Chapter 1

Intelligent Food Design

1.1 Why design food?

Computers have transformed the design of everything from cars to coffee cups. Now the food industry faces the same revolution, with intelligent computer models being used in the design, production and marketing of food products. The combined market capitalisation of the world's biggest food, cosmetics, tobacco, clothing and consumer electronics companies is $2,000 million [45]. This market includes 16% of the world's 500 richest companies. Many of these "fast-moving consumer goods" companies[1] now apply intelligent computer models to the design, production and marketing of their products. Manufacturers aim to develop and produce high-volumes of these commodities with minimum costs, maximum consumer appeal, and of course, maximum profits. Most products have limited lifetimes following the fashions of the consumer-driven marketplace, requiring continual design and innovation. With food and drink, little is known about many of the underlying characteristics and processes: why do some apples taste better than others? How "crunchy" is the perfect apple? Product development and marketing must therefore be rapid, flexible and use raw data alongside existing expert knowledge.

1.2 Intelligent systems, product design, and the food industry

In order to define our area of interest more precisely, we now consider intelligent food design as the overlap between three fields: intelligent data analysis; the food industry; and product design. These are each substantial research domains in their own right, and time does not allow us to give more than a cursory survey of each. Figure 1.1

[1] Here, "fast-moving" refers to items that are purchased by large numbers of retail customers on an every day basis.

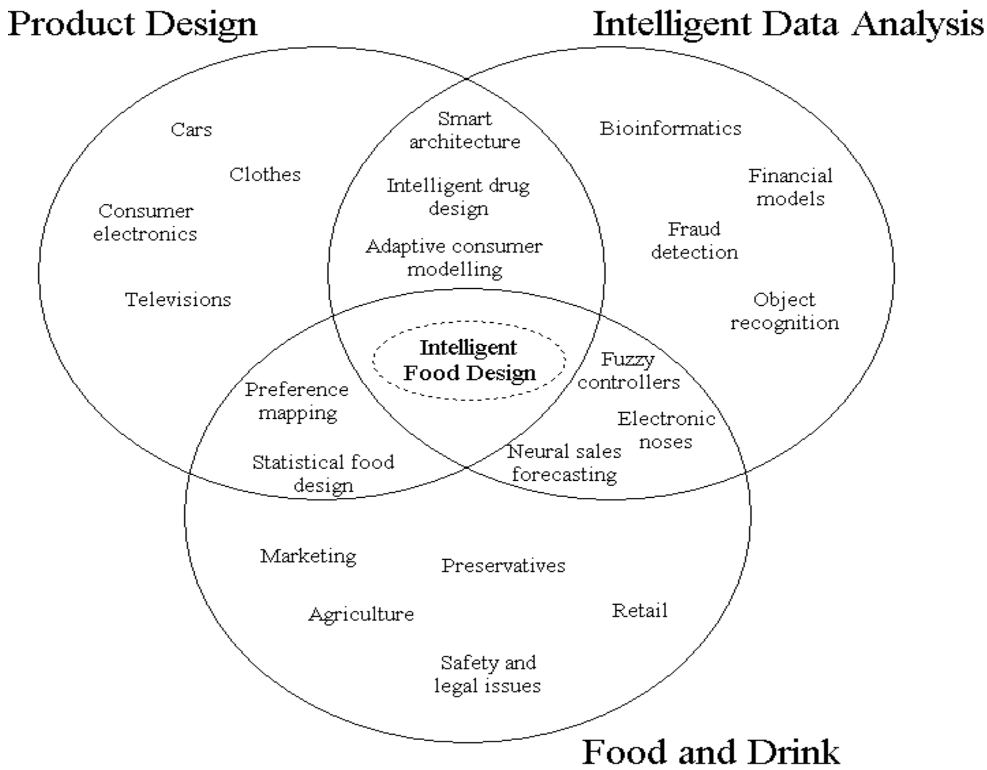

Figure 1.1: Three research domains

shows examples of each domain and each overlap of domains, to act as a map for the remainder of this section. Note that we have published parts of Section 1.2 in a review paper [30].

1.2.1 Product design

In this thesis, we are treating customer satisfaction as the principle behind good food design. However, the customer has not always been the concern of designers and manufacturers. At the height of his powers, Henry Ford famously said:

> "People can have the Model T in any colour — so long as it's black"

By the mid 1920's, Ford's insistence that one model was sufficient for all consumers had contributed to the end of the total dominance of the US car market that his company had enjoyed [47]. Competitors such as General Motors offered an increasingly wide variety of models, attracting customers who wanted glamour and style, rather than an old-fashioned workhorse.

Since then, industry has put the customer at the heart of product design, aiming to manufacture and market products that meet customers' demands. However, there is a continuing debate about how best to satisfy customers while keeping cost to a minimum. One influential approach is "quality engineering", proposed by Dr. Genichi Taguchi. Taguchi revolutionised product design in the 1980's, by changing the way quality is perceived. He said:

> The quality of a product is the (minimum) loss imparted by the product to society from the time the product is shipped. [123]

In this context, "society" refers to both the manufacturing company and their customers. The aim is to measure not only the direct cost to the manufacturer of production, but also the hidden costs of replacing or repairing faulty goods, warranty costs, customer dissatisfaction, resultant loss of market share and so on. The total cost is therefore the manufacturing cost plus the quality loss [96]. Minimising this total cost should result in satisfied customers and profitable manufacturers.

Taguchi defines the "quality loss function" (QLF) as a mathematical expression of the financial loss to a company due to the variability of its products. The QLF is defined as a quadratic, with a minimum cost corresponding to the ideal product and costs rising as the product characteristics move away from those of the ideal.

