Emotive Segmentation Methodology

MARKETING SERVICES | VisualDNA™

Introduction

Emotive personality traits are at the core of understanding online audiences, their feelings, motivations and values. Those traits are harder to identify than preferences and demographics, therefore they are more difficult to target effectively.

When advertisers initiate campaigns, their first task is often to develop sophisticate daudience profiles. These include features such as demographic segments, affluence and interests, but also attitudes towards well-being, family and emotional state, or in general, emotive personality traits.

Traditional data providers derive users' preferences from their online activity and demographic data that is available to them, but would not normally have access to emotive personality data.

This is the problem that VisualDNA set out to solve with *emotive segmentation.*

In this white paper an alternative method of emotive segmentation is presented. The method uses engaging personality quizzes based on advanced psychological profiling theory to identify both self-declared preferences and latent personality traits.

VisualDNA's emotive segmentation methodology is hereby described in detail, including the production of targetable emotive data tags and the key roles of quiz design and factor analysis in their creation. The mathematical foundations of the scoring algorithm are included in the appendix.

Quiz design

Visual quiz design is a complex process which aims to marry a valuable proposition for the quiz taker with reliable profiling for advertisers and publishers. While the immediate gain for the quiz taker is a thorough personality report, the longer-term gain is a personalised web experience. Or to put it more simply, being understood.

Verticals

The first step of quiz design is defining its verticals. Each vertical is designed to facilitate understanding of a distinct personality trait or life skill, and normally composed of three to seven questions. These approaches are demonstrated below using two of our personality quizzes: *Who Am I?*, and *VisualDNA PersonalityTest* (which can be viewed at whoami.visualdna.com and personality.visualdna.com, accordingly).

Who Am I?

The *Who Am I?* quiz design draws on the five factor theory of personality (Costa & McCrae, 1992 [2]) to measure openness to new experiences, conscientiousness, extraversion, agreeableness and neuroticism. This model is widely accepted as the most comprehensive in explaining individual differences in personality, and so provides a solid foundation both for capturing data and providing an engaging user experience. Users receive a comprehensive report detailing their percentage scores across each of the five factors, and personalised feedback on how their personality type might affect things such as their taste, self-control, sociability, levels of composure and outlook on life. Visual questions were based on scales from the International Personality Item Pool (IPIP; Goldberg, 1999 [8]). As in the IPIP, we use multiple questions and also reverse questions (identifying inverse expression of the same characteristic) in order to more precisely identify traits. Six questions were designed for the measurement of each personality factor, using the granular facets they are constructed from.

VisualDNA Personality Test

The *VisualDNA Personality Test* aims to extend beyond psychological profiling and explore the way these manifest in the users' life skills. It is designed to give users an insight into various aspects of their life, from time management to financial behaviour, and provides advice on how to effect change. As such, it explores seven verticals: general interests, aspirations and dreams, dealing with stress, love, finances, resourcefulness and state of mind. These were chosen to reflect our users' preferences, as well as those of advertisers and publishers.

Personality verticals are based on psychology profiling science, while life skills verticals are based on our considerable experience in user profiling. For example, the 'resourcefulness' vertical measures users' potential for achieving tasks, and applies Csikszentmihalyi's model for the experience of 'Flow' state (Csik szentmihalyi, 1997 [3]). As simple Likert scales are capable of identifying the potential for flow experiences at work (Eisenberger, Jones & et al. , 2005 [6]), questions were designed to represent ordinal scales as images. Similarly, for the 'dealing with stress' vertical, we also draw on the five factor model of personality (AKA Big Five; Costa & McCrae, 1992 [2]). However, in this case we focus on the two personality characteristics that are associated with positive emotion and actions (extraversion and openness to new experiences).

Question design

Visual questions are fun, simple, fast and intuitive to answer. Serious and emotionally difficult questions can be asked using humour and euphemisms, without compromising the clarity of the option meaning. However, some concepts are hard to express in images, either because they are too abstract or, alternatively, too specific. In these cases we use text-based questions. A comprehensive approach is key, both to designing questions that cover a variety of life stages and to providing answers to those questions, while also being as economical as possible with the number of options (which never exceeds 15). Manual tagging is used in addition to the statistical methods described below. Weights are assigned to different options to ensure that multiple similar answers must be given before any characteristic is assumed.

