Joint Call

EXPLORING NOVEL OPPORTUNITIES FOR DATA SCIENCE IN BIOMEDICAL RESEARCH

Joint Funding Call
Crick-Turing Biomedical Data Science Awards

Guidance Notes for Applicants

The Alan Turing Institute (the Turing) the Francis Crick Institute (the Crick) is launching a call for Crick-Turing Biomedical Data Science Awards. The call aims to catalyse productive collaborative research between biomedical investigators and data scientists. Awards will provide funding for Post-doctoral level researchers working in the field of AI and data science to be based at Crick for up to 12 months to pilot research projects that have been collaboratively developed with a biomedical investigator to apply data science approaches to biomedical challenges.

For completeness, these Guidance Notes contain all information about the call as well as essential guidance for applicants.

Contents

Background

The Francis Crick Institute is a world-leading multi-disciplinary biomedical research institute bringing together outstanding scientists from all disciplines to carry out fundamental research to improve the health and quality of people's lives. Research at the Crick aims to understand why disease develops and to find new ways to diagnose, prevent and treat a range of illnesses − such as cancer, heart disease and stroke, infections and neurodegenerative diseases.

Over recent years we have experienced advances in technology and an exponential increase in the generation of new datasets, many of which are characterised by their volume, variety, and complexity. Such datasets include, but are not limited to, genomic data derived from high-throughput sequencing and screening technology, single-cell data based on flow cytometry, proteomics, metabolomics, and a range of quantitative and qualitative data arising from microscope imaging of cells and tissues. The application of data science approaches and innovative analyses of these datasets have the exciting potential to improve the diagnosis, treatment, and prognosis of a range of medical diseases.

Scope

Proposals are invited from all post-doctoral level researchers (not limited to the Turing university partner networks) working in the field of AI and data science to pilot research projects that have been collaboratively developed with a biomedical investigator to apply data science approaches to biomedical challenges. It is envisioned that these pilot projects would lead to more ambitious follow on projects.

Applicants can apply to address or contribute to the addressing of one of the biomedical challenges identified in **Appendix A** that would benefit from the application of advanced data science approaches, including statistical modelling, machine learning and AI.

Eligibility

This call is open to all post-doctoral level researchers (not limited to the Turing university partner networks) working in the field of data science and AI.

You can be bought out from a current project or employed directly for the duration of the project.

Applicants are particularly encouraged to apply if they have experience in (but not limited to):

- Applications of data science to challenging problems in life science research
- Advanced statistical modelling for complex high-dimensional data
- Modern machine learning approaches to image analysis
- Modelling nonlinear and stochastic dynamics

Applicants that have not traditionally worked in the biomedical sector are also encouraged to apply in order to promote diversity of ideas.

The Crick and the Turing are actively committed to promoting equality and diversity.

Funding available

Applications are expected to be for small scoping or pilot projects for Post-doctoral researchers to work on a full time or part time basis at the Crick for 3, 6, 9, or 12 months,

embedded within a lab at the Crick, and benefitting from the outstanding research environment that it provides. Successful applicants will also have access to facilities at Turing and can attend a range of events, seminars, reading groups, and workshops held at the Institute – **see a selection of what is happening at Turing on our events pages**.

Eligible costs may include:

- Research assistant salary including on-costs (either in the form of a buy-out or through direct employment by the Turing)
- Travel and subsistence, and other meeting costs where relevant.
- Research consumables directly attributable to the project (e.g. high-performance computing)
- Research equipment essential for the project

We anticipate funding 5 to 10 awards with start dates in early 2020.

How to apply

To apply for funding please submit a short Expression of Interest (EoI) Form online via the smart survey link: https://www.smartsurvey.co.uk/s/84NLW/ by **11:00 GMT 23rd October 2019**. Applicants will be required to complete the following sections as indicated:

- Applicant details – full name and university affiliation
- Please indicate which principal investigator (PI) at Crick you would like to collaborate with (see Appendix A for list of PIs and potential projects).
- Summary of why your experience is relevant to the challenge area and link to research page (100 words max).