Note that there are no absolute specification limits, within which the product is deemed perfect and outside of which it is rejected. The goal in design and manufacturing then becomes to minimise the variability of the items produced.

Closely related to Taguchi's quality engineering is the principal of "robust design". Random errors are inevitable in any measurement and in any production process. It is therefore advisable to design products and processes robustly, so that the random errors have little effect [96]. Two of the main aims in robust design are to measure the effect of noise (e.g. using the signal-to-noise ratio), and to design experiments containing many factors (e.g. using orthogonal arrays).

Belavendram [10] describes an analysis of differences between Japanese and US Sony television sets, which inspired much of Taguchi's work. In the late 1970's, Sony ran two factories producing colour TV sets to the same specification: one in Japan and one in America. A study revealed that American consumers preferred the sets built in Japan to the ones built in America. One component was to blame: the colour density circuit. Although both factories produced circuits within the same specification limits, the distribution of the Japanese circuits was peaked, with most of them near the (optimal) mean. The distribution of the American circuits was almost flat, although still within the design specifications. By shifting the emphasis from specified bounds to a quadratic loss function, the perceived colour quality of the TV sets was increased.

Implicit in all of the work described above is the existence of a gold standard against which every item can be measured. Quality (high or low) is defined as how far from optimal a particular item is. But in the food industry, a key difficulty is identifying just what a "perfect" item is.

1.2.2 Food and drink

The food and drink industry could be stretched to include everyone from a subsistence farmer to a TV chef; this section is limited to discussing the history of the large-scale production of processed foods, including agriculture, ingredient selection, manufacturing and marketing. This historical background provides a context for our later discussions of the current food industry (Section 1.2.6), and food design in particular (Section 1.2.5).

Agriculture originated in many places at many times. One of the earliest steps was when animals, such as deer, were first domesticated around 20,000 years ago in the Middle East [41]. Oxen and sheep were soon added. Gradually, farmers changed from harvesting wild crops to planting and irrigating selected crops. Although the Egyptians ground grain to produce flour from around 8000 BC, it wasn't until 1776 that the first steam-driven mill was built in London, making flour milling one of the first modern industrial food process.

The combined effects of mass-production, greater urbanisation and a greater variety of readily available foodstuffs motivated research into food preservation. After noting that "an army marches on its stomach", Napoleon offered a large prize to the inventor of a practical food preserving method. In 1810, Nicolas Appert won the prize by showing that food could be preserved by placing it in a glass container, heating it, and then sealing the container. In England, a year later, John Hall replaced fragile glass with the more robust tin, producing the first commercial canned food [57].

In 1861 Louis Pasteur developed his eponymous sterilisation technique, which protects food by heating it to kill dangerous microbes, then removing the air and sealing the food in a container. More recently, flash-pasteurisation has been developed, which uses a short burst of high temperature, which not only sterilises liquids but also inhibits damaging enzyme action in beverages such as orange juice [57].

Modern freeze-drying techniques were developed in the early 20$^{\text{th}}$ century, shortly followed by the invention of the domestic refrigerator and freezer [31]. Birds Eye (a subsidiary of Unilever) drove the industry forward, moving from freezing fish and green vegetables to manufacturing value-added frozen foods[2], such as beef burgers, fish fingers and frozen ready meals. By expanding production through the 1950's, Unilever also found it necessary to create demand for these novel products through substantially increased advertising [31]. By 2001, ready meals were the fastest growing area of frozen food in the UK, with emphasis on high quality, (so called) authentic ethnic products,

[2] I.e. foods with increased economic value to the customer, due to processing, distribution etc.

with good value for money [22].

Cox et al. [31] also describe the innovation of chilled ready-meals in the 1980's. The process of freezing and de-frosting causes disintegration of the food, leading to poor flavour and texture. But chilling (rather than freezing) leads to very short shelf-lives, so chilled ready-meals require far more sophisticated supply chain management systems. With the introduction of electronic point of sale (EPOS) systems, and purpose-built computerised stock control systems, stock levels could be much more tightly controlled, minimising wastage. Furthermore, chilled ready-meals require microwave cooking, and so could only survive in the market once microwave ovens became widespread. This also required novel package design, including active packaging, which controls the amount of microwave radiation that reaches and heats the food. Thus many technological factors had to be in place before chilled ready-meals could attain the high-selling levels now enjoyed by the food industry.

1.2.3 Intelligent systems

Having considered a brief history of food production, we now turn to intelligent computer systems, before we combine these two themes in Section 1.2.6. Intelligent systems such as neural networks, fuzzy logic and genetic algorithms, mimic human skills such as the ability to learn from incomplete information, to adapt to changing circumstances, to explain their decisions and to cope with novel situations. These systems are being used to tackle a growing range of problems, from credit card fraud detection and stock market prediction to medical diagnosis and weather prediction [54].

Many advantages are commonly cited for intelligent systems, and the main theme of this thesis is to test whether or not these advantages apply to analysing food design data. The following list is adapted from Goonatilake and Khebbal [53]:

Ability to learn Intelligent systems can learn directly from data, without detailed prior knowledge.

Flexibility Intelligent systems can be applied to a wide range of problems.

Explanatory power Many intelligent systems are designed to produce human-readable outputs, such as fuzzy rule systems.

High- and low- level reasoning Some intelligent systems combine high-level strategies and goal-planning with low-level pattern recognition to solve complex problems.

Adaptability Many intelligent systems can respond to changes in a problem without having to start again from scratch.

Ability to deal with complexity Intelligent systems can work with high-dimensional problems, very large data sets and very complicated models.

