Computer Science

IB Foundations of Data Science

MRes AI4ER

Data Science part 1

Lecturer

Dr Damon Wischik

ARTWORK: TAMAR COHEN, ANDREW J E
ON A PAGE FROM A HIGH SCHOOL YEAR

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE 11 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

WHAT TO READ NEXT

Still the Sexiest Profe

Regulating the internet giants

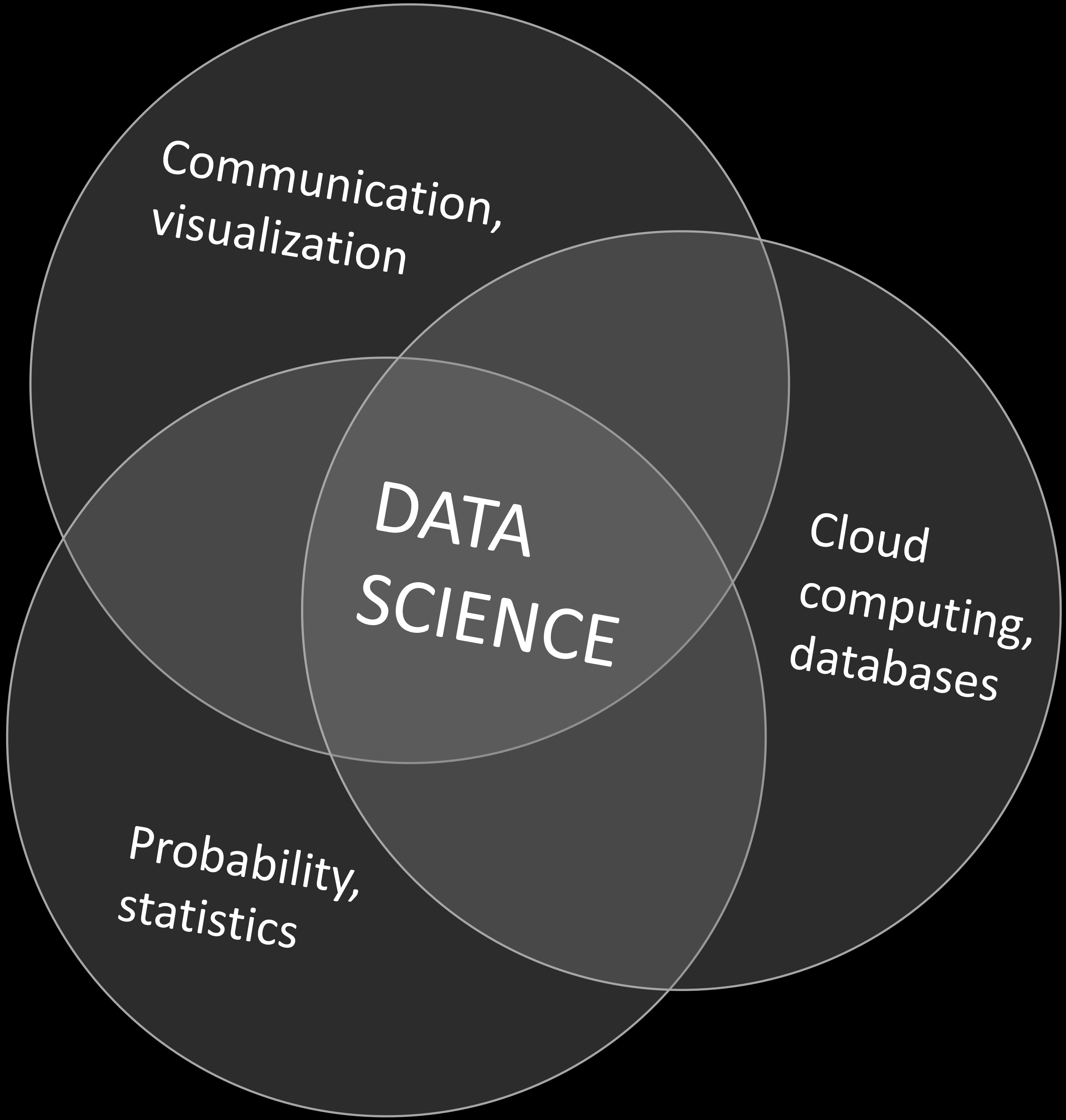
The world's most valuable resource is no longer oil, but data

