d}x = 1 \implies Z = \frac{1}{\lambda}$$

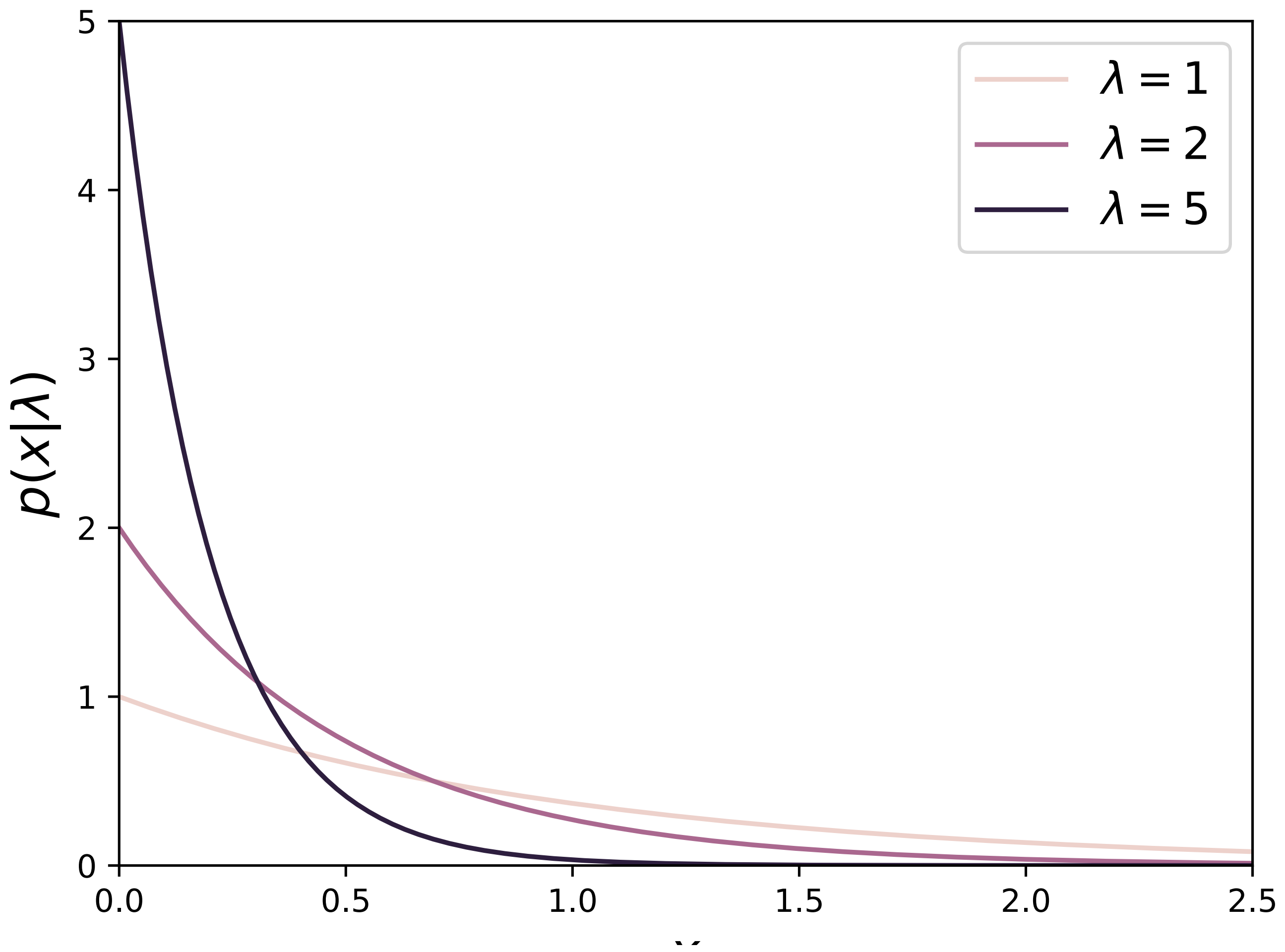
Model 2: A gamma distribution

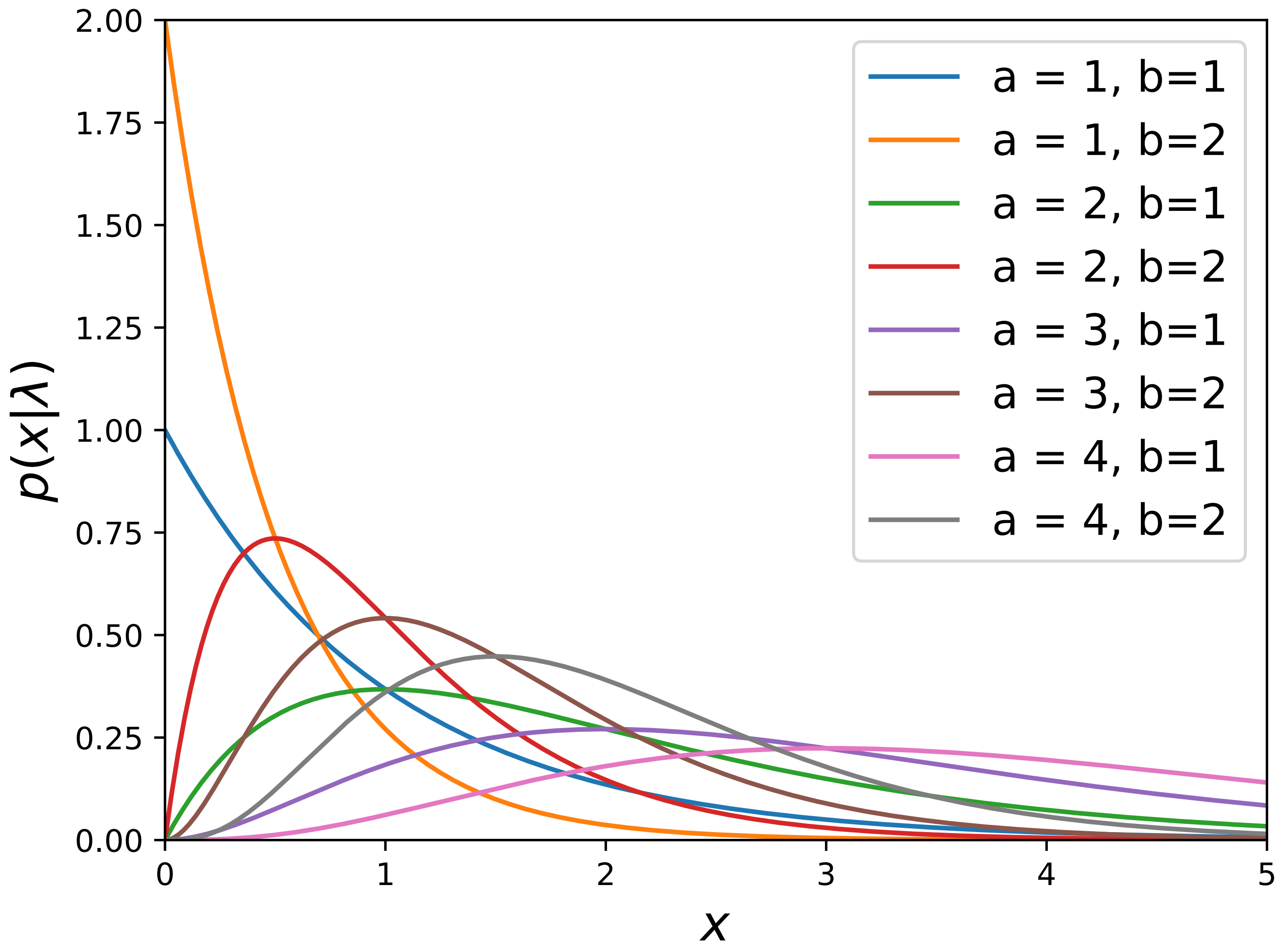
$$p(x|\alpha, \beta) = \frac{1}{Z} x^{\alpha-1} e^{-\beta x}, \quad Z = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

More than 1 parameter

How to model multiple observations?







Joint probability of *sequence* $\{x_1, x_2, \ldots, x_N\}$:

$$p(x_1, x_2, ..., x_N)$$

Marginal independence

Probability of each observation *independent* (doesn't depend on other observations) and *identical* (from the same distribution).

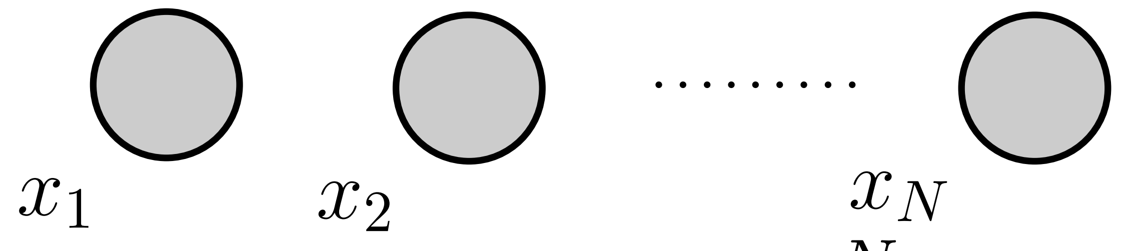
$$p(x_1 | x_i) = p(x_1) \quad \text{for all } x_i \neq x_1$$

Conditional independence

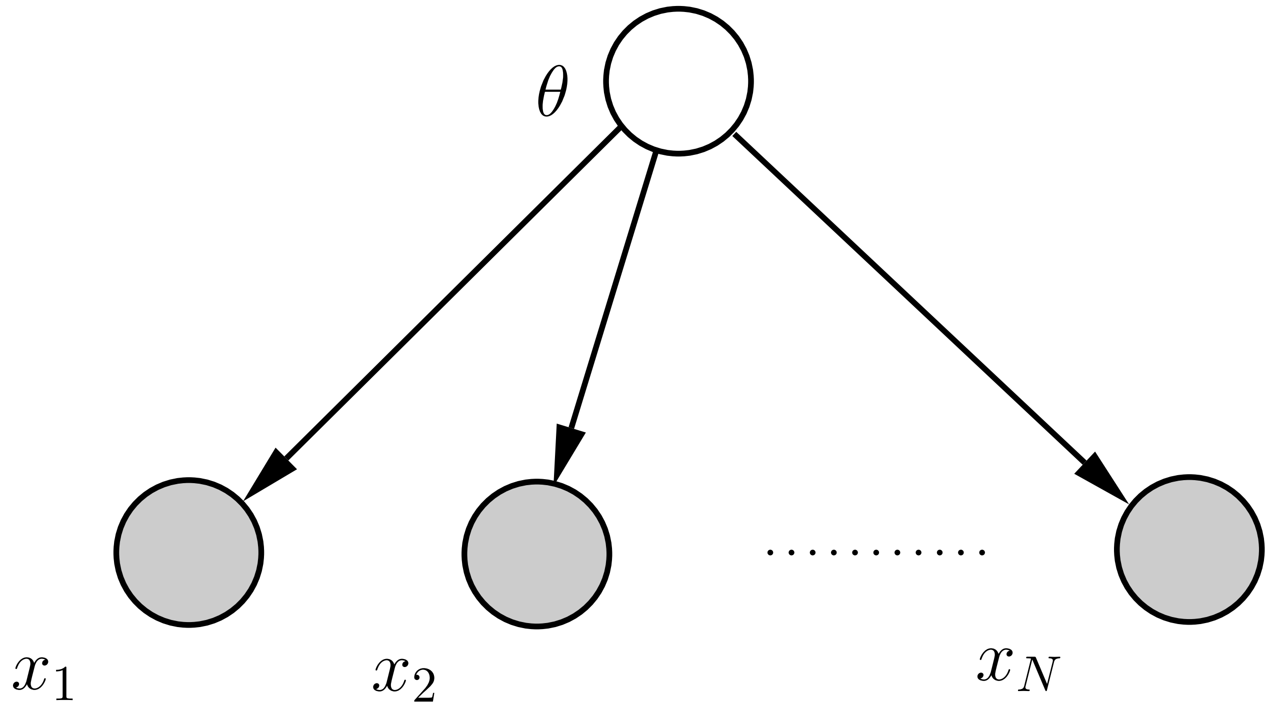
Independence *conditional* on extra information

