Final Project Submission

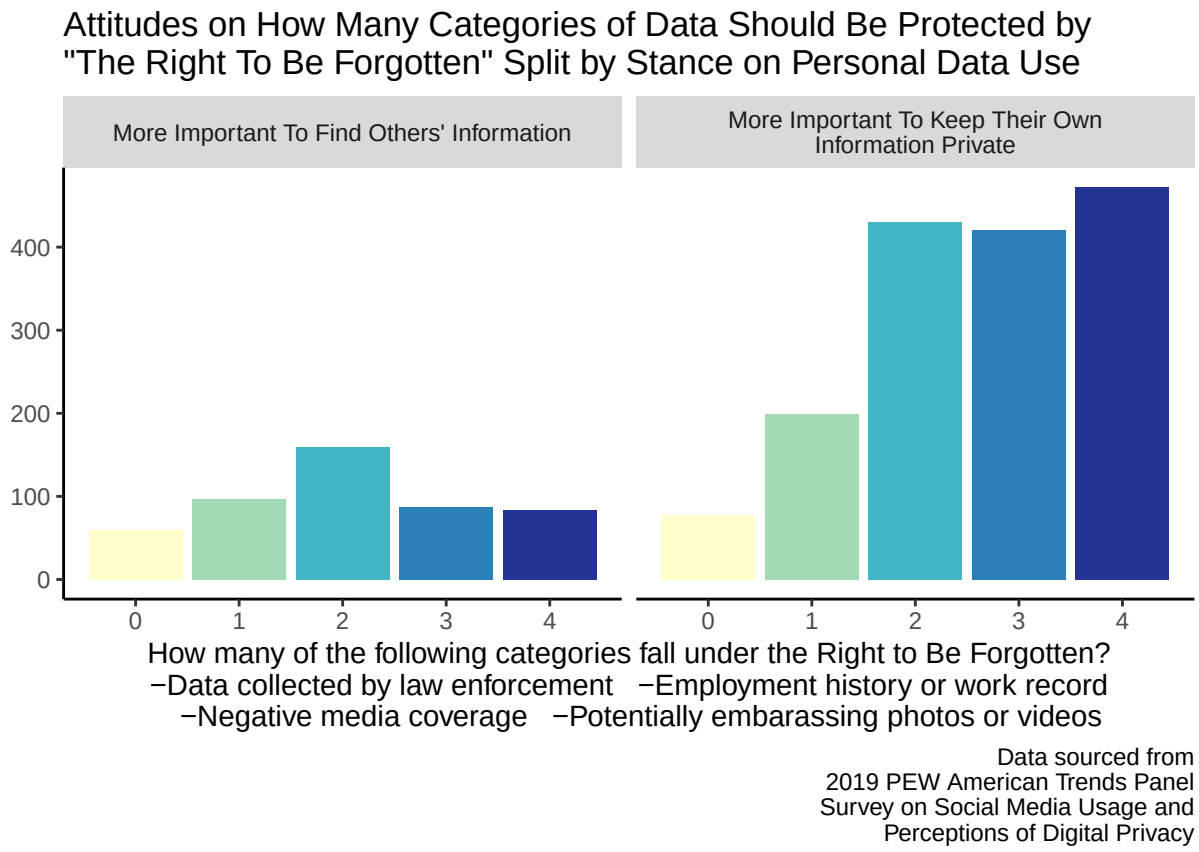
Jake Chanenson and Adriana Knight

5/12/2021

Introduction

Our physical and digital lives have been increasingly entwined during the first two decades of the twenty first century as we enter into the digital age. A direct consequence of this development is the advent of what social psychologist Shoshana Zuboff calls "Surveillance Capitalism" – an economic model where a user's personal data is commodified for profit generation. The commodification of personal data comes at the cost of user privacy as companies hoover up any and all quantifiable metrics about a given user in an effort to serve them more ads and services that are tailored to them. As such, it is now more important than ever to understand how the American people feel about issues surrounding the use of their data and their privacy. We chose two aspects of digital privacy to focus on: public sentiment surrounding the Right to Be Forgotten (RTBF) – a privilege for people to remove photos, documents, and videos of themselves from public internet search – and the privacy threat of Facial Recognition Technologies.

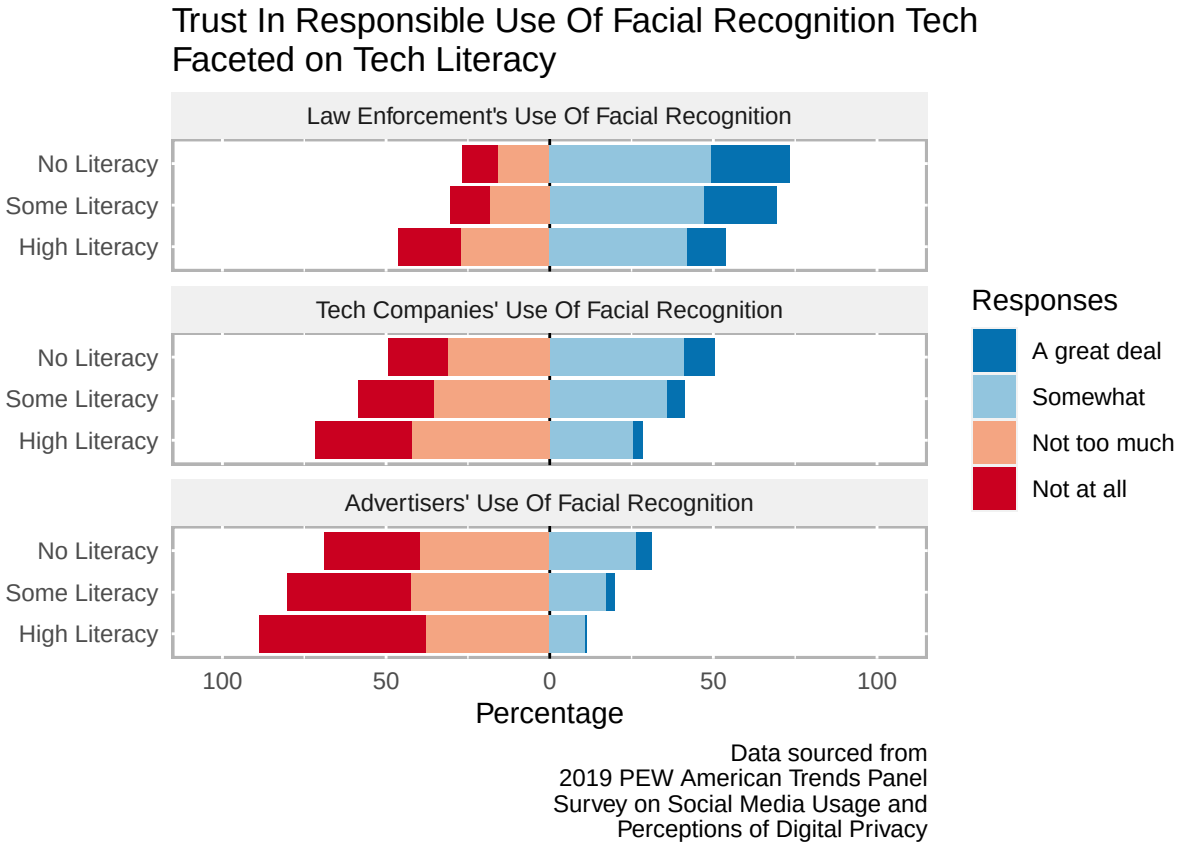
Key Figure 1



Explanation of Key Figure 1

BLAH BLAH BLAH BLAH..... BLAH BLAH BLAH BLAH BLAH

Key Figure 2



Explanation of Key Figure 2

BLAH BLAH BLAH BLAH..... BLAH BLAH BLAH BLAH BLAH

Appdendix

Description of Dataset

The dataset we used is the 2019 PEW Survey on American Trends on Technology. The full citation for this data set is as follows:

