Pew Research Center

FOR RELEASE JULY 8, 2021

Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020

When discussing the election, Black Protestant pastors disproportionately urged voter turnout and registration

BY *Dennis Quinn and Aaron Smith*

FOR MEDIA OR OTHER INQUIRIES:

Dennis Quinn, Computational Social Scientist
Andrew Grant, Communications Associate

202.419.4372

www.pewresearch.org

RECOMMENDED CITATION

Pew Research Center, July, 2021, "Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

About Pew Research Center

Pew Research Center is a nonpartisan fact tank that informs the public about the issues, attitudes and trends shaping the world. It does not take policy positions. The Center conducts public opinion polling, demographic research, content analysis and other data-driven social science research. It studies U.S. politics and policy; journalism and media; internet, science and technology; religion and public life; race and ethnicity; global attitudes and trends; and U.S. social and demographic trends. All of the Center's reports are available at www.pewresearch.org. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

How we did this

This analysis is based on the texts of 12,832 sermons shared online by 2,143 U.S. religious congregations (nearly all of them Christian churches) delivered between Aug. 31 and Nov. 8, 2020 – a period that included the 2020 U.S. presidential election on Nov. 3 and the Sunday following Election Day. The dataset includes sermons from 438 evangelical Protestant congregations, 388 mainline Protestant congregations, 235 Catholic parishes and 205 historically Black Protestant congregations. The remaining congregations could not be reliably classified, belong to other Christian traditions (such as Orthodox Christian denominations) or belong to other faiths. While we collected sermons from a sample of congregations in our database, *all* sermons found on those websites were included in the analysis.

After identifying a comprehensive list of websites of U.S. churches using the Google Places API, we deployed a custom-built web scraper to navigate through each church's website, find any pages with sermons (in audio, video or text form), download each sermon along with the date on which it was delivered, and transcribe it from audio to text, if necessary. We then developed a machine learning classifier to identify sermons that discussed certain key topics – including the election, the COVID-19 pandemic and racism in America.

For technical and legal reasons, the Center was largely unable to collect sermons that were not posted or embedded directly on a church website. In particular, sermons shared or streamed solely on church Facebook accounts could not be included in this research. Sermons shared on YouTube accounts also were not included, unless they were embedded directly on the church's own website.

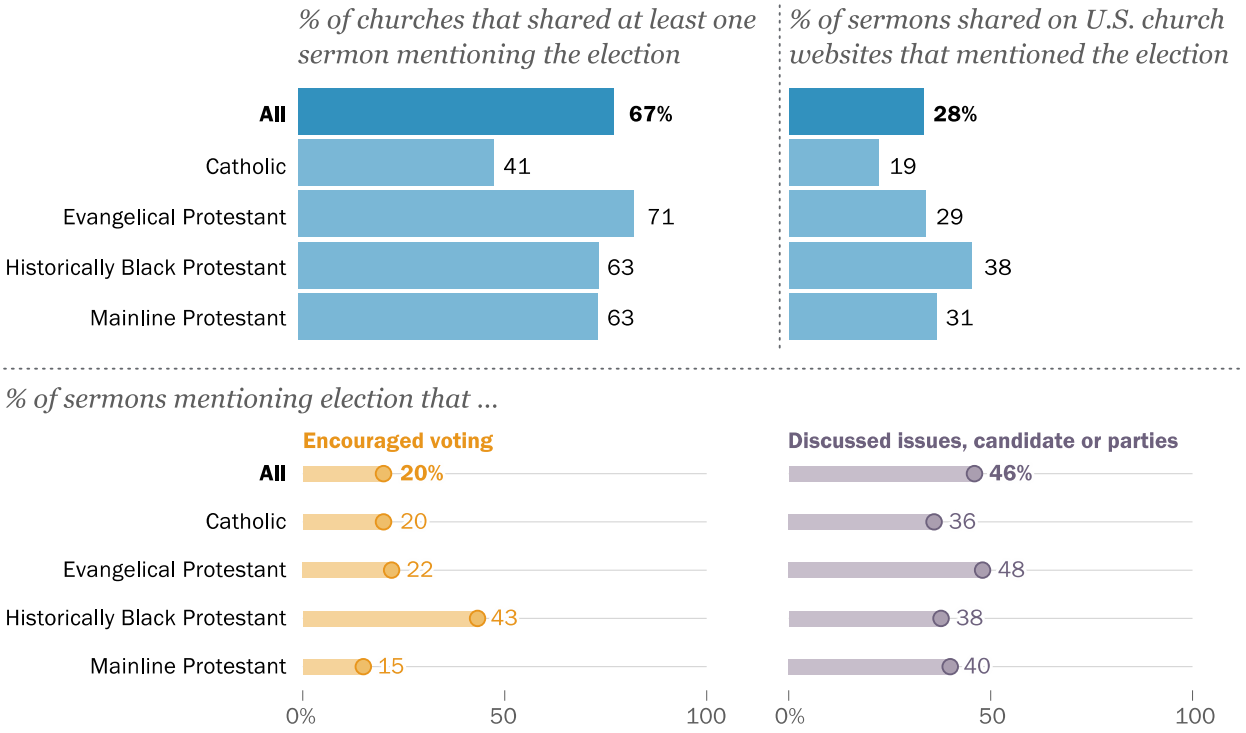
See Appendix A **[Link to appendix A]** for more details on how the congregations included in this study differ from congregations nationwide. See the Methodology **[Link to methodology]** for additional technical information on how this study was conducted.

Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020

When discussing the election, Black Protestant pastors disproportionately urged voter turnout and registration

Religious belief is key to many Americans' political identities, but the public is divided on whether clergy should preach about politics from the pulpit. So, when pastors across the country addressed their flocks last fall, how did they discuss an election that many Americans viewed as historically important?

Two-thirds of congregations heard at least one sermon mentioning the election during fall of 2020



Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

A new Pew Research Center analysis finds that among churches that posted their sermons, homilies or worship services online between Aug. 31 and Nov. 8, 2020, two-thirds posted at least one message from the pulpit mentioning the election. But these rates varied considerably among the four major Christian groups included in the analysis: 41% of Catholic congregations in the database heard at least one sermon mentioning the election, compared with 63% of both mainline Protestant and historically Black Protestant congregations and 71% of evangelical Protestant congregations.

Moreover, the content of the messages tended to differ. Roughly half of all evangelical Protestant sermons mentioning the election discussed specific issues, parties or candidates (48%), the highest share among the four major Christian groups. And, in discussing the election, evangelical pastors tended to employ language related to evil and punishment at a greater rate, using words and phrases such as "Satan" or "hell" at least twice as often as other clergy did. Evangelical pastors also were more likely to use the phrase "pray [for our] president" when discussing the election.

By contrast, historically Black Protestant pastors were by far the most likely to encourage voting and voter turnout: 43% of historically Black Protestant sermons mentioning the election either explicitly encouraged voting or discussed the election in a manner that assumed listeners would vote, roughly double the share of any other group. And when historically Black Protestant pastors discussed the election, they tended to use words or phrases related to voting or voter rights – such as "suppress[ion]," "early voting" and "register [to] vote" – more often than pastors from other groups.

Definitions and analytic frames used in this report

U.S. churches vary widely in the structure of their services and how much of those services they post online. Some post just the sermon. Others post the sermon and part of the service. Still others post the entire service. In many cases, the beginning and end of a sermon are not clearly labeled in the text, audio or video files on a church's website. As a result, the automated tools used for this analysis cannot isolate sermons from other elements of religious services with precision.

In this report, **an "online sermon" refers to a portion of a religious service posted on a church website that contains a commentary from the pulpit** but sometimes may include other parts of the service as well.

This report also uses two different frames for comparison, depending on the focus of the analysis. Some findings are based on the share of *all sermons* that have certain characteristics (e.g., "28% of sermons delivered during the study period referenced the election"). Other findings are based on the share of *all congregations* that heard discussion of a topic in *any* of their sermons (e.g., "67% of all congregations heard at least one sermon mentioning the election during the study period").

Although most congregations posted at least one sermon mentioning the election at some point during the study period, relatively few pastors openly stumped for particular candidates or parties. Indeed, explicit endorsements from the pulpit were rare enough that researchers could not develop a machine learning model that would reliably identify such language across all sermons in the database. However, in a sample of 535 sermons mentioning the election that researchers examined while attempting to train such a model, 61 seemed clearly to favor either Republicans or Democrats, even if they did not mention parties or candidates by name.