$$p(x_1, x_2 | x_i) = p(x_1 | x_i) p(x_2 | x_i)$$

for $x_i \neq x_1, x_i \neq x_2$



$$p(x_1, x_2, ..., x_N) = \prod_{n=1}^{N} p(x_n)$$



$$p(x_1, x_2, ..., x_N | \theta) = \prod_{n=1}^{N} p(x_n | \theta)$$

Forward models

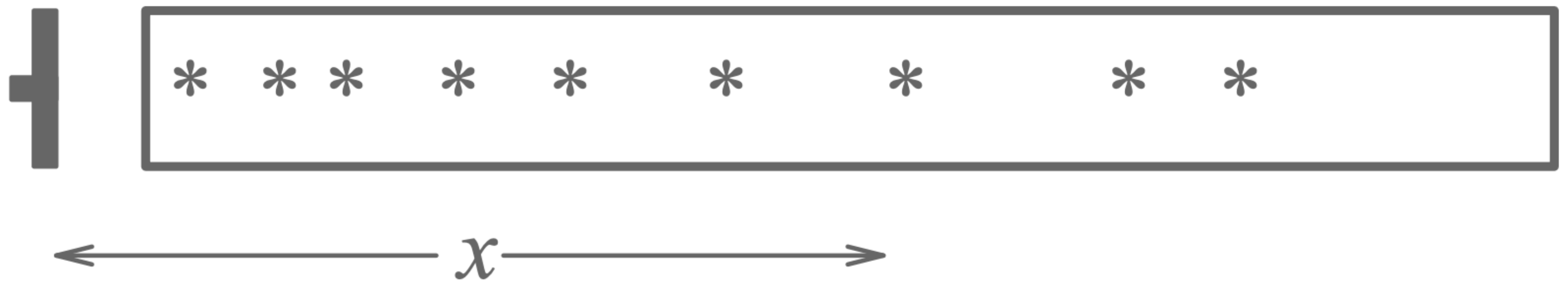
You build a model of radioactive decay.

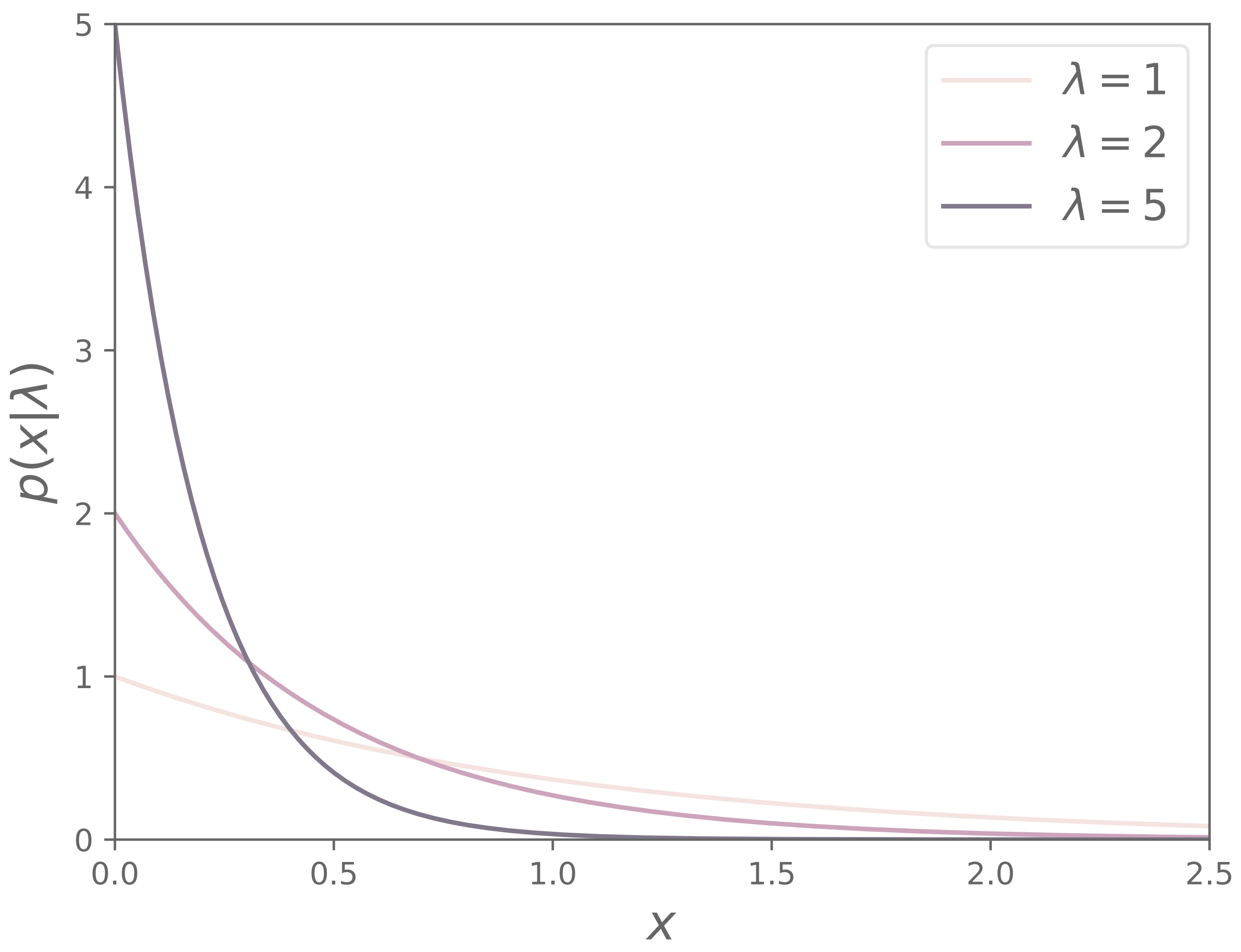
Model 1: Particles decay x cm from the source, following an exponential distribution:

$$p(x|\lambda) = \frac{1}{Z} e^{-\lambda x}$$

Variable — Parameter — Normalization constant

$$\int_0^\infty \frac{1}{Z} e^{-\lambda x} \, \mathrm{d}x = 1 \implies Z = \frac{1}{\lambda}$$





How to model multiple observations?

$$p(x_1, x_2, ..., x_N|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \frac{1}{Z} e^{-\lambda x_n} = \lambda^N e^{-\lambda \sum_{n=1}^{N} x_n}$$

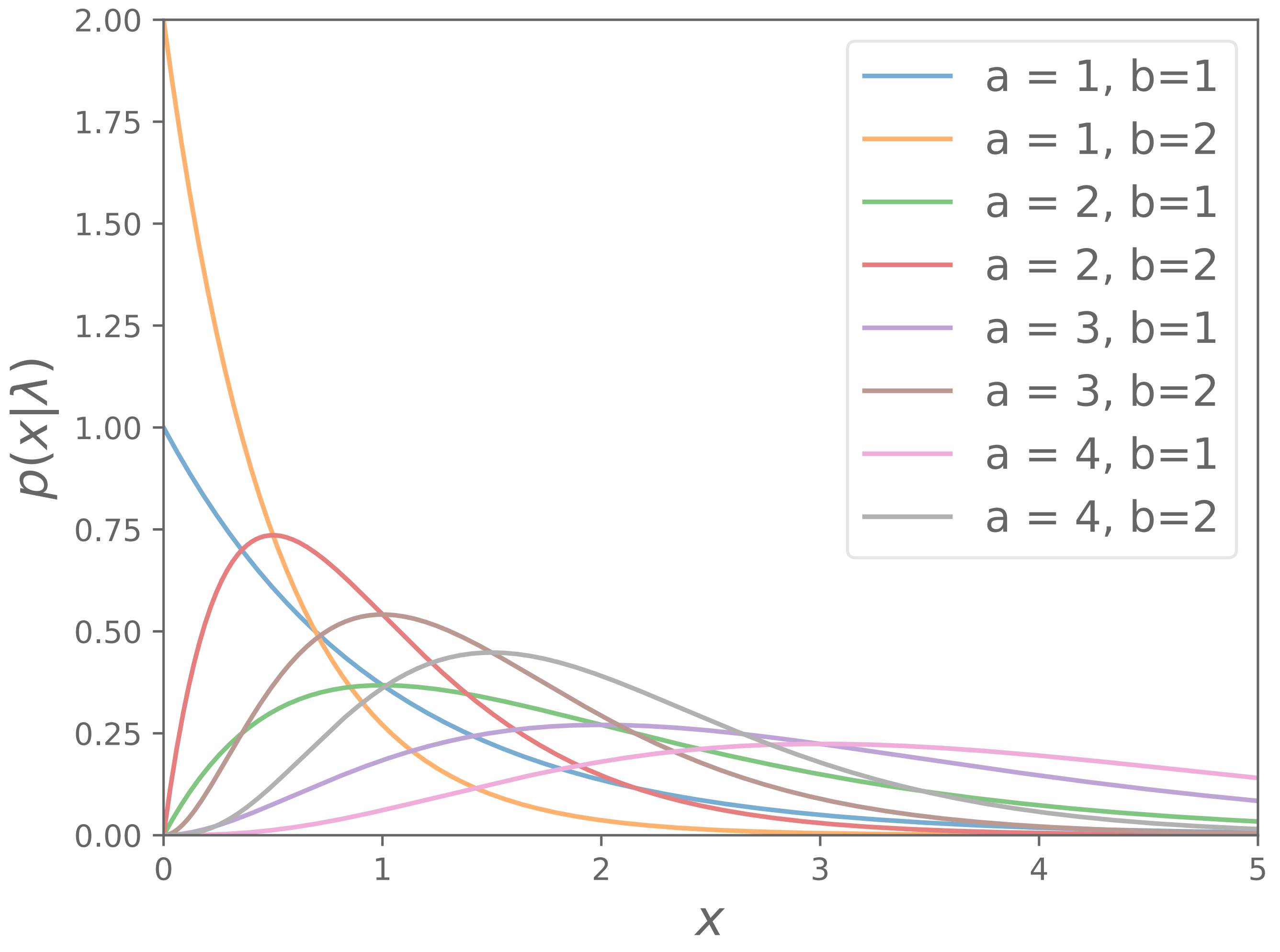
Forward models

Model 2: A gamma distribution

$$p(x|\alpha, \beta) = \frac{1}{Z} x^{\alpha-1} e^{-\beta x}, \quad Z = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

More than 1 parameter

How to model multiple observations?



