# UNIVERSITY OF CAMBRIDGE

# Information, Inference, and Energy

A symposium to celebrate the work of
Professor Sir David MacKay FRS

# Programme

## Monday 14th

| | | |
|---|---|---|
| 09:15 | 09:30 | Welcome & introduction |
| 09:30 | 10:00 | David Spiegelhalter |
| 10:10 | 10:50 | Tom Counsell |
| 10:50 | 11:20 | *Coffee & Tea* |
| | | |
| 11:20 | 12:00 | Amin Shokrollahi |
| 12:00 | 12:40 | Jossy Sayir |
| 12:40 | 12:50 | *Brief reminiscences I* |
| 13:00 | 14:00 | *Lunch* |
| | | |
| 14:00 | 14:40 | Mark Lynas |
| 14:40 | 15:20 | Brendan Frey |
| 15:20 | 15:30 | *Brief reminiscences II* |
| 15:30 | 16:00 | *Coffee & Tea* |
| | | |
| 16:00 | 16:40 | Oliver Stegle |
| 16:40 | 17:20 | Richard Durbin |
| 17:20 | | *(end of Monday programme)* |
| | | |
| 19:15 | 19:55 | *Pre-dinner drinks (Trinity College)* |
| 20:00 | 22:30 | *Dinner (Trinity College)* |

## Tuesday 15th

| | | |
|---|---|---|
| 09:30 | 10:10 | John Hopfield |
| 10:10 | 10:50 | Carlos Brody |
| 10:50 | 11:20 | *Coffee* |
| 11:20 | 12:00 | Andreas Herz |
| 12:00 | 12:40 | Per Ola Kristensson |
| 12:40 | 12:50 | *Brief reminiscences III* |
| 13:00 | 14:00 | *Lunch* |
| 14:00 | 14:40 | John Winn |
| 14:40 | 15:20 | Iain Murray |
| 15:20 | 16:00 | Geoff Hinton |
| 16:00 | 16:30 | Surprise talk & closing remarks |
| 16:30 | 17:30 | *Coffee and cake reception* |
| 17:30 | | *Departure* |

# Abstracts

## Communicating risk using whole numbers

**David Spiegelhalter**

David Mackay's "Sustainable Energy – without the hot air" showed the power of translating complex quantities into whole numbers that can be added and easily compared. Similar ideas have been tried in communicating risk, the units being the Micromort for acute risks and the Microlife for chronic risks. I shall show how these can be used to see how recklessly dangerous, or how boringly safe, your life is.

## The Global Calculator: Trying to apply 'Sustainable Energy - without the hot air' to the world

**Tom Counsell**

Could 9 billion people have a rich world standard and style of living and yet avoid the worst effects of climate change, if only we built enough nuclear power stations? or covered enough desert with solar panels? or all used low energy light bulbs? or efficient cars? Inspired by David Mackay's 'Sustainable Energy - without the hot air' a team managed by Sophie Hartfield at the UK Department of Energy and Climate Change tried to build an interactive calculator to help people to figure out the answer. This talk will cover their journey and the results of their attempt.

## Chordal Codes

**Amin Shokrollahi**

Modern electronic devices consist of a multitude of IC components: the processor, the memory, the RF modem and the baseband chip (in wireless devices), and the graphics processor, are only some examples of components scattered throughout a device. The increase of the volume of digital data that needs to be accessed and processed by such devices calls for ever faster communication between these IC's. Faster communication, however, often translates to higher susceptibility to various types of noise, and inevitably to a higher power consumption in order to combat the noise. This increase in power consumption is, for the most part, far from linear. In this talk I will give a short overview of problems encountered in chip-to-chip communication, and will advocate the use of a novel class of codes, called "Chordal Codes", to solve those problems.

## Codes for efficient data storage on DNA molecules

**Jossy Sayir**

Over the past 6 months I have been working with David MacKay and with Nick Goldman (EBI) on developing a coding system for storing data efficiently on DNA molecules. In this talk, I will focus on David's initial approach to the problem, which provides an information theoretic background for the achievable storage density and reliability. The assumption for this model is a pure packet loss model with noiseless recovery of packets, which David called the "fountain channel". This supposes that lower level coding can fully compensate for synthesis, sequencing and amplification errors such as insertions, deletions and substitutions. An overview of the literature on coding for channels with insertions, deletions and substitutions reveals surprisingly few publications, the majority of which co-authored by David MacKay, and none of which providing unbounded reliability in the information theoretic sense. Hence, I will discuss a model for DNA storage that combines the packet loss with the noisy nature of the recovered packets, and present coding strategies for exploiting this channel that David, Nick and I have been developing.

## Mind the Gap (between science and society) - the case studies of climate, nuclear and GMOs

**Mark Lynas**

Scientists often have a very different idea of risk and benefit from policymakers and the general public. With special attention to three case studies - climate change, nuclear power and genetically modified crops - Mark Lynas looks at the gaps between what scientists understand and what everyone else thinks, and how these gaps can lead to sub-optimal policy outcomes in the real world.

## Why Medicine Needs Deep Learning

**Brendan Frey**

My research on deep inference and learning reaches back to the wake-sleep algorithm, published in 1995, and the paper that David MacKay and I wrote in 1996 showing that belief propagation in graphs with cycles can be used for accurate inference. Most of my time is now spent on genomic medicine. Deep learning will transform medicine, but not in the way that many advocates think. The amount of data times the mutation frequency divided by the biological complexity and the number of hidden variables is small, so downloading a hundred thousand genomes and training a neural network won't cut it. There is what I call a "genotype-phenotype gap" and the value of closing this gap exceeds Google's $200B ad market. I believe that the only way to bridge this gap is to build machine learning systems that properly incorporate biological knowledge. I'll describe this approach, which is being pursued by dozens of young investigators, has improved our ability to "read the genome", and will, I believe, be an indispensable component in the future of medicine.

