Modeling Relative Importance of Text and Visuals in Multi-Modal Media for Human Cognition

Meia Alsup, Sravya Bhamidipati, Anelise Newman

[1]6.804 / 9.66 Massachusetts Institute of Technology (MIT)
Professor Joshua Tenenbaum

{malsup, sravyab, apnewman}@mit.edu

Contents

Abstract. *Many modern visualizations are multimodal–they are composed of both textual and visual elements that work together to convey the main message of the piece. In this work, we seek to extend Latent Dirichlet Allocation to multimodal media in order to better understand the relative importance that humans place on textual and graphic media when viewing infographics. In particular, we model how the importance of textual versus iconographic elements varies with viewing time. First, we formulate a model that extracts a topic distribution over both the textual and visual elements of a visualization and combines them with a mixing parameter alpha. The text-based distribution is derived using LDA, while the visually-based distribution combines category data for each icon in an infographic. Then, for 12 infographics, we collect human data and tune our model's parameters. We find that our model is best able to model human intuition when humans are exposed to the infographics for 1 second, and that our models diverge from human intuition when humans are exposed to the graphics for longer. These observations are likely explained by confounding variables and insufficient controls.*

1. Introduction

The classic use case of Latent Dirichlet Allocation, or LDA[1], is to identify the topics present in a corpus of textual documents. Given that most media today is composed of complex graphics that combine text, images, and more, we are interested in extending LDA to build a more relevant model for understanding how these components play together. To do this, we make use of a data set of infographics containing both visual and textual components. We categorize both the textual and visual components independently, and compare the combined results with humans' intuitions after they have been exposed to the infographic for a variable number of seconds.

Questions we consider when designing our model include:

1. How can we model visual media with an LDA-like framework?

2. How can we weight the relative importance of textual versus visual elements when doing topic classification?

3. Is there a temporal component to how people do infographic topic classification? Do people's topic predictions change when given more/less time to explore a graphic, and can we incorporate this into our model?

In order to gather the human data that we need to answer these questions, we ran a user study. We exposed humans to 12 infographics for time intervals of 1, 5, or 25 seconds and asked them to identify the topics present in each graphic. We hypothesize that when the viewing time is short, the visual content will be a better match for human classifications, whereas if the viewing time is long, the textual content will better match human determinations. The reasoning behind this is that text is less ambiguous than icons or other visual elements, but it takes longer to read text than it takes to catch a glance of an image.

The paper is organized as follows. Section 2 talks about related work. Section 3 introduces the data set that we use in our model-building and analysis. Section 4 discusses our methodology for selecting infographics for our user study, topic modeling of the textual components, topic modeling of the image components, and the user study itself. Section

5 shares the results of our study and section 6 takes a deeper look at the results to draw insights and conclusions. Section 7 outlines our contributions. Section 8 is an appendix with additional information about our procedure and data collection, as well as some techniques that we tried out that did not make it into our model.

2. Related Work

2.1. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a probabilistic generative model which focuses on topic modelling. It takes as input a corpus of text documents and n, the number of topics the model should output. It produces as output a set of n topics, each of which is a probability distribution over all the words in the corpus. Each document is modeled as a probability distribution over the topics. Several libraries exist for performing LDA effectively, such as the Python library gensim[1].

2.2. Understanding multimodal visualizations

Some prior work has focused on what elements of multimodal graphics attract people's attention and how these fixations can be modeled computationally. Eye tracking data collected from subjects looking at multimodal visualizations shows that people fixate most on the title (a textual element) and next on human-recognizable objects (visual elements) [2]. Furthermore, when asked to identify whether a visualization is familiar or not, people's fixation patterns are significantly different for memorable versus non-memorable visualizations. For non-memorable images, people tend to explore the image more and read the text instead of relying on visual data [2]. This suggests that humans' encoding and understanding of visualizations contains an interplay between text and visual elements that is time-dependent.

Although a lot of work in the Computer Vision space has focused on natural images, some work has focused on understanding and summarizing non-natural, multimodal images. [3] attempts to summarize infographics in a text-driven away by predicting topic labels based on the infographic text and then locating icons that match these topics. However, this approach is motivated by ease of engineering (predicting topics from text is arguably easier than from images), not because this is how the human brain works.

3. Data Set

In this project, we draw heavily from the augmented Visally29K dataset[2] [4]. This is a dataset of over 60k infographics scraped from the visualization hosting site Visual.ly, augmented with annotations and detections for both visual and textual elements. Here are the parts of the dataset that were particularly important for this project:

1. The main corpus of full-size infographics.
2. Text transcriptions for all infographics generated by Google's OCR API (used as input to LDA).
3. A subset of 1400 infographics with relatively clean human-annotated bounding boxes for all icons in the image.

[1] https://radimrehurek.com/gensim/
[2] http://visdata.mit.edu/

4. Noisy, model-generated bounding boxes for the icons of every infographic in the dastaset.
5. A dataset of 250k icons scraped from Google (not from particular infographics), along with category labels for each icon (used to fine-tune a CNN for icon recognition).

This dataset was ideal for our project because it provided a curated set of visualizations and extensive information about the visual and textual elements that they contain, which we could use as building blocks when creating our model.

4. Methodology

4.1. Topic modeling for text

We ran LDA on text extracted from the infographics in order to generate text-only topic predictions for them. We used the Python library gensim to perform the LDA. We also used a few pre-processing libraries available in the gensim package.

Pre-processing

Our original data set had over 60,000 infographics. During the first pass, we threw out infographics that were not in English. We did this by only including infographics where the text contained at least one word that was English with high probability and excluding infographics containing at least one word that was high probability Spanish. This preliminary language sweep narrowed our data set down to about 47,000 infographics.

From there, we further processed the documents by tokenizing the words. We removed the punctuation, numbers, and special characters and made all words lowercase. We removed all words under 3 characters, removed all stop words, and removed all words that were not in the English dictionary (this included several mis-spelled words). Our ground truth English dictionary used for filtering was also provided by the gensim package.

Training LDA

Now that we had the cleaned set of tokens pertaining to each infographic, we were able to put together a vocabulary for all the infographics. We threw out words that appeared in over 50% of documents since they would not add much information other than noise. We then also filtered down to the 10,000 most common words for our actual input into the LDA training model.

In training, we varied the number of categories produced by LDA from 10 to 50. We ultimately decided that the parameter 20 resulted in the most coherent and understandable categories.

Defining Topics

A summary of the 20 topics output by our text-based LDA model can be found in the appendix (section 8.1). For each topic, we display the top 20 most probable words associated with that topic, as well as a hand-crafted topic name. As a group, we examined the associated the word distributions and determined appropriate category labels for each distribution.

After examining the topics, we determined that two were noise and two others could be logically combined. Thus, the final topics we settled on are: health, economy and government, celebrations and family, internet and social media, international world, shopping and travel, online shopping, marketing, education, mobile phones and devices, food, cars and driving, culture and colors, money and finances, energy and environment, house and home, and business and technology.

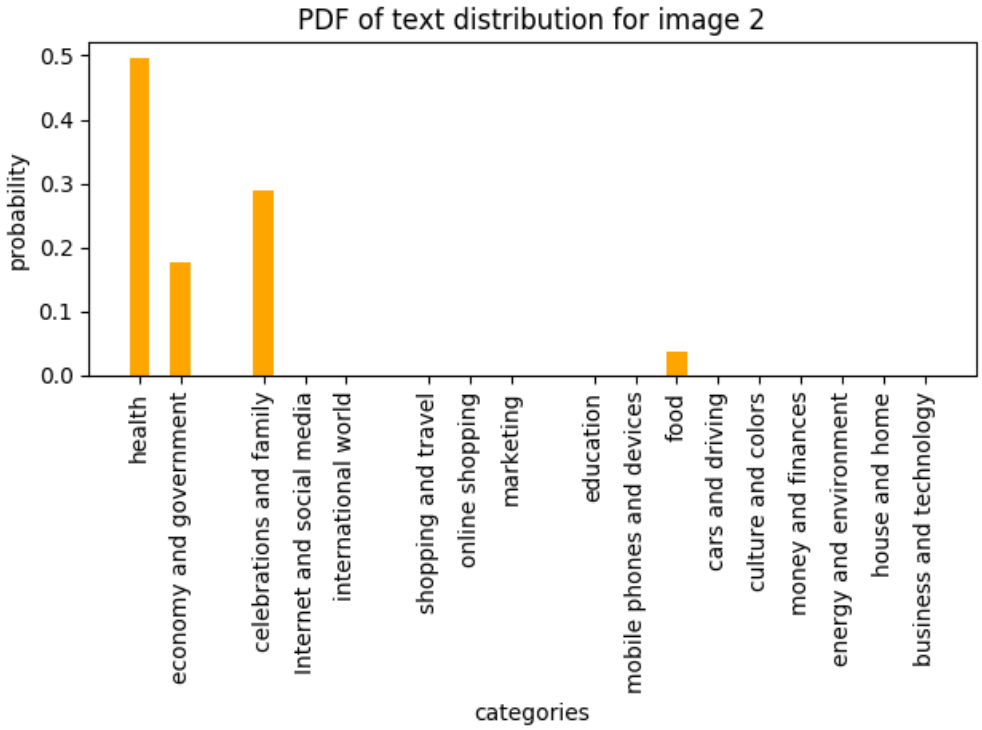


Figure 1. Sample predicted categories from trained LDA model for infographic 2 (SuperDad). All infographics we used in our study can be found in section 8.3 of the appendix.

4.2. Topic modeling for icons

Our original goal was to develop a version of LDA that worked on images instead of text. We attempted the following technique: first, we fine-tuned a CNN to categorize icons, then we clustered the icons based on their feature vectors as output by the CNN, then we used the cluster ids of an infographic's icons as the words in its document in order to run LDA. After looking at the output of this method, we found that the results were very hard to interpret. This is largely because many of the icons tended to cluster based on low-level features like color or presence of basic shapes, not based on higher-level visual or semantic similarity. While we have a few ideas on how to improve this, for this project we moved on to a different technique which gave us cleaner and more reliable data for modeling the interplay between images and text. More information about our original approach and future work on how we can improve it can be found in section 8.2 of the appendix.

We decided to create the visual category distributions manually by collecting human data on each individual icon in our infographics of interest. We used the human-

labeled bounding boxes from the Visual.ly data set to crop out each icon from each of the 12 infographics we planned to use in our user study, which resulted in approximately 200 icons. We created a pipeline to collect the labels of the top 3 categories (in order) that best reflected the icon. The category options were the top 5 categories outputted from the textual LDA so that the textual and visual category distributions would line up. One team member who did not have prior exposure to the full infographics was shown each of these icons one-by-one and was asked to select the most relevant categories. For example, the icon of people holding briefcases and on their phones was labeled 1) mobile phones and devices 2) internet and social media and 3) business and technology. Although we collected the top three categories for each icon, we realized the third one was usually a stretch compared to the other two and the first was usually a much better indicator than the second. We thus created the visual topic distribution for a single infographic by weighting the topic labels for each icon [2, 1, 0] and then summing over all its icons and normalizing.

Consider the example of an infographic with 2 icons (there were on average 16 icons per infographic in our user-study data set; we use 2 here for illustrative purposes). One icon was labeled 1) health, 2) mobile phones and devices and 3) internet and social media, while the other icon was labeled 1) mobile phones and devices, 2) internet and social media, and 3) business and technology. Our resulting categorical distribution for the full infographic based on the images is [1+2=3: mobile phones and devices, 2: health, 1: internet and social media], or [.5: mobile phones and devices, .33 health, .17 internet and social media] with 0 weight for the remaining categories.

4.3. Infographic Selection

We selected 12 infographics from the Visual.ly dataset to show to humans during our user study. We started by filtering down to the infographics for which we had human labeled icon bounding boxes (1,400 total) and computed the intersection of these with the infographics we used for LDA training. We then went through and hand-selected 12 infographics of similar dimension and text to icon ratio, in order to help control for some of the confounding variables that might arise due to variance in these parameters.

4.4. Human experiments

We set up a human experiment in order to collect human intuition around what topics are present in each infographic[3]. Each test subject was shown all 12 infographics in a randomized order where 4 infographics were shown for 1 second, 4 were shown for 5 seconds, and 4 were shown for 25 seconds. The combination of infographics and viewing time was different for each subject. For each infographic, subjects were given a cover story designed to illicit a distribution over what topics they thought were present in that infographic (see instructions in 2). Their options were the top 5 categories predicted by the LDA topic model. In the case that less than 5 topics were given any weighting by the LDA prediction, we chose one of the remaining topic categories at random to bring the number to 5. The same 5 categories shown to humans in the study were the ones used when creating the visual topic model. Figure 2 displays an example of the screen each subject would be asked to input their predictions into.

Our task was sent out to about 65 participants, of which about 50 completed the

[3] See an example of our task ui at $http://cocosci-study.herokuapp.com/?im_id=3n_s=5subj_id=9$

task in full. As a result, each (infographic, viewing duration) pair has abut 16-20 data points.

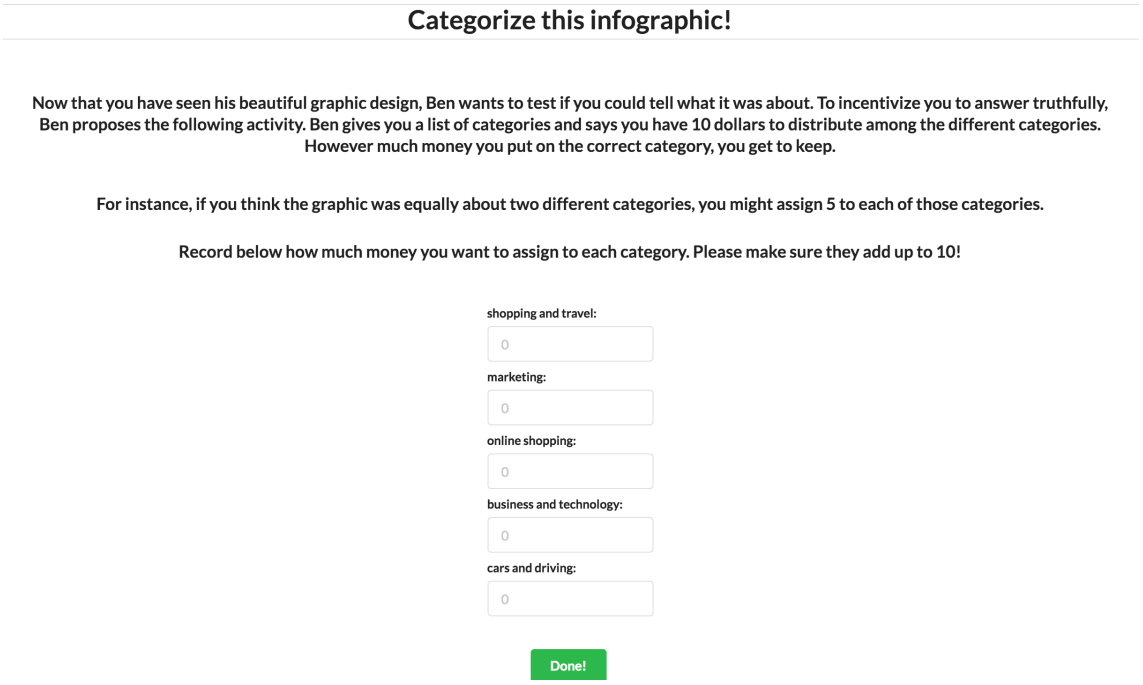


Figure 2. Human Experiment Study

From this set of approximately 20 distributions, we generated the final human distribution on a per-viewing-time basis by first normalizing each individual's input in order to correct for any scenarios where a subject's data did not add to 10. This maintained even weighting among each human's input. Then, we summed together the 20 human inputs and re-normalized to get a final human prediction for each infographic and time pair.

4.5. Model building

We parametrize the relative importance of the textual and image components by a parameter α. Let P_{text} be the textual topic distribution output by LDA, let P_{icon} be the visual topic distribution created from an image's icon labels, and let P_{human} be the distribution gathered from the user study. We want to set alpha so that the combination of the textual and visual topic distributions best match human judgement. In other words, we want:

$$P_{human} = \alpha P_{text} + (\alpha - 1) P_{icon}$$

Figure 3 shows an example of all 5 relevant distributions for infographic 0 (three empirical human distributions, text distribution, and visual distribution).

We optimize α by minimizing a loss function, for which we try both Mean Squared Error and Divergence. Due to the nature of our distributions with many zero values, the optimizer has a hard time converging on a valid α when we use Divergence as the loss function. As a result, our analysis is based on the results of the Mean squared error.

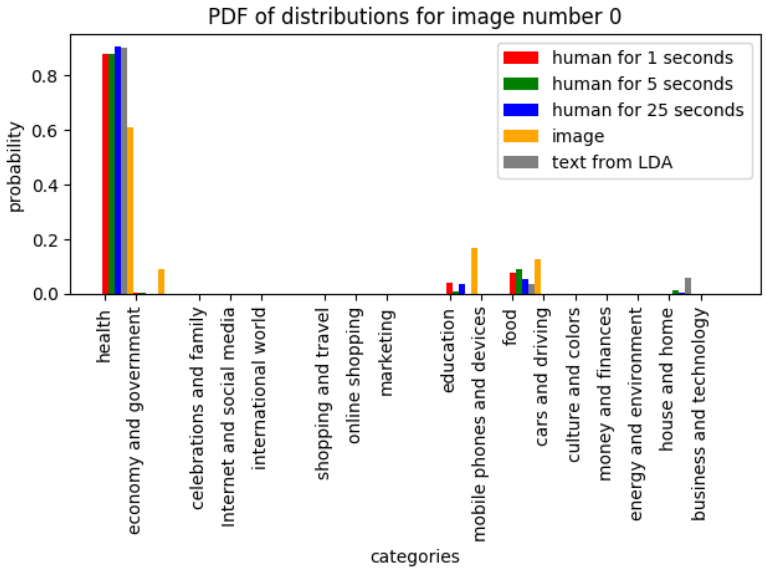


Figure 3. Probability Distribution for Infographic 0

MSE was calculated across all 20 possible categories as the difference between the probability distributions for each of categories. We calculate the MSE for 20 different values of α in order to generate a plot such as Figure 4 which shows the MSE versus α for each of the three time periods: 1 second, 5 seconds, and 25 seconds.