How to model multiple observations?

$$p(x_1, x_2, ..., x_N|\alpha, \beta) = \prod_{n=1}^{N} p(x_n|\alpha, \beta) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^N \prod_{n=1}^{N} x_n^{\alpha-1} e^{-\beta x_n}$$

Probabilistic models

SCIENTISTS FOR PALESTINE

We are mostly concerned with models which look like

$$p(\mathbf{x} \mid \boldsymbol{\theta})$$

In many cases x refers to an *observation* and refers to a set of *parameters*.

Probability

$$\boldsymbol{\theta}$$

$$\downarrow$$

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$$

Statistics

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \}$$

$$\downarrow$$

$$\boldsymbol{\theta}$$

*Sometimes we refer to $\{p(\mathbf{x}|\theta)\}_{\theta \in \Theta}$ as a model, other times we refer to $p(\mathbf{x}|\theta)$ for a single θ as the model

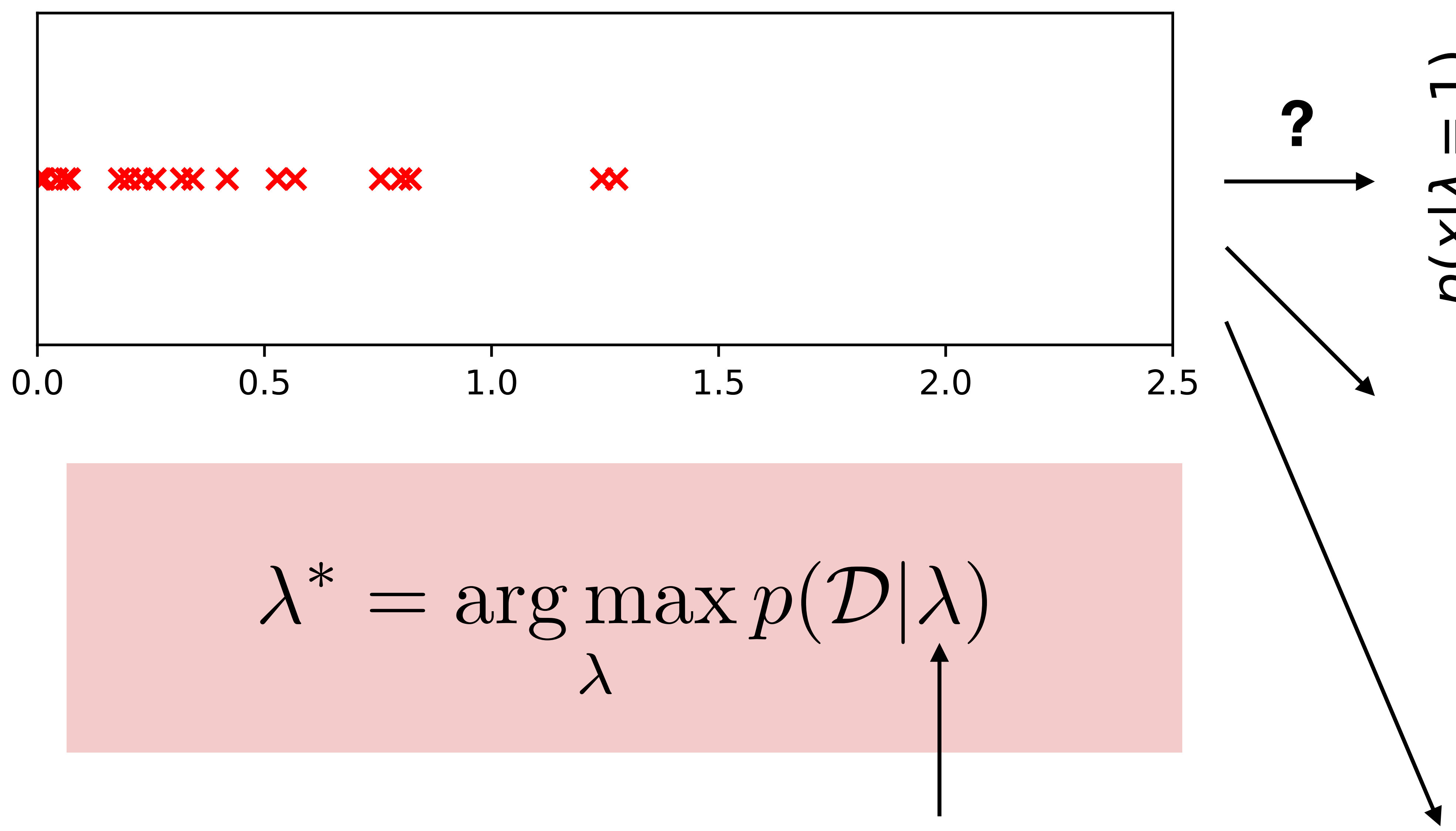
Maximum likelihood

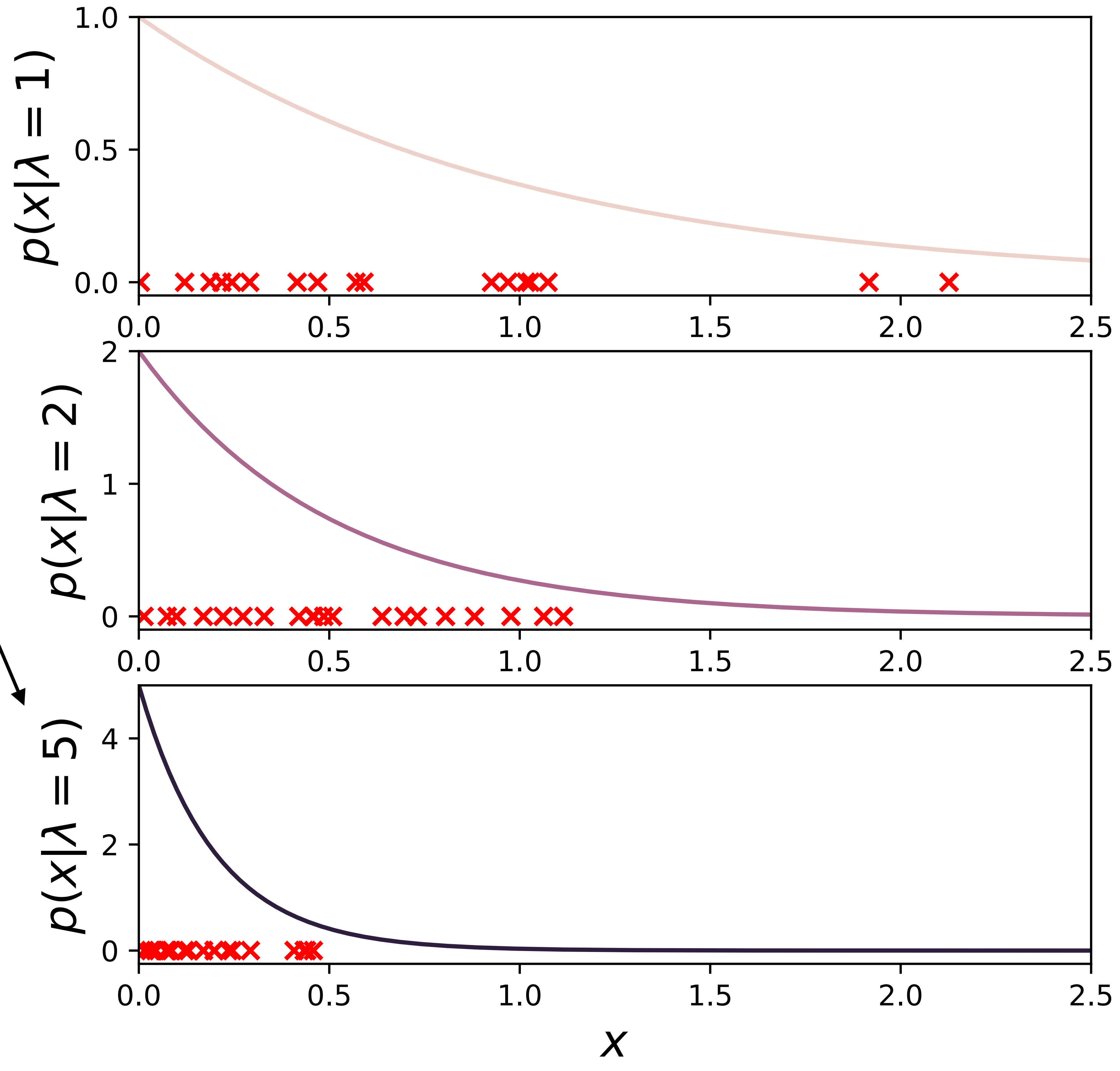
$$\mathcal{D} = \{x_1, x_2, ..., x_N\}$$

Forward model: generate **data given parameters**

Inference: find **parameters given data**

Idea: enumerate every possible forward model and see if it matches the data





$$\lambda^* = \arg\max_\lambda p(\mathcal{D}|\lambda)$$

Likelihood: function of the parameter λ

We say the likelihood of λ given $\mathcal{D}$, never the likelihood of $\mathcal{D}$ given λ!

Maximum log-likelihood

SCIENTISTS FOR PALESTINE

We find the maximum by setting[1] the derivative to 0

$$\lambda^* = \arg\max_\lambda p(\mathcal{D}|\lambda) \implies \frac{\partial}{\partial \lambda} p(\mathcal{D}|\lambda^*) = 0$$

We prefer to find the maximum of the log-likelihood

$$\lambda^* = \arg\max_\lambda p(\mathcal{D}|\lambda) \iff \lambda^* = \arg\max_\lambda \log p(\mathcal{D}|\lambda)$$

$$\begin{aligned}\lambda^* &= \arg\max_\lambda \log p(\mathcal{D}|\lambda) \\ &= \arg\max_\lambda \log \prod_{n=1}^{N} p(x_n|\lambda) \\ &= \arg\max_\lambda \sum_{n=1}^{N} \log p(x_n|\lambda)\end{aligned}$$

Why?

