Syllabus for IB Data Science

Using a probability model to describe data

Models that depend on linear combinations of features → Parameter interpretation and identifiability

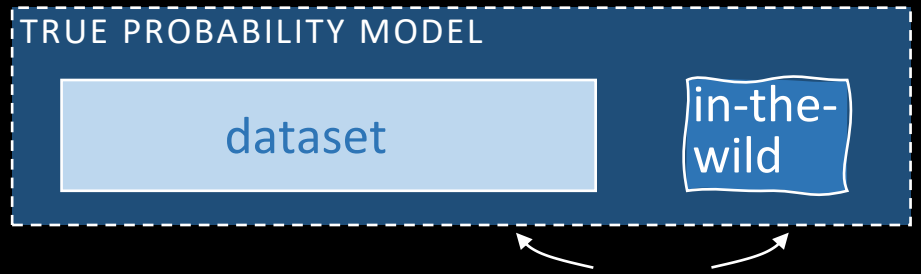
Fitting a model to the data

Maximum likelihood estimation → Fitting via least squares

Does the model generalize?

GENERALIZATION

I trained my model on a dataset. Will it work well on in-the-wild data?

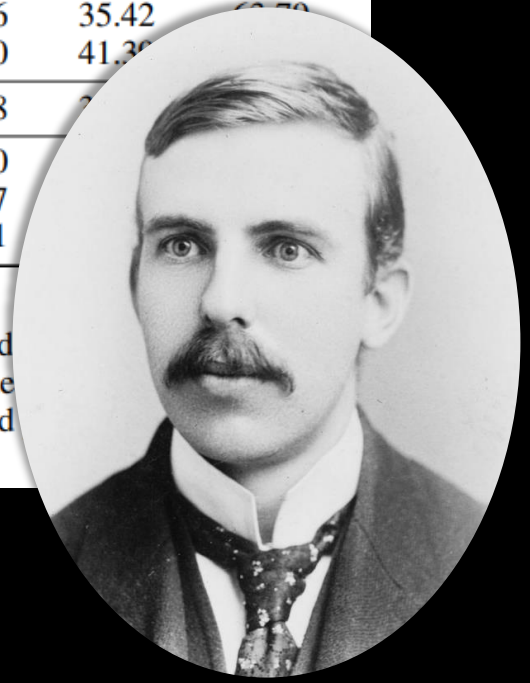


We'll assume that dataset & in-the-wild satisfy the same Laws of Nature.
(Otherwise it's impossible to say anything useful.)

Table 2: Results on HotpotQA distractor (dev). (+hyperlink) means usage of extra hyperlink data in Wikipedia. Models beginning with "–" are ablation studies without the corresponding design.

Model	Ans EM	Ans F_1	Sup EM	Sup F_1	Joint EM	Joint F_1
Baseline [53]	45.60	59.02	20.32	64.49	10.83	40.16
DecompRC [29]	55.20	69.63	N/A	N/A	N/A	N/A
QFE [30]	53.86	68.06	57.75	84.49	34.63	59.61
DFGN [36]	56.31	69.69	51.50	81.62	33.62	59.82
SAE [45]	60.36	73.58	56.93	84.63	38.81	64.96
SAE-large	66.92	79.62	61.53	86.86	45.36	**71.45**
HGN [14] (+hyperlink)	66.07	79.36	60.33	87.33	43.57	71.03
HGN-large (+hyperlink)	69.22	82.19	62.76	88.47	47.11	**74.21**
BERT (sliding window) variants						
BERT Plus	55.84	69.76	42.88	80.74	27.13	58.23
LQR-net + BERT	57.20	70.66	50.20	82.42	31.18	59.99
GRN + BERT	55.12	68.98	52.55	84.06	32.88	60.31
EPS + BERT	60.13	73.31	52.55	83.20	35.40	63.41
LQR-net 2 + BERT	60.20	73.78	56.21	84.09	36.56	63.68
P-BERT	61.18	74.16	51.38	82.76	35.42	
EPS + BERT(large)	63.29	76.36	58.25	85.60	41.3	
CogLTX	65.09	78.72	56.15	85.78		
– multi-step reasoning	62.00	75.39	51.74	83.10		
– rehearsal & decay	61.44	74.99	7.74	47.37		
– train-test matching	63.20	77.21	52.57	84.21		

Results. Table 2 shows that CogLTX outperforms most of previous method solutions on the leaderboard.[4] These solutions basically follow the frame results from sliding windows by extra neural networks, leading to bounded to insufficient interaction across paragraphs.



Most ML papers don't state an inductive claim.

Perhaps the authors haven't thought hard enough to be able to state one?

Perhaps they prefer to leave you, the reader, to make the inference?