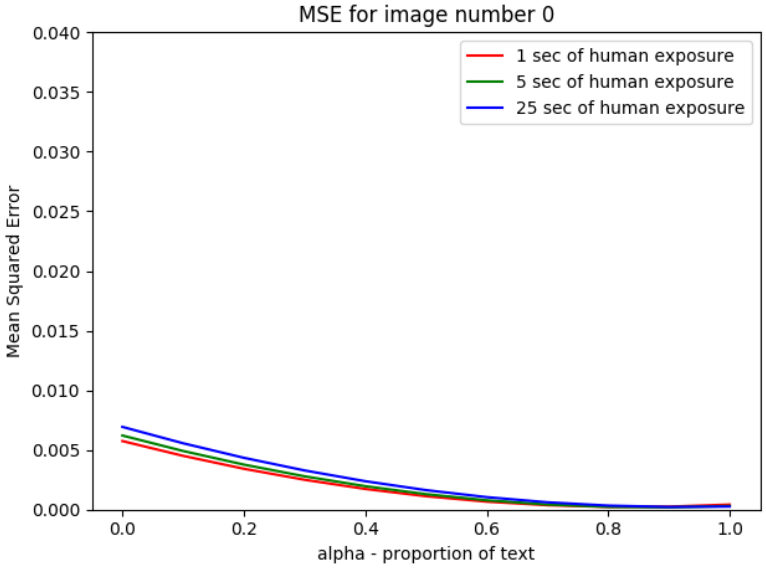


Figure 4. MSE for Infographic 0

See the contribution section for a link to the optimization code used to calculate the optimal α values.

5. Results

5.1. Optimal Mixing Parameter (α)

We determined the ideal mixing parameter α by averaging the results of the MSE's for each of the 12 infographics.

The results are shown in Figure 5.

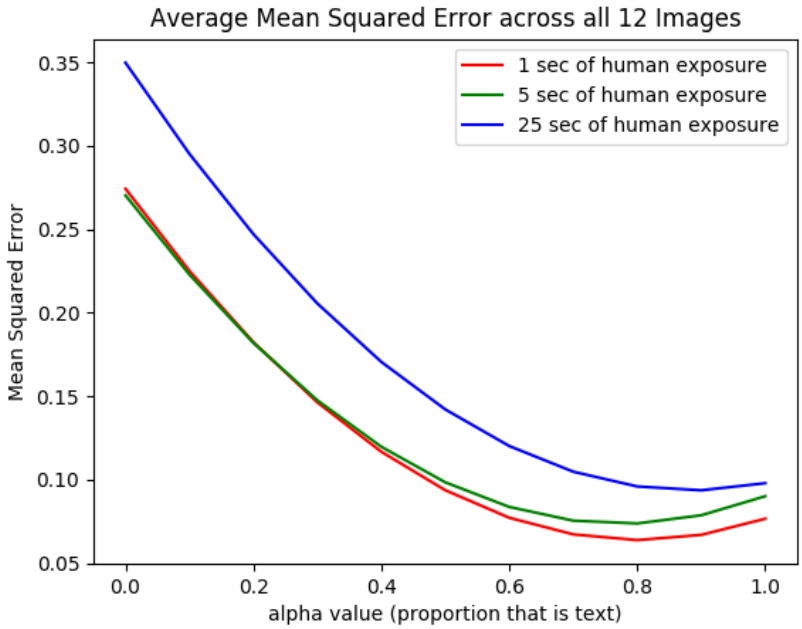


Figure 5. Optimal α

$\alpha = 0.802$ is optimal for 1 seconds with MSE loss of 0.0638. $\alpha = 0.775$ is optimal for 5 seconds with MSE loss of 0.0736. $\alpha = 0.885$ is optimal for 25 seconds with MSE loss of 0.0935.

Our hypothesis was that α would increase with time since we thought more time spent looking at an image would mean that people had time to absorb the stronger signals coming from the textual data. While this was true from 5 seconds to 25 seconds, the opposite was true from 1 to 5 seconds.

5.2. Deep Dive

While analyzing our data, we found that our model's performance varied a lot infographic-to-infographic. Thus, here we investigate some specific infographics in order to better understand the successes and failures of our model. The actual infographics we used can be found in the appendix.

Infographic 11

Infographic 11 can be found in Figure 25 in the appendix.

In this infographic, the categories predicted by the images and text are almost exactly the same. Furthermore, the categories predicted by the humans match the text/image predictions almost exactly. Therefore, across the different exposure times and across the

different values of α, Mean Squared Error is quite low. In this case, our LDA and image labels are good predictors of the categories predicted by humans.

Although MSE does not vary much with α in this case, we see an ideal α value of around 0.6, indicating that even when text and images are well-aligned, text is a better predictor of categories than images.

We saw about 5 infographics total exhibit patterns similar to that of infographic 11 where the model performs quite well in matching human intuition to the predicted distributions.

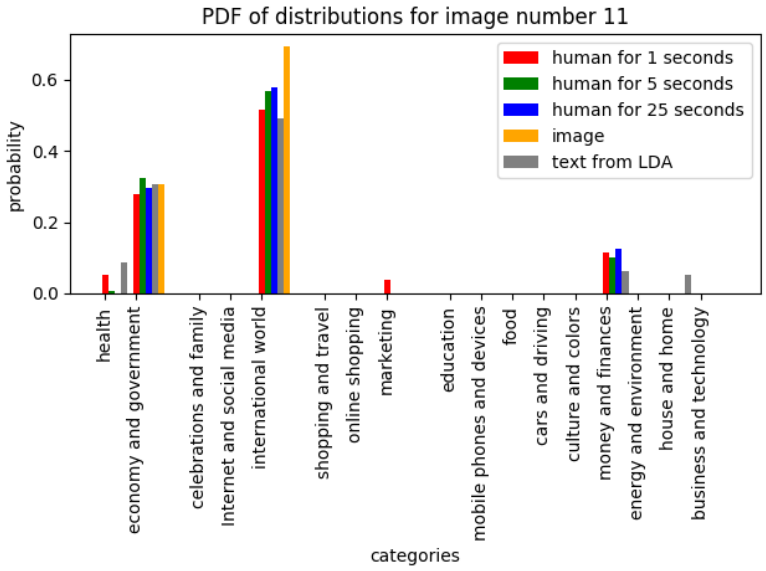


Figure 6. Probability Distributions for Infographic 11

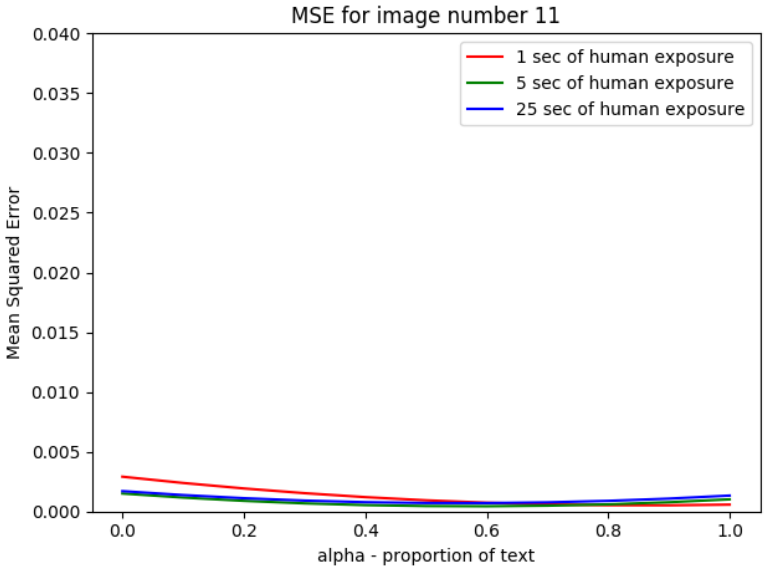


Figure 7. MSE for Infographic 11

Infographic 8

Infographic 8 can be found in Figure 22 in the appendix.

This is an example where the categories predicted by the text and images diverge significantly. Upon examination of the infographic, it is clear that the images convey that the infographic is about lemonade. Only once one reads the text, such as the header "Toronto's Daily Weather Data for 2000", does it become clear that the lemonade cups are metaphors for weather forecasts. We see that for viewing times of 1 or 5 seconds, human predictions correspond to the visual cues, whereas after 25 seconds of close observation human predictions correspond more to the text.

This was the only infographic where the categories for the text and image components were wildly different and led to divergent Mean Squared Error curves. This infographic does reflect our original hypothesis that a small α is appropriate for short viewing times and vice versa.

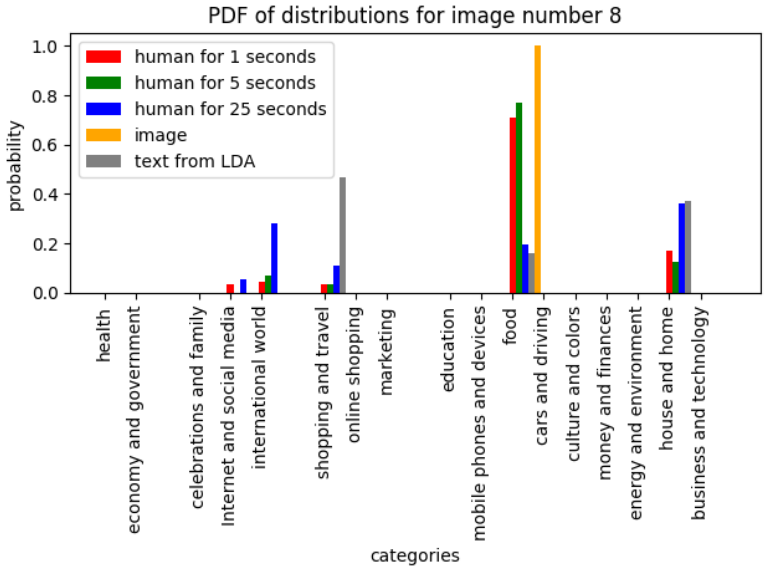


Figure 8. Probability Distributions for Infographic 8

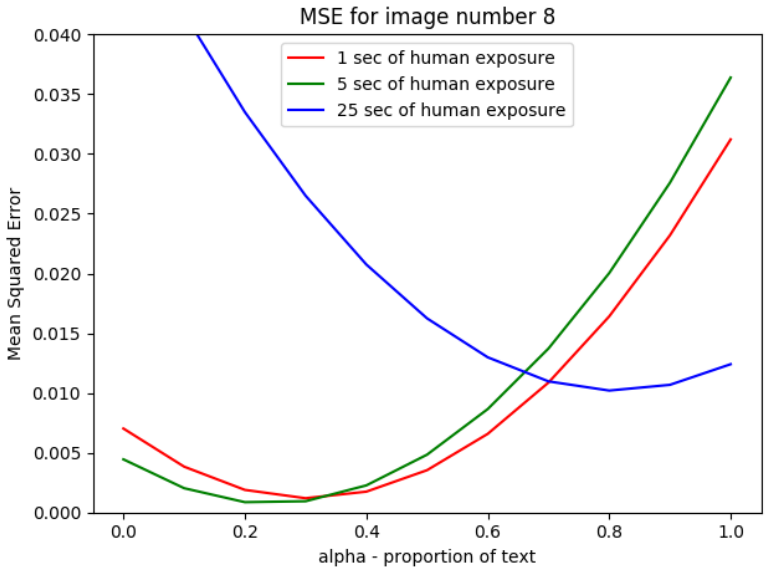


Figure 9. MSE for Infographic 8

Infographic 1

Infographic 1 can be found in Figure 15 in the appendix.

This infographic exposes a flaw in our model. In this infographic, human distributions are very consistent regardless of time exposure. All participants thought the image was about the environment. However, the actual icons contained in the image were not necessarily symbolic of the environment, so the image distribution is far off from human intuition. What's interesting here is that more subtle cues such as color scheme and bolded "E-Green" and "E-Waste" text likely tipped participants off that this infographic was about the environment. As a result, our model of simply incorporating the icons as the only source of visual content is likely insufficient to reflect human response to the infographic's visuals. A more correct model would need to account for global visual features including color scheme and variations in size and emphasis on words/icons.

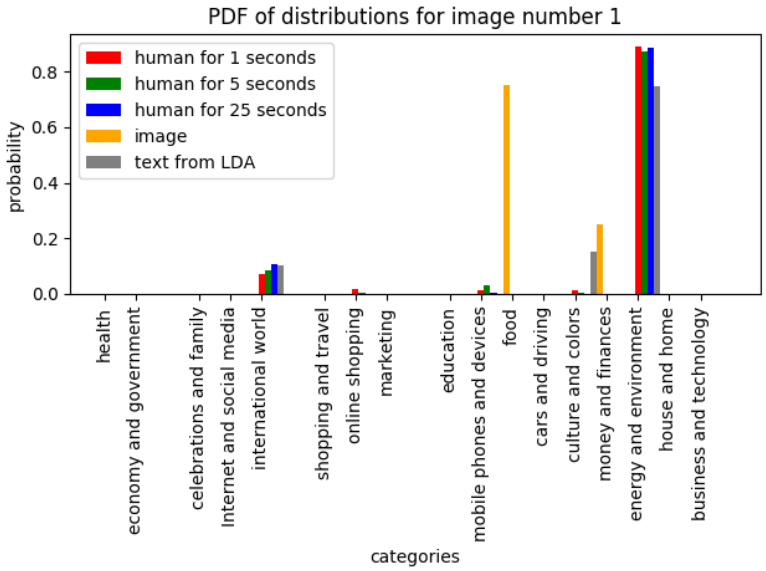


Figure 10. Probability Distributions for Infographic 1

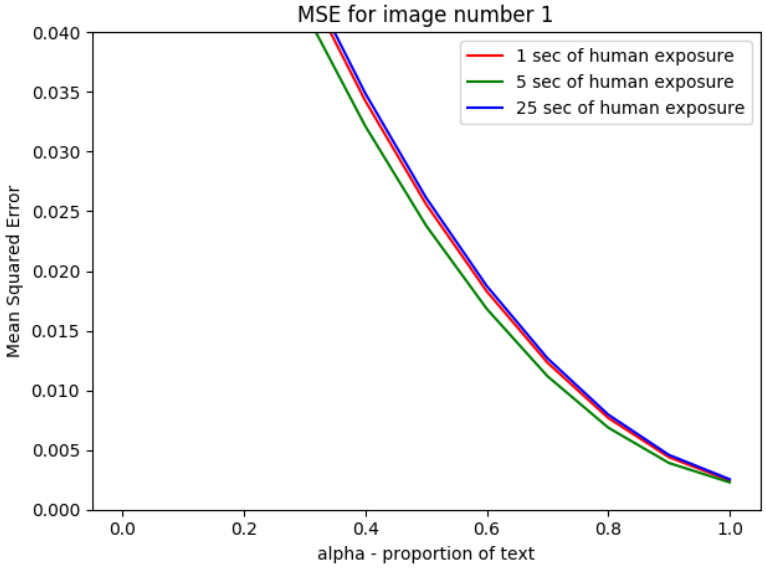


Figure 11. MSE for Infographic 1

Infographic 4

Infographic 4 can be found in Figure 18 in the appendix.

This infographic on the health benefits of Wine vs Beer shows a scenario where more time spent looking at an infographic is associated with a larger divergence between human cognition and our text-image predictions. This points to another limitation of our model compared to human cognition. LDA works by predicting categories based on probability distributions of words across documents. Although to any human it is clear the theme of this infographic is health, most of the words on the page are related to wine, beer, and food. After spending enough time to understand the infographic, humans are able to

recognize that the motivating theme of the infographic is health, but the text distribution predicts ”food” as the most probable category.