Pastors also discussed other prominent issues during the period. About eight-in-ten congregations in the database (83%) heard at least one sermon touching on the COVID-19 pandemic, while 44% heard at least one reference to racism in America. Catholic congregations stood out as the least likely to mention any of the topics analyzed in this study during the services or homilies they shared online.

In discussing racism in America, evangelical pastors disproportionately used oblique phrases such as "racial tension." Meanwhile, clergy in mainline Protestant and historically Black Protestant congregations tended to discuss this issue using more direct terms like "anti-racism" and "White supremacist."

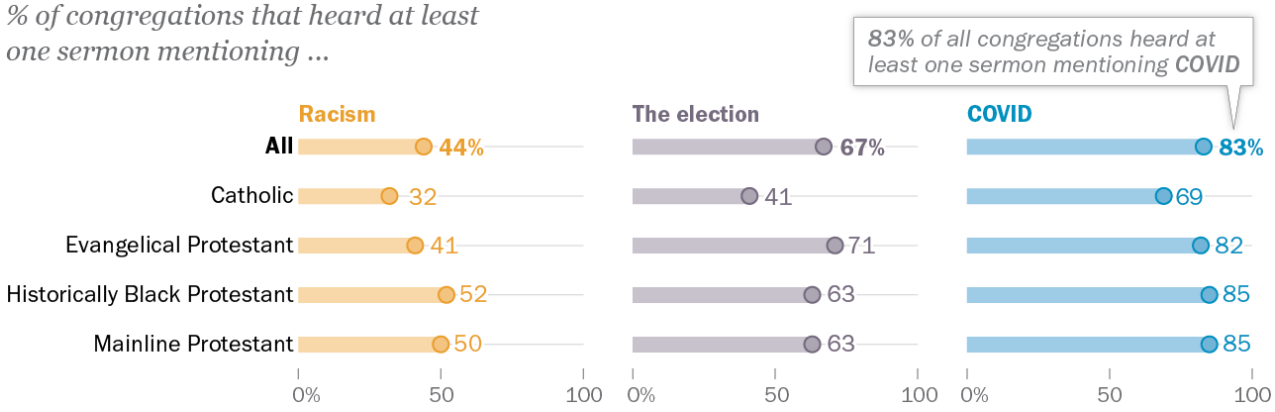
These are among the main findings of an analysis of 12,832 sermons, homilies or full services delivered to 2,143 American congregations between Aug. 30 and Nov. 8, 2020 – a period when many congregations were streaming their services online due to the COVID-19 pandemic. This study builds on an earlier Center report, which examined sermons shared online in mid-2019.

It is important to note that the sermons included in this dataset are not necessarily representative of the messages delivered in all U.S. religious congregations, for a variety of reasons. First, this analysis focuses on Christian churches and does not include other religious traditions. Moreover, not all Christian churches make their sermons publicly available online – and those that do place their sermons online may choose selectively, posting some but not others. Nonetheless, the sermons database provides a window into what churchgoing Americans heard in the pews – physical or virtual – during a historic moment in American civic life.

A majority of churches shared or livestreamed sermons discussing the election and COVID-19 pandemic during the fall of 2020

The 2020 election, the COVID-19 pandemic and nationwide protests over systemic racism and police violence against Black Americans dominated news cycles in the latter half of 2020. This analysis finds that they featured prominently in U.S. sermons as well. (See the Center's topic pages for more on the 2020 election and the COVID-19 pandemic.)

During the fall of 2020, most congregations heard sermons mentioning the election and COVID-19 pandemic



Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

A majority of congregations included in this study heard at least one sermon mentioning the COVID-19 pandemic (83%) or the 2020 election (67%). And a substantial minority of congregations (44%) heard at least some discussion of racism in America. Still, pastors of different religious traditions discussed each topic at different rates.

For instance, mainline Protestant and historically Black Protestant congregations were more likely than evangelical Protestant or Catholic congregations to hear discussion of racism from the pulpit during this time period. Conversely, evangelical churchgoers were the most likely to hear discussion of the election.

Catholic priests were consistently the least likely to mention any of these three issues in their sermons, homilies or services shared online. Fewer than half of Catholic congregations in the database heard a single mention of the election (41%) or racism (32%) during the 10-week study

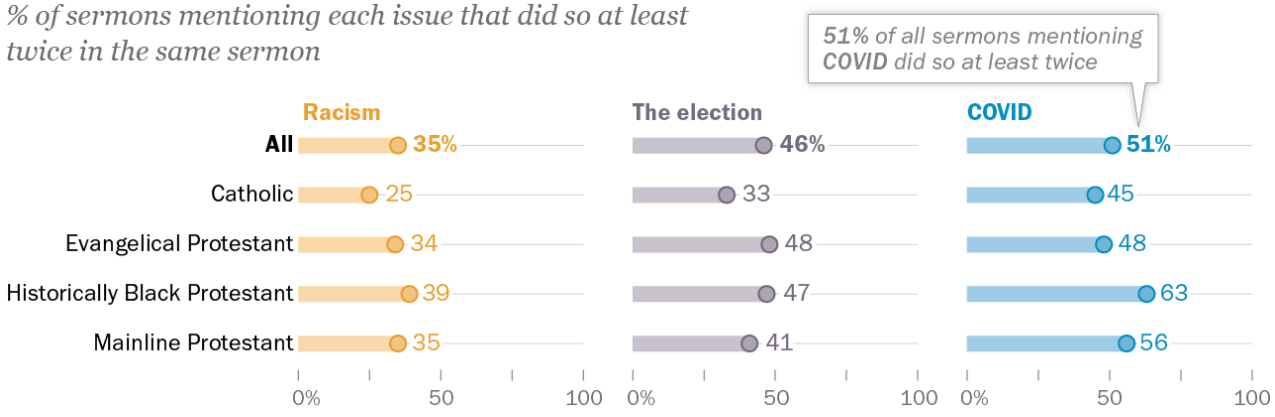
period. And although 69% of Catholic congregations heard at least one mention of the pandemic, congregations belonging to the other three major Christian traditions were at least 10 percentage points more likely to hear messages from the pulpit about the coronavirus.

Pastors often mentioned these topics multiple times in the same sermon

To better understand how heavily pastors focused on each topic, researchers broke each sermon down into smaller segments of 250 words each, or fewer in the case of a sermon's final segment (the median sermon in this collection had 26 such segments, and a segment of that length often occupied one to two minutes of speaking time). The research team then used a combination of labeling by human coders and statistical modeling to determine how many of these individual segments mentioned the three major topics examined in this study.

Sermons mentioning the pandemic often did so multiple times

% of sermons mentioning each issue that did so at least twice in the same sermon



Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

Using a similar technique on a set of sermons delivered in the spring of 2019, the Center found that when pastors discussed abortion, they tended to do so only glancingly. Just one-quarter of all sermons that mentioned abortion did so in more than one 250-word segment.

In contrast, pastors tended to mention the topics examined in this study with greater regularity. Some 35% of sermons where the pastor discussed racism – and 46% of those that mentioned the election – did so in at least two separate 250-word segments.

Pastors were particularly likely to discuss the COVID-19 pandemic at some length: 51% of sermons mentioning this topic included references to the pandemic in two or more 250-word segments. And certain groups were especially likely to make the pandemic a recurring theme in their services. Some 56% of sermons by pastors in mainline Protestant churches that mentioned the pandemic (and 63% of those by pastors in historically Black Protestant churches) did so at least twice.

Nearly half of election-related sermons discussed specific issues, candidates or parties; one-in-five encouraged voting

Researchers also assessed what share of sermons discussing the election encouraged listeners to vote, as well as what share discussed specific issues, candidates or political parties. (The Supreme Court, abortion or taxes would all be examples of an issue). Among the 28% of sermons that discussed the election, roughly half (46%) discussed specific issues, parties or political candidates while 20% encouraged listeners to vote. (For more on how we identified these topics, see the Methodology. **[link to methodology]**) Translated to the congregational level, this means that 23% of *all congregations in the database* heard at least one sermon during the fall of 2020 encouraging them to vote, while 43% heard at least one discussing parties, issues or candidates.

Researchers also attempted to identify instances in which pastors openly encouraged their congregants to vote for a specific party or candidate. However, such explicit admonitions were rare and, as a result, the Center was unable to systematically identify them across the database. But among a sample of 535 segments of sermons that discussed the election, researchers labeled 35 as advocating for Republicans and 26 as advocating for Democrats. This included some cases in which pastors named a candidate or party as well as cases in which they advocated a clearly partisan array of policy positions.

