Version 13 January 2014
This is a pre-edited version of a book section. The final version is published as:

Grommé, Francisca. 2016. "Data Mining 'Problem Youth': Looking Closer But Not Seeing Better." In *Digitizing Identities: Doing Identity in a Networked World*, edited by Irma Van der Ploeg and Jason Pridmore, 163–83. London: Routledge.

Data Mining 'Problem Youth': Looking Closer But Not Seeing Better

ABSTRACT. This chapter examines how a Dutch city attempted to use data mining to profile 'problem youth'. It challenges zooming as the metaphor that guides the use of this statistical technique. To see something from close by, it is argued, is a situated practice. Instead of presenting the object in more detail, a new object is brought into being. Two modes of situated improvisation are identified. These involve the interplay of artefacts, bodies of knowledge, and normativities. Attending to metaphors in practice may thus be a useful starting point to change the terms by which digital identities are produced.

Introduction

In 2011, the Dutch municipality of Burgcity conducted a pilot study about data mining. It aimed to find out whether this statistical technique could be used to improve its understanding of 'problem youth', loosely defined by the city as youth below twenty-three years of age who are likely to commit minor offences such as vandalism, littering or shop theft.[1] In particular, it aimed to learn whether a combination of municipal data, police data and commercial data about consumption could lead to new insights for youth crime policy.

Especially salient in this pilot study was the policy makers' use of the metaphor of 'zooming in'. The policy makers expected data mining to provide knowledge that was local, particular and timely. This understanding of data mining is not unique to Burgcity. Indeed, proponents of data mining promise increased detail and granularity. As was stated in a project plan, data mining techniques would generate "local theories" that "the general theories of social science" cannot provide.

Vision metaphors are never innocent, however, as Donna Haraway has famously argued (1991). The metaphor of zooming draws on the imagery of a mechanical lens, which reveals more detail about an object. The resulting 'close-ups' are assumed to have a high truth status because of their implied precision. In this chapter, I am interested in how digital data are put to use according to a rationale of zooming in to profile problem youth. Profiles can be understood as identities that are ascribed to youth and used a basis for government intervention (Pridmore, this volume).

I aim to challenge zooming as an underlying principle in data mining practices and scholarly and professional accounts. I do so because sticking to this metaphor

[1] My informants used 'problem youth' and 'at risk youth' intertwiningly. To avoid confusion, I will only refer to 'problem youth'. Although I will henceforth use the term without quotation marks, I emphasize that it is by no means a natural category and that applying this category can negatively affect individuals and groups (see Bowker and Star 1999). Some of my informants shared this view.

risks mobilization of data mining in an argument for the acquisition of ever more personal data. Namely, if a technology by itself can provide ever more detailed representations of youth, all it needs are more or better data. Furthermore, the metaphor obscures the normativities that are part of the practice of data mining. With this I mean that it diverts attention from the 'goods' and 'bads' that come into play when 'detailed knowledge' is produced.

My questions are how zooming in was done in the Burgcity data mining pilot and what norms were embedded in and produced through these practices. I attend to the bodies of knowledge, discursive practices and artefacts that were part of zooming in. Adopting a material-semiotic approach, I focus on the heterogeneous relations that bring objects, such as problem youth, into being (Mol 2002; M'charek 2013). From this approach, it follows that by zooming in one does not simply see the same thing in more detail. Instead, the practices of zooming in bring new objects into being (Strathern, 2005). I draw on Charles Goodwin's work to describe data miners' practices as 'situated improvisation' (1995; 1996).

This chapter is based on fieldwork that I conducted before, during and after the Burgcity pilot for a period of eighteen months. I focus on the interactive sessions at the core of the pilot, in which policy makers and corporate experts analysed the data. We will follow the participants through various attempts to zoom in. First, from zooming in to the level of the sub-city district, we learn about two modes of situated improvisation the participants engaged in: evocation and comparison. Second, from zooming in to the level of the neighbourhood we learn from the trouble that the participants ran into: they could not zoom in any further without losing sight of the problem youth. This part if the pilot draws out the regimes of evidence that are part of zooming in.

Attending to the practicalities of zooming in also allows us to learn that it is normative work. I show that results needed to be made relevant as surprises; that the 'lens' focused on the neighbourhood as a source for truth; that city youth were constituted as a norm; that detail is produced by the application of general categories; and that zooming in includes making judgements about good knowledge for government. These normativities suggest that taking seriously the metaphors by which technologies are brought into practice might be a good starting point to change the terms by which digital identities are produced.

Zooming In as Situated Improvisation

Mining for Local Knowledge

Ever more digital data are available for analysis. People produce increasing amounts of data through activities as simple as browsing on the internet and using a chip card on public transport. This development is joined by a growing capacity to search and analyse these data. Data mining is a statistical technique often used for the analysis of big datasets. It is commonly referred to as "the automatic or semi-automatic process of discovering patterns in data" (Witten, Eibe, and Hall 2011, 5), or "the application of specific algorithms for extracting patterns" (Fayyad, Piatetsky-Shapiro, and Smyth 1996, 39). In everyday usage, it can refer to both software and analytical skills.

Data mining is argued to challenge traditional science, because, in contrast with statistical techniques such as regression analysis, the software allows the analyst to search for relations in the data without defining hypotheses and limiting the number

of variables in advance (Witten, Eibe, and Hall 2011; Hildebrandt 2008). Instead, an algorithm is applied to automatically find co-occurrences in a large dataset that can comprise thousands of variables.[2] Industry therefore advertises data mining as 'digging' into the data to find 'nuggets of gold' (yet other metaphors).

It needs to be noted that, although data mining provides analysts with new possibilities, it is often practiced as a combination of old and new statistical techniques. In the application studied in this chapter, for instance, conventional geodemographic marketing techniques are combined with data mining algorithms.[3] Furthermore, in practice, patterns are often not found automatically, but rely on the expert's insight to choose variables (Ang and Goh 2011). Following boyd and Crawford, it therefore seems appropriate to approach data mining not as a "higher form of intelligence and knowledge that can generate insights that were previously impossible", but as a mythology as well as a technological development (2012, 663).

