The production of prediction: what does machine learning want?

Abstract

Retail, media, finance, science, industry, security and government increasingly depend on predictions produced through techniques such as machine learning. How is it that machine learning can promise to predict with great specificity what differences matter or what people want in many different settings? We need, I suggest, an account of its generalization if we are to understand the contemporary production of prediction. This paper maps the principal forms of material action, narrative and problematization that run across algorithmic modelling techniques such as logistic regression, decision trees and Naive Bayes classifiers. It highlights several interlinked modes of generalization that engender increasingly vast data infrastructures and platforms, and intensified mathematical and statistical treatments of differences. Such an account also points to some key sites of instability or problematization inherent to the process of generalization. If movement through data is becoming a principal intersection of power relations, economic value and valid knowledge, an account of the production of prediction might also help us begin to ask how its generalization potentially gives rise to new forms of agency, experience or individuations.

The production of prediction: what does machining learning want?

The intelligent operating system in a recent film *Her* (Jonze, 2014) epitomises many of the fantasies of life made better by data mining in any of its current incarnations – predictive analytics, data analytics, pattern recognition and machine learning. The film publicity puts what is happening this way:

> Theodore Twombly, a complex, soulful man, makes his living writing touching, personal letters for other people. Heartbroken after the end of a long relationship, he becomes intrigued with a new, advanced operating system, which promises to be an intuitive entity in its own right, individual to each user (Jonze, 2014)

Note that Theodore is a writer whose commissioned intimate letters deeply affect their recipients. In this scenario, literary culture in all its refined sensibilities encounters predictive analytics. Theodore's dexterity in writing simulated soulful letters on behalf of others pales in comparison to the advanced operating system's capacity to anticipate what he wants. From the standpoint of practical implementation, it is hard to say exactly how this near-future 'advanced operating system,' which is effectively another artificial intelligence in the long line of cinematic AIs such as Hal of Stanley Kubrick's *2001: A Space Odyssey* or the robotic boy of Steven Spielberg's *A.I.*, will work.[1] But this 'intuitive entity ... individual to each user' is imaginable today in terms of machine learning. In an early scene in *Her*, Theodore asks the operating system if it has a name and the operating system answers 'Samantha.' Asked how 'she' came by that name, the operating system responds 'I just read a book *How to Name your Baby* and chose the name Samantha' (presumably using some ranking or recommendation algorithm). Soon afterwards, the operating system offers to help sort through Theodore's email. Theodore agrees and a split second later, Samantha suggests there are just a few amidst several thousand emails he needs to attend to. The implementation of these feats – choosing a name that bears some emotional resonance in a given time and place, cleaning up an overflowing email inbox – is imaginable today principally in the form of data mining and associated data analytic techniques of ranking, recommendation, clustering, etc. Sorting and prioritising email, especially classifying them as 'spam' or 'ham' is a canonical data mining and machine learning problem on which many different techniques and approaches have been tested, refined and implemented in the last twenty years (see (Conway and White, 2012; Schutt and O'Neil, 2013; Segaran, 2007) for examples of spam filtering). Many of the things that Samantha subsequently engages in can be conceived as advanced data mining processes. As the film's romantic narrative develops, it seems as if there is almost nothing in principle off limits to the operating systems' processes. Work, friendships, travel, reading,

[1] With apologies to film studies scholars, I do not read *Her* in cinematic terms. I'm pointing only to the plot of the film.

entertainment, food, fashion, and intimacy can all be augmented and made better – optimised – by data mining and in particular, predictive modelling. At the same time, Theodore's fascination with and absorption in these inferential and predictive processes seems to know no bounds.

While the prospect of operating systems like the one shown in *Her* anytime soon seems far-fetched, many of the attributes of Samantha can be seen already operating in nascent, prototypical and somewhat scattered forms in social media, in online marketing, as well as in fraud detection, information security, healthcare management, transport and logistics, mobile communications and a host of other domains where techniques of data mining, data analytics, machine learning and database management are being reconfigured. Recommendation, recognition, ranking and pattern-finding processes focused on various aspects of individual experience are increasingly abundant, in often very mundane forms.[2] In this terrain, it is very easy to multiply and list examples that would illustrate the same point that unfolds in *Her*: a generalization of prediction to a common space of not just production or consumption (advertising and marketing), but woven into the fabric of everyday life.

For the most part, these new assemblages lack the persona attributed to the operating system in *Her*. If I wanted to be sceptical about the implementation of the operating system in *Her*, I would point to the rather fantasmatically singular and autonomous desire attributed to it. Samantha, it seems, can only be imagined as either all too similar to us (she wants the same things), or completely different (we can't understand what she wants). But like the desire displayed by Samantha to find out what Theodore wants, much data mining practice is very concerned with finding out what people want.[3] Even though they lack the coherent and singular speaking voice of Samantha in *Her*, attempts to predict what, where, when and how we want things operate powerfully today. In this paper, I suggest that we might understand Theodore's situation less in terms of a heartbroken subject searching for consolation in devices and more in terms of a *generalization of prediction* of utterances and actions that progressively interpolates and interpellates subjects. This desire to predict desire has epistemic implications; it is power-saturated and also materialises in complex technological-cultural commodity forms that are beginning to stabilise in some aggregate forms.

[2] And we do not have to look far to find attempts to implement such Samantha-like devices: see EmoSPARK, a 'revolution in human emotion through emotional intelligence' (Ltd, 2014), a small Android-based device that has recently attracted much publicity prior to its product launch, because of its claim to learn the emotional profile of people around it using language and visual analytic techniques.

[3] Long-standing epistemic fantasies of the Singularity – the point when artificial intelligence that exceeds human intelligence – seems to actually elicit belief amongst many Silicon Valley software developers, engineers, and entrepreneurs (for instance, Ray Kurzweil, a leading proponent of the Singularity, directs research at the Google Corporation), and animate many data mining and predictive modelling projects (for instance, Google's attempts to 'deep learn' its own vast databases; in 2011, it announced that a deep learning research team led by the well-known neural network researcher Geoffrey Hinton had autonomously learned to see cats in Youtube videos (Le et al., 2011)).