Predictably, political discussions reached a crescendo during the week of the election. Although 28% of all sermons delivered over the entirety of the study period mentioned the election in some way, that share rose to 49% of all sermons delivered during the week preceding the election – including 61% of sermons given that week in historically Black Protestant congregations. Also, it appears that sermons mentioning the election were, on the whole, as likely as other sermons to have some scriptural framing. Fully 96% of all sermons that touched on the election mentioned at least one book of the Bible by name, compared with 95% of sermons that did not mention the election.

Different Christian groups used distinctive terminology when discussing the election

To better understand the tenor of what congregants heard in sermons during the fall of 2020, researchers analyzed the words that pastors of each group used most disproportionately – relative to other Christian groups – when discussing topics such as the election. In discussing the election, evangelical pastors were disproportionately likely to use the words "Satan" and "hell," while historically Black Protestant pastors focused heavily on voter turnout and registration.

To conduct this analysis, the research team first identified all the 250-word segments from a given Christian group that discussed a topic – for instance, all segments of evangelical Protestant sermons that mentioned the election.

Next, we calculated the share of those segments that used a certain word or phrase. Finally, we calculated that same value for all the sermon segments of *other* groups that discussed the same topic, and we divided the former by the latter. This statistic represents how many times more often a word or phrase appears when pastors in one Christian tradition discuss a topic relative to when pastors in the other Christian traditions discuss that same topic. Common conjunctions, prepositions and articles (such as and, but, of, in, to, from, a, the) were removed for this analysis, and many words were reduced to their roots. For example, the words "election" and "elected" would be reduced to "elect-." In performing this analysis, researchers also removed any words or phrases used in fewer than 1% of all segments.

While the preceding section of this report examined the prevalence of broad topics within sermons as a whole, this analysis focuses on the short (250-word) segments of sermons that contain pertinent mentions of those topics. This focus is necessary because most Christian services contain core elements – such as traditional prayers, a reading from scripture or the giving of communion – that are far more statistically distinct from other groups' services than any differences in how they discuss a topic like politics. Focusing on the short segments that mention the election removed many of these liturgical elements, allowing other differences to become apparent.

When discussing election, evangelical pastors disproportionately mentioned 'hell' and 'Satan'; pastors in historically Black churches more likely than others to urge voting

When pastors in evangelical Protestant congregations discussed the election, they disproportionately used phrases related to prayer and to forces of evil. Six of the 10 most distinctive terms in their sermons included the word "pray," including variations of the phrase "pray … president" such as "pray for our president" or "pray for the president." Sermons in

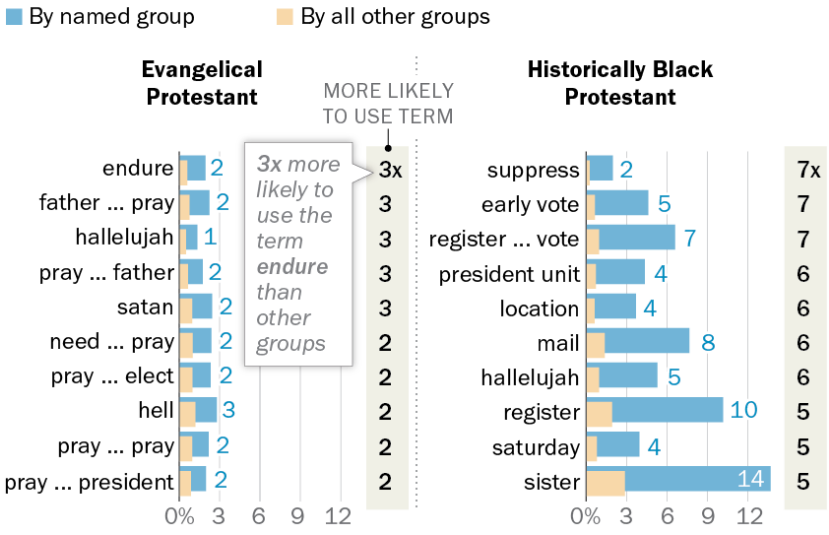
evangelical congregations also disproportionately used terms such as “Satan” or “hell” when discussing the election.

It is important to note that even though these terms were *distinctive* to evangelical sermons mentioning the election, they were not especially *common* in evangelical sermons. The 10 most distinctive terms in evangelical sermons discussing the election were all used in fewer than 5% of segments discussing the election.

By contrast, historically Black Protestant pastors disproportionately used words related to voter suppression, registration and turnout. The word “suppress” (along with common variants such as “suppressed” or “suppression”) was the single most distinctive term used by pastors in historically Black Protestant churches when discussing the election. They also urged their congregations to vote, using words and phrases like “early vot[ing],” “mail” and “register … vote.” Further, some of these phrases were fairly common. For instance, historically Black protestant pastors used the word “register” in 10% of segments that mentioned the election.

Clergy in evangelical and historically Black Protestant churches used different language than other groups when discussing the election

% of sermon segments using term when discussing the election



Note: The words in this analysis were “stemmed” or converted to their roots. Common words (such as most prepositions) and words used by more than 95% or fewer than 1% of all sermon segments were removed.
Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches).
“Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020”

PEW RESEARCH CENTER

Catholic and mainline Protestant sermons touching on the election, by comparison, were primarily distinguished by language related to their respective religious practices – for instance, “Mass,” “bishop” and the word “Catholic” in Catholic sermons, and “communion” in mainline Protestant sermons. This indicates that although these groups may have used some distinguishing language in discussing

the election, that language was *less distinctive* than the usual hallmarks of a Catholic or mainline Protestant service.

Direct quotes from sermons discussing politics and the 2020 election

"We also established that anytime we endeavor to rebuild like the Israelites, we will discover that we will face opposition. For them, opposition came in the form of a Samaritan named Sanballat and others like Tobiah, who tried to keep them from rebuilding. For us, opposition comes in the form of voter suppression, voter intimidation, systemic injustice and a president whose tyrannical leadership chips away at our democracy on a daily basis." – Historically Black Protestant sermon

"If all lives matter and individual lives matter, then there's no such thing as being pro-abortion, no such thing as being pro racism, or ignoring human trafficking, or discrimination or prejudice. Those are the things that, if all lives matter, should be our priority, right? Therefore, I would encourage everyone: You need to vote biblical morality and values, if all life matters." – Evangelical Protestant sermon

"Perhaps, then, today we need to look beyond the chaos of Tuesday's election and settle instead on the overreaching truth of our lives on Earth. That is what St. Paul told the Thessalonians: 'Thus we shall always be with the Lord. Therefore, console one another with these words.' So dead or alive, we are always with the Lord in the Gospel. Today, we are reminded to be ready for anything in life, to be a people prepared not only to deal with the pandemic and a messed up presidential election, but to remember that we are to follow on the path of those wise virgins. To have not only our own lamps lit, but to have extra oil with us just in case. Like it or not, we need to be prepared to meet the Lord when he does call us home." – Catholic sermon

"We simplify everything as if all these complex issues could be boiled down into a right and a wrong. It's preposterous. And then we define each other by these absurd categories that we have created, and what happens is, we don't know each other anymore. We're defined by labels instead of seeing each other as human beings, and we stop listening. And that, my friends, can get very dangerous as this election approaches. If we define each other solely by our politics and our slogans and our words, we cease to listen to one another, and we are in danger of becoming just like the scribes and the elders and the Pharisees. So I ask you: Don't give me any words. I don't want to hear a slogan. But do tell me this: Do you serve the less fortunate than you? Do you take time out of your life to help those in need, to do something that is solely not for you, but for someone else?" – Mainline Protestant sermon

Quotations have been lightly edited for readability.