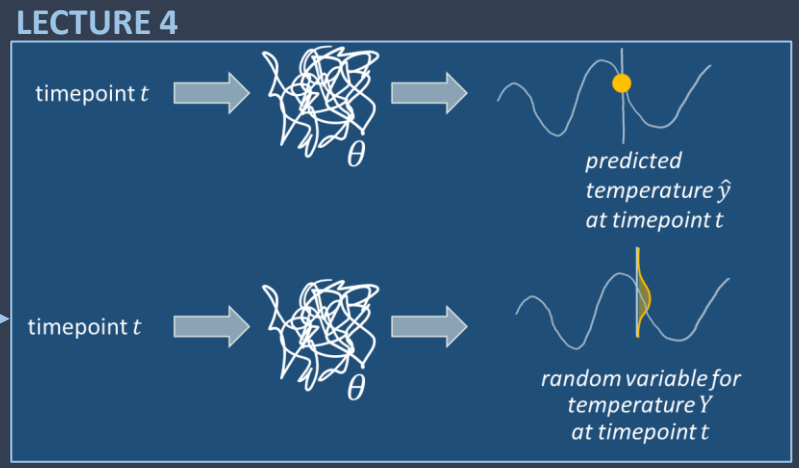
"All science is either physics or stamp-collecting."

Ernest Rutherford (1871–1937)

FITTING

How do we measure how well a model “works”?

- prediction accuracy, or mean square error [MLRD]
- log likelihood is a more general metric



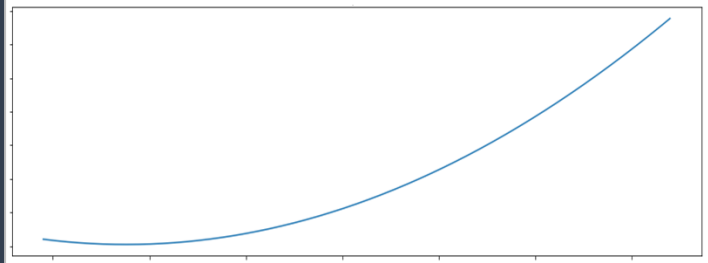
So, let’s fit our model’s parameters by maximimum likelihood estimation.
Then it’ll surely work well on in-the-wild data! 🤪

FITTING AND MODEL CHOICE

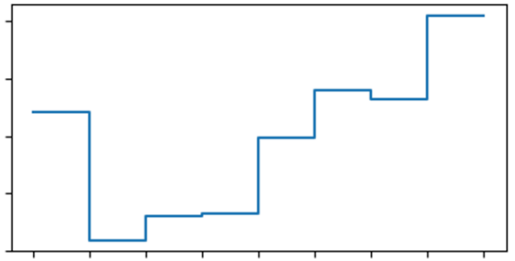
LECTURE 8

How should we choose between models?

$\text{temp} \approx \alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t + \delta t^2$



$\text{temp} \approx \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma_{\text{decade}}$



Choosing between models can be seen as estimating a discrete parameter …

1. Define a full model that has a "switch" parameter m specifying which submodel to use

```
def ry(m,θ₁,θ₂):
  if m == 1:
    return ry1(θ₁)
  else: # m==2
    return ry2(θ₂)
```

$$\Pr_Y(y; m, \theta_1, \theta_2) = \begin{cases} \Pr_Y^1(y; \theta_1) & \text{if } m = 1 \\ \Pr_Y^2(y; \theta_2) & \text{if } m = 2 \end{cases}$$

2. Then we can estimate it using maximum likelihood estimation: $\max_{m,\theta_1,\theta_2} \Pr(\text{data}; m, \theta_1, \theta_2)$

$$= \max\left\{\max_{\theta_1} \Pr^1(\text{data}; \theta_1), \max_{\theta_2} \Pr^2(\text{data}; \theta_2)\right\}$$

The mle for m is the model with the highest likelihood, $\hat{m} = \arg\max_m \Pr(\text{data}; m)$

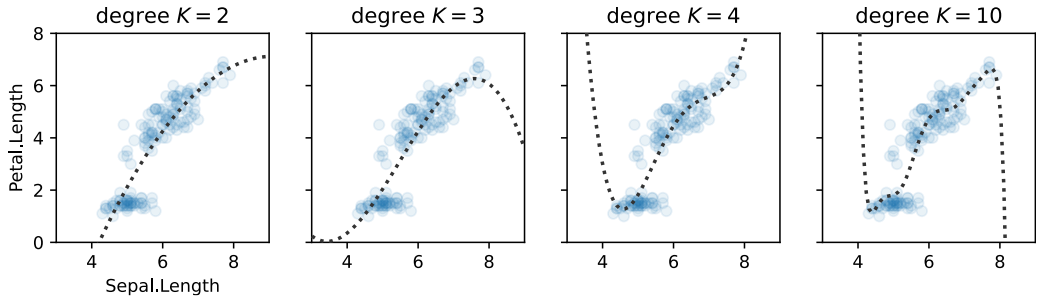
MODEL COMPLEXITY

When we're choosing between models of different complexity, the more complex model will always score better.

LECTURE 5

What degree polynomial?

PL ≈ polynomial in SL of degree K



LECTURE 10

What model for the number of heads?