Pew Research Center. (2019). American Trends Panel (W49). Retrieved from https://www.dropbox.com/sh/adyrtaju2jd7a2d/AAC2fmHoYs2SwVYKqCIkTxOsa?dl=0

The data set is a sample of 4272 respondents of people 18 years or older living in the US. This sample consisted of both English and Spanish-language survey takers. The methodology also details the stratified sampling. I have copied the statement from the methodology section below:

> The ATP subsample was selected by grouping panelists into seven stratum 1. Non-internet panelists. There were 691 total panelists in this stratum and they are sampled at a rate of 100% 2. HS or less panelists. There were 2,027 total panelists in this stratum and they are sampled at a rate of 98.9%. 2,005 panelists were selected for Wave 49. 3. Hispanic, Unregistered or Non-volunteers. There were 5,312 total panelists in this stratum and they are sampled at a rate of 44.8%. 2,380 panelists were selected for Wave 49. 4. Black or 18-34 panelists. There were 1,253 total panelists and they are sampled at a rate of 18.2%. 228 panelists were selected for Wave 49. 5. Other panelists. There were 4,176 total panelists and they are sampled at a rate of 13.5%. 564 panelists were selected for Wave 49.

Variables Used and Their Questions Phrasing Key Figure 1

- `PUBLICDATA_W49` - "Today a wide range of information about people is searchable online. Do you think it is more important for people to have the ability to. . . "
- `RTBF` - "Do you think that ALL Americans should have the right to have the following information about themselves removed from public online search results?"
  - `RTBFa_W49` - "Data collected by law enforcement, such as criminal records or mugshots"
  - `RTBFb_W49` - "Information about their employment history or work record"
  - `RTBFc_W49` - "Negative media coverage"
  - `RTBFd_W49` - "Potentially embarrassing photos or videos"

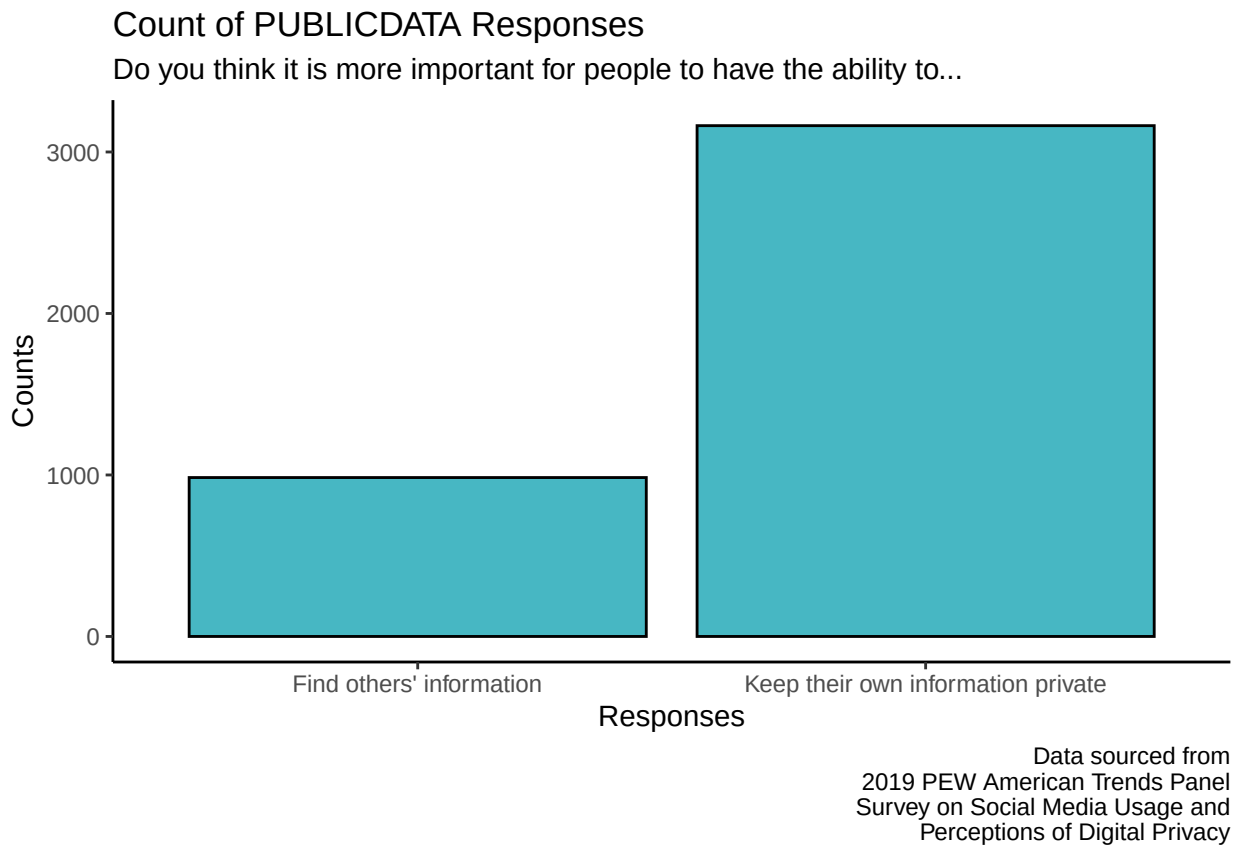
Key Figure 2

- `FACE1` - "How much have you heard or read about the development of automated facial recognition technology that can identify someone based on a picture or video that includes their face?
- `FACE2` - "Based on what you know, how effective do you think facial recognition technology is at the following things?"
  - `FACE2a_W49` - "Accurately identifying individual people"
  - `FACE2b_W49` - "Accurately assessing someone's gender"
  - `FACE2c_W49` - "Accurately assessing someone's race"
- `FACE3` - "How much, if at all, do you trust the following groups to use facial recognition technology responsibly?"
  - `FACE3a_W49` - "Advertisers"
  - `FACE3b_W49` - "Technology companies"
  - `FACE3c_W49` - "Law enforcement agencies"
- `FACE4` - "In your opinion, is it acceptable or unacceptable to use facial recognition technology in the following situations?"