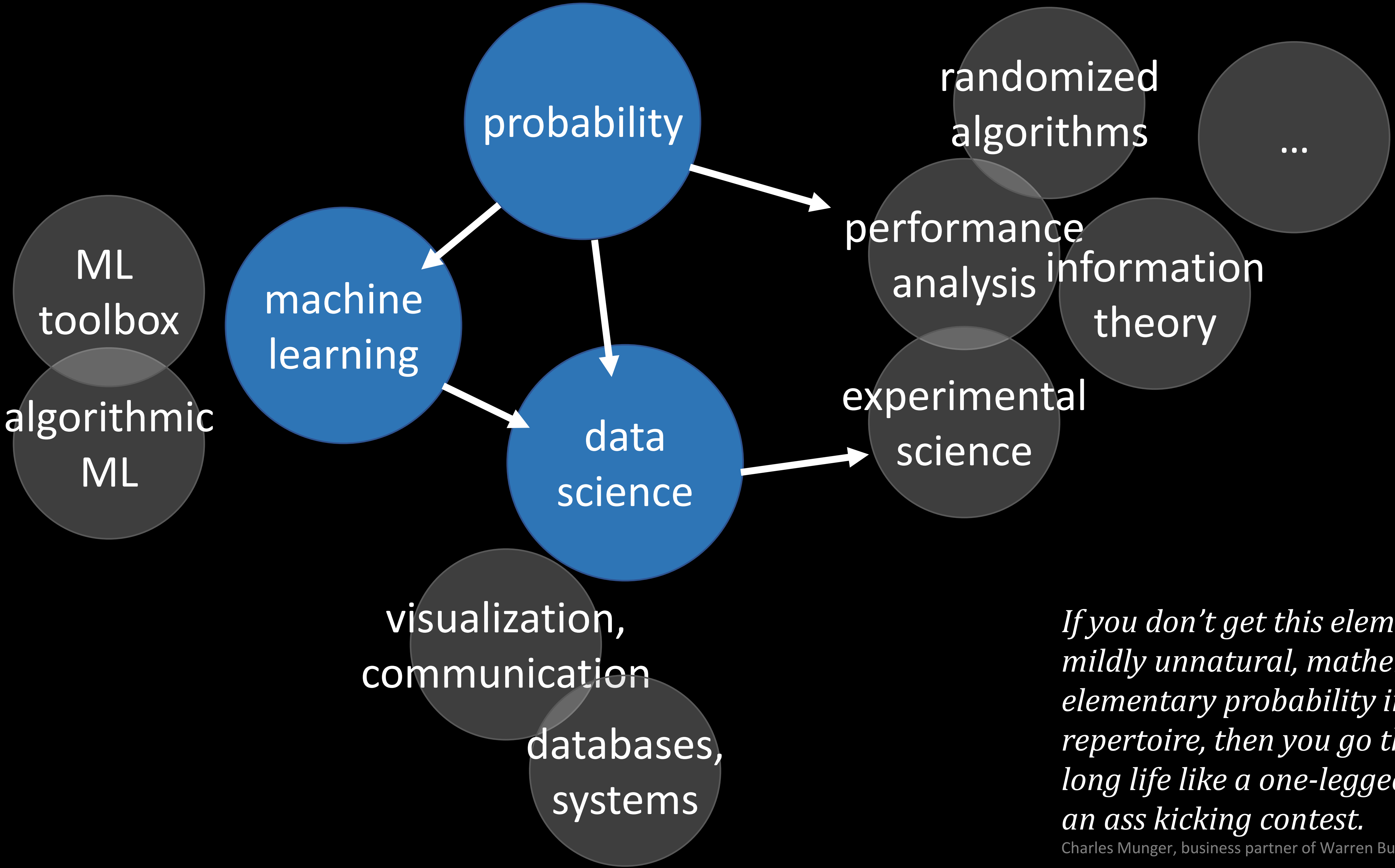
The data economy demands a new approach to antitrust rules

Print edition | Leaders >

May 6th 2017

A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are


Communication, visualization
DATA SCIENCE
Cloud computing, databases
Probability, statistics



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.

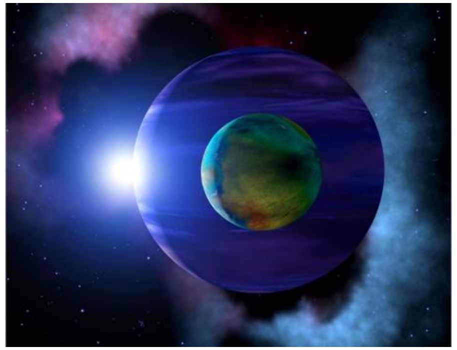
Charles Munger, business partner of Warren Buffett.

Signal may be from first 'exomoon'