In policy practice, data mining is most often used to create profiles: sets of correlations that can be used to identify or represent individuals or groups (Hildebrandt 2008, 19). It is applied in a variety of policy domains, such as anti-terrorism programs, programs against tax evasion, welfare policy and, as this chapter describes, local crime policy. In these cases, the promise of data mining is specificity: more data and diverse data sources generate a closer view. Policy makers are seldom interested in general patterns, policy ethnographies show (Choy 2005; Yanow 2002). They are in search of particularities, and commercial data mining is brought in for this purpose. This is also reflected in the marketing analytics industry, which has specialized in the provision of ever more locally specific knowledge (Phillips and Curry 2003).

Empirical scholarly work on the use of data mining and large datasets by (local) policy agencies mainly describes the limitations of data mining. It is argued that digital data do not necessarily lead to descriptions of individuals and groups that are more precise. In Gary T. Marx's words, users of this technology "see hazily (but not darkly) through the lens" (2005, 339). A fundamental reason is that profiles give very little information about individuals and small groups of people since they aggregate data. The characteristics of the profile may therefore not be applicable to all the individuals it represents (Curry 2004; also see Custers 2004). It is also argued that the data do not represent individuals correctly because government datasets are often incomplete or simply wrong (Pleace 2007). Moreover, not all types of large datasets are valuable or reliable. This is the case for some social media data, such as statements on Twitter (boyd and Crawford 2012). In acquiring data, furthermore, local governments are limited by privacy restrictions. An example is the use of police data. Lastly, the possibilities of acquiring commercial datasets are often limited for financial reasons and because companies are not always allowed or willing to share them (boyd and Crawford 2012; Pleace 2007).

Localized and Embodied Visions

[2] In Burgcity, Data Inc. applied a nearest-neighbour associative algorithm. This algorithm is understood to be suitable for a wide variety of practical applications because it allows for the analysis of databases with many missing values.

[3] Some would argue that this is not 'real' data mining. In using the term data mining I adopt my informants' terminology.

These considerations put data mining into perspective. The reason I raise them here, however, is because they bring to mind a similar issue in anthropology. In *Partial Connections* (2005), Marilyn Strathern identifies the problem that there never seems to be sufficient data to adequately describe social life. Whenever one looks closer or from another angle there still does not seem to be enough data. One would expect complexity to decrease when a smaller part of society is examined; instead, there always seems to be something new to discover. The similarity between these two practices – data mining and anthropological observation – is that they are both based on the premise that seeing better requires more or better data.

Strathern takes issue with the idea that there are an infinite number of perspectives, each requiring more data. She argues that all perspectives exist as "localized, embodied visions" (40). These visions do not present different versions of reality, but enact the object in different, partially connected ways (also see Mol 2002). To change perspective, therefore, does not amount to seeing part of a whole. For zooming in this means that one does not see a part of the same object in more detail. When scientists, professional analysts, policy makers or others zoom in, they bring an object into being.

This is a relevant distinction because it avoids the notion that technologies by themselves reveal the characteristics of populations that exist independently of the practices by which they are mapped (M'charek 2005; Ruppert 2011). This notion of technology is problematic in academic discourse because it feeds into the idea that lack of data or analytical power obstructs perfect close visions. It therefore fails to question data mining in terms other than those of an information problem. I suggest examining how zooming in is done in practice in order to overcome this limitation.

Situated Improvisations

To be sure, to insist on a critical attitude towards metaphors of vision is not new.[4] Donna Haraway is especially known for her contribution to this project. In *Situated Knowledges* (1991), she argues that, with contemporary visualizing technologies, "Vision in this technological feast becomes unregulated gluttony; all perspective gives way to infinitely mobile vision, which no longer seems just mythically about the god-trick of seeing everything from nowhere, but to have put the myth into ordinary practice" (189). Claims for scientific objectivity are based on a 'view from nowhere', thus creating a science that is detached and totalizing. What is needed, Haraway argues, is a commitment to situated knowledges: knowledges that are understood as local, embodied and partial.

Inspired by this body of work, Christopher Gad and Peter Lauritsen (2009) suggest that social scientists with an interest in surveillance practices such as digital identification should study them as situated phenomena. To observe something, actors in the field of surveillance need to draw together bodies of knowledge, artefacts and human actors at a certain place and time. Surveilled objects thus achieve their status in terms of the relations between heterogeneous entities (M'charek 2013).

I use this notion of situatedness to study zooming in as a metaphor comparable to Haraway's vision metaphor of 'seeing everything from nowhere'. Goodwin's work

[4] Surveillance Studies has had its own issues with the Panopticon, a model for visual surveillance that Michel Foucault (1994 [1975]) adopted as a theoretical model (see for instance Lyon (2006)). In this chapter, I am less concerned with theoretical models in social science, and more with knowledge practices.

about situated improvisations (1995) provides a point of departure for understanding how the participants of the Burgcity pilot zoomed in on youth. Goodwin coined the term to describe how an object of knowledge emerges through the interplay of screens, common knowledge, lay theories, everyday artefacts and professional repertoires (606). Importantly, with improvisation I do not refer to random actions, but to actions that are informed by bodies of knowledge, artefacts and emerging priorities at a certain place and time.

Following Goodwin, I will attend to talk and gestures by which the relations between these heterogeneous entities were achieved. For instance, Goodwin describes how marine biologists highlighted the information presented to them on a screen by pointing at it with a pencil or by making comments such as "that nice feature again" (262). In his work on archaeology, he shows that some of the talk makes some results more relevant than others; it sets out the conditions of relevance (1996). Furthermore, from Goodwin's analysis of the Rodney King trial we learn that some actors may use the status of their profession to influence enquiries. This is what he refers to as professional privilege (1996).

Data Mining Problem Youth: A Pilot Study

Marketing Intelligence Methods

Burgcity is a medium sized Dutch city of about two hundred thousand inhabitants. The pilot study took place at the Department of Community safety, which employed sixteen persons at the time I was present. Data Inc. is a data analyst company employing no more than ten persons. It had mainly accepted assignments from the Dutch police force and corporations in the Netherlands. Data Inc. provided the software, collected and cleaned the data, and performed the analyses.