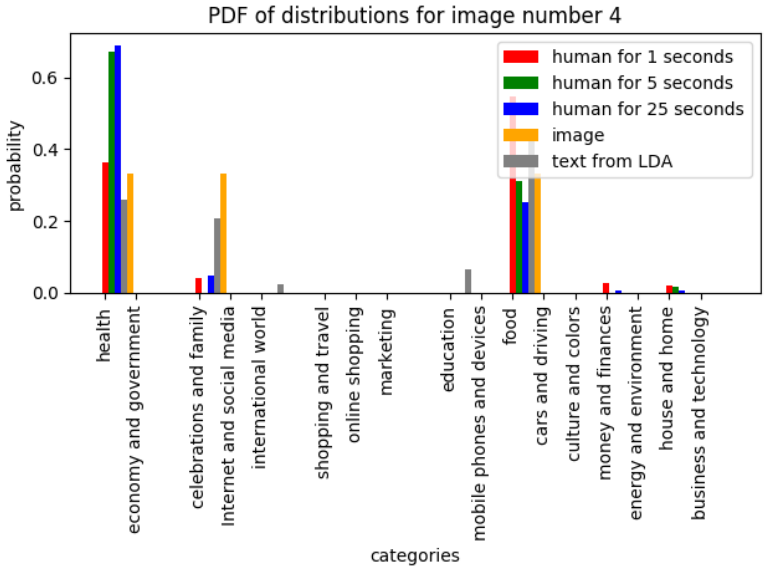


Figure 12. Probability Distributions for Infographic 4

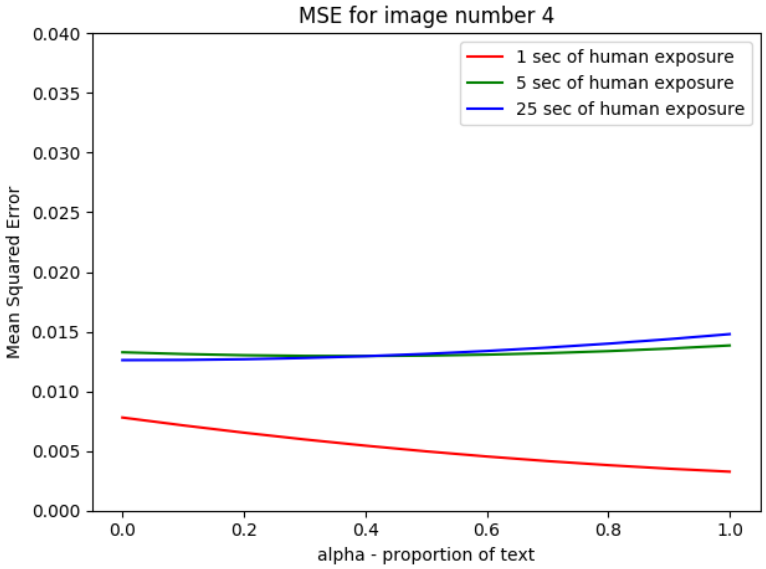


Figure 13. MSE for Infographic 4

All the distributions and MSE curves for all 12 infographics, including those not shown here, can be found in the appendix.

6. Discussion

Overall, the results on these 12 infographics give us some interesting data about the speed and sophistication of human cognition. However, our model suffers from many confounding variables that keep it from being a reliable predictor of multimodal visual

understanding. We explore what we can learn from the data and what further analysis could be done to better model how humans process multimodal media.

First, we were surprised by how little time is required for people to have a pretty good idea of what an infographic is about. We feared that 1 second would be too little time to extract any meaningful data from the visual. On the contrary, for most of the infographics for which we collected data, human predictions after 1 second were quite similar to predictions after 25 seconds.

Another important trend is that the textual data has proven relatively more important than images in determining topics in almost all cases where the infographic is "simple". Here we define simple to mean that the information contained in the images and icons support each other rather than offering divergent information. We see that in all cases except for infographics 8 and 9 (weather as lemonade / India Can't Feed her People) that the textual data is more important across the board. As a result, we conclude that humans tend to prioritize the textual information they consume over the visual information they consume.

We also note that in many cases our model performs the same or worse at predicting human data in the 25-second case than in the shorter cases. This could be because there is very little need for longer exposure to the images to actually read the text in detail and it is easy to decipher the message quickly. However, this could also be a sign that LDA is not that good of an approximation for how humans interpret textual data. It is possible that the LDA model is crude enough that it corresponds most closely to human impressions after 1 second because the textual LDA cannot understand the high level themes that humans pick up on with more time to observe. This is particularly evident in Infographic 4 (Wine and Beer). We see that while our LDA text model is good at producing the what in the text, it is not good at inferring the why or the how. As a result, human predictions better match the text and image models at smaller amounts of time.

By contrast, for confusing infographics such as Infographic 8 (lemonade and weather), time was far more important in realizing the images were being used as metaphors rather than literals. With more time on these images, we get results that match the text and image predictions better to human cognition. While humans get the topics right very quickly in almost all other cases, when the visual cues are not directly correlated with the purpose of the entire image humans are clearly confused.

There are several confounding variables and alternative approaches that could lead to more conclusive results. For example, one issue is that our visual probability distributions are not well-aligned with the true information content of the visuals in that color schemes and relative icon sizes are not captured in any way. One improvement would be to add a third variable capturing the effects of color scheme and design. We also could normalize the category probabilities predicted for each icon by the size of the icons. Larger icons are probably more important in determining categories than smaller icons; but we did not account for this in our model. We could also make the sample size larger than 1 for the generation of these manual labels as this would likely lead to a probability distribution more reflective of the general population's thoughts on categories based on icons.

Further, several of our mean squared error curves are very similar. Given that our sample size for each scenario is around 16-20, it is likely that the differences are not

actually that indicative of actual trends in many cases. We would like to expand our study sample size to be much larger in future iterations to be able to make stronger conclusions.

We would also like to explore alternatives to LDA to better understand the relative importance of topics in text. As mentioned in the Wine/Beer infographic analysis, LDA captures the what but now the how or why. As a result, it would be interesting to supplement the model so that the textual categories are more aligned with human cognition.

Further, we see that the infographics overall were incredibly different from each other. While we controlled for dimensions / shape and rough relative image versus text content, it is possible that choosing infographics more similar to each other would eliminate some confounding variables and allow us to draw more informative conclusions, albeit for a more limited set of media. If we really wanted to dig into a low-level understanding of how the brain weights text versus visual content, we could try using much more controlled stimuli than real-world infographics, such as simple icon-word pairings.

Overall, we make some intersting observations about human cognition with respect to multimodal visualizations: humans tend to weight text over visuals, and in all but the most confusing cases they are able to digest the main idea of an infographic quickly, within a second, even before fully absorbing the text.

7. Contributions

Please check out our github to see the code supporting our analysis. All relevant components are linked from the README: https://github.com/meiaalsup/lda_infographics

The following parts of the project were a group effort: ideation and brainstorming, collecting data for the user study, analysis, and writing the report.

Anelise designed and coded the user study. She explored LDA for icons using CNN-based feature vectors and clustering and also set up the pipeline to hand-label icons to create the visual topic model.

Meia developed the topic model for the textual data using LDA, wrote the optimization code for the mixing parameter α, and built out the distribution and error visualizations.

Sravya did the hand-labeling for the visual topic model.

We want to give a special shout-out to Jon Gauthier for meeting with us and providing valuable insights and feedback as we developed our project plan. We would also like to thank the entire course staff for making this project possible.

References

[1] L. L. et. al, "An overview of topic modeling and its current applications in bioinformatics." `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/`, 2016. [Online; accessed 14-December-2018].

[2] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall.," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 519–528, 2016.

[3] Z. Bylinskii, S. Alsheikh, S. Madan, A. Recasens, K. Zhong, H. Pfister, F. Durand, and A. Oliva, "Understanding infographics through textual and visual tag prediction," in *arXiv preprint arXiv:1709.09215*, 2017.

[4] S. Madan, Z. Bylinskii, M. Tancik, A. Recasens, K. Zhong, S. Alsheikh, H. Pfister, A. Oliva, and F. Durand, "Synthetically trained icon proposals for parsing and summarizing infographics," in *arXiv preprint arXiv:1807.10441*, 2018.

8. Appendix

8.1. LDA Text-Based Topics

Table 1 shows top words and our names for our 20 textual LDA topics. After looking at the data, we figured out that category 2, "work and positivity", and category 10, "beach", were mostly noise, so we discarded them. We also decided to combine category 0, "health care", and category 6, "body and health", because they were substantially similar.

8.2. Training a CNN for icon labels

We used tools from Computer Vision to extend LDA to apply to images. Unfortunately, the outputs of this technique were too noisy to use reliably while analyzing the user study data. However, we think this technique is promising and could be improved in order to produce an end-to-end model for human infographic understanding.

8.2.1. Methodology

First, we wanted to obtain an efficient representation of an icon that captured only the most relevant data about it. We did this by extracting a feature vector from a CNN. As in [4], we fine-tuned a Resnet-18 CNN pre-trained on ImageNet to categorize icons. We used the 250k labelled icons included in the Visual.ly dataset as training data. We then removed the last layer from the network and used the second-to-last layer as a length-512 feature vector[4].

In order to feed these elements into LDA, we needed to convert them into one of many discrete token ids. We did this by clustering the feature vectors using k-means clustering with $k = 1000$. In order to save on computational intensity, we calculated clusters based only on the icons with clean, human-annotated bounding boxes (1400 infographics). We then categorized each computationally-detected icon from the rest of the dataset (45k infographics) into the closest cluster.

Finally, we used LDA to generate topics.

8.2.2. Results

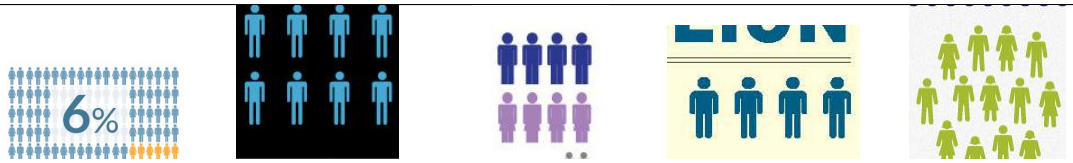
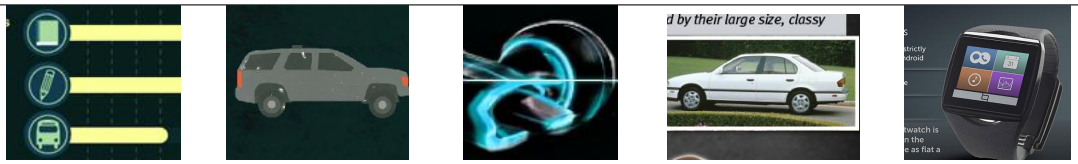
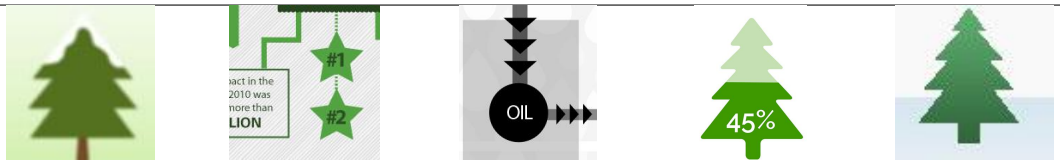
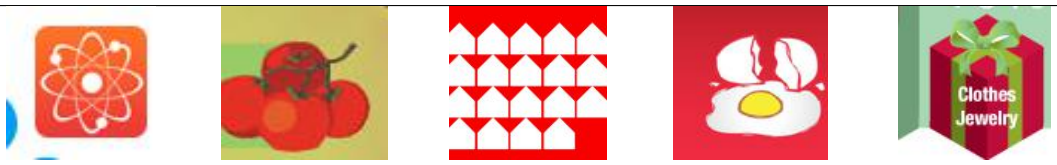
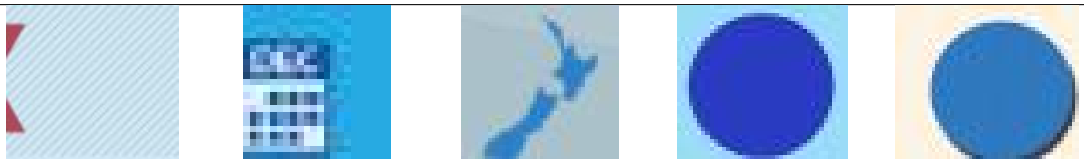
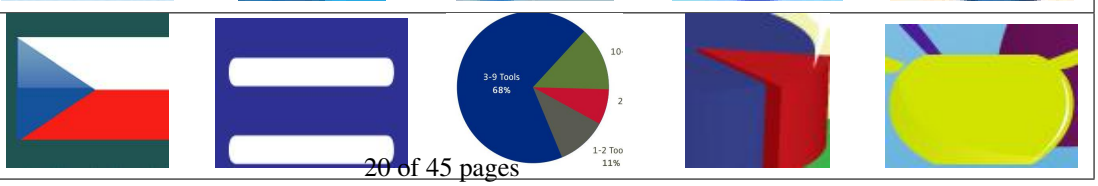
Figure 2 shows some of the icon clusters produced by our method. Each row shows sample icons from a single cluster. As can be seen in the figure, there is significant variation in the quality of the clusters. Although some contain very similar and semantically significant icons, such as groups of people or a particular flag, others are only grouped by low-level, semantically insignificant features like color or basic shape, or are very noisy. As such the output is hard to interpret. Thus, we decided to use human-annotated icon categories as input to our model instead of trying to generate accurate automatic information.

[4] code available at https://github.mit.edu/apnewman/icon-fine-tune/

Table 1. Text-based topics produced by LDA

Id	Topic name	Top 20 words
0	health care	women, health, people, age, million, children, care, men, years, year, medical, americans, child, average, likely, more, adults, are, drug, percent
1	economy and government	new, million, state, states, billion, tax, total, national, united, jobs, north, growth, number, economic, york, government, public, years, federal, industry
2	work and positivity	time, work, job, people, new, good, like, need, hours, want, best, know, think, help, way, feel, things, get, start, life
3	celebrations and family	day, wedding, christmas, love, family, wine, average, kids, party, gift, gifts, baby, people, flowers, dress, days, women, night, guests, home
4	Internet & social media	social, media, users, search, google, twitter, online, use, people, web, website, page, content, site, internet, million, video, sites, post, share
5	international world	world, countries, united, south, china, london, international, top, india, country, australia, france, europe, million, new, english, usa, city, germany, number
6	body and health	body, skin, blood, brain, heart, risk, help, sleep, pain, hair, cause, loss, use, stress, cancer, disease, teeth, common, weight, people
7	shopping and travel	travel, shopping, holiday, online, sales, new, retail, year, average, hotel, day, top, spend, shoppers, spending, free, shipping, price, stores, shop
8	online shopping	people, online, want, need, are, use, yes, know, service, card, money, help, like, customers, buy, pay, information, offer, good, time
9	marketing	marketing, content, brand, marketers, use, email, create, lead, product, sales, brands, leads, design, campaign, audience, quality, media, advertising, digital, customers
10	beach	feet, light, size, long, miles, park, white, square, shoes, sea, inches, great, beach, area, cut, look, black, island, diamond, head
11	education	data, students, college, school, education, information, university, technology, learning, online, degree, high, public, student, use, schools, study, test, new, skills
12	mobile phones & devices	mobile, phone, million, users, billion, internet, devices, video, device, apple, use, market, online, android, phones, tablet, digital, new, games, year
13	food	food, cup, eat, foods, calories, fat, sugar, super, healthy, protein, vitamin, bowl, team, milk, diet, cheese, drink, tea, league, eating
14	cars & driving	car, insurance, safety, driving, vehicle, cars, drivers, injuries, road, accidents, drive, injury, driver, auto, cost, motor, speed, engine, vehicles, traffic
15	culture and colors	gold, design, music, color, game, games, won, best, red, win, series, blue, art, history, black, movie, popular, first, world, new
16	money and finances	average, credit, debt, financial, cost, money, pay, year, total, price, income, market, rate, home, loan, value, years, investment, costs, billion
17	energy and environment	energy, solar, power, million, year, gas, carbon, electricity, green, tons, waste, paper, use, emissions, production, billion, fuel, consumption, years, natural
18	house and home	water, home, air, use, dog, clean, pet, heat, coffee, house, kitchen, food, room, hot, heating, cleaning, dogs, save, space, pets
19	business and technology	business, businesses, companies, small, data, management, cloud, customer, employees, services, service, company, new, security, support, customers, development, technology, software, survey

Table 2. Clusters of icons produced by using k-means clustering on icons' CNN feature vectors. We were hoping that this technique would group icons that looked similar at a high enough level that they also shared a semantic meaning. This table shows the range of quality present in the clusters.

Description	Example icons from cluster
Clear semantic category: groups of people	
Clear semantic category: British flags	
Noisy semantic category: cars	
Noisy semantic category: trees	
Low-level visual category: (connected) dots	
Low-level visual category: red grids	
Color category: orange-red	
Color category: green	
Noise	
Noise	

8.2.3. Future ideas

Given more time, there are some techniques that we would try in order to make image-based LDA more robust and useful.

First, we notice in our results that color plays an outsized role in clustering icons; some clusters appear to have as their primarily feature that they contain icons of the same color. Although color can be an important clue in determining the semantic meaning of an icon (a recycling sign might be disproportionately likely to be green, for instance), it is often less important than shape for determining icon significance. One hypothesis for improving our process is that we could process all icons as grayscale images, eliminating color as a factor.

We could also use an image's raw pixels as input to the k-means clustering algorithm instead of a CNN-produced feature vector. We opted not to use raw pixels at the outset because this representation is not robust to horizontal flips, translations, inverted color schemes, etc. However, the clusters produced from the neural net vectors tend to be pretty noisy, which could indicate that the extra level of processing and abstraction introduced noise that interfered with coherent clustering and LDA. Using raw pixels with appropriate forms of normalization could lead to clusters that are more homogeneous and thus to better LDA predictions.

We realized while doing our user study that extracted icons alone do not convey all the visual information in an infographic. For instance, for one of our sample infographics, the green background color of the visualization was a strong clue that it was about the environment. In a future iteration of this technique, we would try to incorporate the entire visual into the pipeline in some way, possibly by including a thumbnail of the whole image or some random crops of the image along with the extracted icons.

Finally, in order for this automated method to work, it would be necessary to come up with a robust way of interpreting icon clusters/topics or of discarding noisy clusters.

8.3. Infographics used in human study (with bounding boxes)

Below, we include the 12 infographics that we asked people about in our user study, as well as the human-annotated bounding boxes corresponding to the extracted visual elements.

8.4. All distributions and MSE curves

Below, we include the probability distributions and MSE curves associated with all 12 infographics (note they are 0-indexed).

Figure 14. Infographic 0. Topic options: health, house and home, food, education, economy and government.