By Paul Rincon
Science editor, BBC News website

27 July 2017 | Science & Environment

Share



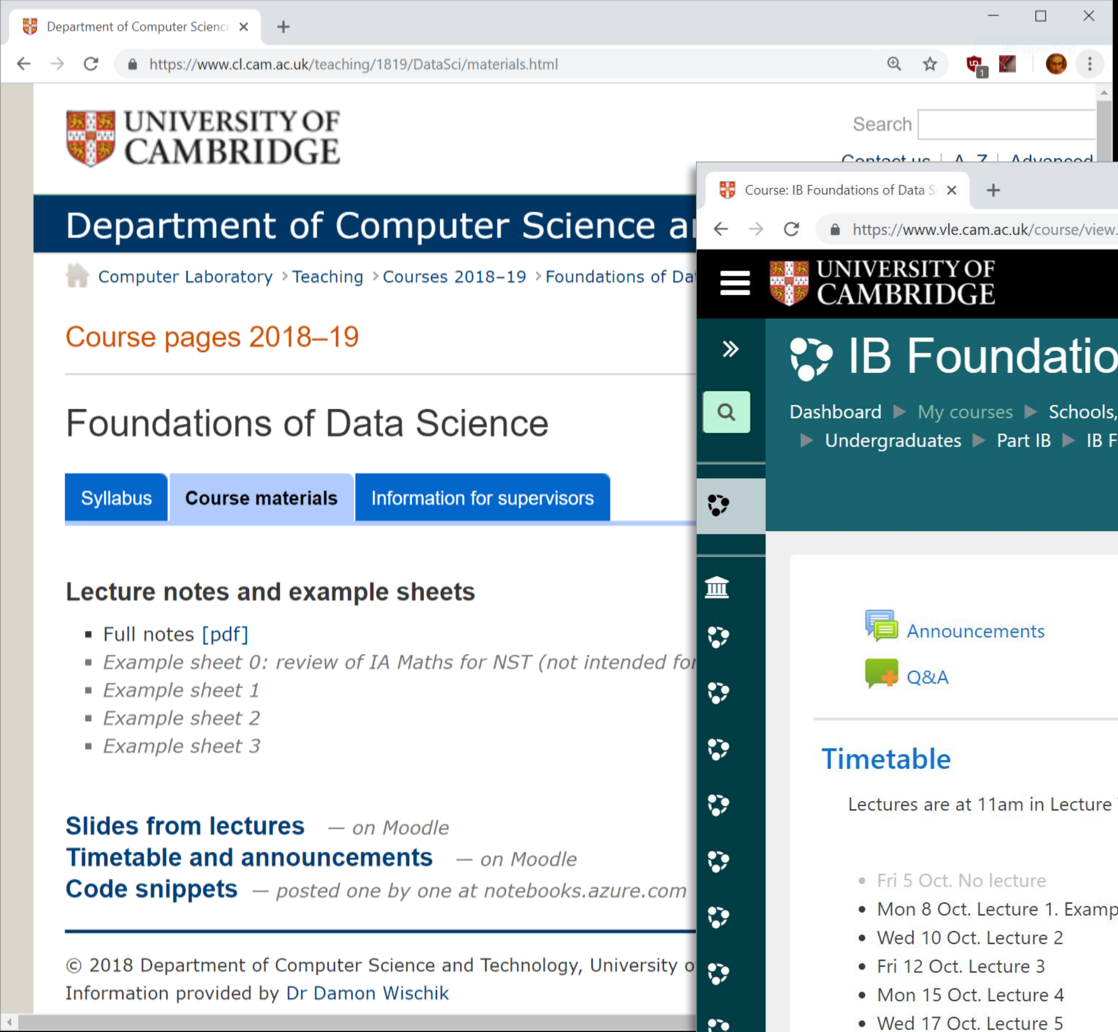
The work by Dr Kipping, his Columbia colleague Alex Teachey and citizen scientist Allan R Schmitt, assigns a confidence level of four sigma to the signal from the distant planetary system. The confidence level describes how unlikely it is that an experimental result is simply down to chance. If you express it in terms of tossing a coin, it's equivalent to tossing 15 heads in row.

But Dr Kipping said this is not the best way to gauge the potential detection.

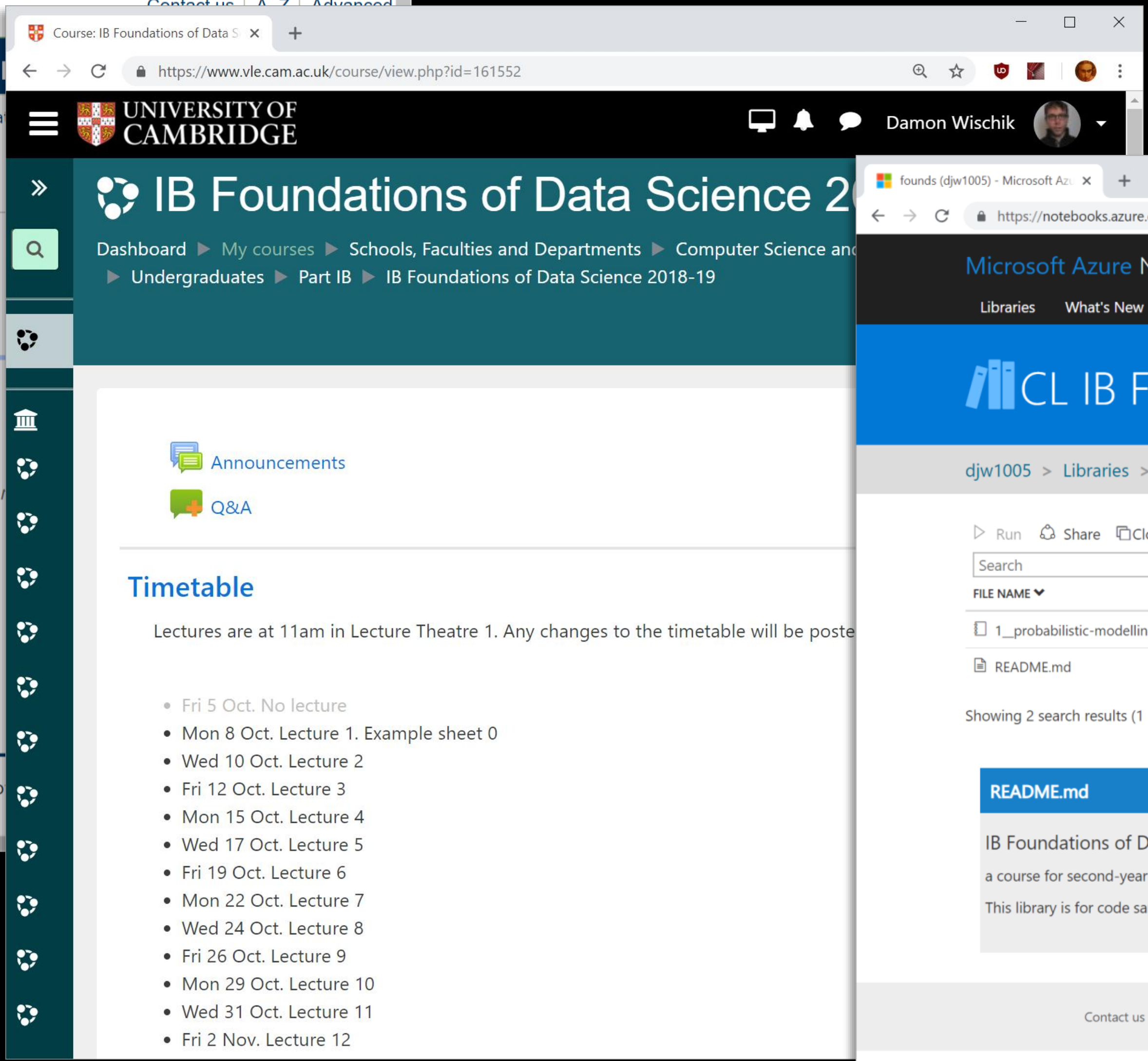
He told BBC News: "We're excited about it... statistically, formally, it's a very high probability. But do we really trust the statistics? That's something unquantifiable. Until we get the measurements from Hubble, it may as well be 50-50 in my mind."