The case description and the case study that follow are based on observations of the pilot meetings, the city's policy makers' everyday routines, and interviews. As the reader might suspect, Burgcity and Data Inc. are fictitious names, as are the names of all other organizations, places and persons mentioned in this chapter (with the exception of the international corporation that delivered the commercial data used in this pilot: Experian). I agreed to anonymity because confidential information regarding suspects, their backgrounds and crime control practices often passed my eyes on the Burgcity work floor. I also agreed because I felt a responsibility to protect the livelihood of my informants. I believe that this is an especially sensitive issue because both the company and the city were involved in an experimental activity, making themselves vulnerable as they treaded uncertain ground.[5]

The pilot study's aim, as stated in Data Inc.'s project proposal, was to develop "more efficient and effective approaches"[6] to non-criminal nuisance and minor criminal offences committed by youth. The use of the Experian dataset, which contains 'lifestyle data', was advertised as the pilot's main innovative feature. 'Marketing intelligence methods' were to lead to new insights into problem youth, their life worlds and motivations. If one knows that the "most troublesome youth" wear Nikes, a new approach in policy might be to involve Nike stores in a campaign against vandalism, one of the project's initiators argued in a 2010 presentation.

[5] For this reason I have chosen not to include the full titles of the policy documents that I cite in the bibliography. Instead, I mention the document types in the text.

[6] All fieldwork quotes have been translated from Dutch by the author.

The Experian lifestyle dataset incorporates about fifteen hundred variables on basic information about household income and age; interests and activities, such as membership to a sports club; and media usage, such as internet usage and newspaper subscriptions. These variables also include pre-set profiles, referred to as Mosaic groups, such as 'free spirits', 'mini machos' or 'digital families'. The data come from sources such as property or telephone registrations and surveys. Experian datasets are sold for marketing purposes, to develop strategies for the collection of bills and debts, and to manage financial risk.

These data were combined with police data about suspects and offences and with municipality data including variables such as age, income and school attendance. A remark about the police dataset is in order here. It contains records about individuals suspected of one or more offences. Even if police investigation has pointed out that an individual was not involved in an incident, he or she remains registered. For the purposes of the pilot, young suspects were used as an approximation of problem youth, which raises the obvious issue of whether these data can really help Burgcity to learn about the group it is interested in.

Data mining was expected to help the city understand problems more "thoroughly and quickly", understand patterns that characterize problem youth at neighbourhood level, identify the groups that need attention, and identify and influence causal relations. Thus, it would allow the city's policy makers to "zoom in" on local youth and generate "local theories" (Interview, former CEO Data Inc., November 12, 2010). In policy maker Anna's words, it could be another method to "reach out" to Burgcity's inhabitants (Interview, September 5, 2011).[7]

Pilot Results

Prior to the pilot, Burgcity's Department of Community Safety had already considered data mining to improve the city's 'information position'. As the department's information specialist put it, "data mining is definitely going to happen in Burgcity, there is no doubt about it. The head of our department asks me how data mining is coming along every week" (Interview, April 21, 2011). So, when the department head was asked by an innovation platform to participate in the pilot, he consented. The funding was supplied by a grant from the Dutch Ministry of Internal Affairs.

The innovation platform chose to focus on problem youth, which was and still is a hot topic in Dutch crime policy. Next, it was up to Burgcity to select a case study. It chose its newest and demographically youngest city district, Molendistrict, as a case study. Molendistrict has about forty thousand inhabitants. The district seemed to suit the purposes of the pilot because one third of Burgcity's youth lives here and nuisances in the area are of considerable concern for the city's staff at the Department of Community Safety.

Data Inc. insisted that the policy makers' input was crucial to asses which results were of value for Molendistrict. Three of Burgcity's civil servants were involved in the pilot: Mieke, Anna and Liesbet. Mieke is the department's information specialist, Anna the department's specialist on youth, and Liesbet the district manager of Molendistrict. The group would ideally explore the possibilities of the software

[7] The participants focused on finding patterns. They did not aim to introduce an automated signal.

together to learn what insights could be gained by using it. Data Inc. suggested taking an "iterative approach", whereby Burgcity's policy makers would formulate questions, find answers and "learn how to ask better questions" (Fieldwork notes, February 28, 2011).[8]

Next to preparatory and strategic meetings, the city's civil servants and Juriaan, Data Inc.'s analyst, met four times for interactive sessions in Burgcity's city hall. During these sessions, Juriaan operated a laptop and the city employees faced him. Everybody in the small meeting rooms could see the computer interface on an LCD screen. During the meetings, a question for analysis would be formulated by the policy makers or by Juriaan. Juriaan would then attempt to answer the question, showing the results on the screen.

Three types of results were generated. First, frequency tables presented counts, for example, the number of suspects living in each neighbourhood. Second, profile analysis was used to compare two or more populations. The resulting table displayed the variables for which significant differences between the populations were found. Third, the data mining system was used to generate 'hotspot' maps: clusters of offences committed in a certain location on the map.

The pilot did not result in a set of stable and generally acknowledged conclusions about problem youth and youth crime policy. Data Inc. did present a number of conclusions in the final report, amongst others about the types of neighbourhoods that have a relatively high risk of youth delinquency, such as neighbourhoods in which household income varies strongly (Data Inc. 2012, internal document). In the end, however, the report stated that it could not characterize problem youth using the lifestyle data due to privacy restrictions (on which I will elaborate later in the chapter).

In what follows, I describe how zooming in was done and what bodies of knowledge and objects informed the efforts to obtain a closer view. I attempt to understand how, once the results had been presented on the screen, a close view of youth "emerged through the interplay between a domain of scrutiny (…) [and] a set of discursive practices (…), being deployed with a specific activity" (Goodwin 1996, 606). Two themes that regularly returned in the sessions are presented to the reader: the place of residence of youth that commit offences in Molendistrict and the use of lifestyle data to characterize youth that commit offences. I have chosen to elaborate on these themes because they were discussed most frequently and thoroughly during the sessions, and therefore are the most instructive about the practice of zooming in. For each of these themes, I focus on the participation of one of the policy makers.