The art and science of image research

A number of steps are taken to minimise assumptions of understanding through careful image selection. Our image specialists maintain an even emotional balance across images in terms of colour balance, focus, zoom and level of abstraction, and multiple stages of in-house review ensure that concepts we aim to convey are reliably perceived.

Descriptive statistical analysis plays a crucial role in allowing us to ensure a high level of accuracy. The distribution of answers is regularly checked for anomalies, and questions are fine-tuned during testing phases in response to this. When a concept is being measured across multiple questions, the distributions of the resulting scores are assessed. Responses to conceptually complex questions are also compared with more straightforward text-based questions designed to measure similar traits.

Factor analysis plays a crucial role at this stage, allowing us to check that users are answering in a way that reflects the groupings that we have aimed to identify. This technique is described in detail in the next section. In addition to such analyses, we collect qualitative data from participants through user testing and surveys to capture feedback on the quiz experience.

Image tagging

Holistic statistical methods, such as factor analysis, are mainly used to extract deep, latent personality traits, which are not self-declarative. In parallel, factual self-declarative options are also tagged manually. For example: the question 'How do you travel to work in the morning?' can yield hard factual answers, as well as contribute to understanding the quiz taker's attitude to well-being and sport; the question 'How do you prefer to listen to music?' can indirectly help to unfold the quiz taker's attitude to morals (in addition to factual information).

Statistical Analysis

This section presents an emotive scoring algorithm developed by VisualDNA with the goal of accurately assigning emotive tags to users.

Prior to introducing this algorithm, emotive characteristics were identified by directly examining user answers. For example, when a user clicks the 'Going to concerts' image as a response to the question 'What would you prefer to do on a Saturday night?', we would assign to them a 'Music Lover' tag. This method is suitable for capturing self-declarative aspects, but struggles to measure deeper dimensions like openness, extraversion or agreeableness.

To measure these, we turn to the field of *psychometrics* - the study of psychological measurement.

Psychometrics consists of a set of approaches for scoring abilities, attitudes and personality traits. One of the main statistical tools is a method called *Factor analysis* [7].

Factor analysis is a technique that identifies the hidden or latent variables that generate observed data. In our case, this means that given a dataset of quiz answers, factor analysis can find the unobserved personality traits that lead users to choose particular observed answers.

Factor analysis

Factor analysis is utilised to extract latent personality information from quiz answers. It is a two step process: the first looks at the dataset as a whole and identifies its underlying latent structure; the second maps a user's quiz answers onto this latent structure.

The following is a high level description of our use of factor analysis. The technical reader is encouraged to consult the appendix where the technical details of this process are supplied.

First step: extracting factors
A typical VisualDNA quiz consists of over forty questions, with 2-15 answer options each. The first step of factor analysis is to extract the underlying factors which span a large dataset of quiz answers. Factor analysis looks for similarities in user answers. The

algorithm understands that the images for 'Concerts' and 'Guitar' represent similar concepts, because many users that picked one also picked the other. A factor is defined by a set of images with similar click patterns.

The following is an example of the top five images that make up the 'Music' factor; image size corresponds to factor loadings.

We derive factors for each of the quiz verticals (please refer to the quiz design section). The number of factors identified within each vertical depends on the variation of answers from user to user for questions in the vertical.

What would spark a conversation with you?
Right now, what matters most in y
What do you most enjoy doing in
How do you most like to be entertained?
What will you treat yourself to next?

Second step: scoring users by factors

With the factors identified, the algorithm assigns every user a score on each factor. Doing this turns the above binary representation of the quiz into a more complex representation in terms of scores and factors. So, for example, the binary representation becomes: 0

User id	Music lover	Contemplator	Extravert	: : :
123	0. 5	1. 78	-0. 87	: : :
456	2. 5	-0. 45	-1. 7	: : :
789	0. 2	1. 88	1. 87	: : :

As marketers are interested in targeting audiences with particular traits, e. g. 'Music Lovers', we then translate factor scores into binary tags that are assigned to users. Since factor analysis captures latent personality information, we call tags derived from factors *emotive tags*. Each factor produces two emotive tags - one for high positive scores on the factor and one for low negative scores on the factor. An emotive tag is assigned to a user if their factor score is above a particular threshold, as determined by our in-house algorithm.

