AI & Soc (2018) 33:347–356
https://doi.org/10.1007/s00146-017-0730-7

ORIGINAL ARTICLE

What is wrong about Robocops as consultants? A technology-centric critique of predictive policing

Martin Degeling[1] · **Bettina Berendt**[2]

Received: 10 January 2017 / Accepted: 8 May 2017 / Published online: 22 May 2017

Abstract Fighting crime has historically been a field that drives technological innovation, and it can serve as an example of different governance styles in societies. Predictive policing is one of the recent innovations that covers technical trends such as machine learning, preventive crime fighting strategies, and actual policing in cities. However, it seems that a combination of exaggerated hopes produced by technology evangelists, media hype, and ignorance of the actual problems of the technology may have (over-)boosted sales of software that supports policing by predicting offenders and crime areas. In this paper we analyse currently used predictive policing software packages with respect to common problems of data mining, and describe challenges that arise in the context of their socio-technical application.

Keywords Predictive policing · Data mining · Privacy · Big data

1 Introduction

In public debates and news reports, predictive policing is often explained with reference to the 2002 motion picture "Minority Report" that portrays a specialized police unit that arrests criminals before they commit a crime. But, the way the *precrime* unit operates, based on visions of three human mutants, has nothing to do with the data-driven approaches of actual predictive policing. Besides that, the fascination for crime prediction makes reporters ignore the main storyline of the film that might be closer to reality: ultimately the program is discontinued since the visions can be misinterpreted and also show false positives, meaning that the persons arrested would not necessarily have committed a crime.

In the real world, *predictive policing* refers to a variety of techniques used by police departments to generate and act on crime *probabilities*, often referred to as *predictions*. These non-binary probabilities are in most cases calculated by software programs that analyse previously recorded data and use machine learning algorithms to make assumptions about future developments. Perry et al. (2013) categorised the existing approaches of predictive policing as following:

1. Methods for predicting places and times of crimes.
2. Methods for predicting offenders and identifying individuals likely to commit crimes.
3. Methods for predicting perpetrators' identities.
4. Methods for predicting victims of crimes.

✉ Martin Degeling
degeling@cs.cmu.edu

Bettina Berendt
bettina.berendt@cs.kuleuven.be

[1] Institute for Software Research, School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Wean Hall 4121, Pittsburgh, PA 15213, USA

[2] Declarative Languages and Artificial Intelligence Group, Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium

The majority of predictive policing technology used in police departments in the western hemisphere falls into the first two categories. The highest adoption can be found in the United States (Perry et al. 2013: 57ff). While only a few states in Europe such as Germany and Switzerland[1] make use of predictive policing software, other states such as Italy[2], France[3] are still evaluating potential benefits of different products.

Each software is based on different assumptions about data and data mining processes, which are mostly disregarded when they are introduced in public policy debates. Based on the analysis of research papers and news reports about the software and its impact, we want to shed light on common misconceptions of data mining, the social constructs that are built into the algorithms and the risks that arise from their application.

On the one hand, predictive policing algorithms deliver statistics about the incidence of certain crimes (such as burglaries) in certain areas. This knowledge can serve to direct resources to patrolling these *hot-spot* areas more intensely than others, e.g. controlling passers-by more frequently than elsewhere. Predictions of which areas are more at risk are assumed to be more precise than those based on traditional techniques of crime mapping. But the measures taken such as stopping or stopping-and-frisking people are the same as in traditional police work, and they have to obey the same restrictions of reasonable suspicion and proportionality.

On the other hand, predictive policing is perceived as an application area of "Big Data" processing in which automated decision-making is superior to human decision-making, a perception that is pushed by software vendors, some academics, and popular-science authors such as Mayer-Schönberger and Cukier (2013). This view implies that technology is capable of solving nearly every problem of society—Morozov (2013) termed it *solutionism*—and often ignores the socio-technical contexts to which the technology is being applied. Predictive policing can change police work and its consequences on those that are meant to be protected, dramatically—not always in the way its inventors intended.

In this paper we discuss and question both perspectives. Considering the bold promises—more accurate statistics, better decision making, radical crime reduction—we describe the limitations of data mining in general and with regard to specific predictive policing software we argue that, for predictive policing to be useful, a socio-technical and not just a techno-centric perspective has to be taken.

The rest of this article is organized as follows: first we describe four predictive policing applications that are currently in use, two of them for predicting place and time of crimes and two of them predicting offenders. Then we discuss problems connected to each software at different necessary steps of data mining such as the selection and collection of data points, and common pitfalls of machine learning algorithms.

2 Examples of predictive policing software

As described above, the majority of available predictive policing software either predicts places and times of crimes, or it focusses on identifying likely offenders. Within each group there are multiple types of theories that inspired software developments: those that calculate the probability of a crime based on previous crimes in an area and those that take into account various features of a geographic space.

This geo-spatial crime prediction is widely adopted within the United States and Europe and is based on geographic information systems (GIS) that have been used since the 1960s. Its methods are based on crime research that shows that the location of crimes is not random but can be used for strategic analysis and planning of resources (Chainey and Ratcliffe 2013). In recent years, computer software enhanced the situation rooms with algorithms that are able to process large amounts of data and new statistical methods that make use of predictive models. Two widely used models that are also implemented in the software products described below are the *near repeat theory* and *Risk Terrain Modelling* (RTM).

The second category of predictive policing applications, to predict offenders, have not yet been as widely adopted. The idea of these systems is to calculate the likeliness that a given person will commit a crime or is prone to behaviour that puts others at risk. These approaches are closer to the often positively referenced "Minority Report", but they are also put under high scrutiny by privacy and human rights advocates.