Once we receive your EoI, we will introduce you to the PI at Crick to progress to the next stage of collaboratively developing a full application. Applicants are encouraged to submit EoIs as soon as possible as we may be able to put you in touch with the PI at Crick sooner to allow more time for application development.

Full applications must be made online via the Turing's Flexigrant portal: https://ati.flexigrant.com (which will open soon). Applicants will be required to complete the following sections on the Flexigrant portal directly or as a PDF upload to the portal, as indicated:

- **Project Summary: maximum of 200 words:**
  - *Please provide the project title and duration* (*Please note that successful projects must start no-later than 1st May 2020)*
  - *Briefly describe the scientific problem/challenge to be addressed and the context i.e. describe the problem, why it is a problem, and what specific elements of the problem this project trying to address.*

- **Brief description of the proposed data to be modelled/analysed: maximum of 100 words:**
  - *Please provide details of the dataset underpinning the project, its availability, how it was created and what will it include.*

- **Research proposal: maximum 2 pages, plus an additional page for bibliographic references if needed, covering:**
    - *A Case for Support structured with the following headings, in order: aims and objectives of the project, tools/approaches/methodologies to attain stated objectives, identify anticipated deliverables and potential outcomes arising from the research and understanding of potential follow-on activities.*

- **Budget: high-level summary covering:**
    - *Total amount of funding being requested for the project and high-level breakdown, with following budget lines made clear:*
        - *Research assistant salary*
        - *Travel and subsistence, and other meeting costs where relevant.*
        - *Research consumables directly attributable to the project (e.g. high-performance computing)*
        - *Research equipment essential for the project (e.g. laptop)*

- **Ethics Approval and Consent for Data Use**
    - *Please note that if your application is successful, you will be required to submit evidence of approval from your institution's Ethics Committee (or equivalent) prior to the award beginning.*
    - *All patient data that is to be used in research must be fully consented and anonymised: please tick this box to confirm that you have the necessary consent and anonymization in place.*
    - *Please note that proposals must utilise existing data and should not be reliant on new data collection.*

- **Letters of Support**
    - *Please upload a letter of support from either the Principal Investigator of the Post-doctoral researcher, Head of Department or from the Research Office (or equivalent), confirming that they have read this application, have approved the costs submitted, and support Post-doctoral researcher undertaking the placement at the Crick should this application be successful.*
    - *Please upload a letter of support from host organisation (the Crick) PI confirming support of this project.*

Text should be single-spaced, with page margins of at least 2cm and font size no smaller than Times New Roman point 11 (or equivalent). Please note that any documents that exceed the guidelines on length, above, may be automatically truncated before being submitted for peer review.

Applications must be submitted no later than **11:00 GMT on 27th November 2019.** Only applications submitted through the Turing's Flexigrant system will be accepted for processing. The application submitted through Flexigrant will be taken to be the final version, and will be the version used for assessment.

Assessment

All applications will be assessed by a specially-constituted multidisciplinary Panel,

against the following criteria:

- Quality of the case for support
- Availability and quality of data
- Feasibility of the plan to achieve objectives within the timeframe of the project
- Potential added value of combined expertise of Post-doctoral researcher and PI
- Understanding of potential follow-on activities
- Value for money

The Panel will be chaired by Darren Wilkinson.

What we will do with your information

The personal information that you provide within the application will specifically be used for administering this call. The information will be viewed by Crick and Turing staff and selection panel members, and your information will not be used for any other purpose without your specific consent.

Application and award timetable

Opening date for applications	10th October 2019
Closing date for EoI	11:00 GMT 23th October 2019
Invitations for full applications	25th October 2019
Closing date for full application	11:00 GMT 27th November 2019
Offers to successful candidates	w/c 9th December 2019

For questions regarding the application process, please contact:
Rebecca Babb (rbabb@turing.ac.uk), Programme Manager (The Alan Turing Institute).

Post-award information - Project Monitoring and Reporting

Each awarded project team will be required to produce short reports to review progress against objectives, monitor outcomes and deliverables, and mitigating risks and issues where necessary. The frequency of reporting intervals will be dependent on the project duration and be required no more than once a month.