$X \sim \text{Bin}(n, ½)$

versus

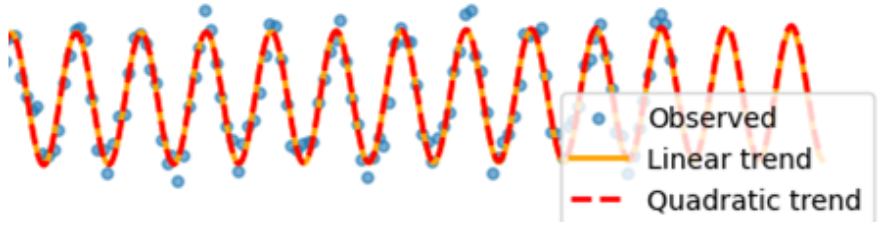
$X \sim \text{Bin}(n, \theta)$ for some $\theta \in [0,1]$

LECTURE 8

Linear or quadratic trend term?

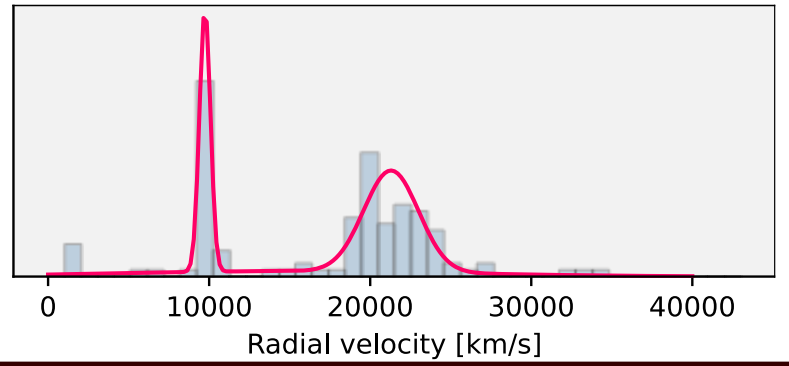
$\texttt{temp} \approx \alpha + \beta_1 \sin(2\pi\texttt{t}) + \beta_2 \cos(2\pi\texttt{t}) + \gamma\texttt{t}$

$\texttt{temp} \approx \alpha + \beta_1 \sin(2\pi\texttt{t}) + \beta_2 \cos(2\pi\texttt{t}) + \gamma\texttt{t} + \delta\texttt{t}^2$



LECTURE 3

How many clusters, in a Gaussian mixture model?

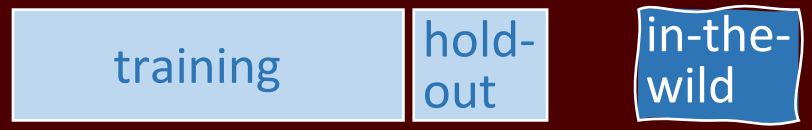


OVERFITTING AND HOLDOUT EVALUATION

A too-complex model will typically fit the dataset well, but generalize poorly to in-the-wild data.

How can we detect and avoid too-complex models?

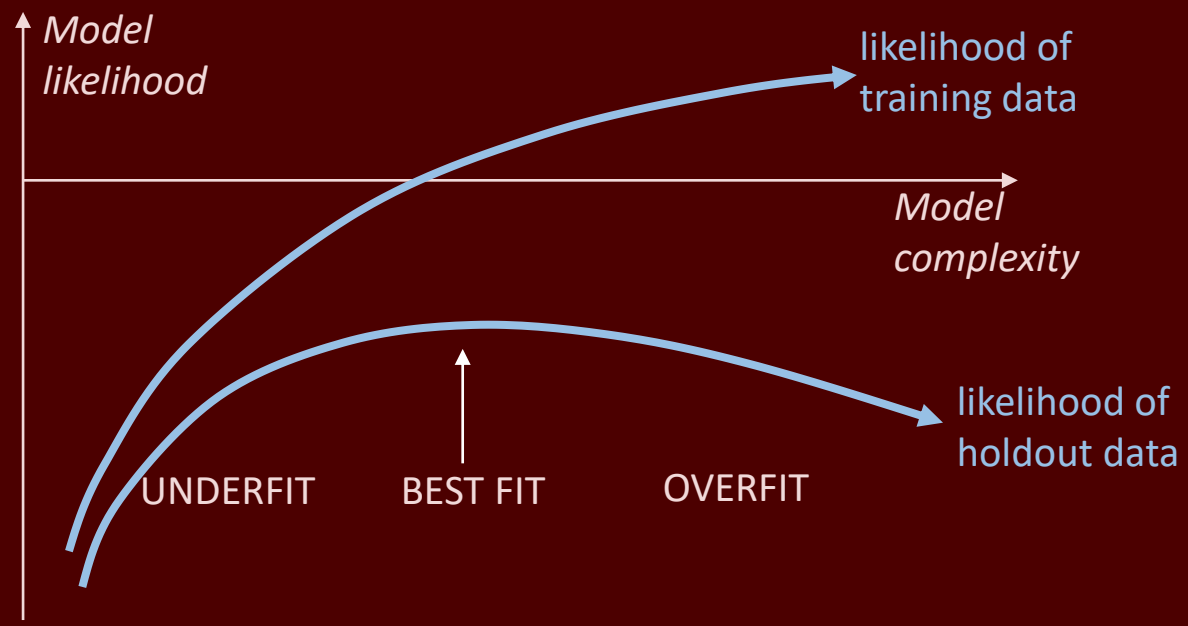
We can approximate in-the-wild performance by holdout performance.