2.1 PredPol: near repeats

PredPol[4] is a well-known company on the predictive policing market. It makes use of theories about crime

[1] Zurich (Switzerland) and Munich (Germany) are using the Software Precobs to predict burglaries. The software is developed by IFMPT http://www.ifmpt.com/.

[2] The Transcrime Research Center tried to predict burglaries in Rome, Milano and Bari for the year 2014. Report available online http://www.academia.edu/download/39022476/Transcrime_Research_in_Brief_Prevedere_i_furti_in_abitazione.pdf.

[3] In France, the technology was evaluated in 2015 http://www.20minutes.fr/societe/1612375-20150521-viols-agressions-cambriolages-nouvel-algorithme-gendarmes-predire-crime.

[4] See http://predpol.com/.

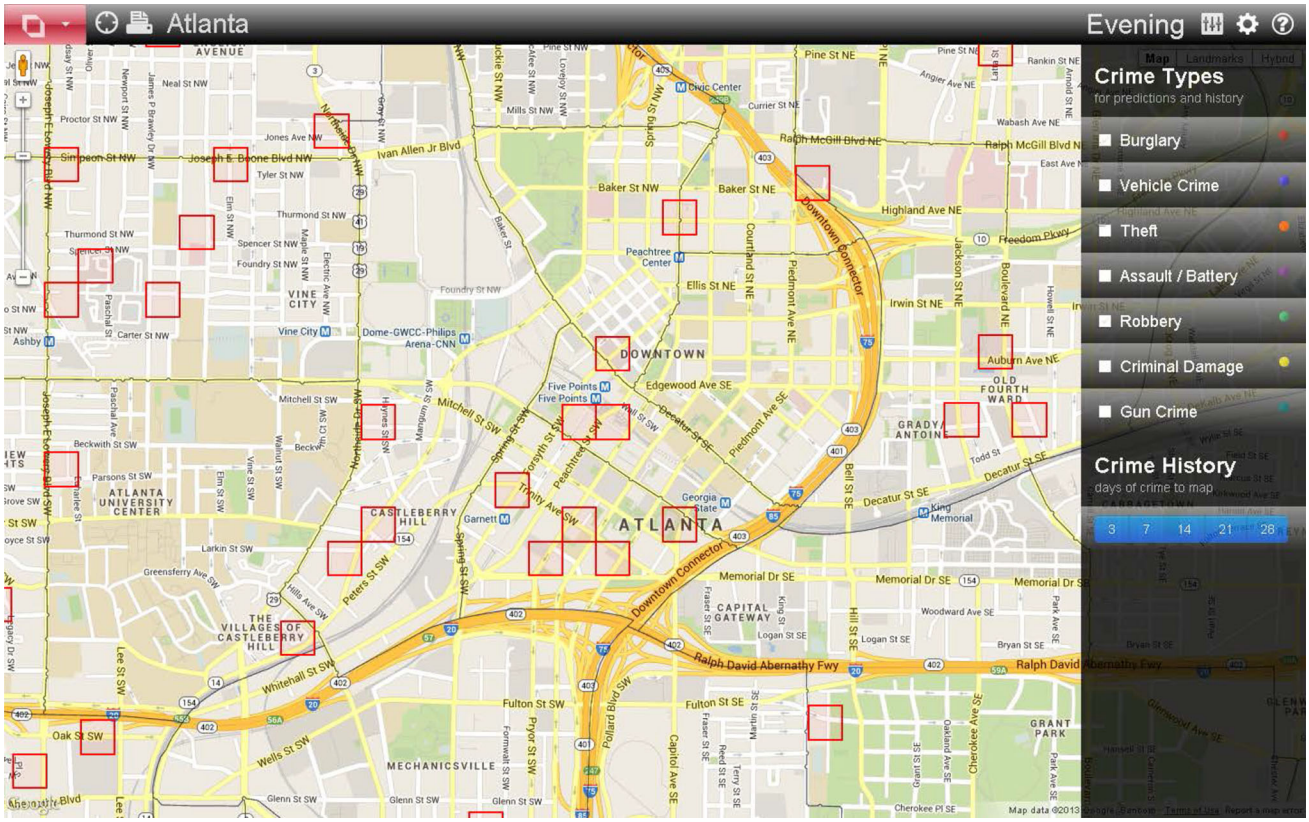

Fig. 1 PredPol screenshot from a product presentation (colour figure online)

patterns and sells a web application (see Fig. 1) to police departments. It not only shows where crimes are reported, but also makes predictions about possible future crime hotspots to guide resource allocation of police units.

The company was founded by researchers of the University of California Los Angeles. According to the company PredPol, their only product, is used by more than 60 police departments in the United States. It is sold as "Software as a Service" meaning that it is not run locally in each police department but on central servers controlled by the company. Local authorities can access the application with a regular web browser.

The software is based on an algorithm that was adapted to crime forecasting from seismology (Mohler and Short 2012). It makes use of the near repeat theory, assuming that after the first occurrence of an event, the likeliness of a repeated event of the same category increases, comparable to aftershocks of earthquakes. Research has shown that the theory works for some serial crimes such as burglaries that often occur in short period of time and close proximity to the original crime scene (Townsley et al. 2003; Johnson 2008). The hypothesis about the motivation is that successful burglars commit multiple burglaries in one night and are likely to return to the same neighbourhood on subsequent days.

2.2 Hunchlab: Risk Terrain Modelling

Hunchlab (see Fig. 2) is developed by Azavea[5] and its goal and appearance are similar to PredPol and, according to a company brochure, it also implements the near repeat theory but, more importantly, integrates other approaches such as RTM to improve the results (Azavea 2015).