Zooming In from City to District

Surprise

The first story is about Liesbet's question for the technology. It demonstrates the particularities of zooming in from the city to the district level.[9] Liesbet is the district manager for Molendistrict. It is her job to facilitate communication between the central city and Molendistrict's local politicians, interest groups and case workers. As

[8] This process is referred to as the CRISP-DSM method, see Custers (2004).

[9] The levels that my informants worked with are, from high to low: city, (sub-city) district, postal code, neighbourhood, sub-neighbourhood, postal code-6 (about twenty households), individual.

she puts it, she defends the district's interests at city level. Of all the participants in the pilot, she visits the district most frequently.

This is what data mining should answer for Liesbet: "Where in Molendistrict do the young people who commit offences live? And how can useful policies be developed for youth who commit offences in Molendistrict but live in other districts?" (Data Inc., internal document, June 18, 2010). These questions require zooming in. It is necessary to look closer, she explains, because, although she is familiar with the area and its inhabitants, it is difficult to learn about the causes of youth crime. Statistics, she argues, often obscure local circumstances because they aggregate data. Furthermore, because the neighbourhood is relatively new (construction started in the early 1980s), she feels that the social networks that should theoretically help her to stay informed do not yet exist.

To find the answer to Liesbet's question, Juriaan starts out by producing a frequency table about Molendistrict. A frequency table is a basic list that presents the number of occurrences in a query.

> Liesbet, Mieke and Anna take a good look at the results presented on the LCD screen. It shows a list of numbers: the number of young suspects in Molendistrict, broken down by neighbourhood. "Neighbourhood H, this surprises me," Liesbet says. "Neighbourhood K makes sense because it has a shopping mall. But neighbourhood H is only a small residential area."
>
> Anna asks her to show neighbourhood H on the map on the wall. Liesbet points at it: "There it is, the park is also part of it, nowadays." "Is there a sports park nearby?" Anna wonders. This is not the case, but Liesbet notes that there is a swimming pool in the area.
>
> Anna has another suggestion: perhaps these are conflicts between neighbours? "No, not in neighbourhood H," Liesbet replies, "this is not a neighbourhood known for fights between neighbours." Next, Liesbet wants to know more about neighbourhood H: "This is making me curious, what types of offences were committed here?" (Fieldwork notes, April 7, 2011)[10]

This tells us that not all results count. A result counts when it is *surprising*, as in the case of neighbourhood H. Surprise, in Goodwin's terms, is a condition of relevance for data mining. A close view is a view that is revealing. It provides a look under the surface.

Evoking

The group returns to the question of problem youth's place of residence at the end of the meeting. This time they take a different approach. To learn more about the questions that neighbourhood H raises, they need to know more about the registered suspects. Therefore, Juriaan creates a frequency table that shows information about the suspects for offences committed in Molendistrict. This table is different from the previous one because it does not present youth living in Molendistrict but youth that have committed an offence there.

[10] The interactive meetings could not be recorded due to the sensitive nature of the data. The quotations from these meetings are based on notes that were typed out within forty-eight hours of the meetings.

> Many of the young suspects come from other districts, Liesbet notes. She points at the map: "Neighbourhood G, this is quite far away, it is in another district." But "distance is not so important to youth," she adds. Juriaan agrees: "Youth have a large radius of activity." They look a little longer at the list. "These are neighbourhoods with a lot of social housing," Liesbet notes. "Indeed, one of these neighbourhoods is a weak neighbourhood," Juriaan adds. (Fieldwork notes, April 7, 2011)

So, youth travel from other places to Molendistrict. They might come from neighbourhoods in other districts with fewer facilities, the group adds, and when they come to Molendistrict they cause a nuisance.

This is the type of account that was often generated in this pilot: propositions, statements, descriptions and explanations that loosely hang together. An important part of producing an account of Molendistrict's suspects is 'evocation'; the data needs to come to life and make sense, to be made 'tellable' (see also Curry 2004; Ziewitz 2011). This first of all required maps. During the same session described in the first fragment above, Anna asks Liesbet to show neighbourhood H on the paper map on the wall. Almost every meeting room has one of these maps. They are large, about one and a half meters by one meter, and detailed. They show the borders of the districts and neighbourhoods as designated by the Dutch central government (National Statistics Netherlands or *CBS*).

The map displays what the screen does not: the borders between neighbourhoods, the distances between the neighbourhoods, and a detailed street plan. The data mining software can plot the results on a digital map, but this is an oblique map: it does not show borders and specificities. The detailed paper map helps Liesbet locate the neighbourhood and tie it to her previous experiences: for instance, neighbourhood H is not known for conflicts between neighbourhoods. Maps also help to fill in the particulars of the neighbourhood. They show whether there is a park nearby, or a shopping mall. Furthermore, they show distances, revealing that youth travel. So pointing at a map is part of bringing the data to life.

The second, and related, part of making the data come to life is to tie the results to policy theories, categories and local knowledge. This might be policy's equivalent of what Mariana Valverde refers to as administrative knowledge in legal practices: "in-between knowledge" used by officials that is neither scientific nor lay (Valverde 2003, 20). By talking through these policy theories, categories and local knowledges, the results on the screen are pieced together. For instance, neither the frequency tables nor the maps give information about social housing, yet the policy officers apply this characteristic to the area on the map. Subsequently, they apply the notion that neighbourhoods with social housing are poorer, and poorer neighbourhoods are more troublesome. Juriaan's additional remark about 'weak neighbourhoods', moreover, pieces the account together. He refers to the forty neighbourhoods that the national government has pointed out as the most problematic in the country in 2007 (also known as *Vogelaarwijken*).

My point here is not that the city's knowledge base is not 'scientific' enough. Policy practitioners need at hand knowledge to do their jobs. Rather, I argue that this mode of improvisation shows that producing such accounts under the banner of more granularity introduces more general categories. In this example, moreover, a neighbourhood category invokes judgments about youth behaviour.