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.

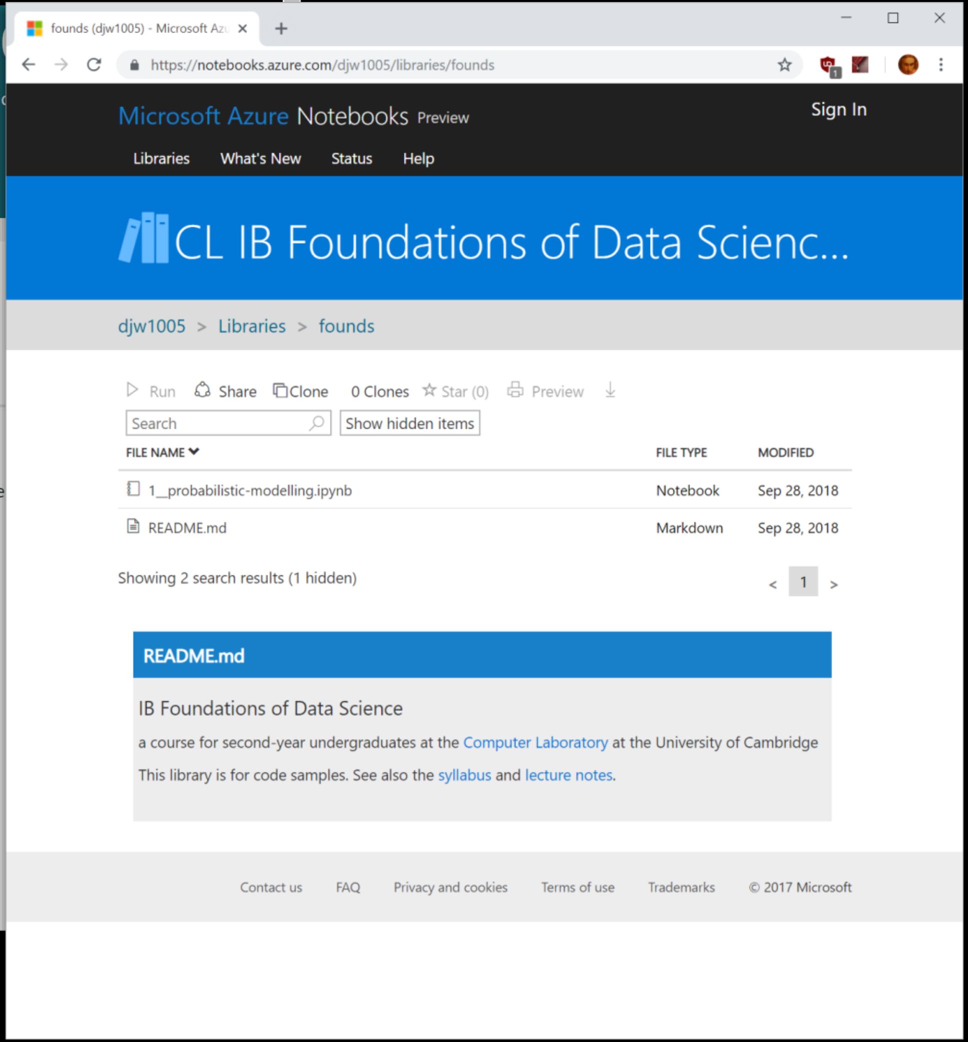
Charles Munger, business partner of Warren Buffett.



CONTENT
notes,
example sheets,
practical walkthroughs



EPHEMERA
timetable, slides,
announcements, Q&A

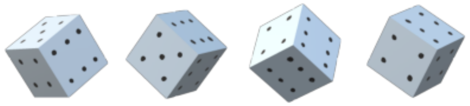


code snippets from lectures

Notes, handed out on Monday

IB Foundations of Data Science

Damon Wischik, Computer Science, Cambridge University



Contents

What is data science? iv

I Learning with probability models 1

1 Specifying and fitting models 2
- 1.1 Maximum likelihood estimation . . . 2
- 1.2 Numerical optimization . . . 5
- 1.3 Random variables in code . . . 7
- 1.4 Random variables in maths . . . 9
  - 1.4.1 Standard random variables . . . 10
  - 1.4.2 Specifying numerical random variables . . . 11
  - 1.4.3 Independence . . . 12
- 1.5 Learning generative models . . . 13
- 1.6 Supervised learning . . . 16
- 1.7 Supervised learning and prediction loss * . . . 20

2 Feature spaces / linear regression 21
- 2.1 Fitting a linear model . . . 22
- 2.2 Features . . . 24
  - 2.2.1 One-hot coding . . . 24
  - 2.2.2 Non-linear response . . . 25
  - 2.2.3 Periodic patterns . . . 25
  - 2.2.4 Discovering features . . . 26
- 2.3 Linear mathematics . . . 30
- 2.4 Linear regression and least squares . . . 31
- 2.5 Confounded features . . . 32

- Material in the notes is examinable (except sections marked *)
- The notes include a few more examples than we cover in lectures
- I am preparing DRAFT expanded notes; they contain extra material that is not examinable

Example sheet 0

Remembering IA Maths for NST

Foundations of Data Science—DJW—2019/2020

Foundations of Data Science builds on the probability theory you learnt in IA *Maths for the Natural Sciences Tripos.* All of the questions below (apart from the last two) are taken from that course. Please look through and make sure you can still answer them! Solutions will be provided.