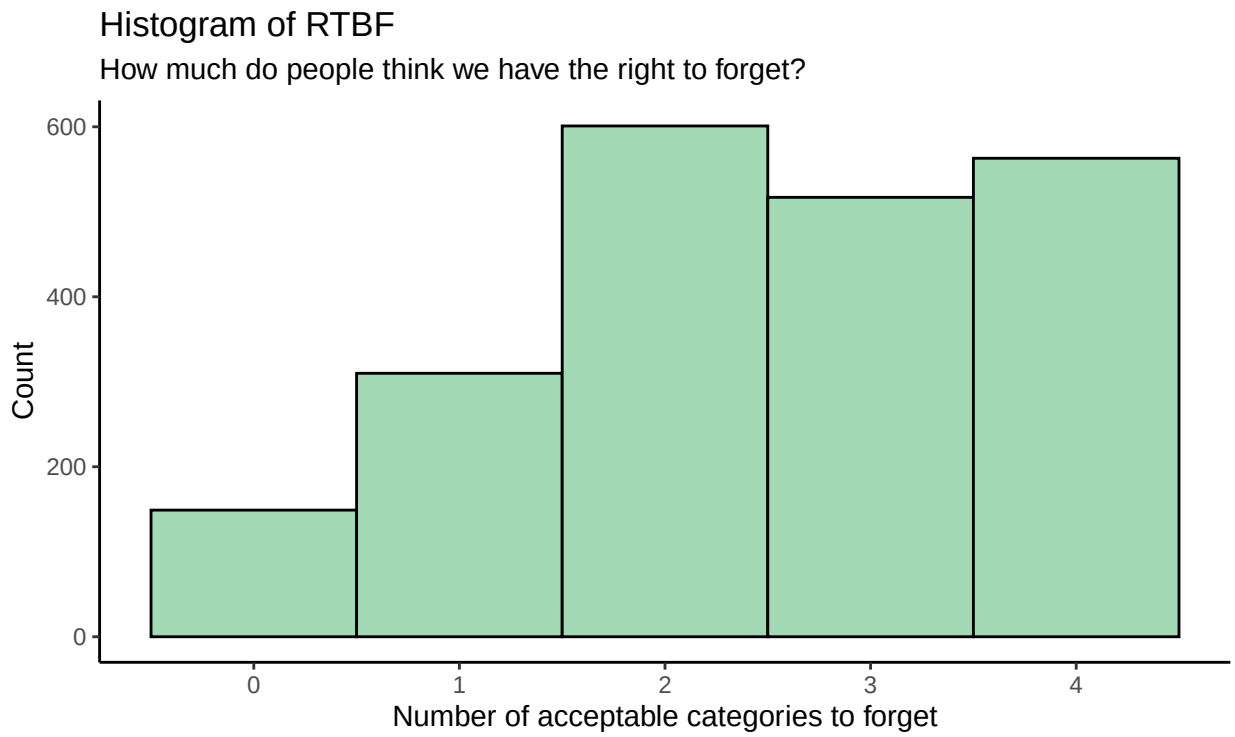
  - `FACE4a_W49` - “Law enforcement agencies assessing potential security threats in public spaces”
  - `FACE4b_W49` - “Companies automatically tracking the attendance of their employees”
  - `FACE4c_W49` - “Advertisers seeing how people respond to public advertising displays”
  - `FACE4d_W49` - “Apartment building landlords tracking who enters or leaves their buildings”
- `KNOW1_W49` - "If a website uses cookies, it means that the site...
- `KNOW3_W49` - “When a website has a privacy policy, it means that the site...”
- `KNOW4_W49` - “What does it mean when a website has ‘https://’ at the beginning of its URL, as opposed to ‘http://’ without the ‘s’?”
- `KNOW7_W49` - “The term ‘net neutrality’ describes the principle that...”
- `KNOW8_W49` - “Many web browsers offer a feature known as ‘private browsing’ or ‘incognito mode.’ If someone opens a webpage on their computer at work using incognito mode, which of the following groups will NOT be able to see their online activities?”
- `KNOW9_W49` - “Some websites and online services use a security process known as two-step or two-factor authentication. Which of the following images is an example of two-factor authentication?”
- `CONCERNCO_W49` - “How concerned are you, if at all, about how companies are using the data they collect about you?”