No one claims that all of these characteristics apply to all intelligent systems. For example, many intelligent systems techniques are often regarded as black boxes, with little explanatory power (e.g. neural networks [85, p. 85]); others scale very badly as the number of features increase (e.g. rule induction [85, p. 280]). However, the ability to build models directly from data, when there is no known underlying process, is common to all intelligent systems, and gives them their power.

Many of the characteristics listed above make intelligent systems particularly applicable to the problems of food design. Every foodstuff has its own attributes, so no one model could ever encompass all food design issues, making flexibility beneficial. Many of the underlying aspects of food preferences are not fully understood, so learning directly from data with few preconceptions is an advantage. It is vital that the predictions of an intelligent system are understood, in order to justify changes to products or processes. Ideally, this understanding can be acquired with the help of a self-explaining system.

Recently, the emphasis of intelligent data analysis research has been on analysing very large data sets, driven by applications such as the human genome project, credit card transactions and stock market trading (e.g. Fayyad et al. [43]). However, food design research is often forced to use very small data sets. Past work on small data sets includes:

Space shuttle O-ring failure Just 23 records with five features were available from which to learn. The fatal failure in 1986 was not predicted largely because the analysis required a large degree of extrapolation, and failed to include correspondingly large error bounds. Later analysis, including Draper's [36], used more sophisticated Bayesian modelling which revealed very large error bounds at the point where the prediction had to be made.

Wage disputes Bergadano et al. [12] used information such as past pay rise levels, pension contributions and length of paid holidays to predict whether a proposed contract would be accepted or not. Concise rules were induced from the 57 records with 16 features that were available.

Classifying odours Baldwin et al. [4] used fuzzy logic to extract rules from an electronic nose. Rather than aiming purely for model accuracy, they developed a system to characterise odours using terms that people naturally use, such as "fruity" and "nutty".

Machine diagnostics Skurichina [119] compared the performance of several classifiers on a range of small data sets, including the prediction of mechanical faults and of maintenance requirements. The small training sets lead to unstable classifiers, and Skurichina described several stabilising methods, such as noise injection and boosting.

Note that although these examples describe analysis of small data sets, in most cases larger sets would be available if required. For example, the fuzzy odour describer [4] is currently being applied to a far larger data set, with the cited work being a prototype.

1.2.4 Intelligent systems for product design

Having outlined the three domains, it is time to consider the various ways they overlap, as shown in Figure 1.1. We start with the combination of product design and intelligent systems.

Bates and Wynn [7] stress the need for a good understanding of the underlying system when developing products or processes. They demonstrate how adaptive radial basis function networks[3] can be used to model complex engineering systems in a computationally efficient manner. This speed, combined with sufficient accuracy, allows the designer to explore options and tradeoffs in the design.

Bentley's book [11] includes many examples of design tools based on artificial evolution. For example, Funes and Pollack [50] describe a system which learns through evolution to design simple structures made from Lego. Examples include a bridge, a crane capable of supporting a load, and a table. With no prior information about forces, the system learnt how to produce novel designs within the given constraints. While the examples could be dismissed as being trivial, the principle of producing original designs, perhaps to be used as inspiration for a human designer, is attractive.

O'Reilly and Testa describe several "smart tools" to aid architectural designers [95]. Their work uses genetic programming, agent-based systems, and artificial life. These are applied to surface design; to assigning functions to different parts of a building; and to producing novel town-planning models. The aim behind all the work was to provide tools to aid a designer, rather than to automatically design everything from scratch.

Even the best designed and produced commodity will not sell if customers choose not to buy it. It is therefore vital to consider the marketplace throughout the design and production process. Greene and Smith [56] describe a system which models how and why consumers make purchasing decisions, by using genetic algorithms to derive decision rules. This works by evolving a set of rules describing the decision processes of individual consumers, or groups of similar consumers. These rules will initially be

[3] Radial basis function (RBF) networks are described in Section 2.2.2.

random, and therefore very poor, but the genetic algorithm will gradually improve their quality, through natural selection. The system takes into account the individual consumer's historic purchase decisions as well as the current situation. Greene and Smith give the example of deciding whether or not to rent a particular residential property. The same modelling process could be used to model food purchasing decisions, producing rules such as:

If (Price < £1.50) AND (Icing Colour is White) THEN (Purchase Cake)

The essential benefit that intelligent systems bring to product design is the possibility of developing accurate predictive models rapidly, allowing the designers to explore effectively a range of new product designs. This is particularly important when designing consumer goods such as food, when the characteristics of an optimum[4] product are often far from obvious. The emphasis is on combining the creativity of (human) designers with the speed and power of (computer) models.

1.2.5 Designing food products

Having discussed product design in general, we must ask how designing food is different from designing any other product. At one level, there is no distinction: we specify the optimum product through experimentation, and then design a manufacturing process. But what is an "optimum" food product? At the lowest level, let us consider the optimum apple for a Mr John Doe. Suppose we find that he prefers a particular shade of green, with a certain degree of crunchiness, a hint of bitterness and a touch of sweetness. But of course, Mrs Doe might prefer a redder apple with more sweetness. In principle, we could continue until we had designed six billion apples, one for every man, woman and child — but this seems unreasonable. Instead, we might try to group people according to their preferences, and then produce one optimal apple for each group.

But how do we optimise preference? Consider the problem of designing an aeroplane wing: we want to maximise lift while minimising the cost, the drag, the weight and so on. This produces a complex multi-objective optimisation problem, but all the relevant factors can be measured relatively easily. With food, we want to maximise preference (while minimising cost etc.), but measuring preferences reliably is difficult (see Section 1.4). The fact that there is no clearly defined, easy-to-measure goal makes a direct application of such function optimisation impossible.