This is also relevant for the outcome of these sessions: namely, that youth who commit offences in Molendistrict are not from Molendistrict but from other, 'bad'

neighbourhoods. This outcome was partly informed by local politics. At the time of the pilot, Liesbet was involved in a discussion about the construction of new facilities in Molendistrict, such as practice rooms and bars. The dominant notion was that more facilities would help to reduce complaints about youth behaviour, as youth could engage in more activities. Liesbet reasoned that facilities might in fact cause more complaints because problem youth would come from other places to use them.

Comparing

At the next meeting, the group returns to the theme of the Molendistrict suspects. The question is how neighbourhoods within Molendistrict differ. Juriaan performs a 'profile analysis' – a comparison between young suspects living in Molendistrict neighbourhoods with all inhabitants of the district. He states:

> Now we see that of all persons we are looking at, all 772 [registered suspects aged twenty-three and younger], 5.3 per cent lives in neighbourhood K. Of all Molendistrict inhabitants, 3.4 per cent lives in neighbourhood K. This is a difference, but not dramatically so. The difference in percentage is 1.9. (Fieldwork notes, April 26, 2011)

Not everything counts as a surprise. From Juriaan's quote above, we learn that surprise also comes with difference. Neighbourhood youth need to be distinguished from district youth.

Comparison is another mode of situated improvisation, next to evocation. In order to zoom in one needs to find a difference. This is built into Data Inc.'s software as 'profile analysis'. As with evocation, however, the results on the screen do not make sense by themselves. This is an instance in which the pilot's participants aimed to zoom in from city to district:

> Juriaan is asked to compare young suspects living in Molendistrict with suspects living in the city as a whole. First, the results on the screen show that seventeen per cent of the Molendistrict group is suspected of vandalism, against eleven per cent in the city. "So it must be the case that city youth commit different types of offences, whereas Molendistrict youth mostly commit vandalism," Liesbet argues.
>
> Next, Juriaan turns to the absolute numbers. He shows that there were about 7,000 young suspects in Burgcity in 2007-2010, of which about 600 suspects live in Molendistrict. Juriaan asks Liesbet for the total number of inhabitants. Liesbet answers that 40,000 people live in the district, compared to 200,000 in the city. "So the city is only five times as big, while the number of suspects is about ten times larger." This means that, relatively, Molendistrict youth are not that bad, Juriaan concludes. (Fieldwork notes, April 7, 2011)

The issue of size is central in this fragment and Liesbet's knowledge of the numbers is basic, but crucial. In this case, comparison serves to estimate the size of the problem of Molendistrict youth delinquency. Interestingly, the group concludes that the problem of youth delinquency is relatively small.

At other times, a paper map was used to compare. Liesbet was not always sure of the neighbourhoods' sizes in terms of numbers of inhabitants, but a map could be used to estimate the geographical size. In the one meeting in which no maps were present, Anna got up, walked to the whiteboard and drew a map for Juriaan. She did so to demonstrate to Juriaan the relatively large size of a neighbourhood compared to the district as a whole.

The two modes of improvisation, evocation and comparison, indicate that the results on the screen did not establish close views on their own. Part of the work of zooming in only started when the results had appeared on the screen. It was done by applying professional and everyday knowledges and reading paper maps. Problem youth therefore, was by no means an identity established by algorithms alone.

Zooming in, moreover, was normative work. First, the use of maps in this section draws attention to a, perhaps obvious, characteristic of this practice: the idea of formal neighbourhoods as units for policy. Consequently, the nature of magnification is geographical; it focuses on a magnification of neighbourhood processes. This is telling, as income or educational levels could have also served as focal points for zooming in. Instead, the project group focused on the physical and spatial characteristics of the neighbourhood, such as facilities.

Neighbourhoods are taken to be the determining factor in many urban processes, and the neighbourhood is central in policy practice in the Netherlands as a whole, as exemplified by national programs that focus on 'weak neighbourhoods'. This focus on the neighbourhood derives from an era of social science in which it was also a marker of social class. As David Phillips and Michael R. Curry note, it brings into practice the notion that "you are where you live" (2003, 143). The latter was especially true for the Burgcity policy makers, as they were not only looking for pragmatic information to sell products, but they wanted "deeper information" to "get close to youth" (Interview, November 15, 2010).

Second, through comparison the entity representing the 'whole' is constituted as the norm. So when Molendistrict youth are compared to city youth, the city is the norm. In this case, Molendistrict's deviation from the norm was positive: the problem of youth delinquency was, comparatively speaking, relatively small.

Zooming In to Neighbourhoods

Lifestyle Analysis: A Discovery

For Anna, the city's policy specialist on youth, results on district level are not specific enough. She wants to know what characterizes problem youth in specific neighbourhoods. This closer view of problem youth, however, was never established in the pilot study. In fact, Data Inc. finally suggested zooming out in order to zoom in.

Anna participated in the meetings as the city's project's manager. Her main concern was to find out whether the software could be of added value to policy. Her point of departure was 'the funnel model' (Interview, July 18, 2011; Burgcity 2012, internal document). According to this model, twenty per cent of all youth is in the lower end of the funnel's cone; these youth are at risk of committing an offence. The four per cent in the tip of the cone is out of the police's and government's reach. Anna's aim was to find proactive approaches for this four per cent.

To this end, Anna contributes the following question to the pilot study: "What types of problem youth can be found in the neighbourhoods in terms of combinations

of characteristics and behaviour?" (Data Inc., June 18, 2010, internal document). This question is to be answered with the lifestyle data. During the meetings, Anna poses her question with some humour: "Do girls that read *Tina* [a Dutch teen magazine] set fire to litterbins more often?" On April 26, 2011, the group finally discovers something:

> Partly joking, Anna introduces what has been her main question for the past few meetings: "Which offences do girls that read *Tina* commit? Or, in other words, can we find a relation between offences and the Mosaic [Experian] data?" When Juriaan does not reply straight away, she rephrases the question: "So can you say that people who own a Jaguar or whose parents own a Jaguar are guilty of different offences than youth whose parents drive a mini?"
>
> Juriaan's first step in generating an answer to this question is to find the characteristics that distinguish young suspects in Molendistrict from all inhabitants of Burgcity. He presents a long list of characteristics that differentiate these youth on the LCD screen. We only see the characteristics for which the difference is statistically significant. Strong correlations with high values are shown at the top of the list; weak correlations are at the bottom. Among the latter are the Experian lifestyle variables.
>
> Anna and her colleagues had been waiting for Juriaan to use the lifestyle data, so quite excitedly they ask him to scroll down. Suddenly Anna exclaims: "The sport darts! Yes, so the pilot is a success!" (Fieldwork notes, April 26, 2011)

Anna finds what she is looking for: an association between lifestyle and youth crime and nuisance. To be sure, she does not believe this relation to be causal. In fact, the nature of the relationship is not relevant to her. But if it would be possible to find a 'segment' of Molendistrict problem youth that could be located in a darts club, that reads *Tina*, or whose fathers drive Jaguars, she would know how to reach them (Interview, November 15, 2010).