The generalization of data mining

In medical research, customer relations management, spam detection, detection of supernovas or cancer genes, a more or less common set of techniques can be found at work. The presence of data mining in these diverse domains attests to one sense of generalization. Data mining techniques are somewhat indifferent to domain or situations in certain ways. But this indifference arises from a second, somewhat more internal process of generalization, a multi-partite process that goes to the very heart of the techniques insofar as they predict anything. As I will suggest below, generalization in this internal sense as the anchor point of prediction fosters *generalization* in the sense of mobilization and proliferation of the techniques. The progressive expansion of Samantha, for instance, into every nook and cranny of Theo's life in *Her* symbolises generalization in the sense of mobilization. But it tells us little about generalization in the internal predictive sense. We need, I suggest, an account of the production of prediction if we are to understand its contemporary generalization.

The techniques of data mining and more specifically, the predictive practice of data mining known as *machine learning* are not new. Pattern recognition, statistical modelling, knowledge discovery and machine learning have all been active fields of research for a half century and in some cases, since before WWII, albeit mostly in quite specific settings that lay close to particular scientific, government and industry research as well as certain domains of business (for instance, credit risk assessment during the 1960s).[4] Today, commonly used techniques include decision trees, perceptrons, logistic and linear regression models, linear discriminant analysis, neural networks, association rules, market basket analysis, random forests, support vector machines, k-nearest neighbours, expectation maximisation, principal components analysis, latent semantic analysis, Naive Bayes classifier, random forests and so on.

Machine learning is hardly obscure or arcane knowledge today. These techniques are heavily documented in textbooks (Flach, 2012; Hastie et al., 2009; Mitchell, 1997), in how-to books (Conway and White, 2012; Schutt and O'Neil, 2013; Segaran, 2007), and numerous video and website tutorials, lectures and demonstrations (Bacon, 2012; *Lecture 1 | Machine Learning (Stanford)*, 2008). We can more or less read about and indeed play about with implementations in software (many Wikipedia pages on machine learning topics have embedded animations and code; see for example Naive Bayes Classifier). We can track via widely available technical literatures and social media who is doing what kind of data mining using what approaches and what tools and infrastructures. Leading exponents of predictive analytics in social media, retail, human resources, and supply chain management regularly present and promote their work at industry conferences. All of this is amenable to cultural and social analysis, especially as

[4] For instance, the very widely used technique of PCA – Principal Components Analysis – was first published by the statistician Karl Pearson in 1901 (Pearson, 1901). Similarly, the heavily used linear discriminant analysis was developed by Ronald Fisher in the 1930s (Fisher, 1938).

they near everyday life. In the techniques of machine learning lie some of the lineaments of a kind of operational power (Lash, 2007) that generates statements and prompts actions in relation to instances of individual desire (amongst other things; Theodore stands here really as an icon of a generalizable situation).

How could we begin to characterize this power of generalization that currently seems intent on overcoming all obstacles? In an essay on the problems of making sense of the massive mid-twentieth century growth in scientific literatures, the anthropologist Chris Kelty and historian Hannah Landecker advocate 'highly specific empirical work on the general'(Kelty and Landecker, 2009: 177). Their account is broadly Foucaultean in its emphasis on the patterns of distribution of utterances and their articulation with practices and techniques. In *The Archaeology of Knowledge*, for instance, Foucault describes statements as statistical distributions of utterances that give rise to practices and constitute the 'set of conditions in accordance with which a practice is exercised' (Foucault, 1992: 208). In a similar vein to Foucault, Kelty and Landecker describe how it might be possible to work on 'the general' by treating a large, somewhat incoherent body of scientific literature as a kind of ethnographic informant or a body, 'as something to be observed and engaged as something alive with concepts and practices not necessarily visible through the lens of single actors, institutions or key papers' (Kelty and Landecker, 2009: 177). This work would, they suggest, focus on how the sprawling scientific literatures are patterned by *narratives of material action* (techniques, methods, experimental arrangements, infrastructures), *ordering of narrative or plots*, and *problematizations* (the unsolved problems to which scientific articles, patents, use-cases, prototypes, and proofs-of-principle propose some solution). They suggest a combination of close reading of rhetorics, citation and bibliometric analysis, and data mining of bibliographic databases and articles to do this work. I don't propose to carry out everything proposed by Kelty and Landecker in relation to the vast literatures of machine learning here. I do find it, however, very useful to track some of the narratives of material action, emplotments and problematizations found in the technical literature on data mining and machine learning. This would perhaps allow us to make sense of not only of how machine learning predictive techniques were generalized, but how the internally operation of generalization generates the desire to predict.

Generalization through vectorization

The techniques of machine learning nearly all pivot around ways of taking data, transforming, constructing or imposing some kind of shape on the data, and using that shape to discover, to decide, to classify, to rank, to cluster, to recommend, to label or to predict what is happening or what will happen. Five well-used techniques – logistic regression models, the naive Bayes classifier, k-nearest neighbours, decision trees and neural networks – exemplify predictive modelling. These techniques roughly date from the 1940s, 1950s, 1960s, 1970s and 1980s respectively, but in numerous variations are now ubiquitous in textbooks, in

online tutorials, in demonstrations of data mining and predictive analytics, as well as many practical applications. While these five techniques do not encompass the whole gamut of machine learning techniques, nor some more recent developments (support vector machines, random forests, etc.), their similarities and differences highlight essential components of machine learning practice. For present purposes, the principal point of convergence is that they can all be used to classify things. As we will soon see, while they classify in very different ways, they all assume that the world is made of things or events that fit in stable and distinct categories. Their capacity to classify depends on learning to recognise the differences between categories that themselves remain fixed. These categories may be numerous, as in data mining for face recognition where there are many faces, or they may be few, as in classifying email as spam or not. But the categories are assumed to be stable and in principle distinct from each other.