For supervisors: it isn't intended that you supervise this example sheet.

Question 1. A card is drawn at random from a pack. Event A is 'the card is an ace', event B is 'the card is a spade', event C is 'the card is either an ace, or a king, or a queen, or a jack, or a 10'. Compute the probability that the card has (i) one of these properties, (ii) all of these properties.

Question 2. A biased die has probabilities p, $2p$, $3p$, $4p$, $5p$, $6p$ of throwing 1, 2, 3, 4, 5, 6 respectively. Find p. What is the probability of throwing an even number?

Question 3. Consider drawing 2 balls out of a bag of 5 balls: 1 red, 2 green, 2 blue. What is the probability of the second ball drawn from the bag being blue given that the first ball was blue if (i) the first ball is replaced, (ii) the first ball is not replaced?

Question 4. Two cards are drawn from a deck of cards. What is the probability of drawing two queens, given that the first card is not replaced?

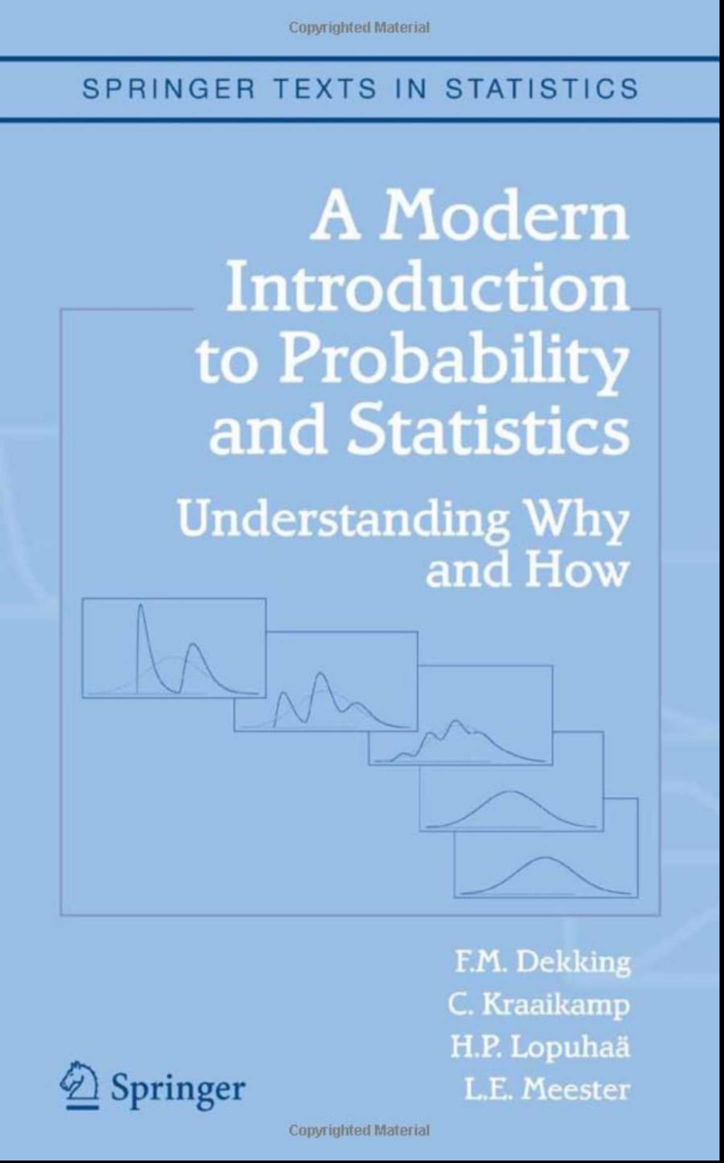
Question 5. A screening test is 99% effective in detecting a certain disease when a person has the disease. The test yields a 'false positive' for 0.5% of healthy persons tested. Suppose 0.2% of the population has the disease. (i) What is the probability that a person whose test is positive has the disease? (ii) What is the probability that a person whose test is negative actually has the disease after all?

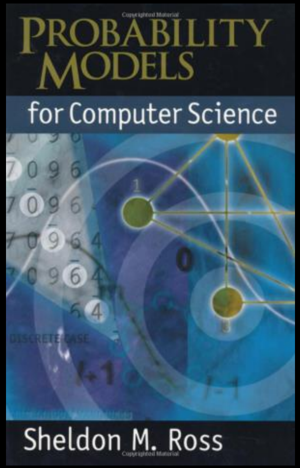
Question 6. What is the probability that in a room of r people at least two have the same

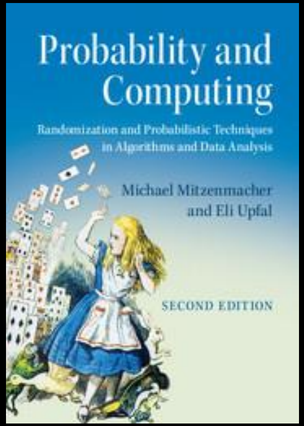
- Example sheet 0
  review of IA Maths,
  not for supervision,
  solutions will be provided