Appendix A – Biomedical Challenges

The following Crick investigators have expressed an interest in being involved in this initiative, and have provided a very brief outline of potential areas of interest, to be developed into a full proposal in collaboration with interested Turing researchers.

Katie Bentley (Cellular Adaptive Behaviour Laboratory)

We have a lot of lightsheet microscopy data imaging mouse retinas to better understand vascular malformations in eye disease. Some of the datasets are just too big for us to handle, e.g. tracking nuclei across a huge field of view to see if any changes to blood vessel cell movement could be seen. Also, we have a half finished image analysis pipeline designed to segment vascular morphology in 3D by creating a surface mesh around the vessel shapes in large 3D image stacks (potentially TBs big) that we need help completing, optimising and calibrating, or even just redesigning from scratch. We plan to then use these quantified and segmented imaging datasets to drive large scale computer simulations of cells moving and growing the retinal vessels initiated on the meshes. We are lacking the big data expertise to cope with the huge files and 3D large scale nature of the data, so see much potential for longer term collaboration with Turing groups.

Domingue Bonnet (Haematopoietic Stem Cell Laboratory)

We are trying at present to quantify changes in the bone marrow microenvironment due to leukemia invasion. We are using 3D multi-colour (6-7) fluorescence imaging of bone marrow section from immunodeficient mice transplanted with or not leukemia (we are as well as trying to develop image mass cytometry). Having a software that can help us with image analysis and identification of different cell type will be great. AI or machine learning might help.

James Briscoe (Developmental Dynamics Laboratory)

The assignment of cell function and identity in tissues depends on Gene Regulatory Networks – recursively connected control networks that determine gene expression. The development of methods that provide high dimensional single cell resolution genomic data from tens of thousands of cells raises the potential of inferring the structure and activity of these networks and reconstructing cell differentiation trajectories. To this end, computational approaches from graph theory, machine learning etc need to be adapted and investigated.

Julian Downward (Oncogene Biology Lab)

Integrating spatial relationships in single cell data analysis and visualisation.
We have been studying ways to improve the effectiveness of drugs that inhibit immune checkpoints in lung cancer using mouse models. To do this, we use Imaging Mass Cytometry (IMC) to produce highly multiplexed images (~40 markers) of the tumour microenvironment with single cell resolution. We can use this technology to identify numerous immune cell types, stromal cells and tumour cells in the tissue, as well as obtain information about the maturation, activation, signalling and proliferation state of these cells. But perhaps even more importantly, we get information about their location, whether the leukocytes are interspersed with the tumour cells, restricted to

the stroma or excluded to the tumour periphery. We can measure the distance to the nearest blood vessel. And we can study the cell communities and social networks; for example, which cell types are often found in each other's proximity or instead show avoidance.
But how do we extract all such spatial relationships from the data? And how can we visualise both the phenotypic and spatial features of a tissue in a way that it is easily interpretable? Will it allow us to compare multiple images across a few treatment groups or correlate with clinical outcome? Can deep learning help to extract features within the spatial relationship of cells and tissue, that we have not yet thought of?

Florencia Iacaruso (Neuronal Circuits and Behaviour Laboratory)

My lab focuses on the study of the neuronal circuit mechanisms underlying multisensory processing. In particular, we are interested in understanding how the brain combines visual and auditory information in order to perceive the surrounding world and to guide behaviour. We perform neuronal recordings in behaving mice using imaging and electrophysiological techniques. These techniques allow us to study the activity of a few hundred neurons simultaneously at single cell resolution.
I believe that partnering with a group from the Turing would be an exciting opportunity to discuss our existing projects and develop new ideas. Furthermore, I hope that the partnership with the Turing will help us find new exciting techniques to analyse our data. The types of questions we would like to address are:
1) *How does the high dimensional neuronal representation relate to the mouse's ongoing behaviour?*
We would like to apply multidimensionality reduction techniques and/or explore new methods to relate the recorded neuronal activity to the mouse behavioural state which is simultaneously recorded, for example locomotor activity, facial movements and performance during a behavioural task.
2) *How does neuronal activity change during multisensory perceptual learning?*
In order to understand how neuronal representations change during learning, we need to take into account the mouse ongoing behaviour that can contribute to a seemingly noisy representation. We would like to develop new methods to uncover patterns of activity associated with multisensory processing that arise during learning.
3) *In the future we would like to develop models of circuit function incorporating the patterns of connectivity that we observe in the brain between different types of excitatory and inhibitory neurons.* The experimental side of this project will take longer and therefore a collaboration would be more fruitful on a longer timescale.