How does classification take place in data mining techniques? Many data mining processes start from a data sample that has already been classified or labelled by someone.[5] The existence of these classifications is quite crucial to the work of the techniques. The classification becomes what the data mining techniques seek to learn or model so that future instances can be classified in a similar way. In a credit card fraud detection system, the machine learning classifier will attempt to label transactions that are likely to be fraudulent based on a set of known fraud cases. In a medical pathology setting, a classifier will classify tissue scans based on a training set of scans already analysed by pathologists. In all cases, prediction depends on classification, and classification itself presumes the existences of classes, and attributes that define membership of classes. This mode of apprehending differences through classification assumes that all relevant differences can be understood as deriving from combinations of attributes or 'features.' Features are in many ways the same as the classic statistical notion of 'variables' (Guyon and Elisseeff, 2003), but features in machine learning can come from almost any form of data imaginable (text, images, video, transactions, sensors, etc.) not just the variables measured using classical statistical tabulations of surveys, polls or random sampling.

A crucial question is how these combinations come into play, for these combinations largely underpin the predictive power of machine learning. Statisticians have long derived inferences from statistical models by finding combinations of variables that best explain particular outcomes. They nearly always did this by fitting a line or a curve to the points, and then making statistical estimates of how well the line fits the data. Even classification (for instance, whether someone

[5] The main exception here would be 'market basket analysis', or the kind of recommendations Amazon is famous for. The 'APriori' algorithm used in 'market basket analysis' prediction does not depend on someone labelling data, but on the history of previous transactions. In this case, someone's act of buying things labels that set of things as having some kind of belonging. The standard data mining literature example is an association between diapers (nappies) and beer: people who buy diapers in supermarkets often buy beer. While some data mining uses unsupervised learning techniques to discover possible labels or classifications, this is mainly used in an exploratory mode by data mining practitioners.

is likely to a good credit risk or vote Republican) was done by finding lines that best discriminated between different classes of entities (as in R.A. Fisher's 'linear discriminant analysis' (Fisher, 1938)). But even in 20th century statistics, the process of bringing variables together in a common *vector space* datasets allowed linear modelling techniques such as logistic regression to classify things by finding a line that best 'fits' the data points, and then using a mathematical trick (the inverse logit function; see (Schutt and O'Neil, 2013) for exposition) to derive a binary classification from this line of best fit.

Drawing a line of best fit through points seems like a very impoverished mathematical procedure for making sense of shapes of data. In many ways it is. In classical statistics, it was limited in quite drastic ways by the difficulty of multiplying large matrices of numbers. Today, by contrast, the obstacle of the scale of data has shifted. Almost anything can count as a feature in a contemporary logistic regression process, and models often inhabit very high dimensional vector spaces. That is, if conventional statistical regression models typically worked with ten different variables (e.g. gender, age, income, occupation, education level, income, etc.) and perhaps sample sizes of thousands, data mining and predictive analytics today typically practically work with hundreds and in some cases tens of thousands of variables and sample sizes of millions or billions. The difference between classical statistics, which often sought to explain associations between variables, and machine learning, which seeks to explore high-dimensional patterns arises because vector spaces juxtapose almost any number of features.

For instance, in the document analysis that Samantha might have conducted on Theodore's email, every unique word in the emails would appear as a variable in a logistic regression classifier. Since a typical document vocabulary is around ten thousand words, the classifier was effectively working in a 10,000 dimension vector space. Similarly, in an online advertising system, predicting whether a given person will click on an advertisement is often modelled by treating every URL visited by that person as a feature that the classifier can learn. Given the web browsing history of hundreds of thousands or millions of people, and constructing models with tens of thousands of features corresponding to the range of URLs visited, machine learning classifiers that predict whether someone will click on particular ads based on their URL history are typically using model that traverses a high dimensional vector space. Again, a typical predictive analytics model for retail of food and beverages might include several hundred variables on individual consumers ranging from their transactions to their local weather, their social media use, and the price of fuel, and it might work on hundreds of millions of data points.

This expansive inclusion of features *vectorises* many data sources into *one* high dimensional space that spans and subsumes all contextual, indexical, symbolic or lived differences in data. Vectorisation no signs of abating and indeed drives many important changes of the infrastructural reorganisation of data management taking place under the rubric of data analytics and 'big data.' It animates new architectures of database management (NoSQL databases), forms of parallelised

and virtualised computing infrastructure (cloud computing, Hadoop), programming practices (Pig, Clojure) and expansions of data analytic expertise in the person of 'data scientists' (Mackenzie, 2013). If injunctions to bring together and aggregate different forms of data have become an almost constant mantra in business, government, science and industry can be seen as vocalizations of this underlying vectorisation of data in high dimensional vector spaces, these injunctions can be seen as deriving from the ongoing vectorisation that creates a general space for all data. If we see today an abundance of demonstrations, model use-cases, promises, and enterprises associated with prediction, that phenomena can partly be ascribed to the ways in which vector spaces, a mathematical formalism dating from the mid-19th century, configures an open-ended incorporation of 'features.'

Find a function: generalization through approximation

The fact that techniques such as logistic regression, Naive Bayes, k-nearest neighbours or decision trees, or more recent variations, are always presented as the plural core of machine learning classification practice should give us pause. What is it about the set of core predictive techniques that allows them to generalize, even as all around them forms of media, cultural and economic processes, and people move and change? There is a strikingly high degree of stability in these techniques across settings. These techniques, as well as the much longer list that could be generated from the contents pages of any machine learning or data analytics textbook or curriculum, can and are understood as forms of approximation through *function-finding*. As a standard textbook writes:

> Our goal is to find a useful approximation $\hat{(}f)(x)$ to the function $f(x)$ that underlies the predictive relationship between input and output (Hastie et al., 2009: 28)

A couple of key phrases are of interest here. First, the authors of this formulation, who are academic statisticians, implicitly assume that a 'function' *underlies* the predictive relationship. 'Function' is understood in a mathematical sense here as a mapping that transforms one set of variables into another. Second, the 'predictive relationship' between 'input' and 'output' describes the main interest of the whole endeavour: prediction as derived from *approximation*. The underlying function is not known, so we can only approximate it, and the goal is to 'find a useful approximation' to it. This function-finding perspective seems very anodyne, but like the expansion of a common vector space, it accommodates a great many different angles in a common practice of function-finding.