Furthermore, there is a more complex interaction between ingredients and food products, compared to non-food components and products. For example, if a (non-food) product designer decides that a larger widget is required, then a larger widget

[4]Where "optimum" could be with respect to an individual's preferences, a groups preferences, health benefits etc.

can, in all likelihood, be built. But if a food designer decides that sweeter tomatoes are required, how is this to be achieved? Can sweeter tomatoes be bred reliably? Would they still taste satisfactory? Or is it acceptable just to add sugar to ordinary tomatoes?

Moreover, personal differences go beyond sensory attributes of foods. For example, the health benefits of an apple may be important to one consumer but irrelevant to another. Consumers may choose products based on nutritional properties, hygiene or toxicological status, functional properties, familiarity, brand or product image, as well as cost (in terms of money or time or effort). Of course, all of these extra factors are based on the consumers *perception* of nutrition, function etc., which may be different from the true values.

Customers' decisions to buy products are based largely on their personal preferences[5]. It seems logical then, to analyse these preferences and manufacture foods that match them more closely. Van Gennert et al. [131] describe a generic three-stage plan to develop more appealing, and hence more profitable, food:

1. discover which food attributes affect consumer preferences and to what degree;

2. find the chemical and physical properties of the product related to these attributes;

3. adjust these properties to develop more appealing products.

Van Gennert et al. report only on the first of these three stages, leaving the others for future work. However, it is this first stage that we are most interested in this work. In common with many food design experiments, two complementary panels were used. The first was a trained sensory panel, who were given various tomatoes, and asked to define and use a series of sensory attributes (such as colour, flavour, texture, etc.). The second was an untrained preference panel, who simply express a preference for each tomato. These panels are described in more detail in Section 1.4. Both groups tend to suffer from "taste-fatigue" if presented with too many products, so such experiments are very limited.

The results were then analysed using principal component analysis (PCA) to rank the sensory attributes in order of importance when predicting preferences. This stage of the work is relatively straightforward. Although training a descriptive sensory panel can be expensive, it is a widely used process, as is analysing panel results (e.g. Stone and Sidel [121]).

The second and third stages of food design present more of a problem, which is largely unsolved to date. One attempt at tackling stage two (matching physical and chemical properties to product attributes) is outlined by Schonkopf et al. [114], who

[5] Or on the perceived personal preferences of the end consumer, if the customer is making the purchase for someone else.

analysed the design of cheese and other dairy products. They used analysis of variance (ANOVA) to design a series of experiments in which several factors in the production of cheese were varied. Then both ANOVA and PCA were used to analyse these results, with the aim of discovering which factors had the maximum influence on various key attributes of the food, such as taste, texture, juiciness and graininess.

Recent advances in genomics open a fourth possible stage to food design: genetic modification of crops (or animals) to produce ingredients with the desired physical and chemical properties. This is obviously at a very early stage: most GM trials concern disease resistance or greater resilience to drought, but in principle, and given public acceptance, one could conceive of plants being modified to make their fruit or leaves taste nicer. Perhaps the ultimate model of food preferences would be a mapping from food DNA profiles to food sales levels. As a further fantasy, one could sample consumers' own DNA, use this to predict their preferences[6], and build a model predicting ideal food DNA profiles from human DNA profiles. Needless to say, these issues are beyond the scope of this current work.

Currently, one widely used approach to food design is "preference mapping"[7]. It is used to compare a range of products to identify the most important attributes, and to identify gaps or overlaps in the existing market. By including new product formulations, preference mapping can also help optimise product design.

McEwan [78] describes two versions of preference mapping: internal and external. Internal preference mapping uses PCA to project consumer preferences onto a low dimensional space, to aid interpretation. Each consumer is then represented as a vector in this space, and these vectors can then be clustered to identify market segments. It is possible then to place the products into the same space, by correlating the PCA dimensions with the sensory attributes, but the usefulness of this is open to debate [27, 78]. The problem is that trained sensory panellists may perceive products differently to untrained preference panellists. Further, the optimum projection found by PCA is optimal with respect to the preferences, and there is no theoretical guarantee that the same projection is ideal for the products.

External preference mapping uses both preference and sensory data to build a linear model which predicts preferences from sensory attribute values. First, PCA is used to project the sensory data into a low dimensional space. Then regression analysis is used to predict preferences from the product attributes in this low dimensional space. This produces an explanation of *why* each consumer prefers the products they do, in terms of the sensory features identified by the sensory panel. We discuss preference mapping in more detail in Section 3.4.1.

[6] insofar as these are genetically influenced

[7] also known as perceptual mapping, brand mapping or product mapping

1.2.6 Intelligent systems in the food industry

Briefly moving away from design issues, we turn to the rest of the food industry, and consider how it has exploited recent advances in intelligent systems.

During the fermentation of beer, numerous chemical compounds are formed and broken down, their levels varying over time. Stopping the process too early or too late wastes time and money, as the beer must be discarded. Traditional approaches to this type of monitoring and control problem are discussed by Bimbenet and Trystram [14]. They include time-based process control (where each stage is of fixed duration) and off-line monitoring, where samples from the ongoing process are repeatedly removed and analysed, to provide feedback to determine when a process is complete. A third option is on-line monitoring, where sensors are placed within the production line, and must therefore be sterile.

Gardner et al. [51] describe an on-line system which monitors the concentration of certain chemicals found in beer during fermentation. An array of chemoresistive sensors was used, each designed to produce a distinct response to a range of chemicals. The outputs from the sensor array were then fed into a neural network. The array was placed in a beer-fermenting tank, and a variety of gases were passed over it. The neural network was then trained to detect trace quantities of diacetyl (a key indicator of fermentation) in the presence of other chemicals, such as ethanol. After training on a number of samples, the system was tested with some novel samples. Overall, the new system was found to significantly outperform the traditional chemometric fingerprinting. This allows fine control over the fermentation process, minimising wastage.