A compendium (Caplan and Kennedy 2011) describes RTM as "an approach to risk assessment in which separate map layers representing the influence and intensity of a crime risk factor at every place throughout a geography is created in a GIS. Then all map layers are combined to produce a composite "risk terrain" map with values that account for all risk factors at every place throughout the geography". While the near repeat theory focuses on endogenous factors like repetitive behaviour, RTM takes only exogenous factors into account such as the position of certain landmarks.

The compendium combines results from a number of studies on environmental context factors of specific crime types that co-occur. For example, it states that street prostitution often takes place at roads that allow drivers to slow down or near bars where prostitutes can rest (Caplan and Kennedy 2011: p. 61). Each of these context factors

[5] https://www.hunchlab.com/.

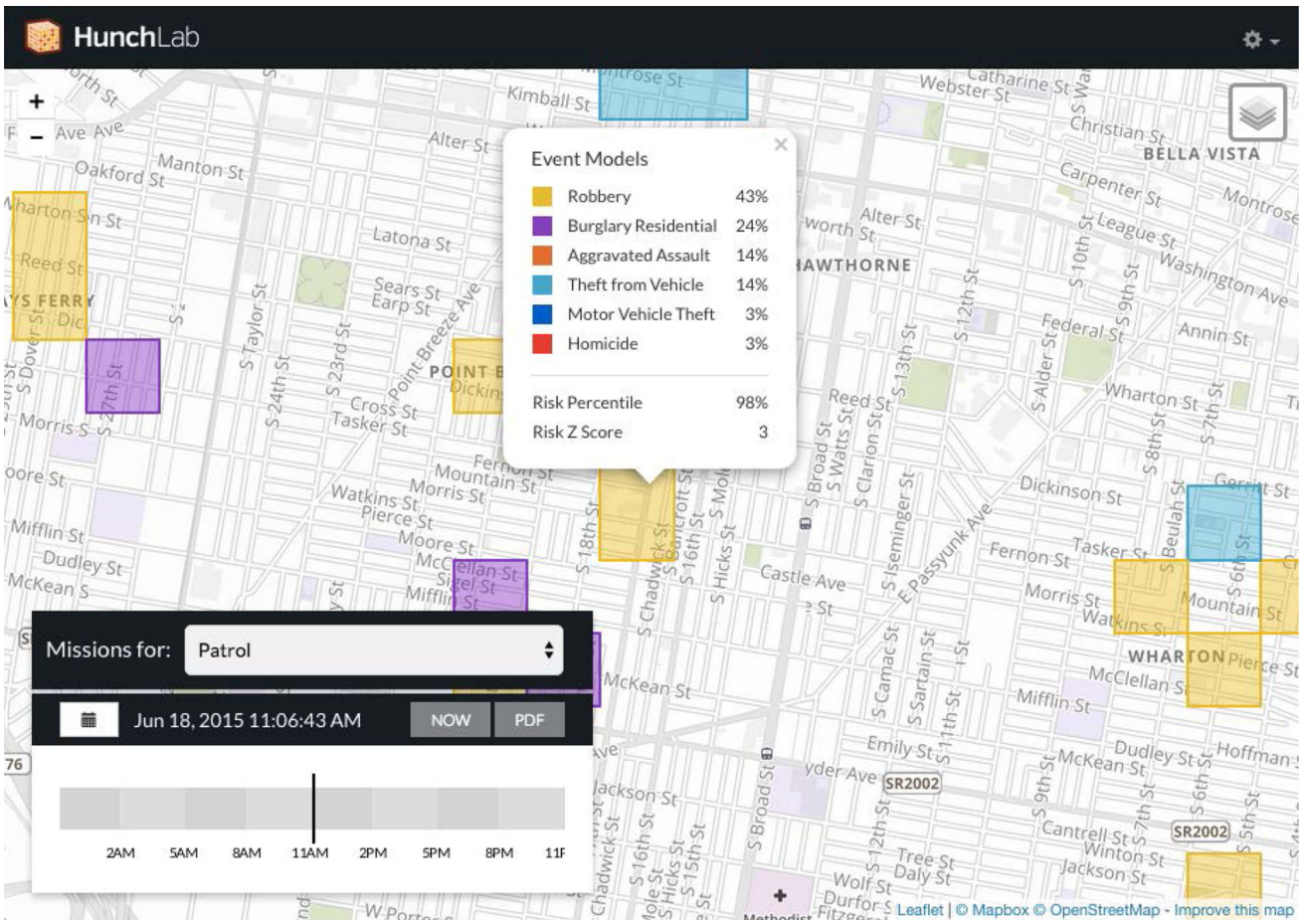

Fig. 2 Hunchlab screenshot from a product presentation (colour figure online)

can be quantified in different layers of a map like a road map that specifies the type of road or Point-of-Interest-Maps that lists bars, night clubs or banks.

2.3 Chicago's heat list: predicting offenders

The second type of predictive policing applications are those that look at the persons committing crimes instead of focusing on time and location of crimes.

In Chicago, the third-most populous city and the one with the highest number of gun violence in the US, the police department analyses networks of those arrested to calculate the likelihood of someone being involved in a serious crime (Gorner 2013). The resulting heat list is said to be based on research on the connections between homicide victims in Chicago in 1998 (Papachristos 2009). The original research analysed the social networks related to gang-related homicides based on past police records. The implementation in Chicago takes the idea of the relevance of social networks in crimes and makes no assumptions about possible crimes. Instead, it analyses various data points collected by the police[6] to compile a list of names of persons that are likely to be involved in major crimes. The police program also outlines that a list of influentials is created that may have an effect on persons on the heat list (McCarthy and Garry 2013). Police officers then single out individuals on the list and hand them a "Custom Notification Letter" that warns them about the charges they might face would they further engage in criminal activities.

