Personalization and probabilities: the expression of propensities in online grocery shopping

Adrian Mackenzie[[1]](#footnote-2)

July 2017

Many accounts of big data systems assume that they target individuals. Personalization, with all the risks of discrimination and bias it entails, has been the focus of inquiry. This paper suggests that personalization occurs against a more expansive background that in principle is not reducible to individual attributes and desires, but includes many different forms of relations in flux. It describes the 'personalization' of an online grocery shopping recommender system. It reconstructs an account of how a recommender system constructs an ordered list of a small number of grocery items of personal relevance for each of the millions of online grocery shoppers in a major UK supermarket chain. Drawing on a theory of probability proposed by the philsopher of science Karl Popper, it suggests that personal relevance is inevitably bound up with relations to others, and the relations between things. Using a mixture of discourse analysis and code-based reconstruction of key elements of the recommender system, it suggests that personalization is one provisional stabilization of a much more open-ended weave of propensities associated with people and things in contemporary big data configurations. The paper explores how in the context of recommender systems the constitutive incompleteness of shopping lists, their propensity to expand or change, becomes more important than their capacity to be personalized.

This paper philosophically reconstructs (**???**) an online grocery ordering system undergoing an upgrade from a demographically-based recommendation system to a 'personal relevance' model. The reconstruction is empirical in certain respects. It departs from an ethnographic moment: being a member of the audience at an industry/academic conference where a data scientist was describing how recommendations were created for online grocery orders. Starting from this highly mediated and fragmentary encounter with a recommender system and one of its designers, the paper explores the problem the system addresses. It draws on anthropological and sociological research concerning shopping and list-making to situate the recommender system amongst the social ordering practices of shopping and lists. It makes use of archaeological approaches, in the sense developed by Michel Foucault (Foucault, [1972](#ref-Foucault_1972)) in order to identify and map important functional statements and practices configuring the knowledge and order generated by such systems over time. It conducts several small-scale code-based experiments in order to reconstruct, using widely available code resources resources such as API (Application Programmer Interfaces) and software libraries for machine learning, some prototypical elements of the system in question. Throughout, it draws on the philosophy of science and media theory to frame an engagement with data, algorithms and recommendations in terms of probabilities and their worldly enactments.

Like some recent work in science and technology studies, anthropology and media studies (**???**; Bogost, [2012](#ref-Bogost_2012); Marcus, [2014](#ref-Marcus_2014)), through this combination of approaches, the paper allows its own reconstruction to be provisionally shaped by an experiential encounter with the object it describes. At times, this mode of empirical philosophical engagement with its objects may well seem overly concerned with the technical minutiae of the groceries and recommender systems, and lacking in sociological rigour or empirical evidence. With some forbearance on the part of readers, however, the main philosophical argument of the paper should resonate sociologically. It concerns social order, or the consistency and regularity of relating and acting amid a world in where probabilities and calculations based on probabilities play a more pervasive role.

The main argument of the paper concerns how probability operates in Big Data situations. While calculations of probability have long-standing importance in many settings (insurance, medicine, experimental and field sciences, operations research, risk analysis, engineering design, public health, economic modelling, etc.), probabilities have taken on a much thicker situational texture on digital platforms associated with the web and social media. In recent research and opinion on data practice, personalization has been seen as the linchpin of predictive modelling. Big data discourse in its promissory mode attributes potency to personalization: 'most important, using big data we hope to identify specific individuals rather than groups' (Mayer-Schönberger and Cukier, [2013](#ref-Mayer-Schonberger_2013)). Similarly, Joseph Turow and co-authors conclude in their critical account of the transformation of retail space: 'through it all, knowingly and not, and away from the spotlights of fierce social debate, retailers are encouraging daily routines that accept data-driven personalization as a centrifugal public' (Turow et al., [2015](#ref-Turow_2015): 476). Analysis of internet filter bubbles and the growth of predictive platforms in industry and government make similar points (**???**; Pariser, [2011](#ref-Pariser_2011)).

The reconstruction of the recommender system suggests a different way of seeing predictive personalization by expanding the key underpinning notion of probability. Probabilities in practice -- in this case, in recommender system -- diverge from probability theory and indeed from statistics as a science. Operational probabilities differ from epistemic probabilities in their position in the world, as part of infrastructures, transactions, and event-structures. Taking their complex siting into account, and accepting their *technical* reality, we might come to understand operative probabilities along the lines of the philosopher of science Karl Popper understood as a 'world of propensities, as an unfolding process of realizing possibilities and of unfolding new possibilities' (Popper, [1990](#ref-Popper_1990): 19). Given this understanding of operational probability, our understanding of what is enacted in contemporary data analytics, especially in association with platforms (Gillespie, [2010](#ref-Gillespie_2010)) where the operationality of probabilities are heightened, might shift focus away from personalization.

The broad implication of this expansion concerns how we approach social order or regularities amidst change or transformation. Proponents and critics of Big Data present the advent of large volumes of data and analytical techniques (machine learning in particular) as transforming for better or worse organisations, institutions, economic processes and social life. Personalization has been the pivotal element, they argue, of this transformation. I suggest too that a transformation or conversion event is occurring, but one which might also, even in the midst of much ongoing personalization, open up the possibility of experiencing the world differently, and conceiving of social ordering processes somewhat afresh. Even if the argument is centred on grocery shopping, the analysis of how a world of propensities continuously and unevenly converts into realizations applies across the board, from the most physical to the most ideal or abstract settings. The average everydayness of the example here -- grocery shopping -- allows some first hand engagement with the messiness, the entanglements, and potentials of this conversion event in ways that other interesting settings -- social media for instance -- do not.

# B. The conversion from demographic to personalized recommendation

At one of the many industry-meets-academia events occurring in increasingly data science-oriented higher education institutions in the UK, speakers from industry, government and commerce described their work with predictive models.[[2]](#footnote-3) Shreena Patel, a PhD graduate in statistics and operations research, works as a data scientist for DunnHumby, a well-known [customer science company](https://www.dunnhumby.com/) (dunnhumby, [2017](#ref-dunnhumby_2017)). Her work at DunnHumby focuses on online grocery shopping at the supermarket chain Tesco. Speaking to an audience of statisticians and operations researchers, Patel focused on the development and operation of predictive models underlying shopping list recommendations. Her presentation was filled with graphs, numbers, and tables concerning ongoing development of the 'Have you forgotten?' recommender system.

Against the background of the sheer number of commodities and their distribution of prices, Patel's presentation presents two opportunities. First, by recounting, contextualising and commenting on the main steps in making the 'Have your forgotten?' list, we might follow some of the predictive sense-making done by data scientists and customer analytics teams working with transactional data in a typical commercial setting. Patel mentioned many of these steps only fleetingly in the presentation, for they are largely taken for granted as part of predictive analytic practice. Second, Patel focused on the renovation and updating of long-standing data-mining practice via a much more explicitly 'big data' and 'machine learning' oriented implementation. Her presentation concerns what we might term, for reasons that will gradually emerge, a 'conversion event' in which a long-standing recommender system was replaced by a new, explicitly 'big data'-style system concerned with 'personally relevant' recommendations. What stands out from the presentation is the normality of the recommender system: it is part of a long-standing and ongoing transformation or conversion of grocery shopping. From a sociological standpoint, the interest lies less in specific technical innovations and more on how the ordering work done by recommender systems relates to problems of social order more generally.