In the meeting, Anna continues to discuss how these findings could be used to inform policy. She reasons that, on the basis of these and similar patterns on the neighbourhood level, one could initiate a marketing campaign. One could address darts associations, for instance, to reduce vandalism. Anna does not expect this approach to lead to drastic reductions of vandalism rates; however, if offences would decrease from one hundred to eighty incidents, this would be a satisfactory result. "We can ask a marketing intern whether this is a useful tool for a communicative government," Anna concludes (Fieldwork notes, April 26, 2011).

How Surprise Disappeared

Anna's discovery did not make it to the end report. When I asked Juriaan about it shortly after the pilot group's last meeting, he did not remember the finding. In fact, Juriaan had discarded it soon after its discovery, arguing that there were simply not enough young suspects living in the individual Molendistrict neighbourhoods to perform such an analysis.

There was an additional problem. The police had not granted the project group access to the names and individual addresses of the young suspects due to privacy restrictions. This was a problem because the lifestyle information needed to be related

to the suspects on the basis of residential address, while the police had only provided information about the neighbourhoods these youth live in. The lifestyle data therefore also needed to be aggregated to neighbourhood level.[11] Accordingly, a correlation between darts and registered suspects only meant that a number of people played darts in the suspects' neighbourhood; it did not mean that the suspects actually played darts themselves.

Anna was disappointed. Not only were the results statistically significant, but in a previous meeting one of Juriaan's Data Inc. colleagues had argued that even if a difference in terms of lifestyle data was found on the basis of only four suspects in one neighbourhood and ten in another, this could be an interesting lead for a policy maker. Even though the numbers are small, the difference between four and ten is telling, so this analyst argued.

There were two ways in which using small numbers as a basis for policy were discussed. The first of these comes from the world of marketing. As Juriaan explained:

> Well, in marketing, it's like this … When you are flyering, for instance, it is about making very small differences. So if you start a campaign that leads to a difference [in sales] of one per cent in one neighbourhood and 0.7 per cent in the other … you could say this is insignificant but it actually is a very good result. (Fieldwork notes, February 28, 2011)

Marketing had inspired Anna to think about lifestyles in the first place. To Juriaan she responded: "It is the same for us, really. I mean, even if one person is prevented from becoming a suspect or a criminal, this would be a welcome result." In terms of the darts club example, even if just one person matched the darts profile, this could be interesting.

The second way in which the use of small numbers as evidence for policy was discussed comes from policing. In detective practices, a difference between four and ten provides a 'lead', It means that one has a starting point for investigation. For instance, if the profile for pickpocketing is a tennis player that reads glossy magazines, a detective can start investigating tennis clubs.

At work here are 'professional regimes of evidence' (Kahn 2013): professional standards for the type of evidence acceptable as a basis for intervention. In this case, the regime of evidence for policy had not yet crystallized. Juriaan, in his role as analytical expert, had final decision making power over the acceptability of the regimes of evidence that were applied. In other words, he had the professional privilege (Goodwin 1996) to decide on this, as he stated in the final pilot project meeting:

> Bart: So what about interventions, as in the *Tina* example?
>
> Juriaan: Well, if we are talking about a neighbourhood of 300-800 people, and maximum five per cent are registered suspects, should one start a campaign? (Fieldwork notes, September 26, 2011)

[11] The original lifestyle data were aggregated to postal code-6 level, which means that this dataset provided information about the average characteristics of about twenty persons.

The question was put forward by Bart, a member of the innovation platform that supported the pilot. In response to Juriaan's question, Bart, Anna, Liesbet and Mieke all shook their heads.

Crime, Juriaan argued, is too serious a topic to flyer for if one can only expect the numbers to range from ten to twenty persons. Moreover, they could not even be certain about the actual relation between the young suspects and the lifestyle data, as explained above. Juriaan furthermore rejected the detective logic; in the case of the Burgcity pilot, the results were not reliable enough to justify a visit to the darts club (Interview April 13, 2012).

Zooming out

In the final meeting, Juriaan and his senior colleague Frank argued that lifestyle analysis was not a feasible option. Nevertheless, Frank suggested changing to a "helicopter perspective". He advised to acquire data about young suspects in other cities. Youth from similar neighbourhoods in different cities, in terms of income similarities, for instance, could be grouped together and compared to other types of neighbourhoods. More variety between neighbourhoods and larger numbers would benefit the analysis and lead to statistically significant results: "to uncover the real processes on a micro level, one needs more material for comparison" (Fieldwork notes, September 26, 2011). Once a pattern was found for a general type of neighbourhood, it could be applied to Molendistrict neighbourhoods. It follows that to zoom in, they would first need to 'zoom out'.

The reasoning here seems surprisingly similar to what Data Inc. and Burgcity's policy makers had earlier referred to as 'general social science'. With this, they referred to national statistics and criminological theory, as well as knowledge of a more universal kind (not specifically tailored to the neighbourhood). This was the type of knowledge that "should inform policy", Frank argued.

Data Inc. seemed to have decided that the regime of evidence suitable for policy should be based on large numbers and statistical significance. Ironically, this was exactly the type of analysis that Data Inc. had promised to avoid at the outset of the pilot when it proposed formulating 'local theories'. Anna therefore argued against Data Inc.'s new suggestion, claiming that this type of analysis lacks specificity. For instance, it overlooked the fact that Molendistrict does not have sufficient facilities for youth between twelve and eighteen years of age. With that, the pilot meetings came to an end and Data Inc. commenced writing the end report.