2.4 Beware: threat scores

Another system that calculates the risks of individuals is "Beware", a software sold by West Corp.[7] Its target group are emergency first responders that are thought to be informed about potential risks that could be expected at the places they are going by looking at the threat score. When

[6] "Among the factors are the extent of a person's rap sheet, his or her parole or warrant status, any weapons or drug arrests, his or her acquaintances and their arrest histories — and whether any of those associates have been shot in the past" (Gorner 2013).

[7] https://www.west.com/safety-services/public-safety/powerdata/beware/.

emergency call centres receive a call, "Beware automatically runs the address. The search returns the names of residents and scans them against a range of publicly available data to generate a colour-coded threat level for each person or address: green, yellow or red" (Jouvenal 2016). The company claims that they also use data from social media to calculate the threat score. However, the use of the software is criticized because neither are people made aware of the score that is assigned to them, nor does the police department have any insight into how the score is calculated.

3 Predictive analytics caveats, and their application to predictive policing methods

In order to understand the epistemic and computational basis of "predictive policing" and each implementation, one needs to understand "predictive analytics". These are computational methods that predict likely outcomes in the future based on patterns that have been machine-learned from past data and (normally) validated against other past data.

3.1 Data mining with classifiers

All the systems described above make use of some kind of classifier. A classifier can be viewed as a decision rule: if (you observe) *this* then (think or do) *that*. This decision rule may be based on anything: intuition, prejudice, past human experience, or statistics. In the latter case, we say that the classifier has been "machine-learned" (or "data-mined"). Near repeat classifies a series of reports of criminal activities, RTM classifies properties of a location and the heat list as well as Beware classify information about an individual, e.g. their social network or other properties.

The key characteristic of predictive analytics is that predictions about future cases are made on the basis of statistical regularities in past cases. To what extent this is "good" depends on a number of criteria. People will have different opinions about the relative importance of the following criteria, but we believe that as decision makers (deciding on whether to apply predictive analytics in a given domain, or on whether to support such application as a politician or as a citizen) should at least be aware of them. Whatever one's beliefs and value judgements: the ideas that "the data speak for themselves" and that the predictions are "objective" and thereby avoid human limitations and biases are overly naive.

3.2 Measuring accuracy

The first set of criteria can be summarized by asking "how good are the predictions?". A classifier learned by a given algorithm from given data needs to be evaluated with respect to its accuracy (or some other measure of goodness of prediction). The reason is simple: a classifier that makes too many wrong predictions is useless and inacceptable.

Let us first look at a possible set of cases where the true nature of the person is known, and we have several descriptive features. Assume a dataset such as the one shown in Table 1.

A (very simple) classifier model that could be machine-learned from these data is that everyone with a red jumper and sweaty hands is a criminal, as is everyone wearing sandals and having a high-pitched voice. This model would be 100% accurate on the training data. (The statement is always true for the cases in which one of the premises holds.) These numbers, however, do not tell us how the model would perform when used to predict on new data. The latter is the yardstick of evaluation, and it is indeed accuracy (and other measures) on test data (that are disjoint with training data) that needs to be reported.[8] Table 2 shows the basic structure of evaluation in counts of entities from the set on which the classifier is tested, again using fictitious data.

In this fictitious example, the classifier that was trained on some historical data (or created otherwise), when applied to a new dataset of 1010 people, classifies 104 of them as criminals and 906 as innocent. These two datasets are known as "training data" and "test data", respectively. The prediction is correct in 904 cases (4 + 900), so the *accuracy* of the model is 904/1010 = 89.5%. However, the *precision* of the model for the class "criminals" is only 3.8%: out of 104 individuals classified as criminal, only four really are, the others are considered false positives. On the other hand, it does find 40% (4/10) of all those that really are criminal; this is the *recall* for the class "criminals". Compare this to the baseline predictor "always predict no": this will have an accuracy of 99% (1000/1010), but a precision, respectively, recall of 0 for the class of criminals. These results point to further pitfalls: the need to have balanced datasets (ideally with as many positive as negative examples), and the need to work on different datasets (to avoid overfitting the model to type of observations).

To be open to criticism and allow for improvement, evaluations of predictive policing—like those of other data mining applications—should be transparently documented

[8] This example has been heavily oversimplified for the purposes of a non-technical introduction. Any realistic classifier learning would take better account of noise, etc. The use of training and test data is also somewhat more involved in practice. Accessible (even if technical) introductions can be found in Witten et al. (2011). Their teaching materials are available at http://www.cs.waikato.ac.nz/ml/weka/book.html, see Chapter 5 on evaluation.

Table 1 A fictitious dataset for learning a classifier

ID	Skin colour	Colour of jumper	Shoes	Hands	Voice	Criminal?
1	Green	Red	Boots	Sweaty	Normal	Yes
2	Green	Red	Flip-flops	Sweaty	Deep	Yes
3	Green	White	Sandals	Dry	High-pitched	Yes
4	Green	Yellow	Sandals	Normal	High-pitched	Yes
5	Green	White	High heels	Dry	Normal	No
6	Green	White	Flip-flops	Dry	Normal	No
7	Blue	White	boots	Dry	Normal	No
–	(Not green)	–	–	–	–	(All no)

Table 2 A fictitious *confusion matrix* for evaluating a classifier

Individuals are	Classified as criminals	Classified as innocent	Row sum
Indeed criminals	True positives: 4	False negatives: 6 (falsely assumed to be innocent)	True total number of criminals: 10
In fact innocent	False positives: 100 (falsely assumed to be criminal)	True negatives: 900	True total number of innocents: 1000
Column sum	Positives: 104	Negatives: 906	1010

and ideally replicable by others. The predictive policing systems described fail on at least one of these points.