This suggests we should simply choose the model that maximizes the holdout likelihood

With this approach, there's an inner loop where we fit a model to the training data, and an outer loop where we choose between fitted models.

QUESTION. Can we do it all in a single loop?

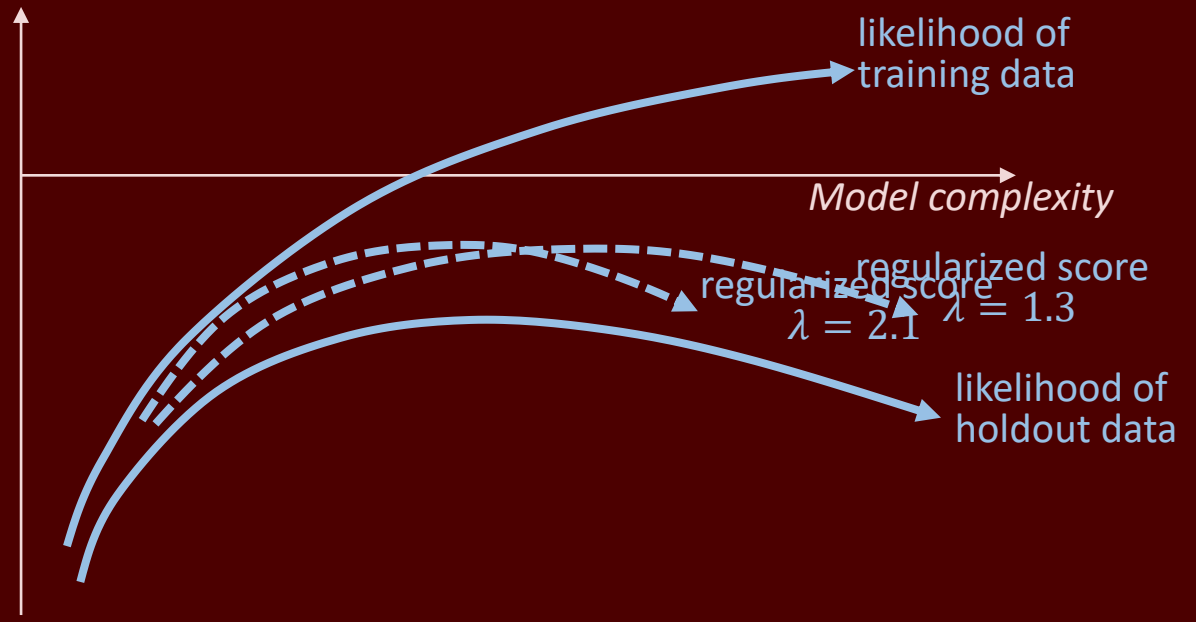


REGULARIZATION

Some models have a natural continuous measure of complexity.
In such cases, we can maximize the *regularized* score, which penalizes complex models:

$$\max_{\theta} \{\log \Pr(\text{data}; \theta) - \lambda \, \text{complexity}(\theta)\}$$

- Gradient descent will choose a tradeoff between fit and complexity.
  We're baking "too much complexity is bad" into the training itself.
- We control the relative importance by specifying $\lambda > 0$.
- (We'll still want to tune λ to maximize holdout likelihood,
  but this maximization is pretty forgiving.)



BAYESIANIST MODEL CHOICE

Suppose we want to choose between several models.
We're uncertain which is the correct model.

BAYESIANIST EPISTEMOLOGY

Whenever there's an unknown parameter, you should express your uncertainty about it by treating it as a random variable.

Q. What are we uncertain about?
Depends on the problem

Q. How do we represent unknowns?
Answer: As random variables, with a prior

Q. What do we report?
Answer: The posterior distribution of the quantity of interest

Q. How do we find this?
Answer: Using Bayes's rule

Exercise 8.3.3 (Bayesian classification)

There are two types of expense claims, legitimate and fraudulent.
The legitimate claim sizes are $\sim \text{Exp}(\lambda_L)$ and the fraudulent ones are $\sim \text{Exp}(\lambda_F)$ where $\lambda_L = 0.1$ and $\lambda_F = 0.02$.
In my prior experience, 99% of claims I've seen are legitimate.
A new claim comes in, for an amount £x. Is it likely to be fraudulent?

What are we uncertain about?

whether the new claim is fraudulent

How do we represent uncertainty?

$$\text{Let } M = \begin{cases} \ell & \text{if the new claim is legitimate} \\ f & \text{if it's fraudulent} \end{cases}$$

What is my prior?

$$\Pr_M(\ell) = 0.99 \quad \text{and} \quad \Pr_M(f) = 0.01$$

What is the posterior I want to report?