Scoring example

Here are examples of actual user answers which score differently on the music factor. The following image answer combination generates the highest positive score on the music factor:

The following image answer combination generates a score of 2 on the music factor:

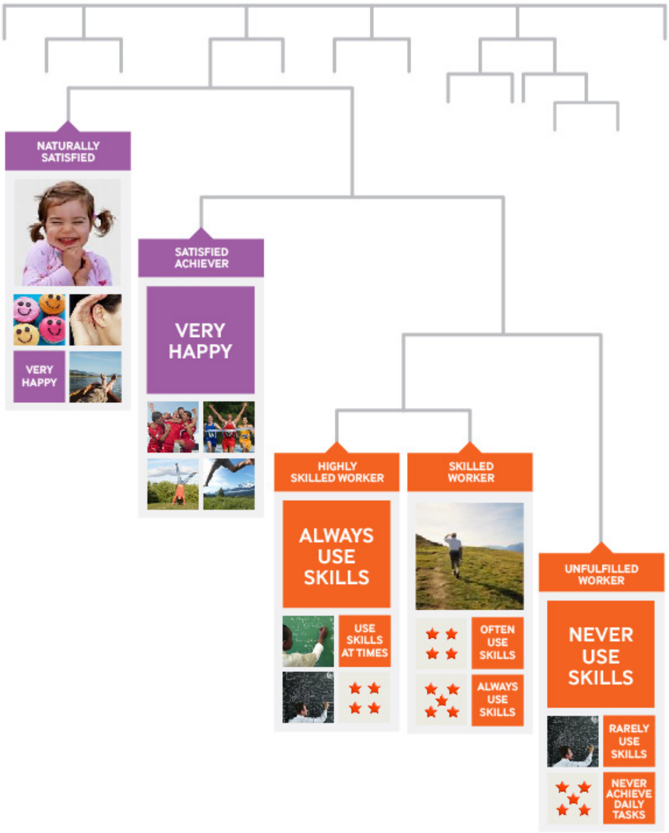
Finally, the following image answer combination generates a score of 1. 68 on the music factor:

All users that scored higher than 1. 68 on the music factor were assigned the 'Music Lover' tag. Above this threshold, the higher the score, the higher the intersection between the actual image choices and the music factor.

Clustering tags

Factor analysis produces a large and complex list of tags that are often related to each other. To make our tag cloud easy to navigate and to eliminate redundancy, we cluster tags together using a hierarchical clustering algorithm.

The following dendrogram illustrates how we can contextualise interpretation of the different emotive tags using cluster analysis. We can see that emotive tags related to levels of use of skills at work are grouped together. We can also see that the tag for using skills frequently is closest to the tag for achievement-based satisfaction, and satisfaction in general. This implies a connection between level of skill use at work and general satisfaction.



Inference

VisualDNA uses two main data sources: personality quiz profiles and our media partners' data. Our media partners are a network of websites that share with VisualDNA the anonymised browsing data of their users. VisualDNA thus has two types of users: ones that have completed a quiz - also called *measured users* - and ones that have not - *inferred users.*

As of May 2013 we have 0.5 million measured users and 140 million inferred users active on our network every month.

Since measured users have completed the quiz, they are assigned emotive tags. Using the internet browsing patterns that we have for both types of users we infer emotive tags for inferred users. VisualDNA uses a proprietary inference system that performs daily updates on our pool of users. Inference allows us to successfully assign emotive tags at scale.

Conclusion

Our goal is to facilitate real time online understanding of users by describing them with tags that reect their personality traits.

We achieve this goal with a process that starts with designing insightful quizzes, and collecting user profiles at scale. The next step is using factor analysis to extract latent personality dimensions, and transform quiz answers to factor score representation. All users are then assigned emotive tags, which in turn seed inference of emotive tags for users who have not answered the quiz.

Through a mixture of careful psychological analysis and advanced statistical techniques, VisualDNA is able to create very rich digital profiles of our users, capturing self-declared information along with deep, latent personality traits.

VisualDNA emotive segments are available for targeting through all major platforms, including DoubleClick, Turn, AppNExus and Videology.

To see VisualDNA emotive segmentation in practice, visit WHYANALYTICS (why.visualdna.com), our free web analytics tool that uses emotive segmentation to profile website audiences in real time.