Online grocery shopping at [Tesco](https://www.tesco.com/groceries/?icid=dchp_groceriesshopgroceries), the largest supermarket chain in the UK, includes recommendations for further grocery purchases under the title of 'Have you forgotten?'. When Tesco customers shop for groceries online, a list of five recommendations appear at the checkout stage. The recommendations are the product of a recommender system, an important category of operational device in big data (see for instance (**???**) for analysis of Netflix recommendation; (Morris, [2015](#ref-Morris_2015)) or (Seaver, [2015](#ref-Seaver_2015)) for an account of music recommendations). The question 'have you forgotten?' is followed by a list of some grocery items that could have been or are usually on a shopping list. The title of the suggestions is a bit misleading: the recommender system, as we will see, is not concerned with forgetting, with the many slips and oversights associated to which shopping is prone, but rather with items that customers had not selected perhaps because they had never thought of buying them in the first place.

# C. Shopping and a theory of list orderings

The anthropologist Daniel Miller has argued (Miller, [2012](#ref-Miller_2012)) that shopping takes place as a negotiation of discrepancies between normative and actual social order (for instance, between the ideals of health emblematised by organic products and commitments to thrift embodied in generic products). More provocatively, Miller claims that 'shopping is largely a technology for the expression of love' (85). What Miller suggests here is that shopping practices attempt to resolve difference between how people think they should live and how they actually live: 'we have to watch how shopping helps resolve these discrepancies between the normative and the actual, but we also need some ideas of where the normative comes from in the first place' (Miller, [2012](#ref-Miller_2012): 72). Whatever theory we have of how normative social order comes about, the key argument is that shopping inhabits a space between normative/idea and actual order. Importantly, and this will be a key consideration for recommender systems, shopping is not particularly personal or individual. Grocery shopping in particular is woven through with forms of social order concerning family and other forms of social grouping, and their associated relations (love, etc.).

Shopping lists provide important clues to how local social order is constituted, maintained, and repaired in shopping. If social ordering is done in shopping, then shopping list as an inscriptive device stabilises certain forms of ordering at the expense of other possible orderings. The shopping list, whether written on the back of an envelope, or saved as a list in an online grocery shopping system, lies at the intersection of logistical flows, infrastructural orderings, and lively negotiations around actual and normative social orders.[[3]](#footnote-4) Shopping lists, therefore are intersectional ordering devices that encapsulates a universe of possible references, and a teeming multitude of propensities with an actual local order (in both senses of that term: an ordering and an imperative). As people shop, either by trawling along aisles packed with thousands of products, or scrolling down screens or searching for particular brands amidst search results, lists inscribe some kind of order that filters or reduces the excessive propensities, claims, dazzle and distraction of commodities to the practicalities of domestic economy.

Hand-written shopping lists undoubtedly have ongoing importance and display interestingly mixed ordering practices (see the montage of handwritten shopping lists at [Grocery List](http://www.grocerylist.org/)). But online shopping lists take shape at the intersection of web and internet infrastructures, supply chain logistics, individualised practices and *habitus,* and increasingly, the predictive operations of recommender systems. Just as the aisles and shelves of a supermarket present a densely-woven semiotics of objects competing for visual attention by offering distinctions of taste, thrift, expedience, novelty, indulgence, health and increasingly, online shopping recommender systems generate lists that seek to align people to commodities that they otherwise might have little relation to (see (Turow et al., [2015](#ref-Turow_2015)) for an overview).

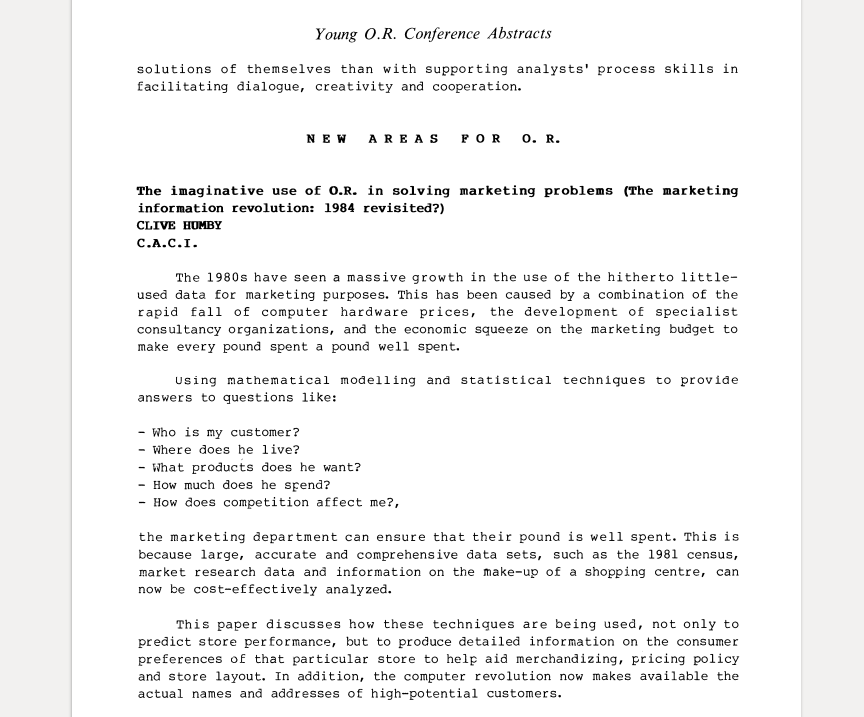
If, as the social anthropologist Jack Goody argues, lists are historically primary as forms of writing in urbanizing cultures (Goody, [1986](#ref-Goody_1986)), and if list-making and its later variants (e.g. tables) precede discursive and narrative writing practices, then we might expect lists to function as powerful social ordering devices. More recent sociological work on lists (see (de Goede et al., [2016](#ref-deGoede_2016)) for an overview) explore the social and political potency of lists as ordering devices. As literary scholars suggest (see (Mainberger, [2003](#ref-Mainberger_2003))), even if lists have often been de-valued as literary forms, list-making commonly appears in literary form whenever writing seeks to address, name, group or evoke totality, profusion, excess or abundance. From both anthropological and literary perspectives, shopping lists have excellent reasons to exist: they are deeply rooted in organisational life and infrastructures. Lists weave together people, infrastructures, things, and places.[[4]](#footnote-5)

Finally, and more formally, we might also consider some of the formal properties of lists as orderings. All lists imply order, but what kind of order? Mathematicians distinguish lists from sets on the basis of a distinction focused on order. Formally, both lists and sets are collections, but a list also contains a mapping of its elements to an ordering, usually the natural or counting numbers. Sometimes this order is explicit. Numbers are written down one side of the list. Often the order is implicit to the spatial arrangement of the list. This formal idea of the list as a mapping is helpful since it suggests relational possibilities. Mappings occasion relations between sets, collections, group or multiplicities. Mappings between list order and social order are manifestly highly contingent. A flight checklist differs greatly from a personal to-do list, even if they share many formal properties. Listing practices in a given social setting, therefore, always index an engagement between the formal properties of the list and the necessarily more fluid dynamics of social order. From the perspective of social order, shopping lists, as they are co-constructed by a supermarket chain's recommender system, might be important intersectional zones for the conversion of average everydayness through data-analytic propensities.