Additional Figures For Key Figure 1

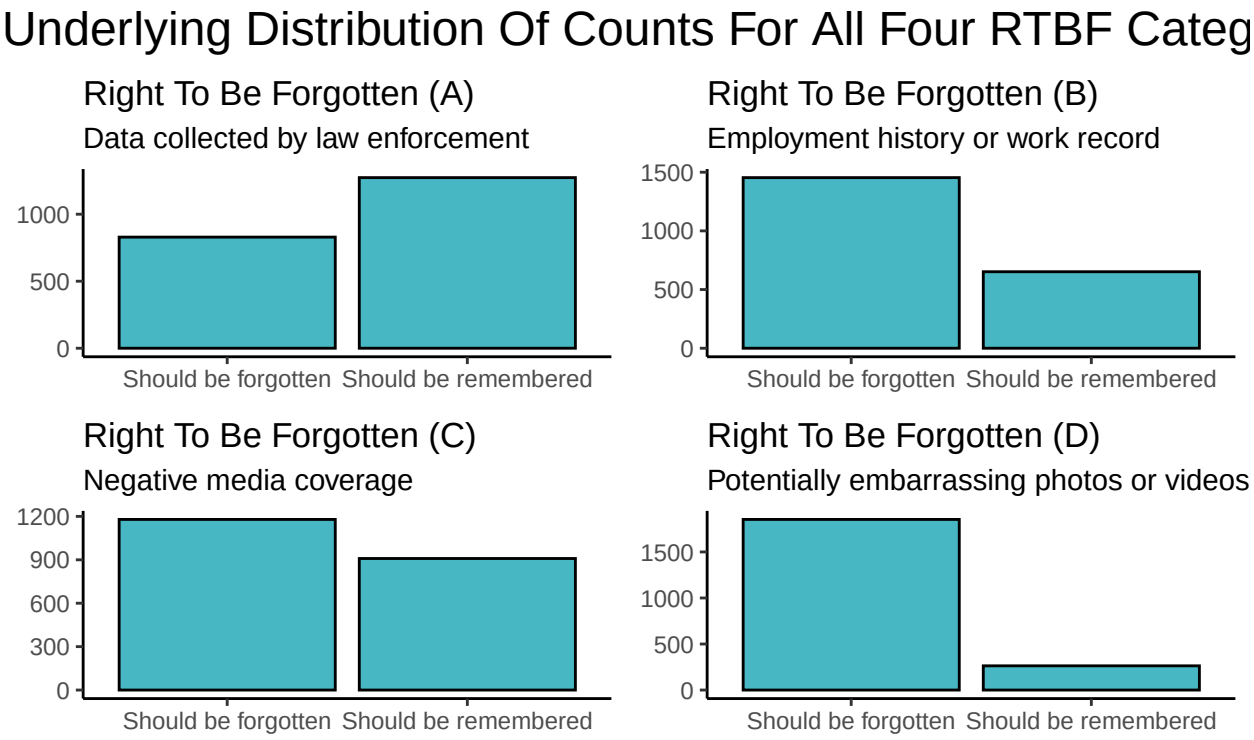
1. Count Of PUBLICDATA Responses



2. How Much Data Do People Think We Have The Right To Forget



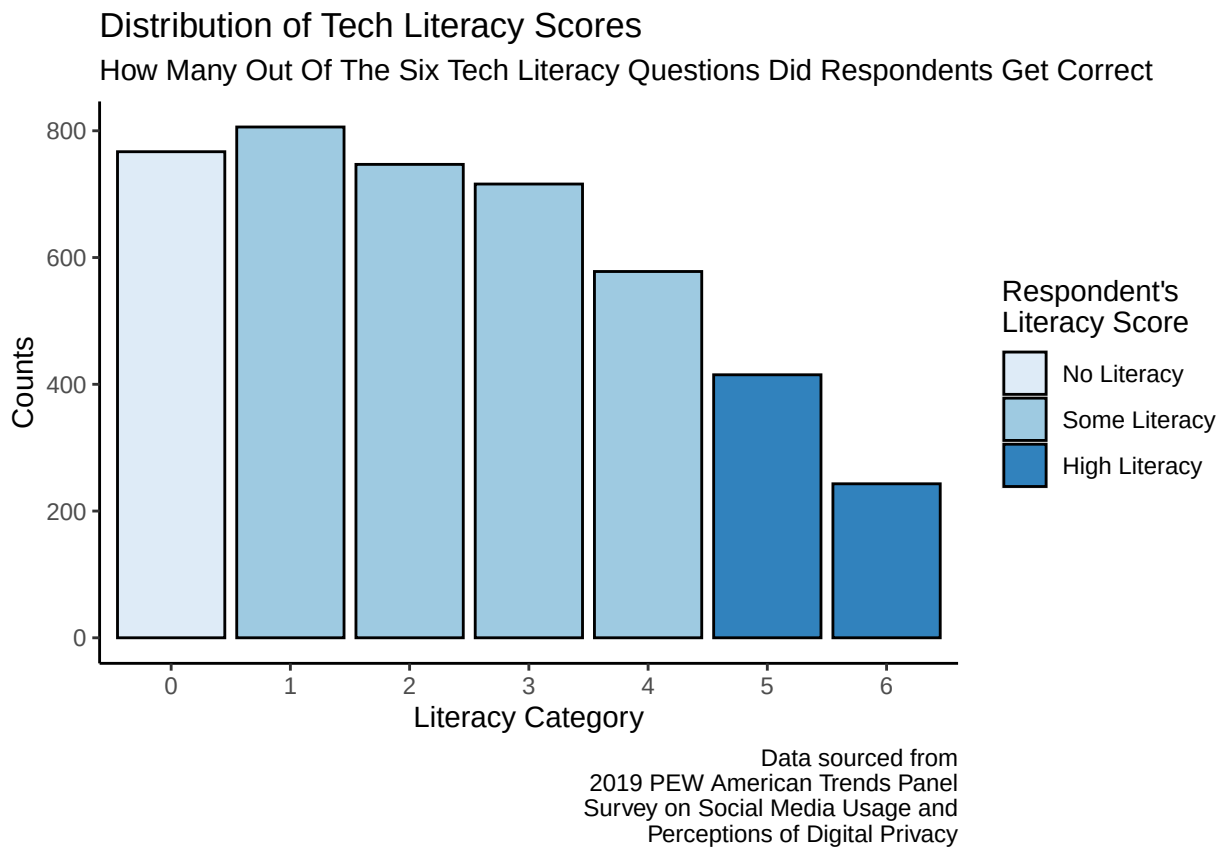
3. Underlying Distribution Of Counts For All Four RTBF Categories