Although a broken biscuit may taste the same as a whole one, customers are reluctant to buy such damaged goods. It is therefore in the interest of the manufacturer to ensure that the goods leave the plant at as high a standard as possible, even at the cosmetic level. Using people to visually inspect large numbers of items on a production line is very expensive as well as unreliable, due to finite attention spans and limited visual acuity. Non-visual inspection, such feeling the edge of the product, may damage delicate foodstuffs, as well as introducing bacteria. A non-intrusive camera is, of course, completely aseptic.

Gunasekaran and Ding [59] describe an automated product inspection system, based on a camera connected to a neural network. The system was initially trained to distinguish between whole and broken crackers, and performed very well. However, this task is relatively easy, because all the crackers on the production line were (supposed to be) the same size and shape, so that a simple template-matching operation was all that was required. A more challenging task was also described, which involved detecting damaged almonds, again on a continuous production line. Because even undamaged almonds vary considerably in size and shape, the new neural network had to perform a more sophisticated task. Nevertheless, the system still performed well, demonstrating the flexibility of these models. Such systems can detect features invisible to the human eye, and although these features may not be apparent to customers, they may still affect the taste of the product, or its shelf life. Similar systems are used widely to detect flaws in glass, wood veneer, and silicon wafer production (e.g. [107]).

Many foods are heat-treated to kill bacteria, but the varying attributes of the food make controlling the temperature of sterilisation units very difficult. Singh and Ou-Yang [118] describe a process plant where a simple PID (proportional-integrative-derivative) controller had been used to control the sterilisation unit. They developed a fuzzy logic controller that out-performed the PID controller. The fuzzy rules used were derived from the (human) operator's intuition, with membership functions derived by trial and error. A typical fuzzy rule used by the system was: "If the holding tube inlet temperature is slightly above the set point temperature and the holding tube inlet temperature is rising slowly then slightly close the steam valve." Fuzzy logic (unlike the inherently linear PIDs) can deal successfully with complex, non-linear systems. The fuzzy controller consisted of a relatively small number of simple rules, and this simplicity led to faster response times. Zhang and Litchfield [137] provide further discussion of fuzzy control in food production.

A key aim of forecasting food sales by supermarkets is to predict trends in customers' purchases, which then guides stock control, marketing decisions, staffing levels, etc. Thiesing and Vornberger [127] describe a study to analyse and predict weekly sales of a range of products sold in a German supermarket. They used a neural network, with inputs such as the previous few weeks' sales levels, national holidays during the week, product promotions and price changes. The predictions made by the network were better than several alternative models, such as moving averages or a static model. One advantage of neural networks over conventional statistics is their ability to cope with chaotic time series predictions. Conventional time-series models, such as ARMA and ARIMA[8], are effective with non-chaotic series, but fail on more complex, non-stationary series.

1.2.7 Intelligent systems for food design

Finally, we reach the centre of Figure 1.1 (and of this thesis): using intelligent systems to aid the design of food. Specific tools can be developed which tackle particular problems within this area: modelling with small data sets; performing both regression and segmentation; selecting important features; detecting outliers; and optimising product designs. These are all examples of intelligent data analysis, discussed throughout this work.

[8] Autoregressive (integrated) moving average

One example of this is using Bayesian belief networks to model the relationships between sensory and preference values as described in our paper [29]. Bayesian models have the useful feature that they can easily be inverted: having learnt how to predict preferences from sensory scores, we can then use the model to predict sensory scores for a hypothetically perfect preference score. In the rest of this thesis, outlined at the end of this chapter, we analyse further approaches to intelligent food design.

1.3 Data analysis

Having placed intelligent food design in context, it is time to consider the issues of data gathering, and first, of data analysis. Hand defines data analysis as the extraction of information from data in order to answer a given question [61]. Such data analysis includes attempts either to summarise data or to make predictions from data, and to do so with or without an underlying theory. In this thesis, we concentrate on building predictive models without a pre-specified underlying theory. This allows us to ask questions such as "How can we improve product X to please consumer Y?", without first having to produce a complete theory of food preferences, in terms of psychology, physiology, biochemistry, and so on. Furthermore, we aim to build models that are easy to interpret, to aid our understanding of the processes that underlie food preferences.

Food manufacturers are principally concerned with maximising their profits. It is assumed that people buy the food that they (or their family) prefer, although other factors such as price do have a large influence. This raises the question: which foods do people prefer, and why? A first attempt at answering this question might be to build a model to predict sales levels from food properties, both intrinsic (e.g. ingredients, chemical properties etc.) and extrinsic (e.g. packaging, price and advertising). However, acquiring detailed sales records is often impossible for food manufacturers. When food retailers such as supermarkets also produce their own brands, they become competitors to non-retail manufacturers, such as Unilever, Nestlé and Proctor and Gamble, and so may be unwilling to share such information. If we assume that preferences and sales levels are correlated, then a second attempt at answering the question is to predict preferences from food properties. However, preferences can only depend on properties that are discernible to the consumers, so we can, and indeed must, ignore other properties, such as ultraviolet absorbency or chemical composition. These may be correlated with preference, but cannot directly explain it, and our goal is both modelling and explanation. Thus our final answer to the question "which foods do people prefer and why?" is to build a model to predict consumer preferences from sensory properties of foods. We must therefore measure both of these.