TBA - short paragraph summarising the social theory of list ordering; make sure to add stuff on the ordering-remapping of order in the list;

# D. Archaeology of recommendations: from 1984 to 2007

The main narrative of Patel's presentation concerned a shift from a well-established loyalty-card based data-mining model to a predictive, probabilistic, personally relevant re-writing of the shopping list in almost-realtime. The changes Patel described are increasingly widespread and common. While they are configured in Tesco-specific ways by DunnHumby (and this reflects a longer history), they are also broadly typical. Tesco is the largest supermarket chain in UK, famous amongst retailers for its customer loyalty and targeted marketing programme known as 'Tesco Clubcard,' which started in 1991. DunnHumby -- founded by operations researchers Edwina Dunn and Clive Humby -- is said to have convinced the CEO of Tesco sometime in 1991 that a loyalty card program could change the supermarket chain's relationship to its customers. Clive Humby's academic publications are hard to track down, but an early paper given at the *Conference of Young Operational Researchers* in Nottingham in 1984 suggests the direction that he, DunnHumby and later Tesco would take in constructing lists (O’Keefe, [1984](#ref-OKeefe_1984)). The abstract for Humby's presentation pre-figures an ongoing trajectory for data mining techniques aimed at eliciting detailed information on individual customer preferences.



Abstract from a Clive Humby presentation (Humby,1984)

While Tesco succeeded in data-mining its customers using demographic segmentation, and perhaps became the UK's biggest supermarket with the help of data-mining in the 1990s, the shopping environment in 2017 is markedly different (Turow, [2017](#ref-Turow_2017)). It is no longer organised around campaigns involving special offers or redemption of points for demographically segmented loyal customers (DunnHumby made heavy use of UK Census data). It can no longer rely on placement of goods in carefully chosen locations in stores. It needs or at least might want to continuously up-sell and cross-sell to customers who might hardly ever visits the supermarket itself.

How could we characterise the shift from ClubCard data mining to online grocery markets? If 'Tesco is the clear winner in the online grocery market, in fact it takes almost 50p of every £1 spent on food shopping on the internet' [How SEO helps Tesco to dominate the online grocery market](https://econsultancy.com/blog/64841-how-seo-helps-tesco-to-dominate-the-online-grocery-market)(Silverwood-Cope, [2014](#ref-Silverwood-Cope_2014)), then Tesco itself has undergone some kind of conversion? Patel described her work at DunnHumby as shifting the recommender system from a 'rules-based list' to a 'relevancy model.' The relevancy model changed how the 'Have You Forgotten' list was constructed. This shift is a typical of broader re-organisation of prediction in which existing data analysis practices are being re-distributed and intensified in particular ways. The shift in models results in a more probabilistic structuring of lists. The ostensible banality of the 'Have you forgotten?' recommendations betrays micrological signs of a topologically complex predictive infrastructure that intervenes in list-based social ordering to reconfigure alignments between people and things.

This is not a drastic or abrupt event. Academic researchers first began writing about recommender systems in the mid-1990s. From the outset they highlighted a potential shift from demographic-driven market research or data mining techniques to personalized recommendations. For instance, writing in 1997 in a special section of the *Communications of the ACM* on recommender systems, Paul Resnick and Hal Varian (at that time Dean of Information Sciences at UC Berkeley, but currently Chief Economist at Google), made much of this personalisation (Resnick and Varian, [1997](#ref-Resnick_1997)). Resnick and Varian emphasise the need to distinguish the emerging practices from data mining:

In everyday life, we rely on recommendations from other people. ... Recommender systems augment this natural social process. In a typical recommender system, people provide recommendations as inputs, which the system then aggregates and directs to appropriate recipients. In some cases the primary transformation is in the aggregation; in others the system’s value lies in its ability to make good matches between the recommenders and those seeking recommendations.(Resnick and Varian, [1997](#ref-Resnick_1997): 56)

Writing just after the advent of web-based e-Commerce, Resnick and Varian imagined recommender systems augmenting the 'collaborative filtering' that people do when they write and read reviews of products and services (for instance, on the travel website, TripAdvisor; (Scott and Orlikowski, [2012](#ref-Scott_2012))). In 1997, Resnick and Varian expected that 'people [would] provide recommendations' and recommender systems would aggregate and rank recommendations for recipients. A slightly later review, (Schafer et al., [2001](#ref-Schafer_2001)), diagrams an augmented 'natural social process' with a range of elements, technologies, inputs and outputs, with varying degrees of personalization.

# E. Experimenting with probabilistic conversions

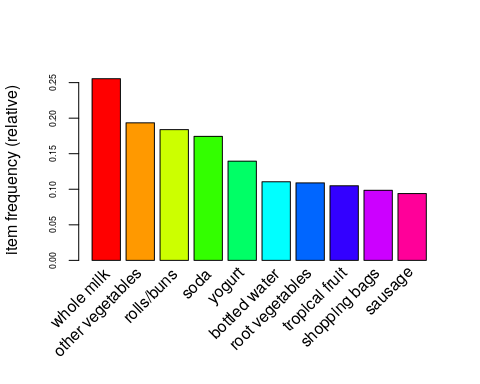
This section of the paper returns directly to Patel's presentation, and the character of the newly implemented recommender system. In the discussion that follows, the description of components of the new recommender system -- predictive models and their parameters, infrastructural provisioning, product similarity matrices -- and the problems of recommendation -- too many products, changes over time, unstable propensities of things in their associations with people -- together suggest that personalization, and the 'personal relevance model' have a troubled relation to social order. I develop an analysis centred on probabilities and processes of probabilisation -- rendering situations calculable as probabilities -- in exploring Patel's account of the recommender system. The motivation to focus on probabilities and probabilisation is explicitly philosophically reconstructive. It represents an attempt to reframe the operation of the recommender system in a way freed from lingering incompatibilities between calculation and social life (Dewey, [1957](#ref-Dewey_1957): 26) .