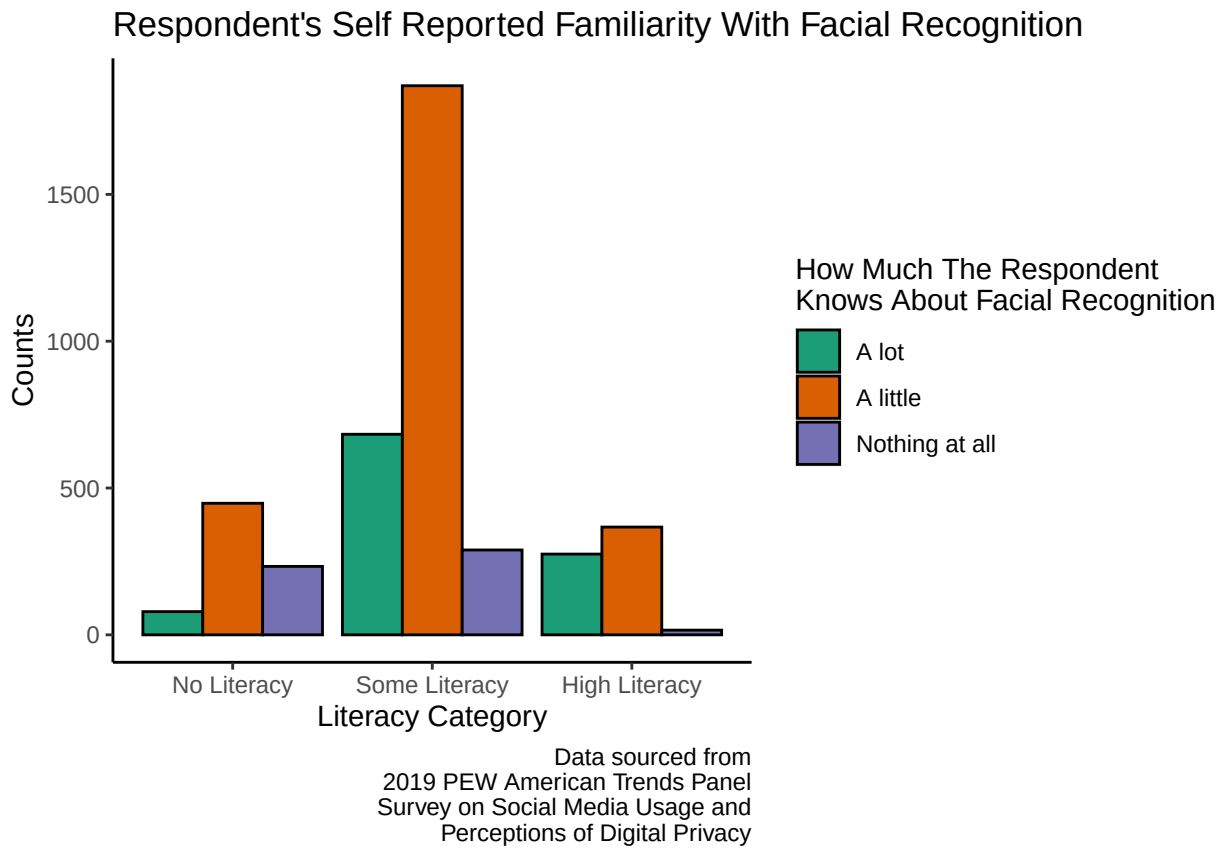
Data sourced from
2019 PEW American Trends Panel
Survey on Social Media Usage and
Perceptions of Digital Privacy

Additional Figures For Key Figure 2

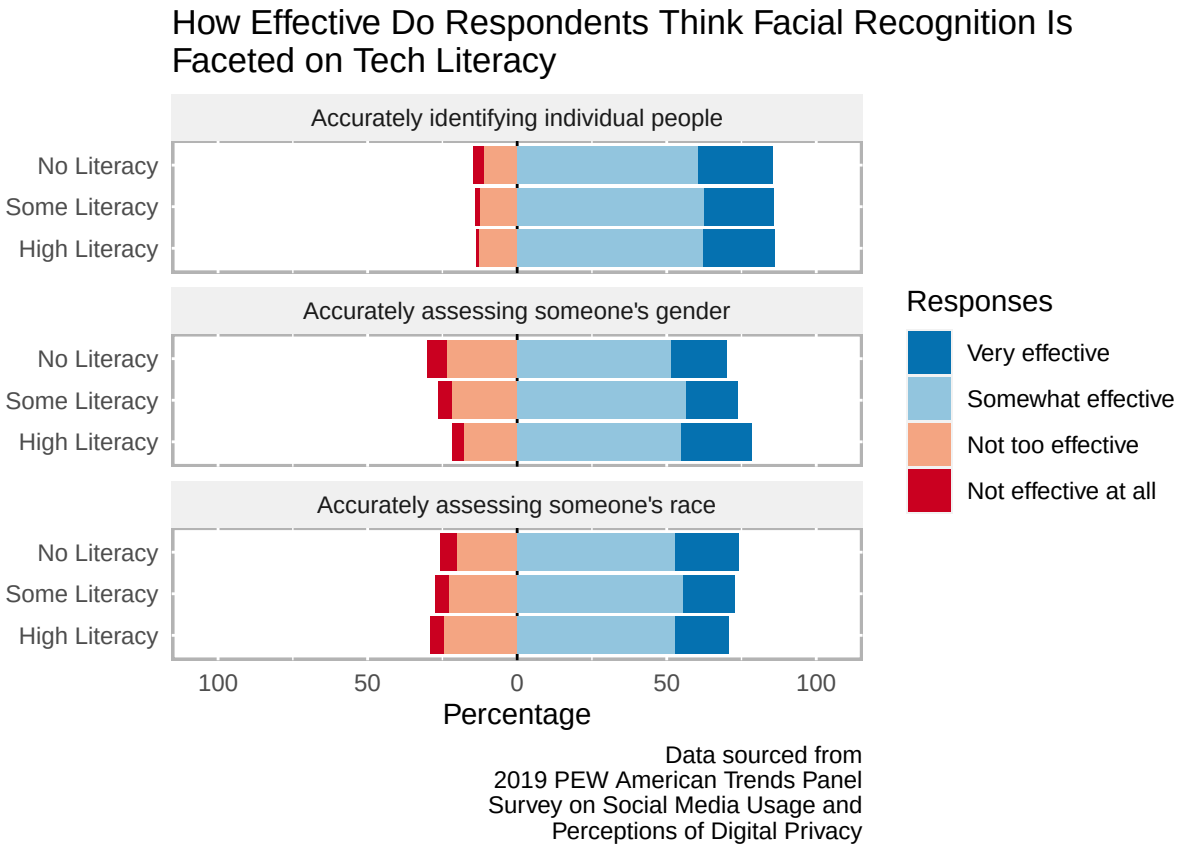
1. Underlying Distribution of Tech Literacy