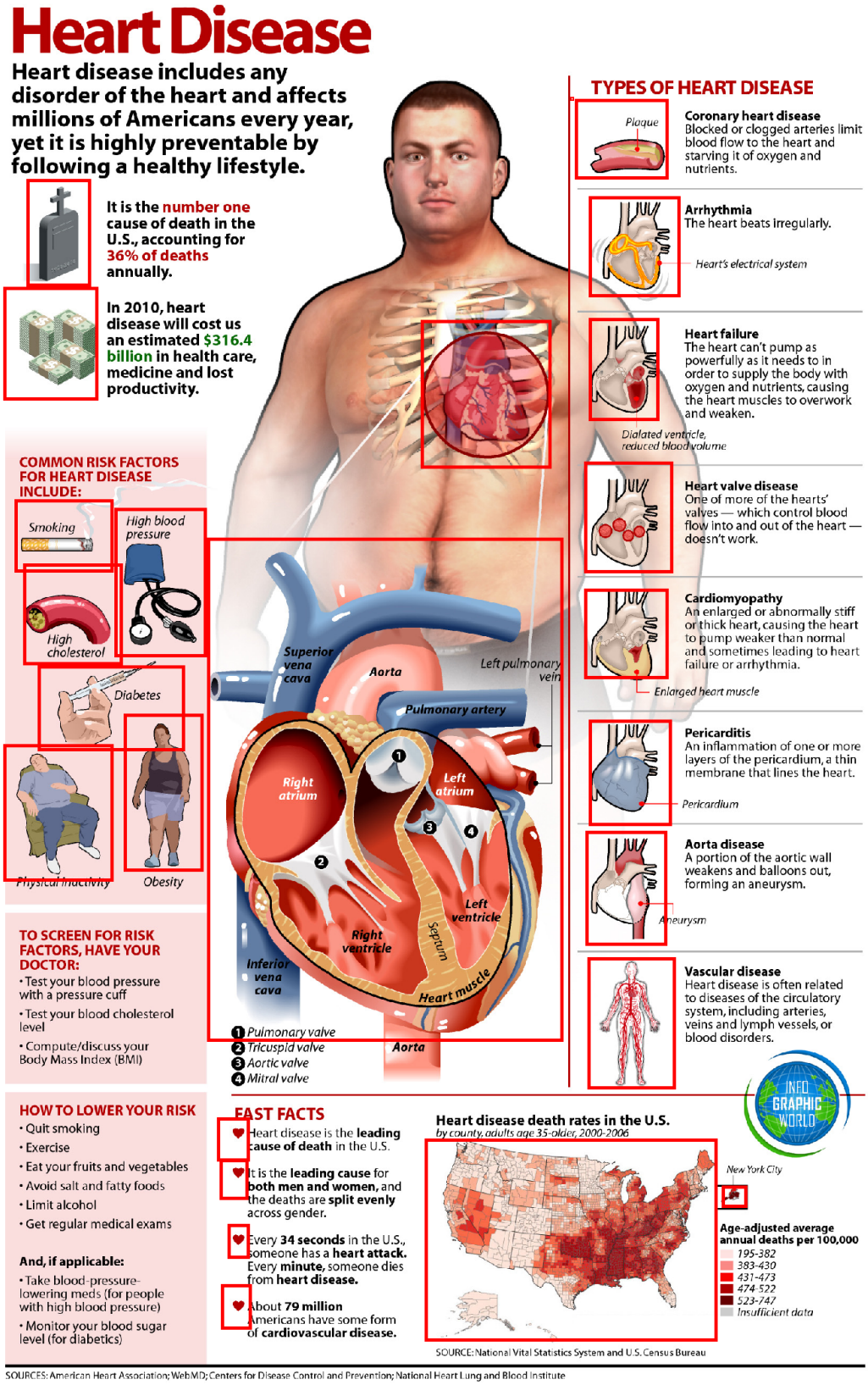


Figure 15. Infographic 1. Topic options: energy and environment, culture and colors, international world, mobile phones and devices, online shopping.

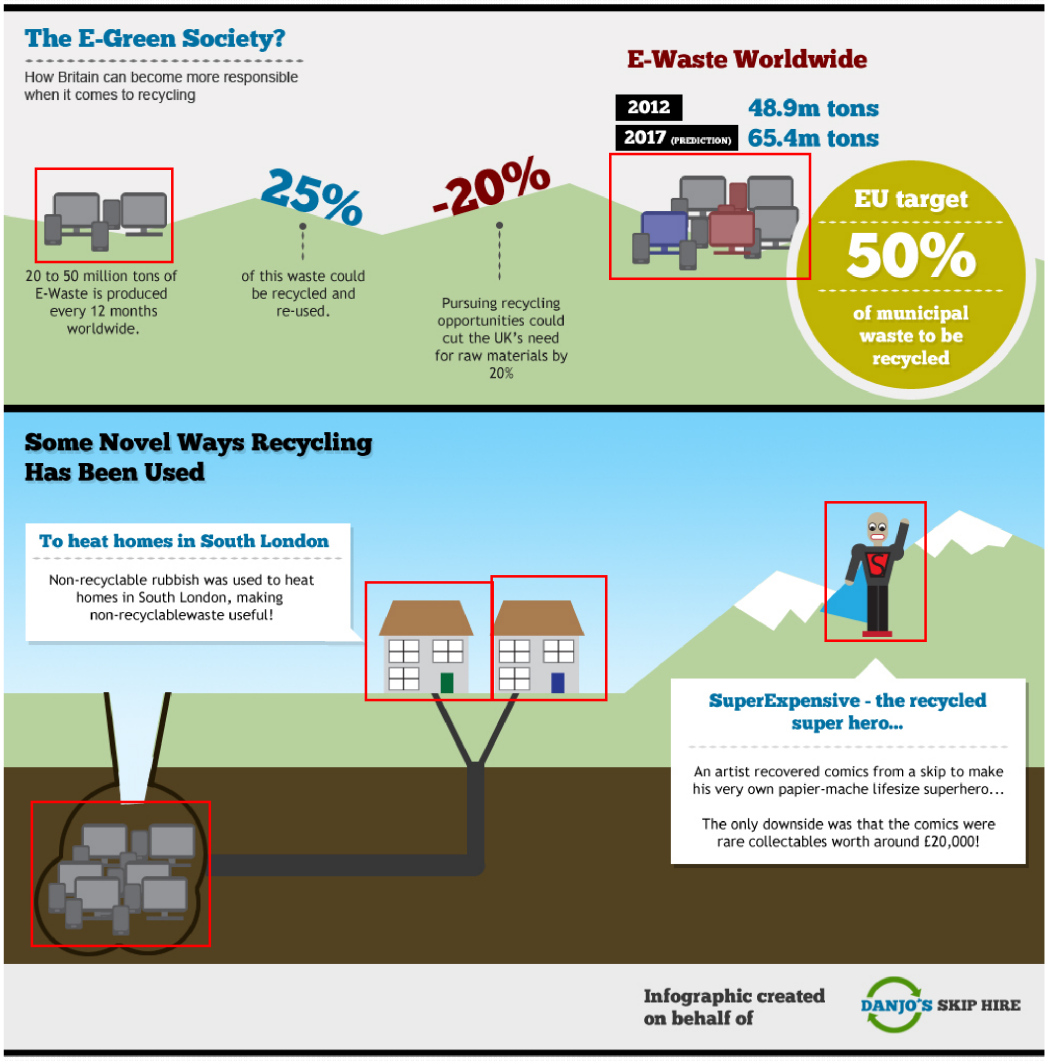


Figure 16. Infographic 2. Topic options: health, celebrations and family, economy and government, food, shopping and travel.

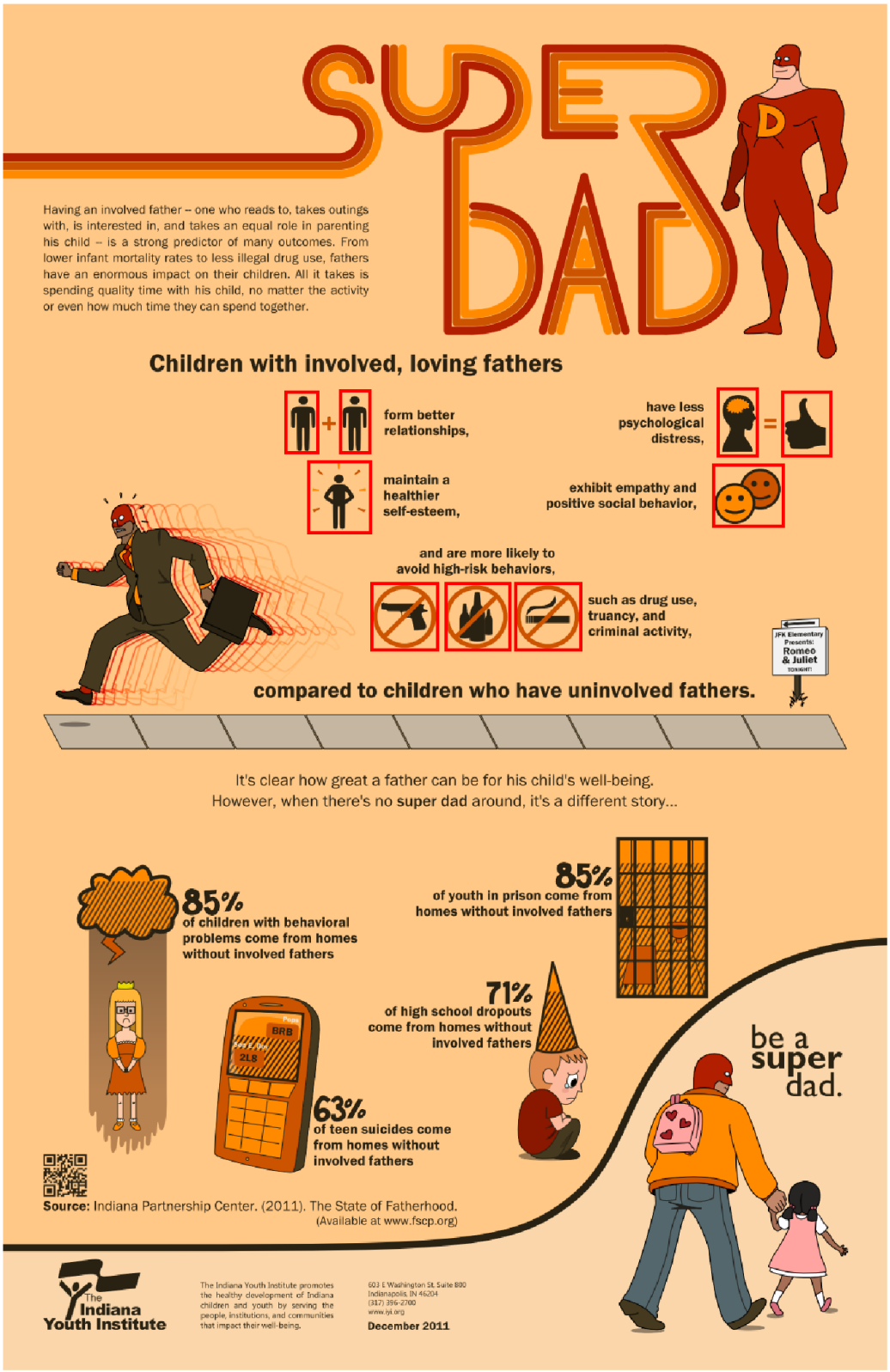


Figure 17. Infographic 3. Topic options: business and technology, shopping and travel, online shopping, marketing, cars and driving.

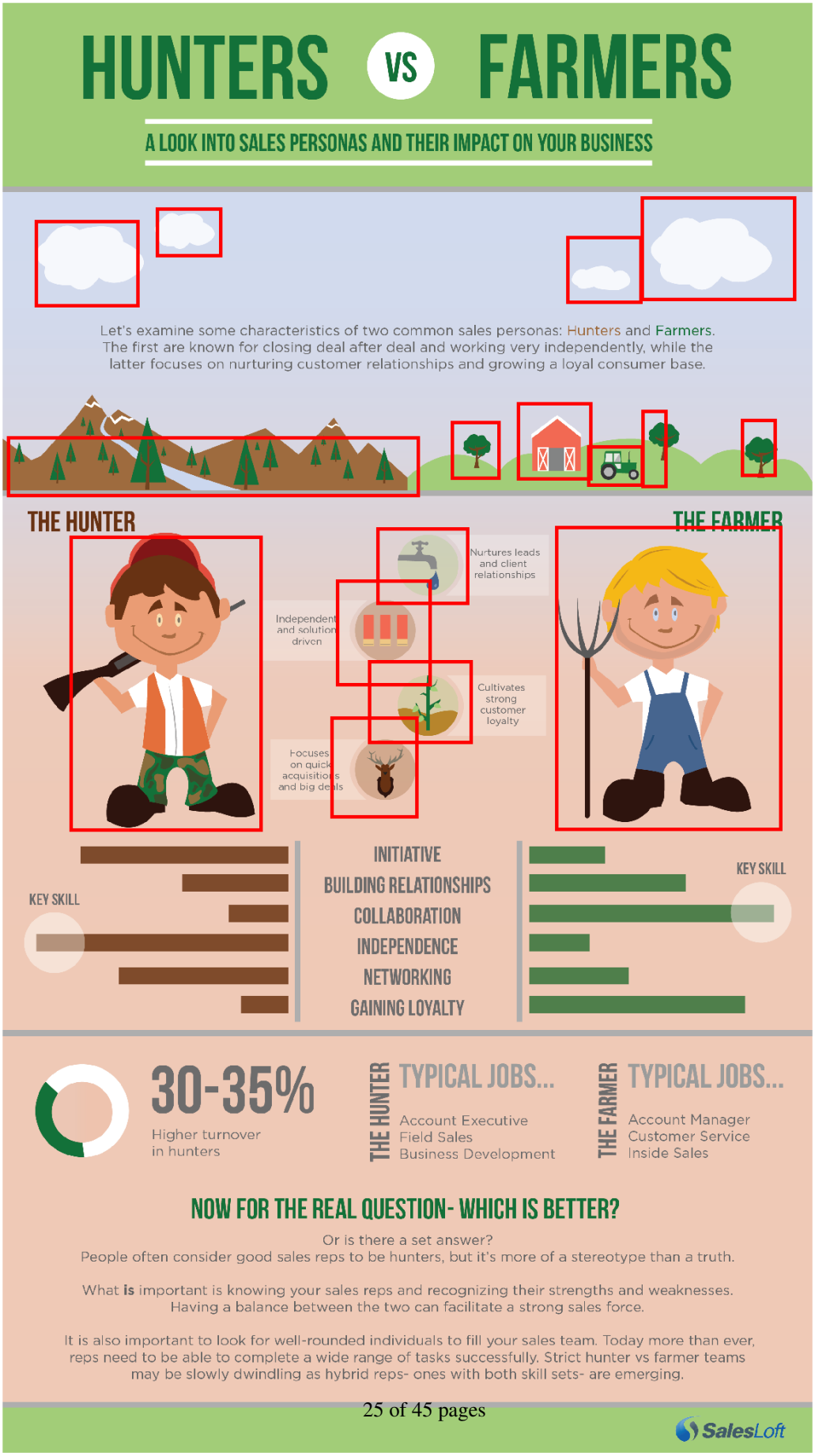


Figure 18. Infographic 4. Topic options: food, celebrations and family, health, money and finances, house and home.

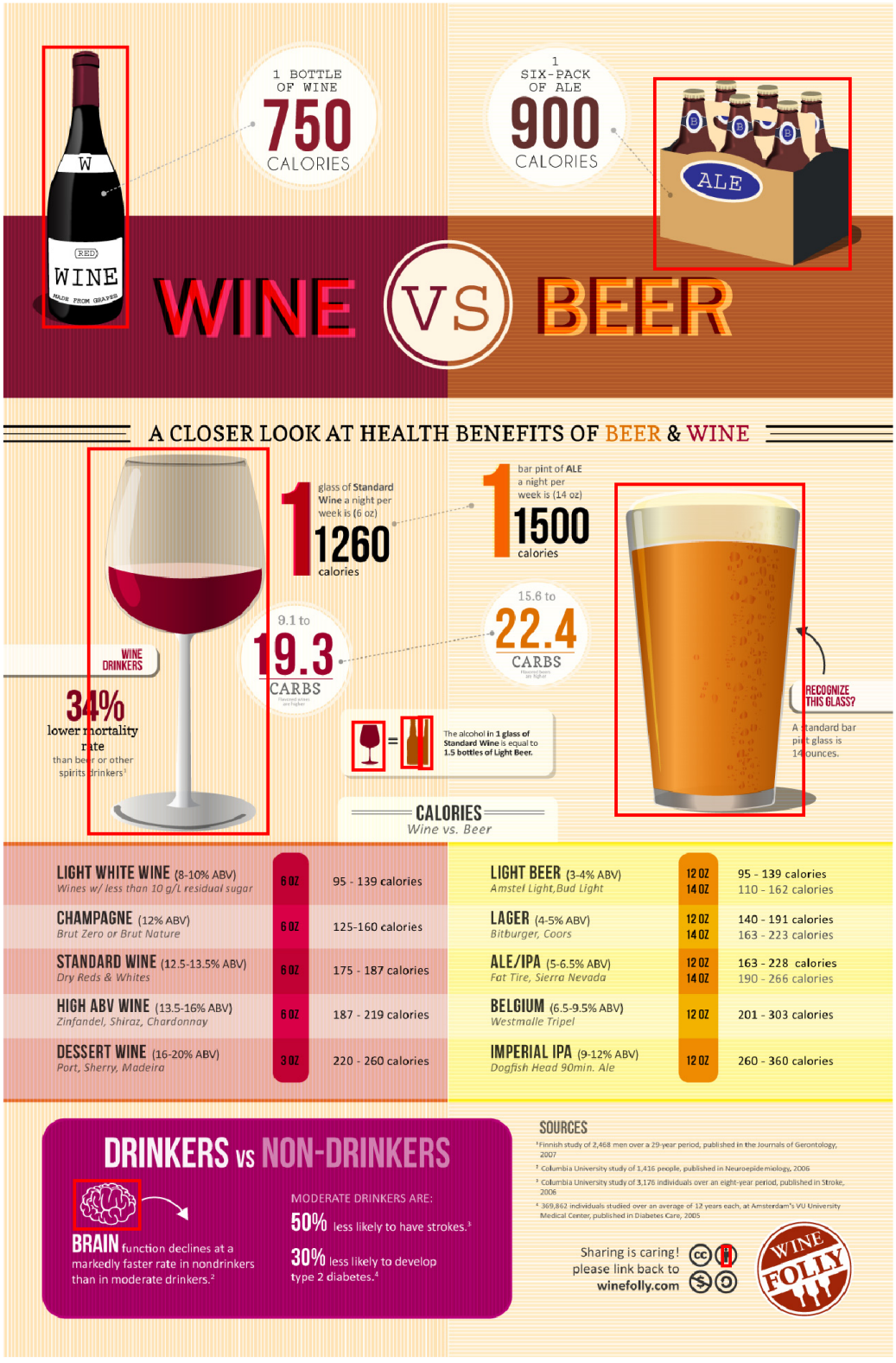


Figure 19. Infographic 5. Topic options: cars and driving, health, shopping and travel, economy and government, culture and colors.