1) Small likelihoods → numerical underflow
2) Derivatives of sums easier than derivatives of products

$$\frac{\partial}{\partial \lambda}\left[f_1(\lambda) f_2(\lambda) f_3(\lambda)\right] = \frac{\partial f_1}{\partial \lambda} f_2(\lambda) f_3(\lambda) + f_1(\lambda) \frac{\partial f_2}{\partial \lambda} f_3(\lambda) + f_1(\lambda) f_2(\lambda) \frac{\partial f_3}{\partial \lambda}$$

Which do you prefer?

$$\frac{\partial}{\partial \lambda}\left[\log\left(f_1(\lambda) f_2(\lambda) f_3(\lambda)\right)\right] = \frac{\partial \log f_1}{\partial \lambda} + \frac{\partial \log f_2}{\partial \lambda} + \frac{\partial \log f_3}{\partial \lambda}$$

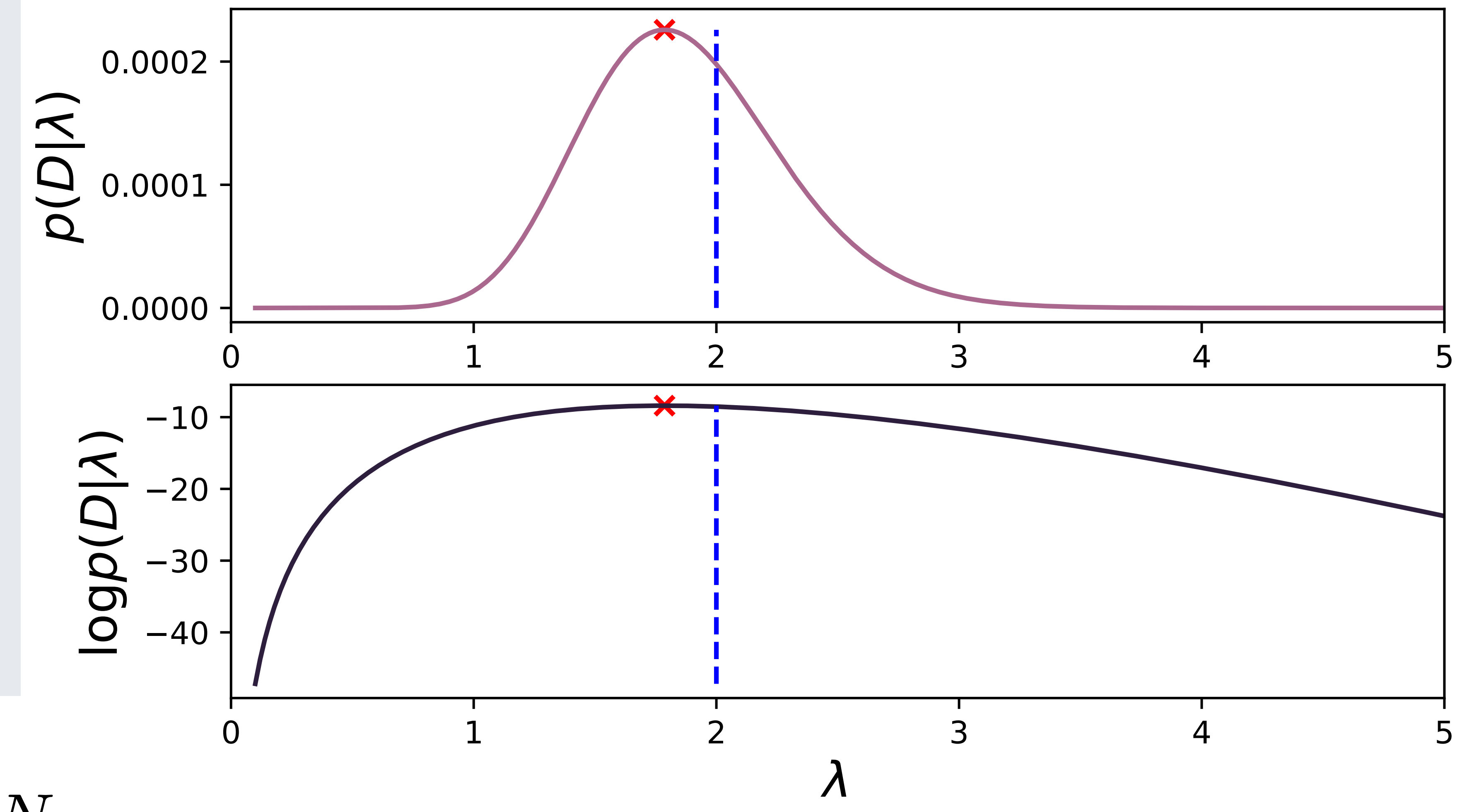
[1] Note we should also check that the Hessian is positive-definite, but for typical distributions this is not necessary, since members of the exponential family are log-concave in the parameters.

Maximum Likelihood Example

You build a model of radioactive decay.

Model 1: Particles decay x cm from the source, following an exponential distribution:

$$p(x|\lambda) = \frac{1}{Z}e^{-\lambda x}$$



$$\lambda^* = \arg\max_\lambda \log p(\mathcal{D}|\lambda) = \arg\max_\lambda \sum_{n=1}^{N} \log p(x_n|\lambda)$$

$$= \arg\max_\lambda \sum_{n=1}^{N} \log \lambda e^{-\lambda x_n} = \arg\max_\lambda N \log \lambda - \sum_{n=1}^{N} \lambda x_n$$

$$\frac{\partial}{\partial \lambda}\left(N \log \lambda - \sum_{n=1}^{N} \lambda x_n\right) = \frac{N}{\lambda} - \sum_{n=1}^{N} x_n = 0 \qquad \implies \lambda^* = \frac{1}{\frac{1}{N}\sum_{n=1}^{N} x_n}$$

The Exponential Family

Which distributions have analytical maximum likelihood solutions? Most of the distributions we have looked at are fairly similar. They have three main components:

$$p(x|\boldsymbol{\theta}) = \underbrace{\frac{1}{Z(\boldsymbol{\theta})}}_{\text{normalizer}} \cdot \overbrace{b(x)}^{\text{fnc of } x} \cdot \underbrace{\exp\left\{\boldsymbol{\theta}^\top \mathbf{t}(x)\right\}}_{\text{exp of linear fnc of } \boldsymbol{\theta}}$$

Natural parameters[1] θ and *sufficient statistics* $\mathbf{t}(x)$.

This may seem like an odd choice, but it has some very handy properties, which allow for **lightning fast** computation. At the ML solution

Model expectation

$$\mathbb{E}_{p(x|\boldsymbol{\theta})}\left[\mathbf{t}(x)\right] = \frac{1}{N}\sum_{n=1}^{N} \mathbf{t}(x_n)$$

Data expectation

Method of moments: Use a mean value mapping to recover the ML parameters, tractably

$$\tau(\boldsymbol{\theta}) = \mathbb{E}_{p(x|\boldsymbol{\theta})}\left[\mathbf{t}(x)\right] \implies \boldsymbol{\theta}^* = \tau^{-1}\left(\frac{1}{N}\sum_{n=1}^{N} \mathbf{t}(x_n)\right)$$

[1] There are many different names and notations for this, so beware!

Examples

SCIENTISTS FOR PALESTINE

Bernoulli

$$p(x|\pi) = \pi^x (1-\pi)^{1-x}$$

Uniform

$$p(x) = \mathbb{I}[x \in [0,1]]$$

Poisson

$$p(x|\lambda) = \frac{1}{Z}\frac{\lambda^x}{x!}, \quad Z = e^\lambda$$

Categorical

$$p(\mathbf{x}|\boldsymbol{\pi}) = \prod_{i=1}^K \pi_i^{x_i}$$

Exponential

$$p(x|\lambda) = \frac{1}{Z}\exp\{-\lambda x\}, \quad Z = \frac{1}{\lambda}$$

Gamma

$$p(x|\alpha,\beta) = \frac{1}{Z}x^{\alpha-1}\exp\{-\beta x\}, \quad Z = \frac{\Gamma(\alpha)}{\beta^\alpha}$$

Gaussian

$$p(x|\mu,\sigma^2) = \frac{1}{Z}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad Z = \sqrt{2\pi\sigma^2}$$

Dirichlet

$$p(x|\alpha,\beta) = \frac{1}{Z}\prod_{i=1}^K x_i^{\alpha-1}, \quad Z = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$$

Bayesian Inference

We just learnt about maximum likelihood, where we solved

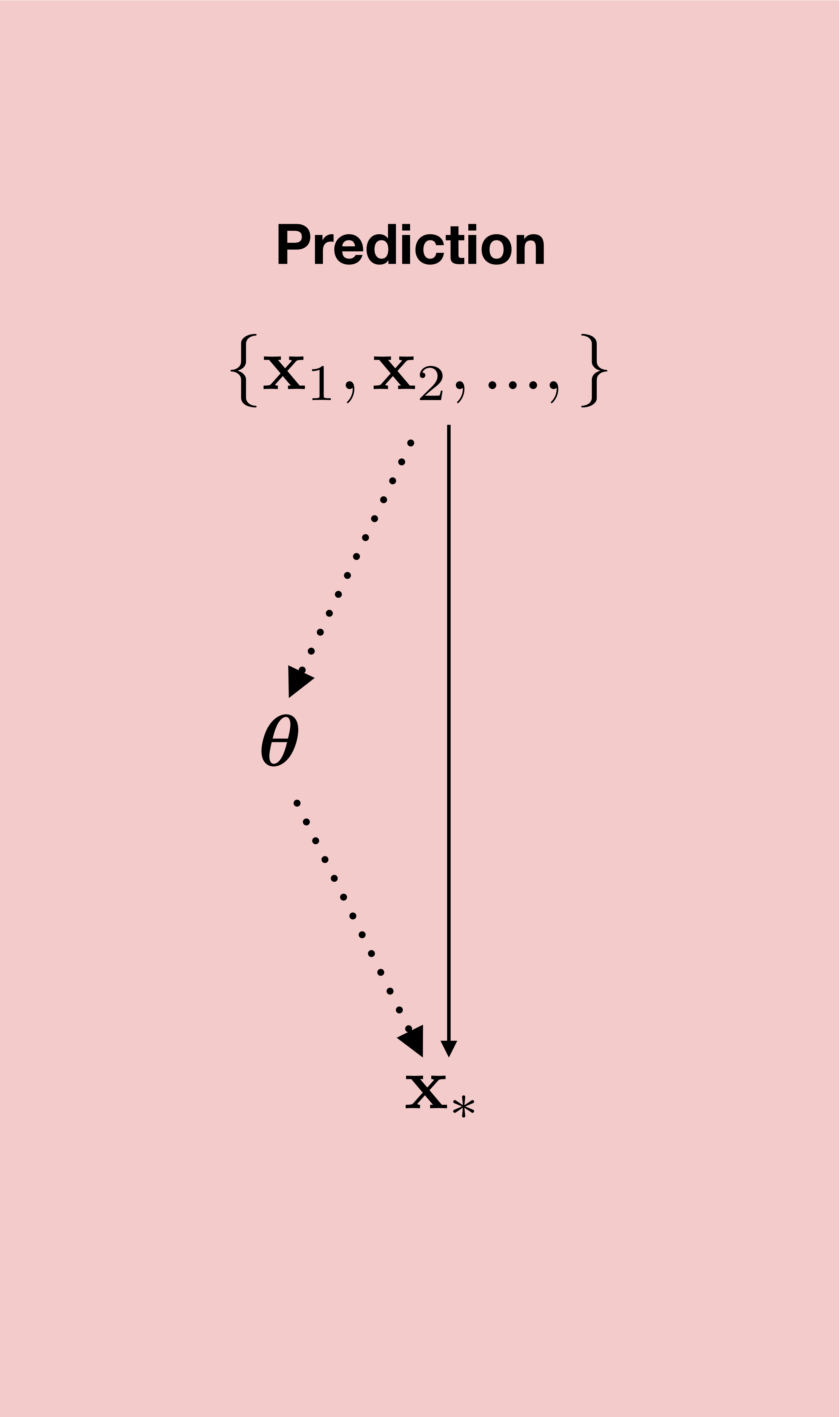
$$\theta_{\mathrm{ML}} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

- Is this the best we can do?
- **It is almost certainly wrong[1]:** $p(\theta_{\mathrm{ML}} = \theta_{\mathrm{true}}) = 0$
- Our model is almost certainly wrong as well

Why do we want θ_{ML} anyway?

