JOANNE HINDS

BEHAVIOUR PREDICTION: THE CHALLENGES AND OPPORTUNITIES OF BIG DATA

We base many of our decisions about other people on assessments such as what we think their personalities are like, or how they may behave in certain situations. Our ability to judge others accurately can have profound consequences in terms of who we socialise with, date or employ.

In physical environments, we use 'cues' such as a person's voice, dress or demeanour to form our judgments. Online, we may use their Facebook profiles, blogs or tweets. The online equivalent of these cues are often referred to as 'digital footprints' or 'digital traces'. These provide opportunities to analyse individuals' attributes and behaviour at mass scale and over long periods of time. So, as our interactions with technology continue to increase, can data be used to infer who we are and how we might behave?

PREDICTING BEHAVIOUR

Using data to predict behaviour has many applications including healthcare, marketing, and criminal investigation. In recent years, academics within psychology and computer science have examined the extent to which individuals' information can be inferred from their digital data. In particular, researchers have attempted to predict individuals' personality traits and demographic attributes.

Personality traits are emotions and behaviours that make up an individual's idiosyncratic disposition. The 'Big Five' (also known as the Five-Factor or OCEAN model) is the most popular approach currently used by researchers when measuring personality. Assessments consist of self-report questionnaires, which evaluate how highly individuals score across five dimensions as follows:

OPENNESS Have a variety of interests/hobbies, enjoy travel/adventure and are comfortable with change.

CONSCIENTIOUSNESS Highly organised, possess leadership skills, prefer planned activity over spontaneous behaviour.

EXTRAVERSION Sociable with many friends, outgoing and talkative, likely to participate in sports.

AGREEABLENESS Highly compliant, forgiving, cooperative and may be perceived as being a pushover.

NEUROTICISM Prone to depression, anxiety, low self-esteem as well as general negative emotions toward situations.

Demographic attributes can relate to any aspect concerning an individual's background characteristics or socioeconomic status. Predicting individuals' demographic attributes is well established in areas such as computer forensics and computational linguistics which often use text-based sources to predict an individual's age and gender. More recently, researchers have used digital data to predict other attributes such as location, occupation, level of education, sexual orientation, and political preferences.

Studies have also used digital data to successfully predict election results and reactions or opinions to events such as the Arab Spring and even box office revenue for films. For example, in the latter case, Márton Mestyán and colleagues demonstrated that the popularity of a movie could be predicted by editor and viewer activity on the film's Wikipedia entry.

HOW BEHAVIOUR CAN BE PREDICTED ONLINE

Similar to the way in which humans use cues to formulate opinions of other people, computer algorithms use 'features', where digital traces (e.g., Facebook likes, number of followers) are analysed to establish the strength to which they are associated with particular attributes (e.g., age, extraversion, location). Typically, this 'experiment' is performed on a subset of data, and then this subset is used to 'train' an algorithm to predict said attributes from the remainder of the dataset. The accuracy of the algorithm's performance then informs the researchers how successful their prediction was.

The ability to predict individuals' personal information, preferences and behaviour can have welcome effects and positive outcomes. Recommender systems such as Netflix or Amazon provide users with suggestions of films and products that individuals are likely to enjoy based on their previous activity.

Likewise, targeted marketing derived from our previous behaviour can be useful when individuals are exposed to advertisements that align with their personal preferences. However, such approaches are far from perfect and can sometimes be inconvenient or annoying. For instance, users who have purchased household items on Amazon go on to receive countless ads for the same items months later.

PREDICTING BEHAVIOUR: ETHICS AND CHALLENGES

Unfortunately, predicting information from digital data can extend beyond mere irritation to unintended or malicious consequences. For instance, individuals who are similar (in age, location, interests etc.) tend to be friends with, or connected with each other. Indeed, the notion that birds of a feather flock together (also known as 'homophily') is a truism often reflected in individuals' online social networks.

These patterns in online human networks can therefore create the potential for shadow profiling – where an individual's undisclosed or private information is revealed or inferred from data accessed through other people within their network. Recent research has emphasised the dangers of shadow profiling by demonstrating the potential to infer the sexuality of non-users of social networking sites.

The potential for shadow profiling highlights just one example of the type of ethical challenges surrounding the privacy and security of peoples' data.

The introduction of the EU General Data Protection Regulation (GDPR) that came in to force in May 2018, attempts to address more recent issues in terms of how personal data is handled. And whilst more up-to-date regulations are certainly necessary and beneficial, it is incredibly difficult to know the true extent to which technology will impact our lives (and our data) in the coming years.

The scale of the challenge is demonstrated by current estimates that predict around 30 billion online devices will be connected to each other by 2020. At the same time as these devices are generating data, data breaches occurring across banking, healthcare and technology companies (e.g., the WannaCry ransomware) have demonstrated the widespread threats to people's data across numerous industries.

In the case of the Cambridge Analytica scandal, data from approximately 87 million individuals' Facebook accounts were collected without their explicit consent.

Data like these were supposedly used to create targeted advertisements, such as those which attempted to influence people's voting preferences in the 'Vote Leave' campaign in Britain's European Referendum, and Donald Trump's 2016 presidential election.

Amidst concerns of how data are collected, used, shared and what true 'informed consent' really is, many people feel uncomfortable with the notion that their devices are 'listening' or that their behaviour is constantly being monitored or analysed. Whilst it is an exciting time for technological advancement and social science, organisations and cybersecurity practitioners face some complex challenges when it comes to handling data carefully and reinforcing trust in using technology.

Dr Joanne Hinds is a Research Associate at the University of Bath. Her work focuses on predicting information and behaviour from digital traces using psychological and computational techniques.