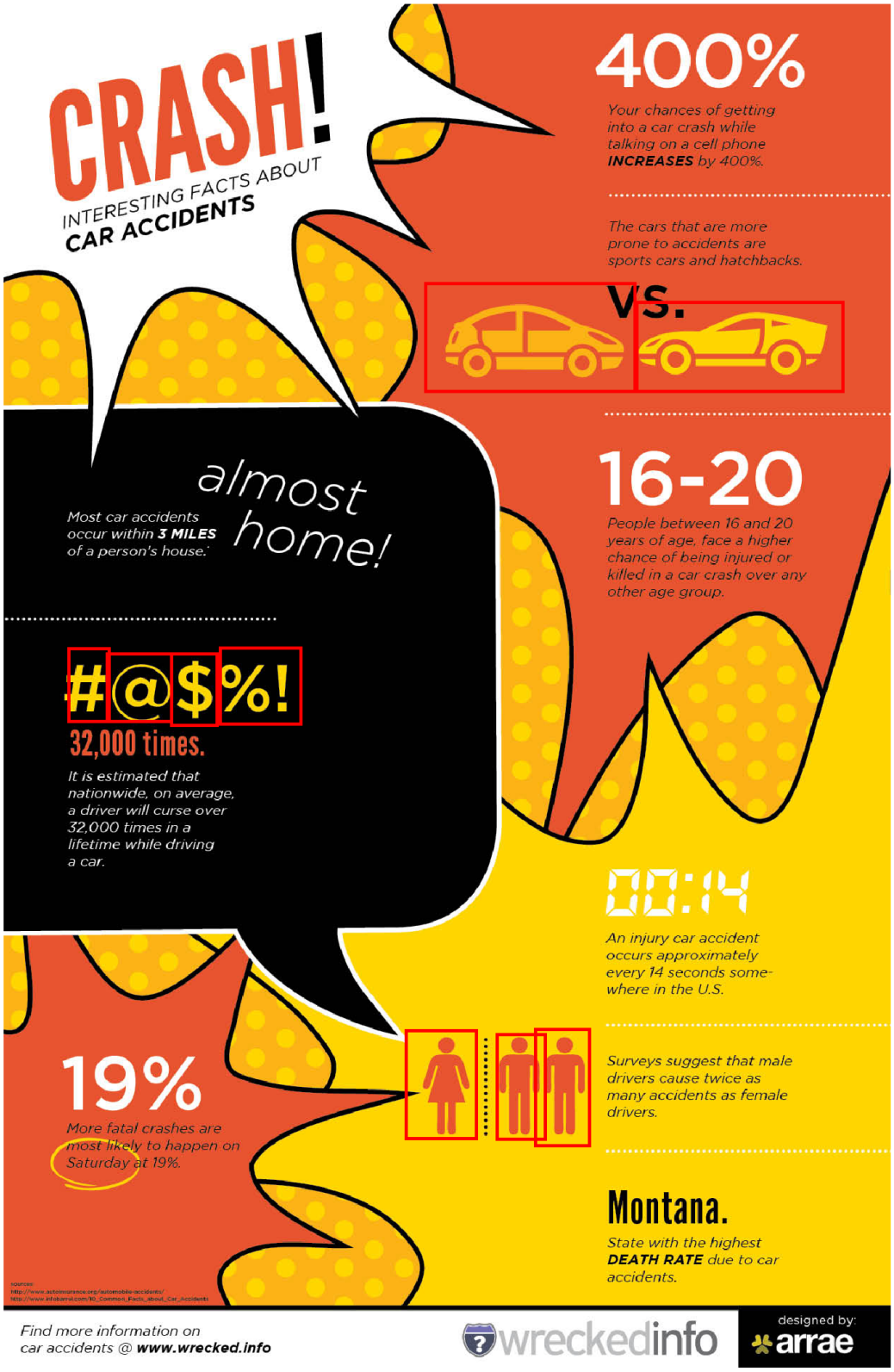


Figure 20. Infographic 6. Topic options: business and technology, mobile phones and devices, international world, online shopping, shopping and travel.

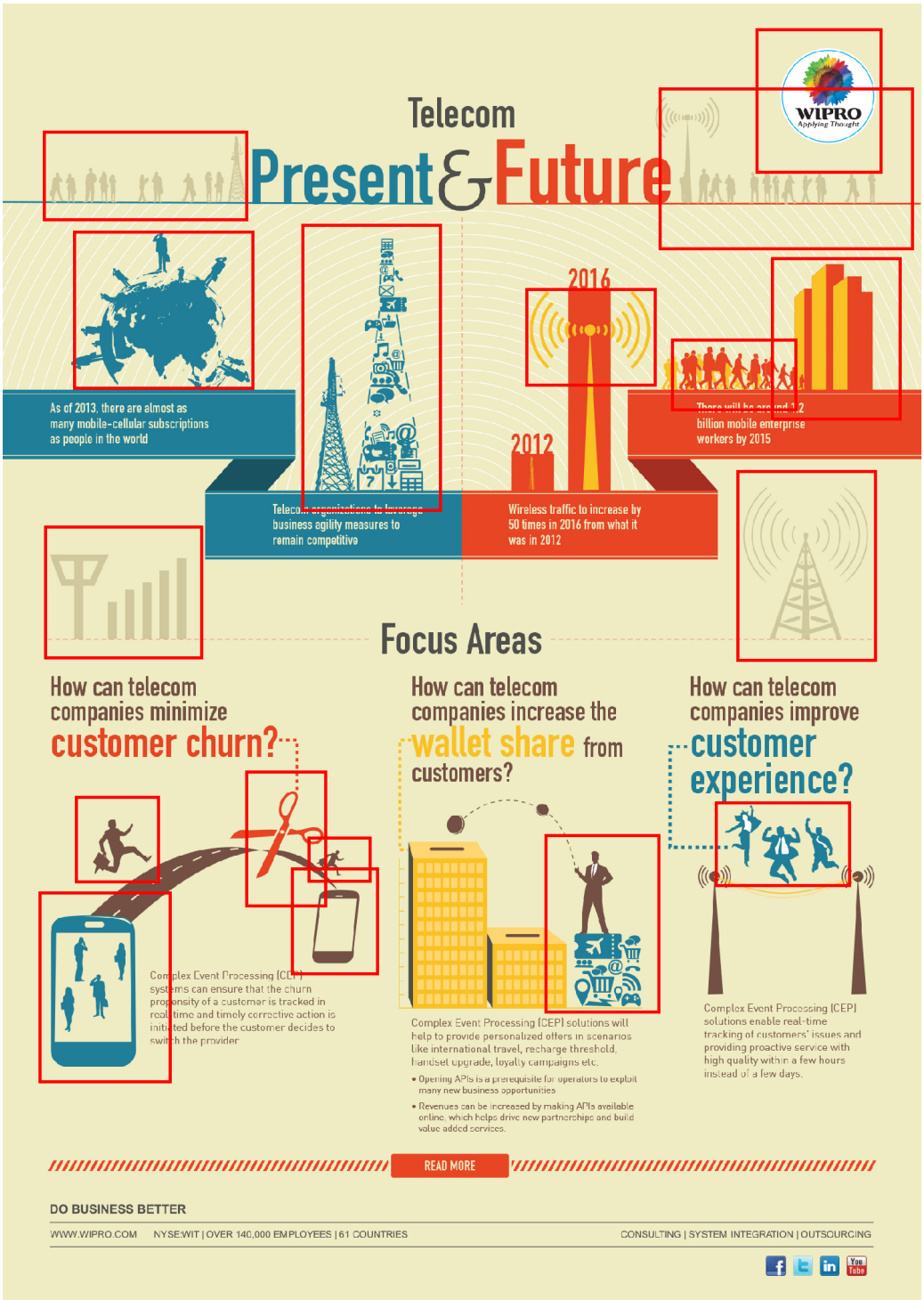


Figure 21. Infographic 7. Topic options: celebrations and family, online shopping, culture and colors, international world, Internet and social media.

Figure 22. Infographic 8. Topic options: shopping and travel, house and home, food, Internet and social media, international world.

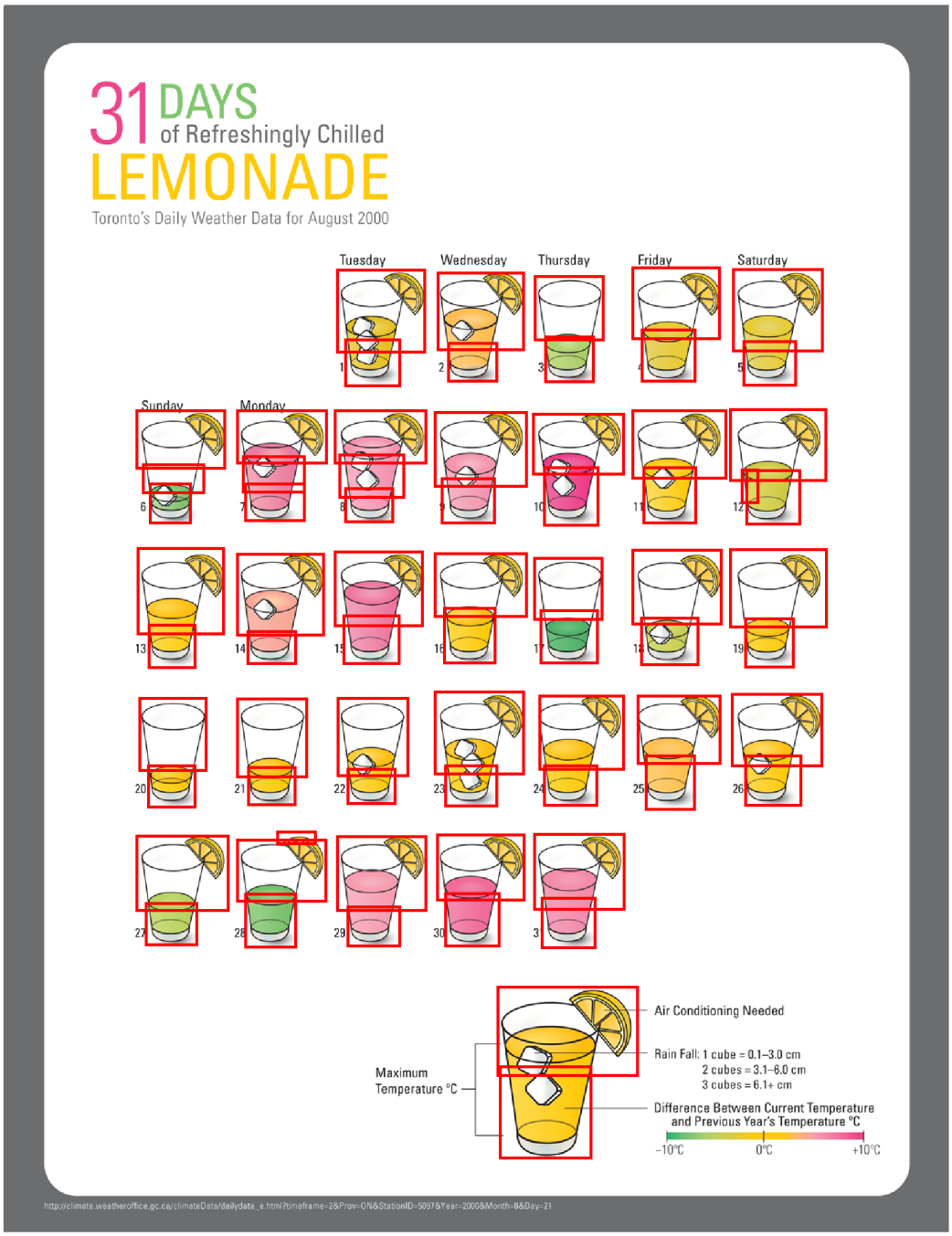


Figure 23. Infographic 9. Topic options: food, health, energy and environment, international world, house and home.

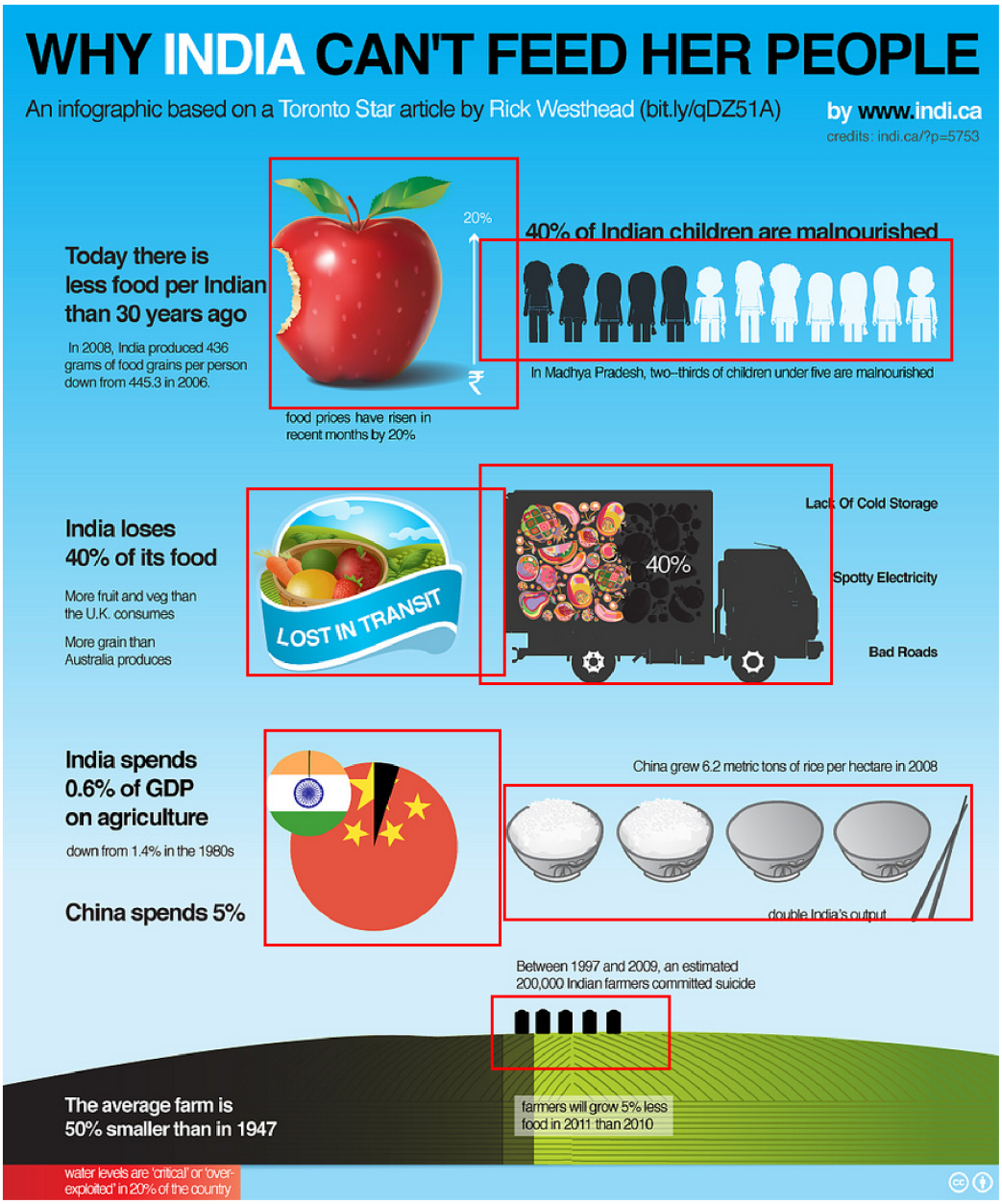


Figure 24. Infographic 10. Topic options: cars and driving, mobile phones and devices, economy and government, health, international world.