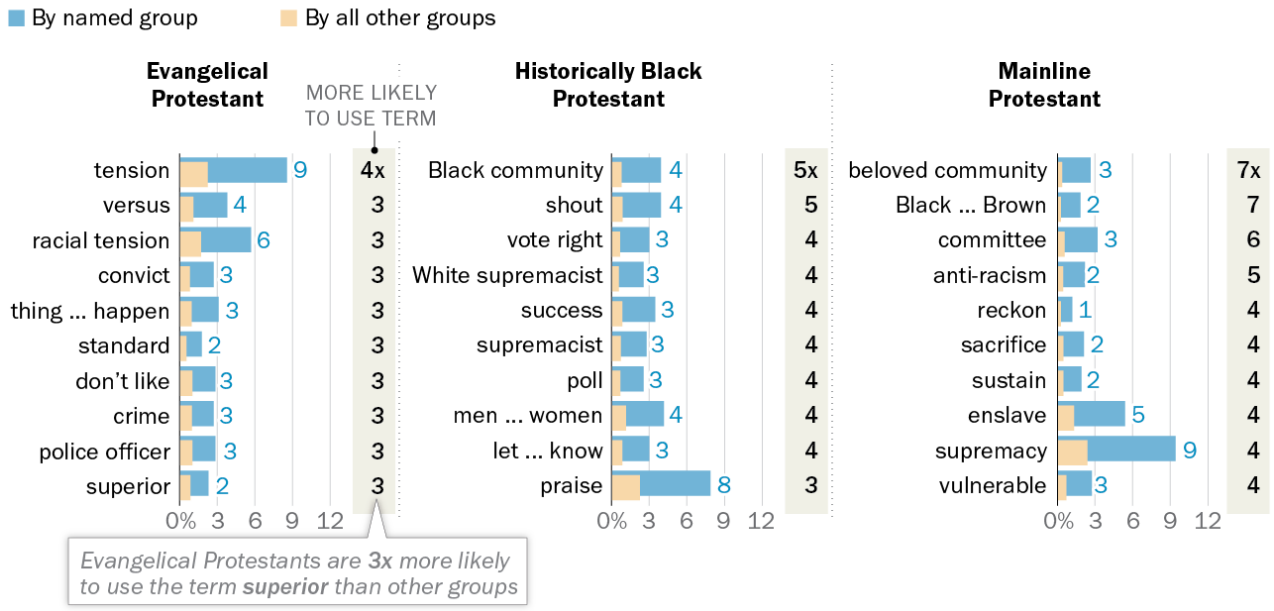
In discussing racism in America, mainline Protestant clergy urged anti-racism, while clergy in historically Black Protestant churches discussed voting and White supremacy

These groups also used distinctive language to discuss racism in America. Pastors in mainline and historically Black Protestant congregations tended to address racism and racial justice directly. For instance, the most distinctive terms used by mainline Protestant pastors included "supremacy" and "anti-racism," and the most distinctive terms used by Black Protestant clergy included phrases like "White supremacist" and "Black community."

Evangelical pastors, by comparison, often used more oblique language to describe racism. Terms such as "tension" and "racial tension" are among the most distinctive terms in evangelical sermons mentioning racism in America. Evangelicals also used terms like "police officer," "crime" and "convict" about three times as often as other pastors when discussing racism.

In discussing racism, mainline and historically Black Protestant sermons were more likely to mention anti-racism and White supremacy

% of sermon segments using term when discussing racism



Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

Catholic sermons were once again distinguished by words common to Catholic homilies or services such as "[Pope] Francis" or "Bishop." For a full list of each group's most distinctive terms, see Appendix B. **[link to appendix B]**

Direct quotes from sermons discussing racism in America

"I don't know when in the world Black lives are gonna matter to some White people, and I don't know why they love Black culture but can't stand some of the Black people. I don't know why Black women who have literally given everything to this world still can't find a safe space for them to grow, nourish and flourish, or even live. And since I know I already jumped out there, can I just tell you how I'm really feeling? I don't know how it's possible for there to have been a $12 million wrongful death lawsuit that was settled, but have nobody responsible for the death. I don't know how sheetrock could get more justice than one of our beautiful Black sisters. And I don't know how many more miscarriages of justice we will have to endure before Black people give birth to a response that might turn this country upside down." – Historically Black Protestant sermon

"Our original sin, then, according to critical race theory, is whiteness. … Salvation from that fall begins when the oppressors become woke – you heard that term, 'woke' recently? This is what's going on in the NFL, and that's why I won't be watching it. When they see and repent from their own sins as oppressors and begin to dismantle the inherently oppressive structures of their culture, they're woke. In other words, you destroy your country's history, you tear down all the the statues of White people. And since Whites can't see their own racism, according to the CRT [critical race theory], they need to learn to see the world through the CRT lens. Only then will racial equality be possible. For CRT, then, salvation comes through law, not grace. Forgiveness only comes after complete and ongoing capitulation. Ultimately, CRT is rooted in a blend of two worldviews: Marxism – the oppressed and the oppressor – and secular humanism – the belief that humanity is capable of self-fulfillment and rescue by self-effort apart from God. However, once God is removed, so is any objective standard defining what it means to live a fulfilled, moral life." – Evangelical Protestant sermon

"The murder of George Floyd has blown open the terrible evil of individual and institutional racism that serves the dominant culture so well. Whether blatant or hiding menacingly under the surface, Roxane Gay wrote that we Blacks live with the knowledge that a hashtag is not a vaccine for White supremacy. We live with the knowledge that, still, no one is coming to save us. The rest of the world yearns to get back to normal. For Black people, normal is the very thing from which we yearn to be free." – Catholic sermon

"But we have to understand that we are called. There is work to be done. We are called to begin that healing. We're called to begin that healing. The way that we overcome systemic racism, the way we overcome systemic poverty, we overcome Bible abuse against homosexuals and queer people, the way we overcome transphobia and sexism, the way we overcome people who don't have a living wage because of the color of their skin or the sex that they were born or identify as, the way we overcome all of that – is coming together and healing across this great divide that has been created. And we do that by coming together prepared for the road ahead." – Mainline Protestant sermon

Quotations have been lightly edited for readability.

Appendix A: Congregations that share sermons online are disproportionately evangelical and have larger audiences; sermons are longer than in 2019 study

Among the congregations included in this study for which researchers were able to identify a religious tradition, 52% were evangelical Protestant; 22% were mainline Protestant; 4% were Catholic; and 1% were historically Black Protestant churches. The database also contained a small number of congregations belonging to smaller Christian groups or non-Christian religions; for more, see the table below.

In drawing the sample of congregations used in this study, the Center oversampled smaller groups such as historically Black Protestant congregations to ensure there was enough data from each group to analyze. To account for this, researchers subsequently adjusted each group to represent its original prevalence in the database – a process called weighting. The percentages presented above are weighted, which is why they do not directly align with the raw number of congregations described earlier in this report.

The congregations included in this study were also larger than congregations nationwide. Roughly half (51%) had more than 200 members, compared with 34% of all congregations nationwide that have more than 200 members, according to the 2012 National Congregations Study.[1] Finally, 59% of all congregations in the database were in geographies considered "suburban" under a Pew Research Center classification scheme devised in 2019. A further 28% were rural, and 13% were urban. (The community type designations used in this report are designed to reflect Americans' self-reported community types; about four-in-ten – 43% – of Americans say they live in a suburban area.)

[1] Chaves, Mark, Shawna Anderson, and Alison Eagle. 2012. "National Congregations Study." Duke University Department of Sociology.

Weighted and unweighted makeup of congregations and sermons from each religious tradition

Raw number and weighted % of all sermons or congregations in the database

Religious tradition	Raw number of sermons	Weighted share of sermons	Raw number of congregations	Weighted share of congregations	Sermons per congregation
Evangelical Protestant	2,990	55%	438	52%	7
Unclassifiable*	2,568	4	376	4	7
Mainline Protestant	2,204	19	388	22	6
Catholic	1,297	3	235	4	6
Unmatched to InfoGroup*	1,240	17	175	15	7
Historically Black Protestant	990	1	205	1	5
Other faiths*	663	0	167	1	4
Orthodox Christian*	481	0	75	0	6
Other Christian*	352	0	70	0	5
Mormon*	18	0	4	0	5
Jewish*	17	0	8	0	2
Buddhist*	12	0	2	0	6

*Not analyzed on its own.
Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches that posted sermons online).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

The messages analyzed in this study are longer than those the Center collected in 2019

The median length of a sermon, homily or service analyzed in this study was 40 minutes, compared with 37 in the Center's 2019 analysis of online sermons. However, this difference varied dramatically by group: Catholic messages, for example, ran a median of 27 minutes in the 2020 study period, compared with 14 minutes in 2019, and mainline Protestant messages were 36 minutes at the median in 2020, compared with 25 minutes in 2019. This suggests that many Catholic and mainline Protestant congregations were streaming full services online in the fall of 2020, rather than posting standalone sermons or homilies.