Problem Youth

At the outset of the paper I proposed that by zooming in with data mining one does not obtain a more detailed view of the same object. Rather, zooming in brings a new object into being. In Burgcity, profiles of problem youth were never stabilized, that is, accepted by all participants and included in the end report as an outcome of the pilot study. Yet, we observed several tentative profiles in this case study. In the first instance of zooming in described in this chapter, for example, problem youth were evoked with crime statistics and physical and spatial neighbourhood characteristics, such as the presence of swimming pools. Later on, we learned how problem youth came into being by assembling more general categories and aggregates, such as weak

neighbourhoods and city youth. When Data Inc. suggested zooming out, problem youth were related to youth in similar neighbourhoods in other cities; they became part of a national trend.

A particularly contentious issue is how corporate data and methods affect government policies. The ways in which corporations categorize and differentiate in order to increase profits may be at par with egalitarianism and the provision of social justice as principles of government (Gandy 2007). I will further elaborate on the profiles of problem youth that were created in Burgcity with this issue in mind. First, I discuss the use of pre-set Experian lifestyle profiles. Next, I discuss the profiles in terms of their regimes of evidence.

The Experian lifestyle profiles played an important role in this case study. The application of consumer profiles was expected to reveal a more personalized and closer view of youth. Problem youths would be consumers, to be identified by their media usage or their parents' cars. Yet, the profiles did not always seem to easily fit into a crime control environment, as is illustrated by the following fieldwork fragment. A local police officer attended part of an interactive meeting at Anna's invitation. She suggested how the Experian categories might be used:

> Police officer: Well, I think you can learn from this [the Experian data]. Those "quiet radio listeners" will be very annoyed by kids playing soccer outside, and they will surely call the city or the police and say: "I can't hear the radio because of the noise."
>
> Liesbet: I think you should see this as a marketing profile (…) It is a characteristic of the neighbourhood, where residents use the radio more often than the internet. (Fieldwork notes April 26, 2013)

"Quiet radio listeners" refers to an Experian profile. Here it was used to learn about the types of people that file complaints with the police and the city about youth. The short dialogue illustrates how marketing profiles can be used to explain and predict behaviour. The profile's name invited connotations about a certain type of person: quiet and peaceful (see also Curry 2004).

The fragment also indicates that practitioners do not necessarily accept the categories they are presented with. Liesbet argued that these profiles are tailored to marketing usage, not to understand other types of behaviour – let alone those of problem youth. Marketing variables and categories with less obvious titles, however, may be more easily integrated into government practice. In another meeting, for instance, Liesbet argued that persons in the category of "two times average income and self-employed" complain more often and insist that the government should solve their problem (Fieldwork notes, April 21, 2013).

With regard to the regimes of evidence that were applied, we learn two things about corporate intervention. The first is that Data Inc. approached problem youth as an information problem, thereby justifying the collection of more data (Schinkel 2009). It argued for the collection of more data as it aspired to a universal truth value in the guise of large numbers and statistical significance.

Second, we should nevertheless be careful in assuming that marketing methods simply 'contaminate' government practices. With regard to the regime of evidence, Data Inc. decided that a more 'traditional' social science analysis on the basis of more cases would be more suitable for local government. It was therefore Data Inc. that reified the ethics of government intervention. In Data Inc.'s view, a

local government cannot deal lightly with the issue of youth crime. Juriaan and Frank emphasized the city's responsibility for careful and effective action in the field of youth crime policy. Burgcity, in contrast, was rethinking its own role in marketing terms, as a 'communicative government'.

Conclusions

My interest in this chapter was in how a digital identity of problem youth was created according to the rationale of zooming in. When brought into practice, I argued, this metaphor suggests that digital representations of youth have a high truth status. It therefore justifies the collection of ever more data and the use of profiles. I set out to challenge zooming in as a data mining metaphor by showing how it was done in practice and by drawing out the norms that were embedded in and produced through this work.

The Burgcity case shows the limitations of the metaphor. It had one obvious limitation: there simply were not enough registered suspects in a neighbourhood to perform an analysis on. Yet, this did not discourage the analysts; Data Inc. argued that simply more data were needed.

I demonstrated, moreover, that data mining is not a practice based solely on digital data. Far from a smooth and technical operation, data mining was a situated practice. I identified two modes of situated improvisation: evocation and comparison. These were conducted by the interplay of screens, professional knowledges, paper maps, local politics and regimes of evidence. By relating these heterogeneous entities, one does not acquire a better view of a smaller part of the same object, but a new object of intervention is brought into being. In this case, problem youth shifted from relations between categories of administrative everyday knowledge and objects in the neighbourhood, such as swimming pools, to relations between youth from comparable neighbourhood in different cities.

We learn about several normativities that were part of zooming in. First, results needed to be made relevant as surprises. Useful knowledge was unexpected knowledge. Yet, the results could not be too unexpected because data mining results needed to be made tellable. Furthermore, surprise depended on difference: local problem youth needed to be distinguished from the problem youth in the larger geographical area they are part of, such as a city or district. This requirement also produced a norm: local problem youth is invariably constituted as a deviation, and the city or the district as the norm.

Second, zooming in had a focus: useful knowledge was knowledge at the level of the neighbourhood. This had a consequence for problem youth: they were taken to 'be where they lived' (Curry 2003). Because crime policy is focused on neighbourhoods in Burgcity, categories of good and bad neighbourhoods already existed. These labels were easily transferred to the young suspects living in them.

Third, as the previous remark also indicates: zooming in depended on the application of more general categories and aggregated data. Problem youth identities only became more particular by assembling generalities. The categories and knowledges that were applied were normatively laden themselves: they told of good and bad neighbourhoods, and they were mobilized in relation to local policy discussions about facilities (as these might attract youth from 'weak neighbourhoods'). Furthermore, standardized profiles were introduced from the domain of marketing, thereby equating problem youth identities with consumer

identities. In this pilot, however, these profiles did not stabilize. It also needs to be noted that the policy makers did not accept every standardized profile.