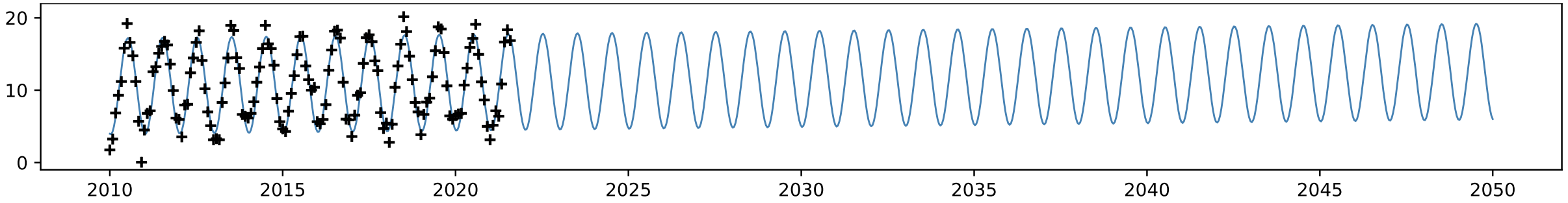
$$\Pr_M(f \mid x)$$

i.e. $\mathbb{P}(M = f \mid x)$

How should we choose between two models?

Modeller 1: $\text{Temp} \sim \alpha + \beta \sin\big(2\pi(t+\phi)\big) + \gamma(t-2000) + N(0,\sigma^2)$

Modeller 2: $\text{Temp} \sim \alpha' + \beta' \sin\big(2\pi(t+\phi')\big) + N(0,\sigma'^2)$



What are we uncertain about?

Which model is correct (and also all nine unknown parameters)

How do we represent uncertainty? With random variables.

Introduce a "switch" random variable M saying which model is correct, M=1 or M=2.
Invent a prior for M, and for the other nine parameters.
Pr(data | params) = Pr($\text{temp}_1,\ldots,\text{temp}_n$ | M=m, $\alpha,\beta,\varphi,\gamma,\sigma,\alpha',\beta',\varphi',\sigma'$)

What do I want to report?

First find the posterior distribution of (M,α,...) given the data.
Then report the marginal of M, $\mathbb{P}$(M=1 | data).

TIP: in multi-parameter problems, use Bayes's rule on *all* the unknowns simultaneously

BAYESIANIST MODEL CHOICE

Suppose we want to choose between several models.
We're uncertain which is the correct model.

our prior expresses our preconceptions about in-the-wild data

Introduce a "switch" parameter. Let the correct model be M, with prior $\Pr_M(m)$.

$$\Pr_M(m|\text{data}) = \kappa \Pr_M(m) \Pr(\text{data}|m)$$

We could simply report our posterior for M.
Or we could report a point estimate. The MAP estimator says to pick the m that maximizes the posterior likelihood:

$$\begin{aligned}\hat{m} &= \arg\max_m \Pr_M(m|\text{data}) \\ &= \arg\max_m \{\log \Pr(\text{data}|m) + \log \Pr_M(m)\}\end{aligned}$$

a regularizer, that penalizes models that are inconsistent with our prior beliefs

IB Data Science syllabus

Using a probability model to describe data

How well does my model fit the dataset?

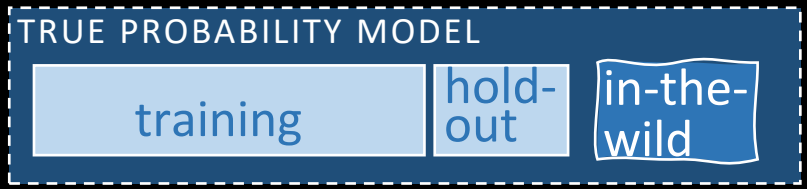
Note: model choice and parameter estimation are the same thing

How well does my model generalize to in-the-wild data?

[EMPIRICIST] We can use holdout evaluation to approximate in-the-wild performance, without making any assumptions (other than that there *is* a common true distribution).

[BAYESIANIST] If we have a prior belief about the true distribution, we should use it. In this case, we don't need holdout evaluation.

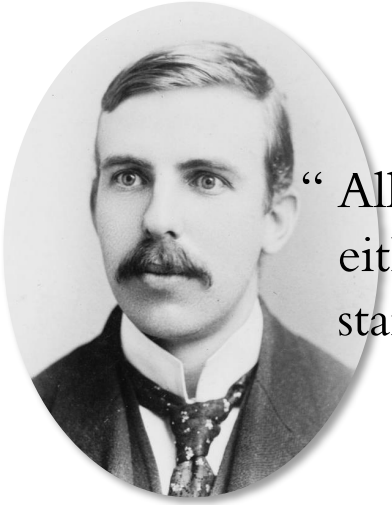
[FREQUENTIST / HYPOTHESIS TESTING] …



INDUCTION IS NOT DEDUCTION

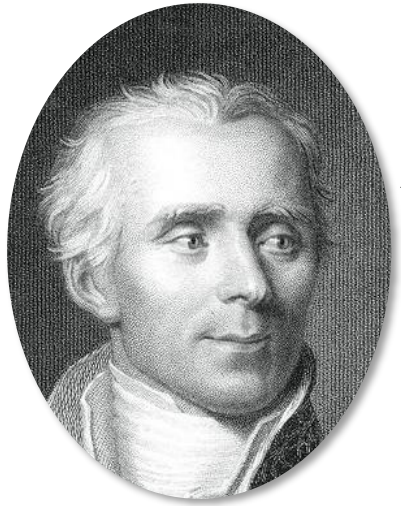
I have data, and I've learned something from it. What can I say about the future?
Philosophers call this *The Problem of Induction*.