The models used for PredPol are the ones that have seen the most scientific analysis. Mohler et al. (2012) reported an accuracy of about 35% for their model. The low accuracy (compared to our example) is related to the fact that only a fraction of crimes can be considered a "near repeat", depending on what is considered "near". In addition, the accuracy varies between crime types, e.g. burglaries are more likely to result in near repeats than vehicle thefts.

One of the rare real-life evaluations of predictive policing that evaluates hot-spot and RTM models reported an accuracy of around 25% when everything occurring in a 400 by 400-foot cell was considered "near" and up to 68% with the size parameter set to 800 foot. PredPol itself focusses on cells of 500 by 500-foot (Hunt et al. 2014: p. 35).

According to Drawve (2014: p. 21), the benefit of true positives outweighs these losses. Therefore, one could argue that a lack of accuracy when predicting the location of crimes as well as false positives only create financial harm, because officers go out to prevent a crime that is not going to happen anyway. While this might hold true with regard to the police side of the events, those effected by wrongful assumptions often feel at least uncomfortable. News reports (Gorner 2013) criticized that the heat list used in Chicago contains a lot of false-positives leading to false accusations and reported embarrassing visits by police officers. A more data-based evaluation of the effects of the program showed that individuals on the list are "not more or less likely to become a victim of a homicide or shooting than the comparison group" (Saunders et al. 2016: p. 1). Instead, they found that those individuals on the list were more likely to be arrested for gun violence, not because of the increased risk they pose, but because of the fact that officers, who investigated shootings, turned to the list as a way to find possible.

An anecdotal example from a news report about the threat scores mentions that a person was given a "yellow" threat level most likely because of who lived at that same address before he had moved there. The author concludes: "Even though it's not me that's the yellow guy, your officers are going to treat whoever comes out of that house in his boxer shorts as the yellow guy" (Jouvenal 2016). A study, which compared nine tools that calculate individual risk, found their accuracy to be limited and concluded that "after almost five decades of developing risk prediction tools, the evidence increasingly suggests that the ceiling of predictive efficacy may have been reached with the available technology" (Yang et al. 2010).

3.3 Effects of how data are created

A drawback of any evaluation, like those described above, is that it depends on the specific data used for the evaluation. After all, the—unavoidable—assumption is that the classifier will generalize to unknown and future data from the known past data, as it has been shown to generalize from training data to test data. It is also considered necessary to show not only an evaluation on one dataset, but on multiple because the danger of overfitting, being too

closely determined by the training data, is always present. Therefore, the second set of questions revolves around these data. Barocas and Selbst (2016) presented a set of relevant questions to be asked of data, shown in italics in the following paragraphs and extended/adapted to our domain of interest.

In general, what a model learns depends on the examples to which it has been exposed. But how are these examples defined in the first place? *What is the target variable, what are the class labels*, and how and by whom are they assigned to instances (such as behaviours or people)? In some policing-related data, this is relatively straightforward: a behaviour or person is criminal/guilty if and only if this has been established by the criminal justice system, and not criminal/not guilty otherwise. But this very ontological dependence on performative speech acts points to the pragmatic challenges of *labelling the examples in the training (and test) data:* who did the labelling, and under what circumstances or constraints?

First, definitions themselves may be treacherous, as an example from drone warfare illustrates: if "militant" is essentially defined as "any military-age male whom we kill, even when we know nothing else about them"[9], high success rates are guaranteed when the question is whether drone strikes killed militants. Second, and returning to an example from predictive policing scenarios, police officers in the field have to instantiate legal terms that are by definition indeterminate, i.e. label examples. Ferguson (2015) illustrates how the introduction of "Big Data" access and analytics may shift the instantiation of "reasonable suspicion" (which is needed under US law to stop and search a person), and how this may run counter to the purposes of laws formulated when police were operating under different epistemological conditions. Third, external conditions may influence the labelling, as when police units have to meet certain performance indicators (e.g. number of expulsions from public places) or when private prison organisations lobby to keep incarceration rates high (Justice Policy Institute 2011). Recent research has also shown that data produced by police departments are often *s*ubject to manipulation (Eterno et al. 2016).

According to the PredPol homepage, the current algorithm makes use of only three variables: crime type, crime location, and date and time, to predict future crimes on maps in rectangles. In the study presented by Mohler et al. (2012), the data used to perform the analysis were from the San Fernando Valley in Los Angeles, a residential area on flat terrain with a rectangular street grid and predominantly detached housing. This reduces the need to take additional aspects of the area into consideration, so the algorithm solely considers the parameters time and space. Geographical factors like streets or natural barriers are not considered, although research has shown that while some offenders commit crimes close to their home, others commute, raising the importance of the environment in predicting future crime scenes (Meaney 2004; Trotta 2010). Similarly, Townsley et al. (2003) found in a study on burglaries in Brisbane that a large housing homogeneity is of great importance for the near repeat theory.

Additional constraints regard the accuracy of the data that is used to train the models. As noted above, the near repeat theory depends on the distance (in space and time) between two events. If these are not accurate, e.g. because police officers do not report addresses or the system is not able to translate these addresses into geo-locations, the accuracy of the predictions is reduced. Short et al. (2009) found that their algorithms proved to be useful when "near" was defined as less than 100 m, while crimes seemed to be randomly distributed when this parameter was set to 4000 m. In addition, studies on the near repeat theory and burglaries have shown that timing is an important factor: many burglaries occur within a few hours in close proximity. The accuracy of the predictions, therefore, also heavily depends on how quickly victims report those crimes.

