## Alessandro Bozzon

# Data science with a twist

Dr Alessandro Bozzon is an assistant professor at the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) and an expert on social data science. He is in charge of the Social Glass Project, which aims to achieve a better understanding of urban environments with the help of social media.

Internet and the web have revolutionized the way we create and access information. The web now hosts huge amount of data covering all aspects of human activity. However, being created by people, data are often influenced by their creator's views and biases. 'To make better computer systems that are personalized and inclusive, we need to understand more about the relationship between people and the data they produce. To achieve this at scale, often means moving out of our comfort zone as computer scientists,' says Dr Alessandro Bozzon. 'My vision is of a new generation of web data management systems that combine the cognitive and reasoning abilities of individuals and crowds, with the computational powers of machines, and the value of big amounts of heterogeneous web data'.
Existing theories about how people interact with the world can help. Personality psychology, for example, but carried out with computational methods. 'How to ascertain the personality traits of individuals without having to ask subjects to sit through hours of questionnaires? If we can automate that process at scale, with comparable results and accuracy, we can create systems that are able to adapt to your personality based on your behaviour and on the data you produce, and thus make your experience with them more fulfilling and engaging.'
One way to do this is to look at writing style. For instance, American psychologist James Pennebaker has shown that everyday language can reflect personality traits. Simply put, the frequency with which you use certain words can say something about your gender, your age or whether you are confident or insecure, for example. 'This is not as straightforward as counting words,' explains Bozzon. 'We know that people behave differently in different environments, and this reflects in the way they produce data,' he says. 'People have an official persona and an informal one. You have to take this systematic bias into account, especially in domain-specific settings. We are after robust and reliable methods to understand the properties of web data and of the people creating them, and to apply that understanding to create better systems.' Bozzon's work is currently applied in three societal and industrial domains: the urban environment, cultural institutions and the enterprise. He is involved in exciting projects in all three of them.

**Social Glass**

The Social Glass project looks at the urban fabric through the lens of social media. Every day, people generate huge amounts of data on social media about life in the city. By tapping into that source of information, it is possible to identify and study both incidents and structural problems, such as demographic pressure and tourism trends. Through social media, citizens can take an active part in improving life in the city, and ultimately their own well-being. 'With Social Glass we want to help making the city more liveable, and we are creating computer science tools for that purpose,' says Bozzon. The Social Glass approach goes beyond the way social media are currently being used by municipalities. 'Twitter is now used widely for interaction with the public, but we are not exploiting the full value of social data yet.'
Last summer, Social Glass experiments were carried out during the Sail Event in Amsterdam, where Amsterdam was the temporary centre of the nautical world, attracting 2.3 million visitors over five days. 'The Sail experiment was all about crowd management,' says Bozzon. 'In cooperation with the research group lead by Professor Serge Hoogendoorn, we monitored flows of people with the help of different techniques, such as cameras, sensors, as well as human sensors on social media. The project showed the added value of social web data, which can help complementing cold and dry numbers with explanations. Why do cameras detect a long queue forming in a certain area? Is it just because an event is popular, or is there something else going on?' In such a case, more information is needed. 'You want that information as fast as possible and as accurate as possible. We are now testing this at event scale, but we are working towards solutions at city scale.'
Some 60,000 Twitter and Instagram users were engaged online by the Sail event, producing a total of 226,000 posts. 'To help interpreting the event, we looked for user activities that gave an idea of how crowded people thought it was, looking for instance for words like

# About Alessandro Bozzon

Dr Alessandro Bozzon studied Computer Science & Engineering at the Politecnico di Milano, where he also obtained a doctorate in Information Engineering in 2009. He joined the TU Delft in 2012, where he is a member of the web Information Systems group at the EEMCS Faculty. He is part of the Delft Data Science Initiative, an investigator at the Amsterdam Institute for Advanced Metropolitan Solutions (AMS), and a Faculty Fellow with the IBM Benelux Centre of Advanced Studies.

'My background is in database and information retrieval, topics that I teach in Delft,' says Dr Alessandro Bozzon. 'I provide the next generation of engineers with the knowledge required to process and valorise large amount of data at scale.'

His research is centred on social data science. 'That is data science with a twist, the twist being that we acknowledge the central role that people have in their relationship with data.' Bozzon studies data-driven user modeling and engagement methods to create inclusive systems that improve the well-being of people. 'By understanding you and your goals, we can create systems that adapt to you, thus giving you a fulfilling and engaging experience.'

"full" or "busy".' Bozzon then applied user modelling techniques to determine if users were residents or visitors. 'In this way we could tell more about the difference in perception of crowdedness between residents and non-residents.' As it turned out, residents were less vocal. 'Perhaps tourists are not used to the big city, whereas resident experience such an event as a novelty, something outside normality.'