In order to do this, we need a working notion of probability. I draw directly on the work of Karl Popper for this. In an essay written towards the end of his career, Popper presents an alternative account of probabilities as real processes, in the same way that contemporary sciences might see forces as real (Popper, [1990](#ref-Popper_1990)). Somewhat counter-intuitively, Popper does not identify probability with either degree of belief (likelihood) or frequency of events. Instead he suggests 'they should be regarded as *inherent in a situation* ' (Popper, [1990](#ref-Popper_1990): 14). On this account, probabilities are tendencies towards realization inherent in a situation. Probabilities have or indeed are *propensities*, tendencies to realize the event (11). While this somewhat abstract philosophical account of probability as propensity might seem to be anchored far away from the concerns of online grocery shopping, Patel's presentation of the personal relevance model and the ways in which it seeks to calculate probabilities of purchase. The series of prototyping reconstructions I am about to discuss will explore the tenability of this rather radical alternative to regular narratives of personalization.[[5]](#footnote-6)

## Apriori and conditional probabilities

Popper suggests that one of the reasons probability has been difficult to work with philosophically, even as it pervades scientific thought, is that a certain limited domain of probability practice focused on dice rolls, coin tosses, urns with balls and other seemingly random events has occupied centre stage in concepts of probability. Even if it has been enormously productive and transformative, it is insufficient because it privileges *absolute* probability at the expense of *conditional* probabilities. 'We need,' Popper urges, '*a calculus of relative or conditional probabilities* as opposed *a calculus of absolute probabilities*' (Popper, [1990](#ref-Popper_1990): 16). Relative or conditional probabilities are propensities that depend on other events for their own realization. All events require other events, so all probabilities are conditional, even if probability calculations have typically preferred to determine the weight of possibilities devoid of the inevitable physical conditioning of their realization.

Viewed from the standpoint of probabilities, many predictive systems are highly crafted arrangements for the calculation of conditional probabilities. For instance, the first element of the new recommender system consisted in a change of the underpinning algorithms and model. Patel described a move from 'a rules-based system' to a 'personal relevance model' as the basis of the new recommender system. It is likely that what Patel describes as the 'rules-based system' refers to the extremely well-known apriori algorithm or association rules learning technique, developed by computer scientists Rakesh Agrawal and Ramakrishnan Srikanti working at IBM Research Alameda in the early 1990s (Agrawal et al., [1994](#ref-Agrawal_1994)). A now-classic approach to 'market basket analysis,' it was listed as one of the top ten data mining algorithms in a survey conducted amongst data miners (Wu et al., [2008](#ref-Wu_2008)) and usually attracts a chapter in data-mining and machine learning textbooks (e.g. (**???**)). The interest of apriori for our purposes is that it begins to address the problem of understanding large numbers of shopping transactions as a matter of conditional probability.



Frequency of association betwene items in the Grocery dataset

The notion of conditional probability at work in aprior is relatively simple, and assumes that the frequency of transactions provides the best guide to what shoppers are likely to do. The apriori algorithm finds sets of items that commonly occur together in transactions. In this sense, is still oriented by a notion of absolute probability, inflected by some elements of conditional probability. Commonly occurring sets are expressed as 'association rules.' For instance, using the Groceries dataset in the R package arules, the apriori algorithm counts frequencies of purchase in the overall set of all items purchased in a supermarket (the Groceries dataset was acquired from a 'local German supermarket' (**???**)). Figure @ref(fig:arules-ex) shows how often the most common items appear. Whole milk appears most frequently.

## lhs rhs support confidence lift  
## [1] {whole milk} => {other vegetables} 0.075 0.29 1.5   
## [2] {whole milk} => {rolls/buns} 0.057 0.22 1.2   
## [3] {whole milk} => {yogurt} 0.056 0.22 1.6   
## [4] {whole milk} => {root vegetables} 0.049 0.19 1.8   
## [5] {whole milk} => {tropical fruit} 0.042 0.17 1.6

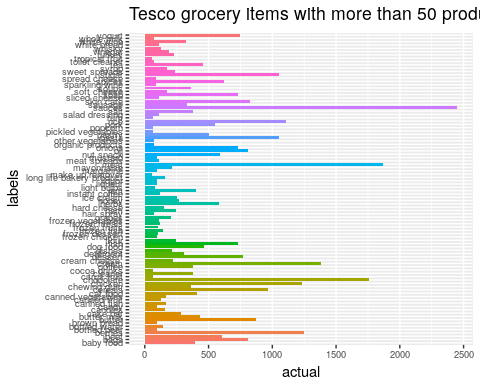
The first five association rules for the Groceries dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lhs |  | rhs | support | confidence | lift |
| [1] | {whole milk} | => | {other vegetables} | 0.07 | 0.29 | 1.5 |
| [2] | {whole milk} | => | {rolls/buns} | 0.06 | 0.22 | 1.2 |
| [3] | {whole milk} | => | {yogurt} | 0.06 | 0.22 | 1.6 |
| [4] | {whole milk} | => | {root vegetables} | 0.05 | 0.19 | 1.8 |
| [5] | {whole milk} | => | {tropical fruit} | 0.04 | 0.17 | 1.6 |

The association rule-based model for the Groceries dataset illustrates some characteristic ordering practices of 1990s-style data-mining approaches to lists. The association between items is expressed in the form of *rules* whose importance is can be ranked by how frequently items are found in the same purchase. Although the apriori model does not use the language of probability calculus, the 'rules,' a term derived from older decision support literature, rank associations between items. When we run the apriori algorithm on the Groceries dataset, the resulting set of rules (shown in Table @ref(tab:arules-ex2)) indicates that milk and vegetables has a stronger association than milk and buns. The calculations of support (how frequently the association appears in the dataset), confidence (how often the rule applies in the dataset) and lift (a ratio between the support and the independent frequencies of the items) in the table attempt to measure the strength of these associations in different ways.

## 2500 sauces: Apriori meets the API

Even as apriori elicits associations between things, it struggles with the propensity of commodities to multiply. A simple illustration of the combinatorial problem faced by recommender system can be developed by bringing the Grocery dataset together with the actual list of items that sells in its online grocery shopping site. If we take all the items in the Grocery dataset and paste them into the 'shopping list' box on the Tesco grocery website (or as I did, run them as searches on the TescoLab Product API (Tesco, [2016](#ref-Tesco_2016))), each of the 169 items in the Grocery dataset yields dozens and sometimes thousands of products from Tesco online.



Tesco surplus

Items in the Groceries dataset proliferate into a Tesco's list of branded products. The 169 items of the Groceries dataset expand into roughly 38114 Tesco items (see Figure @ref(fig:tesco-ex)). There is, I would suggest, a logistic proliferation that recommender systems encounter. The multiple listed in the Grocery dataset becomes more open in this setting through new impersonal intersections that could find themselves in associations of various strengths. If we think just of the almost 2500 sauces or 1200 rice products supplied by the Tesco API, then a tremendous number of associations between sauce and rice are possible. The propensities of actual sauce and rice products to find themselves together in an individual purchase vary greatly. The proliferation of things on the shelves of supermarket or grocery warehouse confronts produces a combinatorial problem for data-mining machine learning approaches such as association rules. 38114 products (actually Patel mentioned 200,000 products) can be combined in a vast number of ways. If a typical shopping list has 20 items, then there are 1.711594e+73 possible lists. Some of this vast number of possible shopping lists have propensities or tendencies to realization that are close to zero. But others with somewhat higher propensities might furnish the basis of interesting recommendations.