Mathematical Appendix

In mathematical terms, suppose that we observe the variables $x = (x_1, x_2, \ldots, x_n)$. The factor analysis model assumes that these variables were generated by the variables $y = (y_1, y_2, \ldots, y_k)$ where $k \leq n$ and each $y_i \sim N(0, 1)$ is standard normal.

The x_i's are related to y_i's via the following assumption:

$$xi\,|y \sim N(\mu_i + \Lambda \cdot y, \sigma_i)$$ [1]

and the assumption of conditional independence:

$$p(x = a\,|y) = \prod_{i=1}^{n} p(x_i = a_i\,|y)$$

where $a = (a_1, \ldots, a_n) \in R^n$.

The $n \times k$ matrix Λ is called the *loadings matrix* and is a parameter of the model. Equation 1 formalises the idea that the observed data is generated by unobserved hidden phenomena. x is the observed data. Roughly speaking, it is generated by the unobserved y by first sampling y from a standard normal and then transforming it according to Λ. In this sense, Λ explains how the unobserved phenomena generates the observed data.

The above form of factor analysis cannot be applied directly to VisualDNA quizzes because our observed data is categorical (each question forms a categorical variable), breaking the assumption of equation 1. To apply the model, we represent user answers in a dummied table format (see [9]) and generate the tetrachoric correlation matrix (see [5]) for this table. To generate this matrix, we assume that each of our binary variables is actually generated by a continuous one as follows:

$$x_i = \begin{cases} 1 \text{ if } x^i_c > t_i\,, \\ 0 \text{ else,} \end{cases}$$

where x^i_c is a latent continuous variable and t_i is some threshold. The tetrachoric correlation matrix is the correlation matrix of the variables x^i_c.

Thus the input to our model is the dummied representation of user answers and its output is the loadings matrix Λ.

There are several methods to solve for Λ. Examples include several E-M algorithms [13] and a Bayesian approach coupled with a latent trait model [11]. We implemented an in-house method based on an iterative application of Principal Component Analysis [10].

From Λ to factors

Columns of Λ correspond to factors and its rows to quiz answers. At this point of the analysis, factors are mathematical abstractions. To transform them into meaningful personality traits, we examine the highest and lowest loadings in each column. We then study the answers identified by the rows of these loadings. These answers form a set of images that allow us to interpret factors. Each factor is assigned meaning by our team of psychologists.

There are two technical steps that our system performs before we begin manual factor identication. The first is a *rotation* of the loadings matrix. Briefly, many Λ's satisfy equation 1: if Λ is a solution to equation 1, so is any rotation of Λ (see [12] for more details). We want to pick a rotation that will make the columns of Λ maximally orthogonal. This matrix form makes factor interpretation possible.

With the factors rotated, we measure their quality. This is done by comparing Λ with a loadings matrix that results from performing factor analysis on random data. We only keep factors whose top and bottom loading are significantly different from random loadings. In practice, this means that we keep factors with enough loadings above certain critical values. The main ideas behind this approach are described in detail in [1] and [15].

Factor scores

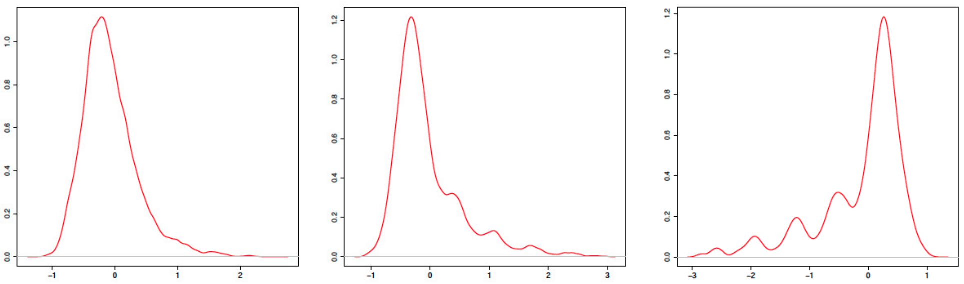
Once the factors are identified, we assign each quiz attempt a score for each factor. There are several methods for assigning factor scores [14]. We picked an approach that allows us to perform scoring in real time and on large volumes of data. Let $x = (x_1, x_2, \ldots, x_n)$ be a binary representation of a quiz attempt. Then the transformation into the latent space - the vector space spanned by the y_i's - is given by

$$y = (y_1, \ldots, y_k) = t(x_1, \ldots, x_n) \cdot \Lambda$$

where $t : R^{n \times 1} \to R^{1 \times n}$ is the transpose transformation.