Options

1. We are actually interested in knowing θ
2. We don't care: want to generate new samples x_*



[1] For continuous parameterizations

Bayesian Inference

SCIENTISTS FOR PALESTINE

Let's think about the distribution $p(x_*|\mathcal{D})$

We can compute it from *known* quantities:

$$p(x_*|\mathcal{D}) = \int p(x_*, \theta|\mathcal{D})\,\mathrm{d}\theta$$
$$= \int p(x_*|\theta)p(\theta|\mathcal{D})\,\mathrm{d}\theta$$

Weighted average

Forward model

This approach is called *generative modeling*

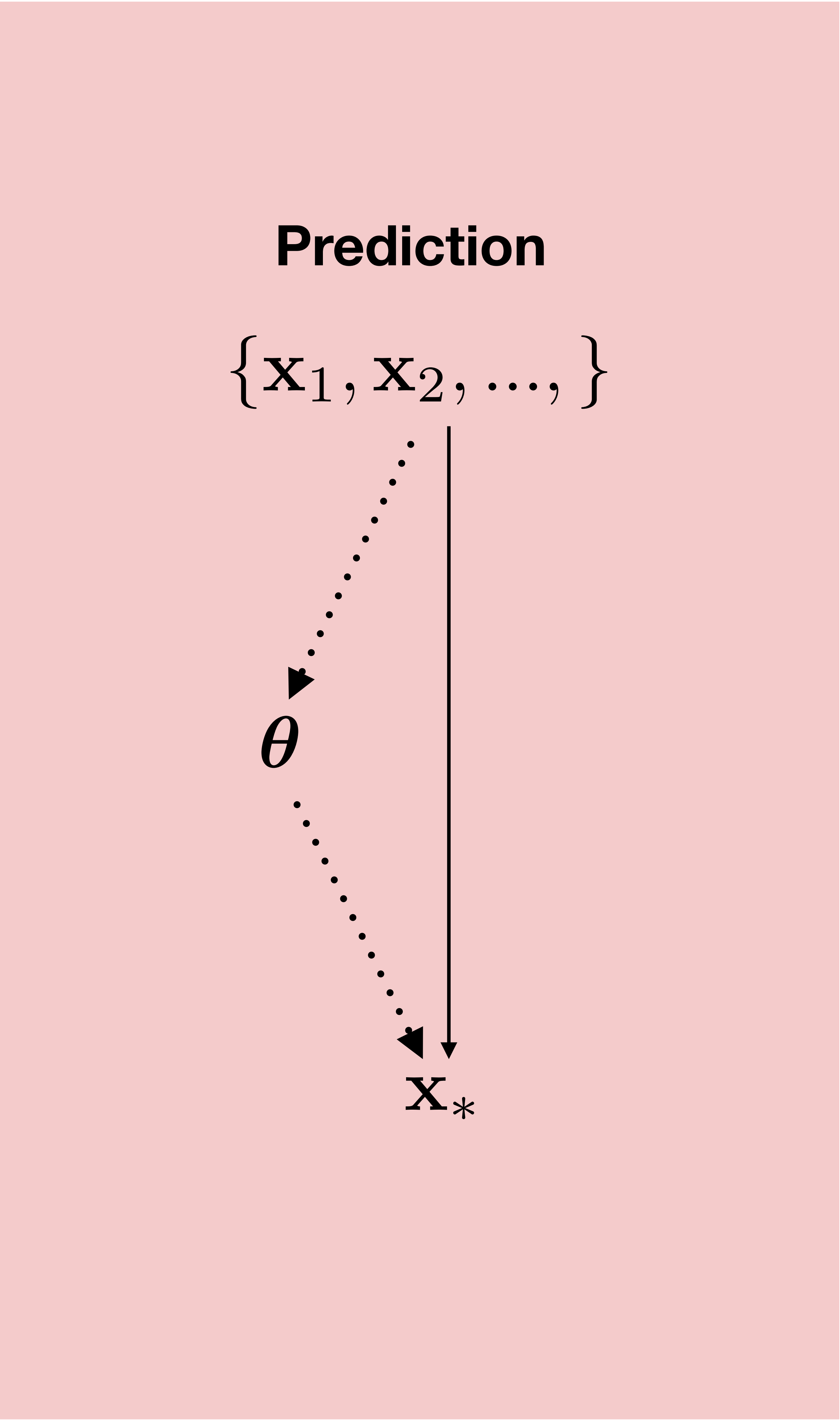
Likelihood Prior

Posterior distribution $$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\,\mathrm{d}\theta}$$

Evidence/marginal likelihood

Read, "*the probability of the parameters, given the data*".

More descriptive than point estimate θ_{ML}.

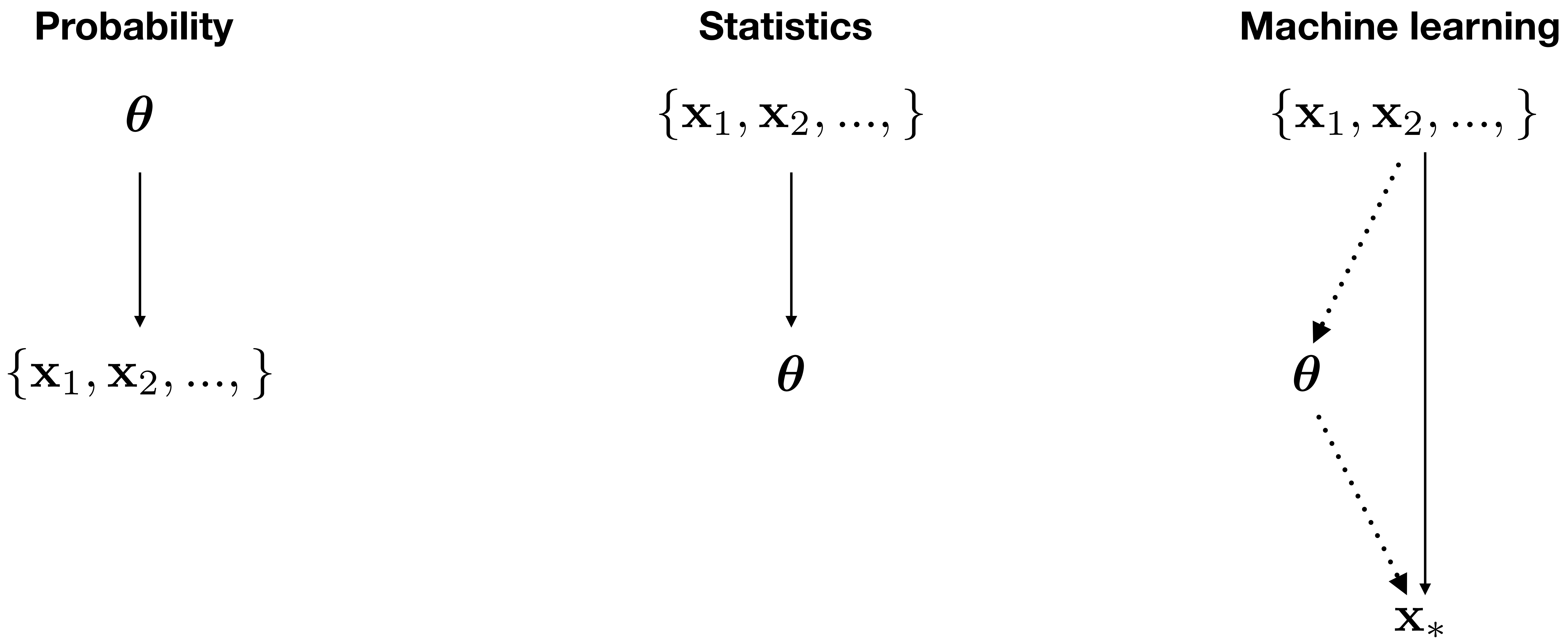


Probabilistic models

We are mostly concerned with models which look like

$$p(\mathbf{x} \mid \boldsymbol{\theta})$$

In many cases x refers to an *observation* and refers to a set of *parameters*.



*Sometimes we refer to $\{p(\mathbf{x}|\theta)\}_{\theta \in \Theta}$ as a model, other times we refer to $p(\mathbf{x}|\theta)$ for a single θ as the model

Example: The Bent Coin[1]

You are given a bent coin. You flip it N times. It lands heads H times.