# New statistical approaches to disentangle single-cell diversity

**Oliver Stegle**

Many key biological processes are driven by differences in the regulatory landscape between single cells. Recent technical developments have enabled the transcriptomes and epigenomes of hundreds of cells to be assayed in an unbiased manner, opening up the possibility to identify and study new physiologically relevant, sub-populations of cells. In this talk I will discuss statistical advances to elucidate the factors that drive single-cell heterogeneity. Accurate and scalable latent variable models allow dissecting single-cell transcriptome and epigenome studies, thereby disentangling biological variation from technical and confounding factors. I will illustrate these approaches in applications to large datasets consisting of tens of thousands of cells and discuss their relationship to methods from population genetics.

# New statistical methods for the analysis of genome variation data

**Richard Durbin**

The last two decades have seen an exponential growth in the quantity of DNA sequencing data, at a rate more than twice as fast as Moore's law for growth in computing power. Individual sequences at one position in the genome are related by descent from a common ancestor in a tree, but this genealogy is made much more complex across the whole genome by recombination, which means that you inherit different sections of your genome from different recent ancestors. A clean mathematical model exists for the full genealogy but it is too complex to use for inference with data sets available today. I will discuss various algorithmic and representational innovations introduced to model and exploit genetic relatedness in analysing ever larger amounts of genome data, illustrating them with our work on the Haplotype Reference Consortium, which has assembled over 65,000 whole genome sequences with data at nearly 40 million variable sites.

## Emergence, dynamics, and behaviour

**John Hopfield**

How do the principles of psychology relate to the activities of 1011 nerve cells in the human brain? Large physical systems generally have emergent collective behaviours. Real computers—the brain included—compute by following a change of computer 'state' with time. When that change of state can be described by collective variables, robust dynamics emerge whose mathematical description can be entirely unlike that of the underlying microscopic variables. We illustrate the useful emergent computational dynamics of two elementary neural networks. The first deals with the problem of variable cadence when recognizing dynamical sensory patterns spread over time. The second achieves goal-directed behaviour when multiple candidate goals are simultaneously present.

## Neural substrates of decision-making in rats

**Carlos Brody**

The most common behavioral observation in decision-making, experienced both in our daily lives and in laboratory settings, is that easy decisions (where we are likely to choose the correct response) are done quickly, whereas difficult decisions (less likely to choose correctly) are much slower. An appealingly simple model was proposed in the behavioral literature many decades ago to account for this observation. This model, sometimes known as the "gradual accumulation of evidence" model, has been used to explain many behavioral data sets. Does the brain implement something well approximated by this model? If so, how does the brain's network of neurons actually carry out the implementation? We have been using using a combination of computational and experimental approaches with rats to try to answer these questions.

## Decoding the population activity of grid cells for spatial localization and goal-directed navigation

**Andreas Herz**

Mammalian grid cells discharge when an animal crosses the points of an imaginary hexagonal grid tessellating the environment. I will show how animals can navigate by reading out a population vector of such activity patterns across multiple spatial scales. The theory explains key experimental results about grid cells, makes testable predictions for future physiological and behavioural experiments, and provides a mathematical foundation for the concept of a "neural metric" for space. For goal-directed navigation, the proposed allocentric grid cell representation can be readily transformed into the egocentric goal coordinates needed for planning movements.

Joint work with Martin Stemmler and Alexander Mathis.

## Next-generation text entry

**Per Ola Kristensson**

Text entry is a common everyday computing task. However, despite its ubiquitousness it is difficult to devise an efficient text entry method that users are willing to adopt. In this talk I will explain the narrow design space of text entry research and make the case that successful next-generation text entry methods are likely to be based on designs that merge behavioural solution principles with information engineering techniques. I will exemplify this principle with several new text entry methods we have developed for a variety of use-cases.

## Democratising data science

**John Winn**

There has never been a higher demand for data science and data scientists. In science, in medicine, in business, in sport, in every area of life, we now have the tools to accumulate large quantities of data, but we lack the corresponding tools to process and understand it. I will talk about work that we have done in Microsoft Research to try to build such tools, the challenges that we have encountered and suggest how data science might be made much more broadly accessible in the future.

## Pseudo-Marginal Slice Sampling

**Iain Murray**

This talk is about sampling from probability distributions in a challenging setting: we can only compute noisy (but unbiased) estimates of the probability density function at individual settings of variables we choose. Previous work on such "pseudo-marginal" methods have been used in inference problems in genetics, continuous time stochastic processes, hierarchical models, and "doubly-intractable" distributions. I'll present algorithms that are easier to apply than previous work and sometimes work a lot better.

Joint work with Matt Graham.

## Can sensory cortex do backpropagation?

**Geoff Hinton**

Stochastic gradient descent in multilayer networks of neuron-like units has led to dramatic recent progress in a variety of difficult AI problems. Now that we know how effective backpropagation can be in large networks it is worth reconsidering the widely held belief that the cortex could not possibly be doing backpropagation. Drawing on joint work with Timothy Lillicrap, I will go through the main objections of neuroscientists and show that none of them are hard to overcome if we commit to representing error derivatives as temporal derivatives. This allows the same axon to carry information about the presence of some feature in the input and information about the derivative of the cost function with respect to the input to that neuron. It predicts spike-time dependent plasticity and it also explains why we cannot code velocity by the rate of change of features that represent position.