On the following page are some examples of the density estimations for three factor scores:

Observe that while the theoretical model predicts a $N(0, 1)$ distribution for the y_i's,in practice the distributions of factor scores are far from normal and vary from factor to



factor.

This makes the problem of finding a threshold that determines when to assign an emotive tag to a user not trivial. VisualDNA has developed an effective in-house method for automatically determining these thresholds. We note in passing that the process of assigning each user a score, determining the cut offs and assigning emotive tags in essence clusters users into groups defined by emotive tags. In fact, using factor analysis for clustering purposes is a generalisation of the familiar k-means clustering algorithm [4].

Detailed summary of process

We finish with a summary of the whole process.

1. Collect quiz data

2. Create the tetrachoric correlation matrix from the dummied representation

3. Solve for the loadings matrix Λ

4. Rotate to make it easy to interpret

5. Identify high quality factors

6. Name the top and bottom score range of each factor

7. Transform the dummied answers into the latent variable space

8. Determine thresholds for the top and bottom ends of each factor

9. Assign emotive tags to all measured users

10. Clustering algorithm to find similarity between emotive tags

11. Infer emotive tags for inferred users

References

[1] Amy S Beavers, John W Lounsbury, Jennifer K Richards, Schuyler W Huck,Gary J Skolits, and Shelley L Esquivel. Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research & Evaluation*, 18(6):2.

[2] P.T. Costa and R.R. McCrae. Revised neo personality inventory and neo five-factor inventory professional manual. *Psychological Assessment Resources,*1992.

[3] M Csikszentmihalyi. *Finding flow: The psychology of engagement with every-day life*. Basic Books, 1997.

[4] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[5] DR Divgi. Calculation of the tetrachoric correlation coecient. *Psychometrika*, 44(2):169{172, 1979.15

[6] Robert Eisenberger, Jason R Jones, Florence Stinglhamber, Linda Shanock,and Amanda T Randall. Flow experiences at work: For high need achievers alone? *Journal of Organizational Behavior*, 26(7):755{775, 2005.

[7] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4:272{299, 1999.

[8] L. R. Goldberg. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *In Personality Psychology in Europe, Vol. 7. Tilburg University Press.*, 1999.

[9] Melissa A Hardy and Alan Bryman. *Handbook of data analysis*. SAGE Publications Limited, 2004.

[10] Ian T Jollie. *Principal component analysis*. Springer verlag, 2002.

[11] Martin Knott and David J Bartholomew. *Latent variable models and factor analysis - A Unified Approach*. Number 7. Edward Arnold, 2011.

[12] Michael B Richman. *Rotation of principal components. Journal of climatology,* 6(3):293{335, 1986.

[13] Donald B Rubin and Dorothy T Thayer. Em algorithms for ml factor analysis. Psychometrika, 47(1):69{76, 1982.

[14] Ledyard R Tucker. Relations of factor score estimates to their use. *Psychometrika,* 36(4):427{436, 1971.

[15] Brett Williams, Ted Brown, and Andrys Onsman. Exploratory factor analysis: A five-step guide for novices. *Journal of Emergency Primary Health Care,*8(3):1, 2012.

About VisualDNA

VisualDNA brings together psychology and big data to deliver new levels of customer understanding.

Our vision is to transform the way people are understood online by providing a new layer of data to the digital ecosystem. Our large team of data scientists works with our psychologists to harness the largest psychographic database in the world, plus 30 years of academic research into human understanding and online behaviour. We provide rapid and in-depth understanding of who an organisation's online customers are and how best to communicate with them.

VisualDNA technology is integrated with all major communication platforms from Display, Email, CRM and Search to onsite personalisation, which means online brands can create optimised campaigns and customer journeys with minimal implementation effort. Our global network reaches more than 300 million consumers, capturing insightful data on their online behaviour. Using powerful modelling techniques, our data scientists have transformed this data into more than 110 million deep profiles that combine demographic information with interests, purchasing intent and other psychographic information drawn from responses to visual quizzes. In the UK, we make 45 million profiles available through 15 major advertising technology platforms, helping businesses access the actionable insights needed to achieve optimal marketing results.

For more information,
please visit www. visualdna. com/marketing-services
or call +44 20 7734 7033

Follow us on Twitter @VisualDNA