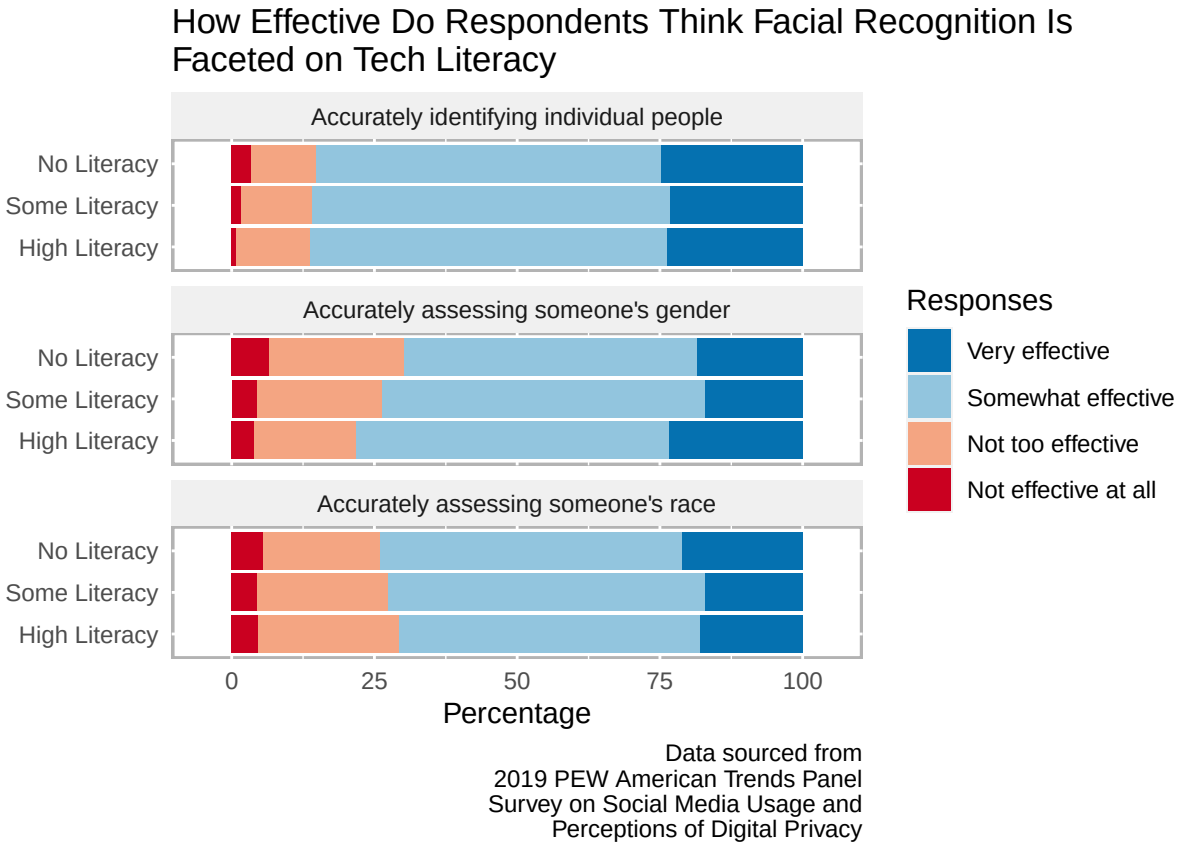


2. Self Reported Facial Recognition Familiarity

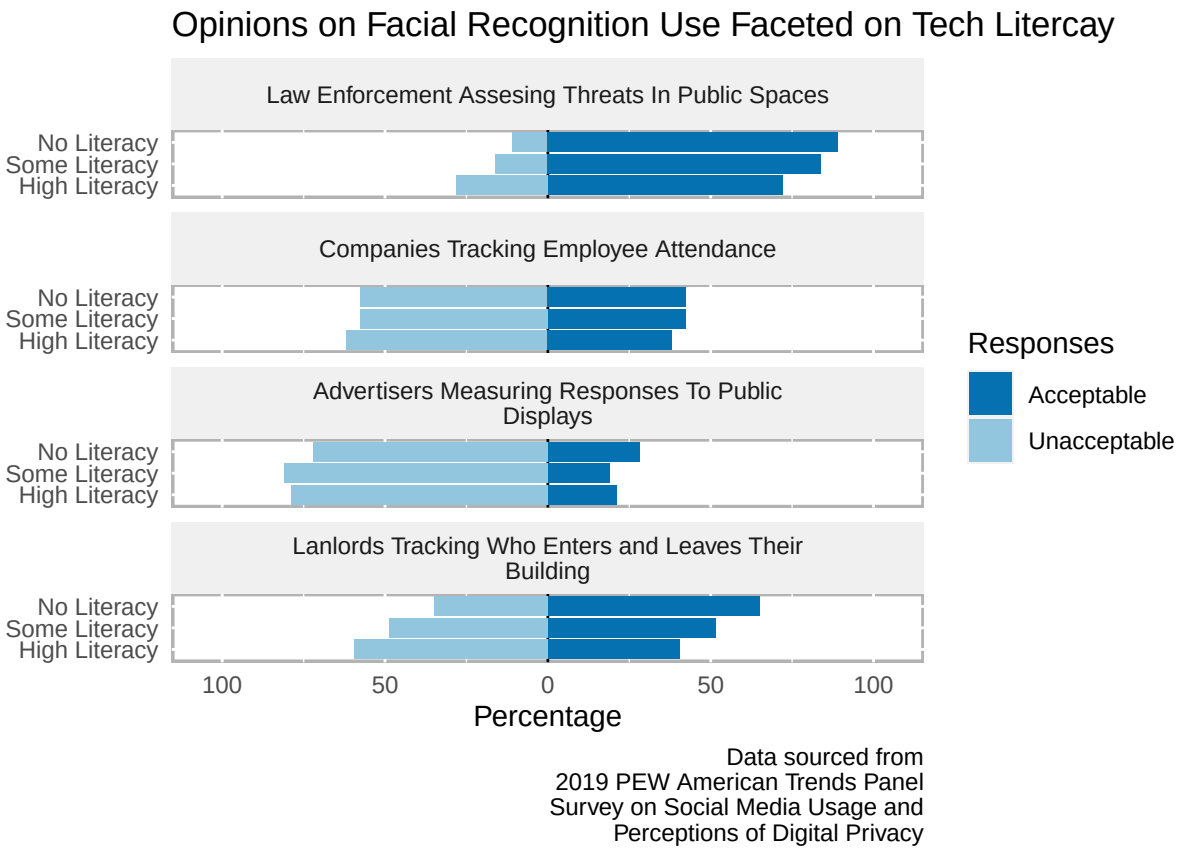


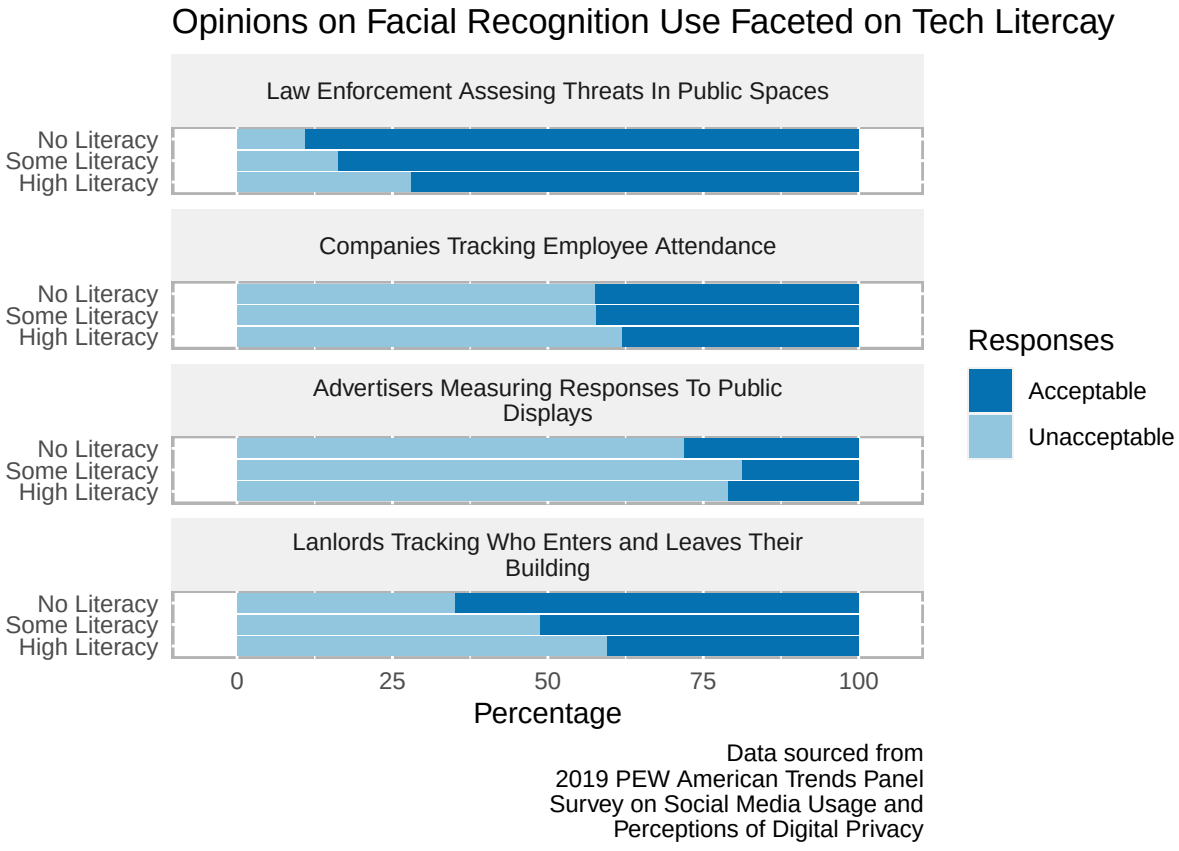
3. Respondent's Understanding of Facial Recognition





4. What Situations Respondents Think It Is Acceptable To Use Facial Recognition





Additional Tables For Key Figure 1

Table 1: Summary Statistics For Key Figure 1

sumRTBF	Stance On RTBF	Freq
0	Data Should Be Private	89
1	Data Should Be Private	213
2	Data Should Be Private	442
3	Data Should Be Private	430
4	Data Should Be Private	479
0	Data Should Be Public	60
1	Data Should Be Public	97
2	Data Should Be Public	159
3	Data Should Be Public	87
4	Data Should Be Public	84

Table 2: Numeric Summary That Involves Uncertainty For Key Figure 1

publicDataNum	mean	sd
0	2.603146	1.183437
1	2.078029	1.248072

Additional Tables For Key Figure 2

Table 3: Summary Statistics For Key Figure 2

Group	Item	low	high	mean
High Literacy	Law Enforcement's Use Of Facial Recognition	46.32238	53.67762	2.461659
High Literacy	Tech Companies' Use Of Facial Recognition	71.51800	28.48200	2.018779
High Literacy	Advertisers' Use Of Facial Recognition	88.73239	11.26761	1.607199
Some Literacy	Law Enforcement's Use Of Facial Recognition	30.48829	69.51171	2.797539
Some Literacy	Tech Companies' Use Of Facial Recognition	58.59468	41.40532	2.237396
Some Literacy	Advertisers' Use Of Facial Recognition	80.26995	19.73005	1.843192
No Literacy	Law Enforcement's Use Of Facial Recognition	26.79612	73.20388	2.862136
No Literacy	Tech Companies' Use Of Facial Recognition	49.51456	50.48544	2.415534
No Literacy	Advertisers' Use Of Facial Recognition	68.73786	31.26214	2.069903

Table 4: Numeric Summary That Involves Uncertainty For Key Figure 2

Group	Item	mean	sd
High Literacy	Law Enforcement's Use Of Facial Recognition	2.461659	0.9325284
High Literacy	Tech Companies' Use Of Facial Recognition	2.018779	0.8149991
High Literacy	Advertisers' Use Of Facial Recognition	1.607199	0.6952671
Some Literacy	Law Enforcement's Use Of Facial Recognition	2.797539	0.9198941
Some Literacy	Tech Companies' Use Of Facial Recognition	2.237396	0.8672941
Some Literacy	Advertisers' Use Of Facial Recognition	1.843192	0.7908986
No Literacy	Law Enforcement's Use Of Facial Recognition	2.862136	0.9030985

Group	Item	mean	sd
No Literacy	Tech Companies' Use Of Facial Recognition	2.415534	0.8921139
No Literacy	Advertisers' Use Of Facial Recognition	2.069903	0.8637574

Discussion of Uncertainty & Inference

Key Figure 1

By grouping the data by `PUBLICDATA_W49` results and running an ANOVA test on each group's `sumRTBF` count variables, we are able to say with extremely high confidence (on a P-value that is extremely close to 0 with a value of `2e-16`) that there is a statistically significant difference between public and private data oriented respondents in the average number of categories of data they deem worthy of falling under the Right to Be Forgotten.