More importantly, the limited conditioning of probability of the association rules approach suggests a compelling reason for personalization. Even if the association rules and the apriori algorithm generate recommendable sets of frequency-weighted associations, an apriori-based recommender system contains no statistical model. Its probabilisation of shopping is incomplete since it only works on associations between things. The tendency of some milk to find itself in a shopping basket alongside bread attests to an important commonality of practice in certain parts of the world. But the conditional probabilities implicit in the model do not bring in much of the world and its teeming sieve of propensities moving towards realization. These associations are not trivial, but they are very open-ended. Put in terms of the big data conversion, the rules-based system lacks, in principle, a means of crystallizing a limited or enclosed set of possibilities.

## The list in its mapping to probabilities

Predictive models are central technical elements in many big data arrangements because they provide a way of constructing conditional probabilities that incorporate more of the situation in which any propensity -- to purchase a bottle of Vueve Cliquot champagne for instance -- is inherent. In the new recommender system described by Patel, the business goal is to extend the list of the items customers have selected for purchase with a few well-chosen items. In order to do this, the list of items selected for purchase will be extended by recommendations that have, according to a predictive model, the most chance of 'conversion'; that is, actual purchase. The predictive model carries the burden of calculating the conditional probability of purchase given everything known about the situation at this moment in time.

Patel introduced the new 'personal relevance model' with a data graphic familiar to machine learners and statistical modellers. (A sketch of her graph appears in Figure @ref(fig:models).) The graph plots the *precision* -- the proportion of the recommended products that customers actually purchase -- for several different statistical models. The graph indexes the 'causal efficacy' of the recommender system, its capacity to include and transform propensities or 'real potentialities' into operational events or purchases. Although she did not dwell on it, Patel used the graph to compare the previous rule-based recommender systems with some of the 'personal relevance model' alternatives -- logistic regression, random forests, gradient boosting and a few others -- in terms of their predictions and how those predictions turned out. Patel dismissed most of the models quite quickly and focused only on one, the logistic regression model, which did as well or slight other than other alternatives.

The models Patel so briefly mentioned are all 'classifiers,' predictive models that 'classify' particular events or outcomes by calculating their probability of membership of some class of event or outcome. Much of the core architecture of the machine learning classifier models received scant reference in Patel's presentation. It is most likely that in DunnHumby's work, the classifier at the heart of the recommender system calculates for a given customer the most likely products to be purchased. The classes of events are binary: recommended or not recommended. Products are allocated to one of these classes depending on their calculated probability. A probability greater than 0.5 typically would be recommended. The logistic regression model generates probabilities of purchase for each product for each customer.

The crucial point here is that the shopping list written by the customer is extended, quasi-invisibly, in the recommender system to potentially any of the 250,000 items in the Tesco inventory. But the basis of this re-writing of the grocery list is an intricate conditional probability statement. Even if Patel assumed that the audience understood the working of the logistic regression model, the bulk of her presentation concerned the obstacles and problems that arise in trying to probabilise recommendations in ways that lead to the much-desired 'conversion events' or sales.

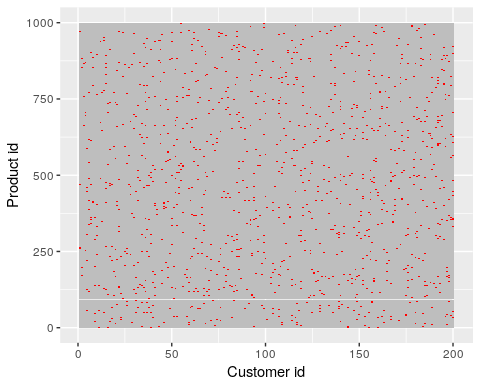
## Repeating sufficiently often to measure

I will not discuss in detail the different dimensions of probabilisation implicit in the architecture of the new model. It is worthwhile, however, highlighting some key elements that point directly to the troubled notion of 'personal relevance' and personalization more generally. These include the problem of repetition, attempts to create new views of the data, experiments on the platform, and the temporalities of calculation.

One fundamental difficulty here is that the propensity of a customer to buy a recommended product at a particular point in time is changing. Popper's notion of probabilities as propensities as tendencies to realization inherent in a situation directly implies this. Measuring propensities and expressing these measurements as numbers between 0 and 1 becomes difficult because the situation is change and the propensities themselves change. As Popper observes, 'the propensities cannot be measured because the relevant situation changes and cannot be repeated' (Popper, [1990](#ref-Popper_1990): 17). No doubt the contingencies associated with grocery shopping are legion, perhaps more so than other kinds of products (books, music, films), and any attempt to measure propensities associated with particular commodities will encounter many dynamic changes.

One way in which recommender systems attempt to address the problem of the relational dynamics of propensities is by looking for repetitions. Patel mentioned that the new model uses '52 weeks of data' for each customer. In including this data in the model, the assumption is that the probability that a specific customer will buy a specific product will be shaped by previous purchases. The history of previous purchases constitutes forms of repetition that imply a stabilisation of propensities (high or low; e.g. a vegan customer will never have purchased chicken products, so the measured propensities for any of the 1000 or so chicken products sold be Tesco will remain close to 0).

When the logistic regression model includes the 52 weeks of previous purchases, the conditional probability calculation undertaken by the recommender system ramifies tremendously in several respects. Each of Tesco's 200,000 products becomes a variable in the classifier. Practically, most of the these variables will not influence the calculation of the probabilities for recommendations very much. As Patel observed, again assuming that this would be obvious to the audience of data scientists, 'we have lots of zeros'. -- points to the difficult terrain of 'good feature' thinking: the vast, nearly empty matrix of customer-product associations. Given that most things in Tesco remain relatively unknown to each other and to customers. A matrix that records associations between individual people and different products is bound to be mostly empty. Say Tesco has 1 million online customers. Each online shopper has bought some selection of the 200,000 products. The customer-product data matrix will be 2e+11 in size. The product-customer matrix, the basic vector-space in which all recommender systems operate, remains very sparse and unpopulated. Given that any one customer is likely to only have bought 100 or so different products, the probability matrix will be close to 99.95% empty. The data is, as Patel put it, 'massively unbalanced' (any dataset where the items of interest are much rarer than some normal values is said to be unbalanced) and such imbalance would heavily bias the recommender system towards common and only impersonally relevant suggestions, suggestions that might not produce the desired conversion experience.



Sparsity of the customer-product association matrix

Since so many people buy milk, the recommendation system might end up always recommending milk products. So the data needs to be 'corrected' by, as Patel reports, removing -- under-sampling -- some of the data for common purchases. Secondly, even if a recommendation system finds uncommon items that are good recommendations, they should neither be too cheap ('low spend'), nor, as we will see too similar to what the customer has already ordered.