The probability the coin lands heads is π, what is the posterior $p(\pi|\mathcal{D})$?

$$p(\pi|\mathcal{D}) = \frac{p(\mathcal{D}|\pi)p(\pi)}{p(\mathcal{D})} = \frac{\left[\prod_{i=1}^N p(x_i|\pi)\right] p(\pi)}{p(\mathcal{D})}$$

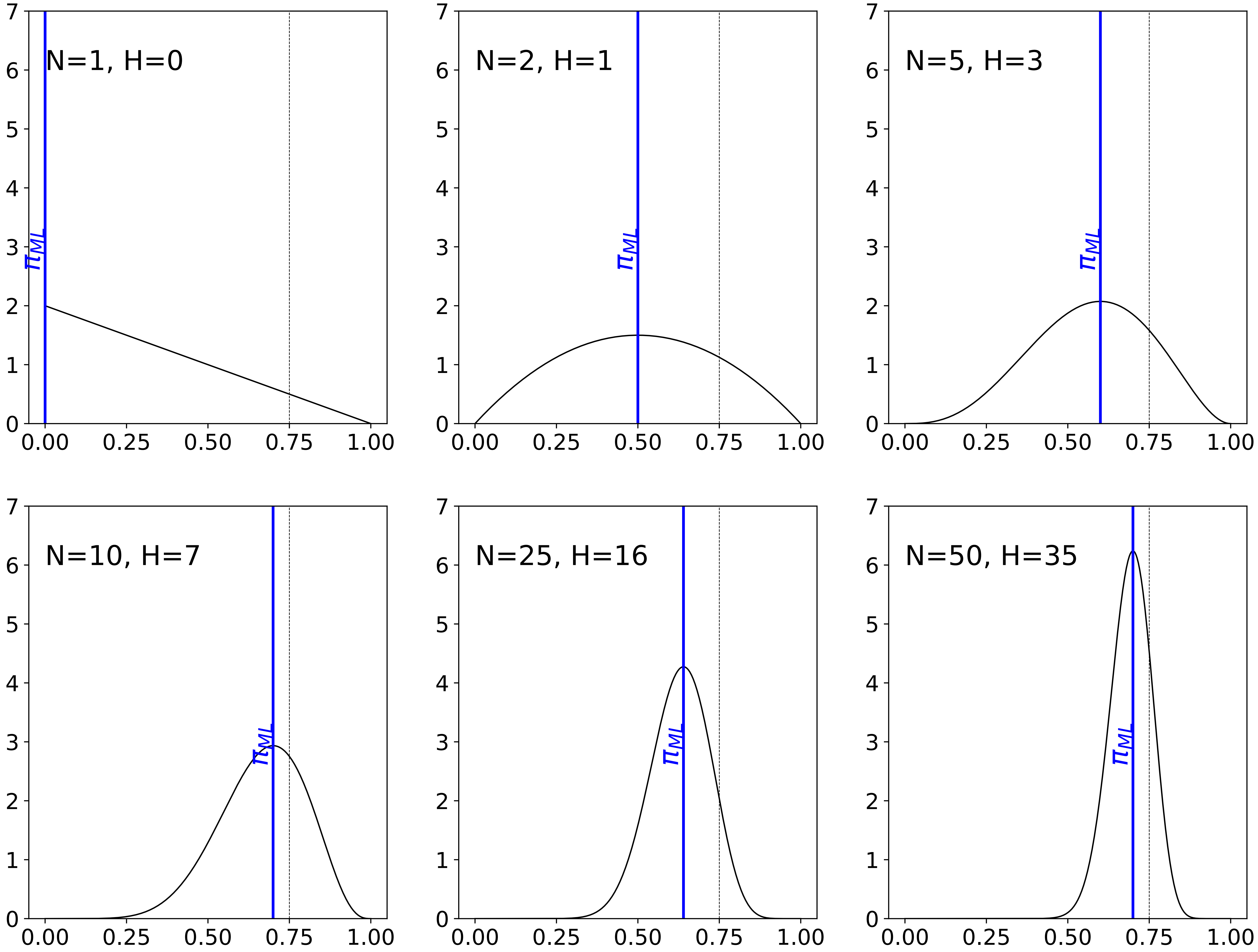
We need a prior on π, let's pick a uniform distribution

$$\begin{aligned} p(\pi|\mathcal{D}) &= \frac{\left[\prod_{i=1}^N \pi^{x_i}(1-\pi)^{1-x_i}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} \\ &= \frac{\left[\pi^H (1-\pi)^{N-H}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} \\ &= \frac{1}{Z}\pi^H(1-\pi)^{N-H} \end{aligned}$$

[1] This is the original inference problem studied by Thomas Bayes in 1763.

Example: The Bent Coin[1]

D = [0 1 0 1 1 1 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1
1 1 1 1 1 0 0 0 1 0 0 1 1]



Example: The Bent Coin[1]

Notice how the posterior is 'less temperamental' than the likelihood function.

$$p(\pi|\mathcal{D}) = \frac{\left[\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} = \frac{\left[\pi^H (1-\pi)^{N-H}\right] \mathbb{I}[\pi \in [0,1]]}{p(\mathcal{D})} = \frac{1}{Z}\pi^H(1-\pi)^{N-H}$$

Next we need to figure out the *marginal likelihood*

$$Z = p(\mathcal{D}) = \int p(\mathcal{D}, \pi)\, \mathrm{d}\pi = \int \underbrace{p(\mathcal{D}|\pi)}_{\text{likelihood}}\ \underbrace{p(\pi)}_{\text{prior}}\, \mathrm{d}\pi.$$

The marginal likelihood is an instance of the famous Beta integral[1]

$$p(\mathcal{D}) = \int_0^1 \pi^H (1-\pi)^{N-H}\, \mathrm{d}\pi = B(H+1, N-H+1) = \frac{H!(N-H)!}{(N+1)!}$$

$$\boxed{p(\pi|\mathcal{D}) = \frac{(N+1)!}{H!(N-H)!}\pi^H(1-\pi)^{N-H}}$$

Don't worry if this integral scares you. It frightens me too! Resources such as Wolfram Alpha, Wikipedia, the Bishop book, and the MacKay book are handy.

[1] $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\, \mathrm{d}t$

Example: The Bent Coin[1]

The posterior has the form of a *Beta distribution*

$$\mathrm{Beta}(\pi|\alpha,\beta) = \frac{1}{Z(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \qquad Z(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

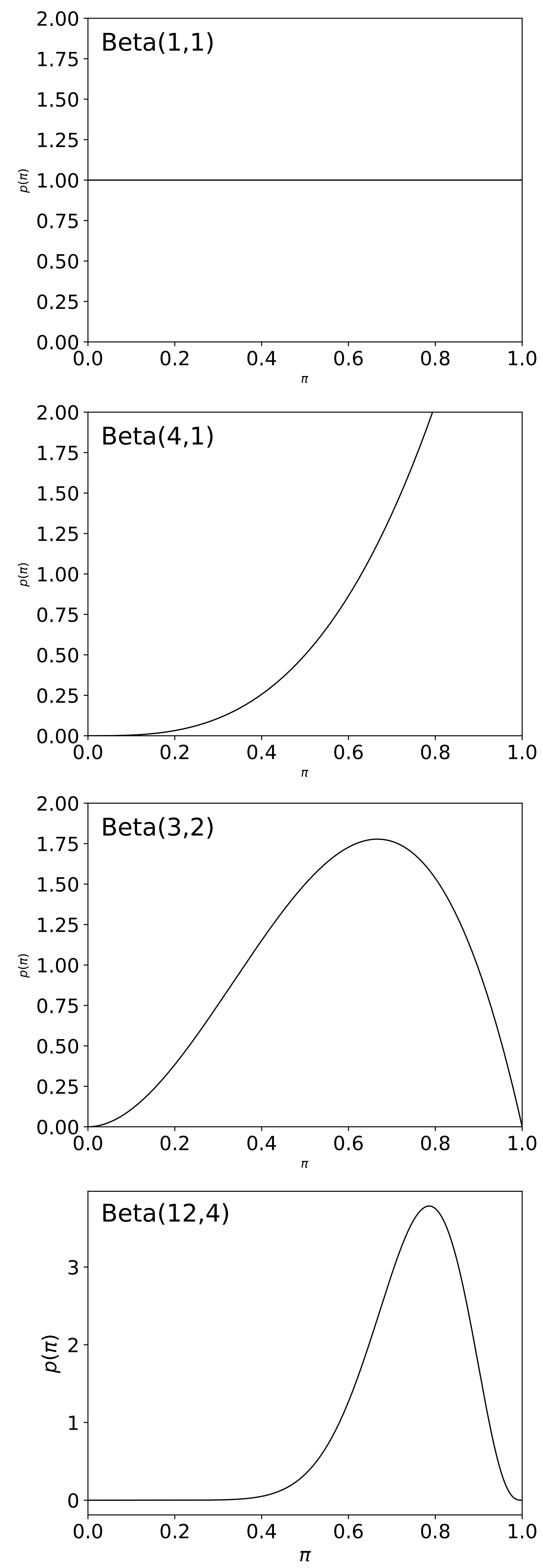
- The Beta distribution is a probability distribution over probabilities.
- The two parameters α and β control the shape of the distribution.
- The *Gamma* function[1] satisfies $\Gamma(\alpha) = (\alpha-1)!$ and $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

Posterior predictive distribution: *Laplace's rule of succession*

$$p(x_* = \mathrm{head}|\mathcal{D}) = \int \underbrace{p(x_* = \mathrm{head}|\pi)}_{\text{forward likelihood}} \underbrace{p(\pi|\mathcal{D})}_{\text{posterior}} \, \mathrm{d}\pi$$

$$= \int_0^1 \pi \cdot \frac{\pi^H(1-\pi)^{N-H}}{p(\mathcal{D})} \, \mathrm{d}\pi$$

$$= \frac{H+1}{N+2}$$



[1] $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$

ML/MAP/Bayes'

Sometimes, instead of the ML estimate people take the *maximum a posteriori*

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\theta|\mathcal{D})$$

$$= \arg\max_{\theta} \underbrace{\log p(\mathcal{D}|\theta)}_{\text{log-likelihood}} + \underbrace{\log p(\theta)}_{\text{log prior}} - \cancel{\log p(\mathcal{D})}$$

As data goes to infinity, MAP → ML

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} \sum_{n=1}^{N} \log p(x_n|\theta) + \log p(\theta)$$

Infinite data limit: ML = MAP = Bayes'

$$p(x_*|\mathcal{D}) = \int p(x_*|\theta)p(\theta|\mathcal{D})\,\mathrm{d}\theta \overset{N\to\infty}{=} \int p(x_*|\theta)\delta(\theta - \theta_{\mathrm{ML}})\,\mathrm{d}\theta = \int p(x_*|\theta)\delta(\theta - \theta_{\mathrm{MAP}})\,\mathrm{d}\theta$$

1 The word "conjugate" comes from conjugal, meaning the relationship of a married couple

Conjugacy

SCIENTISTS FOR PALESTINE

When the posterior and prior have the same form, we call it *conjugacy*[1]*.* The *exponential family* admits conjugate pairs:

Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})^N} \cdot \exp\left\{\boldsymbol{\theta}^\top \sum_{n=1}^N \mathbf{t}(x_n)\right\} \cdot \prod_{n=1}^N b(x_n)$$

Prior

$$p(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\tau}, \nu)} \cdot \frac{1}{Z(\boldsymbol{\theta})^\nu} \cdot \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\tau}\right\}$$

$$\begin{aligned} p(\boldsymbol{\theta}|\mathcal{D}) &\propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= \underbrace{\frac{1}{Z(\boldsymbol{\theta})^N} \exp\left\{\boldsymbol{\theta}^\top \sum_{n=1}^N \mathbf{t}(x_n)\right\} \left[\prod_{n=1}^N b(x_n)\right]}_{\text{likelihood}} \cdot \underbrace{\frac{1}{Z(\boldsymbol{\tau}, \nu)} \frac{1}{Z(\boldsymbol{\theta})^\nu} \exp\left\{\boldsymbol{\theta}^\top \boldsymbol{\tau}\right\}}_{\text{prior}} \\ &\propto \frac{1}{Z(\boldsymbol{\theta})^{N+\nu}} \exp\left\{\boldsymbol{\theta}^\top \left(\boldsymbol{\tau} + \sum_{n=1}^N \mathbf{t}(x_n)\right)\right\} \end{aligned}$$