It is IMPOSSIBLE to mathematically *deduce* the true probability model from a dataset,
so don't hanker after rigorous proofs for why one school of induction is better than another.



" All science is either physics or stamp-collecting."

Ernest Rutherford
(1871-1937)



What's the chance the sun will rise tomorrow? $\frac{x+1}{n+2}$

Pierre-Simon Laplace
(1749-1827)

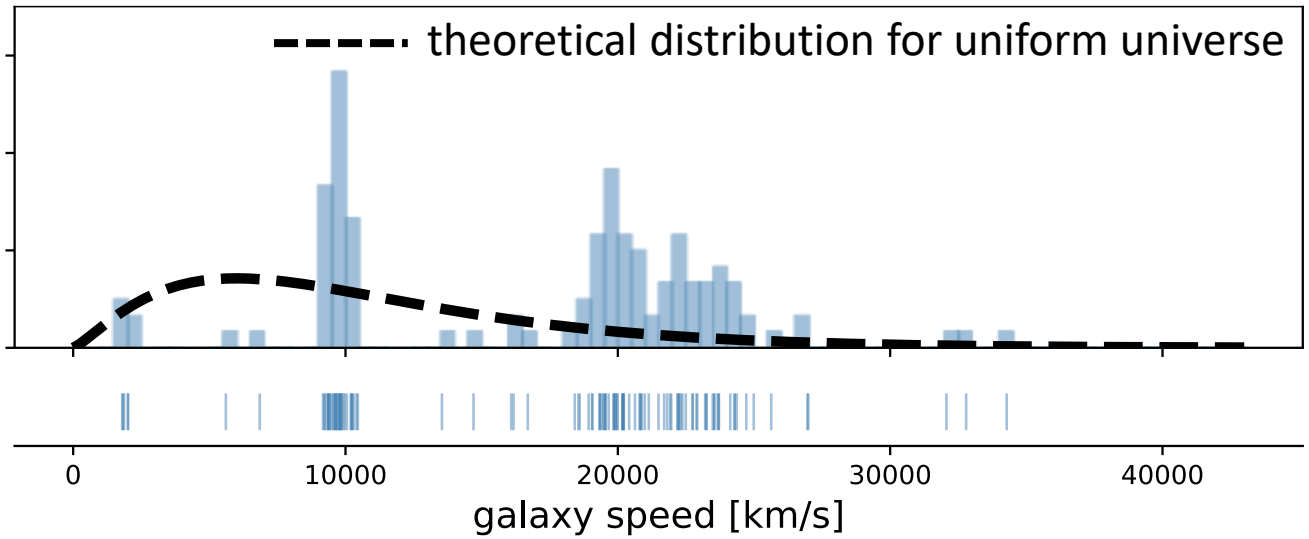
" Induction is the glory of Science and the scandal of Philosophy."