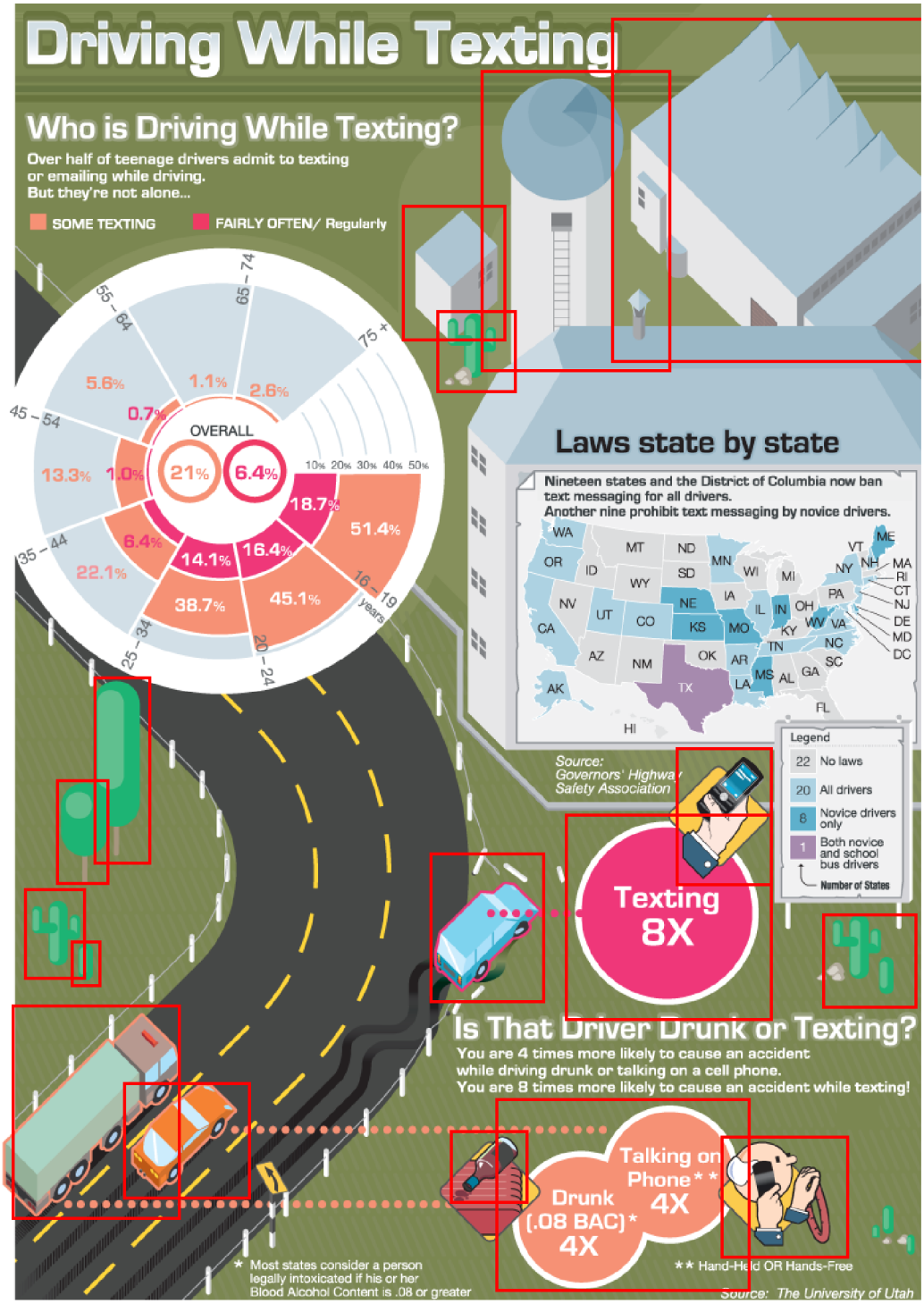
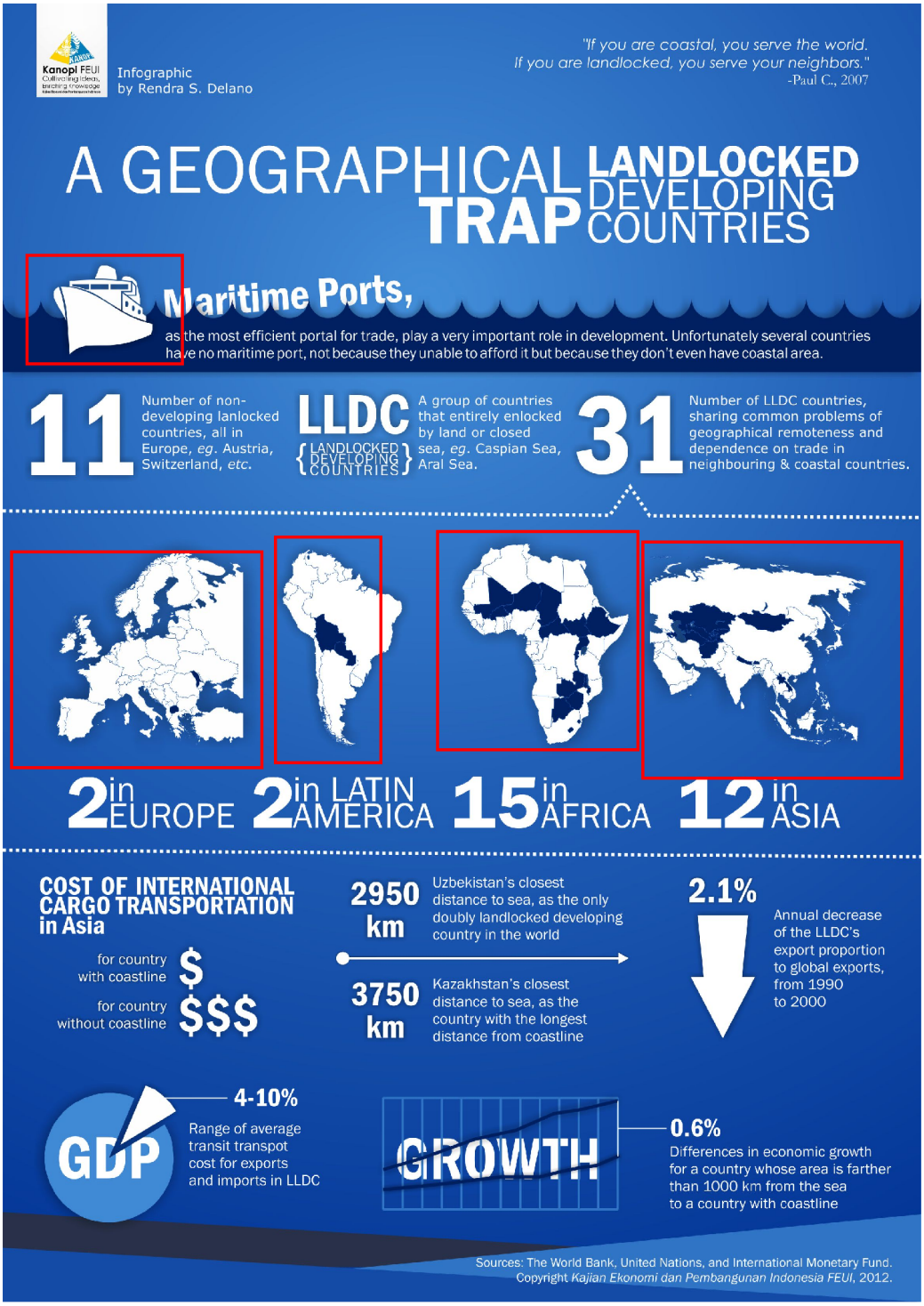


Figure 25. Infographic 11. Topic options: international world, economy and government, health, money and finances, marketing.