By comparison, evangelical Protestant sermons ran a median of 41 minutes during the 2020 study period, only slightly longer than the median of 39 minutes in 2019. And finally, historically Black

Protestant messages – which ran a median of 54 minutes in 2019 – were about nine minutes shorter in the 2020 study period, at 45 minutes.

Appendix B: Words or phrases most disproportionately used by clergy of each group

Pastors of each group used distinctive language when discussing each topic

Words or phrases most disproportionately used by pastors of each group when discussing each topic

Evangelical Protestant	Mainline Protestant	Catholic	Historically Black Protestant
Terms used in discussing the election			
Endure	Elect ... eri	Mass	Suppress
Father pray	Eri	Catholic	Early vote
Hallelujah	Space	Bishop	Register ... vote
Pray Father	Almighty	Priest	President ... unit
Satan	Mary	Dignity	Location
Need pray	Creation	Mary	Mail
Pray ... elect	Hear ... prayer	Hear ... prayer	Hallelujah
Hell	Communion	Conscience	Register
Pray ... pray	Gracious	Lord ... lord	Saturday
Pray ... president	Office	Mother	Sister
Terms used in discussing the COVID-19 pandemic			
Pride	Sanctuary	Parish	Hallelujah
Unto	Reverend	Lord hear	Reverend
Devil	Gratitude	Lord ... lord	Unto
Nebuchadnezzar	Bread	Mass	Thank ... lord
Satan	Hear ... prayer	Bishop	Prais ... god
Bible	Worship ... servic	Catholic	Thank god
Know ... say	Generosity	Father ... son	Wash
Verse	Grieve	Pray ... lord	God bless
Daniel	Lord ... mercy	Hear ... prayer	Midst ... pandem
Guy	Compass	Priest	Sister
Terms used in discussing racism in America			
Tension	Belov ... community	Lord ... lord	Black ... community
Versus	Black ... brown	Pray ... lord	Shout
Racial ... tension	Committee	Francis	Vote ... right
Convict	Anti-racism	Parish	White ... supremacist
Thing ... happen	Reckon	Bishop	Success
Standard	Sacrifice	Hear ... prayer	Supremacist
Don't ... like	Sustain	Catholic	Poll
Crime	Enslave	Let ... pray	Men ... women
Police ... officer	Supremacy	Almighty	Let ... know
Superior	Vulnerable	Mary	Praise

"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"
PEW RESEARCH CENTER

Methodology

This study is based on an analysis of 12,832 sermons, homilies or worship services delivered between Aug. 31 and Nov. 8, 2020, and collected from the websites of 2,143 churches found via the Google Places application programming interface (API), a tool that provides information about establishments, geographic locations or points of interest listed on Google Maps. Pew Research Center data scientists collected these sermons between Nov. 12 and Dec. 21, 2020. The collection program used a machine learning model to identify webpages likely to contain sermons. It also used a set of specially designed algorithms to collect media files with dates from those pages, identify the files containing sermons, and transcribe those files for further analysis.

Researchers conducted this process on a sample of all churches found on Google Places. The sample was designed to ensure that researchers had enough cases to analyze sermons from smaller Christian traditions, including historically Black Protestant congregations. The list of churches is the same as the Center used in its 2019 study, meaning that new churches started since that time could not be included.

Here is a brief description of the main steps in the data collection process. Each is described in greater detail in corresponding sections of the methodology that follow.

[For each of the five following links, please link to section currently linked internally]

Finding every church on Google Maps: Pew Research Center's data scientists began by identifying every institution labeled as a "church" in the Google Places API, including each institution's website (if it had one). This yielded an initial pool of 478,699 institutions. This list contained many non-congregations and duplicative records, which were removed in subsequent stages of the data collection process.

Determining religious tradition, size, and predominant race or ethnicity: The churches found via the Google Places API lacked critical variables like denomination, size or predominant racial composition. To obtain these variables, Center researchers attempted to match each church found on Google Places to a database of religious congregations maintained by InfoGroup, a targeted marketing firm. This process successfully matched 262,876 congregations and captured their denomination, size and racial composition – where available – from the InfoGroup database.

Identifying and collecting sermons from church websites: Data scientists deployed a custom-built software system (a "scraper") to the websites of a sample of all churches in the initial dataset – regardless of whether they existed in the InfoGroup database – to identify, download and transcribe the sermons they share online. This program navigated to pages that appeared likely to contain sermons and saved every dated media file on those pages. Files dated between Aug. 31 and Nov. 8, 2020, were downloaded and transcribed. Researchers then coded a subset of these transcripts to determine whether they contained sermons and trained a machine learning model to remove files that did not contain sermons from the larger dataset.

Evaluating data quality: The resulting list of congregations with sermons online differs from congregations nationwide in critical ways, and is far smaller than the 478,699 institutions the Center initially found on Google Places. Of those congregations, 19,085 were selected to have their websites searched for sermons, and of that sample 2,143 made it into the final sermons dataset – meaning the scraper was able to successfully find and download sermons from their websites. Of the 2,143 churches in the final dataset, the Center was able to match 1,968 with variables derived from InfoGroup data, such as their religious tradition.

To properly contextualize these findings, researchers needed to evaluate the extent of these differences and determine the scraper's effectiveness at finding sermons when they were present. To accomplish this, researchers manually examined the websites of 253 churches randomly sampled from the Center's database in search of sermons, deployed the scraper to those same websites, and compared the results. The scraper successfully found 23% of all sermons (a large body of sermons were missed due to the Center's lack of access to the YouTube API).

Classifying sermons by topic: Each sermon was divided into segments small enough for a researcher to read, and a subset of these segments was examined individually and coded (i.e., labeled) to indicate whether the pastor discussed the election, COVID-19 and racism in America. Researchers then trained machine learning classifiers to identify these topics across the whole database. Among sermons that discussed the election, researchers repeated this process to identify instances where a pastor encouraged congregants to vote, or discussed specific issues, candidates or parties.

Finding every church on Google Maps

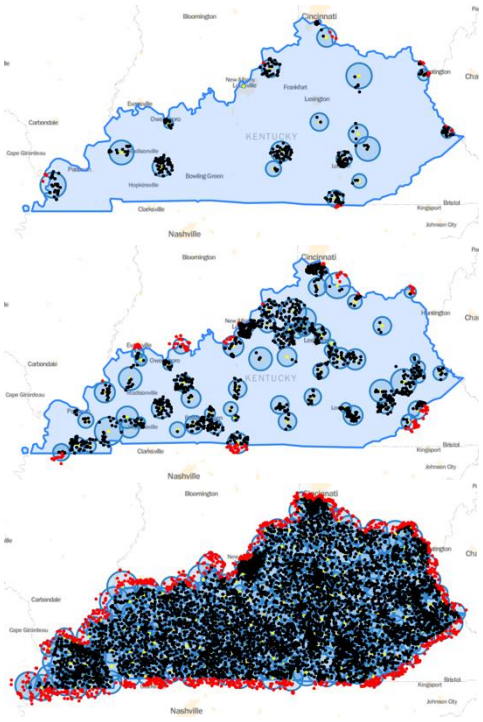
To build a comprehensive database of U.S. churches, Center researchers designed an algorithm that exhaustively searched the Google Places API for every institution labeled as a "church" in the United States. At the time of searching, Google offered only search labels that hewed to specific groups, such as "church" or "Hindu temple." As a result, researchers could not choose a more inclusive term and ultimately used "church" to cover the lion's share of religious congregations in the U.S. Researchers used Google Places because the service provides websites for most of the institutions it labels as churches.

The program searched each state in the country independently. It began by choosing a point within the state's area, querying the API for churches around that point, and then drawing a circle around those churches. The algorithm then marked off that circle as searched, began again with a new point outside the circle, and repeated this process until the entire state was covered in circles. Researchers dictated that results should all be returned in order of distance from the query point, regardless of other factors like prominence. This means that for each query, researchers could deduce that there were no omitted results *closer* to the center point of the query than the farthest result returned by the API.

In practice, researchers could have used the farthest result to draw the coverage areas, but often used a closer one in an effort to be conservative.[2] The algorithm relied on geographic representations of each state –

How Pew Research Center searched each state for churches

Researchers used the locations of churches to determine which areas were searched, and continued searching each state until the entire area was covered



"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

[2] As of summer 2018, the Google Places API only returned a maximum of 60 results per search. If 50 or fewer results were returned, the full search area was considered to have been fully searched successfully. If more than 50 results were returned, the program roughly determined the density of the search area by looking at the distance between the 40th and 50th result. In cases where the distance between these

called “shapefiles” – that are publicly available from the U.S. Census Bureau.