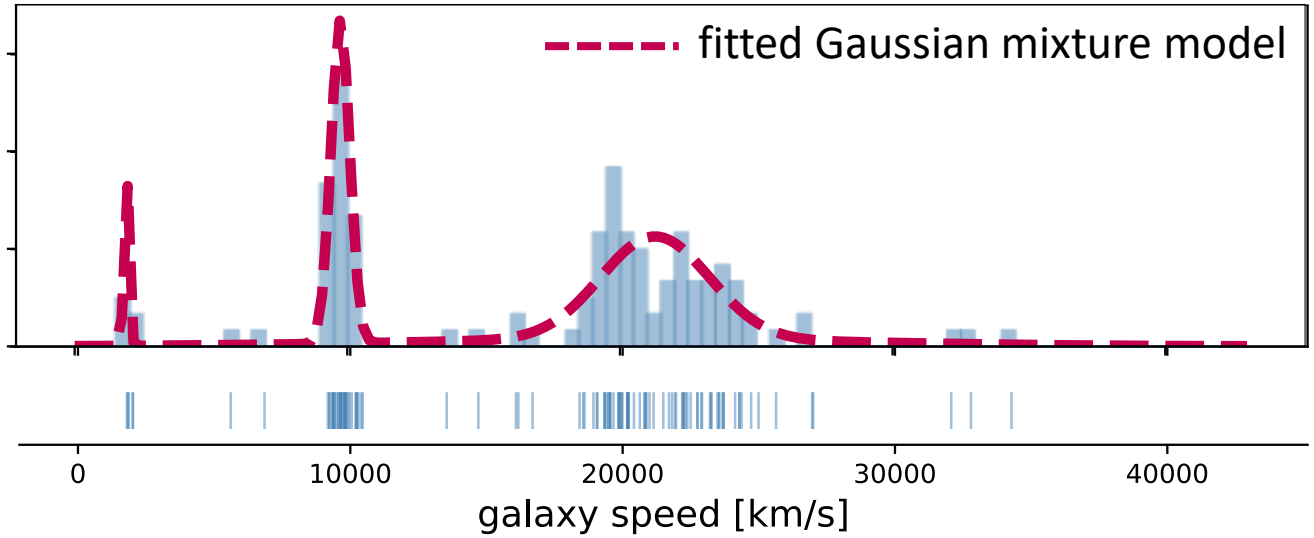
C.D. Broad
1926

IB Data Science syllabus — maths skills

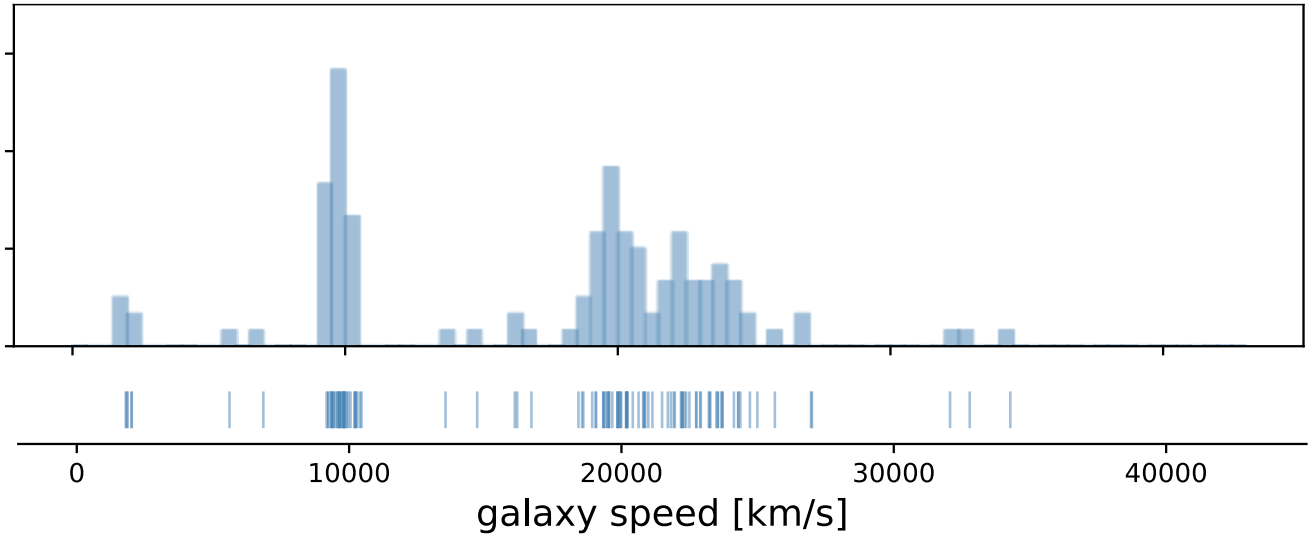
Using a probability model to describe data	▪ familiarity with a range of models ▪ linear models; parameter interpretation
How well does my model fit the dataset?	▪ maximum likelihood estimation ▪ numerical optimization
How well does my model generalize to in-the-wild data?	▪ Bayes's rule ▪ Monte Carlo ▪ bespoke distributions

This chart shows the distribution of the speeds of 120 galaxies, from a survey of the Corona Borealis region.
Postman, Huchra, Geller (1986)