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## PUBLICDATA_W49    2  143.1   71.57   50.46 <2e-16 ***
## Residuals      2137 3031.3    1.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 2132 observations deleted due to missingness
```

I would like to introduce one small caveat that does not delegitimize these findings, but does complicate them slightly. Due to the nature of the dummy variable `sumRTBF`, we may know how many categories, on average, respondents identify as being worthy of protection. However, we do not know *which* categories the respondents select. There are sixteen[1] possible ways a respondent could answer the four Right to Be Forgotten questions. We know with certainty that participating respondents demonstrated trends in how many categories chosen, but we do *not* know with certainty if there are trends in *which* categories respondents chose.

Key Figure 2

Using a 1-way ANOVA test on each of the three facets in key figure 2 – Law Enforcement's Use Of Facial Recognition, Tech Companies' Use Of Facial Recognition, and Advertisers' Use Of Facial Recognition – we are able to say with extremely high confidence (on a P-value that is extremely close to 0 with a value of `1.66e-14`, `5.37e-11`, and `<2e-16` respectively) that there is a statistically significant difference in the responses of those categorized as No Literacy Some Literacy High Literacy in all three facets!

ANOVA Summary For Law Enforcement's Use Of Facial Recognition

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## groupKnow      2     58  29.167      32 1.66e-14 ***
## Residuals   3719   3389   0.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 550 observations deleted due to missingness
```

ANOVA Summary For Tech Companies' Use Of Facial Recognition

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## groupKnow      2   37.5  18.763    23.8 5.37e-11 ***
## Residuals   3719 2932.1   0.788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 550 observations deleted due to missingness
```

ANOVA Summary For Advertisers' Use Of Facial Recognition

[1] calculated using combinatorics: `(4 choose 0) + (4 choose 1) + (4 choose 2) + (4 choose 3) + (4 choose 4) = 16`

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## groupKnow      2   52.9  26.462   40.89 <2e-16 ***
## Residuals   3719 2406.9   0.647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 550 observations deleted due to missingness
```

Discussion of Analytic Choices

Our analytical choices, while earnestly centered around most accurately representing the "truth," are imperfect. We will proceed with an explanation and critique of our analytical choices using Pierre Bourdieu's "Public Opinion Does Not Exist" and chapters 1 and 8 of *Data Visualization: A Practical Introduction* by Kieran Healy.

Per Pierre Bourdieu's article "Public Opinion Does Not Exist," our primary critique of the analytic choices that both we and the source of our data set (the Pew Research Center) made is that we are relying heavily on the survey data yielded by the 2019 Pew American Trends Panel survey as being accurately representative of "the public."

First, to echo Bourdieu's writings, we would like to problematize our use of the idea of "the public" at all; the data we have been manipulating, and the visualizations produced therein, are by no means necessarily representative of "the public." The public, as we would like to think of it, does not exist. Per the topline we received as part of this data set, the *actual* body of people surveyed is 5,869 members of the American Trends Panel: a national, probability-based online panel of non-institutionalized adults living in households in the United States, including Alaska and Hawaii. The fact that this is a pre-selected, probability-based panel highlights that the body of participants does not encompass every single person living in the United States, nor does every person living in the United States who interacts in some meaningful way with technology and digital privacy fall under those categorizations. Pew Research Center is a meticulous, thorough group known for producing robust data as well as analyses of said data. Though, there will always be blind spots, which we must be mindful of when discussing* what* exactly our work produced.

In that regard we are guilty, also, of using this data as a means of making some grand statement about the status of society as a means of gauging "where we are" in regard to digital privacy. Bourdieu writes that "[the public opinion poll] creates the idea that a unanimous public opinion exists in order to legitimate a policy, and strengthen the relations of the force upon which it is based or make it possible" (Bourdieu 125). Not every person has or is even *able* to produce an opinion on every single question posed by the survey; this much is proved by the sheer volume of NAs in response to certain questions. We also do not know *what* those NAs even mean; Bourdieu defines the "consensus effect" to be the tendency to overlook the proportion of respondents who gave "no reply" homogenizes the populus and ignores the political reasons for certain demographics to not respond depending on the variety of question. We do not know, truly, what those NAs mean; we made the decision to largely strike them from our visualizations to help tell a better story, but that narrative in turn excludes some body whose voices we neither know nor understand, which could in turn shift our general understanding of the field of digital privacy as a whole.

The last point of contention or concern within our analysis of our data is the constantly shifting landscape of digital literacy as a whole. We are still only a few mere decades into the digital age, meaning that there is a notable generational gap in literacy, skill sets, and general understanding of the tools offered and threats posed by the digital sphere. One of our visuals does aim to gauge technical literacy, using a dummy summary count variable, though we are cautious to take it as the unchallenged truth. For those and other variables, respondents may feign understanding of topics, or answer in line to how they think they should. As such, we can at best be only cautiously optimistic about the results our analyses yield, knowing that our variables are at best approximations, and we are representing not the hypothetical, homogenous "public" but rather a pre-selected sample group.

Moving into our *visual* analyses, we turn to Healy, particularly the ideals outlined within chapters 1 and 8. Based on those readings, we took great pain thinking about presentation and accessibility in our design. The primary sins of bad design, as Healy describes them, are bad taste, bad perception, and bad data.

We have done our best to address the source of data for analysis already above, and by no means cherry picked our data to further a narrative that we would prefer to tell over the one the data produced with our manipulations. For the first two, we have done our best to present the information in as clear and concise a way as we could manage. Leaning into ideas from chapter 8, we wanted to maintain not only high quality of visuals (no wasted ink, clear and regular visuals) but also accessibility. We used pre-built color palettes to display our data - sequential color palettes for unordered categorical data and diverging color palettes for ordered categorical data. In addition, it was important that we make sure our visuals were widely accessible by identifying color-blind friendly palates in colorbrewer via `display.brewer.all(colorblindFriendly = TRUE)` to select palettes that would work even for those with different abilities or perceptions. Our goal was to keep our color palette and our visual decisions at large consistent with the underlying data, visually pleasing, and accessible to all.

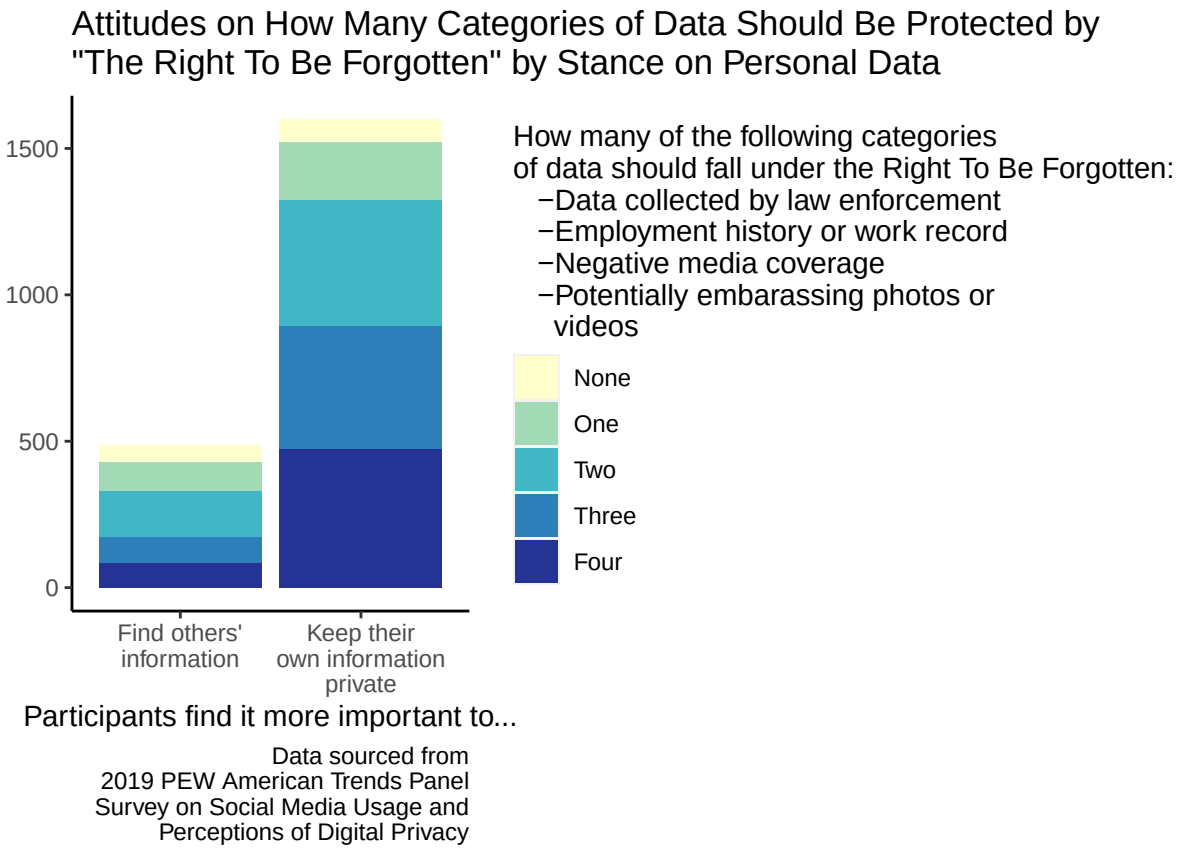
Discussion Of Other Ways We Could Have Made Our Key Figure