Many other slight adjustments, tweaks and local modifications would probably occur around the logistic regression model. The model's recommendations themselves have propensities that need to be adjusted to the scale of values of grocery retail, to the available infrastructures, and to the capacities of the online grocery system to merge recommendations into a grocery order in a timely fashion. While raw data from Tesco Online feeds into the model every hour, a recommendation list for each customer is only generated once a week. Customers only shop online every few days at the most, and in some cases, only every few weeks. To continually update the top 200 recommendations for every customer would be computationally expensive. Patel briefly mentioned specific infrastructural elements such as hadoop (Apache Software Foundation, [2009](#ref-ApacheSoftwareFoundation_2009)). The logistics of running a predictive model affect listing practices. The possibility of adjusting the recommendations for every customer weekly derives from an infrastructure capable of collecting data, assimilating that data to a predictive model able to generate approximately 200 million relevant recommendations, where 'relevance' depends on a matrix of probabilities of associations between people and things that shifts in time. Hadoop and its legion of 'big data' variants (mahout, spark, hive, pig, yarn) operationalise flows of data at an infrastructural rather than analytic scale. In many respects, the quickly glossed-over infrastructural deployment of DunnHumby's relevancy model is the primary conversion event: the logistic regression model at the heart of the recommender system is no longer an analytical device but an operational one.

## Platform Experiments that reduce interfering propensities

In his account of probabilities as physical propensities, Popper emphasises the specific configurations established in laboratory experiments:

experiments work ... by creating, at will, artificial conditions that either exclude, or reduce to zero, all the interfering and disturbing propensities (Popper, [1990](#ref-Popper_1990): 23).

Like many accounts of laboratory practice (**???**, **???**, (**???**), **???** - epistopics ref), Popper understands experiments as stabilizing the conditions so that phenomena can be saved from their own variability. Experiments play an important role in conversion events, or in the process of bring new predictive models to bear on shopping practice. A crucial consideration for a predictive model in a recommender system will be how to experimentalize its own predictions so that they are not interfered with or disturbed by other propensities. This is obviously an inordinately difficult challenge, since, as we have seen, grocery shopping takes place at the intersection of normative and actual social orders, individual and group belongings, and logistically global economic processes.

|  |  |
| --- | --- |
| Test A | Test B |
| Control A | Control B |
| ------ | ------ |

Nonetheless, experimentalization, the practice of creating conditions that reduce disturbing propensities, is very much part of the convernsion between models. For instance, the predictions of the recommender system themselves are the subject of experimental. Patel described the deployment of the relevancy model experimentally in a random A/B controlled trial on the Tesco website. All customers were allocated to one of four categories as shown in the table. In the A/B trial, an identical shopping list, or even the same person, might receive recommendations from different models. The randomised application of competing predictive models, draws on protocols for randomised clinical trials first developed in the 1960's, and is widely used in social media platforms and hence in the implementation and observation of the effects of recommender systems.[[6]](#footnote-7) Random allocation of customers to the four categories (Test A, Test B, Control A, Control B) adds a layer of probability to the recommender system in the name of statistical validation of the effects or the 'uplift' of the model.[[7]](#footnote-8) Ironically, the effects of a predictive model cannot be known in advance. They can only be observed experimentally.

It is perhaps significant that this random allocation of customers to control and test groups occurs without any relation to the particular profile or propensities of the customer. It seeks to statistically validate the effects -- the *uplift* -- of the model on conversion events by allowing the effects of the model on what people do to be measured. The uplift refers to conversion events associated with the same groups of people. Effectively an experiment in inhabiting two different worlds, the randomised control trial sets up a feedback look between the predictive model (the logistic regression), and the world.

## The openness of the data -- new possibilities

Whatever disturbing factors or interferences the experimental trials of different predictive models reduce, a broader and ineluctable stream of propensities remains in operation. Popper's account of probabilities as propensities presents openness as concomitant to realization:

What may happen in the future .. is, to some extent, open. There are many possibilities trying to realize themselves, but few of them have a very high propensity, given the initial conditions (Popper, [1990](#ref-Popper_1990): 22).

There are two ways of reading this statement. One is to see Popper as simply stating one of the consequencies of his physicalist understanding of probabilities as tendencies inherent in a situation. Another is to focus on the pivotal phrase 'possibilities trying to realize themselves' by asking what 'trying to realize' might mean. I would suggest that Patel's description of DunnHumby's work on the new personal relevance model is just such a 'trying to realize' in important ways.

Several times, Patel emphasized the importance of 'constructing good features' from the data, and much of her presentation concerned different attempts to construct 'good features.' A 'feature' in the context of machine learning refers to a variable included in a predictive model. A 'good feature' is one that somehow improves the accuracy, precision, specificity or any of the panoply of measures applied to machine learning models in practice. The precise meaning of 'good', however, remains open to many different attempts, experiments and variations. I find it interesting that the main efforts that DunnHumby made to construct good features were not closely associated with personhood, but with the relations -- the propensities -- between things.

Patel, for instance, addressed the problem of 'basket similarity,' the problem that the recommender system might recommend something similar to items already in a customer's basket. A customer might be willing to substitute a similar item for something they have already chosen, but they are more likely to accept a recommendation for an item that complements rather than duplicates items in their list. How could the recommender system sense the possible substitutions and complementaries between things for a given customer, especially since a person's sensibilities and susceptibilities concerning groceries will be subtly shaped by wider social groups, orderings and circumstances? It could only do that if had some sense of the relation between items already in the list and the field of possible recommendations. The 'personal relevance' model, therefore, sought to include the relation between recommendations and items already ordered. The data scientists constructed a new feature from the data concerning how the substitutabilities and complementarities between product. Taking all the baskets of items purchased, rather in the way that association rules does, a new feature, what DunnyHumby termed 'self-learning substitutes' was derived. The 'self learning substitutes' feature of the relevancy model draws on another data matrix, the product similarity matrix, itself generated from all previous recommendations that have led to conversion events (that is, 'a user clicked as the response'). A customer's recommendation list might already include Tesco Ginger Nuts 300g. The 'self-learning substitution model' seeks to replace list items with similar but higher value items: Mcvitie's Ginger Moments 400g.

This new feature is complex. Patel mentioned that it took the form of a 'design matrix of 14,000 columns.' This means that 14,000 new explanatory variables were added to the logistic regression model producing the recommendations. Each recommendation would be subtly re-weighted by this complex derived feature. The complexity of the model itself introduces new instabilities, new 'possibilities trying to realize themselves' and the DunnyHumby data scientists used 'L1 regularization to drive coefficients down to zero', or, put more simply, to exclude variables there were not contributing much to the predictions produced by the model.

Even this added feature, a complex addition in many respects, does not exhaust DunnHumby's attempts to re-shape the recommender system. They were just beginning, Patel reported, to address complementarities between things. Nutella complements bananas, but the converse is not necessarily true since bananas need not complement nutella, especially if I have nut allergies. From the standpoint of customer data science, Patel emphasised the operational significance of predicting complementarities: 'complementarity drives conversion.' Bringing the matrix of similarities expressed as numerical associations between products into the model improves the conversion of recommendations into purchases.