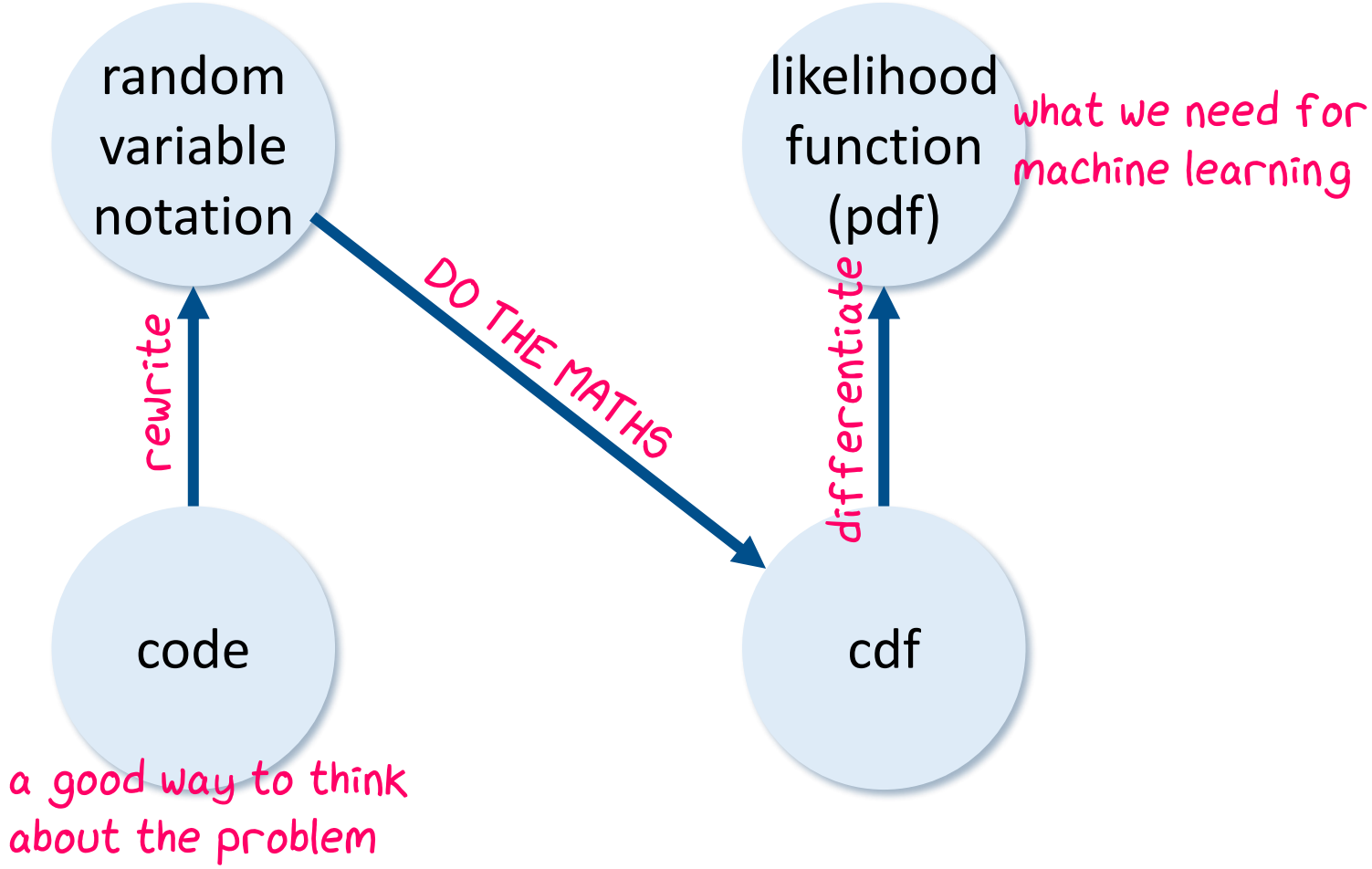

fitted Gaussian mixture model
0
10000
20000
30000
40000
galaxy speed [km/s]

What's the best distribution we can find, to model this dataset?



Bespoke probability distributions part I: from code to likelihood (for continuous random variables)

There are four ways to specify a distribution:



Exercise 5.3.2

Find the pdf of the random variable generated by this code:

```
u = np.random.uniform()
x = - np.log(u) / λ        # λ>0
```

STRATEGY

Try to write our probability in terms of simple standard random variable (for which we can look up the cdf)

Break it down so that the random variables are on the left (so we can use the textbook cdf)

Step 1: random variable notation

$U \sim U[0,1]$

$X = -\frac{1}{\lambda} \log U$

Step 2: X is a continuous r.v., so find its cdf

cdf for X is $\mathbb{P}(X \le x) = \mathbb{P}(-\frac{1}{\lambda} \log U \le x)$ by defn of X

$= \mathbb{P}(U \ge e^{-\lambda x})$

$= 1 - \mathbb{P}(U < e^{-\lambda x})$

$= 1 - e^{-\lambda x}$ assuming $x \ge 0$, given $\lambda > 0$.

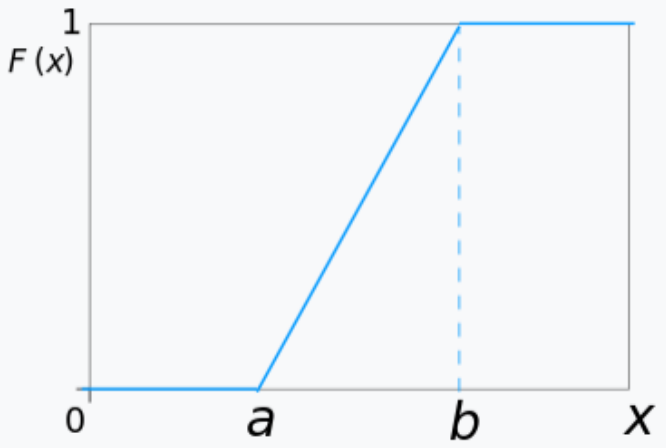
$$\mathbb{P}(U < e^{-\lambda x}) = \begin{cases} 0 & \text{if } e^{-\lambda x} < 0 \\ \frac{e^{-\lambda x} - 0}{1-0} & \text{if } \dots \\ 1 & \text{if } e^{-\lambda x} > 1 \end{cases}$$

Step 3: differentiate cdf to get pdf

$\frac{d}{dx}(1 - e^{-\lambda x}) = \lambda e^{-\lambda x}$.

Wikipedia: Uniform distribution

Cumulative distribution function



Notation	$\mathcal{U}_{[a,b]}$
Parameters	$-\infty < a < b < \infty$
Support	$x \in [a, b]$
PDF	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
CDF	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$