George Kassiotis (Retroviral Immunology Laboratory)

Although the human genome is (nearly completely) sequenced and annotated, the human transcriptome is still incompletely mapped. This is due to the enormous combinatorial potential of alternative start and stop sites and splicing of RNA molecules. We are interested in computational assemblies of RNAs that overlap with the abundant repetitive transposable elements in the genome and RNAs of lymphocyte antigen receptors.

Jonny Kohl (State-Dependent Neural Processing Laboratory)

• *Using machine-learning approaches to quantify state-dependent changes in social interactions:* Social behaviours, such as parenting, mating and aggression, are

intricately linked to animal's internal (physiological) states. Changes in behavioural parameters during social interactions can therefore be used as powerful readouts for internal state changes. This project will implement and improve machine learning approaches to detect subtle behavioural changes during social interactions in mice.
• *Using 3D image registration techniques for integration of multimodal imaging data:* Understanding information processing in the brain increasingly requires integration of different types of data (anatomical, physiological, molecular) into a common framework. In this project, we will build an image processing pipeline to register data from different imaging modalities (MRI, wide-field imaging, neural circuit tracing, gene expression data) into a common 3D reference space. This brain atlas will then be used to identify brain areas and circuits affected by hormones.

John McCauley (WHO Collaborating Centre for Reference and Research on Influenza, at the Crick)

Every year we acquire the sequences of about 20,000 new influenza genomes. We would like to explore whether we can use these to track the steps in the circulation of various new influenza viruses as they spread around the world.

Kathy Niakan (Human Embryo and Stem Cell Laboratory)

Project 1: Integration of single-cell omics analysis of early mammalian embryos and embryonic stem cells from multiple experiments and laboratories: Advances in single-cell molecular profiling allow for the measurement of different features at the single-cell level (RNA transcripts, DNA methylation, chromatin accessibility, etc.). These datasets are likely to reveal subtle differences between cells that single features cannot capture. However, the integration of different omics data produced by different labs is challenging. Can we use machine learning approaches to facilitate this integration and contribute to the construction a more refined notion of cell identity?
Project 2: Learning relationships between samples from raw sequencing data: Pre-processing and alignment of raw sequencing data is time consuming and requires manual intervention. Can we use machine learning approaches to reveal sample relationships from raw sequencing files? Can we use interpretable machine learning techniques to determine which features are responsible for sample differences?

Vassilis Pachnis (Development and Homeostasis of the Nervous System Laboratory)

The enteric nervous system (ENS) plays an essential role in gut function and homeostasis, yet its development, from bipotent progenitor cells to mature neurons and glial cells, is not completely understood. We have generated large scale single cell RNA-seq data sets measuring expression levels for ~20,000 genes in 100s or 1000s of cells. Together with ATAC-seq, proteomics and imaging data, this provides a rich and challenging data set for data integration, analysis and mathematical modelling.
Single cell data analysis is still in active development. Key questions are associated with dimensionality reduction and trajectory inference and segmentation/clustering. We are currently using methods from graph theory and machine learning to retrieve finer grained information on cell populations, their differentiation pathways and the determinants of cell fate.
We are also interested in how neurons migrate to form the enteric ganglia, interacting groups of 10 to 100 cells. Here, we hypothesize that the underlying pattern formation is controlled by a Turing/Gierer-Meinhard reaction-diffusion system with the neurons

migrating in response to the resulting neuro-attractant gradient.