User modelling is essential to give meaning to web data. 'The city belongs to the people, so understanding who they are, how they live, and their cultural background gives you the key to interpretation. For example, if you are looking at Twitter, you have to know who actually is on Twitter. If that turns out to be mostly men in their forties, you still have go to Instagram to get the opinion of teenagers, for example.' That still leaves other groups unaccounted for too. 'You have to find ways to gather data that are inclusive. Not just observing data, but making sure you include the population that is of interest to your purpose. We call this process social sensing, an important part of the Social Glass project.'

Using social data also has the advantage of getting information fast. 'Nowadays municipalities still send out questionnaires to citizens, or compile a report on tourism once a year. By that time things have changed again. We want to shorten the time between the need for information and its actual availability.' There are now plans to apply the Social Glass method to larger areas and bigger events, such as the Gay Pride or King's day in Amsterdam. Bozzon is thinking even bigger. By next year he hopes to be working not on an event basis but continually, with the help of tools for the study of long-term dynamics and trends, possibly across cities. 'What I envision is an open repository, a kind of urban knowledge collider, where the needs of all stakeholders in cities are incorporated and all knowledge is made accessible, so that everyone can make use of it.'

**Rijksmuseum**

Cultural institutions are increasingly aware of the intrinsic value of their collections, which goes beyond the objects to include their usage. 'If collections are accessible and well-described, many more people can make use of them than museum visitors alone. So how do you do that at scale?,' wonders Bozzon. This question arose at The Rijksmuseum when it started digitising its print collection. It is one of the largest collections in the world, comprising over 700,000 prints. It was estimated that it would take professional annotators more than 20 years to go through the entire collection. 'Indeed, the Rijksmuseum will still be here in 20 years' time, but how can we speed up the process while guaranteeing annotation quality?' Moreover, the required knowledge is not always available at the museum. Art experts are not usually botanists or ornithologists as well. 'What we

‘Anyone can recognise a rose,’ but you still need an expert in botany to name the exact species’

need are tools for the acceleration of knowledge creation. How can we employ the cognitive ability of people to create knowledge in a systematic and reliable way?’ Human computation and crowdsourcing are key here. Bozzon explains: ‘It is extremely difficult for a machine to recognise a species of birds in a photograph; in a drawing it is almost impossible, because of style and interpretation, for instance.‘
The Accurator platform has been developed, to help the Rijksmuseum – and other museums later on – with annotating its collection. The platform is also used for research into so-called niche sourcing: the identification of those with specialist knowledge within a crowd. ‘Anyone can recognise a rose,’ says Bozzon, ‘but you still need an expert in botany to name the exact species.’ Once you have identified these experts, you also want them to stay involved until the project is finished. ‘Here we come back to personality. People are driven by different motives. That can be money, but also passion or your status within the group. If you are part of a community of passionate amateurs, perhaps the best way to keep you engaged is by having your contribution recognised by other members of your community.’

**Inclusive Enterprise**
In 2014, IBM and TU Delft joined forces in a Collaborative Innovation Center (CIC) on Big Data Science. ‘We are collaborating on several research projects; one of them is the inclusive enterprise. We want to use social data to increase the well-being and engagement of workers,’ says Bozzon. Naturally, companies want to avoid employees stagnating through lack of motivation. ‘Demotivation is often due to a lack of communication. The bigger the organisation, the harder it is for information to percolate through the organisation. You hardly know anything about your colleagues in the next office, your manager knows little about you, and his manager nothing at all.’ So how can data science help here? ‘You can find a lot of useful information about employees and their expertise within the data they produce. About existing networks, regardless of organisational structure, about how they reach out to the world. You can use this information to make people feel appreciated and valued, for example by engaging them in projects they are passionate about.’
Understandably, not all employees are eager to share too much with their employer. ‘That is where we, as computer and data scientists, can come in and make sure that all the safeguards with regards to data protection within the enterprise are met.’ At IBM, Bozzon and his colleagues can study the inclusive enterprise concept at scale, potentially involving the whole population of 400,000 IBM employees. ‘That is quite a challenge. In addition to ‘traditional’ computer science problems, we need to devise engagement techniques that are inclusive, and safe. So participation should be voluntary, and information should only be shared selectively, when it is beneficial to the employee.’
In an exemplary project, it was studied how employees can best be engaged with a gamified learning experience, where they can learn more about the company, its technology and their colleagues. ‘Through rigorous scientific work, we have shown the suitability of this approach, and discovered how different personalities are attracted to different kinds of incentives. A common misunderstanding about gamification is that all you have to do is give people a badge at the end. It is not; it is about being adaptive and inclusive.’
Bozzon believes the inclusive enterprise is the way forward. ‘Newer generations give great importance to their relations with the company, feeling good about what they do, and about the company’s understanding of their potential and strengths. They care about personal growth and influence. This new perspective brings novel challenges, and excellent opportunities for great research with societal impact.’