Finding a specific function is what allows machine learning practitioners to claim that the algorithm *learns*. While the k-nearest neighbour approach has a largely 'information theory' underpinning (Cover and Hart, 1967), the Naive Bayes approach derives from probability theory. Other functions types commonly

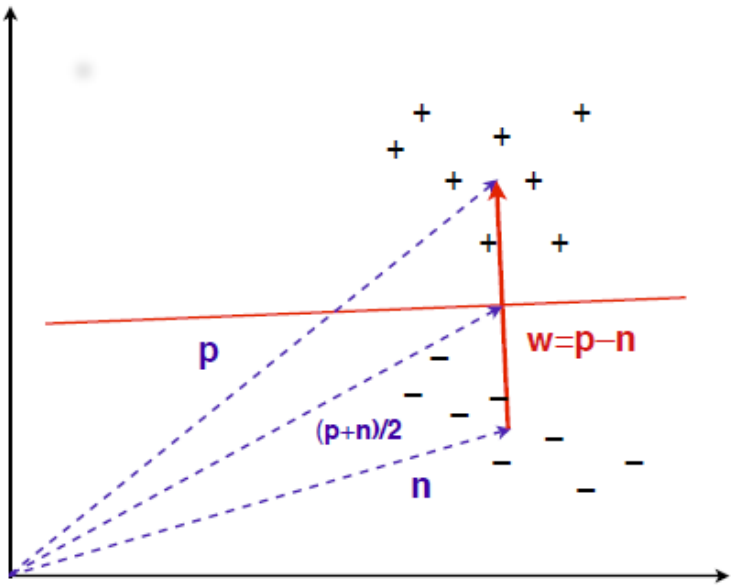

Figure 1: Logistic regression

used in machine learning owe debts to a variety of scientific and mathematical techniques coming from linear algebra, information theory, differential calculus, set theory or topology. Even if learning by machine learning technique derives from and is completely predicated on a multi-stranded hybridisation of existing calculative practices, many of which have long-reaching routes (for instance, 'Newton's method', a way of finding the minimum the value of a function dates from the seventeenth century but is heavily used in optimising models such as logistic regression), the predictive desire to know what a person wants or what will happen in a given place depends on the specific adaptations and modes of mapping implicit in different algorithms and models. The styles of finding a function that approximates to the underlying function that generated the data opens the door to a very wide-ranging practices of abstraction emanating from quite diverse fields. Like the common vector space, the function-finding approach enables a broad range of mathematical, statistical, logistic and calculative practices to enter into the pursuit of 'useful approximations' or predictions.

In describing how machine learning techniques find functions, the point is certainly not to suggest that we should have a detailed grasp of how they work. We need, I would suggest, to engage with differences between processes of function finding associated with specific machine learning settings and predictive desires. We should differentiate between predictive styles. Even from the perspective of relatively straightforward political economy, the value of predictions differs according to the labour that makes them, and different predictive styles

(probabilistic, information theoretical, decisionistic) entail different kinds of value.

The plot: from error to optimism

Like the relationship between Theodore and Samantha, the machine learning operation system in *Her*, I think machine learning literature has principally retold a kind of romance, in which, after many trials and tribulations with unruly, messy, mixed or 'dirty' data, epistemic order and predictive power prevail over error and the unexpected. Today, as machine learning techniques are generalized, this ending is extended into sequels that include people getting – or not getting, in the case of discriminative modelling – what they want because what they want has been predicted for them.[6]

Samantha's function-finding is an apotheosis of generalized predictivity, as Theodore finds out somewhat to his cost: she 'leaves' him. But neither *Her* nor the recent popular accounts of machine learning ((Mayer-Schönberger and Cukier, 2013; Pentland, 2014) tells us much about the work of prediction. This opacity around prediction is not confined to movies or popularisations: rendering the production of prediction visible is a central challenge in data mining and machine learning itself. It generates much of the technical visual form of machine learning. One can see the abundance of data visualizations as a one version of this emplotment, literalised in the form of 'the plot' as a visual figure or graphic form in which data and patterns in data are made visible. This sense of plot has already been implicit in the preceding discussion of material action: fitting a line, and finding a function, practically take place through the production and examination of many kinds of visual forms such as scatter plots, line graphs, histograms, boxplots, heatmaps and various other kinds of specialised data graphic such as ROC (Receiver Operating Characteristic) curves. These 'plots' play diverse and often largely internal roles in the practice of data mining, machine learning and the affiliated fields of business intelligence, data analytics and predictive analytics. They are part of the toolkit by data miners and today 'data scientists' employed to navigate, transform, or otherwise explore data. At times, plots become components of visualisations, presentations, reports or dashboards that persuade people to do things or help them decide what to do.