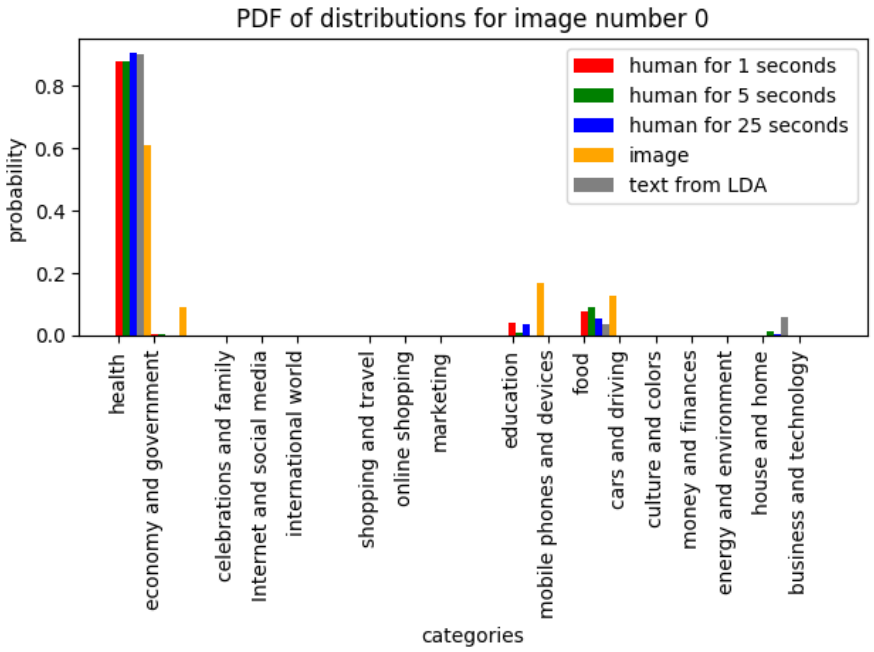


Figure 26. Probability Distributions for Infographic 0

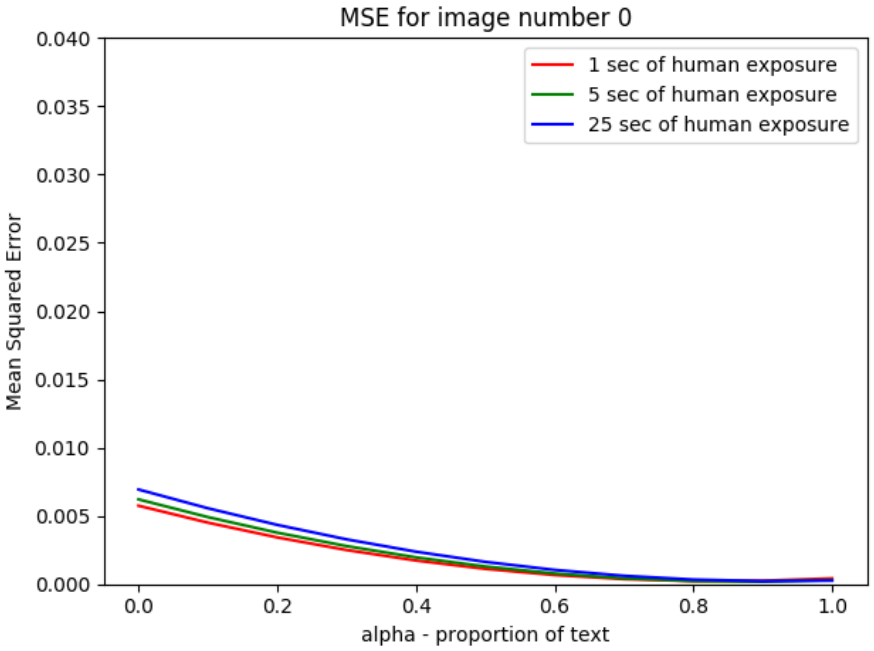


Figure 27. MSE for Infographic 0

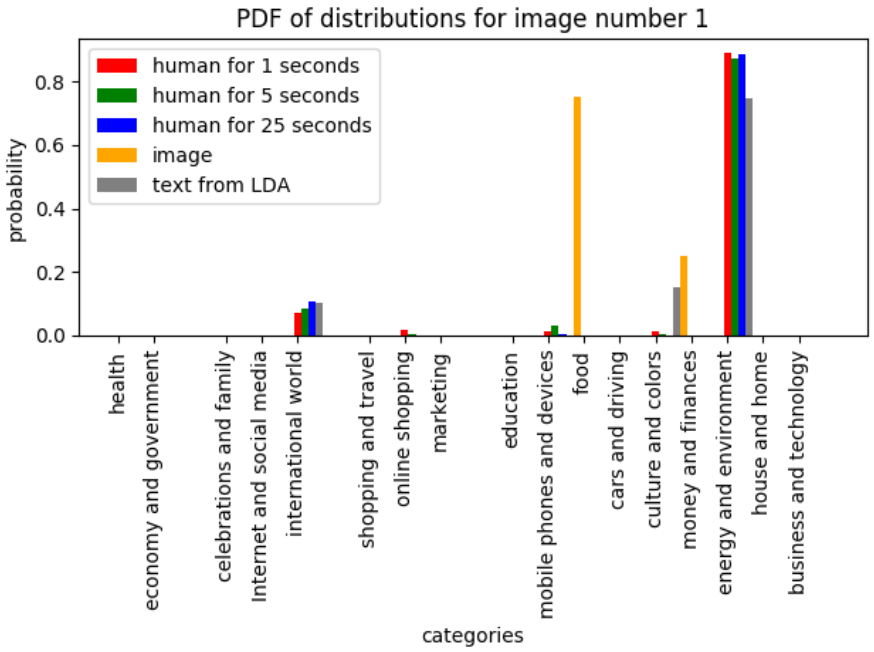


Figure 28. Probability Distributions for Infographic 1

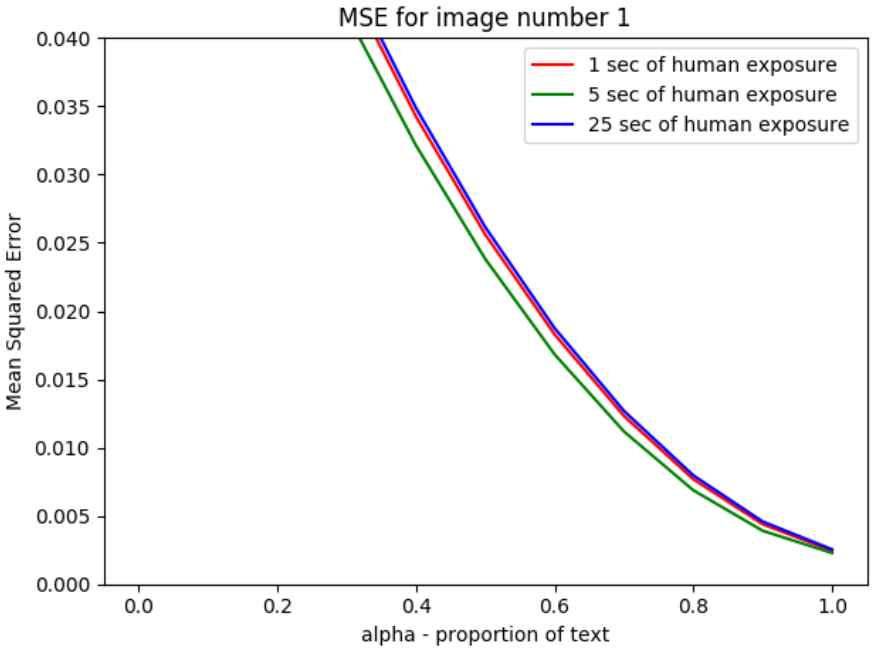


Figure 29. MSE for Infographic 1

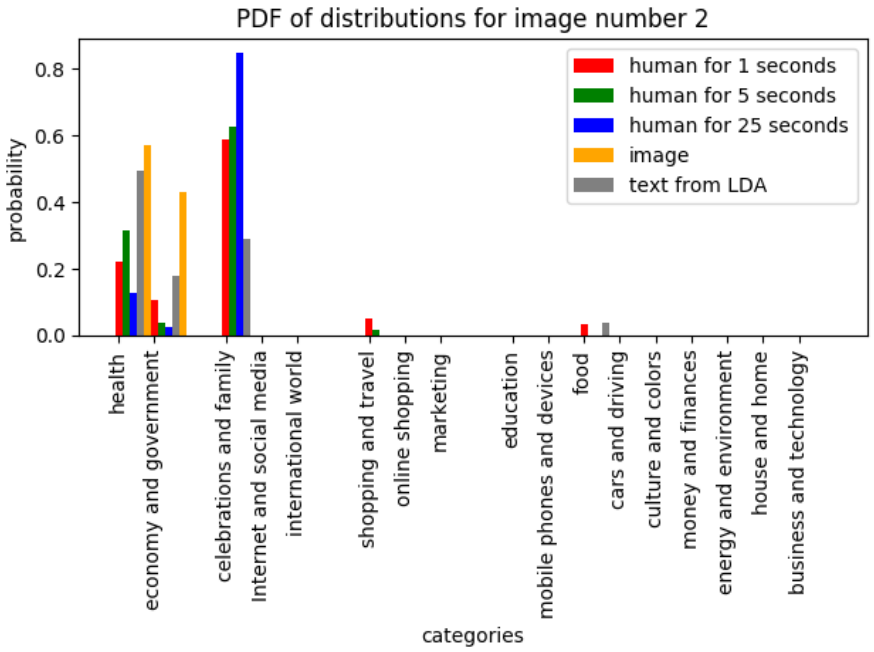


Figure 30. Probability Distributions for Infographic 2

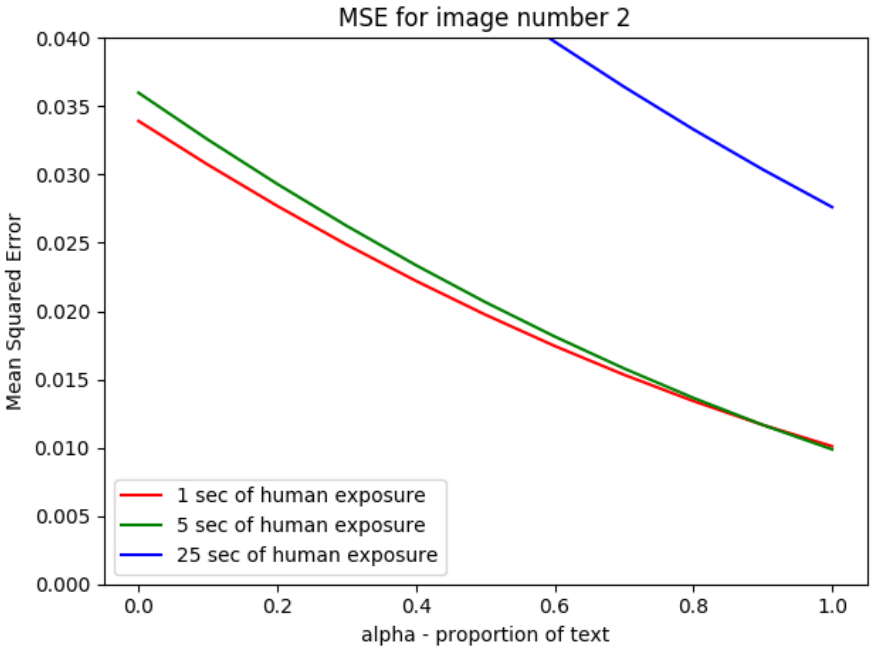


Figure 31. MSE for Infographic 2

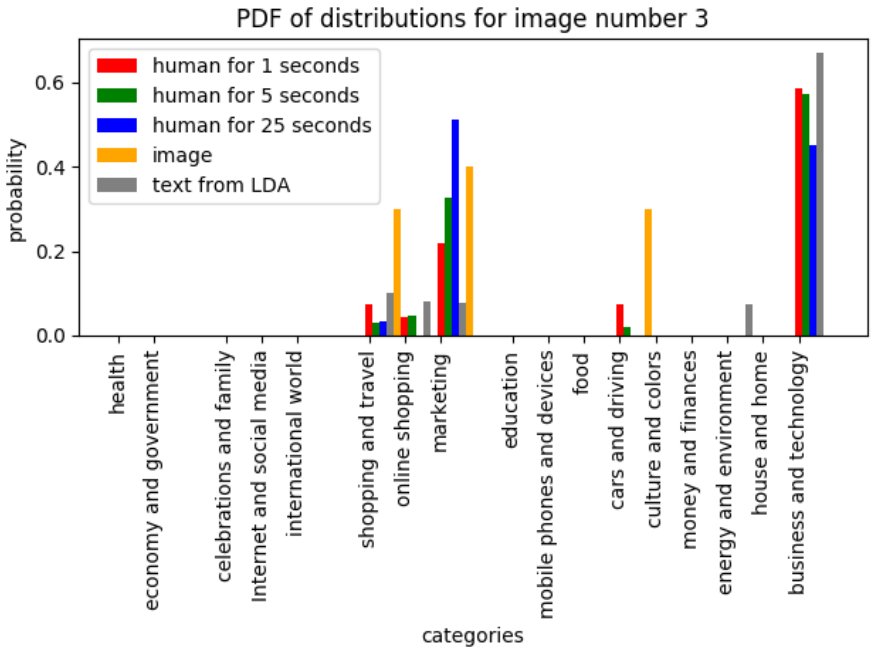


Figure 32. Probability Distributions for Infographic 3

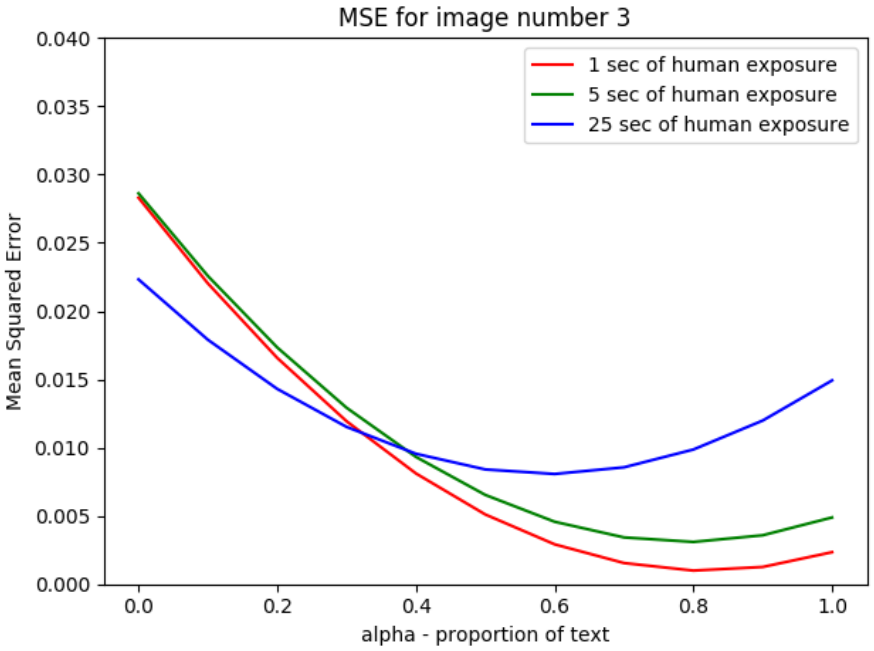


Figure 33. MSE for Infographic 3

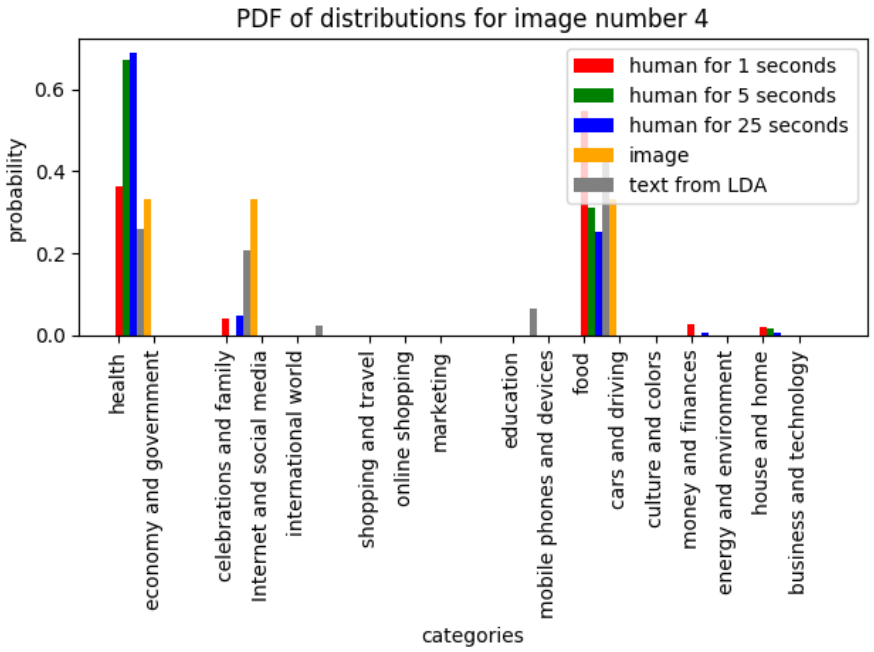


Figure 34. Probability Distributions for Infographic 4

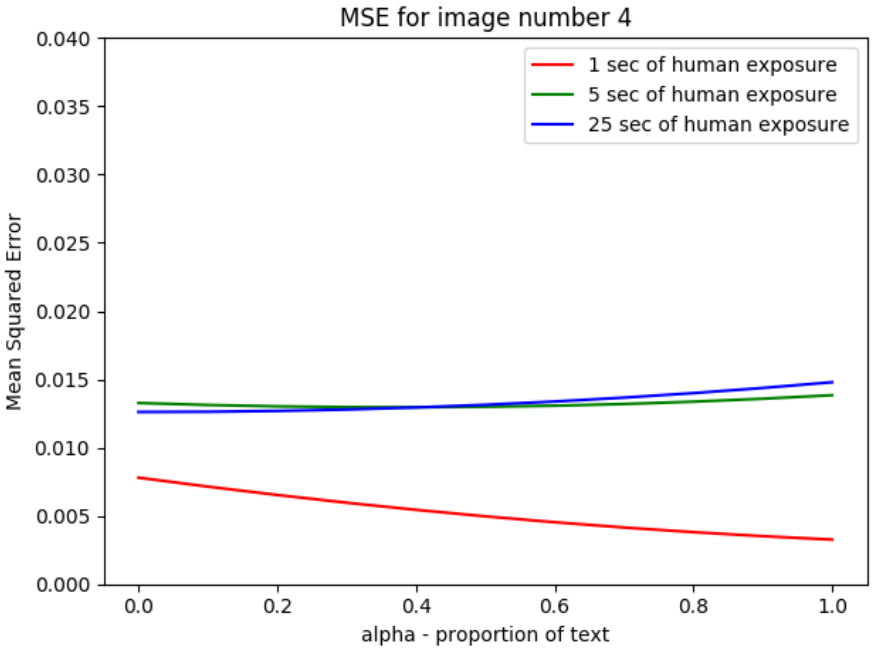


Figure 35. MSE for Infographic 4

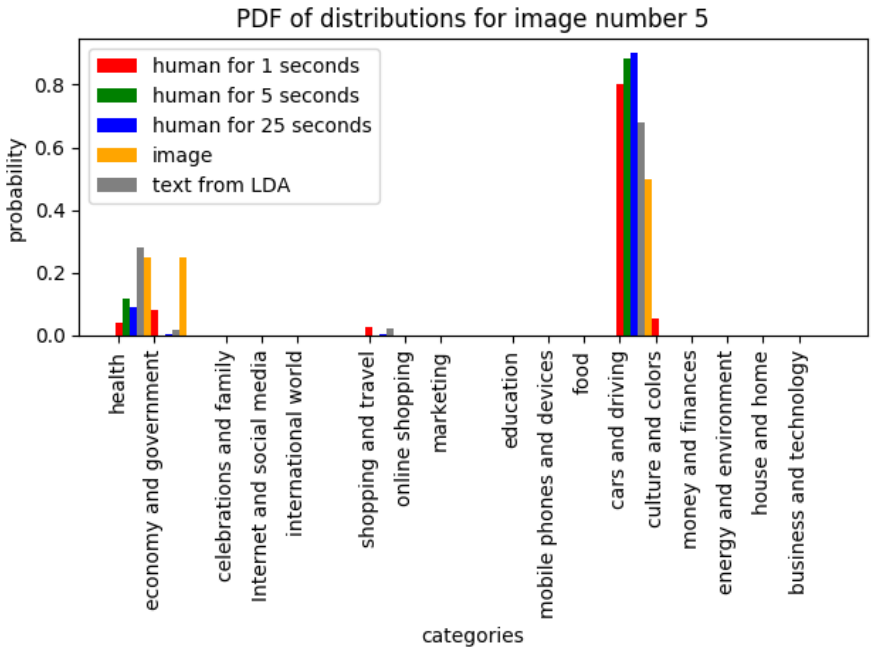


Figure 36. Probability Distributions for Infographic 5

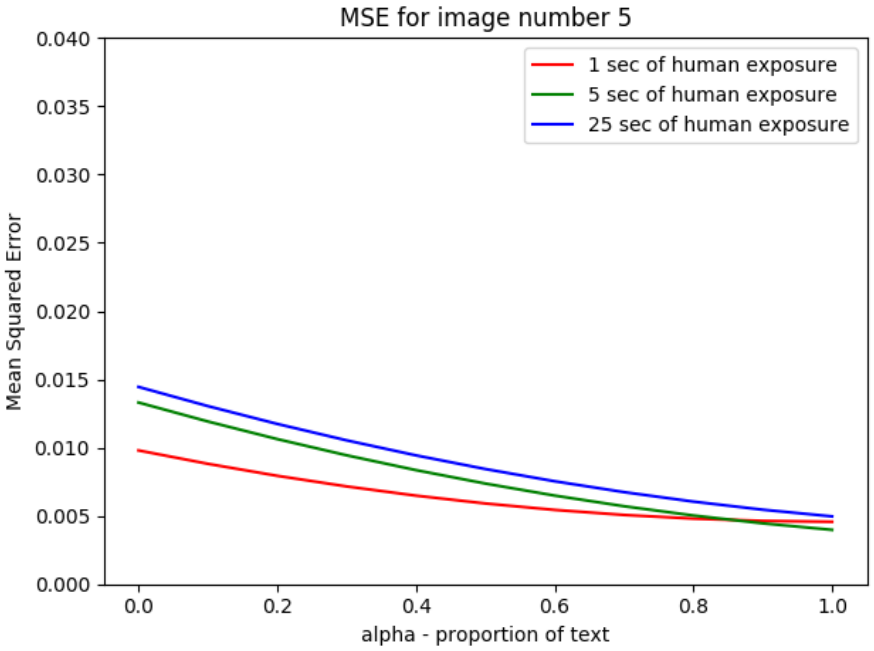


Figure 37. MSE for Infographic 5