Fourth, establishing a closer look depended on decisions about what counts as good evidence for policy practice. Two regimes of evidence were introduced into Burgcity's policy practice: a detective regime and a marketing regime. The marketing regime was appealed most to because it related to the idea of local government as a 'communicative government'. Data Inc., however, decided that government intervention needs a more 'scientific basis'. Intervention on the basis of small numbers would not be effective or justifiable. In this case, a corporation reified what it thought of as the ethics of government intervention.

If anything, these findings point out that improving the use of data mining for a better view of problem youth involves not only decisions about which digital data to use and how much. Producing knowledge about problem youth that can inform a fair policy practice will include rethinking the issues above. At the very least, it involves rethinking the relation of policy knowledge to social science and the corporate sector.

To conclude, a range of metaphors circulates digital identification practices such as data mining. Aside from zooming in, actors use 'connecting the dots' (Amoore and De Goede, 2008), 'deep knowledge', and 'obscured knowledge hidden in the data'. These metaphors, I suggest, help perform the seemingly endless analytical possibilities of these technologies. We need to attend to them as situated practices in order to change the terms by which digital identities are produced.

References

Ang, Rebecca P. and Goh, Dion H. 2013. "Predictive Juvenile Offending: Comparing Different Data Mining Methods." *International Journal of Offender Therapy and Comparative Criminology* 57 (2): 191-207.

Amoore, Louise and De Goede, Marieke. 2008. "Transactions After 9/11: The Banal Face of the Preemptive Strike." *Transactions of the Institute of British Geographers* 33 (2): 173-185.

Bowker, Geoffrey C. and Star, Susan Leigh. 1999. *Sorting Things Out: Classification and its Consequences.* London: MIT Press.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data." *Information, Communication & Society* 15 (5): 662-679.

Choy, Timothy K. 2005. "Articulated Knowledges: Environmental Forms After Universality's Demise." *American Anthropologist* 107 (1): 5-18.

Curry, Michael R. 2004. "The Profiler's Question and the Treacherous Traveler: Narratives of Belonging in Commercial Aviation." *Surveillance and Society* 1 (4): 475-499.

Custers, Bart. 2004. *The Power of Knowledge. Ethical, Legal, and Technical Aspects of Data Mining and Group Profiling in Epidemiology*. Nijmegen: Wolf Legal Publishers.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17 (3): 37-54.

Foucault, Michel. *Discipline & Punish. The Birth of the Prison.* New York: Random House.

Gad, Christopher and Peter Lauritsen. 2009. "Situated Surveillance: An Ethnographic Study of Fisheries Inspection in Denmark." *Surveillance and Society* 7 (1): 49-57.

Gandy, Oscar H. 2007. "Data Mining and Surveillance in the post-9/11 Environment." In *The Surveillance Studies Reader*, edited by Sean P. Hier and Josh Greenberg, 147-157. Maidenhead: McGraw-Hill.

Goodwin, Charles. 1996. "Professional Vision." *American Anthropologist* (3): 606-633.

———. 1995. "Seeing in Depth." *Social Studies of Science* 25 (2): 237-274.

Haraway, Donna. 1991. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." In *Simians, Cyborgs, and Women. The Reinvention of Nature*, 183-202. London: Free Association Books.

Hildebrandt, Mireille. 2008. "Defining Profiling: A New Type of Knowledge?" In *Profiling the European Citizen: Cross-Disciplinary Perspectives*, edited by Mireille Hildebrandt and Serge Gutwirth, 17-46. Springer.

Kahn, Jonathan. 2013. *Race in a Bottle: Law, Commerce and the Production of Race*. Lecture June 24, 2013. University of Amsterdam.

Lyon, David. 2006. *Theorizing Surveillance: The Panopticon and Beyond*. Devon UK, Willan Publishing.

Marx, Gary T. 2005. "Seeing Hazily (but Not Darkly) through the Lens: Some Recent Empirical Studies of Surveillance Technologies." *Law & Social Inquiry* 30 (2): 339-399.

M'charek, Amade. 2005. *The Human Genome Diversity Project. An Ethnography of Scientific Practice.* Cambridge: Cambridge University Press.

M'charek, Amade. 2013. "Beyond Fact or Fiction: On the Materiality of Race in Practice." *Cultural Anthropology* 28 (3): 420-442.

Mol, Annemarie. 2002. *The Body Multiple: Ontology in Medical Practice*. Durham and London: Duke University Press.

Phillips, David and Michael R. Curry. 2003. "Privacy and the Phenetic Urge: Geodemographics and the Changing Spatiality of Local Practice." In *Surveillance as Social Sorting. Privacy, Risk and Digital Discrimination*, edited by David Lyon, 31-56. London: Routledge.

Pleace, Nicholas. 2007. "Workless People and Surveillant Mashups: Social Policy and Data Sharing in the UK." *Information, Communication & Society* 10 (6): 943-960.

Pridmore, Jason. 2014. [ADD: references from this volume]. Routledge.

Ruppert, Evelyn. 2011. "Population Objects: Interpassive Subjects." *Sociology* 45 (2): 218-233.

Schinkel, Willem. 2009. "De Nieuwe Preventie. Actuariële Archiefsystemen en de Nieuwe Technologie van de Veiligheid." *Krisis* (2): 1-21.

Strathern, Marilyn. 2005. *Partial Connections.* Updated Edition. Walnut Creek: Altamira Press.

Witten, Ian H., Frank Eibe, and Mark A. Hall. 2011. *Data Mining: Practical Machines Tools and Techniques*. Amsterdam: Morgan Kaufmann Publishers.

Yanow, Dvora. 2002. *Constructing "Race" and "Ethnicity" in America: Category-Making in Public Policy and Administration*. Armonk, N.Y.: M.E. Sharpe.

Ziewitz. 2011. *How to Think about an Algorithm: Notes from a Not-Quite Random Walk*. Draft paper, September 29, http://ziewitz.org/papers/ziewitz_algorithm.pdf (accessed January 10, 2014).