There are many modelling techniques available, both "old statistical" methods (linear regression, PCA etc.) and "new machine learning" methods (support vector machines, neural networks etc.), with more techniques being published all the time. There can be a tendency to treat statistics and machine learning[9] as opposing camps, using different methods to achieve different goals. This work focuses on data analysis. If one were to be cynical, one could say that statistics is data analysis taught in statistics departments, while machine learning is data analysis taught in computer science departments. The difference is one of emphasis (theoretical, rigorous and conservative, vs. experimental, heuristic and often ad hoc) rather than of goals, and it has been suggested that statistics and artificial intelligence either form [13], or should form [46], a complementary combination. Instead of ineffectually trying to prove that the two approaches are distinct, and that one approach is better than the other, we use tools from both fields in this work, to better understand the process of modelling very small data sets.

1.4 Data gathering

Most of the data being used in this work comes from two sources: sensory panels and preference panels. Stone and Sidel's book [121] covers many aspects of preference and sensory evaluation, some of which are now discussed.

1.4.1 Sensory panels

Many years ago, when there were limited product ranges, limited competition, and limited technology, food companies relied on individual experts to describe the sensory attributes of their products. More recently, the emphasis has shifted towards using sensory panels. These consist of ordinary consumers, who are then trained in sensory analysis for one particular product. A sensory panel is a group of typically 10–20 people, who are initially selected based on discriminatory ability. Typically, 30% of initial candidates are found to have very poor discrimination between the products under consideration.

The panel derives their own descriptors of product attributes, such as "nutty odour" or "chewy texture", which can then be systematically used to describe different varieties of the product. The panel typically develops between 30 and 50 descriptors after discussion and analysis [121, p.209]. In some studies, panellists have been forbidden from using some range of attributes (e.g. odour). However, subsequent analysis then showed that the same information tended to appear in other attributes derived by the panel. In most cases therefore, there are no restrictions placed on the nature or number of descriptors used, allowing the panellists to use whichever attributes seem

[9] and pattern recognition, artificial intelligence, and so on.

most appropriate to them. Morrot, Brochet and Dubourdieu [87] recently showed that the perceived odour of wine depended to a large extent on the perceived colour: white wine that had been dyed red smelled the same as red wine, according to most subjects. While this highlights the difficulties of sensory analysis, it also suggests that the panel must be free to choose its own descriptors.

Having derived these product descriptors, members of the sensory panel are then presented with a variety of different products, selected to represent a wide range of flavours, colours, etc. They then measure each sample by scoring it for each attribute, again with discussion between the panel members. The ideal sensory panel should produce absolutely consistent and uniform results, allowing the panel to be treated as an instrument. The advantage of using a "human instrument" is clear: the target customers are (of course) human, so characterising food based on attributes that are discernible by humans and regarded as important by humans, is more direct than using non-human instruments. However:

> ...humans are poor judges of absolutes but very good judges of relative differences [121, p.219].

We are attempting to find a relationship between sensory and preference scores, and so we need absolute scores for each product, independent of other products that we may or may not have measured. For our purposes therefore, humans are unreliable instruments.

This emphasises the importance of carrying out multi-product trials: the sensory panel can compare products, rather than attempting to describe one product in isolation. Using multiple products with absolute scores also provides clearer information about the product space. The products a panel are presented with typically include the manufacturer's own brand(s), the main competitors' brand(s), and a number of experimental products that have been designed to cover the space of products. Subsequent analysis should then reveal where in this space the ideal product lies, where "ideal" refers to a product that (at least some) consumers are most likely to buy.

1.4.2 Preference panels

There are two approaches to measuring people's preferences: direct and indirect measurement. Direct measurement relies on asking consumers to express their preferences explicitly. This is often done using preference panels, explained below, or via focus groups, home tests and so on. Indirect methods measure consumer behaviour, such as facial expressions while eating, speed of consumption, or detailed retail sales records. Indirect methods are generally regarded as being more valid, because they do not rely on introspection, but the information is harder to get[10]. The data used in this work comes from direct methods.

The preference panel is a large group of *untrained* people, typically 100-500 potential consumers, who are brought in "off the street" specifically for a trial. They are individually presented with a few samples and are then asked to score each one on a simple preference scale. The most common scale used is the "9-point hedonic scale", which ranges from one (dislike extremely) to nine (like extremely) with the middle at five (neither like nor dislike). Unlike the sensory panel, no training is given and no discussion between panellists is allowed, so the results will be entirely subjective and vary from panellist to panellist. The relatively large panel size should accurately reflect the diverse tastes of consumers, allowing market segments to be identified while smoothing out individual discrepancies.

The products presented to the preference panel are typically those presented to the sensory panel, or a subset of them. This allows sensory and preference scores of the same products to be measured. In typical studies, each preference panellist considers the products, four or five at a sitting, over a total of two days. Any more testing might lead to taste-fatigue and so limits the total number of products in studies. The first sample given is a dummy — preference panellists tend to give disproportionately high scores to first sample tasted. One sample is repeated (unbeknownst to the panellists) to allow consistency checks within each panellist's scoring. This can be used in later analysis to (for example) remove preference panellists who gave widely different scores to the same product on different occasions.

1.4.3 Other issues

There are some known limitations with these direct panel approaches. Typically, the panels meet in controlled sensory laboratories, so their behaviour is likely to be different from the table, the kitchen or the supermarket. In the supermarket, packaging and pricing are likely to be important influences; in the kitchen, preparation time and pre-cooked appearance may be important; and at the table, taste and post-cooked appearance may be important. It is only the last of these that is typically measured in preference panels, and even this is still not "natural". For example, when eating at home, the product in question is likely to be eaten with other foods, possibly with added sauces, salt and pepper, and maybe over- or under-cooked. All of these will greatly affect the eating experience.