Drop terms not containing θ

Normalisation is easy: just compare with prior

$$\nu \to N + \nu \qquad \boldsymbol{\tau} \to \boldsymbol{\tau} + \sum_{n=1}^N \mathbf{t}(x_n) \qquad Z(\boldsymbol{\tau}, \nu) \to Z\left(\boldsymbol{\tau} + \sum_{n=1}^N \mathbf{t}(x_n), N + \nu\right)$$

Lightning fast computation: $O(N)$

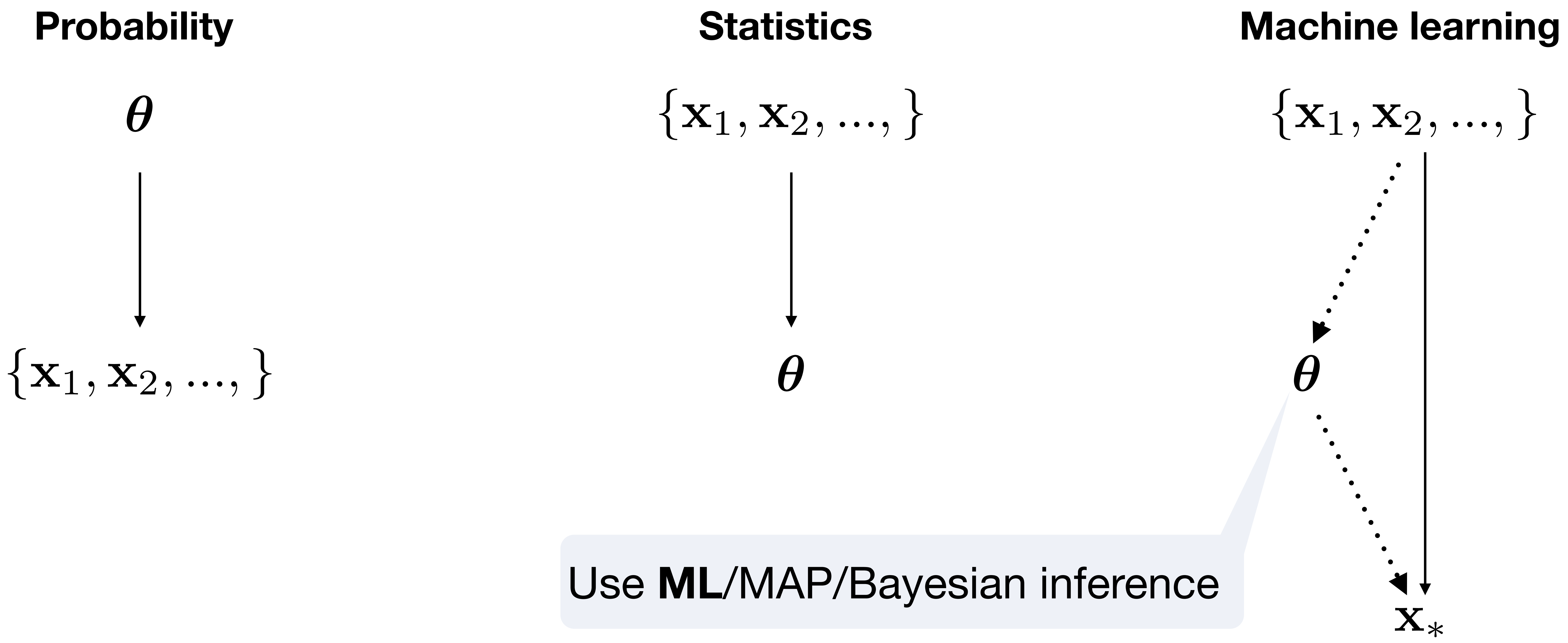
1 The word “conjugate” comes from conjugal, meaning the relationship of a married couple

Probabilistic models

We are mostly concerned with models which look like

$$p(\mathbf{x} \mid \boldsymbol{\theta})$$

In many cases x refers to an *observation* and refers to a set of *parameters*.



*Sometimes we refer to $\{p(\mathbf{x}|\theta)\}_{\theta\in\Theta}$ as a model, other times we refer to $p(\mathbf{x}|\theta)$ for a single θ as the model

Model Comparison

Model comparison

SCIENTISTS FOR PALESTINE

Say I have some data, how do I pick a likelihood and a prior, aka models? Pick a few different *models*, and then find the posterior distribution over the models given the data.

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)$$

Typically, we just want **one** model → MAP inference

Furthermore, the *model prior* is usually flat → MAP = ML

$$\arg\max_{\mathcal{M}_i} p(\mathcal{D}|\mathcal{M}_i) = \arg\max_{\mathcal{M}_i} \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{M}_i)\,\mathrm{d}\boldsymbol{\theta}$$

But hang on, this is just the marginal likelihood/evidence!

$$p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M}_i) = \frac{p(\mathcal{D}|\boldsymbol{\theta},\mathcal{M}_i)p(\boldsymbol{\theta}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}$$

Best model has highest evidence!

Example: Laplace versus Gauss

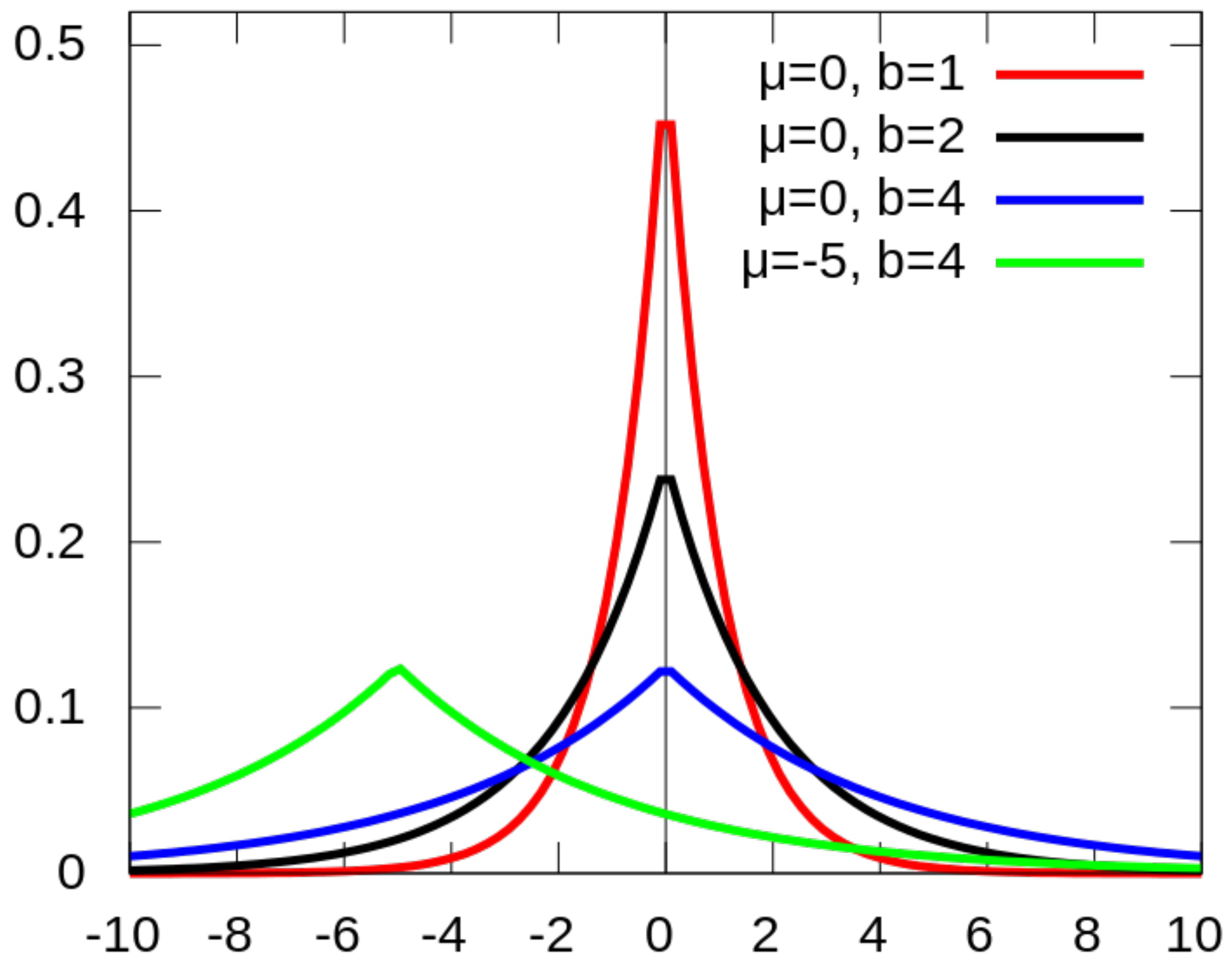
We have zero mean and unit variance data $\{x_1, \ldots, x_N\}$.

Laplacian

$$p(x|\mathcal{M}_1) = \frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|x|\right\}$$

$$\log p(\mathcal{D}|\mathcal{M}_1) = \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|x_n|\right\}\right)$$