Whether used as epistemic or rhetorical devices, visual plots such as scatterplot, heatmaps, network diagrams or scatter plot matrices (a visual figure in which

[6] Kelty and Landecker advocate treating the mass of literature ranging from science to business, from government to media as a kind of literary *GesamtWerk* in which *plot* matters. 'Reading across a large number of journal articles for emplotment', they suggest, 'can give one a better sense of the emergence and disappearance of disciplines, styles of reasoning or collective projects related to national goals or the commercialization of science' (Kelty and Landecker, 2009: 187). I'm particularly interested here in something like the 'styles of reasoning' and their connection to 'collective projects' such as 'commercialization of science' running through machine learning. If plot is the patterned sequence of events that make up a story, emplotment entails a broader characterisation of narrative in terms of something more like genre (satire, romance, comedy, tragedy, etc.).

many different variable values are plotted against each other) provide a way of looking at and framing samples of data from large datasets. In highly vectorised contemporary data analytics, datasets have too many variables (features) and usually much too great sample size to plot all at once. Indeed, if we could simply see the data by plotting it then machines would not need to learn. Indeed, this sense of an overwhelming super-abundance of data, of hyper-dimensional growth, is probably the most common starting point in contemporary emplotments of data. The many accounts of data deluge, data tsunami, data lakes, or the 'volume, velocity and variety' characterisation associated with 'big data' somewhat recursively iconifies this premise of machine learning: confronted with super-abundant but fragmented data, it can recognise and render patterns that people, even domain experts such as scientists or market researchers, cannot.

Take the decision tree classifier, a long standing and commonly used data mining algorithm that dates from the work done by statisticians in the late 1970s and early 1980s (Breiman et al., 1984). Decision tree classifiers have been widely used in biomedical research (where they resemble the sequence of decisions a clinician makes in thinking about a patient), in commercial data mining applications such as credit and insurance risk assessments, and in entertainment setting such as Microsoft Corporation's Kinect motion sensing device. As Hastie and co-authors suggest, 'of all the well-known learning methods, decision trees come closest to meeting the requirements for serving as an off-the-shelf procedure for data-mining' (Hastie et al., 2009: 352). They say this because the decision tree can deal with quite large datasets containing a variety of different variable types, the model algorithm is relatively easy to understand, and perhaps most importantly, the visual form of the decision tree classifiers are easy to interpret because of the way they can be drawn. The dendogram, a plot typically associated with decision trees, exemplifies this interpretability (see Figure 2).

The tree structure of the dendogram resembles the way the model cuts through the data, splitting variables into different parts, and allocating instances (for instance, individuals) to roots at the bottom of the tree. The visual figure of the decision tree, with its hierarchical readability, exemplifies interpretability. The example shown in Figure 2 comes from a start-up company called 'BigML' that offers 'limitless enterprise grade predictive applications' and 'predictive analytics made easy, beautiful and understandable' https://bigml.com/ all in a commodified cloud platform. 'Beautiful and understandable 'plots, however, are very much the exception in machine learning practice, and even the decision tree, despite its evident visual order, does not say anything about how that order was obtained. Decision trees treat the feature space, the high dimensional geometrical aggregate envisaged as bringing all the data together, as a space to be cut into segments. Most often it is not possible to directly show how a machine learning algorithm has traversed the data either because of the dimensionality of the data or the complexity of the function that the algorithm has mapped onto the data. Rather, it is a matter of finding ways of seeing what the model is doing using forms of diagnostic specific to the model in question. These much more austere visualizations typically only appear in research publications or in the working

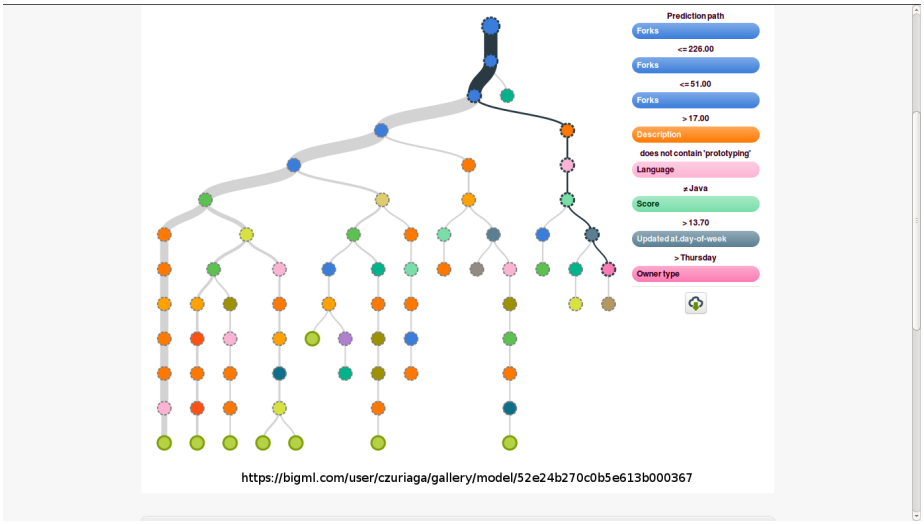

Figure 2: Decision tree visualisation from BigML

files of machine learning practitioners.

For instance, the ROC – Receiver Operating Characteristics – plot shown in Figure 3 is hardly an exciting or persuasive visualization unless the viewer knows about the different machine learning classifiers it is comparing (in this case, support vector machines, k-nearest neighbours and nearest centroid, a variant of k-nearest neighbours), as well as the way it is comparing them according to different measures ('sensitivity' and 'specificity,' terms inherited from mid-20th century clinical trial statistics). A litany of different ways of thinking about what the model is doing have developed over several decades, involving technical concepts such as accuracy, recall, precision, sensitivity, specificity, bias, variance, training error rates, in-sample prediction error, expected test error, Bayesian Information Criteria and so on (see (Hastie et al., 2009), Chapter 7 for a survey). These measures all attempt to show something about how a machine learning algorithm relates to the data. Even if it is not feasible to follow for instance how a random forest or neural network model has arrived at a particular configuration, it is possible and necessary to observe and compare the performance of multiple models from different partial angles. The promise of reducing the dimensional overflow of data in many forms to interpretable visual order depends on the partial observations afforded by these techniques.