Lucia Prieto-Godino **(**Neural Circuits & Evolution Laboratory)

We work on the evolution of neural circuits. Although as neuroscientists we are learning more about how brains of model organisms compute information (in particular using mice, flies, and worms), we know almost nothing about how they evolve. Different species display divergent behaviours, but how do neural circuits change across species to produce this divergent behaviours? And which gene modifications are responsible for this circuit changes? We are addressing these questions by using as a model the olfactory system of the larvae of closely related fly species. We have shown that the larvae of these species display different odour guided behaviours, despite their relatively close evolutionary relationships. Furthermore, they constitute powerful models because their olfactory systems share organisational principles with other animals, including humans, but they are much simpler, consisting of just 21 individually identifiable olfactory sensory neurons, and their cognate downstream partners. In the lab, we are investigating: 1) How are homologous neurons differentially connected across species? We have recently acquired an electron microscopy volume of the entire brain of one of these species, and we are reconstructing its neuronal circuits at synaptic resolution. These results will be compared to an already existing dataset for another species. 2) How do these neurons respond differently to the same odours? We are recording the population activity of these neurons across species, and relating these responses to the odour stimulus statistics found in their natural environments . 3) How are neuronal genes evolving across species? We are performing single cell RNA sequencing in olfactory neurons across species. Together our research programme is generating a lot of novel data in a completely unexplored territory. Each of the projects would benefit from advanced mathematical approaches for data analysis to extract the maximum amount of information. Additionally, in the future, integration of these different types of data could bring a coherent understanding of how neural circuits evolve at the genetic, network and behavioural level.

Erik Sahai (Tumour Cell Biology laboratory)

Our group is interested in how the response to cancer therapies is determined by the combination of inter-cellular variation in cell signalling states and extra-cellular variation in the local tumour microenvironment. We rely heavily on imaging approaches to tackle these problems and generate large amounts of single cell resolved dynamic information about the activity of signalling pathways and cell migration. This is already starting to reveal different signalling dynamics between genetically identical cancer cells and heterogeneity in response to therapies. We would love have a better understanding of the causes of intercellular variation in signalling dynamics and how they might relate to whether a cell ultimately lives or dies when challenged with a therapy. In another strand of the work, we are studying how cell migration and cell-cell interactions determine the higher order spatial organisation of tumours. Relevant papers: 1. Hirata et al., Cancer Cell 2015, 2. Jenkins et al., PLoS Computational Biology 2015, 3. Ege, Dowbaj et al., Cell Systems 2018.

Paola Scaffidi (Cancer Epigenetics Laboratory)

1) *Transcriptional stochasticity and cellular plasticity in cancer*. The project would involve the development of methods to infer transcriptional kinetics from single-cell

RNAseq analysis and model how cells take advantage of transcriptional stochasticity to adapt to stress. Relevant references:
https://www.ncbi.nlm.nih.gov/pubmed/15166174;
https://www.nature.com/articles/s41586-018-0836-1
2) *Predicting epigenetic vulnerability in cancer*. Cellular states are maintained by a complex network of proteins (epigenetic regulators) that cooperate with each other to ensure correct gene expression programs. The project would involve analysis of large pharmacogenomics and expression data from cancer cell lines and use of machine learning approaches to i) identify epigenetic regulators that act as functional hubs in the network and ii) identify network-level transcriptional alterations that predict drug sensitivity.

Andreas Schaefer (Neurophysiology of Behaviour Laboratory)

We have one project that would benefit from strong machine learning and/or natural language processing input. A senior Post-doctoral researcher in the lab (Mihaly Kollo) has been developing a software called "Heron", which is a web-service developed for scientists, a map of scientific studies. It applies a combination of AI methods to extract and link methodological data from published literature, and provides a web interface that provides natural, intuitive navigation of the vast database of experimental details. Heron is developed and tested in partnership with Crick researchers to help researchers extend their interdisciplinary collaborations and improve reproducibility. This project requires especially natural language processing experience ideally (but not essential) with some biomedical background.