Furthermore, preference panellists are forced to express opinions for every product, even when they have no clear preferences to express. A median preference response (e.g. "neither like nor dislike") is not the same as having no preference ("don't care").

[10] As noted earlier, detailed sales records are often impossible for food manufacturers to obtain.

Forcing preference panellists to express an opinion may therefore lead to a bias towards the median.

Three main factors influence most choices related to food products:

1. the product;
2. the consumer; and
3. the situation.

When designing experiments, the investigators must always balance the ease of measuring preferences and sensory scores against the invalidity of an artificial situation.

Once both sensory and preference data has been obtained, it is analysed to determine which sensory attributes best distinguish the different preferences. For example, suppose the preference panel gave two samples significantly different scores. If the sensory panel gave both of them the same grade for some attribute, e.g. texture, then this attribute is a poor predictor of quality as it fails to discriminate between products with different preference scores. A further complication is that the preference panellists are individuals with their own tastes. Thus a sensory attribute may distinguish between one subset of panellists' preferences but not another. Therefore, in Section 3.9 we select attributes for different subsets of preference panellists independently. If a correlation can be found between one or more of the sensory panel attributes and the (subset of) preference panel scores, then this can be used to guide future product design and marketing.

Besides the two panels, a third source of data may sometimes be acquired: instrumental data. The nature of instrumental data is product-specific, but may include digital images, acoustic imaging or chemical fingerprinting. Instrumental data is typically cheaper, but by definition, it is further removed from human tastes than the two panels. This means that the data will be harder to use when predicting preferences. For example, suppose that ultrasonic resonance is perfectly correlated with preference. Given that people cannot detect ultrasonic resonance, this cannot *explain* the preferences: the correlation is most likely to be coincidental, or related to some other sensory feature.

Moreover, it is likely that preferences are influence by a large number of features that may be weakly correlated, such as "minty odour" and "minty appearance". To measure such a wide range of features using instruments may well prove to be considerably harder than using a single device that can measure them all at once: a human.

The entire data-gathering process is very expensive and very time-consuming, and depends on human perception, which lead to the most striking and important features of the data sets: they are small, sparse and noisy.

1.4.4 Description of data sets

Unilever Research[11] has provided three sets of data for this work. For reasons of commercial confidentiality, the exact nature of the foods cannot be revealed, but a brief description of each data set is given below. In common with Unilever's global strategy, all three product ranges are mass produced, and aimed at a mass market. The "meat" and "beverage" data sets were used to investigate and develop the various techniques used; the resultant methods were then applied to the "vegetable" set as a blind trial. The preference data sets used here have been pre-processed by Unilever; the raw scores were adjusted according to the results from a repeated presentation of a single product, which can help correct "drift" in expressed preferences during an experiment lasting two days.

Premium meat product

The sensory panel was presented with 42 products, and derived 65 sensory features, covering appearance, flavour, aroma and texture, both before and during eating. The preference panel of 240 consumers was then presented with 16 of these products.

S_m is the meat sensory data, a 42×65 array.

P_m is the meat preference data, a 16×240 array.

The sensory data S_m can be broken into two parts: first, a labelled set of 16×65 values (with known preferences), and second, an unlabelled set of 26×65 values (with unknown preferences). This combination of labelled and unlabelled data allows us to use semi-supervised learning methods, which we discuss further in Chapter 2.

Processed vegetable sauce

The sensory panel was presented with 17 products, and derived 35 sensory features covering appearance, aroma and flavour. All 17 products were then presented to a preference panel of 211 consumers. One consumer failed to complete the task, and measured only 16 of the products. Rather than try to fill in this piece of missing data, that consumer's entire record was removed, leaving 210 consumers.

S_v is the vegetable sensory data, a 17×35 array.

P_v is the vegetable preference data, a 17×210 array.

Hot beverage

The sensory panel was presented with 20 products, and derived 8 sensory features, covering the appearance of the ingredients of a hot drink. The preference panel of 450 consumers was then presented with all of these products.

[11] The research division of Unilever plc

S_b is the beverage sensory data, a 20×8 array.

P_b is the beverage preference data, a 20×450 array.

Other data

In some experiments, further data sets from the UCI data repository [16] or artificial data sets are used. Hand et al. [62] warn that creating artificial data can be misleading, because it is hard to create realistic distributions. However, during exploratory data analysis, it is often instructive to create data which is "easy" to analyse (e.g. low noise, low dimensionality) for proof-of-concept experiments. By then increasing the realism of the data (adding noise, increasing the dimensionality) we can move towards the real data. While accepting that it will always be a poor approximation, this shift towards reality may give insight into how the algorithm under investigation will perform on real data. The aim is not to create a copy of the real data, but to provide total control over the number of records and features. These are key issues in much of the current work: how do algorithms perform given few records? Or many features? Or noisy data?

1.5 Thesis outline

This thesis contains three technical chapters on data analysis techniques, all motivated by the analysis of small data sets. In line with the aims of this work, each chapter compares statistical techniques with machine learning techniques, often using computationally intensive methods. Particular emphasis is given to model complexity, because the smallness of the data sets makes it easy to produce models with greater complexity than the data actually support (although it is perfectly possible for this to occur with large data sets as well).

The process of modelling consumer preferences can be approached in many fashions. One possible sequence is:

1. Select a set of sensory features;
2. Divide consumers into groups with similar preferences;
3. Identify and remove consumers with inconsistent preferences;
4. Build regression models predicting preference scores from sensory scores.