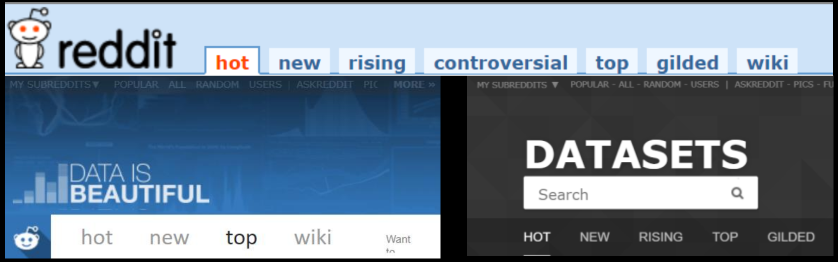
- Example sheets 1, 2, 3
  for supervision

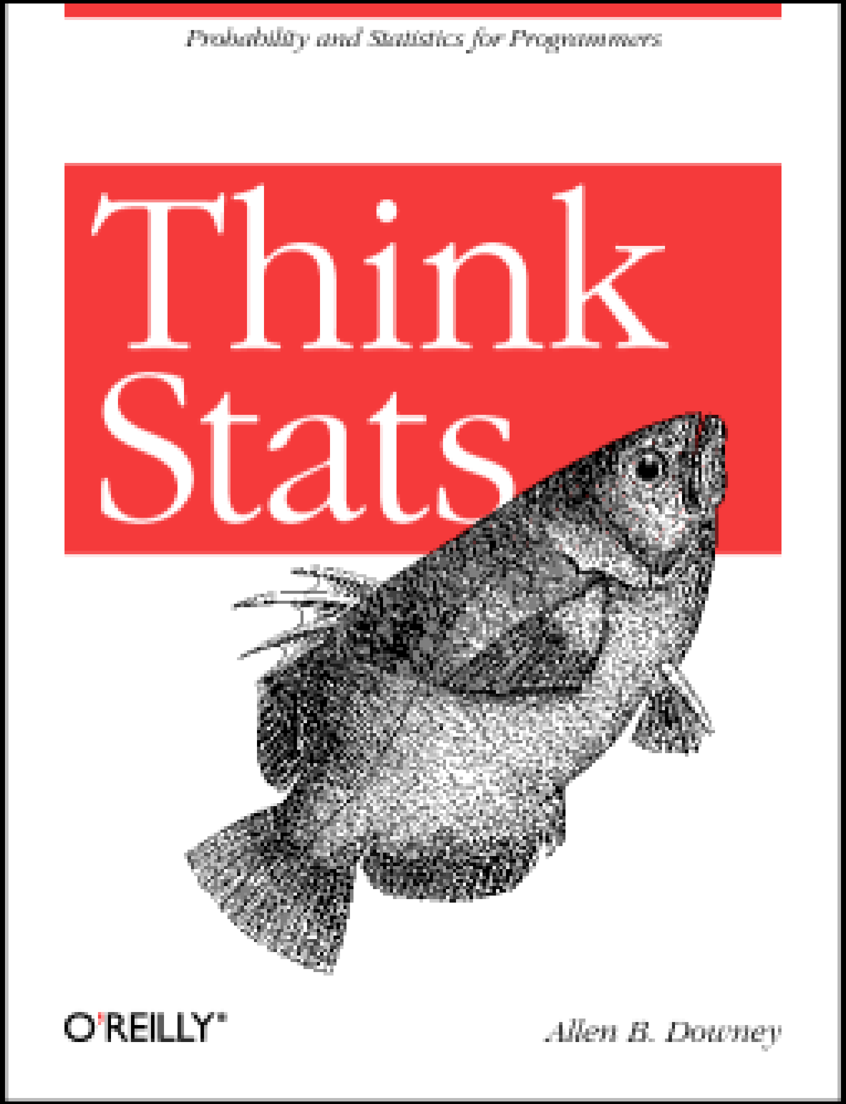
- Ideas explained in the example sheets are examinable


Copyrighted Material
SPRINGER TEXTS IN STATISTICS
A Modern Introduction to Probability and Statistics
Understanding Why and How
F.M. Dekking
C. Kraaikamp
H.P. Lopuhaä
L.E. Meester
Springer
Copyrighted Material


PROBABILITY MODELS
for Computer Science
Sheldon M. Ross


Probability and Computing
Randomization and Probabilistic Techniques in Algorithms and Data Analysis
Michael Mitzenmacher and Eli Upfal
SECOND EDITION


reddit
hot
new
rising
controversial
top
gilded
wiki
DATA IS BEAUTIFUL
hot
new
top
wiki
DATASETS
Search
HOT
NEW
RISING
TOP
GILDED

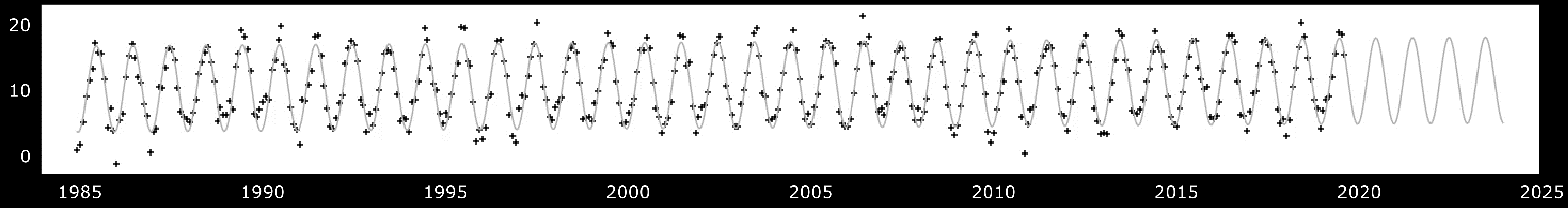

Probability and Statistics for Programmers
Think Stats
O'REILLY
Allen B. Downey

I. Learning with probability models

Most machine learning and data science tools boil down to

1. write out a probability model
2. learn from data (fit the model / estimate its parameters)

Monthly average temperatures [°C] at Cambridge station



Probability models describe everything that *might have* happened, so you can interpret the significance of what *did* happen.