# F. Conclusion

The 'minute and extensive scrutinizing' (Dewey, [1957](#ref-Dewey_1957)) of a recommender system undergoing a conversion from an older data-mining approach to a current 'personal relevance model' approach addresses a specific situation: online grocery shopping (not shopping in general) and, even more micro-sociologically, the ordering of lists.

The theory of shopping lists constructed at the beginning of the paper was a scaffolding component. It draws on anthropology, sociology and mathematics to suggest that we see shopping lists as the site of a complex negotiation between different social orders, between norms and actualities. The list itself as an inscriptive device functions as an ordering device, a device that accommodates an immense variety of specific habits, dispositions, values and associations. The mapping element of the list, the order that arranges the set of items according to different interests, values, and aims, is the operational focus of recommender systems.

The personalization of recommendations has been a defining characteristic of many big data conversion events. Rightly, the effects of all-consuming personalization have been the target of critics and critical research. The archaeology of the Tesco recommender system, including its early experiments with demographic data-mining, its adoption of a rule-based 'market basket' analysis system, and its current implementation of a 'personal relevance model' certainly follows that trajectory. But the conversion event is not reducible to personalization.

Probabilization precedes personalization. When we observe what is practically done to construct a personal relevance model and apply it to a shopping list, we see, conditional probabilities, predictive models, the importance of data accumulated over time, platform experiences, the rhythms of infrastructural maintenance, and open-ended construction of new features in data, all of which negotiate shifting and complicated associations between people and things, people and people, and things with things. If social order is made of associations, is to be social is to associate, then the forms of association taking shape in this and other similar conversion events might be understood as understand as operationalizing probability in matrices of propensity. As Popper argues, probabilities are being realized in the same way that fields of force were realized in modern physics. Probabilization does not take place in physics laboratories but on platforms, combining databases, compute capacity, interfaces and global supply chain logistics.

The fact that probabilisation largely takes the form of personalization tends to obscure the re-framing of social order taking place. This ordering depends on the mappings that inscriptive devices such as shopping lists display. Only because shopping is already somewhat ordered by lists can a partial re-mapping, provisionally conducive to new events, take place. The conversion entailed in this re-mapping is not just about personalization. It is more ontological than that since it includes relations running multi-laterally between people and things in time. It is less ontological, however, in that it is not a sudden transformation, but an ongoing process of adjusting, scaffolding, intervening and configuring orders.

Reconstructions of specific conversion events associated with big data can work in different ways. The combination of social theory, archaeology, philosophy and model/data prototypes used in this paper sketches one possible way of addressing the trouble, the disturbance or instability that seems to affect many discussions of big data.

# References

Agrawal R, Srikant R and others (1994) Fast algorithms for mining association rules. In: *Proc. 20th int. conf. very large data bases, VLDB*, pp. 487–499. Available from: <https://www.it.uu.se/edu/course/homepage/infoutv/ht08/vldb94_rj.pdf> (accessed 28 September 2016).

Almquist E and Wyner G (2001) Boost your marketing ROI with experimental design. *Harvard Business Review* 79(9): 135–141. Available from: <http://offermaxima.com/pdfs/BoostYourROI.pdf> (accessed 10 October 2016).

Apache Software Foundation (2009) Welcome to Apache Hadoop! Available from: <http://hadoop.apache.org/> (accessed 30 October 2009).

Bogost I (2012) *Alien phenomenology, or what it’s like to be a thing*. U of Minnesota Press. Available from: <http://books.google.co.uk/books?hl=en&lr=&id=MwaK2aUclo8C&oi=fnd&pg=PP2&dq=ian+bogost+alien&ots=bRcDxTYd2w&sig=IyStFB_1583RBe48OKZzhEh5_h8> (accessed 7 August 2013).

de Goede M, Leander A and Sullivan G (2016) Introduction: The Politics of the List. 34(1): 3–13. Available from: <http://dx.doi.org/10.1177/0263775815624561> (accessed 15 February 2017).

Dewey J (1957) *Reconstruction in Philosophy*. 1st Beacon paperback ed. Boston: Beacon Press.

Dunn J (2016) Introducing FBLearner Flow: Facebook’s AI backbone. *Facebook Code*. Available from: <https://code.facebook.com/posts/1072626246134461/introducing-fblearner-flow-facebook-s-ai-backbone/> (accessed 13 March 2017).

dunnhumby (2017) Dunnhumby. Available from: <https://www.dunnhumby.com/> (accessed 17 March 2017).

Foucault M (1972) *The archaeology of knowledge and the discourse on language*. New York: Pantheon Books.

Gillespie T (2010) The politics of ’platforms’. *New Media & Society* 12(3): 347–364. Available from: <http://nms.sagepub.com/cgi/doi/10.1177/1461444809342738> (accessed 5 September 2014).

Goody J (1986) *The Logic of Writing and the Organization of Society*. Cambridge University Press.

Hansen M (2015) Prediction in the Wild. In: Grusin RA (ed.), *The nonhuman turn*, 21st century studies, Minneapolis & London: University of Minnesota Press.

Lazzarato M (2014) *Signs and Machines: Capitalism and the Production of Subjectivity*. Cambridge, MA: Semiotext (e). Available from: <http://mitpress.mit.edu/books/signs-and-machines> (accessed 17 February 2015).

Lo VS (2002) The true lift model: A novel data mining approach to response modeling in database marketing. *ACM SIGKDD Explorations Newsletter* 4(2): 78–86. Available from: <http://dl.acm.org.ezproxy.lancs.ac.uk/citation.cfm?id=772872> (accessed 10 October 2016).

Lury C, Parisi L and Terranova T (2012) Introduction: The Becoming Topological of Culture. *Theory, Culture & Society* 29(4-5): 3–35. Available from: <http://tcs.sagepub.com/content/29/4-5/3> (accessed 5 December 2012).

Mainberger S (2003) *Die Kunst des Aufzählens: Elemente zu einer Poetik des Enumerativen*. Reprint 2011. Berlin ; New York: De Gruyter.

Marcus G (2014) Prototyping and Contemporary Anthropological Experiments With Ethnographic Method. *Journal of Cultural Economy* 7(4): 399–410. Available from: <http://dx.doi.org/10.1080/17530350.2013.858061> (accessed 10 October 2016).

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Boston: 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).

Miller D (2012) *Consumption and its Consequences*. Cambridge ; Malden, MA: Polity.

Morris JW (2015) Curation by code: Infomediaries and the data mining of taste. *European Journal of Cultural Studies* 18(4-5): 446–463. Available from: <http://journals.sagepub.com/doi/abs/10.1177/1367549415577387> (accessed 3 February 2017).

Neilson B (2012) Five theses on understanding logistics as power. *Distinktion: Scandinavian Journal of Social Theory* 13(3): 322–339. Available from: <http://www.tandfonline.com/doi/abs/10.1080/1600910X.2012.728533> (accessed 30 October 2013).

O’Keefe B (1984) Young Operational Research Conference. Abstracts. University of Nottingham, 3-5 April 1984. *The Journal of the Operational Research Society* 35(7): 659–671.