However, this list is an oversimplification, as it suggests that these stages are independent, when they are not. For example, when selecting a set of sensory features for regression, the quality of the set depends on the regression model chosen, and vice versa. The optimum feature set for a linear regression model may be different to the optimum set for a non-linear model. Similarly, the result of grouping the consumers depends on which consumers we remove from the data set as being outliers. Therefore, the order in which the ideas are presented in the thesis (and the sections below) is just one path through the modelling process. Alternative paths are discussed in Chapter 5.

1.5.1 Feature selection and regression

All three data sets have very few records (i.e. ≤ 20 products), and two of them have many more features than records. To build predictive and interpretable models, we need to reduce the number of features. We can either select a subset of the available features, or select some combination of features to reduce the dimensionality. In food design, it is important to understand which features drive consumer preferences, so we argue that feature subset selection is essential.

In Chapter 2, we discuss several common statistical approaches for feature selection, such as forward sequential selection. Machine learning approaches discussed include stochastic searches such as simulated annealing. The emphasis in published work in the field is on detecting *relevant* features, all of which are then used in building predictive models. One problem with small data sets is the risk of selecting too many features, even if they are all apparently relevant. This can lead to building models that are too complex, and fail to capture the underlying relationship between sensory attributes and preference scores. This is especially true if the features are weakly correlated.

Chapter 2 also includes a discussion of regression methods, focussing on a comparison between linear and non-linear approaches. Given that we are selecting features in order to perform regression, and that the performance of a regression model depends on the features selected, feature selection and model building are inseparable.

The principal question at the core of this chapter is:

> How can we avoid building models that are too complex, when we only have a very small data set?

Two solutions are suggested that depend on the type of data available. In general, a stochastic search combined with suitable complexity penalty is shown to be effective. If extra unlabelled[12] data is available, then a semi-supervised approach may be more suitable. We propose and analyse two new algorithms using semi-supervised learning, one for feature selection, and one for creating regression ensembles.

In some cases, and given a sufficiently good set of features, non-linear regression models are shown to outperform linear models, although at considerably greater computational expense. In other cases, linear models are more accurate than non-linear models, according to leave one out cross validation.

[12] Unlabelled records correspond to products with known sensory attribute scores, but unknown preference scores.

Note that there are two forms of complexity under investigation. Firstly, introducing more raw features increases the model complexity; and secondly, allowing more parameters per feature (such as moving from a quadratic to a cubic polynomial) also increases the model complexity. The latter introduces more features but these are functionally related to the raw, measured features from which they are derived.

1.5.2 Cluster analysis

Given that consumers have distinct preferences, we cannot select just one set of features and build just one regression model. Instead, we must identify market segments and target appropriate products to appropriate consumers. We can then build a separate regression model for each cluster of preference panellists. In Chapter 3, we discuss several clustering methods, focussing on Gaussian mixture models and k-means clustering. An approach specific to product design — preference mapping — is also described, and a non-linear variant is introduced.

A universal problem in cluster analysis is to determine how many clusters are present in the data. Several traditional approaches to this are discussed, such as those derived from information theory and a novel consistency-based approach.

Two research questions are raised. Firstly we repeat the question from Chapter 2:

> How can we avoid building models that are too complex, when we only have a very small data set?

In this context, "complex" refers to the number of clusters, which determines the number of free parameters that we must estimate. Secondly, we observe that stochastic clustering methods produce different models each time, leading to the question:

> How can we choose between or combine different cluster models, generated from the same data?

The first question is answered using two approaches, namely those derived from information theory and a measure of solution consistency. These give similar results, although it is impossible to determine the true number of clusters present. "True" in this context refers to how many distinct groups of consumers exist in the population, but this is unknown and immeasurable. Moreover, it could be argued that any such grouping of people is artificial and therefore there is no "true" number of clusters. Instead, we must aim to find clusters that allow us to make useful predictions of preferences and to predict the corresponding sensory properties.

In answer to our second question, we describe new algorithms to distinguish between cluster models. One approach measures the consistency with which a particular solution is found, and chooses the cluster model with greatest consistency. Another approach combines clustering with regression, and uses the regression error to select a clustering solution. These provide new ways to choose between two models that are distinct, but as good as each other purely in terms of cluster analysis.

1.5.3 Outlier detection

It is known that many preference panellists produce inconsistent and inaccurate results. Some people are not skilled at (or are incapable of) distinguishing between widely differing products, so will find it impossible to give well-founded preference scores. Furthermore, we have no knowledge of the distribution of these scores, and so we have no noise model. Therefore, we need to remove or accommodate these scores, or else the models we build will be distorted and misleading. In Chapter 4, we discuss several current approaches, including discordancy tests, a support vector approach, classifier stability, and layered removal. Many of these require the use of parameters selected by the analyst.

Two questions are raised:

> Can we use the fact that we are performing clustering to aid the detection of outliers?

and

> Can we reduce the number of arbitrary parameters to be selected, such as decision thresholds?

We present a novel outlier detection approach, which answers both questions by combining cluster analysis with outlier detection, without the need for any extra arbitrary parameters. We highlight the flexibility of this approach by using it to reject outliers from a combination of clustering and regression models.

1.6 Conclusions

Throughout this thesis, we argue that the analysis of small data sets leads to different issues than the analysis of the more common, larger data sets. Overfitting is always a risk, but becomes exaggerated by the lack of data. Effective feature selection is always difficult, but becomes almost impossible when only a few records are available. We address the issues by combining computationally intensive machine learning methods with more conservative statistical approaches, and we propose several novel algorithms to solve these unusual problems.

This chapter has highlighted current trends in the food industry, and how intelligent systems can be used to aid the designers. One common feature is that technology is often a driving force, whether it is domestic freezers and microwaves or computerised stock control systems. The application of intelligent systems to the design of food is one further step down the same road.