Researchers used a previous version of this algorithm in fall 2015 to collect an early version of the database. The early version of the algorithm was less precise than the version used in 2018, but compensated for that imprecision by plastering each area – in that case counties, not states – with dramatically more searches than were needed. The 2015 collection yielded 354,673 institutions, while the 2018 collection yielded 385,675. Researchers aggregated these two databases together for this study, counting congregations that shared the same unique identifier only once. Excluding these duplicates, the aggregated database included 478,699 institutions.

Determining religious tradition, size, and predominant race or ethnicity

This initial search process produced a comprehensive list of institutions labeled as churches on Google Places. But the resulting database contained almost no other information about these institutions – such as their denomination, size or predominant race or ethnicity. To acquire these variables, Center data scientists attempted to find each church listed in Google Places in an outside database of 539,778 congregations maintained by InfoGroup, a targeted marketing firm.

Researchers could not conduct this operation by simply looking for congregations in each database that shared the same name, address or phone number, because congregations may have names with ambiguous spellings or may change their addresses or phone numbers over time. A simple merging operation would fail to identify these “fuzzy” matches. To account for this ambiguity, human coders manually matched 1,654 churches from the Center’s database to InfoGroup’s, and researchers trained a statistical model to emulate that matching process on the remainder of the database.

The matching involved multiple stages:

1. **Limiting the number of options coders could examine:** As a practical matter, coders could not compare every church in the Center’s database to every church in InfoGroup’s. To reduce the number of options presented to each coder, researchers devised a set of rules that delineated what congregations in the InfoGroup database could *plausibly* be a match for any given record in the Center’s collection. This process is known as blocking.

results was small, where a small mistake in the coverage area could exclude institutions), the program was more conservative in determining the area that was successfully covered. If the 40th and 50th results were less than 200 meters apart, the program used the 15th result to determine the successful coverage area; for 200-500 meters, the 25th result was used; for 500-1,000 meters, the 35th result was used; and for a distance of over 1,000 meters, the 45th result was used.

For any given church in the Center's database, the blocking narrowed the number of plausible matches from InfoGroup's database to only those that shared the same postal prefix (a stand-in for region). Next, researchers constructed an index of similarity between each church in the Center's database and each plausible match in the InfoGroup database. The index consisted of three variables summed together, each normalized to a 0-1 range. The variables were:

a. The distance in kilometers between each two churches' GPS coordinates.

b. The similarity of their names, using the Jaro distance.

c. The similarity of their addresses, using the Jaro-Winkler distance.

These three variables were then summed together, and coders examined the 15 options with the greatest similarity values (unless two churches shared the same phone number and postal prefix, in which case they were always presented to the coders as an option regardless of their similarity value). In the rare event that there were fewer than 15 churches in a postal prefix area, coders were presented all churches in that postal area.

2. **Manually choosing the correct match for a sample of churches:** A group of five coders then attempted to match a sample of 2,900 congregations from the Center's database to InfoGroup's. In 191 cases where coders were unsure of a match, an expert from the Center's religion team adjudicated. Overall, coders successfully matched 1,654 churches. Researchers also selected a sample of 100 churches to be matched by every coder, which researchers used to calculate inter-rater reliability scores. The overall Krippendorf's Alpha between all five coders was 0.85, and the individual coders' alpha scores – each judged against the remaining four and averaged – ranged from 0.82 to 0.87.

3. **Machine learning and automated matching:** As noted above, this process generated 1,654 matches between the two datasets. It also generated 41,842 non-matches (each option that the coders did not choose was considered a non-match). Center researchers used these examples to train a statistical model – a random forest classifier in Python's scikit-learn – that was then used to match the remaining churches in the collection.

   Researchers engineered the model to have equal rates of precision (the share of items identified as a match that were truly matches) and recall (the share of true matches that were correctly identified as such). This means that even while there was an error rate, the model neither overestimated nor underestimated the true rate of overlap between the databases. The

model's average 5-fold cross-validated precision and recall were 91%, and its accuracy (the share of all predictions that were correct) was 99%.

To apply the model to the remaining data, researchers had to replicate the blocking procedure for all 478,699 churches in the Center's database, presenting the model with a comparable number of options to those seen by the coders. Researchers also calculated several other variables for use in the model that coders did not have access to, but in which the model might find statistical value.

The model's features (variables) were: the distance between each pair of churches; the *ranked* distance between each pair (whether each was the closest option, the second closest, etc.); the similarity of their names using the Jaro distance; the similarity of their addresses using the Jaro-Winkler distance; a variable denoting whether they shared the same phone number; and one variable each for the most commonly appearing words from church names in the Pew Research Center database, denoting the cumulative number of times each word appeared across both names.

Center data scientists applied this model to each church in the Center's database, successfully identifying a match for 262,876 in the InfoGroup database. For each matched church, researchers merged the congregation's denomination, predominant race or ethnicity, and number of members into the database, where these variables were available.

Once the Center merged these variables into the database, researchers categorized InfoGroup's religious groups into one of 14 groups: evangelical Protestant, mainline Protestant, historically Black Protestant, Catholic, Orthodox Christian, Mormon, Jehovah's Witness, other Christian, Jewish, Muslim, Hindu, Buddhist, other faiths, and unclassifiable.

Protestant congregations with identifiable denominations were placed into one of three traditions – the evangelical tradition, the mainline tradition or the historically Black Protestant tradition. For instance, all congregations flagged as affiliated with the Southern Baptist Convention were categorized as evangelical Protestant churches. All congregations flagged as affiliated with the United Methodist Church were categorized as mainline Protestant churches. And all congregations flagged as affiliated with the African Methodist Episcopal Church were categorized as churches in the historically Black Protestant tradition.

In some cases, information about a congregation's denominational affiliation was insufficient for categorization. For example, some congregations were flagged simply as "Baptist - other" (rather

than “Southern Baptist Convention” or “American Baptist Churches, USA”) or “Methodist - other” (rather than “United Methodist” or “African Methodist Episcopal”).

In those instances, congregations were placed into categories in two ways. First, congregations were categorized based on the Protestant tradition that most group members identify with. Since most Methodists are part of mainline Protestant churches, a Methodist denomination with an ambiguous affiliation was coded into the mainline Protestant category. Second, if the congregation was flagged by InfoGroup as having a mostly African American membership (and the congregation could potentially be considered historically Black Protestant) the denomination was categorized in the historically Black Protestant group.

For example, congregations flagged simply as “Baptist - other” were coded as evangelical Protestant congregations (since most U.S. adults who identify as Baptist are affiliated with evangelical denominations, according to the 2014 U.S. Religious Landscape Study), unless the congregation was flagged as having a mostly African American membership, in which case it was placed in the historically Black Protestant tradition. Similarly, congregations flagged as “Methodist - other” were coded as mainline congregations (since most U.S. adults who identify as Methodist are affiliated with mainline Protestant denominations), unless the congregation was flagged as having a mostly African American membership, in which case it was placed in the historically Black Protestant tradition.

Complete details about how denominations were grouped into traditions are provided in the appendix to the Center’s 2019 report “The Digital Pulpit.”

Identifying and collecting sermons from church websites

Although the database now contained a list of church websites along with data about the characteristics of each congregation, the Center was faced with the challenge of identifying and collecting the sermons posted by these churches online. Researchers designed a custom scraper – a piece of software – for this task. The scraper was designed to navigate through church websites in search of files that appeared to be sermons, download them to a central database and transcribe them from audio to text if needed.

Sampling and weighting

Rather than scrape every church website in the database – which would have taken a great deal of time while offering few statistical benefits – Center researchers scraped the websites of a sample of the entire database. The sample was drawn to ensure adequate representation of each major Christian tradition, as well as congregations that did not match to InfoGroup, for which the Center did not have a tradition or denomination. The Center assigned each record in the database to one of seven strata. The strata were:

- Catholic
- Historically Black Protestant
- Mainline Protestant
- Evangelical Protestant
- Unclassifiable, due to limitations with available data
- Not matched to InfoGroup
- Other – an umbrella category including Buddhist, Mormon, Jehovah's Witness, Jewish, Muslim, Orthodox Christian, Hindu, other Christian or other faiths. (This category was not analyzed on its own, because the original search used only the term "church.")

Researchers then drew a random sample of up to 2,428 records from each stratum, except for historically Black Protestant congregations, of which researchers sample all available congregations to ensure a large enough sample of sermons to analyze.