Most machine learning and data science tools boil down to

1. write out a probability model
2. learn from data (fit the model / estimate its parameters)

fitting ≡ optimization

Programming in the 2.0 stack

Software 1.0 is code we write. Software 2.0 is code written by the optimization based on an evaluation criterion (such as "classify this training data correctly"). It is likely that any setting where the program is not obvious but one can repeatedly evaluate the performance of it (e.g. — did you classify some images correctly? do you win games of Go?) will be subject to this transition, because the optimization can find much better code than what a human can write.

Andrej Karpathy
@karpathy

Gradient descent can write code better than you. I'm sorry.

3:56 PM - 4 Aug 2017

343 Retweets 1,161 Likes

72 343 1.2K

1.1 Maximum likelihood estimation

tl;dr. Assume we have observed data, and we’re told the probability model behind the data. Assume also that this probability model has an unknown parameter, which we wish to estimate.

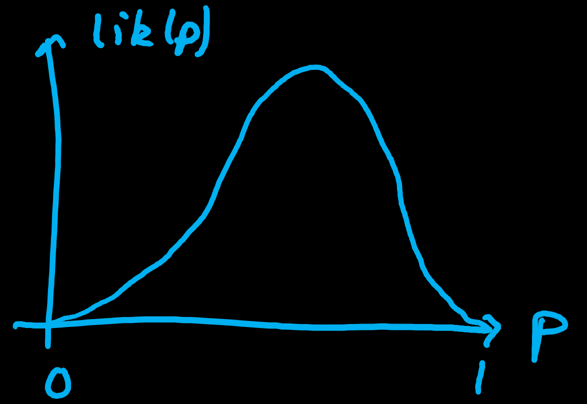
The *likelihood* is the probability of the observed data, viewed as a function of the unknown parameter. The *maximum likelihood estimator* or *mle* is the parameter value that maximizes the likelihood.

Exercise 1.1 (Coin tosses).

Suppose we take a biased coin, and tossed it $n = 10$ times, and observe $x = 6$ heads. Let's use the probability model

$$\mathbb{P}(\text{num.heads} = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0,1,\dots,n\}$$

where p is the probability of heads and $1-p$ is the probability of tails. What is p?



$$\text{lik}(p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\frac{d}{dp}: \quad \binom{n}{x}\left[x p^{x-1}(1-p)^{n-x} - (n-x)p^x(1-p)^{n-x-1} \right]$$

$$\longrightarrow \quad \hat{p} = \frac{x}{n}$$

Or:

$$\log \text{lik}(p) = K + x \log p + (n-x)\log(1-p),$$

K doesn't depend on p

$$\frac{d}{dp}: \quad \frac{x}{p} - \frac{n-x}{1-p} = 0 \quad \longrightarrow \quad \hat{p} = \frac{x}{n}$$

often we write $\text{lik}(p \mid x)$

Exercise 1.2 (The plug-in principle).

Suppose we take a biased coin, and tossed it $n = 10$ times, and observe $x = 6$ heads. Let's use the probability model

$$\mathbb{P}(\text{num.heads} = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x \in \{0, 1, \dots, n\}$$

where p is the probability of heads and $1 - p$ is the probability of tails.

Estimate $h = p/(1-p)$, called the "odds of heads".

HARD WAY. Write the model in terms of h, via $h = \frac{p}{1-p} \Leftrightarrow p = \frac{h}{1+h}$

$$\mathbb{P}(\text{numheads} = x) = \binom{n}{x} \left(\frac{h}{1+h}\right)^x \left(1 - \frac{h}{1+h}\right)^{n-x}$$

$$\longrightarrow \quad \hat{h} = \frac{x}{n-x}$$

EASY WAY. We already found $\hat{p} = \frac{x}{n}$.

Plug this into the formula for h: $\hat{h} = \frac{\hat{p}}{1-\hat{p}} = \frac{x/n}{1-x/n} = \frac{x}{n-x}$

Exercise 1.3 (Estimating multiple parameters).

Suppose we ask $n = 100$ people their views on Brexit, and 37 say Leave, 35 say Remain, and the other 28 don't care. Using the probability model

$$\mathbb{P}(\text{leavers} = x_L,\ \text{remainers} = x_R) = \frac{n!}{x_L!\,x_R!\,(n - x_L - x_R)!} p_L^{x_L} p_R^{x_R} (1 - p_L - p_R)^{n - x_L - x_R}$$

find maximum likelihood estimators for p_L and p_R.

$$\log \text{lik}(p_L, p_R \mid x_L, x_R) = k + x_L \log p_L + x_R \log p_R + (n - x_L - x_R) \log(1 - p_L - p_R)$$

$$\left.\begin{aligned} \frac{\partial}{\partial p_L}: \quad & \frac{x_L}{p_L} - \frac{n - x_L - x_R}{1 - p_L - p_R} = 0 \\ \frac{\partial}{\partial p_R}: \quad & \frac{x_R}{p_R} - \frac{n - x_L - x_R}{1 - p_L - p_R} = 0 \end{aligned}\right\} \qquad \begin{aligned} \hat{p}_L &= x_L / n \\ \hat{p}_R &= x_R / n. \end{aligned}$$

Exercise 1.3 (Estimating multiple parameters).