$$= N \log \frac{1}{\sqrt{2}} - \sqrt{2} \sum_{i=1}^{N} |x_n|$$

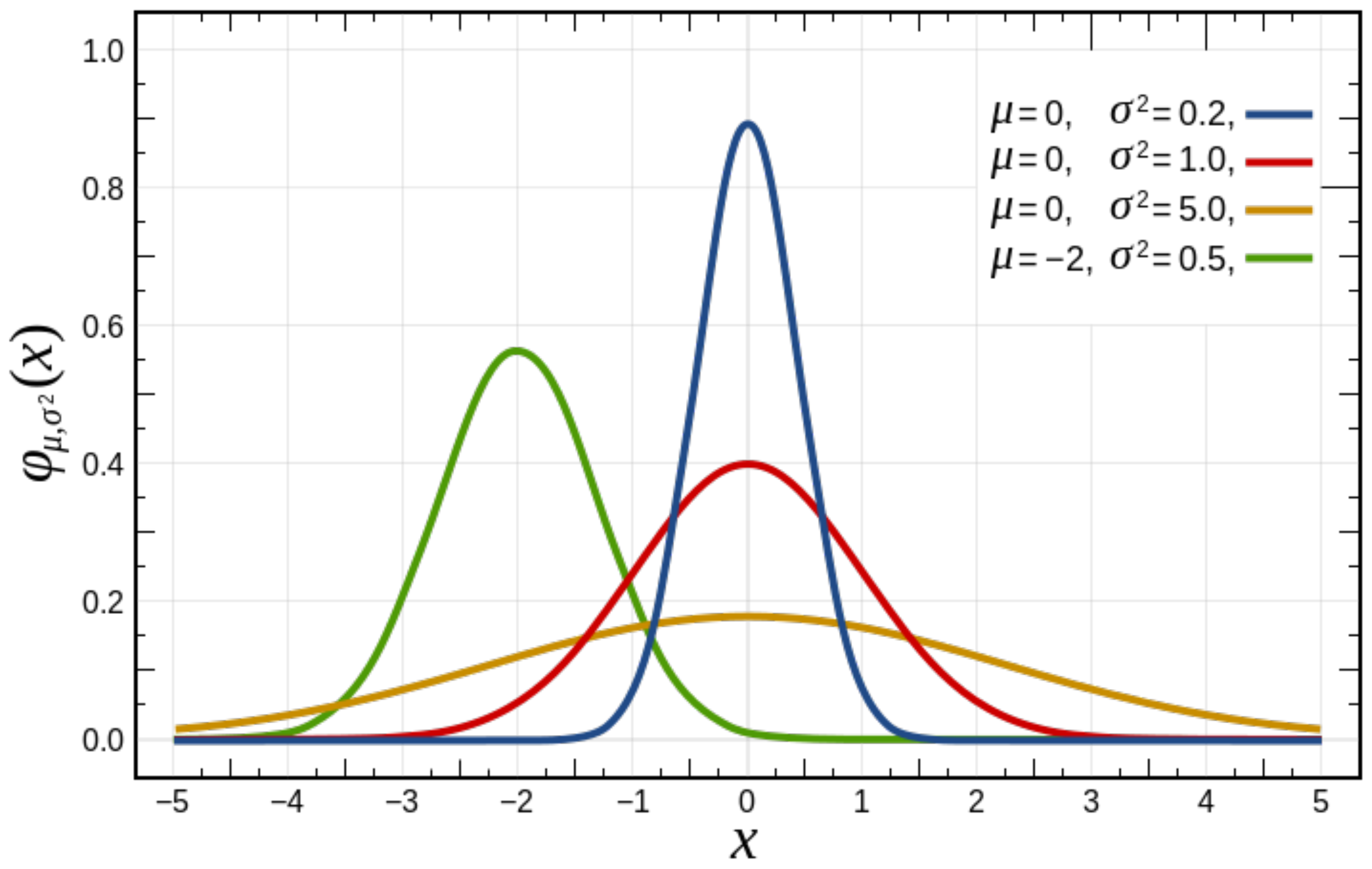


Gaussian

$$p(x|\mathcal{M}_2) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$\log p(\mathcal{D}|\mathcal{M}_2) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x_i^2}{2}\right\}\right)$$

$$= N \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \sum_{i=1}^{N} x_i^2$$

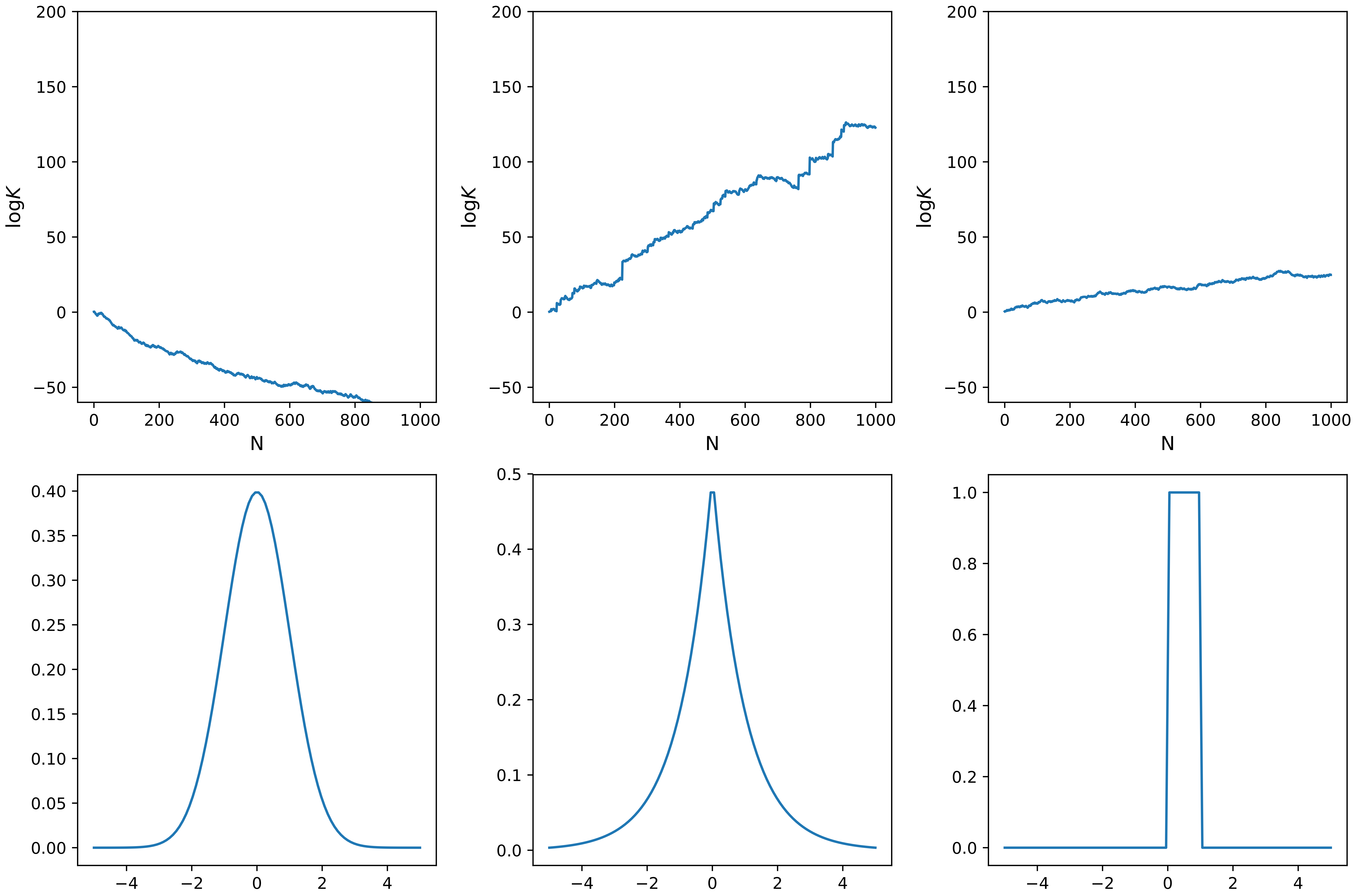


Example: Laplace versus Gauss

Log Bayes' factor

$$\log K = \log \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)}$$

Pierre-Simon Laplace



Carl Friedrich Gauss

Type-II Maximum likelihood

SCIENTISTS FOR PALESTINE

Can optimise model *hyperparameters* too

$$p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\tau}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\tau})p(\boldsymbol{\theta}|\boldsymbol{\tau})}{p(\mathcal{D}|\boldsymbol{\tau})}.$$

$$\boldsymbol{\tau}^* = \arg\max_{\boldsymbol{\tau}} p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{\tau})$$

This goes by the name of *Type-II maximum likelihood*, *empirical Bayes*, or the *evidence approximation*

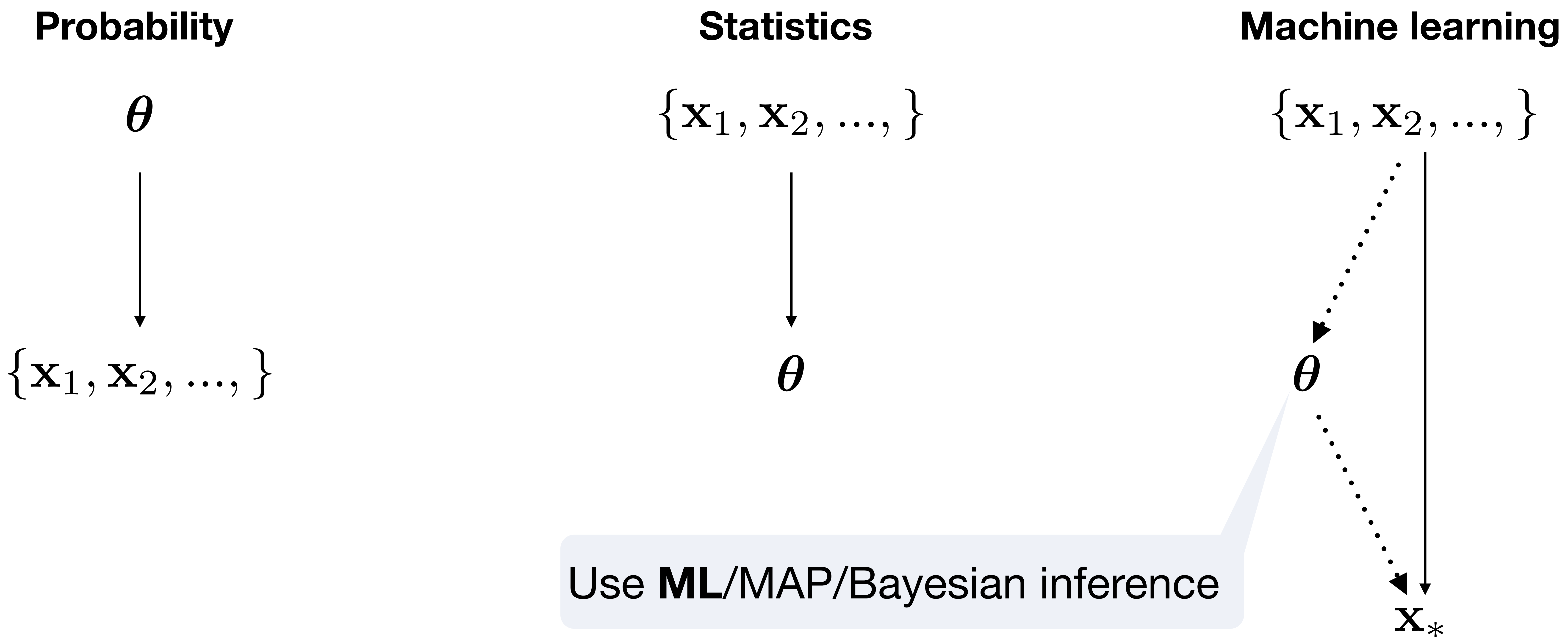
It can be difficult to compute τ^* in closed-form and the optimization landscape is typically highly multimodal. We will see an example of this in linear regression

Recap

We are mostly concerned with models which look like

$$p(\mathbf{x} \mid \boldsymbol{\theta})$$

In many cases x refers to an *observation* and refers to a set of *parameters*.



*Sometimes we refer to $\{p(\mathbf{x}|\theta)\}_{\theta\in\Theta}$ as a model, other times we refer to $p(\mathbf{x}|\theta)$ for a single θ as the model

Summary

SCIENTISTS FOR PALESTINE

This lecture: Machine Learning Basics

- What is Machine Learning?
- Probability Theory
- Probabilistic models
  - Forward models
  - Independence
- Statistical Inference
  - Maximum Likelihood
  - Bayesian Inference
- Modeling paradigms
  - Prediction
- Model comparison

Next lecture: Deep Learning Basics