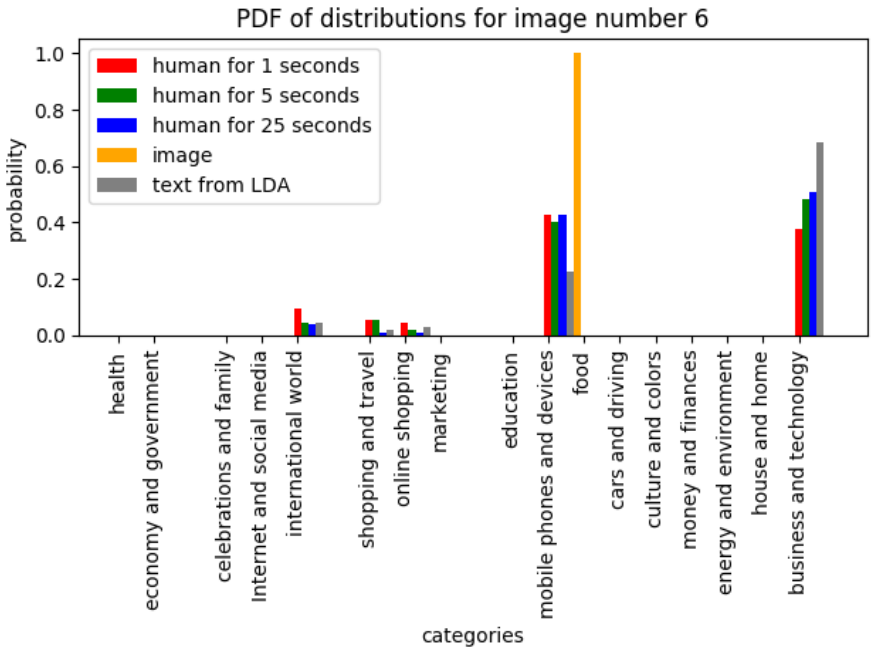


Figure 38. Probability Distributions for Infographic 6

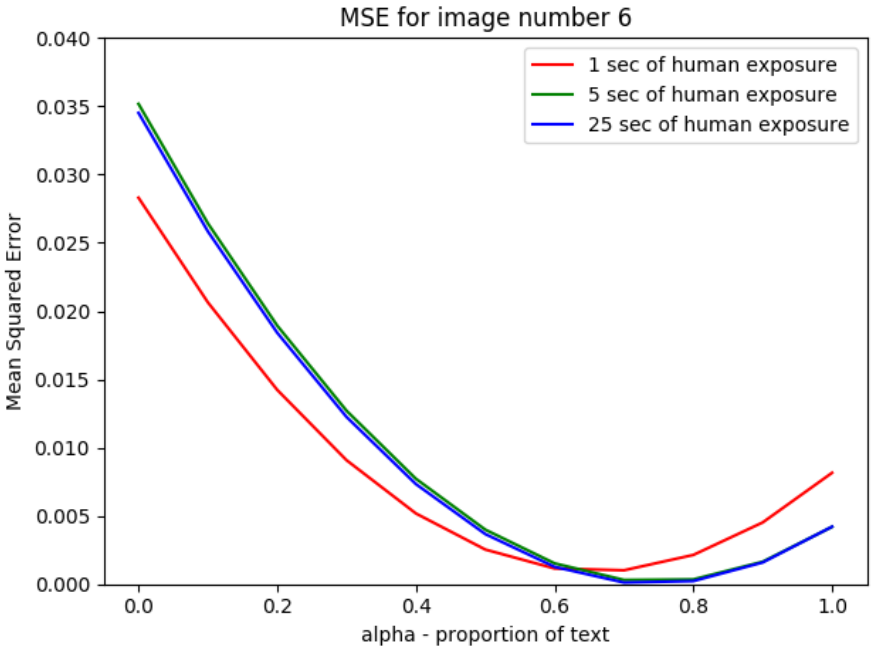


Figure 39. MSE for Infographic 6

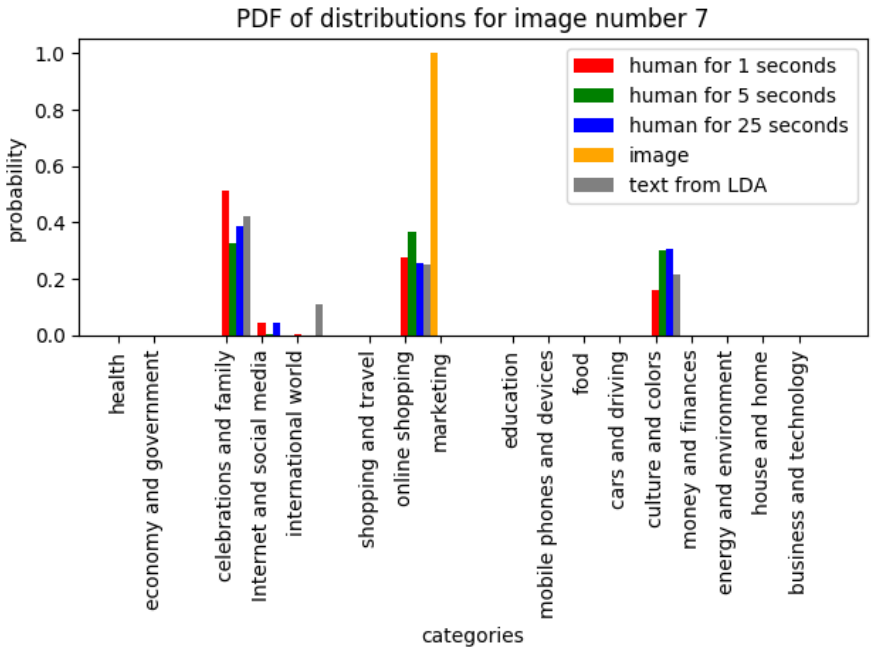


Figure 40. Probability Distributions for Infographic 7

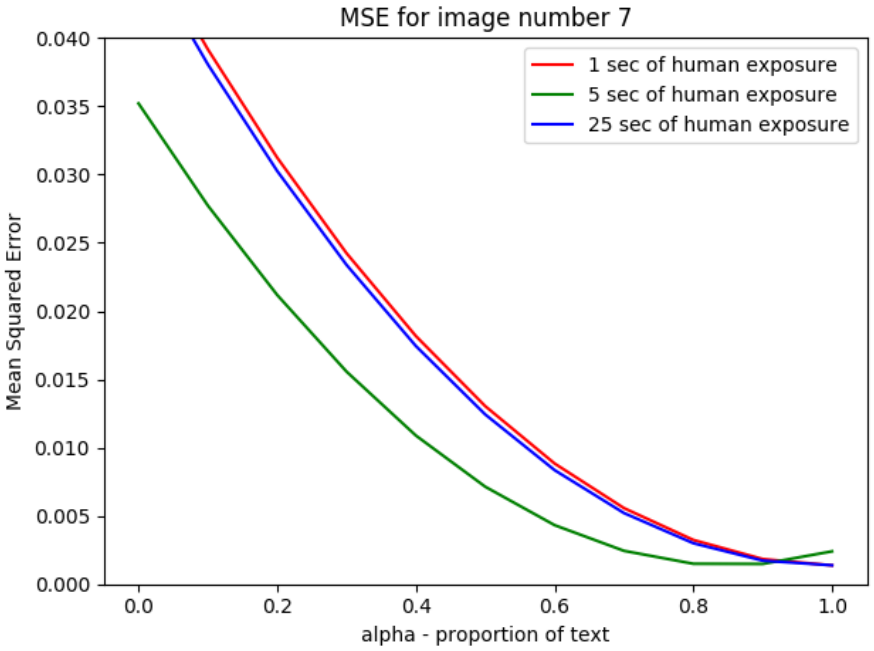


Figure 41. MSE for Infographic 7

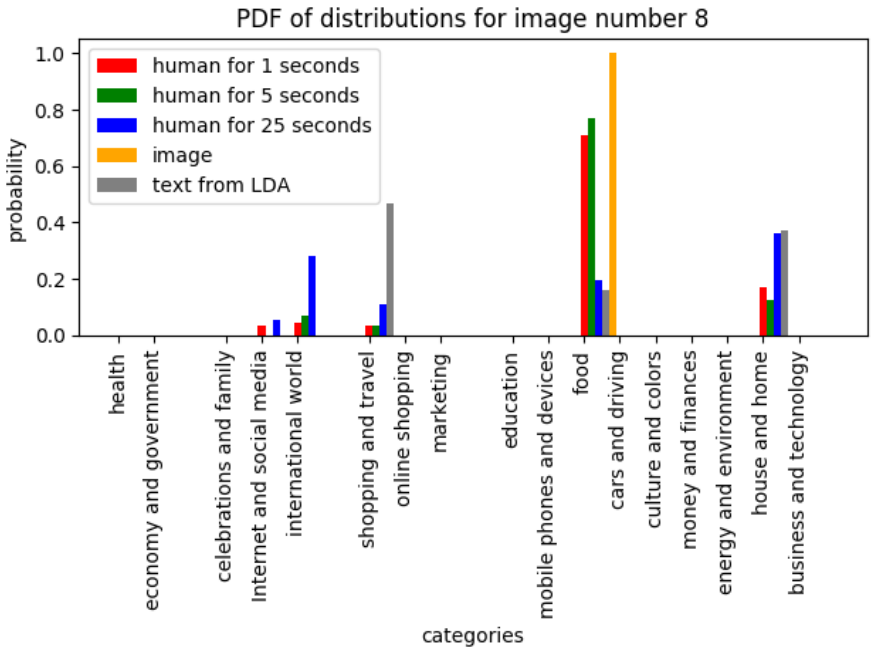


Figure 42. Probability Distributions for Infographic 8

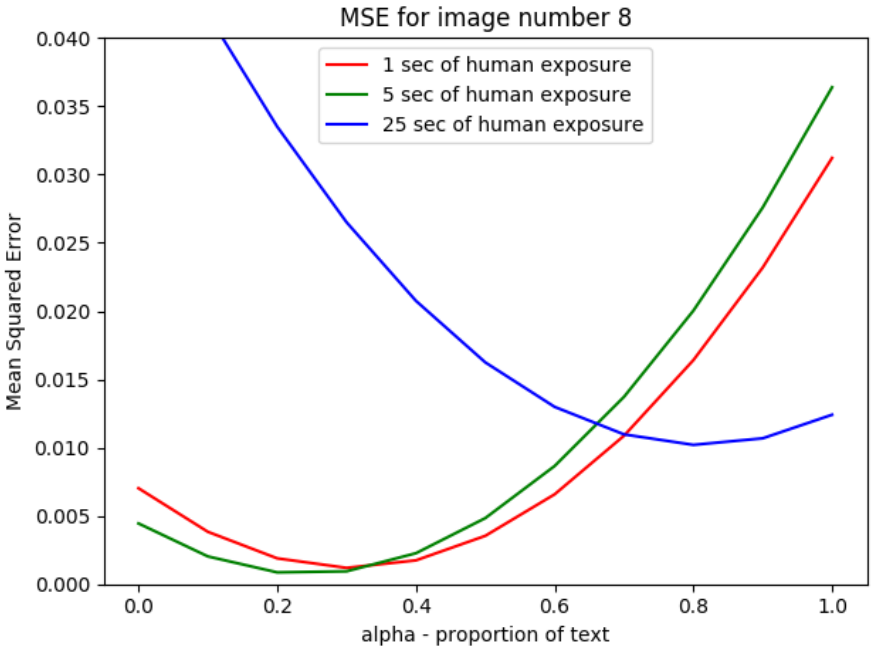


Figure 43. MSE for Infographic 8

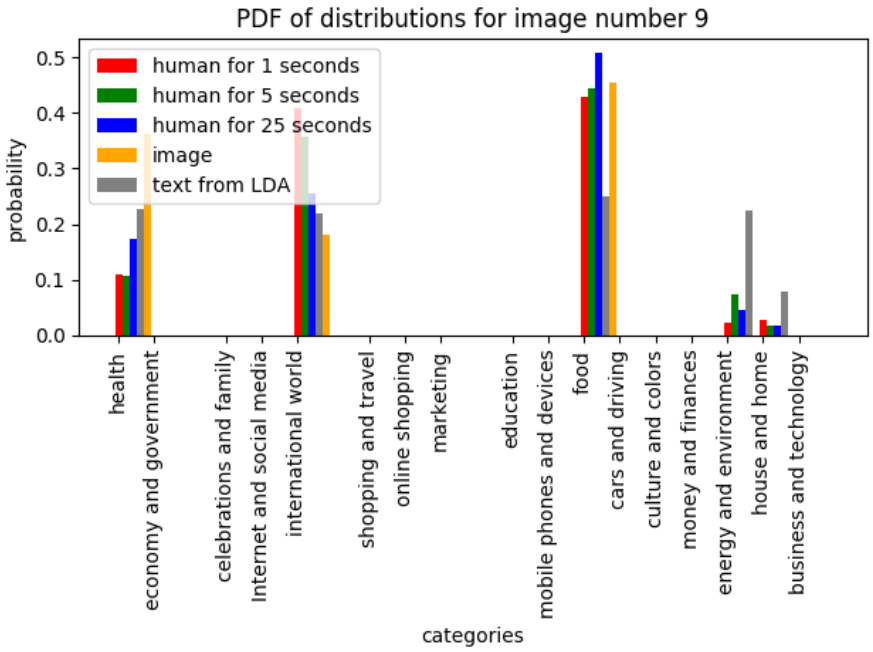


Figure 44. Probability Distributions for Infographic 9

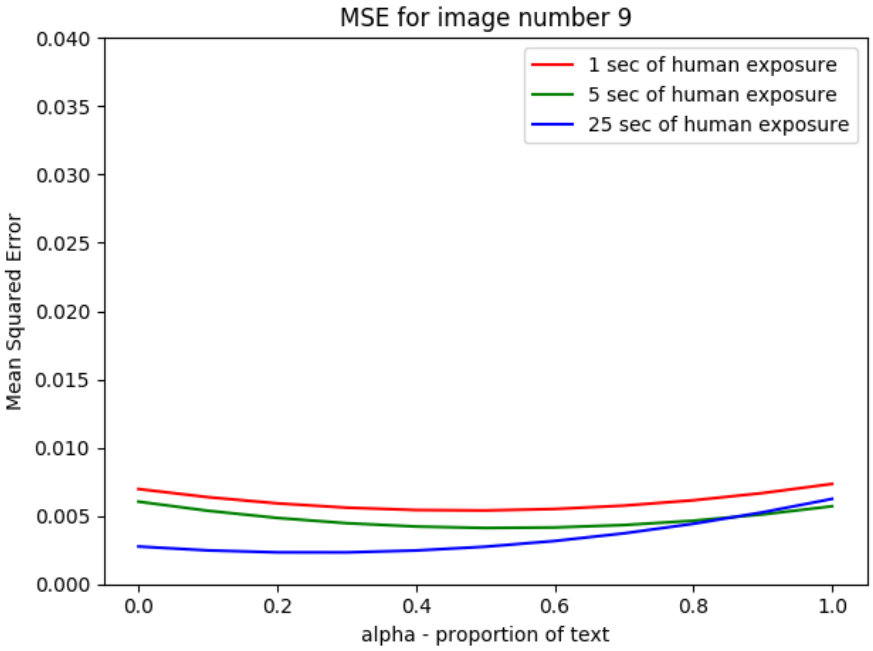


Figure 45. MSE for Infographic 9

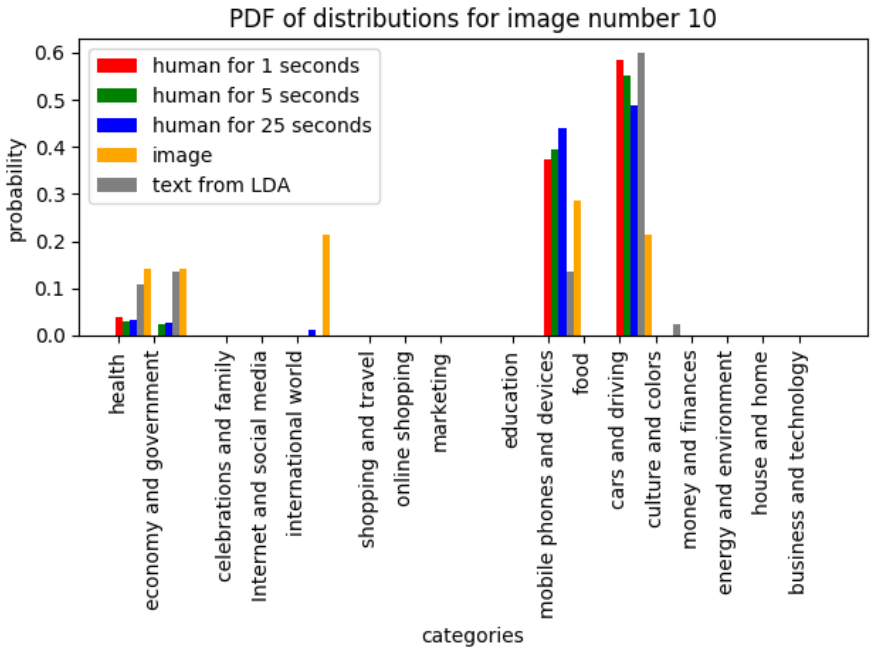


Figure 46. Probability Distributions for Infographic 10

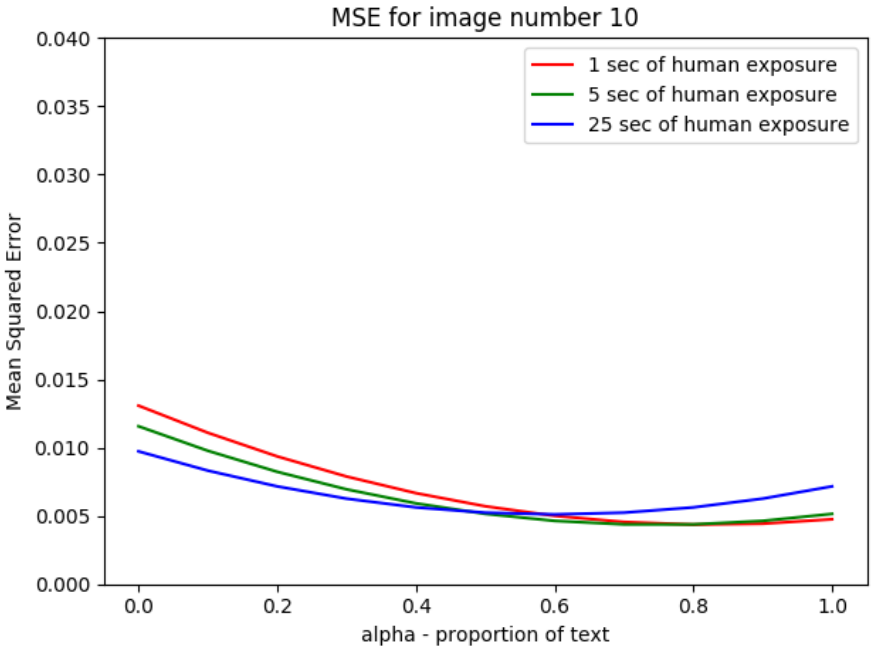


Figure 47. MSE for Infographic 10

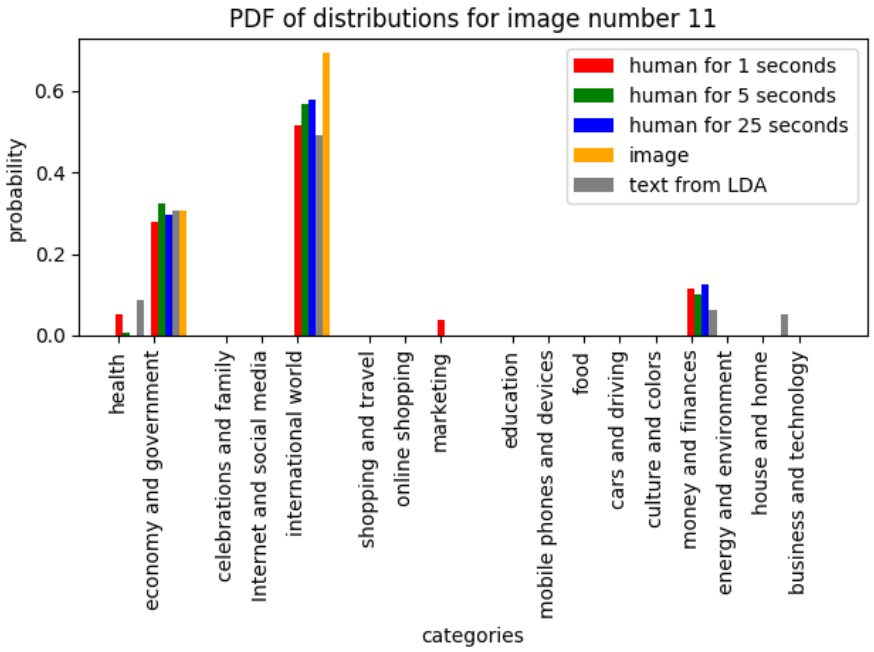


Figure 48. Probability Distributions for Infographic 11

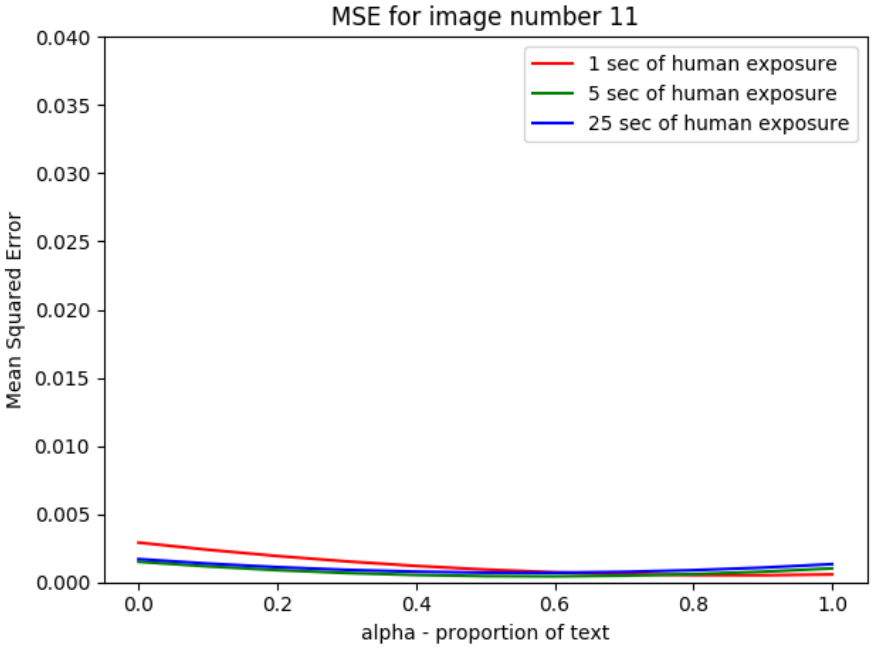


Figure 49. MSE for Infographic 11