This pool of sampled records was then screened to distinguish between multi-site congregations that shared a website and duplicative records, so that duplicative ones could be removed. This was done using the following procedure:

- First, researchers removed churches that were found *only* in the first Google Maps collection (see Google Maps section for more details).

- After that, any records with a website that appeared more than five times in the database were excluded on the grounds that these were likely to include denominational content, rather than that of individual congregations.
- For any remaining records with matching websites, researchers took steps to identify and remove duplicate records that referred to the same actual congregation. Two records were considered to be duplicates if they shared a website *and* met any of the following criteria:

    1. Both records were matched to the same congregation in the InfoGroup database.
    2. Both records had the same street address or census block.
    3. One of the two records lacked both a phone number and a building number in its address.

In any of these three instances, the record with the highest match similarity (as measured by the certainty of the matching model) to InfoGroup or, if none matched to InfoGroup, the most complete address information was retained. Congregations that shared a street address but had different websites were not considered to be duplicates but rather distinct congregations that happened to meet in the same location.

The end result was a sample of 19,085 distinct congregations distributed as follows: evangelical (2,428), Catholic (2,428), mainline (2,428), unclassifiable (2,428), unmatched (2,428), historically Black Protestant (4,517), and an agglomerated "small groups" category (2,428). These congregations were then weighted to once more represent their prevalence in the database.

Any statistics stating the share of sermons that mention a topic have a margin of error between 2.8% and 4%. And any statistics stating the share of all congregations hearing at least one sermon that mentions a topic have a margin of error ranging from 4.4% to 5.8%. (These statistics represent the full range of the margin of error, meaning the appropriate confidence interval is plus or minus *half* the amount reported here.)

It is important to note that the estimates in this report are intended to generalize only to the population of churches with websites that were in the original database, and not the entire population of all Christian churches in the United States (which also includes churches that do not have a website or were not listed in Google maps at the time the database was constructed).

How the scraper worked

Researchers made some early decisions about how the scraper should identify sermons:

1. Every sermon, by definition, had to be associated with a date on the website where it was found. This date was interpreted as its delivery date, an interpretation that generally held true.

Number of cases at each stage of data collection

	Number of cases
Institutions identified on Google Places	478,699
Institutions matched with InfoGroup database	262,876
Congregations selected to be scraped	19,085
Congregations from which the scraper successfully identified and downloaded sermons	2,143

Source: Pew Research Center analysis of sermons delivered Aug. 31-Nov. 8, 2020, and available on church websites (N=12,832 sermons from 2,143 churches that posted sermons online).
"Pastors Often Discussed Election, Pandemic and Racism in Fall of 2020"

PEW RESEARCH CENTER

2. Sermons had to be either hosted on the church's website, or shared through a service, such as Vimeo, that was directly linked from that church's website. This was to ensure we did not incorrectly assign a sermon to a church where it was not delivered. Sermons shared solely on YouTube channels were not included in this report due to increased restrictions on the YouTube API.

3. A sermon had to be hosted in a digital media file, rather than written directly into the contents of a webpage. This is because the scraper had no way of determining whether text written into a webpage was, or was not, a sermon. These files could consist of audio (such as an MP3 file), text (such as a PDF), or video (such as a YouTube link).

Identifying sermons involved two main steps: determining which pages to scrape, and then finding media files linked near dates on those pages. These files – digital media files, displayed near dates, on pages likely to contain sermons – were then transcribed to text if needed, and non-sermons were removed.

Determining which pages to examine

To ensure the scraper navigated to the correct pages, researchers trained a machine learning model that estimated how likely a page was to contain sermons. The model relied on the text in and around a page's URL to make its estimate. In addition to the model – which produced a binary, yes-or-no output – the scraper also looked on church webpages for key words specified by researchers, such as "sermon" or "homily."

Based on a combination of the model's output and the keyword searches, pages were assigned a priority ranging from 0-4. The scraper generally examined every page with a priority above zero, and mostly did so in order of priority.

Finding dated media files on pages flagged for further examination

Once the scraper determined that a page was at least somewhat likely to contain sermons, it visited that page and examined its contents in detail in search of files matching the search criteria described above. In some cases, sermons were housed in a protocol such as RSS – a common format for podcasts – that is designed to feed media files directly to computer programs. In those cases, the sermons were extracted directly, with little room for error. The same was true for sermons posted directly to Vimeo accounts.

But in most cases, sermons were embedded or linked directly within the contents of a page. Although these sermons might be easy for humans to identify, they were not designed to be found by a computer. The scraper used three main methods to extract these sermons:

How we trained a model to identify pages with sermons

To identify pages likely to contain sermons, researchers trained a machine learning classifier – a Linear Support Vector Machine – on pages identified by coders as having sermons on them. In September 2018, coders examined a sample of church websites and identified any links that contained sermons dated between July 8 and Sept. 1, 2018. Coders also examined a random sample of links from these same websites and flagged whether or not the links contained any sermons; most of them did not. Taken together, a set of 906 links was compiled from 318 different church websites, 412 of which were determined to contain sermons and 494 of which did not. Using these links, a classifier was trained on the text of each link, along with any text that was associated with the links for those that had been identified by the scraper. Researchers stripped all references to months out of the text for each link before training the model, so it would not develop a bias toward pages containing the words "July," "August" or "September."

The model correctly identified pages with sermons with 0.86 accuracy, 0.86 precision (the share of cases identified as positive that were correct), and 0.83 recall (the share of positive cases correctly identified). Researchers calculated these statistics using a grouped 5-fold cross validation, where links from the same church were not included in both the test and training sets simultaneously.

1. **Using the page's structure:** Webpages are mostly written in HTML, a language that denotes a page's structure and presentation. Pages written with HTML have a clearly denoted hierarchy, in which elements of the page – such as paragraphs, lines or links – are either adjacent to or nested within one another. An element may be next to another element – such as two paragraphs in a block of text – and each also may have elements nested inside them, like pictures or lines.

   The scraper searched for sermons by examining every element of the page to determine if it contained a single human-readable date in a common date format, as well as a single media file.[3]

2. **Using the locations of dates or media files:** In the event that the scraper could not identify a single element with one date and one media file, it resorted to a more creative solution: finding every date and every media file on the page, and clustering them together based on their locations on a simulated computer screen.

   In this solution, the scraper scanned the entire page for any media files – using a slightly more restrictive set of search terms – and any portions of text that constituted a date.[4] The scraper then calculated each element's "x,y" coordinates, using screen pixels as units. Finally, each media file was assigned to its closest date using their Euclidean distance, except in cases where a date was found in the URL for the page or media file itself, in which case that date was assumed to be the correct one.

3. **Using only the text of the media files:** Finally, the scraper also scanned the page for any media files that contained a readable date *in* the text of their URLs. These were directly saved as sermons.

In addition to the above rules that guided the scraper, researchers also placed some restrictions on the program. These were designed to ensure that it did not endlessly scrape extremely large websites or search irrelevant parts of the internet:

- Researchers did not allow the scraper to examine more than five pages from a website other than the one it was sent to search. This rule allowed for limited cases where a church may link

[3] The scraper counted as media any link containing the following combinations of text: ".mp3," ".mp4," ".m4a," ".aif," ".pdf," ".doc." "vimeo," "youtu" but not "channel," both "video" and "embed" together, both "soundcloud" and "player" together, and any of "download," "content-length," "contentSize," or "content-size." The latter three are pieces of information often included in audio or video data.
[4] These included ".mp3", ".mp4", ".pdf", ".doc", "vimeo", "soundcloud" and either "player" or "track", and "youtube" but neither "channel" nor "user".

to an outside website that hosted its sermons, but prevented the scraper from wandering too far afield from the website in question and potentially collecting irrelevant data.

- There were three cases in which the scraper stopped scraping a website before it had examined all of the pages with priorities above zero: 1) if it had examined more than 100 pages since finding any new sermons; 2) if it had been scraping the same website for more than 10 hours, or 3) if the scraper encountered more than 50 timeout errors.
- Some pages were explicitly excluded from being examined. These mainly included links to common social media sites such as Twitter, links to the home page of an external website, or media files themselves, such as MP3 files.
- The scraper always waited between two and seven seconds between downloading pages from the same website to ensure scraping did not overburden the website.

Finally, the scraper removed duplicative files (those found by multiple methods).