Similar restrictions apply for RTM. Tools such as Hunchlab heavily require accurate and detailed maps of a city. Police departments using this technology, therefore, need to constantly update their dataset and inform the algorithms about any change of usage of buildings, construction sites or the location of events. And the calculation of individual risks also depends on up-to-date information. The anecdote on the outdated threat level above shows that a lack of accuracy in the data set (here knowing that the inhabitant of a building has changed) might have negative impacts on new residents. Besides the constant data flow, those operating predictive policing tools also have to make sure that the reports used as input convey a similar understanding of what happened. Papachristos (2009) notes, for example, that every police department has a different understanding of what is considered a gang-related homicide.

Further, biasing effects can arise from *data collection* generating incorrect, partial or nonrepresentative data. For example, Lum and Isaac (2016) found that PredPol, which makes use of past police reports, over-proportionally targeted poor black neighbourhoods, but not affluent white neighbourhoods that, according to an independent drug-usage study, are likely to have similar amounts of drugs.

Related to this is the *feature selection in data collection*, i.e. which attributes are observed. Model classes often (co-)determine the selection of attributes. For example,

[9] See http://www.nytimes.com/2012/05/29/world/obamas-leadership-in-war-on-al-qaeda.html?pagewanted=1&_r=2 and https://firstlook.org/theintercept/2014/11/18/media-outlets-continue-describe-unknown-drone-victims-militants/.

predictive analytics based on traditional tabular data such as those in the toy example of Tables 1 and 2 focus on individual demographics and behaviour, while social-network analyses such as the Chicago heat list consider individuals to be strongly influenced by their peers.

Summing up, in applications one should ask (a) how constructs are defined and operationalised, (b) how data are collected, and (c) whether there is a validated criminological theory that makes the choice of features and models plausible.

3.4 Assumptions baked into the learning algorithm

By definition, predictive analytics not only describe and summarise: they also generalise. The question then becomes *how* they generalise, and this is where a popular high-level statement, "the data speak for themselves" (Anderson 2008), is shown to be wrong. Every learning algorithm has an *inductive bias*.

For example, many learning algorithms have the inductive bias to favour simpler hypotheses. This is known from general philosophy of science as "Occam's razor" (Blumer et al. 1990). In our toy example, the simple hypothesis that all green-skinned individuals are criminals, would be 67% accurate on the training data. Another popular bias is maximum conditional independence, i.e. the assumption that factors work independently of one another in contributing to their effect (such as making someone likely to commit a crime).

As described above, PredPol is based on only three factors, although other theories with a larger number of factors have also proven useful. Researchers associated with PredPol admit that their algorithm exploits a "mechanistic explanation" of human behaviour: "if an offender encounters a target in the absence of an effective security measure (inhibitor), then he is free to exploit that target" (Short et al. 2010: p. 1). This rationalistic approach is the reason why the near repeat theory is only able to predict crimes that are guided by some determinants, while all crimes, even of the same type, that are not planned are omitted by the algorithms. This inductive bias can also explain why a similar algorithm, applied to other datasets, worked on data in Chicago (gang violence and burglaries) and terrorist attacks in North Ireland, but did not work for Terrorist attacks in Israel and Falludscha (Mohler et al. 2013).

And inductive bias can also be found in the other algorithms. Risk Terrain Modelling assumes only factors that can be defined geographically, while they actually serve as a proxy for assumed group characteristics. For example, with regard to rape, RTM would assign higher risks to campuses, where there is a larger number of women age 16–24, and a proximity to fraternities or places where athletic teams meet (Caplan and Kennedy 2011: p. 61). Though there is a correlation between a campus and the types of people that use it, the increased risk is more social than a geographical feature.

The pragmatic use of classifiers may require one to take into account the differential real-world costs (financial or otherwise) of making different errors. For example, in terrorism prediction, a false negative can be far costlier (in terms of a non-averted attack and its victims) than a false positive (such as the body-search of an innocent person), while for minor offences such as pickpocketing the reverse may be the case. *Cost-sensitive classification* guides a classifier towards regions in the solution space that have fewer of the costly mistakes, thereby also affecting the inductive bias.

Inductive bias is unavoidable, but it expresses assumptions and has consequences. Knowing it helps to assess and question algorithms deployed in predictive policing. Current moves towards more explainable models (e.g., Zeng et al. 2016; Lipton 2016) aim at creating less opaque models, which may also involve more transparency about inductive bias.

3.5 Knowledge-discovery and real-life processes

Neither the collection of data nor the action upon receiving a classifier's predictions operates in an abstract space. The constraints of the real world often not only reduce their effectiveness, but raise questions about whether they should be applied at all.

In their study of Shreveport, Lousiana, Hunt et al. (2014) did not find statistical evidence of a long-term positive effect of their predictive policing approach on crime, though they showed a reduction in costs. The study not only gives several explanations including the use of the wrong algorithms, but also offers alternative explanations such as organisational or study related issues. Similar results might have led the German city of Nuremberg to discontinue the use of a near-repeat based software for burglaries after 6 months in which the success of the software quickly evaporated. News reports cited that the police "saw many burglaries that Precobs did not predict. The burglars looked for new areas and did not behave as expected, though they were clearly serial offenders" (Biermann 2015, own translation). Another recent example of the usage of PredPol in the US underlines that there is a minimum requirement for data; therefore, small cities are less likely to see positive effects.[10] For the heat list

[10] After 3 years a city with around 70,000 discontinued a contract with PredPol with the police chief commenting that "PredPol system may have greater benefit to law enforcement organisations policing much larger geographical jurisdictions where greater variables in crime patterns may exist", while in Milpitas "existing internal processes of tracking crime and identifying potential areas of exposure were often more accurate than results received from PredPol". (Ian Bauer, The Mercury News, 14.07.2016, accessible online at http://www.govtech.com/public-safety/Milpitas-Calif-Police-Department-Nixes-Predictive-Policing-Contract.html).