Pariser E (2011) *The filter bubble: What the Internet is hiding from you*. Penguin UK. Available from: <http://books.google.co.uk/books?hl=en&lr=&id=-FWO0puw3nYC&oi=fnd&pg=PT3&dq=eli+pariser&ots=g2PoCtpQV-&sig=3_CftKt2BPOwLVpT_OFzizJr_-c> (accessed 13 February 2014).

Popper SK (1990) *A World of Propensities*. Bristol: Thoemmes Continuum.

Resnick P and Varian HR (1997) Recommender Systems. *Commun. ACM* 40(3): 56–58. Available from: <http://doi.acm.org/10.1145/245108.245121> (accessed 26 September 2016).

Schafer JB, Konstan JA and Riedl J (2001) E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery* 5(1-2): 115–153. Available from: <http://link.springer.com/article/10.1023/A%3A1009804230409> (accessed 26 September 2016).

Scott SV and Orlikowski WJ (2012) Reconfiguring relations of accountability: Materialization of social media in the travel sector. *Accounting, organizations and society* 37(1): 26–40. Available from: <http://www.sciencedirect.com.ezproxy.lancs.ac.uk/science/article/pii/S0361368211001139> (accessed 10 February 2017).

Seaver N (2015) The nice thing about context is that everyone has it. *Media, Culture & Society* 37(7): 1101–1109. Available from: <http://mcs.sagepub.com.ezproxy.lancs.ac.uk/content/37/7/1101> (accessed 24 September 2015).

Silverwood-Cope S (2014) How SEO helps Tesco to dominate the online grocery market | Econsultancy. *Econsultancy*. Available from: <https://econsultancy.com/blog/64841-how-seo-helps-tesco-to-dominate-the-online-grocery-market/> (accessed 9 February 2017).

Tesco (2016) API – Tesco Labs. Available from: <https://www.tescolabs.com/category/api/> (accessed 20 July 2017).

Tsing A (2009) Supply chains and the human condition. *Rethinking Marxism* 21(2): 148–176. Available from: <http://www.tandfonline.com/doi/abs/10.1080/08935690902743088> (accessed 29 November 2013).

Turow J (2017) *The Aisles Have Eyes: How Retailers Track Your Shopping, Strip Your Privacy, and Define Your Power*. Yale University Press. Available from: <https://books.google.co.uk/books?hl=en&lr=&id=YUrJDQAAQBAJ&oi=fnd&pg=PT25&dq=joseph+turow&ots=mzp4lcnuOR&sig=bSt4dLs0Hun73HoNg5_yEhsecqk> (accessed 17 March 2017).

Turow J, McGuigan L and Maris ER (2015) Making data mining a natural part of life: Physical retailing, customer surveillance and the 21st century social imaginary. *European Journal of Cultural Studies* 18(4-5): 464–478. Available from: <http://journals.sagepub.com/doi/abs/10.1177/1367549415577390> (accessed 17 March 2017).

Wu X, Kumar V, Ross Quinlan J, et al. (2008) Top 10 algorithms in data mining. *Knowledge and Information Systems* 14(1): 1–37.

1. Sociology, Lancaster University [↑](#footnote-ref-2)
2. More than 100 Data Science Institutes have been set up in North America, Europe and UK since around 2012. The traffic between higher education and data analytics in industry is intense and flows in several forms: people funding, research findings, software and technical devices (code) and training. [↑](#footnote-ref-3)
3. The shelves of large contemporary supermarket are the endpoint of global logistic supply chains, in all their logistical, value-transforming and brand-mediated hypercomplexity (see (Neilson, [2012](#ref-Neilson_2012); Tsing, [2009](#ref-Tsing_2009))). Groceries imply a planetary geography of agriculture, industry, transport, communication and financialisation animated by flows of labour and capital. Encounters between this hyper-complex commodity-geography and people, even in the familiar confines of a supermarket, are no simple matter, either for shoppers or for supermarket operators such as Tesco. [↑](#footnote-ref-4)
4. While lists intersect strongly with other meaning and sense-making practices in everyday life and popular culture, they can be refractory to discourse and textual analysis methods built around models of language or speech, with its rules of syntax and grammar. On the one hand, lists are highly associative. They can be interpreted or decoded semiotically, although they exhibit variations in textuality -- how they are written and read -- that thwart semiotic readings. On the other hand, as operational inscriptions, they can be treated ethnomethodologically, as the production of social order in a given setting. We might, in the light of their mutability, permeability and embedding in social order, approach lists from various theoretical angles -- as asignifying semiotics as Maurizio Lazzarato calls it in his *Signs and Machines* (Lazzarato, [2014](#ref-Lazzarato_2014)) or as elements in 'a new order of spatio-temporal continuity for forms of economic, political and cultural life' as Celia Lury puts it in her account of topological turn (Lury et al., [2012](#ref-Lury_2012)). [↑](#footnote-ref-5)
5. The media theorist Mark Hansen has recently applied Popper's account to argue that 'predictive analytics are discoveries of micrological propensities that are not directly correlated with human understanding and affectivity and that do not by themselves cohere into clearly identifiable events' (Hansen, [2015](#ref-Hansen_2015a): 111–112). Hansen's account differs from many widely shared views of big data. In the face of the so much personalization (and Patel's presentation exemplifies this), Hansen attributes a somewhat impersonalizing force to big data. If much predictive practice attempts to elicit 'micrological propensities' from data, Hansen links data to things via probabilities: 'whatever explanatory and causal value predictive analytics of large datasets have is, I suggest, ultimately rooted in this ontological transformation whereby probabilities are understood to be expressions of the actual propensity of things' (Hansen, [2015](#ref-Hansen_2015a): 120). Hansen theorises big data or predictive analytics as an 'ontological transformation' that deploys probabilities in an evermore closely woven and encompassing expression of animated, eventful, propensities of things. Predictive analytics links probabilities as calculations to propensities or the mutable associative agencies of things. The practical and indeed empirical question is whether such transformations or 'conversion' in relations between probabilities and 'the propensities of things' can be detected and articulated in the prosaic setting of shopping lists and recommender systems. 'Conversion' is a preferable term for these changes. Not much hinges on the choice of terms, but 'conversion' happens to be the term used by Patel. It highlights re-orientations in subjects, experience, things, numbers and infrastructures that ranges more widely and are more grounded than Hansen's 'ontological transformation' formulation. [↑](#footnote-ref-6)
6. It is difficult to gauge how much. Facebook reports that it has more than 1 million models operating in its infrastructure. (Dunn, [2016](#ref-Dunn_2016)) [↑](#footnote-ref-7)
7. See (Lo, [2002](#ref-Lo_2002)) for an introduction to the construction of experimental set ups for uplift models. Market researchers and data miners seem to have started adopting a statistically grounded experimental predictive practice sometimes around 2000. See for instance (Almquist and Wyner, [2001](#ref-Almquist_2001)) Response modelling, 'propensity' modelling and uplift modelling all seek to identify associations between 'treatments' or interventions and the 'Responders', the people affected by the treatment. [↑](#footnote-ref-8)