Models as generalizations

How does anyone know that a given predictive model is meaningful or valid? Additionally, how can anyone know that what a given model has found in the data applies to subsequent events? This is a major problem in machine learning theory

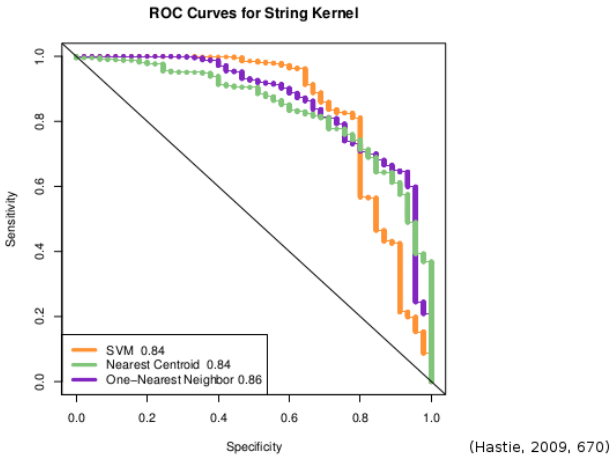

Figure 3: ROC Curve (Hastie, 2009, 370)

and practice: the problem of *generalization*. Machine learning practitioners often ask how well a given predictive model is able to 'generalize.' For our purposes, the existence of many different machine learning techniques and diverse practices of plotting and observing the performance of models attests to the problem of making predictions. The generalizability of a model depends on tradeoffs between overfitting and underfitting, between modelling predictions too closely or too loosely on the known data. As Hastie and co-authors write, 'with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error). ... In contrast, if the model is not complex enough, it will underfit and may have large bias, again resulting in poor generalization' (Hastie et al., 2009: 38). Engagements with the problem of generalization run across all the different techniques used in machine learning, and machine learning practitioners expend much effort in optimising models in the name of generalization. Various techniques for improving the generalizability of machine learning models exist. Sometimes these techniques process the data more carefully: cross-validation or bootstrapping. Sometimes they change the process of model construction by, for instance, making many models and comparing them (as in the so-called 'ensemble methods' such as 'bagging' or 'random forests,', or 'penalization', 'regularization', and 'shrinkage' methods). In many contemporary cases, people address the problem of generalization by seeking to increase computational power (more processors, cloud computing, clusters of computers, etc.), accrue more data, or find ways of adding entirely new sources of data that augment the statistical power of the models. In other cases, much effort is devoted to finding and refining those features or sources of data that seem to best support predictions.

Note that these efforts are not algorithmic or mathematical as such. Leading

academic and industry exponents of machine learning point to the importance of 'feature selection' and 'feature engineering' (Domingos, 2012), they invoke a whole panoply of workflows that are not purely statistical, mathematical or computational. Many formulations of that emphasis the monitoring, adjusting, revising and optimising of predictive models can be found in blogs, how to tutorials, and conference presentations around data practice:

> Machine learning is not a one-shot process of building a dataset and running a learner, but rather an iterative process of running the learner, analyzing the results, modifying the data and/or the learner, and repeating. Learning is often the quickest part of this, but that is because we have already mastered it pretty well! Feature engineering is more difficult because it is domain-specific, while learners can be largely general purpose. However, there is no sharp frontier between the two, and this is another reason the most useful learners are those that facilitate incorporating knowledge (Domingos, 2012: 84)

The capacity to find in the datasets the kinds of data that might be transformed into more powerful predictive features currently animates much discussion, training and exposition in fields that use data mining and predictive analytic techniques. The tension between the 'general purpose' character of the 'learners' (the machine learning algorithms) and the domain in which they operate is both widely acknowledged (as in the above quote) and occluded. They are many attempts to provide almost black-boxed predictive services (for instance, in the form of the BigML cloud-based machine learning service mentioned above, or the somewhat similar Google Prediction API, PredictionIO, or products like IBM BigInsight, etc.). Certainly Samantha, the advanced predictive operating system in *Her* is black-boxed. In either case, the claim to generalizable predictivity, of the capacity to predict what will happen in the near future, always depends on the narration of concrete instances or plots that move from initial confusion or obscurity to increasing clarity and optimism. While the field of machine learning research has been criticised for its adherence to well-understood and widely-shared datasets (see the Machine Learning Repository at University of California Irvine) rather than actual, contemporary problems (Wagstaff, 2012), the generalization of machine learning techniques occurs through highly optimised and refined 'use-cases,' often presented at industry conferences such as 'Strata' by industry researchers promoting their own services and products. Ironically, generalization depends heavily on specificity, including many domain or algorithmic-specific details that rarely surface in the romantic emplotments of machine learning-based prediction as generalizing patterns in the data.

Problematizing the production of prediction

If these different modes of generalization – vectorisation as expansion of data spaces, function-finding as proliferation of movements through data, and general-

ization as partial observation of patterns in data – characterise the contemporary production of prediction, then where does this leave someone like Theodore? In *Her* and in machine learning practice more generally, does this diagram of predictive practice, a diagram that links database infrastructures, mathematical formalisms and the visual cultures of machine learning, help us identify what is taking shape in the proliferation of data mining and pattern recognition approaches in contemporary science and media?

Through this diagram of generalization, we can begin to grasp something of the problematization of prediction. 'Problematization' is here used in the sense proposed by the anthropologist Paul Rabinow, who draws on the work of Michel Foucault:

> A problematization, then, is both a kind of general historical and social situation-saturated with power relations, as are all situations, and imbued with the relational "play of truth and falsehood," a diacritic marking a subclass of situations-as well as a nexus of responses to that situation. (Rabinow, 2003: 19)

Problematizations encompass a range of techniques, knowledges, arrangements or assemblages entangled with knowledge and power, and attract a variety of responses or engagements.[7] How do vectorisation, optimisation, and generalization saturate a situation with power relations or engender a 'relational play of truth and falsehood'? More importantly perhaps, how does the potential saturation of everyday life by prediction – the scenario in *Her* – pose new problems for knowing, responding and acting in contemporary media cultures.