Saunders et al. (2016) showed its inefficiency.[11] Still, those in favour of predictive policing argue that the maps created by algorithms are more accurate than those created by humans (Mohler 2015).

While the use of crime mapping for tactical analysis within police departments is undisputed, many predictive policing approaches also raise privacy concerns as they require additional data collection that involves the physical observation and questioning of people and neighbourhoods, or the intrusion into private spaces comprising the home or private virtual spaces. These activities in themselves may interfere with fundamental rights such as privacy or data protection (Solove 2006; Berendt 2012), and they, therefore, must pass a proportionality test: are they suitable, necessary, and appropriate in relation to the severity of what they try to prevent? Specifically, in relation to "Big Data", the learning of a good classifier requires the collection of *a lot* of data including negative examples, which may lead to a normalisation of comprehensive surveillance regardless of suspicion (Coudert 2015)—a strategy that has come under increasing scrutiny after the Snowden revelations, and that has recently been declared to violate EU law (ECJ 2016).

Similar interferences may obtain, and similar tests should, therefore, be applied when police act upon a prediction. Thus, visits of Chicago police to suspects on their predictive-policing heat list may be measured with the good intention of prevention, but they are also intrusive, are visible to neighbours, and can thereby have various undesired effects (Stroud 2014).

4 Conclusion

Predictive policing technology enjoys positive media coverage and a growing interest by police departments in different countries. We have shown that the positive effects of currently available technology are often exaggerated, and more importantly, that the underlying theories are often not sufficiently substantiated by the evidence. We discussed a number of problems that are connected to the very basic ideas of data mining and knowledge discovery in big data and analysed how they relate to current implementations of predictive policing.

We showed that, instead of a Robocop consultant that can help with every crime, the products available implement highly specialized algorithms that each inherit the bias of its underlying hypotheses. While they are mostly rated by their effectiveness to predict future crimes, they also need to be judged by the implications they have. For this, the software implementing the algorithms has to be publicly available and processes need to be established that allow police to intervene in the data processing. We also highlighted that data mining always depends on the data it is performed on. Since policing is never only about crime the data that is produced by the police may be flawed, by inaccuracies, (implicit) biases or purposeful manipulations that pursue secondary goals.

In addition to these computational considerations, predictive policing measures have to be assessed from a legal perspective. The collection of data, profiling, and the prevention of crimes often interfere with fundamental rights, such as privacy, data protection, and freedom of movement. The preventive measures have to fulfil a number of conditions: any interference with [a fundamental right] must, in addition to being 'prescribed by law' and having a 'legitimate aim', also be proportionate. This means that the interference must pass the following three-part test:

(1) a 'suitability' test, which evaluates whether the measure is reasonably likely to achieve its objectives (effectiveness);
(2) a 'necessity' test, which evaluates whether there are other less restrictive means capable of producing the desired result (least intrusive means); and
(3) a proportionality test 'stricto sensu', which consists of a weighing of interests whereby the consequences on fundamental rights are assessed against the objectives pursued (balance of interests)." (Van Alsenoy et al. 2013: p. 70). In addition, the measures must only serve the prevention of crimes, i.e. not go beyond the scope of the police's tasks.

As Solove (2011) explains, problems arise when, in the interest of "security", due process and the rule of law are disregarded—and these are well-known strategies of power, which are not limited to the realm of "Big Data", but that remain as problematic as they have always been also with "Big Data". To take the influence of technology into account, established methods have to be developed further, as outlined in Citron's (2007) proposals for *technological due process*.

References

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. Wired Mag 16(7):16

Azavea (2015) HunchLab under the hood. https://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.pdf. Accessed 15 May 2017

Barocas S, Selbst AD (2016) Big Data's disparate impact. 104 California Law Review 671. http://ssrn.com/abstract=2477899

Berendt B (2012) More than modelling and hiding: towards a comprehensive view of web mining and privacy. Data Min Knowl Discov 24(3):697–737

[11] The evidence is cynically supported by the fact Chicago saw huge surge in number of homicides in 2016. http://crime.chicagotribune.com/chicago/homicides.

Biermann K (2015) Predictive policing: noch hat niemand bewiesen, dass data mining der Polizei hilft. *Die Zeit*, March 29, sec. Digital. http://www.zeit.de/digital/datenschutz/2015-03/predictive-policing-software-polizei-precobs/komplettansicht

Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1990) Occam's Razor. In: Readings in machine learning. pp 201–204

Caplan JM, Kennedy LW (2011) Risk terrain modeling compendium: for crime analysis. Rutgers Center on Public Security, New Jersey

Chainey S, Ratcliffe J (2013) GIS and crime mapping. Wiley, Hoboken

Citron DK (2007) Technological due process. Wash Univ Law Rev 85:1249–1313

Coudert F (2015) 'Precrime police' is not for 2054, it's for now: how to regulate 'data intensive policing'? In: Amsterdam privacy conference, Amsterdam, 23–26 October 2015

Drawve G (2014) A metric comparison of predictive hot spot techniques and RTM. Justice Q 33:1–29. doi:10.1080/07418825.2014.904393