Key Figure 1

As I emphasized in the original text introducing my image, there were two factors that I found it extremely important to accurately represent within my visual examining respondents' stance on digital privacy and personal data: the scale of the difference between counts for the two binary options to `PUBLICDATA_W49`, and the internal distribution for each of those two categories of respondents on the issue of the Right to Be Forgotten.

I chose fairly quickly to forgo what I would think of as the "reverse" of my final visual: a faceted bar chart broken up by the four categories of information that the survey proposed as falling under the Right to Be Forgotten, showing the breakdown of each category of data's response along the lines of those who prefer public versus private data. Though this visualization would have the advantage of showing a finer-grained understanding of the Right to Be Forgotten (showing each category directly rather than using a dummy variable of sum counts, as I ultimately did) it ultimately did not reveal anything new or interesting about the `PUBLICDATA_W49` responses. Every single response of counts only reflected the same relationship between public and private stances. This helped me narrow down my work to try and prioritize showing internal distinctions between those two categories instead.

A visualization that I considered, but rejected, is one which uses not a dodged bar chart but a stacked one, showing the two categories of public versus private data people as two bars and having those bars be partitioned accordingly to reflect the internal distribution of counts of categories that should be protected. This visualization looks as follows:

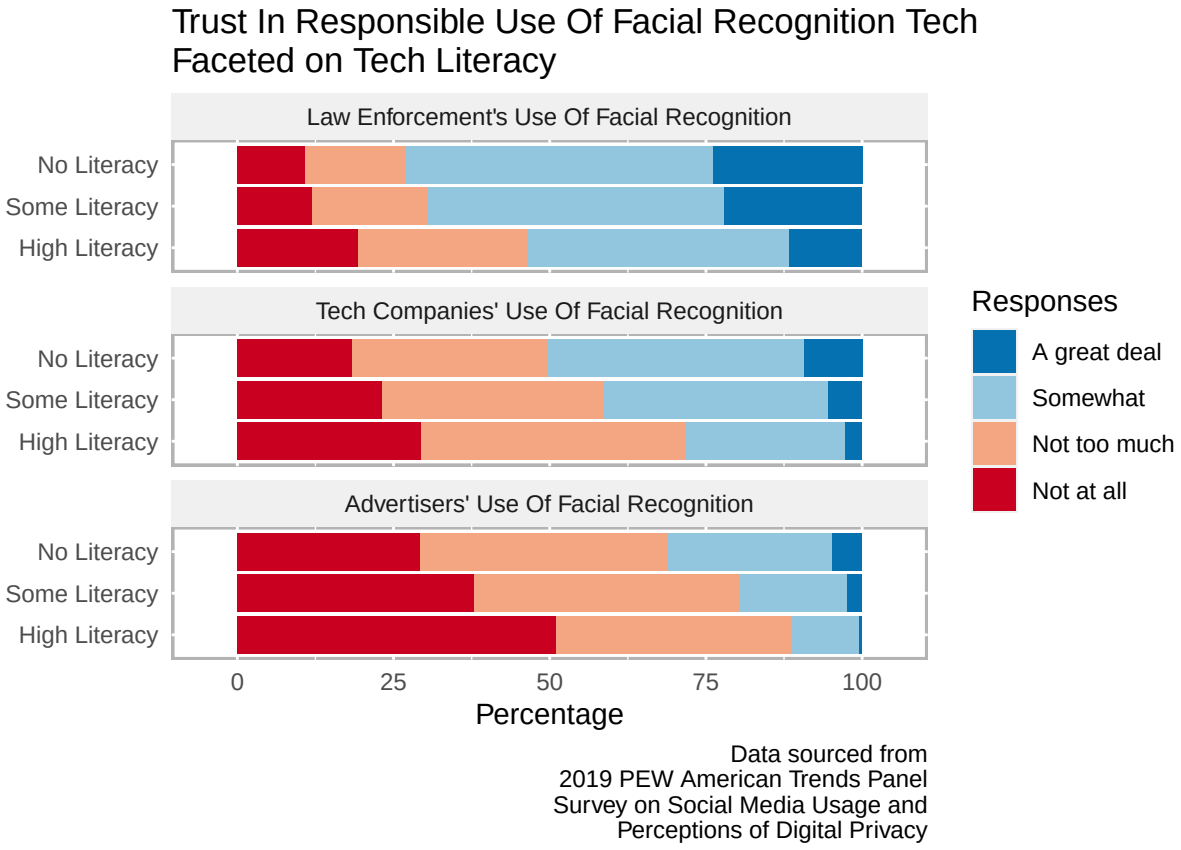


This visualization does both things that I would like it to do: it makes immediately apparent the difference in scale between those who prioritize public data versus private, and it has a clear color-coded breakdown of how many categories of data each class thinks should be protected. However, it is clunky and makes it harder to closely examine the distribution of sum counts than it is in my final graphic. As Healy breaks down, bad design is plagued by bad perception; stacking the boxes for each number makes it hard to directly compare the distribution of counts by not having a common base. This in turn warps our perception of the final data. So while this tells a similar story to the one up above, I ended up making additional aesthetic decisions to better convey that information in a clear and pleasing way.

Key Figure 2

When visualizing my key figure, I knew that I wanted to visualize some form of stacked bar chart to communicate the relative differences in comfort of various groups using facial recognition technology. Professor Ella Foster-Molina sent me an excellent article from Datawrapper making the case against diverging stacked bar charts and offering 100% stacked bar charts as the better alternative. The article argues that "the main problem with diverging bars, however, is comparability." On the whole, I agree; to me, the most compelling point of the piece relates to how neutrals are handled in the diverging stacked bar charts. Putting neutrals in the middle means that (1) none of the bars have a common baseline and (2) the chart is implying that part of the neutrals should be coded as positive and negative respectively. These issues are massively detrimental to making a visualization that is both easy to understand and not misleading. So, in general, the 100% stacked bar chart is the superior choice when visualizing Likert data. That all being said, none of my charts have neutrals. Thus, I think the diverging bar chart looks just fine because the overall positive-negative trend both within and between facets is more pronounced. Below is what the figure would look like with a 100% stacked bar chart.

Link to Datawrapper article: https://blog.datawrapper.de/divergingbars/



Distribution of Work

Jake and Adriana are best of friends and have split the work evenly. [PLACEHOLDER].