Validation and cleaning of scraped files

Researchers conducted a number of steps at various stages of the data collection to clean and validate the scraped files, and to convert them to a machine-readable format that could be used in the subsequent analysis. These steps are described in more detail below.

Removing non-sermons from the collected list of media files

Although the initial scraping process collected dated media files from pages likely to contain sermons, there was no guarantee that these files actually contained sermons. To address this problem, researchers tasked a team of human coders with examining 200 transcribed files – randomly sampled from the database – to determine whether they contained sermons or not. Researchers then trained an extreme gradient boosting model (using the XGBoost package in Python) machine learning model on the results, and used that model to remove non-sermons from the remainder of the database. The model achieved 91% accuracy, 93% recall and 92% precision.

In classifying the files used to train the machine learning model, coders were instructed to consider as a sermon any religious lesson, message or teaching delivered by anyone who appears to be acting as a religious leader in an institution that is at least acting as a religious congregation. They were instructed to *not* include anything that was clearly marked as something other than a sermon (such as a baptism video, Sunday school lesson or religious concert). Sermons with specific audiences (such as a youth sermon) were classified as sermons. In contrast to the Center's 2019 study of online sermons, messages specifically designed for an online audience *were* included in the study if they otherwise met the Center's definition of a sermon.

In determining who qualified as a religious leader, coders could not use the age, gender or race of the speaker, even if there was a reasonable justification for doing so (for instance, a White pastor in a historically Black Protestant denomination, or a woman preaching in a church that ordains only men). Coders were instructed to classify any files that included a sermon along with any other content (such as a song, prayer or reading) as a sermon.

Downloading and transcription

The sermons in the collection varied dramatically in their formatting, audio quality and complexity. Some were complete with podcast-style metadata, while others were uploaded in their raw format. The downloading system attempted to account for this variability by fixing common typographical errors, working around platform-specific formatting or obfuscation, and filling in missing file extensions using other parts of the URL or response headers where possible. Any sermon for which the encoding could be read or guessed was then saved.

Once retrieved, PDFs and other text documents were converted to transcripts with minimal processing using open-source libraries. Multimedia sermons were processed using the FFmpeg multimedia framework to create clean, uniform input for transcription. Video sermons occasionally included subtitles or even different audio streams. When multiple audio streams were available, only the primary English stream was extracted; when an English or unlabeled subtitle stream was available, the first such stream was stored as a distinct type of transcript, but the audio was otherwise handled similarly.

Before transcription could be performed, the extracted media files were normalized to meet the requirements of the transcription service, which imposed constraints on file encoding, size and length. Researchers transcoded all files into the lossless FLAC format and split them into chunks if the file exceeded the service's duration limit. AWS Transcribe returns complex transcripts, including markup that defines each distinct word recognized, the timestamps of the start and end of the word, and the level of confidence in the recognized word.

Evaluating data quality

Evaluating the scraper's performance

To evaluate the scraper's performance, Center researchers manually examined the websites of a random sample of 253 congregations. Each website was assigned a randomly chosen one-week window within the study period, and researchers identified all sermons within that week. The scraper was then deployed to these same websites, and researchers determined whether it had found each sermon identified by researchers.

Of the 128 sermons found by researchers on these NCS church websites, the scraper correctly identified 29. This means the system correctly identified, downloaded and transcribed 23% of all sermons shared on the websites of churches in the database. This relatively low number is partially explained by the fact that the Center was unable to access the YouTube API for this study – of the 128 sermons found by coders, 55 (43%) were shared on YouTube. While some of these may have been directly embedded on church websites – in which case the scraper could find them – many would have been shared only on YouTube channels.

The Center does not view these performance statistics as validating or invalidating the contents of the research. Rather, they are intended to help the reader understand the nature of this limited but interesting window into American religious discourse.

Classifying sermons by topic

To determine whether a sermon discussed any given topic, researchers used machine learning models trained on hand-labeled data. However, any given sermon was generally too long for a researcher to read in one sitting. To address this, researchers divided each sermon into 250-word segments, which were then labeled for each topic of interest. Researchers trained machine learning classifiers (specifically, a distilbert base uncased classifier from the HuggingFace package in Python) to identify each topic. In identifying each topic, every document was coded separately by two coders, and any disagreements were adjudicated by an expert. All performance statistics are based on fivefold cross-validation.

The specific questions answered by coders, any necessary coding notes and the model's performance statistics are as follows:

- **The 2020 election:**
    - Question: "Does this mention the fall 2020 elections, including any references to presidential or congressional candidates, voting, or American political parties?"
    - Precision: 0.86, recall: 0.85
- **The COVID-19 pandemic**:
    - Question: "Does this mention the COVID-19 pandemic, including prevention measures?" (Coders were instructed not to count instances where the speaker

referenced the immediate logistics of holding a service during the pandemic, for instance reminding parishioners to maintain distance in their seats).

- Precision: 0.71, recall: 0.70

- **Racism in America:**

  - Question: "Does the speaker mention racism or racial inequity?"

  - Precision of 0.76, recall: 0.74

Among congregations that discussed the election, researchers trained two more classifiers to identify cases where pastors encouraged congregants to vote, and where pastors discussed specific issues, parties or candidates. Those classifiers performed as follows:

- **Parties, issues or candidates:**

  - Question: "Does the speaker discuss political candidates, parties, or policy issues?" (Only references to candidates, parties or policies in U.S. politics were counted).

  - Precision and recall: 0.79

- **Encouragement to vote:**

  - Question: "Does the speaker explicitly encourage listeners to vote, talk about how they should vote, or speak positively towards voting in a way that clearly encourages it?" (An example of the latter case would be telling listeners to "celebrate their right to vote.")

  - Precision: 0.74, recall: 0.72

Data anomaly during model application

After training and applying the machine learning models, but before beginning analysis, researchers discovered a small number of sermons that appeared twice across the database. Many of these were sermons that were posted twice by the same congregation in different formats – for instance, in both video and audio format. These duplicates were simply removed during analysis; however, a small number of duplicate *sermon segments* remained in the data while researchers were training the classification models. In the case of the main topics – the election, COVID-19,

and racism in America – a single sermon segment appeared twice in the training data out of the 1,356 documents used to train the classifiers. In the case of the subtopics examined *within* discussion of the election, three duplicates appeared out of 599 examples.

Duplicative data could, in theory, inflate a classifier's performance statistics in the event the same document appeared in a training and test set. However, researchers assessed that such a small number of duplicates could not have impacted model performance in any substantial way.

Classifying congregations by religious tradition

Researchers developed a methodology for using the religious denomination information provided by InfoGroup to classify congregations into the major categories (Christian "traditions") used by the Center for analysis. For a full enumeration of how groups were classified, see this appendix from the Center's 2019 report, "The Digital Pulpit."

Acknowledgments

This report was produced by Pew Research Center as part of the Pew-Templeton Global Religious Futures project, which analyzes religious change and its impact on societies around the world. Funding for the Global Religious Futures project comes from The Pew Charitable Trusts and the John Templeton Foundation.

This report is a collaborative effort based on the input and analysis of the following individuals. Find related reports online at pewresearch.org/religion.

Primary Researcher

Dennis Quinn, *Computational Social Scientist*

Research Team

Brian Broderick, *Data Science Engineer*
Gregory A. Smith, *Associate Director of Research*
Emma Remy, *Data Science Analyst*
Besheer Mohamed, *Senior Researcher*
Patrick Van Kessel, *Senior Data Scientist*
Claudia Deane, *Vice President for Research*
Aaron Smith, *Director of Data Labs*
Alan Cooperman, *Director of Religion Research*
Athena Chapekis, *Data Science Assistant*
Meera Ramshankar, *Senior Administrative Coordinator*
Gonzalo Rivero, *Associate Director of Data Labs*
Regina Widjaya, *Computational Social Scientist*
Sono Shah, *Computational Social Scientist*
Rachel Sweeney, *DevOps Engineer*
Ricki Wood, *Content Analyst*
Andrew Perrin, *Research Analyst*

Editorial and Graphic Design

Michael Lipka, *Editorial Manager*
David Kent, *Senior Copy Editor*
Alissa Scheller, *Information Graphics Designer*
Bill Webster, *Information Graphics Designer*
Peter Bell, *Design Director*

Communications and Web Publishing

Stacy Rosenberg, *Associate Director, Digital*
Travis Mitchell, *Digital Producer*
Kelsey Beveridge, *Communications Associate*
Rachel Weisel, *Senior Communications Manager*
Andrew Grant, *Communications Associate*