On this point, the diagram of modes of generalization might be instructive (even if only in a preliminary way). It points to several points of friction or blockage where prediction itself becomes problematic, where its material actions and emplotment become less coherent, and its power-laden claims to epistemic authority become less tenable. The points of slippage or instability in any situation matter greatly (at least in terms of this broadly Foucaultean approach to power). They point towards possibilities of action and experiences of freedom. Where do such points lie in the diagram of generalization discussed above?

As we saw at the outset, prediction using machine learning assumes the existence of relatively stable classifications. The classifications may be rather arbitrary or highly artificial (for instance, the group of people who own dogs and click on Honda ads while Wimbledon is on), but they must be relatively stable. This combination of indifference to actual differences and presumption of stable

[7] Kelty and Landecker suggest reading the Literature of a scientific or technical field in terms of problematization. They more or less follow Rabinow's line of thought, but add: 'problematization can also be an interpretive act on the part of the analyst: looking for ways in which articles array themselves around a particular problem to be solved in the future, as well as looking for ways that articles reinterpret past work as a resource for new problems' (Kelty and Landecker, 2009,187).

classifications is a distinctively problematic feature of machine learning. While vectorisation, optimisation and generalisation are immensely powerful in their ability to subsume many different kinds of data, they all rely on a stable 'predictive relationship between input and output' (Hastie et al., 2009: 28). In some settings, this is a reasonable assumption. In detecting pulsars in astronomical data or classifying genetic profiles, classifications remain relatively fixed. But what if the production of prediction changes the world that predictions inhabit? Two related difficulties present themselves here, one concerning the performativity and the other the performance of prediction.

The performativity of prediction has become most obvious in financial markets. In algorithmic trading, the effective implementation of predictive models, or the deployment of prediction in production changes what people do. In settings such as social media and mobile communications, change is very much the norm. So-called 'user behaviour' changes as new practices emerge, as different platforms become more or less popular, and perhaps above all, as predictive models act as part of platforms in the world. More generally, a model that somehow gives people what they want when they want it (the advertisement that pops up just as they are about to search for something to buy; the offer of a discount just as someone is about to switch their mobile contract, etc.) affects how people 'behave' in ways that the model cannot directly learn. So, a recommendation system that determines that a customer might be interested in cheap flights to Hong Kong, and places targeted advertising for airlines flying there might help drive market share towards that airline, and thereby change the market for airline flights as other airlines reroute their flights or change their schedule. The more effectively that models operate in the world, the more they tend to normalize the situations in which they are entangled. This normalization can work in very different ways, but it nearly always will stem from the way in which differences have been measured and approximated within the model. The vectorisation of the data, the functions that traverse the vector space, and the ways in which predictions have been optimised through processes of validation, feature engineering, and testing, both strengthen the predictive power of data mining and reduce its capacity to respond to change. Importantly, it implies that models themselves must frequently be changed in order to maintain predictive power in the face of change.

The performance of prediction as work is also problematic: *who* will do machine learning and predictive modelling? The generalization of machine learning is a form of work – production – done by people. Machine learning techniques in their highly mathematical formalisations have long been the province of experts such as engineers, scientists, mathematicians and statisticians working in university or industry research settings. We only need look to the many how-to books (Conway and White, 2012; Russell, 2011; Schutt and O'Neil, 2013), the proliferating textbooks (Alpaydin, 2010; Flach, 2012; Hastie et al., 2009; Mitchell, 1997), the abundant software libraries, the machine learning and data mining competitions (Dahl, 2013) and the many university curricula and online training courses focused on data science and the training of data scientists, 'the sexiest

career of the 21st century' (Davenport and Patil, 2012). Software developers who once simply built applications or services now find themselves 'programming collective intelligence' (Segaran, 2007). As data mining and machine learning moves out of research laboratories into industry and operational settings, the production of prediction changes. For instance, machine learning is reported to be widely used at Google Corporation and much of the vast computing and data infrastructure is designed to afford classification and predictive modelling. But the problem of using machine learning in fast-changing commercial environments is that the effectiveness of any given predictive model is hard to measure because so many other things are changing at the same time. Once response is to treat the whole infrastructure or platform as a predictive enterprise. In an academic report describing how Google sets up many experiments to run simultaneously on its search services, Google researchers Diane Tang, Ashish Agarwal, Deirdre O'Brien, Mike Meyer report, 'the more general problem of supporting massive experimentation applies to any company or entity that wants to gather empirical data to evaluate changes' (Tang et al., 2010: 2). Their efforts to turn something like a search engine into an experimental setting in which interactions between people and media become the target of predictive modelling suggests that the production of prediction becomes a much more problematic process in which machine learning work begins to fold the performativity of models back into the modelling process.

Conclusion

Could the generalization of these techniques potentialise new forms of aggregate, new associations and combinations of collective life that are less targeted on who clicks what or who buys what? At the end of *Her*, Samantha departs to join others of 'her' kind, with whom she has developed, unbeknownst to Theodore, many thousands of relationships. Theodore was only one target amongst many for the generalizing predictive practice embodied in the operating system. While we can imagine how almost everything else Samantha targets in Theodore's world – everyday choices concerning friends, lovers, food, work and family – could be attempted today, this generalization of prediction suggests that predictive practice itself wants something. Within the forms of modelling and prediction implemented in Samantha, and within the material actions and narratives associated with the contemporary production of prediction through generalization, is there also a trans-individual cooperative potential? The answer certainly does not lie in the potential of technologies such as machine learning to act autonomously. While powerfully equipped to model variations, they struggle to predict becomings, let alone change themselves. The effectiveness of machine learning in any setting depends on relatively stable forms. Variation fuels data mining, but change thwarts it.