ECJ/Court of Justice of the European Union (2016). Judgment of the Court (Grand Chamber) of 21 December 2016. Tele2 Sverige AB v Post- och telestyrelsen and Secretary of State for the Home Department v Tom Watson and Others. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62015CJ0203

Eterno JA, Verma A, Silverman EB (2016) Police manipulations of crime reporting: insiders' revelations. Justice Q 33(5):811–835. doi:10.1080/07418825.2014.980838

Ferguson AG (2015) Big Data and predictive reasonable suspicion. Univ Pennsylvania Law Rev 1632:327–410

Gorner J (2013) Chicago police use heat list as strategy to prevent violence. Chic Tribune 21:2013. http://articles.chicagotribune.com/2013-08-21/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristos-heat-list

Hunt P, Saunders J, Hollywood JS (2014). Evaluation of the shreveport predictive policing experiment. Research Report RR-531-NIJ. RAND Corp. http://www.rand.org/pubs/research_reports/RR531.html

Johnson SD (2008) Repeat burglary victimisation: a tale of two theories. J Exp Criminol 4(3):215–240. doi:10.1007/s11292-008-9055-3

Jouvenal J (2016) The new way police are surveilling you: calculating your threat 'score'. The Washington Post, January 10. https://www.washingtonpost.com/local/public-safety/the-new-way-police-are-surveilling-you-calculating-your-threat-score/2016/01/10/e42bccac-8e15-11e5-baf4-bdf37355da0c_story.html

Lipton ZC (2016) The mythos of model interpretability. In: ICML 2016 workshop on human interpretability in machine learning (WHI 2016). http://zacklipton.com/media/papers/mythos_model_interpretability_lipton2016.pdf

Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19

Mayer-Schönberger V, Cukier K (2013) Big Data: a revolution that will transform how we live, work and think. John Murray, London

McCarthy GF (2013) Custom notifications in Chicago—pilot program. 13–080 TRH. https://web.archive.org/web/20160414031801/, http://directives.chicagopolice.org/directives-mobile/data/a7a57bf0-13fa59ed-26113-fa63-2e1d9a10bb60b9ae.html

Meaney R (2004) Commuters and marauders: an examination of the spatial behaviour of serial criminals. J Investig Psychol Offen Profil 1(2):121–137. doi:10.1002/jip.12

Mohler G, Short M (2012) Geographic profiling from kinetic models of criminal behavior. SIAM J Appl Math 72(1):163–180. doi:10.1137/100794080

Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2012) Self-exciting point process modeling of crime. J Am Stat Assoc. doi:10.1198/jasa.2011.ap09546

Mohler G et al (2013) Modeling and estimation of multi-source clustering in crime and security data. Ann Appl Stat 7(3):1525–1539

Mohler GO, Short MB, Malinowski S, Johnson M, Tita GE, Bertozzi AL, Brantingham PJ (2015) Randomized controlled field trials of predictive policing. J Am Stat Assoc. doi:10.1080/01621459.2015.1077710

Morozov E (2013) To save everything, click here: technology, solutionism, and the urge to fix problems that don't exist. Allen Lane, London

Papachristos AV (2009) Murder by structure: dominance relations and the social structure of gang homicide1. Am J Sociol 115(1):74–128. doi:10.1086/597791

Perry WL, McInnis B, Price CC, Smith S, Hollywood JS (2013) Predictive policing—the role of crime forecasting in law enforcement operations. Research Report RR-233-NIJ. RAND Corp. http://www.rand.org/pubs/research_reports/RR233.html

Saunders J, Hunt P, Hollywood JS (2016) Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. J Exp Criminol. doi:10.1007/s11292-016-9272-0

Short MB, D'Orsogna MR, Brantingham PJ, Tita GE (2009) Measuring and modeling repeat and near-repeat burglary effects. J Quant Criminol 25(3):325–339. doi:10.1007/s10940-009-9068-8

Short Martin B, Jeffrey Brantingham P, Bertozzi Andrea L, Tita George E (2010) Dissipation and displacement of hotspots in reaction-diffusion models of crime. Proc Natl Acad Sci 107(9):3961–3965. doi:10.1073/pnas.0910921107

Solove D (2011) Nothing to hide. The false trade-off between privacy and security. Yale University Press, Yale

Solove DJ (2006) A taxonomy of privacy. Univ Pa Law Rev 154(3):477. GWU Law School Public Law Research Paper No. 129. https://ssrn.com/abstract=667622

Stroud M (2014) The minority report: Chicago's new police computer predicts crimes, but is it racist? The Verge, 19 Feb 2014. http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

Townsley M, Homel Ross, Chaseling Janet (2003) infectious burglaries. A test of the near repeat hypothesis. Br J Criminol 43(3):615–633. doi:10.1093/bjc/43.3.615

Trotta M (2010) Serial offenders' spatial behaviour: revisiting the marauder/commuter dichotomy. In: Presented at the 10th conference of the european society of criminology, Liège. Retrieved from http://orbi.ulg.ac.be/handle/2268/73623

Van Alsenoy B, Kuczerawy A, Ausloos J (2013) Search engines after 'Google Spain': Internet@Liberty or Privacy@Peril? TPRC 41. In: The 41st research conference on communication, information and internet policy. http://ssrn.com/abstract=2321494

Witten IH, Frank E, Hall MA (2011) Data mining. Practical machine learning tools and techniques, 3rd edn. Morgan Kaufmann, Burlington

Yang M, Coid J (2010) The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. Psychol Bull 136(5):740–767. doi:10.1037/a0020473

Zeng J, Ustun B, Rudin C (2016) Interpretable classification models for recidivism prediction. In: Presentation at FATML 2016. New York, 18 November 2017. https://arxiv.org/abs/1503.07810