Research

Dr Heba Sailem
Research Fellow funded by the Wellcome Trust

PAINTING WITH BIG DATA

Every aspect of the world around us is a potential source of data – and now we can collect those data pretty much at will from the food we eat every day, the number of times we exercise, the things we buy, to how we feel. The rapid technological advances make it easier than ever to capture and store these data, any and all of which can and do provide insights into our own life and behaviours. Data on our lifestyle can be used in medical research: it allows us to answer questions such as how our daily activities, interactions with friends, and our environment affect our health. Businesses can also benefit: big data can provide a more personalised experience for customers and maximise organisations' profits. For example, the web pages we browse can be tracked by the browser and used to predict which other products may interest us. The term big data describes the methods and infrastructure that permit the efficient analysis and mining of large-scale datasets.

Although there is a lot to learn from our own data, it is just a small portion of the information we have the potential to collect. Our body is composed of around 30 trillion cells and 200 cell types where almost every cell has the genetic code of 20 thousand genes. The DNA code is unique to every person and provides a template to create different proteins. The proteins made in every cell depend on the tissue and organ type. These cells adopt different shapes and forms to support their function. Using different experimental techniques, we can collect various datasets describing the cells' genetic code and the level of expressed genes. Furthermore, we can observe the behaviour of individual cells under the microscope and the different proteins can be marked to understand what each does in the cells. These different data help us understand the biology underlying cell behaviour and, for example, how this behaviour changes as a result of disease.

RIGHT: Engaging via art: Dr Sailem participated in the design of a scientific quilt to showcase the science at the Big Data Institute

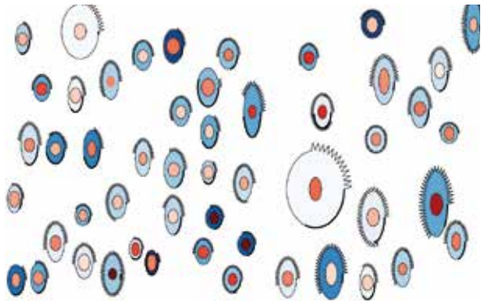
OPPOSITE PAGE: Avatars of breast cancer cells using PhenoPlot

> My research involves developing machine learning methods that mimic human ability in interpreting imaging but in a more systematic and objective manner

It goes without saying that cancer is a disease which is dangerous to life because cells start to behave abnormally. They proliferate more, live longer, and on some occasions start invading the surrounding tissues. Cancer cells tend to dominate other cells in the tissue and impact their function. Cancer becomes deadly when cancer cells colonise other organs which make the tumour difficult to resect. Changes in cancer cell behaviour can be caused by mutations where mistakes in copying DNA results in altered protein functions. One way to understand what the gene is doing in the cell is by perturbing its activity in vitro. For example, we can use methods like CRISPR to suppress the expression of a gene and identify how this can affect cell functions. We can think of proteins as the ingredients of the cells and their physical and chemical interactions give rise to the cell form and functions. An analogy is to take ingredients off a cake recipe one by one to understand each ingredient's role in the cake.

Perturbing every single one of the 20,000 genes in a sample of cells may sound a lot but these experiments are now done on a routine basis. Using robotic microscopy, we can image thousands of isogenic cells after these perturbations. The resulting data provides powerful means of identifying which gene is contributing to carcinogenesis. My research involves developing machine learning methods that mimic human ability in interpreting this imaging but in a more systematic and objective manner. This includes identifying cells and specific features such as length, area, number of neighbouring cells, or the abundance of certain markers. These features can allow the observer to draw inferences about what genes are doing, based on the effect of their perturbations.

Interpreting large imaging data remains a big hurdle and limits what we can learn from large perturbation data. This is the challenge I set out to tackle for my Sir Henry Wellcome Fellowship. I developed KCML, an intelligent system that combines prior knowledge on genes and machine learning to discover new gene functions. Surprisingly, using KCML, I found that smell-sensing genes might play a role in the spread of colon cancer. We have four hundred smell-sensing genes in our nose, allowing us to identify a wide range of scents. These genes can also be activated in other tissues – including the colon – but not much is known about their function in these tissues. My work revealed that perturbing many



smell-sensing genes results in abnormal organisation of colon cells and associates with known colon cancer genes. I was able to validate that the expression of these genes in colon cancer patients correlates with worse outcome. KCML can be applied to different datasets to advance our knowledge of gene functions and identify potential disease biomarkers.

Owing to the complexity of biological systems, one dataset would never provide all the clues. The abundance of big data means that we can increasingly investigate multiple datasets in order to try to explain a specific observation. For me, different types of data are like different colours: I use them to paint a story of cellular behaviour. Like art, data science requires a lot of creativity. Not far from art, I devised PhenoPlot, a first-of-its-kind visualisation method that allows drawing avatars based on measurements extracted from thousands of cancer cells in order to facilitate the understanding of microscopy data and to tell stories explaining their behaviour.

Although genetic changes are believed to be the main factor leading to cancer, they are not the only factor. Cancer can also develop due to changes in the microenvironment of cells, for example how cells are connected and who their neighbours are. During my PhD at the Institute of Cancer Research, I discovered that the shape of the breast cells and their surrounding in culture dishes can have a significant impact on the activity of an oncogenic gene that can turn other genes on or off. For example, when cells are surrounded by many other cells they will have a different response to drugs than when they are spread far away. This work demonstrates the importance of studying the genetic code along with the architecture and form of cells and tissues.

To understand the impact of cell context on its behaviour, I am now working on tissue imaging data from colorectal cancer patient biopsies and resections. Imaging is like a crystal ball that we try to see through to the past and future of cancer cells. I combine methods from computer vision, statistics, and bioinformatics guided by biological knowledge to characterise how the cell surrounding and the interaction between different cell types in the tissue lead to cancer initiation. On the one hand, this can help identify potential targets for patient treatment and on the other hand, it can assist doctors in diagnosing patients. I still do not know what my next painting will look like, but I hope it will bring a brighter future to cancer patients.