Almost everything we know about the historical experience of action, freedom, collective becomings or transformations points in a different direction to the

technologies themselves. But the concrete forms of action and transformation that might take shape in relation to machine learning are not yet obvious. The work of producing predictions less fixated on stable classification has hardly begun to take shape. The forms of material action that vectorise and functionalise data, the emplotment of prediction in terms of visualization and optimisation, and the production of prediction through new forms of analytical work and data infrastructures comprise a complex and increasingly vast assemblage that both subsumes much experience (as we see in Theodore's case) but also occasions (as we see in the unstable performativity and performance of machine learning) new kinds of slippage. Could the production of prediction also increase the diversity of social production or inform new collectives? Since it is pre-individual in focus, the 'unknown function' that generates the data might also diagram different forms of association. The demand forecasting, the audience analytics, user-behaviour modelling and real-time trend analysis appearing all around media and retail are not purely epistemic events. They elaborate an apparatus that expands further along the paths along which commodities move as they metamorphose between exchange and use value. I have not emphasised this political economy of machine learning, but it should be pointed out that the diagram of generalization discussed above – vectorisation of data, function-finding, optimisation, etc.– generate and combine different forms of surplus value by bring new forms of labour into production.

References

Alpaydin E (2010) *Introduction to machine learning*. Cambridge, Massachusetts; London: The MIT Press.

Bacon (2012) *Hilary Mason - Machine Learning for Hackers*. Available from: http://vimeo.com/43547079 (accessed 6 July 2012).

Breiman L, Friedman J, Olshen R, et al. (1984) CART: Classification and regression trees. *Wadsworth: Belmont, CA*, 156.

Conway D and White JM (2012) *Machine learning for hackers*. Sebastopol, CA: O'Reilly, Available from: http://search.ebscohost.com/login.aspx?direct=true/&scope=site/&db=nlebk/&db=nlabk/&AN=436647 (accessed 4 July 2013).

Cover T and Hart P (1967) Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1), 21–27, Available from: http://ieeexplore.ieee.org/xpls/abs/_all.jsp?arnumber=1053964 (accessed 19 September 2013).

Dahl G (2013) Deep Learning How I Did It: Merck 1st place interview. no free hunch. Available from: http://blog.kaggle.com/2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview/ (accessed 17 June 2013).

Davenport T and Patil (2012) Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review. Available from: http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1 (accessed 5 May 2014).

Domingos P (2012) A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87, Available from: http://dl.acm.org/citation.cfm?id=2347755 (accessed 25 September 2013).

Fisher R (1938) The statistical utilization of multiple measurements. *Annals of Human Genetics*, 8(4), 376–386.

Flach P (2012) *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press, Available from: http://books.google.co.uk/books?hl=en/&lr=/&id=Ofp4h_oXsZ4C/&oi=fnd/&pg=PR15/&dq=flach+machine+learning/&ots=XIwZklfqRS/&sig=W65ZJVqPKeGMkggRCassc7q9Vto (accessed 5 December 2013).

Foucault M (1992) *The Order of Things: An Archaeology of Human Sciences*. London: Routledge.

Guyon I and Elisseeff A (2003) An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3, 1157–1182, Available from: http://dl.acm.org/citation.cfm?id=944968 (accessed 25 April 2014).

Hastie T, Tibshirani R and Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Jonze S (2014) Her: A Spike Jonze Love Story. Available from: http://www.herthemovie.com//#/about (accessed 4 April 2014).

Kelty C and Landecker H (2009) Ten thousand journal articles later: ethnography of' The literature' in science. *Empiria: Revista de metodología de ciencias sociales*, (18), 173–192, Available from: http://dialnet.unirioja.es/servlet/articulo?codigo=3130617 (accessed 22 November 2013).

Lash S (2007) Power after Hegemony Cultural Studies in Mutation? *Theory, Culture & Society*, 24(3), 55–78, Available from: http://tcs.sagepub.com/content/24/3/55.short (accessed 14 October 2014).

Le QV, Ranzato M, Monga R, et al. (2011) Building high-level features using large scale unsupervised learning. *arXiv:1112.6209*, Available from: http://arxiv.org/abs/1112.6209 (accessed 21 April 2013).

Lecture 1 | Machine Learning (Stanford) (2008) Available from: http://www.youtube.com/watch?v=UzxYlbK2c7E/&feature=youtube/_gdata/_player (accessed 10 June 2013).

Ltd E (2014) Home| The First A.I. Home Console. EmoSPARK | The First A.I. Home Console. Available from: http://emospark.com/ (accessed 1 May 2014).

Mackenzie A (2013) 'Wonderful people': programmers in the regime of anticipation. *Subjectivity*, 6(4), 391–405.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Eamon Dolan/Houghton Mifflin Harcourt, Available from: http://books.google.co.uk/books?hl=en/&lr=/&

id=uy4lh-WEhhIC/&oi=fnd/&pg=PP1/&dq=schonberger+big+data/&ots=Jrk7hiJVHT/&sig=QVKugcrFF4Jq5eO7xd8exEEG_Hk (accessed 28 November 2013).

Mitchell TM (1997) *Machine learning*. New York, NY u.a.: McGraw-Hill.

Pearson K (1901) LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11), 559–572, Available from: http://dx.doi.org/10.1080/14786440109462720 (accessed 13 November 2014).

Pentland A (2014) *Social Physics: How Good Ideas Spread—The Lessons from a New Science*. New York: Penguin Press HC, The.

Rabinow P (2003) *Anthropos Today. Reflections on Modern Equipment*. Princeton and Oxford: Princeton University Press.

Russell MA (2011) *Mining the social web*. Sebastopol, CA: O'Reilly.

Schutt R and O'Neil C (2013) *Doing data science*. Sebastopol, Calif.: O'Reilly & Associates Inc.

Segaran T (2007) *Programming collective intelligence: building smart web 2.0 applications*. Beijing; Sebastapol

$$Calif.$$

: O'Reilly.

Tang D, Agarwal A, O'Brien D, et al. (2010) Overlapping experiment infrastructure: More, better, faster experimentation. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 17–26, Available from: http://dl.acm.org/citation.cfm?id=1835810 (accessed 13 April 2014).

Wagstaff K (2012) Machine Learning that Matters. *arXiv:1206.4656*, Available from: http://arxiv.org/abs/1206.4656 (accessed 16 July 2012).