Suppose we ask $n = 100$ people their views on Brexit, and 37 say Leave, 35 say Remain, and the other 28 don't care. Using the probability model

$$\mathbb{P}(\text{leavers} = x_L,\ \text{remainers} = x_R) = \frac{n!}{x_L!\, x_R!\, (n - x_L - x_R)!} p_L^{x_L} p_R^{x_R} (1 - p_L - p_R)^{n - x_L - x_R}$$

find maximum likelihood estimators for p_L and p_R.

An estimator is a function of the observed data. You feed in the data, you get out an estimate.

BAD ANSWER

The log likelihood is

$$\log \text{lik} = \kappa + x_L \log p_L + x_R \log p_R + (n - x_L - x_R) \log(1 - p_L - p_R)$$

To find the maximum likelihood estimator for p_L, set the derivative equal to zero:

$$\frac{d}{d} \log \text{lik} = \frac{x_L}{p_L} - \frac{n - x_L - x_R}{1 - p_L - p_R} = 0$$

giving

$$\hat{p}_L = (1 - p_R) \frac{x_L}{n - x_R}$$

Similarly,

$$\hat{p}_R = (1 - p_L) \frac{x_R}{n - x_L}$$

This is not a valid estimator because it involves an unknown parameter p_R.

1. Specifying and fitting models

1.2 Numerical optimization

tl;dr. To find the minimum of a function

$f: \mathbb{R}^K \to \mathbb{R}$,

```
import scipy.optimize

def f(x):
    return …

x0 = […]  # initial guess
x̂ = scipy.optimize.fmin(f, x0)
```

- Choose x_0 wisely
- This function finds a local minimum, perhaps not a global minimum
- See the documentation to control number of iterations, …

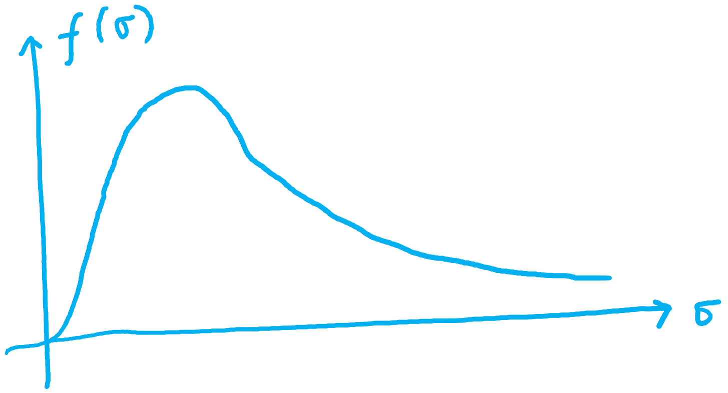
tl;dr. To find the minimum of a function $f : \mathbb{R}^K \to \mathbb{R}$,

```
import scipy.optimize

def f(x):
    return …

x₀ = […] # initial guess
x̂ = scipy.optimize.fmin(f, x₀)
```

Exercise 1.4 (Constrained optimization).
Find the maximum over $\sigma > 0$ of

$$f(\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-3/2\sigma^2}$$

Hint. Instead of maximizing over $\sigma > 0$, maximize over $\tau \in \mathbb{R}$ using the transform $\sigma = e^\tau$.




```
import scipy.optimize
import numpy

π = numpy.pi

def f(σ):
    return numpy.exp(-3/2/numpy.power(σ,2)) \
           / numpy.sqrt(2*π*numpy.power(σ,2))

(τ̂,) = scipy.optimize.fmin(
           lambda τ: -f(numpy.exp(τ)),
           numpy.log(σ0))
σ̂ = numpy.exp(τ̂)
```

```
Optimization terminated successfully.
     Current function value: -0.139702
     Iterations: 13
     Function evaluations: 26

1.7320498691939412
```

Exercise 1.5 (Softmax transformation).
Find the maximum of

$$f(x_1, x_2, x_3) = 0.2 \log x_1 + 0.5 \log x_2 + 0.3 \log x_3$$

over $x_1, x_2, x_3 \in [0,1]$ such that $x_1 + x_2 + x_3 = 1$.

This is called the softmax transformation, and it's widespread in machine learning models.

Instead of optimizing over (x_1, x_2, x_3), we'll optimize over $(\xi_1, \xi_2) \in \mathbb{R}^2$ with

$$x_1 = \frac{e^{\xi_1}}{e^{\xi_1} + e^{\xi_2} + 1} \qquad x_2 = \frac{e^{\xi_2}}{e^{\xi_1} + e^{\xi_2} + 1} \qquad x_3 = \frac{1}{e^{\xi_1} + e^{\xi_2} + 1}$$

The exponentiation ensures we get positive values, even for negative ξ.

The normalization ensures $x_1 + x_2 + x_3 = 1$.

```
def f(ξ):
    ξ1,ξ2 = ξ
    x = numpy.exp([ξ1,ξ2,0])
    x1,x2,x3 = x / sum(x)
    return 0.2*numpy.log(x1) + 0.5*numpy.log(x2) + 0.3*numpy.log(x3)

ξ1,ξ2 = scipy.optimize.fmin(lambda ξ: -f(ξ), [0,0])
x = numpy.exp([ξ1,ξ2,0])
x = x / numpy.sum(x)
```

Optimization terminated successfully. Current function value: 1.02965. Iterations: 63. Function evaluations: 120
array([0.19999474, 0.49999912, 0.30000614])